Play Store App Review Analysis

Sakshant Gongal
Dhawal Khandiat

1. Abstract :

In this paper, we focus on analysing Google Play, the largest Android app store that provides a wide collection of data on features (ratings, price and number of downloads and descriptions related to application functionality. The overall objective of this analysis effort is to provide in-depth insight about intrinsic properties of App repositories in general. This allows us to draw a comprehensive picture of the current situation of the app market in order to help application developers to understand customers' desire and attitude and the trend in the market. To this end, we suggest an analysis approach which examines the given collection of Apps in two directions. In the first direction, we measure the correlation between app features while in the second direction we construct clusters of similar applications and then examine their characteristics in association with features of interest. In our analysis results, we identified a strong correlation between price and number of downloads and similarly between price and rating etc. We also determined that there is a high competition between App providers producing similar applications.

2. Problem Statement :

In this project we have focused on several different attributes of a given application like application name, category, rating, reviews, size, installs, type, price, content rating, genres, last updated, current version, android version. And to find out the most rated and most reviewed apps and also to distinguish between the apps which are either free or paid. We have taken the dataset and observed it nicely and as per our need we have taken various attributes to analyse and further display the result. By doing this, we can clearly and easily observe the dataset. Moreover, I will analyse different attributes given in the dataset.

3. Introduction :

Mobile applications are one of the fastest-growing segments of downloadable software application markets. Out of all of the markets we choose Google Play store due to its increasing popularity and recent fast growth. 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 3.48 million Apps and 226,500 unique sellers in April 2021. This rapid market has, in turn, led to over 500 million users downloading around 40 billion Apps all over the world. Developer and users play key roles in determining the

impact that market interactions have on future technology. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users. In this article, we seek to shed light on the dynamics of the Google Play Store and how we can use different features from this data set for prediction purposes. Using feature extraction from a longitudinal app analysis

will be used to find whether an app will be successful or not. Our Analysis is divided into four phases: data extraction, data cleaning, data visualisation, and EDA. In the first step, we try to do data cleaning on the data set to reduce the error percentage. After the data set is ready, we try to analyse the data set using different plots and remove the stuff not needed from the data set. The last step includes Exploratory Data analysis and visualisation. Finally, we narrate the analysis results to provide a clear vision of the relationship among the areas of interest. We include a detailed discussion of the applicability and future research directions in the last section called Conclusion.

4. Methodology :

Before we start exploring our dataset, look at our analysis approach and steps that are involved in performing EDA:

Importing libraries- First, we imported all the python libraries required for this, which include NumPy for numerical calculations, Pandas for preparing data and Matplotlib and Seaborn for Data Visualization.

Loading the data into data frame- We read the CSV into a data frame and pandas data frame does the work for us. This is one of the most important steps in EDA. Discover and access the data- After loading the data, it is important to discover the dataset. This step is about knowing the data and understanding what has to be done before the data becomes useful. Checking the first rows, last rows, shape of the dataset, columns and their data types are the basic things to look for.

Data Cleaning- Cleaning up the data is the most challenging and time-consuming part of EDA, but it's a crucial step for removing faulty data and filling up missing values. It is important to handle missing values effectively, as they can lead to inaccurate inferences and conclusions.

Data Visualization- Data Visualization is the graphical representation of information and data. It can help in interpreting and understanding the data, identifying trends, highlighting important relations with the help of charts and graphs.

5. Data Overview :

Before performing any operation on the dataset, it is important to understand the data at a high level. Depending on size and type of data, understanding and interpreting data sets can be challenging. After loading data, we observed the dataset by checking a few of the first and last rows. We checked the shape of the dataset and identified that there are 10841 apps(rows) and 13 features(columns) in our dataset. We observed that there are different types of data present in the dataset such as float, string, object. There are categorical variables as well as numeric variables present in the dataset. Each row of the dataset has values for category, rating, reviews and more app features. Here
are the columns of our dataset:

- App - name of the application.
- Category - category of the app.
- Rating - app's rating out of 5.
- Reviews - number of the app's reviews.
- Size - size of the app.
- Install - number of installs of the app.

- Type - whether the app is free or paid.
- Price - price of the app.
- Content Rating - target audience of the app.
- Genres - genre the app belongs to.
- Last Updated - date the app was last updated.
- Current Ver - current version of the app.
- Android Ver - minimum Android version required to run the app.

Now that we got familiar with the dataset, deciding from where to start is the important thing.
Checking the data for missing values is usually a good place to start.

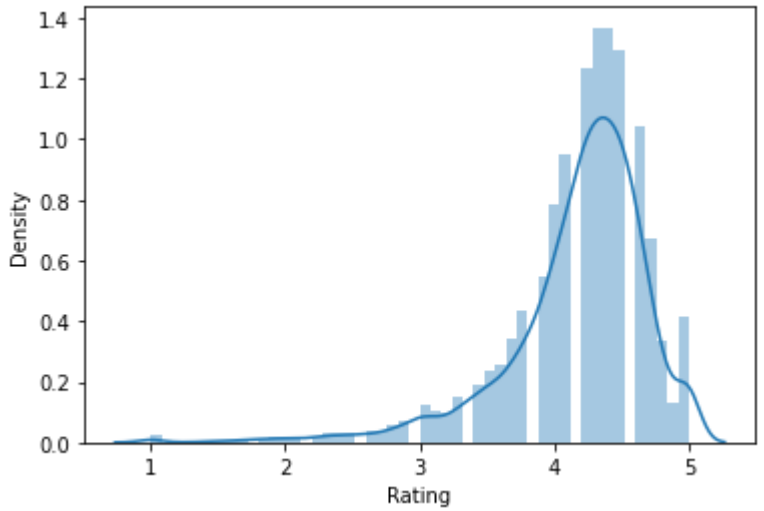
6. Data Preparation :

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is the most important step before performing any analysis. Data preparation usually includes handling missing values, standardizing data formats, enriching data and/or removing outliers. Good data preparation allows for efficient analysis, limiting errors and inaccuracies that can occur to data during processing.

6.1 Treating Missing values :

We started with checking for duplicate data because a huge dataset as in this case contains more than 10, 000 rows often has some duplicate data which might be disturbing. We found that this dataset contains some duplicate values so we removed all the duplicate values from the dataset. Before removing we had 10841 rows of data but after removing the duplicates 9660 rows meaning that we had 1181 rows of duplicate data.

Further we checked for missing values and observed that our dataset contains nearly 1477 null values. We tried to figure out why the data was missing. This is the point at which we get into the part of data science. It is the frustrating part of data science, especially if you are newer to the field. We analysed features one at a time and noticed that the Rating column has 1463 null values which is 99% of total missing values. The other four columns had null values which are less than 10. We decided to fill the missing values of the Rating column because dropping 1463 rows would have affected our analysis. Rating column contains values of the apps rating given by the users so it was not possible for us to predict any value to fill up. We tried to understand the distribution of ratings of other apps from the below distribution plot.



From the above plot, we observed that the distribution of the rating interprets negatively skewed distribution as more values were concentrated towards the right side of the plot. We looked at a few techniques of handling the missing values and decided to replace missing values with median value. In skewed distributions, the median is the best measure because it is unaffected by extreme outliers or non-symmetric distributions of scores. We also analysed other columns one by one and treated the missing values in the best way possible. Sometimes there would be many columns that we never use in such cases dropping is the only solution.
In this case, the columns “Android Ver” and “Current Ver” doesn’t make any sense to us so we just dropped them for this instance.

6.2 Transform Data :

Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood. Here we checked for the datatypes because sometimes the numeric variables are stored as a string. If this is the case, string data is to be converted into integer data only then data can be plotted effectively. In our dataset also we observed the same. Columns like “Size”, “Installs”, “Reviews” and “Price” are numeric variables but their data was stored as a string. So we converted string into integer data after performing some processes for removing special characters present in the values.

“Last Updated” column contained dates and its datatype was string, so we converted it to datetime format.
We noticed the “Installs” column ranges from 0 to 1,000,000,000 and even more i.e it had very high variance and also was highly skewed. It was difficult to plot any correlation with other features. In order to make this data effective we decided to do feature transformation.
The Log Transform is one of the most popular Transformation techniques out there. Log transformation tends to be used most often on skewed distributions. It helps in reducing skewness of the skewed data. In this transform, we took log of values in a column and created a new column. We also performed this transformation for the “Reviews” column as the data had high variance.
After cleaning and preparing data we had a dataset of 8432 apps(rows) and 11(columns). There was no missing data and all the features had their respective datatypes.

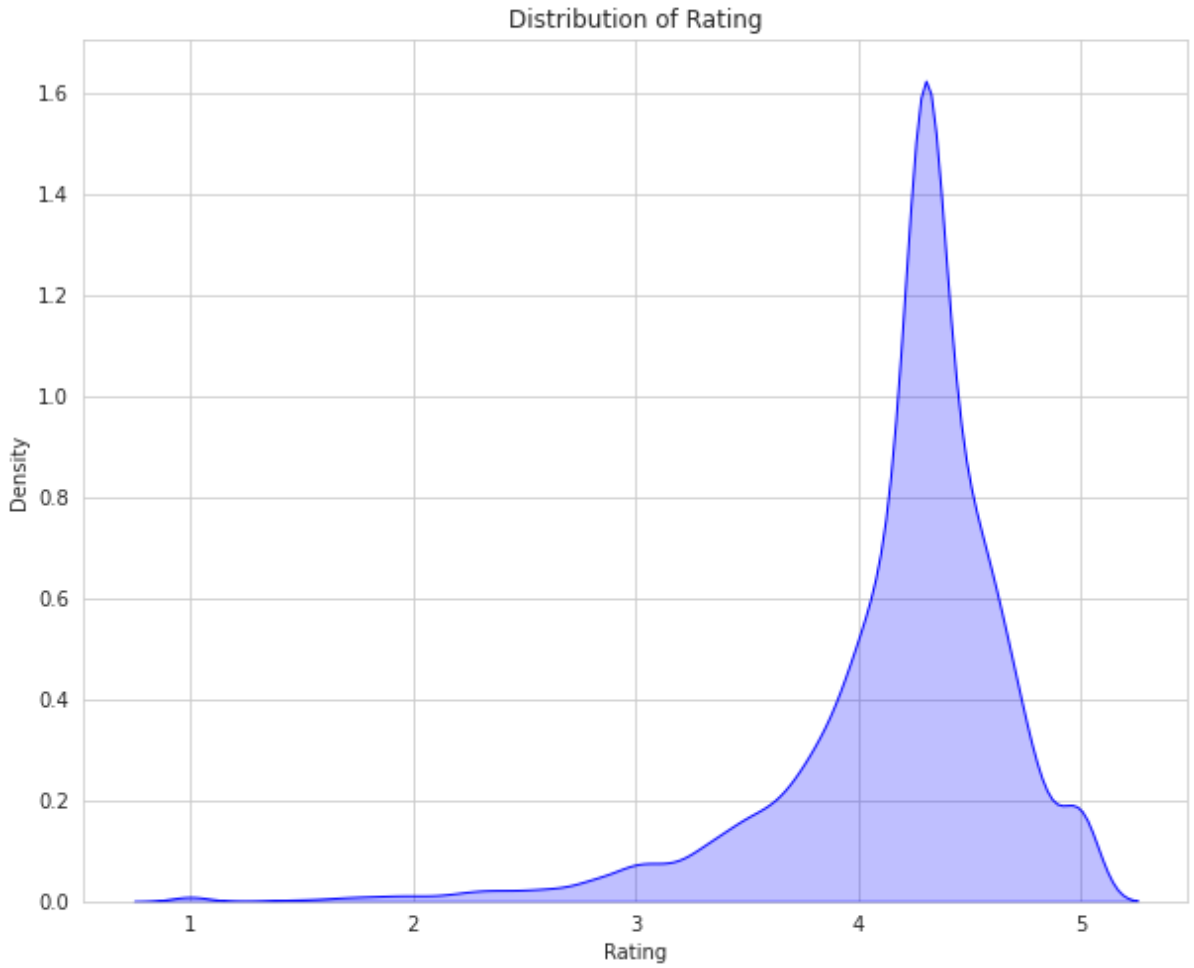
7. Data Visualization :

Now it is time to unveil the real strength of data analysis, i.e., to get an insight, and learn the trend, pattern and get answers to some of the questions related to the dataset. This process helped us figuring out various aspects and relationships among features of the app.

7.1 Data Insights :

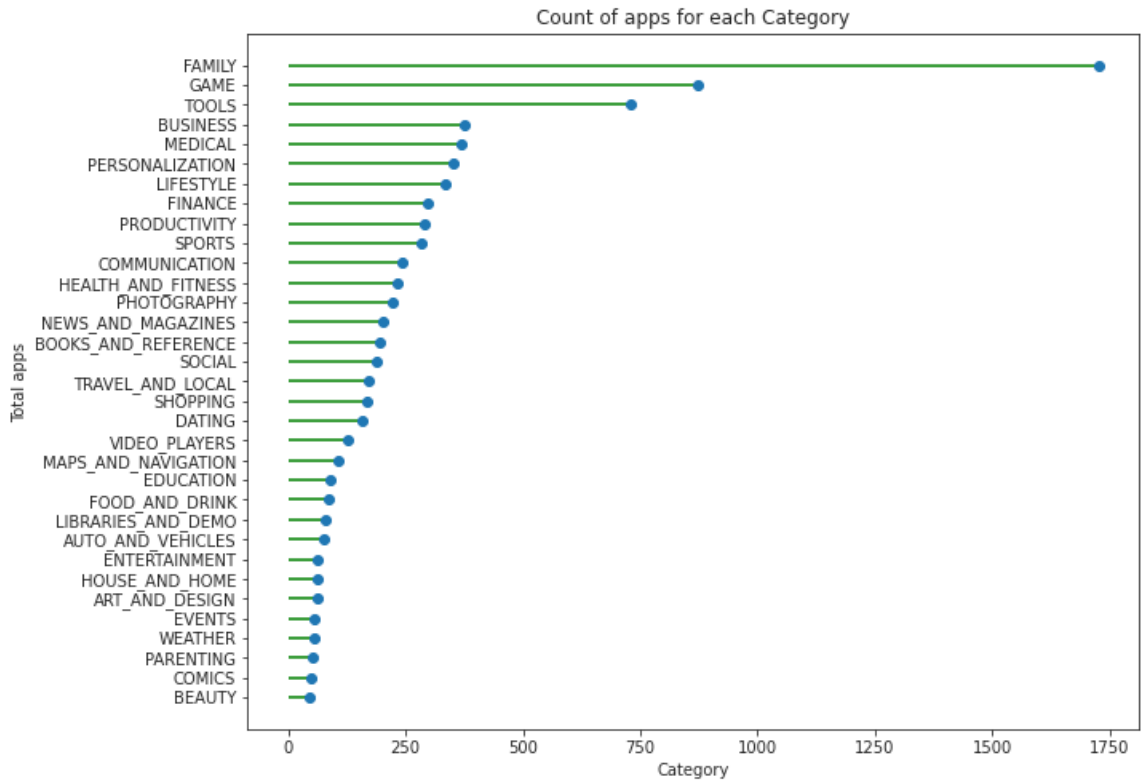
7.1.1 How is the app rating distributed?

App’s rating plays an important role in attracting new users. Research shows that 50% of mobile users won’t consider an app with a 3-star rating. So as an app developer or marketer, you don’t have the luxury of ignoring ratings. They matter quite a lot. Above distribution plot shows that most of the apps in the play store are rated between 3 to 5. We can see that the average rating is quite high, around 4.2 out of 5.



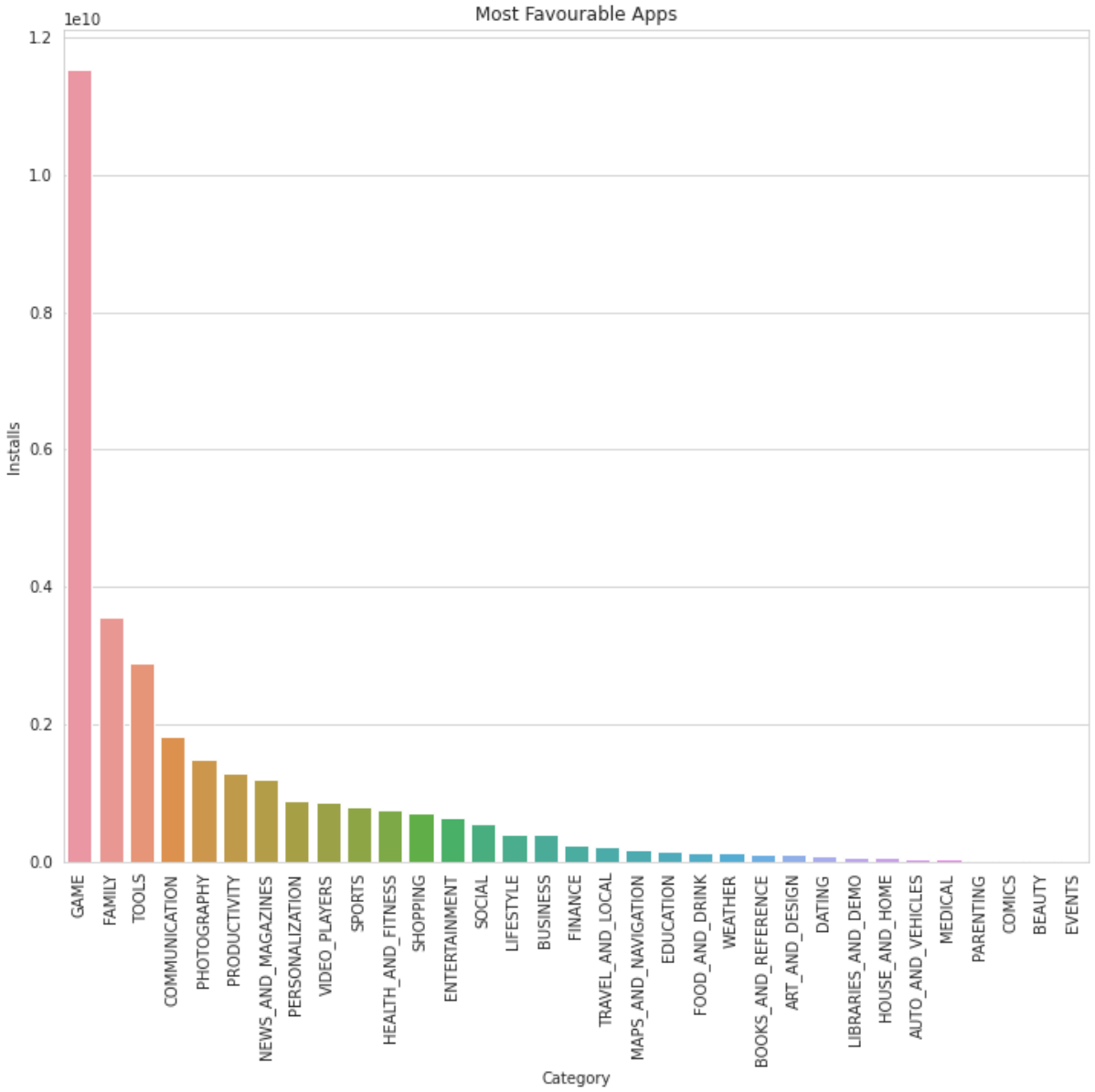
7.1.2 How many apps are there in each category?

Being a developer, we can decide for any type of app to build, but our aim must be to offer usability and engagement to the users.



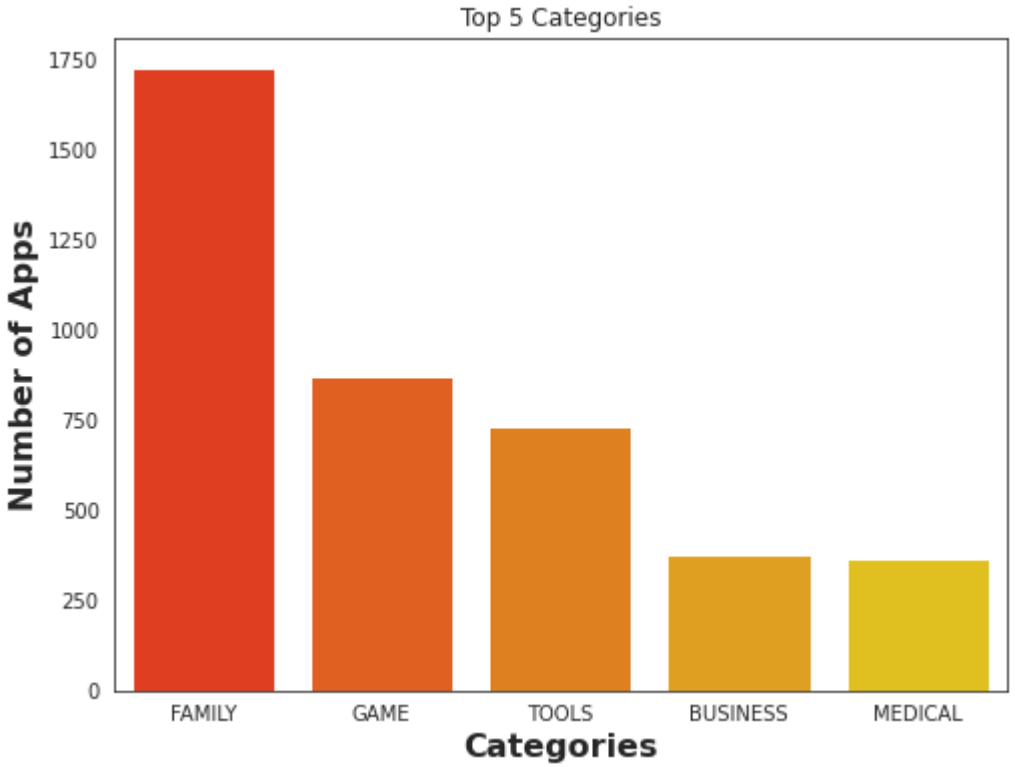
There are 33 different categories present in the play store. We can see that most number of apps in the play store belong to Family and Game categories. This shows that apps in Family and Game categories are more common and have high chances of being successful.

7.1.3 Which category apps are installed most?



Before we saw that the family category has the most number of apps but here it is not the case. When it comes to installs, Game is the most popular category. Entertainment, Communication, Video players, Photography are the other popular categories among smartphone users. Medical category has the least number of installs and it is interesting that it has more apps than many other categories.

7.1.4 Top 5 Categories in Playstore :



In the above plot we have calculated the Top 5 Categories which has highest number of apps preferred by users. From the above plot we can see that the Family category has the most number of apps. Also, the users would like to prefer apps belonging to the family category followed by Game, Tools, Business and Medical. So end user can predict to make such apps which intend towards above these categories.

7.1.4 Which age groups do different categories target?

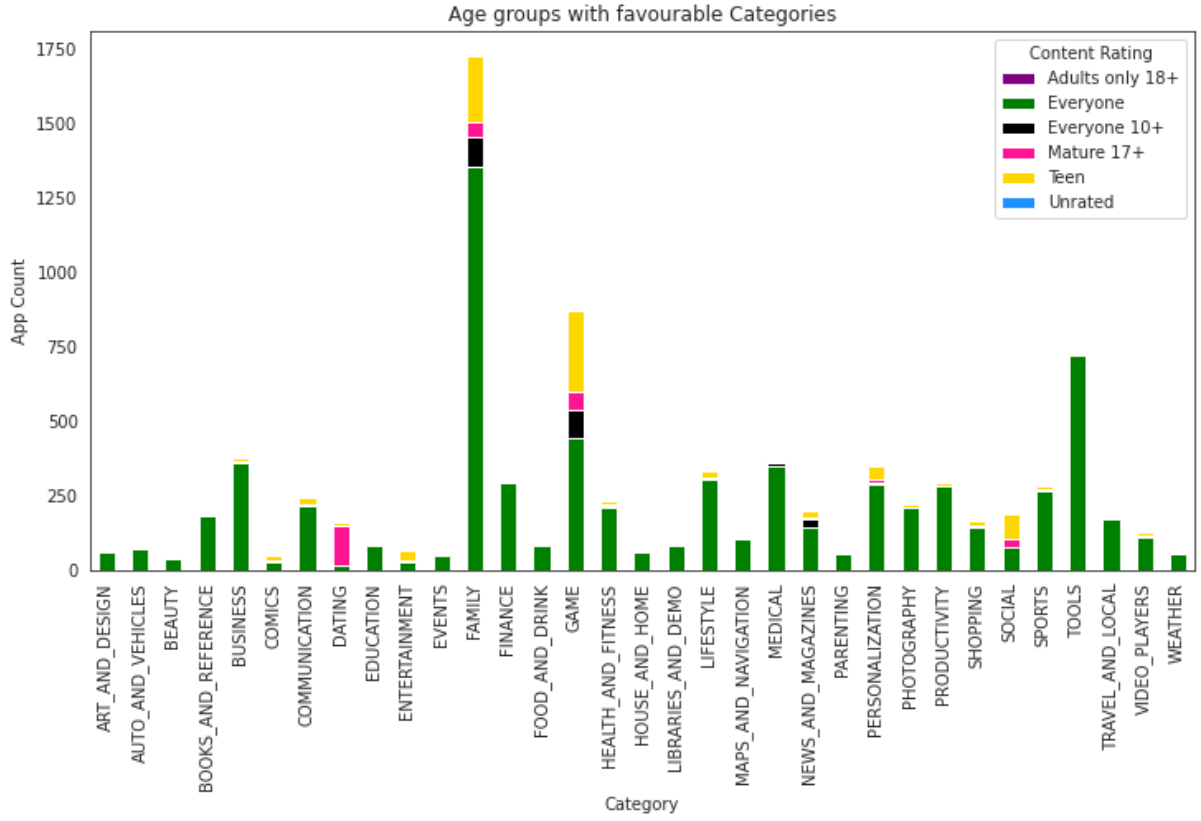
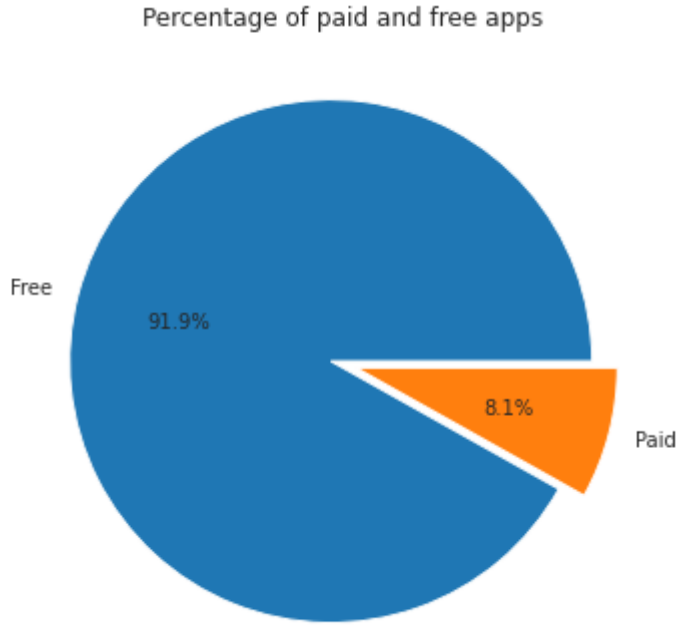


Chart shows every category targets almost all age group audiences. Family, Game and Social categories have more apps for teens as compared to other categories. Dating category has over 90% of apps for the mature audience. Categories such as Tools, Finance, Books and Reference, Education, Weather mostly targets everyone.

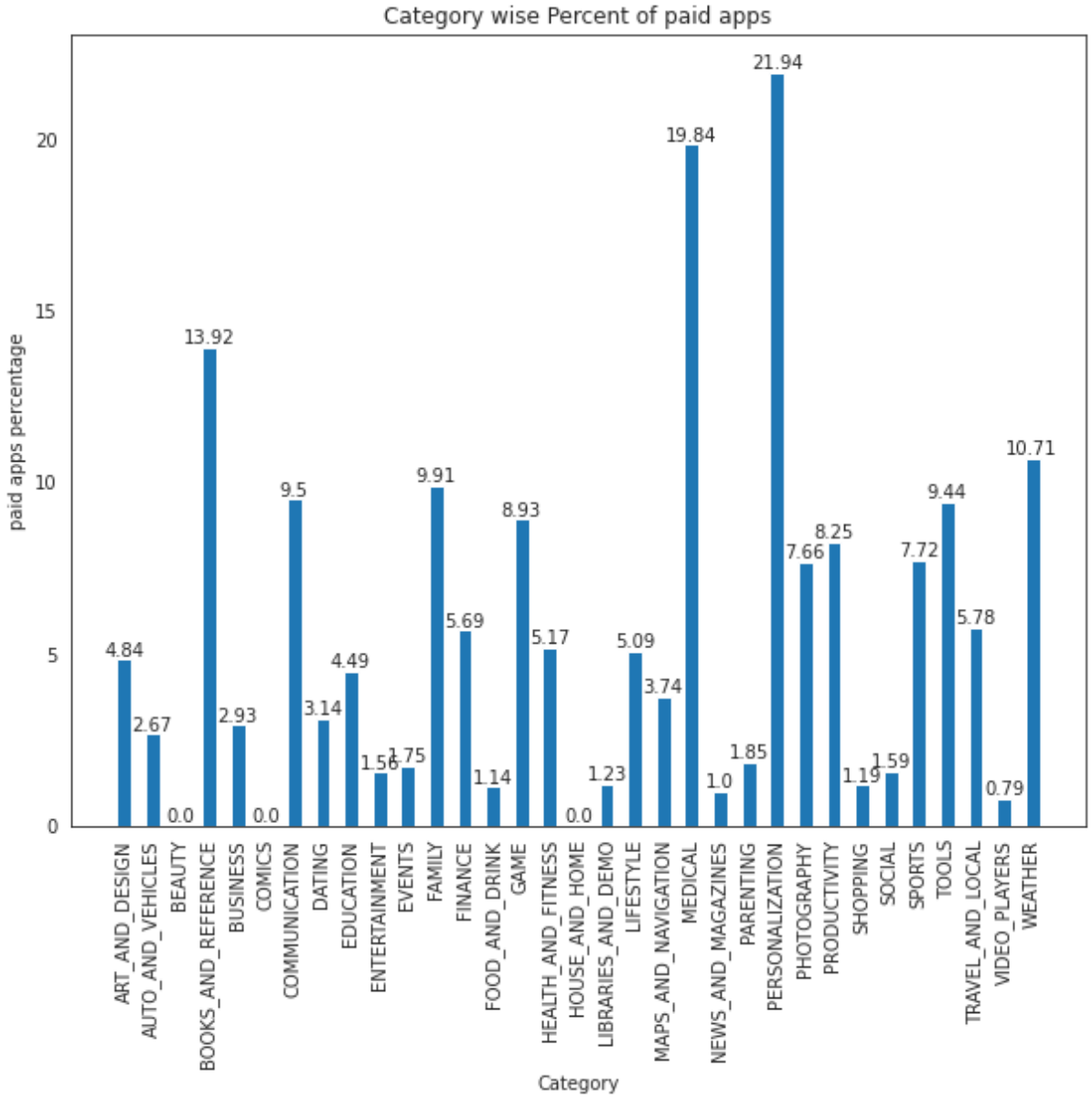
7.1.5 What percentage of apps are paid?



Paid apps are those for which customers pay upfront to download the app whereas free apps are free to download and make money through advertising, in-app purchases or paid subscriptions. Here we saw the percentage of free and paid apps in the play store.

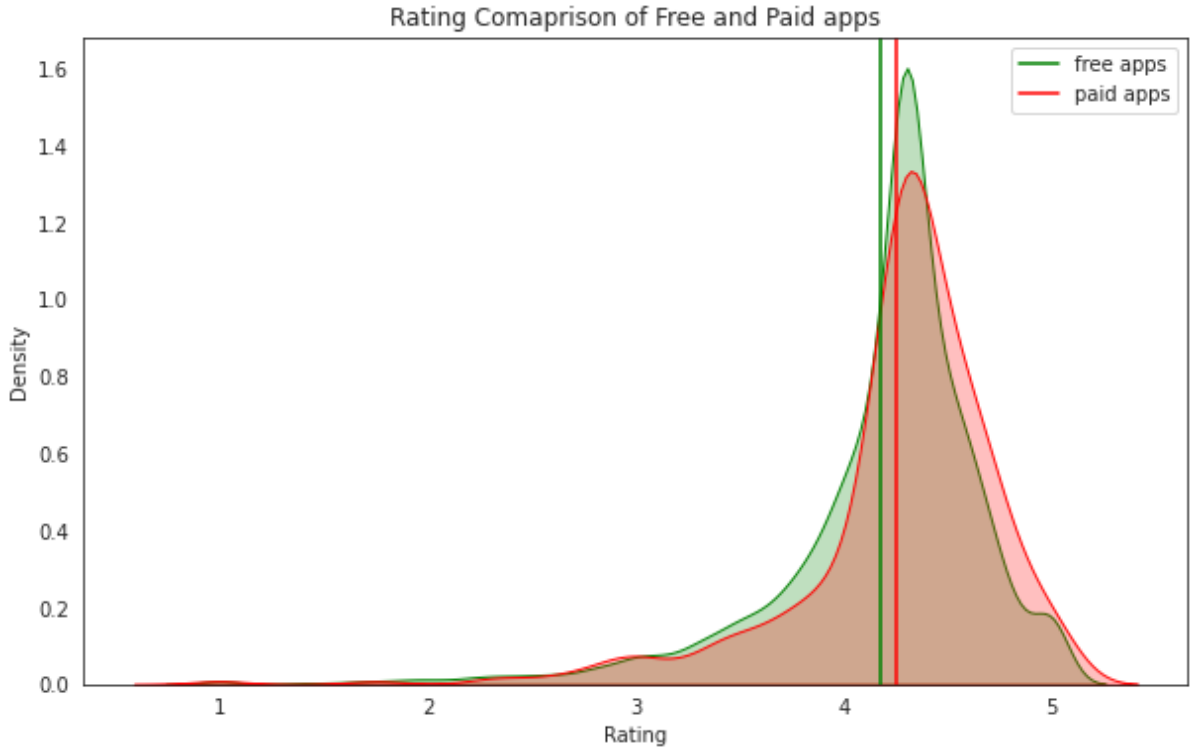
From the above chart we can see that 91.9% of apps are free and 8.1% of apps in the play
store are paid apps. Around 685 apps are paid as compared to 7747 free apps in the play store
dataset which is very less.

7.1.6 Which type of apps are users willing to pay for?



Personalization and Medical categories have a high rate of paid apps as compared to other categories having approximately 22% and 20% of their total apps are paid. We can say that this type of apps generally do well as paid apps, since value is in the apps functionality. Beauty, Comics and House and home categories do not have any paid app.

7.1.7 Do paid apps get better ratings?

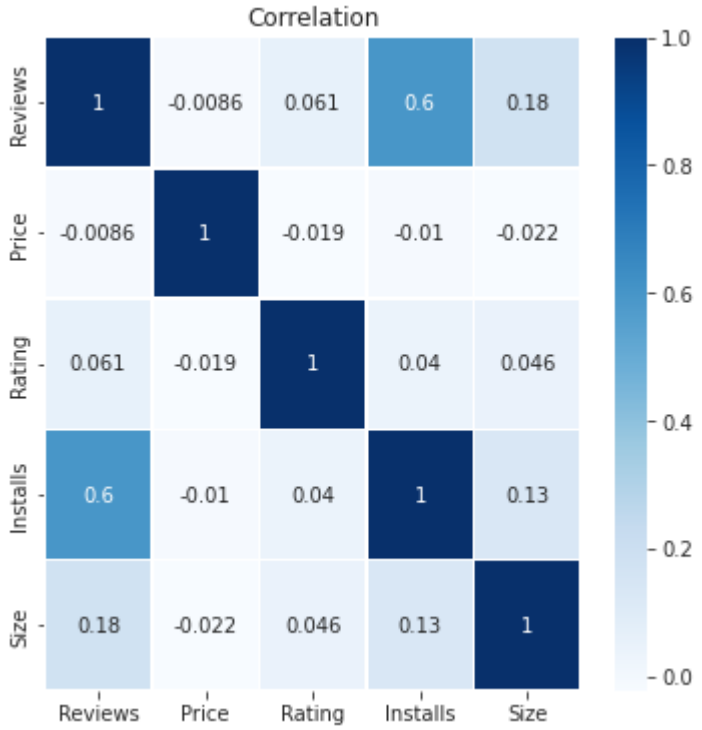


The average rating paid app gets is 4.26 whereas free app gets 4.18 average rating out of 5. We can say that paid apps get better ratings as compared to free apps. The reason for this maybe that paid apps have more loyal and dedicated users. We may not consider this as strong evidence because there are very less paid apps as compared to free apps

.

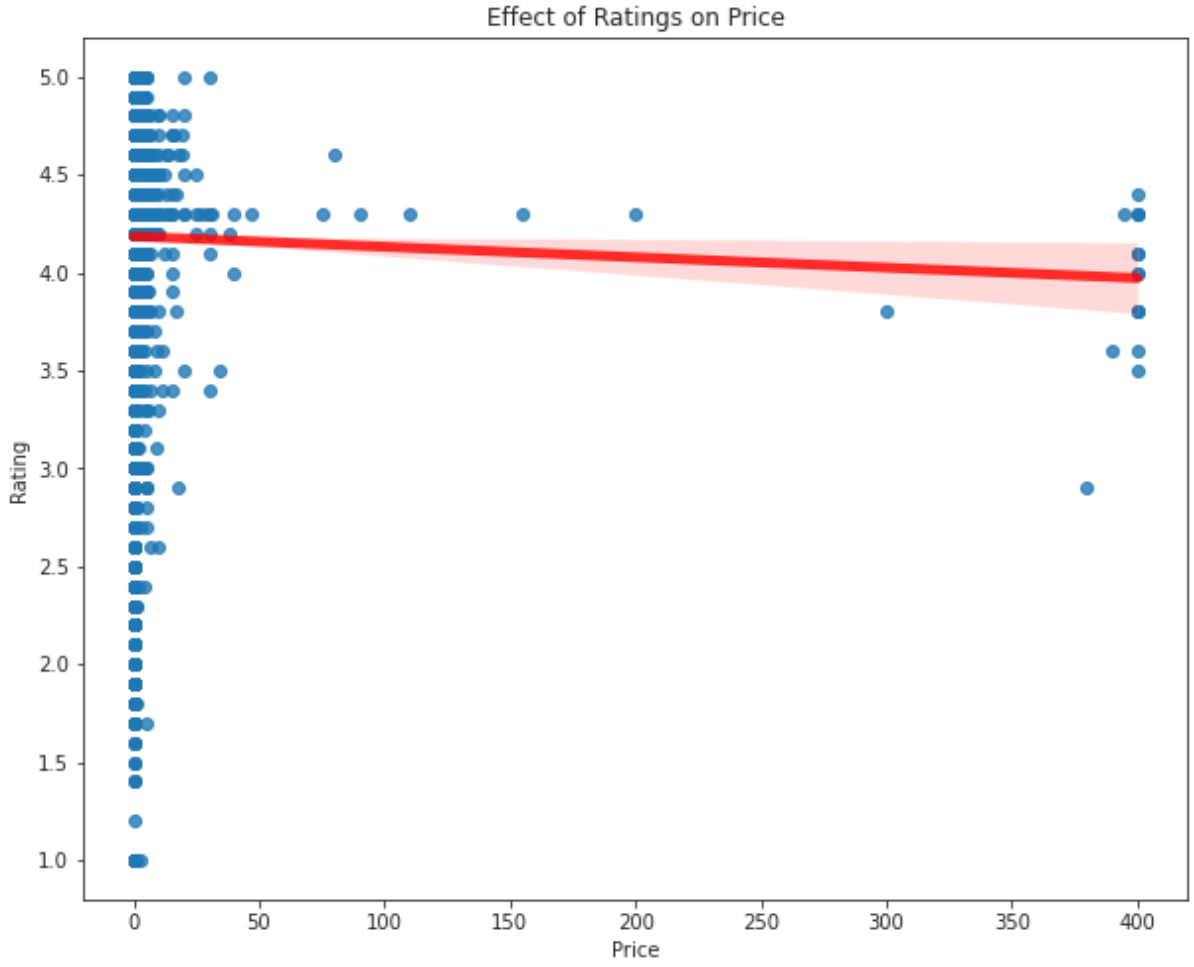
7.2 Visualizing Relations between features :

7.2.1 Correlation matrix :



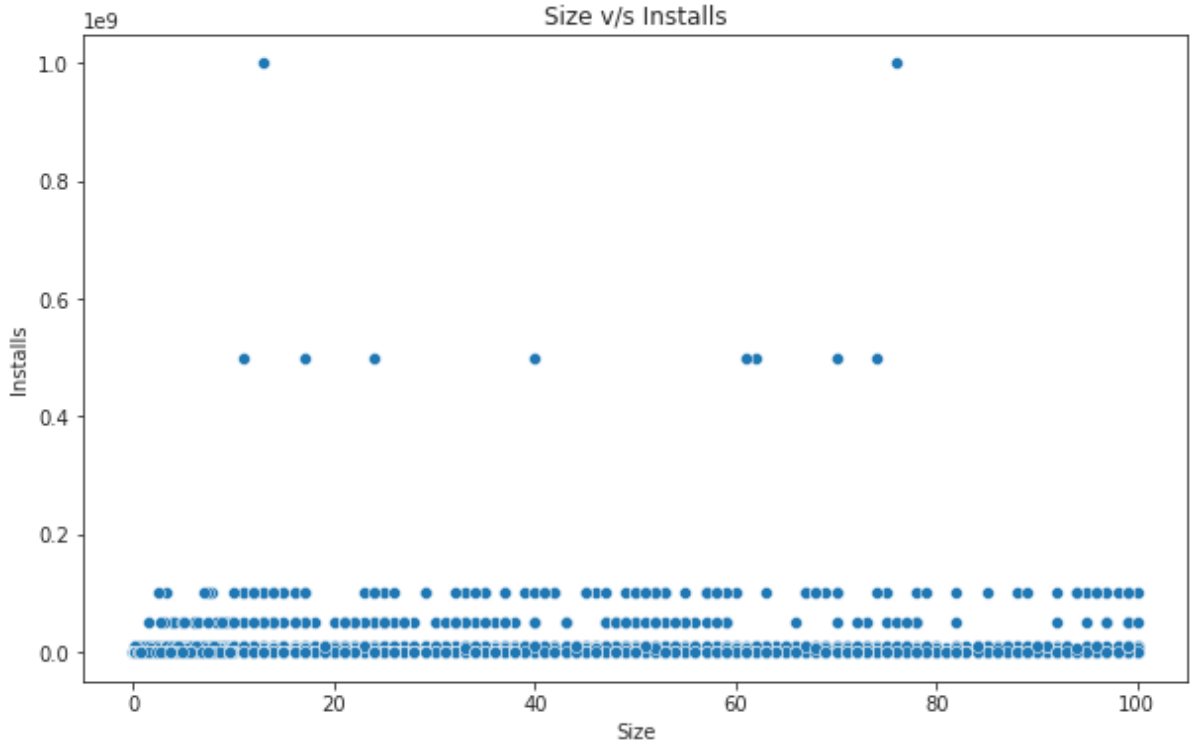
From the above figure we can see that Installs and Reviews have the strongest correlation. This is reasonable because popular apps tend to get more reviews. There is very small correlation found between Installs and other features like Size, Rating, Installs and Price. Price has a negative correlation with all other features but it is very weak and it can be negligible.

7.2.2 Does rating change with increasing price?

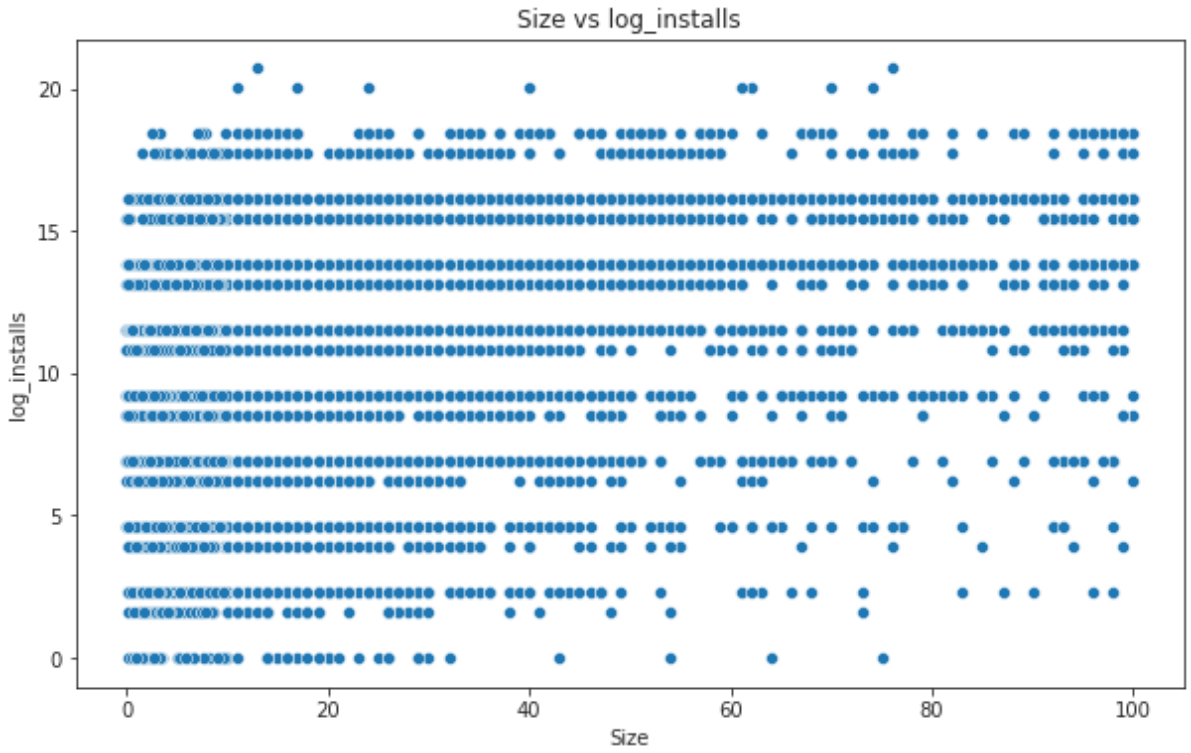


The plot shows that the majority of apps cost below $100. There are less costly apps and the high cost of the app is $400. The red linear regression line in the below plot depicts there is a negative relation between app price and rating. We can say that as the price of app increases there is slight drop in rating of the app.

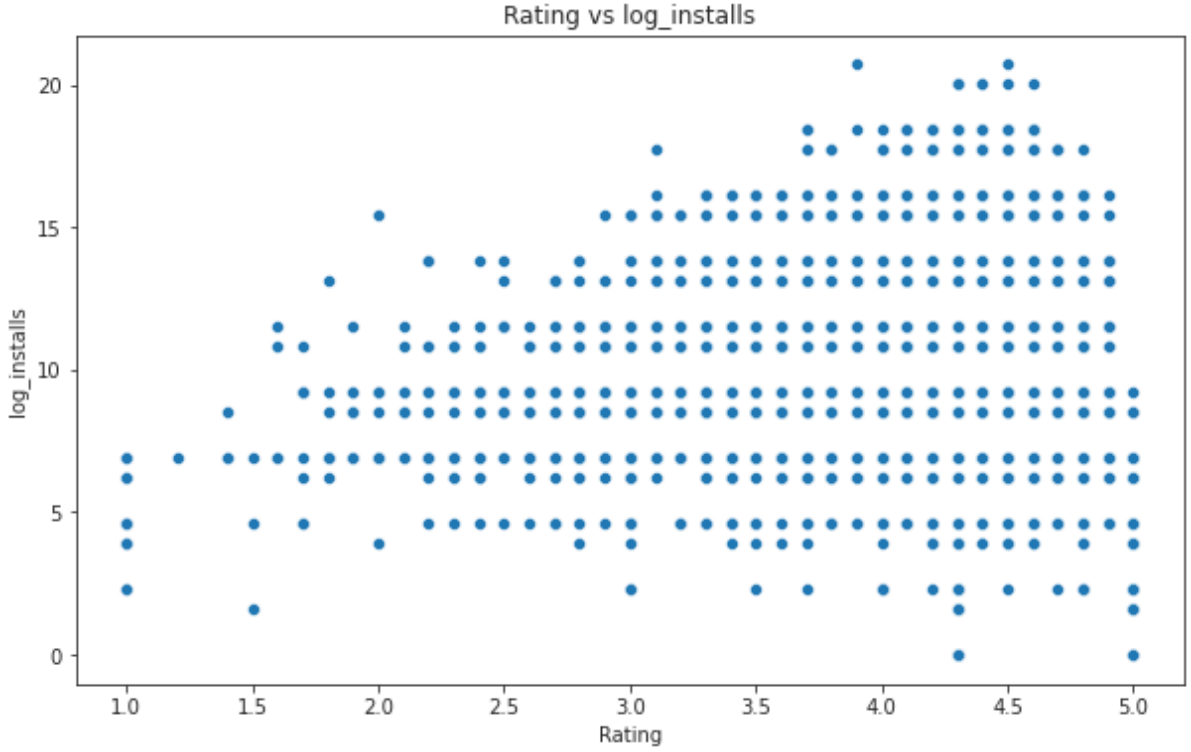
7.2.3 Does the size of an app influence the number of downloads?



Above chart is plotted between Size and original Installs column and this shows why we have done log transform of Installs column as the data is not effectively plotted because installs are highly skewed.
Below graph is plotted after log transformation of Installs and see the difference. The relation between size and installs is not so strong. There are more apps having size less than 20. Users prefer apps that require less space and load faster. We can say that heavy apps are installed less as compared to light apps



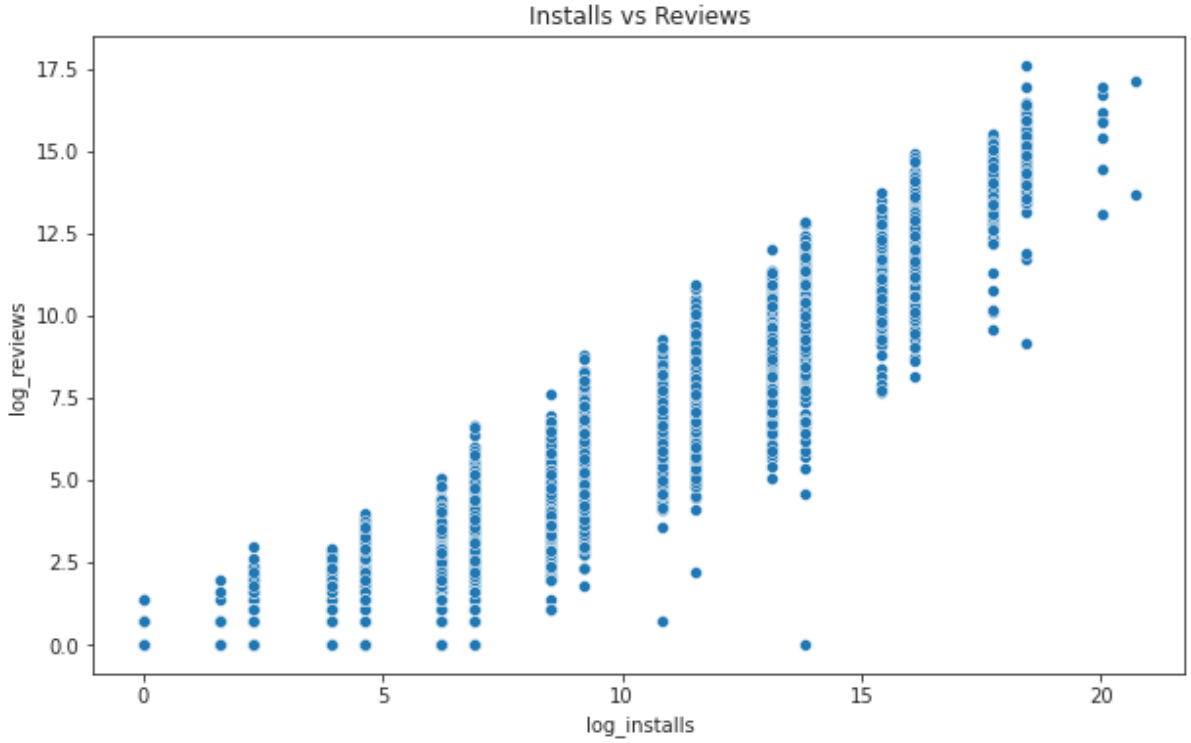
7.2.4 Do higher rated apps attract more users?



We can see that apps rated around 4.5 are installed more. This means higher rating does contribute to more installations. People have a natural tendency to trust the opinion of those around them. This is the reason why users prefer highly rated apps to download.

7.2.5 How do reviews affect the users decision to download an app?

Below graph shows incremental positive relation between Reviews and Installs. We can say
that popular apps which have more installs receive more reviews and this attracts more new
users. When a user wants to know if an app does what it's supposed to do and works well, he
will check the reviews. Other users positive reviews about the app will strengthen his decision to download.



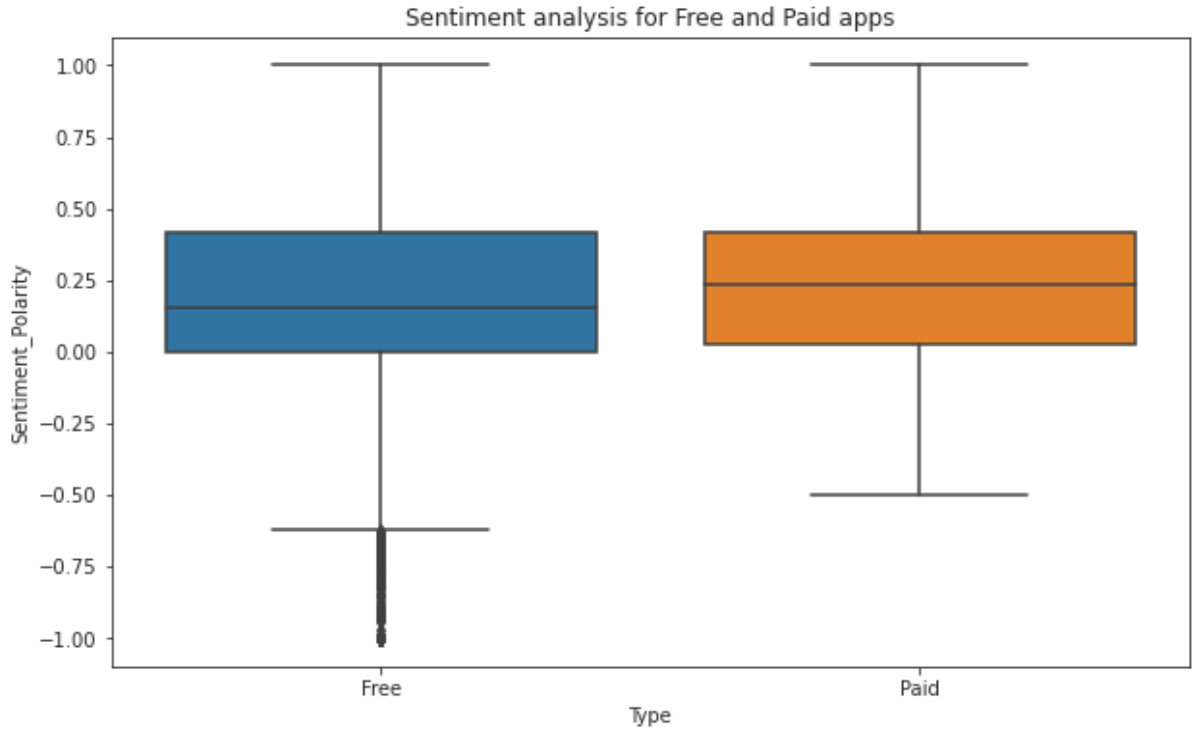
7.2.6 Are app updates important?

We can see from the above plot that most of the apps get frequent updates and they are also installed more. There are very few apps which got updates in 2010, 2011 and 2012. It might be good to update the application as it can be a more engaging way of interaction between the user and the app. We can say that those developers who make their app better over a period of time have a great chance of success.

7.3 Enriching data :

For further analysis we needed more data which was stored in the secondary dataset which we mentioned at the start. This dataset consisted of app reviews and sentiment analysis. We performed the same steps of knowing data, cleaning the data and merged this dataset with the primary one. This dataset consisted of 5 columns namely App, Translated Review, Sentiment, Sentiment Polarity and Sentiment Subjectivity. The dataset contained a lot of missing values. Since we could not predict the missing values, it was better to drop the missing values. The column Translated Review won't be used in analysis, and Sentiment can be identified from the polarity so we also dropped both the columns. Later we merged this dataset with the apps dataset.

7.3.1 Sentiment analysis for free and paid apps :



Free apps get more negative reviews as indicated by the outliers on the negative side in the figure below. Paid apps do not receive extreme negative reviews and median polarity is also higher for paid apps because users are generally more loyal to apps they pay for. We can say that paid apps have better quality because customers who have paid for an app are more likely to demand a premium and frictionless experience.

8. Analysis Result :

From this analysis, we found that there was correlation between app features like rating and reviews, Installs and reviews and installs and rating etc. There was a strong negative correlation between the Price and the number of reviews and between number of installs and price. Most numbers of apps belonged under genres of tools, Entertainment, Education, Business, and Medical. On the basis of the number of installs, we can say that apps from category Game, Communication, Entertainment, video players etc. are most successful amongst all.
On the basis of the number of installs, we divided the apps into two categories: successful and unsuccessful. Decision Tree gave the highest accuracy percentage of 95.32% and the Gaussian Naive Bayes model
gave the lowest accuracy of 88.45%

9. Conclusion :

This data set contains a large amount of data that can be used for various purposes. The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularise the product.

After analysing the dataset we have got answers to some of the serious & interesting questions which any of the android users would love to know. We have also learned that the following things might affect the rating.

- The **Family category** has more apps on the play store but the **Game category** is the most popular category.
- Approx. 91% apps on play store are free apps and Medical and Personalisation apps generally do well as paid apps.
- Users prefer apps that require less space. Bulky apps are downloaded less.
- App ratings and reviews have a significant impact on a user's decision to download or not download an app.
- Updating the app can improve user experience and happy users attract more new users.
- Sentiments in reviews also matters in attracting new users as other user's positive reviews about the app strengthen the decision to download.

Challenges faced during the Project :

- Main challenge in the EDA project is data cleaning/wrangling, but it is the most crucial step while working on EDA projects.
- Generally, data cleaning take 50%-60% time to clean it, and then our data becomes readable and understandable, and then we are free to draw the insights correctly.
- Dealing with the missing and NaN values is another challenge, depends on the data we have, we replace missing or NaN values or else we drop them.
- For replacing we perform some statistical methods like mean, median and median to get the desired result.
- Next challenge was, We found that the columns Installs and Reviews are highly skewed, so by doing thorough research we come up with the method of log transformation, Basically log transformation is used to transform skewed data to approximately conform to normality.