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. To survive and grow in this competitive market we need a great strategy. We have to find some answers from previous market insights. For example- in which niche should we create our app? Should we make our app free or paid? etc.

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 Statement:**

1. 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. Let’s 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?

**Introduction:**

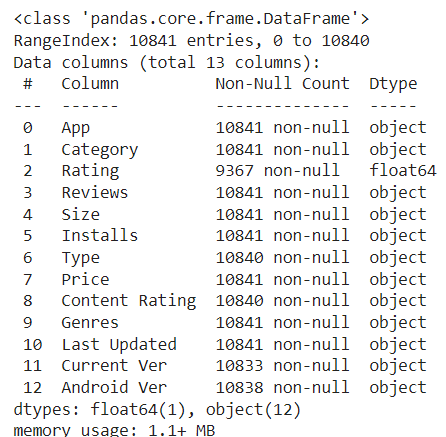
Nowadays, mobile apps are becoming one of the most fundamental parts of our life. With the rapid increase in app usage, the app market is also growing exponentially. Currently, the Google play store is the most dominant android app marketplace. At this time, it contains more than 2.5 million apps and thousands of apps are launched every single day.

In this project we are analysing more than 10,000 play store apps across different categories to get a deeper insight about the play store app market

**Overview of data:**

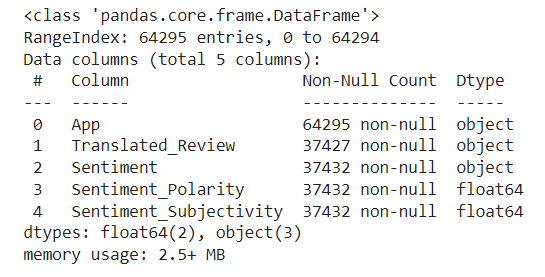
There are two datasets

1. **Play Store Data.csv**- This dataset contains 13 features. Different features indicate a different property of an app. For example- the category of the app, Rating of the app, etc.
2. **Reviews. CSV**-This data set contain 5 features. The 'Translated\_Review' feature contains the pre-processed review of each app. Based on these reviews there are three more features "Sentiment", "Sentiment\_Polarity" and "Sentiment\_Subjectivity".



The Features of **Play Store 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.



The Features of **Reviews Dataset**:

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]. Higher the subjectivity means the review is closer to the opinion of general public.

**Data Preparation and Cleaning**

**Important Steps of Data Cleaning**

• Identifying the null values.

• Identifying the invalid data.

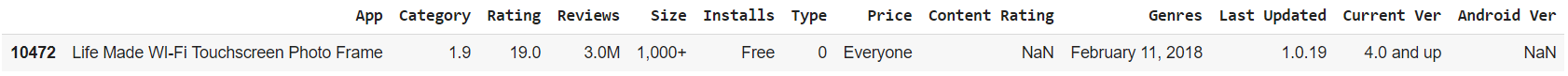
• Removing Symbols.

• Standardizing the data types.

We will begin with Play Store Dataset

Starting univariate analysis with the Reviews Column, and it is found that everything is correct, except for the **3.0M** which doesn’t seem to be a review.

We will explore it by displaying the image



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

So, we will drop this row from the dataset

Columns like Reviews, Size, Installs & Price should have an int or float datatype, but here we can see of object type, so we should convert them to their respective correct type.

* Firstly, we converted **Reviews** column to int type.
* Moving onto **Size** column, **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.
* Now we are only left with **Price** Column, **Price** column had ‘**$**’ symbol present, which was removed and then the column was converted to numeric type.

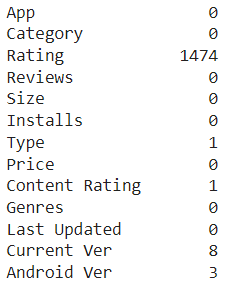
**Content\_Rating** column has relevant variables present in it and thus were retained as it is.

**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**

Null Values of Play Store Dataset



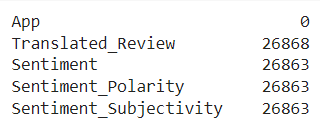
* Rating column contain 1474 NaN values. We cannot drop this much amount of row from dataset because it will lose a huge percentage of information. We replaced all the NaN values with the average of non-null values.
* Type column contain Null value at index 9148 so it was removed.
* 8 null values were 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**

### In the Dataframe there is only one column that contain **UNIQUE** value which cannot be repeated and that column is **App** column. A total of 798duplicates were present in the **App**. So, all the duplicate values were removed from the data set.

After removing "Duplicates" and "null values" from the Dataframe we now have a modified Dataframe with 9648 rows and 15 columns. Earlier it had 10841 rows.

**Handling Null values in User data review Dataframe**



* 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. Earlier it had 64294 rows and 5 columns.

Now finally we are done with this section

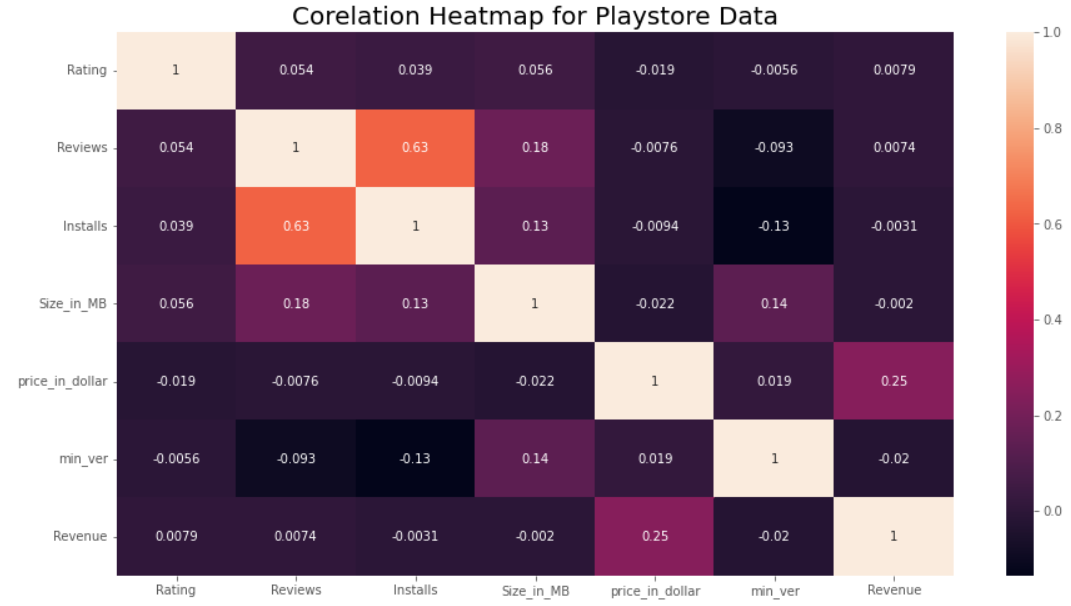
# **Exploratory Analysis and Visualization**

Exploratory data analysis is an approach to analyzing data sets, getting some interesting insights from it, and presenting them in the form of visual methods.

We will be importing matplotlib.pyplot and seaborn for visualizing the data.

Now we will explore our dataset, i.e., to get an insight, and learn the trend, pattern and get answers to some of the questions related to the dataset.

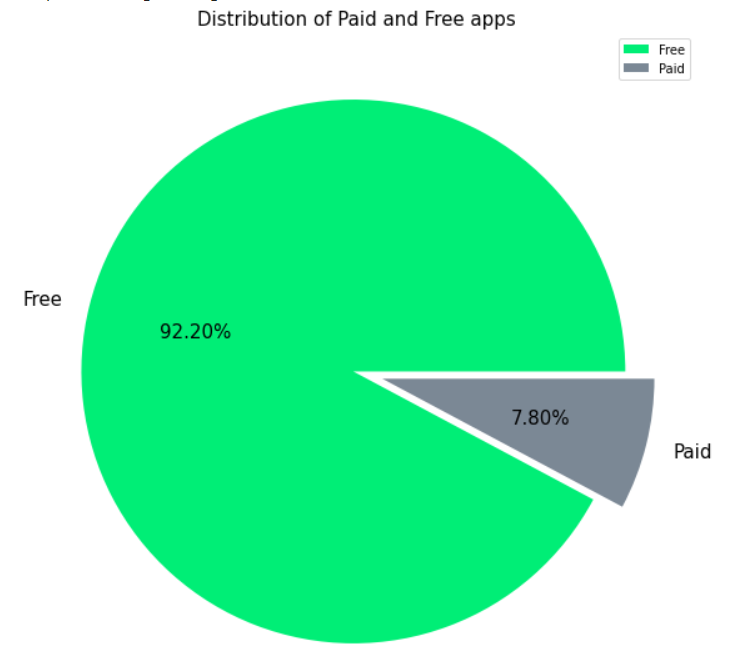
1. **What is the Corelation between different variables?**

****

**Observation**:

* 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.
* The Rating is slightly positively correlated with the Installs and Reviews` column. This indicates that as the average user rating increases, the app installs and number of reviews also increase.

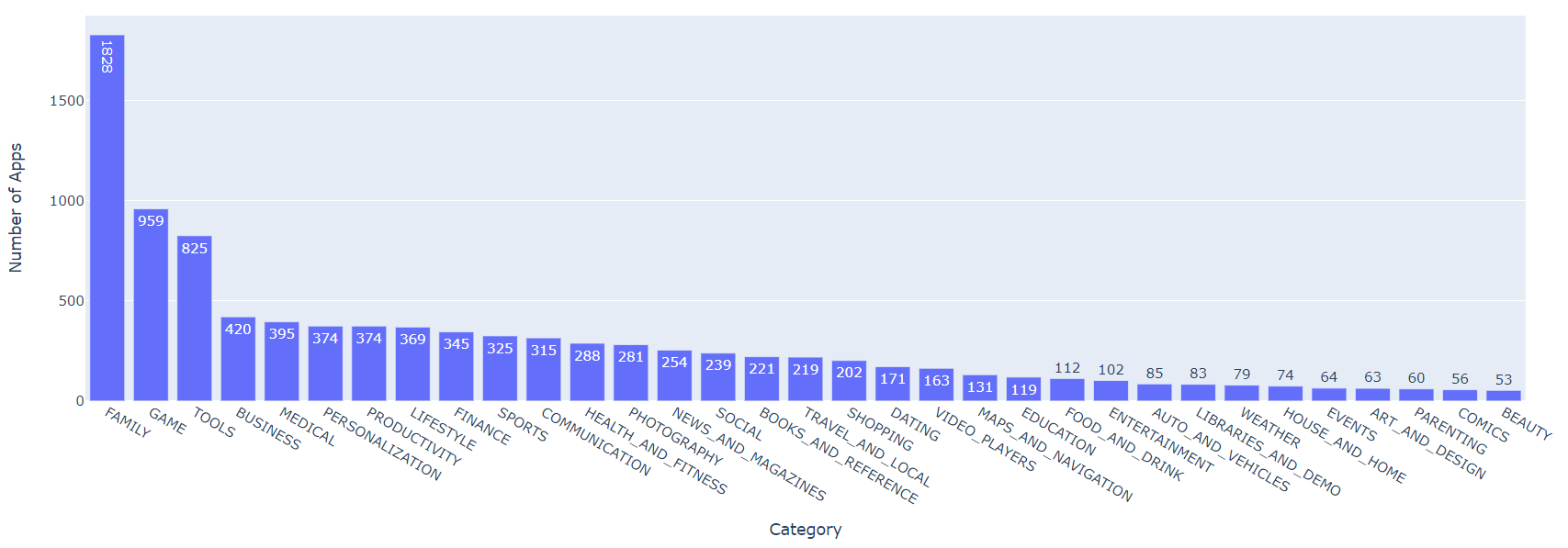
1. **What is the ratio between paid app and free app?**



**Observation:**

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

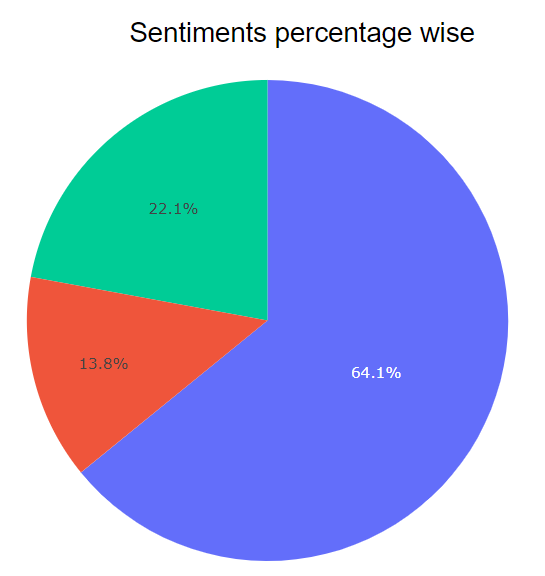
1. **Total Number of apps in each category?**

****

**Observation:**

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

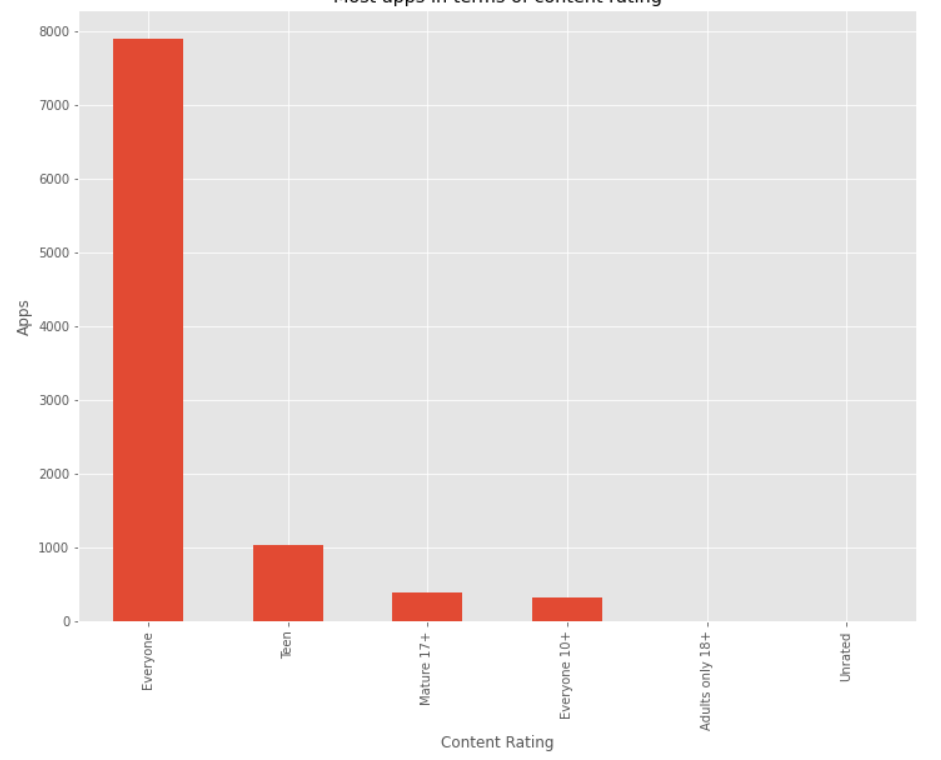
**Beauty** and **Comics** category has least number of apps.

1. **What is the percentage of review sentiments?**



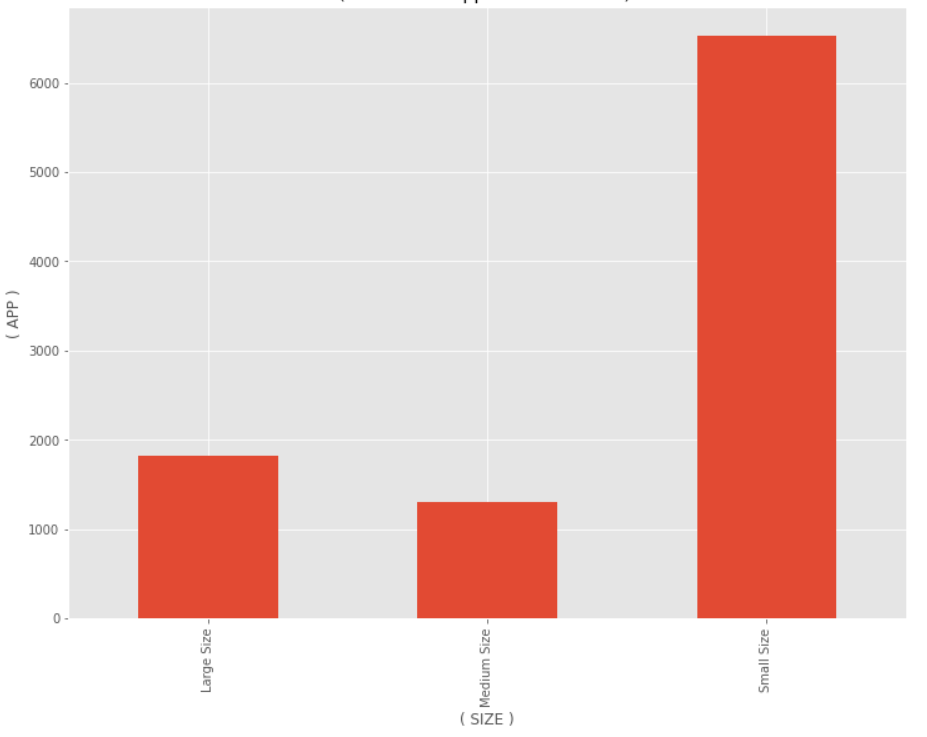
**Observation:**

### 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.

1. **Most apps in terms of content rating?**

**Observation:**

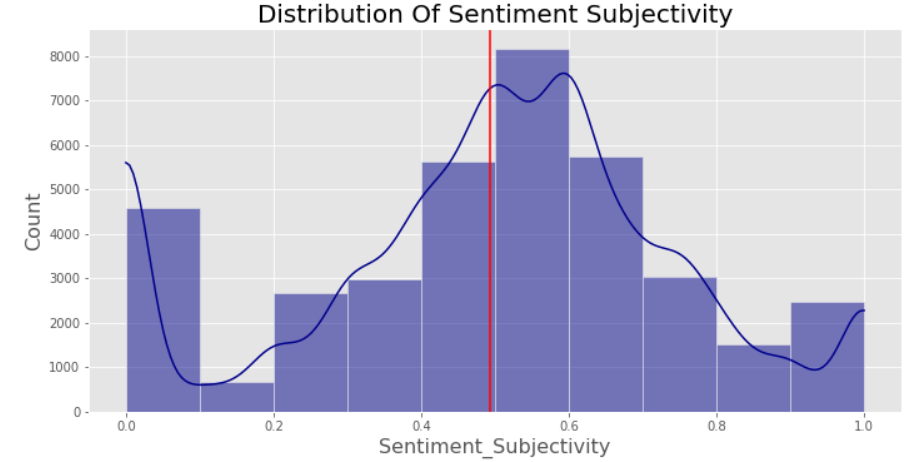
Apps with content rating available for everyone attract a greater number of users.

1. **Number of apps based on size?**

**Observation:**

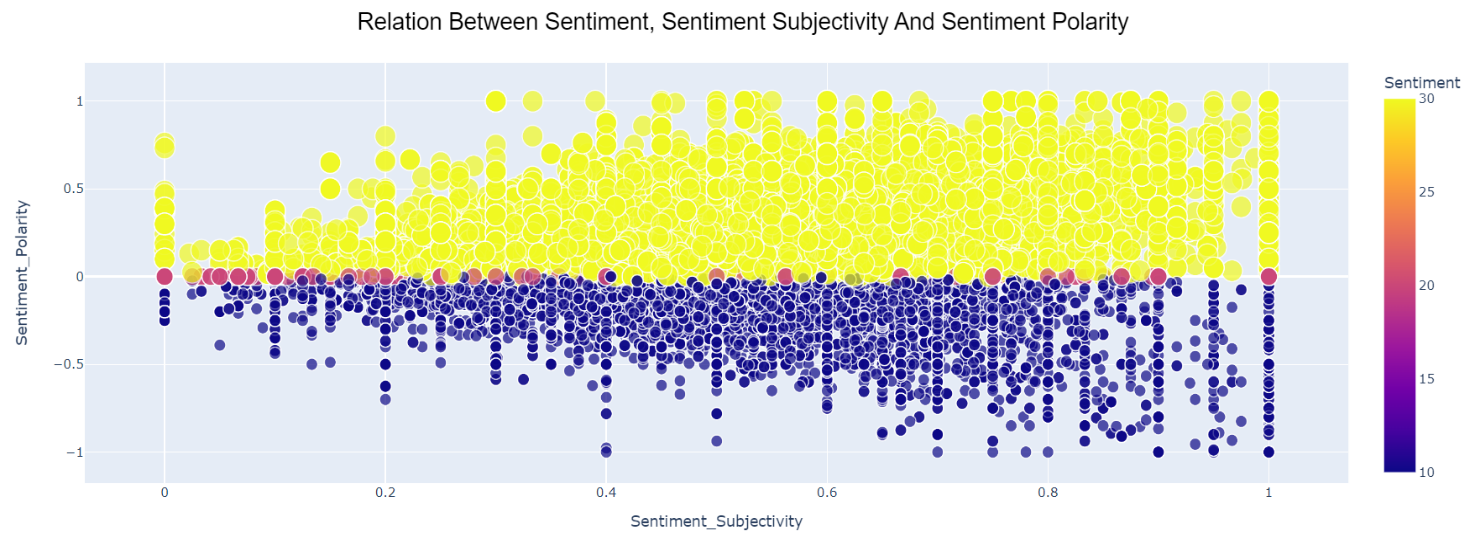
Small size <=30MB, 30MB<Medium size<=60MB, Large size>60MB

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

1. **Let’s us discuss the sentiment subjectivity.**

**Observation:**

* 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 is subjective, close to the opinion of general public.

1. **Relation between Sentiment, Sentiment Subjectivity, Sentiment Polarity?**

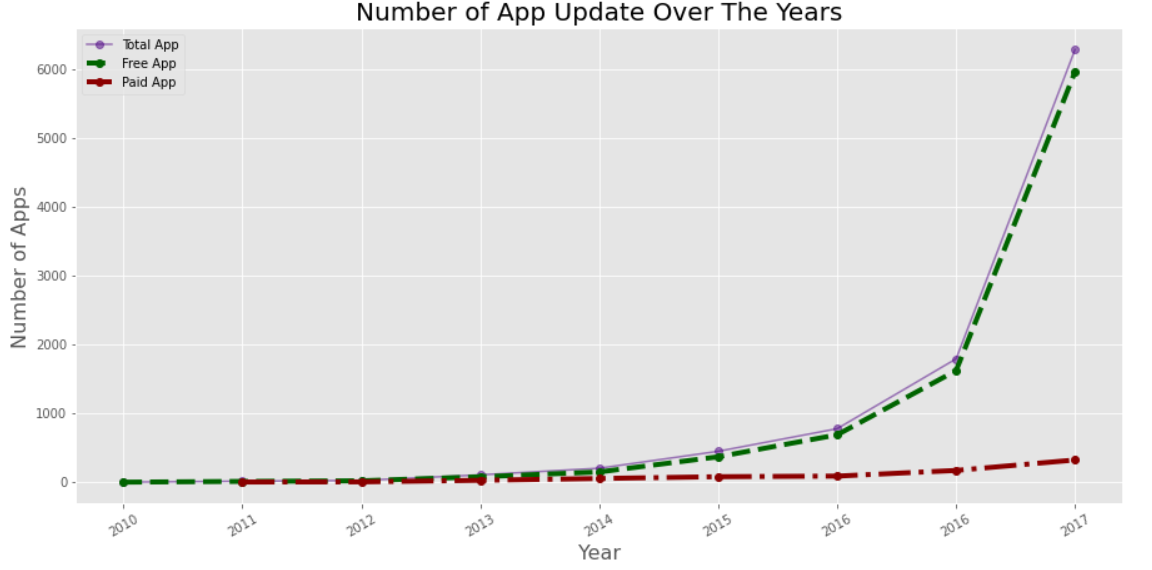
**Notation:**

* "Yellow marker" denotes "positive sentiments"
* "Blue marker" denotes "negative sentiments"
* "Red marker" denotes "neutral sentiments"

**Observation:**

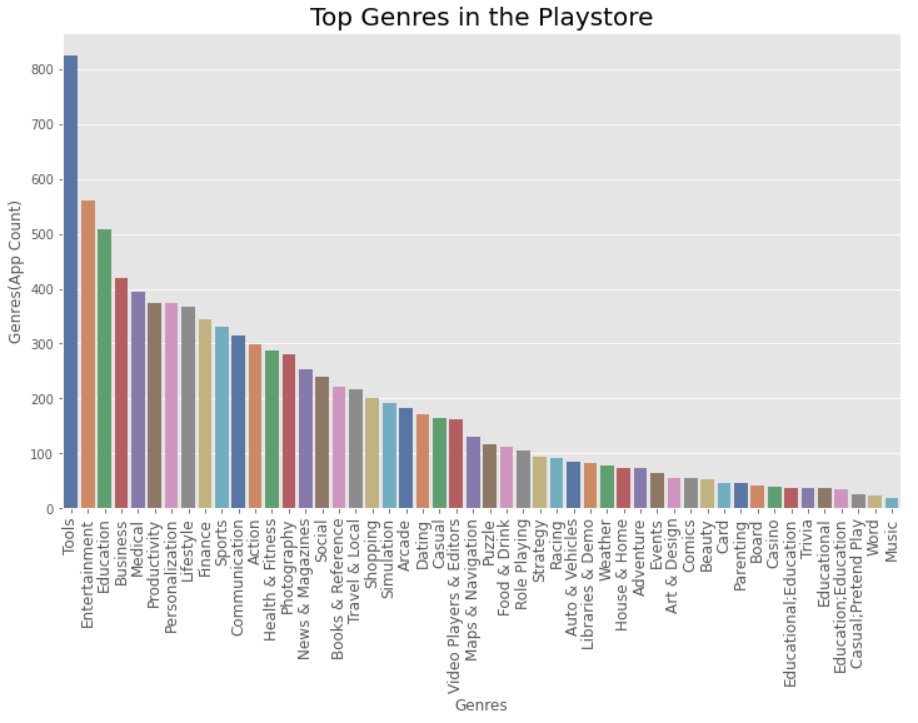
* Sentiment Subjectivity and Sentiment Polarity not always proportional.
* But for positive Sentiment, with increase in Sentiment Subjectivity the Sentiment Polarity is also increase to some extent. And for negative Sentiment, with increase in Sentiment Subjectivity the Sentiment Polarity is decrease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

1. **Distribution of app update over the years?**

****

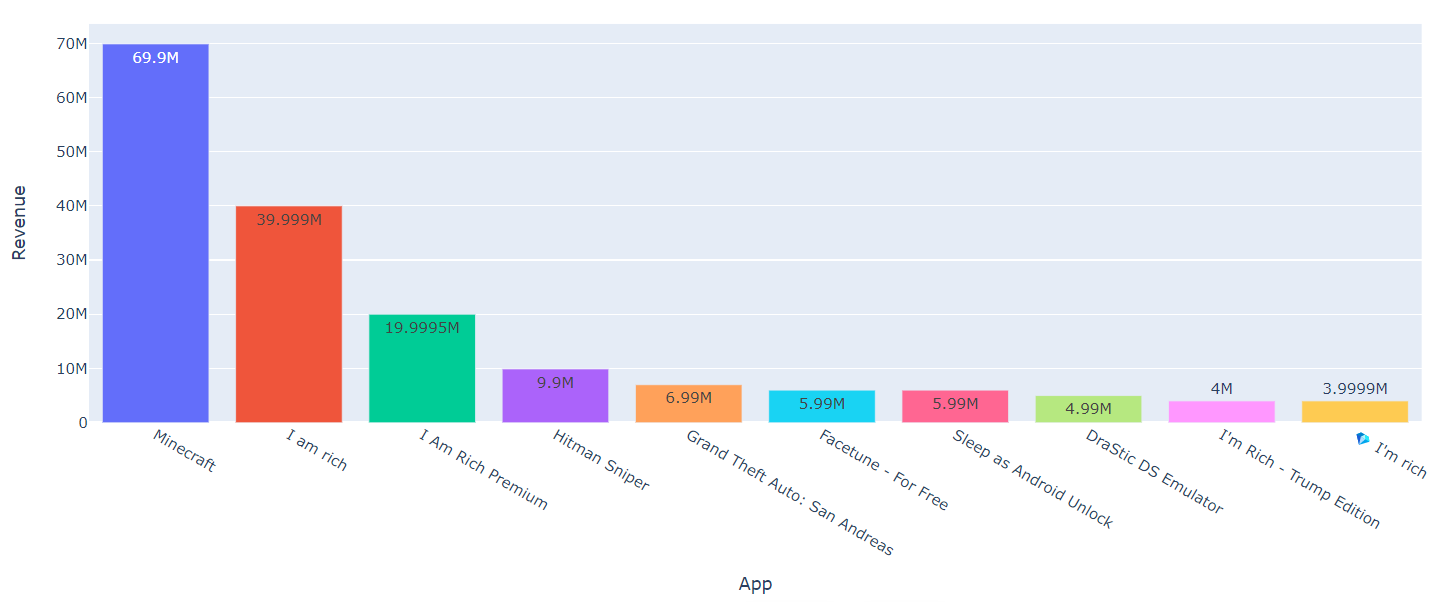
**Observation:**

* 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.

1. ** What is the count of apps in different genres?**

**Observation:**

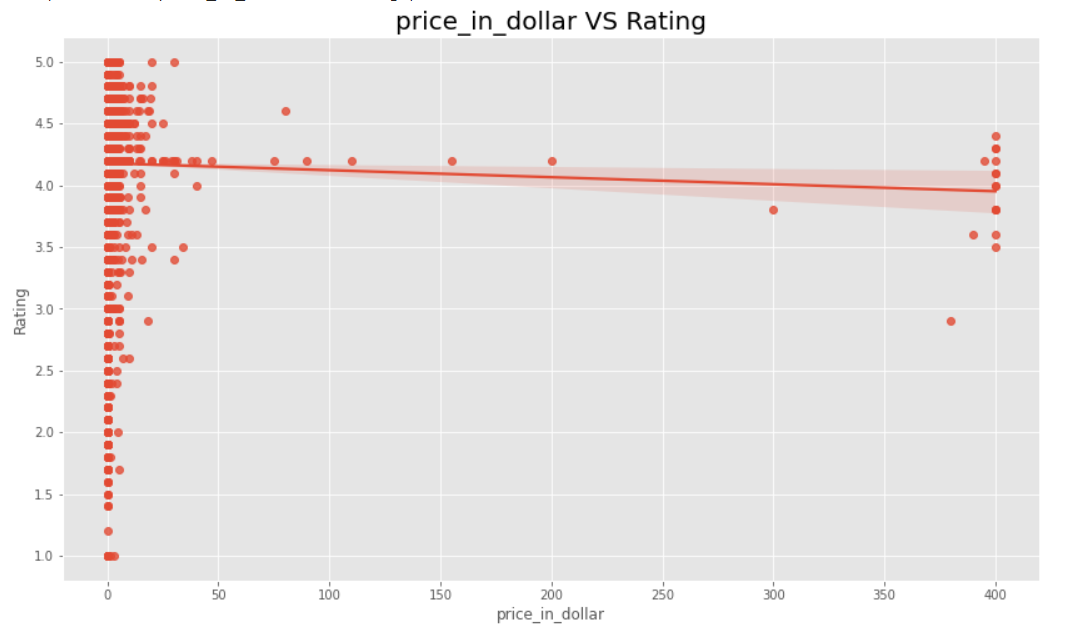
From the above visualization, we can see that the highest number of apps found in the Tools and Entertainment genres followed by Education, Business and many more.

1. ** Top 10 apps in paid type by revenue?**

**Observation:**

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

1. **Does price of the app affect the rating?**

****

**Observation:**

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.