**EDA on Google Play Store review analysis by** AMOL THAKUR

# **Abstract :**

The data from Play Store applications has a huge potential to propel app-making companies to success. For developers to work on and take advantage of the Android market, useful information may be derived.

Category, rating, size, and other information are present for each app (row). Reviews of the Android apps from customers are included in another dataset.

Discover the major elements influencing app engagement and success by exploring and analysing the data.

Introduction :

Google Play, often known as the Google Play Store and formerly known as Android Market, is a digital distribution platform that Google manages and creates. Users may browse and download apps created with the Android software development kit (SDK) and published through Google using this as the official app store for certified devices using ChromeOS as well as certified versions of the Android and its derivatives operating systems. In addition to delivering music, books, movies, and television shows, Google Play has also operated as a digital media shop. The Android and iOS applications and a web browser may be used to access content that has been purchased on Google Play Movies & TV and Google Play Books**.**

Google Play offers applications that may be downloaded for free or for a fee. They may be installed on a device directly from the Google Play website or through the proprietary Google Play Store mobile app for Android devices. Users of devices with certain hardware components, such as a motion sensor (for motion-dependent games) or a front-facing camera, might be targeted by applications that leverage the hardware capabilities of a device (for online video calling). In 2016, there were over 82 billion app downloads from the Google Play Store. In 2017, there were over 3.5 million applications released, however following a purge of apps, the number of users is back up to over 3 million, depending on how severe the purge was.

# **Description of Datasets:**

There are two datasets that we would be dealing with in genreal. One of them is Playstore.csv and other one is User\_reviews.csv all the information regarding these two datasets are given in tables below :

|  |  |  |
| --- | --- | --- |
| No. | Column Name | Description |
| 1 | App | Name of the App |
| 2 | Translated reviews | User reviews for app |
| 3 | Sentiment | User sentiments |
| 4 | Sentiment Polarity | Polarity shows weather the review is negative , positive or neutral |
| 5 | Sentiment Subjectivity | Subjectivity refers to the degree to which a review is related to the app. |

|  |  |  |
| --- | --- | --- |
| No. | Column Name | Description |
| 1 | App | Name of the App |
| 2 | Category | Category of App |
| 3 | Rating | Rating of App |
| 4 | Reviews | Total Number of reviews by users |
| 5 | Size | Size of app |
| 6 | Install | Number of installs |
| 7 | Type | Free or paid |
| 8 | Price | App price |
| 9 | Content Rating | Age creiteria of app |
| 10 | Genres | Genres under which app falls |
| 11 | Last Updated | Last date of update |
| 12 | Current Version | Current android version of app |
| 13 | Android Version | Android version required to run |

**User\_reviews.csv**

**Playstrore.csv**

**3.** **REMOVING UNNECESSARY SYMBOLS(+ , $ , M , k)**

We see that a few items in the columns Installs, Price, and Size include the special characters (+, $, M, and k). The columns are no longer entirely numeric as a result, which makes it challenging to utilise them in subsequent calculations.We're going to clean up our data now. In particular, the "," and "+" special characters in the Installs column and the "$" special character in the Price column must be eliminated.

After removing all the unnecessary symbols it was now time to change datatypes of all these columns to numeric so that it would be easier for us to visualize this data.

**4. REMOVING DUPLICATE VALUES**

The next step was to remove duplicate values , it is necessary to remove duplicate values because they have the tendency to generate false or incorrect insights.

There were alot of duplicate values present in our dataset which we needed to remove , so I removed all the duplicate values that were present while retaining the first original value.

**5. TRANSFORMING VALUES FROM kb TO Mb**

Now it’s time to convert all the values present in kb to mb present in the column ‘Size’ , this is necessary to perform because we wanted to convert all the values to a single format for the ease of computing and visualizing.

For this I divided all the values that were present in kb format with 1000 so that all of them would come across in a single format.

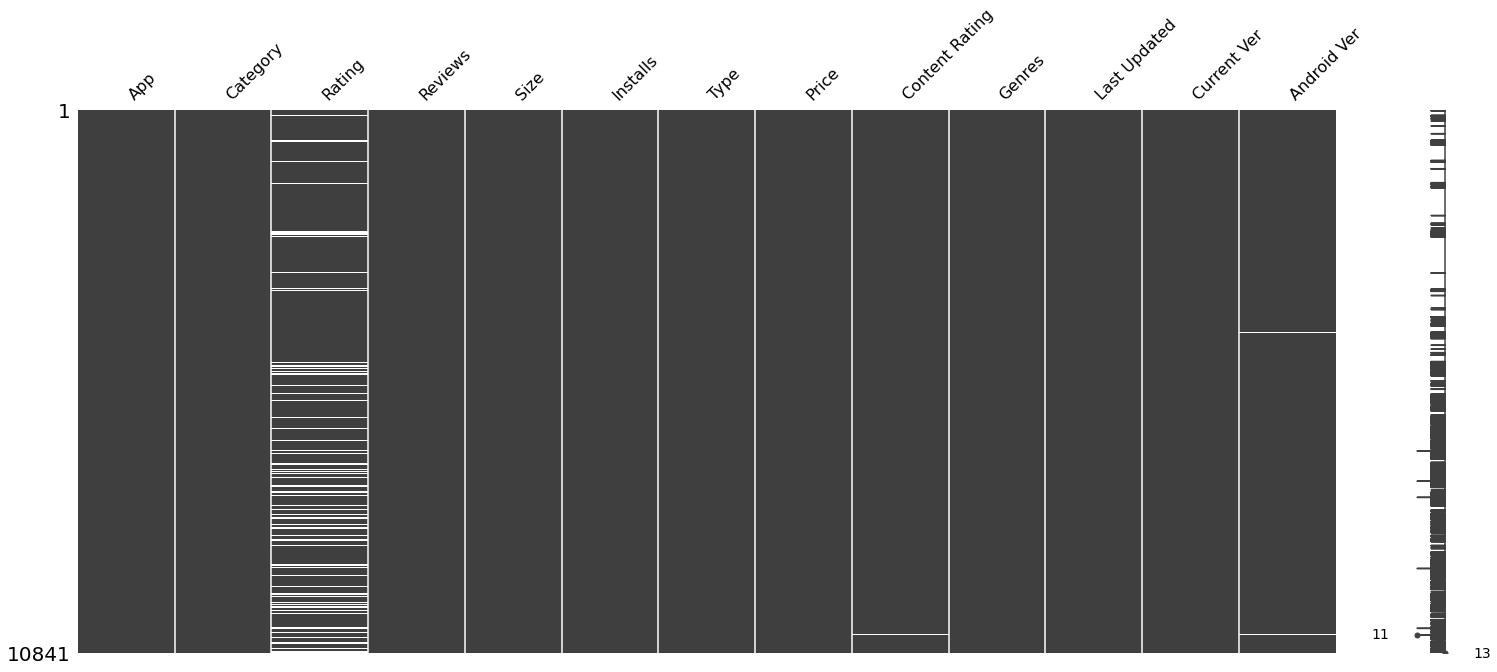
**DATA CLEANING:**

One of the most crucial tasks in any data science endeavour is data cleaning.Despite the fact that it might be a highly time-consuming procedure, it should be always performed.Now on moving further when we take a look at our Playstore.csv dataset we can see there’s a lot of data cleaning required before we perform data visualization.

1. **ANALYSING AND HANDLING NULL VALUES**

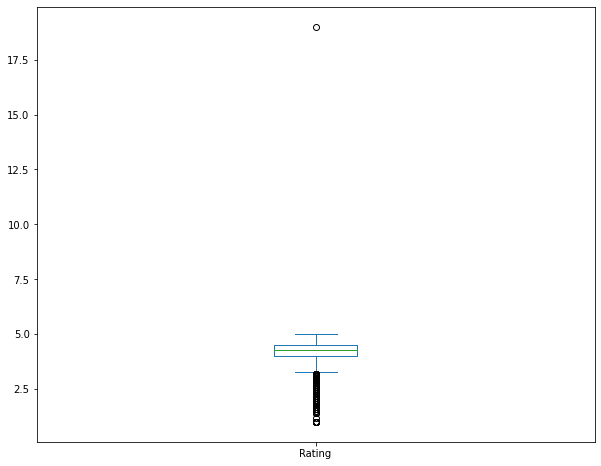
With the help of missigno library we visualized null/missing values to have a better understanding of our data.

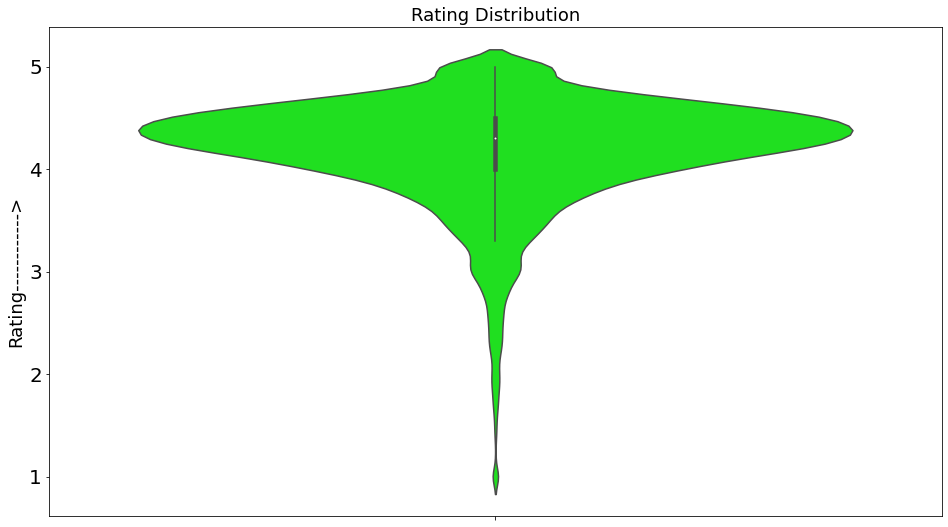
Here we can see that the “Rating” column is missing a lot of values.

****

1. **REMOVING OUTLIERS**

After the above operation now it was the time to remove outliers , i used boxplot to detect any possible outlier that might be present in our dataframe and guess what I found one , it was in column “Rating” , so i removed it using filtering.





**OBSERVATION 3 :**

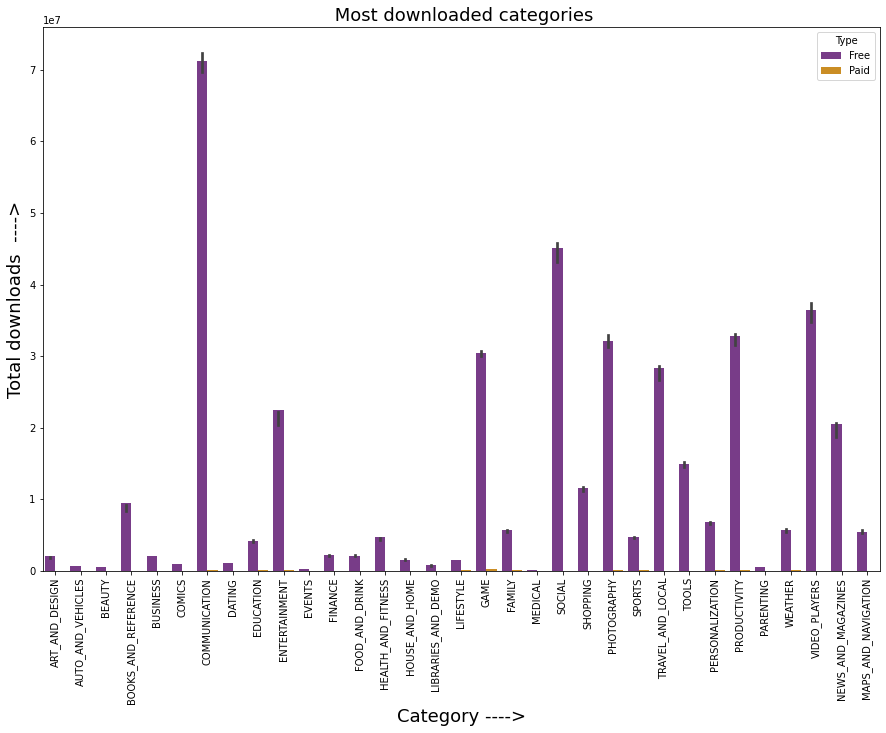
Here we can notice a trend that a large proportion of the top 20 most expensive apps have kind of similar names and price , that might be because of same functionality or origins.

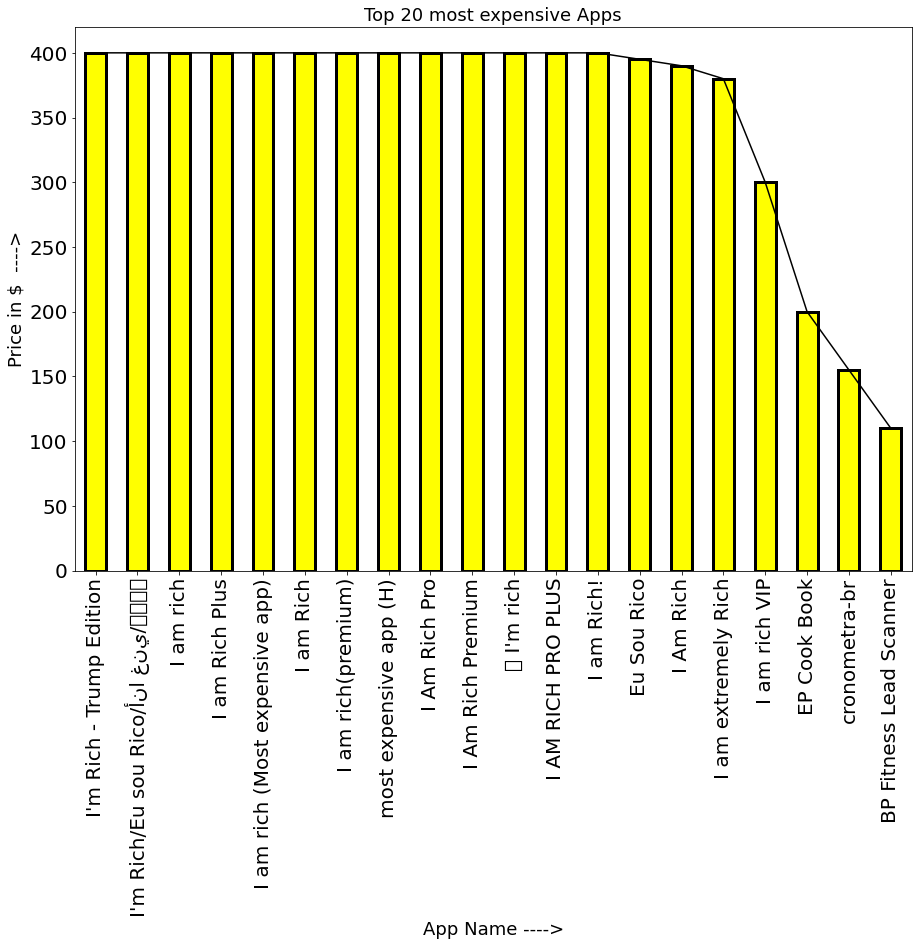
Initially the graph is linear but at 14th observation it starts declining.

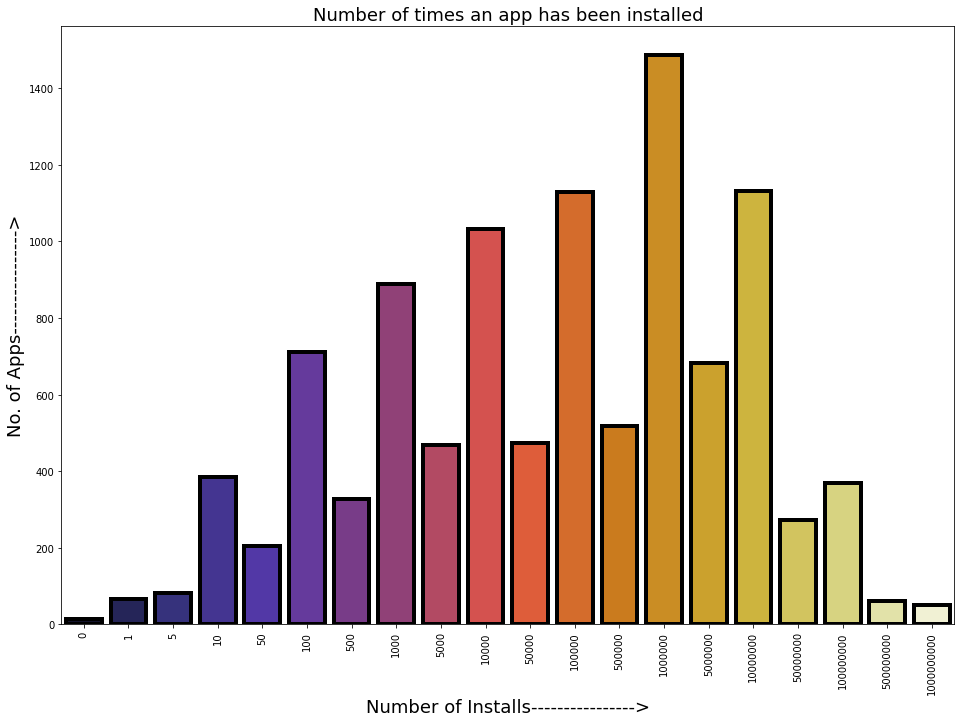
**OBSERVATION 4 :**

Here we can notice that the most downloaded category is COMMUNICATION as people love to communicate via video calls , texts and calls and the least downloaded category is MEDICAL.

Now let's find out what are the most downloaded apps that are present in COMMUNICATION category so we can have an idea why this category tops the charts.







**DATA VISUALIZATION:**

**Data visualisation is a method of presenting information graphically, emphasising patterns and trends in data, and assisting the reader in gaining rapid insights as not all of us might be good in understanding complex set of codes.**

**OBSERVATION 1 :**

Here violin plot is used to study the distribution of ratings, from above plot we can conclude that most of the ratings ranges from 4.3 to 4.4 and that's where the median also lies, because of univariate analysis and to study the distribution we have used violin plot here.

**OBSERVATION 2 :**

Here graph shows the number of apps on Y axis and the number of times those apps have been installed on X axis

The first major observation on y axis is 200 which is coinciding with the value 50 on x axis, this means that there are 200 apps which have been downloaded more than 50 times that are present in our dataset.

The highest value shows that there are near about 1500 apps that have been installed more than a million times.

**OBSERVATION 5 :**

So here are the top 10 most downloaded apps in COMUNICATION category.

Most of the apps that are present in our Top 10 list comes pre installed in a lot of devices which might be the reason for these apps to have near about a billion installs.

**OBSERVATION 6 :**

Even though COMMUNICATION is the most installed category but when it comes to ratings the stats says otherwise, the bar graph above clearly shows that the apps with category EVENTS tops the chart with highest average rating of 4.43 .

The lowest rated category is DATING with an average rating of 3.97.

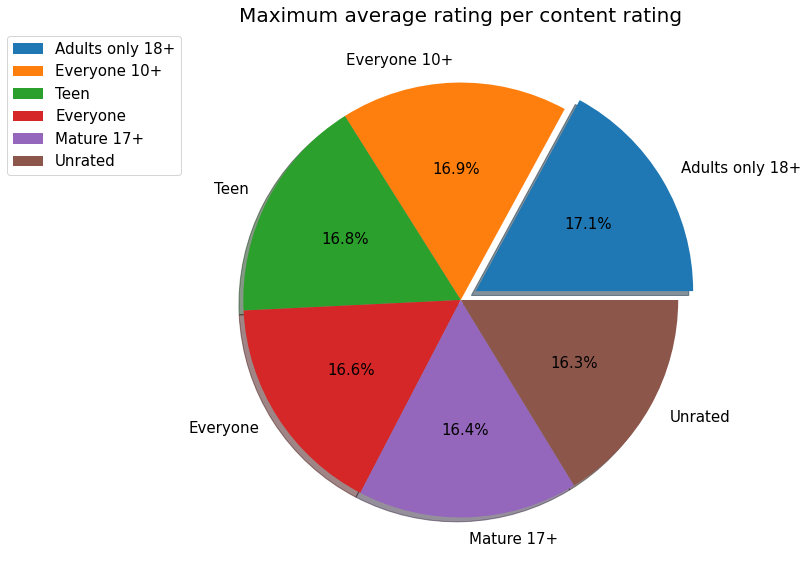
**OBSERVATION 7 :**

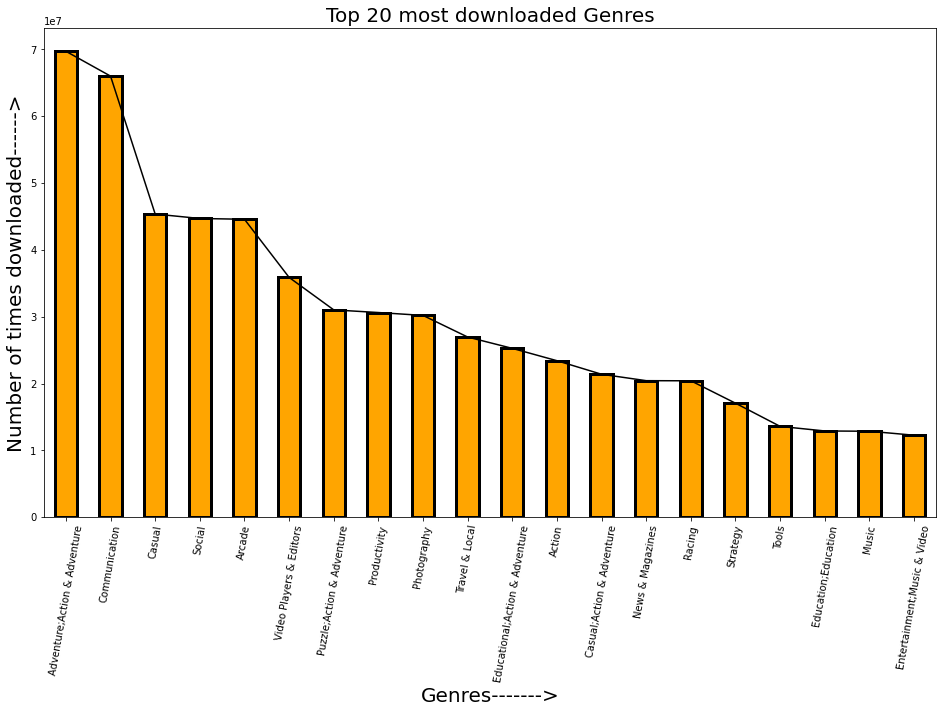
Bar plot above shows the most downloaded genres. It could be noticed that ,here the most downloaded genre is Adventure;Action & Adventure with more than 6.970385e+07 installs and unlike the bar graph which was plotted for categories in which COMMUNICATION topped the list here it is 2nd which is quite interesting .

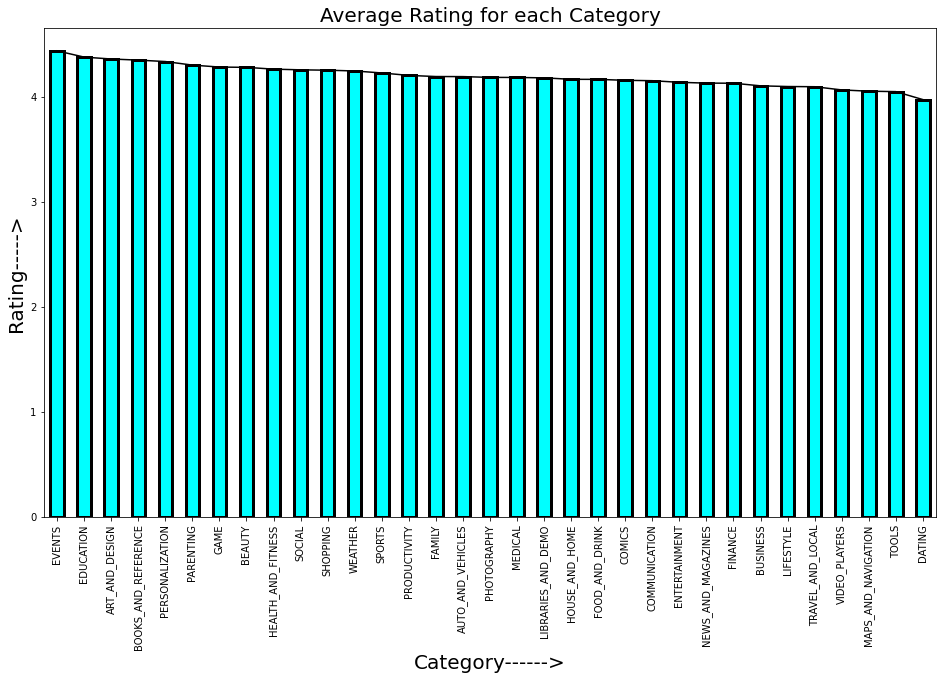
This means that irrespective of some of the values being similar in both 'Category' and 'Genres' columns there's no actual relation between them and both are independent of each other.

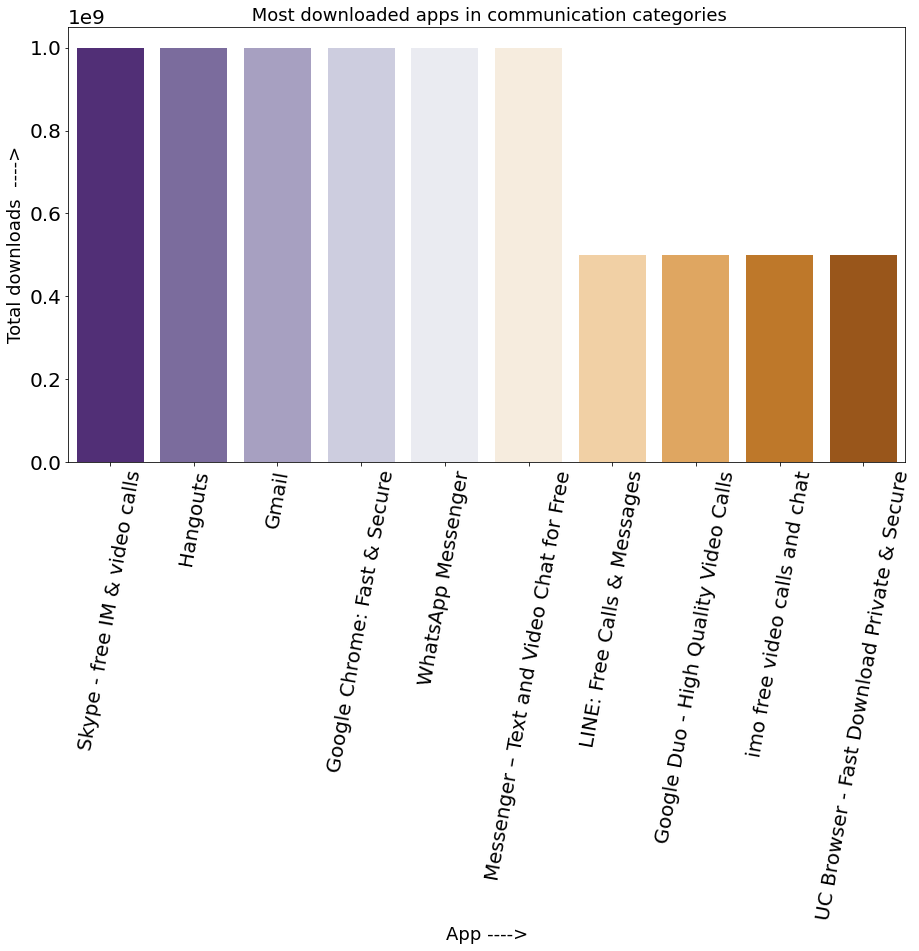
**OBSERVATION 8 :**

Above pie chart shows that content rating 'Adults only 18+' is highest rated among all other categories, but this does not mean that it is the most downloaded content rating, data analysis can be sometimes misguiding or tricky isn't it ?. So let's find out what's the most downloaded content rating is.









**OBSERVATION 9 :**

The above pie chart shows that the most downloaded content rating is "Everyone 10+ " as it acquires 37.7% of area in pie chart , it could be a very useful information for App Devs as they would focus more on building apps that falls in this content rating.

We can also notice that content ratings "Unrated" and "Adults only 18+" are the least downloaded categories , this is a bit surprising because of the fact that "Adults only 18+" was the highest rated content rating in our previous analysis yet it's not the most downloaded one and now it's clear that due to low number of installs the average ratings has spiked up .

**OBSERVATION 10 :**

The vast majority of apps present on google play store are free as we can see from above pie chart ,Free apps(92.6%) are dominating Paid apps(7.4%) because users prefer apps that are free, it's quite understandable from here why devs focus more on building free apps rather than paid.

**OBSERVATION 11 :**

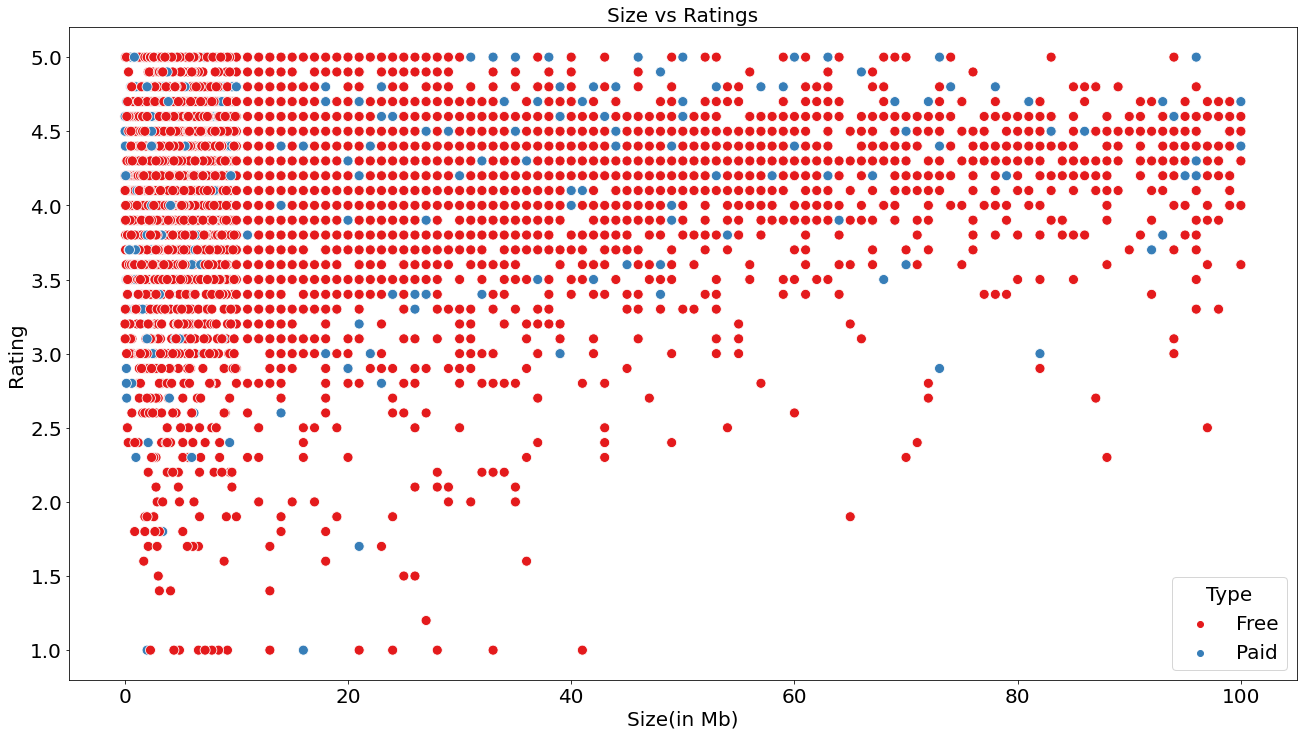
From the above kdeplot it's clear that the probability of an app being rated above 4 stars on play store is high if the app is free as users prefer free app against paid ones.

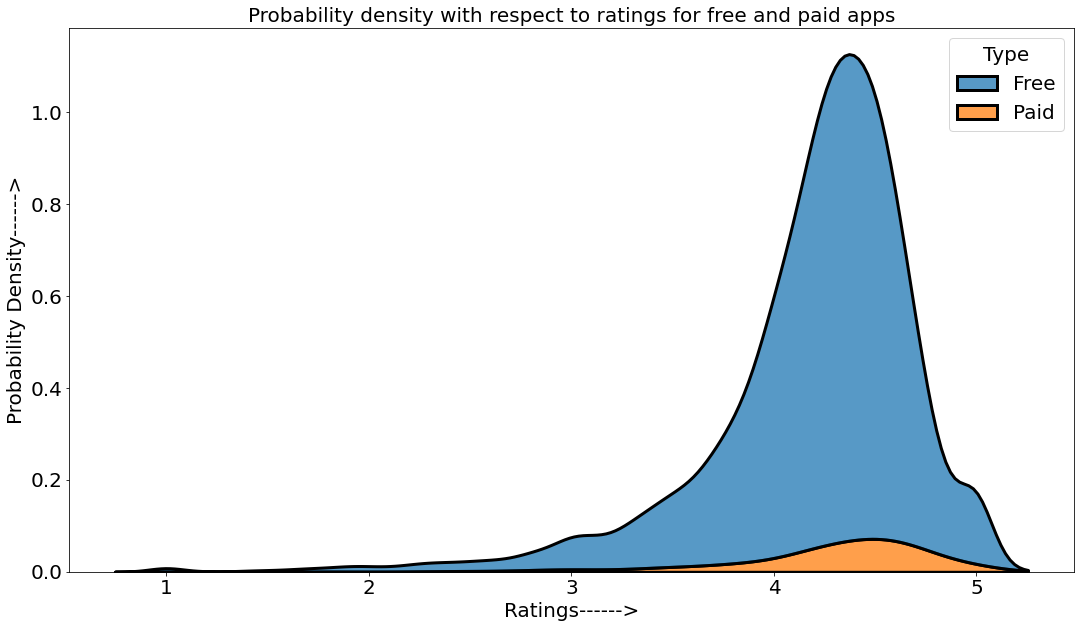
The probability density of a paid app being rated above 4 stars is extremely low as we can see in above kde plot, that might be because of lack of users using paid apps.

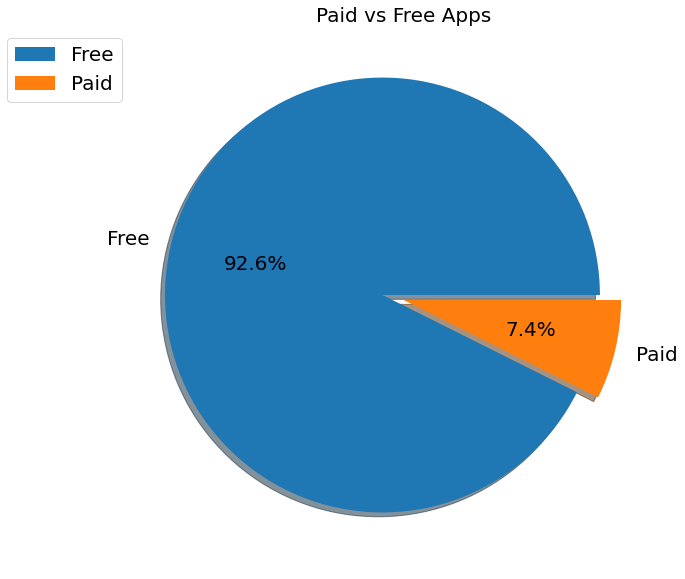
So app devs should focus more on making free apps or affordable paid apps so more users could download them.

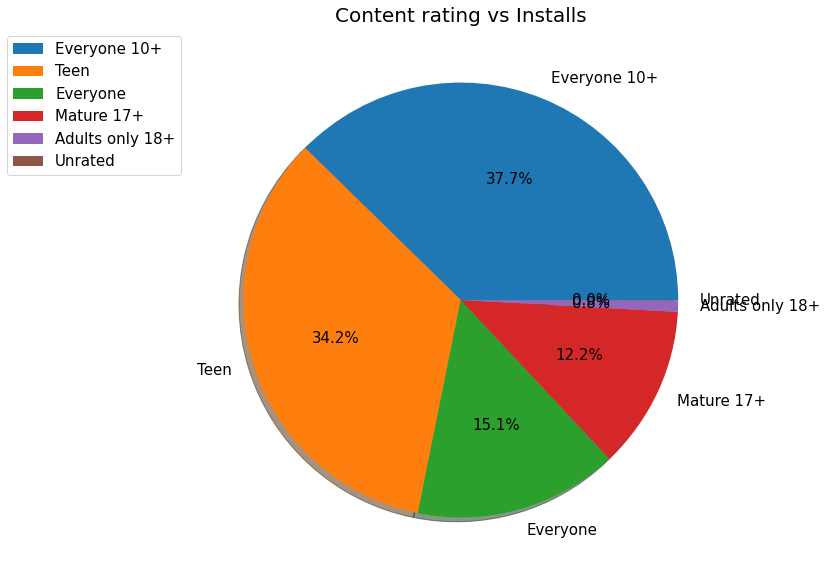
**OBSERVATION 12 :**

Above scatterplot shows the correlation between size and ratings and how these two affects each other.As we notice when the size of an app increases the number of ratings or density of rating decreases that's because users avoid installing apps with big size and prefer using apps with smaller size to save data or storage on their phones.After 40mb the rating density is decreasing which indicates in lack of users or lack of downloads/installs.In case of paid apps the rating density is also slightly decreasing after 50mb which indicates that even paid users don't like apps that are big in size.

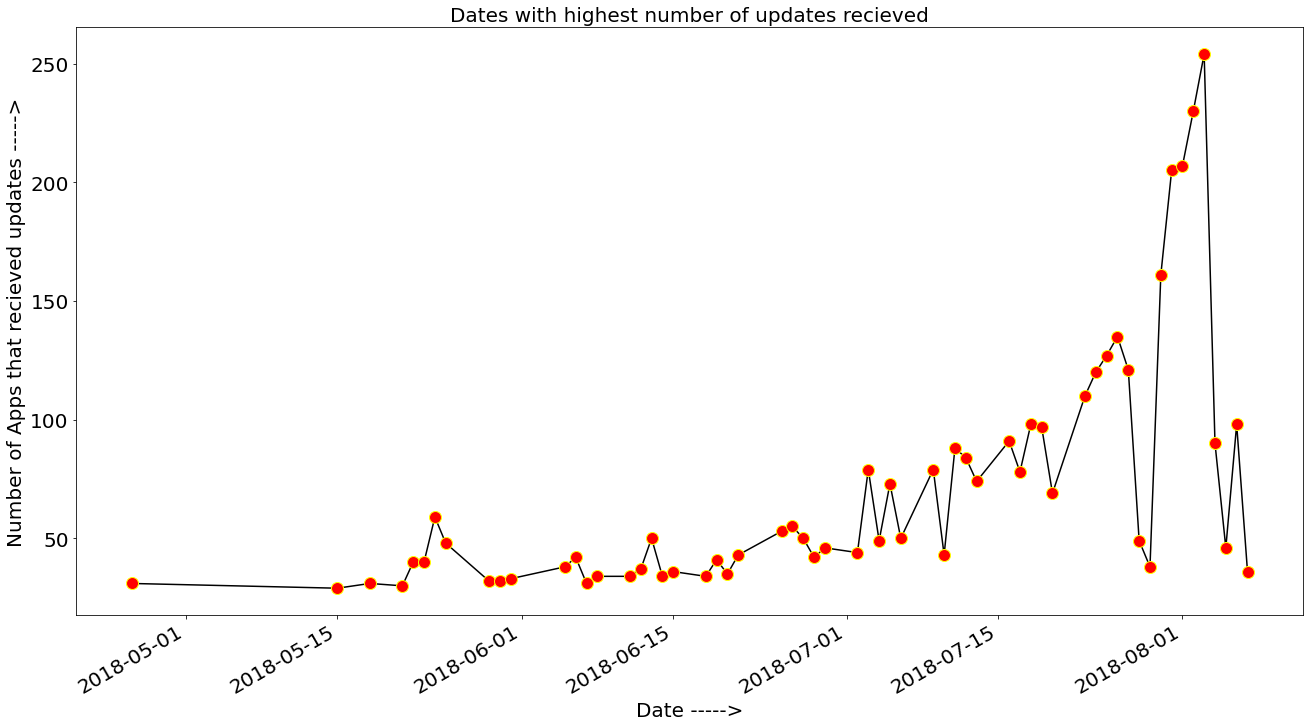












**OBSERVATION 13 :**

From the above scatterplot we can conclude that 'Installs' and 'Size' are inversely proportional to each other because as the size of app increases the number of installs decreases, we can clearly see that after size 40mb the density of Installs is decreasing in case of both free and paid apps. From 60mb to 100mb we can notice extremely low density of installs.

The most dense area of this scatterplot is between 0mb-10mb , this range of apps have the most chances of being installed.

So Devs should focus on building apps with smaller sizes.

**OBSERVATION 14 :**

Line graph above shows the dates when apps on play store received the most amount of updates.

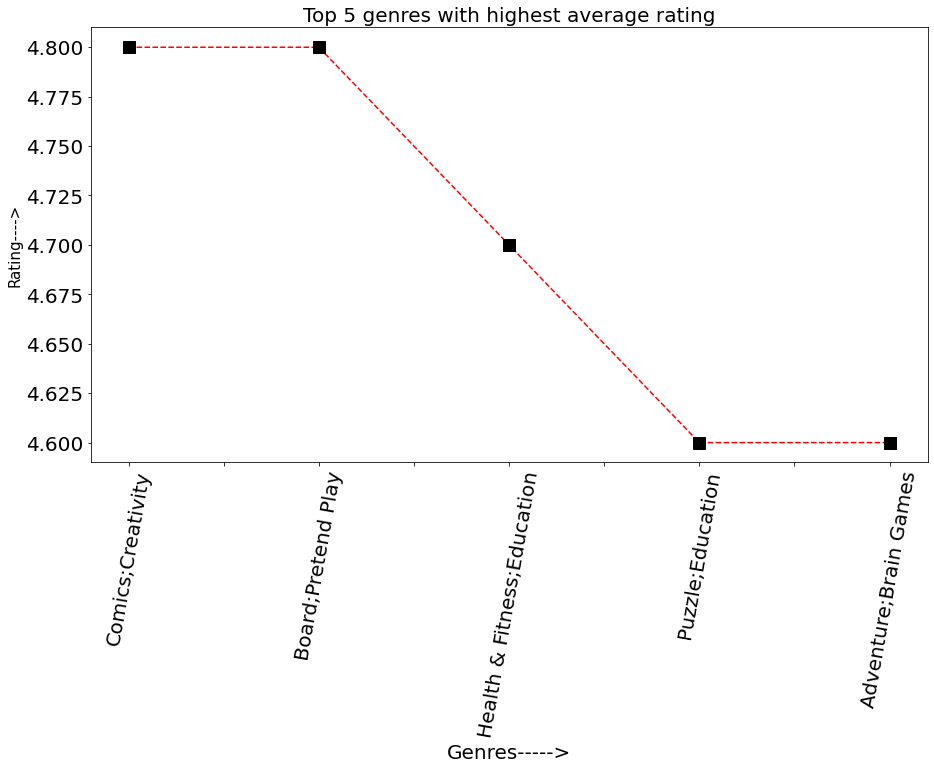
According to this line graph the most number of updates received on google play store was on date 2018-08-03 i.e 254 apps received updates on this day.

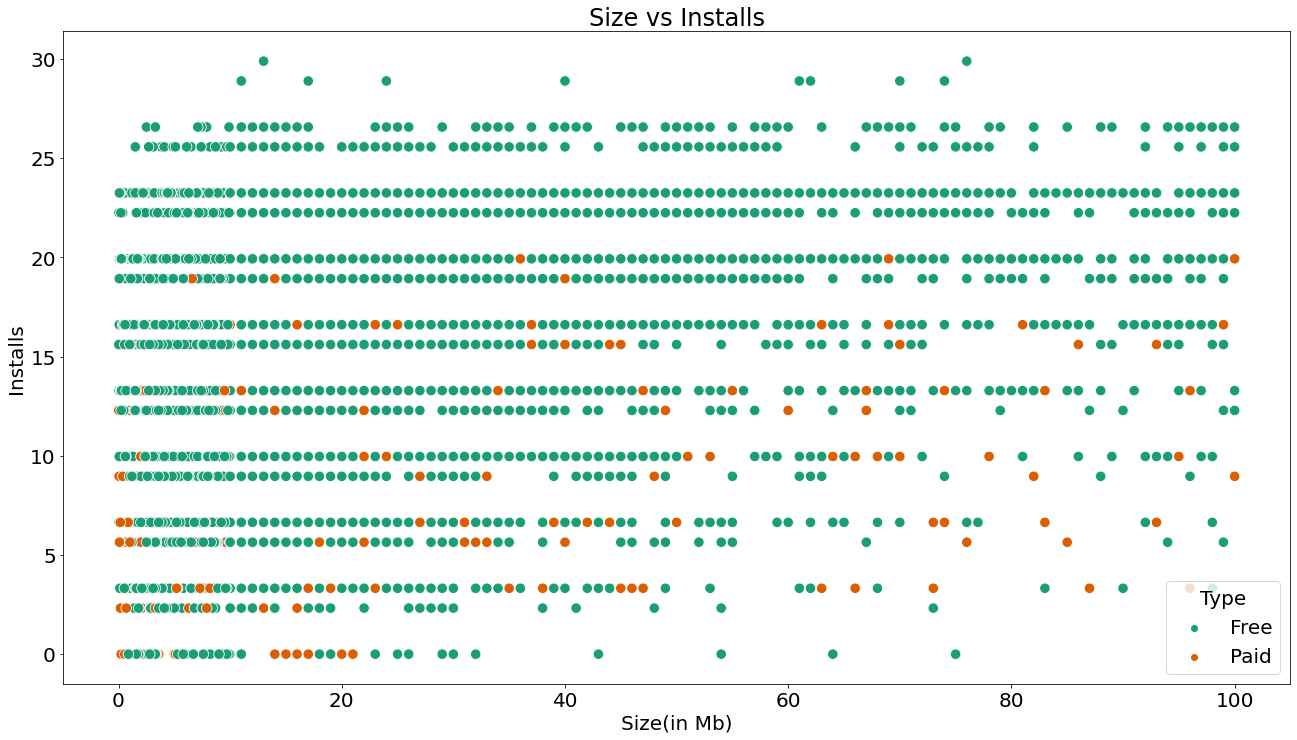
**OBSERVATION 15 :**

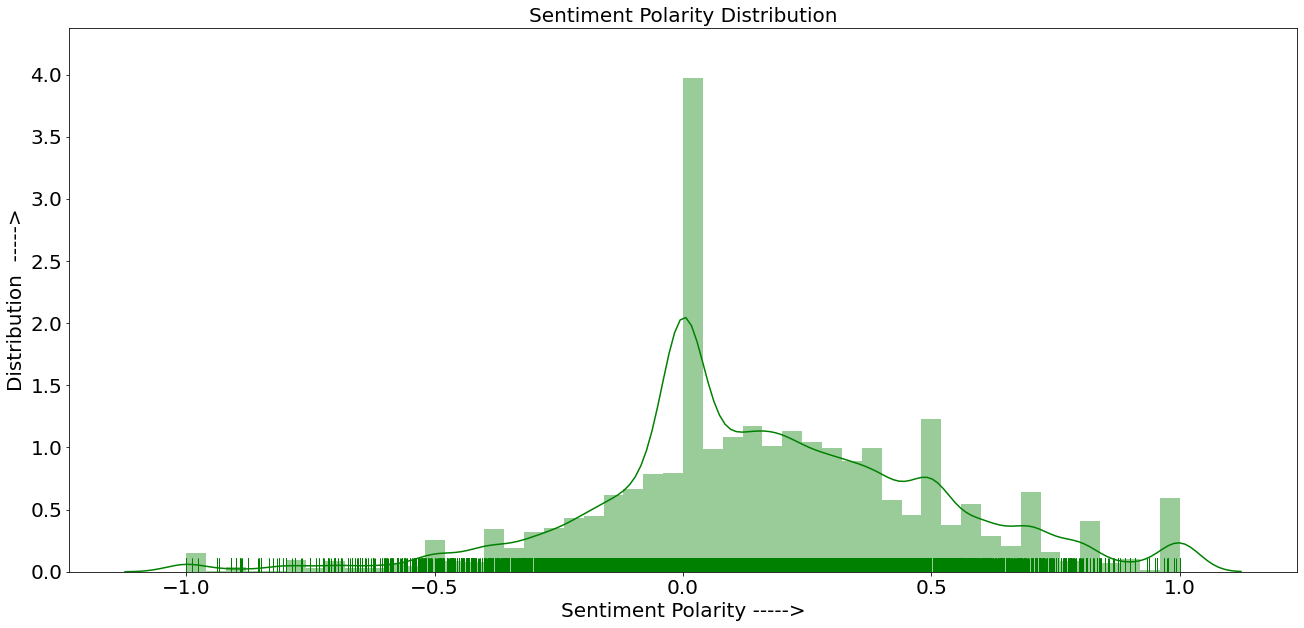
Here are the genres with highest average ratings where Comics;Creativity tops the chart

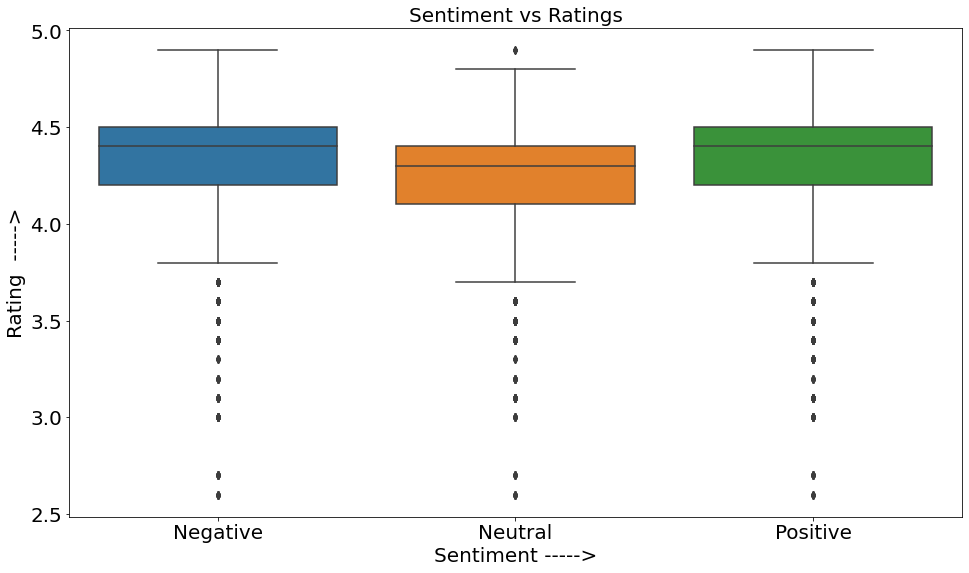
**OBSERVATION 16 :**

From the above heatmap we can conclude that 'Installs' and 'Reviews' share a strong correlation of 0.63 which means that they are directly proportional to each other, in general if the number of reviews will increase then the number of installs will also increase and vice-versa.Price($) column has negative correlation with every column , that's because if price of an app increases then the number of installs will decrease which will further affect ratings and reviews, that's why most of the users prefer using free apps instead of paid ones.In case of Ratings and Reviews they share positive correlation with every column (except Price($)), this is because if ratings and reviews will increase then there will be more probability of users installing a certain app.The correlation between Reviews and Size(in Mb) is also positive (0.24), this might be because more number of reviews helps devs in understanding the actual needs of users which results in an app receiving more updates and features which further increases size of that app.









**OBSERVATION 17 :**

As expected the most number of sentiments are Positive(63.6%) followed by Negative(23.9%) and neutral(12.5%).

This is a good thing for devs as most of the users are satisfied by the services offered by apps.

**OBSERVATION 18 :**

It's kind of unusual that median rating for Negative sentiments(4.4) is higher than median rating for Neutral sentiments(4.3) and is equal to Positive sentiments(4.4). This might be due to the fact that when a user installed an app he initially liked it and gave it more than 4 stars and a positive review but after the next update he/she disliked it because of some reasons and decided to update their review but might have forgot to update the rating that they gave the app initially.

The 2nd reason might be due an inconsistency/error in our data or during data collection.

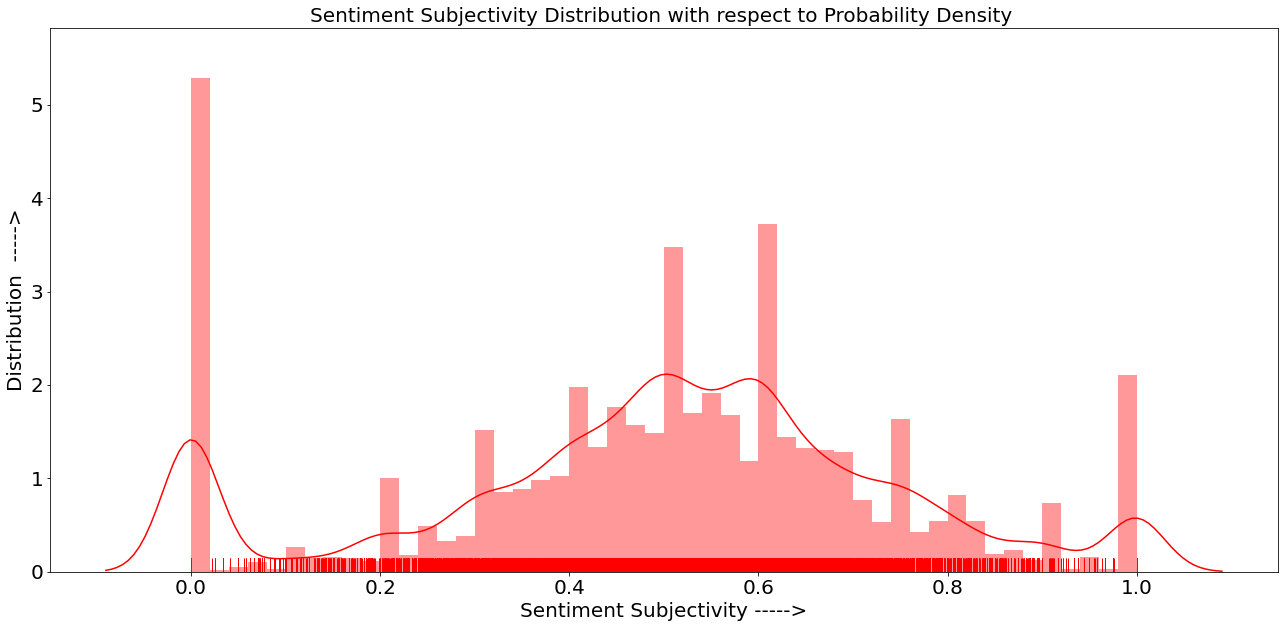
This could also be true that there might some ratings that were generated by bots.

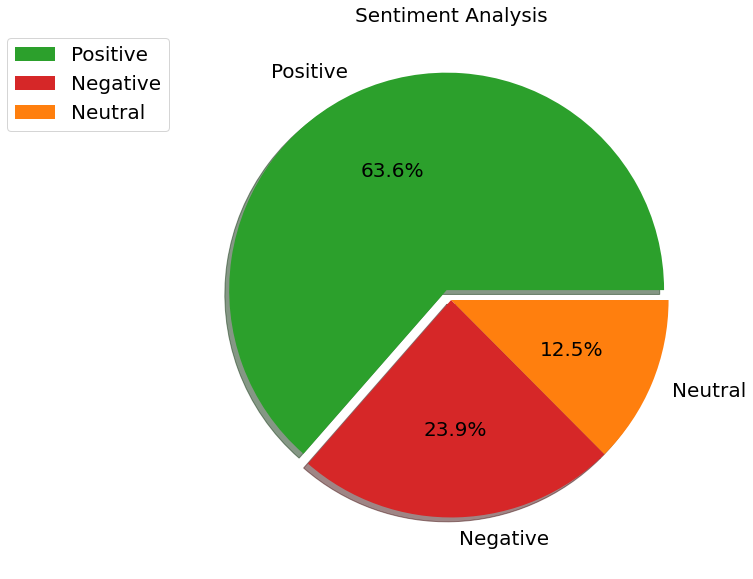
**OBSERVATION 19 :**

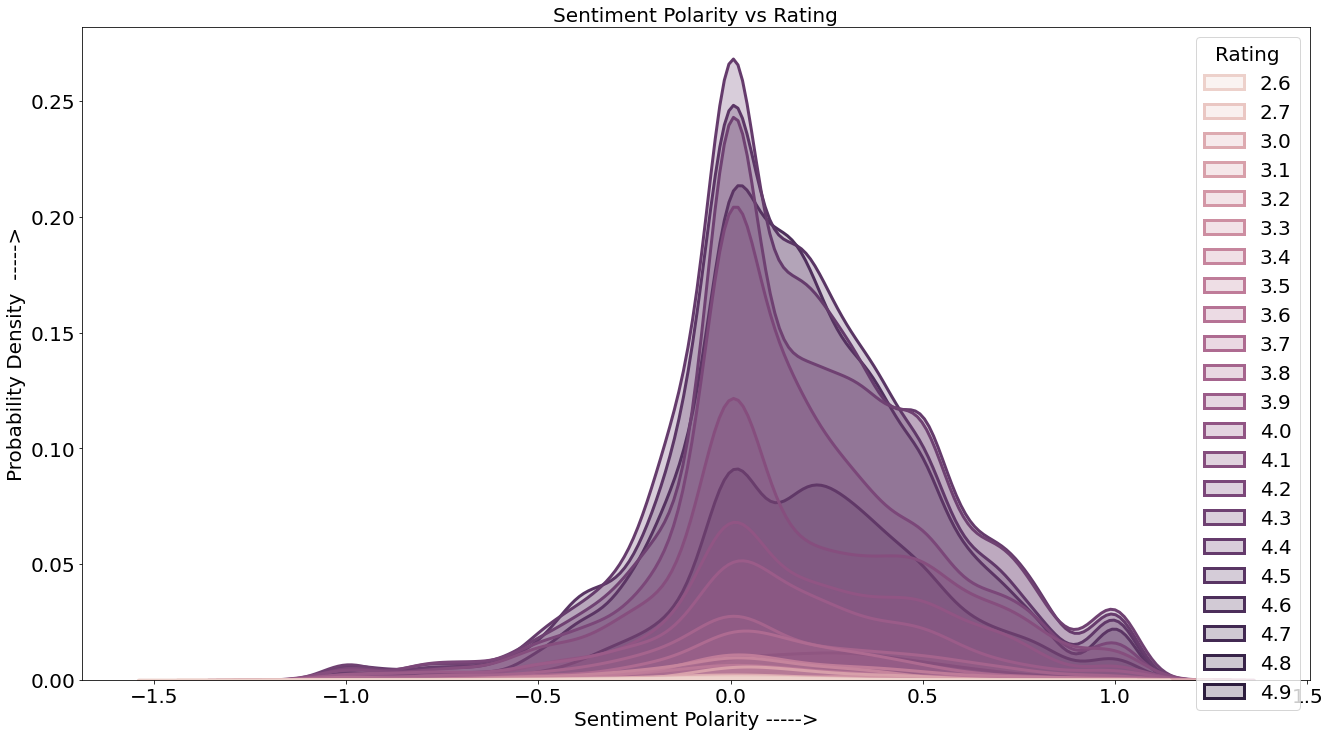
The dist plot here show some interesting and unusual trends. The most common sentiment subjectivity score is 0.0 as it has the highest number of counts ,but the kdeplot shows that even though 0.0 is the most common value, the probability of sentiment subjectivity being between 0.4 to 0.7 is much higher than the probability of it being 0.0 due to combined counts of values present between 0.4-0.7 .

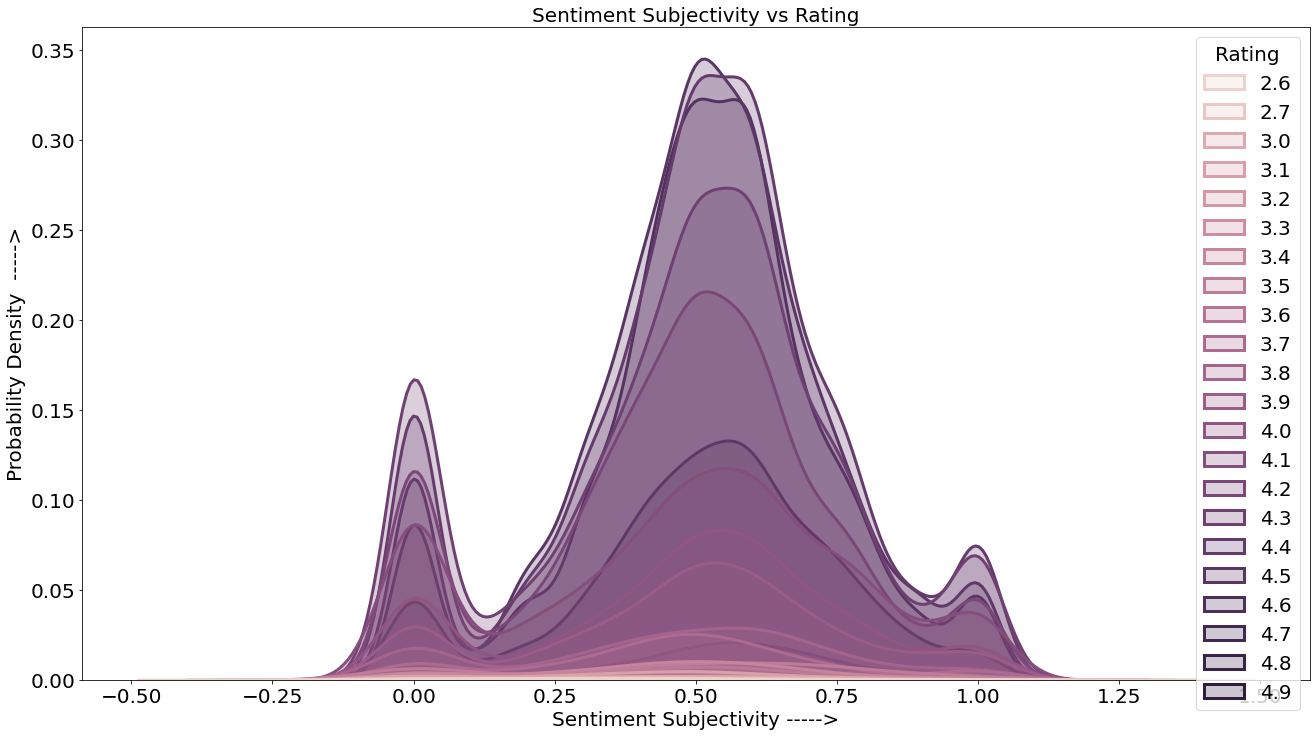
**OBSERVATION 20 :**

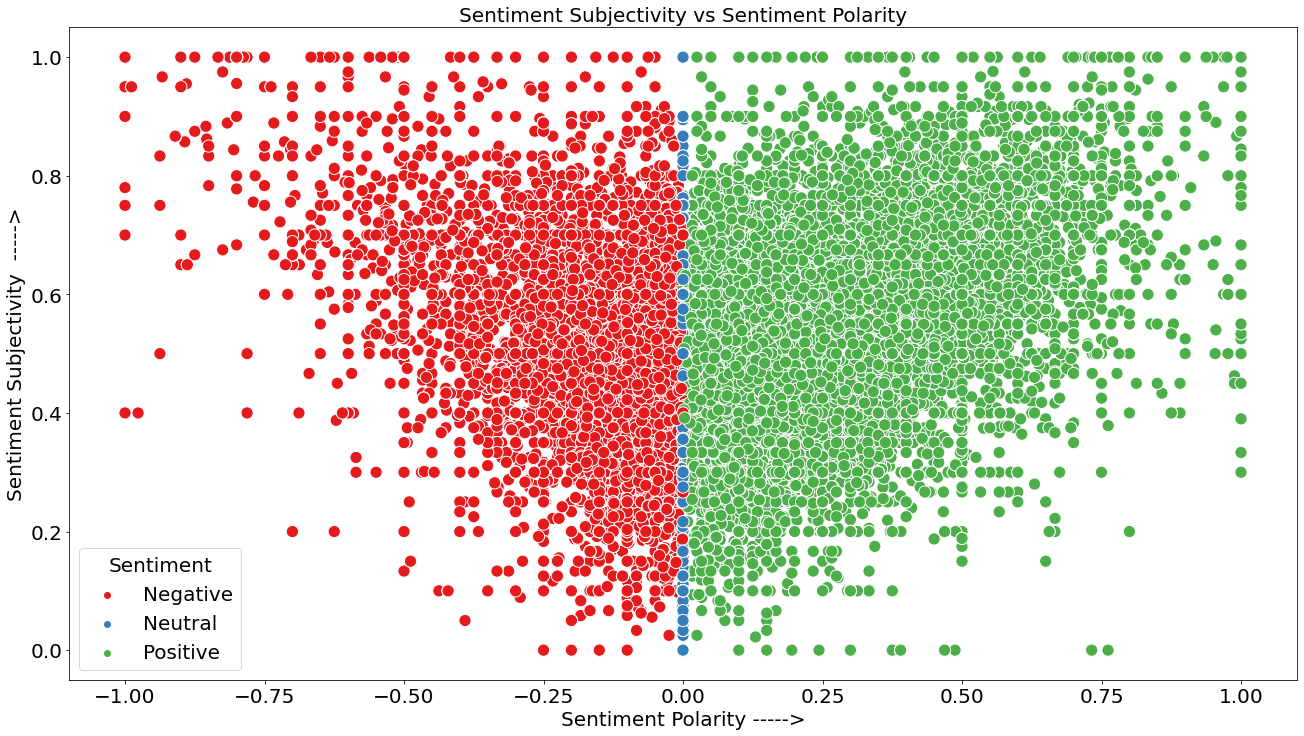
From the above plot we can see that most of the user reviews have a neutral sentiment polarity and the probability of sentiment polarity being 0.0 is the highest which means for any new review the possibility of it being neutral is the highest.











**OBSERVATION 21 :**

Kde plot shows that the probability of user reviews having sentiment polarity greater than 0.0 and sentiment subjectivity being between 0.2 - 1.0 is highest because most of the values lies between Sentiment Polarity(0.0-1.0) and sentiment subjectivity(0.2-1.0), which means that most of the reviews would be positive and slightly or heavily related to the app

We can also see that as the sentiment polarity is increasing towards the positive side the sentiment subjectivity is also increasing which means that the reviews are positive and related to apps. This is a good thing for devs as higher sentiment subjectivity gives them a more brief understanding about the problem users are facing and positive sentiment polarity motivates them to keep on working.

**OBSERVATION 22 :**

After analyzing the above distribution we can see that most of the sentiments are positive and highly subjective to the app.

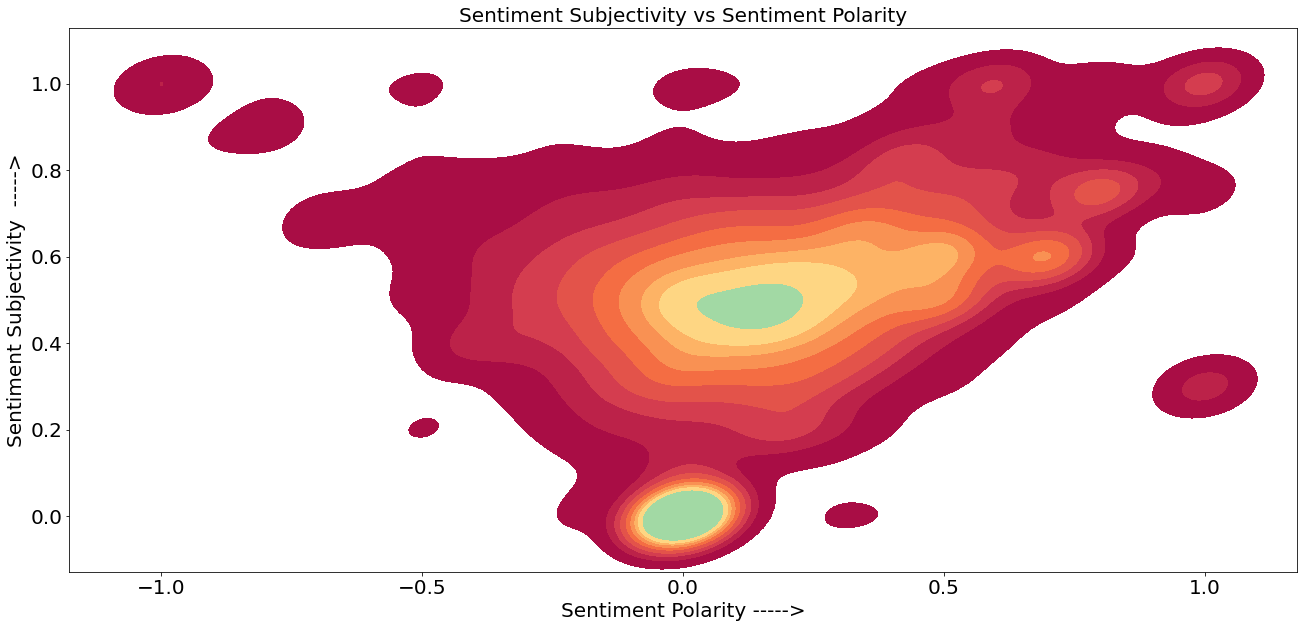
The sentiment polarity is highest between 0.0-0.75 while subjectivity is highest between 0.2 - 0.8 , meanwhile reviews with extremely positive polarity(1.0) and extreme high subjectivity(1.0) or extreme low polarity (-1.0) and extreme low subjectivity (0.0) have low density, which means that there's lack of extremely relevant and extremely irrelevant reviews.

**OBSERVATION 23 :**

The probability of sentiment subjectivity being 0.50 and having rating more than 4 stars is the highest , which means that users whose reviews are moderately subjective to the apps have higher tendency to rate apps more than 4 stars.

**OBSERVATION 24 :**

This graph almost tells us the same thing i.e the tendency of users to rate an app above 4 stars is highest when the sentiment polarity is neutral(0.0) .



**CONCLUSION :**

**After analyzing the data, we were able to determine the answers to a number of significant and intriguing questions that any android user would find fascinating. We carried out numerous operations to obtain insights such as rating distribution, the top 20 most expensive apps, the most downloaded categories according to their type, the top 10 most installed apps in the COMMUNICATION category, and many more. Overall, learning about the app industry was enjoyable and greatly aided my understanding of this domain.**

**OBSERVATION 25 :**

Here we can see the sentiment polarity and sentiment subjectivity are highly related to each other, and are directly proportional to each other this verify the fact that all the above graphs that we plotted to depict the relation between both these variables were correct.



