Play Store App Review Analysis

Submitted by

SUNIL SHURAJ N (192224019)

Guided by

Dr. S.RAMESH

Associate professor

Department of Applied Machine Learning

Department of Computer Science and Engineering,

Saveetha School of Engineering, SIMATS

Thandalam, Chennai

March – 2024

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.

DATASET ANALYSIS

2.1 Google Play store Dataset

The dataset consists of Google play store application and is taken from Almabetter, which is the world's largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

The data set contains the following columns:

- **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.

2.2 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.

ENVIRONMENTAL SETUP

3.1Python

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 is simplest 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 reasons to 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.

3.2Data 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 fillna() 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.

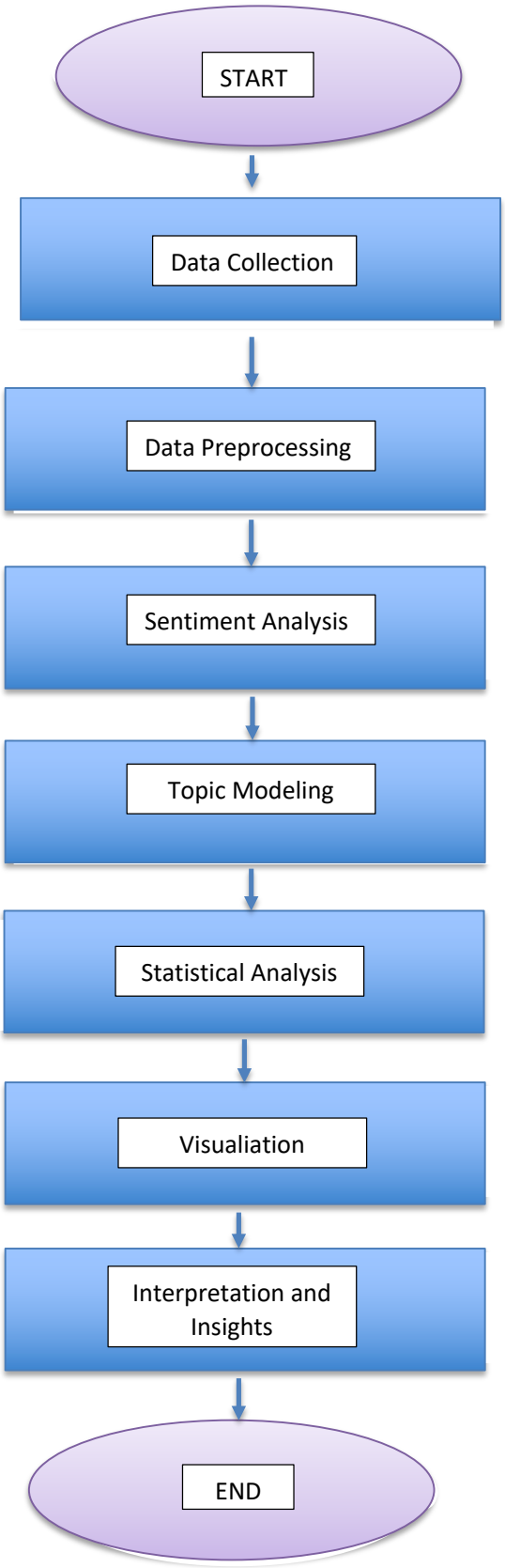
- **Step 8:** The values in the column 'Price' might have the '$' sign in some values and the column is of the datatype 'object'. We will first remove the '$' sign using the **strip()** function and then convert the column into 'int' datatype.

- **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.

DATA FLOW DIAGRAM (OR) ARCHITECTURE DIAGRAM (OR) UML DIAGRAMS



CODE SKELETON

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation
# Step 1: Data Collection
# Code to collect data from Google Play using APIs or web scraping
# Step 2: Data Preprocessing
# Code to clean and preprocess the collected data

# Step 3: Sentiment Analysis
def perform_sentiment_analysis(reviews):
    sia = SentimentIntensityAnalyzer()
    sentiments = []

    for review in reviews:
        sentiment_score = sia.polarity_scores(review)

sentiments.append(sentiment_score['compound'])

    return sentiments

# Step 4: Topic Modeling
def perform_topic_modeling(reviews):
    tfidf_vectorizer = TfidfVectorizer(stop_words='english')
    tfidf = tfidf_vectorizer.fit_transform(reviews)
    lda_model = LatentDirichletAllocation(n_components=5, random_state=42)
    lda_model.fit(tfidf)
    return lda_model

# Step 5: Statistical Analysis
# Code to perform statistical analysis on the data

# Step 6: Visualization
def visualize_data(data):
    # Code to create visualizations using matplotlib, seaborn, etc.
    pass

# Step 7: Interpretation & Insights
# Code to interpret the analysis results and derive insights

# Step 8: Actionable Recommendations
# Code to generate actionable recommendations based on insights

# Step 9: Feedback Loop
# Code to incorporate feedback and insights into the app development process

if __name__ == "__main__":
    # Main code to orchestrate the entire analysis process
    # Call functions for each step of the analysis process
    Pass
```

RESULT 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.

6.1 Free vs Paid

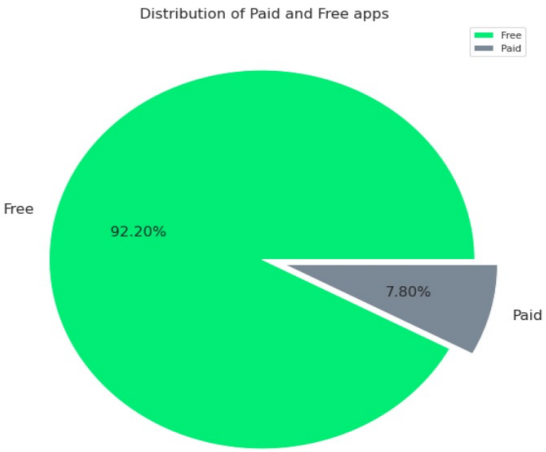


Fig -1: Free vs Paid

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.

6.2 Rating

In the below plot, we plotted the apps Rating

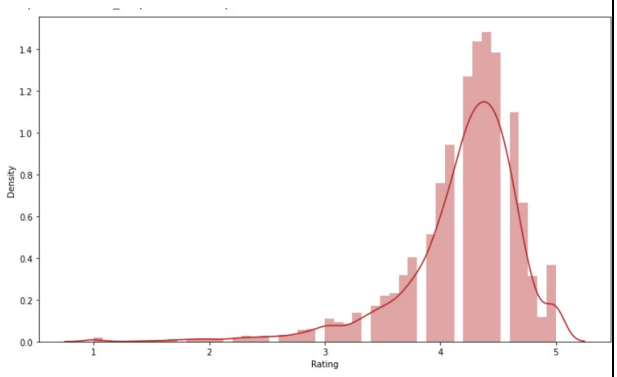


Fig -2: Distribution of App rating

- 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.

6.3 Distribution of App Size

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

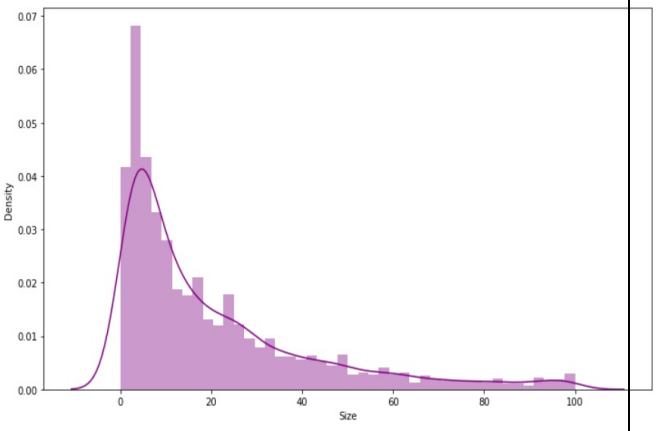


Fig -5: Distribution of App Size

- 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.

6.4 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.

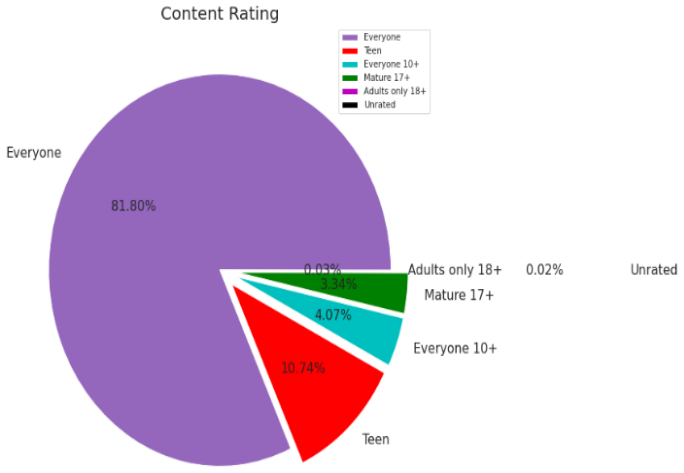


Fig -6: Content rating

6.5 Top Category of Play store

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

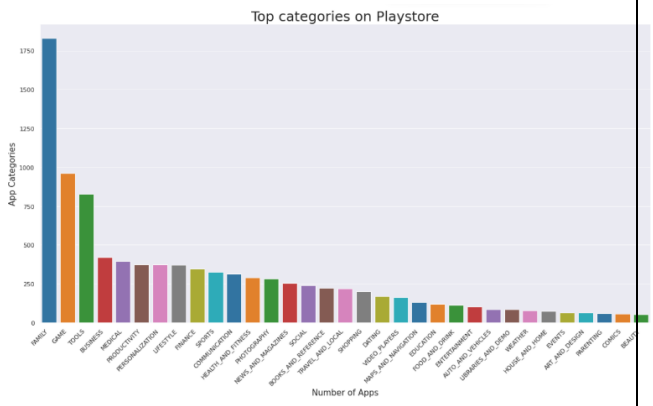


Fig -7: Top Categories on Playstore

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.

6.6 No. of Installs per Category

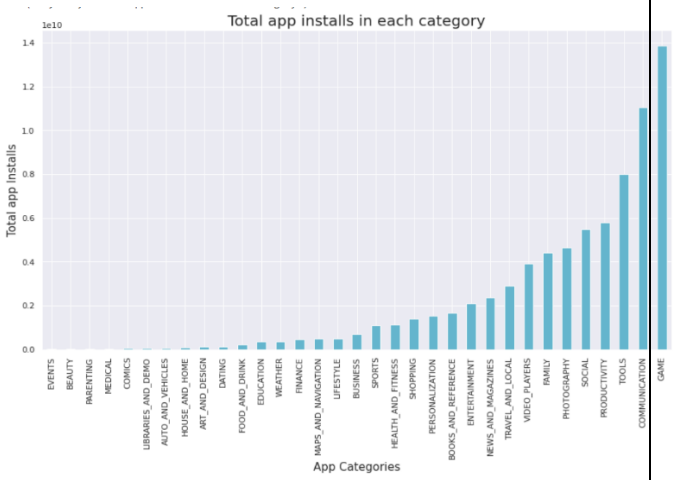


Fig -8: No. of Installs Per Category

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.

6.7 Average App ratings

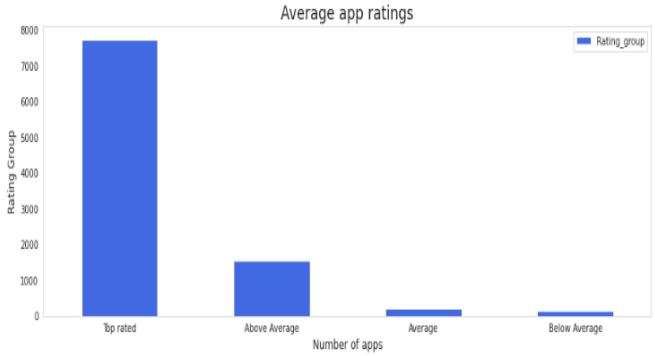


Fig -9: Average App Ratings

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

6.8 Top paid apps per category

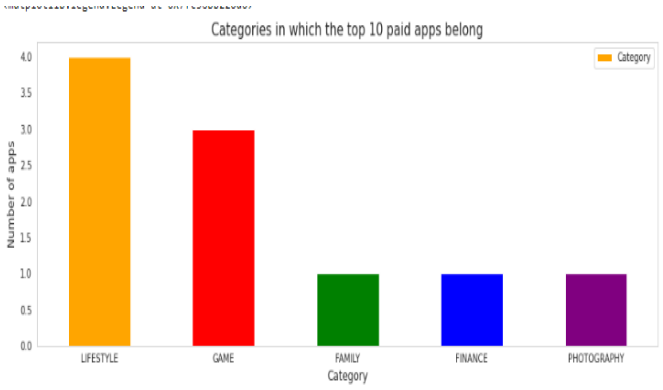


Fig -10: Tops paid app per category

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

6.9 Percentage of User review Sentiments

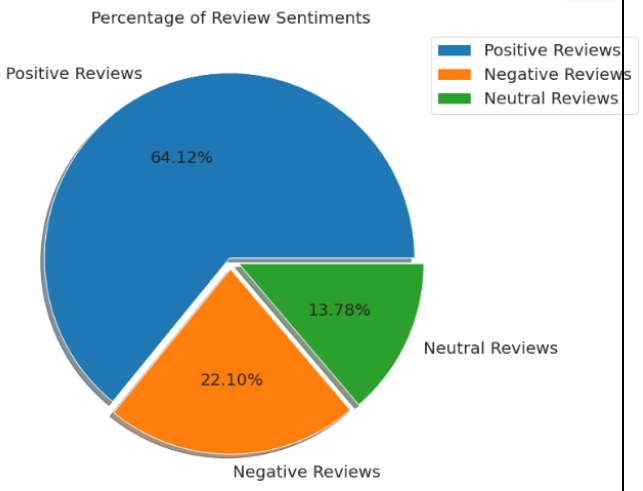


Fig -11: Percentage of User Review Sentiments

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.

6.10 Top 10 positively reviewed Apps

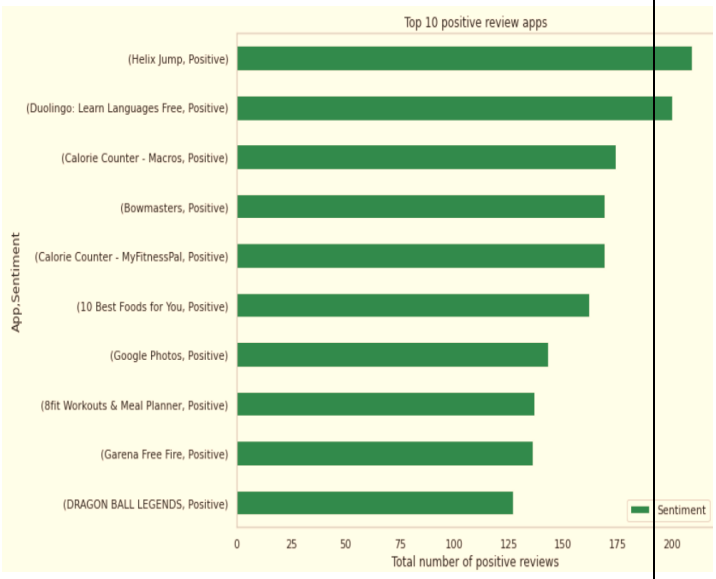


Fig -12: Top 10 Positive Reviewed App

6.11 Top 10 Negative Reviews Apps

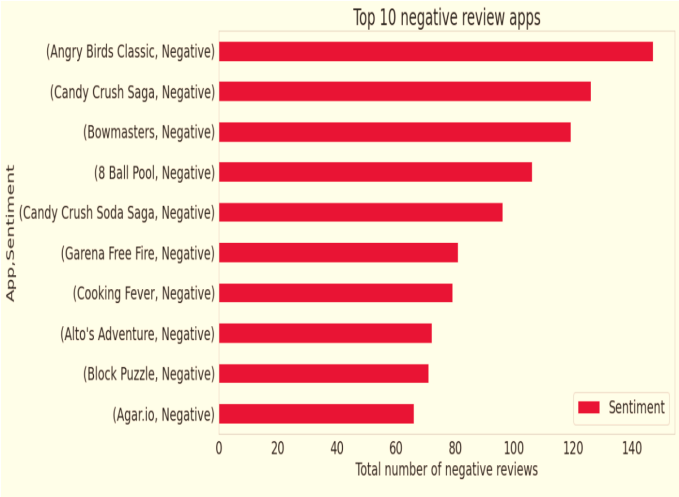


Fig -13: Top 10 Negative Reviewed Apps

6.13 Top 10 Negative Reviews Apps

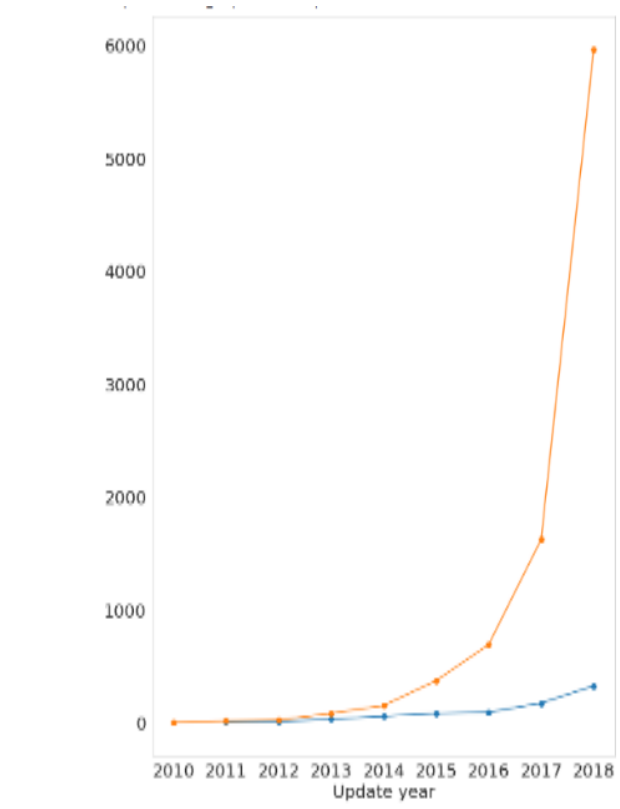


Fig -14: Top 10 Negative Reviewed Apps

6.14Distribution 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.

6.15 Distribution of App update over the Month

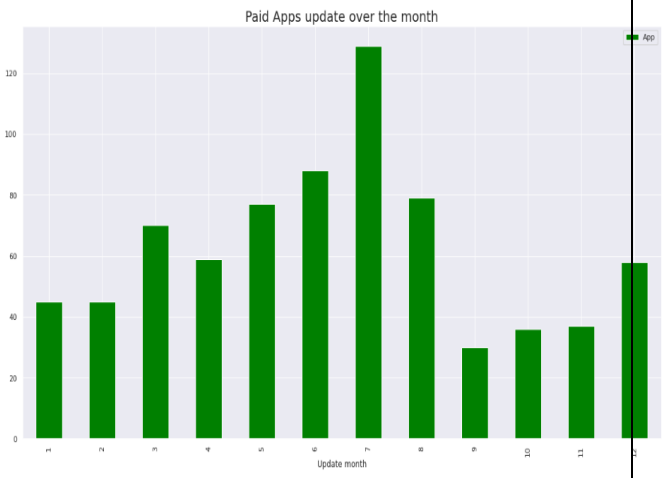


Fig -16: Free Apps update over the month

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.

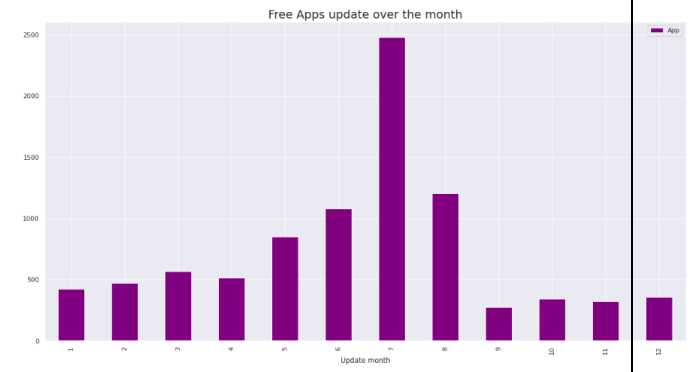


Fig -15: Paid Apps update over the month

6.16 Relationship between sentiment subjectivity proportional to sentiment polarity

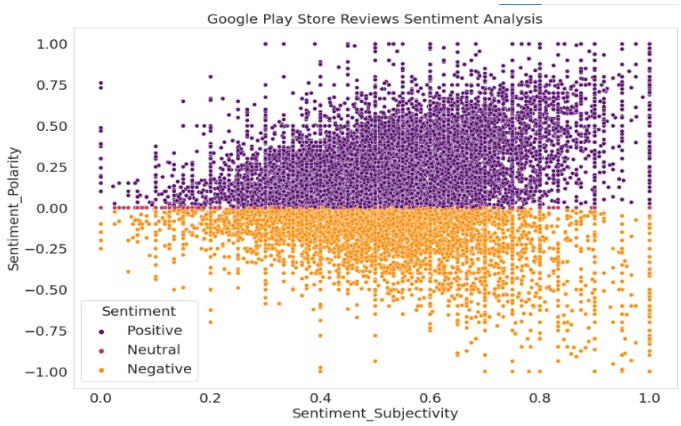


Fig -17: Google play store Reviews Sentiment Analysis

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.

6.17 Distribution of Subjectivity

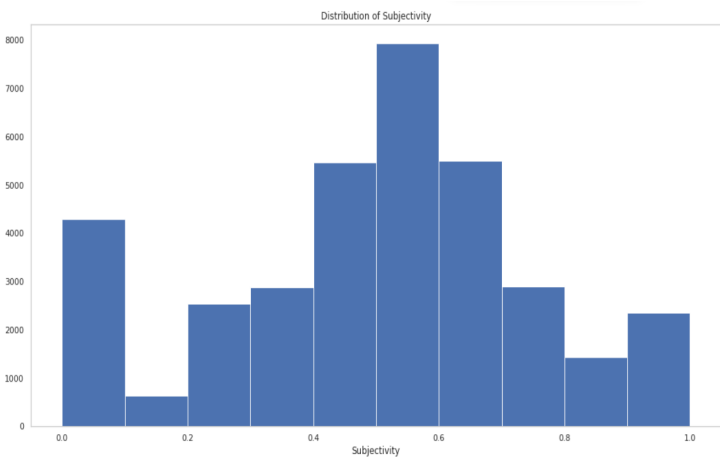


Fig -18: Distribution of subjectivity

`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.

6.18 Relationship between different features of the dataset

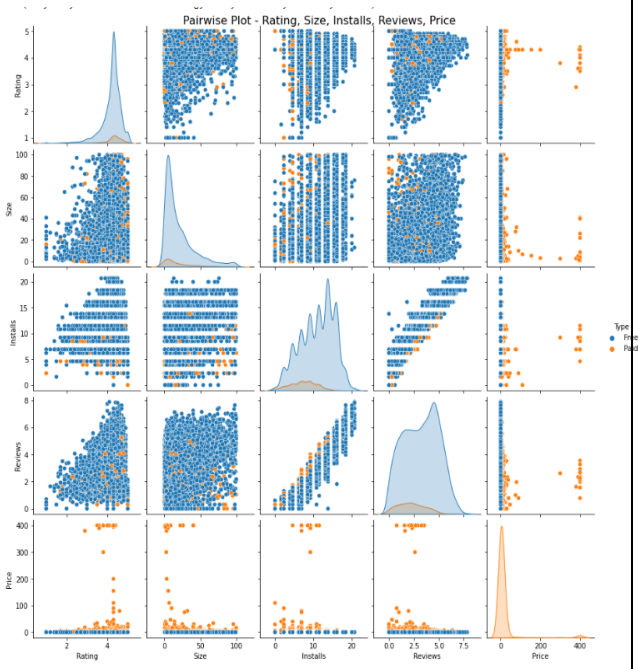


Fig -19: Pair wise plot

- 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.

6.19 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.

6.20 Play store Correlation Heatmap

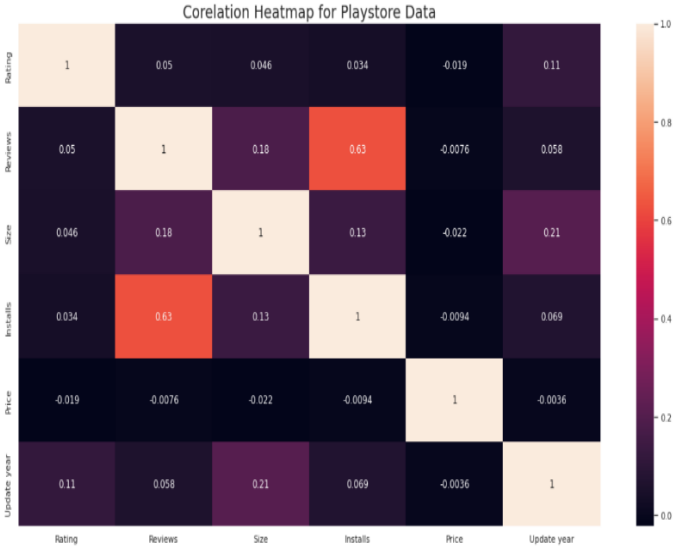


Fig -20: Correlation Heatmap

- 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.

6.21 Merged Data frame Heatmap

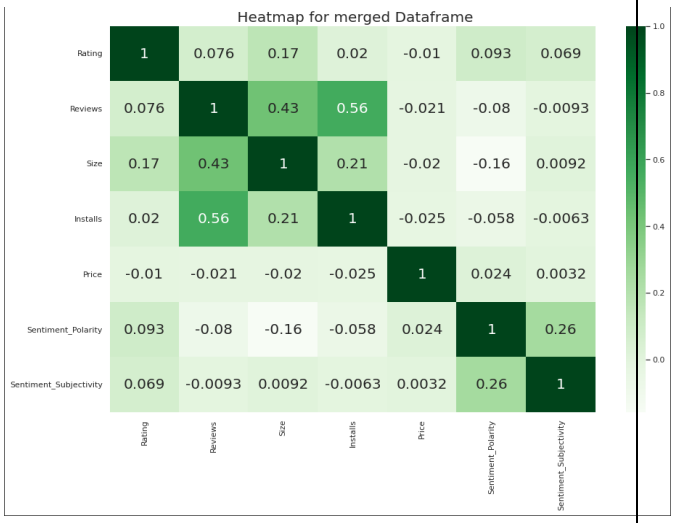


Fig -21: Merged Data frame Heat

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

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