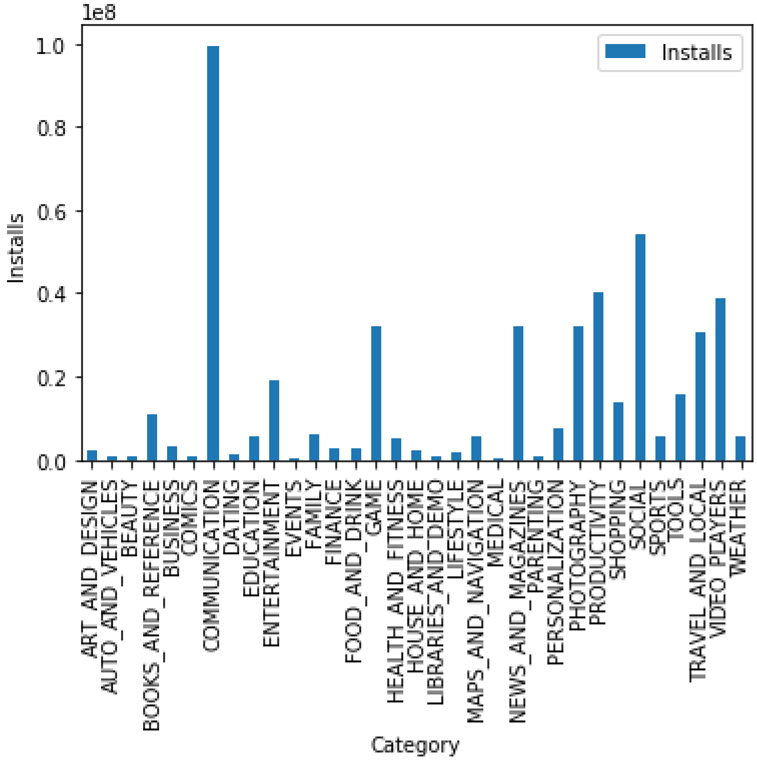
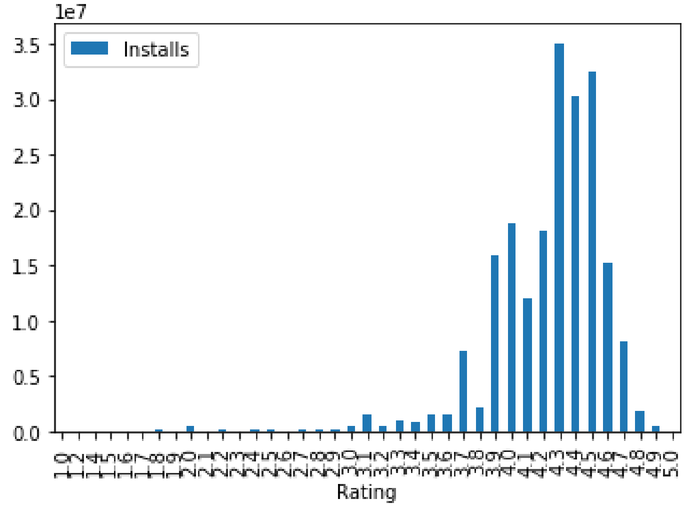
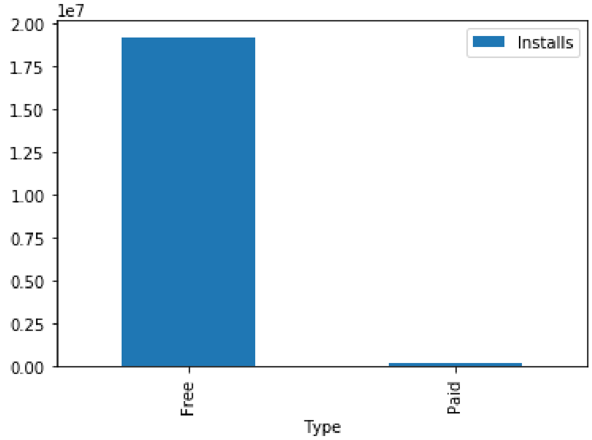
Alex Yom

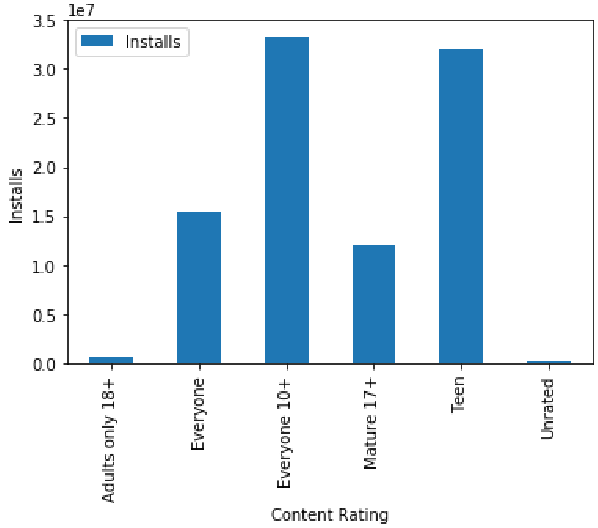
In depth Analysis

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
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Installs
1.0
1.2
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5.0
Rating


1e7
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Installs
Free
Paid
Type


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

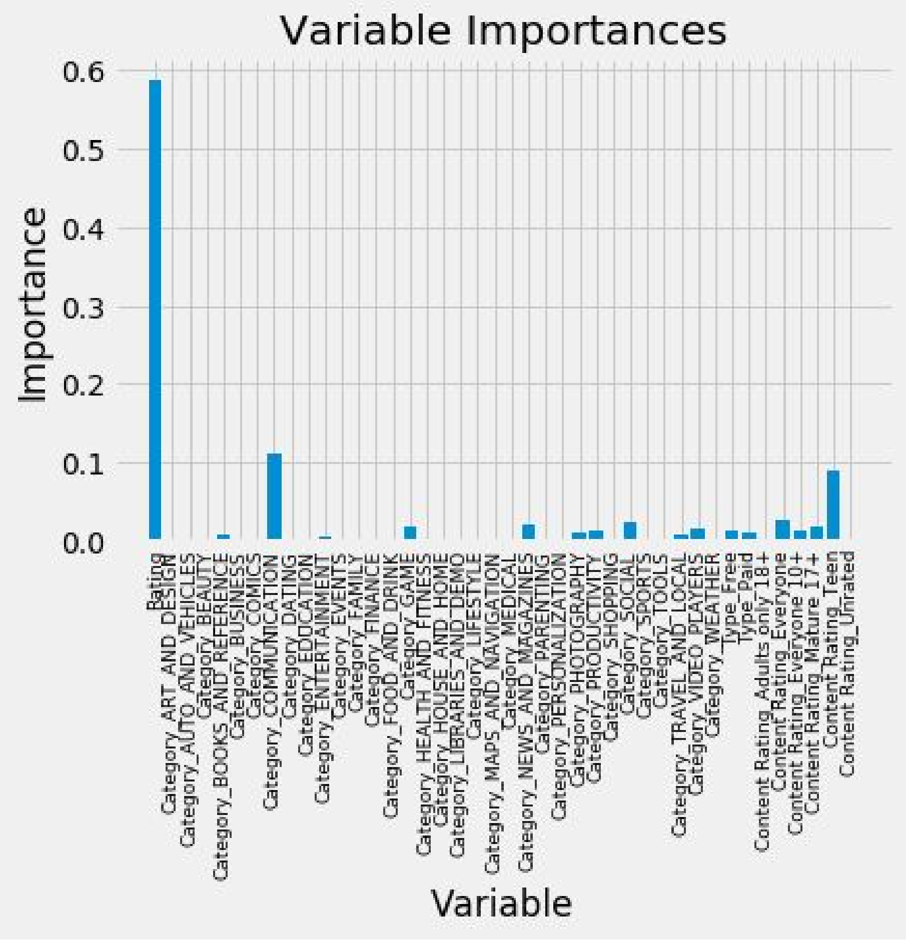
The graphs show that some apps that fulfill certain categories do get downloaded more on average.

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.


Variable Importances
Importance
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Rating
Category_ART_AND_DESIGN
Category_AUTO_AND_VEHICLES
Category_BEAUTY
Category_BOOKS_AND_REFERENCE
Category_BUSINESS
Category_COMICS
Category_COMMUNICATION
Category_DATING
Category_EDUCATION
Category_ENTERTAINMENT
Category_EVENTS
Category_FAMILY
Category_FINANCE
Category_FOOD_AND_DRINK
Category_GAME
Category_HEALTH_AND_FITNESS
Category_HOUSE_AND_HOME
Category_LIBRARIES_AND_DEMO
Category_LIFESTYLE
Category_MAPS_AND_NAVIGATION
Category_MEDICAL
Category_NEWS_AND_MAGAZINES
Category_PARENTING
Category_PERSONALIZATION
Category_PHOTOGRAPHY
Category_PRODUCTIVITY
Category_SHOPPING
Category_SOCIAL
Category_SPORTS
Category_TOOLS
Category_TRAVEL_AND_LOCAL
Category_VIDEO_PLAYERS
Category_WEATHER
Type_Free
Type_Paid
Content Rating_Adults only 18+
Content Rating_Everyone
Content Rating_Everyone 10+
Content Rating_Mature 17+
Content Rating_Teen
Content Rating_Unrated
Variable