What App is That?

Capstone Report

Alex Yom

Table of Contents

1. Introduction
2. Analysis
3. Discussion

Introduction

With so many apps of many different areas out on the market these days, my capstone project will explore what factors may play a role in the app market. Especially in today's society, where the next big thing spreads like wildfire but can be doused in the blink of an eye, people are always trying to stay ahead of the curve. With technology becoming increasingly accessible, cell phones are at the forefront. The dataset for this project consists of all the apps on the google play store with several properties of the apps such as type, price, number of downloads, etc.

The ultimate goal of any app-developer is to produce apps that are downloaded by as many people as possible. Whether the apps are free or not, the more an app is downloaded, the possible revenue increases. This data would be crucial for both current app-developers and developers who looking to release new apps. This data could be used to assess the current market and see which apps are being downloaded most by people. There can be several factors that influence the popularity of the app and this project will seek to identify those relationships. Current app-developers who have released apps on the play store could use this data to see how their app compares to other apps on the market. By viewing the data, developers may see potential changes to make in their app to improve performance and increase downloads. This data could also be used by prospective app-developers and companies who are interested in developing their own apps. This project can show which type of apps are downloaded more often by people and what other factors play a role in the popularity of an app. After comparing and reviewing if there are substantial effects, the information can be used by companies and developers to identify which areas they want to pour out their resources in. For example, if there is a certain category of apps that gets downloaded more frequently, then companies might be inclined to develop those kinds of apps.

Other companies could also use to data to select which app they would potentially like to sponsor and put their advertisements in. By putting an advertisement for their company in an app that many people have, the greater the exposure for the advertising company.

This dataset and project could also interest someone who is looking to see the effect technology is having on our generation. With the access to phones reaching unparalleled levels

from kids as young as elementary school to elders having access to phones. With more and more people having phones, means a wider population of people to download apps. Researchers could use this project to see which apps and what kinds of apps are being downloaded most by people and conduct a social experiment.

To see which factors influence the downloads of the apps, I plan to test the relationship between such factors such as rating, price, and category on the amount of downloads. It will be interesting to assess and compare the correlations between each factor on the number of downloads. On a more macro-scale, I plan to visualize if there is a large disparity between which kinds of apps are downloaded more often than other categories as this could be useful information as well.

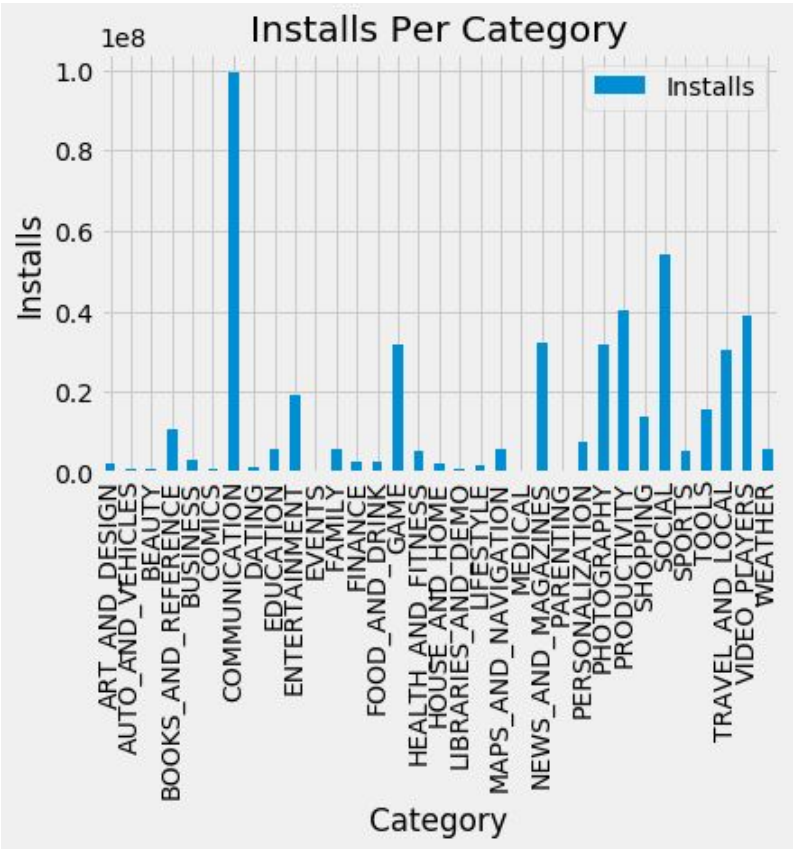
The first step in beginning my capstone project is to review and investigate my dataset. My dataset is information regarding the applications on the google play store. Each application is listed along with the ratings, amount of reviews, number of installs, etc. My goal of the project is to investigate how certain characteristics about these applications can affect the number of installations for an app. Before diving into the trends and correlations, I had to make sure the dataset was clean and ready to be used for analysis.

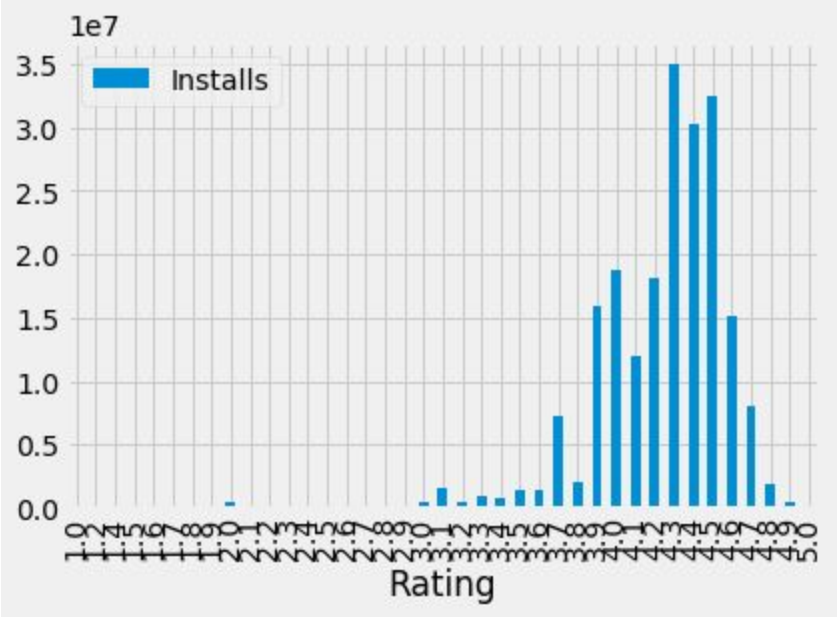
This dataset was obtained from kaggle.com, which is a website database that provides a wide variety of datasets. At first glance, the dataset seemed to be very clean. It looked like the long and tedious process of tidying data and data wrangling would not be needed for this dataset. That was not the case however, after some preliminary exploration of the data. The main dependant variable for this project, the number of installs per app, had to be changed from a float to an integer to facilitate graphing. The values under the 'Installs' column contained the "+" and the "," symbols which led to complications when trying to graph the values. All symbols were removed and then the type was changed to a numeric value. The dataset also contained NaN values that could have posed problems when trying to group values and graph so they were removed. The NaN values appeared in multiple rows and columns and hindered the visualization process so they were removed from the dataset. Being such a large dataset, I had to categorize the apps and group them by the type of category they classified as.

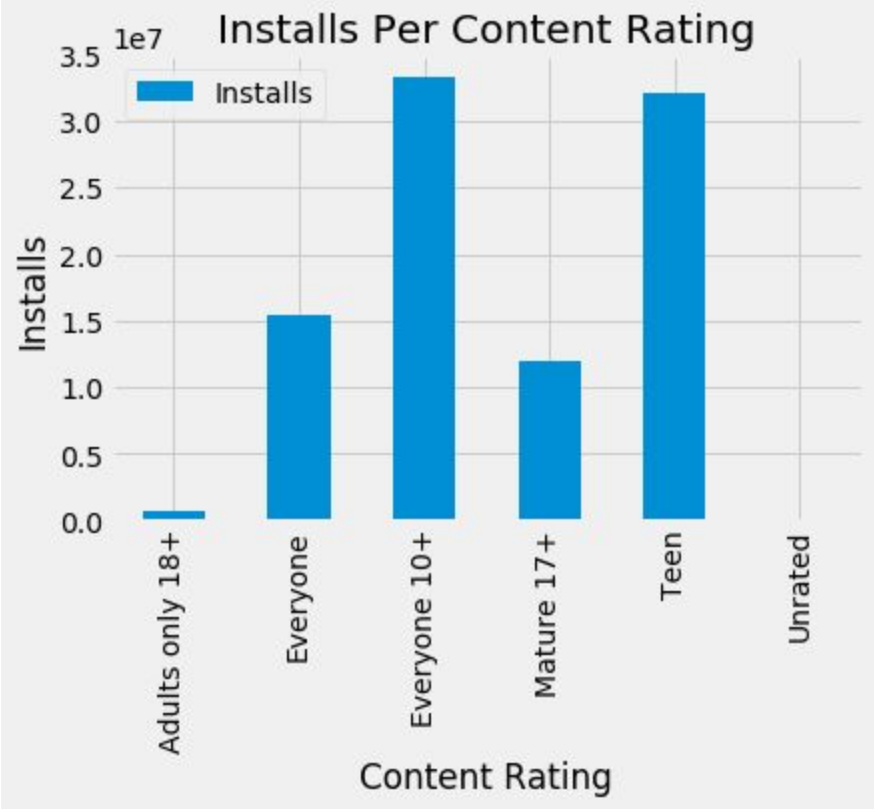
This would allow the graphs to maintain reliability and also could be used to show what categories are downloaded more often.

Analysis

After reorganizing the data so that it could be visualized, some bar graphs were produced to see the difference in installations. When comparing the types of categories to the amount of downloads, it is clear that communication apps are downloaded by a substantial margin over other kinds of apps. Social, productivity, and video playing apps followed behind respectively. Interestingly, it seemed like apps that had a rating of 4.3-4.5 out of 5 were the most installed apps. It would be assumed that the apps that have the best rating would be downloaded the most but that does not appear to be the case. Apps that are free are downloaded more than apps that require payment. This makes sense knowing that there are far more apps on the play store that are free than apps that are paid. Apps that are rated for everyone 10+ and teen are the most downloaded apps. The graphs are included below for reference.

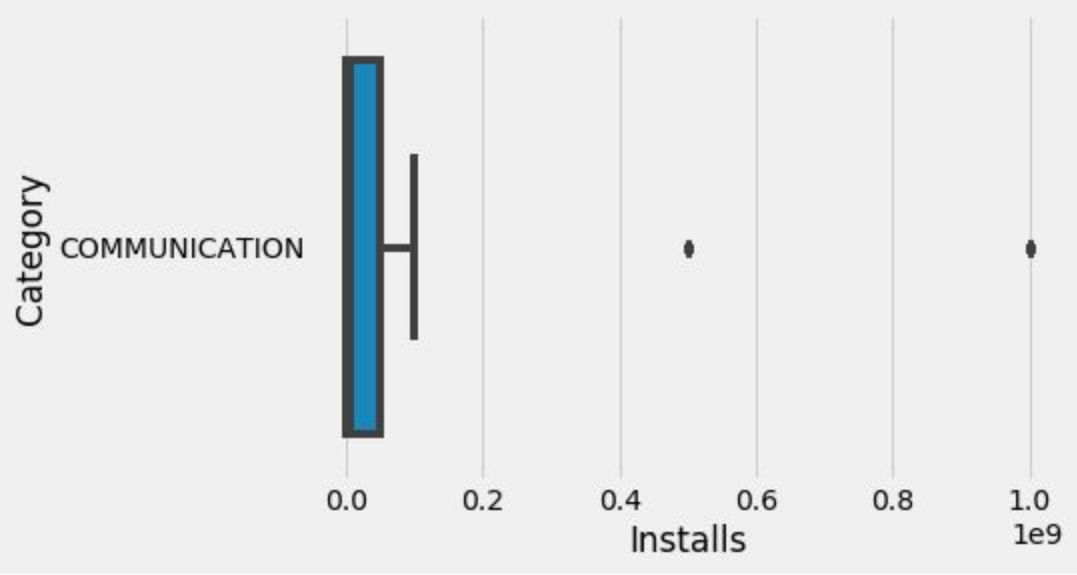


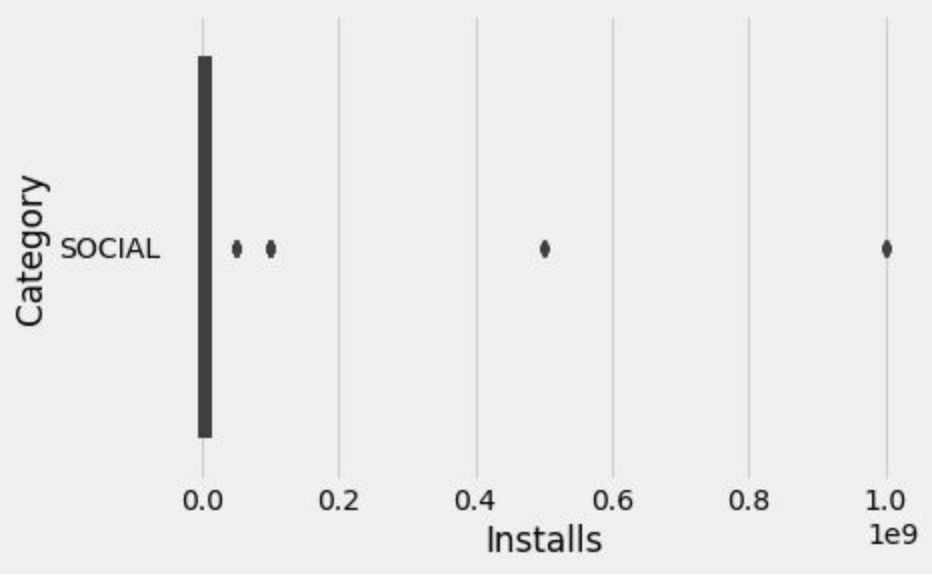

1e7
Installs
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rating

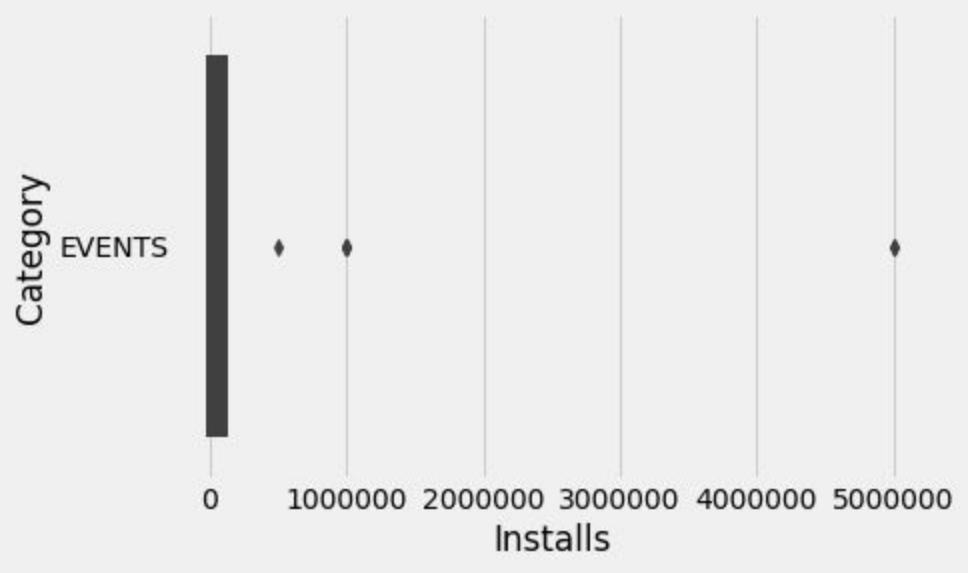

Installs Per Content Rating
1e7
Installs
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Installs
Adults only 18+
Everyone
Everyone 10+
Mature 17+
Teen
Unrated
Content Rating

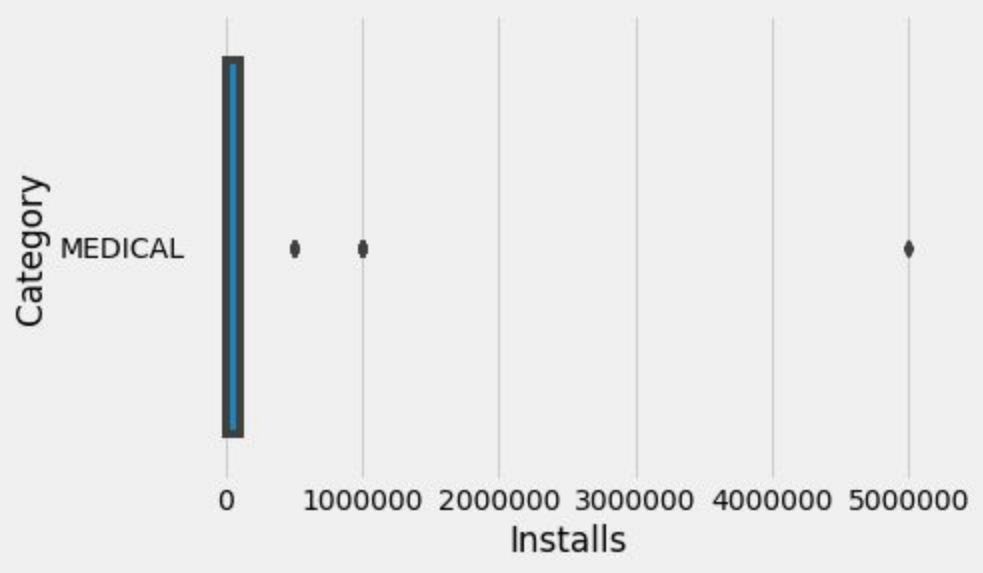
The graphs show that some apps that fulfill certain categories do get downloaded more on average.

After doing some preliminary analysis and the variety of features that could potentially play a role in the number of installations, I thought diving into each specific category could illuminate some interesting information. Plotting box plots for each category would help visualize the totality of every app under their respective categories to get a more refined representation of their data. Below are the boxplots for the two most downloaded categories and two least downloaded categories on average. The communication category has an average of 99,534,273 downloads across all apps. The social category recorded an average of 54,323,712 downloads. The events category recorded an average of 354,431 downloads and medical apps recorded an average of 152,016 downloads.




Category
SOCIAL
0.0
0.2
0.4
0.6
0.8
1.0
1e9
Installs


Category
EVENTS
0
1000000
2000000
3000000
4000000
5000000
Installs



As pictured above, the boxplots that are plotted for the specific categories do not look like traditional boxplots. This can be credited to the high variance in the data of each category. There is a wide spectrum between the least downloaded app and the most downloaded app and few apps that can skew the numbers to create such boxplots. This is not necessarily bad because this shows just how wide of a spectrum there is when trying to predict how many installations an app will receive. There is a nature of extreme unpredictability in an app's popularity. Another takeaway from these plots is that although some categories, such as communication apps and social apps, project to be very popular, there are communication and social apps that are less popular than medical or events apps.

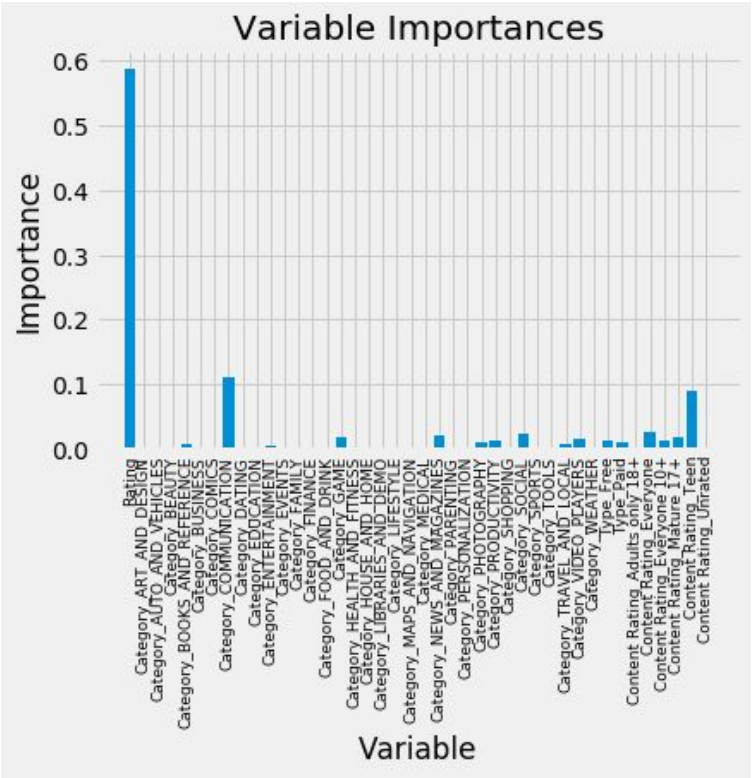
Plotting graphs of the information present in the dataset was very useful in visualizing the information and rendered the data into a very easy to interpret platform. The next step after exploring the data available would be to see if we could use the data that we have and use it to make predictions for the future. That is where the techniques of machine learning can be applied. There are many machine learning techniques that could be used and it was a matter of which technique would be optimized with our data. KNN could be used if we had few samples and few features. Deep learning could be used if we had many features and many samples. SVM could be used if we had many features but few samples. Random forests could be used if we had few features but many samples. Because our dataset had many samples but a only a few features, I thought random samples would be applicable in this case. The random forest

technique was applied to this dataset in order to see if we could use the data in order to predict the possible amount of downloads for apps in the future as well. The random forest algorithm would learn from our data by combining decision trees to make a forest. By using the information we have, random forests would be able to learn the regression and trends of the data and potentially output predictions.

Before separating the data into training and testing sets to prepare for forests, I had to clean the dataset to include only the features that were relevant. After removing the features that would not be used, I had to get dummies for the features that would be used for the random forest. I completed the random forest ensemble after separating the data into training and testing sets. The results yielded a mean absolute error of 26,079,931.46 downloads with an accuracy of -1572542.47 %.

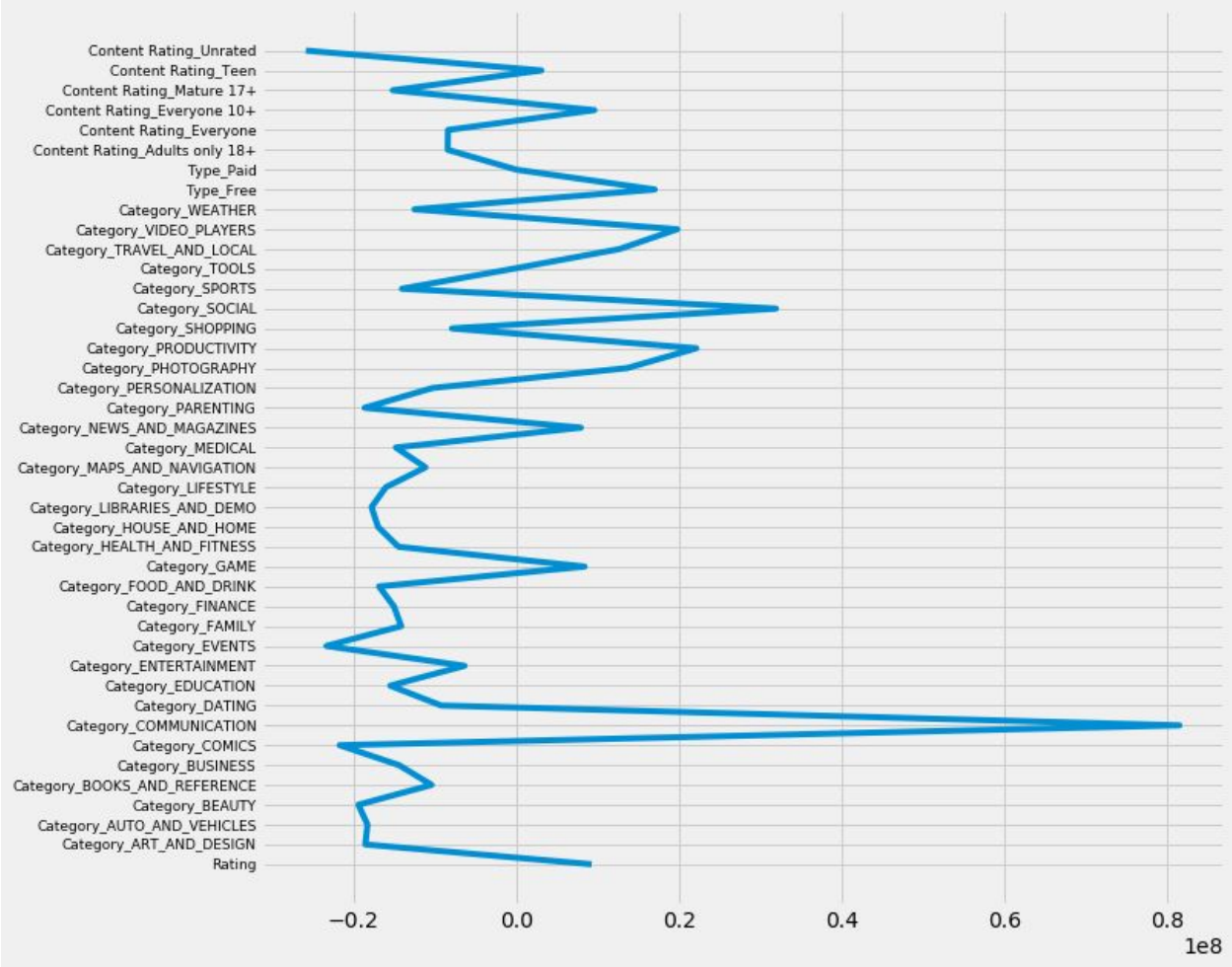
I also thought it would be interesting to see which features had the biggest influence. The rating and the communication category features had the biggest influence. Random forests were ran again only using these two features this time. The mean absolute error for this ensemble was 27,612,929.97 downloads with an accuracy of -1471099.45%.

A graph was plotted to show the importance of the features in a visual way to show which features produced impactfulness on installs. The graph shows that rating and a category classified as communication are the two biggest factors. The content rating of an app being ‘teen’ was a close third.



Although it looks as though the rating value of an app and the categories seem like they can have large impacts on predicting the installation numbers, it is important to remember the results from our random forests. By taking the characteristics of the most popular apps and trying to release an app that has very high potential, following our results from the random forests, there is a huge swing of nearly 26,000,000 downloads.

A lasso regression was fit unto the data as well, to further validate our findings. The goal was to see if the results of the lasso regression would coincide with the results of the random forests. As shown below, the lasso regression predicted that apps belonging to the social and communication category would play the biggest influence in installations.


Content Rating_Unrated
Content Rating_Teen
Content Rating_Mature 17+
Content Rating_Everyone 10+
Content Rating_Everyone
Content Rating_Adults only 18+
Type_Paid
Type_Free
Category_WEATHER
Category_VIDEO_PLAYERS
Category_TRAVEL_AND_LOCAL
Category_TOOLS
Category_SPORTS
Category_SOCIAL
Category_SHOPPING
Category_PRODUCTIVITY
Category_PHOTOGRAPHY
Category_PERSONALIZATION
Category_PARENTING
Category_NEWS_AND_MAGAZINES
Category_MEDICAL
Category_MAPS_AND_NAVIGATION
Category_LIFESTYLE
Category_LIBRARIES_AND_DEMO
Category_HOUSE_AND_HOME
Category_HEALTH_AND_FITNESS
Category_GAME
Category_FOOD_AND_DRINK
Category_FINANCE
Category_FAMILY
Category_EVENTS
Category_ENTERTAINMENT
Category_EDUCATION
Category_DATING
Category_COMMUNICATION
Category_COMICS
Category_BUSINESS
Category_BOOKS_AND_REFERENCE
Category_BEAUTY
Category_AUTO_AND_VEHICLES
Category_ART_AND_DESIGN
Rating
−0.2
0.0
0.2
0.4
0.6
0.8
1e8

Discussion

The data analysis and visualization of this data highlight the unpredictable nature of trying to speculate how popular an app will be. The boxplots illustrated that no matter the category of an app, there is no certainty that the app will have success. The shape of the boxplots show that because of the huge quantity of apps out on the market, the most popular ones that are nearly ubiquitous are mostly anomalies . Although communication apps have the highest average downloads, it would still be possible for a popular medical app to have more downloads than a random communication app. The results of the random forest model portrayed the rating and categorization as communication as the two most influential features in predicting downloads. Meanwhile, the results of the lasso regression model selected the communication and social categories as the two most important factors, with rating falling below several other features. This disparity between the two methods further reinforces the difficulty in predicting installation numbers. The random forest predictor model resulted in a mean absolute error of about 26,000,000 downloads. While this potential plus/minus seems very large, it may not be in a different context. If a company was looking to launch a new communication app that they felt confident in and they ran this random forest, looking at the current average number of downloads for communication apps which comes in at over 99,000,000 downloads, the company might look to move forward with the app.

While the analysis and predictive models were able to locate the key features that influence the installation numbers, it remains to be seen if there are other factors that were not included in this dataset that could have a more profound impact on the number of downloads. One particular feature that would be interesting to further investigate in the future would be to look into the different age groups of people and see where the most downloads come from. It might be more effective for companies to target a specific age group and select the specific features that are most prevalent in said age group. By specifying the age group and running the machine learning tools again, the results may engineer a more accurate result. Adding other features may improve the machine's ability to learn and make more accurate predictions.