## **EDA Capstone Project on Play Store Review Analysis**

**Monica Patel**

**(Data science Enthusiast at Alma Better Pro Member Bangalore)**

**Abstract**:

The Project “Play Store App and Review Analysis” is based on discovering key understandings about the different categories of apps available on the Google Play Store, the installation per category, the count of reviews per app, and the sentimental of the reviews per app.

In this analysis, we’re provided with a dataset with some records. We did a basic inspection, and data cleaning to remove suspicious data and to avoid an error. We’ve used different plots to visualize our analysis in the easiest way. This can be used to know which type or category of app is installed the most, count of the applications available per category, distribution of the rating how much of these apps are paid or are available for free, to know the sentiment of the reviews, the polarity of the sentiment.

***Keywords: EDA, Google Play Store, Data visualization, and exploration.***

**1.Introduction**

The Google Play Store started life as the “Android Market” in 2008. It launched alongside the very first Android devices, and its purpose was to distribute apps and games.The Android Market was extremely basic at the beginning. It didn’t support paid apps and games until 2009. However, as the Android platform grew, so did the Android Market. By 2012, it featured over 450,000 Android apps and games.

By this time, Google’s ecosystem had expanded greatly compared to the humble beginnings of the Android Market. In fact, the Android Market was just one of the company’s online markets. At the time, this was the only place that Google had to sell goods. As the company’s hardware efforts grew, it was time for a new store. Posting

1. Problem Statement

Google Play Store is one of the most used applications on any android phone as it allows the user to get any app that he or she requires. It has many categories to choose the type of application that the user wants to install on his device. Doing a data analysis on the data of the google play store along with the user review dataset will help us to get to know the most dominant category among all the available ones. It also allows us to track the number of installs per category and the most popular application among people of different age groups, the polarity of the sentiment of the reviews given to each application. It can also help us to fetch various information about the different applications like the size, the rating, the version available, and the price of the applications if a paid version of them is available. All this information can help us to fetch meaningful insights from the dataset provided.

The data provided to us have two datasets one about the application information and the second one telling us the reviews on each application available. After combing the two datasets together the number of observations went over 7 lakhs which is distributed among 17 columns and it is a mix of categorical and numeric values. Explore and analyze the data to discover key understandings.

The Problem statements are

1. What are the top categories on Play Store?
2. Are majority of the apps Paid or Free?
3. How importance is the rating of the application?
4. Which categories from the audience should the app be based on?
5. Which category has the most no. of installations?
6. How does the last update has an effect on the rating?
7. How does the count of apps varies by Genres?
8. How are ratings affected when the app is a paid one?

3.EDA on given Data set

There are two dataset:

1. Play Store Data(App, Category, Rating,Review,Size,Install,Type,current rating ,genres , Last update,Current Var ,Android Var)
2. User Review Data(App, Sentiment ,Sentiment Polarity, Sentiment Subjectivity)

By diagnosing the data frame, we know that:

* There are 13 columns of properties with 10841 rows of data.
* Column 'Reviews', 'Size', 'Installs' and 'Price' are in the type of 'object'
* Values of column 'Size' are strings representing size in 'M' as Megabytes, 'k' as kilobytes and also 'Varies with devices'.
* Values of column 'Installs' are strings representing install amount with symbols such as ',' and '+'.
* Values of column 'Price' are strings representing price with symbol '$'.

. This EDA project aims to discover patterns that lead to a successful application on the Google Play Store. This will be done by analyzing the historical data collected from the Google Play Store according to the dataset source.

Google Play-Store Dataset columns description:-

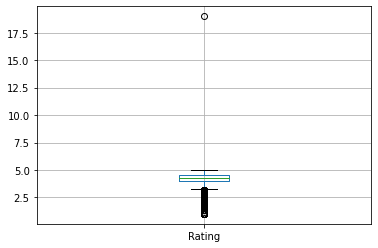
* App: Application name
* Category: Category the app belongs to
* Rating: Overall user rating of the app (as when scraped)
* Reviews: Number of user reviews for the app (as when scraped)
* Size: Size of the app (as when scraped)
* Installs: Number of user downloads/installs for the app (as when scraped)
* Type: Paid or Free
* Price: Price of the app (as when scraped)
* Content: Rating Age group the app is targeted at - Children / Mature 21+ / Adult
* Genres: An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, or Family genres.
* Last Updated: Date when the app was last updated on Play Store (as when scraped)
* Current Ver: Current version of the app is available on the Play Store (as when scraped)
* Android Ver: Min required Android version (as when scraped)

User Review Dataset column description:-

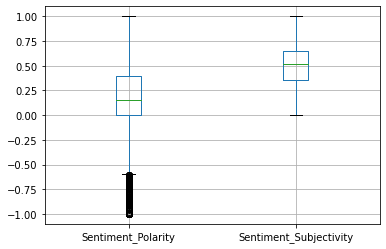
* App: Application name
* Translate Reviews: Reviews on various applications given by the user
* Sentiment: Positive, Negative, or Neutral
* Sentiment Polarity: Tell us about the polarity of the sentiment of the translated reviews that we received.
* Sentiment Subjectivity: Tell us about the subjectivity of the various reviews according to the translated review.
* Dataset shape of Playstore –(10841 Rows and 13 columns)
* Dataset shape of user reviews –(64295 Rows and 5 columns)

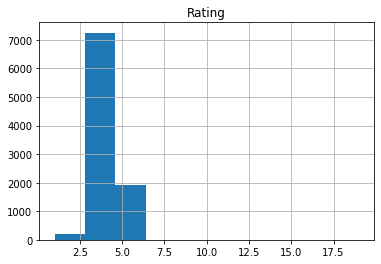
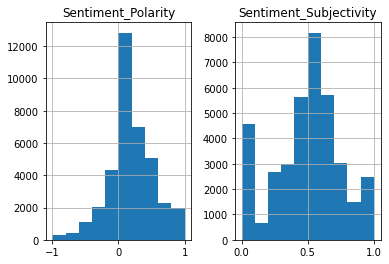
**DATA VISUALIZATION**

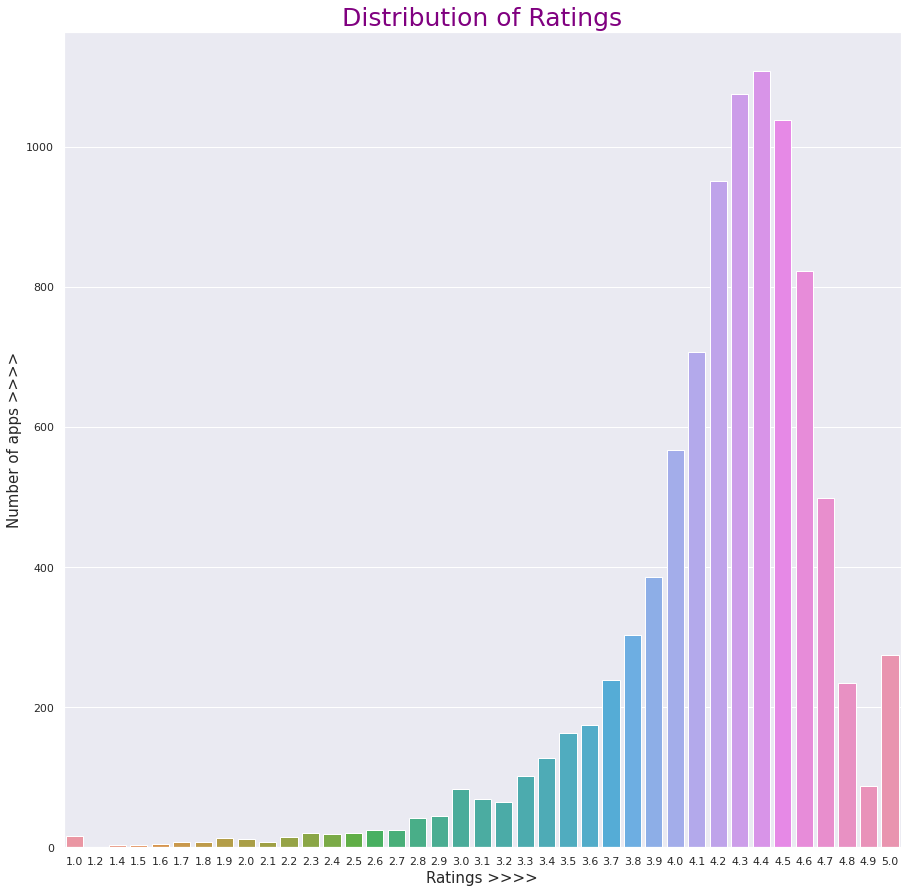
1. **Play store outliers-This is the outlier of playstore which has shown**
2. **Above the 17.5 so we clean the data and the**
3. **Result is below .**
4. **After cleaning the outlier been removed .**
5. **playstore\_df.drop([10472], inplace=True)**

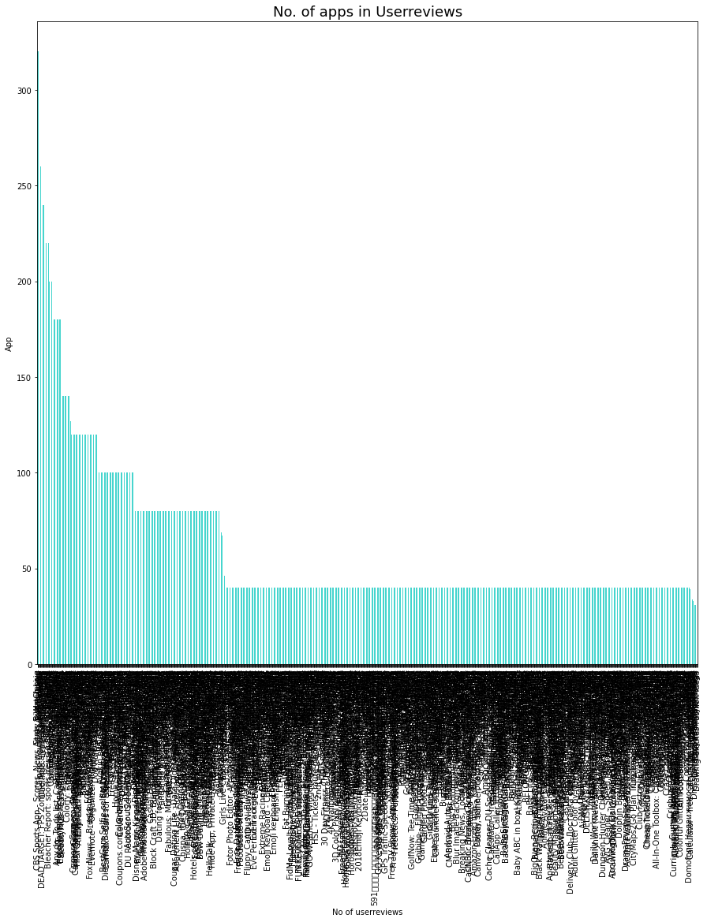


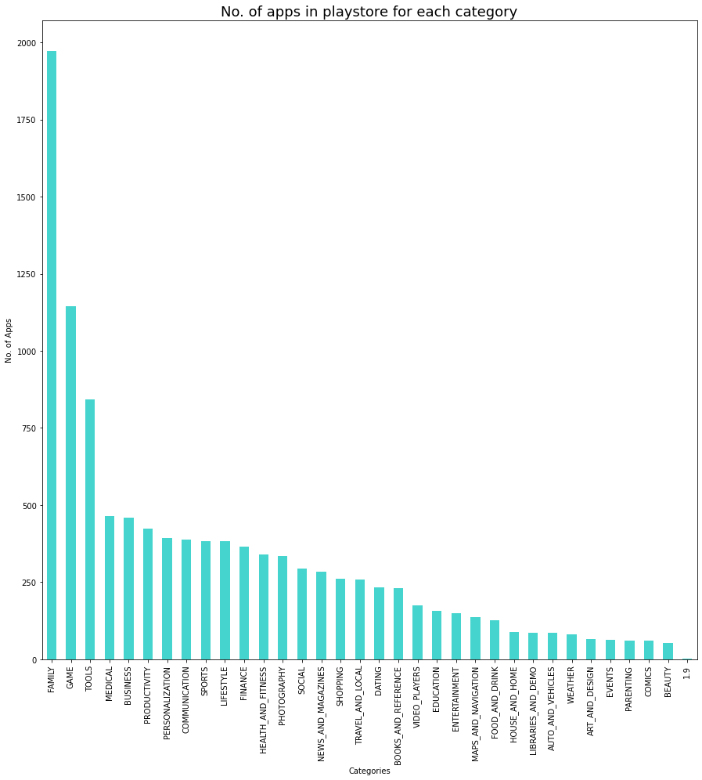
1. User reviews outliers-Positive and Negative  
   sentiments are more and  
   Neutral are least



1. 
2. 

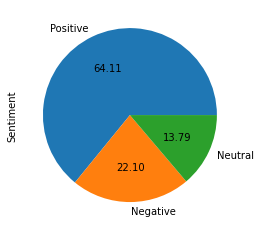


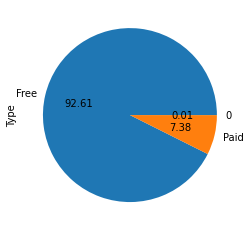




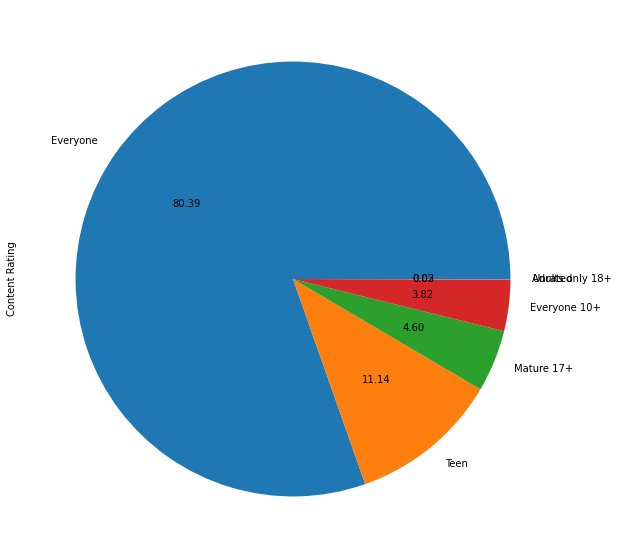


* POSITIVE SENTIMENT – 64.11
* NEGATIVE -22.10
* NETUTRAL -13.79
* Sentiment analysis is a technique with which we can identify and determine if data indicates a positive, negative or neutral It aids in understanding customer feedback and is mostly used by several companies to analyse brand and product reviews. It also helps in finding out the underlying sentiment in a text.

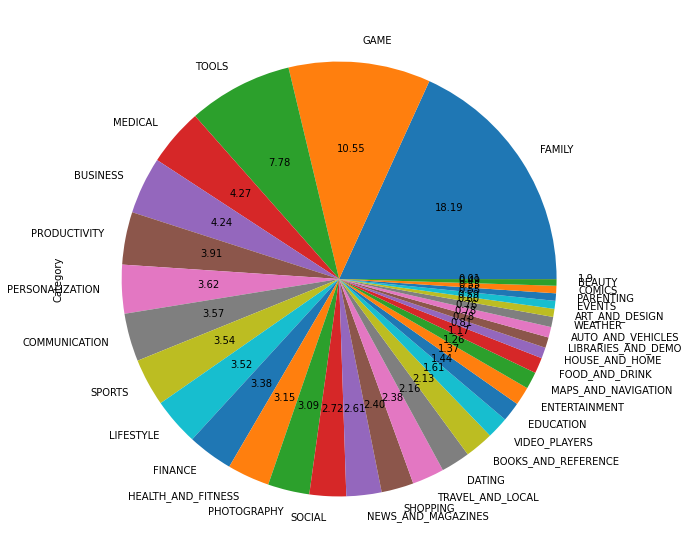




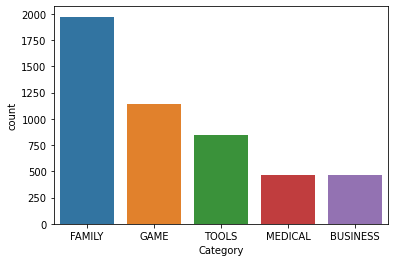
* Free apps are ruling the market by huge percentage that is 92.6%
* Percentage of paid apps present in Play store is 7.4% which is very low.
* 80.39 is highly used by everyone.
* 11.14 is for teen
* 4.60 mature person used which are 17+
* 3.82 use by everyone who are 10+
* 0.02 is for adult.

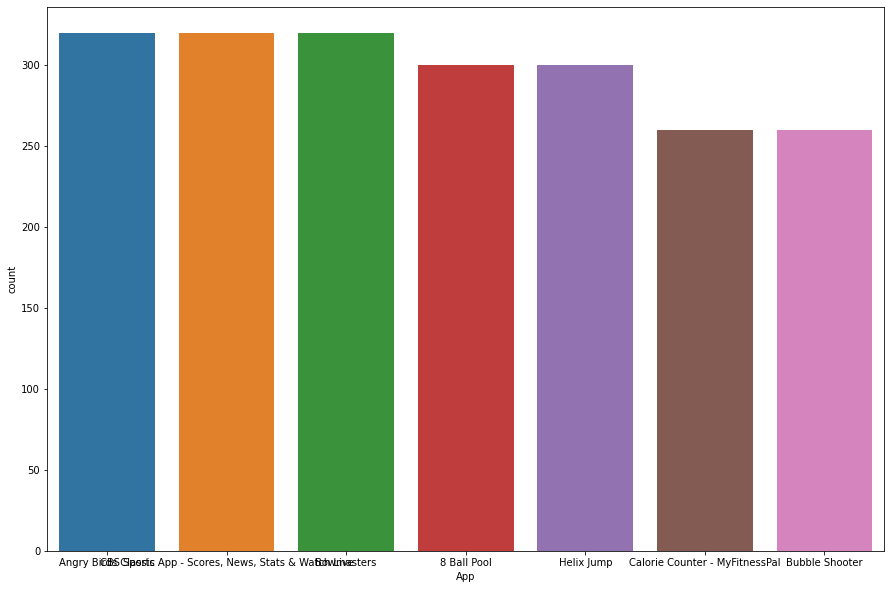


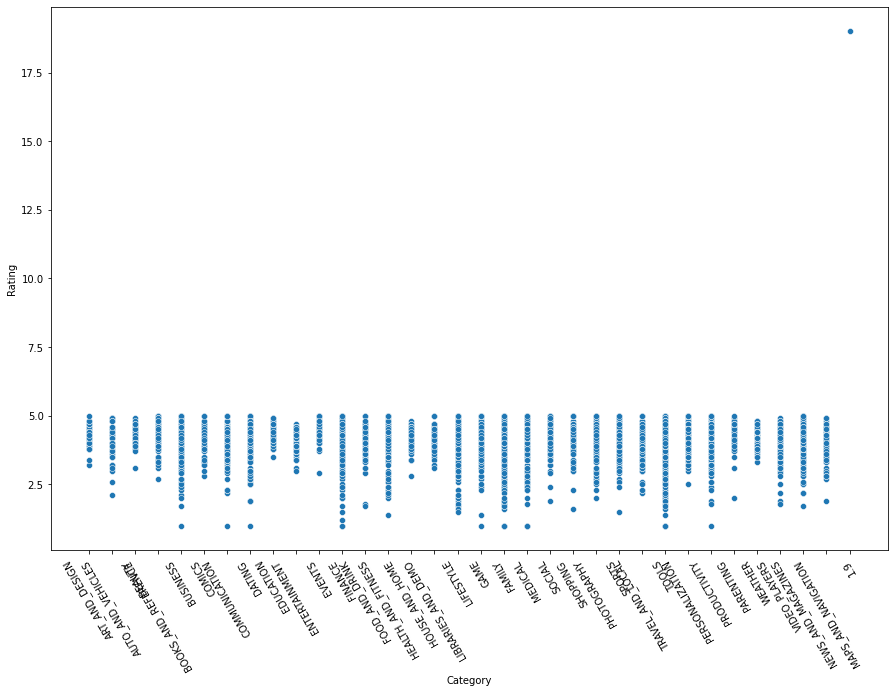
Family’ category has highest number of apps in play store i.e. 1971 apps (18.2%).   
The category which has second highest number of apps present in play store is ‘Game’ and number is 1144apps. Lowest number of apps present in play store is ‘Beauty’ and the number is 53 apps.

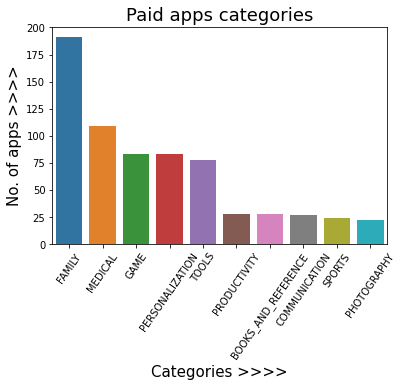


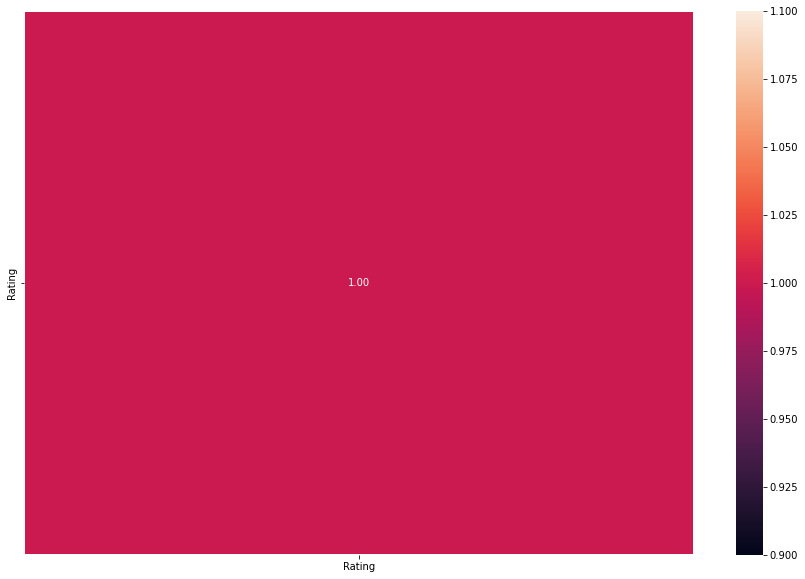
3 category from top 5 categories, Medical, Games  
and Tools are of specific  
micro-niche



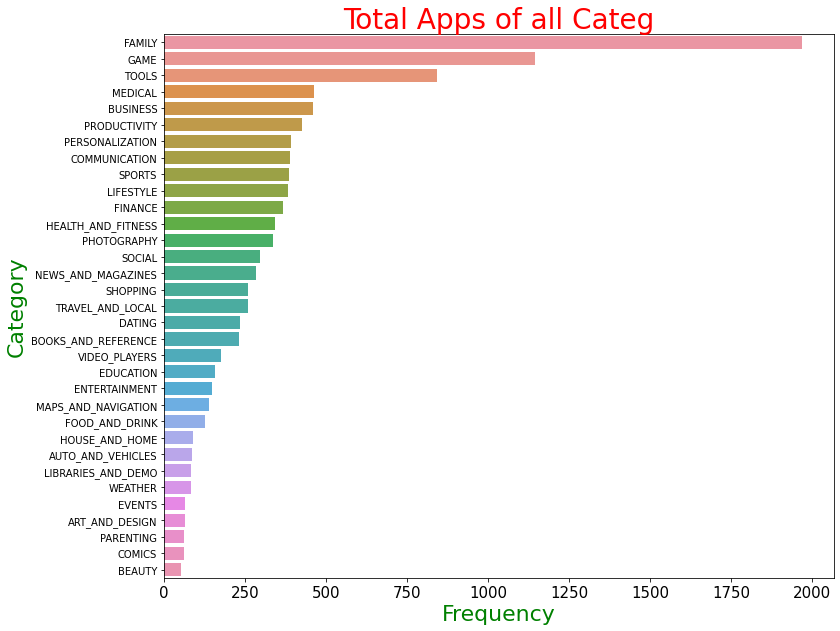




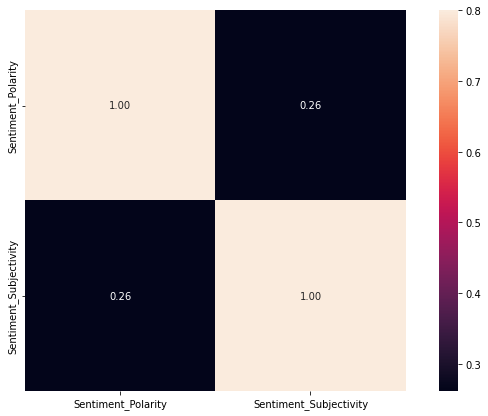


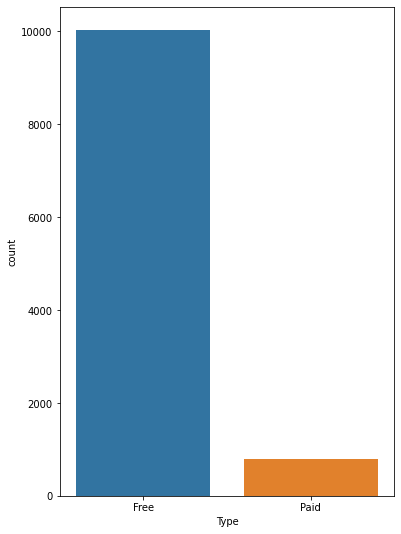


* Apps with 1,000,000 Installs  
  ➢ Game and Family category has most number of  
  apps with 1000000+ installs  
  ➢ ‘Photography’ category also has higher  
  downloads but it has low apps compare to  
  other top categories



* Correlation between Sentiment polarity and sentiment
* subjectivity is
* 0.26 and 1.00 are positively as well as strong correlated
* with each other.





* Size of Free and Paid Apps  
  The size of paid apps are  
  comparatively lower then Size of free  
  apps, it may because of paid apps  
  developers charging for efficiency of  
  apps and free app developers wants  
  to cut the cost of development by  
  neglecting size efficiency.

CONCLUSIONS  
After undergoing these algorithms and process, we concluded that our hypothesis is true. Meaning you can predict the app ratings, however significant preprocessing must be done before you start the classification and regression processes.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! This shows that given the

Size, Type, Price, Content Rating, and Genre of an app,  
we can predict about 92% accuracy if an app will have  
more than 100,000 installs and be a hit on the Google  
Play Store

* As we can see applications that have the highest number of install comes from the communication, photography, and social category.
* It can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in a maximum number of the case, shows a proportional behavior, when variance is too high or low
* Most successful category based on the number of installs is the communication category.
* So 64,65% of the reviews come under positive, 23.18% are negative while the rest comes under neutral
* It can be seen that a maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this, we can conclude that a maximum number of users give reviews to the applications, according to their experience.
* As we can see applications that have the highest price come from the social, tools, medical, and personalization categories.
* The Category that has the highest number of reviews are social, games, photography, family, and communication.
* As we can see from the above plots: The maximum number of apps present in the google play store comes under Tools, Entertainment, and Education Genres but as per the installation and requirement in the market plot, the scenario is not the same. Maximum installed apps come under Communication, Tools, and Productivity Genres.

References-

* GeeksforGeeks
* Stack Overflow
* Medium
* kaggle