Capstone Project – 1

Project Title : Play store app review analysis

By

Pawar Sapana

Points for Discussion

- Introduction
- Data Summary
- Steps Involved in EDA
- Factors affecting on App Rating
- App Size
- Factors affecting number of Installs
- Most Installed Apps
- Type (Free/Paid)
- Most Costly apps
- Correlation heatmap
- Sentiment Analysis
- Sentiment Count
- Sentiment Polarity vs Sentiment Subjectivity
- Conclusion

Introduction

- Exploratory Data Analysis (EDA)

In this project we have done exploratory data analysis on google play store data to get some meaningful insights.

The analysis of google play application aided to build most reliable and more interactive applications.
This would be very useful for app developer to build an application focussed on certain discussed category in this analysis.
This analysis will definitely help in building the application with precise and accurate objective.

Data Summary

We had two different datasets

1. **Play Store Apps Data**

   In given Play store dataset there are total 13 columns and 10841 rows.

It contains all the information of different types of applications like name of app, category of app, ratings from the user, size, installs, type of app(free/paid), price of app, content, genres, last updated date, current version of app available on play store and minimum required android version.

2. **User Reviews Data**

   In this dataset there are 5 columns and 64295 rows.

It contains information regarding the reviews given by users, like review, sentiment of that review, sentiment polarity and sentiment subjectivity of that particular reviews.

Steps involved in EDA

- Data exploration
- Null values treatment
- Data imputation and manipulation
- Outlier treatment
- Data visualization
- Trends and correlations
- Final summary of conclusion

Questions

We will analyse the Play Store data by answering following questions:

1.How distribution of Rating and size look like?

2.What is the average rating per category?

3.Which category has high number of installs? Get 5 most installed apps with corresponding number of installs.

4.Which category have most paid type of apps? Find out the top 5 apps having highest price.

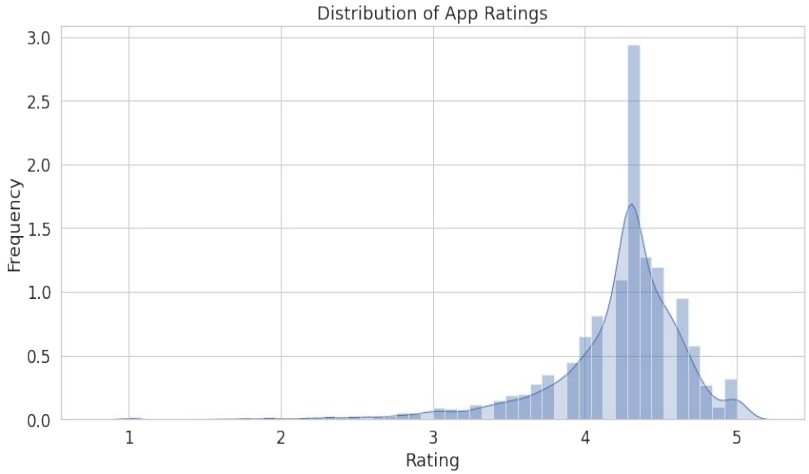
5.What is the percentage of paid and free apps?

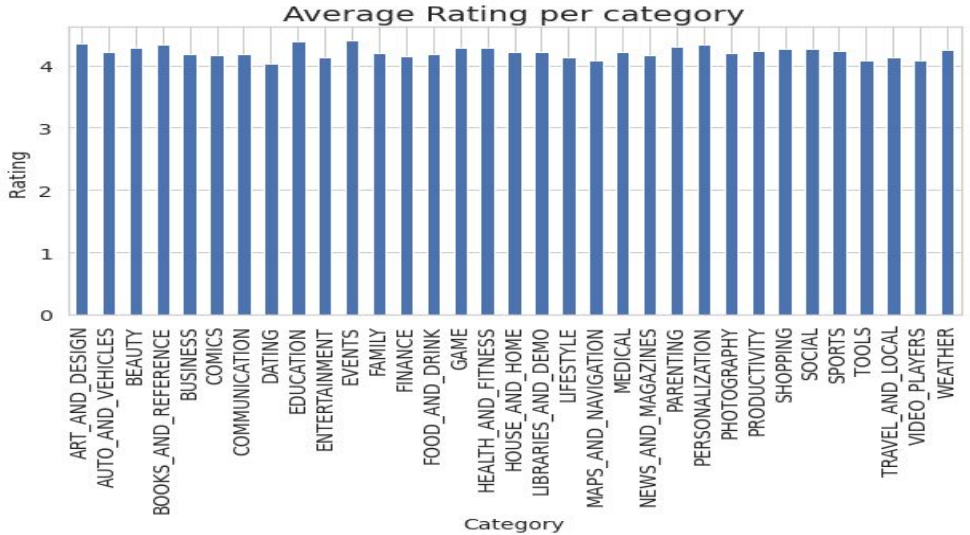
6.What are the installs per content rating of apps?

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

App Ratings

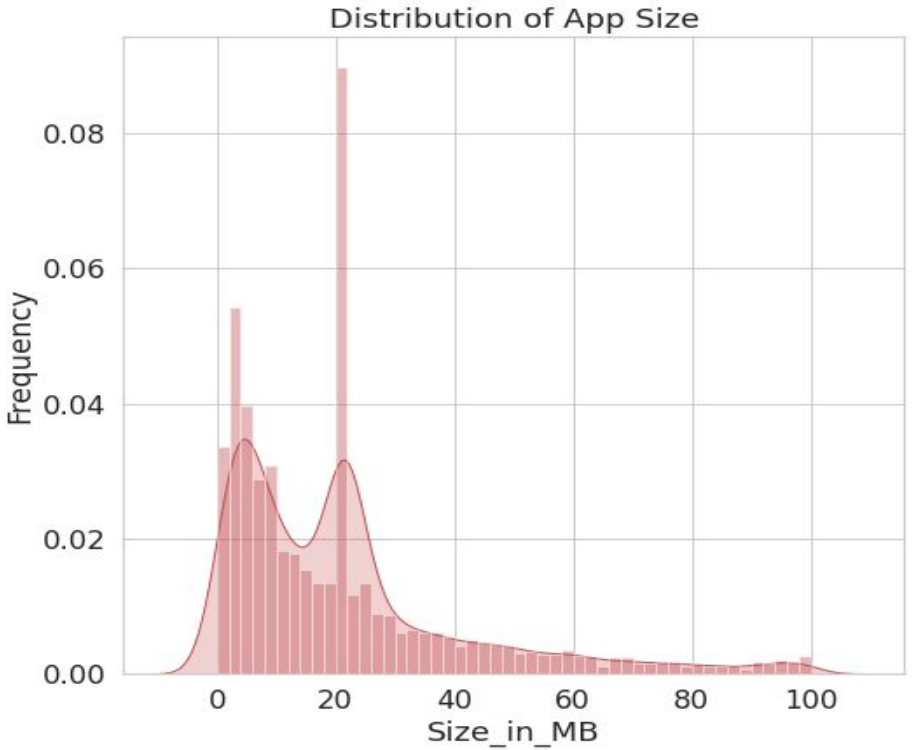
- Most of the ratings of apps given by users are in between 3 to 5.
- Most number of apps are rated at 4.3
- All categories of apps have more than 4 average rating.
- Event category has highest average rating with average rating of 4.395313





App Size

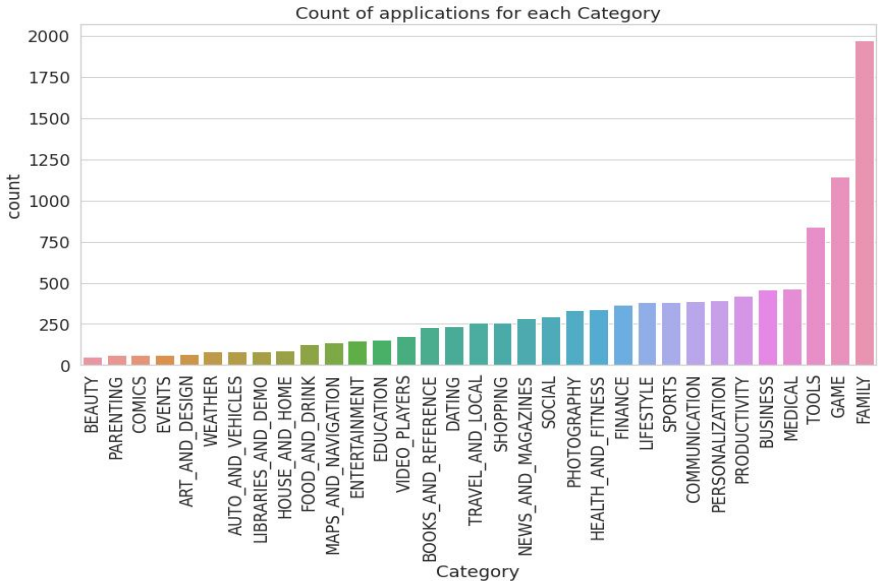
- Maximum number of applications present in the dataset are of small size.

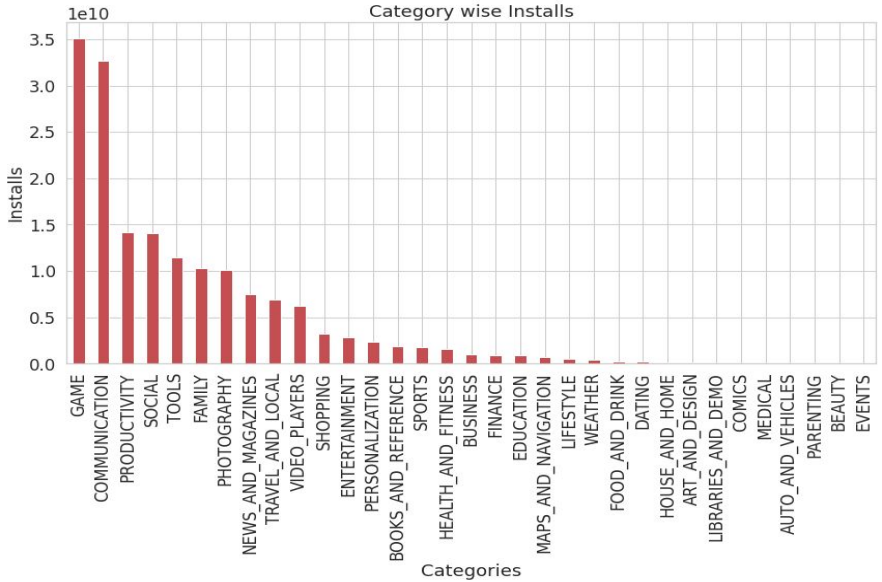


Installs

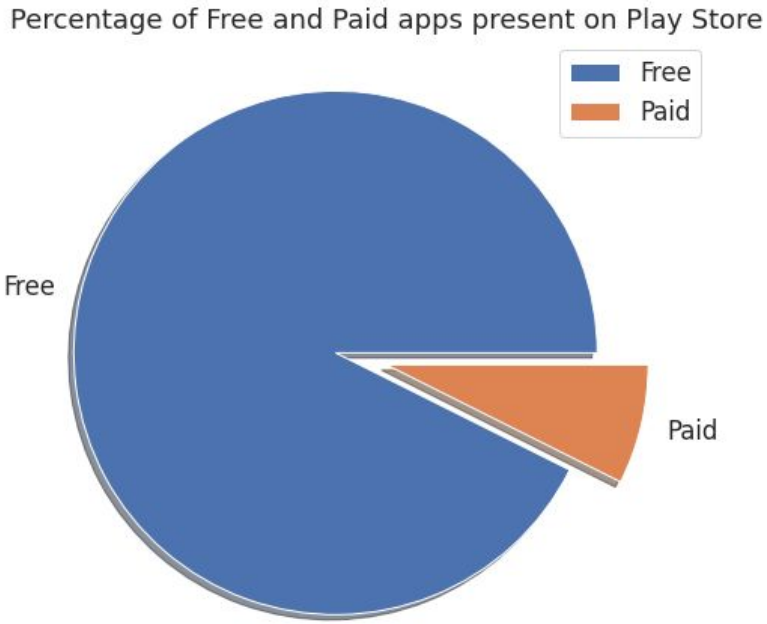
Category vs Installs

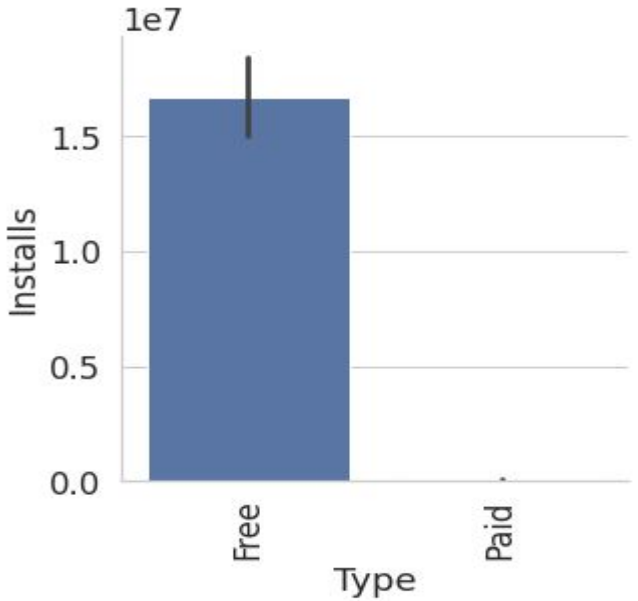
Maximum number of apps present in google play store come under Family, Game and tools but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps comes under Game, Communication, Productivity and Social.





❖ Type (Free/Paid) vs Installs





- About 92% apps are free and 8% apps are of paid type.
- Free apps are installed more than paid apps.

❖ Content Rating vs Installs

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

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

❖ Reviews vs Installs

- Number of installs is positively correlated with reviews with correlation 0.64

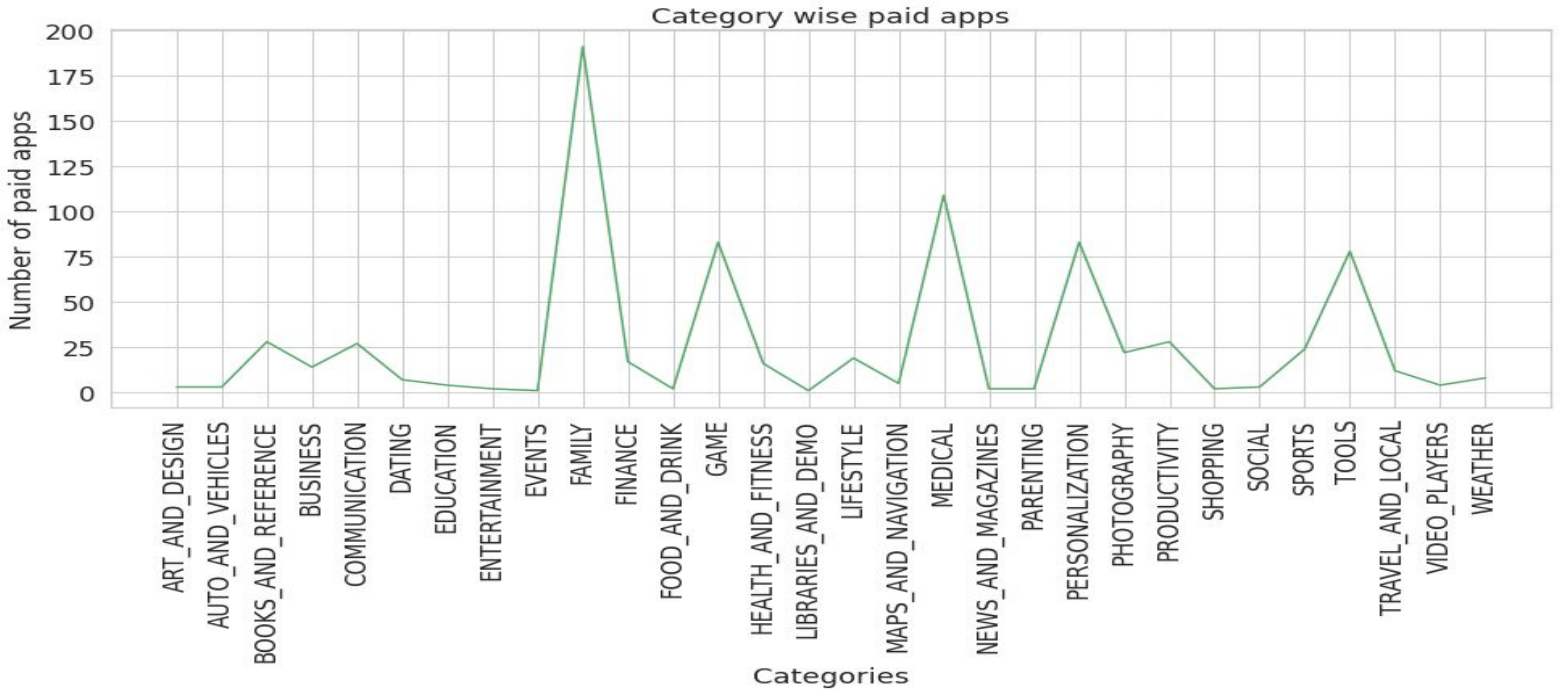
Most installed Apps

- Subway Surfers, Facebook, Messenger and Google Drive are the most installed apps.

	App	Installs
0	Subway Surfers	1.000000e+09
1	Facebook	1.000000e+09
2	Messenger- Text and Video Chat for Free	1.000000e+09
3	Google Drive	1.000000e+09

❖ Category wise paid apps

- The category ‘Family’ has the highest number of paid apps.



❖ Most costly Apps on Play store

- The app “I’m Rich — Trump Edition” from the category ‘Lifestyle’ is the most costly app priced at $400

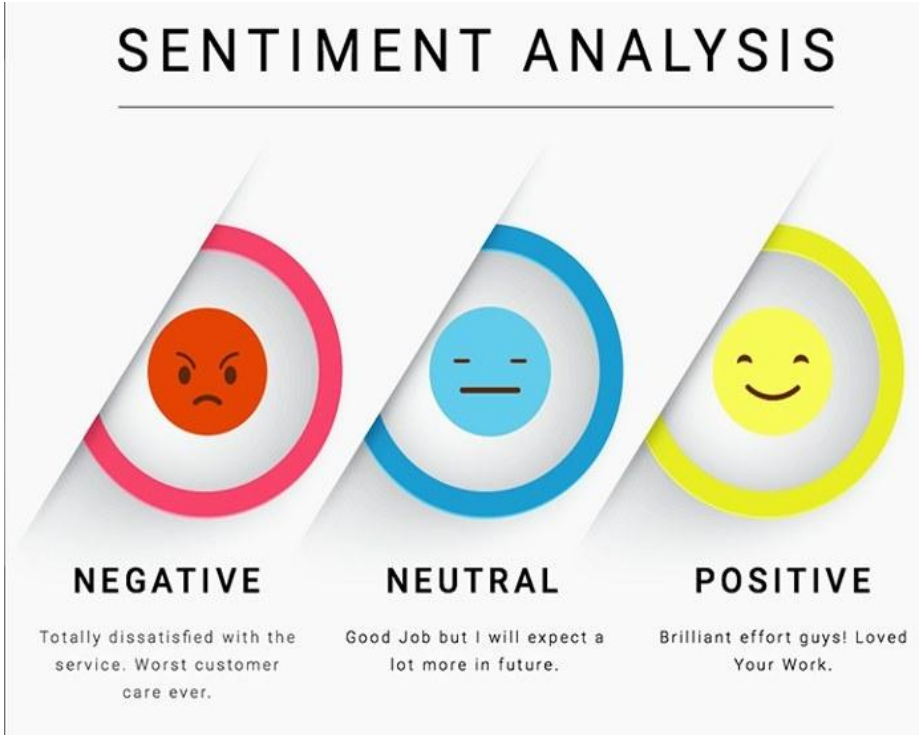
	App	Price	Category
0	I'm Rich - Trump Edition	400.00	LIFESTYLE
1	I am rich(premium)	399.99	FINANCE
2	I AM RICH PRO PLUS	399.99	FINANCE
3	I'm Rich/Eu sou Rico/أنا غني/我很有錢	399.99	LIFESTYLE
4	I am Rich Plus	399.99	FAMILY

Correlation Heatmap

- Correlation heatmap is graphical representation of 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 Installs are positively correlated with Reviews with correlation 0.64
- Price is slight negatively correlated with Rating and Reviews.

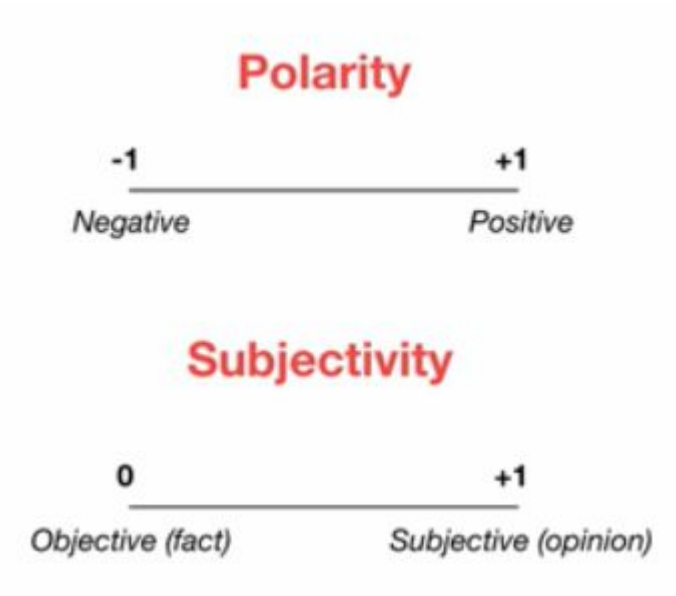
Sentiment Analysis

- We had done sentiment analysis on user_review data.
- Analyzing user sentiments towards apps through their review comments and ratings can be economically profitable to app developers.
- This user data contains name of App, Review given by the user, Sentiment of review, Sentiment Polarity and and Sentiment Subjectivity.

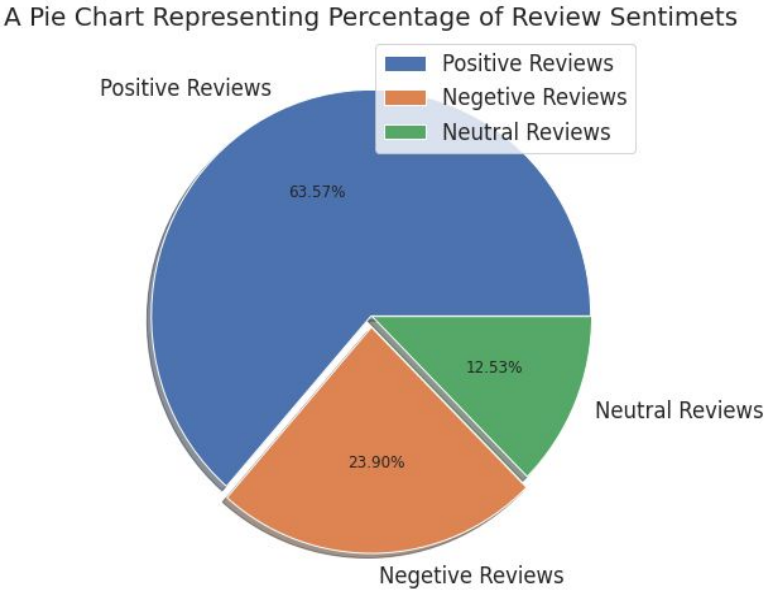


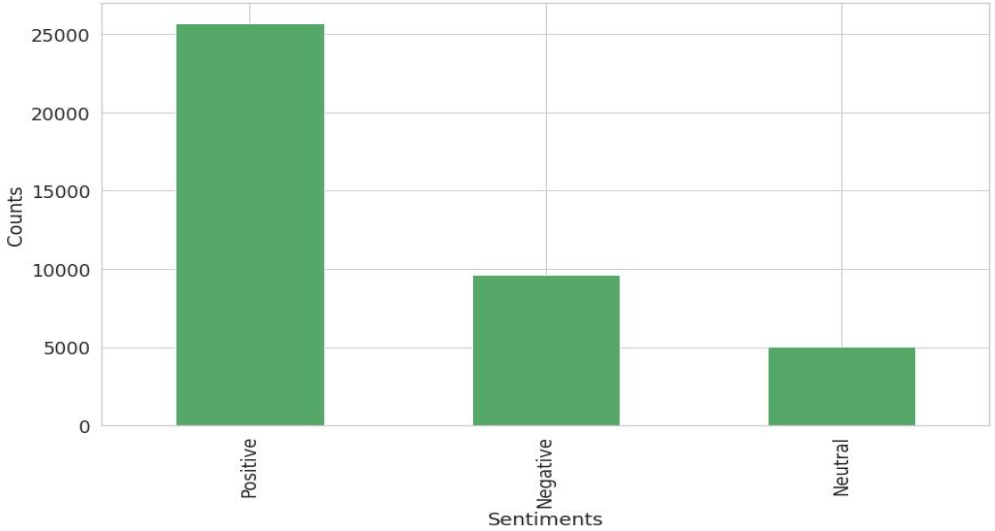
Understanding Features

- **App** - Name of App
- **Sentiment** - A view or opinion that is held or expressed. It can be Positive, Negative or Neutral.
- **Sentiment Polarity** -Sentiment polarity for an element defines the orientation of the expressed sentiment, i.e., it determines if the text expresses the positive, negative or neutral sentiment of the user about the entity in consideration. The polarity score is a float within the range [-1.0, 1.0]
- **Sentiment Subjectivity** - Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. Subjectivity lies between [0.0,1.0]. 0.0 is very objective and 1.0 is very subjective.



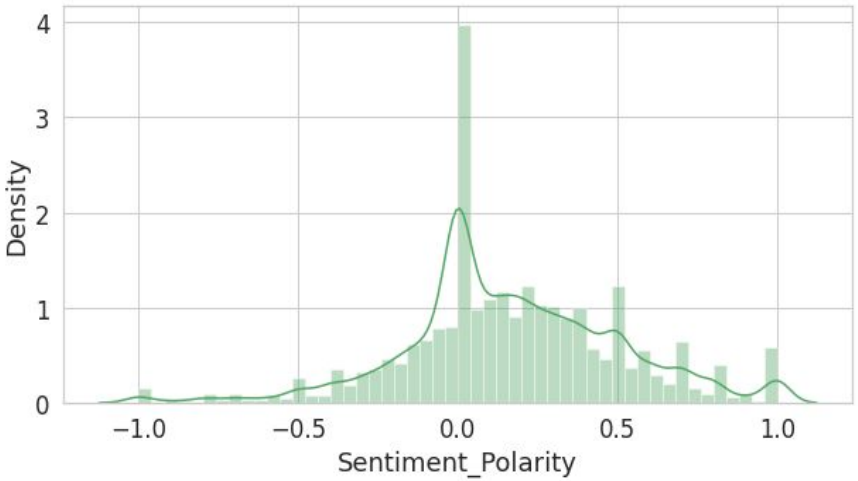
❖ Sentiment

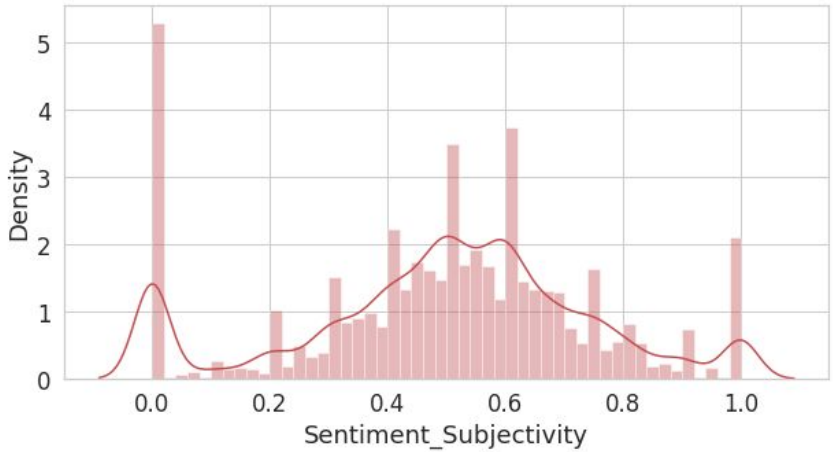




- Most of the reviews are of Positive Sentiment, while Negative and Neutral have low number of reviews.
- Out of total reviews there are 63.57% reviews are positive, 23.90% reviews are Negative and 12.53% reviews are of Neutral Sentiment.

❖ Distribution of Sentiment Polarity and Subjectivity



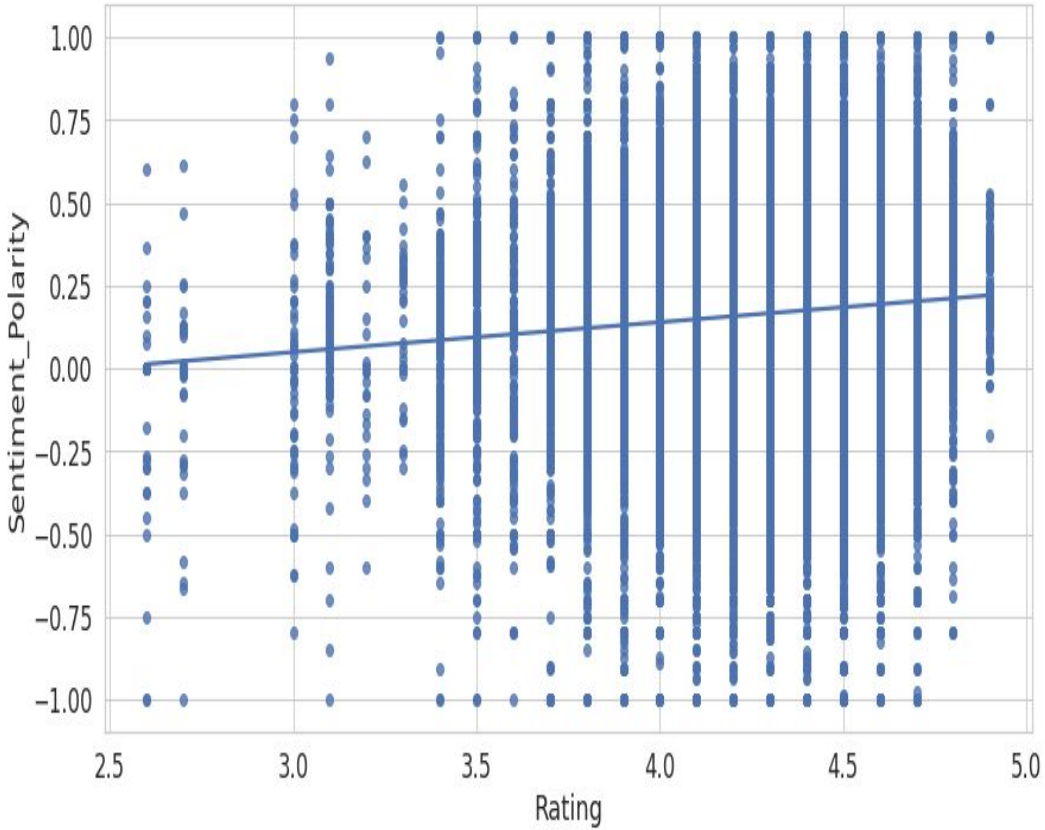


- As we can see in the above distplots most of the reviews fall in [-0.50,0.75] Polarity scale, extremely negative or positive sentiments are significantly low.

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

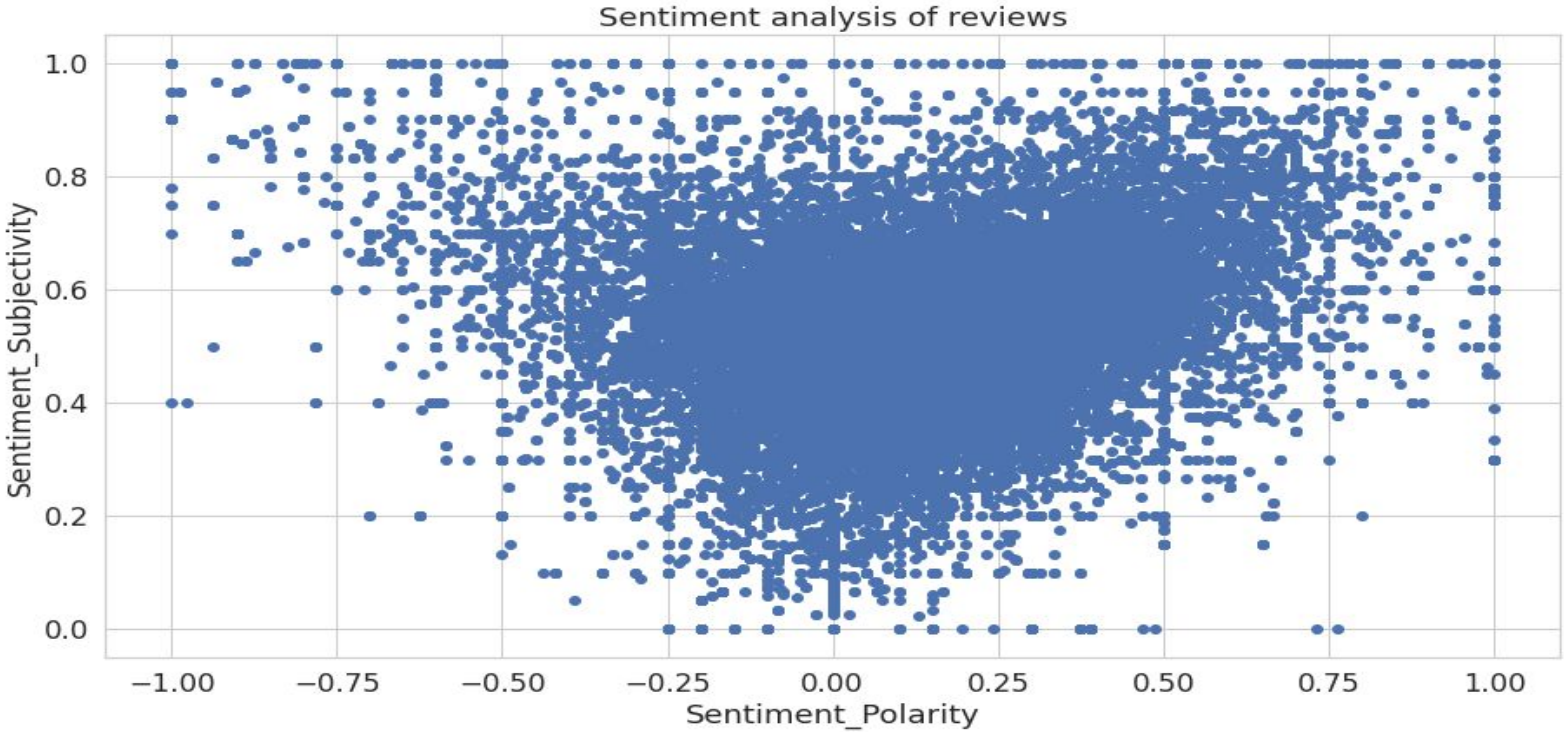
❖ Rating vs Sentiment Polarity

- After merging Play Store dataset and user_review dataset on App.
- We had plotted a regplot between Rating and Sentiment Polarity.
  .
- The correlation between sentiment polarity and rating is not as strong as we thought, though we can see the trend there.



❖ 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.



❖ Conclusions from scatter plot between Sentiment Polarity and Sentiment Subjectivity

- Polarity is concentrated mostly in the center and 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.
- 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 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.

Conclusion

- Most of the apps have rating in between 4 and 5.
- Most numbers of apps are rated at 4.3
- All categories of apps have more than 4 average rating.
- Maximum number of applications present in the dataset are of small size.
- Majority of the apps come into these three categories, Family, Game, and Tools.
- Maximum number of apps present in google play store come under Family, Game and tools but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps comes under Game, Communication, Productivity and Social.
- Subway Surfers, Facebook, Messenger and Google Drive are the most installed apps.
- About 92% apps are free and 8% apps are of paid type.
- The category ‘Family’ has the highest number of paid apps.
- Free apps are installed more than paid apps.
- The app “I’m Rich — Trump Edition” from the category ‘Lifestyle’ is the most costly app priced at $400
- Content having Everyone only has most installs, while unrated and Adults only 18+ have less installs.
- Number of installs is positively correlated with reviews with correlation 0.64

Conclusions (Continued..)

- Most of the reviews are of Positive Sentiment, while Negative and Neutral have low number of reviews.
- 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.
- Most of the reviews show a mid-range of negative and positive sentiments.
- 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 Polarity is not highly correlated with Sentiment Subjectivity.

The analysis of Google Play Store application aided to build most reliable and more interactive applications. This would be very useful for app developer to build an application focussed on certain discussed category in this analysis. This analysis will definitely help in building the application with precise and accurate objectives.

Thank You