

The background of the slide features a stylized world map in shades of blue. Overlaid on the map is a network of white lines connecting various points, resembling a global data or communication network. The map is centered, with the Americas on the left and Europe/Africa on the right.

AN IN-DEPTH ANALYSIS OF GOOGLE PLAY STORE DATA

ALVARO HU

THE PURPOSE

- The objective of this project is to give an informed and statistically significant opinion to mobile developers
 - What project should they work on next?
 - Which will give me the most downloads?
 - Which will give me the most money?



OR



THE DATA

- Provided in two CSV files on Kaggle.com
 - 1st: 10,000 Apps with the name, rating, category/genre, file size, Android version, when it was last updated, and number of Installs and Reviews
 - 2nd: 60,000 separate reviews for about 900 different apps in the 1st dataset. Comes with preprocessed sentiment



WRANGLING STEPS

- Number of installs and file size
 - Getting rid of “+”, “M”, and “k”
 - Divide by 1,000 for k

100,000+



100000

19M

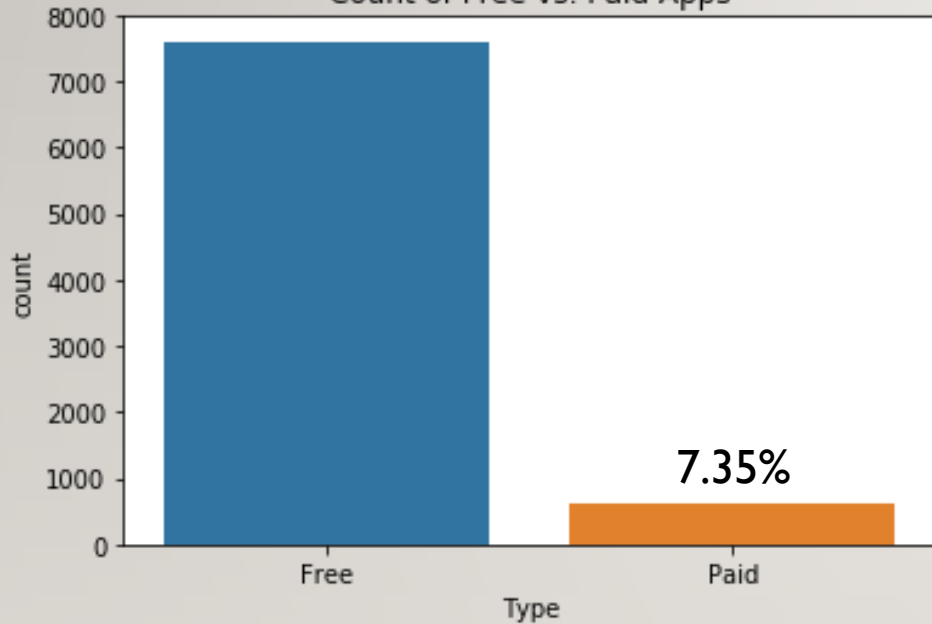


19.0

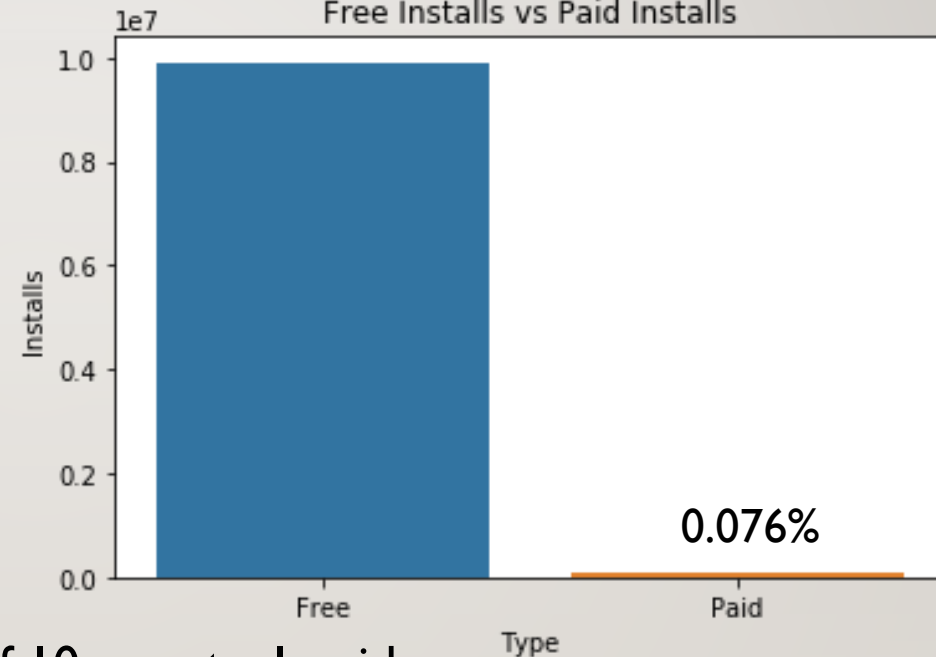
- Fill “Varies with Device” with Median
- All numbers are Integers and not Strings
- Many rows in 2nd dataset full of NaNs. Lost 23,000 observations

FREE VS PAID APPS

Count of Free vs. Paid Apps



Free Installs vs Paid Installs



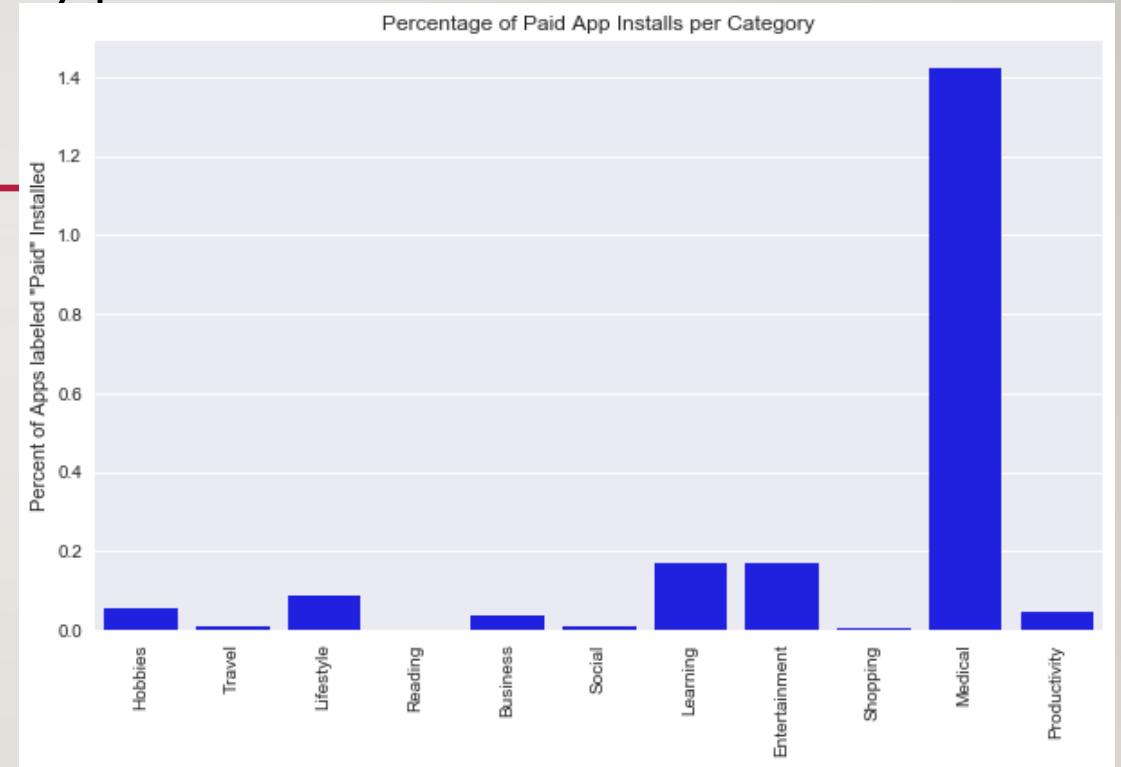
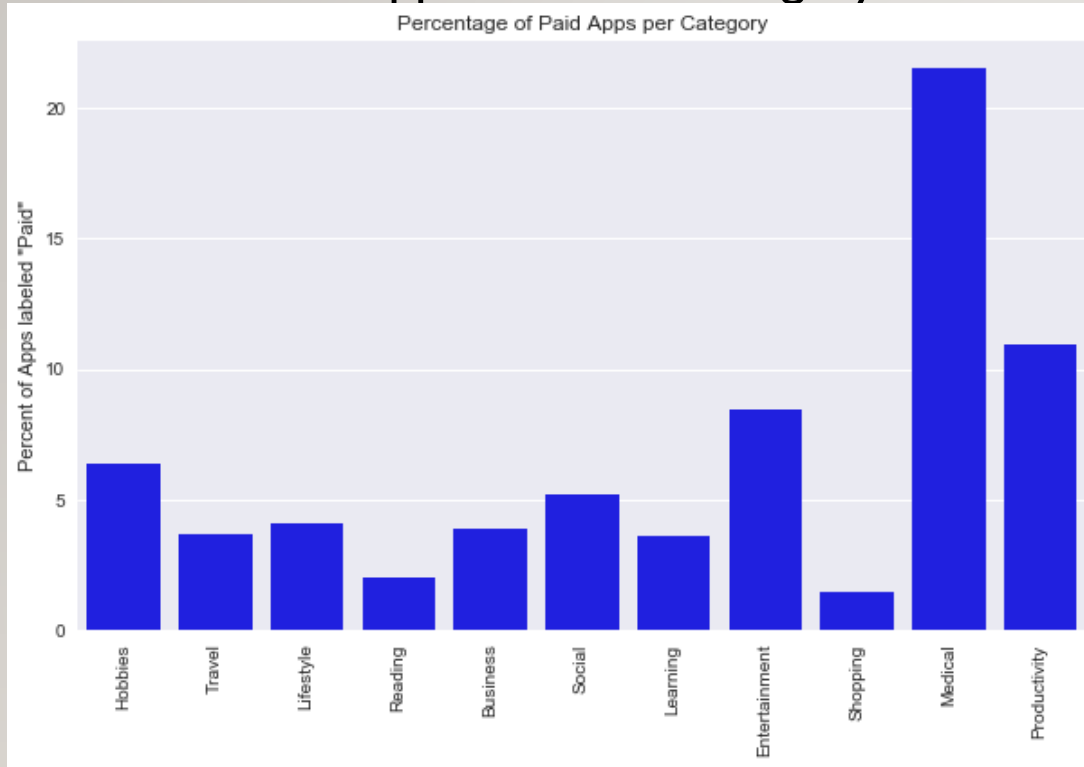
- Free apps outweigh paid apps at a rate of 10 apps to 1 paid app
- But are **installed** at a rate of 100 to 1
 - Metrics for success for installs of paid apps is lower

CATEGORY EXAMPLES

- 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

ANALYZING THE DATA

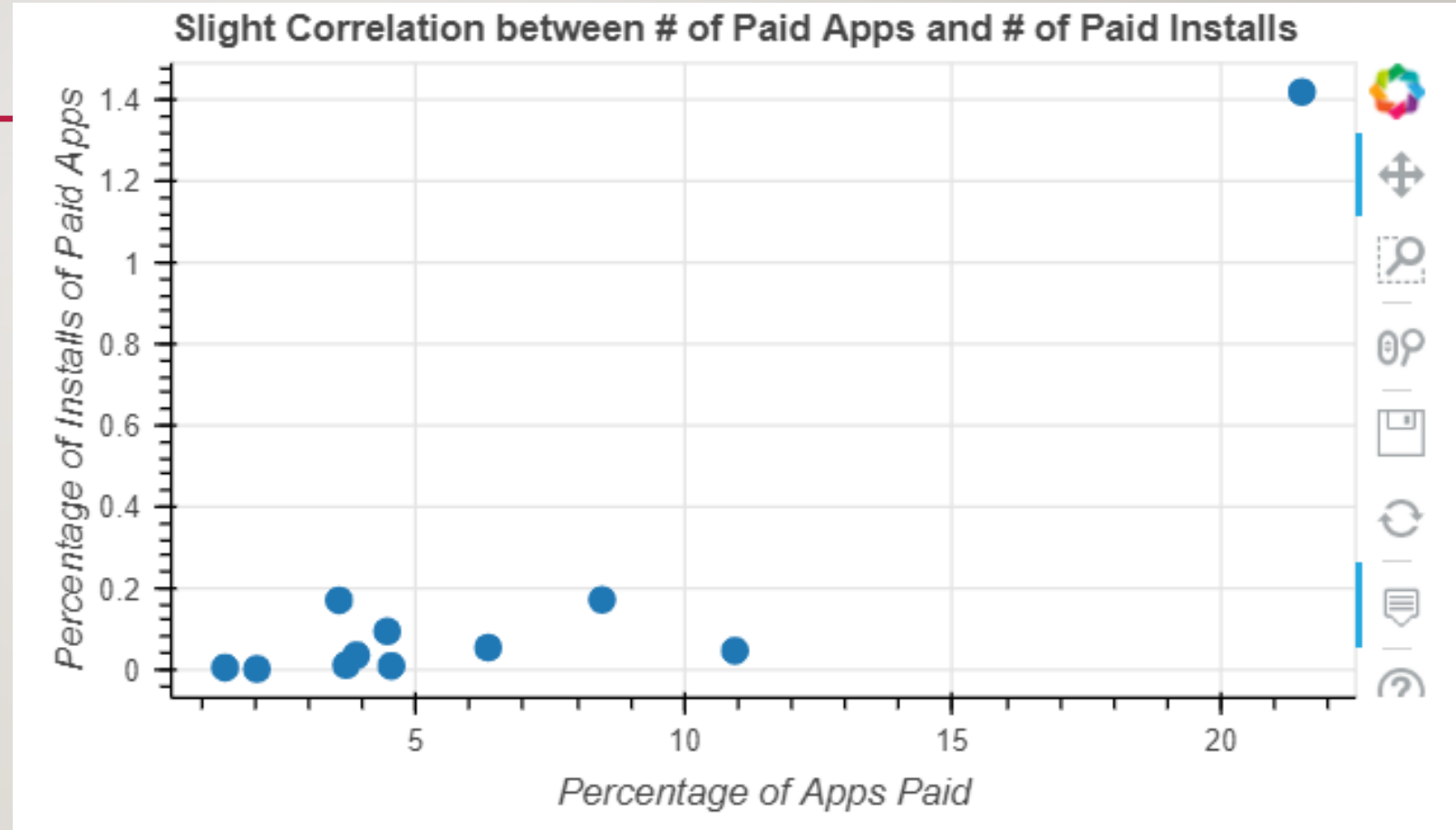
- Which apps, based on category, are most likely paid for?



- Medical Apps both have the highest percentage of paid apps **and** have the highest percentage of installs of said apps.
- Medical apps tend to be more for necessity, so people are willing to pay for them

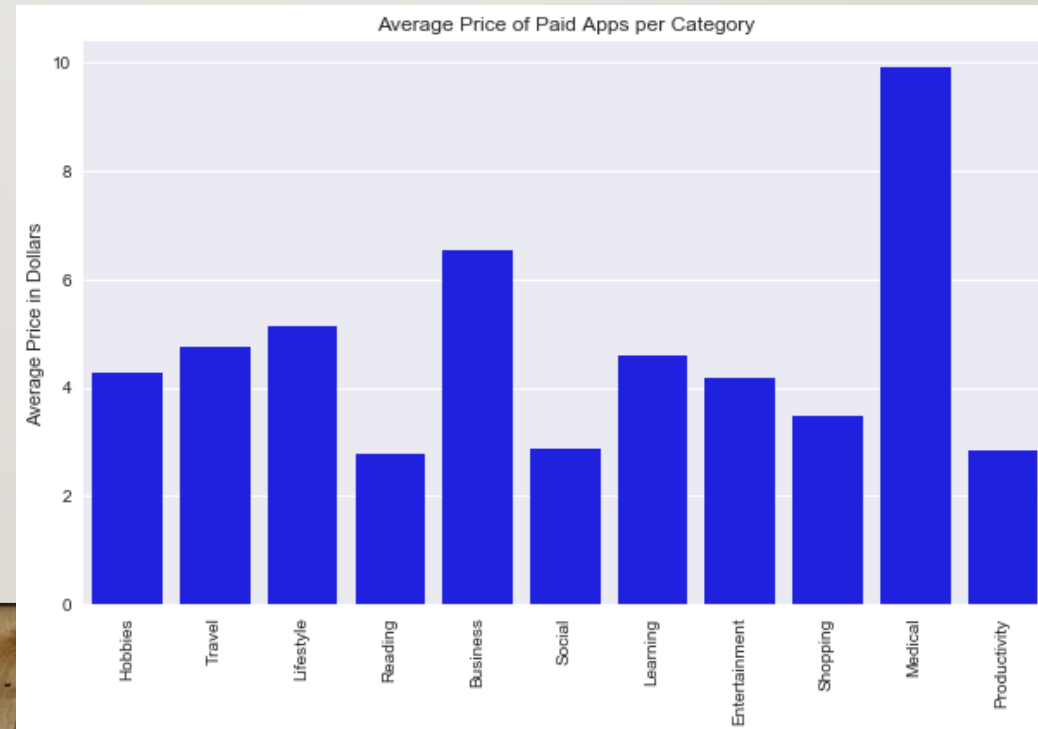
R^2 OF 0.789 FOR THE II CATEGORIES

- People generally put apps behind a paywall if they know there is a higher chance that they will be installed
 - Some apps intrinsically work better to pay for them immediately rather than in-app



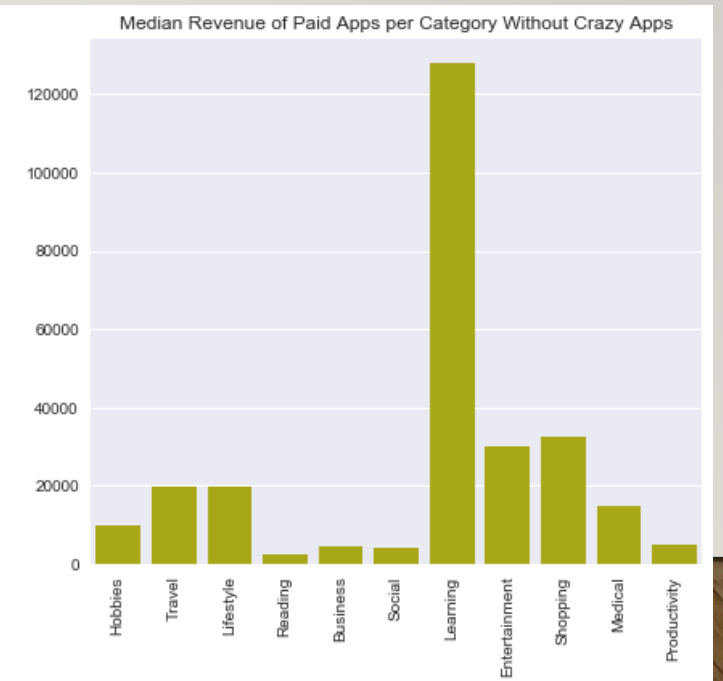
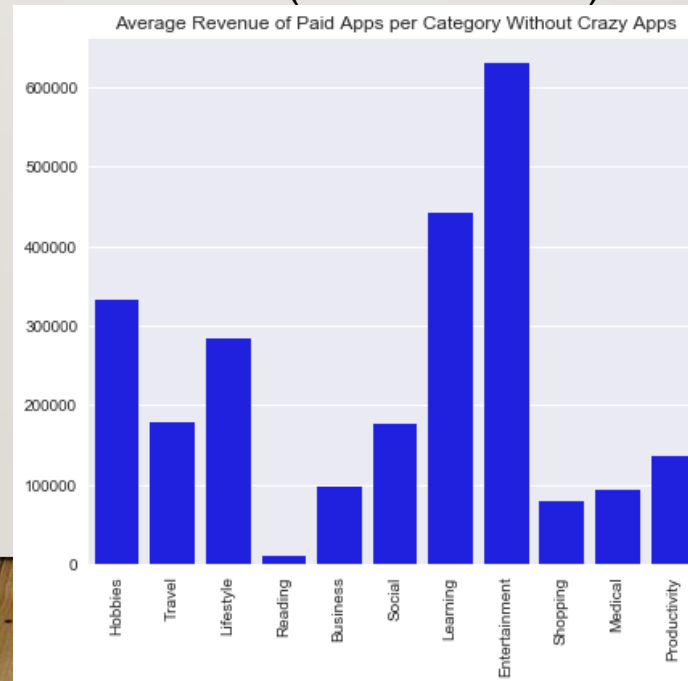
AVERAGE PRICE OF PAID APPS

- Business and Lifestyle categories had unusually high averages
 - Many “I am Rich” apps that cost \$200-\$400.
 - These most likely were not purchased at full price, so I dropped them for my analysis.
- After dropping these apps:
 - Medical apps have highest average price and installed most frequently
 - Business apps are 2nd at \$6.50 per app



REVENUE IS A DIFFERENT STORY

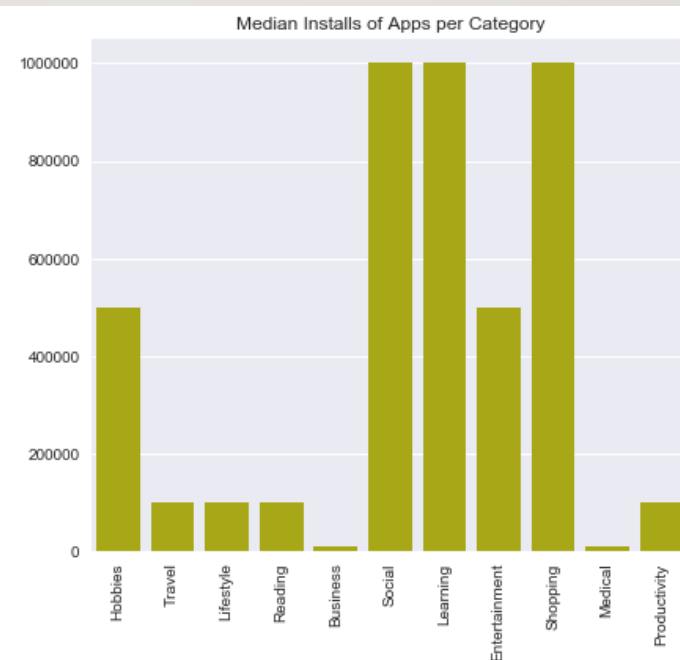
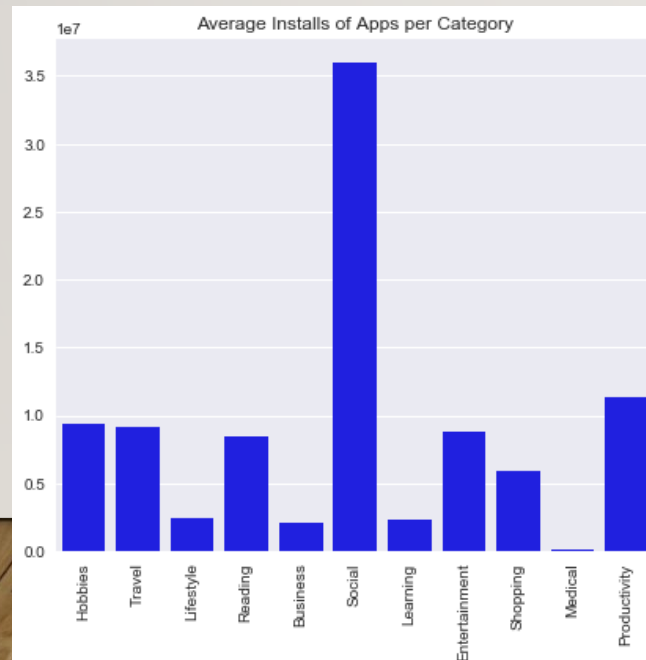
- Obtained revenue field by multiplying Installs by price
 - Two assumptions:
 - All apps purchased at full price
 - Number of installs closer to label than not (because of the +)
- Learning Apps rank highly in both average and median
 - Looking closer:
there are only 6 paid apps
but they all did fairly well



LEARNING APPS STILL RANKS HIGHLY FOR MEDIAN NUMBER OF INSTALLS OF **ALL** APPS

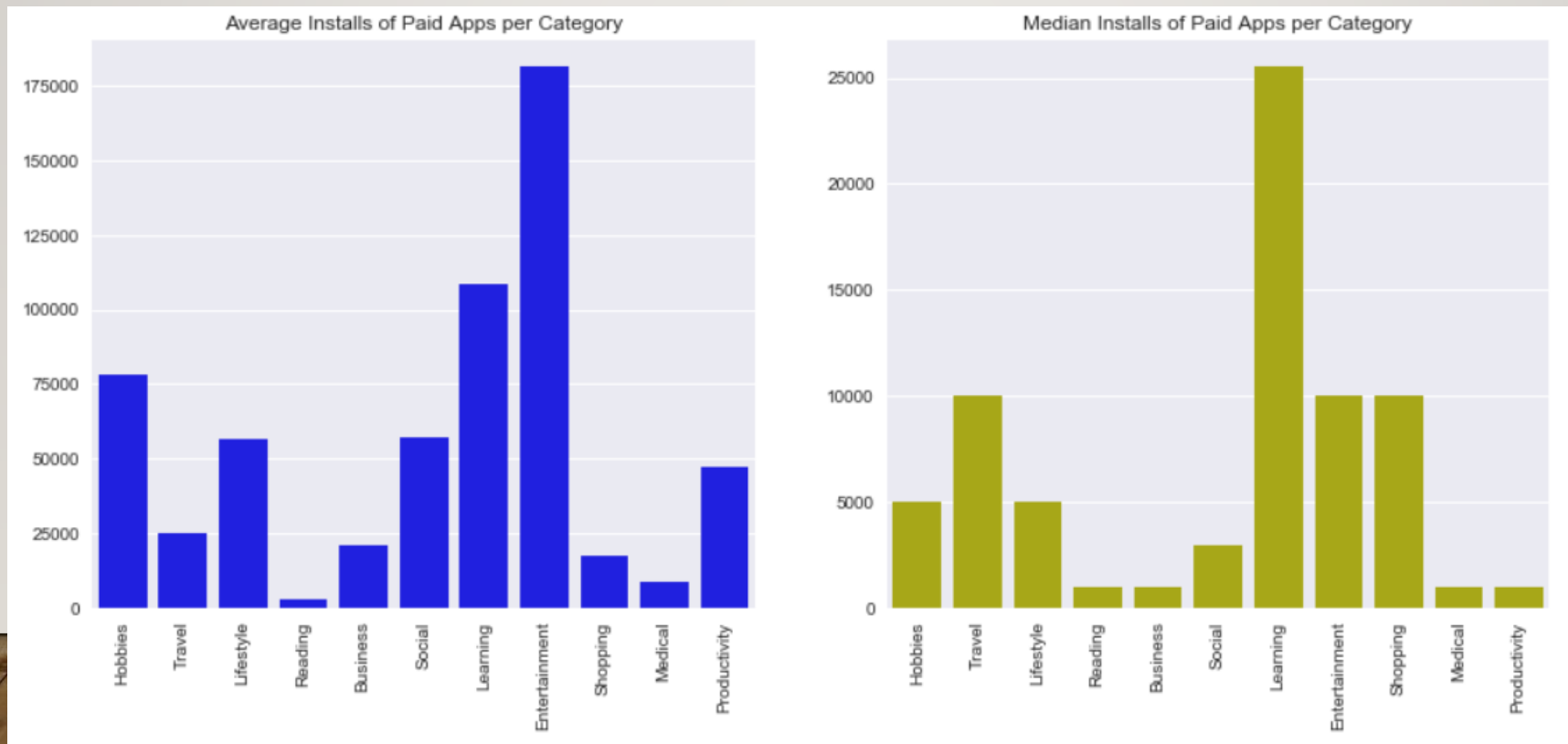
- Some sort of educational app may be good for developers to make if they want it to be downloaded a lot.
- Social apps seem like they could be a good choice, but looking at the top 10 apps in this category, we see that they are all by large companies like Google and Facebook

App	Rank
Messenger – Text and Video Chat for Free	1
WhatsApp Messenger	2
Skype - free IM & video calls	3
Google Chrome: Fast & Secure	4
Gmail	5
Hangouts	6
Google+	7
Instagram	8
Facebook	9
UC Browser - Fast Download Private & Secure	10



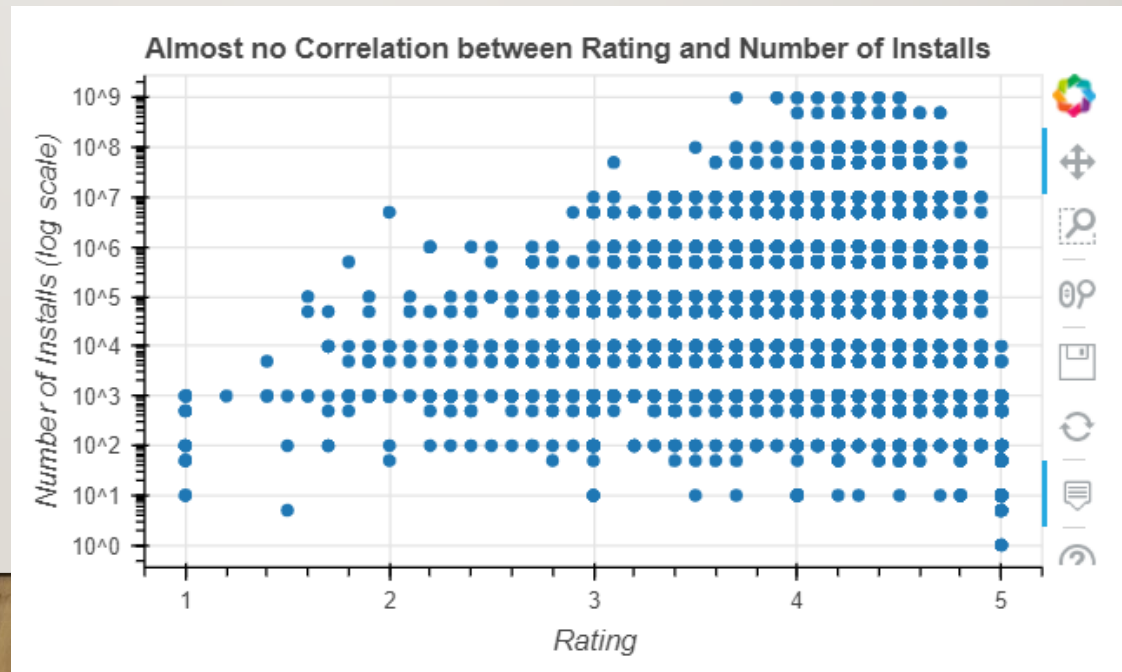
INSTALLS OF PAID APPS

- Social apps gives up top spot to Entertainment for average number of installs
 - Tells us that social app installs are from free apps.



DO RATINGS AFFECT THE INSTALLATION RATE?

- Using the Rating column provided by Play Store
- Plotted on log scale
 - Seems like there is a correlation, but low number of apps with high installs makes it look this way
 - Actual R^2 is only 0.007



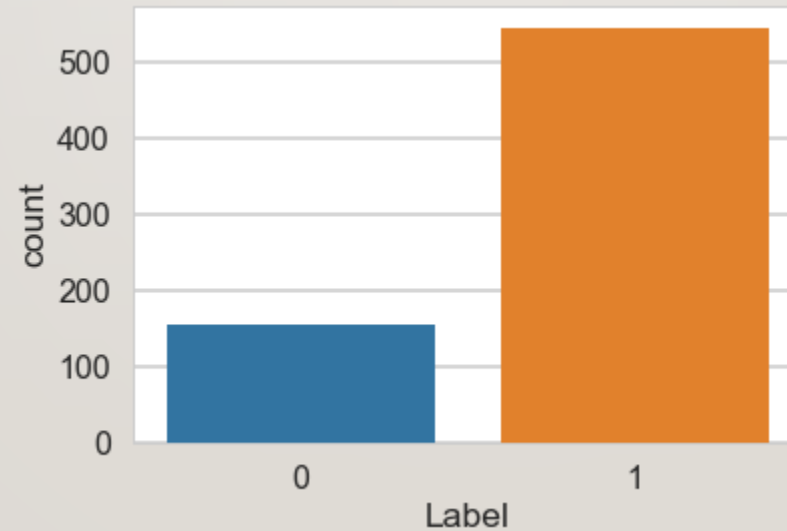
ANY CORRELATIONS OF MY CALCULATED FIELDS?

	Paid Install Percent	Size	Percent of Apps Paid	Number of Installs	Median Installs	Average Revenue	Median Revenue
Paid Install Percent	1.000000	-0.031083	0.882214	-0.291257	-0.309394	-0.234666	-0.089364
Size	-0.031083	1.000000	-0.009141	-0.054185	0.171366	0.319175	-0.230978
Percent of Apps Paid	0.882214	-0.009141	1.000000	-0.143792	-0.353803	-0.012397	-0.182487
Number of Installs	-0.291257	-0.054185	-0.143792	1.000000	0.434327	-0.042173	-0.302877
Median Installs	-0.309394	0.171366	-0.353803	0.434327	1.000000	0.327441	0.533757
Average Revenue	-0.234666	0.319175	-0.012397	-0.042173	0.327441	1.000000	0.502133
Median Revenue	-0.089364	-0.230978	-0.182487	-0.302877	0.533757	0.502133	1.000000

- 3 that I found interesting
 - Percent of Apps paid vs Paid Install Percent: already discussed
 - Percent of Apps Paid vs Median Installs: The more Paid apps in a category, the less likely they will be installed
 - Median number of Installs vs Median Revenue: The more Installs in general, the more money the apps will make

NATURAL LANGUAGE PROCESSING

- Hand labeled around 700 of the 30,000 reviews as positive or negative
- Used a “voting system” to determine the class of the other 29,300
- More positive than negative in training set, but not necessary to over or under sample



NATURAL LANGUAGE PROCESSING (CONT.)

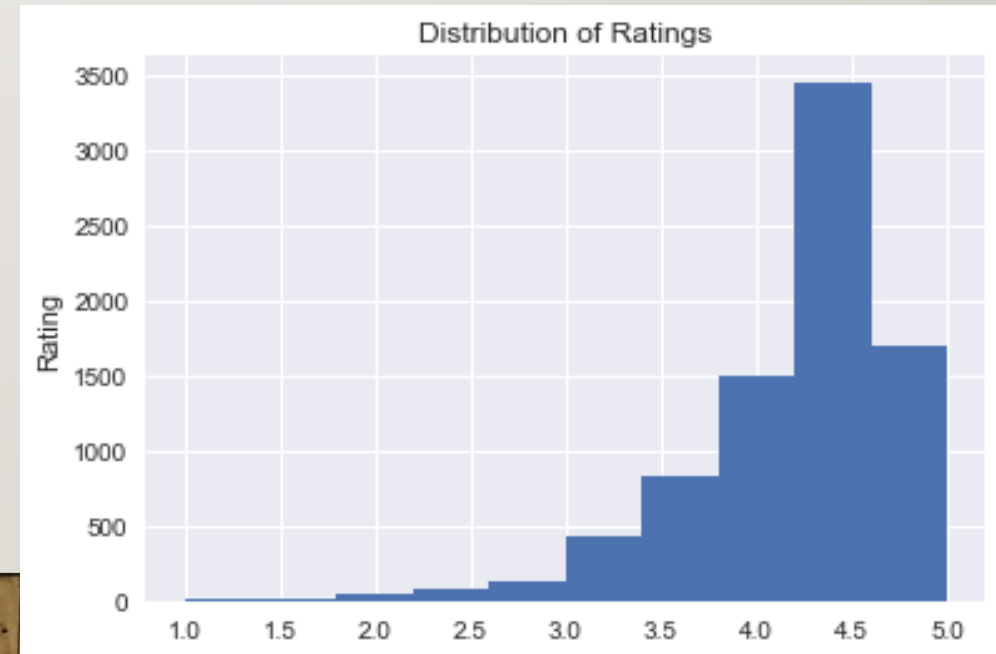
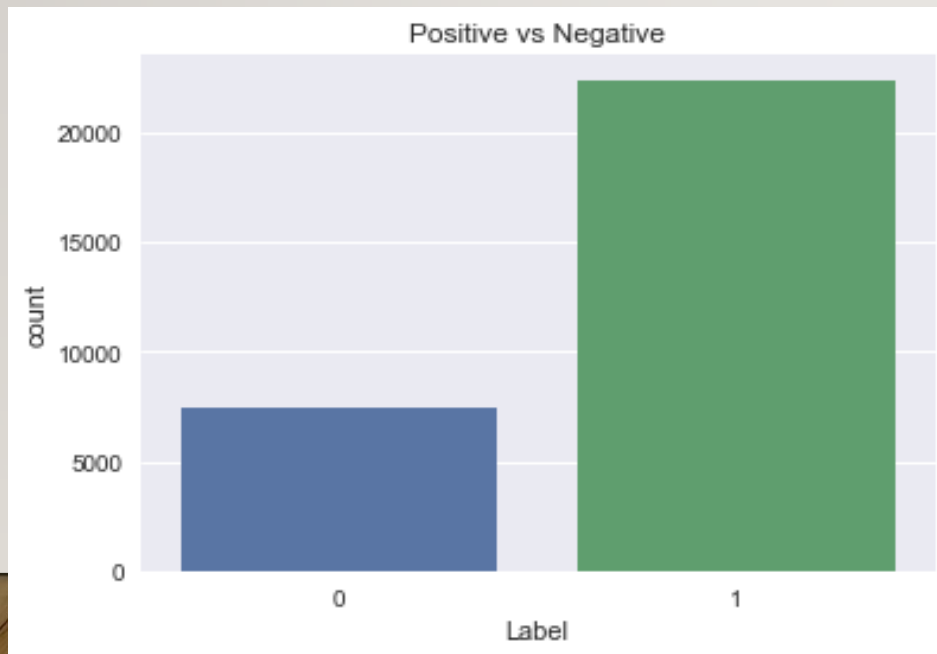
- Used Countvectorizer and a LemmaTokenizer to obtain counts of all lemmatized words in training set
- Used 5 as min_df
- Utilized MultinomialNB, BernoulliNB, GaussianNB, Logistic Regression, and SGD Classifier
 - Used VotingClassifier. If 3/5 models predicted one class, then final prediction is that class
- Accuracy of 0.86
- Precision of 0.86
- Recall of 0.85

These scores are pretty good



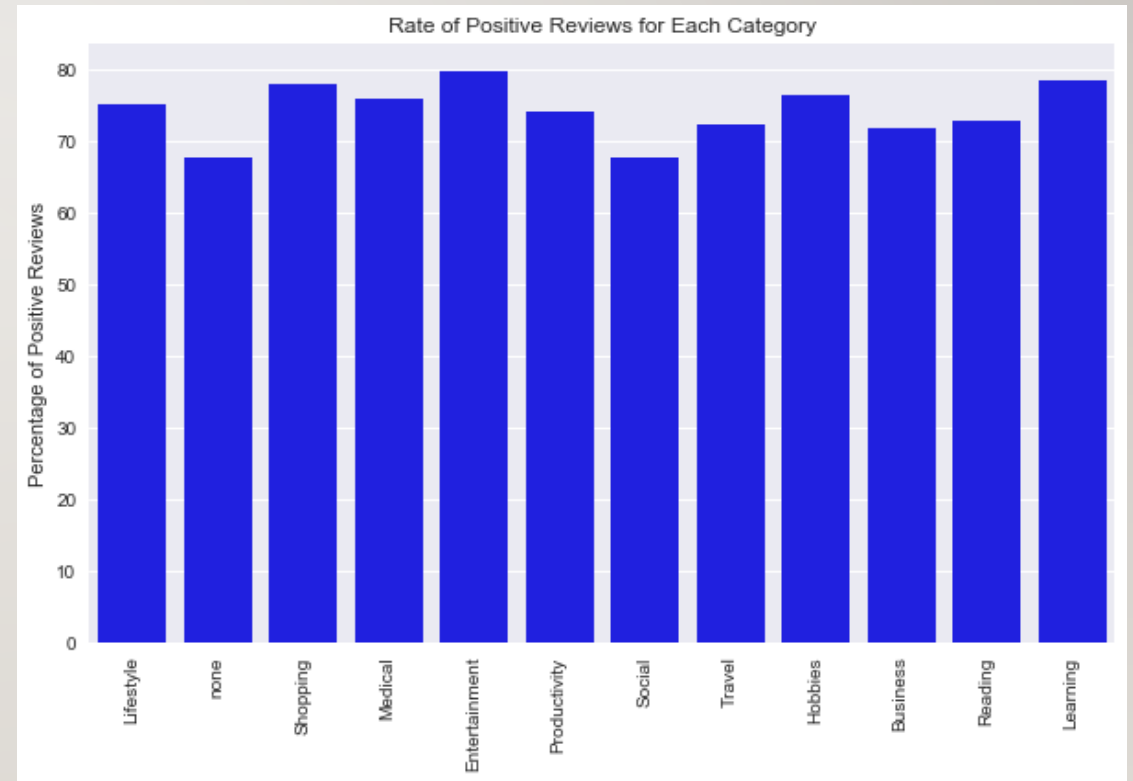
MOST REVIEWS HAD POSITIVE SENTIMENT

- This makes sense, as most apps had a rating of 4.5 or greater



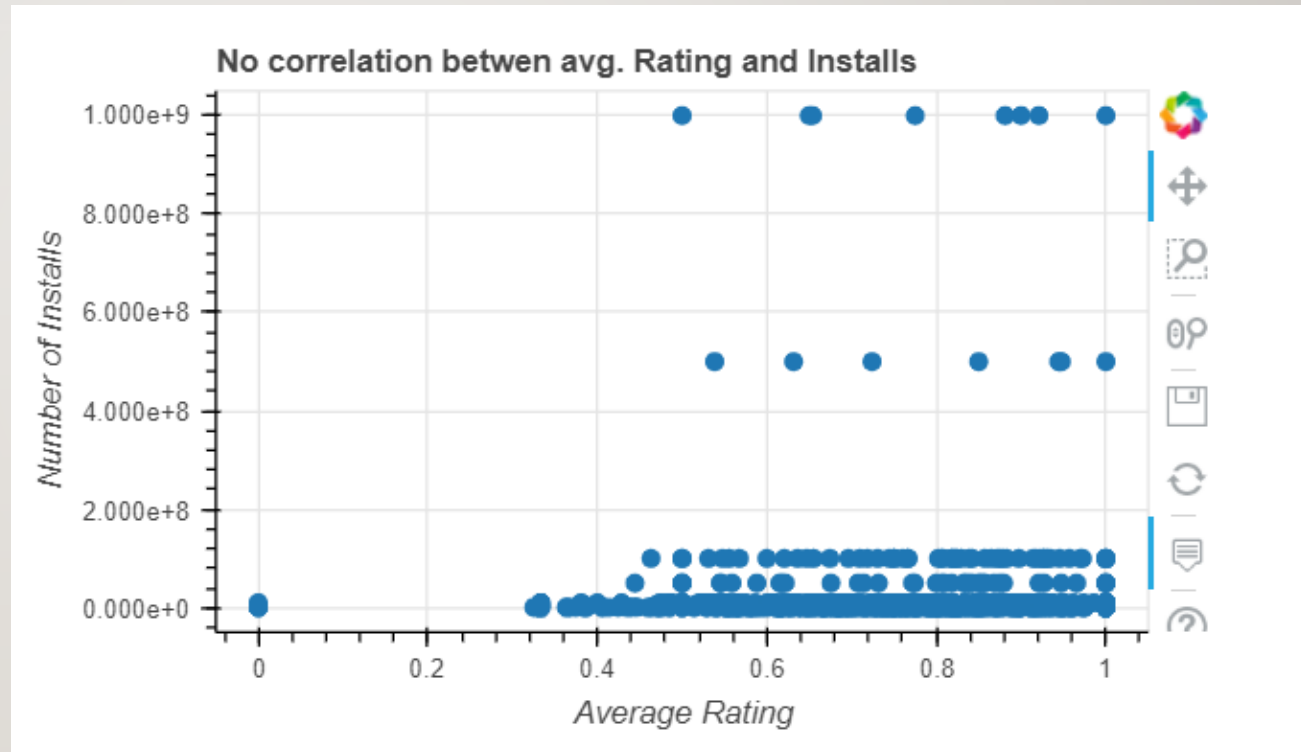
SOCIAL APPS HAVE THE LOWEST PERCENTAGE OF POSITIVE RATINGS

- We have already decided that average rating is not correlated to installation rate
- But is there a correlation with average **percentage of positive ratings**?



USING MY NLP REVIEWS, I CONFIRM THERE IS NO CORRELATION

- $R^2 = 0.003$

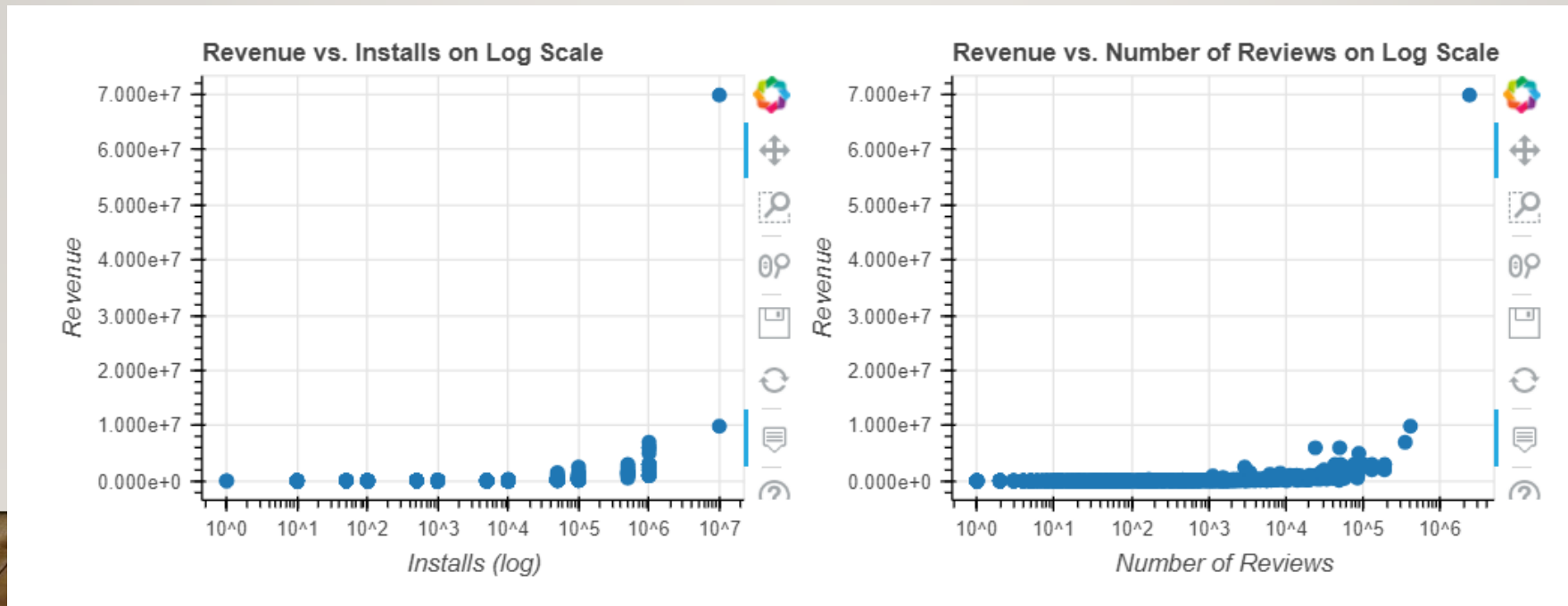


HERE ARE SOME ACTUAL CORRELATIONS

- More installs also means more revenue
- More reviews also means more revenue

$$R^2 = 0.62$$

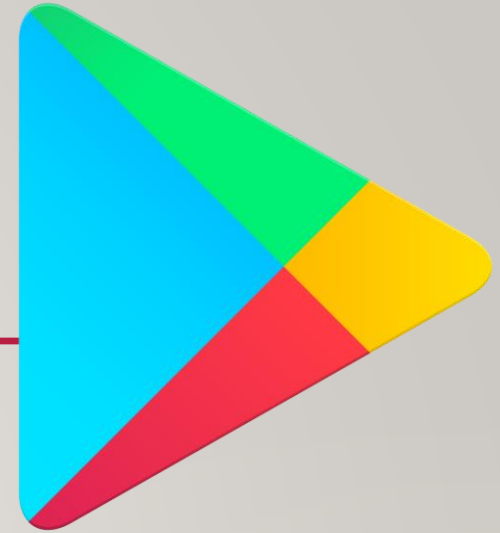
$$R^2 = 0.97!$$



INFERENTIAL STATISTICS

- Hypothesis testing to get statistical significance of findings
 - Paid Medical Apps are installed at a significantly higher rate than non-Medical apps
 - Entertainment Apps may not have a higher average revenue compared to non-Entertainment apps
 - $p = 0.052$, so just above alpha level of 0.05.
 - Still likely significant
 - Social Apps are installed more often than non-social apps, and it's not even close
 - $P = 0.000$
 - Confirmed with statistical significance that no correlation between number of installs and average rating

CONCLUSIONS



- Social apps get the most amount of installs
 - But all made by 1st party developers. Not much room
- Medical apps have highest average cost **and** are installed at highest rate compared to free apps
 - If you have a good idea, you can make good money here
- Entertainment apps average installs of paid apps are highest
- Learning apps have highest median number of installs
 - Could be good for developers
- If your app gets bad reviews, it doesn't necessarily mean it will fail
 - So give it a shot and publish that app!

