Capstone Project-1

EDA Play Store App Review Analysis

Team Members:

Md Asad Alam

(E-mail: mdasadalam354@gmail.com)
(github: https://github.com/asadalam1/play_store_app_analysis)

Ejaz Alam

(E-mail: ejazalam9006@gmail.com)
(github: https://github.com/EjazAlam9006/play-store-app-reveiw-analysis)

Pranjal Jha

(E-mail: sujeetkumarjha37@gmail.com)
(github:https://github.com/pranjaljha25/play_store_data_analysi-EDA-)

Abstract:

The Google play store is one of the largest and most popular Android app store. A few thousands of new applications are regularly uploaded on Google play store. A huge number of developers working freely on designing the apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application.

1. PROBLEM STATEMENT

We were given two datasets, one with basic information and the other with user reviews data of respective applications. 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

Play store is an Android Market which serves as the official app store for devices running on the Android Operating system. Developed and Operated by Google, launched on 6th March, 2012. Approximately 3.48 million apps are in the Play store. Play store apps have their own features such as Ratings, Reviews, Size and more. From the problem statement given, we should analyze the given database and should come up with the key factors that increased the number of users, long term usage etc., the objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

2.1 GOOGLE PLAY STORE DATASET

We have used Play Store data from the team capstone project dashboard. This data set contains 13 different features that can be used for predicting key factors responsible for app engagement & success stories.

This dataset consists 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.

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.

2.2USER 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.
- **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.

2.3DATA CLEANING AND PREPARATION

Data preprocessing and cleaning of raw data is the most important part of any data analysis to get more accurate insights. Preprocessing can help with completeness and comparability. 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 preprocessing because most real-world data contains inaccurate and invalid data. Data can also be uncertain i.e. there can be some missing values.

The available data are 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.

Steps Involved :-

- We have imported various important libraries such as numpy, pandas, matplotlib, warning and seaborn. We also imported both given csv files into Google Colab notebook .

- We need to get the basic overview of our datasets to get that we write can function such as shape, column etc for e.g. info() that will display about all the columns: Data type, Count of non-null values, in the play store dataset.

- 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.

- 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.

- 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.

- 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.

- 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.

- 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 **strip()** function and then convert the column into 'int' datatype.

- Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

- 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.

- 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 WHAT PERCENTAGE OF APPS ARE FREE AND PAID:

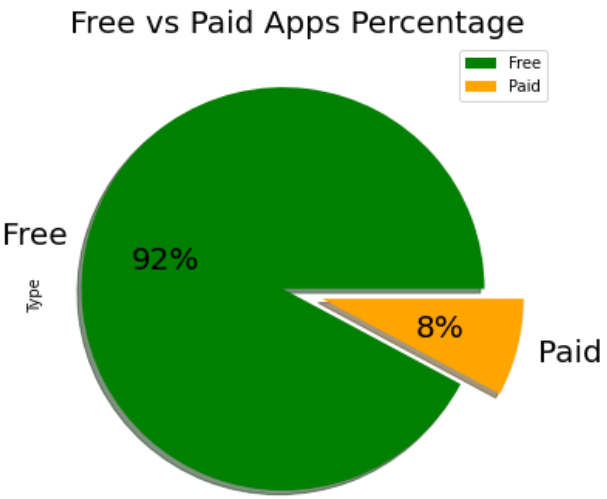


Fig -1: Free vs Paid

Here we can see that 92% apps are free, and 8% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

3.2 COUNT OF APPS IN EACH CATEGORY

- In the below plot, we plotted the count of apps category wise with its type Free/Paid.
- So, there are all total 33 categories in the dataset.

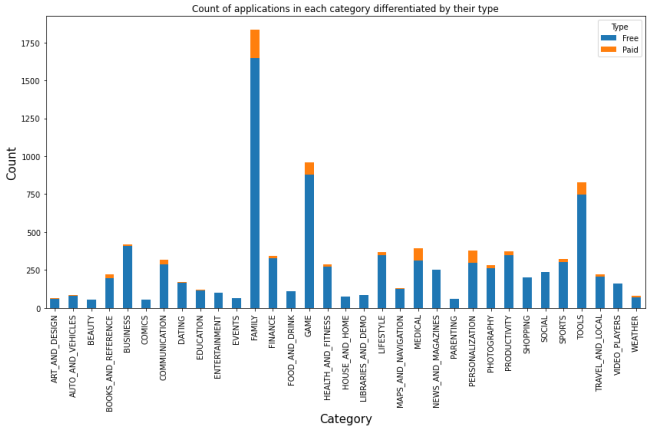
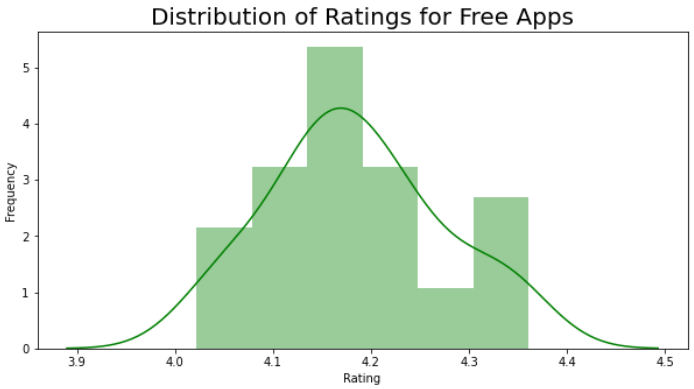


Fig -2: Count of apps in each category

- It looks like certain app categories h ave more free apps available for do wnload than others.
- In our dataset, the majority of apps i n Family, Games,Tools and Busines s are Free.
- At the same time Family, Medical, Games, Personalization and Tools h ad the most number of paid apps ava ilable for download.

3.3 AVERAGE RATING OF FREE AND PAID APPS

We have plotted the distribution plot for the Ratings of Free and Paid apps so to determine the average rating of apps present in the dataset.



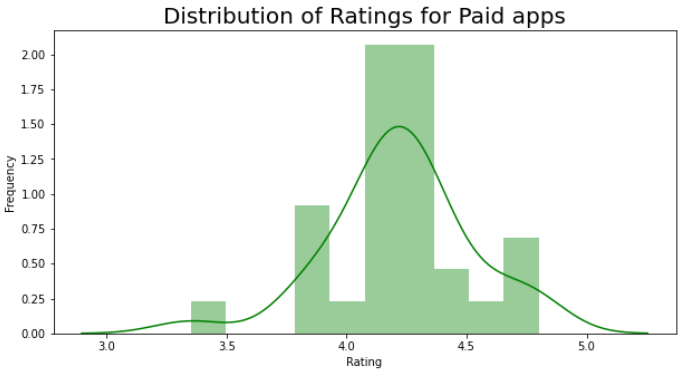


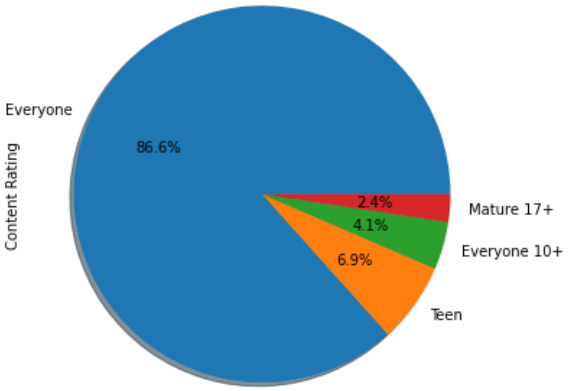
Fig -5: Distribution of App Rating for Free and Paid Apps

- Here we can see that the average ratings of free apps as approx. 4.1 out of 5 and for paid apps it is approx. 4.2 out of 5.
- Thus we can conclude that paid apps are slightly better rated as compared to free apps.
- However there are some paid apps whose average rating is below 3.5 which isn't the case with free apps.

3.4 COMPARING NO. OF INSTALLS AND NO. OF APPS AVAILABLE IN PLAY STORE BY ITS CONTENT RATING

A majority of the apps (82%) in the play store are can be used by everyone. The

remaining apps have various age restrictions to use it.



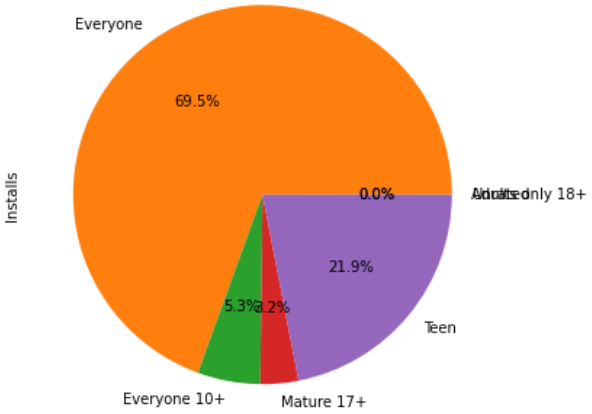


Fig -6: Content rating

- There are majority of apps available for Everyone followed by 10.7% of apps for Teens, 4% for Mature 17+ and 3.3% for 10+.
- And there are very few apps available for Adults 18+ and Unrated.

- We can notice that there are only 10.7% apps available for Teens but it accounts for 21% of total app installs, hence it is evident that demand of apps for Teens is very high.

3.5 CATEGORIES OF APPS THAT ARE GETTING INSTALLED THE MOST

This tells us the category of apps that has the maximum number of installs.
The Game, Communication and Tools categories has the highest number of installs compared to other categories of apps

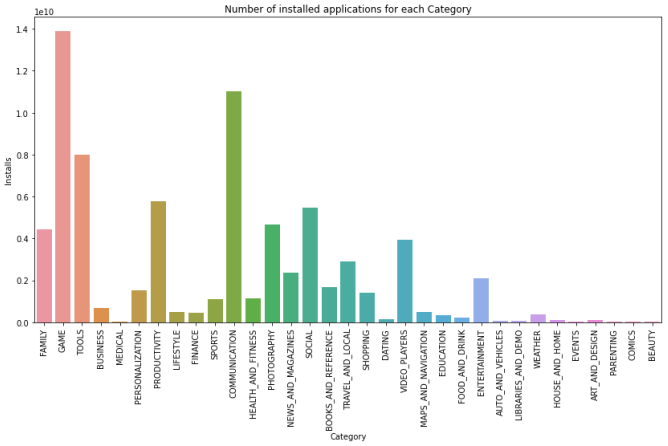


Fig -7: Category wise installs of application

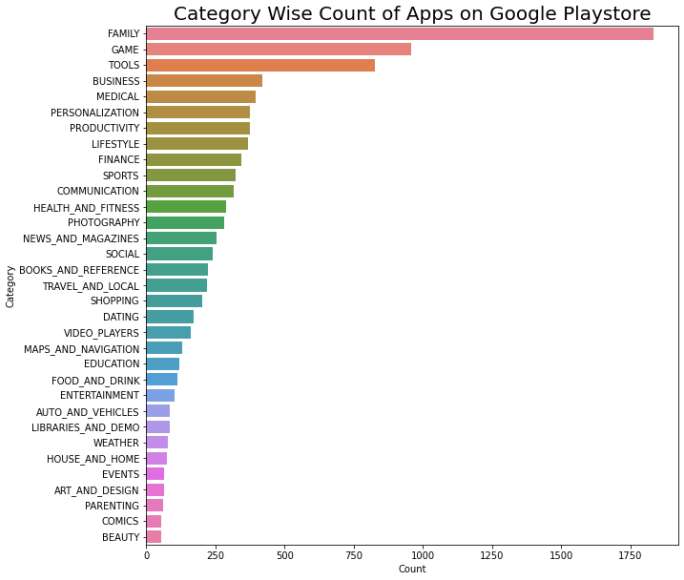


Fig -8: Category wise counts of application

We drew this two plots to compare the present scenario in which category the demand for apps is high as compared to its availability.

- As we can see from the above two plots: Maximum number of apps present in Google play store comes under Family, Games, Tools, and Business but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps comes under Games, Communication, Tools and productivity.
- Here we can also see that no of apps available in Communication and Social category is very less as compared to user demands as its installs are very high.
- Here we can also see that no of apps available in Communication and Social category is very less as compared to user demands as its installs are very high.

3.6 How does size impact the number of installs of any application

As we can notice, there is a high variance in the number of installs (minimum no. of installs is 0 and maximum is 1billion). To remove this we are adding a new column to data frame, which is the log of number of installs.

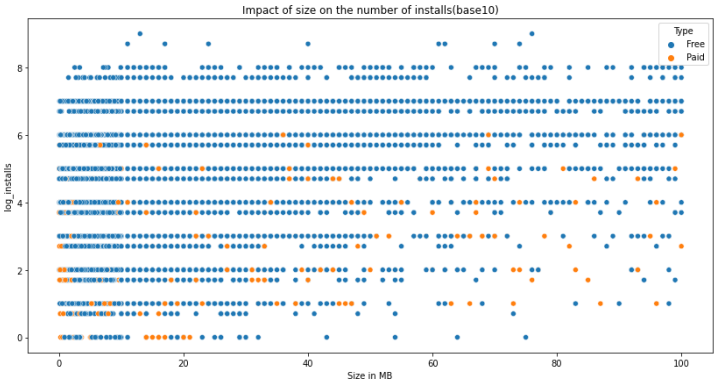


Fig -9: Impact of size on the number of installs

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

We plotted the Histogram of number of apps vs its size to see the composition of apps in the Play Store according to its size.

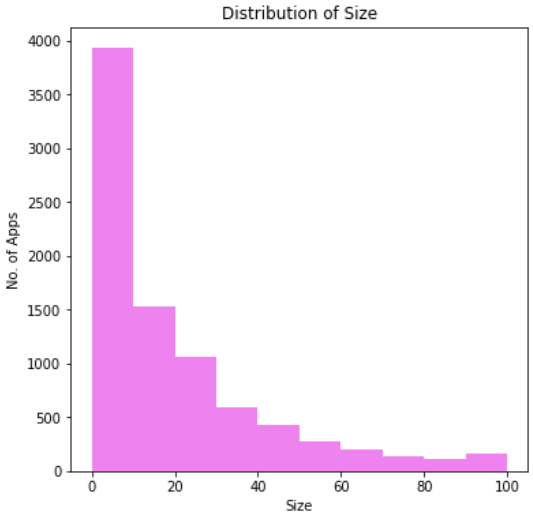


Fig -10: Distribution of size

- Majority of apps in the Play Store are of size less than 20MB.

- As we move further in Size axis the no. of apps drastically goes down.

3.7 TOP REVENUE GENERATED BY PAID APPS DEPENDING ON ITS CATEGORY

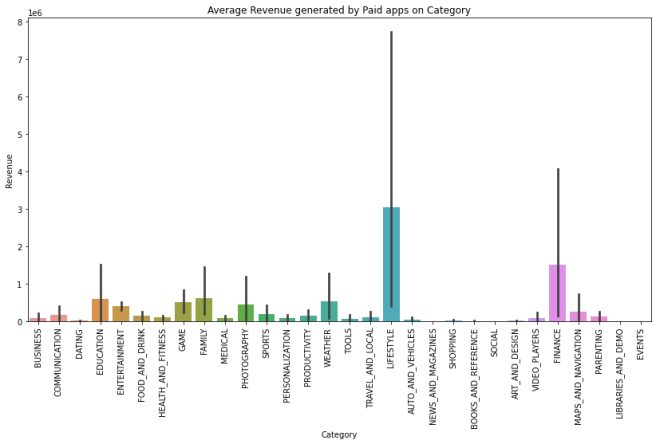


Fig -10: Tops paid app per category

- From the above, we can conclude that Lifestyle Category has generated the highest average revenue followed by Finance, Family and Education.

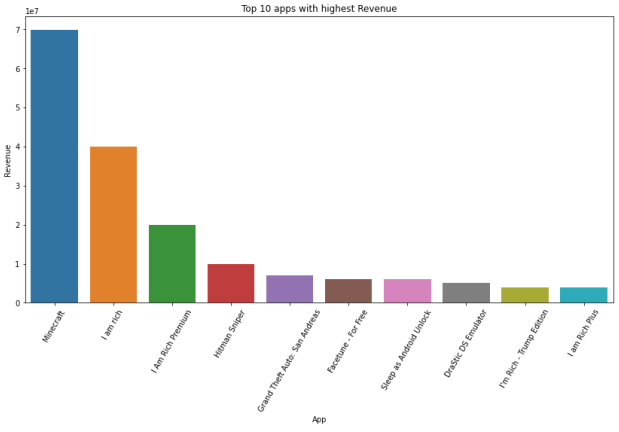


Fig -10: Tops 10 paid apps

- Minicraft, I am Rich, I am Reach Premium are the top revenue generating apps.

3.8 DRAWING 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 is drawn on the Numerical columns only, so in our data frame we had one numerical column Ratings and we have converted Reviews, Size, Installs and Price so the correlation is drawn on the above mentioned columns.

Fig -11: Correlation Heatmap

- We can see that there is a strong positive correlation between Reviews and Installs, More the app is being installed the more it will be reviewed.
- Also there is light positive correlation between Installs and Size, As People tend to install apps which consumes less memory in their device.
- Price is slightly negatively correlated with Rating, Review, Size and Installs.
- Rating is slightly positively correlated with Installs, Size and Reviews.

3.9 DISTRIBUTION OF TYPE OF REVIEWS IN THE DATASET

We have created a pie chart representing percentage of review sentiments and a bar plot representing category wise sentiments of apps.

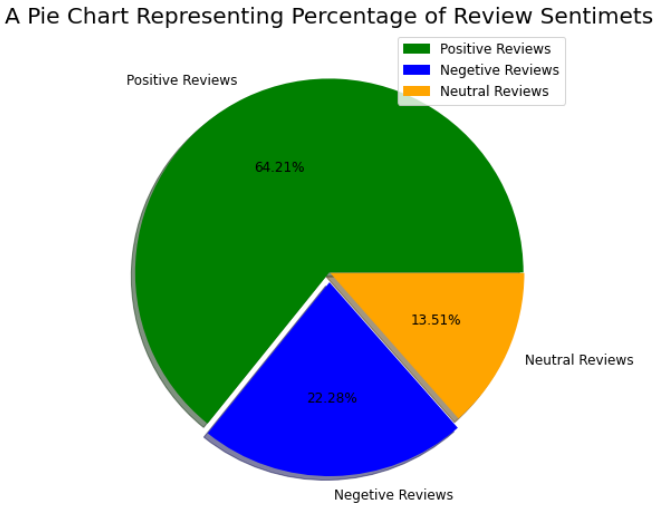


Fig -12: Percentage of Review Sentiment

- Majority of the apps (64.2%) in the Play Store is Positively Reviewed and 22.2% are Negatively Reviewed and the rest 13.5% are Neutral Reviewed**.**

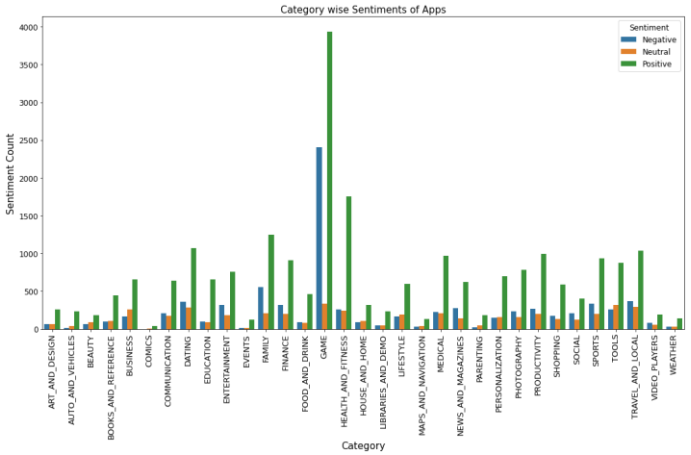


Fig -13: Category wise sentiments of apps

- Looking Category wise Games has both highest Positive Review and Negative Review at the same time.

3.10 Analyzing Sentiment Polarity and Sentiment Subjectivity

- After cleaning our both the data frame we merged the both data frames one of app data and other of review data on the ‘App’ column.
- In the merged data frame, we have three new columns i.e.. Sentiment, Sentiment Polarity and Sentiment Subjectivity. Sentiment basically determines the attitude or the emotion of the writer, i.e., whether it is positive or negative or neutral. Sentiment Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Sentiment Subjectivity generally refer to personal opinion, emotion or judgment, which lies in the range of [0, 1].
- We have plotted Distribution of Subjectivity and correlation heatmap for Sentiment Subjectivity and Sentiment Polarity.

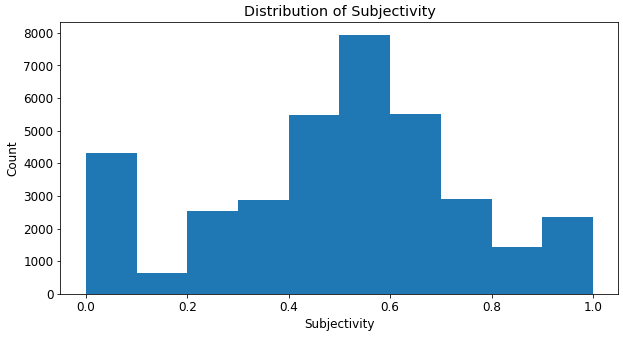


Fig -14: Distribution of Subjectivity

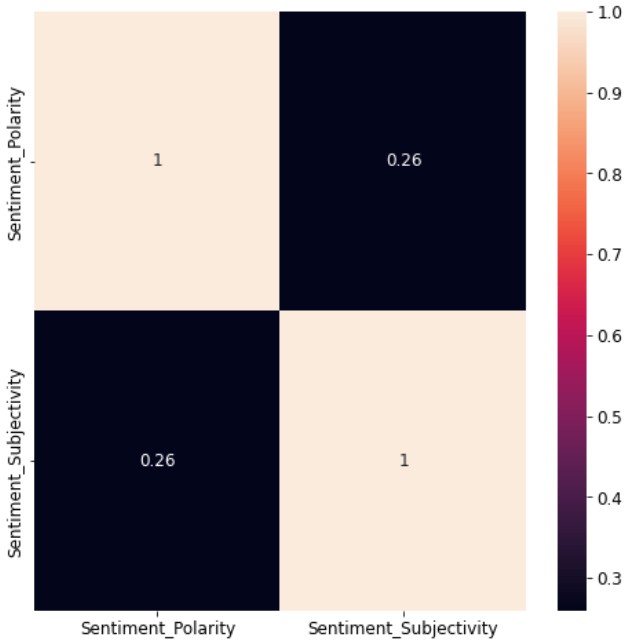


Fig -15: Subjectivity-Polarity Heatmap

- 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 and is not factual.
- From the above Heatmap it can be concluded that sentiment subjectivity is not always correlated to sentiment polarity.

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:-

- Majority of the apps in the Play Store is **Free (~92%).**

- **Family, Games, Tools, and Business** Category holds the major chunk of apps in the Play Store.
- Most competitive category is **Family and Games**.
- Category of Apps getting installed the most is **Games, Communication, and Tools**.
- Demands of the **Communication Category** apps is higher .
- **Lifestyle** Category has generated the **highest revenue** through installs.
- Percentage of apps with no **age restriction is ~82%**
- Only **11%** of apps are available for **Teens** but it accounts for **21% of total app installs**.
- **Minecraft** app has generated the highest revenue of **~69Million$** with over **~10Million installs.**
- Average Rating of Paid apps is **4.2** and for Free apps is **4.1** .
- **Bulky** apps are **less installed** by the user.
- Majority of the apps in the Play Store are of size **less than 20MB**.
- There is a strong **positive correlation** between **Reviews and Installs**.
- **Price** is **negatively correlated** with **Installs**.
- Majority of the apps **(64.2%)** in the Play Store is **Positively Reviewed**.
- **Sentiment Subjectivity** is not correlated with **Sentiment Polarity**.

References~

- GeeksforGeeks
- Analytics Vidhya
- Stackoverflow
- Python libraries documentation

Thank You