Team ACE

Google Play Store

Analysis on Google Play store Apps

Table of Contents

INTRODUCTION

In an emerging world of technology, mobiles play a vital role. Every mobile user, be it iOS or Android have developed applications not only for entertainment but also to improve lifestyle. Having such a wide range of applications and users we decided to analyze google play store dataset. Our research question is to identify the factors that affect the ratings of the applications and the measures needed to improve the same.

DATA

DATA SOURCE

The data used in our analysis is a preloaded raw data from 'www.kaggle.com' named 'Google Play store Apps'. The dataset has 13 variables and 10841 observations.

Following are the variables in in our dataset:

App: [character] Application name

Category: [character] Category the app belongs to.

Rating: [numeric] Overall user rating of the app (as when scraped)

Reviews: [numeric] Number of user reviews for the app (as when scraped)

Size: [numeric] Size of the app (as when scraped)

Installs: [numeric] Number of user downloads/installs for the app (as when scraped)

Type: [character] Paid or Free

Price: [numeric] Price of the app (as when scraped)

Content Rating: [character] Age group the app is targeted at - Children / Mature 21+ / Adult

Genres: [character] An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.

Last Updated: [date] Date when the app was last updated on Play Store (as when scraped)

Current Ver: [character] Current version of the app available on Play Store (as when scraped)

Android Ver: [character] Min required Android version (as when scraped)

DATA CLEANING

To perform analysis, we had to change the data types of few variables which include Price, Installs, Size, Last-Updated. By doing so, we converted the raw dataset into a technically consistent data.

EXPLORATORY DATA ANALYSIS

1. Summary of 'googleplaystore'

Following is the summary of the dataset after the process of cleaning

```
                                          App             Category      Rating         Reviews                  Size           Installs     Type
ROBLOX                                    :   9   FAMILY      :1972   Min.   :1.000   Min.   :      0   Varies with device:1695   1,000,000+ :1579   Free:1003
CBS Sports App - Scores, News, Stats & Watch Live:   8   GAME        :1144   1st Qu.:4.000   1st Qu.:     38   11M               : 198   10,000,000+:1252   NaN :
8 Ball Pool                               :   7   TOOLS       : 843   Median :4.300   Median :   2094   12M               : 196   100,000+   :1169   Paid:  80
Candy Crush Saga                          :   7   MEDICAL     : 463   Mean   :4.192   Mean   : 444153   14M               : 194   10,000+    :1054
Duolingo: Learn Languages Free            :   7   BUSINESS    : 460   3rd Qu.:4.500   3rd Qu.:  54776   13M               : 191   1,000+     : 907
ESPN                                      :   7   PRODUCTIVITY: 424   Max.   :5.000   Max.   :78158306   15M               : 184   5,000,000+ : 752
(Other)                                   :10795   (Other)     :5534   NA's   :1474                      (Other)           :8182   (Other)    :4127
     Price                Content.Rating          Genres        Last.Updated            Current.Ver              Android.Ver
Min.   :  0.000   Adults only 18+:   3   Tools          : 842   3-Aug-18 : 326   Varies with device:1459   4.1 and up        :2451
1st Qu.:  0.000   Everyone       :8714   Entertainment: 623   2-Aug-18 : 304   1                 : 842   4.0.3 and up      :1501
Median :  0.000   Everyone 10+   : 414   Education    : 549   31-Jul-18: 294   1.1               : 276   4.0 and up        :1375
Mean   :  1.027   Mature 17+     : 499   Medical      : 463   1-Aug-18 : 285   1.2               : 185   Varies with device:1362
3rd Qu.:  0.000   Teen           :1208   Business     : 460   30-Jul-18: 211   2                 : 165   4.4 and up        : 980
Max.   :400.000   Unrated        :   2   Productivity : 424   25-Jul-18: 164   1.3               : 145   2.3 and up        : 652
                                         (Other)      :7479   (Other)  :9256   (Other)           :7768   (Other)           :2519
```

2. Descriptive Analysis on Price

To analyze the effect of the price variable, we performed a descriptive analysis on 'Price'. To do so, our actual data had a character '$' before every value. Therefore, to convert the data type into 'numerical', we used the function 'gsub' to eliminate the '$' symbol.

After cleaning the data, we performed a summary operation and also found the summary of prices based on the 'Category' of the Apps.

Category <fctr>	Price <dbl>
BEAUTY	0.00000000
COMICS	0.00000000
HOUSE_AND_HOME	0.00000000
LIBRARIES_AND_DEMO	0.01164706
NEWS_AND_MAGAZINES	0.01406360
SHOPPING	0.02107692
ENTERTAINMENT	0.05355705
SOCIAL	0.05413559
VIDEO_PLAYERS	0.05977143
FOOD_AND_DRINK	0.06677165
ART_AND_DESIGN	0.09184615
EDUCATION	0.11512821
DATING	0.13431624
AUTO_AND_VEHICLES	0.15847059
PARENTING	0.15966667
TRAVEL_AND_LOCAL	0.19360465
MAPS_AND_NAVIGATION	0.19671533
HEALTH_AND_FITNESS	0.19747801
COMMUNICATION	0.21483204
GAME	0.25113636
SPORTS	0.26041667
TOOLS	0.31702254
PERSONALIZATION	0.39275510
WEATHER	0.39536585
PHOTOGRAPHY	0.40062687
BUSINESS	0.40276087
BOOKS_AND_REFERENCE	0.51848485
PRODUCTIVITY	0.59181604
FAMILY	1.23467546
EVENTS	1.71859375
MEDICAL	3.11006479
LIFESTYLE	6.18028796
FINANCE	7.92576503

From the above results, we can say that applications under the 'Beauty' category is Free of cost whereas applications under 'Finance' category are the most expensive application in google play store.

3. Descriptive analysis on Category vs Reviews

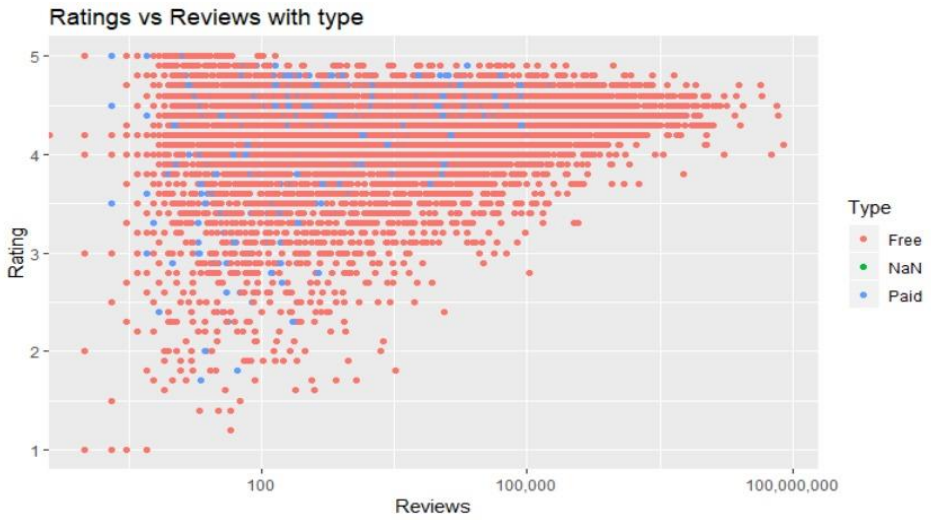
To find the category of app with the highest and the lowest rating, we performed a descriptive analysis on these variables.

Category <fctr>	Reviews <dbl>
EVENTS	2515.906
MEDICAL	3425.432
BEAUTY	7476.226
LIBRARIES_AND_DEMO	12201.388
AUTO_AND_VEHICLES	13690.188
PARENTING	15972.183
ART_AND_DESIGN	26376.000
BUSINESS	30335.983
DATING	31159.308
LIFESTYLE	33724.565
HOUSE_AND_HOME	45186.193
FINANCE	47952.809
COMICS	56387.933
FOOD_AND_DRINK	69947.480
BOOKS_AND_REFERENCE	95060.905
HEALTH_AND_FITNESS	111125.346
WEATHER	178106.524
SPORTS	184453.565
NEWS_AND_MAGAZINES	192229.198
FAMILY	208025.522
MAPS_AND_NAVIGATION	223790.175
PERSONALIZATION	227923.827
TRAVEL_AND_LOCAL	242705.112
EDUCATION	253819.141
PRODUCTIVITY	269143.809
TOOLS	324062.923
ENTERTAINMENT	397168.819
SHOPPING	442466.238
VIDEO_PLAYERS	630743.931
PHOTOGRAPHY	637363.134
GAME	1385858.697
SOCIAL	2105903.125
COMMUNICATION	2107137.623

From the results, we see that application under 'Communication' category are the most reviewed with an average of 2107137.623 reviews.

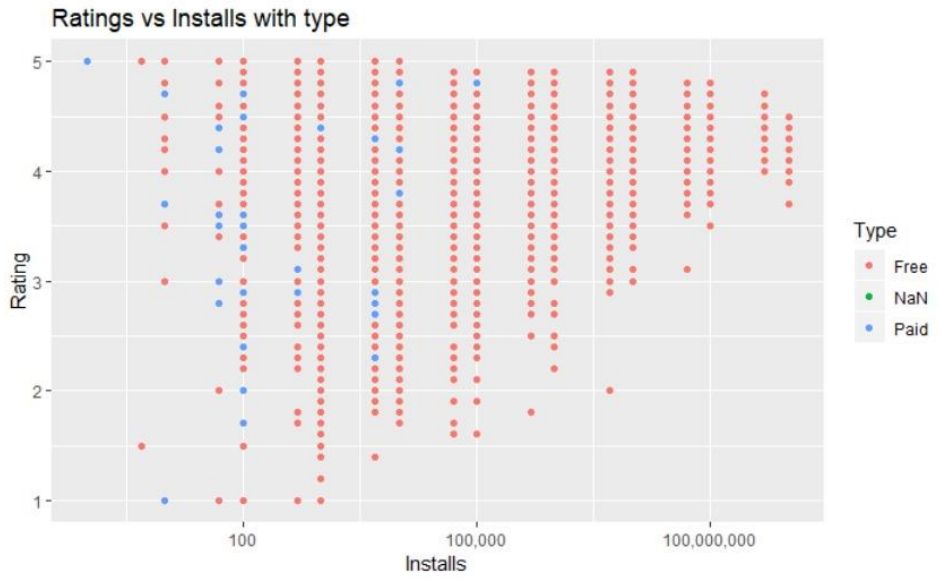
4. Application Ratings vs Reviews w.r.t Type

By creating a scatter plot, we can see that most of the applications that have received rating is under the 'Free' category, there are also few paid applications which are seen with the good rating

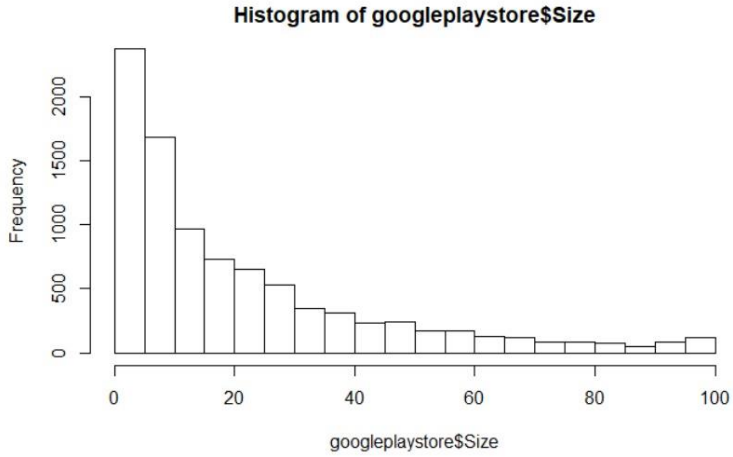


5. Application Ratings vs Installs with respect to type

Here, from the graph below we can state that applications which have 100,000+ installs are under the 'Free' category, there are also few paid applications which has less than 10,000 downloads

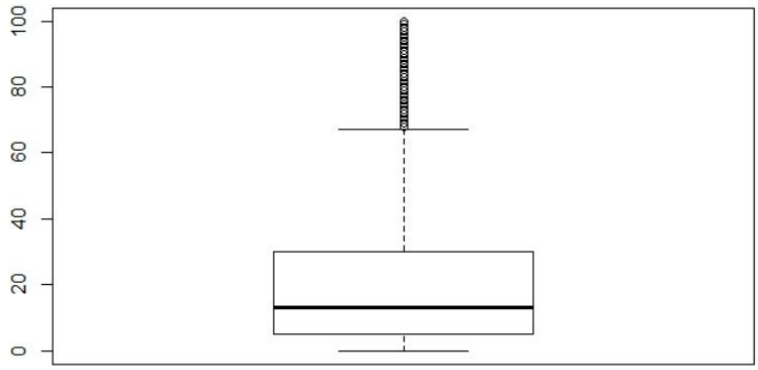


6. Histogram of Size variable
   The below histogram shows the skewness in size along with the frequency and it rightly skewed



7. Boxplot of Size variable

The below figure shows the box plot which has interquartile range of the size variable, we can also see the outliers



CONCLUSION

In the exploratory analysis, we have analyzed the relationship between different variables and their inter-relation. This would further improve our Prediction analysis to build a precise prediction modal. In our prediction model, we will be predicting the 'Ratings' for the applications based on the various factors affecting 'Ratings'. To understand this further, it is essential to perform exploratory analysis to identify the various variables affecting 'Ratings' variables.