Alvaro Hu

Springboard

November 11, 2018

An In-Depth Analysis of Google Play Store Data

I. The Proposal

Almost 20 years into the 21st century, pretty much everything can be done within the confines of a virtual environment. And many of these virtual spaces can be downloaded from the internet in a matter of seconds in the form of an app. It’s hard to believe that only 25 years ago, the internet was not accessible by even the richest people in society. You want to take notes or make a grocery list? There’s an app for that. You want to watch Football on your phone while waiting for your haircut? There’s an app for that. You want to keep track of your stock holdings to make sure you don’t lose money on an investment? There’s an app for that too. There seems to be an app for everything, yet all over the world, especially in regions like Silicon Valley or metropolitan areas like Dallas/ Fort Worth, there are creative innovators designing new and improved versions of apps every day. With my analysis, I hope to be able to steer mobile app developers towards making apps that will yield in the highest returns, both in amount of installs and in revenue earned.

I will use two CSV files available on Kaggle.com. The first one provides me with over 10,000 different apps, and the second provides me with over 60,000 different reviews for the various apps. These are not every review for each app, rather it has provided me with up to 100 of the “most relevant” reviews for various apps. After doing a join on the dataframes, I found out that, of the 10,000 apps in the first CSV file, only about 900 of them were represented in the review dataframe, but this might still be able to serve my purpose. Some of the more important features in the first dataset are the name of the app, the number of reviews, the rating out of 5, the category the app is in, the price of the app, and the number of installs. The number of installs is a tricky feature because it rounds down to the nearest “benchmark”. For example, and app may have 1,850,000 downloads, but it will be filed under the “1,000,000+” benchmark.

The important feature in the second CSV file is the actual review. One thing to take note of for the review is that it has been translated from whatever language it was originally in, and some of these reviews can look questionable at best. There is a provided “sentiment” column for whether the review is positive, negative, or neutral, but I will not use this column. Rather, I will use my own natural language processing algorithm to predict the sentiment of the review.I can use these features to figure out the answers to my questions.

My client is every single developer out there looking for what their next project should be focused on, depending on if they want recognition (downloads) or money (revenue). By providing an in-depth analysis of what genres and types of apps give the most downloads in general, then it would alleviate some of the decision making process for developers when they are deciding on what to work on, or what project they want to help fund or hop in on.

The way I will be approaching this problem is by checking either the mean or the median installs for each of the categories of apps (which is around 30). I will also be basing the revenue aspect of this project on the proportions when compared to the entire category. For example, I will check to see how the average or mean amount of installs for apps in the Productivity category changes when the app is no longer free, so that I can adequately recommend to the developer whether or not it would be safe to charge for the service or app. As for whether or not negative reviews impact sales and installs, I can join the two datasets on the name of the app, and figure out if, overall, a majority negative reviews has the outcome of having less downloads, and if that depends on if the app is paid or if it is free.

II. The Wrangling

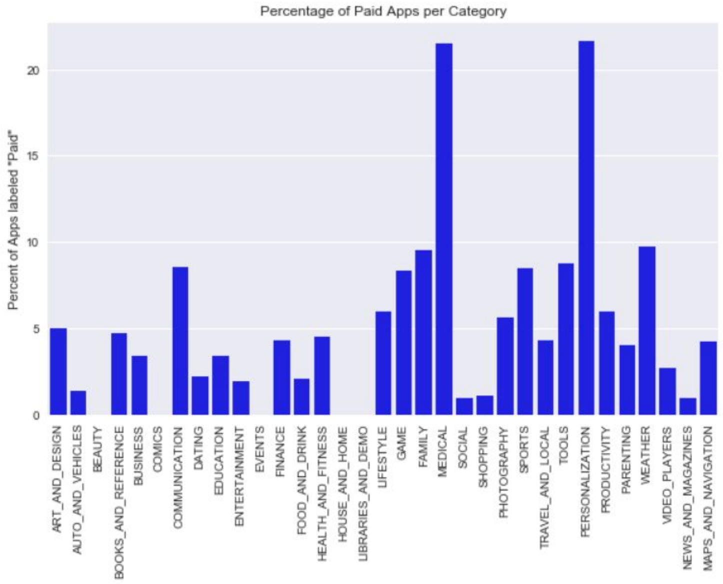
As stated earlier, I downloaded the data from Kaggle.com as csv files. Though data from Kaggle is usually preprocessed pretty well, there were still some issues that I needed to work out on my own. For example, there were still a few null values, but because there were so few, in a dataset of over 10,000 observations, I just decided to drop them. Another thing I had to do was clean up some of the columns to make sure that they could be used in mathematical operations and plot correctly. One such example of this was removing the "+" sign and commas in "10,000,000+" installs. A similar thing I had to do was convert the size column, the size of the app in bytes, into an integer. This was tricky because there were M and k, depending on if it was in Megabytes or kilobytes, so I couldn't just drop the letters. There were also quite a few apps with "varies with device" as its value for size. I could not just drop these either because many of them had over 10,000,000 downloads, meaning they are probably important. So what I did was, I looped through all values in that column, and if it ended in M, I removed the M, but if it ended in k, I would remove the k *and* divide the resulting number by 1000, to get all the numbers in Megabytes. If the value didn't end in M or k, I added NaN. From there, I replaced all the NaN with the median of the set, since the distribution was very skewed.

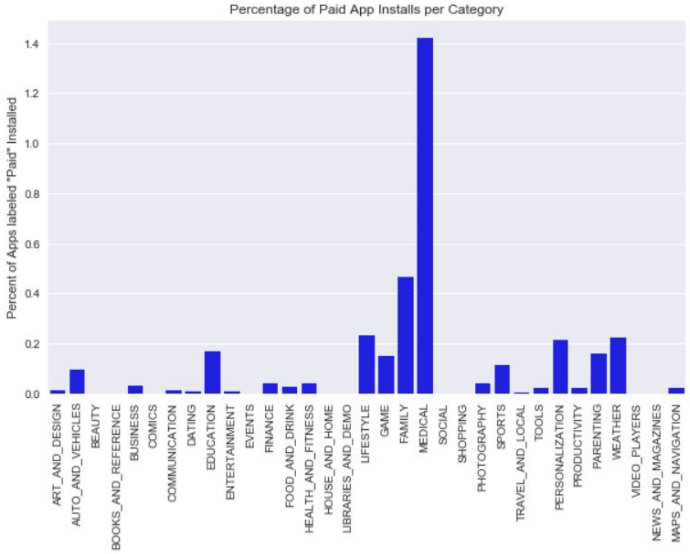
In the second dataset, I found the provider of the data to be very misleading. Almost 23,000 of the 60,000 rows were filled with null values, besides the name of the app. I judged these rows useless and dropped them. There were also a few more null values in the "review" column, and judging by the fact that this was the only useful column to me, I dropped it.

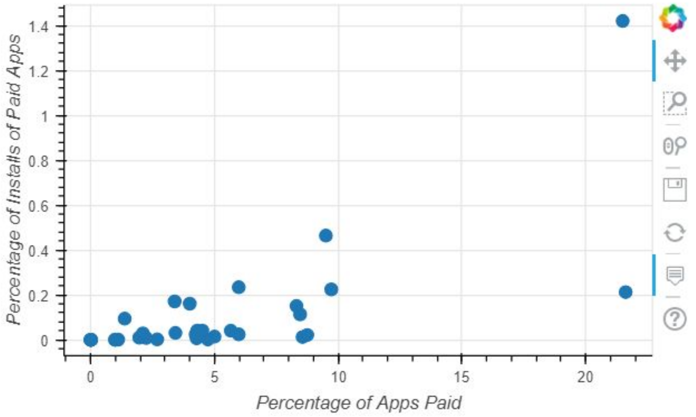
III. Exploratory Data Analysis

Here is where I really delved into the data to see if there are any patterns. I decided to spend a sizable amount of time in the EDA section, because there was just so much that I could do with this dataset. The first thing that I wanted to do was to see which category of apps were the most likely to be paid for. From that analysis, I yielded this barplot. You can clearly see that **medical** apps and **personalization** apps had the highest percentage of paids apps, at over 20% of all apps in that category.

Just because there are many paid apps in a category does not mean that people are downloading said apps. So as a follow up to the previous analysis, I made another barplot that told me what percentage of all installs in each category were paid for. I yielded the figure to the right. Paid medical apps accounted for 1.4% of all installs of medical apps, and personalization apps did nothing very special in this category. However, **Family** apps were shown to have a fairly high percentage of paid installs. I then followed up even further and made a scatter plot of the data to find that a higher percentage of apps that are paid for in a certain category does result in a higher percentage of installs of paid apps in said category. Here is the scatter plot. As you can

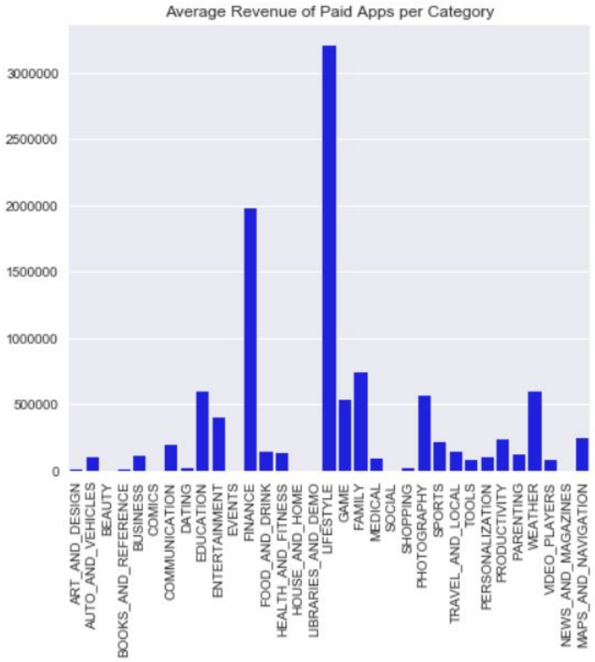


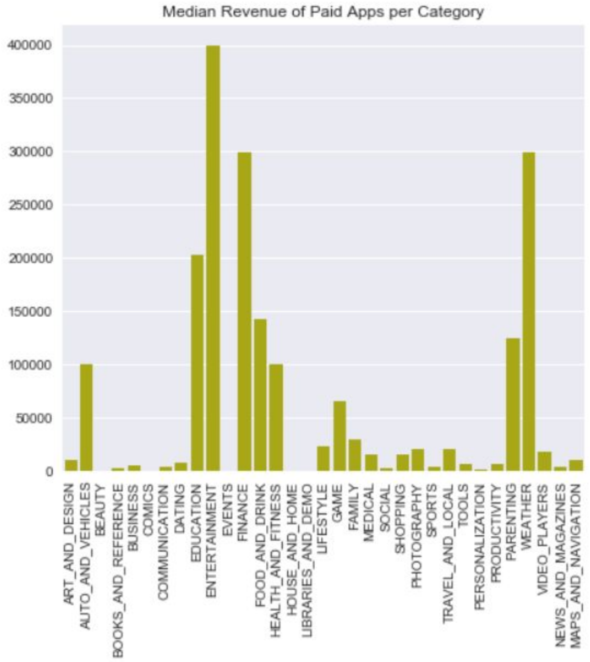




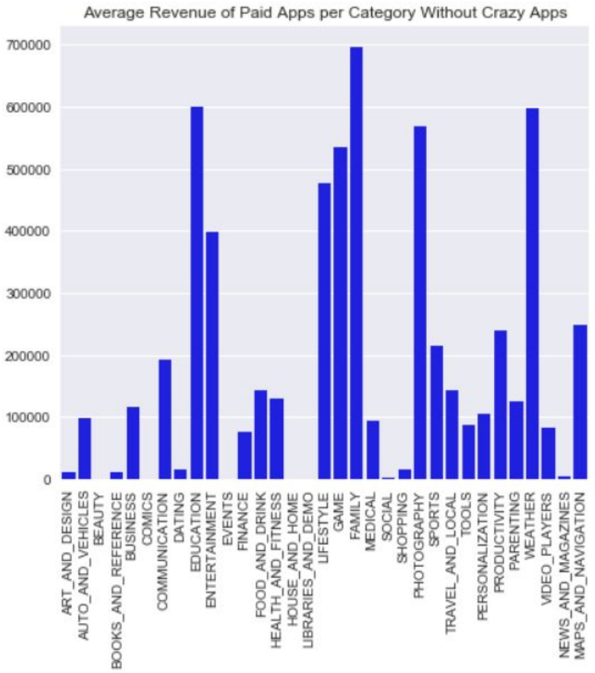
see, there is fairly clearly a positive correlation between the two. This correlation is actually 0.5.

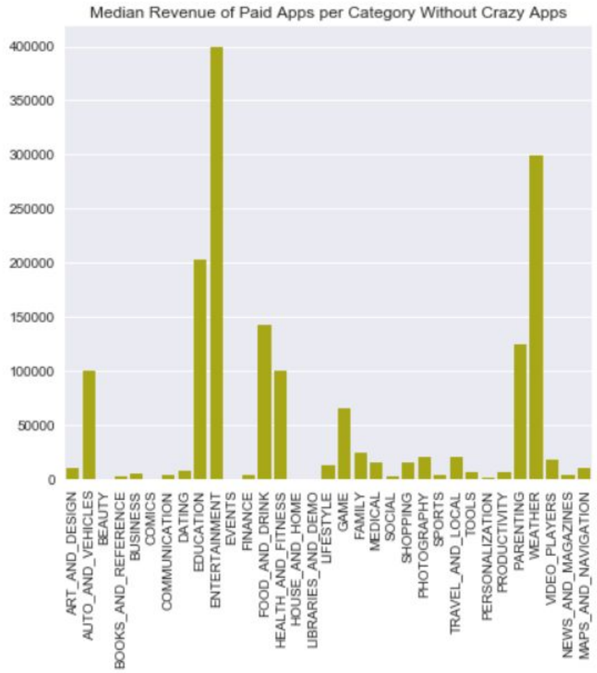
My next goal was to find the average revenue of apps in different categories. To do this, I needed to create a revenue column, and to do so, I multiplied the price by the number of installs. Again, I must reiterate that the number of installs is not a very accurate measure. From this point, I took the average and median revenue of each category, as a few very highly downloaded apps in a category might cause the average to shift significantly.





As you can see, the **finance** and **lifestyle** apps have very high average revenues, but do not rank very highly in the median revenue. I looked into the finance and lifestyle categories and found that there were several apps that cost a whopping $300, and a few of them actually got many downloads, the most being 100,000. So, I removed them and got the graphs below. Now, as you can see, those two categories do not rank highly on either of the plots. The main thing to draw from this is that the **entertainment** category ranks very highly in both the average and median revenue, which means it may be a good option for developers to make if they are planning on making a lot of money. Although family has a high average, we can look into the median revenue plot and see that it is actually quite low. Looking further into it, we see that there are 1 or 2 apps, namely Minecraft, that has over 10,000,000+ downloads and costs $7, so evidently, that was enough to cause a major change in the plot below.





So far, the two revelations we have found are that **Entertainment apps could possibly have the highest the revenue**, and that **medical apps are the most likely to be paid for.**

The next thing I wanted to do was to see was if there were any correlations in the data. The first thing I did was check whether or not there was a correlation between the rating of an app and the number of installs, and based on the correlation coefficient, there wasn't. So a higher rated app does not necessarily mean it will be more successful. I then performed natural language processing on the reviews and obtained a label of positive or negative sentiment. I then took the average of the reviews (so the percentage of reviews that were positive). What I found was that most apps fell into 80% positively rated and then I checked for correlations to the data. I did not find any. I did find correlations between the number of installs and the revenue and the number of reviews and the revenue. But this might just be because these are all number based.

IV: Inferential Statistics

I then performed hypothesis testing on these findings. My results are as follows:

1. There are significantly less paid apps than there are free apps at a confidence level of 0.05.
2. Paid medical apps are installed at a higher rate than other paid apps at a confidence level of 0.05.
3. Apps in the family category have a higher mean revenue than apps not in that category at a confidence level of 0.05.
4. Communication apps have the highest install rate, but most of them are from Google.
5. Health and Fitness apps do not have the highest median number of installs of paid apps with more than 10 in the category.
6. There is no significant correlation present between the average positive rating and number of installs.
7. There is a significant correlation between the number of installs and revenue and the number of reviews and the revenue.
8. There is only 1 principal component.