**Capstone Project-1 Technical Documentation**

**Play Store App Review Analysis**

**Data science trainees,**

**AlmaBetter, Bangalore**

--------------------------------------------------------------------------\*\*\*--------------------------------------------------------------------

**Navneet Sawhney -** [navneetksawhney@gmail.com](mailto:navneetksawhney@gmail.com)

**Tinu J Pooppally -** [tinupooppally@gmail.com](mailto:tinupooppally@gmail.com)

**Shubham V Jadhav -** [jadhavshubham7743@gmail.com](mailto:jadhavshubham7743@gmail.com)

**Manoj Chetry -** [chetrymanoj97@gmail.com](mailto:chetrymanoj97@gmail.com)

**Biswanath Das -** [Biswanath2594@gmail.com](mailto:Biswanath2594@gmail.com)

**GitHub Link~~**

**ABSTRACT**

For any business to succeed these days, going digital has become paramount and getting an application for your business can lead it towards more profits. None of us can now imagine our lives without using apps in our mobile phones. Think of any problem you are facing in your life, you will find a solution of same in form of an app specially curated to fix the issue for you. As per latest Google Play stats, there are 3.48 million apps currently at the Google Play Store. The number is constantly rising as around 3,739 apps are added to the Play Store every single day.

Due to constant rising numbers of apps, it becomes imperative to understand the parameters and patterns related to app creation from a business point of view. The Play store datasets are quite intriguing as they consist of details like number of installations, app reviews, sentiment ploarity and so on. In this project we digged deeper into the datasets by doing an Exploratory Data Analysis to see certain patterns that can be helpful to different category of businesses.

We began by in-depth cleaning of the datasets and then we merged them to create one dataset. After that we did a generalised analysis to get numerous insights. We particularly focussed on the customer behaviour and what components affect the decision of the user to install the app. With the information gathered, we further tried to see if for medical category creating an app would be benefical. If so, what kinds of apps have a higher probability of being liked by the audience.

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

# Problem Statement

The data considered for the Analysis is the Google Play Store Dataset which actually consists of two datasets. The datasets consists of different features/columns related to certain apps. An Exploratory Data Analysis will be performed on both datasets. The initial step would be to perform deep cleaning of both the data sets. Cleaning is the most important step of Data Science cycle as it will help us in proper further analysis and visualisation. Also, it can later be used for Data modelling and apply Machine Learning Algorithms.

Our main aim is to structure the data and then see numerous patterns and trends across different features. We will do a general study of dataset through visualisations and draw some preliminary conclusions. After that we will delve deeper into seeing the scope of medical apps and also the consumer behaviour in cased of apps related to the healthcare/medical category. Also we will make final conclusions about the probability of the success of a new app based on the gathered insights.

# 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 India’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, we will examine various qualities like rating, free or paid and so forth utilizing Hive and after that we 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.
* **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:** A value from 0 to 1 indicating the subjectivity of the review. Lower values indicate the review is based on factual information, and higher values indicate the review is based on personal or public opinions or judgements.

# EXPLORATORY DATA ANALYSIS ( EDA)

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

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

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

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

# 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.

* **Step 1:** First of all we handled duplicate value, As we can see there are 483 rows with duplicate values so we need to delete them. After handling the duplicate values we can see here the number of rows have reduced from 10841 to 9660. So we have successfully handled the duplicate values.
* **Step 2**: We write a function playstoreinfo (), that will display 5 attributes about all the columns: Data type, All values, Null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.
* **Step 3**: We drop the columns ‘Current Ver’, ‘Android Ver’ from our dataset using the notna() function of the pandas library.
* **Step 4**: 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 ‘free’ to replace corresponding row index.
* **Step 5**: The Rating column contains 1463 NaN values which accounts to apprximately 13.5% of the rows in the entire dataset. It is not practical to drop these rows because by doing so, we will loose a large amount of data, which may impact the final quality of the analysis. The NaN values in this case can be imputed by the aggregate (mean or median) of the remaining values in the Rating column.
* **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:** We can see that the ‘Reviews’ column despite being a numerical indicator is of the ‘object’ data type, we will convert this to ‘float’ data type using the astype(float) function.
* **Step 8:** 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 replace() functions and using astype(float) function..
* **Step 9:** 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 ‘float’ datatype.
* **Step 10:** We also change ‘Last updated’ column type to datetime using to\_datetime() function.
* **Step 11:** We write a function Userinfo(), that will display 5 attributes about all the columns: Data type, count of non-null values, null values ,number of unique values in that column and percentage of null value in that columns in the User review dataset
* **Step12:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26868 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.

## Number Of Apps Per Category

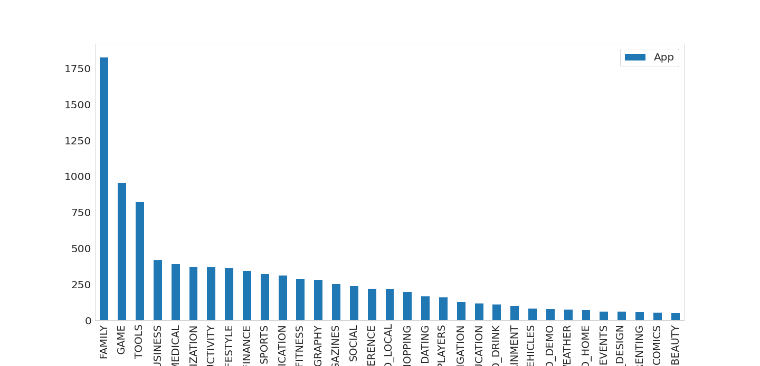
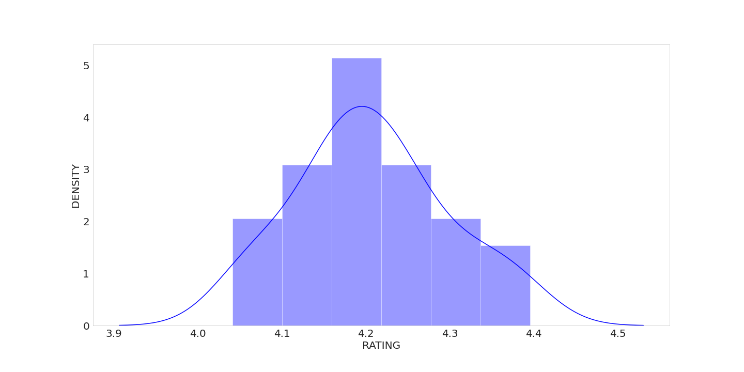


Fig -1: Max number of apps per category

The above bar graph represents the distribution of number of apps in different categories in the Play Store. It can be infered that FAMILY Category has the maximum number of Apps.

## Rating

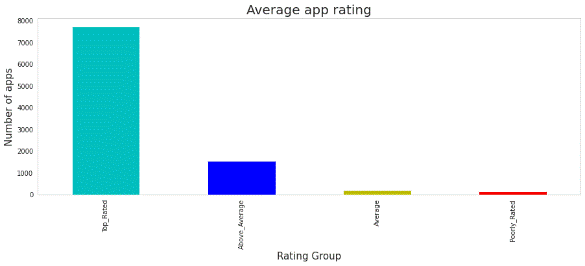
In the below plot, we plotted the apps Rating



**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 distplot 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.

## Average app rating

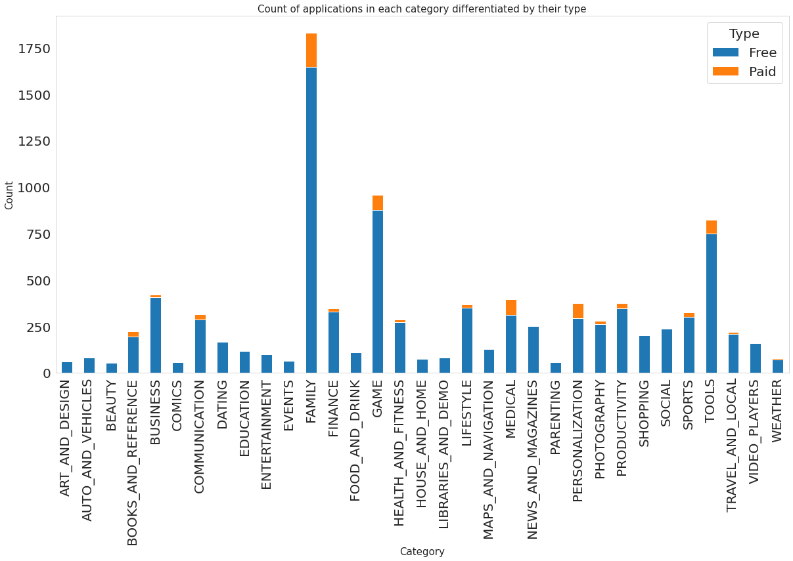


**Fig-5: Average app rating**

Nearly 8000 apps are top rated and about 2000 apps are above average.

From the above analysis we can conclude,

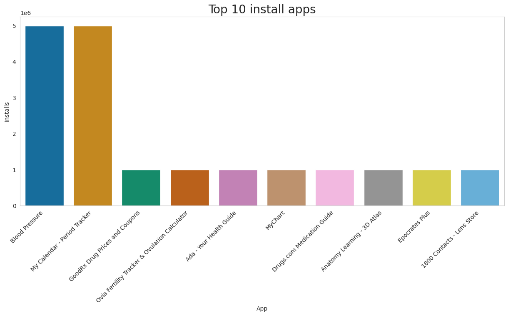
1. For all the categories, the average rating is above 4 stars.
2. Most of the Apps are "Top\_Rated" as we can see in the above visualisation.
   * **Count of app in each category based on their type**



**Fig -6: count of app by type.**

It can be inferred that certain app categories have more free apps available for download than others. In our dataset, the majority of apps in Family, Games and Tools, as well as Social categories are free for users to install.

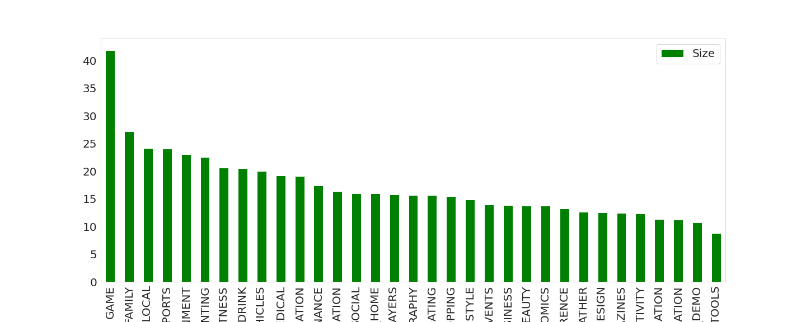
* **Top 10 apps in medical category**



**Fig -7: top 10 medical app**

We can see the top 10 Medical Category apps in the above visualization.

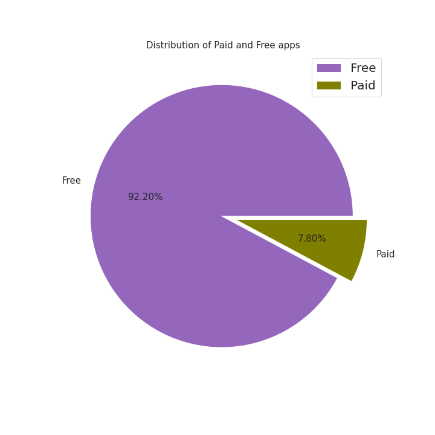
* **Optimal app size**



**Fig -8: Average app size by category**

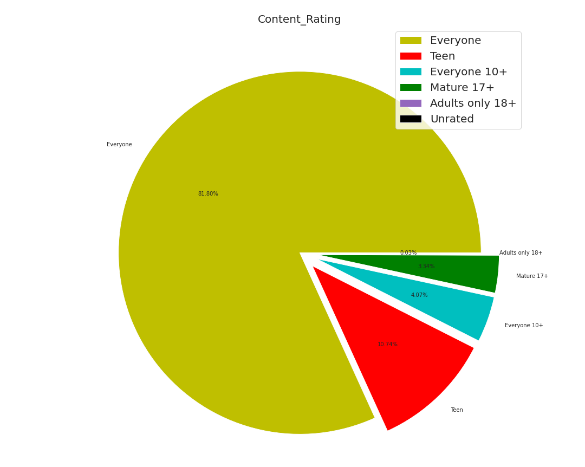
This tells us the category of apps that has the maximum number of installs.

* The average App Size is in the range 10-25 Mb.
* The Game category has the maximum average App Size.
* Tools has the least average App Size.
* The average App Size of Medical category stands at 19 Mb
* **Distribution of Paid and Free apps**



**Fig -9: Distribution of paid and free apps**

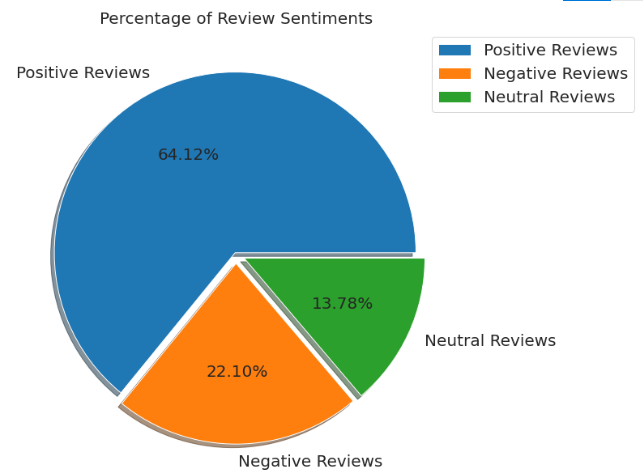
**It can be inferred that around 92% Apps are free and the 8% apps are paid.**



**Fig -10: Distribution of content rating.**

As per the above pie chart, around 82% Apps are created for Everyone and the least being Unrated i.e. 0.02 followed by adults only 18+ i.e. 0.03.

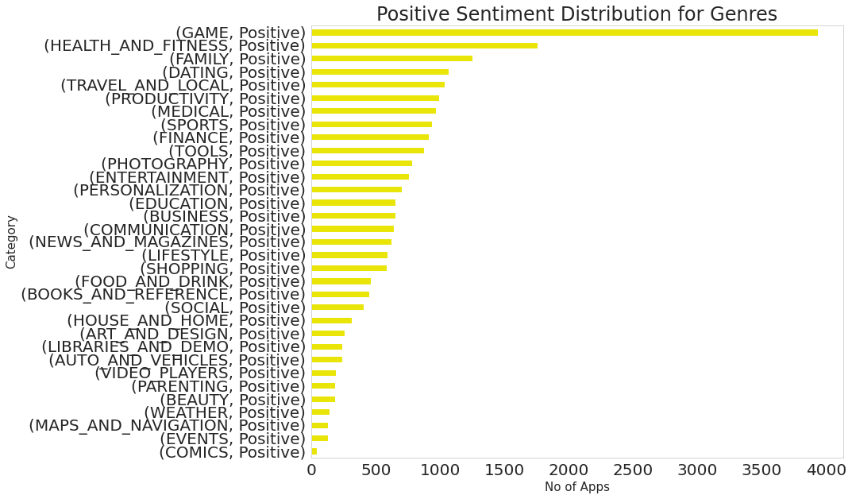
## Percentage of User review Sentiments



**Fig -11: Percentage of User Review Sentiments**

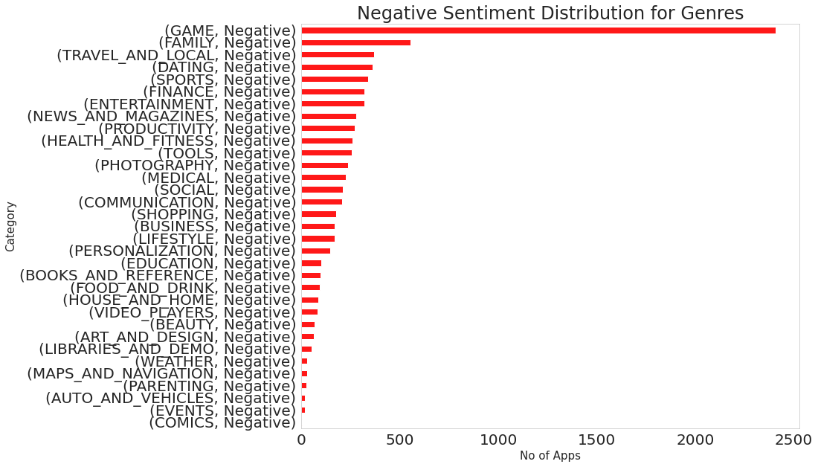
From the above pie chart, we can say that most of the apps that are present on the play store has received positive review by the user while there are some apps which have negative reviews as well.

* **TOP POSITIVE SENITMENT RATED FOR GENRES**



**Fig -12: Top Positive Reviewed Genres**

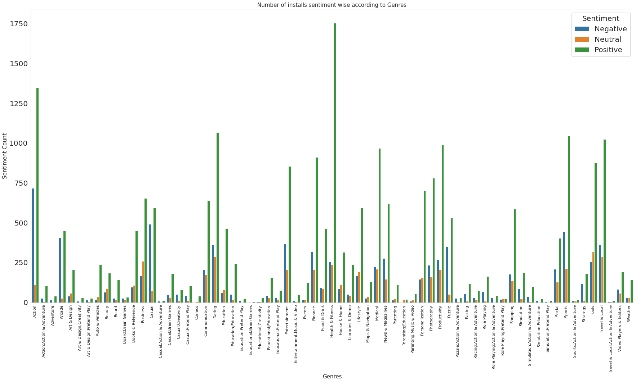
* **TOP NEGATIVE SENITMENT RATED FOR GENRES**



**Fig -13: Top Negative Reviewed Genres**

The graphs showcase that Games has the most number of positive and negative reviews which can be a result of it having the highest number of reviews within the category distribution. But a difference can be seen when moving to the next one, where Health and Fitness holds the second place in positive reviews followed by Family and vice versa in the negative sentiment distribution

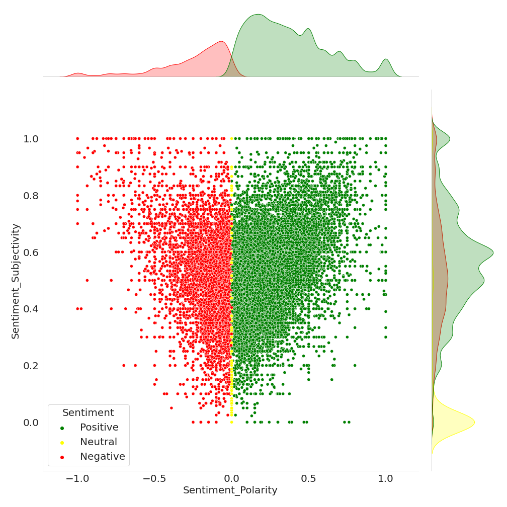
## Distribution of sentiment within the genres



**Fig -14: App install sentiment wise by genres.**

It can be seen that Health and Fitness has the highest number of positive reviews which is followed by Action. But it is also worthwhile to note that Action has the highest number of negative reviews as well while the ratio is much less for Health and Fitness.

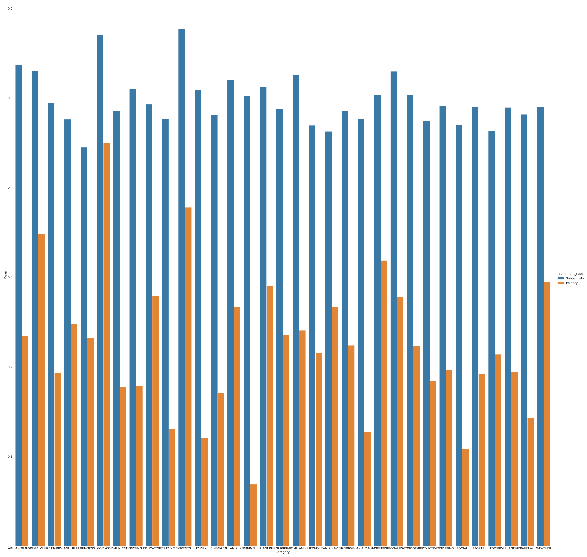
## Relationship between sentiment subjectivity proportional to sentiment polarity



**Fig -17: Proportional check**

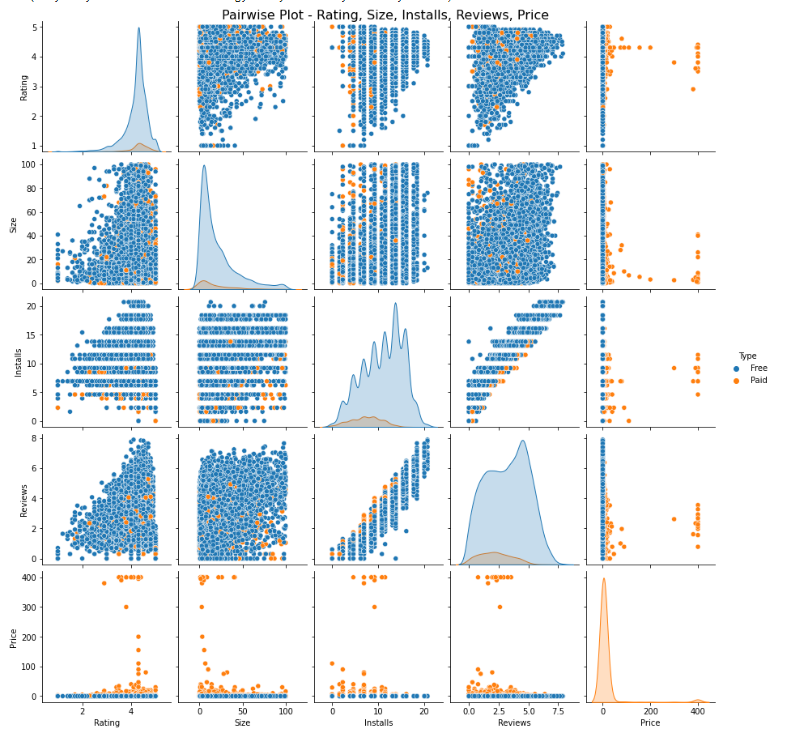
From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, show a proportional behavior, when variance is too high or low.

## Distribution of Subjectivity and polarity within DIFFERENT LOCATIONS categories.



**Fig -18: Distribution of sentiment**

## Relationship between different features of the dataset



**Fig -19: Pair wise plot**

* Most of the App are Free.
* Most of the Paid Apps have Rating around 4
* As the number of installations increases the number of reviews of the particular app also increases.
* Most of the Apps are light-weighted.

## Correlation Heatmap

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1. Correlation between two random variables or bivariate data does not necessarily imply a causal relationship.

## 

**Fig -20: Merged Data frame Heatmap**

* There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
* The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviews and installs fall slightly.
* The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs, and number of reviews also increase.
* Sentiment Polarity is not highly correlated with Sentiment Subjectivity.

**CONCLUSION**

1. The app's name should accurately describe its value propositions. because majority of successful application have this quality.
2. Launching the apps in the category which having more and easy user reach such as **"FAMILY"**, **"GAME"**etc.
3. As per our analysis most of the apps are free it's around **92%**, so if possible try to **launch the app with "Free" type**. as it's **increases the user engagement.**
4. We also seen that the number of installs are correlated with the rating of the application as number of installs increases so as the application rating.
5. One important inside we get with respect to the size of the app is that as the size of the application increases the installation of the app descrease. So, if we release a new apps in market **make sure it's under 20MB.**
6. Content rating also affect the user engagement as more restructed your content rating is the more restricted your user engagement. So try to keep user rating as **"Everyone"**.
7. Make sure the app will get the update at regular interval, as it's an important factor for user engagement and performance of the application. In our analysis we seen that most of the apps will get their app update at **July month**.
8. As we seen form subjectivity most of the reviews as the objective, So for successful apps it's more important to keep eye on the user reviews and early resolution of the problems.
9. We also seen from our analysis that there is strong relationship between Install and reviews. for the new apps reviews are the important tool for increase the user engagement.
10. For the apps it's more important have the android version compactibility **above version 4**.
11. It can be seen that Health and Fitness has the highest number of positive reviews which is followed by Action.

So, from our analysis we suggest that client's must be aware of above points beside creating a medical healthcare app

Thank You