Capstone Project-1 Submission

PLAY STORE APP REVIEW ANALYSIS

Pranav Vilasrao Balpande

Kartik Subhashrao Dhande

Sanket Rajendra Bhosale

Kartik Anilrao Pisudde

Data science trainees,

AlmaBetter, Bangalore

----------***----------

Pranav Vilasrao Balpande - pranav.balpande@gmail.com

Kartik Subhashrao Dhande – dhandekartik07@gmail.com

Sanket Rajendra Bhosale - sanketbhosale0023@gmail.com

Kartik Anilrao Pisudde - pisuddekartik@gmail.com

GitHub Link~~

Pranav Vilasrao Balpande :-
https://github.com/pranav4536/capstone-project-1-

Kartik Subhashrao Dhande :-
https://github.com/KartikDhande007/play_store_app_review_analysis

Sanket Rajendra Bhosale :-
https://github.com/sanket-bhosale-12/Capston-project-1---Play-store-app-analysis

Kartik Anilrao Pisudde :-

https://github.com/kartik-pisudde/capstone-project-playstore-review-analyze

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 Alma-better , 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:

- **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

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing 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 preprocessing 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 drop the columns 'Current Ver', 'Android Ver' and 'last updated' from our dataset using the drop () function of the pandas' library.

- **Step 4**: 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 Fillan() function.

- **Step 5:** 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 6:** 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 7:** 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 8:** 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.

- **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

- **Step 10:** 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.

- **Step11:** 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 FREE VS PAID

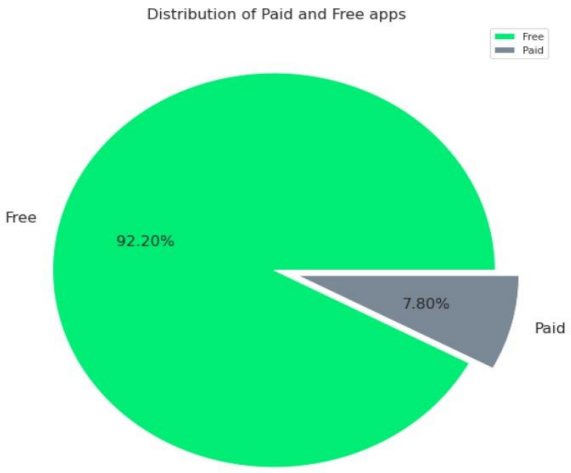

Fig -1: Free vs Paid

Here we can see that 92.2% apps are free, and 7.80% apps are paid on Google Play Store, so we can say that Most ofthe apps are free on Google Play Store.

3.2 RATING

In the below plot, we plotted the apps Rating

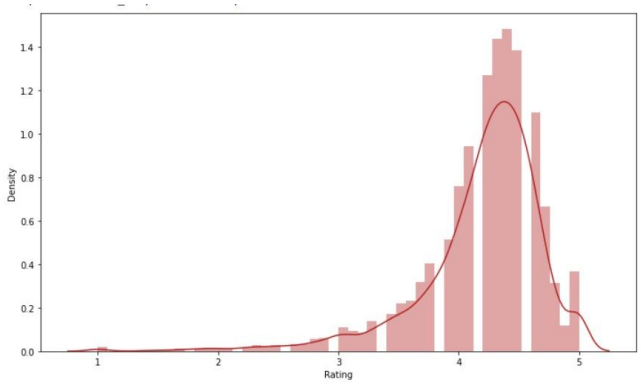

Fig -2: Distribution of App rating

- 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 50% of the apps have an average rating of above 4.3, and the rest below 4.3.
- From the distplot visualizations, it is clear that the ratings are left skewed.
- We know that if the variable is skewed, the mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation of the majority of the values in the variable.

3.3 DISTRIBUTION OF APP SIZE

The below curve represents the variation of the size of apps available on Google Play store

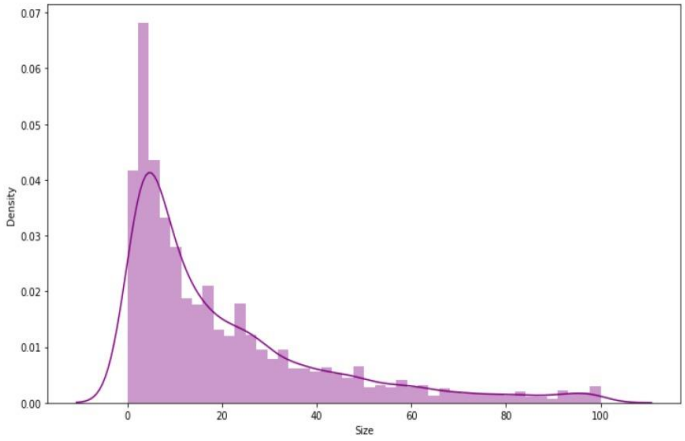

Fig -5: Distribution of App Size

- It is clear from the visualizations that the data in the **Size** column is skewed towards the right.
- Also, we see that a vast majority of the entries in this column are of the value **Varies with device**, replacing this with any central tendency value (mean or median) may give incorrect visualizations and results. Hence these values are left as it is.

3.4 UPDATED PAID APPS

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.

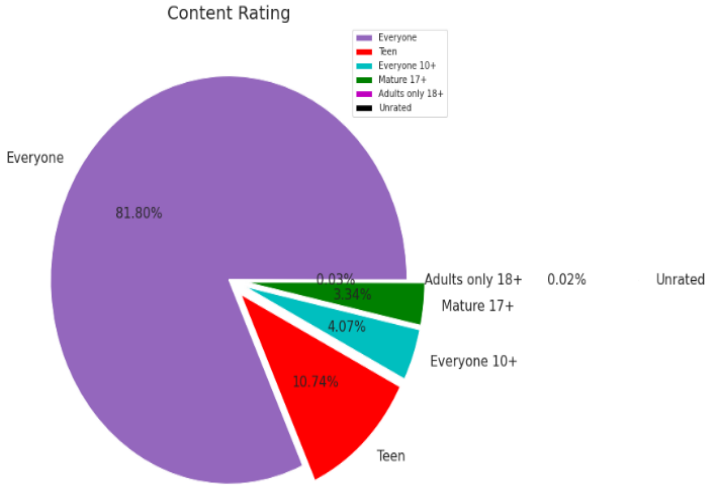

Fig -6: Content rating

3.5 TOP CATEGORY OF PLAY STORE

There are lot of category wise apps are available on play store so the below curve show hoe the apps are distributed.

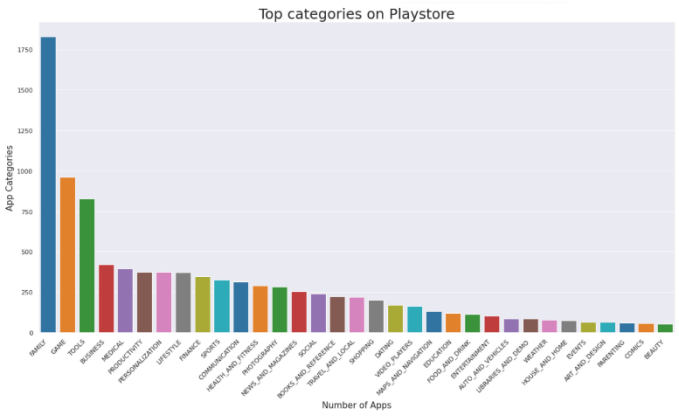

Fig -7: Top Categories on Playstore

So, there are all total 33 categories in the dataset from the above output we can come to a conclusion that in play store most of the apps are under FAMILY & GAME category and least are of EVENTS & BEAUTY Category.

3.6 No. of Installs per Category

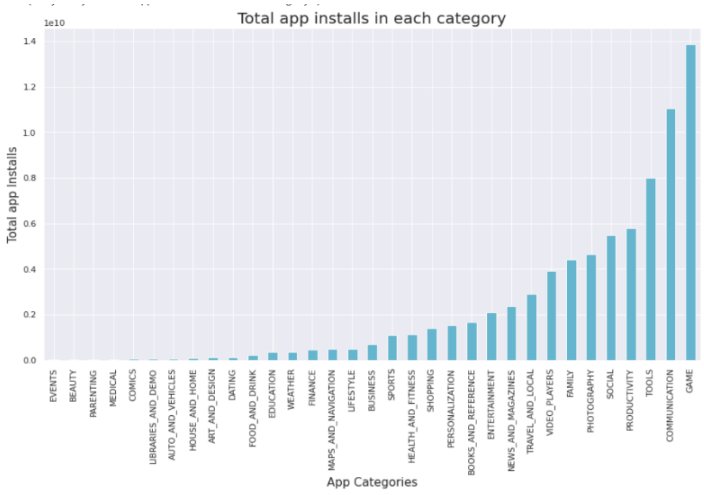

Fig -8: No. of Installs Per Category

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.

3.7 Average App Ratings

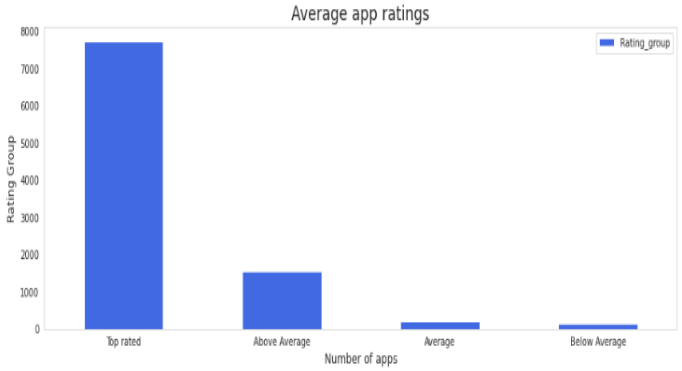

Fig -9: Average App Ratings

The rating available in the dataset is distributed so we can represent the ratings in a better way if we group the ratings between certain intervals. Here, we can group the rating as follows:

- 4-5: Top rated
- 3-4: Above average
- 2-3: Average
- 1-2: Below average

3.8 Top Paid Apps per Category

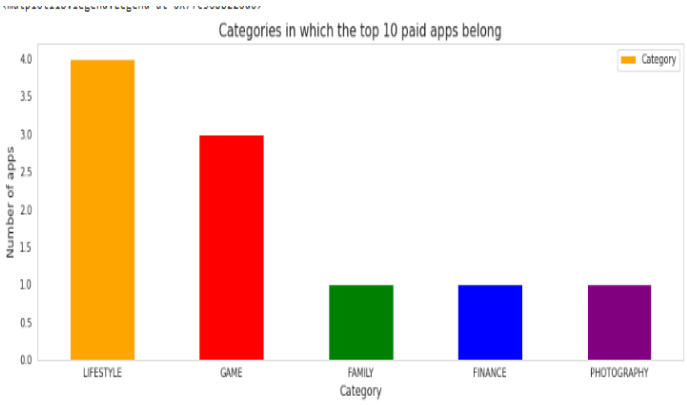

Fig -10: Tops paid app per category

From the above, we can conclude that most of the paid apps are present in the lifestyle and game category.

3.9 PERCENTAGE OF USER REVIEW SENTIMENTS

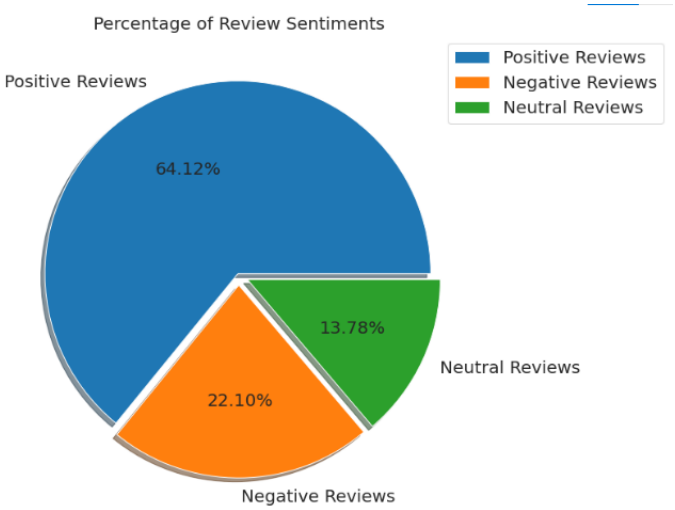

Fig -11: Percentage of User Review Sentiments

From the above pie chart, we can say that most of the apps that are present on the play store has received positive review by the user while there are some apps which have negative reviews as well.

3.10 TOP 10 POSITIVELY REVIEWED APPS

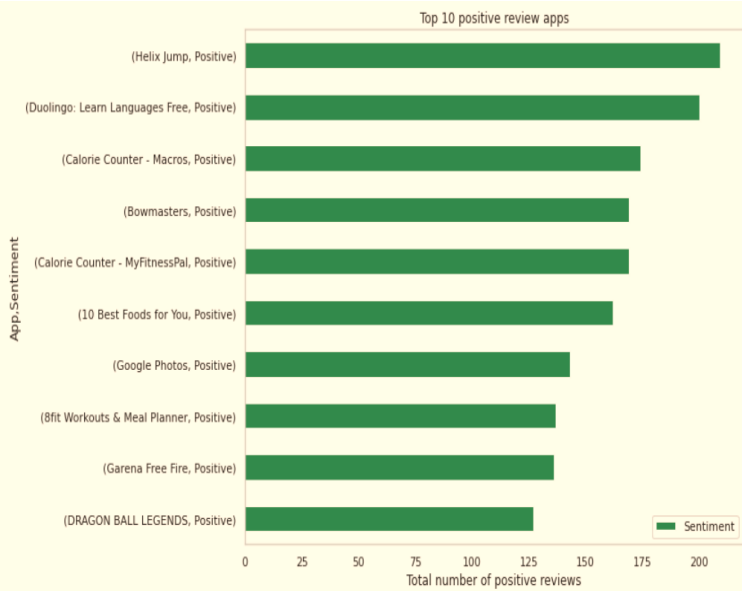

Fig -12: Top 10 Positive Reviewed App

3.11 TOP 10 NEGATIVE REVIEWS APPS

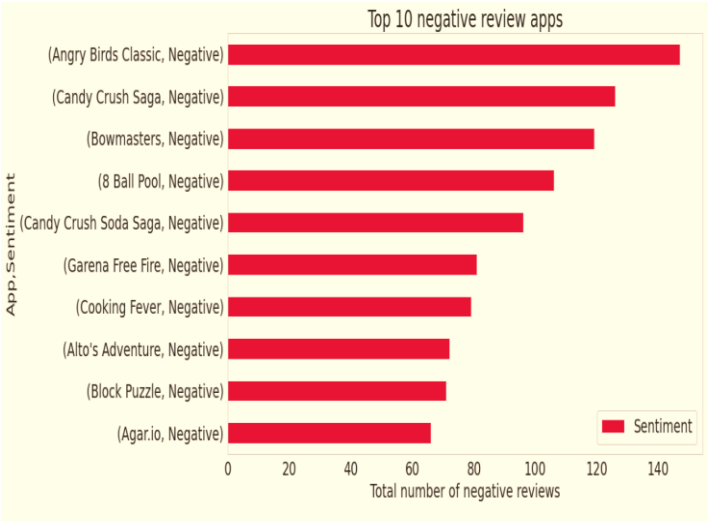

Fig -13: Top 10 Negative Reviewed Apps

3.12 TOP 10 NEGATIVE REVIEWS APPS

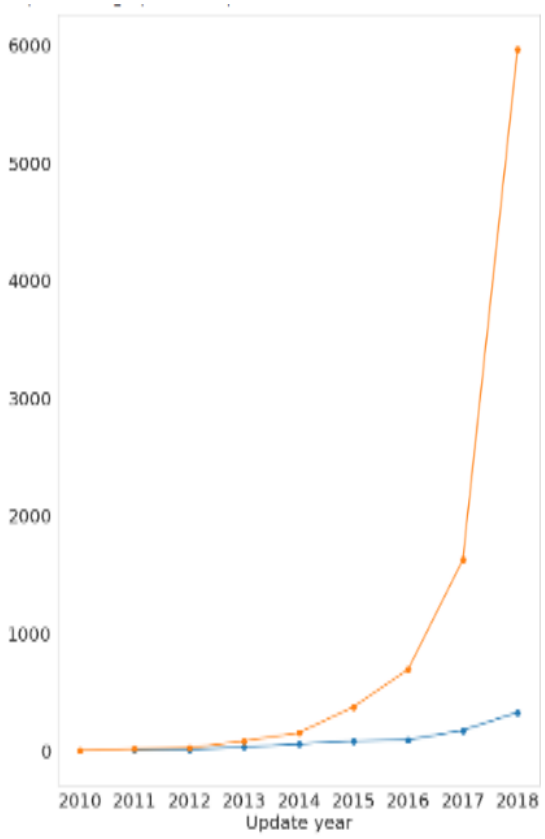

Fig -14: Top 10 Negative Reviewed Apps

3.13 DISTRIBUTION OF APP UPDATE OVER THE YEAR

In the above plot, we plotted the apps updated or added over the years comparing Free vs. Paid, by observing this plot we can conclude that before 2011 there were no paid apps, but with the years passing free apps has been added more in comparison to paid apps, by comparing the apps updated or added in the year 2011 and 2018 free apps are increases from 80% to 96% and paid apps are goes from 20% to 4%. So, we can conclude that most of the people are after free apps.

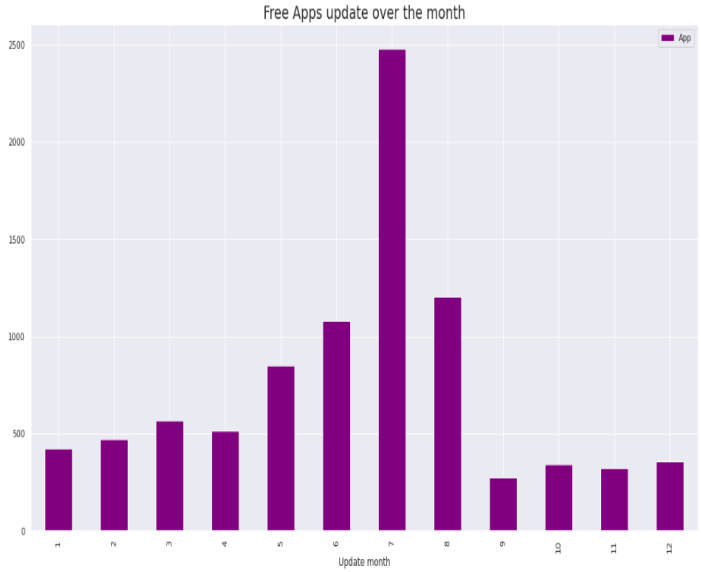

Fig -15: Paid Apps update over the month

3.14 DISTRIBUTION OF APP UPDATE OVER THE MONTH

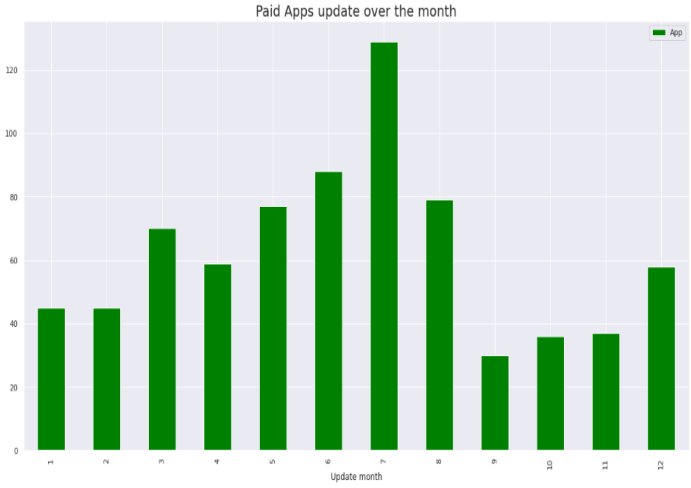

Fig -16: Free Apps update over the month

In this data almost 50% apps are added or updated on the month of July, 25% of apps are updated or added on the month of August and rest of 25% remaining months.

Most of the paid apps too updates in the month of July same as free app.

3.15 RELATIONSHIP BETWEEN SENTIMENT SUBJECTIVITY PROPORTIONAL TO SENTIMENT POLARITY

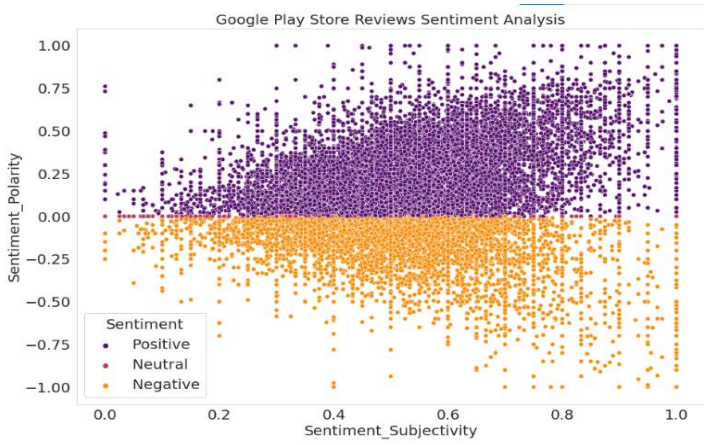

Fig -17: Google play store Reviews Sentiment Analysis

From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, show a proportional behavior, when variance is too high or low.

3.16 DISTRIBUTION OF SUBJECTIVITY

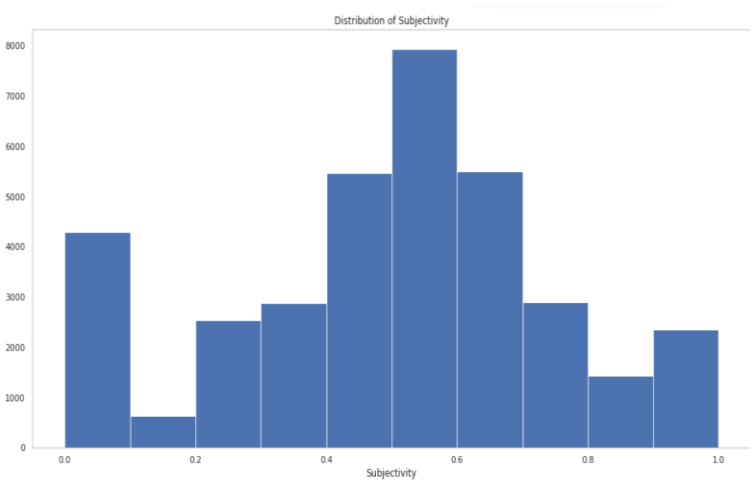

Fig -18: Distribution of subjectivity

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.

3.17 RELATIONSHIP BETWEEN DIFFERENT FEATURES OF THE DATASET

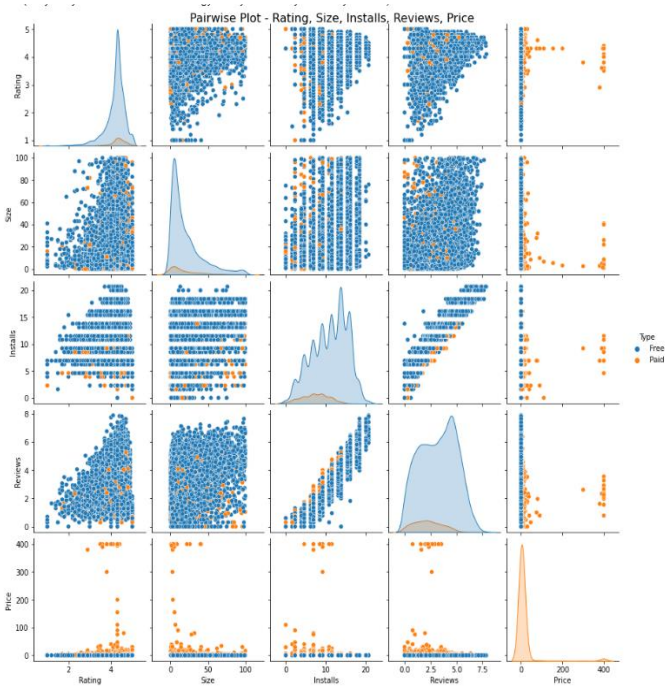

Fig -19: Pair wise plot

- Most of the App are Free.
- Most of the Paid Apps have Rating around 4
- As the number of installations increases the number of reviews of the particular app also increases.
- Most of the Apps are light-weighted.

3.18 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.

3.19 PLAY STORE CORRELATION HEATMAP

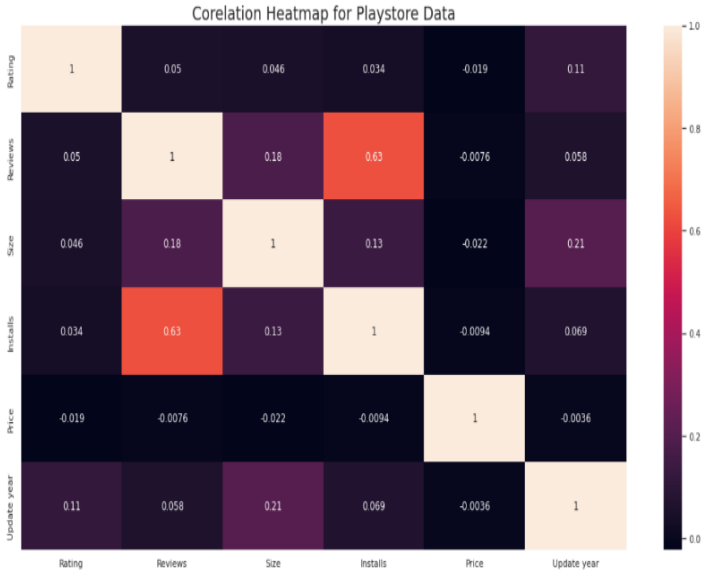

Fig -20: Correlation Heatmap

- There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
- The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviews and installs fall slightly.
- The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs, and number of reviews also increase.

3.20 MERGED DATA FRAME HEATMAP

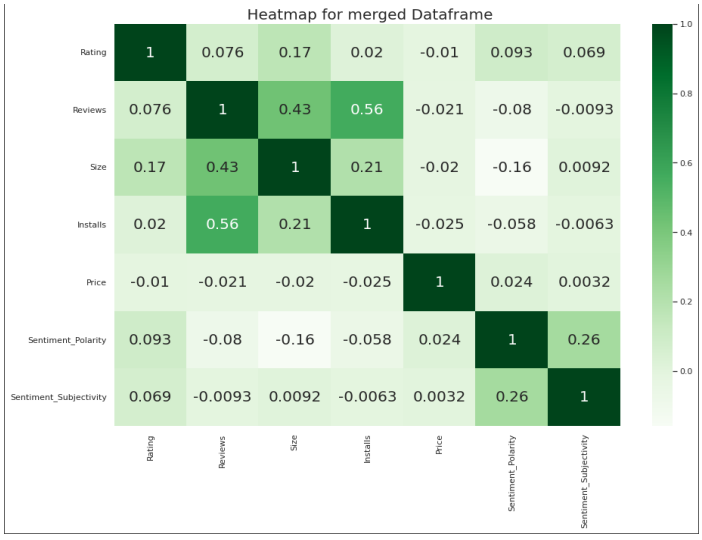

Fig -21: Merged Data frame Heatmap

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.

- Percentage of free apps = ~92%
- Percentage of apps with no age restrictions = ~82%
- Most competitive category: Family
- Family, Game and Tools are top three categories having 1906, 926 and 829 app count.
- 8783 Apps are having size less than 50 MB. 7749 Apps are having rating more than 4.0 including both type of apps.
- Category with the highest average app installs: Game
- Percentage of apps that are top rated = ~80%
- There are 20 free apps that have been installed over a billion time
- There are 20 free apps that have been installed over a billion time
- Minecraft is the only app in the paid category with over 10M installs. This app has also produced the most revenue only from the installation fee.
- Category in which the paid apps have the highest average installation fee: Finance
- The median size of all apps in the play store is 12 MB.
- The apps whose size varies with device has the highest number average app installs.
- The apps whose size is greater than 90 MB has the highest number of average user reviews, ie, they are more popular than the rest.
- Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.
- Overall sentiment count of merged dataset in which Positive sentiment count is 64%, Negative 22% and Neutral 13%.
- Sentiment Polarity is not highly correlated with Sentiment Subjectivity.

- Tools, Entertainment, Education, Business and Medical are top Genres.

References~

- GeeksforGeeks
- Analytics Vidhya
- Stack overflow
- Towards data science
- Python libraries documentation
- Data camp
- 1. Researchgate.net
- 2. https://www.academia.edu

Thank You