Play store apps review analysis

Naved mansuri
Data science trainees,
AlmaBetter

Abstract

It has been observed that the significant growth of the mobile application market has a great impact on digital technology.

For the EDA part, we analyzed our dataset and created several Bar Charts to visualize the relationship between each attribute. Then we created a heat map of installs with other numeric features from there we could conclude that numeric features such as the number of ratings as well as subjectivity are all uncorrelated with installs. Additionally, we did an analysis on the features App Name, Content Rating, and Type and reached some valid conclusion. Furthermore, we did review analysis on the 170000 reviews obtained from scraping, using those reviews we created a dictionary of words that appeared in the reviews and recorded their occurrences and illustrated that using a Bar Chart.

From those Bar Charts, we observed the variance of the first 25 most frequently used word associated with each rating ranging from 1 through 5. To further investigate that, we created a dictionary of positive and negative words manually, using information from the dataset

Introduction

Play store is estimated to be around 2.6 million applications in March 2018 subsequent to outperforming 1 million apps in July 2013. The huge number of apps in the play store and the numbers of the app released every day make it quite competitive for the app developers, the companies who are building an app to come up with a unique idea that will definitely be bought by the end users. Because at the end of the day if the app does not perform well in the android market, then all the hard work behind building the app will go in vain. As the mobile industry is growing rapidly it is increasing the level of competition however, increased competition also leads to increased chances of failure. So, the developers need to do enough research as an enormous amount of time, effort and the money are invested into the process, so business cannot afford an app failure

Data Description

1. Play store data

Attribute	**Description**
App	Name of the app
Category	Category of available apps

Constant rating	Constant of audience
Installs	Number of installs
Rating	Rating of apps
Last update	Date of update
Type	Paid/free
Size	Size of app

2. User review data

Attribute	**Description**
App	Name of the app
Transitioned review	Comment text entered by user
Sentiment	Positive, negative and neutral

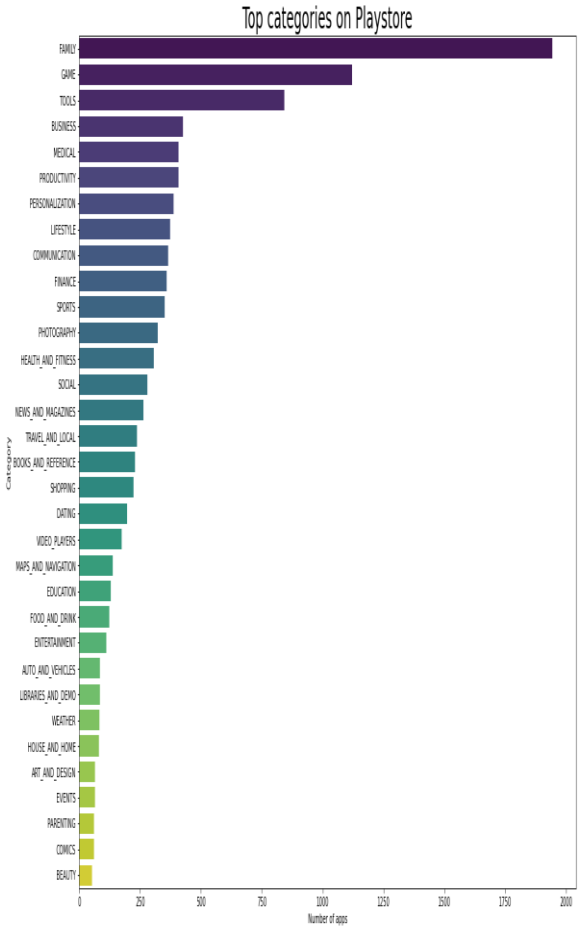
Data Processing

The raw data extracted needed to be pre-processed to turn it into some valuable information. To be able to perform EDA and run algorithms on our dataset, installs, total number of ratings, rating distribution were converted to integers. Sizes of app having kilobytes were converted to megabytes to have consistent units. Apps with varying sizes (those that vary with device) were set to the average size of the rest of the apps. Later the categorical attributes like categories/genre, content rating and type were label encoded

Exploratory Data Analysis

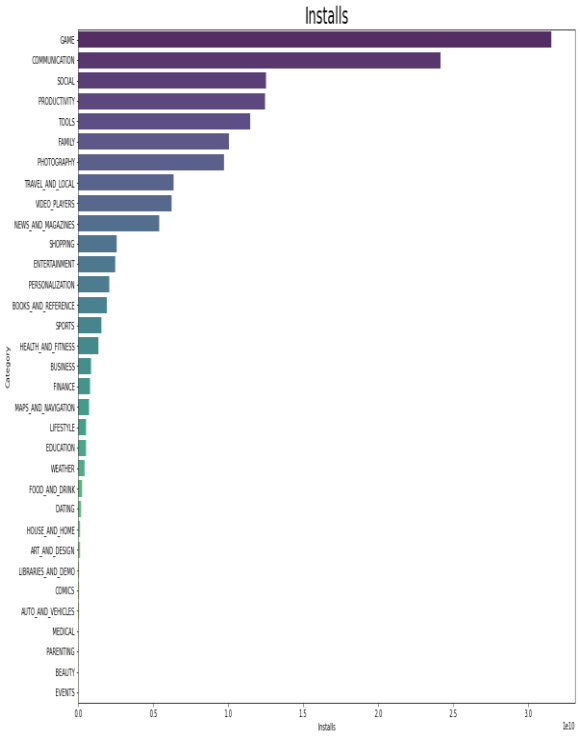
Category App

The Category column in our dataset has 47 different types of categories. Shows the bar chart of the number of categories. Also, from the figure it is apparent that the Tools category is the most dominant in the Play Store. It came to our attention that games were not classified under a single label, instead they were very specific. The following contains all the different types of games: Action, Casual, Arcade, Puzzle

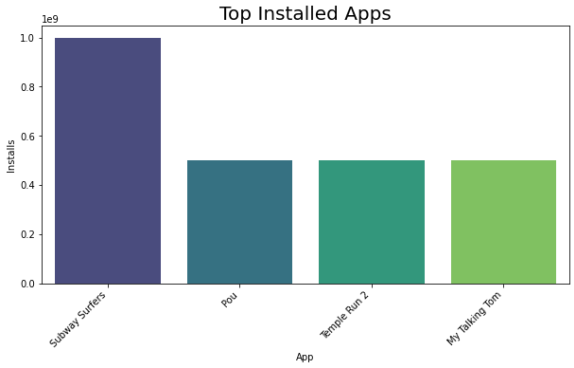


Category vs Installs

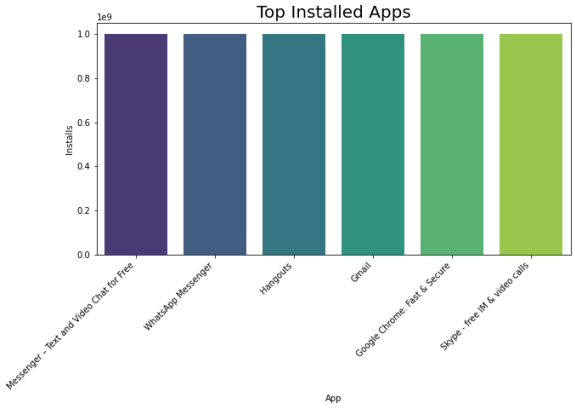
In contrast to our previous results, comparing the top ten most dominant categories by installs tells a slightly different story. Figure shows that while Games still rule the Play Fig Bar chart of top 10 categories by installs Store, Tools, previously the 2nd most dominant by app numbers, has been dethroned by Communication which was the 8th most dominant. Similarly, the Education category which previously occupied the 3rd spot has now fallen by six places to the 9th spot. Hence, looking at categories by apps might be misleading for a developer. A developer wanting to attract a large user base should pick a category based on the number of installs and not by the number of apps in the Play Store. Figure 4.3 compares all the categories against the number of installs.



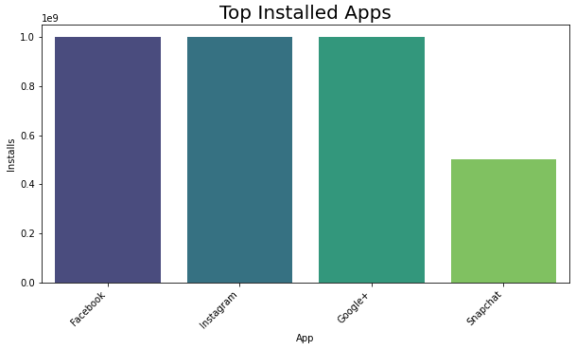
Categorical analysis in game category



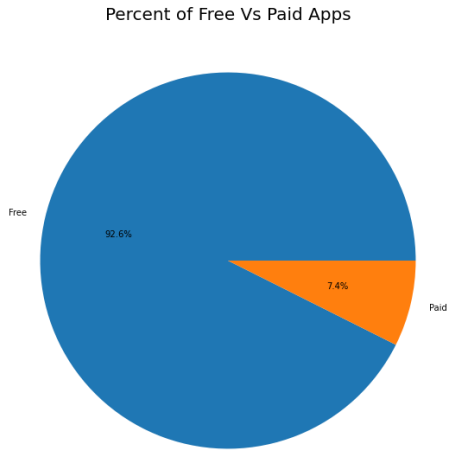
Categorical analysis in comm. category



Categorical analysis in social category



Paid apps and free apps



Approximately 92% of the apps in the play store are free to install. Approximately 82% of the app in the play store has to age restrictions to install and use the app. The rest of them apps have certain age restrictions on it. Around 11% of the apps are rated as “Teen”, which means that the user must be at least a teenager (13 years old) to install and use the respective app. Around 4% of the apps are rated as “Mature 17+”, and around 3% of the apps as “Everyone 10+”.

Average app rating

The median user rating for all the apps in the play store is 4.3. To simplify the analysis, the average user ratings were divided into 4 categories as follows:

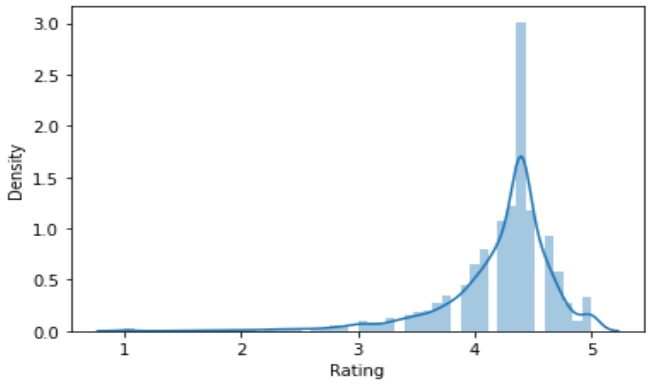
Rating: 4-5 = Top rated

Rating: 3-4 = above average

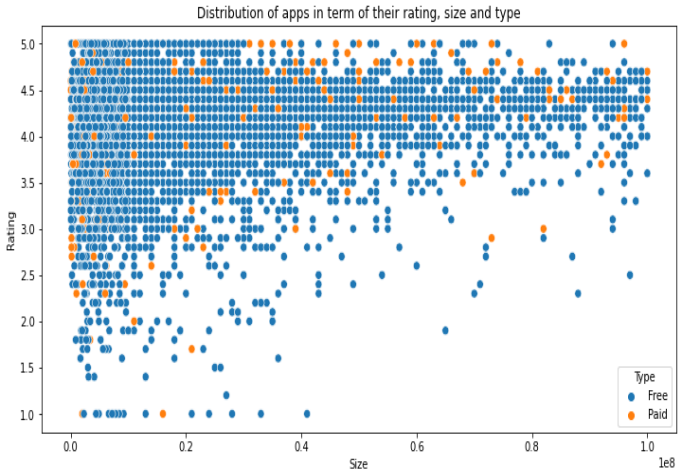
Rating: 2-3 = Average

Rating: 1-2 = below average

Approximately 80% of the apps in the play store are top rated. From this we can say that the users are satisfied with the services provided by 80% of the apps.



Distribution of ratting, size and type



The size of an app in our database varies from 100 MB to 0.0083 MB. We can analyses the size of the apps if we can group

them into certain intervals. The visualization below gives the number of apps present in each size group. The higher this number, the higher is the competition. The majority of the apps in the play store are in the size range of 1-20 MB, but when it comes to popularity, the apps which are bulky are more popular than the former.

Most common word use for apps review

The word clouds can be used to get a visual representation of any textual data, in this case the user reviews. The higher the number of times a particular word is repeated, the bigger and bolder it gets. The words which are not repeated as much have a smaller text height and are not in bold. Hence the word clouds can be used to get a bird's eye view of all the textual data in the dataset. The following is the word cloud of all the user reviews.

This can be a used on any subset on the dataset which contains textual data. In this case, we can analyze the mood of the general audience and get a picture of what they are saying about the app.

Analysis Summary

- Percentage of free apps = ~92%
- Most competitive category: Family
- Category with the highest number of installs: Game
- Category with the highest average app installs: Communication
- Percentage of apps that are top rated = ~80%
- The median size of all apps in the play store is 12 MB.
- The apps whose size varies with device has the highest number average app installs.
- The apps whose size is greater than 90 MB has the highest number of average user reviews, i.e., they are more popular than the rest.

Challenges Faced

- Reading the dataset and comprehending the problem statement.
- Handling the error, duplicate and NaN values in the dataset.
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.

Different Python libraries used to complete this EDA

- Pandas
- NumPy
- DateTime
- Matplotlib.Pyplot
- Seaborn
- Wordcloud
- Stopwords

Conclusion

These are some of the aspects that the developer should research before proceeding with the app development. By conducting a simple exploratory data analysis (EDA) on the play store dataset, we not only eliminate avoidable risks of failure, but we may also be able to provide better ideas for building the app. These are some of the aspects that the developer should research before proceeding with the app development. By conducting a simple exploratory data analysis (EDA) on the play store dataset, we not only eliminate avoidable risks of failure, but we may also be able to provide better ideas for building the app.

References

GeeksforGeeks

Stackoverflow

Towards data science

Python libraries documentation