# Στατιστική Ανάλυση Δεδομένων

## Εξάμηνο 1ο, 2016-2017


# Ανάλυση Αθέτησης Πληρωμών Πιστωτικών Καρτών, Απρίλιος - Σεπτέμβριος 2005, Ταϊβάν

**Καραποστολάκης Σωτήριος**

**ΑΕΜ: 666**

**Περιεχόμενα**

# 1. Εισαγωγή

Η στατιστική ανάλυση που ακολουθεί έγινε στα πλαίσια του μαθήματος Στατιστική Ανάλυση Δεδομένων και αφορά την εξαγωγή συμπερασμάτων από ένα σύνολο δεδομένων συναλλαγών πιστωτικών καρτών. Τα στοιχεία του συνόλου δεδομένων αφορούν το εξάμηνο Απρίλιος-Σεπτέμβριος 2005, ενώ η χώρα που έλαβε χώρα η καταμέτρηση αυτών είναι το Ταϊβάν. Ο βασικός λόγος της καταγραφής τους ήταν για την εξαγωγή συμπεράσματος σχετικά με το εάν ένας πελάτης επρόκειτο να αθετήσει τις πληρωμές του ή όχι.
Τα δεδομένα συλλέχθηκαν και επεξεργάστηκαν για ερευνητικούς σκοπούς. Πιο συγκεκριμένα για την εξόρυξη γνώσης, με την πρόβλεψης της φερεγγυότητας ενός πελάτη. Η καλύτερη πρόβλεψη για ένα πελάτη έγινε με τη χρήση ενός νευρωνικού δικτύου που χρησιμοποιήθηκε για αυτό το σκοπό.
Το σύνολο δεδομένων αποτελείται από 30000 πελάτες. Ο διαχωρισμός των δεδομένων έγινε με βάση 23 ανεξάρτητων μεταβλητών και 1 δυαδικής εξαρτημένης μεταβλητής που χαρακτήριζε εάν ο πελάτης αθετεί ή όχι τις πληρωμές του. Η τιμή που παίρνει η εξαρτημένη μεταβλητή όπως αναφέραμε είναι προϊόν πρόβλεψης, μέσω ενός μοντέλου μηχανικής μάθησης. Οι ανεξάρτητες μεταβλητές είναι οι εξής:

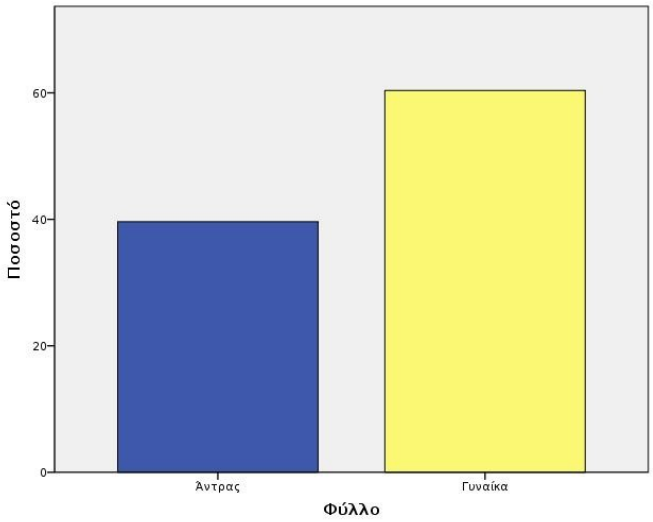
1. Μέγιστο όριο αγορών(Συμπεριλαμβανομένου και της οικογενείας του πελάτη)
2. Φύλλο
3. Εκπαίδευση
4. Κατάσταση γάμου
5. Ηλικία
6. Αριθμός μηνών καθυστέρησης πληρωμών για το μήνα Σεπτέμβριο 2005
7. Αριθμός μηνών καθυστέρησης πληρωμών για το μήνα Αύγουστο 2005
8. Αριθμός μηνών καθυστέρησης πληρωμών για το μήνα Ιούλιο 2005
9. Αριθμός μηνών καθυστέρησης πληρωμών για το μήνα Ιούνιο 2005
10. Αριθμός μηνών καθυστέρησης πληρωμών για το μήνα Μάιο 2005
11. Αριθμός μηνών καθυστέρησης πληρωμών για το μήνα Απρίλιο 2005
12. Ποσό λογαριασμού μήνα Σεπτεμβρίου 2005
13. Ποσό λογαριασμού μήνα Αυγούστου 2005
14. Ποσό λογαριασμού μήνα Ιουλίου 2005
15. Ποσό λογαριασμού μήνα Ιουνίου 2005
16. Ποσό λογαριασμού μήνα Μαΐου 2005
17. Ποσό λογαριασμού μήνα Απριλίου 2005
18. Ποσό αποπληρωμής μήνα Σεπτεμβρίου 2005
19. Ποσό αποπληρωμής μήνα Αυγούστου 2005
20. Ποσό αποπληρωμής μήνα Ιουλίου 2005
21. Ποσό αποπληρωμής μήνα Ιουνίου 2005
22. Ποσό αποπληρωμής μήνα Μαΐου 2005
23. Ποσό αποπληρωμής μήνα Απριλίου 2005

Για την καλύτερη κατανόηση των δεδομένων δημιουργήθηκαν 2 επιπλέον μεταβλητές. Η πρώτη αφορούσε το μέσο ποσό οφειλής για το εξάμηνο Απρίλιος-Σεπτέμβριος, ενώ η δεύτερη το μέσο ποσό αποπληρωμής για το εξάμηνο Απρίλιος-Σεπτέμβριος.
Κατά την επεξεργασία θεωρήθηκε θεμιτό η προσπάθεια ανακάλυψης και κάποιων μη-μετρήσιμων χαρακτηριστικών. Τα εν λόγω χαρακτηριστικά όμως κρίθηκαν ότι έχουν σημαντική επιρροή στα δεδομένα. Με τη χρήση της κατάλληλης τεχνικής εξάχθηκαν 2 καινούργιες μεταβλητές όπου η πρώτη αφορούσε την οικονομική δύναμη ενός πελάτη, συνυπολογίζοντας τις μεταβλητές του ορίου αγορών και την καθυστέρηση στην αποπληρωμή του μήνα Απριλίου. Η δεύτερη περιγράφει την τιμιότητα ενός πελάτη. Ο υπολογισμός έγινε λαμβάνοντας υπόψη όλες τις μεταβλητές μέτρησης

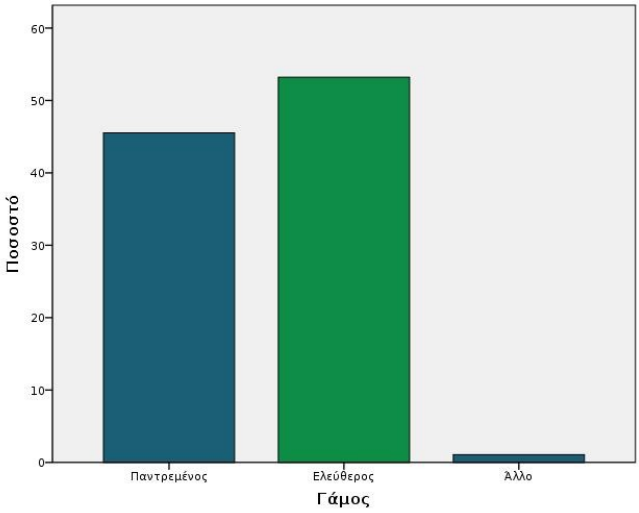
της καθυστέρησης πληρωμής για το κάθε μήνα. Στο τελευταίο κομμάτι ασχοληθήκαμε με τεχνικές ομαδοποίησης των πελατών με βάση πληροφοριών του συνόλου δεδομένων.

# 2. Περιγραφική Στατιστική

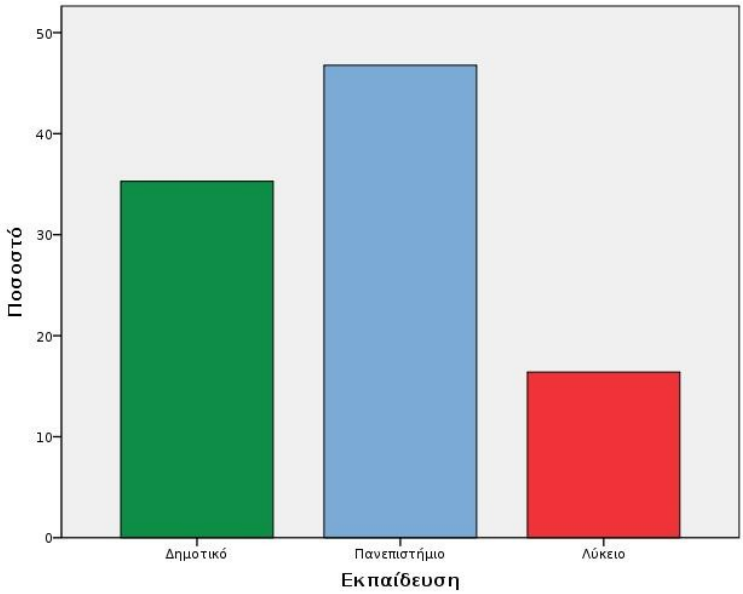
Στο κεφάλαιο αυτό επιδιώκεται μία ανάλυση του συνόλου δεδομένων, μέσω ορισμένων τεχνικών περιγραφικής στατιστικής. Στόχος αποτελεί η καλύτερη κατανόησή του.
Με μία πρώτη παρατήρηση βγαίνει το συμπέρασμα ότι ο αριθμός των γυναικών υπερτερεί του αριθμού των αντρών(σχήμα 1). Στον τομέα της εκπαίδευσης το μεγαλύτερο πλήθος είναι απόφοιτοι πανεπιστημίου, ακολουθούν οι απόφοιτοι δημοτικού και τελευταίοι είναι οι απόφοιτοι λυκείου(σχήμα 3). Ενώ όσον αναφορά την κατάσταση γάμου παρατηρείται μία μικρή υπεροχή των μη-παντρεμένων(ελεύθερων)(σχήμα 2).




**Σχήμα 1: Ραβδόγραμμα φύλλου**




**Σχήμα 2: Ραβδόγραμμα κατάστασης γάμου**




**Σχήμα 3: Ραβδόγραμμα εκπαιδεύσεως**

Στον πίνακα 1 απαριθμούνται ορισμένα στατιστικά στοιχεία για την ηλικία και το όριο αγορών κάθε πελάτη. Αξίζει να σημειωθεί ότι στο όριο αγορών συνυπολογίζεται το συνολικό οικογενειακό

όριο. Παρατηρούμε ότι το μέσο όριο αγορών είναι τα 5359 δολάρια. Ενώ η μέση ηλικία κυμαίνεται στα 35.5 έτη περίπου.

| | Ελάχιστο | Μέγιστο | Μέσος Όρος |
|---|---|---|---|
| Ηλικία | 21 | 79 | 35.49 |
| Όριο αγορών | 320(δολάρια) | 32000(δολάρια) | 5359.4(δολάρια) |

**Πίνακας 1: Στοιχεία ηλικίας και ορίου αγορών**

Μέχρις στιγμής ασχοληθήκαμε με κάποια δημογραφικά χαρακτηριστικά. Τώρα θα παρουσιαστούν κάποια περιγραφικά στοιχεία για το πως συμπεριφέρονται σαν πελάτες. Με βάση του πίνακες 2 και 3 παρατηρείται ότι όσο περνάνε οι μήνες οι οφειλές των πελατών αυξάνονται όλο και περισσότερο. Στην αντίπερα όχθη και στο τομέα των εξοφλήσεων το μέσο ποσό που δίνεται παρουσιάζει αρκετές αυξομειώσεις. Αρχικά φθίνει όμως όσο περνούν οι μήνες αρχίζει και αυξάνεται μάλλον λόγω της αύξησης των οφειλών τους.

| Μήνας | Μέγιστο | Μέσος Όρος |
|---|---|---|
| Απρίλιος | 30773.2 | 1205 |
| Μάιος | 29669.4 | 1289.9 |
| Ιούνιος | 28530.7 | 1384.3 |
| Ιούλιος | 53250.8 | 1504.4 |
| Αύγουστος | 31485.7 | 1573.7 |
| Σεπτέμβριος | 30864.3 | 1639.1 |

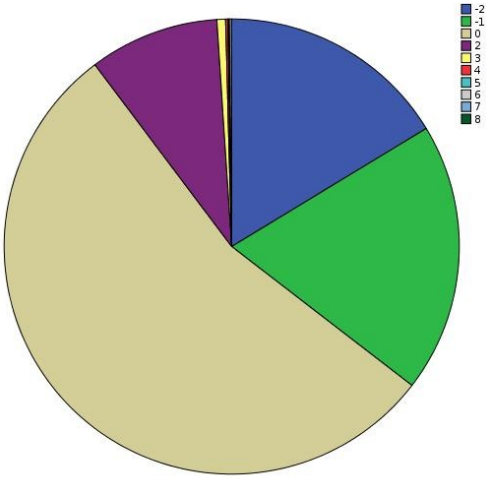
**Πίνακας 2: Στοιχεία οφειλών(δολάρια)**

| Μήνας | Μέγιστο | Μέσος Όρος |
|---|---|---|
| Απρίλιος | 16917.3 | 166.8 |
| Μάιος | 13648.9 | 153.5 |
| Ιούνιος | 19872 | 154.4 |
| Ιούλιος | 28673.2 | 167.2 |
| Αύγουστος | 53896 | 189.4 |
| Σεπτέμβριος | 27953 | 181.2 |

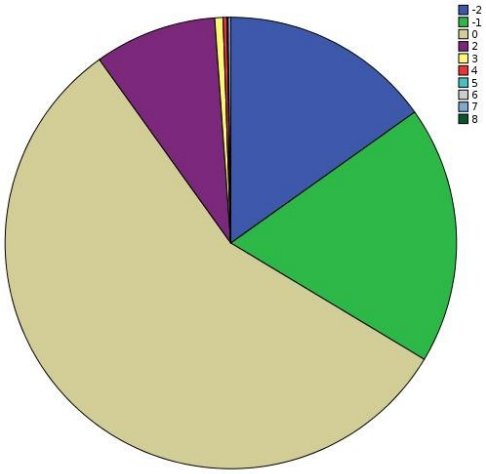
**Πίνακας 3: Στοιχεία εξοφλήσεων(δολάρια)**

Το σύνολο δεδομένων όπως γνωρίζουμε περιέχει τις καθυστερήσεις του κάθε πελάτη για την εξόφληση του λογαριασμού τον κάθε μήνα. Από την ανάλυση των συγκεκριμένων χαρακτηριστικών προκύπτει το συμπέρασμα ότι περίπου το 50% των πελατών καθυστερούν τις

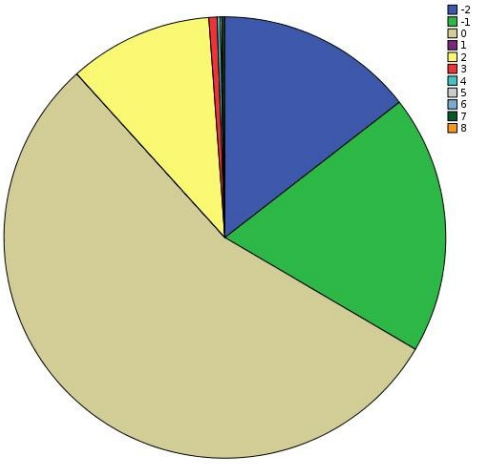
πληρωμές τους 1 μήνα. Ενώ το ποσοστό των εμπρόθεσμων εξοφλήσεων κυμαίνεται κατά τη διάρκεια του εξαμήνου γύρω στο 20%. Είναι άξιας αναφοράς ότι το ποσοστό των πελατών με κάποιο διακανονισμό το μήνα Απρίλιο υπολογίζεται στο 16.3%, αντιθέτως το μήνα Σεπτέμβριο σημειώνεται στο 9.2%. Δηλαδή ακολουθεί πτωτική τάση κατά τη διάρκεια του εξαμήνου. Στα διαγράμματα που ακολουθούν αντικατοπτρίζονται οι καθυστερήσεις για το κάθε μήνα.



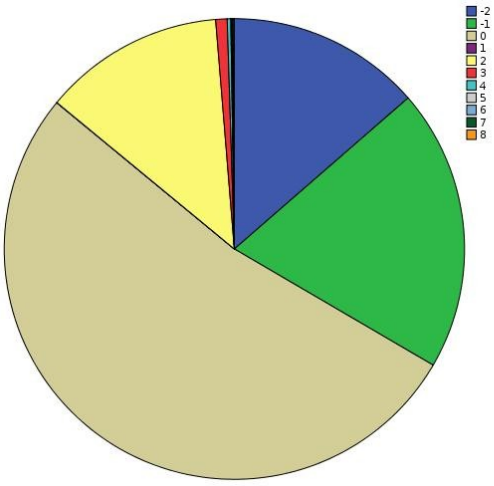

**Σχήμα 4: Καθυστερήσεις Απριλίου**



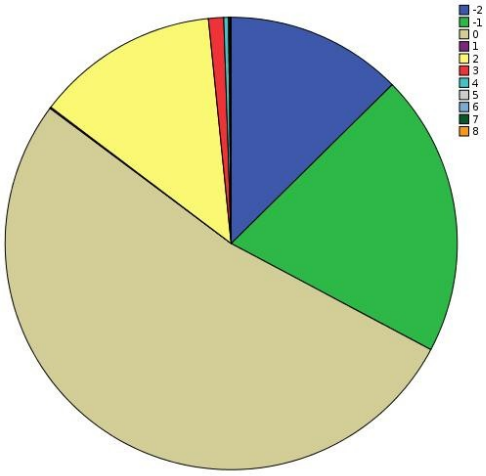

**Σχήμα 5: Καθυστερήσεις Μαΐου**



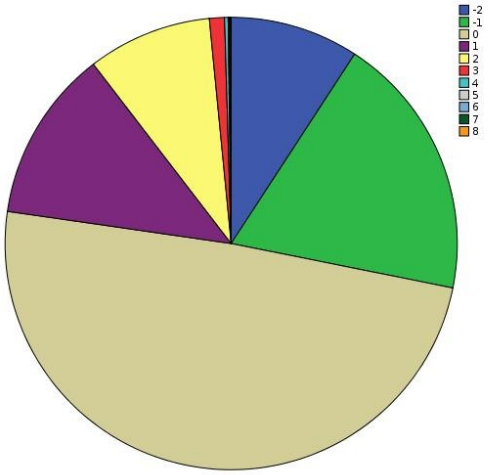

**Σχήμα 6: Καθυστερήσεις Ιουνίου**




**Σχήμα 7: Καθυστερήσεις Ιουλίου**



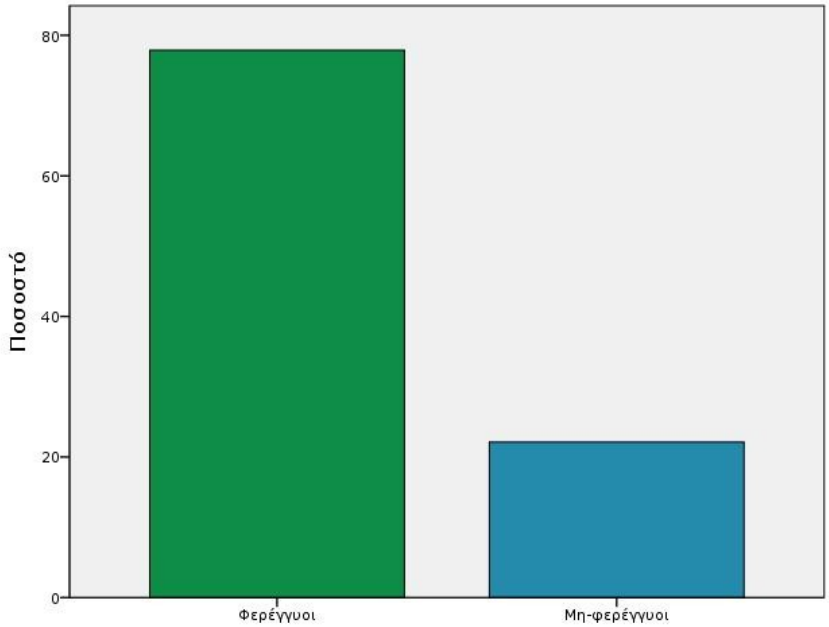

**Σχήμα 8: Καθυστερήσεις Αυγούστου**




**Σχήμα 9: Καθυστερήσεις Σεπτεμβρίου**

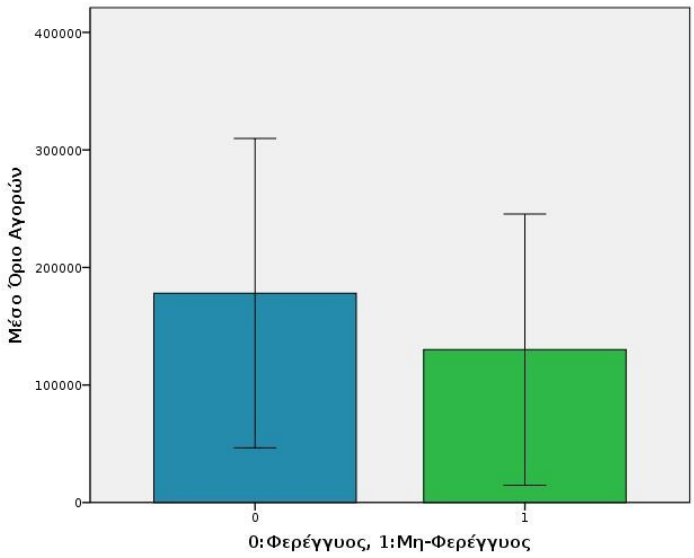
Σχετικά με τα νούμερα που αντιστοιχούν σε κάθε κομμάτι του διαγράμματος. Ο αριθμός -1 υποδηλώνει ότι ο πελάτης είναι απολύτως φερέγγυος με τις πληρωμές του. Ο αριθμός -2 υποδηλώνει ότι ο πελάτης βρίσκεται σε κάποιο πρόγραμμα ρύθμισης των οφειλών του. Ο αριθμός 0 δηλώνει καθυστέρηση ενός μηνός, ο αριθμός 1 καθυστέρηση 2 μηνών. Η ίδια λογική ακολουθείται και για μεγαλύτερες τιμές.

Η ανάλυση του συνόλου δεδομένων μέχρι στιγμής επικεντρώθηκε στην κατανόηση του πληθυσμού που γίνεται η έρευνα. Τώρα θα παρουσιαστούν κάποια περιγραφικά στοιχεία σχετικά με την πρόβλεψη που μας ενδιαφέρει στο δείγμα. Με άλλα λόγια σχετικά με το εάν επρόκειτο να αθετήσουν ή όχι τις πληρωμές τους. Με βάση το σχήμα 10 το ποσοστό των φερέγγυων πελατών υπολογίζεται στο 77.9%, ενώ στην αντίπερα όχθη το ποσό των μη-φερέγγυων υπολογίζεται στο 22.1%.




**Σχήμα 10: Αναπαράσταση συμπεριφοράς πελατών**

Το όριο αγορών ενός πελάτη δίνει μία πρώτη γεύση σχετικά με τη δυνατότητα αποπληρωμής μίας πιστωτικής κάρτας. Αυτό εξάγεται από την υπόθεση ότι ένας πελάτης με μεγάλο όριο αγορών θα είναι πιο δυνατός οικονομικά και από κάποιον αντίστοιχο με χαμηλότερο όριο. Βασιζόμενοι σε αυτό στο σχήμα 11 συγκρίναμε το μέσο όριο αγορών με βάση τη φερεγγυότητα ενός πελάτη και καταλήξαμε ότι οι μη-φερέγγυοι πελάτες έχουν πιο χαμηλό όριο αγορών.




**Σχήμα 11: Αναπαράσταση μέσου ορίου αγορών σε σχέση με φερεγγυότητα πελάτη**

**Τεχνικός Σχολιασμός**

Η εξαγωγή των στατιστικών στοιχείων έγινε με χρήση των επιλογών frequencies και descriptives που παρέχει το SPSS στην κατηγορία Descriptives Statistics. Η εξαγωγή των διαγραμμάτων έγινε με επιλογή barcharts ή piecharts κατά την εφαρμογή των frequencies και descriptives.

# 3. Στατιστική Συμπερασματολογία

Αντικείμενο του κεφαλαίου είναι η εξαγωγή πιο πολύπλοκων στατιστικών συμπερασμάτων από το σύνολο δεδομένων. Αρχικά θα παρουσιαστούν ορισμένες συσχετίσεις ανάμεσα σε κάποιες κατηγορικές και κάποιες αριθμητικές μεταβλητές του δείγματος. Επόμενο στάδιο είναι η ανάλυση των μέσων όρων ορισμένων μεταβλητών ανάμεσα στους διαφορετικούς πληθυσμούς που χωρίζουν ορισμένες μεταβλητές το σύνολο δεδομένων. Ακολουθεί η εξαγωγή των παραγόντων γύρω από τους οποίους ομαδοποιούνται κάποιες μεταβλητές και τελευταίο κομμάτι είναι η εφαρμογή κάποιων τεχνικών ομαδοποίησης μέσω των πληροφοριών των δεδομένων. Οι συγκεκριμένες τεχνικές εφαρμόζονται για να μετρηθεί κατά πόσο γίνεται σωστή κατανομή των πελατών σε αυτούς που επρόκειτο να αθετήσουν τις πληρωμές τους και σε αυτούς που είναι φερέγγυοι μέσω της στατιστικής. Τελευταίο κομμάτι αποτελεί η εφαρμογή τεχνικών clustering. Ο λόγος της χρησιμοποίησης τους είναι για τη δημιουργία κάποιων ομάδων πελατών που μοιάζουν στη συμπεριφορά τους.

## 3.1. Ποιοτικές Συσχετίσεις

Η πρώτη συσχέτιση που ελέγξαμε είναι αυτή του βαθμού εκπαίδευσης σε συνάρτηση με το φύλλο. Ο λόγος είναι η κατανόηση του μορφωτικού επιπέδου στο σύνολο των γυναικών και ανάμεσα στο σύνολο των αντρών. Στον πίνακα 4 αναφέρονται τα αποτελέσματα της συσχέτισης φύλλου-εκπαίδευσης. Με μία πρώτη ματιά παρατηρούμε ότι η μόρφωση των γυναικών είναι ανώτερη από την αντίστοιχη των αντρών. Η διαφορά βρίσκεται στο δημοτικό, όπου οι άντρες εμφανίζουν ποσοστό της τάξεως του 36.6%, ενώ οι γυναίκες 34.4%. Η διαφορά αυτή παρουσιάζεται αντίθετη στην εκπαίδευση πανεπιστημίου. Εκεί οι άντρες έχουν ποσοστό 45.7%, ενώ οι γυναίκες 47.8%.

| | | | Εκπαίδευση | | | |
|---|---|---|---|---|---|---|
| | | | Δημοτικό | Πανεπιστήμιο | Λύκειο | Σύνολο |
| Φύλλο | Άντρας | % επί το φύλλο | 36.6% | 45.2% | 16.7% | 39.6% |
| | | % επί της εκπαιδεύσεως | 41.1% | 38.3% | 40.5% | |
| | Γυναίκα | % επί το φύλλο | 34.4% | 47.8% | 16.2% | 60.4% |
| | | % επί της εκπαιδεύσεως | 58.9% | 61.7% | 59.5% | |
| Σύνολο | | % επί το φύλλο | 35.3% | 46.8% | 16.4% | |

**Πίνακας 4: Σχέση εκπαίδευσης-φύλλου**

Προχωρώντας θεωρήσαμε σωστό να ελέγξουμε εάν υπάρχει σχέση ανάμεσα στο φύλλο και τη φερεγγυότητα ενός πελάτη. Τα αποτελέσματα φαίνονται στον πίνακα 4. Το αρχικό συμπέρασμα που εξάγεται είναι ότι οι γυναίκες είναι κατά 4% πιο φερέγγυες από τους άντρες. Συγκεκριμένα το 75.8% από τους άντρες είναι φερέγγυοι, ενώ το αντίστοιχο ποσοστό στις γυναίκες κυμαίνεται στο 79.2%. Η εν λόγω διαφορά φανερώνεται και στους πληθυσμούς των φερέγγυων και μη-φερέγγυων. Στον πληθυσμό των φερέγγυων το 38.6% είναι άντρες, ενώ το 61.4% είναι γυναίκες. Στους μη-φερέγγυους πελάτες τα ποσοστά είναι 43.3% και 56.7%. Μπορεί κανείς υπολογίζοντας τη διαφορά να καταλήξει στο συμπέρασμα που αναφέραμε.

| | | | Αθέτηση Πληρωμών | | |
|---|---|---|---|---|---|
| | | | Όχι | Ναι | Σύνολο |
| Φύλλο | Άντρας | % επί το φύλλο | 75.8% | 24.2% | 39.6% |
| | | % επί την αθέτηση πληρωμών | 38.6% | 43.3% | |
| | Γυναίκα | % επί το φύλλο | 79.2% | 20.8% | 60.4% |
| | | % επί την αθέτηση πληρωμών | 61.4% | 56.7% | |
| Σύνολο | | % επί το φύλλο | 77.9% | 22.1% | |

**Πίνακας 5: Σχέση φύλλου-φερεγγυότητας**

Επόμενη σχέση που μελετήσαμε ήταν αυτή της εκπαιδεύσεως με την αθέτηση πληρωμών. Ο λόγος που μας οδήγησε σε κάτι τέτοιο είναι διότι θέλαμε να δούμε εάν η αύξηση του μορφωτικού επιπέδου αύξανε και τη συνέπεια στις πληρωμές, η οποία μετουσιώνεται στην ανάλυσή μας ως φερεγγυότητα στις πληρωμές. Πιστεύουμε ότι ένας πελάτης με μεγαλύτερη μόρφωση θα έχει μεγαλύτερες πιθανότητες να είναι συνεπείς στις πληρωμές του. Πρώτον λόγω πιθανόν καλύτερης δουλειάς. Και δεύτερον λόγω αυξημένων ηθικών αξιών που πολλές φορές ενισχύει η μόρφωση. Στον πίνακα 6 βλέπουμε τα αποτελέσματα. Όπως βλέπουμε τα στατιστικά είναι αρκετά μακριά από τις αρχικές μας προσδοκίες. Από το δημοτικό στην μετάβαση στο λύκειο έχουμε μείωση της φερεγγυότητας κατά 6.4%, η οποία αυξάνεται ξανά κατά 1.5% στο πανεπιστήμιο. Στη γλώσσα των αριθμών αυτό μεταφράζεται ως εξής. Το 80.8% των απόφοιτων δημοτικού είναι φερέγγυοι στις πληρωμές τους, ενώ το 19.2% προβλέπεται να αθετήσουν τις πληρωμές τους. Στους απόφοιτους λυκείου το 74.8% προβλέπεται να μην αθετήσουν τις πληρωμές τους, ενώ το 25.2% ναι. Τέλος από τους απόφοιτους πανεπιστημίου το 76.3% χαρακτηρίζονται ως φερέγγυοι, ενώ το 23.7% ως μη φερέγγυοι.

| | | | Αθέτηση Πληρωμών | | |
|---|---|---|---|---|---|
| | | | Όχι | Ναι | Σύνολο |
| Εκπαίδευση | Δημοτικό | % επί της εκπαιδεύσεως | 80.8% | 19.2% | 35.3% |
| | | % επί της αθέτησης πληρωμών | 36.6% | 30.7% | |
| | Πανεπιστήμιο | % επί της εκπαιδεύσεως | 76.3% | 23.7% | 46.8% |
| | | % επί της αθέτησης πληρωμών | 45.8% | 50.2% | |
| | Λύκειο | % επί της εκπαιδεύσεως | 74.8% | 25.2% | 16.4% |
| | | % επί της αθέτησης πληρωμών | 15.8% | 18.6% | |
| Σύνολο | | % επί της εκπαιδεύσεως | 77.9% | 22.1% | |

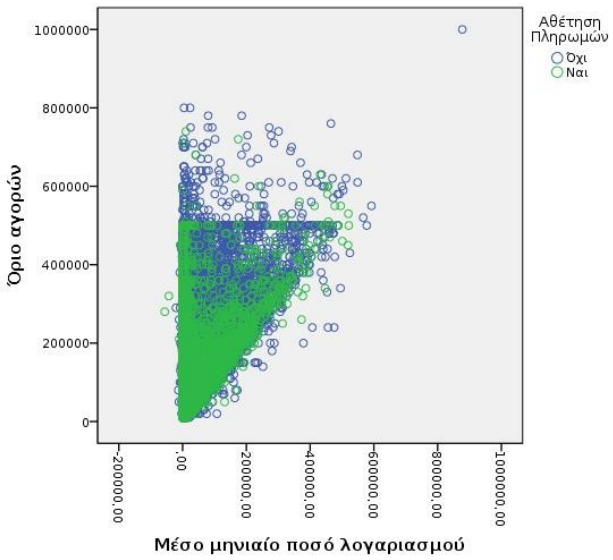
**Πίνακας 6: Σχέση εκπαίδευσης-αθέτησης πληρωμών**

**<u>Τεχνικός Σχολιασμός</u>**

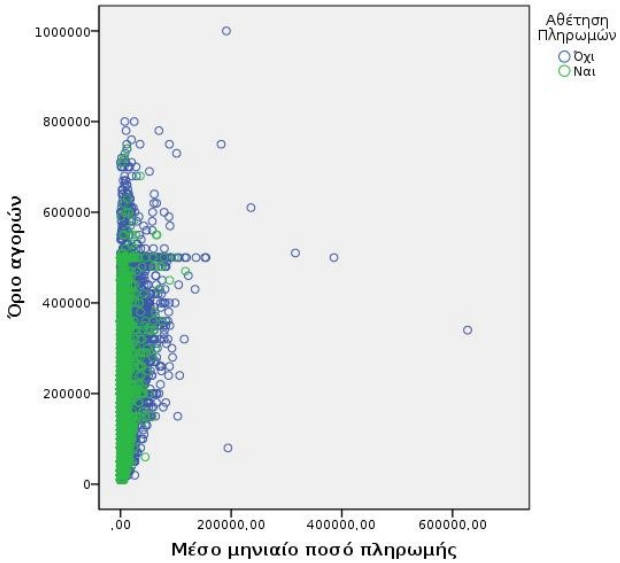
Η συσχετίσεις που παρουσιάστηκαν μέχρι στιγμής έγιναν με χρήση της τεχνικής Crosstabs του SPSS. Η συγκεκριμένη τεχνική έχει σαν μέτρο σημαντικότητας(δηλαδή εάν υπάρχει πραγματική σχέση ανάμεσα σε 2 χαρακτηριστικά) το $\chi^2$. Όλες οι συσχετίσεις είχαν sig = 0.000 < 0.05, οπότε υπήρχε ισχυρή σχέση ανάμεσα στις 2 μεταβλητές.

## 3.2. Ποσοτικές Συσχετίσεις

Στον τομέα των ποσοτικών συσχετίσεων περιλαμβάνονται οι συσχετίσεις μεταβλητών με συνεχείς τιμές. Στο σύνολο έγινε έλεγχος αρκετών συσχετίσεων. Συγκεκριμένα εξετάστηκαν οι συσχετίσεις ανάμεσα στα εξής χαρακτηριστικά: ηλικία, όριο αγορών, μέσο μηνιαίο ποσό πληρωμής και μέσο μηνιαίο ποσό λογαριασμού. Από τις συγκεκριμένες συσχετίσεις παρατηρήθηκε ότι υπάρχει μέτρια συσχέτιση στις εξής σχέσεις: όριο αγορών-μέσο μηνιαίο ποσό πληρωμής, όριο αγορών-μέσο μηνιαίο ποσό λογαριασμού και μέσο μηνιαίο ποσό λογαριασμού-μέσο μηνιαίο ποσό πληρωμής. Το μέσο ποσό πληρωμής αφορά κατά τον μέσο όρο χρημάτων που δίνουν οι πελάτες το μήνα, ενώ λέγοντας μέσο ποσό λογαριασμού εννοείται το μέσο ποσό που οφείλουν οι πελάτες στο εξάμηνο που εξετάζουμε. Και οι δύο αυτές μεταβλητές κατασκευάστηκαν για την καλύτερη κατανόηση του συνόλου δεδομένων. Στα σχήματα που ακολουθούν(σχήμα 12, 13 και 14) παρατηρούμε πως συμπεριφέρονται οι πελάτες ως προς τις πληρωμές τους με βάση αυτές τις σχέσεις.

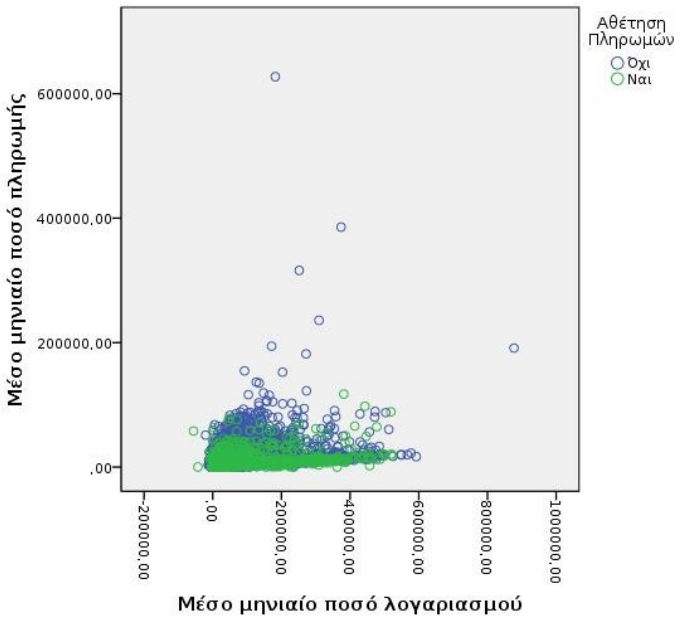



**Σχήμα 12: Σχέση ορίου αγορών-μέσο ποσό λογαριασμού**




**Σχήμα 13: Σχέση ορίου αγορών-μέσο ποσό πληρωμής**

Στο σχήμα 12 συμπεραίνουμε ότι οι πελάτες που προβλέπεται να αθετήσουν τις πληρωμές τους τείνουν να έχουν πιο χαμηλό όριο αγορών και μέσο μηνιαίο ποσό λογαριασμού. Ενώ οι φερέγγυοι πελάτες παρουσιάζουν πιο ομαλή συμπεριφορά με αρκετούς να έχουν και μεγαλύτερο όριο αγορών και υψηλότερο μέσο μηνιαίο ποσό λογαριασμού.
Στο σχήμα 13 παρατηρούμε το ίδιο μοτίβο όπου οι μη-φερέγγυοι πελάτες συγκεντρώνονται στην περιοχή με μικρότερο μέσο όριο αγορών και μέσο μηνιαίο ποσό πληρωμής. Στον πληθυσμό των φερέγγυων παρατηρείται ότι υπάρχει μία μεγαλύτερη εξάπλωση με μεγαλύτερα όρια αγορών και μέσα μηνιαία ποσά πληρωμής.




**Σχήμα 14: Σχέση μέσου μηνιαίου ποσού λογαριασμού-πληρωμής**

Με βάση το σχήμα 14, καταλήγουμε ότι και εδώ οι μη-φερέγγυοι ακολουθούν ένα σύνηθες πρότυπο. Ενώ οι φερέγγυοι συμπεριφέρονται πιο ομοιόμορφα. Δηλαδή όπως και στα προηγούμενα 2 σχήματα οι φερέγγυοι είναι πιο οικονομικά δυνατοί από τους αντίστοιχα μη-φερέγγυους.
Ο τελευταίος έλεγχος που πραγματοποιήθηκε για την ανακάλυψη συσχετίσεων είναι ανάμεσα στις μεταβλητές μηνιαίας πληρωμής και στις μεταβλητές μηνιαίου λογαριασμού. Σχετικά με τις μηνιαίες μεταβλητές λογαριασμού παρατηρήθηκαν ισχυρές θετικές συσχετίσεις ανάμεσά τους. Δηλαδή όσο ανεβαίνει το ποσό του ενός μήνα ανεβαίνει και το ποσό του άλλου μήνα. Κάτι τέτοιο είναι απολύτως λογικό, διότι αν η οφειλή ενός μήνα δεν εξοφληθεί προστίθεται στο ποσό του επόμενου. Ο επόμενος μήνας όμως έχει ήδη ένα ποσό, καθώς ο χρήστης λογικά θα χρησιμοποιήσει ξανά την κάρτα του. Άρα το μέγεθος και του επομένου μήνα στις περισσότερες περιπτώσεις είναι μεγαλύτερο από του προηγούμενου. Αξίζει να σημειωθεί ότι οι συσχέτιση της οφειλής ενός μήνα με των προηγούμενων φθίνει όσο γίνεται σύγκριση με προηγούμενους μήνες, παραμένοντας όμως σε υψηλά επίπεδα. Σχετικά με τις συσχετίσεις ανάμεσα στις μεταβλητές μηνιαίου ποσού λογαριασμού και μηνιαίου ποσού πληρωμής παρατηρούνται μικρές συσχετίσεις. Δηλαδή δεν υπάρχει κάποια μορφής έγκυρου μοτίβου που να περιγράφει το πως εξελίσσεται η μία ως προς την άλλη. Τέλος ανάμεσα στις μεταβλητές μηνιαίου ποσού πάλι συναντήσαμε μικρές -θετικές πάντα- συσχετίσεις. Δηλαδή ο τρόπος που θα συμπεριφερθεί η μία δεν επηρεάζει την άλλη.

**<u>Τεχνικός Σχολιασμός</u>**

Οι συσχετίσεις των ποσοτικών μεταβλητών έγιναν με χρήση της τεχνικής Bivariate του SPSS. Πιο συγκεκριμένα επειδή είχαμε scale μεταβλητές χρησιμοποιήθηκε ο συντελεστής συσχέτισης του Pearson. Όσες συσχετίσεις παρουσίαζαν συντελεστή κατώτερες του 0.3 θεωρήθηκαν αδύναμες, οι συσχετίσεις με συντελεστή ανώτερο του 0.3 θεωρήθηκαν μέτριες και τέλος οι συσχετίσεις ανώτερες του 0.5 ισχυρές. Όλες ήταν θετικές δηλαδή υπάρχει αναλογία στις αυξομειώσεις τους.
Ενδεικτικό παράδειγμα ο πίνακας συσχέτισης(πίνακας 7) ανάμεσα στις μεταβλητές όριο αγορών, ηλικία, μέσο ποσό οφειλής και μέσο ποσό αποπληρωμής

**Correlations**

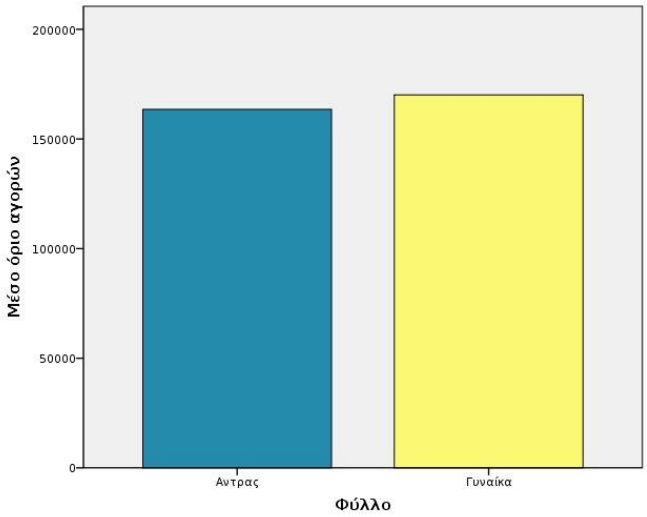
| | | Όριο αγορών | Ηλικία | Μέσο ποσό πληρωμής | Μέσο ποσό λογαριασμού |
|---|---|---|---|---|---|
| Όριο αγορών | Pearson Correlation | 1 | ,145** | ,353** | ,302** |
| | Sig. (2-tailed) | | ,000 | ,000 | ,000 |
| | N | 30000 | 30000 | 30000 | 30000 |
| Ηλικία | Pearson Correlation | ,145** | 1 | ,041** | ,055** |
| | Sig. (2-tailed) | ,000 | | ,000 | ,000 |
| | N | 30000 | 30000 | 30000 | 30000 |
| Μέσο ποσό πληρωμής | Pearson Correlation | ,353** | ,041** | 1 | ,344** |
| | Sig. (2-tailed) | ,000 | ,000 | | ,000 |
| | N | 30000 | 30000 | 30000 | 30000 |
| Μέσο ποσό λογαριασμού | Pearson Correlation | ,302** | ,055** | ,344** | 1 |
| | Sig. (2-tailed) | ,000 | ,000 | ,000 | |
| | N | 30000 | 30000 | 30000 | 30000 |

**. Correlation is significant at the 0.01 level (2-tailed).

**Πίνακας 7: Πίνακας συσχέτισης με Pearson's μετρική**

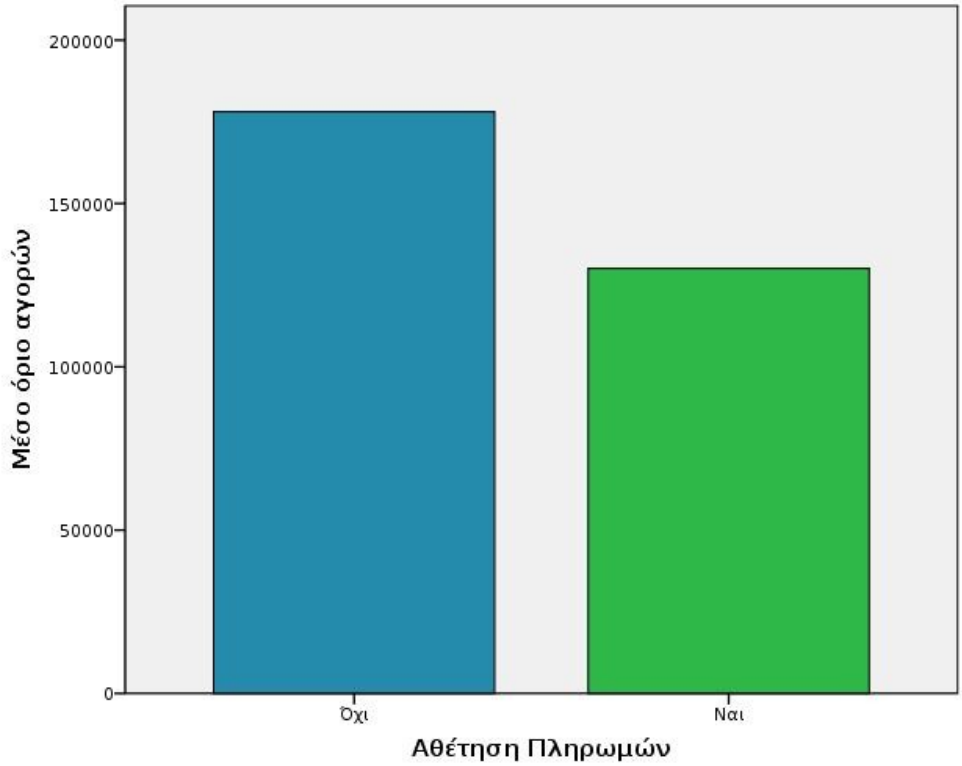
## 3.3. Σχέσεις Μέσων Όρων

Αντικείμενο του κεφαλαίου αυτού είναι η εξέταση της σημαντικότητας των διαφορών των μέσων όρων ορισμένων πληθυσμών που χωρίζουν μερικές μεταβλητές το σύνολων των πελατών.
Η πρώτη ανάλυση που έλαβε χώρα αφορούσε εάν υπάρχει σημαντική διαφορά στο μέσο όριο αγορών μίας γυναίκας από αυτό ενός άντρα. Το συμπέρασμα(σχήμα 15) που εξάχθηκε είναι ότι υπάρχει σημαντική διαφορά στους μέσους όρους, με τις γυναίκες να έχουν υψηλότερο μέσο όρο ορίου αγορών έναντι των αντρών. Οι άντρες έχουν μέσο όριο αγορών 5069 δολάρια, ενώ οι γυναίκες 5272.6 δολάρια. Δηλαδή οι γυναίκες είναι ελάχιστα αλλά σημαντικά πιο οικονομικά δυνατές από τους άντρες.




**Σχήμα 15: Σχέση μέσου ορίου αγορών-φύλλου**

Από τις προηγούμενες μετρήσεις μας παρατηρήσαμε ότι υπάρχει μία υπόνοια ότι οι οικονομικά ασθενέστεροι είναι και πιο πιθανό να αθετήσουν τις πληρωμές τους. Έτσι θέλαμε να δούμε εάν υπάρχει σχέση του μέσου ορίου αγορών με το εάν επρόκειτο να αθετήσουν ή όχι τις πληρωμές τους. Παρατηρήθηκε ότι υπάρχει σημαντική διαφορά(Σχήμα 16). Οι φερέγγυοι πελάτες έχουν υψηλότερο όριο αγορών, από τους μη-φερέγγυους. Σε τεχνικούς όρους αυτό μεταφράζεται ως 5521 δολάρια για τους φερέγγυους και 4033 δολάρια για τους μη-φερέγγυους. Αυτό πρακτικά σημαίνει ότι είναι πιθανό οι πιο αδύναμοι οικονομικά να είναι πιο δύσκολο να είναι ακριβείς στις πληρωμές τους.




**Σχήμα 16: Σχέση μέσου ορίου αγορών-αθέτηση πληρωμών**

**<u>Τεχνικός Σχολιασμός</u>**

Η τεχνική που ακολουθήθηκε στην προηγούμενη σύγκριση είναι η Independent-Samples T Test. Η συγκεκριμένη τεχνική λαμβάνει 2 μεταβλητές. Μία συνεχή(scale) και μία κατηγορική. Η κατηγορική οφείλει να χωρίζει το σύνολο δεδομένων σε 2 υπό-σύνολα. Η τεχνική αυτή υπολογίζει τους μέσους όρους με στόχο να φανεί εάν υπάρχει σημαντική διαφορά ανάμεσά στους μέσους όρους των δύο πληθυσμών. Το διάστημα εμπιστοσύνης είναι 95% και στις δύο περιπτώσεις. Αναλυτικά το πως εξάγεται το συμπέρασμα για τη σημαντικότητα:

1. Ελέγχεται εάν το sig για τις διασπορές είναι ίσες. Και στις δύο περιπτώσεις sig=0.000<0.05. Άρα δεν είναι ίσες.
2. Λαμβάνοντας υπόψη το βήμα 1 επόμενο στάδιο είναι ο έλεγχος του sig για διαφορετικές διασπορές. Και στις δύο περιπτώσεις sig=0.000<0.05. Συνεπώς υπάρχει σημαντική διαφορά ανάμεσα στους μέσους όρους των 2 πληθυσμών.

Πολλοί άνθρωποι εξαιτίας της φύσης της εργασίας του κυρίως, ενδέχεται να μεταβάλλουν τον τρόπο που συμπεριφέρονται σαν πελάτες ως προς τις οφειλές τους. Στα πλαίσια αυτά θέλαμε να δούμε εάν υπάρχει σημαντική διαφορά ανάμεσα στο μέσο όρο της οφειλής τον Απρίλιο και το Σεπτέμβριο. Τα αποτελέσματα που καταλήξαμε είναι ότι τον Απρίλιο η μέση οφειλή είναι 1205 δολάρια, ενώ τον Σεπτέμβριο είναι 1587 δολάρια. Συνεπώς υπάρχει σημαντική διαφορά.
Σχετικά με τα ποσά εξοφλήσεως παρατηρείται ότι υπάρχει και εδώ σημαντική διαφορά με το μέσο ποσό για τον Απρίλιο να διαμορφώνεται στα 161.6 δολάρια, ενώ για το Σεπτέμβριο στα 175.5 δολάρια. Το συμπέρασμα που προκύπτει είναι ότι όσο περνούν οι μήνες το μέσο χρέος ανεβαίνει όμως το μέσο ποσό που εξοφλείται μειώνεται.

**Τεχνικός Σχολιασμός**

Η τεχνική που εφαρμόστηκε είναι η Pair-Samples T Test. Η τεχνική αυτή λαμβάνει 2 μεταβλητές και υπολογίζει τους μέσους όρους τους στο σύνολο των πελατών. Ο στόχος είναι για την ανακάλυψη σημαντικής σχέσης ανάμεσα στους μέσους όρους τους. Το διάστημα εμπιστοσύνης είναι 95%. Το συμπέρασμα εξάγεται ύστερα από τα εξής βήματα:

1. Γίνεται έλεγχος του sig της συσχέτισης. Εδώ sig=0.000<0.05. Άρα υπάρχει σημαντική σχέση και δικαιολογείται η χρήση Pair-Samples T Test.

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | Ποσό λογαριασμού Σεπτεμβρίου-Ποσό λογαριασμού Απριλίου | 30000 | .803 | .000 |

**Πίνακας 8: Πίνακας σημαντικότητας συσχέτισης**

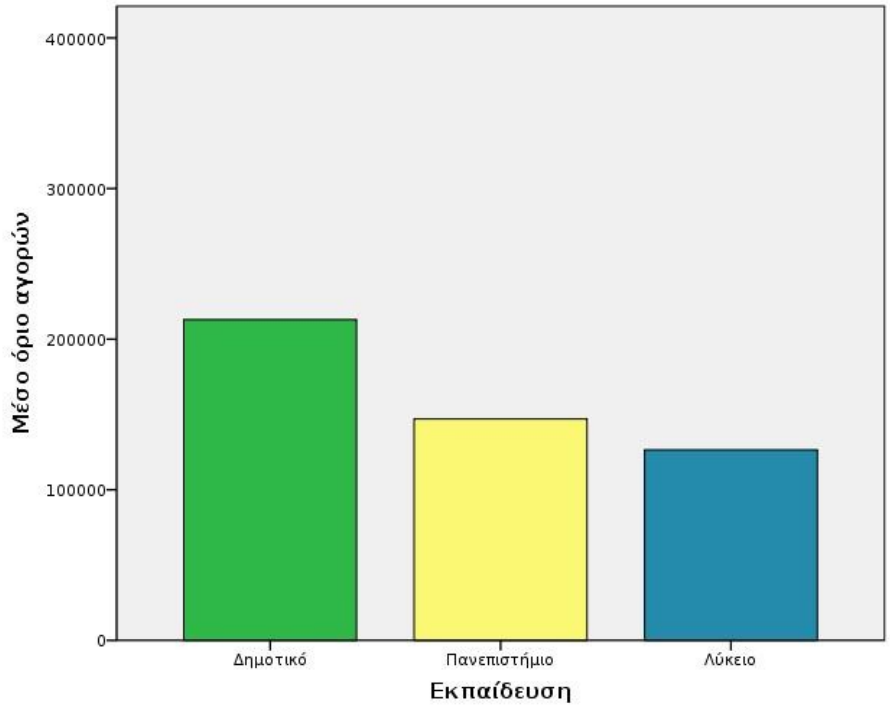
2. Ελέγχεται το sig των διαφορών, εδώ sig=0.000<0.05. Άρα οι μέσες τιμές έχουν διαφορά.

**Paired Samples Test**

| | | Paired Differences | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | | | | |
| | | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | t | df | Sig. (2-tailed) |
| Pair 1 | Ποσό λογαριασμού Απριλίου-Ποσό λογαριασμού Σεπτεμβρίου | 12351.571 | 43922.422 | 253.586 | 11854.531 | 12848.610 | 48.708 | 29999 | .000 |

**Πίνακας 9: Πίνακας αποτελεσμάτων paired t-test**

Το υψηλό μορφωτικό επίπεδο είναι αρκετές φορές προπομπός υψηλού εισοδήματος για κάποιο πελάτη. Ο λόγος είναι οι αυξημένες δυνατότητες εργασίας. Επιπρόσθετα δικαιούται βάση νόμου υψηλότερες αποδοχές. Στο πλαίσιο αυτό θεωρήσαμε πρέπον να συγκρίνουμε το μέσο όρο του ορίου αγορών ενός πελάτη με βάση την εκπαίδευση που έχει περάσει. Στο σχήμα 17 βλέπουμε το μέσο όρο των ορίων αγορών για τα διαφορετικά μορφωτικά επίπεδα.




**Σχήμα 17: Σχέση μέσου ορίου αγορών-εκπαίδευση**

Συμπεραίνουμε ότι το μέσο όριο αγορών για τους απόφοιτους δημοτικού βρίσκεται στην πρώτη θέση. Στη δεύτερη θέση ακολουθούν οι απόφοιτοι πανεπιστημίου, ενώ στην τελευταία θέση είναι οι απόφοιτοι Λυκείου. Σε αριθμούς αυτό μεταφράζεται ως 6601 δολάρια για το δημοτικό, 4558.9 δολάρια για το πανεπιστήμιο και 3923 δολάρια. Συνεπώς η υπόθεση που κάναμε δεν συμβαδίζει με τα αποτελέσματα.

**Τεχνικός Σχολιασμός**

Η συγκεκριμένη ανάλυση έγινε με χρήση Oneway-Anova. Κατά την εφαρμογή χρησιμοποιήθηκαν όλες οι τιμές της μεταβλητής εκπαίδευσης. Εμφανίζονται όμως μόνο 3(δημοτικό, λύκειο, πανεπιστήμιο). Για τις άλλες δεν γνωρίζουμε την περιγραφή και καταλαμβάνουν αμελητέο ποσοστό στο σύνολο δεδομένων. Αρχικά ελέγχουμε το sig για τις διασπορές. Εδώ έχουμε sig=0.000<0.05. Άρα έχουμε σημαντική διαφορά στις διασπορές των ομάδων. Για να δούμε ποιες διαφορές είναι σημαντικές εφαρμόζουμε post-hoc. Οι μετρικές που εφαρμόζονται είναι οι Tuckey HSD, Games-Howell και Dunnett t. Από το post-hoc test βγαίνει το συμπέρασμα ότι υπάρχει σημαντική διαφορά ανάμεσα στους 3 μέσους όρους. Ο λόγος είναι διότι sig=0.000<0.05, σε όλες τις περιπτώσεις.

## 4. Εξαγωγή Παραγόντων

Αντικείμενο του κεφαλαίου αυτού είναι η παρουσίαση της ανάλυσης που έγινε για την εξαγωγή κάποιων μη-μετρήσιμων μεταβλητών από το σύνολο δεδομένων. Πολλές φορές σε αρκετά σύνολα δεδομένων υπάρχουν ορισμένοι άξονες γύρω από τους οποίους κινούνται κάποιες μεταβλητές του

δείγματός μας. Οι άξονες αυτοί συνήθως είναι κάποια μη-μετρήσιμα χαρακτηριστικά, και μπορούν να συνοψίσουν το σύνολο δεδομένων μέσω αυτών. Για τον υπολογισμό τους χρησιμοποιήθηκε η Factor Analysis και ύστερα από πολυάριθμες εφαρμογές της τεχνικής προέκυψαν 2 νέες μεταβλητές.

## 1η Ανάλυση

Κατά την πρώτη ανάλυση χρησιμοποιήθηκαν όλες οι συνεχείς μεταβλητές(scale). Δηλαδή οι μεταβλητές:

- Ηλικία
- Όριο αγορών
- Αριθμός μηνιαίας καθυστέρησης Απριλίου-Σεπτεμβρίου
- Ποσό οφειλής Απριλίου-Σεπτεμβρίου
- Ποσό πληρωμής Απριλίου-Σεπτεμβρίου

Δεν χρησιμοποιήθηκαν οι μεταβλητές avgBill και avgPay, οι οποίες κατασκευάστηκαν και μετρούν το μέσο ποσό χρέους για τους 6 μήνες και το μέσο ποσό πληρωμής για τους 6 μήνες αντίστοιχα. Κατά την ανάλυση παρατηρήθηκαν ισχυρές θετικές συσχετίσεις ανάμεσα σε ορισμένες μεταβλητές. Επίσης αυτές οι μεταβλητές είχαν και χαμηλή συμμετοχικότητα, δηλαδή εξηγούνταν μικρό ποσοστό των διασπορών των μεταβλητών μέσω των παραγόντων. Έτσι θεωρήθηκε πρέπον να αφαιρεθούν και να εφαρμοστεί νέα ανάλυση. Αξίζει να σημειωθεί ότι μέσω αυτής της ανάλυσης προέκυπταν 4 διαστάσεις(μεταβλητές).

## 2η Ανάλυση

Η δεύτερη ανάλυση έγινε με χρήση των υπολοίπων μεταβλητών που απέμειναν από την πρώτη ανάλυση και είναι οι εξής:

- Όριο αγορών
- Ηλικία
- Αριθμός μηνιαίας καθυστέρησης Απριλίου-Σεπτεμβρίου
- Ποσό οφειλής Απριλίου-Σεπτεμβρίου
- Ποσό πληρωμής Απριλίου

Κατά τη συγκεκριμένη ανάλυση παρατηρήθηκαν υψηλές θετικές συσχετίσεις, συνεπώς αφαιρέθηκαν και αυτές οι μεταβλητές. Οπότε ακολούθησε νέα ανάλυση.

## 3η Ανάλυση

Κατά την τρίτη ανάλυση χρησιμοποιήθηκαν πάλι το σύνολο των υπολοίπων μεταβλητών. Αυτές είναι:

- Όριο αγορών
- Ηλικία
- Αριθμός μηνιαίας καθυστέρησης Απριλίου-Σεπτεμβρίου
- Ποσό πληρωμής Απριλίου

Σε αυτή την απόπειρα δεν υπήρξε κάποια άλλη παραβίαση. Επομένως ήτανε και η τελευταία ανάλυση που εφαρμόστηκε. Οι τομείς που προέκυψαν ήταν 2. Ο πρώτος συμπεριελάμβανε τις μεταβλητές ηλικία, ποσό αποπληρωμής Απριλίου και όριο αγορών. Ονομάστηκε οικονομική δύναμη του πελάτη. Δηλαδή κατά πόσο είναι ικανός να ανταπεξέλθει στις υποχρεώσεις του έγκαιρα. Ο δεύτερος τομέας περιελάμβανε όλα τα χαρακτηριστικά για τις καθυστερήσεις πληρωμών του κάθε μήνα και ονομάστηκε τιμιότητα πελάτη. Δηλαδή κατά πόσο ο πελάτης θα ήτανε πιστός στις πληρωμές του.
Από την ανάλυση παρατηρήθηκε ότι το δείγμα είναι αρκετό για την εξαγωγή συμπερασμάτων. Επιπρόσθετα υπάρχουν ισχυρές συσχετίσεις ανάμεσα στις μεταβλητές και γενικά η συσχέτιση που υπάρχει είναι διάχυτη στο σύνολο δεδομένων. Το πιο σημαντικό συμπέρασμα είναι ότι υπάρχουν σημαντικές συσχετίσεις ανάμεσα στα χαρακτηριστικά όλου του συνόλου δεδομένων.
Στα αρνητικά από την χρήση μόνο των διαστάσεων για την εξήγηση της πληροφορίας της κάθε μεταβλητής παρατηρείται ότι έχουμε μείωση της ακρίβειας. Δηλαδή χάνεται αρκετή πληροφορία. Ενδεικτικό είναι ότι για τη μεταβλητή της ηλικίας εξηγείται μόνο το 21% της συνολικής διασποράς του χαρακτηριστικού.
Στα θετικά είναι ότι οι 2 παράγοντες εξηγούν συνολικά το 62.5% της συνολικής διασποράς του συνολικού πληθυσμού. Ο πιο σημαντικός παράγοντας είναι αυτός της οικονομικής δύναμης που εξηγεί το 49.2%. Το συμπέρασμα που προκύπτει είναι ότι η συμπεριφορά ενός πελάτη ως προς την αμεσότητα των πληρωμών του εξαρτάται κατά το ήμισυ από την αρχική εντύπωση που δίνει. Λέγοντας εντύπωση εννοούμε από το μέγεθος της οικονομικής ικανότητας και τη συμπεριφορά του στον πρώτο μήνα του εξαμήνου. Συνοπτικά η αρχή είναι η ήμισυ του παντός.
Ο δεύτερος παράγοντας συγκεντρώνει το 14.2% και εκφράζει την πιστότητα ή αλλιώς την συνέπεια στις πληρωμές ενός πελάτη. Δηλαδή την συμπεριφορά του κατά τη διάρκεια του εξαμήνου.

**Τεχνικός Σχολιασμός**

Όπως προαναφέραμε η τεχνική που εφαρμόστηκε για την ανάλυση ήταν η Factor Analysis. Η τεχνική για μείωση των διαστάσεων είναι η Principal Component Analysis. Ο πρώτος έλεγχος που γινόταν ήταν στο πίνακα Anti-Image Matrices ποιες μεταβλητές εμφάνιζαν στη διαγώνιο τιμή <0.5. Στην πρώτη ανάλυση οι μεταβλητές που αφορούν το ποσό πληρωμής για τους μήνες Μάιο-Σεπτέμβριο αφαιρέθηκαν, σύμφωνα με αυτό το περιορισμό. Οι μεταβλητές αυτές παρουσίασαν επίσης χαμηλή τιμή στον πίνακα συμμετοχικότητας(communalities), συνεπώς δεν υπήρξε πιθανή χαμένη πληροφορία.
Ένας άλλος έλεγχος είναι ο πίνακας συσχετίσεων(correlation matrix). Εδώ αφαιρέθηκαν όσες μεταβλητές εμφάνιζαν τιμές >0.9. Ο λόγος είναι διότι η τιμή determinant εμφάνιζε τιμές <0.00001. Αυτό το πρόβλημα εμφανίστηκε στη δεύτερη ανάλυση όπου αφαιρέθηκαν οι μεταβλητές για τα ποσά οφειλής το κάθε μήνα από το εξάμηνο.
Μία σωστή ανάλυση επίσης προϋποθέτει ο πίνακας με το δείκτη KMO να έχει τιμές >0.5. Η ανάλυση που καταλήξαμε εμφανίζει τιμή 0.86. Επιπρόσθετα sig = 0.000 < 0.05, το οποίο σημαίνει ότι υπάρχουν σημαντικές συσχετίσεις.

| **KMO and Bartlett's Test** | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | | .860 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 148513.709 |
| | df | 36 |
| | Sig. | .000 |

**Πίνακας 10: Πίνακας καταλληλότητας τελευταίας ανάλυσης**

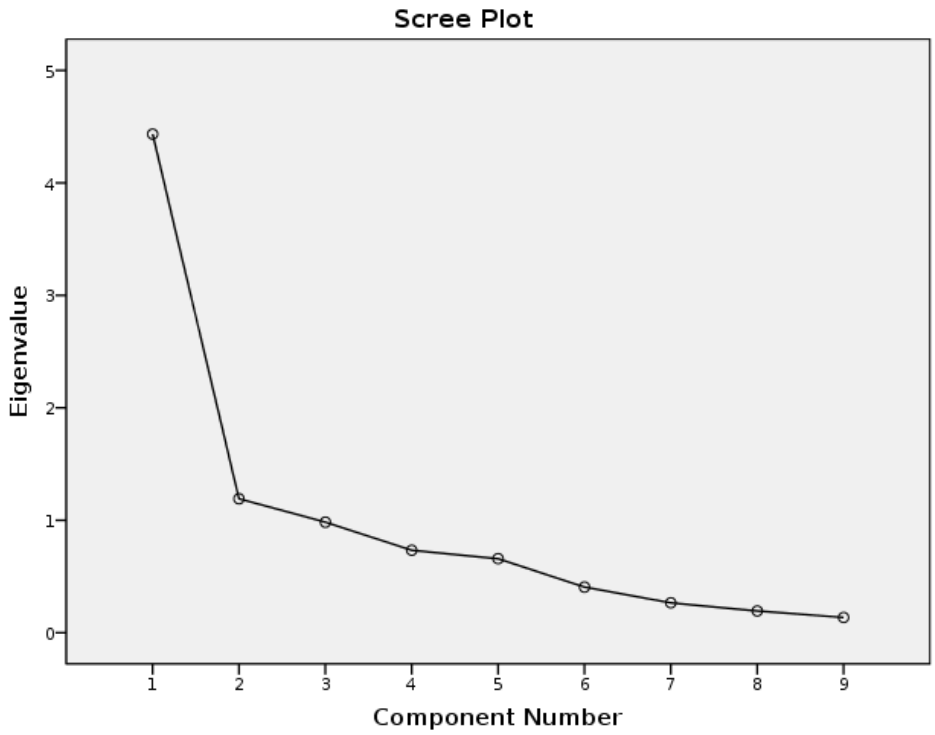
Τελευταίος έλεγχος που γίνεται είναι ο πίνακας με τα φορτία. Στόχος είναι η κάθε μεταβλητή να φορτώνεται σε ένα μόνο παράγοντα, το οποίο επιτεύχθηκε.

**Rotated Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| PAY_4 | .896 | |
| PAY_5 | .886 | |
| PAY_3 | .873 | |
| PAY_2 | .836 | |
| PAY_6 | .836 | |
| PAY_0 | .714 | |
| PAY_AMT6 | | .732 |
| LIMIT_BAL | | .698 |
| AGE | | .458 |

**Πίνακας 11: Πίνακας κατανομής μεταβλητών στους παράγοντες**

Ο αριθμός των κρυμμένων διαστάσεων του συνόλου δεδομένων μπορεί να επικυρωθεί και μέσω του σχήματος 18 όπου μετά τον αριθμό 2 βλέπουμε η καμπύλη να αποκτά μία πιο επίπεδη καμπυλότητα.




**Πίνακας 18: Scree plot factor analysis**

## 5. Ανάλυση Διακρίσεων

Το σύνολο δεδομένων μας όπως γνωρίζουμε έχει μία μεταβλητή η οποία περιέχει μία πρόβλεψη σχετικά με το αν ο πελάτης επρόκειτο να αθετήσει ή όχι τις πληρωμές του. Είναι προϊόν πρόβλεψης ενός νευρωνικού δικτύου. Στο κεφάλαιο αυτό στόχος είναι η παρουσίαση των αποτελεσμάτων μίας τεχνικής που έγινε με σκοπό να κατηγοριοποιηθούν οι πελάτες στις 2 κατηγορίες, μέσω των πληροφοριών του συνόλου δεδομένων.
Η διάκριση έγινε με χρήση της ανάλυσης διακρίσεων. Χρησιμοποιήθηκαν οι συνεχείς μεταβλητές του δείγματος, και το χαρακτηριστικό του φύλλου με κατάλληλη τροποποίηση. Τα αποτελέσματα που εξάγαμε είναι ενθαρρυντικά και το ποσοστό κατανομής στις 2 ομάδες είναι ικανοποιητικό(πίνακας 12). Συγκεκριμένα το 72.5% των περιπτώσεων κατατάχθηκε κανονικά.

**Αποτελέσματα ταξινόμησης**

| | | Αθέτηση Πληρωμών | Προβλεπόμενη κατηγορία Όχι | Ναι | Σύνολο |
|---|---|---|---|---|---|
| Πραγματική κατηγορία | Πλήθος | Όχι | 17767 | 5597 | 23364 |
| | | Ναι | 2637 | 3999 | 6636 |
| | % | Όχι | 76.0 | 24.0 | 100 |
| | | Ναι | 39.7 | 60.3 | 100 |

**Πίνακας 12: Πίνακας κατανομής των περιπτώσεων**

Σύμφωνα με τον πίνακα το 76% κατατάχθηκε σωστά ως φερέγγυοι πελάτες. Το 24% κατατάχθηκε λανθασμένα ως μη-φερέγγυος, ενώ στην πραγματικότητα είναι φερέγγυος. Στον τομέα των φερέγγυων το 39.7% ενώ επρόκειτο να αθετήσουν τις πληρωμές τους κρίθηκαν φερέγγυοι. Τέλος το 60.3% κατατάχθηκαν σωστά ως μη-φερέγγυοι.
Από τις μεταβλητές του δείγματος αυτή που έχει την μεγαλύτερη διαχωριστική ικανότητα είναι το χαρακτηριστικό που δείχνει την καθυστέρηση πληρωμής για το μήνα Σεπτέμβριο. Δηλαδή με βάση την τελευταία τους συμπεριφορά μπορούμε να κρίνουμε καλύτερα ποιος επρόκειτο να αθετήσεις ή όχι τις πληρωμές του.

**<u>Τεχνικός Σχολιασμός</u>**

Η τεχνική που ακολουθήθηκε είναι η discriminal analysis, του SPSS. Χρησιμοποιήθηκαν όλες οι συνεχείς μεταβλητές(scale), ενώ από τις κατηγορικές μόνο το φύλλο. Η χρησιμοποίησή του έγινε μέσω μετατροπής του σε dummy μεταβλητή.
Επίσης έγινε χρήση και η τεχνική cross-validation. Η συγκεκριμένη τεχνική εξειδικεύεται για να δούμε πως γενικεύεται ένα μοντέλο. Δηλαδή πως συμπεριφέρεται για ένα νέο σύνολο δεδομένων. Κάθε φορά που φτιάχνεται το μοντέλο με ένα υπό-σύνολο δεδομένων, κρατιέται ένα μικρό άλλο υπό-σύνολο δεδομένων το οποίο δοκιμάζεται σε αυτό το μοντέλο. Τα αποτελέσματα σχετικά με την ακρίβεια των ταξινομήσεων δεν διέφεραν σχεδόν καθόλου. Η σωστή ταξινόμηση μέσω cross-validation άγγιξε το 72.5%(πίνακας 13). Επομένως η διαφορά είναι ελάχιστη(0.1%), αλλά στον τραπεζικό τομέα όχι αμελητέα.

**Classification Results**

| | | Αθέτηση Πληρωμών | Predicted Group Membership Όχι | Ναι | Total |
|---|---|---|---|---|---|
| Original | Count | Όχι | 17767 | 5597 | 23364 |
| | | Ναι | 2637 | 3999 | 6636 |
| | % | Όχι | 76.0 | 24,0 | 100 |
| | | Ναι | 39.7 | 60.3 | 100 |
| Cross-validated | Count | Όχι | 17761 | 5603 | 23364 |
| | | Ναι | 2643 | 3993 | 6636 |
| | % | Όχι | 76.0 | 24.0 | 100 |
| | | Ναι | 39.8 | 60.2 | 100 |

a. 72,6% of original grouped cases correctly classified.

b. 72,5% of cross-validated grouped cases correctly classified.

**Πίνακας 13: Πίνακας με συνολικά αποτελέσματα ταξινόμησης**

Σχετικά με το ποιες μεταβλητές είναι πιο σημαντικές για το διαχωρισμό συμβουλευτήκαμε τον πίνακα 14 και την τιμή με τη μεγαλύτερη απόλυτη τιμή.

**Standardized Canonical Discriminant Function Coefficients**

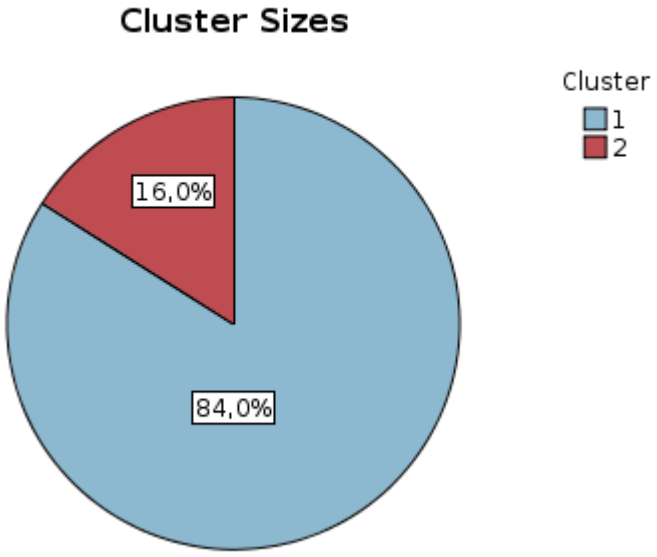
| | Function 1 |
|---|---|
| LIMIT_BAL | -.057 |
| AGE | .114 |
| PAY_0 | .747 |
| PAY_2 | .165 |
| PAY_3 | .099 |
| PAY_4 | .026 |
| PAY_5 | .045 |
| PAY_6 | .008 |
| BILL_AMT1 | -.347 |
| BILL_AMT2 | .081 |
| BILL_AMT3 | .014 |
| BILL_AMT4 | -.031 |
| BILL_AMT5 | -.003 |
| BILL_AMT6 | .050 |
| PAY_AMT1 | -.090 |

| | |
|---|---|
| PAY_AMT2 | -,036 |
| PAY_AMT3 | -.006 |
| PAY_AMT4 | -.030 |
| PAY_AMT5 | -.039 |
| PAY_AMT6 | -.014 |
| male | .050 |

**Πίνακας 14: Πίνακας σημαντικότητας μεταβλητών**

## 6. Ομαδοποίηση

Η ανάλυση διακρίσεων έγινε με σκοπό να δούμε πως μέσω της πληροφορίας του συνόλου δεδομένων κατατάσσεται ένας πελάτης σε φερέγγυος ή μη-φερέγγυος. Το αποτέλεσμα ήταν ικανοποιητικό. Σε αυτό το κεφάλαιο στόχος είναι να δούμε πως μέσω των πληροφοριών του συνόλου δεδομένων δημιουργούνται τυχόν ομάδες πελατών. Δηλαδή εάν υπάρχει κάποια περαιτέρω ομαδοποίηση με βάση τη συμπεριφορά τους.
Το αποτέλεσμα που έτρεξε δεν είναι αρκετά ενθαρρυντικό. Η ποιότητα των ομαδοποιήσεων είναι χαμηλού μεγέθους. Ωστόσο αποτελεί άξιο σχολιασμού. Το πρώτο συμπέρασμα που εξάγουμε είναι ότι πάλι έχουμε τον διαχωρισμό σε δύο ομάδες(σχήμα 19). Το σχήμα και τα ποσοστά μας προϊδεάζουν ότι οι ομάδες αφορούν την αθέτηση ή όχι των πληρωμών τους.




**Σχήμα 19: Σύνολο ομάδων και ποσοστό πελατών που ανήκει στην κάθε μία**

Σχετικά με τις μεταβλητές που επιδρούν περισσότερο στην δημιουργία της κάθε ομάδας, διακρίνουμε 10 μεταβλητές. Διακρίνονται σε 2 κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τις μεταβλητές με το ποσό εξόφλησης τον κάθε μήνα. Δηλαδή το ποσό που δίνει ένας πελάτης έχει

μεγάλη σημασία στην πρόβλεψη της ομάδας που θα καταταχθεί. Η άλλη κατηγορία μεταβλητών με μεγάλη επιρροή στον καθορισμό της ομάδας είναι ορισμένα χαρακτηριστικά για το ποσό οφειλής τους. Πιο συγκεκριμένα οι μεταβλητές από τον Ιούνιο μέχρι και τον Σεπτέμβριο. Παρατηρούμε ότι σε αυτή την περίπτωση, δεν παίζουν σημαντικό ρόλο όλες οι μεταβλητές παρά μόνο αυτές που είναι πιο κοντά στο τέλος του εξαμήνου. Μία ερμηνεία είναι ότι όσο περνούν οι μήνες αυξάνεται το ποσό λογαριασμού. Επίσης το ποσό του μηνιαίου λογαριασμού επηρεάζεται από το χρηματικό ποσό που δόθηκε τον προηγούμενο και από την εξόφληση ή όχι του προηγούμενου μήνα.
Το συμπέρασμα που καταλήγουμε είναι ότι χρησιμοποιώντας τα ποσοστά η διάκριση ανάμεσα σε φερέγγυους και μη-φερέγγυους πελάτες δεν βρίσκεται πολύ κοντά στην πραγματικότητα. Θυμίζουμε ότι σύμφωνα με την μεταβλητή του συνόλου δεδομένων το ποσοστό για τους φερέγγυους πελάτες είναι 77.9%, ενώ για τους μη-φερέγγυους είναι 22.1%. Ενώ εδώ οι φερέγγυοι είναι στο 84% και οι μη-φερέγγυοι στο 16%.

**Τεχνικός Σχολιασμός**

Η τεχνική που εφαρμόστηκε για την ομαδοποίηση είναι η twostep cluster. Η συγκεκριμένη τεχνική είναι ιδανική για clustering. Μπορεί να επεξεργαστεί κάθε τύπου μεταβλητή. Ως προς τον αριθμό των αποστάσεων χρησιμοποιήθηκε το log-likelihood. Ο διαχωρισμός των clusters γίνεται με χρήση του BIC. Ο αριθμός των clusters αφέθηκε στην κρίση του twostep cluster. Σχετικά με την ποιότητα του clustering χαρακτηρίζεται αρκετά χαμηλή παίρνοντας τιμή poor.

## 7. Επίλογος

Στο κεφάλαιο αυτό θέλουμε να κάνουμε μία ανασκόπηση των πιο σημαντικών συμπερασμάτων μας. Χρησιμοποιήθηκαν αρκετές τεχνικές για την ανάλυση του συνόλου δεδομένων. Στόχος ήταν η βέλτιστη εξέτασή του από όλες τις σκοπιές.
Η αρχική παρατήρηση που προκύπτει είναι ότι το ποσό των φερέγγυων είναι σχεδόν τετραπλάσιο των μη-φερέγγυων. Έπειτα ότι οι γυναίκες υπερτερούν των αντρών στο δείγμα και είναι πιο φερέγγυες από τους άντρες. Σχετικά με το μέσο όριο αγορών παρατηρήθηκε ότι οι πιο φερέγγυοι είναι και δυνατότεροι οικονομικά, δηλαδή διαθέτουν υψηλότερο όριο αγορών.
Από των έλεγχο των συσχετίσεων εξάγαμε ορισμένες μέτριες συσχετίσεις ανάμεσα στο όριο αγορών, το μέσο ποσό λογαριασμού και το μέσο ποσό πληρωμής. Η συσχετίσεις είναι ανά 2 σε αυτές τις μεταβλητές και σημαίνουν ότι όσο αυξομειώνει η μία αυξομειώνει και η άλλη.
Κατά την ανάλυση κρίθηκε απαραίτητο η εξέταση για τυχόν κρυφούς παράγοντες. Βρέθηκαν 2, οι οποίοι είναι η συνέπεια ενός πελάτη ως προς τις πληρωμές του και η οικονομική δύναμη η οποία εκφράζει την αρχική εντύπωση που δίνει ο πελάτης για την εκπλήρωση των λογαριασμών του.
Το τελικό στάδιο αφορούσε την ομαδοποίηση του συνόλου δεδομένων με βάση των πληροφοριών του και όχι της εξαρτημένης μεταβλητής. Από την ανάλυση εξάγαμε το συμπέρασμα ότι γίνεται ικανοποιητική κατανομή των πελατών στην προβλεπόμενη συμπεριφορά τους. Ενώ ως προς τη δημιουργία των ομάδων βλέπουμε ότι σχηματίζονται δύο ομάδες με κριτήριο την φερεγγυότητα στις πληρωμές τους όπως είναι αρμόδια η εξαρτημένη μεταβλητή για το συγκεκριμένο λόγο.