Brianna Griffin
MAT 265

Final Project

I. Introduction

A. For my final project, I will be using Python, R, and Tableau as technologies to analyze a data set that I found online. The data set that I found is from Kaggle, and originally contained 5 different CSV files. The context of the data is Udemy Courses. Udemy is an online platform in which you can take courses in a variety of subjects. These courses are either free or of charge. I will first clean the data set then analyze it, perform some statistical analysis, linear modeling, and visualize some of the results and findings.

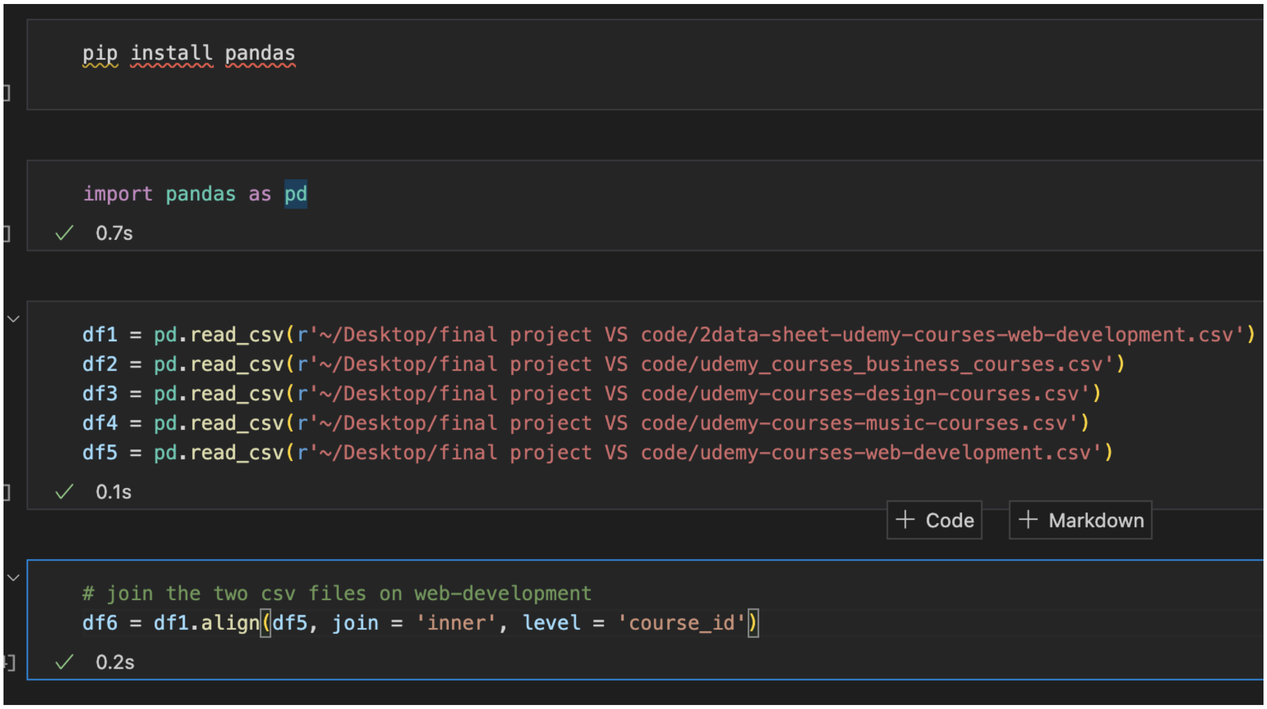
II. Project

A. Cleaning of the Data

Here are the columns for the data set.

course_id	Unique id for the course (string)
course_title	The title of the Udemy course. (String)
url	The URL of the Udemy course. (String)
price	The price of the Udemy course. (Float)
num_subscribers	The number of subscribers for the Udemy course. (Integer)
num_reviews	The number of reviews for the Udemy course. (Integer)
num_lectures	The number of lectures in the Udemy course. (Integer)
level	The level of the Udemy course. (String)
Rating	The rating of the Udemy course. (Float)
content_duration	The content duration of the Udemy course. (Float)
published_timestamp	The timestamp of when the Udemy course was published. (Datetime)
subject	The subject of the Udemy course. (String)

To begin, the CSV files were separated by subject. Using the *Pandas* package in Python, I imported all of the CSV files. Since there were two files for the web-development, I joined those two on course_id since this is unique for the course. The code for this is below:



```
pip install pandas
```

```
import pandas as pd
```

✓ 0.7s

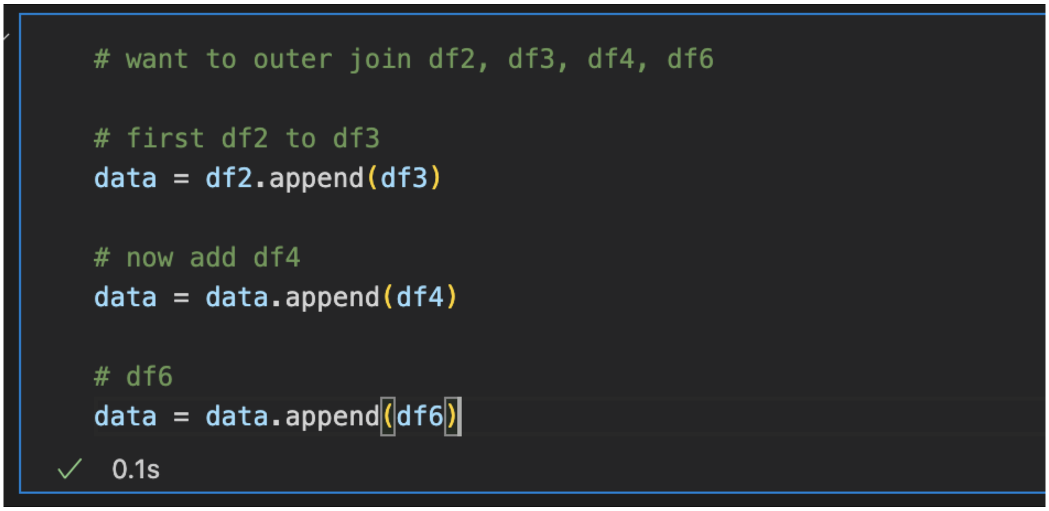
```
df1 = pd.read_csv(r'~/Desktop/final project VS code/2data-sheet-udemy-courses-web-development.csv')
df2 = pd.read_csv(r'~/Desktop/final project VS code/udemy_courses_business_courses.csv')
df3 = pd.read_csv(r'~/Desktop/final project VS code/udemy-courses-design-courses.csv')
df4 = pd.read_csv(r'~/Desktop/final project VS code/udemy-courses-music-courses.csv')
df5 = pd.read_csv(r'~/Desktop/final project VS code/udemy-courses-web-development.csv')
```

✓ 0.1s

```
# join the two csv files on web-development
df6 = df1.align(df5, join = 'inner', level = 'course_id')
```

✓ 0.2s

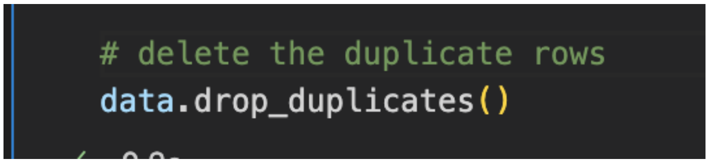
Now, I am going to join all of the data frames together so that it is easier to perform analysis on the data. Hence, we do not need to write so many lines of code.



```
# want to outer join df2, df3, df4, df6

# first df2 to df3
data = df2.append(df3)

# now add df4
data = data.append(df4)

# df6
data = data.append(df6)
```

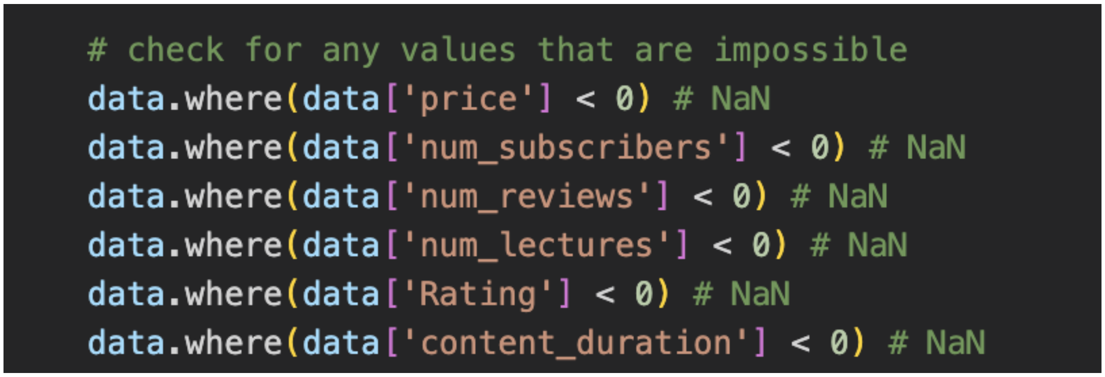
✓ 0.1s

Now, all of the data is together in one data frame and is named data!
I then dropped the duplicate rows. There were 10 of them.



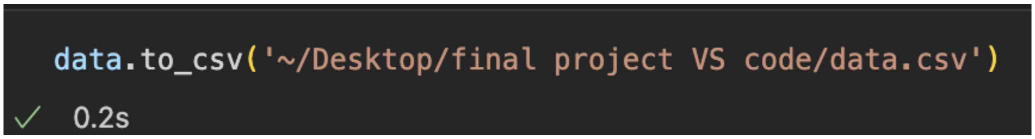
```
# delete the duplicate rows
data.drop_duplicates()
```

Finally, I will simply check for any observations that do not make sense. In the context of this data set this just means negative values for numerical data.



```
# check for any values that are impossible
data.where(data['price'] < 0) # NaN
data.where(data['num_subscribers'] < 0) # NaN
data.where(data['num_reviews'] < 0) # NaN
data.where(data['num_lectures'] < 0) # NaN
data.where(data['Rating'] < 0) # NaN
data.where(data['content_duration'] < 0) # NaN
```

Each where clause resulted in a matrix of Null values. Therefore, none of the observations are unreasonable and we can keep them all in the data frame. I will end by exporting the data frame as a CSV file.



```
data.to_csv('~/Desktop/final project VS code/data.csv')
✓ 0.2s
```

B. Data Analysis (R)

I will begin by doing some analysis on the data in R. I will need to import the CSV file into R to begin this. Attached Below is a PDF where I do some analysis of the data set. It allowed me to familiarize myself with the data set. This PDF includes the R code and its output. Data Analysis PDF

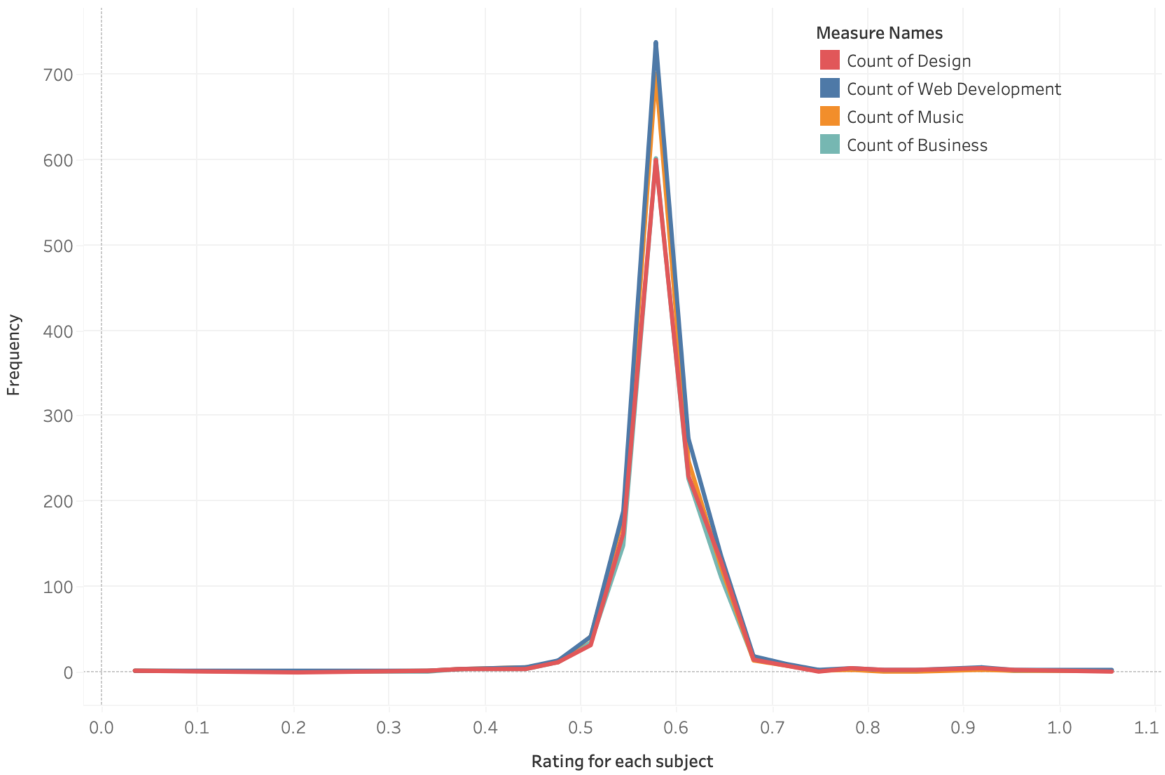
C. Statistics

Here is a PDF of the code and explanations of the statistics that I ran using the data. I created a model to predict the rating of the web development courses using the variables from the other courses. I made CSV tables with the model scores that I will use in the next part of the project which is data visualization.

Final Project - Statistical Analysis

D. Data Visualization

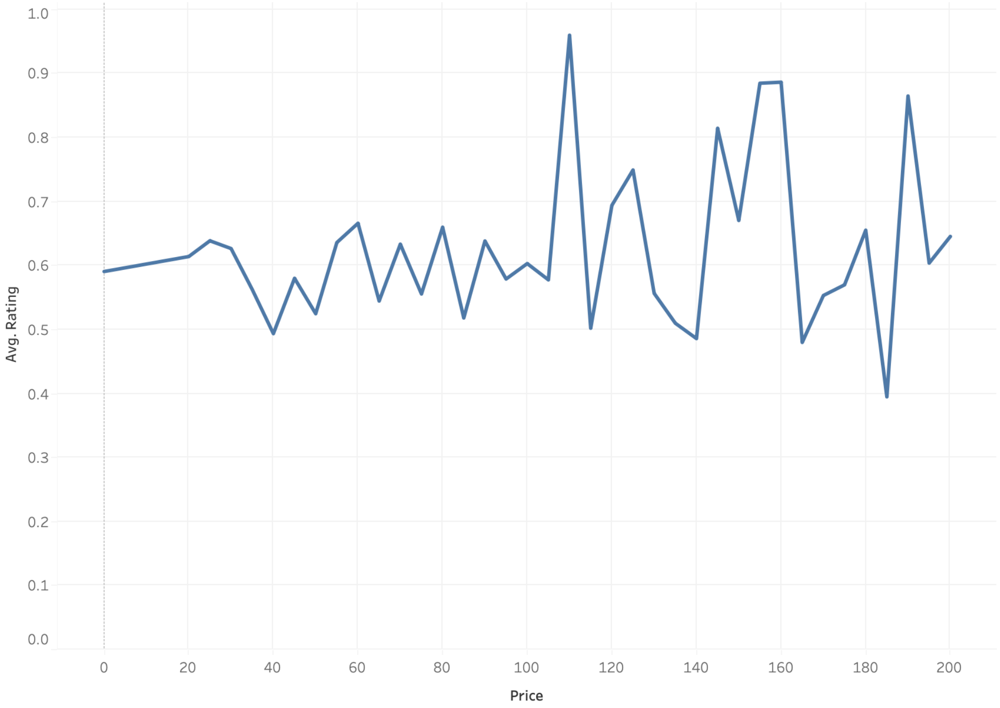
To end the project, I will perform some data visualization using Tableau as a technology. Hopefully these graphs will help interpret everything that I have done as a conclusion.



For the first graph, I graphed the rating for each subject along with its frequency. It is seen that they follow a similar pattern. Courses on web development, the ones predicted by the model, have the highest number of courses at the maximum which is something to note. Also, for web development, one of the courses was rated to be 105% which is not possible, but the model does not know that. Some error is expected.

The next visualization that I made compared the price of the Udemy course to rating. In this graph you are able to change some key categorical variables such as subject and level. I chose to add filters since the original plot between price and rating was very noisy.

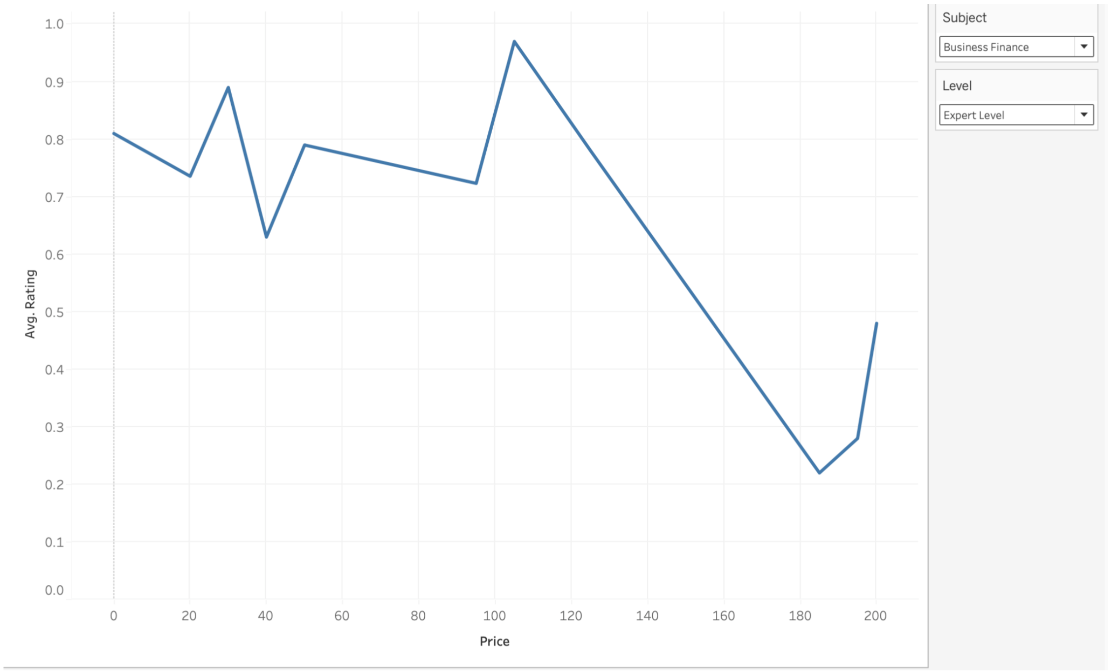
I.E.



There is no real trend to be seen.

Some key things that I found interesting are below. They involve separating the data into smaller groups. This is key since the data set is so large.

- Expert level Business courses



Subject

Business Finance

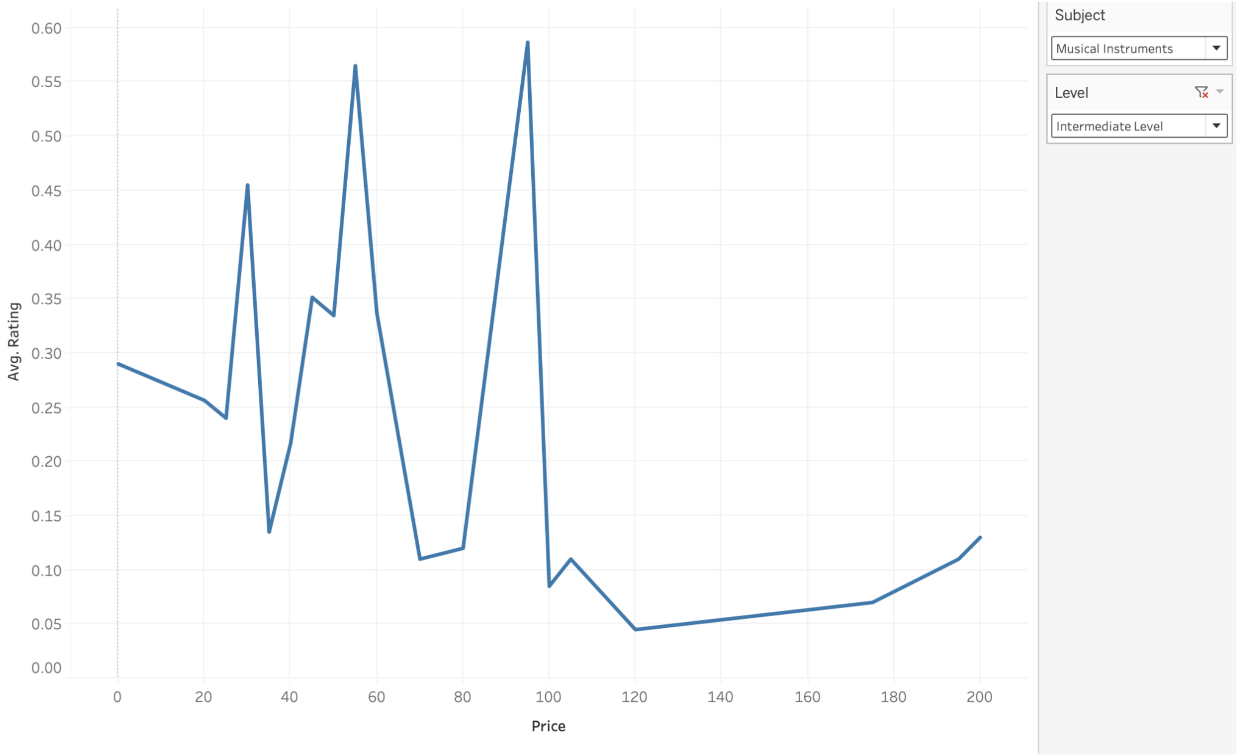
Level

Expert Level

The cheaper courses had a higher rating on average!

- Musical Instruments Intermediate Courses:

From the graph, I found that the courses that are averagely priced are the most popular. They are rated the best.



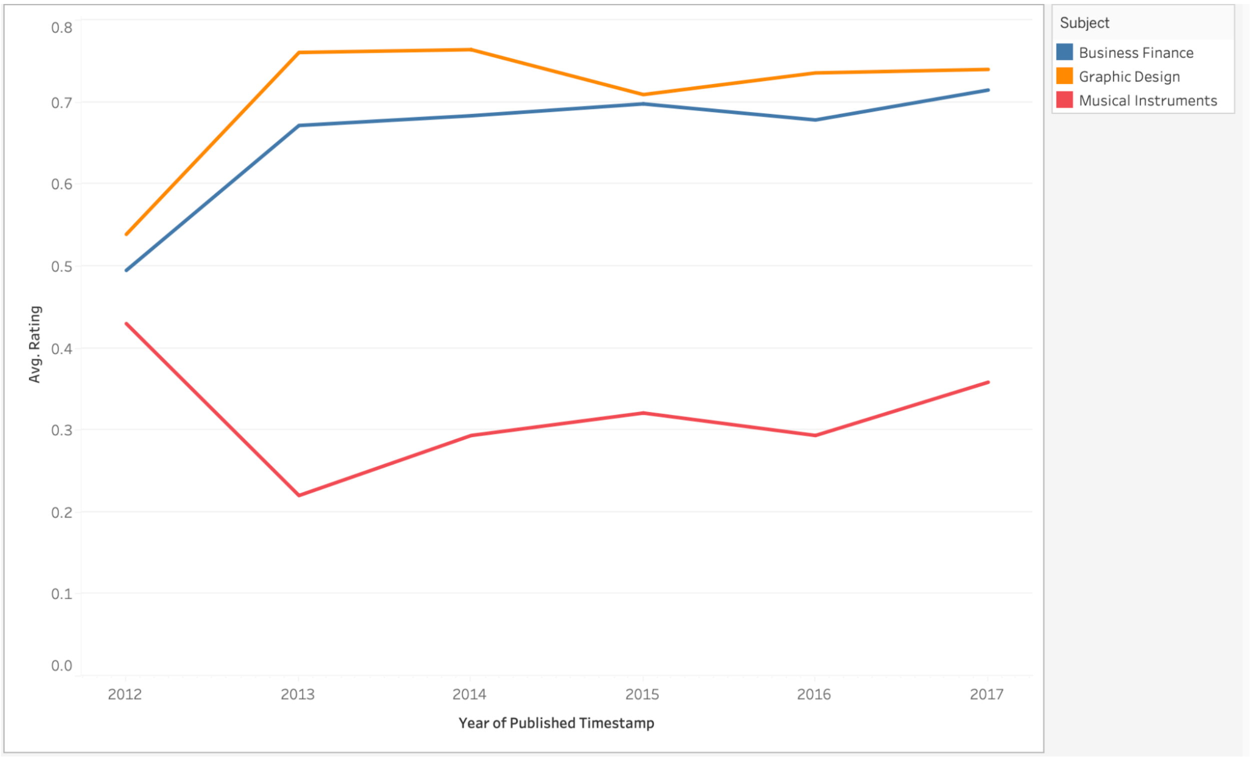
Subject

Musical Instruments

Level

Intermediate Level

The last visualization that I did was looking at the trend of the average ratings per subject over the years. The date is the year that the course was published.



It is seen that Business and Graphic Design follow a very similar pattern. Musical instruments also closely follow the same trend except in 2013 those courses saw a decrease in their ratings rather than an increase like the other two subjects.

III. Conclusion

A. In conclusion, I enjoyed applying skills I learned in this course and other courses at UCSB to an outside data set. I was able to try new things and combine many different techniques and technologies that I've been wanting to experiment with. I began by cleaning the data using the Pandas packing in Python. Then after initially analyzing the data, I built a linear model in order to calculate missing values. I finished my project with some visualizations to help grasp and visualize the large data set. It is important to visualize the interrelationships between two variables and look at small subsets of data when dealing with a large data set like the one I used for this project.

IV. Ranking of Past Assignments (most important to least)

1. Week 7 - MySQL commands new to you
2. Week 4 - Discover Patterns with MySQL
3. Week 8 - Outliers
4. Week 3 - 2nd Project in MySQL
5. Week 6 -MidTerm Presentation
6. Week 2 - 1st Project in MySQL

7. Week 9 - Random Sampling

V. Resources

   A. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_csv.html
   B. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html
   C. https://www.kaggle.com/datasets/thedevastator/udemy-courses-revenue-generation-and-course-anal?resource=download&select=Entry+Level+Project+Sheet+-+3.1-data-sheet-udemy-courses-web-development.csv
   D. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.drop_duplicates.html