

Софийски университет „Св. Климент Охридски“

Факултет по математика и информатика

Специалност: Информационни системи,

Курс 2, Група 1

Курсов проект

по Статистика и емпирични методи

на тема:

**Зависимостта между човека и мобилните
телефони**

Изготвен от: Кристина Георгиева Цекова

Факултетен номер: 71 852

Част 1. Избор на данни. Задаване на цели на проекта.

Начинът ни на живот е пряко свързан със социалните мрежи, които можем да използваме във всеки един момент, най – вече благодарение на мобилните си телефони. И именно фактът, че те са станали част от ежедневието ни, ме накара да избира точно тази тема за проекта си. За осъществяването на целите му направих проучване - чрез анкета. Разпространих я сред хора на различна възраст и събраните данни обобщих в таблица.

Главната цел на този проект е да се определи дали хората наистина са зависими от своите мобилни телефони и това отразява ли се на отношенията им с техните близки и приятели. Ще намеря отговор и на въпросите:

- За какво най – често хората използват телефоните си?
- Коя е най – използваната от тях социална мрежа?
- Каква такса средно заплащат на месец, за да използват услуги от мобилните си оператори?

Анкетата се състои от 11 въпроса. Общият брой на събраните отговори е 130, като на някои въпроси може да се посочват и повече от един отговор.

Най – напред, последователно ще направя анализ на всяка променлива (всеки въпрос). След това ще направя проверка за зависимост между отделните въпроси. И накрая ще направя заключение, като отговоря на въпросите по – горе.

Самата анкета може да намерите, като отворите следния линк:

https://docs.google.com/forms/d/e/1FAIpQLSd5ixGS68ztPB1339gIpLHzKoH0QLWqQ6uuzbn8b9z71Qn3rw/viewform?usp=sf_link

Част 2. Описателен анализ на променливите. Анализ на взаимодействието им.

За да можем да използваме данните и да ги изследваме с помощта на R, първо трябва да ги въведем. Тъй като ще въведем таблицата с отговори, която е с формат .xlsx, то първо трябва да инсталирам пакетите за обработка на такива файлове. Следният код показва как става въвеждането на данните от таблицата, като функцията `library(readxl)` зарежда пакетите от библиотеката `readxl`.

```
~/Project71852/ ↗  
> library(readxl)  
> myResults <- read_excel("C:/Users/Krisi/Downloads/myForm_71852.xlsx")  
> colnames(myResults) <- paste("Question", 1:ncol(myResults), sep = "_")  
> str(myResults)  
tibble [130 x 11] (s3: tbl_df/tbl/data.frame)
```

Получава се структура от данни `tibble`, която е специален случай на `data.frame` и е предназначена за набора от пакети `tidyverse`. За разлика от `data.frame`, при `tibble` редовете и колоните може да са с различна дължина.

За по – добра четливост задавам имена на колоните: “Question_”, като долната черта е последвана от номера на колоната.

2.1 Анализ на едномерна променлива – графично представяне и статистики за локацията и дисперсията.

Анкетата съдържа 11 въпроса, които включват три непрекъснати числови величини (като две от тях са от един и същи тип), една дискретна числова величина и останалите са категорийни, като има и ординални, и номинални.

- Категорийните променливи носят по – малко информация, отколкото числовите. Но за сметка на това представят по – голяма стабилност за прогнозните модели. За анализ на този вид променливи трябва да представим данните в таблици и след това да ги изобразим графично, като най – подходящ инструмент за тази цел е `barplot`. Може да се използва и `pie`.
- Числовите променливи ще изследвам по следния начин:
 - Оценка на центъра (локацията) на разпределение.
 - средна стойност(очакване)
 - медиана
 - мода

- Оценка на вариацията на разпределение.
 - Обхват (Range)
 - Вариация (дисперсия) и стандартно отклонение
 - The five number summary

Ще изобразя графично разпределението на отговорите на съответните въпроси, представляващи числови променливи. А след изследването на въпросите поотделно ще направя и тест, който ще покаже дали са нормално разпределени, или не.

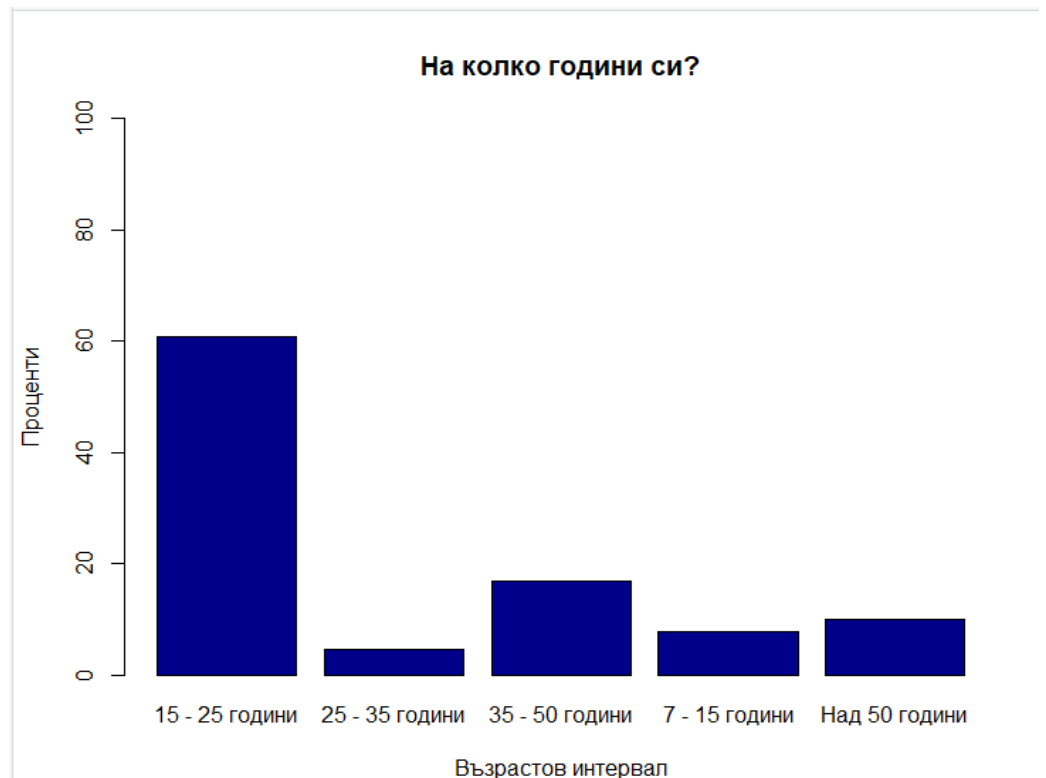
❖ Въпрос 1: На колко години си?

Този въпрос представлява категорийна величина, тъй като анкетираните трябва да изберат от петте отговора в кой възрастов интервал попадат, т.е. възрастта им може да попада в 5 различни категории.

Код на R:

```
~/Project71852/ ↗  
> table_q1 <- table(myResults$Question_1) # използваме table(), за да представим данните в таблица  
> table_q1  
  
15 - 25 години 25 - 35 години 35 - 50 години 7 - 15 години Над 50 години  
79          6          22          10          13  
> barplot(round(prop.table(table_q1)*100, 2), col = "darkblue", main = "На колко години си?", xlab = "Възрастов интервал",  
+         ylab = "Проценти", ylim = c(0, 100))
```

Графика:



Отговорите на този въпрос са 130.

От графиката се вижда, че от анкетираните:

- о около 60% са на възраст 15 – 25 години
- о 5% са тези в интервала 25 – 35 години
- о 10% над 50 години
- о 10% от анкетираните са в интервала 7 – 15 години
- о 15% са в интервала 25 – 50 години

❖ Въпрос 2: Имаш ли мобилен телефон, таблет или друго устройство, което използваш за комуникация и за връзка с Интернет?

След прочитането на отговорите от втория въпрос се вижда, че всичките 130 човека са отговорили с „Да, имам“, което означава, че 100% от анкетираните притежават мобилен телефон или друго устройство за комуникация и връзка с Интернет. Това можем да го покажем, използвайки следния код на R:

```
~/Project71852/ ↗  
> table_q2 <- table(myResults$Question_2)  
> factor(table_q2)  
Да, имам.  
130  
Levels: 130
```

Функцията factor() показва, че данните са разпределени в 130 нива (130 отговора) и в случая на всяко едно ниво има един и същ отговор.

❖ Въпрос 3: По колко часа на ден в интервала 0 - 24 използваш телефона си?

Третият въпрос представлява числова непрекъсната променлива, която ще изследвам, следвайки стъпките, споменати по – горе. Започвам с въвеждането на всички отговори във вектор. След това ще намеря средната стойност, медианата и модата, както и центъра на разпределение:

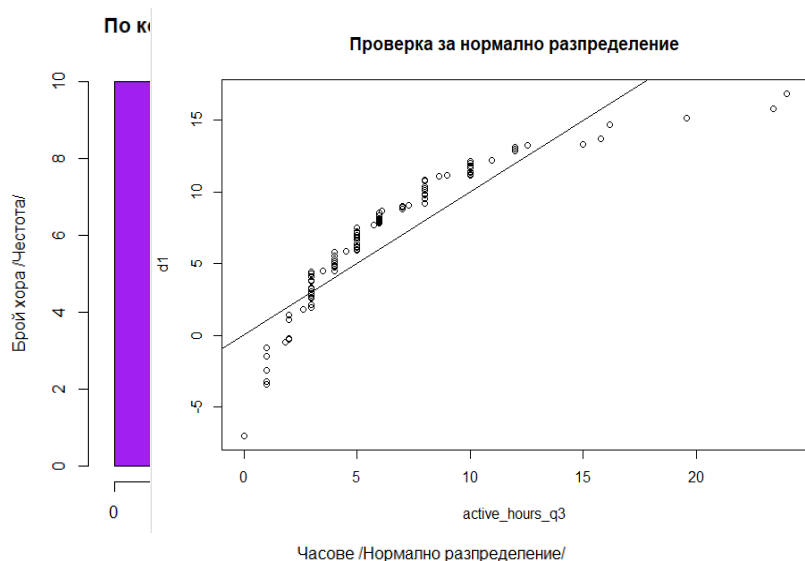
```
~/Project71852/ ↗
> active_hours_q3 <- c(5,5,9,8,6,5,1,6,1,2,4,6,6,5,6,24,8,3,3,2,6,3,5,3,5,5,2,5,10,1,8,8,3,16,2,3,3,10,
+ 3,4,3,12,8,8,7,3,16,15,10,5,10,10,4,5,6,7,4,1,4,5,12,1,10,24,3,4,5,10,8,5,10,4,4,12,10,
+ 8,3,3,2,2,7,4,4,5,6,7,10,10,3,15,8,5,6,6,3,4,10,3,3,2,22,12,1,8,5,4,18,0,3,5,2,11,9,8,
+ 1,6,6,6,8,3,7,3,6,12,8,5,4,8,6)
> length(active_hours_q3)
[1] 130
> mean(active_hours_q3) # средна стойност
[1] 6.338462
> median(active_hours_q3) # медиана
[1] 5
> table_q3 <- table(active_hours_q3) # мода - най - често срещана стойност
> names(table_q3)[table_q3 == max(table_q3)]
[1] "3"
> # Следващите две функции описват центъра на разпределение
> summary(active_hours_q3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000  3.000   5.000   6.338   8.000   24.000
> quantile(active_hours_q3, prob = seq(0.1, 0.9, by = 0.1))
10% 20% 30% 40% 50% 60% 70% 80% 90%
2.0  3.0  4.0  5.0  5.0  6.0  8.0  9.2 11.1
```

Сега ще опиша вариацията (дисперсията) и стандартното отклонение, което ще покаже колко далече са наблюденията от очакването. Освен това чрез хистограма ще изобразя разпределението на данните графично.

Код на R:

```
~/Project71852/ ↗
> # Вариация (дисперсия) на разпределението
> range(active_hours_q3) # range - показва най - голямата и най - малката стойност
[1] 0 24
> var(active_hours_q3) # дисперсия
[1] 19.14037
> sd(active_hours_q3) # стандартно отклонение
[1] 4.374971
> fivenum(active_hours_q3)
[1] 0 3 5 8 24
> # Графично представяне
> hist(table_q3, main = "По колко часа на ден в интервала 0 - 24 използваш телефона си?", xlab = "часове /нормално разпределение/",
+ ylab = "Брой хора /Честота/", col = "purple")
>
> d1 <- rnorm(n = 10^2, mean = mean(active_hours_q3), sd = sd(active_hours_q3))
> qqplot(active_hours_q3, d1, main = "Проверка за нормално разпределение")
> abline(a = 0, b = 1) # чертае линия
```

Графики:



Повечето от хората прекарват между 0 – 5 часа дневно над телефоните си, като най – честият отговор е 3 часа. Изчислявайки средната стойност от всички отговори, става ясно, че средно на човек се падат по около 6 часа на ден, което е доста тревожен факт. Освен това

отговорите не са нормално разпределени, както се вижда от втората графика, но това ще докажем по – късно с помощта на shapiro.test.

❖ Въпрос 4: С каква цел най - често използваш мобилното си устройство?

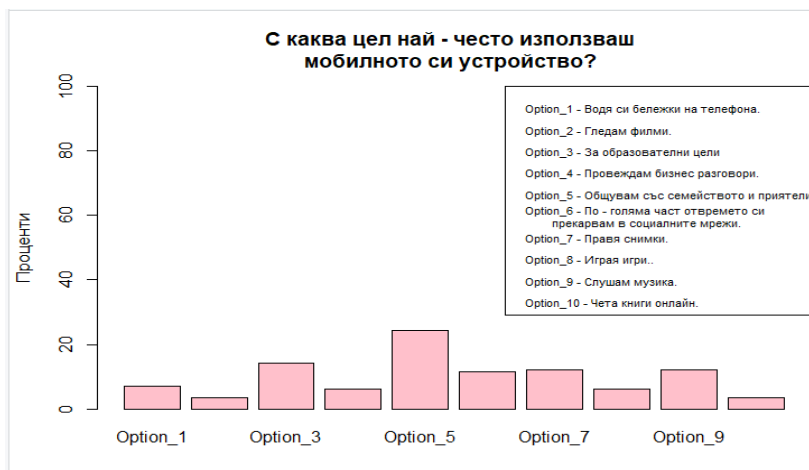
Този въпрос представлява категорийна величина, състояща се от 10 категории. Всеки може да избира повече от един отговор. За да не изписвам дългите наименования на отговорите, ще си ги преименувам с answer_1, answer_2, ... , answer_10. Въвеждам данните по следния начин:

```
~/Project71852/   
> answer_1 <- "общувам със семейството \ни приятелите си."  
> answer_2 <- "провеждам бизнес разговори."  
> answer_3 <- "Играя игри."  
> answer_4 <- "гледам филми."  
> answer_5 <- "чета книги онлайн."  
> answer_6 <- "Водя си бележки на телефона."  
> answer_7 <- "Правя снимки."  
> answer_8 <- "Слушам музика."  
> answer_9 <- "По - голяма част от\времето си прекарвам \в социалните мрежи."  
> answer_10 <- "За образователни цели."  
> # функция rep() - обединява всички отговори и тяхното процентно разпределение  
> question_4 <- c(rep(answer_1, 88), rep(answer_2, 22), rep(answer_3, 22), rep(answer_4, 12),  
+ rep(answer_5, 12), rep(answer_6, 25), rep(answer_7, 44), rep(answer_8, 44),  
+ rep(answer_9, 42), rep(answer_10, 51))  
> table_q4 <- table(question_4) # правим таблица с отговорите  
> str(table_q4)  
'table' int [1:10(1d)] 25 12 51 22 88 42 44 22 44 12  
- attr(*, "dimnames")=List of 1  
..$ question_4: chr [1:10] "Водя си бележки на телефона." "гледам филми." "За образователни цели." "Играя игри." ...
```

За графичното представяне на данните използвам barplot и добавям легенда:

```
~/Project71852/   
> barplot(round(prop.table(table_q4)*100, 2), col = "pink", main = "С каква цел най - често използваш \мобилното си устройство? ",  
+ ylim = c(0, 100), ylab = "проценти")  
> row.names(table_q4) <- paste("option", 1:nrow(table_q4), sep = "_")  
> legend(x = "topright", legend = c("option_1 - Водя си бележки на телефона.",  
+ "option_2 - Гледам филми.", "option_3 - За образователни цели", "option_4 - Провеждам бизнес разговори.",  
+ "option_5 - Общувам със семейството и приятелите си.", "option_6 - По - голяма част от времето си  
+ прекарвам в социалните мрежи.", "option_7 - Правя снимки.", "option_8 - Играя игри.", "option_9 - Слушам музика.",  
+ "option_10 - Чета книги онлайн."), col = "black", bty = "n", cex = 0.7, text.width = 5)
```


Графика:



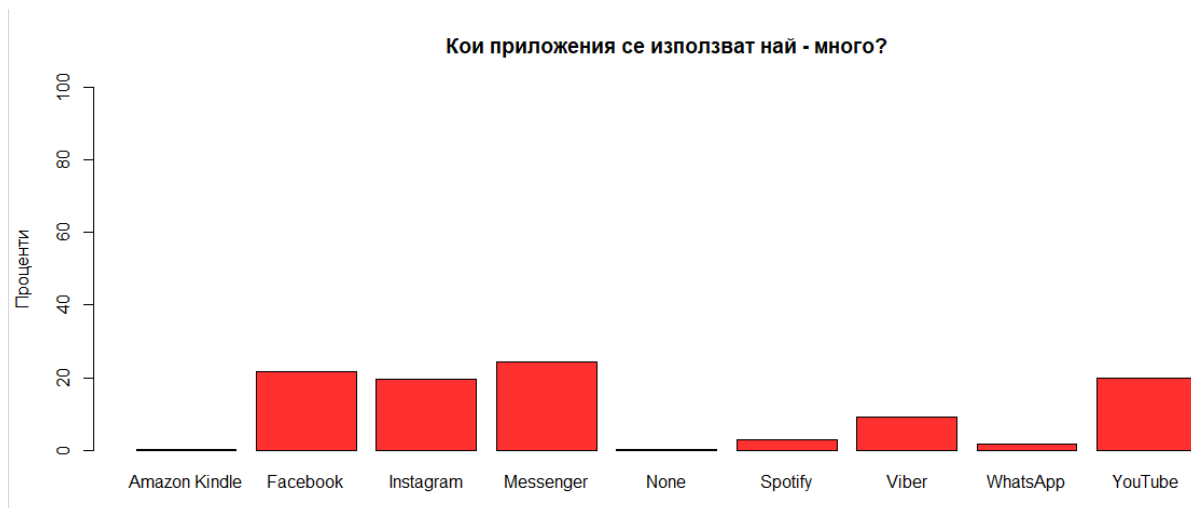
Изводът е, че телефоните се използват главно за общуване с близките – около 24% . На второ място – за образователни цели: 16%. След това – за слушане на музика и правене на снимки – под 15%. И най – малко за гледане на филми и четене на книги онлайн – едва 3%.

❖ Въпрос 5: От изброените приложения и социални мрежи, отбележи тези, които най - често използваш.

Въпросът също представлява категорийна променлива. В случая може да се избира повече от един възможен отговор. Код на R:

```
~/Project71852/   
> question_5 <- c(rep("Facebook", 73.1), rep("Instagram", 66.2), rep("Twitter", 0.0), rep("YouTube", 67.7),  
+ rep("Viber", 31.5), rep("Messenger", 82.3), rep("WhatsApp", 6.2), rep("Spotify", 10.8),  
+ rep("Amazon kindle", 1.5), rep("None", 1.5))  
>  
> table_q5 <- table(question_5)  
> str(table_q5)  
'table' int [1:9(1d)] 1 73 66 82 1 10 31 6 67  
- attr(*, "dimnames")=List of 1  
..$ question_5: chr [1:9] "Amazon kindle" "Facebook" "Instagram" "Messenger" ...  
> table_q5  
question_5  
Amazon Kindle      Facebook      Instagram      Messenger      None      Spotify      Viber      whatsapp  
1                73                66                82                1                10                31                6  
YouTube  
67  
> barplot(table_q5, col = "firebrick1", main = "кои приложения се използват най - много?", ylim = c(0,100), ylab = "проценти")
```

Графика:




Изводът е, че най – използваните приложения са: Messenger – 24%, Facebook – около 22%, YouTube – около 21% и веднага след него е Instagram – почти 20%. Най – малко използваното приложение от изброените е: Amazon Kindle – едва 2% - толкова от анкетираните не използват нито едно от посочените

приложения. От което можем да направим извод, че повечето от хората използват социалните мрежи.

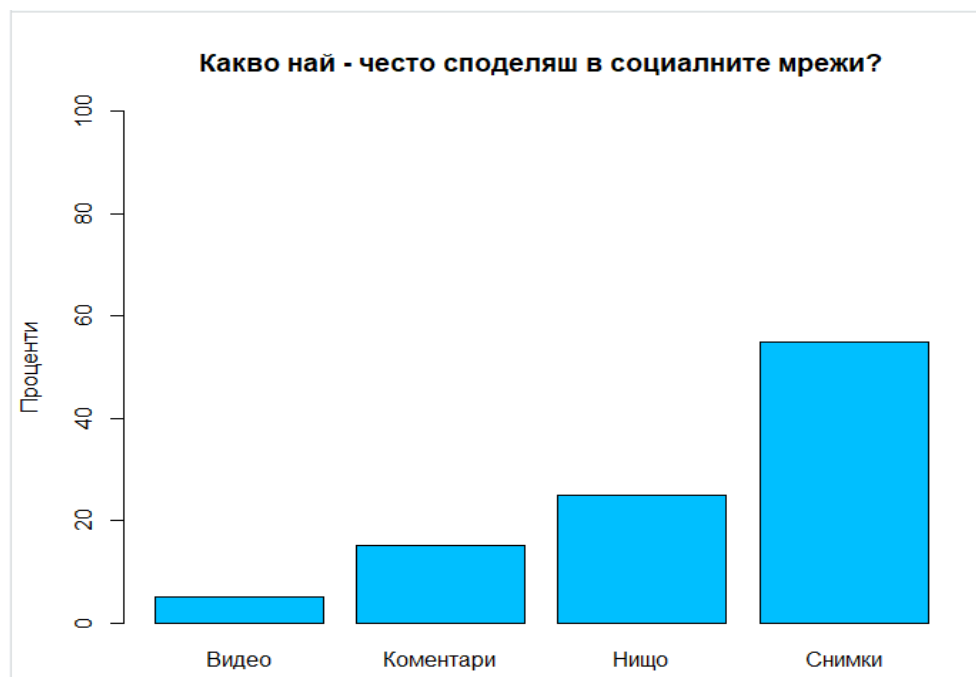
❖ Въпрос 6: Какво най - често споделяш в социалните мрежи?

В този въпрос имаме 4 различни категории, които описват дейността на анкетираните.

Код на R:

```
~/Project71852/   
> # Графично представяне  
> barplot(round(prop.table(table_q6)*100, 2), col = "deepskyblue1", main = "какво най - често споделяш в социалните мрежи?",  
+         ylim = c(0, 100), ylab = "проценти")  
- attr(*, "dimnames")=List of 1  
..$ question_6: chr [1:4] "Видео" "коментари" "нищо" "снимки"  
> table_q6  
question_6  
Видео  Коментари      Нищо    Снимки  
      5         15       25      55
```

Графика:



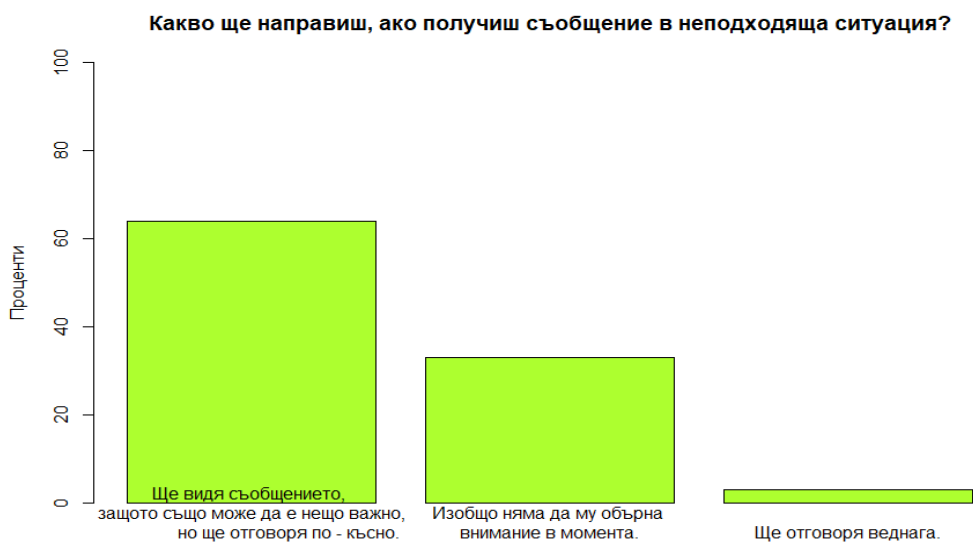
От графиката се вижда, че най – напред хората споделят снимки в социалните мрежи – 55 на брой. Най – малко – видео - едва 5 от всички анкетирани. От тях 25 на брой пък казват, че не споделят нищо. Останалите 15 човека обичат да пишат и коментари.

❖ **Въпрос 7: Представи си, че си на важна среща, или имаш неотложен ангажимент, който трябва да свършиш. В същия момент получаваш съобщение. Какво ще направиш?**

Целта на въпрос 7 от анкетата е да постави хората в реална ситуация, в която трябва да изберат между това да отговорят на съобщение, получено на техните мобилни телефони по време на важна среща, или да го игнорират. Въвеждане на данните за този въпрос става с помощта на кода:

```
~/Project71852/ <
> question_7 <- c(гер("ще отговоря веднага.", 3), гер("\пще видя съобщението, \пзащото също може да е нещо важно,
+             но ще отговоря по - късно.", 64), гер("Изобщо няма да му обърна \пвнимание в момента.", 33))
> table_q7 <- table(question_7)
> round(prop.table(table_q7)*100, 2)
question_7
\пще видя съобщението, \пзащото също може да е нещо важно,\п             но ще отговоря по - късно.             64
                                     Изобщо няма да му обърна \пвнимание в момента.             33
                                     Ще отговоря веднага.             3
> barplot(round(prop.table(table_q7)*100, 2), col = "greenyellow", main = "какво ще направиш, ако получиш съобщение в неподходяща ситуация?",
+         ylab = "проценти", ylim = c(0, 100))
```

Графика:



От графиката се вижда, че 64% от анкетираните първо биха завършили ангажиментите си и след това ще обърнат внимание на съобщението, което са получили. Едва около 3% ще отговорят веднага, а 33% дори не биха позволили едно съобщение да провали срещата им.

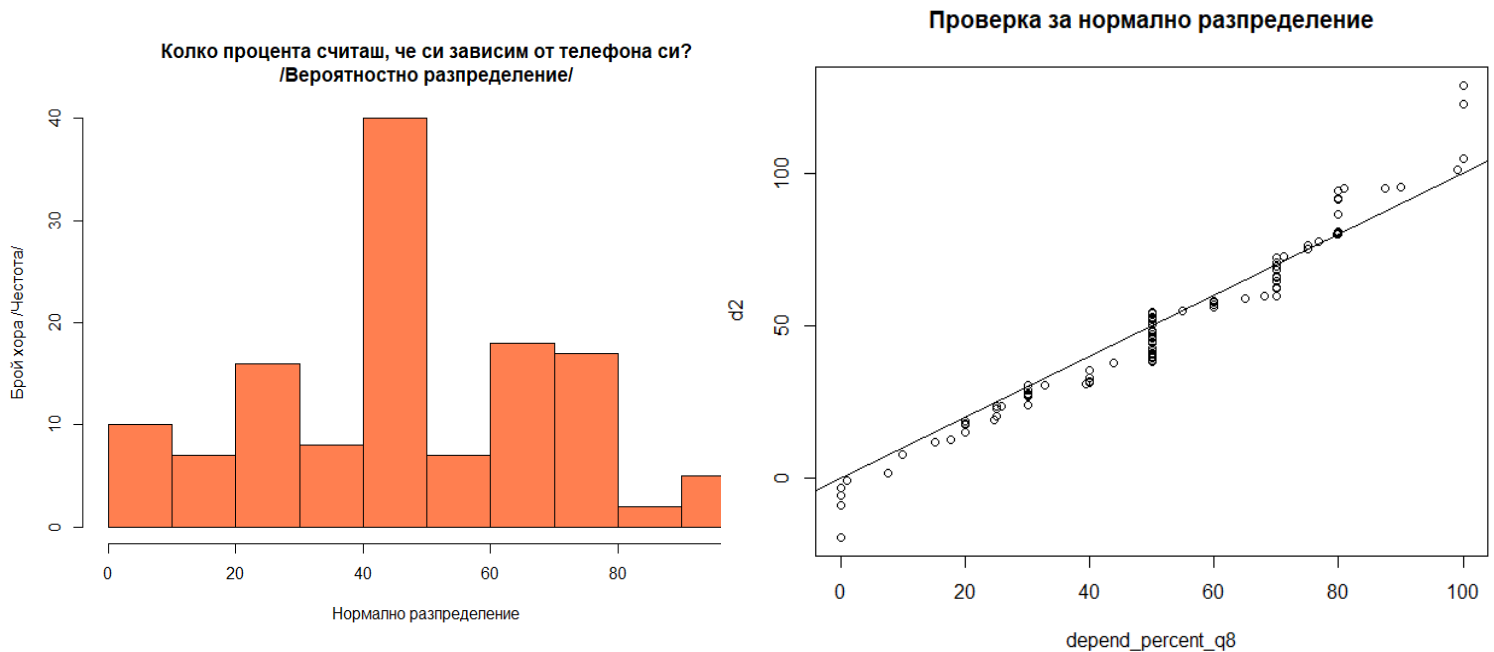
❖ Въпрос 8: В рамките на интервала 0 - 100 колко процента считаш, че си „зависим“ от телефона си?

Въпрос 8 представлява числова непрекъсната променлива, която ще анализирам по вече познатия начин. Първо, въвеждам данните и намирам средната им стойност, медианата и модата. Правя и описание на центъра на разпределението:

```
~/Project71852/
> depend_percent_q8 <- c(10,50,60,70,40,50,0,50,40,30,30,60,50,50,80,79,30,50,25,50,50,40,80,40,50,30,50,
+ 0,80,60,20,50,70,75,16,15,20,80,30,30,65,60,70,70,75,80,70,50,50,70,60,50,50,50,5,
+ 80,70,30,50,0,50,25,50,40,75,100,70,30,90,30,50,0,50,50,90,50,80,70,30,40,10,70)
~/Project71852/
> # Вариация на разпределението
> range(depend_percent_q8) #обхват
[1] 0 100
> var(depend_percent_q8) #дисперсия
[1] 579.2909
> sd(depend_percent_q8) #стандартно отклонение
[1] 24.06846
> fivenum(depend_percent_q8)
[1] 0 30 50 70 100
> # Графично представяне
> hist(depend_percent_q8, main = "Колко процента считаш, че си зависим от телефона си?\nВероятностно разпределение/",
+ xlab = "Нормално разпределение", ylab = "Брой хора /честота/", col = "coral")
>
> d2 <- rnorm(n = 10^2, mean = mean(depend_percent_q8), sd = sd(depend_percent_q8))
> qqplot(depend_percent_q8, d2, main = "Проверка за нормално разпределение")
> abline(a = 0, b = 1) # чертае линия
```

Сега ще проверя каква е вариацията на разпределението и стандартното отклонение. Накрая изобразявам разпределението на данните:

Графики:




Най – често срещаната стойност е 50%, т.е. половината от анкетираните се смятат за зависими от телефоните си.

❖ Въпрос 9: Смяташ ли, че мобилните устройства те отдалечават от приятелите и семейството ти?

Въпрос 9 представлява категорийна променлива, която предоставя три категории, от които анкетираните трябва да изберат само една.

Код на R:

```
~/Project71852/   
> question_9 <- c(rep("да", 10), rep("не", 79), rep("донякъде", 33))  
> table_q9 <- table(question_9)  
> round(prop.table(table_q9)*100, 2)  
question_9  
   да  донякъде   не  
  8.20   27.05  64.75  
> barplot(round(prop.table(table_q9)*100, 2), col = "orange", main = "Смяташ ли, че мобилните устройства \nte отдалечават от близките ти?",  
+         ylab = "проценти", ylim = c(0, 100))
```

Графика:



Очевидно, повечето от тях смятат, че мобилните телефони не са проблем, който може да застраши отношенията им с останалите – около 65%. Едва 8% пък са на обратното мнение. А 27% смятат, че вероятно това може да се окаже проблем.

❖ Въпрос 10: Колко време - 0 - 24 часа, би могъл/ла да издържиш без да използваш мобилния си телефон?

Този въпрос също е непрекъсната величина, тъй като хората имат право да въвеждат произволи числа в интервала 0 – 24 часа.

Първо, намирам средната стойност, медианата и модата. Също така правя описание на центъра на разпределени.

```
~/Project71852/ ↗  
> no_mobiles_q10 <- c(24,5,24,3,24,24,24,12,23,24,12,23,24,10,24,5,20,10,24,24,24,12,0,24,12,24,24,24,10,21,12,18,24,  
+ 24,24,12,12,1,24,6,24,17,10,5,10,24,5,0,12,12,23,24,12,24,1,24,24,24,24,20,10,24,3,5,2,4,24,24,16,  
+ 12,24,10,20,5,12,12,10,24,24,24,24,24,24,12,4,24,10,10,6,24,9,12,24,24,24,24,24,24,11,10,24,24,  
+ 23,3,10,10,12,24,4,24,24,24,24,24,24,4,24,10,24,24,24,20,22,24,24,24,15,4,4)  
> length(no_mobiles_q10)  
[1] 130  
> mean(no_mobiles_q10) #средна стойност  
[1] 16.92308  
> median(no_mobiles_q10) #медиана  
[1] 23  
> table_q10 <- table(no_mobiles_q10) #мода  
> names(table_q10)[table_q10 == max(table_q10)]  
[1] "24"  
> # Център на разпределението  
> summary(no_mobiles_q10)  
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.        
  0.00   10.00   23.00   16.92   24.00   24.00        
> quantile(no_mobiles_q10, prob = seq(0.1, 1.0, by = 0.1))  
10% 20% 30% 40% 50% 60% 70% 80% 90% 100%        
  .4  10  12  12  23  24  24  24  24  24  24
```

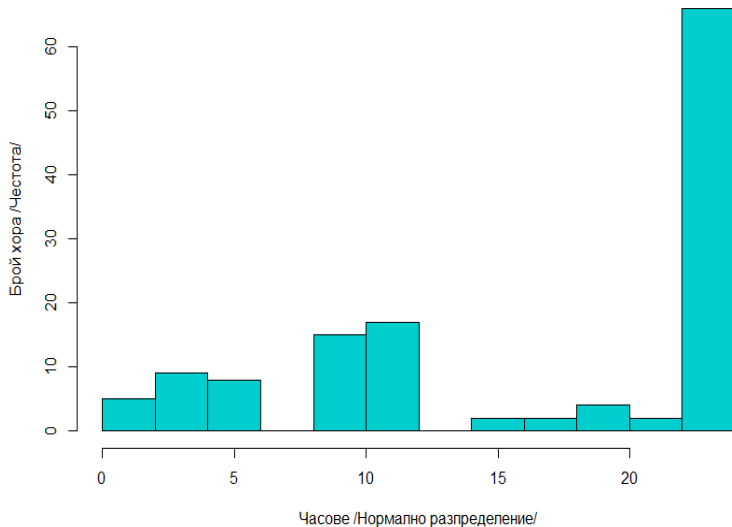
След това вариацията на разпределение (дисперсия), стандартното отклонение и изобразявам графично данните.

```
~/Project71852/ ↗  
> # Вариация на разпределението  
> range(no_mobiles_q10) #обхват  
[1] 0 24  
> var(no_mobiles_q10) #дисперсия  
[1] 65.68396  
> sd(no_mobiles_q10) #стандартно отклонение  
[1] 8.104564  
> fivenum(no_mobiles_q10)  
[1] 0 10 23 24 24
```

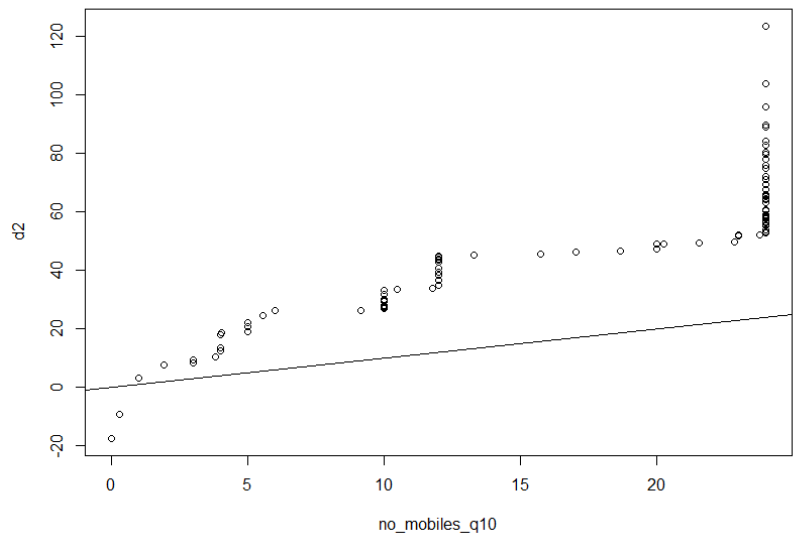
```
~/Project71852/ ↗  
> # Графично представяне  
> hist(no_mobiles_q10, main = "Колко часа, би могъл/ла да издържиш \nбез да използваш мобилния си телефон?  
+ /Вероятностно разпределение/", xlab = "часове /Нормално разпределение/",  
+ ylab = "Брой хора /честота/", col = "cyan3")  
> # Проверка за нормално разпределение  
> d3 <- rnorm(n = 100, mean = mean(no_mobiles_q10), sd = sd(no_mobiles_q10))  
> qqplot(no_mobiles_q10, d2, main = "Проверка за нормално разпределение")  
> abline(a = 0, b = 1)
```

Графики:

Колко часа, би могъл/ла да издържиш
без да използваш мобилния си телефон?
/Вероятностно разпределение/



Проверка за нормално разпределение



В заключение, можем да кажем, че все пак има хора, които не са зависими от мобилните си телефони, т.е. могат да издържат цял ден без тях. В случая те са над 60%. Това може да се определи и от факта, че най – често срещаната стойност, която те посочват, е 24 часа.

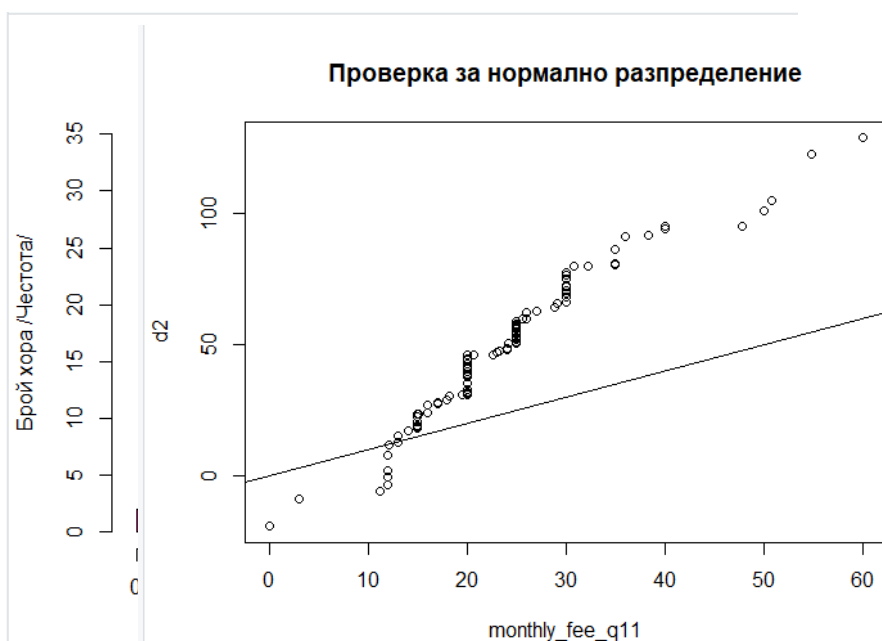
❖ Въпрос 11: Каква месечна такса плащаш, за да ползваш услугите на мобилния си оператор ?

Този въпрос представлява дискретна числова променлива, защото не е определен интервалът, в който е необходимо се отговаря. След обобщението на отговорите се вижда, че те са в интервала 0 – 60 лв. Следователно можем да я разглеждаме като непрекъсната променлива и да я изследваме по познатия начин:

```
~/Project71852/
> monthly_fee_q11 <- c(15,13,56,35,20,20,10,14,25,30,15,12,15,19,36,25,20,20,25,25,25,0,20,20,25,25,30,20,25,24,30,
+ 20,26,30,15,25,35,30,12,52,40,20,50,30,32,24,25,25,20,12,26,22,15,13,25,25,29,20,18,30,30,36,26,
+ 20,26,25,25,25,50,50,15,20,25,17,25,28,25,12,30,20,17,17,12,40,20,12,12,20,35,25,40,30,16,25,30,30,25,30,
+ 20,16,23,23,17,40,40,25,20,29,30,20,16,15,20,20,60,20,0,18,24,20,23,24,30,20,20,15,13,31,15,35)
> length(monthly_fee_q11)
[1] 130
> mean(monthly_fee_q11) #средна стойност
[1] 24.21538
> median(monthly_fee_q11) #медиана
[1] 24
> table_q11 <- table(monthly_fee_q11) #мода
> names(table_q11)[table_q11 == max(table_q11)]
[1] "20"
```

```
~/Project71852/
> # Вариация на разпределението
> range(monthly_fee_q11) #обхват
[1] 0 60
> var(monthly_fee_q11) #дисперсия
[1] 95.93775
> sd(monthly_fee_q11) #стандартно отклонение
[1] 9.794782
> fivenum(monthly_fee_q11)
[1] 0 19 24 30 60
> # Графично представяне
> hist(monthly_fee_q11, main = "Каква месечна такса заплащаш?\nВероятностно разпределение/",
+ xlab = "Такса /Нормално разпределение/", ylab = "Брой хора /честота/", col = "deeppink")
> # проверка за нормално разпределение
> d3 <- rnorm(n = 108, mean = mean(monthly_fee_q11), sd = sd(monthly_fee_q11))
> qqplot(monthly_fee_q11, d2, main = "Проверка за нормално разпределение")
> abline(a = 0, b = 1)
```

Графики:



такса 10–20 лв. и 25–30лв.

Изводи:

Най – често срещаната стойност сред отговорите е 20 лв.

Средната стойност от всички отговори е около 24 лева.

Около 30% от хората плащат на месец между 20-25лв.

Между 20–25% заплащат

2.2. Проверка за нормално разпределение.

След като вече въведох и изобразих данните, ще използвам `shapiro.test` за всеки въпрос, за да проверя дали са равномерно разпределени, а за тези, за които вече съм направила проверка, ще докажа, че тя е правилна. Тъй като на втория въпрос отговорът е един, а именно „Да, имам“, то считаме, че отговорите са нормално разпределени.

Нека хипотезата H_0 гласи, че отговорите са равномерно разпределени.

Нивото на съгласие е $\alpha = 0,05$.

```
~/Project71852/ ↗
> shapiro.test(table_q1)

      shapiro-wilk normality test

data:  table_q1
W = 0.72425, p-value = 0.01689

> shapiro.test(table_q3)

      shapiro-wilk normality test

data:  table_q3
W = 0.85482, p-value = 0.0101

> shapiro.test(table_q4)

      shapiro-wilk normality test

data:  table_q4
W = 0.87528, p-value = 0.1151

> shapiro.test(table_q5)

      shapiro-wilk normality test

data:  table_q5
W = 0.8341, p-value = 0.04964

> shapiro.test(table_q6)

      shapiro-wilk normality test

data:  table_q6
W = 0.92708, p-value = 0.5774
```

```
~/Project71852/ ↗
> shapiro.test(table_q7)

      shapiro-wilk normality test

data:  table_q7
W = 0.99991, p-value = 0.9819

> shapiro.test(depend_percent_q8)

      shapiro-wilk normality test

data:  depend_percent_q8
W = 0.96358, p-value = 0.001469

> shapiro.test(table_q9)

      shapiro-wilk normality test

data:  table_q9
W = 0.96429, p-value = 0.6369

> shapiro.test(no_mobiles_q10)

      shapiro-wilk normality test

data:  no_mobiles_q10
W = 0.78822, p-value = 2.027e-12

> shapiro.test(monthly_fee_q11)

      shapiro-wilk normality test

data:  monthly_fee_q11
W = 0.92202, p-value = 1.384e-06
```

Забелязваме, че стойностите за p са различни. Тези, чиито p – value са по – големи от нивото на съгласие, считаме разпределенията им за нормални. За нормално разпределените променливи ще използваме параметрични тестове, а за останалите – непараметрични. Забелязваме, че стойността на p за Въпрос 4, Въпрос 6, Въпрос 7 и Въпрос 9 са над нивото на съгласие. Следователно, можем да считаме, че отговорите на тези въпроси са нормално разпределени.

2.1. Анализ на взаимодействието между две променливи.

- ✓ Категорийна vs Категорийна
- ✓ Категорийна vs Числова
- ✓ Числова vs Категорийна
- ✓ Числова vs Числова

За изпълнението на тази част ще направя проверка за това дали има зависимост между две променливи, като използвам различни хипотези. За графичното им изобразяване ще използвам различни техники.

1. Категорийна vs Категорийна

Ще изследвам: Въпрос 4: С каква цел най - често използваш мобилното си устройство? и Въпрос 6: Какво най - често споделяш в социалните мрежи? – и двата представляват категорийни величини.

```
~/Project71852/
> sample_q4 <- sample(x = question_4, size = 100, replace = TRUE)
> table(sample_q4, question_6)
```

sample_q4	Видео	Коментари	Нищо	Снимки
Водя си бележки на телефона.	1	1	4	4
Гледам филми.	1	0	1	0
За образователни цели.	1	3	4	5
Играя игри.	0	1	1	2
Общувам със семейството \ни приятелите си.	0	1	3	12
По - голяма част от\пвремето си прекарвам \пв социалните мрежи.	0	2	5	10
Правя снимки.	0	4	3	8
Провеждам бизнес разговори.	1	1	1	4
Слушам музика.	1	2	3	7
Чета книги онлайн.	0	0	0	3

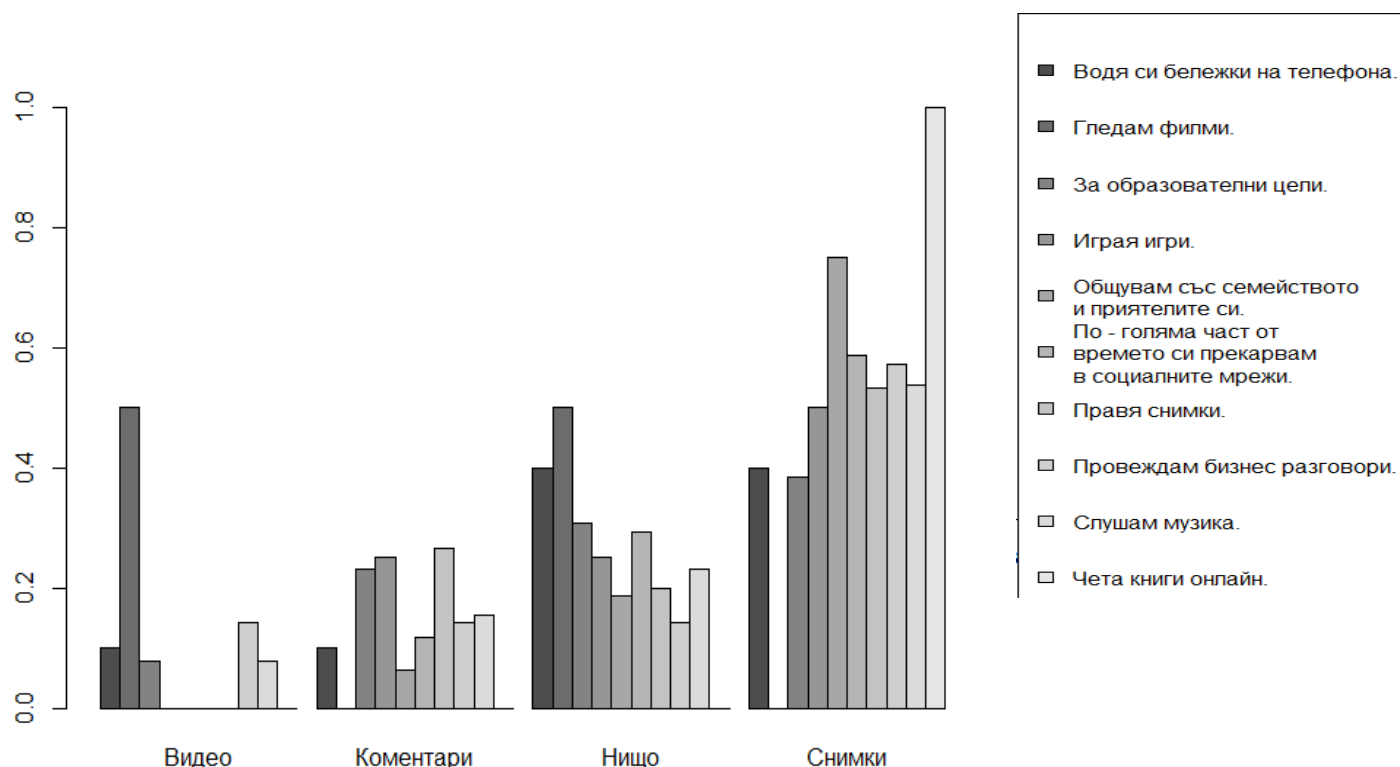
```
> prop.table(x = table(sample_q4, question_6), margin = 1)
```

sample_q4	Видео	Коментари	Нищо	Снимки
Водя си бележки на телефона.	0.10000000	0.10000000	0.40000000	0.40000000
Гледам филми.	0.50000000	0.00000000	0.50000000	0.00000000
За образователни цели.	0.07692308	0.23076923	0.30769231	0.38461538
Играя игри.	0.00000000	0.25000000	0.25000000	0.50000000
Общувам със семейството \ни приятелите си.	0.00000000	0.06250000	0.18750000	0.75000000
По - голяма част от\пвремето си прекарвам \пв социалните мрежи.	0.00000000	0.11764706	0.29411765	0.58823529
Правя снимки.	0.00000000	0.26666667	0.20000000	0.53333333
Провеждам бизнес разговори.	0.14285714	0.14285714	0.14285714	0.57142857
Слушам музика.	0.07692308	0.15384615	0.23076923	0.53846154
Чета книги онлайн.	0.00000000	0.00000000	0.00000000	1.00000000

Започвам със създаването на извадка от Въпрос 4, тъй като на него могат да се дават по няколко отговора и дължината на целия вектор е доста по – голяма от тази на question_6. След това представям данните от двата въпроса в обща таблица и чрез prop.table() става ясно колко процента в коя група попадат. Вижда се, че има зависимост между някои от отговорите. С помощта на

```
~/Project71852/
> # Графично представяне
> barplot(prop.table(x = table(sample_q4, question_6), margin = 1), beside = TRUE)
```

barplot представям данните графично:



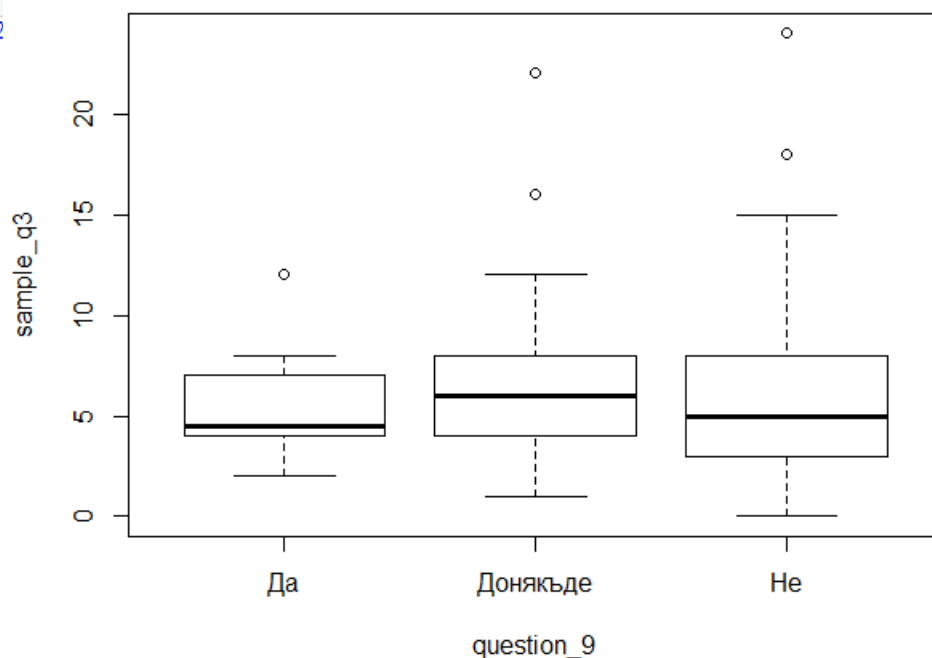
От графиката и от таблицата става ясно, че хората, които използват мобилните си телефони най – често, за да общуват със семейството и приятелите си, също ги използват и за споделяне на снимки в социалните мрежи – 0,75%. Тези пък, които ги използват за да си водят бележки, споделят и видео – 0,8%, и коментари – 0,02%, но в повечето случаи не са особено активни – 0,3%. Хората, прекарващи по – голяма част от времето си в социалните мрежи, споделят най – вече снимки – почти 0,6%.

2. Категорийна vs Числова

Ще изследвам: Въпрос 3: По колко часа на ден в интервала 0 - 24 използваш телефона си? и Въпрос 9: Смяташ ли, че мобилните устройства те отдалечават от приятелите и семейството ти? с помощта на boxplot. Тази техника се използва за обобщаване на данните и бързо определяне на това дали те са симетрични или имат outliers.

```
~/Project71852/
```

```
> sample_q3 <- sample(x = active_hours_q3, size = 12)
> tt <- boxplot(sample_q3~question_9)
```



Удебелената черна линия във всяка група е медианата. От двете страни на медианата са съответно първи и трети квантил. Дължините на опашките пък представляват минималната и максималната стойност.

В първата група медианата се доближава до минималната стойност. Във втората група медианата съвпада със средната стойност от отговорите. А в третата група стойностите са най – големи, защото отговорът „Не“ преобладава и както може да се види от графиката, максималната стойност е по – далеч от медианата.

Изводът е, че мобилните телефони не пречат на отношенията между хората.

3. Числова vs Категорийна

Нека да разгледаме: Въпрос 8: В рамките на интервала 0 - 100 колко процента считаш, че си „зависим“ от телефона си? и Въпрос 1: На колко години си?

Ще използвам Wilcoxon signed rank test понеже въпросите не са с нормално разпределени отговори, защото минималната стойност на p за двата вектора е $0,01689 < \alpha = 0,05$. Нулевата хипотеза H_0 е следната: Зависимостта на хората от телефоните има общо с възрастта им.

Алтернативната хипотеза H_1 е: Няма нищо общо.

Нивото на съгласие е $\alpha: 0,05$.

```
~/Project71852/ ↗
> wilcox.test(x = depend_percent_q8, y = table_q1, alternative = "greater", conf.int = TRUE)

wilcoxon rank sum test with continuity correction

data: depend_percent_q8 and table_q1
W = 498.5, p-value = 0.02049
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 7.999983      Inf
sample estimates:
difference in location
 30.00007
```

Стойността на $p = 0.02049 < 0.05$. Въпреки, че доверителният интервал е 95%, изводът е, че няма значение на каква възраст е един човек, за да може да развие зависимост към мобилния си телефон.

4. Числова vs Числова

Първо, ще направя един пример с корелационен анализ. След това още един пример с регресионен анализ.

4.1. Корелационен анализ

Въпрос 3: По колко часа на ден в интервала 0 - 24 използваш телефона си?

Въпрос 8: В рамките на интервала 0 - 100 колко процента считаш, че си „зависим“ от телефона си?

Изобразявам графично връзката между двете променливи. Вижда се, че тя не е линейна и не мога да използвам линеен модел, за да моделирам връзката.

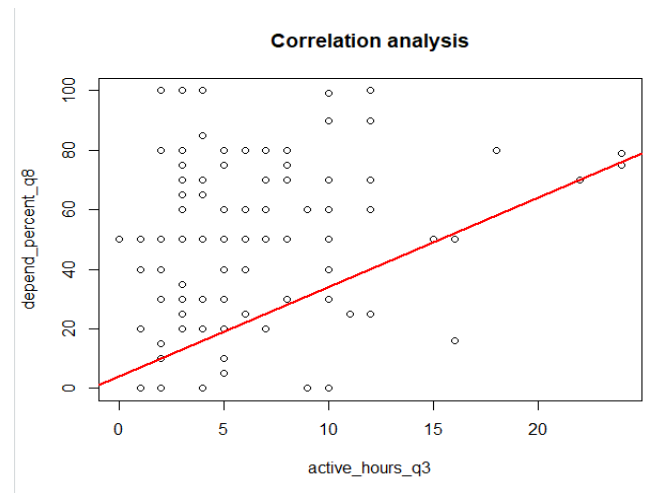
```
~/Project71852/ ↗
> rho <- round(cor(active_hours_q3, depend_percent_q8), 3) #коефициент на корелация
> par(mfrow = c(1, 1))
> plot(active_hours_q3, depend_percent_q8, main = "Correlation analysis")
> abline(a = 4, b = 3, col = "red", lwd = 2)
> cor(active_hours_q3, depend_percent_q8)
[1] 0.2333105
> cor.test(active_hours_q3, depend_percent_q8, method = "spearman")

spearman's rank correlation rho

data: active_hours_q3 and depend_percent_q8
S = 279633, p-value = 0.006802
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.2362773
```

Ще направя корелационен анализ, за да измеря силата на връзката им.

ρ_{ho} представлява коефициент на корелация и в случая той показва, че има много слаба корелация (зависимост) между тези два въпроса. Това може да се види и на графиката:



Изводът, който може да се направи е, че броят на часовете, които хората прекарват на телефоните си, не може да покаже със сигурност дали те са зависими от тях, или не.

4.2. Регресионен анализ

Въпрос 3: По колко часа на ден в интервала 0 - 24 използваш телефона си?

Въпрос 11: Каква месечна такса плащаш, за да ползваш услугите на мобилния си оператор ?

Ще изследвам двата въпроса с помощта на линейна регресия, за да видя дали има връзка между тях. Нулевата хипотеза е: Хората, които използват повече мобилните си телефони, плащат по – големи такси. Алтернативната хипотеза - H_1 - е, че това не е така.

H_0 се отхвърля при стойност на $p\text{-value} < 0.05$.

Най – напред въвеждам данните от двата въпроса в data frame (DF), като първо изравнявам дължините на двата вектора. Построяваме линейния модел по следния начин:

```
~/Project71852/ ↗  
> DF <- data.frame(active_hours_q3, monthly_fee_q11)  
> model <- lm(active_hours_q3~monthly_fee_q11)  
> model  
  
Call:  
lm(formula = active_hours_q3 ~ monthly_fee_q11)  
  
Coefficients:  
    (Intercept)  monthly_fee_q11  
          5.40803             0.03842
```

Коефициентът пред monthly_fee_q11 е 0.03842:

$$\text{active_hours_q3} = 0.03842 * \text{monthly_fee_q11} + 5.40803$$

След това ще проверим до колко този модел описва добре данните и какви са оценките на коефициентите му:

```
~/Project71852/ > summary(model)

Call:
lm(formula = active_hours_q3 ~ monthly_fee_q11)

Residuals:
    Min       1Q   Median       3Q      Max
-6.5607 -3.1765 -0.9844  1.6218 17.6314

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.40803    1.02689   5.266 5.71e-07 ***
monthly_fee_q11 0.03842    0.03933   0.977  0.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

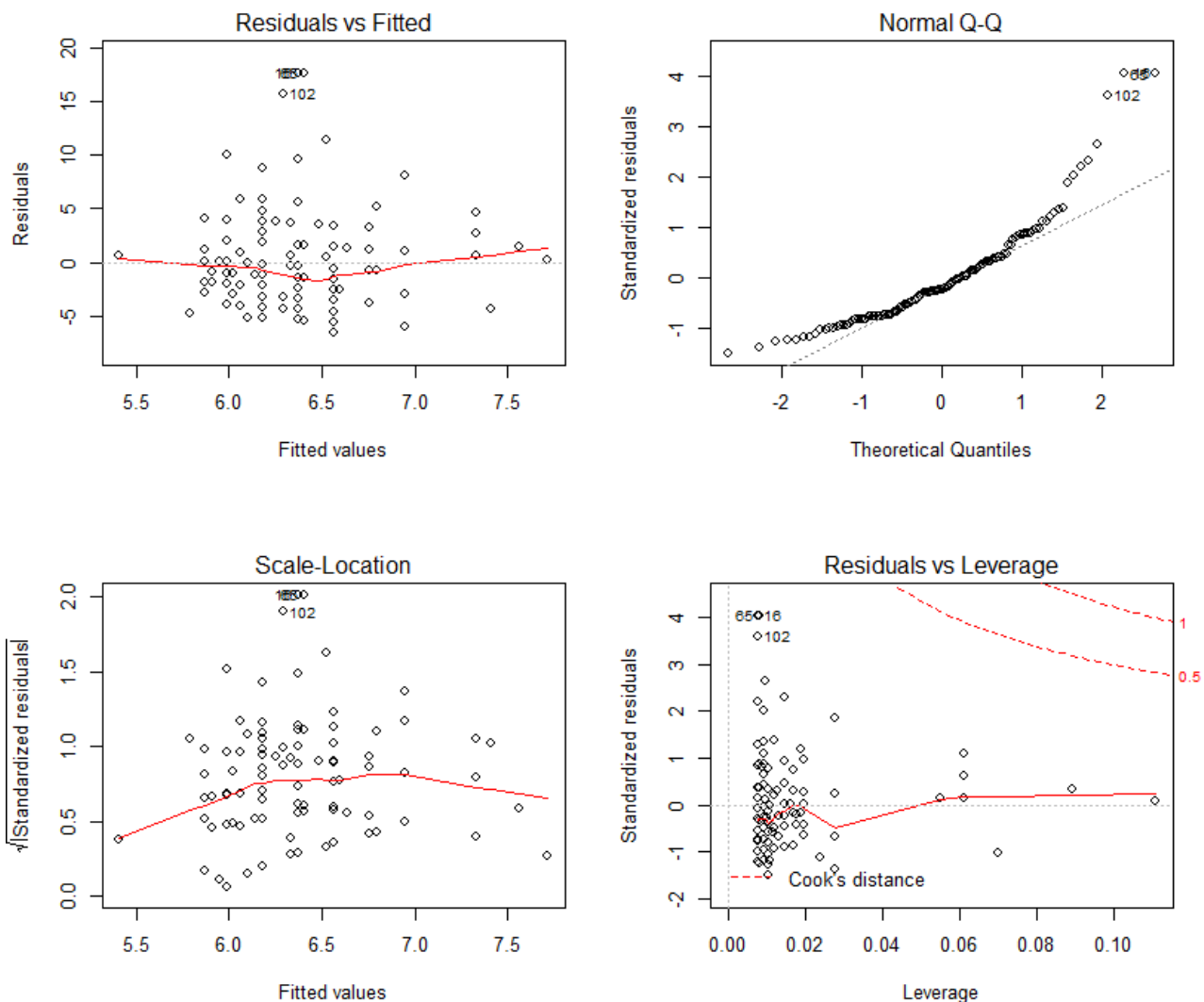
Residual standard error: 4.376 on 128 degrees of freedom
Multiple R-squared:  0.0074,    Adjusted R-squared:  -0.0003548
F-statistic: 0.9542 on 1 and 128 DF,  p-value: 0.3305
```

Стойностите на всички коефициенти са различни от 0. Тъй като стойността на p (p – value) в колоната Pr(>|t|) за единия параметър е $5.71e-07 \ll 0.05$, то той е статистически значими. Коефициентът пред monthly_fee_q11 обаче е $0.33 > 0.05$. Тоест, той може да отпадне от анализа.

Статистиката “Multiple R-squared“ или “Adjusted R-squared“ показват колко добре моделът описва данните. И двете приемат стойности в интервала [0, 1]. Стойността на първата статистика е далеч от 1 и това означава, че моделът не е толкова добър. Стойността на “Adjusted R-squared“ е отрицателна, тъй като тя приема и статистически незначими стойности.

Използвам функцията resid() за създаване на списък с остатъците и след това правя графика за тях

```
~/Project71852/ > resid(lm(active_hours_q3~monthly_fee_q11))
 1      2      3      4      5      6      7      8      9     10     11     12
-0.98437791 -0.90753172 1.44027507 1.24716013 -0.17649340 -1.17649340 -4.79226242 0.05404518 -5.36860889 -4.56072438 -1.98437791 0.13089138
13     14     15     16     17     18     19     20     21     22     23     24
 0.01562209 -1.13807030 -0.79126297 17.63139111 1.82350660 -3.17649340 -3.36860889 -3.36860889 -4.36860889 0.59196855 -3.17649340 -1.17649340
25     26     27     28     29     30     31     32     33     34     35     36
-3.36860889 -1.36860889 -1.56072438 -4.17649340 -1.36860889 3.66981421 -5.56072438 1.82350660 1.59296801 -3.56072438 10.01562209 -4.36860889
37     38     39     40     41     42     43     44     45     46     47     48
-3.75283987 -3.56072438 4.13089138 -4.40603253 -2.94495536 -3.17649340 4.67081366 1.43927562 1.36242942 0.66981421 -3.36860889 9.63139111
49     50     51     52     53     54     55     56     57     58     59     60
 8.82350660 4.13089138 -1.40703199 3.74666040 4.01562209 -1.90753172 -1.36860889 -0.36860889 0.47769872 -2.17649340 -5.09964721 -2.56072438
61     62     63     64     65     66     67     68     69     70     71     72
-1.56072438 5.20873703 -5.40703199 3.82350660 17.59296801 -3.36860889 -2.36860889 -1.36860889 2.67081366 0.67081366 -0.98437791 3.82350660
73     74     75     76     77     78     79     80     81     82     83     84
-2.36860889 -2.06122411 5.63139111 3.51612181 1.63139111 -2.86910862 -3.56072438 -4.17649340 -4.06122411 0.93877589 -1.86910862 -2.94495536
85     86     87     88     89     90     91     92     93     94     95     96
-1.17649340 0.13089138 1.13089138 3.82350660 3.24716013 -3.36860889 8.05504464 1.43927562 -1.02280101 -0.36860889 -0.56072438 -3.56072438
97     98     99     100    101    102    103    104    105    106    107    108
-2.36860889 3.43927562 -3.17649340 -3.02280101 -4.29176270 15.70823730 5.93877589 -5.94495536 1.05504464 -1.36860889 -2.17649340 11.47769872
109    110    111    112    113    114    115    116    117    118    119    120
-6.56072438 -3.17649340 -1.02280101 -3.98437791 4.82350660 2.82350660 0.28658268 -5.17649340 0.59196855 -0.09964721 -0.33018579 1.82350660
121    122    123    124    125    126    127    128    129    130
-3.29176270 0.66981421 -3.56072438 -0.17649340 5.82350660 2.01562209 -0.90753172 -2.59914748 2.01562209 -0.75283987
```



Residuals vs Fitted – Графика на разсейване на остатъците по оста y и прогнозните отговори по оста x . Използва се за откриване на нелинейност, неравномерни отклонения и външни разходи. Вижда се, че остатъците са разпръснати на случаен принцип около остатъчния ред $= 0$. Това предполага, че връзката е линейна. Освен това остатъците образуват „хоризонтална лента“ около остатъчната линия и това означава, че вариантите на условията за грешки са равни. Има няколко остатъци, които се открояват от основния модел, но това не винаги е показател за наличие на проблем.

Normal Q-Q – Метод на най – малките квадрати, който показва, че остатъците не са нормално разпределени спрямо прекъснатата линия, тъй като се наблюдават тежки опашки. Това може да се докаже и чрез `shapiro.test(res)`, където `res` е променлива за остатъците.

Scale – Location – Тази графика (на мащабното местоположение) е подобна на първата, но опростява анализа на предположението за хомоскедастичност. Той отнема квадратния корен на абсолютната стойност на стандартизираните остатъци вместо да начертава самите остатъци. Червената линия е приблизително хоризонтална. Това означава, че средната величина на стандартизираните остатъци не се променя много като функция на `fitted` стойностите. Освен това тази линия не се променя особено в зависимост от тези стойности. Значи променливостта на величините не се променя много като функция от `fitted` стойностите.

Residuals vs Leverage – Тази графика се използва за откриване на хетероскедастичност и нелинейност. Разпространението на стандартизираните остатъци не трябва да се променя като функция на `leverage`. В случая тази функция намалява, което показва наличието на хетероскедастичност. Точките с висок `leverage` (лост) могат да повлияят на модела, тъй като тяхното изтриване би могло да го промени изцяло. Затова разглеждаме разстоянието на Кук, което измерва ефекта от изтриването на точка върху комбинирания вектор на параметъра. В този случай няма точки извън пунктираната линия, затова не можем и да определим тяхното влияние.

С разглеждането на тези няколко графики установихме, че в този линеен модел има наличие на хетероскедастичност, тъй като остатъците не са разпръснати равномерно по цялата графика. Освен това не е изпълнено условието за нормално разпределение на грешката, което означава, че оценките на коефициентите на линейната регресия не са от най - добрите. А с помощта на `dwtest()` установяваме, че няма наличие на автокорелация на грешките, защото $p - value = 0,4392 > 0,05$.

```
~/Project71852/ ↗  
> library(lmtest)  
> dwtest(model)  
  
Durbin-watson test  
  
data: model  
DW = 1.9754, p-value = 0.4392  
alternative hypothesis: true autocorrelation is greater than 0
```


Част 3: Заключение.

С помощта на различните подходи – тестове и графики за анализ на променливите и взаимодействието между тях, стигнах до някои важни изводи, които отговарят на въпросите, поставени още в началото. Целта на проекта беше да се установи дали хората са зависими от мобилните си телефони и това влияе ли на отношенията им с близките.

1. Резултатите показват, че всички – 130 анкетирани, притежават мобилен телефон, който най – често използват за общуване с приятелите и семейството си.
2. Най – използваната социална мрежа според отговорите е Messenger, която е именно с такава цел.
3. Според анкетираните обаче фактът, че прекарват повече време на телефоните си, не е от значение за отношенията им с околните.
4. Това, че на ден човек прекарва средно по 6 часа, използвайки мобилния си телефон, не го прави зависим.
5. Средната месечна такса, която плащат хората, е около 24 лв.

В заключение може да се каже, че няма логическа свързаност между отговорите на въпросите. Човек винаги може да избегне зависимостта между него и мобилния телефон.

Използвана литература:

simpleR - Using R for Introductory Statistics – J. Verzani

Упражнения по СЕМ - практикум