ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Συγκριτική μελέτη και οπτικοποίηση δεδομένων της πανδημίας Covid19 με χρήση τεχνικών πολυμεταβλητής ανάλυσης

Βασίλειος Ν. Μουζάκης

Διπλωματική εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς
ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην
Εφαρμοσμένη Στατιστική

ΠΕΙΡΑΙΑΣ

ΙΟΥΝΙΟΣ 2023

UNIVERSITY OF PIRAEUS

SCHOOL OF FINANCE AND STATISTICS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Comparative study and visualization of Covid19 epidemic data using multivariate analysis techniques

Vasileios N. Mouzakis

MSc Dissertation
submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

PIRAEUS

JUNE 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίαση του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Χαράλαμπος Ευαγγελάρας (Αναπληρωτής Καθηγητής)
- Αθανάσιος Ρακιντζής (Επίκουρος Καθηγητής)

Abstract

The subject of the present MSc Dissertation is the comparison of the progress and the consequences of the Covid19 pandemic between countries and between continents. The first case of the Covid19 pandemic was recorded in Wuhan, China in December 2019 and since then the personal, social, financial life of all the people of the world changed dramatically.

The aim of this thesis is the presentation of the progress and the consequences of the Covid19 pandemic as well as the comparison of them between countries and between continents by analyzing data. In the first part, we present some studies which have been conducted recently aiming at the comparison of the consequences of the Covid19 pandemic in different countries. In the second part, the theory of the multivariate analysis, which will be used, is presented briefly. In the third part, the multivariate analysis methods, which have been presented in the second part, are implemented in the R programming language and then they are applied to real datasets. The results of the analysis are presented along with several comments/suggestions that may be proved useful for policy makers.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως θέμα την σύγκριση της εξέλιξης και των συνεπειών της πανδημίας Covid19 τόσο σε επίπεδο χωρών όσο και σε επίπεδο ηπείρων. Το πρώτο κρούσμα της πανδημίας Covid19 εμφανίστηκε για πρώτη φορά τον Δεκέμβριο του 2019 στην Κίνα και από τότε η ζωή όλων των ανθρώπων του πλανήτη άλλαξε ριζικά σε προσωπικό, κοινωνικό και οικονομικό επίπεδο.

Ο στόχος της εργασίας είναι η παρουσίαση και σύγκριση των διαφόρων συνεπειών της πανδημίας Covid19 τόσο μεταξύ χωρών όσο και μεταξύ ηπείρων μέσω επεξεργασίας μεγάλου όγκου δεδομένων. Στο πρώτο μέρος παρουσιάζονται διάφορες μελέτες που έχουν γίνει πρόσφατα σε δεδομένα Covid19 με στόχο την σύγκριση των επιπτώσεων της πανδημίας Covid19 σε διαφορετικές χώρες. Στο δεύτερο μέρος παρουσιάζεται συνοπτικά η θεωρία των μεθόδων πολυμεταβλητής ανάλυσης δεδομένων που θα χρησιμοποιηθούν. Στο τρίτο μέρος γίνεται εφαρμογή αυτών των μεθόδων σε πραγματικά δεδομένα με την χρήση της γλώσσας R. Τα αποτελέσματα της ανάλυσης παρουσιάζονται μαζί με μερικά σχόλια/προτάσεις τα οποία μπορεί να φανούν χρήσιμα στους φορείς χάραξης πολιτικής.

Στην οικογένεια μου,

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον επιβλέποντα μου κ. Μάρκο Κούτρα, καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για την ανάθεση του θέματος και την συνεχή καθοδήγηση που μου παρείχε καθ' όλη τη διάρκεια της συγγραφής της διπλωματικής εργασίας ώστε να λάβει την παρούσα μορφή της. Ευχαριστώ επίσης τα μέλη της τριμελούς επιτροπής, για τον χρόνο που αφιέρωσαν στην διόρθωση της εργασίας. Τέλος ευχαριστώ θερμά την οικογένεια μου για την οικονομική και ψυχολογική υποστήριξη που μου έχει προσφέρει.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Εισαγωγή

Η πανδημία Covid19 επηρέασε την υγεία όλου του κόσμου και την παγκόσμια οικονομία. Η ανθρώπινη υγεία επηρεάστηκε καθώς χιλιάδες άνθρωποι είτε αρρώστησαν είτε έχασαν την ζωή τους εξαιτίας της εξάπλωσης της πανδημίας. Τα πιο σύνηθη συμπτώματα αυτής της πολύ διαδεδομένης μόλυνσης είναι πυρετός, κρυολόγημα, βήχας, πόνος στα κόκκαλα, αναπνευστικά προβλήματα τα οποία μπορούν να καταλήξουν σε πνευμονία. Με στόχο την μείωση της εξάπλωσης της πανδημίας Covid19 πολλές χώρες επέβαλλαν εγκλεισμό των κατοίκων στο σπίτι, έκαναν την χρήση μάσκας προσώπου υποχρεωτική, απαγόρευσαν τις συγκεντρώσεις κ.α. Αυτά τα μέτρα εφαρμόστηκαν με στόχο κάθε κράτος να σταματήσει τον ρυθμό αύξησης της καμπύλης μετάδοσης της πανδημίας Covid19 η οποία αυξανόταν εκθετικά. Επιπλέον, λόγω της καραντίνας που εφαρμόστηκε πολλοί άνθρωποι σταμάτησαν να πηγαίνουν στον χώρο εργασίας τους το οποίο είχε σημαντικό αντίκτυπο στις επιχειρήσεις και στο διεθνές εμπόριο. Μερικοί τομείς που ζημειώθηκαν αρκετά ήταν ο κατασκευαστικός τομέας και ο τουρισμός. Με βάση τα παραπάνω καταλαβαίνουμε ότι αξίζει να μελετηθούν και να συγκριθούν οι επιπτώσεις που είχε η πανδημία Covid19 στην ζωή των πολιτών διαφορετικών κρατών.

Κεφάλαιο 1

1. ΣΥΓΚΡΙΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΕΠΙΠΤΩΣΕΩΝ ΤΗΣ ΠΑΝΔΗΜΙΑΣ COVID19 ΣΥΜΦΩΝΑ ΜΕ ΠΡΟΣΦΑΤΕΣ ΜΕΛΕΤΕΣ

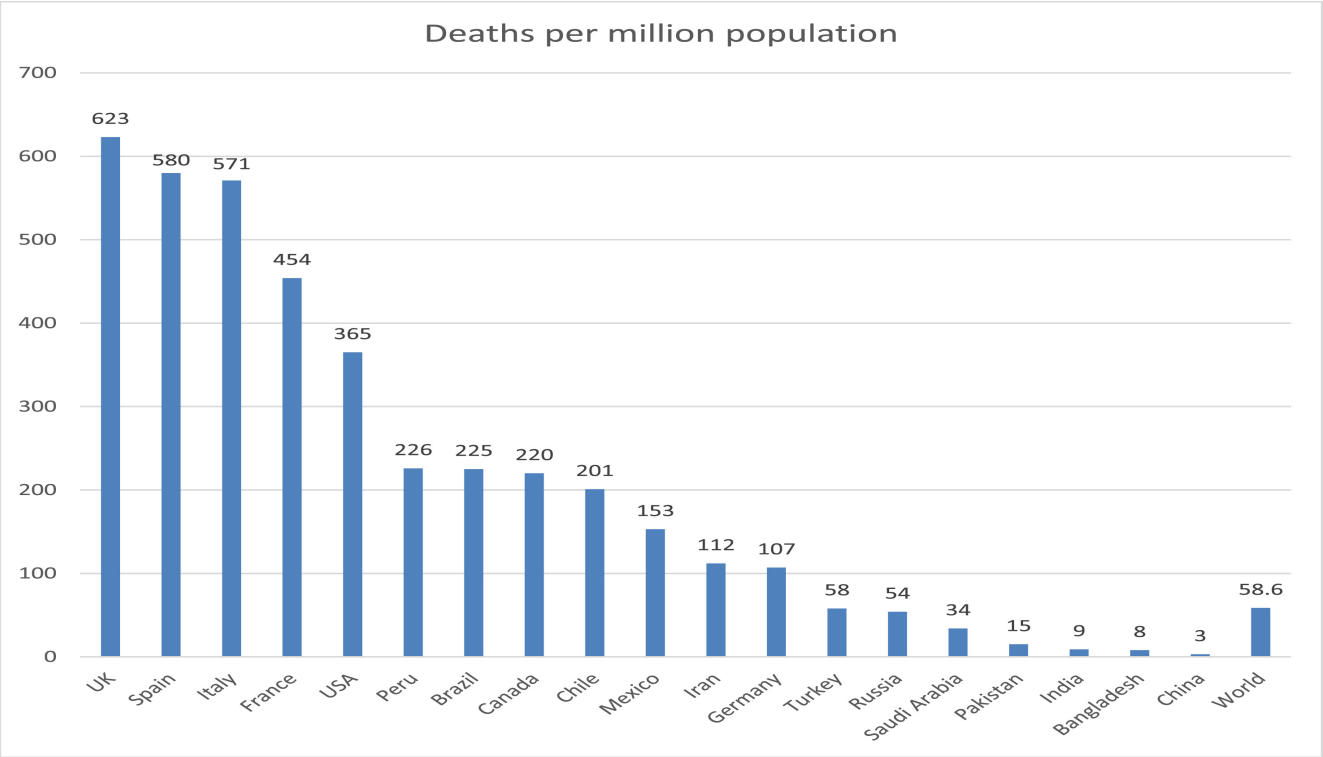
1.1 Επιπτώσεις πανδημίας σε όλο τον κόσμο

Η πανδημία COVID-19 αποτελεί μια συνεχή, ασύμμετρη και απρόβλεπτη απειλή για την ανθρωπότητα, επηρεάζοντας έντονα την ποιότητα της ζωής όλων των πολιτών του κόσμου. Ακόμη και αυτή την χρονική περίοδο που υπάρχουν τα εμβόλια κατά του κορωνοϊού τα οποία έχουν χορηγηθεί σε μεγάλο μέρος του παγκόσμιου πληθυσμού, πολλές χώρες προσπαθούν να αναρρώσουν από τις σφοδρές επιπτώσεις που έχει προκαλέσει. Πιο συγκεριμένα, ο κορωνοϊός είχε τεράστιες επιπτώσεις σε αρκετούς τομείς της ζωής του ανθρώπου.

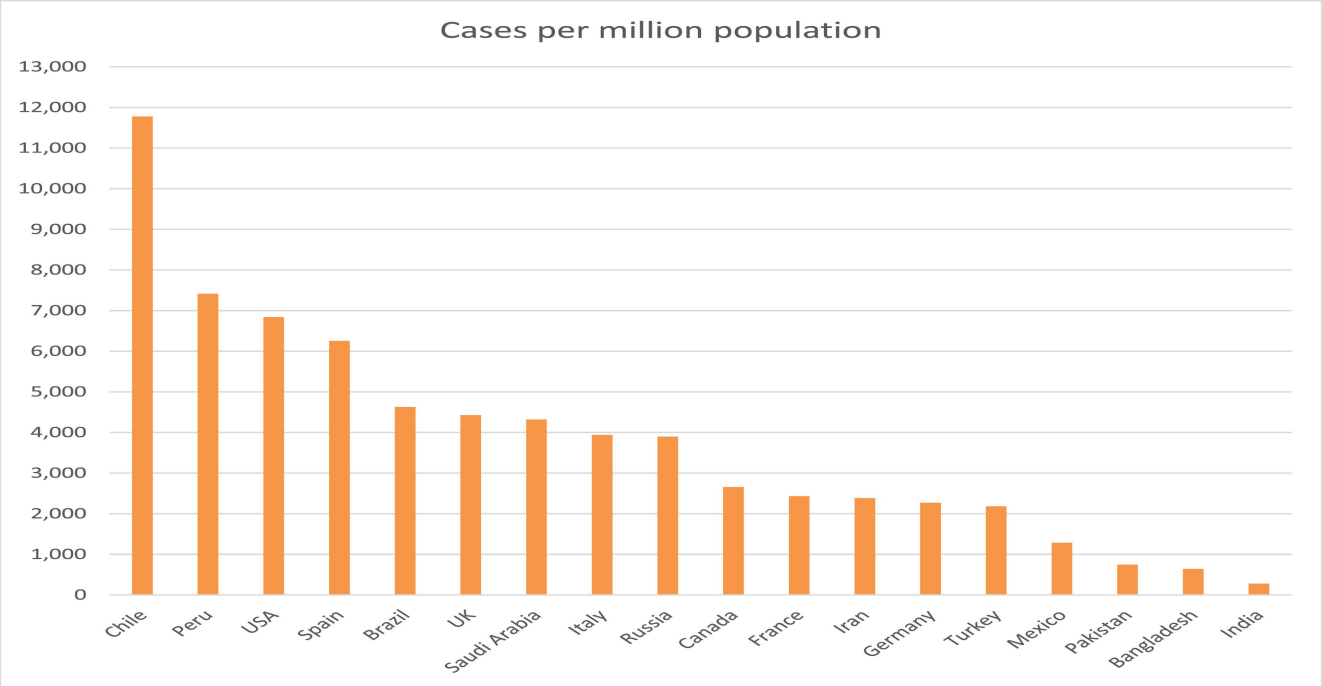
Πρώτα από όλα, ο κορωνοϊός επηρέασε σε μεγάλο βαθμό την υγεία των ανθρώπων όλου του πλανήτη μιας και είναι μια ασθένεια αρκετά μεταδοτική και θανατηφόρα. Ο Παγκόσμιος Οργανισμός Υγείας ανακοίνωσε ότι στις 31 Ιανουαρίου 2020 υπήρχαν 9.847 κρούσματα και 213 θάνατοι, στις 29 Φεβρουαρίου 2020 υπήρχαν 85.961 κρούσματα και 2.941 θάνατοι, στις 31 Μαρτίου 2020 υπήρχαν 754.933 κρούσματα και 36.522 θάνατοι, στις 30 Απριλίου 2020 υπήρχαν 3.096.626 κρούσματα και 217.896 θάνατοι. Οι αναφορές αυτές δείχνουν ότι υπήρξε αύξηση των κρουσμάτων τον Φεβρουάριο κατά 872.97%, τον Μάρτιο κατά 878.22% και τον Απρίλιο κατά 410.18%. Επιπλέον, οι παραπάνω αναφορές δείχνουν αύξηση και στον αριθμό των θανάτων τον Φεβρουάριο κατά 1.387,26%, κατά 1.241,82% τον Μάρτιο και κατά 596,61% τον Απρίλιο. Με βάση αυτά τα ποσοστά παρατηρούμε ότι υπάρχει εκθετική εξάπλωση του ιού ([14]). Είναι επίσης σημαντικό αξιολογήσουμε την πληροφορία που μπορούμε να αντλήσουμε από επιδημιολογικά δεδομένα λαμβάνοντας υπόψιν τον πληθυσμό της κάθε χώρας. Το Σχήμα 1.1 μας δείχνει τον αριθμό των θανάτων ανά εκατομμύριο κατοίκων στις 18 χώρες που είχαν το μεγαλύτερο αριθμό κρουσμάτων μαζί με τον μέσο όρο θανάτων ανά εκατομμύριο ανθρώπων παγκοσμίως και την χώρα στην οποία εμφανίστηκε για πρώτη φορά ο κορονοϊός.

Από το Σχήμα 1.1 βλέπουμε ότι το Ηνωμένο Βασίλειο έχει τους περισσότερους θανάτους ανά εκατομμύριο κατοίκων από ότι οι υπόλοιπες χώρες. Ακόμη, βλέπουμε ότι το Ηνωμένο Βασίλειο ακολουθούν η Ισπανία, η Ιταλία, και η Γαλλία. Επιπροσθέτως, η Γερμανία αποτελεί ακραία παρατήρηση μιας και έχει τον δεύτερο υψηλότερο πληθυσμό σε όλη την Ευρώπη, το οποίο αποδεικνύει ότι τα μέτρα αντιμετώπισης της πανδημίας Covid19 που εφαρμόστηκαν ήταν πολύ πιο αποδοτικά. Επιπλέον, από το Σχήμα 1.1 βλέπουμε ότι η Ινδία είχε μικρό αριθμό θανάτων ανά εκατομμύριο κατοίκων, το οποίο δείχνει ότι η επιβολή των αυστηρών lockdown και η κοινωνική αποστασιοποίηση που επιβλήθηκαν την τελευταία εβδομάδα του Μαρτίου 2020 ήταν αποτελεσματικά ([7]).

Ο αριθμός των κρουσμάτων σε μια χώρα καθορίζει την διασπορά του ιού στους κατοίκους της και ταυτόχρονα την αποτελεσματικότητα των κυβερνητικών μέτρων που έχουν ως στόχο την αντιμετώπιση της πανδημίας Covid19. Το Σχήμα 1.2 δείχνει τον αριθμό κρουσμάτων ανά εκατομμύριο κατοίκων στις 18 χώρες με την χειρότερη επιδημιολογική εικόνα. Από το Σχήμα 1.2 βλέπουμε ότι η Χιλή είναι πρώτη με βάση τον αριθμό κρουσμάτων ανά εκατομμύριο κατοίκων. Αξίζει να αναφερθεί ότι κάποια χώρα μπορεί να εμφανίζει χαμηλό αριθμό κρουσμάτων ανά εκατομμύριο κατοίκων είτε γιατί τα κυβερνητικά μέτρα που εφαρμόζονται με στόχο την αντιμετώπιση της πανδημίας Covid19 είναι αποτελεσματικά είτε λόγω του χαμηλού αριθμού διαγνωστικών ελέγχων Covid19 ([7]).



Σχήμα 1.1: Οπτικοποίηση αριθμού θανάτων ανά εκατομύριο κατοίκων στις 18 χώρες που είχαν το μεγαλύτερο αριθμό κρουσμάτων



Σχήμα 1.2: Αριθμός κρουσμάτων ανά εκατομμύριο κατοίκων στις 18 χώρες με την χειρότερη επιδημιολογική εικόνα

Αναμφισβήτητα, ο κορωνοϊός επηρέασε σε μεγάλο βαθμό και την παγκόσμια οικονομία. Πολλές επιχειρήσεις ήρθαν αντιμέτωπες με μεγάλα ελλείματα και λόγω του κορωνοϊού διαταράχθηκε η εφοδια-

στική αλυσίδα τους. Αυτό έγινε επειδή οι κυβερνήσεις μέσω των μέτρων αντιμετώπισης του κορωνοϊού επέβαλλαν το κλείσιμο των εργοστασίων και τον περιορισμό του εμπορίου. Παρακάτω θα δούμε κάποιες οικονομικές συνέπειες του κορωνοϊού σε διάφορες χώρες.

α. ΗΠΑ: Στις ΗΠΑ ο δείκτης ανεργίας τον Απρίλιο του 2020 έφτασε το 14.7% αφού περισσότεροι από 20 εκατομμύρια Αμερικανοί πολίτες έχασαν την δουλειά τους. Ακόμη, την εβδομάδα που τελείωσε στις 9 Μαΐου του 2020 2.9 εκατομμύρια Αμερικανοί έκαναν αίτηση για να μπορούν να λαμβάνουν το επίδομα ανεργίας εξαιτίας του γεγονότος ότι έχασαν την δουλειά τους. Επιπλέον, το Υπουργείο Εμπορίου των Η.Π.Α ανέφερε στις 29 Απριλίου ότι υπήρξε δραστική μείωση στο ΑΕΠ τους πρώτους 3 μήνες του 2020. Ειδικότερα το ΑΕΠ μειώθηκε κατά 4.8% το πρώτο τρίμηνο του 2020 το οποίο αποτέλεσε την μεγαλύτερη συρρίκνωση από τότε που εμφανίστηκε η παγκόσμια οικονομική κρίση του 2007-2009 ([16]).

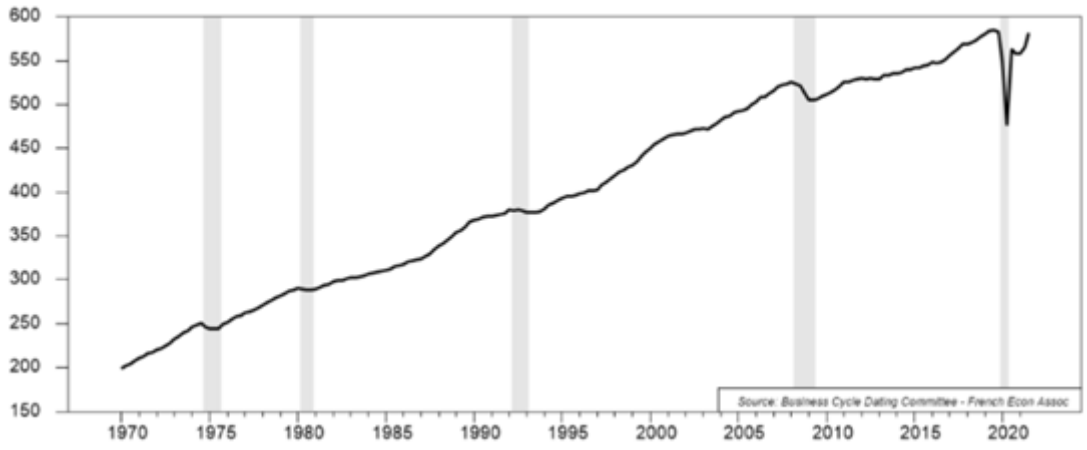
β. Ασία: Στις 15 Απριλίου του 2020 το ΔΝΤ προειδοποίησε ότι οι οικονομίες στην Ασία δεν θα έχουν καμμία ανάπτυξη για το έτος 2020, για πρώτη φορά στα τελευταία 60 χρόνια με τον τρίτο τομέα της οικονομίας (είναι ο τομέας της οικονομίας που εμπεριέχει τις μη καπιταλιστικές επιχειρήσεις και τις ιδιωτικές ή άτυπες ενώσεις, δηλαδή περιέχει οργανώσεις που δεν έχουν ως πρωτεύοντα σκοπό την επιδίωξη κέρδους αλλά δραστηριοποιούνται με γνώμονα την αλληλεγγύη και σκοπό την κοινή ωφέλεια) συγκεκριμένα να βρίσκεται υπό πίεση ([16]).

γ. Κίνα: Το ΑΕΠ της Κίνας της οποίας η οικονομία αποτελεί την δεύτερη μεγαλύτερη οικονομία παγκοσμίως μειώθηκε κατά 6.8% την περίοδο του Ιάνουάριου έως και Μαρτίου, το οποίο είναι μεγαλύτερο από το 6.5% που είχαν προβλέψει οι αναλυτές. Αυτή η μείωση θα μπορούσε κανείς να πει ότι ήταν ξαφνική αν αναλογιστεί ότι το τέταρτο τρίμηνο του 2019 υπήρξε 6% αύξηση του ΑΕΠ της Κίνας. Ακόμη τα δεδομένα που δημοσιεύτηκαν στις 16 Μαρτίου και αφορούσαν τους πρώτους δυο μήνες του 2020 έδειξαν ότι η εργοστασιακή παραγωγή της Κίνας μειώθηκε με τον πιο γρήγορο ρυθμό που υπήρξε μέσα σε τριείς δεκαετίες ([16]).

δ. Ευρώπη: Φυσικά, ο κορωνοϊός γονάτισε και την οικονομία της Ευρώπης, αν αναλογιστούμε ότι η Ευρωπαϊκή Ένωση εφήρμοσε οικονομικά μέτρα στήριξης συνολικής αξίας τουλάχιστον τριών τρισεκαττομυρίων ευρώ ([16]).

ε. Βουλγαρία: Πιο συγκεκριμένα, τα μέτρα αντιμετώπισης για τον κορωνοϊό που επιβλήθηκαν στην Βουλγαρία είχαν σημαντικό αντίκτυπο στην οικονομία της χώρας. Σύμφωνα με τα στατιστικά που δημοσιεύτηκαν από το γραφείο ευρέσεως εργασίας της Βουλγαρίας (Employment Agency of Bulgaria) 87.063 άνθρωποι έμειναν άνεργοι τον Απρίλιο του 2020 από τους οποίους οι 58.744 (οι οποίοι αντιστοιχούν στο 67.4% των ατόμων που έμειναν άνεργοι τον Απρίλιο) δήλωσαν ότι έχασαν την δουλειά τους λόγω του κορωνοϊού. Τον Απρίλιο του 2020 ο δείκτης ανεργίας έφτασε στο 8.9% το οποίο αντιστοιχεί σε αύξηση 3.3 ποσοστιαίων μονάδων από την τιμή αυτού του δείκτη τον Απρίλιο του 2019 (5.6%). Τον Μάϊο του 2020 υπήρξαν σχεδον 300.000 άνεργοι το οποίο αντιστοιχεί σε αύξηση μεγαλύτερη από 68% σε σχέση με τον προηγούμενο χρόνο. Ακόμη, τον ίδιο μήνα 31.478 πολίτες έχασαν τις δουλείες τους. Επιπλέον, μπορούμε να καταλάβουμε το πόσο επηρέασε ο κορωνοϊός την ανεργία στην Βουλγαρία αν αναλογιστούμε ότι το τελευταίο τρίμηνο του 2019 ο συνολικός αριθμός των ανέργων ήταν μικρότερος από 140.000. Επιπροσθέτως, σύμφωνα με το γραφείο ευρέσεως εργασίας της Βουλγαρίας οι κλάδοι που απέλυσαν τους πιο πολλούς υπαλλήλους ήταν: τα ξεχνοδοχεία, τα catering, το εμπόριο και ο κατασκευαστικός κλάδος. Στην κορυφή αυτής της λίστας είναι ο κλάδος που ανήκουν τα επαγγέλματα που προσφέρουν υπηρεσίες αντί για αγαθά (service sector) ο οποίος βασίζεται πολύ στην ανθρώπινη επαφή και για αυτό τον λόγο επηράστηκε πάρα πολύ από το lockdown. Επιπροσθέτως, ο οργανισμός National Network for Children ανέφερε ότι περισσότερα από 6500 παιδιά σε 3200 οικογένειες ζουν σε συνθήκες απόλυτης φτώχειας ([9]).

ζ. Γαλλία: Η γαλλική οικονομία βίωσε μια δυνατή ανάκαψη στο τρίτο τρίμηνο του 2020. Παρόλο που αυτό ήταν κάτι μηχανικό μετά από την μεγάλη ύφεση, η ανάπτυξη του ΑΕΠ παρέμεινε πολύ κοντά στο μηδεν κατά μέσο όρο για τρια συνεχόμενα τρίμηνα σημειώνοντας ακόμη και αρνητικές τιμές το τελευταίο τρίμηνο του 2020. Παρ' όλα αυτά, η ανάρρωση της οικονομίας κατά την διάρκεια του καλοκαιριού του 2021 επιβεβαιώνει ότι η οικονομία αυτή την στιγμή βρίσκεται σε φάση ανάρρωσης από το τρίτο τρίμηνο του 2020.



Σχήμα 1.3: ΑΕΠ και επιχειρηματικός κύκλος στην Γαλλία

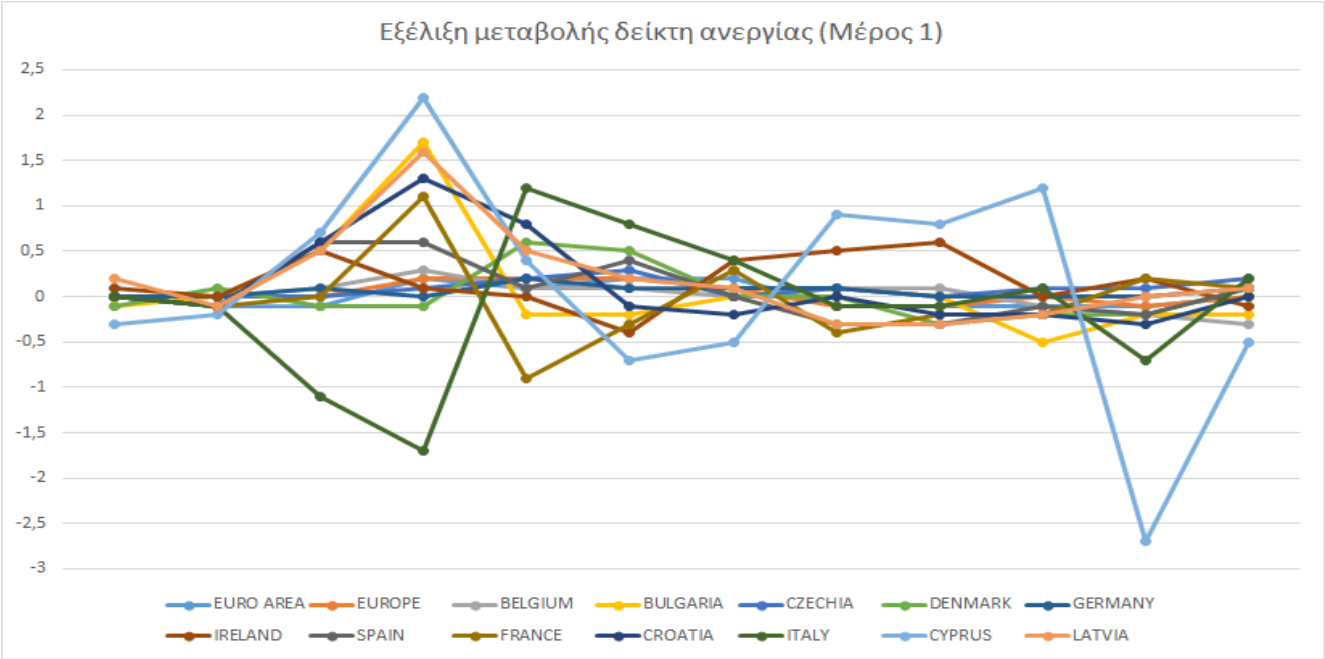
Επιπλέον, το σοκ που βίωσε η γαλλική οικονομία το 2020, δεν ήταν σε καμμία περίπτωση κάτι συνηθισμένο. Πιο συγκεκριμένα, η περίοδος της ύφεσης από την κορύφωση της έως και την λήξη της διήρκησε δυο τρίμηνα γεγονός που την καθιστά την πιο σύντομη που υπήρξε από το 1970 ενώ η μέση διάρκεια μιας ύφεσης είναι τέσσερα τρίμηνα. Παρ' όλ' αυτά, η ύφεση αυτή είναι η μεγαλύτερη από πλευράς βάθους μιας και η μείωση του ΑΕΠ έφτασε το 18.4% ([13]).

	GDP	Duration	Depth	Severity
Peak	1974q3			
Trough	1975q3	4	-2.7	5.4
Peak	1980q1			
Trough	1980q4	3	-0.7	1
Peak	1992q1			
Trough	1993q1	4	-0.9	1.8
Peak	2008q1			
Trough	2009q2	5	-3.9	9.7
Peak	2019q4*			
Trough	2020q2*	2	-18.4	18.4

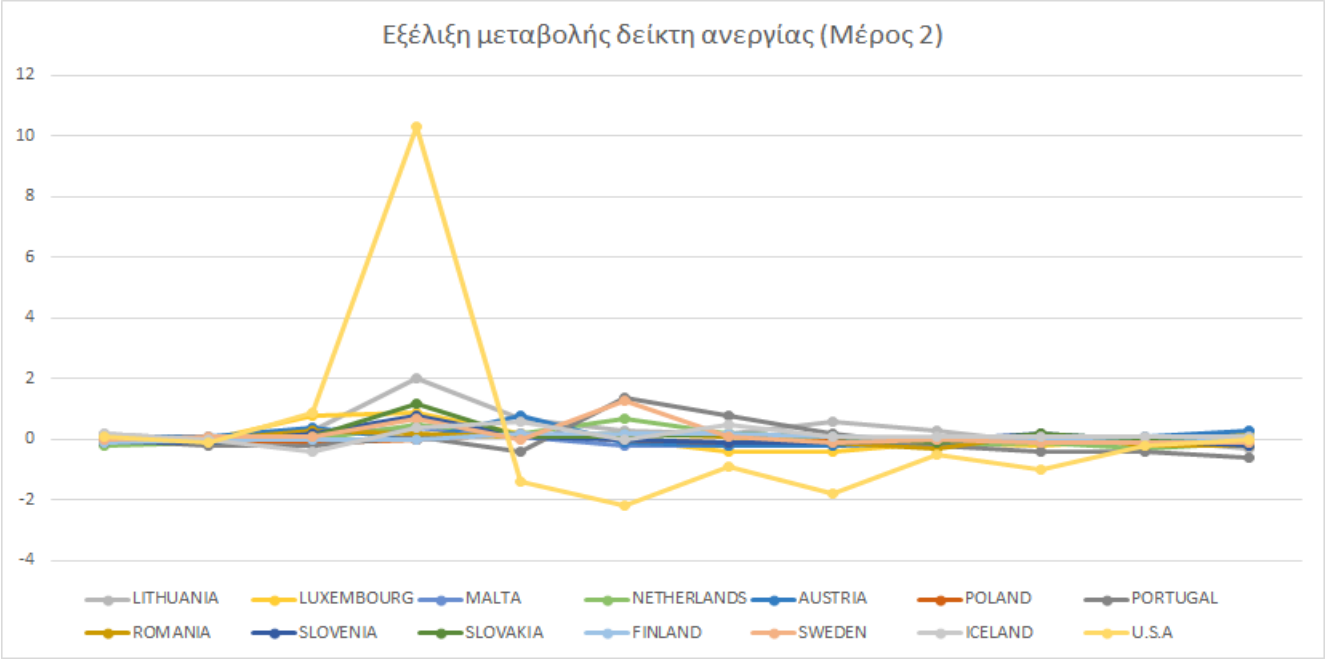
Πίνακας 1.1: Χρονολογική σειρά οικονομικών κύκλων στην Γαλλία

1.2 Σύγκριση Ευρωπαϊκών χωρών και Η.Π.Α ως προς την επίπτωση που είχε ο κορωνοϊός στον δείκτη ανεργίας το 2020

Θα δώσουμε στην συνέχεια το Σχήμα 1.4 και το Σχήμα 1.5 στα οποία θα απεικονίζονται οι μεταβολές στον δείκτη ανεργίας σε κάποιες Ευρωπαϊκές χώρες και στις ΗΠΑ.



Σχήμα 1.4: Εξέλιξη της μεταβολής του δείκτη ανεργίας (Μέρος 1)



Σχήμα 1.5: Εξέλιξη της μεταβολής του δείκτη ανεργίας (Μέρος 2)

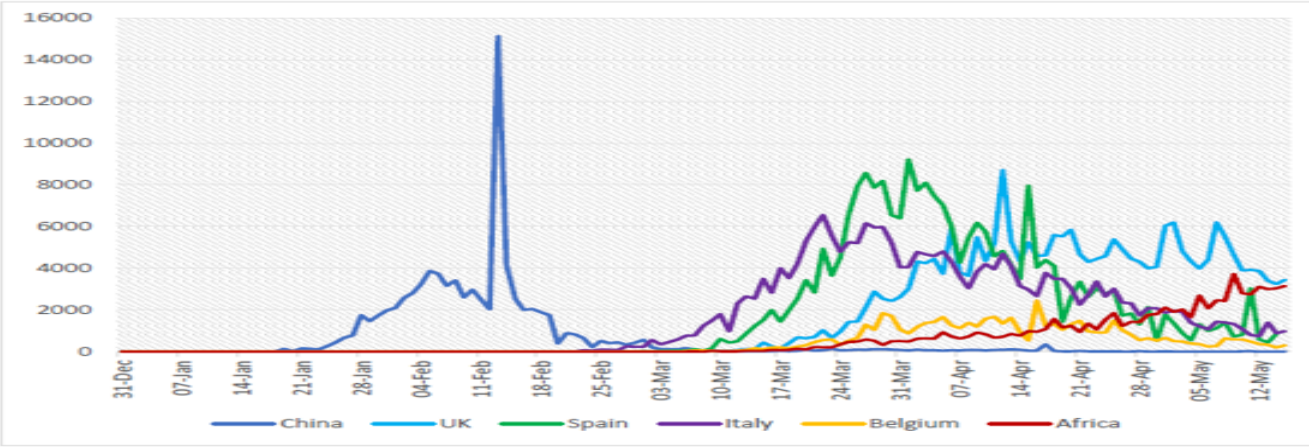
Από τα Σχήματα 1.4,1.5 βλέπουμε ότι ο δείκτης ανεργίας στις Η.Π.Α αυξήθηκε πολύ περισσότερο και πολύ πιο γρήγορα από ότι στις Ευρωπαϊκές χώρες. Επιπλέον, βλέπουμε ότι στις Η.Π.Α ο δείκτης ανεργίας μειώθηκε πολύ πιο σύντομα από ότι στις Ευρωπαϊκές χώρες. Ακόμη, από το Σχήμα 1.4 βλέπουμε ότι στην Ιταλία και στην Κύπρο ο δείκτης ανεργίας αυξήθηκε σύντομα και ότι σε αυτές τις χώρες ο δείκτης ανεργίας άργησε πολύ να σταθεροποιηθεί ([10],[11],[12]).

1.3 Σύγκριση οικονομικών συνεπειών Ευρωπαϊκών κρατών, Κίνας, Η.Π.Α και χωρών της Αφρικής

Πολλά κράτη, με στόχο να μειώσουν την μεταδοτικότητα του κορωνοϊού, έκλεισαν τα σύνορα τους, επέβαλλαν μερικά ή ολικά lockdown στις οικονομίες τους, επέβαλλαν κλείσιμο των επιχηρήσεων τους. Όμως αυτά τα μέτρα είχαν τρομερό αντίκτυπο στην οικονομία της Αφρικής όσον αφορά την μείωση της παραγωγής και του εμπορίου τόσο μέσα όσο και ανάμεσα στις χώρες της. Πιο συγκεκριμένα, αυτά τα μέτρα έχουν φέρει μεγάλη πίεση σε σχεδόν όλους τους αναπτυξιακούς τομείς πολλών οικονομιών και συνεπώς στο συνολικό τους εισόδημα.

Ενώ οι συνέπειες του κορωνοϊού θα μπορούσαν να είναι όμοιες σε περιφερειακό και εθνικό επίπεδο στην Ευρώπη και στην Ασία αναλόγως των τομέων που χτυπήθηκαν, εξαιτίας της έλλειψης οικονομικής αντοχής και επέκτασης, η Αφρική βιώνει πολύ σοβαρές συνέπειες λόγω του κορωνοϊού για πολλούς λόγους. Πρώτα από όλα, οι χώρες της Ευρώπης, η Κίνα και οι Η.Π.Α ξεκίνησαν να έχουν πιο νωρίς τα πρώτα κρούσματα κορωνοϊού από ότι η Αφρική. Με στόχο την καταπολέμηση του κορωνοϊού μείωσαν τις εμπορικές συναλλαγές που είχαν με την Αφρική με αποτέλεσμα να μειωθούν οι εξαγωγές που γίνονται από την Αφρική κάνοντας μεγάλη ζημιά στις χώρες της Αφρικής που έχουν σημαντική συνεισφορά στην παγκόσμια αλυσίδα αξίας. Επιπλέον, τα μέτρα καταπολέμησης του κορωνοϊού σε αυτές τις περιοχές είχαν σαν αποτέλεσμα να μειωθούν αισθητά οι επενδύσεις που γίνονται από το εξωτερικό για την Αφρική.

Ο δεύτερος λόγος για τον οποίο η οικονομία της Αφρικής ζημιώθηκε παραπάνω εξαιτίας του κορωνοϊού από ότι οι οικονομίες των άλλων χωρών, ήταν ότι όταν το ποσοστό μόλυνσης από κορωνοϊό στην Ευρώπη, στην Κίνα, στις Η.Π.Α ξεκίνησε να μειώνεται στην Αφρική γινόταν το αντίθετο. Απόδειξη αυτού είναι το Σχήμα 1.6. Αυτό είχε σαν αποτέλεσμα εκείνη την περίοδο να αίρονται τα μέτρα αντιμετώπισης του κορωνοϊού και να ανοίγουν ξανά πολλές επιχειρήσεις εκτός Αφρικής ενώ στην Αφρική υπήρχε μεγάλη πιθανότητα να υπάρξει κάποια ύφεση μιας και υπήρχε μεγάλη πιθανότητα να μειωθεί η παραγωγή και το εμπόριο, αν το ποσοστό μόλυνσης αυξανόταν ([6]).



Σχήμα 1.6: Ημερήσια κρούσματα Covid19 στην Αφρική, Κίνα και σε επιλεγμένες χώρες της Ευρώπης μέχρι 15/5/20

Κεφάλαιο 2

2. ΣΥΝΟΠΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ ΤΗΣ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ

2.1 Ανάλυση Κυριών Συνιστωσών

Η μέθοδος των κυρίων συνιστωσών (Principal Components Analysis) είναι μια μέθοδος η οποία έχει σκοπό να δημιουργήσει γραμμικούς συνδιασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδιασμοί να είναι ασυσχέτιστοι μεταξύ τους αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Το κέρδος από μια τέτοια διαδικασία είναι πως:

1. Από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, κάτι το οποίο για ορισμένες στατιστικές μεθόδους είναι περισσότερο χρήσιμο.
2. Αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης τότε αυτό σημαίνει πως αντί να έχουμε p μεταβλητές όπως είχαμε αρχικά, έχουμε λιγότερες, με κόστος βέβαια ότι χάνουμε κάποιο (ελπίζουμε μικρό) ποσοστό της συνολικής μεταβλητότητας.
3. Η μέθοδος μας επιτρέπει να αναγνωρίσουμε δίνοντας ονόματα στις καινούργιες μεταβλητές (συνιστώσες) παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές.

Η μέθοδος των κυρίων συνιστωσών στηρίζεται στη Φασματική Ανάλυση ενός τετραγωνικού πίνακα. Αυτό σημαίνει πως μπορούμε να χρησιμοποιήσουμε είτε τον πίνακα διακυμάνσεων είτε τον πίνακα συσχετίσεων που είναι στην ουσία ο πίνακας διακυμάνσεων των τυποποιημένων δεδομένων.

Στην συνέχεια θα περιγράψουμε τον τρόπο με τον οποίο μπορούμε να βρούμε τις κύριες συνιστώσες. Έστω ότι έχουμε ένα σύνολο δεδομένων με k μεταβλητές $(X_1, X_2, \ldots, X_k)$ και δημιουργούμε τις κύριες συνιστώσες $(Y_1, Y_2, \ldots, Y_k)$ οι οποίες είναι γραμμικός συνδιασμός των αρχικών μεταβλητών. Δηλαδή, οι Y_i έχουν την παρακάτω μορφή:

$$
\begin{aligned}
Y_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1k}X_k \\
Y_2 &= \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2k}X_k \\
&\vdots \\
Y_k &= \alpha_{k1}X_1 + \alpha_{k2}X_2 + \cdots + \alpha_{kk}X_k
\end{aligned}
$$

Υπό μορφή πινάκων μπορεί να γραφτεί ως $\boldsymbol{Y} = \boldsymbol{AX}$ όπου $\boldsymbol{Y}, \boldsymbol{X}$ είναι διανύσματα $k \times 1$ και $\boldsymbol{A}$ είναι $k \times k$ πίνακας με στοιχεία:

$$
\boldsymbol{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1k} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{k1} & \alpha_{k2} & \ldots & \alpha_{kk} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha_1} & \boldsymbol{\alpha_2} & \ldots & \boldsymbol{\alpha_k} \end{bmatrix}
$$

όπου $\boldsymbol{\alpha_j}$ είναι το διάνυσμα στήλη με στοιχεία $\boldsymbol{\alpha'_j} = \begin{bmatrix} \alpha_{j1} & \alpha_{j2} & \dots & \alpha_{jk} \end{bmatrix}$ για $j = 1, \dots, k$.

Η διασπορά του συνόλου $\boldsymbol{N}$ κατά μήκος του διανύσματος $\boldsymbol{\alpha_j}$ ορίζεται ως

$$Dis_{\alpha_j}(\boldsymbol{N}) = \boldsymbol{\alpha_j}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\alpha_j}$$

όπου ο πίνακας $\boldsymbol{Z}$ είναι ο πίνακας κεντρικοποιημένων δεδομένων.

Το διάνυσμα $\boldsymbol{\alpha_j}$ θέλουμε να είναι τέτοιο έτσι ώστε να μεγιστοποιείται η ποσότητα $Dis_{\alpha_j}(\boldsymbol{N})$. Είναι όμως φανερό ότι αν πολλαπλασιάσουμε όλες τις συντεταγμένες του διανύσματος $\boldsymbol{\alpha_j}$ με τον ίδιο αριθμό t τότε προκύπτει

$$Dis_{t\boldsymbol{\alpha_j}}(\boldsymbol{N}) = (t\boldsymbol{\alpha_j})'\boldsymbol{Z}'\boldsymbol{Z}(t\boldsymbol{\alpha_j}) = t^2(\boldsymbol{\alpha_j}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\alpha_j}) = t^2 Dis_{\boldsymbol{\alpha_j}}(\boldsymbol{N})$$

Επομένως η διασπορά που προσπαθούμε να μεγιστοποιήσουμε μπορεί να αυξάνεται απεριόριστα και για αυτό τον λόγο θέτουμε ως περιορισμό το διάνυσμα $\boldsymbol{\alpha_j}$ να έχει μέτρο ίσο με 1, δηλαδή να ισχύει $\sum_{i=1}^{k} \alpha_{ji}^2 = \boldsymbol{\alpha'_j}\boldsymbol{\alpha_j} = 1$.

Επομένως το πρόβλημα εύρεσης των κυρίων συνιστωσών είναι το πρόβλημα της εύρεσης των στοιχείων του πίνακα $\boldsymbol{A}$. Έχουμε όμως έναν επιπλέον περιορισμό ο οποίος είναι ότι οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς τη διακυμανσή τους, δηλαδή η πρώτη να έχει τη μεγαλύτερη διακύμανση, η δεύτερη τη δεύτερη μεγαλύρερη διακύμανση κ.ο.κ.

Ας δουλέψουμε για την πρώτη κύρια συνιστώσα $Y_1 = \boldsymbol{\alpha'_1} \cdot \boldsymbol{X}$. Είναι σαφές πως $Var(Y_1) = \boldsymbol{\alpha_1}' \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{\alpha_1}$ όπου $\boldsymbol{\Sigma}$ ο πίνακας διακυμάνσεων του τυχαίου διανύσματος $\boldsymbol{X}$. Επομένως για να βρούμε το $\boldsymbol{\alpha_1}$ θα πρέπει να μεγιστοποιήσουμε την $Var(Y_1)$ με τον περιορισμό πως $\boldsymbol{\alpha_1}' \cdot \boldsymbol{\alpha_1} = 1$ δηλαδή θα μεγιστοποιήσουμε τη συνάρτηση:

$$L(\boldsymbol{\alpha_1}) = \boldsymbol{\alpha_1}' \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{\alpha_1} - \lambda(\boldsymbol{\alpha_1}' \cdot \boldsymbol{\alpha_1} - 1)$$

όπου λ είναι ο πολλαπλασιαστής Lagrange.

Χρησιμοποιώντας παραγώγους διανυσμάτων βρίσκουμε πως

$$\frac{\partial L(\boldsymbol{\alpha_1})}{\partial \boldsymbol{\alpha_1}} = 2(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})\boldsymbol{\alpha_1} = 0$$

και επομένως αντιστοιχεί στο να λύσουμε την εξίσωση:

$$\boldsymbol{\Sigma}\boldsymbol{\alpha_1} = \lambda\boldsymbol{\alpha_1}$$

η οποία είναι η εξίσωση των ιδιοδιανυσμάτων του πίνακα $\boldsymbol{\Sigma}$ όπου λ είναι η ιδιοτιμή. Δηλαδή κάθε ζεύγος ιδιοτιμής και του ιδιοδιανύσματος που τη συνοδεύει είναι λύση της εξίσωσης και άρα έχουμε k δυνατές λύσεις. Από αυτές πρέπει να διαλέξουμε ποια οδηγεί σε μεγαλύτερη διακύμανση. Η διακύμανση του $\boldsymbol{Y}$, θα είναι ίση με λ, και επομένως αρκεί να διαλέξουμε το ζεύγος ιδιοτιμής και ιδιοδιανύσματος που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή.

Με παρόμοια επιχειρήματα μπορούμε να δούμε πως για όλες τις κύριες συνιστώσες τα διανύσματα $\boldsymbol{\alpha_j}$ που χρειαζόμαστε θα αντιστοιχούν στα ιδιοδιανύσματα της j σε φθίνουσα σειρά ιδιοτιμής. Φυσικά για την εύρεση των υπόλοιπων κύριων συνιστωσών χρειάζεται να προσθέσουμε έναν ακόμη περιορισμό: ότι οι κύριες συνιστώσες είναι ασυσχέτιστες με τις προηγούμενες.

Επομένως:

- Για να κατασκευάσουμε τις κύριες συνιστώσες χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα $\boldsymbol{\Sigma}$ που χρησιμοποιούμε.

- Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν στην δεύτερη κύρια συνιστώσα κ.ο.κ.

- Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί. Έτσι αν συμβολίσουμε με λ_j την j μεγαλύτερη ιδιοτιμή τότε ισχύει: $Var(Y_j) = \lambda_j$.

- Όπως είπαμε και πριν οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και άρα ο πίνακας διακυμανσής τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές λ_j.

- Η συνολική διακύμανση των κυρίων συνιστωσών θα είναι η ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών εξαιτίας των ιδιοτητών του ίχνους συμμετρικού και τετραγωνικού πίνακα. Δηλαδή θα ισχύει $tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda})$ και άρα η συνολική διακύμανση διατηρείται.

- Η ποσότητα $\frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i}$ μας δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η j συνιστώσα. Είναι ευνόητο πως αν κάποιος πάρει όλες τις συνιστώσες τότε θα διατηρήσει όλη τη διακύμανση, ενώ αν τελικά παραλείψει κάποιες συνσιτώσες κάποιο ποσοστό διακύμανσης θα χαθεί. Προφανώς συμφέρει να διατηρούμε τις πρώτες συνιστώσες που εξηγούν μεγαλύτερο μέρος της διακύμανσης.

Αν θέλουμε να μην επηρεάζουν οι μονάδες μέτρησης τα αποτελέσματα της αναλυσής μας τότε πρέπει να χρησιμοποιήσουμε τον πίνακα συσχετίσεων αντί του πίνακα διακυμανσής $\mathbf{\Sigma}$. Αυτή είναι μια καλή κίνηση γιατί οι συσχετίσεις δεν αλλάζουν όταν αλλάξουν οι μονάδες μέτρησης ή η κλίμακα. Επίσης, με αυτό τον τρόπο δίνεται το ίδιο βάρος σε όλες τις μεταβλητές καθώς όλα τα στοιχεία της διαγωνίου είναι 1 και άρα τα προβλήματα που δημιουργεί ο πίνακας διακύμανσης μπορούν να ξεπεραστούν.

Η συσχέτιση ανάμεσα στην i κύρια συνιστώσα Y_i και την j αρχική μεταβλητή X_j δίνεται από τον τύπο

$$\rho(Y_i, X_j) = \frac{\alpha_{ij}\sqrt{\lambda_i}}{s_j^2}$$

όπου όπως πριν α_{ij} είναι ο συντελεστής της μεταβλητής X_j στην κύρια συνιστώσα Y_i και s_j^2 είναι η διακύμανση της μεταβλητής X_j. Ο υπολογισμός όλων των συντελεστών συσχέτισης ανάμεσα στις καινούργιες και τις παλιές μεταβλητές είναι σημαντικός για να μπορέσουμε να κάνουμε την ερμηνεία των κυρίων συνιστωσών.

Στην συνέχεια θα περιγράψουμε τα κριτήρια με τα οποία αποφασίζουμε τον αριθμό των συνιστωσών που θα κρατήσουμε. Συνήθως στην πράξη δεν διατηρούμε όλες τις κύριες συνιστώσες αλλά τις πρώτες m από αυτές και αγνοούμε τις υπόλοιπες. Σε μια τέτοια περίπτωση χάνουμε πληροφορία. Το ποσοστό της διακύμανσης των αρχικών δεδομένων που εξηγούμε με τις πρώτες m κύριες συνιστώσες είναι:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_m}{\sum_{i=1}^{k} \lambda_i}$$

Με βάση τα παραπάνω καταλαβαίνουμε πως, αφού βρούμε τις ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα τους θα πρέπει να αποφασίσουμε πόσες κύριες συνιστώσες θα κρατήσουμε και δυστυχώς σε αυτό το ερώτημα δεν υπάρχει εύκολη ούτε κοινώς αποδεκτή απάντηση. Παρακάτω θα παρουσιαστούν μερικά κριτήρια στα οποία μπορούμε να βασιστούμε ώστε να μπορέσουμε να απαντήσουμε σε αυτό το ερώτημα:

- Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

  Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο (π.χ 80%) και διαλέγουμε τόσες συνιστώσες ώστε αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από το στόχο που βάλαμε.

- Κριτήριο Kaiser

  Το κριτήριο αυτό λέει να πάρουμε τόσες ιδιοτιμές όσες είναι μεγαλύτερες από $\bar{\lambda} = \frac{\sum_{j=1}^{k} \lambda_j}{k}$ δηλαδή μεγαλύτερες από τη μέση τιμή των ιδιοτιμών. Στην περίπτωση που δουλεύουμε με πίνακα συσχετίσεων ισχύει $\bar{\lambda} = 1$ και άρα διαλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές που είναι μεγαλύτερες της μονάδας.

- Scree plot

  Το Scree plot είναι ένα γράφημα που έχει στον οριζόντιο άξονα των x τη σειρά και στον κάθετο άξονα των y την τιμή της κάθε ιδιοτιμής. Το κριτήριο αυτό προτείνει να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο, στην ουσία μέχρι να διαπιστώσουμε ότι αρχίζει να αλλάζει η κλίση ([2]).

2.2 Παραγοντική Ανάλυση

Η Παραγοντική Ανάλυση είναι μια στατιστική μέθοδος που έχει σκοπό να βρει την ύπαρξη κοινών παραγόντων ανάμεσα σε ενα σύνολο μεταβλητών το οποίο μελετάμε. Έτσι, εκφράζοντας αυτούς τους παράγοντες (οι οποίοι δεν είναι μια υπαρκτή ποσότητα αλλά την κατασκευάζουμε για τις ανάγκες μας) μπορούμε:

1. να μειώνουμε τις διαστάσεις του προβλήματος. Αντί να δουλεύουμε με τις αρχικές μεταβλητές μπορούμε να δουλέψουμε με λιγότερες αφού οι παράγοντες είναι έτσι κατασκευασμένοι ώστε να διατηρούν όσο γίνεται την πληροφορία που υπήρχε στις μεταβλητές.

2. να δημιουργήσουμε νέες μεταβλητές, τους παράγοντες, στις οποίες μπορούν να αποδοθούν υποκειμενικές ερμηνείες σε μη μετρήσιμες ποσότητες όπως η ευφυΐα ή η ελκυστικότητα ενός προϊόντος στο Μάρκετινγκ.

3. να εξηγήσουμε τις συσχετίσεις που υπάρχουν στα δεδομένα, για τις οποίες έχουμε υποθέσει ότι οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων που δημιούργησαν τα δεδομένα.

Αξίζει να αναφέρουμε το πιο διαδεδομένο μοντέλο Παραγοντικής Ανάλυσης που είναι το ορθογώνιο. Στο ορθογώνιο μοντέλο της Παραγοντικής Ανάλυσης υποθέτουμε πως οι όποιες συσχετίσεις μεταξύ των μεταβλητών οφείλονται αποκλειστικά στην ύπαρξη κάποιων παραγόντων τους οποίους δεν ξέρουμε και θέλουμε να εκτιμήσουμε.

Έτσι, υποθέτουμε πως οι p μεταβλητές μας μπορούν να γραφούν ως γραμμικός συνδιασμός των k παραγόντων, δηλαδή

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon}$$

όπου

- $\boldsymbol{X}$ είναι το διάνυσμα των αρχικών μεταβλητών μεγέθους $p \times 1$ (υποθέτοντας ότι έχουμε p μεταβλητές)

- $\boldsymbol{\mu}$ είναι το διάνυσμα των μέσων μεγέθους $p \times 1$

- $\boldsymbol{L}$ είναι ένας πίνακας $p \times k$ όπου ℓ_{ij} είναι η επιβάρυνση (loading) του παράγοντα F_j στη μεταβλητή X_i

- $\boldsymbol{F}$ είναι ένας $k \times 1$ πίνακας με τους παράγοντες

- $\boldsymbol{\varepsilon}$ είναι το σφάλμα ή μοναδικός παράγοντας. Το σφάλμα ε_i είναι μια ποσότητα που σχετίζεται με την i μεταβλητή και είναι το μέρος της μεταβλητής το οποίο δεν μπορεί να εξηγηθεί από τους παράγοντες.

Μπορούμε να υποθέσουμε πως όλες οι μεταβλητές έχουν μέσο όρο 0 οπότε το διάνυσμα $\boldsymbol{\mu}$ δεν χρειάζεται στο παραπάνω μοντέλο (αυτό μπορεί να επιτευχθεί εύκολα αφαιρώντας από κάθε μεταβλητή τη μέση της τιμή). Επίσης είναι προφανές ότι $k < p$, δηλαδή ο αριθμός των παραγόντων πρέπει να είναι μικρότερος του αριθμού των μεταβλητών γιατί αλλιώς δεν θα είχε νόημα η Παραγοντική Ανάλυση. Σύμφωνα με τα παραπάνω, υποθέτουμε ότι κάθε μεταβλητή μπορούμε να την γράψουμε στη μορφή:

$$\begin{aligned} X_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1k}F_k + \varepsilon_1 \\ X_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2k}F_k + \varepsilon_2 \\ &\vdots \\ X_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pk}F_k + \varepsilon_p. \end{aligned}$$

Αξίζει να αναφερθεί ότι:

- Το παραπάνω μοντέλο αν και μοιαζει με ένα γραμμικό μοντέλο έχει μερικές διαφορές. Κατ' αρχάς τα X_i δεν είναι παρατηρήσεις αλλά μεταβλητές. Αφετέρου το δεξί μέλος της εξίσωσης δεν είναι παρατηρήσιμο και έτσι πρέπει να εκτιμηθεί.
- Οι παράγοντες F_i μπορούν να γραφτούν και αυτοί ως γραμμικός συνδιασμός των μεταβλητών. Αυτό είναι χρήσιμο να γίνεται όταν θέλουμε να δημιουργήσουμε νέες μεταβλητές. Οι συντελεστές κάθε παράγοντα όταν εκφράζουμε τις μεταβλητές ως γραμμικό συνδιασμό των παραγόντων καλούνται επιβαρύνσεις ενώ αντίστοιχα οι συντελεστές κάθε μεταβλητής όταν εκφράζουμε κάθε παράγοντα ως γραμμικό συνδιασμό των μεταβλητών καλούνται συντελεστές των score (factor scores coefficients).
- Οι παράγοντες έχουν την ίδια διακύμανση.

Ένα πολύ βασικό κομμάτι του παραγοντικού μοντέλου είναι οι υποθέσεις που πρέπει να γίνουν. Αυτές είναι:

1. $\mathbb{E}(\boldsymbol{F}) = \mathbf{0}$
2. $\mathrm{Cov}(\boldsymbol{F}) = \mathbb{I}$ (δηλαδή οι παράγοντες μεταξύ τους είναι ορθογώνιοι)
3. $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$
4. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ όπου $\boldsymbol{\Psi}$ είναι ένας διαγώνιος της μορφής:

$$\begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \psi_p \end{bmatrix}$$

5. $\mathrm{Cov}(\varepsilon_i, F_j) = \mathbf{0}$ για κάθε $i \neq j$

Από τις παραπάνω υποθέσεις του μοντέλου προκύπτει

$$\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{X}) = \text{Cov}(\boldsymbol{LF} + \boldsymbol{\varepsilon}) = \boldsymbol{L}\text{Cov}(\boldsymbol{F})\boldsymbol{L}' + \text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{LL}' + \boldsymbol{\Psi}$$

καθώς από τις υποθέσεις του μοντέλου η συνδιακύμανση μεταξύ $\boldsymbol{F}$ και $\boldsymbol{\varepsilon}$ είναι μηδέν. Συνεπώς, βλέπουμε πως ο πίνακας διακύμανσης μπορεί να διασπαστεί σε δυο μέρη, το πρώτο είναι το κομμάτι που ερμηνεύουν οι κοινοί παράγοντες και ονομάζεται κοινή διασπορά (communality) και το δεύτερο είναι το κομμάτι που οφείλεται στους μοναδικούς παράγοντες και άρα το μοντέλο δεν μπορεί να ερμηνεύσει και ονομάζεται ιδιαιτερότητα (specificity).

Στην Παραγοντική Ανάλυση σκοπός μας είναι να εκτιμήσουμε τους πίνακες $\boldsymbol{L}$ και $\boldsymbol{\Psi}$, να αναπαραστήσουμε δηλαδή τον πίνακα διακύμανσης του πληθυσμού.

Θα πρέπει να γνωρίζουμε πως η Παραγοντική Ανάλυση αλλά και άλλες τεχνικές πολυμεταβλητής ανάλυσης δεδομένων έχουν καλά αποτελέσματα όταν οι μεταβλητές είναι υψηλά συσχετισμένες. Οπότε είναι καλό να ελένξουμε αν τα δεδομένα που καλούμαστε να αναλύσουμε είναι πολύ συσχετισμένα ή όχι για να κρίνουμε αν είναι ωφέλιμο να εφαρμόσουμε μια μέθοδο πολυμεταβλητής ανάλυσης ή όχι. Ένα μέτρο για να συγκρίνουμε το σχετικό μέγεθος των συντελεστών συσχέτισης σχετικά με τους μερικούς συντελεστές συσχέτισης είναι το Kaiser-Meyer-Olkin στατιστικό που υπολογίζεται από τον τύπο

$$KMO = \frac{\sum \sum_{i \neq j} r_{ij}^2}{\sum \sum_{i \neq j} r_{ij}^2 + \sum \sum_{i \neq j} \alpha_{ij}^2}$$

όπου r_{ij} και α_{ij} είναι οι δειγματικοί συντελεστές συσχέτισης και μερικής συσχέτισης αντίστοιχα. Αν η τιμή του είναι μεγάλη τότε τα δεδομένα μας είναι κατάλληλα για Παραγοντική Ανάλυση. Τιμές κάτω από 0.5 είναι πολύ κακές τιμές. Στην πράξη τιμές γύρω στο 0.8 θεωρούνται αρκετά καλές για να προχωρήσουμε.Μικρότερες τιμές αποτελούν ένδειξη ότι η Παραγοντική Ανάλυση δεν θα μας δώσει ικανοποιητικά αποτελέσματα.

Στην συνέχεια, θα δούμε τους τρόπους με τους οποίους μπορούμε να αποφασίσουμε τον αριθμό των παραγόντων που θα υπάρχουν στην ανάλυση μας καθώς και πως τους εκτιμάμε. Ένα από τα βασικά ερωτήματα στην Παραγοντική Ανάλυση είναι ο καθορισμός του αριθμού των παραγόντων που θα χρησιμοποιήσουμε. Όπως είπαμε και προηγουμένως ο αριθμός αυτός δεν είναι γνωστός και υπάρχουν διάφορες μέθοδοι για να εκτιμηθεί. Πολλά στατιστικά πακέτα επιτρέπουν στον ερευνητή να καθορίσει εκ των προτέρων τον αριθμό αυτό αλλά γενικά αυτό γίνεται κυρίως για λόγους ευκολίας.

Για να βρεθεί ο αριθμός των παραγόντων ο ερευνητής μπορεί να χρησιμοποιήσει παρόμοιες τεχνικές με αυτές που είδαμε στην Ανάλυση Κυρίων Συνιστωσών. Δηλαδή, τις τιμές των ιδιοτιμών του πίνακα διακύμανσης - συνδιακύμανσης, τιμές που εξηγούν κάποιο ποσοστό της διακύμανσης ή το Scree plot.

Είναι ευνόητο ότι ο αριθμός των παραγόντων χρειάζεται να καθοριστεί πριν γίνει η εκτίμηση τους. Επομένως κάποιος θα μπορούσε να εργαστεί με διαδοχικά αυξανόμενο αριθμό παραγόντων και να κρατήσει το μοντέλο με βάση κάποιο κριτήριο καλής προσαρμογής τέτοια κριτήρια είναι:

- Από τον πίνακα των επιβαρύνσεων μπορεί κάποιος να εκτιμήσει τον πίνακα $\boldsymbol{\Sigma}$. Οι αποκλίσεις του πραγματικού πίνακα με τον εκτιμημένο (συνήθως ονομάζεται reproduced matrix) θα πρέπει να είναι μικρές.
- Έλεγχος λόγου πιθανοφανειών αν οι εκτιμήσεις έχουν γίνει με την μέθοδο της μέγιστης πιθανοφάνειας. Τέτοιοι έλεγχοι στηρίζονται σε υποθέσεις για την κατανομή του πληθυσμού.

Αξίζει να σημειωθεί πως:

- Η ερμηνεία των παραγόντων μπορεί να εξαρτάται και από τον αριθμό τους, δηλαδή προσθέτοντας παράγοντες αυτοί να παύουν να έχουν την ίδια ερμηνεία.

- Για μερικές μεθόδους εκτίμησης υπάρχει περιορισμός στον αριθμό των παραγόντων που μπορούν να εκτιμηθούν.

Οι δυο βασικές μέθοδοι εκτίμησης που χρησιμοποιούνται στην πράξη είναι η μέθοδος των κυρίων συνιστωσών και η μέθοδος μέγιστης πιθανοφάνειας, τις οποίες θα δούμε στην συνέχεια αναλυτικά.

1. **Εκτίμηση με την μέθοδο των κυρίων συνιστωσών**

   Η εκτίμηση με τη μέθοδο των κυρίων συνιστωσών βασίζεται στη Φασματική Ανάλυση του πίνακα διακύμανσης (συσχέτισης). Όταν λέμε πως θέλουμε να εκτιμήσουμε τις παραμέτρους του παραγοντικού μοντέλου εννοούμε πως θέλουμε να εκτιμήσουμε τα στοιχεία του πίνακα επιβαρύνσεων $\boldsymbol{L}$ και τα στοιχεί της διαγωνίου του πίνακα $\boldsymbol{\Psi}$. Επιπλέον, το πλήθος των στοιχείων του πίνακα $\boldsymbol{L}$ έχει να κάνει με το πλήθος των παραγόντων που έχουμε υποθέσει πως υπάρχουν. Επομένως, ο σκοπός μας είναι να βρούμε πίνακες $\hat{\boldsymbol{L}}, \hat{\boldsymbol{\Psi}}$ για τους οποίους ο πίνακας $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}}$ να είναι κοντά στον πίνακα δειγματικής διακύμανσης (ή στον πίνακα συσχέτισης αν εργαζόμαστε με τυποποιημένα δεδομένα).

   Από την Φασματική Ανάλυση ενός πίνακα διακύμανσης γνωρίζουμε πως μπορούμε να γράψουμε τον πίνακα $\boldsymbol{\Sigma}$ στη μορφή $\boldsymbol{\Sigma} = \boldsymbol{AA}'$ όπου $\boldsymbol{A} = \boldsymbol{\Pi\Lambda}^{1/2}$, $\boldsymbol{\Lambda}$ είναι ο διαγώνιος πίνακας που περιέχει στη διαγώνιο τις ιδιοτιμές και $\boldsymbol{\Pi}$ είναι ο πίνακας με στήλες τα ιδιοδιανύσματα του πίνακα $\boldsymbol{\Sigma}$. Επομένως, αν χρησιμοποιήσουμε ως $\hat{\boldsymbol{\Lambda}} = \boldsymbol{\Pi\Lambda}^{1/2}$ τότε μπορούμε να αναπαραστήσουμε πλήρως τον πίνακα $\boldsymbol{\Sigma}$. Στην πράξη δουλεύουμε με το δειγματικό πίνακα διακύμανσης $\boldsymbol{S}$.

   Αν το πλήθος των παραγόντων είναι k συμπίπτει με το πλήθος p των μεταβλητών, επιτυγχάνουμε την πλήρη αναπαράσταση του δειγματικού πίνακα διακύμανσης (συσχέτισης) και επομένως οι εκτιμήσεις των ψ_i είναι 0, δηλαδή οι παράγοντες εξηγούν όλη τη διακύμανση.

   Αν $k < p$ τότε ο πίνακας $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}'$ δεν μπορεί να αναπαραστήσει πλήρως τον αρχικό πίνακα διακύμανσης. Έτσι σε αυτή την περίπτωση μπορούμε να εκτιμήσουμε και τις ιδιαιτερότητες ως:

$$\hat{\psi}_i = s_i^2 - \sum_{j=1}^{p} L_{ij}^2$$

   όπου L_{ij} είναι το στοιχείο της i-οστής γραμμής και j-οστής στήλης του πίνακα $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}'$, δηλαδή η επιβάρυνση του j παράγοντα στην i μεταβλητή, $j = 1, \ldots, k$ και $i = 1, \ldots, p$. Ο δεύτερος όρος στο δεξί μέλος της ισότητας είναι η κοινή διασπορά της μεταβλητής.

   Μερικές χρήσιμες παρατηρήσεις σχετικά με τη μέθοδο κυρίων συνιστωσών είναι οι εξής:

   - Αν χρησιμοποιήσουμε πολλούς παράγοντες ($k \cong p$) μπορούμε να αναπαραστήσουμε πλήρως τον αρχικό πίνακα. Σε αυτή όμως την περίπτωση δεν έχουμε κερδίσει κάτι σημαντικό αφού χρησιμοποιήσαμε πολλούς παράγοντες και στην ουσία απλά μετασχηματίσαμε τα δεδομένα μας.
   - Δεν υπάρχει περιορισμός ως προς τον αριθμό των παραγόντων που μπορούμε να εκτιμήσουμε με τη μέθοδο κυρίων συνιστωσών.

2. **Εκτίμηση με την μέθοδο της μέγιστης πιθανοφάνειας**

   Για να χρησιμοποιήσουμε τη μέθοδο της μέγιστης πιθανοφάνειας χρειάζεται να κάνουμε κάποιες υποθέσεις σχετικά με τον πληθυσμό από όπου προήλθαν τα δεδομένα μας. Συγκεκριμένα υποθέτουμε πως τα σφάλματα ακολουθούν Πολυμεταβλητή Κανονική Κατανομή με διάνυσμα μέσων το μηδενικό διάνυσμα και πίνακα διακύμανσης το διαγώνιο πίνακα $\boldsymbol{\Psi}$, δηλαδή $\boldsymbol{\varepsilon} \sim N_p(\boldsymbol{0}, \boldsymbol{\Psi})$. Επομένως το διάνυσμα των τυχαίων μεταβλητών $\boldsymbol{X}$ δοθέντος του διανύματος των παραγόντων

ακολουθεί Πολυδιάστατη Κανονική Κατανομή, δηλαδή $\boldsymbol{X} \mid \boldsymbol{F} \sim N_p(\boldsymbol{LF}, \boldsymbol{\Psi})$ και άρα αν υποθέσουμε πως και οι παράγοντες προέρχονται από Πολυδιάστατη Κανονική Κατανομή, δηλαδή $\boldsymbol{F} \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ προκύπτει πως $\boldsymbol{X} \sim N_k(\boldsymbol{LF}, \boldsymbol{LL'} + \boldsymbol{\Psi})$

Αν έχουμε ένα δείγμα από Πολυμεταβλητή Κανονική Κατανομή μπορεί να δειχτεί ότι η πιθανοφάνεια είναι συνάρτηση του πίνακα διακύμανσης $\boldsymbol{\Sigma}$ του πληθυσμού, πιο συγκεκριμένα

$$\ell(\boldsymbol{X}, \boldsymbol{\Sigma}) = -\frac{n}{2}\left[p \ln(2\pi) + \ln|\boldsymbol{\Sigma}| + tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right]$$

όπου n είναι το μέγεθος του δείγματος, p ο αριθμός των μεταβλητών και $\boldsymbol{\Sigma}$ ο δειγματικός πίνακας διακυμάνσεων. Στον παραπάνω τύπο δεν περιέχεται το διάνυσμα των μέσων $\boldsymbol{\mu}$ αφού αυτό δεν επηρεάζει το μοντέλο μας (ή ισοδύναμα έχουμε κεντρικοποιήσει όλες τις μεταβλητές ώστε να έχουν μέση τιμή 0). Για να εκτιμήσουμε το μοντέλο με τη μέθοδο μέγιστης πιθανοφάνειας πρέπει να μεγιστοποιήσουμε τη συνάρτηση

$$\ell(\boldsymbol{X}, \boldsymbol{L}, \boldsymbol{\Psi}) = -\frac{n}{2}\left[p \ln(2\pi) + \ln|(\boldsymbol{LL'} + \boldsymbol{\Psi})| + tr((\boldsymbol{LL'} + \boldsymbol{\Psi})^{-1} \cdot \boldsymbol{S})\right]$$

ως προς $\boldsymbol{L}$ και $\boldsymbol{\Psi}$. Αν το μοντέλο έχει k παράγοντες τότε ο πίνακας $\boldsymbol{L}$ έχει $p \times k$ στοιχεία ενώ ο πίνακας $\boldsymbol{\Psi}$ επειδή είναι διαγώνιος έχει p στοιχεία. Συνολικά έχουμε $(p+1) \cdot k$ παραμέτρους ενώ ο πίνακας $\boldsymbol{\Sigma}$ από όπου ξεκινάμε έχει $\frac{p}{2} \cdot (p+1)$ διαφορετικά στοιχεία. Για να έχει λύση λοιπόν πρέπει να βάλλουμε περιορισμό στο k δηλαδή στον αριθμό των παραγόντων που μπορούμε να εκτιμήσουμε. Επομένως, με τη μέθοδο μέγιστης πιθανοφάνειας υπάρχει περιορισμός στον αριθμό των παραγόντων που μπορούμε να εκτιμήσουμε κάτι το οποίο δεν υπήρχε στην μέθοδο κυρίων συνιστωσών. Επίσης χρειαζόμαστε έναν ακόμα περιορισμό. Αυτό που συνήθως χρησιμοποιείται (και που χρησιμοποιούν τα περισσότερα στατιστικά πακέτα) είναι ότι ο πίνακας $\boldsymbol{L'\Psi}^{-1}\boldsymbol{L}$ είναι διαγώνιος και τα στοιχεία του είναι σε φθίνουσα σειρά. Η μεγιστοποίηση της πιθανοφάνειας γίνεται με εφαρμογή αριθμητικών μεθόδων και χρήση κριτηρίων τερματισμού.

Θα περιγράψουμε στην συνέχεια την τεχνική της περιστροφής των παραγόντων η οποία μας βοηθάει στην ερμηνεία των παραγόντων. Με την περιστροφή των παραγόντων πετυχαίνουμε οι παράγοντες να έχουν πιο εύκολη ερμηνεία. Με την περιστροφή δεν αλλάζουν κάποια από τα χαρακτηριστικά του μοντέλου όπως η καλή του προσαρμοστηκότητα και το ποσό της διακύμανσης που ερμηνεύει το μοντέλο παρά μόνο οι τιμές των επιβαρύνσεων. Γενικά αν $\boldsymbol{L}$ είναι ένας πίνακας που περιέχει τις επιβαρύνσεις και $\boldsymbol{G}$ είναι ένας ορθογώνιος πίνακας τότε ισχύει πως

$$\boldsymbol{LG}(\boldsymbol{LG})' = \boldsymbol{LGG'L'} = \boldsymbol{LL'}$$

και επομένως και ο πίνακας $\boldsymbol{LG}$ μπορεί να θεωρηθεί ως ένας πίνακας επιβαρύνσεων. Μαθηματικά ο πίνακας $\boldsymbol{G}$ ορίζει έναν ορθογώνιο μετασχηματισμό.

Κάνοντας περιστροφή μπορούμε να εξασφαλίσουμε οι επιβαρύνσεις κάποιων παραγόντων να είναι μεγάλες σε απόλυτη κλίμακα μόνο για κάποιες από τις μεταβλητές και έτσι βλέποντας ποιες μεταβλητές εξαρτώνται από τους παράγοντες να μπορέσουμε να δώσουμε μια ερμηνεία σε αυτούς. Οι πιο συνηθισμένες μέθοδοι περιστροφής είναι:

- Varimax: Προσπαθεί να ελαχιστοποιήσει τον αριθμό των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε παράγοντα.
- Quartimax: Προσπαθεί να ελαχιστοποιήσει τον αριθμό παραγόντων που εξηγούν μια μεταβλητή.
- Equimax: Συνδιασμός των Varimax και Quartimax.

Στην συνέχεια θα δούμε πως μπορούμε να υπολογίσουμε τα scores των παραγόντων. Όπως είπαμε και προηγουμένως ένας από τους σκοπούς της Παραγοντικής Ανάλυσης είναι να μειώσει τον αριθμό των μεταβλητών. Για να επιτευχθεί αυτό μπορούμε να δημιουργήσουμε καινούργιες μεταβλητές, τους παράγοντες, ως γραμμικούς συνδασμούς των αρχικών μεταβλητών. Κάθε παράγοντας μπορεί να γραφεί στη μορφή:

$$\begin{aligned} F_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p \\ F_2 &= \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p \\ &\vdots \\ F_k &= \alpha_{k1}X_1 + \alpha_{k2}X_2 + \cdots + \alpha_{kp}X_p \end{aligned}$$

Οι συντελεστές α_{ij} είναι το score της μεταβλητής X_j στον παράγοντα F_i και δεν πρέπει να συγχέονται με τις επιβαρύνσεις. Όταν το μοντέλο έχει εκτιμηθεί με τη μέθοδο των κυρίων συνιστωσών οι παράγοντες είναι ακριβείς, δηλαδή μπορούν να υπολογιστούν χωρίς σφάλμα. Αντίθετα, στα μοντέλα που είναι εκτιμημένα με τη μέθοδο μέγιστης πιθανοφάνειας χρησιμοποιούνται προσεγγιστικές μέθοδοι. Αξίζει να σημειωθεί πως οι νέες μεταβλητές θα έχουν μέση τιμή 0 και θα είναι ασυσχέτιστες, δεδομένου πως το μοντέλο είναι ορθογώνιο.

Έχοντας εκτιμήσει ένα παραγοντικό μοντέλο και έστω $\boldsymbol{L}$ και $\boldsymbol{\Psi}$ οι εκτιμήσεις μας για τις παραμέτρους αυτού (πριν ή μετά την περιστροφή) μπορούμε να βρούμε τα factor scores δηλαδή τις τιμές των καινούργιων μεταβλητών για κάθε μεταβλητή. Οι μέθοδοι που προσφέρονται είναι πολλές. Αυτές που είναι διαθέσιμες στα περισσότερα στατιστικά πακέτα είναι οι εξής (βλ. [2])

- Regression method: Το διάνυσμα $\boldsymbol{F}$ των καινούργιων μεταβλητών υπολογίζεται μέσω του τύπου:

  $$\boldsymbol{F} = (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'\boldsymbol{X}$$

  Η μέθοδος αυτή βασίζεται στη μέθοδο ελαχίστων τετραγώνων ανάμεσα στις πραγματικές τιμές και αυτές που το παραγοντικό μοντέλο προβλέπει.

- Barlett method: Σε σχέση με την παραπάνω μέθοδο, ο Barlett πρότεινε αντί να χρησιμοποιήσει κανείς την απλή μέθοδο ελαχίστων τετραγώνων να χρησιμοποιήσει την αρχή των σταθμισμένων ελαχίστων τετραγώνων καθώς η διακύμανση δεν είναι η ίδια για όλα τα σφάλματα ε_i. Η μέθοδος εκτίμησης εκτιμά τους παράγοντες ως:

  $$\boldsymbol{F} = (\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}$$

- Η μέθοδος του Anderson: Η μέθοδος αυτή χρησιμοποιεί τον τύπο

  $$\boldsymbol{F} = (\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{L})(\boldsymbol{I} + \boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1/2}\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}$$

2.3 Διαχωριστική Ανάλυση

Η βασική ιδέα της διαχωριστικής ανάλυσης είναι να κατατάξει παρατηρήσεις σε γνωστούς πληθυσμούς με γνωστές κατανομές για κάθε πληθυσμό.

Ας υποθέσουμε ότι έχουμε K πληθυσμούς $\Pi_1, \Pi_2, \ldots, \Pi_K$ όπου $K \geq 2$. Τότε για κάθε πληθυσμό Π_k έχουμε και μια κατανομή $f_k(x)$. Σκοπός της διαχωριστικής συνάρτησης είναι να διαχωρίσει ή να

κατανείμει κάθε παρατήρηση στους K γνωστούς πληθυσμούς - ομάδες. Προφανώς ψάχνουμε για ένα διαχωριστικό κανόνα που μπορεί να κατατάξει σωστά όσο το δυνατόν περισσότερες παρατηρήσεις. Για την εύρεση ενός διαχωριστικού κανόνα θα χρειαστεί να καθορίσουμε τις περιοχές $R_1, R_2, \dots R_K$ οι οποίες αποτελούν διαμέριση του συνόλου των δυνατών αποτελεσμάτων για την τυχαία μεταβλητή $\boldsymbol{X}$ (δειγματικό χώρο της $\boldsymbol{X}$).

Είναι σημαντικό πριν προχωρήσουμε σε οποιοδήποτε κριτήριο ταξινόμησης να ορίσουμε την πιθανότητα λανθασμένης και ορθής ταξινόμησης.

Γενικά η πιθανότητα ταξινόμησης της παρατήρησης $\boldsymbol{x}$ στον πληθυσμό Π_i δεδομένου ότι η παρατήρηση ανήκει στον πληθυσμό Π_r είναι:

$$\mathbb{P}(i, r) := \mathbb{P}(\boldsymbol{X} \in R_i | \boldsymbol{X} \sim f_r(\boldsymbol{x})).$$

- Αν $i \neq r$ τότε η παραπάνω πιθανότητα αποτελεί την πιθανότητα λανθασμένης ταξινόμησης της παρατήρησης $\boldsymbol{x}$ η οποία ταξινομείται στον πληθυσμό Π_i ενώ κανονικά ανήκει στον πληθυσμό Π_r.
- Αν $i = r$ τότε η παραπάνω πιθανότητα αποτελεί την πιθανότητα ορθής ταξινόμησης της παρατήρησης $\boldsymbol{x}$ στον πληθυσμό Π_i.

Στην συνέχεια θα δούμε πως μπορούμε να ταξινομήσουμε παρατηρήσεις σε K πληθυσμούς με βάση τον κανόνα της μέγιστης πιθανοφάνειας. Δεδομένου ότι οι συναρτήσεις πυκνότητας - πιθανότητας ή συναρτήσεις πιθανότητας $f_i(\boldsymbol{x})$ (ανάλογα αν έχουμε συνεχείς ή διακριτές κατανομές) είναι γνωστές για κάθε πληθυσμό i η συνάρτηση πιθανοφάνειας είναι:

$$L(i) = f(\boldsymbol{x}|i) \; \forall \; i = 1, \dots, k$$

Σύμφωνα με τον κανόνα μέγιστης πιθανοφάνειας η παρατήρηση $\boldsymbol{x}$ κατατάσσεται στον πληθυσμό r για τον οποίο ισχύει:

$$L(r) = \max_{1 \leq i \leq k} L(i).$$

Αν οι πληθυσμοί είναι δυο και οι παρατηρήσεις του πρώτου πληθυσμού προέρχονται από $N(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}_1)$ και του δεύτερου προέρχονται από $N(\boldsymbol{\mu_2}, \boldsymbol{\Sigma}_2)$, τότε η πιθανοφάνεια για τον i πληθυσμό δίνεται από τον τύπο:

$$L(i) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_i|^2 \cdot \exp(-(\boldsymbol{x} - \boldsymbol{\mu_i})' \boldsymbol{\Sigma_i}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i})/2)$$

όπου p το πλήθος των μεταβλητών.

Επειδή ο λογάριθμος είναι αύξουσα συνάρτηση η μεγιστοποίηση της παραπάνω πιθανοφάνειας είναι ισοδύναμη με την μεγιστοποίηση της συνάρτησης:

$$\log(L(i)) = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_i| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_i})' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}) \right)$$

Αν αγνοήσουμε τον σταθερό όρο $-\frac{p}{2} \log(2\pi)$ ελαχιστοποιώντας τον όρο $\log(|\boldsymbol{\Sigma}_i| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_i})' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}))$ παρατηρούμε ότι ο κανόνας διαχωρισμού με βάση το κριτήριο της μέγιστης πιθανοφάνειας για τους παραπάνω πληθυσμούς είναι:

- Αν $\log\left(|\boldsymbol{\Sigma}_1| + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_1})' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu_1}) \right) < \log\left(|\boldsymbol{\Sigma}_2| + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_2})' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu_2}) \right)$ τότε η παρατήρηση $\boldsymbol{x}$ κατατάσσεται στον πληθυσμό Π_1.
- Αν $\log\left(|\boldsymbol{\Sigma}_2| + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_2})' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu_2}) \right) < \log\left(|\boldsymbol{\Sigma}_1| + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_1})' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu_1}) \right)$ τότε η παρατήρηση $\boldsymbol{x}$ κατατάσσεται στον πληθυσμό Π_2.

Αξίζει να αναφέρουμε πως:

- Ο όρος $(\boldsymbol{x}-\boldsymbol{\mu_1})'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})$ είναι η απόσταση Mahalanobis μεταξύ της παρατήρησης $\boldsymbol{x}$ και $\boldsymbol{\mu_1}$ στο τετράγωνο.
- Ο όρος $(\boldsymbol{x}-\boldsymbol{\mu_2})'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})$ είναι η απόσταση Mahalanobis μεταξύ της παρατήρησης $\boldsymbol{x}$ και $\boldsymbol{\mu_2}$ στο τετράγωνο.

Στην συνέχεια θα δούμε πως μπορούμε να ταξινομήσουμε παρατηρήσεις με βάση το κριτήριο ελαχιστοποίησης του αναμενόμενου κόστους λανθασμένης ταξινόμησης. Το βασικό στοιχείο της Διαχωριστικής Ανάλυσης είναι η δημιουργία κανόνων απόφασης σχετικά με την κατάταξη παρατηρήσεων σε διάφορους πληθυσμούς. Επομένως, στην ουσία έχουμε να αντιμετωπίσουμε ένα πρόβλημα θεωρίας αποφάσεων. Όταν μπορούμε να ποσοτικοποιήσουμε τις απώλειες λόγω λανθασμένης κατάταξης μπορούμε να γράψουμε το αναμενόμενο κόστος ταξινόμησης που προέρχεται από την r ομάδα ως εξής:

$$ECM_r = \pi_r \sum_{i=1}^{k} C(i|r)\mathbb{P}(i|r)$$

όπου

- $C(i|r)$ είναι το κόστος να κατατάξουμε την παρατήρηση στην i ομάδα ενώ ανήκει στην r.
- $\mathbb{P}(i|r)$ είναι η πιθανότητα να κατατάξουμε την παρατήρηση στην i ομάδα ενώ ανήκει στην r.
- π_r είναι η εκ των προτέρων πιθανότητα να ανήκει μια παρατήρηση στον πληθυσμό r.

Το συνολικό κόστος είναι ίσο με το άθροισμα των επιμέρους ECM_r. Φυσικά επιλέγουμε να κατατάξουμε την παρατήρηση στην ομάδα με το μικρότερο αναμενόμενο κόστος λανθασμένης κατάταξης το οποίο είναι ισοδύναμο με την ελαχιστοποίηση του συνολικού κόστους λανθασμένης κατάταξης.

Όταν έχουμε δυο πληθυσμούς τότε:

$$\begin{aligned} ECM &= C(2|1)\pi_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + C(1|2)\pi_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= C(2|1)\pi_1 \left(1 - \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x}\right) + C(1|2)\pi_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= C(2|1)\pi_1 - \int_{R_1} [C(2|1)\pi_1 f_1(\boldsymbol{x}) - C(1|2)\pi_2 f_2(\boldsymbol{x})] \end{aligned}$$

Επομένως,

$$R_1 : \boldsymbol{x} \text{ τέτοια ώστε να ισχύει } C(2|1)\pi_1 f_1(\boldsymbol{x}) - C(1|2)\pi_2 f_2(\boldsymbol{x}) > 0$$

ή ισοδύναμα

$$R_1 : \boldsymbol{x} \text{ τέτοια ώστε να ισχύει } \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}$$

δηλαδή

- αν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}$ τότε κατατάσσουμε την $\boldsymbol{x}$ παρατήρηση στην πρώτη ομάδα.

- αν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \leq \frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}$ τότε κατατάσσουμε την $\boldsymbol{x}$ παρατήρηση στην δεύτερη ομάδα.

όπου $\boldsymbol{x}$ είναι το διάνυσμα με τα χαρακτηριστικά (μεταβλητές) μιας παρατήρησης, $C(1|2)$ είναι το κόστος που προέρχεται από την λανθασμένη καταχώρηση μιας παρατήρησης στην πρώτη ομάδα ενώ πραγματικά ανήκει στην δεύτερη και $C(2|1)$ είναι το κόστος που προέρχεται από την λανθασμένη καταχώρηση μιας παρατήρησης στην δεύτερη ομάδα ενώ πραγματικά ανήκει στην πρώτη.

Θα δούμε στην συνέχεια πως θα φτάσουμε στη δημιουργία μιας διαχωριστικής συνάρτησης υποθέτωντας κανονικότητα των πληθυσμών. Η πιο συχνή επιλογή για την κατανομή των δεδομένων μιας ομάδας είναι η Πολυμεταβλητή Κανονική Κατανομή. Αρχικά θα θεωρήσουμε ότι οι πίνακες συνδιακύμανσης $\boldsymbol{\Sigma}$ είναι ίσοι και ότι κάθε ομάδα διαφέρει μόνο ως προς τις μέσες τιμές $\boldsymbol{\mu_k}$. Έτσι για την παρατήρηση $\boldsymbol{x_i}$ που ανήκει στο k πληθυσμό έχουμε:

$$f_k(\boldsymbol{x_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2} \cdot |\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{\mu_k})' \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k})$$

Παίρνοντας λογάριθμους στον διαχωριστικό κανόνα έχουμε:

$$\ln \frac{f_1(\boldsymbol{x_i}|\boldsymbol{\mu_1}, \boldsymbol{\Sigma})}{f_2(\boldsymbol{x_i}|\boldsymbol{\mu_2}, \boldsymbol{\Sigma})} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}\boldsymbol{x_i} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})$$

Άρα ο διαχωριστικός κανόνας είναι:

- Αν $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}\boldsymbol{x_i} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) \geq \ln\left(\frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}\right)$ η παρατήρηση $\boldsymbol{x_i}$ ταξινομείται στον πληθυσμό Π_1.
- Αν $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}\boldsymbol{x_i} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) < \ln\left(\frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}\right)$ η παρατήρηση $\boldsymbol{x_i}$ ταξινομείται στον πληθυσμό Π_2.

Σε περίπτωση που οι πίνακες διακύμανσης των δυο πληθυσμών δεν είναι ίσοι, ο κανόνας απόφασης γίνεται:

- Αν $-\frac{1}{2}\boldsymbol{x_i'}\left(\Sigma_1' - \Sigma_2'\right)\boldsymbol{x_i} + \left(\mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1}\right)\boldsymbol{x_i'} - k \geq \ln\left(\frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}\right)$ τότε η παρατήρηση $\boldsymbol{x_i}$ ταξινομείται στον πληθυσμό Π_1.
- Αν $-\frac{1}{2}\boldsymbol{x_i'}\left(\Sigma_1' - \Sigma_2'\right)\boldsymbol{x_i} + \left(\mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1}\right)\boldsymbol{x_i'} - k < \ln\left(\frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}\right)$ τότε η παρατήρηση $\boldsymbol{x_i}$ ταξινομείται στον πληθυσμό Π_2.

όπου $k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right)$

Αξίζει να αναφέρουμε ότι σε πραγματικές εφαρμογές δεν γνωρίζουμε τις τιμές των $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}$ για αυτό τον λόγο παίρνουμε τις εκτιμήσεις τους δηλαδή τα $\boldsymbol{\bar{x}_1}, \boldsymbol{\bar{x}_2}, \boldsymbol{S_1}, \boldsymbol{S_2}$. Επιπλέον, όταν $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$ αντί των $\boldsymbol{S_1}, \boldsymbol{S_2}$ χρησιμοποιούμε τον S_{pooled} ο οποίος είναι ίσος με:

$$S_{pooled} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}S_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}S_2$$

Άρα ο κανόνας απόφασης γίνεται:

- Αν $(\boldsymbol{\bar{x}_1} - \boldsymbol{\bar{x}_2})'\boldsymbol{S_{pooled}}^{-1}\boldsymbol{x_i} - \frac{1}{2}(\boldsymbol{\bar{x}_1} - \boldsymbol{\bar{x}_2})'\boldsymbol{S_{pooled}^{-1}}(\boldsymbol{\bar{x}_1} + \boldsymbol{\bar{x}_2}) \geq \ln\left(\frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}\right)$ η παρατήρηση $\boldsymbol{x_i}$ ταξινομείται στον πληθυσμό Π_1.
- Αν $(\boldsymbol{\bar{x}_1} - \boldsymbol{\bar{x}_2})'\boldsymbol{S_{pooled}}^{-1}\boldsymbol{x_i} - \frac{1}{2}(\boldsymbol{\bar{x}_1} - \boldsymbol{\bar{x}_2})'\boldsymbol{S_{pooled}^{-1}}(\boldsymbol{\bar{x}_1} + \boldsymbol{\bar{x}_2}) < \ln\left(\frac{\pi_2}{\pi_1} \cdot \frac{C(1|2)}{C(2|1)}\right)$ η παρατήρηση $\boldsymbol{x_i}$ ταξινομείται στον πληθυσμό Π_2.

(Πηγές: [2],[1])

2.4 Ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες είναι μια μέθοδος που έχει ως σκοπό να κατατάξει ένα σύνολο παρατηρήσεων σε ομάδες χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Για να το πετύχει εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών με σκοπό να δημιουργήσει ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Μια επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς και παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

Δυο βασικές έννοιες για την ανάλυση κατά συστάδες είναι οι έννοιες της απόστασης και της ομοιότητας. Οι δυο αυτές έννοιες είναι αντίθετες αφού παρατηρήσεις που είναι όμοιες έχουν μεγάλη ομοιότητα και μικρή απόσταση.

Οι βασικότερες και πιο διαδεδομένες προσεγγίσεις είναι:

- Ιεραρχικές μέθοδοι: Ξεκινάμε με κάθε παρατήρηση να είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις που έχουν την πιο μικρή απόσταση. Αν δυο παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα τότε ενώνουμε μια προϋπάρχουσα ομάδα με μια παρατήρηση μέχρι να φτιάξουμε μια ομάδα. Κοιτώντας τα αποτελέσματα διαλέγουμε στις πόσες ομάδες θα σταματήσουμε. Οι ιεραρχικές μέθοδοι με την σειρά τους χωρίζονται σε συσσωρευτικές και διαιρετικές μεθόδους.
- Μη ιεραρχικές μέθοδοι: Θεωρούν ένα πλήθος N παρατηρήσεων και ένα πλήθος k συστάδων και διαμερίζουν τα σημεία στις συστάδες. Τυπικά, το πλήθος των συστάδων k προκαθορίζεται από τον χρήστη. Ξεκινώντας από έναν αρχικό διαχωρισμό, με μια επαναληπτική διαδικασία, τα σημεία μετακινούνται από μια συστάδα σε μια άλλη. Ο σχηματισμός των συστάδων γίνεται με τέτοιο τρόπο, ώστε να βελτιστοποιείται ένα κριτήριο διαχωρισμού. Στόχος είναι να δημιουργηθούν συστάδες οι οποίες περιέχουν όμοια αντικείμενα ενώ τα αντικείμενα διαφορετικών συστάδων να είναι ανόμοια.

Στην συνέχεια θα δούμε την έννοια της απόστασης. Η απόσταση είναι μια θεμελιώδης έννοια στην Πολυμεταβλητή Ανάλυση και όχι μόνο για την ανάλυση δεδομένων. Σκοπός της απόστασης είναι να μετρήσει πόσο απέχουν δυο παρατηρήσεις.

Έστω δυο διανύσματα $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$ και $\boldsymbol{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jp}) \in \mathbb{R}^p$. Για να είναι μια συνάρτηση $d_{ij} = d(\boldsymbol{x_i}, \boldsymbol{x_j})$ συνάρτηση απόστασης θα πρέπει να ικανοποιεί τις εξής ιδιότητες:

- $d_{ij} \geq 0 \; \forall \; i, j$ και για $i = j \iff d_{ij} = 0$
- $d_{ij} \leq d_{is} + d_{sj}$
- $d_{ij} = d_{ji}$

Δίνονται στην συνέχεια τα πιο συνηθισμένα μέτρα απόστασης για δυο διανύσματα $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$ και $\boldsymbol{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jp}) \in \mathbb{R}^p$

- Ευκλείδια απόσταση:

$$d_{ij} = d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

- Απόσταση του Pearson:

$$d_{ij} = d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\sum_{r=1}^{p} \left(\frac{x_{ir} - x_{jr}}{s_r} \right)^2}$$

όπου $s_r = \left(\frac{1}{p-1} \sum_{i=1}^{p} (x_{ir} - \bar{x_r})^2 \right)^{\frac{1}{2}}$

- Απόσταση Manhattan:

$$d_{ij} = \sum_{r=1}^{p} |x_{ir} - x_{jr}|$$

- Απόσταση Minkowski:

$$d_{ij} = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}|^{\lambda} \right)^{\frac{1}{\lambda}}$$

όπου $\lambda \geq 1$ δεδομένη παράμετρος.

- Απόσταση Chebyshev:

$$d_{ij} = \max_{1 \leq r \leq p} |x_{ir} - x_{jr}|$$

Στην συνέχεια θα δούμε την έννοια της ομοιότητας. Τα μέτρα ομοιότητας μπορεί να χρησιμοποιηθούν για να μας δείξουν αν δυο παρατηρήσεις είναι όμοιες ή ανόμοιες μεταξύ τους. Δηλαδή αναμένουμε μεγάλη τιμή στο μέτρο ομοιότητας για τις παρατηρήσεις που μοιάζουν πολύ, ενώ για τις ανόμοιες παρατηρήσεις αναμένουμε μικρή.

Ας υποθέσουμε ότι για κάθε ζεύγος παρατηρήσεων $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$ και $\boldsymbol{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jp}) \in \mathbb{R}^p$ ορίζεται ένας πραγματικός αριθμός $s_{ij} = s(\boldsymbol{x_i}, \boldsymbol{x_j})$ που είναι η ομοιότητα ανάμεσα στην i και j παρατήρηση, έτσι ώστε να ισχύουν οι παρακάτω τρεις ιδιότητες:

- $s_{ij} \geq 0 \; \forall \; i, j$ και για $i = j \Rightarrow \; s_{ij} = 1$
- $s_{ij} \leq 1$
- $s_{ij} = s_{ji}$

Το πιο γνωστό μέτρο ομοιότητας για ποσοτικές παρατηρήσεις είναι η απόλυτη τιμή του δειγματικού συντελεστή συσχέτισης του οποίου ο τύπος είναι:

$$s_{ij} = \frac{\sum_{r=1}^{p} (x_{ir} - \bar{x_{i\cdot}})(x_{jr} - \bar{x_{j\cdot}})}{\left(\sum_{r=1}^{p} (x_{ir} - \bar{x_{i\cdot}})^2 \sum_{r=1}^{p} (x_{jr} - \bar{x_{j\cdot}})^2 \right)^{1/2}}$$

όπου $\bar{x_{i\cdot}} = \frac{1}{p} \sum_{r=1}^{p} x_{ir}$ και $\bar{x_{j\cdot}} = \frac{1}{p} \sum_{r=1}^{p} x_{jr}$

Ας υποθέσουμε ότι τα δεδομένα που μας ενδιαφέρουν έχουν διχοτομικές μεταβλητές, δηλαδή για την κάθε παρατήρηση έχουμε μια δυαδική αναπαράσταση, όπου ο αριθμός 1 θα υποδηλώνει την παρουσία αυτού του χαρακτηριστικού/μεταβλητής και το 0 την απουσία του.

Για κάθε δυο παρατηρήσεις/αντικείμενα x και y κατασκευάζουμε έναν πίνακα ομοιότητας έτσι ώστε να μπορέσουμε να υπολογίσουμε κάποια μέτρα ομοιότητας.

α	b
c	d

Πίνακας 2.1: Πίνακας ομοιότητας

όπου:

- α : το πλήθος των μεταβλητών που το αντικείμενο x έχει την τιμή 1 και το y την τιμή 1
- b: το πλήθος των μεταβλητών που το αντικείμενο x έχει την τιμή 1 και το y την τιμή 0
- c: το πλήθος των μεταβλητών που το αντικείμενο x έχει την τιμή 0 και το y την τιμή 1
- d: το πλήθος των μεταβλητών που το αντικείμενο x έχει την τιμή 0 και το y την τιμή 0

Παρακάτω δίνονται μερικά μέτρα ομοιότητας.

- Απλός συντελεστής αντιστοίχισης (Sokal and Michener, 1958)

$$s(x, y) = \frac{a + d}{a + b + c + d}$$

- Rogers and Tarimoto (1960)

$$s(x, y) = \frac{a + d}{(a + d) + 2(b + c)}$$

- Sokal and Sneath (1963)

$$s(x, y) = \frac{2(a + d)}{2(a + d) + (b + c)}$$

- Jaccard (1908)

$$s(x, y) = \frac{a}{a + b + c}$$

- Dice and Sorensen (1948)

$$s(x, y) = \frac{2a}{2a + b + c}$$

- Sokal and Sneath (1963)

$$\frac{a}{a + 2(b + c)}$$

Έστω τώρα ότι οι μεταβλητές μας είναι ονομαστικές με πολλές διαφορετικές τιμές οι οποίες δεν επιδέχονται καμμιά μορφή ιεράρχησης. Εαν x, y οι παρατηρήσεις που εξετάζουμε ως προς τις p μεταβλητές και έστω ότι οι τιμές τους συμπίπτουν σε m από αυτά τα γνωρίσματα τότε ο συντελεστής ομοιότητας ορίζεται από την σχέση:

$$s(x, y) = \frac{m}{p}$$

2.5 Συσσωρευτικές μέθοδοι συσταδοποίησης

Στο πρώτο βήμα ενός συσσωρευτικού αλγορίθμου θεωρούμε ότι η κάθε παρατήρηση αποτελεί μια ομάδα και στη συνέχεια βρίσκουμε τις δυο πλησιέστερες παρατηρήσεις εντοπίζοντας στον πίνακα αποστάσεων D που εμφανίζεται η μικρότερη απόσταση και συγχωνεύουμε τις αντίστοιχες παρατηρήσεις για να αποτελέσουν μια ομάδα. Έτσι καταλήγουμε σε μια ομαδοποίηση που αποτελείται από $n - 1$ ομάδες, μια ομάδα σε δυο στοιχεία και $n-2$ ομάδες του ενός στοιχείου. Συνεχίζεται ο αλγόριθμος εως ότου όλα τα άτομα να βρεθούν σε μια και μοναδική ομάδα αποτελούμενη από n στοιχεία. Παρακάτω παρουσιάζεται ο αλγόριθμος σε μορφή βημάτων:

1. Ξεκινώντας θεωρούμε καθεμιά από τις παρατηρήσεις σαν να είναι μια ξεχωριστή ομάδα και υπολογίζουμε τον πίνακα αποστάσεων (εναλλακτικά ομοιότητας) μεταξύ τους (για όλες τις ομάδες).

2. Εντοπίζουμε στον πίνακα τη μικρότερη δυνατή απόσταση (ή εναλλακτικά τη μεγαλύτερη τιμή ενός μέτρου ομοιότητας).

3. Συνενώνουμε τις παρατηρήσεις με τη μικρότερη απόσταση (ή τη μεγαλύτερη τιμή ομοιότητας) μειώνοντας έτσι τον αριθμό των ομάδων κατά 1. Υπολογίζουμε ξανά τον πίνακα με τις ομάδες που έχουν προκύψει και βρίσκουμε ξανά την μικρότερη απόσταση και ενώνουμε τις δυο ομάδες που αντιστοιχούν σε αυτή την απόσταση.

4. Αν δεν έχουν μπει όλες οι παρατηρήσεις σε μια ομάδα επαναλαμβάνουμε τα βήματα 2 και 3 αλλιώς ο αλγόριθμος τερματίζεται.

Με βάση όσα έχουμε αναφέρει μέχρι στιγμής μπορεί να δει κανείς ότι έχουμε ορίσει μέτρο απόστασης μεταξύ δυο παρατηρήσεων όμως δεν έχουμε ορίσει μέτρο απόστασης μεταξύ μιας παρατήρησης και μιας ομάδας. Για αυτό τον λόγο θα δούμε μερικές συσσωρευτικές μεθόδους στις οποίες ορίζεται κάθε φορά ένα διαφορετικό μέτρο απόστασης μεταξύ μιας παρατήρησης και μιας ομάδας.

1. Μέθοδος της απλής συνένωσης (Single Linkage Method):

   Η μέθοδος αυτή γνωστή και ως μέθοδος του πλησιέστερου ή κοντινότερου γείτονα υπολογίζει την απόσταση ανάμεσα σε δυο ομάδες ως τη μικρότερη απόσταση από μια παρατήρηση μέσα στη μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Παρόλο που έχει κάποιες χρήσιμες μαθηματικές ιδιότητες, συνήθως δημιουργεί ομάδες που δεν είναι συμπαγείς, μερικές πολύ μεγάλες ομάδες και κάποιες άλλες πολύ μικρές. Προκειμένου να ενώσουμε δυο ομάδες υπολογίζουμε την ελάχιστη απόσταση $\min d_{ij}$.

2. Μέθοδος της πλήρους συνένωσης (Complete Linkage Method):

   Η μέθοδος αυτή γνωστή και ως μέθοδος του μακρινότερου γείτονα υπολογίζει την απόσταση ανάμεσα σε δυο ομάδες ως τη μεγαλύτερη απόσταση από μια παρατήρηση μέσα σε μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Οι ομάδες που δημιουργούνται με αυτή τη μέθοδο είναι συνήθως συμπαγείς και μεγάλες, όμως η μέθοδος αυτή αρκετά συχνά αποτυγχάνει να ξεχωρήσει κάποιες πολύ συμπαγείς μικρές ομάδες.

3. Average Between Groups:

   Στην περίπτωση αυτή η απόσταση είναι ο μέσος όλων των αποστάσεων που προκύπτουν όταν ενώσουμε τις δυο ομάδες.

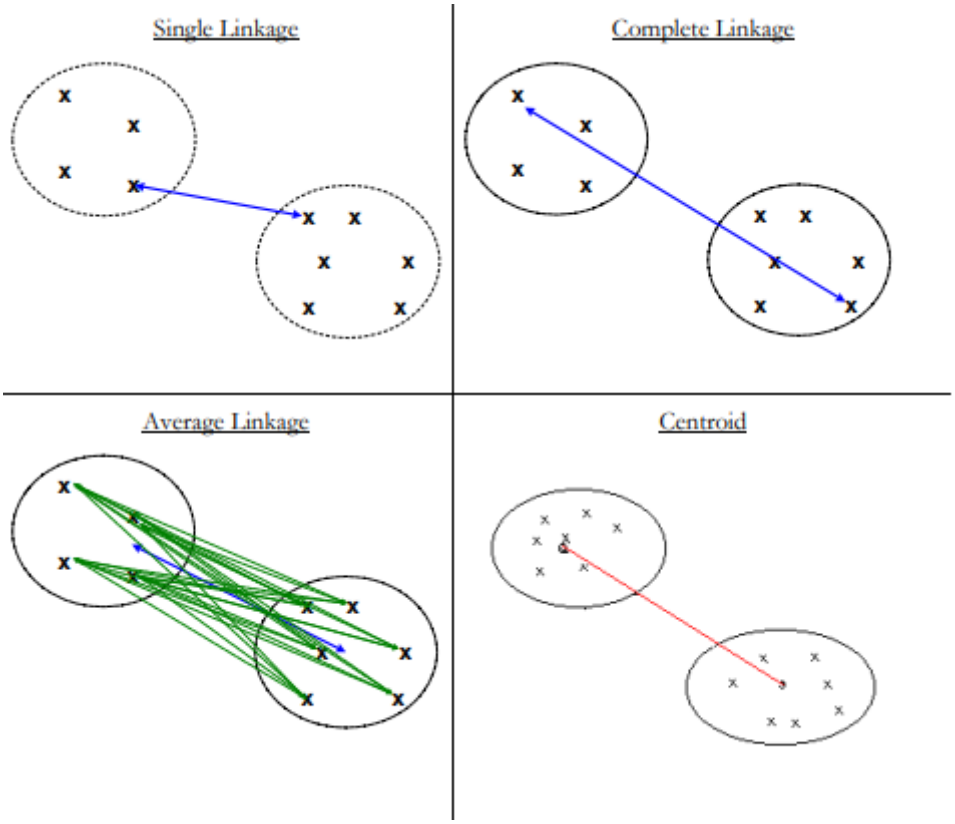
4. Μέθοδος των κέντρων βάρους:

   Σε αυτή τη μέθοδο η απόσταση των δυο ομάδων είναι ίση με την απόσταση του κέντρου βάρους των ομάδων. Οι ομάδες που δημιουργούνται είναι συμπαγείς και ελλειπτικές. Ένα μειονέκτημα της μεθόδου είναι ότι εφαρμόζεται μόνο σε ποσοτικά δεδομένα λόγω της χρήσης της Ευκλείδιας απόστασης.

5. Μέθοδος του Ward:

   Η μέθοδος αυτή διαφέρει από τις υπόλοιπες διότι δεν υπολογίζει την απόσταση ανάμεσα στις ομάδες αλλά είναι σχεδιασμένη με τέτοιο τρόπο ώστε να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Ένα μέτρο ομοιογένειας που χρησιμοποιείται είναι το άθροισμα των τετραγώνων των σφαλμάτων. Αυτή η μέθοδος εφαρμόζεται μόνο σε ποσοτικά δεδομένα και δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων. Επειδή δίνει τα καλύτερα αποτελέσματα χρησιμοποιείται πολύ συχνά στην πράξη.

Στο Σχήμα 2.1 μπορεί κανείς να δει πως λειτουργούν και υπολογίζουν την απόσταση διάφορες μέθοδοι:



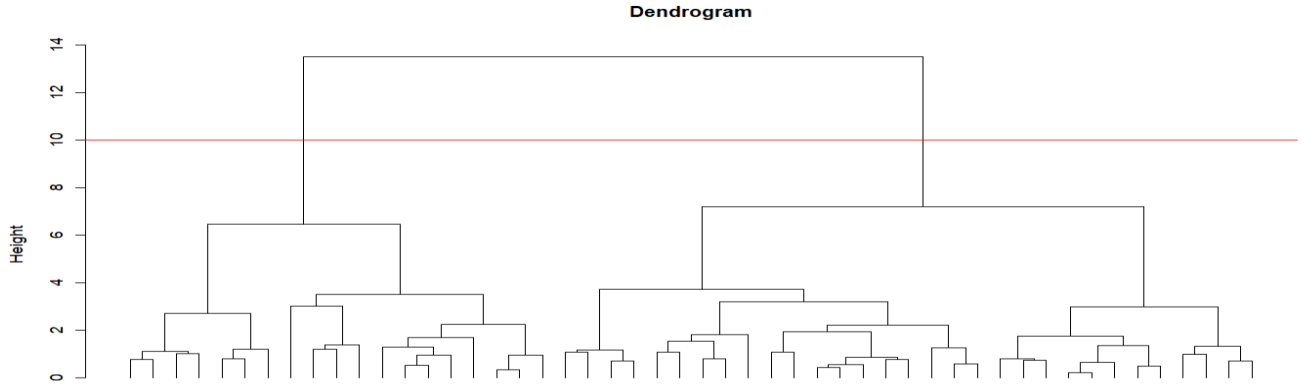
Σχήμα 2.1: Διαγραμματική απεικόνιση μεθόδων υπολογισμού αποστάσεων μεταξύ ομάδων

Αξίζει να αναφέρουμε ότι υπάρχει και μια άλλη κατηγορία ιεραρχικών μεθόδων η οποία εκτελεί ακριβώς την αντίστροφη διαδικασία από τις συσσωρευτικές. Θεωρούμε αρχικά όλες τις παρατηρήσεις σαν μέλη μιας ενιαίας ομάδας. Στη συνέχεια η αρχική αυτή ομάδα διαιρείται σε δυο υποομάδες οι οποίες έχουν τη μεγαλύτερη ανομοιότητα. Η διαδικασία αυτή του διαδοχικού διαχωρισμού των ομάδων επαναλαμβάνεται έως ότου φτάσουμε στο σημείο όλες οι ομάδες να περιέχουν μια μόνο παρατήρηση.

Κλείνοντας αυτή την ενότητα αξίζει να αναφέρουμε το βασικό κριτήριο με το οποίο επιλέγουμε το πλήθος των ομάδων. Τις περισσότερες φορές επιλέγουμε το πλήθος των ομάδων με βάση την εικόνα του δενδρογράμματος διότι είναι ένας πολύ απλός και πρακτικός τρόπος καθορισμού του βέλτιστου πλήθους των ομάδων. Το δενδρόγραμμα αρχίζει με τόσες ομάδες όσες είναι οι παρατηρήσεις και τελειώνει σε μια ομάδα η οποία περιλαμβάνει όλες τις παρατηρήσεις, διότι προκύπτει από μια ιεραρχική συσσωσευρευτική μέθοδο.

Σε αυτό οι οριζόντιες γραμμές δηλώνουν συνδυασμούς ομάδων παρατηρήσεων, ενώ στον κάθετο άξονα καταγράφεται η ποσότητα (απόσταση ή μέτρο ομοιότητας) κατά την οποία οι ομάδες συνδυάζονται. Ουσιαστικά, κάθε επίπεδο ενός δενδρογράμματος ορίζει ένα βήμα του αλγορίθμου. Στο σημείο εκείνο του δενδρογράμματος που παρατηρείται ότι γίνεται συγχώνευση ομάδων που απέχουν αισθητά πάρα πολύ μπορούμε να φέρουμε μια παράλληλη γραμμή ως προς τον οριζόντιο άξονα και να δούμε σε πόσα σημεία τέμνει το δενδρόγραμμα. Εστω ότι το πλήθος αυτών των σημείων είναι k, τότε η επιλογή των k ομάδων αποτελεί μια λογική επιλογή του βέλτιστου πλήθους των ομάδων. Για να κατανοήσουμε καλύτερα το πως γίνεται η επιλογή του βέλτιστου αριθμού ομάδων μπορούμε να δούμε το Σχήμα 2.2. Πιο συγκεκριμένα, στο Σχήμα 2.2 μπορούμε να δούμε ότι στο ύψος ίσο με 10 συγχωνεύονται δυο ομάδες οι οποίες απέχουν αισθητά πολύ μεταξύ τους. Αν τράβήξουμε μια παράλληλη γραμμή ως προς τον οριζόντιο άξονα στο ύψος ίσο με 10 θα δούμε ότι αυτή τέμνει το δενδρόγραμμα σε 2 σημεία. Άρα, το

βέλτιστο πλήθος ομάδων είναι ίσο με δυο. Είναι ευνόητο πως τα αποτελέσματα που προκύπτουν από μια τέτοια διαδικασία υπόκεινται στην κρίση του ερευνητή γιατί η εκτίμηση του αν μια απόσταση είναι πολύ μεγάλη ή όχι είναι κάτι τελείως υποκειμενικό.



Σχήμα 2.2: Παράδειγμα ενός δενδρογράμματος

2.6 Μη Ιεραρχικές Μέθοδοι συσταδοποίησης

Μια άλλη κατηγορία μεθόδων είναι οι λεγόμενες μη ιεραρχικές μέθοδοι. Εδώ θεωρείται ότι ο αριθμός των ομάδων είναι από πριν γνωστός. Ο στόχος αυτών των μεθόδων είναι να ομαδοποιήσουν n παρατηρήσεις που έχουμε σε *k* ομάδες, όπου το *k* είναι καθορισμένο από την αρχή από τον ερευνητή. Χρησιμοποιούμε έναν επαναληπτικό αλγόριθμο για να τοποθετήσουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην εκάστοτε παρατήρηση.

Ουσιαστικά ο τρόπος λειτουργίας ενός τέτοιου αλγορίθμου είναι είτε να θεωρεί *k* συγκεκριμένες παρατηρήσεις (μητρικά σημεία) και γύρω από αυτά να ταξινομούνται οι υπόλοιπες παρατηρήσεις έως ότου διαμορφωθούν οι επιθυμητές ομάδες είτε να ξεκινούν με έναν αρχικό διαμερισμό των παρατηρήσεων σε *k* ομάδες και στη συνέχεια να μετακινούνται οι παρατηρήσεις μεταξύ των ομάδων μέχρι να επιτευχθεί ο καλύτερος διαμερισμός. Αυτές οι μέθοδοι, ενώ δουλεύουν ικανοποιητικά με μεγάλα δείγματα, επηρεάζονται αρκετά από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

Για τον τρόπο δημιουργίας των μητρικών σημείων υπάρχουν διάφορες μέθοδοι. Ενδεικτικά αναφέρουμε τρεις:

1. Αριθμούμε τις παρατηρήσεις από το 1 έως το *n* και δημιουργούμε *k* διαφορετικούς τυχαίους αριθμούς από το 1 έως το *n* και επιλέγουμε τις παρατηρήσεις που αντιστοιχούν σε αυτούς τους αριθμούς.

2. Επιλέγουμε τις πρώτες *k* στη σειρά παρατηρήσεις.

3. Διαχωρίζουμε με κάποιο υποκειμενικό τρόπο τις παρατηρήσεις σε *k* ομάδες και θεωρούμε τα κέντρα βάρους των ομάδων μητρικά σημεία.

Από τις πιο γνωστές μη ιεραρχικές μεθόδους είναι αυτή που προτάθηκε από τον Forgy και ο αλγόριθμος K-means (μέθοδος Mac Queen).

Παρακάτω παρουσιάζονται τα βήματα του αλγορίθμου K-means:

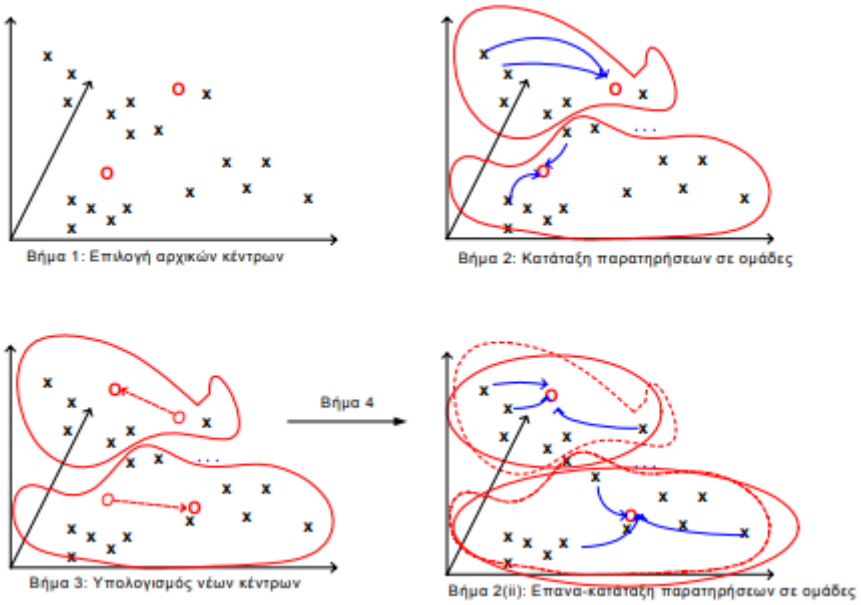
1. Καθόρισε ένα αρχικό συνόλο από *k* μητρικά σημεία χρησιμοποιώντας *k* από τις *n* παρατηρήσεις που είναι διαθέσιμες.

2. Κατάταξε την καθεμιά από τις εναπομείνουσες $n - k$ παρατηρήσεις στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.

3. Μετά από κάθε τοποθέτηση παρατήρησης, υπολόγισε ξανά τα μητρικά σημεία (κέντρα) της καινούργιας πλεον ομάδας.

4. Αν τα μητρικά σημεία δεν είναι διαφορετικά από τα παλιά σταμάτα αλλιώς πήγαινε στο βήμα 2.

Στο Σχήμα 2.3 φαίνεται πως δουλεύει ο αλγόριθμος (βλ. [2],[3],[5],[4]).



Σχήμα 2.3: Περιγραφή της μεθόδου K-means

Κεφάλαιο 3

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΠΟ ΤΗΝ ΕΦΑΡΜΟΓΗ ΤΩΝ ΜΕΘΟΔΩΝ

3.1 Πληροφορίες για το σύνολο δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την εφαρμογή των μεθόδων πολυμεταβλητής ανάλυσης δεδομένων αντλήθηκε από το site Our World in Data [15] και περιέχει δεδομένα που αφορούν 129 χώρες. Πιο συγκεκριμένα, περιέχει τις ημερήσιες τιμές των παρακάτω μεταβλητών από 13-5-20 έως και 18-6-22 για κάθε μια από τις 129 χώρες.

1. iso_code: Διεθνής αναγνωρισμένος κωδικός για κάθε χώρα ο οποίος αποτελείται από 2 ή 3 γράμματα. Παραδείγματος χάρη, για το Αφγανιστάν αυτός ο κωδικός είναι AFG.
2. continent: Η ήπειρος στην οποία ανήκει η κάθε χώρα.
3. location: Η χώρα
4. date: Η ημερομηνία
5. population: Ο πληθυσμός της χώρας (παραμένει σταθερός για κάθε χώρα).
6. life_expectancy: Το προσδόκιμο ζωής (παραμένει σταθερό για κάθε χώρα).
7. total_cases_per_million: Ο αριθμός των συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού την αντίστοιχη ημέρα.
8. new_cases_per_million: Ο ημερήσιος αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού την αντίστοιχη μέρα.
9. population_density: Η πληθυσμιακή πυκνότητα. Η πληθυσμιακή πυκνότητα είναι ο αριθμός των κατοίκων διαιρεμένος με την επιφάνεια της κάθε χώρας (παραμένει σταθερή για κάθε χώρα).
10. diabetes_prevalence: Το ποσοστό των ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017 (παραμένει σταθερό για κάθε χώρα).
11. total_deaths_per_million: Ο αριθμός των συνολικών θανάτων από Covid19 ανά εκατομμύριο πληθυσμού την αντίστοιχη μέρα.
12. new_deaths_per_million: Ο ημερήσιος αριθμός θανάτων από Covid19 ανά εκατομμύριο πληθυσμού την αντίστοιχη μέρα.
13. median_age: Η διάμεση ηλικία του πληθυσμού (παραμένει σταθερή για κάθε χώρα).
14. aged_70_older: Το ποσοστό του πληθυσμού που είναι πάνω από 70 ετών (παραμένει σταθερό για κάθε χώρα).
15. cardiovasc_death_rate: Ο δείκτης θνησιμότητας από καρδιαγγειακές ασθένειες το 2017 (παραμένει σταθερός για κάθε χώρα).
16. gdp_per_capita: Ακαθάριστο Εγχώριο Προϊόν (παραμένει σταθερό για κάθε χώρα).

17. human_development_index: Σύνθετος δείκτης ο οποίος ποσοτικοποιεί την μέση πρόοδο του ανθρώπου στους τομείς: μακροζωία και υγεία, μόρφωση, καλό επίπεδο ζωής (παραμένει σταθερός για κάθε χώρα).

18. reproduction_rate: Ο δείκτης R (Είναι μια ποσότητα η οποία μας δείχνει την δυνατότητα του ιού να μεταδοθεί. Πιο συγκεκριμένα, το R δηλώνει τον κατά μέσο όρο αριθμό των ανθρώπων που μπορεί να μολύνει ένα άτομο που φέρει τον ιο. Παραδείγματος χάρη, αν το R = 0.7 αυτό σημαίνει πως κατα μέσο όρο κάθε 10 άτομα που πάσχουν από Covid19 θα μολύνουν 7 άλλα άτομα.).

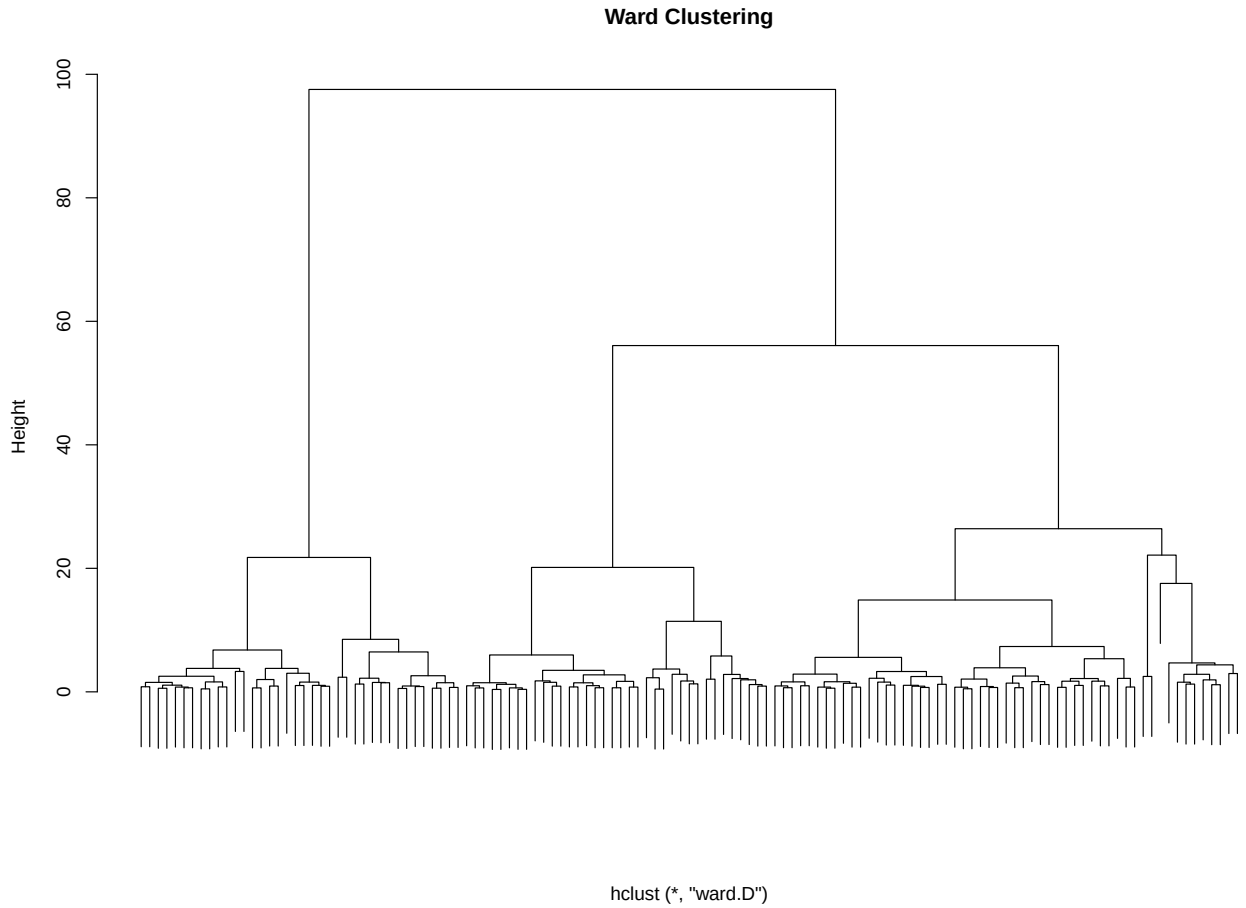
19. stringency_index: Δείκτης με τον οποίο ποσοτικοποιείται η σφοδρότητα των κυβερνητικών μέτρων που έχουν στόχο την καταπολέμηση της πανδημίας Covid19.

20. hospital_beds_per_thousand: Ο αριθμός των κρεβατιών στα νοσοκομεία ανά 1000 άτομα (παραμένει σταθερός για κάθε χώρα).

3.2 Ανάλυση κατά Συστάδες

Για την μελέτη της εξέλιξης της πανδημίας αν πάμε τυφλά και μελετήσουμε απλά τις μεταβλητές που σχετίζονται απευθείας με την εξέλιξη της πανδημίας (total_cases_per_ million, new_cases_per_million, total_deaths_per_million, new_deaths_per_million, reproduction_rate, stringency_index) τότε η εικόνα που θα μας δώσουν τα αποτελέσματα δεν θα είναι η πραγματική, γιατί οι χώρες που υπάρχουν στην ανάλυση μπορεί να έχουν άλλα χαρακτηριστικά που τις διαφοροποιούν με αποτέλεσμα να παρατηρούμε αυτή την εξέλιξη των μεταβλητών λόγω άλλων χαρακτηριστικών. Επομένως, για να αναλύσουμε τα δεδομένα που έχουμε στην διαθεσή μας θα ήταν πρώτα λογικό να βρούμε, με βάση τις υπόλοιπες μεταβλητές δηλαδή τις μεταβλητές population, life_expectancy, population_density, diabetes_prevalence, median_age, aged_70_older, cardiovasc_death_rate, gdp_per_capita, human_development_index, hospital _beds_per_thousand, τις ομαδοποιήσεις που υπάρχουν μεταξύ των χωρών και έπειτα να μελετήσουμε σε κάθε ομάδα που είναι ομοιογενής ως προς τις προαναφερθείσες μεταβλητές, πως εξελίχθηκε η πανδημία Covid19 και να δούμε αν υπάρχουν διαφοροποιήσεις.

Εφαρμόζοντας την μέθοδο Ward στις τυποποιημένες τιμές των μεταβλητών population, life_ expectancy, population_density, diabetes_prevalence, median_age, aged_70_older, cardiovasc_ death_rate, gdp_per_capita, human_development_index, hospital _beds_per_thousand μπορούμε να δούμε πόσες ομάδες χωρών υπάρχουν.



Σχήμα 3.1: Διάγραμμα Ward για την εύρεση του πλήθους των ομαδοποιήσεων των χωρών

Από το Σχήμα 3.1 μπορούμε να δούμε ότι υπάρχουν 3 ομάδες χωρών. Έπειτα θα εφαρμόσουμε την μέθοδο k-means για να δούμε σε ποια ομάδα ανήκει η κάθε χώρα. Τα αποτελέσματα συνοψίζονται στους πίνακες 3.1 - 3.5.

Αργεντινή	Αυστραλία	Αυστρία	Λευκορωσία	Βέλγιο
Βοσνία Ερζεγοβίνη	Βουλγαρία	Καναδάς	Κροατία	Κύπρος
Τσεχία	Δανία	Εστονία	Φιλανδία	Γαλλία
Γερμανία	Ελλάδα	Ουγγαρία	Ισλανδία	Ιρλανδία
Ισραήλ	Ιταλία	Ιαπωνία	Λετονία	Λιθουανία
Λουξεμβούργο	Μάλτα	Ολλανδία	Νέα Ζηλανδία	Νορβηγία
Πολωνία	Πορτογαλία	Ρουμανία	Ρωσία	Σιγκαπούρη
Σλοβακία	Σλοβενία	Νότια Κορέα	Ισπανία	Σουηδία
Ελβετία	Ουκρανία	Ηνωμένο Βασίλειο	Η.Π.Α	Ουρουγουάη

Πίνακας 3.1: Χώρες που ανήκουν στην πρώτη ομάδα

Αλβανία	Αλγερία	Αζερμπαϊτζάν	Μπαχάμες	Μπαχρέιν
Μπανγκλαντές	Βολιβία	Βραζιλία	Μπρουνέι	Χιλή
Κίνα	Κολομβία	Κόστα Ρίκα	Εκουαδόρ	Αίγυπτος
Ελ Σαλβαδόρ	Γκαμπόν	Γεωργία	Γουατεμάλα	Γουιάνα
Ονδούρα	Ινδία	Ινδονησία	Ιράν	Ιράκ
Τζαμάικα	Ιορδανία	Καζακστάν	Κουβέιτ	Κιργιζία
Λίβανο	Λιβύη	Μαλαισία	Μαυρίκιος	Μεξικό
Μολδαβία	Μαρόκο	Νικαραγουά	Ομάν	Παναμάς
Παραγουάη	Περού	Φιλιππίνες	Κατάρ	Σαουδική Αραβία
Σρι Λάνκα	Ταϊλάνδη	Τυνησία	Τουρκία	Βενεζουέλα

Πίνακας 3.2: Χώρες που ανήκουν στην δεύτερη ομάδα (Μέρος 1)

Ενωμένα Αραβικά Εμιράτα	Δομινικανή Δημοκρατία	Τρινιναντ και Τομπάγκο

Πίνακας 3.3: Χώρες που ανήκουν στην δεύτερη ομάδα (Μέρος 2)

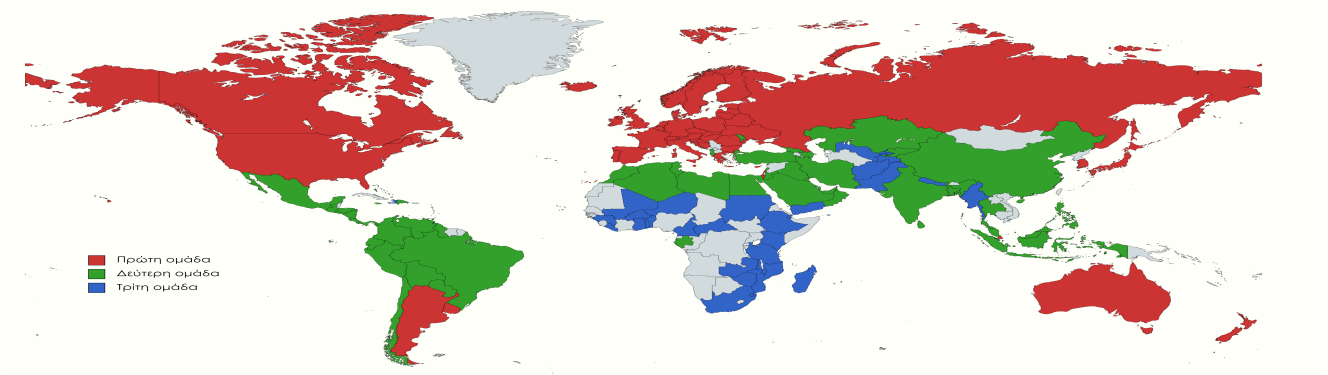
Αφγανιστάν	Μπενίν	Μπουρκίνα Φάσο	Καμερούν
Υεμένη	Ζάμπια	Τζιμπουτί	Σουαζιλάνδη
Αιθιοπία	Γκάνα	Γουινέα	Αϊτή
Κένυα	Λιβερία	Μαγαδασκάρη	Μαλάουι
Μάλι	Μοζαμβίκη	Μιανμάρ (Βιρμανία)	Νεπάλ
Νιγηρία	Πακιστάν	Νότια Αφρική	Σουδάν
Τατζικιστάν	Τανζανία	Τογκό	Ουζμπεκιστάν

Πίνακας 3.4: Χώρες που ανήκουν στην τρίτη ομάδα (Μέρος 1)

Ζιμπάμπουε	Cape Verde (Πράσινο Ακροτήρι)	Κεντροαφρικανική Δημοκρατία

Πίνακας 3.5: Χώρες που ανήκουν στην τρίτη ομάδα (Μέρος 2)

Η πληροφορία που δίνεται από τους πίνακες 3.1 - 3.5 μπορεί να φανεί στο Σχήμα 3.2.



Σχήμα 3.2: Απεικόνιση των τριών ομάδων στον παγκόσμιο χάρτη

Από το Σχήμα 3.2 βλέπουμε ότι η πρώτη ομάδα περιέχει όλες τις Ευρωπαϊκές χώρες, τον Καναδά, τις Η.Π.Α, την Αργεντινή και την Αυστραλία. Αυτό σημαίνει ότι η πρώτη ομάδα περιέχει κατά κύριο λόγο χώρες οι οποίες είναι πολύ ανεπτυγμένες στον τομέα της υγείας, της οικονομίας, της εκπαίδευσης. Επιπλέον, από το Σχήμα 3.2 βλέπουμε ότι η δεύτερη ομάδα περιέχει χώρες της Ανατολής, χώρες που ανήκουν στην Νότια Αμερική, χώρες που ανήκουν στην Βόρεια Αφρική. Αυτό σημαίνει ότι η δεύτερη ομάδα περιέχει κατά κύριο λόγο χώρες οι οποίες είναι μέτρια ανεπτυγμένες στον τομέα της υγείας, της οικονομίας, της εκπαίδευσης. Τέλος, από το Σχήμα 3.2 βλέπουμε ότι η τρίτη ομάδα κατά κύριο λόγο περιέχει χώρες της Αφρικής οι οποίες δεν είναι καθόλου ανεπτυγμένες στον τομέα της υγείας, της οικονομίας, της εκπαίδευσης.

Αξίζει να επισημάνουμε ότι δοκιμάστηκε να εφαρμοστεί η Παραγοντική Ανάλυση αλλά δεν ήταν ξεκάθαρα τα αποτελέσματα και δεν εντοπίστηκαν κάποιοι παράγοντες που να μπορούσαν να ερμηνευτούν.

3.3 Περιγραφική Στατιστική

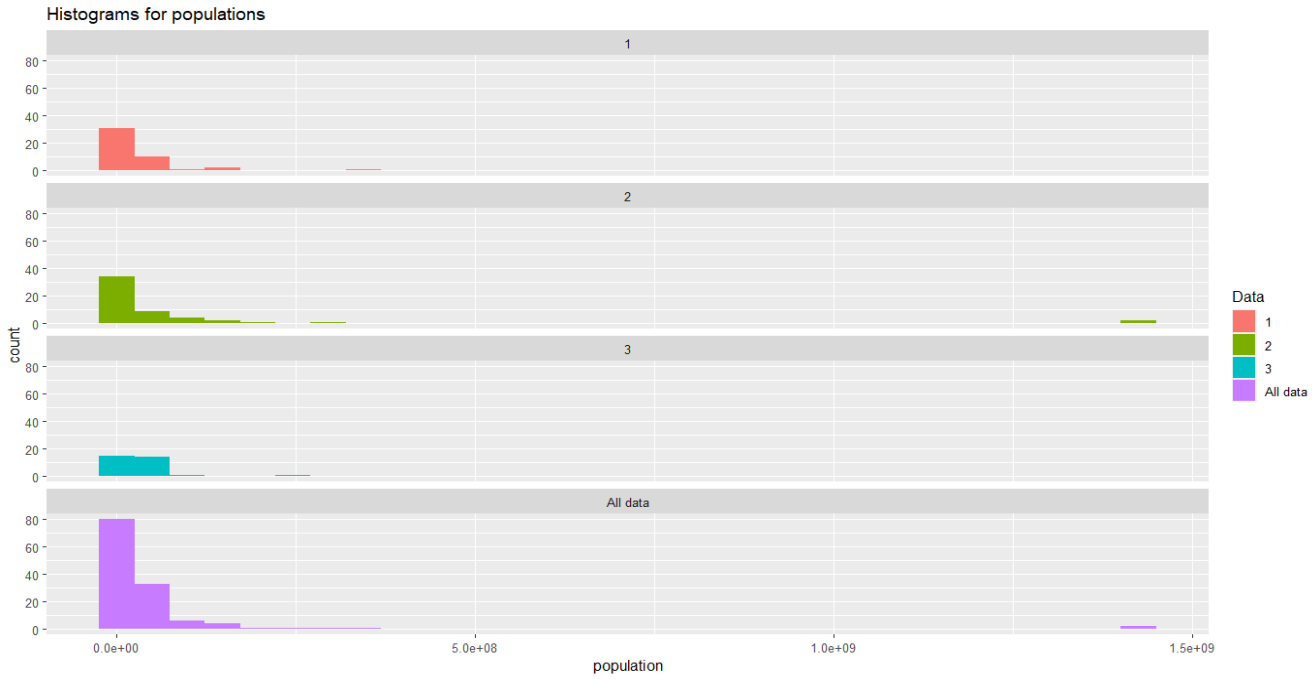
Στους πίνακες 3.6 - 3.21 δίνονται μερικά περιγραφικά μέτρα για τις 3 ομάδες καθώς και το σύνολο των δεδομένων.

- Μεταβλητή population

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	30598142	87654645	34834546	55058012
Τυπική απόκλιση	56814543	271501151	43798352	179595069
Ελάχιστη τιμή	370335	407906	587925	370335
25% ποσοστημόριο	4986526	4351267	12222232	5193416
Διάμεσος	9578168	11148278	25252722	12079472
75% ποσοστημόριο	38307726	43533592	37090456	40099462
Μέγιστη τιμή	336997624	1425893464	231402116	1425893464

Πίνακας 3.6: Περιγραφικά μέτρα για την μεταβλητή population

Από τον πίνακα 3.6 βλέπουμε ότι ο μέσος πληθυσμός της δεύτερης ομάδας είναι πολύ μεγαλύτερος από τον μέσο πληθυσμό των ομάδων 1 και 3. Επιπλέον, παρατηρούμε ότι η διάμεσος του πληθυσμού της ομάδας 3 είναι μεγαλύτερη από αυτή των ομάδων 1 και 2.



Σχήμα 3.3: Ιστογράμματα για την μεταβλητή population

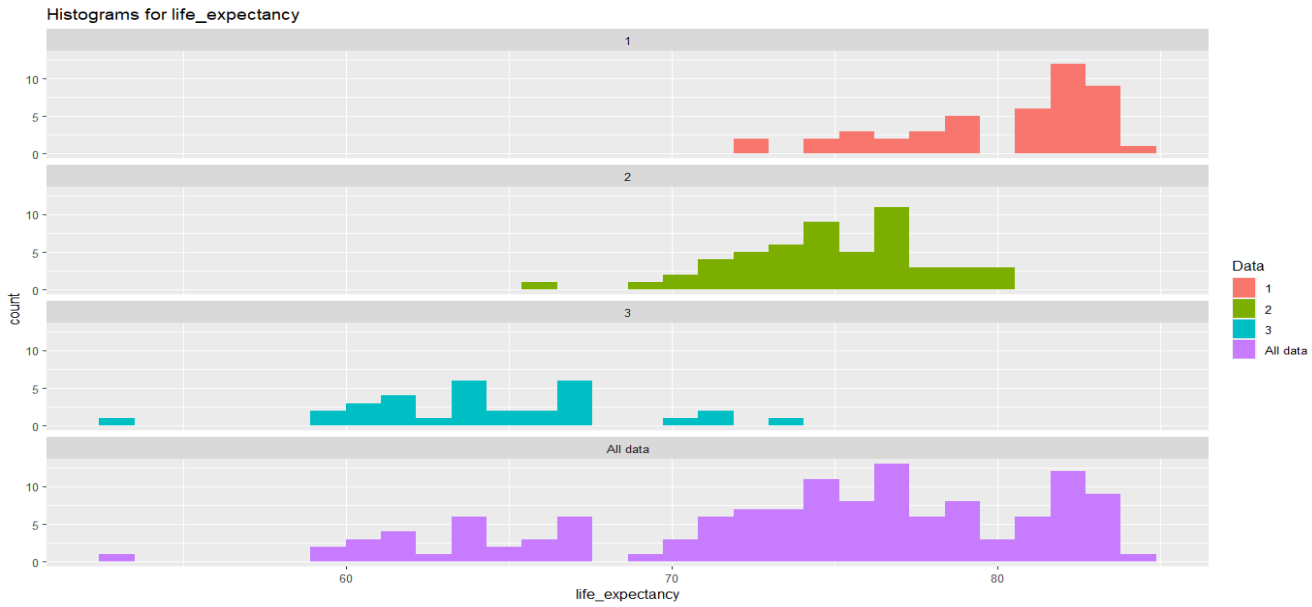
Από το Σχήμα 3.3 βλέπουμε ότι στο ιστόγραμμα της δεύτερης ομάδας υπάρχουν χώρες που έχουν πάρα πολύ υψηλό πληθυσμό. Αυτές οι χώρες είναι η Κίνα και η Ινδία. Άρα, ο πληθυσμός της Κίνας και της Ινδίας είναι πάρα πολύ υψηλός. Επειδή η Κίνα και η Ινδία έχουν πάρα πολύ υψηλό πληθυσμό, η διάμεσος και τα ποσοστημόρια είναι πιο αξιόπιστα περιγραφικά μέτρα από ότι η μέση τιμή για την μεταβλητή population.

- Μεταβλητή life_expectancy

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	80.24533	74.93566	64.43290	74.26395
Τυπική απόκλιση	3.225796	2.914619	4.127502	6.869709
Ελάχιστη τιμή	72.06	66.47	53.28	53.28
25% ποσοστημόριο	77.91	73.00	61.59	71.10
Διάμεσος	81.54	75.05	64.13	75.29
75% ποσοστημόριο	82.53	76.98	66.87	78.93
Μέγιστη τιμή	84.63	80.28	72.98	84.63

Πίνακας 3.7: Περιγραφικά μέτρα για την μεταβλητή life_expectancy

Από τον πίνακα 3.7 βλέπουμε ότι η πρώτη ομάδα έχει το υψηλότερο μέσο προσδόκιμο ζωής και ότι η δεύτερη ομάδα έχει το αμέσως μεγαλύτερο μέσο προσδόκιμο ζωής. Η αλήθεια είναι ότι κάτι τέτοιο το αναμέναμε καθώς όπως αναφέραμε και πιο πριν η πρώτη ομάδα περιλαμβάνει χώρες που είναι πολύ ανεπτυγμένες τους τομείς της υγείας, της οικονομίας, της εκπαίδευσης, η δεύτερη ομάδα περιλαμβάνει χώρες που είναι μέτρια ανεπτυγμένες στους προαναφερθέντες τομείς και η τρίτη ομάδα περιλαμβάνει χώρες που δεν είναι ανεπτυγμένες στους προαναφερθέντες τομείς. Επιπλέον, βλέπουμε ότι η διάμεσος του προσδόκιμου ζωής στην πρώτη ομάδα είναι η υψηλότερη και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή προσδόκιμο ζωής.



Σχήμα 3.4: Ιστογράμματα για την μεταβλητή life_expectancy

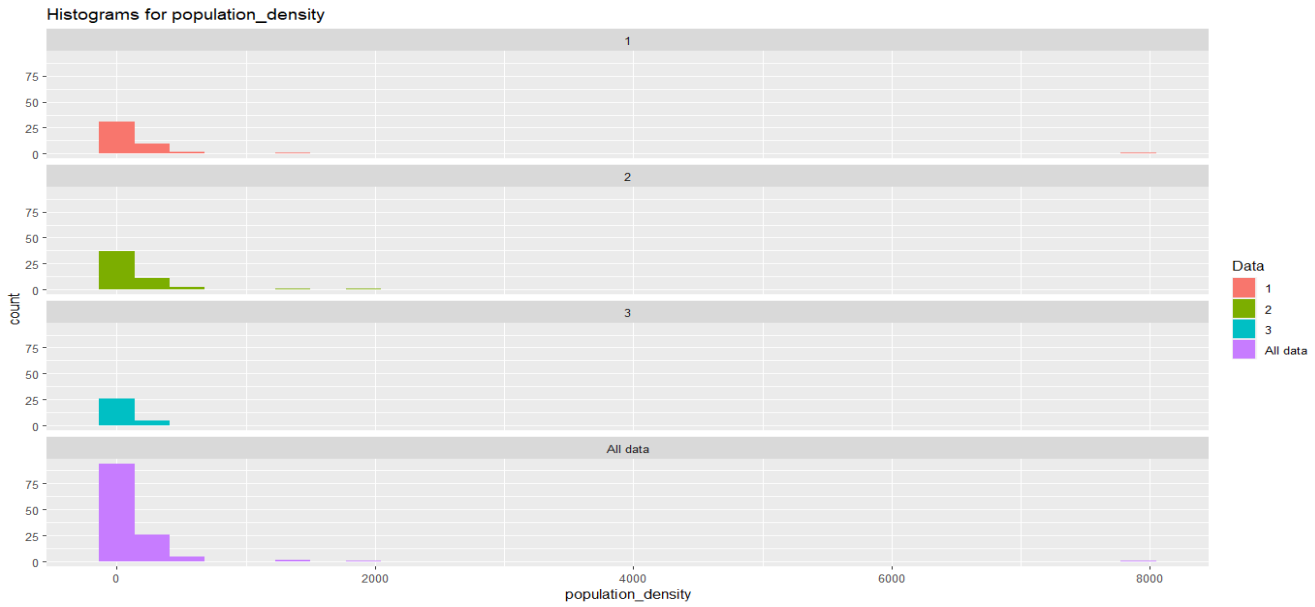
Από το Σχήμα 3.4 βλέπουμε ότι οι χώρες της πρώτης ομάδας λαμβάνουν πολλές τιμές που είναι υψηλές στην μεταβλητή life_expectancy. Επιπλέον, βλέπουμε ότι οι χώρες της δεύτερης ομάδας λαμβάνουν πολλές τιμές που είναι σχετικά υψηλές στην μεταβλητή life_expectancy και ότι οι χώρες της τρίτης ομάδας λαμβάνουν πολλές τιμές που είναι αρκετά χαμηλές στην μεταβλητή life_expectancy. Επίσης, βλέπουμε ότι στο ιστόγραμμα της τρίτης ομάδας υπάρχει μια ακραία παρατήρηση στην αριστερή "ουρά". Αυτή η ακραία παρατήρηση αντιστοιχεί στην Κεντροαφρικανική Δημοκρατία. Άρα, το προσδόκιμο ζωής στην Κεντροαφρικανική Δημοκρατία είναι πάρα πολύ χαμηλό. Λόγω αυτής της ακραίας παρατήρησης, η διάμεσος και τα ποσοστημόρια είναι πιο αξιόπιστα περιγραφικά μέτρα από ότι η μέση τιμή για την μεταβλητή life_expectancy.

- Μεταβλητή population_density

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	329.37109	183.45819	88.63003	211.5699
Τυπική απόκλιση	1180.58282	322.12650	81.87022	729.1793
Ελάχιστη τιμή	3.202	3.623	7.479	3.202
25% ποσοστημόριο	31.212	36.253	43.340	37.728
Διάμεσος	93.105	88.125	64.281	81.347
75% ποσοστημόριο	205.8590	157.8340	102.0335	143.366
Μέγιστη τιμή	7915.731	1935.907	398.448	7915.731

Πίνακας 3.8: Περιγραφικά μέτρα για την μεταβλητή population_density

Από τον πίνακα 3.8 βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη μέση πληθυσμιακή πυκνότητα και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη μέση πληθυσμιακή πυκνότητα. Επιπλέον, βλέπουμε ότι η διάμεσος της πληθυσμιακής πυκνότητας στην πρώτη ομάδα είναι η υψηλότερη και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή πληθυσμιακή πυκνότητα.



Σχήμα 3.5: Ιστογράμματα για την μεταβλητή population_density

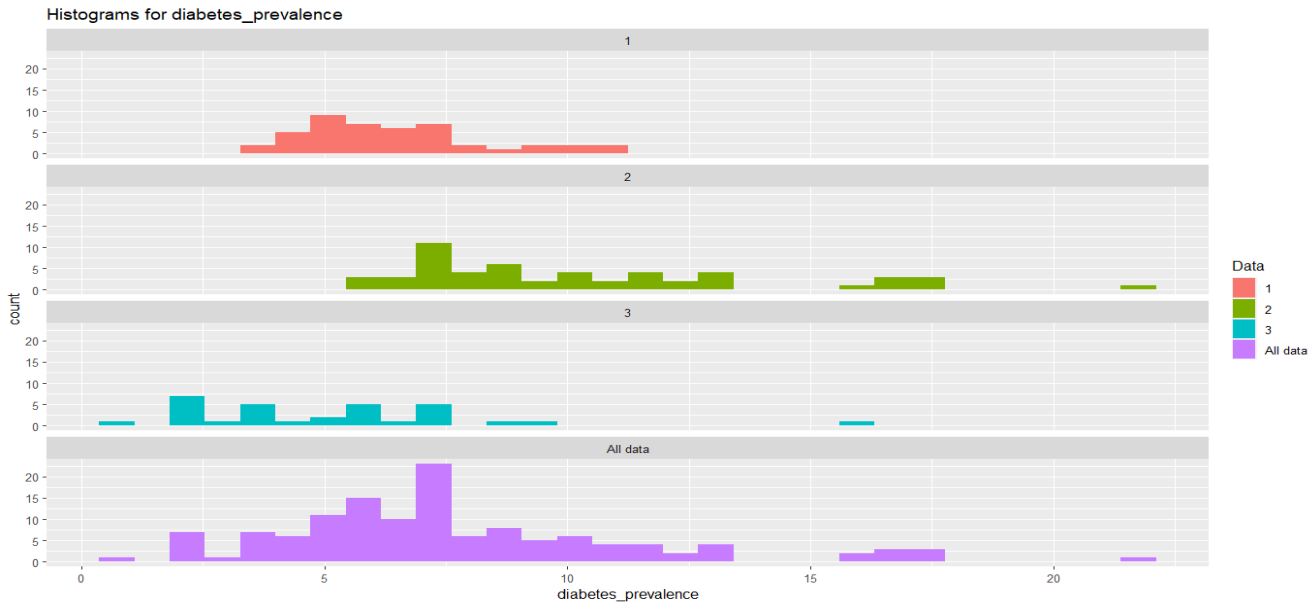
Από το Σχήμα 3.5 βλέπουμε ότι στο ιστόγραμμα της πρώτης ομάδας υπάρχει μια ακραία παρατήρηση στην δεξιά "ουρά". Αυτή η ακραία παρατήρηση αντιστοιχεί στην Σινγκαπούρη. Άρα, η Σινγκαπούρη έχει πάρα πολύ υψηλή πληθυσμιακή πυκνότητα. Λόγω αυτής της ακραίας παρατήρησης, η διάμεσος και τα ποσοστημόρια είναι πιο αξιόπιστα περιγραφικά μέτρα από ότι η μέση τιμή για την μεταβλητή population_density.

- Μεταβλητή diabetes_prevalence

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	6.437333	10.273585	5.183226	7.712093
Τυπική απόκλιση	1.922714	3.757348	2.929800	3.722202
Ελάχιστη τιμή	3.28	5.55	0.99	0.990
25% ποσοστημόριο	5.07	7.14	2.67	5.350
Διάμεσος	5.91	8.87	4.97	7.11
75% ποσοστημόριο	7.29	12.13	6.88	9.590
Μέγιστη τιμή	10.99	22.02	15.67	22.020

Πίνακας 3.9: Περιγραφικά μέτρα για την μεταβλητή diabetes_prevalence

Από τον πίνακα 3.9 βλέπουμε ότι η δεύτερη ομάδα έχει το υψηλότερο μέσο ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017 και ότι η πρώτη ομάδα έχει το αμέσως μεγαλύτερο μέσο ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017. Επιπροσθέτως, βλέπουμε ότι η δεύτερη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει το ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017 και ότι η πρώτη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει το ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017.



Σχήμα 3.6: Ιστογράμματα για την μεταβλητή diabetes_prevalence

Από το Σχήμα 3.6 βλέπουμε ότι υπάρχουν πολλές χώρες στην πρώτη ομάδα που έχουν αρκετά χαμηλό ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017. Επιπλέον, βλέπουμε ότι υπάρχουν πολλές χώρες στην δεύτερη ομάδα που έχουν μέτριο ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017. Ακόμη, βλέπουμε ότι το ιστόγραμμα της δεύτερης ομάδας περιλαμβάνει και κάποιες ακραίες παρατηρήσεις. Αυτό σημαίνει πως στην δεύτερη ομάδα υπάρχουν κάποιες χώρες με πολύ υψηλό ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017. Από την άλλη μεριά, βλέπουμε ότι υπάρχουν πολλές χώρες στην τρίτη ομάδα που έχουν πολύ χαμηλό ποσοστό ατόμων ηλικίας 20 - 79 ετών που έπασχαν από διαβήτη το 2017.

- Μεταβλητή median_age

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	41.67556	30.53585	20.83226	32.08992
Τυπική απόκλιση	3.596346	4.460127	3.587421	8.918545
Ελάχιστη τιμή	30.6	20.0	15.1	15.10
25% ποσοστημόριο	39.70	27.60	18.45	25.30
Διάμεσος	42.4	30.6	19.6	31.9
75% ποσοστημόριο	43.5	33.5	23.4	40.30
Μέγιστη τιμή	48.2	40.1	29.1	48.20

Πίνακας 3.10: Περιγραφικά μέτρα για την μεταβλητή median_age

Από τον πίνακα 3.10 βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη μέση τιμή στην μεταβλητή διάμεση ηλικία και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη μέση τιμή στην μεταβλητή διάμεση ηλικία.

Σχήμα 3.7: Ιστογράμματα για την μεταβλητή median_age

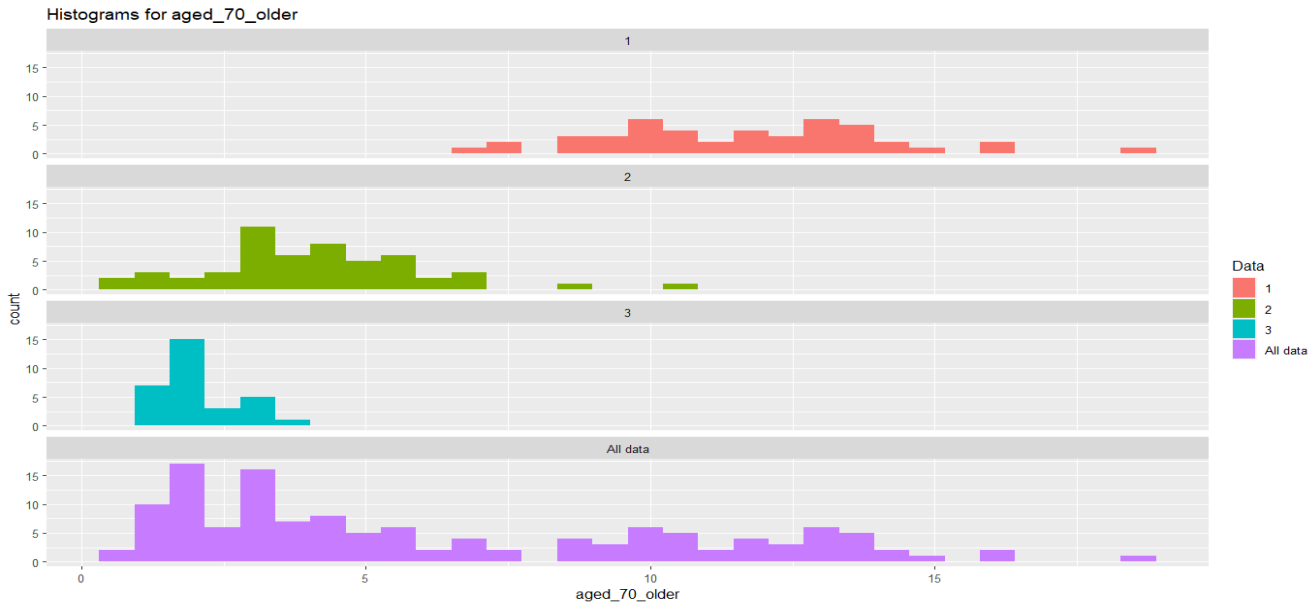
Από το Σχήμα 3.7 βλέπουμε ότι η πρώτη ομάδα περιέχει πολλές χώρες στις οποίες η διάμεση ηλικία είναι πολύ υψηλή. Επιπλέον, από το Σχήμα 3.7 βλέπουμε ότι η δεύτερη ομάδα περιέχει πολλές χώρες στις οποίες η διάμεση ηλικία είναι μέτρια. Ακόμη, από το Σχήμα 3.7 βλέπουμε ότι η τρίτη ομάδα περιέχει πολλές χώρες στις οποίες η διάμεση ηλικία είναι πολύ μικρή.

- Μεταβλητή aged_70_older

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	11.680533	4.116528	2.073774	6.26424
Τυπική απόκλιση	2.4832665	1.9120466	0.6038856	4.490738
Ελάχιστη τιμή	7.049	0.526	1.337	0.526
25% ποσοστημόριο	9.7880	2.8830	1.6345	2.382
Διάμεσος	11.690	3.915	1.882	4.458
75% ποσοστημόριο	13.2720	5.2000	2.3155	10.129
Μέγιστη τιμή	18.493	10.244	3.437	18.493

Πίνακας 3.11: Περιγραφικά μέτρα για την μεταβλητή aged_70_older

Από τον πίνακα 3.11 βλέπουμε ότι οι χώρες της πρώτης ομάδας έχουν το υψηλότερο μέσο ποσοστό πληθυσμού που είναι πάνω από 70 ετών και ότι οι χώρες της δεύτερης ομάδας έχουν το αμέσως μεγαλύτερο μέσο ποσοστό πληθυσμού που είναι πάνω από 70 ετών. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει το ποσοστό πληθυσμού που είναι πάνω από 70 ετών και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει το ποσοστό πληθυσμού που είναι πάνω από 70 ετών.



Σχήμα 3.8: Ιστογράμματα για την μεταβλητή aged_70_older

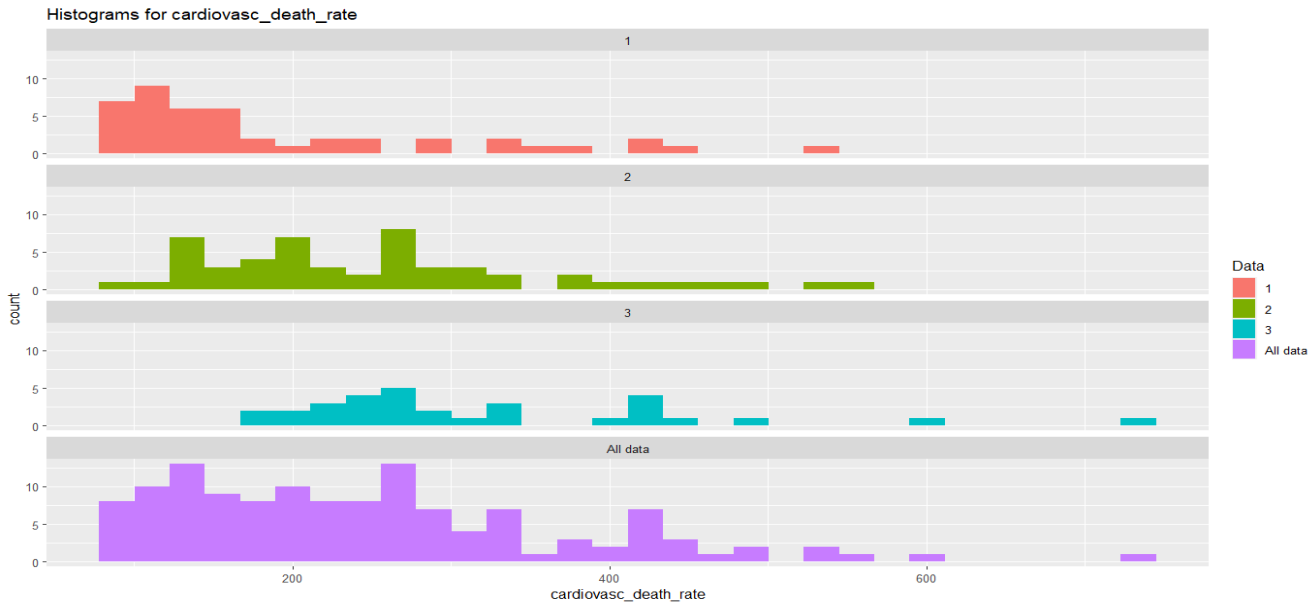
Από το Σχήμα 3.8 βλέπουμε ότι η πρώτη ομάδα περιέχει πολλές χώρες στις οποίες το ποσοστό του πληθυσμού που είναι πάνω από 70 ετών είναι μέτριο. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα περιέχει αρκετές χώρες στις οποίες το ποσοστό του πληθυσμού που είναι πάνω από 70 ετών είναι αρκετά υψηλό. Από την άλλη μεριά, βλέπουμε ότι η δεύτερη ομάδα περιέχει πολλές χώρες στις οποίες το ποσοστό του πληθυσμού που είναι πάνω από 70 ετών είναι αρκετά χαμηλό. Επιπροσθέτως, βλέπουμε ότι η τρίτη ομάδα περιέχει πολλές χώρες στις οποίες το ποσοστό του πληθυσμού που είναι πάνω από 70 ετών είναι πολύ χαμηλό.

- Μεταβλητή cardiovasc_death_rate

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	192.4721	254.7038	321.5944	249.0695
Τυπική απόκλιση	114.9481	111.3213	126.2744	125.4063
Ελάχιστη τιμή	79.370	85.755	182.219	79.37
25% ποσοστημόριο	114.3160	171.2850	235.1735	151.09
Διάμεσος	145.183	235.954	272.509	227.485
75% ποσοστημόριο	253.7820	304.1950	414.5125	317.84
Μέγιστη τιμή	539.849	559.812	724.417	724.42

Πίνακας 3.12: Περιγραφικά μέτρα για την μεταβλητή cardiovasc_death_rate

Από τον πίνακα 3.12 βλέπουμε ότι η τρίτη ομάδα έχει τον υψηλότερο μέσο δείκτη θνησιμότητας από καρδιαγγειακές ασθένειες το 2017 και ότι η δεύτερη ομάδα έχει τον αμέσως μεγαλύτερο μέσο δείκτη θνησιμότητας από καρδιαγγειακές ασθένειες το 2017. Επιπλέον, βλέπουμε ότι η τρίτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει τον δείκτη θνησιμότητας από καρδιαγγειακές ασθένειες το 2017 και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει τον δείκτη θνησιμότητας από καρδιαγγειακές ασθένειες το 2017.



Σχήμα 3.9: Ιστογράμματα για την μεταβλητή cardiovasc_death_rate

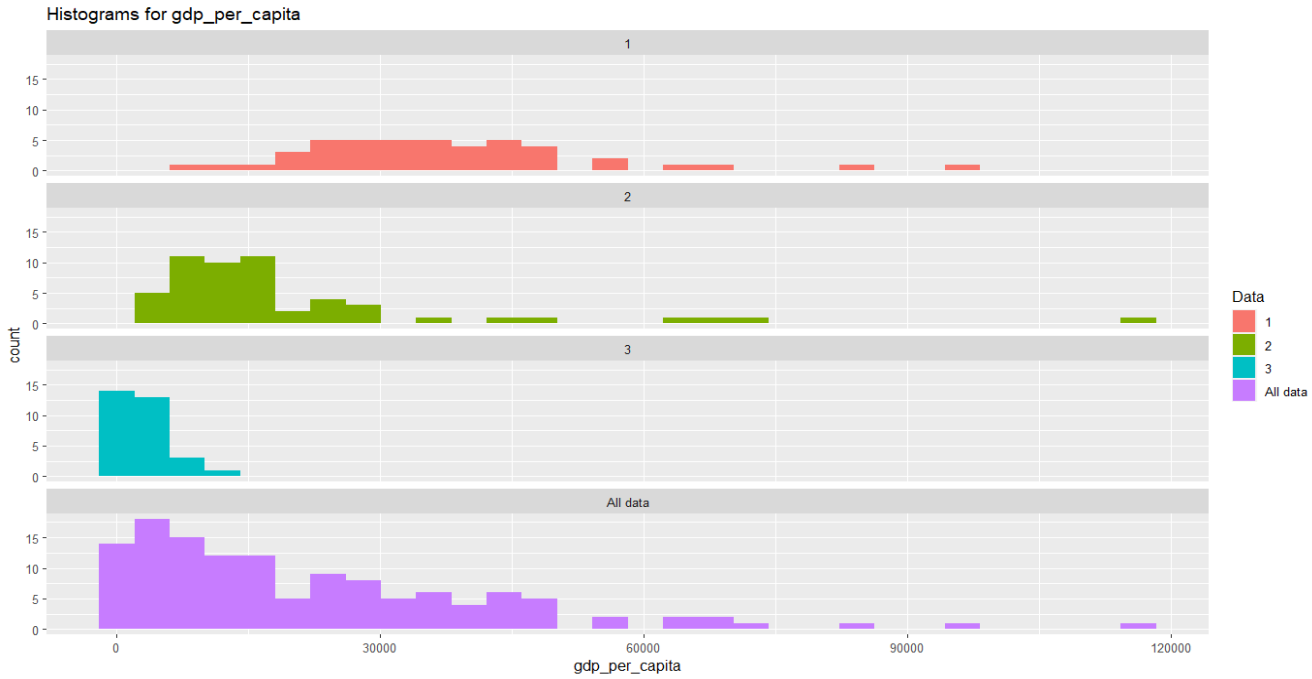
Από το Σχήμα 3.9 βλέπουμε ότι η πρώτη ομάδα περιέχει πολλές χώρες που έχουν πολύ μικρό δείκτη θνησιμότητας από καρδιαγγειακές ασθένειες το 2017. Επιπλέον, βλέπουμε ότι η δεύτερη και η τρίτη ομάδα περιέχουν πολλές χώρες που έχουν σχετικά μικρό δείκτη θνησιμότητας από καρδιαγγειακές ασθένειες το 2017.

- Μεταβλητή gdp_per_capita

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	37466.602	20219.608	3108.696	22124.08
Τυπική απόκλιση	17155.686	20596.333	2482.567	21137.16
Ελάχιστη τιμή	7894.393	3393.474	661.240	61.2
25% ποσοστημόριο	26777.56	8337.49	1566.16	6222.6
Διάμεσος	35220.084	14103.452	2064.236	15847.42
75% ποσοστημόριο	45229.245	22267.037	3958.441	32605.9
Μέγιστη τιμή	94277.96	116935.60	12294.88	116935.6

Πίνακας 3.13: Περιγραφικά μέτρα για την μεταβλητή gdp_per_capita

Από τον πίνακα 3.13 βλέπουμε ότι η πρώτη ομάδα έχει το υψηλότερο μέσο Α.Ε.Π και ότι η δεύτερη ομάδα έχει το αμέσως μεγαλύτερο μέσο Α.Ε.Π. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στο Α.Ε.Π και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στο Α.Ε.Π. Τα παραπάνω συμπεράσματα τα αναμέναμε, γιατί πιο πριν αναφέραμε ότι η πρώτη ομάδα περιέχει χώρες που έχουν πολύ ανεπτυγμένη οικονομία, η δεύτερη ομάδα περιέχει χώρες που έχουν μέτρια ανεπτυγμένη οικονομία και η τρίτη ομάδα περιέχει χώρες που δεν έχουν ανεπτυγμένη οικονομία.



Σχήμα 3.10: Ιστογράμματα για την μεταβλητή gdp_per_capita

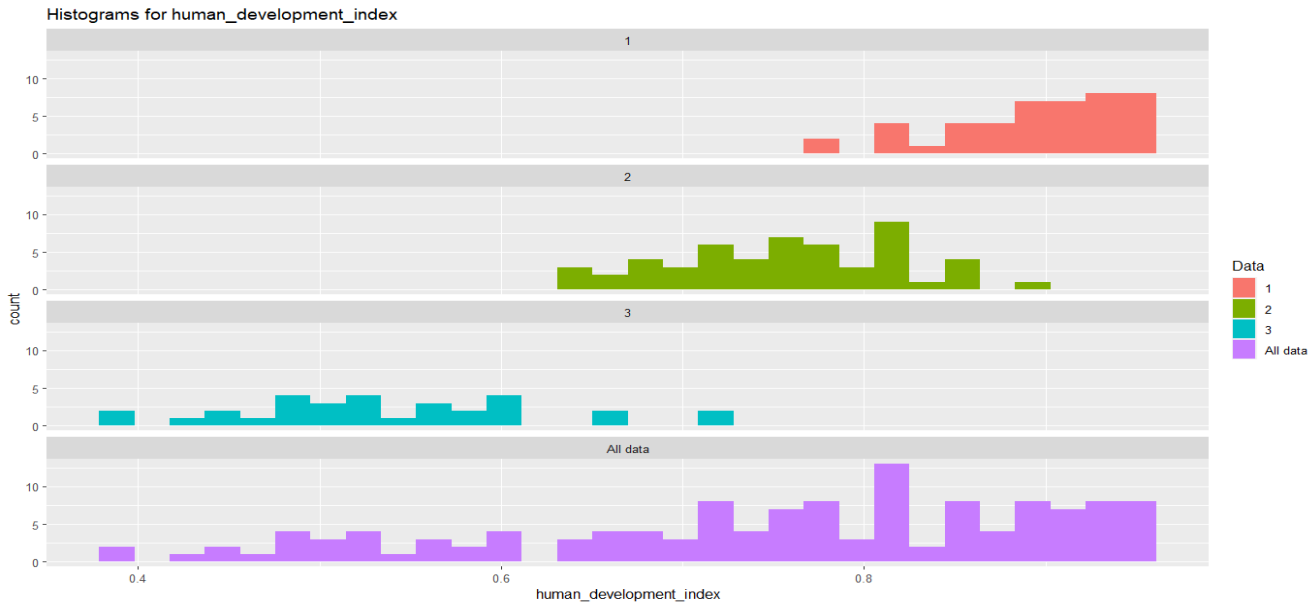
Από το Σχήμα 3.10 βλέπουμε ότι η δεύτερη και η τρίτη ομάδα περιέχουν πολλές χώρες που έχουν πολύ χαμηλό Α.Ε.Π. Από την άλλη μεριά, βλέπουμε ότι η πρώτη ομάδα περιέχει πολλές χώρες που έχουν αρκετά χαμηλό Α.Ε.Π.

- Μεταβλητή human_development_index

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	0.8955111	0.7572264	0.5401613	0.7533023
Τυπική απόκλιση	0.04827319	0.06245141	0.08297313	0.1487197
Ελάχιστη τιμή	0.779	0.632	0.394	0.3940
25% ποσοστημόριο	0.8640	0.7180	0.4815	0.6650
Διάμεσος	0.904	0.759	0.528	0.779
75% ποσοστημόριο	0.9320	0.8100	0.5925	0.8800
Μέγιστη τιμή	0.957	0.890	0.720	0.9570

Πίνακας 3.14: Περιγραφικά μέτρα για την μεταβλητή human_development_index

Από τον πίνακα 3.14 βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη μέση τιμή στον δείκτη που ποσοτικοποιεί την μέση πρόοδο του ανθρώπου στους τομείς: μακροζωία και υγεία, μόρφωση, καλό επίπεδο ζωής και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη μέση τιμή στον δείκτη που ποσοτικοποιεί την μέση πρόοδο του ανθρώπου στους τομείς: μακροζωία και υγεία, μόρφωση, καλό επίπεδο ζωής. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στον δείκτη που ποσοτικοποιεί την μέση πρόοδο του ανθρώπου στους τομείς: μακροζωία και υγεία, μόρφωση, καλό επίπεδο ζωής και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στον δείκτη που ποσοτικοποιεί την μέση πρόοδο του ανθρώπου στους τομείς: μακροζωία και υγεία, μόρφωση, καλό επίπεδο ζωής.



Σχήμα 3.11: Ιστογράμματα για την μεταβλητή human_development_index

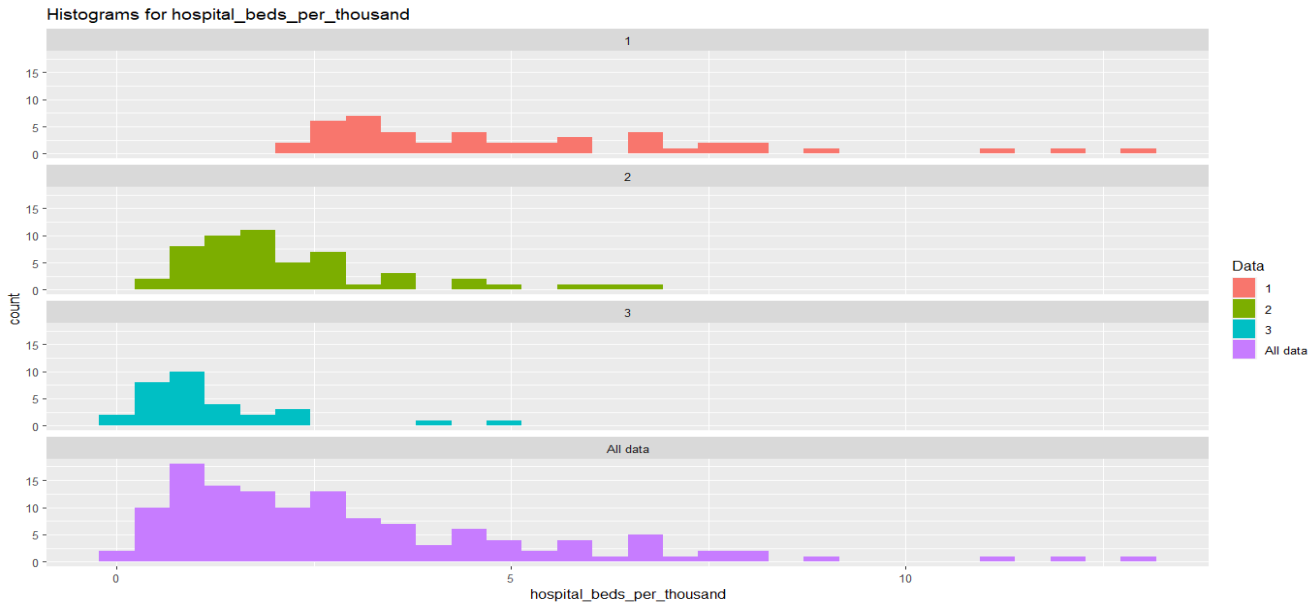
Από το Σχήμα 3.11 βλέπουμε ότι η πρώτη ομάδα περιέχει πολλές χώρες που έχουν υψηλές τιμές στον δείκτη που ποσοτικοποιεί την μέση πρόοδο του ανθρώπου στους τομείς: μακροζωία και υγεία, μόρφωση, καλό επίπεδο ζωής. Επιπλέον, βλέπουμε ότι η δεύτερη ομάδα πειρέχει πολλές χώρες που έχουν σχετικά υψηλές τιμές στον προαναφερθέντα δείκτη. Από την άλλη μεριά, βλέπουμε ότι η τρίτη ομάδα περιέχει πολλές χώρες που έχουν χαμηλές τιμές στον προαναφερθέντα δείκτη.

- Μεταβλητή hospital_beds_per_thousand

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	5.108244	2.219623	1.155484	2.971558
Τυπική απόκλιση	2.621140	1.411474	1.061806	2.464386
Ελάχιστη τιμή	2.22	0.53	0.10	0.100
25% ποσοστημόριο	2.99	1.30	0.50	1.200
Διάμεσος	4.50	1.71	0.80	2.3
75% ποσοστημόριο	6.62	2.81	1.40	4.000
Μέγιστη τιμή	13.05	6.70	4.80	13.050

Πίνακας 3.15: Περιγραφικά μέτρα για την μεταβλητή hospital_beds_per_thousand

Από τον πίνακα 3.15 βλέπουμε ότι η πρώτη ομάδα έχει τον υψηλότερο μέσο αριθμό κρεβατιών ανά 1000 άτομα και ότι η δεύτερη ομάδα έχει τον αμέσως μεγαλύτερο μέσο αριθμό κρεβατιών ανά 1000 άτομα. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό κρεβατιών ανά 1000 άτομα και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό κρεβατιών ανά 1000 άτομα. Τα παραπάνω συμπεράσματα τα αναμέναμε γιατί πιο πριν είχαμε αναφέρει ότι η πρώτη ομάδα περιέχει χώρες που είναι πολύ ανεπτυγμένες στον τομέα της υγείας, η δεύτερη ομάδα περιέχει χώρες που είναι μέτρια ανεπτυγμένες στον τομέα της υγείας, η τρίτη ομάδα περιέχει χώρες που δεν είναι καθόλου ανεπτυγμένες στο τομέα της υγείας.



Σχήμα 3.12: Ιστογράμματα για την μεταβλητή hospital_beds_per_thousand

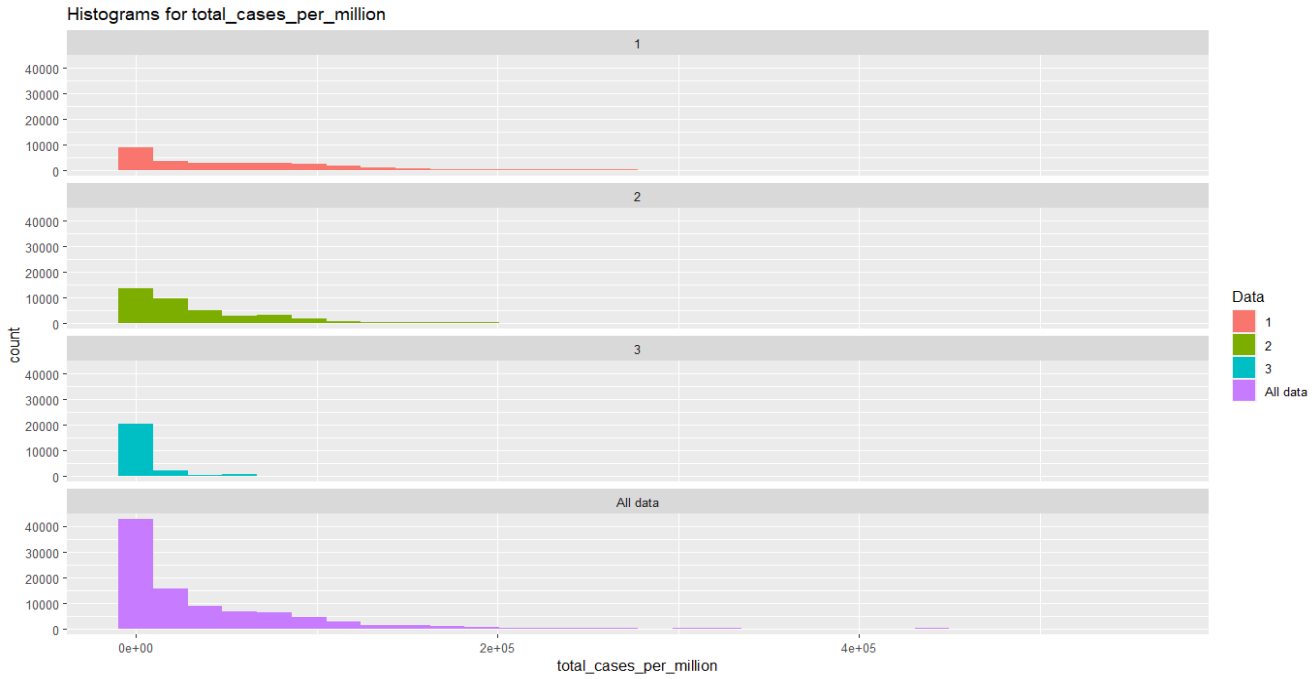
Από το Σχήμα 3.12 βλέπουμε ότι η πρώτη ομάδα περιέχει αρκετές χώρες που έχουν σχετικά μικρό αριθμό κρεβατιών στα νοσοκομεία ανά 1000 άτομα. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα περιέχει και χώρες που έχουν μέτριο αριθμό κρεβατιών στα νοσοκομεία ανά 1000 άτομα αλλά και κάποιες χώρες που έχουν υψηλό αριθμό κρεβατιών στα νοσοκομεία ανά 1000 άτομα. Από την άλλη μεριά, βλέπουμε ότι η δεύτερη και η τρίτη ομάδα περιέχουν πολλές χώρες που έχουν πολύ χαμηλό αριθμό κρεβατιών στα νοσοκομεία ανά 1000 άτομα.

- Μεταβλητή total_cases_per_million

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	90981.867	40460.450	6124.575	49822.12
Τυπική απόκλιση	111372.16	54303.60	13543.38	81641.44
Ελάχιστη τιμή	128.382	15.000	2.187	2.2
25% ποσοστημόριο	8425.246	4719.337	555.982	2188.7
Διάμεσος	55919.401	21800.366	1659.953	15318.83
75% ποσοστημόριο	116177.198	58076.135	4438.314	66573.5
Μέγιστη τιμή	555189.86	441395.38	98701.37	555189.9

Πίνακας 3.16: Περιγραφικά μέτρα για την μεταβλητή total_cases_per_million

Από τον πίνακα 3.16 βλέπουμε ότι η πρώτη ομάδα έχει τον υψηλότερο μέσο αριθμό συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει τον αμέσως μεγαλύτερο μέσο αριθμό συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού. Ακόμη, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού.



Σχήμα 3.13: Ιστογράμματα για την μεταβλητή total_cases_per_million

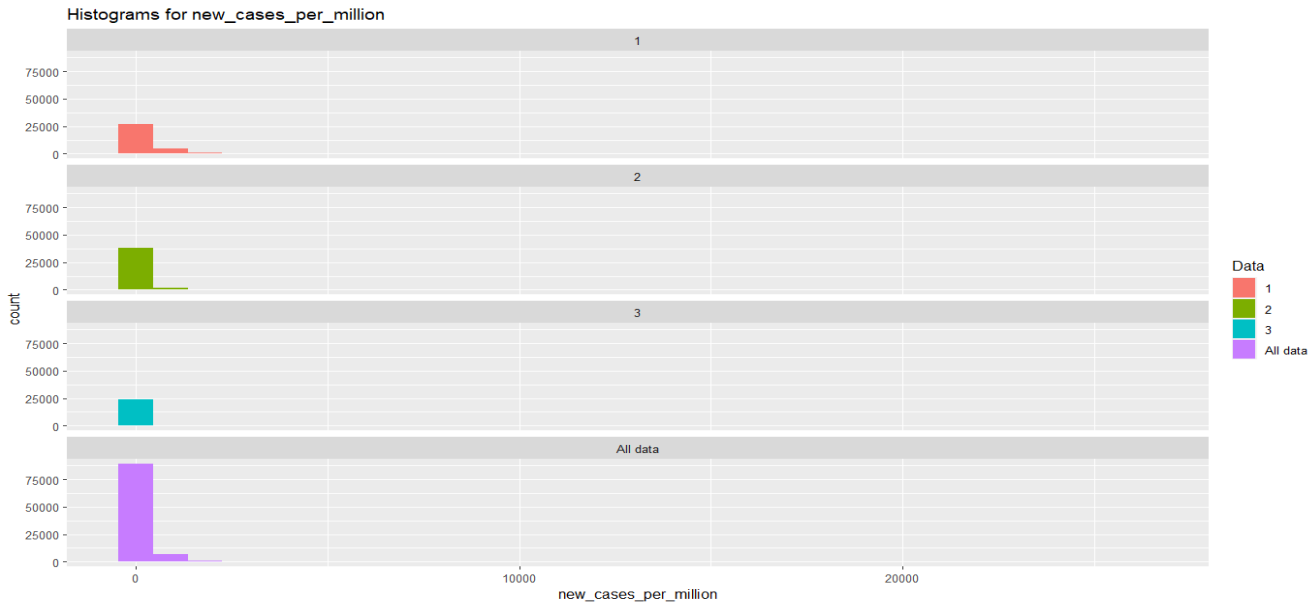
Από το Σχήμα 3.13 βλέπουμε ότι τα ιστογράμματα που αντιστοιχούν στην πρώτη και στην δεύτερη ομάδα έχουν μακριά δεξιά "ουρά". Αυτό σημαίνει ότι στην πρώτη και στην δεύτερη ομάδα υπάρχουν χώρες στις οποίες υπήρξαν μέρες όπου ο αριθμός συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού ήταν αρκετά υψηλός. Από την άλλη μεριά, βλέπουμε ότι στο ιστόγραμμα της τρίτης ομάδας η δεξιά "ουρά" είναι αρκετά πιο στενή. Αυτό σημαίνει πως στις χώρες της τρίτης ομάδας ο αριθμός συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού ήταν χαμηλός.

- Μεταβλητή new_cases_per_million

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	404.2426	129.0260	16.3838	197.8971
Τυπική απόκλιση	921.35049	378.90925	66.56145	617.0036
Ελάχιστη τιμή	0	0	0	0.000
25% ποσοστημόριο	16.26675	5.32200	0.00000	1.809
Διάμεσος	110.864	39.343	1.486	28.477
75% ποσοστημόριο	366.2793	129.7620	8.6280	149.611
Μέγιστη τιμή	26186.094	23100.819	2498.618	26186.094

Πίνακας 3.17: Περιγραφικά μέτρα για την μεταβλητή new_cases_per_million

Από τον πίνακα 3.17 βλέπουμε ότι η πρώτη ομάδα έχει τον υψηλότερο μέσο αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει τον αμέσως μεγαλύτερο μέσο αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού.



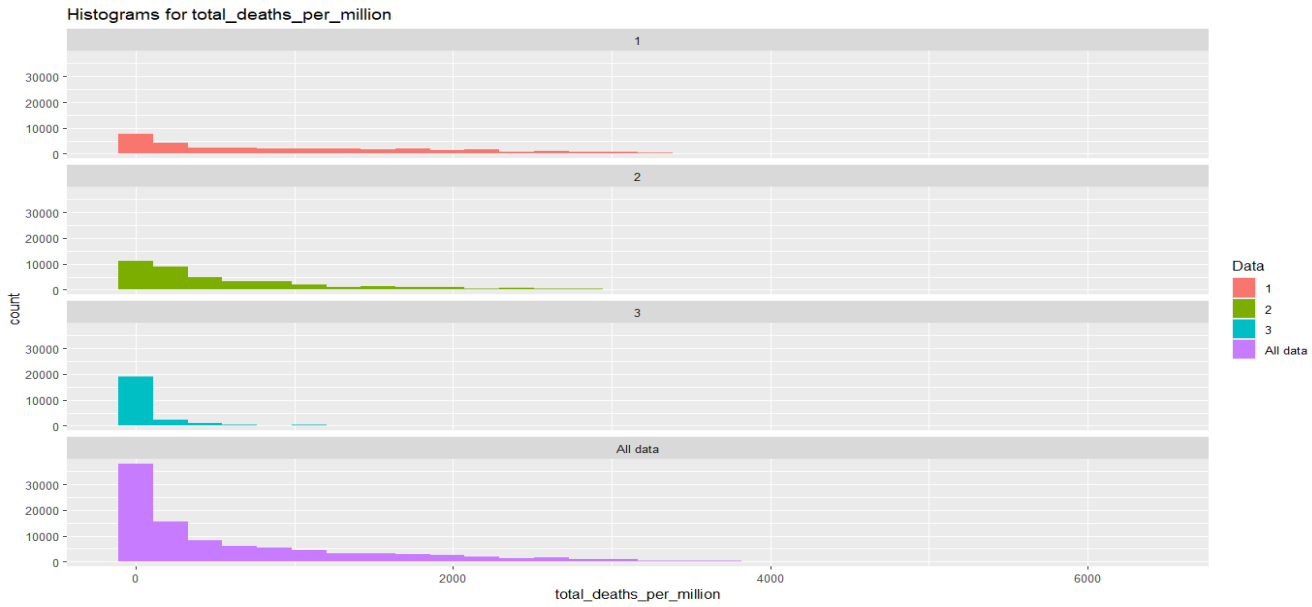
Σχήμα 3.14: Ιστογράμματα για την μεταβλητή new_cases_per_million

- Μεταβλητή total_deaths_per_million

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	1099.4983	694.1226	112.5795	695.7226
Τυπική απόκλιση	1079.8073	943.3745	248.6464	962.1537
Ελάχιστη τιμή	3.781	0.355	0.031	0.031
25% ποσοστημόριο	143.5865	86.2945	11.3070	37.715
Διάμεσος	825.8575	335.7530	29.1710	241.444
75% ποσοστημόριο	1768.908	942.380	71.631	1017.542
Μέγιστη τιμή	5405.854	6328.904	1710.476	6328.904

Πίνακας 3.18: Περιγραφικά μέτρα για την μεταβλητή total_deaths_per_million

Από τον πίνακα 3.18 βλέπουμε ότι η πρώτη ομάδα έχει τον υψηλότερο μέσο αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει τον αμέσως μεγαλύτερο μέσο αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει τον αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού.



Σχήμα 3.15: Ιστογράμματα για την μεταβλητή total_deaths_per_million

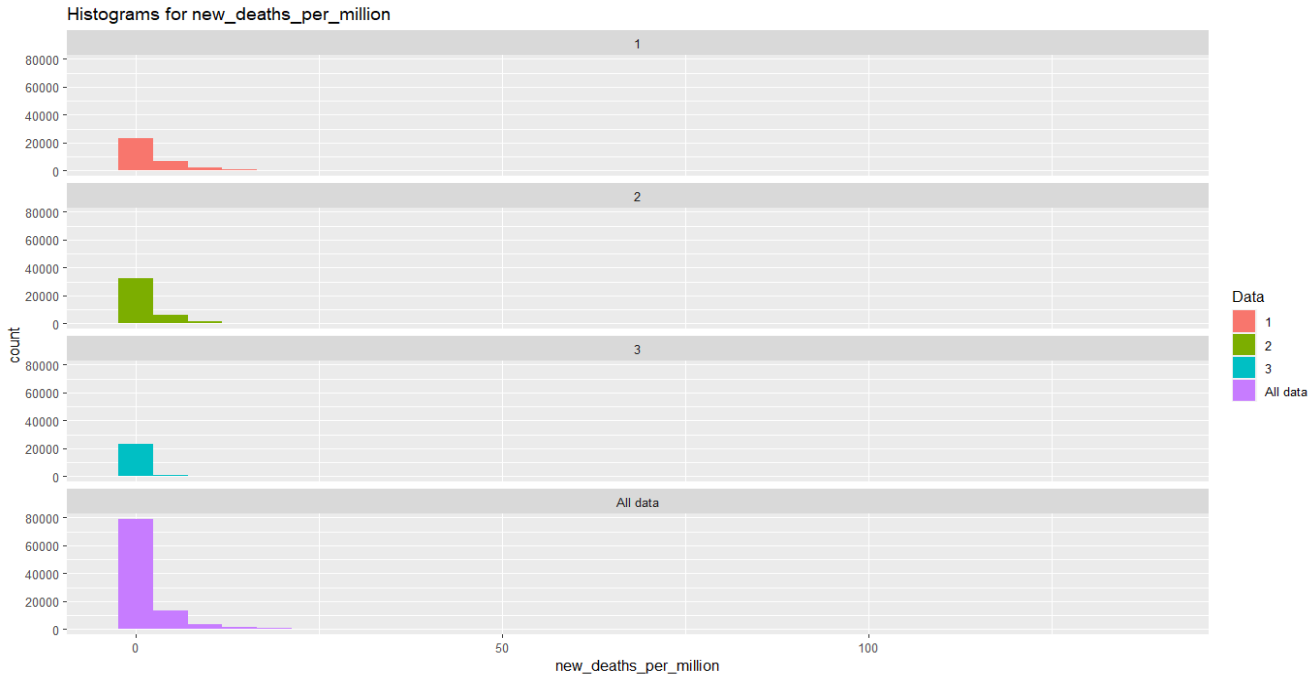
Από το Σχήμα 3.15 βλέπουμε ότι τα ιστογράμματα που αντιστοιχούν στην πρώτη και δεύτερη ομάδα έχουν πολύ μεγάλη δεξιά "ουρά". Αυτό σημαίνει πως στην πρώτη και στην δεύτερη ομάδα υπάρχουν χώρες στις οποίες υπήρξαν μέρες που ο αριθμός συνολικών θανάτων ανά εκατομμύριο πληθυσμού ήταν υψηλός. Από την άλλη μεριά, βλέπουμε ότι στο ιστόγραμμα της τρίτης ομάδας η δεξιά "ουρά" είναι αρκετά πιο στενή. Αυτό σημαίνει ότι οι χώρες της τρίτης ομάδας είχαν χαμηλό αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού.

- Μεταβλητή new_deaths_per_million

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	2.6698188	1.6541050	0.2708107	1.675849
Τυπική απόκλιση	4.416400	3.436379	1.003566	3.565815
Ελάχιστη τιμή	0	0	0	0.000
25% ποσοστημόριο	0	0	0	0.000
Διάμεσος	0.835	0.460	0.000	0.266
75% ποσοστημόριο	3.362	1.863	0.116	1.722
Μέγιστη τιμή	74.011	137.092	20.873	137.092

Πίνακας 3.19: Περιγραφικά μέτρα για την μεταβλητή new_deaths_per_million

Από τον πίνακα 3.19 βλέπουμε ότι η πρώτη ομάδα έχει τον υψηλότερο μέσο ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει τον αμέσως μεγαλύτερο μέσο ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει τον ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει τον ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού.



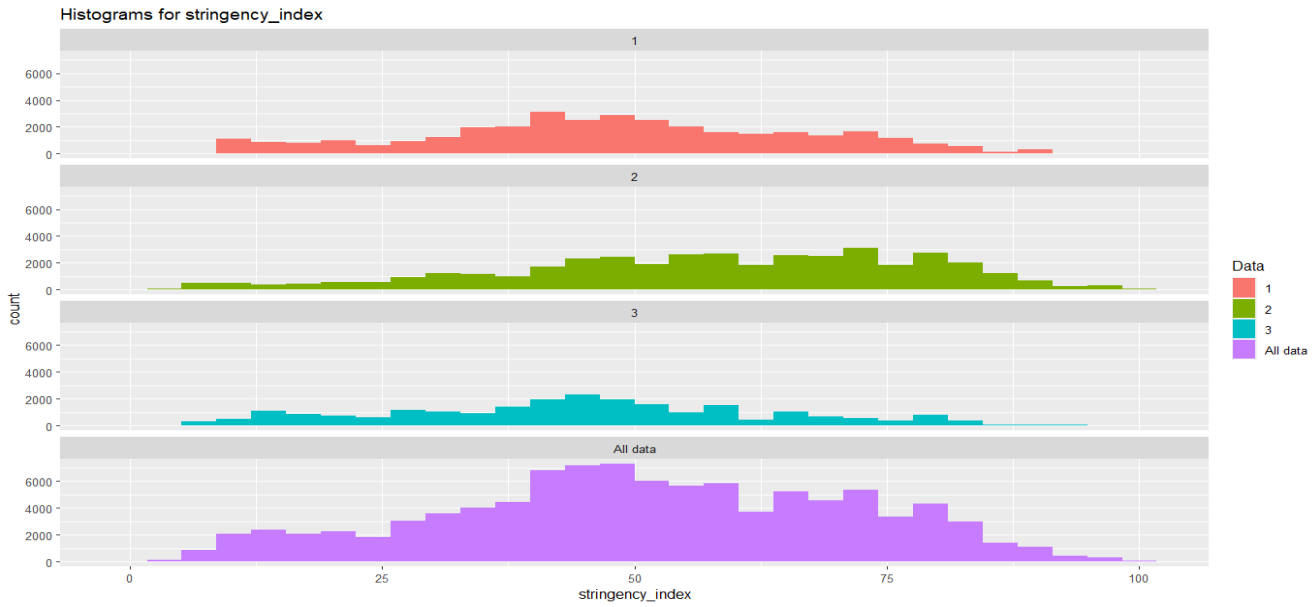
Σχήμα 3.16: Ιστογράμματα για την μεταβλητή new_deaths_per_million

- Μεταβλητή stringency_index

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	48.01687	57.56302	44.76385	51.16097
Τυπική απόκλιση	18.42990	20.11846	19.02050	20.04656
Ελάχιστη τιμή	5.56	0.00	2.78	0.00
25% ποσοστημόριο	36.15	44.54	31.48	37.96
Διάμεσος	47.22	59.71	44.44	50.93
75% ποσοστημόριο	61.46	72.69	57.41	66.67
Μέγιστη τιμή	96.30	100.00	97.22	100.00

Πίνακας 3.20: Περιγραφικά μέτρα για την μεταβλητή stringency_index

Από τον πίνακα 3.20 βλέπουμε ότι η δεύτερη ομάδα έχει την υψηλότερη μέση σφοδρότητα των κυβερνητικών μέτρων και ότι η πρώτη ομάδα έχει την αμέσως μεγαλύτερη μέση σφοδρότητα των κυβερνητικών μέτρων. Επιπλέον, βλέπουμε ότι η δεύτερη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει την σφοδρότητα των κυβερνητικών μέτρων και ότι η πρώτη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει την σφοδρότητα των κυβερνητικών μέτρων. Επιπροσθέτως, βλέπουμε ότι και στις 3 ομάδες η μέγιστη σφοδρότητα των κυβερνητικών μέτρων ήταν πάρα πολύ υψηλή. Αυτό σημαίνει ότι και στις 3 ομάδες υπήρξαν μέρες όπου η σφοδρότητα των κυβερνητικών μέτρων ήταν πάρα πολύ υψηλή.



Σχήμα 3.17: Ιστογράμματα για την μεταβλητή stringency_index

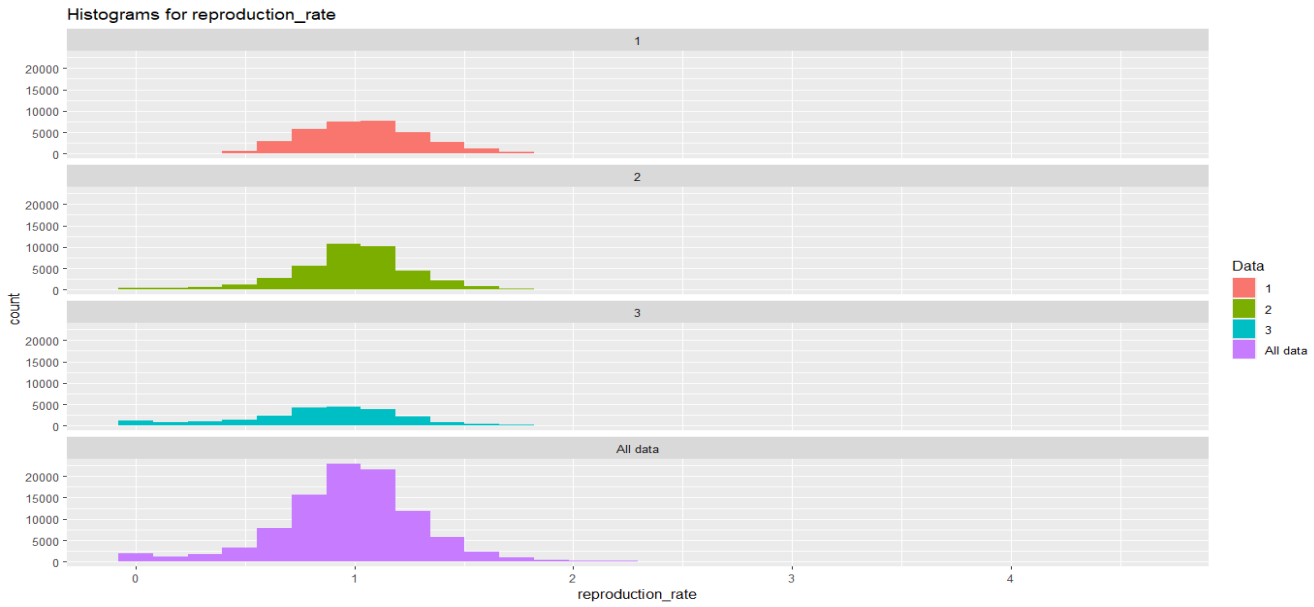
Από το Σχήμα 3.17 βλέπουμε ότι σε όλες τις ομάδες υπήρξαν μέρες όπου η σφοδρότητα των κυβερνητικών μέτρων ήταν χαμηλή, μέτρια, υψηλή. Επιπροσθέτως, βλέπουμε ότι στην τρίτη ομάδα στις πιο πολλές μέρες η σφοδρότητα των κυβερνητικών μέτρων ήταν μέτρια. Από την άλλη μεριά, στην πρώτη και στην δεύτερη ομάδα βλέπουμε ότι τις πιο πολλές μέρες η σφοδρότητα των κυβερνητικών μέτρων ήταν είτε μέτρια είτε υψηλή.

- Μεταβλητή reproduction_rate

	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ανεξάρτητα από ομάδα
Μέση τιμή	1.0455432	0.9917531	0.8729517	0.9819638
Τυπική απόκλιση	0.2879025	0.3097536	0.4155919	0.3378424
Ελάχιστη τιμή	-0.01	0.00	-0.03	-0.030
25% ποσοστημόριο	0.85	0.85	0.66	0.810
Διάμεσος	1.03	1.00	0.90	0.99
75% ποσοστημόριο	1.21	1.15	1.11	1.160
Μέγιστη τιμή	4.56	4.26	4.41	4.560

Πίνακας 3.21: Περιγραφικά μέτρα για την μεταβλητή reproduction_rate

Από τον πίνακα 3.21 βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη μέση μεταδοση του ιού και ότι η δεύτερη ομάδα είχε την αμέσως μεγαλύτερη μέση μετάδοση του ιού. Επιπλέον, βλέπουμε ότι η πρώτη ομάδα έχει την υψηλότερη διάμεσο στην μεταβλητή που εκφράζει την μετάδοση του ιού και ότι η δεύτερη ομάδα έχει την αμέσως μεγαλύτερη διάμεσο στην μεταβλητή που εκφράζει την μετάδοση του ιού. Επιπροσθέτως, βλέπουμε ότι η μέγιστη μετάδοση του ιού και στις 3 ομάδες είναι πολύ μεγαλύτερη από το 1. Αυτό σημαίνει πως και στις 3 ομάδες υπήρξαν μέρες όπου η μετάδοση του ιού ήταν πάρα πολύ υψηλή.



Σχήμα 3.18: Ιστογράμματα για την μεταβλητή reproduction_rate

Από το Σχήμα 3.18 βλέπουμε ότι τα ιστογράμματα των τριών ομάδων μοιάζουν πολύ με το ιστόγραμμα της Κανονικής Κατανομής με μέση τιμή 1. Επιπλέον, από το Σχήμα 3.18 βλέπουμε ότι στα ιστογράμματα των ομάδων 2 και 3 η αριστερή "ουρά" είναι αρκετά μακριά. Αυτό σημαίνει πως στις ομάδες 2 και 3 υπάρχουν χώρες στις οποίες υπήρξαν μέρες που η μετάδοση του ιού ήταν χαμηλή (R<1). Επιπλέον, από τα ιστογράμματα της πρώτης και της δεύτερης ομάδας μπορούμε να συμπεράνουμε ότι σε αυτές τις ομάδες υπάρχουν χώρες στις οποίες υπήρξαν αρκετές μέρες όπου η μετάδοση του ιού ήταν υψηλή (R>1).

3.4 Ανάλυση Κυρίων Συνιστωσών

Πριν εφαρμοστεί η Ανάλυση Κυρίων Συνιστωσών στο σύνολο δεδομένων έγιναν κάποιες τροποποιήσεις. Πιο συγκεκριμένα, υπολογίσαμε την μέση τιμή κάθε μεταβλητής σε κάθε χώρα και για κάθε μήνα που ανήκει στο χρονικό διάστημα 13-5-2020 έως 18-6-22. Αυτό είχε σαν αποτέλεσμα να δημιουργηθούν 156 νέες μεταβλητές. Έπειτα, εφαρμόστηκε η Ανάλυση Κυρίων Συνιστωσών στο τροποποιημένο σύνολο δεδομένων χρησιμοποιώντας τον πίνακα των τυποποιημένων δεδομένων R.

Στον Πίνακα 3.22 μπορούμε να δούμε το αθροιστικό ποσοστό της μεταβλητότητας του συνόλου δεδομένων που εξηγείται για τις 10 πρώτες κύριες συνιστώσες.

Αύξων αριθμός κύριας συνιστώσας	Αθροιστικό ποσοστό μεταβλητότητας των αρχικών δεδομένων που εξηγείται
1	33.57690%
2	45.50873%
3	53.67055%
4	60.68885%
5	65.43660%
6	68.80200%
7	71.94215%
8	74.85202%
9	77.07571%
10	78.95905%

Πίνακας 3.22: Αθροιστικά ποσοστά μεταβλητότητας του συνόλου δεδομένων για τις πρώτες 10 κύριες συνιστώσες

Το πλήθος των ιδιοτιμών που είναι μεγαλύτερες από 1 είναι ίσο με 4, οπότε με βάση το κριτήριο Kaiser θα επιλέξουμε τις 4 πρώτες κύριες συνιστώσες. Για να μπορέσουμε να ερμηνεύσουμε τις 4 πρώτες κύριες συνιστώσες θα πρέπει να υπολογίσουμε τις συσχετίσεις μεταξύ των scores των κυρίων συνιστωσών και των αρχικών μεταβλητών. Επειδή το πλήθος των μεταβλητών είναι αρκετά μεγάλο στον Πίνακα 3.23 δίνονται μόνο οι συσχετίσεις που είναι κατά απόλυτη τιμή μεγαλύτερες ή ίσες από το 0.50.

Μεταβλητή	Κύρια συνιστώσα	Συντελεστής γραμμικής συσχέτισης
Oct_2020_tot_cases_per_million	PC1	-0.561346112
Nov_2020_tot_cases_per_million	PC1	-0.783528175
Dec_2020_tot_cases_per_million	PC1	-0.871353411
Jan_2021_tot_cases_per_million	PC1	-0.896238298
Feb_2021_tot_cases_per_million	PC1	-0.89561416
Mar_2021_tot_cases_per_million	PC1	-0.898403126
Apr_2021_tot_cases_per_million	PC1	-0.907195311
May_2021_tot_cases_per_million	PC1	-0.907516598
Jun_2021_tot_cases_per_million	PC1	-0.897308829
Jul_2021_tot_cases_per_million	PC1	-0.894920311
Aug_2021_tot_cases_per_million	PC1	-0.89584914
Sep_2021_tot_cases_per_million	PC1	-0.896164928
Oct_2021_tot_cases_per_million	PC1	-0.900163683
Nov_2021_tot_cases_per_million	PC1	-0.904380359
Dec_2021_tot_cases_per_million	PC1	-0.904208278
Jan_2022_tot_cases_per_million	PC1	-0.887333799
Feb_2022_tot_cases_per_million	PC1	-0.836598854
Mar_2022_tot_cases_per_million	PC1	-0.770589933
Apr_2022_tot_cases_per_million	PC1	-0.733590791
May_2022_tot_cases_per_million	PC1	-0.72333078
Jun_2022_tot_cases_per_million	PC1	-0.717438289
Sep_2020_new_cases_per_million	PC1	-0.535425576
Oct_2020_new_cases_per_million	PC1	-0.742258944
Nov_2020_new_cases_per_million	PC1	-0.777440191
Dec_2020_new_cases_per_million	PC1	-0.767211244

Jan_2021_new_cases_per_million	PC1	-0.716650686
Feb_2021_new_cases_per_million	PC1	-0.714785734
Mar_2021_new_cases_per_million	PC1	-0.70966939
Apr_2021_new_cases_per_million	PC1	-0.67743
Oct_2021_new_cases_per_million	PC1	-0.527931708
Nov_2021_new_cases_per_million	PC1	-0.630131937
Dec_2021_new_cases_per_million	PC1	-0.613957848
Jan_2022_new_cases_per_million	PC1	-0.652222875
May_2020_total_deaths_per_million	PC1	-0.513981016
Jun_2020_total_deaths_per_million	PC1	-0.548559993
Jul_2020_total_deaths_per_million	PC1	-0.557997463
Aug_2020_total_deaths_per_million	PC1	-0.551799797
Sep_2020_total_deaths_per_million	PC1	-0.548923997
Oct_2020_total_deaths_per_million	PC1	-0.579030581
Nov_2020_total_deaths_per_million	PC1	-0.697422204
Dec_2020_total_deaths_per_million	PC1	-0.816395675
Jan_2021_total_deaths_per_million	PC1	-0.873597559
Feb_2021_total_deaths_per_million	PC1	-0.890821318
Mar_2021_total_deaths_per_million	PC1	-0.898279472
Apr_2021_total_deaths_per_million	PC1	-0.898728669
May_2021_total_deaths_per_million	PC1	-0.896823821
Jun_2021_total_deaths_per_million	PC1	-0.896706821
Jul_2021_total_deaths_per_million	PC1	-0.890106627
Aug_2021_total_deaths_per_million	PC1	-0.883894684
Sep_2021_total_deaths_per_million	PC1	-0.883410886
Oct_2021_total_deaths_per_million	PC1	-0.88344646
Nov_2021_total_deaths_per_million	PC1	-0.883769975
Dec_2021_total_deaths_per_million	PC1	-0.883763972
Jan_2022_total_deaths_per_million	PC1	-0.882814803
Feb_2022_total_deaths_per_million	PC1	-0.882523757
Mar_2022_total_deaths_per_million	PC1	-0.885342743
Apr_2022_total_deaths_per_million	PC1	-0.889040935
May_2022_total_deaths_per_million	PC1	-0.891342148
Jun_2022_total_deaths_per_million	PC1	-0.892379183
Oct_2020_new_deaths_per_million	PC1	-0.695760498
Nov_2020_new_deaths_per_million	PC1	-0.753220726
Dec_2020_new_deaths_per_million	PC1	-0.775163673
Jan_2021_new_deaths_per_million	PC1	-0.799747029
Feb_2021_new_deaths_per_million	PC1	-0.800266062
Mar_2021_new_deaths_per_million	PC1	-0.713557508
Apr_2021_new_deaths_per_million	PC1	-0.655882245
May_2021_new_deaths_per_million	PC1	-0.573956944
Nov_2021_new_deaths_per_million	PC1	-0.521188566
Dec_2021_new_deaths_per_million	PC1	-0.576024683
Jan_2022_new_deaths_per_million	PC1	-0.664534155
Feb_2022_new_deaths_per_million	PC1	-0.776041115
Mar_2022_new_deaths_per_million	PC1	-0.612540415
Oct_2020_reprod_rate	PC1	-0.502036083

Oct_2021_reprod_rate	PC1	-0.571374163
Nov_2021_reprod_rate	PC1	-0.510934172
May_2020_str_ind	PC2	0.668876599
Jun_2020_str_ind	PC2	0.762060952
Jul_2020_str_ind	PC2	0.789701315
Aug_2020_str_ind	PC2	0.789867151
Sep_2020_str_ind	PC2	0.765716992
Oct_2020_str_ind	PC2	0.704689046
Jun_2021_str_ind	PC2	0.579438131
Jul_2021_str_ind	PC2	0.68605411
Aug_2021_str_ind	PC2	0.660900033
Sep_2021_str_ind	PC2	0.593121183
Dec_2020_str_ind	PC3	-0.567010564
Jan_2021_str_ind	PC3	-0.593256566
Feb_2021_str_ind	PC3	-0.564559874
Mar_2021_str_ind	PC3	-0.539960246
Apr_2021_str_ind	PC3	-0.519821907
May_2021_str_ind	PC3	-0.524475367
Jun_2021_str_ind	PC3	-0.526226801
Oct_2021_str_ind	PC3	-0.500553062
Nov_2021_str_ind	PC3	-0.529037705
Dec_2021_str_ind	PC3	-0.585751698
Jan_2022_str_ind	PC3	-0.618051139
Feb_2022_str_ind	PC3	-0.559547852
May_2020_tot_cases_per_million	PC4	-0.658685336
Jun_2020_tot_cases_per_million	PC4	-0.687953173
Jul_2020_tot_cases_per_million	PC4	-0.703307345
Aug_2020_tot_cases_per_million	PC4	-0.676948883
Sep_2020_tot_cases_per_million	PC4	-0.65046533
Oct_2020_tot_cases_per_million	PC4	-0.622971658
May_2020_new_cases_per_million	PC4	-0.594663954
Jun_2020_new_cases_per_million	PC4	-0.640866362
Jul_2020_new_cases_per_million	PC4	-0.509185851

Πίνακας 3.23: Συσχετίσεις μεταξύ αρχικών μεταβλητών και κυρίων συνιστωσών

Με βάση τις συσχετίσεις μεταξύ των αρχικών μεταβλητών και των κυρίων συνιστωσών που φαίνονται στον Πίνακα 3.23 μπορούμε να πούμε ότι για την:

α. Πρώτη κύρια συνιστώσα:

Η πρώτη κύρια συνιστώσα σχετίζεται

- αρνητικά με τα συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 – Ιούνιος 2022
- αρνητικά με τον ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Σεπτέμβριος 2020 – Απρίλιος 2021 και στο χρονικό διάστημα Οκτώβριος 2021 – Ιανουάριος 2022.

- αρνητικά με τους συνολικούς θανάτους ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάιος 2020 – Ιούνιος 2022.
- αρνητικά με τον ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 – Μάϊος 2021 και στο χρονικό διάστημα Νοέμβριος 2021 – Μάρτιος 2022.
- αρνητικά με την εκτίμηση του αριθμού των ατόμων που αναμένονταν να μολυνθούν από τον ιό σε περίπτωση που ερχόντουσαν σε επαφή με ένα κρούσμα τον Οκτώβριο 2020 και στο χρονικό διάστημα Οκτώβριος 2021 – Νοέμβριος 2021.

β. Δεύτερη κύρια συνιστώσα

Η δεύτερη κύρια συνιστώσα σχετίζεται θετικά με την σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα: Μάϊος 2020 – Οκτώβριος 2020, Ιούνιος 2021 – Σεπτέμβριος 2021.

γ. Τρίτη κύρια συνιστώσα

Η τρίτη κύρια συνιστώσα σχετίζεται αρνητικά με την σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022.

δ. Τέταρτη κύρια συνιστώσα

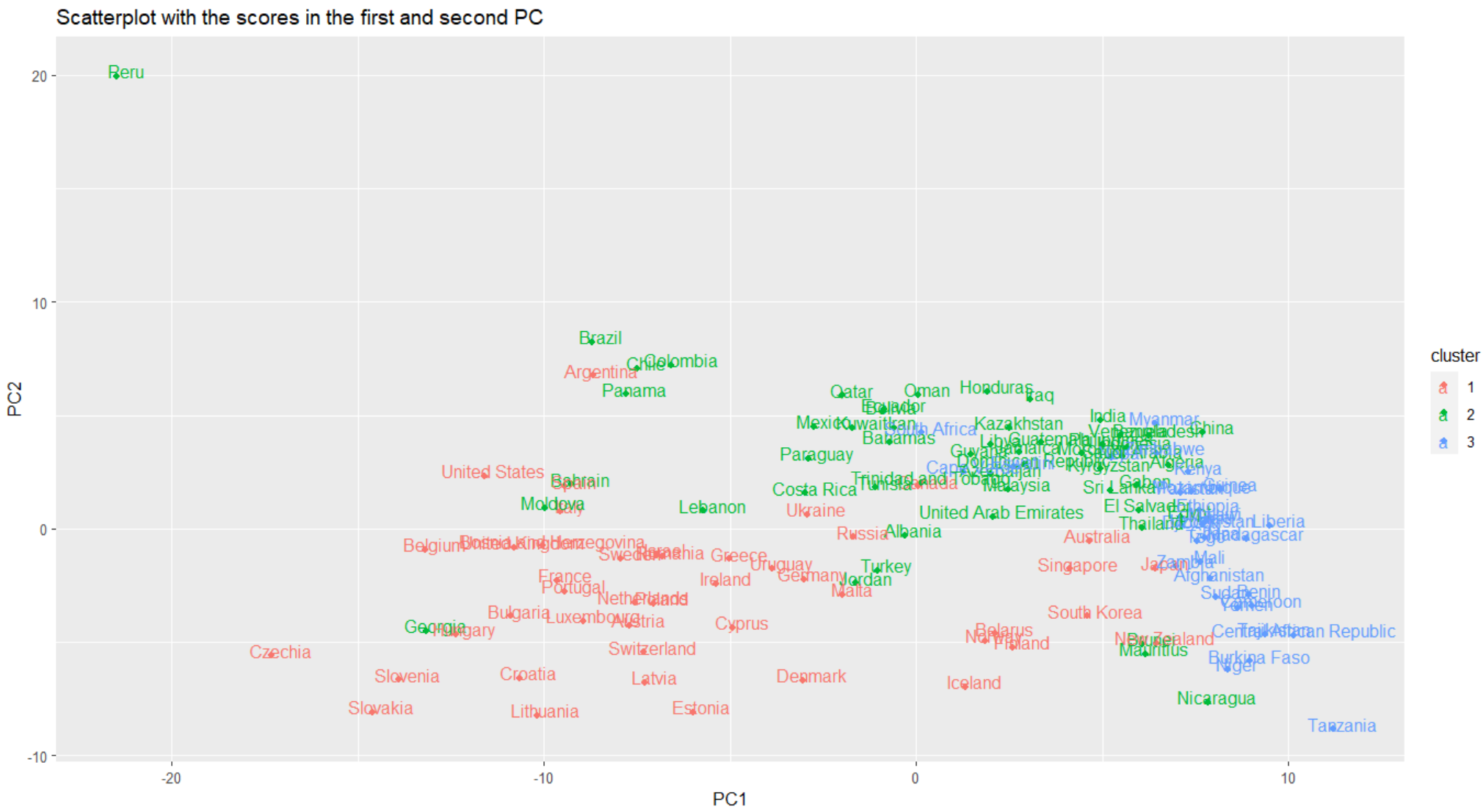
Η τέταρτη κύρια συνιστώσα σχετίζεται αρνητικά με

- τα συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Οκτώβριος 2020.
- τον ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Ιούλιος 2020.

Με βάση τα παραπάνω συμπεράσματα η ερμηνεία των κυρίων συνιστωσών είναι η ακόλουθη:

- Η πρώτη κύρια συνιστώσα εκφράζει το πόσο καλή ήταν η επιδημιολογική εικόνα μιας χώρας.
- Η δεύτερη και η τρίτη κύρια συνιστωσα σχετίζονται με την σφοδρότητα των κυβερνητικών μέτρων. Πιο συγκεκριμένα, η δεύτερη κύρια συνιστώσα εκφράζει υψηλή ήταν η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 – Οκτώβριος 2020, Ιούνιος 2021 – Σεπτέμβριος 2021, ενώ η τρίτη κύρια συνιστώσα εκφράζει το πόσο χαμηλή ήταν η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022.
- Η τέταρτη κύρια συνιστώσα σχετίζεται κυρίως με τον αριθμό κρουσμάτων. Πιο συγκεκριμένα, εκφράζει πόσο χαμηλά ήταν τα συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Οκτώβριος 2020 και πόσο χαμηλά ήταν τα ημερήσια κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Ιούλιος 2020.

Στην συνέχεια θα δούμε κάποια scatter plots (διαγράμματα διασποράς) στα οποία θα απεικονίζονται τα scores όλων των χωρών στις κύριες συνιστώσες με σκοπό να δούμε τις πιθανές διαφοροποιήσεις που υπήρξαν ανάμεσα σε χώρες αλλά και ομάδες χωρών.



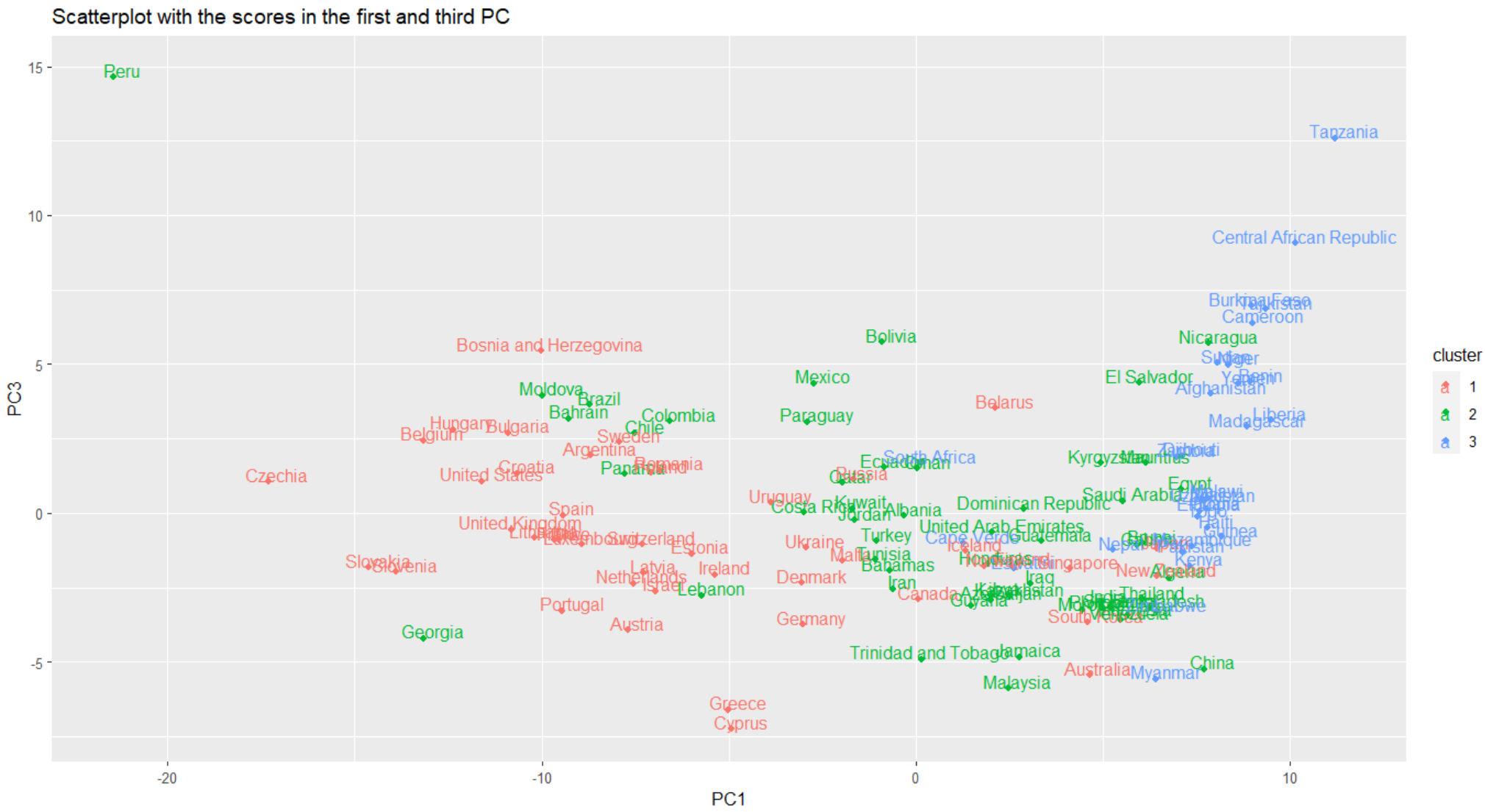
Σχήμα 3.19: Scatter plot (Διάγραμμα διασποράς) με τα scores των χωρών στην πρώτη και στην δεύτερη κύρια συνιστώσα

Από το Σχήμα 3.19 βλέπουμε επίσης ότι υπάρχουν χώρες της πρώτης ομάδας (κόκκινο χρώμα) που παίρνουν υψηλές τιμές στην πρώτη κύρια συνιστώσα και χώρες που παίρνουν χαμηλές τιμές στην πρώτη κύρια συνιστώσα. Πιο συγκεκριμένα, βλέπουμε ότι η Αυστραλία, η Σιγκαπούρη, Ιαπωνία, Νότια Κορέα, Νέα Ζηλανδία, Λευκορωσία, Ισλανδία λαμβάνουν υψηλές τιμές στην πρώτη κύρια συνιστώσα το οποίο σημαίνει ότι σε αυτές τις χώρες η επιδημιολογική εικόνα ήταν αρκετά καλή. Δηλαδή αυτό σημαίνει ότι οι προαναφερθείσες χώρες εμφάνισαν χαμηλό συνολικό αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Ιούνιος 2022, χαμηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στα χρονικά διαστήματα Σεπτέμβριος 2020 - Απρίλιος 2021, Οκτώβριος 2021 - Ιανουάριος 2022, χαμηλό συνολικό αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Ιούνιος 2022, χαμηλό ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Μάϊος 2021 και στο χρονικό διάστημα Νοέμβριος 2021 - Μάρτιος 2022, χαμηλή μετάδοση του ιού τον Οκτώβριο 2020 και στο χρονικό διάστημα Οκτώβριος 2021 - Νοέμβριος 2021. Από την άλλη μεριά, παρατηρούμε ότι η Τσεχία, η Σλοβενία, η Σλοβακία, η Κροατία, η Λιθουανία, η Ουγγαρία, η Βουλγαρία, το Βέλγιο λαμβάνουν χαμηλές τιμές στην πρώτη κύρια συνιστώσα το οποίο σημαίνει ότι σε αυτές τις χώρες η επιδημιολογική εικόνα ήταν αρκετά κακή. Δηλαδή αυτό σημαίνει ότι στις προαναφερθείσες χώρες εμφάνισαν υψηλό συνολικό αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Ιούνιος 2022, υψηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στα χρονικά διαστήματα Σεπτέμβριος 2020 - Απρίλιος 2021, Οκτώβριος 2021 - Ιανουάριος 2022, υψηλό αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Ιούνιος 2022, υψηλό ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Μάϊος 2021 και στο χρονικό διάστημα Νοέμβριος 2021 - Μάρτιος 2022, υψηλή μετάδοση του ιού τον Οκτώβριο 2020 και στο χρονικό διάστημα Οκτώβριος 2021 - Νοέμβριος 2021. Επιπροσθέτως παρατηρούμε ότι στην πρώτη ομάδα υπάρχουν χώρες που λαμβάνουν πολύ χαμηλές τιμές στην δεύτερη κύρια συνιστώσα και χώρες που λαμβάνουν μέτριες τιμές στην δεύτερη κύρια συνιστώσα. Πιο συγκεκριμένα, παρατηρούμε ότι η Τσεχία, η Σλοβενία, η Σλοβακία, η Κροατία, η Λιθουανία, η Λετονία, η Εσθονία, η Δανία, η Ισλανδία λαμβάνουν πολύ χαμηλές τιμές στην δεύτερη κύρια συνιστώσα. Αυτό σημαίνει πως σε αυτές τις χώρες η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 - Οκτώβριος 2020, Ιούνιος 2021 - Σεπτέμβριος 2021 ήταν πολύ χαμηλή. Αντιθέτως, βλέπουμε ότι οι Η.Π.Α και η Αργεντινή λαμβάνουν μέτριες τιμές στην δεύτερη κύρια συνιστώσα. Αυτό σημαίνει πως σε αυτές τις χώρες η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 - Οκτώβριος 2020, Ιούνιος 2021 - Σεπτέμβριος 2021 ήταν μέτρια.

Τέλος, από το Σχήμα 3.19, βλέπουμε ότι οι περισσότερες χώρες της δεύτερης ομάδας (πράσινο χρώμα) λαμβάνουν υψηλές τιμές στην πρώτη κύρια συνιστώσα το οποίο σημαίνει ότι είχαν καλή επιδημιολογική εικόνα. Δηλαδή οι περισσότερες χώρες της δεύτερης ομάδας εμφάνισαν χαμηλό συνολικό αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Ιούνιος 2022, χαμηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στα χρονικά διαστήματα Σεπτέμβριος 2020 - Απρίλιος 2021, Οκτώβριος 2021 - Ιανουάριος 2022, χαμηλό συνολικό αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Ιούνιος 2022, χαμηλό ημερήσιο αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Μάϊος 2021 και στο χρονικό διάστημα Νοέμβριος 2021 - Μάρτιος 2022, χαμηλή μετάδοση του ιού τον Οκτώβριο 2020 και στο χρονικό διάστημα Οκτώβριος 2021 - Νοέμβριος 2021. Από την άλλη μεριά, η Γεωργία και το Περού λαμβάνουν πολύ χαμηλές τιμές στην πρώτη κύρια συνιστώσα το οποίο σημαίνει ότι είχαν κακή επιδημιολογική εικόνα. Δηλαδή η Γεωργία και το Περού, εμφάνισαν υψηλό συνολικό αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Ιούνιος 2022, υψηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στα χρονικά διαστήματα Σεπτέμβριος 2020 - Απρίλιος 2021, Οκτώβριος 2021 - Ιανουάριος 2022, υψηλό συνολικό αριθμό θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Ιούνιος 2022, υψηλό αριθμό ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Μάϊος 2021 και στο χρονικό διάστημα Νοέμβριος 2021 - Μάρτιος 2022, υψηλή μετάδοση του ιού τον Οκτώβριο 2020 και στο χρονικό διάστημα Οκτώβριος 2021 - Νοέμβριος 2021. Επιπλέον, παρατηρούμε ότι οι περισσότερες χώρες

της δεύτερης ομάδας λαμβάνουν μέτριες τιμές στην δεύτερη κύρια συνιστώσα το οποίο σημαίνει ότι στις περισσότερες χώρες της δεύτερης ομάδας η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 - Οκτώβριος 2020, Ιούνιος 2021 - Σεπτέμβριος 2021 ήταν μέτρια. Εξαίρεση αποτελεί το Περού το οποίο όπως βλέπουμε λαμβάνει πολύ υψηλή τιμή στην δεύτερη κύρια συνιστώσα το οποίο σημαίνει πως η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 - Οκτώβριος 2020, Ιούνιος 2021 - Σεπτέμβριος 2021 ήταν πολύ υψηλή.

Από το Σχήμα 3.19 βλέπουμε ότι οι χώρες της τρίτης ομάδας (μπλε χρώμα) παίρνουν υψηλές τιμές στην πρώτη κύρια συνιστώσα. Αυτό σημαίνει ότι οι χώρες της τρίτης ομάδας είχαν πολύ καλή επιδημιολογική εικόνα. Δηλαδή οι χώρες της τρίτης ομάδας εμφάνισαν χαμηλό συνολικό αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Ιούνιος 2022, χαμηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στα χρονικά διαστήματα Σεπτέμβριος 2020 - Απρίλιος 2021, Οκτώβριος 2021 - Ιανουάριος 2022, χαμηλό αριθμό συνολικών θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Ιούνιος 2022, χαμηλό αριθμό ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Οκτώβριος 2020 - Μάϊος 2021 και στο χρονικό διάστημα Νοέμβριος 2021 – Μάρτιος 2022, χαμηλή μετάδοση του ιού τον Οκτώβριο 2020 και στο χρονικό διάστημα Οκτώβριος 2021 - Νοέμβριος 2021. Επιπλέον παρατηρούμε ότι υπάρχουν χώρες της τρίτης ομάδας οι οποίες παίρνουν πολύ χαμηλές τιμές στην δεύτερη κύρια συνιστώσα και χώρες της τρίτης ομάδας που παίρνουν μέτριες τιμές στην δεύτερη κύρια συνιστώσα. Πιο συγκεκριμένα, από το Σχήμα 3.19 μπορούμε να δούμε ότι η Τανζανία, η Μπουρκίνα Φάσο, η Νιγηρία, Κεντροαφρικανική Δημοκρατία λαμβάνουν πολύ χαμηλές τιμές στην δεύτερη κύρια συνιστώσα. Αυτό σημαίνει ότι στις προαναφερθείσες χώρες η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 - Οκτώβριος 2020, Ιούνιος 2021 - Σεπτέμβριος 2021 ήταν πολύ χαμηλή. Από την άλλη μεριά, βλέπουμε ότι η Νότια Αφρική, η Cape Verde (Πράσινο Ακρωτήρι), η Μιανμαρ (Βιρμανία) λαμβάνουν μέτριες τιμές στην δεύτερη κύρια συνιστώσα. Αυτό σημαίνει ότι στις προαναφερθείσες χώρες η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Μάϊος 2020 - Οκτώβριος 2020, Ιούνιος 2021 - Σεπτέμβριος 2021 ήταν μέτρια.

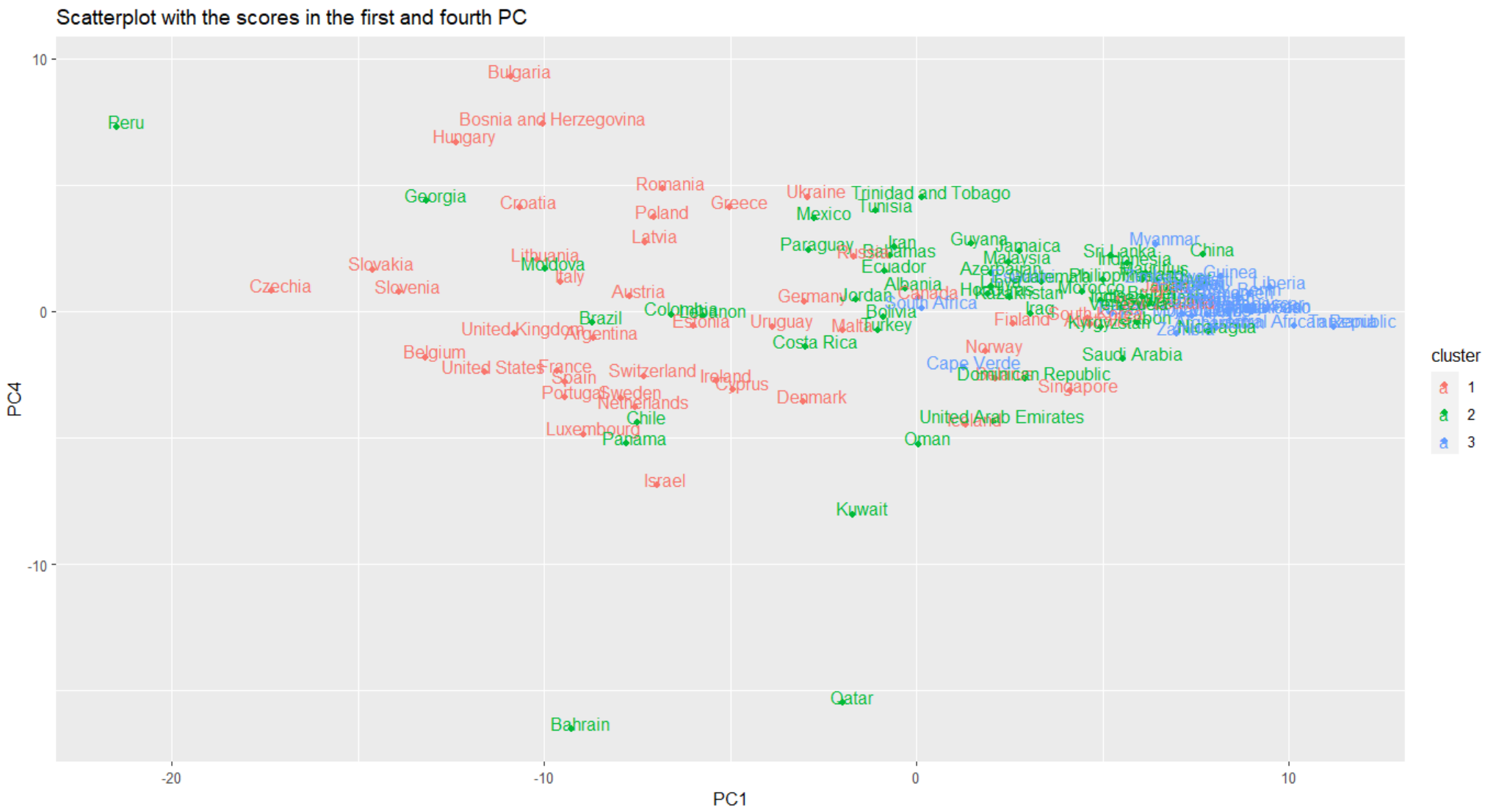


Σχήμα 3.20: Scatter plot (Διάγραμμα διασποράς) με τα scores των χωρών στην πρώτη και στην τρίτη κύρια συνιστώσα

Από το Σχήμα 3.20 παρατηρούμε ότι οι περισσότερες χώρες της πρώτης ομάδας λαμβάνουν χαμηλές τιμές στην τρίτη κύρια συνιστώσα. Αυτό σημαίνει ότι σε αυτές τις χώρες η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022 ήταν υψηλή. Αξίζει να σημειωθεί ότι από το Σχήμα 3.20 βλέπουμε ότι η Ελλάδα και η Κύπρος είχαν την πιο υψηλή σφοδρότητα των κυβερνητικών μέτρων από όλες τις χώρες που συμμετείχαν στην ανάλυση στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022.

Από το Σχήμα 3.20 παρατηρούμε ότι οι περισσότερες χώρες της δεύτερης ομάδας λαμβάνουν σχετικά χαμηλές τιμές στην τρίτη κύρια συνιστώσα. Αυτό σημαίνει ότι στις περισσότερες χώρες της δεύτερης ομάδας η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022 ήταν σχετικά υψηλή. Εξαίρεση σε αυτό αποτελεί το Περού που όπως βλέπουμε στο Σχήμα 3.20 λαμβάνει την πιο υψηλή τιμή στην τρίτη κύρια συνιστώσα από όλες τις χώρες. Αυτό σημαίνει ότι στο Περού η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022 ήταν πολύ χαμηλή.

Από το Σχήμα 3.20 παρατηρούμε ότι υπάρχουν χώρες της τρίτης ομάδας που λαμβάνουν υψηλές τιμές στην τρίτη κύρια συνιστώσα και χώρες της τρίτης ομάδας που λαμβάνουν χαμηλές τιμές στην τρίτη κύρια συνιστώσα. Πιο συγκεκριμένα, βλέπουμε ότι η Τανζανία, η Κεντροαφρικανική Δημοκρατία, η Μπουρκίνα Φάσο, το Καμερούν λαμβάνουν υψηλές τιμές στην τρίτη κύρια συνιστώσα το οποίο σημαίνει ότι σε αυτές τις χώρες η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022 ήταν χαμηλή. Από την άλλη μεριά, βλέπουμε ότι η Μιανμάρ (Βιρμανία) λαμβάνει χαμηλή τιμή στην τρίτη κύρια συνιστώσα το οποίο σημαίνει ότι σε αυτήν τη χώρα η σφοδρότητα των κυβερνητικών μέτρων στα χρονικά διαστήματα Δεκέμβριος 2020 – Ιούνιος 2021 και Οκτώβριος 2021 – Φεβρουάριος 2022 ήταν υψηλή.



Σχήμα 3.21: Scatter plot (Διάγραμμα διασποράς) με τα scores των χωρών στην πρώτη και στην τέταρτη κύρια συνιστώσα

Από το Σχήμα 3.21 βλέπουμε ότι σχεδόν όλες οι χώρες της πρώτης ομάδας λαμβάνουν υψηλές τιμές στην τέταρτη κύρια συνιστώσα το οποίο σημαίνει ότι σχεδόν όλες οι χώρες της πρώτης ομάδας εμφάνισαν χαμηλά συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Οκτώμβριος 2020 και χαμηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Ιούλιος 2020.

Από το Σχήμα 3.21 βλέπουμε ότι σχεδόν όλες οι χώρες της δεύτερης ομάδας λαμβάνουν υψηλές τιμές στην τέταρτη κύρια συνιστώσα το οποίο σημαίνει ότι σχεδόν όλες οι χώρες της δεύτερης ομάδας εμφάνισαν χαμηλά συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Οκτώβριος 2020 και χαμηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Ιούλιος 2020. Εξαίρεση στα παραπάνω αποτελούν το Μπαχρέιν και το Κατάρ όπου όπως βλέπουμε από το Σχήμα 3.21 λαμβάνουν πολύ χαμηλές τιμές στην τέταρτη κύρια συνιστώσα. Αυτό σημαίνει ότι οι προαναφερθείσες χώρες εμφάνισαν υψηλά συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 - Οκτώβριος 2020 και υψηλό αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020– Ιούλιος 2020.

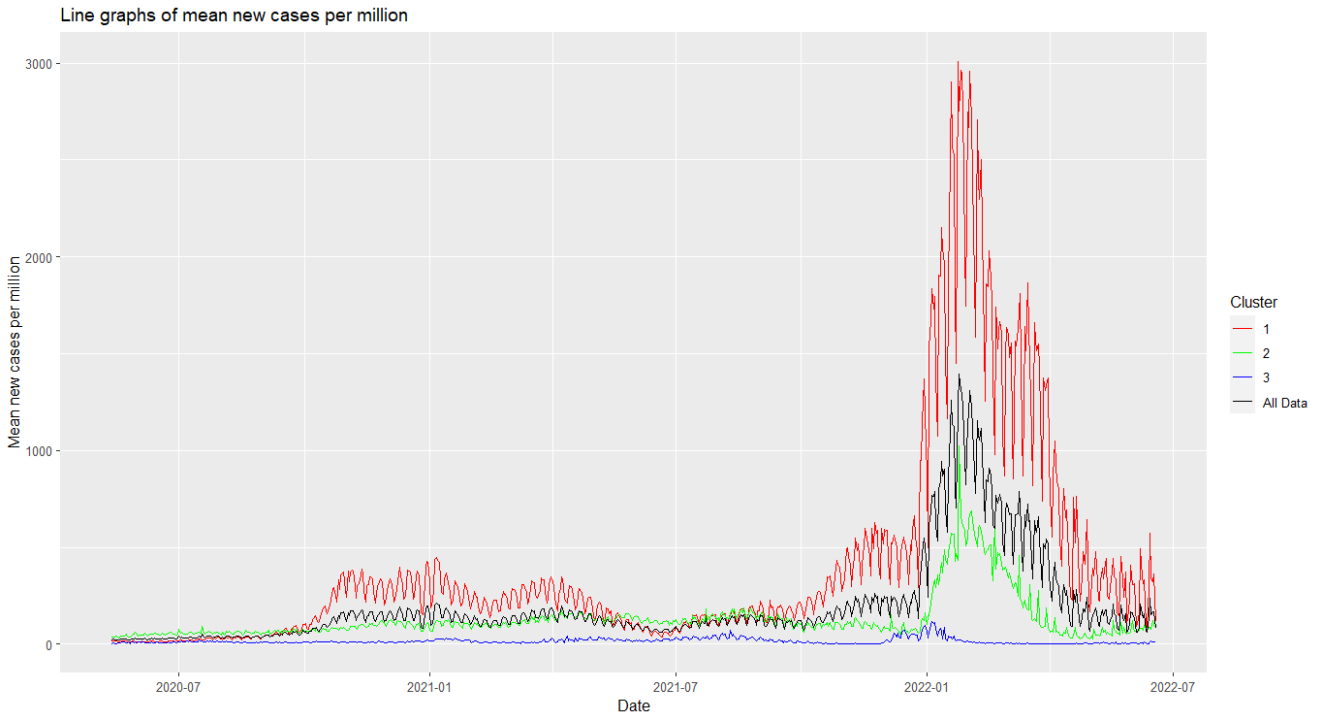
Στο Σχήμα 3.21 βλέπουμε ότι οι χώρες της τρίτης ομάδας είναι πολύ συγκεντρωμένες στο σημείο (7.5 , 0). Επιπλέον, παρατηρούμε ότι οι χώρες της τρίτης ομάδας λαμβάνουν αρκετά υψηλές τιμές στην τέταρτη κύρια συνιστώσα το οποίο σημαίνει ότι οι χώρες της τρίτης ομάδας εμφάνισαν αρκετά χαμηλά συνολικά κρούσματα ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Οκτώμβριος 2020 και αρκετά χαμηλό ημερήσιο αριθμό κρουσμάτων ανά εκατομμύριο πληθυσμού στο χρονικό διάστημα Μάϊος 2020 – Ιούλιος 2020.

Κεφάλαιο 4

4. ΣΥΝΟΛΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

Σε αυτό το κεφάλαιο θα δούμε τα ευρήματα που παρουσιάστηκαν στην προηγούμενη ενότητα ανά μεταβλητή.

- Μεταβλητή new_cases_per_million

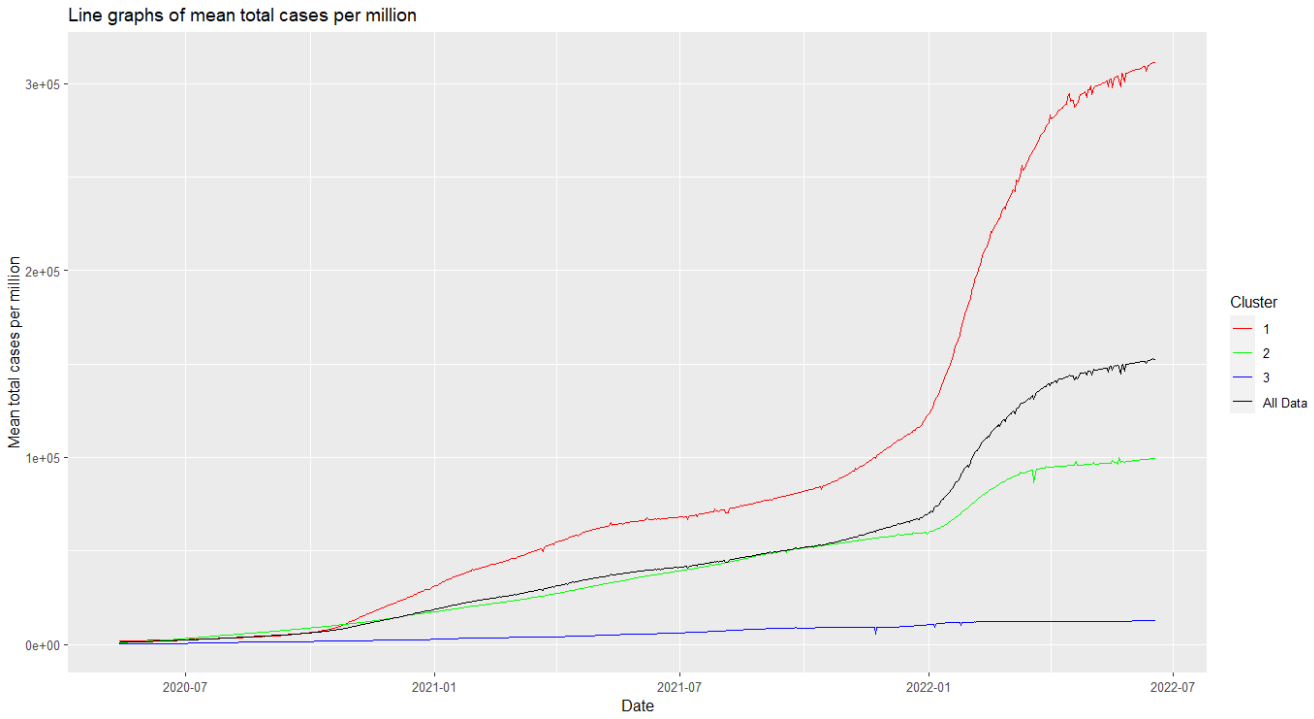


Σχήμα 4.1: Εξέλιξη μέσου αριθμού ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού σε όλα τα δεδομένα και ανά ομάδα.

Από το Σχήμα 4.1 βλέπουμε ότι ο μέσος αριθμός ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες που μελετήσαμε (μαύρο χρώμα) παρέμεινε σχετικά σταθερός έως και τον Ιανουάριο του 2022. Από τον Ιανουάριο του 2022 και μετά βλέπουμε ότι εμφανίστηκε μια απότομη αύξηση στον μέσο αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού. Επιπλέον, βλέπουμε ότι ο μέσος αριθμός ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της πρώτης ομάδας (κόκκινο χρώμα) παρέμεινε σχετικά σταθερός έως και τον Οκτώβριο του 2021. Από τον Οκτώβριο του 2021 και μετά, βλέπουμε ότι εμφανίστηκε μια πολύ απότομη αύξηση στον μέσο αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της πρώτης ομάδας. Επιπροσθέτως, βλέπουμε ότι ο μέσος αριθμός ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της δεύτερης ομάδας (πράσινο χρώμα) παρέμεινε σχετικά σταθερός έως και τον Ιανουάριο του 2022. Από τον Ιανουάριο του 2022 και μετά, βλέπουμε ότι εμφανίστηκε μια απότομη αύξηση στον μέσο αριθμό ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της δεύτερης ομάδας η οποία ήταν αισθητά πιο χαμηλή από ότι αυτή που εμφανίστηκε στην πρώτη ομάδα. Από την άλλη μεριά, βλέπουμε ότι ο μέσος αριθμός ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της τρίτης ομάδας ήταν αισθητά πιο χαμηλός από ότι ο μέσος

αριθμός ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της πρώτης και της δεύτερης ομάδας. Ακόμη, βλέπουμε ότι ο μέσος αριθμός ημερήσιων κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες της τρίτης ομάδας δεν εμφάνισε κάποια σημαντική αύξηση καθ' όλη την διάρκεια του χρονικού διαστήματος που μελετήσαμε.

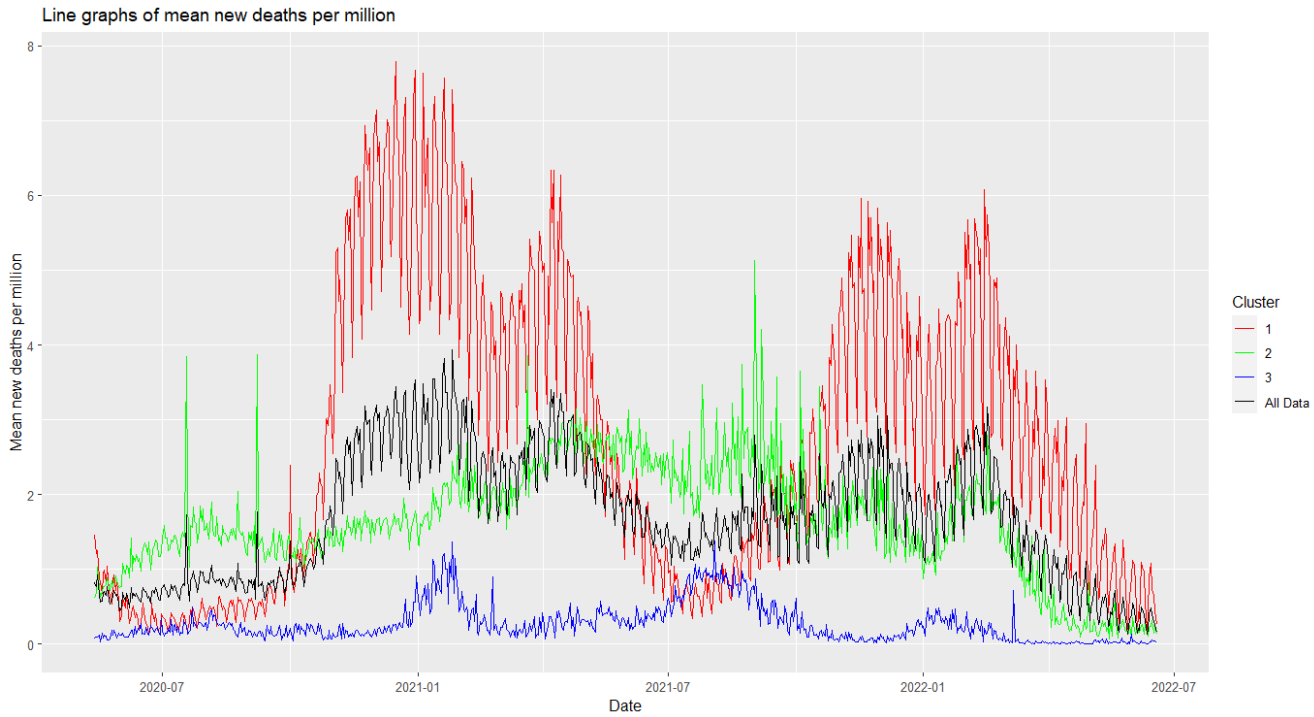
- Μεταβλητή total_cases_per_million



Σχήμα 4.2: Εξέλιξη μέσου αριθμού συνολικών κρουσμάτων ανά εκατομμύριο πληθυσμού σε όλα τα δεδομένα και ανά ομάδα.

Από το Σχήμα 4.2 βλέπουμε ότι ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες που μελετήσαμε (μαύρο χρώμα) αυξανόταν με χαμηλό ρυθμό έως και τον Ιανουάριο του 2022. Από τον Ιανουάριο του 2022 και μετά βλέπουμε ότι ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στις χώρες που μελετήσαμε αυξανόταν με μεγαλύτερο ρυθμό από ότι αυξανόταν πριν. Επιπλέον, βλέπουμε ότι ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στην πρώτη ομάδα (κόκκινο χρώμα) ήταν αισθητά πιο υψηλός από ότι στην δεύτερη και στην τρίτη ομάδα από τον Ιανουάριο του 2021 και μετά. Επιπροσθέτως, βλέπουμε ότι από τον Ιανουάριο του 2022 και μετά ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στην πρώτη ομάδα εμφάνισε ραγδαία αύξηση. Ακόμη, βλέπουμε ότι ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στην δεύτερη ομάδα (πράσινο χρώμα) αυξανόταν με χαμηλό ρυθμό έως και τον Ιανουάριο του 2022. Από τον Ιανουάριο του 2022 έως και τον Απρίλιο του 2022, βλέπουμε ότι ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στην δεύτερη ομάδα αυξανόταν με μεγαλύτερο ρυθμό από ότι αυξανόταν πιο πριν. Από την άλλη μεριά, βλέπουμε ότι ο μέσος συνολικός αριθμός κρουσμάτων ανά εκατομμύριο πληθυσμού στην τρίτη ομάδα (μπλε χρώμα) ήταν αισθητά πιο χαμηλός από ότι στις υπόλοιπες ομάδες καθ' όλη την διάρκεια του χρονικού διαστήματος που μελετήσαμε και ότι η αύξηση που εμφάνισε ήταν ανεπαίσθητη.

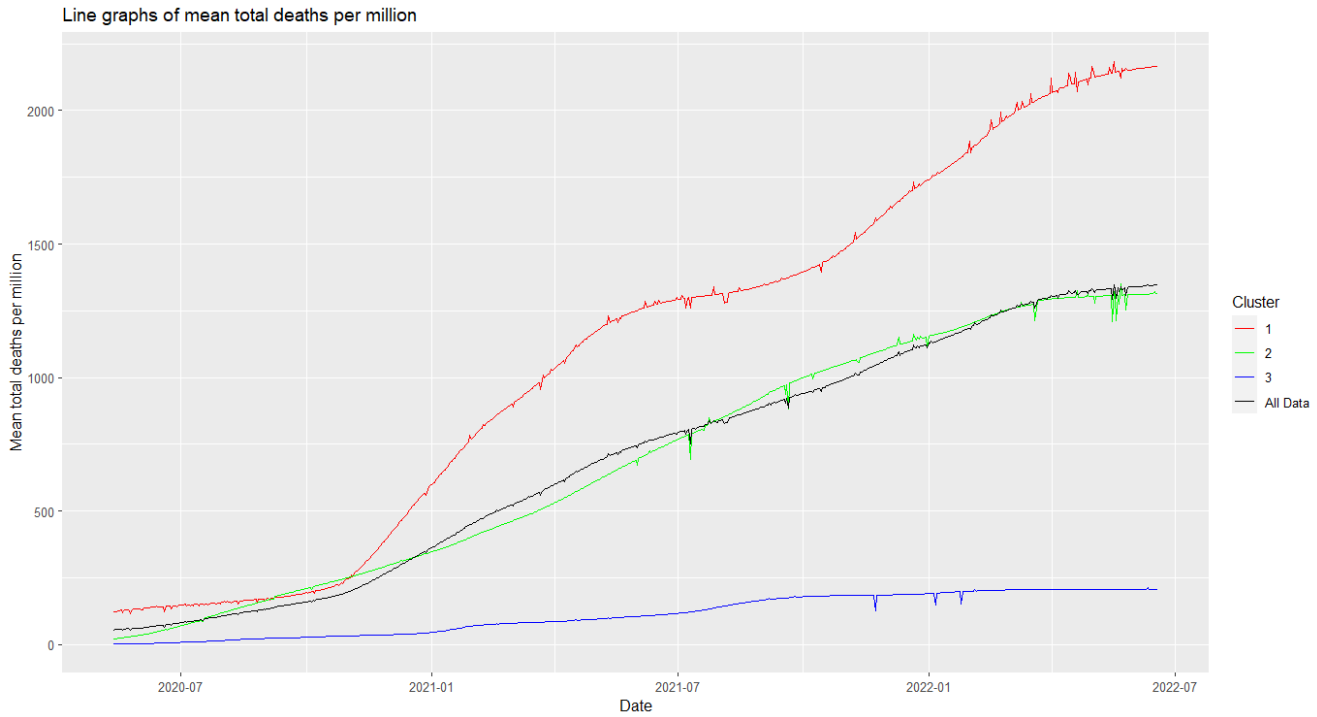
- Μεταβλητή new_deaths_per_million



Σχήμα 4.3: Εξέλιξη μέσου αριθμού ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού σε όλα τα δεδομένα και ανά ομάδα.

Από το Σχήμα 4.3 βλέπουμε ότι ο μέσος αριθμός ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στις χώρες που μελετήσαμε (μαύρο χρώμα) εμφάνισε κάποιες απότομες αυξήσεις. Πιο συγκεκριμένα, βλέπουμε ότι τον Ιανουάριο του 2021 και στο χρονικό διάστημα Οκτώβριος 2021 - Ιανουάριος 2022 ο μέσος αριθμός ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στις χώρες που μελετήσαμε εμφάνισε σημαντική αύξηση. Επιπλέον, βλέπουμε ότι ο μέσος αριθμός ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στις χώρες της πρώτης ομάδας (κόκκινο χρώμα) εμφάνισε πολύ απότομες αυξήσεις τον Ιανουάριο 2021 και τον Ιανουάριο 2022. Επιπροσθέτως, βλέπουμε ότι ο μέσος αριθμός ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στις χώρες της δεύτερης ομάδας (πράσινο χρώμα) αυξανόταν με ομαλό ρυθμό. Ακόμη, βλέπουμε ότι οι αυξήσεις που εμφανίστηκαν στον μέσο αριθμό ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στις χώρες της πρώτης ομάδας ήταν πολύ πιο απότομες από ότι οι αυξήσεις που εμφανίστηκαν στον προαναφερθέντα αριθμό στην δεύτερη ομάδα. Από την άλλη μεριά, βλέπουμε ότι ο μέσος αριθμός ημερήσιων θανάτων ανά εκατομμύριο πληθυσμού στις χώρες της τρίτης ομάδας δεν εμφάνισε κάποια ιδιαίτερη αύξηση καθ' όλη την διάρκεια του χρονικού διαστήματος που μελετήσαμε.

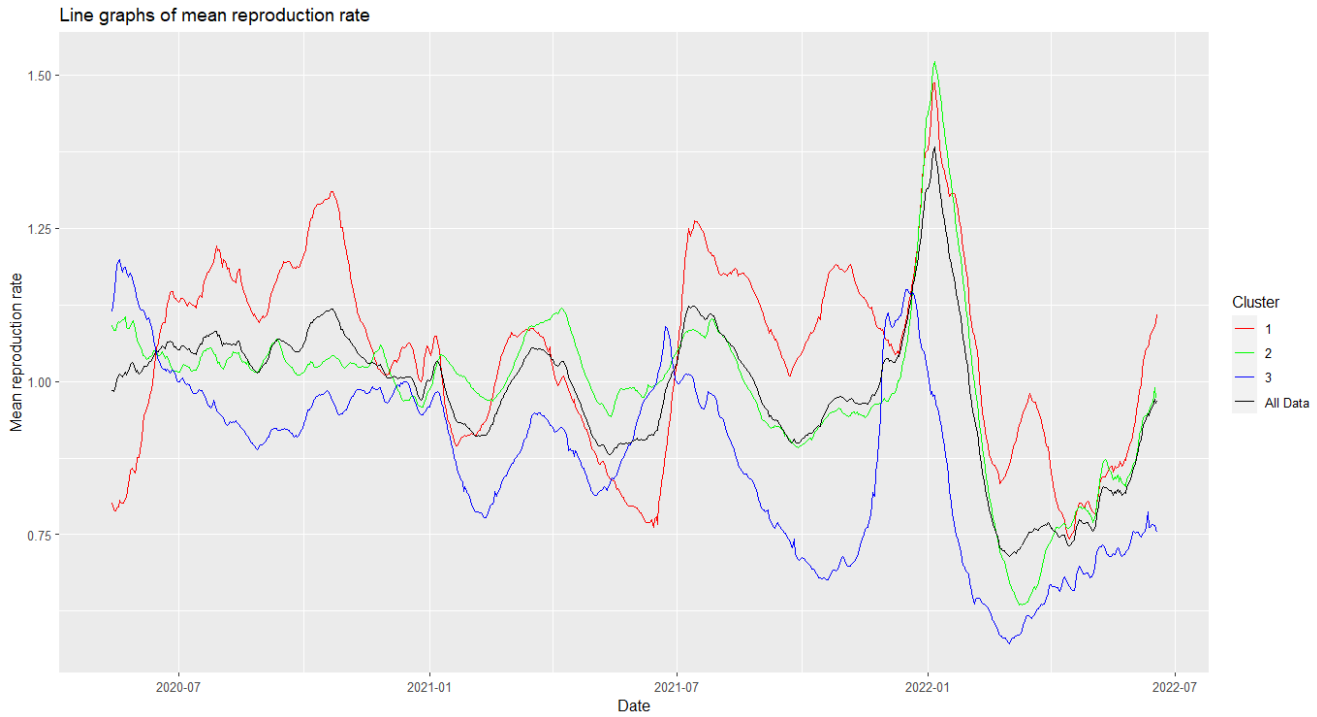
- Μεταβλητή total_deaths_per_million



Σχήμα 4.4: Εξέλιξη μέσου αριθμού συνολικών θανάτων ανά εκατομμύριο πληθυσμού σε όλα τα δεδομένα και ανά ομάδα.

Από το Σχήμα 4.4 βλέπουμε ότι ο μέσος συνολικός αριθμός θανάτων ανά εκατομμύριο πληθυσμού στις χώρες που μελετήσαμε (μαύρο χρώμα) αυξανόταν σταθερά και με γραμμικό τρόπο. Επιπλέον, βλέπουμε ότι ο μέσος συνολικός αριθμός θανάτων ανά εκατομμύριο πληθυσμούς στις χώρες της πρώτης ομάδας (κόκκινο χρώμα) αυξανόταν με πολύ μεγάλο ρυθμό από τον Οκτώβριο 2020 και μετά. Επιπροσθέτως, βλέπουμε ότι ο μέσος συνολικός αριθμός θανάτων ανά εκατομμύριο πληθυσμού στις χώρες της δεύτερης ομάδας (πράσινο χρώμα) αυξανόταν σταθερά και με γραμμικό τρόπο καθ' όλη την διάρκεια του χρονικού διαστήματος που μελετήσαμε. Από την άλλη μεριά, βλέπουμε ότι ο μέσος συνολικός αριθμός θανάτων ανά εκατομμύριο πληθυσμού στις χώρες της τρίτης ομάδας (μπλε χρώμα) αυξανόταν με πολύ μικρό ρυθμό.

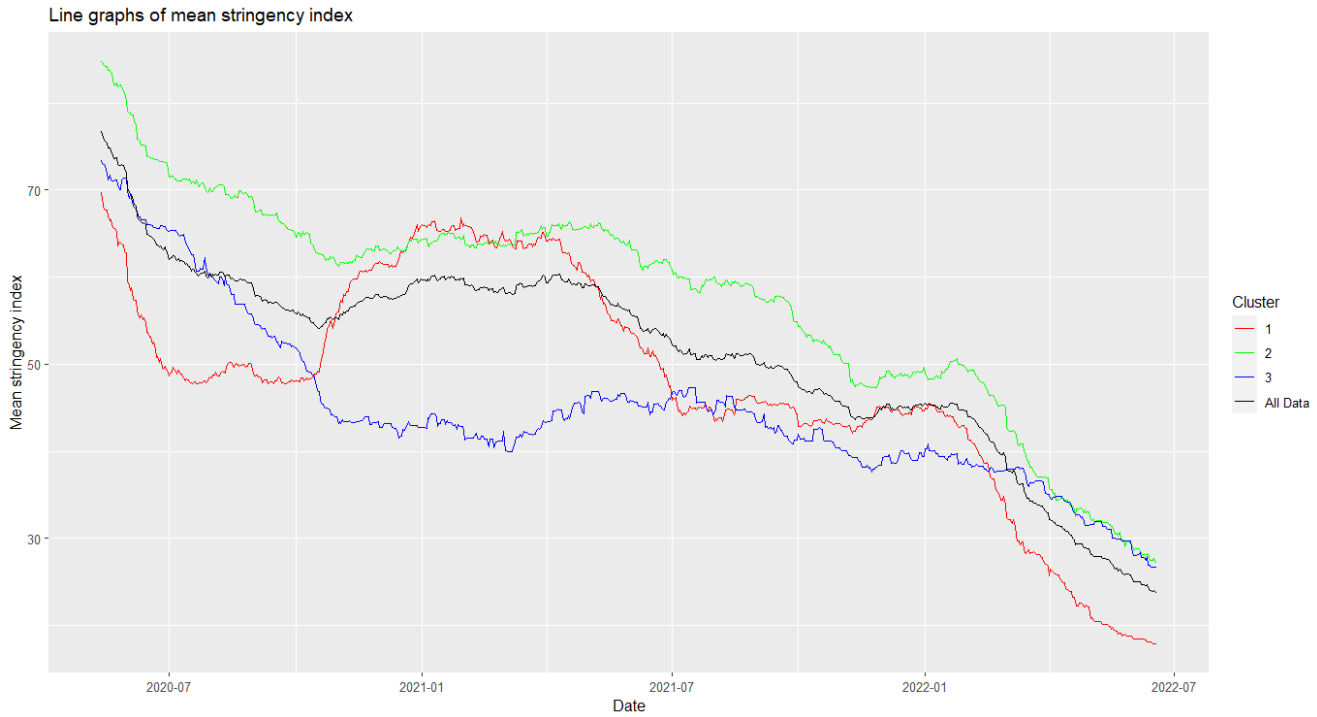
- Μεταβλητή reproduction_rate



Σχήμα 4.5: Εξέλιξη μέσης τιμής δείκτη R σε όλα τα δεδομένα και ανά ομάδα.

Από το Σχήμα 4.5 βλέπουμε ότι η μέση τιμή του δείκτη R στις χώρες που μελετήσαμε (μαύρο χρώμα) εμφάνισε απότομη αύξηση τον Ιανουάριο του 2022 με αποτέλεσμα η μέση του δείκτη R να ξεπεράσει την μονάδα. Αυτό σημαίνει ότι η μέση μετάδοση του ιού στις χώρες που μελετήσαμε τον Ιανουάριο του 2022 ήταν υψηλή. Επιπλέον, η μέση τιμή του δείκτη R στις χώρες της πρώτης ομάδας (κόκκινο χρώμα) και της δεύτερης ομάδας (πράσινο χρώμα) εμφάνισε απότομη αύξηση τον Ιανουάριο του 2022 με αποτέλεσμα να ξεπεράσει την μονάδα. Αυτό σημαίνει ότι η μέση μετάδοση του ιού στις χώρες της πρώτης και της δεύτερης ομάδας τον Ιανουάριο του 2022 ήταν υψηλή. Από την άλλη μεριά, βλέπουμε ότι η μέση τιμή του δείκτη R στις χώρες της τρίτης ομάδας (μπλε χρώμα) ήταν αισθητά πιο μικρή από ότι η μέση τιμή του δείκτη R στην πρώτη και στην δεύτερη ομάδα. Η μέση τιμή του δείκτη R στις χώρες της τρίτης ομάδας εμφάνισε απότομη αύξηση λίγο πριν τον Ιανουάριο του 2022 με αποτέλεσμα η μέση τιμή του δείκτη R να ξεπεράσει την μονάδα. Αυτό σημαίνει ότι η μέση μετάδοση του ιου στις χώρες της τρίτης ομάδας λίγο πριν τον Ιανουάριο του 2022 ήταν υψηλή. Ακόμη, από το Σχήμα 4.5 μπορούμε να δούμε ότι η μέση τιμή του δείκτη R στην τρίτη ομάδα λίγο πριν τον Ιανουάριο του 2022 ήταν πολύ πιο μικρή από ότι η μέση τιμή του δείκτη R στην πρώτη και στην δεύτερη ομάδα την αντίστοιχη χρονική στιγμή. Αυτό σημαίνει ότι λίγο πριν τον Ιανουάριο του 2022 η μέση μετάδοση του ιού στις χώρες της τρίτης ομάδας ήταν πολύ μικρότερη από ότι η μέση μετάδοση του ιού στις χώρες της πρώτης και δεύτερης ομάδας την αντίστοιχη χρονική στιγμή.

- Μεταβλητή stringency_index



Σχήμα 4.6: Εξέλιξη μέσης τιμής της σφοδρότητας των κυβερνητικών μέτρων που είχαν ως στόχο την καταπολέμηση της πανδημίας Covid19 σε όλα τα δεδομένα και ανά ομάδα.

Από το Σχήμα 4.6 βλέπουμε ότι η μέση τιμή της σφοδρότητας των κυβερνητικών μέτρων που είχαν ως στόχο την καταπολέμηση της πανδημίας Covid19 στις χώρες που μελετήσαμε (μαύρο χρώμα) στην αρχή της υπό μελέτης χρονικής περιόδου ήταν υψηλή. Επιπλέον, βλέπουμε ότι αυτή η μέση τιμή με την πάροδο του χρόνου μειωνόταν. Επιπροσθέτως, βλέπουμε ότι η μέση τιμή της σφοδρότητας των κυβερνητικών μέτρων που είχαν ως στόχο την καταπολέμηση της πανδημίας Covid19 στις χώρες τις πρώτης ομάδας (κόκκινο χρώμα), στις χώρες της δεύτερης ομάδας (πράσινο χρώμα), στις χώρες της τρίτης ομάδας (μπλε χρώμα) στην αρχή της υπό μελέτης χρονικής περιόδου ήταν υψηλή. Ακόμη, βλέπουμε ότι αυτές οι 3 μέσες τιμές με την πάροδο του χρόνου μειώνονταν. Εξαίρεση αποτελεί το χρονικό διάστημα Οκτώβριος 2020 - Ιούλιος 2021 στο οποίο όπως βλέπουμε υπήρξε μια αύξηση στην μέση τιμή της σφοδρότητας των κυβερνητικών μέτρων που είχαν ως στόχο την καταπολέμηση της πανδημίας Covid19 στις χώρες της πρώτης ομάδας. Επίσης, βλέπουμε ότι η μέση τιμή της σφοδρότητας των κυβερνητικών μέτρων που είχαν ως στόχο την καταπολέμηση της πανδημίας Covid19 στις χώρες της δεύτερης ομάδας ήταν για μεγάλο χρονικό διάστημα υψηλότερη από τις αντίστοιχες μέσες τιμές της πρώτης και της δεύτερης ομάδας. Αυτό σημαίνει ότι οι χώρες της δεύτερης ομάδας για μεγάλο χρονικό διάστημα είχαν υψηλότερη μέση σφοδρότητα των κυβερνητικών μέτρων που είχαν ως στόχο την καταπολέμηση της πανδημίας Covid19 από ότι οι χώρες της πρώτης και της τρίτης ομάδας.

Παράρτημα

```
library(readr)
library(dplyr)
library(randomcoloR)
library(ggplot2)
data1 = read_csv("https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-
covid-data.csv")
data1 = as.data.frame(data1)
data1$date = as.character(data1$date)
list_of_countries = unique(data1$location)
starting_date = rep(0,length(list_of_countries))
ending_date = rep(0,length(list_of_countries))
for (i in 1:length(list_of_countries)){
  k = data1$date[data1$location == list_of_countries[i]];
  starting_date[i] = k[1]
  ending_date[i] = k[length(k)]}
cbind(unique(data1$location),starting_date,ending_date)
data1 = subset(data1, data1$location!="Western Sahara")
starting_date = starting_date[-240]
ending_date = ending_date[-240]
list_of_countries = list_of_countries[list_of_countries != "Western Sahara"]
countries_2021 = list()
df = data.frame("location" = unique(data1$location),"start_date" = starting_date
                ,"end_date" = ending_date)
for (i in 1:length(list_of_countries)) {
  if ((grepl("2021",df$start_date[i],fixed=TRUE) == TRUE) |
      (grepl("2021",df$end_date[i],fixed=TRUE) == TRUE)){
    countries_2021 = append(countries_2021,df$location[i])}}
for (country in countries_2021) {
  data1 = subset(data1, data1$location != country)}
for (country in countries_2021) {
  list_of_countries = list_of_countries[list_of_countries != country]}
list_of_countries
drop_countries = c("Africa","Asia","Europe","European Union","International",
                   "North America","Oceania","South America","High income",
                   "Low income","Lower middle income","Upper middle income",
                   "World")
drop_countries = c(drop_countries,"Saint Helena","Samoa","Solomon Islands","Vanuatu"
,"Wallis and Futuna")
for (country in drop_countries) {
  data1 = subset(data1, data1$location != country)}
for (country in drop_countries) {
  list_of_countries = list_of_countries[list_of_countries != country]}
starting_date = rep(0,length(list_of_countries))
ending_date = rep(0,length(list_of_countries))
for (i in 1:length(list_of_countries)){
```

```
  k = data1$date[data1$location == list_of_countries[i]];
  starting_date[i] = k[1]
  ending_date[i] = k[length(k)]}
cbind(unique(data1$location),starting_date,ending_date)
max_starting_date = max(as.Date(starting_date))
min_ending_date = min(as.Date(ending_date))
starting_date = as.data.frame(starting_date)
rownames(starting_date) = list_of_countries
for (country in list_of_countries) {
  if(starting_date[country,]!= max_starting_date){
    count1 = 1
    while (data1$location[count1]!=country) {
      count1 = count1 + 1}
    count2 = count1
    while (data1$date[count2]!=max_starting_date) {
      count2 = count2 + 1}
    count2 = count2 - 1
    rows_to_delete = count1:count2
    data1 = data1[-rows_to_delete,]}}
i = 1
while (i <= length(list_of_countries)){
  country = list_of_countries[i]
  data_c = data1[data1$location == country,]
  if(data_c$date[length(data_c$date)] != min_ending_date){
    start1 = 1
    while ((data1$location[start1]!=country) == TRUE) {
      start1 = start1 + 1}
    data_c = data1[data1$location == country,]
    ind = 1
    steps = 0
    while (as.Date(data_c$date[ind])<=as.Date(min_ending_date)) {
      steps = steps + 1
      ind = ind + 1}
    start1 = start1 + steps
    finish1  = start1
    while((data1$location[finish1] == country) & (is.na(data1$location[finish1]) == FALSE)){
      finish1 = finish1 + 1}
    finish1 = finish1 - 1
    rows_to_delete = start1:finish1
    data1 = data1[-rows_to_delete,]}
  i = i + 1}
starting_date = rep(0,length(list_of_countries))
ending_date = rep(0,length(list_of_countries))
for (i in 1:length(list_of_countries)){
  k = data1$date[data1$location == list_of_countries[i]];
  starting_date[i] = k[1]
  ending_date[i] = k[length(k)]}
cbind(unique(data1$location),starting_date,ending_date)
sort((colMeans(is.na(data1)))*100)
final_dataset = data1[,c("iso_code","continent","location","date","population"
```

```
                    ,"life_expectancy","total_cases_per_million","new_cases_per_million",
                    "population_density","diabetes_prevalence","total_deaths_per_million",
                    "new_deaths_per_million","median_age","aged_70_older",
                    "cardiovasc_death_rate","gdp_per_capita",
                    "human_development_index","reproduction_rate",
                    "stringency_index","hospital_beds_per_thousand")]
final_dataset = na.omit(final_dataset)
drop_countries = c("Vietnam","Timor","Mongolia","Laos","Cambodia","Bhutan","Grenada",
"Belize","Barbados","Fiji","Uganda","Seychelles","Gambia","Burundi","Botswana")
for (country in drop_countries) {
  final_dataset = subset(final_dataset, final_dataset$location != country)}
for (country in drop_countries) {
  list_of_countries = list_of_countries[list_of_countries != country]}
list_of_countries = unique(final_dataset$location)
final_dataset = subset(final_dataset, final_dataset$location != "Suriname")
list_of_countries = list_of_countries[list_of_countries != "Suriname"]
dataset.clust = final_dataset
dataset.clust = dataset.clust[,c("continent","location","population","life_expectancy",
"population_density","diabetes_prevalence","median_age","aged_70_older","cardiovasc_death_rate",
"gdp_per_capita","human_development_index","hospital_beds_per_thousand")]
for (country in list_of_countries) {
  i = 1
  while (dataset.clust$location[i] != country) {
    i = i + 1}
  i = i + 1
  j = i
  while (dataset.clust$location[j] == country & (j <= nrow(dataset.clust))) {
    j = j + 1}
  j = j - 1
  dataset.clust = dataset.clust[-(i:j),]}
dataset.clust.num = dataset.clust[,-c(1,2)]
wardsmethod = hclust(dist(scale(dataset.clust.num,scale = TRUE)),method="ward.D")
plot(wardsmethod,labels = FALSE,main = "Ward Clustering",xlab="")
data.kmeans = kmeans(scale(dataset.clust.num,scale = TRUE),centers = 3,algorithm = "MacQueen")
dataset.clust$cluster = c(3,2,2,1,1,1,2,2,2,2,1,1,3,2,1,2,2,1,3,3,1,3,3,2,2,2,2,1,1,1,1,3,2,2,2,2,1,3,3,1,1,2,2,1,3,1,
2,3,2,3,2,1,1,2,2,2,2,1,1,1,2,1,2,2,3,2,2,1,2,3,2,1,1,3,3,2,3,1,2,2,2,2,3,3,3,1,1,2,3,1,2,3,2,2,2,2,1,1,2,1,1,2,1,1,
1,3,1,1,2,3,1,1,3,3,2,3,2,2,2,1,2,1,1,1,3,2,3,3,3)
d_1 = data.frame(location = unique(final_dataset$location))
d_1$May_2020_tot_cases_per_million = NA
d_1$Jun_2020_tot_cases_per_million = NA
d_1$Jul_2020_tot_cases_per_million = NA
d_1$Aug_2020_tot_cases_per_million = NA
d_1$Sep_2020_tot_cases_per_million = NA
d_1$Oct_2020_tot_cases_per_million = NA
d_1$Nov_2020_tot_cases_per_million = NA
d_1$Dec_2020_tot_cases_per_million = NA
d_1$Jan_2021_tot_cases_per_million = NA
d_1$Feb_2021_tot_cases_per_million = NA
d_1$Mar_2021_tot_cases_per_million = NA
d_1$Apr_2021_tot_cases_per_million = NA
```

```
d_1$May_2021_tot_cases_per_million = NA
d_1$Jun_2021_tot_cases_per_million = NA
d_1$Jul_2021_tot_cases_per_million = NA
d_1$Aug_2021_tot_cases_per_million = NA
d_1$Sep_2021_tot_cases_per_million = NA
d_1$Oct_2021_tot_cases_per_million = NA
d_1$Nov_2021_tot_cases_per_million = NA
d_1$Dec_2021_tot_cases_per_million = NA
d_1$Jan_2022_tot_cases_per_million = NA
d_1$Feb_2022_tot_cases_per_million = NA
d_1$Mar_2022_tot_cases_per_million = NA
d_1$Apr_2022_tot_cases_per_million = NA
d_1$May_2022_tot_cases_per_million = NA
d_1$Jun_2022_tot_cases_per_million = NA
for (i in 1:length(d_1$location)) {
  loc = d_1$location[i]
  data_init = subset(final_dataset, final_dataset$location == loc)
  data_sub = subset(data_init, grepl("2020-05",data_init$date) == TRUE)
  d_1$May_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-06",data_init$date) == TRUE)
  d_1$Jun_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-07",data_init$date) == TRUE)
  d_1$Jul_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-08",data_init$date) == TRUE)
  d_1$Aug_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-09",data_init$date) == TRUE)
  d_1$Sep_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-10",data_init$date) == TRUE)
  d_1$Oct_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-11",data_init$date) == TRUE)
  d_1$Nov_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2020-12",data_init$date) == TRUE)
  d_1$Dec_2020_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-01",data_init$date) == TRUE)
  d_1$Jan_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-02",data_init$date) == TRUE)
  d_1$Feb_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-03",data_init$date) == TRUE)
  d_1$Mar_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-04",data_init$date) == TRUE)
  d_1$Apr_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-05",data_init$date) == TRUE)
  d_1$May_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-06",data_init$date) == TRUE)
  d_1$Jun_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-07",data_init$date) == TRUE)
  d_1$Jul_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-08",data_init$date) == TRUE)
  d_1$Aug_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-09",data_init$date) == TRUE)
```

```
  d_1$Sep_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-10",data_init$date) == TRUE)
  d_1$Oct_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-11",data_init$date) == TRUE)
  d_1$Nov_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2021-12",data_init$date) == TRUE)
  d_1$Dec_2021_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2022-01",data_init$date) == TRUE)
  d_1$Jan_2022_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2022-02",data_init$date) == TRUE)
  d_1$Feb_2022_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2022-03",data_init$date) == TRUE)
  d_1$Mar_2022_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2022-04",data_init$date) == TRUE)
  d_1$Apr_2022_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2022-05",data_init$date) == TRUE)
  d_1$May_2022_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  data_sub = subset(data_init, grepl("2022-06",data_init$date) == TRUE)
  d_1$Jun_2022_tot_cases_per_million[i] = mean(data_sub$total_cases_per_million)}
d_1$May_2020_new_cases_per_million = NA
d_1$Jun_2020_new_cases_per_million = NA
d_1$Jul_2020_new_cases_per_million = NA
d_1$Aug_2020_new_cases_per_million = NA
d_1$Sep_2020_new_cases_per_million = NA
d_1$Oct_2020_new_cases_per_million = NA
d_1$Nov_2020_new_cases_per_million = NA
d_1$Dec_2020_new_cases_per_million = NA
d_1$Jan_2021_new_cases_per_million = NA
d_1$Feb_2021_new_cases_per_million = NA
d_1$Mar_2021_new_cases_per_million = NA
d_1$Apr_2021_new_cases_per_million = NA
d_1$May_2021_new_cases_per_million = NA
d_1$Jun_2021_new_cases_per_million = NA
d_1$Jul_2021_new_cases_per_million = NA
d_1$Aug_2021_new_cases_per_million = NA
d_1$Sep_2021_new_cases_per_million = NA
d_1$Oct_2021_new_cases_per_million = NA
d_1$Nov_2021_new_cases_per_million = NA
d_1$Dec_2021_new_cases_per_million = NA
d_1$Jan_2022_new_cases_per_million = NA
d_1$Feb_2022_new_cases_per_million = NA
d_1$Mar_2022_new_cases_per_million = NA
d_1$Apr_2022_new_cases_per_million = NA
d_1$May_2022_new_cases_per_million = NA
d_1$Jun_2022_new_cases_per_million = NA
for (i in 1:length(d_1$location)) {
  loc = d_1$location[i]
  data_init = subset(final_dataset, final_dataset$location == loc)
  data_sub = subset(data_init, grepl("2020-05",data_init$date) == TRUE)
  d_1$May_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
```

```
data_sub = subset(data_init, grepl("2020-06",data_init$date) == TRUE)
d_1$Jun_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2020-07",data_init$date) == TRUE)
d_1$Jul_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2020-08",data_init$date) == TRUE)
d_1$Aug_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2020-09",data_init$date) == TRUE)
d_1$Sep_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2020-10",data_init$date) == TRUE)
d_1$Oct_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2020-11",data_init$date) == TRUE)
d_1$Nov_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2020-12",data_init$date) == TRUE)
d_1$Dec_2020_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-01",data_init$date) == TRUE)
d_1$Jan_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-02",data_init$date) == TRUE)
d_1$Feb_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-03",data_init$date) == TRUE)
d_1$Mar_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-04",data_init$date) == TRUE)
d_1$Apr_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-05",data_init$date) == TRUE)
d_1$May_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-06",data_init$date) == TRUE)
d_1$Jun_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-07",data_init$date) == TRUE)
d_1$Jul_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-08",data_init$date) == TRUE)
d_1$Aug_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-09",data_init$date) == TRUE)
d_1$Sep_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-10",data_init$date) == TRUE)
d_1$Oct_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-11",data_init$date) == TRUE)
d_1$Nov_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2021-12",data_init$date) == TRUE)
d_1$Dec_2021_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2022-01",data_init$date) == TRUE)
d_1$Jan_2022_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2022-02",data_init$date) == TRUE)
d_1$Feb_2022_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2022-03",data_init$date) == TRUE)
d_1$Mar_2022_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2022-04",data_init$date) == TRUE)
d_1$Apr_2022_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2022-05",data_init$date) == TRUE)
d_1$May_2022_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
data_sub = subset(data_init, grepl("2022-06",data_init$date) == TRUE)
d_1$Jun_2022_new_cases_per_million[i] = mean(data_sub$new_cases_per_million)}
```

```
d_1$May_2020_total_deaths_per_million = NA
d_1$Jun_2020_total_deaths_per_million = NA
d_1$Jul_2020_total_deaths_per_million = NA
d_1$Aug_2020_total_deaths_per_million = NA
d_1$Sep_2020_total_deaths_per_million = NA
d_1$Oct_2020_total_deaths_per_million = NA
d_1$Nov_2020_total_deaths_per_million = NA
d_1$Dec_2020_total_deaths_per_million = NA
d_1$Jan_2021_total_deaths_per_million = NA
d_1$Feb_2021_total_deaths_per_million = NA
d_1$Mar_2021_total_deaths_per_million = NA
d_1$Apr_2021_total_deaths_per_million = NA
d_1$May_2021_total_deaths_per_million = NA
d_1$Jun_2021_total_deaths_per_million = NA
d_1$Jul_2021_total_deaths_per_million = NA
d_1$Aug_2021_total_deaths_per_million = NA
d_1$Sep_2021_total_deaths_per_million = NA
d_1$Oct_2021_total_deaths_per_million = NA
d_1$Nov_2021_total_deaths_per_million = NA
d_1$Dec_2021_total_deaths_per_million = NA
d_1$Jan_2022_total_deaths_per_million = NA
d_1$Feb_2022_total_deaths_per_million = NA
d_1$Mar_2022_total_deaths_per_million = NA
d_1$Apr_2022_total_deaths_per_million = NA
d_1$May_2022_total_deaths_per_million = NA
d_1$Jun_2022_total_deaths_per_million = NA
for (i in 1:length(d_1$location)) {
  loc = d_1$location[i]
  data_init = subset(final_dataset, final_dataset$location == loc)
  data_sub = subset(data_init, grepl("2020-05",data_init$date) == TRUE)
  d_1$May_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-06",data_init$date) == TRUE)
  d_1$Jun_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-07",data_init$date) == TRUE)
  d_1$Jul_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-08",data_init$date) == TRUE)
  d_1$Aug_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-09",data_init$date) == TRUE)
  d_1$Sep_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-10",data_init$date) == TRUE)
  d_1$Oct_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-11",data_init$date) == TRUE)
  d_1$Nov_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-12",data_init$date) == TRUE)
  d_1$Dec_2020_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-01",data_init$date) == TRUE)
  d_1$Jan_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-02",data_init$date) == TRUE)
  d_1$Feb_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-03",data_init$date) == TRUE)
```

```
  d_1$Mar_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-04",data_init$date) == TRUE)
  d_1$Apr_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-05",data_init$date) == TRUE)
  d_1$May_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-06",data_init$date) == TRUE)
  d_1$Jun_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-07",data_init$date) == TRUE)
  d_1$Jul_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-08",data_init$date) == TRUE)
  d_1$Aug_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-09",data_init$date) == TRUE)
  d_1$Sep_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-10",data_init$date) == TRUE)
  d_1$Oct_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-11",data_init$date) == TRUE)
  d_1$Nov_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-12",data_init$date) == TRUE)
  d_1$Dec_2021_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-01",data_init$date) == TRUE)
  d_1$Jan_2022_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-02",data_init$date) == TRUE)
  d_1$Feb_2022_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-03",data_init$date) == TRUE)
  d_1$Mar_2022_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-04",data_init$date) == TRUE)
  d_1$Apr_2022_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-05",data_init$date) == TRUE)
  d_1$May_2022_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-06",data_init$date) == TRUE)
  d_1$Jun_2022_total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)}
d_1$May_2020_new_deaths_per_million = NA
d_1$Jun_2020_new_deaths_per_million = NA
d_1$Jul_2020_new_deaths_per_million = NA
d_1$Aug_2020_new_deaths_per_million = NA
d_1$Sep_2020_new_deaths_per_million = NA
d_1$Oct_2020_new_deaths_per_million = NA
d_1$Nov_2020_new_deaths_per_million = NA
d_1$Dec_2020_new_deaths_per_million = NA
d_1$Jan_2021_new_deaths_per_million = NA
d_1$Feb_2021_new_deaths_per_million = NA
d_1$Mar_2021_new_deaths_per_million = NA
d_1$Apr_2021_new_deaths_per_million = NA
d_1$May_2021_new_deaths_per_million = NA
d_1$Jun_2021_new_deaths_per_million = NA
d_1$Jul_2021_new_deaths_per_million = NA
d_1$Aug_2021_new_deaths_per_million = NA
d_1$Sep_2021_new_deaths_per_million = NA
d_1$Oct_2021_new_deaths_per_million = NA
d_1$Nov_2021_new_deaths_per_million = NA
```

```
d_1$Dec_2021_new_deaths_per_million = NA
d_1$Jan_2022_new_deaths_per_million = NA
d_1$Feb_2022_new_deaths_per_million = NA
d_1$Mar_2022_new_deaths_per_million = NA
d_1$Apr_2022_new_deaths_per_million = NA
d_1$May_2022_new_deaths_per_million = NA
d_1$Jun_2022_new_deaths_per_million = NA
for (i in 1:length(d_1$location)) {
  loc = d_1$location[i]
  data_init = subset(final_dataset, final_dataset$location == loc)
  data_sub = subset(data_init, grepl("2020-05",data_init$date) == TRUE)
  d_1$May_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-06",data_init$date) == TRUE)
  d_1$Jun_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-07",data_init$date) == TRUE)
  d_1$Jul_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-08",data_init$date) == TRUE)
  d_1$Aug_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-09",data_init$date) == TRUE)
  d_1$Sep_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-10",data_init$date) == TRUE)
  d_1$Oct_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-11",data_init$date) == TRUE)
  d_1$Nov_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2020-12",data_init$date) == TRUE)
  d_1$Dec_2020_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-01",data_init$date) == TRUE)
  d_1$Jan_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-02",data_init$date) == TRUE)
  d_1$Feb_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-03",data_init$date) == TRUE)
  d_1$Mar_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-04",data_init$date) == TRUE)
  d_1$Apr_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-05",data_init$date) == TRUE)
  d_1$May_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-06",data_init$date) == TRUE)
  d_1$Jun_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-07",data_init$date) == TRUE)
  d_1$Jul_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-08",data_init$date) == TRUE)
  d_1$Aug_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-09",data_init$date) == TRUE)
  d_1$Sep_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-10",data_init$date) == TRUE)
  d_1$Oct_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-11",data_init$date) == TRUE)
  d_1$Nov_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2021-12",data_init$date) == TRUE)
  d_1$Dec_2021_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
```

```
  data_sub = subset(data_init, grepl("2022-01",data_init$date) == TRUE)
  d_1$Jan_2022_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-02",data_init$date) == TRUE)
  d_1$Feb_2022_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-03",data_init$date) == TRUE)
  d_1$Mar_2022_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-04",data_init$date) == TRUE)
  d_1$Apr_2022_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-05",data_init$date) == TRUE)
  d_1$May_2022_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  data_sub = subset(data_init, grepl("2022-06",data_init$date) == TRUE)
  d_1$Jun_2022_new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)}
d_1$May_2020_str_ind = NA
d_1$Jun_2020_str_ind = NA
d_1$Jul_2020_str_ind = NA
d_1$Aug_2020_str_ind = NA
d_1$Sep_2020_str_ind = NA
d_1$Oct_2020_str_ind = NA
d_1$Nov_2020_str_ind = NA
d_1$Dec_2020_str_ind = NA
d_1$Jan_2021_str_ind = NA
d_1$Feb_2021_str_ind = NA
d_1$Mar_2021_str_ind = NA
d_1$Apr_2021_str_ind = NA
d_1$May_2021_str_ind = NA
d_1$Jun_2021_str_ind = NA
d_1$Jul_2021_str_ind = NA
d_1$Aug_2021_str_ind = NA
d_1$Sep_2021_str_ind = NA
d_1$Oct_2021_str_ind = NA
d_1$Nov_2021_str_ind = NA
d_1$Dec_2021_str_ind = NA
d_1$Jan_2022_str_ind = NA
d_1$Feb_2022_str_ind = NA
d_1$Mar_2022_str_ind = NA
d_1$Apr_2022_str_ind = NA
d_1$May_2022_str_ind = NA
d_1$Jun_2022_str_ind = NA
for (i in 1:length(d_1$location)) {
  loc = d_1$location[i]
  data_init = subset(final_dataset, final_dataset$location == loc)
  data_sub = subset(data_init, grepl("2020-05",data_init$date) == TRUE)
  d_1$May_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-06",data_init$date) == TRUE)
  d_1$Jun_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-07",data_init$date) == TRUE)
  d_1$Jul_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-08",data_init$date) == TRUE)
  d_1$Aug_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-09",data_init$date) == TRUE)
```

```
  d_1$Sep_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-10",data_init$date) == TRUE)
  d_1$Oct_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-11",data_init$date) == TRUE)
  d_1$Nov_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2020-12",data_init$date) == TRUE)
  d_1$Dec_2020_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-01",data_init$date) == TRUE)
  d_1$Jan_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-02",data_init$date) == TRUE)
  d_1$Feb_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-03",data_init$date) == TRUE)
  d_1$Mar_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-04",data_init$date) == TRUE)
  d_1$Apr_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-05",data_init$date) == TRUE)
  d_1$May_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-06",data_init$date) == TRUE)
  d_1$Jun_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-07",data_init$date) == TRUE)
  d_1$Jul_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-08",data_init$date) == TRUE)
  d_1$Aug_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-09",data_init$date) == TRUE)
  d_1$Sep_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-10",data_init$date) == TRUE)
  d_1$Oct_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-11",data_init$date) == TRUE)
  d_1$Nov_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2021-12",data_init$date) == TRUE)
  d_1$Dec_2021_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2022-01",data_init$date) == TRUE)
  d_1$Jan_2022_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2022-02",data_init$date) == TRUE)
  d_1$Feb_2022_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2022-03",data_init$date) == TRUE)
  d_1$Mar_2022_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2022-04",data_init$date) == TRUE)
  d_1$Apr_2022_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2022-05",data_init$date) == TRUE)
  d_1$May_2022_str_ind[i] = mean(data_sub$stringency_index)
  data_sub = subset(data_init, grepl("2022-06",data_init$date) == TRUE)
  d_1$Jun_2022_str_ind[i] = mean(data_sub$stringency_index)}
d_1$May_2020_reprod_rate = NA
d_1$Jun_2020_reprod_rate = NA
d_1$Jul_2020_reprod_rate = NA
d_1$Aug_2020_reprod_rate = NA
d_1$Sep_2020_reprod_rate = NA
d_1$Oct_2020_reprod_rate = NA
d_1$Nov_2020_reprod_rate = NA
```

```
d_1$Dec_2020_reprod_rate = NA
d_1$Jan_2021_reprod_rate = NA
d_1$Feb_2021_reprod_rate = NA
d_1$Mar_2021_reprod_rate = NA
d_1$Apr_2021_reprod_rate = NA
d_1$May_2021_reprod_rate = NA
d_1$Jun_2021_reprod_rate = NA
d_1$Jul_2021_reprod_rate = NA
d_1$Aug_2021_reprod_rate = NA
d_1$Sep_2021_reprod_rate = NA
d_1$Oct_2021_reprod_rate = NA
d_1$Nov_2021_reprod_rate = NA
d_1$Dec_2021_reprod_rate = NA
d_1$Jan_2022_reprod_rate = NA
d_1$Feb_2022_reprod_rate = NA
d_1$Mar_2022_reprod_rate = NA
d_1$Apr_2022_reprod_rate = NA
d_1$May_2022_reprod_rate = NA
d_1$Jun_2022_reprod_rate = NA
for (i in 1:length(d_1$location)) {
  loc = d_1$location[i]
  data_init = subset(final_dataset, final_dataset$location == loc)
  data_sub = subset(data_init, grepl("2020-05",data_init$date) == TRUE)
  d_1$May_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-06",data_init$date) == TRUE)
  d_1$Jun_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-07",data_init$date) == TRUE)
  d_1$Jul_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-08",data_init$date) == TRUE)
  d_1$Aug_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-09",data_init$date) == TRUE)
  d_1$Sep_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-10",data_init$date) == TRUE)
  d_1$Oct_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-11",data_init$date) == TRUE)
  d_1$Nov_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2020-12",data_init$date) == TRUE)
  d_1$Dec_2020_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-01",data_init$date) == TRUE)
  d_1$Jan_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-02",data_init$date) == TRUE)
  d_1$Feb_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-03",data_init$date) == TRUE)
  d_1$Mar_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-04",data_init$date) == TRUE)
  d_1$Apr_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-05",data_init$date) == TRUE)
  d_1$May_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-06",data_init$date) == TRUE)
  d_1$Jun_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
```

```
  data_sub = subset(data_init, grepl("2021-07",data_init$date) == TRUE)
  d_1$Jul_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-08",data_init$date) == TRUE)
  d_1$Aug_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-09",data_init$date) == TRUE)
  d_1$Sep_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-10",data_init$date) == TRUE)
  d_1$Oct_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-11",data_init$date) == TRUE)
  d_1$Nov_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2021-12",data_init$date) == TRUE)
  d_1$Dec_2021_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2022-01",data_init$date) == TRUE)
  d_1$Jan_2022_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2022-02",data_init$date) == TRUE)
  d_1$Feb_2022_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2022-03",data_init$date) == TRUE)
  d_1$Mar_2022_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2022-04",data_init$date) == TRUE)
  d_1$Apr_2022_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2022-05",data_init$date) == TRUE)
  d_1$May_2022_reprod_rate[i] = mean(data_sub$reproduction_rate)
  data_sub = subset(data_init, grepl("2022-06",data_init$date) == TRUE)
  d_1$Jun_2022_reprod_rate[i] = mean(data_sub$reproduction_rate)}
d_1$cluster = c(3,2,2,1,1,1,2,2,2,2,1,1,3,2,1,2,2,1,3,3,1,3,3,2,2,2,2,1,1,1,1,3,2,2,2,2,1,3,3,1,1,2,2,1,3,1,2,3,
2,3,2,1,1,2,2,2,2,1,1,1,2,1,2,2,3,2,2,1,2,3,2,1,1,3,3,2,3,1,2,2,2,2,3,3,3,1,1,2,3,1,2,3,2,2,2,2,1,1,2,1,1,2,1,1,1,3,
1,1,2,3,1,1,3,3,2,3,2,2,2,1,2,1,1,1,3,2,3,3,3)
X = d_1[,-c(1,158)]
d_1.pca = prcomp(x = X,center = TRUE,scale. = TRUE)
cumulat.var = (cumsum(d_1.pca$sdev^2)/(sum(d_1.pca$sdev^2)))*100
d_1 = cbind(d_1,d_1.pca$x[,1:4])
d_1$cluster = as.character(d_1$cluster)
ggplot(d_1,aes(x = PC1,y = PC2,col = cluster)) +geom_point() +
  geom_text(label = d_1$location,nudge_x = 0.25, nudge_y = 0.25,) +
  ggtitle("Scatterplot with the scores in the first and second PC")
ggplot(d_1,aes(x = PC1,y = PC3,col = cluster))+geom_point()+
  geom_text(label = d_1$location,nudge_x = 0.25, nudge_y = 0.25, ) +
  ggtitle("Scatterplot with the scores in the first and third PC")
ggplot(d_1,aes(x = PC1,y = PC4,col = cluster))+geom_point()+
  geom_text(label = d_1$location,nudge_x = 0.25, nudge_y =0.25,) +
  ggtitle("Scatterplot with the scores in the first and fourth PC")
final_dataset$cluster = NA
for (i in 1:nrow(dataset.clust)) {
  loc = dataset.clust$location[i]
  final_dataset$cluster[final_dataset$location == loc] = rep(dataset.clust$cluster[i],
  sum(final_dataset$location == loc))}
result = aggregate(. ~ cluster, dataset.clust[,-c(1,2)], FUN = function(x) c(mean = mean(x), sd = sd(x),
median = median(x), min = min(x), max = max(x), quant = quantile(x, probs = c(0,0.25,0.5,0.75,1))))
summary(dataset.clust[,-c(1,2,ncol(dataset.clust))])
dataset.clust$cluster = as.factor(dataset.clust$cluster)
```

```
dataset.clust |>
  ggplot(aes(population, fill = cluster)) +geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"),name = 'Data') +
  ggtitle("Histograms for populations")
dataset.clust$cluster = factor(dataset.clust$cluster)
dataset.clust |>
  ggplot(aes(life_expectancy, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) +facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"),name = 'Data') +
  ggtitle("Histograms for life_expectancy")
dataset.clust |>
  ggplot(aes(population_density, fill = cluster)) +geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for population_density")
dataset.clust |>
  ggplot(aes(diabetes_prevalence, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for diabetes_prevalence")
dataset.clust |>
  ggplot(aes(median_age, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for median_age")
dataset.clust |>
  ggplot(aes(aged_70_older, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for aged_70_older")
dataset.clust |>
  ggplot(aes(cardiovasc_death_rate, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for cardiovasc_death_rate")
dataset.clust |>
  ggplot(aes(gdp_per_capita, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for gdp_per_capita")
dataset.clust |>
```

```
  ggplot(aes(human_development_index, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for human_development_index")
dataset.clust |>
  ggplot(aes(hospital_beds_per_thousand, fill = cluster)) + geom_histogram() +
  geom_histogram(data = dataset.clust |> mutate(cluster = gl(1, nrow(dataset.clust),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for hospital_beds_per_thousand")
result = aggregate(. ~ cluster, final_dataset[,-(1:4)], FUN = function(x) c(mean = mean(x), sd = sd(x),
median = median(x), min = min(x), max = max(x), quant = quantile(x, probs = c(0,0.25,0.5,0.75,1))))
summary(final_dataset[,c("total_cases_per_million","new_cases_per_million",
"total_deaths_per_million","new_deaths_per_million","stringency_index","reproduction_rate")])
final_dataset$cluster = as.factor(final_dataset$cluster)
final_dataset |>
  ggplot(aes(total_cases_per_million, fill = cluster)) + geom_histogram() +
  geom_histogram(data = final_dataset |> mutate(cluster = gl(1, nrow(final_dataset),
  labels = c('All data')))) +facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"),name = 'Data') +
  ggtitle("Histograms for total_cases_per_million")
final_dataset |>
  ggplot(aes(new_cases_per_million, fill = cluster)) + geom_histogram() +
  geom_histogram(data = final_dataset |> mutate(cluster = gl(1, nrow(final_dataset),
  labels = c('All data')))) +facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for new_cases_per_million")
final_dataset |>
  ggplot(aes(total_deaths_per_million, fill = cluster)) + geom_histogram() +
  geom_histogram(data = final_dataset |> mutate(cluster = gl(1, nrow(final_dataset),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for total_deaths_per_million")
final_dataset |>
  ggplot(aes(new_deaths_per_million, fill = cluster)) + geom_histogram() +
  geom_histogram(data = final_dataset |> mutate(cluster = gl(1, nrow(final_dataset),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for new_deaths_per_million")
final_dataset |>
  ggplot(aes(stringency_index, fill = cluster)) + geom_histogram() +
  geom_histogram(data = final_dataset |> mutate(cluster = gl(1, nrow(final_dataset),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for stringency_index")
final_dataset |>
  ggplot(aes(reproduction_rate, fill = cluster)) + geom_histogram() +
  geom_histogram(data = final_dataset |> mutate(cluster = gl(1, nrow(final_dataset),
  labels = c('All data')))) + facet_wrap(~ cluster, ncol = 1) +
```

```
  scale_fill_discrete(breaks = c("1", "2","3","All data"), name = 'Data') +
  ggtitle("Histograms for reproduction_rate")
df_line_chart = data.frame(date = rep(seq(as.Date("2020-05-13"), as.Date("2022-06-18"),by = "days"),
each = 3))
df_line_chart$cluster = rep((1:3),length((seq(as.Date("2020-05-13"), as.Date("2022-06-18"),by = "days"))))
df_line_chart$total_cases_per_million = NA
df_line_chart$total_deaths_per_million = NA
df_line_chart$new_cases_per_million = NA
df_line_chart$new_deaths_per_million = NA
df_line_chart$stringency_index = NA
df_line_chart$reproduction_rate = NA
for (i in 1:nrow(df_line_chart)) {
  date = df_line_chart$date[i]
  cl = df_line_chart$cluster[i]
  data_sub = final_dataset[final_dataset$date == date & final_dataset$cluster == cl, ]
  df_line_chart$total_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  df_line_chart$total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  df_line_chart$new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
  df_line_chart$new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  df_line_chart$stringency_index[i] = mean(data_sub$stringency_index)
  df_line_chart$reproduction_rate[i] = mean(data_sub$reproduction_rate)}
df_line_chart$cluster = as.factor(df_line_chart$cluster)
df_line_chart_all = data.frame(date = seq(as.Date("2020-05-13"), as.Date("2022-06-18"),by = "days"))
df_line_chart_all$total_cases_per_million = NA
df_line_chart_all$total_deaths_per_million = NA
df_line_chart_all$new_cases_per_million = NA
df_line_chart_all$new_deaths_per_million = NA
df_line_chart_all$stringency_index = NA
df_line_chart_all$reproduction_rate = NA
for (i in 1:nrow(df_line_chart_all)) {
  date = df_line_chart_all$date[i]
  data_sub = final_dataset[final_dataset$date == date, ]
  df_line_chart_all$total_cases_per_million[i] = mean(data_sub$total_cases_per_million)
  df_line_chart_all$total_deaths_per_million[i] = mean(data_sub$total_deaths_per_million)
  df_line_chart_all$new_cases_per_million[i] = mean(data_sub$new_cases_per_million)
  df_line_chart_all$new_deaths_per_million[i] = mean(data_sub$new_deaths_per_million)
  df_line_chart_all$stringency_index[i] = mean(data_sub$stringency_index)
  df_line_chart_all$reproduction_rate[i] = mean(data_sub$reproduction_rate)}
df_line_chart_all = df_line_chart_all[rep(seq_len(nrow(df_line_chart_all)), each = 3), ]
df_line_chart$cluster = as.factor(df_line_chart$cluster)
df_line_chart_all$cluster = "All data"
df_line_chart_all = df_line_chart_all[, c("date","cluster","total_cases_per_million",
"total_deaths_per_million","new_cases_per_million","new_deaths_per_million","stringency_index",
"reproduction_rate")]
d_new = rbind(df_line_chart,df_line_chart_all)
ggplot(d_new, aes(x = date, y = new_cases_per_million, color = cluster)) + geom_line() +
  labs(x = "Date", y = "Mean new cases per million", color = "Cluster") +
  ggtitle("Line graphs of mean new cases per million") +
 scale_color_manual(values = c("red", "green", "blue", "black"), labels = c("1", "2", "3", "All Data"))
ggplot(d_new, aes(x = date, y = total_cases_per_million, color = cluster)) + geom_line() +
```

```
  labs(x = "Date", y = "Mean total cases per million", color = "Cluster") +
  ggtitle("Line graphs of mean total cases per million") +
 scale_color_manual(values = c("red", "green", "blue", "black"), labels = c("1", "2", "3", "All Data"))
ggplot(d_new, aes(x = date, y = new_deaths_per_million, color = cluster)) + geom_line() +
  labs(x = "Date", y = "Mean new deaths per million", color = "Cluster") +
  ggtitle("Line graphs of mean new deaths per million") +
 scale_color_manual(values = c("red", "green", "blue", "black"), labels = c("1", "2", "3", "All Data"))
ggplot(d_new, aes(x = date, y = total_deaths_per_million, color = cluster)) + geom_line() +
  labs(x = "Date", y = "Mean total deaths per million", color = "Cluster") +
  ggtitle("Line graphs of mean total deaths per million") +
 scale_color_manual(values = c("red", "green", "blue", "black"), labels = c("1", "2", "3", "All Data"))
ggplot(d_new, aes(x = date, y = reproduction_rate, color = cluster)) + geom_line() +
  labs(x = "Date", y = "Mean reproduction rate", color = "Cluster") +
  ggtitle("Line graphs of mean reproduction rate") +
 scale_color_manual(values = c("red", "green", "blue", "black"), labels = c("1", "2", "3", "All Data"))
ggplot(d_new, aes(x = date, y = stringency_index, color = cluster)) + geom_line() +
  labs(x = "Date", y = "Mean stringency index", color = "Cluster") +
  ggtitle("Line graphs of mean stringency index") +
 scale_color_manual(values = c("red", "green", "blue", "black"), labels = c("1", "2", "3", "All Data"))
```

Για τον παραπάνω κώδικα αξιοποιήθηκαν οι πηγές [19], [18], [17], [8]

Βιβλιογραφία

[1] Ιωαννίδου Κυριακή (2011). *Διαχωριστική Ανάλυση και εφαρμογές*. Προπτ/κή Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών.

[2] Καρλής Δημήτρης (2005). *Πολυμεταβλητή Στατιστική Ανάλυση*. Εκδόσεις Σταμούλη.

[3] Κύρκος Ευστάθιος (2015). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων:* Κεφ. 11. Ανάλυση Συστάδων. Εκδόσεις Κάλλιπος, Open Academic Editions.

[4] Παπούλια Καλλιόπη (2020). *Μέθοδοι Πολυμεταβλητής Ανάλυσης Επιχειρηματικών Δεδομένων*. Μεταπτ/κή Διπλωματική Εργασία, Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Π.Μ.Σ Εφαρμοσμένη Στατιστική.

[5] Στρατινάκης Νικόλαος (2018). *Εφαρμοσμένη Ανάλυση Συστάδων*. Μεταπτ/κή Διπλωματική Εργασία, Πολυτεχνείο Κρήτης Σχολή Μηχανικών Παραγωγής και Διοίκησης, Π.Μ.Σ Εφαρμοσμένα Μαθηματικά για μηχανικούς.

[6] Gondwe Grace (2020). Assessing the impact of COVID-19 on Africa's economic development. *United Nations Conference on Trade and Development,* Geneva 2020.

[7] Shams Shahbaz A., Haleem Abid and Javaid Mohd (2020). Analyzing COVID-19 pandemic for unequal distribution of tests, identified cases, deaths, and fatality rates in the top 18 countries. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* **14** (5), 953-961.

Διαδικτυακές πηγές

[8] Ron Ammar. *randomcoloR: Generate Attractive Random Colors*. 2019. URL: https://CRAN.R-project.org/package=randomcoloR.

[9] George Bogdanov. *The impact of the COVID-19 crisis on poverty and social exclusion in Bulgaria*. https://ec.europa.eu/social/BlobServlet?docId=22934&langId=en. 2020.

[10] Eurostat. *Euro area unemployment at 7.3% EU at 6.5%*. 2020. URL: https://ec.europa.eu/eurostat/documents/2995521/10662618/3-01042020-AP-EN.pdf/be3d73ee-6715-824b-2c23-f0512f12bdc6?t=1585657691000.

[11] Eurostat. *Euro area unemployment at 7.3% EU at 6.6%*. 2020. URL: https://ec.europa.eu/eurostat/documents/2995521/10294960/3-03062020-AP-EN.pdf/b823ec2b-91af-9b2a-a61c-0d19e30138ef?t=1591123422000.

[12] Eurostat. *Euro area unemployment at 7.9%*. 2020. URL: https://ec.europa.eu/eurostat/documents/2995521/10568643/3-01092020-BP-EN.pdf/39668e66-2fd4-4ec0-9fd4-4d7c99306c98?t=1598882965000.

[13] Laurent Ferrara and Valerie Mignon. *The Covid-19 recession in France: The trough is behind us, but let's stay vigilant*. 2021. URL: https://cepr.org/voxeu/columns/covid-19-recession-france-trough-behind-us-lets-stay-vigilant.

[14] Hector Florez and Sweta Singh. *Online dashboard and data analysis approach for assessing COVID-19 case and death data*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7450472/. 2020.

[15] GitHub. *Our World in Data*. URL: https://github.com/owid.

[16] Rosamond Hutt. *The economic effects of the coronavirus around the world*. 2020. URL: https://www.weforum.org/agenda/2020/02/coronavirus-economic-effects-global-economy-trade-travel/.

[17] Hadley Wickham, Romain François and Lionel Henry. *dplyr: A Grammar of Data Manipulation*. 2022. URL: https://CRAN.R-project.org/package=dplyr.

[18] Hadley Wickham, Jim Hester and Jennifer Bryan. *readr: Read Rectangular Text Data*. 2022. URL: https://CRAN.R-project.org/package=readr.

[19] Hadley Wickham, Carson Sievert and Springer. *Ggplot2 : Elegant Graphics For Data Analysis*. 2016. URL: https://ggplot2.tidyverse.org.