**Софийски университет „Св. Климент Охридски“**

*Факултет по математика и информатика*

*Специалност: “Информационни системи”*

*Курс: 2, Група: 3*

*Дисциплина: “Статистика и емпирични методи - практикум”*

**КУРСОВ ПРОЕКТ**

Тема: Отношението на пазара на смартфоните

***Изготвил:***

Никола Петров Кирилов(ФН: 71986)

***Преподавател:***

Методий Кандиларов

София 2021

**1. Избор на данни и цели на проекта**

Пазарът на смартфони често е много динамичен и пряко свързан с използваемостта на даден телефон и отношението на собственика му към него. Много често малки търговци и производители имат нужда от качествен анализ на пазара, за да могат адекватно да отговарят към изискванията на потребителите, затова моята тема е отношението в пазара на смартфоните.

Направих своя анкета и я пратих на познати, общият брой на запитаните хора е 117 човека. Анкетата може да бъде разгледана на следния линк:

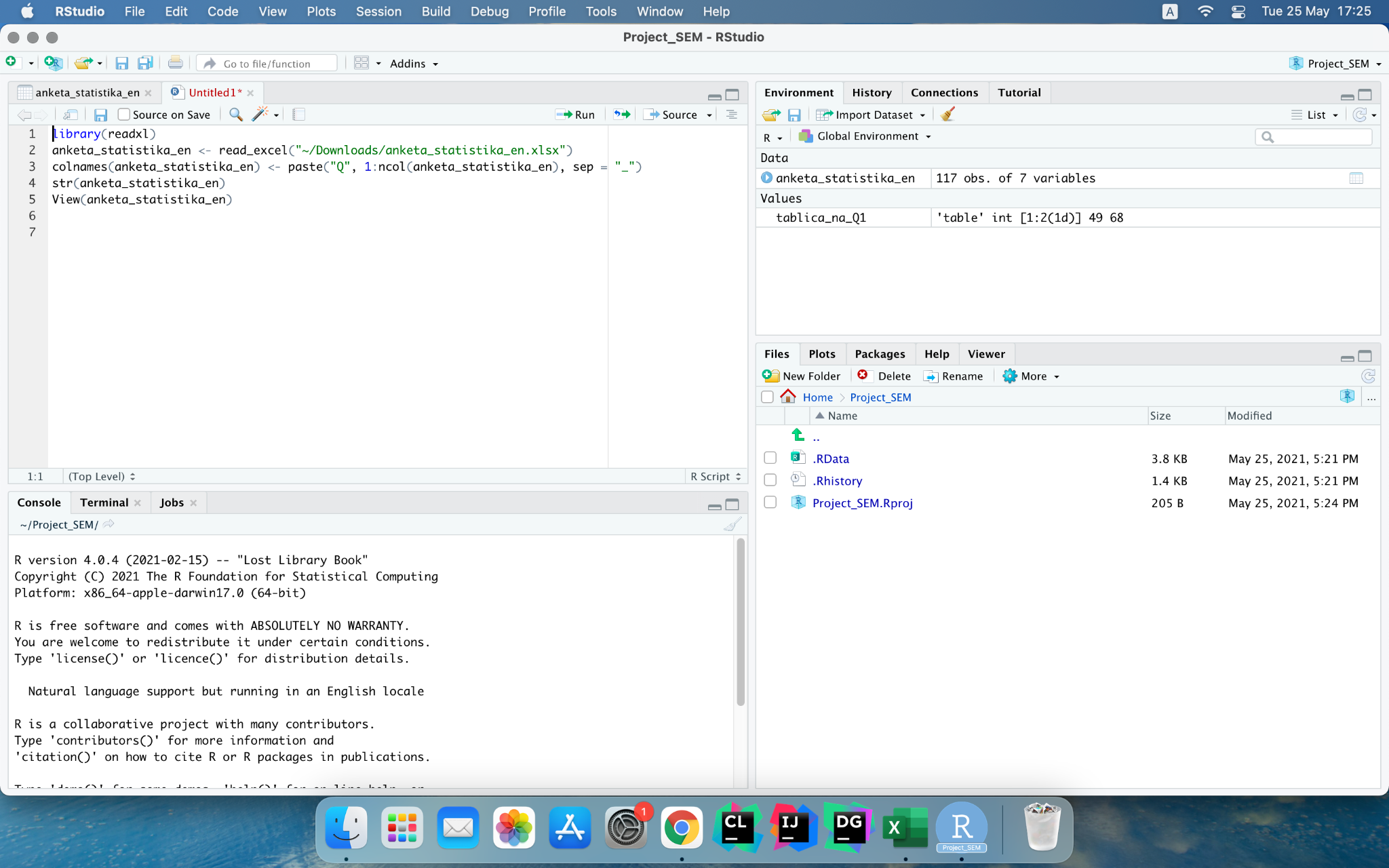
<https://docs.google.com/forms/d/15Z7_k8ZPBk39syInUQuf49_W2h--PFjSdUyBr6m4Dxc/edit>

Ще направя описателен анализ върху получените резултати от запитването съгласно изискванията: анализ на едномерна променлива, анализ на взаимодействието между две променливи.

**2. Въвеждане на данните и анализ на едномерните променливи**

За да въведем толкова много данни ще ни е нужна библиотеката “readxl”, която ни позволява да въведем всички данни по колони в нашата програма без да се налага да ги пишем ръчно.Тъй като R Studio даде проблем с кодировката на данните на кирилица, преведох цялата таблица на английски език, което ще ми позволи да ги вмъкна, което става със следния код:

\***ФРАГМЕНТИТЕ ОТ КОД В ПРОЕКТА ЩЕ БЪДАТ ОЦВЕТЕНИ В ЗЕЛЕНО**



***Въпрос 1: На колко години си?***

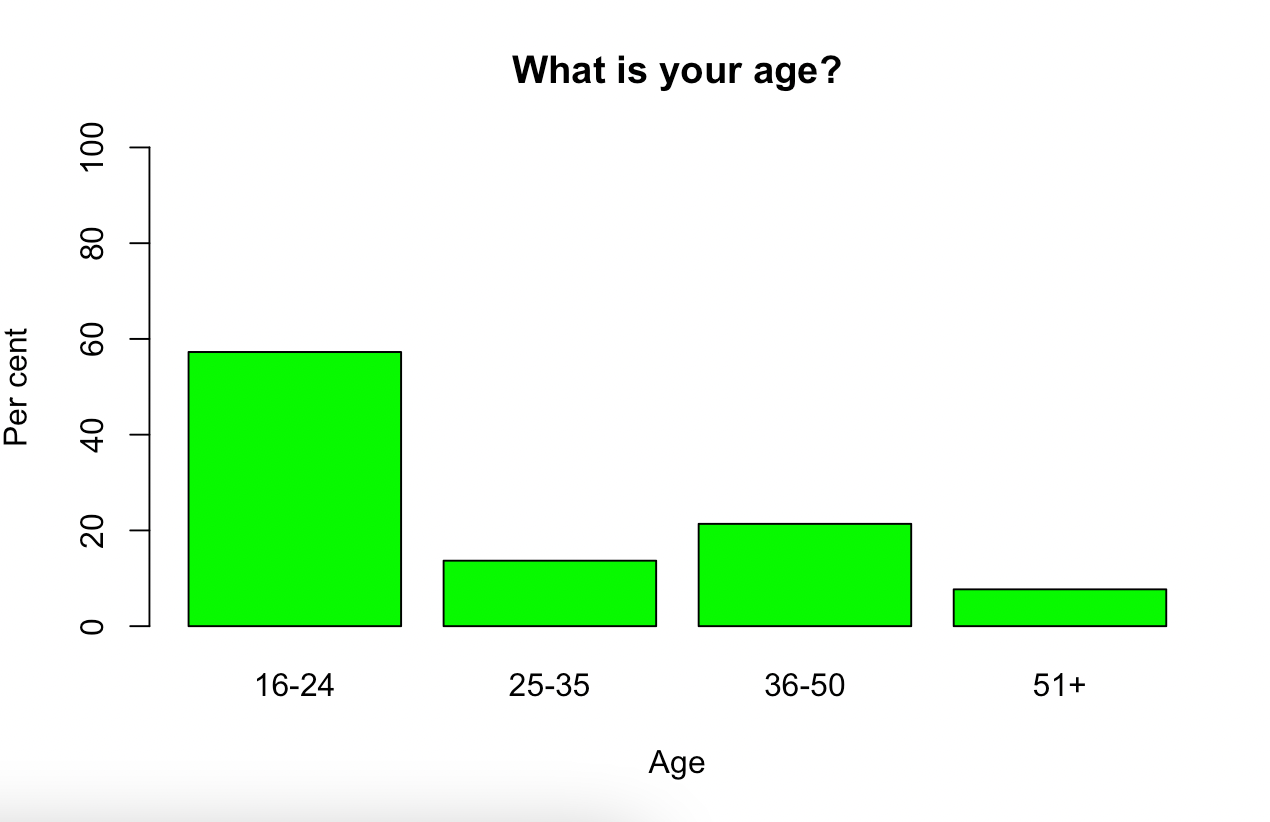
Възможните отговори на въпросът описват на категорна променлива, в която годините са представени като интервали. Отново ще използваме функцията **table().**

table\_age <- table(anketa\_statistika\_en$Q\_2)

prop\_table\_age <- prop.table(table\_age)

За следният анкетен въпрос ще използвам **barplot**, с който много ясно се вижда разпределението на тази категорна величина.

barplot(height = (prop\_table\_age)\*100, col = "green",xlab = "Age", ylab = "Per cent", ylim = c(0,100), main = "What is your age?")



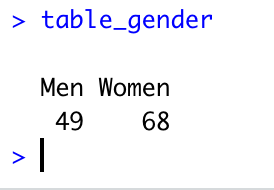
Очевидното от диаграмата е, че най-голямата част от анкетираните (близо 60%)са хората във възрастова група 16-24, които са и най-активната и бързо развиваща се част. Около 23% са във възрастовата група 36-50, около 12% са 25-35, и около 5%

***Въпрос 2: Какъв е вашият пол?***

При категорните променливи най-добре се вижда тяхната честота посредством таблици, за това използвам функцията **table()**, а с **prop.table()** взимам процентното разпределение.

table\_gender <- table(anketa\_statistika\_en$Q\_1)

table\_ gender



Сега ще използваме **piechart**, за да видим графично процентно разпределение:

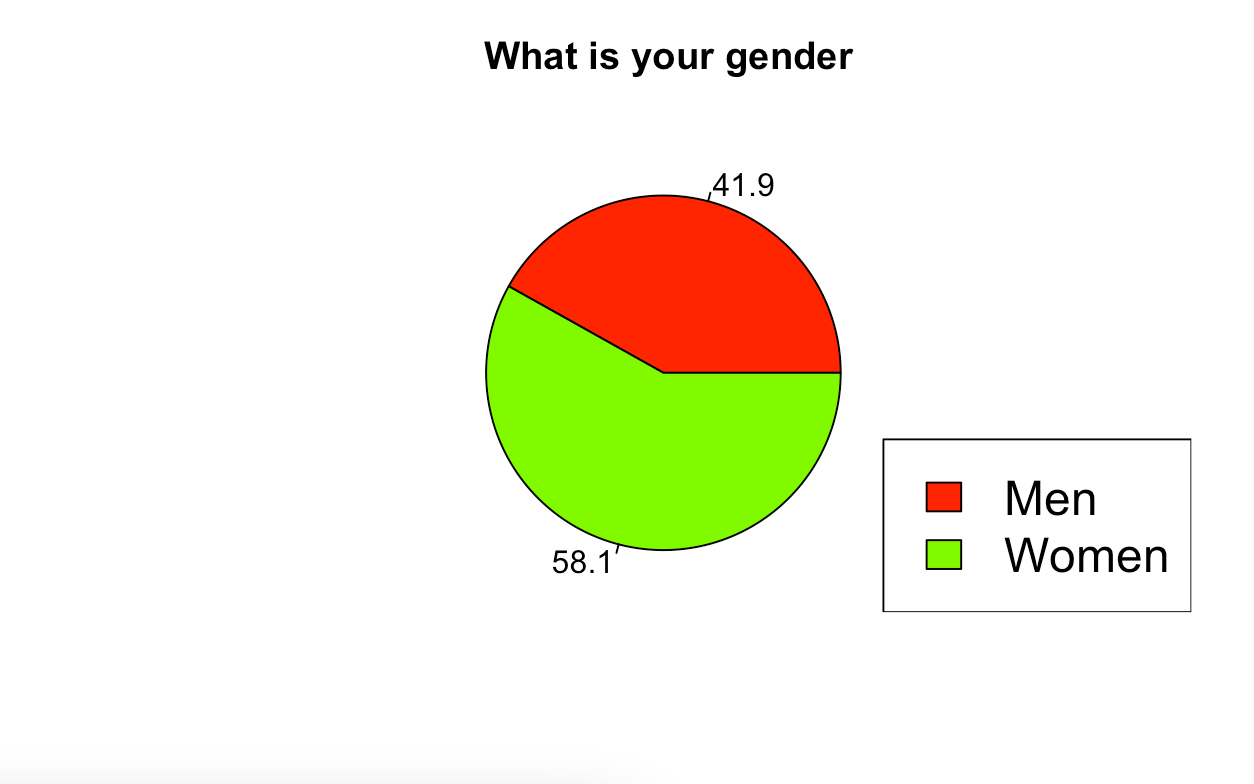
piepercent\_gender <- round(100\*table\_gender/sum(table\_gender), 1)

pie(table\_gender, labels = piepercent\_gender, main = " What is your gender", col =

rainbow(n = length(table\_gender)\*2))

legend(x = "bottomright", legend = c("Men", "Women"),

cex = 1.5, fill = rainbow(length(table\_gender) \*2))



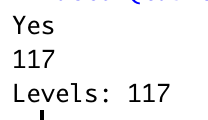
Става ясно, че близо 58,1% от анкетираните са жени, а останалите 41,9% са мъже.

***Въпрос 3: Притежаваш ли смартфон?***

Може би най-незаинтригуващият въпрос от всички. За него имам 117 отговора ,,Да”, които смятам да представя с **factor()**, който ще го раздели на групи, а в нашия случай ще ни помогне с обобщаването на данните. Очевидното в случая е, че всички от анкетираните притежават смартфон, което е знак, че целите на анализа са оправдани.

> table\_smartphone <- table(anketa\_statistika\_en$Q\_3)

> factor(table\_smartphone)



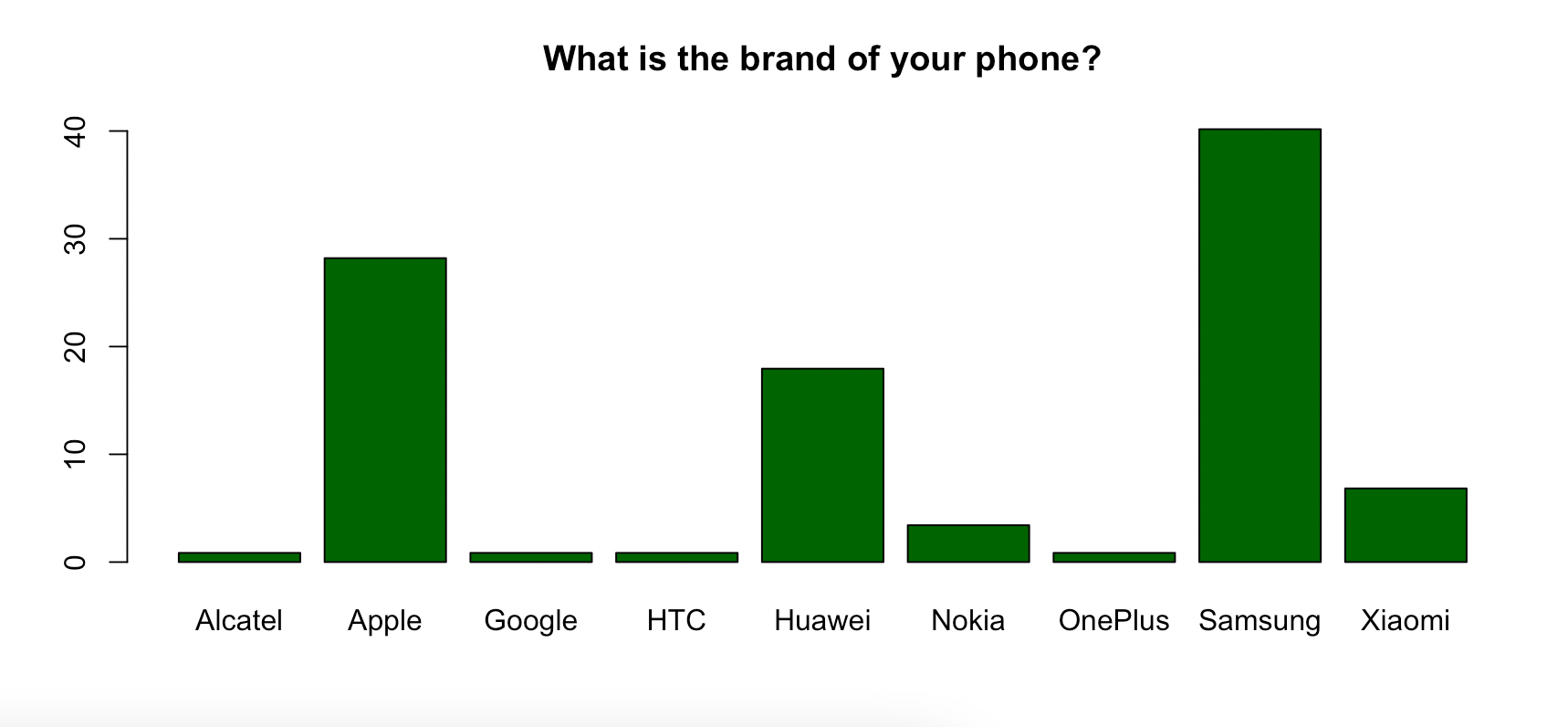
***Въпрос 4: Каква марка е?***

Отново категорна променлива, като пак ще използвам **table()** и **prop.table()** за по-добро онагледяване на данните.

table\_brands <- table(anketa\_statistika\_en$Q\_4)

prop\_table\_brands <- prop.table(table\_brands)

barplot(height =prop\_table\_brands\*100, col = "darkgreen", , main = "What is the brand of your phone?")



От така дадените данни можем да кажем, че Samsung е първенец с 48 броя, Apple са втори с 34, след тях се нареждат Huawei с 21, Xiaomi са четвърти с 8, Nokia с 4, а останалите марки имат по 1 брой.

Нека сега покажем и **piechart** диаграма за по-лесна разбираемост на процентното отношение на пазарния дял(взето от моята анкета)

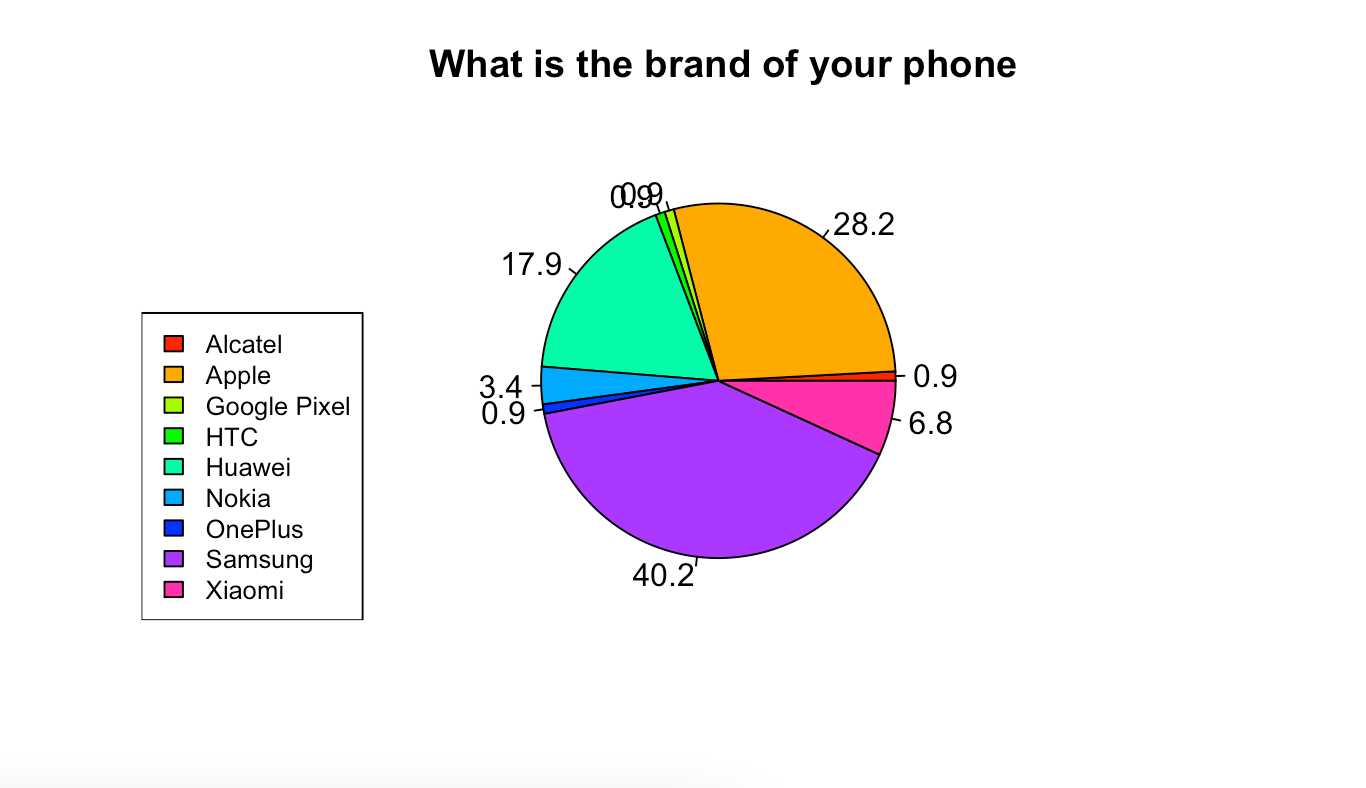
piepercent\_brands <- round(100\*table\_brands/sum(table\_brands), 1)

pie(table\_brands, labels = piepercent\_brands, main = " What is the brand of your phone", col =

rainbow(n = length(table\_brands)))

legend(x = "bottomleft", legend = c("Alcatel","Apple", "Google Pixel", "HTC", "Huawei", "Nokia","OnePlus", "Samsung", "Xiaomi"),

cex = 0.8, fill = rainbow(length(table\_brands)))



***Въпрос 5: Колко инча е вашият смартфон?***

Отговорите на този въпрос представляват числова непрекъсната променлива. Започвам с въвеждането на всички данни във вектор.

smartphone\_size <- smartphone\_size <- anketa\_statistika\_en$Q\_8

След това искам да намеря модата, а именно най-срещаната стойност сред изброените.

modeFunction <- function(x)

{

res\_table <- table(x)

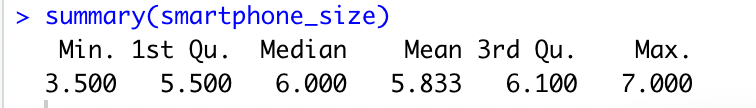
return(names(res\_table)[res\_table == max(res\_table)])

}

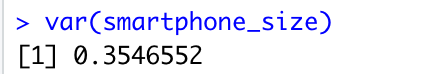
modeFunction(smartphone\_size)



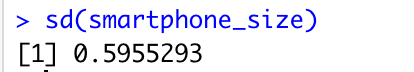
След като приложа функцията виждам, че модата е 5.5, което значи, че най-предпочитаният размер на екрана от потребителите е 5.5 инча.

Сега ще приложа **summary** - описателна статистика за централните тенденции на извадката.

Сега намирам **вариацията**(**дисперсията**)



След това намирам **стандартно** **отклонение**.

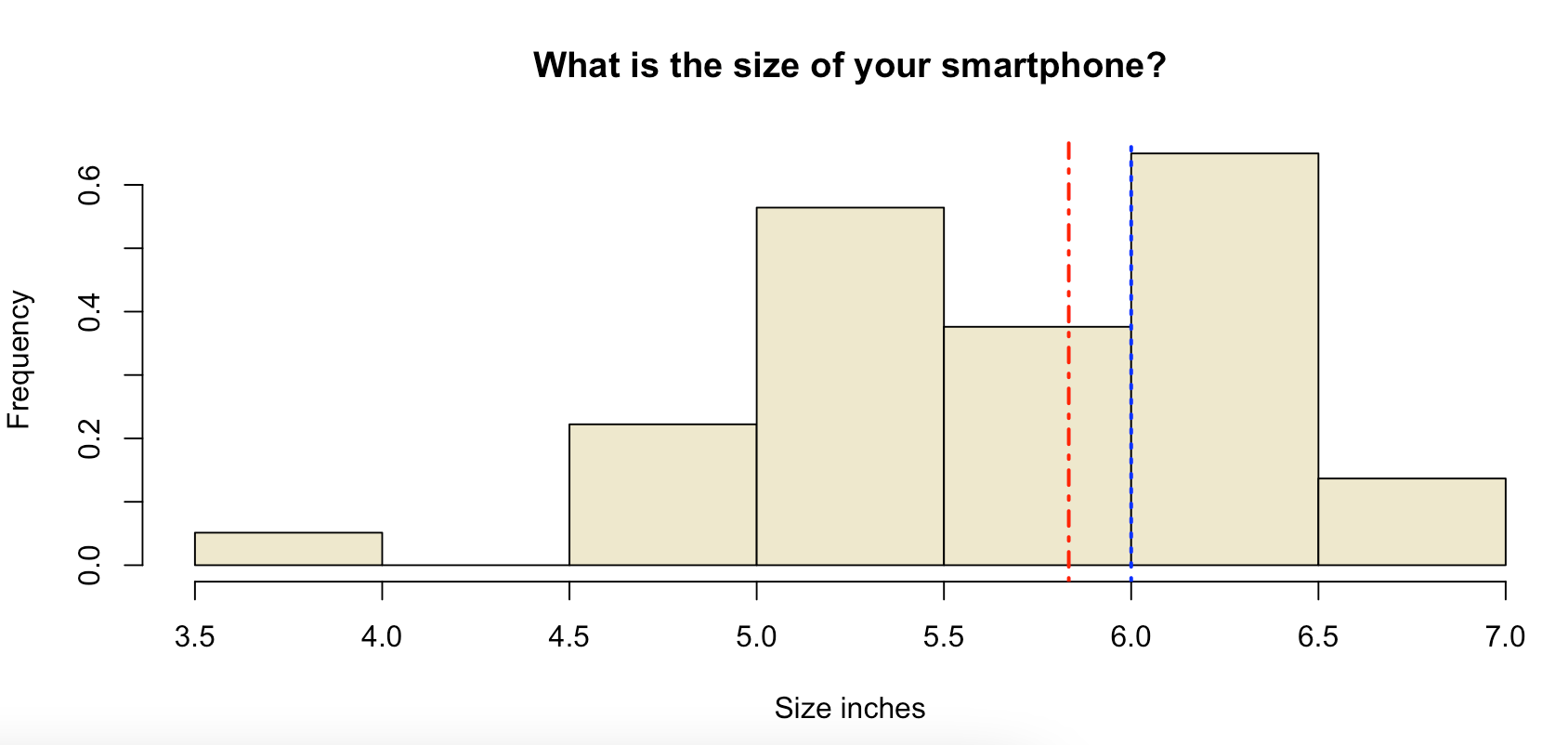


Нека сега видим графичното представяне на данните с хистограма.

hist(smartphone\_size, main = "What is the size of your smartphone?", xlab = "Size inches", ylab = "Frequency", col = "cornsilk2", prob = T)

abline(v = mean(smartphone\_size), lwd = 2, lty = 4, col = "red")

abline(v = median(smartphone\_size), lwd = 2, lty = 3, col = "blue")

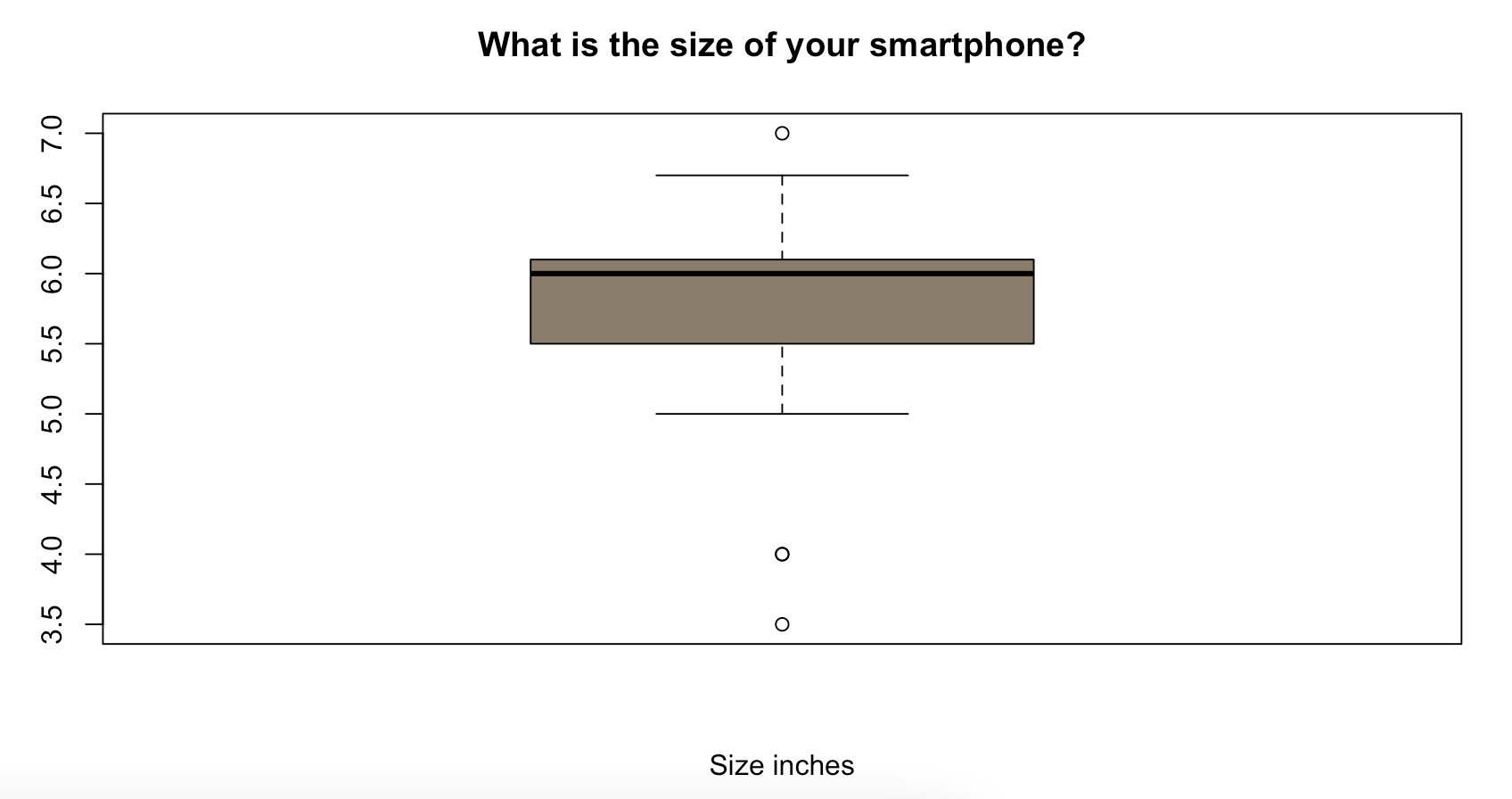


На графиката в червено е начертана медианата, а в синьо е начертана средната стойност. Средната стойност е **5.833333,** а медианата е точно **6**

Сега ще конструираме **boxplot**, за потенциално откриване на **outliers.**

boxplot(smartphone\_size, col = "bisque4", main = "What is the size of your smartphone?",

xlab = "Size inches")



Видно от **boxplot** диаграмата засякохме 3 **outlier-и.**

***Въпрос 6:Кое е водещ фактор за Вас при закупуването на нов смартфон?***

answer\_1 = "Battery"

answer\_2 = "Camera"

answer\_3 = "Speed"

answer\_4 = "Brand"

answer\_5 = "Memory"

answer\_6 = "Durability"

answer\_7 = "Price/quality"

answer\_8 = "Operating system"

answer\_9 = "Price"

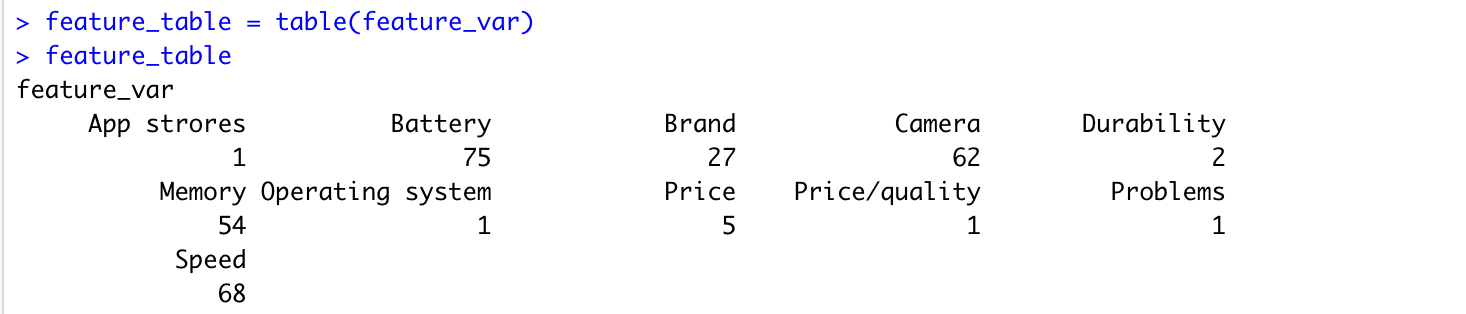
answer\_10 = "App stores"

answer\_11 = "Problems"

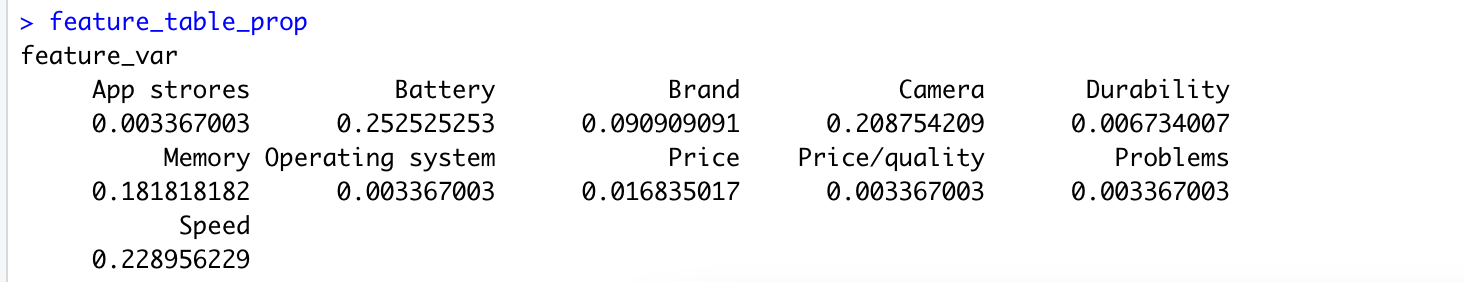
feature\_var = c(rep(answer\_1, 75), rep(answer\_2, 62), rep(answer\_3, 68), rep(answer\_4, 27), rep(answer\_5, 54), rep(answer\_6, 2),

rep(answer\_7, 1), rep(answer\_8, 1), rep(answer\_9, 5), rep(answer\_10, 1), rep(answer\_11, 1))

feature\_table = table(feature\_var)

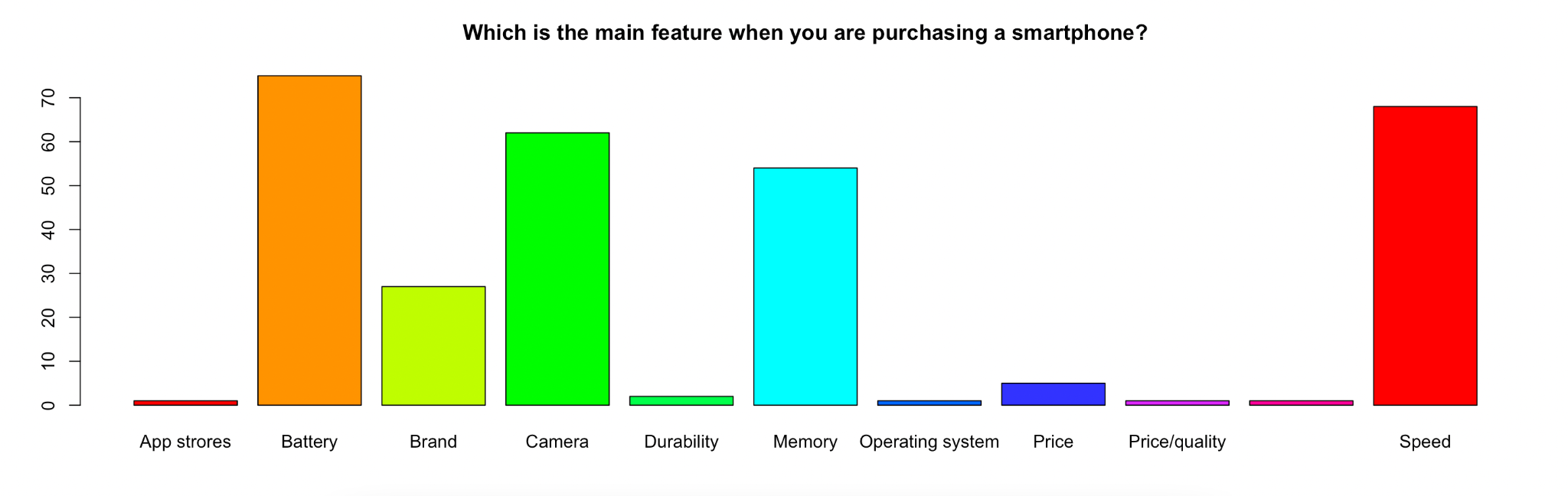


feature\_table\_prop <- prop.table(feature\_table)



Представям графично чрез **barplot,** отнасящо се за броя отговори 0-75

barplot(height = feature\_table, col = rainbow(10), main = "Which is the main feature when you are purchasing a smartphone?")



Изводът е, че батерията, заедно със скоростта са най-важните фактори определящи покупката на даден смартфон съответно със 75 и 69 отговори при запитването. Камера и памет със съответно 62 и 54 , а останалите са: Марка-27, Цена-5, Устойчивост - 2, Пазар за апове - 1, Съотношението цена/ качество - 2, Операционна система - 1, Проблем със стария - 1.

Сега ще покажа и процентното отношение с **piechart()**

piepercent\_features <- round(100\*feature\_table/sum(feature\_table), 1)

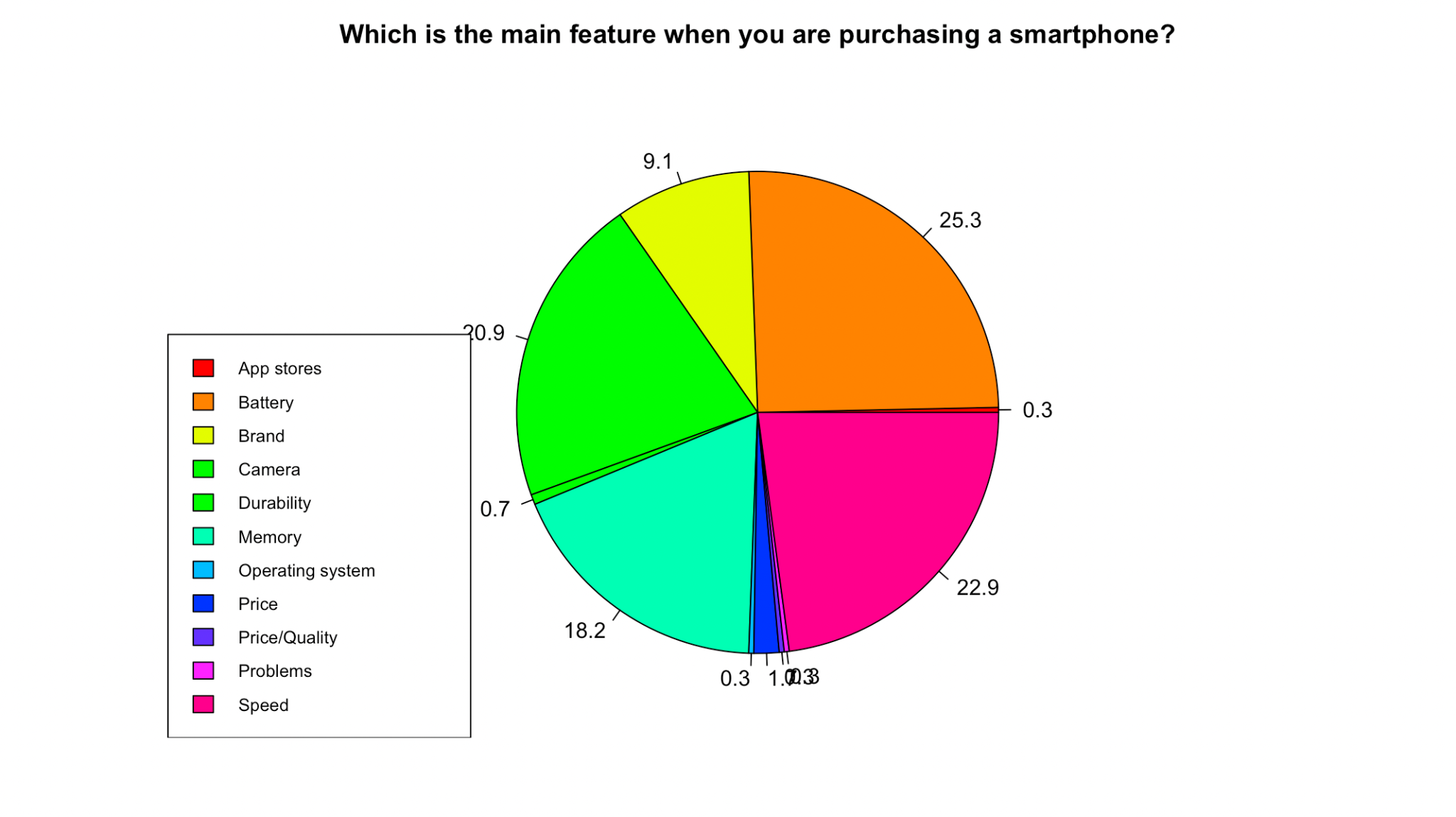
pie(feature\_table, labels = piepercent\_features, main = "Which is the main feature when you are purchasing a smartphone?",

col = rainbow(n = length(feature\_table)))

legend(x = "bottomleft", legend = c("App stores","Battery","Brand","Camera","Durability",

"Memory","Operating system","Price","Price/Quality","Problems", "Speed"), cex = 0.8,

fill = rainbow(length(feature\_table)))



***Въпрос 7: През колко години закупувате ново устройство?***

Сега вкарваме стойностите от анкетата на създаден от нас вектор за тази цел.

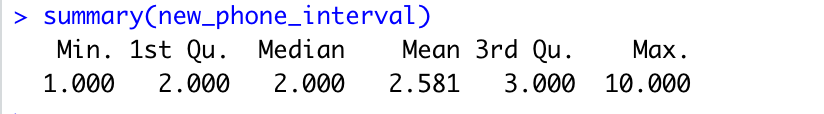
new\_phone\_interval <- new\_phone\_interval <-anketa\_statistika\_en$Q\_6

Първо ще намерим модата - най-често срещаната стойност.



Установих, че модата на този въпрос е 2, значи повечето хора сменяват своите апарати на всеки 2 години.

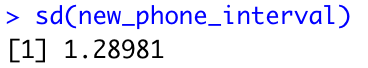
Сега прилагаме summary, която трябва да ни даде информация за минималната стойност 1ви квартил, медианата, 3тия квартил и максималната стойност.



Сега нека открием дисперсията(разпределението)



Намирам стандартното отклонение, то се явява оценка на дисперсията, което ни показва колко далеч се се намират наблюденията от очакването, и представлява корен квадратен от дисперсията.



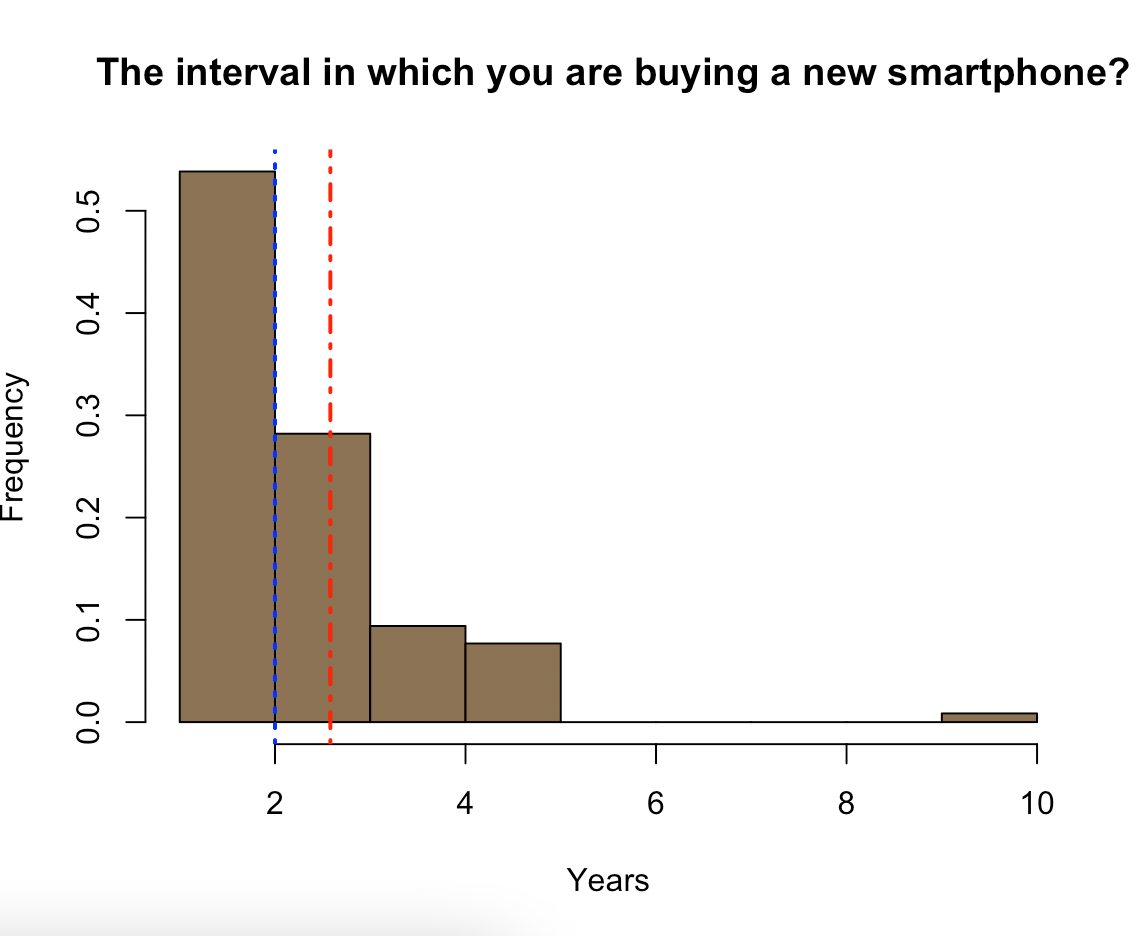
Графичното представяне както досега показвам с хистограма:

hist(new\_phone\_interval, main = " The interval in which you are buying a new smartphone?",

xlab = "Years", ylab = "Frequency", col = "burlywood4", prob = T)

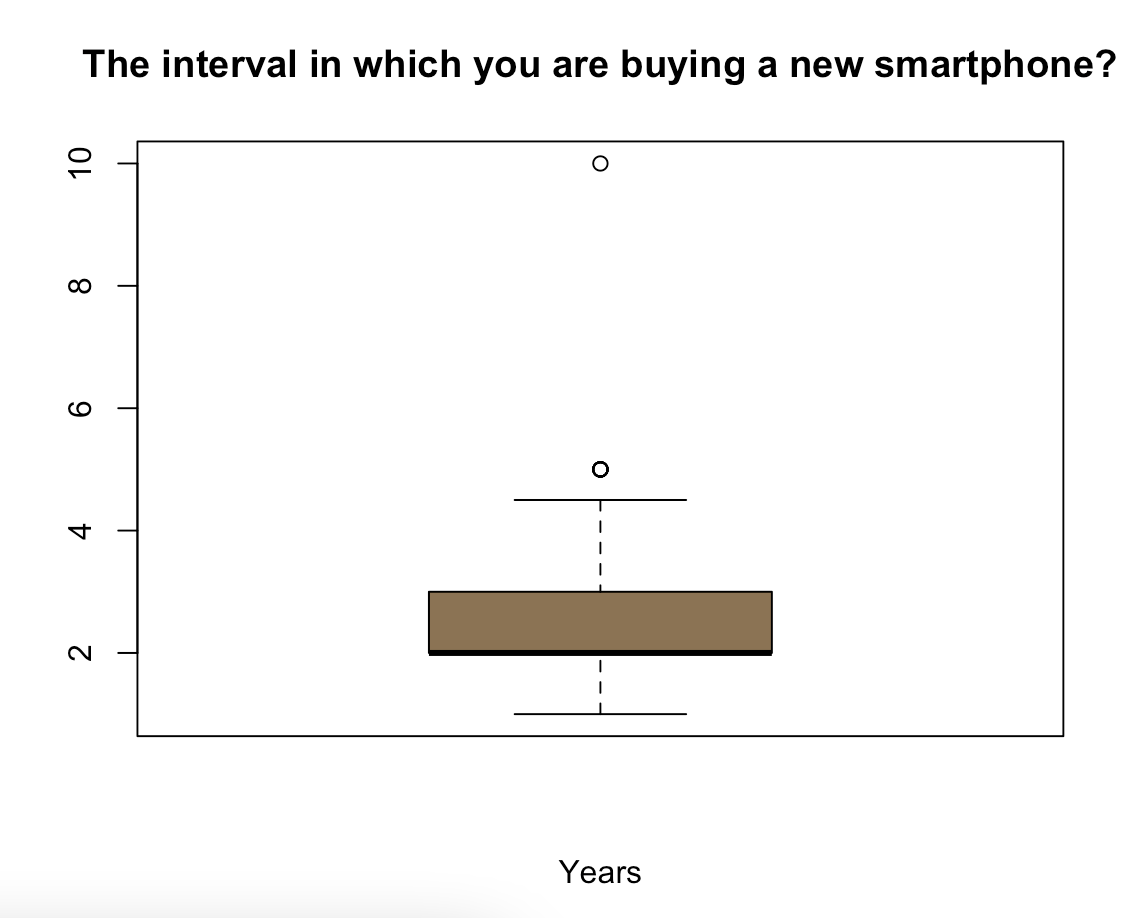
abline(v = mean(new\_phone\_interval), lwd = 2, lty = 4, col = "red")

abline(v = median(new\_phone\_interval), lwd = 2, lty = 3, col = "blue")



Със синя пунктирана линия е показана **медианата**, а с червена пунктирана линия е показана **средната стойност.**

И тук ще конструираме **boxplot**, с цел засичане на **outlier-и**:



На графиката се вижда, че имам засечени два потенциални outlier-а.

***Въпрос 8:От (0-10) оценете колко удовлетворени сте от своето устройство?***

Отговорите на последния въпрос от анкетата също представляват числова променлива и отново ще анализираме по установения досега ред.

Както винаги първо сформираме вектор от стойностите:

customer\_satisfication <- anketa\_statistika\_en$Q\_7

Сега намирам модата, което е най-срещаната стойност от анкетата.

modeFunction <- function(x) {

res\_table <- table(x)

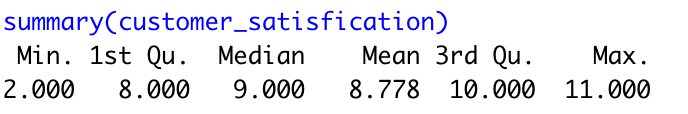
return(names(res\_table)[res\_table == max(res\_table)])

}

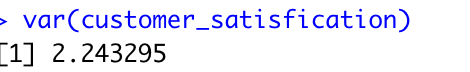


Модата е 10, което като заключение можем да кажем, че хората предимно са доволни от работата на своето устройство.

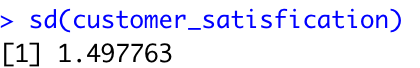
Сега прилагаме summary, която трябва да ни даде информация за минималната стойност 1ви квартил, медианата, 3тия квартил и максималната стойност.



Сега нека открием дисперсията(разпределението)



Намирам стандартното отклонение



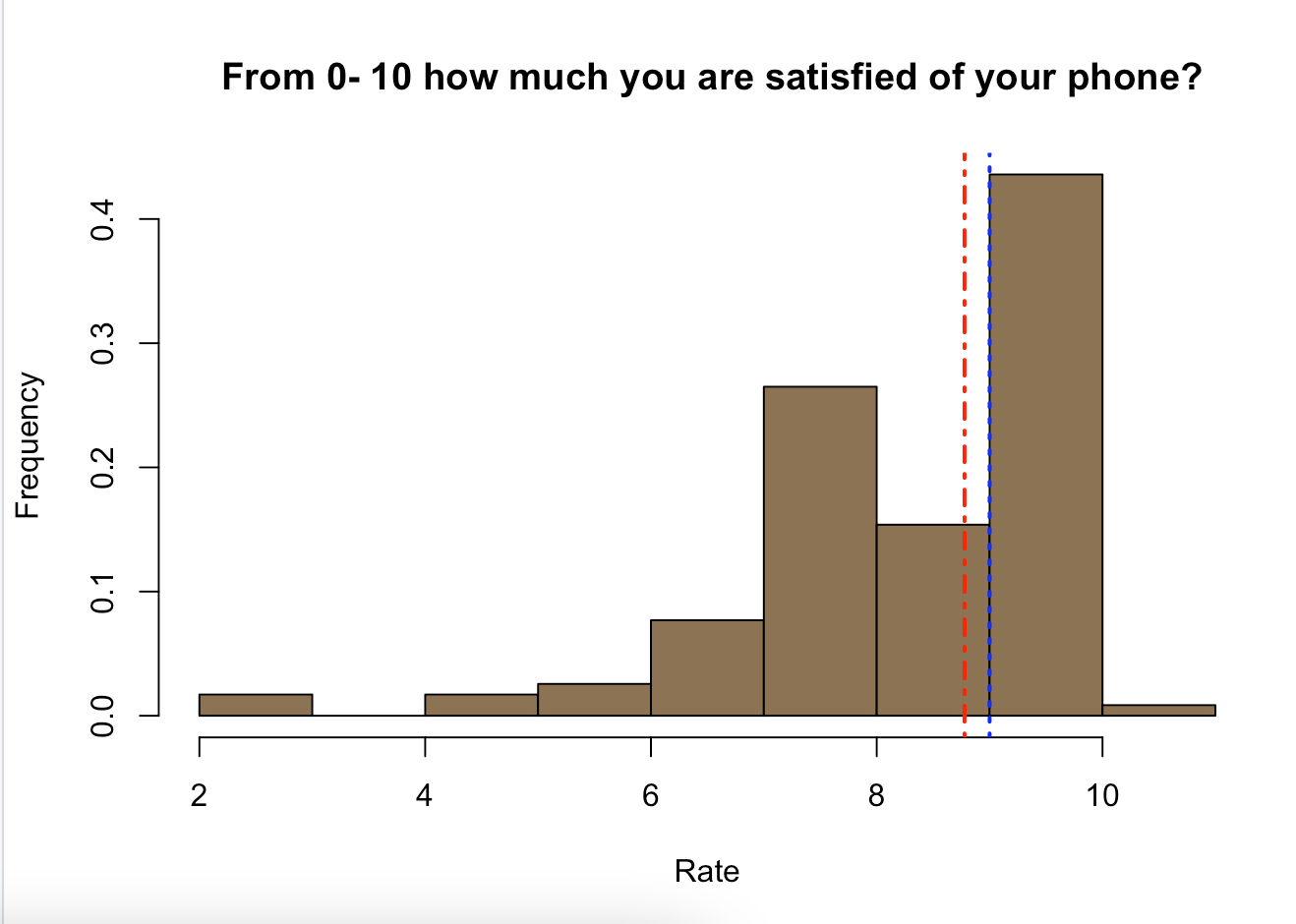
Графичното представяне ще покажа с хистограма:

hist(customer\_satisfication, main = " From 0- 10 how much you are satisfied of your phone?",

xlab = "Rate", ylab = "Frequency", col = "burlywood4", prob = T)

abline(v = mean(customer\_satisfication), lwd = 2, lty = 4, col = "red")

abline(v = median(customer\_satisfication), lwd = 2, lty = 3, col = "blue")

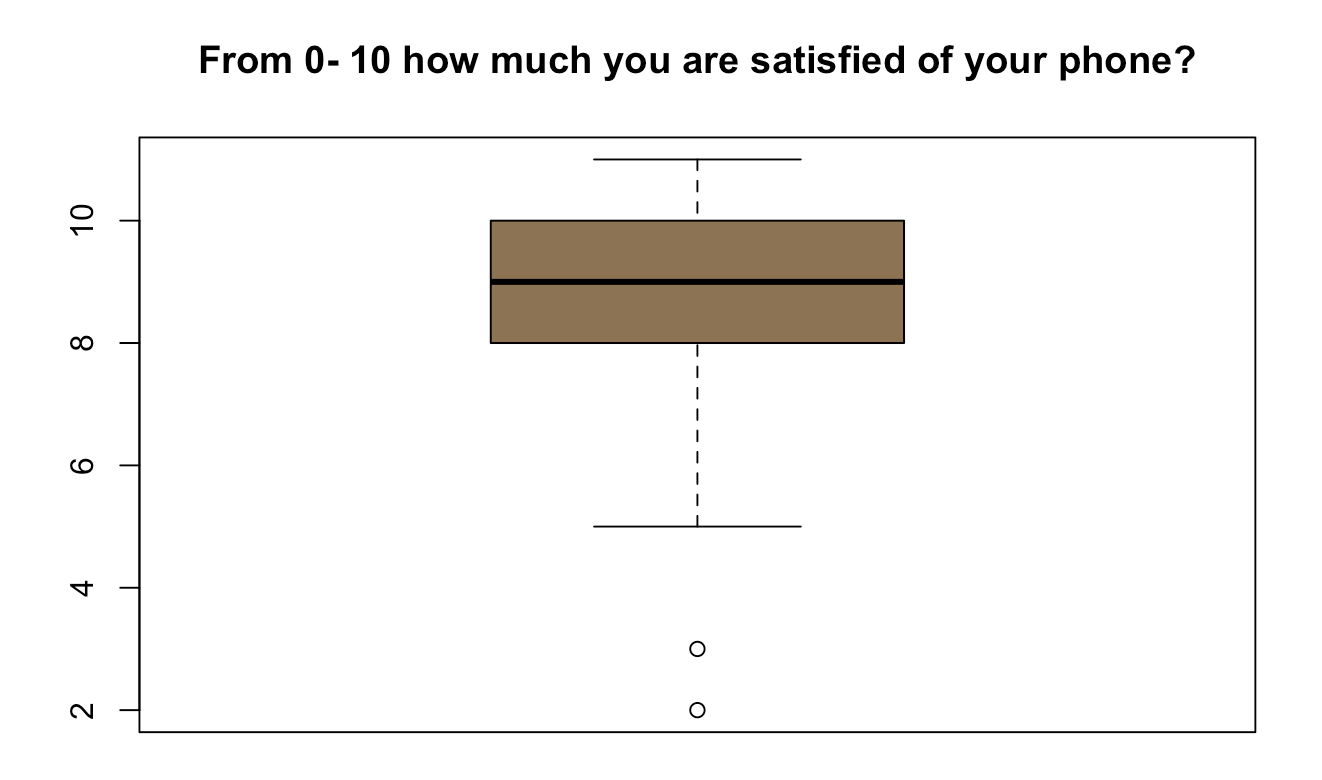


Със синя пунктирана линия е показана **медианата(9)**, а с червена пунктирана линия е показана **средната стойност(8.777778).**

Построяваме boxplot с цел засичане на **outlier-и**:

boxplot(customer\_satisfication, col = "burlywood4", main = "From 0- 10 how much you are satisfied of your phone?",

xlab = "")



Засечени са 2 потенциални **outlier-и.**

**3. Анализ на взаимодействието между две променливи**

**Нека започнем с корелационен анализ**

**3.1 Числова срещу числова**

Нека изберем следните въпроси:

Въпрос 5: Колко инча е вашият смартфон?

Въпрос 8:От (0-10) оценете колко удовлетворени сте от своето устройство?

**3.1.1 Корелационен анализ**

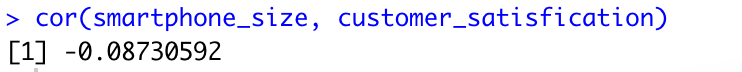
Първо нека видим **корелационния** **анализ** на двете променливи:

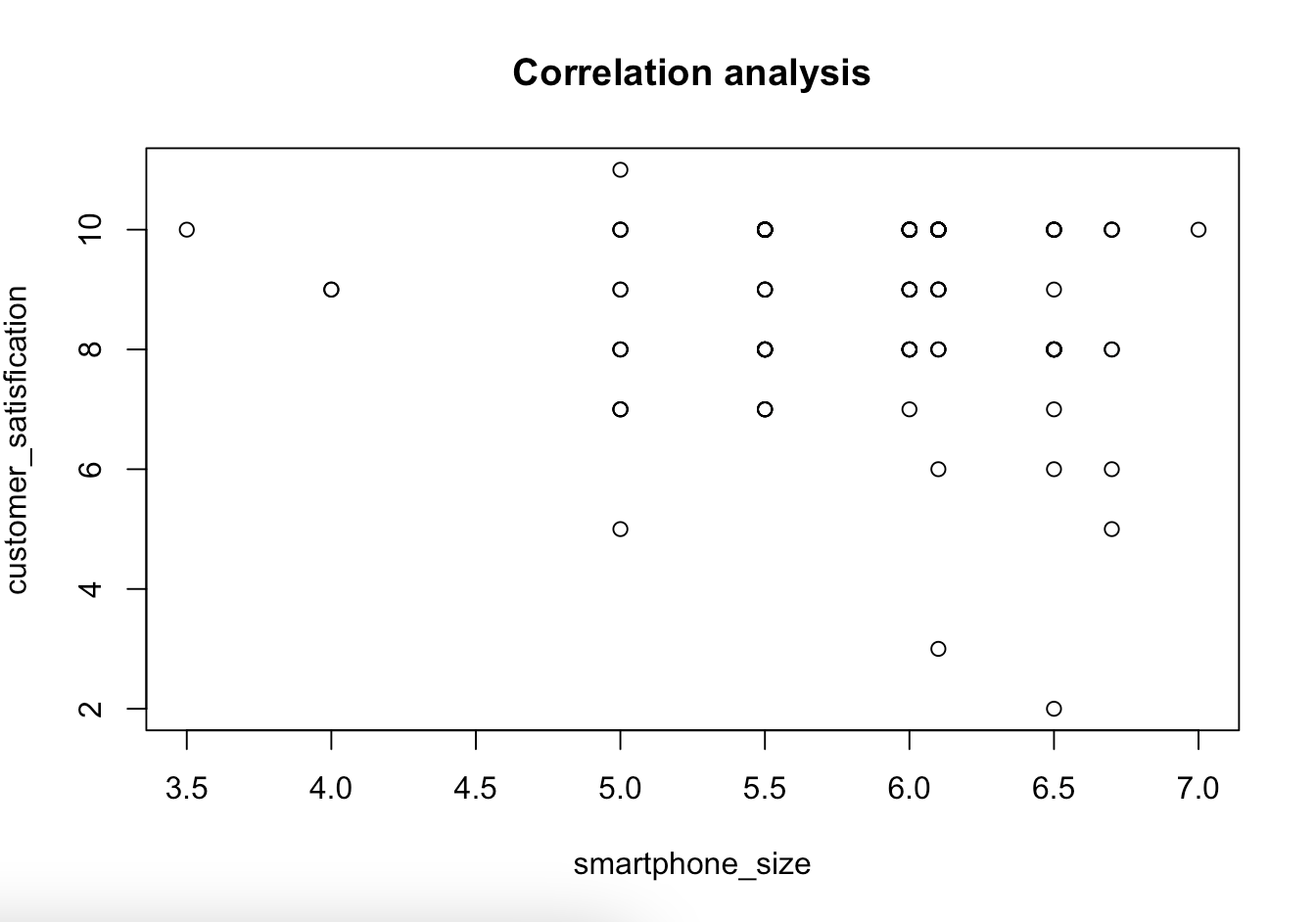
rho <- round(cor(smartphone\_size, customer\_satisfication), 3) #коефициент на корелация

par(mfrow = c(1, 1))

plot(smartphone\_size, customer\_satisfication, main = "Correlation analysis")

cor(smartphone\_size, customer\_satisfication)





**rho** е нашият коефициент на корелация и както се вижда и на графиката корелацията е слаба и на базата на така полученият резултат нямаме основание да твърдим, че големината на смартфона има влияние върху удовлетворението от устройството.

**3.1.2 Линейна регресия:**

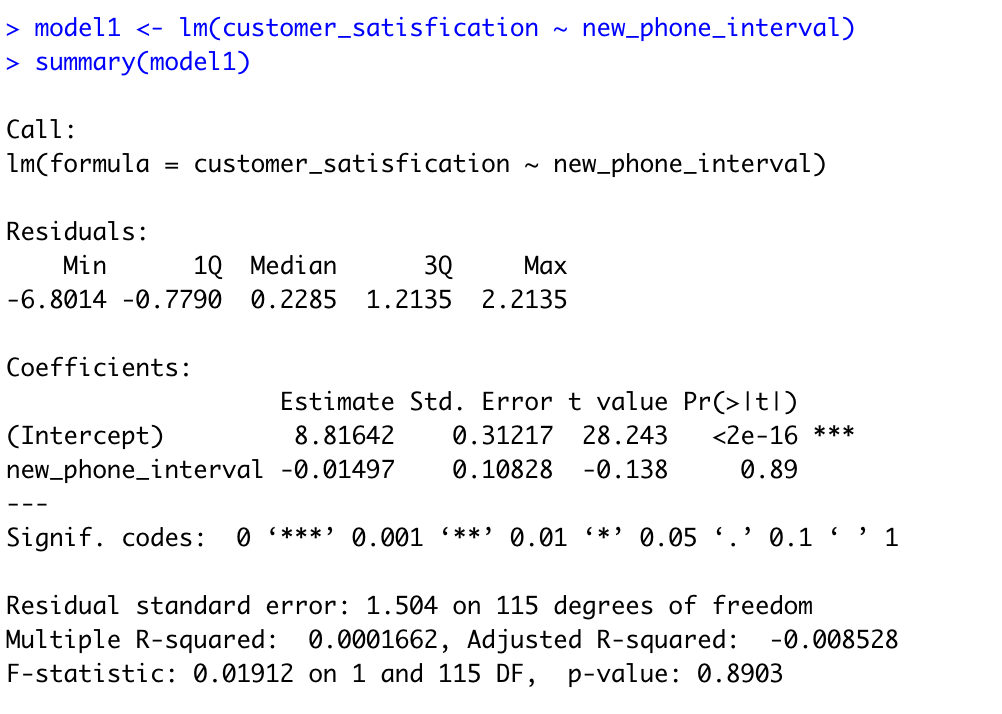
Въпрос 7: През колко години закупувате ново устройство?

Въпрос 8:От (0-10) оценете колко удовлетворени сте от своето устройство?

Целта на анализа е да тестваме до колко адекватно е да се опитваме да предскажем удовлетвореността на потребителите на база на данните през колко години сменят мобилното си устройство.

Нека въведем данните от въпросите в един **data** **frame**:

DF <- data.frame(new\_phone\_interval, customer\_satisfication)



От таблицата, изведена чрез функцията **summary** на нашия модел виждаме, че коефициентът пред променливата *new-phone\_interval* не е статистически значим, тъй като съответстващото му p-value>0.05 и съответно тази променлива не може да бъде използвана за предсказване на удовлетвореността на потребителите.

**Multiple/Adjusted** **R-squared** е процентът от дисперсията на данните, които се обясняват от модела и приема стойности [0,1]. За да смятаме, че моделът добре описва данните стойността на R-squared трябва да е близо до 1.

В моят случай, както се показа от горния анализ, стойността на R-squared е доста малка, което означава, че нашият модел не може да бъде използван за адекватно предсказване на стойностите на нивото на удовлетвореност на потребителите.

Това показва, че задоволството на хората от техните мобилни устройства се влияе от други фактори, различни от периода на закупуване на нов смартфон.

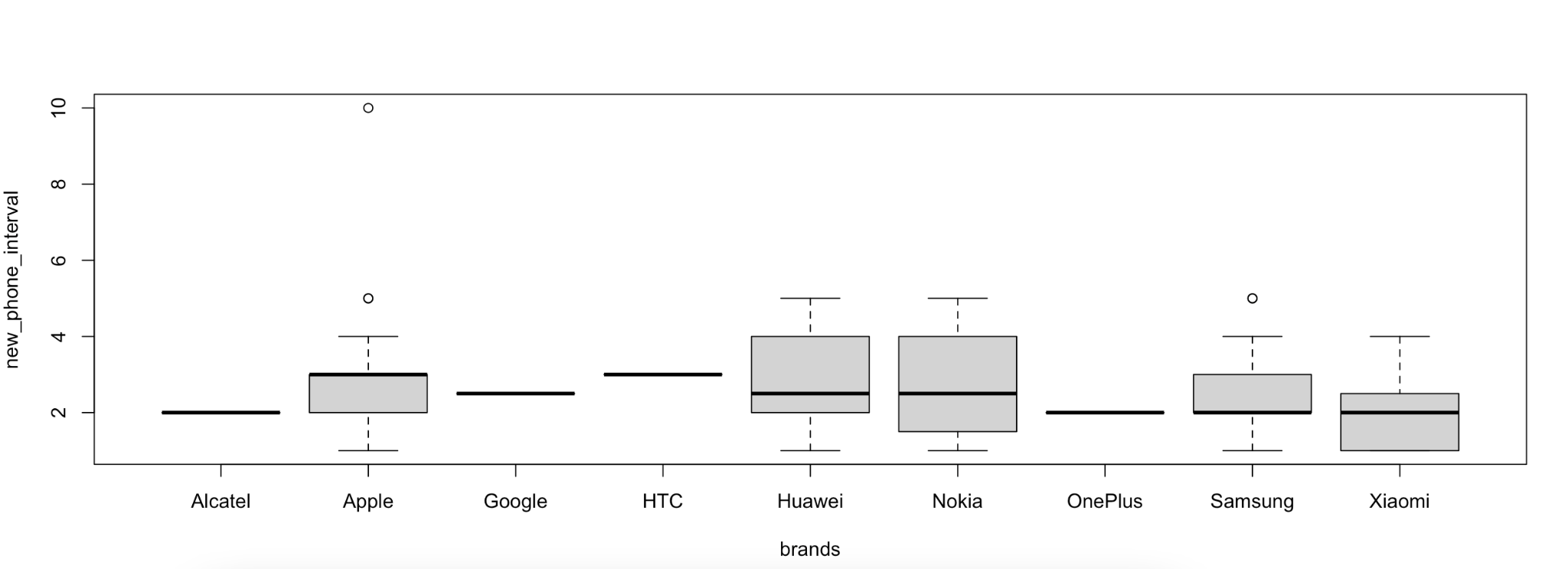
**3.2 Числова срещу категорна**

Ще изследвам връзките между променливите **new\_phone\_interval**(годините през, които се закупува ново устройство) и **brands**(марка на смартфона)

Представям графично с boxplot:

brands <-anketa\_statistika\_en$Q\_4

new\_phone\_interval\_vs\_table\_brands<-boxplot(new\_phone\_interval~brands)

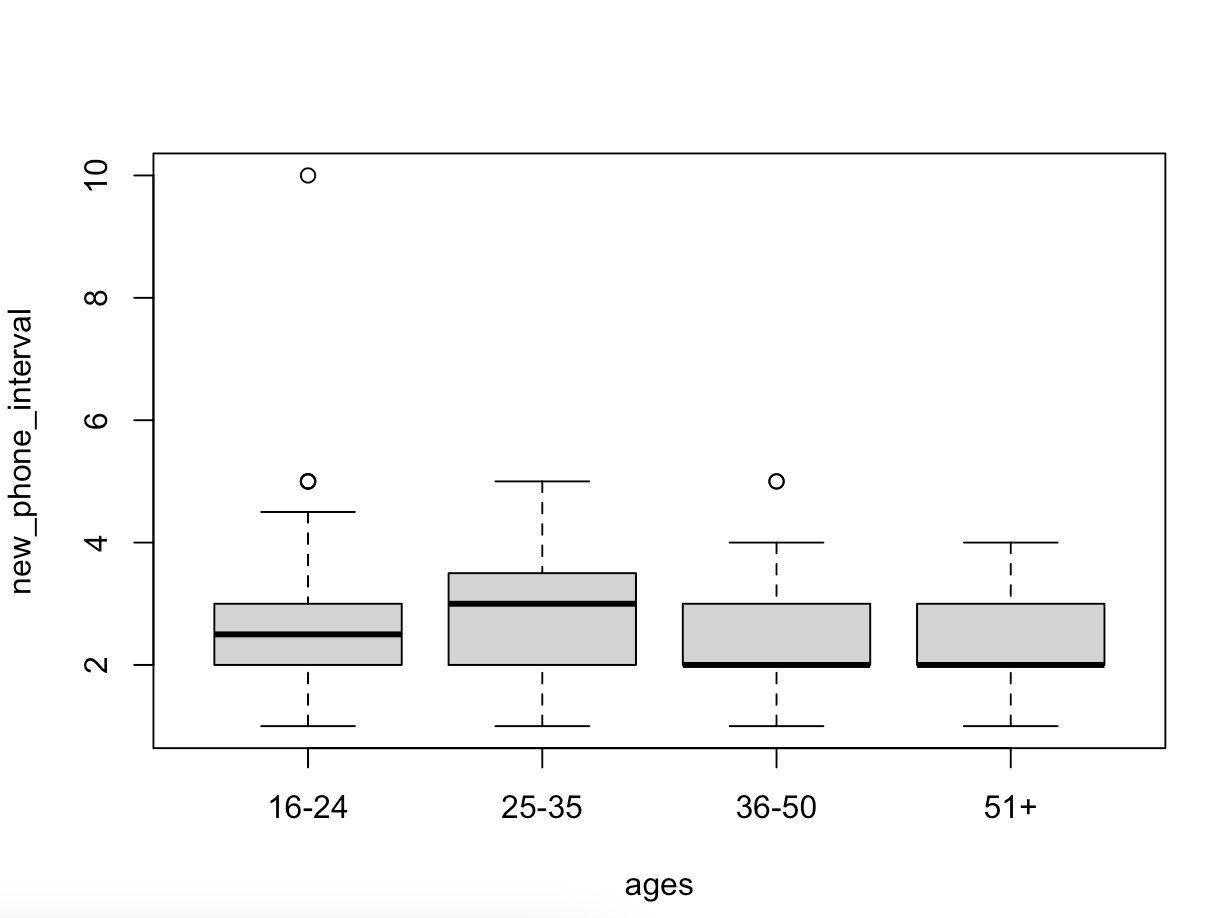


Удебелената черна линия във всяка една група е медианата. От двете страни са съответно първи и трети квартили. От графиката се вижда, че хората ползващи марките Samsung, Huawei, Nokia, Xiaomi сменят своите устройства през около 2 години. Виждаме също, че потребителите на Apple, чиито резултати са средно 3 години, използват един телефон по-дълго време от гореспоменатите марки. Останалите няколко групи са с малко наблюдения и не можем да заявим категорично тяхната значимост.

**3.3 Категорна срещу числова**

Ще изследвам връзките между променливите **new\_phone\_interval**(годините през, които се закупува ново устройство) и **ages**(възрастова категория на потребител)

Изобразявам графично:



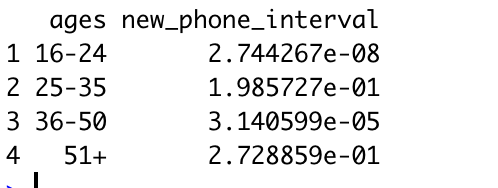
Удебелената черна линия във всяка една група е медианата. От двете страни са съответно първи и трети квартили. От графиката се вижда, че хората между 25-35 години заменят своите смартфони през най-много време, последвани от хората на възраст 16-24, останалите групи сменят своите смартфони доста по-често, през около 2 години.

За изследване на разликата в локациите ще използвам **One-way ANOVA**, ако някое от неговите предположения бъде нарушено ще да използвам **Kruskal test**.

Първо ще проверя за нормалност на разпределението на данните в различните групи с функцията **aggregate()** и **shapiro-wilk test.**

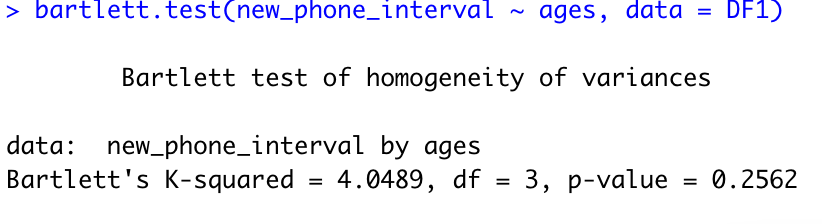
aggregate(new\_phone\_interval ~ ages, data = DF1, FUN = function(x) {shapiro.test(x)$p.value})

Първото

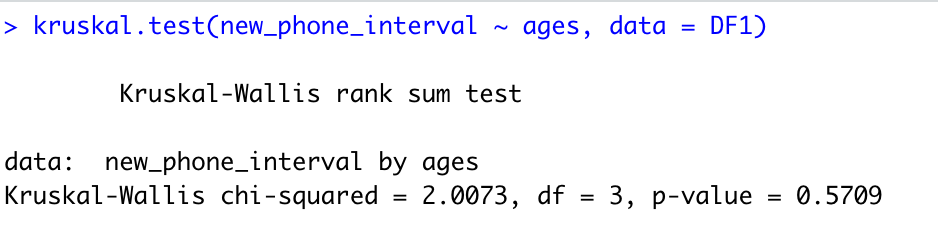
****

От така нагледната таблица се вижда, че стойностите на p-value в група 2 и група 4 са по-големи от alpha = 0.05 и можем да твърдим, че в тези групи, данните имат нормално разпределение. В другите две групи обаче, стойността на p-value <0.05 и имаме основание да отхвърлим нулевата хипотеза за нормалност на разпределението, в полза на алтернативната, че данните идват от разпределение, което се отличава от нормалното.

Сега ще използвам **bartlett,** за да проверя хомогенността и дисперсията между различните групи.



Стойността на p-value > 0.05 => Нямаме основание да отхвърлим нулевата хипотеза за равенство на дисперсиите и можем да смятаме, че дисперсиите в групите са равни ( хомоскедастичност).



**Kruskal-Wallis test** също показва, че стойността на **p-value**>0.05, което потвърждава, че не можем да твърдим, че има значими разлики между дисперсиите на групите.

**3.4 Категорна срещу категорна**

Това сравнение не беше упоменато в изискванията на проекта, но все пак аз искам да го представя, защото от него можем да направим генерален извод за използваемостта на всяка марка от различните възрастови групи.

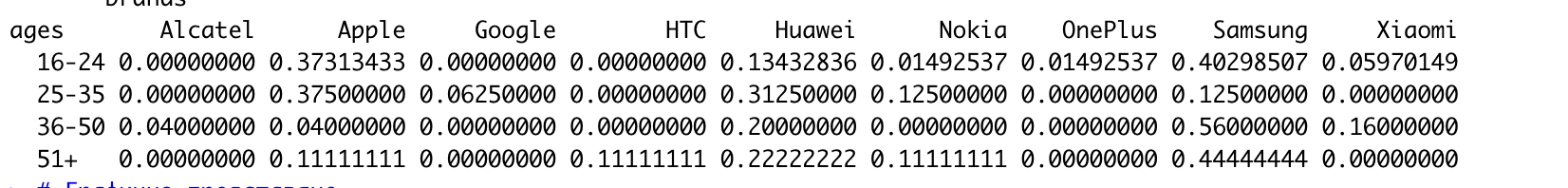
Затова взимаме двете променливи :

ages <- anketa\_statistika\_en$Q\_2

brands <-anketa\_statistika\_en$Q\_4

table(ages, brands)

prop.table(x = table(ages, brands), margin = 1)

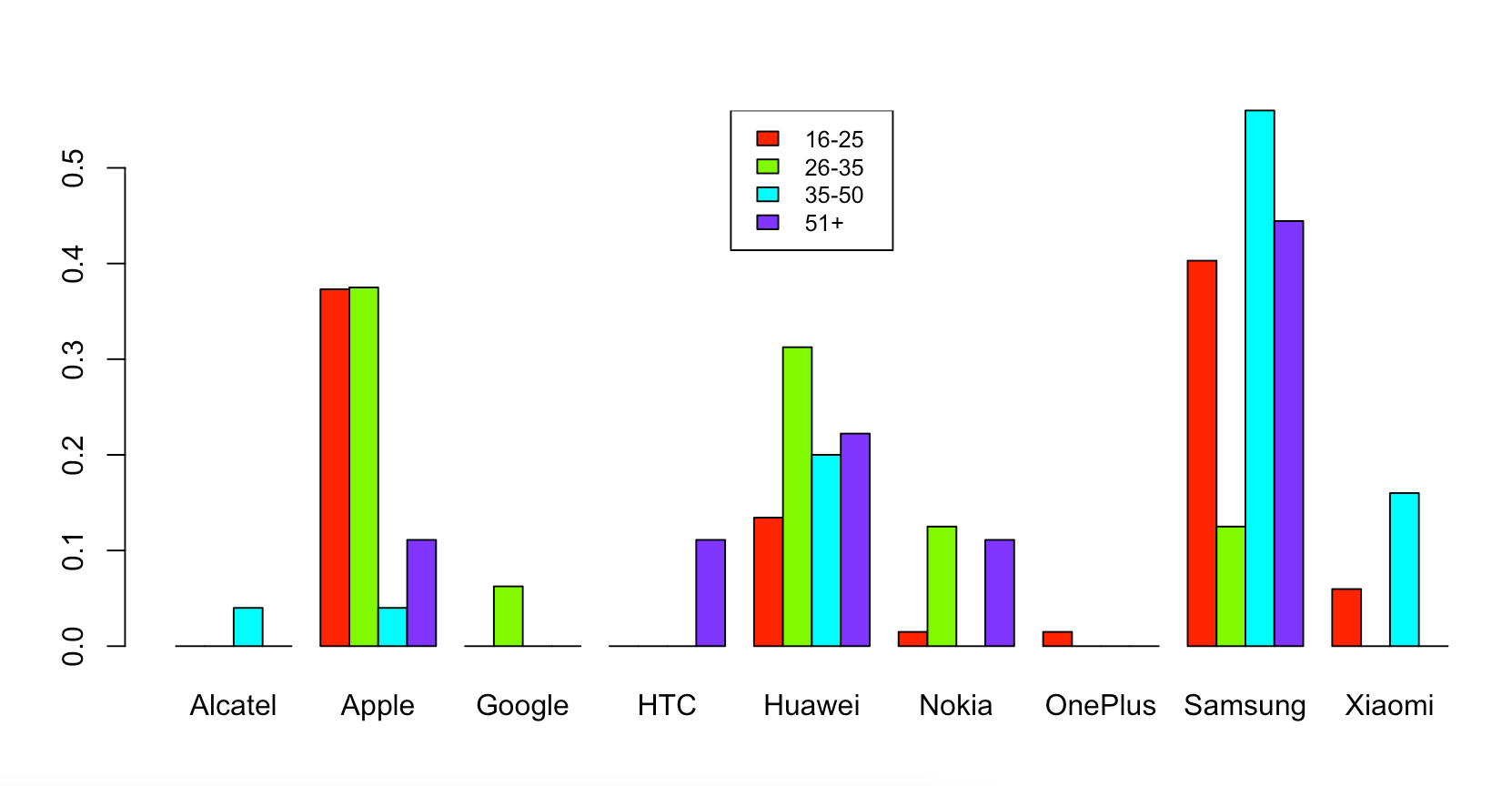


И графично:

barplot(prop.table(x = table(ages, brands), margin = 1),col = rainbow(length(table(ages))) ,beside = TRUE)

legend(x = "top", legend = c("16-25","26-35", "35-50", "51+"),

cex = 0.8, fill = rainbow(length(table(ages))))



Гледайки графиката установяваме предимството на Apple в групaтa “25-35” - 40%. Apple и Samsung разпределят почти еднакви дялове в “16-25” - 40%, Samsung е първенец в групата “35-50” с 56%, следван от Huawei - 20% и Xiaomi - 16%. В групата “51+” Samsung- 45%, Huawei-22%.

**4. Заключение**

С помощта на различни подходи и графично представяне направихме някои важни изводи за пазара на мобилни устройства.

Много важен извод е това, че всички от анкетираните притежават смартфон, което прави анализа наистина стойностен.

Анализът на разпределението на мобилните марки спрямо възрастовите интервали би бил полезен за производителите и малките търговци, за да осъществяват качествен контрол върху покупко-продажбите на аксесоари за смартфоните.

Също така бяха анализирани и важни фактори, определящи покупката на един смартфон (приоритетните функционалности на един телефон Батерия, Скорост, Памет и тн.), според предпочитанията на потребителите, което е насока за производителите на смартфони.

Стигнахме до извода, че не можем да твърдим, че възрастта на потребителя може да определи какъв е интервалът за смяна на смартфоните.

Още заключихме, че нямаме основание да твърдим, че големината на смартфона както и периода на закупуване на нов смартфон имат влияние върху удовлетворението на потребителите от устройството.

Беше постигната целта на анализа - да онагледи разпределението на мобилния пазар в България.

Използвани материали:

Записки от упражнения на Методий Кандиларов:

<https://www.facebook.com/groups/1166815187109672>

<https://www.investopedia.com/terms/t/t-test.asp>

<http://www.sthda.com/english/wiki/one-way-anova-test-in-r>

<https://www.scribbr.com/statistics/anova-in-r/>

<https://www.statmethods.net/graphs/bar.html>

<https://cran.r-project.org/web/packages/xlsx/xlsx.pdf>

<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/bartlett.test>

<http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r>

<http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/>