Alex Yom
Capstone Project Draft

Capstone Project Draft

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. 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. There does not appear to be much data cleaning to do as this is a rather clean dataset. I will explore the different relationships that could be found within this data.

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 download, 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.

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.

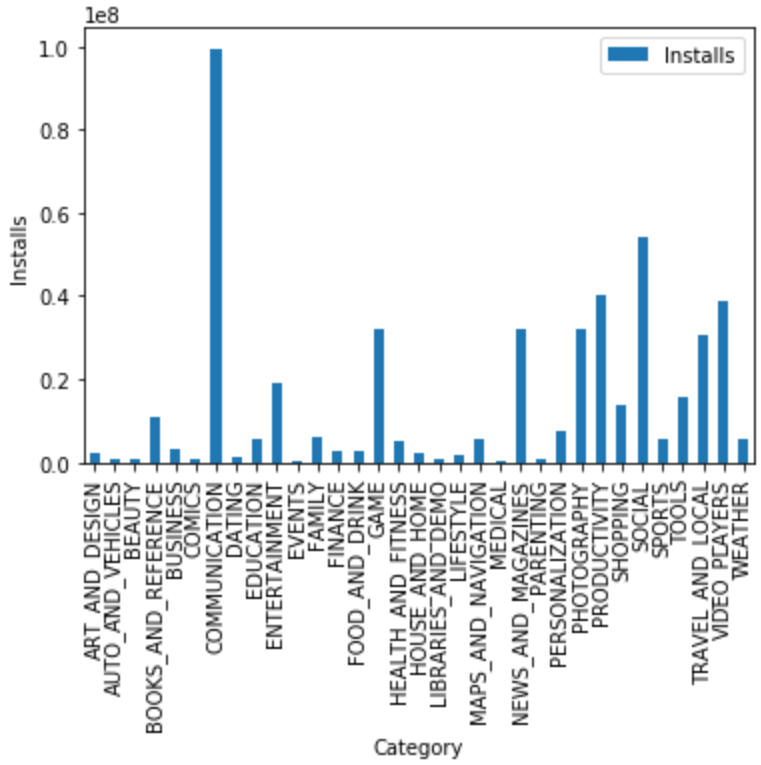
My capstone report will detail the introduction to my project and detail the different steps I take in this project. A slideshow containing the noteworthy findings and using a multitude of graphs should suffice in presenting the information and data. I plan to use a variety of graphs to show the different relationships and correlations that are found in this project.

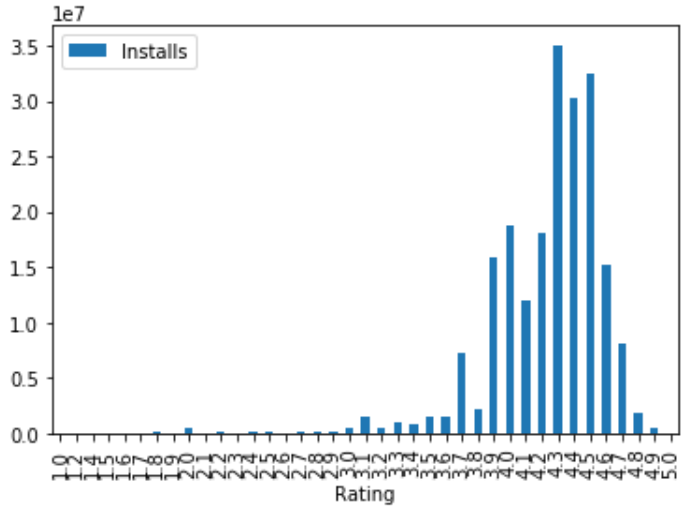
The first step in beginning my capstone project is to review and investigate my dataset. My dataset is information regarding the applications on the google play store. Each application is listed along with the ratings, amount of reviews, number of installs, etc. My goal of the project is to investigate 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.

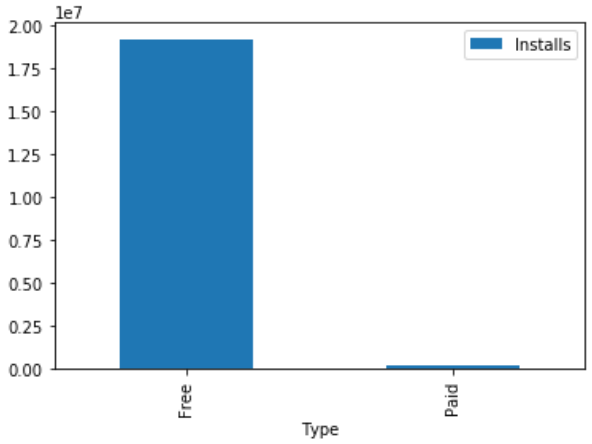
This dataset was obtained from kaggle.com, which is a website database that provides a wide variety of datasets. At first glance, the dataset seemed to be very clean. It looked like the long and tedious process of tidying data and data wrangling would not be needed for this dataset. That was not the 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.

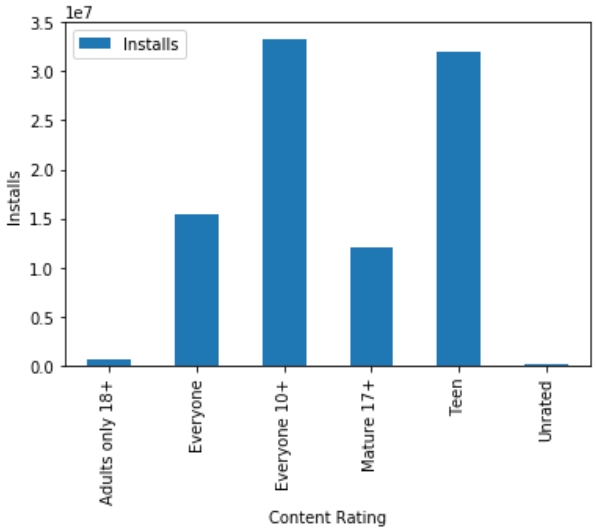
After reorganizing the data so that it could be visualized, some bar graphs were produced to see the difference in installations. When comparing the types of categories to the amount of downloads, it is clear that communication apps are downloaded by a substantial

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
Rating


1e7
Installs
Free
Paid
Type


1e7
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.