Play store App Review Analysis

JAVED AHMAD,
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.

• 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.

Introduction

- Play store is an app market for Android operating systems developed and regulated by Google, offering apps of wide variety and purposes, currently there are 3.48 million apps on play store and the number is still growing by an impressive rate. Its home to app kinds of app and provide them majorly free of cost. It is trusted by billions around the world due it its security policies and surveys. Thus studying this large amount of data makes it important to understand trends and get other valuable Insights from a developers point of view.

Logos representing play store are:

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.

Methodology

We collected the data of play store and user reviews from website of alma better. Our basic approach was to make a copy of original clean the data and make it ready for data analysis and data visualization. Steps will be explained in details below.

Step 1:- Data obtaining

- Google play store apps data frame and user review data frame were obtained from official alma better website. Data was then imported into running environment i.e., google collab in this case.

Step 2:- Data overview

- Necessary copies of data were made to save the parent data from any permanent modifications. General layout of data was studied using .info (), .describe () and .head() methods.

Step 3:- Data cleaning

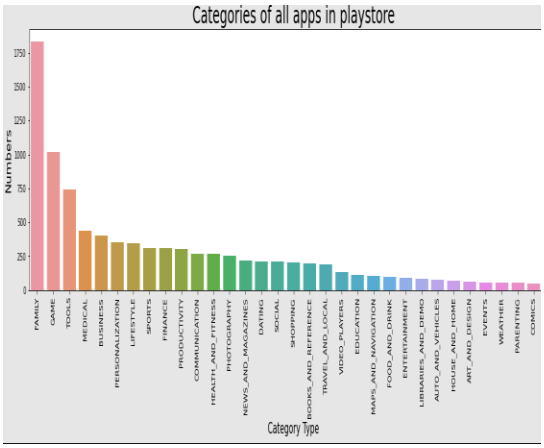
- Function “complete_info()” was made to review the dataset again and again after processing to determine the null values ,duplicates and data types.
- Duplicates were determined and dropped from the data, using ‘.dropna()” method.
- Some columns were processed and null values were replaced by median if applicable or else dropped in order to make the data complete.

–

- **Step 4:- Data preparation**
- Converting some columns datatype in more suitable datatypes, appropriate for data analysis i.e.
- Size: - It was in from of object (M, K, varies with device, etc.) with this units thus function was made to convert them all into simple numerical string and then column was converted to float using ".to_numeric()" method. KB is converted to MB using formula mb= kb/1024 in function.
- Installs: - It was in the form of object e.g., 500,000,000+, which was converted to number after replacing '+' and 'comma' with suitable characters, to get 500000000 in integer format.
- Price: - It was in format of object from with unit "$" thus converted to basic numerical value by removing "$" and converting it to float.
- Reviews: - This column was clean thus directly converted it to integer using ".astype(int)" function.

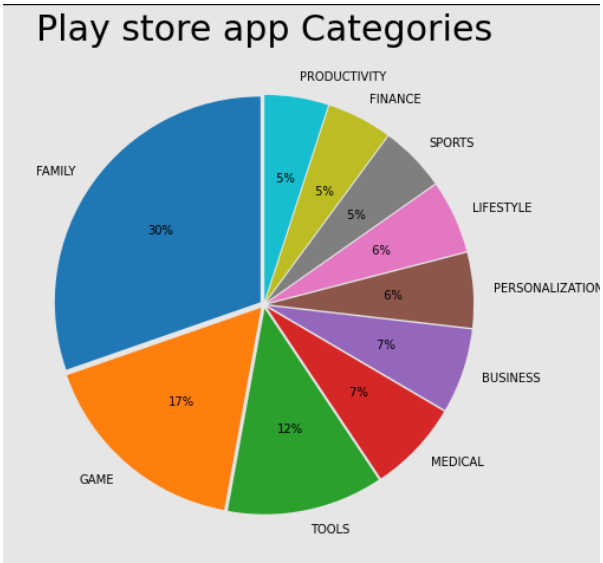
- **Step 5:- Data wrangling & data visualization**
- Data was explored based on categories, in total 33 unique categories were found in the data.

- **Number of Apps with categories**



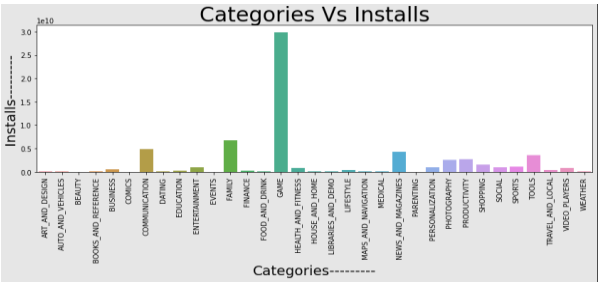
- Based on the 33 categories data set result most application belong to Family,Game and Tools categories and least contribution of Comics and Beauty category in google play store.

- **Top 10 Categories**



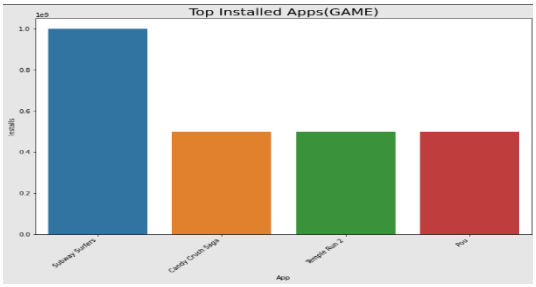
- From above pie chart of top 10 Categories 29% apps belong to family,17% app belong to game categories

- **Top Installed Categories vs Categories**



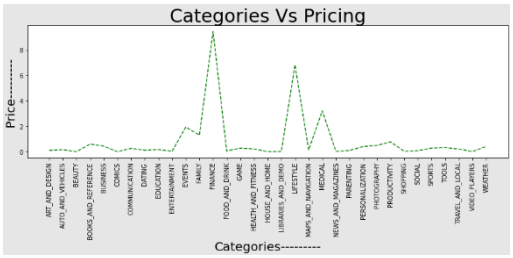
- We can see that most of the apps has been installed from 'Game' category.

- **Top Installed gaming Apps**



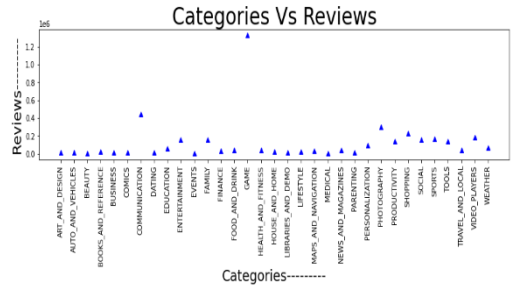
- From the above plot we can see that the most installed games from the 'Game' Category are 'Subway Suffers', 'Candy Crush Saga',’Temple Run 2’and 'Pou'.

- **Apps Categories vs Pricing**



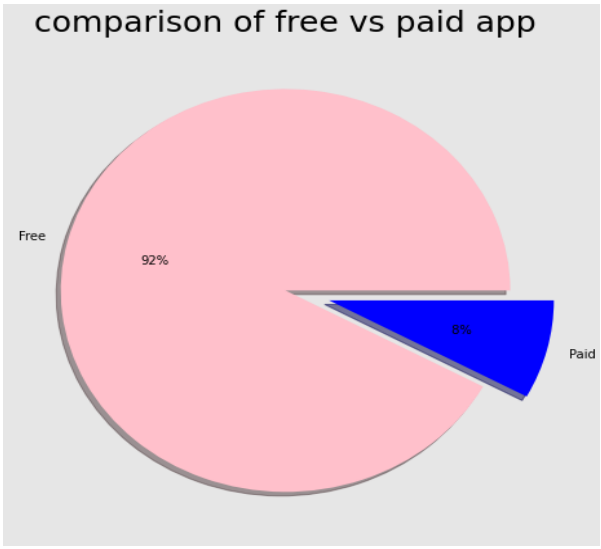
- By above graph it visualize that the prices are high of those apps who belong to Finance and Lifestyle categories.

- **Categories vs Reviews Chart**



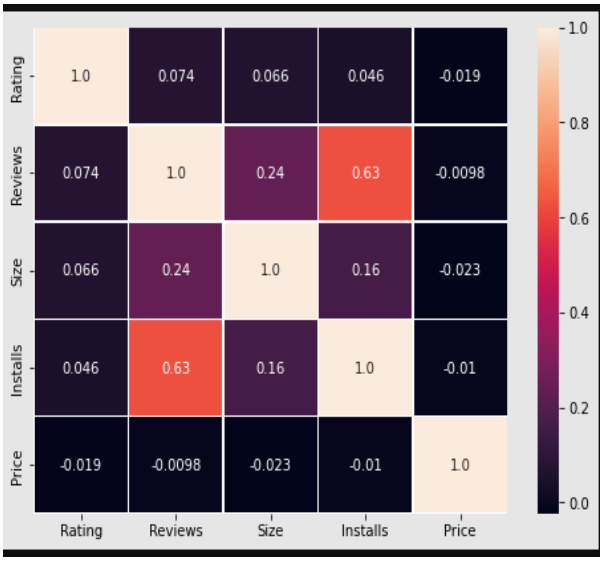
- The reviews are Game categories app is so high..

- **Free Vs Paid Applications**



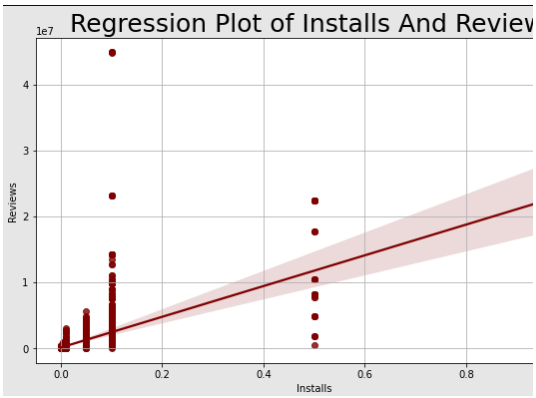
- From the pie chart we can see that 92% of apps in google play store are free and 8% apps are paid.

- **Correlation Heatmap of Columns**



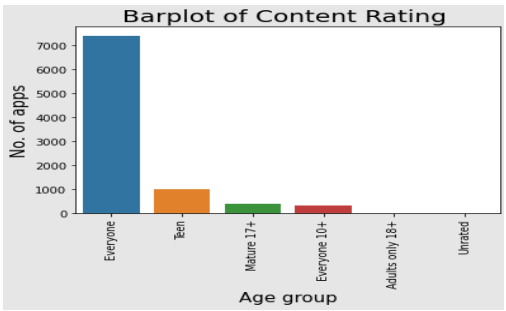
- From the above heatmap we can see that the 'Reviews' and 'Installs' columns has most correlation.
- It is much more obvious that a higher number of installs has a higher number of reviews. There is a negative correlation between price and install apps, with the price of the app influencing the number of installation of the app.

- **Regression analysis**



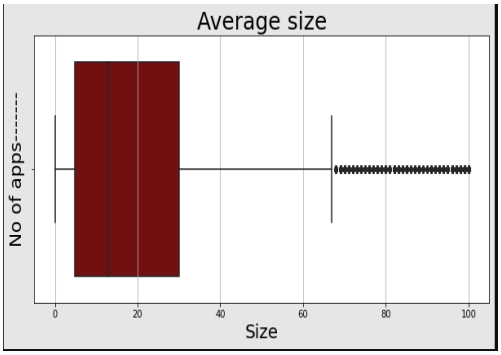
- By above Regression plot it clearly seen that Installs and Reviews are highly corelated.

- **NO. OF Apps vs Content Ratings**



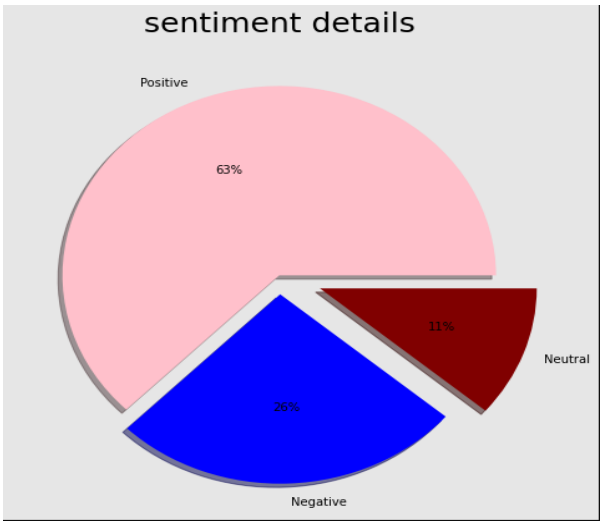
- From above plot we can seen the most apps come under the Everyone content rating. Content rating of adult and unrated apps are very less**.**

- **Box plot of Apps according to size**



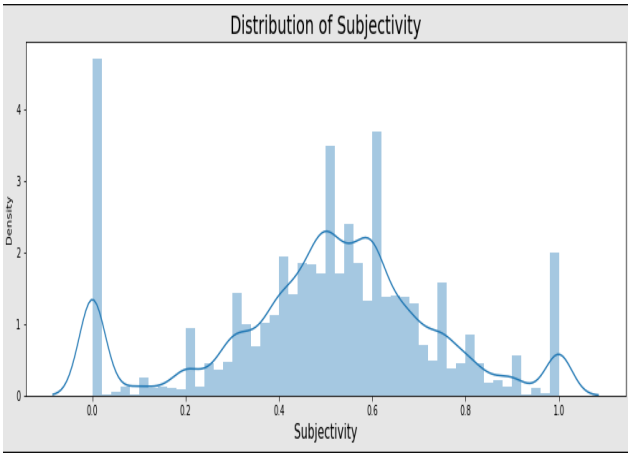
- .By above box plot we observe that almost 50% apps size lies between 5 MB to 30 MB. Some apps are also present which size greater than the 70 MB.

- **Percentage of Review Sentiments**



- By above pie chart we can say that positive reviews is high that is 63% ,negative reviews are 26% and neutral reviews are 11%.

- **Distribution subjectivity**



- The 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.

- ## Conclusion:

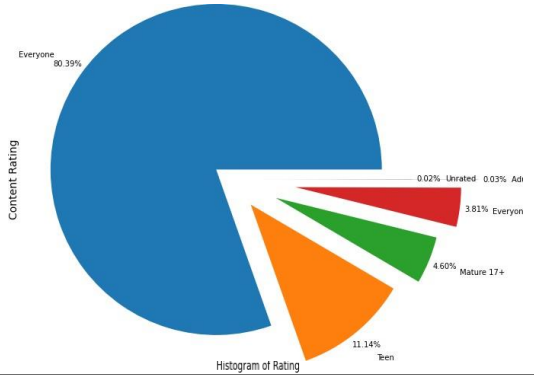
- After analyzing the dataset we have got answers to some of the serious & interestingfacts which any of the android users would love to know.

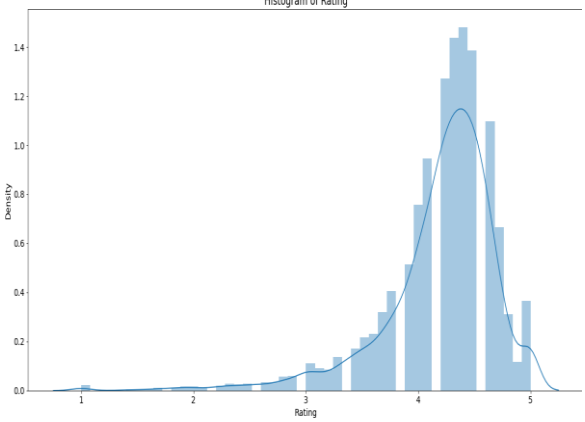
- Top Categories on Google Play store.
- Which Category Content is downloaded more?
- Which category of apps has the most number of installs?
- What are the Top installed Games?
- Distribution of the size of apps
- Comparison between Free and Paidapps
- Which category apps are expensive?
- Distribution of the Price of the apps
- Which Categories Apps with the highest number of reviews?

References-

1. Stackoverflow
2. GeeksforGeeks
3. Matplotlib.org

g
,
:





- 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.

- 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%.

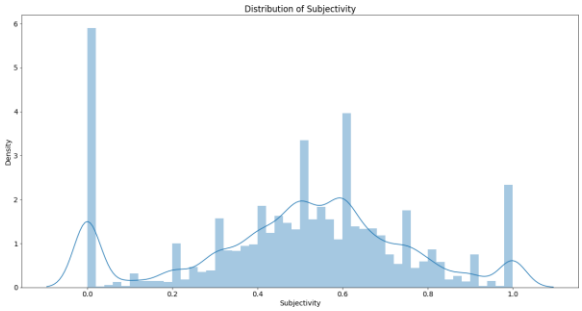
7.2. Impact of ‘Rating

- Adult’s only and unrated contents areleast in this plot, 0.03% and 0.02% respectively.
- So it can be concluded that most ofthe contents are generic.

8. User sentiment analysis:

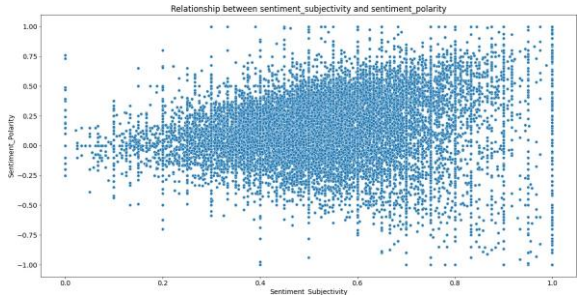
Distribution of Subjectivity:

- Subjectivity lies mostly between 0.5and 0.65.



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

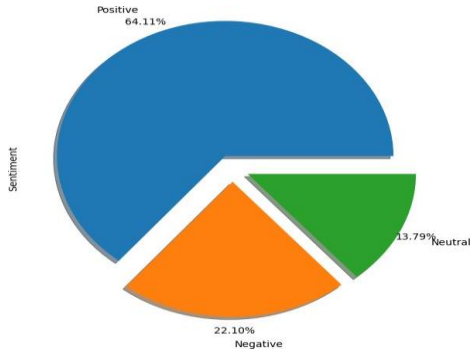
Relationship between Subjectivity and polarity:



From the above scatter plot it can be concluded that sentiment subjectivity isnot 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-

3. Stackoverflow
4. GeeksforGeeks
5. matplotlib.org