Alvaro Hu

Springboard

December 6, 2018

Determining Successful Categories of Google Play Store Apps

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.

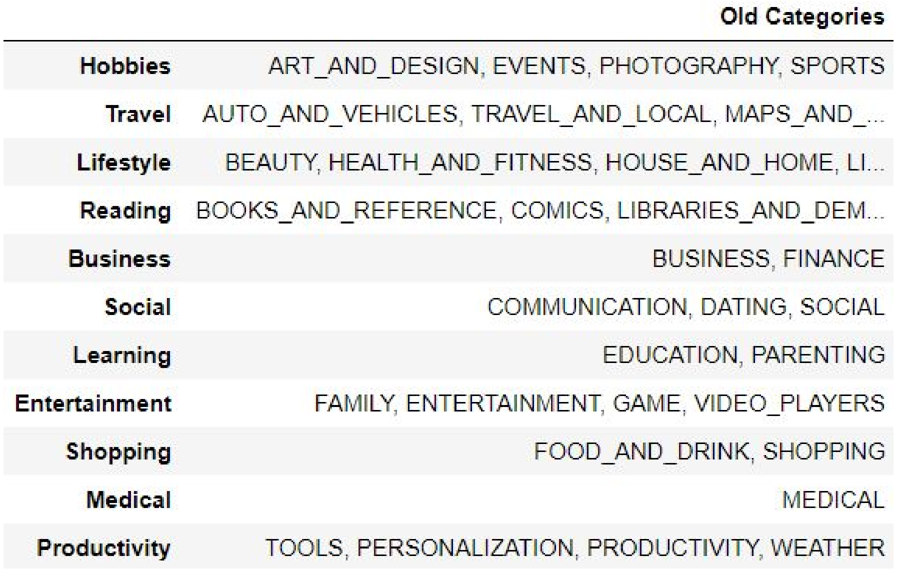
My client is every single developer out there looking for what their next project should be focused on, depending on if they want recognition (downloads) or money (revenue). By providing an in-depth analysis of what genres and types of apps give the most downloads in general, then it would alleviate 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.

To deal with the 33 different categories, I consolidated some of them together. Here is a list of the categories that I put together and the overarching category that it fell under:



	Old Categories
Hobbies	ART_AND_DESIGN, EVENTS, PHOTOGRAPHY, SPORTS
Travel	AUTO_AND_VEHICLES, TRAVEL_AND_LOCAL, MAPS_AND_...
Lifestyle	BEAUTY, HEALTH_AND_FITNESS, HOUSE_AND_HOME, LI...
Reading	BOOKS_AND_REFERENCE, COMICS, LIBRARIES_AND_DEM...
Business	BUSINESS, FINANCE
Social	COMMUNICATION, DATING, SOCIAL
Learning	EDUCATION, PARENTING
Entertainment	FAMILY, ENTERTAINMENT, GAME, VIDEO_PLAYERS
Shopping	FOOD_AND_DRINK, SHOPPING
Medical	MEDICAL
Productivity	TOOLS, PERSONALIZATION, PRODUCTIVITY, WEATHER

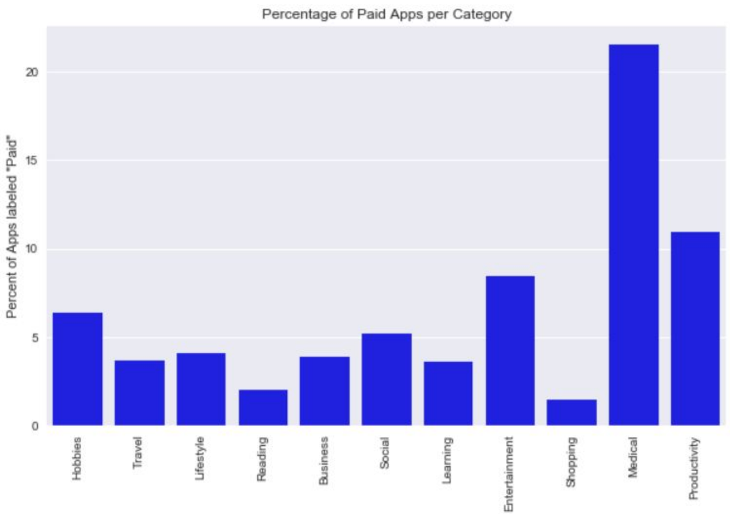
Some examples from each of the 11 categories are as follows:

- Social: WhatsApp, Gmail, Messenger
- Entertainment: Netflix, YouTube, Games
- Medical: Essential Anatomy, VeinSeek
- Learning: Duolingo, TED, Khan Academy
- Hobbies: Sketch, Photo Editor
- Travel: Used Cars, CityBus, Lyft
- Lifestyle: Beauty Selfie Cam, Beauty Tips
- Reading: Kindle, Wikipedia, Ebook Reader
- Business: Indeed, OfficeSuite, FB Ads Manager
- Shopping: Amazon, McDonald’s
- Productivity: Google, Calculator

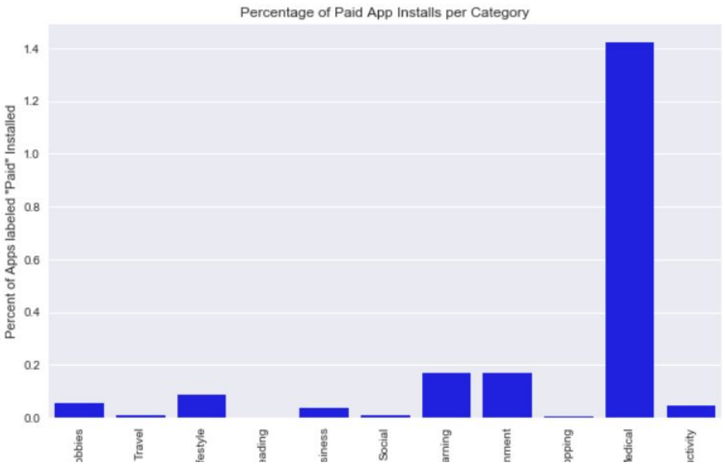
In the second dataset, I found the provider of the data to be very misleading. Almost 23,000 of the 60,000 rows were filled with null values, besides the name of the app. I judged these rows useless and dropped them. There were also a few more null values in the “review” column, and judging by the fact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

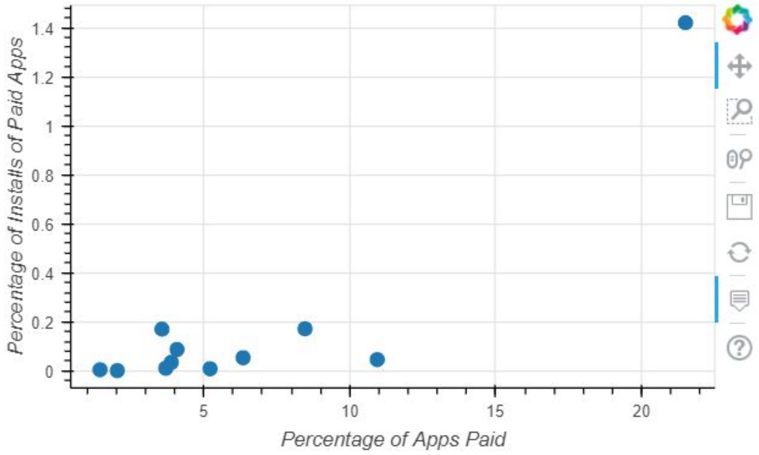
III. Exploratory Data Analysis

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 I needed to do was consolidate some of the categories. There were 33 categories to begin with, and I would preferably get that number to around 10, especially with many of the categories being fairly similar. 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



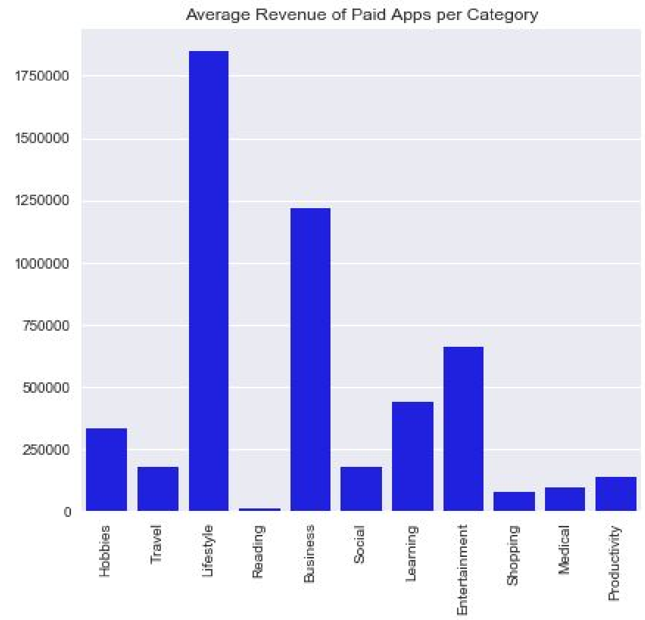
Productivity apps had the highest percentage of paids apps, at over 20% of all apps in that category.

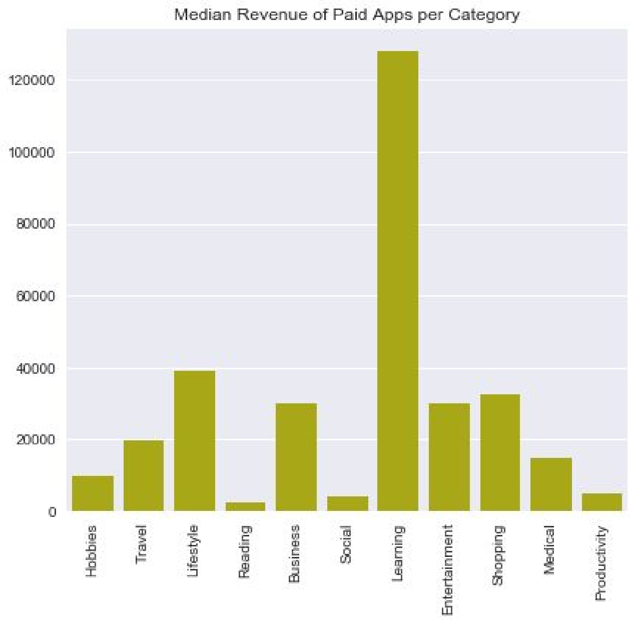




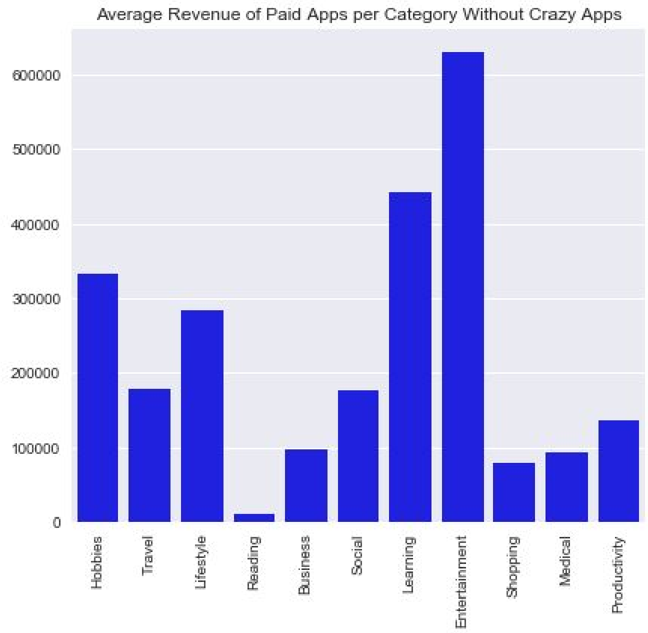
Just because there are many paid apps in a category does not mean that people are downloading said apps. So as a follow up to the previous analysis, I made another barplot that told me what percentage of all installs in each category were paid for. I yielded the figure to the right. Paid medical apps accounted for 1.4% of all installs of medical apps, and personalization apps did nothing very special in this category. 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 0.78.

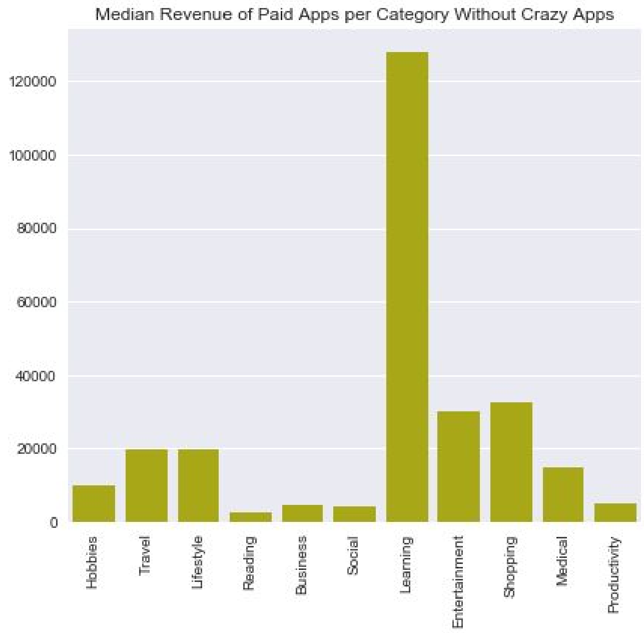
My next goal was to find the average revenue of apps in different categories. To do this, I needed to create a revenue column, and to do so, I multiplied the price by the number of installs. Again, I must reiterate that the number of installs is not a very accurate measure. From this point, I took the 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 **business** and **lifestyle** apps have very high average revenues, but do not rank very highly in the median revenue. I looked into the business 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. Now, you can think for yourself, but I personally believe that it would be quite hard to find 100,000+ people willing to spend $300 on a useless app, so what is probably going on here is that the app was on sale, where a majority of the installs might come from. I cannot really trust these installs, so I will drop them. Now, as you can see, those two categories do not rank highly on either of the plots. The main thing to draw from





this stage of analysis is that the **entertainment and learning** categories 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. 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 and paid app installs than there are free apps at a confidence level of 0.05.
    a. What I think this means: People are more willing to install free apps.
2. Paid **medical** apps are installed at a higher rate than other paid apps at a confidence level of 0.05.
    a. People are willing to pay for apps if they are for work, even if they have a higher average cost.

3. Apps in the **entertainment** category could not be said to have a significantly higher average revenue than other apps, but the p-value was 0.057, so it was very close to the alpha level.
    a. It is likely that people will buy paid video games, but a few very successful games could be swaying this a lot.
4. **Social** apps have the highest install rate, but most of them are from Google.
    a. Most of the installs come from free apps, and most of them come from 1st party developers, which leads me to believe that this isn't a great place for developers to enter.
5. There is no significant correlation present between the rating of an app and number of installs.
    a. An app can be very poorly rated, but that is not that determining factor for its success.
6. There is a significant correlation between the number of installs and revenue and the number of reviews and the revenue.
    a. If an app is getting reviewed a lot, then it has a higher chance of being successful when it comes to paid installs. This must mean it is polarizing enough either way that it gets bought.

V: Machine Learning

In this section, I used natural language processing to try and guess the sentiment of the review from the second data set. There were several steps to complete this. The first thing that I did was to hand label a few of the reviews, so that the algorithm has a training set to work with. Since there were over 30,000 reviews, I had to do quite a few labels; 700 to be exact. After labelling the data, I converted all of the reviews to a format that can be used by a machine learning algorithm. To do this, I used the CountVectorizer() to transform all of the words in each word into a feature that is either 1, if the word is in the review, or 0, if it is not. I then used the Counter from the collections package to count up the number of times that each word shows up, and used a cutoff for the amount of times that the word had to show up in order to use it in the training set. For this cut off I used 5, since it was the elbow in the graph plotting the number of documents and the percentage of words that show up on said documents.

After having all of the words in a matrix, I used several machine learning algorithms to take the training set and predict on the rest of the reviews. The 5 algorithms I used were MultinomialNB, BernoulliNB, Logistic Regression, GaussianNB, and an SGD Classifier. These 5 classifiers had varying degrees of success. I then used a voting classifier to take the results of all 5 of these algorithms and use a majority vote to determine the label of the review. Below is a table of the precision, accuracy, and recall for all 6 algorithms.

Algorithm	MultinomialNB	BernoulliNB	Logistic Regression	GaussianNB	SGD Classifier	Voting Classifier
Precision	0.856	0.854	0.847	0.88	0.866	0.863
Accuracy	0.814	0.79	0.824	0.614	0.776	0.859
Recall	0.911	0.899	0.953	0.604	0.858	0.851

As you can see, the accuracy, precision, and recall are all highest in the voting classifier. After doing this, I applied the algorithm to the rest of the reviews.

Some findings I found from this data are that it is most likely that an app will be rated positively 80% of the time. This makes sense in the context that the average rating was about a 4.0 out of 5.0. Another thing I found was that the correlation between the percentage of positive ratings and the number of installs is virtually nonexistent, coming in at and R^2 of 0.003. There were also only 10 apps in the reviews dataset that had a matching app in the app dataset that were paid, so I was unable to determine correlation between the positive rating rate and the revenue of an app. But from the few apps provided, it did seem as though there were no correlation, with an R^2 of 0.0001.

VI. Conclusion

To conclude, I would like to state some of my findings and the recommendations that I would make to me clients. The first conclusion is that social apps get the highest average number of installs, and it isn't even close. After looking more into the apps in this dataset, I found that most of these apps are made by huge billion dollar companies like Facebook and Google, and I don't think that if you wanted to break into the app space, you would make an app for socializing to do so.

The next conclusion is that doctors and people in the medical world are very willing to purchase apps to help them with their jobs, even if they have to pay for such apps. Medical apps have both the highest percentage of installs that are paid for and the highest average price per app. I also personally know a start up that works provides an app that allows doctors to communicate directly with their patients through a web interface. If the client has a good idea for a medical app, I would recommend that they try it out.

Entertainment apps had the highest average number of installs for paid apps. This was heavily swayed by an extremely successful app (Minecraft). Learning apps, however, had the highest median number of installs. In the same vein as in medical apps, I would say if you had a good idea for an educational app, then go for it.

Plus, if worst comes to worst, I found with statistical significance that there was no correlation between the number of installs of an app and the average rating. So give the app a shot and see how it goes!