Capstone Project- 1

Play Store App Review Analysis

Sudhanshu Kumar

Abstract

The Google Play Store is constantly inundated with just few thousand new applications, and an ever large number of developers are working independently or collaboratively to achieve their goals while facing an ultimate challenge from around the world. Since the majority of Play Store apps are free, it is difficult to understand how in-app purchases, advertisements, and membership fees contribute to an app's success. In this method, rather than the amount of revenue generated, an app's success is typically determined by the number of installations it has received as well as the customer reviews it has received over time. User-provided feedback known as application ratings is a crucial criterion for software assessment. However, because of inadequate or absent ballots, these scores are frequently skewed. Furthermore, there are big variations between user reviews and ratings in numbers. This analysis uses machine learning algorithms to forecast app ratings on the Google Play Store. The information for the Google Play store software that I obtained from Kaggle has been subjected to my best efforts at data analysis and prediction. I have attempted to identify links between several features contained in my dataset utilizing machine learning algorithms, such as whether an application is free or paid, information regarding user reviews, and program rating.

Problem Statement

Information is sourced from a dataset from the Google Play store. There are various listings for each app in a series. Here on data collection, we will perform exploratory study, which would be a crucial stage in the machine learning process because it not only aids in making very preliminary business choices but also prepares the data for additional modelling for use in machine learning algorithms. Our goal will be to organise the data, clean it up, and highlight any tendencies that we see that can help us make very early predictions about the likelihood that what a recently introduced app would be successful.

# **Google Play store and User Review Analysis**

Information is sourced from a dataset from the Google Play store. There are various listings for each app in a series. Here on data collection, we will perform exploratory study, which would be a crucial stage in the machine learning process because it not only aids in making very preliminary business choices but also prepares the data for additional processing used in machine learning techniques. Our goal will be to organise the data, clean it up, and highlight any tendencies that we see that can help us make very early predictions about the likelihood that what a recently introduced app would be successful.

**GOOGLE PLAY STORE DATASET**

The dataset, which would be drawn from Almabetter, the biggest international network for data analysts to discover, analyse, and share information, is made up of the Google Play Store applications.

### **The data set contains the following columns:**

· **App :**This Column contains the name of the app

· **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.

· **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.

· **Reviews:** This column contains the number of people that have given their feedback for the app.

· **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.

· **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.

· **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.

· **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.

· **Content Rating:** It indicates the targeted audience of the app and their age group.

· **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.

· **Last updated:** This column contains the info about the date on which the last update for the app was launched.

· **Current version:** Contains information about the current version of the app available on the play store.

**Android version:** Contains information about the version of the android OS on which the app can be installed

**Data Cleaning**

Preprocessing is crucial for transforming unprocessed data into a more usable state. Completeness and compellability can both be improved through preprocessing. You can check whether or not specific values were recorded, for example. You'll also discover how reliable the information is. It might be useful for determining how consistent the values are. Preprocessing is necessary because most real-world data are unclean. Data can be chaotic, meaning it may include outliers or just be generally inaccurate. Data may also include null values, or it may be incomplete.

Step 1 -

We created a function called play store info () that will show five characteristics about each column in the google playstore dataset: the data type, the number of non-null values, the number of null values, the total unique entries in that column, and the percentage of missing value in that column.

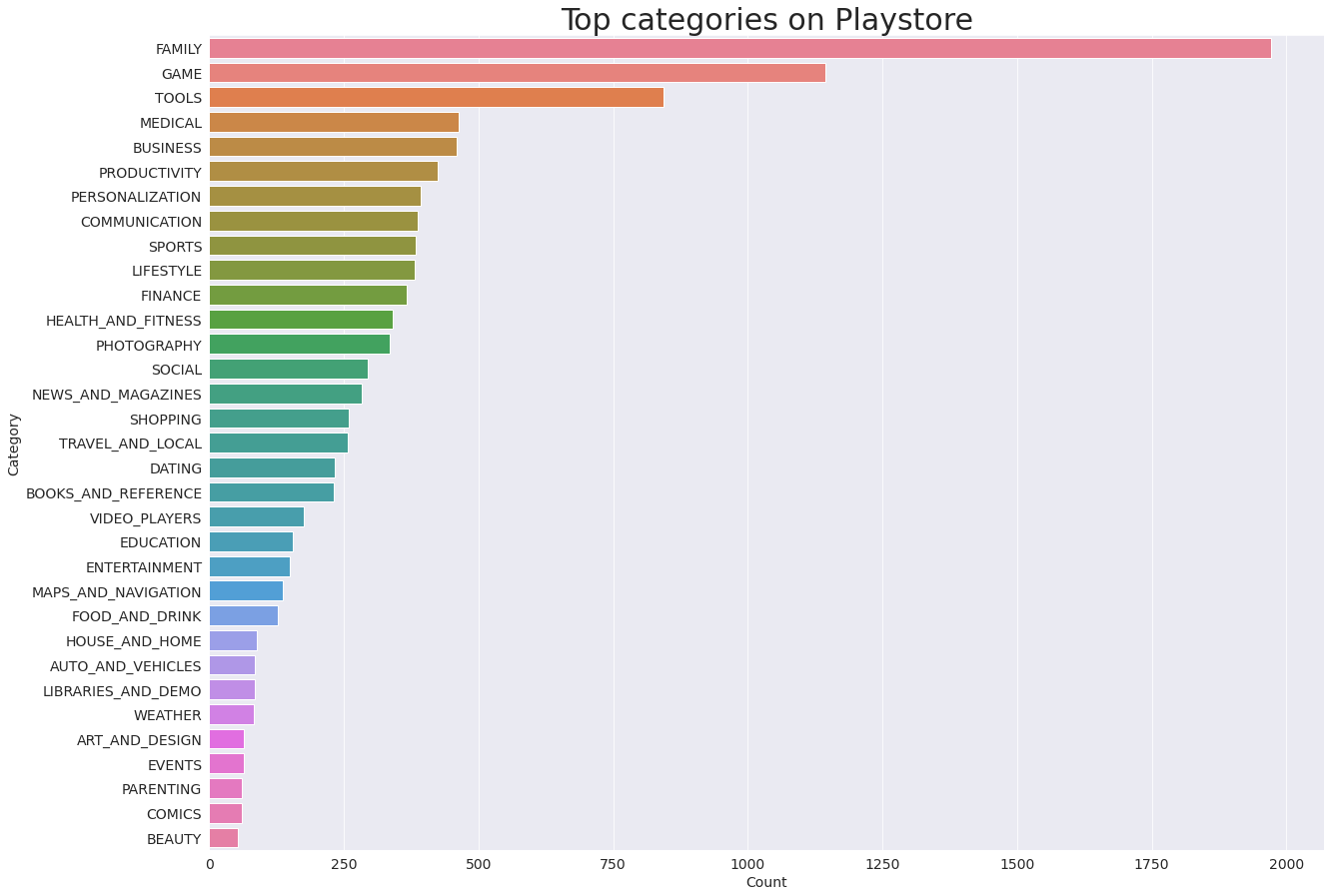
Step 2 -

The first column we look at is "Type," and we can see that it only has one null value. We looked into this row and discovered that it is a free app from the play store. To complete this number, we just use analytics library's filllna() function.

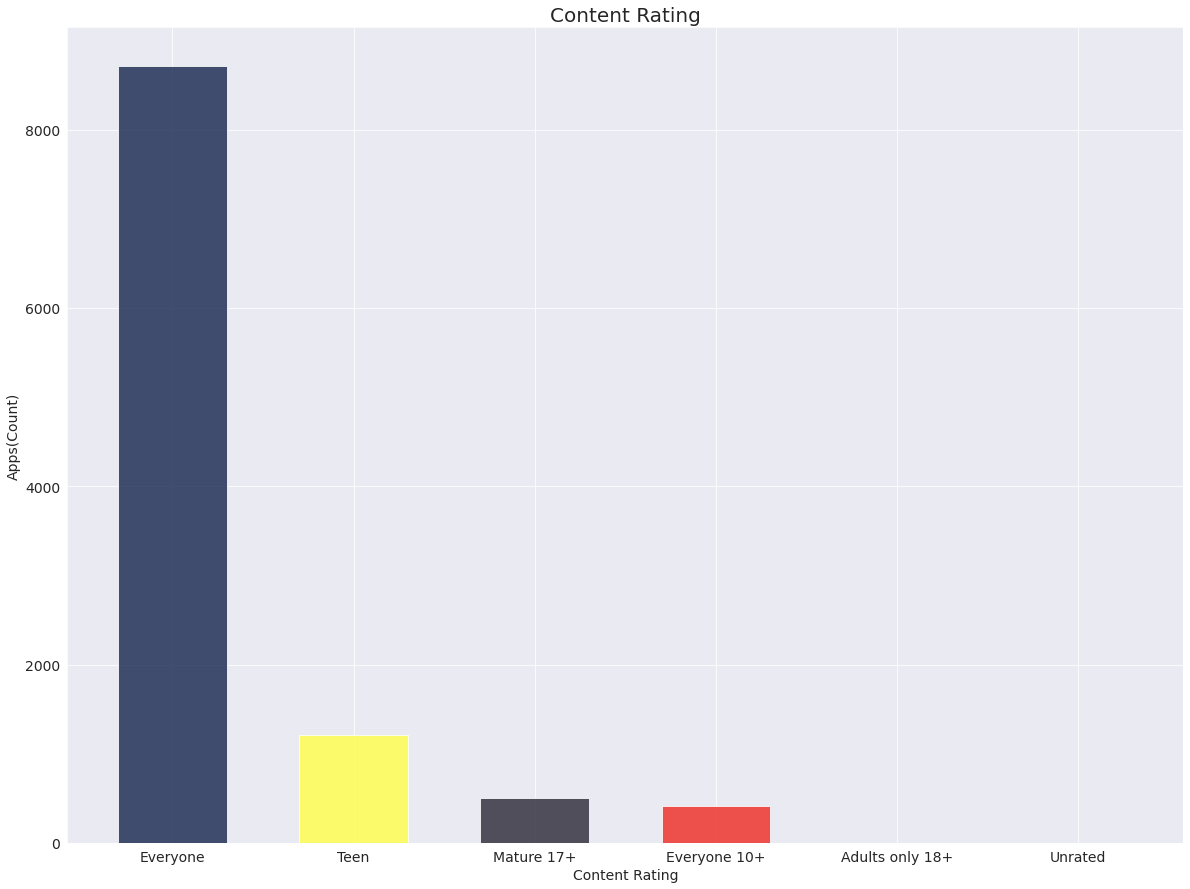
**EXPLORATORY DATA ANALYSIS**

Any project that involve data analysis or information science should begin with data exploration, or EDA. Investigating a collection to find anomalies and patterns (outliers) and developing assumptions based on our understanding of the information is the process of exploratory data analysis. (EDA). EDA entails producing statistical results for the dataset's numerical data and developing various visualization tools to aid with data comprehension. With the aid of an example dataset, we shall understand EDA in this post. For this, we'll utilise Python and the Pandas package.

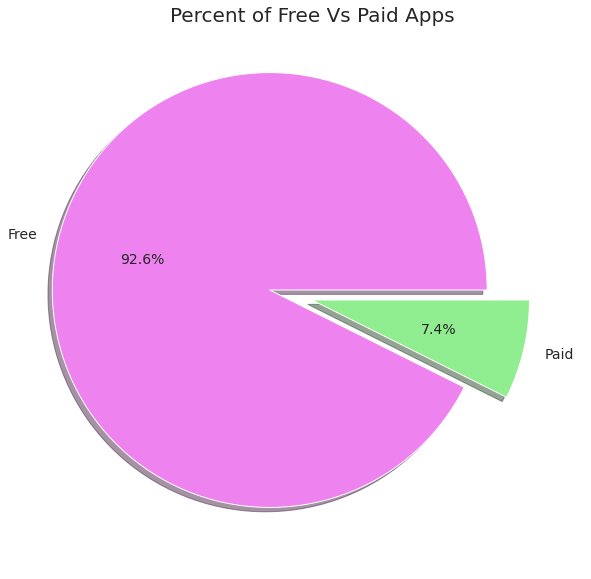
* **Finding top Categories on Play Store**



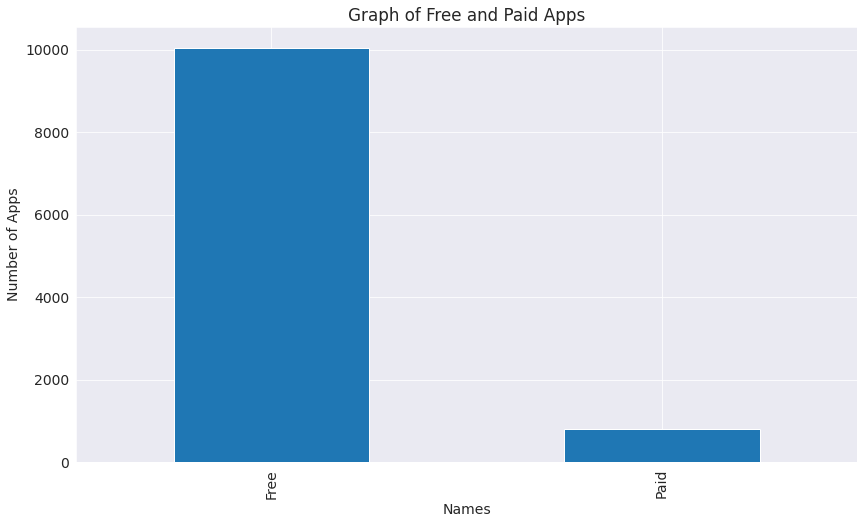
\* **Finding the Ratings of the Content**



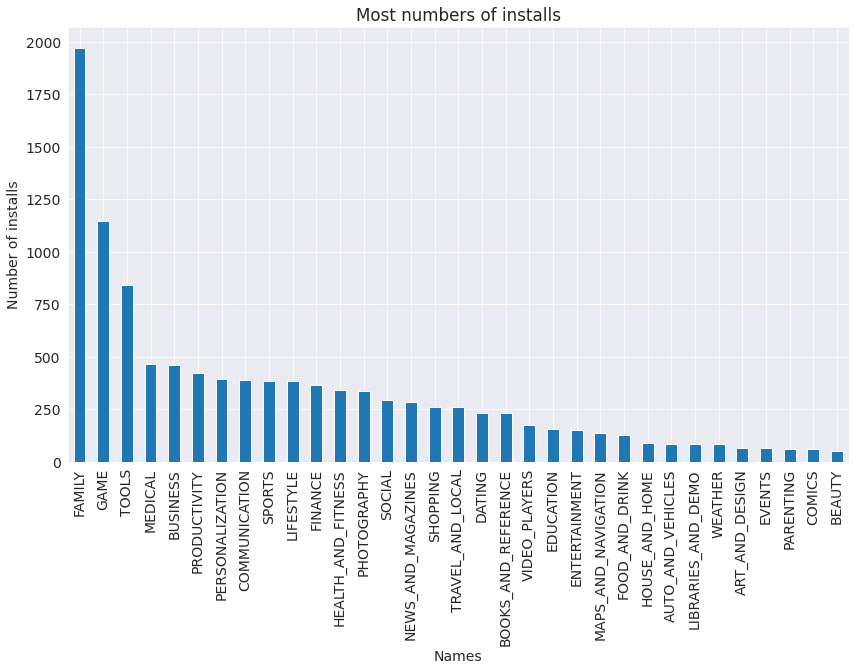
**\* Percentage of Free Vs Paid Apps**



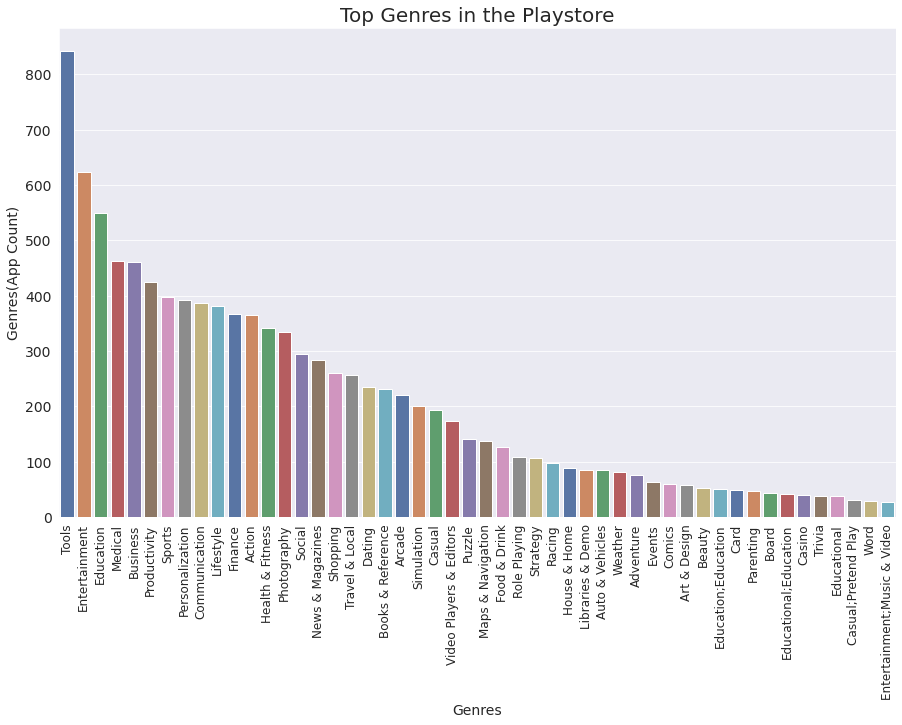
* **Bar Graph of Free And Paid Apps**



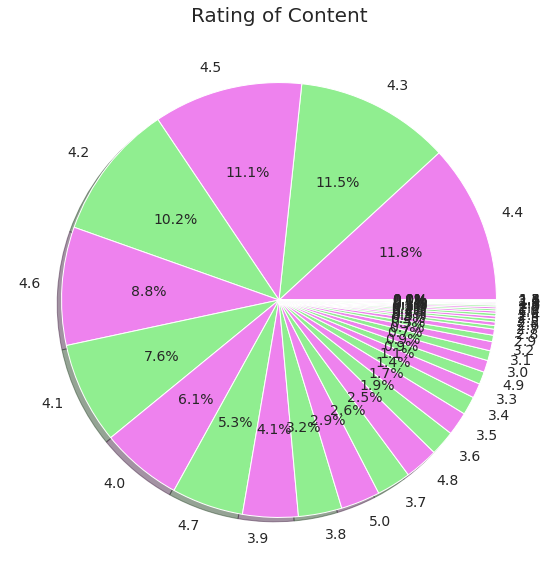
* **Category has most Number of Install**



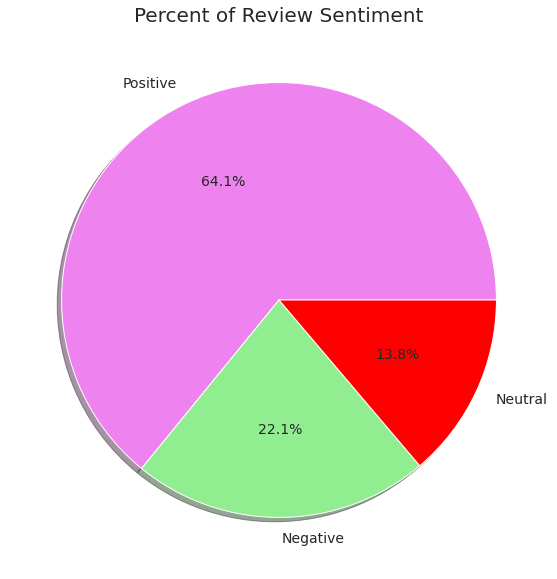
* **Top Genres in Play Store**

****

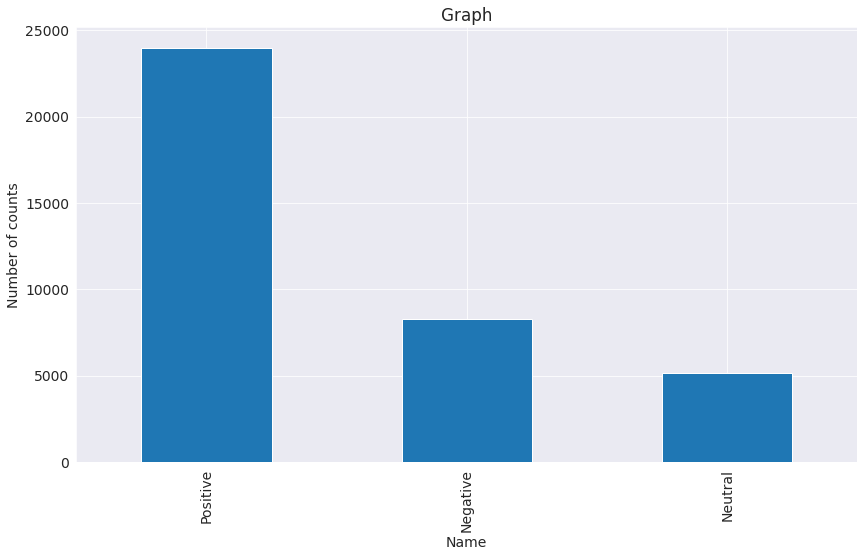
* **Rating Of Content**

****

**\* Percentage of Review Sentiments**

****

* **Printing Sentiments**

****