Play Store App Review Analysis

Sapana Pawar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

1. The main primary aspect of the analysis is to identify various categories and the irrespective application available on the google play store. While creating an application there are certain factors that need to be considered.
From the analysis it will find factors required and their respective needs. The current market trend is also displayed. According to this factor we can summarize the market and it will help to create applications according to the latest trend.

2. Another dataset contains customer reviews of the android apps.
Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language.
In Google Play Store, users very often rely on the opinions of others before downloading an application and its reputation could depend entirely on them. This makes analysis of users' reviews very interesting for application owners to make future decisions. In this paper, we are interested in analyzing users' reviews on Play Store using sentiment analysis.

Keywords:Exploratory Data Analysis,Play store apps,Sentiment analysis.

1.Problem Statement

The Play Store app's data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.

Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.

The main object of this EDA is to explore and analyze the features of the dataset in order to help the developers to understand the trends within the market and the end user needs towards the application, and to discover key factors responsible for app engagement and success.

Following are the features present in the provided datasets.

- App - Application name
- Category - Category the app belongs to
- Rating - Overall user rating of the app (as when scraped)
- Reviews - Number of user reviews for the app
- Size - Size of the app
- Installs - Number of user downloads/installs for the app

- Type - Paid or Free
- Price - Price of the app
- Content Rating - Age group the app is targeted at - Children / Mature 21+ / Adult
- Genres - An app can belong to multiple genres (apart from its main category).
- Last updated - Date when the app was last updated on Play Store
- Current ver - Current version of the app available on Play Store
- Android Ver - Minimum required Android Version
- Translated_Review - Review given by user
- Sentiment - A view or opinion that is held or expressed. It can be Positive, Negative or Neutral.
- Sentiment Polarity - It defines the orientation of the expressed sentiment. (Range[-1.0,1.0])
- Sentiment Subjectivity - Subjectivity quantifies the amount of personal opinion and factual information contained in the text. (Range[0.0,1.0])

2. Introduction

In the android world there are n number applications available on the android market. The android market is also known as Google Play Store.

Play Store contains multiple applications, these multiple applications are used for various purposes and for some it is also used for daily purposes, social media apps like whatsapp, Facebook etc. Android is the dominant mobile operating system today with most of all mobile devices running Google’s OS. The Google Play Store is the largest and most popular Android app store. The aim of our project is to gather and analyse detailed information on apps in the Google Play Store in order to provide details on app features and the current state of the Android app market.

Mobile app reviews are key drivers when it comes to the success of your new app.The purpose of analysis is to find out which apps dominate the most and what are the important factors to create an application. For a developer to create such apps needs data that can be gathered in this analysis.

3. Exploratory Data Analysis(EDA)

First we performed EDA on Play Store App Review data

- **Importing Python Libraries and loading dataset**

We have imported some important python libraries to perform EDA on a given dataset. For analysis,manipulation and imputation purposes, libraries like Numpy and Pandas are imported.

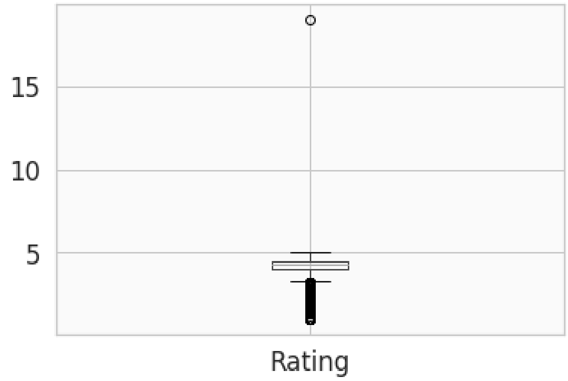
Visual Representation of data makes it easier to identify and share real time trends, outliers and new insights about the information represented in data. For data visualization python libraries like Matplotlib and Seaborn are imported.

- **Exploration**

Data exploration gives some basic information about the dataset. It can include functions like .head and .tail(gives the top and last 5 rows of data), .info, .shape and .describe functions give the basic information,shape and descriptive summary of the dataset.

- **Outliers detection and removal**

After data exploration we looked for outliers present in data by plotting a box plot. An outlier is defined as a data point that is located outside the whiskers of the box plot.



By plotting a box plot over the data we got an outlier in the Rating column. Outlier Rating value is 19, Since rating lies between [0.0,5.0] we dropped it.

- **Null values treatment**

If null/missing values are present in the dataset that could affect our analysis. In our dataset columns namely Rating, Type, Current ver and Android Ver contains 1474,1,8 and 2 respectively.

For null values present in the numerical column (Rating) we have replaced them with the median value of that column. And for null values present in the categorical columns (Type, Current Ver, Android Ver) we have replaced them with the mode of that particular column.

After imputation now our dataset is free from all the null values.

- **Correcting Datatypes of Columns**

On observing the dataset we found that some of the columns(Price,Reviews,Installs and Size) have integer/ float values but the datatype of those columns is of 'Object' type. So to do analysis we have to change the datatype of those columns to int/float.

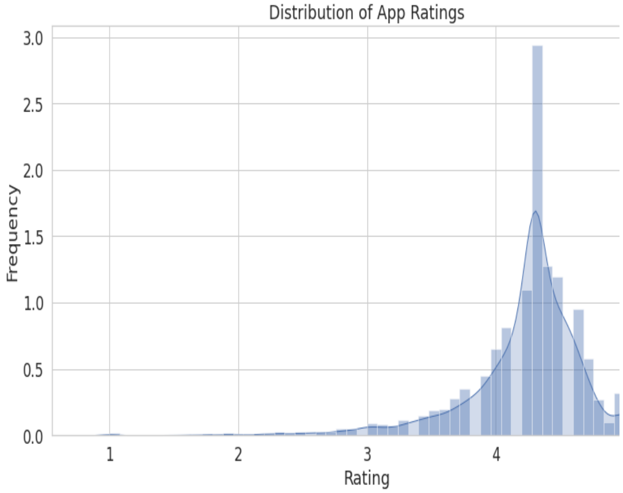
Before converting the datatype of columns we had removed some signs and characters, like ‘$’ in the Price column, ‘+’ in the Installs column and ‘M’,’k’ present in the Size column . And after that we changed the datatype of those columns. Also we changed the datatype of the ‘Last_Updated’ column from string to Datetime.

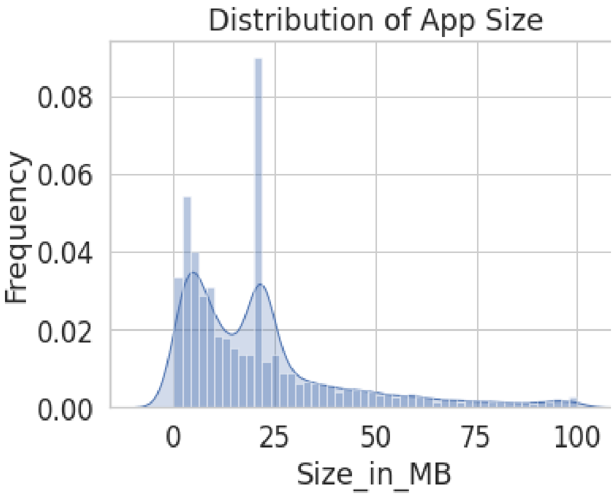
● Data Visualization

To perform EDA on a given dataset we had prepared some questions and answered them to get some meaningful conclusions from it.

Following are the questions that we have prepared and answered.

1.How does the distribution of Rating and size look like?

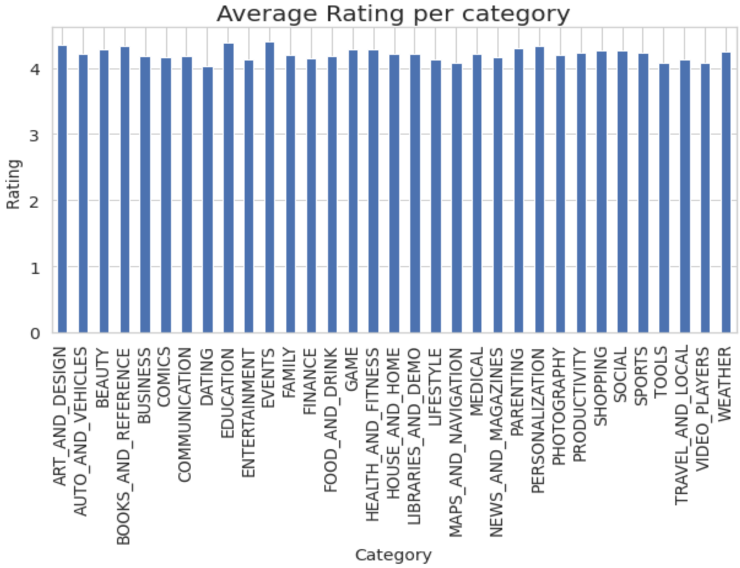




From above visualization we can conclude that

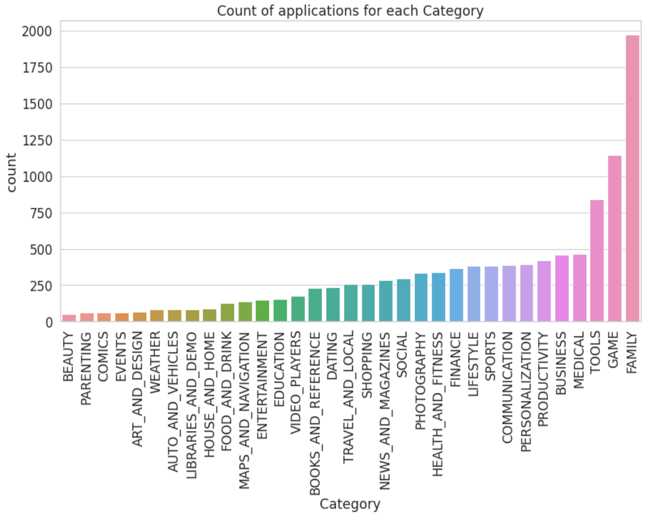
- Most of the apps have ratings between 3 and 5.
- Maximum number of applications present in the dataset are of small size.

2.What is the average rating per category?

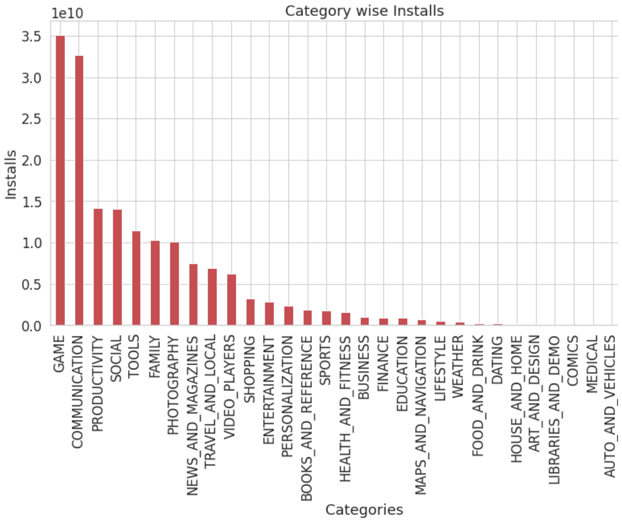


- All categories of apps have more than 4 average ratings.

3.Which category has a high number of installs?

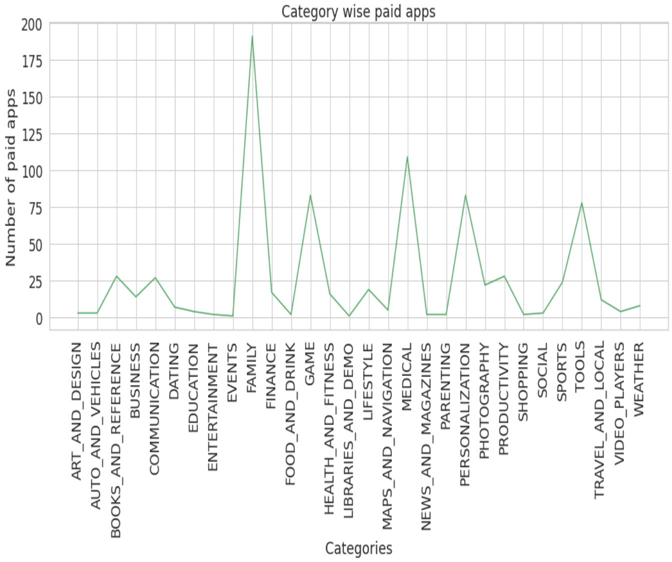


- The categories Family, Game and tools have the most apps on the Play Store.



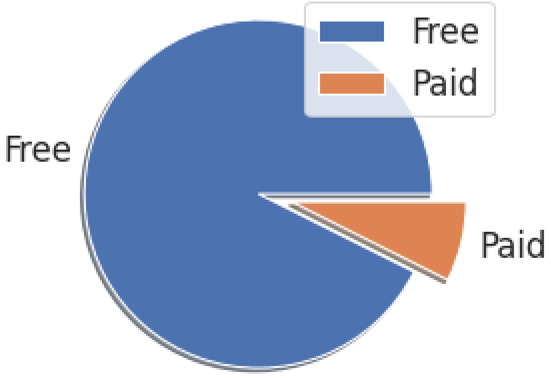
- Maximum number of apps present in Google Play Store comes under Family, Game and tools but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps come under Game, Communication, Productivity and Social.
- Subway Surfers, Facebook, Messenger andGoogle Drive are the most installed apps.

4.Which category has the most paid type of apps?



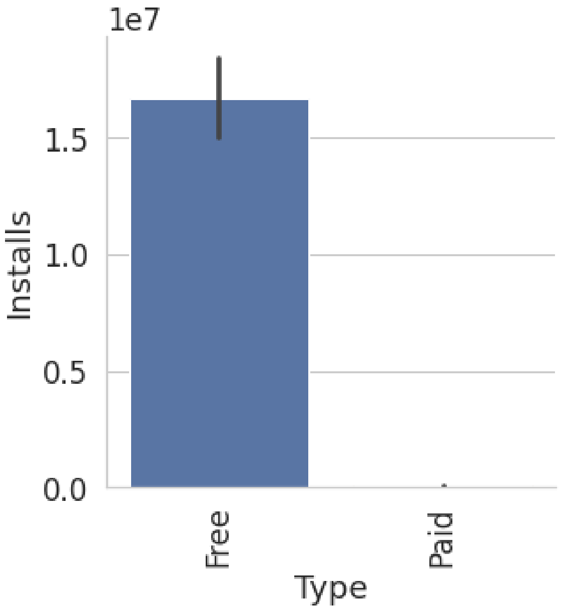
- The category ‘Family’ has the highest number of paid apps.
- The app “I’m Rich — Trump Edition” from the category ‘Lifestyle’ is the most costly app priced at $400

5.What is the percentage of paid and free apps?



- About 92% apps are free and 8% apps are of paid type.

Type vs Installs



- Free apps are installed more than paid apps.

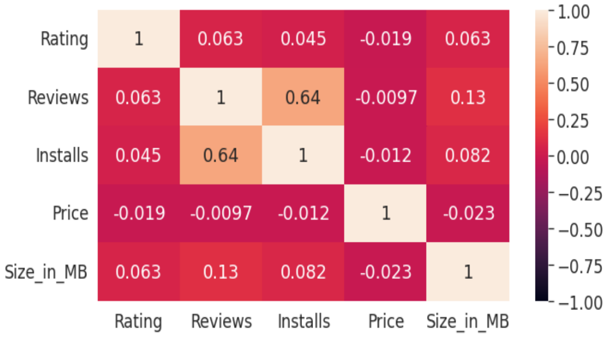
6.What are the installs per content rating of apps?

	Content Rating	Installs
1	Everyone	1.141567e+11
2	Teen	3.471635e+10
3	Everyone 10+	1.323388e+10
4	Mature 17+	5.524491e+09
5	Adults only 18+	2.000000e+06
6	Unrated	5.050000e+04

- Content having Everyone only has most installs, while unrated and Adults only 18+ have less installs.

7.Visualize the correlation between all the columns with the help of a heatmap.

Correlation heatmap is a graphical representation of a correlation matrix representing correlation between different variables. The value of correlation can take any value from -1 to 1.



- From this correlation heatmap we can see how variables are correlated with each other.
- Number of Reviews are positively correlated with number of Installs with correlation 0.64

Now we perform EDA on user_review data

Analyzing user sentiments towards apps through their review comments and ratings can be economically profitable to app developers.
We had done sentiment analysis on user_review data.

- **Null values Treatment**

On exploring user review data with the help of some pandas library functions, we observed that four out of five columns contain near about 41% null values, So we dropped them because it has no use.
After removing Null values from the dataframe, the data frame contains 37427 rows and 5 columns.

- **Outliers detection and removal**

We checked outliers for Sentiment Polarity, and Sentiment Subjectivity columns by seeing the minimum and maximum value present in the columns.
Values of sentiment polarity lie between -1 to 1, so there is no outlier present.
Values of sentiment subjectivity lie between 0 to 1, so there is no outlier present.

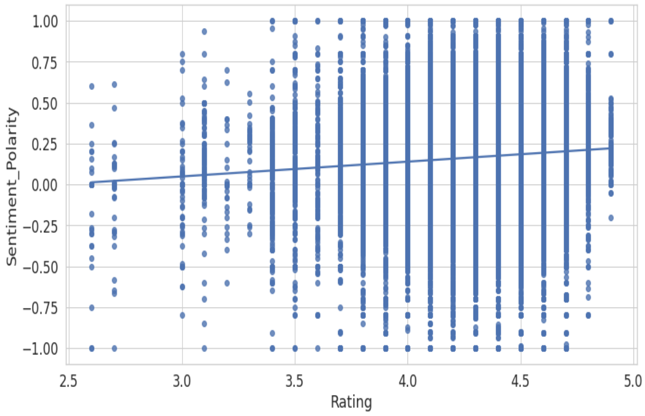
- **Merging Dataframes on App**

We merged Play store App data and user_review data on the App column.
On merging datasets, many duplicate rows are generated. So we dropped duplicates by keeping the first ones. Now the shape of merged_df is (40414,18)

- **Correlations, Trends and conclusions**

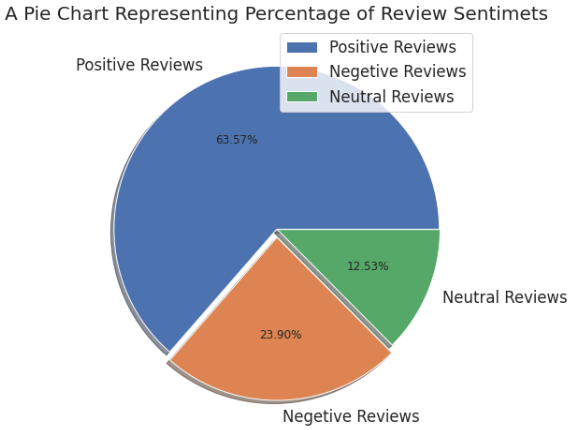
We have created some plots from the data with the help of powerful data visualization libraries like Matplotlib and Seaborn to understand data better.
By visualizing, we got some interesting trends and conclusions. Following are the graphs that we have plotted and made some conclusions from.

Regplot between Rating and Sentiment Polarity



The correlation between the sentiment and the rating is not as strong as we thought, though we can see the trend there.

Percentage of Review Sentiment

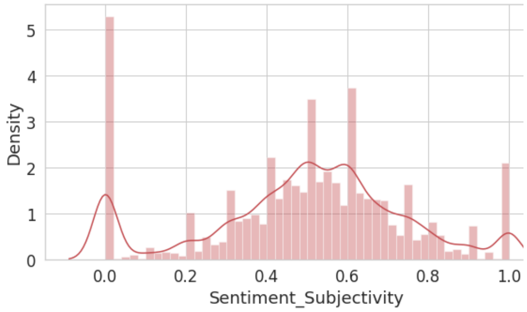


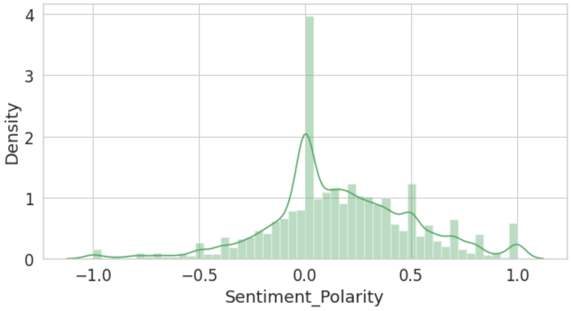
- Most of the reviews are of Positive Sentiment, while Negative and Neutral have a low number of reviews.

Distributions of Sentiment Polarity and Subjectivity

- Most of the reviews fall in [-0.50,0.75] Polarity scale, extremely negative or positive sentiments are significantly low.

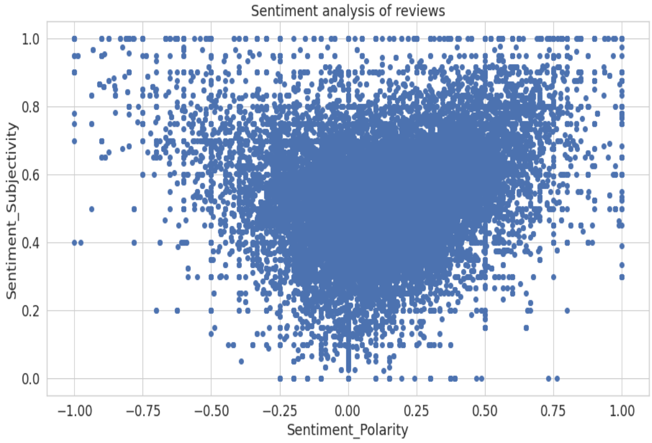
- Most of the reviews fall in the [0.3,0.8] Subjectivity scale.





Sentiment Polarity vs Sentiment Subjectivity

Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.



Above figure represents a plot of polarity and subjectivity for all the reviews. While the polarity is concentrated mostly in the center, the subjectivity is spread out across the graph. This indicates that our collection of reviews shows a wide range of subjectivity and most of the reviews fall in [-0.50,0.75] polarity scale implying that the extremely negative or positive sentiments are significantly low.

While the users have shared their complete opinions as well as facts about the apps, most of the reviews show a mid-range of negative and positive sentiments. In the graph, reviews with low subjectivity are concentrated at the center of the polarity range [-1, +1] and the reviews with high subjectivity are scattered across the polarity range [-1, +1]. This is understandable because a fact (low subjectivity) is more likely to be neutral (with polarity 0) and an opinion (high subjectivity) is more likely to have a diverse range of negative to positive sentiments.

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

Challenges Faced

- Handling Null Values
- Changing the datatype of misclassified datatypes of columns

Conclusion

That's it! We reached the end of our exercise.
Starting with importing important Python libraries so far we have done Exploration, Null values treatment, Outlier detection and removal, Correcting datatypes of columns and Data visualization.
In this analysis we plotted some interesting graphs which show trends and correlation between the variables.
The analysis of Google Play Store applications aided to build more reliable and more interactive applications. This would be very useful for app developers to build an application focussed on certain discussed categories in this analysis. This analysis will definitely help in building the application with precise and accurate objectives.

References

1. Stack Overflow
2. GeeksforGeeks