Abstract

Google play store is one of the largest and most popular Android app stores. It contains huge amount of data that can be used to make an optimal model and can be used for identification of trends and future challenges. In this EDA project we have used two raw data sets of Google Play Store one is of play store attributes and other is of user reviews from Kaggle. The first dataset contains 13 different attributes and the second dataset contains the 5 other features that can be used for data manipulation and its analysis.

EDA is the first and most important step to solve any data science problem.It also gives valuable insights into the pattern and information it has to convey. It is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.The best EDA gives the interesting insights about your data.We have followed all the steps which are essential while doing exploratory data analysis(EDA). After cleaning the data we have plotted various graphs using different python libraries such as matplotlib and seaborn that show relation between different attributes and we have successfully gained some good insights with it that can help the industries to capture the android market.

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

1. Introduction

Google Play is a website and app that gives users access to download and purchase apps, books, games, movies, music, and other content from Google for Android devices.

Android is the dominant mobile operating system today with about 85% of all mobile devices running Google's OS. The Play Store is the largest and most popular app store for android.

The Google Play Store started life as the "Android Market" in 2008. It launched alongside the very first Android devices, and its purpose was to distribute apps and games.The Android Market at the initial days didn't support paid apps and games until 2009. However, as the Android platform grew, so did the Android Market. By 2012, it featured over 450,000 Android apps and games. In this article, we seek to shed light on the dynamics of the Google Play Store and how we can use different features from this data set for prediction purposes.

The main aim is to do this with the help of a sentimental analysis and data visualization that will analyze customer needs and suggest the developers best app for developing. The analysis is to be done using the survey of the user download behavior on the apps across all the categories on the google play store. Mobile app stores are becoming extremely lucrative. Android is expanding as an operating system and Mobile app industry is increasing in significantly and thus giving rise to more competitions to the ones who are new to the industry and that are creating applications. Hence, for a developer to know the recent trends, competition is important so that the value of their app in the store does not degrade. Google play store is a digital distribution service and it allows users to browse and download different apps. It is the official store of apps for the android operating system. Play store is additionally a platform which offers music, digital media, books, movies and tv programs. Due to the competition in the market and also expansion of the play store,

in order to help our developer understand what kinds of apps are likely to attract more users and what is the motivating factor for the people to download an app for example some people have motivation factor as

1.Low price,

2.Small size of apps,

3.High user reviews and rating

We analyze and research relevant data. They will be getting to know the success rate and they will get to decide what features should be added or modified and what should be maintained according to the current trend and future prediction on an app. Hence we found this topic interesting and convincing for our project work.

Analysis Methodology

Our Analysis is divided into four phases:

Data extraction

Data cleaning

Data visualization

Interpret results

Conclusion

First, we collect the data from the given data set, in the form of a csv file of both the data set i.e play store data and review data. In the next step, we try to do data cleaning on the data set to reduce the error percentage in the data set which includes null values and duplicate values. After the data set is ready, we try to analyze the data set using different plots and remove the stuff not needed from the data set. The last step includes the use of different classification algorithms and visualization of models on the data set to see which one gives the highest percentage of accuracy. Finally, we narrate the analysis results to provide a clear vision of the relationship among the areas of interest.At the end we include a detailed discussion of the applicability and future research directions called Conclusion and future work.

Google Play Store Dataset

1. **App**- It tells us about the name of the application.
2. **Category** - It tells us about the category to which an application belongs.
3. **Rating**- It tells us about the ratings given by the users for a specific application.
4. **Reviews** - It tells us about the total number of users who have given a review for the application.
5. **Size**- It tells us about the size of the application on the mobile phone.
6. **Installs** - It tells us about the total number of installs/downloads for an application.

7. **Type** - It tells us whether the application is free or a paid one.
8. **Price**- It tells us about the price of the application.
9. **Content_Rating**- It tells us about the target audience for the application.
10. **Genres** - It tells us about the various other categories to which an application can belong.
11. **Last_Updated** - It tells us about when the application was updated.
12. **Current_Ver** - It tells us about the current version of the application.
13. **Android_Ver** - It tells us about the android version which can support the application on its platform.

Problem Definition:

The Play Store apps data has enormous potential to drive the app-making businesses to success .Android is expanding as an operating system exponentially and Mobile app industry is increasing significantly and thus giving rise to more competitions to the ones that are creating applications and which are new to this industry .Due to the heavy competition in the market and also expansion in order to help our developer understand what kinds of apps are likely to attract more users and what is future demand of app category wise and what is the motivating factor for the people to download an app. We analyze and research the data from the past years. Which will help in analyzing market trends and peoples approach to install applications. This will be helpful For the app development industry where they can analyze the downloads and demand of an app in the category which they want to enter or which industry is to be explored in the forthcoming years to capture good market share.

The Problem statements are:

1. What are the top categories on the Play Store?
2. Which category has the most no. of installations?
3. How does app rating affect the application?
4. How Size, Reviews, Installs and Price of apps are correlated?
5. What is the Relation between app category and app price?
6. How does the last update have an effect on the rating?
7. How does sentiment polarity and subjectivity affect the size,rating,and review?
8. What is the average rating of apps
9. Are the majority of the apps Paid or Free?
10. How are Installs and Rating correlated?
11. . Which are the most expensive apps and name their corresponding categories?
12. Which are Top revenue generated apps and what are their corresponding categories?
13. What is the relation between sentiment polarity and sentiment subjectivity?
14. How Content rating affects number of installs?
15. How are ratings affected when the app is a paid one?
16. How does the size of an app affect the ratings and number of installs?

2. Data Cleaning and Preparation

Data cleaning and preprocessing is an important task. It transforms raw data into a more understandable, useful and efficient format.

2.1 Why is Data Cleaning and Preprocessing required?

Data can be noisy i.e. the data can contain outliers, null or simply out of the trend values. So It's necessary to treat them before the exploratory data analysis.

2.2 Available Dataset:

The given dataset Play Store Data and User Reviews 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.

2.3 Steps Performed:

Step1:- We wrote a function getinfo (), that will display 6 attributes about all the columns: Data type, total number of records, Count of null values, Count of non null values, percentage of null value and unique values in that columns in the play store dataset and user reviews dataset.

Step2:- From step1, we observed that the "Rating"column has the most number of null values. Since the number of missing values in this column is approx 15% of the total records. So, it's not a good idea to drop the null values for the "Rating"column. To deal with null value in this case, we replaced the null with mode of the Rating Column.

Step3:- There were some special characters in the "Installs"column like $ and + and MB and KB was represented as M and k in the size column. Using a for loop, we removed these special characters + and $ from the Price and Installs column respectively. For the M and K , we converted KBs into MBs using a function so that our data is consistent. We dropped other rows where value is null since those records are very few. So it will not affect the dataset

Step4:- 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.

Step5:- Using the getinfo() function, we checked , there were no null values now. But there were some duplicates, We have dropped the duplicates from both the data set.

3. Exploratory Data Analysis

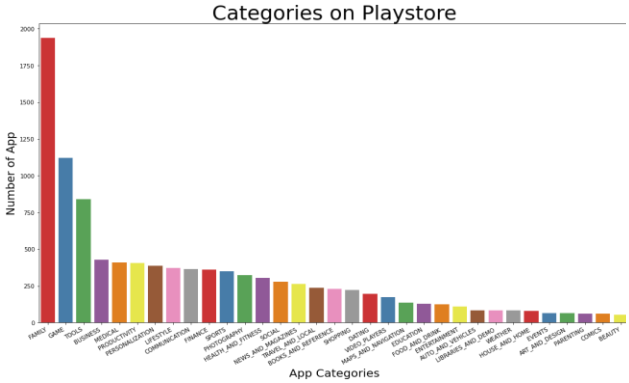
Exploratory Data Analysis(EDA) is the very first and most important step in every Data Science project. After cleaning the raw data EDA involves the process of discovering the patterns which may give us good insights and that can help in future challenges. EDA is the process of analyzing the dataset to discover trends, and outliers, and form hypotheses based on our understanding of the dataset.

EDA also involves creating statistics for numerical data in the dataset, finding correlation between numerical values and creating various graphical representations, plots by using visual methods to understand the data in a short and better way. In this article, we will understand EDA with the help of a Google play store dataset. We will use python and its essential libraries and will try to generate some good insights with it.

3.1 Top Categories in play store

We know that all apps belong to a particular category on play store. On the basis of our analysis we have observed that the "Family"

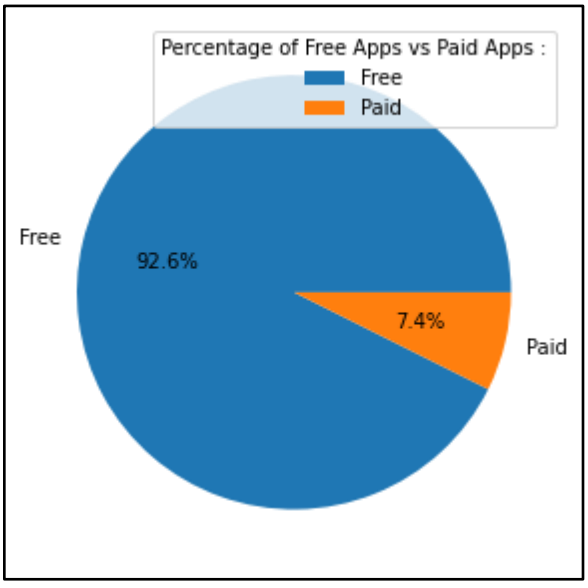
category contains the highest number of applications on play store.



From the above plotted bar graph we conclude that out of 33 categories, the top most category is “Family” and least is “Beauty”.

3.2 Paid vs Free Apps

We all know that the majority of the audience prefer to use free apps instead of paid one. So we have created the pie chart to determine the percentage of the paid and free apps that playstore have.

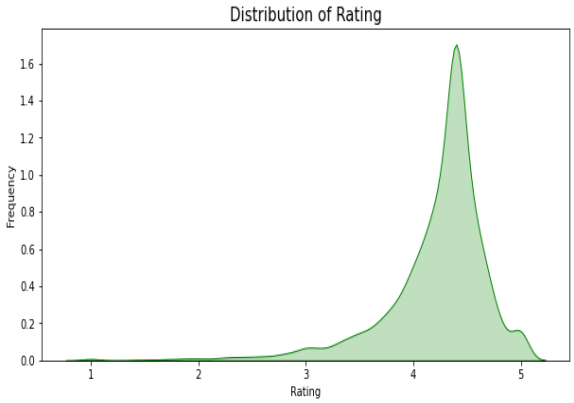


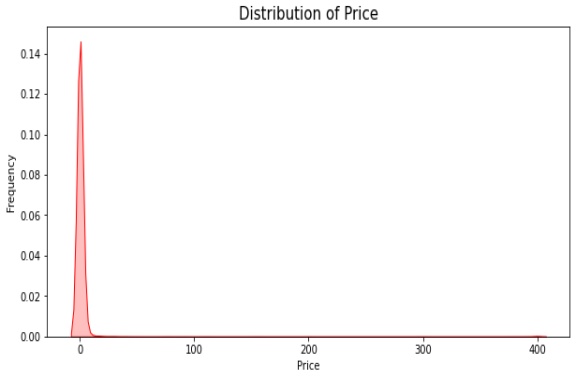
We can clearly see that there are only 7.4% paid apps available on play store and the rest 92.6% are free.

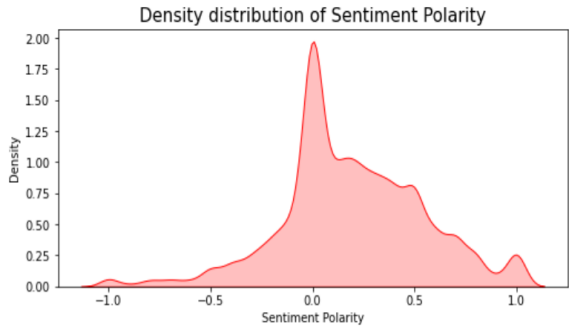
3.3 Density/Frequency Distribution

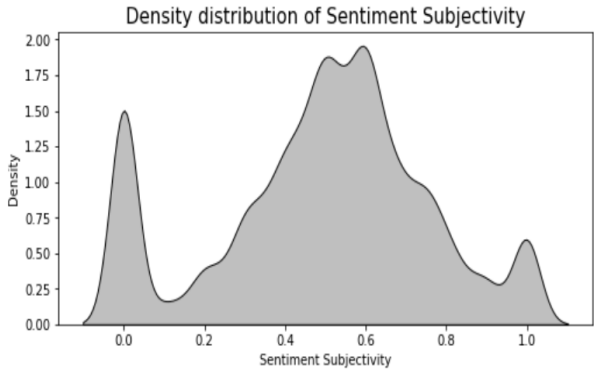
A density plot is a representation of the distribution of a numeric variable that uses a kernel density estimate to show the probability density function of the variable.Density Distribution graphs shows the range where the maximum value lies.

In our analysis we have plotted the density distribution graphs of rating, price, sentiment polarity and sentiment subjectivity.



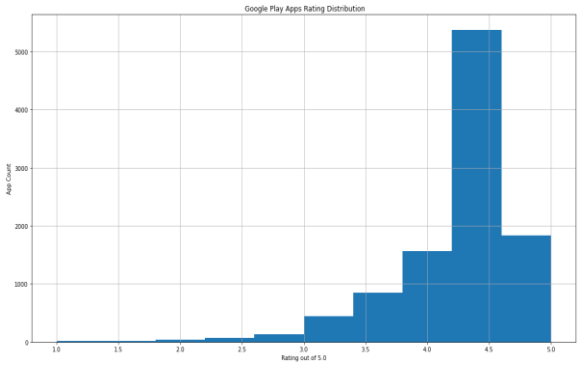






From the above density distribution plots we can say that:

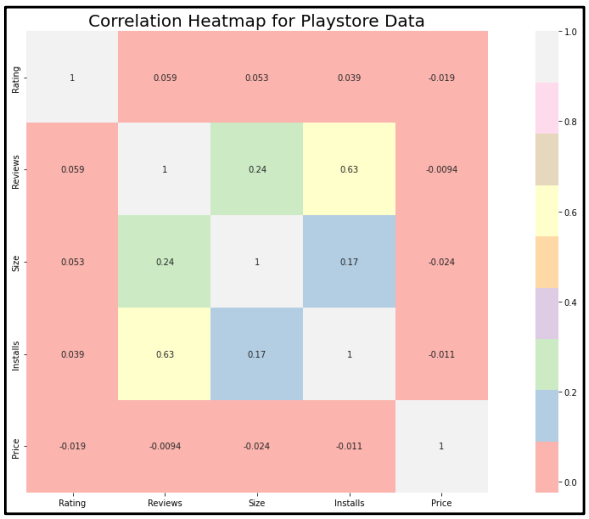
1. **Mean rating** of our dataset is **4.2178813** on the play store.
2. **Mean price** of our dataset is **1.0315609** which means that the majority of the apps are **free (0$)** on google play store.
3. The polarity score lies in the range of **[-1,1]**. Anything below a score of -0.05 we tag as negative and anything above 0.05 we tag as positive Sentiment score. Here we can see from our calculations and the Sentiment Polarity Density distribution graph that the **Mean Sentiment Polarity Score is 0.18886801** which resembles a good average sentiment score(Majority of the users liking the apps).
4. The subjectivity is a float within the range **[0.0, 1.0]** where **0.0 is very objective and 1.0 is very subjective.** As per our analysis and plotted graph the **Mean Sentiment Subjectivity Score is 0.49093045.** That means around 50% users are sharing personal opinions while others 50% are just sharing the factual information in reviews.



3.4 Correlation Heatmap

Correlation heatmaps are the plot that visualize the strength of relationships between numerical attributes.

3.4.1 Correlation Heatmap of Play Store Data:



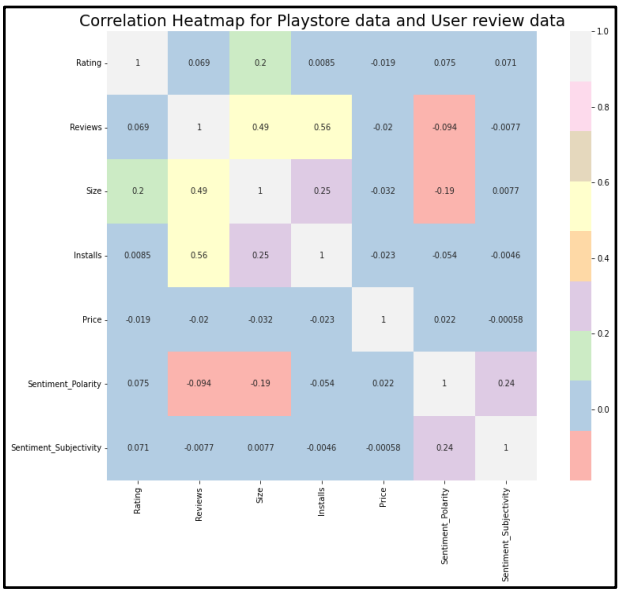
From the above heatmap we have concluded that:

1. There is a positive correlation between the Reviews and Installs column i.e (0.63). Higher the number of reviews ,higher the total number of downloads. Which means customers download a given app more if it is reviewed by large no.of people.
2. The Price is slightly negatively correlated with the Rating, Reviews, and Installs. Which means if the price of an app increases, the average rating, total number of reviews and Installs fall down.
3. The Rating is slightly positive correlated with the Installs and Reviews column.

This shows that if the Rating of an app increases then it will also increase the downloads and reviews of a given app.

4. Hence we can conclude that increasing the review and Rating count in the app may increase the market share.
5. To capture the market more rapidly The launch price of an app should be less at the starting and then can be increased with time.

3.4.2 Correlation Heatmap for Play Store Data and User Review Data:



While doing our analysis we were curious how both the data frames are related to each other. Is there any positive or negative correlation between them? For that purpose we have merged both the data frames for obtaining results and we have observed that:

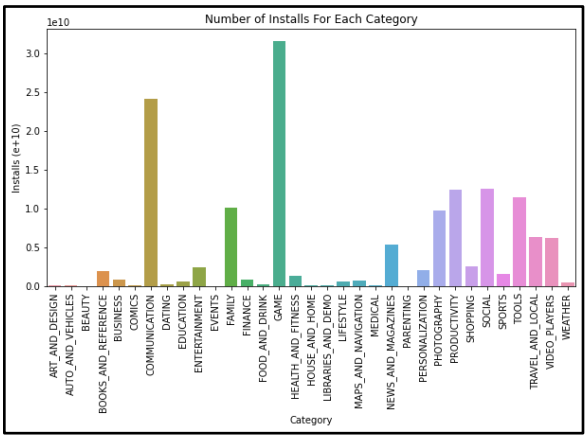
1. Size and sentiment polarity are negatively correlated(-0.19). There may be a reason when the size of an app increases people start disliking the app because it consumes more storage, takes more RAM and needs a high speed connection for its execution.
2. There is a positive correlation between reviews and number of installs(0.56) because as the reviews increase, people start noticing the app and install them.
3. There is a slightly positive correlation(0.24) between sentiment polarity and sentiment subjectivity that means if users share the positive reviews (sentiment polarity) then there are many chances that users are sharing their personal opinion and not factual information(Sentiment Subjectivity).

3.5 Highest Installs Category

To maximize the profit, companies want to know what type of apps are trending in the market and which category they are from, so that they will develop their apps from the same category and capture the audience's interest.

While performing the experiment we have plotted the bar graph between category and Installs and we observed that the “GAME” category has the highest number of installs followed by the “COMMUNICATION”.

3.5.1 No of Installs for each category:



Following inferences has been observed from the Category vs Installs bar plot:

1. From the above barplot, we can see that most of the downloaded apps are from the categories of 'Game' and 'Communication'.Earlier, we also saw that the price of gaming apps is the least among all the categories.This can be meaningfully attributed towards the **new wave** of gaming as a career choice.

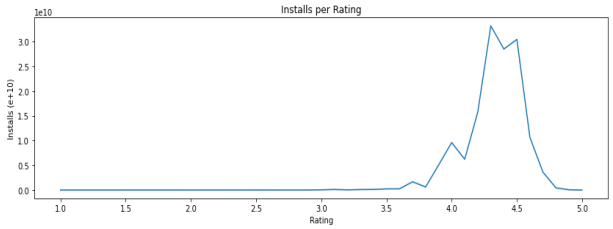
Owing to this transition, companies are more eager to launch products/apps in the gaming category. This can be confirmed with the information that gaming category has one of the highest no of apps on playstore.

2. We can highlight that the business category has very few installs despite listing one of the highest no of apps on playstore.

3.6 Relation between Rating and Installs

This plot is basically oriented towards a marketing perspective. How much a company should invest for a good rating so that the app can capture more audience.

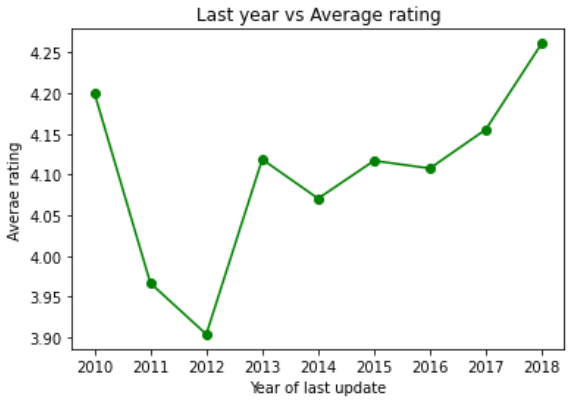
3.6.1 Installs vs ratings



1. Here, we can observe that higher the rating, more the no of installs. **But this correlation slightly changes after 4.5 ratings**. The probable reason could be
    a. Increase in the size of an app
    b. Increase in the price of an app
2. One interesting insight is after 4.0, the ratings are slightly dipped. The reason is the increase in price! But as soon as the prices are decreased, the ratings are again up.

3.7 Effect of Last Updated Year on Rating

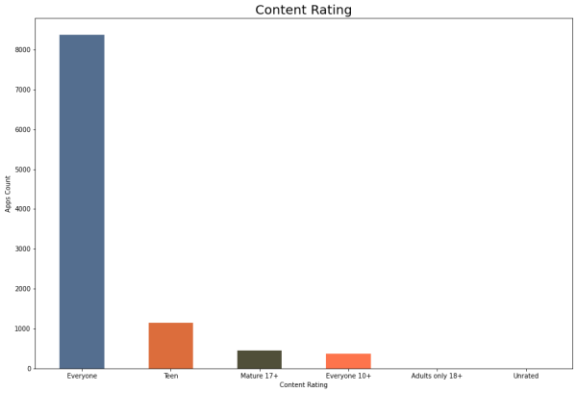
We all personally experience that we do not like any app/companies which do not provide regular updates. So we have plotted the line plot and tried to get some useful insights between the last updated year of an app and its average rating.



From this line plot we have concluded that:

1. Average app rating is gradually increasing after 2016. That means apps that have regular updates have higher average ratings.
2. The graph is falling from 2010 to 2012. This may be due to bad user experience, people are started giving poor ratings due to non regular updates or many other reasons too.

3.8 Content Rating vs Number of App



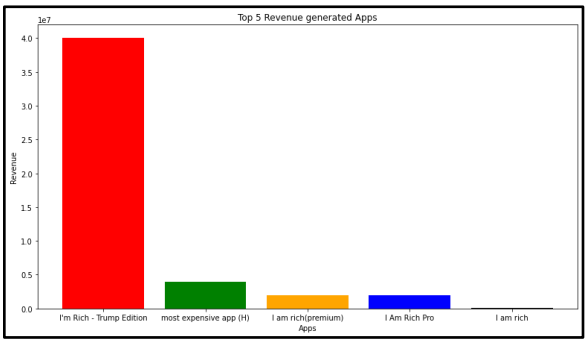
Content ratings in play store are used to describe the minimum maturity level of content in apps. While manipulating and visualizing our data we have observed that:

1. Mostly 90% of total apps are targeting audiences in **every age group** and hence open for everyone.

2. Very few(less than 500 apps) are catering to only the adult population i.e Mature 17+.

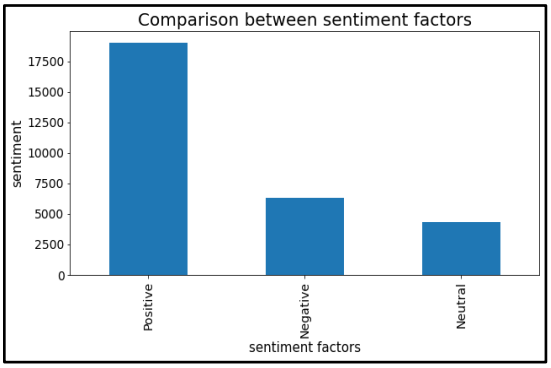
3.9 Maximum Revenue Generated

We tried to utilize the information given in 'installs' and 'price', to calculate the revenue generated by apps. Top 5 revenue-generated apps are **"I'm Rich-Trump Edition" app that has generated the maximum revenue** till now and followed by **"most expensive app (H)", "I am rich(premium)", "I Am Rich Pro", "I am rich".**

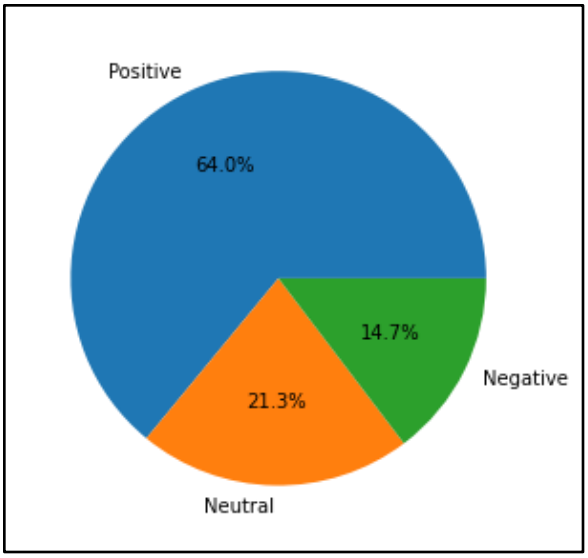


Total revenue generated by these top paid apps are **39999000.0, 4000000.0, 1999950.0, 1999950.0, 39999.0 dollars($)** respectively.

3.10 Sentiment Percentage Of Reviews

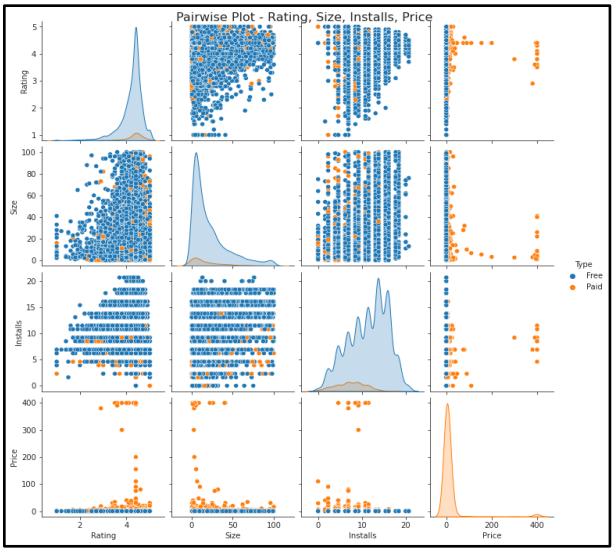


Sentiment plays a very crucial role in identifying whether people are liking the app or not by determining the overall sentiment polarity of the review. We have also determined the percentage of Positive, Negative and Neutral reviews on the play store.



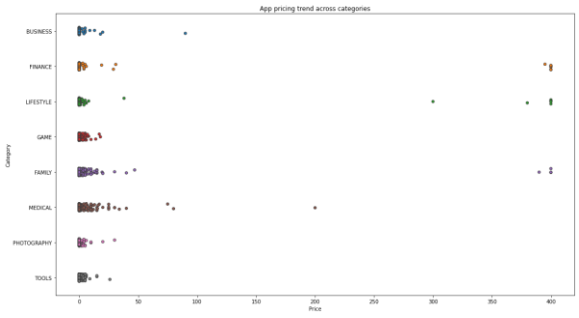
From the above pie chart, we can say that most of the apps on google play store have received positive reviews(64%) by the user, while some of the apps have received negative reviews as well(approx 15%).

3.11 Bivariate Analysis



3.12 Price vs Category

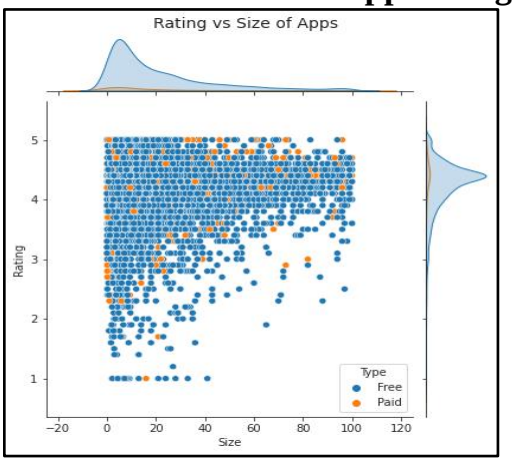
We are very curious to know which category has a higher number of paid apps and what apps people are preferring paid one or free one.



From our analysis from the give datasets we came to the conclusion that:

1. Many factors to be considered when selecting the right pricing strategy for your mobile app. It is important to reevaluate the app price before entering the market.Is it worth the price or not.
2. Here we can see that Different categories of apps demand different price ranges. Some apps that are simple and easy are free, whereas apps in category FAMILY LIFESTYLE FINANCE and MEDICAL are high in price.
3. All Game apps are comparatively low in price, maybe that's the reason game apps have more downloads, as we have seen earlier.

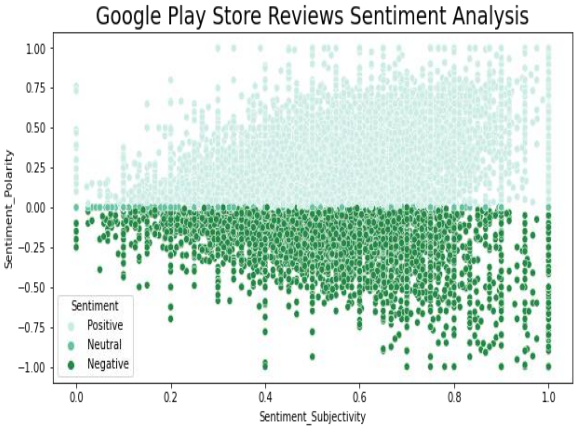
3.13 Effect Of Size On App Rating



It can be observed from the above graph that:

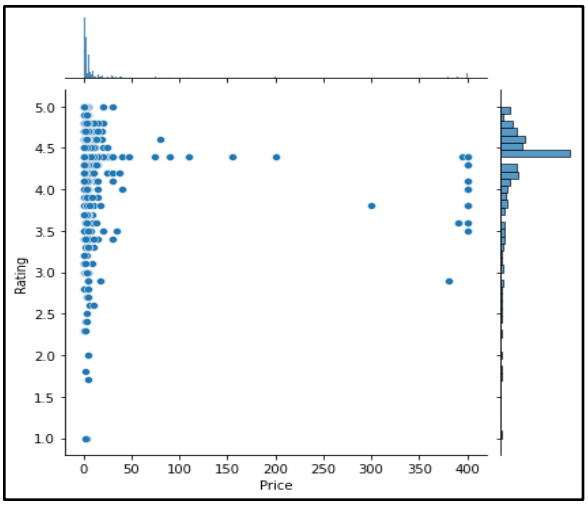
1. People generally prefer apps with less size due to data and/size constraints. As the size of an app increases, its rating decreases.
2. Surprisingly, there are few apps whose size is close to 100 MBs but has 4/5 stars as a review. Though this number is less, we can't ignore the fact that the **rating also depends upon the content that the app** is serving.
3. Also, some of the paid apps have less ratings. So price and ratings are poorly correlated.

3.14 Sentiment Polarity vs Sentiment Subjectivity



- As we can observe from the above graph that sentiment subjectivity mostly lies in the range of 0.5 to 0.8. which means people are giving reviews more opinion and experience based rather than factual.
- Sentiment subjectivity is mostly scattered around -0.5 to 0.75 this shows that polarity is not always proportional to sentiment subjectivity but in the maximum number of cases it shows a proportional behavior.

3.15 Effect Of Price On Rating



1. Here, we can observe that for high priced apps, the ratings are mostly more than 3.0.
2. But for free or apps with price less than 50 dollars, ratings do not usually follow any specific pattern.
3. They are skewed from 1 to 5. The reasons for such a diverse rating pattern could be user experience or size.

Conclusions:

1. **Reviews** and **Installs** share positive correlation while **Price** and **Rating** share negative correlation.
2. **Art and design** have the most number of installs.
3. Developing apps within **Family** and **Lifestyle** categories can be aimed for more profit i.e high revenue .
4. 61% of people have positive sentiments while approx 15% reacted negatively which is quite low in comparison.(Rest are Neutral).
5. Compared with Free and paid apps, 92.12% apps are Free and 7.81% apps are paid.
6. As **Everyone** content rating contains all age group people , it has maximum i.e 81.80% apps.
7. Maximum number of apps belong to the **Family** , **Game** and **Tools** category.
8. The category **Game** is a potential unsaturated space for all developers, as it has a maximum number of installs.
9. People love to download apps from **Tools** , **Entertainment** , **Education** , **Business** and **medical** genres.
10. Average rating of apps on the play store is 4.17 which is quite good.
11. Users prefer to pay for apps that are light weighted.
12. Paid apps that are higher in size may not perform well in the market.
13. Users tend to download a given app more if it has been reviewed by a large number of people.
14. People tend to review harsher reviews for paid apps.
15. There is a positive correlation between **Installs** and **Rating**.
16. To develop an app which results with high rating needs to get updated with the latest version keeping it optimally sized.
17. It’s good to develop a free app and have a content rating for everyone.

References:

1. GeeksforGeeks
2. Analytics Vidhya
3. AlmaBetter Class material
4. Pandas and Numpy libraries
5. Stack overflow
6. YouTube
7. Researchgate.net
8. W3schools.com