**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 избера 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:

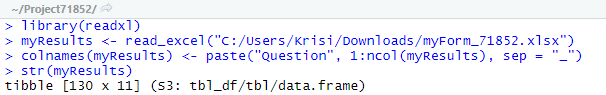
* За какво най – често хората използват телефоните си?
* Коя е най – използваната от тях социална мрежа?
* Каква такса средно заплащат на месец, за да използват услуги от мобилните си оператори?

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 въведа 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.

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 ралична дължина.

За по – добра четливост задавам имена нa 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)

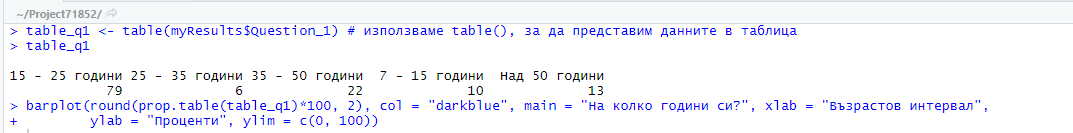
- Вариация (дисперсия) и стандартно отклонение

- 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 преставлява 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:



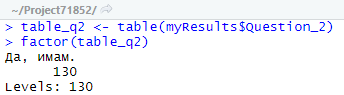
Графика:

Отговорите на този въпрос са 130.

От графиката се вижда, че от анкетираните:

* около 60% са на възраст 15 – 25 години
* 5% са тези в интервала 25 – 35 години
* 10% над 50 години
* 10% от анкетираните са в интервала 7 – 15 години
* 15% са в интервала 25 – 50 години
* Въпрос 2: Имаш ли мобилен телефон, таблет или друго уствойство, което използваш за комуникация и за връзка с Интернет?

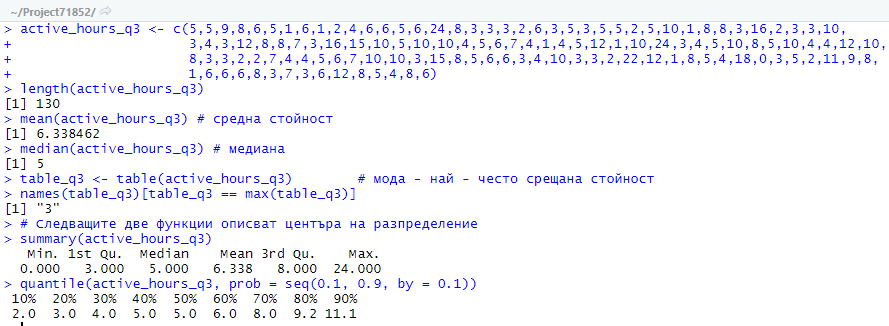
След прочитането на отговорите от втория въпрос се вижда, че всичките 130 човека са отговорили с „Да, имам“, което означава, че 100% от анкетираните притежават мобилен телефон или друго устройство за комуникация и връзка с Интернет. Това можем да го покажем, използвайки следния код на R:



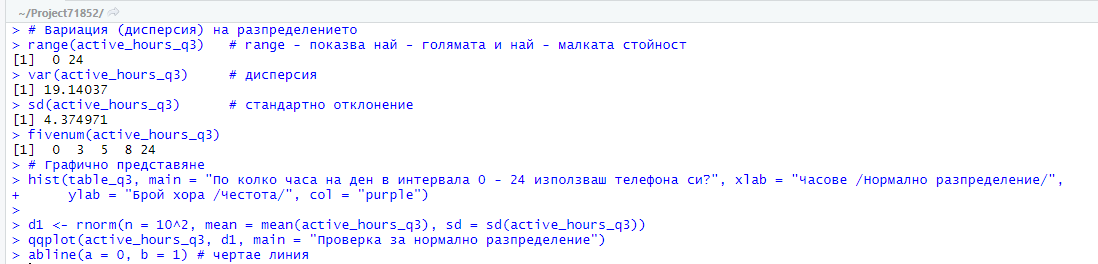
Функцията factor() показва, че данните са разпределени в 130 нива (130 отговора) и в случая на всяко едно ниво има един и същ отговор.

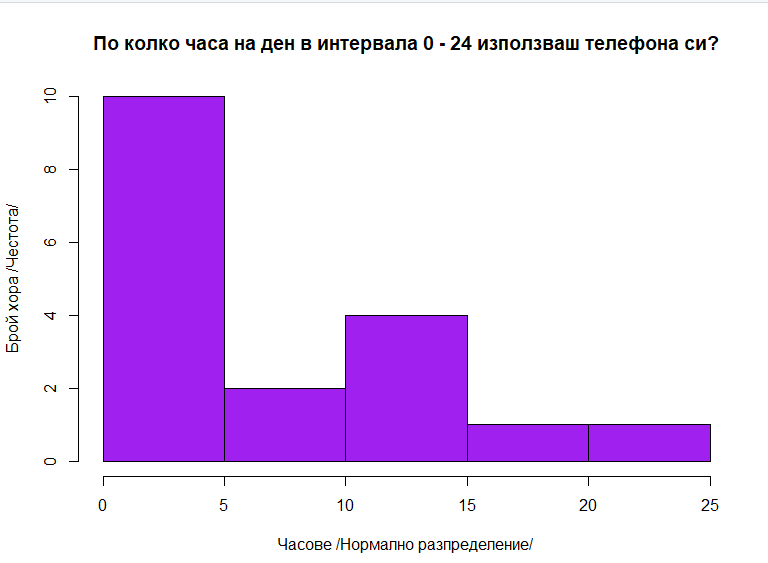
* Въпрос 3: По колко часа на ден в интервала 0 - 24 използваш телефона си?

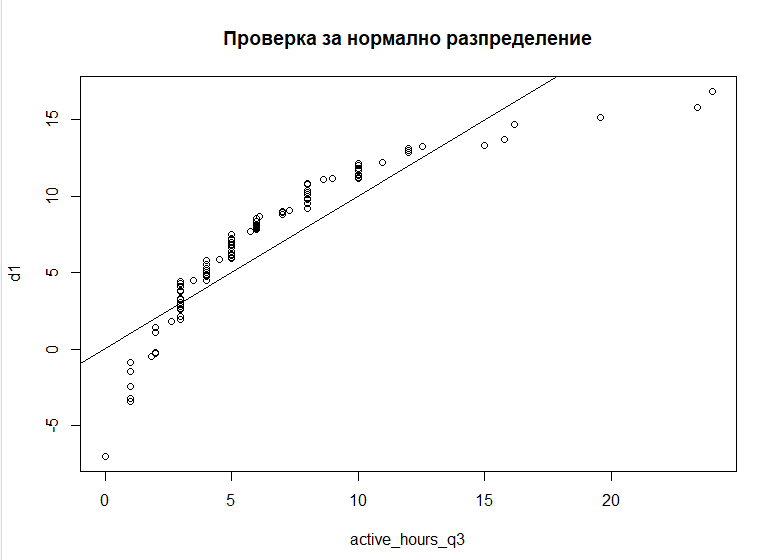
Третият въпрос представлява числова непрекъсната променлива, която ще изследвам, следвайки стъпките, споменати по – горе. Започвам с въвеждането на всички отговори във вектор. След това ще намеря средната стойност, медианата и модата, както и центъра на разпределение:



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.

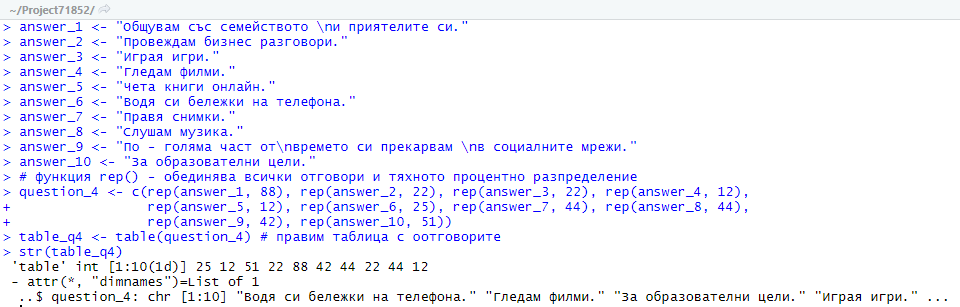
Код на R:

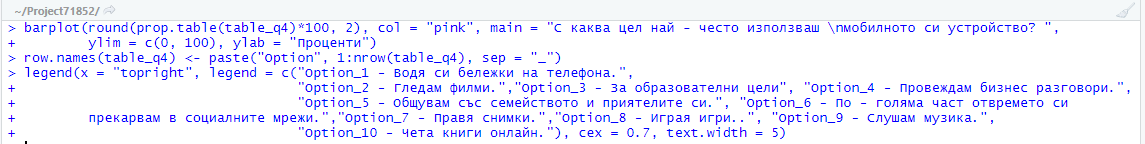
Графики:



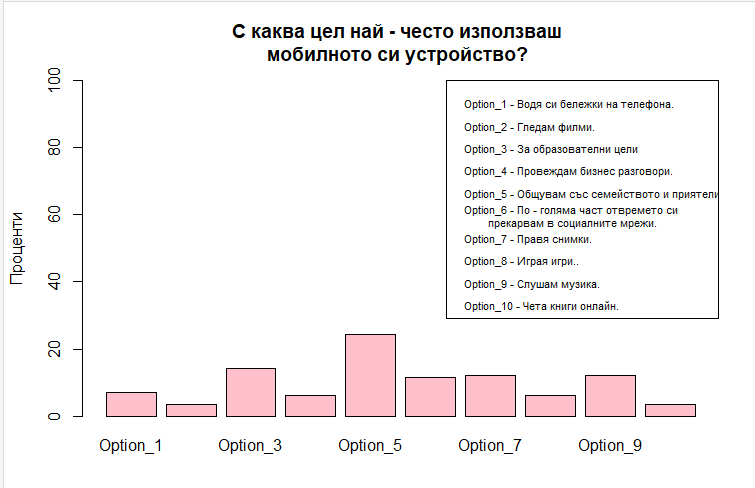
Повечето от хората прекарват меду 0 – 5 часа дневно над телефоните си, като най – честият отговор е 3 часа. Изчислявайки средната стойност от всички отговори, става ясно, че средно на човек се падат по около 6 часа на ден, което е доста тревожен факт. Освен това отговорите не са нормално разпределени,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

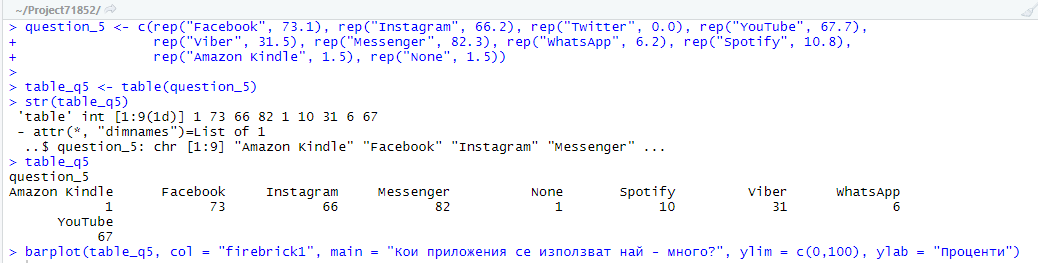
За графичното представяне на данните използвам barplot и добавям легенда:

Графика:

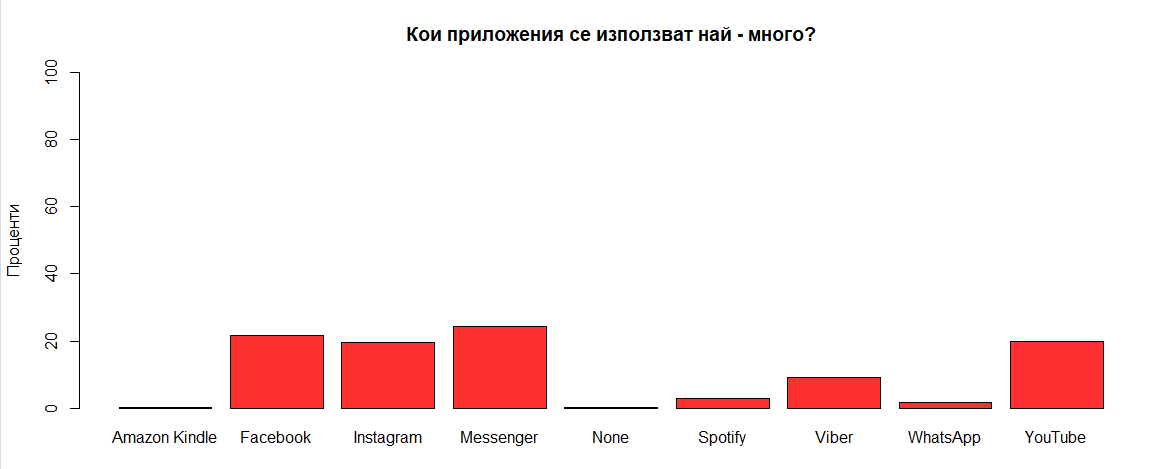


Изводът е, че телефоните се използват главно за общуване с близките – около 24% . На второ място – за образователни цели: 16%. След това – за слушане на музика и правене на снимки – под 15%. И най – малко за гледане на филми и четене на книги онлайн – едва 3%.

* 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:

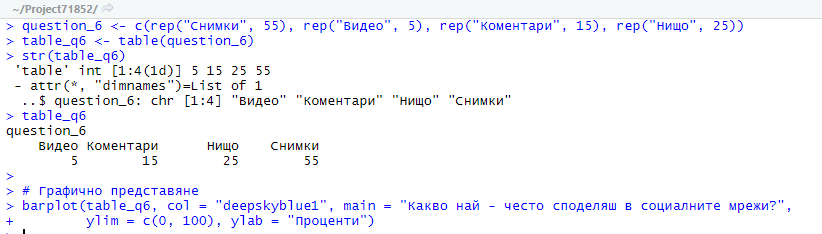
Графика:

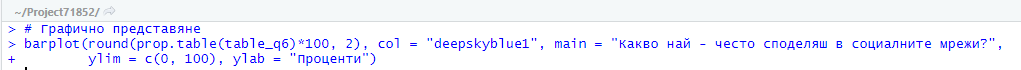


Изводът е, че най – използваните приложения са: Messenger – 24%, Facebook – около 22%, YouTube – около 21% и веднага след него е Instagram – почти 20%. Най – малко използваното приложение от изброените е: Amazon Kindle – едва 2% - толкова от анкетираните не използват нито едно от посочените приложения.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

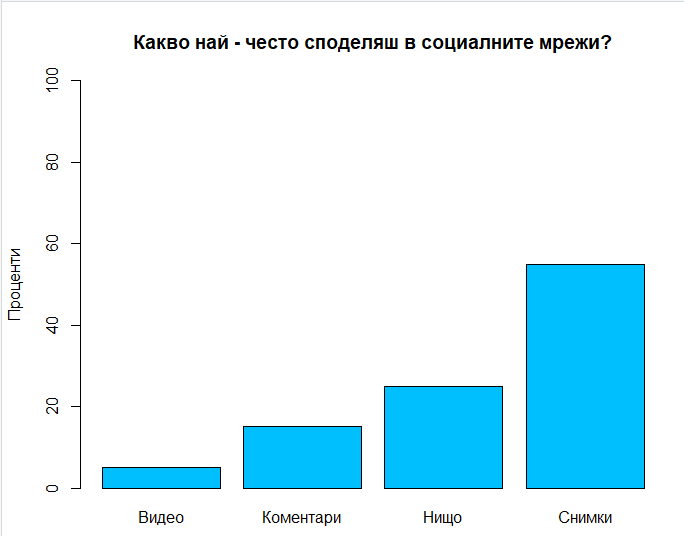
* Въпрос 6: Какво най - често споделяш в социалните мрежи?

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

Код на R:

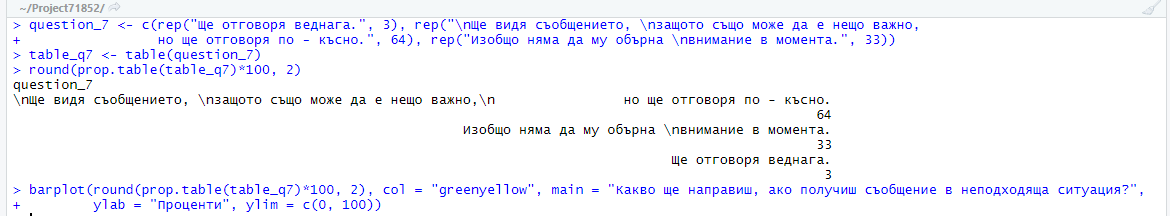


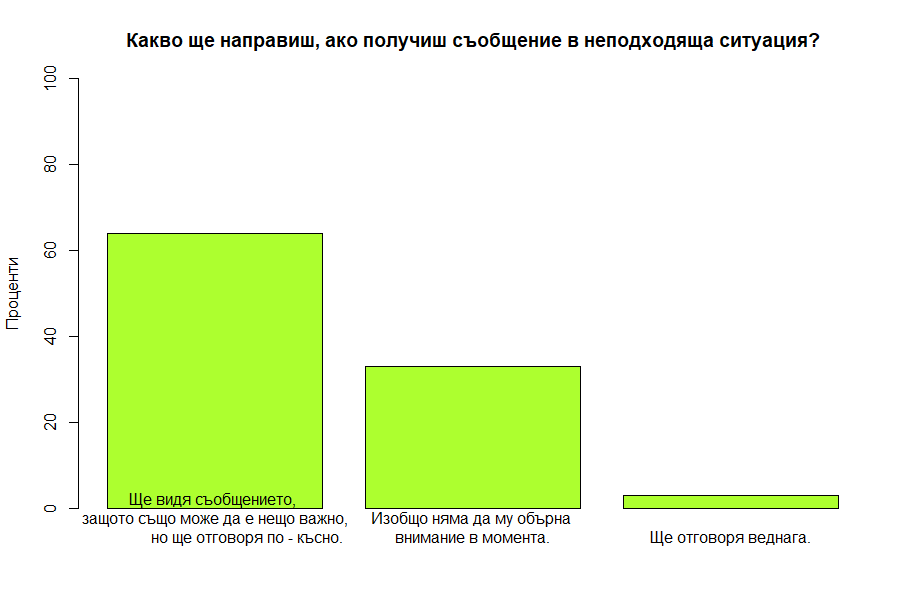
Графика:



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?

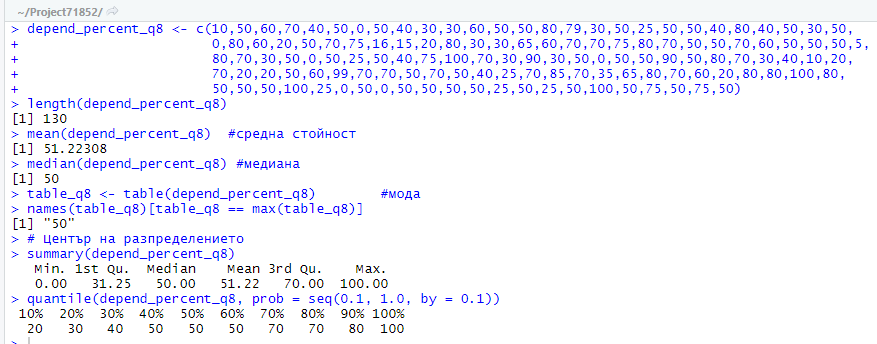
Цялта на въпрос 7 от анкетата е да постави хората в реална ситуация, в която трябва да изберат между това да отговорят на съобщение, получено на техните мобилни телефони по време на важна среща, или да го игнорират. Въвеждане да данните за този въпрос става с помощта на кода:

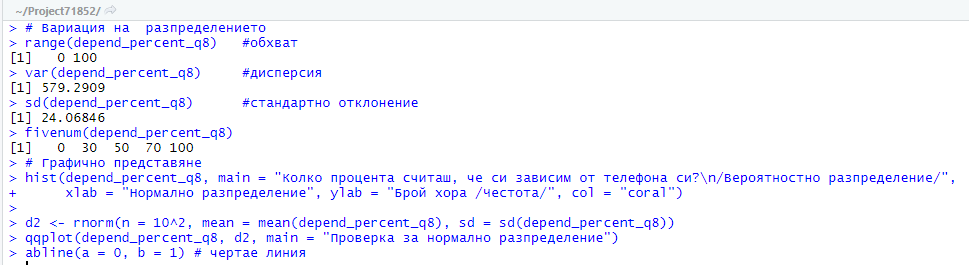


Графика:

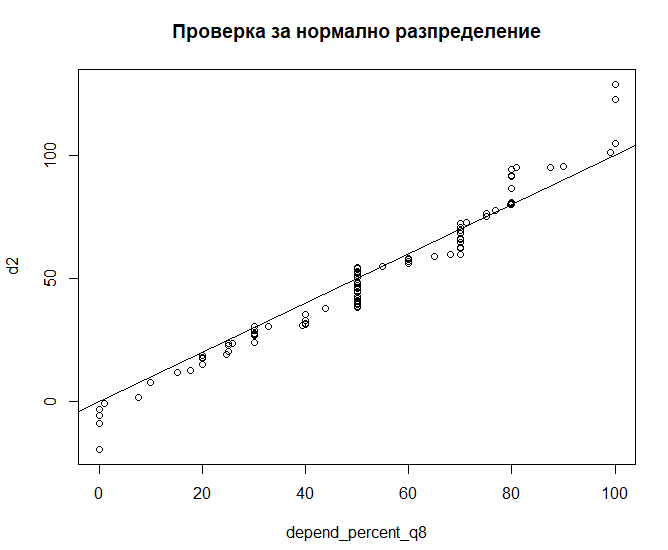
От графиката се вижда, че 64% от анкетираните първо биха завършили ангажиментите си и след това ще обърнат внимание на съобщението, което са получили. Едва около 3% ще отговорят веднага, а 33% дори не биха позволили едно съобщение да провали срещата им.

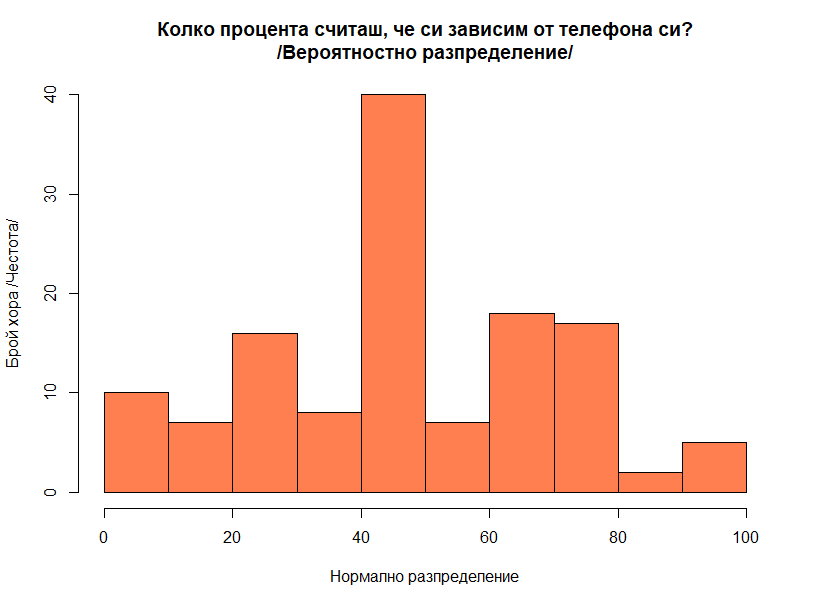
* Въпрос 8: В рамките на интервала 0 - 100 колко процента считаш, че си ,,зависим" от телефона си?

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:

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

Графики:

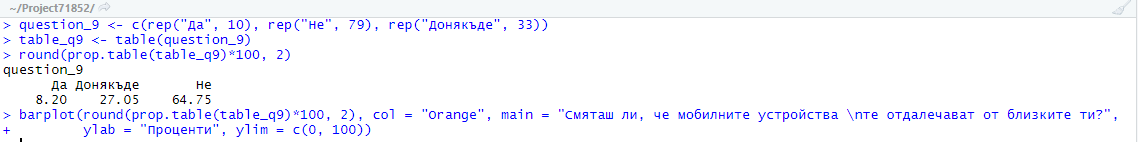




Най – често срещаната стойност е 50%, т.е. половината от анкетиранитаните се смятат за зависими от телефоните си.

* 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:

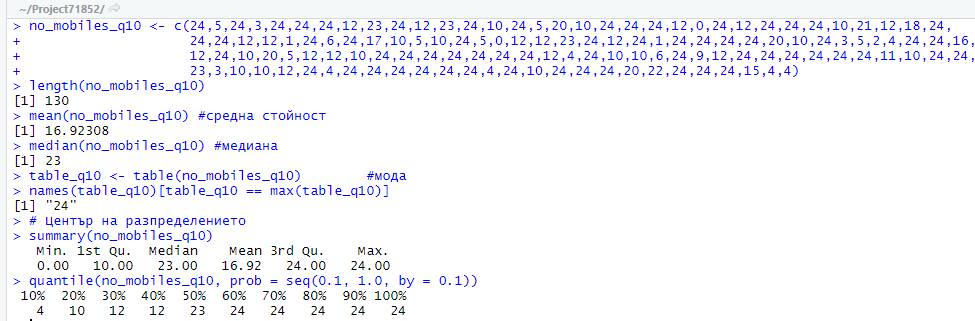


Графика:

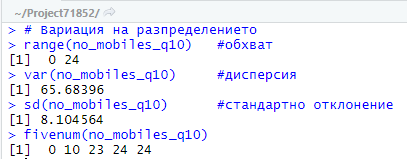
Очевидно, повечето от тях смятат, че мобилните телефони не са проблем, който може да застраши отношенията им с останалите – около 65%. Едва 8% пък са на обратното мнение. А 27% смятат, че вероятно това може да се окаже проблем.

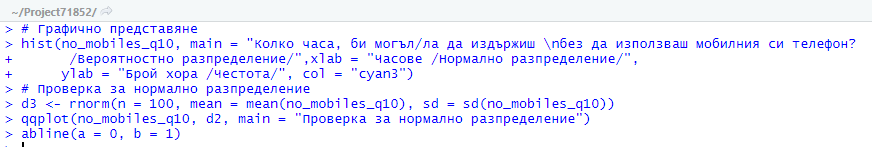
* Въпрос 10: Колко време - 0 - 24 часа, би могъл/ла да издържиш без да използваш мобилния си телефон?

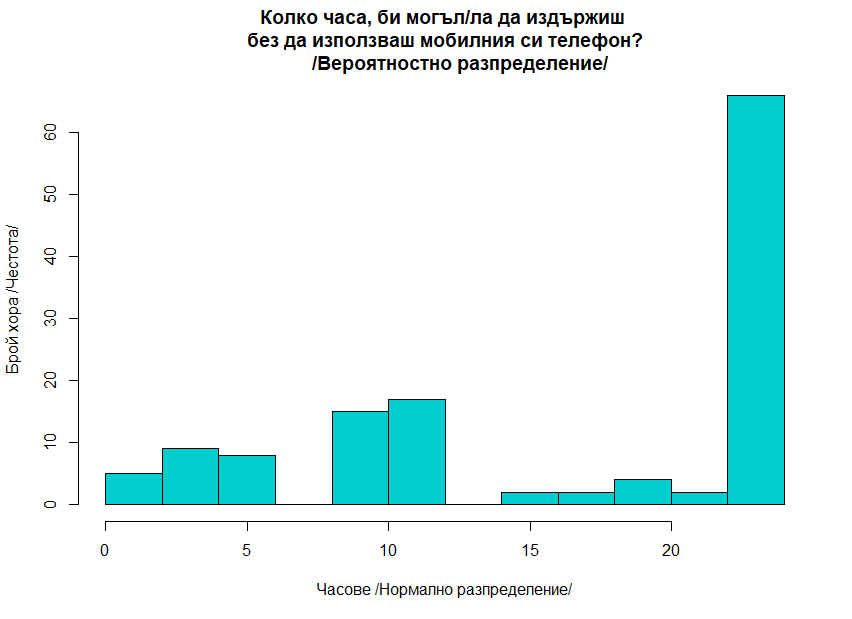
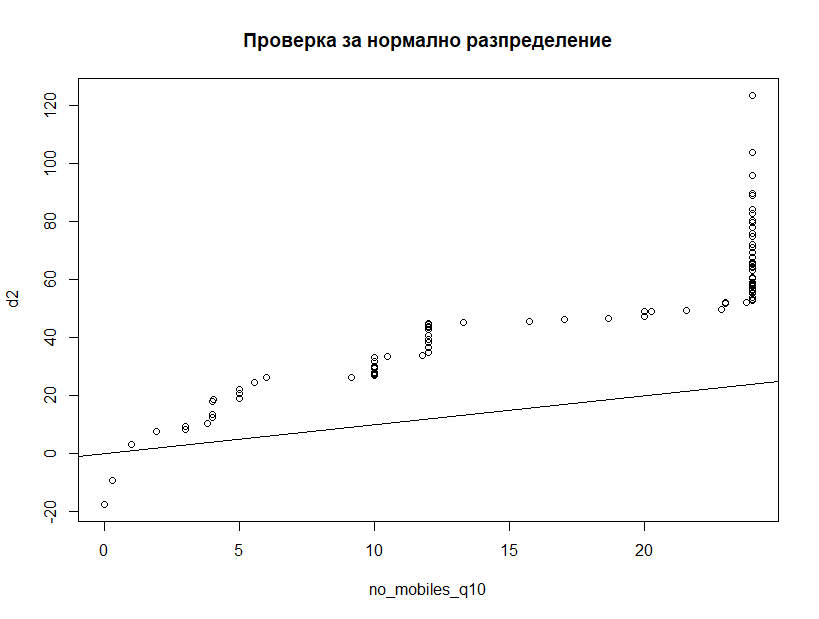
Този въпрос също е непрекъсната величина, тъй като хората имат право да въвеждат произволи числа в интервала 0 – 24 часа.

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

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

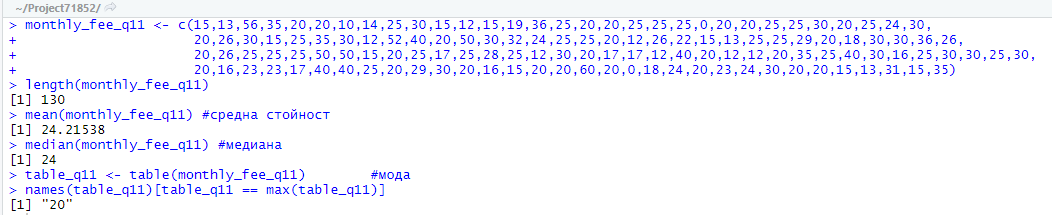


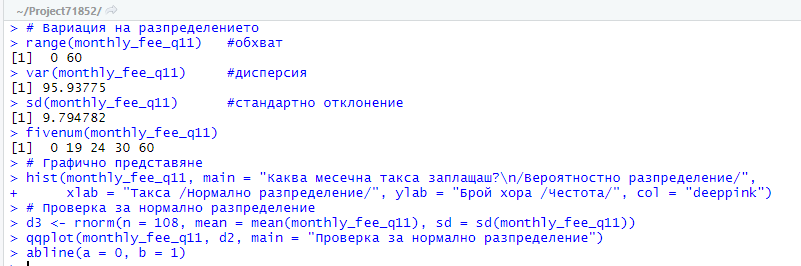


Графики:

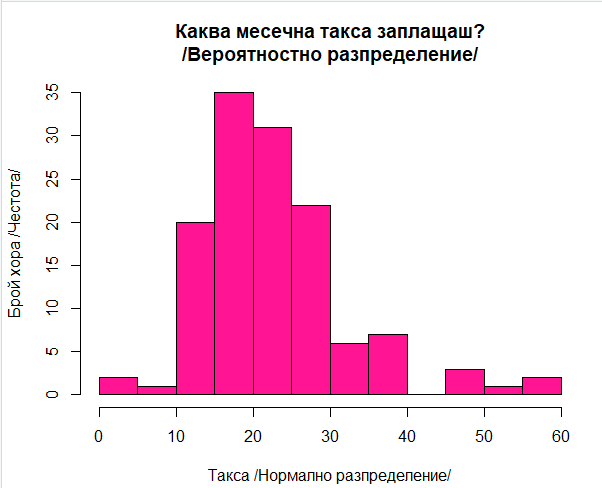
В заключение, можем да кажем, че все пак има хора, които не са зависими от мобилните си телефони, т.е. могат да издържат цял ден без тях. В случая те са над 60%. Това може да се определи и от факта, че най – често срещаната стойност, която те посочват, е 24 часа.

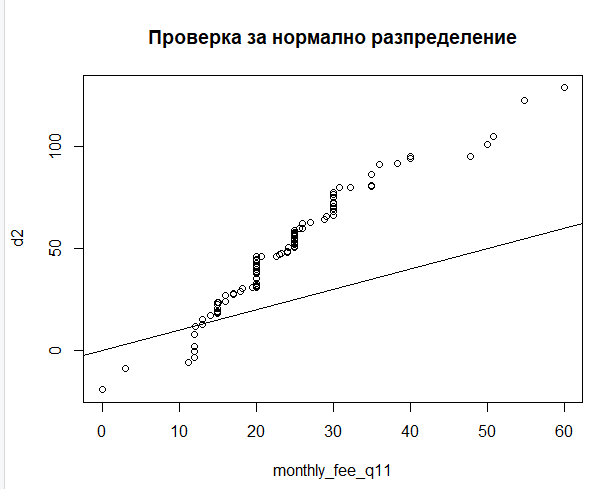
* Въпрос 11: Каква месечна такса заплащаш, за да използваш услугите на мобилиния си оператор ?

Този въпрост 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:



Графики:





Изводи:

Най – често срещаната стойност сред отговорите е 20 лв.

Средната стойност от всички отговори е около 24 лева.

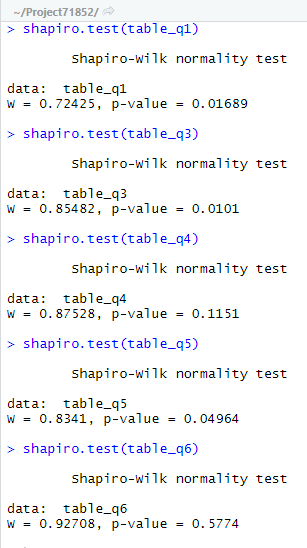
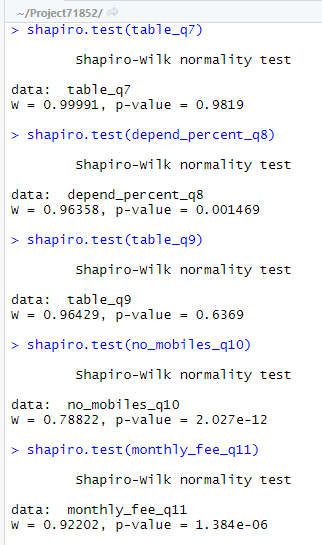
Около 30% от хората заплащат на месец между 20 - 25лв.

Между 20 – 25% заплащат такса 10 – 20 лв. и 25 – 30лв.

*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 отговоите са нормално разпределени.

Нека хипотезата H0 гласи, че отговорите са равномерно разпределени. Нивото на съгласие е alpha = 0,05.



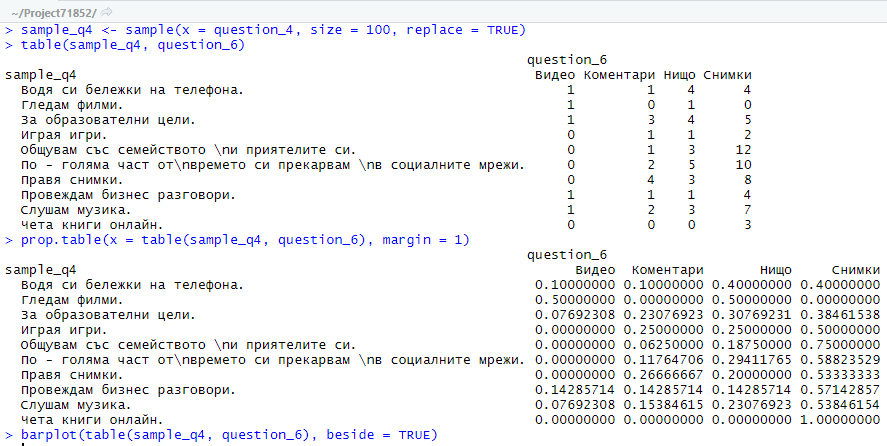
Забелязваме, че стойностите за p са различни. Тези, чиито p – value са по – големи от нивото на съгласие, считаме разпределенията им за нормални. За нормално разпределените променливи ще използваме параметрични тестове, а за останалите – непараметрични. Забелязваме, че стойността на p за Въпрос 4, Въпрос 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

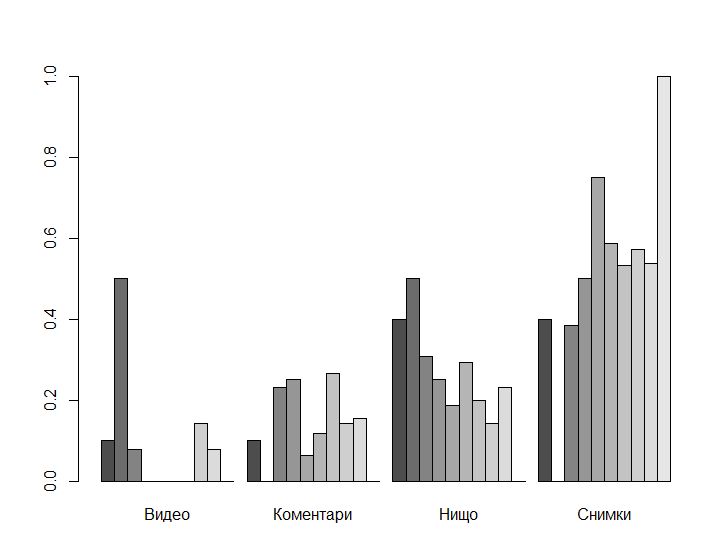
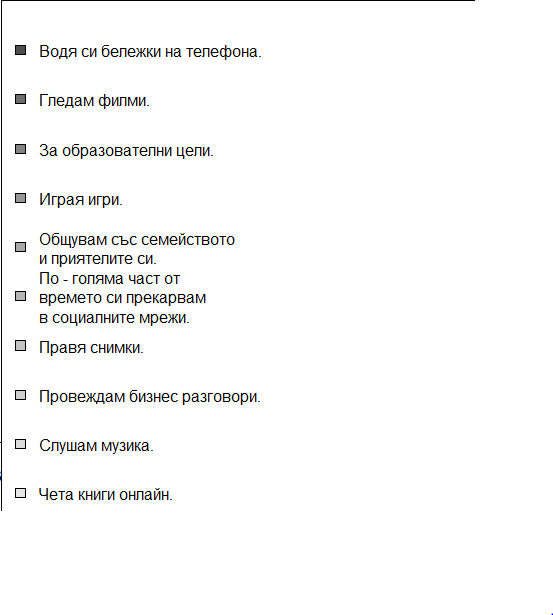
За изпълнението на тази част ще направя проверка за това да ли има зависимост между две променливи, като използвам различни хипотези. За графичното им изобразяване ще използвам различни техники.

**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.

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 barplot представям данните графично:

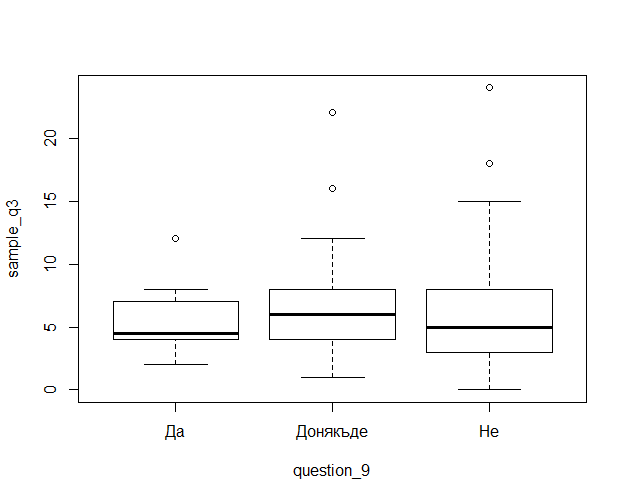




От графиката и от таблицата става ясно, че хората, които използват мобилните си телефони най – често, за да общуват със семейството и приятелите си, също ги използват и за споделяне на снимки в социалните мрежи – 0,75%. Тези пък, които ги използват за да си водят бележки, споделят и видео – 0,8%, и коментари – 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.



Удебелената черна линия във всеяка 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.

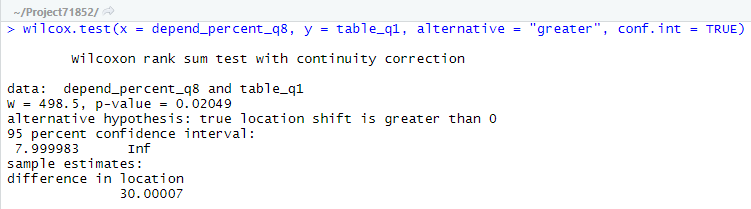
**3. Числова vs Категорийна**

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 H0 е следната: Зависимостта на хората от телефоните има общо с възрастта им.

Алтернативната хипотеза H1 е: Нямат нищо общо.

Нивото на съгласие е alpha: 0,05.



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**

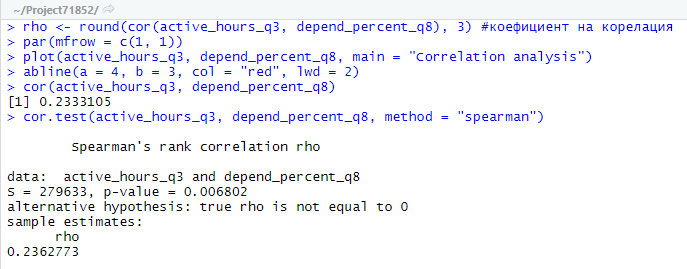
Първо, ще направя един пример с корелационен анализ. След това още един пример с регресионен анализ.

*4.1. Корелационен анализ*

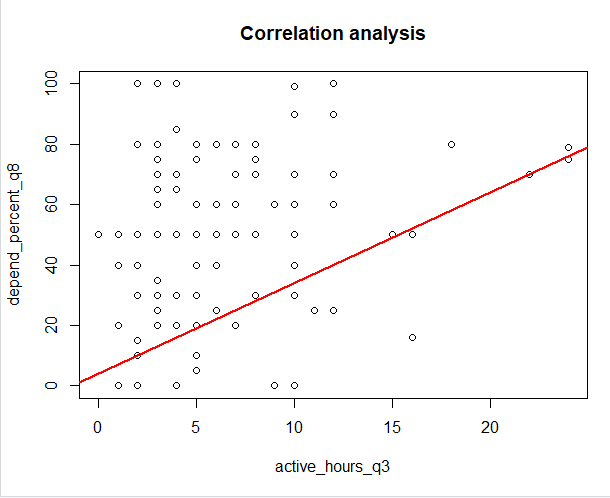
Въпрос 3: По колко часа на ден в интервала 0 - 24 използваш телефона си?

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.



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



rho 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.

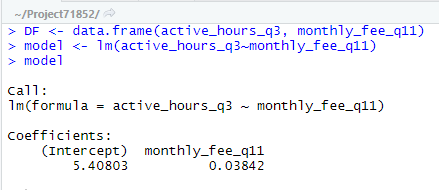
*4.2. Регресионен анализ*

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

Въпрос 11: Каква месечна такса заплащаш, за да използваш услугите на мобилиния си оператор ?

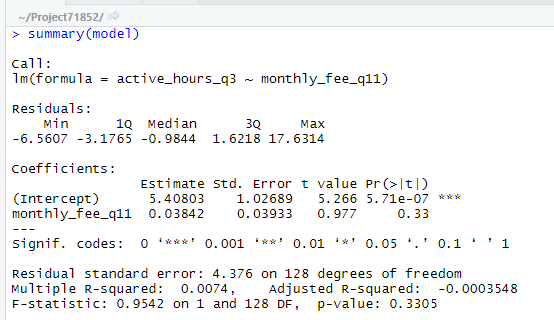
Ще изследвам двата въпроса с помощта на линейна регресия, за да видя дали има връзка между тях. Нулевата хипотеза е: Хората, които използват повече мобилните си телефони, плащат по – големи такси. Алтернативната хипотеза - H1 - е, че това не е така.

H0 се отхвърля при стойност на p-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:

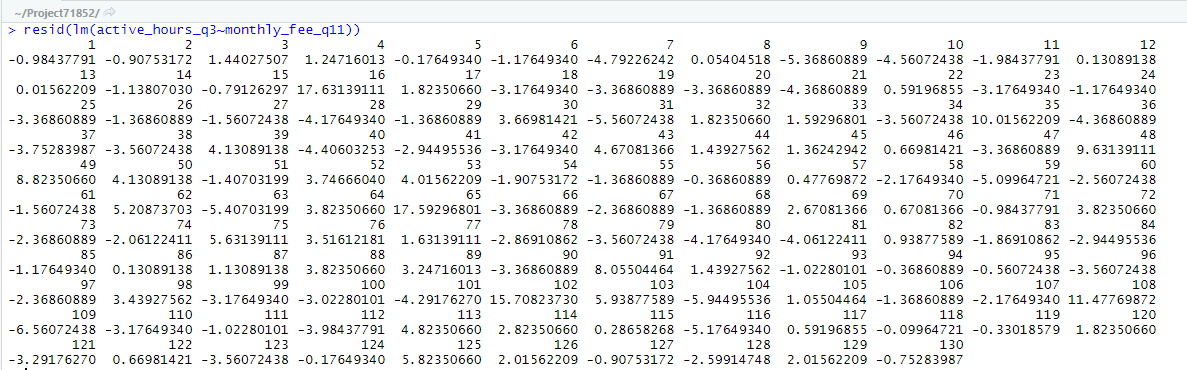
Коефициентът пред monthly\_fee\_q11 е 0.03842:

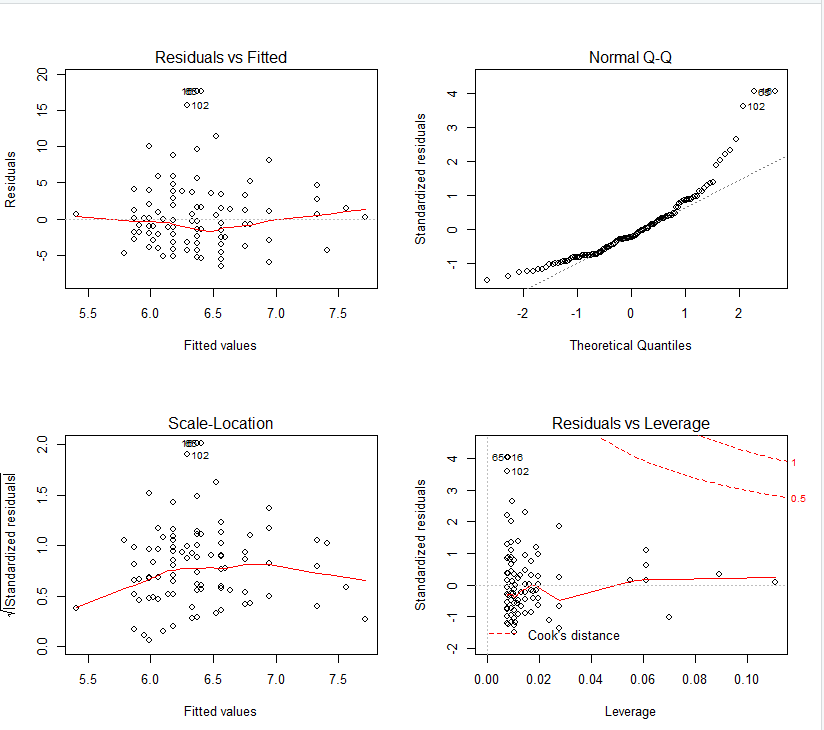
active\_hours\_q3 = 0.03842\*monthly\_fee\_q11 + 5.40803

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

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 << 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

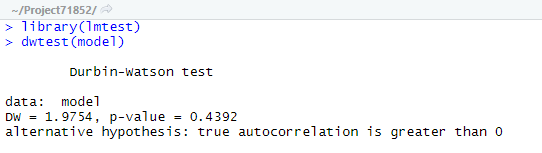


*Residuals vs Fitted* – Графика на разсейване на остатъците по оста y и прогнозните отговори по оста x. Използва се за откриване на нелинейност, неравномерни отклонения и външни разходи. Вижда се, че остатъците са разпръснати на случаен принцип около остатъчния ред = 0. Това предполага, че връзката е 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 старадартизираните 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 изоплзва 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.

**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 прекарават 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 заплащат 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*