EXPLORATORY DATA ANALYSIS - GOOGLE PLAYSTORE APPS

EAS 503 – Programming and Database Fundamentals for Data Scientists

Submitted by Group 1

Karun Parashar
Roshan Ramesh
Kavya Anantha Rao

Project Guide
Dr. Mohammad Zia

ABSTRACT:

To understand the most significant factors that app developers consider while developing a new app and the criteria they set as a benchmark, to define success.

INTRODUCTION:

The aim of this project was to identify the important metrics/factors that define a successful application in the real world.

DATA:

The data set was procured from https://www.kaggle.com/lava18/google-play-store-apps. It consists of two .CSV files.

Table	Record Count	Table Description
Play table (Google_Play_Store)	9660	Stores details of the applications on Google PlayStore such as the app name, category, genre, type, content rating, etc.
UR table (Google_Play_Store_User_Reviews)	1074	Stores the first 'most relevant' 100 reviews for each app. It also contains the sentiment and the sentiment scores

EXPLORATORY DATA ANALYSIS:

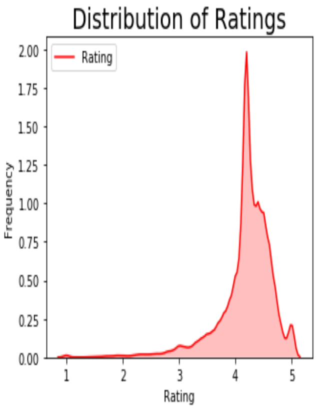
- ***Data Cleaning & Pre-Processing:***
  - Nan's were removed from the UR table and the user ratings that were NA's were replaced by the mean of the apps that had non-NA ratings
  - Unclassified type from Installs column was removed
  - Nan's were removed from the Play table for the column 'type'
  - Invalid category was deleted from the Play table

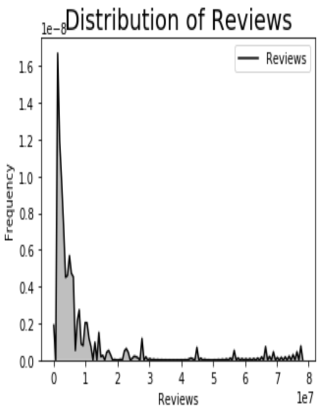
- ***Table Creation & Modification***
  - The 'trending_apps' table was created with the required columns and order to perform the necessary analysis, and was in turn updated to have data in upper case for standardization
  - The 'app_cat' table was created so that it contained the distribution of apps across various categories

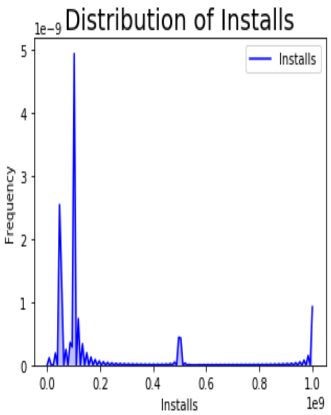
  - The datatype of 'Installs' and the ‘Reviews’ columns were changed to numeric and integer respectively
  - The table 'common_apps' was created to get the matching records from 'trending_apps' and 'UR', obtain the average polarity and subjectivity scores along with their respective average rating and reviews

- ***Analysis:***

  - Distributions of certain columns to understand how the data was spread

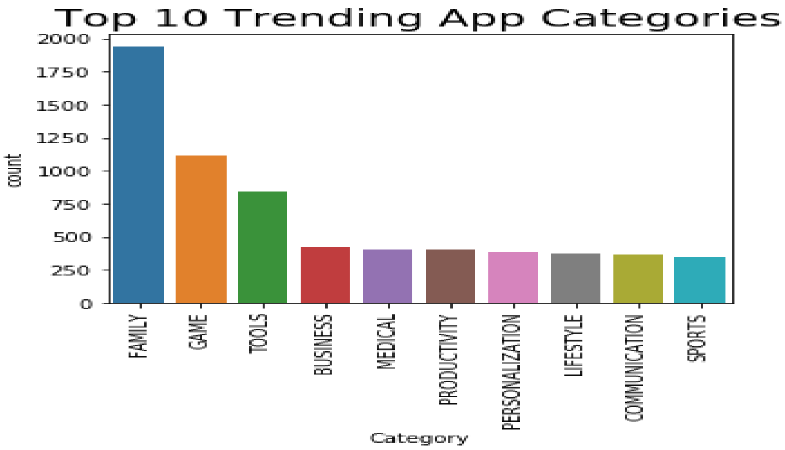


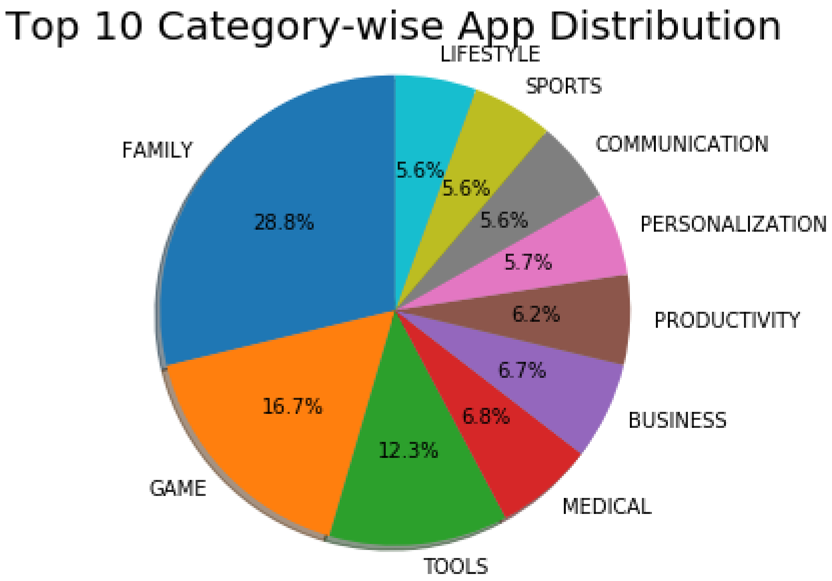




Insights: Most of the apps have an average user rating ranging from approximately 3.8 to 4.8, while most apps haven’t been given a user review.

  - Determined the Top N ('User defined') categories based on number of apps belonging to each category
  - A text field was created for entering the number of categories for which a user wanted to see the app distribution and the corresponding bar chart was plotted
  - A pie chart was also plotted that represents the category-wise app distribution of the Top N categories



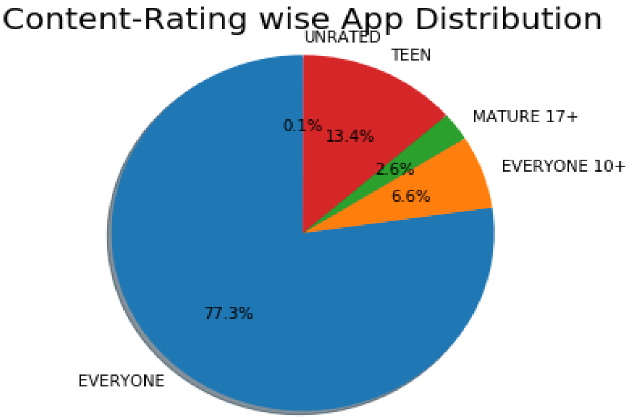


Insights: The 'Family', 'Game' and 'Tools' categories account for almost 57.8% of the total number of apps

○ A dropdown was created which enabled a user to select a specific category for visualizing the scatter plots that indicates the app name, category, genre, content rating, user rating, review, etc., on hover

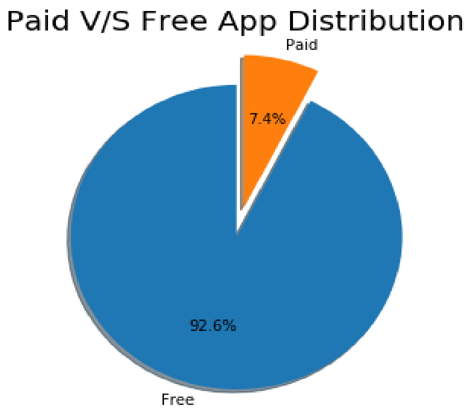
Insights: An app that had a user rating doesn't necessarily indicate that it's a widely popular app. On must also account for the number of users who have reviewed it and those who installed it. This would indicate how the app is performing in the current market, against other competitors

○ Percentage distribution of apps with respect to the content rating was analyzed



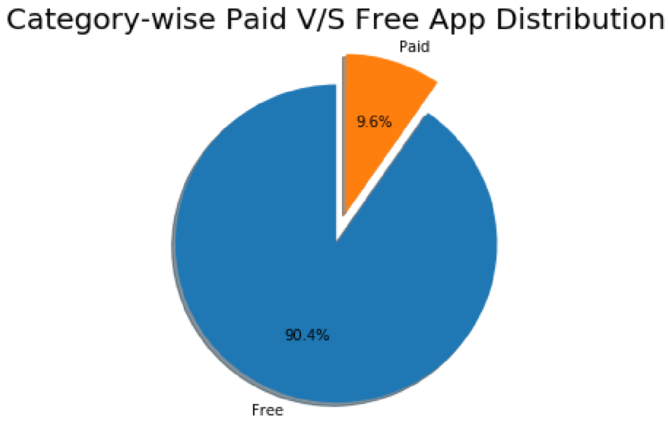
Insights: 77% of the apps are marketed for 'Everyone'

- Overall distribution of paid and free apps was plotted



<u>Insights:</u> Almost 93% of apps are freely accessible to everyone in the world

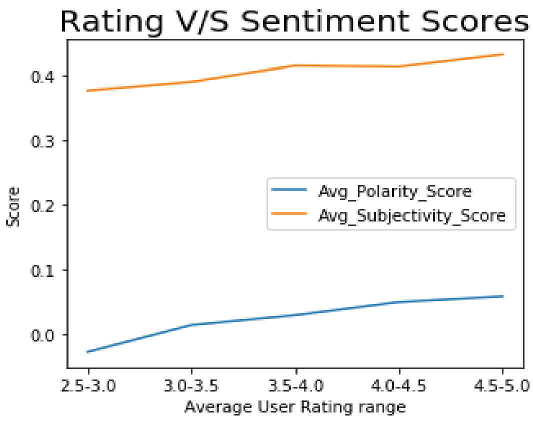
- A dropdown was displayed that enabled a user to select a category. The distribution of paid and free apps belonging to the selected category was plotted



<u>Insights:</u> 90% of apps belonging to the 'FAMILY' category are free

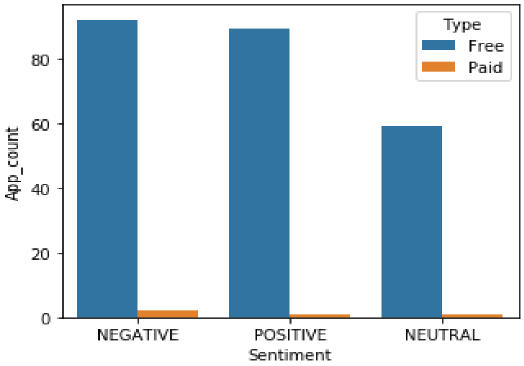
- Sentiment polarity and subjectivity scores are defined as follows:
  - ***Polarity:*** The emotion expressed in the sentence. It can be positive, negative or neutral. It ranges from -1 to 1
  - ***Subjectivity:*** The text in an explanatory article which must be analyzed in context. It ranges from 0 to 1

- User ratings were binned/grouped into 5 parts to better understand the relationship between the sentiment polarity and subjectivity score for apps belonging to a specific user rating group. Average Polarity and Subjectivity scores across all 5 user rating groups were plotted



Insights: A linear relation exists between the average rating and the polarity/subjectivity score which is natural, since the better polarity score an app has, it is very highly likely that a user has rated it relatively better

- The distribution of free and paid apps of a specific category across different sentiments was plotted to understand how users felt about free and paid apps



Insights: There's an almost equal distribution of free apps of the 'TOOLS' category that have been talked about in both a positive and negative way

CONCLUSION:

- A high user rating doesn't necessarily indicate an extremely popular app. User reviews and/or number of installs also need to be considered to correctly estimate how well an app is performing in the current market
- People tend to naturally prefer a free app irrespective of the category it belongs, rather than a paid app that accomplishes the same task
- A high user rating indicates a high sentiment polarity and subjectivity score
- Developers need to make sure that all the above factors have been taken into consideration before they release an app for a target audience

FUTURE RESEARCH DIRECTIONS:

- Top 'N' trending apps can be defined with respect to a specific category and content rating combination
- Sentiment analysis can be performed to capture the words that indicate how a user perceived different apps
- Word clouds can help us better understand the words that users most frequently use while reviewing apps. This will help developers tweak apps by incorporating the necessary changes so that their issues of various are addressed

REFERENCE:

- The data set was procured from https://www.kaggle.com/lava18/google-play-store-apps
- Visualization libraries used - Plotly Express, Matplotlib and Seaborn