Capstone Project

On

Play Store App Review Analysis

by

1. Subodh Shankar Dooganavar
2. Kasmin Talukdar
3. Mohammed Maaz Ansari
4. Muktesh Singh

Introduction

- Currently, the Google play store is the most dominant android app marketplace. At this time, It contains more than 2.5 million apps and thousands of apps are launched every single day.

- Since a number of the dominant player with several quality apps are already present in the market, there is tough competition for newcomers. To survive and grow in this competitive market we need a great strategy.

- To find the answer to many such questions we are going to do a detailed analysis of over ten thousand apps in Google Play across different categories.

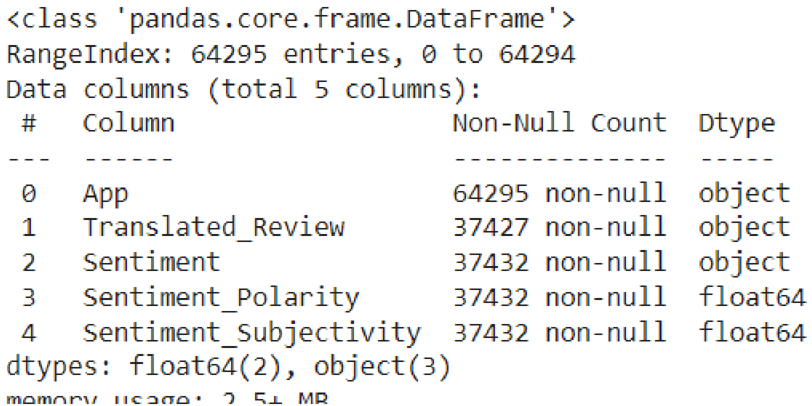
Problem Statements:

1. What is the Corelation between different variables?
2. What is the ratio between paid app and free app?
3. Total Number of apps in each category?
4. What is the percentage of review sentiments?
5. Most apps in terms of content rating?
6. Number of apps based on size?
7. Lets us discuss the sentiment subjectivity.
8. Relation between Sentiment, Sentiment Subjectivity, Sentiment Polarity?
9. Distribution of app update over the years?
10. What is the count of apps in different genres?
11. Top 10 apps in paid type by revenue?
12. Does price of the app affect the rating?

Data Summary

Details of dataset – 1) **Play Store Data.csv**
2) **Reviews. csv**

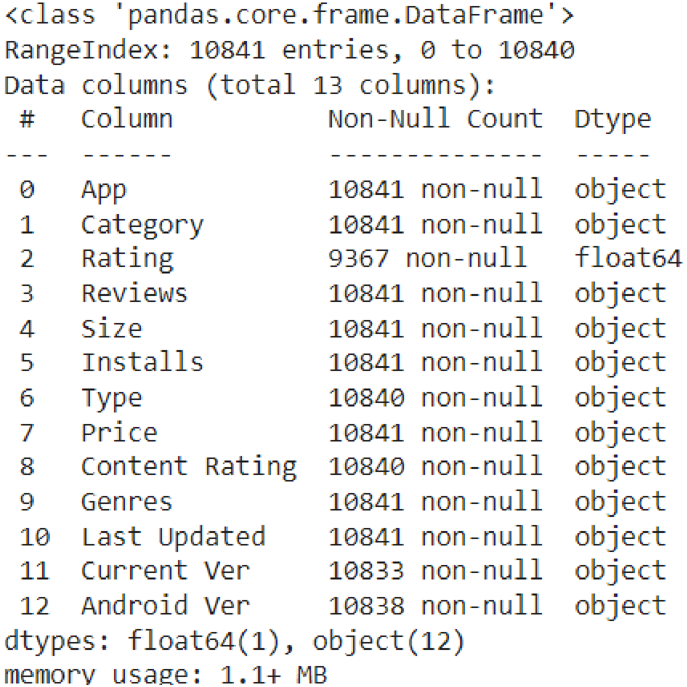
Reviews. csv



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64295 entries, 0 to 64294
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   App                     64295 non-null  object 
 1   Translated_Review       37427 non-null  object 
 2   Sentiment               37432 non-null  object 
 3   Sentiment_Polarity      37432 non-null  float64
 4   Sentiment_Subjectivity  37432 non-null  float64
dtypes: float64(2), object(3)
memory usage: 2.5+ MB
```

1. **App** - describes the name of each apps.
2. **Translated_Review**-English translation of the user reviews.
3. **Sentiment** - It gives the emotion of the reviewer related with his review.It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
4. **Sentiment_Polarity** - The polarity of the review.It ranges from [-1 to 1]. -1 means "negative sentiment" and 1 means "positive sentiment"
5. **Sentiment_Subjectivity** - How the opinion of a particular reviewer is aligned with the opinion of the general public.It ranges from [0 to 1]. Heigher the subjectivity means the review is closer to the opinion of general public.

Play Store Data.csv



```
<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
```

Features in the Dataset

1. **App** - describes the name of each apps.
2. **Category** - Category at which the app belongs.
3. **Rating** -Average rating of the app received from its users.
4. **Reviews** - The total number of reviews got by the apps from its users.
5. **Size**- Memory size occupied by the app in the mobile device.
6. **Installs**- The total number of downloads for the application.
7. **Type** - whether the app is free or paid
8. **Price**-If the app is paid, what is the cost required to install the app.
9. **Content Rating** - It specifies weather the app is suitable for all age group or not.
10. **Genres**- the other categories to which the app can belong.
11. **Last Updated**- Last updated date of the app.
12. **Current Ver**-The current version of the app.
13. **Android Ver**-The Android Version which can support the application.

Data Cleaning

Important Steps of Data Cleaning

•Identifying the null values.

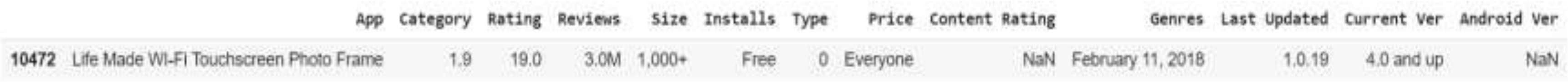
•Identifying the invalid data.

•Removing Symbols.

•Standardizing the data types.

To start with **Play Store Data.csv**

- We began our univariate analysis with the Reviews Column, and found out that everything was correct, except for the **3.0M** which doesn't seem to be a review.
- We then explored it and found the below image.



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up	NaN

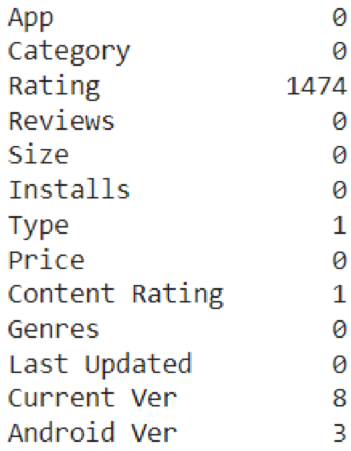
•It is evident from the above image that rowno.10472 has garbage value, as all the columns have mismatch of data.

•We thus decided to drop this row from the dataset

Data Cleaning(continued)

- **Reviews** column was converted to numeric type.
- **Size** column had ‘**M**’, ‘**k**’ & ‘**Varieswithdevice**’ where ‘M’ & ‘k’ were striped and ‘**Varieswithdevice**’ was replaced with np.NaN and then converted to a numeric variable.
- **Installs** column had ‘**+**’ & ‘**,**’ characters present in it, which were removed, and then the column was converted to numeric type.
- **Type** column has only two strings i.e., Free & Paid which were relevant and retained as it is.
- **Price** column had ‘**$**’ symbol present, which was removed and then the column was converted to numeric type.
- **Content_Rating** column has relevant variables present in it and thus were retained as it is.
- **Genres** column also had relevant variables present in it and thus were retained as it is.
- **Last_Updated** column has the dates in string format, these were converted to date time format.

Handling null values



```
App                  0
Category             0
Rating            1474
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
Content Rating       1
Genres               0
Last Updated         0
Current Ver          8
Android Ver          3
```

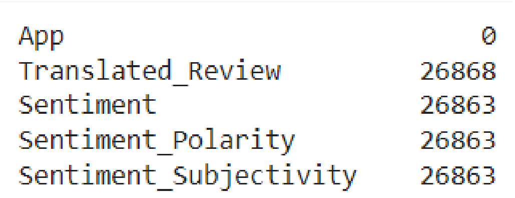
- Rating column contain 1474 NaN values. We cannot drop this much amount of row from dataset because it will lose a huge percentage of information. We replaced all the NaN values with the average of non-null values.
- Type column contain Null value at index 9148 so it was removed.
- 8 null values were present in the 'Current Ver' column. All of them were removed.
- 2 null values were present in the ‘Android Ver' column. Both of them were removed.

Removing Duplicates

A total of **798** duplicates were present in the App column. So all the duplicate values were removed from the data set.

After removing "Duplicates" and "NaN values" from the Dataframe we now have a modified Dataframe with 9648 rows and 15 columns.

Handling Null values in User data review Dataframe

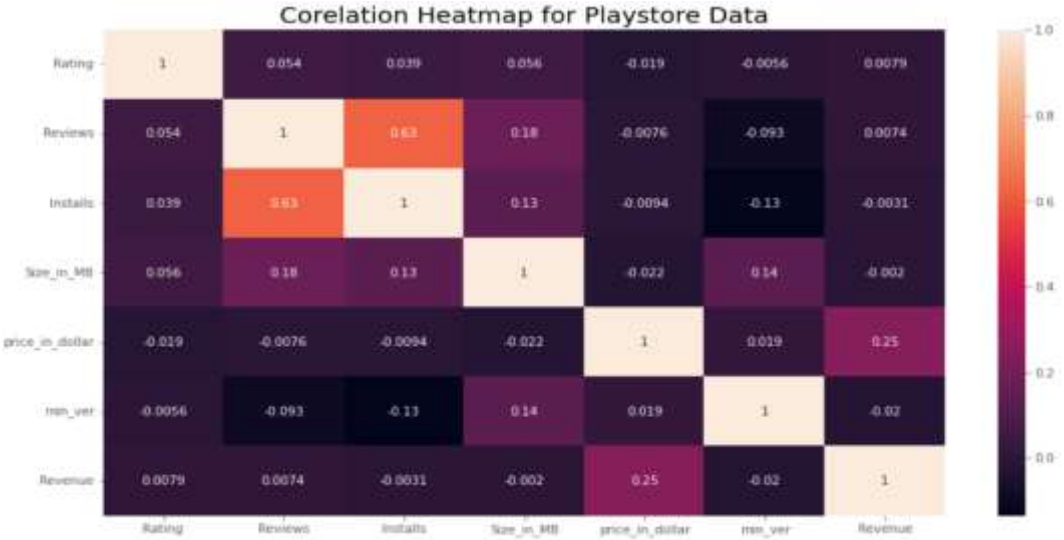


```
App                           0
Translated_Review         26868
Sentiment                 26863
Sentiment_Polarity        26863
Sentiment_Subjectivity    26863
```

- There are 26868 Null Values in T**ranslated Review** column.
- Removing NaN values from **Translated_Review** column, because the rows containing NaN values are of no use and we cannot impute null values for this column. If there is no review then there will be no sentiment.
- Therefore, We will remove all the rows that contains NaN values in Translated_Review column.

After removing the null values from Dataframe we have modified Dataframe with 37427 rows and 5 columns.

Corelation between different variables

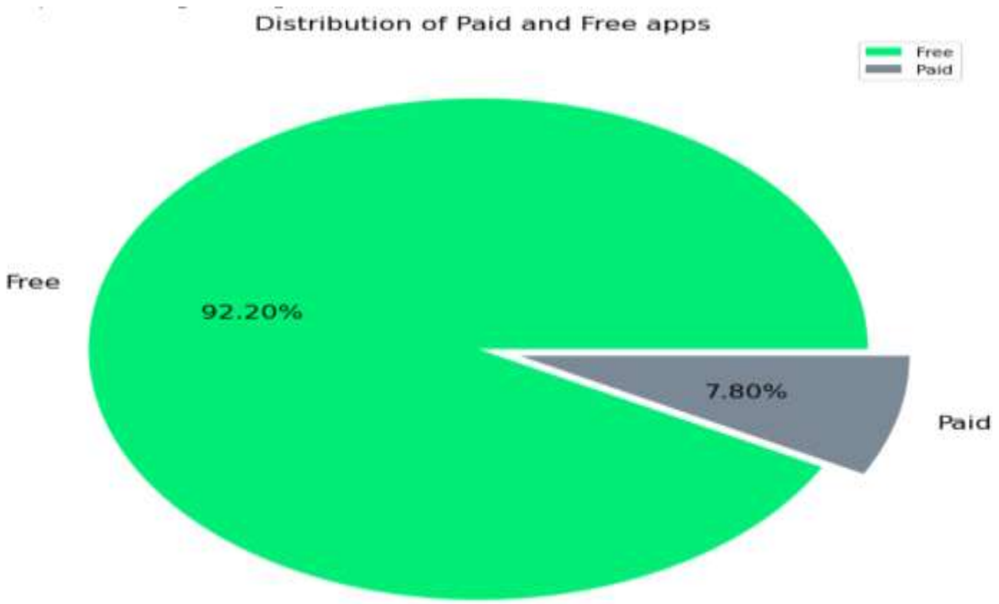


Findings

•There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.

•The Price is slightly negatively correlated with the Rating, Reviews, and Installs.` This means that as the prices of the app increases, the average rating, total number of reviews and Installs fall slightly.

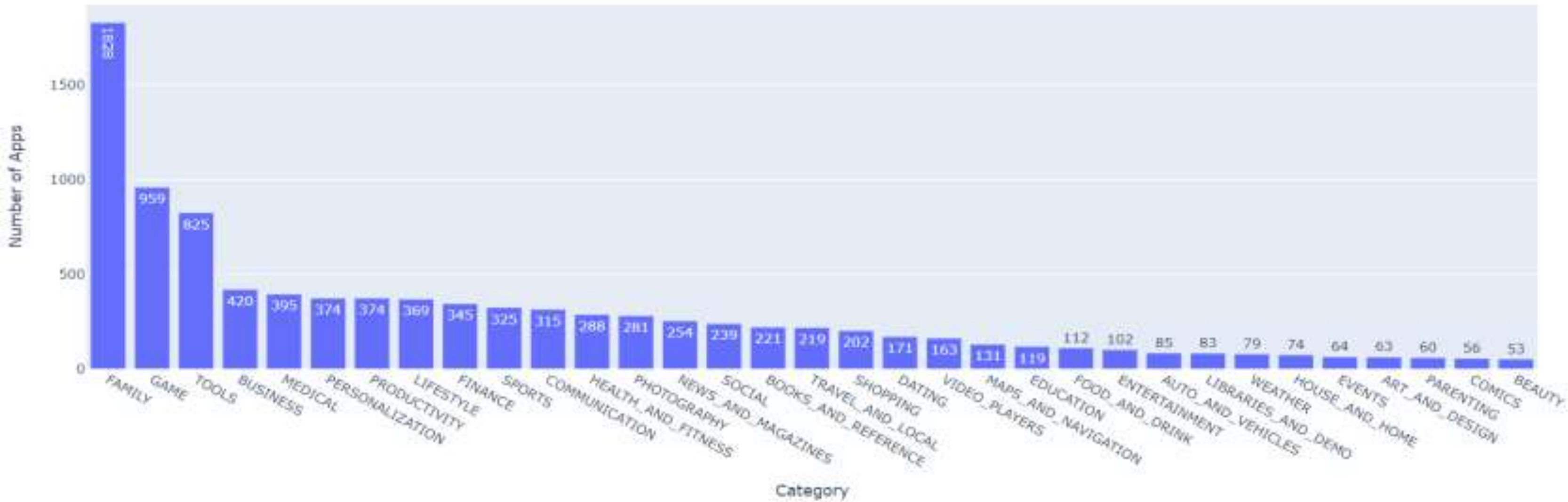
What is the ratio between paid app and free app?



Findings

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

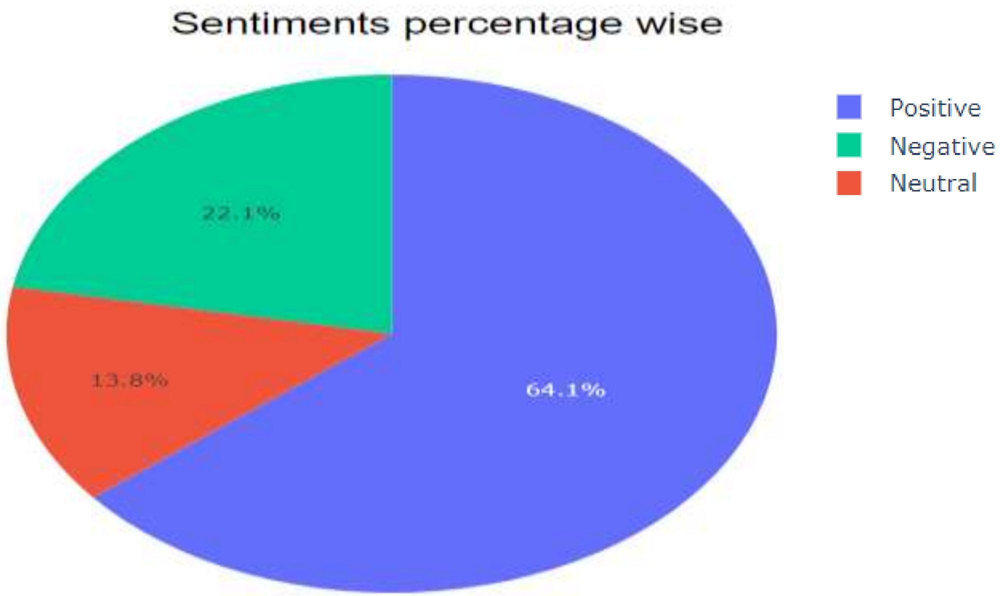
Total number of apps in each category?



Findings

As we can see most number of apps in the Play store are of **Family** category followed by **Game** and **Tools** category.
Beauty and **Comics** category has least number of apps.

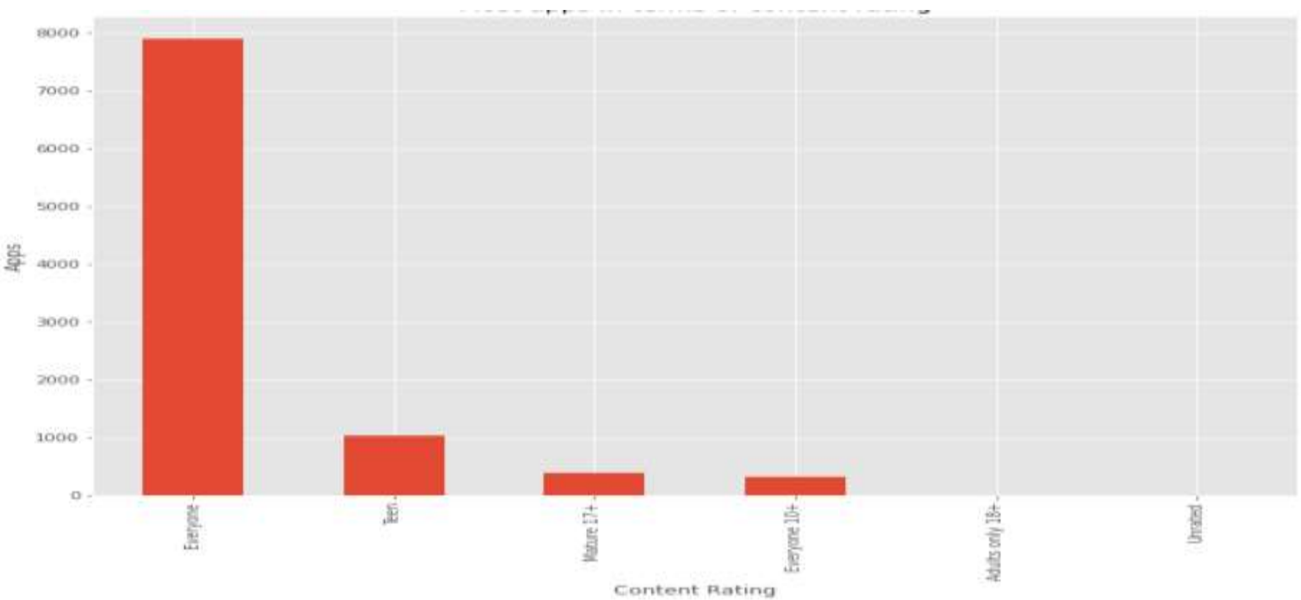
What is the percentage of review sentiments?



Findings

Reviews obtained from customers about play store apps **64.1%** are of **Positive** sentiment followed by **Negative** review which is **22.1%** and **13.8%** reviews are of **Neutral** type.

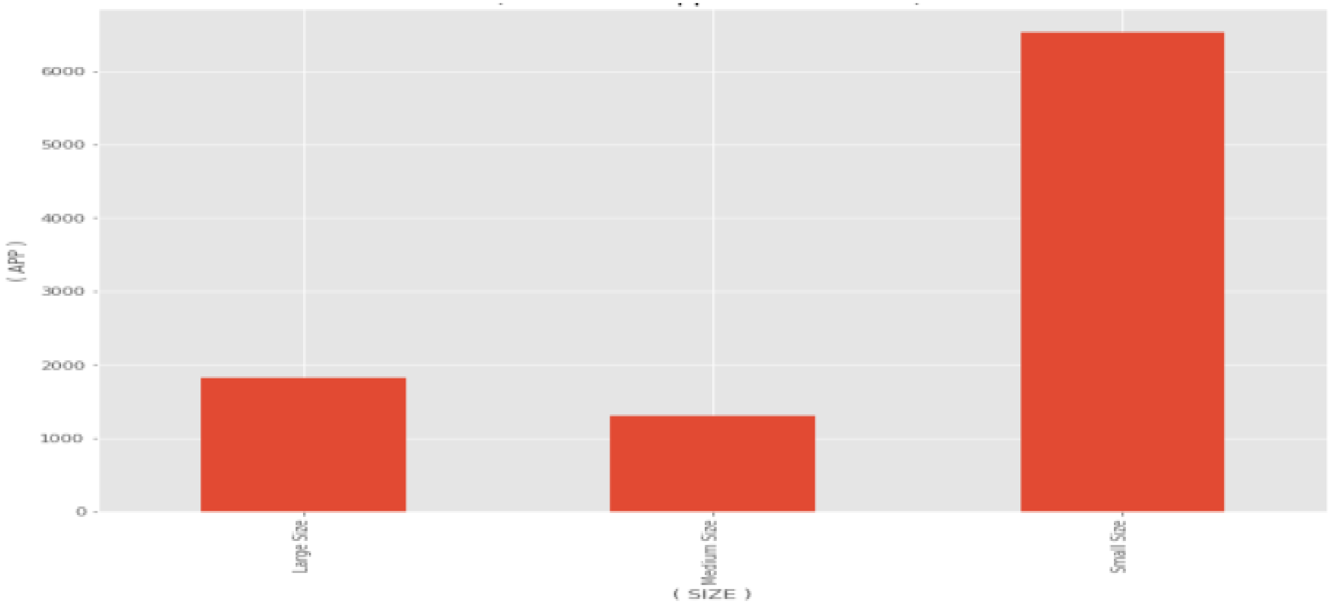
Most apps in terms of content rating?



Findings

Apps with content rating available for everyone attract more number of users.

Number of apps based on size?

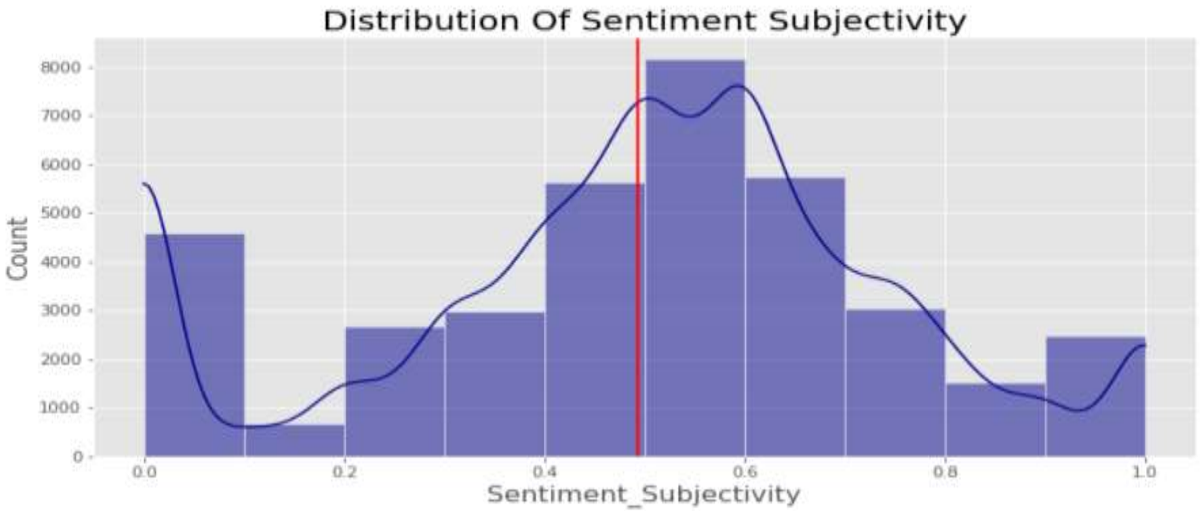


Findings

Small size <=30MB, 30MB<Medium size<=60MB, Large size>60MB

Most apps are based on **Small Size(67.6%)**, followed by **Medium Size(18.9%)** and **Large Size(13.5%)**.

Histogram of Subjectivity

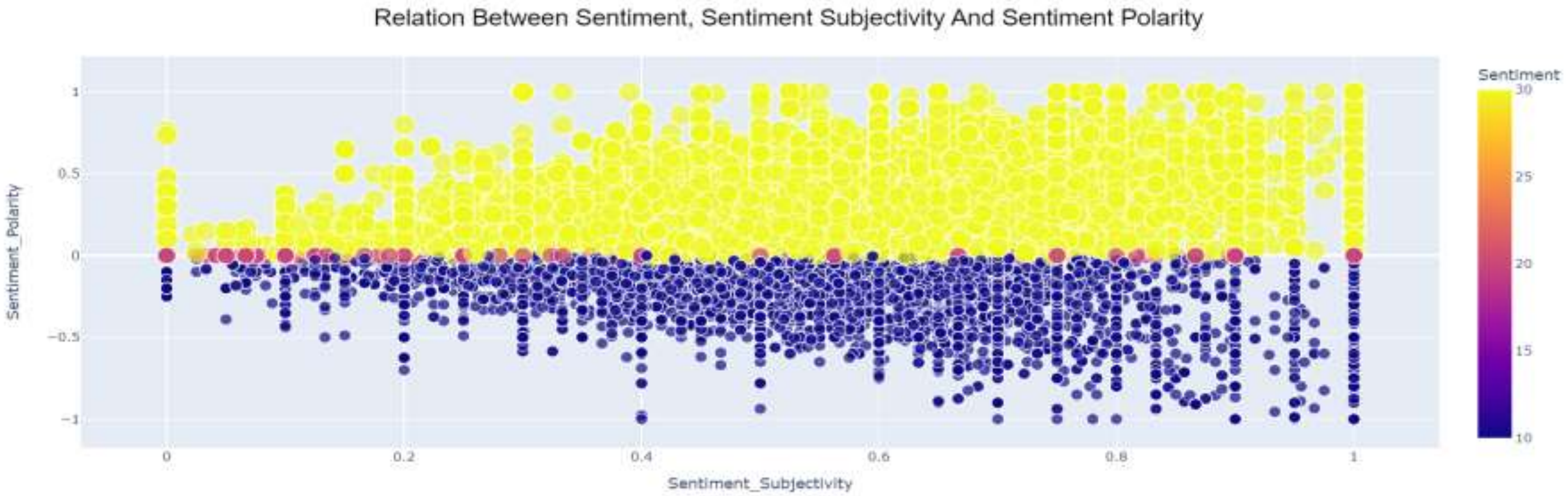


Findings

•The Sentiment Subjectivity of maximum reviews lies between 0.4 to 0.7. Average Sentiment Subjectivity is close to 0.5.

•It indicates that maximum of the reviews are subjective, close to the opinion of general public.

Relation between Sentiment, Sentiment Subjectivity, Sentiment Polarity?

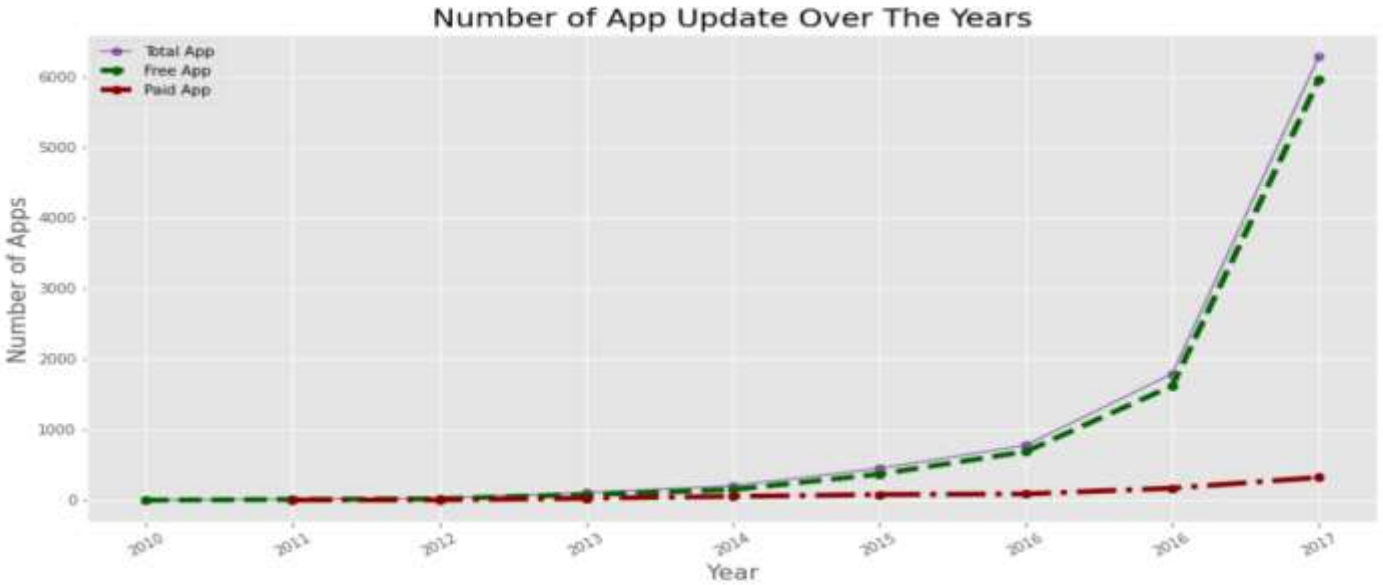


Findings

Yellow = Positive sentiment, Blue = Negative Sentiment, Red = Neutral Sentiment

•Sentiment Subjectivity And Sentiment Polarity not always proportional.
•But for positive Sentiment, with increase in Sentiment Subjectivity the Sentiment Polarity is also increase to some extent. And for negative Sentiment, with increase in Sentiment Subjectivity the Sentiment Polarity is decrease to some extent.

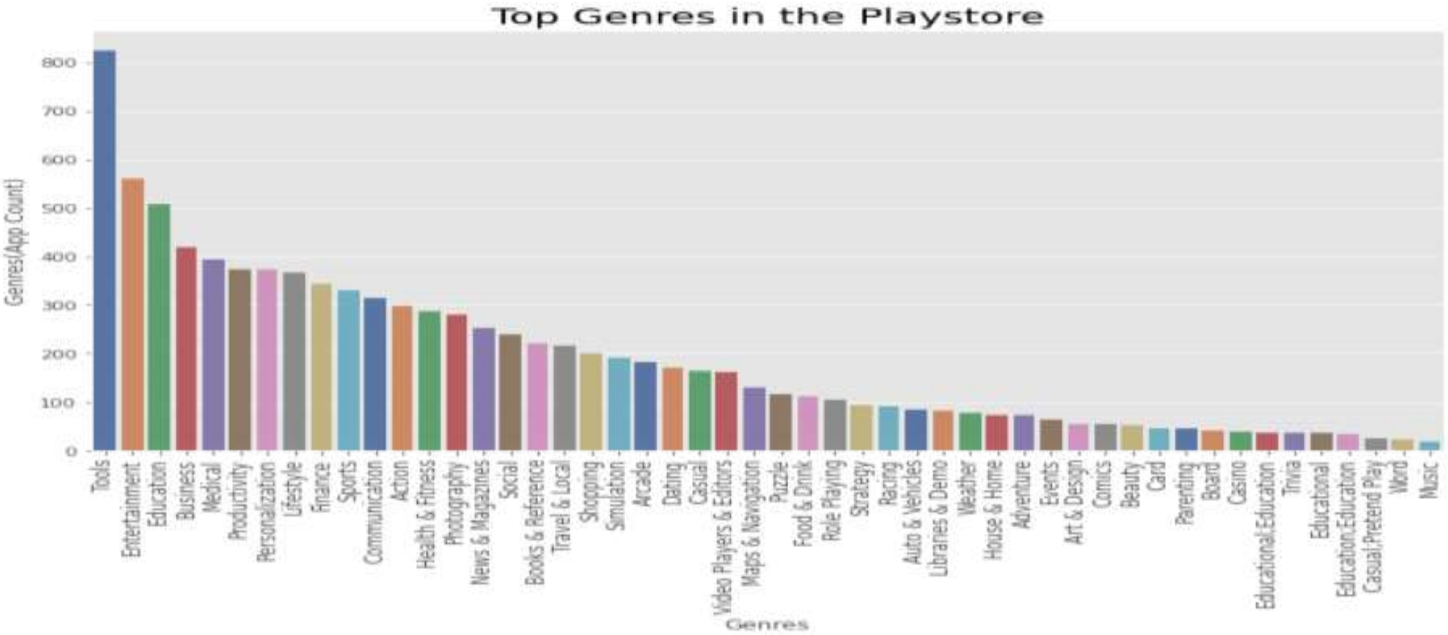
Distribution of app update over the years?



Findings

- Before 2011 there was no paid app.
- With year the total app increased rapidly, but the percentage of paid app is very less as compared to free app. Most of the apps are free.

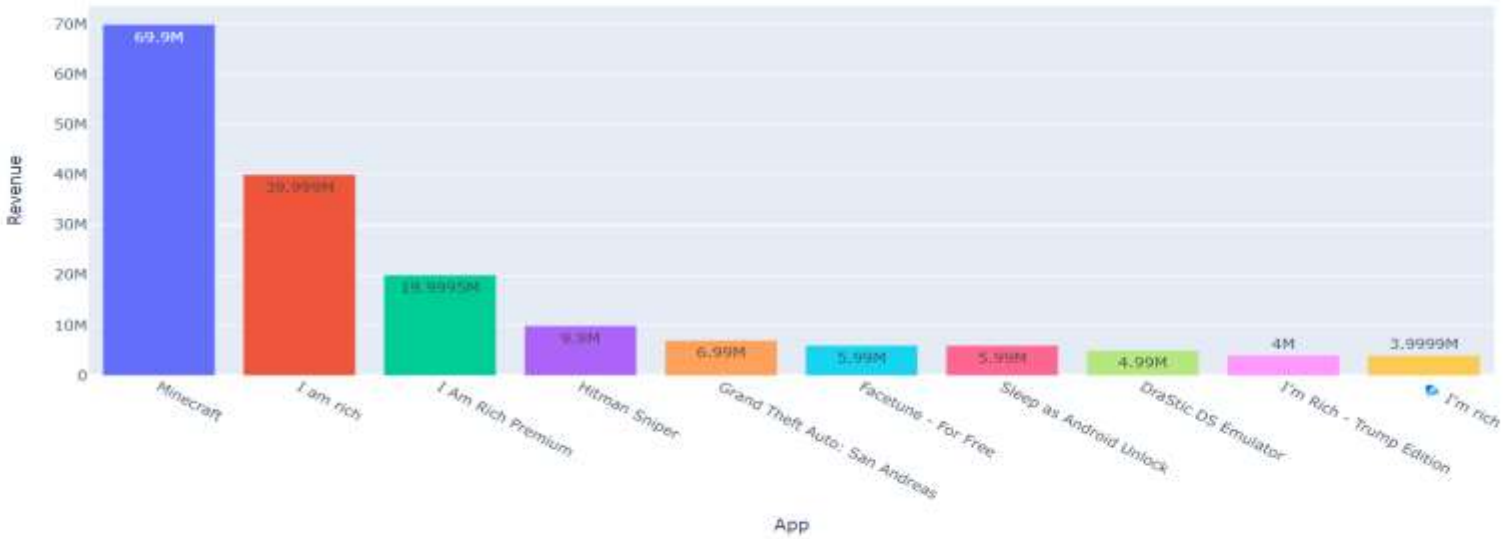
What are the count of apps in different genres?



Findings

From the above visualization, we can see that the Highest Number of Apps found in the Tools and Entertainment genres followed by Education, Business and many more.

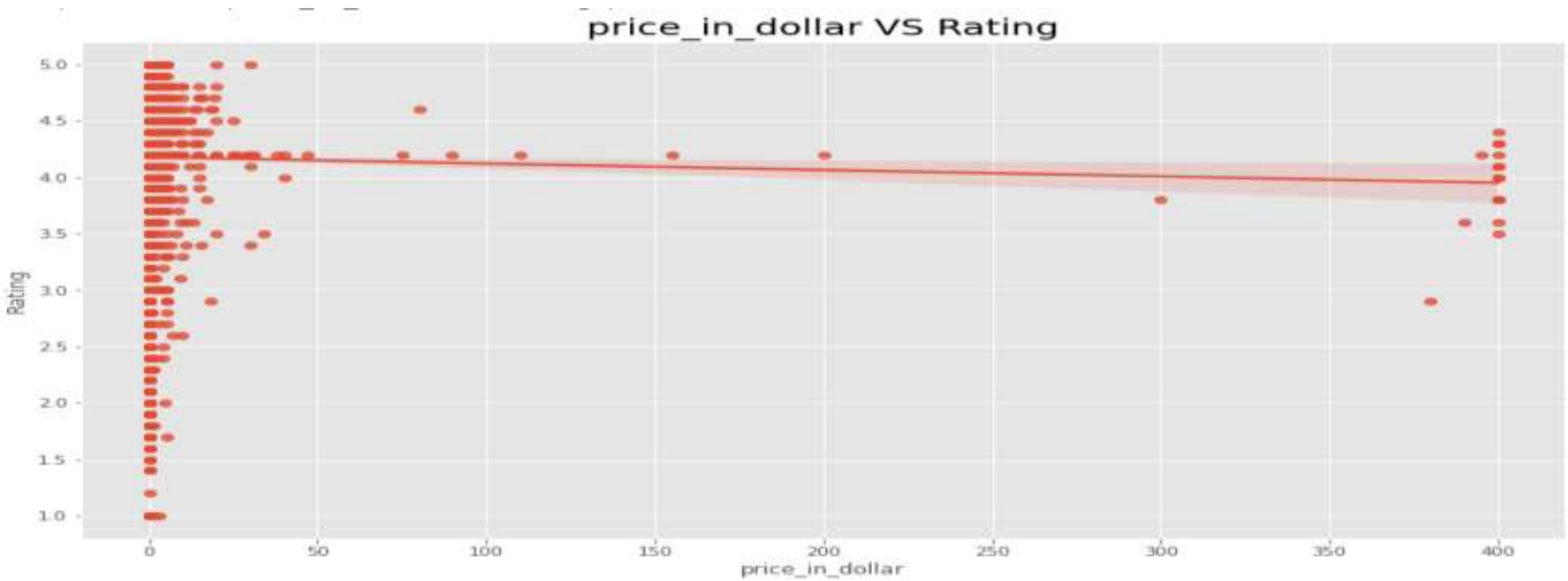
Top 10 apps in paid type by revenue?



Findings

App which has generated most revenue through download is **Minecraft** making **$69.9Millions**.

Does price of the app affect the rating?



Findings

Yes, as the price increases ratings received seems to decrease even below the average rating in the appstore.

Conclusion

In this project we analyzed more than 10,000 play store apps across different categories to get a deeper insight about the play store app market where we observed some interesting strategies for growing the app businesses.

1. Some categories like art and design, parenting, comics, beauty has very less competition hence it is easier to grow by focusing on such niches.

2. Free apps are way more in number than paid apps so it is better to focus on free app as to increase the number of installs and gain more popularity.

3. Most of the apps are small in size(less than 30MB) it is better to create apps that consume less memory.

4. Apps with content available for everyone will attract more users which increases the chance of getting more installs.

5. Apps with regular updates get more number of installs. So we should focus on updating app frequently.

Challenges faced

- Most of the features not in proper data type along with that for some of the features like size all the observations are not in same unit.
- Rating column has 1474 null values. If we drop hugh amount of observation then so we will loose so much of information. On the other hand after replacing this much amount of null values by mean or median it affects the accuracy of the analysis.
- In User review dataset except app column all the features contains 42% of null values.
- The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis, if we had atleast 70%-80% of the data available in the merged data frames.