**Capstone Project-1 Submission**

**Play Store App Review Analysis**



**Yotish Lakhanpal**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Email-** [**yotish.lakhnpal@gmail.com**](mailto:yotish.lakhnpal@gmail.com)

**Github link – https://github.com/y0tish/Play-Store-App-Review-Analysis.git**

**Abstract-** On the Google Play store, a few thousand new applications are constantly uploaded. Unrestricted design work is being done by a large number of designers to create successful apps. A developer needs to be aware of whether or not they are moving in the right direction given the immense difficulty they face on a global scale. Since the majority of Play Store apps are free, it is quite difficult to understand how in-app purchases, in-app advertisements, and memberships contribute to an app's success. In this method, rather than the amount of revenue generated, an application's success is typically determined by the number of installations it has received as well as the customer reviews it has received over time. The goal of this experiment is to provide data that will assist creators to better understand what user wants and, in turn, aid in the product's widespread adoption. We have made an effort to understand the connections between many characteristics, such as whether an application is free or paid, how it is rated by users, how big it is, and how much it costs.

1. **Problem Statement:** One dataset contains general information, whereas the other contains user reviews for the relevant software. To determine the crucial factors that affect app engagement and success, we must analyze and assess the data in both datasets. These are the some attributes of given dataset

**The contents of Play Store Data are:**

* App: It contains the name of the app with a short description (optional).
* Category: This section gives the category to which an app belongs. In this dataset, the apps are divided among 33 categories.
* Size: The disk space required to install the respective app.
* Rating: The average rating given by the users for the respective app. It can be in between 1 and 5.
* Reviews: The number of users that have dropped a review for the respective app.
* Installs: The approximate number of times the respective app was installed.
* Type: It states whether an app is free to use or paid.
* Price: It gives the price payable to install the app. For free type apps, the price is zero.
* Content rating: It states which age group is suitable to consume the content of the respective app.
* Genres: It gives the genre(s) to which the respective app belongs.
* Last updated: It gives the day in which the latest update for the respective app was released.
* Current Version: It gives the current version of the respective app.
* Android Version: It gives the android version of the respective app.

From above dataset we are going to extract some of these insights:

1. What is percentage of Free and Paid applications?
2. How many apps are belonging to what content Rating?
3. How many apps are belonging to which Category?
4. What are the top Categories in play store in most Number of Installs?
5. What are the Top apps in that Category which has most number of installs?
6. Which Category has most number of reviews?
7. Which Category has most number of bigger apps?
8. How the different sizes of apps make differences in different attributes of apps?
9. Which Category is performing well in market is decided by rating analysis.

**The contents of User Reviews are:**

* App: It contains the name of the app with a short description (optional).
* Translated\_Review: It contains the English translation of the review dropped by the user of the app.
* Sentiment: It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
* Sentiment\_Polarity: It gives the polarity of the review. Its range is [-1, 1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
* Sentiment\_Subjectivity: This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0, 1]. Higher the subjectivity, closer is the reviewer’s opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

The above date set is going to helps in analyzing about which apps have most number of positive or negative reviews and also tells the sentiment polarity towards the free and paid apps.

**2. Introduction:** In the current environment, it is clear that mobile apps play a significant part in every person's life. A designer needs to be aware of whether they are moving in the right direction or not in the face of immense challenge from all around the world. The application designers may need to find out a way to maintain their current position in order to maintain this money and their position in the market. The dataset containing 10k Play Store apps is accessible to study the android market. It can be looked at to analyze several categories, including those related to family, communication, entertainment, tools, gaming, health and fitness, education, etc. In this project, we look at the many data set characteristics that influence how well-liked an application is. We concentrated on providing answers to queries like, "What makes an app popular?" "What should the app's price and size be?" and "Are there any user sentiment trends?" We have two csv files in our data set for data analysis: Play Store user reviews data. First, we looked at the data from the Play Store, which has 10841 rows and 13 columns, as well as 64295 rows and 5 columns of data from user reviews. . We need to explore and analyze the data to discover key factors responsible for app engagement and success.

**3. Exploratory Data Analysis: Exploratory data analysis (EDA) is used to examine data sets for patterns and abnormalities (outliers), formulate hypotheses based on our understanding of the information, and describe their key characteristics. Data visualization techniques are frequently employed in exploratory data analysis. In any project involving data analysis or data science, it is a crucial stage. It assists in deciding how to best modify data sources to obtain the desired results.**

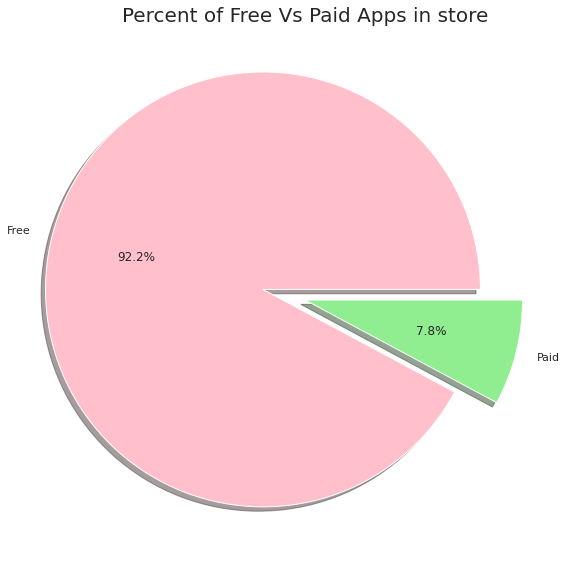
The following are the various steps involved in the EDA process:

1. **Problem Statement** -We'll discuss ideas and analyze the provided data set. We'll examine the traits it has and attempt to conduct a theoretical examination of their significance for this issue.
2. **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.
3. **Univariate Analysis** -It is the most straight forward method of data analysis. In order to start, we would choose just one quality and thoroughly examine it. It deals with no co-relation whatsoever, and its primary goal is to describe. Data is taken, reviewed, and patterns are discovered within the data.
4. **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.
5. **Multivariate Analysis** -This is done when more than two variables have to be analyzed simultaneously.
6. **Data Cleaning** -We shall clean the dataset and handle the missing data, outliers and categorical variables
7. **Testing Hypothesis** -We shall check if our data meets the assumptions required by most of the multivariate techniques.

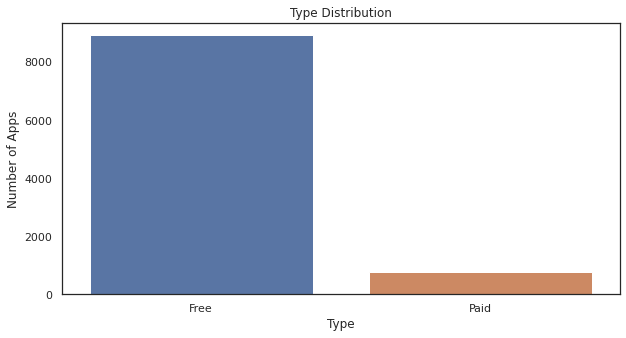
## 4. Steps Involved:

The first and foremost step is to load the given dataset and for that we have to import the important libraries like Pandas, Numpy, Matplotlib and Seaborn. After importing the dataset, we may begin the exploration process, but first we must determine whether the dataset is ready to support many exploration operations or not. To that end, let's first have a look at the structure and way the data is arranged. Firstly we check about the information of the datasets and find out that how many missing or null values are there. A general exploration of datasets is been conducted by checking all the columns and a rough idea is generated about what this dataset is all about and what we can do from it.

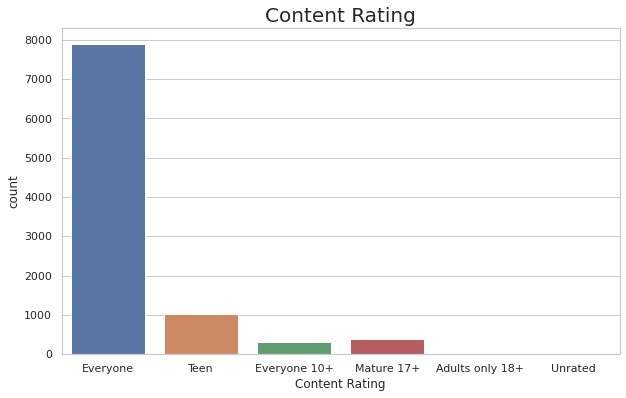
* **Data Cleaning and Transforming:** There are 13 columns in one dataset with 10841 rows. The column Review had type object and some values were in million but the million was written as ‘M’, Hence this ‘M’ was removed and million was multiplied where it was written and column data type was changed to float. In column of size the values was in string and different sizes were give in Mb and Kb, all the sizes were changed to Mb and the column data type was changed to float type. The install column was also in object type it is also changed to float type and signs like ‘+’ and ‘,’ is also removed. The values of price column had a ‘$’ sign in it, it was removed and the type is again changed to float type.
* **Null Values Treatment:** Our dataset contains some null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.
* **Exploratory Data Analysis: Mostly are analysis is Categorical wise, mainly we saw which category is performing well in market but before that we saw the percentage of free apps versus paid apps**



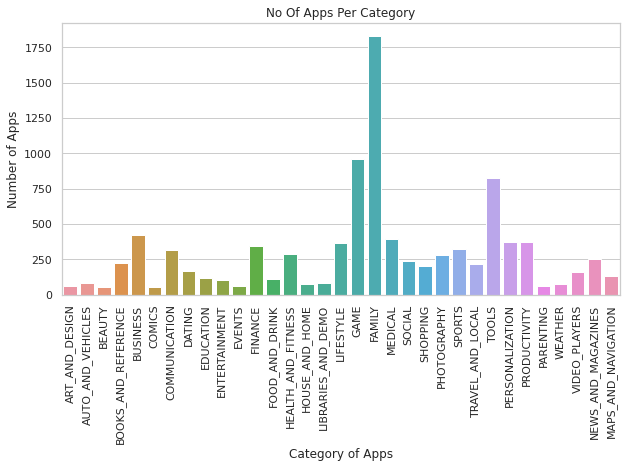
Most of the apps are free in the given dataset provided, Less than 10% apps are paid in play store dataset.



1. **Content Rating:**

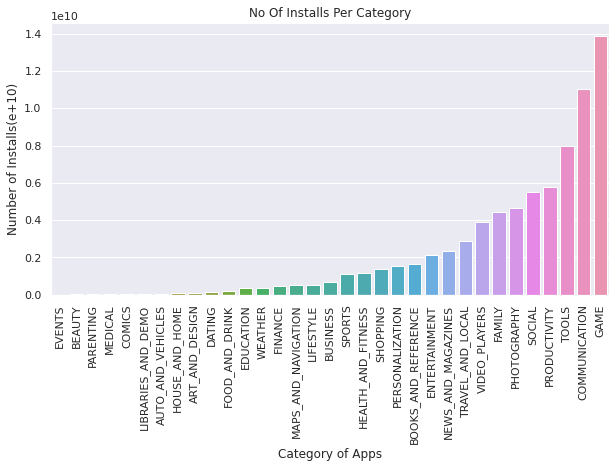
Most of the apps are for everyone to use and there was not a single app which is only for adults. 

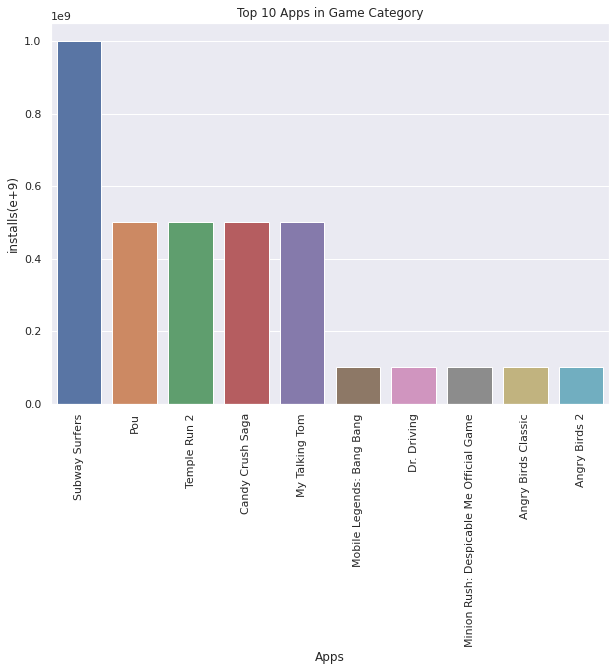
1. **Number of Apps in each Category:**

Most No of Apps in the play store is from Categories **FAMILY, GAMES AND TOOLS.** ****

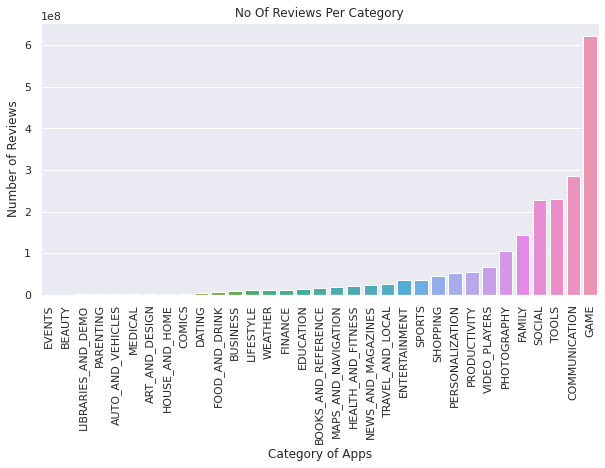
1. **Number of Installs in each Category:**

The most no of Installs is for **GAME** whereas for category **FAMILY** has less no of Installs. The **COMMUNICATION** Category has quite less no of apps but this category has **2nd** in most Number of Installs.

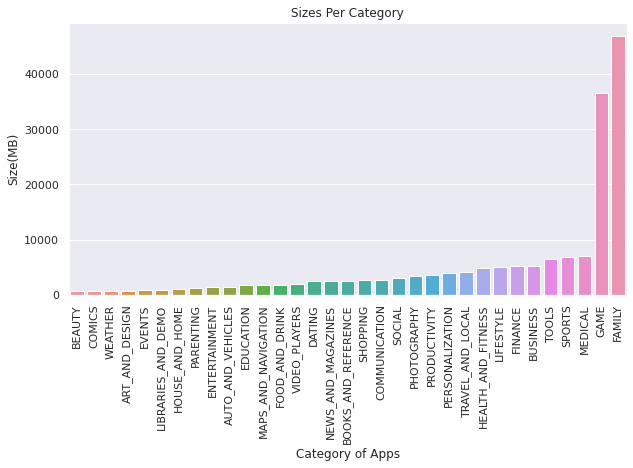
As the Category Game as most of number of install we will see Top 10 Games in installs and will see their attributes.

Most of the Game apps are light weighted, offline games and single player games. Also the Rating of these apps is quite remarkable.

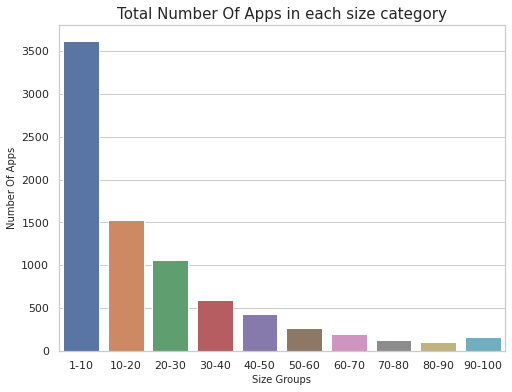
1. **Number of reviews per category**:

Most Number of reviews is also for **GAME** Category followed by **COMMUNICATION** and **TOOLS.** Results are matching with most number of installs, there can be some relation between them. For now we can say that the category which has highest no of installs gets highest no of review. 

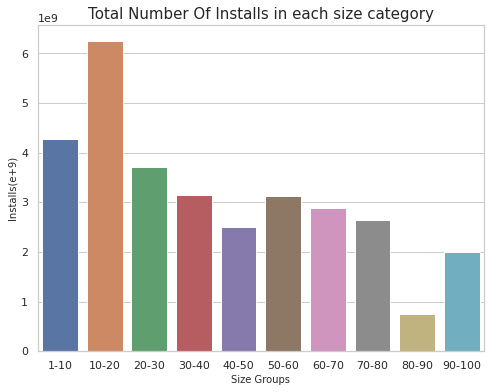
1. **Size Analysis:**

Firstly the size analysis is done on categorical basis. Most number of heavy apps in sizes is in categories **FAMILY** and **GAME**. 

We have grouped the sizes in 10 sub-groups varies from 1-10, 10-20, 20-30 and so on (all sizes are in Mb) and then we saw how various group sizes make an impact on the various attributes of apps.

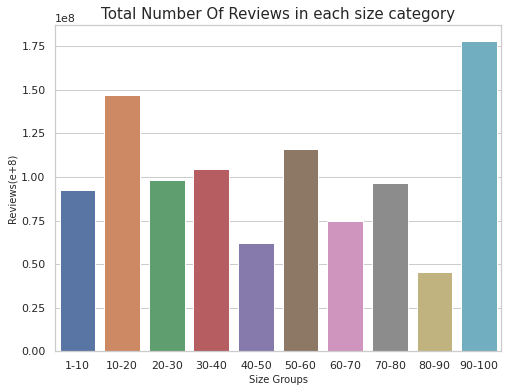
Firstly we saw that which sub group has most number of apps. 

Most number of apps are in sub group of 1-10(Mb).

Secondly we saw that which sub group has most number of Installs. 

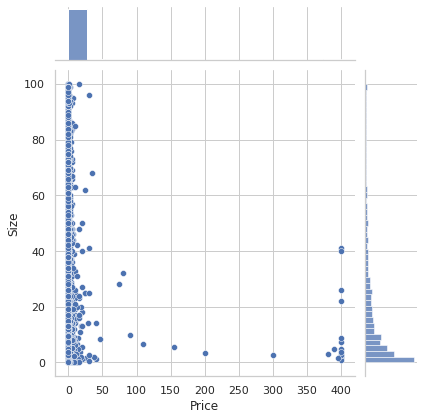
User **prefers** to install those apps which has a size between **(~10MB to ~ 20MB).**

Lastly we saw that which sub group has most number of reviews

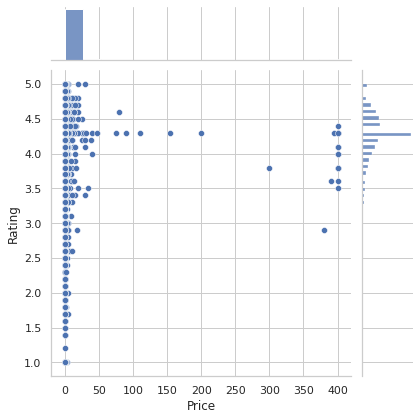


As **(~10MB to ~ 20MB)** sizes apps have more installs therefore they also have more number of reviews. But **(~90MB to ~ 100MB)** have most number of reviews although these categories have least number of apps and also the installs are not that remarkable.

Now will see about the sizes of paid apps and free apps and will conclude how a size of a paid app can be a vital attribute when it comes to the performance of app in the market.

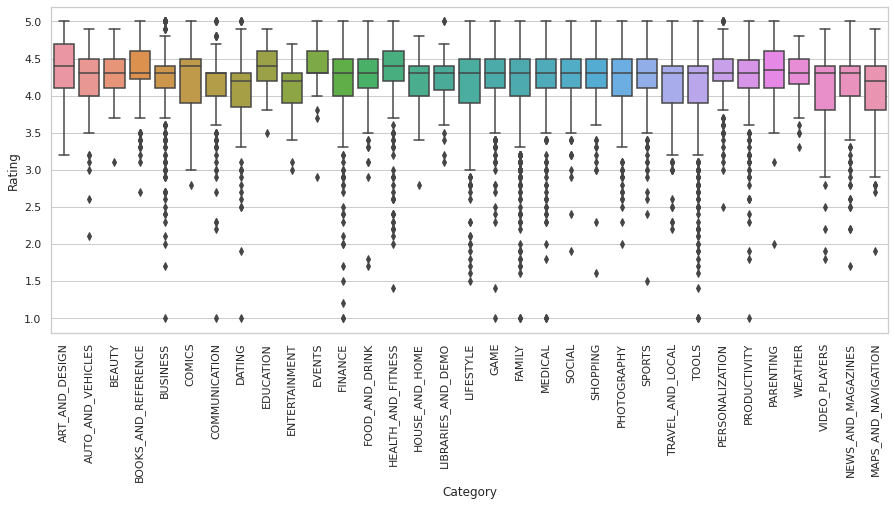


Here we plot 2 joint-plots one is between price and sizes and other is between price and rating

****

After Comparing both plots we can conclude that most paid apps are smaller in size this means that most paid apps are designed and developed to cater to specific functionalities. Most of the paid apps which are light weighted are highly rated. Hence light weighted paid apps will perform well in market.

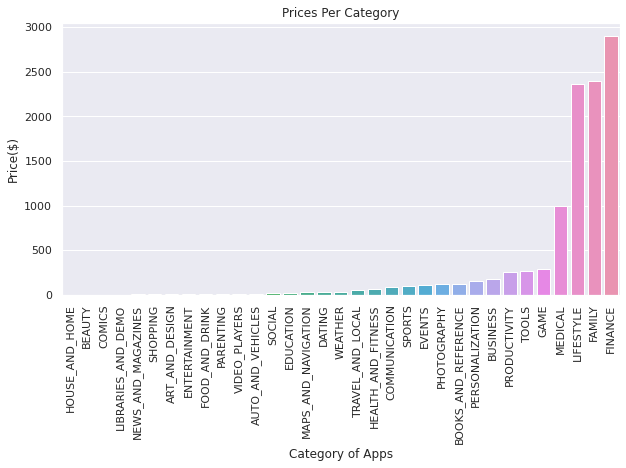
1. **Rating Analysis:**

****

For rating analysis we have plotted a box plot to see which category is performing well in market.

Education, Health & Fitness, Art & Design are the categories which are performing well in the market as their average rating are higher from average rating of all other categories. Business, Comics, Tools, Finance are the categories in which 50% of the apps have lower rating then their respective median rating or average rating.

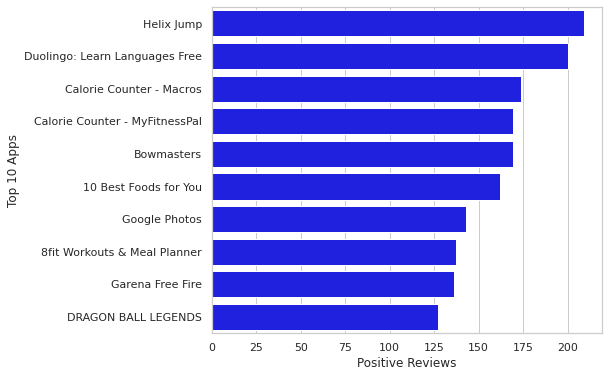
1. **Paid apps in each Category:** We also see which category contains most expensive apps.

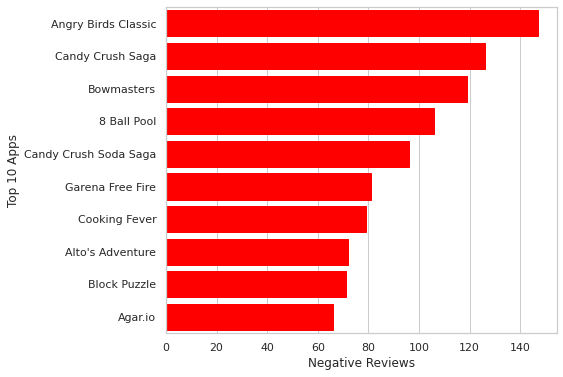


Most Paid apps is for category **FINANCE** and **LIFESTYLE** but their ratings are not so good (result from rating analysis), some of the **FAMILY** apps are there too in top ten most expensive apps

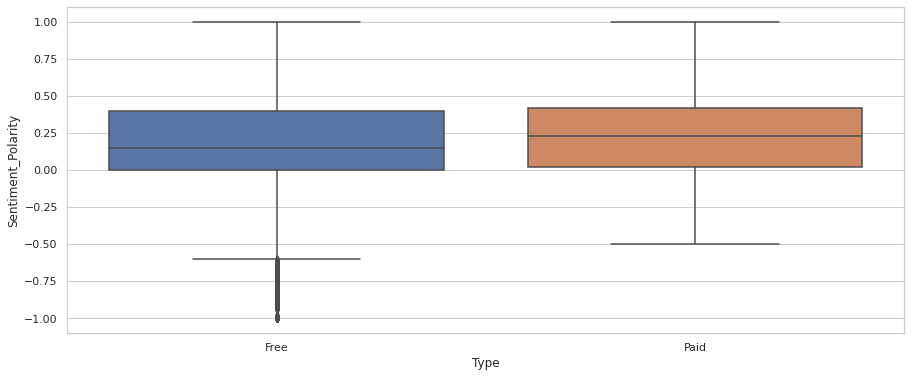
**Now we are analyzing the second dataset provided to us.**

* In the first Plot we see a comparison of top 10 apps having positive and negative reviews. In the Positive review section, the apps are from different Categories but In the Negative review section, the apps are mostly from GAME Category.





**In the size analysis we saw that (~90MB to ~ 100MB) have** most number of reviews although this category has least number of apps and also the installs are not that remarkable. This is because mostly heavy weighted apps are in GAME Category and this Category got most no of Negative Reviews, we can say that the games apps which comes in this size group may not perform well in market.

In the next plot we plot the box plot between subject polarity versus type of apps

We can clearly see the outliers in the free apps which are more towards the negative polarity meaning negative review. Hence we can conclude that users are harsher towards free apps whereas users are more tolerant when they are paying for it.

**5. Challenges Faced:**

1. Understanding the problem statement and reading the dataset. Data cleansing was the main problem we faced.
2. Handling the error, duplicate and Nan values in the dataset.
3. Those rows were dropped which has Nan values from both dataset, minimizing our data.
4. There is so much more which can be explored. We have the current version and the Android version available, which can be investigated in greater detail and lead to additional analysis where we can explain how these things affect and need to be considered when designing apps for users.

**6. Conclusions:**

* More Than **90%** of Apps are Free In play store and almost all apps have content rating ‘Everyone’.
* The most competitive Category is **FAMILY.**
* The most no of Installs (in sum) is for **GAME**. The **COMMUNICATION** Category has quite less no of apps but this category has 2nd in most No of Installs.
* Light weight size games apps perform well in market.
* Most Number of reviews are also for **GAME** Category followed by **COMMUNICATION** and **TOOLS.**
* Most number of heavy apps in sizes is in category **FAMILY** and **GAME.**
* **(~1MB to~10MB)** size group has **more no of apps** but their **Installs and review are less**.
* User **prefers** to install those apps which has a size between **(~10MB to ~ 20MB).**
* . **(~90MB to ~ 100MB) have** most number of reviews although this category has least number of apps and also the installs are not that remarkable**. (This is because mostly heavy weighted apps are in GAME Category and this Category got most no of Negative Reviews)**, we can say that the games apps which comes in this size group may not perform well in market.
* Majority of the paid apps that are highly rated have small sizes.
* Users prefer to pay for apps that are light-weighted.
* **Education, Health & Fitness, Art & Design** are the categories which are performing well in the market as their **average rating are higher from average rating of all categories**.
* **Business, Comics, Tools, Finance** are the categories in which **50%** of the apps have lower rating then their respective median rating or average rating. Interestingly dating apps also does not perform well in market.
* User are harsher towards free apps

**7. Refrences:**

* GeeksforGeeks
* Stack overflow
* Alma better Verified projects on play store app review analysis.
* Python libraries documentation