**Play Store App Review Analysis**

**Team Members**

Sourav Karmakar, T.Ompriya Subudhi, Thasniya Mansoora, Meenu Benny,

*Data science Enthusiast, Almabetter*

**Abstracts:**

Today around 85% cell phones are running with the help of Android OS. Google play store plays major role in building today’s android app empire. Android is fundamentally an open-source Operating system. It offers engineers to build their own android app and make our day-to-day life easy and beautiful. So, to assist organizations with acknowledging how they can develop their business really in each viewpoint and become familiar with the requirements of the clients to develop their products further and design their app according to customer’s expectations.

**1.Problem Statement**

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.

Each app (row) has values for category, rating, size, and more. Another dataset

contains customer reviews of the android apps.

Explore and analyze the data to discover key factors responsible for app engagement and success.

The main objective of this project is to deal with the data provided by the

company and to analyze the data in every aspect possible so to help them match their idea and help them to convert new and retain old customers and make reasonable growth.

*We have been provided with two datasets:*

* 1. Play Store Data.csv
  2. User Reviews.csv

***Play Store Data.csv column elements: -***

**App** - Name of the Application

**Category** - Category of the Application

**Rating** - Rating given to the Application

**Reviews** - No of reviews given to the Application

**Size** - Size of the Application

I**nstalls** - No of downloads of the Application

**Type** - Free or Paid

**Price** - Price of the Application if it is paid

**Content Rating** - It is Age appropriate or Not

**Genres** - Type of Genre the Application belongs to

**Last Updated** - When the last time the Application is Updated

**Current Ver** - Current version of the Application

**Android Version** - Minimum Android version required to run the Application

Let's understand what things are in Play Store data set

**App** - Name of the Application

**Category** - Category of the Application

**Rating** - Rating given to the Application

**Reviews** - No of reviews given to the Application

**Size** - Size of the Application

**Installs** - No of downloads of the Application

**Type** - Free or Paid

**Price** - Price of the Application if it is paid

**Content Rating** - It is Age appropriate or Not

**Genres** - Type of Genre the Application belongs to

**Last Updated** - When the last time the Application is Updated

**Current Ver** - Current version of the Application

**Android Version** - Minimum Android version required to run the Application

**2. User Review Data.csv column elements:**

**App** – An app name

**Translated Review:-** Reviews being given by consumer

**Sentiment –** Sentiment given to an app by users ( i.e. Positive, Neutral, Negative)

**Sentiment Polarity** – The polarity of sentiment measures how negative or positive

the context is. In the data we have, the polarity ranges from +1(Positive) to

-1(Negative).

**Sentiment Subjectivity** - The subjectivity of a sentiment is how likely that sentiment

is to be based on data or factual information, versus personal opinions or public

notions

**Problem Questions:**

1. Best Rated App in Every Category
2. Top 5 Categories which has overall best ratings
3. Top Genres in Play store
4. Average Rating in Genres
5. best paid app of top genres
6. Free & Paid app of top genres
7. Last Update impact on Rating
8. content rating column
9. Find relationship between Rating & Reviews
10. Sentiment analysis between free and paid app
11. Top 5 categories of review sentiments
12. Most reviewed apps
13. Apps which have rating more than 4 and have reviews greater than 10000
14. variation of review with month in the year 2018
15. Sentiment column according to category and rating

**2. Introduction**

Android is an open-source operating system which has maximum potential to influence the whole world with its revolutionary features. Each application engineer has the opportunity to contribute to this android era.

In 2007, Google released a first beta version of the Android Software Development Kit (SDK) and the first commercial version of Android 1.0 (with name Alpha), was released in September 2008.Google consolidated the market with android in 2007 which have allowed designers to foster android applications and it opened the business sectors of chances for both the producers and the client. It creates and improves many new organization and lots of impossible difficult problems were solved for this android revolution.

Data has been provided by AlmaBetter.

From these analysis clients can understand the needs of their customers in different aspects of product development and it will help them to solve the problems with the market superiors.

The main objective of this project is to deal with the data provided by the company and to analyze the data in every aspect possible. It will help company to understand consumer behaviors, their buying pattern and give more clarity to developers regarding future important decisions.

**3 Types of Android Applications**

*Background Services* and *Intent Receivers Applications*

*Foreground Background Applications*

*Intermittent Applications*

The type of Application where we will do the analysis can be any of the above.

**4 Steps Involved in Data Cleaning:**

**EDA** (Exploratory Data Analysis) on two given Datasets.

Play Store Data (['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver'])

User review (['App', 'Translated Review', 'Sentiment', 'Sentiment Polarity', 'Sentiment Subjectivity'])

*After analyzing both of these datasets:*

* There are 10841 rows and 13 columns in Play Store Data and 64295 rows and 5 columns in User Review.
* Except for Rating all the column values are object type which is type string.
* Value of the last Updated column is in string.
* Value of ‘Size’ is in string.
* So float64 type column value is 1 and object type column value is 12.

**Treatment of null value**

Evacuation of invalid qualities is essential for each column to set compelling investigation yield up to end negative information and work on helpful information.

**Treatment of Duplicate Rows**

After evacuation of invalid qualities another significant undertaking is the expulsion of copy column values on the off chance that they exist. With the expulsion of copies, we can now push

ahead with the issue explanation without having pointless increased information.

**Conversion of necessary column values from string to required datatype:**

This is the essential task to get authentic data portrayal and examination too. Concerning what is going on beside Rating every single other column has a string type data type which isn't required. For example: Size we need to have this portion in float or int to analyze, equivalently with Last Updated column we truly need this column value in Datetime Format and not string to get fundamental data visualization.

**Replacing the missing values with mean, median or mode**

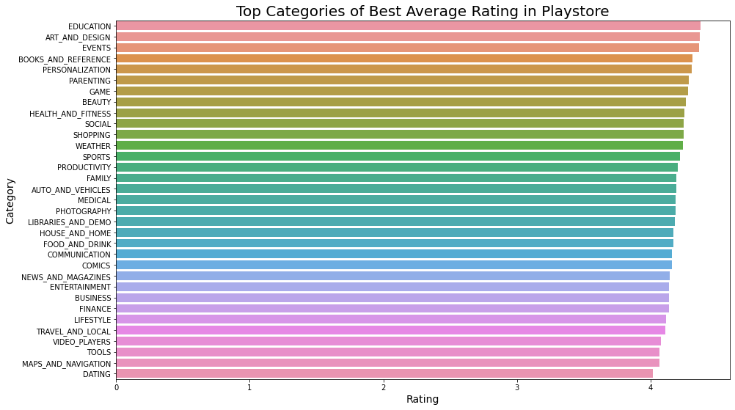
This strategy is additionally significant as far as getting output visualization. We can replace the missing value with either of the three depending on the data in the dataset. In this project we are displacing the missing values with mean value in Rating column.

**5.Top 5 Categories which have Overall Best rating**

**Top Categories of Apps**

|  |
| --- |
| * FAMILY 1939 |
| * GAME 1121 |
| * TOOLS 841 |
| * BUSINESS 427 |
| * MEDICAL 408 |
| * PRODUCTIVITY 407 |
| * PERSONALIZATION 386 |
| * LIFESTYLE 373 |
| * COMMUNICATION 366 |
| * FINANCE 360 |
| * SPORTS 351 |
| * PHOTOGRAPHY 322 |
| * HEALTH\_AND\_FITNESS 306 |
| * SOCIAL 280 |
| * NEWS\_AND\_MAGAZINES 264 |
| * TRAVEL\_AND\_LOCAL 237 |
| * BOOKS\_AND\_REFERENCE 229 |
| * SHOPPING 224 |
| * DATING 196 |
| * VIDEO\_PLAYERS 175 |
| * MAPS\_AND\_NAVIGATION 137 |
| * EDUCATION 130 |
| * FOOD\_AND\_DRINK 124 |
| * ENTERTAINMENT 111 |
| * AUTO\_AND\_VEHICLES 85 |
| * LIBRARIES\_AND\_DEMO 84 |
| * WEATHER 82 |
| * HOUSE\_AND\_HOME 80 |
| * EVENTS 64 |
| * ART\_AND\_DESIGN 64 |
| * PARENTING 60 |
| * COMICS 60 |
| * BEAUTY 53 |

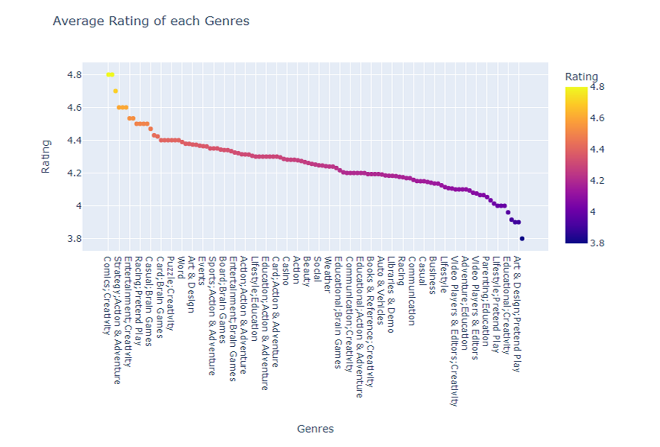
* From the above data we can conclude that there are 1939 applications which fall under the FAMILY category.
* Also top 5 categories of Application in the play store are FAMILY, GAME, TOOLS, BUSINESS, MEDICAL.
* Also, there are only 53 applications which fall under the BEAUTY category.
* We can also conclude that there is significant difference between the top two categories FAMILY 1939 Apps and GAME 1121 Apps. This shows that there is more application
* developers who develop Apps under FAMILY Category.



* FAMILY which category has maximum apps in play store (1931). It has overall average rating 4.19 and ranked 15th according to overall average rating. But interestingly, 2nd most maximum apps category 'GAME' ranked 5th with 4.27 average rating.
* Interestingly any of the top 5 maximum average rating apps does not match with top 5 most produced apps category.
* Top 5 most produced app category = FAMILY, GAME, TOOLS, BUSINESS, MEDICAL
* Top 5 maximum average rating app = EDUCATION, ART\_AND\_DESIGN, EVENTS, BOOKS\_AND\_REFERENCE, PERSONALIZATION

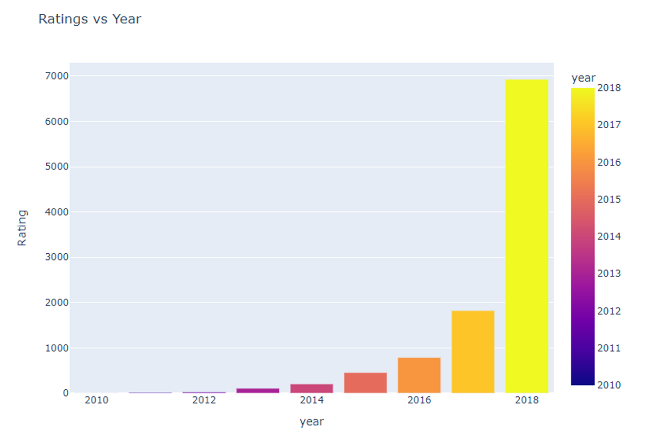
**All categories have average rating above 4.0 and 'Dating' has lowest average rating 4.01**

**6. Top 5 Genres which have best average Rating**



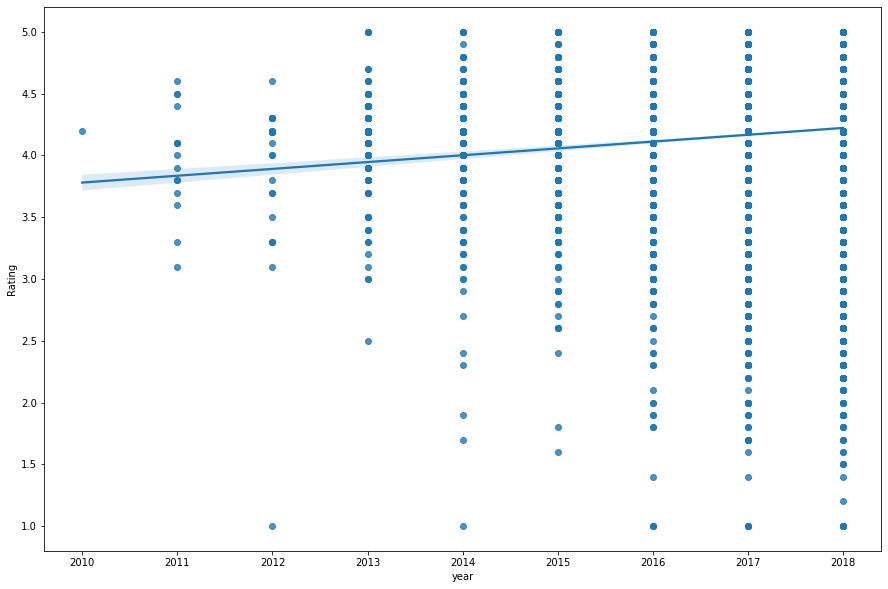
* From the scatter plot from plotly we can conclude that both the Genre-’Board Pretend Play' and 'Comics; Creativity' is having the highest Average Rating of 4.8
* There is no direct relation between quantity of apps under a genre and rating.
* Here Top 5 genres according to maximum average reviews are: - Comics; Creativity Board; Pretend Play, Health & Fitness, Education, Strategy, Action & Adventure, Puzzle; Education
* Genre- ‘Parenting; Brain and Games’ has got lowest Average Rating.

**7. Last Update Impacts on Trend of Rating**



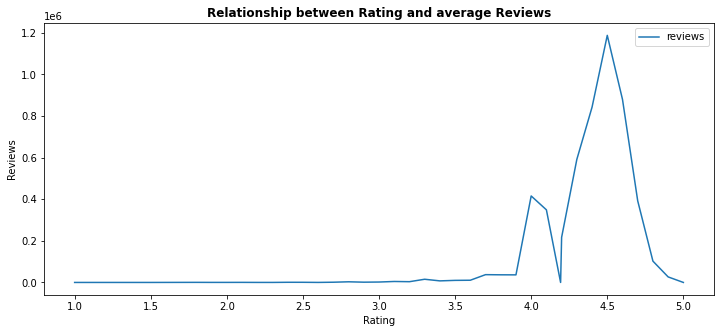
* From the above graph from plotly we can conclude that there are more number ratings given to the applications which were updated recently in 2018, no. of rating is 6929 than those applications which were updated in 2017 no.of rating 1823. This shows with the latest update user reviews response increases for both less or more ratings.

*How Last Update impacts on trend?*

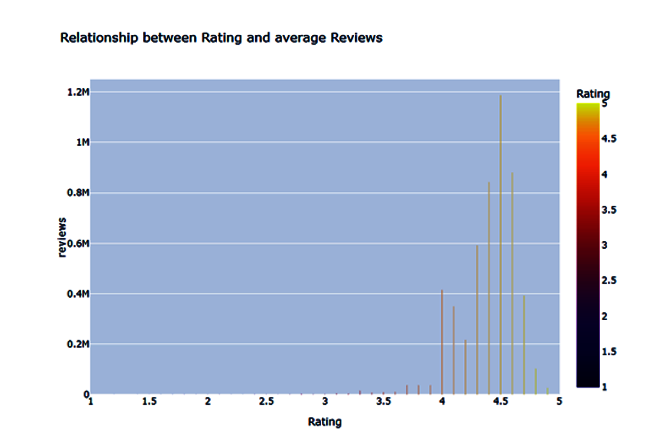


* From these above graphs we can conclude that
* rating is increasing in a proportionate manner with the last updated time.
* So, from this we can be sure that with the latest updating the reviewers are giving better rating.

**8.Relationship between Rating and Average Reviews**

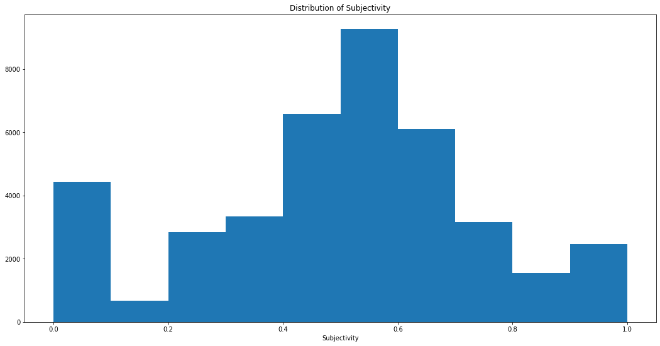


*Let's clarify it more by using plotly*



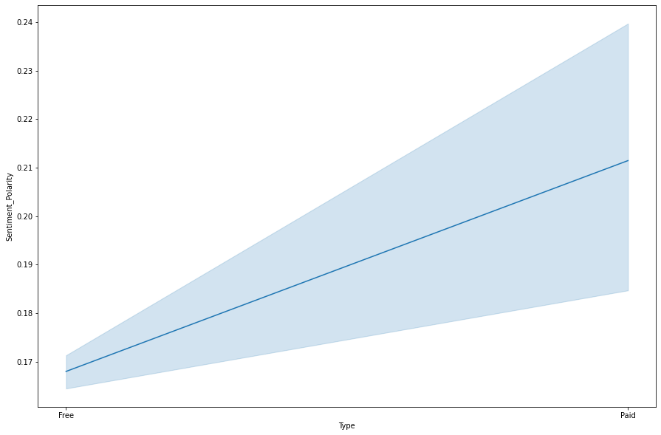
* from the above graph we can conclude that as the rating increases the average reviews for each rating also increases.
* Also we can observe some deviation after 4.5 rating as we can see 4.5 rating has maximum no.of reviews of 1.1Mand reviews increases in quite proportionate manner with the increase of rating.
* but after 4.5 rating reviews eventually decreases and we can see only around 26586 reviews.
* It means maximum reviews does not imply maximum rating, so now we have to know review sentiment for understanding it better

*Let's find the sentiment subjectivity*



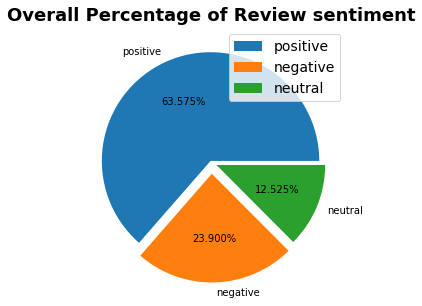
* It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications, according to their experience.

*Let’s see how sentiment polarity varies with Free and paid Apps?*



* people express more sentiment towards paid app rather than free app

*Overall percentage of review sentiment*

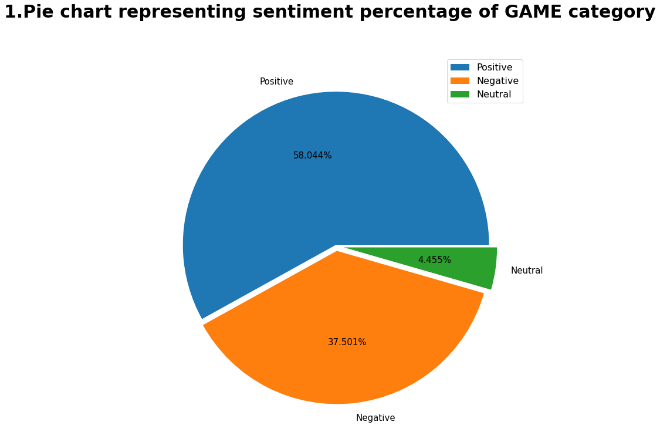


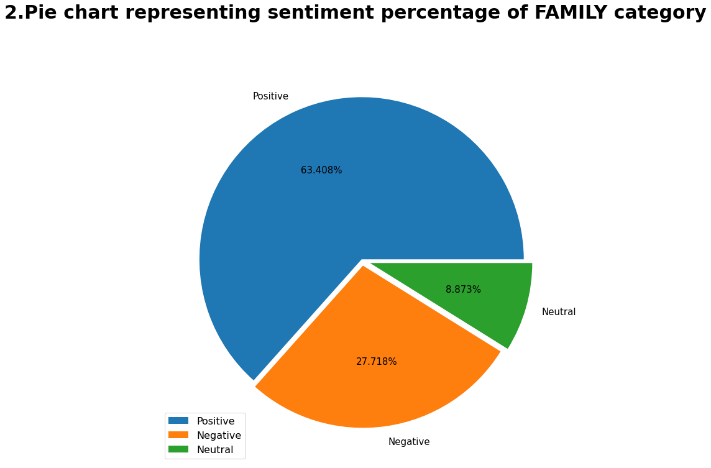
**9.Top 5 Categories Review sentiments**

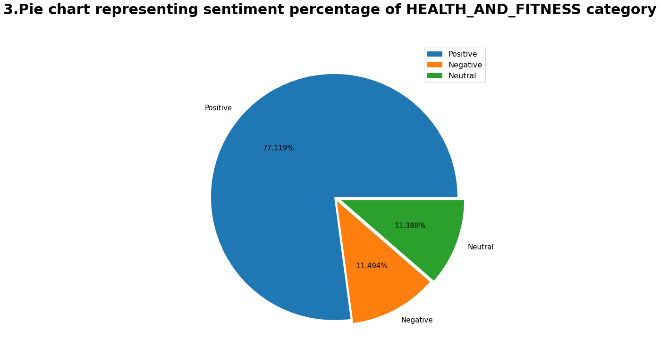
we got 'GAME', 'FAMILY’, 'HEALTH\_AND\_FITNESS', 'TRAVEL\_AND\_LOCAL', 'TOOLS'

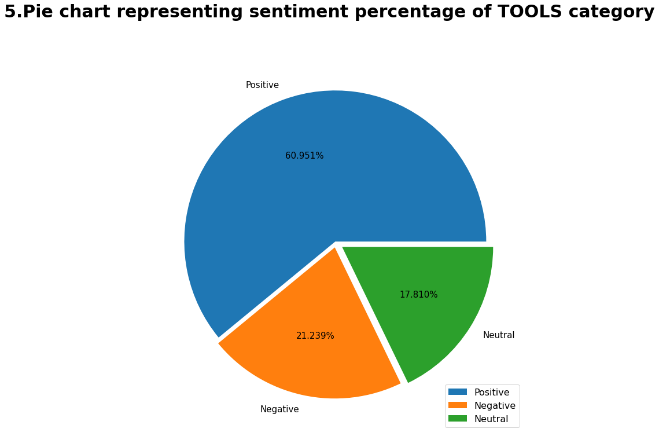
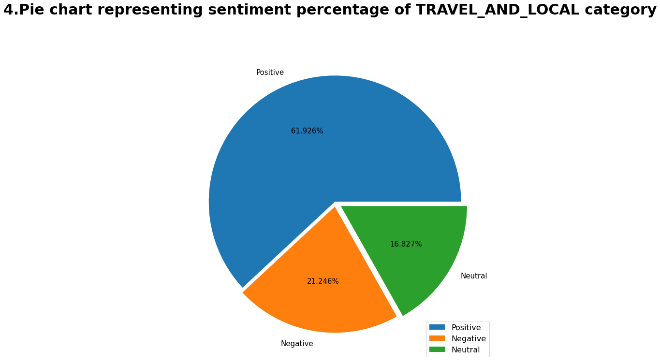
as most popular category which has most translated reviews

*Let's analyze Sentiments percentage of Each Categories*



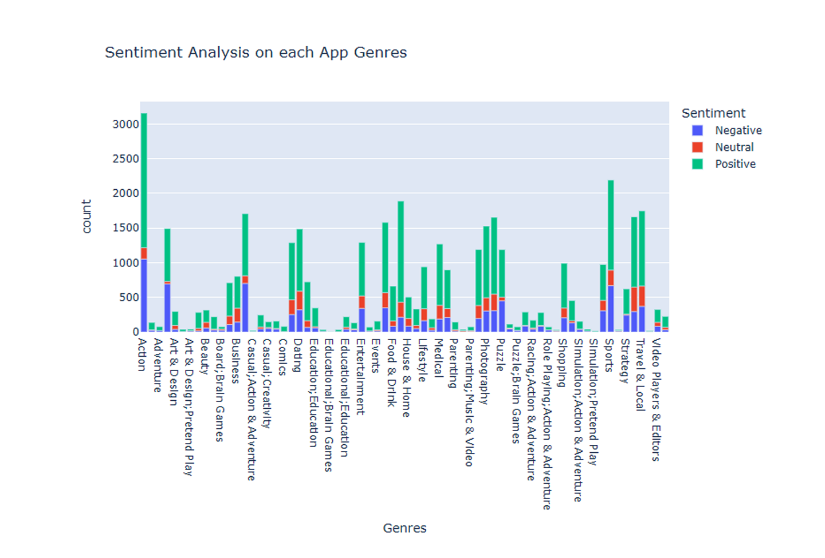
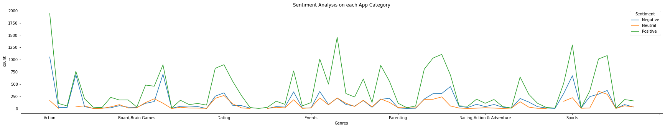






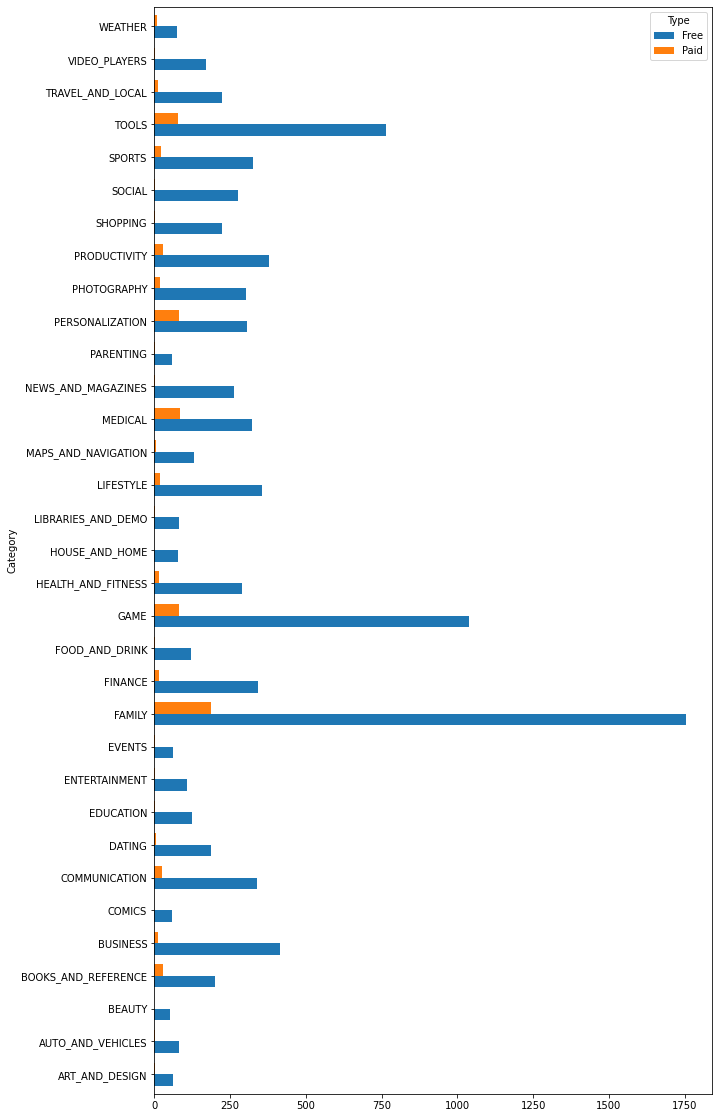
* From the above figure we conclude that 'HEALTH\_AND\_FITNESS' Category has got highest positive percentage of 77.119% and negative sentiment percentage of 11.494% and neutral percentage of 11.388%
* Also, we found that top Category GAME has less positive sentiment percentage of 58.04% than its competitor.
* so, most review does not suggest most positive sentiments
* Most negative sentiments from the top translated app category have been received by GAME CATEGORY THIS shows that even if GAME app has the highest translated reviews but in positive sentiment percentage it is lower than its counterparts.

**10.Let's do sentiment analysis in each app GENRES**



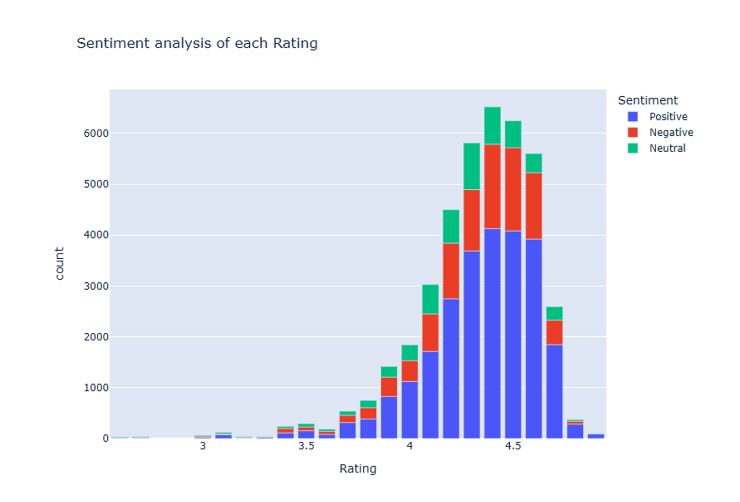
* From Above bar plot we can conclude that Action Category has got most positive sentiment in terms of count but if we take percentage of positive vs negative sentiment, we will find that HEALTH\_AND\_FITNESS has got the highest positive sentiment percentage.
* Again, if we count on the basis of the count of positive sentiment, we can see that Category comic has got least number of positive sentiment and no negative sentiment as such.

**11. Let’s check if the Category is free or paid**



* From the above graph we can conclude that,
* FAMILY category has the most free and paid app on the google play store.
* We can see SOCIAL app is always free, like ENTERTAINMENT, EVENT, EDUCATION, COMIC, and more.
* The second place in free app is GAME category app followed by the TOOLS category app.
* Google Play Store has more free apps than paid apps.

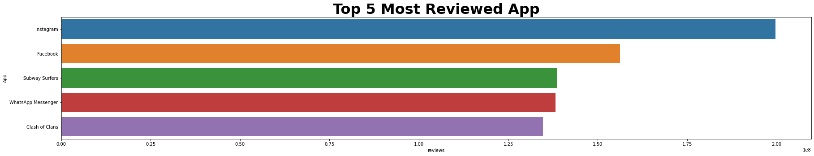
**12.** **Let’s look the sentiment column according to Rating**

* Top Apps are rated 4.4, 4.5, 4.6, 4.3, 4.2****
* From the above graph we can conclude that most of the reviews are Positive.
* Neutral reviews are in the middle range.
* Less people have commented negatively about google play store apps.
* 4.4 rating apps have most positive reviews, from 2.5 to 3.1 there are almost equal numbers of positive, neutral, negative reviews.

**13.** **Most Reviewed Apps**

From the data we see that ROBLOX ,8 Ball Pool and Zombie Catchers apps are occurring frequently.

*Let's see most reviewed top 5 apps.*

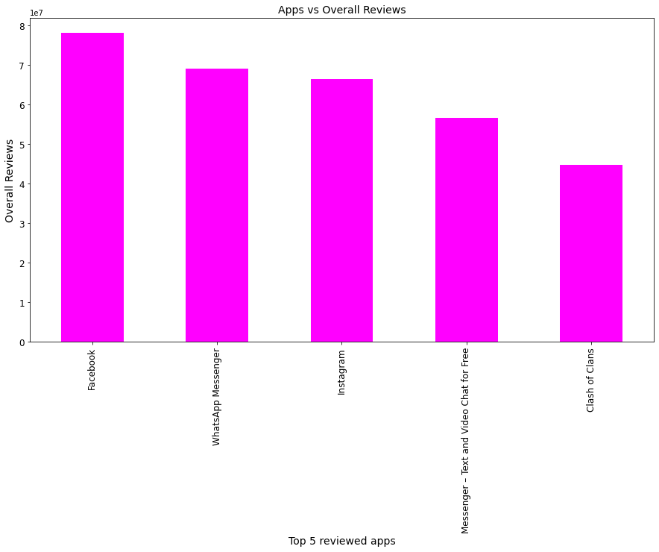
****

* Here we see that, Instagram, Facebook, Subway surfers, WhatsApp Messenger, Clash of Clans have highest reviews.
* All these apps have overall reviews between 19cr and 13cr.

*Now let's see what are the top 5 apps which have average highest reviews*

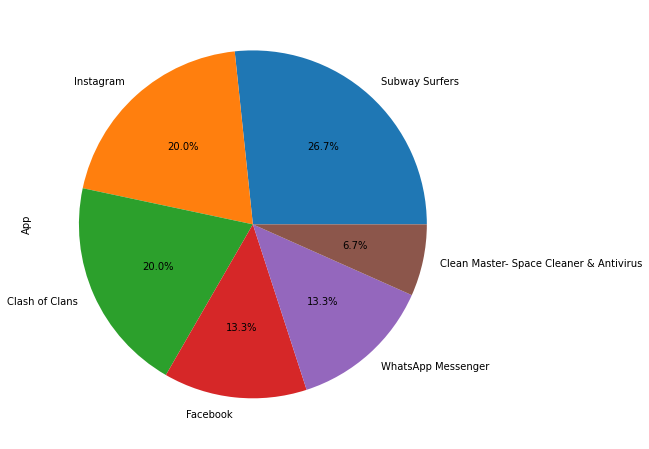
* Here we see that, WhatsApp Messenger, Instagram, Messenger – Text and Video Chat for Free & Clash of Clans have highest reviews.
* All these apps have overall reviews between 4.4e^7 and 7.9e^7.

*Now let’s analyses it by plotting a graph between app and reviews*



* From this above graph we concluded that Facebook, WhatsApp Messenger and Instagram app have more reviews.
* These apps are falls under the category of Communication and Game.
* There is different outcome of ranking between top 5 total reviewed apps and top 5 average reviewed apps. But 'Facebook, Instagram, WhatsApp messenger, Clash of clans' these apps are common in both outcomes. so, we can conclude that these 4 apps are most trending and popular among masses

**14.** **Apps which have rating more than 4 and have reviews greater than 10,000**

****

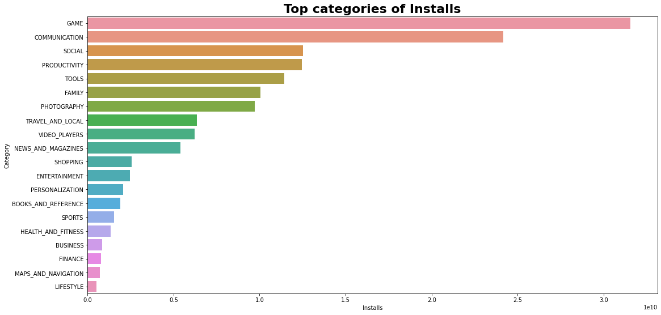
* From the above pie chart, we found that Subway Surfers app have highest rating along with highest reviews.
* From most reviewed apps, we have seen that Facebook has more reviews but when both rating and reviews comes into picture Instagram app capture it.

**15.** **Variation of review with month in the year 2018**

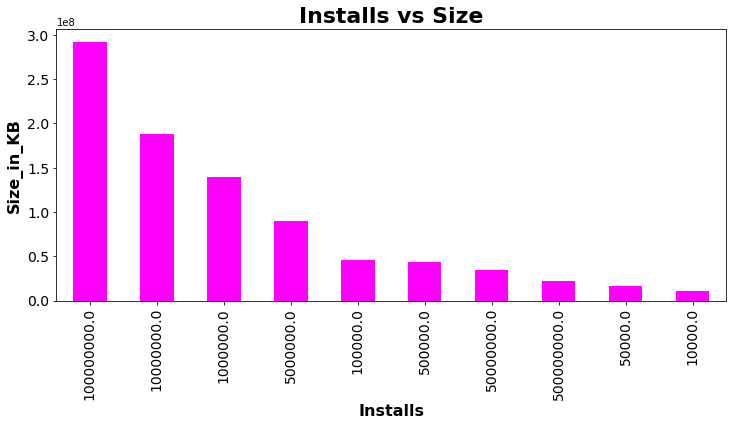
****

* From this bar plot we come to the point that apps have highest review in the month of January and lowest review in February and March.

**16.** **Factors affecting installs**

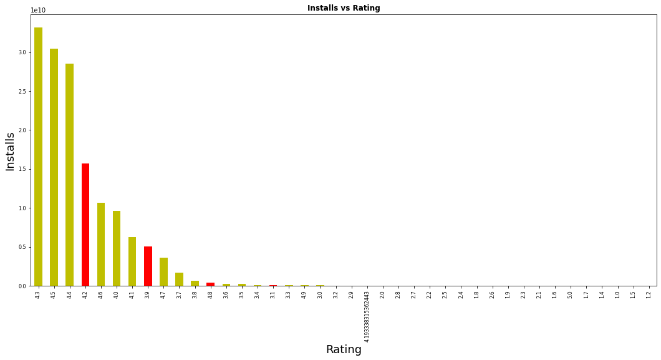
****

* Top Categories are Games, Communication, Social, Productivity, Tools etc.
* Top 5 categories have over 1cr installations in play store
* *b. Installs vs size (let’s check how size of the app affects installs)*
* Most heavy app has also got 10 cr installations
* Amongst Top 10 heavy apps 4 apps are above 1 cr installations and 4 apps are above 1 lakhs installation

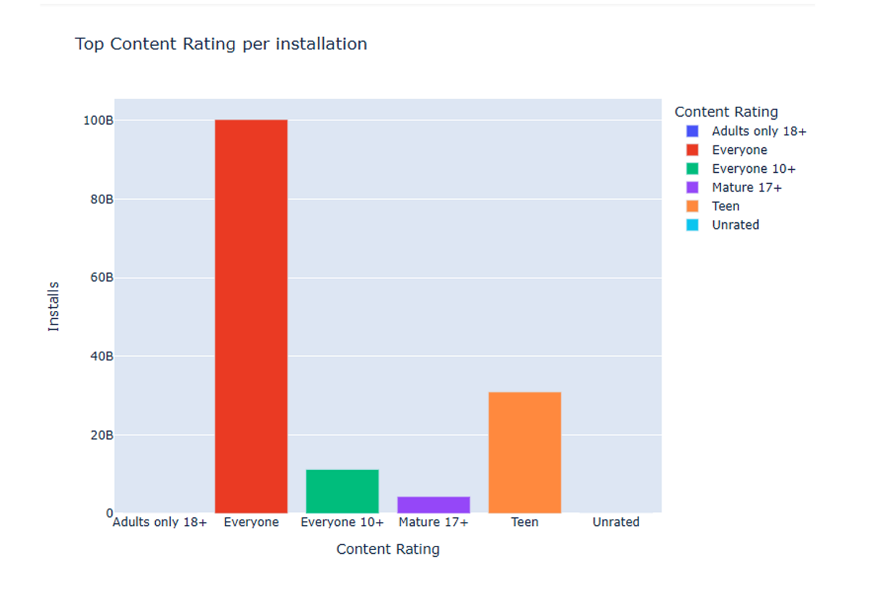


* This graph concludes that Heavy app does not less attractive for customer for installation.
* AS among top 10 heavy apps 4 apps are over 1 cr installations and 4 are above 1 lakh installations

*c. installs vs Rating (let’s see how rating of an app affects installs)*

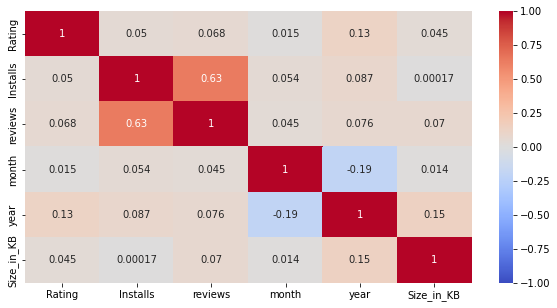


* Top Ratings apps which have most installs 4.3, 4.5, 4.4, 4.2, 4.6, 4.0, 4.1, 3.9, 4.7, 3.7
* Top 3 Ratings are 4.3,4.5,4.4 (yellow bars)
* Interestingly below 3.7 and above 4.7rated every apps enjoys very less installs compared to top 10 ratings
* It suggests good rating does not imply good installations. Certain Rating apps are only enjoying most installation

**17.Top Content Rating Per Installations**

* From above graph by plotly we can conclude that App Having 'Content Rating' - Everyone is having maximum number installation of \*\*100.228B\*\*.
* Also, we can conclude that there are \*\*4.29B\*\* installation for application having Content Rating of Everyone 10+.
* We can also conclude that there are no such noticeable installation for Application having Content Rating of Adults only 18+ and Unrated.

**18.Co-relation between installs, ratings and size**

****

* Here we find out the correlation between rating, install, month, year, size and reviews
* Correlation coefficients of install with rating, reviews, month, year, and size are 0.05 ,0.63,0.054,0.087,0.00017 respectively
* So, the install is mostly correlated to reviews
* Install is least correlated to size

**19.Conclusion:**

After doing this project I got many interesting conclusions. First, I have done the removal of null value from rows and columns and the same goes with the removal of duplicates from the datasets. Then did the formatting for each of the required columns in each dataset.

App with the category Family and the genre tools are in large numbers. But interestingly Category EDUCATION and Genres 'Board;Pretend Play' and 'Comics;Creativity' both get the best average rating in playstore.

App Rating is directly proportional with the recent update. From this I can be sure that with the latest update the reviewers are giving better rating.

I also found that Ratings are directly proportional with Average Reviews till 4.5. But after that I observe some deviation and eventually reviews decreases after 4.5 ratings. 4.5 Rating apps have maximum 1.1 million reviews.

I found that most of the opinion on sentiment subjectivity lies high in the range 0.4 to 0.7. It means people give review after some experience.

I also found people's sentiment are more towards paid app less towards free app

GAME category has maximum no of translated reviews but, most positive percentage acquired by HEALTH\_AND\_FITNESS category and GAME has most negative reviews among top 5 categories. So, it concludes that most reviews do not suggests most positive sentiment percentage.

Action Genres got most strong positive sentiment among all genres. It suggests that people prefer Action genres app far more than any other genres.

google playstore has more free apps than paid. and Family category has most free and paid apps in playstore. MEDICAL category has highest number of paid apps considering quantity.

Social,entertainment,event,education,comic\* etc these popular category apps are always free.

I found that in terms of both Top Total and Average Apps Reviews outcomes Facebook,Instagram,whatsapp messanger,Clash of Clans these apps are common. so I can conclude that these 4 apps are most popular and trending among masses.

Subway Surfers app has highest rating along with highest reviews followed by Instagram,clashofclans,Facebook,whatsapp messanger,clean master.

Highest reviewed apps are those which are last updated in january,2018 and got most review in February and march update. It concludes that Most of the companies are majorly updating their apps in January month, so in that months they get most reviews.

I also found that 4.4 rated apps are most reviewed apps in google playstore and from 2.5 to 3.1 there are almost equal numbers of positive,negative,neutral reviews. It states that neither top rating nor less rating makes an app good or bad.

GAME,COMMUNICATION,SOCIAL these category apps got most installation(above 1 crore).

Most heavy app got 10cr installations, among top 10 heavy apps 4 apps got above 1 cr installations so, it concludes that customers still preferred heavy apps.

Most Installed, Rated, reviewed positive sentiment apps are always in range between 4 and 4.5 rating.

Most of the apps which are used by the users have a content rating of ‘Everyone’.

From install,rating,reviews co-relation graph I also found install and reviews are positively co-related means when Installation is more reviews are more.

**Important Points to be noted:**

All active apps on play store have an an Average Rating of 4.32.

Also, we can see that after merging both of the dataset the maximum Average Rating is 4.9.

Also, the average sentiment Polarity is 0.16 and average sentiment subjectivity is 0.49

Also, we have noticed that the average size of the Application available on play store is 21933.38 KB

lastly, all of the calculations and graphs in this project have accuracy in the range of 75% to 80%.

**20.References**

GeekforGeeks

Kaggle

Analytics Vidya