ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑ & ΥΠΗΡΕΣΙΕΣ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΦΑΡΜΟΓΗ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΕΝΑΝΤΙ ΤΟΥ ΙΟΥ SARS-CoV-2

Αντώνιος Μπαλάσκας

Πειραιάς, Ιανουάριος 2023

UNIVERSITY OF PIRAEUS DEPARTMENT OF DIGITAL SYSTEMS

MASTER PROGRAM IN INFORMATION SYSTEMS AND SERVICES

DATA ANALYSIS AND APPLICATION OF MACHINE LEARNING AGAINST VIRUS SARS-CoV-2

Antonios Balaskas

Piraeus, Greece, January 2023

Στην οικογένειά μου

Ευχαριστώ πολύ την Παναγιώτα Δεσποτάκη για την συντακτική επιμέλεια της παρούσας διπλωματικής εργασίας.

Ανάλυση Δεδομένων και Εφαρμογή Μηχανικής Μάθησης Έναντι του Ιού SARS-CoV-2

Περίληψη

Η παρούσα εργασία επικεντρώνεται στον αντίκτυπο της νόσου με το διεθνές όνομα SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) που προκαλεί την ασθένεια COVID-19, η οποία ευθύνεται για την τρέχουσα πανδημία του COVID-19. Συγκεκριμένα χρησιμοποιείται ο αλγόριθμος μηχανικής μάθησης ARIMA (Auto-Regressive Integrated Moving Average), που είναι μοντέλο αυτοπαλινδρομικού ολοκληρωμένου κινητού μέσου όρου, το οποίο βασίζεται σε δεδομένα χρονοσειρών για την πρόβλεψη μελλοντικών τιμών. Η ανάλυση έγινε με την χρήση της γλώσσας προγραμματισμού Python παρουσιάζοντας την εξάπλωση της νόσου παγκοσμίως και κυρίως τα στοιχεία που αφορούν την Ελλάδα. Επίσης, με κατάλληλες οπτικοποιήσεις υποδεικνύεται η διάδοση του ιού και αναλύονται τα δεδομένα εμβολιασμού κατά της πανδημίας του COVID-19. Τέλος, παρουσιάζεται ένας συνοπτικός πίνακας με στατιστικά δεδομένα με την χρήση του εργαλείου οπτικοποίησης δεδομένων Tableau.

Data Analysis and Application of Machine Learning Against Virus SARS-CoV-2

Abstract

The present study focuses on the impact of the disease known internationally as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), which causes COVID-19 and is responsible for the ongoing pandemic. Specifically, the study will use the ARIMA (Auto-Regressive Integrated Moving Average) machine learning algorithm to predict future values, based on time-series data. The analysis will be conducted using the Python programming language, with a focus on global data, including data specific to Greece. The study will also include appropriate visualizations to illustrate the spread of the virus and analysis of vaccination data related to the COVID-19 pandemic. The study will conclude with a summary table of statistical data, using the Tableau data visualization tool.

Πίνακας περιεχομένων

Πίνακας εικόνων

ΚΕΦΑΛΑΙΟ 1

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

1.1 Αναδρομή του SARS-CoV-2 και η σημαντικότητα της Μηχανικής Μάθησης

Η νόσος του κορονοϊού (COVID-19) είναι μια μεταδοτική νόσος που προκαλείται από τον Ιό SARS-CoV-2 και αναφέρθηκε για πρώτη φορά στην Wuhan της Κίνας στα τέλη του 2019. Η επιδημία έκτοτε έχει εξελιχθεί σε πανδημία, φθάνοντας σε χώρες πολύ πιο πέρα από την Κίνα. Με περίπου 540 εκατομμύρια περιπτώσεις και 6 εκατομμύρια θανάτους παγκοσμίως (Ιούνιος 2022), η πανδημία COVID-19 είναι το σύγχρονο στοιχείο ανησυχίας σε όλο τον κόσμο διότι ο αντίκτυπος αυτής της νόσου είναι τεράστιος. Στις 30 Ιανουαρίου 2020, η Επιτροπή Έκτακτης Ανάγκης Διεθνών Κανονισμών της Υγείας, δηλαδή ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) ανακοίνωσε το ξέσπασμα της νόσου που προέκυψε από τον κορονοϊό SARS-CoV-2 (ασθένεια κορονοϊού 2019, γνωστή και ως οξεία αναπνευστική νόσος 2019-nCoV), ως «επείγουσα ανάγκη δημόσιας υγείας σε διεθνές επίπεδο». Η νόσος του COVID-19, προκαλεί σοβαρή οξεία αναπνευστική ανεπάρκεια, έχει εξαπλωθεί σε όλες τις ηπείρους και δημιούργησε μία άνευ προηγουμένου κρίση δημόσιας υγείας. Τον Οκτώβριο του 2020, το κέντρο ιατρικής στο Πανεπιστήμιο Johns Hopkins ανέφερε συνολικά περισσότερο από 1 εκατομμύριο θανάτους καθώς οι λοιμώξεις από τον COVID-19 παγκοσμίως ξεπέρασαν τα 40 εκατομμύρια. Αυτή η γρήγορη εξάπλωση οφείλεται στο γεγονός ότι ο ιός μεταδίδεται από άτομο σε άτομο πολύ εύκολα μέσω του βήχα, του φτερνίσματος και των σταγονιδίων του αναπνευστικού. Συνήθως εκδηλώνεται με συμπτώματα πυρετού, βήχα, δύσπνοιας και μπορεί να έχει σοβαρές συνέπειες όπως πνευμονία, πολυοργανική ανεπάρκεια ακόμη και θάνατο. Σήμερα, έχει αναπτυχθεί μια ξεκάθαρη λύση για το COVID-19 με την χρήση κατάλληλων εμβολίων όπως της Pfizer-BioNTech, το οποίο είναι ένα mRNA εμβόλιο κατά της λοίμωξης COVID-19 που αναπτύχθηκε από την γερμανική εταιρεία BioNTech σε συνεργασία με την αμερικανική εταιρεία Pfizer. Είναι το πρώτο εμβόλιο που εγκρίθηκε από μία αυστηρή αρχή για χρήση έκτακτης ανάγκης και συνάμα το πιο διαδεδομένο με ποσοστό αποτελεσματικότητας κατά 95% σε μια συνεχιζόμενη κλινική δοκιμή μεγάλης κλίμακας. Η συντριπτική πλειοψηφία των χωρών πριν την ανάπτυξη των εμβολίων

λάμβανε μέτρα για την πρόληψη της μετάδοσης αυτού του ιού με την χρήση απαγόρευσης της κυκλοφορίας για την αποφυγή συνωστισμών. Αυτό συνέβαινε λόγω της αβεβαιότητας στον τρόπο μετάδοσης του SARS-CoV-2 και της υψηλής βεβαιότητας για τον υψηλό δείκτη θνησιμότητας του λοιμογόνου αυτού ιού. Δεδομένου του αυξανόμενου φόρτου υποθέσεων, υπάρχει επείγουσα ανάγκη να αυξηθεί η ιατρική και η οικονομική ενίσχυση για την αντιμετώπιση αυτής της κρίσιμης ασθένειας. Ως εκ τούτου, η επιστημονική πρόκληση είναι να εντοπιστούν, μέσω συμπερασμάτων και προσομοίωσης, μέτρα που θα μπορούσαν να προσφέρουν μια καλύτερη προστασία με λιγότερο κοινωνικό κόστος. Η αυξανόμενη έμφαση στη μηχανική μάθηση και οι τεχνικές εκμάθησης σε ιατρικούς τομείς μπορούν να παρέχουν το κατάλληλο περιβάλλον για να υπάρξει αλλαγή και βελτίωση. Αυτή η εργασία επικεντρώνεται στον αντίκτυπο της μηχανικής μάθησης, στην πανδημία του COVID-19 όπου η μηχανική μάθηση έχει αποδειχθεί ανεκτίμητη για την πρόβλεψη κινδύνων σε πολλούς τομείς και από τότε που ξεκίνησε η εξάπλωση του ιού, η εφαρμογή της βοηθάει στην καταπολέμηση της ιογενούς πανδημίας.

Η επιστημονική κοινότητα σε όλο τον κόσμο συλλέγει, μοιράζει δεδομένα και νέες ανακαλύψεις για τον ιό. Εκατοντάδες ερευνητικές ομάδες συνδυάζουν τις προσπάθειές τους να συλλέξουν δεδομένα και να αναπτύξουν λύσεις καθημερινώς. Ξεκινώντας από αυτό, οι κύριοι στόχοι αυτής της εργασίας είναι να εμβαθύνει στον τρόπο εφαρμογής των τεχνικών μηχανικής μάθησης σε διαφορετικά επιστημονικά πεδία που πλήττονται από την πανδημία για να βοηθήσει στην καταπολέμηση του κορονοϊού. Η έγκαιρη θεραπεία με τη βοήθεια των υπολογιστών, με την χρήση αλγορίθμων μηχανικής μάθησης αποδείχθηκαν επειγόντως απαραίτητοι. Η εφαρμογή τους θα μπορούσε να μειώσει σε μεγάλο βαθμό τις προσπάθειες των κλινικών δοκιμών και να επιταχύνει τη διαδικασία προσυμπτωματικού ελέγχου και διάγνωσης, καθώς και την μείωση της ανθρώπινης παρέμβασης στην ιατρική πράξη. Τέλος, η εφαρμογή μηχανικής μάθησης στην εργαστηριακή έρευνα επέτρεψε στην επιτάχυνση της ανάπτυξης των εμβολίων κατά του ιού SARS-CoV-2 και επίσης βοηθάει για την ανακάλυψη νέων φαρμάκων που έχουν πιθανή δράση κατά του κορονοϊού.

1.2 Επιδημιολογικά Μοντέλα

Στη μελέτη ασθενειών, είναι πολύ συνηθισμένο να βλέπουμε επιδημιολογικά μοντέλα να χρησιμοποιούνται για να αξιολογήσουν τους τρόπους ανάπτυξης των μολυσματικών ασθενειών και να προβλέψουν τα μελλοντικά αποτελέσματα των επιδημιών. Ένα συγκεκριμένο μοντέλο, γνωστό ως μοντέλο Susceptible-Infected-Removed (SIR) υπολογίζει τον θεωρητικό αριθμό ατόμων που έχουν προσβληθεί από συγκεκριμένη νόσο σε ένα κλειστό πληθυσμό πάνω από ένα α χρονικό διάστημα. Αυτό το μοντέλο περιλαμβάνει μια σειρά από εξισώσεις που χρησιμοποιούν τον αριθμό ευπαθών (S), μολυσμένων (I) και αναρρωμένων (R) ατόμων και έχει χρησιμοποιηθεί για την μοντελοποίηση της εξάπλωσης της νόσου COVID-19 στα αρχικά στάδια της πανδημίας στην Κίνα καθώς και σε άλλες χώρες όπως η Νότια Κορέα, η Ινδία, η Αυστραλία, οι ΗΠΑ, και η Ιταλία. Άλλες μελέτες έχουν επεκτείνει ακόμη περισσότερο αυτό το μοντέλο, χρησιμοποιώντας το μοντέλο Susceptible-Exposed-Infected-Removed (SEIR). Αυτό είναι παρόμοιο με το μοντέλο SIR αλλά ενσωματώνει και τον εκτεθειμένο (E) πληθυσμό στους υπολογισμούς του. Το SEIR μοντέλο έχει επίσης χρησιμοποιηθεί για την πρόβλεψη της εξάπλωσης του COVID-19 σε χώρες όπως είναι η Κίνα και η Ιταλία. Συνολικά, οι ερευνητές μπόρεσαν να παράγουν έγκυρα αποτελέσματα και προβλέψεις για τις αντίστοιχες χώρες που μελετήθηκαν.

1.3 Τα είδη της Μηχανικής Μάθησης

Η Μηχανική Μάθηση είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης (AI) που εξελίχθηκε από την αναγνώριση προτύπων, της οποίας τα δεδομένα μπορούν να δομηθούν για την κατανόηση τους από τους χρήστες. Πρόσφατα, πολλές εφαρμογές έχουν αναπτυχθεί χρησιμοποιώντας τη Μηχανική Εκμάθηση σε διάφορους τομείς, όπως η υγειονομική περίθαλψη, στις τραπεζικές συναλλαγές, στον στρατιωτικό εξοπλισμό, στο διάστημα κ.λπ. Επί του παρόντος, η Μηχανική Μάθηση είναι ένας ταχέως εξελισσόμενος και συνεχώς αναπτυσσόμενος τομέας. Προγραμματίζει υπολογιστές που χρησιμοποιούν δεδομένα για τη βελτιστοποίηση της απόδοσής τους. Μαθαίνει τις παραμέτρους να βελτιστοποιούν τα προγράμματα υπολογιστών χρησιμοποιώντας τα δεδομένα εκπαίδευσης ή τις προηγούμενες εμπειρίες τους. Χρησιμοποιώντας τα δεδομένα, μπορεί

ακόμα και να προβλέψει το μέλλον. Η μηχανική μάθηση βοηθά επίσης στην κατασκευή ενός μαθηματικού μοντέλου χρησιμοποιώντας τα στατιστικά στοιχεία των δεδομένων.
Ο κύριος στόχος της είναι να μαθαίνει από τα δεδομένα χωρίς καμία ανθρώπινη παρέμβαση και να μας δίνει το επιθυμητό αποτέλεσμα από την αναζήτηση διαφόρων τάσεων/μοτίβων στα δεδομένα. Είναι ευρέως ταξινομημένη σε τέσσερις τύπους:

• Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning).

• Μη εποπτευόμενη Μηχανική Εκμάθηση (Unsupervised Machine Learning).

• Ημι-Εποπτευόμενη Μηχανική Εκμάθηση (Semi-Supervised Machine Learning).

• Ενισχυτική Μηχανική Μάθηση (Reinforcement Machine Learning).

Εποπτευόμενη Μηχανική Εκμάθηση

Η εποπτευόμενη μάθηση είναι ένα μοντέλο μηχανικής μάθησης που έχει δημιουργηθεί για να παρέχει προβλέψεις. Αυτός ο αλγόριθμος εκτελείται λαμβάνοντας ως είσοδο ένα σύνολο δεδομένων και επίσης γνωστές απαντήσεις ως έξοδο για την εκμάθηση του μοντέλου παλινδρόμησης/ταξινόμησης. Αναπτύσσει προγνωστικά μοντέλα από αλγόριθμους ταξινόμησης και τεχνικές παλινδρόμησης.
Η κατηγοριοποίηση προβλέπει διακριτές απαντήσεις. Εδώ, ο αλγόριθμος κατηγοριοποιεί σε δύο ή περισσότερες τάξεις για κάθε παράδειγμα. Αν γίνεται μεταξύ δύο τάξεων τότε ονομάζεται δυαδική κατηγοριοποίηση και αν γίνεται μεταξύ δύο ή περισσότερων κλάσεων τότε ονομάζεται κατηγοριοποίηση πολλαπλών τάξεων. Οι εφαρμογές της κατηγοριοποίησης περιλαμβάνουν την αναγνώριση γραφής με το χέρι και την ιατρική απεικόνιση.
Η παλινδρόμηση προβλέπει συνεχείς αποκλίσεις. Εδώ, οι αλγόριθμοι επιστρέφουν μια στατιστική τιμή. Τα πιο γνωστά είδη τεχνικών παλινδρόμησης είναι:

• Γραμμική παλινδρόμηση.

• Λογιστική παλινδρόμηση.

Μη εποπτευόμενη Μηχανική Εκμάθηση

Σε αντίθεση με την εποπτευόμενη μάθηση, δεν υπάρχει επόπτης εδώ αλλά έχουμε μόνο εισερχόμενα δεδομένα. Εδώ, ο βασικός στόχος είναι να βρεθούν ορισμένα μοτίβα με την μεγαλύτερη συχνότητα εμφάνισης. Σύμφωνα με την στατιστική, αυτό ονομάζεται εκτίμηση πυκνότητας. Μία από τις μεθόδους για την εκτίμηση της πυκνότητας είναι η συσταδοποίηση. Σε αυτήν σχηματίζονται τα δεδομένα εισόδου σε συστάδες ή ομάδες. Οι υποθέσεις γίνονται έτσι ώστε οι συστάδες να ανακαλύψουν που θα ταιριάξουν αρκετά καλά με μια κατηγοριοποίηση. Αυτό είναι μια προσέγγιση καθοδηγούμενη από τα δεδομένα και λειτουργεί καλύτερα όταν παρέχεται μεγάλος όγκος δεδομένων. Για παράδειγμα, οι ταινίες στο Netflix προτείνονται με βάση την αρχή της συσταδοποίησης ταινιών, όπου πολλές παρόμοιες ταινίες ομαδοποιούνται με βάση την ταινία που παρακολούθησε πρόσφατα ο χρήστης. Ανακαλύπτει κυρίως τα άγνωστα μοτίβα στα δεδομένα αλλά τις περισσότερες φορές αυτές οι προσεγγίσεις είναι αδύναμες σε σύγκριση με την εποπτευόμενη μάθηση.

Ημι-εποπτευόμενη Μηχανική Εκμάθηση

Η ονομασία «ημι-εποπτευόμενη μάθηση» προέρχεται από το γεγονός ότι τα δεδομένα που χρησιμοποιούνται είναι μεταξύ εποπτευόμενης και μη εποπτευόμενης μηχανικής μάθησης. Ο ημι-εποπτευόμενος αλγόριθμος έχει την τάση για μάθηση τόσο από δεδομένα με ετικέτα όσο και από δεδομένα χωρίς ετικέτα. Η ημι-εποπτευόμενη μηχανική μάθηση προσφέρει υψηλή ακρίβεια με ελάχιστη εργασία επεξήγησης. Η εκμάθηση αυτή επίσης χρησιμοποιεί κυρίως δεδομένα χωρίς ετικέτα σε συνδυασμό με δεδομένα με ετικέτα για να δώσει καλύτερους ταξινομητές.

Ενισχυτική Μηχανική Μάθηση

Η ενισχυτική μάθηση μαθαίνει τη συμπεριφορά της από μια μέθοδο δοκιμής και λάθους σε ένα δυναμικό περιβάλλον. Εδώ, το πρόβλημα λύνεται με την ανάληψη κατάλληλης δράσης μιας ορισμένης κατάστασης για τη μεγιστοποίηση της παραγωγής και για την απόκτηση των επιτευχθέντων αποτελεσμάτων. Στην Ενισχυτική Μάθηση, ο αλγόριθμος ενίσχυσης μαθαίνει δοκιμάζοντας πολλούς τρόπους και κάνοντας αρκετά λάθη αλλά στην πορεία μαθαίνοντας από τα λάθη αυτά κάνει πολύ λιγότερα. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού όπως η κίνηση των ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους. Το μοντέλο αυτό αποτελείται από:

• ένα διακριτό σύνολο καταστάσεων περιβάλλοντος, S.

• ένα διακριτό σύνολο ενεργειών πράκτορα, A.

• ένα σύνολο βαθμωτών σημάτων ενίσχυσης συνήθως {0;1} ή τους πραγματικούς αριθμούς.

1.4 Μοντέλα Μηχανικής Μάθησης

Στον τομέα της μηχανικής μάθησης έχουν αναφερθεί πολλά μοντέλα που προβλέπουν αποτελεσματικά την εξάπλωση του COVID-19. Από κλασικά μοντέλα όπως το αυτοπαλινδρομικό μοντέλο κινητών μέσων (Auto-Regressive Integrated Moving Average-ARIMA) όπως και του επαναλαμβανόμενου τεχνητού νευρωνικού δικτύου γνωστό και ως Long-Short Term Memory (LSTM). Επιπλέον, οι ερευνητές έχουν πραγματοποιήσει συγκριτικές μελέτες παρατήρησης πολλαπλών διαφορετικών μοντέλων μηχανικής μάθησης ταυτόχρονα και συγκρίνοντας τα αποτελέσματα πρόβλεψής τους. Τα νευρωνικά δίκτυα αρχίζουν να γίνονται πιο δημοφιλή λόγω της ικανότητάς τους να αποτυπώνουν χωρικές σχέσεις μέσα στα μοντέλα τους. Αυτό σχετίζεται άμεσα με τον COVID-19 επειδή ο αντίκτυπος της νόσου ποικίλλει ανάλογα με την παρατηρούμενη τοποθεσία διότι αρκετές τοποθεσίες έχουν διαφορετικούς πληθυσμούς, υγειονομικούς κανονισμούς κ.λπ.

ΚΕΦΑΛΑΙΟ 2

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

2.1 Προετοιμασία Δεδομένων

Καθώς η μελέτη μας περιλαμβάνει αλγόριθμο μηχανικής μάθησης σχετικά με την εξάπλωση του COVID-19 παγκοσμίως, τα δεδομένα ήταν μια βασική πτυχή της έρευνας. Τα δεδομένα που χρησιμοποιήθηκαν είναι από το πανεπιστήμιο των ΗΠΑ Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) και αφορούν την περίοδο από 22 Ιανουαρίου 2020 έως 30 Ιουνίου 2022. Τα δεδομένα ανανεώνονται καθημερινώς στο παρακάτω σύνδεσμο όπου ανήκει στην πιο διάσημη πλατφόρμα φιλοξενίας κώδικα το GitHub https://github.com/CSSEGISandData/COVID-19. Στην πρώτη φάση της ανάλυσης χρησιμοποιήθηκε το αρχείο time_series_covid19_confirmed_global.csv όπου περιλαμβάνει σχεδόν όλες τις χώρες του κόσμου μαζί με τις επαρχίες τους, το γεωγραφικό πλάτος και μήκος της κάθε χώρας (συντεταγμένες), και τα συνολικά επίσημα κρούσματα του COVID-19 όπου δηλώνει η κάθε χώρα από 22 Ιανουαρίου 2020 έως 30 Ιουνίου 2022 αθροιστικά. Το περιβάλλον ανάπτυξης που χρησιμοποιείται σε αυτή την εργασία για την φόρτωση και την επεξεργασία των αρχείων .csv για να εξαχθούν χρήσιμα δεδομένα με την γραφή κατάλληλου κώδικα σε γλώσσα Python είναι το Jupyter Notebook. Το Jupyter Notebook είναι μια εφαρμογή web ανοιχτού κώδικα που μπορεί να χρησιμοποιηθεί για να δημιουργηθεί και να μοιραστούν έγγραφα που περιέχουν κώδικα, εξισώσεις, απεικονίσεις και κείμενο. Δύο διάσημες βιβλιοθήκες ανοιχτού κώδικα της Python που χρησιμοποιούνται ευρέως για την πρώτη επεξεργασία των δεδομένων είναι η Pandas και η NumPy. Η Pandas χρησιμοποιείται για εργασίες επιστήμης/ανάλυσης δεδομένων και μηχανικής μάθησης όπου τα δεδομένα αυτά είναι συνήθως αποθηκευμένα σε υπολογιστικά φύλλα ή σε βάσεις δεδομένων. Είναι χτισμένο πάνω στην NumPy, η οποία παρέχει υποστήριξη για μαθηματικές πράξεις με πολυδιάστατους πίνακες και είναι το κατάλληλο εργαλείο για εξερεύνηση, εκκαθάριση και επεξεργασία δεδομένων. Στο περιβάλλον της βιβλιοθήκης της Pandas ένας πίνακας δεδομένων ονομάζεται DataFrame. Η NumPy (Numerical Python) είναι μια βιβλιοθήκη

της Python ανοιχτού κώδικα που χρησιμοποιείται σχεδόν σε κάθε τομέα της επιστήμης και της μηχανικής. Είναι το καθολικό πρότυπο για την εργασία με αριθμητικά δεδομένα στην Python και βρίσκεται στον πυρήνα των επιστημονικών οικοσυστημάτων Python και PyData. Η βιβλιοθήκη NumPy περιέχει πολυδιάστατες δομές δεδομένων πινάκων. Η NumPy μπορεί να χρησιμοποιηθεί για την εκτέλεση μιας μεγάλης ποικιλίας μαθηματικών πράξεων σε πίνακες. Προσθέτει ισχυρές δομές δεδομένων στην Python που εγγυώνται αποτελεσματικούς υπολογισμούς με πίνακες και παρέχει μια τεράστια βιβλιοθήκη μαθηματικών συναρτήσεων υψηλού επιπέδου που λειτουργούν σε αυτούς τους πίνακες.

2.2 Διερευνητική Ανάλυση Δεδομένων

Η πρώτη επαφή με το αρχείο time_series_covid19_confirmed_global.csv ήταν να εισαχθεί με την χρήση Pandas στο Jupyter Notebook. Το αρχείο περιλαμβάνεται από 285 γραμμές και 895 στήλες. Παρατηρείται ότι οι ημερομηνίες αναγράφονται ως ονόματα στηλών συνεπώς πρέπει να υπάρξει κατάλληλη επεξεργασία του πολυδιάστατου πίνακα για να υπάρξουν ίδιοι τύποι δεδομένων στις στήλες. Αυτό γίνεται πάλι με την χρήση της Pandas που καθορίζει τις ημερομηνίες ως τύπο ημερομηνίας. Υπήρξε λοιπόν εξαγωγή όλων των στηλών, επιλέχθηκαν μόνο όλες οι ημερομηνίες και αποθηκεύτηκαν σε μια νέα μεταβλητή όπου αυτή θα είναι μία νέα στήλη δεδομένων τύπου ημερομηνίας. Αυτή η τεχνική επεξεργασίας του πίνακα είναι ιδιαίτερα σημαντική διότι δίνει την δυνατότητα να χρησιμοποιήσουμε τα δεδομένα για μεμονωμένες χώρες κάτι που δεν γινόταν στην αρχική μορφή του αρχείου. Οι χώρες σε σύνολο είναι 285 και έχει υπολογιστεί το σύνολο των επιβεβαιωμένων κρουσμάτων από 22 Ιανουαρίου 2020 έως 30 Ιουνίου 2022. Στην συνέχεια εξήχθησαν τα επιβεβαιωμένα κρούσματα του COVID-19 όπου είναι αθροιστικά για την Ελλάδα για κάθε ημερομηνία όπου στις 30 Ιουνίου 2022 είναι συνολικά 3.676.502. Έπειτα δημιουργήθηκε μια μικρή λίστα που περιέχει δεδομένα του COVID-19 για την Ιταλία, ΗΠΑ, Ισπανία, Γερμανία και Ελλάδα και στη συνέχεια παράχθηκε μια απλή οπτικοποίηση με τα συνολικά κρούσματα των παραπάνω χωρών. Ακολουθούν εικόνες των παραπάνω βημάτων επεξεργασίας που αναφέρθηκαν.

```
0                Afghanistan
1                    Albania
2                    Algeria
3                    Andorra
4                     Angola
                 ...
280       West Bank and Gaza
281     Winter Olympics 2022
282                    Yemen
283                   Zambia
284                 Zimbabwe
Name: Country/Region, Length: 285, dtype: object
```

Εικόνα 1 Johns Hopkins Github csv data

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	6/21/22	6/22/22	6/23/22	6/24/22	6/25/22	6/26/22	6/27/22	6/28/22	6/29/22	6/30/22
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	181808	181912	181987	182033	182072	182149	182228	182324	182403	182528
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	277663	277940	278211	278504	278793	279077	279077	279167	280298	280851
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	265993	266006	266015	266025	266030	266038	266049	266062	266073	266087
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	43449	43774	43774	43774	43774	43774	43774	43774	43774	43774
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	99761	99761	99761	99761	99761	99761	99761	101320	101320	101320

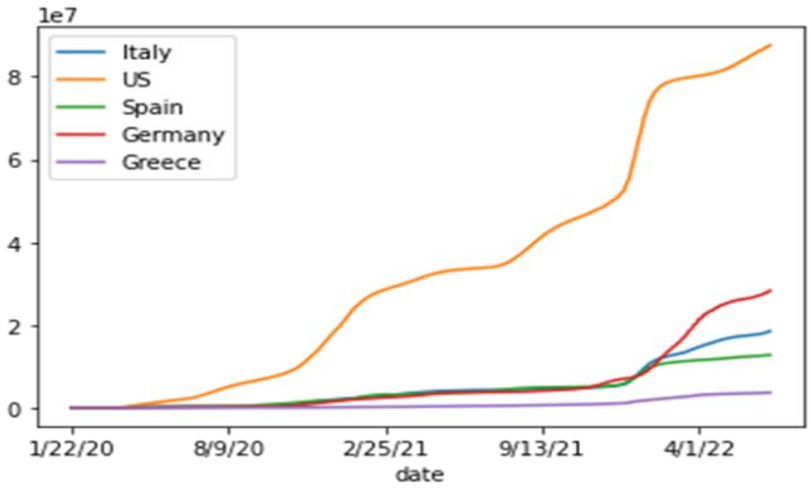
5 rows × 895 columns

Εικόνα 2 Country/Region Data Column

```
pd_raw[pd_raw['Country/Region']=='Greece'].iloc[:,4::].sum(axis=0)
```

```
1/22/20          0
1/23/20          0
1/24/20          0
1/25/20          0
1/26/20          0
             ...
6/26/22    3616874
6/27/22    3624556
6/28/22    3644889
6/29/22    3661004
6/30/22    3676502
Length: 891, dtype: int64
```

Εικόνα 3 Total COVID-19 cases in Greece



Εικόνα 4 Total COVID-19 cases on countries list

Παρατηρείται από το διάγραμμα ότι οι ΗΠΑ είχαν την πιο ραγδαία αύξηση κρουσμάτων COVID-19 σε σχέση με τις υπόλοιπες τέσσερις χώρες, στη συνέχεια ακολουθεί η Γερμανία, η Ιταλία, η Ισπανία και τέλος με τα λιγότερα κρούσματα η Ελλάδα. Στη συνέχεια παρουσιάζεται αναλυτικότερο διάγραμμα με τις πέντε αυτές χώρες. Προς το παρόν διενεργείται μια σύντομη διερευνητική ανάλυση δεδομένων για την αξιολόγηση-κατανόηση των δεδομένων αυτών με σκοπό την εξαγωγή γνώσεων ή βασικών χαρακτηριστικών. Συνεπώς προβήκαμε σε μια αρχική γραφική και μη γραφική ανάλυση. Είναι σημαντική η διερευνητική ανάλυση δεδομένων διότι επιτρέπει στον χρήστη να αναλύσει τα δεδομένα πριν καταλήξει σε οποιαδήποτε υπόθεση. Επίσης διασφαλίζει ότι τα αποτελέσματα που παράγονται είναι έγκυρα και εφαρμόζονται πάνω στους εκάστοτε στόχους που έχουν δημιουργηθεί. Η επόμενη κίνηση που εφαρμόστηκε ήταν η αλλαγή της ημερομηνίας σε κατάλληλο τύπο δεδομένων. Επειδή η ημερομηνία που υπήρχε μέχρι στιγμής δεν ήταν συμβατή με την ευρωπαϊκή μορφή δηλαδή yyyy-mm-dd έπρεπε να μετατραπεί σε αυτή την συγκεκριμένη μορφή ημερομηνίας. Αφού μετατράπηκε η ημερομηνία στην κατάλληλη μορφή διορθώθηκαν και τα ονόματα των στηλών και αφαιρέθηκαν οι συντεταγμένες διότι δεν χρησίμευαν για την ώρα στην ανάλυση που διεξάγεται μπορούμε να προχωρήσουμε την ανάλυση. Τέλος, καταλήξαμε σ’ έναν τελικό πίνακα που υπάρχουν μόνο οι απαραίτητες πληροφορίες που χρειάζονται για την ανάλυση με τις κατάλληλες αλλαγές στους τύπους δεδομένων των μεταβλητών που εξετάζονται. Συγκεκριμένα υπάρχει η στήλη date τύπου datetime, η στήλη state τύπου object, η στήλη country τύπου επίσης object και η στήλη confirmed όπου δηλώνει το άθροισμα των κρουσμάτων COVID-19 τύπου float. Ακολουθεί ένα παράδειγμα για το άθροισμα των κρουσμάτων της Ελλάδας για τις τελευταίες πέντε ημέρες.

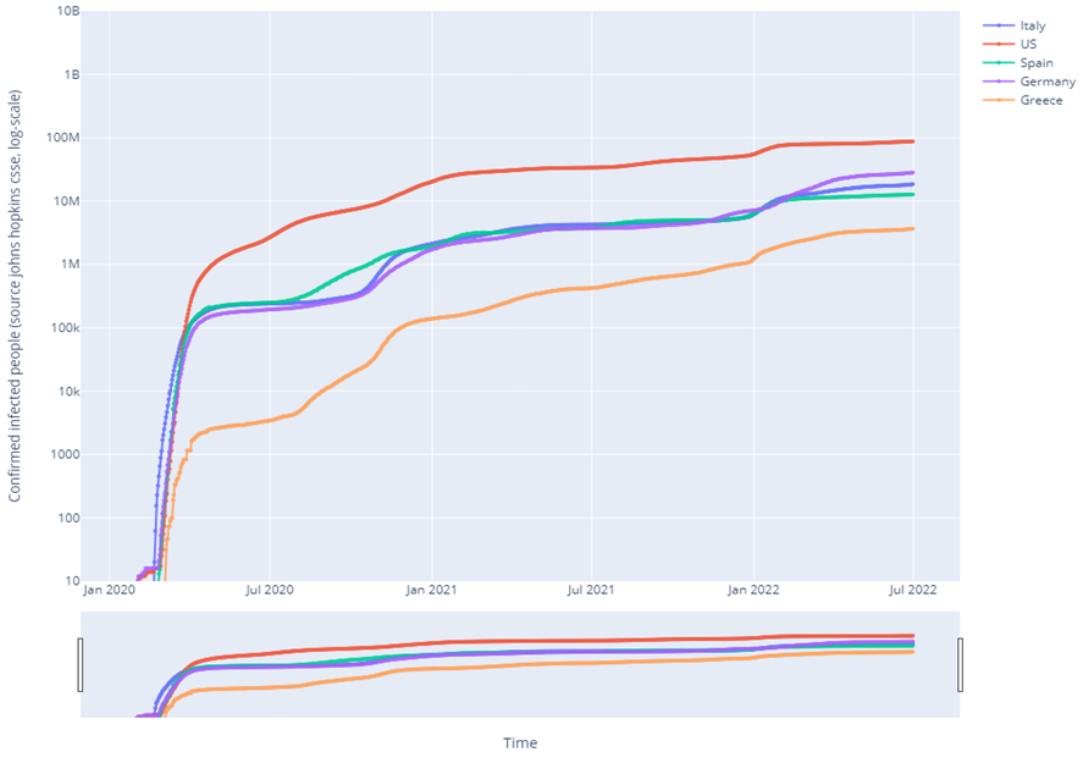
	date	state	country	confirmed
2022-06-26		no	Greece	3616874
2022-06-27		no	Greece	3624556
2022-06-28		no	Greece	3644889
2022-06-29		no	Greece	3661004
2022-06-30		no	Greece	3676502

Εικόνα 5 Covid-19 Total Cases of the Last 5 Days of June 2022 in Greece

2.3 Αναλυτική οπτικοποίηση δεδομένων-Δυναμικός πίνακας εργαλείων

Το επόμενο βήμα της ανάλυσης είναι η δημιουργία μιας αναλυτικής οπτικοποίησης του τελικού πίνακα με τις πέντε χώρες. Για την κατασκευή αυτού του γραφήματος με την χρήση της γλώσσας Python χρησιμοποιήθηκαν οι βιβλιοθήκες Matplotlib, Seaborn και Plotly. Η Matplotlib είναι μια βιβλιοθήκη για τη δημιουργία δισδιάστατων γραφημάτων πινάκων στην Python. Έχει τις ρίζες της στο MATLAB αλλά είναι ανεξάρτητο από αυτό. Συγκεκριμένα έχει χρησιμοποιηθεί το matplotlib.pyplot που είναι μια συλλογή από συναρτήσεις εντολών που κάνουν την Matplotlib να λειτουργεί όπως το MATLAB. Στη συνέχεια εισάγεται και η βιβλιοθήκη Seaborn που είναι μια βιβλιοθήκη οπτικοποίησης δεδομένων της Python που βασίζεται στην Matplotlib. Παρέχει μια διεπαφή υψηλού επιπέδου για τη σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφικών και γενικά βοηθάει στην κατανόηση και εξερεύνηση των δεδομένων. Τέλος, χρησιμοποιείται η βιβλιοθήκη Plotly που δημιουργεί διαδραστικά γραφήματα όπως για παράδειγμα γραμμικά γραφήματα, διαγράμματα διασποράς, γραφήματα εμβαδών, ραβδώσεων, γραφικών πλαισίων, ιστογραμμάτων, χάρτες θερμότητας, πολλαπλών αξόνων, πολικών γραφημάτων και γραφημάτων με φυσαλίδες. Συνεπώς έχει δημιουργηθεί ένα διαδραστικό διάγραμμα με την χρήση της Plotly διότι η Matplotlib παράγει μόνο στατικά διαγράμματα ενώ με την Plotly δίνεται η δυνατότητα με την χρήση του κέρσορα πάνω σε ένα σημείο του διαγράμματος μιας χώρας να εμφανίζεται το ακριβές σύνολο των κρουσμάτων σε μια συγκεκριμένη ημερομηνία στο σημείο αυτό. Συνεπώς, παρουσιάζεται αναλυτικά στο επόμενο διάγραμμα με την κατάλληλη ανάπτυξη κώδικα με την βοήθεια της βιβλιοθήκης Plotly η οπτικοποίηση των κρουσμάτων COVID-19 των πέντε χωρών.

2.4 Δυναμική οπτικοποίηση δεδομένων με την χρήση της βιβλιοθήκης Plotly



Εικόνα 6 Total COVID-19 Cases of the Countries List With Plotly Visualization Tool

Όπως παρατηρούμε το παραπάνω δυναμικό διάγραμμα διασποράς (scatter plot) που κάθε σημείο δεδομένων αναπαρίσταται ως σημείο δείκτη, η θέση του οποίου δίνεται από τις στήλες x και y απεικονίζει με ακρίβεια τα συνολικά κρούσματα COVID-19 για τις πέντε χώρες που επιλέχθηκαν και παράλληλα δίνετε η δυνατότητα να οριστεί ένα συγκεκριμένο χρονικό διάστημα (ρυθμιστικό εύρος) για να μελετηθούν τα νούμερα των κρουσμάτων μεταξύ αυτών των πέντε χωρών ή και μεμονωμένα ανάμεσα στον Ιανουάριο του 2020 και τον Ιούνιο του 2022. Αναλύοντας το διάγραμμα της επιδημιολογικής εξέλιξης του COVID-19 παρατηρείται ότι τείνει να διαμορφωθεί μια καμπύλη σχήματος S. Αυτή η καμπύλη ονομάζεται σιγμοειδής (sigmoid curve) ή λογιστική καμπύλη (logistic curve). Αυτού του είδους η καμπύλη συνήθως εμφανίζεται σε μοντέλα πληθυσμιακής μεταβολής-

αύξησης. Συνεπώς ένας άμεσος τρόπος μελέτης της επιδημίας του COVID-19 είναι η χρήση στατικών μοντέλων που βασίζονται σε καμπύλες S που εξαρτώνται από πολλές παραμέτρους. Αυτές οι παράμετροι μπορούν να υπολογιστούν χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων με ή χωρίς λειτουργία βάρους και το αντίστοιχο λογισμικό για προσαρμογή καμπύλης. Ένας άλλος τρόπος για την εύρεση των παραμέτρων είναι η επίλυση συστημάτων μη γραμμικών εξισώσεων και ο υπολογισμός των μέσων όρων των λύσεων. Επίσης η σιγμοειδής συνάρτηση χρησιμοποιείται στην μηχανική μάθηση και συγκεκριμένα στη λογιστική παλινδρόμηση και στα τεχνητά νευρωνικά δίκτυα. Η εξαγωγή συμπερασμάτων του συγκεκριμένου διαγράμματος που αναπαράχθηκε είναι σχετικά εμφανής. Οι ΗΠΑ βρίσκονται στην πρώτη θέση με συνολικά κρούσματα 87.623.593 στην δεύτερη θέση βρίσκεται η Γερμανία με 28.293.960 συνολικά κρούσματα ακολουθεί η Ιταλία με 18.523.111 συνολικά κρούσματα στην τέταρτη θέση είναι η Ισπανία με 12.734.038 συνολικά κρούσματα και στην τελευταία θέση βρίσκεται η Ελλάδα με 3.676.502 συνολικά κρούσματα. Οι συγκεκριμένοι αριθμοί ισχύουν μέχρι τις 30 Ιουνίου του 2022.

ΚΕΦΑΛΑΙΟ 3

Η ΤΕΧΝΟΛΟΓΙΑ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΠΡΟΒΛΕΨΗ ΤΟΥ COVID-19 – ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ARIMA

3.1 Αυτοπαλινδρομικό Μοντέλο Κινητού Μέσου Όρου (ARIMA)

Παρά το γεγονός ότι αναπτύχθηκαν μοντέλα για τη διάγνωση και την πρόγνωση του COVID-19, η έλλειψη των μοντέλων πρόβλεψης καθιστά δύσκολη την έγκαιρη ανίχνευση. Δεδομένου του αυξανόμενου φόρτου υποθέσεων, υπάρχει επείγουσα ανάγκη να αυξηθούν οι κλινικές δεξιότητες με την υποστήριξη μοντέλων μηχανικής μάθησης. Για παράδειγμα, η πρόβλεψη σοβαρών/κρίσιμων περιπτώσεων πριν εμφανιστεί το σύμπτωμα, μπορεί αποτελεσματικά να εξοικονομήσει ιατρικούς πόρους, εντοπίζοντας από τις πολλές ήπιες περιπτώσεις τις λίγες που θα εξελιχθούν σε κρίσιμη ασθένεια. Επιπλέον, η πρόβλεψη και η παρακολούθηση της συνεχιζόμενης αύξησης των κρουσμάτων COVID-19, μπορεί να βοηθήσει τους υπεύθυνους στα συστήματα λήψης αποφάσεων, αναπτύσσοντας στρατηγικές σχεδιασμού στο δημόσιο σύστημα υγείας για την αποφυγή θανάτων. Οι ερευνητικές πρακτικές που υπάρχουν σε αυτή την ενότητα περιλαμβάνουν την εφαρμογή της μηχανικής μάθησης και των στατιστικών μοντέλων με στόχο την παροχή υποστήριξης στην υγειονομική περίθαλψη και στον έγκαιρο εντοπισμό κρίσιμων και σοβαρών συμπτωμάτων.

Αν και τα εμβόλια μπορούν πλέον να χρησιμοποιηθούν για την πρόληψη του COVID-19, εξακολουθούν να υπάρχουν δεκάδες χιλιάδες νέα επιβεβαιωμένα κρούσματα καθημερινά στον κόσμο. Σε αυτήν την περίπτωση, η πρόληψη και η προετοιμασία των υπηρεσιών υγειονομικής περίθαλψης είναι κρίσιμης σημασίας. Η μοντελοποίηση και η μελλοντική πρόβλεψη του ημερήσιου αριθμού των νέων επιβεβαιωμένων κρουσμάτων και θανάτων είναι σημαντική διότι βοηθάει το σύστημα υγείας να παρέχει τις υπηρεσίες υγειονομικής περίθαλψης για τους πρόσφατα επιβεβαιωμένους ασθενείς και να επέμβει εγκαίρως για την αποκατάστασή τους. Συνεπώς, τα στατιστικά μοντέλα πρόβλεψης θα μπορούσαν να είναι ωφέλιμα για την πρόβλεψη και τον έλεγχο αυτής της παγκόσμιας πανδημίας.

Η μοντελοποίηση και η πρόβλεψη των καθημερινών επιβεβαιωμένων κρουσμάτων και θανάτων λόγω του COVID-19 μπορεί να βοηθήσει την παροχή υπηρεσιών υγειονομικής

περίθαλψης τροφοδοτώντας την με περισσότερες πληροφορίες σχετικά με τον αριθμό των καθημερινών νέων επιβεβαιωμένων κρουσμάτων και θανάτων, ώστε το υγειονομικό προσωπικό να μπορέσει να προετοιμαστεί εκ των προτέρων σχετικά με τις διαδικασίες οπλισμού, τροφοδοσίας και διασφάλισης τις ποιότητας των υπηρεσιών τους. Ως εκ τούτου, η κατασκευή χρήσιμων μοντέλων ARIMA (Αυτοπαλινδρομικό Μοντέλο Κινητών Μέσων – Auto Regressive Integrated Moving Average) για την πρόβλεψη των καθημερινών επιβεβαιωμένων κρουσμάτων και θανάτων του COVID-19 είναι πολύ σημαντική και μπορεί να βοηθήσει σε αυτό το πρόβλημα. Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσων είναι στοχαστικά μοντέλα τα οποία βοηθάνε να αναλύσουμε και να προβλέψουμε την εξέλιξη κάποιου μετρήσιμου μεγέθους. Ωστόσο, λόγω της υψηλής μεταβλητότητας των δεδομένων, η απλή χρήση αυτών των δεδομένων για την προσαρμογή των μοντέλων ARIMA δεν μπορεί να δημιουργήσει ένα τέλειο μοντέλο. Σε αυτό το κεφάλαιο χρησιμοποιήθηκε το μοντέλο ARIMA για την πρόβλεψη των καθημερινών νέων επιβεβαιωμένων κρουσμάτων και θανάτων του COVID-19 από τις 24 Φεβρουαρίου 2020 έως τις 30 Ιουνίου 2022. Στη συνέχεια, επιλέγοντας το καλύτερο μοντέλο ARIMA με βάση την μικρότερη τιμή του Akaike Information Criterion (AIC) όπου είναι ένας εκτιμητής του σφάλματος πρόβλεψης και ως εκ τούτου σχετικής ποιότητας στατιστικών μοντέλων για ένα σύνολο δεδομένων. Συνεπώς, λαμβάνοντας υπόψη μια συλλογή μοντέλων για τα δεδομένα, ο εκτιμητής AIC εκτιμά την ποιότητα κάθε μοντέλου, σε σχέση με κάθε ένα από τα άλλα μοντέλα. Τέλος, αφού έχει προσδιοριστεί το βέλτιστο μοντέλο ARIMA στη συνέχεια γίνεται μια πρόβλεψη πάνω στο τελικό γράφημα για τις επόμενες 30 ημέρες.

3.2 Θεωρητικό υπόβαθρο του μοντέλου ARIMA

Τα μοντέλα παλινδρόμησης (regression models) περιέχουν μια εξαρτημένη μεταβλητή ως συνάρτηση κάποιων άλλων ανεξάρτητων μεταβλητών. Όταν δουλεύουμε με γραμμικά μοντέλα παλινδρόμησης η συνάρτηση αυτή είναι γραμμική δηλαδή δίνεται ως γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών. Τα αυτοπαλινδρομούμενα μοντέλα (Auto Regressive models, AR) είναι μοντέλα γραμμικής παλινδρόμησης όπου έχουμε ως εξαρτημένη μεταβλητή την τυχαία μεταβλητή της χρονοσειράς σε μια χρονική στιγμή t, x_t και ως ανεξάρτητες μεταβλητές θεωρούμε την τυχαία μεταβλητή της χρονοσειράς σε προηγούμενους χρόνους x_{t-1}. Ο αριθμός των υστερήσεων λέγεται η τάξη (order) του αυτοπαλινδρομούμενου μοντέλου. Ένα αυτοπαλινδρομούμενο μοντέλο τάξης p συμβολίζεται ως AR(p). Το γενικό γραμμικό μοντέλο για την πρόβλεψη στάσιμης χρονοσειράς είναι το αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου (Auto Regressive Moving Average, ARMA) που περιέχει τον κινούμενο μέσο. Μια χρονοσειρά είναι στάσιμη (stationary) όταν δεν υπάρχει συστηματική μεταβολή στο μέσο (δεν περιέχει τάση), δεν υπάρχει συστηματική μεταβολή στη διακύμανση, και εάν οι περιοδικές μεταβολές έχουν αφαιρεθεί. Το μέρος του κινούμενου μέσου (MA) είναι τάξης q και το μοντέλο ορίζεται ως ARMA(p,q). Όταν η χρονοσειρά προσδιορίζεται από το ολοκληρωμένο αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου ή ολοκληρωμένο μικτό μοντέλο (Auto Regressive Integrated Moving Average Model, ARIMA) τότε το μοντέλο συμβολίζεται ως ARIMA(p,1,q) όπου ο δείκτης 1 δηλώνει ότι οι πρώτες διαφορές της χρονοσειράς προσδιορίζεται από το μοντέλο ARMA(p,q). Η προσαρμογή του μοντέλου AR και η πρόβλεψη βασίζεται πάνω σε κάποιες προϋποθέσεις που πρέπει να τηρούνται. Για την επιλογή της τάξης p του μοντέλου AR υπάρχουν δύο κριτήρια. Το πρώτο βασίζεται στη μερική αυτοσυσχέτιση (partial autocorrelation, PAC) που ορίζεται για κάθε υστέρηση t, που εδώ τη συμβολίζουμε p γιατί αναφέρεται στην τάξη του μοντέλου AR. Η συνάρτηση μερικής αυτοσυσχέτισης παρέχει μια καθαρή εικόνα της σειράς των εξαρτήσεων για τα μεμονωμένα στοιχεία των υστερήσεων. Επίσης, η τάξη του AR μοντέλου μπορεί να προσδιοριστεί και από διάφορα κριτήρια πληροφορίας που βασίζονται στη πιθανοφάνεια (likelihood) των δεδομένων με βάση αυτό το μοντέλο. Το πιο γνωστό κριτήριο που χρησιμοποιείται ευρέως για στατιστικά συμπεράσματα είναι το κριτήριο πληροφορίας του Akaike (Akaike Information Criterion, AIC) όπως αναφέρθηκε και στην εισαγωγή το

οποίο πήρε το όνομά του από τον Ιάπωνα Hirotugu Akaike, ο οποίος και το όρισε. Το κριτήριο ορίζεται ως

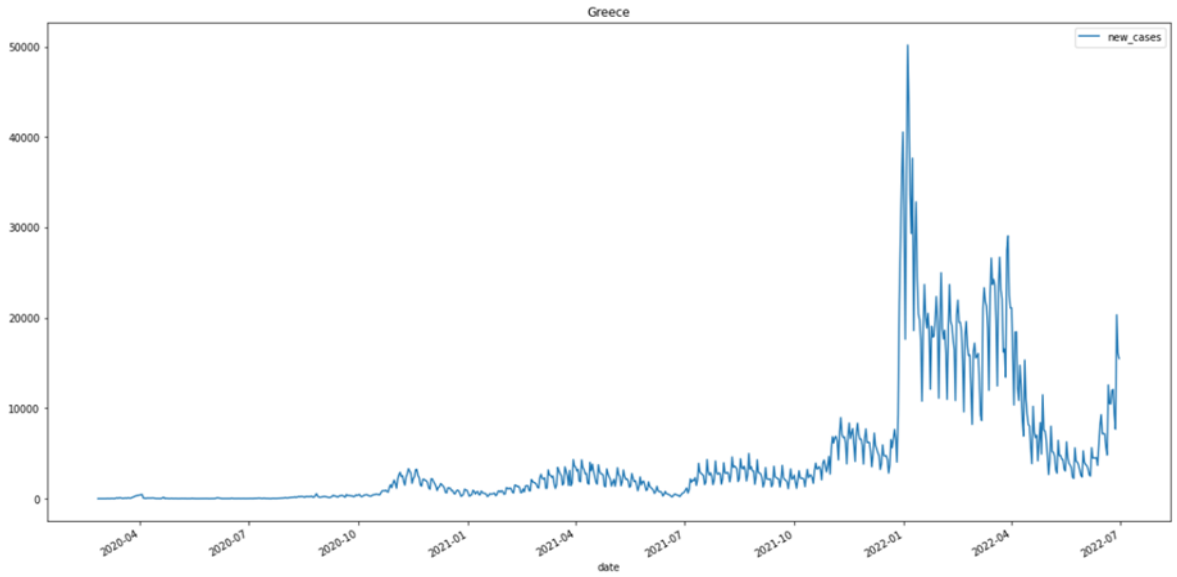
$$AIC(p) = \ln(s_z^2) + \frac{2p}{n}$$

όπου n είναι το μήκος της χρονοσειράς και s_z^2 η διασπορά των υπολοίπων. Παρατηρούμε από την εξίσωση ότι όσο μεγαλώνει η τάξη p του AR μοντέλου τα σφάλματα ή υπόλοιπα (residuals) γίνονται μικρότερα, εάν μικραίνει το s_z^2 για μεγάλες τάξεις το AR μοντέλο προσαρμόζεται σε διακυμάνσεις που σχετίζονται με το λευκό θόρυβο δηλαδή την μεταβολή των τιμών του μεγέθους που παρατηρείται είναι εντελώς τυχαία και η εξεύρεση τρόπων μείωσης ή απομάκρυνσης του θορύβου από τα δεδομένα αποτελεί σημαντικό μέρος της αναλυτικής διαδικασίας. Υπολογίζοντας το κριτήριο AIC για ένα μεγάλο αριθμό τάξεων μοντέλου επιλέγουμε εκείνη την τάξη p που δίνει την ελάχιστη τιμή του AIC. Επιλέγοντας την μικρότερη τιμή του AIC τότε έχουμε το βέλτιστο μοντέλο σε σχέση με τα υπόλοιπα μοντέλα.

3.3 Μεθοδολογία και Αποτελέσματα εφαρμογής του μοντέλου ARIMA

Όπως παρουσιάστηκε στα προηγούμενα κεφάλαια, μοντελοποιώντας την μελλοντική πρόβλεψη του ημερήσιου αριθμού νέων επιβεβαιωμένων κρουσμάτων και θανάτων είναι ένα σημαντικό βήμα που μπορεί να βοηθήσει το υγειονομικό σύστημα να παρέχει τις υπηρεσίες περίθαλψης για τα πρόσφατα επιβεβαιωμένα κρούσματα. Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή τη μελέτη προήλθε από την επιστημονική διαδικτυακή ομάδα Our World in Data (OWID) του Πανεπιστημίου της Οξφόρδης. Το σύνολο δεδομένων περιλαμβάνεται από πολλές χρονοσειρές με πληροφορίες από τα καθημερινά επιβεβαιωμένα κρούσματα και τους θανάτους του COVID-19 παγκοσμίως από τις 24 Φεβρουαρίου 2020 έως τις 30 Ιουνίου 2022. Το σύνολο δεδομένων αποτελείται από 197.930 σειρές και 67 στήλες. Το πρώτο πράγμα που πρέπει να γίνεται στην αρχή μετά την εισαγωγή του συνόλου δεδομένων είναι ένας καθαρισμός των δεδομένων έτσι ώστε να προχωρήσει η μελέτη. Επίσης, στη συνέχεια έγιναν κάποιες αλλαγές στο τύπο των δεδομένων για την ευκολία της ανάλυσης. Επιπρόσθετα, το σύνολο δεδομένων περιέχει πληροφορίες για 6 ηπείρους και 238 χώρες συμπεριλαμβανομένου και της Ελλάδας όπου θα μελετηθεί μεμονωμένα και θα εφαρμοστεί το μοντέλο ARIMA. Για την μελέτη της πρόβλεψης αυτής επιλέχθηκαν οι στήλες date και new_cases και στη συνέχεια παράχθηκε

μια οπτικοποίηση για αυτές τις δύο στήλες. Η στήλη new_cases αναφέρεται στα νέα επιβεβαιωμένα κρούσματα του COVID-19. Παρακάτω φαίνεται το σχετικό χρονοδιάγραμμα ή η χρονοσειρά των νέων επιβεβαιωμένων κρουσμάτων του COVID-19 για την Ελλάδα.



Εικόνα 7 New COVID-19 cases in Greece

Όπως παρατηρείται, υπάρχουν μερικές μεγάλες διακυμάνσεις στη στήλη new_cases. Θα γίνει μια προσπάθεια πρόβλεψης για αυτά τα κύματα. Πρωτίστως, θα πρέπει να διαγραφούν δεδομένα που λείπουν χρησιμοποιώντας την κατάλληλη μέθοδο στην Python με την χρήση της βιβλιοθήκης Pandas. Στη συνέχεια, πρέπει να ελεγχθεί κατά πόσο είναι στάσιμη ή μη στάσιμη η χρονοσειρά. Τι είναι όμως μια στάσιμη χρονοσειρά; Μια στάσιμη χρονοσειρά είναι εκείνη της οποίας οι στατιστικές ιδιότητες όπως η μέση τιμή, η διακύμανση, η συνδιακύμανση δεν ποικίλλουν με το χρόνο ή αυτές οι ιδιότητες στατιστικών στοιχείων δεν είναι συνάρτηση του χρόνου. Με άλλα λόγια, η στασιμότητα στις χρονοσειρές σημαίνει επίσης χρονοσειρά χωρίς τάση ή εποχιακά στοιχεία. Οι στάσιμες χρονοσειρές είναι πιο εύκολο για τα στατιστικά μοντέλα να προβλέψουν αποτελεσματικά και με ακρίβεια. Οι τύποι στάσιμων χρονοσειρών είναι οι εξής:

1. Αυστηρή Στασιμότητα – Ικανοποιεί τον μαθηματικό ορισμό μιας στατικής διαδικασίας και ο μέσος όρος, η διακύμανση και η συνδιακύμανση δεν είναι συνάρτηση του χρόνου.
2. Εποχική Στασιμότητα – Χρονοσειρά που εμφανίζει εποχικότητα.
3. Στασιμότητα με Τάση – Χρονοσειρά που παρουσιάζει τάση.

Αξίζει να σημειωθεί ότι μόλις αφαιρεθεί η εποχικότητα και η τάση, οι χρονοσειρές θα είναι αυστηρά σταθερές. Πως θα ελεγχθεί όμως εάν η χρονοσειρά είναι στάσιμη ή όχι; Ένας τρόπος είναι με κατάλληλες οπτικοποιήσεις. Οι πιο βασικές μέθοδοι για την ανίχνευση σταθερότητας βασίζονται στη γραφική παράσταση των δεδομένων και στον οπτικό έλεγχο για τάσεις και εποχιακά στοιχεία. Ένας άλλος τρόπος είναι με στατιστικές δοκιμές όπως το Augmented Dickey-Fuller Test. Οι στατιστικές δοκιμές κάνουν ισχυρές υποθέσεις για τα δεδομένα. Μπορούν να χρησιμοποιηθούν μόνο για να ενημερώσουν τον βαθμό στον οποίο μια μηδενική υπόθεση μπορεί να απορριφθεί ή να μην απορριφθεί. Το αποτέλεσμα πρέπει να ερμηνευθεί για να έχει νόημα ένα δεδομένο πρόβλημα. Ωστόσο, παρέχουν έναν γρήγορο έλεγχο και επιβεβαιωτικά στοιχεία ότι η χρονοσειρά είναι στάσιμη ή μη. Η δοκιμή Augmented Dickey-Fuller είναι ένας τύπος στατιστικής δοκιμής που ονομάζεται δοκιμή μοναδιαίας ρίζας. Στη θεωρία πιθανοτήτων και τη στατιστική, μια μοναδιαία ρίζα είναι ένα χαρακτηριστικό ορισμένων στοχαστικών διεργασιών (όπως τυχαίους περιπάτους) που μπορεί να προκαλέσει προβλήματα στα στατιστικά συμπεράσματα που περιλαμβάνουν μοντέλα χρονοσειρών. Με έναν απλό όρο, η μοναδιαία ρίζα είναι μη στάσιμη αλλά δεν έχει πάντα μια συνιστώσα τάσης. Η δοκιμή ADF διεξάγεται με τις ακόλουθες υποθέσεις.

1. Μηδενική υπόθεση (H0): Η σειρά είναι μη στάσιμη ή η σειρά έχει μοναδιαία ρίζα.
2. Εναλλακτική υπόθεση (H1): Η σειρά είναι στάσιμη ή η σειρά δεν έχει μοναδιαία ρίζα.

Εάν η μηδενική υπόθεση δεν απορριφθεί, αυτό το τεστ μπορεί να αποδείξει ότι η σειρά είναι μη στάσιμη.

Προϋποθέσεις απόρριψης μηδενικής υπόθεσης (H0)

Εάν η στατιστική δοκιμή < Κρίσιμη τιμή και η τιμή $p < 0,05$ – τότε απορρίπτεται η μηδενική υπόθεση (H0), δηλαδή, η χρονοσειρά δεν έχει μοναδιαία ρίζα, που σημαίνει ότι είναι στάσιμη. Παρακάτω εφαρμόστηκε ο έλεγχος Augmented Dickey-Fuller με την χρήση της Python για την στήλη new_cases και τα αποτελέσματα είναι τα εξής:

```
Test Statistic                -1.9934
p-value                        0.2895
#Lags Used                    19.0000
Number of Observations Used  812.0000
dtype: float64
```

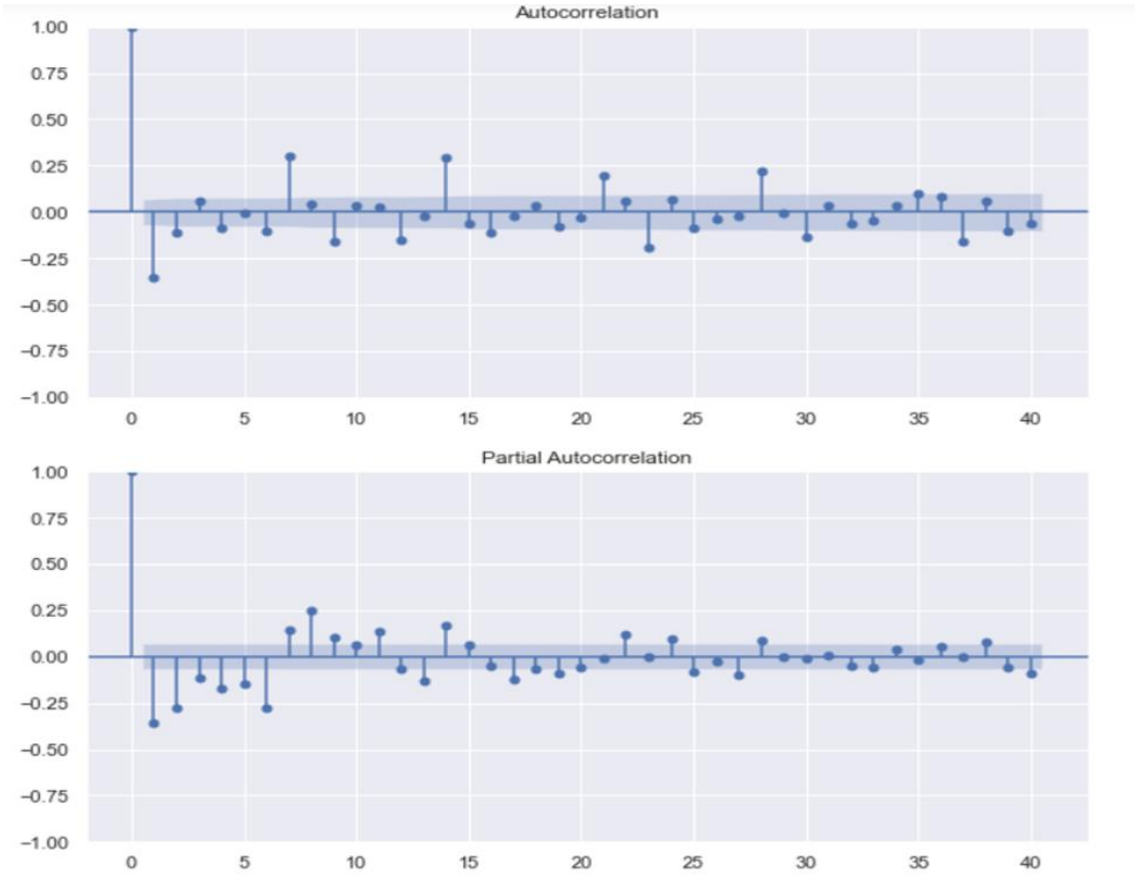
Εικόνα 8 Augmented Dickey-Fuller Test1

Παρατηρούμε ότι η τιμή της στατιστικής δοκιμής είναι -1.9934 πού είναι μικρότερη από την τιμή του p-value αλλά το p-value δεν είναι μικρότερο του 0.05 συνεπώς δεν απορρίπτεται η μηδενική υπόθεση H0 συνεπώς η χρονοσειρά είναι μη στάσιμη. Για να γίνει στάσιμη μια χρονοσειρά χρησιμοποιούνται πολλές τεχνικές μια από αυτή που θα εφαρμοστεί είναι να πάρουμε τον λογάριθμο των τιμών αυτών και να βρούμε τις διαδοχικές διαφορές μεταξύ τους. Αφού έγινε η παραπάνω διαδικασία με τον λογάριθμο θα γίνει πάλι ο έλεγχος Augmented Dickey-Fuller για να δούμε εάν έγινε η χρονοσειρά στάσιμη. Τα αποτελέσματα του ελέγχου είναι τα ακόλουθα:

```
Test Statistic                -7.3039
p-value                        0.0000
#Lags Used                    21.0000
Number of Observations Used  809.0000
dtype: float64
```

Εικόνα 9 Augmented Dickey-Fuller Test2

Όπως βλέπουμε η τιμή της στατιστικής δοκιμής είναι -7.3039 που είναι μικρότερη από την τιμή του p-value που ισούται με 0 και αυτό είναι μικρότερο από το 0.05 συνεπώς απορρίπτεται η μηδενική υπόθεση H0 άρα η χρονοσειρά μετατράπηκε σε στάσιμη. Το επόμενο βήμα της ανάλυσης είναι να δημιουργηθούν τα διαγράμματα της αυτοσυσχέτισης (Autocorrelation Function, ACF) και της μερικής αυτοσυσχέτισης (Partial Autocorrelation Function, PACF). Αυτό είναι το σημαντικότερο βήμα στην υλοποίηση του μοντέλου ARIMA διότι τα διαγράμματα αυτά χρησιμοποιούνται για τον προσδιορισμό των παραμέτρων εισόδου για το μοντέλο ARIMA. Παρακάτω παρουσιάζονται αυτά τα δύο διαγράμματα.



Εικόνα 10 ACF and PACF Plots

Στα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης παρατηρείται ότι η χρονοσειρά των πρώτων διαφορών που περιγράφει τα νέα κρούσματα του COVID-19 δεν παρουσιάζει τάσεις και η αυτοσυσχέτισή της τίνει γρήγορα στο 0. Το γεγονός ότι κάποιες αυτοσυσχετίσεις για κάποιες υστερήσεις είναι στατιστικά σημαντικές σημαίνει ότι η χρονοσειρά των διαφορών δεν είναι λευκός θόρυβος αλλά στάσιμη χρονοσειρά με ασθενείς αυτοσυσχετίσεις. Επίσης παρατηρούμε ότι το p είναι τουλάχιστον 1 και το πολύ 3 διότι μετά την $3^{η}$ υστέρηση δεν έχουμε σημαντικές στατιστικές σχέσεις ανάμεσα στις παρατηρήσεις μας και τις υστερήσεις και το ίδιο για το q που κυμαίνεται από 1 έως 6.

3.4 Εύρεση του βέλτιστου μοντέλου με την χρήση της διαδικασίας Auto-ARIMA

Στη συνέχεια της ανάλυσης χωρίσθηκαν τα δεδομένα σε train και test data έτσι ώστε να αποτραπεί η υπερβολική προσαρμογή του μοντέλου μας (overfitting), δηλαδή το μοντέλο γίνεται πολύ καλό στην ταξινόμηση των δειγμάτων στο σύνολο εκπαίδευσης. Τα σετ εκπαίδευσης χρησιμοποιούνται συνήθως για την εκτίμηση διαφορετικών παραμέτρων ή για τη σύγκριση διαφορετικών επιδόσεων των μοντέλων. Το σύνολο δεδομένων δοκιμών (test data) χρησιμοποιείται μετά την ολοκλήρωση της εκπαίδευσης. Τα δεδομένα εκπαίδευσης και δοκιμής (training and test data) συγκρίνονται για να ελεγχθεί ότι το τελικό μοντέλο λειτουργεί σωστά. Δεν υπάρχει καθορισμένη κατευθυντήρια γραμμή ή μέτρηση για τον τρόπο διαχωρισμού των δεδομένων. Μπορεί να εξαρτάται από το μέγεθος της αρχικής δεξαμενής δεδομένων ή τον αριθμό των προβλέψεων σε ένα μοντέλο πρόβλεψης. Η πιο κοινή αναλογία διαχωρισμού των δεδομένων είναι 80:20 δηλαδή το 80% των δεδομένων θα είναι για εκπαίδευση και το υπόλοιπο 20 % για τις δοκιμές. Στη ανάλυση την οποία μελετάμε χωρίσθηκαν 652 σειρές δεδομένων για την εκπαίδευση και οι υπόλοιπες 180 (δηλαδή οι τελευταίοι 6 μήνες 1 Ιανουαρίου 2022 μέχρι 30 Ιουνίου 2022) για την δοκιμή του μοντέλου ARIMA. Αφού χωρίσθηκαν τα δεδομένα σε train και test θα χρησιμοποιηθεί η συνάρτηση auto-ARIMA. Η διαδικασία auto-ARIMA επιδιώκει να εντοπίσει τις βέλτιστες παραμέτρους για ένα μοντέλο ARIMA. Το Auto-ARIMA λειτουργεί πραγματοποιώντας δοκιμές διαφοροποίησης και στη συνέχεια προσαρμόζοντας μοντέλα εντός των ορίων των καθορισμένων παραμέτρων εισαγωγής start_p, start_q, max_p, max_q. Προκειμένου να βρεθεί το βέλτιστο μοντέλο, το auto-ARIMA βελτιστοποιεί για ένα δεδομένο κριτήριο πληροφοριών, ένα από τα AIC, BIC, HQIC, (Akaike Information Criterion, Bayesian Information Criterion, Hannan-Quinn Information) και επιστρέφει το μοντέλο ARIMA που ελαχιστοποιεί την τιμή και στα τρία αυτά κριτήρια πληροφορίας. Η παρακάτω εικόνα δείχνει τα αποτελέσματα της εφαρμογής της συνάρτησης auto-ARIMA.

```
Best model:  ARIMA(3,0,5)(2,0,0)[12]
Total fit time: 167.024 seconds
```

SARIMAX Results

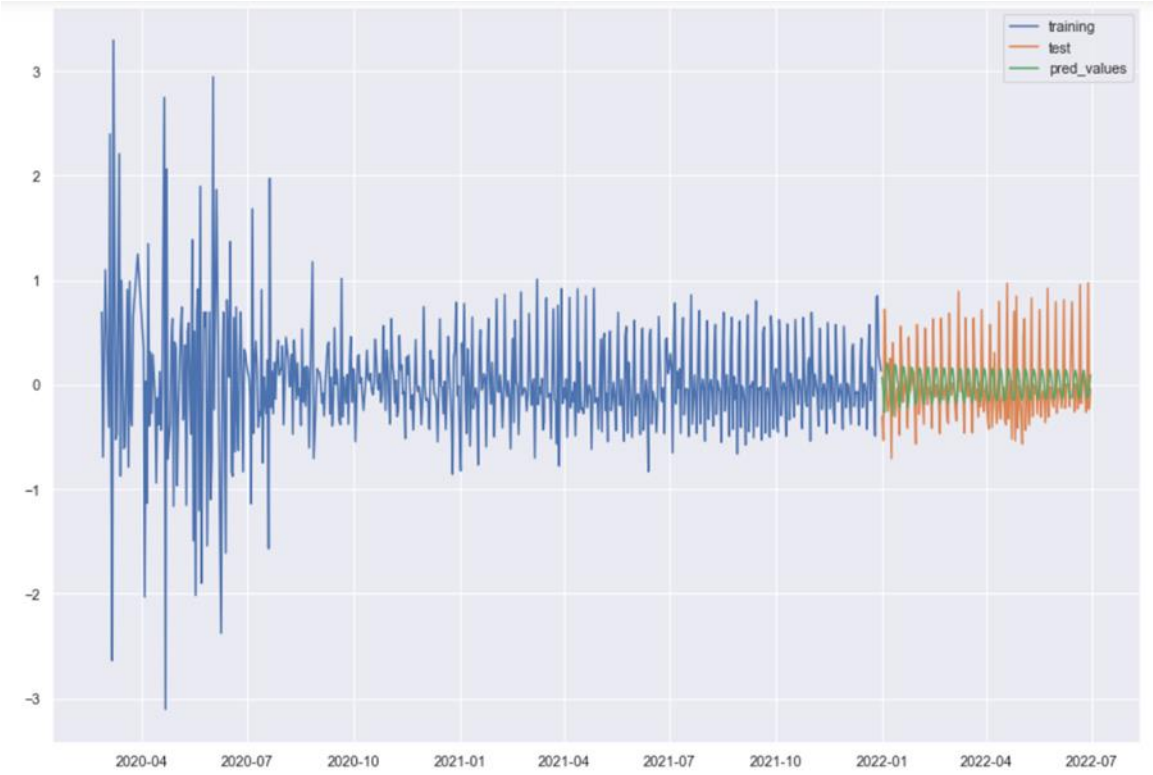
Dep. Variable:	y	**No. Observations:**	831
Model:	SARIMAX(3, 0, 5)x(2, 0, [], 12)	**Log Likelihood**	-456.006
Date:	Mon, 05 Sep 2022	**AIC**	934.012
Time:	11:36:40	**BIC**	985.961
Sample:	0	**HQIC**	953.932

Εικόνα 11 SARIMAX Model

Παρατηρείται ότι το βέλτιστο μοντέλο για το σύνολο δεδομένων που εξετάζουμε είναι το ARIMA (3,0,5) x (2,0,0)[12] και συγκεκριμένα το SARIMAX (3,0,5) x (2,0,0,12). Ο Εποχιακός Αυτόματος Παλινδρομικός Ενσωματωμένος Κινητός Μέσος (Seasonal Auto-Regressive Integrated Moving Average) με εξωγενείς παράγοντες, ή SARIMAX, είναι μια επέκταση της κατηγορίας μοντέλων ARIMA. Διαισθητικά, τα μοντέλα ARIMA συνθέτουν δύο μέρη: τον αυτοπαλινδρομικό όρο (AR) και τον όρο κινούμενου μέσου όρου (MA) όπως έχει αναφερθεί προηγουμένως. Ο πρώτος βλέπει την τιμή ταυτόχρονα ως ένα σταθμισμένο άθροισμα προηγούμενων τιμών και ο τελευταίος μοντελοποιεί την ίδια τιμή και ως σταθμισμένο άθροισμα αλλά προηγούμενων υπολειμμάτων (συνεισφέρει στην αποσύνθεση χρονοσειρών). Υπάρχει επίσης ένας ολοκληρωμένος όρος (I) για τη διαφορά των χρονοσειρών. Μια συντομογραφία για τα μοντέλα SARIMA είναι η εξής:

$$SARIMA\ (p,d,q)x\ (P,D,Q,S)$$

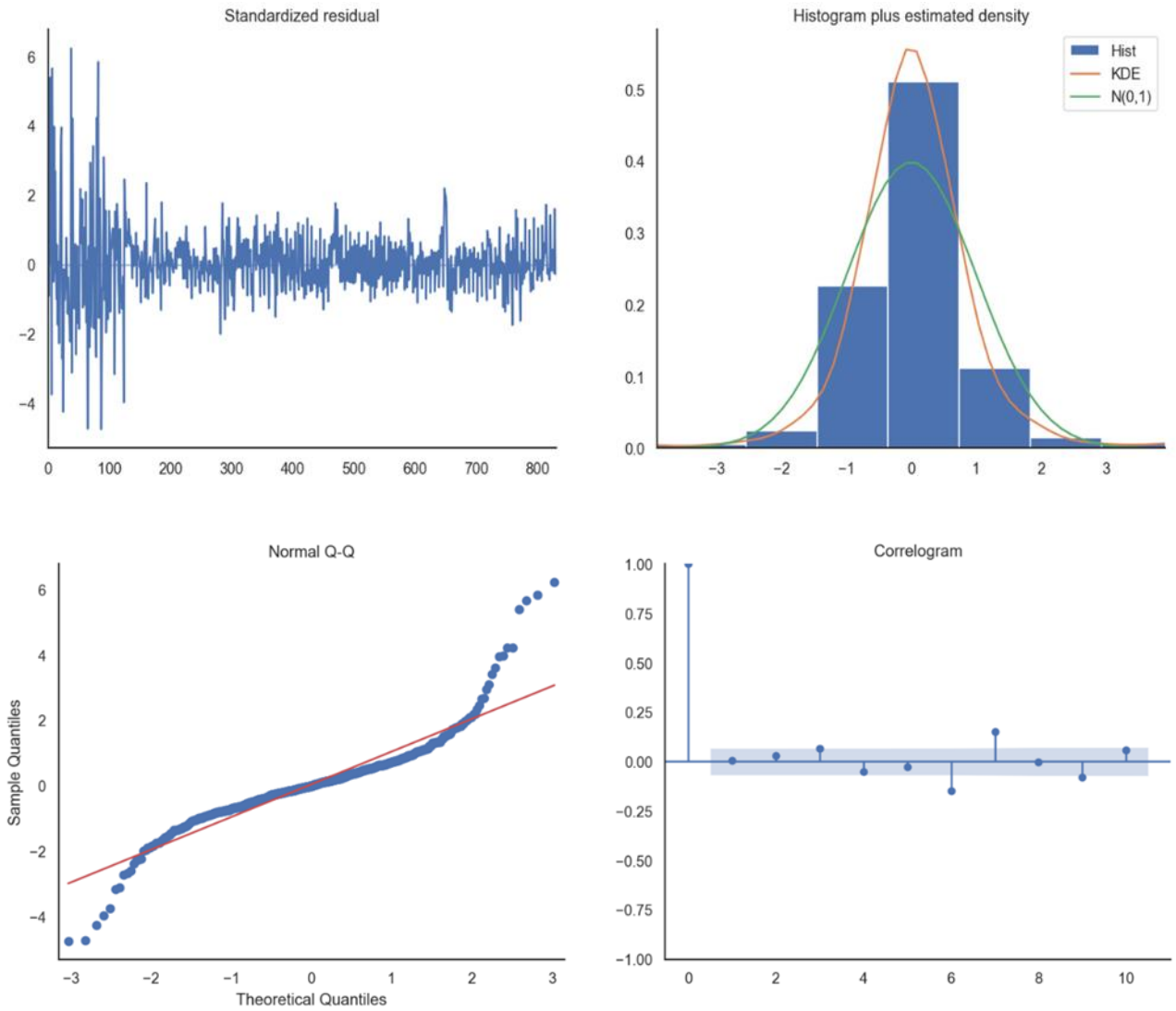
Όπου p = μη εποχική αυτοπαλινδρομική σειρά (AR), d = μη εποχική διαφοροποίηση, q= σειρά μη εποχικού κινητού μέσου όρου (MA), P = εποχική αυτοπαλινδρομική σειρά AR, D = εποχική διαφοροποίηση, Q = εποχικός κινητός μέσος (MA) και S = μήκος επαναλαμβανόμενου εποχιακού μοτίβου. Προσθέτοντας αυτά τα εποχιακά στοιχεία AR και εποχιακά στοιχεία MA, η SARIMA λύνει το πρόβλημα εποχικότητας. Συνεπώς το AR δηλαδή το p=3, το d=0 και το MA δηλαδή το q=5, το P=2, το D=0, το Q=0 και το S=12 (12 μήνες). Στο παραπάνω πίνακα αποτελεσμάτων η συνάρτηση πιθανοφάνειας ισούται με -456.006 και η ελάχιστη τιμή AIC που βρέθηκε ισούται με 934.012 για τις 831 παρατηρήσεις όπου εξετάσθηκαν. Έπειτα, θα επακολουθήσει η πρόβλεψη μέσω κατάλληλης οπτικοποίησης των νέων κρουσμάτων του COVID-19 για την περίοδο 1 Ιανουαρίου 2022 μέχρι 31 Ιουλίου 2022.



Εικόνα 12 SARIMAX Prediction Plot

3.5 Επικύρωση μοντέλου – Υπολειμματικά Διαγνωστικά

Για να προσδιορίσουμε την καλή προσαρμογή του μοντέλου, μπορούμε να εξετάσουμε τα υπολείμματά του χρησιμοποιώντας την τυπική υπόθεση ότι θα πρέπει κανονικά να κατανέμονται γύρω στο 0. Μπορούμε να το ελέγξουμε αυτό κοιτάζοντας τα διάφορα διαγράμματα που δείχνουν την κατανομή των υπολειμμάτων.



Εικόνα 13 Normality Test Plots

Στα διαγράμματα, τα υπολείμματα φαίνεται να κατανέμονται κανονικά γύρω από το 0 που είναι η συνθήκη που χρειαζόμαστε. Στο πρώτο διάγραμμα εμφανίζονται τα τυποποιημένα υπολείμματα. Στο δεύτερο διάγραμμα μπορούμε να δούμε ότι η κατανομή έχει μια Gaussian μορφή, αλλά είναι πιο μυτερή, δείχνοντας μια εκθετική κατανομή με κάποια ασυμμετρία. Εάν η γραφική παράσταση έδειχνε μια κατανομή που ήταν σαφώς μη Gaussian, θα υποδηλώνει ότι οι υποθέσεις που έγιναν από τη διαδικασία μοντελοποίησης ήταν ίσως λανθασμένες και ότι μπορεί να απαιτείται διαφορετική μέθοδος μοντελοποίησης. Στο τρίτο διάγραμμα είναι η γραφική παράσταση Q-Q, όπου αυτή η γραφική παράσταση ποσοτήτων, συγκρίνει δύο κατανομές και μπορεί να χρησιμοποιηθεί για να δει πόσο παρόμοιες ή διαφορετικές τυχαίνει να είναι. Η γραφική παράσταση Q-Q μπορεί να χρησιμοποιηθεί για τον γρήγορο έλεγχο της κανονικότητας της κατανομής των υπολειπόμενων σφαλμάτων. Οι τιμές ταξινομούνται και συγκρίνονται με μια εξιδανικευμένη κατανομή Gauss. Η σύγκριση εμφανίζεται ως διάγραμμα διασποράς όπου η αντιστοίχιση μεταξύ των δύο κατανομών εμφανίζεται ως διαγώνια γραμμή από κάτω αριστερά προς τα πάνω δεξιά της γραφικής παράστασης. Η γραφική παράσταση Q-Q που παράχθηκε δείχνει ότι η κατανομή είναι φαινομενικά κανονική με μερικές προσκρούσεις και ακραίες τιμές. Στο τέταρτο και τελευταίο διάγραμμα είναι η οπτικοποίηση της αυτοσυσχέτισης για τα υπολειπόμενα σφάλματα. Ο άξονας x δείχνει την υστέρηση και ο άξονας y δείχνει τη συσχέτιση μεταξύ μιας παρατήρησης και της μεταβλητής υστέρησης, όπου οι τιμές συσχέτισης είναι μεταξύ -1 και 1 για αρνητικές και θετικές συσχετίσεις αντίστοιχα. Από το διάγραμμα αυτό δεν παρατηρείται μια προφανή τάση αυτοσυσχέτισης σε όλο το διάγραμμα. Μπορεί να υπάρχει κάποια θετική αυτοσυσχέτιση που αξίζει περαιτέρω έρευνας στην υστέρηση 0 που φαίνεται σημαντική. Τέλος, ένα βασικό βήμα σε κάθε μοντέλο μηχανικής μάθησης είναι η αξιολόγηση της ακρίβειας του μοντέλου. Για αυτό τον λόγο υπολογίστηκαν οι μετρήσεις μέσου τετραγωνικού σφάλματος, μέσου απόλυτου σφάλματος, η ρίζα του μέσου τετραγωνικού σφάλματος και η μέτρηση R-Squared ή συντελεστής προσδιορισμού όπου χρησιμοποιούνται όλες αυτές οι μετρήσεις για την αξιολόγηση της απόδοσης του μοντέλου στην ανάλυση παλινδρόμησης.

Το μέσο απόλυτο ποσοστό σφάλματος (Mean Absolute Percentage Error - MAPE) είναι ένας από τους πιο συχνούς δείκτες απόδοσης ή ένας βασικός δείκτης απόδοσης (Key Performance indicator – KPI) για τη μέτρηση της ακρίβειας πρόβλεψης. Το μέσο απόλυτο

ποσοστό σφάλματος είναι το άθροισμα των μεμονωμένων απόλυτων σφαλμάτων διαιρούμενο με τη ζήτηση. Με άλλα λόγια, είναι ο μέσος όρος των ποσοστών σφαλμάτων.

$$MAPE = \frac{1}{n}\sum \frac{|e^t|}{d_t}$$

Πολλοί αλγόριθμοι (ειδικά για μηχανική μάθηση) βασίζονται στο Μέσο Τετράγωνο Σφάλμα (Mean Squared Error - MSE), το οποίο σχετίζεται άμεσα με το RMSE το οποίο θα αναφερθεί παρακάτω. Πολλοί αλγόριθμοι χρησιμοποιούν τον MSE καθώς είναι γρήγορος στον υπολογισμό του και ευκολότερος στον χειρισμό από τον RMSE. Ο εκτιμητής αυτός μετρά τον μέσο όρο των τετραγώνων των σφαλμάτων δηλαδή τη μέση τετραγωνική διαφορά μεταξύ των εκτιμώμενων τιμών και της πραγματικής αξίας. Το MSE είναι μια συνάρτηση κινδύνου στην ουσία , που αντιστοιχεί στην αναμενόμενη τιμή του τετραγώνου της ζημίας σφάλματος.

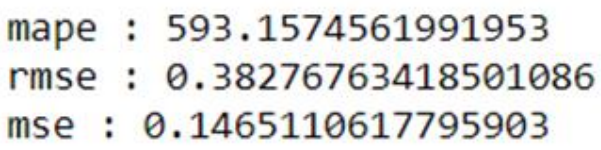
$$MSE = \frac{1}{n}\sum e_t^2$$

Η ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error - RMSE) ορίζεται ως η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Είναι ένα μέτρο ακρίβειας, για τη σύγκριση σφαλμάτων πρόβλεψης διαφορετικών μοντέλων για ένα συγκεκριμένο σύνολο δεδομένων. Όσο χαμηλότερο είναι το RMSE, τόσο καλύτερα το μοντέλο ταιριάζει σε ένα σύνολο δεδομένων. Το RMSE χρησιμοποιείται συχνότερα επειδή μετριέται στις ίδιες μονάδες με την εξαρτημένη μεταβλητή. Αντιθέτως, το MSE μετριέται σε μονάδες που είναι το τετράγωνο της εξαρτημένης μεταβλητής.

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

Τέλος, τα παρακάτω αποτελέσματα MAPE, RMSE, MSE υποδεικνύουν κατά πόσο ήταν ακριβής η πρόβλεψη που εφαρμόστηκε προηγουμένως με το μοντέλο ARIMA. Το MAPE βγήκε 593.15 που σημαίνει ότι τα σφάλματα είναι "πολύ μεγαλύτερα" από τις πραγματικές

τιμές (π.χ. η πραγματική τιμή είναι 1, εμείς προβλέπουμε 3, άρα το MAPE είναι 200%). Ωστόσο, το MAPE έχει πολλές παγίδες ως μέτρο σφαλμάτων διότι παράγει άπειρες ή απροσδιόριστες τιμές όταν οι πραγματικές τιμές είναι μηδέν ή κοντά στο μηδέν, επομένως συχνά δεν συνιστάται ως η καλύτερη επιλογή για την αξιολόγηση της ακρίβειας του μοντέλου. Το RMSE είναι 0.38. Συνήθως οι αποδεκτές τιμές του RMSE κυμαίνονται μεταξύ 0.2 και 0.5 που δείχνουν ότι το μοντέλο μπορεί να προβλέψει σχετικά με ακρίβεια τα δεδομένα. Το RMSE είναι πάντα μη αρνητικό και όσο πιο κοντά βρίσκεται στο 0 τόσο καλύτερη είναι η πρόβλεψη του εκάστοτε μοντέλου όταν συγκρίνουμε μοντέλα πρόβλεψης. Το MSE είναι 0.14 που υποδηλώνει ότι υπάρχει καλή προσαρμογή του μοντέλου αφού όσο πιο κοντά είναι η τιμή του MSE στο 0, τόσο πιο ακριβές είναι το μοντέλο. Ωστόσο, δεν υπάρχει «καλή» τιμή για το MSE. Είναι μια απόλυτη τιμή που είναι μοναδική για κάθε σύνολο δεδομένων και μπορεί να χρησιμοποιηθεί μόνο για να δείξει εάν το μοντέλο έχει γίνει περισσότερο ή λιγότερο ακριβές από μια προηγούμενη εκτέλεση.



Εικόνα 14 Forecast Accuracy Measures

ΚΕΦΑΛΑΙΟ 4

ΣΤΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ ΤΩΝ ΔΙΑΘΕΣΙΜΩΝ ΕΜΒΟΛΙΩΝ ΕΝΑΝΤΙΑ ΣΤΟΝ COVID-19

4.1 Η ανάπτυξη των εμβολίων κατά του COVID-19

Σε αυτή την ενότητα εφαρμόστηκε μια διερευνητική ανάλυση δεδομένων για τα διαθέσιμα εμβόλια κατά του COVID-19 με δεδομένα από χρονοσειρές με τις κατάλληλες οπτικοποιήσεις. Βασικός στόχος αυτής της ανάλυσης είναι να παρουσιάσει τα διαφορετικά είδη των εμβολίων που υπάρχουν για την αντιμετώπιση του COVID-19, να υπολογίσει το πλήθος των ανθρώπων που έλαβαν αυτά τα εμβόλια και σε ποιες χώρες κατά τη διάρκεια του χρονικού διαστήματος Δεκεμβρίου του 2020 μέχρι 30 Ιουνίου 2022. Τα δεδομένα αντλήθηκαν από το Github repository της επιστημονικής διαδικτυακής ομάδας Our World in Data (OWID) που εστιάζει σε μεγάλα παγκόσμια προβλήματα όπως υπαρξιακούς κινδύνους, μέτρηση ανισοτήτων σε θέματα φτώχειας, ασθενειών, πείνας, πολέμων κτλ. Η ερευνητική ομάδα εδρεύει στο Πανεπιστήμιο της Οξφόρδης. Όπως αναφέρθηκε πριν υπάρχουν πολλά είδη εμβολίων κατά του COVID-19, διαφορετικά μεταξύ τους από διαφορετικές εταιρίες παραγωγής και αντιμετωπίζουν τον ιό με διαφορετικές προσεγγίσεις. Αναλυτικότερα έχουμε τα εξής δημοφιλή εμβόλια:

- PFIZER: Η Pfizer Inc. αναπτύσσει, κατασκευάζει και πουλά προϊόντα υγειονομικής περίθαλψης παγκοσμίως. Προσφέρει φάρμακα και εμβόλια σε διάφορους θεραπευτικούς τομείς. Το εμβόλιο που ανέπτυξε πωλείται με την επωνυμία Comirnaty και βασίζεται στο mRNA που αναπτύχθηκε από τη γερμανική εταιρεία βιοτεχνολογίας BioNTech. Για την ανάπτυξή της, η BioNTech συνεργάστηκε με την αμερικανική εταιρεία Pfizer για τη διεξαγωγή κλινικών δοκιμών, διαχείριση εφοδιαστικής αλυσίδας και κατασκευής. Είναι εγκεκριμένο για χρήση σε άτομα ηλικίας πέντε ετών και άνω σε ορισμένες δικαιοδοσίες, δώδεκα ετών και άνω σε ορισμένες δικαιοδοσίες, και για άτομα δεκαέξι ετών και άνω σε άλλες δικαιοδοσίες, για την παροχή προστασίας έναντι του COVID-19, που προκαλείται από μόλυνση απο τον ιό SARS-CoV-2.
- ASTRAZENECA: Η Moderna, Inc., είναι μια εταιρεία βιοτεχνολογίας κλινικού σταδίου, αναπτύσσει θεραπείες και εμβόλια με βάση το αγγελιοφόρο RNA για τη

θεραπεία μολυσματικών ασθενειών, ανοσο-ογκολογίας, σπάνιων ασθενειών και καρδιαγγειακών παθήσεων. Το εμβόλιο Oxford – AstraZeneca για τον COVID-19, με την κωδική ονομασία AZD1222, πωλείται μεταξύ άλλων με τις επωνυμίες Covishield και Vaxzevria, είναι ένα εμβόλιο ικού φορέα για την πρόληψη του COVID-19. Αναπτύχθηκε στο Ηνωμένο Βασίλειο από το Πανεπιστήμιο της Οξφόρδης και τη βρετανο-σουηδική εταιρεία AstraZeneca. Είναι εγκεκριμένο για χρήση σε άτομα ηλικίας δεκαοχτώ ετών και άνω μέχρι εξήντα ετών σε ορισμένες δικαιοδοσίες

- MODERNA: Η Moderna, Inc., είναι μια εταιρεία βιοτεχνολογίας κλινικού σταδίου, αναπτύσσει θεραπείες και εμβόλια με βάση το αγγελιοφόρο RNA για τη θεραπεία μολυσματικών ασθενειών, ανοσο-ογκολογίας, σπάνιων ασθενειών και καρδιαγγειακών παθήσεων. Το εμβόλιο Moderna COVID-19, που πωλείται με την επωνυμία Spikevax, είναι ένα εμβόλιο COVID-19 που αναπτύχθηκε από την αμερικανική εταιρεία Moderna, το Εθνικό Ινστιτούτο Αλλεργιών και Λοιμωδών Νοσημάτων των Ηνωμένων Πολιτειών (NIAID) και την Biomedical Advanced Research και την Αναπτυξιακή Αρχή (BARDA). Ανάλογα με τη δικαιοδοσία, επιτρέπεται η χρήση του σε άτομα ηλικίας έξι μηνών, δώδεκα ετών ή δεκαοκτώ ετών και άνω. Παρέχει προστασία έναντι του COVID-19 που προκαλείται από μόλυνση από τον ιό SARS-CoV-2.
- NOVAVAX: Η Novavax, Inc., μαζί με τη θυγατρική της, τη Novavax AB, μια εταιρεία βιοτεχνολογίας τελευταίου σταδίου, επικεντρώνεται στην ανακάλυψη, ανάπτυξη και εμπορευματοποίηση εμβολίων για την πρόληψη σοβαρών μολυσματικών ασθενειών. Το εμβόλιο Novavax COVID-19, που πωλείται με τις εμπορικές ονομασίες Nuvaxovid και Covovax, μεταξύ άλλων, είναι ένα εμβόλιο υπομονάδας COVID-19 που αναπτύχθηκε από τη Novavax και τον Συνασπισμό για Καινοτομίες Επιδημικής Ετοιμότητας (CEPI).
- JANSSEN VACCINES: Η Janssen Vaccines είναι μια εταιρεία βιοτεχνολογίας που ειδικεύεται στα εμβόλια και τις βιοφαρμακευτικές τεχνολογίες. Δημιουργήθηκε όταν η Johnson & Johnson εξαγόρασε την ολλανδική εταιρεία βιοτεχνολογίας Crucell με έδρα το Λέιντεν και την τοποθέτησε στο φαρμακευτικό της τμήμα. Το εμβόλιο Janssen COVID-19, που πωλείται με την επωνυμία Jcovden, είναι ένα

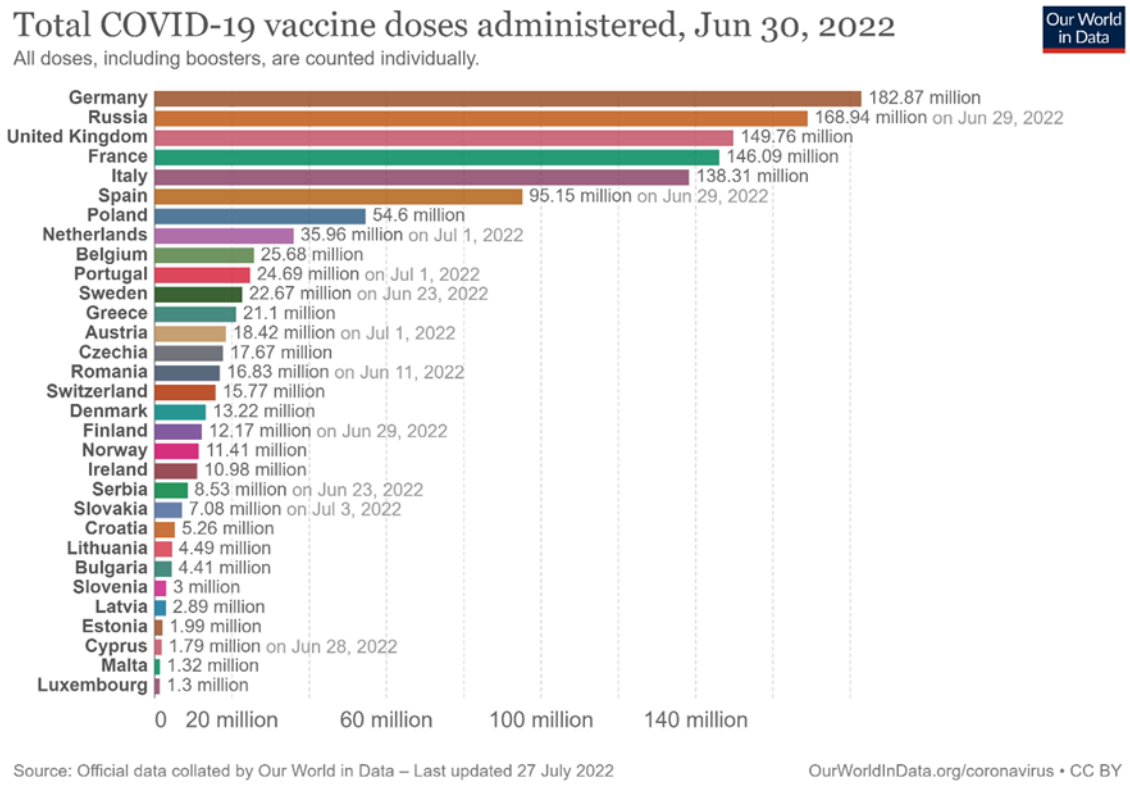
εμβόλιο κατά του COVID-19 που αναπτύχθηκε από την Janssen Vaccines στο Λέιντεν της Ολλανδίας και τη βελγική μητρική της εταιρεία Janssen Pharmaceuticals, μια θυγατρική εταιρεία. της αμερικανικής εταιρείας Johnson & Johnson.

Η έρευνα των εμβολίων έχει διεξαχθεί σε όλο τον κόσμο από την αρχή της έξαρσης της πανδημίας του COVID-19. Ακόμη και πριν από το ξέσπασμα του COVID-19, υπήρχαν έρευνες για τα εμβόλια κατά των κορονοϊών όπως το σοβαρό οξύ αναπνευστικό σύνδρομο (SARS) και το αναπνευστικό σύνδρομο της Μέσης Ανατολής (MERS). Αυτή η γνώση των κορονοϊών παρείχε κάποια βάση για την νέα έρευνα για τα εμβόλια ενάντια του COVID-19 και αυτό επιτάχυνε την πρόοδο ανάπτυξης των εμβολίων. Από τα τέλη του 2020, πολλές χώρες είχαν ξεκινήσει προγράμματα εμβολιασμού και βασίζονταν σε αυτή τη στρατηγική για να δημιουργήσουν σε μεγάλο βαθμό έναν τοίχο ανοσίας έτσι ώστε να σταματήσει η μετάδοση του COVID-19 και να καταστήσει τα ευάλωτα άτομα πιο ανθεκτικά στη νόσο. Υπάρχουν τέσσερις κύριες κατηγορίες εμβολίων κατά του COVID-19, ολόκληρος ο ιός, υπομονάδα πρωτεΐνης, ιικός φορέας και νουκλεϊκό οξύ (RNA και DNA). Τα εμβόλια με ολόκληρο τον ιό είναι συμβατικά εμβόλια που χρησιμοποιούνται για να πυροδοτήσουν μια ανοσολογική απόκριση. Μπορεί να είναι είτε αδρανοποιημένα εμβόλια είτε ζωντανά εξασθενημένα εμβόλια ως μια εξασθενημένη μορφή του ιού. Η υπομονάδα πρωτεΐνης χρησιμοποιεί κομμάτια του παθογόνου - συχνά θραύσματα πρωτεΐνης για την ενεργοποίηση της ανοσολογικής απόκρισης. Τα εμβόλια νουκλεϊκού οξέος χρησιμοποιούν γενετικό υλικό είτε RNA είτε DNA για να παρέχουν στα κύτταρα τις οδηγίες για την παραγωγή του αντιγόνου. Στην περίπτωση του COVID-19, αυτό είναι συνήθως η πρωτεΐνη ακίδας του ιού. Η Comirnaty (ή Pfizer-BioNTech COVID-19 εμβόλιο) και το εμβόλιο της Moderna, που χρησιμοποιούνται ευρέως Παγκοσμίως είναι αυτού του τύπου. Τα εμβόλια ιικών φορέων λειτουργούν δίνοντας στα κύτταρα γενετικές οδηγίες για παραγωγή αντιγόνων αλλά διαφέρουν από τα εμβόλια νουκλεϊκού οξέος στο ότι χρησιμοποιούν έναν αβλαβή ιό, διαφορετικό από αυτό που στοχεύει το εμβόλιο, για να παραδώσει αυτές τις οδηγίες στο κύτταρο. Το εμβόλιο Comirnaty και το εμβόλιο Moderna είναι ευρέως εγκεκριμένα στις Ευρωπαϊκές χώρες και χρησιμοποιούνται από τον Δεκέμβριο του 2020. Ωστόσο, παρόλο που αυτά τα εμβόλια είναι εγκεκριμένα και προμηθεύονται από πολλές

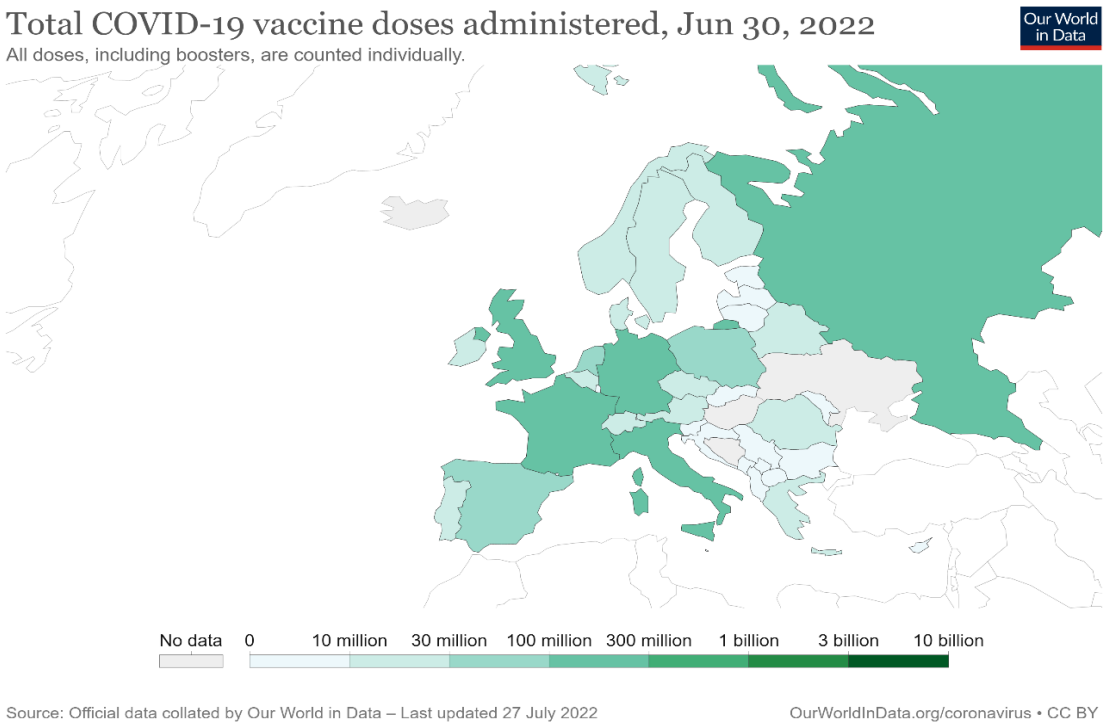
χώρες, το ποσοστό εμβολιασμού δεν ήταν στις αρχές τόσο υψηλό όσο αναμενόταν προηγουμένως και ποικίλλει στις χώρες της Ευρώπης.

4.2 Στατιστική ανάλυση των εμβολιασμών σε παγκόσμιο και ευρωπαϊκό επίπεδο

Το παρακάτω σχήμα απεικονίζει αθροιστικά τις δόσεις των εμβολιασμών στην Ευρώπη από τις 16 Ιανουαρίου 2021 μέχρι τις 30 Ιουνίου 2022. Οι περισσότερες δόσεις εμβολίων έγιναν στην Γερμανία (182.87 εκατομμύρια δόσεις), ακολουθεί η Ρωσία (168.94 εκατομμύρια δόσεις) και το Ηνωμένο Βασίλειο (149.76 εκατομμύρια δόσεις) κ.α.

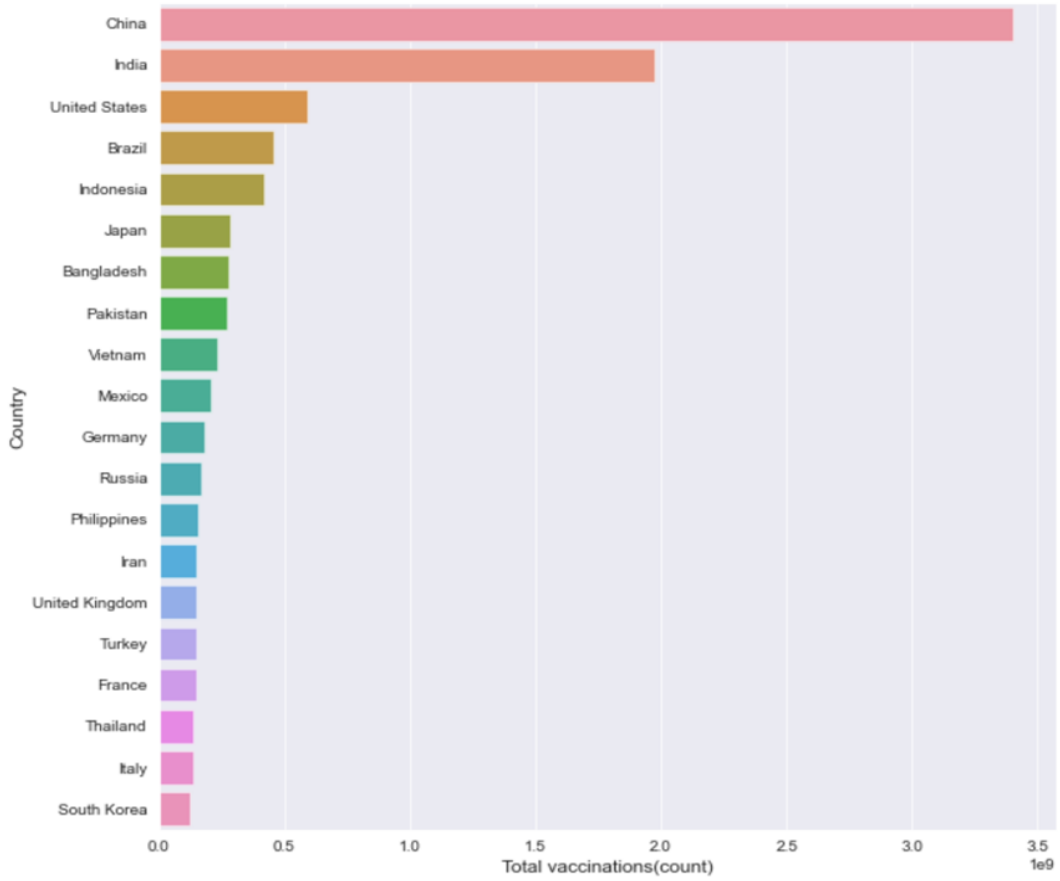


Εικόνα 15 Total COVID-19 Vaccine Doses in Europe 1



Εικόνα 16 Total COVID-19 Vaccine Doses in Europe 2

Για την παρακάτω ανάλυση χρησιμοποιήθηκαν με την χρήση της γλώσσας προγραμματισμού Python οι βιβλιοθήκες Pandas, NumPy, Matplotlib, Seaborn και η SciPy που χρησιμοποιείται για επιστημονικούς και τεχνικούς υπολογισμούς. Αρχικά πρέπει να απλοποιηθεί το σύνολο των δεδομένων εισόδου που αποτελείται από χρονοσειρές με την χρήση της βιβλιοθήκης της Pandas έτσι ώστε να εξεταστεί το σύνολο των ανθρώπων που έχουν εμβολιαστεί. Στη συνέχεια χρησιμοποιήθηκε η μέθοδος της βιβλιοθήκης της Pandas που ονομάζεται to_datetime για να βοηθήσει να μετατραπεί η σειρά της ημερομηνίας σε Pandas datetime αντικείμενο για να διευκολύνει την ανάλυση όταν αναλύονται δεδομένα χρονοσειρών. Επιπρόσθετα, έγινε και έλεγχος για τυχόν κενές τιμές και έλεγχος για τον τύπο των μεταβλητών. Έπειτα εφαρμόστηκε ένας καθαρισμός δεδομένων διότι υπήρχαν στην στήλη country οι ήπειροι και η οικονομική δυνατότητα των ανθρώπων (High-Low Income) τα οποία δεν χρησιμεύουν στην ανάλυση. Συνεπώς αφού δημιουργήθηκε ο τελικός πίνακας που περιέχει την πληροφορία που θέλουμε να αναλυθεί μπορεί να προχωρήσει η ανάλυση για την εξαγωγή των οπτικοποιήσεων. Παρακάτω παρατηρείται μια πρώτη οπτικοποίηση με την χρήση ενός ραβδογράμματος (bar plot) που δείχνει ποιες χώρες έχουν τους περισσότερους εμβολιασμούς παγκοσμίως.



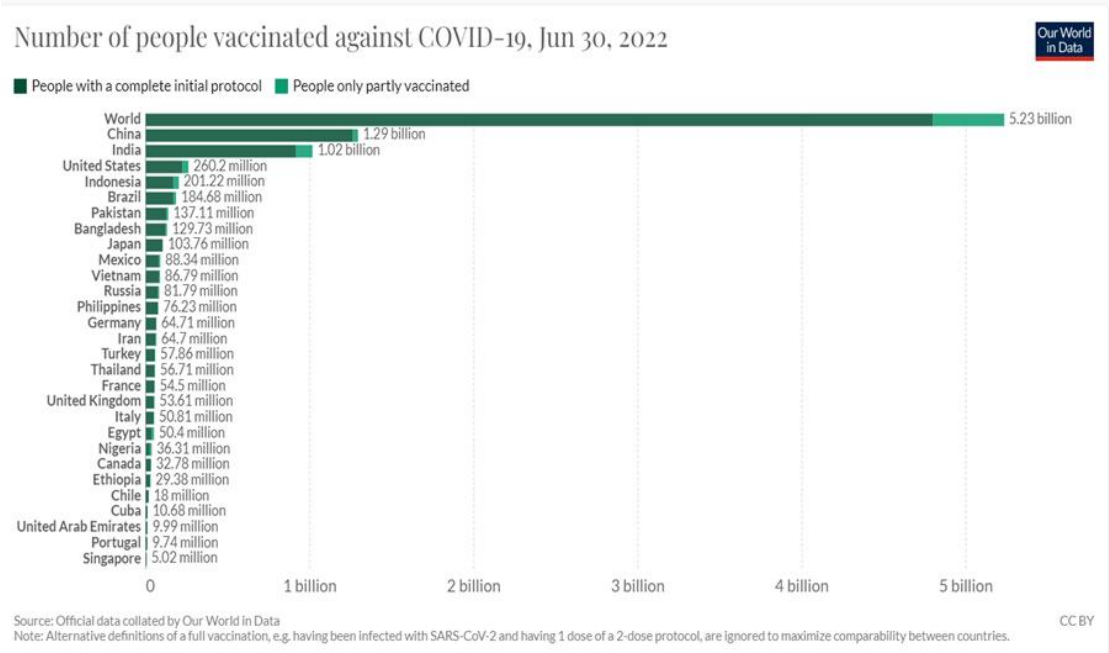
Εικόνα 17 Total COVID-19 Vaccine Doses Worldwide

Η στήλη Total Vaccinations περιέχει το συνολικό αριθμό των δόσεων που χορηγήθηκαν. Για τα εμβόλια που απαιτούν πολλαπλές δόσεις, υπολογίζεται κάθε μεμονωμένη δόση. Εάν ένα άτομο λάβει μία δόση του εμβολίου, αυτή η μέτρηση αυξάνεται κατά 1. Εάν λάβει μια δεύτερη δόση, αυξάνεται ξανά κατά 1. Εάν λάβουν μια τρίτη/αναμνηστική δόση, αυξάνεται ξανά κατά 1. Όπως παρατηρείται στο παραπάνω ραβδόγραμμα η Κίνα είναι στην πρώτη θέση με κοντά στις 3.5 δισεκατομμύρια δόσεις εμβολίων. Στη δεύτερη θέση βρίσκεται η Ινδία με 2 δισεκατομμύρια δόσεις και στην τρίτη θέση οι Ηνωμένες Πολιτείες της Αμερικής με κάτι παραπάνω από 500 εκατομμύρια δόσεις. Στη συνέχεια υπάρχει η στήλη Full_Vaccinations που αφορά τον συνολικό αριθμό ατόμων που έλαβαν όλες τις δόσεις που προέβλεπε το αρχικό πρωτόκολλο εμβολιασμού. Εάν ένα άτομο λάβει την πρώτη δόση ενός εμβολίου 2 δόσεων, αυτή η μέτρηση παραμένει η ίδια. Εάν λάβουν τη δεύτερη δόση, η μέτρηση αυξάνεται κατά 1. Επομένως, παρουσιάζονται παρακάτω οι πρώτες 20 χώρες παγκοσμίως με τα περισσότερα άτομα που ολοκλήρωσαν τον εμβολιασμό τους σύμφωνα με το ισχύων πρωτόκολλο της χώρας.

```
country
India             914479703.0
United States     222123223.0
Brazil            168924041.0
Indonesia         168251795.0
Pakistan          126718132.0
Bangladesh        119422697.0
Japan             102348805.0
Vietnam            81100223.0
Mexico             79947470.0
Russia             74673362.0
Philippines        70532636.0
Germany            63347864.0
Iran               57860379.0
Turkey             53088281.0
France             52905213.0
Thailand           52891174.0
United Kingdom     50126662.0
Italy              47948889.0
South Korea        44628221.0
England            42164176.0
Name: Full_vaccinations(count), dtype: float64
```

Εικόνα 18 Total Amount of People with Full Vaccination

Αξίζει να αναφερθεί πως η Κίνα δεν υπάρχει στον πίνακα αυτόν διότι στην συγκεκριμένη στήλη (Full_vaccinations) υπάρχουν κενές (NaN) τιμές. Όμως, σύμφωνα με την ιστοσελίδα https://ourworldindata.org/covid-vaccinations η Κίνα μέχρι τις 27 Ιουνίου 2022 είχε 1.26 δισεκατομμύρια πλήρους εμβολιασμούς και 33 εκατομμύρια μερικός εμβολιασμένους. Παρακάτω υπάρχει η ανάλογη εικόνα.

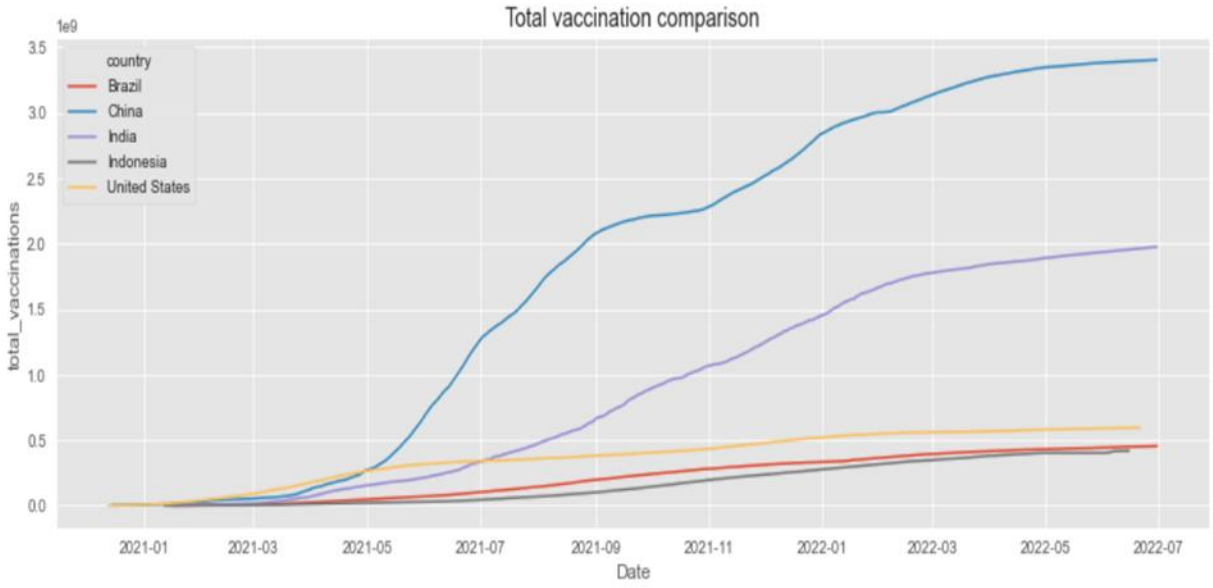


Εικόνα 19 Number of People Vaccinated Against COVID-19

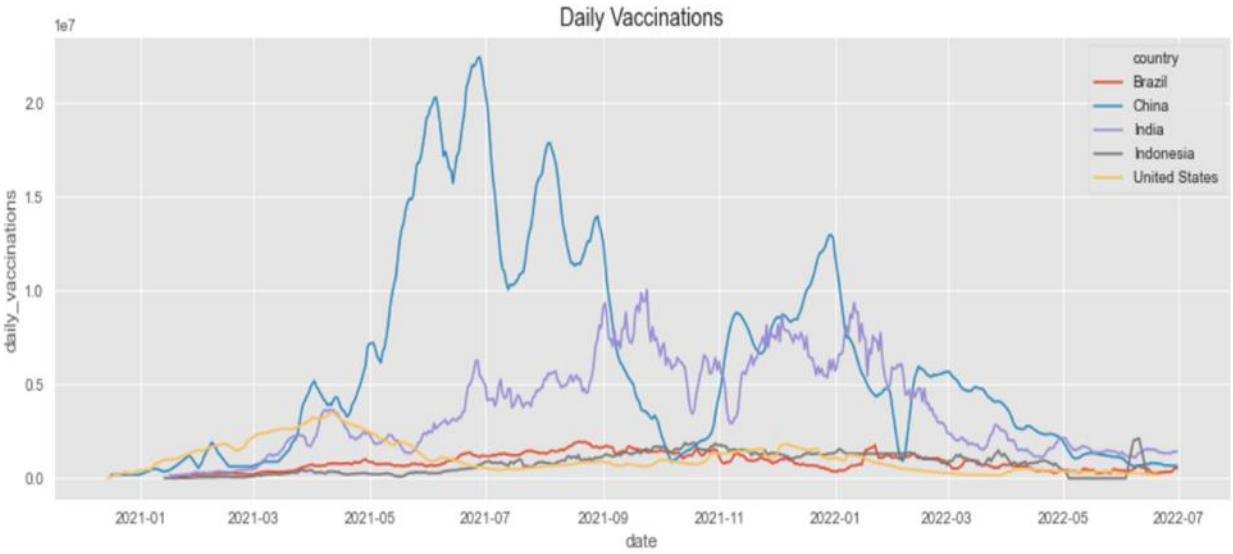
```
country
China            3.402622e+09
India            1.975335e+09
United States    5.937395e+08
Brazil           4.548542e+08
Indonesia        4.175223e+08
Japan            2.849892e+08
Bangladesh       2.774992e+08
Pakistan         2.701681e+08
Vietnam          2.321260e+08
Mexico           2.091793e+08
Name: Total_vaccinations(count), dtype: float64
```

Εικόνα 20 Top 10 Countries with the most successful Vaccinations

Στη συνέχεια θα δούμε σε παγκόσμια επίπεδο τις πέντε πρώτες χώρες με τους περισσότερους συνολικούς εμβολιασμούς μέχρις τις 30 Ιουνίου 2022 και παρακάτω τους ημερήσιους εμβολιασμούς.



Εικόνα 21 Top 5 Countries with the most successful Vaccinations

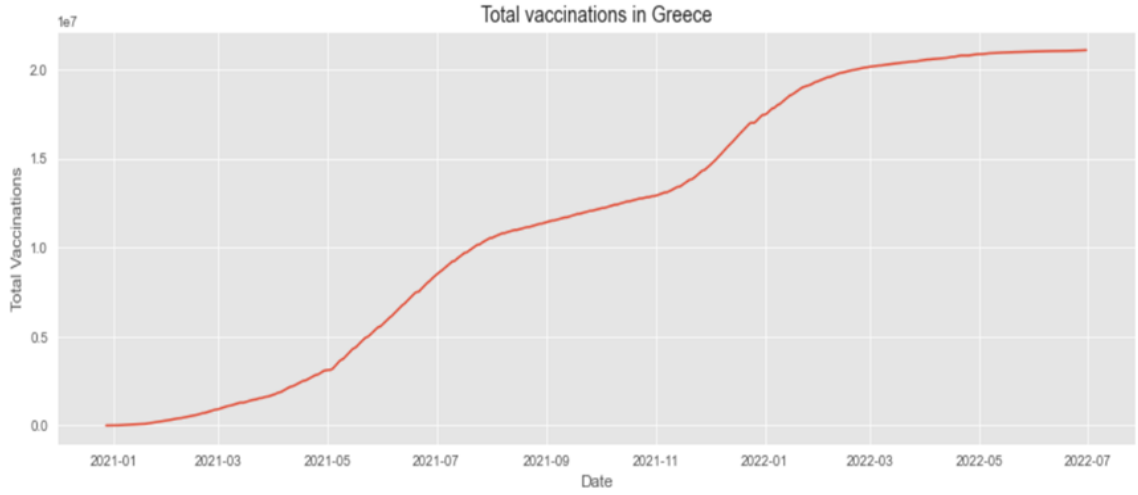


Εικόνα 22 Top 5 Countries with the most Daily Vaccinations

Όπως παρατηρούμε η Κίνα έχει τους περισσότερους ημερήσιους εμβολιασμούς με διαφορά από την δεύτερη Ινδία. Οι υπόλοιπες 3 χώρες (ΗΠΑ, Βραζιλία, Ινδονησία) κινούνται στα ίδια επίπεδα ημερησίων εμβολιασμών.

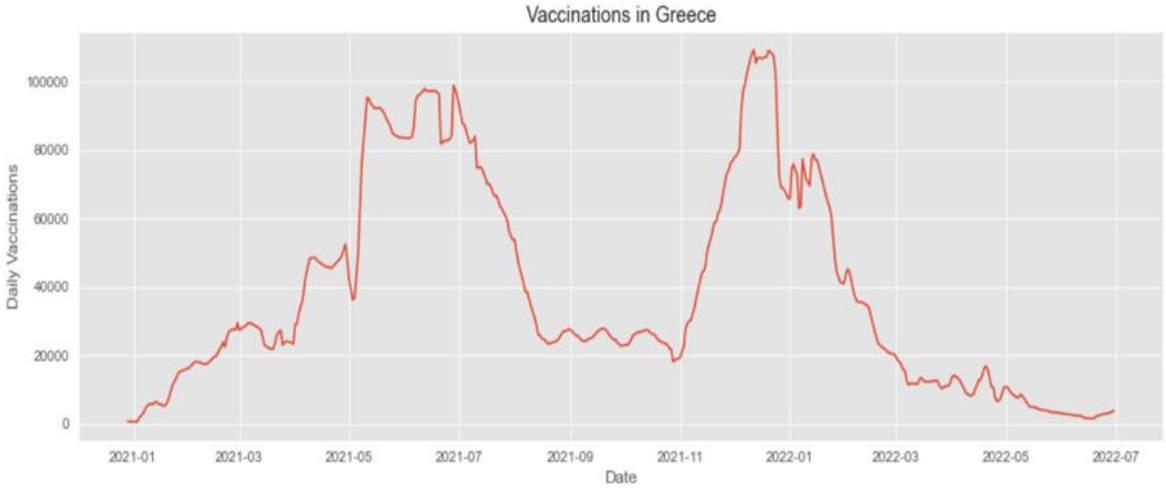
4.3 Ανάλυση πορείας των εμβολιασμών για την Ελλάδα

Για την Ελλάδα συγκεκριμένα υπάρχουν παρακάτω τρείς οπτικοποιήσεις. Η πρώτη αφορά τους συνολικούς εμβολιασμούς μέχρι τις 30 Ιουνίου 2022 όπου μέχρι τότε ήταν 21.097.421



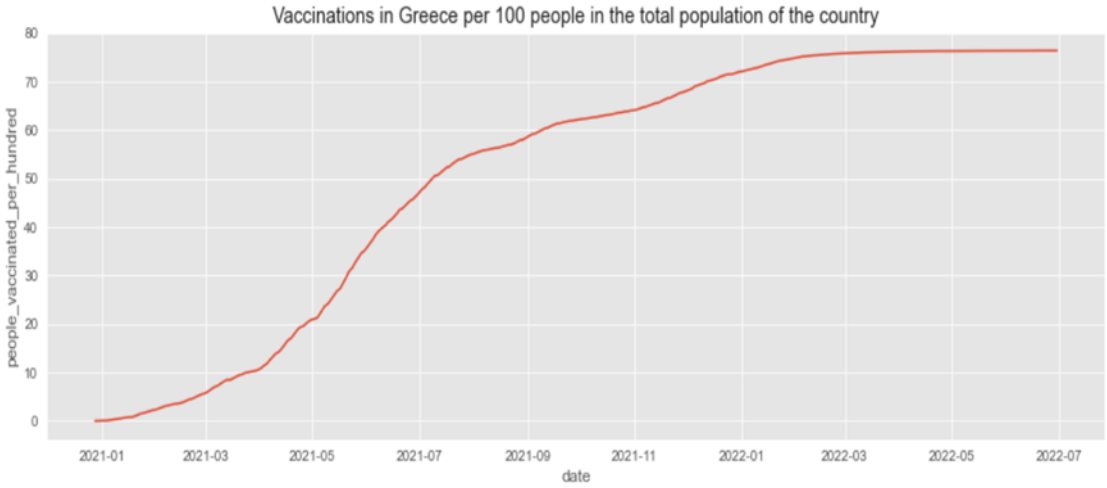
Εικόνα 23 Total Vaccinations in Greece

Η δεύτερη οπτικοποίηση αφορά τους καθημερινούς εμβολιασμούς που έγιναν στην Ελλάδα μέχρι τις 30 Ιουνίου 2022.



Εικόνα 24 Daily Vaccinations in Greece

Η τρίτη οπτικοποίηση αφορά τους συνολικούς εμβολιασμούς στην Ελλάδα ανά 100 άτομα στο σύνολο του πληθυσμού της χώρας. Μέχρι τις 30 Ιουνίου 2022 το ποσοστό των πλήρων εμβολιασμένων στην Ελλάδα ήταν 73,04% όπως φαίνεται και παρακάτω.



Εικόνα 25 Total Vaccinations in Greece per 100 people in the total population of the country

4.4 Μελέτη των διαθέσιμων εμβολίων ανά εταιρεία παραγωγής

Σε αυτή την ενότητα θα εφαρμοστεί μια σύντομη διερευνητική ανάλυση δεδομένων πάνω στα διαθέσιμα εμβόλια από διαφορετικές εταιρείες. Παρακάτω παρουσιάζονται ονομαστικά τα διαθέσιμα εμβόλια που υπάρχουν στο σύνολο δεδομένων που αναλύεται και η οπτικοποίηση με το άθροισμα των χορηγήσεών τους σε παγκόσμιο επίπεδο. Τέλος, υπάρχουν άλλες δύο οπτικοποιήσεις που δείχνουν το άθροισμα των διαφορετικών εμβολίων ανά εταιρεία παραγωγής σε επίπεδο Ευρωπαϊκής Ένωσης. Όλα τα παρακάτω δεδομένα ισχύουν μέχρι και τις 30 Ιουνίου 2022.

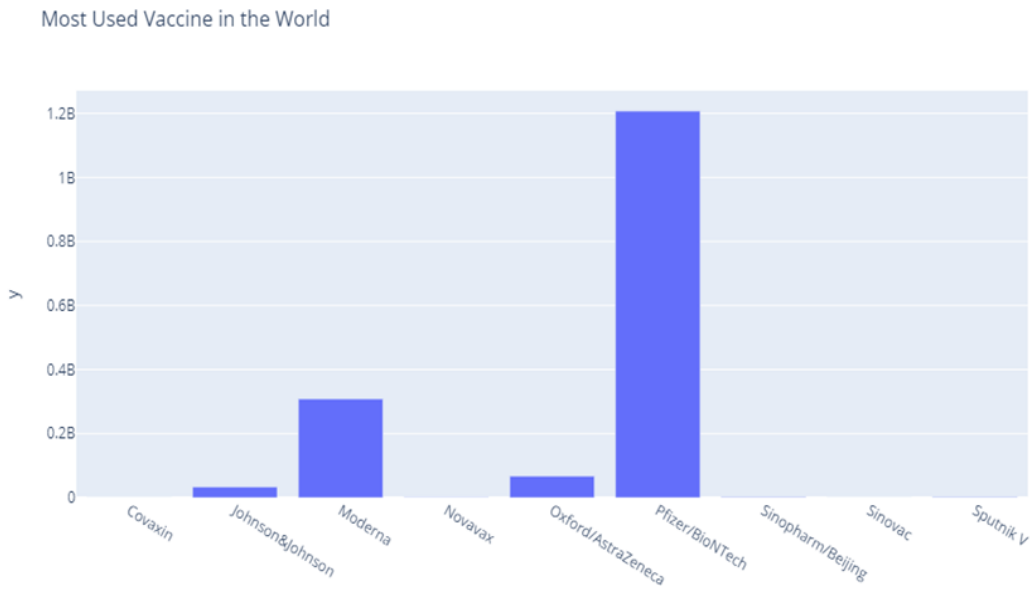
```
Pfizer/BioNTech
Johnson&Johnson
Moderna
Novavax
Covaxin
Oxford/AstraZeneca
Sinopharm/Beijing
Sinovac
Sputnik V
```

Εικόνα 26 Available Vaccines

```
vaccine
Covaxin                         53
Johnson&Johnson           33349143
Moderna                  308263285
Novavax                    1025307
Oxford/AstraZeneca        67133848
Pfizer/BioNTech         1207215345
Sinopharm/Beijing         2313766
Sinovac                      4811
Sputnik V                 1845102
Name: total_vaccinations, dtype: int64
```

Εικόνα 27 Total Vaccinations per Manufacturer

Όπως παρατηρείται σε παγκόσμιο επίπεδο το πιο χορηγούμενο εμβόλιο είναι της εταιρείας Pfizer/BioNTech με συνολικές χορηγήσεις 1.2 δισεκατομμύρια. Το δεύτερο σε σειρά είναι της εταιρείας Moderna με συνολικές χορηγήσεις 300 εκατομμύρια. Ακολουθεί το εμβόλιο της Oxford/AstraZeneca με 67 εκατομμύρια χορηγήσεις και έπειτα της Johnson&Johnson με 33 εκατομμύρια χορηγήσεις. Από τα δεδομένα αυτά προκύπτει ότι το πιο δημοφιλές εμβόλιο για την καταπολέμηση του COVID-19 ανήκει στην εταιρεία Pfizer/BioNTech με διαφορά. Παρακάτω υπάρχει και η σχετική οπτικοποίηση.

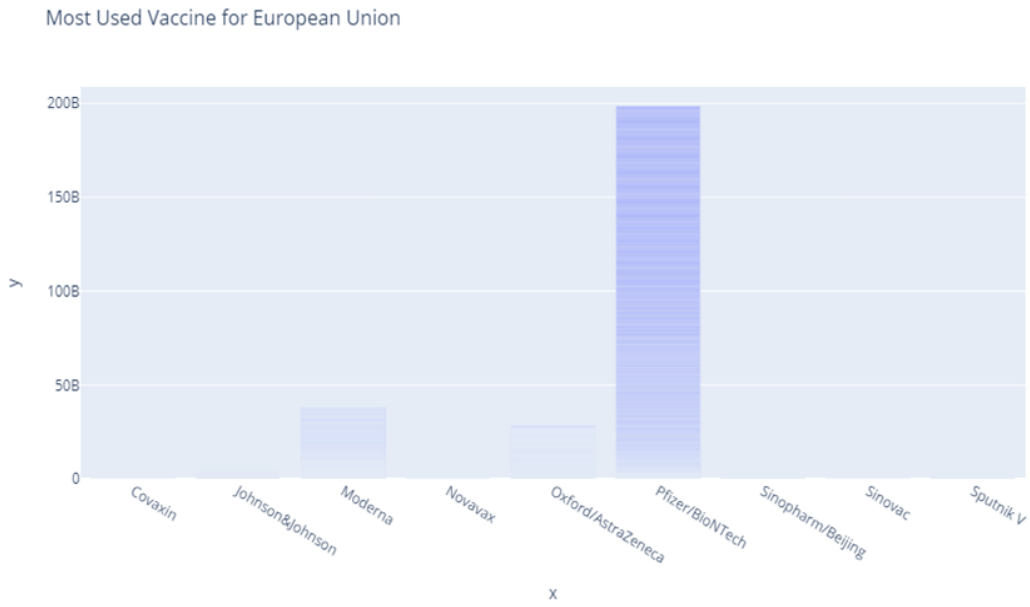


Εικόνα 28 Most Used Vaccines

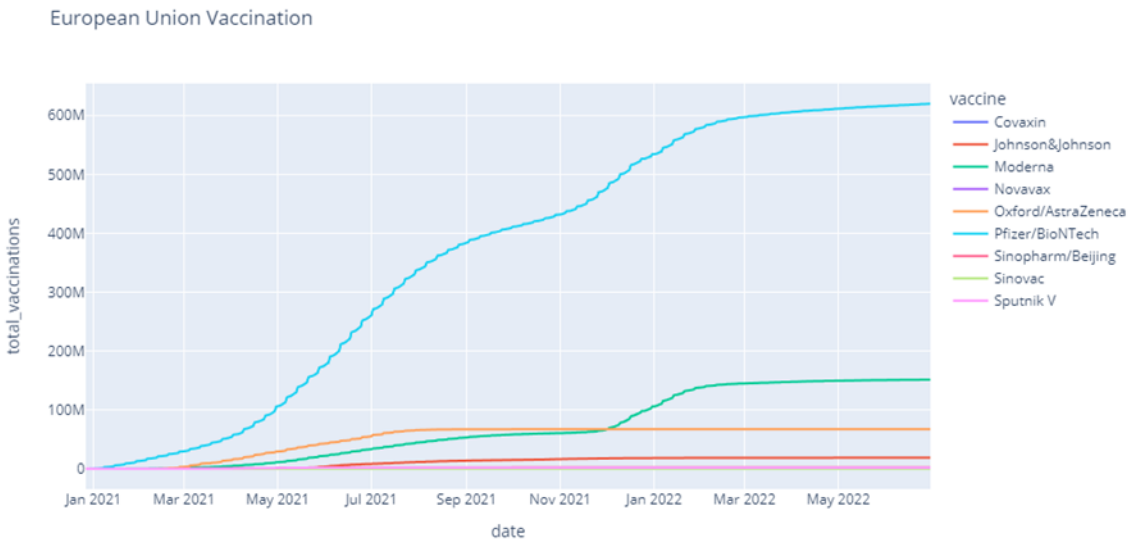
European Union	Covaxin	53
	Johnson&Johnson	18626105
	Moderna	151302886
	Novavax	240752
	Oxford/AstraZeneca	67133848
	Pfizer/BioNTech	619955808
	Sinopharm/Beijing	2313766
	Sinovac	4811
	Sputnik V	1845102

Εικόνα 29 Most used Vaccines in European Union1

Σε Ευρωπαϊκό επίπεδο παρατηρείται η ίδια εικόνα των πιο χορηγούμενων εμβολίων όπως και στο παγκόσμιο επίπεδο. Στην πρώτη θέση βρίσκεται πάλι το εμβόλιο της Pfizer/BioNTech με 620 εκατομμύρια συνολικές χορηγήσεις και στη δεύτερη θέση το εμβόλιο της Moderna με 150 εκατομμύρια συνολικές χορηγήσεις και ούτω καθεξής. Παρακάτω υπάρχουν δύο οπτικοποιήσεις. Η μια είναι ένα ραβδόγραμμα το οποίο δεν έχει έντονο χρώμα διότι προγραμματίστηκε στην Python έτσι ώστε να παρέχει αθροιστικές πληροφορίες για τις χορηγήσεις του κάθε εμβολίου σε κάθε σημείο του γραφήματος με την μετακίνηση του κέρσορα, σε αντίθεση με το προηγούμενο ραβδόγραμμα που αφορά τις συνολικές χορηγήσεις σε παγκόσμιο επίπεδο χωρίς αναλυτικές πληροφορίες ανά ημέρα. Παρακάτω εικονίζεται το γράφημα γραμμών το οποίο παρέχει τις ίδιες πληροφορίες με το ραβδόγραμμα σε συναρτήσει του χρόνου.



Εικόνα 30 Most used Vaccines in European Union2



Εικόνα 31 Total Vaccinations in European Union per Manufacturer

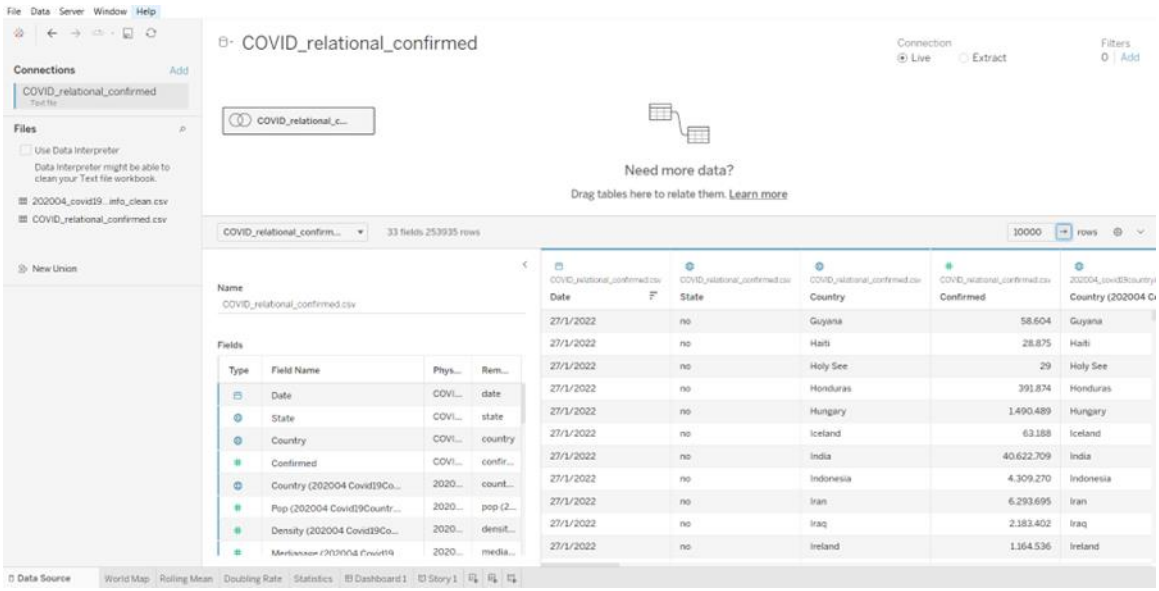
Όπως παρατηρείται και από αυτό το γράφημα το εμβόλιο της Pfizer/BioNTech είχε ραγδαία αύξηση στην Ευρωπαϊκή Ένωση μεταξύ των άλλων από την αρχή της χορήγησης των εμβολίων ενάντια στον COVID-19. Το εμβόλιο της Oxford/AstraZeneca ξεκίνησε ελάχιστα πιο δυναμικά να χορηγείται έναντι του εμβολίου της Moderna αλλά στην συνέχεια και πιο συγκεκριμένα από τον Δεκέμβριου του 2021 το εμβόλιο της Moderna ξεπέρασε το εμβόλιο της Oxford/AstraZeneca. Πλέον η χορήγηση των εμβολίων έχει στασιμότητα διότι έχει καλυφθεί ένα μεγάλο κομμάτι του παγκόσμιου πληθυσμού. Συγκεκριμένα μέχρι τις 30 Ιουνίου 2022 έχουν εμβολιαστεί πλήρως 4.8 δισεκατομμύρια άνθρωποι παγκοσμίως το όποιο είναι περίπου το 60% του πληθυσμού της Γης. Γι' αυτόν τον παραπάνω λόγο οι καμπύλες των εμβολίων τείνουν να γίνουν επίπεδες με την πάροδο του χρόνου.

ΚΕΦΑΛΑΙΟ 5

ΟΠΤΙΚΟΠΟΙΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗΝ ΧΡΗΣΗ ΤΟΥ TABLEAU

5.1 Εισαγωγή στο Tableau

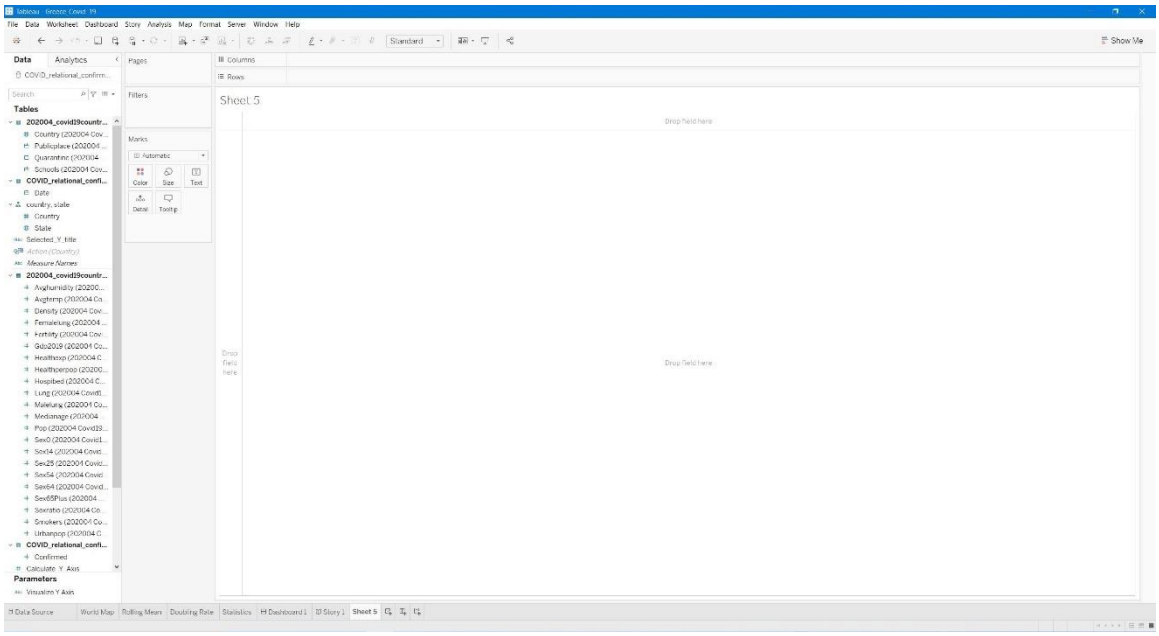
Το Tableau είναι ένα λογισμικό που χρησιμοποιείται για οπτικοποίηση και ανάλυση δεδομένων. Είναι ένα εργαλείο που μπορεί να διευκολύνει την ανάλυση δεδομένων και είναι αρκετά φιλικό προς τον χρήστη. Οι οπτικοποιήσεις έχουν τη μορφή φύλλων εργασίας (Worksheets) όπως στο Microsoft Excel. Αρχικά, για να αναλυθούν τα δεδομένα πρέπει πρώτα να εισαχθούν στο Tableau. Η εισαγωγή του Data Source το οποίο θα αναλυθεί όπως φαίνεται παρακάτω είναι το csv αρχείο COVID_relational_confirmed. Αφού εισαχθεί φαίνονται αυτομάτως τα πεδία τα οποία περιέχει το Data Source και έχουν αναγνωριστεί καταλλήλως από το Tableau ο τύπος δεδομένων του.



Εικόνα 32 Tableau Data Source

Γενικά το Tableau χωρίζει τα δεδομένα σε διαστάσεις (Dimensions) και σε μέτρα (Measures). Οι διαστάσεις περιέχουν ποιοτικές τιμές (όπως ονόματα, ημερομηνίες ή γεωγραφικά δεδομένα). Μπορούμε να χρησιμοποιήσουμε διαστάσεις για να κατηγοριοποιήσουμε, να τμηματοποιήσουμε και να αποκαλύψουμε τις λεπτομέρειες στα δεδομένα μας. Οι διαστάσεις επηρεάζουν το επίπεδο λεπτομέρειας στην προβολή. Τα μέτρα περιέχουν αριθμητικές, ποσοτικές τιμές που μπορούμε να μετρήσουμε. Όλη η λογική είναι να μεταφέρουμε απλά με τον κέρσορα τις κατάλληλες διαστάσεις και μετρήσιμα που θέλουμε να αναλύσουμε είτε στις γραμμές (Rows) είτε στις στήλες (Columns) είτε στο κέντρο του φύλλου εργασίας (Worksheet) όπως φαίνεται παρακάτω σε ένα κενό φύλλο εργασίας.

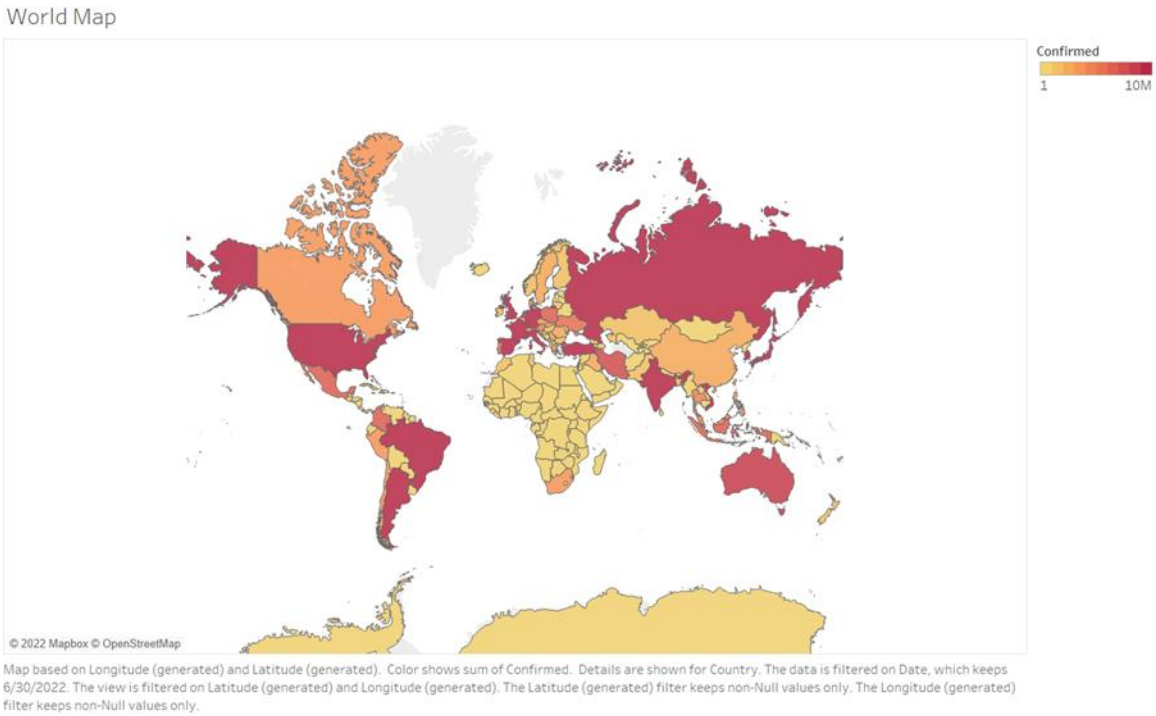
Επίσης μπορούν να δημιουργηθούν νέα μετρήσιμα ή αλλιώς υπολογισμούς (Calculations) και να ορίσουμε ως φίλτρο (Filters) μια διάσταση ή ένα μέτρο έτσι ώστε να μας βοηθήσουν στην ελαχιστοποίηση του μεγέθους των δεδομένων για λόγους αποτελεσματικότητας, γρήγορης φόρτωσης των δεδομένων, στον καθαρισμό των υποκείμενων δεδομένων, στην αφαίρεση άσχετων μελών διάστασης και στον ορισμό μετρήσεων ή εύρος ημερομηνιών για αυτό που θέλουμε να αναλύσουμε.



Εικόνα 33 Tableau Blank Worksheet

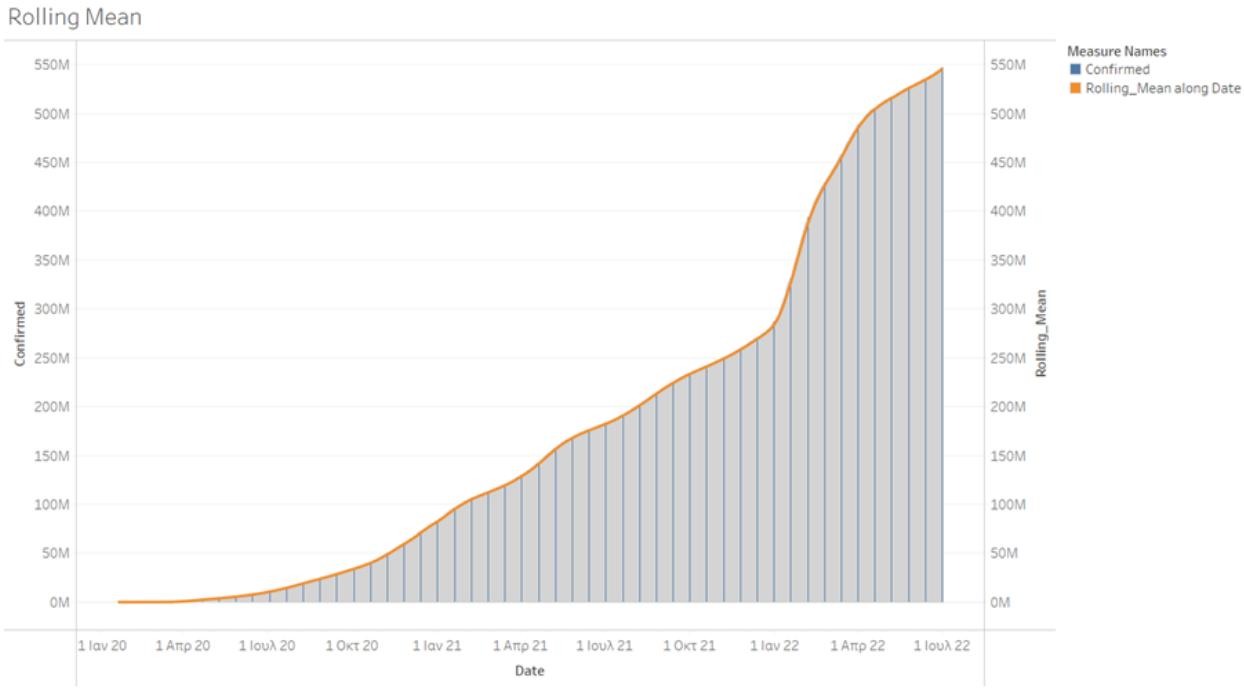
5.2 Στατιστικές οπτικοποιήσεις με την χρήση του Tableau

Η πρώτη ανάλυση που έγινε ήταν η οπτικοποίηση του παγκόσμιου χάρτη με χρωματισμό της κάθε χώρας από 10 παραλλαγές του πορτοκαλί χρώματος, ανάλογα με το σύνολο του μεγέθους των κρουσμάτων που έχουν επιβεβαιωθεί σε κλίμακα από το 1 μέχρι τα 10 εκατομμύρια. Όπως και στις προηγούμενες αναλύσεις που διεξήχθησαν τα παρακάτω αποτελέσματα είναι μέχρι τις 30 Ιουνίου 2022. Παρακάτω είναι το αποτέλεσμα.



Εικόνα 34 Total Covid-19 Cases World Map

Παρατηρείται ότι η Βραζιλία, η Αργεντινή, οι Ηνωμένες Πολιτείες της Αμερικής, η Κεντρική και Δυτική Ευρώπη, η Ρωσία, η Αυστραλία, η Ινδία και η Ιαπωνία έχουν αρκετά έντονο χρώμα το οποίο υποδεικνύει και αρκετά κρούσματα ανά 10 εκατομμύρια κατοίκους. Με τον κέρσορα πάνω σε οποιαδήποτε χώρα μπορούμε να δούμε αναλυτικά το συνολικό αριθμό των κρουσμάτων αυτής της χώρας με την βοήθεια του εργαλείου tooltip. Μια δεύτερη οπτικοποίηση είναι η χρονοσειρά των συνολικών κρουσμάτων παγκοσμίως μέχρι τις 30 Ιουνίου 2022 και συγχρόνως υπολογίζεται και ο κινητός μέσος όρος (Rolling Mean) όπου φαίνεται παράλληλα με την ημερομηνία. Ο κινητός μέσος όρος στην τεχνική ανάλυση παρέχει χρήσιμες πληροφορίες για την πορεία ενός μετρήσιμου μεγέθους. Φυσικά, επειδή το μετρήσιμο το οποίο αναλύουμε είναι αθροιστικό ο κινητός μέσος όρος θα έχει αυξητική τάση. Παρακάτω είναι το σχετικό διάγραμμα.

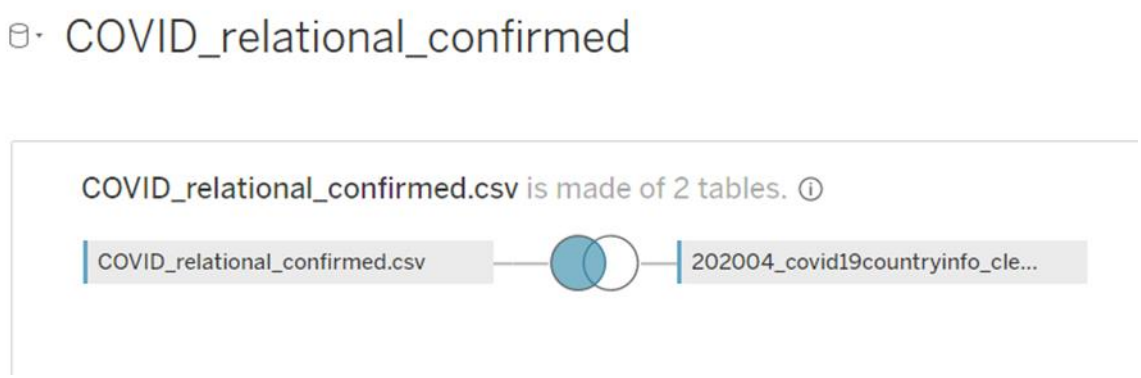


Εικόνα 35 Total Cases Worldwide Rolling Mean

Όπως παρατηρείται στο παραπάνω διάγραμμα τα κρούσματα φτάνουν τα 547 εκατομμύρια μέχρι τις 30 Ιουνίου 2022 και φαίνεται μια απότομη αύξηση των κρουσμάτων παγκοσμίως από την αρχή του έτους του 2022 πράγμα φυσιολογικό διότι η μετάλλαξη Όμικρον του κορωνοϊού ήταν σε έξαρση εκείνη την περίοδο και μεταδιδόταν ευκολότερα σε σχέση με την μετάλλαξη Δέλτα που προϋπήρχε.

5.3 Σύνδεση συνόλων δεδομένων στο Tableau και οπτικοποίηση αποτελεσμάτων

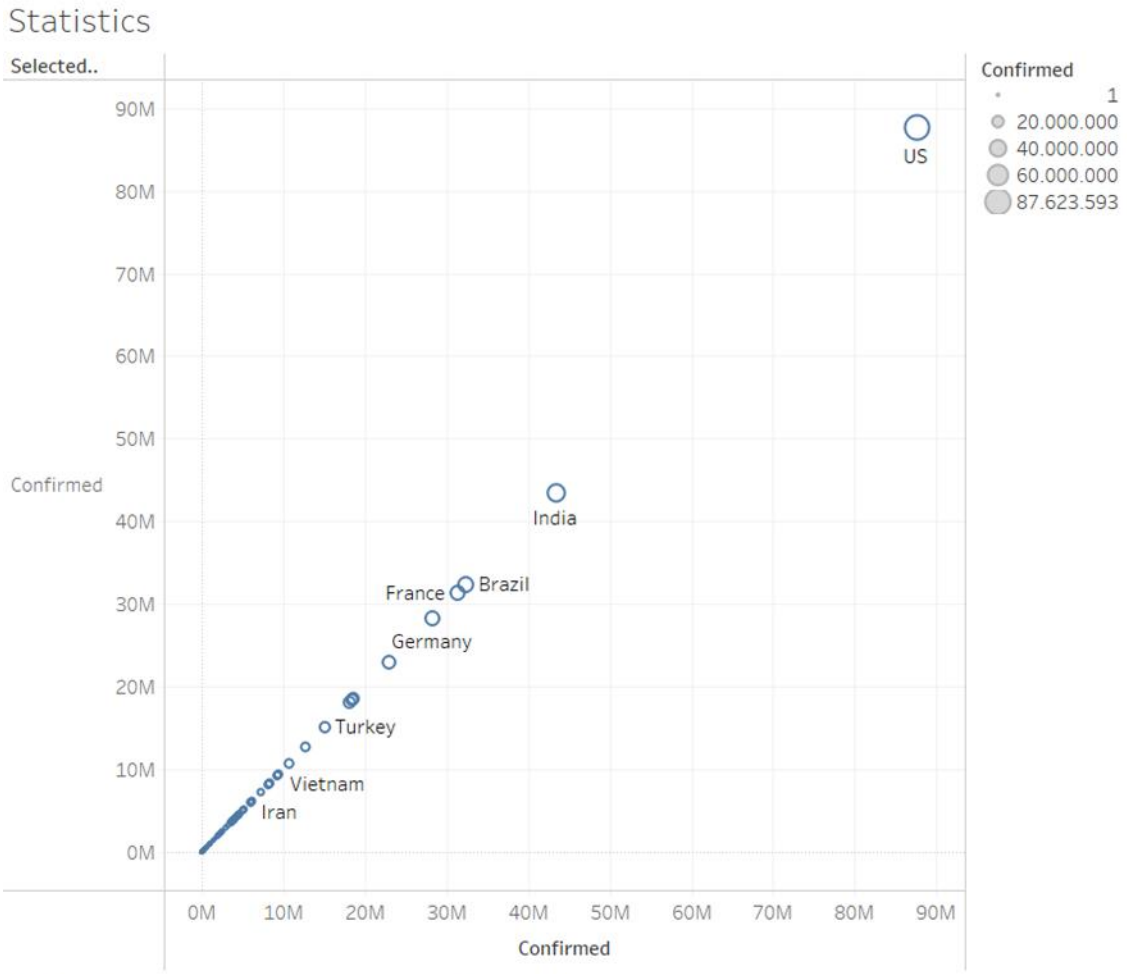
Στην πορεία της ανάλυσης έχει εισαχθεί ένα νέο Data Source το οποίο παρέχει στατιστικά δεδομένα για κάθε χώρα όπου μπορούμε να δούμε τις διαφορετικές συμπεριφορές της κάθε χώρας παραδείγματος χάριν πόσους καπνιστές έχει μια χώρα, πόσο πυκνοκατοικημένη είναι, ακόμη και πόσα νοσοκομειακά κρεβάτια διαθέτει μια χώρα ανά χίλους κατοίκους το οποίο είναι πολύ σημαντικό μετρήσιμο για την περίθαλψη των ασθενών που μολύνονται βαριά από τον COVID-19. Επίσης το συγκεκριμένο Data Source συνδέθηκε στο Tableau με Left Join δηλαδή στο Left Join φέραμε όλες τις εγγραφές του πρώτου Data Source ακόμα και αν δεν υπάρχουν εγγραφές που ταιριάζουν με το δεύτερο Data Source στο πεδίο που συσχετίζονται.



Εικόνα 36 Tableau Data sources Connection

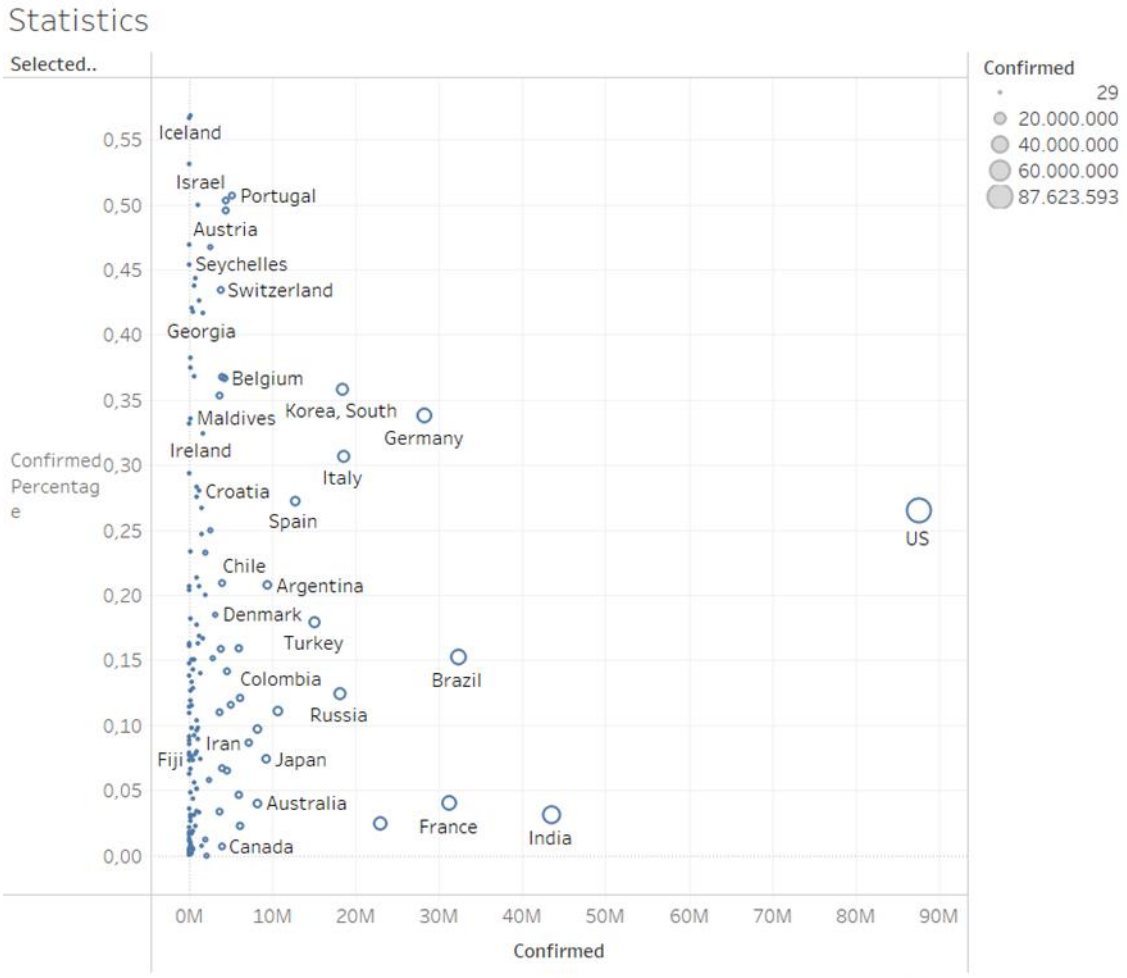
Παρακάτω θα παρουσιαστούν συνολικά πέντε ενδιαφέροντα διαγράμματα διασποράς για κάθε χώρα που συνδέονται άμεσα με την νόσο του COVID-19.

Το πρώτο διάγραμμα διασποράς είναι το άθροισμα των συνολικών κρουσμάτων ανά χώρα (opened cases, closed cases and deaths).



Εικόνα 37 Total Covid-19 Cases Statistics

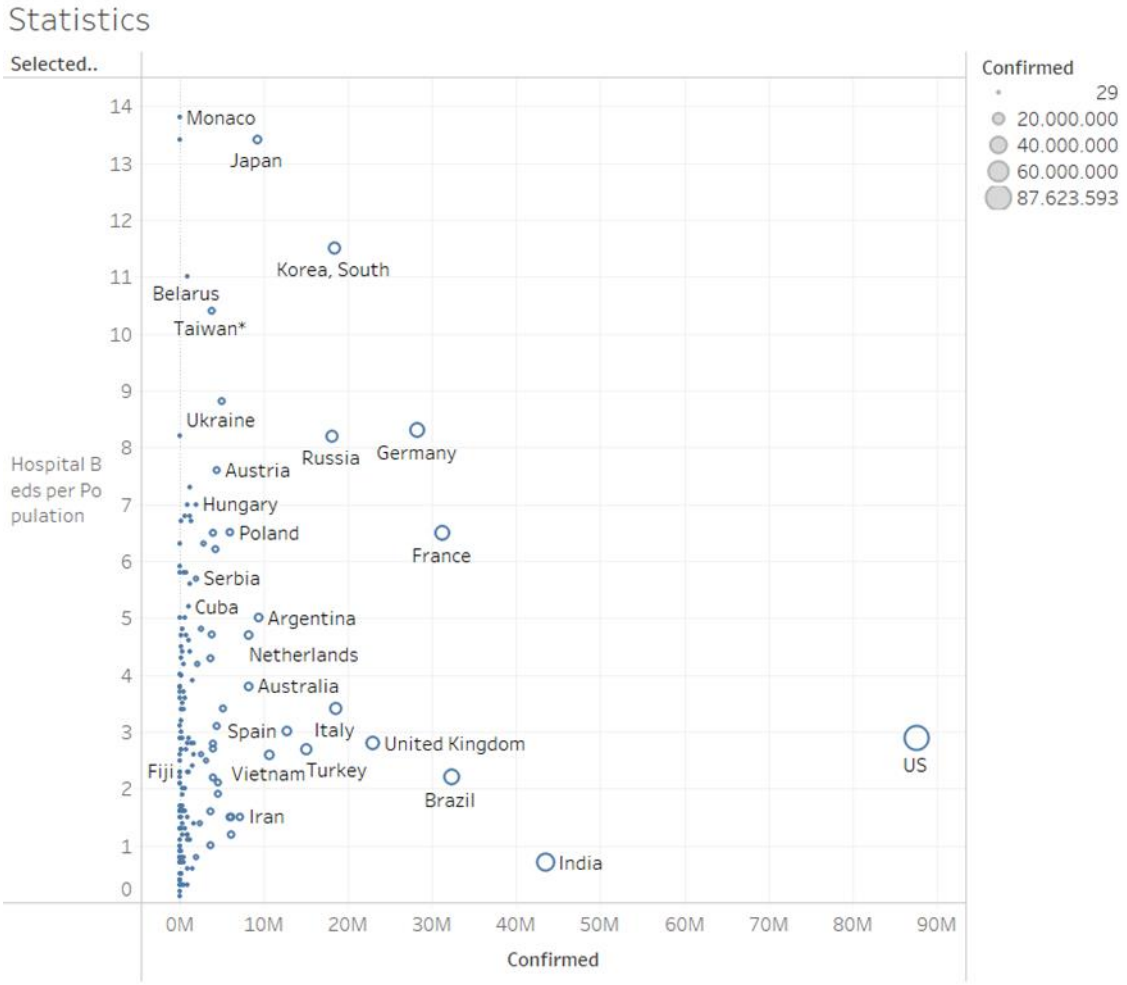
Το δεύτερο διάγραμμα παρουσιάζει το ποσοστό της διαίρεσης ανάμεσα στον συνολικό αριθμό των επιβεβαιωμένων κρουσμάτων όπως υπάρχει στο προηγούμενο διάγραμμα και στον συνολικό πληθυσμό της εκάστοτε χώρας.



Sum of Confirmed vs. Calculate_Y_Axis broken down by Selected_Y_title. Size shows sum of Confirmed. The marks are labeled by Country. The data is filtered on Date and Action (Country). The Date filter keeps 6/30/2022. The Action (Country) filter keeps 199 members. The view is filtered on Calculate_Y_Axis, which keeps non-Null values only.

Εικόνα 38 Total Covid-19 Cases Percentage

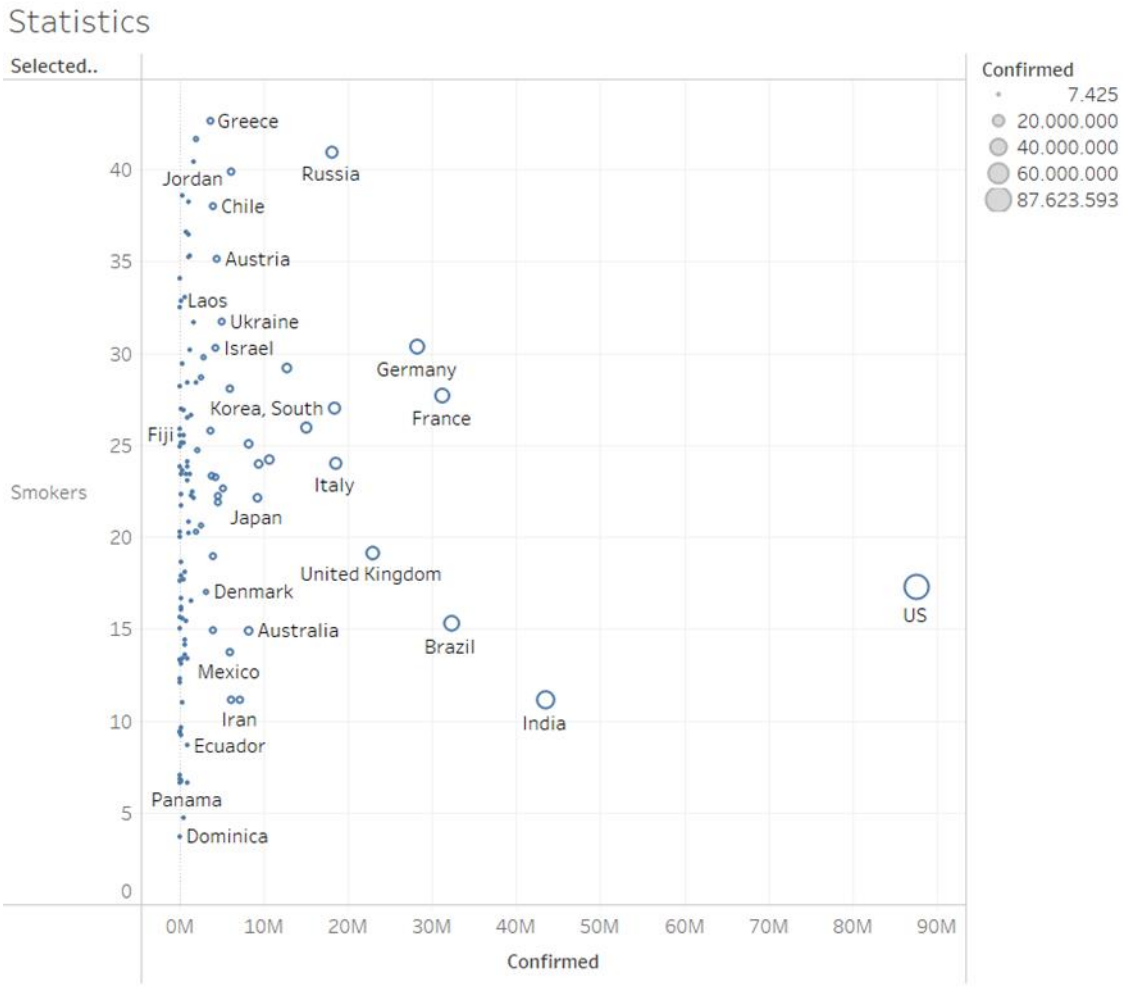
Το τρίτο διάγραμμα αφορά το σύνολο των νοσοκομειακών κρεβατιών που είναι διαθέσιμο για κάθε χώρα ανά χίλιους κατοίκους. Τα νούμερα αυτά υπάρχουν ήδη στο δεύτερο Data Source και δεν χρειάστηκε να υπολογίσουμε κάτι.



Sum of Confirmed vs. Calculate_Y_Axis broken down by Selected_Y_title. Size shows sum of Confirmed. The marks are labeled by Country. The data is filtered on Date and Action (Country). The Date filter keeps 6/30/2022. The Action (Country) filter keeps 199 members. The view is filtered on Calculate_Y_Axis, which keeps non-Null values only.

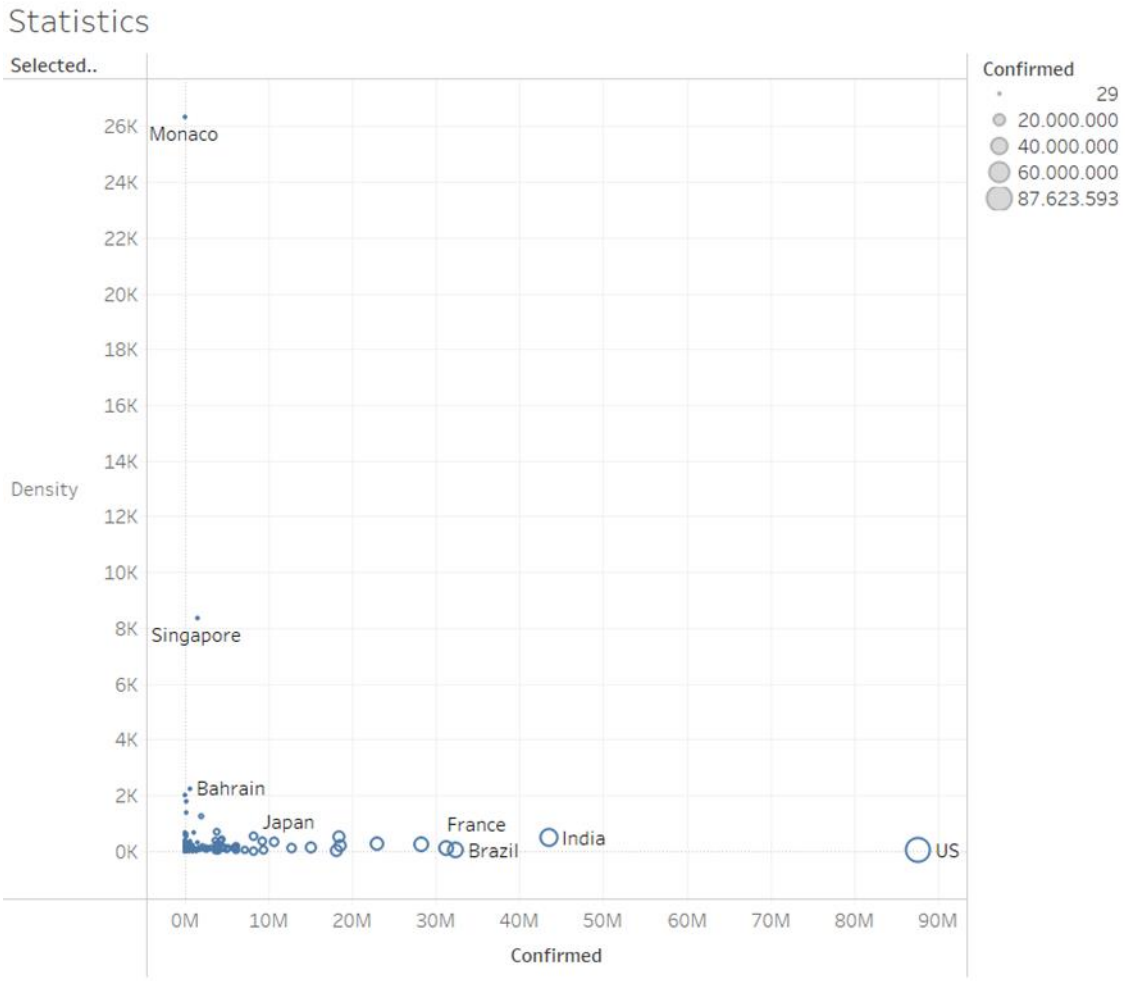
Εικόνα 39 Total Hospital Beds per Population

Το τέταρτο διάγραμμα αφορά το ποσοστό των καπνιστών της κάθε χώρας ανάλογα του πληθυσμού της. Και εδώ αυτό το στατιστικό δεδομένο υπήρχε έτοιμο στο Data Source.



Εικόνα 40 Total Smokers

Το πέμπτο και τελευταίο διάγραμμα δείχνει το ποσοστό των κατοίκων της κάθε χώρας ανά τετραγωνικό μέτρο. Και αυτό ήταν έτοιμο από το Data Source.

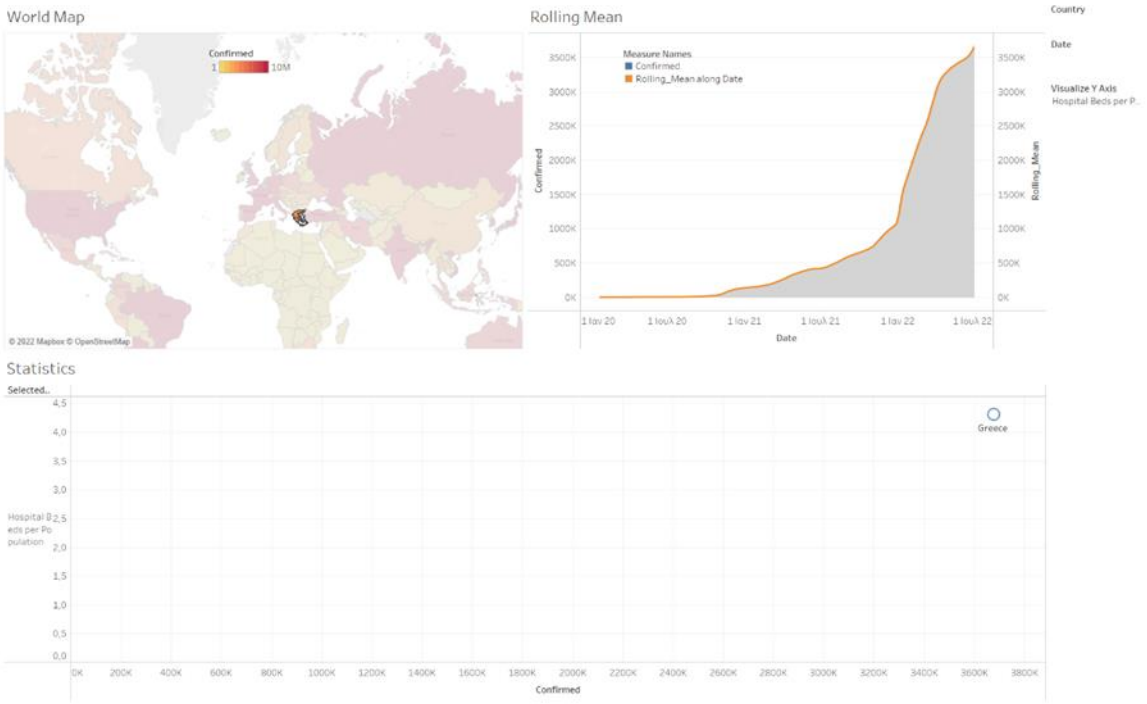


Sum of Confirmed vs. Calculate_Y_Axis broken down by Selected_Y_title. Size shows sum of Confirmed. The marks are labeled by Country. The data is filtered on Date and Action (Country). The Date filter keeps 6/30/2022. The Action (Country) filter keeps 199 members. The view is filtered on Calculate_Y_Axis, which keeps non-Null values only.

Εικόνα 41 Density per Square Meter

5.4 Tableau Dashboard

Τέλος, το Tableau προσφέρει την δυνατότητα να παρουσιαστούν μαζί διαφορετικά φύλλα εργασιών. Το μέρος του Tableau που το κάνει αυτό λέγεται ταμπλό ή αλλιώς Dashboard. Ένα Dashboard είναι μια συλλογή από πολλές προβολές, που επιτρέπουν να συγκρίνουμε μια ποικιλία δεδομένων ταυτόχρονα. Για παράδειγμα, εάν έχουμε ένα σύνολο προβολών που ελέγχεται κάθε μέρα, μπορεί να δημιουργηθεί ένα Dashboard που θα εμφανίζει όλες τις προβολές ταυτόχρονα, αντί να περιηγηθούμε σε ξεχωριστά φύλλα εργασίας. Τα δεδομένα σε φύλλα εργασίας και του Dashboard συνδέονται άρρηκτα. Όταν τροποποιείται ένα φύλλο εργασίας, το Dashboard που τα περιέχει αλλάζει και το αντίστροφο. Τόσο τα φύλλα εργασίας όσο και το Dashboard ενημερώνονται με τα πιο πρόσφατα διαθέσιμα δεδομένα από την πηγή δεδομένων. Παρακάτω θα παρουσιαστεί το Dashboard αποκλειστικά για την Ελλάδα.



Εικόνα 42 Tableau Dashboard for Greece

Όπως παρατηρείται υπάρχει μόνο για την Ελλάδα στο πάνω αριστερά διάγραμμα το άθροισμα των συνολικών κρουσμάτων ανά ημέρα που ανέρχεται στις 30 Ιουνίου 2022 στα 3.676.502 και στο κάτω διάγραμμα το σύνολο των νοσοκομειακών κρεβατιών που είναι διαθέσιμο για την Ελλάδα ανά χίλιους κατοίκους. Συγκεκριμένα τα κρεβάτια είναι 4.3 ανά 1000 κατοίκους. Φυσικά, μπορούμε στο κάτω διάγραμμα να επιλέξουμε τους καπνιστές της Ελλάδας ή το ποσοστό των κατοίκων της Ελλάδας ανά τετραγωνικό μέτρο όπως είδαμε προηγουμένως.

Όλες αυτές οι δυνατότητες οπτικοποίησης που προσφέρει το Tableau βοηθάνε να παρατηρηθούν αναλυτικά τα διάφορα μετρήσιμα μεταξύ τους, το πως συνδέονται, να κατανοηθεί επίσης τι κρύβεται από πίσω από τα διάφορα νούμερα που υπάρχουν και που υπάρχει το μεγαλύτερο πρόβλημα έτσι ώστε να προταθούν λύσεις για τα προβλήματα αυτά. Τέλος, είναι ένας βολικός τρόπος παρουσίασης των δεδομένων ο οποίος είναι και φιλικός στον χρήστη για την δημιουργία τέτοιων οπτικοποιήσεων αλλά και οπτικά κομψός στα άτομα όπου γίνεται η παρουσίαση.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. The Lancet (Jan. 2020)."A novel coronavirus outbreak of global health concern" URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7135038/
2. World Health Organization (WHO). "COVID-19 Vaccines". URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines
3. Tableau Software, LLC, Salesforce Company URL: https://www.tableau.com/
4. The Python Software Foundation URL: https://www.python.org/
5. Project Jupyter URL: https://jupyter.org/
6. Our World In Data, Coronavirus (COVID-19) Vaccinations, URL: https://ourworldindata.org/covid-vaccinations
7. Brownlee, Jason (Dec. 2016). "What Is Time Series Forecasting?" Machine Learning Mastery. URL: https://machinelearningmastery.com/time-series-forecasting/
8. CDC (Apr. 2021a). "Frequently Asked Questions". URL: https://www.cdc.gov/coronavirus/2019-ncov/faq.html
9. Cooper, I., A. Mondal, and C. Antonopoulos (June 2020). "A SIR Model Assumption for the Spread of COVID-19 in Different Communities". US National Library of Medicine National Institutes of Health. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7321055/
10. Hopkins, Johns (Apr. 2021). "Coronavirus Resource Center". Johns Hopkins University. URL: https://coronavirus.jhu.edu
11. Kufel, T. (June 2020). "ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries". Equilibrium.Quarterly Journal of Economics and Economic Policy 15, pp. 181–204. URL: https://ideas.repec.org/a/pes/ierequ/v15y2020i2p181-204.html
12. Tian, Y., I. Luthra, and X. Zhang (July 2020). "Forecasting COVID-19 cases using Machine Learning models". MedRxiv. URL: https://www.medrxiv.org/content/10.1101/2020.07.02.20145474v1.article-info

13. Adithyan, Nikhil (Aug. 2020). “Covid-19 Analysis with Python” Towards Data Science. URL: https://medium.com/codex/covid-19-analysis-with-python-b898181ea627
14. Yiu, Tony (Apr. 2020). “Understanding ARIMA (Time Series Modeling)”. Towards Data Science. URL: https://towardsdatascience.com/understandingarima-time-series-modeling-d99cd11be3f8
15. Edwards, Gavin (Apr. 2018) “Machine Learning An Introduction”. Towards Data Science. URL: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0
16. Zeroual, A. et al. (Nov. 2020). “Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study”. Chaos, Solitons and Fractals 140.110121. URL: https://www.sciencedirect.com/science/article/pii/S096007792030518X
17. Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, Cameron Appel, Lucas Rodés-Guirao and Max Roser (2020) "Coronavirus Pandemic (COVID-19)". OurWorldInData.org. URL: https://ourworldindata.org/coronavirus
18. Coronavirus Disease (COVID-19) - events as they happen. Library Catalog, URL: www.who.int
19. Countries where Coronavirus has spread - Worldometer. Library Catalog, URL: www.worldometers.info
20. Chand, Eswar (May 2020). “Timeseries Forecasting of Covid-19 ARIMA” Kaggle. URL: https://www.kaggle.com/code/eswarchandt/timeseries-forecasting-of-covid-19-arima