Capstone Project-EDA

Play Store App Review Analysis

Abstract:-

The Play Store apps 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.

Explore and analyse the data to discover key factors responsible for app engagement and success.

Key Words:-

Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Android market, customer reviews, **A**pp engagement and success.

The data set-1 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.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10840 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10840 non-null  object 
 1   Category        10840 non-null  object 
 2   Rating          9366 non-null   float64
 3   Reviews         10840 non-null  object 
 4   Size            10840 non-null  object 
 5   Installs        10840 non-null  object 
 6   Type            10839 non-null  object 
 7   Price           10840 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10840 non-null  object 
 10  Last Updated    10840 non-null  object 
 11  Current Ver     10832 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.2+ MB
```

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.

Data Cleaning and Preparation

Pre-processing is important into transitioning raw data into a more desirable format. Undergoing the pre-processing 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 pre-processing 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.

```
App                0
Category           0
Rating          1463
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     0
Genres             0
Last Updated       0
Current Ver        8
Android Ver        2
dtype: int64
```

```
App                0
Category           0
Rating             0
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     0
Genres             0
Last Updated       0
Current Ver        8
Android Ver        2
dtype: int64
```

- **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' data type.

- **Step 7:** The 'Installs' column values contain the characters '+' and ',' which are going to prevent us from converting this column into a numeric data type. We will get rid of these using the strip () and replace () functions.

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

- **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.

EXPLORATORY DATA ANALYSIS

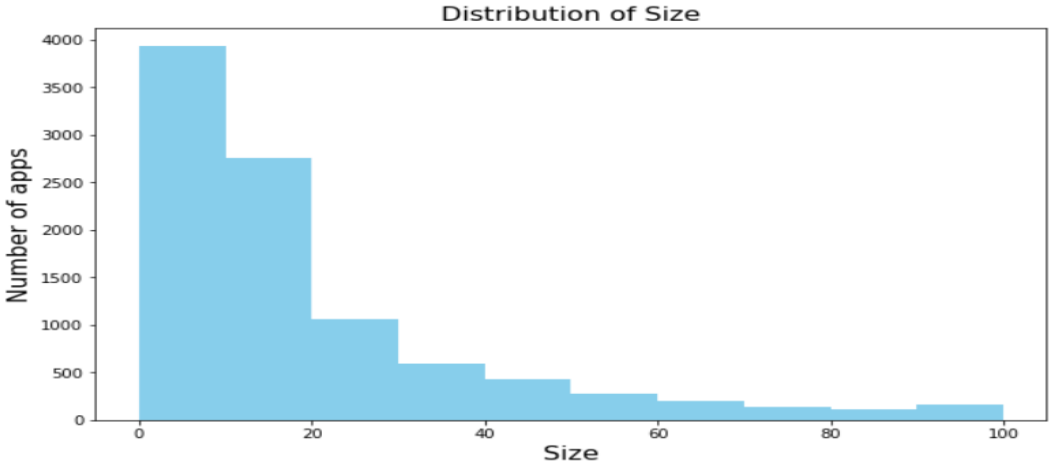
Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

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.

Free vs. Paid-

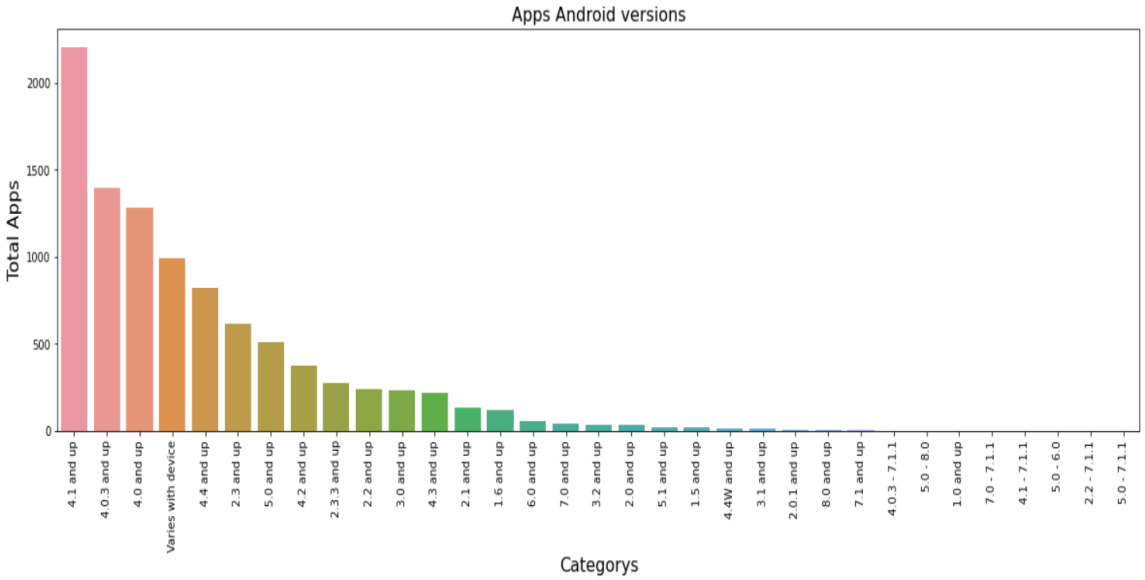
Here we can see that 92.2% apps are free, and 7.80% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

Apps size distributions-



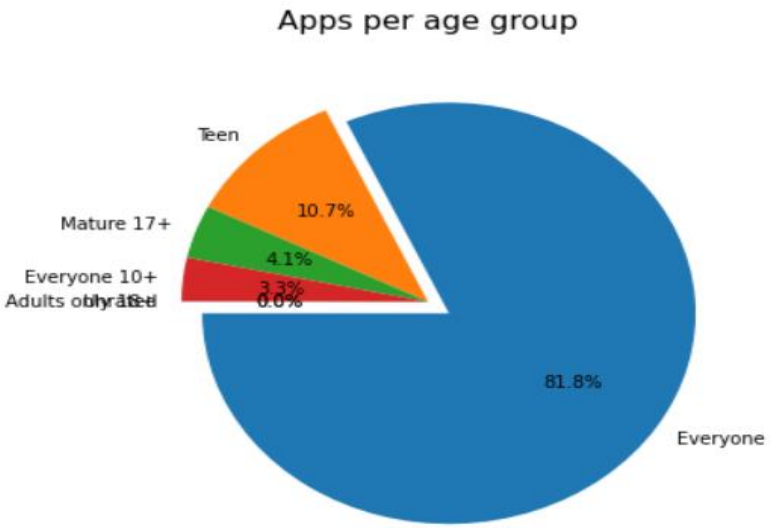
Most of the app's size between 0 to 20Mb, most of the apps are small in size

Distribution of Apps based on versions-



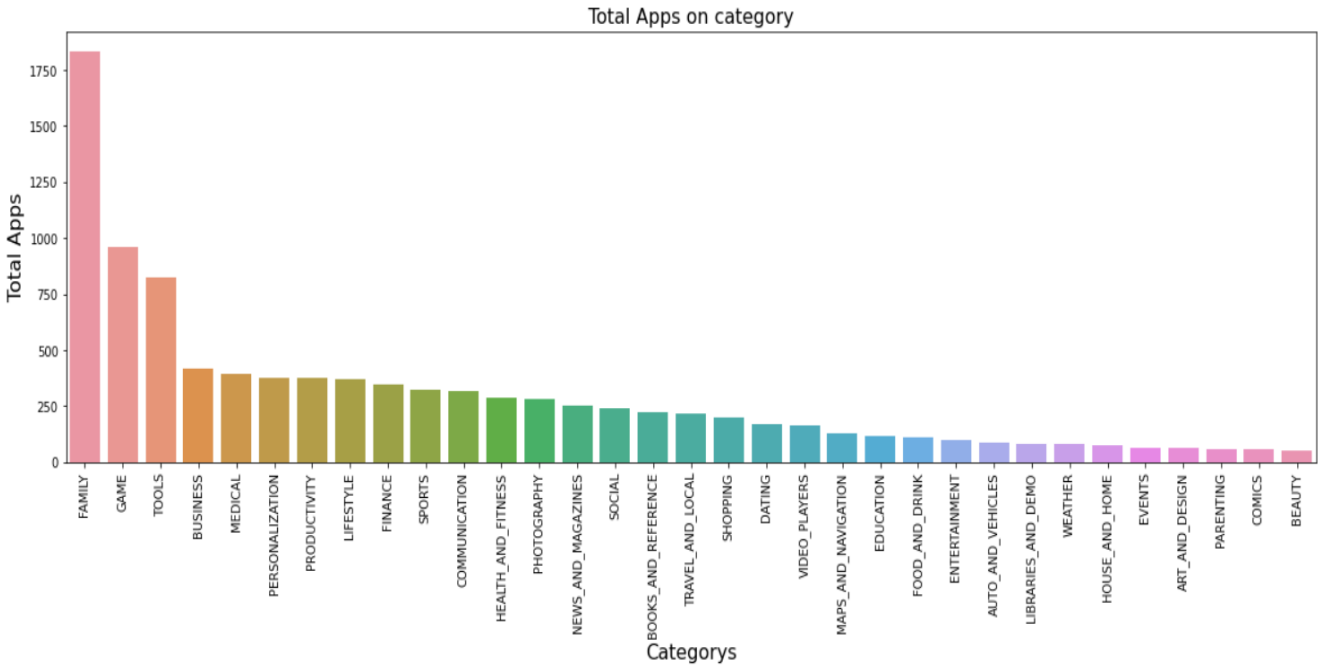
- It is clear from the visualizations that the data in the **Size** column is skewed towards the right.
- Also, we see that a vast majority of the entries in this column are of the value **Varies with device**, replacing this with any central tendency value (mean or median) may give incorrect visualizations and results. Hence these values are left as it is.

Apps per age group-



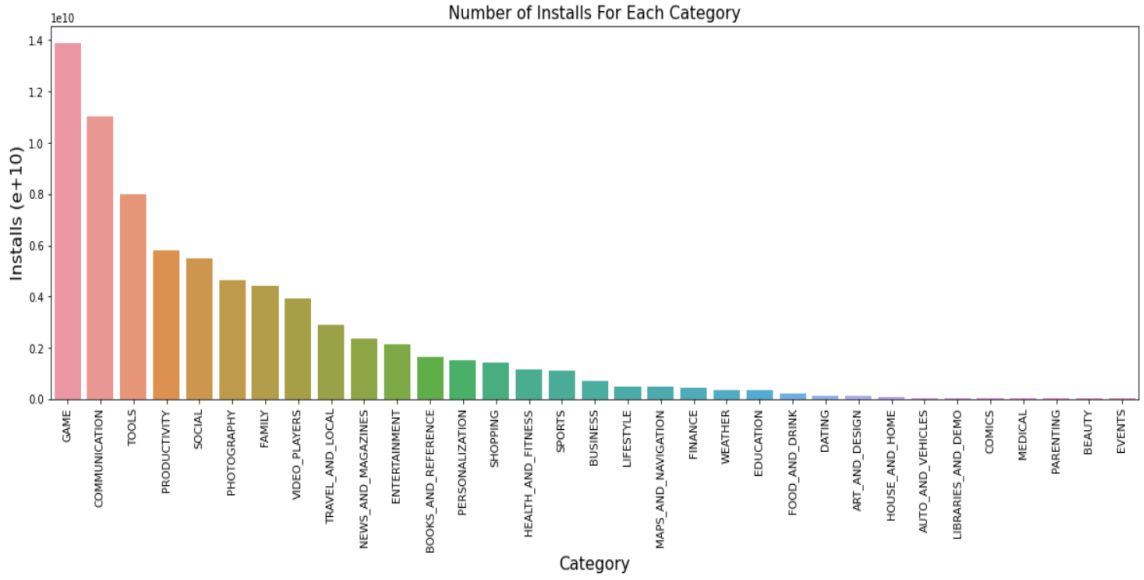
A majority of the apps (82%) in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

Top Category of Play store-



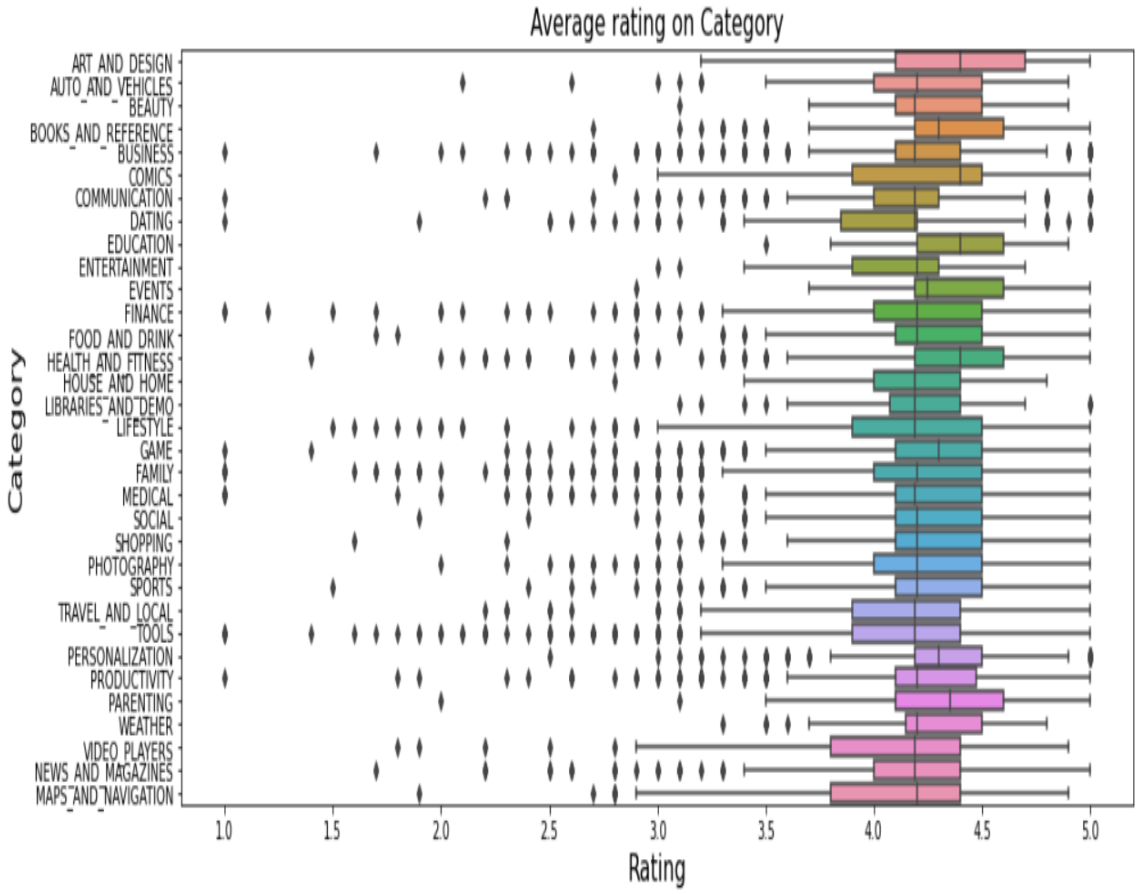
So, there are all total 33 categories in the dataset From the above output we can come to a conclusion that in play store most of the apps are under FAMILY & GAME category and least are of EVENTS & BEAUTY Category

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.

Average ratings on category-



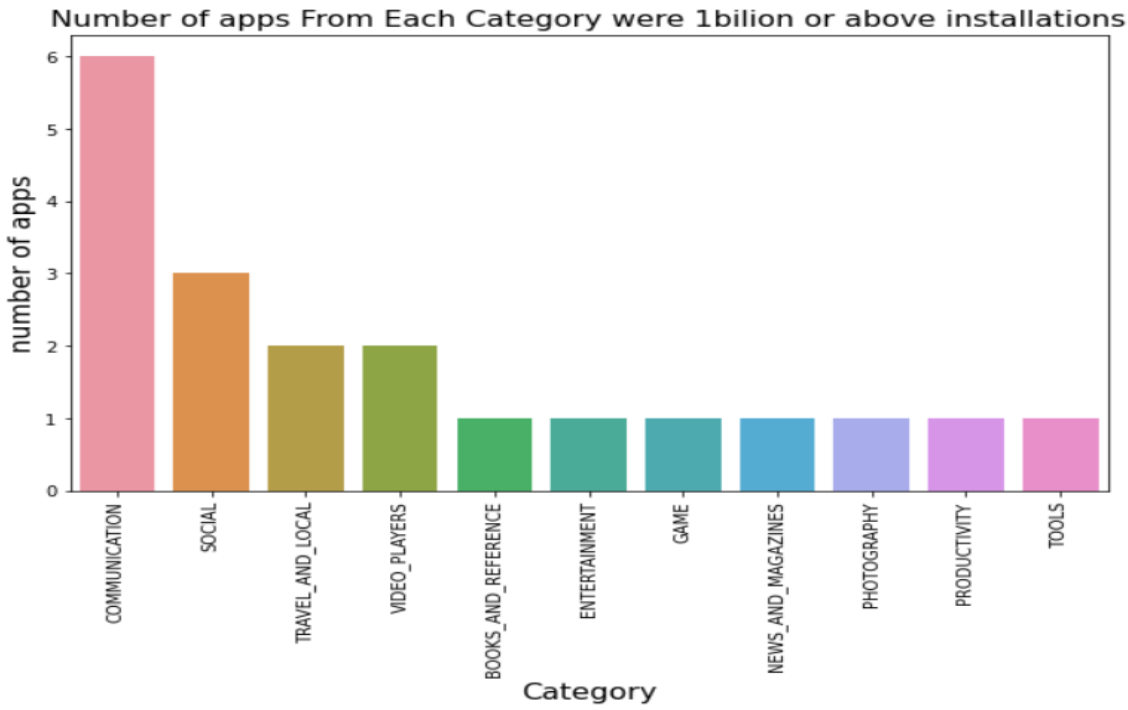
The rating available in the dataset is distributed so we can represent the ratings in a better way if we group the ratings between certain intervals. Here, we can group the rating as follows:

- 4-5: Top rated
- 3-4: Above average
- 2-3: Average
- 1-2: Below average

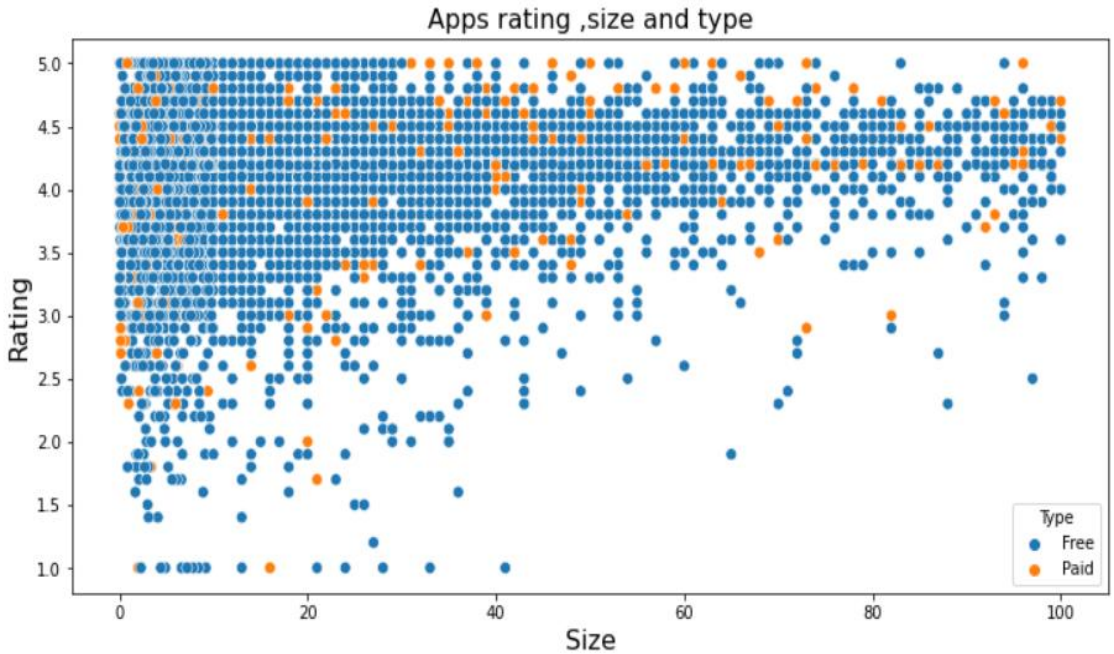
Top paid apps-

	App	Price	Category
4367	I'm Rich - Trump Edition	400.00	LIFESTYLE
9934	I'm Rich/Eu sou Rico/أنا غني/我很有錢	399.99	LIFESTYLE
5359	I am rich(premium)	399.99	FINANCE
5358	I am Rich!	399.99	FINANCE
5373	I AM RICH PRO PLUS	399.99	FINANCE

Categories were apps have above 1bilion installations-

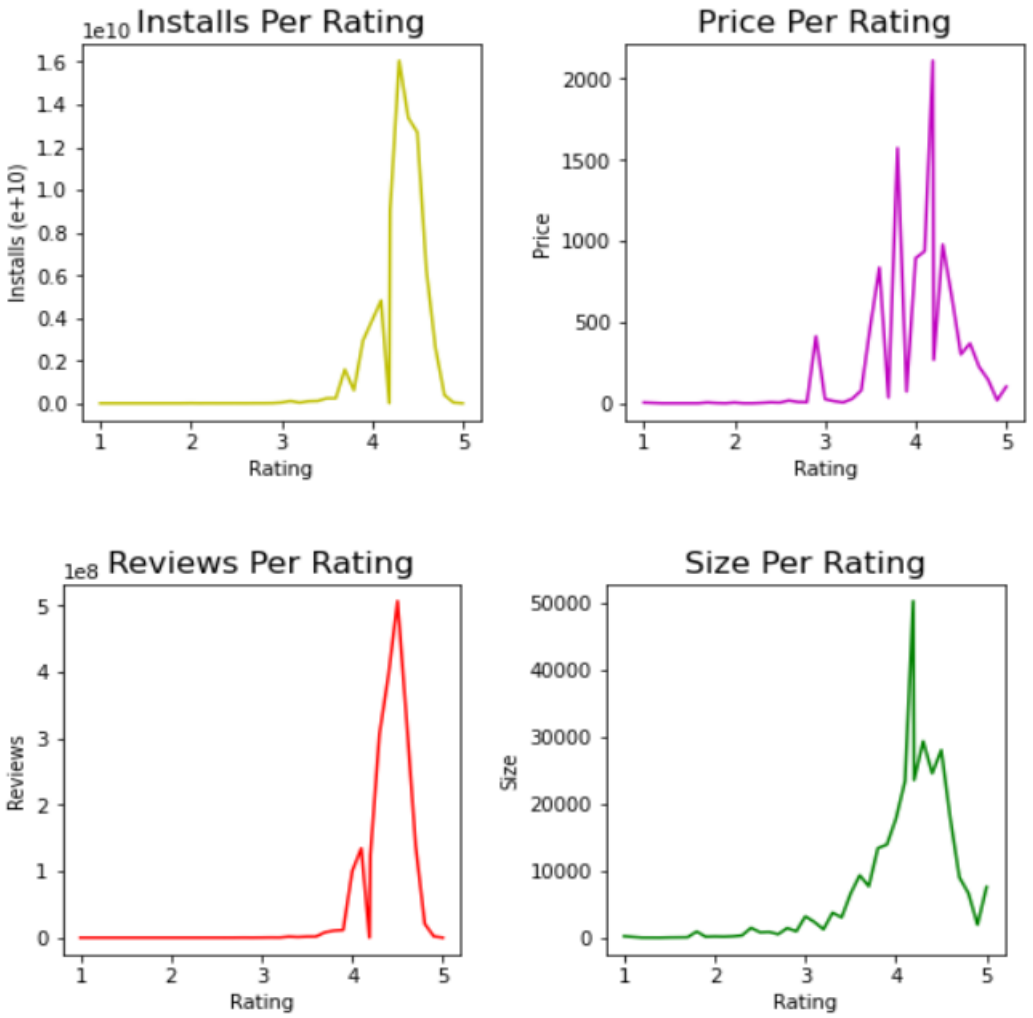


Apps rating distribution on Size and type-



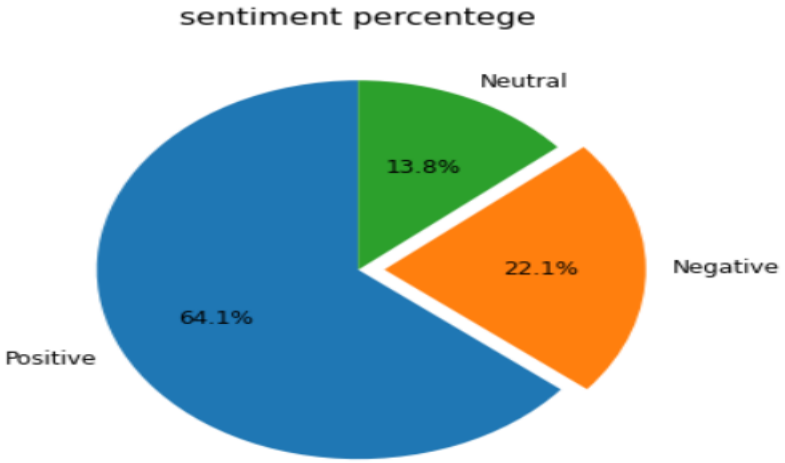
From this scatter plot, we can imply that majority of the free apps are small in size and having high rating. While for paid apps, we have quite equal distribution in term on size and rating.

Distribution of rating along with price, size, review, installs-



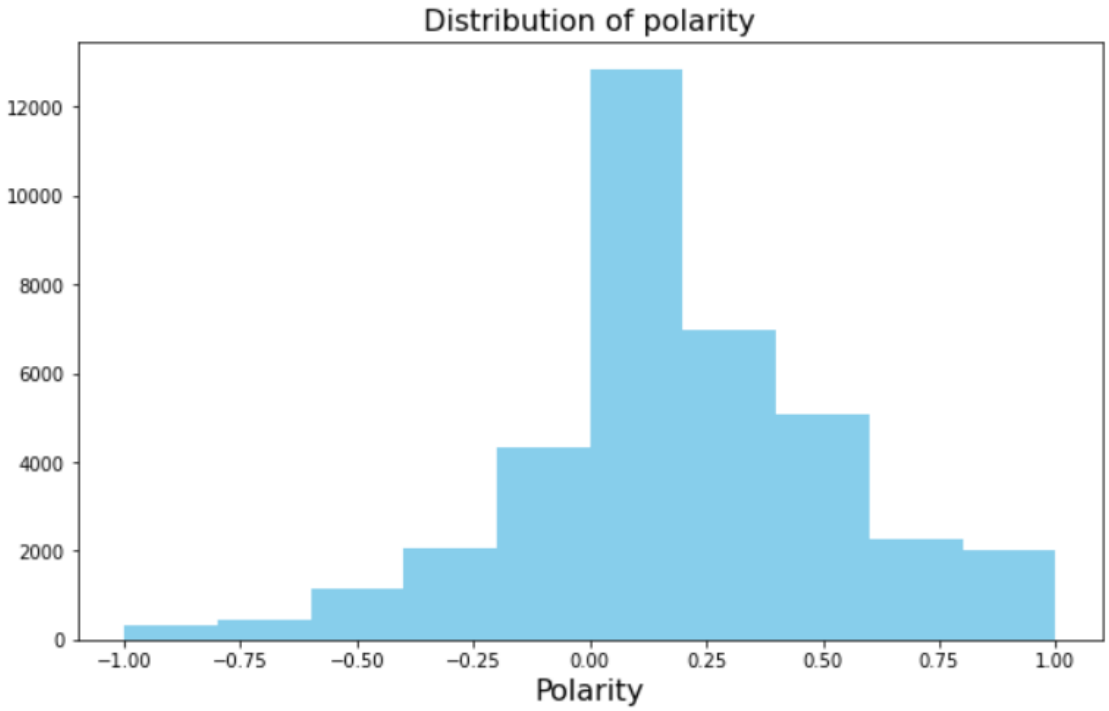
From the above plotting's, we can imply that most of the apps with higher rating range of 4.0 - 4.7 are having high amount of reviews, size, and installs. In terms of price, it doesn't reflect a direct relationship with rating, as we could see a fluctuation in term of pricing even at the range of high rating

Percentage of User review Sentiments-



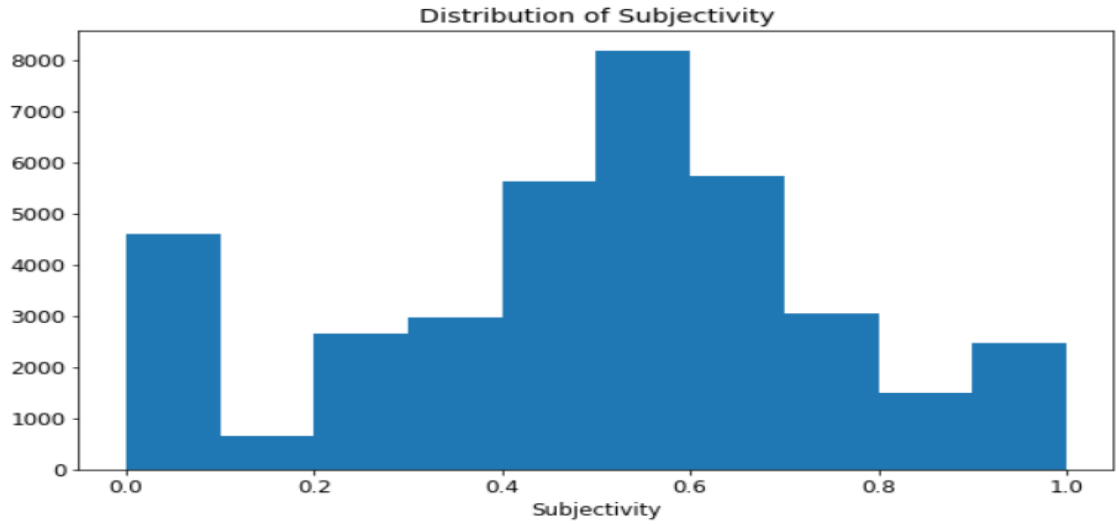
From the above pie chart, we can say that most of the apps that are present on the play store has received positive review by the user while there are some apps which have negative reviews as well.

Distribution of polarity-



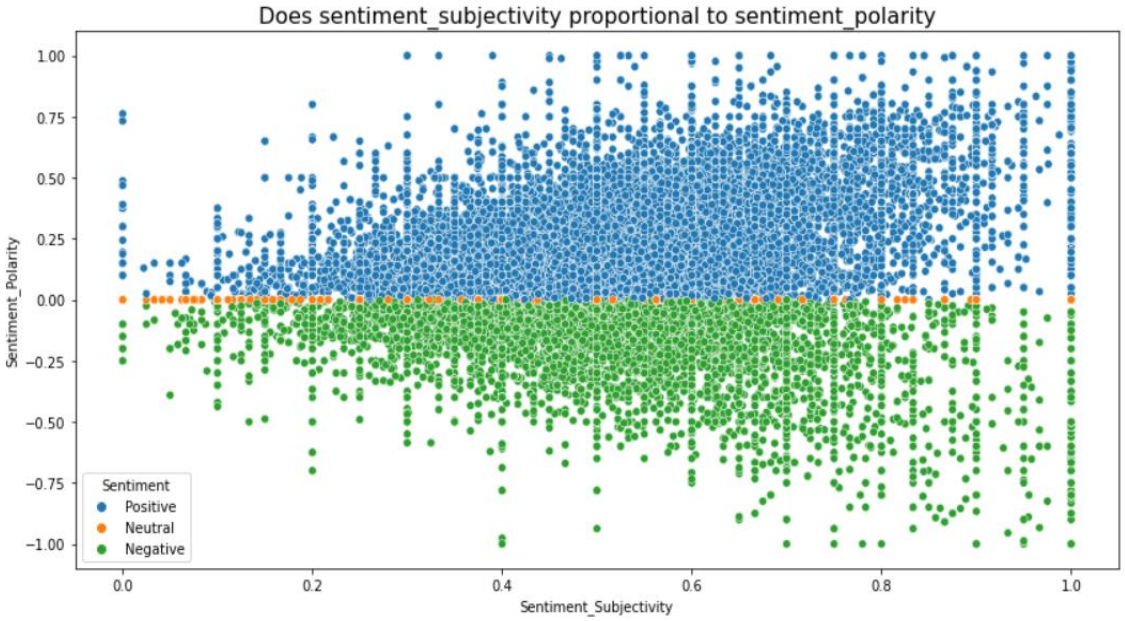
It can be seen that the maximum number of sentiment polarity lies between 0.0 to 0.50. Most people have positive sentiment

Distribution of Subjectivity-



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.

Relationship between sentiment subjectivity proportional to sentiment 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 cases, show a proportional behaviour, when variance is too high or low.

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.

The Google Play Store Apps report provides some useful insights regarding the trending of the apps in the play store. As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like GAME, COMMUNICATION, and TOOL even though the amount of available apps from these categories are twice as much lesser than the category FAMILY. The trending of these apps are most probably due to their nature of being able to entertain or assist the user. Besides, it also shows a good trend where we can see that developers from these categories are focusing on the quality instead of the quantity of the apps.

Other than that, the charts shown above actually implies that most of the apps having good ratings of above 4.0 are mostly confirmed to have high amount of reviews and user installs. There are some spikes in term of size and price but it shouldn't reflect that apps with high rating are mostly big in size and pricy as by looking at the graphs they are most probably are due to

some minority. Furthermore, most of the apps that are having high amount of reviews are from the categories of SOCIAL, COMMUNICATION and GAME like Facebook, WhatsApp Messenger, Instagram, Messenger – Text and Video Chat for Free, Clash of Clans etc.

Eventhough apps from the categories like GAME, SOCIAL, COMMUNICATION and TOOL of having the highest amount of installs, rating and reviews are reflecting the current trend of Android users, they are not even appearing as category in the top 5 most expensive apps in the store (which are mostly from FINANCE and LIFESTYLE). As a conclusion, we learnt that the current trend in the Android market is mostly from these categories which assisting, communicating and entertaining app

- Percentage of free apps = ~92%
- 8783 Apps are having size less than 50 MB. 7749 Apps are having rating more than 4.0 including both types 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
- Mine craft 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 have the highest number average app installs.
- The apps whose size is greater than 90 MB has the highest number of average user reviews, i.e., 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
- 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

THANK YOU

-NAVJOTKHATRI