

Στατιστική Υπολογιστική

Βασίλειος Παπαποστόλου - 3176

Χρήστος Περχανίδης - 3194

Μαρία Εσκιόγλου - 3237

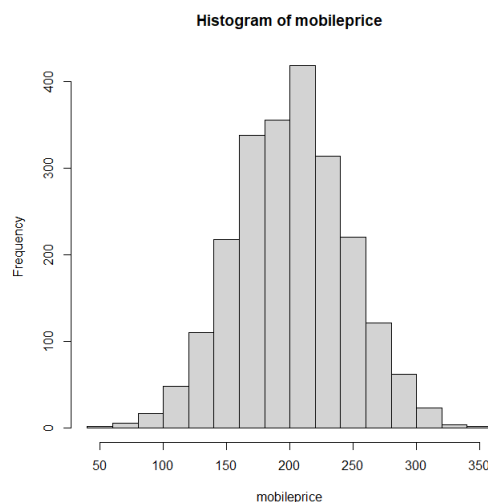
27 Ιουνίου 2021

Listing 1: R output

psraid	usr_r	sex	q10c
Min. :100004	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:101927	1st Qu.:1.000	1st Qu.:1.000	1st Qu.:1.000
Median :104000	Median :2.000	Median :2.000	Median :1.000
Mean :135230	Mean :1.938	Mean :1.565	Mean :1.148
3rd Qu.:200585	3rd Qu.:2.000	3rd Qu.:2.000	3rd Qu.:1.000
Max. :202531	Max. :3.000	Max. :2.000	Max. :2.000
	NA's :55		NA's :1
q14a	q14b	q14c	q14d
Min. :1.000	Min. :1.00	Min. :1.000	Min. :1.000
1st Qu.:1.000	1st Qu.:1.00	1st Qu.:1.000	1st Qu.:1.000
Median :2.000	Median :1.00	Median :1.000	Median :2.000
Mean :1.715	Mean :1.38	Mean :1.311	Mean :1.746
3rd Qu.:2.000	3rd Qu.:2.00	3rd Qu.:2.000	3rd Qu.:2.000
Max. :2.000	Max. :2.00	Max. :2.000	Max. :2.000
NA's :335	NA's :335	NA's :337	NA's :335
q29	age	mar	educ
Min. :1.000	Min. :18.00	Min. :1.000	Min. :1.000
1st Qu.:1.000	1st Qu.:37.00	1st Qu.:1.000	1st Qu.:3.000
Median :2.000	Median :52.00	Median :1.000	Median :5.000
Mean :1.529	Mean :50.98	Mean :2.676	Mean :4.528
3rd Qu.:2.000	3rd Qu.:64.00	3rd Qu.:5.000	3rd Qu.:6.000
Max. :2.000	Max. :97.00	Max. :7.000	Max. :7.000
NA's :1827	NA's :51	NA's :16	NA's :16
empl	inc	mobileprice	
Min. :1.000	Min. :1.000	Min. : 47.8	
1st Qu.:1.000	1st Qu.:3.000	1st Qu.:170.2	
Median :2.000	Median :5.000	Median :201.4	
Mean :2.359	Mean :5.042	Mean :201.3	
3rd Qu.:3.000	3rd Qu.:7.000	3rd Qu.:231.2	
Max. :7.000	Max. :9.000	Max. :345.3	
NA's :20	NA's :408		

Στα δεδομένα `CellPhones_values` υπάρχει μια βασική εξαρτημένη μεταβλητή (απόκρισης) η τιμή του τηλεφώνου. Οι κατηγορικές μεταβλητές (παράγοντες) που αναγράφονται στη συλλογή δεδομένων παίρνουν θετικές ακέραιες τιμές.

Αρχικά με την εντολή **`summary(dataset)`** θα δείξουμε τα περιγραφικά μέτρα για μερικές από τις μεταβλητές του συνόλου δεδομένων μας. Για κάθε μία από τις αναγραφόμενες μεταβλητές μπορούμε να δούμε την ελάχιστη και την μέγιστη τιμή, την μέση τιμή, την διάμεσο και το 1ο και 3ο τεταρτημόριο. Συγκεκριμένα για την βασική εξαρτημένη μεταβλητή (απόκρισης), την `mobileprice` μπορούμε να πούμε ότι η μέγιστη τιμή ενός κινητού που αγοράστηκε και υπάρχει στο σύνολο δεδομένων είναι η 346.3 και η μικρότερη 47.8. Στην ίδια μεταβλητή μπορούμε να διακρίνουμε ότι η μέση τιμή και η διάμεσος έχουν παρόμοιες τιμές (201.4 και 201.3) άρα υπάρχει συμμετρία, όπως φαίνεται και στο ιστόγραμμα 1.



Σχήμα 1: Ιστόγραμμα της μεταβλητής `mobileprice`

Μετά την ανάγνωση των μεταβλητών, ξεχωρίσαμε κάποιες που θεωρούμε ότι συμβάλλουν περισσότερο στην τιμή του κινητού τηλεφώνου. Αυτές είναι οι:

1. `psraid`: ID
2. `usr_r`: Η περιοχή κατοικίας του χρήστη. Οι πιθανές τιμές είναι η 1, οπότε ο χρήστης μένει σε Urban περιοχή, 2 είναι η Suburban και 3 η Rural.
3. `sex`: Το φύλο του χρήστη δηλαδή 1 για αρσενικό και 2 για θηλυκό.
4. `age`: Η ηλικία.
5. `mar`: Η οικογενειακή κατάσταση του χρήστη. Υπάρχουν 7 επιλογές για τους παντρεμένους, αυτούς που συζούν, τους χωρισμένους, τους εν

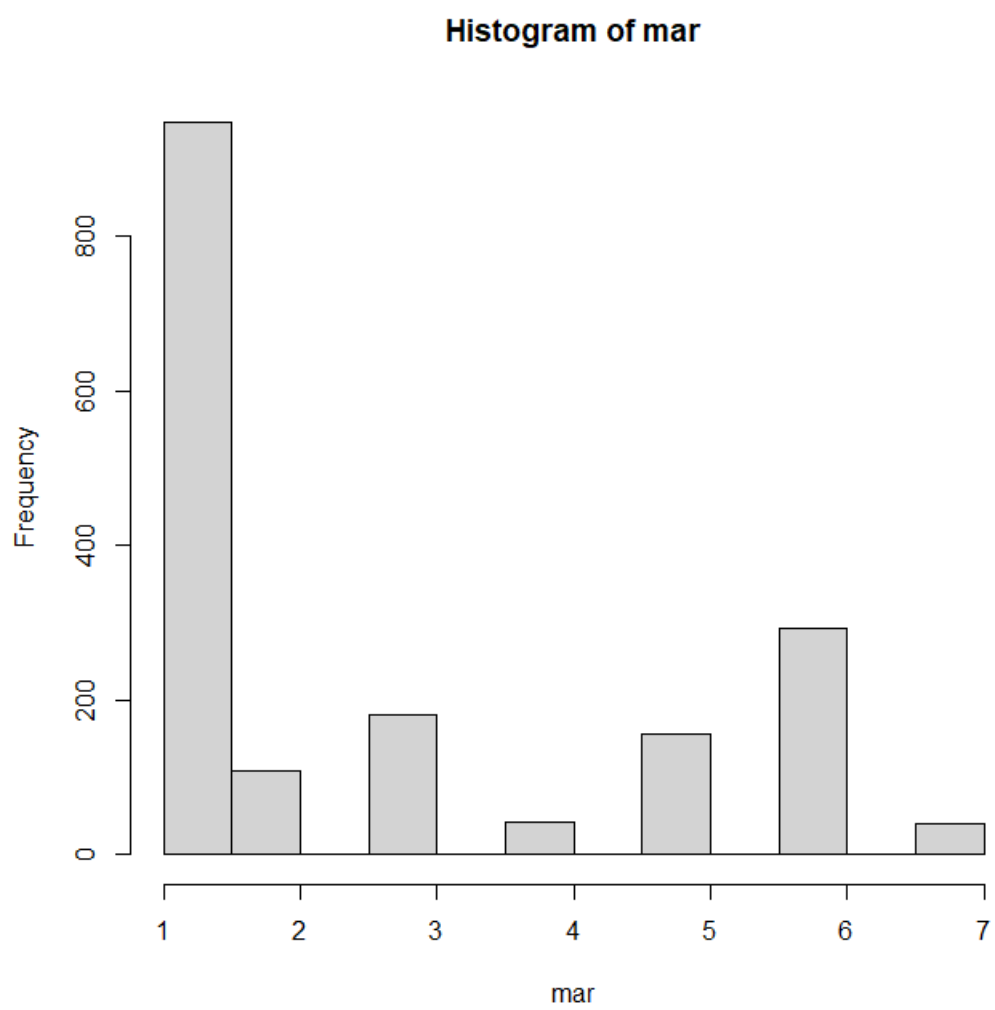
διάσταση, αυτούς που έχουν χηρέψει, αυτούς που δεν έχουν παντρευτεί ποτέ και τους ελεύθερους.

6. educ: Την εκπαίδευση.
7. empl: Το επάγγελμα.
8. inc: Το εισόδημα του κάθε χρήστη.

Με την εντολή **summary(main_df)** θα δούμε τα περιγραφικά μέτρα του υποσυνόλου των μεταβλητών που επιλέξαμε. Από την διαφορά μέσης τιμής και διαμέσου μπορούμε να υποθέσουμε ότι στην μεταβλητή mar θα υπάρχει ασυμμετρία κάτι που επιβεβαιώνεται από το ιστόγραμμα 2.

Listing 2: R output

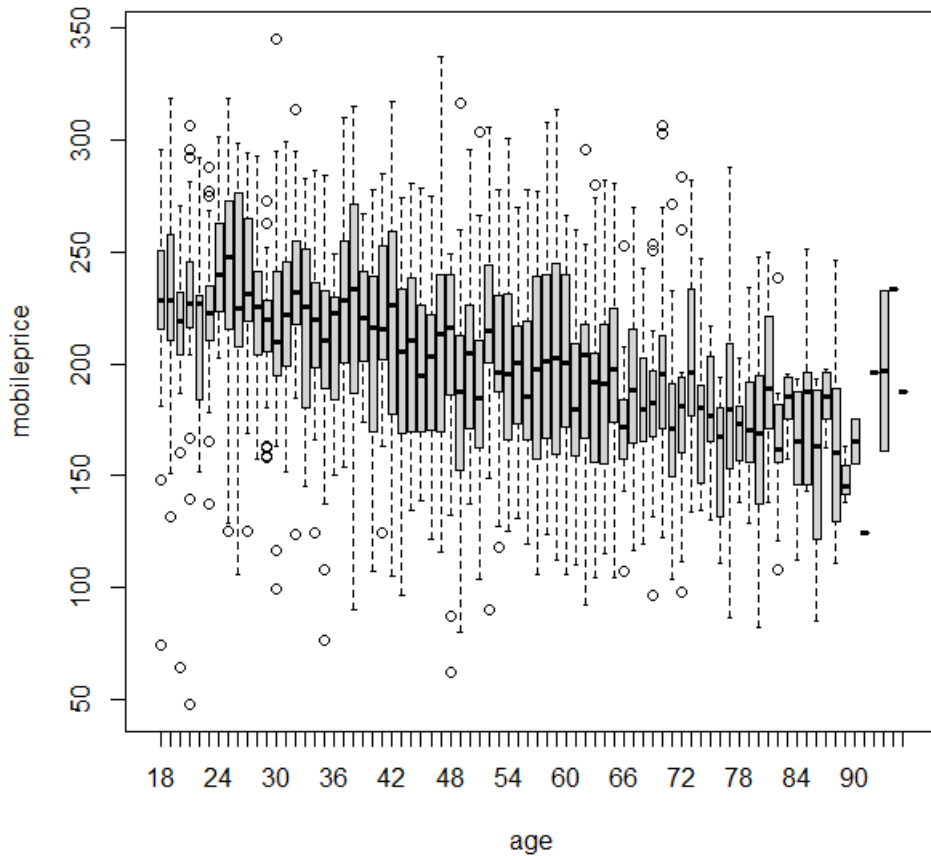
psraid	usr_r	sex	age
Min. :100004	Min. :1.000	Min. :1.000	Min. :18.00
1st Qu.:102034	1st Qu.:1.000	1st Qu.:1.000	1st Qu.:35.00
Median :104127	Median :2.000	Median :2.000	Median :51.00
Mean :136415	Mean :1.938	Mean :1.542	Mean :49.75
3rd Qu.:200646	3rd Qu.:2.000	3rd Qu.:2.000	3rd Qu.:62.00
Max. :202530	Max. :3.000	Max. :2.000	Max. :97.00
mar	educ	empl	inc
Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:1.000	1st Qu.:3.000	1st Qu.:1.000	1st Qu.:3.000
Median :1.000	Median :5.000	Median :2.000	Median :5.000
Mean :2.654	Mean :4.563	Mean :2.267	Mean :5.049
3rd Qu.:5.000	3rd Qu.:6.000	3rd Qu.:3.000	3rd Qu.:7.000
Max. :7.000	Max. :7.000	Max. :7.000	Max. :9.000
mobileprice			
Min. : 47.8			
1st Qu.:173.2			
Median :204.1			
Mean :203.9			
3rd Qu.:233.7			
Max. :345.3			



Σχήμα 2: Ιστόγραμμα της μεταβλητής mar

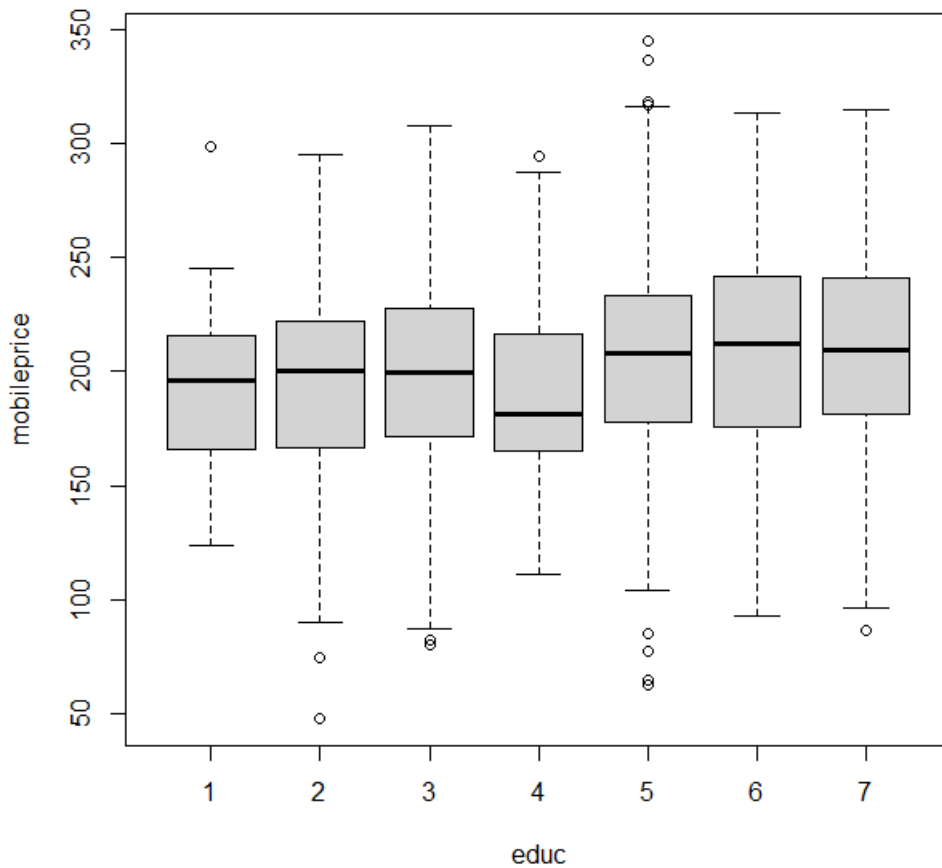
Μελέτη των boxplot

- age: Η πρώτη μεταβλητή είναι η ηλικία. Γνωρίζουμε ότι το ύψος του κουτιού ισούται με το ενδοτεταρτημοριακό εύρος και ότι πέφτει 1.5 φορές κάτω από το πρώτο τεταρτημόριο ή πάνω από το τρίτο θεωρείται έκτοπο σημείο η κατάλοιπο. Στην περίπτωση μας τα outliers συμβολίζονται με κυκλάκια και φαίνεται να υπάρχουν στις περισσότερες ηλικίες. Παρατηρούμε ότι όσο αυξάνεται η ηλικία η διάμεσος ελαττώνεται. Τα περισσότερα εύρη ηλικιών φαίνεται να έχουν μεγάλη μεταβλητότητα, ωστόσο καθώς αυξάνεται η ηλικία πάλι παρατηρείται μία μικρή μείωση στην μεταβλητότητα.



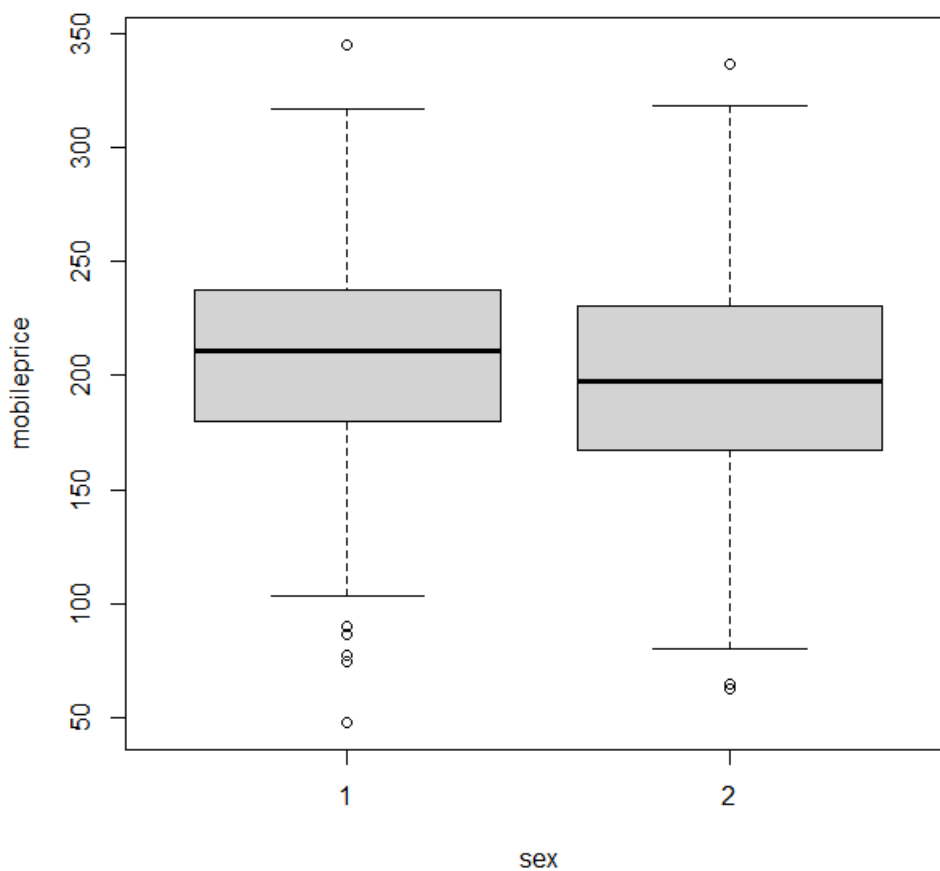
Σχήμα 3: Boxplot της μεταβλητής age

- educ: Στην δεύτερη μεταβλητή παρατηρούμε ότι το ενδοτεταρτημοριακό εύρος είναι παρόμοιο στις περισσότερες τιμές, δηλαδή η μεταβλητότητα δεν είναι πολύ διαφορετική. Και πάλι έχουμε outliers, με μεγαλύτερη συγκέντρωση να εμφανίζεται στην τιμή 5, όπου αντιστοιχεί στην περίπτωση που ένας χρήστης έχει πάει σε ένα κολλέγιο, δηλαδή δεν έκανε τετραετές σπουδές. Η τιμή της διαμέσου τείνει να γίνει ευθεία σε όλες τις περιπτώσεις, με μία μικρή απόκλιση στην τέταρτη επιλογή, δηλαδή στους μη εργαζόμενους χρήστες.



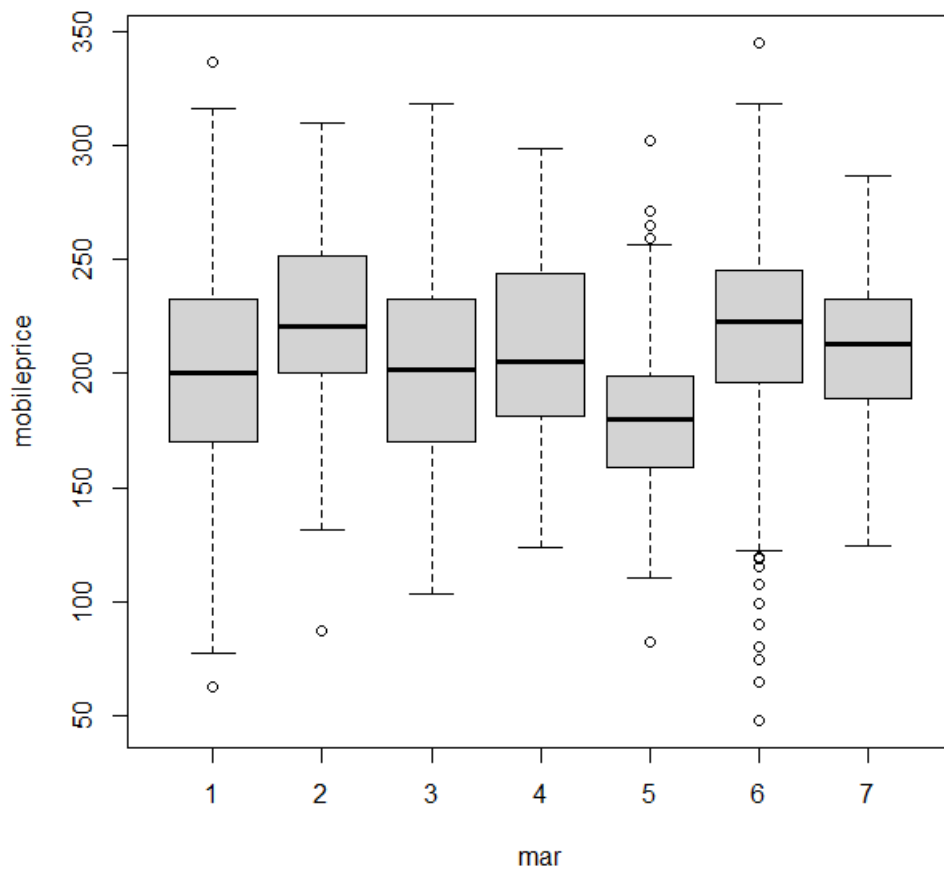
Σχήμα 4: Boxplot της μεταβλητής educ

- sex: Στην μεταβλητή για το φύλο έχουμε μόνο δύο επιλογές, αυτή του αρσενικού και του θηλυκού. Στο θηκόγραμμα τους παρατηρείται μία μικρή άνοδος της διαμέσου στους άντρες σε σχέση με τις γυναίκες, ωστόσο το ενδοτεταρτομοριακό εύρος φαίνεται ίδιο. Και οι δύο τιμές εμφανίζουν outliers, με μία ελαφρώς αυξημένη τάση στους άντρες κάτω από το πρώτο τεταρτημόριο.



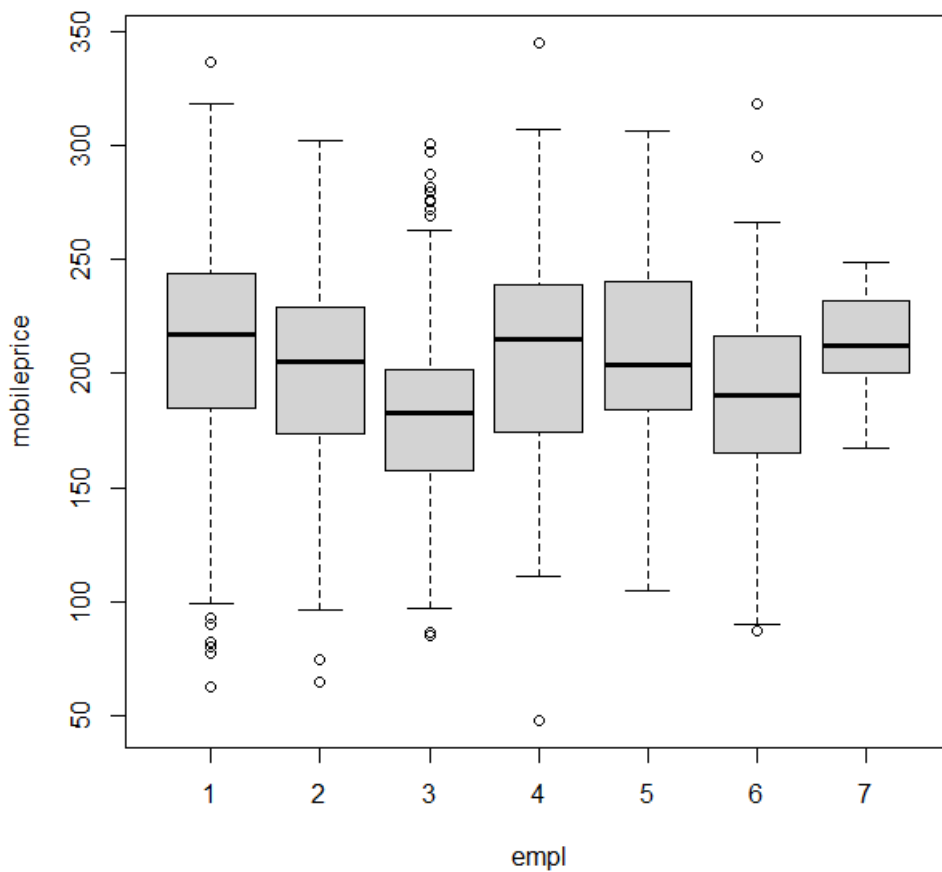
Σχήμα 5: Boxplot της μεταβλητής sex

- mar: Στην μεταβλητή για την οικογενειακή κατάσταση έχουμε επτά επιλογές. Στο θηκόγραμμα τους παρατηρείται ότι η διάμεσος δεν έχει μια ανοδική ή καθοδική πορεία, ενώ το ενδοτεταρτομοριακό εύρος και η μεταβλητότητα έχουν αυξομειώσεις. Οι περισσότερες τιμές εμφανίζουν outliers, με μία ελαφρώς αυξημένη τάση στους χρήστες που δεν έχουν παντρευτεί ποτέ κάτω από το πρώτο τεταρτημόριο. Τέλος, οι τιμές 3 και 7 που αντιστοιχούν στους χωρισμένους και στους ελεύθερους δεν εμφανίζουν outliers.



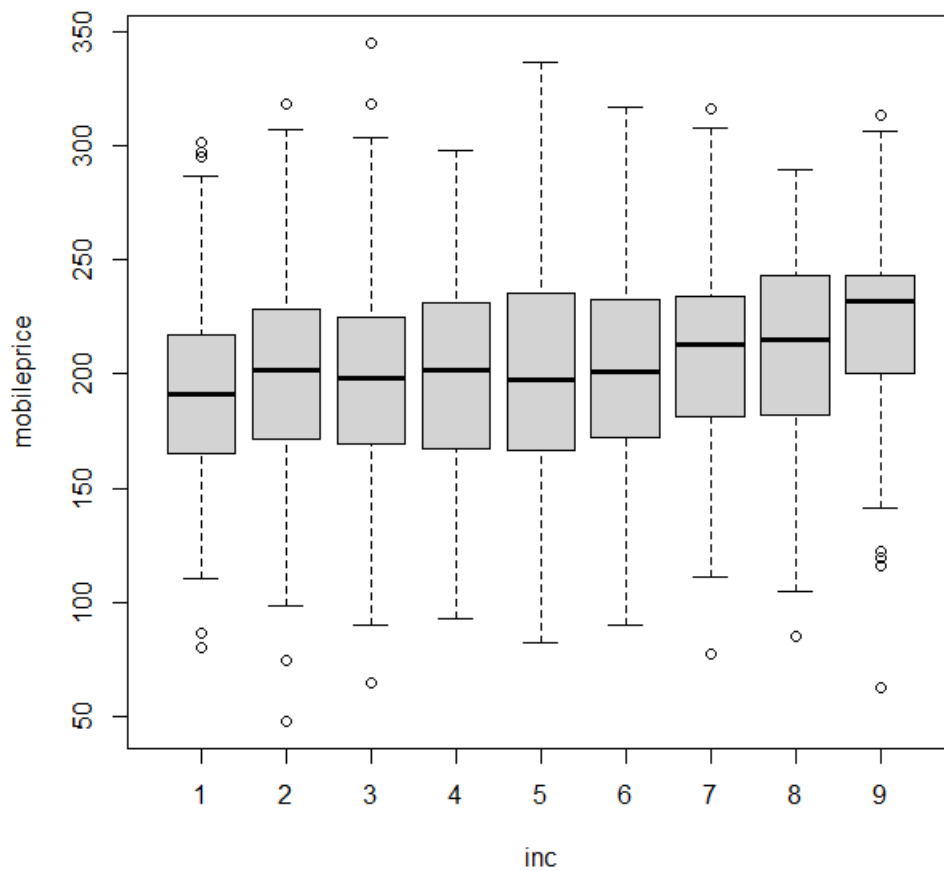
Σχήμα 6: Boxplot της μεταβλητής mar

- empl: Στην μεταβλητή για την επαγγελματική κατάσταση έχουμε επτά επιλογές. Στο θηκόγραμμα τους παρατηρείται ότι η διάμεσος δεν έχει μια ανοδική ή καθοδική πορεία, ενώ το ενδοτεταρτομοριακό εύρος και η μεταβλητότητα έχουν αυξομειώσεις. Οι περισσότερες τιμές εμφανίζουν outliers, με μία ελαφρώς αυξημένη τάση στους χρήστες που δουλεύουν full-time κάτω από το πρώτο τεταρτημόριο και στους συνταξιούχους πάνω από το τρίτο τεταρτημόριο. Τέλος, οι τιμές 5 και 7 που αντιστοιχούν στους self-employed και στους μαθητές/φοιτητές δεν εμφανίζουν outliers, ενώ οι τελευταίοι δεν εμφανίζουν και μεγάλη μεταβλητότητα.



Σχήμα 7: Boxplot της μεταβλητής empl

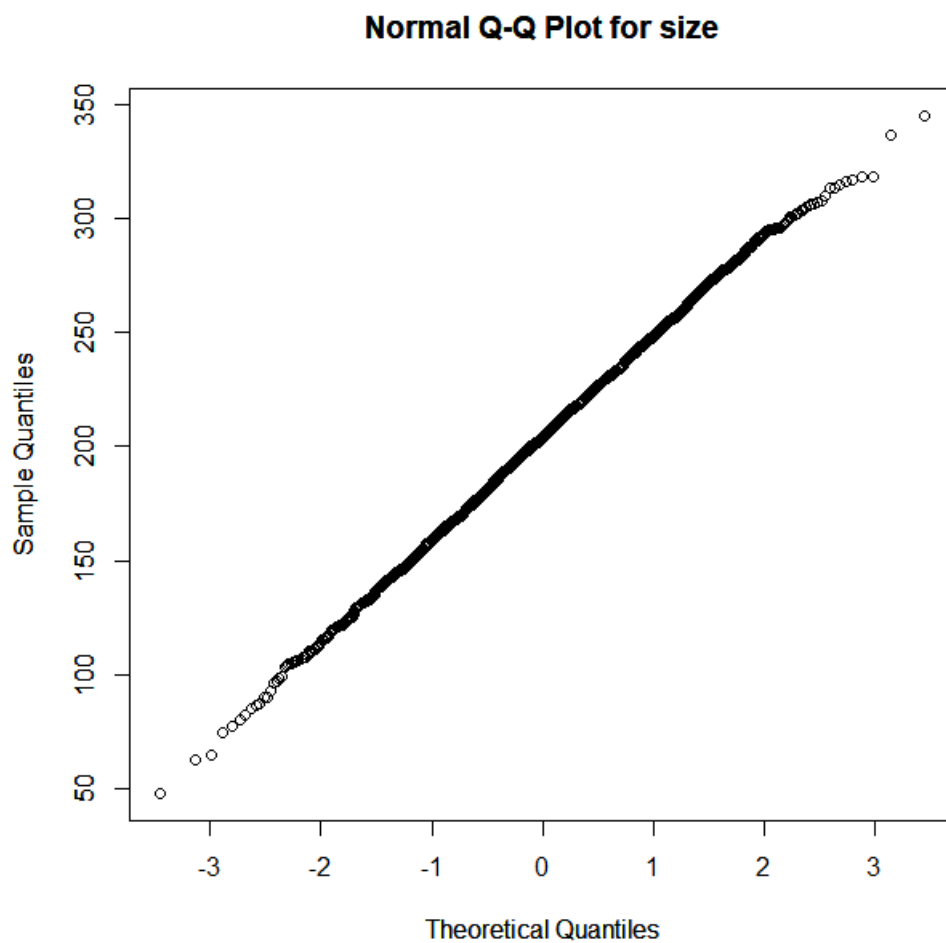
- inc: Στην μεταβλητή για την εισοδηματική κατάσταση έχουμε εννιά επιλογές. Στο θηκόγραμμα τους παρατηρείται μία μικρή άνοδος της διαμέσου, ενώ το ενδοτεταρτομοριακό εύρος και η μεταβλητότητα δεν εμφανίζουν σημαντικές αυξομειώσεις. Οι περισσότερες τιμές εμφανίζουν outliers, ωστόσο είναι μεμονομένα σημεία. Τέλος, οι τιμές 4 και 5 που αντιστοιχούν σε αυτούς με εισόδημα ανάμεσα στις 30.000\$ με 40.000\$ και 40.000\$ με 50.000\$ δεν εμφανίζουν outliers.



Σχήμα 8: Boxplot της μεταβλητής mobileprice

Έλεγχος Κανονικότητας

Για τον έλεγχο της κανονικότητας, σε πρώτη ανάγνωση και με γραφικό τρόπο εφαρμόσαμε την Q-Q plot. Ουσιαστικά, κάνουμε αντιπαραβολή των ταξινομημένων ποσοστών των δεδομένων με ταξινομημένα ποσοστά της κανονικής κατανομής. Όπως φαίνεται και από το παρακάτω διάγραμμα τα δεδομένα που έχουμε για την μεταβλητή φαίνεται να εφαρμόζουν σωστά πάνω στην ευθεία με ελάχιστες αποκλίσεις. Επομένως, μπορούμε να υποθέσουμε ότι η μεταβλητή ακολουθεί την κανονική κατανομή.



Σχήμα 9: Q-Q plot της μεταβλητής mobileprice

Σχέση Εξαρτημένης Μεταβλητής με Παράγοντες

Στο παραπάνω output της R βλέπουμε τα αποτελέσματα στην εντολή **tapply** πάνω στις διάφορες σημαντικές μεταβλητές που κρίναμε και στην εξαρτημένη τιμή `mobileprice`. Μερικές παρατηρήσεις που μπορούμε να κάνουμε είναι οι εξής:

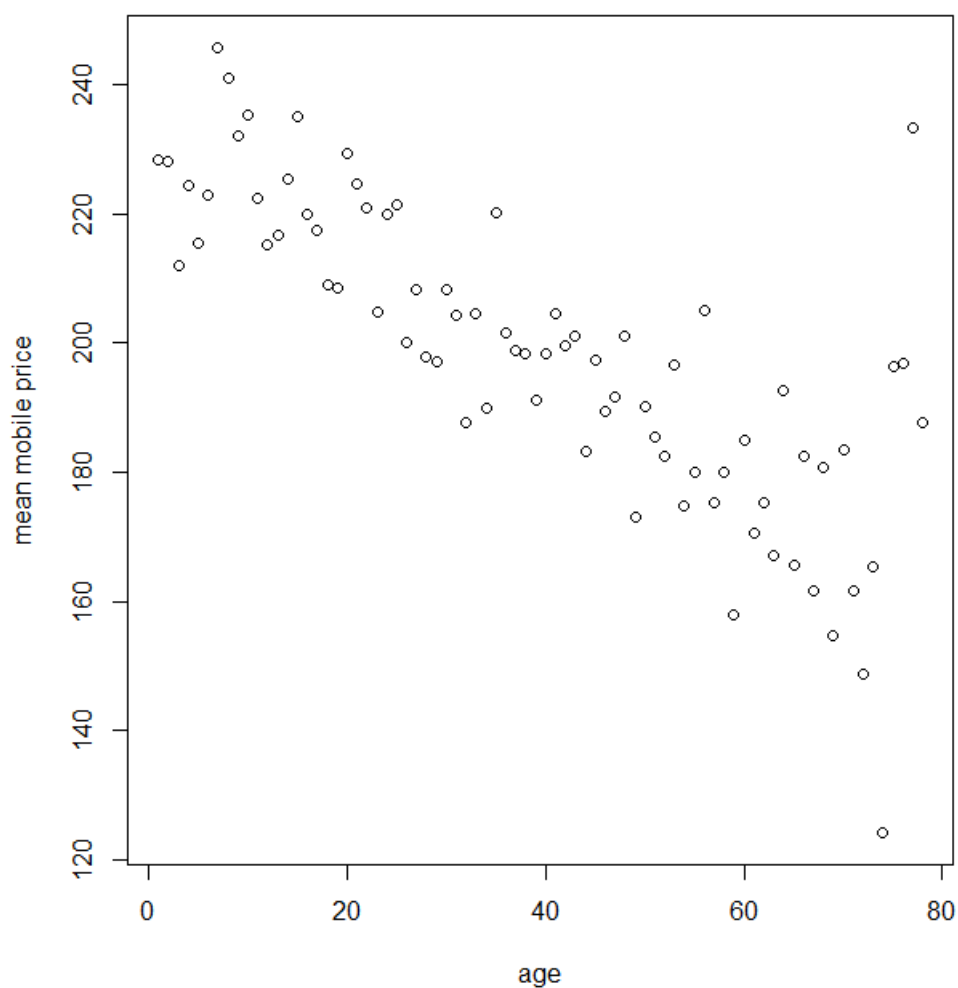
- Η μέση τιμή της αξίας ενός τηλεφώνου είναι κατά 10 ευρώ μεγαλύτερη στους άντρες από ότι στις γυναίκες.
- Στην δεύτερη σχέση που εξετάζουμε την μόρφωση του κάθε χρήστη παρατηρείται μία ανοδική πορεία στην μέση τιμή της αξίας του κινητού με μία μικρή απόκλιση στην τιμή 4, που αντιστοιχεί στους χρήστες που έχουν παρακολουθήσει τεχνικό σχολείο.
- Η ίδια άνοδος παρατηρείται και στην σχέση μεταξύ της εξαρτημένης μεταβλητής και του παράγοντα του εισοδήματος. Όπως θα περιμέναμε, οι χρήστες με εισόδημα μεγαλύτερο από 150.000\$ είναι αυτοί που μπορούν να διαθέσουν περισσότερα χρήματα για την απόκτηση ενός κινητού τηλεφώνου, όπως φαίνεται και από την μέση τιμή. Αντίστοιχα, η μικρότερη μέση τιμή παρατηρείται σε αυτούς με εισόδημα κάτω των 10.000\$.
- Στην μεταβλητή του επαγγέλματος η μεγαλύτερη μέση τιμή παρατηρείται στους επαγγελματίες που έχουν δουλειά πλήρους ωραρίου και η μικρότερη στους συνταξιοδοτημένους.
- Τέλος, η μεγαλύτερη μέση τιμή παρατηρείται σε αυτούς που συζούν και η μικρότερη στους χήρους.

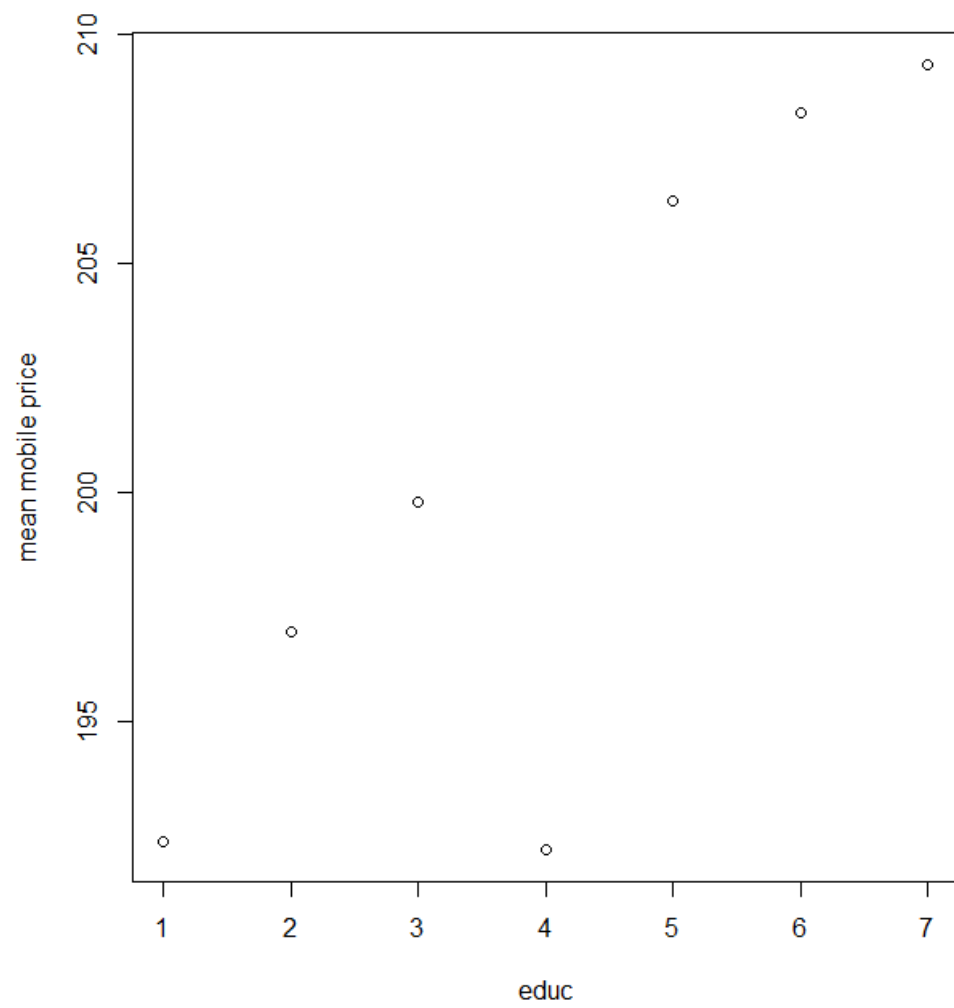
Listing 3: R output

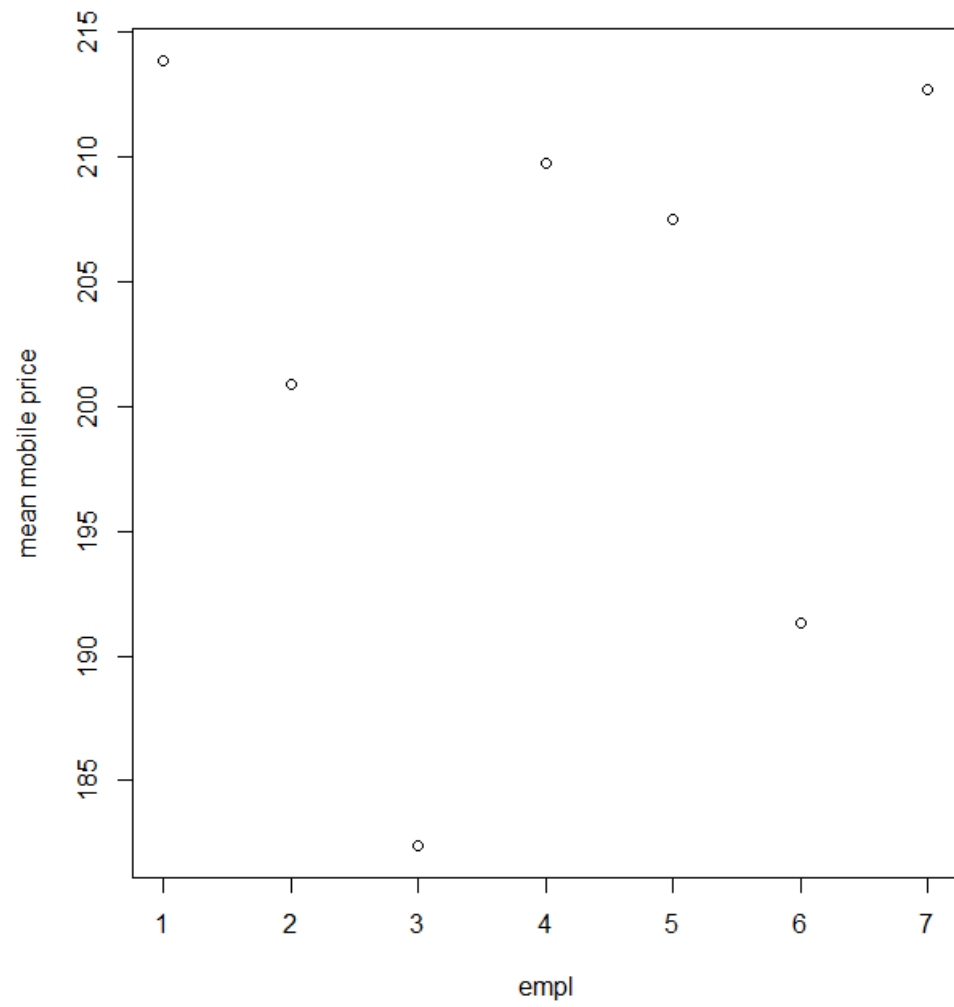
```
> tapply(mobileprice, sex, mean)
      1      2
209.3172 199.3089
> tapply(mobileprice, educ, mean)
      1      2      3      4      5      6      7
192.3444 196.9558 199.7908 192.1857 206.3578 208.2871 209.3319
> tapply(mobileprice, inc, mean)
      1      2      3      4      5      6      7
192.9958 201.3726 198.2603 201.4384 203.2799 202.4445 208.6763
      8      9
210.2543 219.9899
> tapply(mobileprice, empl, mean)
      1      2      3      4      5      6      7
213.8701 200.8823 182.3972 209.7516 207.5179 191.3347 212.7300
> tapply(mobileprice, mar, mean)
      1      2      3      4      5      6      7
200.4541 223.8306 203.1867 210.9163 182.4462 217.9437 207.9350
```

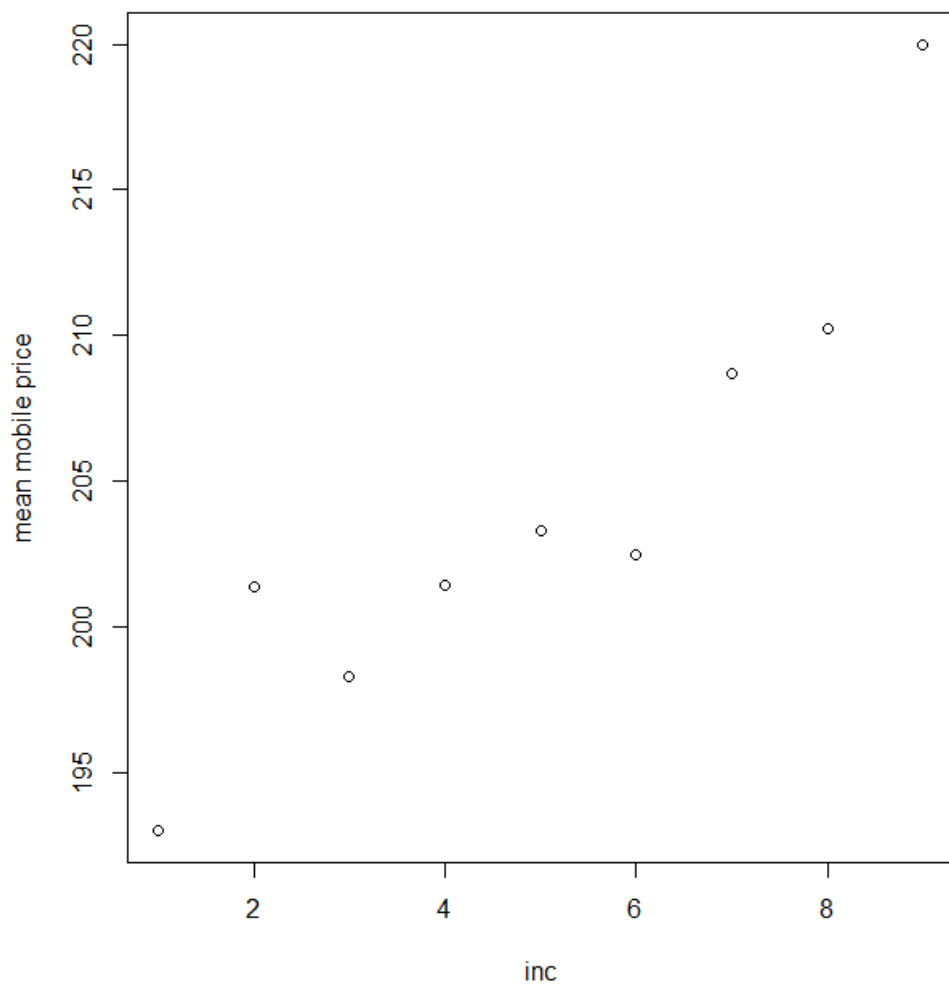
Διαγράμματα Σχέσεων

Στα παρακάτω διαγράμματα φαίνονται κάποιοι βασικοί παράγοντες συναρτήσει της μέσης τιμής αγοράς του κινητού τηλεφώνου. Καταλαβαίνουμε ότι υπάρχει αρνητική συσχέτιση της τιμής με την ηλικία, καθώς όσο αυξάνεται η ηλικία του αγοραστή τόσο το διάγραμμα παρουσιάζει πτωτική τάση. Αντίθετα στις μεταβλητές του εισοδήματος, της μόρφωσης και της εργασίας του αγοραστή παρουσιάζεται θετική συσχέτιση.







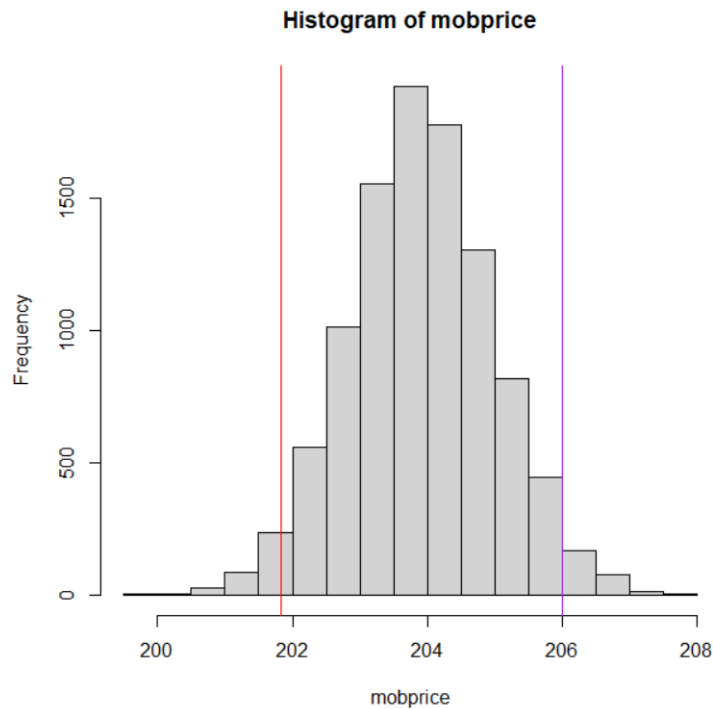


Bootstrap

Σε αυτό το σημείο, εφαρμόσαμε την Bootstrap για τον υπολογισμό των διαστημάτων εμπιστοσύνης χωρίς να υποθέσουμε κανονική κατανομή. Με την εντολή **quantile(b,c(0.025,0.975))** βρίσκουμε το διάστημα εμπιστοσύνης 95%. Το Bootstrap 95% δ.ε. για τη μέση τιμή της αξίας του κινητού από όπου προέρχεται το δείγμα είναι [201.83, 206.01].

Listing 4: R output

```
> # Bootstrap for mobile prices
> mobprice <- numeric(10000)
> for (i in 1:10000) {
+   mobprice[i] <- mean(sample(mobileprice,length(mobileprice),replace=T))}
> hist(mobprice)
> quantile(mobprice,c(0.025,0.975))
   2.5%    97.5%
201.8303 206.0072
> abline(v=quantile(mobprice, 0.025),col="red")
> abline(v=quantile(mobprice,0.975),col="purple")
> plot(tapply(mobileprice, age, mean), xlab = "age", ylab = "mean mobile price")
```



Σχήμα 10: Histogram της μεταβλητής mobileprice - Διάστημα Εμπιστοσύνης

Πίνακας Συσχετίσεων Spearman

Πρόκειται για ένα συμμετρικό πίνακα με τις συσχετίσεις Spearman όλων των ζευγαριών από το υποσύνολο των μεταβλητών που επιλέξαμε στο main_df. Χρησιμοποιούμε την μέθοδο Spearman, καθώς μπορεί να εφαρμοστεί σε κάθε είδους ποσοτικές μεταβλητές αλλά και σε μεταβλητές διάταξης.

Listing 5: R output

```
> cor(main_df, method="spearman")
      psraid      usr_r      sex
age      mar
psraid    1.000000000 -0.08324133 -1.167061e-01 -0.28604744
0.12142879
usr_r    -0.083241326  1.000000000  4.009169e-02
0.07198940 -0.08526036
sex    -0.116706124  0.04009169  1.000000e+00  0.06389970
0.01939229
age    -0.286047444  0.07198940  6.389970e-02
1.000000000 -0.20547266
mar     0.121428789 -0.08526036  1.939229e-02 -0.20547266
1.000000000
educ     0.009378352 -0.09130135  3.083057e-05 -0.01661946 -0.14445439
empl    -0.107718254  0.02251194  1.555397e-01  0.22486420
0.11892002
inc    -0.008177675 -0.06642817 -1.214430e-01 -0.03786753 -0.38022114
mobileprice 0.651051715 -0.11168451 -1.194202e-01 -0.38176631
0.08687417
      educ      empl      inc mobileprice
psraid    9.378352e-03 -0.10771825 -0.008177675  0.65105172
usr_r    -9.130135e-02  0.02251194 -0.066428172 -0.11168451
sex      3.083057e-05  0.15553969 -0.121442950 -0.11942018
age     -1.661946e-02  0.22486420 -0.037867530 -0.38176631
mar     -1.444544e-01  0.11892002 -0.380221136  0.08687417
educ     1.000000e+00 -0.20111406  0.451226869  0.10921151
empl    -2.011141e-01  1.000000000 -0.337288817 -0.17370103
inc      4.512269e-01 -0.33728882  1.000000000  0.14945270
mobileprice 1.092115e-01 -0.17370103  0.149452700  1.000000000
```

Έλεγχος για Στατιστικά Σημαντική Συσχέτιση Spearman

Φαίνεται αρκετά αρνητική συσχέτιση καθώς ο συντελεστής είναι -0.4. Επίσης, η τιμή της p-value είναι αρκετά μικρή, (≤ 0.05) άρα η συσχέτιση της τιμής με την μεταβλητή της ηλικίας είναι στατιστικά σημαντική.

Listing 6: R output

```
> cor.test(mobileprice, age, method = "spearman")

Spearman's rank correlation rho

data:  mobileprice and age
S = 1274872374, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.3817663

Warning message:
In cor.test.default(mobileprice, age, method = "spearman") :
  Cannot compute exact p-value with ties
```

Έλεγχος για Στατιστικά Σημαντική Συσχέτιση Kendall

Χρησιμοποιούμε το συντελεστή συσχέτισης Kendall λόγω της εμφάνισης επαναλαμβανόμενων τιμών.

- Στην πρώτη συσχέτιση ανάμεσα στην αξία του τηλεφώνου και το εισόδημα από το αποτέλεσμα της μεθόδου Kendall μπορούμε να αντιληφθούμε ότι η συσχέτιση είναι στατιστικά σημαντική αφού η τιμή της p-value είναι πολύ μικρότερη από το 0.05. Η τιμή του συντελεστή είναι θετική και αρκετά ασθενής ($\tau = 0.1$).
- Στην δεύτερη συσχέτιση ανάμεσα στην αξία του τηλεφώνου και την περιοχή κατοικίας από το αποτέλεσμα της μεθόδου Kendall μπορούμε να αντιληφθούμε ότι η συσχέτιση είναι στατιστικά σημαντική αφού η τιμή της p-value είναι πολύ μικρότερη από το 0.05. Η τιμή του συντελεστή είναι αρνητική και αρκετά ασθενής ($\tau = - 0.08$).
- Στην τρίτη συσχέτιση ανάμεσα στην αξία του τηλεφώνου και το φύλο του αγοραστή από το αποτέλεσμα της μεθόδου Kendall μπορούμε να αντιληφθούμε ότι η συσχέτιση είναι στατιστικά σημαντική αφού η τιμή της p-value είναι πολύ μικρότερη από το 0.05. Η τιμή του συντελεστή είναι αρνητική και αρκετά ασθενής ($\tau = - 0.097$).
- Στην τέταρτη συσχέτιση ανάμεσα στην αξία του τηλεφώνου και την εκπαίδευση του αγοραστή από το αποτέλεσμα της μεθόδου Kendall μπορούμε να αντιληφθούμε ότι η συσχέτιση είναι στατιστικά σημαντική αφού η τιμή της p-value είναι πολύ μικρότερη από το 0.05. Η τιμή του συντελεστή είναι θετική και αρκετά ασθενής ($\tau = - 0.08$).
- Στην πέμπτη συσχέτιση ανάμεσα στην αξία του τηλεφώνου και την επαγγελματική κατάσταση του αγοραστή από το αποτέλεσμα της μεθόδου Kendall μπορούμε να αντιληφθούμε ότι η συσχέτιση είναι στατιστικά σημαντική αφού η τιμή της p-value είναι πολύ μικρότερη από το 0.05. Η τιμή του συντελεστή είναι αρνητική και αρκετά ασθενής ($\tau = - 0.131$).

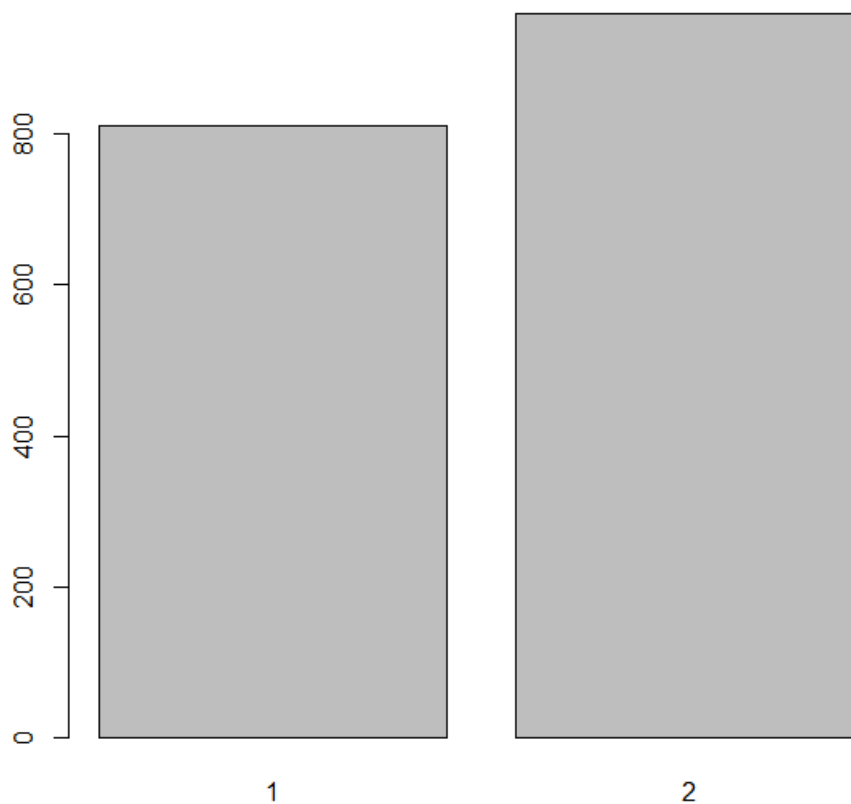
Επομένως όλες οι μεταβλητές του υποσυνόλου είναι στατιστικά σημαντικές κάτι που μπορούμε να κατανοήσουμε καθώς τόσο το εισόδημα, η περιοχή κατοικίας, το φύλο, η επαγγελματική κατάσταση και η εκπαίδευση μπορούν να συμβάλλουν στην στην οικονομική ευημερία ή όχι του αγοραστή και να δικαιολογήσουν το ποσό που διαθέτει ο καθένας για την αγορά ενός κινητού τηλεφώνου.

Ερώτημα: Επηρεάζει το φύλο την τιμή αγοράς ενός κινητού τηλεφώνου.

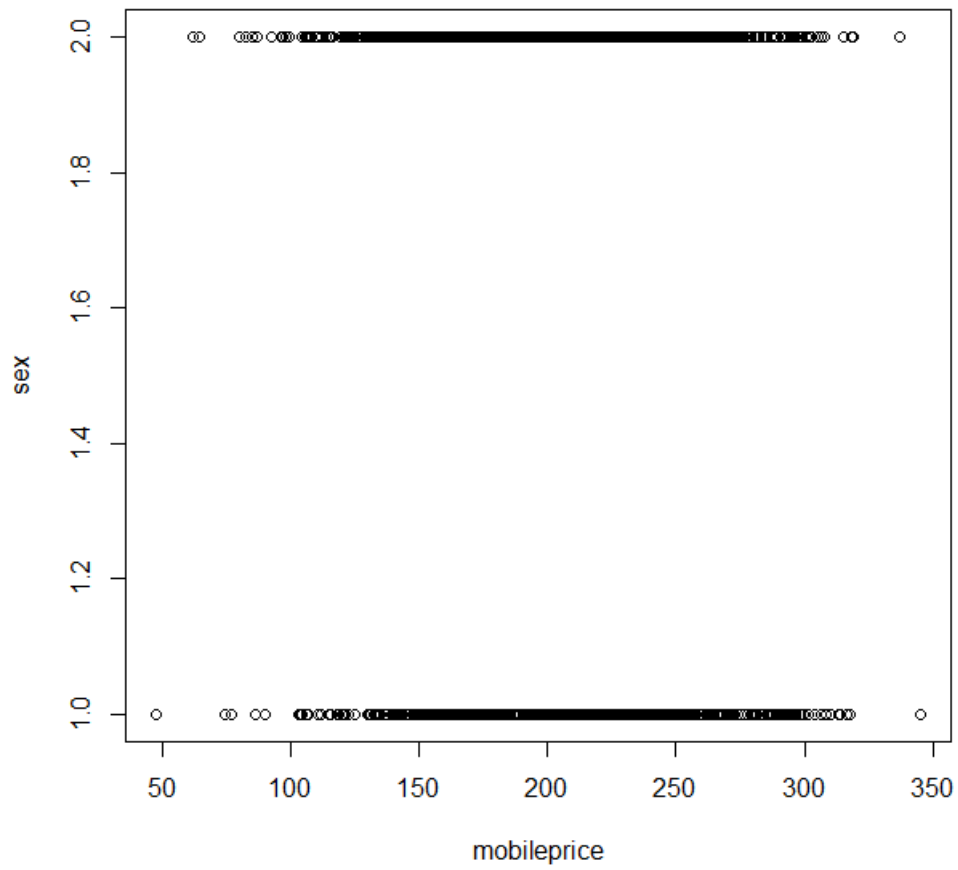
Μεταβλητή Απόκρισης: αριθμός ανδρών και γυναικών

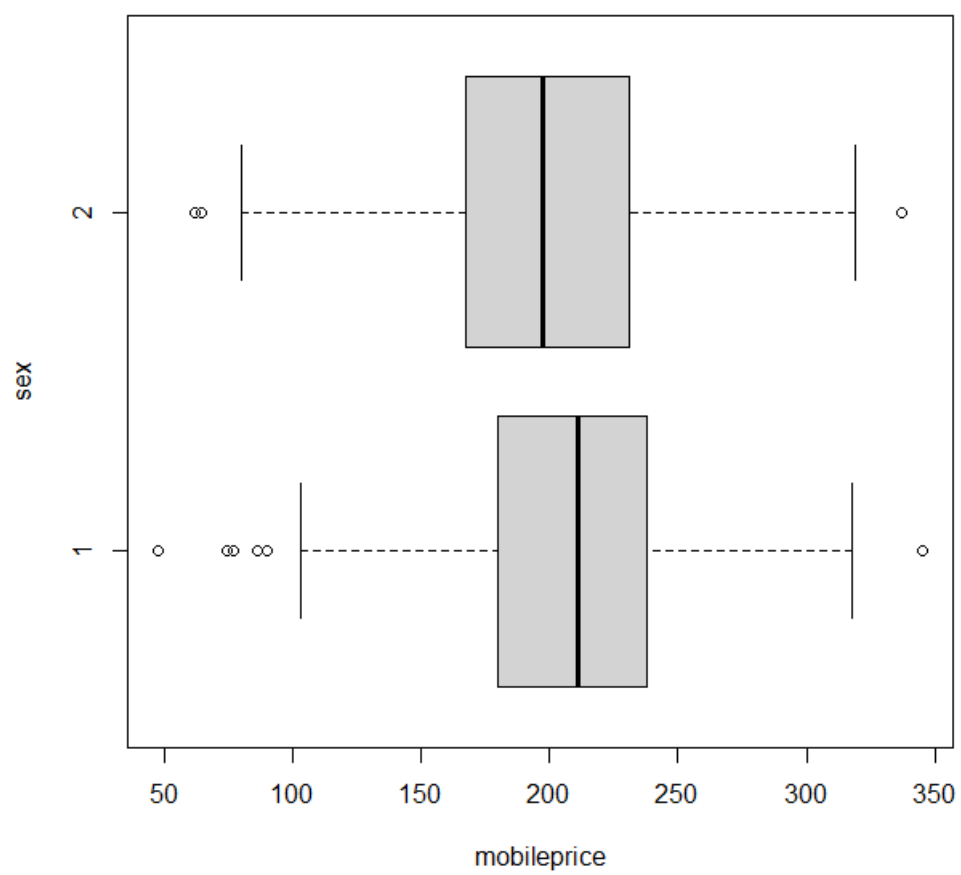
```
> table(sex)
sex
 1    2
810 959
```

Ερμηνευτική μεταβλητή ανεξάρτητη: τιμή κινητού τηλεφώνου



Σχέση της μεταβλητής απόκρισης με την ερμηνευτική





Ανάλυση Διακύμανσης - Διασποράς

Αρχικά τρέχουμε Anova για τις μεταβλητές που έχουμε ξεχωρίσει. Όπως παρατηρούμε από τα αποτελέσματα το κάθε υπόδειγμα ξεχωριστά συνεισφέρει σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής. Ο λόγος F είναι ο λόγος διακυμάνσεων και όσο μεγαλύτερος είναι (1), τόσο το υπόδειγμα εξηγεί μεγάλο ποσοστό της συνολικής διακύμανσης. Έτσι για τις μεταβλητές που έχουμε επιλέξει βγάζουμε το συμπέρασμα ότι συνεισφέρουν σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής.

```
> one.way <- aov(mobileprice~educ, data = main_df)
> summary(one.way)
              Df Sum Sq Mean Sq F value    Pr(>F)
educ             1   35773    35773   18.08 2.23e-05 ***
Residuals      1767 3496687     1979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> one.way <- aov(mobileprice~inc, data = main_df)
> summary(one.way)
              Df Sum Sq Mean Sq F value    Pr(>F)
inc             1   58029    58029   29.51 6.33e-08 ***
Residuals      1767 3474430     1966
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> one.way <- aov(mobileprice~empl, data = main_df)
> summary(one.way)
              Df Sum Sq Mean Sq F value    Pr(>F)
empl             1   64591    64591   32.91 1.13e-08 ***
Residuals      1767 3467869     1963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> one.way <- aov(mobileprice~mar, data = main_df)
> summary(one.way)
              Df Sum Sq Mean Sq F value    Pr(>F)
mar             1   15790    15790    7.934 0.00491 **
Residuals      1767 3516669     1990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> one.way <- aov(mobileprice~age, data = main_df)
> summary(one.way)
              Df Sum Sq Mean Sq F value    Pr(>F)
age             1  453233   453233  260.1 <2e-16 ***
Residuals      1767 3079226     1743
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


ANOVA

Θέλουμε να εξετάσουμε αν οι επιδράσεις των 5 παραγόντων και η αλληλεπίδραση τους στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντικές.

```
> model <-aov(mobileprice~sex * age * inc * mar * usr_r)
> summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	43984	43984	25.984	3.82e-07	***
age	1	435339	435339	257.181	< 2e-16	***
inc	1	39145	39145	23.125	1.65e-06	***
mar	1	6100	6100	3.604	0.05782	.
usr_r	1	15808	15808	9.339	0.00228	**
sex:age	1	245	245	0.144	0.70391	
sex:inc	1	956	956	0.565	0.45237	
age:inc	1	194	194	0.115	0.73487	
sex:mar	1	9565	9565	5.650	0.01756	*
age:mar	1	446	446	0.264	0.60776	
inc:mar	1	662	662	0.391	0.53193	
sex:usr_r	1	3495	3495	2.064	0.15095	
age:usr_r	1	1628	1628	0.962	0.32683	
inc:usr_r	1	2332	2332	1.377	0.24071	
mar:usr_r	1	1021	1021	0.603	0.43737	
sex:age:inc	1	28	28	0.017	0.89693	
sex:age:mar	1	2	2	0.001	0.97102	
sex:inc:mar	1	2089	2089	1.234	0.26682	
age:inc:mar	1	8875	8875	5.243	0.02215	*
sex:age:usr_r	1	404	404	0.238	0.62538	
sex:inc:usr_r	1	1627	1627	0.961	0.32698	
age:inc:usr_r	1	2059	2059	1.216	0.27025	
sex:mar:usr_r	1	2216	2216	1.309	0.25268	
age:mar:usr_r	1	1513	1513	0.894	0.34465	
inc:mar:usr_r	1	3	3	0.002	0.96836	
sex:age:inc:mar	1	3684	3684	2.177	0.14031	
sex:age:inc:usr_r	1	7070	7070	4.176	0.04114	*
sex:age:mar:usr_r	1	808	808	0.478	0.48960	
sex:inc:mar:usr_r	1	751	751	0.444	0.50549	
age:inc:mar:usr_r	1	3	3	0.002	0.96771	
sex:age:inc:mar:usr_r	1	132	132	0.078	0.78023	
Residuals	1737	2940277	1693			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Όπως φαίνεται και από την πάνω φωτογραφία οι 4 από τους 5 παράγοντες είναι στατιστικά σημαντικοί καθώς το p-value είναι μικρότερο από το 0.05. Συγκεκριμένα, οι μεταβλητές του φύλου, της ηλικίας, του εισοδήματος και της περιοχής διαμονής είναι στατιστικά σημαντικές ενώ η μεταβλητή της οικογενειακής κατάστασης δεν είναι στατιστικά σημαντικές καθώς το p-value είναι

λίγο μεγαλύτερο από το 0.05. Από τις αλληλεπιδράσεις οι περισσότερες δεν είναι στατιστικά σημαντικές εκτός από τις αλληλεπιδράσεις του φύλου και της οικογενειακής κατάστασης, της ηλικίας του εισοδήματος και της οικογενειακής κατάστασης και του φύλου της ηλικίας του εισοδήματος και της περιοχής διαμονής.

Αναλυτικότερα Αποτελέσματα ANOVA

```
> summary.lm(model)

Call:
aov(formula = mobileprice ~ sex * age * inc * mar * usr_r)

Residuals:
    Min       1Q   Median       3Q      Max
-176.313  -24.719   -0.173   24.457  142.083

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    314.777012  123.932074   2.540  0.0112 *
sex             -34.667670   77.010874  -0.450  0.6526
age             -2.511952    2.301983  -1.091  0.2753
inc             -6.211395    22.620806  -0.275  0.7837
mar             18.230210    27.506852   0.663  0.5076
usr_r          -52.612728    59.930007  -0.878  0.3801
sex:age          0.756624    1.454517   0.520  0.6030
sex:inc          1.787908    14.522499   0.123  0.9020
age:inc          0.267364    0.425066   0.629  0.5294
sex:mar        -12.509227    17.098276  -0.732  0.4645
age:mar         -0.172160    0.583043  -0.295  0.7678
inc:mar         -4.479912    5.332024  -0.840  0.4009
sex:usr_r       30.020786    37.064725   0.810  0.4181
age:usr_r        1.145454    1.115563   1.027  0.3047
inc:usr_r        8.724344    11.202050   0.779  0.4362
mar:usr_r       -5.848108    13.347254  -0.438  0.6613
sex:age:inc     -0.083194    0.276520  -0.301  0.7636
sex:age:mar      0.123611    0.355029   0.348  0.7278
sex:inc:mar      3.753768    3.450105   1.088  0.2767
age:inc:mar      0.054363    0.115935   0.469  0.6392
sex:age:usr_r   -0.697770    0.693937  -1.006  0.3148
sex:inc:usr_r   -5.099318    7.159375  -0.712  0.4764
age:inc:usr_r   -0.203603    0.209380  -0.972  0.3310
sex:mar:usr_r    3.126315    8.269272   0.378  0.7054
age:mar:usr_r    0.025345    0.278785   0.091  0.9276
inc:mar:usr_r    1.205578    2.620502   0.460  0.6455
sex:age:inc:mar -0.052607    0.072869  -0.722  0.4704
sex:age:inc:usr_r 0.114913    0.134735   0.853  0.3938
sex:age:mar:usr_r 0.008153    0.168862   0.048  0.9615
sex:inc:mar:usr_r -0.880877    1.704491  -0.517  0.6054
age:inc:mar:usr_r -0.014401    0.057115  -0.252  0.8010
sex:age:inc:mar:usr_r 0.010033    0.035953   0.279  0.7802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.14 on 1737 degrees of freedom
Multiple R-squared:  0.1676,    Adjusted R-squared:  0.1528
F-statistic: 11.29 on 31 and 1737 DF,  p-value: < 2.2e-16
```

Το μοντέλο έχει χαμηλή ερμηνευτική ισχύ, από το r-squared προκύπτει ότι εξηγεί το 16.76% της μεταβλητότητας της εξαρτημένης. Γενικά δε θέλουμε μεγάλα κατάλοιπα σε απόλυτη τιμή (δείγμα κακής προσαρμογής), ωστόσο στο υπόδειγμα μας έχουμε. Οι περισσότεροι συντελεστές είναι στατιστικά μη σημαντικοί αφού η τιμή της p-value είναι μεγαλύτερη από 0.05 και δεν συνεισφέρουν στην ερμηνεία της εξαρτημένης μεταβλητής. Ωστόσο, η τιμή της p-value του υποδείγματος είναι πολύ μικρότερη από το 0.05 επομένως το υπόδειγμα συνεισφέρει σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής. Στο καινούριο μοντέλο όλες οι μεταβλητές είναι στατιστικά σημαντικές εκτός από την μεταβλητή της οικογενειακής κατάστασης καθώς η τιμή της p-value είναι μεγαλύτερη από 0.05.

```
> model2 <- aov(mobileprice ~ sex + age + inc + mar + usr_r)
> summary(model2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	43984	43984	25.916	3.95e-07	***
age	1	435339	435339	256.511	< 2e-16	***
inc	1	39145	39145	23.065	1.70e-06	***
mar	1	6100	6100	3.594	0.05815	.
usr_r	1	15808	15808	9.314	0.00231	**
Residuals	1763	2992083	1697			

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.lm(model2)
```

Call:

```
aov(formula = mobileprice ~ sex + age + inc + mar + usr_r)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-179.575	-24.741	-0.616	24.243	140.037

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	250.70377	6.29495	39.826	< 2e-16	***
sex	-6.41566	1.98542	-3.231	0.00125	**
age	-0.84205	0.05658	-14.883	< 2e-16	***
inc	2.20924	0.45385	4.868	1.23e-06	***
mar	0.83822	0.52039	1.611	0.10741	
usr_r	-4.33910	1.42174	-3.052	0.00231	**

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.2 on 1763 degrees of freedom
Multiple R-squared:  0.153,    Adjusted R-squared:  0.1506
F-statistic: 63.68 on 5 and 1763 DF,  p-value: < 2.2e-16
```

Δημιουργούμε έτσι και το τρίτο μοντέλο όπου όλες οι μεταβλητές είναι στατιστικά σημαντικά και όλες συνεισφέρουν στην ερμηνεία της εξαρτημένης μεταβλητής. Η τιμή της p-value του υποδείγματος είναι πολύ μικρότερη από το 0.05 επομένως το υπόδειγμα συνεισφέρει σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής.

```
> model2 <-aov(mobileprice~sex + age + inc + mar + usr_r)
> summary(model2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	43984	43984	25.916	3.95e-07 ***
age	1	435339	435339	256.511	< 2e-16 ***
inc	1	39145	39145	23.065	1.70e-06 ***
mar	1	6100	6100	3.594	0.05815 .
usr_r	1	15808	15808	9.314	0.00231 **
Residuals	1763	2992083	1697		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.lm(model2)
```

Call:
aov(formula = mobileprice ~ sex + age + inc + mar + usr_r)

Residuals:

Min	1Q	Median	3Q	Max
-179.575	-24.741	-0.616	24.243	140.037

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	250.70377	6.29495	39.826	< 2e-16 ***
sex	-6.41566	1.98542	-3.231	0.00125 **
age	-0.84205	0.05658	-14.883	< 2e-16 ***
inc	2.20924	0.45385	4.868	1.23e-06 ***
mar	0.83822	0.52039	1.611	0.10741
usr_r	-4.33910	1.42174	-3.052	0.00231 **

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.2 on 1763 degrees of freedom
Multiple R-squared:  0.153,    Adjusted R-squared:  0.1506
F-statistic: 63.68 on 5 and 1763 DF,  p-value: < 2.2e-16
```

Post hoc Tests

Με την παράμετρο `method` καθορίζουμε ποιο test χρησιμοποιήσαμε. Στην περίπτωση μας επιλέξαμε το Tukey HSD (Honestly Significant Difference). Στατιστικά σημαντική διαφορά με `p-value` μικρότερο από το 0.05 μεταξύ των μεταβλητών `mobileprice` και `educ`. Σε όλα τα παρακάτω test η μόνη συσχέτιση που βρέθηκε ανάμεσα στα επίπεδα είναι στο εισόδημα.

```
> PostHocTest(aov(mobileprice~factor(sex)), method="hsd")
```

```
Posthoc multiple comparisons of means : Tukey HSD  
95% family-wise confidence level
```

```
$`factor(sex)`  
      diff      lwr.ci  upr.ci    pval  
2-1 -10.0083 -14.16699 -5.8496 2.5e-06 ***
```

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> PostHocTest(aov(mobileprice~factor(educ)),method="hsd")
```

```
Posthoc multiple comparisons of means : Tukey HSD
95% family-wise confidence level
```

```
$`factor(educ)`
      diff      lwr.ci    upr.ci    pval
2-1  4.6113695 -23.1920800  32.41482  0.9990
3-1  7.4463103 -18.4728484  33.36547  0.9797
4-1 -0.1587302 -31.6463616  31.32890  1.0000
5-1 14.0133644 -12.0532731  40.08000  0.6909
6-1 15.9426268 -10.2256213  42.11087  0.5493
7-1 16.9874140  -9.7634171  43.73824  0.4977
3-2  2.8349408 -10.0630789  15.73296  0.9951
4-2 -4.7700997 -26.8161211  17.27592  0.9955
5-2  9.4019949  -3.7898859  22.59388  0.3508
6-2 11.3312573  -2.0602825  24.72280  0.1606
7-2 12.3760445  -2.1209583  26.87305  0.1527
4-3 -7.6050404 -27.2213102  12.01123  0.9142
5-3  6.5670541  -1.9650440  15.09915  0.2583
6-3  8.4963165  -0.3413482  17.33398  0.0688 .
7-3  9.5411037  -0.8960448  19.97825  0.0992 .
5-4 14.1720946  -5.6386310  33.98282  0.3461
6-4 16.1013570  -3.8428769  36.04559  0.2063
7-4 17.1461441  -3.5565662  37.84885  0.1807
6-5  1.9292624  -7.3320077  11.19053  0.9964
7-5  2.9740495  -7.8241389  13.77224  0.9837
7-6  1.0447872  -9.9964308  12.08601  1.0000
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

> PostHocTest(aov(mobileprice~factor(inc)),method="hsd")

Posthoc multiple comparisons of means : Tukey HSD
95% family-wise confidence level

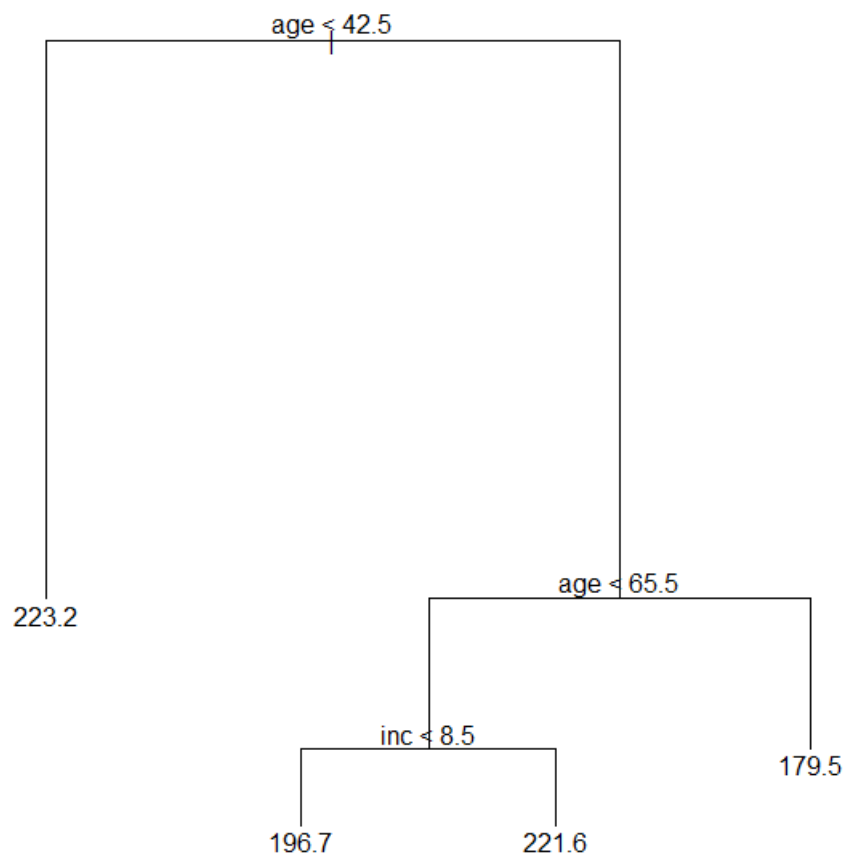
$`factor(inc)`
      diff      lwr.ci      upr.ci      pval
2-1  8.37679678 -7.1752937 23.92889 0.76327
3-1  5.26449932 -9.5712577 20.10026 0.97403
4-1  8.44260919 -6.6994532 23.58467 0.72695
5-1 10.28411666 -5.0966526 25.66489 0.48985
6-1  9.44873946 -4.4422534 23.33973 0.46498
7-1 15.68054114  0.9604125 30.40067 0.02671 *
8-1 17.25851107  1.7064206 32.81060 0.001690 *
9-1 26.99414783 10.2459548 43.74234 2.2e-05 ***
3-2 -3.11229746 -17.0737103 10.84912 0.99889
4-2  0.06581241 -14.2206643 14.35229 1.00000
5-2  1.90731988 -12.6319175 16.44656 0.99998
6-2  1.07194268 -11.8810896 14.02497 1.00000
7-2  7.30374436 -6.5347364 21.14223 0.78308
8-2  8.88171429 -5.8386417 23.60207 0.63205
9-2 18.61735105  2.6385109 34.59619 0.00925 **
4-3  3.17810987 -10.3250595 16.68128 0.99836
5-3  5.01961733 -8.7506983 18.78993 0.96938
6-3  4.18424013 -7.8993526 16.26783 0.97769
7-3 10.41604182 -2.6122243 23.44431 0.24090
8-3 11.99401174 -1.9674011 25.95542 0.16011
9-3 21.72964851  6.4471258 37.01217 0.00036 ***
5-4  1.84150747 -12.2582783 15.94129 0.99998
6-4  1.00613027 -11.4516216 13.46388 1.00000
7-4  7.23793195 -6.1380946 20.61396 0.75863
8-4  8.81590188 -5.4705748 23.10238 0.60237
9-4 18.55153864  2.9714916 34.13159 0.00692 **
6-5 -0.83537720 -13.5822038 11.91145 1.00000
7-5  5.39642449 -8.2492376 19.04209 0.95036
8-5  6.97439441 -7.5648430 21.51363 0.86057
9-5 16.71003118  0.8978879 32.52217 0.02906 *
7-6  6.23180168 -5.7095431 18.17315 0.79355
8-6  7.80977161 -5.1432607 20.76280 0.63298
9-6 17.54540837  3.1782410 31.91258 0.00485 **
8-7  1.57796992 -12.2605109 15.41645 0.99999
9-7 11.31360669 -3.8566933 26.48391 0.33278
9-8  9.73563677 -6.2432034 25.71448 0.61928

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```


Δενδρόγραμμα

Σύμφωνα με το παρακάτω δενδρόγραμμα ο πιο σημαντικός παράγοντας φαίνεται να είναι η ηλικία. Για τους αγοραστές με ηλικία κάτω των 42.5, η μέση τιμή αγοράς ενός κινητού είναι 223.2. Για τους αγοραστές ανάμεσα στις ηλικίες 42.5 και 65.5 ο δεύτερος σημαντικότερος παράγοντας είναι το εισόδημα.



Μοντέλα

```
> model<-glm(as.numeric(sex)~mobileprice, quasipoisson)
> model<-lm(as.numeric(sex)~mobileprice)
> summary(model)
```

Call:

```
lm(formula = as.numeric(sex) ~ mobileprice)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.7363	-0.5277	0.3744	0.4548	0.6237

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.7957777	0.0550170	32.64	< 2e-16 ***
mobileprice	-0.0012441	0.0002636	-4.72	2.54e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4954 on 1767 degrees of freedom

Multiple R-squared: 0.01245, Adjusted R-squared: 0.01189

F-statistic: 22.28 on 1 and 1767 DF, p-value: 2.543e-06

Χρησιμοποιούμε την quasipoisson. Ο όρος που μας ενδιαφέρει η μόρφωση δηλαδή του αγοραστή είναι στατιστικά σημαντική, δηλαδή φαίνεται η σημαντική επίδραση της ανεξάρτητης μεταβλητής.

```
> model<-glm(mobileprice~educ, quasipoisson)
> summary(model)
```

Call:

```
glm(formula = mobileprice ~ educ, family = quasipoisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-12.7611	-2.1680	0.1005	2.0567	8.9113

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.255555	0.015541	338.173	< 2e-16 ***
educ	0.013541	0.003185	4.251	2.24e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 9.707628)

Null deviance: 17857 on 1768 degrees of freedom
Residual deviance: 17681 on 1767 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

Ο όρος που μας ενδιαφέρει το εισόδημα δηλαδή του αγοραστή είναι στατιστικά σημαντική, δηλαδή φαίνεται η σημαντική επίδραση της ανεξάρτητης μεταβλητής.

```
> model<-glm(mobileprice~inc)
> summary(model)

Call:
glm(formula = mobileprice ~ inc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-151.312   -29.467    1.668    28.163   146.398

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  191.5966     2.4967  76.739  < 2e-16 ***
inc           2.4350     0.4482   5.432 6.33e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1966.288)

    Null deviance: 3532459  on 1768  degrees of freedom
Residual deviance: 3474430  on 1767  degrees of freedom
AIC: 18440

Number of Fisher Scoring iterations: 2
```

Ο όρος που μας ενδιαφέρει η οικογενειακή κατάσταση δηλαδή του αγοραστή είναι στατιστικά σημαντική, δηλαδή φαίνεται η σημαντική επίδραση της ανεξάρτητης. μεταβλητής.

```
> model<-glm(mobileprice~mar)
> summary(model)
```

Call:

```
glm(formula = mobileprice ~ mar)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-160.903	-31.013	0.649	30.935	136.597

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	200.0751	1.7207	116.275	< 2e-16 ***
mar	1.4380	0.5105	2.817	0.00491 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1990.192)

Null deviance: 3532459 on 1768 degrees of freedom
Residual deviance: 3516669 on 1767 degrees of freedom
AIC: 18462

Number of Fisher Scoring iterations: 2

.

Ο όρος που μας ενδιαφέρει η εκπαίδευση δηλαδή του αγοραστή είναι στατιστικά σημαντική, δηλαδή φαίνεται η σημαντική επίδραση της ανεξάρτητης.

```
> model<-glm(mobileprice~educ)
> summary(model)

Call:
glm(formula = mobileprice ~ educ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-149.026   -30.126    1.474    30.217   140.204

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  191.3125     3.1419  60.890  < 2e-16 ***
educ          2.7567     0.6484   4.252 2.23e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1978.883)

    Null deviance: 3532459  on 1768  degrees of freedom
Residual deviance: 3496687  on 1767  degrees of freedom
AIC: 18451

Number of Fisher Scoring iterations: 2
```

Ο όρος που μας ενδιαφέρει το φύλο δηλαδή του αγοραστή είναι στατιστικά σημαντική, δηλαδή φαίνεται η σημαντική επίδραση της ανεξάρτητης.

```
> model<-glm(mobileprice~sex)
> summary(model)

Call:
glm(formula = mobileprice ~ sex)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-161.517   -30.409    -0.509    29.891   137.891

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   219.325      3.436   63.83  < 2e-16 ***
sex           -10.008      2.120   -4.72 2.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1974.236)

    Null deviance: 3532459  on 1768  degrees of freedom
Residual deviance: 3488475  on 1767  degrees of freedom
AIC: 18447

Number of Fisher Scoring iterations: 2
```