Bedroom

# Анализ на данните

* App - Качествени
* Category - Качествени
* Rating - Количествени и Дискретни
* Reviews - Количествени и Дискретни
* Size - Качествени и Непрекъснати
* Installs - Количествени и Дискретни
* Type - Качествени
* Price - Количествени и Непрекъснати
* Content.Rating - Качествени
* Genres - Качествени
* Last.Updated - Качествени
* Current.ver - Количествени
* Android.ver - Качествени

# Анализ на колоните от дата фрейма apps

### Type - 92.60% - Free, 7.38 - Paid  
cbind(round(prop.table(table(apps$Type))\*100, 2))

## [,1]  
## Free 92.51  
## Paid 7.49

### Category  
cbind(round(prop.table(table(apps$Category))\*100, 2))

## [,1]  
## ART\_AND\_DESIGN 0.76  
## AUTO\_AND\_VEHICLES 0.82  
## BEAUTY 0.48  
## BOOKS\_AND\_REFERENCE 1.86  
## BUSINESS 3.18  
## COMICS 0.63  
## COMMUNICATION 2.73  
## DATING 2.24  
## EDUCATION 1.42  
## ENTERTAINMENT 1.16  
## EVENTS 0.49  
## FAMILY 20.92  
## FINANCE 3.44  
## FOOD\_AND\_DRINK 1.09  
## GAME 12.60  
## HEALTH\_AND\_FITNESS 2.89  
## HOUSE\_AND\_HOME 0.72  
## LIBRARIES\_AND\_DEMO 0.80  
## LIFESTYLE 3.62  
## MAPS\_AND\_NAVIGATION 1.23  
## MEDICAL 4.19  
## NEWS\_AND\_MAGAZINES 2.19  
## PARENTING 0.57  
## PERSONALIZATION 3.62  
## PHOTOGRAPHY 3.05  
## PRODUCTIVITY 3.04  
## SHOPPING 2.32  
## SOCIAL 2.29  
## SPORTS 3.20  
## TOOLS 8.20  
## TRAVEL\_AND\_LOCAL 2.07  
## VIDEO\_PLAYERS 1.50  
## WEATHER 0.66

### Genres  
cbind(round(prop.table(table(apps$Genres))\*100, 2))

## [,1]  
## Action 4.17  
## Action;Action & Adventure 0.21  
## Adventure 0.88  
## Adventure;Action & Adventure 0.17  
## Adventure;Brain Games 0.01  
## Adventure;Education 0.03  
## Arcade 2.41  
## Arcade;Action & Adventure 0.17  
## Arcade;Pretend Play 0.01  
## Art & Design 0.69  
## Art & Design;Creativity 0.08  
## Art & Design;Pretend Play 0.03  
## Auto & Vehicles 0.82  
## Beauty 0.48  
## Board 0.49  
## Board;Action & Adventure 0.04  
## Board;Brain Games 0.18  
## Board;Pretend Play 0.01  
## Books & Reference 1.86  
## Books & Reference;Education 0.03  
## Business 3.18  
## Card 0.50  
## Card;Action & Adventure 0.03  
## Card;Brain Games 0.01  
## Casino 0.41  
## Casual 2.07  
## Casual;Action & Adventure 0.19  
## Casual;Brain Games 0.17  
## Casual;Creativity 0.09  
## Casual;Education 0.04  
## Casual;Music & Video 0.03  
## Casual;Pretend Play 0.39  
## Comics 0.62  
## Comics;Creativity 0.01  
## Communication 2.73  
## Dating 2.24  
## Education 5.40  
## Education;Action & Adventure 0.08  
## Education;Brain Games 0.01  
## Education;Creativity 0.06  
## Education;Education 0.53  
## Education;Music & Video 0.04  
## Education;Pretend Play 0.28  
## Educational 0.40  
## Educational;Action & Adventure 0.05  
## Educational;Brain Games 0.08  
## Educational;Creativity 0.06  
## Educational;Education 0.43  
## Educational;Pretend Play 0.23  
## Entertainment 5.80  
## Entertainment;Action & Adventure 0.04  
## Entertainment;Brain Games 0.08  
## Entertainment;Creativity 0.03  
## Entertainment;Education 0.01  
## Entertainment;Music & Video 0.21  
## Entertainment;Pretend Play 0.03  
## Events 0.49  
## Finance 3.44  
## Food & Drink 1.09  
## Health & Fitness 2.89  
## Health & Fitness;Action & Adventure 0.01  
## Health & Fitness;Education 0.01  
## House & Home 0.72  
## Libraries & Demo 0.80  
## Lifestyle 3.61  
## Lifestyle;Pretend Play 0.01  
## Maps & Navigation 1.23  
## Medical 4.19  
## Music 0.19  
## Music & Audio;Music & Video 0.01  
## Music;Music & Video 0.04  
## News & Magazines 2.19  
## Parenting 0.45  
## Parenting;Brain Games 0.01  
## Parenting;Education 0.04  
## Parenting;Music & Video 0.06  
## Personalization 3.62  
## Photography 3.05  
## Productivity 3.04  
## Puzzle 1.40  
## Puzzle;Action & Adventure 0.06  
## Puzzle;Brain Games 0.23  
## Puzzle;Creativity 0.03  
## Puzzle;Education 0.01  
## Racing 1.07  
## Racing;Action & Adventure 0.18  
## Racing;Pretend Play 0.01  
## Role Playing 1.33  
## Role Playing;Action & Adventure 0.06  
## Role Playing;Brain Games 0.01  
## Role Playing;Pretend Play 0.06  
## Shopping 2.32  
## Simulation 2.35  
## Simulation;Action & Adventure 0.14  
## Simulation;Education 0.04  
## Simulation;Pretend Play 0.04  
## Social 2.29  
## Sports 3.38  
## Sports;Action & Adventure 0.03  
## Strategy 1.24  
## Strategy;Action & Adventure 0.03  
## Strategy;Creativity 0.01  
## Strategy;Education 0.01  
## Tools 8.20  
## Travel & Local 2.06  
## Travel & Local;Action & Adventure 0.01  
## Trivia 0.35  
## Video Players & Editors 1.49  
## Video Players & Editors;Creativity 0.03  
## Video Players & Editors;Music & Video 0.01  
## Weather 0.66  
## Word 0.31

### Content.Rating - 80.38% - Everyone, 11.14% - Teen, ...  
cbind(round(prop.table(table(apps$Content.Rating))\*100, 2))

## [,1]  
## Adults only 18+ 0.03  
## Everyone 79.86  
## Everyone 10+ 4.11  
## Mature 17+ 4.76  
## Teen 11.23  
## Unrated 0.01

### Rating  
cbind(round(prop.table(table(apps$Rating))\*100, 2))

## [,1]  
## 1 0.21  
## 1.2 0.01  
## 1.4 0.04  
## 1.5 0.04  
## 1.6 0.05  
## 1.7 0.10  
## 1.8 0.10  
## 1.9 0.16  
## 2 0.14  
## 2.1 0.10  
## 2.2 0.18  
## 2.3 0.26  
## 2.4 0.25  
## 2.5 0.25  
## 2.6 0.31  
## 2.7 0.30  
## 2.8 0.54  
## 2.9 0.56  
## 3 0.96  
## 3.1 0.80  
## 3.2 0.78  
## 3.3 1.27  
## 3.4 1.49  
## 3.5 1.98  
## 3.6 2.04  
## 3.7 2.56  
## 3.8 3.52  
## 3.9 4.31  
## 4 6.22  
## 4.1 7.69  
## 4.2 9.57  
## 4.3 10.58  
## 4.4 10.80  
## 4.5 10.64  
## 4.6 8.36  
## 4.7 5.45  
## 4.8 2.79  
## 4.9 1.11  
## 5 3.48

### Installs  
cbind(round(prop.table(table(apps$Installs))\*100, 2))

## [,1]  
## 1 0.04  
## 5 0.12  
## 10 0.87  
## 50 0.72  
## 100 3.92  
## 500 2.55  
## 1000 8.93  
## 5000 5.43  
## 10000 12.54  
## 50000 5.65  
## 100000 13.42  
## 500000 6.35  
## 1000000 16.85  
## 5000000 6.92  
## 10000000 10.67  
## 50000000 1.90  
## 100000000 2.60  
## 500000000 0.39  
## 1000000000 0.13

### Android.version  
cbind(round(prop.table(table(apps$Android.Ver))\*100, 2))

## [,1]  
## 1.0 and up 0.03  
## 1.5 and up 0.19  
## 1.6 and up 1.13  
## 2.0 and up 0.35  
## 2.0.1 and up 0.09  
## 2.1 and up 1.46  
## 2.2 and up 2.67  
## 2.3 and up 7.32  
## 2.3.3 and up 3.04  
## 3.0 and up 2.73  
## 3.1 and up 0.10  
## 3.2 and up 0.40  
## 4.0 and up 14.35  
## 4.0.3 - 7.1.1 0.03  
## 4.0.3 and up 15.45  
## 4.1 - 7.1.1 0.01  
## 4.1 and up 24.98  
## 4.2 and up 4.12  
## 4.3 and up 2.52  
## 4.4 and up 10.43  
## 4.4W and up 0.08  
## 5.0 - 6.0 0.01  
## 5.0 - 8.0 0.03  
## 5.0 and up 6.34  
## 5.1 and up 0.22  
## 6.0 and up 0.58  
## 7.0 - 7.1.1 0.01  
## 7.0 and up 0.50  
## 7.1 and up 0.03  
## 8.0 and up 0.06  
## Varies with device 0.72

## Summaries

summary(apps$Rating)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 4.000 4.300 4.174 4.500 5.000

summary(apps$Installs)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 10000 100000 8417734 1000000 1000000000

summary(apps$Type)

## Length Class Mode   
## 7729 character character

summary(apps$Content.Rating)

## Length Class Mode   
## 7729 character character

summary(apps$Last.Updated)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2010 2017 2018 2017 2018 2018

### Първите 10 наблюдения

head(apps,10)

## App Category  
## 1 Photo Editor & Candy Camera & Grid & ScrapBook ART\_AND\_DESIGN  
## 2 Coloring book moana ART\_AND\_DESIGN  
## 3 U Launcher Lite â\200“ FREE Live Cool Themes, Hide Apps ART\_AND\_DESIGN  
## 4 Sketch - Draw & Paint ART\_AND\_DESIGN  
## 5 Pixel Draw - Number Art Coloring Book ART\_AND\_DESIGN  
## 6 Paper flowers instructions ART\_AND\_DESIGN  
## 7 Smoke Effect Photo Maker - Smoke Editor ART\_AND\_DESIGN  
## 8 Infinite Painter ART\_AND\_DESIGN  
## 9 Garden Coloring Book ART\_AND\_DESIGN  
## 10 Kids Paint Free - Drawing Fun ART\_AND\_DESIGN  
## Rating Reviews Size Installs Type Price Content.Rating  
## 1 4.1 159 19.0 10000 Free 0 Everyone  
## 2 3.9 967 14.0 500000 Free 0 Everyone  
## 3 4.7 87510 8.7 5000000 Free 0 Everyone  
## 4 4.5 215644 25.0 50000000 Free 0 Teen  
## 5 4.3 967 2.8 100000 Free 0 Everyone  
## 6 4.4 167 5.6 50000 Free 0 Everyone  
## 7 3.8 178 19.0 50000 Free 0 Everyone  
## 8 4.1 36815 29.0 1000000 Free 0 Everyone  
## 9 4.4 13791 33.0 1000000 Free 0 Everyone  
## 10 4.7 121 3.1 10000 Free 0 Everyone  
## Genres Last.Updated Current.Ver Android.Ver  
## 1 Art & Design 2018 1.0.0 4.0.3 and up  
## 2 Art & Design;Pretend Play 2018 2.0.0 4.0.3 and up  
## 3 Art & Design 2018 1.2.4 4.0.3 and up  
## 4 Art & Design 2018 Varies with device 4.2 and up  
## 5 Art & Design;Creativity 2018 1.1 4.4 and up  
## 6 Art & Design 2017 1.0 2.3 and up  
## 7 Art & Design 2018 1.1 4.0.3 and up  
## 8 Art & Design 2018 6.1.61.1 4.2 and up  
## 9 Art & Design 2017 2.9.2 3.0 and up  
## 10 Art & Design;Creativity 2018 2.8 4.0.3 and up

## Ще разгледаме типът на приложенията

apps <- apps[is.na(apps$Type)==FALSE, ]  
table(apps$Type)

##   
## Free Paid   
## 7150 579

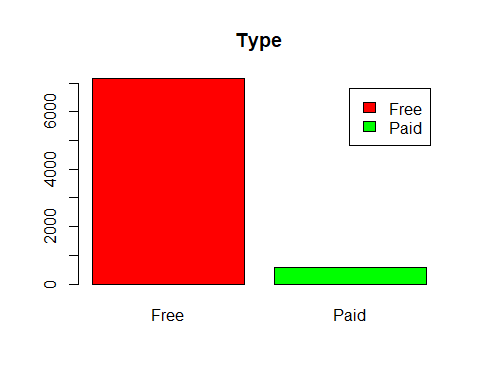
table(apps$Type)["Free"]

## Free   
## 7150

table(apps$Type)["Paid"]

## Paid   
## 579

barplot(table(apps$Type),  
 beside=TRUE,  
 col = c("red","green"),  
 legend.text = TRUE,  
 main = "Type")



### Забелязжаме, че безплатните приложения са значително повече от платените

## Tаблици

table(apps$Category)

##   
## ART\_AND\_DESIGN AUTO\_AND\_VEHICLES BEAUTY   
## 59 63 37   
## BOOKS\_AND\_REFERENCE BUSINESS COMICS   
## 144 246 49   
## COMMUNICATION DATING EDUCATION   
## 211 173 110   
## ENTERTAINMENT EVENTS FAMILY   
## 90 38 1617   
## FINANCE FOOD\_AND\_DRINK GAME   
## 266 84 974   
## HEALTH\_AND\_FITNESS HOUSE\_AND\_HOME LIBRARIES\_AND\_DEMO   
## 223 56 62   
## LIFESTYLE MAPS\_AND\_NAVIGATION MEDICAL   
## 280 95 324   
## NEWS\_AND\_MAGAZINES PARENTING PERSONALIZATION   
## 169 44 280   
## PHOTOGRAPHY PRODUCTIVITY SHOPPING   
## 236 235 179   
## SOCIAL SPORTS TOOLS   
## 177 247 634   
## TRAVEL\_AND\_LOCAL VIDEO\_PLAYERS WEATHER   
## 160 116 51

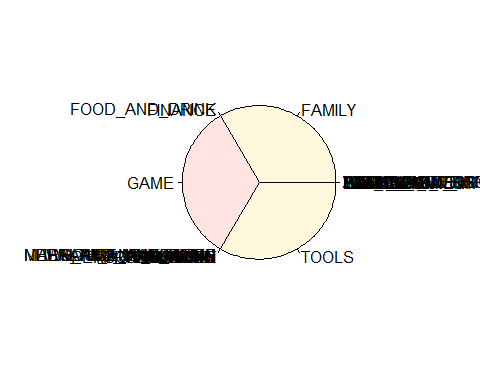
table(apps$Category)>500

##   
## ART\_AND\_DESIGN AUTO\_AND\_VEHICLES BEAUTY   
## FALSE FALSE FALSE   
## BOOKS\_AND\_REFERENCE BUSINESS COMICS   
## FALSE FALSE FALSE   
## COMMUNICATION DATING EDUCATION   
## FALSE FALSE FALSE   
## ENTERTAINMENT EVENTS FAMILY   
## FALSE FALSE TRUE   
## FINANCE FOOD\_AND\_DRINK GAME   
## FALSE FALSE TRUE   
## HEALTH\_AND\_FITNESS HOUSE\_AND\_HOME LIBRARIES\_AND\_DEMO   
## FALSE FALSE FALSE   
## LIFESTYLE MAPS\_AND\_NAVIGATION MEDICAL   
## FALSE FALSE FALSE   
## NEWS\_AND\_MAGAZINES PARENTING PERSONALIZATION   
## FALSE FALSE FALSE   
## PHOTOGRAPHY PRODUCTIVITY SHOPPING   
## FALSE FALSE FALSE   
## SOCIAL SPORTS TOOLS   
## FALSE FALSE TRUE   
## TRAVEL\_AND\_LOCAL VIDEO\_PLAYERS WEATHER   
## FALSE FALSE FALSE

prop.table(table(apps$Category))

##   
## ART\_AND\_DESIGN AUTO\_AND\_VEHICLES BEAUTY   
## 0.007633588 0.008151119 0.004787165   
## BOOKS\_AND\_REFERENCE BUSINESS COMICS   
## 0.018631130 0.031828180 0.006339759   
## COMMUNICATION DATING EDUCATION   
## 0.027299780 0.022383232 0.014232113   
## ENTERTAINMENT EVENTS FAMILY   
## 0.011644456 0.004916548 0.209212058   
## FINANCE FOOD\_AND\_DRINK GAME   
## 0.034415836 0.010868159 0.126018890   
## HEALTH\_AND\_FITNESS HOUSE\_AND\_HOME LIBRARIES\_AND\_DEMO   
## 0.028852374 0.007245439 0.008021736   
## LIFESTYLE MAPS\_AND\_NAVIGATION MEDICAL   
## 0.036227196 0.012291370 0.041920041   
## NEWS\_AND\_MAGAZINES PARENTING PERSONALIZATION   
## 0.021865701 0.005692845 0.036227196   
## PHOTOGRAPHY PRODUCTIVITY SHOPPING   
## 0.030534351 0.030404968 0.023159529   
## SOCIAL SPORTS TOOLS   
## 0.022900763 0.031957562 0.082028723   
## TRAVEL\_AND\_LOCAL VIDEO\_PLAYERS WEATHER   
## 0.020701255 0.015008410 0.006598525

pie(prop.table(table(apps$Category)>500))



### Забелязваме че има най-много приложения от категории Dating,Education,Health\_And\_Fitnes,Entertainment и Finance

sum(apps$Category=="ART\_AND\_DESIGN",as.numeric(apps$Reviews),na.rm=TRUE)

## [1] 2277524203

sum(apps$Category=="SPORTS",as.numeric(apps$Reviews),na.rm=TRUE)

## [1] 2277524391

sum(apps$Category=="GAME",as.numeric(apps$Reviews),na.rm=TRUE)

## [1] 2277525118

sum(apps$Category=="FAMILY",as.numeric(apps$Reviews),na.rm=TRUE)

## [1] 2277525761

sum(apps$Category=="TOOLS",as.numeric(apps$Reviews),na.rm=TRUE)

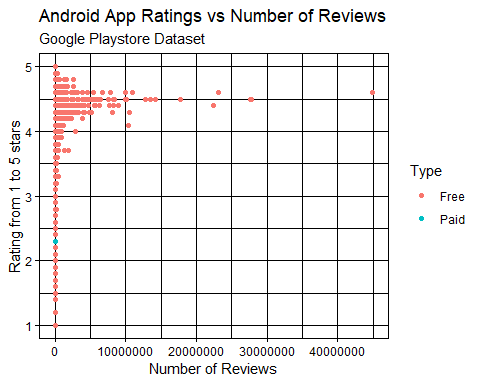
## [1] 2277524778

sum(apps$Category=="FINANCE",as.numeric(apps$Reviews),na.rm=TRUE)

## [1] 2277524410

## Ще анализираме връзките между Rating, Reviews и типа на приложението

ggplot(apps, aes(x=Reviews, y=Rating)) +  
 geom\_point(aes(col=Type)) +  
 labs(title="Android App Ratings vs Number of Reviews", subtitle="Google Playstore Dataset", y="Rating from 1 to 5 stars", x="Number of Reviews") +  
 theme\_linedraw()



cor(apps$Rating,as.numeric(apps$Reviews))

## [1] 0.07982279

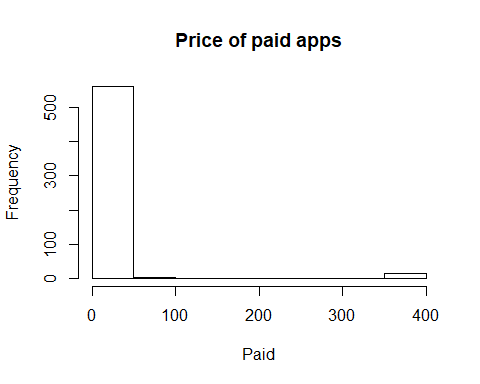
### Забелязваме, че повечето безплатни приложения имат рейтинг между 4 и 5,както и че има ясно изразена връзка между Reviews и Rating

## Ще разгледаме само платените тъй като базплатните са значително повече и не можем да получим нова информация от тях

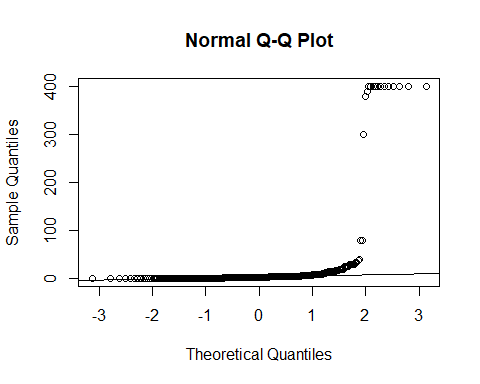
Paid<-apps$Price[apps$Price>0]  
summary(Paid)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.990 1.865 2.990 15.052 4.990 400.000

hist(Paid,  
 main = "Price of paid apps")



qqnorm(Paid)  
qqline(Paid)



shapiro.test(head(Paid,5000))

##   
## Shapiro-Wilk normality test  
##   
## data: head(Paid, 5000)  
## W = 0.19914, p-value < 0.00000000000000022

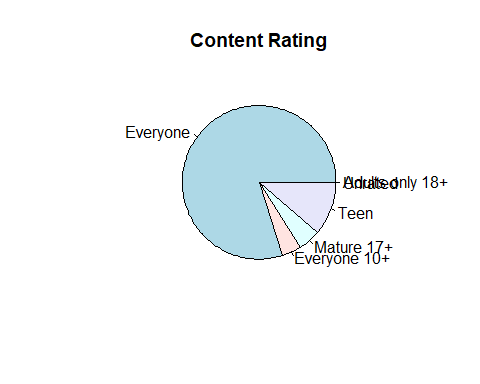
t.test(Paid)

##   
## One Sample t-test  
##   
## data: Paid  
## t = 5.846, df = 578, p-value = 0.000000008424  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 9.995227 20.109540  
## sample estimates:  
## mean of x   
## 15.05238

От хистограмата и шапиро теста забелязваме, че платените приложения са с нормално разпределение и, 95% от данните за цените на платените приложения имат доверителен интервал [4.17, 5.80]

## Piе и честотна таблица за Content.Rating

pie(table(apps$Content.Rating),  
 main="Content Rating")



table(apps$Content.Rating)

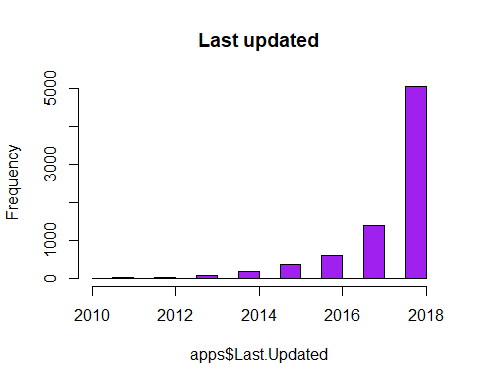
##   
## Adults only 18+ Everyone Everyone 10+ Mature 17+   
## 2 6172 318 368   
## Teen Unrated   
## 868 1

prop.table(table(apps$Content.Rating))

##   
## Adults only 18+ Everyone Everyone 10+ Mature 17+   
## 0.0002587657 0.7985509121 0.0411437443 0.0476128865   
## Teen Unrated   
## 0.1123043084 0.0001293828

### Oт графиката виждаме, че рейтинга на съдржанието на повечето приложения е подходящо за всички

hist(apps$Last.Updated,  
 main = "Last updated",  
 col = "purple")



### Забелязваме, че повечето приложения са последно обновени през 2018, както и че разпределението е експоненциално

wilcox.test(as.numeric(lastUpdated),conf.int =FALSE)

##   
## Wilcoxon signed rank test  
##   
## data: as.numeric(lastUpdated)  
## V = 1, p-value = 1  
## alternative hypothesis: true location is not equal to 0

**Най-високо оценено приложение за Dating**

datingApps=apps[apps$Category == "DATING", ]  
head(datingApps[datingApps$Rating == max(datingApps$Rating),],1)

## App Category Rating Reviews Size Installs  
## 613 American Girls Mobile Numbers DATING 5 5 4.4 1000  
## Type Price Content.Rating Genres Last.Updated Current.Ver Android.Ver  
## 613 Free 0 Mature 17+ Dating 2018 3.0 4.0.3 and up

**Най-ниско оценено приложение за Dating**  House party - live chat, DATING, 1, 1, 9.2, 10, Free, 0, Mature 17+, Dating, 2018, 3.52, 4.0.3 and up

**Търсим приложение, което го няма и приложение, което го има**

datingApps[grepl("Tinder", datingApps$App), ]

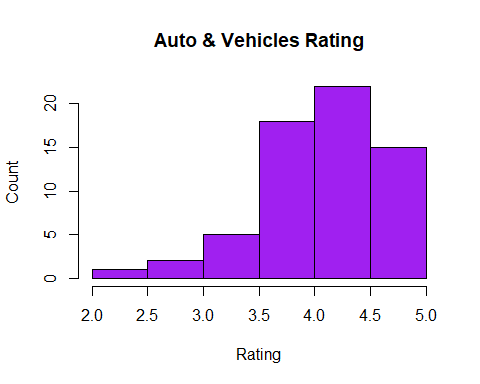
## [1] App Category Rating Reviews   
## [5] Size Installs Type Price   
## [9] Content.Rating Genres Last.Updated Current.Ver   
## [13] Android.Ver   
## <0 rows> (or 0-length row.names)

datingApps[grepl("Daddy", datingApps$App), ]

## App Category Rating Reviews  
## 528 Sugar Daddy Dating App DATING 2.5 277  
## 535 Gay Sugar Daddy Dating & Hookup â\200“ Sudy Gay DATING 4.1 2212  
## 587 Gay Sugar Daddy Dating & Hookup â\200“ Sudy Gay DATING 4.1 2212  
## Size Installs Type Price Content.Rating Genres Last.Updated  
## 528 5.7 100000 Free 0 Mature 17+ Dating 2017  
## 535 41.0 100000 Free 0 Mature 17+ Dating 2018  
## 587 41.0 100000 Free 0 Mature 17+ Dating 2018  
## Current.Ver Android.Ver  
## 528 3.0.0 4.1 and up  
## 535 2.0.5 4.0.3 and up  
## 587 2.0.5 4.0.3 and up

### Броят на игрите между 1 и 10 милиона изтегляния и извадка в графика на техния рейтинг

autoApps <-apps[apps$Category=="AUTO\_AND\_VEHICLES", ]  
autoApps = autoApps[autoApps$Installs >= 1000000 | autoApps$Installs <= 5000000,]  
hist(autoApps$Rating, main="Auto & Vehicles Rating", xlab="Rating", ylab="Count",col="purple")



shapiro.test(head(autoApps$Rating, 5000))

##   
## Shapiro-Wilk normality test  
##   
## data: head(autoApps$Rating, 5000)  
## W = 0.90936, p-value = 0.0002065

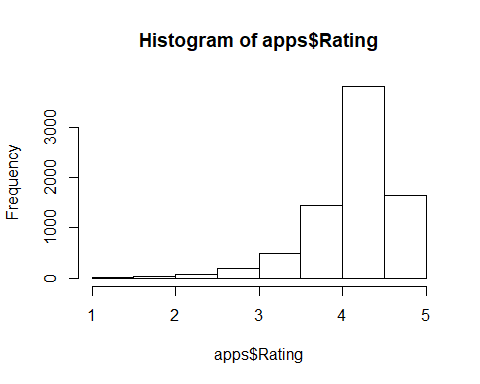
t.test(autoApps$Rating)

##   
## One Sample t-test  
##   
## data: autoApps$Rating  
## t = 58.23, df = 62, p-value < 0.00000000000000022  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 4.005235 4.290003  
## sample estimates:  
## mean of x   
## 4.147619

**Забелязваме, че рейтинга на приложенията от категория Auto & Vehicles имат нормално разпределение и са с доверителен интервал [4.13,4.38]**

## Ще разгледаме по-подбробно информацията за рейтинг на п всички риложенията

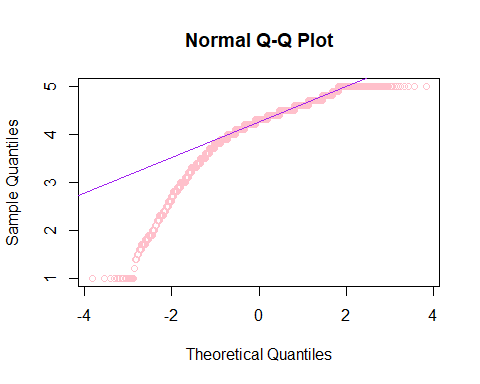
hist(apps$Rating)



shapiro.test(head(apps$Rating,5000))

##   
## Shapiro-Wilk normality test  
##   
## data: head(apps$Rating, 5000)  
## W = 0.86317, p-value < 0.00000000000000022

qqnorm(apps$Rating,col = 'pink')  
qqline(apps$Rating,col = 'purple')



wilcox.test(apps$Rating,conf.int = TRUE)

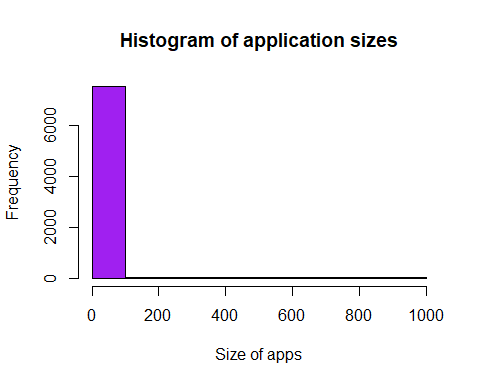
##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: apps$Rating  
## V = 29873000, p-value < 0.00000000000000022  
## alternative hypothesis: true location is not equal to 0  
## 95 percent confidence interval:  
## 4.249971 4.250014  
## sample estimates:  
## (pseudo)median   
## 4.249961

Забелязваме, че 95% от данните за рейтинга на приложенията имат доверител интервал [4.29,4.3]

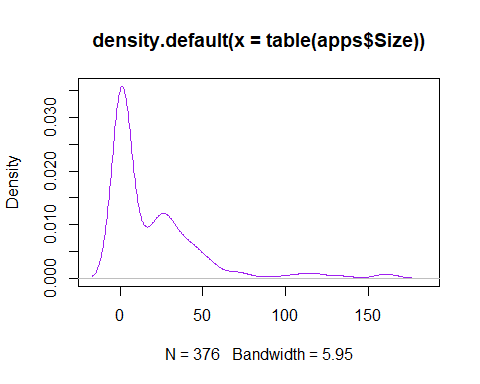
## Хистограма и графика на плътността на размерите на всички приложения

стойностите са конвертирани от string в numeric

hist(apps$Size, col="purple", xlab="Size of apps", main="Histogram of application sizes" )



plot(density(table(apps$Size)), col="purple")



shapiro.test(head(apps$Size,5000))

##   
## Shapiro-Wilk normality test  
##   
## data: head(apps$Size, 5000)  
## W = 0.30962, p-value < 0.00000000000000022

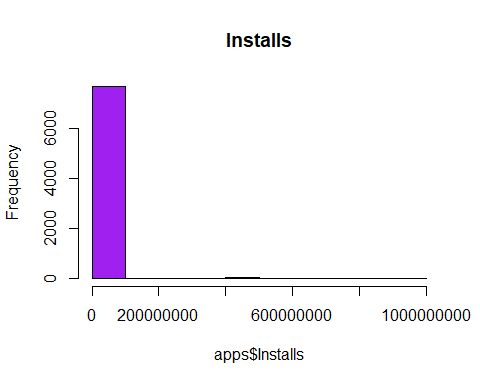
wilcox.test(apps$Size,conf.int = TRUE)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: apps$Size  
## V = 29873000, p-value < 0.00000000000000022  
## alternative hypothesis: true location is not equal to 0  
## 95 percent confidence interval:  
## 20.50001 21.85000  
## sample estimates:  
## (pseudo)median   
## 21.10006

След проверка забелязваме, че 95% от данните за размера на приложеништа попадат в интервала [17.5,19.9]

## Ще разгледаме данните за инсталации на приложения

hist(apps$Installs,  
 main = "Installs",  
 col='purple')



shapiro.test(head(apps$Installs,5000))

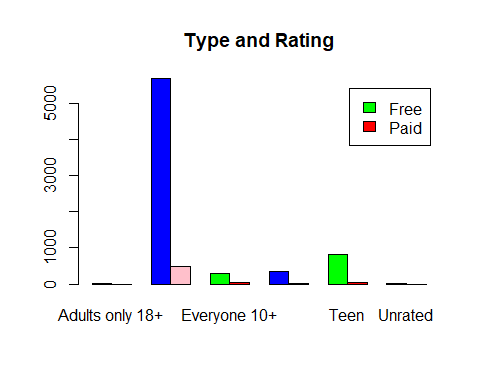
##   
## Shapiro-Wilk normality test  
##   
## data: head(apps$Installs, 5000)  
## W = 0.16953, p-value < 0.00000000000000022

wilcox.test(apps$Installs,conf.int = TRUE)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: apps$Installs  
## V = 29873000, p-value < 0.00000000000000022  
## alternative hypothesis: true location is not equal to 0  
## 95 percent confidence interval:  
## 525000 550000  
## sample estimates:  
## (pseudo)median   
## 525000

след тестовете забелязваме, че инсталациите са с екпоненциално разпределение и имат доверителен интервал [750000,1000000]

barplot(table(apps$Type,apps$Content.Rating),  
 legend.text = TRUE,  
 beside=TRUE,  
 col =c("green","red","blue","pink"),  
 main = "Type and Rating")

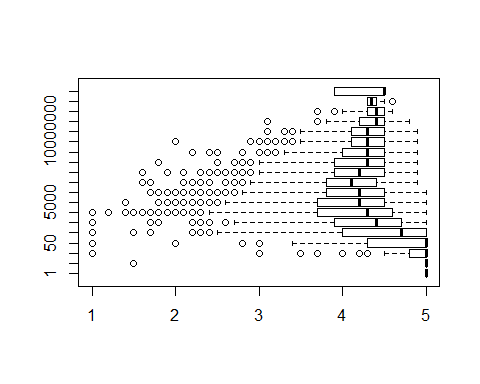


## Ще разгледаме връзката между инсталациите и рейтинга

cor(apps$Rating,apps$Installs)

## [1] 0.05270479

boxplot(apps$Rating~apps$Installs,  
 horizontal=TRUE)

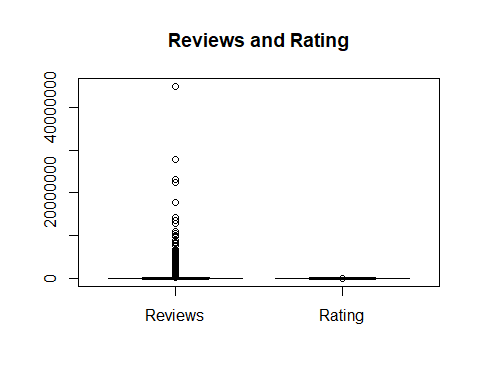


Набллудаваме, че приложенията с по-висок рейтинг имат повече инсталации

# —————————————->DA SE OPRAVI OTTUK

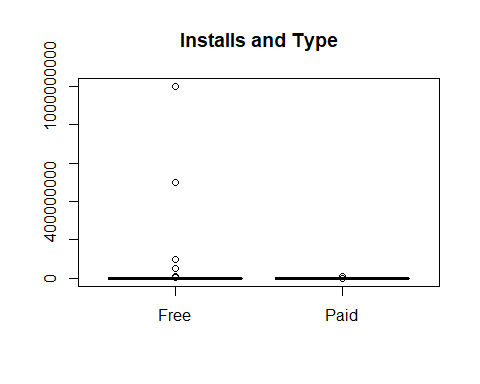
# сравнение на непрекъсната и дискретна

boxplot(as.numeric(apps$Reviews) ,apps$Rating,  
 beside= TRUE,  
 notch =TRUE,  
 names=c("Reviews","Rating"),  
 main="Reviews and Rating")



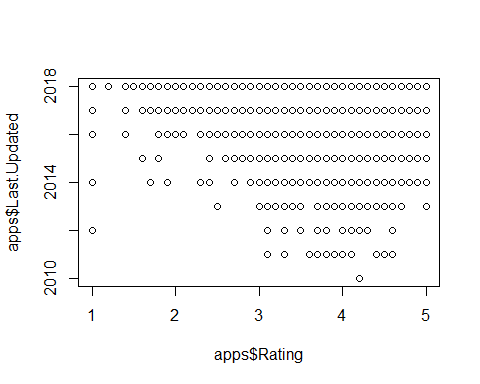
**Наблюдаваме, че медианите не се отдалечават рязко следователно данните са симетрични**

boxplot(apps$Installs~apps$Type,  
 main="Installs and Type")



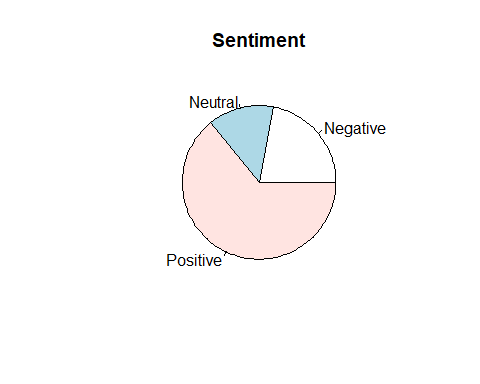
**Наблюдаваме, че безплатните апове имат повече изтегляния**

plot(apps$Last.Updated~apps$Rating)



Данните показват, че аповете с по-висок рейтинг са ъпдейтвани по-скоро

pie(table(appsReview$Sentiment),  
 main="Sentiment")

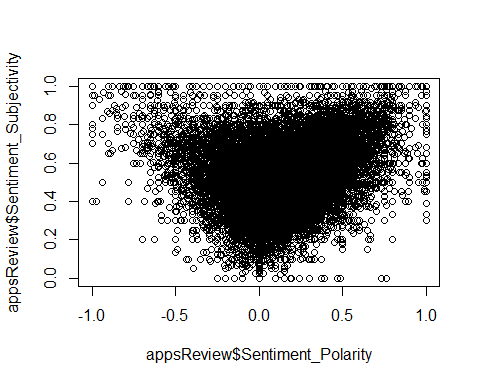


prop.table(table(appsReview$Sentiment))

##   
## Negative Neutral Positive   
## 0.2209607 0.1379301 0.6411092

Забелязваме, че повечето ревюта са с положителна

plot(appsReview$Sentiment\_Polarity,appsReview$Sentiment\_Subjectivity)

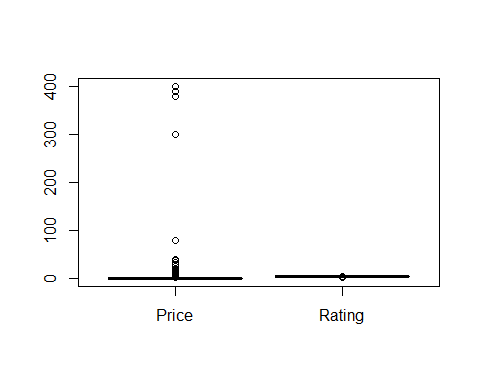


cor(appsReview$Sentiment\_Polarity,appsReview$Sentiment\_Subjectivity)

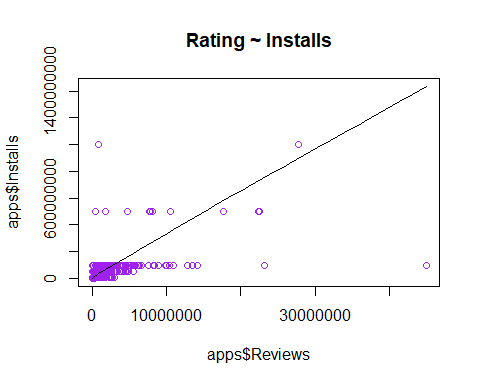
## [1] 0.2616511

Наблюдаваме, че има силно изразена корелация между субективните и полярните ревюта

boxplot(apps$Price,apps$Rating, name="PRICE & RATING", names = c("Price", "Rating"))

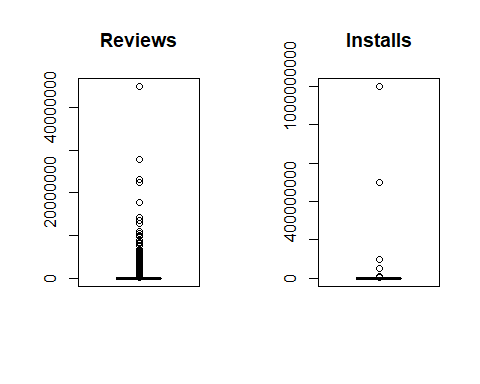
 Наблюдаваме, че аповете с по-ниски цени имат по-висок рейтинг

# Ще изследваме връзката между reviews и броят изтегляния на приложенията



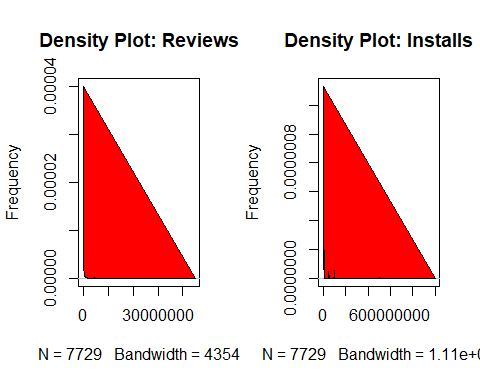
Графиката показва че по-високият брой ревюта на приложенията има общо с по-високия брой изтегляния Правим боксплот за да проверим за outliers

par(mfrow=c(1, 2)) # divide graph area in 2 columns  
boxplot(as.numeric(apps$Reviews), main="Reviews") # box plot for 'reviews'  
boxplot(apps$Installs, main="Installs") # box plot for 'installs'



**density plot**

par(mfrow=c(1, 2)) # divide graph area in 2 columns  
plot(density(as.numeric(apps$Reviews)), main="Density Plot: Reviews", ylab="Frequency") # density plot for 'reviews'  
polygon(density(as.numeric(apps$Reviews)), col="red")  
plot(density(apps$Installs,na.rm=TRUE), main="Density Plot: Installs", ylab="Frequency") # density plot for 'installs'  
polygon(density(apps$Installs,na.rm=TRUE), col="red")



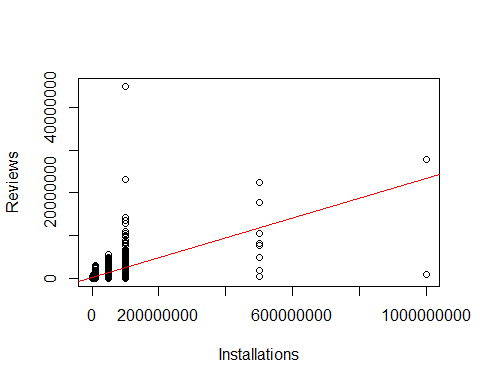
Корелация

apps <- apps[is.na(apps$Installs)==FALSE,]  
cor(as.numeric(apps$Reviews), apps$Installs)

## [1] 0.6261946

Стойността на корелацията върви към 1, което означава че има връзка между ревютата и броя изтегляния в права пропорционалност linear model

plot(apps$Installs, as.numeric(apps$Reviews), xlab="Installations", ylab = "Reviews")  
linearMod <- lm(Reviews ~ Installs, data=apps) # build linear regression model on full data  
abline(linearMod, col="red") #линия означаваща регресията



print(linearMod)

##   
## Call:  
## lm(formula = Reviews ~ Installs, data = apps)  
##   
## Coefficients:  
## (Intercept) Installs   
## 98788.39000 0.02327

Зависимостта е Installs = Intercept + (β ∗ Reviews) Installs = 98946.3399 + (0.2327 \* Reviews) Сега имаме линеен модел и направихме формула която да предсказва стойността на Installs ако имаме стойност на Reviews

summary(linearMod)

##   
## Call:  
## lm(formula = Reviews ~ Installs, data = apps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22491569 -100977 -98825 -98129 42468058   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 98788.3900028 16756.2986991 5.896 0.00000000389 \*\*\*  
## Installs 0.0232704 0.0003296 70.600 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1453000 on 7727 degrees of freedom  
## Multiple R-squared: 0.3921, Adjusted R-squared: 0.392   
## F-statistic: 4984 on 1 and 7727 DF, p-value: < 0.00000000000000022

ДО ТУК

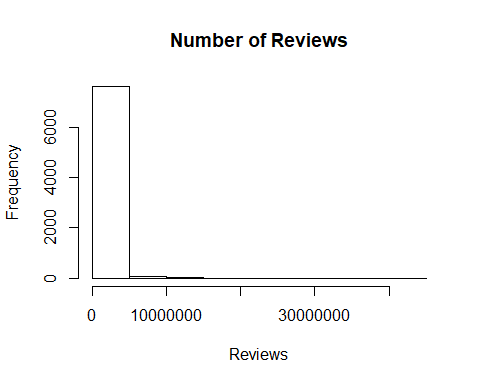
Size <- apps$Size  
Reviews <- apps$Reviews  
Rating <- apps$Rating[apps$Rating > 0]  
Installs <- apps$Installs[apps$Installs > 0]

# Ще разгледаме само платените тъй като базплатните са значително повече и не можем да получим нова информация от тях

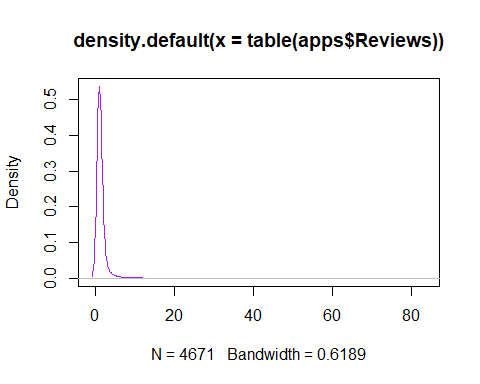
summary(Reviews)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 108 2328 294673 38961 44893888

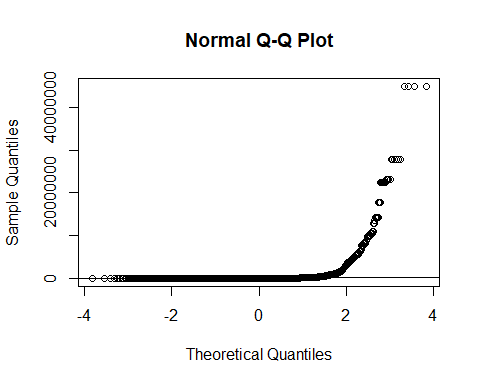
hist(Reviews,  
 main = "Number of Reviews")



plot(density(table(apps$Reviews)), col="purple")



qqnorm(Reviews)  
qqline(Reviews)



shapiro.test(head(Reviews, 5000))

##   
## Shapiro-Wilk normality test  
##   
## data: head(Reviews, 5000)  
## W = 0.16645, p-value < 0.00000000000000022

t.test(Reviews)

##   
## One Sample t-test  
##   
## data: Reviews  
## t = 13.904, df = 7728, p-value < 0.00000000000000022  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 253127.4 336217.7  
## sample estimates:  
## mean of x   
## 294672.6

wilcox.test(Reviews, conf.int = TRUE)

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: Reviews  
## V = 29873000, p-value < 0.00000000000000022  
## alternative hypothesis: true location is not equal to 0  
## 95 percent confidence interval:  
## 13081.5 15635.5  
## sample estimates:  
## (pseudo)median   
## 14370.5

От хистограмата и t-теста забелязваме, че броя на отзивите за приложенията за автомобили е с нормално разпределение и, яе 95% от данните за цените на платените приложения имат доверителен интервал [3297.4, 28203.7] inversii-kakva e veroqtnostta edno prilojenie da ima rating 4.0/5.0 R wants the s. d. as the parameter, not the variance. We’ll need to take a square root!

1 -pnorm(4.0, mean=mean(Rating), sd=sqrt(var(Rating)))

## [1] 0.6252316

kakva e veroqtnostta edno prilojenie da ima poveche ot 10000 reviews

1 -pnorm(10000, mean=mean(Reviews), sd=sqrt(var(Reviews)))

## [1] 0.560716

kakva e veroqtnostta edno prilojenie da ima size>25.0

1 -pnorm(25.0, mean=mean(Size), sd=sqrt(var(Size)))

## [1] 0.5522594

names(apps)[which(sapply(apps, is.numeric))]

## [1] "Rating" "Reviews" "Size" "Installs"   
## [5] "Price" "Last.Updated"

NumericData <- data.frame(apps$Rating, apps$Size)  
names(NumericData) <- c("Ratings", "Size")  
NumericTable <- as.table(as.matrix(NumericData))

При извадка от 30 приложения, има хипотеза, че средния размер на приложението е 20мб.

SizeSample <- sample(Size, 30)  
t.test(SizeSample)

##   
## One Sample t-test  
##   
## data: SizeSample  
## t = 4.5416, df = 29, p-value = 0.00009051  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 9.86466 26.02867  
## sample estimates:  
## mean of x   
## 17.94667

t.test(SizeSample, mu = 30, alternative = "two.sided")

##   
## One Sample t-test  
##   
## data: SizeSample  
## t = -3.0502, df = 29, p-value = 0.00485  
## alternative hypothesis: true mean is not equal to 30  
## 95 percent confidence interval:  
## 9.86466 26.02867  
## sample estimates:  
## mean of x   
## 17.94667

правим извадка от 25 приложения с техните Size. Chrez t.test проверяваме дали хипотезата е вярна. p-value > 0.05, no 30 vliza v doveritelniq interval, ot kydeto sledva, че хипотезата е логична.

Твърди се, че при извадка от 40 приложения, средния брой изтегляния на дадено приложение е 100000. Вярна ли е хипотезата?

InstallsSample <- sample(Installs, 40)  
t.test(InstallsSample)

##   
## One Sample t-test  
##   
## data: InstallsSample  
## t = 1.7648, df = 39, p-value = 0.08543  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -709972.3 10425177.3  
## sample estimates:  
## mean of x   
## 4857603

# p<0.5

t.test(InstallsSample, mu = 0.1, alternative = "less")

##   
## One Sample t-test  
##   
## data: InstallsSample  
## t = 1.7648, df = 39, p-value = 0.9573  
## alternative hypothesis: true mean is less than 0.1  
## 95 percent confidence interval:  
## -Inf 9495320  
## sample estimates:  
## mean of x   
## 4857603

Правим извадка от 40 приложения и вземаме техният брой инсталации. 100000 се намира в доверителния интервал. Оттук следва, че хипотезата не е правилна.