****

**Capstone Project-Module 1**

**Play Store App Review Analysis**

**Team :**

**Brijesh R. Patel**

**Ghanshyam Menariya**

---------------------------------------------------------\*\*\*------------------------------------------------------------

**Brijesh R. Patel -** [brp16793@gmail.com](mailto:brp16793@gmail.com)

**Ghanshyam Menariya -** [ghanshyammenariya50@gmail.com](mailto:ghanshyammenariya50@gmail.com)

**GitHub Link~~**

**Brijesh R. Patel -** https://github.com/brpatel24/play-store-app-reviews-analysis

**Ghanshyam Menariya -** <https://github.com/ganny55/play-store-review.git>

# Problem Statement

Data is taken from the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in data science cycle, as it not only helps in taking very initial business decisions Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

# Google Play store Dataset

The dataset consists of Google play store application and is taken from Almabetter , which is the world’s largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application.

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.

# User Review Dataset

* User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:
* **App:** Contains the name of the app with a short description (optional).
* **Translated Review:** It contains the English translation of the review dropped by the user of the app.
* **Sentiment:** It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
* **Sentiment Polarity:**It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
* **Sentiment Subjectivity:** This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer’s opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

# Data Cleaning and Preparation

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.

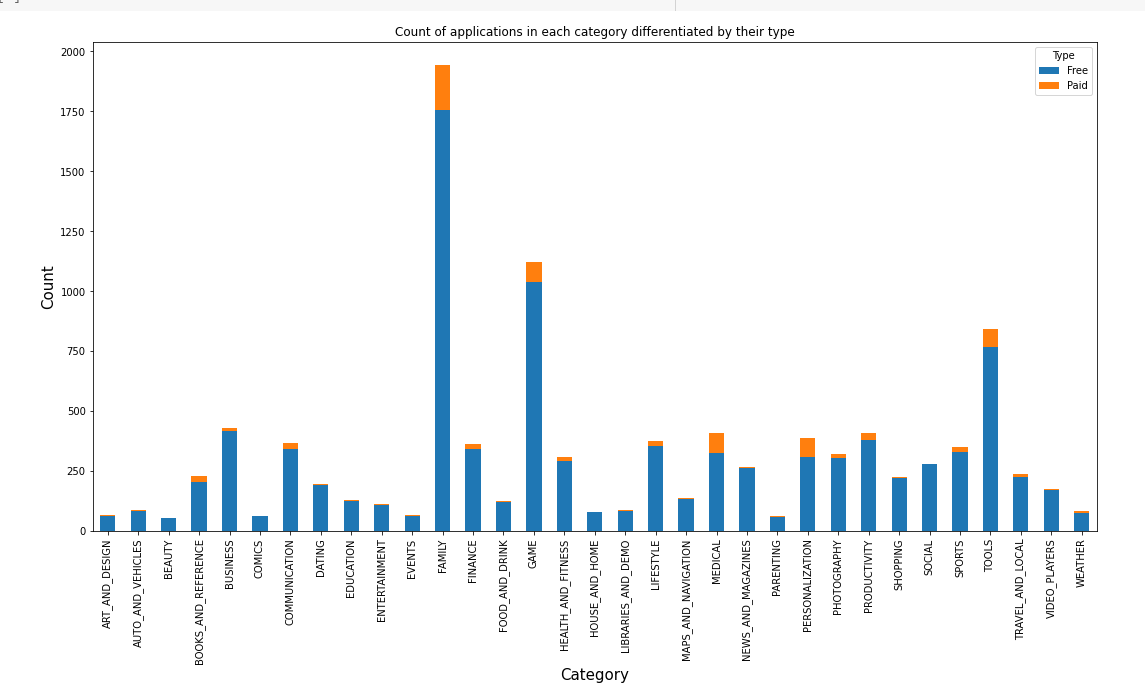
The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis

# EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

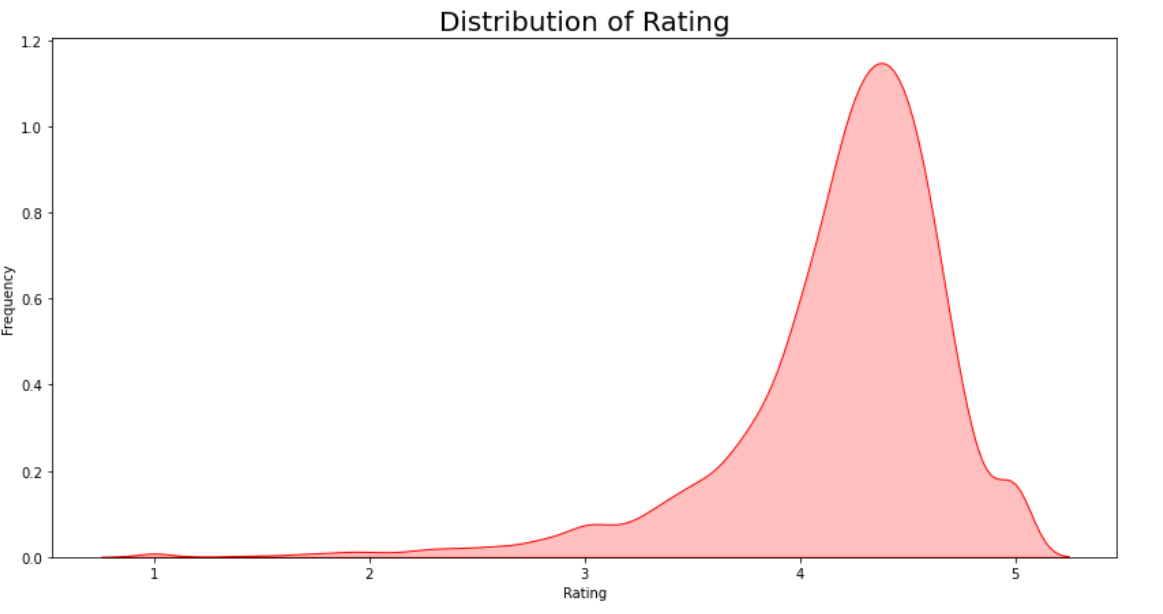
EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

**count of applications in each category differentiated by their type**



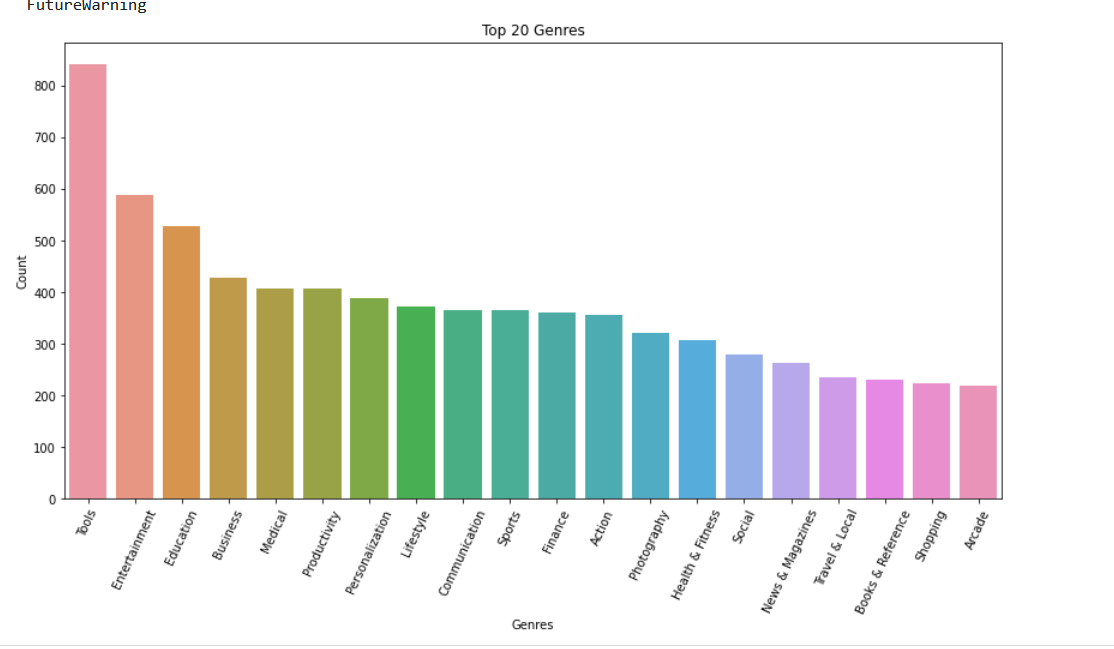
It looks like certain app categories have more free apps available for download than others. In our dataset, the majority of apps in Family, Games and Tools, as well as Social categories were free to install. At the same time Family, Personalization and Medical categories had the biggest number of paid apps available for download.

**Distribution of App Rating**



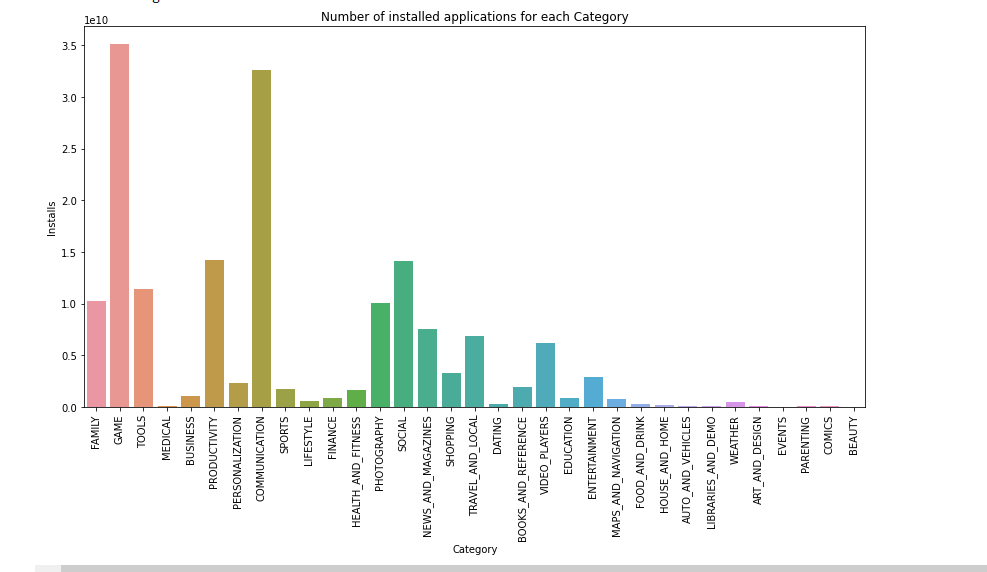
Average rating of application in store is around 4.3, which is very high. This plot can be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

Top 20 genres

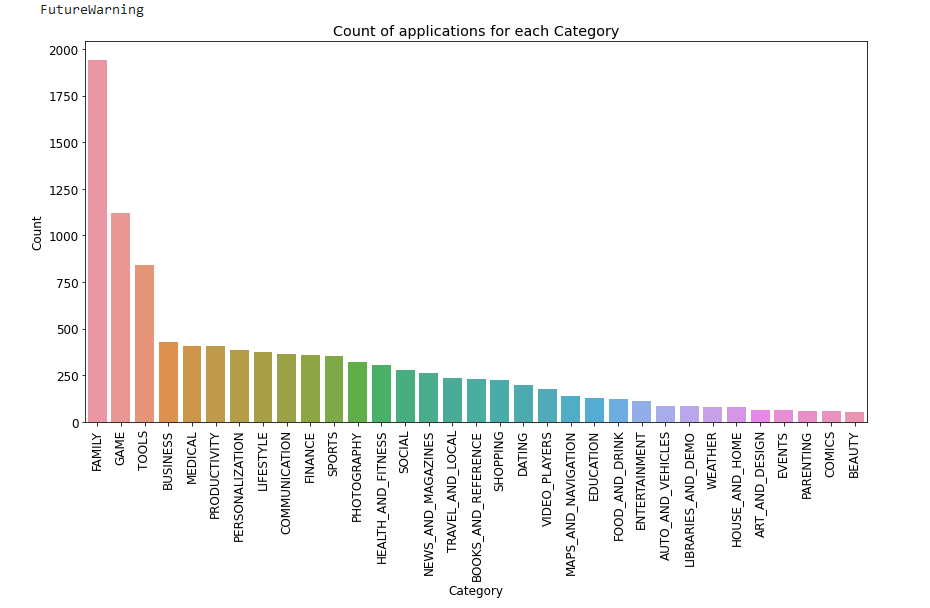


As well as we can sea the tools and entertainment genres have most of count

**Number of Installed applications for each category**

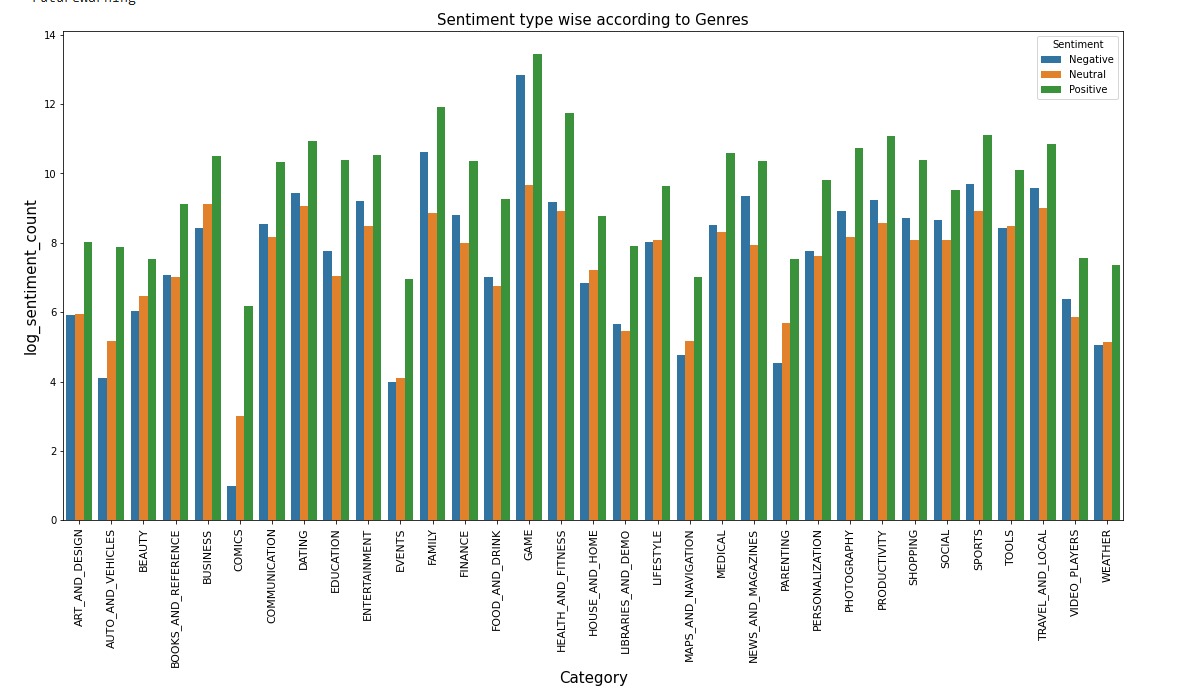


Count of applications for each Category



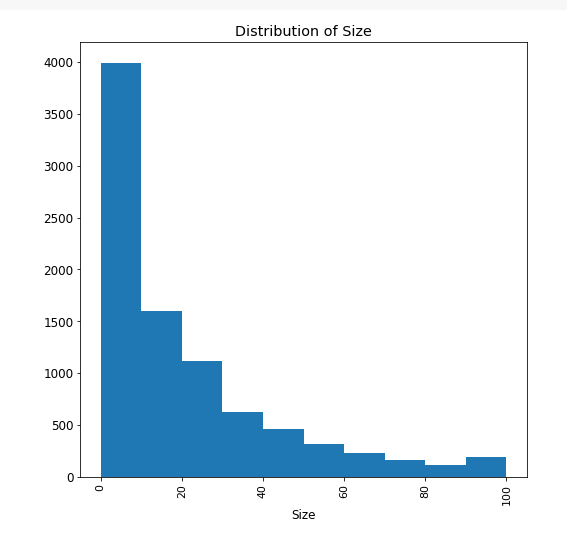
From the above two plots we can conclude that, maximum number of apps present in google play store comes under Family, Games and Tools Category but as per the installations and requirements in the market place, this is not the case. Maximum installed apps comes under Games, Communication and Tools.

**distribution of type of reviews, category wise in the dataset**



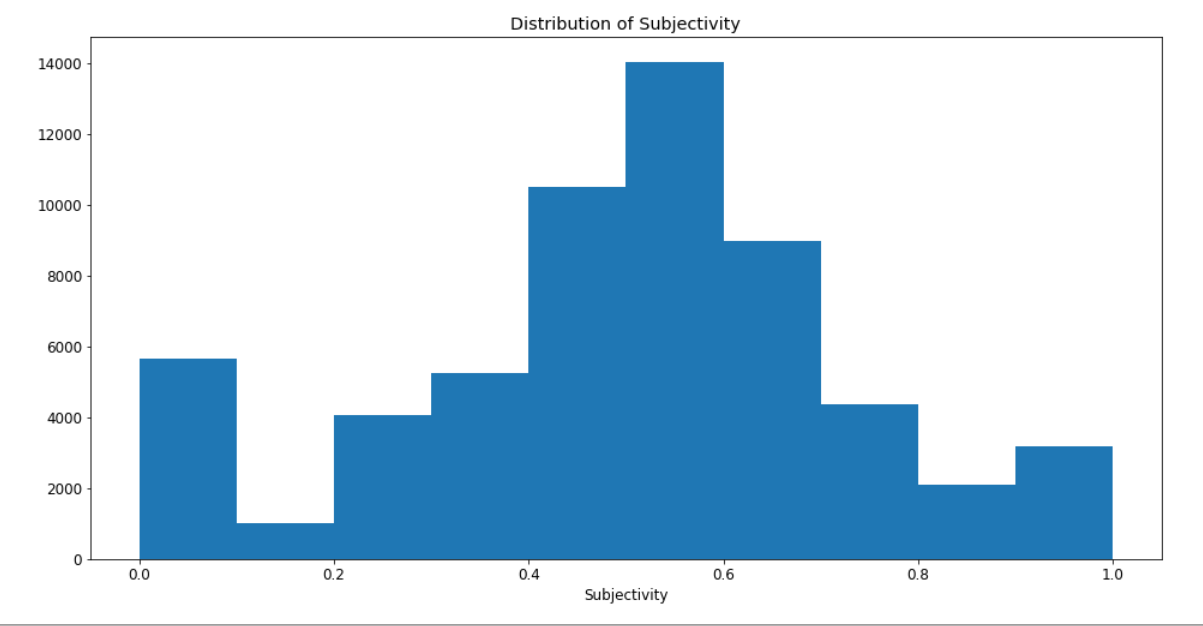
It can be seen from the above plot that the number of positive reviews are way higher than negetive and neutral ones.

**Distribution of App Size**



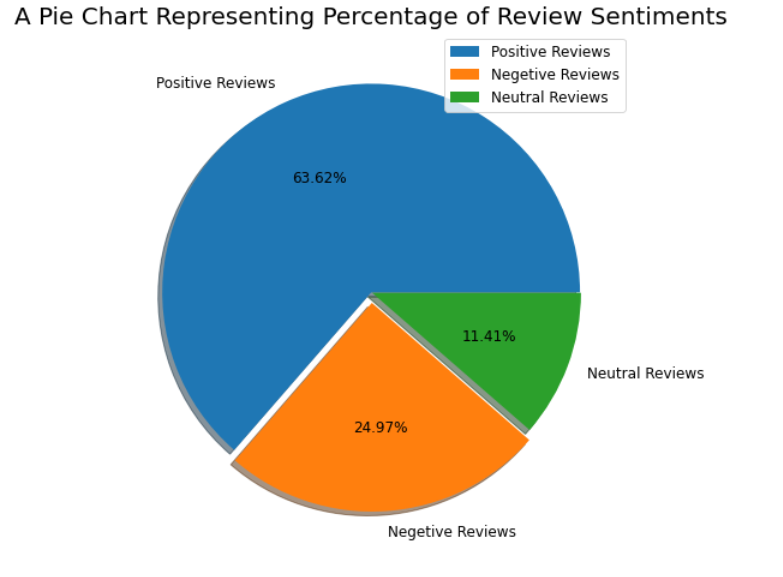
It can be seen from the above plot that maximum application’s size lies between 0–10 MB.

**Distribution of Subjectivity**

****

**It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications according to their experience.**

**Representing Percentage of Review Sentiments**

****

**It can be seen from the above plot that the number of positive reviews are way higher than negative and neutral ones.**

**Conclusion**

The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product. Dataset can also be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store