### Play Store App Review Analysis

**KULDEEP SINGH, Data Science Trainee,**

**AlmaBetter, Bangalore**

# **Abstract**

The google play store is one of the largest and most popular Android app stores. It has an enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.

We have used two raw data sets from Google Play Store. First data set contains 13 different features like category, rating, size, and more. Another dataset contains customer reviews of the android apps. These datasets can be used for predicting whether an app will be successful or not. This data set is scraped from the Google Play Store. This journal gives detailed information on feature extraction and the complete Data visualization done on this data set.

**1. Problem Statement**

### 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.

### One dataset contains each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.

### Our goal is to explore and analyze the data to discover key factors responsible for app engagement and success.

**2.Analysis Methodology**

Our analysis approach is divided into three phases: knowing data, data cleaning, and data analysis and visualization. In the ﬁrst step, we have two raw datasets given by the AlmaBetter, we looked at the structure and the manner in which the data is organized. Then we did basic data cleaning. After that we performed data analysis and visualization

**2.1. Knowing Data**

### In the first phase, We had to check and see if the datasets are ready for performing several exploration operations or not and also had to look at the structure and the manner in which the data are organized.Here , we have two different datasets having data related to google play store apps. We have used pandas library’s attributes and methods that are:

* **shape**: It gives a number of rows and columns in a tuple.
* **columns**: It returns column labels.
* **nfo()**: It gives a concise summary of a DataFrame.
* **head()**: Returns the first n rows.
* **tail()**: Returns the last n rows

**Data Summary of first dataset**

### In the first dataset each app (row) has values for category, rating, size, and more.

* **shape** : There are 10841 rows and 13 columns.

**columns:**  
1. App : Names of android apps.

2. Category : Category of apps.

3. Ratings : average rating given by the user.

4. Reviews : Number of Reviews.

5. Size : size of app in bytes.

6. Installs : number of installs or downloads.

7. Type : Free or Paid.

8. Price : price of app.

9. Content Rating : suitable age group.

10. Genres : genres of app.

11. Last Updated : last updated date.

12. Current Ver : current version available.

13. Android Ver : suitable android versions.

**Data summary of second dataset**

This dataset contains customer reviews of the android apps.

* **Shape** : There are 64295 rows and 5 columns

**columns** :  
1. App : Name of app

2. Translated\_Review : Reviews given by users.

3. Sentiment : Positive, negative or Neutral opinion.

4. Sentiment\_Polarity : Polarity is a float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.

5. 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 [0,1]

****

### **2.2. Data Cleaning**

### Dataset can contain missing data, numerical string value, various cues. If we can clean them, It will make our analyzing process easy. Data Cleaning was the second phase of our analysis approach. In this phase, We needed to fix or remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

**Dealing With NaN values**

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. In Pandas, usually missing values are represented by NaN. Dealing with NaN values or we can say missing values is also part of data cleaning. Methods we used to deal to with NaN values are:-

* **Replacing NaN with Mean, Median or Mode:** This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We calculated the mean, median or mode of the feature and replaced it with the missing values.

## **Assigning An Unique Category:** A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. Like we assigned Unknown genre for rows having missing genre value.

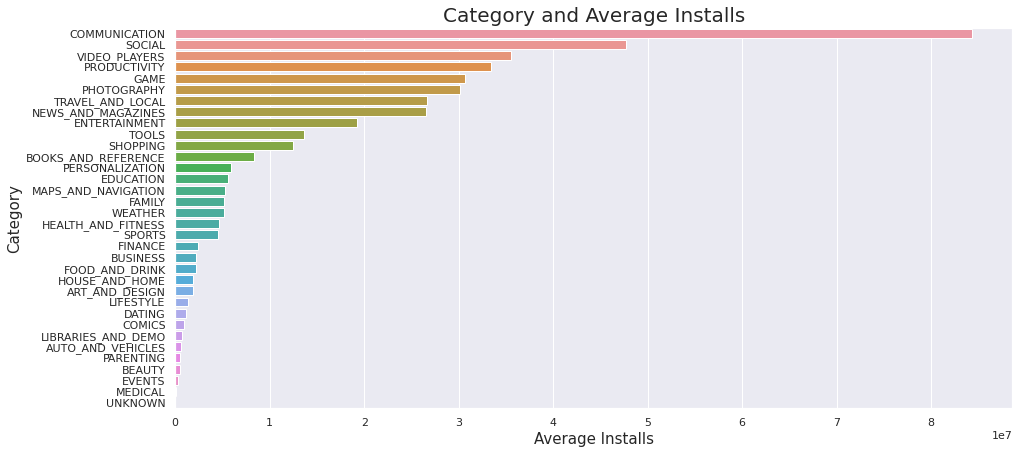
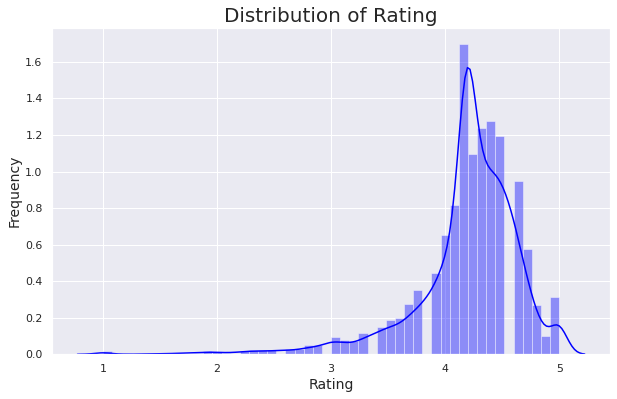
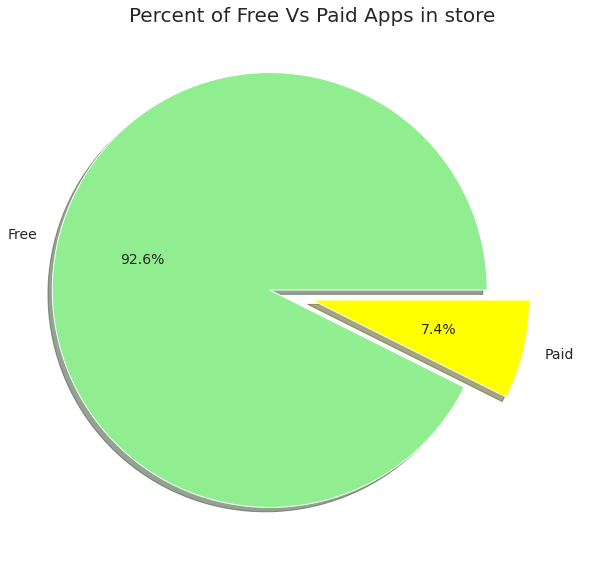
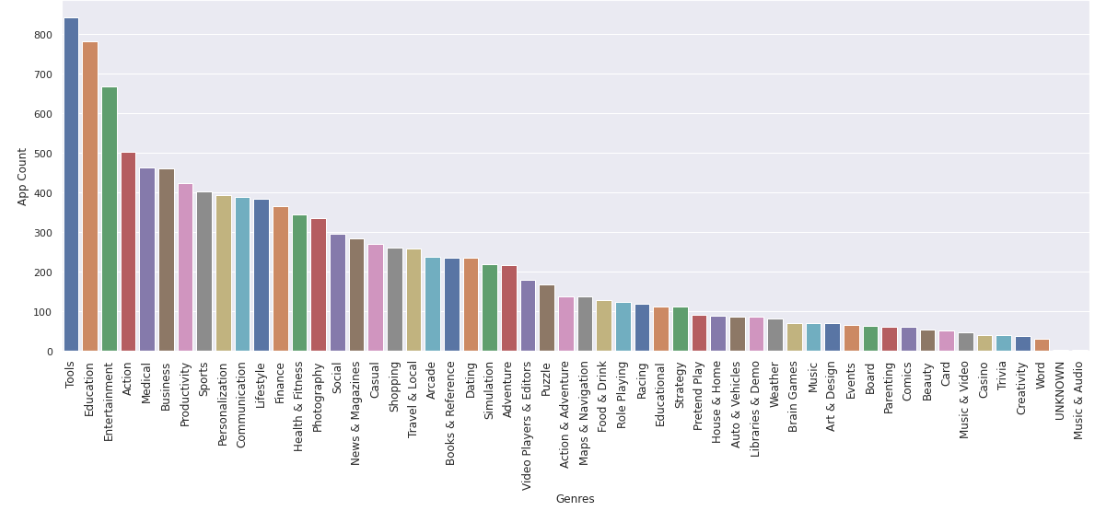
## **Predicting The Missing Values:** We had predicted the missing values for some variables like for ‘Type’we assigned ‘Free’ for every app having price of zero and ‘Paid’ for not zero.

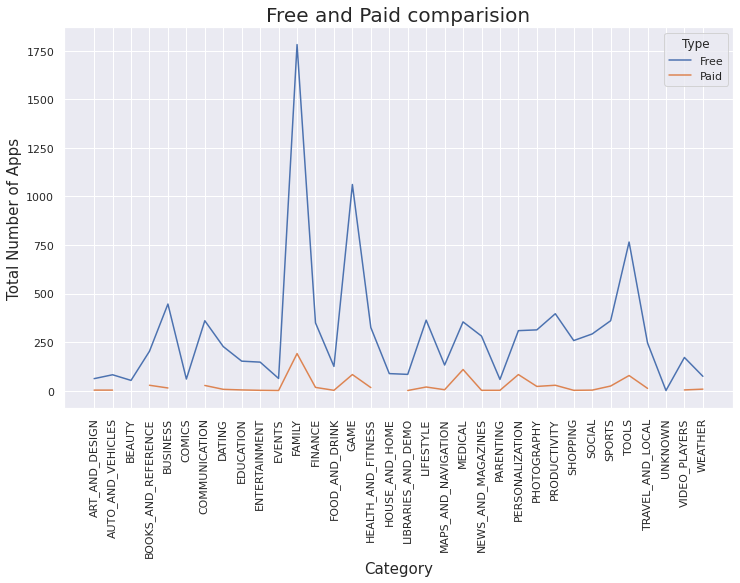
**2.3. Data Analysis And Visualization**

### Data Analysis and Visualization refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. Here we have used several packages from python’s library to complete data analysis and visualization. These packages are **numpy, pandas, matplotlib** and **seaborn.** NumPy is a Python package. It stands for **'**Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of arrays. The NumPy package is the workhorse of data analysis, machine learning, and scientific computing in the python ecosystem. Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

### 

**Instances & Observations**

Instances & Observations we have discovered after Analyzing the datasets that **Family**, **game** and **tools** category has the most number of apps in playstore. Most downloaded category based on average installs is **'Communication'** and least is **'events'** and **'medical'**. **‘Game**’ category is the most downloaded category. **Subway Surfers** is the most downloaded app under the game category. Most of the apps in the google play store are rated between 3.5 to 4.8. **Facebook**, **whatsapp Messenger** and **instagram** apps have the highest number of reviews. 93%(Approx.) of apps in the google play store are free. Only 7.4% are paid. **Minecraft** has made the highest earning, Also Minecraft is the most downloaded paid app. **I'm Rich - Trump Edition** is the most expensive app.There are more than 50 different genres available in playstore. **'Tools'**, **'Education'** and **'Entertainment'** genres have the highest number of apps in playstore. Most of the apps have more positive sentiment reviews. Most of the review's Sentiment\_polarity score is between -0.5 and 0.5. More positive Sentiment\_polarity score than negative. There are more sentiment subjective reviews as compared to neutral which means that the text contains personal opinion rather than factual information.

****

Number of free and paid apps comparison in every category

**3. Conclusion**

After Analyzing the datasets we have got answers to some of the serious & interesting questions which will help any business planning to launch their app on playstore , also it will help existing businesses to improve their apps. We have drawn some actionable insights for developers to work on and it will help a business to capture the android market.

**4. Task Assignment and Acknowledgement**

This EDA capstone project was completed by Kuldeep Singh. The datasets were given by the AlmaBetter.