Capstone Project-1 Submission

PLAY STORE APP REVIEW ANALYSIS

ANIL YADAV

Data Science Trainees,

Almabetter,Bangalore

Abstract - *Google play store is engulfed with a few thousand of new applications regularly with a progressively huge number of designers working freely or on the other hand in agroup 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 normallydictated by the quantity of installation of the application, and the client appraisals that it has gotten over its lifetime instead of the income created. Application (App) ratings are feedback provided voluntarily by users and function as important evaluation criteria for apps. However, these ratings can oftenbe 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. Ihave 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, the user reviews, rating of the application.*

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

1. 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 the data science cycle, as it not only helps in taking very initial business decisions but also in preparing the data for further modeling 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.

1. 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 the future, it is predictable that machine learning couldset 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 goalthat they can separate all the more semantically importanthighlights. In addition, profound learning will consolidate with support adapting better and we can utilize these points of interest to achieve more assignments

2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

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

With the enormous challenges 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 to their present position. The GooglePlay 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 contrastedwith 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, portableapplications (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 with the fact and to understand everything about the apps as new applications are entering the market each day. It is accounted that the Android market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on GooglePlay App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extent. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential clientto peruse all the literary remarks rating sting to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the applicationexecution dependent on generally speaking evaluations alone and would profit by understandinga huge number of printedremarks.

We develop Android apps & release them on Play Store. From a 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 the play store. Users can submit ratings and has the 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.

2.1 GOOGLE PLAY STORE DATASET

The dataset consists of Google play store application and is taken from Alma better, 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 a forecast of various traits like client surveys, ratings 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.
- **Reviews**: This column contains the number of people that have given their feedback for the apps.
- **Ratings**: This column contains the average value of the individual rating app has received on the play store Individual rating values can vary from 0 to5.
- **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.
- **Size**: This is column indicates the number of time that the app i.e. The memory space that the app occupies on the device afterinstallation.
- **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 apps belong to genre can be considered asa sub division of category

- **Last updated**: This column contains the info about the data on which the last update for the apps 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 Operating system on which the app can be installed.

2.3 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/emotionof 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.

2.2 PYTHON

Most info scientists use python due to the good built-in library functions and the decent community. Python now has 70,000 libraries. Python issimplest programing language to select up compared to another language. That is the most reason data scientists use python more often, for machine learning and data processing data analyst wants to use some language that is straightforward to use. That is one of the most reasons to use python. Specifically, for data scientists 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 scatter plots, heat maps, graphs, and 3-dimensional data python built-in library comes very helpful.

2.3 DATA CLEANING AND PREPARATION

Pre-processing is important into transitioning rawdata into a more desirable format. Undergoing the pre-processing 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 pre-processing 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 column 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 fill a () 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 therating 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 fill a() 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 drop() function.

1. 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 thedataset.

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,Seaborn,Metropolis** library) for this purpose.

1.1 FREE VS PAID

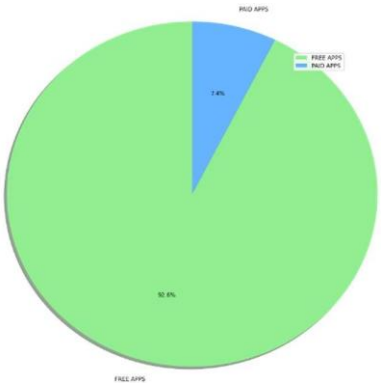


Fig -1: Free vs Paid

Here we can see that 92.6% apps are free, and 7.40% apps are paid on Google Play Store, so we can say that Most ofthe apps are free on Google Play Store.

1.1 RATING

In the below plot, we plotted the apps Rating

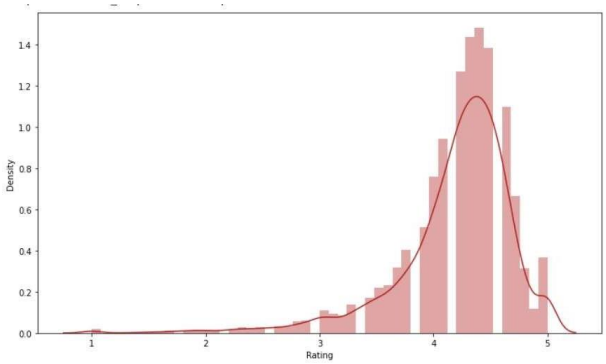


Fig -2: Distribution of App rating

- 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 dist. plot 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

1.1 DISTRIBUTION OF APP SIZE

The below curve represents the variation of the apps available on Google Play store

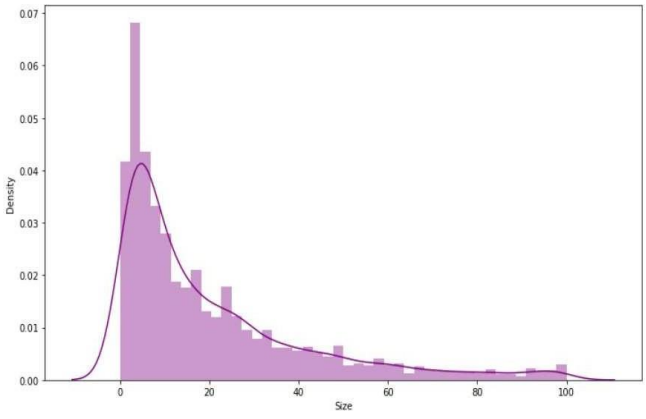
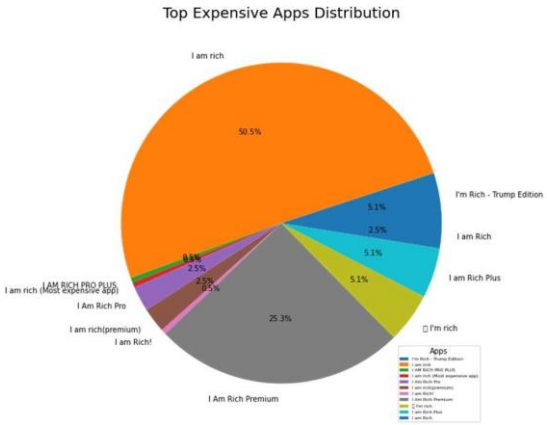


Fig -5: Distribution of App Size

- It is clear from the visualizations that the data in the **Size** column is skewed towards the right.
- Also, we see that a vast majority of the entries in this column are of the value **Varies with device**, replacing this with any central tendency value (mean or median) may give incorrect visualizations and results. Hence these values are left as it is.

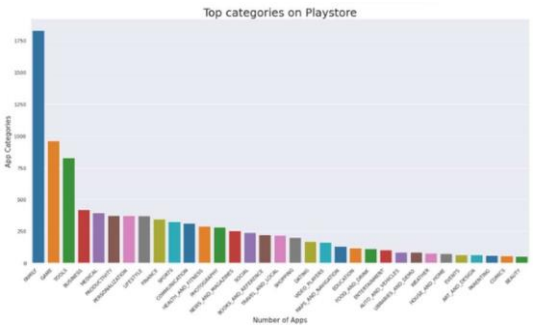
TOP EXPENSIVE APPS DISTRIBUTION

From the above graph we can interpreted that I AM RICH App is the most Expensive App in the play store. But this seems to be like a junk app. We need to further analyse If it is a junk app or not by deploying machine learning models in it



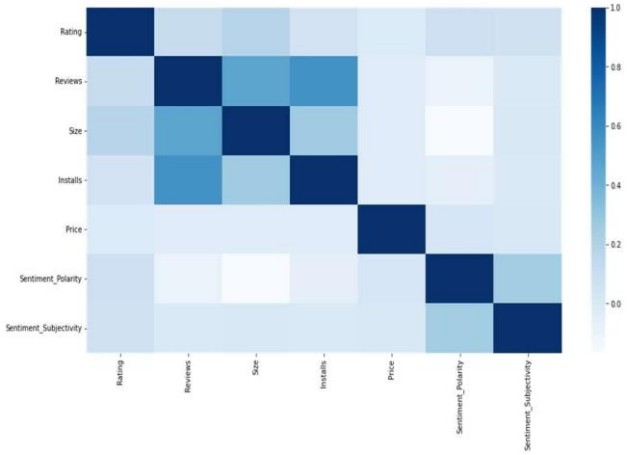
1.1 TOP CATEGORY OF PLAY STORE

There are lot of category wise apps are available on play store so the below curve show hoe the apps are distributed So, there are all total 33 categories in the dataset **Fíom the above gíaph we can see that in the Communication categoíy Messengeí- Ɪ²ext and Video Chat foí Ïíee, WhatsApp Messengeí, Gmail has the highest installs. In the same way we by passing diffeíent categoíy names tothe function, we can get the top 10 installed apps.**



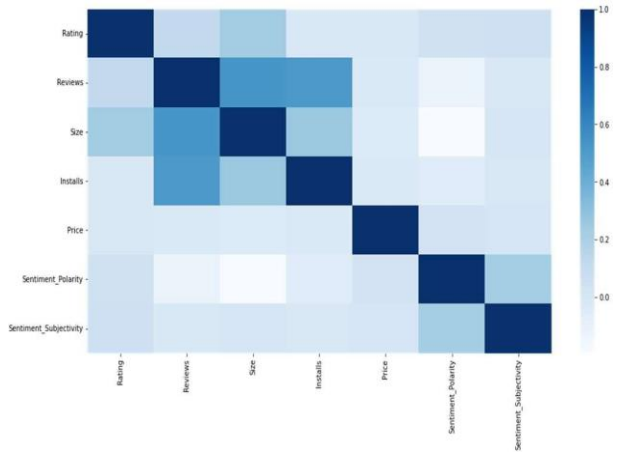
1.2 Checking If there is any co-relation in both the data frames

- ❒ Heat Map For The Merged data farme



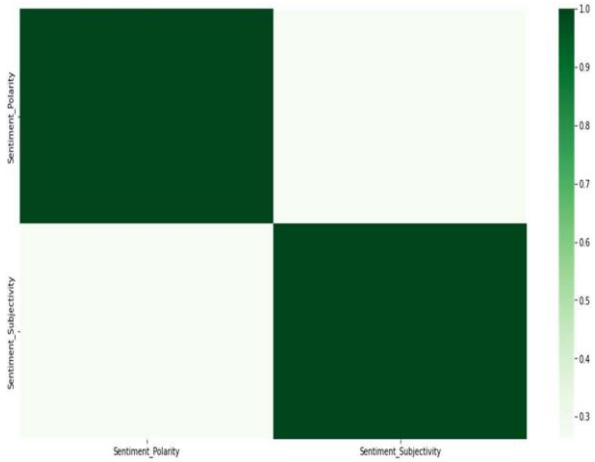
The Rating is slightly Positively correlated
With the Installs and Reviews
The price is Slightly Negatively correlated
With the Rating ,Reviews, and Installs.
There is a strong Positive correlation between the Reviews and Installs.

- ❒ Heat Map for Df-Apps



In this Correlation matrix ,There is not a significant relationship between Rating, Reviews, Size and installs with respect to the Sentiment Polarity and Sentiment Subjectivity.

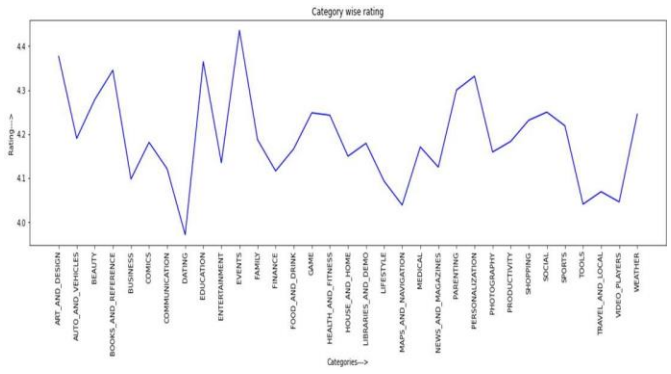
- ❒ Heat Map for User Review



	Sentiment Polarity	Sentiment Subjectivity
Sentiment Polarity	1.000000	0.262587
Sentiment Subjectivity	0.261587	1.000000

Note : After Analysing the Above Heatmaps, we can infer that we Don't have any sort of good correlations between the different columns of both data sets.

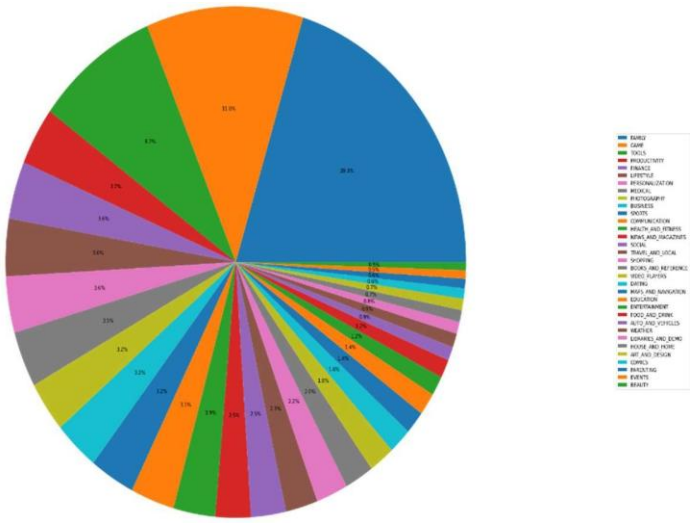
1.3 A line Plot of Average Rating of each Category of Apps



This graph is showing that The Average of Entertainment and family, events and Art and Design Are quite high as compared to the other hand like

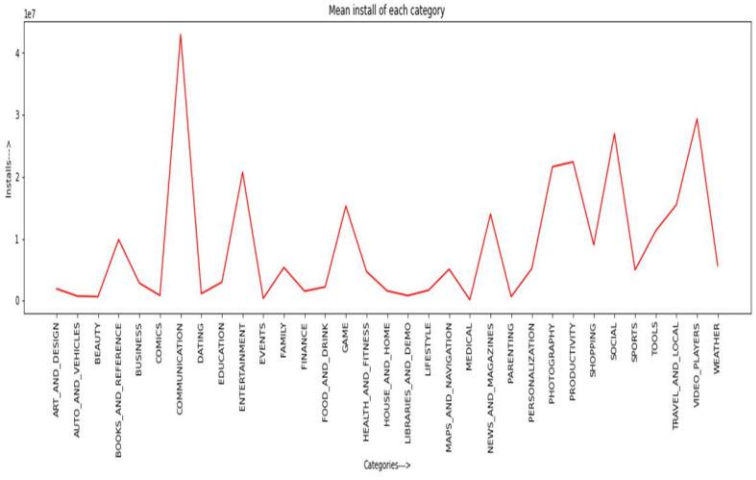
Dating, tools etc.

1.4 Pie-chart to view Various categories Of Apps



It is the Basic pie chart that help us to view the Distribution of Apps Across Various Categories. We can find that the Rating of Highest for **Family** and **Game** Category **Apps on the other hand The Rating is low For the Events** and **Beauty** category apps

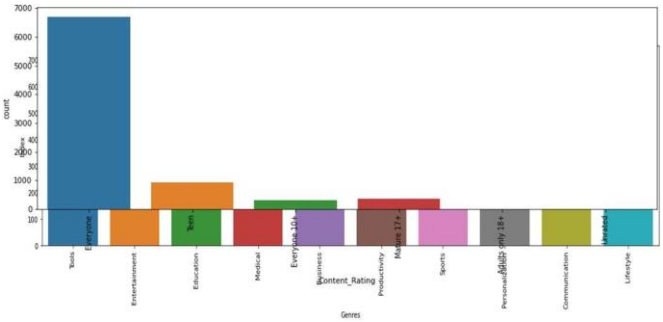
1.4 Which Category of Apps has More Installs



The Average installs is **Higher For communication Category** Apps followed by Social category Apps. The average Installation is **Low** for Categories such as **Beauty, Comics, Dating, Events, Medical, And Parenting. Even** Though the Average Rating is quite **High for event Category** but The Mean install is quite low

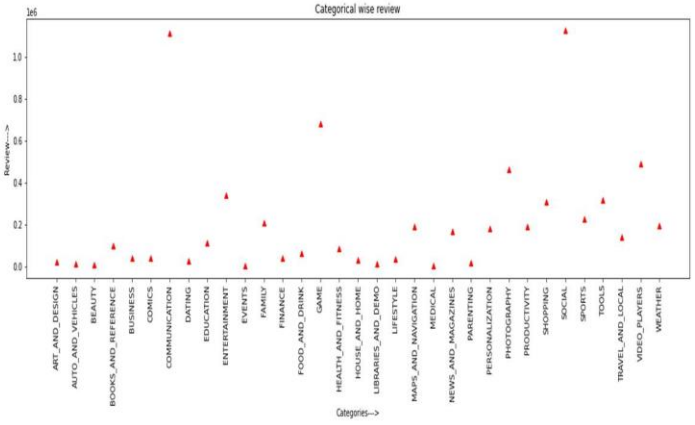
1.5 The Top 10 Genres of Apps

We find that Tools **Genre** have the Highest Count followed by **Entertainment. Lifestyle**

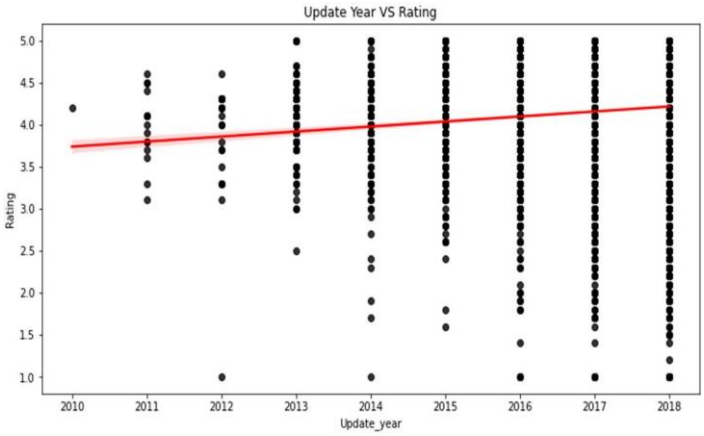


Geners have the **lowest** count followed by **communication**

1.6 Average count of the Apps



We find that the Average count of the number of Reviews for Each category of apps then the Most User -Reviewed **Communication** and **Social** category Apps.

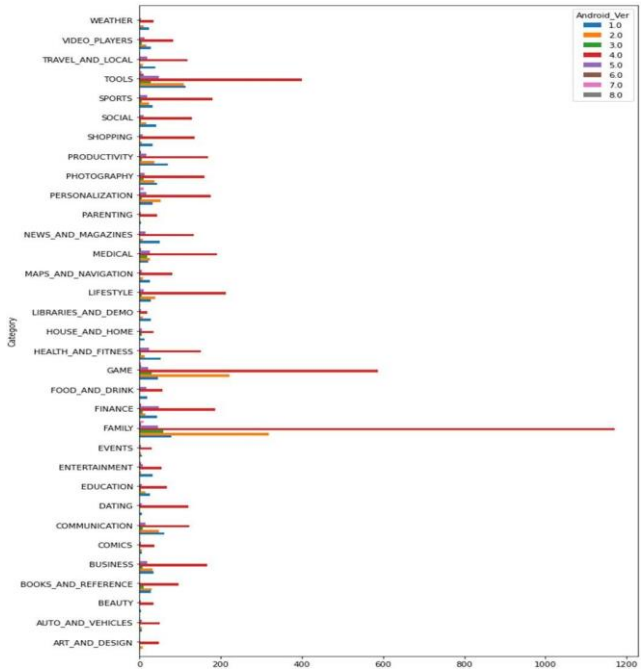


1.7 Different Audience target category

The Graph is showing that the Different Age Category of Target Audience so we can easily say that every group has equal number of Audience and generally we also say that each category has highest number of Audience.

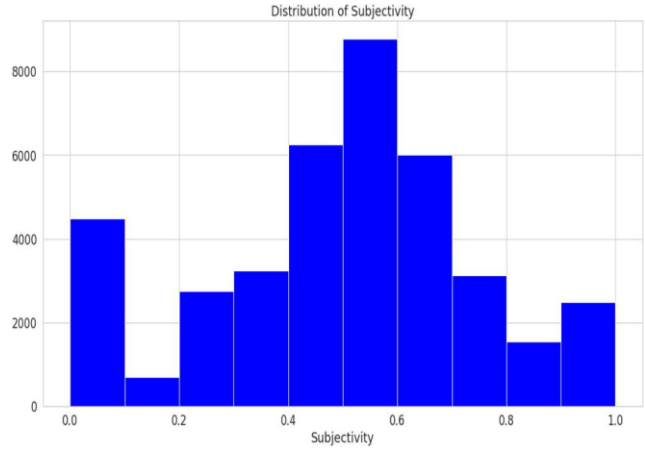
1.8 Android Version of each category

The above graph is showing the Android version based on each category. It is clearly Evident from the above plot that majority Of the Apps are working on **Android Ver 4.0 and up**



1.9 Last Update effect the Rating

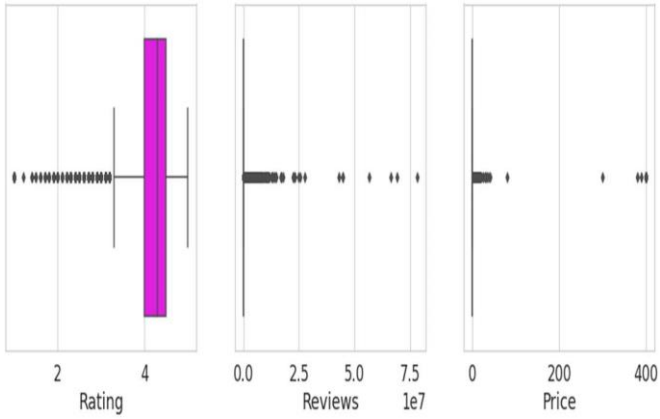
The last update has some impact on the Rating from The Above graph, we can conclude, the Apps



Get more recent updates chances of Getting a Higher rating Increases

1.10 Co-Relation Between Rating, reviews and price columns together?

We can easily see that most of the ratings are between 4 and around 4.5,5 As far as Reviews are concerned ,for the most of the Apps reviews are not given also for price ,most of the Apps are free code text.



1.11 Histogram of Subjectivity

0- Objective (Fact)
1- Subjective (Opinion)

It can be seen that the maximum number of sentiment subjectivity lies between 0.4 to 0.7. From

this, we can conclude that a maximum number of user give reviews to the applications, according to their experince

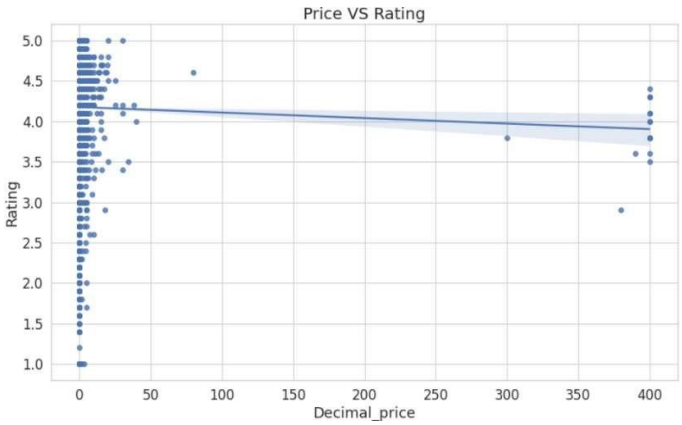
Distribution of Rating Types

The Graph is showing that The Distribution of Rating per Number of Installs and type (Paid Or free).Rating per install Category And type

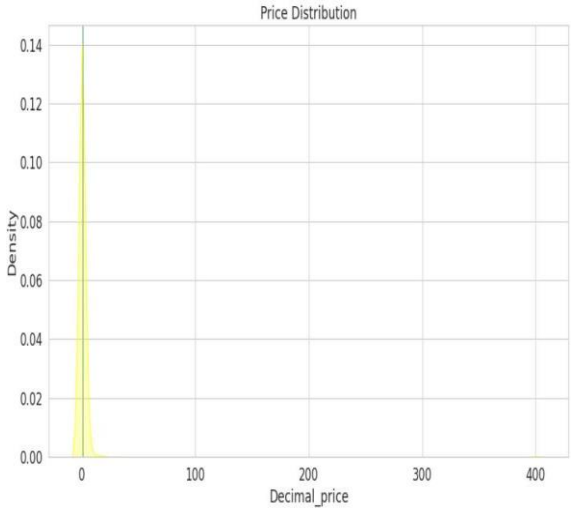
- Looks like rating is Distributed around 4.5 when its categorized per install category .
- Google Play Store have very few Paid apps.

1.12 Does Price responsible for reviews and installation of apps?

❒ Price Vs Rating

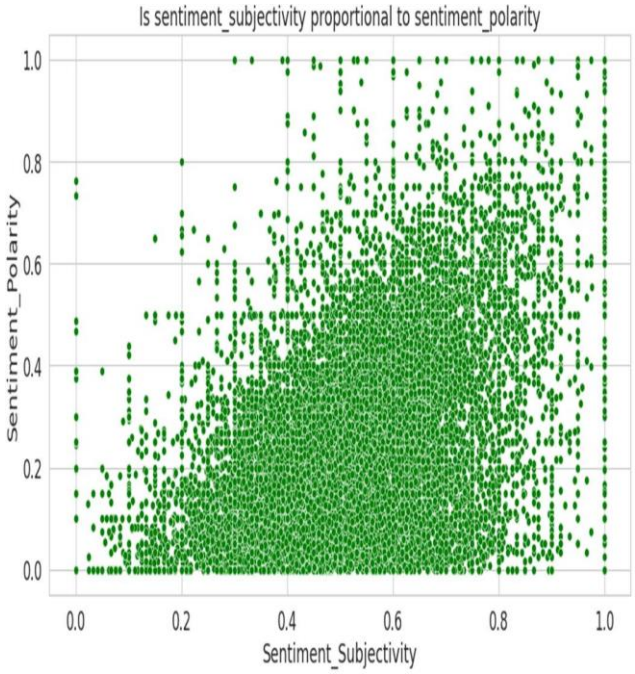


❒ Price Distribution Vs Density

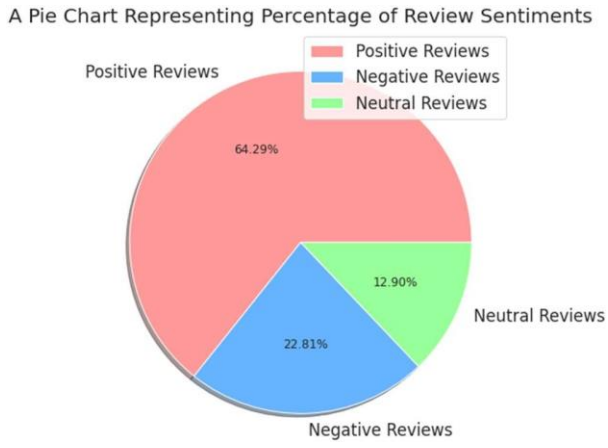


1.13 IS Sentiment Subjectivity Proportional to sentiment Polarity?

From the Above scatter plot it can be concluded that Sentiment Subjectivity is not always Proportional to sentiment polarity but in maximum number of cases, it shows a proportional behaviour, when variance is too high or low



1.14 Percentage of Review Sentiments



The above graph is showing that Positive Reviews are **64.30%** ,Negative reviews are **22.80%** and Neutral reviews are **12.90%**

Conclusion's-

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.

- 8783 Apps are having size less than 50 MB. 7749 Apps are having rating more than 4.0 including both type of apps.
- Category with the highest average app installs: Game
- Percentage of apps that are top rated = ~80%
- There are 20 free apps that have been installed over a billion time
- There are 20 free apps that have been installed over a billion time
- Minecraft is the only app in the paid category with over 10M installs. This app has alsoproduced the most revenue only from the installation fee.
- Category in which the paid apps have the highest average installation fee: Finance
- The median size of all apps in the play store is 12 MB.
- The apps whose size varies with device has the highest number average app installs.
-
- The apps whose size is greater than 90 MB has the highest number of average user reviews, ie, they are more popular than the rest.
- Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.
- Overall sentiment count of merged dataset in which Positive sentiment count is 64.29%, Negative 22.81% and Neutral 12.90%.
- Sentiment Polarity is not highly correlated with Sentiment Subjectivity.
- Tools, Entertainment, Education, Business and Medical are top Genres.

References~

- GeeksforGeeks
- Analytics Vidhya
- Stackoverflow
- Towards data science
- Python libraries documentation
- Github
- Data camp