**Team Capstone Project-1 Submission**

**PLAY STORE APP REVIEW ANALYSIS**

Team curio monks

**Data science trainees,**

**AlmaBetter, Bangalore**

**Presented by:- Email:-**

**1)Adil imam adil.imam12@gmail.com**

**2)Sushil kumar singh sushilsinghrajput.333@gmail.com**

**3)Asadullah pathan asadullaapathan@gmail.com**

**4)Md Sazil mdsazilsharif@gmail.com**

**5)Madhulika Kumari mpradhan1990@yahoo.com**

**GitHub Link:-**

<https://github.com/adilimam12/adilimam12>

<https://github.com/sushilksing808?tab=projects>

<https://github.com/madhulika-1990>

**Abstract:**

Playstore is the app which provides the platform to android apps from where users can download or update the apps as per their requirement. Android is the dominant mobile operating system today, more than 85% of all mobile devices running Google’s OS. The Google Play Store is the largest and most popular Android app store. We focus on analysing Google Play Store, the largest Android app

store that provides a wide collection of data on features (ratings, reviews, type ,install and number of downloads ,) and descriptions related to application functionality. The overall objective of this analysis effort is to provide in-depth insight about play store review analysis in general.

**Introduction**

The Google Play Store started life as the “Android Market” in 2008. It launched alongside the very first Android devices, and its purpose was to distribute apps and games.The Android Market was extremely basic at the beginning. It didn’t support paid apps and games until 2009. However, as the Android platform grew, so did the Android Market. By 2012, it featured over 450,000 Android apps and games.

We live in an era where data is produced and circulated in an enormous amount. Those data can be collected and allow us to infer meaningful results and make well-informed decisions. However, as the number of data increases, it becomes difficult to analyse the data, here we need to visualise the data to help us in conducting data analysis. By using visualisation tools, we can deliver a message to our audience and inform them about our findings

By this time, Google’s ecosystem had expanded greatly compared to the humble beginnings of the Android Market. In fact, the Android Market was just one of the company’s online markets. At the time, this was the only place that Google had to sell goods. As the company’s hardware efforts grew, it was time for a new store. Posting reviews online has become an increasingly popular way for people to express opinions and sentiments towards the products bought or services received. Analysing the large value of online reviews would produce useful actionable knowledge that could be of economic values to vendors and other interested parties.

Many apps are being developed as apps are easy to create and it's lucrative. But it's important for developers to know which apps are loved by customers and are trending in the market so that they develop only those apps and also there is a high competition between app providers producing similar applications. Analysing customer needs is one of the bizarre tasks in the business world today. Hence proposing to analyse data to the developer that what customer is likely to download, which category got the maximum downloads this all plays a crucial role in app development. Generally, customers download apps depending on number of downloads, positive reviews, negative reviews, ratings and comments. So, in this project we are going to help the users by categorising positive, negative and neutral reviews and comments of the particular. We are going to help developers by analysing the desire of the customer through the reviews provided in the feedback section and apps trend in the market to help the organisation & developers. Also provide an idea about app that managed to get maximum and minimum number of downloads and predicting the category of apps that is most likely to be downloaded in the coming years. The dataset of google Play Store for analysing is collected from the dataset.

The purpose of our project is to gather and analyse detailed information on apps in the Google Play Store in order to provide insights on app features and the current state of the Android app market.The Objective of the project to Explore and analyse the data to discover key factors responsible for app engagement and success.

**Problem Definition:**

The Play Store apps data has enormous potential to drive app-making businesses to success .Android is expanding as an operating system and Mobile app industry is increasing in significantly and thus giving rise to more competitions to the ones that are creating applications .Due to the competition in the market and also expansion in order to help our developer understand what kinds of apps are likely to attract more users and what is the motivating factor for the people to download an app we analyse and research relevant data. For the app development industry where they can analyse the downloads and demand off app download in the industry.

1. What are the top categories on the Play Store?
2. Are the majority of the apps Paid or Free?
3. How important is the rating of the application?
4. Which categories from the audience should the app be based on?
5. Which category has the most no. of installations?
6. How does the last update have an effect on the rating?
7. How does the count of apps vary by Genres?
8. How are ratings affected when the app is a paid one?

**3.EDA on given Data set**

There are two dataset:

1. **Play Store Data**(App, Category, Rating,Review,Size,Install,Type,current rating ,genres , Last update,Current Var ,Android Var)
2. **User Review Data**(App, Sentiment ,Sentiment Polarity, Sentiment Subjectivity)

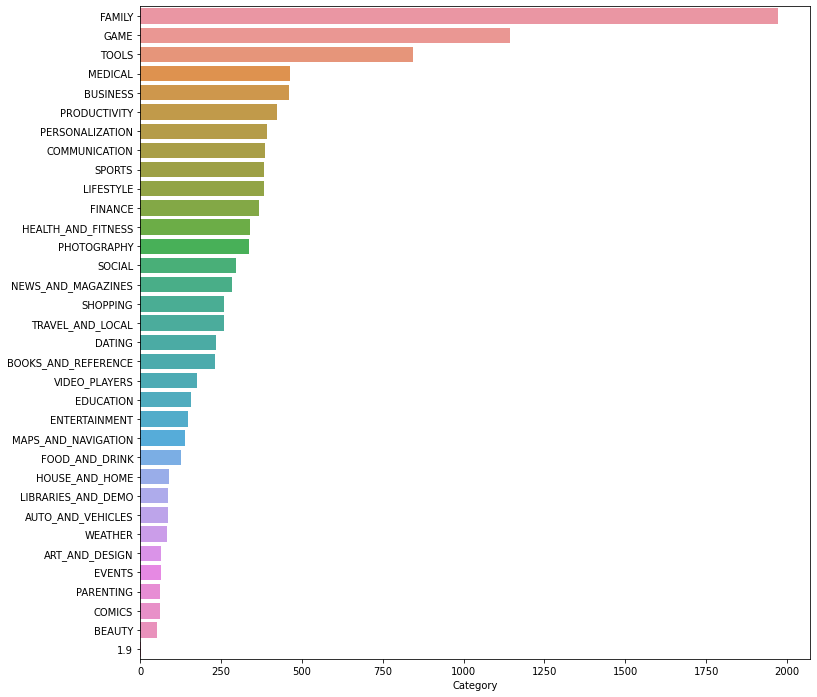
By diagnosing the data frame, we know that:

* There are 13 columns of properties with 10841 rows of data.
* Column 'Reviews', 'Size', 'Instals' and 'Price' are in the type of 'object'
* Values of column 'Size' are strings representing size in 'M' as Megabytes, 'k' as kilobytes and also 'Varies with devices'.
* Values of column 'Initials' are   
  strings representing install amount with symbols such as ',' and '+'
* Values of column 'Price' are strings representing price with symbol '$'.

**Observation1:**

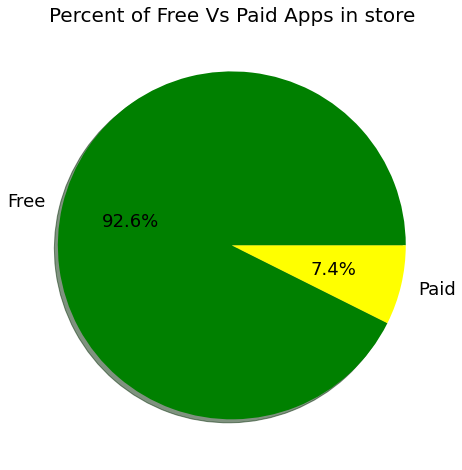
We plotted a graph of top categories on playstore.

So we got 34 categories on this dataset. Lets see which one is the most used categories



The Playstore market is ruled over by the family and game category followed by tools, business, medical, personalisation and so on.

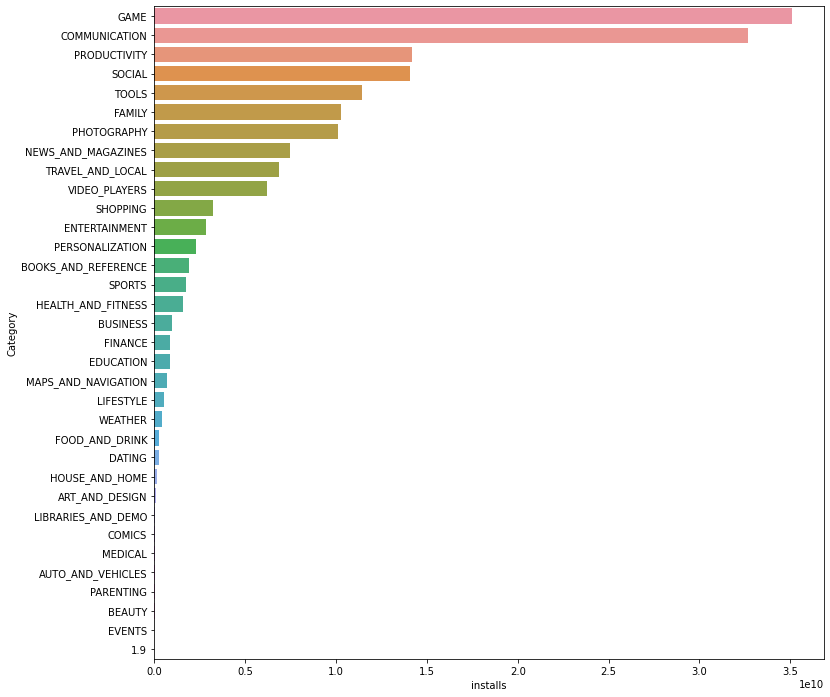
**Observation 2:**

****

The above Pie-Chart represents the percentage of Free apps are 92.6% and paid apps are 7.4% in the dataset.

**Observation 3:**

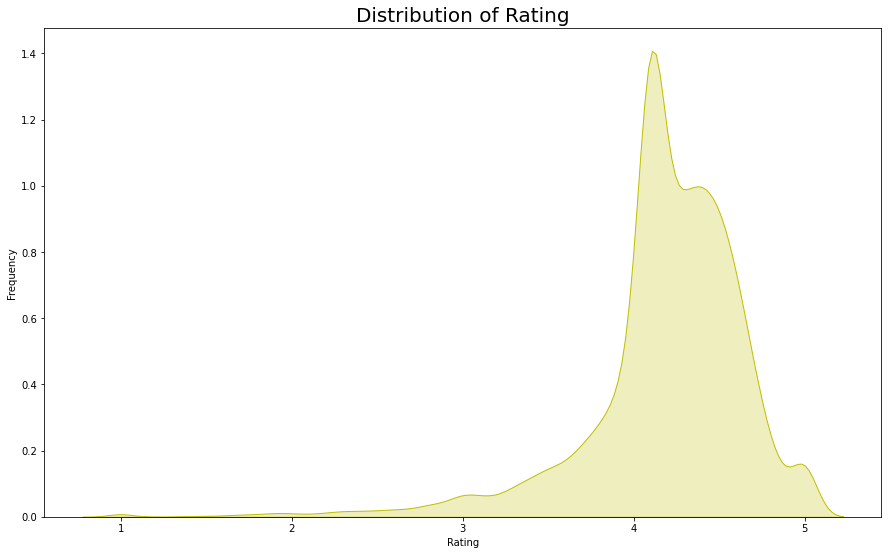
Category with most no of instals



As seen in the graph, the top 5 genres are- tools, entertainment, education, business, medical.

**Observation 4:**

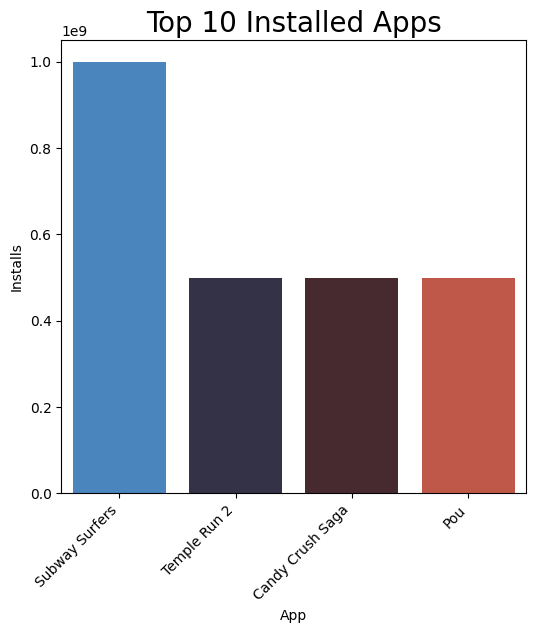
We plotted a graph to check the most frequent rating of the apps.



Most apps have a rating between 4 to 5(mostly 4.3 approx.)

**Observation 5:**

Top 10 Installed games in playstore are

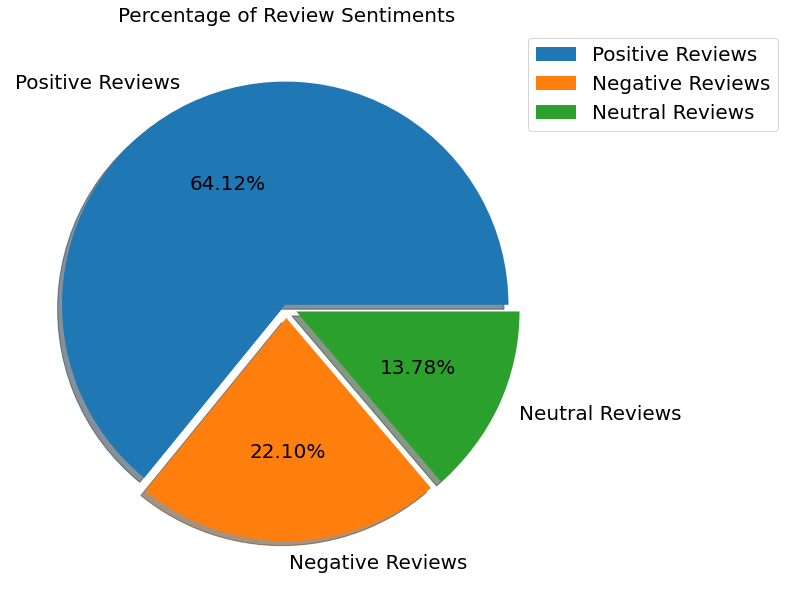


.

From the above graph we can see that in the Game category Subway Surfers,Candy Crush Saga, Temple Run 2 has the highest instals. In the same way we by passing different category names to the function, we can get the top 10 installed apps

**Observation 6:**

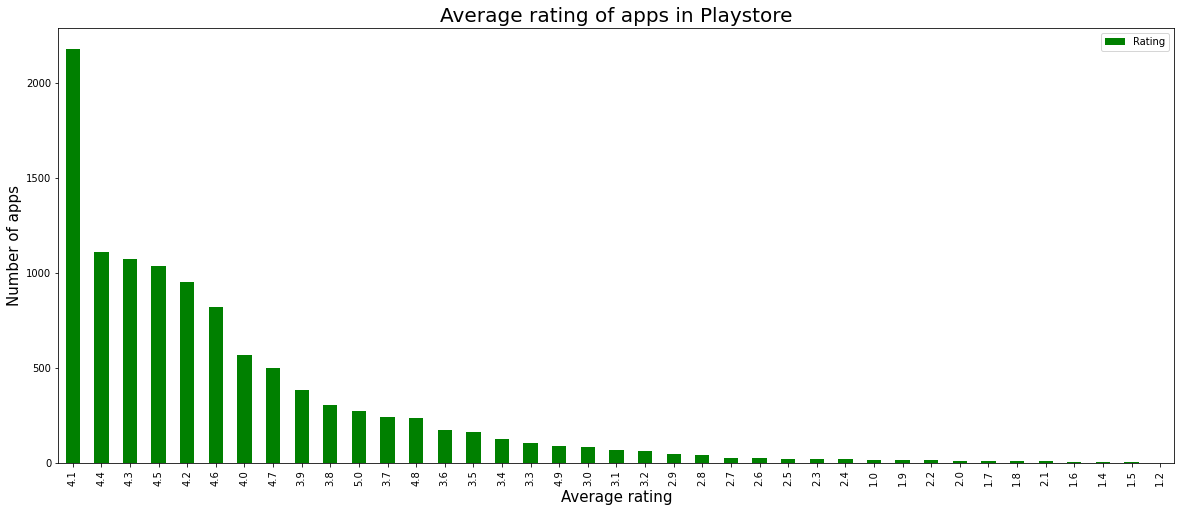
Percentage of Review Sentiments

****

Positive reviews are 64.30% Negative reviews are 22.80% Neutral reviews are 12.90%

**Observation7:**

average rating of apps in playstore

****

We can represent the ratings in a better way if we group the ratings between certain intervals. Here, we can group the rating as follows:

4-5: Top rated

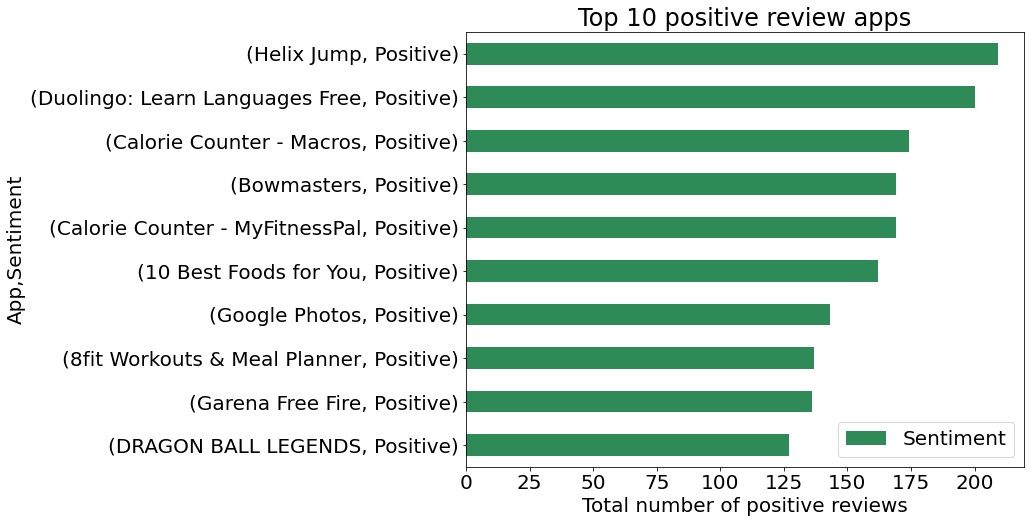
3-4: Above average

2-3: Average

1-2: Below average

**Observation8:-**

Top 10 Apps with highest number of positive reviews

****

In the above graph we can see the top 10 Apps with most no of positive reviews in play store

**Conclusion**

Our major challenge was data cleaning. 13.60% of reviews were NaN values, and even after merging both the dataframes, we could not infer much in order to fill them. Thus we had to drop them. The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis, if we had at least 70% - 80% of the data available in the merged dataframes. User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column. There is so much more which can be explored. Like we have the current version, an android version available which can be explored in detail and we can come out with more analysis where we can tell how these things affect and needs to be kept in mind while developing apps for the users. We can explore the correlation between the size of the app and the version of Android on the number of instals. Machine learning can help us to deploy more insights by developing models which can help us interpret even better. We have left this as future work as this is something where we can work on.

**ACKNOWLEDGEMENT**

This project is presented by-Adil imam , Sushil kumar singh, Md sazil sharif, Madhulika kumari and Asadullah pathan. Our sincere efforts have made us accomplish the task of completing this project. We are extremely grateful to our instructors and mentors who have helped us to grow in this field. We would like to express our sincere gratitude to the celebrated authors whose phenomenal work has been consulted and referred to in our project work. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need. This project helped us improve our skills and enhanced our knowledge.

**REFERENCES**

[**https://blog.jovian.ai/**](https://blog.jovian.ai/)

[**https://www.kaggle.com/**](https://www.kaggle.com/)

[**Analytics Vidhya - Learn Machine learning, artificial intelligence ...**](https://www.analyticsvidhya.com/)