Contents

1 Introduction to the dataset

1.1 Data source

The data source used for this analysis is the *2018 google play store*(https://www.kaggle.com/lava18/google-play-store-apps) collected from Kaggle.

1.2 Description of the dataset

The dataset is a collection of web-scraped data of 10,000 apps from Google Play Store. Google Play Store originally referred as the Android Market, is Google's official store and portal for Android apps, games and other content for Android-powered phone, tablet or Android TV device. As of May 2017, it has over two billion monthly active users, the largest installed base of any operating system, and as of January 2020, the Google Play Store features over 2.9 million apps[13].

The variables of the dataset are as follows:

1) App (Name) – Name/Title of the application
2) Category (App)- Category/Domain to which the app belongs to
3) Rating (App)- Overall user rating of the app
4) Reviews (User)- Number of user reviews for the app
5) Size (App)- Space or memory that the app takes up

6) Installs (App)- Number of user downloads/installs
7) Type (Free/Paid)-Apps may be free or paid depending on the developer's choice
8) Price (App)-Price of the app if not free
9) Content Rating - Age group the app is based off at - Children / Mature 21+ / Adult
10) Genres (Detailed Category)- An app can belong to multiple genres, For eg, a musical family game will belong to Music, Game, Family genres.
11) Last Updated (App)- Date when the app was last updated on Play Store
12) Current Version (App)- Current version of the app available on Play Store

13) Android Version (Support) – minimum version of android it takes to have the app on the device

2 Purpose of the project

- The aim of our project is to find out if we can predict ratings of an app based on different variables and we intend to summarise the different factors that influence the success of an app.These analysis might also help the developer community to build more successful apps by taking accurate data-based decisions, and focusing on those aspects of applications that matters most.
- Also,since this is the first time we are doing data analysis using R, it is a fun way to learn and to strengthen the concepts learned during the course by taking a hands-on approach.

3 Intended audience of the project

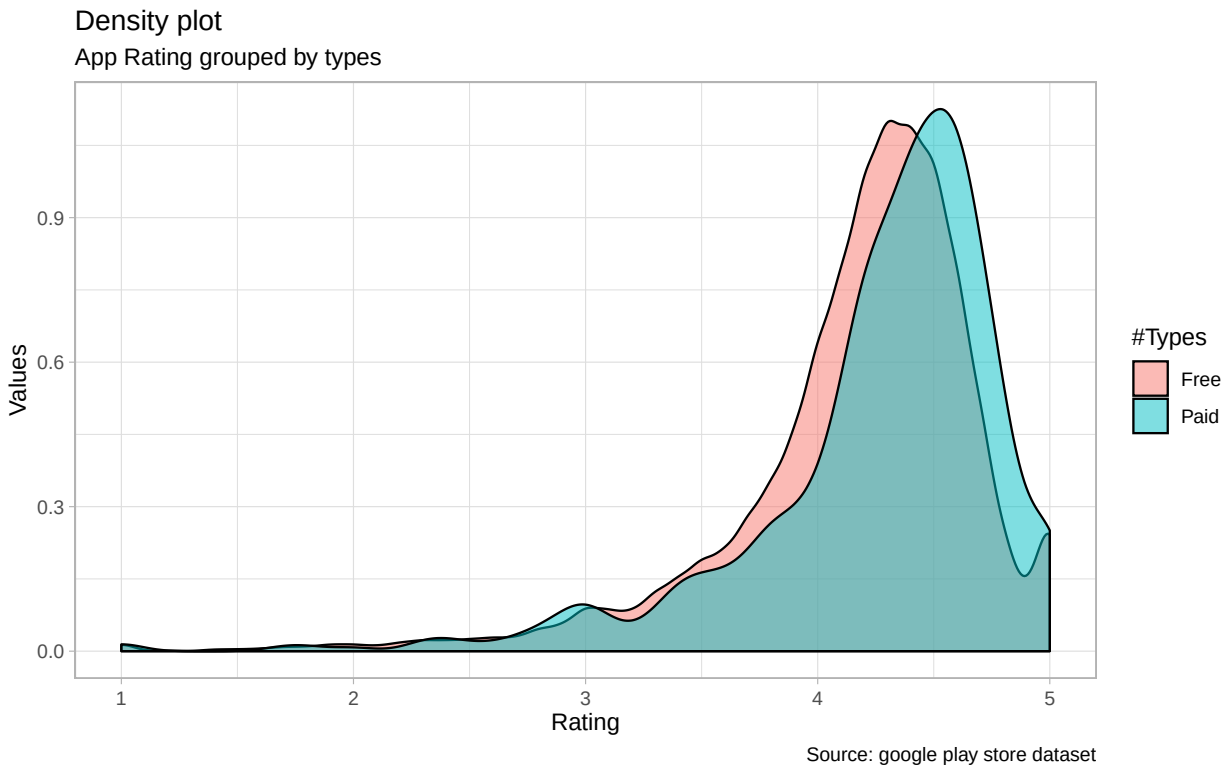
- We believe there's a diverse set of audience who might be interested in our project. As of Febraury 2020, 73.3% of the mobile operating system market share belongs to Android devices[11]. This large community consists of the general public who use android devices and appstore,the developer community and anyone who wants to understand how the app market works.
- This project is primarily intended for the growing developer community. It will help them make data backed decisions before launching their application. Besides developers, it is also helpful for tech journalists, Google Play Store users or any other interested party.

4 Exploratory Data Analysis to find factors for success of an app

Generally, the most successful apps have high ratings and high installs.To look at which app makes it to the top, we consider ratings and installs, so we explore these to find any relationship or trends

4.1 Rating column in depth

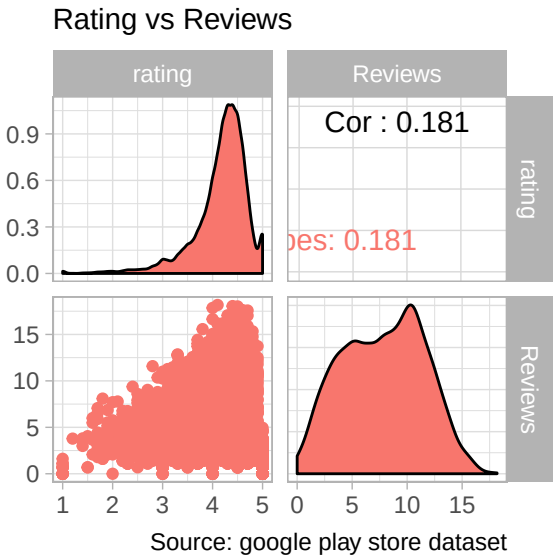
4.1.1 Distribution of rating

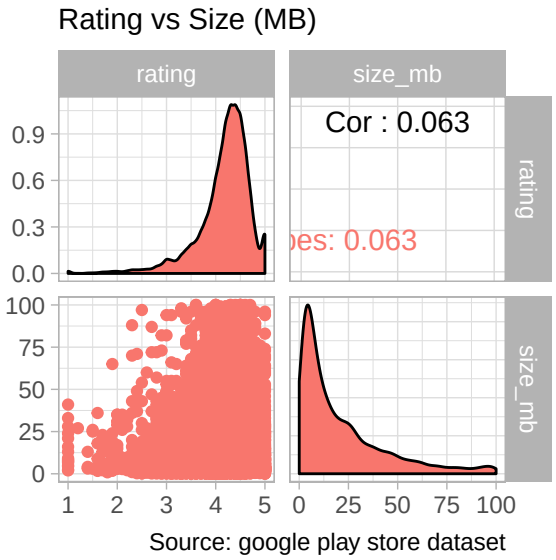


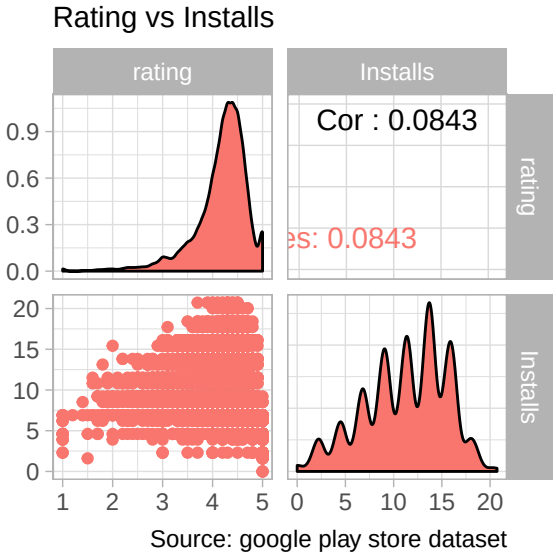
- We can observe that the number of apps with low ratings are less in number, and most apps have a ratings between 3-5.

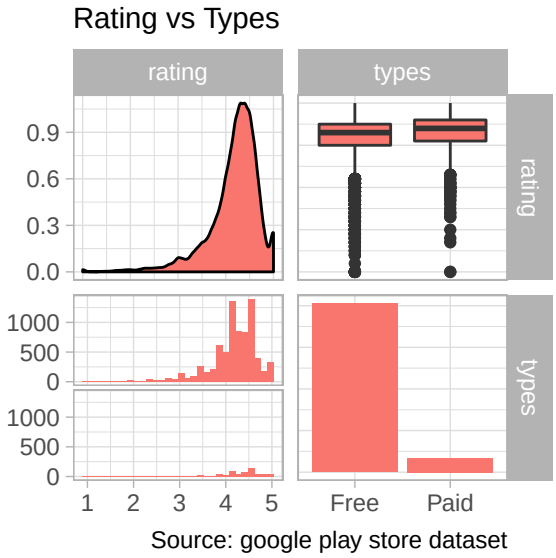
First, we plot correlograms of ratings versus different columns to find any relationship. Correlograms are useful to understand the relationship between different numerical variables. If the correlogram index is 1, it means that the variables are directly proportional to each other.

4.2 Correlogram plots



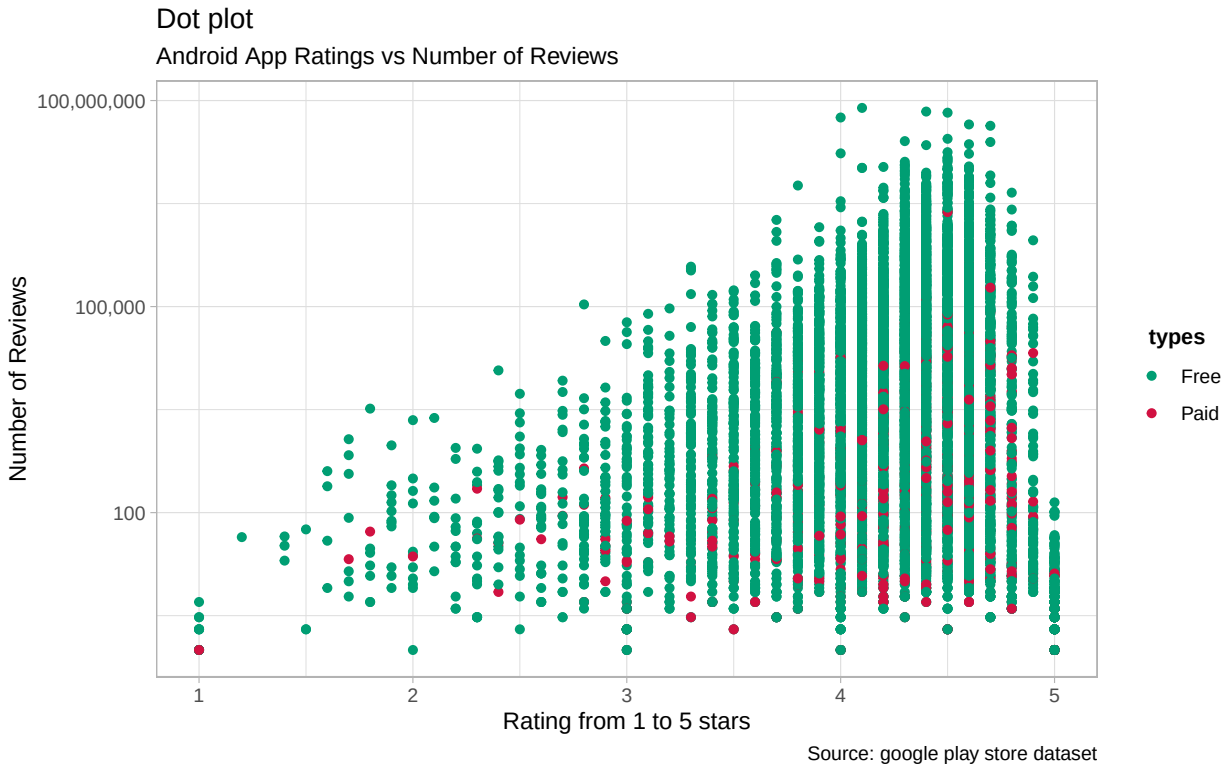






Finding:
+ Each correlation index talks about the relationship between plotted columns. If the index is 1, it means they have linear relationship + Each plot on the diagonal refers to the density plot of the respective column + We can observe that there is no significant linear relationship between rating and the plotted numerical variables.

4.2.1 Plot of reviews vs app ratings

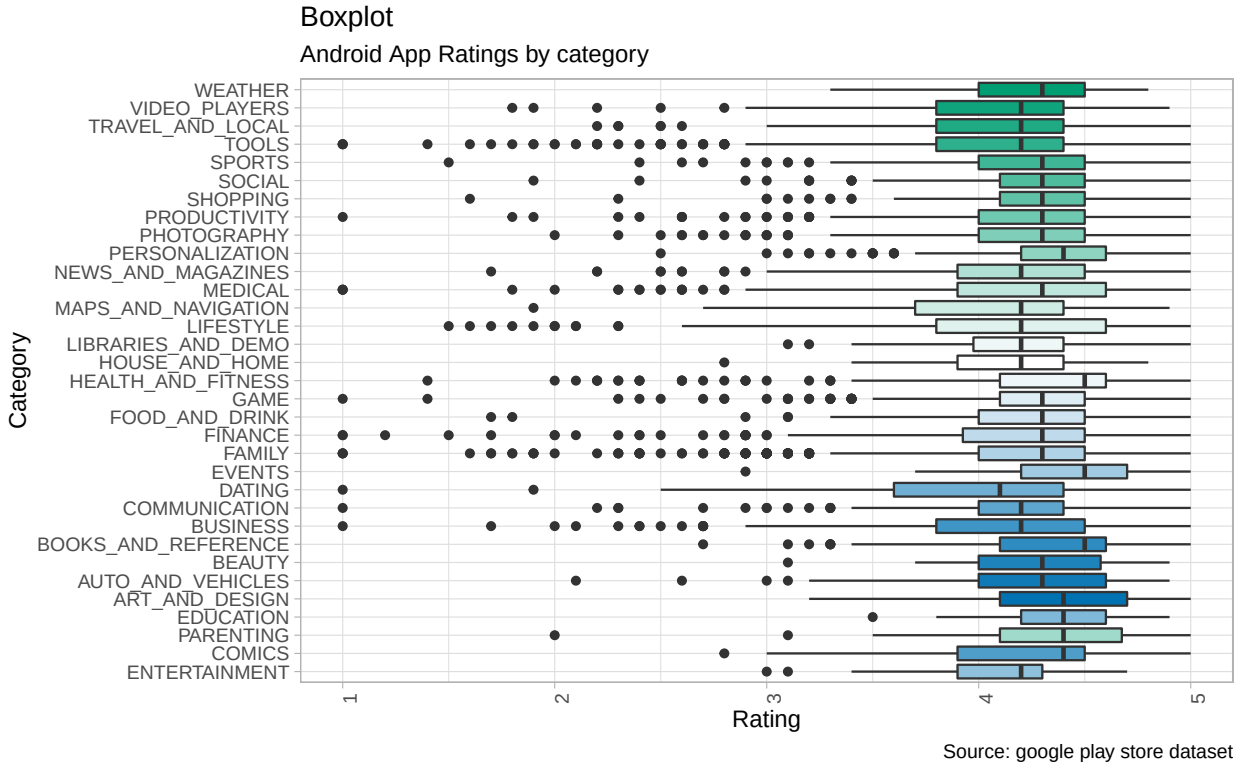


Finding: We can observe that the number of reviews influence the ratings. Generally, as the number of reviews increase, the rating is higher.

We now explore other factors that might potentially influence rating

4.2.2 App rating vs category

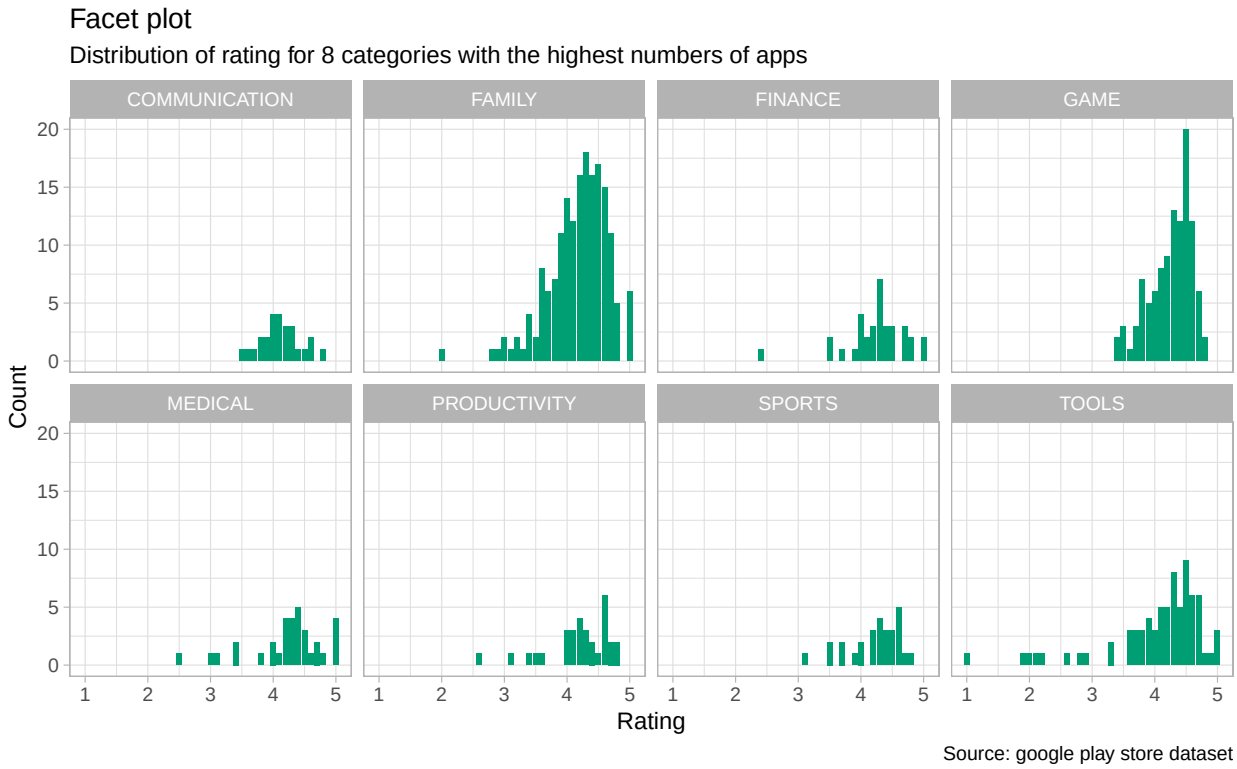
To check the relationship between category and rating, we plotted a box plot with rating on y-axis and category on x-axis.



Finding: This graph shows that for some categories like TOOLS, FAMILY, FINANCE and LIFESTYLE a great majority of applications fall below first quartile. Thus, even though median rating is high, deviation from median is significant.

4.2.3 Distribution of rating for 8 categories with the largest numbers of apps

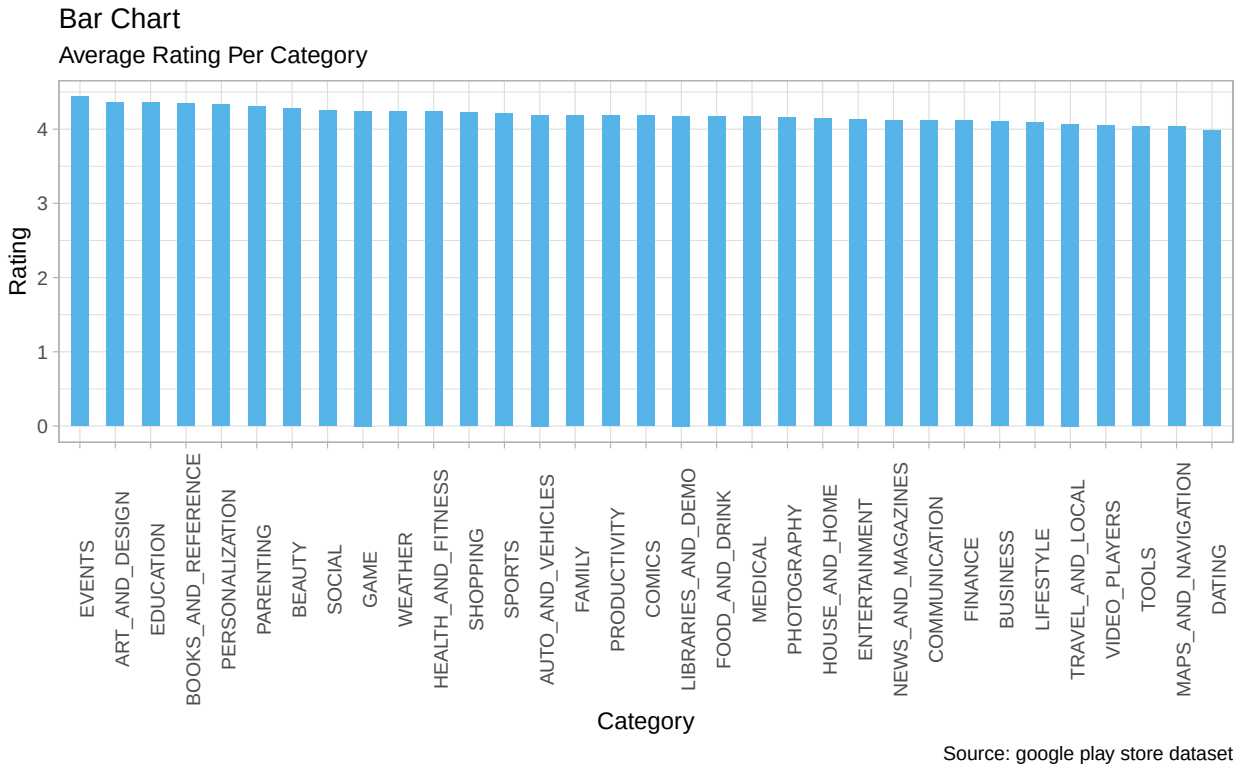
Here we look at the distribution of rating across different categories. We chose 8 categories with the largest number of applications.



Finding: The distribution of rating varies significantly across each category.

4.2.4 Average rating per category

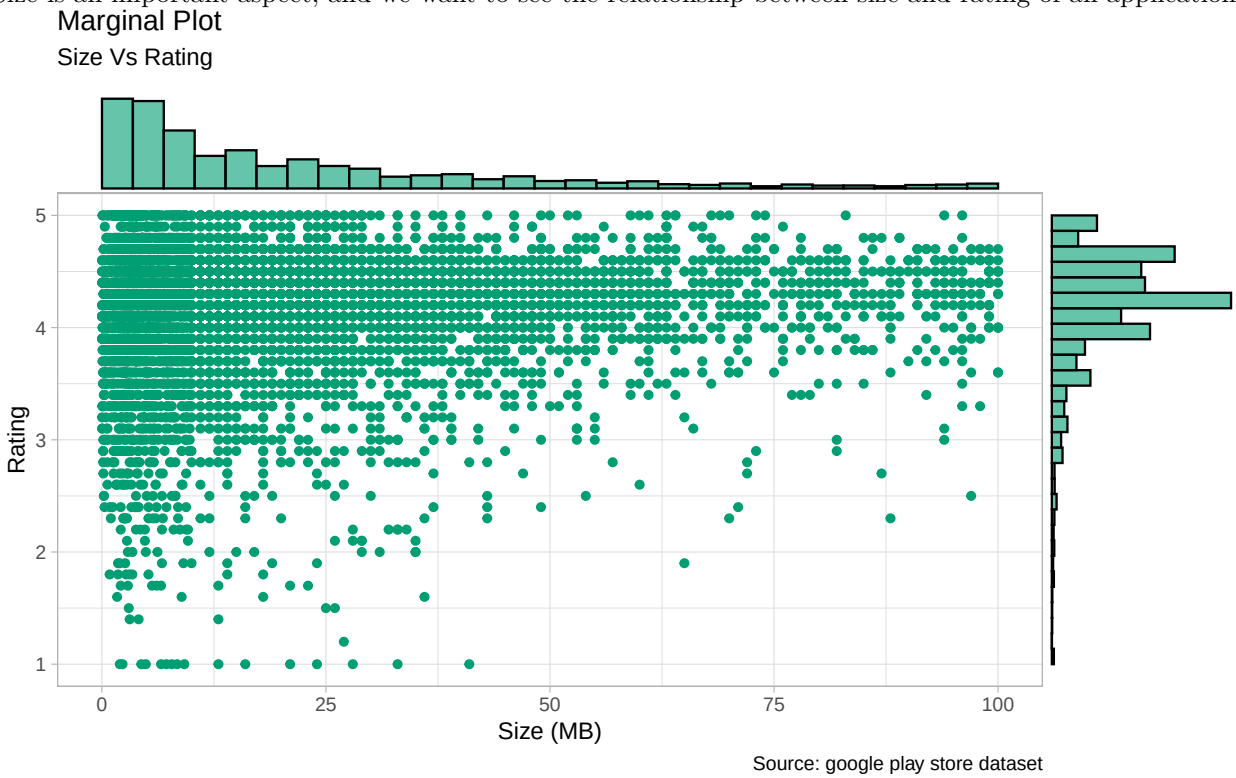
In previous graph we observed that distribution of rating as per category varies significantly. Now we want to find out what the average rating per category is.



Finding: This graph shows that the average rating per category is not very different. Still the "EVENTS" category has the highest average rating, and "DATING" category has the least average rating.

4.2.5 Size and rating

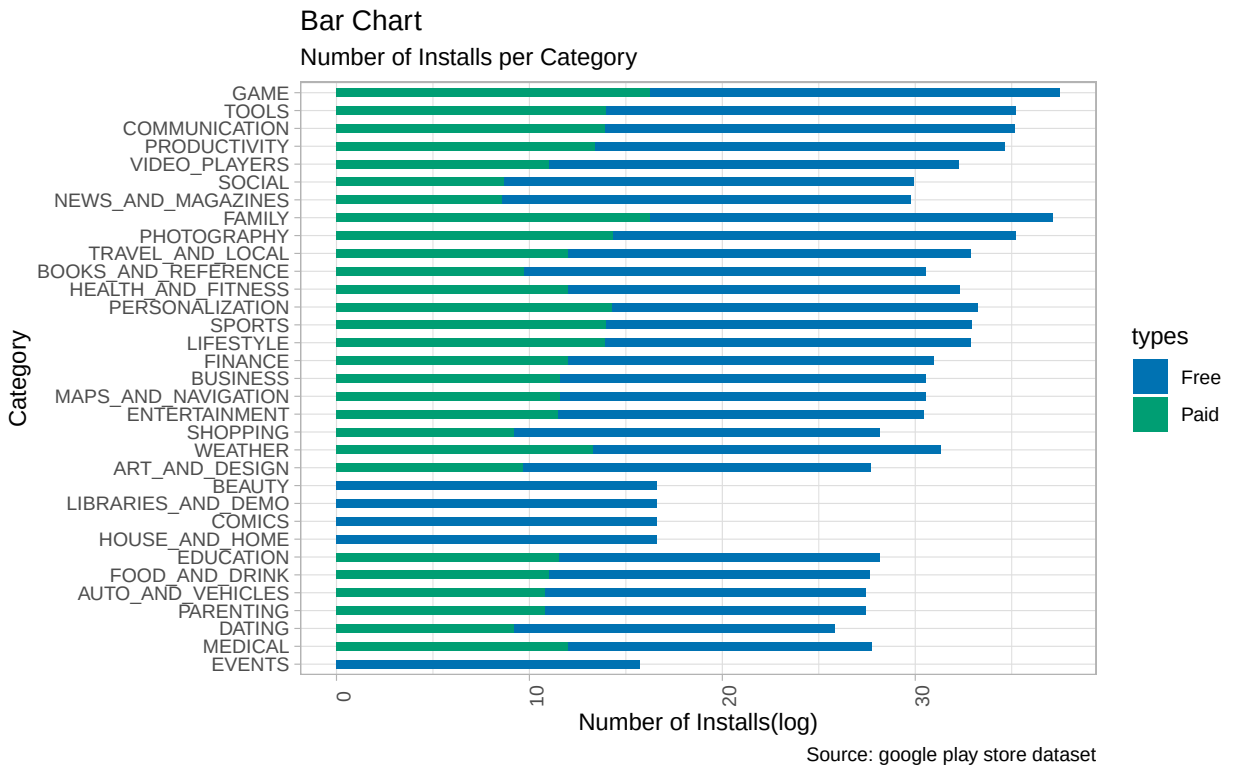
Size is an important aspect, and we want to see the relationship between size and rating of an application.



Finding: We can see that majority of applications with their sizes under 25 MB, have a good rating(4).

4.3 Installs column in depth

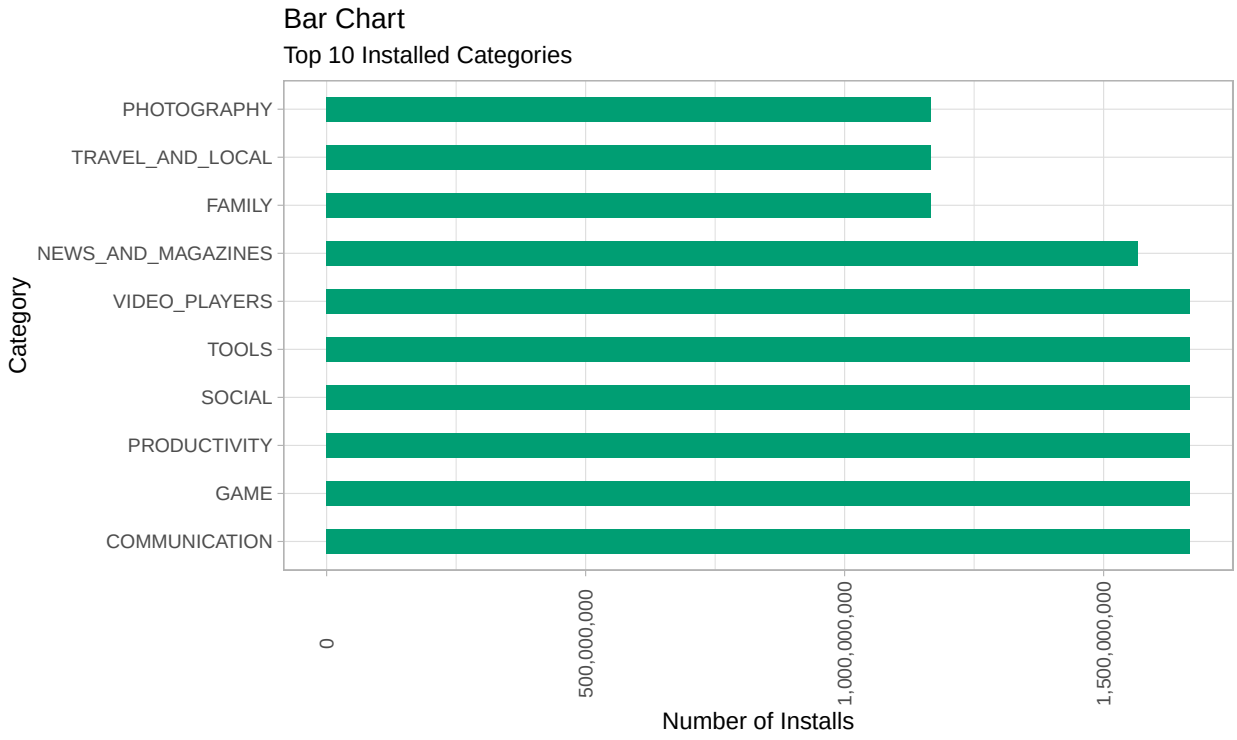
4.3.1 Number of installs per category



Finding: The graph shows the log of number of installs (the values of installs varied from 0 to 1 billion) vs CATEGORY. FAMILY and GAME has the highest number of installs. EVENTS, HOUSE_AND_HOME, COMICS, LIBRARIES_AND_DESIGN and BEAUTY have the least number of installs.

4.3.2 Top 10 installed categories

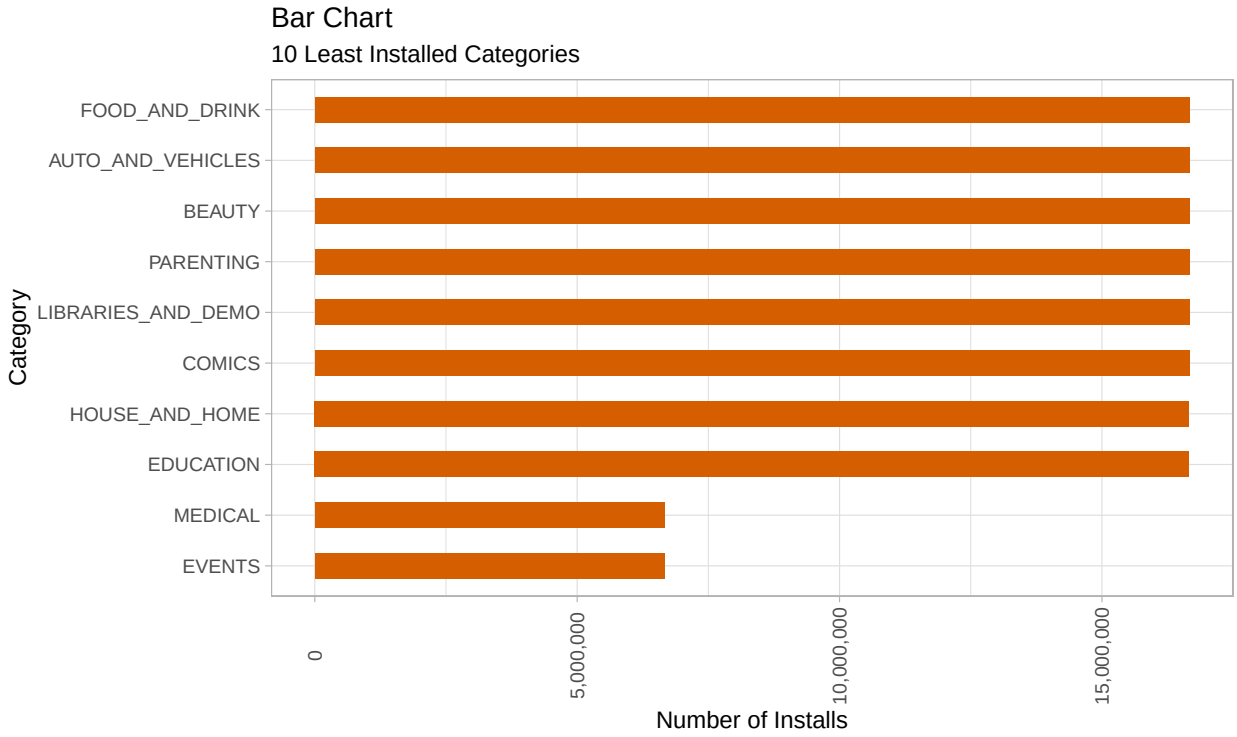
Top 10 categories with greatest number of installs.



Source: google play store dataset

Finding: COMMUNICATION has the highest number of installs.

4.3.3 10 least installed categories

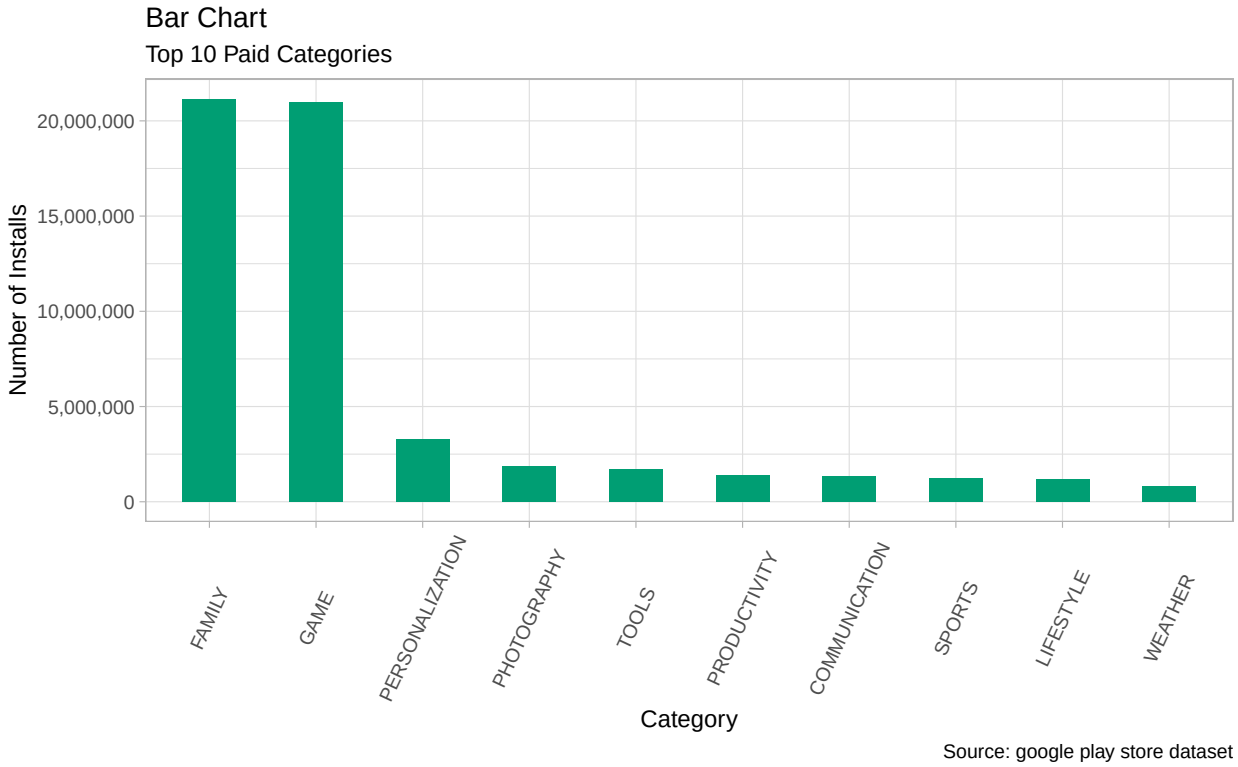


Source: google play store dataset

Finding: Events has the least number of installed applications.

4.3.4 Top 10 paid Categories

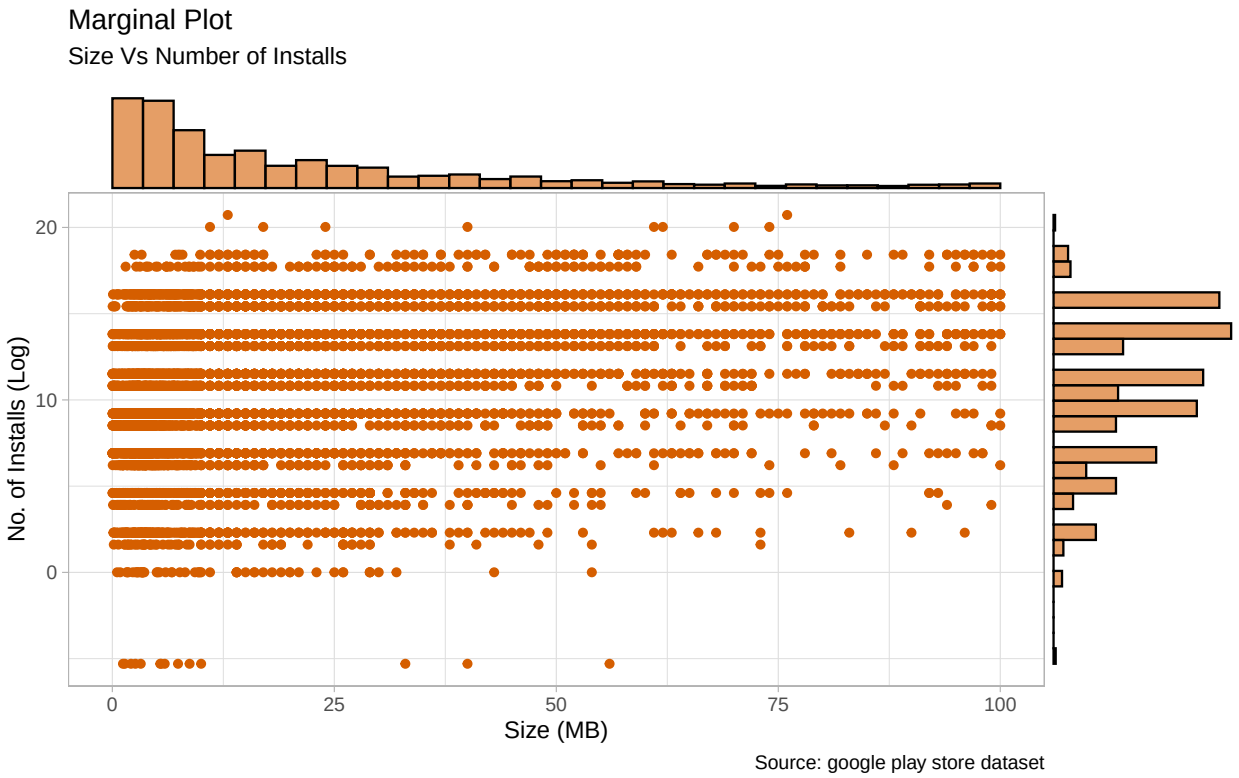
Top 10 categories with the highest number of installs for paid applications.



Finding: FAMILY and GAME has the greatest number of installs.

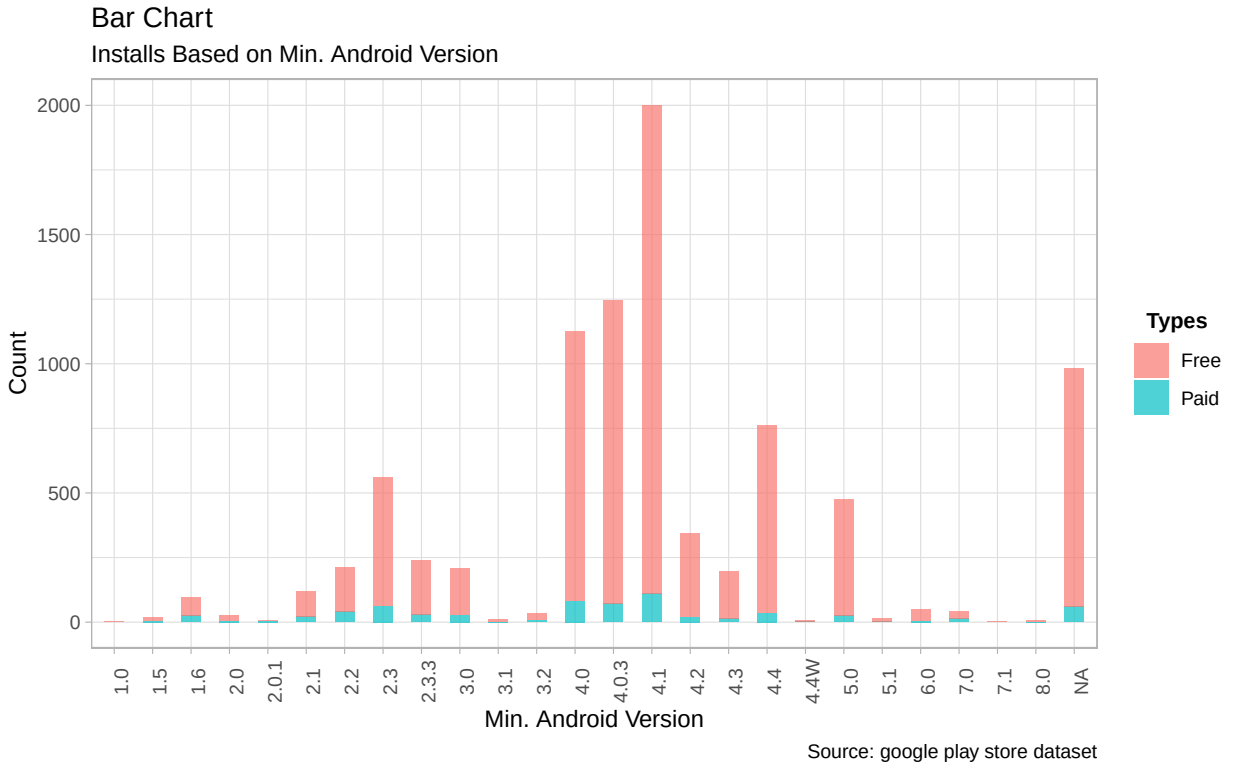
4.3.5 Size Vs. number of installs

Size is an important characteristic of an application. Large applications might reduce the number of installs, as it reduces to targeted audience. We can test this by plotting size against installs.



Finding: Optimally sized Applications, with sizes between 5MB and 30MB, gets the greatest number of installs.

4.3.6 Number of installs based on support by minimum android version



Finding: 4.1 android version has the maximum number of installs.

5 Conclusion

- To get most success from an app, it has to have maximum number of reviews and maximum number of ratings. These are the other trends we found:
  - Apps should atleast support 4.1 android version or more to succeed.This is expected because the majority of users have smart phones with constant android updates.
  - They have high chances of success if the genre is family, game, communication or productivity apps whereas food_and_drink, auto_vehicles categories have very low probability to succeed.
  - The apps which are be free have huge success rates however there are a few exceptions to this.
  - Apps should be optimally sized between 5MB and 30MB.

6 Future work (models)

6.1 What are we trying to achieve (thesis/hypothesis)

The dataset we have contains information about 10000 apps dated till 2018. With the dataset, we plan on using Linear Regression Model, Recursive Partitioning Model and Random Forest to perform predictive analysis. We want to answer the following hypothesis:

1) Find similarities in apps that make it to the top of Play Store. Factors contributing to the success of applications.

2) Can we predict rating of apps based on other parameters such as number of reviews or the size of an app?

To answer these questions, we will be exploring all the variables of this dataset to find if there's any relationship between rating and other variables. We intend to find out which of these variables will play the most important role in predicting rating.

6.2 Why is this important/interesting?

Before we started this project, we took part in a competition called game-jam [12] where we built a game,we had an idea of launching it on Google Play Store.We then thought it would be interesting to see currents trends in the market and to do a detailed analysis to get more insights to the following questions:
1.Factors that influence the success of an app,
2.Which categories are highly installed
3.What are the most famous applications and do they have any trends in common like number of installs, number of reviews, size or android-version? and etc

These analysis inturn will aid the developer community to build successful apps targeting a specific audience.

6.3 How are we going to test the hypothesis?

Since the dataset is dated till 2018. We are thinking to test it by comparing the predictions of our model with the 2019 dataset or the latest dataset.

6.4 Any challenges that we might encounter

There is a lot of useful information that could have given us more insight into the Play Store market e.g. Demographic data could have offered insights into the rating and number of installs of apps, with respect to different regions, different cultures and different trends popular to specific age groups. Also, it would have been interesting to see how different global trends affect the usage of the app, for instance, the current pandemic "covid-19" has called for quarantine across the globe and many people, markets and other companies are relying on smart phones and virtual connections, this will heavily increase the use of many applications, thus deviating from the general trend.

7 Model

- Our main goal is to predict rating of an app based on different parameters. Our EDA confirms that variables like number of installs, number of reviews, category of the app and the type of app (paid/free) highly affect the rating of an app.
- In this section, we try to fit both explanatory and predictive models. We used the cross validation technique to select the best fitting model for our dataset and lastly, we test the models by predicting ratings for a dummy app.

7.1 Correlation using Pearson method

Before creating a model, we will investigate the correlation between different variables to help select the exploratory variables.

```
##
##  Pearson's product-moment correlation
##
## data:  clean_play_store$installs and clean_play_store$rating
## t = 3.6459, df = 8194, p-value = 0.0002681
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.01861078 0.06184077
## sample estimates:
##       cor
## 0.04024461
```

```
##
##  Pearson's product-moment correlation
##
## data:  log10(clean_play_store$installs)^2 and clean_play_store$rating
## t = 10.376, df = 8194, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0924570 0.1351958
## sample estimates:
##       cor
## 0.1138791
```

- **Correlation coefficient between rating and installs**: The correlation coefficient is 0.0402. Using the square of log transform of the installs variable increased the correlation coefficient to 0.1138. This indicates that there is a slight positive relation between the variables which can also be observed in the graphs (given in the next section) where the blue line indicates the best fitted Local Regression model and the red line shows a Recursive Partitioning model.

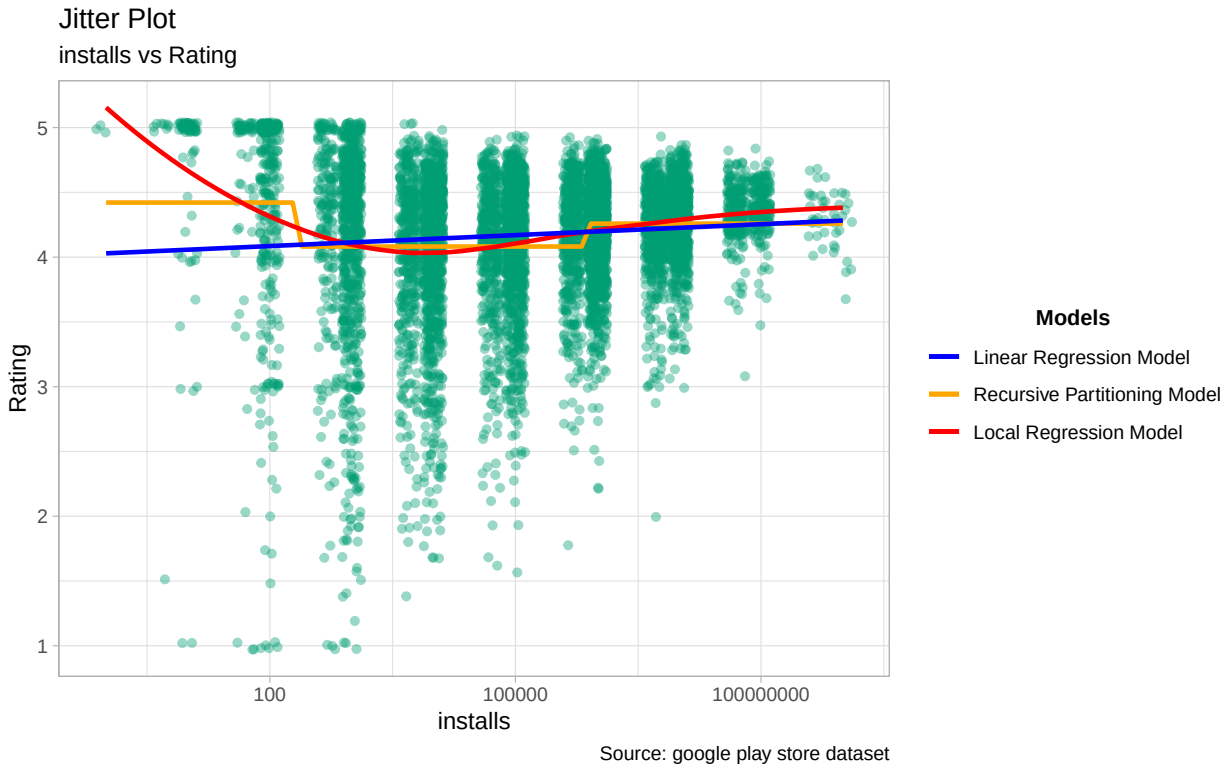
```
##
##  Pearson's product-moment correlation
##
## data:  clean_play_store$reviews and clean_play_store$rating
## t = 5.0001, df = 8194, p-value = 5.849e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03354294 0.07671138
## sample estimates:
##       cor
## 0.05515293
```

```
##
##  Pearson's product-moment correlation
##
## data:  log10(clean_play_store$reviews)^2 and clean_play_store$rating
## t = 18.775, df = 8194, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1822407 0.2237556
## sample estimates:
##       cor
## 0.2030894
```

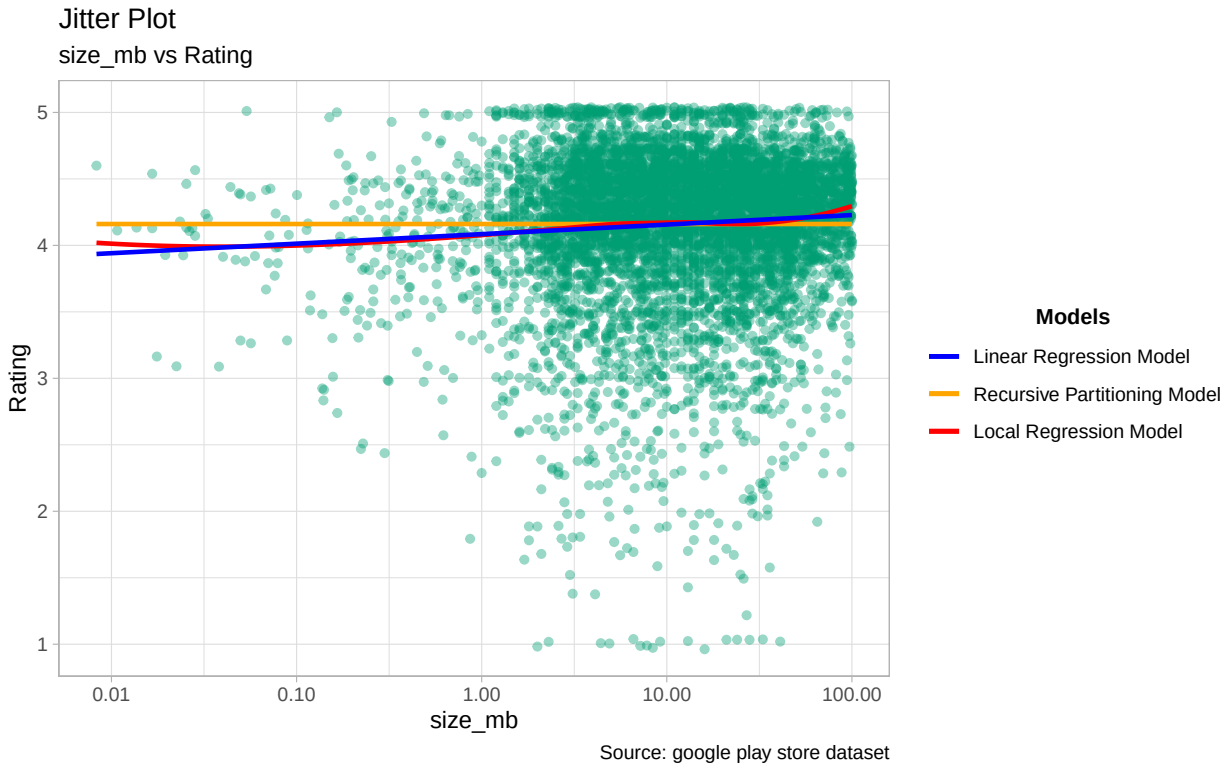
- **Correlation coefficient between rating and reviews**: The correlation coefficient is 0.05515. Using the square of log transform of the installs variable increased the correlation coefficient to .20308.

7.2 Explanatory/Descriptive modeling

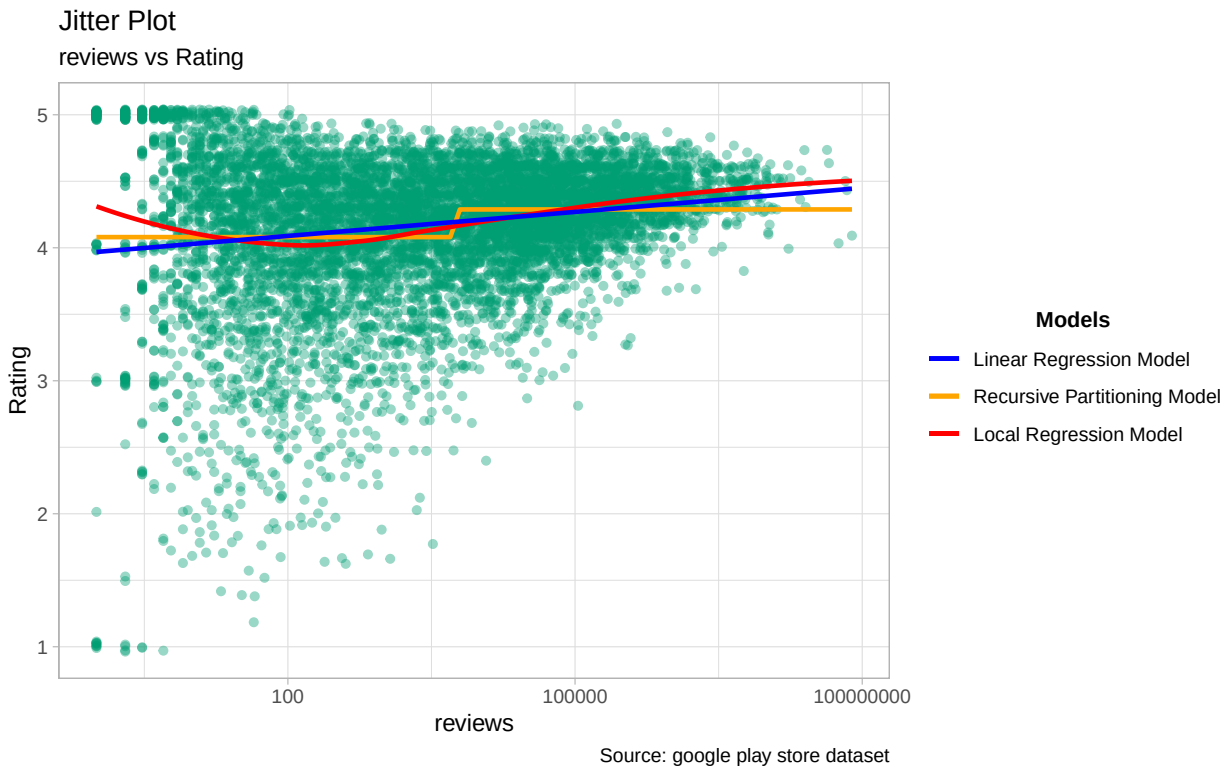
- The following models are used:
- Linear Regression Model
- Local Regression Model
- Recursive Partitioning Model



Source: google play store dataset



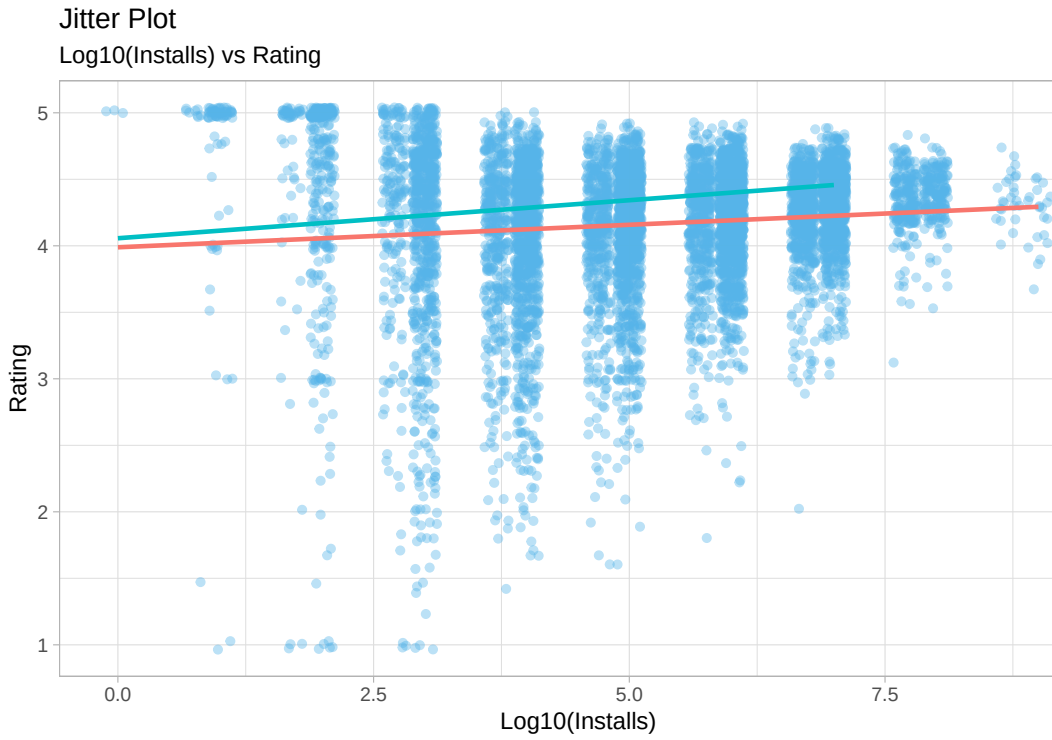
Source: google play store dataset



The above graphs show the fitted Linear Regression model, Local Regression model and Recursive Partitioning model for predicting rating using number of installs, size of the app and number of reviews variables.

7.3 Explanatory/Descriptive modeling (cont.)

Continuing with our explanatory modeling we will fit a model for rating vs installs given taking into consideration whether the app is Free or Paid. We have also calculated sum of square residuals, R-Square value and Root



Mean Square Error for the model.

Observations/Findings: We observed that the rating of paid applications is much higher than the free apps. Which might be indicate its better quality. On the other hand, Free apps have a higher number of installs, thus covering a wide range of audience.

7.4 Cross Validation of models

Choosing the model that gives us the least root mean square error. We will make a comparison between lm and rpart, and try to find which model provides the best analysis for our data

```
## [1] 0.5130422
```

```
## [1] 0.5031791
```

- We sliced the dataset into training and test dataset. Training data det contain 6951 observations (75% of total) and test data set contains 2198 observations (25% of total).
- We trained both lm model and rpart model with the training dataset.
- After training the model we predicted the rating values for test dataset. In order to find which model did a better job of predicting the value of rating, we calculated the root mean square error (RMSE) for the predictions made by both models. RMSE values for lm and rpart are 0.5476666 and 0.5399403 respectively. We can see that RMSE value for rpart is less than RMSE value for lm, implying that rpart can make a better prediction than lm.
- Along with performing a comparison between lm and rpart, we experimented with different explanatory variables. We got the least RMSE by using installs, types, categories and reviews as explanatory variables.

7.5 Making Predictions

We can predict rating of a dummy app by entering different parameters like installs, types, category and reviews

```
##        1
```

```
## 4.294068
```

```
##        1
## 4.132784
```

- **1st dummy application**: First dummy application of type Free with 1 billion installs and 100000 reviews, belonging to TOOLS category has a predicted rating of 4.2884. This seems like a reasonable prediction A Free app with 1 billion installs and 100,000 reviews must be a popular app and is expected to have high rating. But the fact that it belongs to "Tools" category, which has the third least average rating, its rating is not exceptionally high. Its rating much better than average (4.17) but is not among the highest ratings.
- **2nd dummy application**: Second dummy application is of type Free, with 1000 installs and 100 reviews, belonging to EVENTS category, has a predicted rating of 4.0436. This seems like a reasonable prediction. A free app is expected to have higher installs, but as this app does not have high number of installs, indicating that the app is not the best in its segment or it has limited relevance or utility, thus we can expect low rating. But, as it belongs to EVENTS category, which has the highest average rating among all categories, its rating is not among the worst. It has a below average rating but is still not among the wort rated apps.

8 Conclusion

9 References

[1] Dataset;
[2] R markdown ;
[3] Stackoverflow ;
[4] GGally;
[5] Custom color pallete;
[6] Tidying data;
[7] Colors;
[8] TidyVerse;
[9] ggplot2;
[10] Google playstore Kernel by Danilodiogo;
[11] market share ;
[12] Global game jam ;
[13] Play Store Statistics ;