Capstone Project 1

Play Store App Review Analysis

By- Saugata Deb
Harshad Savle
Prasoon Kumar

Introduction

- Android is the most popular operating system in the world, with over 3 billion active users spanning over 190 countries.
- Google Play was launched on March 6, 2012, bringing together Android Market marking a shift in Google's digital distribution strategy .
- Lots of designers and developers work on it to make an app successful on the Play Store.
- There are more than 5 million apps found on Google Play Store.

Contents

Objective

- The main objective of this Project is to gather and understand client demands and help them compete with competitors and also developers to make their Apps better and popular on the Play Store.

Problem Statement

1. Top Categories in Playstore?
2. Top Genres in the Playstore?
3. Top Content Rating per installation?
4. What is the percentage of free and paid Apps in the Play Store?
5. What is the effect of the last update on rating?
6. How does the last update have an effect on the trend of rating?
7. Effect on rating when the application was of type free and paid?
8. Relationship between reviews and rating?
9. relationship between Rating and Average Reviews
10. Average Rating of each App category

Problem Statement

11. Average Rating for each genre
12. What is the distribution of sentiment subjectivity?
13. How sentiment polarity varies with Free and Paid Apps?
14. Different percentages of review sentiments based on two Datasets provided?
15. Different percentages of sentiment analysis on top 5 Reviewed App Categories?
16. Sentiment Analysis on each App Category
17. Sentiment Analysis on the basis of Genres

Description of Dataset

There are two dataset: Play Store Data & User Data

1. **Play Store Data:-**

- **App** - Name of the Application
- **Category** - Category of the Application
- **Rating** - Rating given to the Application
- **Reviews** - No of reviews given to the Application
- **Size** - Size of the Application
- **Installs** - No of downloads of the Application
- **Type** - Free or Paid
- **Price** - Price of the Application if it is paid

Description of Dataset

- **Content Rating**-It is Age appropriate or Not
- **Genres** - Type of Genre the Application belongs to
- **Last Updated** - When the last time the Application is Updated
- **Current Ver** - Current version of the Application
- **Android Version**- Minimum Android version required to run the Application

Description of Dataset

2. User Review Data:

- **App** – An app name
- **Translated_Review**:- Reviews being given by consumer
- **Sentiment** – Sentiment given to an app by users (i.e. Positive, Neutral, Negative)
- **Sentiment Polarity** – The polarity of sentiment measures how negative or positive the context is. In the data we have, the polarity ranges from +1(Positive) to -1(Negative).
- **Sentiment Subjectivity** - The subjectivity of a sentiment is how likely that sentiment is to be based on data or factual information, versus personal opinions or public notions.

Data Cleaning

Data cleaning not just means removing the incorrect data or erroneous data. Many times we get the data which has all kinds of values some of them will cause problems during the analysis of the data and make our predictions incorrect. So we have to make sure our data has no erroneous values.

Data Cleaning Step:

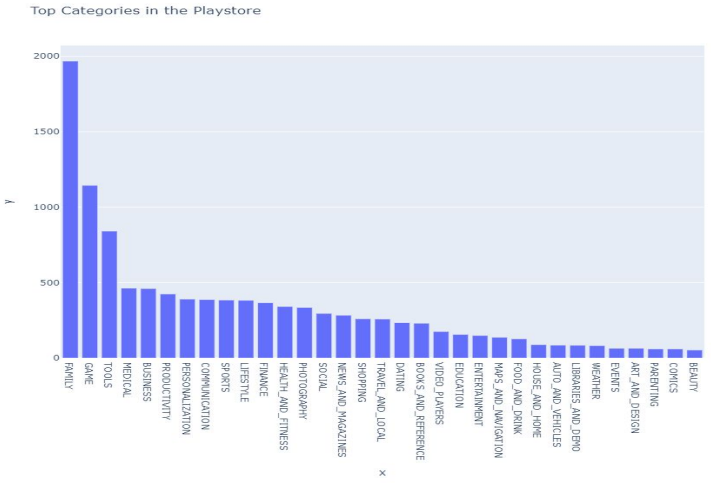
- **Removing unwanted Values** : Deleting of duplicate/incorrect or irrelevant values
- **Handling Missing Values**: Handling missing values in our Dataset
- **Handling Structural Errors**: Fixing mislabelled categories or classes, Types, Strange name conventions
- **Filtering Unwanted Outliers**: Removing incorrect or unwanted outliers
- **Replacing missing values with mean,median or mode**: Replacing missing values with median is the most popular method of replacing missing values

Data Analysis & Visualization

Top Categories in Playstore

- From above graph we can conclude that there is 1939 applications which falls under FAMILY category.
- Also top 5 categories of Application in the playstore are FAMILY,GAME,TOOLS,,BUSINESS,MEDICAL.
- Also there is only 53 applications which falls under BEAUTY category.
- We can also conclude that there is significant difference between the top two categories FAMILY 1939 Apps and GAME 1121 Apps.

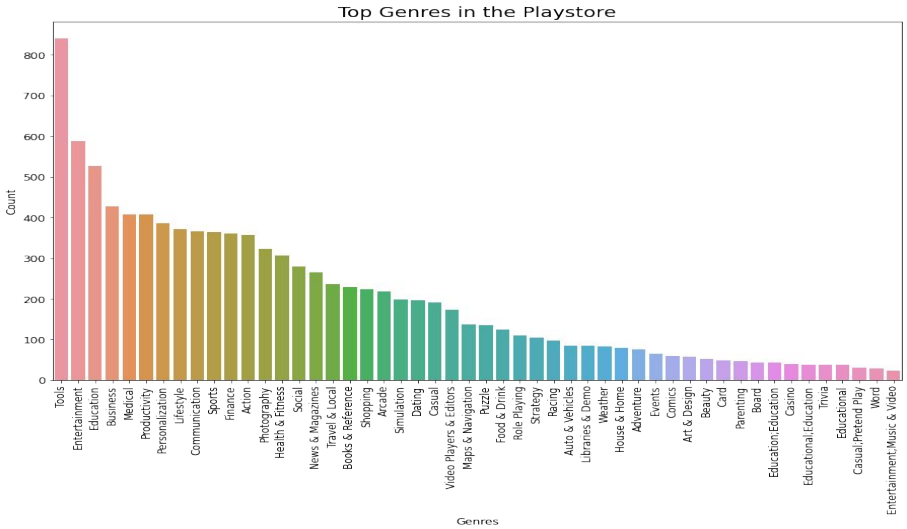
This shows that there are more application developers who develops Apps under FAMILY.



Data Analysis & Visualization(Cont..)

Top Genres in the Playstore

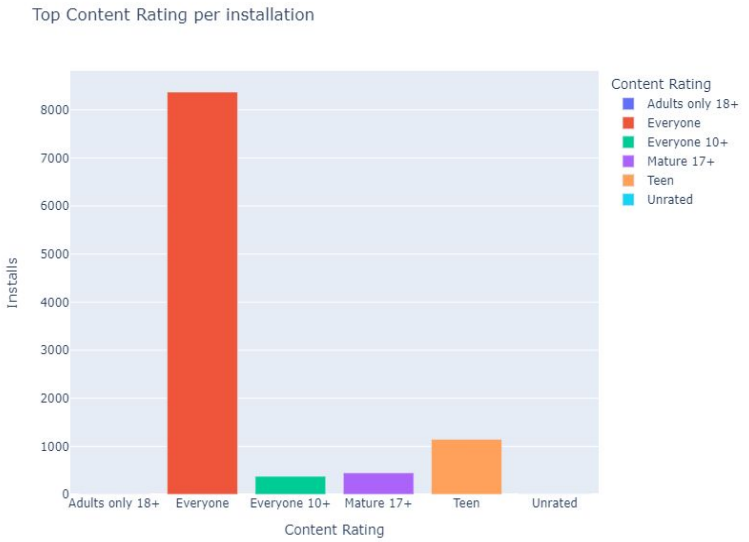
- From above bar plot in plotly we can conclude that maximum application which have been developed falls under App Genre Tools 840.
- Also from above we can observe from above plot that least applications were developed under App Genre Entertainment,Music & Video 23.
- Also top 5 Genres are Tools,Entertainment,Education,business and Medical which are 840,587,427,407,460 in Top 5 Genre name order.



Data Analysis & Visualization(Cont..)

Top Content Rating of each Category of App per installation

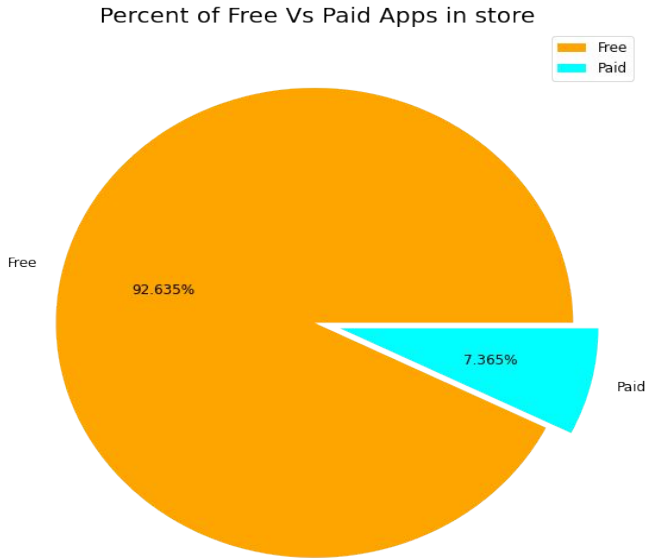
- From above graph by plotly we can conclude that App Having 'Content Rating' - **Everyone** is having maximum number installation of **100.228B**
- Also we can conclude that there are **4.29B** installation for applicatio having Content Rating of **Everyone 10+**.
- We can also conclude that there are no such noticeable installation for Application having Content Rating of **Adults only 18+** and **Unrated**.



Data Analysis & Visualization(Cont..)

What is the percentage of free and paid Apps in the Play Store?

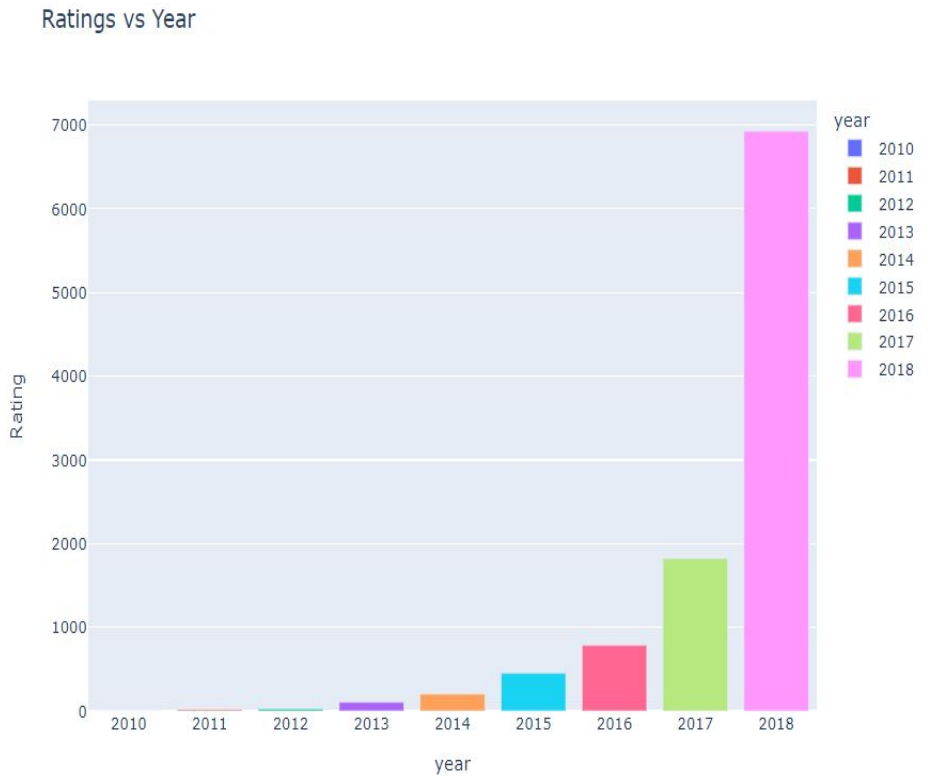
- From the above chart we can conclude that most of the apps available on the playstore are free which are enjoyed by most of the users.
    - Free Apps= 92.635%
    - Paid Apps= 7.365%



Data Analysis & Visualization(Cont..)

Effect of the last update on rating

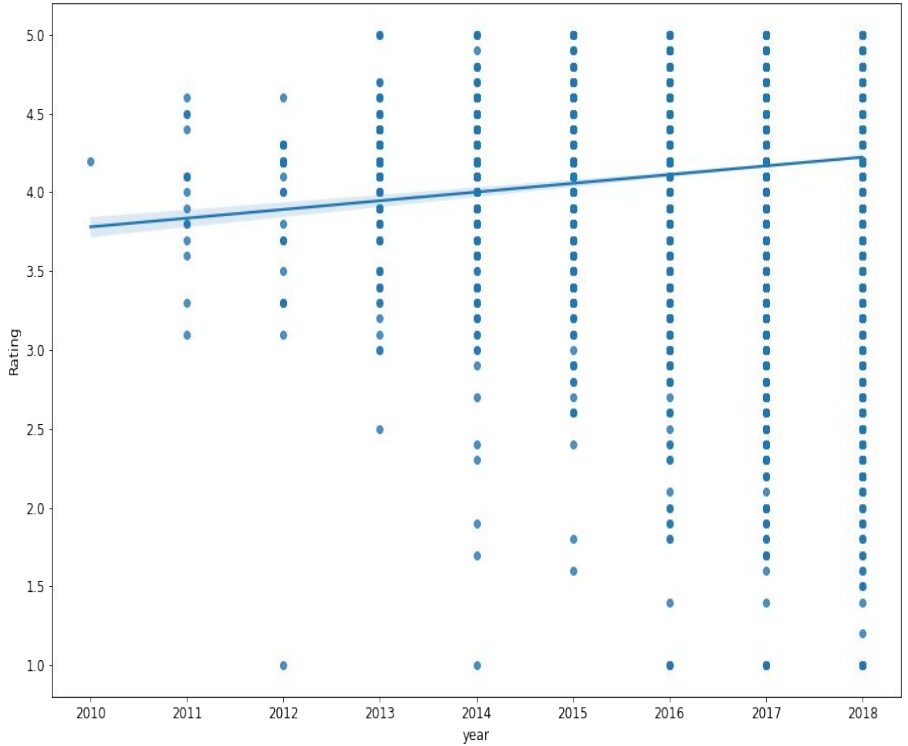
From this graph from plotly we can conclude that there is more number ratings given the application which are updated recently in 2018 no. of rating is 6929 than those application which were updated in 2017 no .of rating 1823. This shows with the latest update user reviews response increases for both less or more ratings.



Data Analysis & Visualization(Cont..)

Effect of the last update on the trends of rating

From this graph from matplotlib and seaborn we can conclude that rating is increasing in a proportionate manner with the last updated time. So from this we can be sure that with the latest update the reviewers are giving better rating.

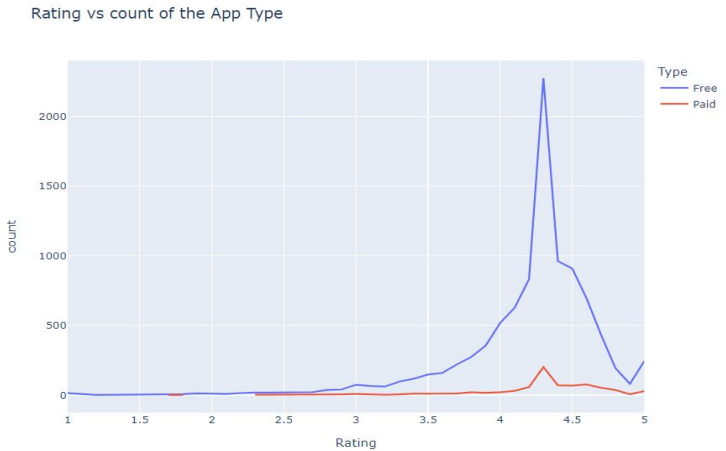


Data Analysis & Visualization(Cont..)

Effect on rating when the application was of type free and 'paid':

From the above graph using plotly we can conclude that the Free Apps has got more rating in terms of number of rating. From this we can also see that the users of free App are way higher than those who uses paid App.

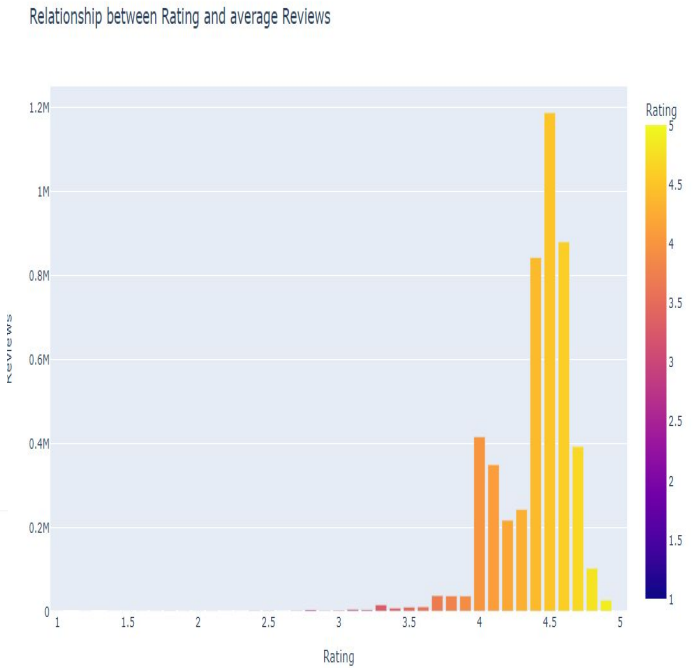
But we try watch very clearly the highest rating for both the Free and Paid Apps are 5. But if we go on finding the average of both the type of App we can see that the Average rating of the Free Apps will be less as compared to that of the Paid Apps . Again for the same that no. of users in Free Apps are way too high . For Example Free App for Rating 4.3 has a number count of 2275 as compared to only 201 count of rating for paid Apps with same Rating 4.3.

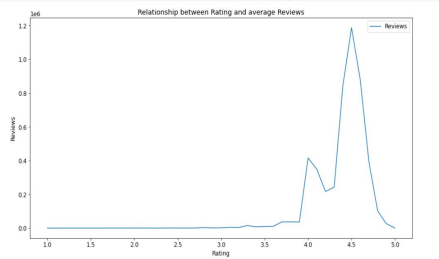


Data Analysis & Visualization(Cont..)

Relationship between Rating and Average Reviews:

- From the above graph we can conclude that as the rating increases the average reviews for each ratings also increases. Also we can observe some deviation after 4.5 rating as we can see 4.5 rating has maximum no.of reviews of 1.1M and reviews increases in quite proportionate manner with the increase of rating but after 4.5 rating reviews eventually decreases and we can see only around 26586 reviews.

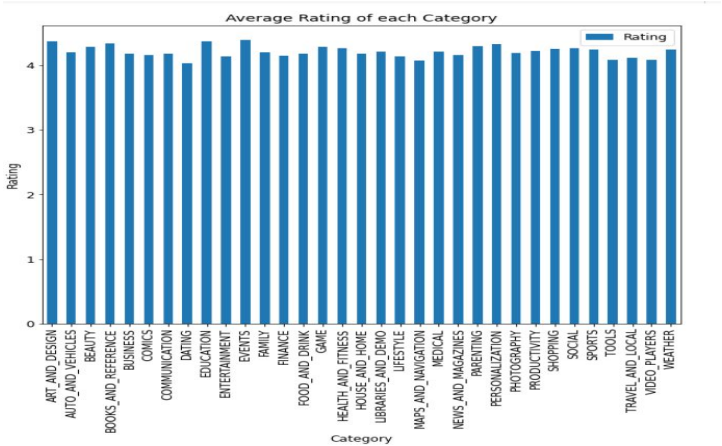


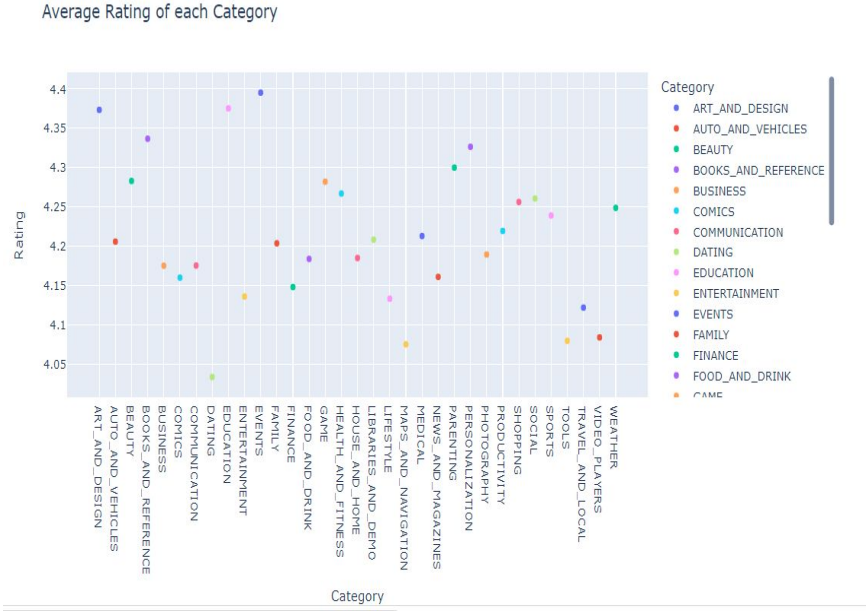


Data Analysis & Visualization(Cont..)

Average Rating of each category:

- From above plot from plotly we can conclude that 'Event' category has got highest Average Rating of 4.39.
- Also we can see that 'Dating' Category has got Lowest Average Rating of '4.033673'

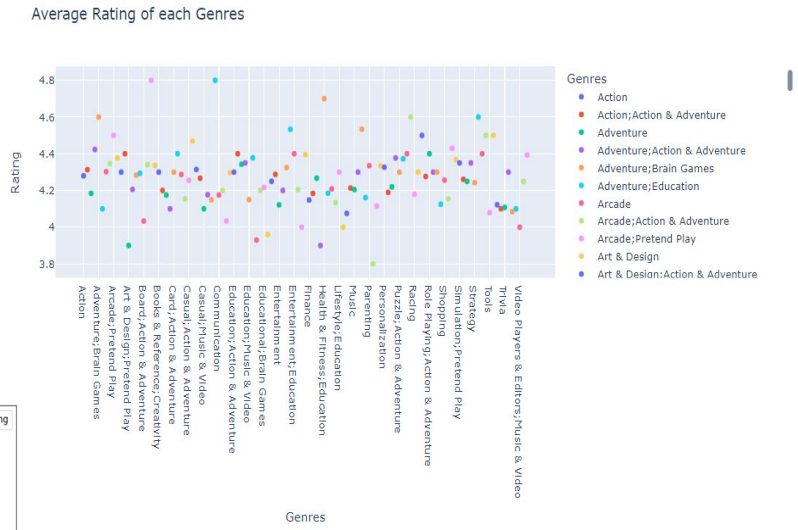


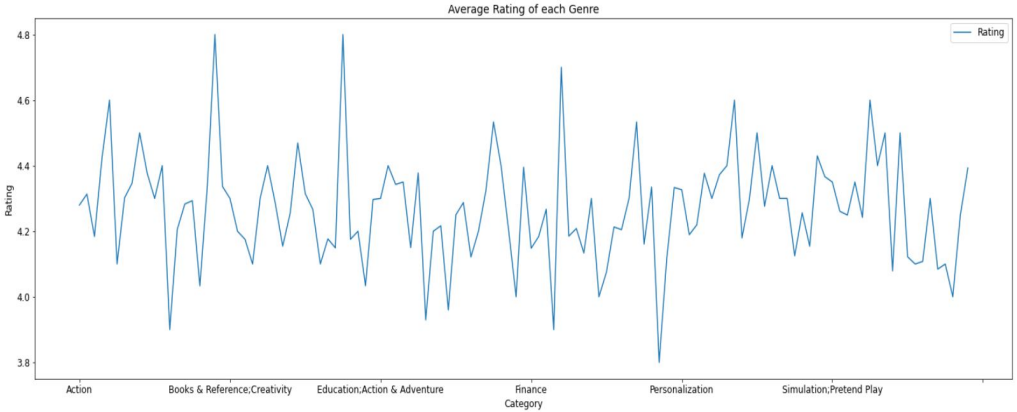


Data Analysis & Visualization(Cont..)

Average Rating of each genre:

From the scatter plot from plotly we can conclude that both the Genre-'Board;Pretend Play' and 'Comics;Creativity' is having the highest Average Rating of 4.8 and Genre-' Parenting;Brain and Games' has got lowest Average Rating.

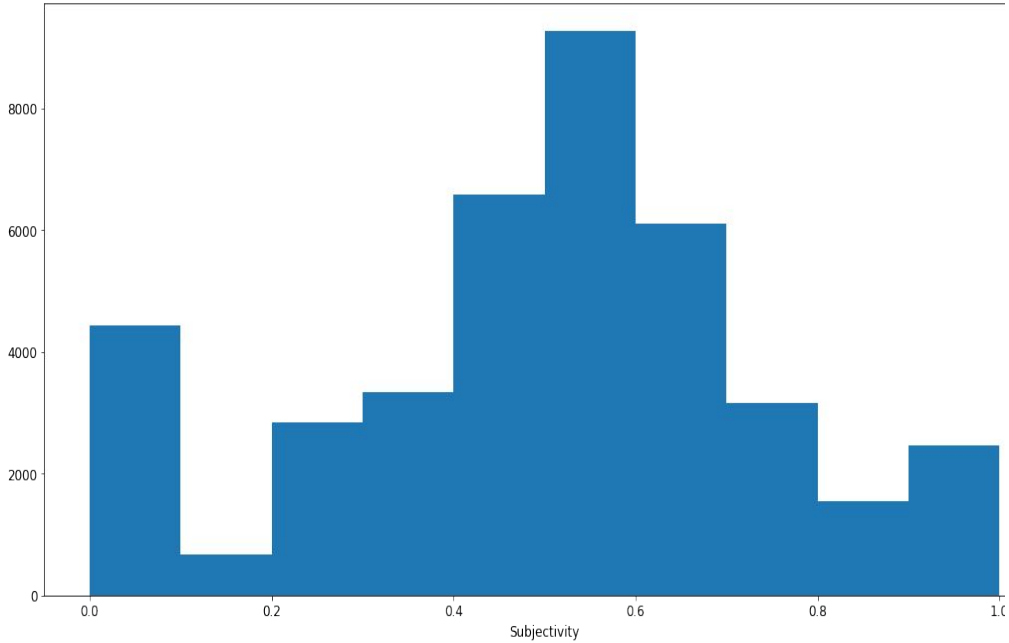




Sentiment Analysis

The distribution of sentiments subjectivity:

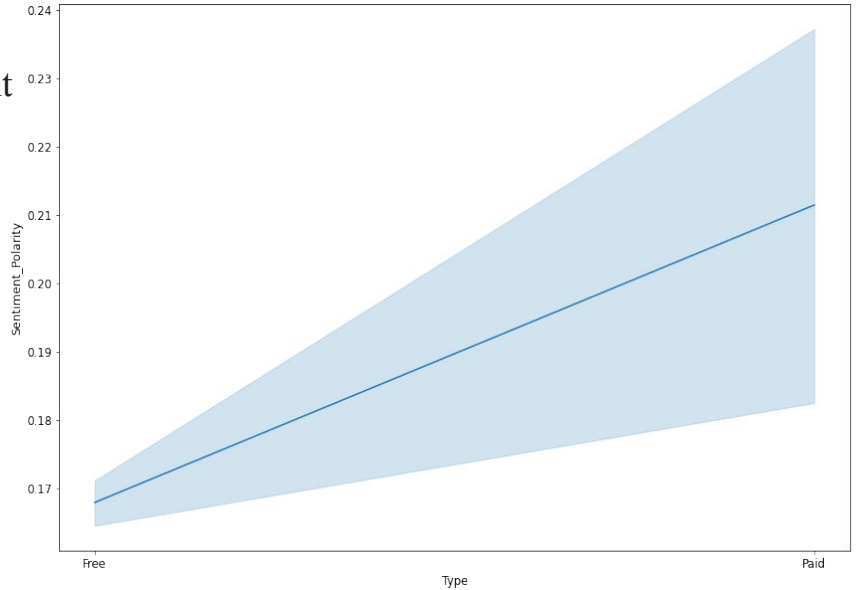
- It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications, according to their experience.



Sentiment Analysis(Cont..)

How sentiment polarity varies with Free and Paid Apps?

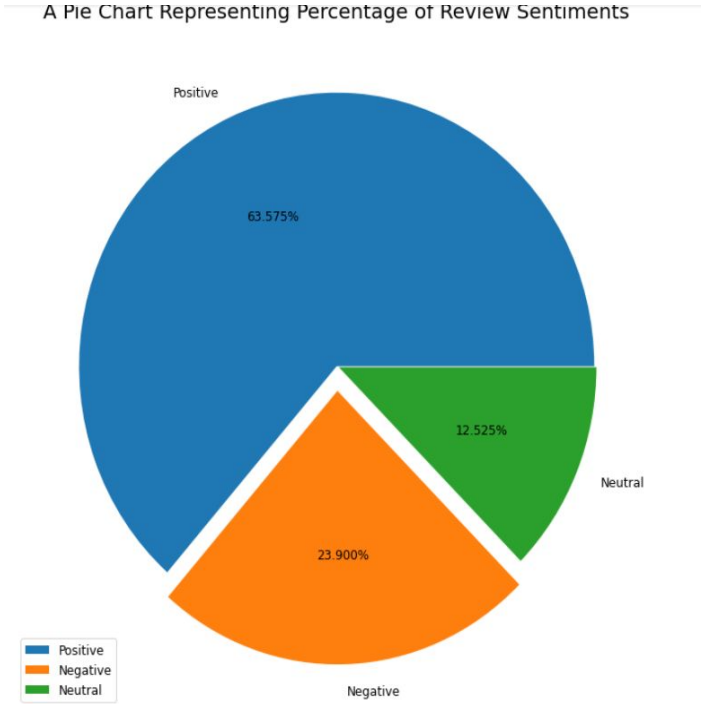
- From the above line plot we can conclude that with increase in sentiment polarity ,the sentiment polarity for the paid app is higher than the sentiment polarity for free app. This means people has more sentiment towards paid App than free App.



Sentiment Analysis(Cont..)

Different percentages of review sentiments based on two Datasets provided?

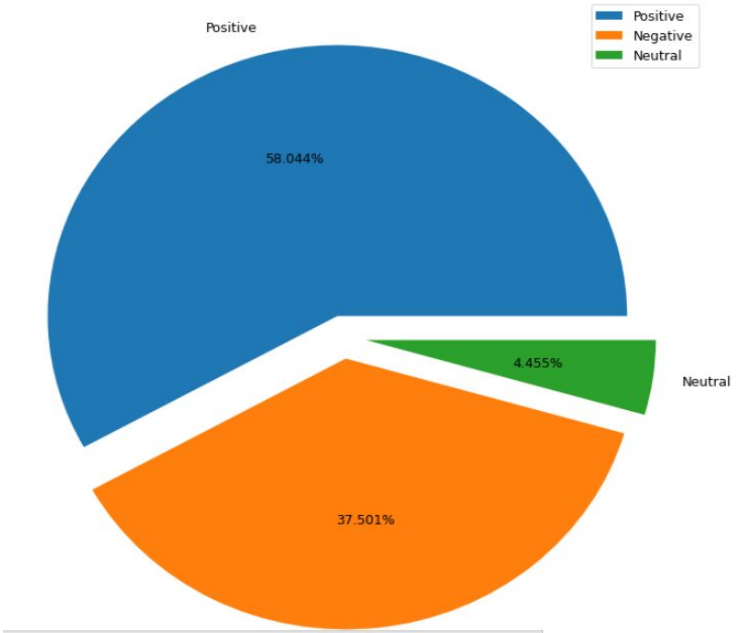
- From this pie Chart we can conclude that most of the sentiment reviews given by the user are positives with 63.625%. But also there is a negative sentiment percentage of 24.976% which is higher than the one with the neutral sentiments with 11.399%. This means app developers needs to convert more negative sentiments to neutral or positive sentiments with their Hard work



Sentiment Analysis(Cont..)

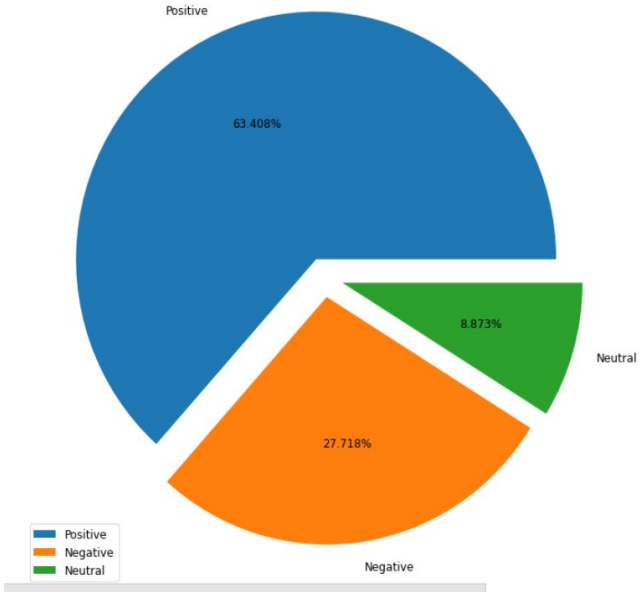
Different percentages of sentiment analysis on top 5 Reviewed Categories

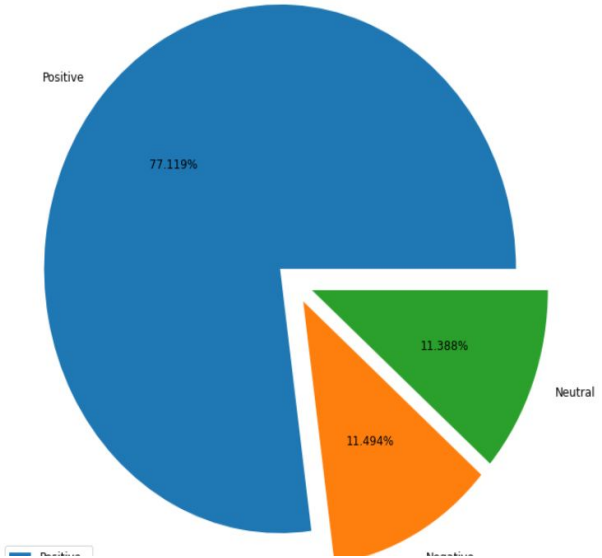
- From the above figure we conclude that 'HEALTH_AND_FITNESS' Category has got highest positive percentage of 77.119% and negative sentiment percentage of 11.494% and neutral percentage of 11.388%
- Also we found that top Category GAME has less positive sentiment percentage of 50.04% than it's competitor.
- Most negative sentiments from the top translated app category has been received by GAME CATEGORY.THIS shows that even if GAME app has the highest translated reviews but in positive sentiment percentage it is lower than it's counterparts.
- Highest percentages of neutral sentiments has been claimed by SPORTS category from the list top 5 App category.



Sentiment Analysis(Cont..)

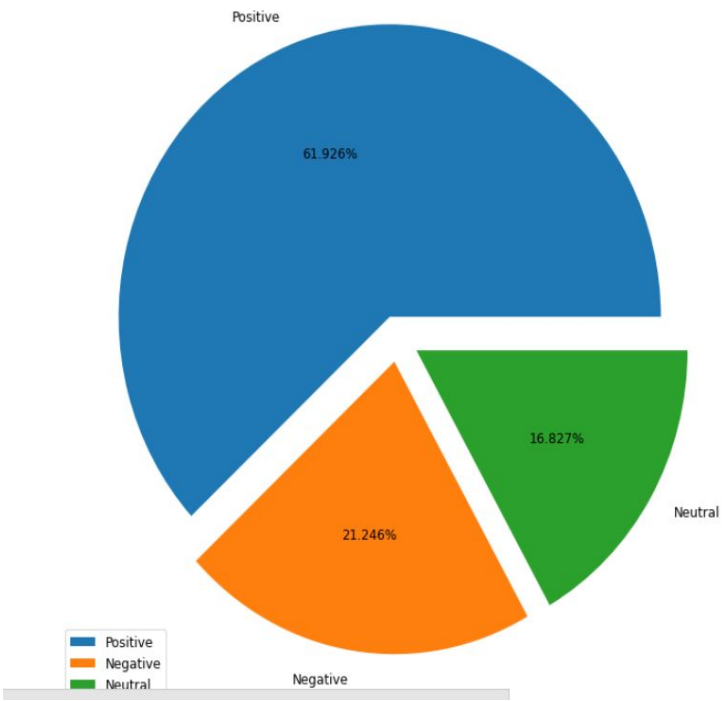
Different percentages of sentiment analysis on top 5 Reviewed Categories

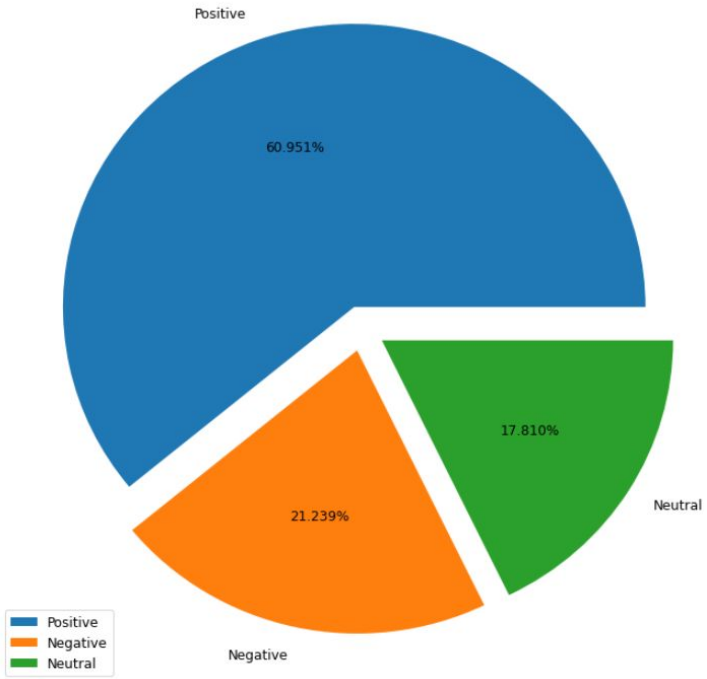




Sentiment Analysis(Cont..)

Different percentages of sentiment analysis on top 5 Reviewed Categories

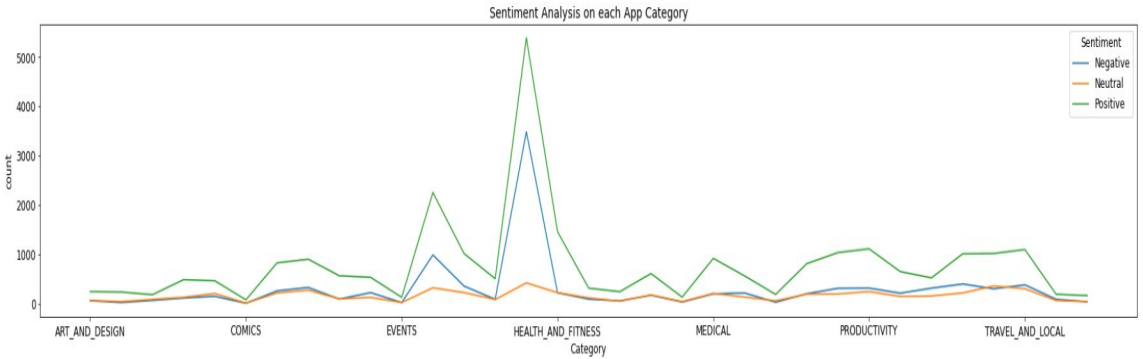


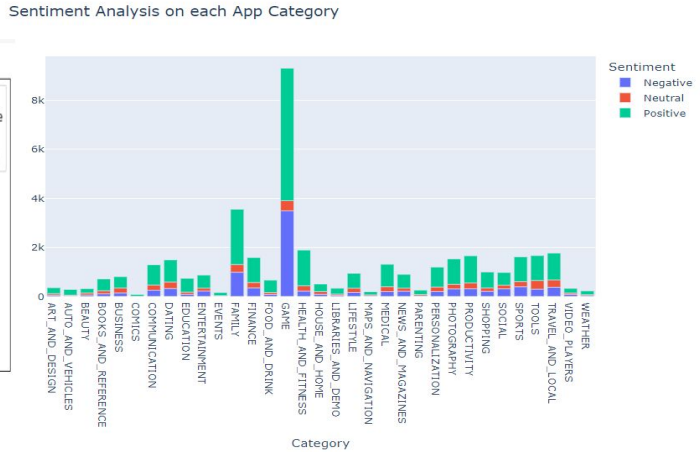


Sentiment Analysis(Cont..)

Sentiment Analysis on each App Category

- From Above bar plot we can conclude that GAME Category has got most positive sentiment in terms of count but if we take percentage of positive vs negative sentiment we will found that HEALTH_AND_FITNESS has got the highest positive sentiment percentage.
- Again if we count on the basis of the count of positive sentiment we can see that Category comic has got least number of positive sentiment and no negative sentiment as such.

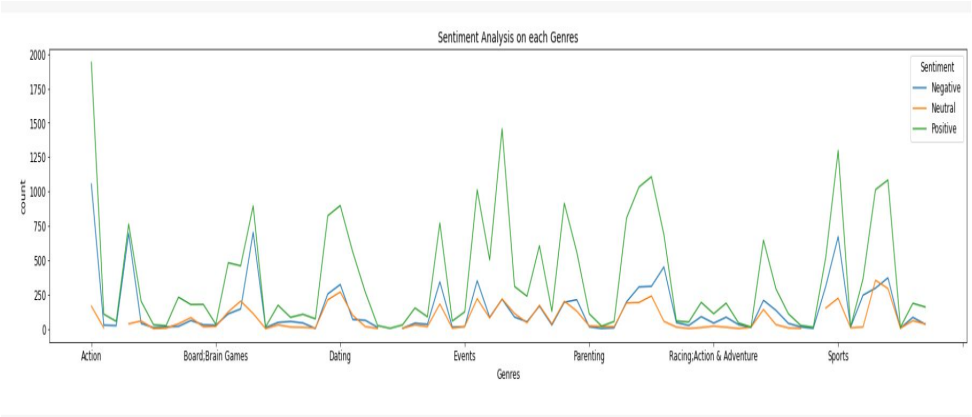


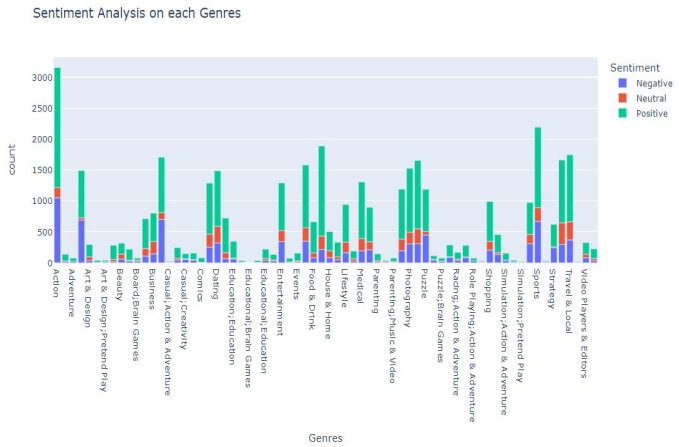


Sentiment Analysis(Cont..)

sentiment Analysis on the basis of Genres

- From Above bar plot we can conclude that **Action** Genre has got most number positive and negative sentiments.
- We can see that **Simulation;Pretend Play** has got least number of positive sentiments
- Also we can notice that '**Education;Brain and Games**' has got no Sentiment at all.





Conclusion:

- So here we come at the end of our project which is play store App Review Analysis.What we have done just take a short recap. First we have done the removal of null value from rows and columns and the same goes with the removal of duplicates from the datasets. Then we did the formatting for each of the required columns in each dataset.
- After analyzing the data we conclude that App with the category Family and the genre tools are in large numbers. Also we can conclude that the number App Rating is directly proportional with the recent update. From this we can see that with all the major updates apps will get more ratings.
- We can also conclude that most of the apps which are used by the users have a content rating of ‘Everyone’.
- In percentage of Free and Paid App Available in the Play Store we can assume that most apps being used by the users are Free. This shows very few users purchase Apps on playstore.
- In rating vs count of App Type we conclude that rating is not get affected even if the app is paid or not but if we go on for finding the average rating we will find that free app will have less average rating compared to paid because of significantly high counts of free Apps as compared to Paid App available in App Store.

Conclusion:

- After moving forward when we performed analysis on sentiment subjectivity we found that most of the opinion on sentiment subjectivity lies high in the range 0.4 to 0.7.
- When we analyzed sentiment polarity for paid and free Apps we noticed that sentiment polarity for free apps is way less than paid Apps.
- In pie presenting the percentages of review sentiment we found that most of the sentiment are positive and neutral review is the lowest. Also in case finding the percentage of sentiments for top 5 Apps we found among top 5 App Category Health and Fitness has received the highest positive sentiments while Game app category has received the highest negative sentiments and Sports App Category has received the highest neutral sentiments.

Important Points to be noted:

1. All active apps on play store has an an Average Rating of 4.32.
2. Also we can see that after merging both of the dataset the maximum Average Rating is 4.9.
3. Also the average sentiment Polarity is 0.16 and average sentiment_subjectivity is 0.49
4. Also we have noticed that the average size of the Application available on playstore is 21933.38 KB
5. lastly, all of the calculations and graphs in this project have accuracy in the range of 75% to 80%.