# Στατιστική Ανάλυση Δεδομένων

*Ανάλυση του "Cell Phone Dataset"*

```
> setwd("C:\\Users\\giorg\\Desktop\\Statistiki ergasia")
> getwd()
# Read to data frame
> cell.phones <- read.csv("Cell_Phones_labels.csv", header = T, sep = ';',
                          stringsAsFactors = T, dec = ",", na.strings = " ")
> head(cell.phones)
> str(cell.phones)
```

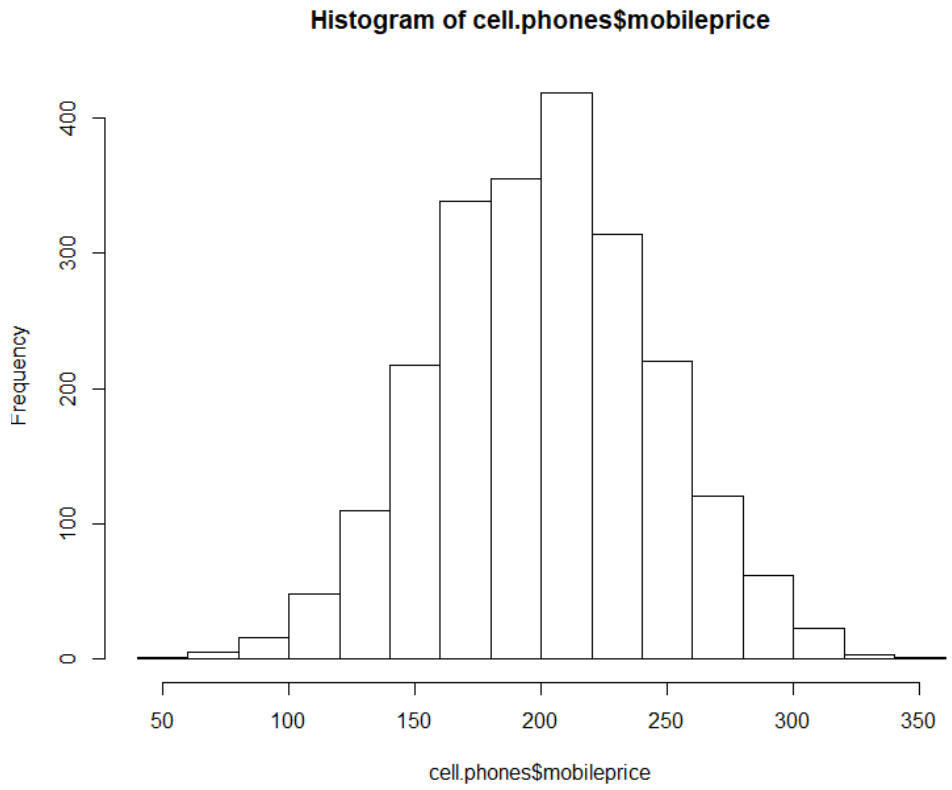
---

Ξεκινώντας την ανάλυση, τρέξαμε ένα απλό summary για να ελέγξουμε κάποια απλά χαρακτηριστικά των δεδομένων μας. Από το πιο σημαντικό, δηλαδή τη τιμή αγοράς μιας νέας συσκευής παρατηρούμε τα παρακάτω.

```
> summary(cell.phones$mobileprice)
```

| *Min.* | *1st Qu.* | *Median* | *Mean* | *3rd Qu.* | *Max.* |
|---|---|---|---|---|---|
| *47.8* | *170.2* | *201.4* | *201.3* | *231.2* | *345.3* |

Παρατηρούμε ότι καθώς η μέση τιμή που ξοδεύει κανείς ανέρχεται στα 201€ . Οι περισσότεροι αγοραστές επέλεξαν συσκευές με τιμή από 170€ εως 230€ ενώ φυσικά υπήρξαν και άτομα που επέλεξαν συσκευή με χαμηλό κόστος στα 48€ ή και πολύ υψηλό στα 345€. Σε κάθε περίπτωση με την εντολή summary βλέπουμε ότι πιθανότατα πρόκειται για κανονική κατανομή στο κόστος των κινητών συσκευών και αυτό μπορούμε να το επαληθεύσουμε στην συνέχεια.



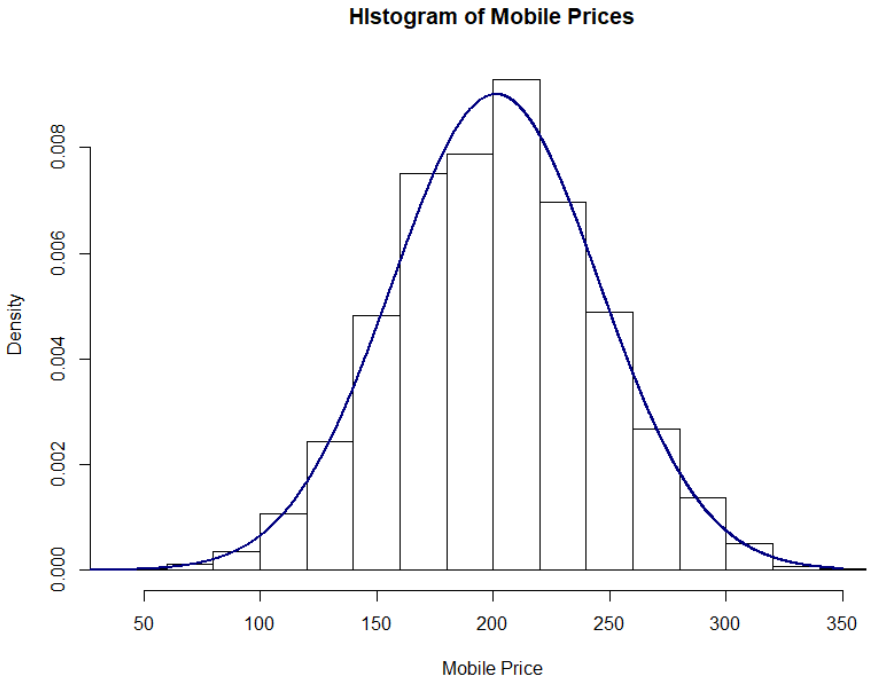

Ένα απλό ιστόγραμμα μας δείχνει καλύτερα οπτικά τις υποθέσεις που κάναμε παραπάνω. Παρατηρούμε λοιπόν ότι όντως πολλά άτομα έχουν αγοράσει τις συσκευές τους σε τιμές κοντά στο κέντρο. Το ιστρόγραμμα μας δείχνει ότι οι τιμές των κινητών τηλεφώνων παρουσιάζουν κανονική κατανομή καθώς μπορούμε να δούμε την γνωστή "καμπάνα" που αποτελεί χαρακτηριστικό της. Αυτό το επαληθεύουμε και στην πορεία με τις εντολές για έλεγχο κανονικότητας.

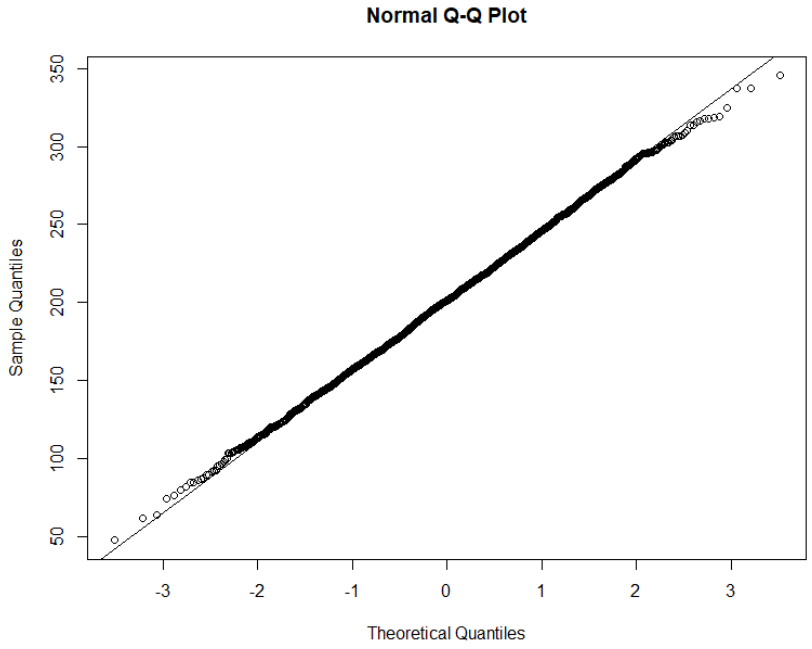
## ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ ΤΩΝ ΤΙΜΩΝ ΑΓΟΡΑΣ ΣΥΣΚΕΥΗΣ

```
> summary(cell.phones$mobileprice)

> hist(cell.phones$mobileprice)

# Check if mobile price is normally distributed

> hist(cell.phones$mobileprice, freq = F, main = "HIstogram of Mobile Prices",

    xlab = "Mobile Price")

# Draw normal dist over histogram

> lines(seq(0,350,.1), dnorm(seq(0,350,.1), mean(cell.phones$mobileprice),

                sd(cell.phones$mobileprice)), col = "navy", lwd = 2)
```

# QQ-plot

```
> qqnorm(cell.phones$mobileprice)
> qqline(cell.phones$mobileprice, lwd = 2, col = "blue")
```



Εκτελώντας τις παραπάνω εντολές παρατηρούμε πως η κατανομή των τιμών των συσκευών είναι κανονική κατανομή. Βλέπουμε ότι υπάρχουν ορισμένα σημεία τα οποία δεν είναι ακριβώς πάνω στην ευθεία ώστοσο η ύπαρξη αυτών δεν μπορεί να μας αποτρέψει να πούμε ότι δεν πρόκειται για κανονική κατανομή.

**Shapiro test για την επαληθευση και με αυτην την εντολή της κανονικότητας .**

```
> shapiro.test(cell.phones$mobileprice)
```

*Shapiro-Wilk normality test*

*data: cell.phones$mobileprice*

***W = 0.99957, p-value = 0.928***

Το μεγάλο p-value που είναι κοντά στο ένα δεν μας επιτρέπει να απορρίψουμε τη μηδενική υπόθεση και καταδεικνύει για ακόμη μία φορά πως έχουμε κανονική κατανομή στις τιμές της μεταβλητής Τιμή Τηλεφώνου (mobile price).

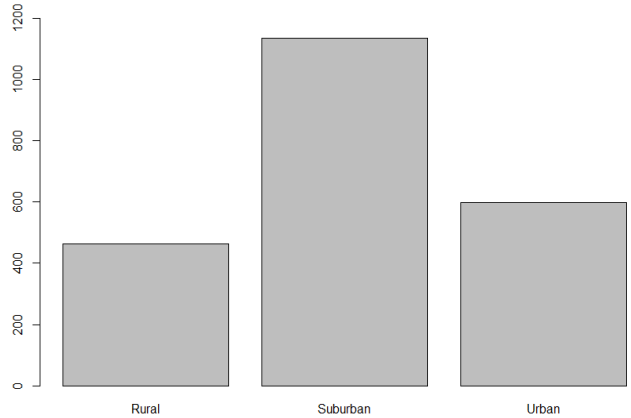
Στην συνέχεια με διαγράμματα διερευνήσαμε με έναν οπτικό τρόπο κάποιες απο τις μεταβλητές

Για λόγους καλύτερης κατανόησης συμπεριλαμβάνουμε και τις αντίστοιχες εντολές summary. Επιπλέον για να επαληθεύσουμε τις όποιες μας υποθέσεις κρίνουμε φρόνιμο το να εκτελέσουμε επιπλέον εντολές που συσχετίζουν τις τιμές των ανεξάρτητων μεταβλητών μας με αυτές των εξαρτημένων.

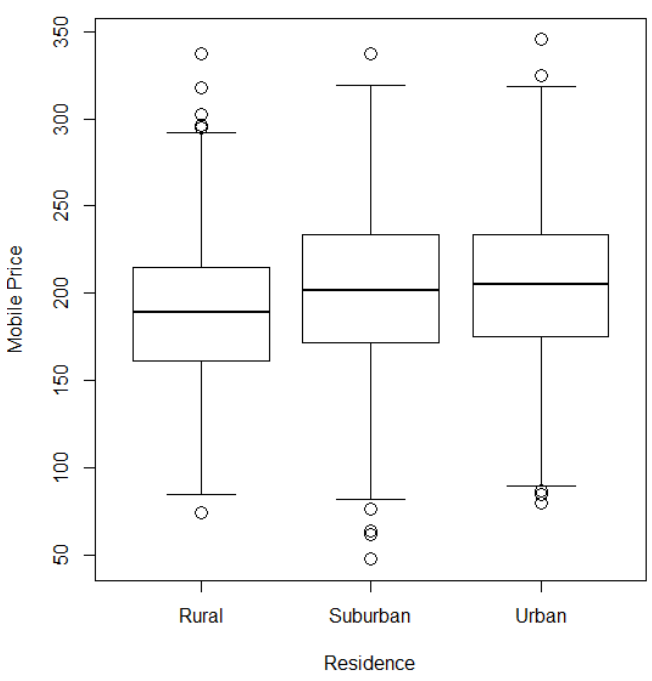
## Τόπος κατοικίας πληθυσμού

```
> plot(cell.phones$usr_r,ylim=range(0:1200))

> summary(cell.phones$usr_r)

> boxplot(cell.phones$mobileprice ~ cell.phones$usr_r, lty = 1, lwd = .8, cex = 1.5,
                xlab = "Residence", ylab = "Mobile Price", border = "black")

> summary(cell.phones$usr_r)
```

| Rural | Suburban | Urban | NA's |
|---|---|---|---|
| 463 | 1135 | 599 | 55 |




**Barplot τόπου κατοικίας**: Εδώ παρατηρούμε πώς η πλειοψηφία του πληθυσμού κατοικεί είτε μεγάλες πόλεις είτε σε προαστιακές περιοχές. Ένα τέταρτο μόνο μόνο του πληθυσμού μας κατοικεί σε αγροτική περιοχή.

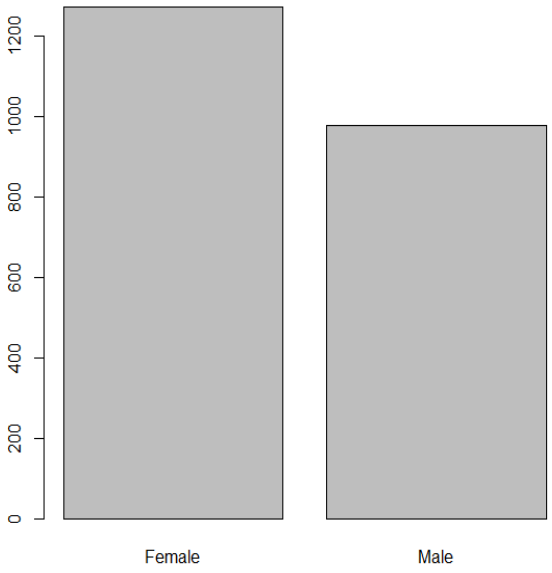


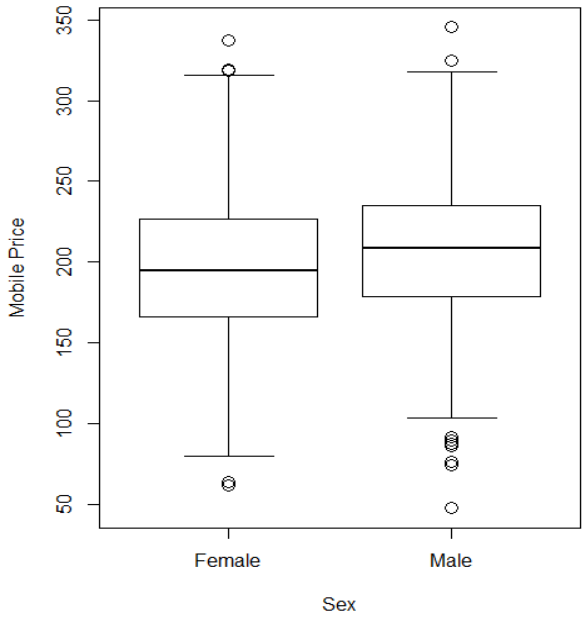

**Boxplot ώστε να ανακαλύψουμε συσχέτιση μεταξύ κατοικίας και αγοραστικών προτιμήσεων κινητών τηλεφώνων:** Παρατηρούμε πως όσοι κατοικούν σε αγροτικές περιοχές ξοδεύουν λιγότερα χρήματα. Η διαμεσος της τιμής κινητού για όσους ζουν σε αγροτική περιοχή είναι μικρότερη. Παρατηρούμε πως η κατοίκηση σε πολύ μεγάλη πόλη δεν επιδρά σημαντικά στην τιμή της συσκευής σε σχέση με το εάν κατοικεί κάποιος σε προάστιο κάποιας μεγάλης πόλης ή σε κάποια μικρότερη πόλη στην γειτονιά αυτής. Αυτό ίσως οφείλεται στο ότι οι ανάγκες και των δύο ομάδων συνδέονται με την ζωή στην πόλη.

## Σχέση Ανδρών/Γυναικών

Καταρχάς πρέπει να ελέγξουμε το ποσοστό των ανδρών και των γυναικών για να επαληθεύσουμε το ότι ερωτήθηκε παρόμοιος αριθμός γυναικών και ανδρών και οπότε το δείγμα μας αντικατοπτρίζει επαρκώς τη διαφορά φύλου.

```
# Plot mobileprice and sex
> summary(cell.phones$sex)
> plot(cell.phones$sex)
> boxplot(cell.phones$mobileprice ~ cell.phones$sex, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Sex", ylab = "Mobile Price", border = "black")
> summary(cell.phones$sex)
Female   Male
 1273    979
> summary(cell.phones$age)
Min.   1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
18.00   37.00    52.00     50.98  64.00   97.00     51
```




Εδώ παρατηρούμε πως οι άνδρες ξοδεύουν περισσότερα χρήματα σε σχέση με τις γυναίκες για την αγορά μιας συσκευής κινητού. Επίσης βλέπουμε ότι από την πλευρά των αντρών υπάρχουν περισσότεροι outliers, δηλαδή άτομα που επέλεξαν να αγοράσουν κινητό σε τιμές κυρίως μικρότερες από το μεγαλύτερο δείγμα του πληθυσμού.

## Σχέση μεταξύ τύπου εργασίας και κινητού τηλεφώνου


Ο τύπος εργασίας και το κόστος κινητού τηλεφώνου μας δίνουν αποτελέσματα που θα περιμέναμε. Πιο συγκεκριμένα παρατηρούμε ότι αυτοί που ξοδεύουν τα περισσότερα είναι οι εργαζόμενοι πλήρους απασχόλησης και οι φοιτητές. Ο λόγος που θα μπορούσε να συμβαίνει αυτό είναι ότι ένας εργαζόμενος πλήρους απασχόλησης έχει αρκετά έσοδα για να καλύψει τις ανάγκες του και να κάνει μεγαλύτερη “σπατάλη” στην αγορά του τηλεφώνου, ενώ αντίστοιχα ένας φοιτητής είναι νεότερος (επομένως και πιο σχετικός με την τεχνολογία) μπορεί να έχει επιχορήγηση από

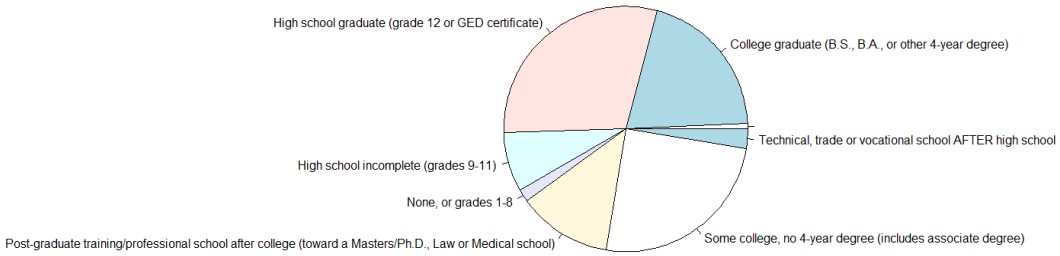
τους γονείς του. Αναμενόμενο μπορούμε να θεωρήσουμε και το γεγονός ότι οι συνταξιούχοι ξοδεύουν λιγότερα χρήματα για αγορά smartphone καθώς ως επί το πλείστον πρόκειται για άτομα μεγάλης ηλικίας που δεν έχουν ίδια επαφή με την τεχνολογία όσο οι νεότεροι.

## Σύμπτυξη επιπέδων μόρφωσης

```
# Plot mobileprice and education

> summary(cell.phones$educ)

> pie(table(cell.phones$educ))

> ANOVA1_EDUC = aov(cell.phones$mobileprice ~ cell.phones$educ)

> levels(cell.phones$educ)

> levels(cell.phones$educ)[c(3,4)]= "LOW LEVEL"

> levels(cell.phones$educ)

> levels(cell.phones$educ)[c(2,6)]= "MID LEVEL"

> levels(cell.phones$educ)

> levels(cell.phones$educ)[c(1,5)]= "MID-HIGH LEVEL"

> levels(cell.phones$educ)

> levels(cell.phones$educ)[c(4)]= "HIGH LEVEL"

> summary(cell.phones$educ)

> plot(cell.phones$educ)

> ANOVA2_EDUC = aov(cell.phones$mobileprice ~ cell.phones$educ)

> anova(ANOVA1_EDUC, ANOVA2_EDUC)

> boxplot(cell.phones$mobileprice ~ cell.phones$educ, lty = 1, lwd = .8, cex = 1.5,

      xlab = "Education Level", ylab = "Mobile Price", border = "black", notch = T)

> summary(aov(cell.phones$mobileprice ~ cell.phones$educ))
```

| Education Level | Frequency |
|---|---|
| College graduate (B.S., B.A., or other 4-year degree) | 453 |
| High school graduate (grade 12 or GED certificate) | 667 |
| High school incomplete (grades 9-11) | 176 |
| None, or grades 1-8 | 38 |

| | |
|---|---|
| Post-graduate training/professional school after college (toward a Masters/Ph.D., Law or Medical school) | 282 |
| Some college, no 4-year degree (includes associate degree) | 562 |
| Technical, trade or vocational school AFTER high school | 58 |
| NA | 16 |




*Γενικότερα σχόλια για το επίπεδο μόρφωσης του πληθυσμού*

Η ύπαρξη πολλών διαφορετικών τιμών στην μεταβλητή αύτη σε συνδυασμό με την ύπαρξη τιμών που μοιάζουν μεταξύ τους (πχ. Η τιμή Some college με την College graduate δεν έχουν σημαντική διαφορά για να τις θεωρούμε διαφορετικές). Ακόμη άλλες τιμές έχουν πολύ λίγες παρατηρήσεις και θα ήταν σκόπιμο να ενωθούν με κάποια άλλη παρεμφερή τιμή της μεταβλητής.

Μετά την κάθε LEVELS εντολή αλλάζει το INDEXING οπότε αναπροσαρμόσει τις τιμές στις παρενθέσεις.

Οι νέες μας τιμές θα είναι οι εξής και θα περιέχουν τις παρακάτω αναφερόμενες τιμές.

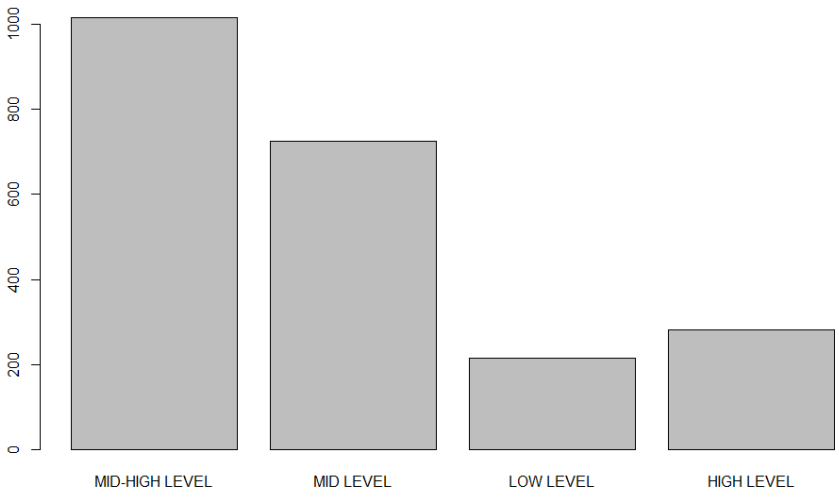
**LOW-LEVEL**: NONE, High school incomplete (Μπορεί να αναφέρεται σε παιδιά που ίσως δεν έχουν αρκετά χρήματα για την αγορά ακριβής συσκευής ή σε κάποιον που έχει απλώς ελλιπή μόρφωση. Για να το εξακριβώσουμε αυτό θα τρέξουμε μία summary για το age για να ελέγξουμε τις ηλικίες του πληθυσμού μας.)

Παρατηρούμε πως δεν υπάρχουν παιδιά. Οπότε το χαμηλό επίπεδο μόρφωσης δεν οφείλεται στην ηλικία των ερωτηθέντων.

**MID LEVEL**: High school graduate, technical

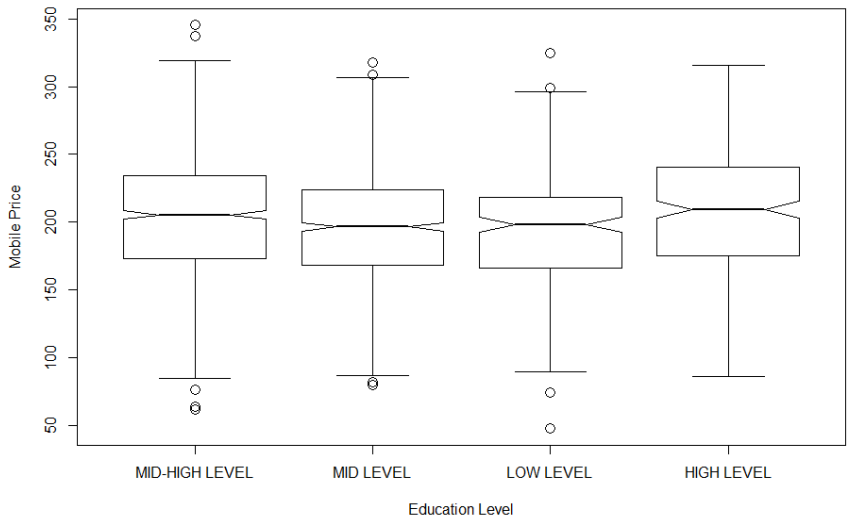
**MID-HIGH LEVEL**: some college graduate, college graduate

**HIGH**: PDH level - Master




Παρατηρώντας τη κατανομή του πληθυσμού μας σύμφωνα με το επίπεδο μόρφωσης του βλέπουμε πως οι περισσότεροι ανήκουν στην MID-HIGH κατηγορία. Ακολουθούν όσοι έχουν MID-LEVEL μόρφωση ενώ αισθητά χαμηλότερος είναι ο αριθμός των όσων έχουν πολύ υψηλή μόρφωση ή καθόλου μόρφωση. Παρόλα αυτά ο αριθμός των περιπτώσεων είναι περί τα διακόσια και για τις δύο αυτές τιμές οπότε και η συσχέτιση της τιμής EDUCATION με την mobileprice θα έχει αξιοπιστία.

## Σχέση Εκπαίδευσης με τιμή αγοράς κινητού




Παρατηρούμε μία ελάχιστη διαφορά μεταξύ των διαφόρων επιπέδων εκπαίδευσης. Πιο συγκεκριμένα, όσοι ανήκουν στις κατηγορίες HIGH LEVEL & MID HIGH LEVEL τείνουν να δίνουν

ελάχιστα περισσότερα χρήματα για την αγορά μιας συσκευής. Αυτό μπορεί να οφείλεται στο ότι χρειάζονται μία συγκεκριμένη λειτουργία που συνήθως κοστίζει λίγο περισσότερο για να την αποκτήσεις (π.χ. λειτουργία κάλυψης δικτύου 5G). Παρόλα αυτά μπορούμε να πούμε πως δεν είναι στατιστικά σημαντική μεταβλητή το επίπεδο της μόρφωσης.

Αποφασίσαμε να εκτελέσουμε έλεγχο ANOVA προκειμένου να δούμε αν ή υπόθεση πως η εκπαίδευση δεν παίζει ρόλο στην τιμή του κινητού ισχύει ή εάν δεν ισχύει και απλώς επηρεάζεται και από άλλους παράγοντες που θα εξεταστούν μετέπειτα κατά τη κατασκευή πιο ολοκληρωμένων μοντέλων.

```
> summary(aov(cell.phones$mobileprice ~ cell.phones$educ))
```

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *cell.phones$educ* | *3* | *33708* | *11236* | *5.747* | *0.000649 ****** |
| *Residuals* | *2232* | *4363420* | *1955* | | |

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

*16 observations deleted due to missingness*

Παρατηρούμε πως το p-value ισούται με 0.00064 < 0,05. Άρα η διαφορά ανάμεσα στις στάθμες του παράγοντα είναι στατιστικά σημαντική. Θα πρέπει να εξερευνήσουμε με ποια άλλη μεταβλητή συνδέεται η EDUC.

Παρακάτω κάνουμε την εντολή ANOVA για να συγκρίνουμε τα δυο μοντέλα (πριν και μετά την σύμπτυξη) και παρατηρούμε ότι το p>0.05 επομένως το μοντέλο είναι πιο απλό και δεν διαφέρει σημαντικά από το προηγούμενο.

## Συσχέτιση εισοδήματος με τιμή κινητού

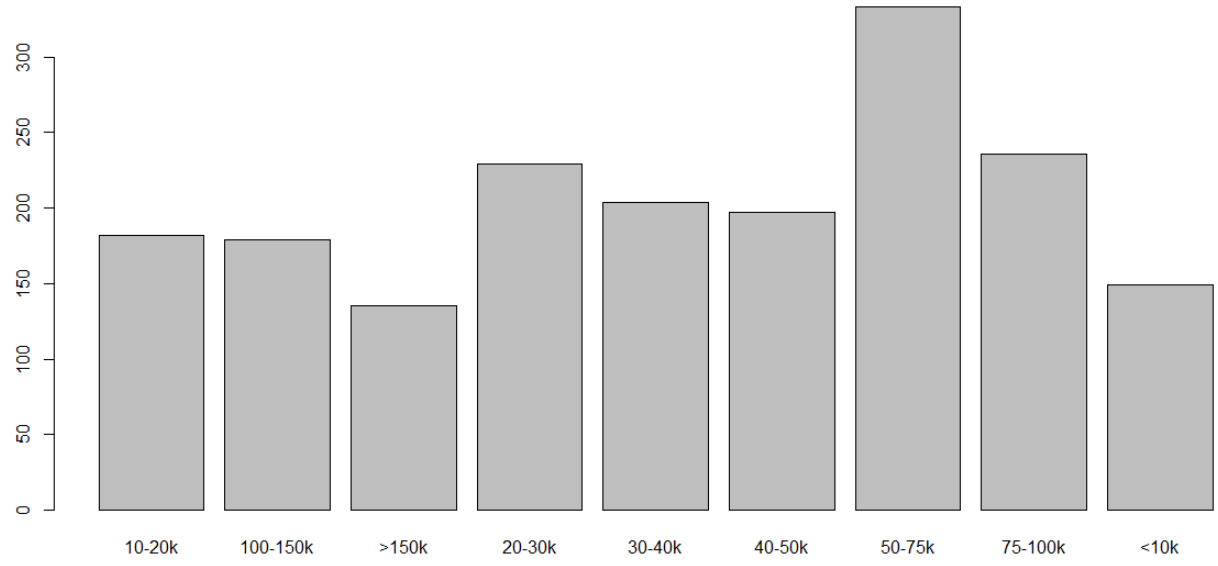
```
# Plot mobileprice and income

> summary(cell.phones$inc)

> levels(cell.phones$inc) <-  c("10-20k", "100-150k", ">150k", "20-30k", "30-40k",
                        "40-50k", "50-75k", "75-100k", "<10k")

> plot(cell.phones$inc)

> boxplot(cell.phones$mobileprice ~ cell.phones$inc, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Income (USD)", ylab = "Mobile Price", border = "black")

> summary(aov(cell.phones$mobileprice ~ cell.phones$inc))
```

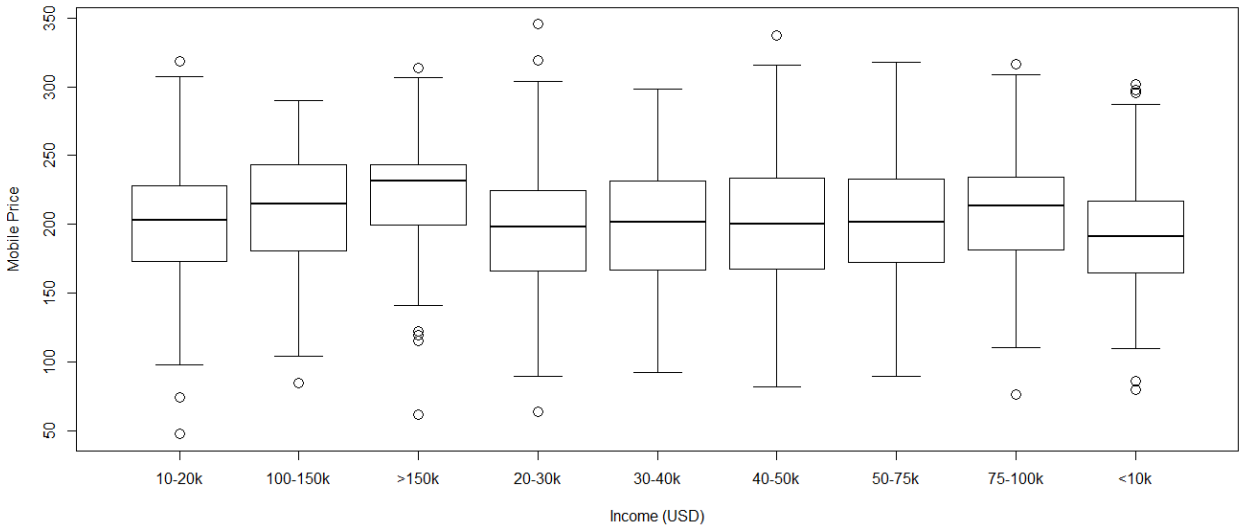
Κάνοντας ένα διάγραμμα παρατηρούμε πως πολύ μεγάλο ποσοστό των ερωτηθέντων δεν απάντησαν σχετικά με το εισόδημα τους. Αυτό θα μπορούσε να μας οδηγήσει στο να μην λάβουμε καθόλου υπόψη την μεταβλητή αυτή. Παρόλα αυτά κρίνουμε πως πρόκειται για μεταβλητή κλειδί

και πως θα υπάρχει ισχυρή συσχέτιση. Για αυτό το λόγω θα αναλύσουμε περαιτέρω τη συσχέτιση αυτή αγνοώντας τις παρατηρήσεις που δεν έχουν κάποια τιμή .

```
> summary(cell.phones$inc)
```

| | | |
|---|---|---|
| *$10,000 to under $20,000* | *$100,000 to under $150,000* | *$150,000 or more* |
| *182* | *179* | *135* |
| *$20,000 to under $30,000* | *$30,000 to under $40,000* | *$40,000 to under $50,000* |
| *229* | *204* | *197* |
| *$50,000 to under $75,000* | *$75,000 to under $100,000* | *Less than $10,000* |
| *333* | *236* | *149* |
| *NA's : 408* | | |






**Συμπεράσματα**: Παρατηρούμε ότι όσοι βγάζουν εως 10.000$ ξοδεύουν λιγότερα χρήματα από όλους για την αγορά συσκευής. Μετά από αυτό το μέγεθος εισοδήματος και εώς τις 75.000$ δεν υπάρχουν σημαντικές αποκλίσεις. Μετά τις 100.000$ όμως παρατηρούμε πώς ξοδεύουν περισσότερα χρήματα από ότι στις προηγούμενες κατηγορίες. Όσοι έχουν εισόδημα πάνω από 150.000$ έχουν και αισθητά πιο υψηλή διάμεσο.

> summary(aov(cell.phones$mobileprice ~ cell.phones$inc))

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *cell.phones$inc* | *8* | *75118* | *9390* | *4.796* | *7.43e-06 **** |
| *Residuals* | *1835* | *3592801* | *1958* | | |

---

*Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

408 observations deleted due to missingness

Επαληθεύουμε και με την ANOVA το γεγονός ότι το εισόδημα επηρεάζει την τιμή αγοράς του κινητού. Παρατηρούμε πως το p-value < 0.05 . Άρα η διαφορά ανάμεσα στις στάθμες του παράγοντα είναι στατιστικά σημαντική .

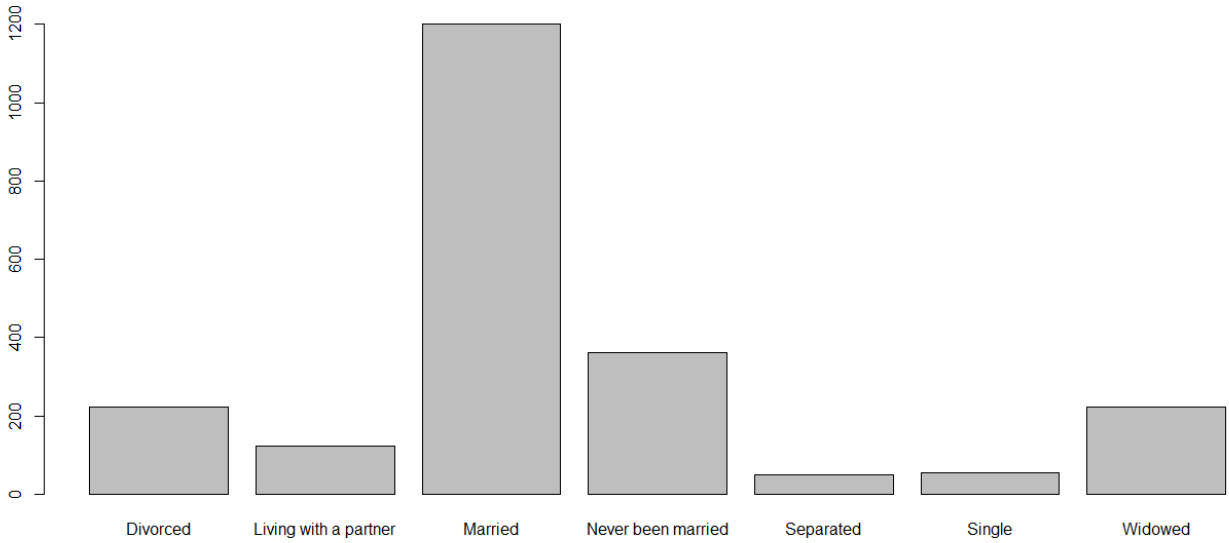
## Συσχέτιση οικογενειακής κατάστασης(Παντρεμένος/η ή όχι) με τιμή κινητού τηλεφώνου

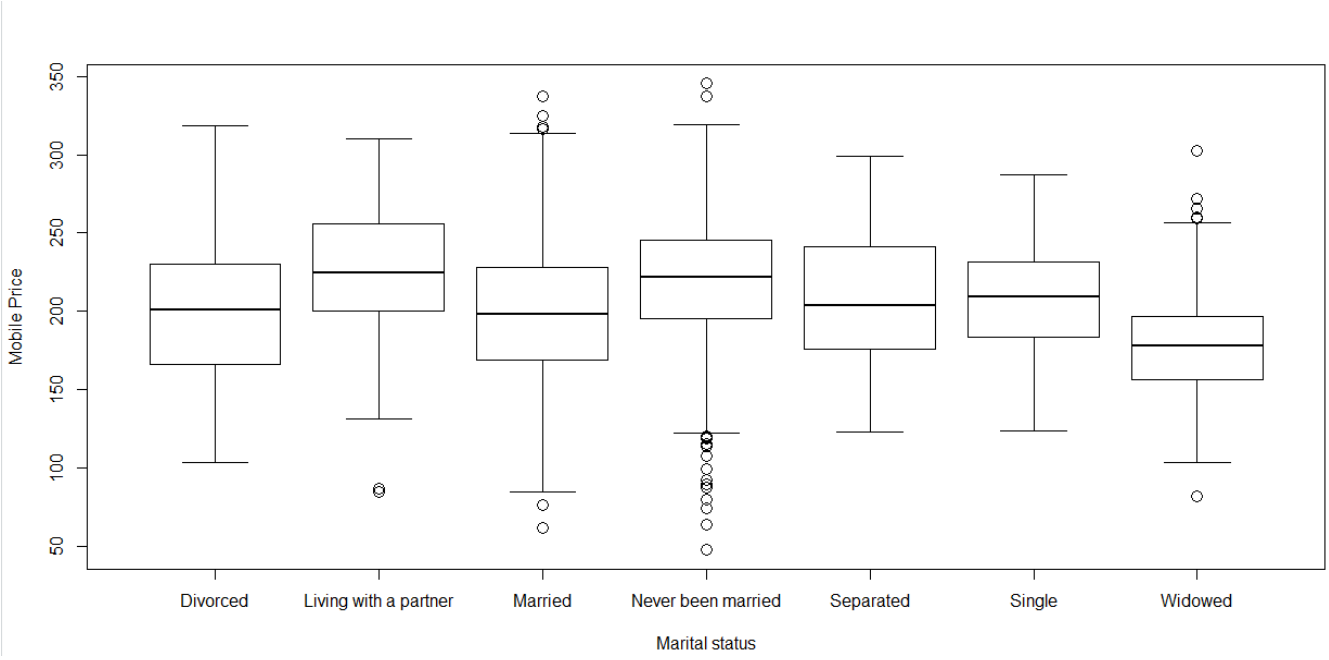
# Plot mobileprice and mar

```
> summary(cell.phones$mar)

> plot(cell.phones$mar)

> summary(cell.phones[cell.phones$mar=="Widowed",])

> boxplot(cell.phones$mobileprice ~ cell.phones$mar, lty = 1, lwd = .8, cex = 1.5,

      xlab = "Marital status", ylab = "Mobile Price", border = "black")

> summary(aov(cell.phones$mobileprice ~ cell.phones$mar))

> summary(cell.phones$mar)
```

| *Divorced* | *Living with a partner* | *Married* | *Never been married* |
|---|---|---|---|
| *223* | *124* | *1200* | *363* |
| *Separated* | *Single* | *Widowed* | *NA's* |
| *50* | *54* | *222* | *16* |

Παρατηρούμε ότι το μεγαλύτερο ποσοστό είναι παντρεμένοι, ακολουθούν οι διαζευγμένοι και οι χήροι. Έπειτα όσοι ζουν με τον σύντροφο τους και τέλος οι χωρισμένοι και οι ελεύθεροι των οποίων όμως ο αριθμός δεν είναι πολύ μικρός ώστε να επηρεάσει αρνητικά την αξιοπιστία των αποτελεσμάτων.




Ακόμη, παρατηρούμε ότι τα λιγότερα χρήματα τα δίνουν οι χήροι. Αυτό μπορεί να οφείλεται στο ότι είναι μεγάλοι σε ηλικία. Επίσης, παρατηρούμε πως σε όσους δεν έχουν παντρευτεί υπάρχουν πολλές τιμές που θεωρούνται outliers κυρίως προς τα κάτω. Θα κάνουμε μία summary για να πάρουμε πιο λεπτομερή έξοδο ανά κατηγορία.

```
> summary(cell.phones[cell.phones$mar=="Widowed",])
```

| Age | Mobile Price |
|---|---|
| Min. : 33.00 | Min. : 82.3 |

| | |
|---|---|
| 1st Qu. : 66.25 | 1st Qu.:156.7 |
| Median : 76.00 | Median :177.9 |
| Mean : 74.39 | Mean :179.3 |
| 3rd Qu. : 82.00 | 3rd Qu.:196.6 |
| Max. :97.00 | Max. :302.7 |
| NA's :20 | NA's :16 |

Είναι λογικό το ότι οι χήροι που είναι και ηλικιωμένοι δεν ξοδεύουν περισσότερα χρήματα. Περισσότερα χρήματα φαίνεται να ξοδεύουν όσοι συζούν με σύντροφο. Παρατηρούμε ακόμη ότι οι παντρεμένοι, οι διαζευγμένοι και οι χωρισμένοι ξοδεύουν ίδια ποσά για κινητά αλλά λιγότερα από όσους είναι σε σχέση, δεν έχουν παντρευτεί ποτέ ή είναι ελεύθεροι.

```
> summary(aov(cell.phones$mobileprice ~ cell.phones$mar))
```

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *cell.phones$mar* | *6* | *265549* | *44258* | *24.06* | *<2e-16 ****|
| *Residuals* | *2229* | *4099631* | *1839* | | |

---

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

*16 observations deleted due to missingness*

Επαληθεύουμε και με την ANOVA το γεγονός ότι το εισόδημα επηρεάζει την τιμή αγοράς του κινητού. Παρατηρούμε πως το p-value < 0.05. Άρα η διαφορά ανάμεσα στις στάθμες του παράγοντα είναι στατιστικά σημαντική .

Ωστόσο δεν προσδιορίζεται από την απλή ANOVA σε ποια επίπεδα του παράγοντα επηρεάζουν τη τιμή αγοράς μιας συσκευής. Στην περίπτωση μας γίνονται 14 συγκρίσεις λόγω το ότι έχουμε 7 επίπεδα και έτσι η πιθανότητα σφάλματος αυξάνει δραματικά. Σε θεωρητικό επίπεδο η πιθανότητα να έχει γίνει έστω ένα σε σφάλμα σε κάποια σύγκριση ανέρχεται στο **1-(0.95)^14= 0.51!.** Για το λόγο αυτό θα ήταν φρόνιμο να χρησιμοποιηθεί μέθοδος Post Hoc Test (π.χ. Tukey) για επιπλέον διερεύνηση των σχέσεων

```
> AOV1=aov(cell.phones$mobileprice ~ cell.phones$mar)

> library(DescTools)

> PostHocTest(aov(cell.phones$mobileprice ~ cell.phones$mar),method="hsd")

> PostHocTest(aov(cell.phones$mobileprice ~ cell.phones$mar),method="duncan")
```

```
> PostHocTest(aov(cell.phones$mobileprice ~ cell.phones$mar),method="hsd")
```

**Post Hoc multiple comparisons of means : Tukey HSD - 95% family-wise confidence level**

| $`cell.phones$mar` | diff | lwr.ci | upr.ci | pval |
|---|---|---|---|---|
| Living with a partner-Divorced | 23.053678 | 8.876447 | 37.230908 | 3.5e-05 *** |
| Married-Divorced | -2.284849 | -11.513736 | 6.944038 | 0.99072 |
| Never been married-Divorced | 15.649939 | 4.881989 | 26.417888 | 0.00037 *** |
| Separated-Divorced | 7.422484 | -12.380645 | 27.225614 | 0.92643 |
| Single-Divorced | 5.547077 | -13.647587 | 24.741741 | 0.97916 |
| Widowed-Divorced | -21.37074 | -33.369630 | -9.371852 | 3.4e-06 *** |
| Never been married-Living with a partner | -7.403739 | -20.567812 | 5.760334 | 0.64315 |
| Separated-Living with a partner | -15.631194 | -36.832798 | 5.570411 | 0.30917 |
| Single-Living with a partner | -17.506601 | -38.141019 | 3.127817 | 0.15830 |
| Widowed-Living with a partner | -44.424419 | -58.613055 | -30.235782 | 5.6e-11 *** |
| Never been married-Married | 17.934788 | 10.353805 | 25.515770 | 1.4e-10 *** |

| | | | | |
|---|---|---|---|---|
| Separated-Married | 9.707333 | -8.559755 | 27.974421 | 0.70279 |
| Single-Married | 7.831926 | -9.773689 | 25.437541 | 0.84618 |
| Widowed-Married | -19.085892 | -28.332291 | -9.839493 | 2.8e-08 *** |
| Separate | -1.875407 | -26.713871 | 22.963056 | 0.99999 |
| Widowed-Separated | -28.793225 | -48.604522 | -8.981928 | 0.00037 *** |
| Widowed-Single | -26.917818 | -46.120908 | -7.714728 | 0.00072 *** |

Ας δούμε όμως και τι αποτελέσματα παράγονται εάν χρησιμοποιήσουμε και ως μέθοδο την Duncan.

```
> PostHocTest(aov(cell.phones$mobileprice ~ cell.phones$mar),method="duncan")
```

**Posthoc multiple comparisons of means : Duncan's new multiple range test - 95% family-wise confidence level**

| $`cell.phones$mar` | diff | lwr.ci | upr.ci | pval |
|---|---|---|---|---|
| Living with a partner-Divorced | 23.053678 | 12.55415 | 33.5532010 | 4.2e-06 *** |
| Married-Divorced | -2.284849 | -8.417676 | 3.8479779 | 0.4651 |
| Never been married-Divorced | 15.649939 | 7.862654 | 23.4372237 | 3.7e-05 *** |
| Separated-Divorced | 7.422484 | -6.432972 | 21.2779410 | 0.3003 |

| | | | | |
|---|---|---|---|---|
| Single-Divorced | 5.547077 | -7.208259 | 18.3024124 | 0.3939 |
| Widowed-Divorced | -21.370741 | -29.76588 | -12.9755990 | 2.4e-07 *** |
| Married-Living with a partner | -25.338527 | -34.3425 | -16.3344996 | 1.4e-09 *** |
| Never been married-Living with a partner<br>. | -7.403739 | -16.151 | 1.3441173 | 0.0971 |
| Separated-Living with a partner | -15.631194 | -30.4651 | -0.7972803 | 0.0385 * |
| Single-Living with a partner | -17.506601 | -32.429228 | -2.5839743 | 0.0203 * |
| Widowed-Living with a partner | -44.424419 | -55.283663 | -33.5651746 | 9.4e-12 *** |
| Never been married-Married | 17.934788 | 12.320384 | 23.5491920 | 2.4e-11 *** |
| Separated-Married | 9.707333 | -3.503262 | 22.9179283 | 0.1552 |
| Single-Married | 7.831926 | -4.486018 | 20.1498698 | 0.2178 |
| Widowed-Married | -19.08589 | -25.230356 | -12.9414279 | 1.4e-09 *** |
| Separated-Never been married | -8.227455 | -20.913844 | 4.4589345 | 0.2036 |

| | | | | |
|---|---|---|---|---|
| Single-Never been married | -10.10286 | -23.017873 | 2.812149 | 0.1279 |
| Widowed-Never been married | -37.020680 | -45.153523 | -28.8878364 | 1.1e-11 *** |
| Single-Separated | -1.875407 | -18.381189 | 14.6303744 | 0.8237 |
| Widowed-Separated | -28.793225 | -43.465285 | -14.1211658 | 4.5e-05 *** |
| Widowed-Single | -26.917818 | -40.805321 | -13.0303147 | 7.1e-05 *** |

Όπως φαίνεται παραπάνω τα αποτελεσματα απο τις δυο μεθοδους έχουν ελάχιστες διαφορές μεταξύ τους και όποτε περιγράφουν τις συσχετίσεις των επιπέδων με τον ίδιο σχετικά τρόπο.

Από τις τιμές συσχέτισης παραπάνω παρατηρούμε πως τα Married, Divorced, Separated και Single έχουν όλα μεταξύ τους τιμές p-value > 0.05. Αυτό μας καταδεικνύει ότι μπορεί να γίνει σύμπτυξη επιπέδων παράγοντα. Λόγω το ότι όμως το Married δεν ταιριάζει διαισθητικά με τα υπόλοιπα επίπεδα, καλύτερο θα ήταν να μην συμπεριληφθει στη νέα ομάδα ώστε να μη χάσει την εννοιολογική της σημασία. Επίσης τα Single, Divorced και Separated έχουν p-value κοντά στην μονάδα ενώ την ίδια στιγμή υποδηλώνουν και παρόμοια οικογενειακή κατάσταση και για αυτό μπορούν να συμπτυχθούν σε ένα νέο επίπεδο που ονομάζουμε single. Ένα ακόμα επίπεδο που θα δημιουργήσουμε είναι το LivingWithPartner όπου θα γίνει η σύμπτυξη των Never been married με το Living with a partner.

Οι εντολές που εκτελέσαμε είναι οι παρακάτω

```
> AOV1=aov(cell.phones$mobileprice ~ cell.phones$mar)

> library(DescTools)

> PostHocTest(aov(cell.phones$mobileprice ~ cell.phones$mar),method="hsd")

> PostHocTest(aov(cell.phones$mobileprice ~ cell.phones$mar),method="duncan")

> levels(cell.phones$mar)

> levels(cell.phones$mar)[c(1,5,6)] = "single"

> levels(cell.phones$mar)
```

```
> levels(cell.phones$mar)[c(2,4)]= "LivingWithPartner"
> levels(cell.phones$mar)
> AOV2=aov(cell.phones$mobileprice ~ cell.phones$mar)
> anova(AOV1,AOV2)
```

Analysis of Variance Table

*Model 1: cell.phones$mobileprice ~ cell.phones$mar*

*Model 2: cell.phones$mobileprice ~ cell.phones$mar*

| *Res.Df* | *RSS* | *Df* | *Sum of Sq* | *F* | *Pr(>F)* |
|---|---|---|---|---|---|
| *1* | *2229* | *4099631* | | | |
| *2* | *2232* | *4107738* | *-3* | *-8107.2* | *1.4693* | *0.221* |

Η σύγκριση των δύο μοντέλων ANOVA όπως φαίνεται παραπάνω μας δίνει p-value>0.05 (0.221) κάτι το οποίο σημαίνει ότι το μοντέλο είναι πιο απλό και δεν διαφέρει σημαντικά από το προηγούμενο.

- **Συσχέτιση ηλικίας με χαρακτηριστικά κινητού** (π.χ. τιμή κινητού, βγάζει ή όχι φωτογραφίες, χρησιμοποιεί ή όχι το κινητό για βίντεο κ.α)
- **Συσχέτιση ηλικίας με τιμή κινητού τηλεφώνου**

```
# Pearson correlation of age, mobileprice
> cor(cell.phones$age, cell.phones$mobileprice, use = "c")
> cor.test(cell.phones$age, cell.phones$mobileprice, method = "pearson")
> cor.test(cell.phones$age, cell.phones$mobileprice, method = "spearman")
> cor.test(cell.phones$age, cell.phones$mobileprice, method = "kendall")
```

***Pearson's product-moment correlation***

*data:  cell.phones$age and cell.phones$mobileprice*

*t = -18.806, df = 2199, p-value < 2.2e-16*

*alternative hypothesis: true correlation is not equal to 0*

*95 percent confidence interval:*

***[ -0.4076559, -0.3356530 ]***

*sample estimates:*

*cor:* ***-0.3722143***

> cor.test(cell.phones$age, cell.phones$mobileprice, method = "spearman")

***Spearman's rank correlation rho***

*data: cell.phones$age and cell.phones$mobileprice*

*S = 2487758055, p-value < 2.2e-16*

*alternative hypothesis: true rho is not equal to 0*

*sample estimates:*

*rho :* ***-0.3999075***

> cor.test(cell.phones$age, cell.phones$mobileprice, method = "kendall")

***Kendall's rank correlation tau***

*data: cell.phones$age and cell.phones$mobileprice*

*z = -18.915, p-value < 2.2e-16*

*alternative hypothesis: true tau is not equal to 0*

*sample estimates:*

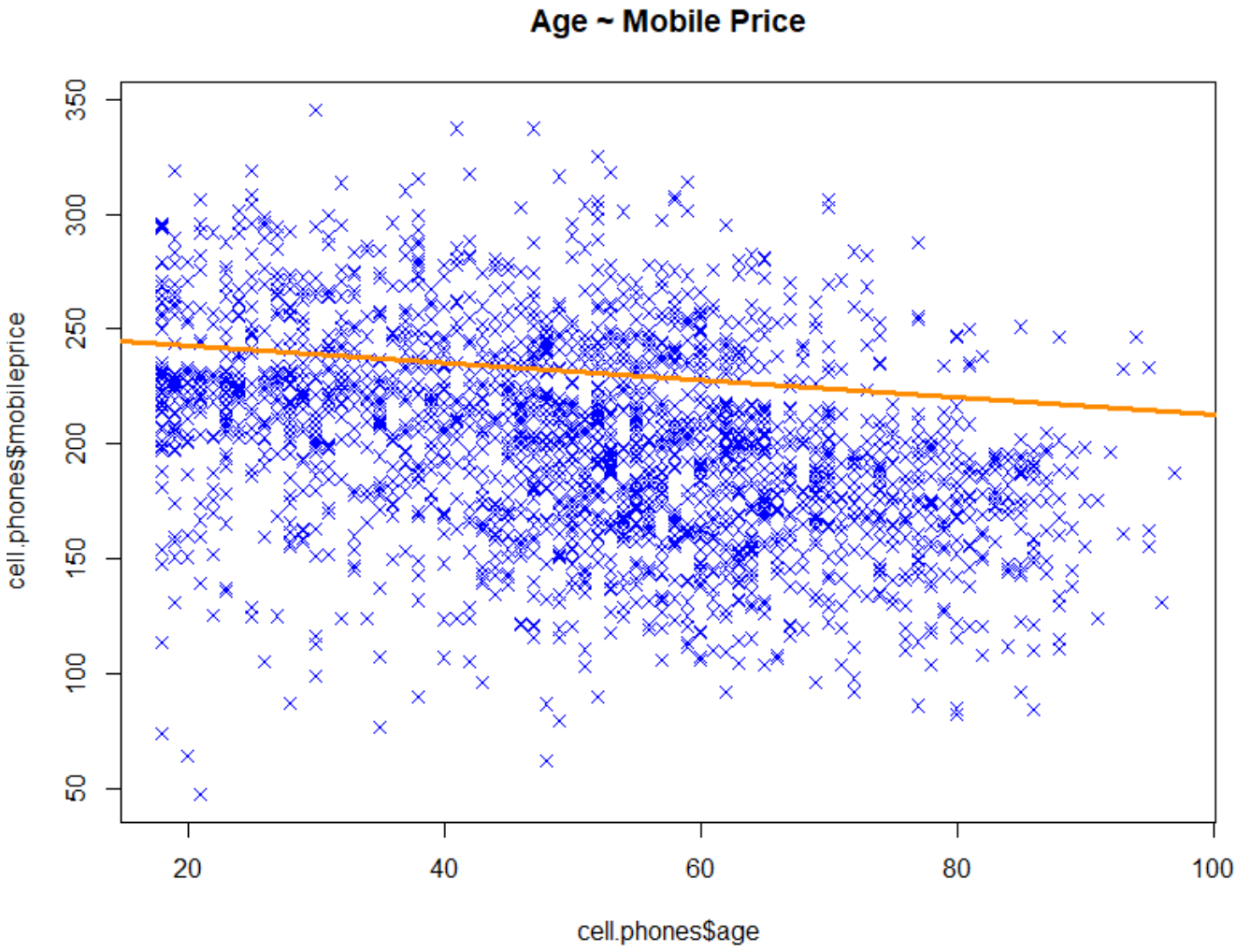
*tau :* ***-0.2711624***

Παρατηρούμε πως το p-value < 0.05. Άρα απορρίπτουμε το ότι δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές. Η συσχέτιση αυτή όπως υποδηλώνουν οι τιμές από τα διάφορα τεστ (**-0.3722, -0.3999, -0.2711**) είναι αρνητική αλλά μικρή. Μπορεί να θεωρηθεί και σχετικά ασήμαντη καθώς δεν ξεπερνά το όριο του **0.4**. Αυτό φαίνεται και στο παρακάτω plot όπου δεν παρατηρείται κάποια πολύ μεγάλη σχέση αλλά βλέπουμε ότι όσο μεγαλώνει η ηλικία υπάρχουν πολύ περισσότερα σημεία προς το κάτω μέρος του γραφήματος.

# plot age with mobileprice

> plot(cell.phones$age, cell.phones$mobileprice, pch = 4, col = "blue", main = "Age ~ Mobile Price")

> abline(b = cor(cell.phones$age, cell.phones$mobileprice, use = "c"), a = 250, lwd = 3, col = "darkorange")

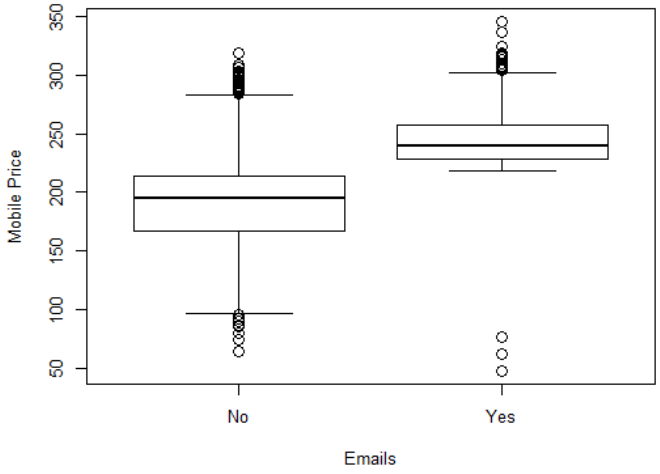


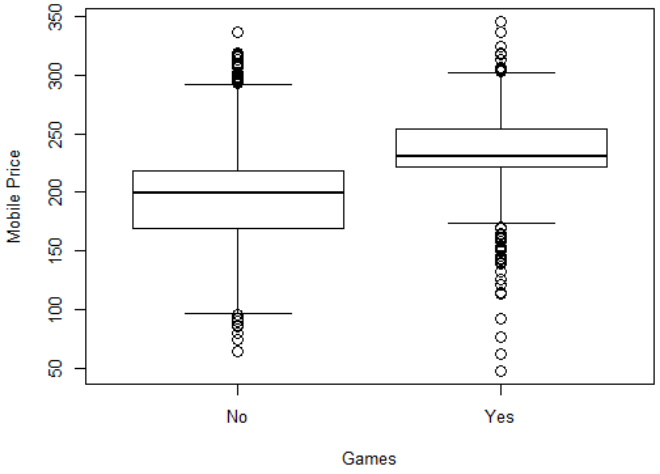

**Εξέταση μερικών ακόμη χαρακτηριστικών που ενδεχομένως εμφανίζουν ενδιαφέρον**

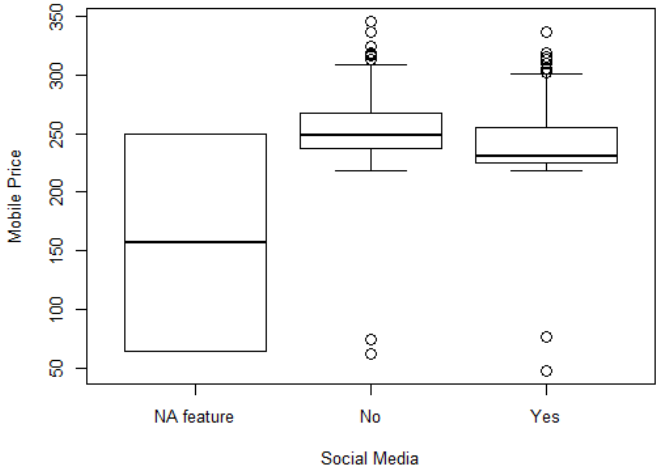
Εξέταση της συσχέτισης των μεταβλητών *"Χρησιμοποιεί το κινητό για Social media", "Χρησιμοποιεί το κινητό για να στείλει Email" , "Νιώθετε πιο ασφαλής επειδή μπορείτε να χρησιμοποιήσετε το κινητό σας για να καλέσετε βοήθεια" , "Χρησιμοποιείτε το κινητό σας για να παρακολουθήσετε βίντεο", "Αν έχει κατεβάσει κάποια εφαρμογή στο κινητό", "Εάν λαμβάνει γραπτά μηνύματα ή όχι ", "παίζετε παιχνίδια στο κινητό σας","Χρησιμοποιείτε το κινητό σας για να μπείτε στο Ιντερνετ"*

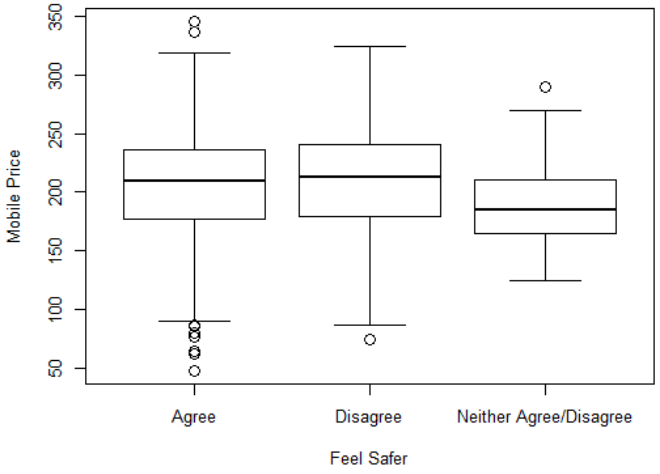
Σε πολλές μεταβλητές έγινε των ονομάτων των τιμών τους για λόγους καλύτερης απεικόνισης.

```
# plot some other features
> levels(cell.phones$q17e) <-  c("NA feature", "No", "Yes")
> levels(cell.phones$q22a) <-  c("Agree", "Disagree", "Neither Agree/Disagree")
> par(mfrow=c(2,2))
> boxplot(cell.phones$mobileprice ~ cell.phones$q14a, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Emails", ylab = "Mobile Price", border = "black", notch = F)
>boxplot(cell.phones$mobileprice ~ cell.phones$q14g, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Games", ylab = "Mobile Price", border = "black", notch = F)
> boxplot(cell.phones$mobileprice ~ cell.phones$q17e, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Social Media", ylab = "Mobile Price", border = "black", notch = F)
> boxplot(cell.phones$mobileprice ~ cell.phones$q22a, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Feel Safer", ylab = "Mobile Price", border = "black", notch = F)
```




**Παρατηρήσεις για τα παραπάνω :**

Οι παράγοντες ***"Χρησιμοποιεί το κινητό για να στείλει Email"*** και ***"Παίζετε παιχνίδια στο κινητό σας"*** επηρεάζουν αισθητά την τιμή αγοράς μιας νέας συσκευής . Αυτό ήταν και ένα αναμενόμενο αποτέλεσμα που επαληθεύτηκε από τα παραπάνω διαγράμματα.

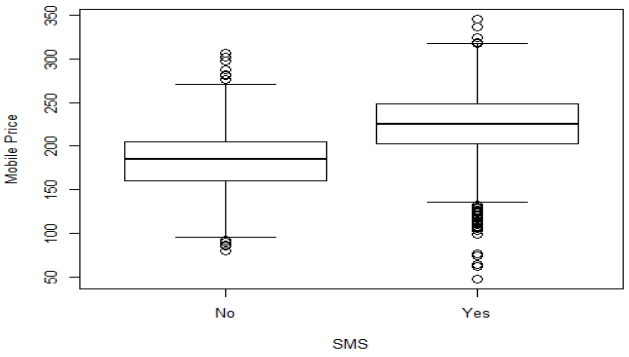
Από την άλλη η χρήση ή μη social media δε δεν φαίνεται να επηρεάζει τη τιμή του κινητού. Το μόνο που την επηρεάζει σε σχέση με αυτή τη μεταβλητή είναι το εάν το κινητό δεν έχει καν τη λειτουργία αυτή. Ωστόσο η μη ύπαρξη αυτής οφείλεται στο ότι τα κινητά δεν είναι μάλλον καν Smartphone οπότε και οι ιδιοκτήτες αυτών δεν έχουν και άλλες ανάγκες πέρα από τη χρήση των Social media.

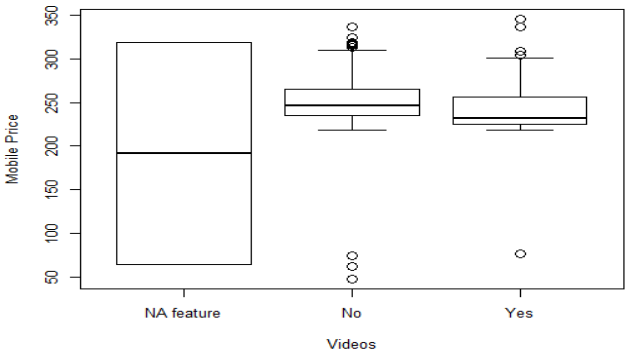
Όσον αφορά τη μεταβλητή **"Νιώθετε πιο ασφαλής επειδή μπορείτε να χρησιμοποιήσετε το κινητό σας για να καλέσετε βοήθεια"** φαίνεται να μην επηρεάζει τη τιμή του κινητού. Αυτό οφείλεται μάλλον στο ότι όλα τα κινητά από τη φύση τους εκτελούν κλήσεις, που αποτελούν και

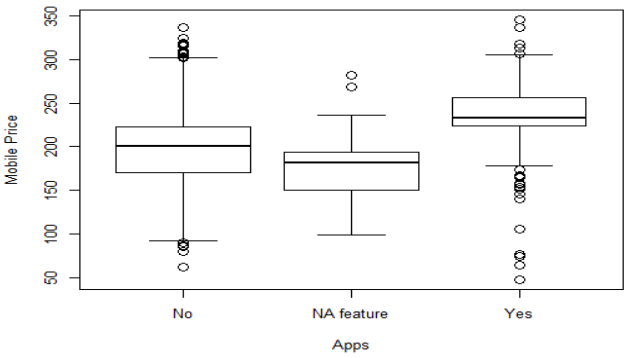
την πρώτη σκέψη όταν πρόκειται να καλέσει βοήθεια. Κανείς δε θα στείλει ένα Mail για παράδειγμα σε περίπτωση έκτακτης ανάγκης.

## Έλεγχος άλλων τεσσάρων μεταβλητών

```
> levels(cell.phones$q17g) <- c("NA feature", "No", "Yes")

> par(mfrow=c(2,2))

> levels(cell.phones$q24) <- c("No", "NA feature", "Yes")

> boxplot(cell.phones$mobileprice ~ cell.phones$q14b, lty = 1, lwd = .8, cex = 1.5,
      xlab = "SMS", ylab = "Mobile Price", border = "black", notch = F)

> boxplot(cell.phones$mobileprice ~ cell.phones$q17g, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Videos", ylab = "Mobile Price", border = "black", notch = F)

> boxplot(cell.phones$mobileprice ~ cell.phones$q24, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Apps", ylab = "Mobile Price", border = "black", notch = F)

> boxplot(cell.phones$mobileprice ~ cell.phones$q14h, lty = 1, lwd = .8, cex = 1.5,
      xlab = "Browsing", ylab = "Mobile Price", border = "black", notch = F)
```




Εδώ μπορούμε να παρατηρήσουμε τα εξής :

Οι μεταβλητές *"Αν έχει κατεβάσει κάποια εφαρμογή στο κινητό", "Εάν λαμβάνει γραπτά μηνύματα ή όχι ", "Χρησιμοποιείτε το κινητό σας για να μπείτε στο Ιντερνετ"*

Η μεταβλητή *"Χρησιμοποιείτε το κινητό σας για να παρακολουθήσετε βίντεο",* φαίνεται να μην επηρεάζει τόσο τη τιμή αγοράς. Θα γίνει ωστόσο και ένας έλεγχος ANOVA για να το σιγουρεψουμε. Στον έλεγχο αυτό θα απομακρυνθουν οι δύο παρατηρήσεις που απάντησαν πως δεν έχουν καν τη λειτουργία αυτή στο κινητό τους

```
> summary(aov(cell.phones$mobileprice[cell.phones$q17g != "NA feature"] ~
 cell.phones$q17g[cell.phones$q17g != "NA feature"]))
```

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| | *2* | *18716* | *9358* | *12.5* | *4.54e-06 **** |
| *Residuals* | *776* | *580935* | *749* | | |

*---*

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

*1473 observations deleted due to missingness*

Παρατηρούμε πως το p-value < 0.05 . Άρα η διαφορά ανάμεσα στις στάθμες του παράγοντα είναι στατιστικά σημαντική και απορρίπτεται η αρχική μας υπόθεση. Εδώ βγαίνει το συμπέρασμα που είπαμε και στις διαλέξεις πως σε περίπτωση αντιπαράθεσης της κοινής λογικής με τα αποτελέσματα ενός Plot χρειάζεται περαιτέρω εξέταση μέσω άλλων μεθόδων.

Η μεταβλητή ***"Αν έχει κατεβάσει κάποια εφαρμογή στο κινητό"*** τόσο στο boxplot όσο και στην ANOVA που τρέχουμε παρακάτω φαίνεται επίσης να επηρεάζει τη τιμή αγοράς μιας νέας συσκευής.

```
> summary(aov(cell.phones$mobileprice ~ cell.phones$q24))
```

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *cell.phones$q24* | *2* | *523711* | *261855* | *154.6* | *<2e-16 **** |
| *Residuals* | *1909* | *3233065* | *1694* | | |

*---*

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

*340 observations deleted due to missingness*

Παρατηρούμε πως το p-value < 0.05. Άρα η διαφορά ανάμεσα στις στάθμες του παράγοντα είναι στατιστικά σημαντική

Η μεταβλητή ***"Εάν λαμβάνει γραπτά μηνύματα ή όχι"*** επίσης φαίνεται να επηρεάζει την τιμή αγοράς. Ο ίδιος σχολιασμός με παραπάνω μπορεί να γίνει και αναφορικά με τα αποτελέσματα της εντολής ANOVA και σε αυτήν τη μεταβλητή.

```
> summary(aov(cell.phones$mobileprice ~ cell.phones$q14b))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| cell.phones$q14b | 1 | 711801 | 711801 | 447.1 | <2e-16 *** |
| Residuals | 1915 | 3048858 | 1592 | | |

---

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

*335 observations deleted due to missingness*

Η μεταβλητή ***"Χρησιμοποιείτε το κινητό σας για να μπείτε στο Ιντερνετ"*** επηρεάζει ξεκάθαρα τη τιμή αγοράς όπως φαίνεται και στο BOXPLOT. Πράγμα αναμενόμενο καθώς καμία φθηνή συσκευή στην χρονολογία του δημοσκοπισματος δεν θα είχε αυτή τη λειτουργία.

## Έλεγχος μεταβλητών που πιστεύουμε ότι ΔΕΝ έχουν επίδραση στην τιμή αγοράς κινητού

Για να επαληθεύσουμε πως δεν υπάρχει άμεση συσχέτιση όλων των μεταβλητών με το κόστος του κινητού δοκιμάσαμε να τρέξουμε την εντολή ANOVA για διάφορες μεταβλητές που φαίνονται μη σχετικές.

### 1. "Χρησιμοποιείτε το κινητό σας για να κάνετε μία φιλανθρωπική δωρεά μέσω μηνύματος"

```
> summary(aov(cell.phones$mobileprice ~ cell.phones$q17d))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| cell.phones$q17d | 1 | 2612 | 2611.8 | 3.402 | 0.0655 . |
| Residuals | 775 | 594952 | 767.7 | | |

---

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

*1475 observations deleted due to missingness*

Παρατηρούμε τιμή p-value> 0.05 οπότε και δε μπορούμε να απορρίψουμε την μηδενική υπόθεση. Επομένως η μεταβλητή αυτή δεν έχει κάποια συσχέτιση με τη μεταβλητή **"ΤΙΜΗ ΚΙΝΗΤΟΥ"**

**2.”Θεωρείτε αγένεια το να σας διακόπτει κανείς σε μία συζήτηση ώστε να μιλήσει στο κινητό του;”**

```
> summary(aov(cell.phones$mobileprice ~ cell.phones$q22e))
```

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *cell.phones$q22e* | *2* | *4264* | *2132* | *1.084* | *0.338* |
| *Residuals* | *1903* | *3743377* | *1967* | | |

*346 observations deleted due to missingness*

Παρατηρούμε τιμή p-value> 0.05 οπότε και δε μπορούμε να απορρίψουμε την μηδενική υπόθεση. Επομένως η μεταβλητή αυτή δεν έχει κάποια συσχέτιση με τη μεταβλητή “ΤΙΜΗ ΚΙΝΗΤΟΥ”

## Κατασκευή μοντέλου και ανάλυση συνδιακύμανσης (ANCOVA)

```
# Ancova model structure using two variables , Age (numerical) & sex(categorical)

> model=lm(cell.phones$mobileprice ~ cell.phones$age * cell.phones$sex )

> summary.aov(model)
```

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *cell.phones$age* | *1* | *597143* | *597143* | *356.256* | *< 2e-16 **** |
| *cell.phones$sex* | *1* | *28728* | *28728* | *17.139* | *3.6e-05 **** |
| *cell.phones$age:cell.phones$sex* | *1* | *1745* | *1745* | *1.041* | *0.308* |
| *Residuals* | *2197* | *3682530* | *1676* | | |

*---*

*Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*51 observations deleted due to missingness*

Παρατηρούμε ότι στο συγκεκριμένο μοντέλο, μεταξύ κόστους κινητού, ηλικίας και φύλου, η αλληλεπίδραση (ηλικίας και φύλου) δεν είναι στατιστικά σημαντική αφού p = 0.308 >0.05 άρα μπορούμε να την απομακρύνουμε απο το μοντέλο.

```
# Simplification of our model

> model2=lm(cell.phones$mobileprice ~ cell.phones$age + cell.phones$sex )

> anova(model,model2)
```

*Analysis of Variance Table*

*Model 1: cell.phones$mobileprice ~ cell.phones$age * cell.phones$sex*

*Model 2: cell.phones$mobileprice ~ cell.phones$age + cell.phones$sex*

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 2197 | 3682530 | | | | |
| 2 | 2198 | 3684275 | -1 | -1744.6 | 1.0409 | **0.3077** |

Στο απλούστερο μοντέλο, όπου έχουμε αφαιρέσει την αλληλεπίδραση παρατηρούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά από το προηγούμενο αφού p=0.3>0.05.

#reviewing our simpler model

> summary.lm(model2)

*Call:*

*lm(formula = cell.phones$mobileprice ~ cell.phones$age + cell.phones$sex)*

*Residuals:*

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -184.352 | -25.511 | 0.167 | 25.240 | 135.117 |

*Coefficients:*

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 43.20433 | 2.75723 | 88.21 | < 2e-16 *** |
| cell.phones$age | -0.87492 | 0.04759 | -18.38 | < 2e-16 *** |
| cell.phones$sexMale | 7.32148 | 1.76851 | 4.14 | 3.61e-05 *** |

---

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

*Residual standard error: 40.94 on 2198 degrees of freedom*

*(51 observations deleted due to missingness)*

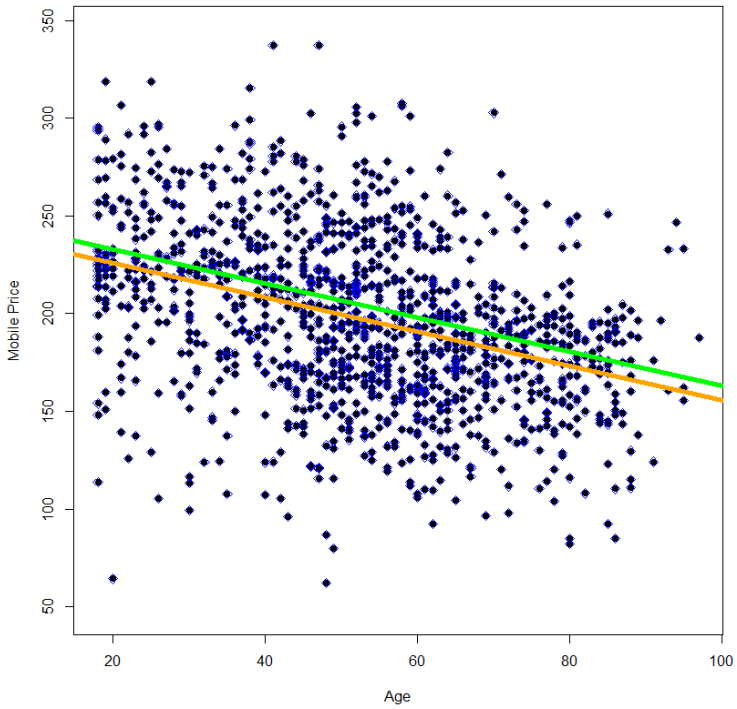
*Multiple R-squared:* ***0.1452****, Adjusted R-squared: 0.1444*

*F-statistic: 186.7 on 2 and 2198 DF, p-value:* ***< 2.2e-16***

Το μοντέλο έχει χαμηλή ερμηνευτική ισχύ, που είναι λογικό καθώς λείπουν αρκετοί παράγοντες, και το R-squared προκύπτει ότι εξηγεί το **14.4%** της μεταβλητότητας της εξαρτημένης. Η συμβολή των μεταβλητών age και sex είναι στατιστικά σημαντική στο μοντέλο καθώς το **p<<0.05**. Από την στήλη estimate παίρνουμε τα εξής δεδομένα που έχουμε ήδη ανακαλύψει παραπάνω. Ο αρνητικός συντελεστής του age μας δείχνει ότι όσο αυξάνεται η ηλικία μειώνεται το κόστος αγοράς του κινητού κατά -0.87 μονάδες. Αντίστοιχα ο θετικός συντελεστής του sexMale που παρατηρούμε στην στήλη Estimate μας δείχνει ότι η μεταβολή από το επίπεδο Female στο Male έχει διαφορά μέσου κόστους τιμής κινητού της τάξης των 7.3 μονάδων.

Οι ευθείες που προκύπτουν φαίνονται στο παρακάτω σχήμα.

```
# visualization of our model

> mobileprices_sex = split(cell.phones$mobileprice,cell.phones$sex)

> age_sex = split(cell.phones$age,cell.phones$sex)

> plot(cell.phones$age,cell.phones$mobileprice,type="n",ylab="Mobile Price",xlab="Age")

> points(age_sex[[1]],mobileprices_sex[[1]],pch=16)

> points(age_sex[[1]],mobileprices_sex[[1]],pch=5,col="blue")

> abline(243.20,-0.874,col="orange",lwd=5)

> abline(243.20 + 7.32,-0.874 , col="green",lwd=5)
```



Εξαιτίας της απλότητας του μοντέλου δεν μπορεί να ερμηνευθεί μεγάλο ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής κάτι που φαίνεται τόσο από το διάγραμμα όσο και από την πολύ χαμηλή τιμή του R-squared (14%) που είδαμε και παραπάνω.

## Βελτίωση του μοντέλου με στόχο την αύξηση της ερμηνευτικής ισχύς του (R-squared).

Έλεγχος μοντέλου που να περιέχει μόνο τη συσχέτιση της τιμής με την ηλικία των αγοραστών.Αντί της εξαρτημένης μεταβλητής , λαμβάνεται ο λογάριθμος αυτής.

```
> model2 = lm(log(mobileprice)~ age)

> summary.lm(model2)
```

*Call:*

*lm(formula = log(mobileprice) ~ age)*

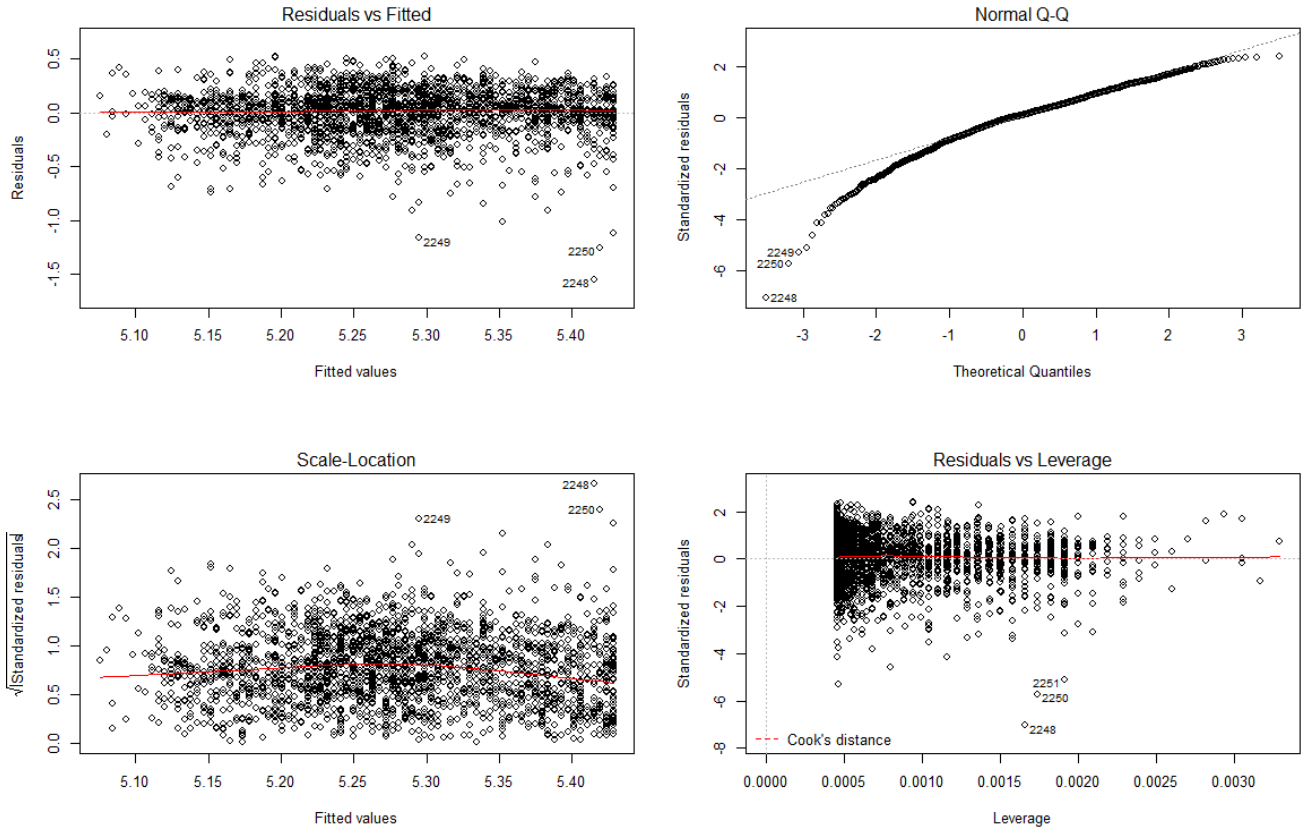
*Residuals:*

| *Min* | *1Q* | *Median* | *3Q* | *Max* |
|---|---|---|---|---|
| *-1.5481* | *-0.1145* | *0.0221* | *0.1404* | *0.5287* |

*Coefficients:*

| | *Estimate* | *Std. Error* | *t value* | *Pr(>\|t\|)* |
|---|---|---|---|---|
| *(Intercept)* | *5.5089178* | *0.0137956* | *399.32* | *<2e-16 **** |
| *age* | *-0.0044672* | *0.0002545* | *-17.55* | *<2e-16 **** |
| *Residual standard error: 0.22 on 2199 degrees of freedom*<br>*(51 observations deleted due to missingness)* | | | | |
| *Multiple R-squared* | *Adjusted R-squared* | *F-statistic* | *p-value* | |
| ***0.1229*** | *0.1225* | *08.1 on 1 and 2199 DF* | ***< 2.2e-16*** | |

Παρατηρούμε πως το μοντέλο μας δεν έχει ισχυρή ερμηνευτική ισχύ καθώς εξηγεί μονό το 12.29 % της μεταβλητότητας της εξαρτημένης μεταβλητής. Ας ελέγξουμε και της 4 συναρτήσεις που περιγράφουν οπτικά το μοντέλο για να σχολιάσουμε το τι παρατηρούμε.




Παρατηρούμε πως δεν υπάρχει κάποιο συγκεκριμένο σχήμα στα κατάλοιπα, εμφανίζουν τυχαιότητα και επίσης απο το διάγραμμα QQ φαίνεται πως δεν ακολουθούν την κανονική κατανομή.

## Πίνακες συχνοτήτων για κατηγορικές μεταβλητές.

```
> freq(cell.phones$educ)
```

```
> freq(cell.phones$q17b)
```

```
> freq(cell.phones$q14a)
```

```
> freq(cell.phones$usr_r)
```

Μεταβλητή επιπέδου εκπαίδευσης :

| | ***Freq*** | ***% Valid*** | ***% Valid Cum.*** | ***% Total*** | ***% Total Cum.*** |
|---|---|---|---|---|---|
| *MID-HIGH LEVEL* | *1015* | *45.39* | *45.39* | *45.07* | *45.07* |
| *MID LEVEL* | *725* | *32.42* | *77.82* | *32.19* | *77.26* |
| *LOW LEVEL* | *214* | *9.57* | *87.39* | *9.50* | *86.77* |
| *HIGH LEVEL* | *282* | *12.61* | *100.00* | *12.52* | *99.29* |
| *<NA>* | *16* | | | *0.71* | *100.00* |
| *Total* | *2252* | *100.00* | *100.00* | *100.00* | *100.00* |

Παρατηρήσεις : Η πλειοψηφία των ερωτηθέντων έχουν μεσαίο ή μεσαίο προς ανώτερο επίπεδο εκπαίδευσης. Δεν παρατηρείται μεγάλη κατανομή δειγμάτων στις ακραίες τιμές HIGH LEVEL και LOW LEVEL.

Μεταβλητή q17b: Χρησιμοποιείτε το κινητό για να ανεβάσετε φωτογραφίες/ βίντεο online; ( 1 = "Yes, do this"/2 = "No, do not do this/Have not done this"/3 = "Cell phone can’t do this")

| | ***Freq*** | ***% Valid*** | ***% Valid Cum.*** | ***% Total*** | ***% Total Cum.*** |
|---|---|---|---|---|---|
| *Cell phone can’t do this* | *5* | *0.64* | *0.64* | *0.22* | *0.22* |
| *No, do not do this/ Have not done this* | *549* | *70.47* | *71.12* | *24.38* | *24.60* |
| *Yes, do this* | *225* | *28.88* | *100.00* | *9.99* | *34.59* |
| *<NA>* | *1473* | | | *65.41* | *100.00* |
| *Total* | *2252* | *100.00* | *100.00* | *100.00* | *100.00* |

Παρατήρηση: υπάρχουν πολύ λίγες παρατηρήσεις για το πρώτο επίπεδο της μεταβλητής. Αυτό οδηγεί στο συμπέρασμα πως δεν έχει μεγάλη ερμηνευτική ισχύ το επίπεδο αυτό της μεταβλητής.

Μεταβλητή : Χρησιμοποιείται το κινητό σας για να στείλετε e-mail?

| | ***Freq*** | ***% Valid*** | ***% Valid Cum.*** | ***% Total*** | ***% Total Cum.*** |
|---|---|---|---|---|---|
| *No* | *1371* | *71.52* | *71.52* | *60.88* | *60.88* |
| *Yes* | *546* | *28.48* | *100.00* | *24.25* | *85.12* |
| *<NA>* | *335* | *-* | *14.88* | *100.00* | *-* |
| *Total* | *2252* | *100.00* | *100.00* | *100.00* | *100.00* |

Παρατήρηση : Παρατηρούμε πως οι περισσότεροι δεν χρησιμοποιούν το κινητό τους για να στείλουν e-mail. Ακόμη παρατηρούμε πως ένα σημαντικό ποσοστό ερωτηθεντων επέλεξαν να μην απαντήσουν την ερώτηση.

Μεταβλητή: Περιοχή κατοικίας (1 = Urban, 2 = Suburban, 3 = Rural)

| | ***Freq*** | ***% Valid*** | ***% Valid Cum.*** | ***% Total*** | ***% Total Cum.*** |
|---|---|---|---|---|---|
| *Rural* | *463* | *21.07* | *21.07* | *20.56* | *20.56* |
| *Suburban* | *1135* | *51.66* | *72.74* | *50.40* | *70.96* |
| *Urban* | *599* | *27.26* | *100.00* | *26.60* | *97.56* |
| *<NA>* | *55* | | *2.44* | *100.00* | |
| *Total* | *2252* | *100.00* | *100.00* | *100.00* | *100.00* |

```
> describe(cell.phones$mobileprice)
```

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 1 | 2252 | 201.3 | 44.28 | 201.4 | 201.1 | 44.9 | 47.8 | 345.3 | 297.5 | 0.01 | -0.11 | 0.93 |

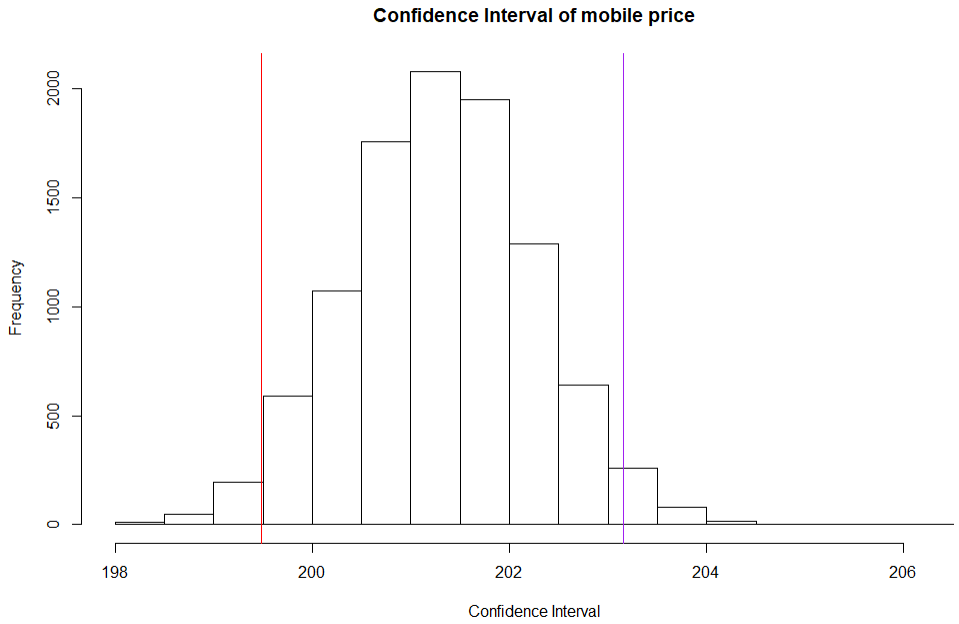
Ένας τρόπος να περιγράψουμε κάποιες από τις σημαντικές μεταβλητές είναι η describe.

Παρατηρούμε για την μεταβλητή MobilePrice πως έχει μέση τιμή 201 και διασπορά 44.28. Η χαμηλότερη απο την υψηλότερη τιμή που παρατηρήθηκε έχει διαφορά 297.5 μονάδες. Επίσης, παρατηρούμε ότι δεν υπάρχει ασυμμετρία και κύρτωση όπως αναμέναμε μιας και σε προηγούμενους ελέγχους είδαμε πως η εξαρτημένη μεταβλητή ακολουθεί κανονική κατανομή.

```
> describe(cell.phones$age)
```

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 1 | 2201 | 50.98 | 18.43 | 52 | 50.86 | 19.27 | 18 | 97 | 79 | 0 | -0.8 | 0.39 |

```
> b =numeric(10000)
> for (i in 1:10000) {b[i] =
  mean(sample(cell.phones$mobileprice, length(cell.phones$mobileprice), replace = T))}
> hist(b,main = "Confidence Interval of mobile price", xlab = "Confidence Interval")
> quantile(b,c(0.025,0.975))
> abline(v=quantile(b,0.025),col="red")
> abline(v=quantile(b,0.975),col="purple")
```




| *2.5%* | *97.5%* |
|---|---|
| *199.4867* | *203.1607* |

Τρέξαμε έναν bootstrap αλγόριθμο και παρατηρήσαμε πως η μέση τιμή της τιμής αγοράς ενός τηλεφώνου είναι με διάστημα εμπιστοσύνης 95% μεταξύ των τιμών 199.48 και 203.16.

```
> model3 = lm(cell.phones$mobileprice~cell.phones$sex+cell.phones$q17b+
        cell.phones$q17g+cell.phones$educ+
        cell.phones$sex*cell.phones$age+cell.phones$q17e)
> summary.lm(model3)
> model.dim = ols_step_both_p(model3)
> model.dim
> summary(model.dim$model)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -36.753 | -11.942 | -1.404 | 6.508 | 83.249 |

Coefficients:

| | Estimate | Std. | Error t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 82.527 | 14.774 | 5.586 | 3.25e-08 |
| cell.phones$sexMale | -1.29357 | 3.5634 | -0.363 | 0.716698 |
| cell.phones$q17bNo, do not do this/Have not done this | 201.8578 | 8.5970 | 23.480 | < 2e-16 |
| cell.phones$q17bYes, do this | 164.5926 | 8.67341 | 18.977 | < 2e-16 |
| cell.phones$q17gNo, do not do this/Have not done this | -40.25953 | 14.80018 | -2.720 | 0.006 ** |
| cell.phones$q17gYes, do this | -39.62109 | 14.8379 | -2.670 | 0.007 ** |
| cell.phones$educMID LEVEL | 0.14452 | 1.52895 | 0.095 | 0.9247 |
| cell.phones$educLOW LEVEL | 3.63956 | 2.56911 | 1.417 | 0.1569 |

| | | | | |
|---|---|---|---|---|
| cell.phones$educHIGH LEVEL | -1.93700 | 1.91819 | -1.010 | 0.3129 |
| cell.phones$age | -0.23628 | 0.06060 | -3.899 | 0.0001 *** |
| cell.phones$q17eNo | 23.58084 | 14.7365 | 1.600 | 0.109983 |
| cell.phones$q17eYes | 25.50193 | 14.734 | 1.731 | 0.0838 . |
| cell.phones$sexMale:cell.phones$age | 0.06959 | 0.08038 | 0.866 | 0.386870 |
| ********** | ********** | ********** | ********** | ********** |
| *Residual standard error: 17.66 on 752 degrees of freedom*<br>*(1487 observations deleted due to missingness)* | | | | |
| ***Multiple R-squared*** | *Adjusted R-squared:* | *F-statistic* | ***p-value*** | |
| ***0.6057*** | *0.5994* | *96.26 on 12 and 752 DF* | ***< 2.2e-16*** | |

*Stepwise Selection Summary*

| *Step* | *Variable* | *Added/Removed* | *Adj. R-Squared* | *R-Square* | *C(p)* | *AIC* | *RMSE* |
|---|---|---|---|---|---|---|---|
| *1* | *cell.phones$q17b* | *addition* | *0.576* | *0.575* | *40.7340* | *6727.7197* | *18.1022* |
| *2* | *cell.phones$age* | *addition* | *0.598* | *0.596* | *7.4910* | *6600.9349* | *17.7191* |
| *3* | *cell.phones$q17g* | *addition* | *0.601* | *0.598* | *4.2550* | *6599.6918* | *17.6819* |

Συμπεράνουμε πως, από όλες τις μεταβλητές και τις αλληλεπιδράσεις τους μόνο οι q17b , q17g και age είναι σημαντικές για τη δημιουργία ενός αποτελεσματικού μοντέλου.

*Residuals:*

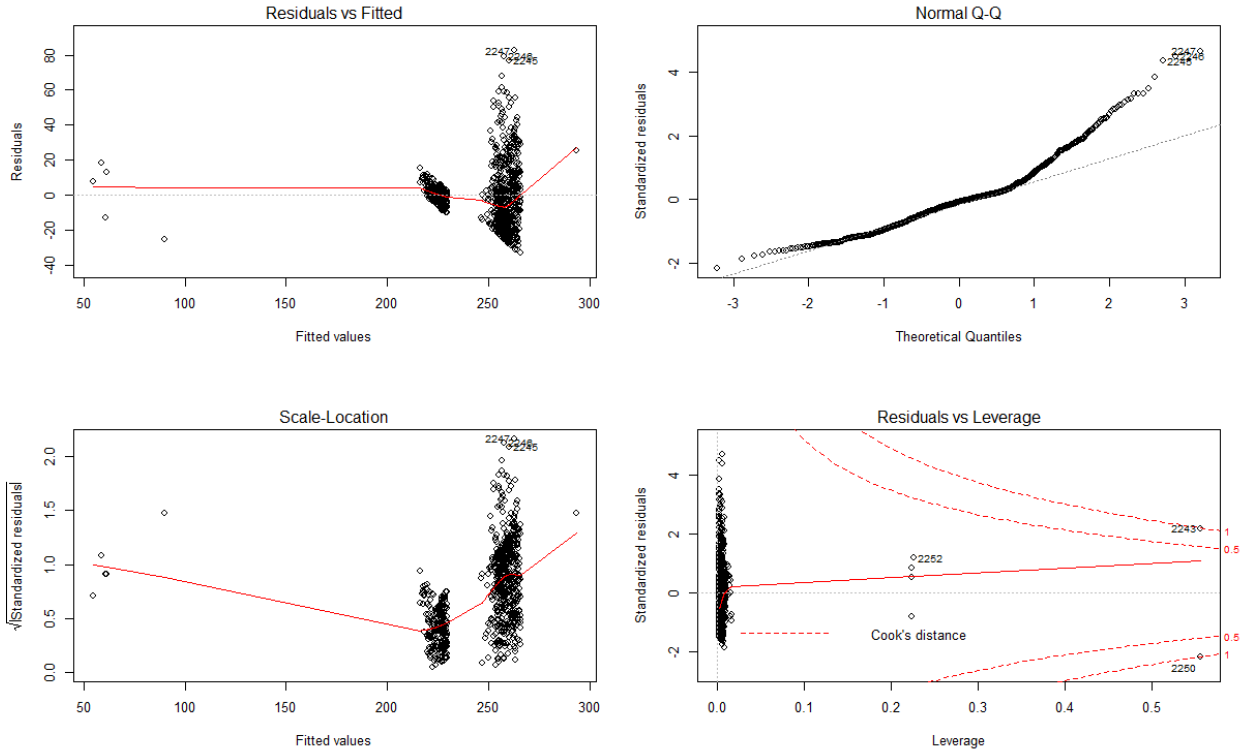
| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| *-32.805* | *-11.654* | *-1.322* | *5.728* | *82.414* |

Coefficients:

| | Estimate | Std. | Error t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 94.6443 | 13.1964 | 7.172 | 1.75e-12 *** |
| *** | *** | *** | *** | *** |
| cell.phones$q17<br>No, do not do this/Have not done this | 203.0733 | 8.3788 | 24.237 | < 2e-16 *** |

| | | | | |
|---|---|---|---|---|
| cell.phones$q17b Yes, do this | 166.6267 | 8.4412 | 19.740 | < 2e-16 *** |
| cell.phones$age | -0.2349 | 0.0430 | -5.463 | 6.33e-08 *** |
| cell.phones$q17g No, do not do this/Have not done this | -28.8930 | 13.2208 | -2.185 | 0.0292 * |
| cell.phones$q17g Yes, do this | -27.7837 | 13.2403 | -2.098 | 0.0362 * |
| *Residual standard error: 17.68 on 762 degrees of freedom* *(1484 observations deleted due to missingness)* | | | | |
| ***Multiple R-squared*** | *Adjusted R-squared* | *F-statistic* | ***p-value*** | |
| ***0.6006*** | *0.598* | *229.2 on 5 and 762 DF* | ***< 2.2e-16*** | |

Το τελικό μας μοντέλο αποτελείται από 3 μεταβλητές, η μία εκ των οποίων είναι κατηγορική με 3 επίπεδα και εξηγεί το 60.06% της μεταβλητότητας της εξαρτημένης μεταβλητής mobileprice. Όλες οι ανεξάρτητες μεταβλητές είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 5%.




Επειδή οι ανεξάρτητες μεταβλητές μας είναι κατηγορικές δεν μας βοηθούν στην ερμηνεία τα πρώτα δύο διαγράμματα, ωστόσο απο το qq plot φαίνεται πως τα κατάλοιπα δεν ακολουθούν την κανονική κατανομή. Τέλος, παρατηρούμε 2 leverage points τα οποία ιδανικά θα έπρεπε να αφαιρεθούν για να μην επηρεάζουν αρνητικά το μοντέλο μας.

**Κλείνοντας,** έχοντας αναλύσει το dataset που μας δόθηκε παρατηρήσαμε ότι είναι αρκετές οι μεταβλητές που μπορούν να επηρεάσουν την τελική τιμή αγοράς ενός κινητού τηλεφώνου. Ξεκινήσαμε με περιγραφική στατιστική για να ανακαλύψουμε την δομή των συνολικών δεδομένων και για να αποκτήσουμε καλύτερη εικόνα της εξαρτημένης μεταβλητής. Βάση ιστογράμματος, παρατηρήθηκε ότι ακολουθείται κανονική κατανομή, κάτι που επιβεβαιώθηκε και απο τα QQ-plots και τα Shapiro-Wilk test.

Έγιναν διαγράμματα για τον εντοπισμό σχέσεων μεταξύ των διαφόρων ανεξάρτητων μεταβλητών με την εξαρτημένη. Παρατηρήθηκε ότι η τιμή αγοράς επηρεάζεται τόσο από το εισόδημα του εκάστοτε αγοραστή όσο και από τις λειτουργίες που επιθυμει να διαθέτει η συσκευή του.

Πραγματοποιήσαμε μερικές επιπλέον στατιστικές μελέτες (ανάλυση διακύμανσης ANOVA) για να επιβεβαιώσουμε αυτά που παρατηρήσαμε απο τα διαγράμματα. Συνεχίζοντας, βάση των αποτελεσμάτων των ελέγχων ANOVA εκτελέσαμε και μερικά PostHocTest για να δούμε ποιες μεταβλητές επηρεάζουν μέσω αλληλεπιδράσεων την τιμή του κινητού. Έγιναν επίσης έλεγχοι συσχετίσεων για να βρεθούν σχέσεις μεταξύ των μεταβλητών. Τέλος, ξεκινήσαμε από ένα σχετικά μοντέλο και προχωρήσαμε στην κατασκευή ενός πιο σύνθετου μέσω Stepwise Regression.