Capstone Project

Play Store App Review Analysis

Team Members:

Sanjay Jaiswal

Sachin Dubey

Jayalaxmi Mekap

Content :

Abstract :

Google Play Store and formerly Android Market, developed by Google. It serves as the official app store for certified devices running on the Android operating system, It is allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google. Google Play also serves as a digital media store, offering music, books, movies, and television programs.

Google Play featured more than 3.5 million Android applications. Android Market was announced by Google on August 28, 2008, and was made available to users on October 22.

Problem Statement:

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

Introduction of Dataset :

There are two datasets

1. **Play Store Data** : This dataset has 10841 observations in it with 13 columns and it is a mix between categorical and numeric values.

2. **User Reviews** : This dataset has 64294 observations in it with 5 columns and it is a mix between categorical and numeric values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

Steps took to Analyse play store App review :

- Importing the necessary Libraries
- Mounting Google Drive and Creating a file path
- Importing Dataset From Drive
- Printing the information about a DataFrame including the index dtype and columns, non-null values and memory usage.
- We are going to use Pandas describe() view some basic statistical details like percentile, mean, std etc.
- Checking the sum of null values present in our dataset
- Data cleaning and handling null values
- Taking necessary columns only
- Merging two dataframes
- Data visualization

Analysis:

Play Store App Analysis:

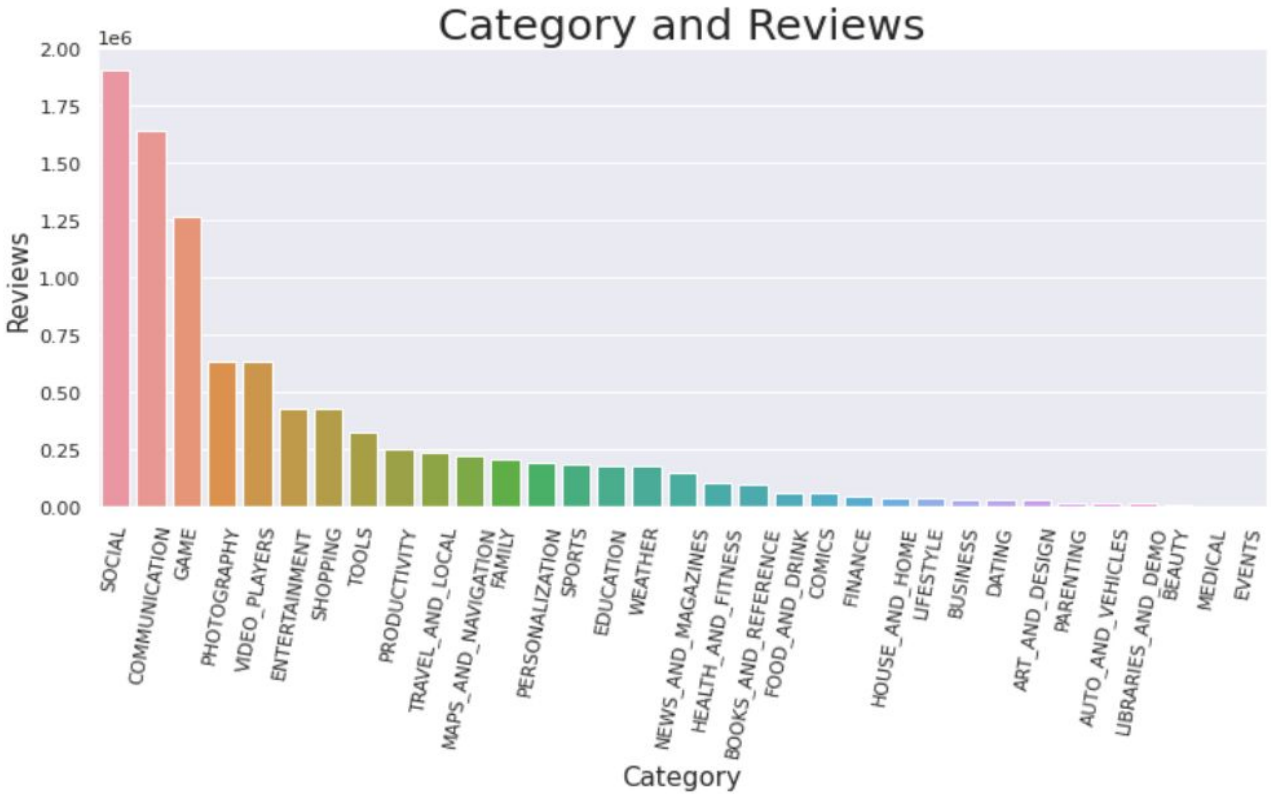
- Average reviews across each category
- Top app category in play store(Count of apps in each category)
- App installed according to category(Number of installed application for each category)
- Top app genres in play store(Count of apps in each genres)
- App installed according to genres(Number of installed applications for each genres)
- Percentage of free vs paid apps in play store
- Content rating
- Analyse the distributions of app rating, app size and app price
- Top earning app in play store
- Average installation of app across the year
- What are the top five installed apps in any category
- How many apps are present in each category according to their version(Free/paid)?
- Rating Vs Type

App Review Analysis:

- Distribution of Sentiment Subjectivity
- Percentage of Review Sentiments
- Distribution of Sentiment Polarity
- Polarity Vs Subjectivity

Average reviews across each category

- In the given plot we can say that maximum average reviews high for social,communication and games category apps.
- And minimum average reviews for beauty,medical and events category apps.

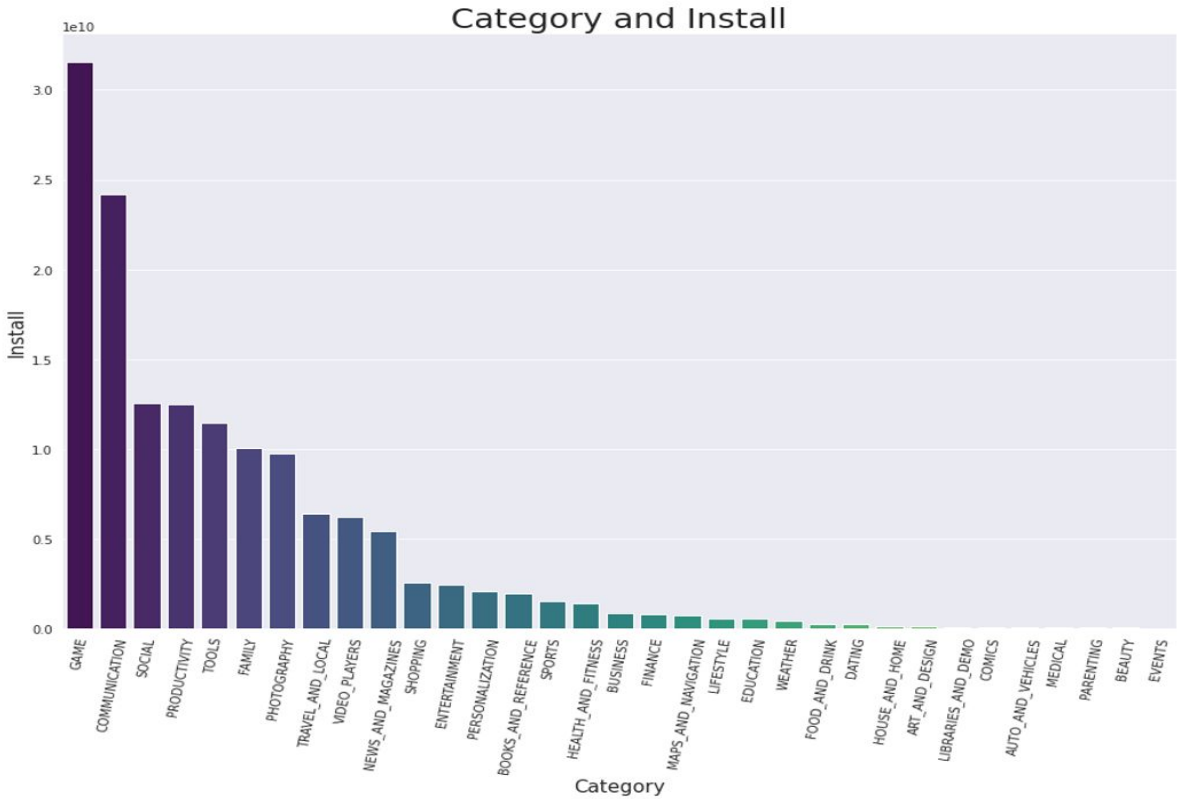


Top app category in play store(Count of apps in each category)

- As we can see the plot maximum no. of apps present in the play store are comes under Family, Games and Tools Category.
- And minimum no. of apps are present in Comics,Parenting,Beauty category.

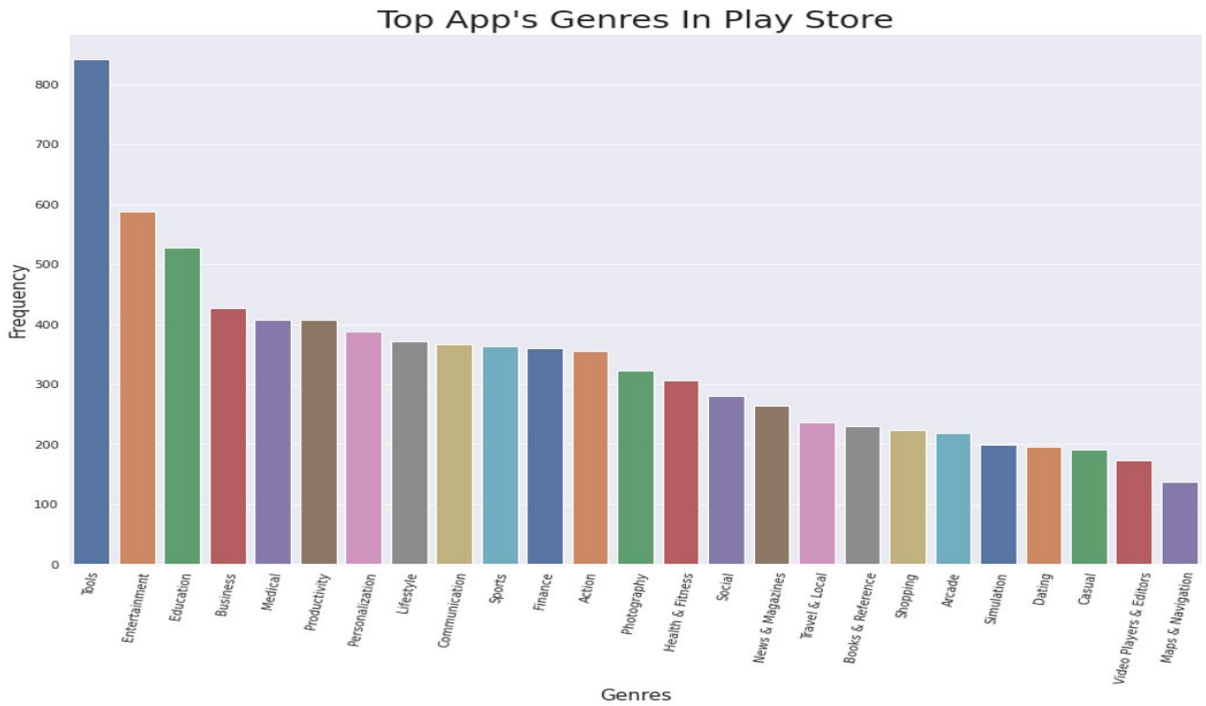
App installed according to category(Number of Installed applications for each category)

- As we can see from the above plot: Maximum number of apps present in google play store comes under Family, Games and Tools Category.
- But as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps comes under Games,Communication and Social Category.



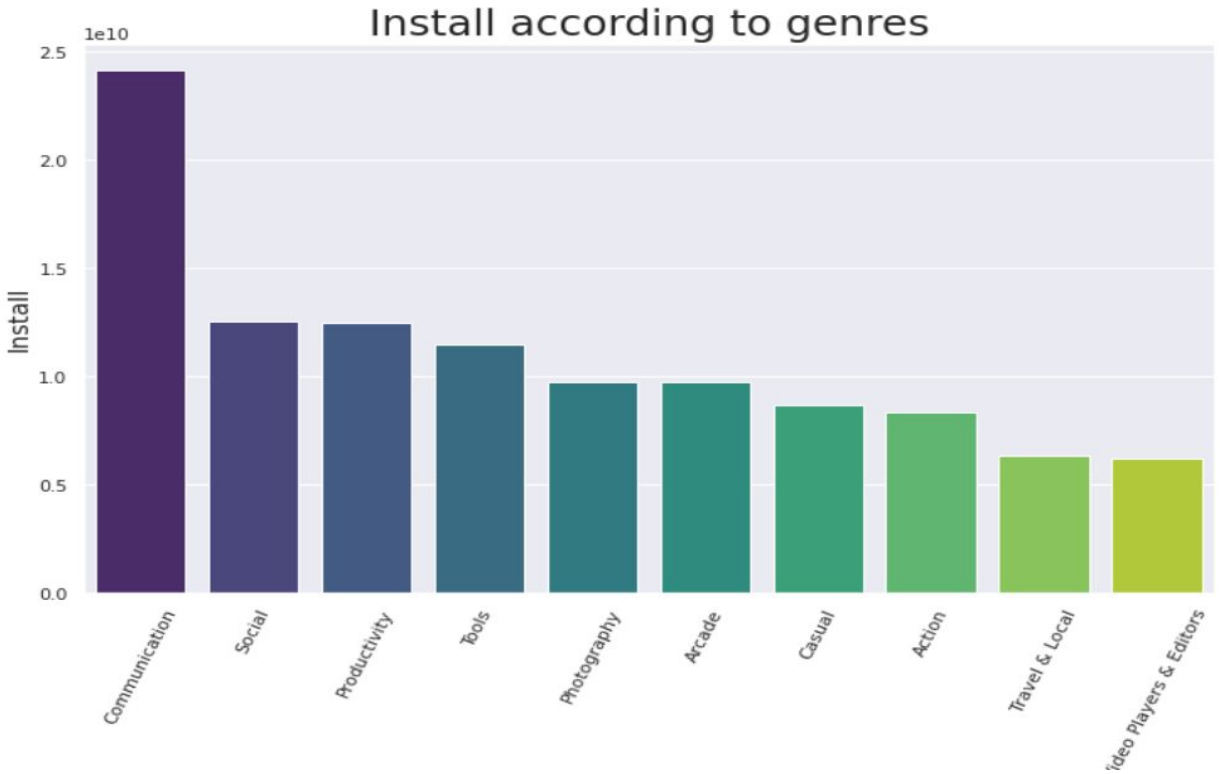
Top app genres in play store(Count of apps in each genres)

- As we can see the plot maximum no. of apps present in the play store are comes under Tools, Entertainment and Education Genres..



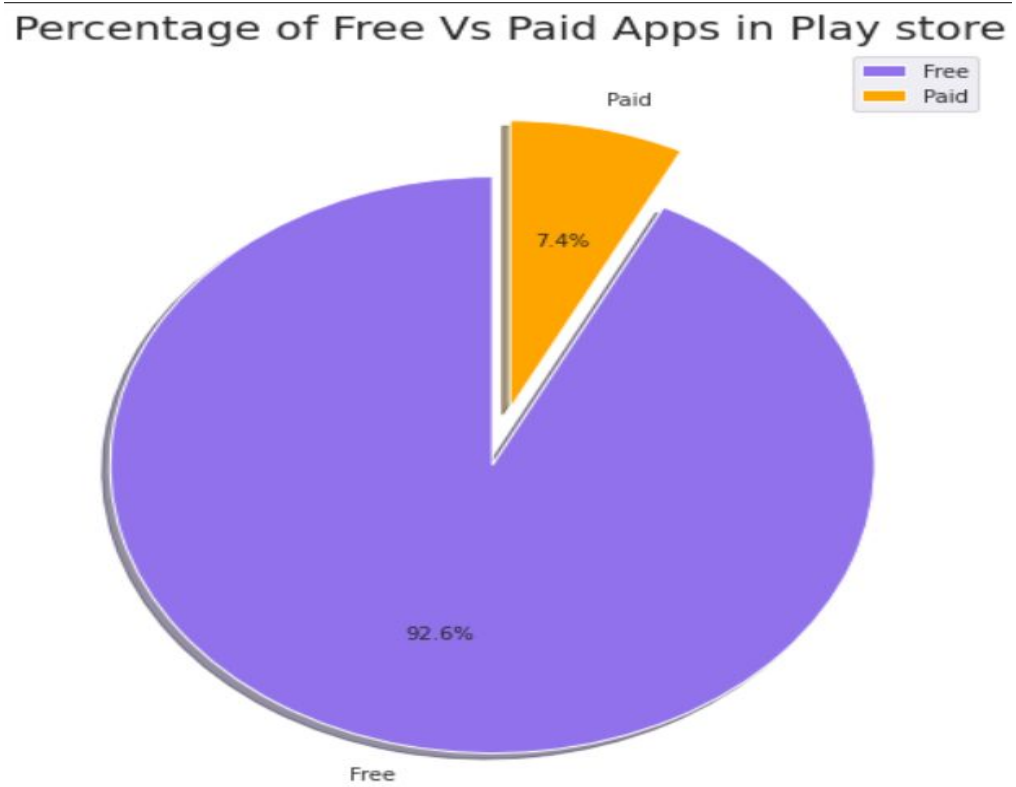
App installed according to genres(Number of installed applications for each genres)

- As we can see the above bar chart maximum number of apps present in google play store comes under Tools, Entertainment and Education Genres.
- But as per the installation and requirement in the market plot, scenario is not the same maximum installed apps comes under Communication,Social and Productivity Genres.



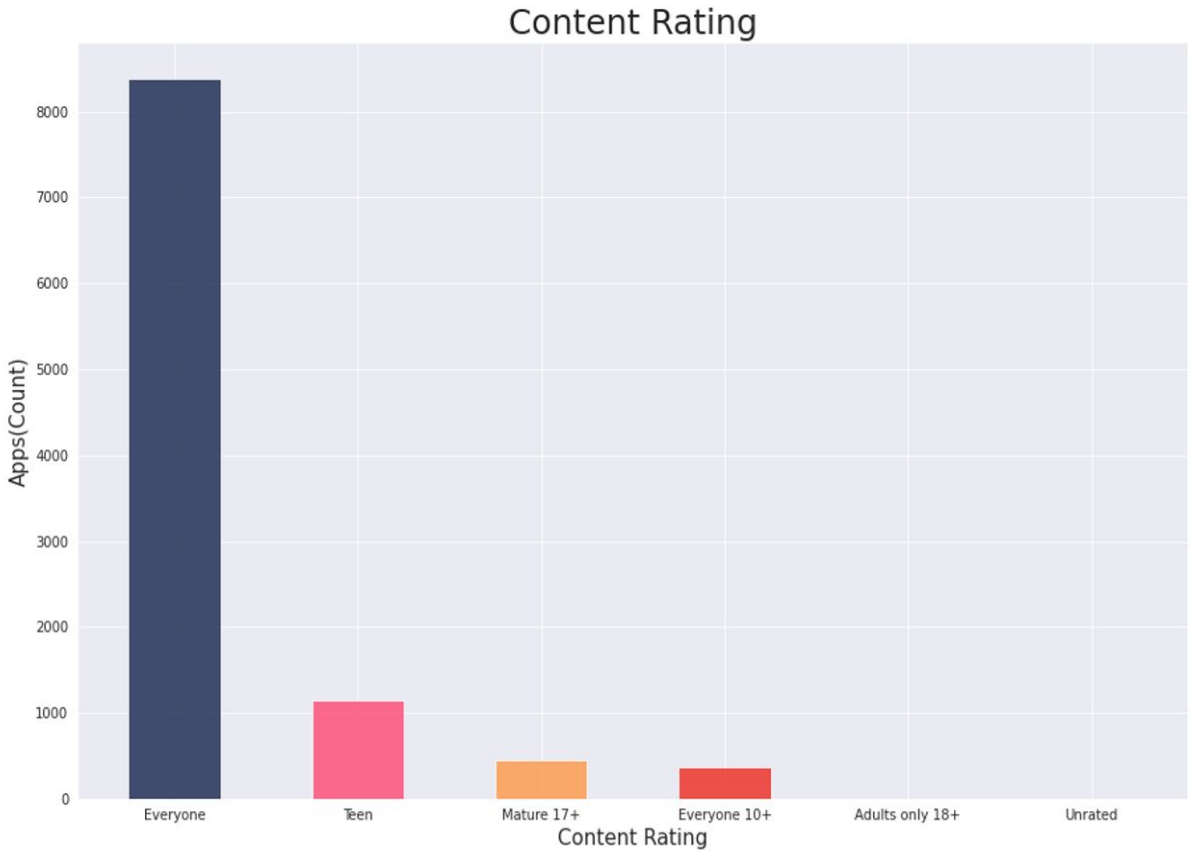
Percentage of free vs paid apps in play store

- As we can see pie chart maximum no. of apps are free version in the given dataset.



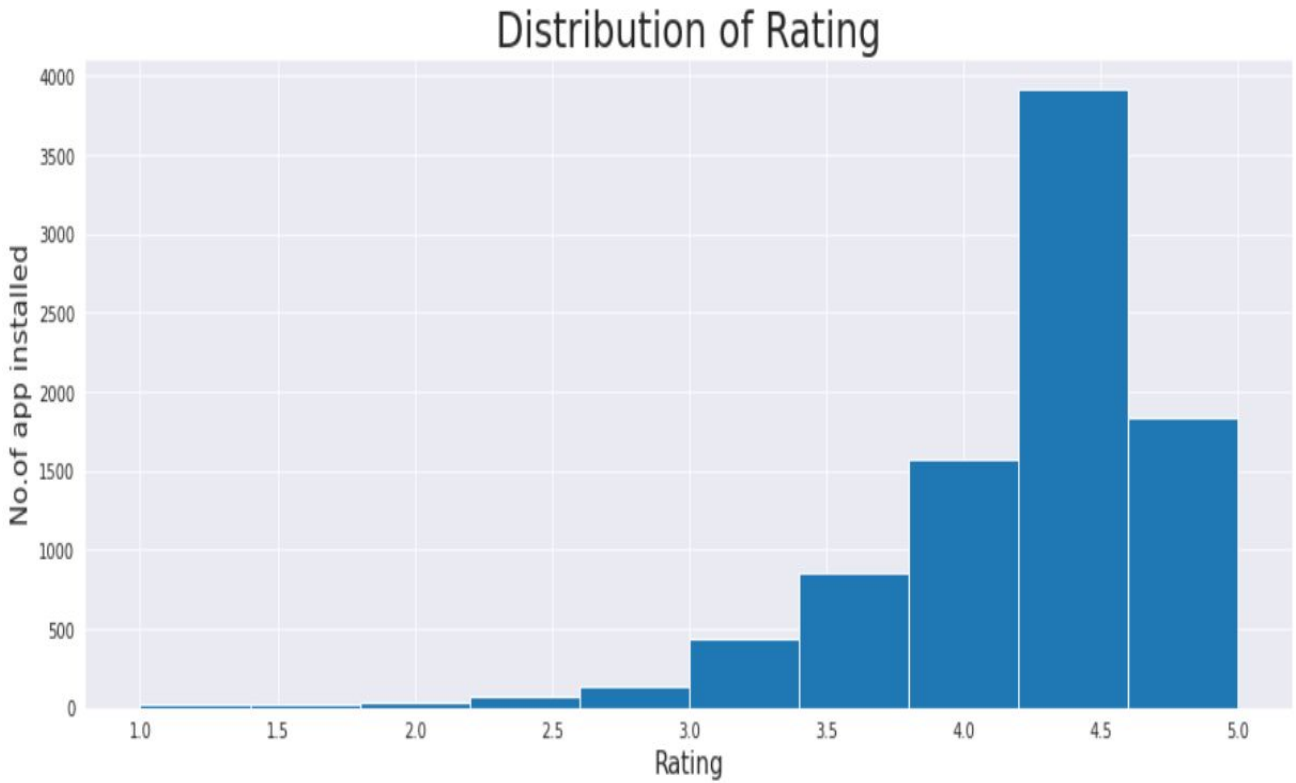
Content Rating

- As we can see the plot maximum no. of apps are in google play store comes under Everyone,Teen,Mature 17+ content rating.
- There is very few apps comes under Adults only 18+,Unrated content rating.



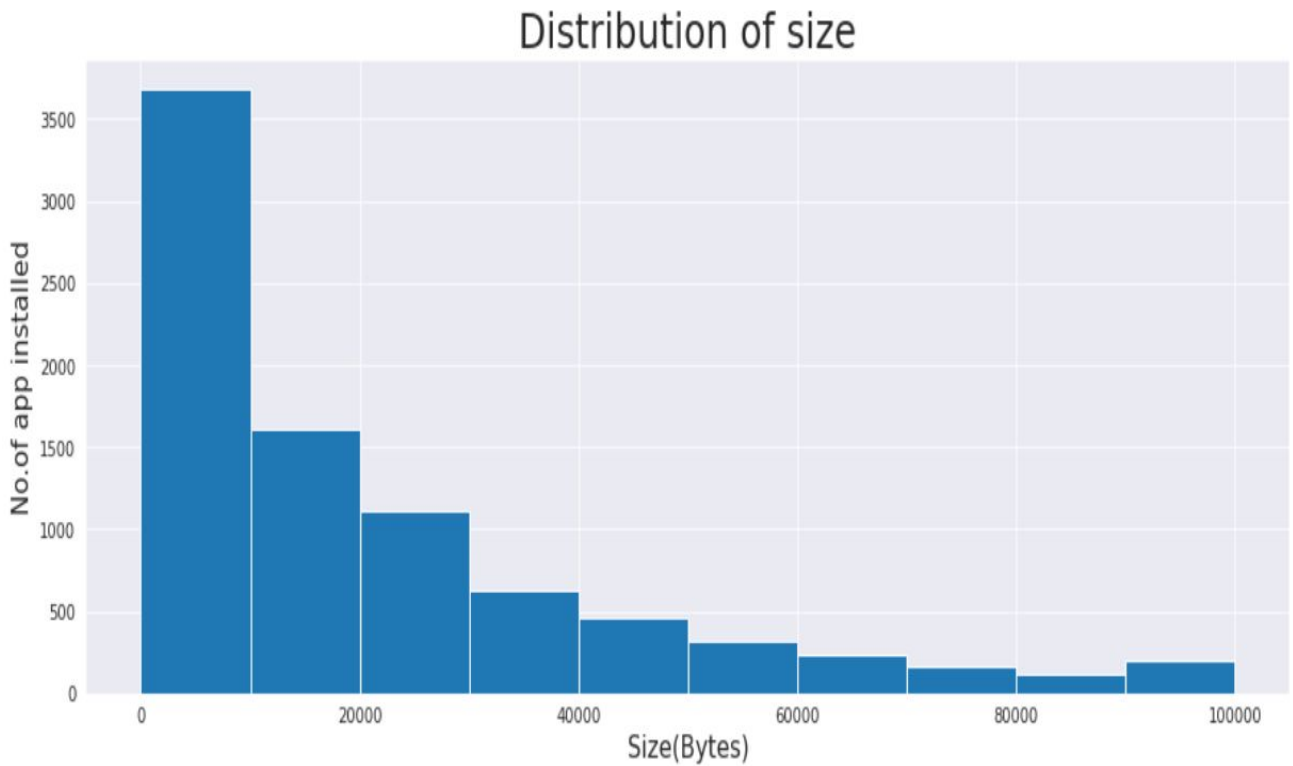
Distribution of App Rating

- As we can see the histogram plot maximum no. of apps installed which has rating lies between 4.0 to 4.5
- Very few apps are installed which has rating below 3.0



Distribution of App Size

- As we see the histogram plot we can say that maximum no. of apps installed which has size below **40,000 kilobyte(40mb)**
- So, most of users preferred less mb size apps.



Distribution of App Price

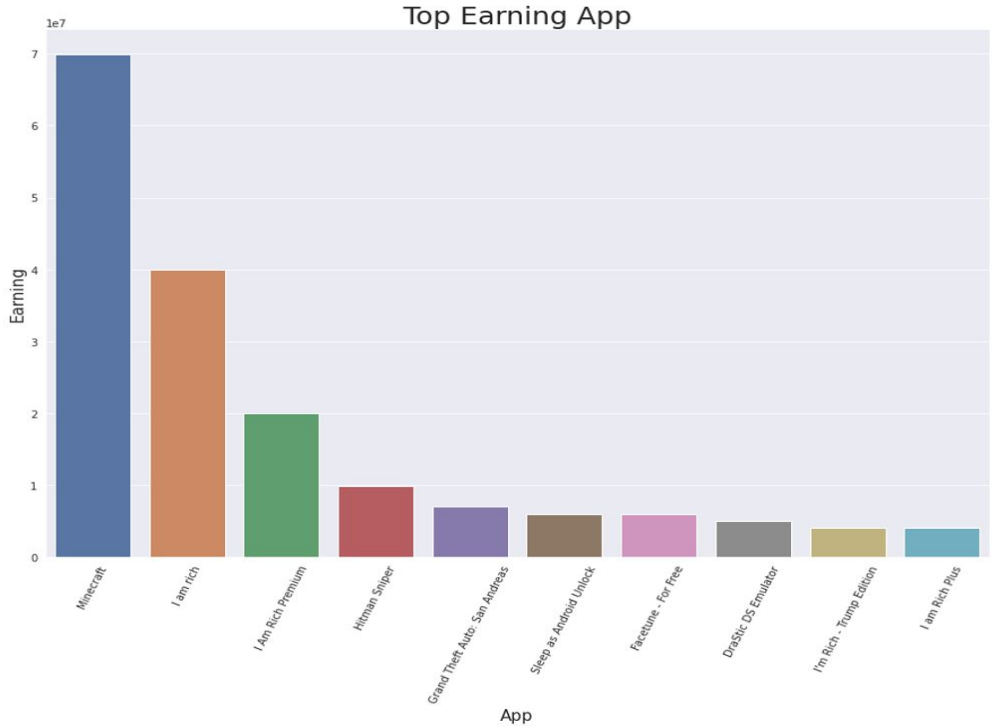
- In the given dataset maximum apps price is $0 and there is very few apps that has price is $399.99 and $400

```
Price
0.00      9593
0.99       146
1.00         3
1.04         1
1.20         1
          ...
379.99       1
389.99       1
394.99       1
399.99      12
400.00       1
Name: Price, Length: 92, dtype: int64
```

Top Earning App in Play Store

```
[ ] top_earning_app
```

	App	Installs	Price	Earning
4347	Minecraft	10000000.0	6.99	69900000.0
5351	I am rich	100000.0	399.99	39999000.0
5356	I Am Rich Premium	50000.0	399.99	19999500.0
4034	Hitman Sniper	10000000.0	0.99	9900000.0
7417	Grand Theft Auto: San Andreas	1000000.0	6.99	6990000.0
5578	Sleep as Android Unlock	1000000.0	5.99	5990000.0
2883	Facetune - For Free	1000000.0	5.99	5990000.0
8804	DraStic DS Emulator	1000000.0	4.99	4990000.0
4367	I'm Rich - Trump Edition	10000.0	400.00	4000000.0
5354	I am Rich Plus	10000.0	399.99	3999900.0



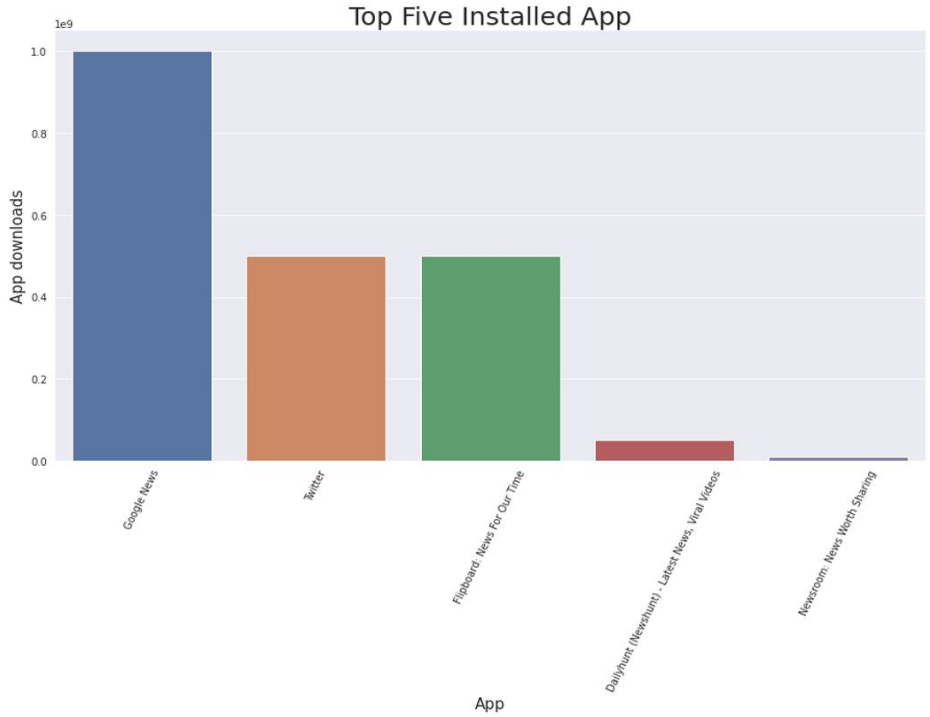
Average installation of app across the year

- As we see the bar plot we can say that maximum app downloaded in year 2018, 2017 and 2016.
- Very few app downloaded in year 2010,2012.

Top five installed apps in any category

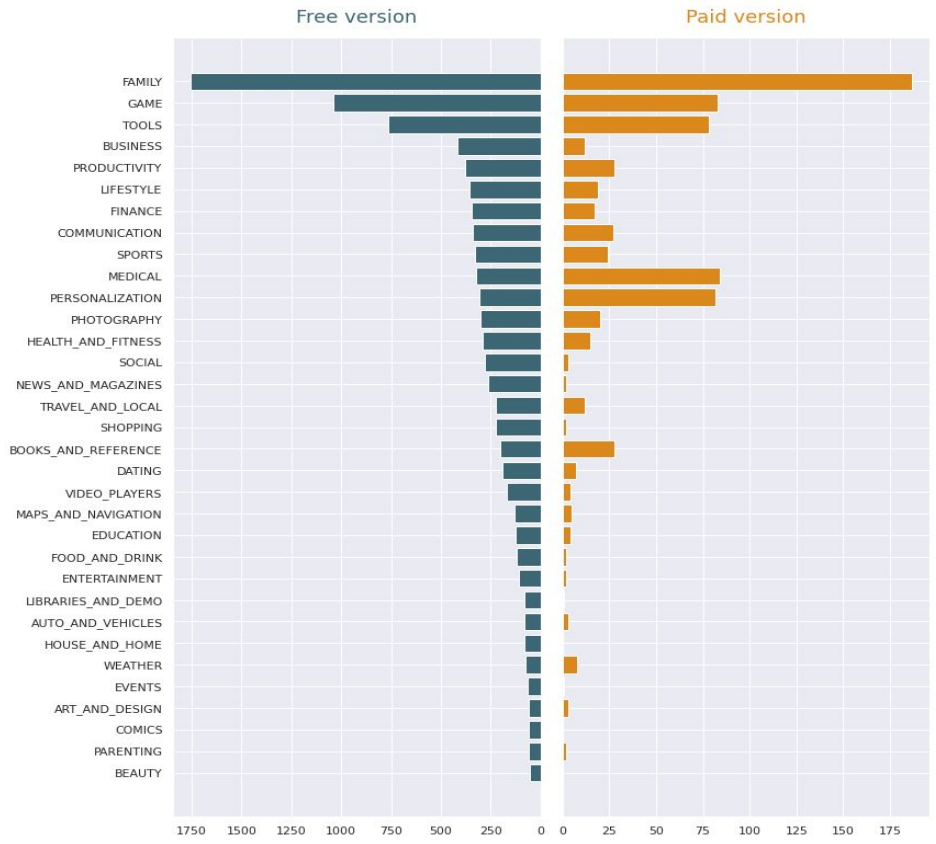
- Here i took NEWS_AND_MAGAZINES category As we can see the plot maximum number of **Google News** app installed.

App	Installs
Google News	1.000000e+09
Twitter	5.000000e+08
Flipboard: News For Our Time	5.000000e+08
Dailyhunt (Newshunt) - Latest News, Viral Videos	5.000000e+07
Newsroom: News Worth Sharing	1.000000e+07



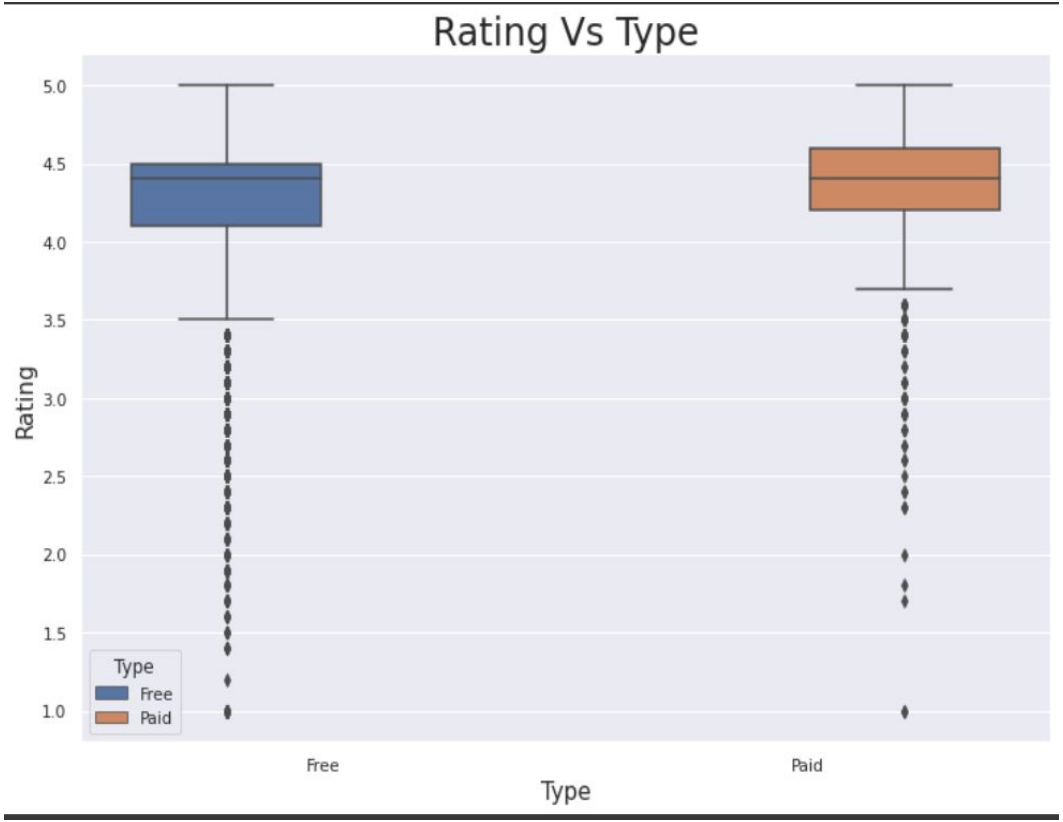
How many apps are present in each category according to their version(Free/paid)?

- As we see the bidirectional bar chart we can say maximum free version and paid version apps are present in family,game,tools category.
- As we move further in chart we also observe in medical , books_and_reference and personalization category paid version apps are more than free version apps



Rating Vs Type

- As we see the box plot we say maximum no. of rating more in paid version apps as compared to free version apps.

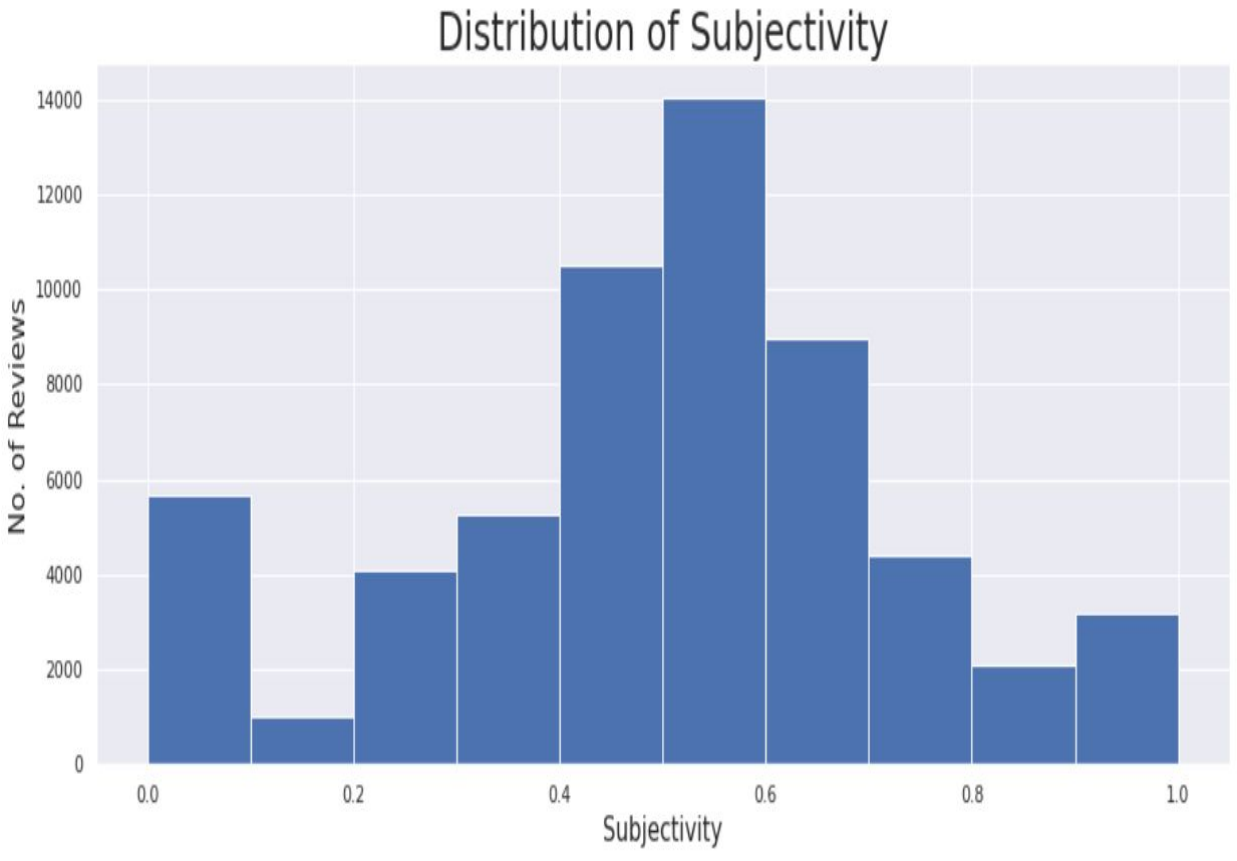


Distribution of Sentiment Subjectivity

- **Sentiment Subjectivity** generally refer to personal opinion, emotion or judgment, which lies in the range of [0,1].

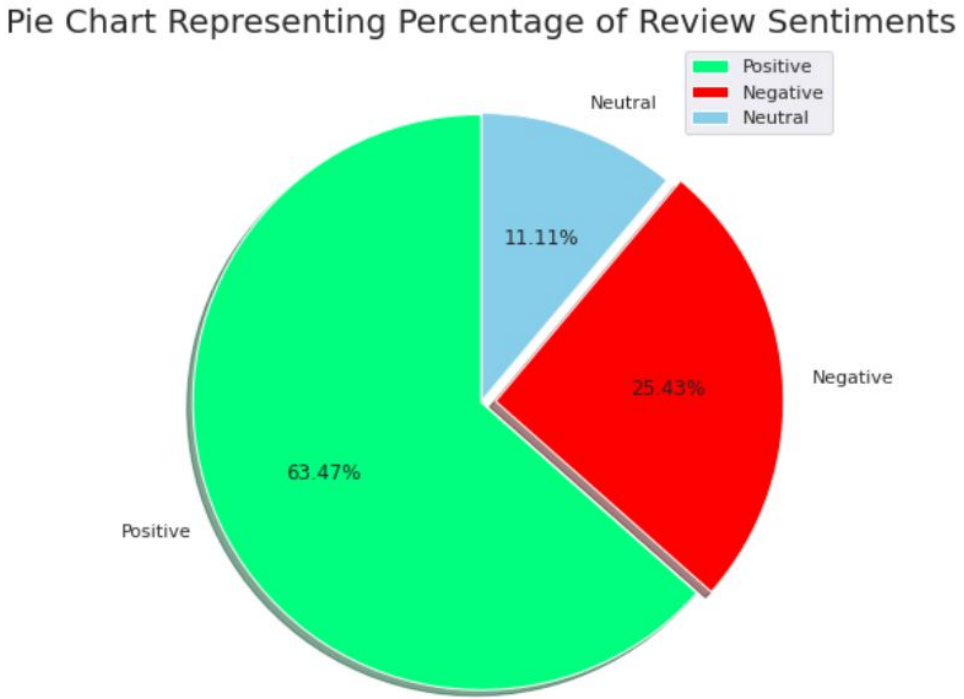
- As we see the histogram plot It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7.

- From this we can conclude that maximum number of users give reviews to the applications, according to their experience.



Percentage of Review Sentiments

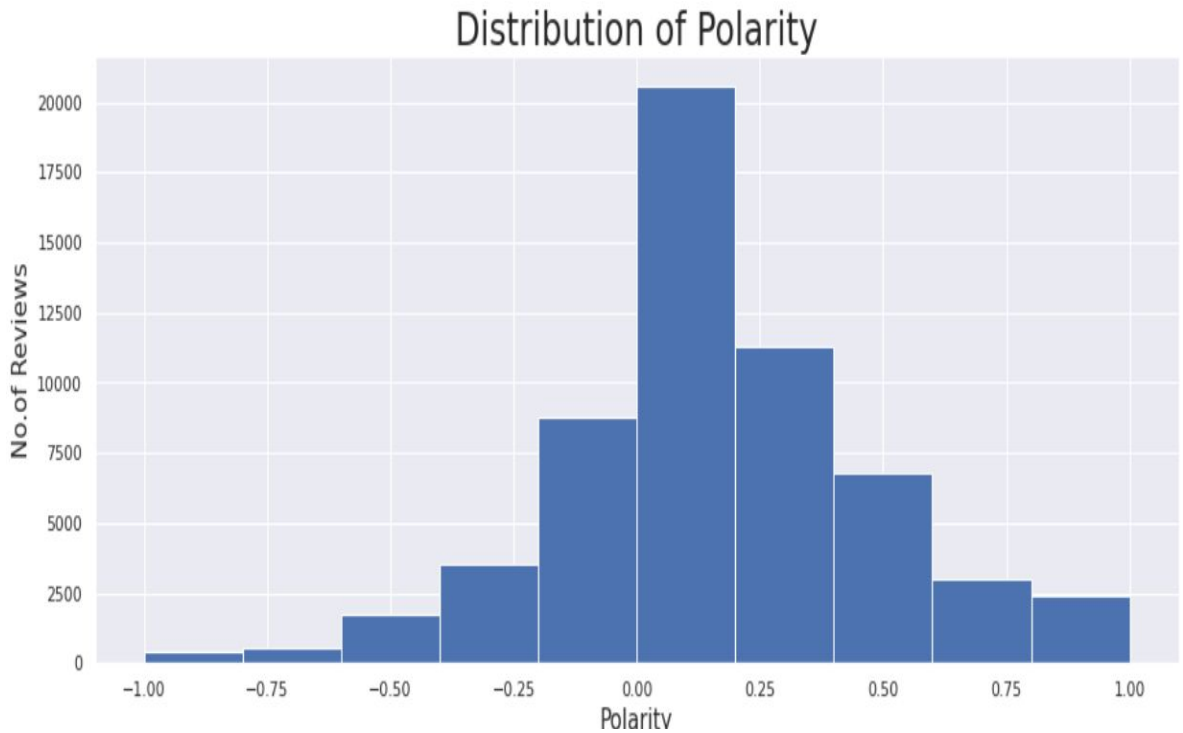
	Response	Sentiment
0	Positive	37554
1	Negative	15045
2	Neutral	6572



Distribution of Sentiment Polarity

- **Sentiment Polarity** is a float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.

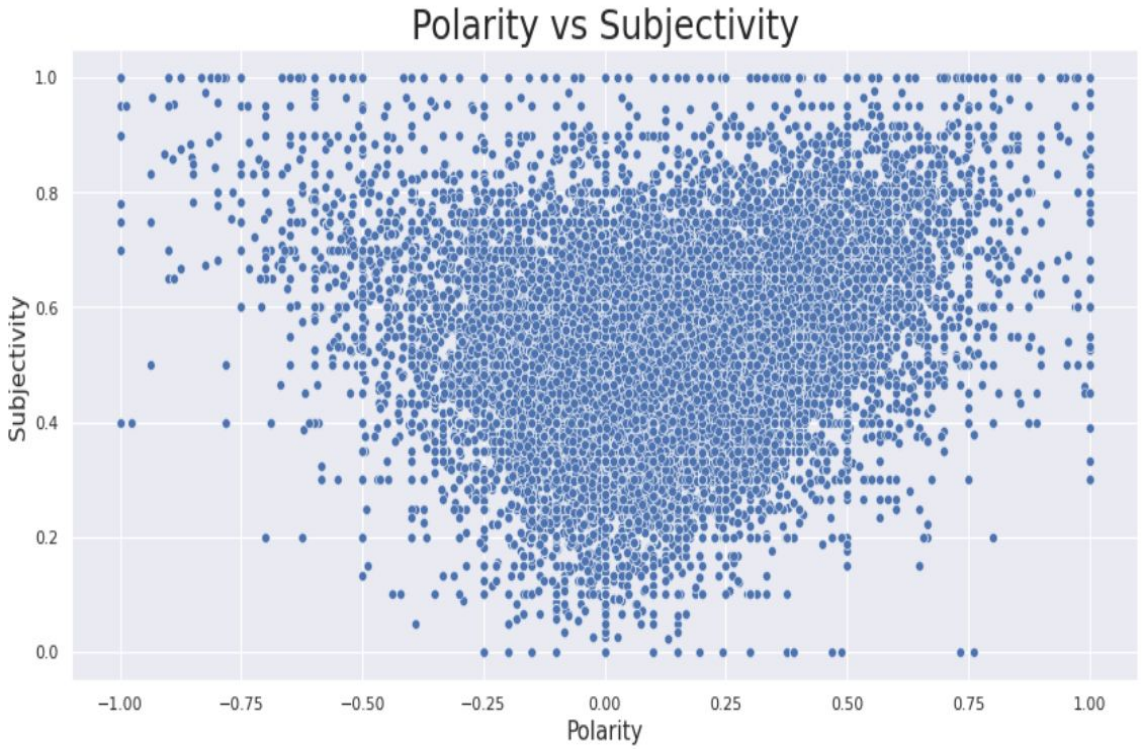
- In this plot we can say that maximum no. of positive reviews lies between 0 to 0.5.



Polarity Vs Subjectivity

- **Sentiment Subjectivity** generally refer to personal opinion, emotion or judgment, which lies in the range of [0,1].

- From the given scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in some number of case Sentiment Subjectivity proportional to Sentiment Polarity.



Correlation

- Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.
- In simple terms, it tells us how much does one variable changes for a slight change in another variable. It may take positive, negative and zero values depending on the direction of the change.

Conclusion:

- We calculated the average reviews across each category and we also calculated top category and top genres of apps in the given dataset.
- Further we also calculated apps installed according to their category and genres. We observe the maximum number of apps present in google play store comes under Tools, Entertainment and Education Genres but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps come under Communication,Productivity and Social Genres.
- We also observe that the percentage of free apps is more than 90% in the given dataset.
- We also draw the pie chart of review sentiments and observe that the percentage of positive sentiments is near about 64%.
- Histogram of sentiment subjectivity and observe the maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that the maximum number of users give reviews to the applications, according to their experience.


THANK
YOU