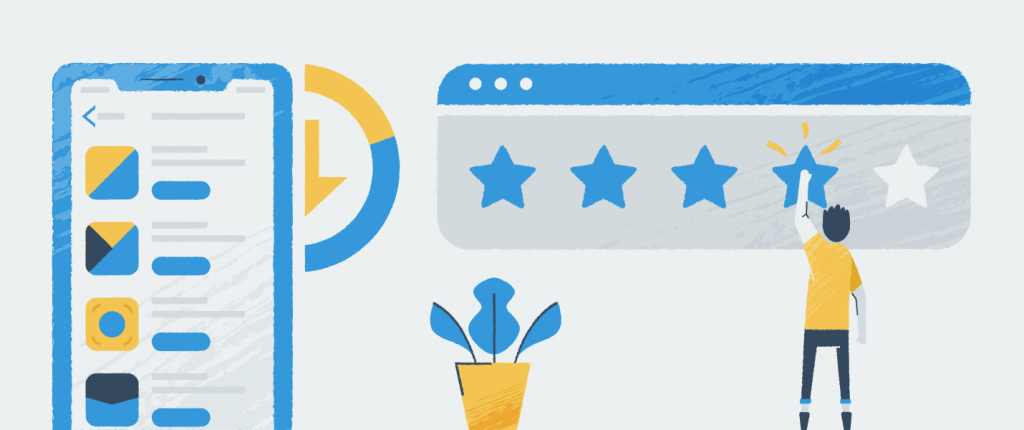
**Capstone Project-1**

**Play Store App Review Analysis**

****

**Data science trainees**

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

**VIVEK KUMAR – vivek.kumar2696@yahoo.com**

**ANKUR SARKAR – ankur04052000@gmail.com**

**SANIYA YUSUF PIRJADE -** [**saniyapirjade786@gmail.com**](mailto:saniyapirjade786@gmail.com)

**PRAVIN CHAVAN -** [**pravinsanjaychavan1999@gmail.com**](mailto:pravinsanjaychavan1999@gmail.com)

**RATNESH KUMAR - ratnesh.kumar157@gmail.com**

**Abstract -** *Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. Application (App) ratings are feedback provided voluntarily by users and function important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. This Study aims to predict the ratings of Google Play Store apps using machine learning Algorithms. I have tried to perform Data Analysis and prediction into the Google Play store application dataset that I have collected from Kaggle. Using Machine Learning Algorithms, I have tried to discover the relationships among various attributes present in my dataset such as which application about the user reviews, rating of the application.*

***Key Words*:** Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, python.

# 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 but also in preparing the data for further modelling for use in machine learning algorithms. 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.

# INTRODUCTION

Python approaches are essential for us to take care of numerous issues. In this paper, we present python models and structures in detail. python has numerous applications in numerous perspectives and has incredible advancement potential.

In future, it is predictable that python could set up ideal speculations to clarify its exhibitions. In the meantime, its capacities of unsupervised learning will be improved since there is much information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize these points of interest to achieve more assignments.

# Google Play store and User Review Analysis

In today’s scenario we can see that mobile apps playing an important role in any individual’s life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that although it creates more than two-fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and rating to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on generally speaking evaluations alone and would profit by understanding a huge number of printed remarks.

We develop Android apps & release on Play Store. As an Developer or say Business Perspective it’s very important to know whether users are enjoying the app or facing any issues. To know this Play Store has a Ratings & reviews section for each app released on play store. Users can submit the ratings and has a freedom to write a review for a particular app. This approach is quite a lengthy to rate & review app i.e. navigate to Play store to submit feedback or redirect leaving a current app workflow to open Play Store App link using URI. We never wanted our customers to leave our application, but with this flow, we are forced to redirect the control to Play store 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. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

### 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.

# Python

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programing language to select up compared to other language. That is the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That is one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open-source library is named panda. As we have seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

# Data Cleaning

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.

* **Step1**: First we used info function to know that how many null values are present in data frame for each column.
* **Step2:** Then we came to know that ratings column has the maximum number of null values.
* **Step3:** After that we used isnull function and replace function to replace those null values with median of the rest of the values of the ratings column.
* **Step4:** After plotting the box plot we came to know that, there is an outlier in the rating column.
* **Step5:** As rating more than 5 is not practically possible so we decided to find the row number in which the outlier is present.
* **Step6:** We found that the outlier is present in the row number 10472.
* **Step7:** we dropped that row(10472).
* **Step8:** Then we again plotted the box plot and saw the outlier was not present.
* **Step9:** we found that there were some unnecessary symbols in the **price** and **install** column such as comma(,), plus(+).
* **Step10:** We used **lambda function** to replace those symbols with “(space)“.
* **Step11:** finally we checked the data frame with **head function** and found it was looking clean and nice.

# EXPLORATORY DATA ANALYSIS

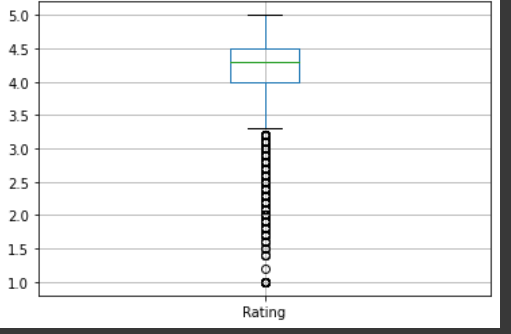
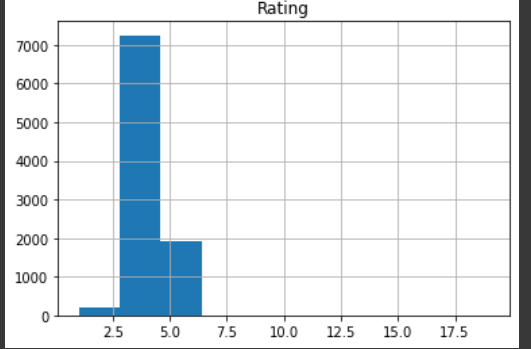
Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

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** **, Numpy, matplot,** library) for this purpose.

## Rating



In the below plot, we plotted histogram and boxplot for rating

**BOXPLOT HISTOGRAM**

**Fig -1:**

* The mean of the average ratings (excluding the NaN values) comes to be 4.2.
* The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3. From this we can say that 50% of the apps have an average rating of above 4.3, and the rest below 4.3.
* From the distplot visualizations, it is clear that the ratings are left skewed.
* We know that if the variable is skewed, the mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation of the majority of the values in the variable.

## Average App ratings

**Average App Ratings**

The rating available in the dataset is distributed so we can represent the ratings in a better way if we group the ratings between certain intervals. Here, we can group the rating as follows:

* 4-5: Top rated
* 3-4: Above average
* 2-3: Average
* 1-2: Below average

**Conclusion~**

Through exploratory data analysis we have observed some trends and have made some assumptions that might lead to app success among the users in the play store.

* Most competitive category: Family
* Family, Game and Tools are top three categories having 1906, 926 and 829 app count.
* Tools, Entertainment, Education, Business and Medical are top Genres.

Thank You