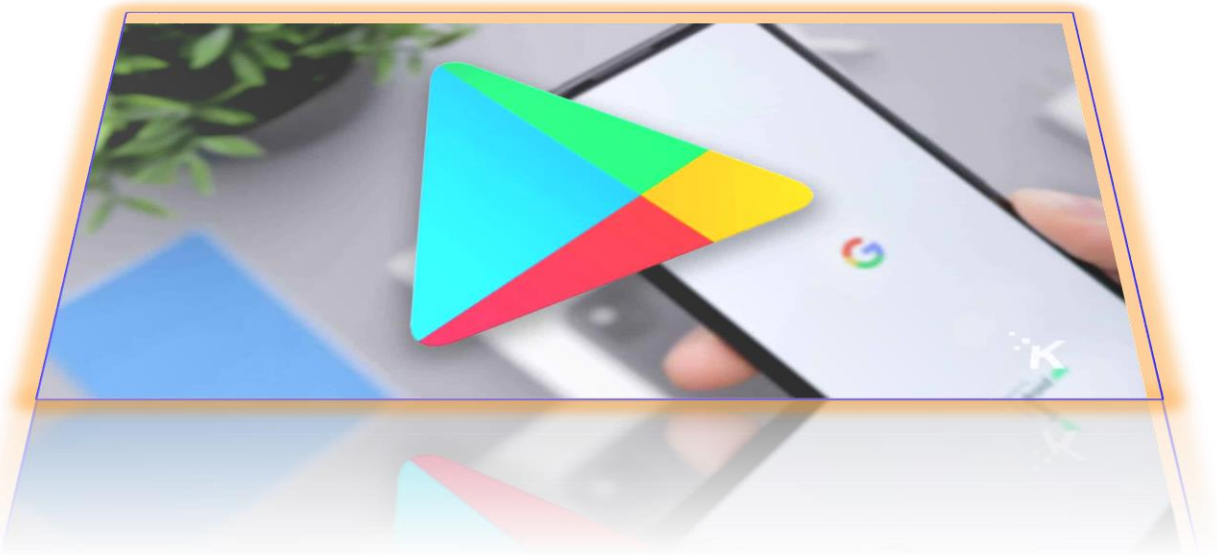
Play Store App Review Analysis



Sanju Khanra

Data science trainees

Alma Better

Abstract - Google play store is engulfed with a few thousandsof 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 challengefrom everywhere throughout the globe. Since most Play Storeapplications 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. Ihave 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 triedto discover the relationships among various attributes presentin 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 careof numerous issues. In this paper, we present machine learning models and structures in detail. Machine learning has numerous applications in numerous perspectives and hasincredible advancement potential.

In future, it is predictable that machine learning couldset 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 thatthe development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile applicationshowcase 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 theirplace in the market the application designers may need to figure out how to stick into their present position. The GooglePlay Store is observed to be the biggest application platform.It has been seen that although it creates morethan 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, portableapplications (Mobile Apps) have turned out to be basic piecesof 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 isaccounted 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 accountsan extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application surveys as a noteworthyimpact 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 andwould profit by understanding a huge number of printedremarks.

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 Almabetter, which is the world's largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched 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:

- ## Play Store Dataset

- **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 issimplest programing 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 reasonsto 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 filllna() 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 as type(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.

- **Step 9:** We write a function Ur 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 User review dataset.

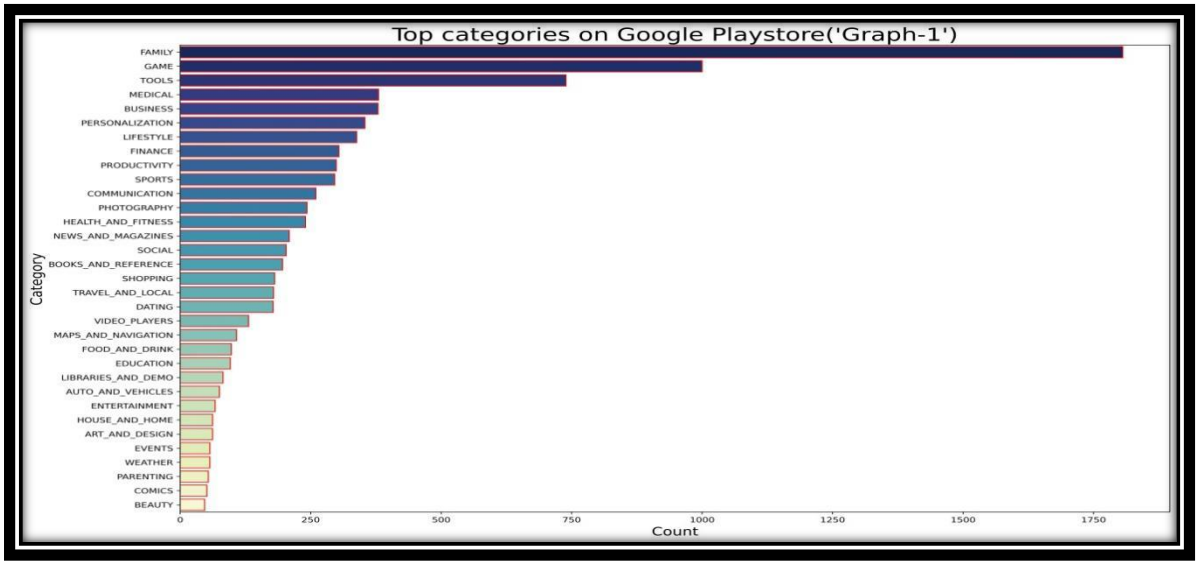
- **Step 10:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using dropna() function.

3 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

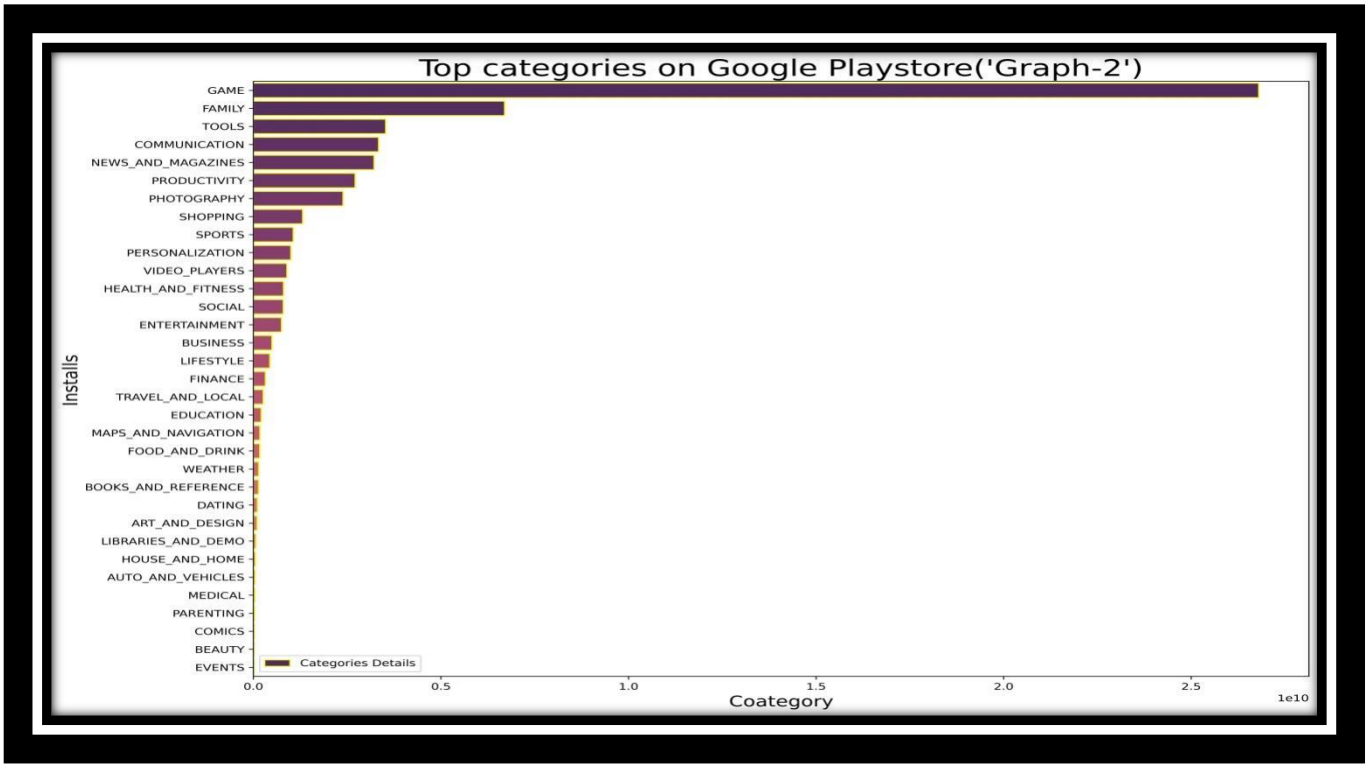
EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

3.1 TOP CATEGORIES ON GOOGLE PLAY STORE



So, there are all total of 33 categories in the dataset from the above output we can come to the conclusion that in the play store most of the apps are under Family & Game category and least are of Beauty & Comics Category.

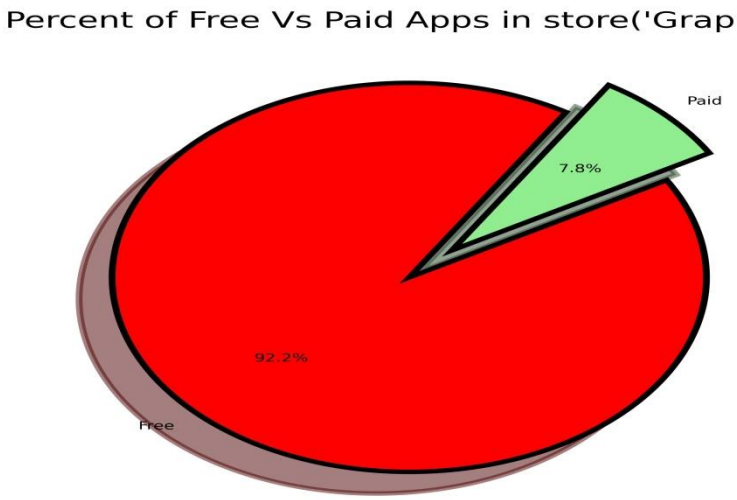
3.2 WHICH CATEGORY APPS HAVE THE GREATEST NUMBER OF INSTALLS?



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

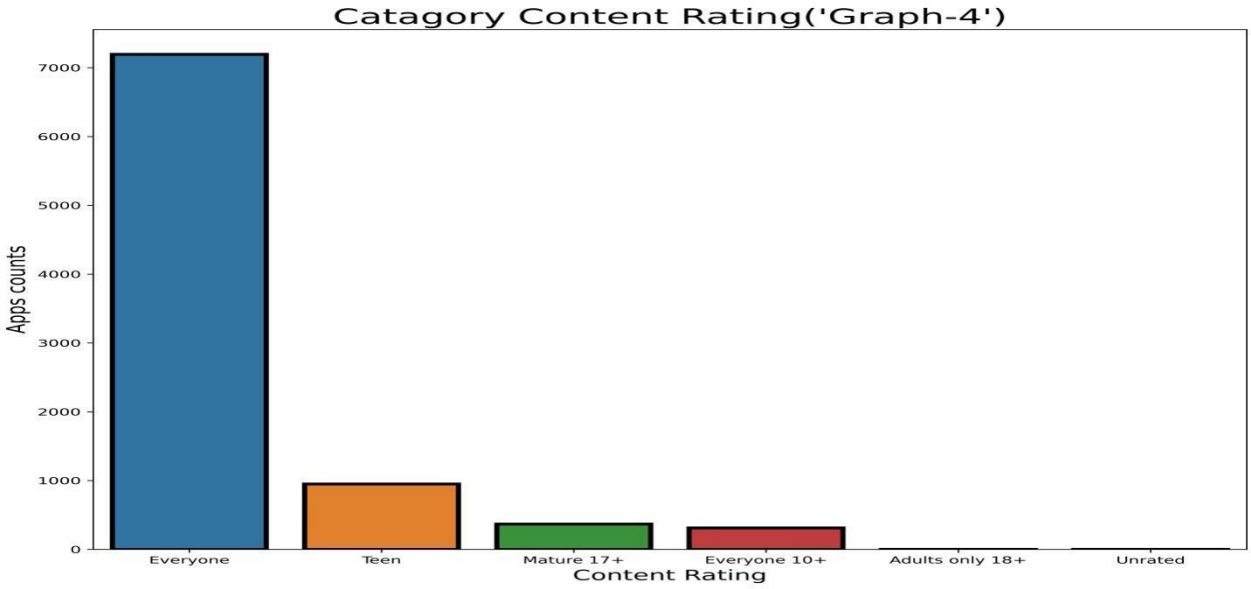
From the above visualization, it can be interpreted that the top categories with the highest installs are Game, Family, Tools, Communication, News & Magazines.

3.3 How much Types apps category percentage are paid or free



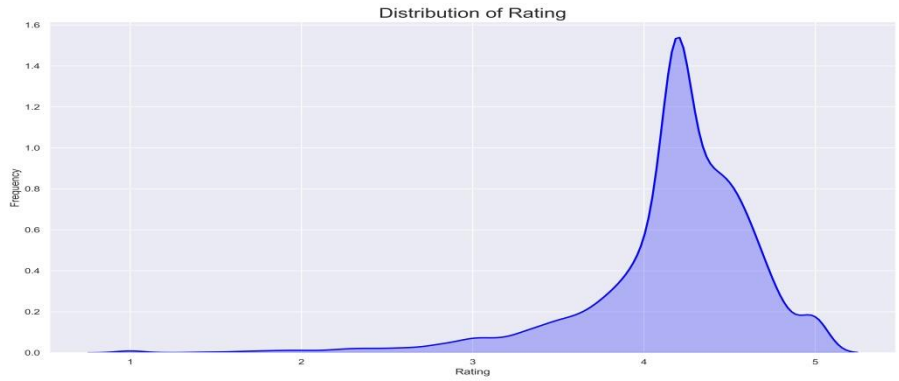
Here we can see that 92.6% apps are free, and 7.4% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

3.4 Which category of Apps from the 'Content Rating' column is found more on the play store?



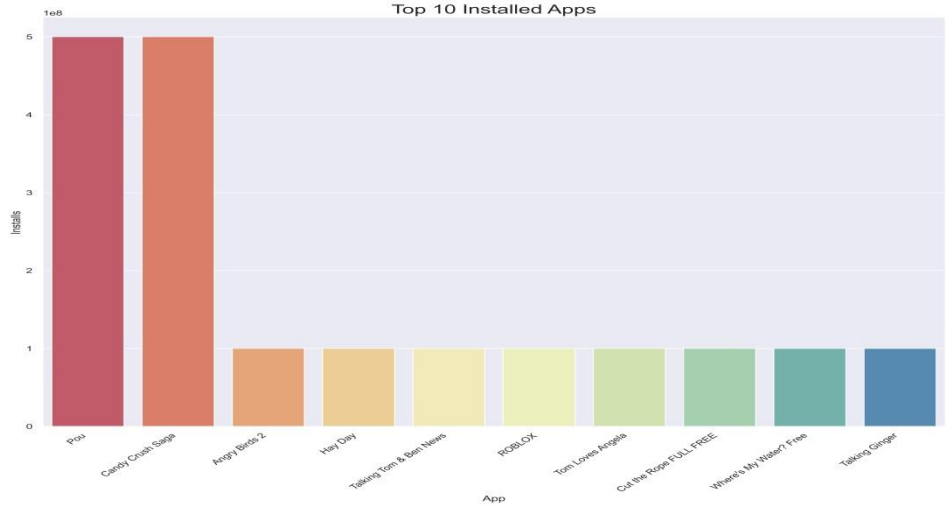
A majority of the apps in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

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 3.5 and 4.8.
- From the kde plot visualizations, it is clear that the ratings are left skewed.

3.6 What are the Top 10 installed apps in any category ?

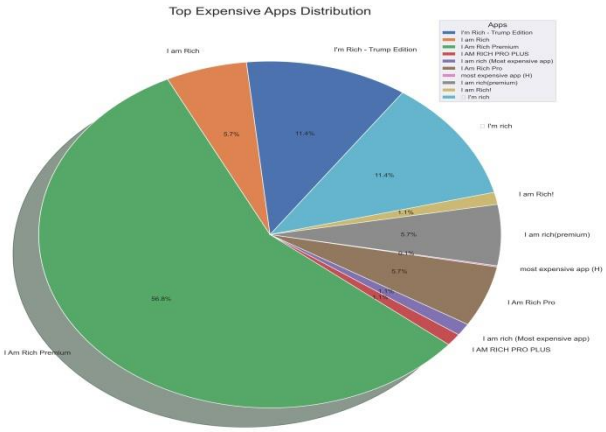


- So, we have to be able to answer this not only for a single category but for many, i.e., we will need to define a function which should be able to return us a nice plot for any Category the name provided by any user as an argument to it.

```
def findtop10(str):
selected_category=clean_df[clean_df["Category"]==str]
top10installed_selected_categorey_app=selected_category.sort_values(by="Installs",ascending=False).head(10)
return top10installed_selected_categorey_app
```

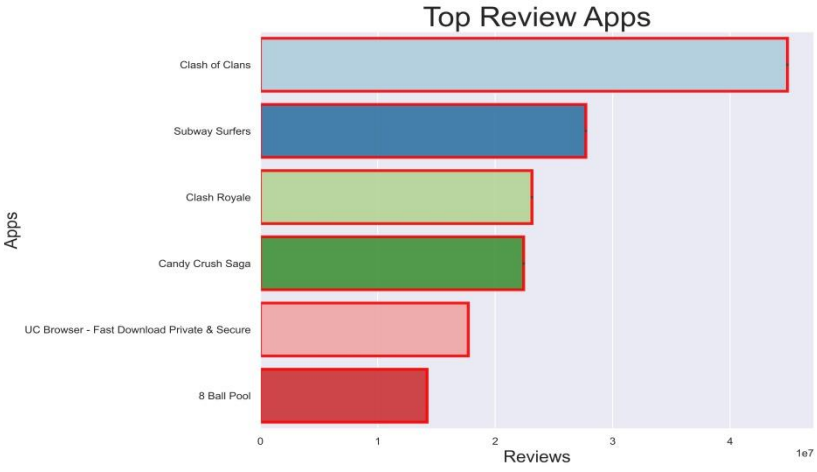
- From the above graph we can see that in the FAMILY category POU, and CANDY CRUSH SAGA has the highest installs. In the same way we by passing different category names to the function, we can get the top 10 installed apps.

3.7 : Which are the top 10 expensive Apps in play store?



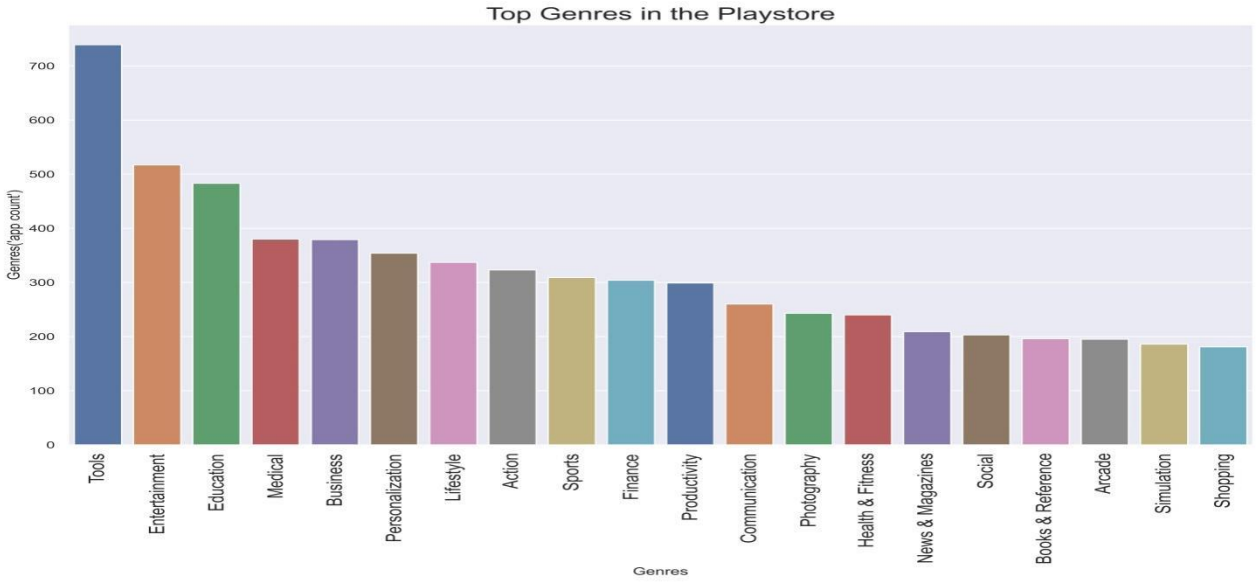
- From the above graph we can interpret that the App I Am Rich Premium is the most expensive app in the GOOGLE play store followed by I am Rich. I Am Rich Premium 56.8% app almost expensive. we also had to drop one row data for this visualization because the language of the app was Chinese and it was messing with the pie chart, In this data frame under 9934 row labels "I'm Rich/EU SOU Rico//أنا غني我很有錢" in this app zero(0)times install that case this unwanted visualization

3.8: Top the Apps with highest number of reviews?



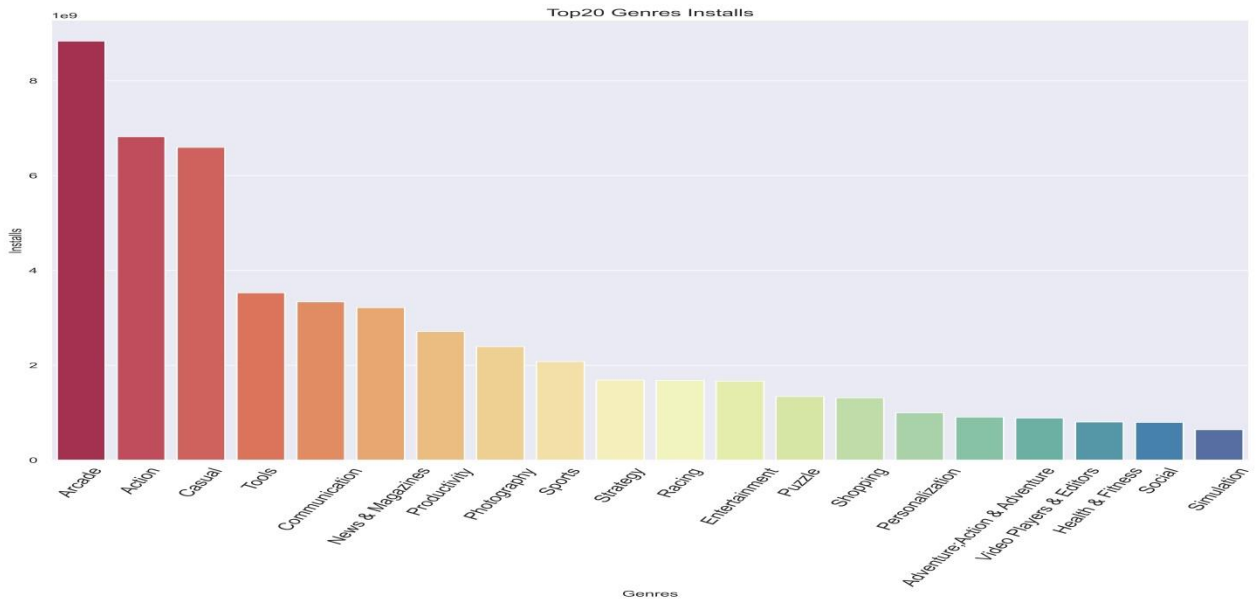
- From the above data frame we can interpret, and come to conclusion that the Apps like Clash of Clans, Subway Surfers, Clash Royale, and Candy Crush Saga , UC Browser - Fast Download Private & Secure,8 Ball Pool, has the highest number of reviews on GOOGLE play store. This top apps under ,Clash of Clans, Subway Surfers are most reviews app. Clash of Clans 44893888 numbers of reviews apps and Subway Surfers 44891723 numbers of reviews app.

3.9. What are the count of Top20 Apps in different genres?



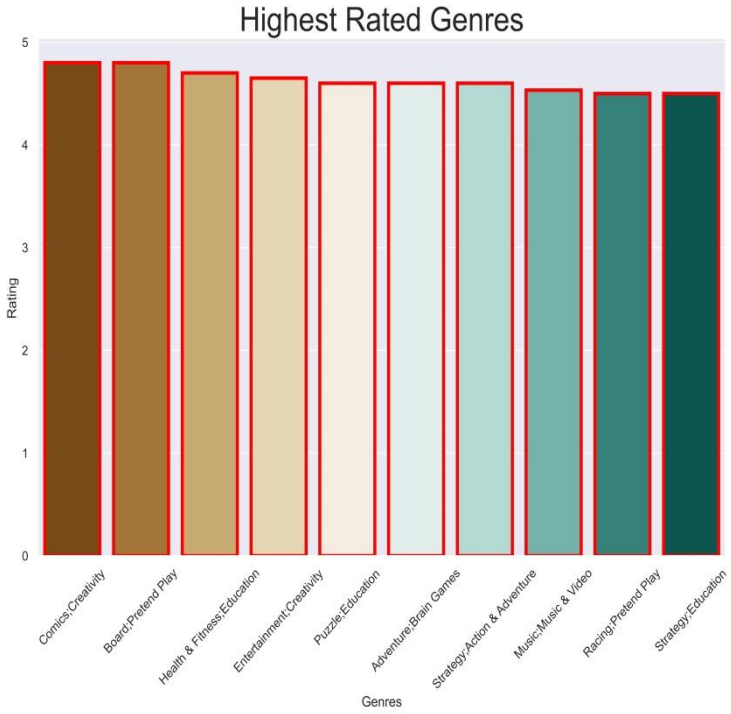
After visualization we can see that the Highest Number of Apps found in the Tools and Entertainment genres followed by Education, Medical and many more .Tools Genres app count 739, Entertainment Genres app count 517, and Education Genres app count 483.

3.10. Which are the Genres that are getting installed the most in top 20 Genres?

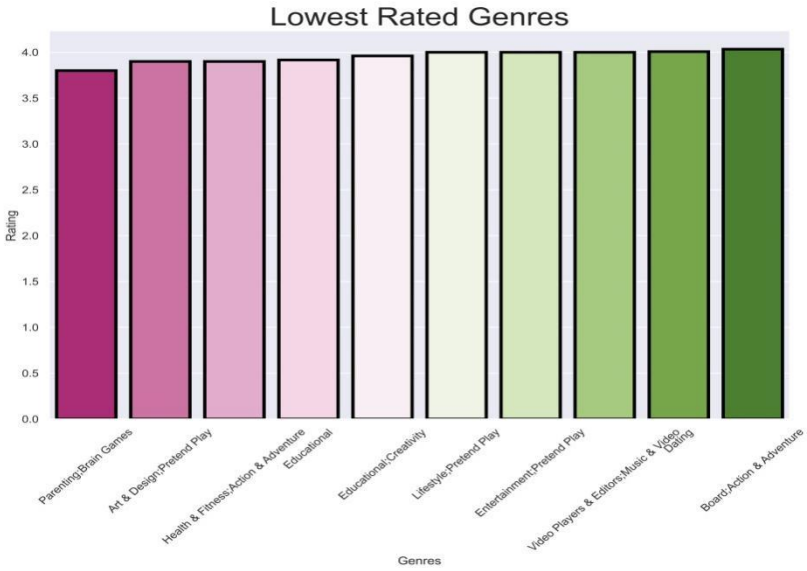


- After visualization we can come to the conclusion that maximum app install comes under Arcade Genres and followed by Action, Casual and Tools Genres, Arcade Genres install 8836079153 times, and
  Action Genres most install times 6818939040.

3.11. Find the highest and the lowest rated Genres

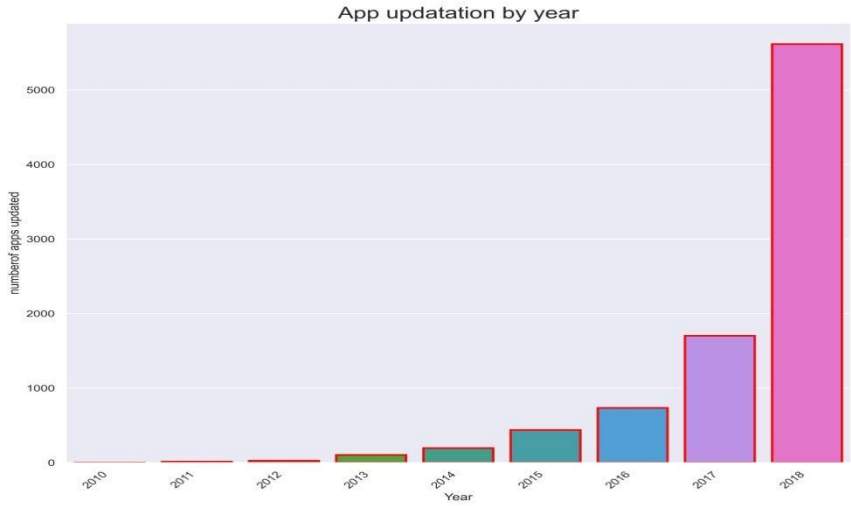


- From the above graph we can see that Comics : Creativity and Board - Pretend Play are the highest rated genres. Comics Creativity Genres rating is 4.8000, Board : Pretend Play rating is 4.8000.



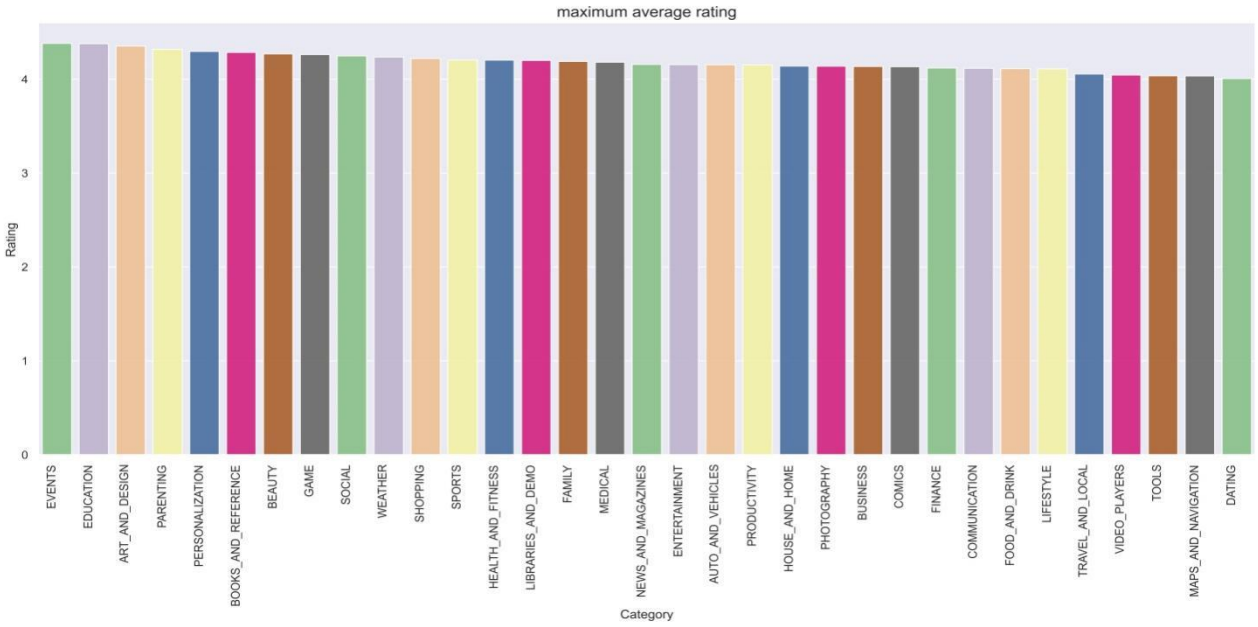
- Graph we can see that Parenting Brain Games is the lowest rated genres. Brain Games is the lowest rated genres this rating is 3.800

3.12. What year more Apps update details "By Year”



- From this plot we can see that a very wide range of app updated in play store during 2017-2018, actually in 2017-2018 under Almost More of the apps will update new version and in this 2018 year 5615 number of apps will be updated, and in this 2017 year 1720 numbers of app will be updated.

3.**13 (i) Before Merging Which Category has highest number of average rating?**

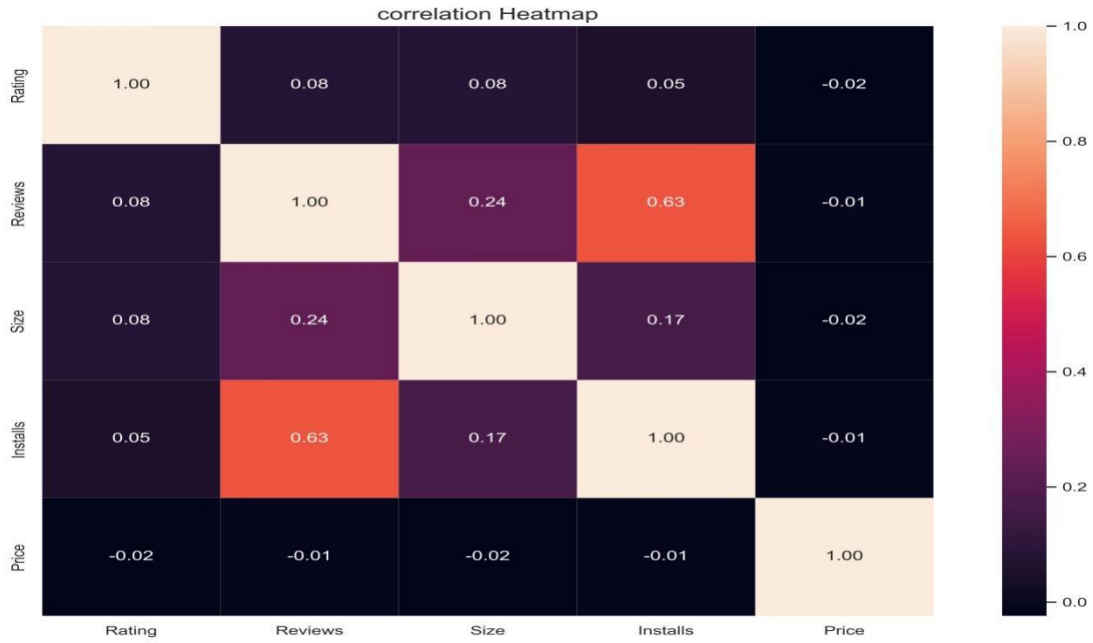


- From the above graph we can see that in the "EVENTS" category has the highest average rating. Category of EVENTS Rating 4.381924, and Category of EDUCATION Rating 4.377999. before merging play store category rating all 33Categories average rating is 4.183

3.13 (ii).User reviews after merging Which Category has highest number of average rating?

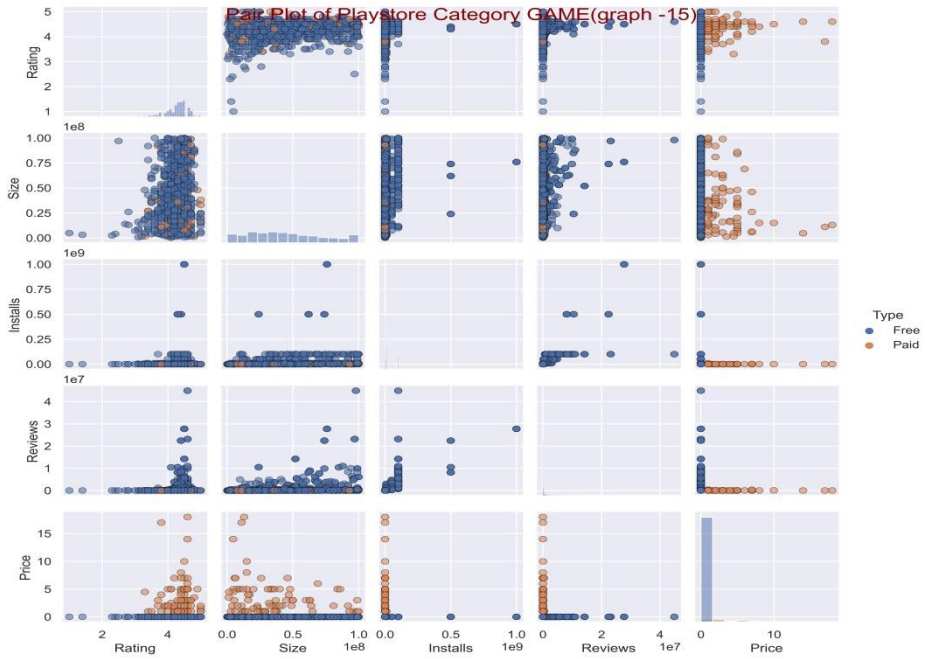
- After merging User reviews than graph we can see that in the "COMICS" category has the highest average rating. Highest COMICS Category rating is 4.7000, and lowest ENTERTAIMENT category rating is 4.05642. User reviews after merging all 33Categories Average rating 4.274.

3.14. Correlation Heat map



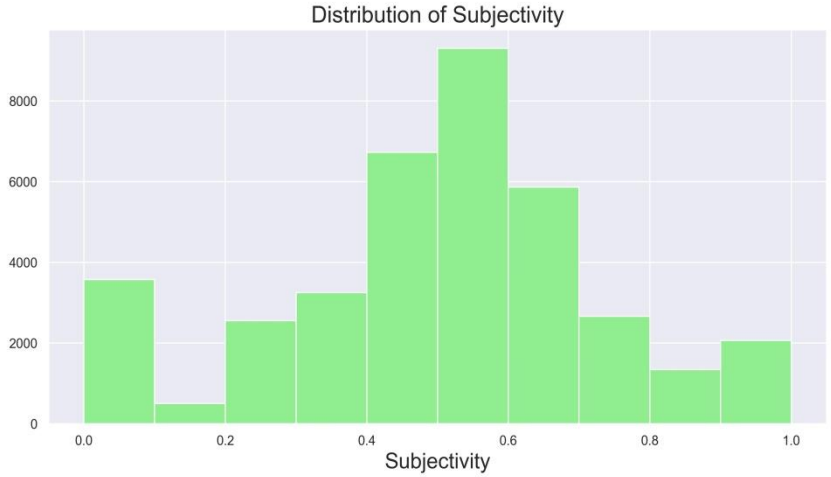
- There is a strong positive correlation between the Reviews and Installs. The Price is slightly negatively correlated with the Rating, Reviews, and Installs. The Rating is slightly positively correlated with the Installs and Reviews.

3.15. Pair Plot



- Game Category Rating, Reviews, Size, Installs and Price Check relationship To Type Free and Paid.

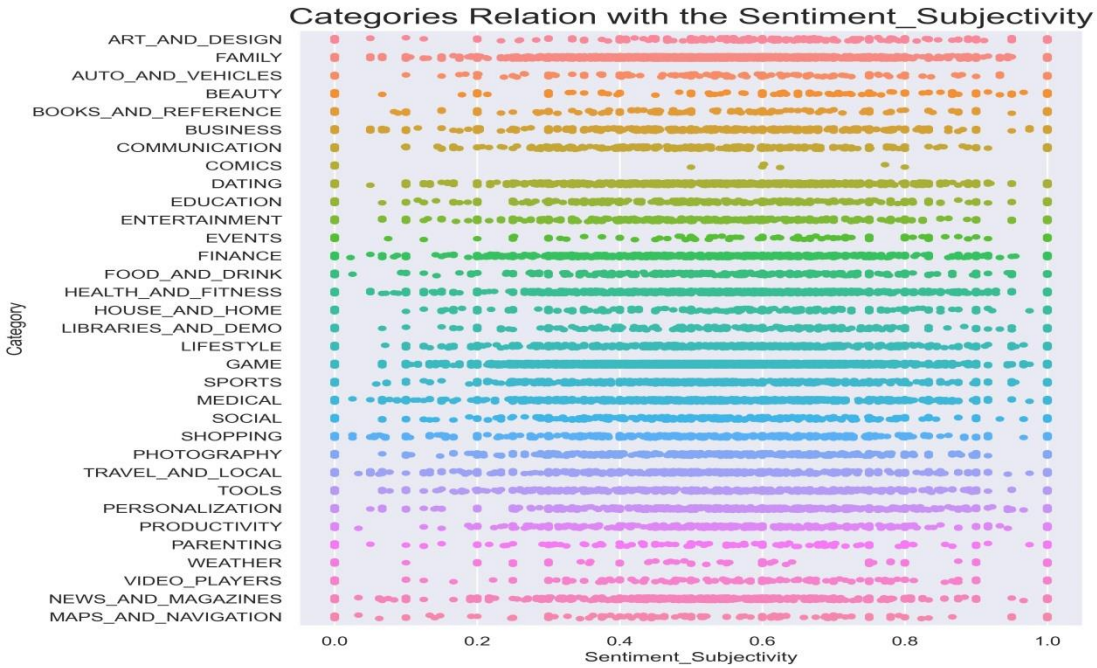
3.16. Distribution of Sentiment subjectivity



0 – objective (fact), 1 – subjective (opinion)

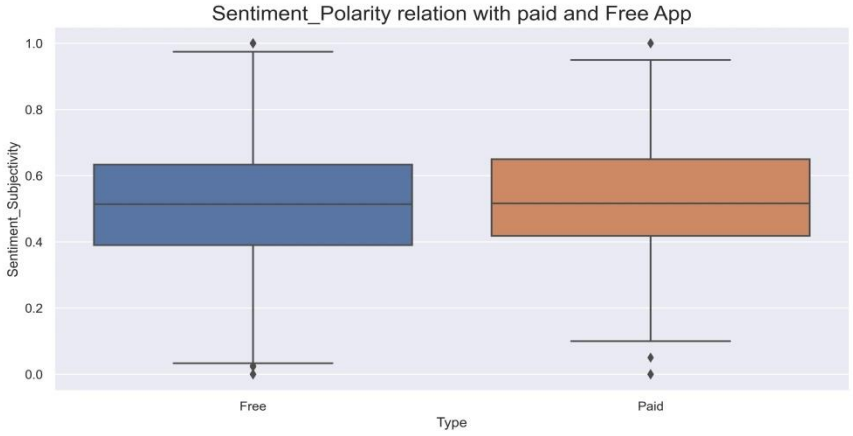
- This graph it can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7.
- So we can conclude that maximum number of users give reviews to the applications, according to their experience

3.17 Categories Relation with the Sentiment Subjectivity



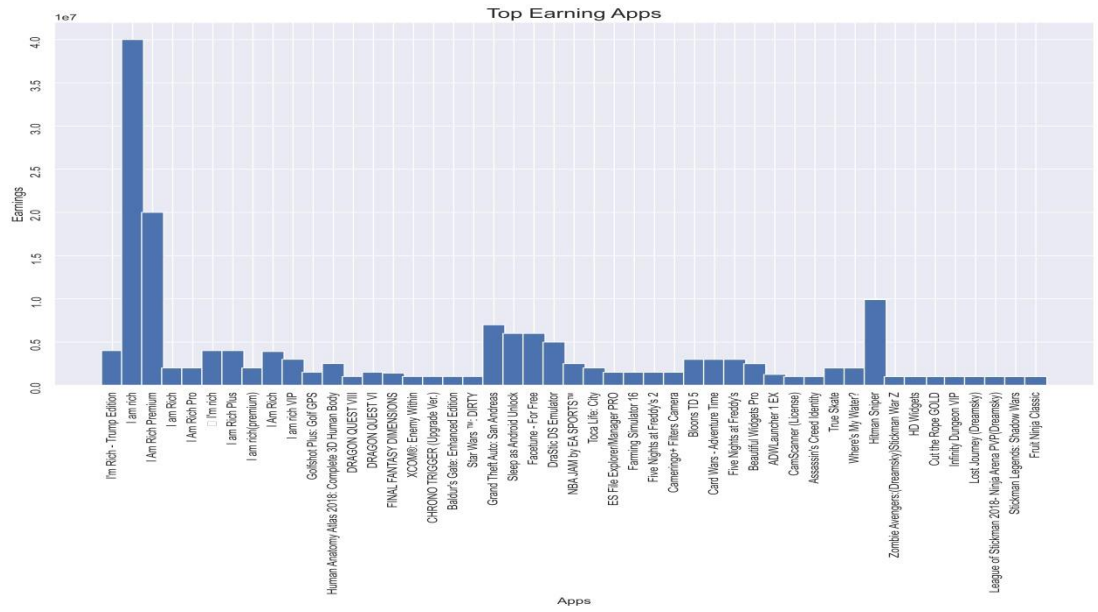
- User sentiment subjectivity ART-AND-DESIGN, FAMILY, play store App category Relation will be check here.
- From the above strip plot
- It can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behavior, when variance is too high or low.

3.18. Sentiment _ Polarity relation with paid and Free App



- In this graph we can come to the conclusion that , Paid type Apps more than Free type Apps Sentiment Subjectivity

3.19 Which are the apps that have made the highest earning?



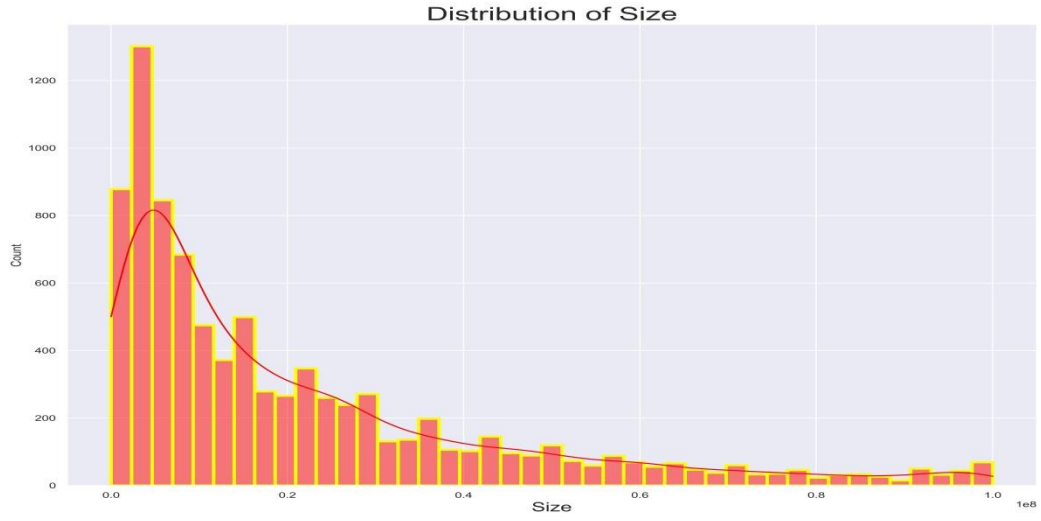
Find Earning generated is given by the formula -

Earning = Installs * Price

The top four apps with highest earnings found on GOOGLE Play store are:-

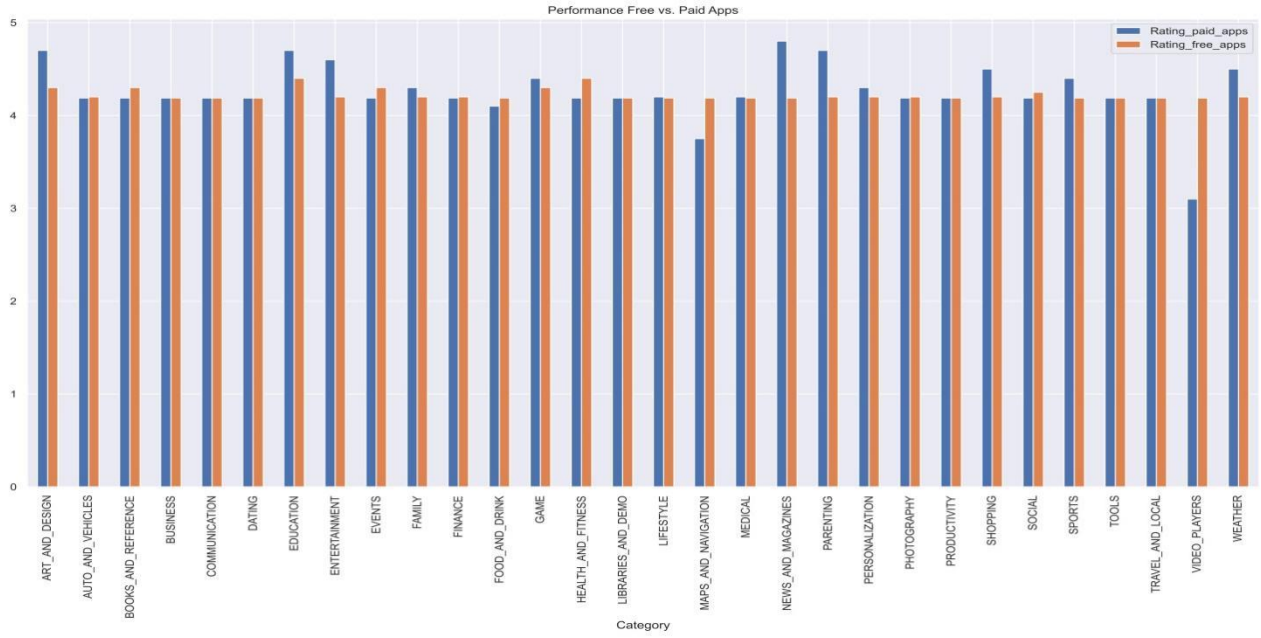
* I am Rich
* I am Rich Premium
* Hit man Sniper
* Grand Theft Auto: San Andreas

3.20. 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

3.21. Are Paid apps worth buying? (Analysis based on Average User Rating)



- It is using Bar plot Looks like paid apps perform marginally better than the free apps, check category under high rating paid apps and free apps and check category under low rated paid apps and free apps

- Let's check High rated paid Category apps NEWS_AND_MAGAZINES Rating 4.800000 and high rating free apps EDUCATION Ratings 4.400000 and low rated paid Category apps VIDEO_PLAYERS Rating 3.100000 and low rated free category apps HOUSE_AND_HOME Rating 4.187877.

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.

3.1) In play store present most of the apps are under Family & Game category and least are of Beauty & Comics Category.

3.2)Category with the highest number of installs is Game and least number of installs i s EVENTS categories.

3.3) 92.2% apps are Free and 7.8% apps are paid in type.

3.4) Most of the apps in the Google play store are rated between3.8 to 4.8.

3.5) Play store maximum number of apps are available for Everyone and then for Teen Everyone content Rating 7195 and Teen Content Rating 915.

3.6) FAMILY category POU,and CANDY CRUSH SAGA has the highest installs 5 Billions times install in two apps.

3.7) In Play store most expensive app is
I'm Rich Trump Edition price actually 400.00 this install only 10000 times but I Am Ri ch Premium is price 399.99 but this maximum time install 50000 times.

3.8) In This play store Game category belong Clash of Clans game is most reviews Estimate 44893888 reviews

3.9) Top20 Apps in different genres highest genres count Tools 739

3. 10) The Genres that are getting installed the most in top 20 Genres Arcade Genres 8836079153 times installs this is the highest install in Play store.

3.11) In Genres highest rating genres is Comics Creativity is Rating 4.800000
And lowest rating genres is Parenting Brain Games Rating 3.800000

3.12) App updated in play store during 2017-
2018 in 2017 update 1702 apps and in 2018 5615 apps will be update.

3.13) The "EVENTS" category has the highest average rating. "EVENTS" category rati ng is 4.381924, before merging play
store category rating all 33Categories average rating is 4.183

After merging User reviews than the "COMICS" category has the highest average rating . Highest COMICS Category rating is 4.7000, and lowest ENTERTAIMENT category Rating is 4.05642. User reviews after merging all 33Categories Average rating 4.274.

3.14) Correlation heatmap

The Price is slightly negatively correlated with the Rating, Reviews, and Installs. The Rati ng is slightly positively correlated with the Installs and Reviews.

3.15) Pair plot Game Category Rating, Reviews, Size, Installs and Price Check relationship To Type Free and Paid.

3.16) sentiment subjectivity lies between 0.4 to 0.7.
So we can conclude that maximum number of users give reviews to the applications, according to their experience.

3.17) In this chart show that Sentiment
Subjectivity in all category GAME and FAMILY is underlying distribution is maximum

3.18) Free type app compare to Paid type app high Sentiment Subjectivity.

3.19)In this Google Play store I am Rich

I am Rich Premium
Hit man Sniper
Grand Theft Auto: San Andreas are top four highest expensive app

3.20)The conclusion that maximum number of applications present in the dataset are of small size.

3.21)Looks like paid apps perform marginally better than the free apps, check category under high rating paid apps and free apps and check category under low rated paid apps and free apps high rated paid Category apps NEWS_AND_MAGAZINES Rating 4.800 000 and high rating free apps EDUCATION Ratings 4.400000 and low rated paid Categ ory apps VIDEO_PLAYERS Rating 3.100000 and low rated free category apps HOUS E_AND_HOME Rating 4.187877.

Thank You