Play store App Review Analysis

Hariharasudhan M,
Data science trainees,
AlmaBetter, Bangalore.

Abstract:

The Play store is the greatest source of Android apps. The most leading and safest apps are published here.

Our features based analysis on the database can help to find the key factors which are responsible for the app engagement, popularity and success.

Keywords: Exploratory Data Analysis (EDA), Reviews, Rating, Distribution.

1. Problem Statement

Data provided by Play store, which is operated and developed by Google. Play store has a million number of similar or different kinds of apps which are tagged by different Genres. These apps are available for both free and paid. They are also providing different features for the users to choose and download the best apps from it.

- App name: This is the column which contains the name of the app.
- Category: The Category is a column by which the apps got separated based on the application and purpose.

They are very useful for us to characterize an app from different points of view and provide to the people across available options. They have been in operation since the year 2008. During this period, they have optimized a lot and are more secure from any android app resources.

The main objective is to build an analysis, which could help the ordinary people to find the best apps based on the previous data. This would in turn help them in matching the right cabs with the right customers quickly and efficiently. Reviews: The distance for the trip requested by the customer

- Size: Size is a factor which is not the same for all the devices, and it fully depends on the developers.
- Installs: Installs gives us the data of total installed users count.
- Type: Type defines whether the app is free or paid.
- Price: Here price is mentioned for the paid apps, for free apps 0 will be given.
- Rating: Rating plays a huge role in finding the correct apps. It was manually given by the users.
- Genres: Genres are like tags, an app can be coming under more than one Genres, based on the usage.

- Android Version: This feature gives us the supported device versions of android.

2. Introduction

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

3. Exploring the database

We have provided with two databases

- Shape of this database is (10841, 13).
- Out of this thirteen columns we have numeric

User reviews database

- Shape of this database is (64295, 5).
- Here there are only two numeric values found,
- Sentiment Subjectivity, Sentiment Polarity.

4. Features selection

The columns are also known as features, one or more different features are grouped together for different analyses to form a data frame.

top_genres_df: It contains the rows of top ‘Genres’ based on the app counts.

genres_free_apps_installs_df: It holds the rows of top ‘Genres’ and corresponding total free app installs count.

genres_paid_apps_installs_df: It holds the rows of top ‘Genres’ and corresponding total paid app installs count.

genres_ratings_df: It takes the rows of top ‘Genres’ and their corresponding mean Rating.

top_50_genres_df: This data frame contains the rows of features top Genres, Free app installs count, Paid app installs count and Rating.

categoty_type_installs_df: It contains the rows of Category wise free and paid apps installed.

price_df: This holds the rows of paid apps’ non null price values.

categrory_price_mean_df: This holds the rows of Category wise mean price of paid apps.

rating_df: This holds the rows of mean price for all the apps.

category_mean_rating_df: It contains the rows of Category wise mean rating of all the apps.

content_rating_df: It holds the rows of different aged people wise ratings count.

non_null_user_reviews_df: It holds the different users' non null Sentiment type, Sentiment Subjectivity distribution and Sentiment Polarity.

sentiment_subjectivity_df: It holds the different users' sentiment subjectivity distribution.

5. Top 'Genres' analysis

The analyze is done between the top 50 Genres which was having more number of downloads

- As we saw here, the Play store having more number of applications in the genres like Tools, Entertainment, Education and etc.,
- The developers are mostly focusing on these genres because of the people's daily basis requirements.
- Genres like Educational, Parenting, Music are having comparatively less amount of apps count.

Top Free apps

- When comparing the both plots, people are showing more interest on free apps like communication, Tools
- Again the Educational, Parenting and Music are the genres in the least

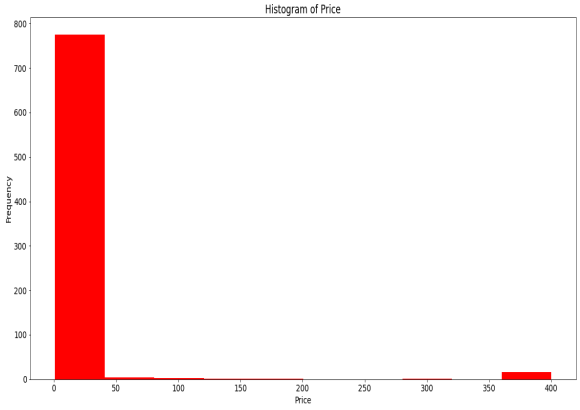
Top Paid apps

- When it's coming, commercial people are preferring apps like Games, Photography, and Personalization.

- People are preferring less on Educational, Event, and Art & Design.

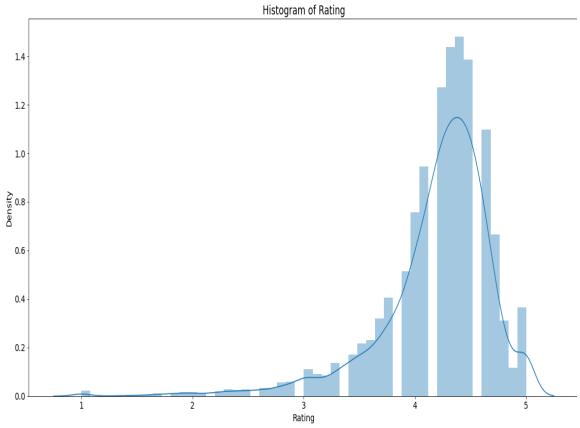
6. Price distribution

- The distribution of Price plotted as we see here.
- Most of the paid apps are in between the price range of 0 to 50 USD.
- Few apps are in the price range of 350 to 400 USD.
- Very few of the apps are priced between 75 to 200 USD.



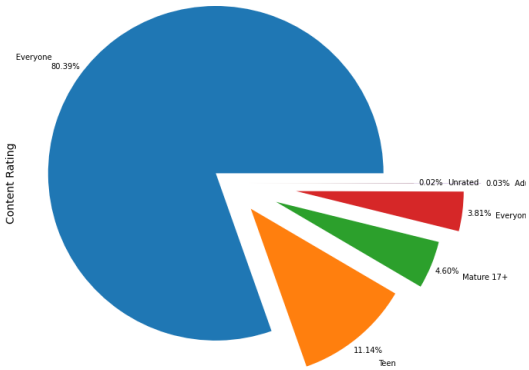
7.1. Rating-Distribution:

- In general Rating is the main scale factor, which measures how much people are satisfied with the product.



- Here, most of the people rated the apps between the ranges of 4 to 5, which can be considered as good.
- Moderate rating lies between 2.5 to 4.
- And below average or poor apps are in the range of 0 to 2.5 which are less in count.

7.2. Impact of 'Rating':

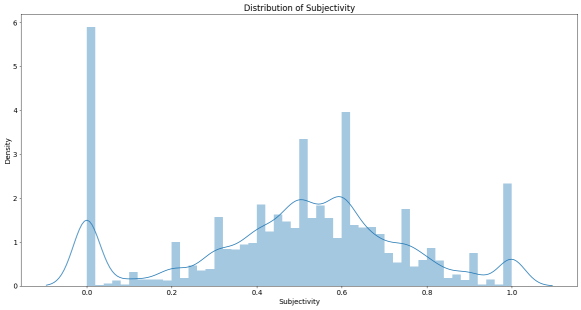


- The content rating shows the results for general contents as high.
- The content rating type 'Everyone' has the most percentage value of 80.39%.
- 'Teen' contents are second in the order with the percentage of 11.14%.
- Adult's only and unrated contents are least in this plot, 0.03% and 0.02% respectively.
- So it can be concluded that most of the contents are generic.

8. User sentiment analysis:

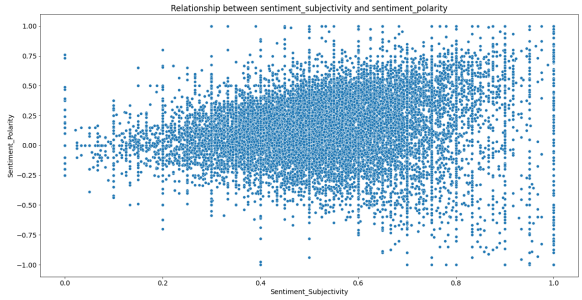
Distribution of Subjectivity:

- Subjectivity lies mostly between 0.5 and 0.65.



- It shows that the average content and apps reviews subjectivity are mostly relevant.
- Subjectivity of 100% has slightly occurred frequently.
- The nearly 0 subjectivity has a considerable amount of frequency.

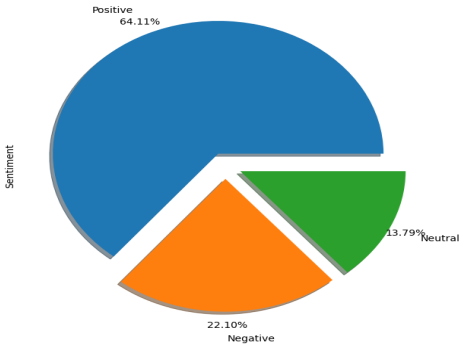
Relationship between Subjectivity and polarity:



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

Sentiment percentage:

- The sentiment plot shows the results for positive reviews as high.
- Sentiment type ‘Positive’ has the most percentage value of 64.11%.
- ‘Negative’ reviews are with the percentage of 22.10%.
- ‘Neutral’ percentage is 13.79%.
-



9. Conclusion:

After analyzing the dataset we have got answers to some of the serious & interesting facts which any of the android users would love to know.

- Top Genres on Google Play store.
- Top Categories on Google Play store.
- Which Category of Content is downloaded more?
- Which category of apps has the most number of installs?
- What are the Top 10 installed apps in different categories?
- Distribution of the ratings of the apps
- Variation between Free and Paid apps
- Which are the top expensive Apps?
- Distribution of the Price of the apps
- Which are the apps that have made the highest-earning?
- Which are the Apps with the highest number of reviews?

References-

1. Stackoverflow
2. GeeksforGeeks
3. matplotlib.org