**Play Store App Review Analysis**

**Krushna Chaure**

**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 amount of data that can be used to make an optimal model. Apps have reshaped nearly every aspect of our lives, from ordering food, making a grocery list, checking account balances, and communicating with others. For everything, there's literally an app. Our everyday activities have been improved and simplified by technology, thereby making things easier. But with the growing Industrial revolution in the form of Digitalization has boosted the development of more and more application. So, we’ll use some Data sets (Specifically Play store data and user review csv files) to analyse the key factors behind the growth, success or failure of an App.

**Introduction:**

In today’s scenario we can see that mobile apps playing an important role in any individual’s life. It is one of the fastest-growing segments of downloadable software application markets It has become more important as the android market has gone to a real amelioration among mankind over the last decade. One of the main reasons for this popularity is the fact that about 81% of the apps are free of cost. The market has increased to over 3.5 million Apps and around 3000+ apps are being added per day as per a Google survey report. Thus, the market, in turn, led to around 5 billion users downloading all over the world. Developers and users play key roles in determining the impact that market interactions have on future technology. With enormous challenge from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. 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 dataset with 10k Play Store applications is available to analyse the market of android. It can be examined to analysis the different category such as family, communication, entertainment, tools, music, camera etc. In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app, is there some trends in user sentiments. In our data set we have two csv files for data analysis: Play Store data User Reviews At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data. We need to explore and analyse the data to discover key factors responsible for app engagement and success.

**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.

**What is Exploratory Data Analysis:**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets for patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset and summarize their main characteristics, often employing data visualization methods. It is an important step in any Data Analysis or Data Science project. It helps determine how best to manipulate data sources to get the answers you need.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better and make it more attractive and appealing.

The following are the various steps involved in the EDA process:

1. **Problem Statement** - We shall brainstorm and understand the given data set. We shall study the attributes present in it and try to do a philosophical analysis about their meaning and importance for this problem.
2. **Hypothesis** - Upon studying the attributes present in the data base, we shall develop some basic hypothesis on which we can work and play with the data to look for the varied results which we can get out of it.
3. **Univariate Analysis** - It is the simplest form of analyzing the data. In this we would initially pick up a single attribute and study it in and out. It doesn't deal with any sort of co-relation and it's major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.
4. **Bivariate Analysis** - This analysis is related to cause and the relationship between the two attributes. We will try to understand the dependency of attributes on each other.
5. **Multivariate Analysis** - This is done when more than two variables have to be analyzed simultaneously.
6. **Data Cleaning** - We shall clean the dataset and handle the missing data, outliers and categorical variables.
7. **Testing Hypothesis** - We shall check if our data meets the assumptions required by most of the multivariate techniques.

**Dataset Description:**

There are two types of Dataset:

1. **Google Play Store Dataset**
2. **User Review Dataset**
3. **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.

1. **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.

**Breakdown of Datasets**:

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

* Importing Analytical necessary library classes for future analysis.
* Reading the csv data file from Google drive.
* Setting figure size for future visualization.
* Removing future warnings in seaborne plots.
* Visualizing all the columns of the respective Data frame.
* Viewing all data information
* Checking the Unique values in the column ( if any)
* Converting the data types to similar objects as the Analysis Demands.
* Formatting the “size” column into a single column in the dataset.
* Eradicating special characters from the dataset columns.

**Examining Null Values:**

The most critical thing from which we can draw some observations is Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at the time of analysing the first thing which we will do is to examine the null or missing values on the Dataset. It is the first step that will make the results “more” accurate &should be handled before it affects the performance of the models that predict the outcome. By plotting a graph it can be seen that missing values are more in Size & Rating columns. Hence, several methods to eradicate those nullvalues.

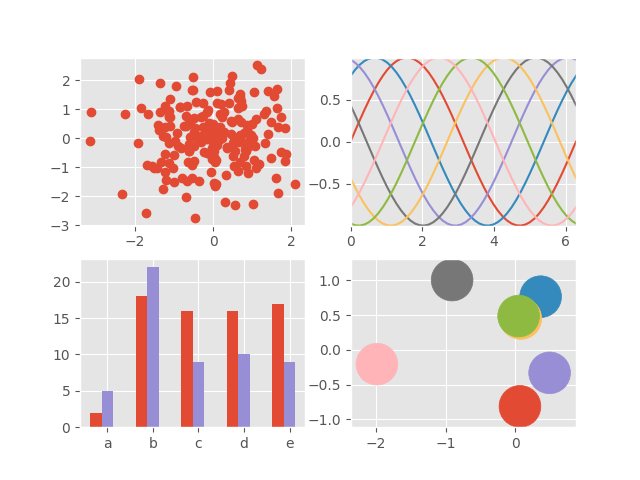
**Data Cleaning:**

Data cleaning is one of the most essential subtask of any data science project. Although it can be a very tedious process, it's worth should never be undermined. By looking at a random sample of the dataset rows (from the above task), we observe that some entries in the columns like Installs, Price and Size have a few special characters (+ , $ ,M , k) .This prevents the columns from being purely numeric, making it difficult to use them in subsequent future mathematical calculations. Ideally, as their names suggest, we would want these columns to contain only digits from [0-9]. Hence, we now proceed to clean our data. Specifically, the special characters "," and "+" present in Installs column and "$" present in Price column need to be removed. By finding all unique values of each row the inappropriate values can be identified. Different methods can then be used for removing them or to change those values accordingly to use them to make predictions better. As the proverb goes by saying “More Data beats clever algorithm, but better data beats more Data” – Peter Norvig. So going with the method firstly we have found the categorical null values and replacing them by a textual string, secondly finding out the numerical ‘Nan’ values & replacing them with the median of that respective column .After we had check the entire datasets for any null values (if, any exists after eradicating). Now, cleaning all the null values we would drop certain labels/columns which is unnecessary for actionable insights. Therefore, we are can proceed for the Exploratory Data Analysis and observations regarding the datasets.

**Data Visualization using charts and plots:**

**Plt.Style.Use** (ggplot):

* A plot style is an object property, similar to line type and color. A plot style can be assigned to an object or a layer. A plot style controls an object's plotted properties, including Color, Size etc.
* GG Plot is a plotting package that provides helpful commands to create complex plots from data in a data frame. It provides a more programmatic interface for specifying what variables to plot, how they are displayed, and general visual properties.

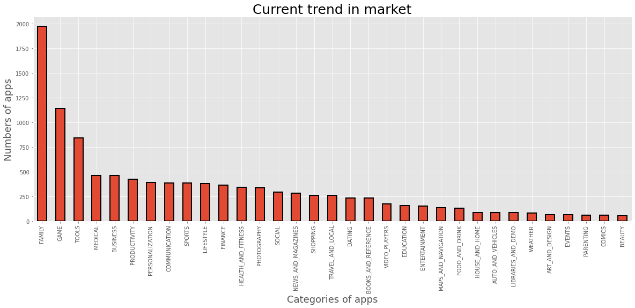


**Fig- ggplot**

* Advantages of style plot is consistent underlying grammar of graphics plot specification at a high level of abstraction. very flexible. theme system for polishing plot appearance (more on this later) mature and complete graphics system. many users, active mailing list.

**1. Bar Chart:**

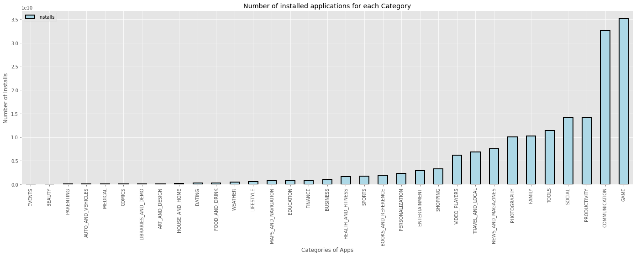
* Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data
* A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. Some bar graphs present bars clustered in groups of more than one, showing the values of more than one measured variable.
* Advantages Bar graph summarises the large set of data in simple visual form. It displays each category of data in the frequency distribution. It clarifies the trend of data better than the table. It helps in estimating the key values at a glance.



**Fig-1: Categories on Playstore**

As above Bar chart we analysis that and finding some insights.

* Family and Game have maximum number of apps available in play store where Comics and Beauty has less number of apps in play store.
* So absolutly trend is goes towards families and gamings types of apps which has maximum numbers of apps available in market.
* So this analysis very helpful to improve to making business decision like where we focused, where is more opportunity to grab the market and where we skipped, where is less market.



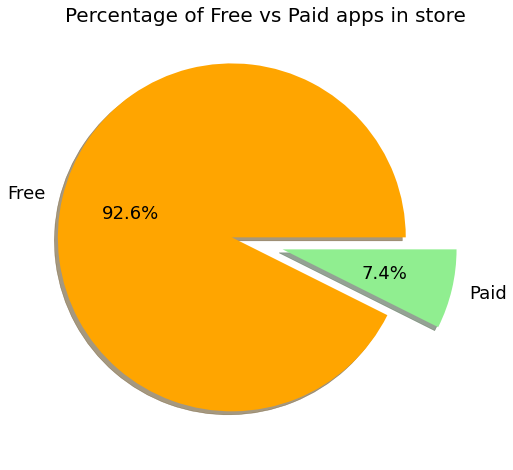
**Fig-2: No. of Installs Per Category**

As above bar chart we analysis that and finding some insights.

* No doubt where gaming and communication type of apps are top in list where they installs user maximum time.
* Events type of apps not installed any user.
* From Beauty to Business type of apps are very less installation.

**2. Pie Chart:**

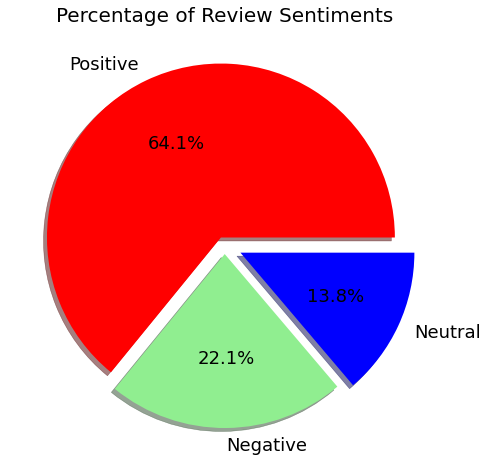
* a circular chart cut by radii into segments illustrating relative magnitudes or frequencies
* A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.
* Advantages of pie chart is represents data visually as a fractional part of a whole, which can be an effective communication tool for the even uninformed audience. It enables the audience to see a data comparison at a glance to make an immediate analysis or to understand information quickly.



**Fig-3: Free vs Paid**

As above pie chart we analysis that and finding some insights.

* There are 92.6% apps are free to use in play store which is almost all the market.
* There are only 7.4% paid apps in play store.

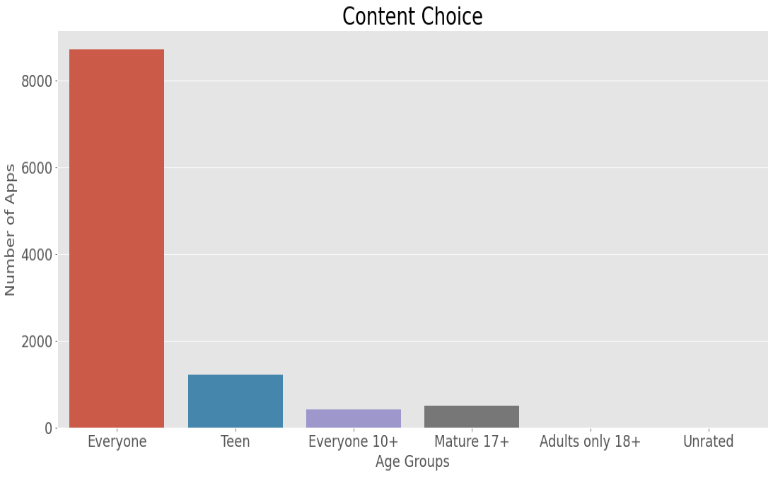
**Fig-4: Percentage of User Review Sentiments**

As above Pie chart we analysis that and finding some insights.

* It clearly shows that Positive feedback has maximum in this list where they grab 64.1% where it means customer are happy and satisfy using apps.
* Where Negative and Neutral feedback less in this which is 22.1% & 13.8% respectively.

**3. Count plot:**

* Count plot is to Show the counts of observations in each categorical bin using bars.
* countplot() method is used to Show the counts of observations in each categorical bin using bars. Parameters : This method is accepting the following parameters that are described below: x, y: This parameter take names of variables in data or vector data, optional, Inputs for plotting long-form data.
* Advantages of Count plot is value counts for a single variable, representing two categorical variable using hue parameter, creating horizontal plots.



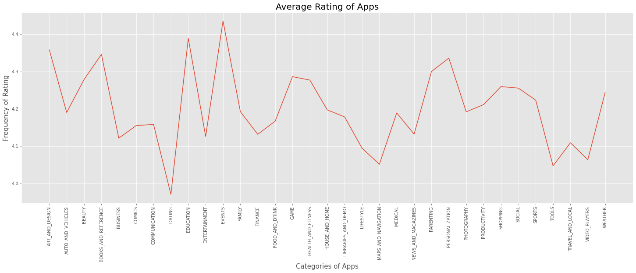
**Fig-5: Content Choice**

As above counterplot we analysis that and finding some insights.

* We see clearly that most of the apps in play store are accessible for everyone, their is any restriction to use this apps.
* There is minimum numbers of apps are in play store where there is use only Teen, Everyone10+, Mature 17+.
* There is no any apps in play store have restriction like they use only Adults only 18+, Unrated.

**4. Line Chart:**

* A line chart is a form of graphical representation of data in the form of points that are joined continuously with the help of a line.
* A line chart consists of a horizontal line i.e. x-axis and a vertical line i.e. y-axis to represent the data.
* Advantages of line chart for showing changes and trends over different time periods. It is also helpful to show small changes that are difficult to measure in other graphs. Line graph is common and effective charts because they are simple, easy to understand, and efficient.



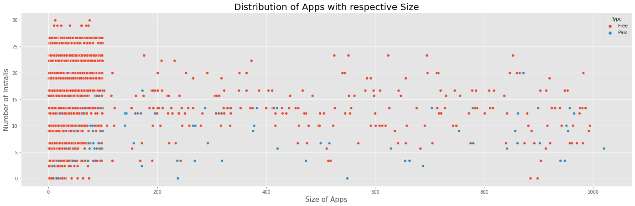
**Fig-6: Average Rating of Apps**

As above Line plot we analysis that and finding some insights.

* Most of the apps rating are between in 4 to 4.5.
* Where highest rating to Education base and Event base apps.
* Where lowest rating got to the Dating and Tools apps.

**5. Scatter Diagram:**

* a two-dimensional graph in rectangular coordinates consisting of points whose coordinates represent values of two variables under study.
* A scatter plot is a diagram where each value in the data set is represented by a dot. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis
* Advantages of scatter diagram is a shows the relationship between two variables. It is the best method to show you a non-linear pattern. The range of data flow, like the maximum and minimum value, can be determined. Patterns are easy to observe.



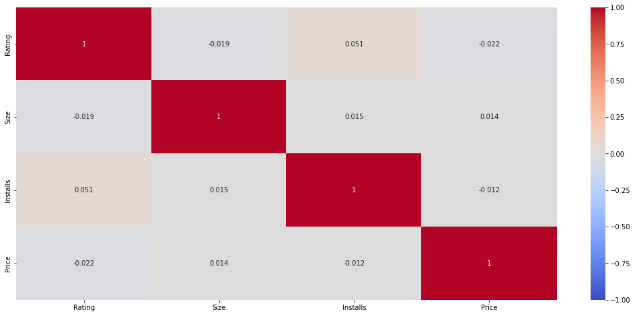
**Fig-7: Distribution of Apps With Respective Size**

As above scatter plot we analysis that and finding some insights.

* It clearly show that where size of apps smaller then it will get higher installation.
* Where size of apps increases then installation decreases.
* Also we look deeply into that there free apps are more installed compare to paid apps.

**6. Heat maps:**

* Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate in relation to each other. A heat map is a data visualization technique that’s a graphical representation of data in two dimension, using colours to demonstrate different factors.
* Practical data science using python in heatmap. Where heatmap contains values representing various shades of the same colour for each value to be plotted. Usually the darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour can also be used.
* Advantages of heat map help you to understand how users and customers interact with your website, giving insight into things like where they're looking and how far down they're scrolling. This information can help you set up your site in a more user-friendly way.



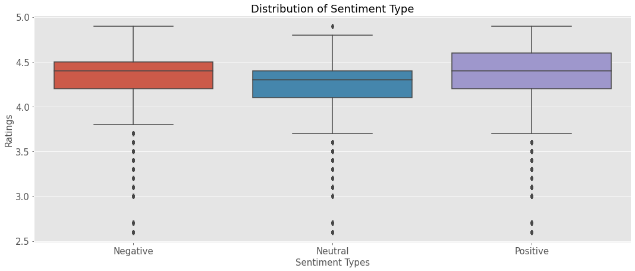
**Fig-8: Correlation Heatmap**

As above Heatmap we analysis that and finding some insights.

* In heatmap we using lot of correlation like Rating, Size, Installs, Price.
* Where we see that there is slightly positive feedback or rating between Installs and Rating and there negative rating between Size and Rating.
* Also there is negative feedback or rating between Price and Rating.

**7. Box Plot:**

* A box plot is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate if a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. Boxplots are also very beneficial when large numbers of data sets are involved or compared.
* A box plot is a highly visually effective way of viewing a clear summary of one or more sets of data. It is particularly useful for quickly summarizing and comparing different sets of results from different experiments. Some interesting findings have came along with it in our project where we used this to understand the the pattern of ratings with sentiments

****

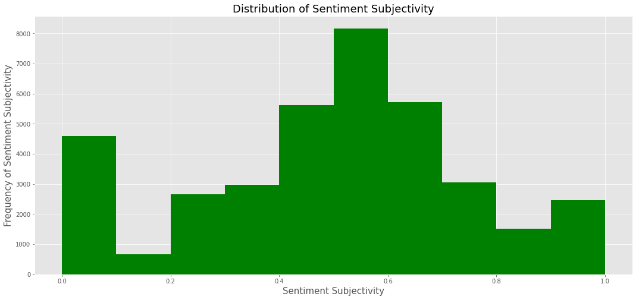
**Fig-9: Distribution of Sentiment Type**

As above Boxplot we analysis that and finding some insights.

* Clearly shows That even its Postive, Negative or Neutral the median(50%) remains near 4.4.
* The Quartile 1(25%) and Quartile 3(75%) in all the cases have a slightly differences.
* So here we deal with customer to study their behaviour like they give 4.4 rating even if the sentiment is negative.

**8. Histogram:**

* A histogram is an approximate representation of the distribution of numerical data.
* A histogram is basically used to represent data provided in a form of some groups. It is accurate method for the graphical representation of numerical data distribution. It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.
* Advantages of histogram is a helps to know and compare the frequency of occurrence of different data ranges. Also its simplicity and versatility. It can be used in many different situations to offer an insightful look at frequency distribution.



**Fig-10: Sentiment Subjectivity**

As above Histogram we analysis that and finding some insights.

* It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7.
* So maximum number of customers give review to the apps according to their experience.

**Conclusion and Future work:**

Thus the app development companies could decide what application should be developed and they can also see the prediction of their developed application. In this they also get to see the categorized reviews of all the application in one interface which will help them decide which app is liked by the users and which apps need to be developed more. The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project. We could add a system that would create application on its own by using the data set and creating the best user interface by the highly rated apps.

**Acknowledgement:**

This project was completed by Krushna Chaure. I am extremely grateful to the celebrated authors whose precious works have been consulted and referred to in this project work. I also wish to convey my appreciation to our peers who provided encouragement and timely support in the hour of need.

**References:**

1. Stackoverflow
2. Geeksforeeks
3. Jovian
4. Matplotlib.org
5. Sciencing.com

**Thank You**