Batch: RProg/MP/18thDec/12-3pm/W

By Amit J Nambiar

R Project

Data Visualization is performed on the following data set

https://www.kaggle.com/shakthidhar/google-play-store-category-wise-top-500-apps/version/1

Acknowledgement - kaggle.com

Google Play stores top 500 apps data based on their rankings on January 2022 for all the available categories.

Features –

Rank
Serial Number

Name
Name of App

Developer
Name of Company App Developer

Category
Classification of Consumer/Customer sector

Size
Size of the App in MB

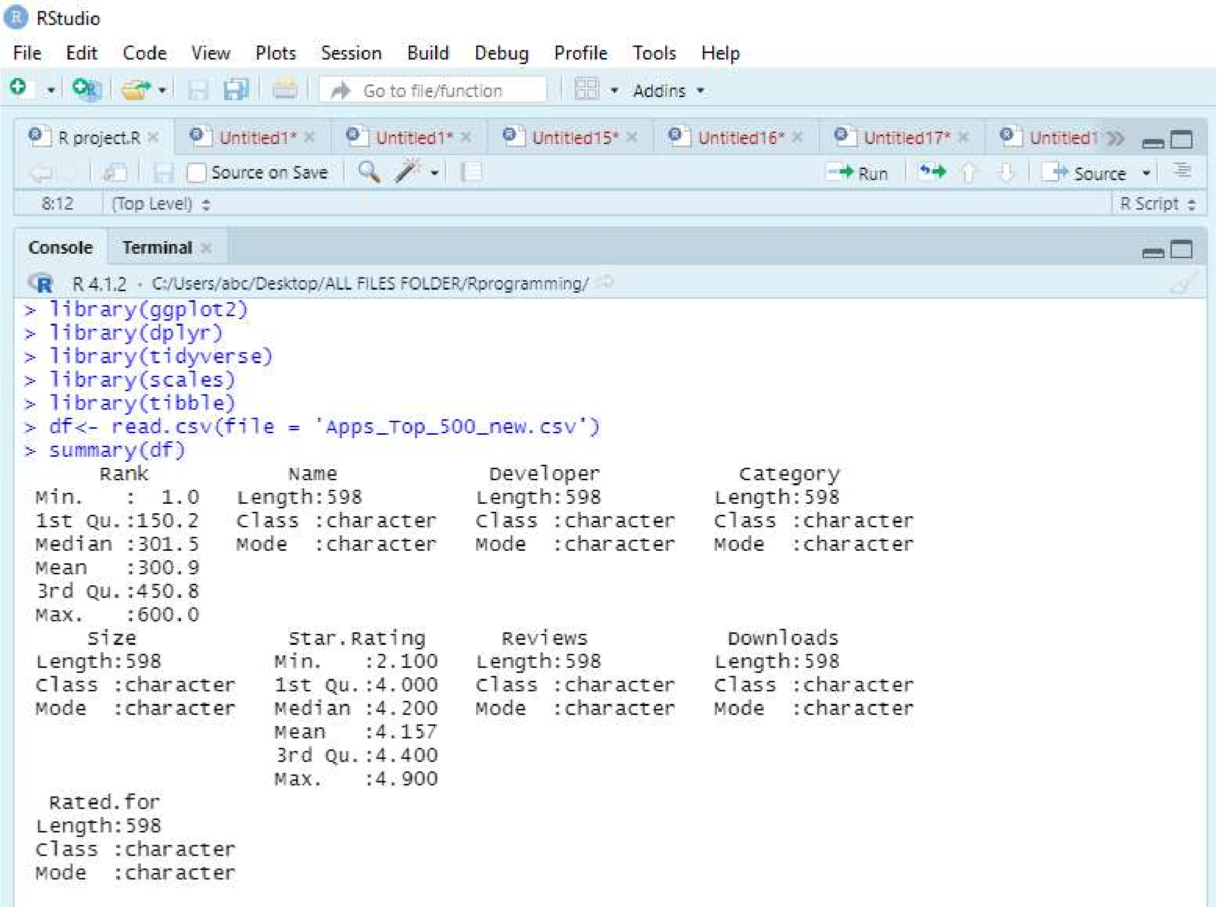
Star.Rating
Rating represented by Stars for showing quality of the app

Reviews
Number of Reviews of people experience after using app

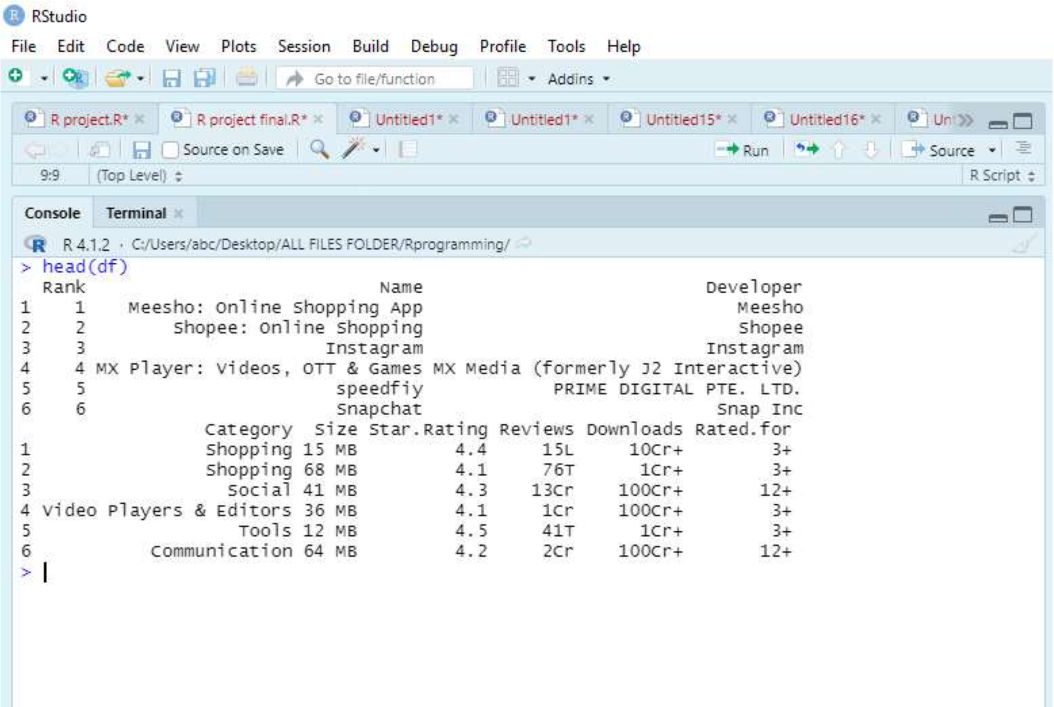
Downloads
Downloads shows how many people have downloaded the app

Rated.for
Rated for shows the age group for whom the apps are created and targeted to

R program to upload dataset in dataframe



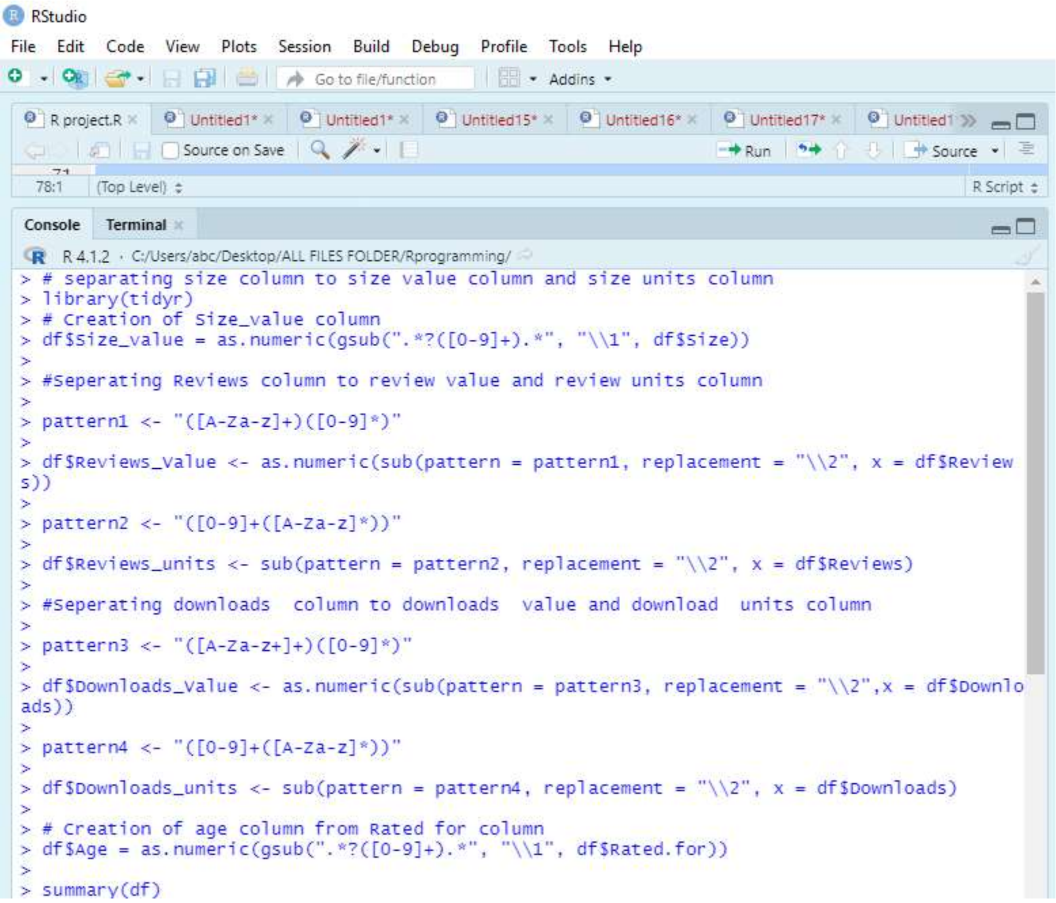
```
> library(ggplot2)
> library(dplyr)
> library(tidyverse)
> library(scales)
> library(tibble)
> df<- read.csv(file = 'Apps_Top_500_new.csv')
> summary(df)
      Rank             Name            Developer           Category        
 Min.   :  1.0   Length:598         Length:598         Length:598        
 1st Qu.:150.2   Class :character   Class :character   Class :character  
 Median :301.5   Mode  :character   Mode  :character   Mode  :character  
 Mean   :300.9                                                           
 3rd Qu.:450.8                                                           
 Max.   :600.0                                                           
     Size            Star.Rating       Reviews           Downloads        
 Length:598         Min.   :2.100   Length:598         Length:598        
 Class :character   1st Qu.:4.000   Class :character   Class :character  
 Mode  :character   Median :4.200   Mode  :character   Mode  :character  
                    Mean   :4.157                                        
                    3rd Qu.:4.400                                        
                    Max.   :4.900                                        
  Rated.for        
 Length:598        
 Class :character  
 Mode  :character  
```



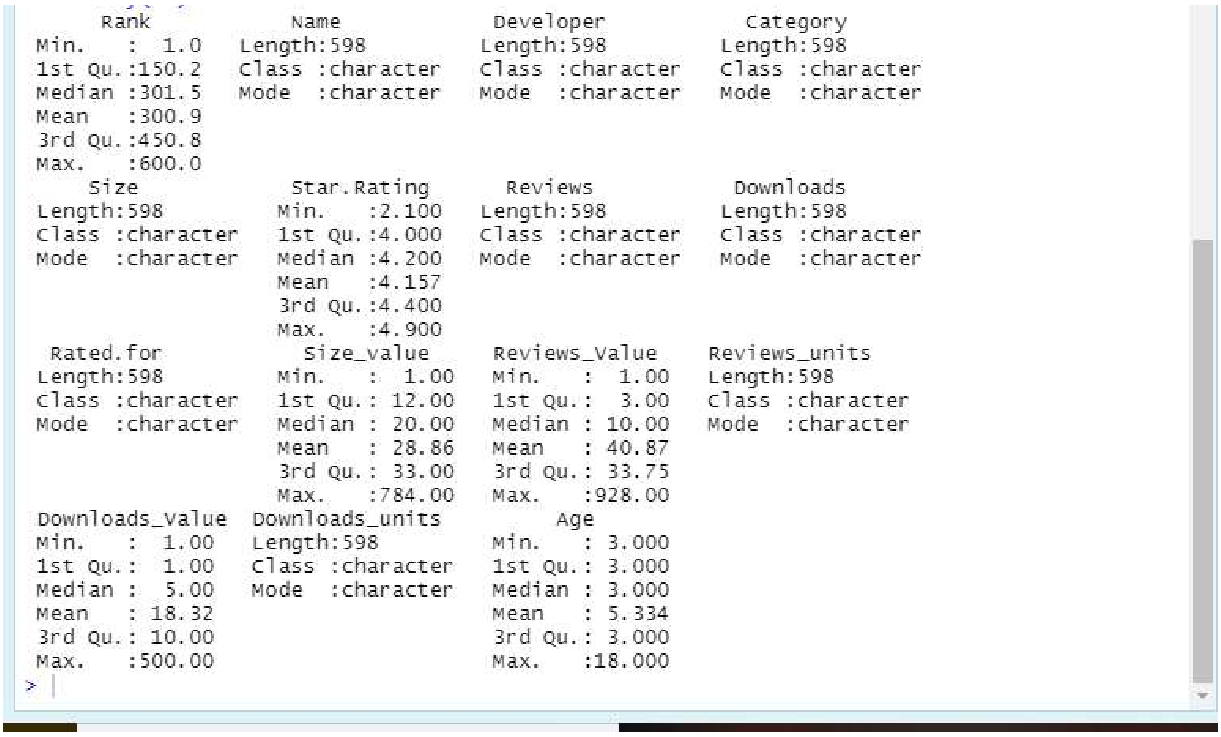
```
> head(df)
  Rank                                Name                                 Developer
1    1       Meesho: Online Shopping App                                    Meesho
2    2           Shopee: Online Shopping                                    Shopee
3    3                         Instagram                                 Instagram
4    4 MX Player: Videos, OTT & Games MX Media (formerly J2 Interactive)
5    5                          speedfiy                      PRIME DIGITAL PTE. LTD.
6    6                          Snapchat                                  Snap Inc
               Category  Size Star.Rating Reviews Downloads Rated.for
1              Shopping 15 MB         4.4     15L    10Cr+        3+
2              Shopping 68 MB         4.1     76T     1Cr+        3+
3                Social 41 MB         4.3    13Cr   100Cr+       12+
4 Video Players & Editors 36 MB         4.1     1Cr   100Cr+        3+
5                 Tools 12 MB         4.5     41T     1Cr+        3+
6         Communication 64 MB         4.2     2Cr   100Cr+       12+
> |
```

From the above it is seen that maximum (7 columns) are character class. Since plots cannot be created on character data it is decided to separate the column values into numeric and character for the following columns by creating additional columns as shown below

Sr No	**Column in dataset**	**Additional column in dataframe**	
		Numeric	**Character**
1	Size	Size_value	-
2	Reviews	Reviews_value	Reviews_units
3	Downloads	Downloads_value	Downloads_units
4	Rated.for	Age	-

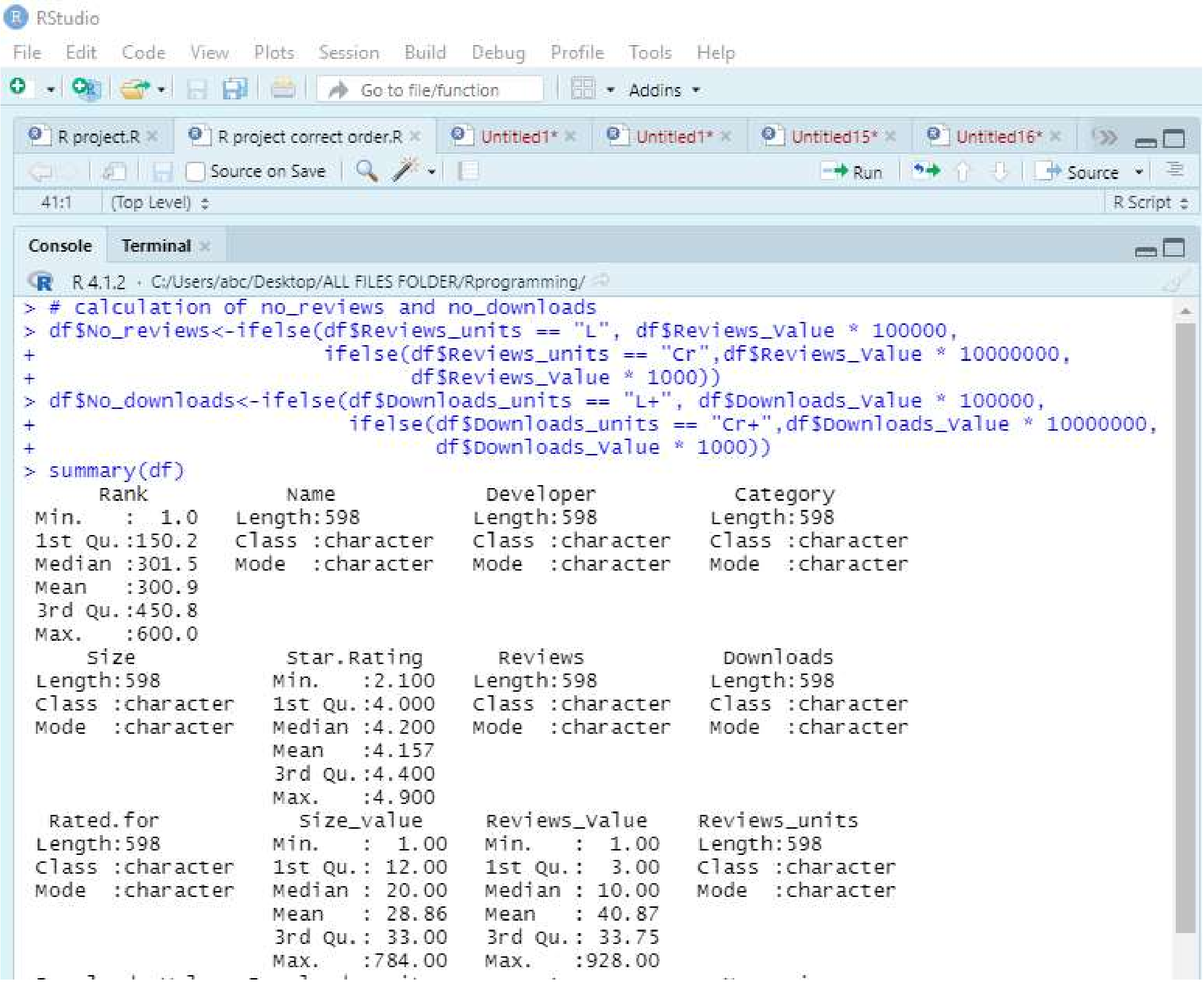


```
> # separating size column to size value column and size units column
> library(tidyr)
> # Creation of Size_value column
> df$Size_value = as.numeric(gsub(".*?([0-9]+).*", "\\1", df$Size))
>
> #Seperating Reviews column to review value and review units column
>
> pattern1 <- "([A-Za-z]+)([0-9]*)"
>
> df$Reviews_value <- as.numeric(sub(pattern = pattern1, replacement = "\\2", x = df$Review
s))
>
> pattern2 <- "([0-9]+([A-Za-z]*))"
>
> df$Reviews_units <- sub(pattern = pattern2, replacement = "\\2", x = df$Reviews)
>
> #Seperating downloads  column to downloads  value and download  units column
>
> pattern3 <- "([A-Za-z+]+)([0-9]*)"
>
> df$Downloads_value <- as.numeric(sub(pattern = pattern3, replacement = "\\2",x = df$Downlo
ads))
>
> pattern4 <- "([0-9]+([A-Za-z]*))"
>
> df$Downloads_units <- sub(pattern = pattern4, replacement = "\\2", x = df$Downloads)
>
> # Creation of age column from Rated for column
> df$Age = as.numeric(gsub(".*?([0-9]+).*", "\\1", df$Rated.for))
>
> summary(df)
```

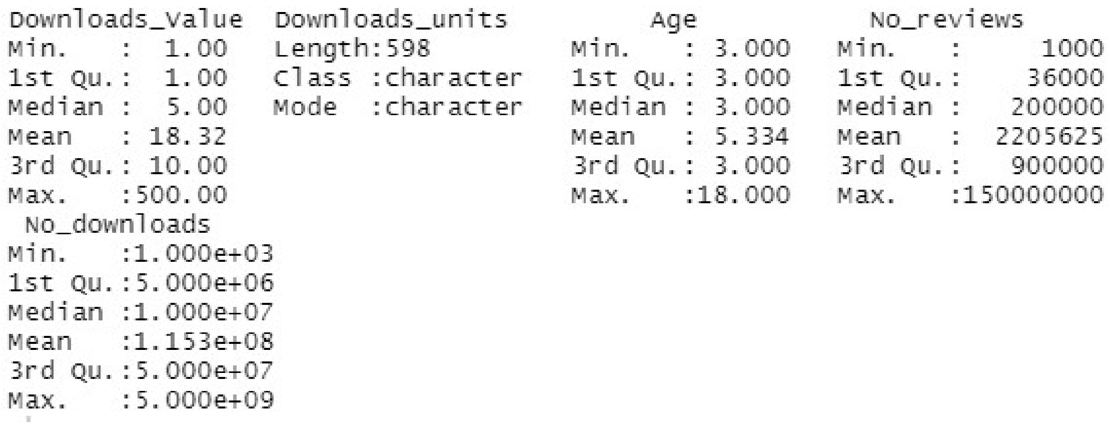


```
      Rank            Name              Developer          Category        
 Min.   :  1.0   Length:598         Length:598         Length:598        
 1st Qu.:150.2   Class :character   Class :character   Class :character  
 Median :301.5   Mode  :character   Mode  :character   Mode  :character  
 Mean   :300.9                                                           
 3rd Qu.:450.8                                                           
 Max.   :600.0                                                           
     Size            Star.Rating      Reviews           Downloads        
 Length:598         Min.   :2.100   Length:598         Length:598        
 Class :character   1st Qu.:4.000   Class :character   Class :character  
 Mode  :character   Median :4.200   Mode  :character   Mode  :character  
                    Mean   :4.157                                        
                    3rd Qu.:4.400                                        
                    Max.   :4.900                                        
  Rated.for           Size_value      Reviews_Value    Reviews_units     
 Length:598         Min.   :  1.00   Min.   :  1.00   Length:598        
 Class :character   1st Qu.: 12.00   1st Qu.:  3.00   Class :character  
 Mode  :character   Median : 20.00   Median : 10.00   Mode  :character  
                    Mean   : 28.86   Mean   : 40.87                     
                    3rd Qu.: 33.00   3rd Qu.: 33.75                     
                    Max.   :784.00   Max.   :928.00                     
 Downloads_Value  Downloads_units         Age        
 Min.   :  1.00   Length:598         Min.   : 3.000  
 1st Qu.:  1.00   Class :character   1st Qu.: 3.000  
 Median :  5.00   Mode  :character   Median : 3.000  
 Mean   : 18.32                      Mean   : 5.334  
 3rd Qu.: 10.00                      3rd Qu.: 3.000  
 Max.   :500.00                      Max.   :18.000  
> 
```

Since Reviews and Downloads are in different units of Thousand, Lac and Crore just separating the numeric and character is not enough. To convert them to a common scale two columns No_reviews and No_downloads are calculated as under.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
41:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

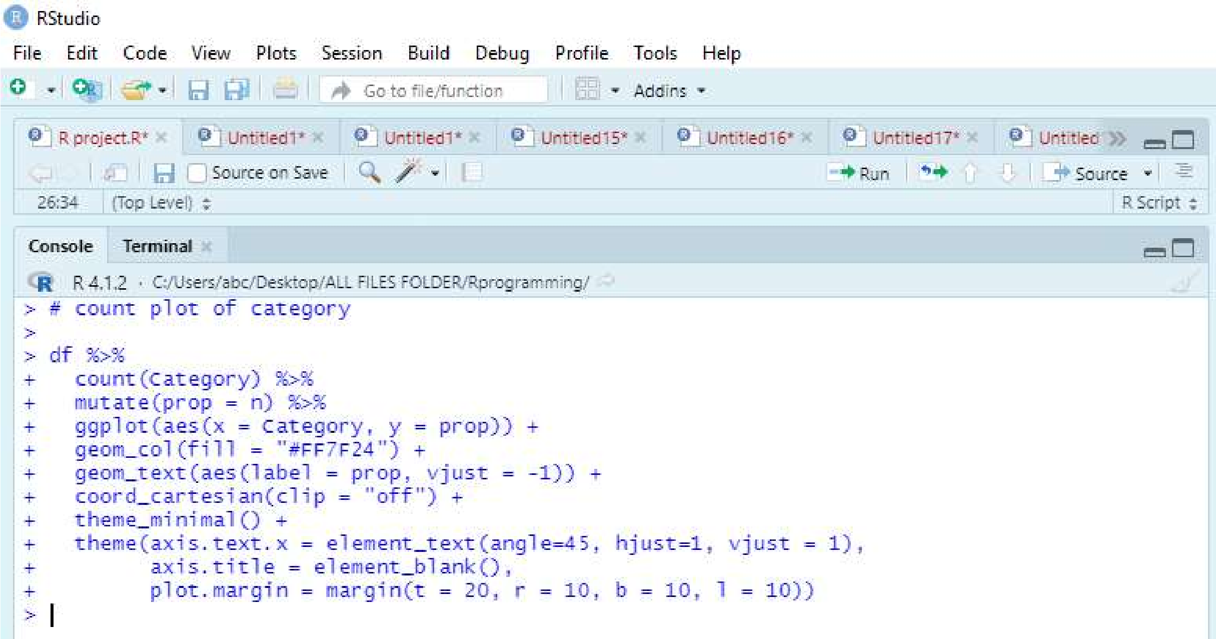
```
> # calculation of no_reviews and no_downloads
> df$No_reviews<-ifelse(df$Reviews_units == "L", df$Reviews_value * 100000,
+                       ifelse(df$Reviews_units == "Cr",df$Reviews_value * 10000000,
+                              df$Reviews_value * 1000))
> df$No_downloads<-ifelse(df$Downloads_units == "L+", df$Downloads_value * 100000,
+                         ifelse(df$Downloads_units == "Cr+",df$Downloads_value * 10000000,
+                                df$Downloads_value * 1000))
> summary(df)
      Rank            Name             Developer          Category        
 Min.   :  1.0   Length:598         Length:598         Length:598        
 1st Qu.:150.2   Class :character   Class :character   Class :character  
 Median :301.5   Mode  :character   Mode  :character   Mode  :character  
 Mean   :300.9                                                           
 3rd Qu.:450.8                                                           
 Max.   :600.0                                                           
     Size            Star.Rating      Reviews           Downloads        
 Length:598         Min.   :2.100   Length:598         Length:598        
 Class :character   1st Qu.:4.000   Class :character   Class :character  
 Mode  :character   Median :4.200   Mode  :character   Mode  :character  
                    Mean   :4.157                                        
                    3rd Qu.:4.400                                        
                    Max.   :4.900                                        
  Rated.for           Size_value      Reviews_value    Reviews_units     
 Length:598         Min.   :  1.00   Min.   :  1.00   Length:598        
 Class :character   1st Qu.: 12.00   1st Qu.:  3.00   Class :character  
 Mode  :character   Median : 20.00   Median : 10.00   Mode  :character  
                    Mean   : 28.86   Mean   : 40.87                     
                    3rd Qu.: 33.00   3rd Qu.: 33.75                     
                    Max.   :784.00   Max.   :928.00                     
```



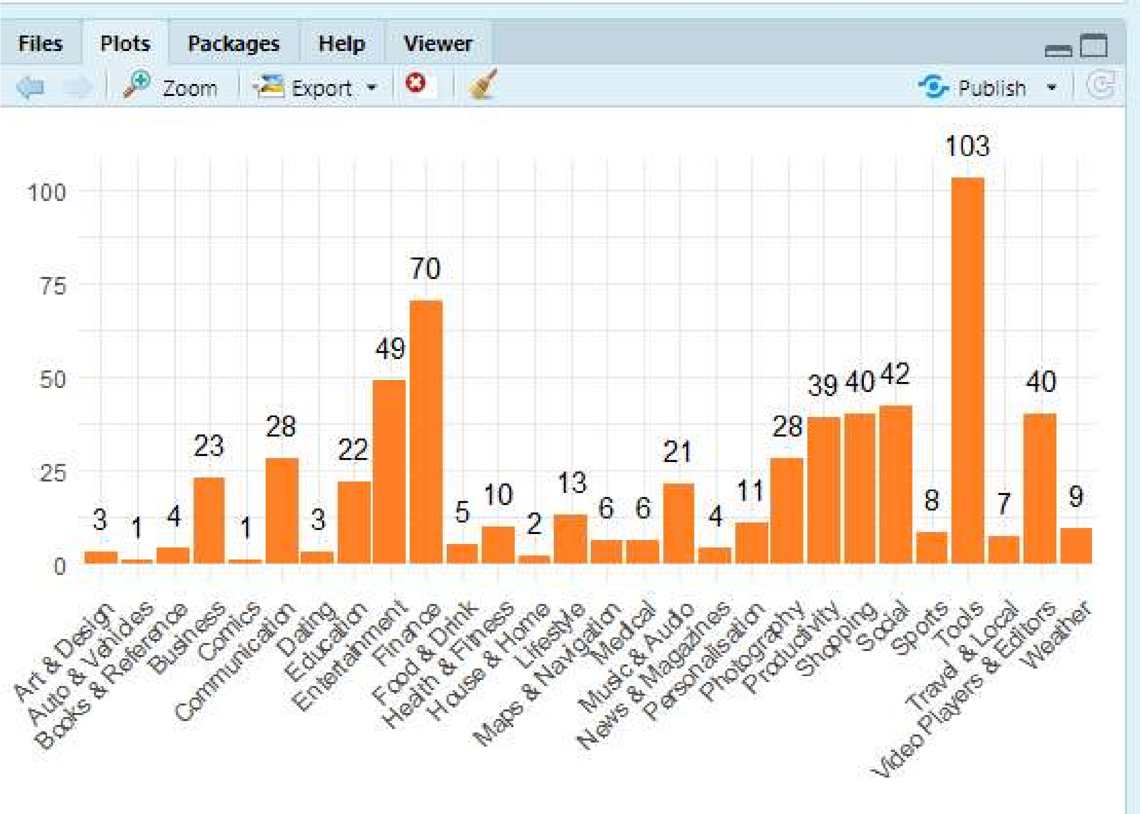
```
Downloads_value  Downloads_units         Age           No_reviews       
Min.   :  1.00   Length:598         Min.   : 3.000   Min.   :     1000  
1st Qu.:  1.00   Class :character   1st Qu.: 3.000   1st Qu.:    36000  
Median :  5.00   Mode  :character   Median : 3.000   Median :   200000  
Mean   : 18.32                      Mean   : 5.334   Mean   :  2205625  
3rd Qu.: 10.00                      3rd Qu.: 3.000   3rd Qu.:   900000  
Max.   :500.00                      Max.   :18.000   Max.   :150000000  
 No_downloads      
Min.   :1.000e+03  
1st Qu.:5.000e+06  
Median :1.000e+07  
Mean   :1.153e+08  
3rd Qu.:5.000e+07  
Max.   :5.000e+09  
```

Univariate analysis

Count plot of **Category**

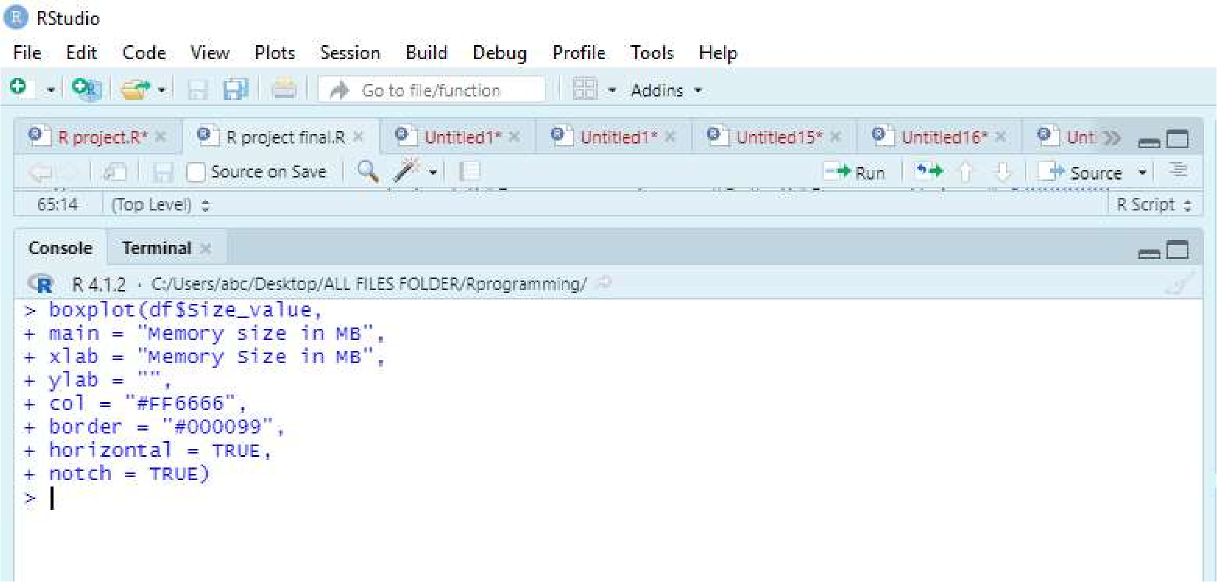


```
> # count plot of category
>
> df %>%
+   count(Category) %>%
+   mutate(prop = n) %>%
+   ggplot(aes(x = Category, y = prop)) +
+   geom_col(fill = "#FF7F24") +
+   geom_text(aes(label = prop, vjust = -1)) +
+   coord_cartesian(clip = "off") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle=45, hjust=1, vjust = 1),
+         axis.title = element_blank(),
+         plot.margin = margin(t = 20, r = 10, b = 10, l = 10))
> |
```

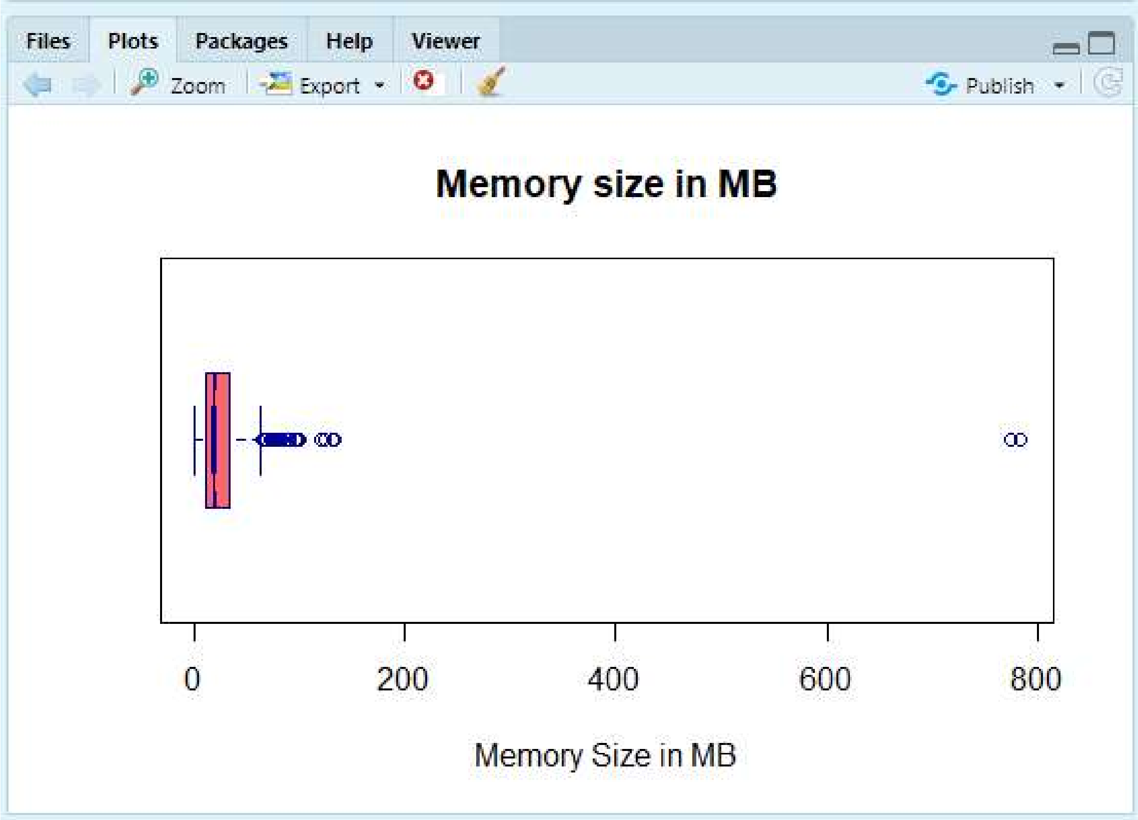


The **maximum** no of apps is in the **Tools** Category (103) followed by **Finance**Category (70)

Box plot of **Size_value**

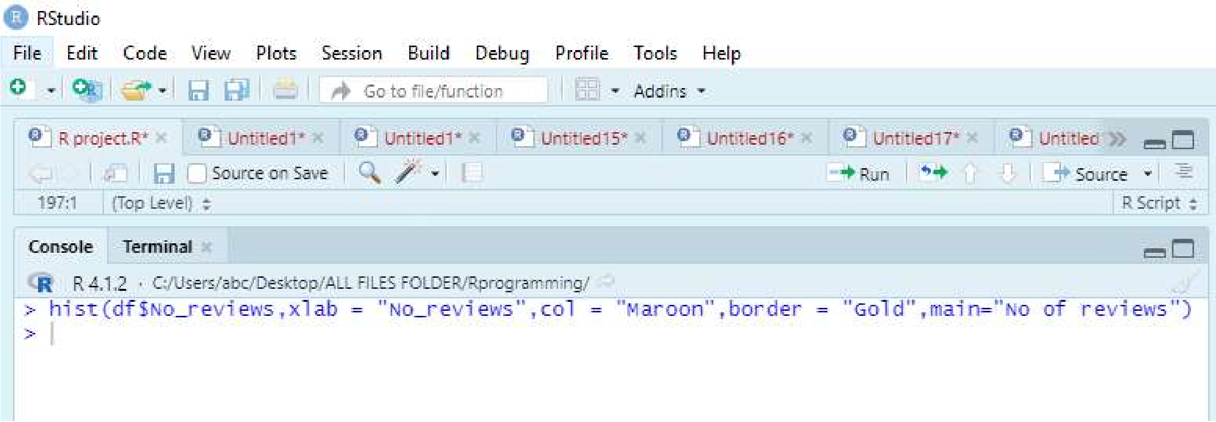


```
> boxplot(df$Size_value,
+ main = "Memory size in MB",
+ xlab = "Memory Size in MB",
+ ylab = "",
+ col = "#FF6666",
+ border = "#000099",
+ horizontal = TRUE,
+ notch = TRUE)
> |
```

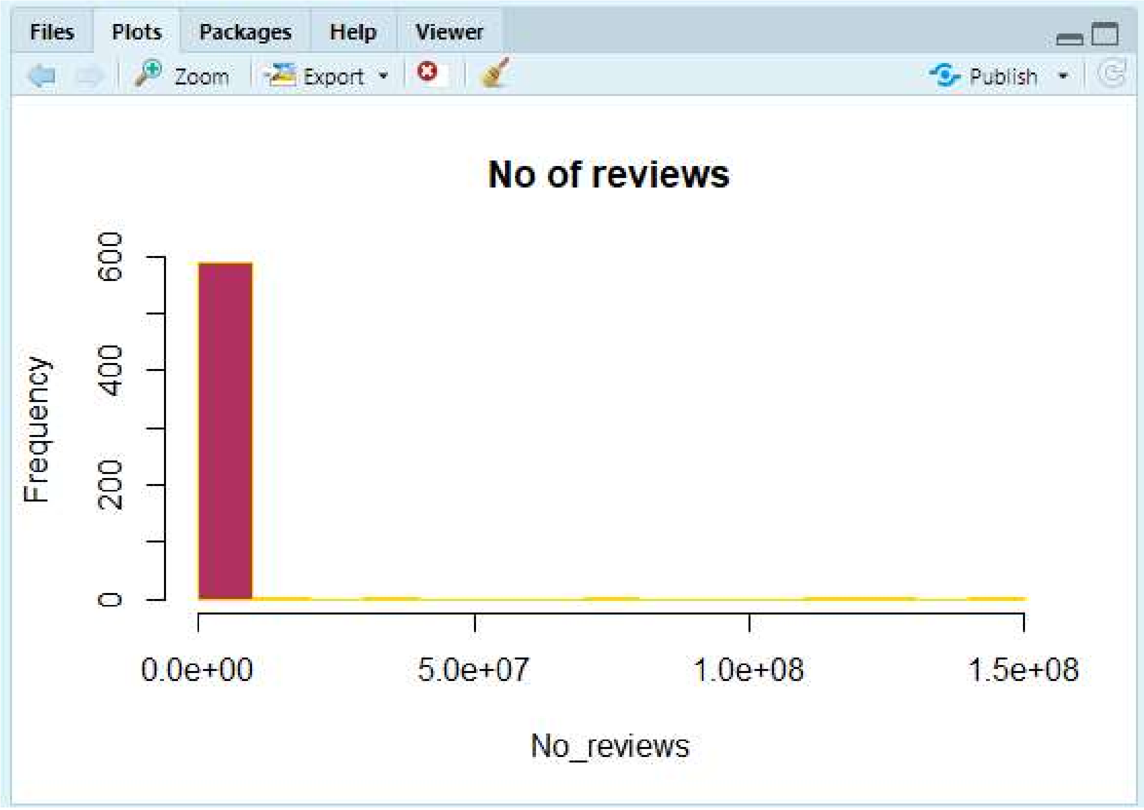


The maximum number of apps have memory size between 0-50 MB. There are 2 outliers between 600MB - 800 MB

Histogram of **No_reviews**



```
> hist(df$No_reviews,xlab = "No_reviews",col = "Maroon",border = "Gold",main="No of reviews")
> |
```

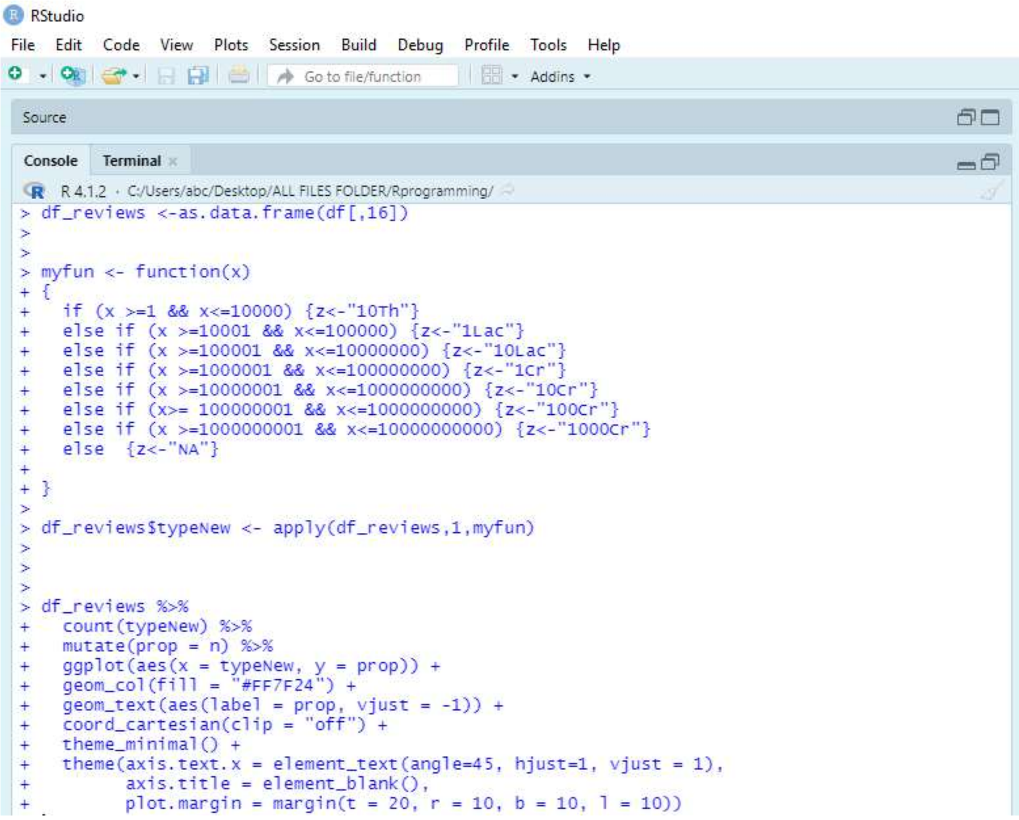


Since the above histogram shows a concentration of values in the first interval it is decided to Classify the no of reviews further as below to get a better picture.

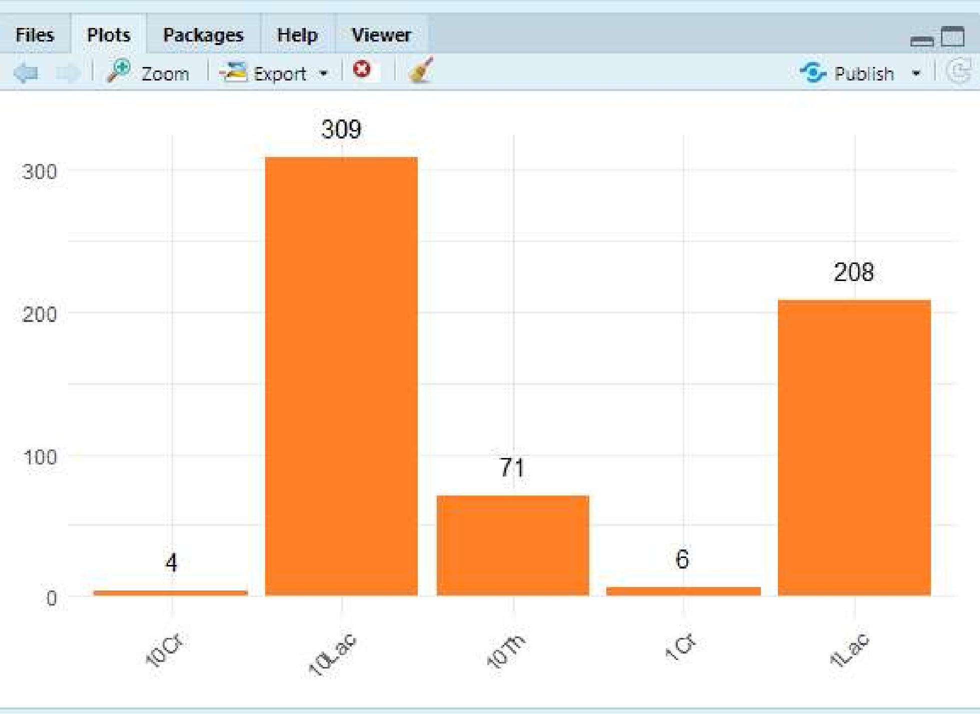
Sr No	No of Reviews	typeNew
1	1-10000	10Th
2	10001-100000	1Lac
3	100001-1000000	10Lac
4	1000001-10000000	1Cr
5	10000001-100000000	10Cr
6	100000001-1000000000	100Cr
7	1000000001-10000000000	1000Cr

A new data frame df_reviews is created with columns (No_reviews ,typeNew)

Count plot of **typeNew**

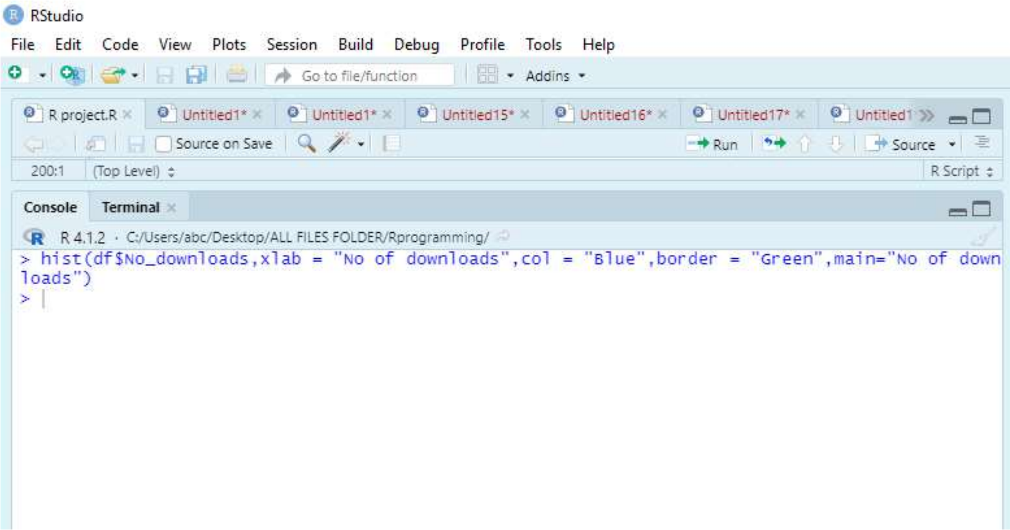

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Source
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> df_reviews <-as.data.frame(df[,16])
>
>
> myfun <- function(x)
+ {
+   if (x >=1 && x<=10000) {z<-"10Th"}
+   else if (x >=10001 && x<=100000) {z<-"1Lac"}
+   else if (x >=100001 && x<=10000000) {z<-"10Lac"}
+   else if (x >=1000001 && x<=100000000) {z<-"1Cr"}
+   else if (x >=10000001 && x<=1000000000) {z<-"10Cr"}
+   else if (x>= 100000001 && x<=1000000000) {z<-"100Cr"}
+   else if (x >=1000000001 && x<=10000000000) {z<-"1000Cr"}
+   else  {z<-"NA"}
+
+ }
>
> df_reviews$typeNew <- apply(df_reviews,1,myfun)
>
>
>
> df_reviews %>%
+   count(typeNew) %>%
+   mutate(prop = n) %>%
+   ggplot(aes(x = typeNew, y = prop)) +
+   geom_col(fill = "#FF7F24") +
+   geom_text(aes(label = prop, vjust = -1)) +
+   coord_cartesian(clip = "off") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle=45, hjust=1, vjust = 1),
+         axis.title = element_blank(),
+         plot.margin = margin(t = 20, r = 10, b = 10, l = 10))
```

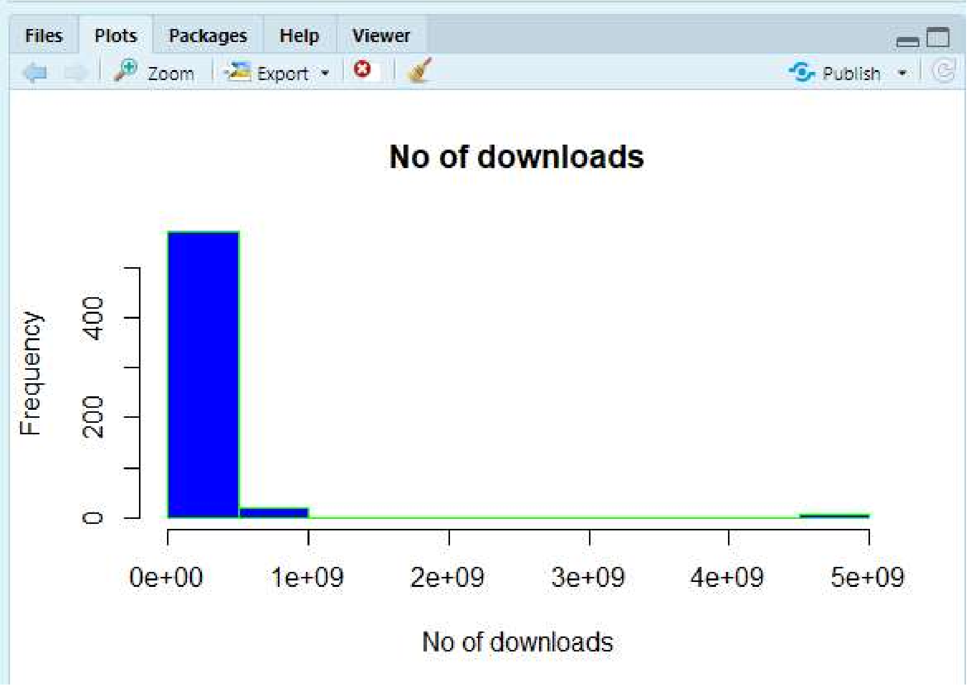


The no of reviews is between (1 lac to 10 Lac) for **309** apps
The no of reviews is between (10 thousand to 1 Lac) for **208** apps

Histogram of **No_downloads**



```
> hist(df$No_downloads,xlab = "No of downloads",col = "Blue",border = "Green",main="No of down
loads")
> |
```

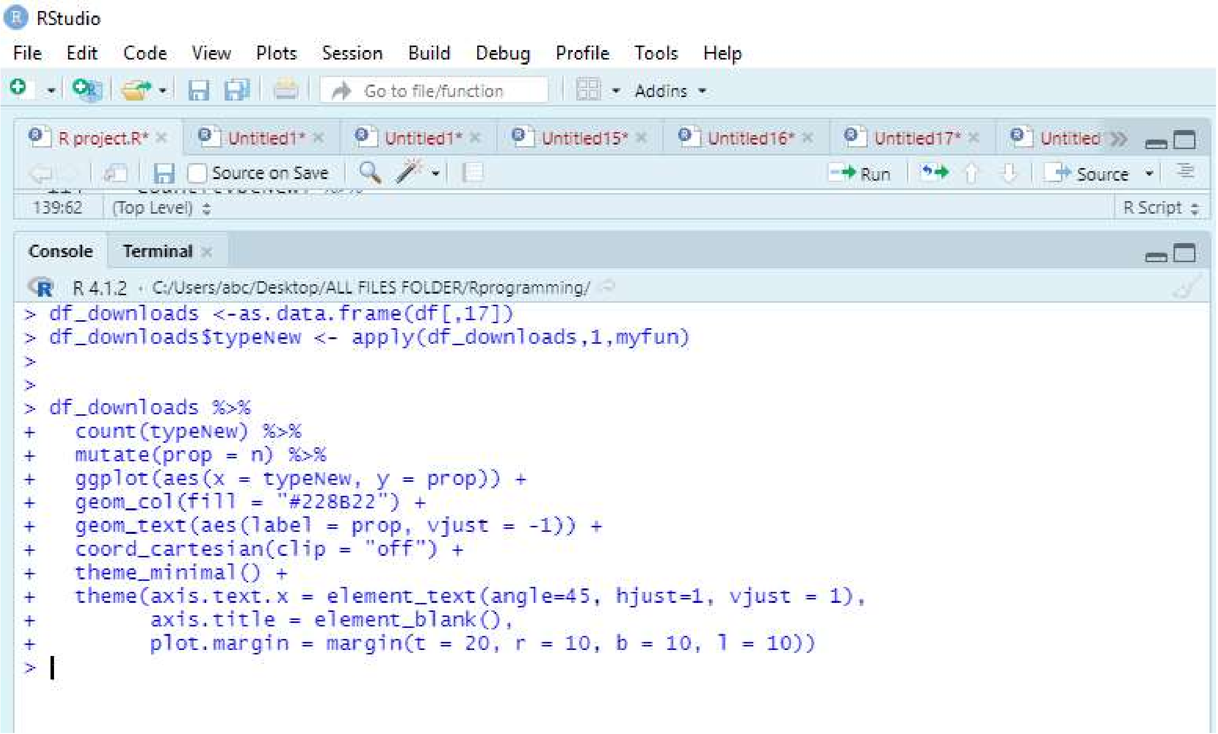


Since the above histogram shows a concentration of values in the first interval it is decided to Classify the no of downloads further as below to get a better picture.

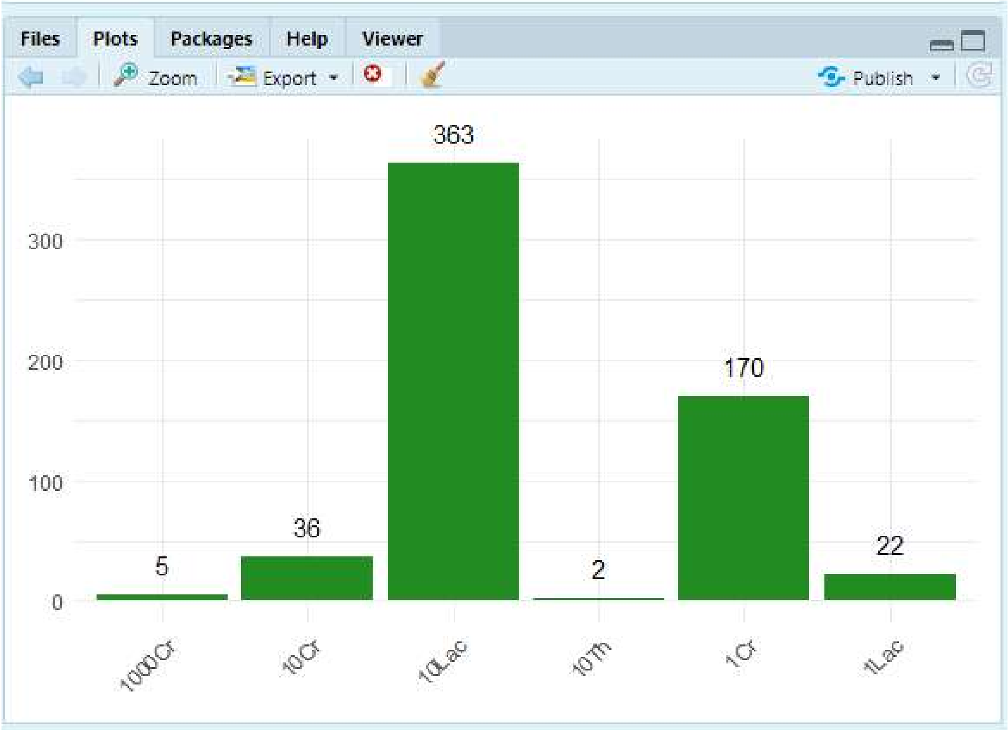
Sr No	No of Downloads	typeNew
1	1-10000	10Th
2	10001-100000	1Lac
3	100001-1000000	10Lac
4	1000001-10000000	1Cr
5	10000001-100000000	10Cr
6	100000001-1000000000	100Cr
7	1000000001-10000000000	1000Cr

A new data frame df_downloads is created with columns (No_downloads, typeNew)

Count plot of **typeNew**

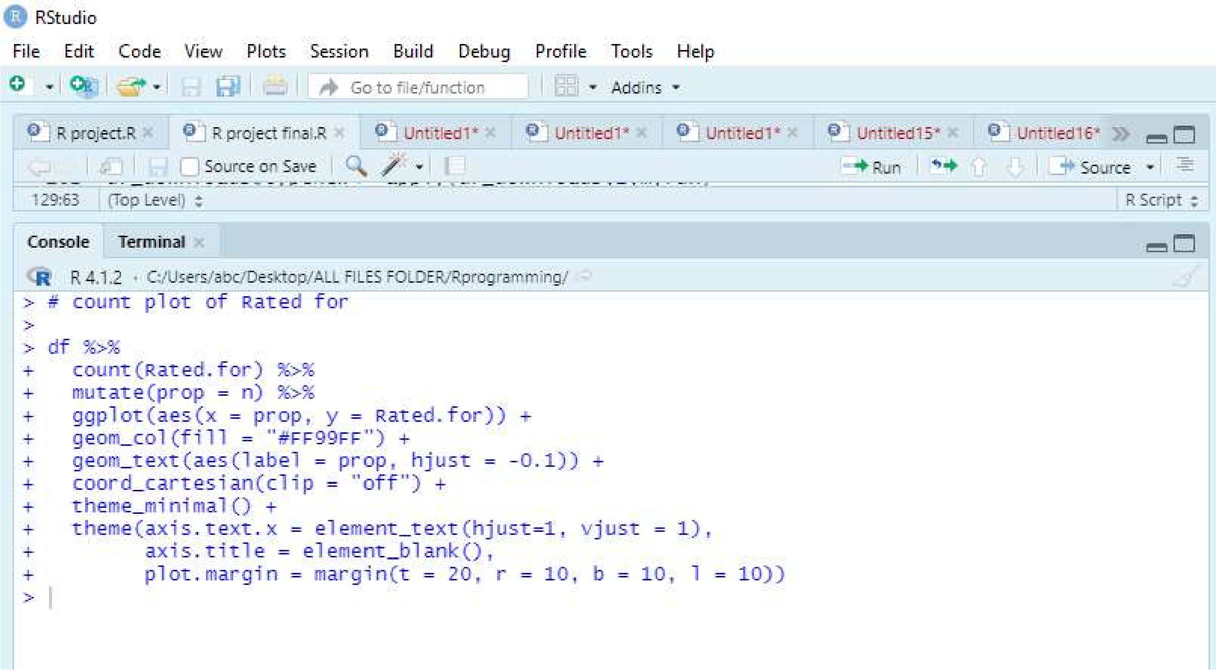


```
> df_downloads <-as.data.frame(df[,17])
> df_downloads$typeNew <- apply(df_downloads,1,myfun)
>
>
> df_downloads %>%
+   count(typeNew) %>%
+   mutate(prop = n) %>%
+   ggplot(aes(x = typeNew, y = prop)) +
+   geom_col(fill = "#228B22") +
+   geom_text(aes(label = prop, vjust = -1)) +
+   coord_cartesian(clip = "off") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle=45, hjust=1, vjust = 1),
+         axis.title = element_blank(),
+         plot.margin = margin(t = 20, r = 10, b = 10, l = 10))
> |
```

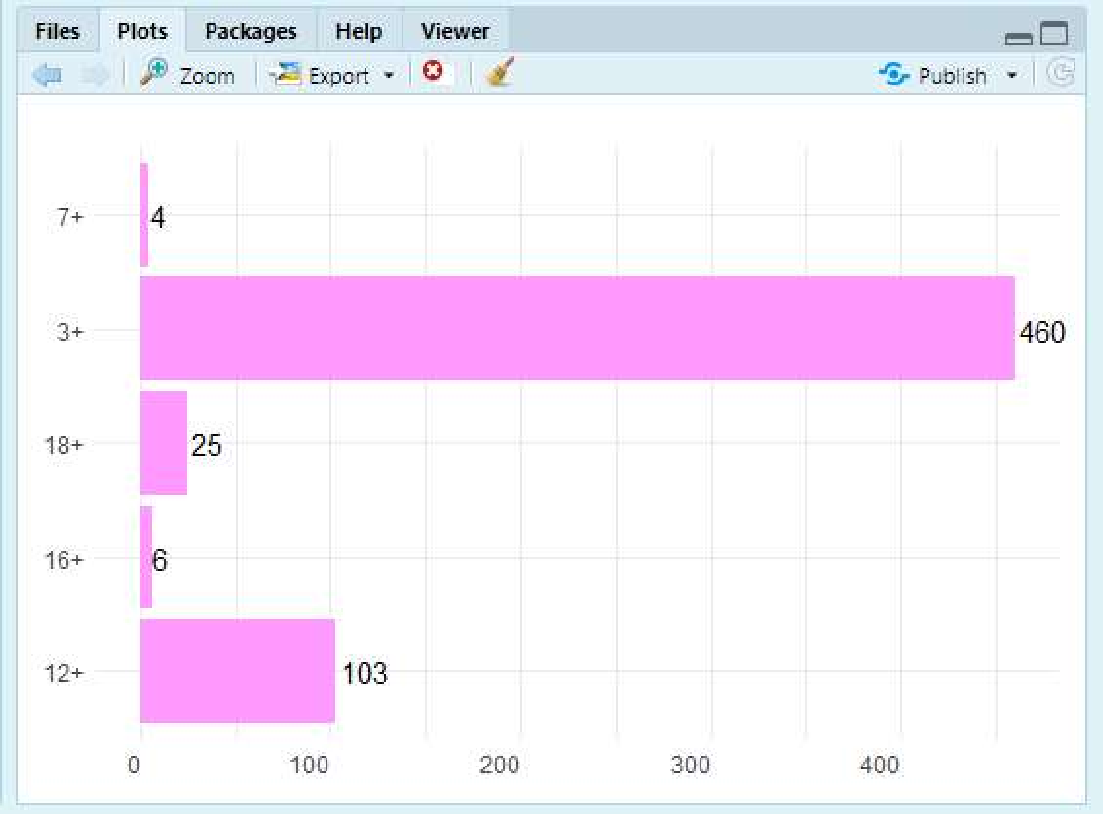


The no of downloads is between (1 Lac to 10 Lac) for **363** apps
The no of downloads is between (10 Lac to 1 Crore) for **170** apps

Count plot of **Rated.for**

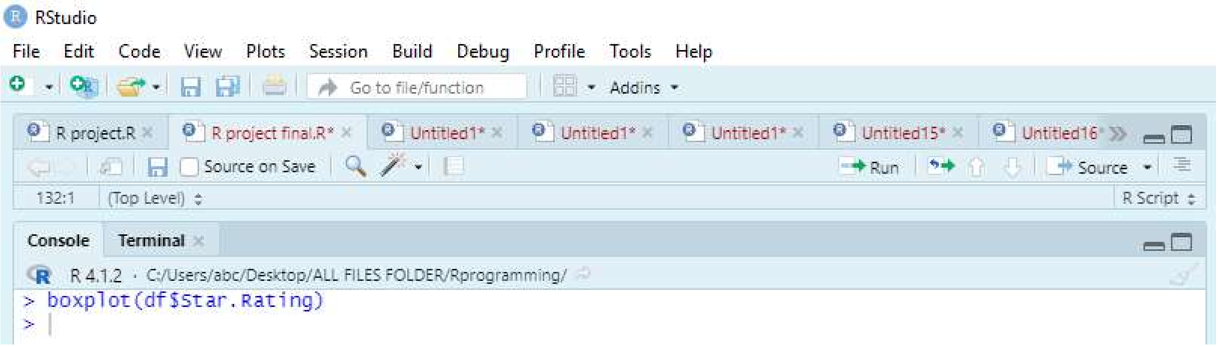


```
> # count plot of Rated for
>
> df %>%
+   count(Rated.for) %>%
+   mutate(prop = n) %>%
+   ggplot(aes(x = prop, y = Rated.for)) +
+   geom_col(fill = "#FF99FF") +
+   geom_text(aes(label = prop, hjust = -0.1)) +
+   coord_cartesian(clip = "off") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(hjust=1, vjust = 1),
+         axis.title = element_blank(),
+         plot.margin = margin(t = 20, r = 10, b = 10, l = 10))
>
```

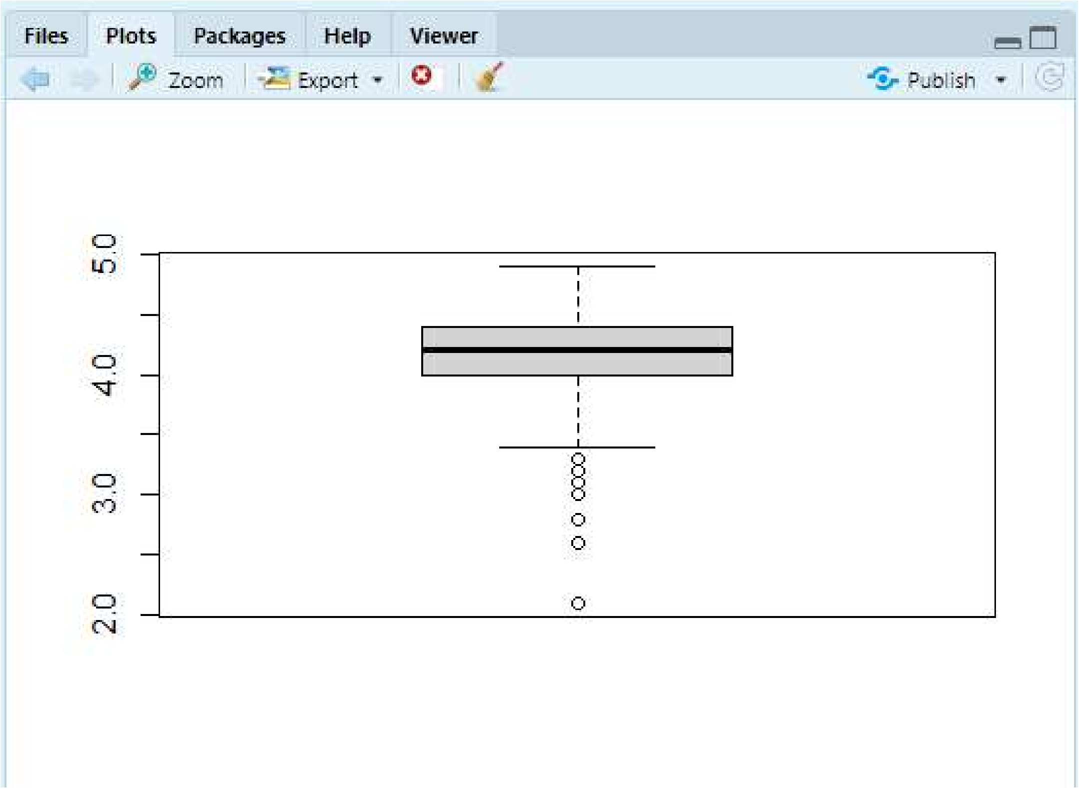


Maximum no of apps **460** is in 3+ Rated.for followed by **103** in 12 + Rated.for

Boxplot of **Star.Rating**



```
> boxplot(df$Star.Rating)
> |
```



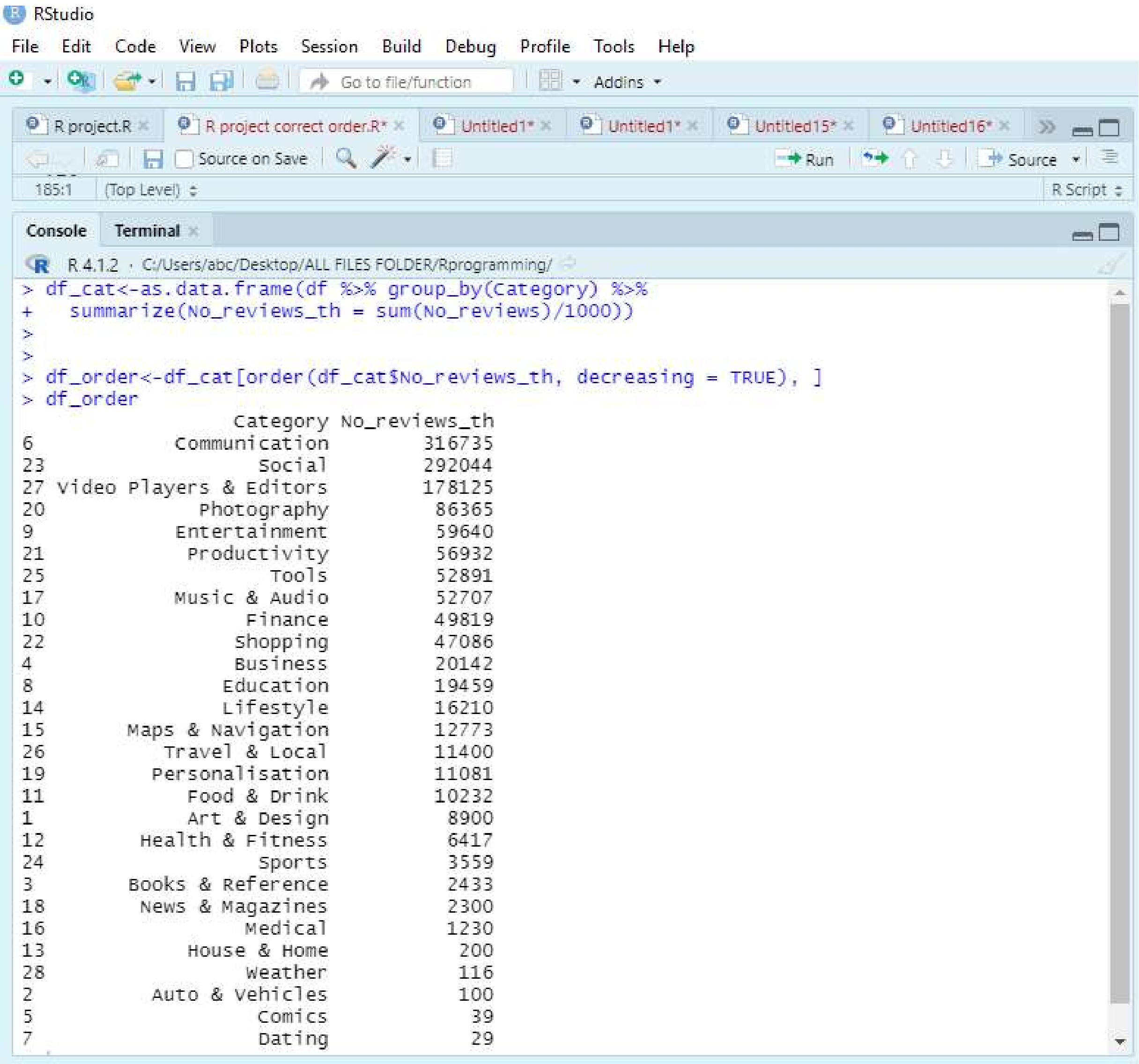
Maximum values are in range of **(4 – 4.5)** There are **seven** outliers with values in range of (**2 to 3.5)**

Bivariate analysis

No_reviews , Category

The sum(No_reviews) for each category is calculated and divided by 1000 for better visualization. A new data frame df_cat with column (Category, No_reviews_th) is created.

The data frame df_cat is sorted in descending order

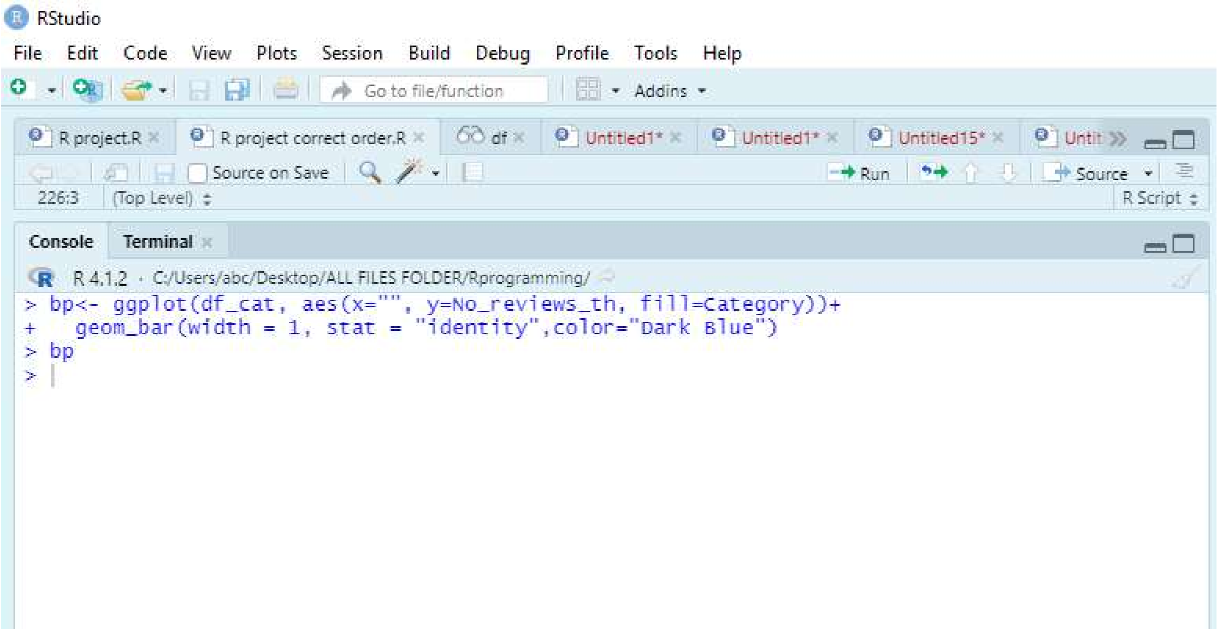


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R*
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
185:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

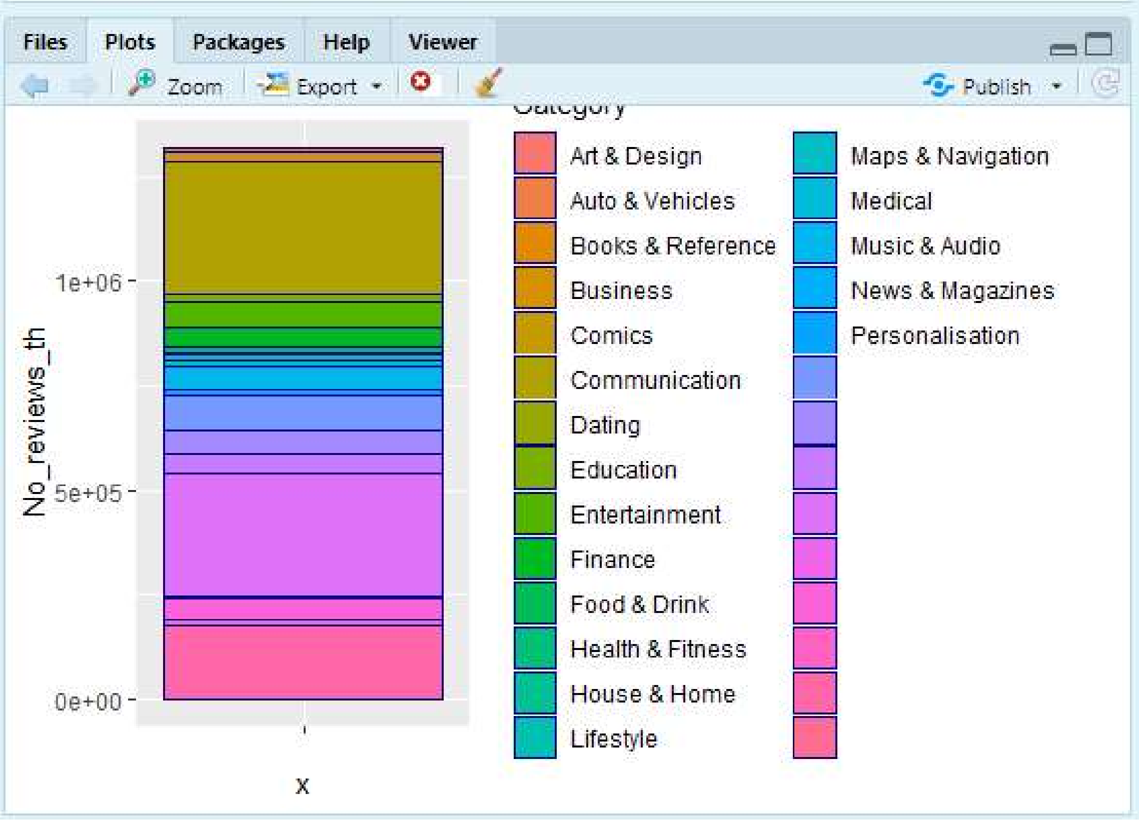
```
> df_cat<-as.data.frame(df %>% group_by(Category) %>%
+   summarize(No_reviews_th = sum(No_reviews)/1000))
>
>
> df_order<-df_cat[order(df_cat$No_reviews_th, decreasing = TRUE), ]
> df_order
                 Category No_reviews_th
6           Communication        316735
23                 Social        292044
27 Video Players & Editors       178125
20            Photography         86365
9           Entertainment         59640
21           Productivity         56932
25                  Tools         52891
17          Music & Audio         52707
10                Finance         49819
22               Shopping         47086
4                Business         20142
8               Education         19459
14              Lifestyle         16210
15       Maps & Navigation        12773
26          Travel & Local        11400
19        Personalisation         11081
11           Food & Drink         10232
1            Art & Design          8900
12       Health & Fitness          6417
24                 Sports          3559
3        Books & Reference         2433
18        News & Magazines         2300
16                Medical          1230
13           House & Home           200
28                Weather           116
2          Auto & Vehicles           100
5                  Comics            39
7                  Dating            29
```

The maximum value of sum of Reviews is in **Communication** Category

The above data can be visualized using ggplot as follows

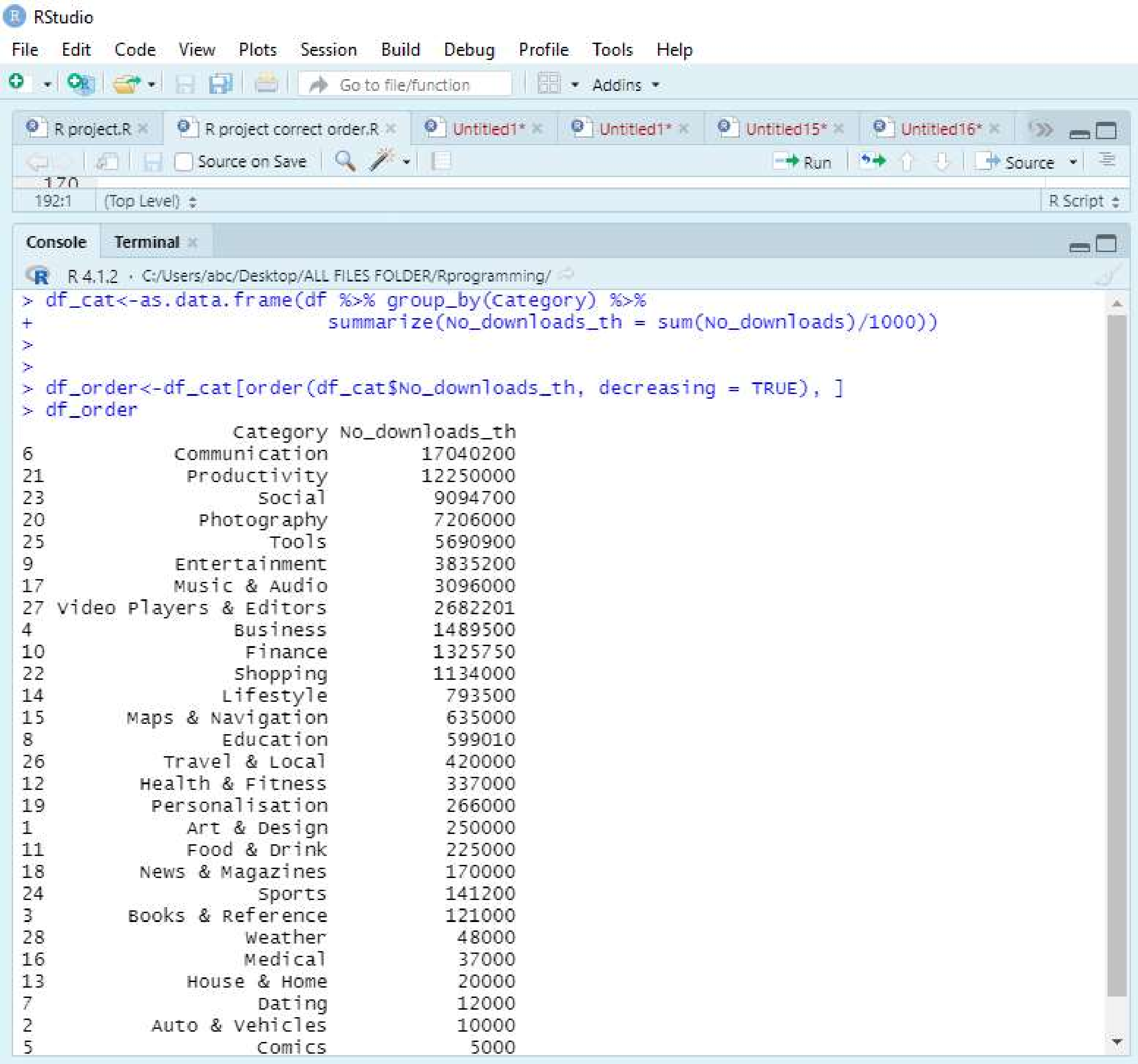

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R
df
Untitled1*
Untitled1*
Untitled15*
Untit
Source on Save
Run
Source
226:3
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> bp<- ggplot(df_cat, aes(x="", y=No_reviews_th, fill=Category))+
+   geom_bar(width = 1, stat = "identity",color="Dark Blue")
> bp
> |
```



No_downloads , Category

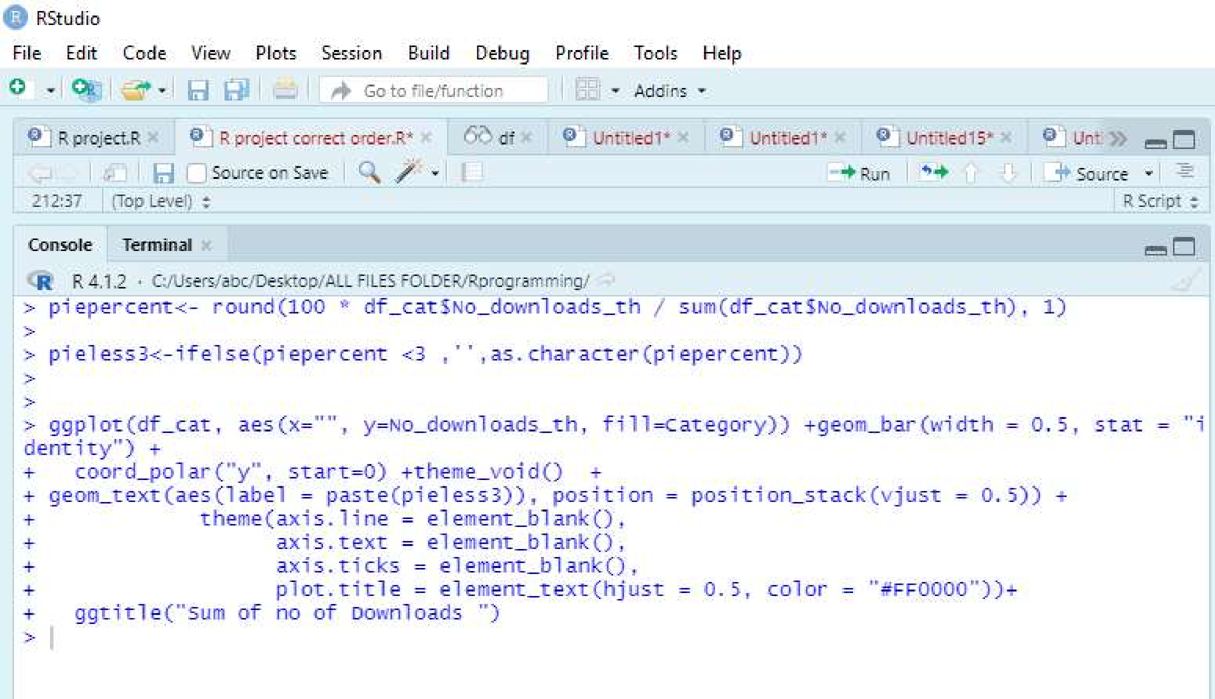
The sum(No_downloads) for each category is calculated and divided by 1000 for better visualization. A new data frame df_cat with column (Category, No_downloads_th) is created. The data frame df_cat is sorted in descending order


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
192:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

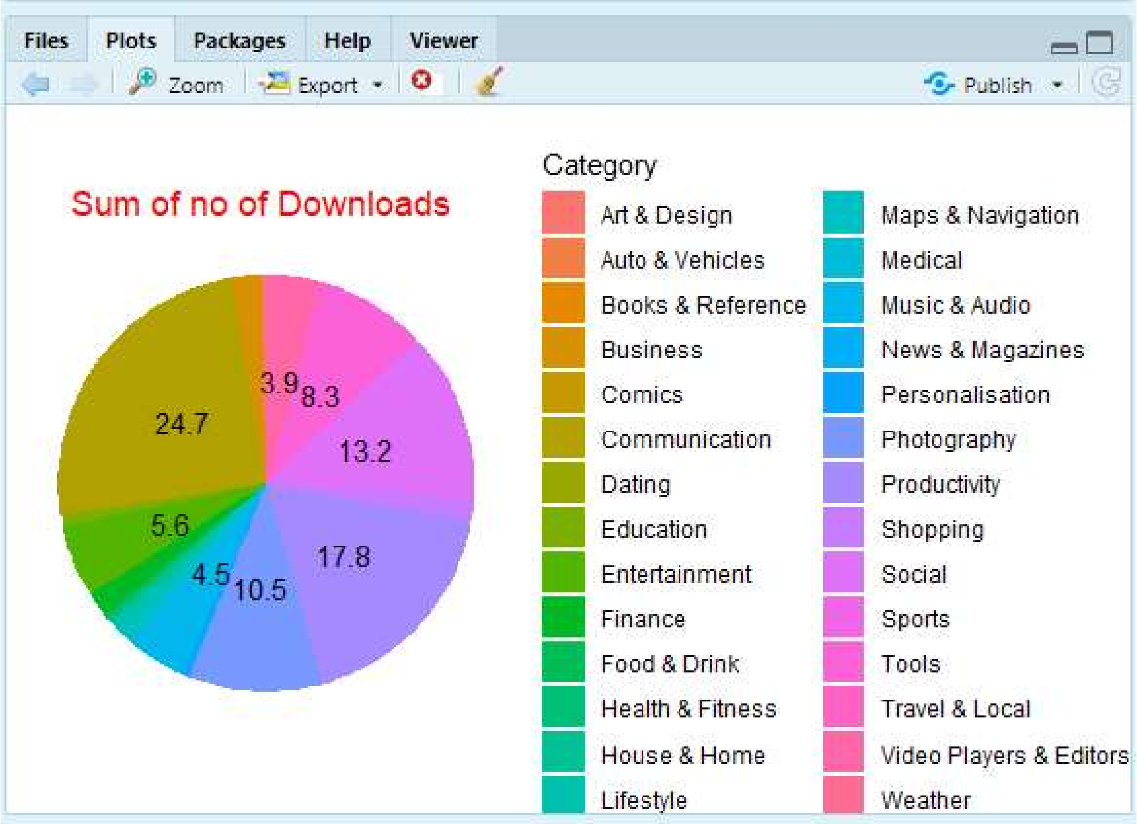
```
> df_cat<-as.data.frame(df %>% group_by(Category) %>%
+                         summarize(No_downloads_th = sum(No_downloads)/1000))
>
>
> df_order<-df_cat[order(df_cat$No_downloads_th, decreasing = TRUE), ]
> df_order
                Category No_downloads_th
6          Communication        17040200
21          Productivity        12250000
23                Social         9094700
20           Photography         7206000
25                 Tools         5690900
9          Entertainment         3835200
17         Music & Audio         3096000
27 Video Players & Editors       2682201
4               Business         1489500
10               Finance         1325750
22              Shopping         1134000
14             Lifestyle          793500
15      Maps & Navigation         635000
8              Education          599010
26         Travel & Local         420000
12       Health & Fitness         337000
19        Personalisation         266000
1            Art & Design         250000
11           Food & Drink         225000
18      News & Magazines          170000
24                 Sports         141200
3      Books & Reference          121000
28                Weather          48000
16                Medical          37000
13           House & Home          20000
7                 Dating           12000
2        Auto & Vehicles           10000
5                 Comics            5000
```

The maximum value of sum of Downloads is in **Communication** Category

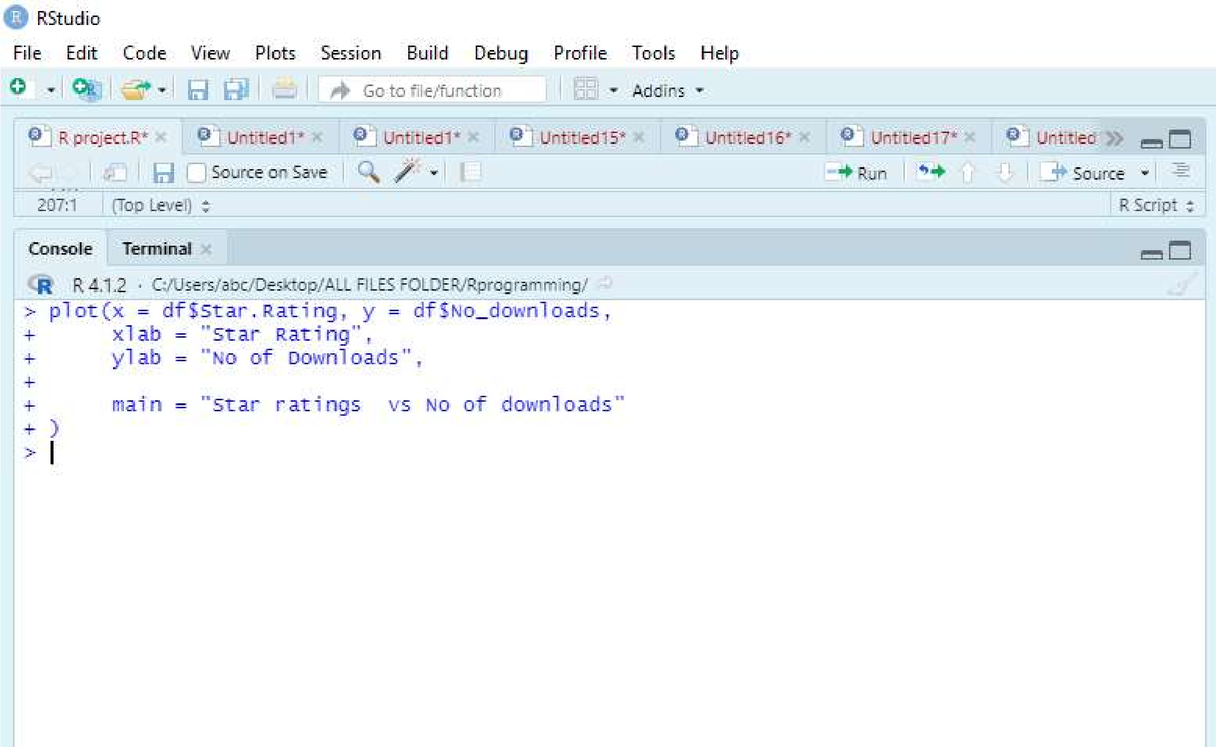
The above data can be visualized in a pie chart as follows


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R*
df
Untitled1*
Untitled1*
Untitled15*
Unti
Source on Save
Run
Source
212:37
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> piepercent<- round(100 * df_cat$No_downloads_th / sum(df_cat$No_downloads_th), 1)
>
> pieless3<-ifelse(piepercent <3 ,'',as.character(piepercent))
>
>
> ggplot(df_cat, aes(x="", y=No_downloads_th, fill=Category)) +geom_bar(width = 0.5, stat = "i
dentity") +
+   coord_polar("y", start=0) +theme_void()  +
+ geom_text(aes(label = paste(pieless3)), position = position_stack(vjust = 0.5)) +
+           theme(axis.line = element_blank(),
+                 axis.text = element_blank(),
+                 axis.ticks = element_blank(),
+                 plot.title = element_text(hjust = 0.5, color = "#FF0000"))+
+   ggtitle("Sum of no of Downloads ")
> |
```

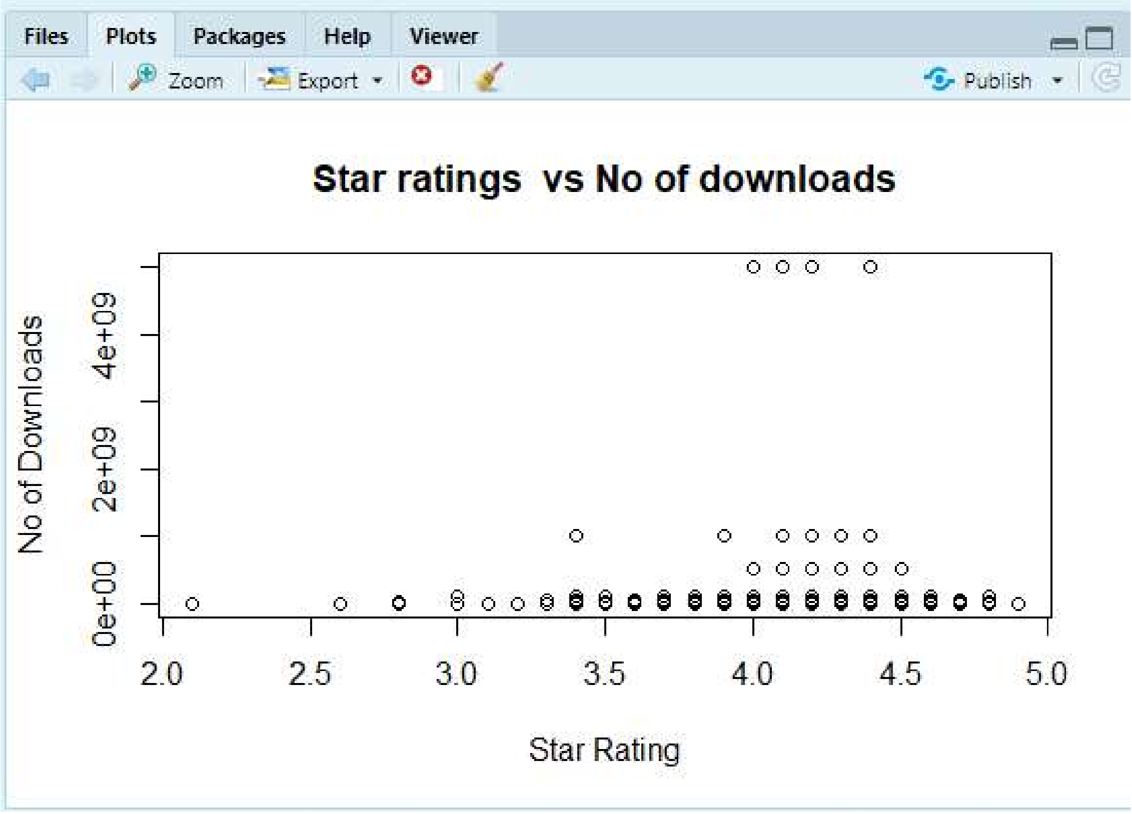


Scatterplot of **Star.Rating** vs **No_downloads**



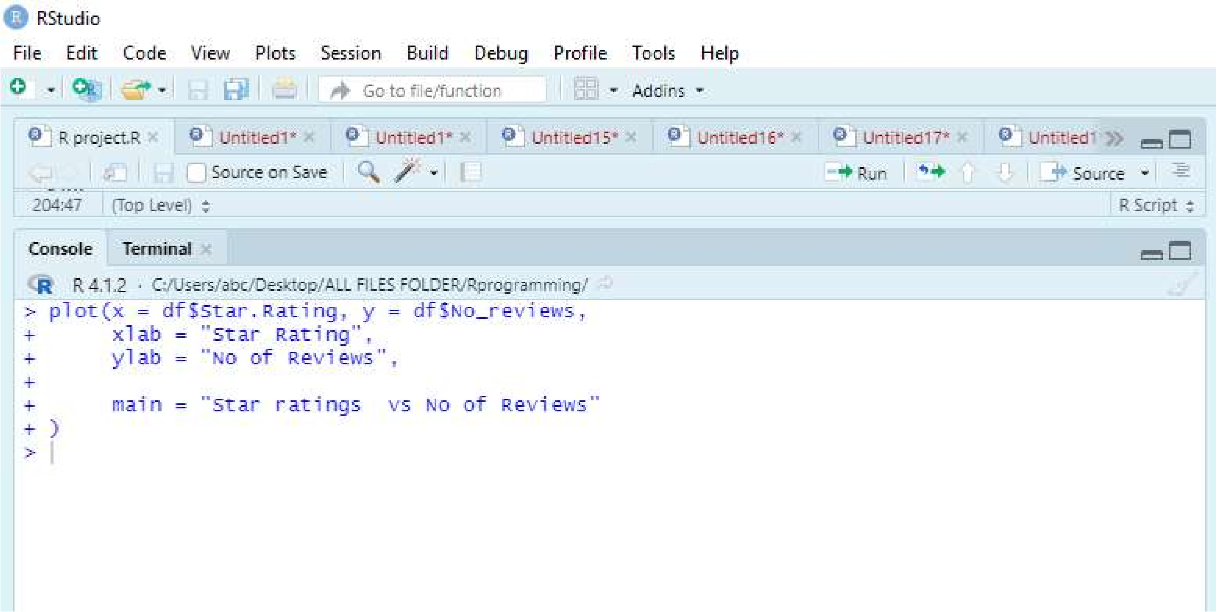
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R* Untitled1* Untitled1* Untitled15* Untitled16* Untitled17* Untitled
Source on Save
Run
Source
207:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> plot(x = df$Star.Rating, y = df$No_downloads,
+      xlab = "Star Rating",
+      ylab = "No of Downloads",
+
+      main = "Star ratings  vs No of downloads"
+ )
> 
```

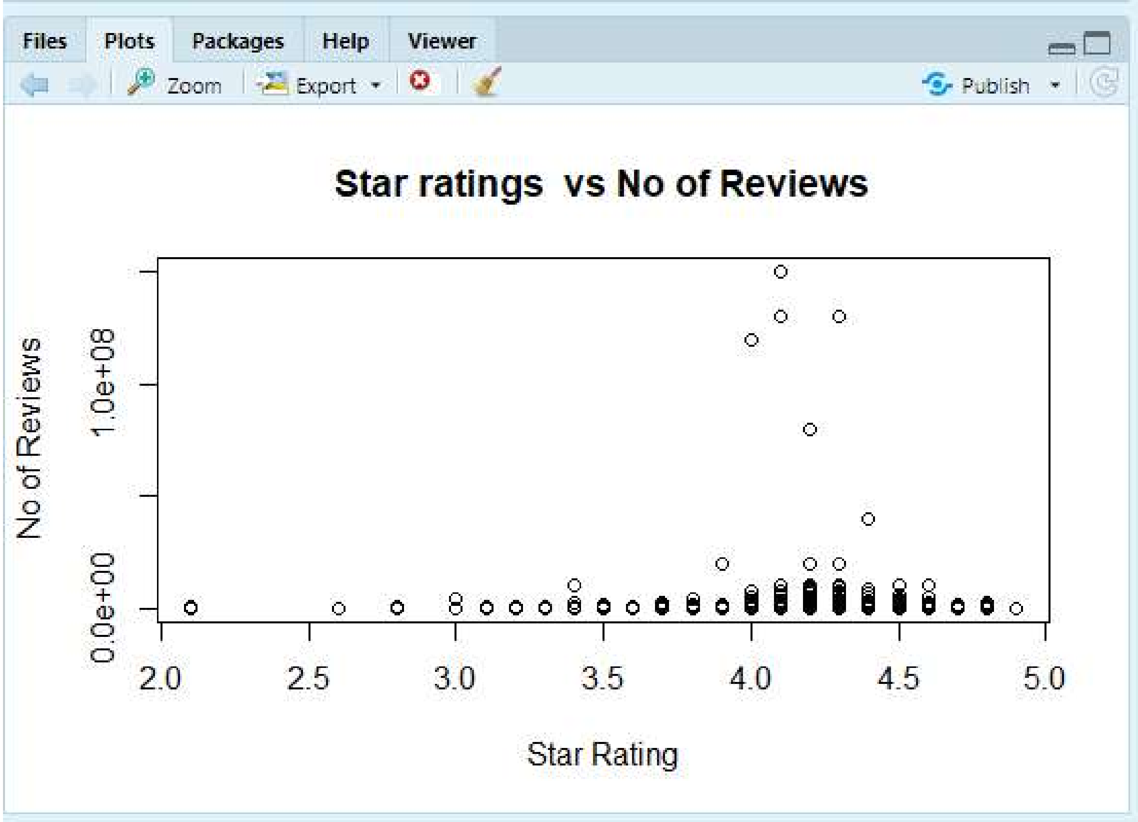


The no of Downloads is **highest** for rating between **4.0 to 4.5**

Scatterplot of **Star.Rating** vs **No Reviews**

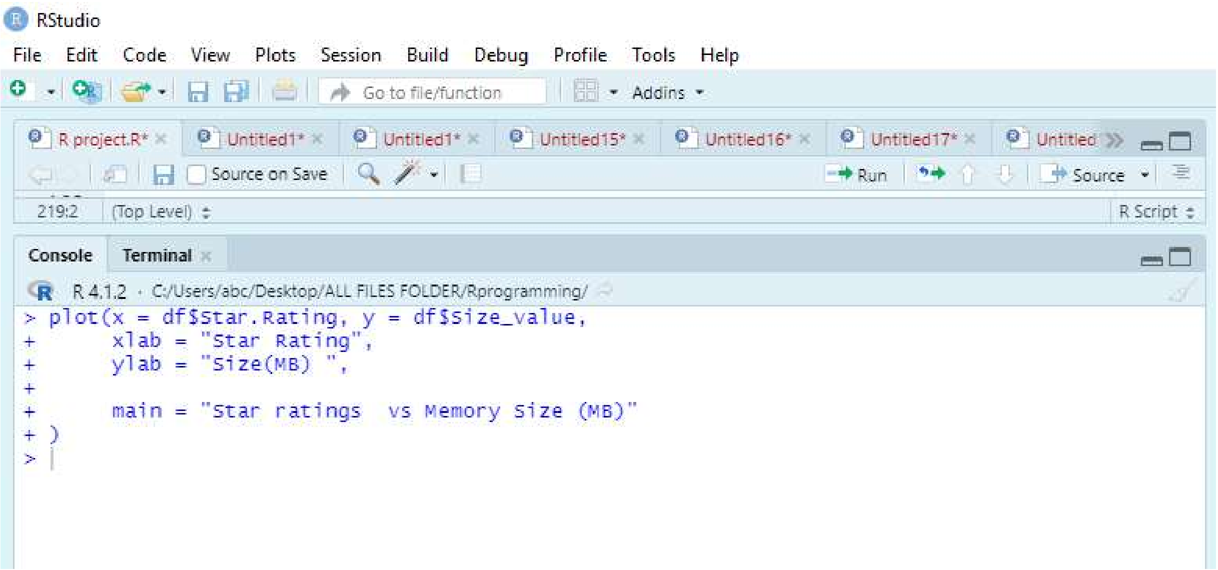


```
> plot(x = df$Star.Rating, y = df$No_reviews,
+      xlab = "Star Rating",
+      ylab = "No of Reviews",
+
+      main = "Star ratings  vs No of Reviews"
+ )
> 
```

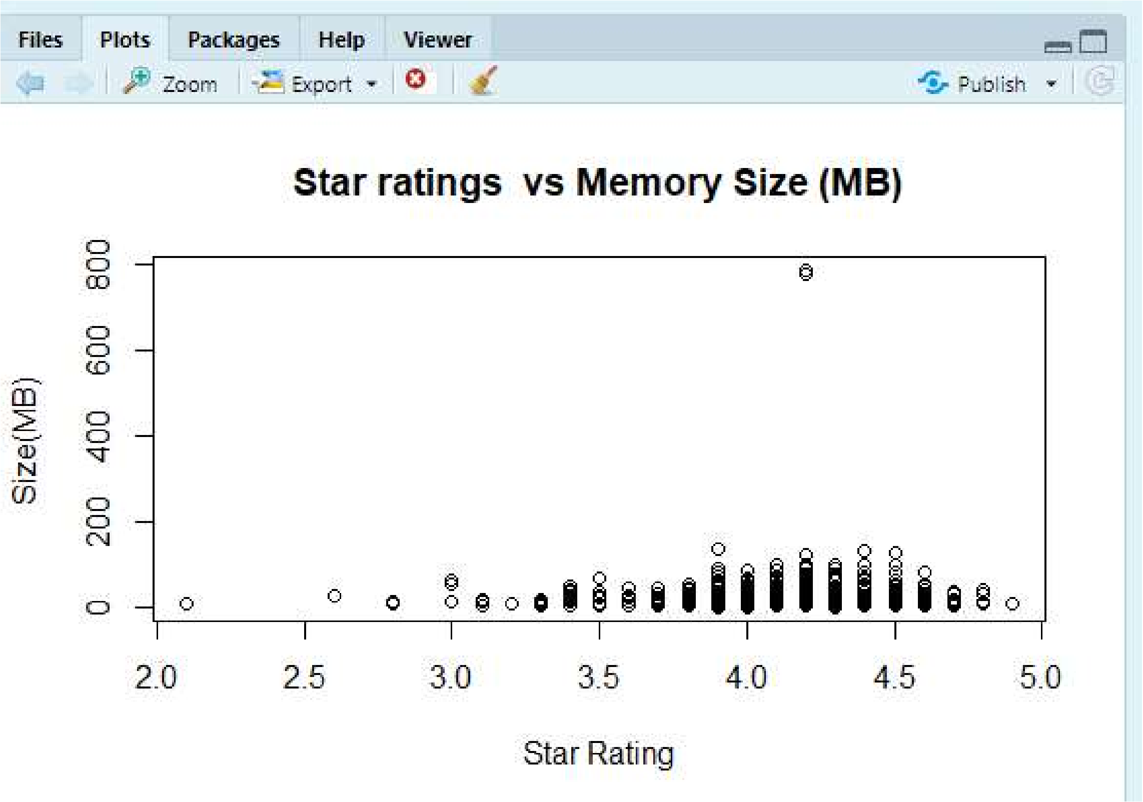


The no of Reviews is **highest** for rating between **4.0 to 4.5**

Scatterplot of **Star.Rating** vs **Size_value**

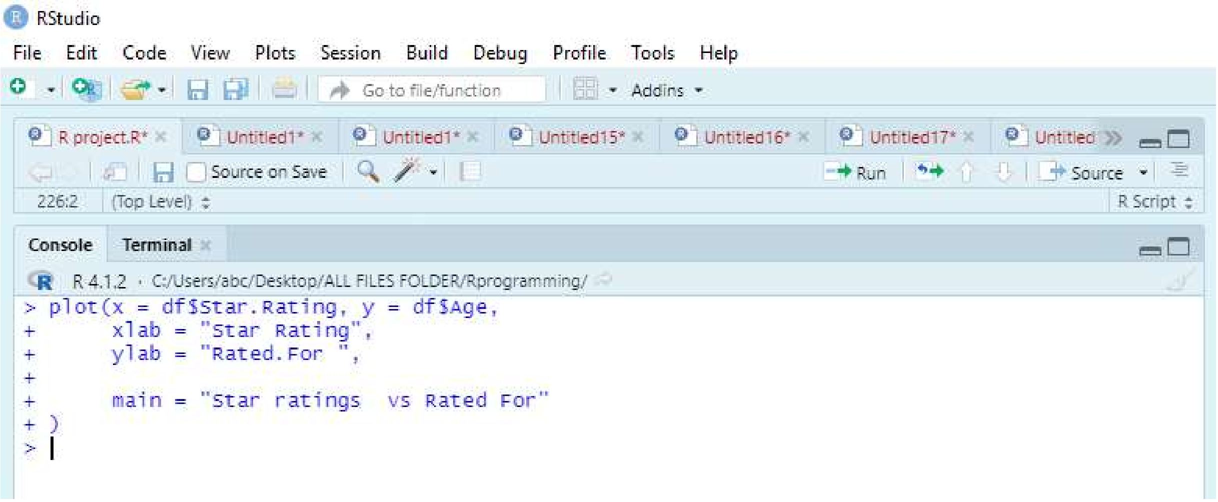


```
> plot(x = df$Star.Rating, y = df$Size_value,
+      xlab = "Star Rating",
+      ylab = "Size(MB) ",
+
+      main = "Star ratings  vs Memory Size (MB)"
+ )
> 
```

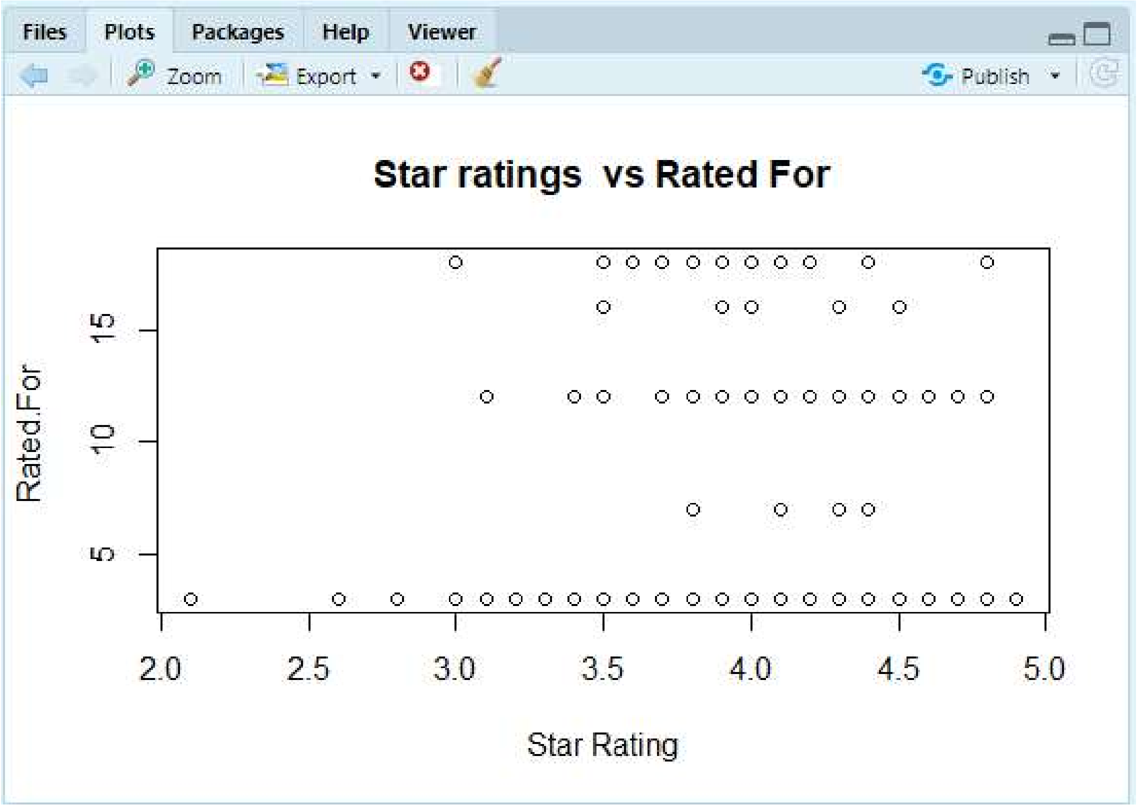


The Star. Rating is distributed across the band of the Memory size (0-100 MB)

Scatterplot of **Star.Rating** vs **Rated.For**


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R* Untitled1* Untitled1* Untitled15* Untitled16* Untitled17* Untitled
Source on Save
Run
Source
226:2
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

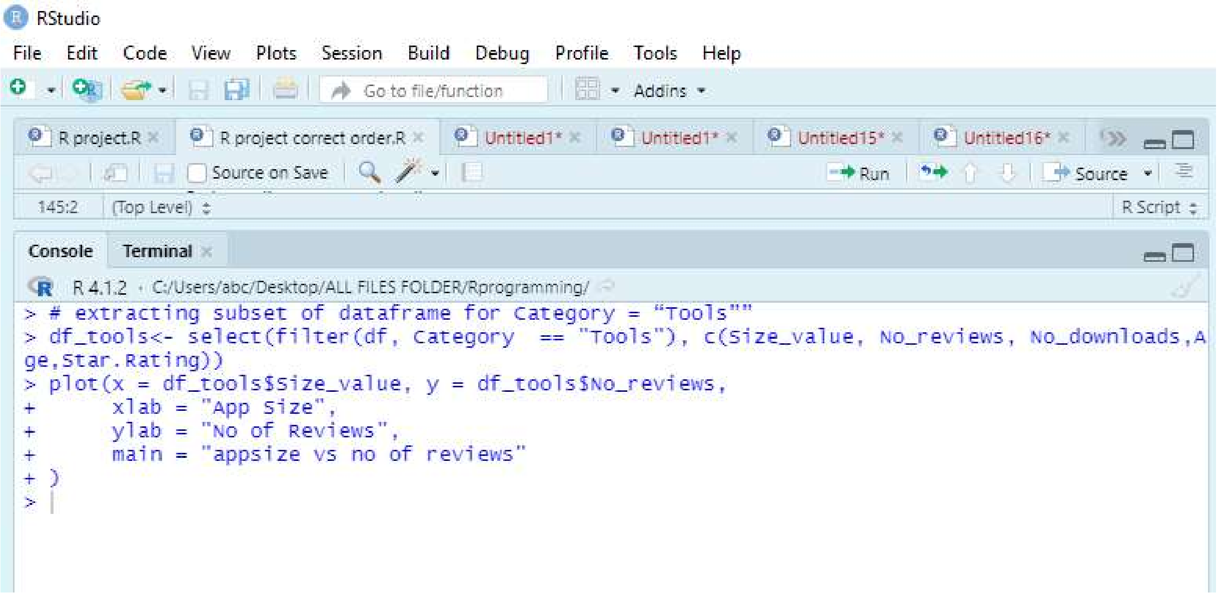
```
> plot(x = df$Star.Rating, y = df$Age,
+      xlab = "Star Rating",
+      ylab = "Rated.For ",
+
+      main = "Star ratings  vs Rated For"
+ )
> |
```



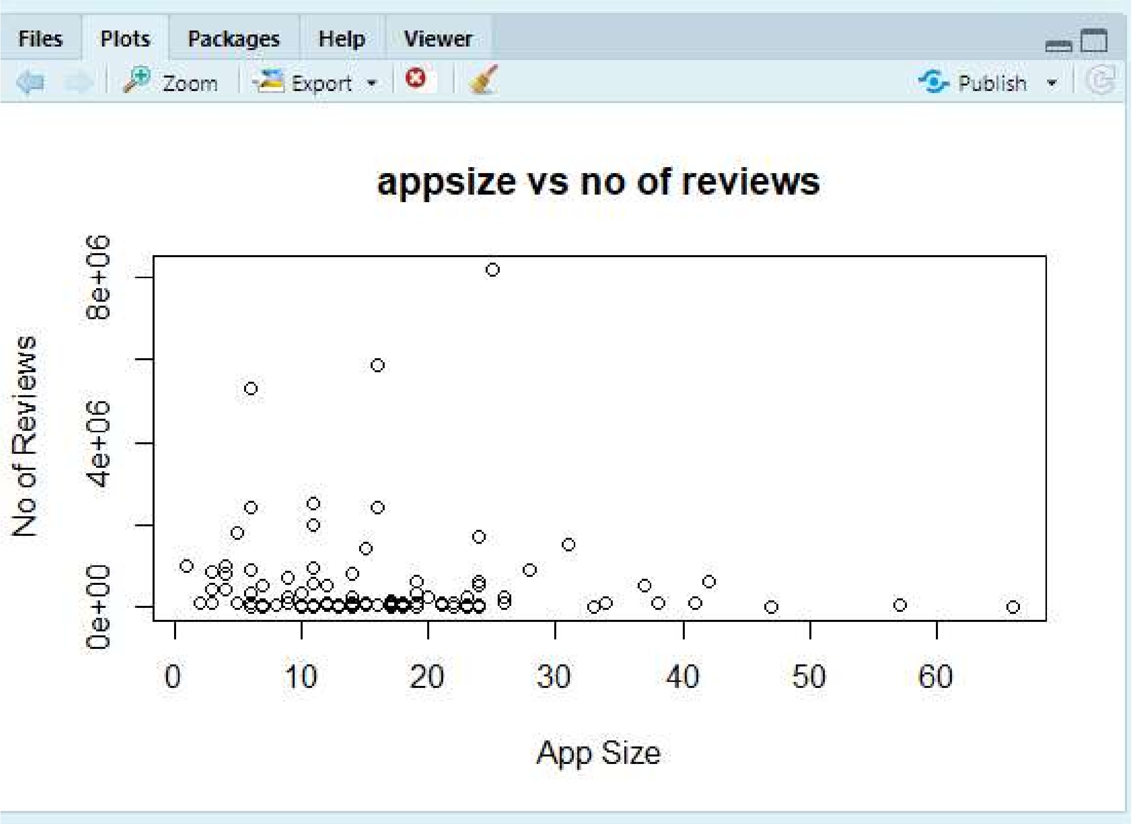
The Star Rating is distributed across all ages

From the scatter plot distribution, a well-defined relationship between any of the variables namely Size_value, Star. Rating, No_downloads, No_reviews and Rated. For cannot be established.

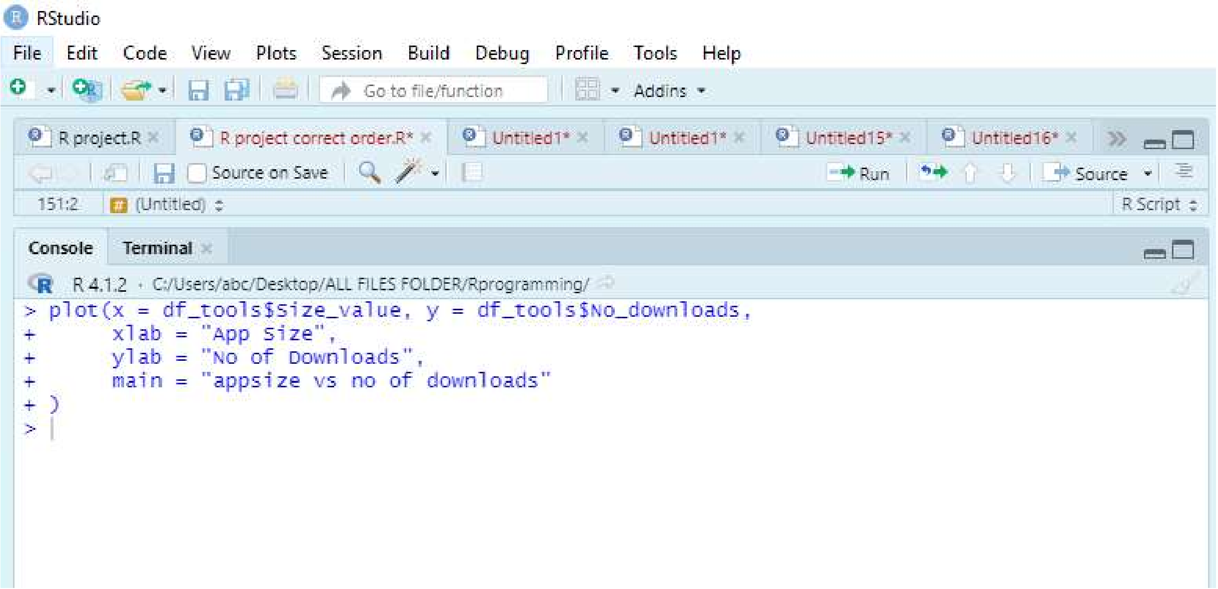
This may be because the apps belong to different categories. Since Category **Tools** have maximum no of apps the subset of dataset with Category**Tools** is extracted in data frame df_tools and data frame df_tools is used for further analysis.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
145:2
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

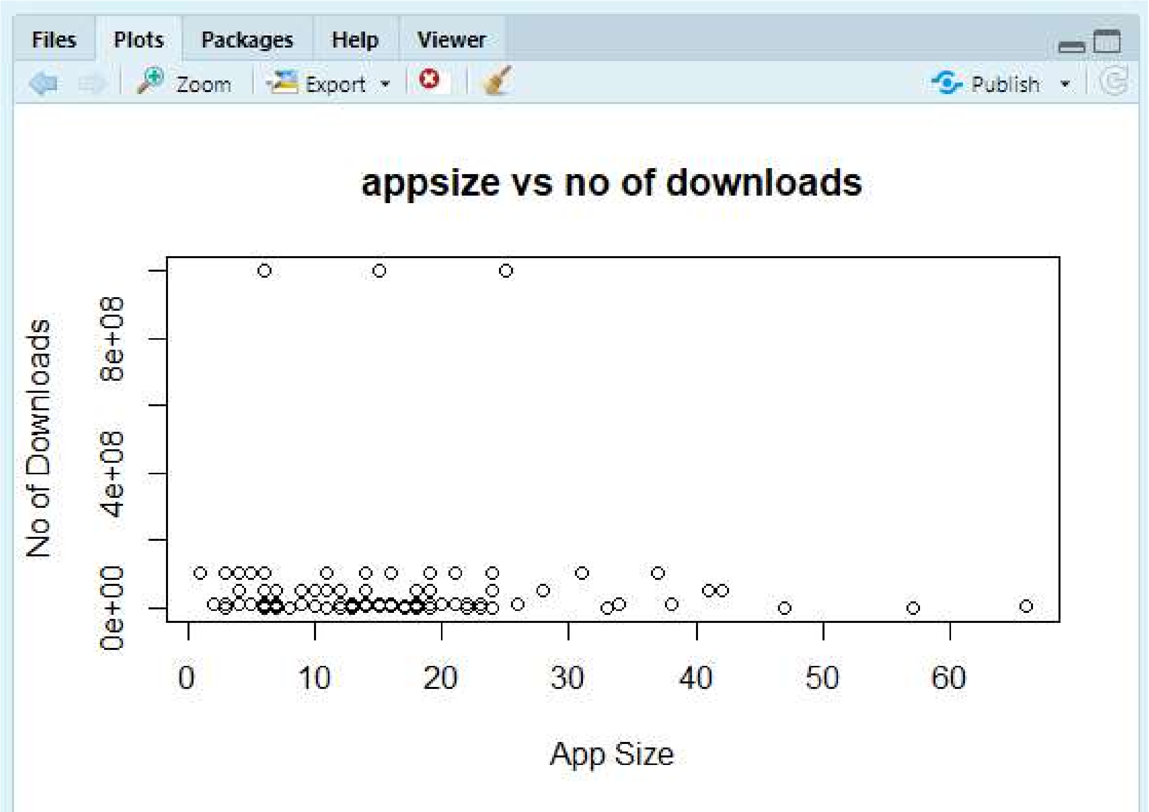
```
> # extracting subset of dataframe for Category = "Tools""
> df_tools<- select(filter(df, Category  == "Tools"), c(Size_value, No_reviews, No_downloads,A
ge,Star.Rating))
> plot(x = df_tools$Size_value, y = df_tools$No_reviews,
+      xlab = "App Size",
+      ylab = "No of Reviews",
+      main = "appsize vs no of reviews"
+ )
> |
```



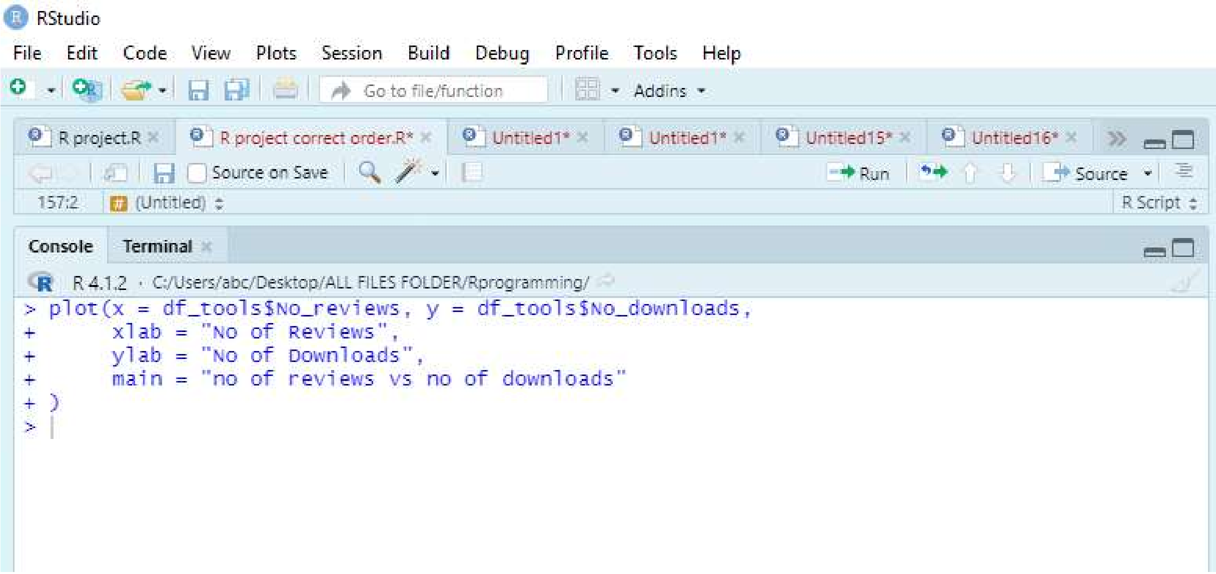
The values are concentrated in the region of **App size (0-30 MB)** and **No of Reviews (0-20 Lac)**


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R*
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
151:2
(Untitled)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> plot(x = df_tools$size_value, y = df_tools$No_downloads,
+      xlab = "App Size",
+      ylab = "No of Downloads",
+      main = "appsize vs no of downloads"
+ )
> 
```

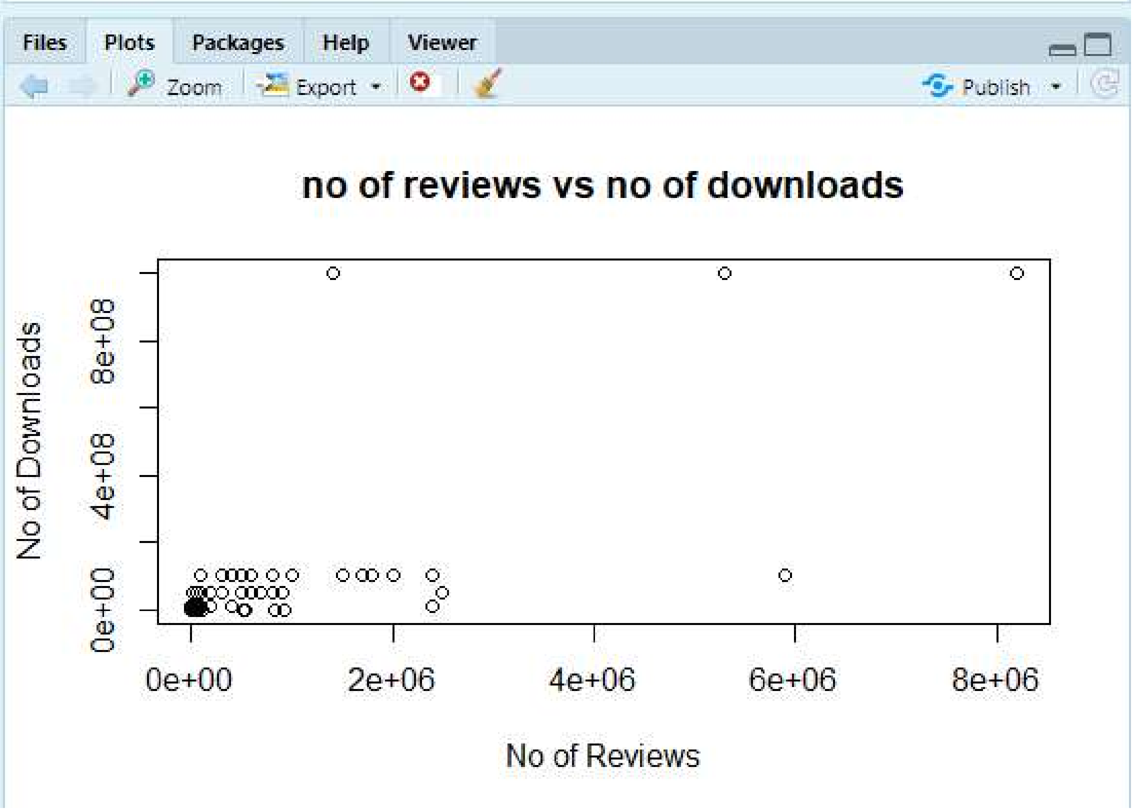


Maximum downloads are of App Size (0-30 MB)



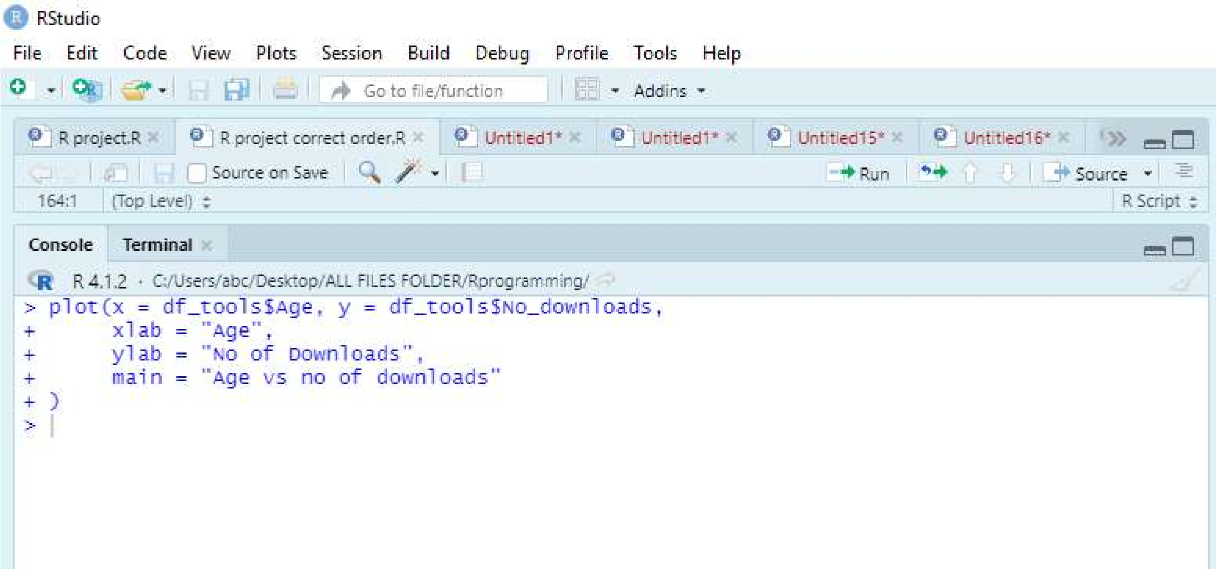
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R*
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
157:2
(Untitled)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> plot(x = df_tools$No_reviews, y = df_tools$No_downloads,
+      xlab = "No of Reviews",
+      ylab = "No of Downloads",
+      main = "no of reviews vs no of downloads"
+ )
> 
```



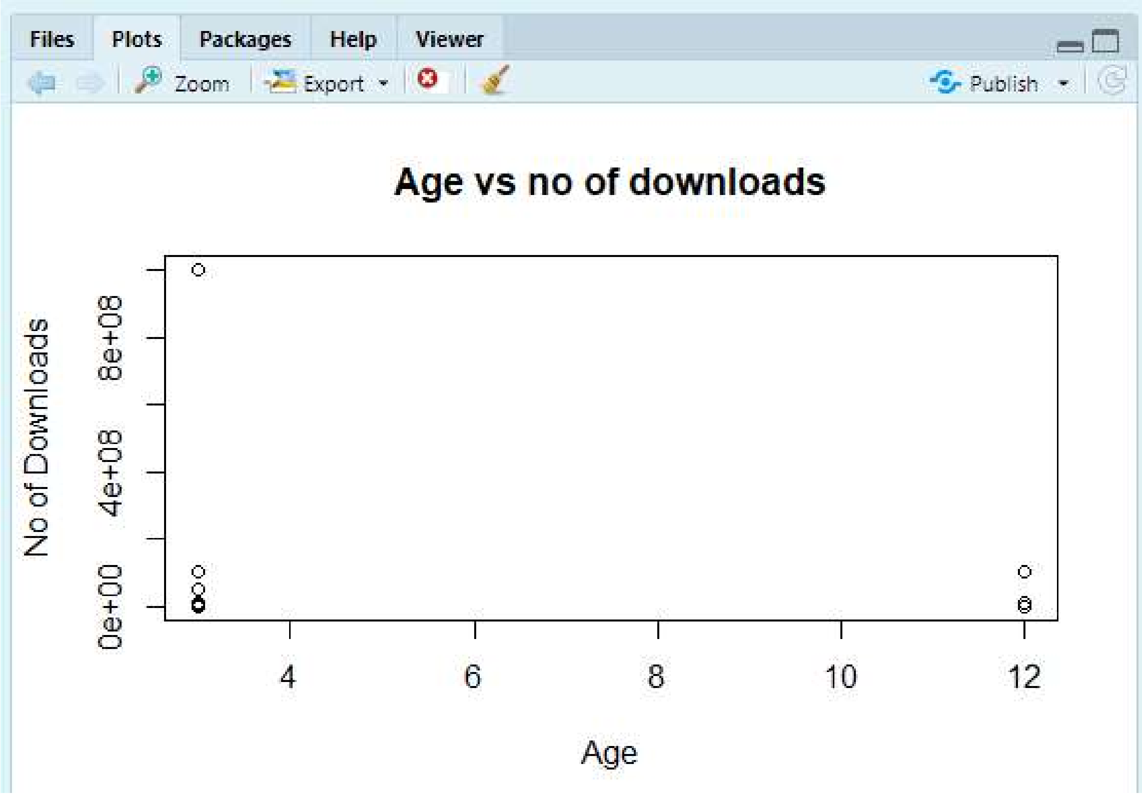
The values are concentrated in the region of **No_reviews (0-20 Lac)** and

No_downloads (0-20 Crores)



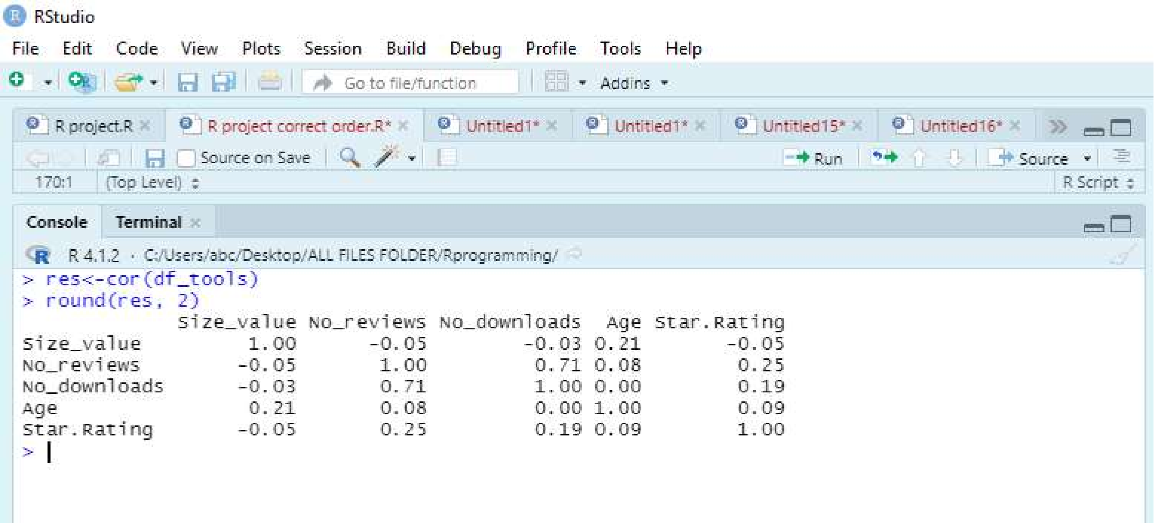
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
164:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> plot(x = df_tools$Age, y = df_tools$No_downloads,
+      xlab = "Age",
+      ylab = "No of Downloads",
+      main = "Age vs no of downloads"
+ )
> |
```



The downloads are **only in 3 + and 12 + Rated.for**

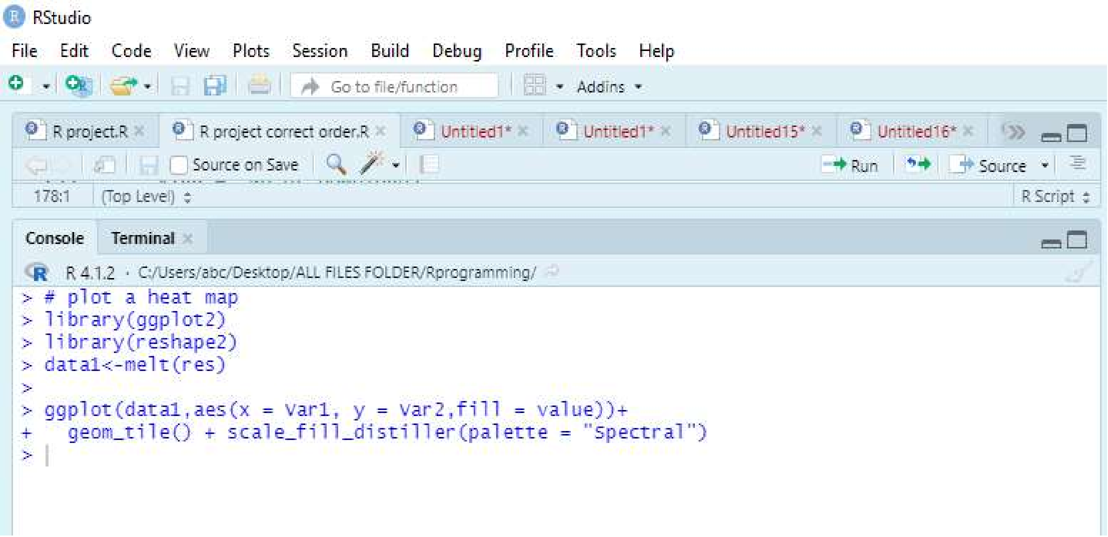
Correlation matrix



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R*
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
170:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

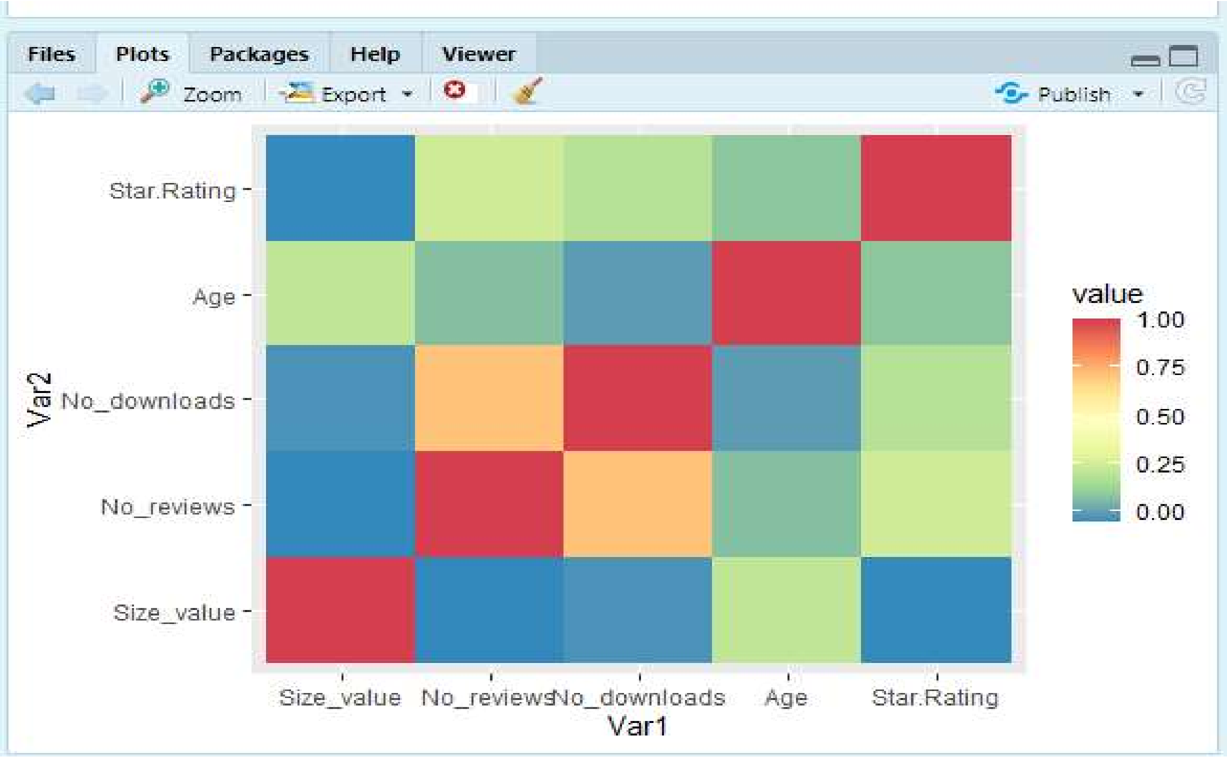
```
> res<-cor(df_tools)
> round(res, 2)
              Size_value No_reviews No_downloads  Age Star.Rating
Size_value          1.00      -0.05        -0.03 0.21       -0.05
No_reviews         -0.05       1.00         0.71 0.08        0.25
No_downloads       -0.03       0.71         1.00 0.00        0.19
Age                 0.21       0.08         0.00 1.00        0.09
Star.Rating        -0.05       0.25         0.19 0.09        1.00
> 
```

Heat Map



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R
Untitled1*
Untitled1*
Untitled15*
Untitled16*
Source on Save
Run
Source
178:1
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> # plot a heat map
> library(ggplot2)
> library(reshape2)
> data1<-melt(res)
> 
> ggplot(data1,aes(x = Var1, y = Var2,fill = value))+
+   geom_tile() + scale_fill_distiller(palette = "Spectral")
> 
```



High Positive correlation is between **No_reviews and No_downloads**

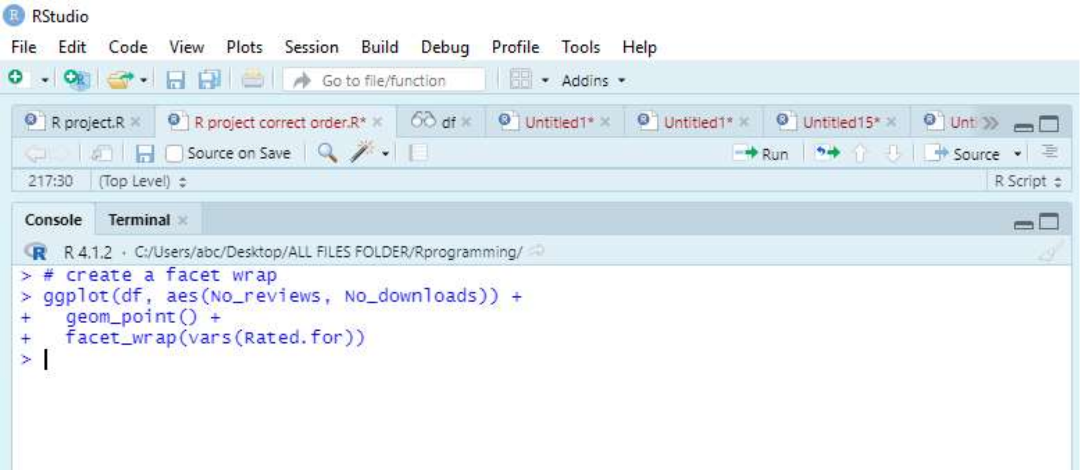
Low Positive correlation is between **Star.Rating** and **No_reviews , No_downloads**

Low Positive correlation is between **Size_value** and **Age**

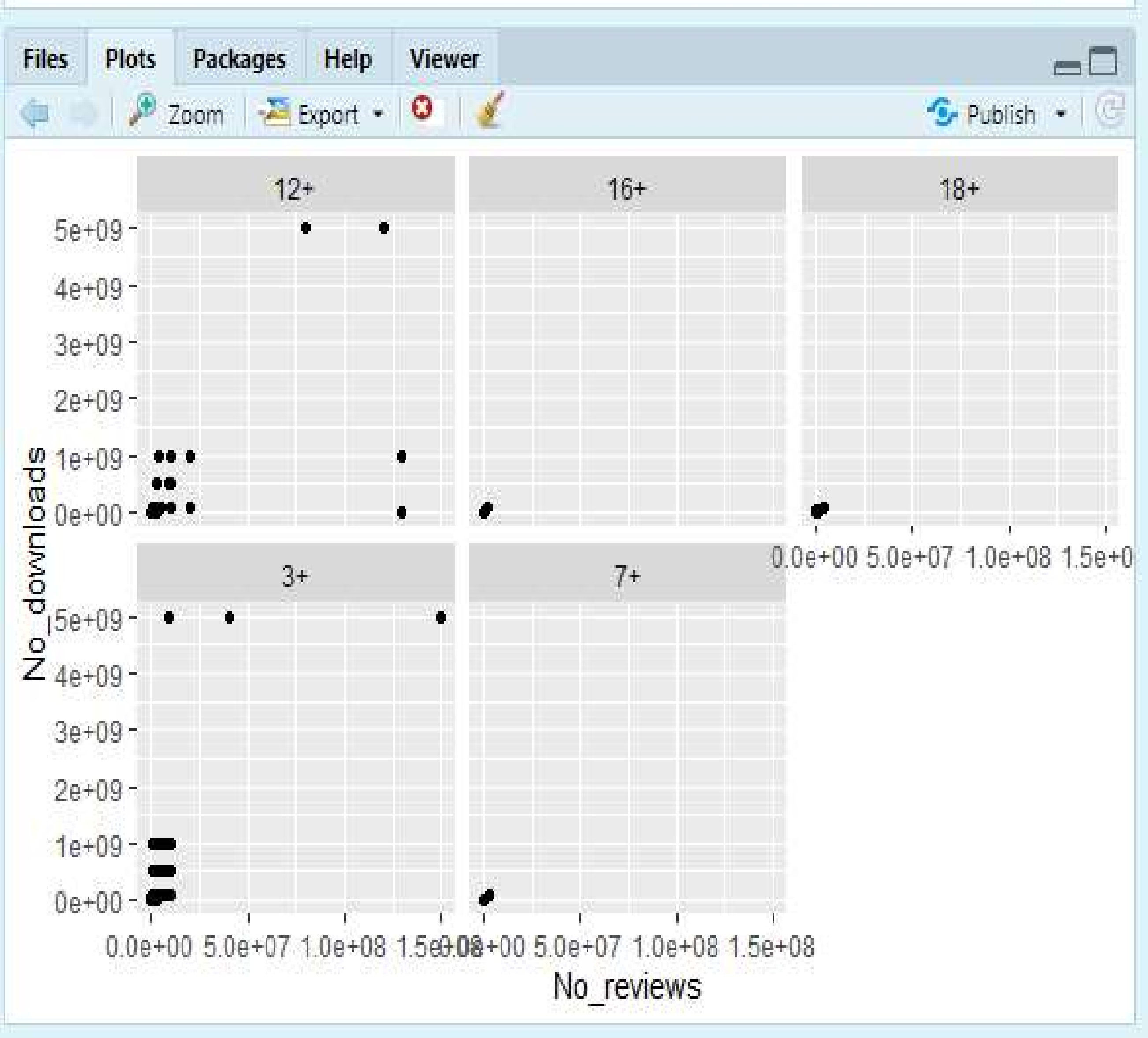
Low Negative correlation between **Size_value** and **No_reviews, No_downloads, Star.Rating**

Facet wrap

Facet wrap is used to find relationships between No_Reviews and No_downloads for categorical variable Rated.for


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R project.R
R project correct order.R*
df
Untitled1*
Untitled1*
Untitled15*
Unti
Source on Save
Run
Source
217:30
(Top Level)
R Script
Console
Terminal
R 4.1.2 · C:/Users/abc/Desktop/ALL FILES FOLDER/Rprogramming/

```
> # create a facet wrap
> ggplot(df, aes(No_reviews, No_downloads)) +
+   geom_point() +
+   facet_wrap(vars(Rated.for))
> |
```



Maximum No_reviews and No_downloads is in 3+ and 12+ Rated. For