Google Play Store Analysis

Eric Gong, Raian Rahman, Tony Li, Russel Tsang

Introduction

This statistical analysis is based on the Kaggle dataset ‘googleplaystore.csv’, which is a collection of web-scraped data of 10000 Play Store apps. Our goal was to analyze this data in order to generate interesting insights within the Android market. We decided to explore relationships between different meaningful attributes such as App Category, User Ratings, # of Reviews, Free vs Paid, and Content Ratings.

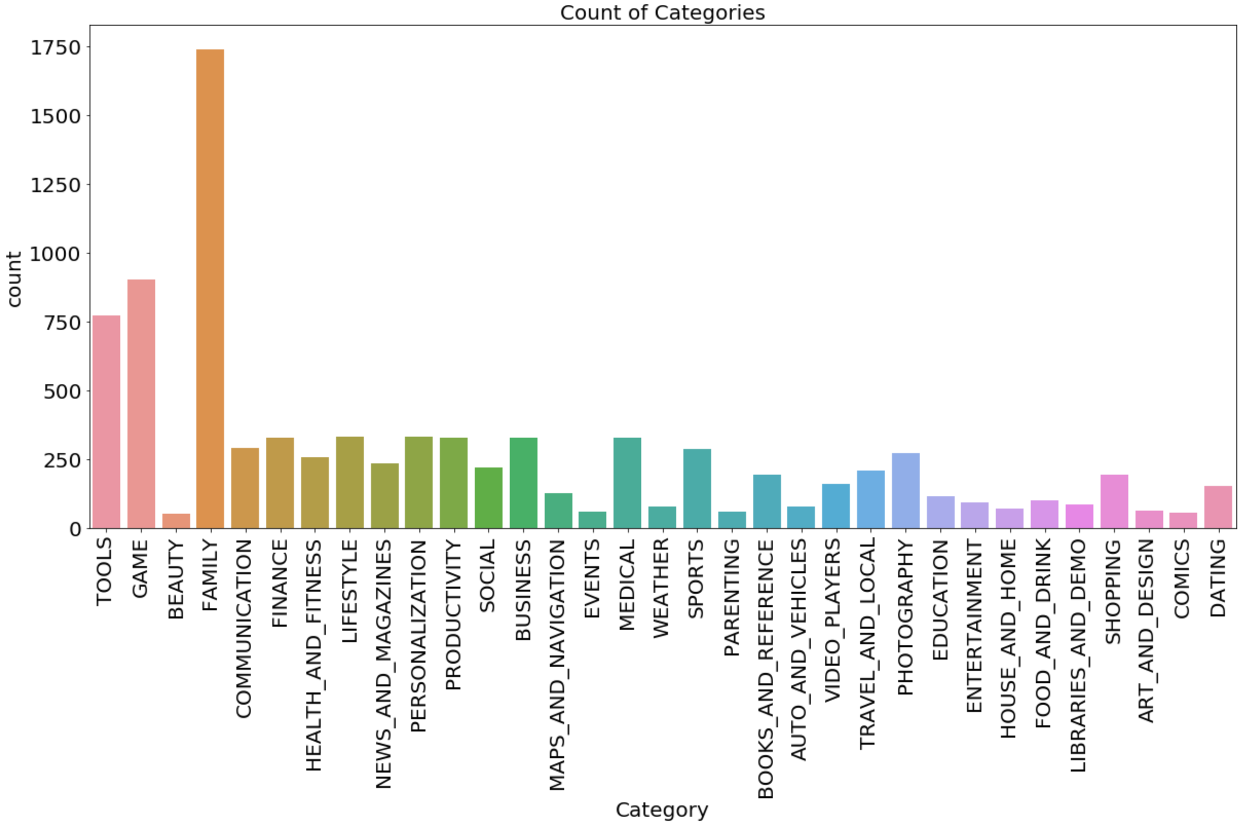
After removing duplicate App names, and categories with little to no apps, we were left with 8881 apps to analyze. One finding that is very visible is that the most popular app category is almost twice as popular as the next most popular app category (Family: 1827 apps vs Game: 927 apps). Through our box plot that compares App Category and User Ratings, we also found a range of median User Ratings across all app categories, spanning from 4.1 - 4.5. We continued to evaluate various significant relationships through additional graphs.

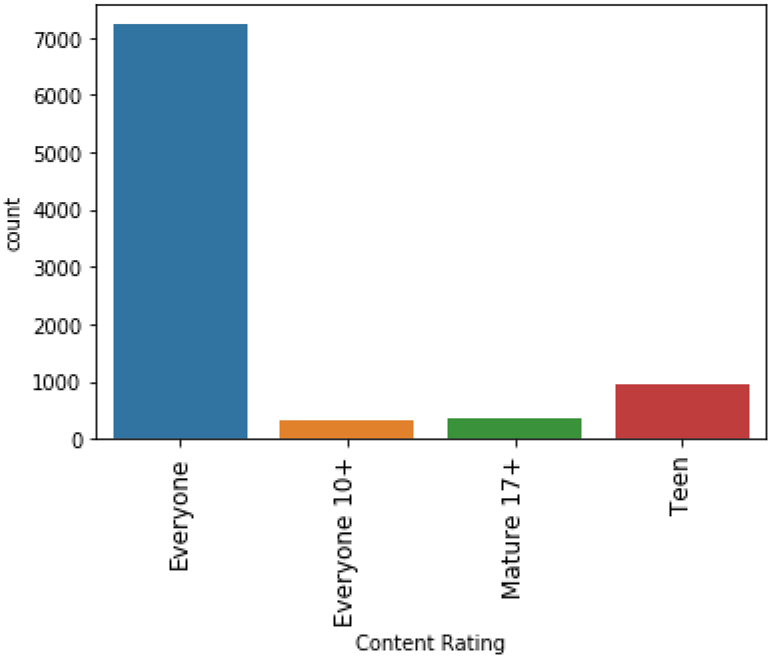
Furthermore, we continued to analyze the Google Playstore by inspecting a sentiment analysis data set through the dataset, ‘googleplaystore_user_reviews’. This dataset contained the first “most relevant” 100 reviews for 1075 apps. It also contained many missing values and values that read as “NaN”. After we cleaned this dataset, we were left with 35899 rows in the excel sheet to further examine. Using python, this dataset was merged with the original dataset in order to conduct an analysis on the amount of positive, neutral, and negative reviews. Lastly, we had also examined the sentiment polarity across different categories.

Initial Questions

Since we are all CIS Majors and Minors, and we are all considering creating Apps in the future, we thought it would be interesting to figure out and analyze what makes an app successful. We first started looking at the category column and tried to find if category was related to rating. Since rating is an important aspect to an app, we decided to base most of our questions on it. We asked questions such as “are certain categories are generally higher rated”, and “are certain categories are reviewed more?” In addition, we wanted to look into how many of each rating there is and how scattered they are. Furthermore, we wanted to evaluate how sentiment analysis plays a role in the success of an app. By exploring these types of relationships, we wanted to uncover possible determinants of a successful mobile application.

Category Popularity

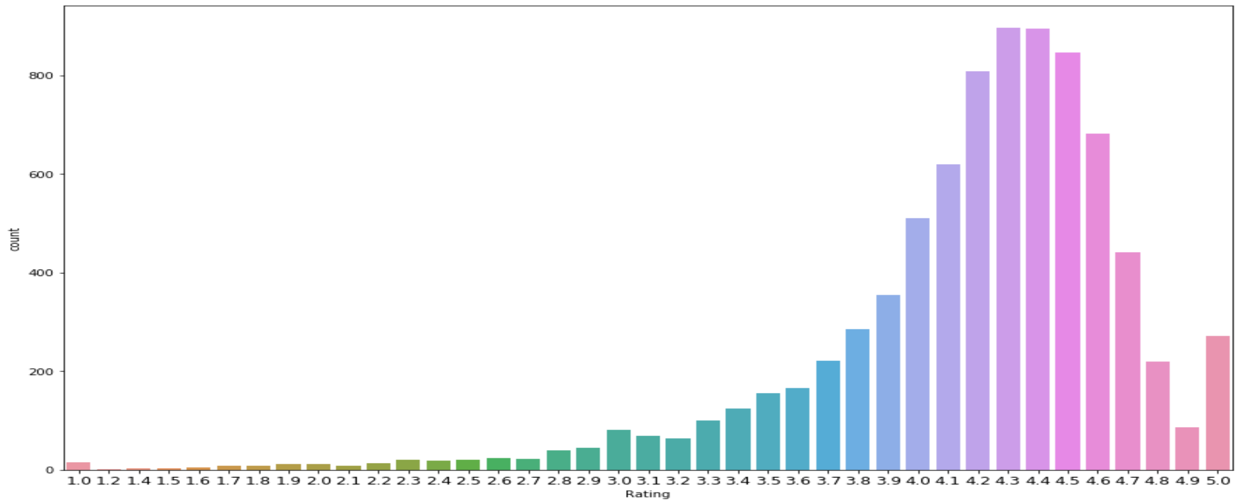


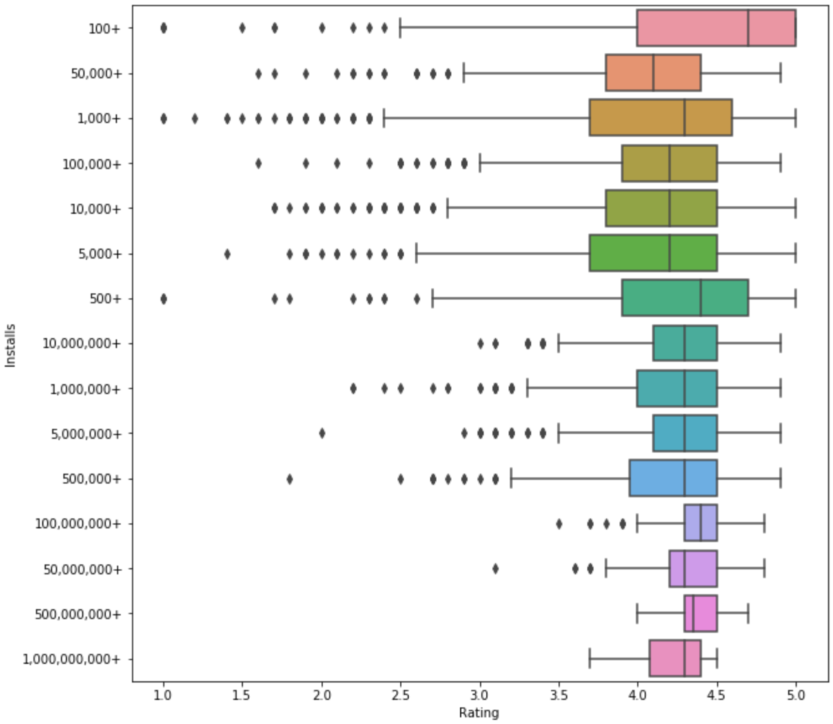


The first graph illustrates the number of apps for each category. By far, the most popular app category is ‘Family’, which is almost 2x as popular as the second highest ‘Game’ category. Perhaps developers assume that there is a high demand for child-friendly apps? Also, this graph may be a reflection of the popularity of these categories outside of the play store.

The second graph supports the results of the first graph: a huge majority of these apps are content suitable for Everyone. It makes sense that developers are targeting this content rating since they would want the biggest consumer base.

Rating and Installations

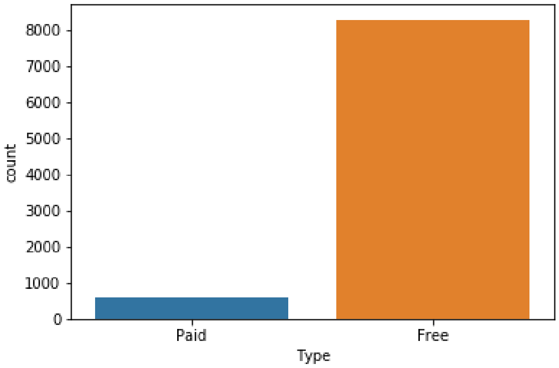




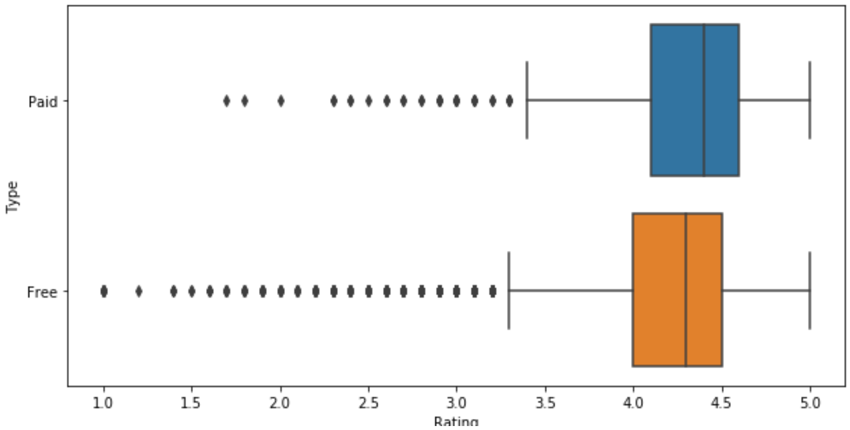
This graph illustrates the number of ratings. A majority of the ratings are in the upper range of the graph while only a few ratings are on the lower end.

For installations, its evident that for Apps with installations of 100+ there is a much higher rating median compared to the rating for apps with a higher amount of installations. This may suggest that after the range of 100+ installations, the rating reviews become slightly more negative.

Free vs Paid

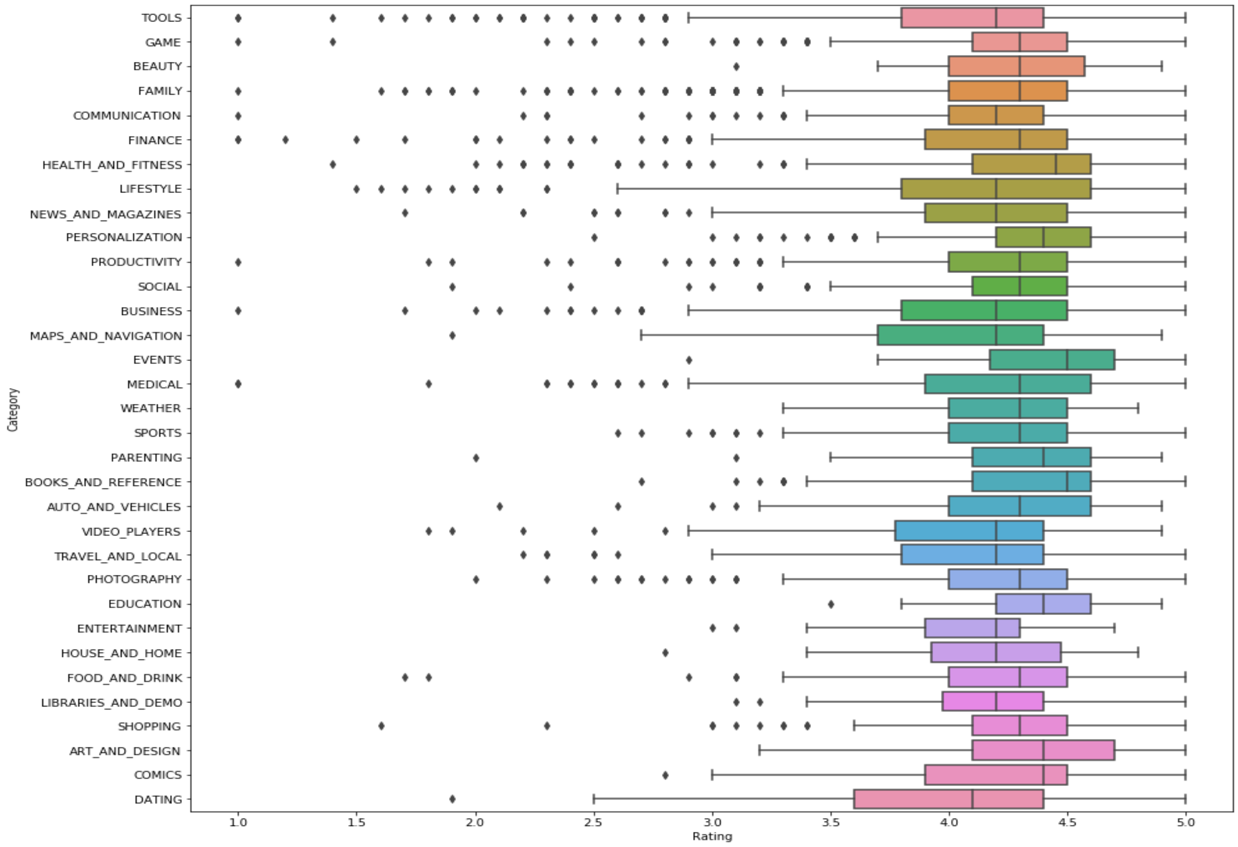


The first graph represents the amount of free apps there are compared to the amount of paid apps. It is clear from the graph that free apps are dominant by a large



margin in the Google Play Store. Although there are many more free apps than paid apps, we still wanted to see if there was any clear differences in ratings between them. After all, it seems natural for a paid app to be better which would result in it being rated higher. The second graph did indeed demonstrate that paid apps tend to be rated higher but only by a minuscule margin.

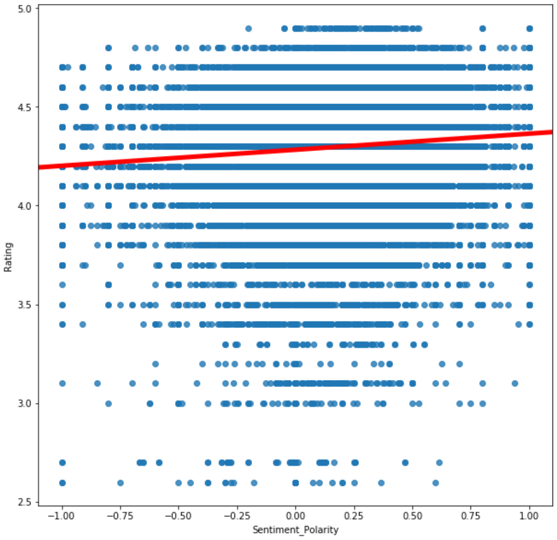
Ratings by App Category



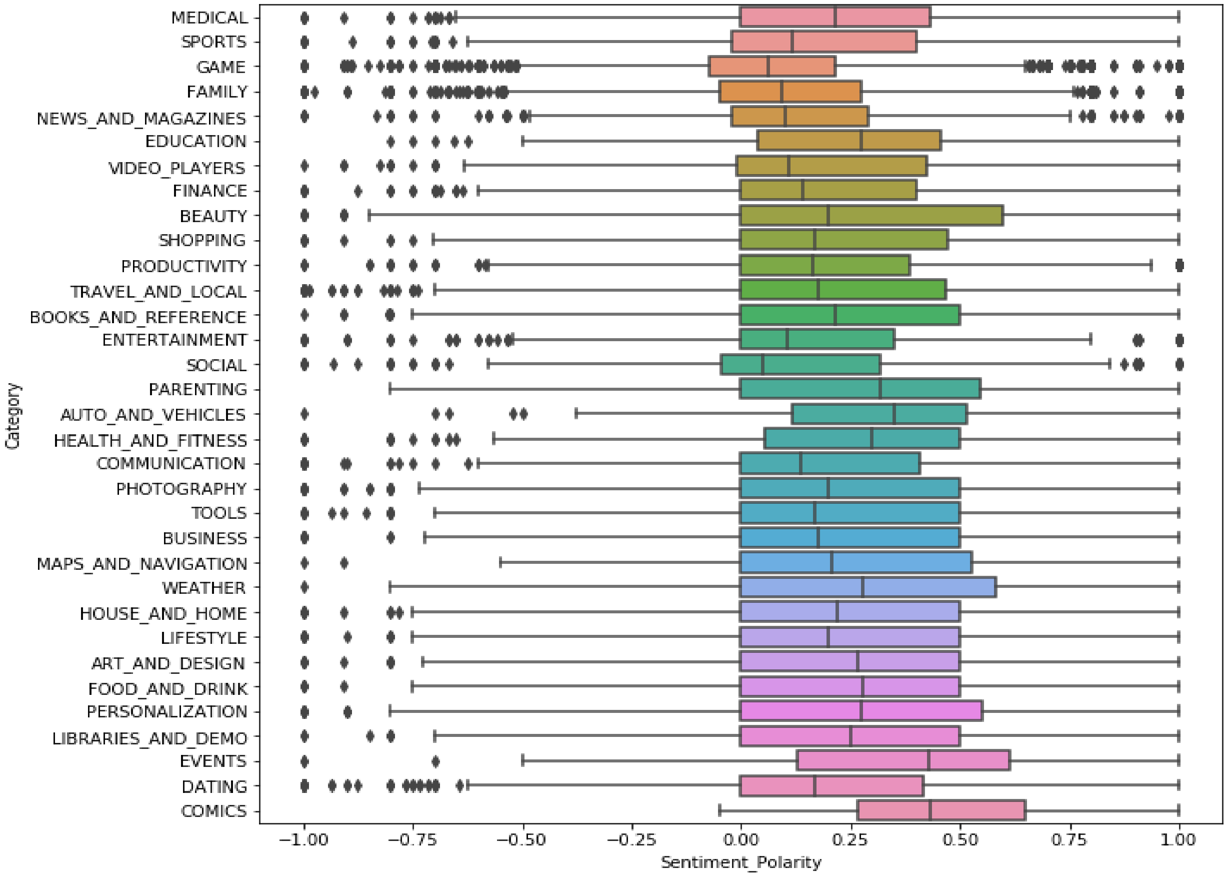
This box plot presents a range of 4.1 - 4.5 for median User Ratings across categories. There are also outliers presented in this graph for each category. This is a considerable spread, where:

- ‘Dating’ apps have a median rating of 4.1
- ‘Books_and_Reference’, ‘Events’, and ‘Health_and_Fitness’ have a median ratings of 4.5

Sentiment Analysis



The correlation between sentiment polarity and ratings is .09683, indicating that there is a weak positive correlation between the two. There are countless data points with a high rating and a low sentiment polarity and vice versa. This may indicate the need for improvement in detecting sentiment polarity from texts. Furthermore, after looking at the boxplots with the categories and sentiment polarity, it can be seen that majority of apps are viewed positively. The 'Events' and 'Comics' categories had the highest sentiment polarity while the 'Social' and 'Game' categories had the lowest sentiment polarity.



Sources

The following datasets were used for analysis from Kaggle:

https://www.kaggle.com/lava18/google-play-store-apps#googleplaystore.csv

https://www.kaggle.com/lava18/google-play-store-apps#googleplaystore_user_reviews.csv