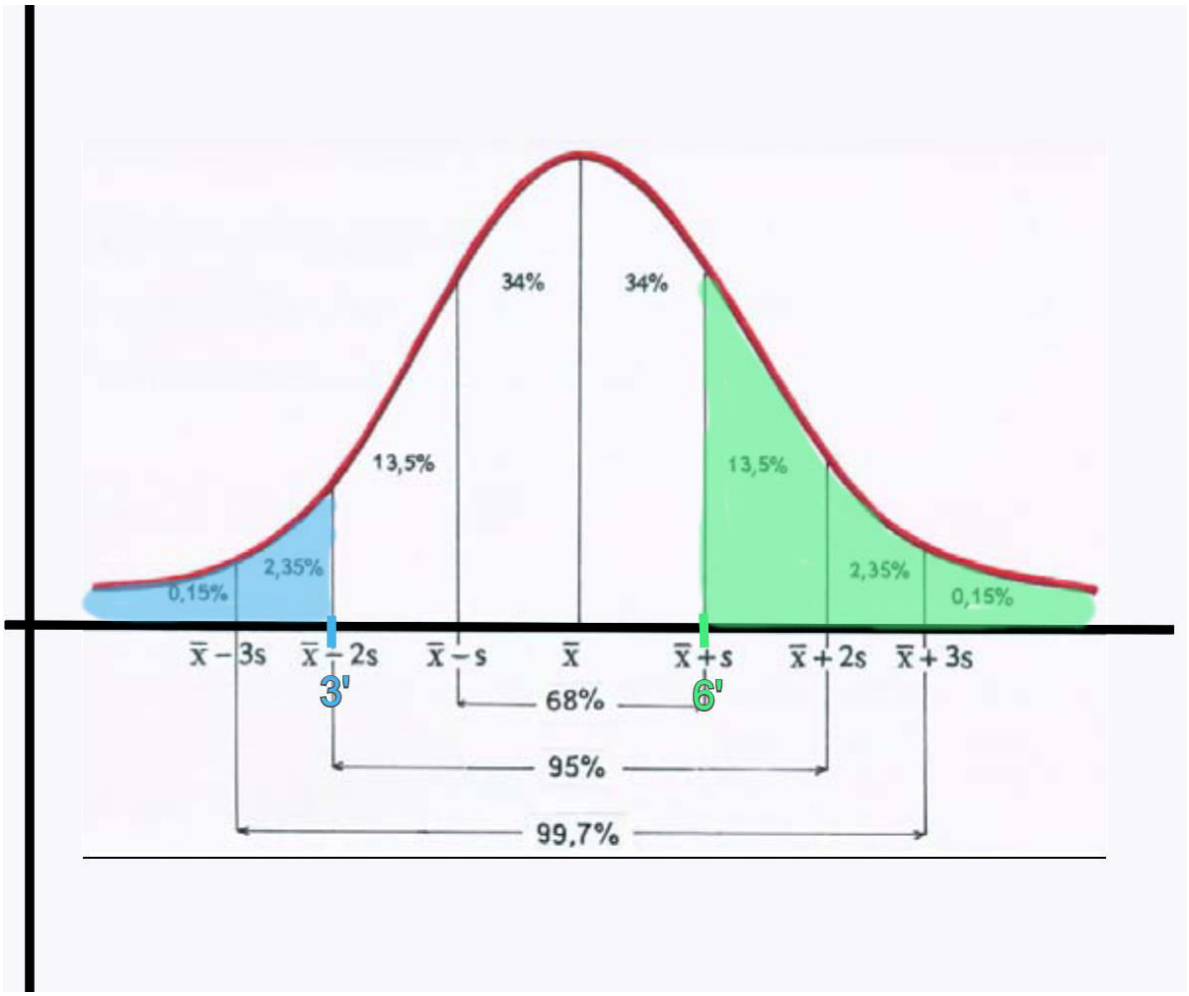
DS7006 - Quantitative Data Analysis
Απαντήσεις - Νίκος Κουγιανός (UEL Number: 2018506)

A. Laboratory Exercise

1. Μια εταιρία κινητής τηλεφωνίας καθορίζει πρόγραμμα τηλεφωνικών συνδιαλέξεων διάρκειας 3 έως 6 λεπτών με τη χαμηλότερη χρέωση. Κατά τη διάρκεια μιας ημέρας με πολλές συνδιαλέξεις βρέθηκε ότι το 2,5% των τηλεφωνημάτων ήταν μικρότερης διάρκειας από 3 λεπτά ενώ το 16% των τηλεφωνημάτων ήταν μεγαλύτερης διάρκειας των 6 λεπτών. Υποθέτουμε ότι η κατανομή των χρόνων συνδιαλέξεων είναι περίπου κανονική.

α) Να βρείτε τη μέση τιμή και την τυπική απόκλιση των χρόνων συνδιαλέξεων.



Εικόνα 1- Γράφημα κανονικής κατανομής

Στην εκφώνηση αναφέρεται ότι η κατανομή των χρόνων συνδιαλέξεων είναι περίπου κανονική. Από την εικόνα 1 (γράφημα κανονικής κατανομής) μπορούμε εύκολα, βασισμένοι και στις υπόλοιπες πληροφορίες της εκφώνησης να βρούμε τη μέση τιμή και την τυπική απόκλιση των χρόνων συνδιαλέξεων. Θα πρέπει επίσης να αναφερθούν και οι ιδιότητες που έχει ένα σύνολο παρατηρήσεων που ακολουθεί κανονική κατανομή:

1. Στο διάστημα $(\bar{x} - s, \bar{x} + s)$ ανήκει περίπου το 68% των παρατηρήσεων, οπότε σε καθένα από αυτά τα υποδιαστήματα ανήκει το 34%.
2. Στο διάστημα $(\bar{x} - 2s, \bar{x} + 2s)$ ανήκει προσεγγιστικά το 95% των παρατηρήσεων, οπότε με απλά μαθηματικά βγαίνει ότι στα διαστήματα $(\bar{x} - 2s, \bar{x} - s)$ και $(\bar{x} + s, \bar{x} + 2s)$ ανήκει το 13.5% αντίστοιχα.
3. Στο διάστημα $(\bar{x} - 3s, \bar{x} + 3s)$ ανήκει το 99.7% των παρατηρήσεων, οπότε με τις ίδιες πράξεις προκύπτει ποσοστό 2.35% σε κάθε υποδιάστημα $(\bar{x} - 3s, \bar{x} - 2s)$ και $(\bar{x} + 2s, \bar{x} + 3s)$.
4. Όπως φαίνεται και στο γράφημα το υπόλοιπο 0.3% κατανέμεται ίσα στις παρατηρήσεις που είναι είτε μικρότερες από $\bar{x} - 3s$ είτε μεγαλύτερες από $\bar{x} + 3s$.
5. Το εύρος είναι περίπου ίσο με έξι τυπικές αποκλίσεις, δηλαδή $R \approx 6s$.

Με τις παραπάνω πληροφορίες έχουμε ζωγραφίσει στο γράφημα με **γαλάζιο** χρώμα τις παρατηρήσεις που είναι μικρότερες από 3 λεπτά (2.5%) και με **πράσινο** χρώμα τις παρατηρήσεις που είναι μεγαλύτερες από 6 λεπτά (16%). Χρησιμοποιώντας και την εικόνα 1 οδηγούμαστε εύκολα σε ένα σύστημα 2 εξισώσεων με 2 αγνώστους:

- $\bar{x} - 2s = 3$
- $\bar{x} + s = 6$

Από το παραπάνω σύστημα χρησιμοποιώντας την μέθοδο της αντικατάστασης προκύπτει ότι $\bar{x} = 5$ και $s = 1$.

Συνεπώς η απάντηση στο ερώτημα α είναι πως η μέση τιμή των χρόνων συνδιαλέξεων είναι 5 λεπτά και η τυπική απόκλιση είναι 1 λεπτό.

β) Αν κατά τη διάρκεια αυτής της ημέρας έγιναν 2000 τηλεφωνήματα να βρείτε:

i) Πόσα τηλεφωνήματα είχαν διάρκεια από 3 έως 5 λεπτά.

Σύμφωνα με την ανάλυση που έχει γίνει στο ερώτημα α, και αφού έχουμε βρει την μέση τιμή και την τυπική απόκλιση του προβλήματος, μπορούμε εύκολα να βρούμε πόσα τηλεφωνήματα είχαν διάρκεια από 3 έως 5 λεπτά. Πιο συγκεκριμένα, χρησιμοποιώντας και πάλι το γράφημα ως αναφορά, θα πρέπει να πάρουμε το διάστημα $(\bar{x} - 2s, \bar{x})$, το οποίο ανήκει στο 47.5% των παρατηρήσεων.

Αφού οι συνολικές παρατηρήσεις είναι 2000, το 47.5% αυτών είναι 900. **Συνεπώς 900 τηλεφωνήματα είχαν διάρκεια από 3 έως 5 λεπτά.**

ii)Πόσα τηλεφωνήματα είχαν διάρκεια μεγαλύτερη από 7 λεπτά.

Με την ίδια μέθοδο, ψάχνουμε το διάστημα που είναι μεγαλύτερο ή ίσο από το ($\bar{x}$ + 2s), το οποίο ανήκει στο 2.5% των παρατηρήσεων που είναι 50. **Συνεπώς 50 τηλεφωνήματα είχαν διάρκεια μεγαλύτερη από 7 λεπτά.**

(Marks : 20)

2. Οι παρατηρήσεις μιας μεταβλητής x μεγέθους 800 ακολουθούν την κανονική κατανομή. Είκοσι παρατηρήσεις είναι μικρότερες του 18 και 128 μεγαλύτερες του 36

α) Να βρείτε κατά προσέγγιση το εύρος του δείγματος.

Από την 5^{η} ιδιότητα που έχει αναφερθεί παραπάνω, σχετικά με το εύρος σε ένα σύνολο παρατηρήσεων που ακολουθούν κανονική κατανομή, γνωρίζουμε ότι το εύρος είναι προσεγγιστικά ίσο με 6 τυπικές αποκλίσεις (R ≈ 6s).

Στην εκφώνηση αναφέρεται ότι 20 από τις 800 παρατηρήσεις είναι μικρότερες του 18. Το 20 όμως είναι ακριβώς το 2.5% του 800 (20/800=0,025), οπότε συμπεραίνουμε ότι $\bar{x}$ - 2s = 18.

Αναφέρεται επίσης ότι 128 από τις 800 παρατηρήσεις είναι μεγαλύτερες του 36, όμως το 128 είναι ακριβώς το 16% του 800 (128/800=0,16) , οπότε συμπεραίνουμε ότι $\bar{x}$ + s = 36.

Προκύπτει πάλι ένα σύστημα 2 εξισώσεων με 2 αγνώστους:

- $\bar{x}$ - 2s = 18
- $\bar{x}$ + s = 36

Από το παραπάνω σύστημα χρησιμοποιώντας τη μέθοδο της αντικατάστασης προκύπτει ότι $\bar{x}$ = 30 και **s = 6.**

Συνεπώς, συμπεραίνουμε ότι το εύρος R του δείγματος προσεγγιστικά είναι 6*6 = 36

β) Να εξετάσετε αν το δείγμα των παρατηρήσεων είναι ομοιογενές.

Συντελεστής μεταβολής ονομάζεται ο λόγος της τυπικής απόκλισης προς τη μέση τιμή, δηλαδή

$CV = s / \bar{x}$, $(\bar{x} \neq 0)$

Ο συντελεστής μεταβολής είναι ένα μέτρο ομοιογένειας, το οποίο χρησιμοποιείται για τη σύγκριση ομάδων τιμών που εκφράζονται είτε σε διαφορετικές μονάδες μέτρησης είτε στην ίδια μονάδα μέτρησης αλλά έχουν σημαντικά διαφορετικές μέσες τιμές.

Ο συντελεστής μεταβολής είναι ανεξάρτητος από τις μονάδες μέτρησης, εκφράζεται επί τοις εκατό, και είναι ένα μέτρο σχετικής διασποράς των τιμών και όχι της απόλυτης διασποράς.

Η σημαντική ιδιότητα που έχει και μας ενδιαφέρει για το συγκεκριμένο ερώτημα είναι πως **ένα δείγμα τιμών Α θεωρείται ομοιογενές όταν $CV_A \leq 10\%$.**

Από το ερώτημα α έχουμε βρει $\bar{x} = 30$ και $s = 6$, οπότε έχουμε $CV = s / \bar{x} = 6/30 = 0.2 =$ **20%. Άρα CV = 20%.**
Συνεπώς <u>όχι</u>, το δείγμα των παρατηρήσεων δεν είναι ομοιογενές.

(Marks : 20)

3. Δημιουργήστε μια συνάρτηση η οποία θα δέχεται ως όρισμα ένα διάνυσμα x, εν συνεχεία θα το ταξινομεί κατά αύξουσα τάξη μεγέθους και θα υπολογίζει τη διάμεσο. (η διάμεσος είναι η μεσαία διατεταγμένη παρατήρηση όταν ο αριθμός των παρατηρήσεων είναι περιττός αριθμός ή το άθροισμα των δύο μεσαίων διατεταγμένων παρατηρήσεων δια 2, όταν ο αριθμός των παρατηρήσεων είναι άρτιος αριθμός.)

Αυτό που ζητάει το συγκεκριμένο ερώτημα είναι στην ουσία η υλοποίηση της συνάρτησης **median** στην R, η οποία σύμφωνα με το επίσημο documentation της R κάνει ακριβώς αυτό που αναφέρεται στην εκφώνηση. Μια σημαντική λεπτομέρεια, <u>η οποία θα υλοποιηθεί σαν επιπρόσθετο feature της συνάρτησης παρ' όλο που δεν αναφέρεται στην εκφώνηση</u>, είναι το δεύτερο input που δέχεται η median και έχει να κάνει με τις NA τιμές. Η median εκτός από το

διάνυσμα που το δέχεται ως 1^{ο} input, δέχεται και μια boolean μεταβλητή na.rm η οποία έχει default τιμή FALSE.
Αν na.rm = FALSE και το διάνυσμα που θα δοθεί στην median έχει μέσα NA τιμές, τότε το αποτέλεσμα της συνάρτησης θα είναι και αυτό NA, ενώ αν na.rm = TRUE τότε η συνάρτηση αφαιρεί τις NA τιμές και επιστρέφει κανονικά την διάμεσο του διανύσματος.
Συνημμένο υπάρχει R αρχείο (*Assignment_2018506_Exercise3_R_Code.r*) που περιέχει την υλοποίηση της ανωτέρω συνάρτησης, όπως και ενδεικτικές εκτελέσεις της συνάρτησης με διάφορα διανύσματα (ακεραίων, δεκαδικών, με NA τιμές κλπ).
Κώδικας R:

```
# Author: Nikos Kougianos
# Date: 13/12/2020

custom_median <- function(x, na.rm = FALSE) {
  # Check if vector has any NA values
  if (anyNA(x)) {
    # Check na.rm if TRUE or FALSE
    if (na.rm == TRUE) {
      # If true, remove NA values from x
      x <- x[!is.na(x)]

    } else {
      # If false, then return NA as result like median function
      return(NA)

    }

  }

  # Sort x
  x = sort(x)

  # Check if length of x is even or odd
  # If even, then we want to return the mean of the 2 middle elements
  # If odd, then we want to return only the middle element
  if (length(x) %% 2 == 0) {
```

```
    middle1 <- x[length(x) %/% 2]
    middle2 <- x[length(x) %/% 2 + 1]
    return(mean(c(middle1, middle2)))
  } else {
    return(x[length(x) %/% 2 + 1])
  }

}

# Sample executions of custom_median function, using integer vectors,
# double vectors, vectors with NA values and vectors without NA values
custom_median(runif(10, 0, 100), FALSE)
custom_median(runif(150, 0, 100), FALSE)
custom_median(sample.int(100, 9), FALSE)
custom_median(sample.int(100, 200, replace=TRUE), FALSE)
custom_median(c(NA,1,5,15.6,NA,6,7,8,9), FALSE)
custom_median(c(NA,1,5,15.6,NA,6,7,8,9), TRUE)
custom_median(c(1,5,15.6,6,7,8,9), FALSE)
```

(Marks : 20)

B. Data Analysis Project

4. Ο παρακάτω πίνακας παρουσιάζει τις μέσες τιμές βαθμολογιών επιβατών που έλαβαν οχτώ ποιοτικά χαρακτηριστικά σε μια έρευνα αξιολόγησης των παρεχόμενων υπηρεσιών ενός οργανισμού αστικών συγκοινωνιών. Για τη βαθμολόγηση των χαρακτηριστικών χρησιμοποιήθηκε μια κλίμακα από 1 έως 5, όπου για τη βαθμολόγηση της σημαντικότητας ισχύει: 1= Πολύ ασήμαντο και 5 = Πολύ σημαντικό, ενώ για τη βαθμολόγηση της ικανοποίησης ισχύει: 1 = Καθόλου ικανοποιημένος και 5 = Πολύ ικανοποιημένος.

	Σημαντικότητα	**Ικανοποίηση**
1. Άνεση οχημάτων	4.1	2.4
2. Καθαριότητα οχημάτων	4.2	1.4
3. Πληροφόρηση στις στάσεις	1.8	3.9
4. Πληροφόρηση εντός οχημάτων	1.3	1.6
5. Ασφάλεια στις στάσεις	4.3	3.3
6. Ασφάλεια εντός οχημάτων	3.8	3.9
7. Συμπεριφορά προσωπικού	1.8	4.5
8. Διαθεσιμότητα εισιτηρίων	3.5	1.9

Εφαρμόζοντας τη μέθοδο Ανάλυσης Τεταρτημορίων βρείτε τα ποιοτικά χαρακτηριστικά που πρέπει να βελτιώσει ο οργανισμός αστικών συγκοινωνιών, σύμφωνα με την άποψη των επιβατών.

Η Ανάλυση Τεταρτημορίων (Quadrant Analysis) είναι μια απλή στην εφαρμογή μέθοδος ανάλυσης ποιοτικών χαρακτηριστικών, η οποία χρησιμοποιεί 2 άξονες κάθετους μεταξύ τους και τοποθετεί τις παρατηρήσεις στα 4 τεταρτημόρια ίδιου εμβαδού με βάση τα χαρακτηριστικά τους.

Χρησιμοποιείται ευρέως από μεγάλες αλλά και μικρότερες εταιρίες, αλλά και σε ανεξάρτητες έρευνες από διάφορους φορείς, προκειμένου να μπορούν να οπτικοποιηθούν με σχετική ευκολία τα υπό αξιολόγηση χαρακτηριστικά, με σκοπό να βγουν συμπεράσματα για το ποια από αυτά χρήζουν άμεσης βελτίωσης, αλλά και ποια από αυτά είναι ήδη σε υψηλό επίπεδο και λόγω της σημαντικότητας τους θα πρέπει να παραμείνουν υψηλά.

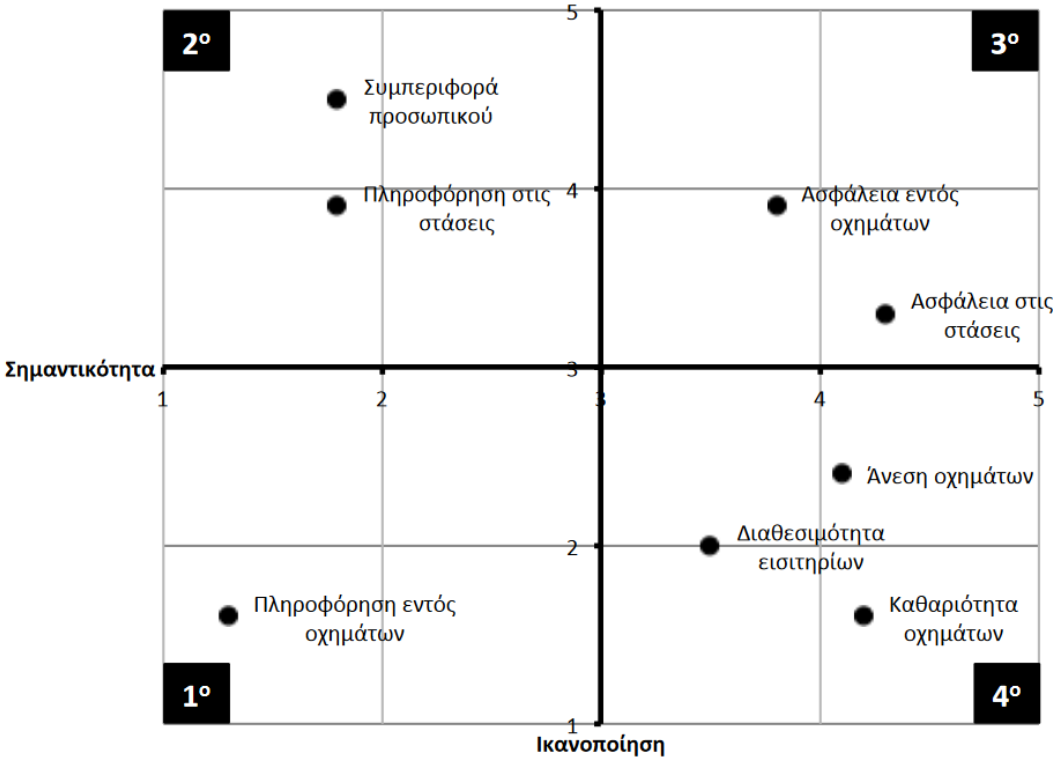
Οι 2 άξονες βάσει των οποίων γίνεται η ανάλυση τεταρτημορίων είναι συνήθως η ικανοποίηση/βαθμολογία του χαρακτηριστικού, και η σημαντικότητα/βάρος η οποία αντιστοιχεί στο εκάστοτε χαρακτηριστικό.

Για την εφαρμογή της ανάλυσης τεταρτημορίων, χρειάζεται να υπάρχουν δεδομένα των ποιοτικών χαρακτηριστικών, όπως είναι ο πίνακας της εκφώνησης. Το πρώτο βήμα στην εφαρμογή της Ανάλυσης Τεταρτημορίων είναι ο καθορισμός των ποιοτικών χαρακτηριστικών (ποιοτικοί δείκτες αξιολόγησης), τα οποία θα αξιολογηθούν από το επιβατικό κοινό. Μέσω έρευνας ερωτηματολογίου, οι επιβάτες δίνουν σε κάθε ποιοτικό χαρακτηριστικό δύο βαθμολογίες, που φαίνονται στον παραπάνω πίνακα: μια ως προς τη σημαντικότητα του χαρακτηριστικού και μια ως προς την ικανοποίηση των επιβατών αναφορικά με την απόδοση του χαρακτηριστικού. Για κάθε χαρακτηριστικό, πρώτα αξιολογείται η σημαντικότητα που έχει σύμφωνα με κάθε χρήστη (επιβάτη), και στη συνέχεια αξιολογείται η ίδια του η απόδοση. Η βαθμολόγηση των ποιοτικών χαρακτηριστικών γίνεται χρησιμοποιώντας την κλίμακα Likert (1932) (kallipos, n.d.).

Ένα μειονέκτημα της Ανάλυσης Τεταρτημορίων είναι τα όρια των τεταρτημορίων, δηλαδή το σημείο τομής των δύο αξόνων του διαγράμματος ικανοποίησης - σημαντικότητας, τα οποία είναι συνήθως αυθαίρετα και το μέγεθος των διαφορών μεταξύ των μέσων βαθμολογιών των ποιοτικών χαρακτηριστικών συνήθως δεν λαμβάνεται υπόψιν.

Η κλίμακα Likert είναι μια κλίμακα συνήθως 5 (ή 7) βαθμών που δίνουν την δυνατότητα σε έναν χρήστη να εκφράσει πόσο πολύ συμφωνεί με έναν ισχυρισμό/γεγονός. Η συγκεκριμένη κλίμακα θεωρεί πως η διαφορά ανάμεσα στους βαθμούς 1-5 (ή 1-7) είναι ίση σε όλα τα στάδια, δηλαδή το 1 διαφέρει από το 2 όσο και το 4 από το 5, είναι δηλαδή γραμμική (McLeod, 2019).

Για να σχεδιαστεί η διαγραμματική απεικόνιση των αποτελεσμάτων της Ανάλυσης Τεταρτημορίων, παίρνουμε τις τιμές των χαρακτηριστικών και χρησιμοποιώντας τους 2 προαναφερθέντες άξονες σχεδιάζουμε το γράφημα:



Εικόνα 2 - Γράφημα Ανάλυσης Τεταρτημορίων

Από το ανωτέρω γράφημα, βγαίνουν τα εξής συμπεράσματα:

- Τα χαρακτηριστικά Άνεση οχημάτων, Διαθεσιμότητα εισιτηρίων και Καθαριότητα οχημάτων, θεωρούνται σημαντικά από τους επιβάτες και έχουν λάβει χαμηλές βαθμολογίες.
- Το χαρακτηριστικό Πληροφόρηση εντός οχημάτων έχει λάβει χαμηλή βαθμολογία αλλά δε θεωρείται σημαντικό.
- Τα χαρακτηριστικά Συμπεριφορά προσωπικού και Πληροφόρηση στις στάσεις έχουν λάβει υψηλή βαθμολογία, αλλά δε θεωρούνται σημαντικά.
- Τα χαρακτηριστικά Ασφάλεια εντός οχημάτων και Ασφάλεια στις στάσεις έχουν λάβει υψηλή βαθμολογία, και θεωρούνται και σημαντικά από τους επιβάτες.

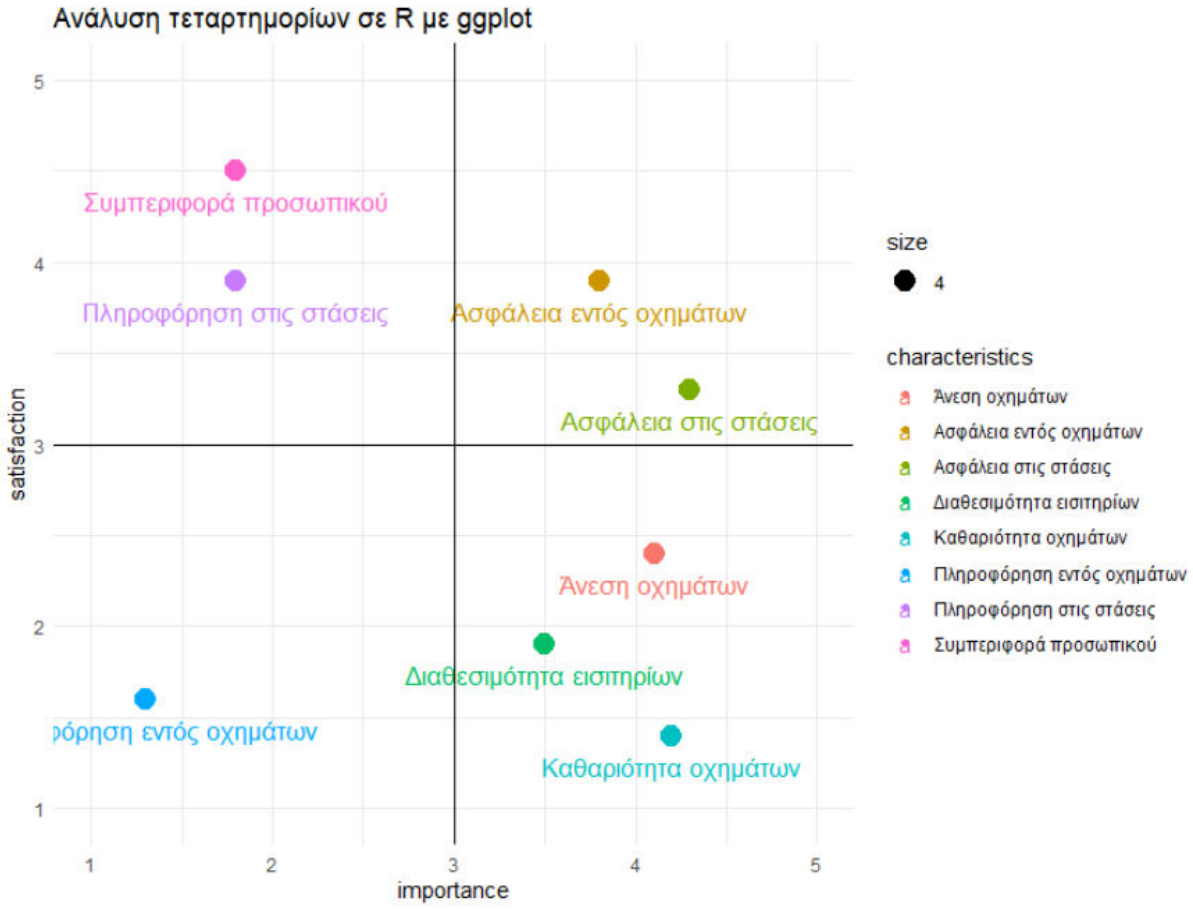
Βάσει της παραπάνω ανάλυσης, μπορούμε να οδηγηθούμε στο εξής πλάνο αναφορικά με τις ενέργειες που πρέπει να κάνει ο οργανισμός αστικών συγκοινωνιών:

- **Η απάντηση στο ερώτημα της εκφώνησης, είναι πως τα ποιοτικά χαρακτηριστικά τα οποία πρέπει άμεσα να βελτιώσει ο οργανισμός είναι η Άνεση οχημάτων, η Διαθεσιμότητα εισιτηρίων, και η Καθαριότητα οχημάτων.**

- Τα χαρακτηριστικά Συμπεριφορά προσωπικού και Πληροφόρηση στις στάσεις δε θεωρούνται σημαντικά, συνεπώς ο οργανισμός μπορεί να πάρει κάποια από την προσοχή (ή/και budget) που έχει αφιερώσει σε αυτά τα 2 χαρακτηριστικά και να τα διαθέσει στα 3 προηγούμενα που θεωρούνται σαφώς πιο σημαντικά.
- Θα πρέπει να γίνει προσπάθεια διατήρησης της υψηλής βαθμολογίας των χαρακτηριστικών του 3ου τεταρτημορίου, καθώς θεωρούνται σημαντικά και θα βοηθήσουν τον οργανισμό να διατηρήσει το ανταγωνιστικό πλεονέκτημα απέναντι σε άλλους ανταγωνιστές.

EXTRA ΥΛΟΠΟΙΗΣΗ ΣΕ R ΤΗΣ ΑΝΑΛΥΣΗΣ ΤΕΤΑΡΤΗΜΟΡΙΩΝ

Χρησιμοποιώντας το εργαλείο της R σε συνδυασμό με τη βιβλιοθήκη ggplot2, βγαίνει ένα γράφημα που είναι παρόμοιο με το παραπάνω, αλλά με περισσότερες λεπτομέρειες (χρώματα, ονόματα):



Εικόνα 3- Ανάλυση τεταρτημορίων σε R με ggplot2

Το παραπάνω γράφημα παράχθηκε από τον παρακάτω R κώδικα, ο οποίος μπορεί να βρεθεί και στο συνημμένο αρχείο *Assignment_2018506_Exercise4_QuadrantAnalysis.r*

```
# Author: Nikos Kougianos
# Date: 14/12/2020

library(ggplot2)

characteristics <- c('Άνεση οχημάτων', 'Καθαριότητα οχημάτων',
                     'Πληροφόρηση στις στάσεις', 'Πληροφόρηση εντός οχημάτων',
                     'Ασφάλεια στις στάσεις', 'Ασφάλεια εντός οχημάτων',
                     'Συμπεριφορά προσωπικού','Διαθεσιμότητα εισιτηρίων')

importance <- c(4.1,4.2,1.8,1.3,4.3,3.8,1.8,3.5)

satisfaction <- c(2.4,1.4,3.9,1.6,3.3,3.9,4.5,1.9)

dataset <- data.frame(characteristics, importance, satisfaction)

graph<-ggplot(dataset,  aes(x=importance,  y=satisfaction,  colour=characteristics,
size = 4)) +
   geom_point() +
   lims(x=c(1,5),y=c(1,5)) +
   theme_minimal() +
   coord_fixed() +
   geom_text(aes(label=characteristics), vjust=2) +
   ggtitle("Ανάλυση τεταρτημορίων σε R με ggplot") +
   geom_vline(xintercept = 3) + geom_hline(yintercept = 3)

graph
```

(Marks : 40)

Αναφορές

kallipos. (χ.χ.). *repository.kallipos*. Ανάκτηση από repository.kallipos: https://repository.kallipos.gr/bitstream/11419/3566/1/02_chapter_7.pdf

McLeod, S. (2019). *simplypsychology*. Ανάκτηση από simplypsychology: https://www.simplypsychology.org/likert-scale.html