

# Play Store App Review Analysis



*Sandip Dey*

*Data science trainees*

*AlmaBetter*

**Abstract** - Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. Application (App) ratings are feedback provided voluntarily by users and function important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. This Study aims to predict the ratings of Google Play Store apps using machine learning Algorithms. I have tried to perform Data Analysis and prediction into the Google Play store application dataset that I have collected from Kaggle. Using Machine Learning Algorithms, I have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application.

**Key Words:** Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Machine Learning.

## 1 PROBLEM STATEMENT

---

Data is taken from the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in data science cycle, as it not only helps in taking very initial business decisions but also in preparing the data for further modelling for use in machine learning algorithms. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app

## 2 INTRODUCTION

---

Machine learning approaches are essential for us to take care of numerous issues. In this paper, we present machine learning models and structures in detail. Machine learning has numerous applications in numerous perspectives and has incredible advancement potential.

In future, it is predictable that machine learning could set up ideal speculations to clarify its exhibitions. In the meantime, its capacities of unsupervised learning will be improved since there is much information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly

unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize these points of interest to achieve more assignments.

## 2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

In today's scenario we can see that mobile apps playing an important role in any individual's life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that although it creates more than two-fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and rating to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on generally speaking evaluations alone and would profit by understanding a huge number of printed remarks.

We develop Android apps & release on Play Store. As an Developer or say Business Perspective it's very important to know whether users are enjoying the app or facing any issues. To know this Play Store has a Ratings & reviews section for each app released on play store. Users can submit the ratings and has a freedom to write a review for a particular app. This approach is quite a lengthy to rate & review app i.e. navigate to Play store to submit feedback or redirect leaving a current app workflow to open Play Store App link using URI. We never wanted our customers to leave our application, but with this flow, we are forced to redirect the control to Play store app.

## 2.2 GOOGLE PLAY STORE DATASET

The dataset consists of Google play store application and is taken from Almbetter, which is the world's largest community for data scientists to explore, analyze and share data.

This dataset is for Web scraped information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future

references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

### The data set contains the following columns:

- **App:** This Column contains the name of the app
- **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
- **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
- **Reviews:** This column contains the number of people that have given their feedback for the app.
- **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
- **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
- **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
- **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
- **Content Rating:** It indicates the targeted audience of the app and their age group.
  - **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
  - **Last updated:** This column contains the info about the date on which the last update for the app was launched.
  - **Current version:** Contains information about the current version of the app available on the play store.
  - **Android version:** Contains information about the version of the android OS on which the app can be installed.
  -

## 2.3 USER REVIEW DATASET

- User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:
- **App:** Contains the name of the app with a short description (optional).
- **Translated Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be 'Positive', 'Negative', or 'Neutral'.
- **Sentiment Polarity:** It gives the polarity of the review. Its range is [-1,1], where 1 means 'Positive statement' and -1 means a 'Negative statement'.
- **Sentiment Subjectivity:** This value gives how close a reviewer's opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer's opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

## 2.4 PYTHON

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programming language to select up compared to other language. That is the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That is one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open-source library is named panda. As we have seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

## 2.5 DATA CLEANING AND PREPARATION

Pre-processing is important into transitioning raw data into a more desirable format. Undergoing the pre-processing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need pre-processing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.

The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

- **Step1:** We write a function `play store info ()`, that will display 5 attributes about all the columns: Data type, count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.
- **Step2:** we start off with the column 'Type' we can see that it has one null value. We checked this row and found out from the play store that it is a free app. We use `fillna()` function of the pandas library to fill this value.
- **Step 3:** We can see that the 'Rating' column has 1474 null values. Due to low variations in the rating values and a lot of repeated values the 'median' would be a suitable statistical indicator to replace the null values with. We calculate the mode of the column using the `median ()` aggregate method, and fill this value in place of null values using the `fillna()` function.
- **Step 4:** We can see that the 'Reviews' column despite being a numerical indicator is of the 'object' data type, we will convert this to 'int' data type using the `astype(int)` function.
- **Step 5:** We can see that the size column, which should be numeric, is of the data type 'object', it also has characters 'k' and 'M' in the values which stand for kilobytes and Megabytes, we will replace the 'k' with 1000 and 'M' with 1000000. Some values also have '+' sign in them, which will be removed. Next, we will convert this column into 'int' datatype.
- **Step 6:** The 'Installs' column values contain the characters '+' and ',' which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the `strip()` and `replace()` functions.
- **Step 7:** The values in the column 'Price' might have the '\$' sign in some values and the column is of the datatype 'object'. We will first remove the '\$' sign using the `replase()` function and then convert the column into 'int' datatype.
- **Step 8:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

- ### 3 EXPLORATORY DATA ANALYSIS

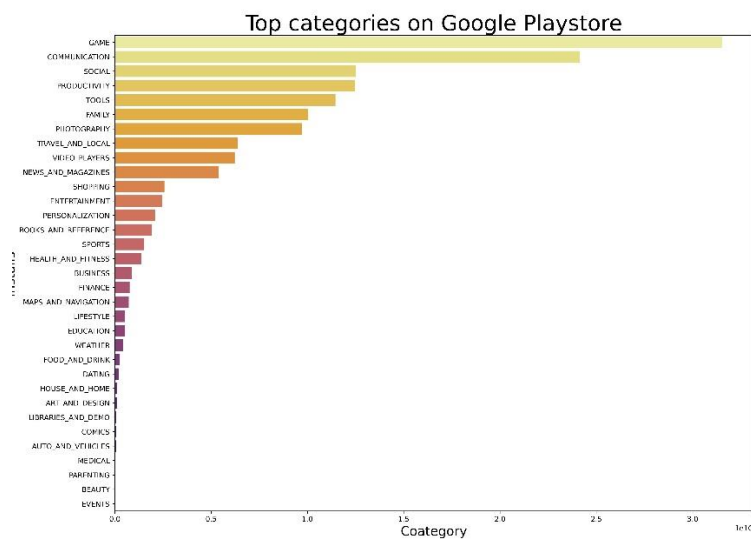
EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

Top categories on Google Playstore

Category	Count
ENTERTAINMENT	1950
GAMES	1150
TOOLS	850
PRODUCTIVITY	450
PERSONALIZATION	430
LIBRARY	420
COMMUNICATION	410
FINANCE	400
HEALTH	390
PHOTOGRAPHY	380
HEALTH_AND_FITNESS	370
SOCIAL	360
NAVY_AND_NAVALYFES	350
ART_AND_VISUAL	340
BOOKS_AND_REFERENCE	330
SHOPPING	320
EDUCATION	310
VIDEO_PLAYERS	300
SPORTS_AND_RECREATION	290
FOOD_AND_DRINK	280
ENTERTAINMENT	270
ART_AND_VISUAL	260
LEARNING_AND_DEVELOPMENT	250
HOUSE_AND_HOME	240
ART_AND_VISUAL	230
ART_AND_VISUAL	220
ART_AND_VISUAL	210
ART_AND_VISUAL	200
ART_AND_VISUAL	190
ART_AND_VISUAL	180
ART_AND_VISUAL	170
ART_AND_VISUAL	160
ART_AND_VISUAL	150
ART_AND_VISUAL	140
ART_AND_VISUAL	130
ART_AND_VISUAL	120
ART_AND_VISUAL	110
ART_AND_VISUAL	100
ART_AND_VISUAL	90
ART_AND_VISUAL	80
ART_AND_VISUAL	70
ART_AND_VISUAL	60
ART_AND_VISUAL	50
ART_AND_VISUAL	40
ART_AND_VISUAL	30
ART_AND_VISUAL	20
ART_AND_VISUAL	10

pg. 7

### 3.2 WHICH CATEGORY APPS HAVE THE GREATEST NUMBER OF INSTALLS?



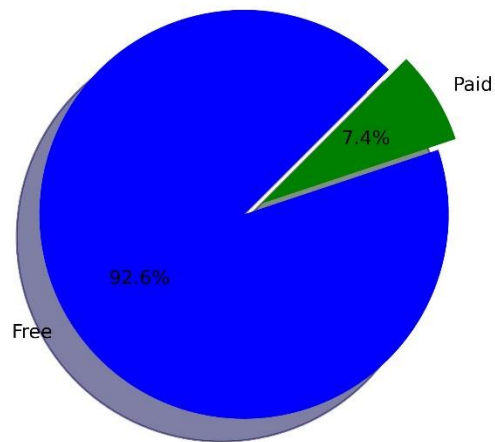
To answer this question we need to create a separate data frame out of our data frame which will contain a grouped value by Category and Installs

From the above visualization, it can be interpreted that the top categories with the highest installs are Game, Family, Communication, News & Magazines, & Tools.

### 3.3 WHICH APP CATEGORY ARE PAID OR FREE WITH PERCENTAGE?

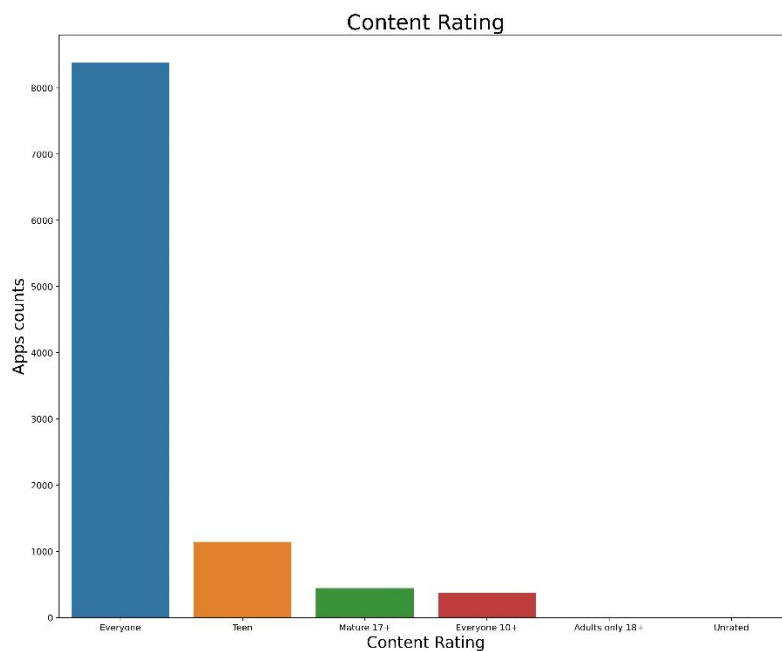


## Percent of Free Vs Paid Apps in store



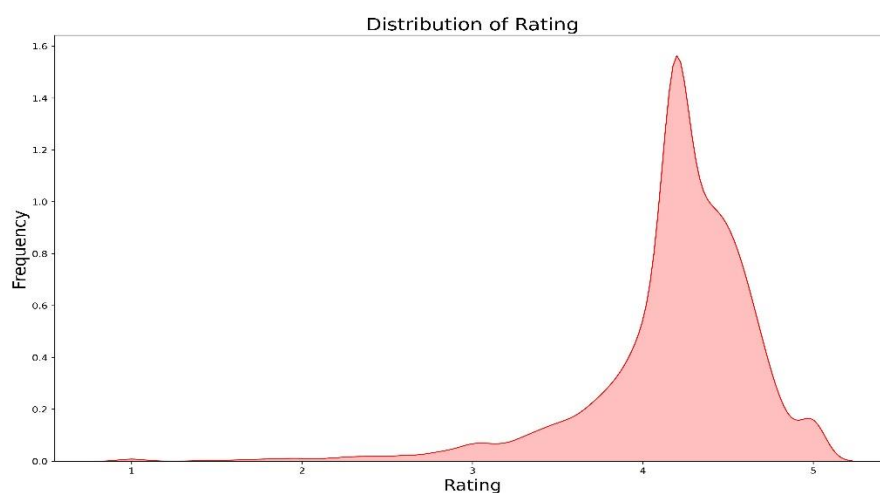
Here we can see that 92.6% apps are free, and 7.4% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

### 3.4 WHICH CATEGORY OF APPS FROM THE 'CONTENT RATING' COLUMN IS FOUND MORE ON THE PLAY STORE?



A majority of the apps in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

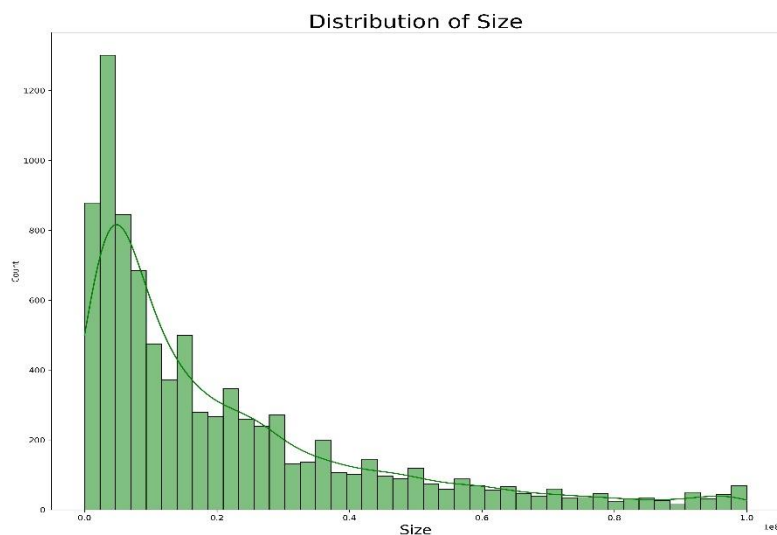
### 3.5 LET'S HAVE A LOOK AT THE DISTRIBUTION OF THE RATINGS OF THE DATA FRAME.



- The mean of the average ratings (excluding the NaN values) comes to be 4.2.
- The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3. From this we can say that most of the apps have an average rating is between 3.8 and 4.8.

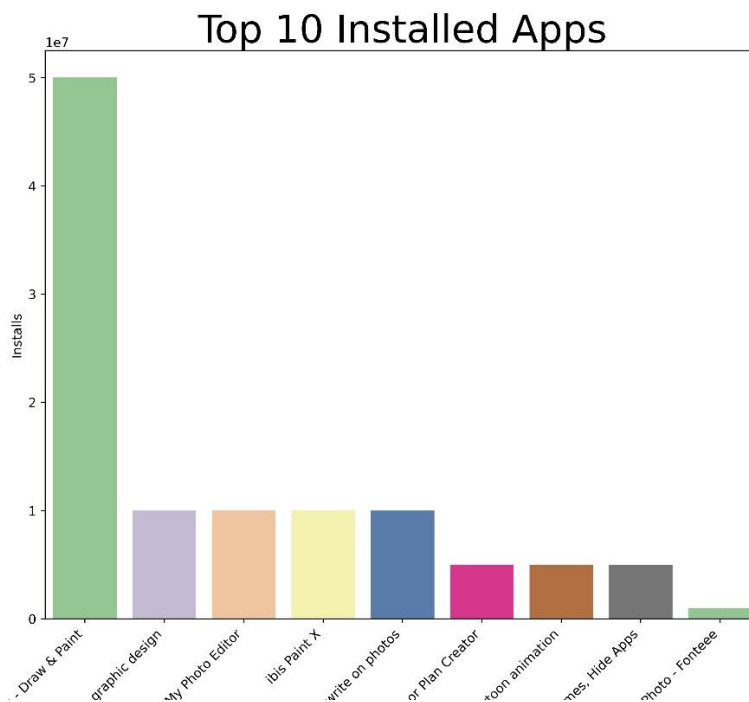
- From the kdeplot visualizations, it is clear that the ratings are left skewed.

### 3.6 LET'S HAVE A LOOK AT THE DISTRIBUTION OF THE SIZE OF THE DATA FRAME.



- It is clear from the visualizations that the data in the Size column is skewed towards the right.
- From the above histogram graph, we can come to the conclusion that maximum number of applications present in the dataset are of small size.

### 3.7 WHAT ARE THE TOP 10 INSTALLED APPS IN ANY CATEGORY?



- So, we have to be able to answer this not only for a single category but for many, i.e., we will need to define a function which should be able to return us a nice plot for any Category the name provided by any user as an argument to it.

```
def findtop10(str):
```

```
    selected_category=clean_df[clean_df["Category"]==str]
```

```
    top10installed_selected_catecorey_app=selected_category.sort_values(by="Installs",ascending=False).head(10)
```

```
    return top10installed_selected_catecorey_app
```

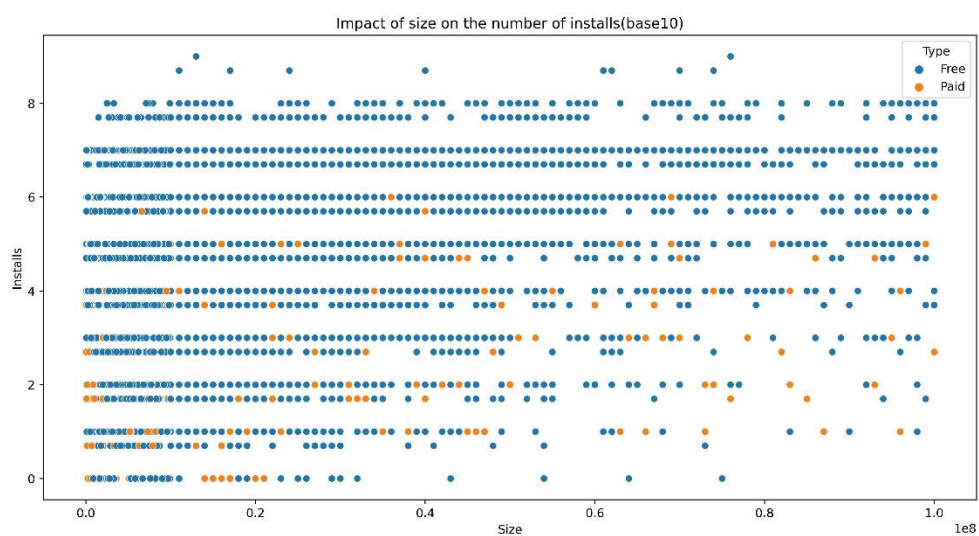
- After we are done with defining the function, it's time to check and see if everything is working fine. So let's test it by passing ART\_AND\_DESIGN category to the above-defined function.
- we can see that in the "ART\_AND\_DESIGN" category's "Sketch - Draw & Paint" has the highest install and Text on Photo-Fontee has lowest install.

### 3.8 WHICH ARE THE APPS HAVING THE HIGHEST NUMBERS OF REVIEWS?



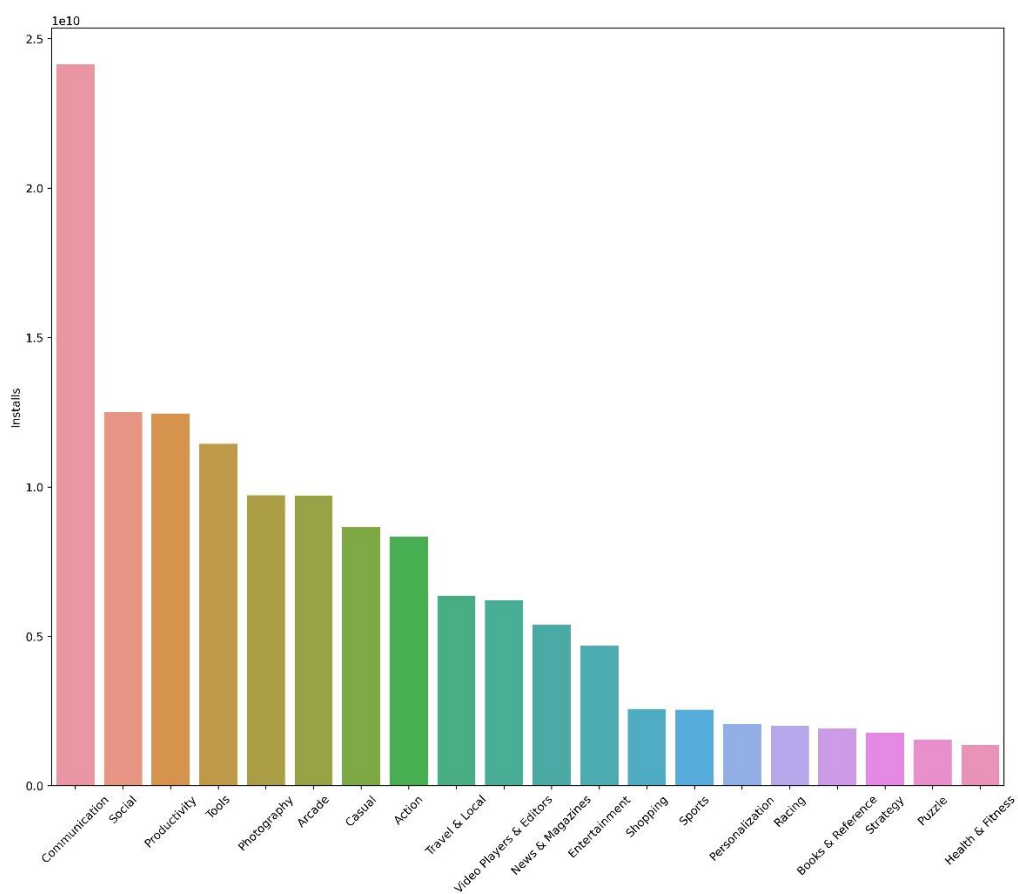
we can see that Facebook has the highest number of reviews and Clash Royale has lowest number of reviews.

### 3.9 HOW DOES SIZE IMPACT ON THE NUMBER OF INSTALLS OF ANY APPLICATION?



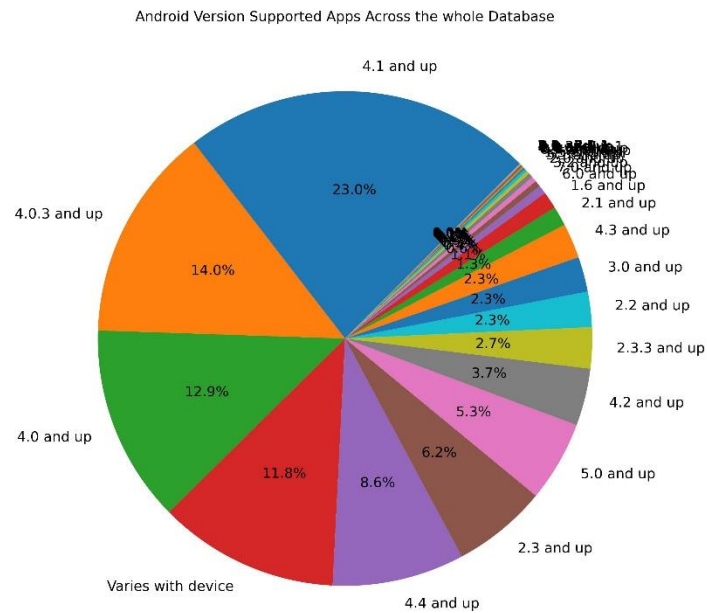
It is clear from the above mentioned plot that size may impact the number of installations. Bulky applications are less installed by the user

### 3.10 WHICH ARE THE GENRES THAT ARE GETTING INSTALLED THE MOST IN TOP 20 GENRES?



From the above plot we can come to the conclusion that maximum app install comes under Communication Genres and followed by Social, Productivity and Tools Genres.

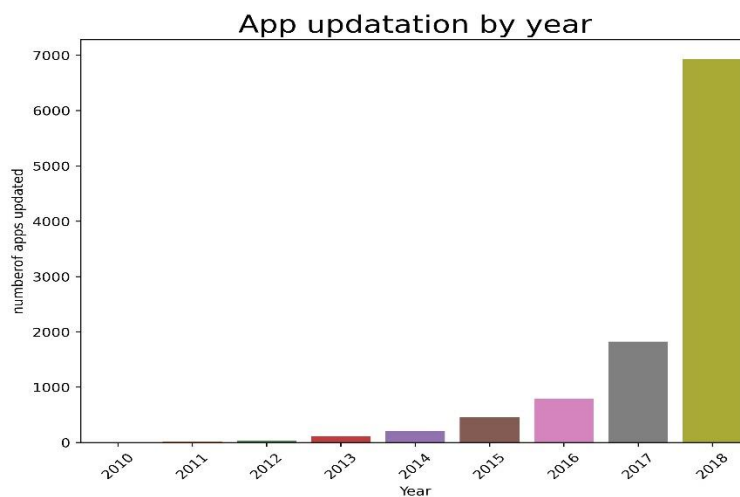
### 3.11 MOST ANDROID VER SUPPORTED APPS IN PLAY STORE.



After identify total distribution percentage on data, given details of more app supported Android OS version Basically android 4.0

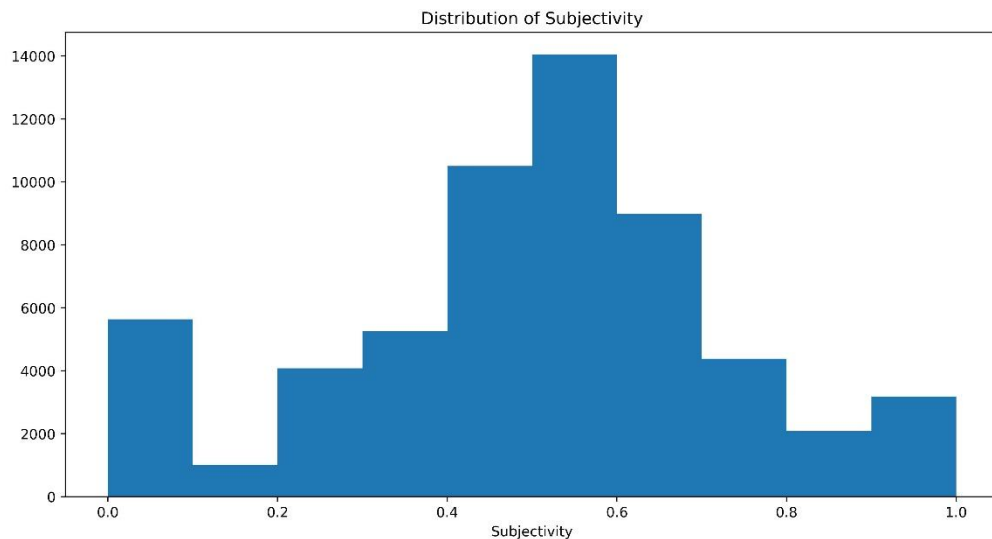
and above version supported app ratio is very higher and more then 60% apps support only on android 4.0 and above.

### 3.12 APP UPDATE DETAILS "BY YEAR".



From this barplot we can see that a very wide range of app updated in play store during 2017-2018.

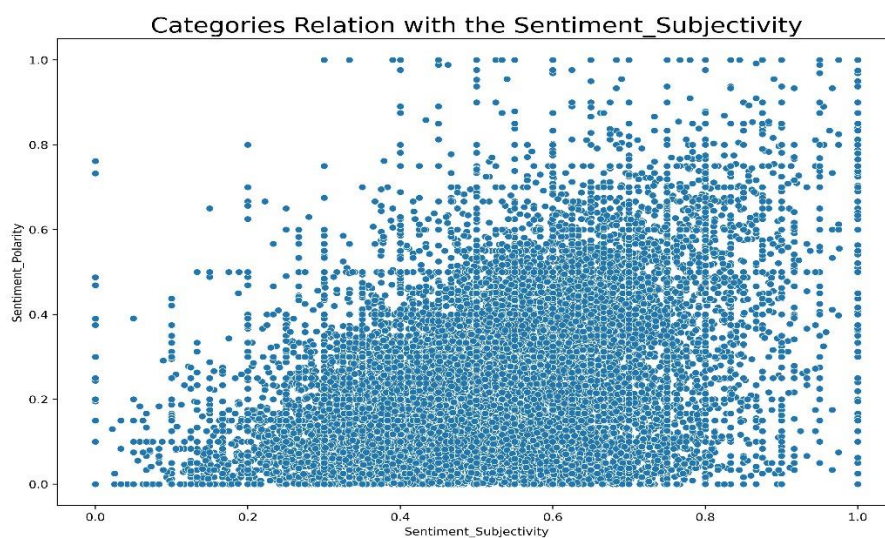
### 3.13 HISTOGRAM OF SUBJECTIVITY.



**0 - objective(fact), 1 - subjective(opinion)**

It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications, according to their experience.

### 3.14 IS SENTIMENT SUBJECTIVITY PROPORTIONAL TO SENTIMENT POLARITY?

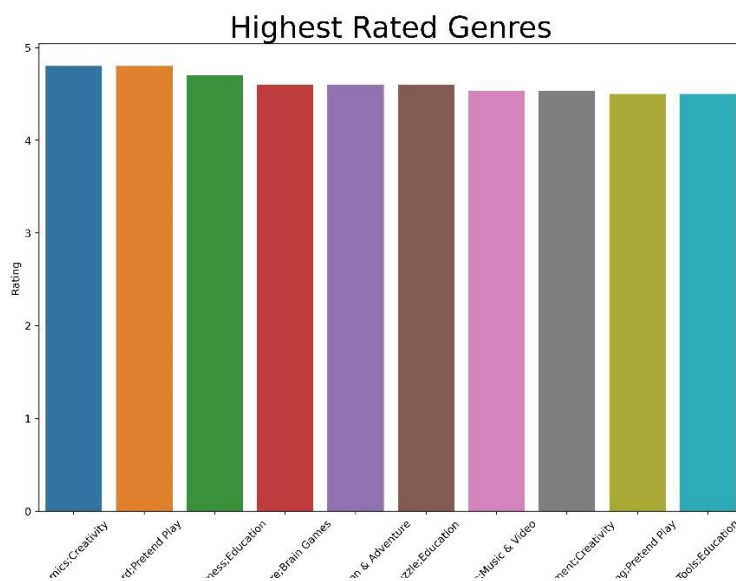




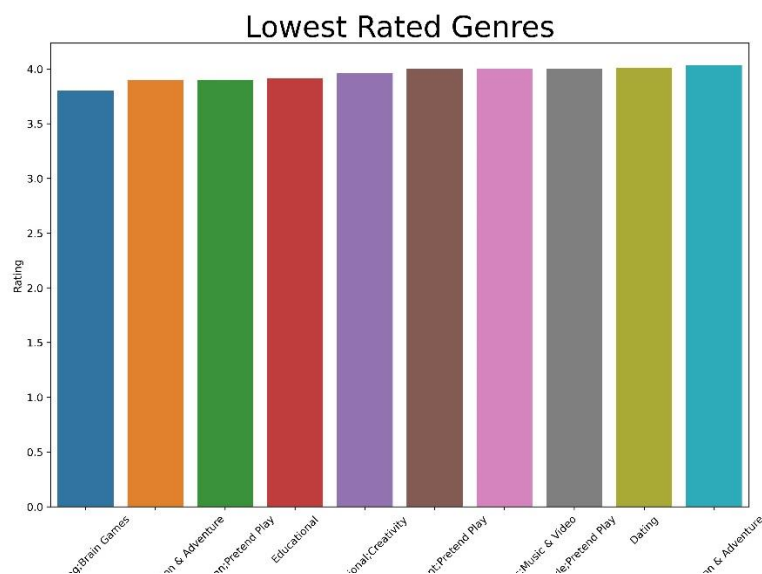
From the above scatter plot

it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behavior, when variance is too high or low.

### 3.15 FIND THE HIGHEST AND THE LOWEST RATED GENRES.



we can see that Comics;Creativity and Board;Pretend Play are the highest rated genres

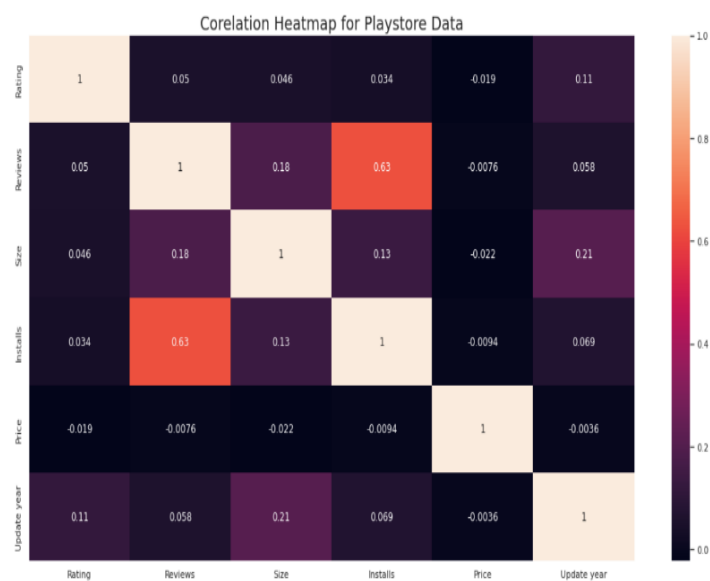


and Parenting;Brain Games is the lowest rated genres

### 3.16 CORRELATION HEATMAP

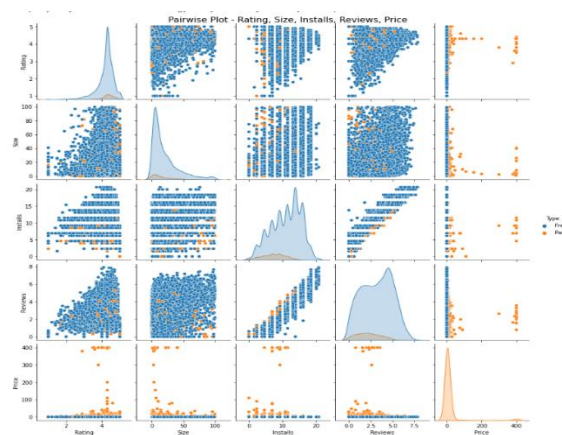
A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1. Correlation between two random variables or bivariate data does not necessarily imply a causal relationship.



Correlation Heatmap

### 3.17 RELATIONSHIP BETWEEN DIFFERENT FEATURES OF THE DATASET



Pair wise plot

## 4 CONCLUSION

---

Through exploratory data analysis we have observed some trends and have made some assumptions that might lead to app success among the users in the play store.

- In play store most of the apps are under Family & Game category and least are of Beauty & Comics Category.
- 92.6% apps are Free and 7.4% apps are paid in type.
- Category with the highest number of installs is Game and least number of installs is EVENTS categories.
- Most of the apps in the google play store are rated between 3.8 to 4.8.
- Education category has a highest mean rating of 4.37 and Dating category has lowest 4.01 rating.
- Family, Game and Tools are top three categories having 1943, 1121 and 843 app count.
- Everyone category has the highest number of apps and Unrated category has the lowest number of apps.
- maximum number of applications present in the dataset are of small size.
- Facebook has the highest number of reviews.
- Bulky applications are less installed by the user.
- Communication Social, Productivity and Tools are top Genres.
- Overall sentiment count of merged dataset in which Positive sentiment count is 64%, Negative 22% and Neutral 14%.
- Maximum number of sentiment subjectivity lies between 0.4 to 0.7.
- It's good to develop a Free type app and having a content rating for Everyone.

- Comics;Creativity and Board;Pretend Play are the highest rated genres and Parenting;Brain Games is the lowest rated genres

# Thank You