Πανεπιστήμιο Πειραιώς
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΚΙΝΗΤΗΣ ΤΗΛΕΦΩΝΙΑΣ

Θεοδώρα Γ. Νίκου

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

Πειραιάς
2022

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. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μάρκος Κούτρας (Επιβλέπων)
- Καθηγήτρια Γεωργία Βερροπούλου
- Αναπληρωτής Καθηγητής Πλάτων Τήνιος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

APPLICATION OF MULTIVARIATE ANALYSIS TECHNIQUES IN MOBILE PHONE DATA

By

Theodora G. Nikou

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, Greece
2022

Στην οικογένειά μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους γονείς μου και τον επιβλέποντα καθηγητή μου για τη στήριξη και την καθοδήγηση που μου προσέφεραν κατά τη διάρκεια της εκπόνησης της διπλωματικής εργασίας.

Περίληψη

Στα πλαίσια της παρούσας διπλωματικής εργασίας εφαρμόζονται μέθοδοι Πολυμεταβλητής Ανάλυσης σε δεδομένα κινητής τηλεφωνίας. Τα δεδομένα αυτά συλλέχθηκαν από πελάτες καταστήματος εταιρίας κινητής τηλεφωνίας το καλοκαίρι του 2020 και αφορούν συνδρομητές οι οποίοι χρησιμοποιούν πρόγραμμα καρτοκινητής της συγκεκριμένης εταιρίας.

Η πολυμεταβλητή στατιστική ανάλυση, πλέον, αποτελεί ένα πολύτιμο εργαλείο στα χέρια των στατιστικών οι οποίοι ενδιαφέρονται να αναλύσουν δεδομένα από επιχειρήσεις. Οι μέθοδοί της διευκολύνουν την εξαγωγή χρήσιμων συμπερασμάτων από μεγάλο όγκο δεδομένων που αποτελείται από πιθανώς εκατοντάδες μεταβλητές, οι οποίες με τη σειρά τους αποτελούνται από χιλιάδες παρατηρήσεις.

Σκοπός της εργασίας, είναι η παρουσίαση της θεωρίας ορισμένων χαρακτηριστικών μεθόδων της Πολυμεταβλητής Ανάλυσης καθώς και η πρακτική εφαρμογή τους σε πραγματικά δεδομένα μίας εταιρίας κινητής τηλεφωνίας. Μέσω της εφαρμογής των μεθόδων αυτών, η εταιρία, συμπυκνώνοντας την πληροφορία που έχει συγκεντρώσει από αυτό το μεγάλο πλήθος παρατηρήσεων και μεταβλητών, μπορεί να δημιουργήσει προφίλ πελατών, κατηγοριοποιώντας τους σε ομάδες, καθώς και να βρει ποιοι παράγοντες επηρεάζουν τις επιλογές των συνδρομητών στην αγορά των προϊόντων της εταιρίας και συγκεκριμένα των επιλογών τους στα προγράμματα καρτοκινητής. Η εταιρία βασιζόμενη στα αποτελέσματα που θα προκύψουν, θα μπορούσε να καθορίσει ένα καινούριο στρατηγικό πλάνο διαπιστώνοντας αν χρειάζεται να γίνει αναδιαμόρφωση των προϊόντων που παρέχει, ανάλογα με τις ανάγκες των συνδρομητών της ή να επιλέξει πως θα κινηθεί από πλευράς μάρκετινγκ στοχεύοντας ξεχωριστά σε κάθε ομάδα πελατών.

Abstract

In the present master thesis, methods of Multivariate Statistics are applied to mobile telephony data. This data was collected by customers from a mobile phone company store in the summer of 2020 and concern subscribers who use a prepaid program from the specific company.

Multivariate statistical analysis is now a valuable tool in the hands of statisticians who are interested in analyzing data from companies. Its methods may be exploited to draw useful conclusions from a large body of data consisting of possibly hundreds of variables, which in turn consist of thousands of observations.

The purpose of this thesis is to present the theory of a number of popular methods of Multivariate Analysis as well as their practical application in real data of a mobile telephony company. Through the application of these methods, the company, by summarizing the information gathered from this large number of observations and variables, can create customer profiles, categorize them into groups, as well as find out what factors influence the subscribers' choices in the market products of the company and specifically their choices in prepaid programs.

The company, based on the results that could be obtained, could define a new strategic plan by determining whether it is necessary to restructure the products it provides, according to the needs of its subscribers or to choose its marketing plan, by analyzing each customer group separately.

Περιεχόμενα

Κεφάλαιο 1

1.1 Εισαγωγή

Το αντικείμενο της διπλωματικής αυτής είναι η εφαρμογή μεθόδων Πολυμεταβλητής Ανάλυσης σε δεδομένα που έχουν συλλεχθεί από συνδρομητές ελληνικής εταιρίας κινητής τηλεφωνίας. Τα δεδομένα συλλέχθηκαν το καλοκαίρι του 2020.

Οι μέθοδοι Πολυμεταβλητής Ανάλυσης που θα εξετασθούν είναι

- Ανάλυση Κατά Συστάδες (Clustering)
- Ανάλυση Κυρίων Συνιστωσών (Principle Component Analysis)
- Ανάλυση Παραγόντων (Factor Analysis)

Η ανάλυση των δεδομένων έγινε με την χρήση του στατιστικού πακέτου IBM SPSS Statistics 24.0

1.2 Παρουσίαση Μεταβλητών

Τα δεδομένα συλλέχθηκαν από πελάτες συγκεκριμένου καταστήματος εταιρίας κινητής τηλεφωνίας. Σε διάστημα δυο μηνών, καταγράφηκαν ορισμένα στοιχεία (μεταβλητές) από τυχαίους συνδρομητές καρτοκινητής που επισκέπτονταν το κατάστημα για να ανανεώσουν την κάρτα τους.

Το δείγμα που συλλέχθηκε αποτελείται από 100 παρατηρήσεις (100 συνδρομητές) για τους οποίους καταγράφηκαν 13 μεταβλητές.

Οι μεταβλητές είναι οι εξής:

- age: Δηλώνει την ηλικία του συνδρομητή
- sex: Κατηγορική μεταβλητή που δηλώνει το φύλο, όπου 1: άνδρας και 2: γυναίκα
- yearpos: Δηλώνει τα χρόνια κατοχής του καρτοκινητού από τον συνδρομητή
- voice: Δηλώνει τα λεπτά ομιλίας που προσφέρει για 30 ημέρες το πρόγραμμα που έχει επιλέξει ο κάθε συνδρομητής.
- sms: Δηλώνει τον αριθμό των μηνυμάτων που προσφέρει για 30 ημέρες το πρόγραμμα που έχει επιλέξει ο κάθε συνδρομητής.
- data: Δηλώνει τον αριθμό των δεδομένων (σε mb) που προσφέρει για 30 ημέρες το πρόγραμμα που έχει επιλέξει ο κάθε συνδρομητής.
- bundle: Κατηγορική μεταβλητή που παίρνει τιμές από το 1 μέχρι το 7 και δείχνει το πρόγραμμα καρτοκινητής που έχει επιλέξει ο συνδρομητής. Η μεταβλητή

αυτή αποτελεί συγκεκριμένους συνδυασμούς των τριών προηγούμενων μεταβλητών. Για το λόγο αυτό θα χρησιμοποιηθεί μόνο αυτή η μεταβλητή στην παρακάτω ανάλυση.

- resvoice: Δείχνει τα λεπτά ομιλίας που απομένουν στο συνδρομητή όταν λήγει το πρόγραμμα μετά το πέρας των 30 ημερών.
- ressms: Δείχνει τον αριθμό των μηνυμάτων που απομένουν στο συνδρομητή όταν λήγει το πρόγραμμα μετά το πέρας των 30 ημερών.
- resdata: Δείχνει τον αριθμό των δεδομένων που απομένουν στο συνδρομητή όταν λήγει το πρόγραμμα μετά το πέρας των 30 ημερών.
- percentagevoiceuse: Δηλώνει το ποσοστό του χρόνου ομιλίας που έχει χρησιμοποιήσει ο συνδρομητής κατά τη διάρκεια των 30 ημερών σε σχέση με τον αρχικό χρόνο ομιλίας που του παρείχε το πρόγραμμα.
- percentagesmsuse: Δηλώνει το ποσοστό των μηνυμάτων που έχει χρησιμοποιήσει ο συνδρομητής κατά τη διάρκεια των 30 ημερών σε σχέση με το αρχικό πλήθος μηνυμάτων που του παρείχε το πρόγραμμα.
- percentagedatause: Δηλώνει το ποσοστό των δεδομένων που έχει χρησιμοποιήσει ο συνδρομητής κατά τη διάρκεια των 30 ημερών σε σχέση με τον αρχικό αριθμό δεδομένων που του παρείχε το πρόγραμμα.

1.3 Περιγραφικά στοιχεία μεταβλητών

Παρακάτω παρουσιάζονται οι πίνακες με τα περιγραφικά μέτρα των μεταβλητών καθώς και τα γραφήματα συχνοτήτων τους.

Παρατηρείται ότι στο δείγμα δεν υπάρχουν ελλειπούσες τιμές.

Statistics

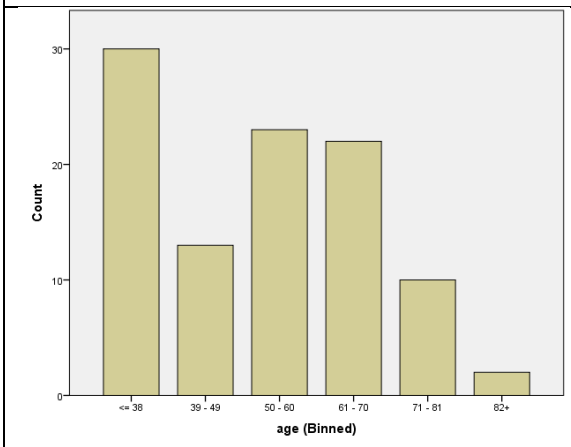
		age	sex	yearpos	voice	sms	data	bundle	resvoice	ressms	resdata
N	Valid	100	100	100	100	100	100	100	100	100	100
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		49,85		4,840	456,00	386,80	519,00		257,57	360,91	215,88
Std. Deviation		17,571		4,4003	319,507	362,566	345,153		282,558	348,320	229,750
Range		74		20,5	900	1170	1100		1197	1172	1125
Minimum		18		,5	300	30	100		0	28	0
Maximum		92		21,0	1200	1200	1200		1197	1200	1125

Statistics

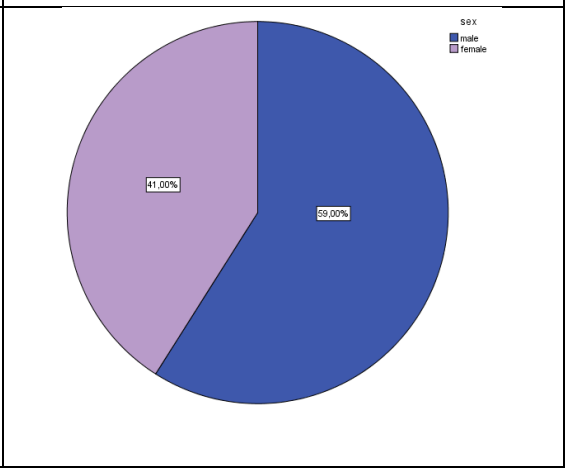
		percentagevoiceuse	percentagesmsuse	percentagedatause
N	Valid	100	100	100
	Missing	0	0	0
Mean		,4982	,0785	,4860
Std. Deviation		,30971	,12785	,44241
Range		1,00	,55	1,00
Minimum		,00	,00	,00
Maximum		1,00	,55	1,00

Από το ιστόγραμμα (σχήμα 1.1) και τον πίνακα με τα περιγραφικά μέτρα παρατηρείται ότι οι συνδρομητές του δείγματος έχουν ηλικία από 18 μέχρι και 92 έτη, με μέσο όρο τα 50 χρόνια. Διαπιστώνεται ότι οι περισσότεροι από τους συνδρομητές βρίσκονται στην κλάση από 18 έως 38 χρονών.

Σχήμα 1.1



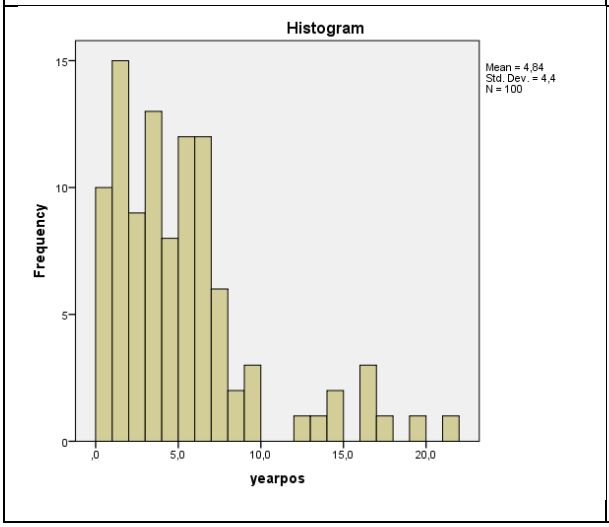
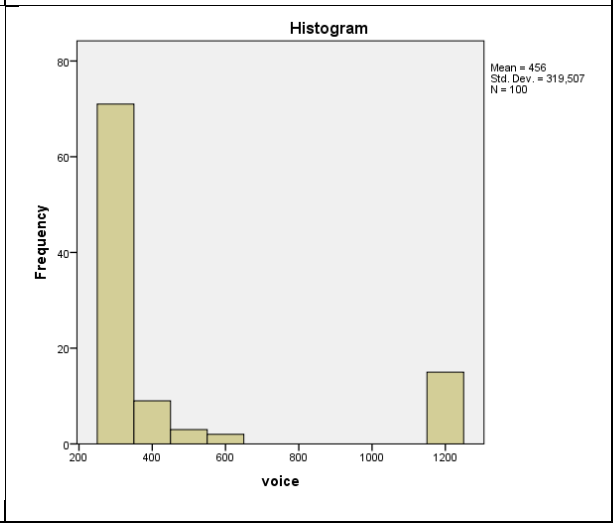
Σχήμα 1.2

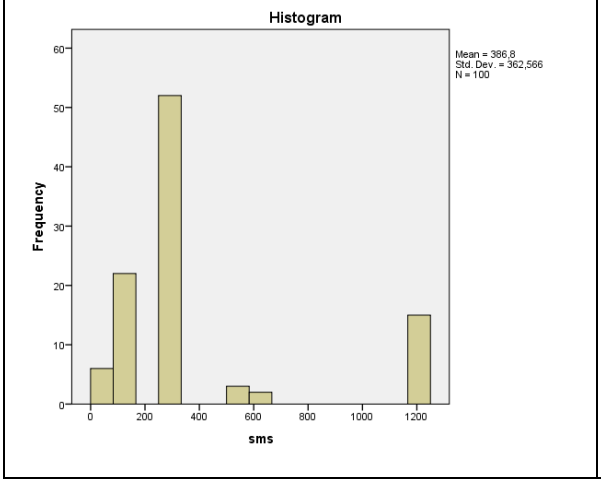
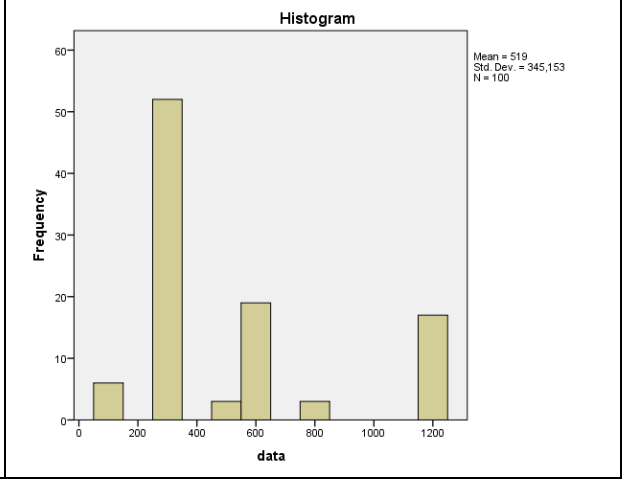


Στο δείγμα υπάρχουν 59 άνδρες και 41 γυναίκες. Δηλαδή, όπως φαίνεται και από το κυκλικό διάγραμμα (σχήμα 1.2) το 59% του δείγματος αποτελείται από άνδρες και το 41% από γυναίκες.

Ο μέσος όρος κατοχής του καρτοκινητού για τους 100 συνδρομητές είναι περίπου πέντε χρόνια.

Παρατηρείται από το γράφημα (σχήμα 1.3) ότι η πλειοψηφία των πελατών έχουν το καρτοκινητό τους μέχρι και εφτά χρόνια.

Σχήμα 1.3	Σχήμα 1.4
	

Σχήμα 1.5	Σχήμα 1.6
	

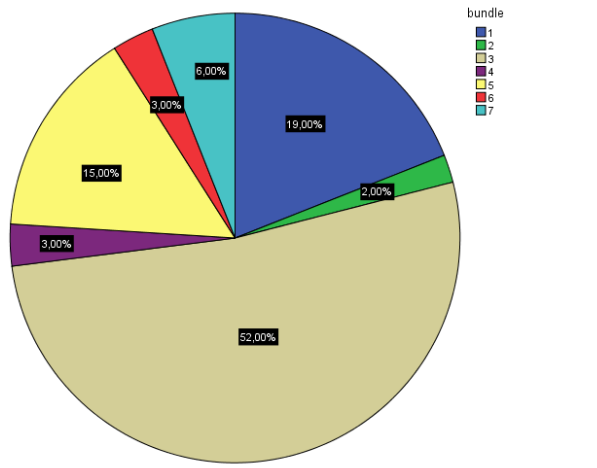
Όπως φαίνεται από τα σχήματα 1.4, 1.5, 1.6 κατά μέσο όρο τα προγράμματα καρτοκινητής που παρέχει η εταιρία προσφέρουν 456 λεπτά ομιλίας, 387 μηνύματα και 519Mb .

Οι περισσότεροι συνδρομητές επιλέγουν πρόγραμμα με 300 λεπτά ομιλίας, 300 μηνύματα και 300 Mb.

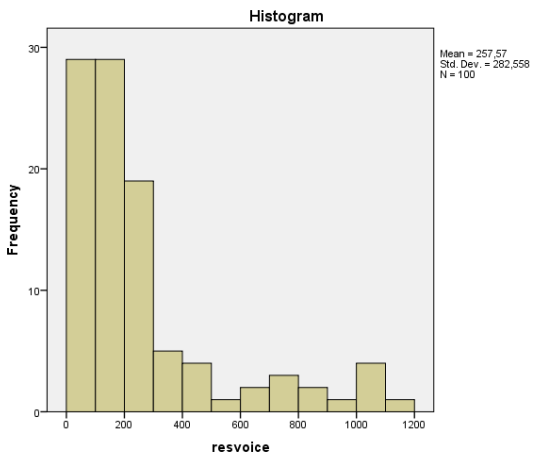
Από το σχήμα 1.7 φαίνεται ότι η πλειοψηφία των συνδρομητών ενεργοποιούν στο καρτοκινητό τους τα προγράμματα ένα, τρία και πέντε. Τα προγράμματα δύο, τέσσερα και έξι τα έχει ενεργοποιήσει πολύ μικρό ποσοστό του δείγματος. Για το λόγο αυτό, τα άτομα του δείγματος που έχουν επιλέξει τα συγκεκριμένα προγράμματα δεν θα

συμπεριληφθούν στην πολυμεταβλητή ανάλυση που θα ακολουθήσει στα επόμενα κεφάλαια.

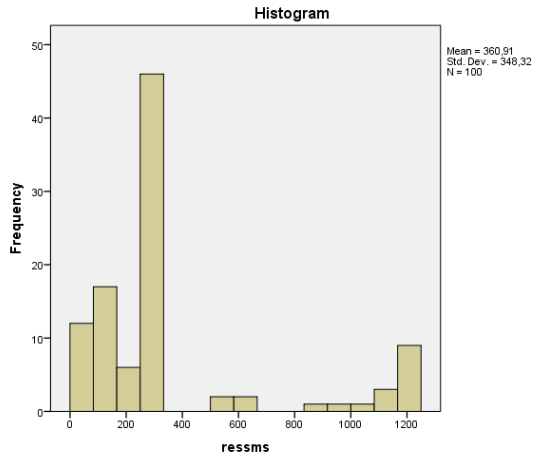
Σχήμα 1.7



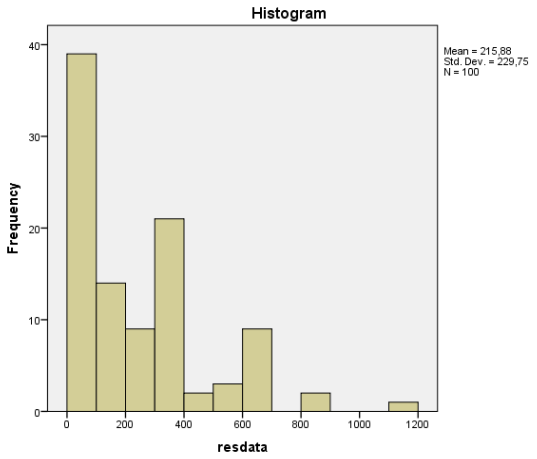
Σχήμα 1.8



Σχήμα 1.9



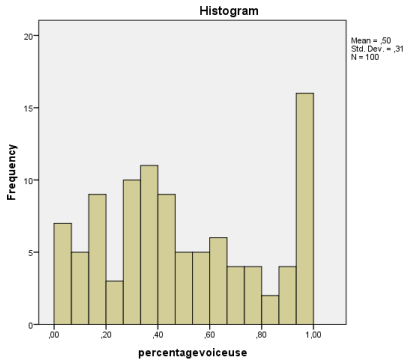
Σχήμα 1.10



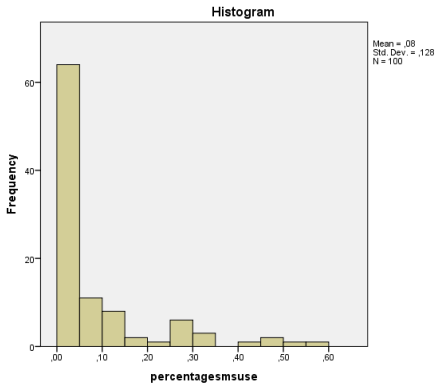
Στον κάθε πελάτη, όπως φαίνεται στα σχήματα 1.8, 1.9,1.10 καθώς και τον πίνακα με τα περιγραφικά μέτρα απομένουν κατά μέσο όρο 283 λεπτά ομιλίας, 348 μηνύματα και 229 mb όταν λήγει το πρόγραμμα του μετά το τέλος των 30 ημερών.

Τέλος, από τα σχήματα 1.11, 1.12, 1.13 και τον πίνακα με τα περιγραφικά μέτρα φαίνεται ότι οι συνδρομητές χρησιμοποιούν κατά μέσο όρο το 50% των λεπτών ομιλίας του προγράμματός τους, το 8% των μηνυμάτων τους και το 49% των δεδομένων τους.

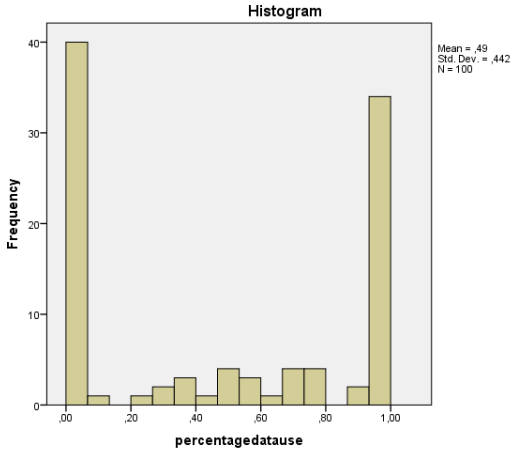
Σχήμα 1.11



Σχήμα 1.2

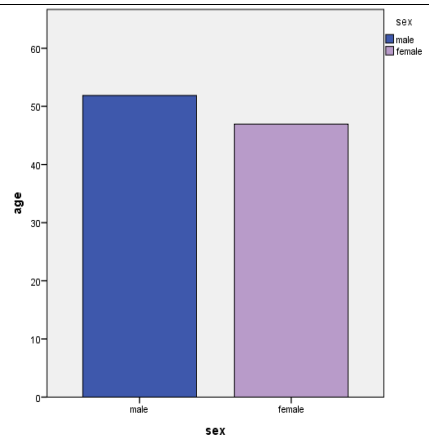
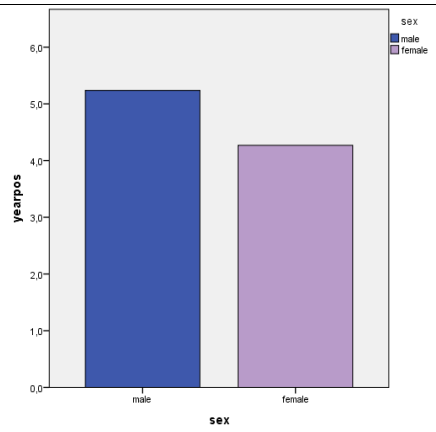
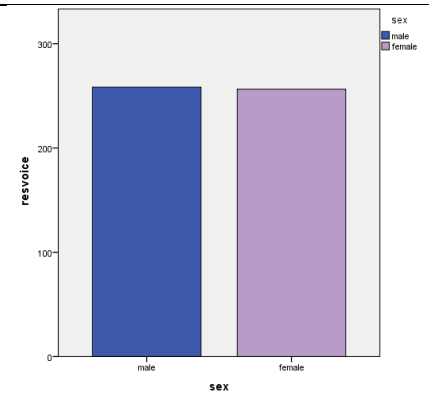
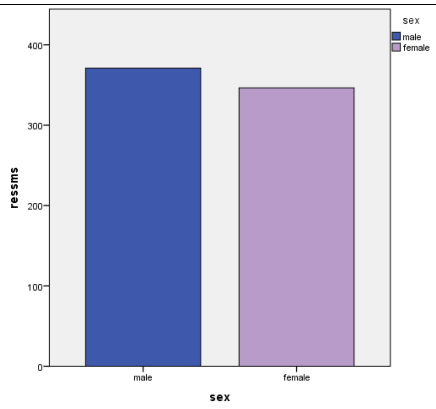


Σχήμα 1.13



Παρατηρείται ότι πολλοί συνδρομητές χρησιμοποιούν όλα τα λεπτά ομιλίας τους και πολλοί δεν χρησιμοποιούν καθόλου μηνύματα. Όσον αφορά στα δεδομένα ίντερνετ, παρατηρούνται μεγάλες συχνότητες σε αυτούς που θα τα χρησιμοποιήσουν όλα αλλά και σε αυτούς που δεν θα χρησιμοποιήσουν καθόλου ίντερνετ.

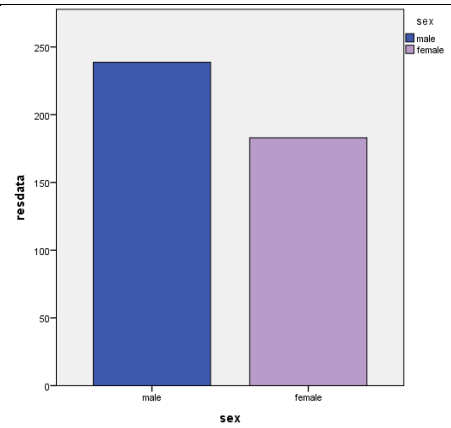
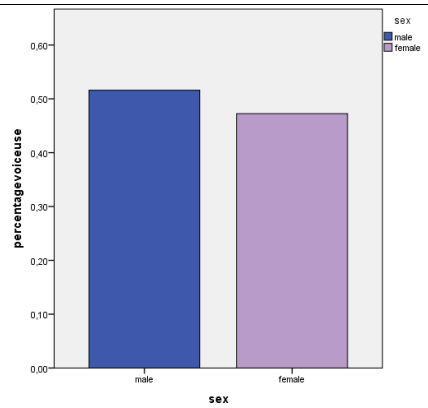
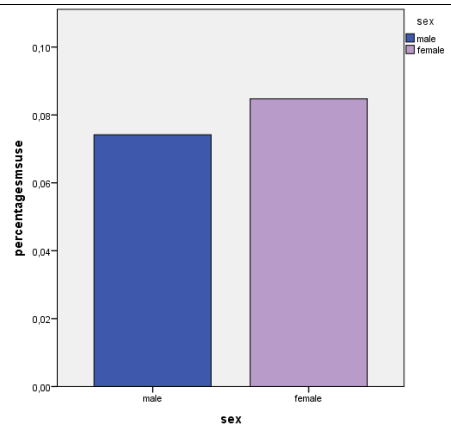
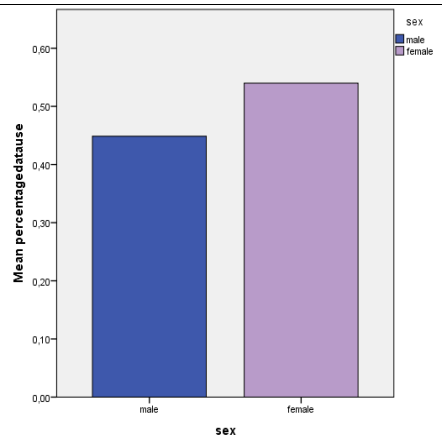
Στη συνέχεια, παρουσιάζονται γραφήματα των μεταβλητών σε σχέση με το φύλο των συνδρομητών.

Σχήμα 1.14	Σχήμα 1.15
	
Σχήμα 1.16	**Σχήμα 1.17**
	

Οι γυναίκες συνδρομήτριες παρουσιάζουν μικρότερη μέση ηλικία σε σχέση με τους άνδρες, όπως φαίνεται στο σχήμα 1.14.

Από το σχήμα 1.15 φαίνεται ότι τα χρόνια κατοχής του καρτοκινητού δεν φαίνεται να διαφοροποιούνται σε σχέση με το φύλο των συνδρομητών. Και οι άνδρες και οι γυναίκες έχουν σχεδόν ίδιο μέσο όρο χρόνων κατοχής, με λίγο χαμηλότερο αυτόν των γυναικών.

Τα υπολειπόμενα λεπτά ομιλίας καθώς και τα υπολειπόμενα μηνύματα δεν φαίνεται να επηρεάζονται από το φύλο του συνδρομητή.(σχήματα 1.16, 1.17)

Σχήμα 1.18	Σχήμα 1.19
	
Σχήμα 1.20	**Σχήμα 1.21**
	

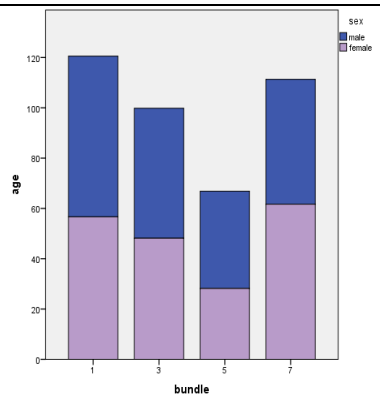
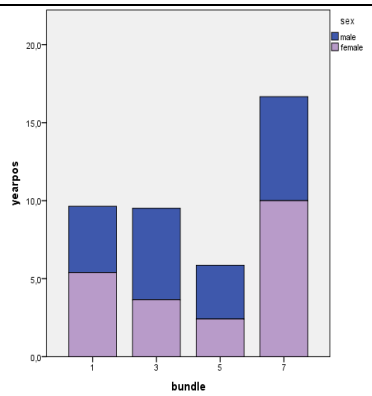
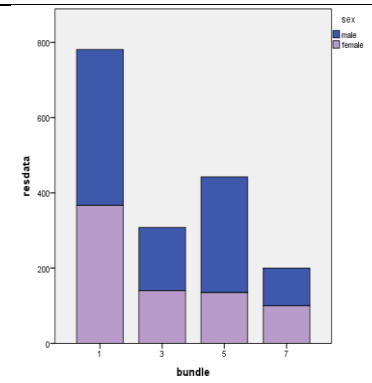
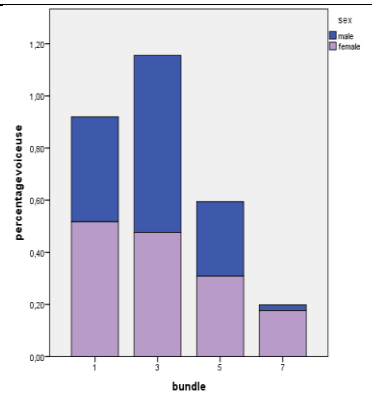
Όσον αφορά στα δεδομένα internet, οι γυναίκες φαίνεται να χρησιμοποιούν σχετικά παραπάνω σε σχέση με τους άνδρες (σχήμα 1.18).

Παρόμοιο ποσοστό χρήσης ομιλίας και μηνυμάτων φαίνεται να έχουν τα δύο φύλα, με λίγο χαμηλότερα το ποσοστό ομιλίας των γυναικών. (σχήματα 1.19, 1.20)

Το ποσοστό χρήσης των δεδομένων των γυναικών φαίνεται να είναι μεγαλύτερο σε σχέση με τους άνδρες, όπως φαίνεται στο σχήμα 1.21.

Στη συνέχεια, παρουσιάζονται τα γραφήματα των μεταβλητών σε σχέση με το πακέτο καρτοκινητής (bundle) που ενεργοποίησαν οι συνδρομητές. Όπως παρουσιάστηκε παραπάνω τα bundles 2,4,6 δεν έχουν ενεργοποιηθεί από

ικανοποιητικό πλήθος συνδρομητών του δείγματος, επομένως θα εξαιρεθούν από την παρακάτω ανάλυση.

Σχήμα 1.22	**Σχήμα 1.23**
	
Σχήμα 1.24	**Σχήμα 1.25**
	

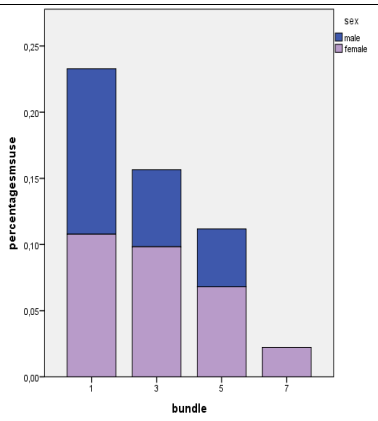
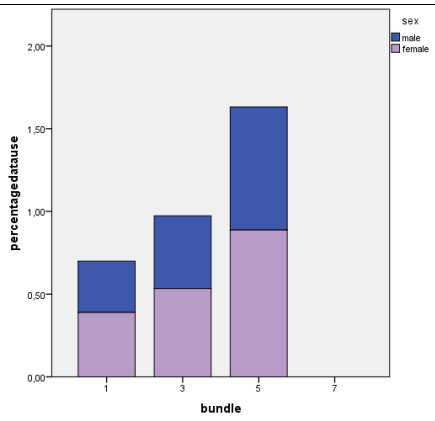
Από το σχήμα 1.22 παρατηρείται ότι οι συνδρομητές που χρησιμοποιούν το bundle 5 έχουν μικρότερη μέση ηλικία από τους υπόλοιπους. Οι γυναίκες φαίνεται να έχουν μικρότερη μέση ηλικία από τους άνδρες, ιδίως στην κατηγορία των ανθρώπων που χρησιμοποιούν το bundle 5.

Από το σχήμα 1.23 παρατηρείται ότι ο μέσος όρος κατοχής του καρτοκινητού των συνδρομητών που χρησιμοποιούν το bundle 7 είναι μεγαλύτερος σε σχέση με αυτών που χρησιμοποιούν τα υπόλοιπα.

Από το σχήμα 1.24 παρατηρείται ότι στους συνδρομητές που χρησιμοποιούν το πρόγραμμα 1 απομένουν περισσότερα δεδομένα ίντερνετ από ότι στους υπόλοιπους παρότι σαν πακέτο έχει λιγότερα δεδομένα από το πρόγραμμα 5. Στις γυναίκες

φαίνεται να απομένει λιγότερο ίντερνετ σε σχέση με τους άνδρες, ειδικά στο πρόγραμμα 5.

Από το σχήμα 1.25 παρατηρείται ότι οι συνδρομητές που χρησιμοποιούν το πρόγραμμα 1 και 3 κάνουν χρήση μεγαλύτερου ποσοστού λεπτών ομιλίας.

Σχήμα 1.26	**Σχήμα 1.27**
	

Από το σχήμα 1.26 παρατηρείται ότι οι συνδρομητές που χρησιμοποιούν το πρόγραμμα 1 κάνουν χρήση μεγαλύτερου ποσοστού μηνυμάτων.

Από το σχήμα 1.27 παρατηρείται ότι οι συνδρομητές που χρησιμοποιούν το πρόγραμμα 5 κάνουν χρήση μεγαλύτερου ποσοστού μηνυμάτων.

Από τον πίνακα συσχετίσεων που φαίνεται παρακάτω προκύπτουν σαν γενική εικόνα τα ακόλουθα συμπεράσματα:

- Η ηλικία των συνδρομητών (age) σχετίζεται με την χρήση των δεδομένων ίντερνετ. Αναλυτικότερα παρουσιάζει θετική συσχέτιση με τη μεταβλητή που δείχνει τα υπολειπόμενα δεδομένα (resdata) και αρνητική συσχέτιση με τη μεταβλητή που δείχνει το ποσοστό χρήσης δεδομένων (percentagedatause). Δηλαδή όσο αυξάνεται η ηλικία τόσο μικρότερη χρήση ίντερνετ γίνεται (μένουν περισσότερα mb) και τόσο μειώνεται το ποσοστό των δεδομένων που έχουν χρησιμοποιηθεί (percentagedatause).
- Ο αριθμός των χρόνων που έχει ο συνδρομητής το καρτοκινητό του δεν σχετίζεται με καμία από τις υπόλοιπες μεταβλητές.

- Το υπόλοιπο του χρόνου ομιλίας (resvoice) σχετίζεται με το υπόλοιπο των μηνυμάτων (resms) και το ποσοστό χρήσης των λεπτών ομιλίας (percentagevoiceuse).
- Το υπόλοιπο των δεδομένων σχετίζεται με το ποσοστό χρήσης των δεδομένων.

Στα επόμενα κεφάλαια περιγράφονται οι στατιστικές μέθοδοι που θα χρησιμοποιηθούν για την στατιστική ανάλυση των δεδομένων, καθώς και η εφαρμογή τους.

Correlation Matrix[a]

		age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
Correlation	age	1,000	,012	-,370	-,419	,461	,009	-,029	-,561
	yearpos	,012	1,000	-,160	-,204	-,106	,029	,088	-,055
	resvoice	-,370	-,160	1,000	,819	-,014	-,625	-,118	,275
	ressms	-,419	-,204	,819	1,000	,005	-,184	-,170	,358
	resdata	,461	-,106	-,014	,005	1,000	,015	-,177	-,704
	percentagevoiceuse	,009	,029	-,625	-,184	,015	1,000	,143	-,037
	percentagesmsuse	-,029	,088	-,118	-,170	-,177	,143	1,000	,206
	percentagedatause	-,561	-,055	,275	,358	-,704	-,037	,206	1,000

a. Determinant = ,014

Κεφάλαιο 2

2.1 Ανάλυση κατά συστάδες

Στόχος της ανάλυσης κατά συστάδες (clustering) είναι η τμηματοποίηση δεδομένων, κατατάσσοντας όμοιες παρατηρήσεις σε ομάδες (συστάδες), βάσει κάποιου αριθμού μεταβλητών.

Η ανάλυση κατά συστάδες χρησιμοποιείται ευρέως από επιχειρήσεις προκειμένου να γίνει ομαδοποίηση των πελατών τους με σκοπό το στοχευμένο marketing (target marketing), δηλαδή την τμηματοποίηση της αγοράς και επικέντρωση στη δημιουργία στρατηγικών για κάθε τμήμα της ξεχωριστά. Στην περίπτωση που μελετάται στα πλαίσια της παρούσας διπλωματικής εργασίας, στόχος είναι η δημιουργία ομάδων συνδρομητών καρτοκινητής, οι οποίοι παρουσιάζουν όμοιες καταναλωτικές συμπεριφορές. Η ομαδοποίηση θα γίνει χρησιμοποιώντας την πληροφορία που συλλέχθηκε από το δείγμα, δηλαδή τις τιμές των μεταβλητών για τον κάθε πελάτη/παρατήρηση.

Η συσταδοποίηση θεωρείται επιτυχημένη όταν οι ομάδες που έχουν δημιουργηθεί αποτελούνται από όσο το δυνατόν πιο όμοιες παρατηρήσεις, ενώ οι διαφορετικές ομάδες παρουσιάζουν τη μεγαλύτερη δυνατή διαφοροποίηση στις παρατηρήσεις.

2.2 Μέτρα Απόστασης

Για να καθοριστεί εάν οι παρατηρήσεις είναι όμοιες ή όχι, χρησιμοποιούνται κατάλληλα μέτρα απόστασης. Τα μέτρα αυτά παίρνουν πολύ μικρές τιμές όταν οι παρατηρήσεις μοιάζουν πολύ μεταξύ τους.

Έστω δείγμα n ατόμων και ότι σε κάθε άτομο παρατηρούνται p χαρακτηριστικά/μεταβλητές, όπου $p \geq 2$.

Από εδώ και στο εξής με $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ θα συμβολίζεται το διάνυσμα των παρατηρήσεων του i ατόμου όπου $i = 1,2,..n$.

α. Ευκλείδεια Απόσταση

Το πιο γνωστό μέτρο απόστασης είναι η ευκλείδεια απόσταση, ο τύπος της οποίας είναι ο:

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

Η ευκλείδεια απόσταση εξαρτάται από την κλίμακα μέτρησης των μεταβλητών.

β. Απόσταση του Pearson

Για το λόγο αυτόν, χρησιμοποιείται εναλλακτικά η απόσταση του Pearson, η οποία δίνεται από τον τύπο:

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

Παρακάτω παρουσιάζονται κάποια επιπλέον μέτρα αποστάσεων που χρησιμοποιούνται για συνεχή δεδομένα:

γ. Απόσταση Manhattan ή City-block metric

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

δ. Απόσταση Minkowski

$$d_{ij} = d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}|^{\lambda}\right)^{\frac{1}{\lambda}},$$ όπου $\lambda \geq 1$ δεδομένη παράμετρος.

ε. Απόσταση max ή Chebyshev

$$d_{ij} = d(\boldsymbol{x_i},\boldsymbol{x_j}) = \max_{r=1,2,\dots,p} |x_{ir} - x_{jr}|.$$

στ. Απόσταση Mahalanobis

$$d_{ij} = d(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\boldsymbol{x_i} - \boldsymbol{x_j})' \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}),$$

όπου Σ είναι ο δειγματικός πίνακας διακύμανσης- συνδιακύμανσης που αντιστοιχεί στα διανύσματα $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})$ και $\boldsymbol{x_j} = (x_{j1}, x_{j2}, \dots, x_{jp})$.

Εναλλακτικά θα μπορούσαν να χρησιμοποιηθούν και τα μέτρα ομοιότητας, για να διαπιστωθεί αν δυο άτομα είναι όμοια μεταξύ τους, αλλά στα πλαίσια της παρούσας διπλωματικής δεν θα αναλυθούν.

2.3 Μέθοδοι Ομαδοποίησης

Υπάρχουν διάφοροι τρόποι προσέγγισης της ομαδοποίησης των δεδομένων, όπως γραφικές μέθοδοι (π.χ. star plots, Chernoff faces) ή συστηματικές μέθοδοι (ιεραρχικές ή μη ιεραρχικές μέθοδοι). Οι τελευταίες, οι οποίες έχουν κάποια μαθηματική βάση, θα παρουσιαστούν στη συνέχεια.

Έστω ότι υπάρχει δείγμα *n* ατόμων και στόχος είναι να ομαδοποιηθούν σε *k* ομάδες.

Στις **ιεραρχικές μεθόδους** ο σχηματισμός των ομάδων γίνεται σταδιακά, παράγοντας μία ιεραρχία δενδροειδούς μορφής όπου στα διάφορα στάδια το πλήθος *k* των ομάδων παίρνει όλες τις δυνατές τιμές από το 1 έως το *n*.

Οι ιεραρχικές μέθοδοι χωρίζονται σε δύο κατηγορίες, τις συσσωρευτικές και διαιρετικές μεθόδους.

Οι συσσωρευτικές μέθοδοι ξεκινώντας από *n* ομάδες, δηλαδή κάθε ομάδα να αποτελείται από ένα άτομο, καταλήγουν έπειτα από διαδοχικές συγχωνεύσεις σε μία ομάδα που αποτελείται από όλα τα άτομα.

Με τις διαιρετικές μεθόδους συντελείται η αντίθετη διαδικασία, δηλαδή ξεκινώντας από μία μόνο ομάδα πλήθους *n* ατόμων, που αποτελείται δηλαδή από όλα τα άτομα, καταλήγουν έπειτα από διαδοχικές διαιρέσεις σε n ομάδες όπου η κάθε μια αποτελείται από μόνο ένα άτομο. Οι διαιρετικές μέθοδοι απαιτούν πολύ περισσότερους υπολογισμούς από τις συσσωρευτικές, γεγονός που τις καθιστά πρακτικά δύσχρηστες. Για το λόγο αυτό δεν θα υπάρξει περαιτέρω ανάλυση για αυτές στην παρούσα διπλωματική.

Μειονέκτημα των ιεραρχικών μεθόδων (κυρίως των διαιρετικών) είναι ότι απαιτούν πολύ χρόνο, μεγάλη μνήμη και υπολογιστική ισχύ για να ολοκληρωθούν, καθώς και ότι ορισμένες φορές δημιουργούνται ομάδες με ανομοιογενές μέγεθος.

Στις **μη ιεραρχικές μεθόδους** τα *n* άτομα ομαδοποιούνται σε *k* ομάδες, όπου το *k*, δηλαδή ο αριθμός των ομάδων έχει καθοριστεί εκ των προτέρων.

Οι επαναληπτικοί αλγόριθμοι που χρησιμοποιούνται στις μη ιεραρχικές μεθόδους είτε θα τοποθετήσουν τις παρατηρήσεις γύρω από *k* συγκεκριμένα άτομα (μητρικά σημεία) δημιουργώντας έτσι τις ομάδες είτε θα ξεκινήσουν με έναν αρχικό διαμερισμό των παρατηρήσεων σε *k* ομάδες και θα μετακινούν τις παρατηρήσεις μέχρι να επιτευχθεί ο καλύτερος δυνατός διαμερισμός.

Όλοι οι αλγόριθμοι των μη ιεραρχικών μεθόδων χρησιμοποιούν την έννοια του κέντρου μιας ομάδας (κέντρου βάρους), δηλαδή τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Αναλυτικότερα, κατά τη διάρκεια της μεθόδου η κατάταξη στις ομάδες γίνεται βάσει της απόστασης (συνήθως ευκλείδειας) κάθε παρατήρησης από το κέντρο μιας ομάδας. Οι αλγόριθμοι διαφοροποιούνται στον τρόπο με τον οποίο γίνεται η ανανέωση των κέντρων των ομάδων και στο πως τοποθετούνται οι παρατηρήσεις σε αυτές.

2.3.1. Ιεραρχικές Μέθοδοι-Συσσωρευτικές μέθοδοι

Γίνεται η υπόθεση ότι υπάρχει δείγμα που αποτελείται από n άτομα : $x_1, x_2, \ldots, x_n$.

Έχει γίνει υπολογισμός των αποστάσεων τους $d_{ij} = d(x_i,x_j)$ με $i,j=1,2,..,n$ οι οποίες έχουν τοποθετηθεί σε έναν πίνακα αποστάσεων $D=[\, d_{ij}]$ ($n \times n$ διαστάσεων)

Η λογική των αλγορίθμων των συσσωρευτικών μεθόδων είναι η εξής:

Αρχικά κάθε άτομο αποτελεί μία ομάδα. Γίνεται εντοπισμός στον πίνακα αποστάσεων του ζεύγους των πλησιέστερων ομάδων. Οι δύο αυτές ομάδες συγχωνεύονται σε μία με αποτέλεσμα ο αριθμός των ομάδων να μειώνεται κατά ένα και στη συνέχεια γίνεται επαναυπολογισμός του πίνακα των αποστάσεων. Στον νέο πίνακα αποστάσεων έχουν διαγραφεί οι γραμμές και οι στήλες που αντιστοιχούσαν στις δύο ομάδες που συγχωνεύτηκαν και έχει προστεθεί μια νέα γραμμή και στήλη που περιέχει τις αποστάσεις της νέας ομάδας που σχηματίστηκε από τις υπόλοιπες. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να έχει δημιουργηθεί μία μόνο ομάδα η οποία περιλαμβάνει όλα τα άτομα.

Παρακάτω περιγράφονται οι συνηθέστερες τεχνικές με τις οποίες μπορεί να υπολογιστεί η απόσταση μιας ομάδας, που έχει δημιουργηθεί είτε από συγχώνευση άλλων ομάδων είτε από συγχώνευση παρατηρήσεων με κάποια άλλη ομάδα.

α .Μέθοδος της απλής συνένωσης (Single Linkage Method)

Στη μέθοδο αυτή, η οποία είναι γνωστή και ως μέθοδος του πλησιέστερου γείτονα – nearest neighbor method, ως απόσταση μεταξύ δύο ομάδων ορίζεται η μικρότερη απόσταση από μία παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Έστω ότι υπάρχουν οι ομάδες R και Q. Τότε η απόστασή τους $d(R,Q)$ ορίζεται από τον τύπο:

$$d(R,Q) = \min_{i \epsilon R, j \in Q} d_{ij}.$$

Όταν δύο ομάδες A και B ενωθούν για να δημιουργήσουν μια νέα ομάδα, έστω $R=(A,B)$, τότε όλες οι αποστάσεις μεταξύ της R και των άλλων ομάδων, έστω Q, προκύπτουν από τον παρακάτω τύπο που χρησιμοποιείται στην ανανέωση του πίνακα αποστάσεων:

$$d(R, Q) = min\{d(A,Q), d(B,Q)\} = \frac{1}{2}(d(A,Q) + d(B,Q)) - \frac{1}{2}|d(A,Q) - d(B,Q)|.$$

Η μέθοδος της απλής συνένωσης αποτελεί την πιο παλιά και απλή όλων των ιεραρχικών μεθόδων. Βασικότερο μειονέκτημα της μεθόδου είναι ότι ο αλγόριθμος δεν μπορεί να διαχωρίσει δύο ομάδες εμφανώς διαφορετικές, οι οποίες όμως έχουν κάποιο σημείο ή σύνολο σημείων που τις «συνδέει» (φαινόμενο της αλυσίδας).

β. Μέθοδος της πλήρους συνένωσης (Complete Linkage Method)

Στη μέθοδο αυτή, η οποία είναι γνωστή και ως μέθοδος του μακρινότερου γείτονα– furthest neighbor method, ως απόσταση μεταξύ δύο ομάδων ορίζεται η μεγαλύτερη απόσταση μίας παρατήρησης της μιας ομάδας με κάποια παρατήρηση από την άλλη ομάδα, δηλαδή

$$d(R,Q) = \max_{i \epsilon R, j \in Q} d_{ij}.$$

Ο τύπος ανανέωσης των αποστάσεων της μεθόδου αυτής έχει τη μορφή:

$$d(R,Q) = max\{d(A,Q), d(B,Q)\} = \frac{1}{2}(d(A,Q) + d(B,Q)) + \frac{1}{2}|d(A,Q) - d(B,Q)|.$$

Αδυναμία της μεθόδου αυτής είναι ότι ο αλγόριθμος μπορεί να μην καταφέρει να ενώσει ομάδες οι οποίες αν και περιέχουν όμοια στοιχεία, περιέχουν έστω και ένα ζευγάρι σημείων που βρίσκονται αρκετά μακριά μεταξύ τους.

γ. Μέθοδος των σταθμισμένων μέσων (Weighted Average Linkage Method)

Στη μέθοδο αυτή ως απόσταση ορίζεται ο μέσος των αποστάσεων όλων των στοιχείων μιας ομάδας με τα στοιχεία της άλλης.

Ο τύπος ανανέωσης των αποστάσεων της μεθόδου αυτής έχει τη μορφή:

$$d(R, Q) = \frac{|A|\, d(A,Q) + |B|\, d(B,Q)}{|A| + |B|}.$$

δ. Μέθοδος των κέντρων βάρους (Centroid Method)

Στη μέθοδο αυτή ως απόσταση ορίζεται η απόσταση των κέντρων των ομάδων.

Αν x_{ir} είναι η r μέτρηση του i αντικειμένου $\boldsymbol{x_i}$, όπου το i παίρνει τιμές από 1 έως n και το r από 1 έως p, το κέντρο βάρους μιας ομάδας R θα είναι το σημείο:

$$\bar{x}(R) = \left(\bar{x}_1(R), \bar{x}_2(R), \ldots, \bar{x}_p(R)\right),$$

το οποίο έχει ως r συντεταγμένη την:

$$\overline{x_r}(R) = \frac{1}{|R|}\sum_{i\in R} x_{ir} \quad \text{για } r=1,2,\ldots,p.$$

Η απόσταση μεταξύ των ομάδων R και Q ορίζεται ως η ευκλείδεια απόσταση μεταξύ των κέντρων βάρους τους, δηλαδή:

$$d(R,Q) = d\left(\bar{x}(R), \bar{x}(Q)\right) = \sqrt{\sum_{r=1}^{p}\left(\bar{x}_r(R) - \bar{x}_r(Q)\right)^2}.$$

Ο τύπος ανανέωσης του πίνακα αποστάσεων είναι ο εξής:

$$d^2(R,Q) = \frac{|A|}{|R|}d^2(A,Q) + \frac{|B|}{|R|}d^2(B,Q) + \frac{|A||B|}{|R|^2}d^2(A,B).$$

Η μέθοδος αυτή παράγει συνήθως ομάδες συμπαγείς και ελλειπτικές. Μπορεί να εφαρμοστεί μόνο σε ποσοτικά δεδομένα, λόγω της χρήσης της ευκλείδειας απόστασης.

ε. Μέθοδος του Ward (Ward's method)

Η μέθοδος αυτή έχει στόχο να ελαχιστοποιήσει τη διακύμανση μέσα στις ομάδες.

Για κάθε παρατήρηση υπολογίζεται η απόστασή της από το κέντρο της ομάδας, χρησιμοποιώντας τις περισσότερες φορές την ευκλείδεια απόσταση. Οι αποστάσεις αυτές αθροίζονται για όλα τα άτομα μιας ομάδας C και προκύπτει η ποσότητα :

$$ESS(C) = \sum_{i\in C}\left(d\left(\boldsymbol{x_i}, \bar{x}(C)\right)\right)^2.$$

Η ποσότητα αυτή ονομάζεται άθροισμα των τετραγωνικών αποκλίσεων (τετραγώνων των αποστάσεων) για την ομάδα C και χρησιμοποιείται ως μέτρο συνεκτικότητας των στοιχείων της ομάδας.

Έστω ότι υπάρχουν k ομάδες. Το συνολικό άθροισμα τετραγωνικών αποκλίσεων, που προκύπτει αν προστεθούν τα αθροίσματα των τετραγωνικών αποκλίσεων για όλες τις ομάδες είναι ίσο με :

$$ESS = ESS_1 + ESS_2 + \ldots + ESS_k.$$

Στην αρχή, δηλαδή όταν κάθε παρατήρηση αποτελεί και μια ομάδα και επομένως η απόστασή της από το κέντρο της είναι 0, το άθροισμα ESS είναι ίσο με 0.

Ο αλγόριθμος, έπειτα από δοκιμές όλων των δυνατών συγχωνεύσεων ανά δύο, τελικά συγχωνεύει τις δύο ομάδες που η ομαδοποίησή τους οδηγεί στη μικρότερη αύξηση του *ESS* (ελάχιστη απώλεια πληροφορίας).

Στο τελικό βήμα του αλγορίθμου όταν τα *n* άτομα θα έχουν συγχωνευτεί σε μια ομάδα το *ESS* ισούται με :

$$ESS=\sum_{j=1}^{N}(\boldsymbol{x_j}-\bar{\boldsymbol{x}})'(\boldsymbol{x_j}-\bar{\boldsymbol{x}}),$$

όπου x_j είναι η πολυδιάστατη μέτρηση του *j*-οστού ατόμου και $\bar{x}$ η μέση τιμή όλων των ατόμων.

Η απόσταση μεταξύ δύο ομάδων υπολογίζεται από τον τύπο :

$$d^2(R,Q)=\frac{2|R||Q|}{|R|+|Q|}d^2\big(\bar{x}(R),\bar{x}(Q)\big),$$

ενώ ο τύπος ανανέωσης έχει τη μορφή :

$$d^2(R,Q)=\frac{1}{2}d^2(A,Q)+\frac{1}{2}d^2(B,Q)-\frac{1}{4}d^2(A,B).$$

Η μέθοδος αυτή δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων και χρησιμοποιείται πολύ συχνά στην πράξη.

Τα βήματα των ιεραρχικών μεθόδων απεικονίζονται συνήθως γραφικά με τη χρήση κάποιου **δενδρογράμματος**. Στο δενδρόγραμμα φαίνονται οι ομάδες που έχουν σχηματιστεί στα διάφορα βήματα της μεθόδου, καθώς και οι αποστάσεις στις οποίες έγιναν οι συνενώσεις. Τα δενδρογράμματα είναι χρήσιμα και για την επιλογή του βέλτιστου αριθμού ομάδων. Αναλυτικότερα, στο σημείο του δενδρογράμματος που παρατηρείται η μεγαλύτερη απόσταση για το επόμενο επίπεδο συνένωσης αν χαραχθεί μια ευθεία γραμμή κάθετη στον άξονα των αποστάσεων, αυτή θα τέμνει το δενδρόγραμμα σε *k* σημεία. Το *k* αποτελεί το βέλτιστο πλήθος ομάδων.

Οι παραπάνω πληροφορίες για τη θεωρία της ανάλυσης των ιεραρχικών μεθόδων έχουν αντληθεί από τις σημειώσεις του Μεταπτυχιακού Εφαρμοσμένης Στατιστικής του Πανεπιστημίου Πειραιώς, Κούτρας (2020), όπου ο αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες.

Για περισσότερες πληροφορίες σχετικά με την θεωρία των ιεραρχικών μεθόδων, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Johnson και Wichern (1998), στο βιβλίο του Manly, (1986) και στο βιβλίο των Everitt και Dunn, (1991).

Για περισσότερες πληροφορίες σχετικά με την θεωρία των αποστάσεων, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο του Manly, (1986). Συμπληρωματικά, μπορεί να ανατρέξει στην ιστοσελίδα http://www.norusis.com/pdf/SPC_v13.pdf.

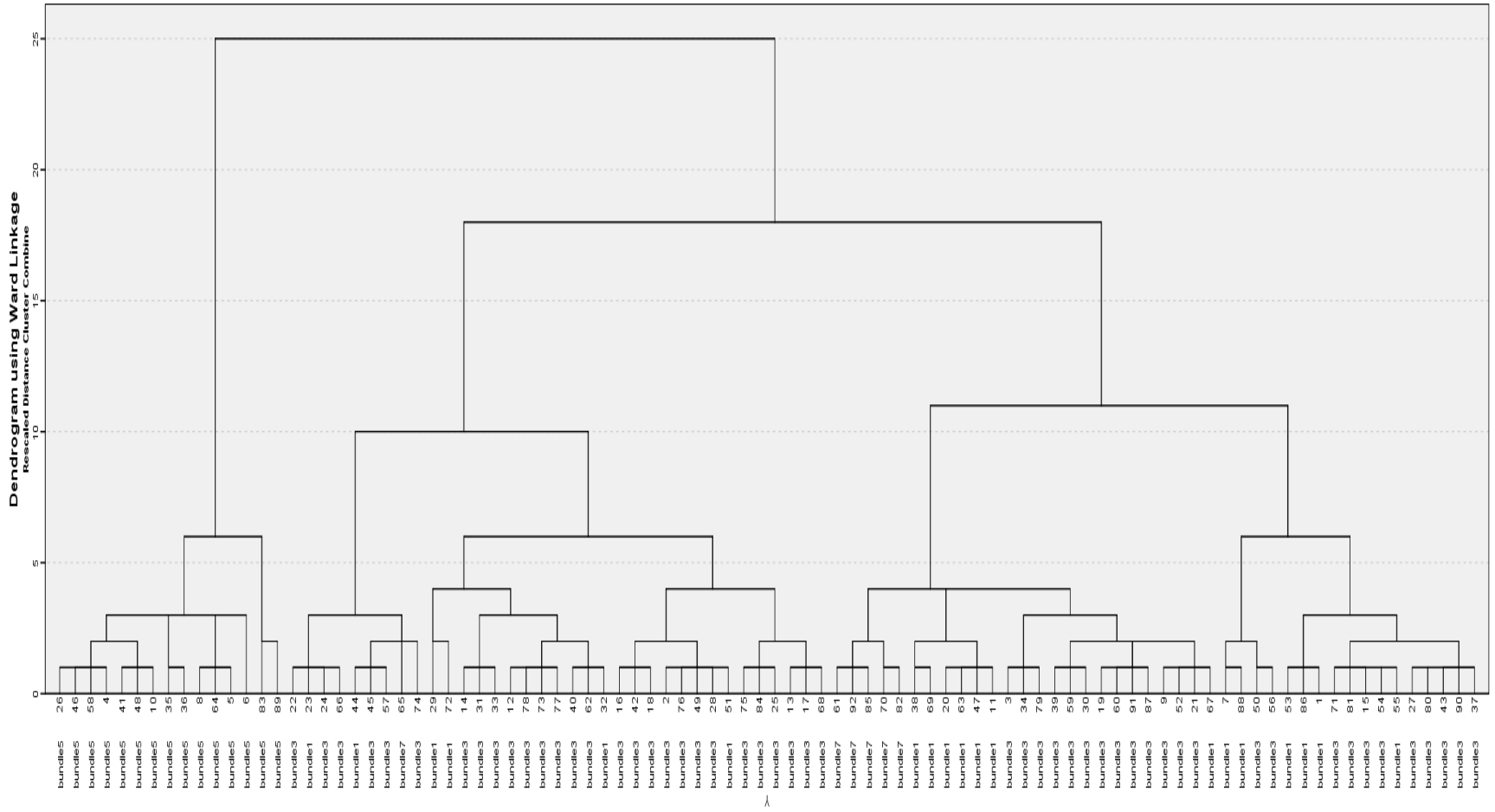
2.3.2 Εφαρμογή Ιεραρχικής μεθόδου

Στη συνέχεια έχει εφαρμοστεί στα δεδομένα η μέθοδος Ward, για τις ποσοτικές μεταβλητές (age, yearpos, resvoice, ressms, resdata, percentagevoiceuse, percentagesmsuse, percentagedatause) με τη χρήση της απόστασης του Pearson, δεδομένου ότι οι μεταβλητές θα πρέπει να τυποποιηθούν, διότι είναι μετρημένες σε διαφορετικές κλίμακες. Όπως αναφέρθηκε παραπάνω, για την ανάλυση δεν έχουν χρησιμοποιηθεί οι παρατηρήσεις- συνδρομητές που έχουν χρησιμοποιήσει τα πακέτα 2,4,6, διότι ήταν πολύ μικρό το δείγμα, άρα όχι αντιπροσωπευτικό. Έχει επιλεχθεί για τη γραφική απεικόνιση μέσω του δενδρογράμματος να φαίνεται στον οριζόντιο άξονα το πρόγραμμα που έχει επιλέξει ο κάθε συνδρομητής.

Από το δενδρόγραμμα που παρουσιάζεται (δενδρόγραμμα 1) στην επόμενη σελίδα φαίνεται ότι ο κατάλληλος αριθμός ομάδων είναι τρεις, αφού εκεί παρουσιάζεται η μεγαλύτερη απόσταση σε σχέση με τις αποστάσεις των άλλων επιπέδων συνένωσης. Η πρώτη ομάδα αποτελείται από 42 παρατηρήσεις, η δεύτερη από 35 και η τρίτη από 15. Στην τελευταία ομάδα κατατάχθηκαν οι παρατηρήσεις- συνδρομητές που χρησιμοποιούν το bundle 5. Στο πρώτο cluster κατατάχθηκαν 13 συνδρομητές που ενεργοποίησαν το bundle 1, 24 συνδρομητές που ενεργοποίησαν το bundle 2 και 5 συνδρομητές που ενεργοποίησαν το bundle 7. Στο δεύτερο cluster κατατάχθηκαν 6 συνδρομητές που ενεργοποίησαν το bundle 1, 28 συνδρομητές που ενεργοποίησαν το bundle 2 και 1 συνδρομητής που ενεργοποίησε το bundle 7. Στο τρίτο cluster κατατάχθηκαν οι 15 συνδρομητές που ενεργοποίησαν το bundle 5.

	cluster			
bundle	**1**	**2**	**3**	**Άθροισμα**
1	13	6		19
3	24	28		52
5			15	15
7	5	1		6
Γενικό Άθροισμα	**42**	**35**	**15**	**92**

Δενδρόγραμμα 1



Στον πίνακα 2.1 παρουσιάζονται οι μέσες τιμές της κάθε μεταβλητής για τα τρία clusters που έχουν δημιουργηθεί. Παρατηρείται ότι οι συνδρομητές που ανήκουν στο πρώτο cluster έχουν μεγαλύτερη μέση ηλικία, ενώ αυτοί που ανήκουν στο τρίτο τη μικρότερη μέση ηλικία. Για τα άτομα του δεύτερου cluster φαίνεται ότι κατέχουν περισσότερα χρόνια το καρτοκινητό τους. Κάποιο εμφανές συμπέρασμα δεν προκύπτει από τη μέση τιμή του υπόλοιπου του χρόνου ομιλίας, των μηνυμάτων ή του ίντερνετ αν γίνει σύγκριση ανάμεσα στα clusters, διότι αυτά εξαρτώνται από τα αρχικά λεπτά, μηνύματα ή δεδομένα του πακέτου αντίστοιχα. Για παράδειγμα, εάν το πακέτο είχε πάρα πολλά λεπτά ομιλίας, είναι πολύ πιθανό ενώ ο συνδρομητής έχει μιλήσει πολλή ώρα, να του υπολείπονται ακόμα πολλά λεπτά λόγω του αρχικού μεγέθους του πακέτου. Με την ίδια λογική προκύπτει το συμπέρασμα και για τα υπολειπόμενα μηνύματα και δεδομένα. Μόνο για το cluster 3, δηλαδή το bundle 5, φαίνεται ότι ενώ σαν πακέτο έχει τα περισσότερα δεδομένα ίντερνετ σε σχέση με τα άλλα, οι συνδρομητές που το χρησιμοποιούν αφήνουν πολύ λίγα υπολειπόμενα. Επιπλέον, οι συνδρομητές που ανήκουν στο cluster 3 χρησιμοποιούν μικρότερο ποσοστό των λεπτών ομιλίας τους από τους υπόλοιπους. Οι συνδρομητές που ανήκουν στο cluster 2 χρησιμοποιούν λίγο μεγαλύτερο ποσοστό των μηνυμάτων του πακέτου τους σε σχέση με τους υπόλοιπους. Τέλος, παρατηρείται ότι οι συνδρομητές που ανήκουν στα clusters 2 και 3 χρησιμοποιούν πολύ μεγάλο ποσοστό των δεδομένων τους, ενώ αυτοί του 1 πολύ μικρό ποσοστό.

Πίνακας 2.1

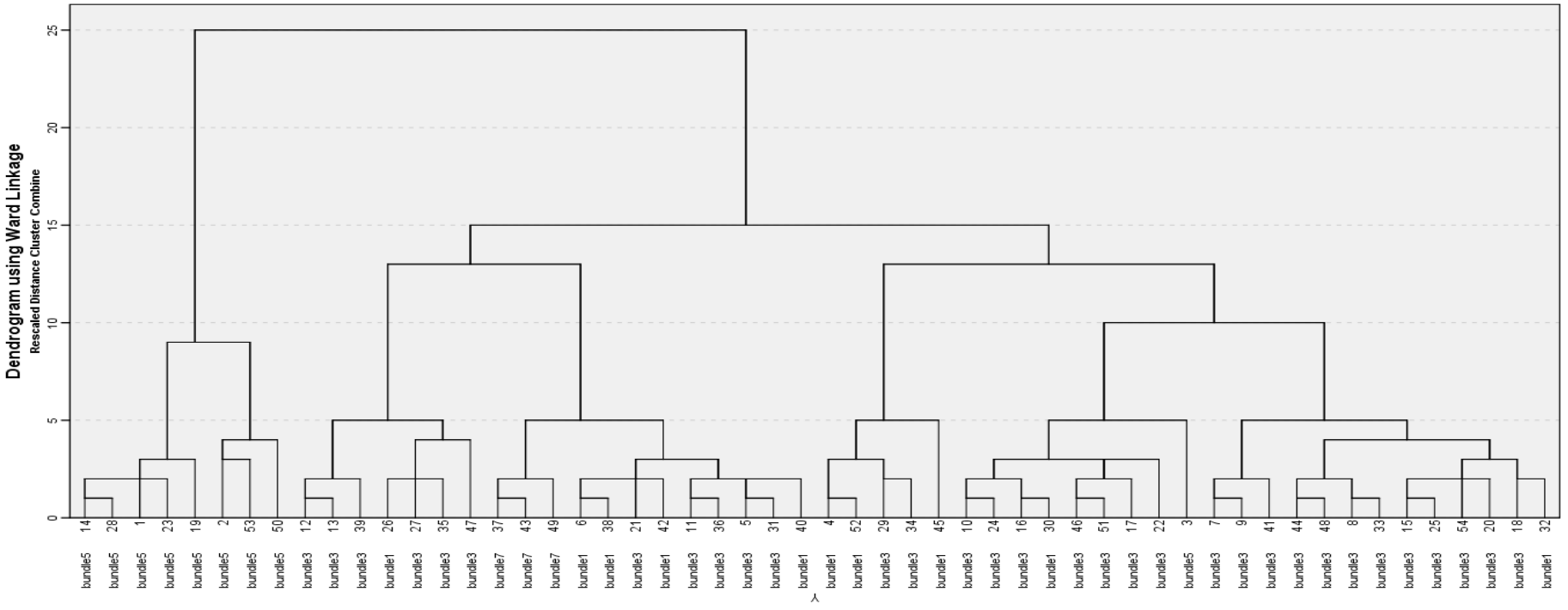
Mean

Ward Method	age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
1	60,12	3,964	156,57	194,10	336,93	,5152	,0508	,0778
2	45,00	6,714	140,17	233,03	58,97	,5397	,1137	,8023
3	34,47	3,033	846,27	1135,87	238,40	,2948	,0534	,8013
Total	50,18	4,859	262,78	362,46	215,12	,4886	,0752	,4714

Στη συνέχεια θα εφαρμοστεί η ιεραρχική μέθοδος συσταδοποίησης ξεχωριστά για τους άνδρες και ξεχωριστά για τις γυναίκες.

Ακολουθώντας ακριβώς την ίδια διαδικασία με παραπάνω (μέθοδος Ward, χρήση της απόστασης του Pearson λόγω ανάγκης τυποποίησης για τις ποσοτικές μεταβλητές, age, yearpos, resvoice, ressms, resdata, percentagevoiceuse, percentagesmsuse, percentagedatause), αλλά επιλέγοντας μόνο τους άνδρες του δείγματος, προκύπτει το δενδρόγραμμα 2.

Δενδρόγραμμα 2



Φαίνεται ότι οι συνδρομητές χωρίζονται σε δύο ομάδες, αυτούς που χρησιμοποιούν το bundle 5 (εκτός από μία παρατήρηση) και στους υπόλοιπους. Στον πίνακα 2.2 παρουσιάζονται οι μέσες τιμές της κάθε μεταβλητής για τα δύο clusters.(δενδρόγραμμα 2)

Πίνακας 2.2

Mean

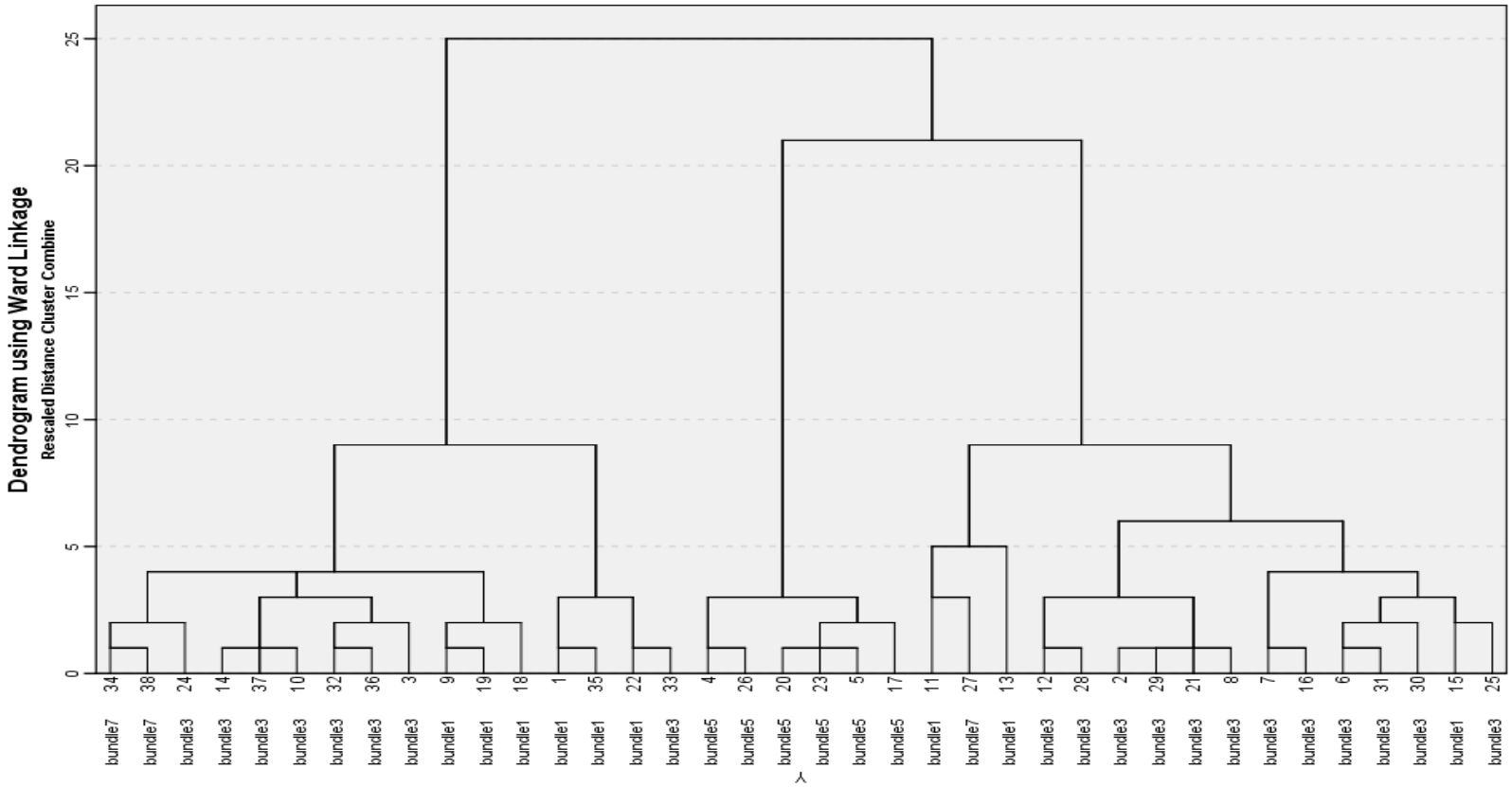
Ward Method	age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
1	34,13	3,375	897,25	1141,50	345,88	,2523	,0488	,7118
2	54,72	5,522	143,26	243,54	213,46	,5731	,0677	,3950
Total	51,67	5,204	254,96	376,57	233,07	,5256	,0649	,4419

Παρατηρείται ότι στο cluster 1, το οποίο αποτελείται από τους άνδρες συνδρομητές που ενεργοποιούν το πρόγραμμα 5 εκτός από έναν, ανήκουν άνδρες μικρότερης ηλικίας, που έχουν το καρτοκινητό στην κατοχή τους λιγότερα χρόνια κατά μέσο όρο, χρησιμοποιούν μικρότερο ποσοστό των λεπτών ομιλίας τους και πολύ μεγάλο ποσοστό των δεδομένων τους.

Ο συνδρομητής που χρησιμοποιεί το bundle 5, αλλά δεν κατατάχθηκε στο cluster 1 έχει μεγαλύτερη ηλικία (75 έτη) και έχει χρησιμοποιήσει μεγάλο ποσοστό των λεπτών ομιλίας του (55%) σε σχέση με το μέσο όρο των συνδρομητών που έχουν καταταχθεί στο cluster 1.

Το αντίστοιχο δενδρόγραμμα για τις γυναίκες συνδρομήτριες είναι το παρακάτω (πάλι ακολουθείται η ίδια διαδικασία αλλά μόνο για το δείγμα των γυναικών).

Δενδρόγραμμα 3



Σε αυτήν την περίπτωση φαίνεται ότι διαμορφώνονται τρία clusters (δενδρόγραμμα 3). Στον πίνακα 2.3 παρουσιάζονται οι μέσες τιμές της κάθε μεταβλητής για τα τρία clusters που έχουν δημιουργηθεί.

Πίνακας 2.3

Mean

Ward Method	age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
1	60,94	3,656	176,81	183,81	367,31	,4469	,0256	,0360
2	42,69	5,813	162,69	210,00	32,44	,4729	,1621	,8502
3	28,17	2,417	829,33	1118,33	134,83	,3089	,0681	,8876
Total	48,08	4,368	273,89	342,39	189,61	,4361	,0898	,5133

Τα συμπεράσματα που προκύπτουν για τις γυναίκες συνδρομήτριες είναι ίδια με τα συμπεράσματα που περιγράφηκαν παραπάνω για το σύνολο του δείγματος (και άνδρες και γυναίκες).

Φαίνεται ότι η διαφορά του δείγματος των ανδρών σε σχέση με αυτό των γυναικών (καθώς και με το συνολικό δείγμα) είναι ότι στους άνδρες δεν διαμορφώνεται μεσαία ηλικιακή ομάδα, η οποία κάνει υψηλή χρήση και ομιλίας και δεδομένων. Έτσι προκύπτει και η διαφοροποίηση στον αριθμό των ομάδων που δημιουργούνται ανάμεσα στα δείγματα.

Λαμβάνοντας υπόψιν τα αποτελέσματα της ανάλυσης σε συστάδες με την ιεραρχική μέθοδο, η εταιρία κινητής τηλεφωνίας θα πρέπει να διαμορφώσει τα προγράμματα της καρτοκινητής ανάλογα, ούτως ώστε αφενός να μπορέσει να εξυπηρετήσει κατάλληλα τις ανάγκες όλων των πελατών της και επομένως αυτοί, έχοντας μείνει ευχαριστημένοι, δεν θα προβούν σε κάποια ενέργεια για αλλαγή σε κάποια εταιρία του ανταγωνισμού και αφετέρου να προσελκύσει καινούριους συνδρομητές.

Όσον αφορά στα προγράμματα, έχει γίνει φανερό, από τις ομάδες που σχηματίστηκαν, ότι το ένα θα πρέπει να έχει πολλά λεπτά ομιλίας και λίγα μηνύματα και δεδομένα ίντερνετ και θα απευθύνεται κατά κύριο λόγο σε συνδρομητές μεγαλύτερων ηλικιών. Ένα δεύτερο πακέτο θα πρέπει να περιέχει πολλά λεπτά ομιλίας, πολλά δεδομένα και αρκετά μηνύματα. Το πακέτο αυτό θα στοχεύει σε μεσαίες ηλικιακές ομάδες. Τέλος, χρειάζεται και ένα τρίτο πρόγραμμα το οποίο θα απευθύνεται στο μικρότερο ηλικιακό κοινό και θα προσφέρει περισσότερο ίντερνετ και λιγότερα

λεπτά ομιλίας και μηνύματα. Με τον τρόπο αυτό ο συνδρομητής δεν θα νιώθει ότι πληρώνει άσκοπα κάποια παροχή που δεν χρησιμοποιεί.

Δεν κρίνεται απαραίτητο να γίνει κάποια διαφοροποίηση των πακέτων ανάμεσα σε άνδρες και γυναίκες συνδρομήτριες. Τα τρία πακέτα που αναφέρθηκαν παραπάνω καλύπτουν πλήρως και τις πιθανές διαφοροποιήσεις των αποτελεσμάτων της ανάλυσης σε άνδρες και γυναίκες.

2.3.3 Μη ιεραρχικές μέθοδοι

Η πιο γνωστή τεχνική μη ιεραρχικής ομαδοποίησης, που θα εφαρμοστεί και παρακάτω είναι μέθοδος MacQueen ή *k*-means method. Η μέθοδος αυτή αρχικά διαμερίζει τα δεδομένα σε *k* ομάδες και υπολογίζει το κέντρο βάρους της κάθε ομάδας. Στη συνέχεια, υπολογίζει την απόσταση της κάθε παρατήρησης από το κέντρο βάρους κάθε ομάδας. Αν ο αλγόριθμος εντοπίσει ότι υπάρχει πλησιέστερο κέντρο βάρους ομάδας στο στοιχείο από αυτό της ομάδας που ήδη ανήκει, τότε τοποθετεί το στοιχείο αυτό στην πλησιέστερη ομάδα και υπολογίζει εκ νέου τα κέντρα βάρους των ομάδων που παρουσίασαν αλλαγή. Η διαδικασία επαναλαμβάνεται μέχρι να σταθεροποιηθούν οι ομάδες.

Ο αλγόριθμος είναι ιδιαίτερα εύχρηστος σε μεγάλα σύνολα δεδομένων, διότι τερματίζει έπειτα από σχετικά μικρό αριθμό επαναλήψεων, δεν κρατά στη μνήμη πολλά στοιχεία, δεν απαιτεί τεράστιες χωρητικότητες ούτε μεγάλη υπολογιστική ισχύ. Επιπλέον, δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων.

Η μέθοδος *k*-means παρουσιάζει και ορισμένα μειονεκτήματα. Αρχικά είναι απαραίτητο να γίνει σωστή επιλογή αρχικών μητρικών σημείων ή αρχικών διαμερίσεων. Αυτό εξασφαλίζεται κάνοντας διαφορετικές δοκιμές ώστε να επιτευχθεί η βέλτιστη λύση. Επιπρόσθετα, υπάρχει περίπτωση δημιουργίας ομάδων με πολύ διασπαρμένα άτομα αν υπάρχουν έκτροπες παρατηρήσεις. Τέλος, ο αλγόριθμος μπορεί να οδηγήσει σε παραπλανητικές ομαδοποιήσεις αν είναι γνωστό εκ των προτέρων ότι ο πληθυσμός που μελετάται διαμερίζεται σε *k* ομάδες αλλά μετά την δημιουργία ομάδων κάποια από αυτές τύχει να μην αντιπροσωπεύεται.

Οι παραπάνω πληροφορίες για τη θεωρία της ανάλυσης των μη ιεραρχικών μεθόδων έχουν αντληθεί από τις σημειώσεις του Μεταπτυχιακού Εφαρμοσμένης Στατιστικής

του Πανεπιστημίου Πειραιώς, Κούτρας (2020), όπου ο αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες.

Για περισσότερες πληροφορίες σχετικά με την θεωρία των μη ιεραρχικών μεθόδων, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Johnson και Wichern (1998).

2.3.4 Εφαρμογή μη ιεραρχικής μεθόδου

Προτού γίνει η εφαρμογή του αλγορίθμου MacQueen ή *k*-means, έγινε τυποποίηση στα δεδομένα, εφόσον οι μεταβλητές όπως αναφέρθηκε και παραπάνω είναι μετρημένες σε διαφορετικές μονάδες μέτρησης (πχ χρόνια, λεπτά ομιλίας, ποσοστά).

Η μέθοδος εφαρμόστηκε για τις ποσοτικές μεταβλητές age, yearpos, resvoice, ressms, resdata, percentagevoiceuse, percentagesmsuse, percentagedatause. Όπως και προηγουμένως, για την ανάλυση δεν έχουν χρησιμοποιηθεί οι παρατηρήσεις-συνδρομητές που έχουν χρησιμοποιήσει τα πακέτα 2,4,6.

Επιπλέον, εφόσον είναι μη ιεραρχική μέθοδος, θα πρέπει να γίνει επιλογή των ομάδων εκ των προτέρων. Για την επιλογή των ομάδων λαμβάνεται υπόψιν ο αριθμός που προέκυψε όταν εκτελέστηκε η ιεραρχική μέθοδος.

Επομένως, εκτελώντας τον αλγόριθμο επιλέγοντας να δημιουργηθούν τρία clusters προκύπτει η εξής κατηγοριοποίηση όπως φαίνεται παρακάτω.

Στο πρώτο cluster έχουν καταταχθεί 44 παρατηρήσεις, στο δεύτερο 34 και στο τρίτο 14.

Number of Cases in each Cluster

Cluster	1	44,000
	2	34,000
	3	14,000
Valid		92,000
Missing		,000

Παρακάτω παρουσιάζεται η κατανομή των παρατηρήσεων που κατατάχθηκε στην κάθε ομάδα σε σχέση με το πακέτο .

Στο πρώτο cluster κατατάχθηκαν 13 συνδρομητές που ενεργοποίησαν το bundle 1, 24 συνδρομητές που ενεργοποίησαν το bundle 2, 1 συνδρομητής που ενεργοποίησε το bundle 5 και 6 συνδρομητές που ενεργοποίησαν το bundle 7. Στο δεύτερο cluster

κατατάχθηκαν 6 συνδρομητές που ενεργοποίησαν το bundle 1, 28 συνδρομητές που ενεργοποίησαν το bundle 2. Στο τρίτο cluster κατατάχθηκαν 14 συνδρομητές που ενεργοποίησαν το bundle 5.

	Clusters			
Bundles	**1**	**2**	**3**	**Άθροισμα**
1	13	6		19
3	24	28		52
5	1		14	15
7	6			6
Γενικό Άθροισμα	**44**	**34**	**14**	**92**

Φαίνεται ότι η κατανομή στις ομάδες είναι σχεδόν ίδια με την κατανομή που προέκυψε έπειτα από την εφαρμογή της ιεραρχικής μεθόδου.

Από τον πίνακα 2.3 φαίνεται ότι το μοντέλο ολοκληρώνει την ομαδοποίηση έπειτα από δέκα επαναλήψεις του αλγορίθμου.

Πίνακας 2.3

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3
1	4,001	2,844	1,436
2	,370	,177	,000
3	,405	,262	,000
4	,124	,082	,000
5	,122	,092	,000
6	,054	,042	,000
7	,102	,084	,000
8	,225	,217	,000
9	,098	,114	,000
10	,110	,132	,000

a. Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is ,098. The current iteration is 10. The minimum distance between initial centers is 5,986.

Πίνακας 2.4

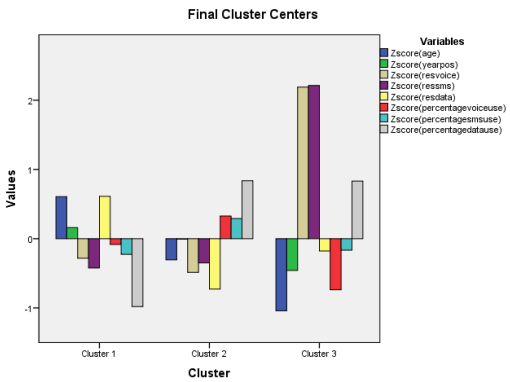
Mean

Clusters	age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
1	60,52	5,545	177,86	213,30	356,45	,4724	,0495	,0524
2	44,47	4,809	120,12	238,91	48,71	,5996	,1158	,8561
3	31,57	2,821	876,14	1131,29	175,07	,2699	,0573	,8541
Total	50,18	4,859	262,78	362,46	215,12	,4886	,0752	,4714

Στον πίνακα 2.4 παρουσιάζονται οι μέσες τιμές της κάθε μεταβλητής για τα τρία clusters που έχουν δημιουργηθεί. Παρατηρείται ότι οι συνδρομητές που ανήκουν στο πρώτο cluster έχουν μεγαλύτερη μέση ηλικία, ενώ αυτοί που ανήκουν στο τρίτο τη μικρότερη μέση ηλικία. Τα άτομα του πρώτου και δεύτερου cluster φαίνεται ότι κατέχουν περισσότερα χρόνια το καρτοκινητό τους. Επιπλέον, οι συνδρομητές που ανήκουν στο cluster 2 χρησιμοποιούν το μεγαλύτερο ποσοστό των λεπτών ομιλίας τους σε σχέση με τους υπόλοιπους. Οι συνδρομητές που ανήκουν στο cluster 3 χρησιμοποιούν αρκετά μικρότερο ποσοστό των λεπτών ομιλίας τους και από τα δύο άλλα clusters. Οι συνδρομητές που ανήκουν στο cluster 2 χρησιμοποιούν λίγο μεγαλύτερο ποσοστό των μηνυμάτων του πακέτου τους σε σχέση με τους υπόλοιπους. Τέλος, παρατηρείται ότι οι συνδρομητές που ανήκουν στα clusters 2 και 3 χρησιμοποιούν πολύ μεγάλο ποσοστό των δεδομένων τους, ενώ αυτοί του 1 πολύ μικρό ποσοστό.

Στο σχήμα 2.1 απεικονίσθηκαν σε bar plots τα τελικά κανονικοποιημένα κέντρα των μεταβλητών στις τρεις ομάδες που δημιουργήθηκαν. Όλα τα κέντρα των μεταβλητών φαίνεται να διαφέρουν από το ένα cluster στο άλλο εκτός από της percentagedatause που στα clusters 2 και 3 είναι σχεδόν ίδια.

Σχήμα 2.1



Στη συνέχεια θα εφαρμοστεί η μέθοδος μόνο για τους άνδρες συνδρομητές του δείγματος. Το πλήθος των clusters που επιλέγεται είναι 2, όπως προέκυψε από την ιεραρχική μέθοδο ομαδοποίησης.

Number of Cases in each Cluster

Cluster	1	8,000
	2	46,000
Valid		54,000
Missing		,000

Πίνακας 2.5

Mean

Clusters	age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
1	34,13	3,125	911,25	1141,00	205,25	,2406	,0492	,8290
2	54,72	5,565	140,83	243,63	237,91	,5751	,0676	,3746
Total	51,67	5,204	254,96	376,57	233,07	,5256	,0649	,4419

Από τους 54 άνδρες συνδρομητές οι 8 κατατάχθηκαν σε μία ομάδα και οι υπόλοιποι 46 στη δεύτερη.

Στον πίνακα 2.5 παρουσιάζονται οι μέσες τιμές της κάθε μεταβλητής για τα δύο clusters που έχουν δημιουργηθεί. Παρατηρείται ότι οι συνδρομητές που ανήκουν στο πρώτο cluster έχουν μικρότερη μέση ηλικία. Τα άτομα του πρώτου cluster φαίνεται ότι κατέχουν κατά μέσο όρο λιγότερα χρόνια το καρτοκινητό τους. Επιπλέον, οι συνδρομητές που ανήκουν στο cluster 2 χρησιμοποιούν μεγαλύτερο ποσοστό των λεπτών ομιλίας. Οι συνδρομητές που ανήκουν στο cluster 2 χρησιμοποιούν λίγο μεγαλύτερο ποσοστό των μηνυμάτων του πακέτου τους. Τέλος, παρατηρείται ότι οι συνδρομητές που ανήκουν στο cluster 1 χρησιμοποιούν πολύ μεγάλο ποσοστό των δεδομένων τους.

Στη συνέχεια θα εφαρμοστεί η μέθοδος μόνο για τις γυναίκες συνδρομήτριες του δείγματος. Το πλήθος των clusters που επιλέγεται είναι 3, όπως προέκυψαν από την ιεραρχική μέθοδο ομαδοποίησης.

Από τις 38 γυναίκες συνδρομήτριες οι 16 κατατάχθηκαν στο πρώτο cluster, άλλες 16 κατατάχθηκαν στο δεύτερο cluster και οι υπόλοιπες 6 στο τρίτο.

Number of Cases in each Cluster

Cluster	1	16,000
	2	16,000
	3	6,000
Valid		38,000
Missing		,000

Πίνακας 2.6

Mean

Clusters	age	yearpos	resvoice	ressms	resdata	percentagevoiceuse	percentagesmsuse	percentagedatause
1	60,94	3,656	176,81	183,81	367,31	,4469	,0256	,0360
2	42,69	5,813	162,69	210,00	32,44	,4729	,1621	,8502
3	28,17	2,417	829,33	1118,33	134,83	,3089	,0681	,8876
Total	48,08	4,368	273,89	342,39	189,61	,4361	,0898	,5133

Στον πίνακα 2.6 παρουσιάζονται οι μέσες τιμές της κάθε μεταβλητής για τα τρία clusters που έχουν δημιουργηθεί. Παρατηρείται ότι οι συνδρομήτριες που ανήκουν στο πρώτο cluster έχουν μεγαλύτερη μέση ηλικία, ενώ αυτές που ανήκουν στο τρίτο τη μικρότερη μέση ηλικία. Τα άτομα του δεύτερου cluster φαίνεται ότι κατέχουν κατά μέσο όρο περισσότερα χρόνια το καρτοκινητό τους. Επιπλέον, οι συνδρομήτριες που ανήκουν στα cluster 1 και cluster 2 χρησιμοποιούν το μεγαλύτερο ποσοστό των λεπτών ομιλίας τους σε σχέση με αυτές του cluster 3. Οι συνδρομήτριες που ανήκουν στο cluster 2 χρησιμοποιούν λίγο μεγαλύτερο ποσοστό των μηνυμάτων του πακέτου τους σε σχέση με τις υπόλοιπες. Τέλος, παρατηρείται ότι τα άτομα που ανήκουν στα clusters 2 και 3 χρησιμοποιούν πολύ μεγάλο ποσοστό των δεδομένων τους, ενώ αυτά του 1 πολύ μικρό ποσοστό. Τα αποτελέσματα που προέκυψαν για τις γυναίκες συνδρομήτριες είναι πολύ παρόμοια με αυτά που είχαν προκύψει για το σύνολο του δείγματος.

Συμπερασματικά, αντίστοιχα αποτελέσματα με την ιεραρχική μέθοδο που εφαρμόστηκε στην προηγούμενη παράγραφο παρατηρούνται και στη μη ιεραρχική μέθοδο. Δηλαδή, θα πρέπει να υπάρχουν πακέτα τα οποία να καλύπτουν τις ανάγκες όλων των συνδρομητών. Παρατηρώντας τις ομάδες που σχηματίστηκαν, φαίνεται ότι θα πρέπει να υπάρχει ένα πακέτο το οποίο θα απευθύνεται κατά κύριο λόγο σε

συνδρομητές μεγαλύτερων ηλικιών, που παρέχει πολλά λεπτά ομιλίας, κάποιον αριθμό μηνυμάτων και κάποιον αριθμό δεδομένων. Επιπλέον θα πρέπει να υπάρχει ένα δεύτερο πακέτο για τις μεσαίες ηλικιακές ομάδες με πολλά λεπτά ομιλίας, πολλά δεδομένα και αρκετά μηνύματα. Τέλος για τις μικρότερες ηλικιακές ομάδες, χρειάζεται και ένα τρίτο πρόγραμμα το οποίο θα προσφέρει περισσότερο ίντερνετ και λιγότερα λεπτά ομιλίας και μηνύματα. Όπως και στην ιεραρχική μέθοδο δεν κρίνεται απαραίτητο να γίνει διαφοροποίηση των πακέτων ανάμεσα στους άνδρες και τις γυναίκες συνδρομήτριες.

Κεφάλαιο 3

3.1 Ανάλυση Κυρίων Συνιστωσών

Στόχος της ανάλυσης των κυρίων Συνιστωσών (Principal Components Analysis) είναι η μείωση των διαστάσεων δεδομένων που αποτελούνται από μεγάλο πλήθος μεταβλητών. Με την υλοποίηση της μεθόδου αυτής οι αρχικές μεταβλητές «συνοψίζονται» από έναν μικρό αριθμό από γραμμικούς συνδυασμούς τους, τις πρώτες κύριες συνιστώσες, οι οποίες περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της αρχικής πληροφορίας των μεταβλητών αλλά είναι και ασυσχέτιστες μεταξύ τους. Αυτό έχει ως αποτέλεσμα τη δυνατότητα της αξιόπιστης παρουσίασης των δεδομένων, μέσω της γραφικής απεικόνισης των πρώτων κύριων συνιστωσών. Επιπλέον, γίνεται οικονομία όσον αφορά στον χώρο αποθήκευσης των δεδομένων.

Αναλυτικότερα, με τη μέθοδο της Ανάλυσης των Κυρίων Συνιστωσών ένα σύνολο μεταβλητών $X_1, X_2, .., X_p$ αντικαθίσταται με ένα μικρότερο πλήθος μεταβλητών $Y_1, Y_2, \ldots$, το οποίο αποτελείται από γραμμικούς συνδυασμούς των αρχικών μεταβλητών και διατηρεί ένα σημαντικό μέρος της πληροφορίας των αρχικών δεδομένων. Πληροφορία των μεταβλητών θεωρείται το ποσοστό της συνολικής διασποράς που θα μεταφερθεί στις $Y_1, Y_2, \ldots$.Η πρώτη μεταβλητή Y_1 λέγεται πρώτη κύρια συνιστώσα και περιέχει τη μέγιστη δυνατή πληροφορία. Αντίστοιχα, η επόμενη στη σειρά μεταβλητή Y_2 λέγεται δεύτερη κύρια συνιστώσα και έχει την αμέσως μεγαλύτερη δυνατή πληροφορία που δεν έχει μεταφερθεί στην Y_1 κ.ο.κ . Οι κύριες συνιστώσες είναι ασυσχέτιστες.

Παρακάτω θα παρουσιαστεί αναλυτικότερα η θεωρία της Ανάλυσης των Κυρίων Συνιστωσών:

Γίνεται η υπόθεση ότι υπάρχει δείγμα n ατόμων ενός πληθυσμού, όπου σε κάθε άτομο παρατηρούνται $p \geq 2$ τυχαίες μεταβλητές – χαρακτηριστικά $X_1, X_2, ..., X_p$.

Έτσι προκύπτει ο πίνακας $n \times p$ διαστάσεων:

$$X=(x_{ij}) = \begin{bmatrix} x_{11} & \cdots x_{1j} & \ldots x_{1p} \\ \vdots & \vdots & \vdots \\ x_{i1} & \ldots x_{ij} & \ldots x_{ip} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots x_{nj} & \ldots x_{np} \end{bmatrix}.$$

Η τιμή x_{ij} αντιστοιχεί στην τιμή που καταγράφηκε για το i άτομο για το χαρακτηριστικό j.

Σκοπός της ανάλυσης είναι η δημιουργία της πρώτης κύριας συνιστώσας *Y*, η οποία προκύπτει ως γραμμικός συνδυασμός των $X_1, X_2, \ldots, X_p$.

Δηλαδή η μορφή της *Y* θα είναι η εξής:

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_p X_p$$

Στόχος είναι ο προσδιορισμός των $\alpha_1, \alpha_2, \ldots, \alpha_p \in \Re$, ώστε οι τιμές (scores) των *n* ατόμων για τη μεταβλητή *Y*, δηλαδή τα

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip}, \qquad i=1,2,\ldots,n \tag{3.1}$$

να διατηρούν το μεγαλύτερο δυνατόν τις αποστάσεις που έχουν τα άτομα ως προς όλες τις αρχικές μεταβλητές.

Αυτό επιτυγχάνεται με τη μεγιστοποίηση της ποσότητας :

$$SS_y = \sum_{r=1}^{n} \sum_{i=r+1}^{n} (y_r - y_i)^2 = \frac{1}{2} \sum_{r=1}^{n} \sum_{i=r+1}^{n} (y_r - y_i)^2$$

Έπειτα από πράξεις προκύπτει η παρακάτω ισότητα:

$$SS_y = nDis(Y), \text{ όπου } Dis(Y) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Η ποσότητα *Dis(Y)* αποτελεί μέτρο μεταβλητότητας για το σύνολο των τιμών $y_1, y_2, \ldots, y_n$.

Επομένως, το πρόβλημα μεγιστοποίησης της ποσότητας SS_y είναι ισοδύναμο με την μεγιστοποίηση της ποσότητας *Dis(Y).*

Λόγω της (3.1) έχουμε ότι :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} (y_i) = \frac{1}{n} \sum_{i=1}^{n} \left(a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip}\right).$$

Εισάγοντας το συμβολισμό :

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \text{ , } j=1, 2, \ldots, p,$$

προκύπτει ότι:

$$\bar{y} = a_1 \bar{x}_1 + a_2 \bar{x}_2 + \ldots + a_p \bar{x}_p, \tag{3.2}$$

αφαιρώντας την (3.2) από την (3.1) προκύπτει η σχέση :

$$y_i - \bar{y} = a_1 (x_{i1} - \bar{x}_1) + a_2 (x_{i2} - \bar{x}_2) + \cdots + a_p \left(x_{ip} - \bar{x}_p\right). \tag{3.3}$$

Εισάγοντας τους συμβολισμούς για τα παρακάτω διανύσματα-στήλες :

$$\boldsymbol{z_i} = \begin{bmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_1 \\ . \\ . \\ . \\ x_{ip} - \bar{x}_p \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_p \end{bmatrix}$$

η σχέση (3.3) μπορεί να γραφτεί:

$$f_i = y_i - \bar{y} = \boldsymbol{z_i'a}, \quad i=1, 2, \ldots, n.$$

Επιπλέον θα εισαχθούν και οι εξής συμβολισμοί:

$$\boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ . \\ . \\ . \\ f_p \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ . \\ . \\ . \\ y_p - \bar{y} \end{bmatrix}.$$

Το διάνυσμα αυτό ονομάζεται διάνυσμα των κεντρικοποιημένων scores.

Ο αριθμός $f_i = y_i - \bar{y}$, $i=1,2,\ldots,n$ συμβολίζει την απόκλιση του score y_i του i ατόμου από το μέσο score $\bar{y}$ των n ατόμων.

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip} \end{bmatrix}.$$

Το διάνυσμα γραμμή $x_i' = (x_{i1}, x_{i2,\ldots} x_{ip})$ περιέχει τις παρατηρήσεις που αφορούν στο i άτομο για κάθε μία από τις p μεταβλητές.

$$\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ . \\ . \\ . \\ \bar{x}_p \end{bmatrix}.$$

Το διάνυσμα γραμμή $\bar{x}$ περιέχει τους δειγματικούς μέσους ανά μεταβλητή

Βάσει των παραπάνω σχέσεων μπορεί να γραφτεί ότι :

$$\boldsymbol{z_i} = \boldsymbol{x_i} - \bar{\boldsymbol{x}}, \boldsymbol{y_i} = \boldsymbol{x_i'a} \text{ και } \boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ . \\ . \\ . \\ f_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{z_1'a} \\ \boldsymbol{z_2'a} \\ . \\ . \\ . \\ \boldsymbol{z_n'a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z_1'} \\ \boldsymbol{z_2'} \\ . \\ . \\ . \\ \boldsymbol{z_n'} \end{bmatrix} \boldsymbol{\alpha} = Z\boldsymbol{\alpha},$$

όπου με Z συμβολίζεται ο παρακάτω πίνακας και ονομάζεται πίνακας κεντρικοποιημένων δεδομένων:

$$Z=\begin{bmatrix} \boldsymbol{z_1'} \\ \boldsymbol{z_2'} \\ . \\ . \\ . \\ \boldsymbol{z_n'} \end{bmatrix} = \begin{bmatrix} x_1-\bar{x}_1 & x_2-\bar{x}_2 & \cdots & x_1-\bar{x}_P \\ x_1-\bar{x}_1 & x_2-\bar{x}_2 & \cdots & x_1-\bar{x}_P \\ & & . & \\ & & . & \\ & & . & \\ x_1-\bar{x}_1 & x_2-\bar{x}_2 & \cdots & x_1-\bar{x}_P \end{bmatrix}.$$

Η ποσότητα

$$Dis(Y)= \sum_{i=1}^{n}(y_i-\bar{y})^2 ,$$

η οποία πρέπει να μεγιστοποιηθεί μπορεί να γραφτεί ως εξής:

$$Dis(Y)= \sum_{i=1}^{n} f_i^2 = \boldsymbol{f'f} = (Z\boldsymbol{\alpha})'(Z\boldsymbol{\alpha}).$$

Η ποσότητα αυτή θα ονομάζεται διασπορά του συνόλου Ν κατά μήκος του διανύσματος *α* και θα συμβολίζεται με

$$Dis_{\alpha}(Y) = Dis(Y) = \boldsymbol{a'}Z'Z\boldsymbol{a}.$$

Για τα αρχικά δεδομένα έπειτα από πράξεις προκύπτει το συμπέρασμα ότι :

$$Dis(N) = tr(Z'Z), \qquad (3.4)$$

Δηλαδή η *Dis(N)* θα ισούται με το ίχνος του πίνακα $Z'Z$ (άθροισμα των διαγώνιων στοιχείων του $Z'Z$)

Η παραπάνω ποσότητα θα ονομάζεται διασπορά του συνόλου σημείων *N*.

Για τη μεγιστοποίησή της διασποράς $Dis_{\alpha}(N)$ θα πρέπει να βρεθεί το κατάλληλο διάνυσμα των συντελεστών :

$$\boldsymbol{\alpha}= \left(a_1, a_2, \dots, a_p\right)'.$$

Αυτό μπορεί να επιτευχθεί πολλαπλασιάζοντας όλες τις συντεταγμένες του α με κάποιον αριθμό. Επομένως η διασπορά μπορεί να αυξάνει απεριόριστα. Για να αποφευχθεί αυτό, τις περισσότερες φορές, λόγω του ότι γίνεται χρήση της Ευκλείδειας απόστασης, εφαρμόζεται ο περιορισμός ότι το μέτρο του διανύσματος α να είναι ίσο με τη μονάδα.

Δηλαδή για το μοναδιαίο διάνυσμα α ισχύει:

$$\|\boldsymbol{a}\| = 1 \Leftrightarrow \boldsymbol{a'a} = 1 \Leftrightarrow \sum_{i=1}^{p} a_i^2 = 1$$

Συμπερασματικά, σκοπός είναι να βρεθεί για δεδομένο πίνακα *Z* διάστασης $n \times p$, το μοναδιαίο διάνυσμα *α* το οποίο μεγιστοποιεί την τετραγωνική μορφή :

$$Dis_{\alpha}(Y) = Dis(Y) = \boldsymbol{a'}Z'Z\boldsymbol{a}$$

Αυτό θα πραγματοποιηθεί με τη χρήση των ιδιοτιμών και των ιδιοδιανυσμάτων του πίνακα $Z'Z$.

Έστω $\lambda_1, \lambda_2, \dots, \lambda_p$ οι μη αρνητικές ιδιοτιμές του πίνακα $Z'Z$.

Αν θεωρηθεί ότι $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ και συμβολιστούν $\boldsymbol{u_1, u_2, \dots, u_p}$ τα αντίστοιχα μοναδιαία ιδιοδιανύσματα τότε:

- το διάνυσμα $\boldsymbol{a}$ που μεγιστοποιεί την τετραγωνική μορφή $\boldsymbol{a}'Z'Z\boldsymbol{a}$ είναι το μοναδιαίο διάνυσμα $\boldsymbol{u_1}= (u_{11}, u_{12}, \ldots, u_{1p})$ που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή λ_1
- η μέγιστη τιμή της $a'Z'Za$ είναι ίση με την ιδιοτιμή λ_1, δηλαδή

$$\max_{\|a\|=1} Dis_a(N) = Dis_{u_1}(N) = \lambda_1.$$

Η μεταβλητή Y_1 για την οποία επιτυγχάνεται η μεγιστοποίηση που περιγράφηκε παραπάνω ονομάζεται πρώτη κύρια συνιστώσα. Δηλαδή, η Y_1 γράφεται ως γραμμικός συνδυασμός των αρχικών μεταβλητών ως εξής

$$Y_1= u_{11}X_1+ u_{12}X_2+ \ldots+ u_{1p}X_p.$$

Αντίστοιχα, το διάνυσμα u_2 που αντιστοιχεί στη δεύτερη στη σειρά μεγαλύτερη ιδιοτιμή λ_2 του πίνακα $Z'Z$ θα δημιουργεί έναν δεύτερο γραμμικό συνδυασμό των αρχικών μεταβλητών ως εξής:

$$Y_2= u_{21}X_1+ u_{22}X_2+\ldots+ u_{2p}X_p\,.$$

Η Y_2 αποτελεί την δεύτερη κύρια συνιστώσα.

Οι υπόλοιπες κύριες συνιστώσες υπολογίζονται αντίστοιχα.

Γενικά, ισχύει η ιδιότητα ότι το ίχνος ενός πίνακα μπορεί να γραφεί ως το άθροισμα των ιδιοτιμών του. Επομένως, από τη σχέση (3.4) προκύπτει ότι :

$$Dis(N) = tr(Z'Z) = \sum_{j=1}^{p} \lambda_j = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

3.2 Τυποποίηση των δεδομένων

Οι γραμμικοί συνδυασμοί $u_1,u_2,\ldots,u_p$ που δημιουργούνται εξαρτώνται από τις μονάδες μέτρησης των μεταβλητών. Για να αποφευχθεί αυτό θα πρέπει να γίνει κανονικοποίηση της κάθε μεταβλητής- χαρακτηριστικού.

3.3 Γεωμετρική ερμηνεία

Μελετώντας μεμονωμένα την πρώτη κύρια συνιστώσα διαπιστώνεται ότι αυτή καθορίζει μια κατεύθυνση (ευθεία ε που ορίζεται από το μοναδιαίο διάνυσμα) τέτοια ώστε οι προβολές των διαθέσιμων σημείων (δεδομένων) επάνω της να βρίσκονται όσο πιο κοντά γίνεται στις πραγματικές θέσεις των σημείων. Η κατεύθυνση αυτή αποτελεί τον πρώτο κύριο άξονα.

Επεκτείνοντας την ιδέα αυτή σε ένα επίπεδο αντί για ευθεία μπορεί να εξηγηθεί γεωμετρικά και η δεύτερη κύρια συνιστώσα. Αναλυτικότερα, στόχος τώρα είναι η όσο το δυνατόν πιο πιστή αναπαράσταση των σημείων (δεδομένων) πάνω στο επίπεδο.

3.4 Επιλογή πλήθους των κυρίων συνιστωσών

Όπως έχει προαναφερθεί στόχος της Ανάλυσης των κυρίων Συνιστωσών είναι η μείωση των διαστάσεων των δεδομένων. Στη διαδικασία αυτής της μείωσης είναι αναμενόμενο να υπάρξει απώλεια πληροφορίας. Επομένως γεννάται το ερώτημα πόσο ποσοστό της πληροφορίας είναι «αποδεκτό» και «επιθυμητό» να διατηρηθεί, δηλαδή ποιο είναι το πλήθος των κύριων συνιστωσών που θα διατηρηθούν. Η απάντηση στο ερώτημα αυτό είναι υποκειμενική και διαφοροποιείται ανάλογα με τα δεδομένα και τον κάθε αναλυτή.

Παρακάτω περιγράφονται κάποια από τα συνηθέστερα κριτήρια που χρησιμοποιούνται για τη λήψη της απόφασης αυτής:

α. Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Σύμφωνα με το κριτήριο αυτό επιλέγεται εκ των προτέρων ένα όριο (ποσοστό) και στη συνέχεια επιλέγεται ο αριθμός των συνιστωσών που εξηγεί μεγαλύτερο ποσοστό διακύμανσης από το όριο που τέθηκε. Με το κριτήριο αυτό ελλοχεύει ο κίνδυνος, αν το όριο έχει οριστεί πολύ υψηλό, να χρειαστεί να διατηρηθεί αρκετά μεγάλο πλήθος συνιστωσών.

β. Κριτήριο του Kaiser

Σύμφωνα με το κριτήριο αυτό διατηρούνται μόνο οι ιδιοτιμές που είναι μεγαλύτερες από τη μέση τιμή των ιδιοτιμών $\lambda_1, \lambda_2, ..., \lambda_p$.

Όταν γίνεται η χρήση του πίνακα συσχετίσεων τότε η μέση τιμή των ιδιοτιμών είναι ίση με 1, άρα όσες ιδιοτιμές είναι μεγαλύτερες του 1 τόσες συνιστώσες σε πλήθος επιλέγονται.

γ. Ποσοστό της διακύμανσης που ερμηνεύεται για κάθε αρχική μεταβλητή

Σύμφωνα με το κριτήριο αυτό επιλέγεται εκ των προτέρων ένα όριο για το ελάχιστο ποσοστό της διακύμανσης της καθεμίας από τις αρχικές μεταβλητές που θα πρέπει να ερμηνεύεται από το «κατάλληλο» πλήθος συνιστωσών.

δ. Scree Plot

Το Scree Plot είναι το γράφημα που απεικονίζει τις ιδιοτιμές βάσει της σειράς μεγέθους τους. Το σημείο στο οποίο το γράφημα γίνεται περίπου παράλληλο με τον άξονα που απεικονίζει τον αριθμό των ιδιοτιμών, υποδεικνύει το πλήθος των συνιστωσών που πρέπει να διατηρηθούν.

Οι παραπάνω πληροφορίες για τη θεωρία της ανάλυσης των κυρίων συνιστωσών έχουν αντληθεί από τις σημειώσεις του Μεταπτυχιακού Εφαρμοσμένης Στατιστικής του Πανεπιστημίου Πειραιώς, Κούτρας (2020), όπου ο αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες.

Επιπλέον, για περισσότερες πληροφορίες σχετικά με την θεωρία της ανάλυσης κυρίων συνιστωσών, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Johnson και Wichern, (1998) και στο βιβλίο του Jolliffe, (2002).

3.5 Εφαρμογή Ανάλυσης Κυρίων Συνιστωσών

Για την εφαρμογή της μεθόδου αυτής χρησιμοποιήθηκαν μόνο οι ποσοτικές μεταβλητές age, yearspos, resvoice, ressms, resdata, percentagevoiceuse, percentagesmsuse, percentagedatause .

Οι μεταβλητές είναι μετρημένες σε διαφορετικές μονάδες μέτρησης επομένως θα γίνει τυποποίηση.

Παρακάτω παρουσιάζεται ο πίνακας συσχετίσεων.

Correlation Matrix[a]

		age	yearpos	resvoice	ressms	resdata	Percentage voiceuse	Percentage smsuse	Percentage datause
Correlation	age	1,000	,012	-,370	-,419	,461	,009	-,029	-,561
	yearpos	,012	1,000	-,160	-,204	-,106	,029	,088	-,055
	resvoice	-,370	-,160	1,000	,819	-,014	-,625	-,118	,275
	ressms	-,419	-,204	,819	1,000	,005	-,184	-,170	,358
	resdata	,461	-,106	-,014	,005	1,000	,015	-,177	-,704
	percentagevoiceuse	,009	,029	-,625	-,184	,015	1,000	,143	-,037
	percentagesmsuse	-,029	,088	-,118	-,170	-,177	,143	1,000	,206
	percentagedatause	-,561	-,055	,275	,358	-,704	-,037	,206	1,000
Sig. (1-tailed)	age		,455	,000	,000	,000	,467	,391	,000
	yearpos	,455		,063	,026	,157	,394	,201	,301
	resvoice	,000	,063		,000	,449	,000	,131	,004
	ressms	,000	,026	,000		,482	,040	,053	,000
	resdata	,000	,157	,449	,482		,444	,046	,000
	percentagevoiceuse	,467	,394	,000	,040	,444		,087	,362
	percentagesmsuse	,391	,201	,131	,053	,046	,087		,024
	percentagedatause	,000	,301	,004	,000	,000	,362	,024	

a. Determinant = ,014

Εφαρμόζοντας την ανάλυση των κυρίων συνιστωσών παρατηρείται για τις μεταβλητές yearpos και percentagesmsuse ότι δεν παρουσιάζουν υψηλή συσχέτιση με καμία από τις υπόλοιπες μεταβλητές, επομένως αναμένεται οι κύριες συνιστώσες να μην εξηγούν μεγάλο μέρος της διακύμανσής τους.

Στον πίνακα 3.1 η πρώτη στήλη δείχνει τον αριθμό των κυρίων συνιστωσών. Η δεύτερη δείχνει τη διακύμανση της κάθε συνιστώσας, δηλαδή την ιδιοτιμή λ. Στην τρίτη στήλη φαίνεται το ποσοστό της διακύμανσης που εξηγεί η κάθε συνιστώσα. Στη τέταρτη στήλη φαίνεται το αθροιστικό ποσοστό της συνολικής μεταβλητότητας που προκύπτει με την προσθήκη της κάθε μεταβλητής. Στις επόμενες στήλες φαίνονται οι αντίστοιχες πληροφορίες για το πλήθος των κυρίων συνιστωσών που επιλέχθηκαν .

Πίνακας 3.1

Total Variance Explained

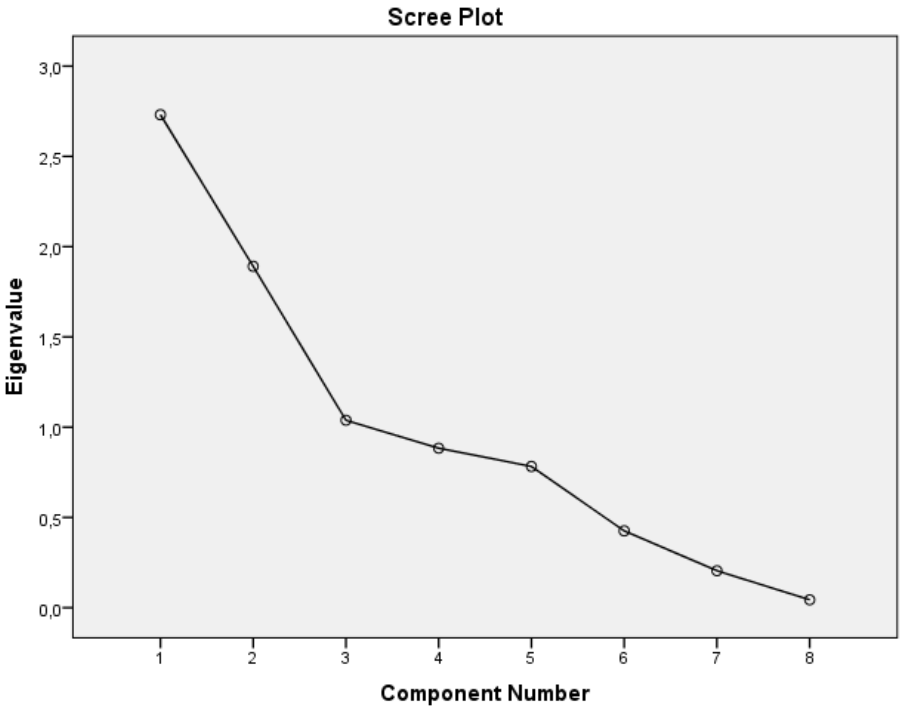
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,731	34,134	34,134	2,731	34,134	34,134
2	1,891	23,636	57,770	1,891	23,636	57,770
3	1,038	12,981	70,751	1,038	12,981	70,751
4	,883	11,044	81,794			
5	,782	9,781	91,575			
6	,426	5,319	96,894			
7	,205	2,560	99,454			
8	,044	,546	100,000			

Extraction Method: Principal Component Analysis.

Σύμφωνα με το κριτήριο του Kaiser και εφόσον για την ανάλυση έχει χρησιμοποιηθεί ο πίνακας συσχετίσεων, επιλέγονται οι κύριες συνιστώσες που αντιστοιχούν σε ιδιοτιμή μεγαλύτερη του 1. Επομένως το πλήθος των συνιστωσών που θα κρατηθεί είναι 3.

Παρατηρείται ότι με τις τρείς κύριες συνιστώσες ερμηνεύεται το 70,8% της συνολικής μεταβλητότητας. Το αθροιστικό αυτό ποσοστό είναι αρκετά ικανοποιητικό.

Το ίδιο συμπέρασμα προκύπτει παρατηρώντας και το scree plot. Το σημείο στο οποίο φαίνεται να οριζοντιοποιείται το γράφημα είναι στις 3 κύριες συνιστώσες.



Παρακάτω παρουσιάζονται οι συσχετίσεις της κάθε μεταβλητής με την κάθε συνιστώσα.

(Το SPSS θεωρεί τα ιδιοδιανύσματα που αντιστοιχούν στα $\lambda_1, \lambda_2, \lambda_3$ και έχουν μέτρο $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}$, δηλαδή τα $\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2, \sqrt{\lambda_3}u_3$)

Πίνακας 3.2

Component Matrix[a]

	Component		
	1	2	3
age	-,728	,317	,124
yearpos	-,171	-,294	,751
resvoice	,803	,511	,159
ressms	,771	,380	-,209
resdata	-,486	,708	-,136
percentagevoiceuse	-,400	-,493	-,597
percentagesmsuse	-,023	-,522	,105
percentagedatause	,732	-,530	-,068

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

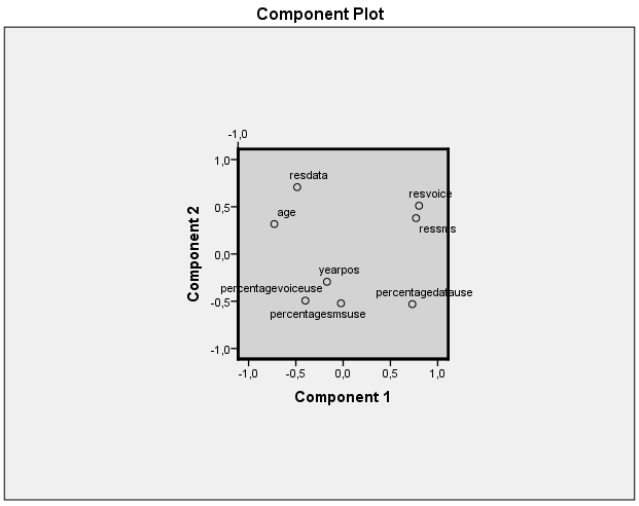
Όπως φαίνεται από τον πίνακα 3.2 η πρώτη κύρια συνιστώσα σχετίζεται ισχυρά με τις μεταβλητές age, resvoice, resms, και percentagedatause.

Η δεύτερη κύρια συνιστώσα σχετίζεται ισχυρά με τις μεταβλητές resvoice, resdata, percentagesmsuse και percentagedatause.

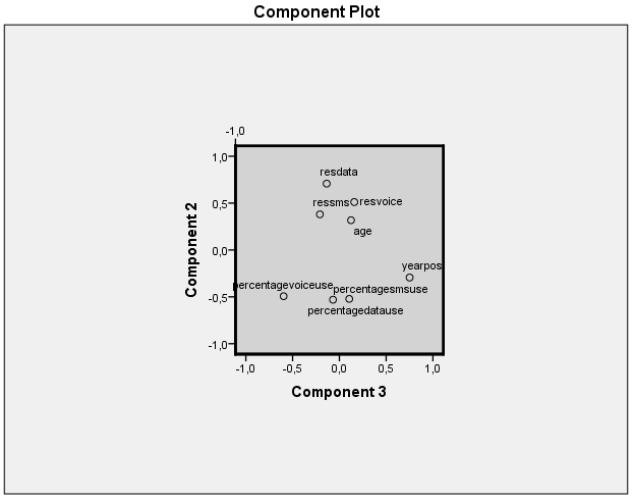
Η τρίτη κύρια συνιστώσα σχετίζεται ισχυρά με τις μεταβλητές yearpos και percentagevoiceuse.

Αυτό επιβεβαιώνεται και από τα σχήματα 3.1 και 3.2. Στο πρώτο γράφημα απεικονίζονται οι συσχετίσεις των μεταβλητών σε σχέση με την πρώτη και δεύτερη συνιστώσα, ενώ στο δεύτερο απεικονίζονται οι συσχετίσεις των μεταβλητών σε σχέση με την πρώτη και τρίτη συνιστώσα. Φαίνονται οι συσχετίσεις που περιγράφηκαν παραπανω:

Σχήμα 3.1

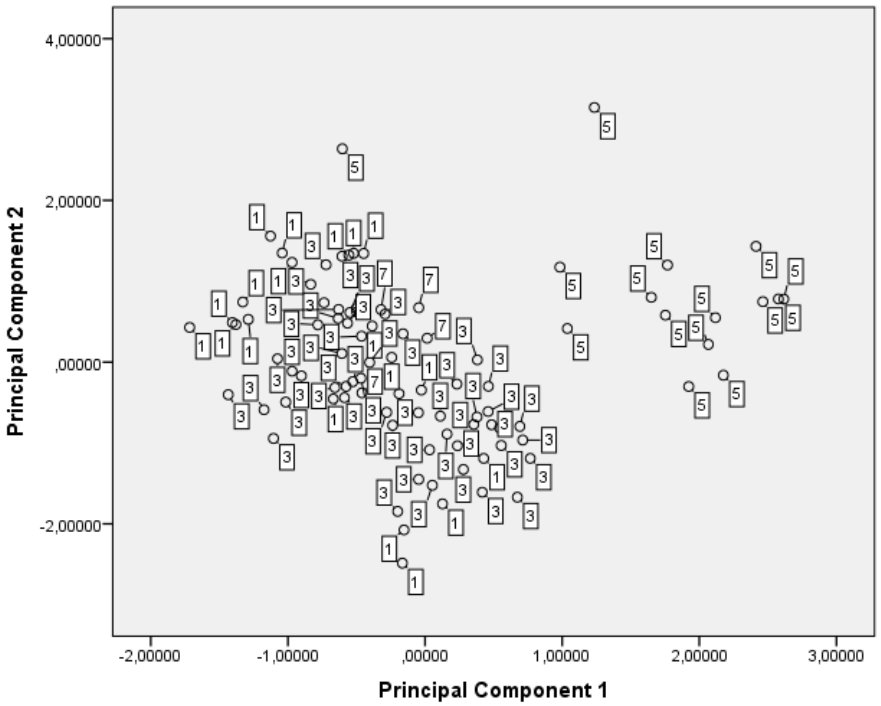


Σχήμα 3.2



Στη συνέχεια, απεικονίζονται στο σχήμα 3.3 οι τιμές της πρώτης και δεύτερης συνιστώσας σε δύο άξονες και πάνω στα σημεία παρουσιάζεται ο αριθμός του προγράμματος στον οποίο αντιστοιχεί η κάθε τιμή. Παρατηρείται ο διαχωρισμός του προγράμματος 5 από όλα τα υπόλοιπα.

Σχήμα 3.3



Προκύπτει το συμπέρασμα, σε συνδυασμό με τις συσχετίσεις που παρατηρήθηκαν στον component matrix ότι οι συνδρομητές που χρησιμοποιούν το bundle 5, δηλαδή ουσιαστικά οι παρατηρήσεις που κατηγοριοποιούνται από την πρώτη κύρια συνιστώσα, έχουν μικρή ηλικία, αφήνουν πολλά αχρησιμοποίητα λεπτά ομιλίας και μηνυμάτων. Επιπλέον, έχουν υψηλό ποσοστό χρήσης δεδομένων ίντερνετ.

Με την εφαρμογή της μεθόδου των κυρίων συνιστωσών μειώθηκαν οι διαστάσεις των δεδομένων αφού πλέον μπορούν να χρησιμοποιηθούν οι 3 κύριες συνιστώσες, οι οποίες είναι και ασυσχέτιστες σε σχέση με τις αρχικές 8 ποσοτικές μεταβλητές των δεδομένων.

Συμπερασματικά, η εταιρία λαμβάνοντας υπόψιν τα αποτελέσματα της εφαρμογής της ανάλυσης των κυρίων συνιστωσών, θα πρέπει να προσαρμόσει τα πακέτα της καρτοκινητής, ώστε να είναι πιο κοντά στις ανάγκες των χρηστών. Αναλυτικότερα, να υπάρχουν bundles τα οποία περιέχουν περισσότερα δεδομένα ίντερνετ και άλλα τα οποία περιέχουν περισσότερη ομιλία και μηνύματα. Με τον τρόπο αυτό ο πελάτης δεν θα πιστεύει ότι πληρώνει χωρίς λόγο κάτι που δεν χρειάζεται.

Κεφάλαιο 4

4.1 Ανάλυση Παραγόντων

Η ανάλυση παραγόντων αποτελεί και αυτή μια τεχνική μείωσης της διάστασης των δεδομένων. Αναλυτικότερα, στόχος της μεθόδου είναι η ελάττωση της διάστασης ενός συνόλου δεδομένων που αποτελείται από έναν αρκετά μεγάλο αριθμό συσχετισμένων μεταβλητών. Αυτό επιτυγχάνεται δημιουργώντας νέες μεταβλητές, ασυσχέτιστες μεταξύ τους, οι οποίες ονομάζονται παράγοντες.

Οι παράγοντες κατασκευάζονται τεχνητά με τη χρήση των αρχικών δεδομένων και δεν αποτελούν απαραίτητα υπαρκτές ποσότητες. Πολλές φορές στους παράγοντες αναγνωρίζουμε κάποιες μη μετρήσιμες ποσότητες, οι οποίες όμως ερμηνεύουν τις συσχετίσεις των αρχικών μεταβλητών. Επομένως, είναι σημαντικό να υπάρχουν συσχετίσεις ανάμεσα στις μεταβλητές που πρόκειται να μελετηθούν, αλλιώς δεν έχει νόημα η εφαρμογή της παραγοντικής ανάλυσης.

Στην παρούσα εφαρμογή που μελετάται σκοπός της ανάλυσης παραγόντων είναι να διαπιστωθεί αν υπάρχουν κάποιοι παράγοντες που επηρεάζουν την συμπεριφορά των πελατών καρτοκινητής.

4.2 Γενικό Μοντέλο Ανάλυσης Παραγόντων

Έστω ότι μελετώνται p συσχετισμένες τυχαίες μεταβλητές $X_1, X_2, \ldots, X_p$. Η προϋπόθεση οι μεταβλητές να είναι συσχετισμένες είναι απαραίτητη προκείμενου ο κάθε παράγοντας να ερμηνεύει παραπάνω από μία από τις αρχικές μεταβλητές.

Η κάθε μεταβλητή μπορεί να γραφεί ως άθροισμα ενός γραμμικού συνδυασμού m παραγόντων ($m<p$, όπου m ο αριθμός των παραγόντων που θα χρησιμοποιηθούν) και μιας ποσότητας που είναι διαφορετική για κάθε μεταβλητή.

Επομένως , προκύπτουν οι ισότητες:

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2$$
$$\vdots \qquad \vdots \qquad \vdots\vdots \qquad \vdots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p$$

Οι τυχαίες μεταβλητές $F_1, F_2, \ldots, F_m$ ονομάζονται παράγοντες.

Τα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ λέγονται σφάλματα ή ειδικοί παράγοντες.

Τα l_{ir} , i=1,2,..., p και j= 1,2,..., m ονομάζονται φορτία (loadings)

Στο μοντέλο της Ανάλυσης Παραγόντων γίνονται οι εξής παραδοχές:

- $E(F_r) = 0$, $V(F_r) = 1$, r=1,2,...,m δηλαδή οι τυχαίες μεταβλητές F_i θεωρούνται τυποποιημένες.
- $E(\varepsilon_\iota) = 0$, $V(\varepsilon_\iota) = \psi_i$, i=1,2,...,p
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$
- $Cov(\varepsilon_i, F_r) = 0$, i=1,2,…,p και r=1,2,…,m
- $Cov(F_r, F_s) = 0$, $r \neq s$ (οι παράγοντες είναι ασυσχέτιστοι- ορθογώνιοι μεταξύ τους)

Έπειτα από πράξεις προκύπτει για τη διακύμανση των τυχαίων μεταβλητών X_i η σχέση:

$$\sigma_i^2 = l_{i1}^2+l_{i2}^2+\ldots+l_{im}^2+\psi_i, \ i=1,2,\ldots,p$$

Η ποσότητα $l_{i1}^2+l_{i2}^2+\ldots+l_{im}^2$ ονομάζεται κοινή διακύμανση (communality) και αποτελεί το μέρος της διασποράς που οφείλεται στους παράγοντες $F_1, F_2, \ldots, F_m$.

Το υπόλοιπο κομμάτι της διασποράς ψ_i το οποίο δεν εξηγείται από τους παράγοντες λέγεται ειδική διασπορά (specific variance)

Η συνδιακύμανση σ_{ij} των τυχαίων μεταβλητών X_i έπειτα από πράξεις δίνεται από την σχέση:

$$\sigma_{ij}=l_{i1}l_{j1}+l_{i2}l_{j2}+\ldots+l_{im}l_{jm}$$

Η συνδιακύμανση της μεταβλητής X_i και του παράγοντα F_r δίνεται από τον τύπο:

$$Cov(X_i, F_r)=l_{ir}.$$

Ο συντελεστής συσχέτισης μεταξύ των μεταβλητών X_i και των παραγόντων F_r δίνεται από τον τύπο:

$$\rho(X_i, F_r)= \frac{l_{ir}}{\sigma_i} .$$

Όταν οι μεταβλητές X_i έχουν τυποποιηθεί τότε ο παραπάνω συντελεστής συσχέτισης δίνεται από τον τύπο $\rho(X_i, F_r)= l_{ir}$, επομένως τα φορτία δείχνουν το βαθμό συσχέτισης των μεταβλητών και των παραγόντων.

Ισοδύναμα, υπό μορφή πινάκων, το μοντέλο μπορεί να γραφεί ως εξής:

$$X-\mu= LF+\varepsilon \qquad (4.1)$$

όπου:

$$\boldsymbol{X}=\begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_p \end{bmatrix}, \boldsymbol{\mu}=\begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_p \end{bmatrix}, L=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix}, F=\begin{bmatrix} F_1 \\ F_2 \\ . \\ . \\ . \\ F_p \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_p \end{bmatrix}.$$

Ο πίνακας διακύμανσης γράφεται :

$$\Sigma = LL' + \Psi \ , \tag{4.2}$$

όπου:

$$\Psi=\begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

4.3 Περιστροφή παραγόντων

Οι παραπάνω πίνακες L και Ψ δεν είναι μοναδικοί.

Έστω ότι υπάρχει πίνακας M (ορθογώνιος) που ικανοποιεί τη σχέση

$$MM' = M'M = I_m \ . \tag{4.3}$$

Ο πίνακας $L_1 = LM$ ικανοποιεί τη σχέση (4.2) για τον ίδιο πίνακα Ψ. Επομένως, προκύπτει ότι:

$$\Sigma = L_1 L_1' + \Psi.$$

Η σχέση (4.1) αντικαθιστώντας με την ισότητα (4.3) γίνεται :

$$X - \mu = L_1(M'F) + \varepsilon$$

Από τη θεωρία πινάκων είναι γνωστό ότι αν ένα διάνυσμα πολλαπλασιαστεί με έναν ορθογώνιο πίνακα τότε αυτό περιστρέφεται στο χώρο. Συνεπώς, για αυτό η χρήση του L_1 αντί του L λέγεται περιστροφή των παραγόντων.

Σκοπός της περιστροφής των παραγόντων είναι η δημιουργία παραγόντων που επιδέχονται καλύτερη ερμηνεία. Οι βασικότερες μέθοδοι περιστροφής είναι οι:

Varimax: Ελαχιστοποιεί τον αριθμό των μεταβλητών που έχουν μεγάλα φορτία για όλους τους παράγοντες.

Quartimax: Ελαχιστοποιεί τον αριθμό των παραγόντων που εξηγούν μια μεταβλητή.

Equimax: Συνδυάζει τις 2 παραπάνω μεθόδους

Oblique: Πραγματοποιεί μη ορθογώνια περιστροφή. Οι παράγοντες που προκύπτουν δεν είναι ασυσχέτιστοι. Για το λόγο αυτό, η ερμηνεία των παραγόντων είναι πιο δύσκολη.

4.4 Μέθοδοι Εκτίμησης των φορτίων

Οι δύο βασικότερες μέθοδοι εκτίμησης των παραγόντων είναι η μέθοδος των κυρίων συνιστωσών και η μέθοδος της μεγίστης πιθανοφάνειας.

α. Η μέθοδος των κυρίων συνιστωσών

Στη μέθοδο αυτή γίνεται χρήση της τεχνικής της ανάλυσης των κυρίων συνιστωσών δημιουργώντας ασυσχέτιστες μεταξύ τους μεταβλητές. Η ανάλυση παραγόντων με τη μέθοδο των κυρίων συνιστωσών δεν αποτελεί τεχνική ελάττωσης της διάστασης των δεδομένων μέσω χρήσης γραμμικών συνδυασμών των αρχικών μεταβλητών (όπως η ανάλυση κυρίων συνιστωσών), αλλά μια τεχνική καθορισμού παραγόντων.

Στη μέθοδο των κυρίων συνιστωσών, με τη προσθήκη επιπλέον παραγόντων δεν αλλάζουν τα φορτία των προηγούμενων παραγόντων. Επιπλέον, η μέθοδος αυτή δεν θέτει περιορισμούς στον αριθμό των παραγόντων. Η μέθοδος εξαρτάται από τις μονάδες μέτρησης, επομένως πρέπει να γίνει επιλογή εάν θα χρησιμοποιηθεί ο πίνακας διακύμανσης ή συσχέτισης. Τέλος, μπορούν να υπολογιστούν ακριβώς οι τιμές-σκορ των παραγόντων με μεθόδους που θα αναφερθούν στη συνέχεια.

Η μέθοδος των κυρίων συνιστωσών προτιμάται

- αν η ανάλυση γίνεται κυρίως για περιγραφικούς σκοπούς
- αν δεν μπορεί να εξασφαλισθεί η κανονικότητα
- αν δεν είναι επιθυμητό η προσθήκη ή απαλοιφή ενός παράγοντα να αλλάζει τα προηγούμενα αποτελέσματα

β. Η μέθοδος της μεγίστης πιθανοφάνειας

Στη μέθοδο αυτή γίνεται η υπόθεση ότι τα χαρακτηριστικά $X_1, X_2, \ldots, X_p$ που μελετώνται ακολουθούν κάποια συγκεκριμένη κατανομή (συνήθως την πολυμεταβλητή κανονική) ούτως ώστε στη συνέχεια να γίνει χρήση της αντίστοιχης από κοινού συνάρτηση πυκνότητας για την εύρεση των εκτιμητριών μέγιστης πιθανοφάνειας των L και Ψ. Ο έλεγχος όμως της υπόθεσης της πολυμεταβλητής κανονικότητας δεν είναι εύκολο να ελεγχθεί. Σε αντίθεση με τη μέθοδο των κυρίων

συνιστωσών, αν το μοντέλο έχει κάποιο πρόβλημα η μέθοδος της μεγίστης πιθανοφάνειας υπάρχει περίπτωση να μην δουλεύει.

Επιπλέον, με τη μέθοδο αυτή υφίσταται περιορισμός στον αριθμό των παραγόντων που μπορούν να δημιουργηθούν (έχει διαπιστωθεί ότι ο μέγιστος αριθμός παραγόντων που μπορούν να εκτιμηθούν είναι $[p/2]$). Οι τιμές-σκορ των παραγόντων δεν μπορούν να υπολογιστούν ακριβώς διότι για τη μεγιστοποίηση της συνάρτησης πιθανοφάνειας απαιτείται η εφαρμογή επαναληπτικών αριθμητικών μεθόδων.

Η μέθοδος της μεγίστης πιθανοφάνειας προτιμάται :

- αν κατά την ανάλυση είναι επιθυμητό να γίνει στατιστική συμπερασματολογία
- αν μπορεί να εξασφαλιστεί η κανονικότητα
- αν υπάρχει ενδοιασμός σχετικά με το αν πρέπει να χρησιμοποιηθεί ο πίνακας διακυμάνσεων ή ο πίνακας συσχετίσεων.

Όταν το μοντέλο της Ανάλυσης Παραγόντων είναι κατάλληλο και το δείγμα αρκετά μεγάλο, οι μέθοδοι δίνουν συνήθως παρόμοια αποτελέσματα.

4.5 Εκτίμηση των τιμών των παραγόντων

Έστω ότι έχει εκτιμηθεί το μοντέλο $\boldsymbol{X-\mu = LF+\varepsilon}$ όπως παρουσιάστηκε παραπάνω.

Πολλές φορές είναι χρήσιμο να υπολογιστούν για κάθε παρατήρηση οι τιμές των παραγόντων.

Με $\hat{L}$ και $\hat{\Psi}$ θα συμβολιστούν οι εκτιμήτριες των L και Ψ αντίστοιχα.

Επομένως, θα πρέπει να εκτιμηθούν τα $F_1, F_2, \ldots, F_m$ στο μοντέλο $\boldsymbol{X-\mu = \hat{L}F+\varepsilon}$ όπου τα στοιχεία του διανύσματος $\boldsymbol{X-\mu}$ και του πίνακα $\hat{L}$ είναι γνωστοί πραγματικοί αριθμοί.

Το διάνυσμα F περιέχει άγνωστες παραμέτρους και το ε είναι τυχαίο διάνυσμα.

Οι συνηθέστερες μέθοδοι εκτίμησης των παραγόντων είναι οι ακόλουθες:

α. Regression Method

Η μέθοδος αυτή βασίζεται στη θεωρία της παλινδρόμησης και συγκεκριμένα στη μέθοδο των ελαχίστων τετραγώνων.

Οι εκτιμήτριες του F δίνονται από τον τύπο:

$$\boldsymbol{F} = \left(\hat{L}'\hat{L}\right)^{-1}\hat{L}'(\boldsymbol{X}-\boldsymbol{\mu}).$$

β .Bartlett Method

Η μέθοδος αυτή βασίζεται στη θεωρία των γενικευμένων ελαχίστων τετραγώνων, διότι γίνεται η υπόθεση ότι η διακύμανση των σφαλμάτων δεν είναι ίδια για όλες τις παρατηρήσεις.

Οι εκτιμήτριες του *F* δίνονται από τον τύπο:

$$\boldsymbol{F}= \left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\hat{L}'\hat{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}).$$

γ. Anderson Method

Η μέθοδος αυτή οδηγεί πάντα σε ασυσχέτιστους παράγοντες.

Οι εκτιμήτριες του F δίνονται από τον τύπο:

$$\boldsymbol{F}= \left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\left(I+\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-\frac{1}{2}}\hat{L}'\hat{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}).$$

4.6 Καθορισμός κατάλληλου πλήθους παραγόντων

Το πλήθος των παραγόντων δεν είναι εκ των προτέρων γνωστό και θα πρέπει να καθοριστεί πριν γίνει η εκτίμησή των παραγόντων. Υπάρχουν διάφορες μέθοδοι για να εκτιμηθεί το πλήθος αυτό.

Ο καθορισμός του αριθμού των παραγόντων θα μπορούσε να γίνει με παρόμοιες τεχνικές που χρησιμοποιήθηκαν και στην ανάλυση κυρίων συνιστωσών για την επιλογή του πλήθους των κυρίων συνιστωσών. Για παράδειγμα θα μπορούσε να γίνει χρήση του scree plot ή να επιλεχθεί κάποιο επιθυμητό ποσοστό της διακύμανσης που θα πρέπει να εξηγείται και να κρατηθούν οι αντίστοιχες ιδιοτιμές του πίνακα διακύμανσης-συνδιακύμανσης.

Επιπλέον, η επιλογή του κατάλληλου πλήθους παραγόντων θα μπορούσε να γίνει βάσει κάποιου κριτηρίου «καλής προσαρμογής» αυξάνοντας διαδοχικά τον αριθμό των παραγόντων. Τέτοια κριτήρια είναι:

- Από τον πίνακα επιβαρύνσεων μπορεί να εκτιμηθεί ο πίνακας *Σ*. Οι αποκλίσεις του πραγματικού πίνακα με τον εκτιμημένο (reproduced matrix) θα πρέπει να είναι μικρές.
- Έλεγχος λόγου πιθανοφανειών αν οι εκτιμήσεις έχουν γίνει με τη μέθοδο μεγίστης πιθανοφάνειας. Τέτοιοι έλεγχοι στηρίζονται σε υποθέσεις για την κατανομή του πληθυσμού.

Οι παραπάνω πληροφορίες για τη θεωρία της ανάλυσης Παραγόντων έχουν αντληθεί από τις σημειώσεις του Μεταπτυχιακού Εφαρμοσμένης Στατιστικής του

Πανεπιστημίου Πειραιώς, Κούτρας (2020), όπου ο αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες.

Για περισσότερες πληροφορίες σχετικά με την θεωρία της ανάλυσης παραγόντων, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Johnson και Wichern (1998), στο βιβλίο του Manly, (1986) και στις Σημειώσεις του Οικονομικού Πανεπιστημίου Αθηνών, του τμήματος Στατιστικής, Καρλής (2003).

Επιπλέον, ο αναγνώστης μπορεί συμπληρωματικά να ανατρέξει στις ιστοσελίδες:

https://thalis.math.upatras.gr/~vpiperig/Mul/factor.pdf

http://www.e-conometrics.gr/2013/11/02/factoranalysis/

https://stats.idre.ucla.edu/spss/output/factor-analysis/

4.7 Εφαρμογή Ανάλυσης Παραγόντων

Για να δικαιολογηθεί η ύπαρξη κοινών παραγόντων θα πρέπει οι μεταβλητές να έχουν μεταξύ τους μεγάλες συσχετίσεις, δηλαδή για να μπορέσει να εφαρμοστεί η μέθοδος της Ανάλυσης Παραγόντων, θα πρέπει να υπάρχει επαρκής συσχέτιση ανάμεσα στις αρχικές μεταβλητές. Για το λόγο αυτό χρησιμοποιούνται ο έλεγχος σφαιρικότητας του Bartlett και η Kaiser-Meyer-Olkin στατιστική συνάρτηση για την καταλληλόλητα των δεδομένων.

Κάνοντας παραγοντική ανάλυση στις μεταβλητές age, yearpos, resvoice, ressms, resdata, percentagevoiceuse, percentagesmsuse percentagedatause για τις παρατηρήσεις-συνδρομητές του δείγματος που χρησιμοποιούν τα bundles 1,3,5,7, χρησιμοποιώντας τη μέθοδο των κυρίων συνιστωσών για την εκτίμηση των φορτίων προκύπτουν τα ακόλουθα αποτελέσματα:

Πίνακας 4.1

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,456
Bartlett's Test of Sphericity	Approx. Chi-Square	372,856
	df	28
	Sig.	,000

Από τον πίνακα 4.1 φαίνεται ότι ο δείκτης KMO έχει αρκετά υψηλή τιμή (0,456). Επιπλέον, το p value που προκύπτει από τον έλεγχο του Bartlett είναι σχεδόν μηδέν, δηλαδή απορρίπτεται η μηδενική υπόθεση της σφαιρικότητας των δεδομένων.

Συμπερασματικά, εξασφαλίζεται η επάρκεια της συσχέτισης ανάμεσα στις μεταβλητές.

Στη συνέχεια, πρέπει να ληφθεί η απόφαση πόσοι παράγοντες πρέπει να χρησιμοποιηθούν.

Στον πίνακα 4.2 φαίνεται το ποσοστό της συνολικής διασποράς που ερμηνεύεται με την προσθήκη κάθε επιπλέον παράγοντα. Έχοντας, ως κριτήριο την προϋπόθεση οι ιδιοτιμές του πίνακα διακύμανσης-συνδιακύμανσης να είναι μεγαλύτερες της μονάδας, φαίνεται ότι κατάλληλη επιλογή είναι οι τρεις παράγοντες. Με τους τρεις παράγοντες ερμηνεύεται το 70,75% της συνολικής διακύμανσης, που αποτελεί αρκετά ικανοποιητικό ποσοστό.

Πίνακας 4.2

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,731	34,134	34,134	2,731	34,134	34,134
2	1,891	23,636	57,770	1,891	23,636	57,770
3	1,038	12,981	70,751	1,038	12,981	70,751
4	,883	11,044	81,794			
5	,782	9,781	91,575			
6	,426	5,319	96,894			
7	,205	2,560	99,454			
8	,044	,546	100,000			

Extraction Method: Principal Component Analysis.

Ωστόσο, ο πίνακας των communalities που παρουσιάζεται παρακάτω δείχνει ότι δεν ερμηνεύονται όλες οι μεταβλητές ικανοποιητικά από τους τρεις παράγοντες (όλες οι τιμές είναι μεγαλύτερες του 0,5 εκτός από αυτή της μεταβλητής pecentagesmsuse). Επομένως φαίνεται να χρειάζεται να προσθέσουμε κάποιον παράγοντα.

Communalities

	Initial	Extraction
age	1,000	,646
yearpos	1,000	,680
resvoice	1,000	,931
ressms	1,000	,783
resdata	1,000	,756
percentagevoiceuse	1,000	,759
percentagesmsuse	1,000	,284
percentagedatause	1,000	,822

Extraction Method: Principal Component Analysis.

Εκτελώντας την ίδια διαδικασία αλλά με τέσσερεις παράγοντες προκύπτει ο πίνακας των communalities:

Communalities

	Initial	Extraction
age	1,000	,709
yearpos	1,000	,815
resvoice	1,000	,950
ressms	1,000	,786
resdata	1,000	,764
percentagevoiceuse	1,000	,818
percentagesmsuse	1,000	,878
percentagedatause	1,000	,823

Extraction Method: Principal Component Analysis.

Πλέον όλες οι τιμές είναι ικανοποιητικές, επομένως η ανάλυση θα συνεχιστεί με τέσσερεις παράγοντες. Το ποσοστό της συνολικής διακύμανσης που ερμηνεύεται τώρα είναι 81,8%.

Ο reproduced correlations matrix είναι ο πίνακας των συσχετίσεων βασισμένος στους παράγοντες που έχουν εξαχθεί. Είναι επιθυμητό οι τιμές του πίνακα αυτού να είναι όσο το δυνατόν πιο κοντά στις τιμές του αρχικού πίνακα συσχετίσεων. Αυτό φαίνεται από τον πίνακα των residuals, όπου απεικονίζεται η διαφορά μεταξύ των δύο αυτών πινάκων που αναφέρθηκαν παραπάνω (reproduced correlations matrix και original correlations matrix). Από τον πίνακα των residuals δεν παρατηρούνται πολύ υψηλές τιμές, όλες είναι μικρότερες κατ' απόλυτη τιμή του 0,26, επομένως ο αρχικός πίνακας συσχετίσεων αναπαράγεται ικανοποιητικά με την ανάλυση με τέσσερεις παράγοντες.

Μια επιπλέον παρατήρηση που θα μπορούσε να γίνει από τον πίνακα συσχετίσεων, είναι ότι οι μεταβλητές yearpos και η percentagesmsuse δεν σχετίζονται με καμία από τις υπόλοιπες μεταβλητές.

Reproduced Correlations

		age	yearpos	resvoice	ressms	resdata	percentage voiceuse	percentage smsuse	percentage datause
Reproduced Correlation	age	,646[a]	,125	-,402	-,467	,562	,061	-,136	-,710
	yearpos	,125	,680[a]	-,168	-,400	-,227	-,235	,236	-,021
	resvoice	-,402	-,168	,931[a]	,780	-,050	-,668	-,269	,305
	ressms	-,467	-,400	,780	,783[a]	-,078	-,371	-,238	,377
	resdata	,562	-,227	-,050	-,078	,756[a]	-,073	-,372	-,722
	percentagevoiceuse	,061	-,235	-,668	-,371	-,073	,759[a]	,204	,009
	percentagesmsuse	-,136	,236	-,269	-,238	-,372	,204	,284[a]	,253
	percentagedatause	-,710	-,021	,305	,377	-,722	,009	,253	,822[a]
Residual[b]	age		-,113	,032	,048	-,101	-,052	,107	,148
	yearpos	-,113		,008	,197	,121	,263	-,148	-,034
	resvoice	,032	,008		,039	,036	,043	,151	-,031
	ressms	,048	,197	,039		,082	,188	,068	-,019
	resdata	-,101	,121	,036	,082		,088	,196	,018
	percentagevoiceuse	-,052	,263	,043	,188	,088		-,061	-,047
	percentagesmsuse	,107	-,148	,151	,068	,196	-,061		-,047
	percentagedatause	,148	-,034	-,031	-,019	,018	-,047	-,047	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 16 (57,0%) nonredundant residuals with absolute values greater than 0.05.

Παρατηρώντας τον component matrix φαίνεται ότι η μεταβλητή percentagedatause ερμηνεύεται και από τους δύο παράγοντες.

Component Matrix[a]

	Component			
	1	2	3	4
age	-,728	,317	,124	,252
yearpos	-,171	-,294	,751	-,368
resvoice	,803	,511	,159	,136
ressms	,771	,380	-,209	-,061
resdata	-,486	,708	-,136	,094
percentagevoiceuse	-,400	-,493	-,597	-,242
percentagesmsuse	-,023	-,522	,105	,771
percentagedatause	,732	-,530	-,068	,034

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Προκειμένου να αποφευχθεί αυτό θα γίνει περιστροφή των παραγόντων με τη μέθοδο varimax. Πλέον, βάσει του rotated component matrix, που έχει προκύψει μετά την περιστροφή, κάθε μεταβλητή εξηγείται κυρίως από έναν μόνο από τους παράγοντες. Οι μεταβλητές age, resdata, percentagedatause εξηγούνται από τον πρώτο παράγοντα, οι resvoice, ressms και percentagevoiceuse από τον δεύτερο, η yearpos από τον τρίτο και η percentagesmsuse από τον τέταρτο. Οι δύο τελευταίες μεταβλητές, οι yearpos και η percentagesmsuse, αποτελούν από μόνες τους ξεχωριστούς παράγοντες, διότι, όπως αναφέρθηκε προηγουμένως, δεν συσχετίζονται με καμία από τις υπόλοιπες μεταβλητές .

Rotated Component Matrix[a]

	Component			
	1	2	3	4
age	-,807	-,138	-,097	,173
yearpos	,063	-,015	-,899	-,040
resvoice	,240	,898	,251	-,152
ressms	,373	,578	,451	-,331
resdata	-,811	,073	,242	-,206
percentagevoiceuse	,113	-,875	,187	-,067
percentagesmsuse	,136	-,068	,022	,925
percentagedatause	,877	,117	,088	,180

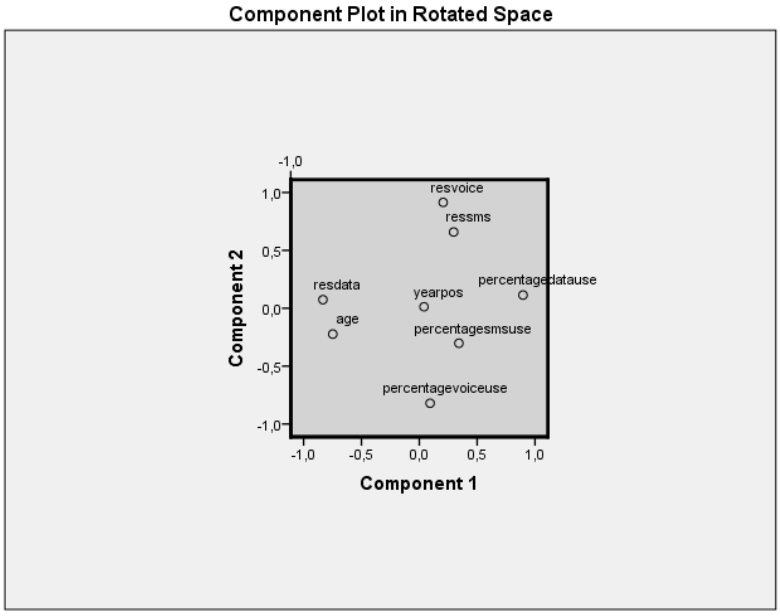
Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Το συμπέρασμα αυτό επαληθεύεται και παρατηρώντας το σχήμα 4.1 (component plot), όπου οι μεταβλητές age, resdata, percentagedatause ταυτίζονται με τον πρώτο παράγοντα, ενώ οι resvoice, ressms και percentagevoiceuse με τον δεύτερο. Οι μεταβλητές yearpos και percentagesmsuse είναι κοντά στην αρχή των αξόνων διότι δεν ταυτίζονται με κανέναν από τους πρώτους δύο παράγοντες.

Σχήμα 4.1



Ο πρώτος παράγοντας εξηγεί την ηλικία, το υπόλοιπο των δεδομένων καθώς και το ποσοστό χρήσης των δεδομένων. Επομένως, ο πρώτος παράγοντας θα μπορούσε να εκφράζει την εξοικείωση του πελάτη με την τεχνολογία.

Ο δεύτερος παράγοντας εξηγεί το υπόλοιπο του χρόνου ομιλίας, το υπόλοιπο των μηνυμάτων και το ποσοστό χρήσης της ομιλίας. Επομένως ο δεύτερος παράγοντας θα μπορούσε να θεωρηθεί ότι εκφράζει την ανάγκη για επικοινωνία.

Συμπερασματικά, η εταιρία θα πρέπει να λάβει υπόψιν της τους δύο παράγοντες που σχηματίζονται. Δηλαδή, να διαμορφώσει πακέτα καρτοκινητής που να ταιριάζουν στο προφίλ χρηστών εξοικειωμένων με τις καινούριες τεχνολογίες, γεγονός που συνεπάγεται ότι έχουν αυξημένη ανάγκη για δεδομένα ίντερνετ και μικρότερα για ομιλία και μηνύματα. Τα άτομα αυτά είναι ως επί το πλείστον άτομα μικρότερων ηλικιών. Επιπλέον, θα πρέπει να διαμορφώσει ξεχωριστά πακέτα καρτοκινητής που να εξυπηρετούν τους χρήστες που έχουν ανάγκη για επικοινωνία, δηλαδή πακέτα με πολλά λεπτά ομιλίας, αρκετά μηνύματα και λιγότερα δεδομένα ίντερνετ. Στη δεύτερη κατηγορία θα μπορούσαν να υπάρχουν και επιπλέον προσφορές για επαγγελματίες, που είναι αυξημένη η ανάγκη τους για λεπτά ομιλίας και μηνύματα. Φυσικά, θα πρέπει να υπάρξει και πακέτο που να καλύπτει τις ανάγκες συνδρομητών που έχουν ανάγκη και για επικοινωνία μέσω τηλεφώνου και μηνυμάτων αλλά και ίντερνετ.

Με αυτόν τον τρόπο επιτυγχάνεται να μπορεί ο κάθε πελάτης να αγοράσει το πακέτο που τον συμφέρει, χωρίς να του «μένουν» παροχές αχρησιμοποίητες και να του δημιουργείται η εντύπωση ότι πληρώνει ασκόπως υπηρεσίες.

Κεφάλαιο 5

5.1Ανακεφαλαίωση

Κατόπιν συλλογής δεδομένων που αφορούν στη χρήση προγραμμάτων καρτοκινητής από πελάτες εταιρίας κινητής τηλεφωνίας, έγινε εφαρμογή μεθόδων Πολυμεταβλητής Ανάλυσης. Οι μέθοδοι που χρησιμοποιήθηκαν ήταν η Ανάλυση κατά συστάδες, η Ανάλυση κυρίων συνιστωσών και η Ανάλυση Παραγόντων.

Συμπερασματικά, από την Ανάλυση κατά Συστάδες, εφαρμόζοντας ιεραρχική μέθοδο με τη χρήση δενδρογράμματος, προκύπτει ότι οι συνδρομητές ανάλογα με τη χρήση που κάνουν στο καρτοκινητό τους, κατατάσσονται σε τρεις ομάδες- κατηγορίες. Η διαφοροποίηση γίνεται βάσει της ηλικίας των συνδρομητών, των χρόνων κατοχής του καρτοκινητού τους και των ποσοστών χρήσης του χρόνου ομιλίας, των μηνυμάτων και των δεδομένων ίντερνετ του μηνιαίου προγράμματος που έχουν επιλέξει. Η ίδια διαφοροποίηση των ομάδων προκύπτει και μετά την εφαρμογή της μη ιεραρχικής μεθόδου K-means έχοντας επιλέξει τρεις ως τον αριθμό ομάδων που θα γίνει η κατάταξη. Η κατανομή των συνδρομητών στις τρεις αυτές ομάδες έχει γίνει με σχεδόν ίδιο τρόπο με την ιεραρχική μέθοδο. Η εταιρία θα μπορούσε να συγκεντρώσει περισσότερα δεδομένα για κάθε (ομοιογενή) ομάδα και με βάση αυτά να καθορίσει την πολιτική marketing που θα εφαρμόσει σε κάθε ομάδα ξεχωριστά.

Από την Ανάλυση των Κυρίων Συνιστωσών προκύπτει ότι σχηματίζονται τρείς κύριες συνιστώσες, ασυσχέτιστες με τις αρχικές μεταβλητές των δεδομένων. Παρατηρείται ο διαχωρισμός ενός συγκεκριμένου προγράμματος καρτοκινητής (συνδρομητές μικρής ηλικίας που χρησιμοποιούν πολλά δεδομένα ίντερνετ και αφήνουν πολλά αχρησιμοποίητα λεπτά ομιλίας και μηνύματα) από τα υπόλοιπα.

Από την Ανάλυση Παραγόντων σχηματίζονται δύο παράγοντες, οι οποίοι επηρεάζουν τη συμπεριφορά των πελατών καρτοκινητής. Ο πρώτος παράγοντας ερμηνεύεται ως η εξοικείωση του πελάτη με την τεχνολογία, ενώ ο δεύτερος ως η ανάγκη του πελάτη για επικοινωνία λόγω εργασίας.

Επιπλέον, η ανάλυση έγινε και ξεχωριστά για τους άνδρες και τις γυναίκες συνδρομητές του δείγματος και δεν φάνηκε να υπάρχουν τόσο σημαντικές διαφορές

ανάμεσα στα δύο φύλα, ώστε να κριθεί ότι η εταιρία πρέπει να δείξει διαφορετική αντιμετώπιση στις δύο υποομάδες.

Συνοψίζοντας τα συμπεράσματα που προέκυψαν από τις τρεις μεθόδους, η εταιρία κινητής τηλεφωνίας θα πρέπει να διαμορφώσει τα προγράμματα καρτοκινητής που προσφέρει. Ένα πακέτο θα πρέπει να προσφέρει πολλά λεπτά ομιλίας, λίγα μηνύματα και λίγα δεδομένα ίντερνετ και θα απευθύνεται κατά κύριο λόγο σε συνδρομητές μεγαλύτερων ηλικιών. Ένα δεύτερο πακέτο θα πρέπει να προσφέρει πολλά λεπτά ομιλίας, πολλά δεδομένα ίντερνετ και αρκετά μηνύματα και θα απευθύνεται κυρίως σε συνδρομητές μεσαίων ηλικιών. Τέλος, ένα τρίτο πακέτο, το οποίο θα στοχεύει τις μικρότερες ηλικιακές ομάδες, θα προσφέρει πολλά δεδομένα ίντερνετ, λίγα λεπτά ομιλίας και λίγα μηνύματα.

Με τον τρόπο αυτό εξασφαλίζεται για τον κάθε πελάτη η παροχή μόνο υπηρεσιών που έχει ανάγκη και χρησιμοποιεί. Αυτό θα έχει σαν αποτέλεσμα ο πελάτης να είναι ευχαριστημένος και να μην «νιώθει» ότι πληρώνει παροχές που δεν χρειάζεται.

Βιβλιογραφία

Ξένη:

Everitt, B. S. and Dunn, G. (1991). *Applied Multivariate Data Analysis* Arnold, New York.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis* Prentice Hall, New Jersey.

Jolliffe, I. T. (2002). *Principal Component Analysis*, second edition, Springer-Verlag.

Manly, B. F. J. (1986). *Multivariate Statistical Methods: A primer* Chapman and Hall, London.

Ελληνική:

Κούτρας Μ., *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση*, Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης (2020), Πανεπιστήμιο Πειραιώς.

Καρλής Δ. *Πολυμεταβλητή Στατιστική Αναλυση*, Σημειώσεις Προπτυχιακού Μαθήματος του τμήματος Στατιστικής (2003) Οικονομικό Πανεπιστήμιο Αθηνών.

Ηλεκτρονική Βιβλιογραφία:

https://stats.idre.ucla.edu/spss/output/factor-analysis/

Norusis, SPSS cluster analysis Xavier University Intro Data Mining for Managers: http://www.norusis.com/pdf/SPC_v13.pdf

Πιπερίγκου Β. Πανεπιστήμιο Πατρών Τμήμα Μαθηματικών: https://thalis.math.upatras.gr/~vpiperig/Mul/factor.pdf

http://www.e-conometrics.gr/2013/11/02/factoranalysis/