Google PlayStore review analysis

Pranit Ransingh

Sridhar Padhy

Data Science Trainees,

Alma Better, Bangalore

**ABSTRACT**

The Google Play store is one of the largest and popular Android application stores. The data in it is enormous and can be very beneficial if used properly. We have been provided with the dataset of Google PlayStore

Our main objective is to perform exploratory data analysis on the given data set and get appropriate conclusion about the ongoing trends and what types of problem can we solve with the given data set.

**DATASET: -**

In our day-to-day life Mobile application have become very useful and are developed to create easy in our life. They are used in almost every field being sports, business, social life, medical and many more. Then main reason of them being popular is almost 83% of the apps are free to download and are downloaded in mere minutes. It creates an ease around the people. The Playstore market has around 3.5 million apps and approx. 3000+ app is being developed very day. The total market has around 5 billion users which makes it a lot more popular. However as blindfolded making of app without analysing it or doing research on it can we too much time consuming can be less efficient. Which might lead to

failure of the project. Due to this EDA can be done on this dataset to give the developers a broad view about the market and DO’s and Don’t of the market to get the maximum out of it and to successfully develop a functional application.

**INTEGRAL METHODOLOGY**

The whole Analysis is divided into the following parts: Description, Breakdown of datasets, Finding and treating the null values & missing values, Data Cleaning after that we will do Exploratory Data Analysis

**DATASET DISCRIPTION**

Now let’s take a look at our data which has total of 2 csv files

Playstore.csv: This csv files contains information of the apps in the PlayStore being it having app name, category, size of the app, version of the app in total we have 13 variables in this data set

User review.csv: This dataset contains reviews of the app along with there positive, neutral and negative sentiments

Before going into data, we will analysis the following observation.

1. Getting the average Rating of the Apps.
2. Checking the count of application in each category
3. Getting the number of installs in each Category
4. Checking Corelation
5. Getting App Size Distribution
6. Getting the number of installs with respect to the size of the application
7. Let’s see the major Type of app distribution in PlayStore.
8. Let’s see the apps dependence on size and type and it effect on rating too.
9. Review sentiments in all the app dataset.
10. Let's see a more depth understanding of sentiment Polarity and Subjectivity.
11. Does sentiments Polarity is proportional to sentiments subjectivity.

**DATASET FEATURES**

There are total 13 feature in the Playstore.csv dataset. The feature include: -

* App: - Name of the App
* Category: - Category under which the App falls.
* Rating: - Application's rating on PlayStore
* Reviews: - Number of reviews of the App.
* Size: - Size of the App.
* Install: - Number of Installs of the App
* Type: - Whether the App is free/paid
* Price: - Price of the app (0 if it is Free)
* Content Rating: - Appropriate Target Audience of the App.
* Genres: - Genre under which the App falls.
* Last Updated: - Date when the App was last updated
* Current Ver:- Current Version of the Application
* Android Ver: - Minimum Android Version required to run the App

**EXAMINING NULL VALUES**

We saw and also got the information about the dataset. Now we have to clean that attributes which have not to be useful. Data Cleaning is one of the most essential part of data analysis. We have to Replace all the Null values with the Average of their Columns or with not null values.

**DATA CLEANING: -**

In this we will clean the dataset which will make it easy for us to make visualisation and analysis. The cleaning process in this dataset included removing the pre and post symbols around the values ad then converting it into numerical form for future use. Here the symbols we are mostly removing are $, M, +, K etc.

**DATA VISUALISATION**

With more than 1 billion active users in 190 countries around the world, Google Play continues to be an important distribution platform to build a global audience. This makes the hell lot of competition in App market.

And also, we can say that before starting any kind of exploration the data cleaning plays a vital role in result and accuracy.

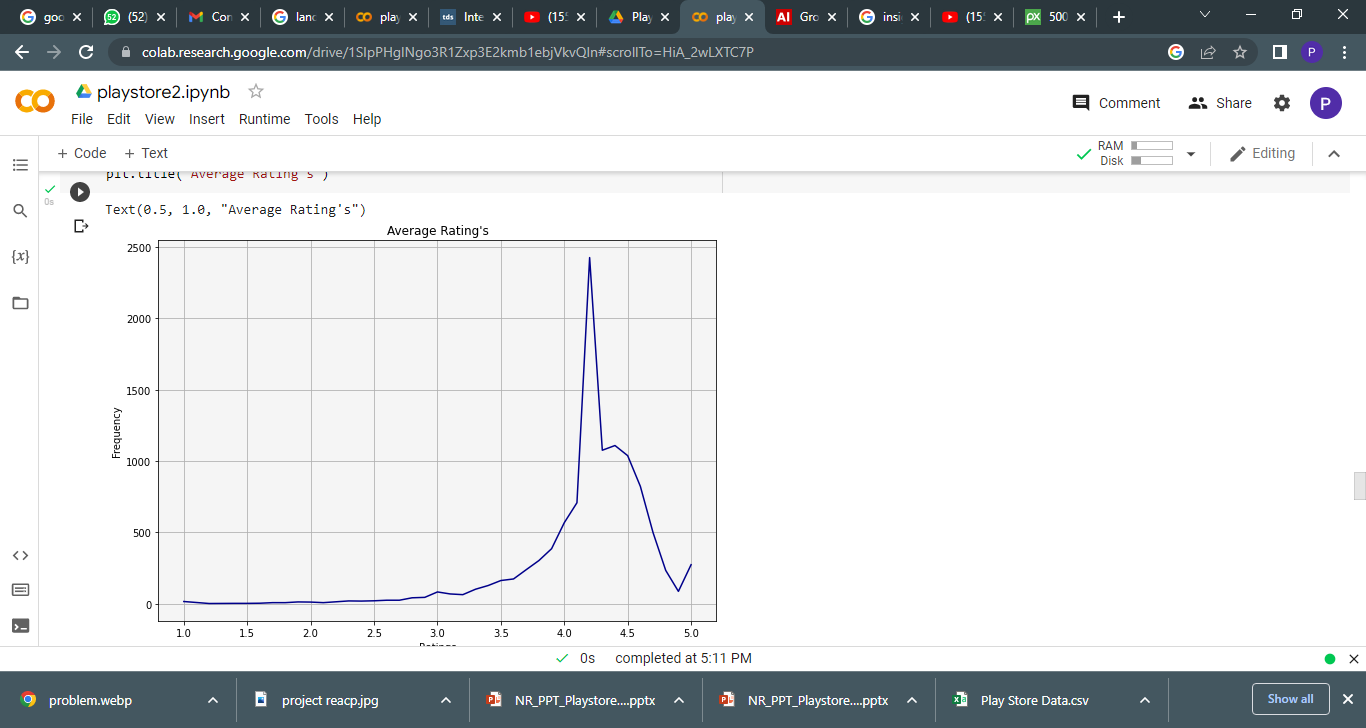
As we all can see the many observations and conclusion that arise from the data visualization and how visualizations make anything interesting with the graphs, plots, chart or maps.

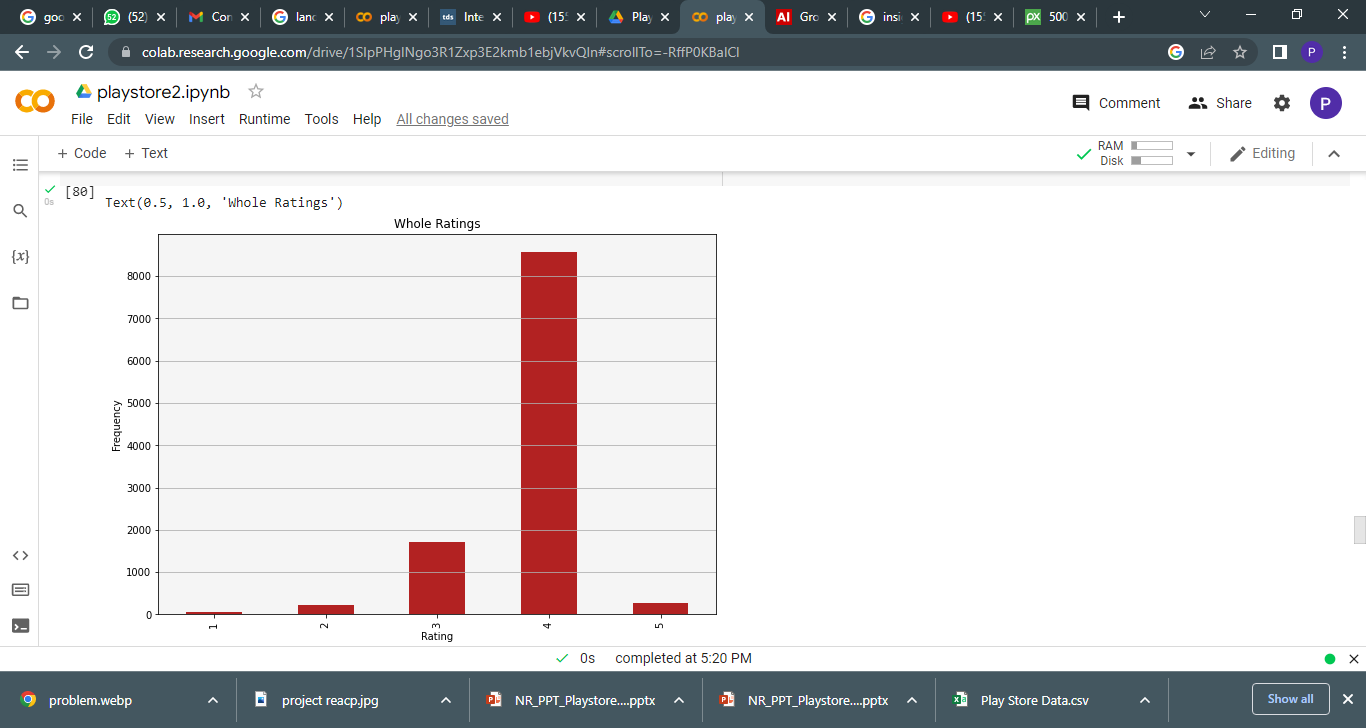
Which simplify the data and can be easily understand the role of the data and their elements in the diverse world of play store.

As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like GAME, COMMUNICATION, and TOOL even though the number of available apps from these categories are twice as much lesser than the category FAMILY. The trending of these apps is most probably due to their nature of being able to entertain or assist the user. Besides, it also shows a good trend where we can see that developers from these categories are focusing on the quality instead of the quantity of the apps.

**OBSERVATION 1**

The First visualisation consists of overall rating of the apps in the dataset. We will be finding out the average rating in the dataset.





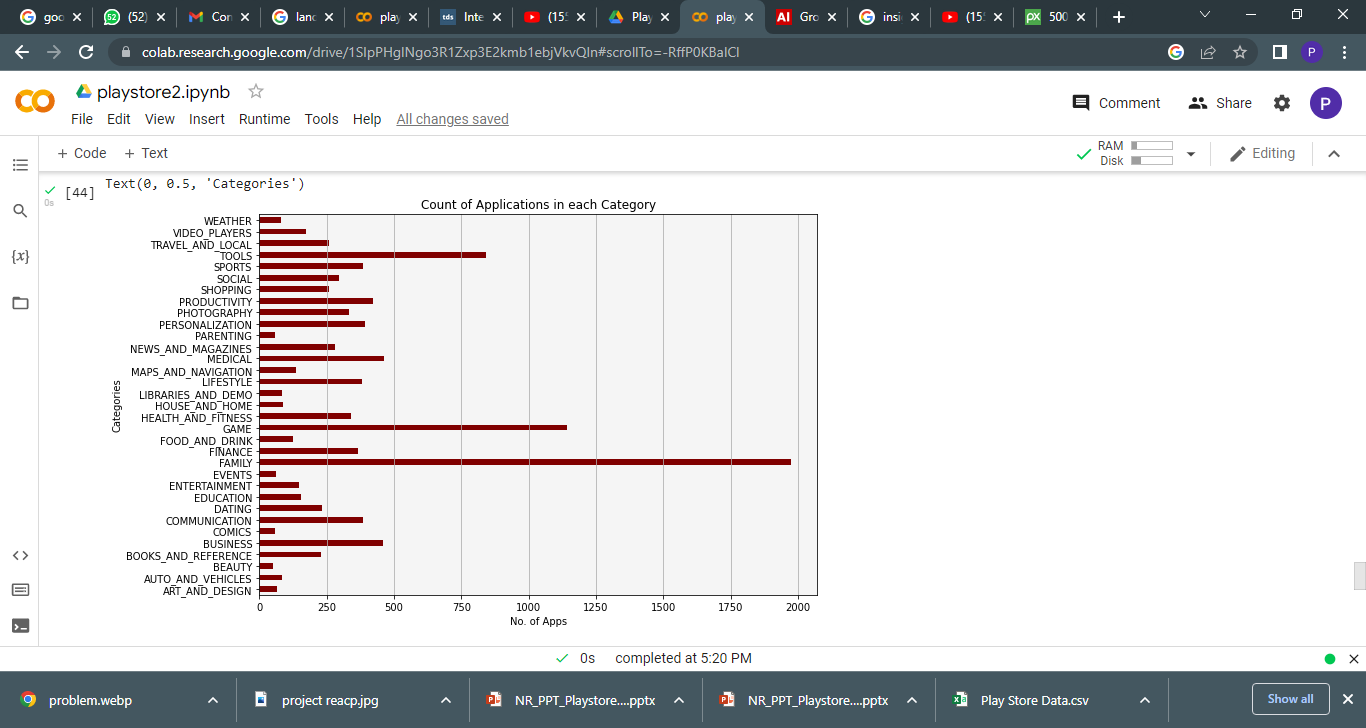
We can observe a gradual increase in the ratings in the graphical representation

We can see that the amongst all the apps most of the application have ratings between 4.1 to 4.4

Also, can be said that as the frequency of app increases the rating also increase or visa-versa.

**OBSERVATION 2**

The second observation consists of finding the total count of the applications in their categories.

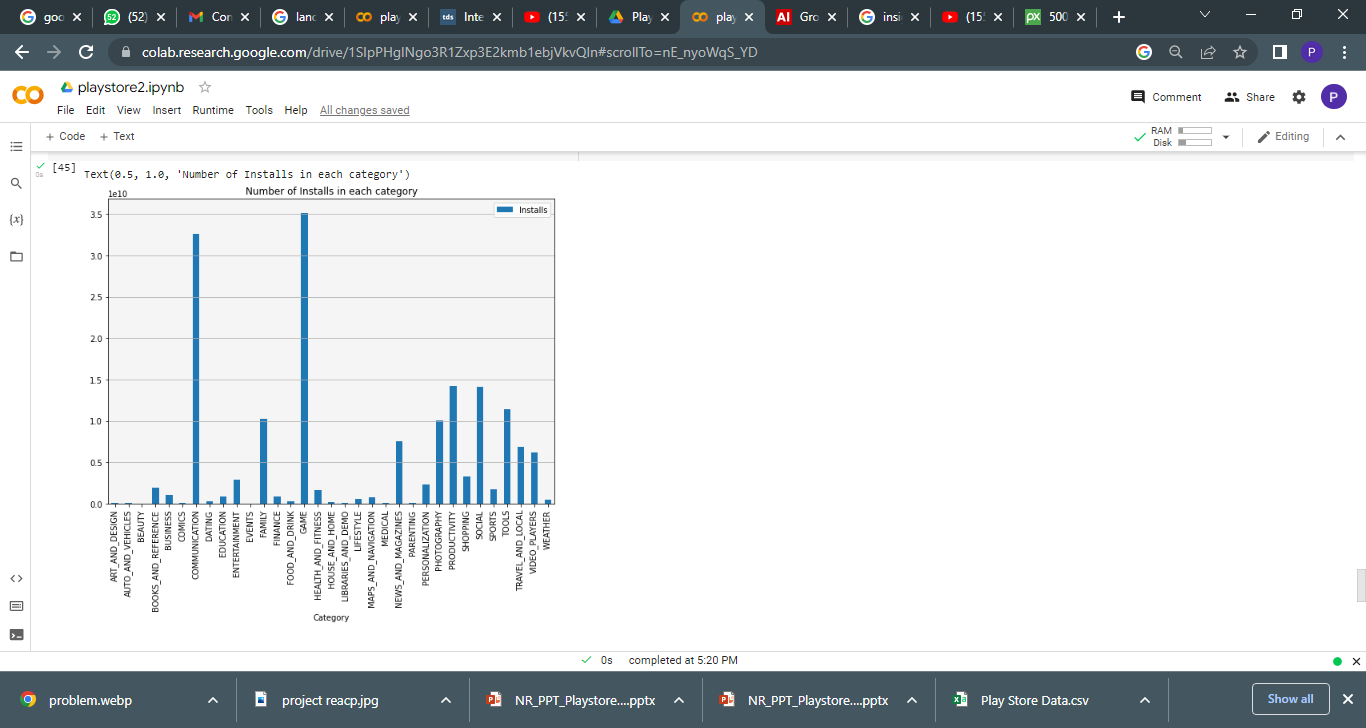


After the visualisation we can observe that we can see from the above visualisation the two categories Family and games have the highest number of applications to download from the play store.

And on the contrary, parenting, beauty, comics are few of the categories with least number of applications to download in play store.

**OBSERVATION 3**

After visualising the count of apps with respect to categories. We should also know the installs in the Playstore with respect to the categories present. We this observation we will get to know which categories has the greatest number of downloads.



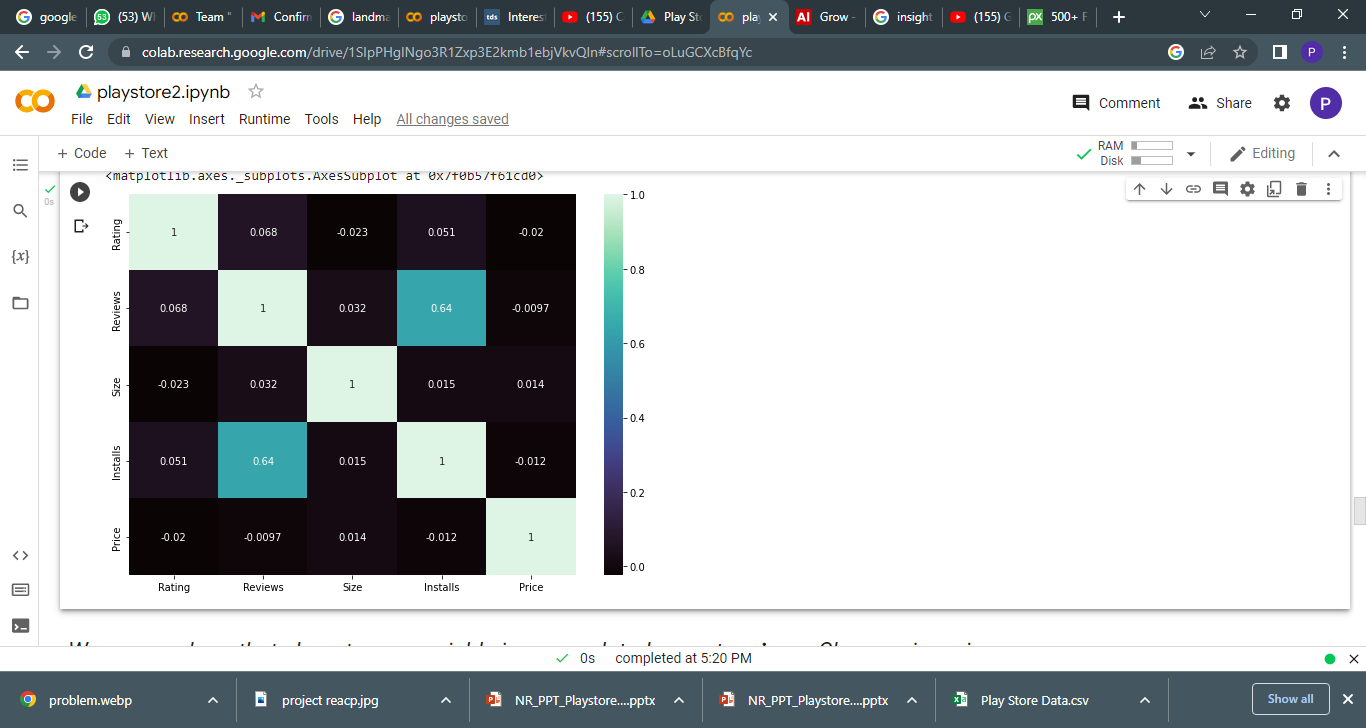
We can see here that the observation here is pretty different form that of count of the application. Form the data set acquired we can visualize and observe that although family had the highest number of applications in category, the installation of those app as very minimal.

We can observe that Games and Communication have the highest install rates, that is people are downloading Games and Communication application on a greater extent.

And as expected Beauty, comics are on an all-time lowest in installs along with events, medical and many more.

**OBSERVATION 4**

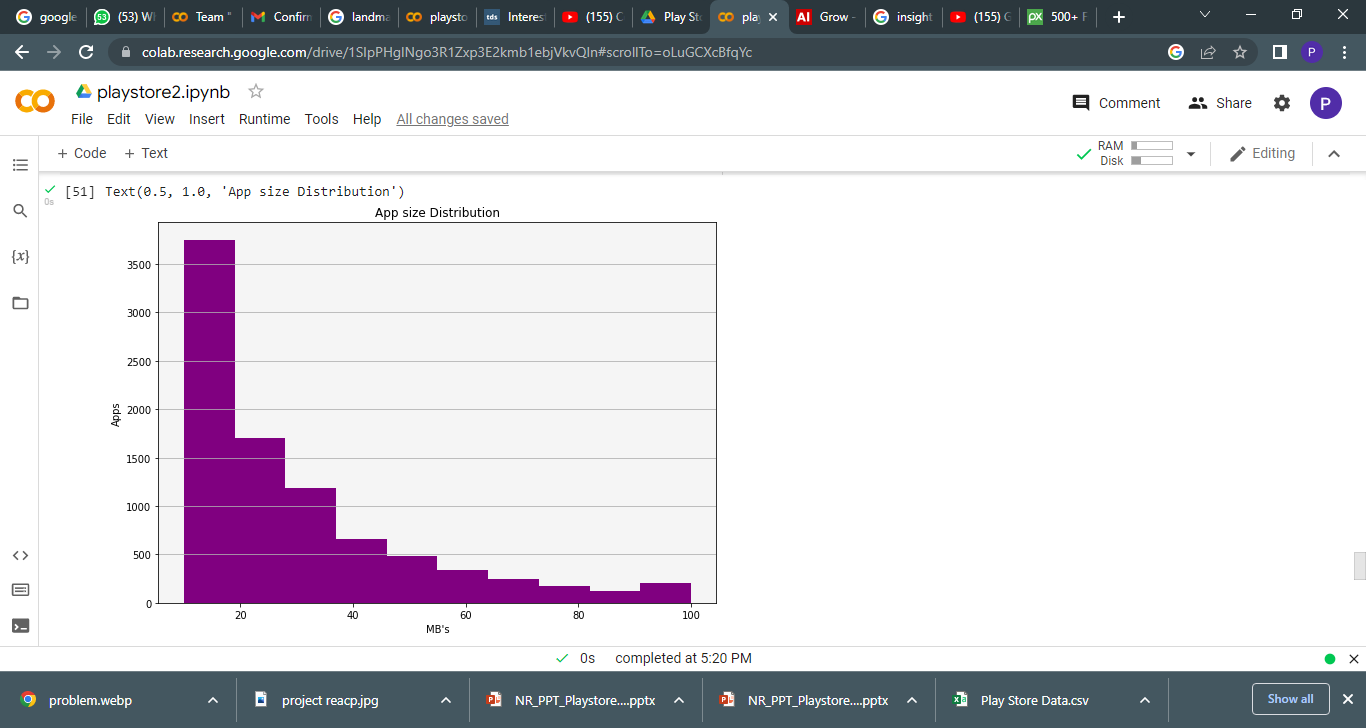
It is important for the developer and analyst to get the corelation graph. We the help of this graph we can understand the corelation between the features that is if the features are co-dependent on each other. This can create difficulty in visualisation.



We can see here that almost every variable in uncorrelated except **reviews**. Changes in reviews seems to have some effect on number of installs here.

**OBSERVATION 5**

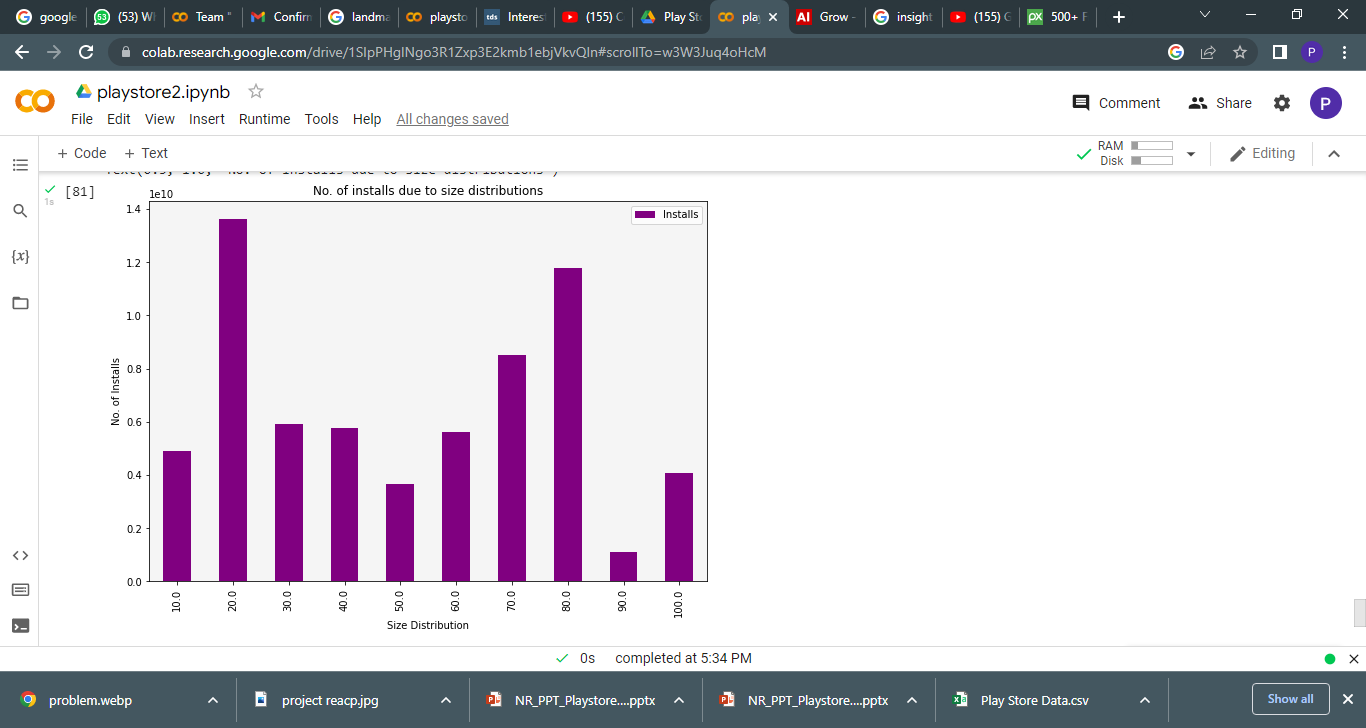
Next observation is about app size distribution. In this visualisation we will get to know about the different division of the size and the number of application present in those distribution



We can see in the visualisation that 10-20 mb size has highest number on application in the Play store and the number gradually decreases when the app size increases resulting in that there is majority application in 10-40 mb size range.

**OBSERVATION 6**

After we get the observation of the app size distribution the next step is to get the count of install in application with respect to size of the application.

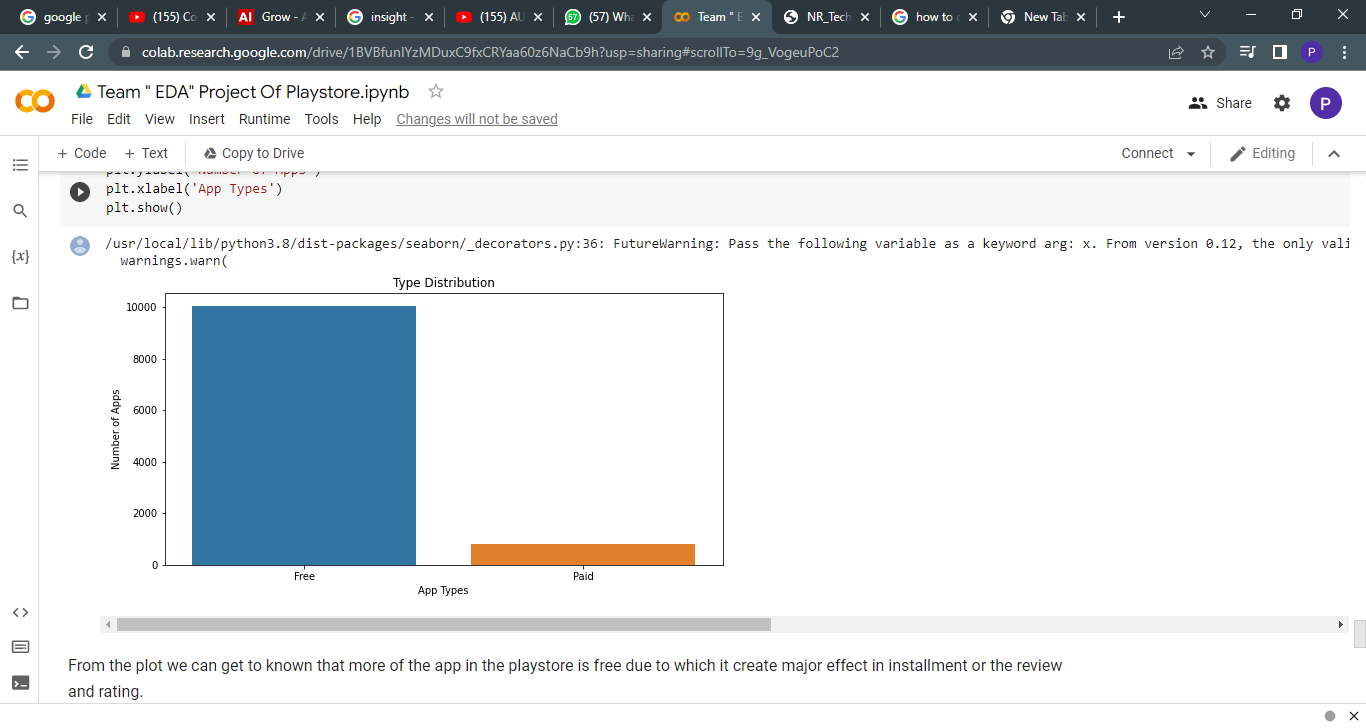


It is a quite interesting observation that although the app distribution in 10 mb bracket is on the higher side the installs is quite less here.

On the other side the 70 -80 mb bracket has higher installs and the 20mb bracket has the highest installs of all.

**OBSERVATION 7**

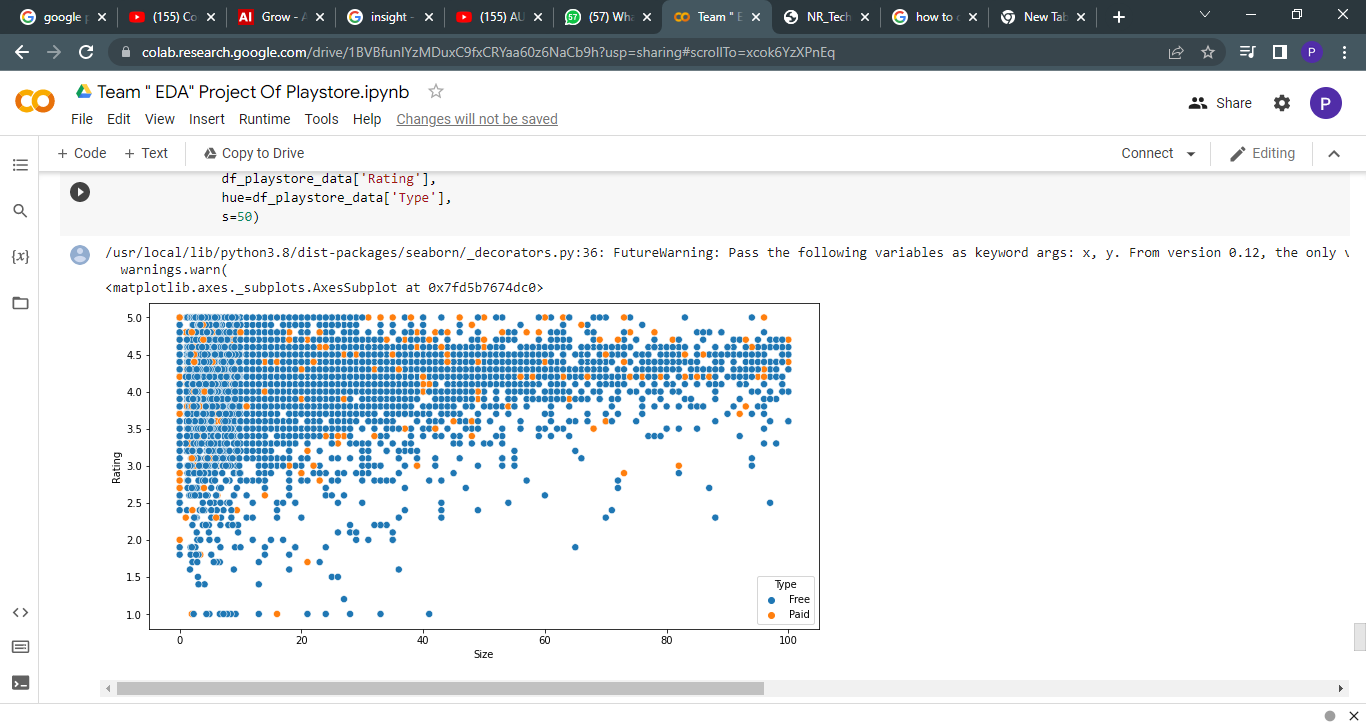
The application in the PlayStore are mainly divided into two parts Free and paid. Making the visitation of the free and paid app will give us the better understanding of the dataset.



From the plot we can get to known that more of the app in the PlayStore is free due to which it creates major effect in installation or the review and rating.

**OBSERVATION-8**

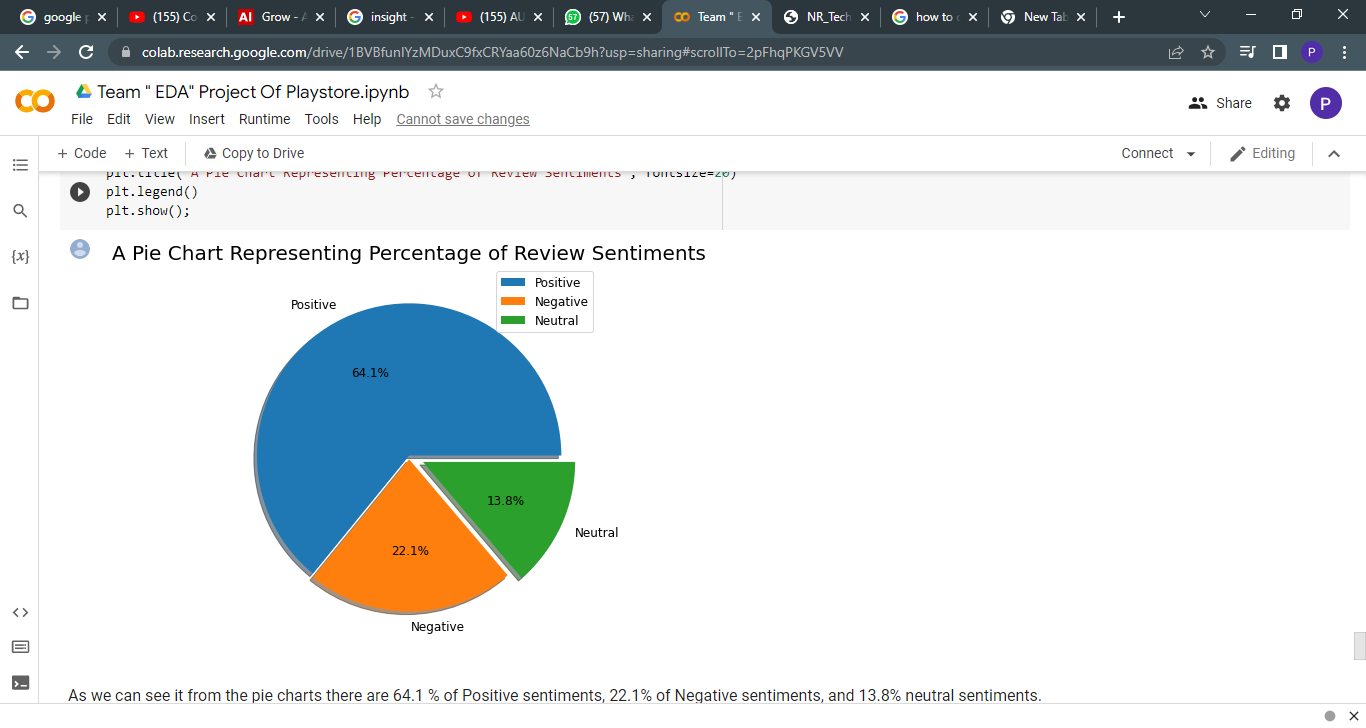
Now let’s see the apps dependence on size and type and it effect on rating too.



From this scatter plot above, we can conclude that majority of the free apps are small in size and having high rating. While for paid apps, we have quite equal distribution in term on size and rating. so, free app and small size app create much more effect than others.

**OBSERVATION-9**

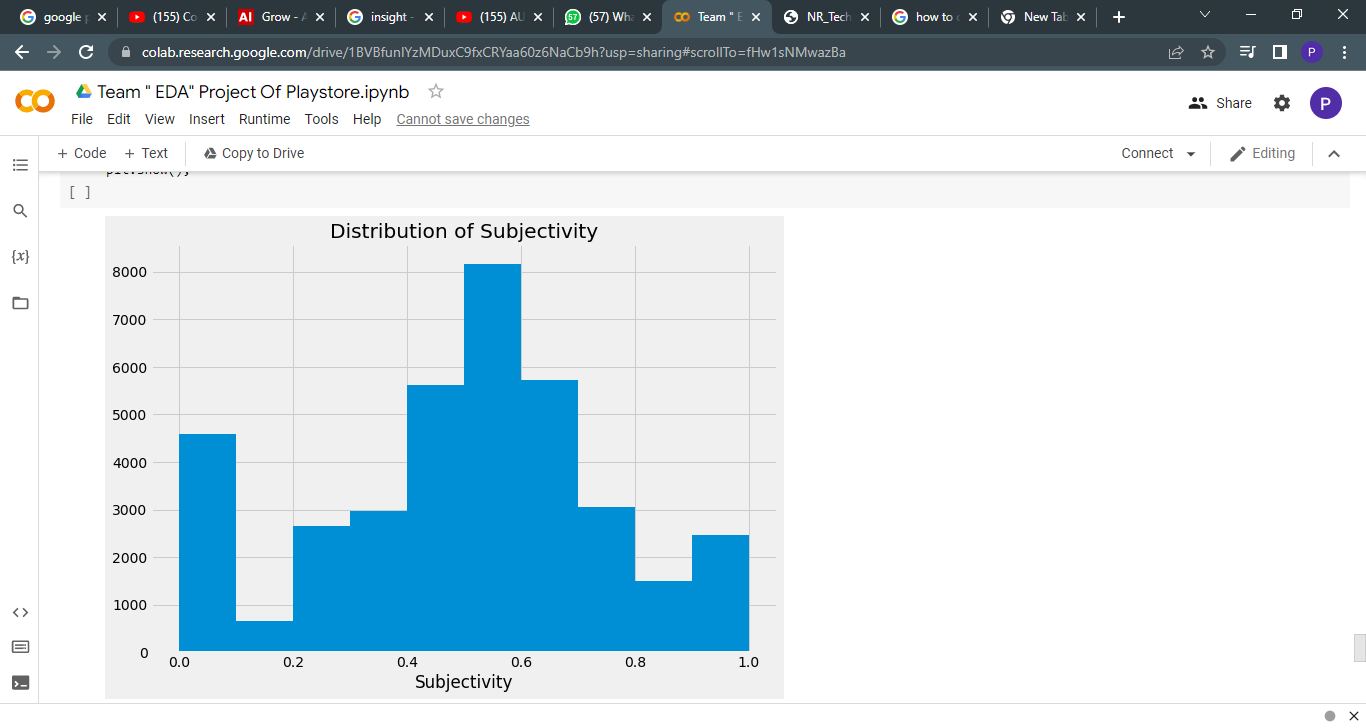
Now let us have a look at Review sentiments in all the app dataset.



As we can see it from the pie charts there are 64.1 % of Positive sentiments, 22.1% of Negative sentiments, and 13.8% neutral sentiments.

**OBSERVATION 10**

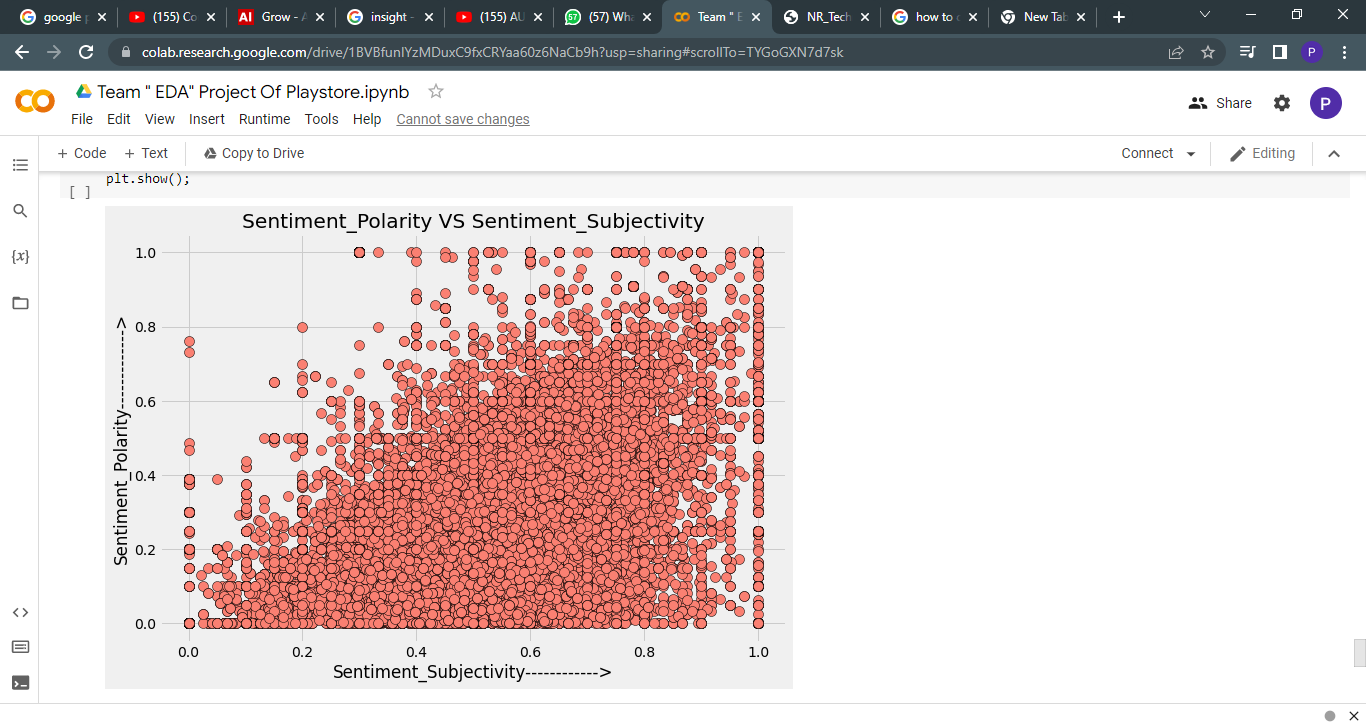
Let's see a more depth understanding of sentiment Polarity and Subjectivity.



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

**OBSERVATION-11**

Let’s have a look at our last observation which is sentiment polarity vs sentiment subjectivity



From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behaviour.

**CONCLUSION AND FUTURE WORK**

With more than 1 billion active users in 190 countries around the world, Google Play continues to be an important distribution platform to build a global audience. This makes the hell lot of competition in App market.

And also, we can say that before starting any kind of exploration the data cleaning plays a vital role in result and accuracy.

As we all can see the many observations and conclusion that arise from the data visualization and how visualizations make anything interesting with the graphs, plots, chart or maps.

Which simplify the data and can be easily understand the role of the data and their elements in the diverse world of play store.

As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like GAME, COMMUNICATION, and TOOL even though the number of available apps from these categories are twice as much lesser than the category FAMILY. The trending of these apps is most probably due to their nature of being able to entertain or assist the user. Besides, it also shows a good trend where we can see that developers from these categories are focusing on the quality instead of the quantity of the apps.

Other than this also can be say that mostly the app has higher number of user installs or review gives the app a good rating and mostly all the apps are around the rating of 4 (8000 apps).Also the size and price of the app create a minute difference but not majorly effect the app with good rating, review even if their app size are high. But due to free in charges we get the higher number of users installs and ratings.

Also, it can be seen from the chart that the sentiments play important role and maximum number of sentiments which are positive around (64.1 %) lies between subjectivity 0.4 to 0.7. From this, we can conclude that the maximum number of the audience give reviews to the applications, according to their experience.

As, we can see that, why exploring data is important before starting to build ML models.

As a conclusion, we learnt that the current trends in the Android market are mostly from these categories which assisting, communicating or entertaining apps.

**Thank YOU…!**