Data Exploration of Udemy

Brianna Boston

DISCOVERY STEPS

PROBLEM STATEMENT

METHODOLOGY + FINDINGS

CONCLUSION + RECOMMENDATION

JOB

My name is Brianna Boston and I just got a job at VCU as an entry level Data Scientist and my first research project was given to me regarding VCU and Udemy.

VCU is interested in adding Udemy as a service offered to specific schools like the school of engineering and business. They were wondering if any other schools would also benefit from having access to Udemy. I had mentioned that I used Udemy frequently while studying to become a Data Scientist. VCU thought it would be perfect to include me in this research project as a way to improve and showcase what i've learned throughout my time as an Undergraduate at VCU.

Problem Statement

Goal: The goal is to help VCU determine which Schools would benefit most from Udemy and provide an overview of how much different courses cost within Udemy.

Benefits: Students will have an additional resource to build upon what they are learning in class. Udemy can also be used by professors as a way to have students apply their learning to real world case studies.

This project is important to me since my goal is to grow as a Data Scientist.

1. This project helps me understand where my strengths are and what I need to work on more.
2. I hope this project helps VCU to make informed decision when contracting a subscription with Udemy.

Brianna Boston

Udemy

Facts about Udemy!

- Udemy is a global marketplace to buy and sell videos that are for educational/professional purposes.
- Udemy 49 million users
- 2021 Revenue : $515 million

Why did I chose Udemy?

- I personally use Udemy, and find the platform to be very helpful, and dangerous.
- I have roughly 25 Udemy videos, and have only watched half of them. I have only finished two Udemy courses.
- Two courses in progress of completion

Given

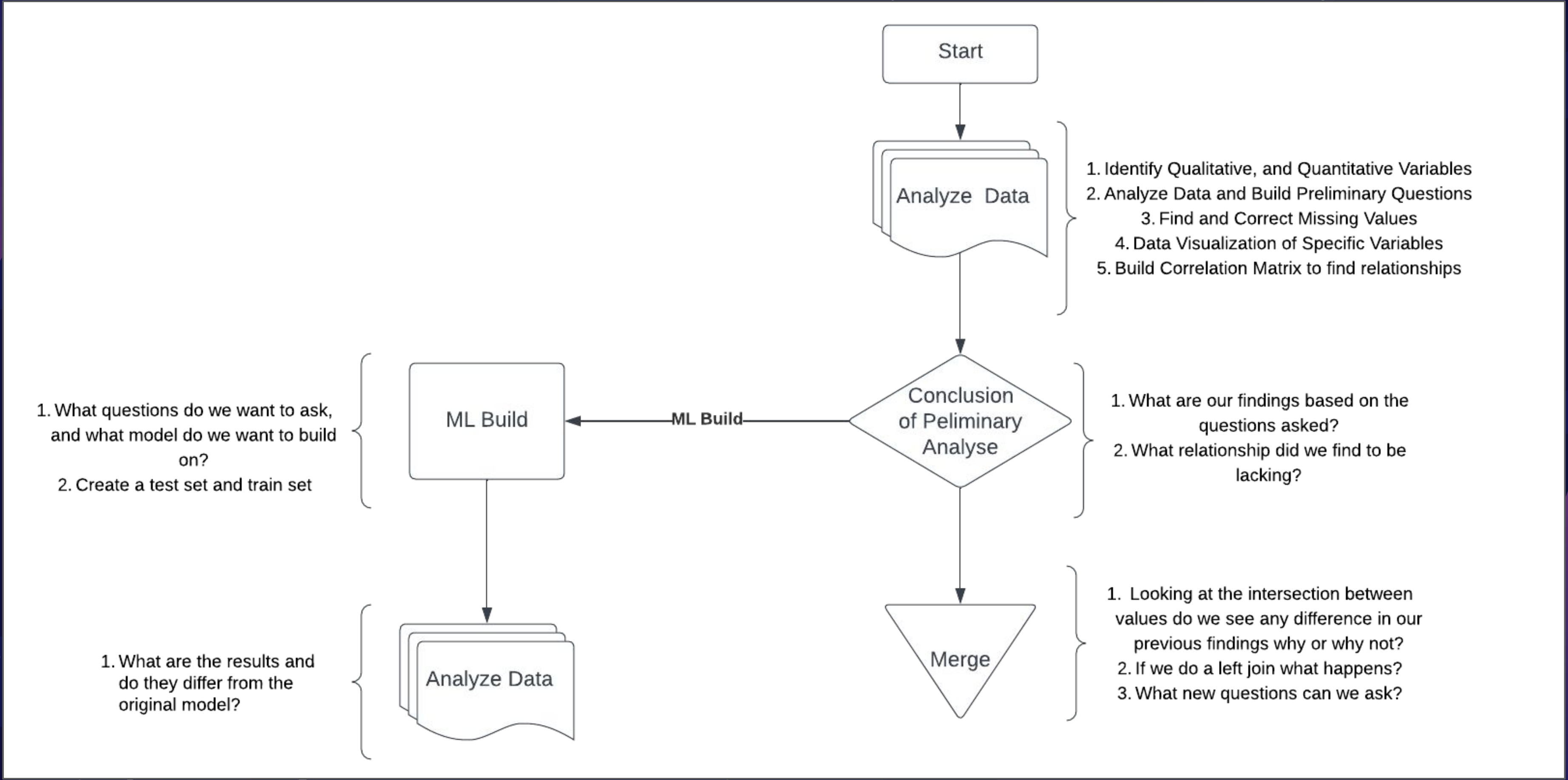
- 5 CSV Files
  - Business Course CSV
  - Design Course CSV
  - Web Development Course CSV
  - Music Course CSV
  - Customer CSV
- Description of Data per CSV
- Usability Score of 10
- Research Ideas

Link to Kaggle: https://www.kaggle.com/datasets/thedevastator/udemy-courses-revenue-generation-and-course-anal

Preliminary Questions

- Which Udemy Course Categories have the highest price on average?
- Which Udemy Course has the highest reviews on average?
- Is there a correlation between video length and price? Why or Why not?
- Is there a correlation above .5? Why or Why not?
- What is the average price spent per customer on Udemy Courses including title, free/paid statues?
- The frequency of money spent on courses, and statues?
- Tree Classification of Predicted Number of Subscribers given numerical columns
- Logistic Regression of Predicted Number of Subscribers given numerical columns

METHODOLOGY + FINDINGS



Start

Analyze Data

1. Identify Qualitative, and Quantitative Variables
2. Analyze Data and Build Preliminary Questions
3. Find and Correct Missing Values
4. Data Visualization of Specific Variables
5. Build Correlation Matrix to find relationships

Conclusion of Peliminary Analyse

1. What are our findings based on the questions asked?
2. What relationship did we find to be lacking?

ML Build

ML Build

1. What questions do we want to ask, and what model do we want to build on?
2. Create a test set and train set

Analyze Data

1. What are the results and do they differ from the original model?

Merge

1. Looking at the intersection between values do we see any difference in our previous findings why or why not?
2. If we do a left join what happens?
3. What new questions can we ask?

Step 1: Understanding our Data

Shape:

- Music: **(680, 12)**
- Web Development: **(1205, 12)**
- Design : **(604, 12)**
- Business : **(1192, 12)**
- Customer: **(3676, 14)**

Missing Values:

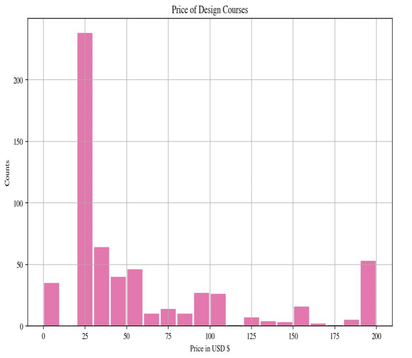
- No missing values to be seen, the <u>usability score</u> seems to be accurate.

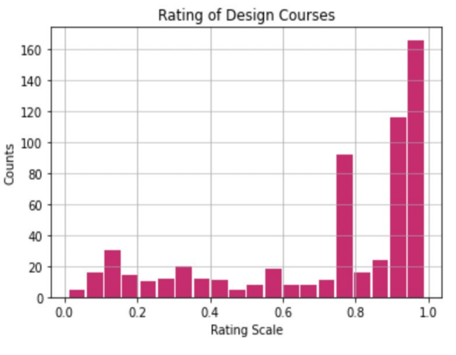
Notice anything interesting?

- All of them have same column names, however Customer has two extra columns (Date, Free/Paid)

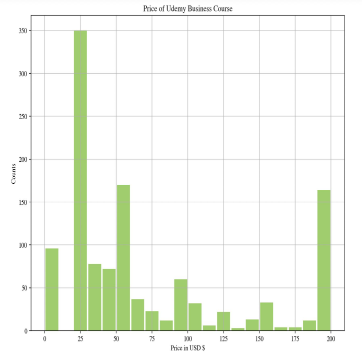
Step 2. Check for histogram on price and rating, what distribution is seen? Mean, Mode, Range?

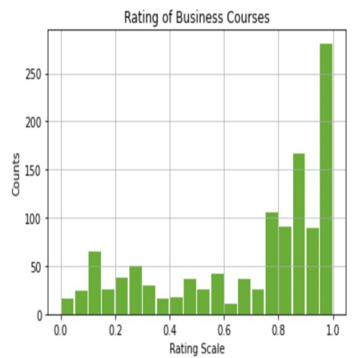
Design





Business





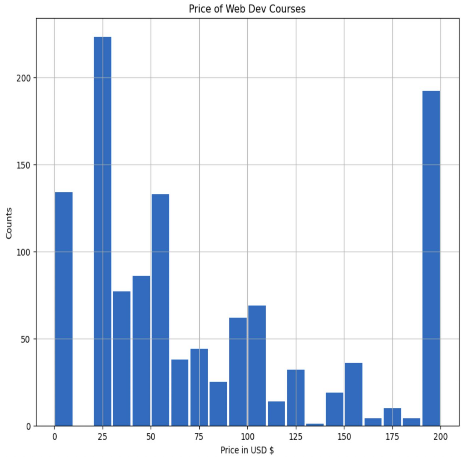
Mode: Price: $25.00, Rating: .98
Mean: Price: $57.89, Rating: .73
Range: Price: $0 - 200.00, Rating: [0,1]

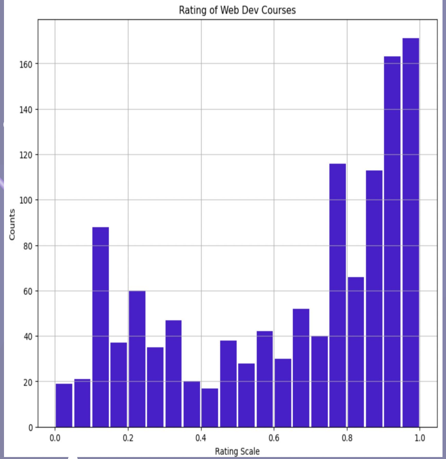
Mode: Price: $25.00, Rating: .99
Mean: Price: $68.69, Rating: .69
Range: Price: $0 - 200.00, Rating: [0,1]

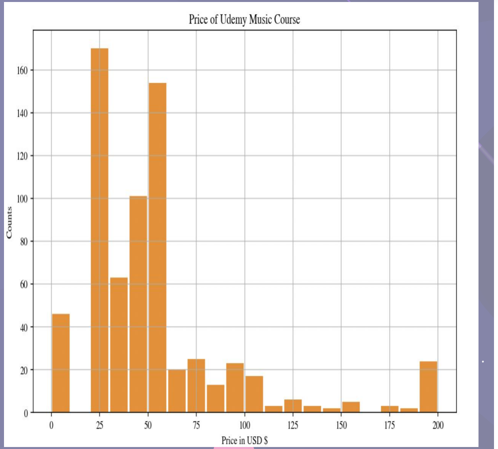
cont…

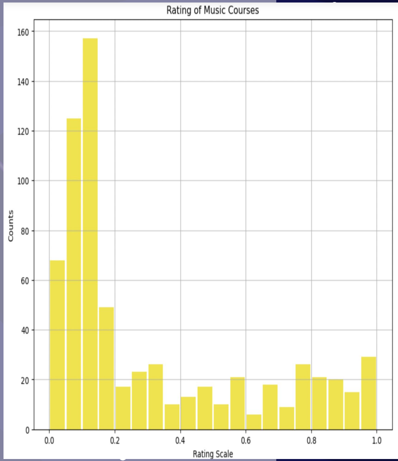
Web.Dev

Music









Mode: Price: $25.00, Rating: .99
Mean: Price: $77.04, Rating: .64
Range: Price: $0 - 200.00, Rating: [0,1]

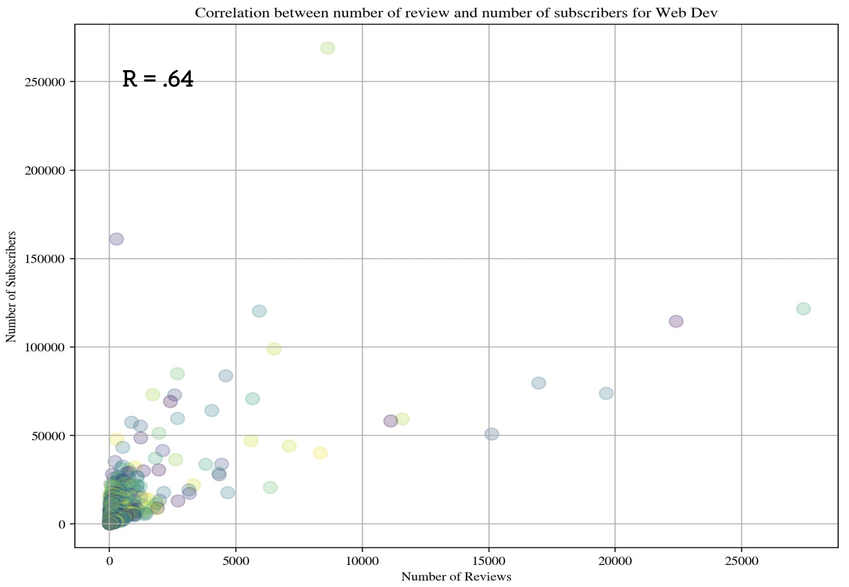
Mode: Price: $25.00, Rating: .10
Mean: Price: $49.56, Rating: .38
Range: Price: $0 - 200.00, Rating: [0,1]

Step 3. Choose a Category to go in depth.

- Web Development

Step 4. Build a Correlation Matrix

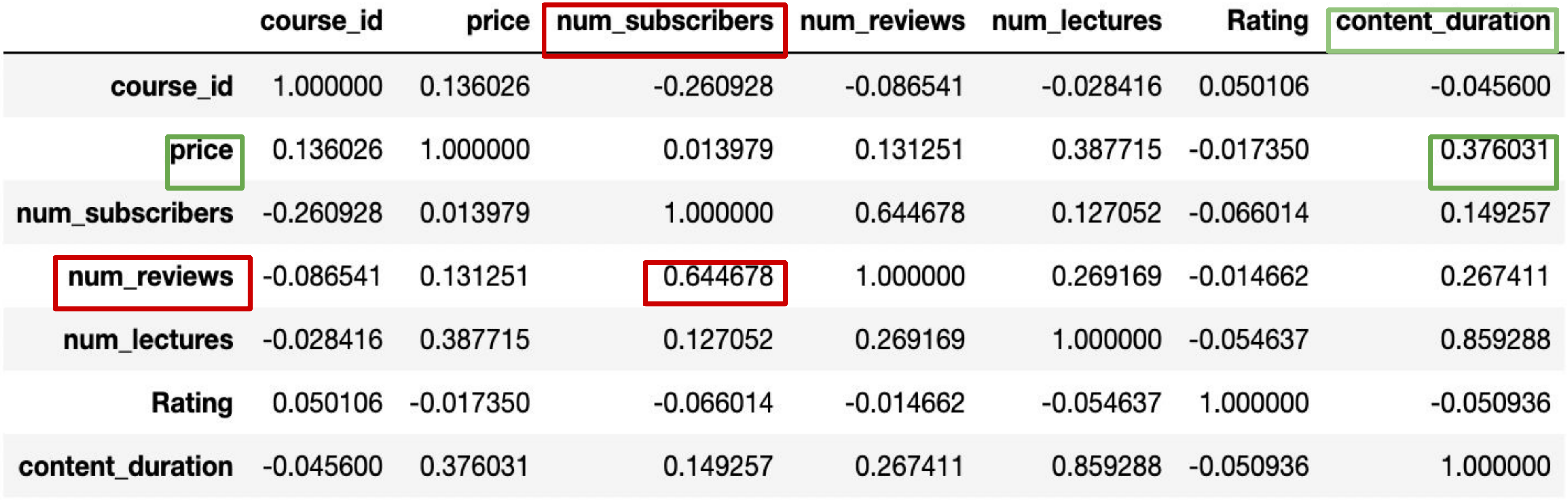
Strong	.8-1	+/-
Moderate	.5 -.8	+/-
Weak	.3 -.5	+/-
No correlation	0 - .3	+/-



Correlation between number of review and number of subscribers for Web Dev

R = .64

Correlation matrix for web.development



	course_id	price	num_subscribers	num_reviews	num_lectures	Rating	content_duration
course_id	1.000000	0.136026	-0.260928	-0.086541	-0.028416	0.050106	-0.045600
price	0.136026	1.000000	0.013979	0.131251	0.387715	-0.017350	0.376031
num_subscribers	-0.260928	0.013979	1.000000	0.644678	0.127052	-0.066014	0.149257
num_reviews	-0.086541	0.131251	0.644678	1.000000	0.269169	-0.014662	0.267411
num_lectures	-0.028416	0.387715	0.127052	0.269169	1.000000	-0.054637	0.859288
Rating	0.050106	-0.017350	-0.066014	-0.014662	-0.054637	1.000000	-0.050936
content_duration	-0.045600	0.376031	0.149257	0.267411	0.859288	-0.050936	1.000000

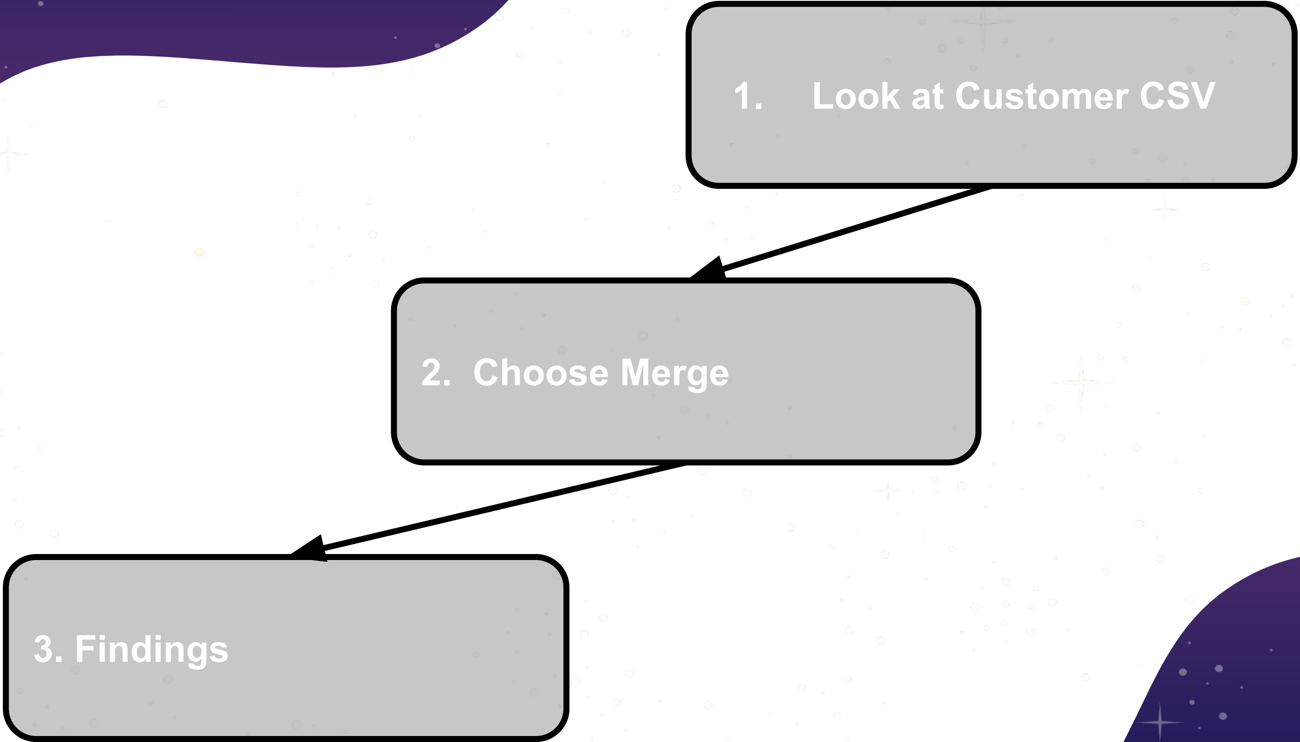
Solutions to Answer Part 1

- Which Udemy Course Categories have the highest price on average?
  - By building a histogram, and using the describe method, Web Development Courses have the highest price on average $77.04
- Which Udemy Course has the highest rating on average?
  - By building a histogram, and using the describe method, Design Courses have the highest Rating on average .73
- Is there a correlation between video length and price? Why or Why not?
  - By building a correlation matrix, we can notice a weak correlation between duration and price.
- Is there a correlation above .5? Why or Why not?
  - Yes, number of subscribers vs number of reviews seems to have a positive moderate correlation.

Noticed Areas to Explore

- Music has negative rating
- Histogram of web dev is exponential
- When building correlation matrix for all other categories, the correlation never changed below moderate for num_subscribers, num_reviews
- High correlation between duration and number of lectures
- Mode for all histograms involving price is the same

Merge



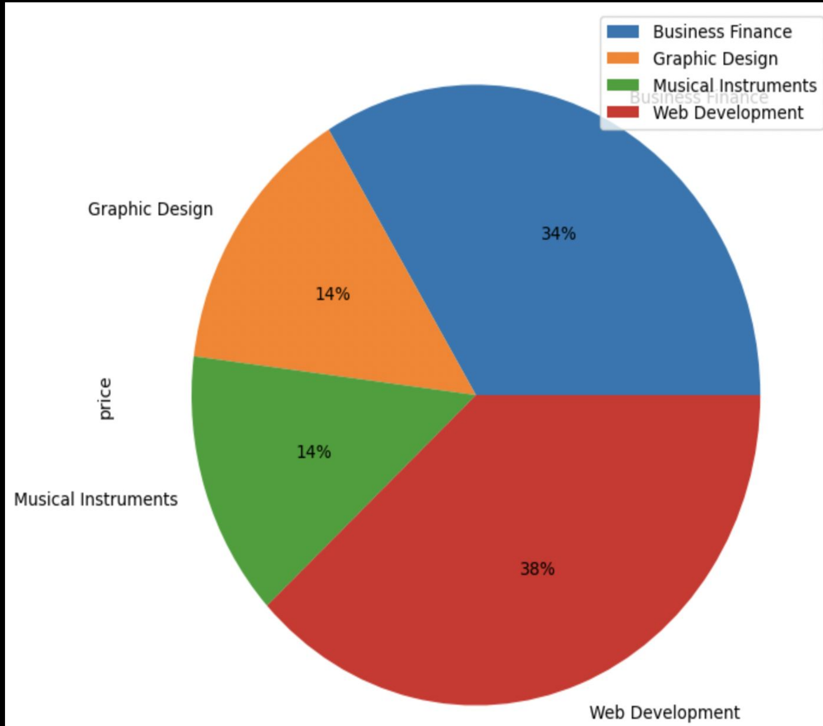
1. Look at Customer CSV

2. Choose Merge

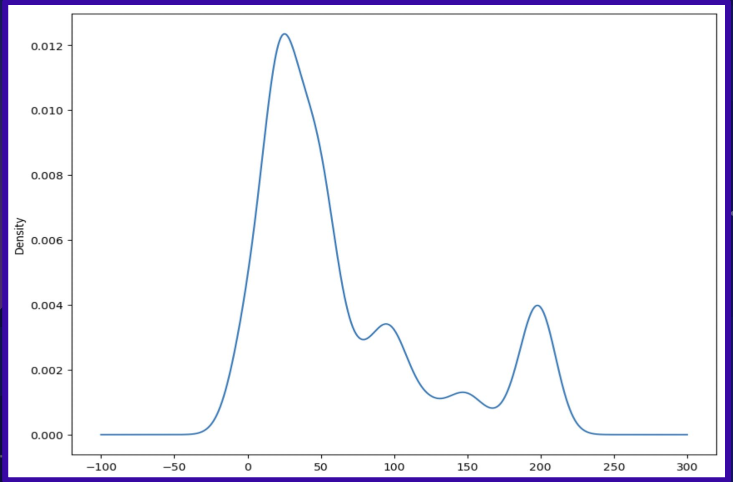
3. Findings

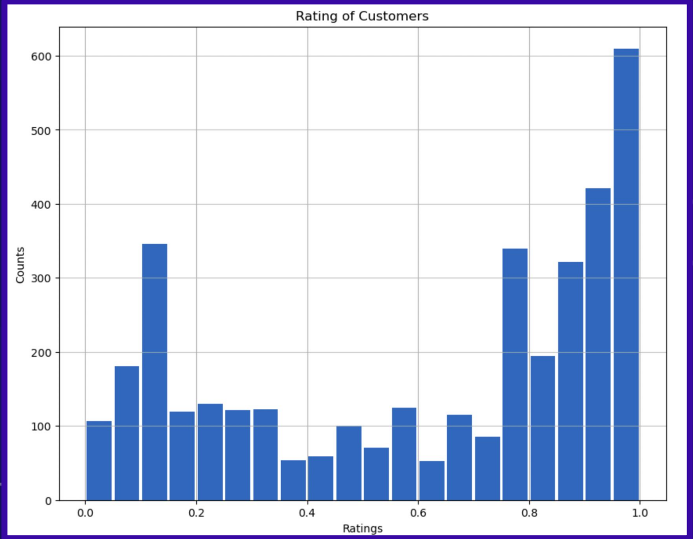
Customer Findings

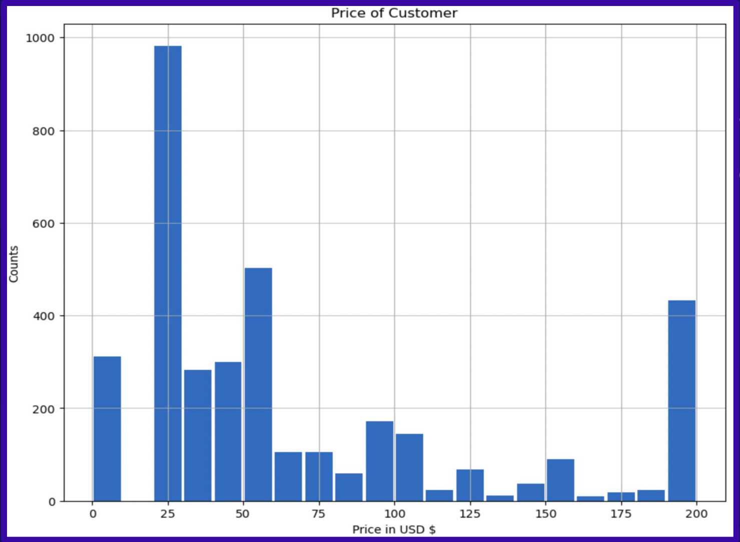
1. Customer CSV
   a. Customer.shape: **(3676, 14)**
   b. Shape follows other Histograms
   c. Two extra Columns (Date, Free/Paid)
2. Price
   a. Mean = $66.11
   b. Mode = $25.00
   c. Range = [0,200]

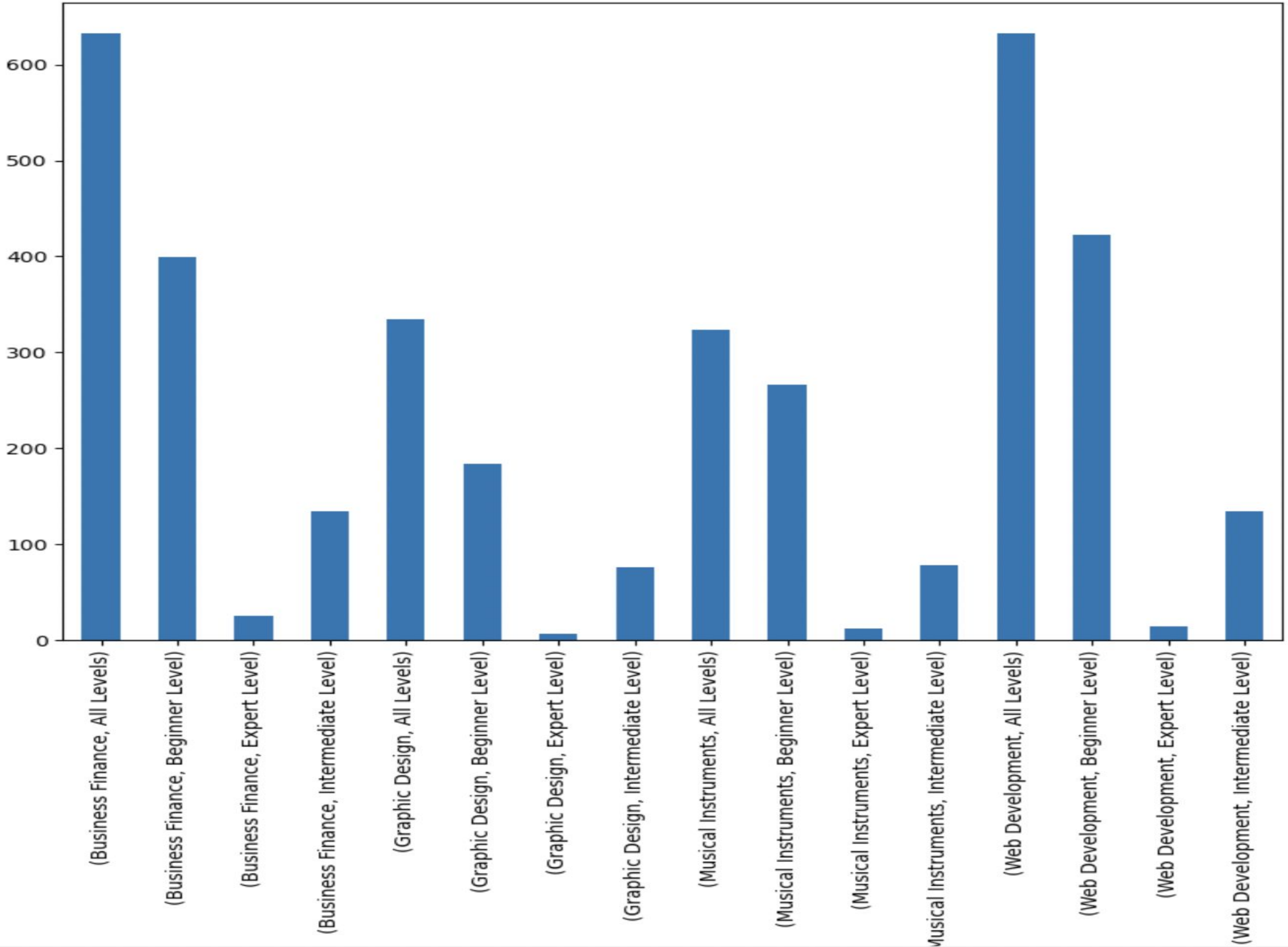


some visuals of customer



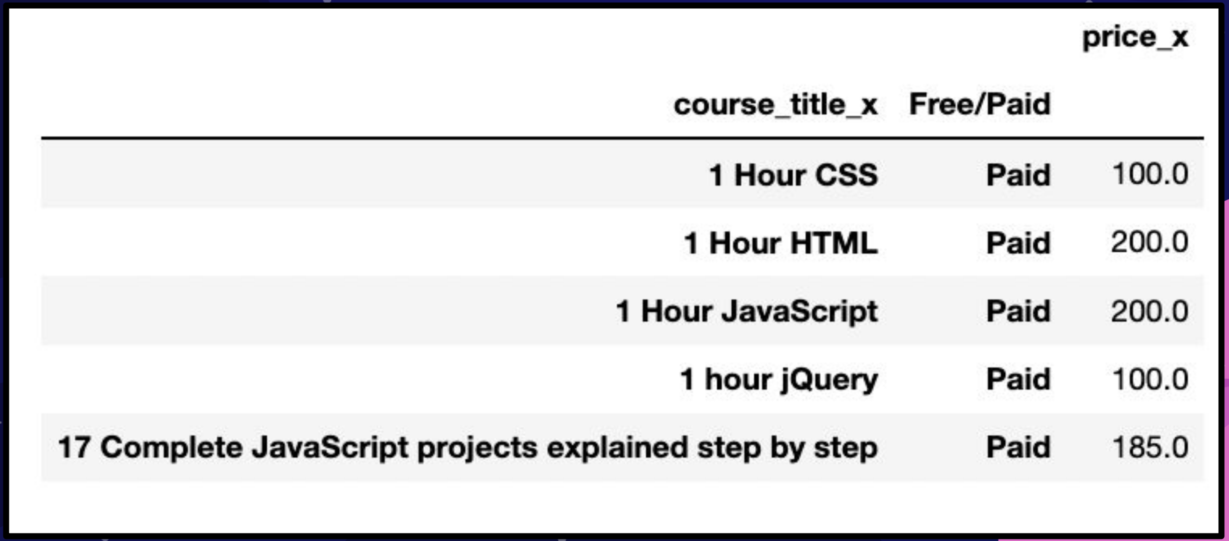





0
100
200
300
400
500
600
(Business Finance, All Levels)
(Business Finance, Beginner Level)
(Business Finance, Expert Level)
(Business Finance, Intermediate Level)
(Graphic Design, All Levels)
(Graphic Design, Beginner Level)
(Graphic Design, Expert Level)
(Graphic Design, Intermediate Level)
(Musical Instruments, All Levels)
(Musical Instruments, Beginner Level)
(Musical Instruments, Expert Level)
Musical Instruments, Intermediate Level)
(Web Development, All Levels)
(Web Development, Beginner Level)
(Web Development, Expert Level)
(Web Development, Intermediate Level)

Merge of customers

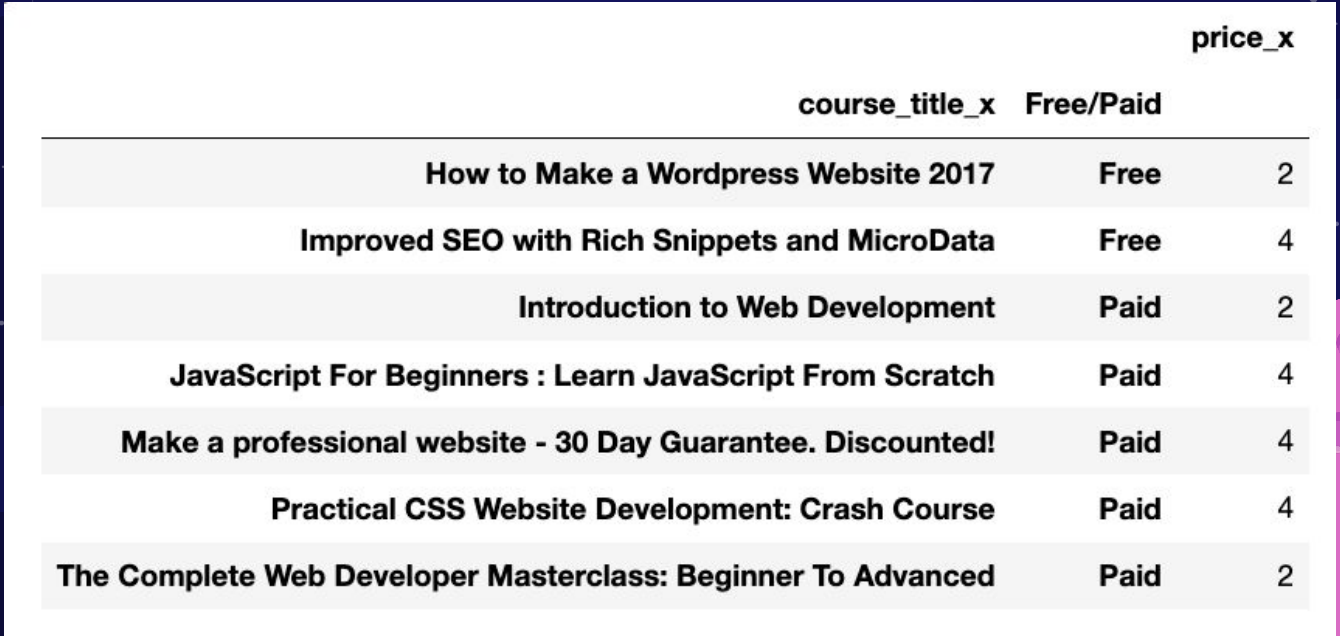
1. Merge Choice: Inner Join since I only want to look at intersection of customers and web development courses
   a. What is the average price spent per customer on Udemy Courses including title, free/paid statues?
   b. First top 5 seen in my data



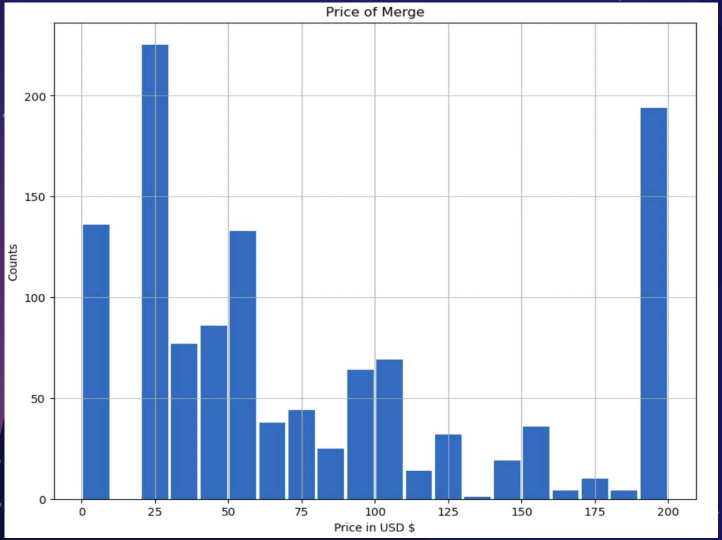
		price_x
course_title_x	Free/Paid	
1 Hour CSS	Paid	100.0
1 Hour HTML	Paid	200.0
1 Hour JavaScript	Paid	200.0
1 hour jQuery	Paid	100.0
17 Complete JavaScript projects explained step by step	Paid	185.0

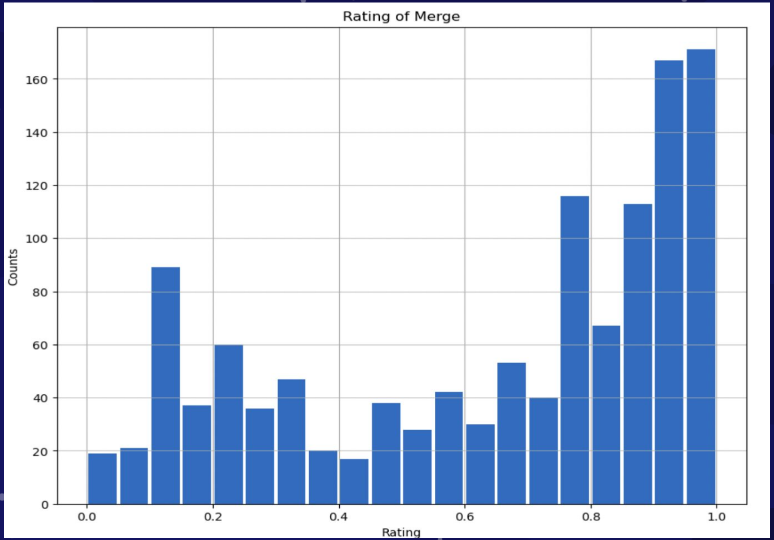
Merge of customers

- The frequency of money spent on courses, and statues?



		price_x
course_title_x	Free/Paid	
How to Make a Wordpress Website 2017	Free	2
Improved SEO with Rich Snippets and MicroData	Free	4
Introduction to Web Development	Paid	2
JavaScript For Beginners : Learn JavaScript From Scratch	Paid	4
Make a professional website - 30 Day Guarantee. Discounted!	Paid	4
Practical CSS Website Development: Crash Course	Paid	4
The Complete Web Developer Masterclass: Beginner To Advanced	Paid	2



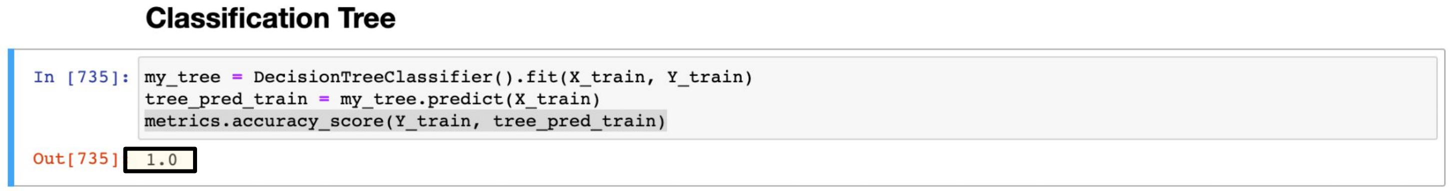


These are the price and rating histograms of the merge, We can notice it follows the trend of our previous histograms.

Machine Learning Models

Step 1. Dropna()
Step 2. Understand what columns we want
Step 3. Create Test,Train Datasets
Step 4. Pick Machine learning model
Step 5. Conclusion

1. Dropped all NaN values since the amount wouldn’t affect our Dataset
2. We have a mixture of categorical and numerical columns
    a. I chose all numerical columns
        i. Why not Categorical?
3. Creating Test, Train Dataset
    a. from sklearn.model_selection import train_test_split
        i. This is why I chose Numerical Variables
4. Decisions Tree Classification
    a. X = price, number of reviews, rating, number of lectures (variable to change)
    b. Y = Number of subscribers
    c. Results:



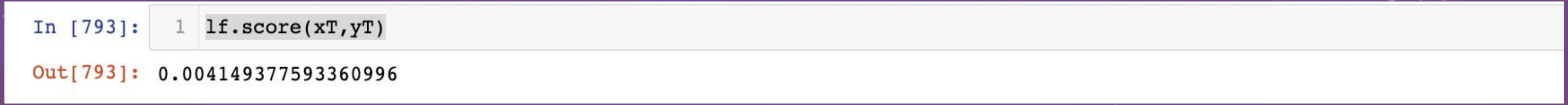
Classification Tree

```
In [735]: my_tree = DecisionTreeClassifier().fit(X_train, Y_train)
          tree_pred_train = my_tree.predict(X_train)
          metrics.accuracy_score(Y_train, tree_pred_train)

Out[735]: 1.0
```

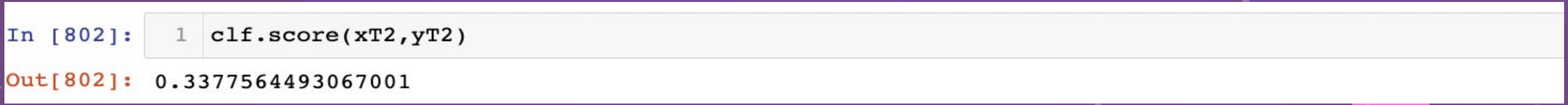
Machine Learning models continued

1. Logistic Regression
   a. Same X,Y variables were used for both Linear and Logistic that were in Class.Tree
   b. Results:



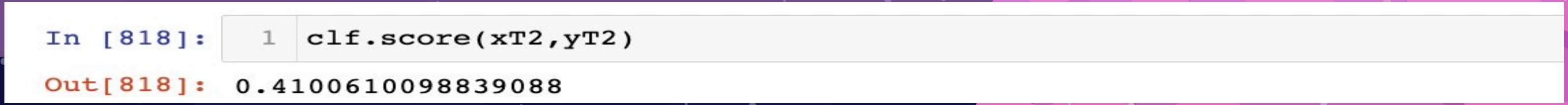
```
In [793]:  lf.score(xT,yT)
Out[793]: 0.004149377593360996
```

2. Linear Regression
   a. Results:



```
In [802]:  clf.score(xT2,yT2)
Out[802]: 0.3377564493067001
```

For Linear Regression we can notice a weak relationship between our variables, if manipulate variables we see an increase: in the relationship between num subscribers and num reviews, price



```
In [818]:  clf.score(xT2,yT2)
Out[818]: 0.4100610098839088
```

Conclusion and Future Work

Recommendations:

- When Students are on Udemy they should look at the amount of people and reviews in the course. We have seen there is a relationship in the reviews to subscribers, however high reviews alone aren't always a good sign.
- Courses regarding Web Development and Business would benefit based on the majority of excellent ratings, and vast diversity of content.
- Further research into Design courses within Udemy, should be conducted based on high ratings, and average minimal cost.
- Music courses are not recommended based on overwhelming negative ratings.

Conclusion:

- If VCU decides to offer Udemy they should market towards the School of Business and the School of Engineering.
- Udemy would benefit certain departments within the School of Arts such as the Computer Science/Graphic Design.

Future Works

- My only issue was time, I don’t feel like I had enough time. If I had more time I would do an in depth analysis on all courses, and merge all of them. I would also use different algorithms like K Means. I had issues with my test and training data so it would be nice to figure out why.
- I do hope to continue to work on this data, so I may be able to take the knowledge from this project and use it for projects to come.
- Open to any suggestions or areas I need to work on!

THANK YOU!