Capstone Project-1 Submission

Play Store App Review Analysis

Ashwini R

Dhanaraj S

Data Science Trainees,

AlmaBetter, Bangalore

---***---

Ashwini R- ashwini.rajendra58@gmail.com

Dhanaraj S- dhanaraj.siddappa81@gmail.com

GitHub Link~~

Ashwini R: - https://github.com/AnshRockstar/Playstore-Apps-Review-Analysis.git

Dhanaraj S: - https://github.com/San13deep/Play-Store-App-Review-Analysis.git

Abstract - *The Google Play Store witnesses a constant influx of several thousand new applications regularly, created by a growing number of independent developers or collaborative teams facing intense competition on a global scale. Given that a majority of Play Store apps are offered for free, the revenue model remains somewhat obscure, relying on in-app purchases, advertisements, and subscriptions to contribute to an application's success. Consequently, an app's success is often measured by its installation numbers and user ratings over its lifetime rather than the generated revenue.*

App ratings, serving as voluntarily provided feedback by users, play a crucial role as an evaluation criterion for apps. However, these ratings can be subject to bias due to inadequate or missing votes. Furthermore, disparities often exist between numeric ratings and the content of user reviews. This study aims to predict the ratings of Google Play Store apps using machine learning algorithms. Through data analysis and prediction, I have explored a dataset from the Google Play Store applications obtained from Kaggle. Utilizing machine learning algorithms, the objective is to unveil relationships among various attributes within the dataset, such as distinguishing between free and paid applications and understanding the impact of user reviews on an app's rating.

***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 methodologies play a crucial role in addressing a multitude of challenges. This paper delves into the intricate details of machine learning models and frameworks. With applications spanning various domains, machine learning exhibits significant potential for further advancements.

In the future, it is anticipated that machine learning will establish optimal theories to elucidate its performances. Simultaneously, improvements in unsupervised learning capabilities are expected as there is a vast amount of data globally, yet it may not be feasible to assign labels to all of it. The projection also includes the anticipation that neural network structures will become increasingly intricate, enabling them to discern more semantically meaningful features. Additionally, it is foreseen that deep learning will synergize more effectively with reinforcement learning, allowing for the utilization of these advantages in accomplishing a broader range of tasks.

2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

In the current landscape, we observe that mobile apps play a pivotal role in the lives of individuals, significantly influencing technological advancements. The evolution of the mobile application industry has had a profound impact on digital innovation. However, amidst the continually expanding mobile app market, there is a noticeable surge in mobile app developers, contributing to the substantial revenue generated by the global mobile app industry.

Given the intense competition on a global scale, it is crucial for developers to ensure they are on the right path. To sustain their revenue and market position, app developers may need to strategize on maintaining their current standing. The Google Play Store stands out as the largest application platform. Despite generating more than twice the downloads compared to the Apple App Store, it is observed to make only half the revenue. Consequently, data from the Play Store was collected to conduct our analysis.

With the rapid evolution of smartphones, mobile applications (Mobile Apps) have become integral aspects of our daily lives. However, keeping abreast of the latest developments and comprehending

every detail about the apps is challenging, given the continuous influx of new applications into the market. In September 2011, it was reported that the Android market had reached half a million applications. Currently, 0.675 million Android apps are available on the Google Play App Store. This abundance of apps presents a significant opportunity for users to choose from a diverse range.

We believe that mobile app users consider online app reviews to be a crucial influence when evaluating paid applications. For a potential customer, it is challenging to sift through all the textual comments and ratings to make an informed decision. Additionally, application developers face difficulties in understanding how to enhance app performance based solely on overall ratings and would benefit from comprehending numerous written comments.

We engage in the development of Android apps, releasing them on the Play Store. From both a developer and business standpoint, understanding whether users are enjoying the app or encountering issues is crucial. To gather this feedback, the Play Store provides a Ratings and Reviews section for each app. Users have the ability to submit ratings and express their thoughts through reviews. However, the conventional method of rating and reviewing an app involves a lengthy process of navigating to the Play Store to provide feedback or being redirected, disrupting the current app workflow to open the Play Store App link using URI. Although we prefer to keep our customers within our application, this process forces us to redirect control to the Play Store app.

2.2 GOOGLE PLAY STORE DATASET

The dataset comprises Google Play Store applications and is sourced from Almabetter, recognized as the world's largest community for data scientists, offering a platform to explore, analyze, and share data.

This dataset comprises web-scraped information on 10,000 Play Store applications, providing a basis for analyzing the Android market. Users can utilize this downloaded dataset to explore various categories in the Android market, such as music and camera applications. With this data, users can predict whether a given application is likely to receive a lower or higher rating. Additionally, this dataset can serve as a valuable resource for future reference and recommendations when proposing new applications.

Furthermore, the offline dataset is selected to ensure precise measurements, as online data undergoes frequent updates. Using this dataset, I plan to analyze various attributes such as ratings, whether an application is free or paid, etc., employing Hive. Subsequently, I will also conduct predictions for various characteristics, including user reviews and ratings. The data set contains the following columns:

- **App:** This column displays the app's name.
- **Category:** This includes the app's category. The category column encompasses 33 distinct values.
- **Rating:** This column comprises the mean rating that the app has garnered on the Play Store. Individual ratings can range from 0 to 5.
- **Reviews:** This column displays the count of individuals who have provided feedback for the app.
- **Size:** This column indicates the app size, representing the amount of memory space the app occupies on the device following installation.
- **Installs:** This column reflects the approximate number of times the app has been downloaded from the Play Store. These values are not absolute but are estimates.
- **Type:** This column only includes two values: free and paid. It signifies whether users need to make a payment to install the app on their device or not.
- **Price:** For paid applications, this column displays the app's price; for free apps, it holds the value 0.
- **Content Rating:** It signifies the app's intended audience and their age group.
- **Genre:** This column specifies the genre to which the app belongs. The genre can be regarded as a subdivision of the category.
- **Last Updated:** This column provides information about the date when the app's latest update was released.
- **Current Version:** This column includes details about the present version of the app accessible on the Play Store.

- **Android version:** This column provides details about the Android OS version on which the app is compatible for installation.

2.3 User Review Dataset

The user reviews data frame consists of 64,295 rows and is structured with 5 columns, identified as follows:

- **App:** Includes the app name along with an optional brief description.
- **Translated Review:** This column contains the English translation of the review provided by the app user.
- **Sentiment:** This column indicates the sentiment or emotion expressed by the writer, categorized as 'Positive,' 'Negative,' or 'Neutral.'
- **Sentiment Polarity:** This column provides the polarity of the review, with a range of [-1, 1], where 1 indicates a 'Positive statement' and -1 signifies a 'Negative statement.'
- **Sentiment Subjectivity:** This value indicates the proximity of a reviewer's opinion to the general public's opinion, with a range of [0,1]. A higher subjectivity implies that the reviewer's opinion aligns closely with the general public, while lower subjectivity suggests that the review contains more factual information.

2.4 Python

Many data scientists prefer Python because of its robust built-in library functions and the supportive community. Python boasts an impressive repository of 70,000 libraries. It stands out as one of the easiest programming languages to learn when compared to others, making it a popular choice among data scientists. The simplicity of Python is a key factor, especially in fields like machine learning and data processing, where analysts seek a user-friendly language. This ease of use is a significant rationale for the prevalent use of Python. Notably, Pandas, a widely used open-source library, holds a special place among data scientists. As observed in our previous assignment, when the need arises to create scatterplots, heatmaps, graphs, or handle 3-dimensional data, Python's built-in library proves to be highly beneficial.

2.5 Data Cleaning and Preparation

Preprocessing plays a crucial role in transforming raw data into a more suitable format. Engaging in the preprocessing process contributes to achieving data completeness and reliability. For example, it allows you to identify whether certain values were recorded or not and assess the reliability of the information. Additionally, preprocessing aids in evaluating the consistency of values. The necessity for preprocessing arises because real-world data is often unclean. Data may exhibit noise, such as outliers or errors in general, and may also be incomplete, featuring missing values.

The data currently accessible is in its raw form and is unsuitable for Exploratory Data Analysis. Therefore, before proceeding with any analysis, we need to thoroughly explore and clean the data to prepare it for further examination.

- **Step1**: We create a function called `play_store_info()` to present five attributes for all the columns in the Play Store dataset: Data type, Count of non-null values, Count of null values, number of unique values in that column, and the percentage of null values in that column.

- **Step2**: Commencing with the 'Type' column, we observe a single null value. Upon investigation, we determined from the Play Store that it pertains to a free app. To address this, we utilize the `fillna()` function from the pandas library to fill in this value.

- **Step 3**: We utilize the `drop()` function from the pandas library to eliminate the columns 'Current Ver', 'Android Ver', and 'Last Updated' from our dataset.

- **Step 4**: Observing the 'Rating' column, we identify 1474 null values. Given the limited variations in the rating values and a high frequency of repeated values, the 'median' serves

as a suitable statistical indicator for replacing the null values. We compute the median of the column using the `median()` aggregate method and utilize the `fillna()` function to substitute null values with this calculated median.

- **Step 5:** Despite the 'Reviews' column being a numerical indicator, it is currently of the 'object' data type. To rectify this, we will convert it to the 'int' data type using the `astype(int)` function.

- **Step 6:** Observing the 'Size' column, which ideally should be numeric, we note that it is currently of the 'object' data type. Additionally, the values contain characters 'k' and 'M,' representing kilobytes and megabytes, respectively. To address this, we will replace 'k' with 1000 and 'M' with 1000000. Furthermore, certain values include the '+' sign, which will be eliminated. Subsequently, we will convert this column to the 'int' data type.

- **Step 7:** The values in the 'Installs' column include the characters '+' and ',', hindering the conversion of this column into a numeric data type. To address this, we will remove these characters using the `strip()` and `replace()` functions.

- **Step 8:** The 'Price' column values may include the '$' sign, and the column is currently of the 'object' data type. Initially, we will eliminate the '$' sign using the `strip()` function, followed by converting the column to the 'int' data type.

- **Step 9:** Addressing duplicates in the 'App' column involves eliminating the number of duplicate rows within this specific column.

- **Step 10:** We create a function called `ur_info()` to showcase five attributes for all the columns in the User Review dataset: Data type, Count of non-null values, Count of null values, the number of unique values in that column, and the percentage of null values in those columns.

- **Step11:** Within the User Review dataset, the columns include App, Translated Review, Sentiment, Sentiment Polarity, and Sentiment Subjectivity. Among these, a total of 26,863 NaN values are identified, and we eliminate them using the `dropna()` function.

3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, abbreviated as EDA, constitutes a crucial phase in any Data Analysis or Data Science endeavor. EDA involves thoroughly examining the dataset to unveil patterns, identify anomalies (outliers), and formulate hypotheses grounded in our comprehension of the dataset.

Exploratory Data Analysis (EDA) encompasses the generation of summary statistics for numerical data within the dataset and the creation of diverse graphical representations to enhance our understanding of the data. This article will delve into EDA using an illustrative dataset, employing the Python language with the Pandas library for this demonstration.

3.1 FREE VS PAID

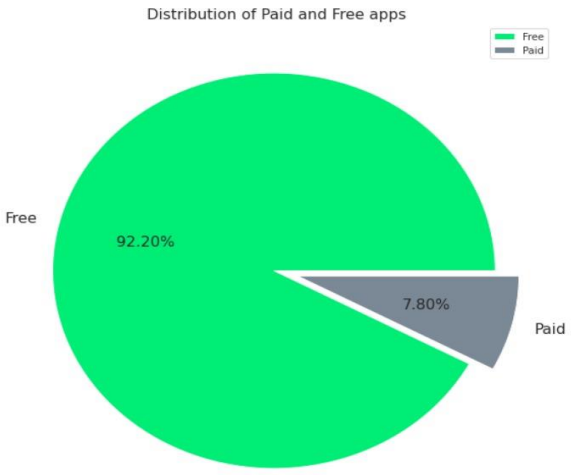

Fig -1: Free vs Paid

Observing the data, it becomes evident that 92.2% of the apps are free, while 7.80% are paid on the Google Play Store. This indicates that the majority of apps on the Google Play Store are free.

3.2 Rating

In the below plot, we plotted the apps Rating

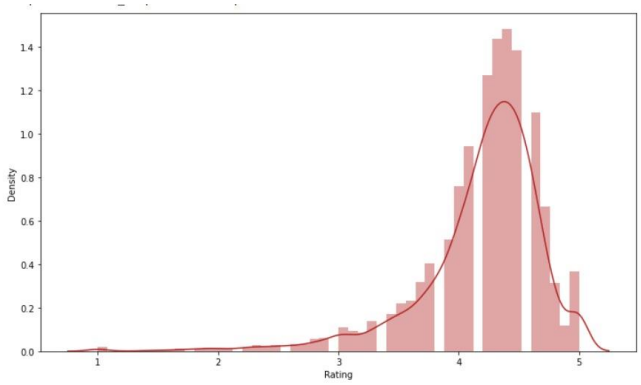

Fig -2: Distribution of App rating

- The average of the mean ratings (excluding NaN values) is calculated to be 4.2.
- The median of the entries (excluding NaN values) in the 'Rating' column is determined to be 4.3. This implies that 50% of the apps possess an average rating above 4.3, while the remaining 50% fall below 4.3.
- The distplot visualizations indicate a clear left-skewed distribution of ratings.
- Understanding that a skewed variable can bias the mean due to values at the distribution's far end, we recognize that the median provides a more representative measure for the majority of 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

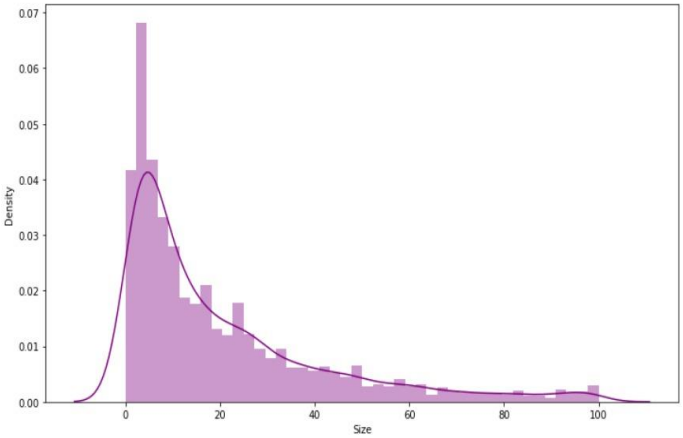

Fig -5: Distribution of App Size

- The visualizations make it evident that the data in the Size column demonstrates a rightward skew.
- Furthermore, a significant proportion of entries in this column are labeled as "Varies with device." Substituting these with a central tendency value (such as mean or median) could lead to inaccurate visualizations and results. Therefore, these values are retained in their original form.

3.4 Updated Paid Apps

The majority of apps (82%) in the Play Store are accessible to users of all age groups, while the remaining apps have different age restrictions.

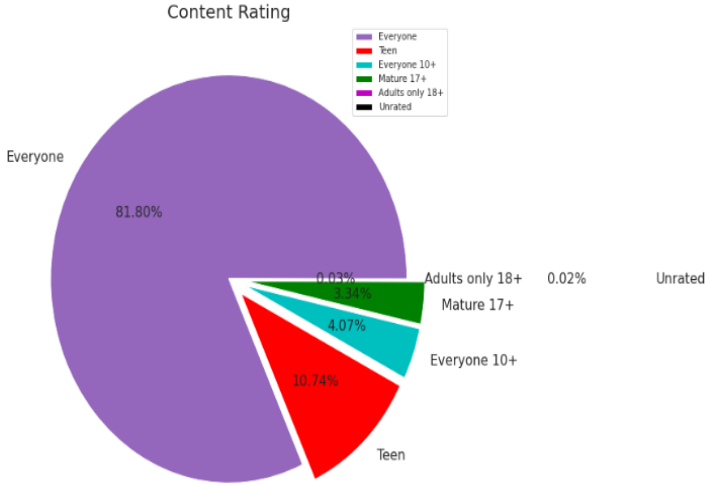

Fig -6: Content rating

3.5 Top Category of Play Store

Numerous apps are available on the Play Store, categorized in various ways. The graph below illustrates how these apps are distributed.

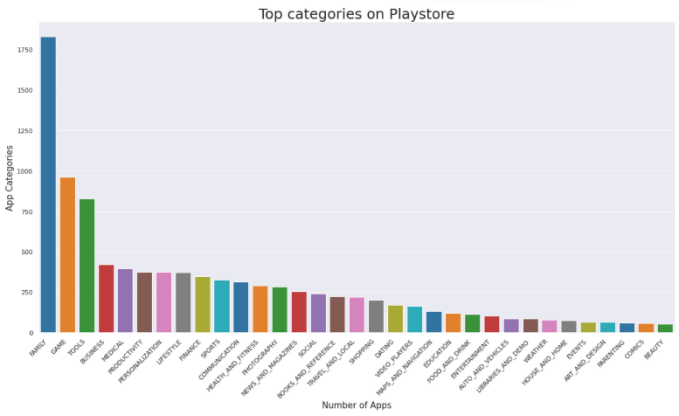

Fig -7: Top Categories on Play Store

Therefore, the dataset comprises a total of 33 categories. From the provided output, it can be deduced that in the Play Store, the most prevalent app categories are FAMILY and GAME, whereas the least represented categories include EVENTS and BEAUTY.

3.6 No. of Installs per Category

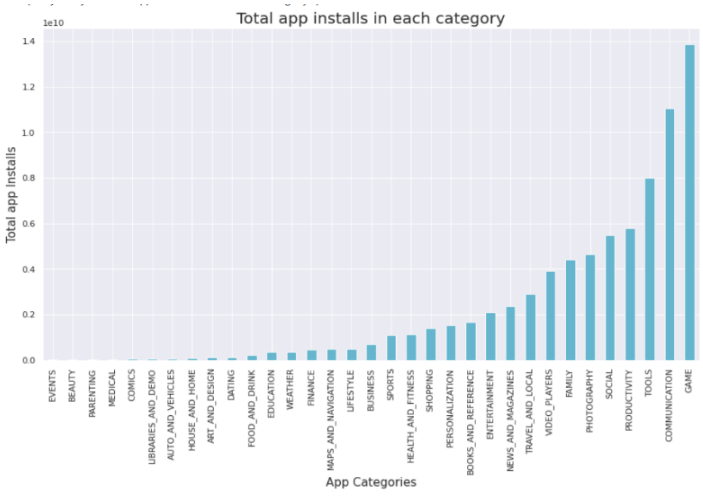

Fig -8: No. of Installs Per Category

This indicates the app category with the highest number of installations. The Game, Communication, and Tools categories boast the highest number of installations compared to other app categories.

3.7 Average App Ratings

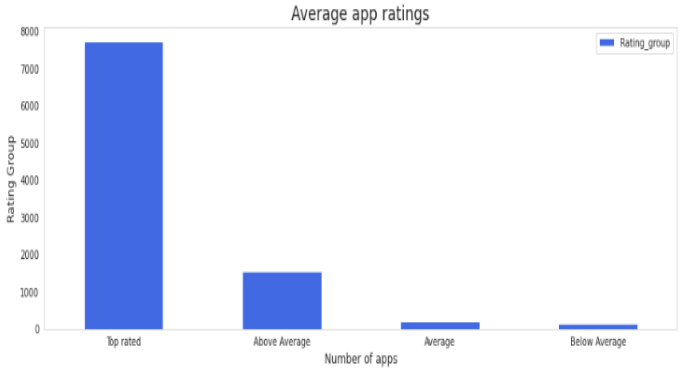

Fig -9: Average App Ratings

The ratings in the dataset exhibit a distributed pattern, and to enhance their representation, we can organize them into intervals. In this context, we can group the ratings as follows:4-5: Top rated

- 3-4: Above average
- 2-3: Average
- 1-2: Below average

3.8 Top Paid Apps per Category

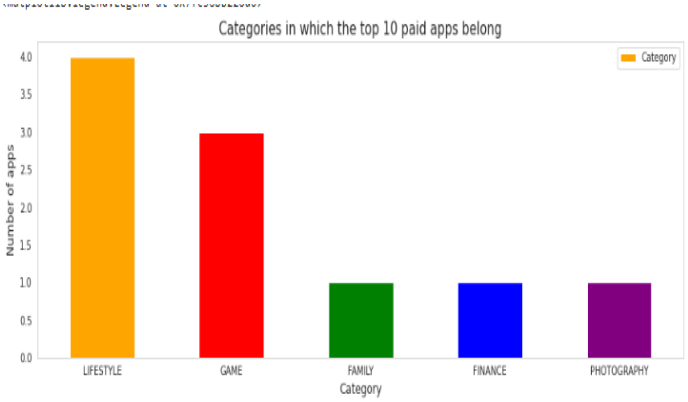

Fig -10: Tops paid app per category

Based on the information above, we can infer that the majority of paid apps fall within the lifestyle and game categories.

3.9 PERCENTAGE OF USER REVIEW SENTIMENTS

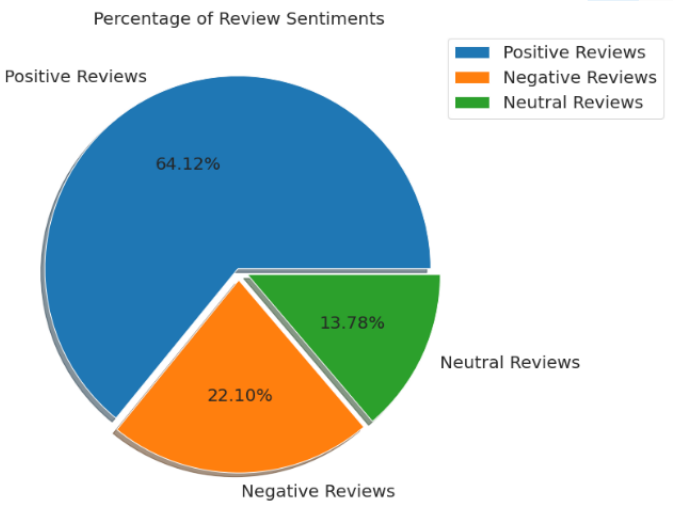

Fig -11: Percentage of User Review Sentiments

Based on the pie chart above, it can be asserted that the majority of apps on the Play Store have received positive reviews from users. However, there are also some apps with negative reviews.

3.10 TOP 10 POSITIVELY REVIEWED APPS

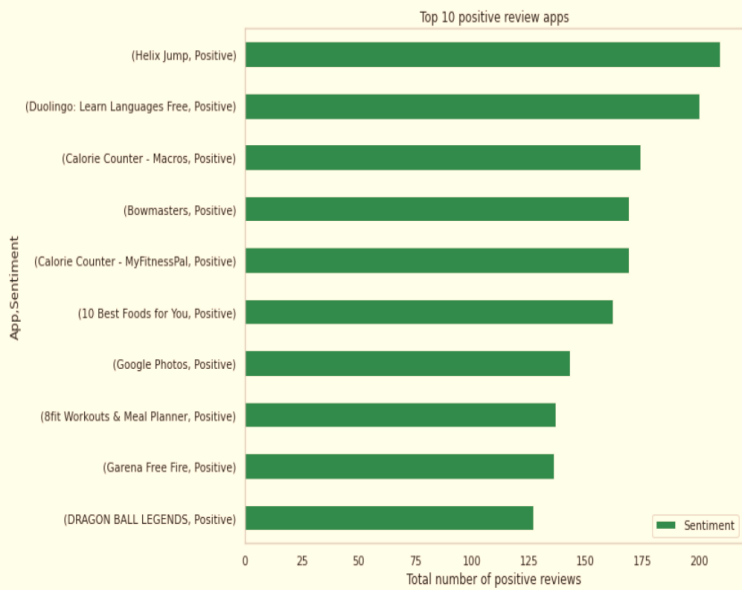

Fig -12: Top 10 Positive Reviewed App

3.11 TOP 10 NEGATIVE REVIEWS APPS

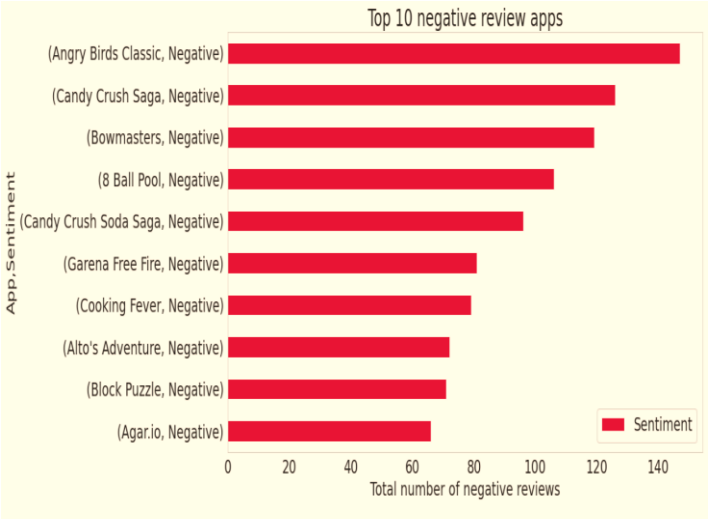

Fig -13: Top 10 Negative Reviewed Apps

3.12 TOP 10 NEGATIVE REVIEWS APPS

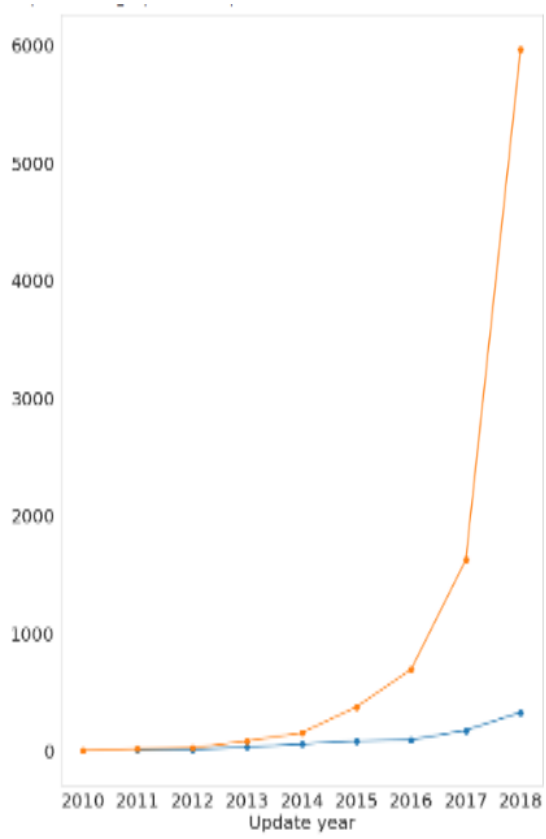

Fig -14: Top 10 Negative Reviewed Apps

3.13 DISTRIBUTION OF APP UPDATE OVER THE YEAR

In the depicted plot, we visualized the apps updated or added over the years, comparing Free vs. Paid categories. From this plot, we can infer that before 2011, no paid apps were available. However, as the years progressed, the addition of free apps outpaced that of paid apps. When comparing the apps updated or added in the years 2011 and 2018, the percentage of free apps increased from 80% to 96%, while paid apps decreased from 20% to 4%. Consequently, we can deduce that a majority of users prefer free apps.

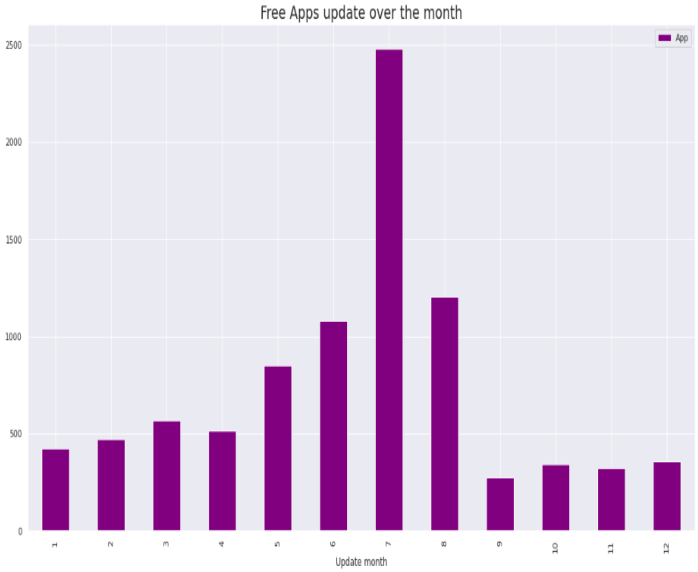

Fig -15: Paid Apps update over the month

3.14 DISTRIBUTION OF APP UPDATE OVER THE MONTH

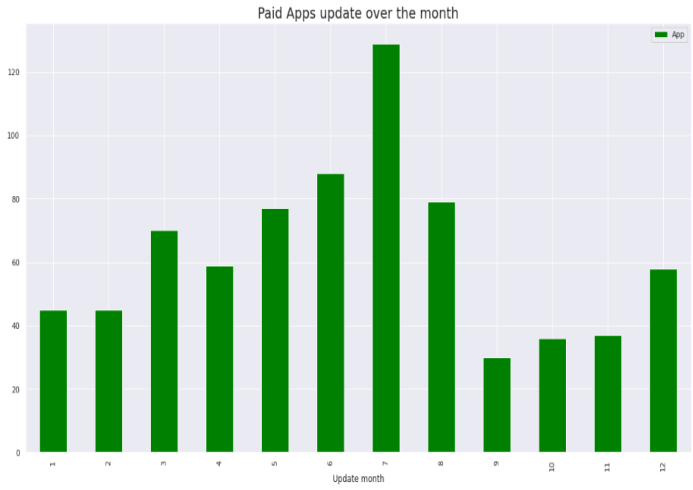

Fig -16: Free Apps update over the month

In this dataset, nearly half of the apps are added or updated in the month of July, followed by 25% in August, and the remaining 25% distributed across the other months.

The majority of paid apps also receive updates in the month of July, similar to free apps.

3.15 RELATIONSHIP BETWEEN SENTIMENT SUBJECTIVITY PROPORTIONAL TO SENTIMENT POLARITY

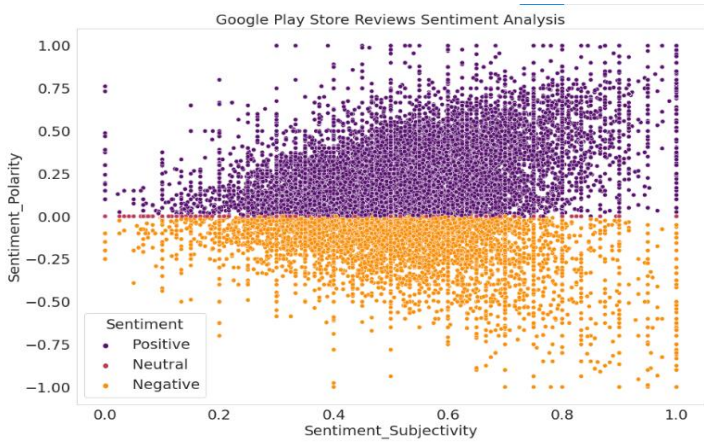

Fig -17: Google play store Reviews Sentiment Analysis

Based on the scatter plot above, it can be inferred that sentiment subjectivity does not always exhibit a proportional relationship to sentiment polarity. However, in the majority of cases, it tends to demonstrate a proportional behavior, especially when the variance is high or low.

3.16 DISTRIBUTION OF SUBJECTIVITY

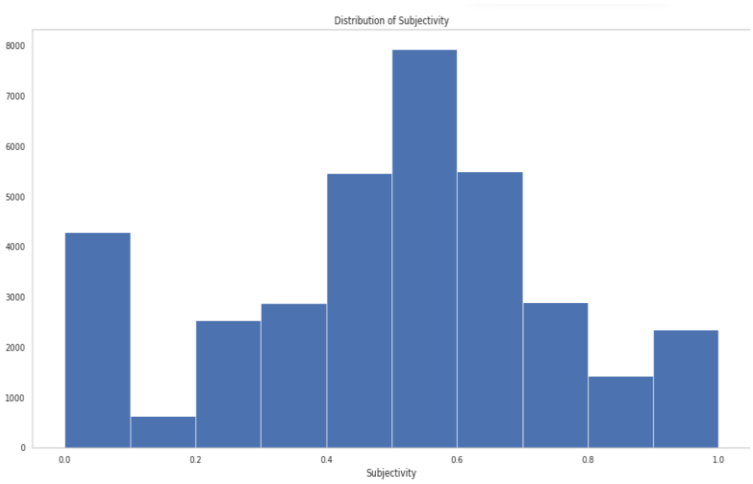

Fig -18: Distribution of subjectivity

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

Observing the data, it is evident that the majority of sentiment subjectivity values fall within the range of 0.4 to 0.7. This suggests that a significant number of users provide reviews for applications based on their experiences.

3.17 RELATIONSHIP BETWEEN DIFFERENT FEATURES OF THE DATASET

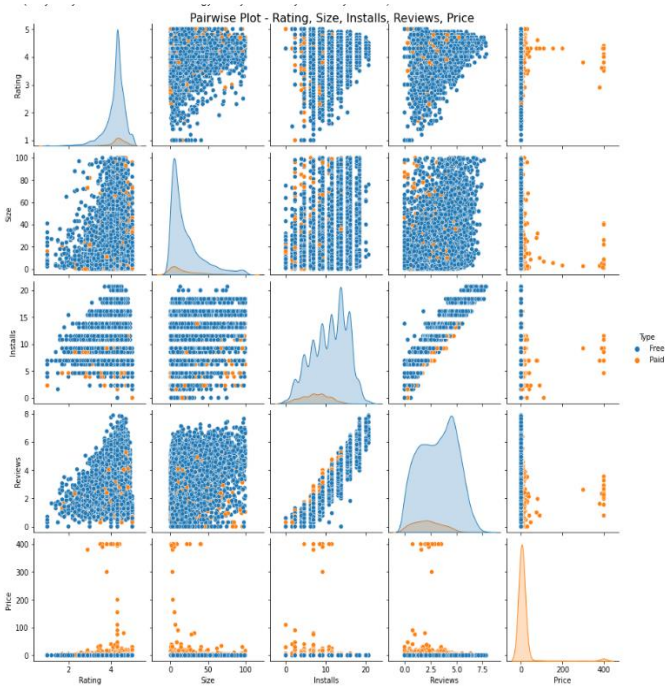

Fig -19: Pair wise plot

- The majority of the apps are available for free.
- The majority of paid apps have a rating of approximately 4.
- With the increase in the number of installations, there is a corresponding increase in the number of reviews for the particular app.
- The majority of the apps have a lightweight design.

3.18 CORRELATION HEATMAP

A correlation matrix is essentially a table that presents correlation coefficients for various variables. This matrix illustrates the correlation between all potential pairs of values in a tabular form. It serves as a potent tool for summarizing extensive datasets and identifying and visualizing patterns within the provided data.

A correlation heatmap visually represents a correlation matrix, illustrating the correlations between various variables. The correlation values can range from -1 to 1. It's important to note that correlation between two random variables or bivariate data does not necessarily imply a causal relationship.

3.19 PLAY STORE CORRELATION HEATMAP

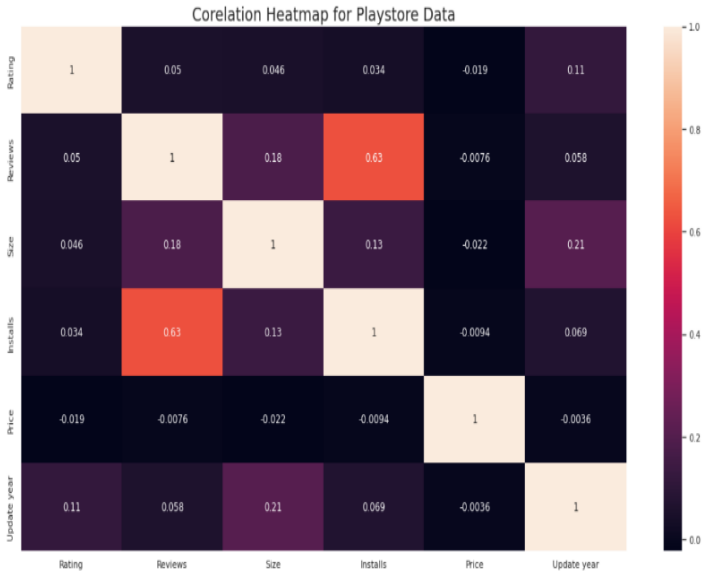

Fig -20: Correlation Heatmap

- A robust positive correlation exists between the Reviews and Installs columns, which is quite evident. As the number of installs increases, so does the user base, resulting in a higher total number of reviews submitted by users.
- The Price exhibits a modest negative correlation with the Rating, Reviews, and Installs. This indicates that as the app prices increase, there is a slight decrease in the average rating, total number of reviews, and installations.
- There is a slight positive correlation between the Rating and the Installs and Reviews columns. This suggests that as the average user rating goes up, there is also an increase in both app installations and the number of reviews.

3.20 MERGED DATA FRAME HEATMAP

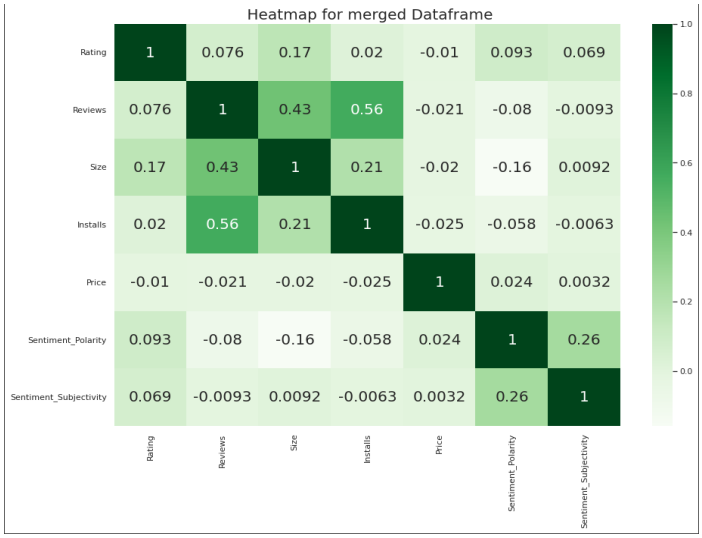

Fig -21: Merged Data frame Heatmap

Conclusion~

Exploring the data has allowed us to identify certain trends and make assumptions that could contribute to the success of an app among users on the Play Store.

- Percentage of free apps = ~92%
- Percentage of apps with no age restrictions = ~82%
- Most competitive category: Family
- 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.

- Family, Game and Tools are top three categories having 1906, 926 and 829 app count.
- Tools, Entertainment, Education, Business and Medical are top Genres.

References~

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


thank
you