**CAPSTONE PROJECT(EDA)**

11/18/2022

**Play Store App Data and User Review Analysis**

Created By

**Chandrashekhar Chaudhari Kishore Yamali**



Email

Chandrashekhar : [chandrashekhar130697@gmail.com](mailto:chandrashekhar130697@gmail.com)

Kishore : [kishoreyamali36@gmail.com](mailto:kishoreyamali36@gmail.com)

GitHub Link



Chandrashekhar : [Chandrashekhar569 (github.com)](https://github.com/Chandrashekhar569)

Kishore : [kishoreyamali (github.com)](https://github.com/kishoreyamali)

A**bstract -** 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 is free or paid, about the user reviews, rating of the application*.*

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

# 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

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

In future, it is predictable that machine learning 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 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.

* **Step1**: We write a function play store info (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.
* **Step2**: we start off with the column ‘Type’ we can see that it has one null value. We checked this row and found out from the play store that it is a free app. We use filllna() function of the pandas library to fill this value.
* **Step 3**: We drop the columns ‘Current Ver’, ‘Android Ver’ and ‘last updated’ from our dataset using the drop() function of the pandas library.
* **Step 4**: We can see that the ‘Rating’ column has 1474 null values. Due to low variations in the rating values and a lot of repeated values the ‘median’ would be a suitable statistical indicator to replace the null values with. We calculate the mode of the column using the median () aggregate method, and fill this value in place of null values using the fillna() function.
* **Step 5:** We can see that the ‘Reviews’ column despite being a numerical indicator is of the ‘object’ data type, we will convert this to ‘int’ data type using the as type(int) function.
* **Step 6:**  We can see that the size column, which should be numeric, is of the data type ‘object’, it also has characters ‘k’ and ‘M’ in the values which stand for kilobytes and Megabytes, we will replace the ‘k’ with 1000 and ‘M’ with 1000000. Some values also have ‘+’ sign in them, which will be removed. Next, we will convert this column into ‘int’ datatype.
* **Step 7:** The ‘Installs’ column values contain the characters ‘+’ and ‘,’ which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the strip() and replace() functions.
* **Step 8:** The values in the column ‘Price’ might have the ‘$’ sign in some values and the column is of the datatype ‘object’. We will first remove the ‘$’ sign using the **strip()** function and then convert the column into ‘int’ datatype.
* **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.
* **Step 10:** We write a function Ur info(), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values ,number of unique values in that column and percentage of null value in that columns in the User review dataset.
* **Step11:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using dropna() function.

# 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.

## 1. Top 10 Installed Apps in Play Store :

* Taking the unique values of the apps.
* Calculating the value counts of the no of installs to make a relation between value counts of the installs and apps.
* Taking the ascending order of the relation to see the top required no of top list of the apps with highest downloads.
* Plotting the graph of the top required no apps and the value counts of the downloads.

Icon

Description automatically generated with low confidence

**AXIS: X = NAMES OF THE APPS , Y = VALUE OF RATINGS IN 10^VARIABLE**

* **Observation** : from the above graph we can clearly see the top 10 apps installed from play store and also we can say that Roblox is the highest installed app and there is only one app which belongs to learning category in top 10 apps and most of the remaining apps belongs to games category in play store.

## Total No of Apps In Each Category:

* Plotting a Bar Graph to show total no of apps belonging to each category .
* For every category there will be certain no apps .
* In this graph we will show total number of apps related to each category.
* First finding the sum of apps for every category
* Making a relation of category with each count by doing a data set
* Finally we plot the data set into a Bar graph

**AXIS : X = APP CATEGORY , Y = NUMBER OF APPS**

**Observation:** from the above graph we can see that apps belongs to family category has

Histogram

Description automatically generated with low confidence

* **Observation** : From the above graph we can clearly say that apps of family category are highest and of beauty are the lowest in the play store.

## **Content Ratings Given By Age Group:**

* Showing the trends of Content Ratings Given by Age Groups By plotting a line graph
* To make relation of content ratings and age groups
* Take unique of age group and the value counts of the age group in the content ratings and finally make a dataset of both the unique values of the content rating and value counts of the content rating
* Plot the dataset obtained into a line graph to see the highest content rating given by the age groups

**A picture containing Word

Description automatically generatedAXIS: X = AGE GROUPS , Y = MAXIMUM COUNT OF THE AGE GROUPS**

**Observation:**

* By seeing this graph we can clearly conclude that content ratings of group everyone is highest in the play store with 8000+ ratings and the ratings given by age group Adults only 18+ is least with only 3 ratings in the play store.

## Top 3 Most Reviewed Apps On Play Store:

* we got the apps in the play store and reviews are assigned to each of the apps.
* we here taking the top three apps which got higher rating by ascending the rating values in the play store dataset and assigning the attributes to the value counts to the names of the apps which are in concern with the ratings.

Chart

Description automatically generated with low confidence

**Observation:**

From the above graph it is clear that Facebook is the most reviewed app on the play store followed by WhatsApp and Instagram.

## Type Of Apps Which Are Free OR Paid

* Using sum counts we can calculate percentages .
* Plotting a pie chart to show the percentage of free and paid apps in the play store.
* Venn diagram

  Description automatically generated with low confidencepercentage of types of apps in play store

Chart, histogram

Description automatically generated

* **Observation** :
* In the following graph we have seen the various type of apps which are paid and free and from the Pie chart its clear that play store consists of most number of free apps with 93% and paid apps with just 7% and from the Bar graph its clear that that apps belons to family category are high and beauty apps are low in play store.

## Types of reviews GIVEN BY users in play store

* In play store dataset we got the sentiments given as positive , negative , null.
* Finding the total sum of the reviews belonged all positive , negative , null we will get the count of each sentiment.
* Plot the count of each sentiment to get the highest user reviews in play store and we can also get the least .
* By plotting the Bar graph we can visualize our data of sentiment with user reviews.

A picture containing bar chart

Description automatically generated

* **Values of types of ratings**

**Chart, sunburst chart

Description automatically generated**

* **Percentage of types of ratings**

**Observation:**

From the above graphs we can conclude the values of each content ratings and also we can say that play store has more number of positive ratings of 64% with 17500+ given by the users and has only 21% negative ratings in play store.

## Sentiment polarity (free vs paid):

plotting the sentiment polarity graph of free and paid apps.

Chart, box and whisker chart

Description automatically generated

**Observation:** From the above graph it is clear that sentiment polarity of free apps is higher than paid apps.

* SENTIMENT POLARITY FOR FREE GOES FROM -1 TO 1 SENTIMENT POLARITY FOR PAID GOES FROM -0.50 TO BETWEEN 0.75 TO 1 AND BOTH THE QUADRILES ARE RANGING FROM 0.00 TO BETWEEN 0.25 AND 0.50.

## Histogram Of Sentiment Subjectivity:

* Plotting a histogram graph to show high subjective scores.
* A diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

Chart, histogram

Description automatically generated

**Observation:** From the above graph we can say that subjectivity score is high between 0.5 and 0.6.

The lowest subjectivity score ranges from 0.1 to 0.2.

## Top 10 Apps With Highest Positive Reviews:

* Plotting a graph to see top 10 highest positive reviewed apps.
* For positive sentiment reviews for each app we have formed a relation between apps ,reviews and the sentiments.
* By taking value counts of total positive reviews and relating the value to the apps in the play store we get the relation between apps , reviews and statements.
* Showing these apps with highest number of positive reviews trend in the form of graph to visualize the data.

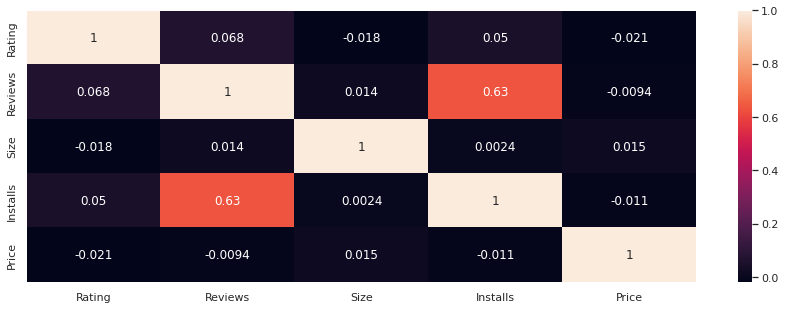
Chart, bar chart

Description automatically generated

**Observation:** From the above graph we can say that Color Notepad Notes has got highest positive reviews on the play store.

## Correlation Between Ratings , Reviews , Size , Installs And Price:

* Complete correlation between two variables is **expressed by either + 1 or -1**. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0.

****

**Observation:** From the above graph we can easily take out the correlation values between any two attributes in sense of all attributes .

* **correlation between reviews is 1 because reviews are proportional to reviews as they are same attributes.**
* **But for correlation between reviews and price is -0.0094 this because they are inverse in the play store data set son one increase another decrease .**
* **This is the correlation between the attributes in the play store dataset.**

# Conclusion

* We have reached at the end of our EDA. We have gone through both datasets and made them ready for the analyzation. Later we performed EDA by statistical measures and using visualizing tools.
* We have provided all important top performing features like - Category, Content Rating, Size and type which is required for the development of App. We also discovered apps which are having most comments of positive sentiment than negative sentiments and neutral sentiments.
* In overall data visualization we plot different type of graph (Pia chart, bar chart, line chart, box plot,) study the whole graph we conclude different segment like top most category lower most category and high customer review and different category according to age
  + we also go with pia chart and here we conclude the percentage of free apps and paid apps which is used the most by customer.
  + further go on these process we also try to do and play with dataset so we can merge the two dataset into one data frame and represent the correlation between them.
* Inspection and exploration of the data frames and checking the data type available in provided data frame.
* Executing various python function and different libraries operation to change data values and convert into required data form.
* Using various data visualisation function and methods for the projection of summarised information from the data frame via whom anyone can easily understand and explain the outputs.

Thank You