Project Description

Description:

I created a CSV file, “dsci_510_dataset.csv,” for this project to document my final dataset. This dataset is formed by combining three separate dataframes described in the Data Sources section below (based on “universities_clean.csv,” “clean_crime_data.csv,” and ”clean_Zip_zhvi.csv”). I also attached the raw dataset for each CSV file before data cleaning for reference. These CSV files are included in the subfolder (“1st”,”2nd”, and “3rd”) in the folder named “Data” in the zip file that I submitted. With this data, I researched the universities in the US by analyzing their rankings, tuition, crime frequency in their zip codes, and house prices in their zip codes. I checked to see if there were correlations between these variables.

Motivation:

As USC students, nobody knows how to pay the highest tuition fees, but living in an area with high rental and crime rates is better than we do. If we could start again, would we make a different choice? How should we choose? Do all good universities have high tuition, living costs, and crime rates that will cause us to end up in similar circumstances no matter how we choose? As a result, these questions motivated me to look into the correlations between universities' rankings, tuition, crime rate, and living prices around the universities in this study. In addition, this data also helped me to perform statistical analysis about descriptive features of variables, such as the average tuition rate of all universities. Last but not least, I found the safest and cheapest universities among the top 50 universities, which is a valuable resource for deciding to choose a university to study.

Questions to Address:

1. Are there any relationships between universities rankings and the housing price of the cities where the universities are located?
2. Are there any relationships between universities rankings and the crime rates of the cities the universities are located in?
3. Are there any relationships between universities tuition rates, crime rate, and housing market of the cities the universities are located in? Perhaps correlation matrix?
4. Among top 50 universities, which is the cheapest & safest choice?
5. EDA showing basic features of this data

Data Sources:

1. A-Z Universities in the United States by Alphabetical Order (scraping)
   https://www.4icu.org/us/a-z/
   I scraped the information (Name, Ranking, Zip Code, University Website, Undergrad Local Tuition, and Undergrad International tuition) of more than 1700 universities in the US. I then coded the tuition into ordinal variables by using the criteria on the same website (criteria shown in Appendix A). I then used the average of coded local and international tuition to get "Average Tuition." The cleaned df1 has the shape (1642,7), and the 7 columns are "University," "Rank," "Zip Code," "Website," "Undergrad Local Tuition," "Undergrad International Tuition", and "Average Tuition."

University	Rank	Zip Code	Website
Massachusetts Institute of Technology	1	'02139'	https://www.4icu.org//reviews/5728.htm
Harvard University	2	'02138'	https://www.4icu.org//reviews/5720.htm
Stanford University	3	'94305'	https://www.4icu.org//reviews/5135.htm
Cornell University	4	'14853'	https://www.4icu.org//reviews/6077.htm
University of California, Berkeley	5	'94720'	https://www.4icu.org//reviews/5143.htm

Undergrad Local Tuition	Undergrad International Tuition	Average Tuition
17	17	17.0
17	17	17.0
17	17	17.0
17	17	17.0
7	15	11.0

2. Crime Data by Zip Code API (API)
   https://zylalabs.com/api-marketplace/data/crime+data+by+zipcode+api/824
   Link to fetch json:
   https://zylalabs.com/api/824/crime+data+by+zipcode+api/583/get+crime+rates+by+zip?zip=
   To increase maintainability, I put the API website and API key outside the function just in case they expire. By this API, I get information about the crime rate of the zip codes of all universities in the first dataset. One of the information provided in the API was the crime frequency in a sentence, such as "A crime occurs every 6 hours 53 minutes." I then re-coded this information to "Frequency in Hours," which is an interval ratio variable. After cleaning, this dataframe has the shape (1553, 6). The columns are "Zip Code," "Overall Grade," "Crime

Frequency," "Violent Crime Rate," "Property Crime Rate," and "Frequency in Hours."

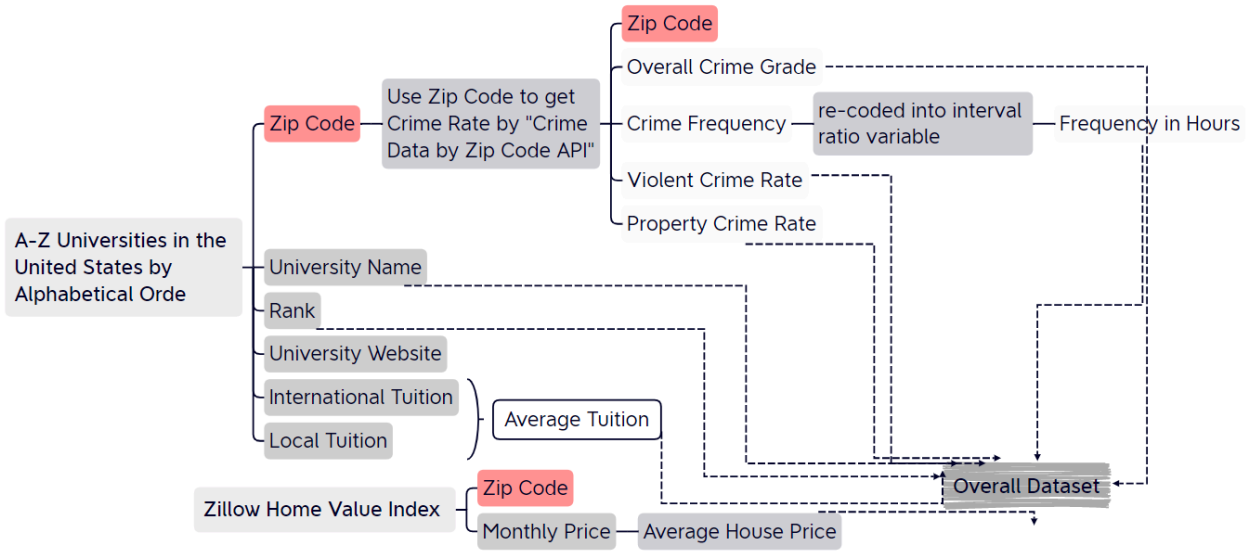
Zip Code	Overall Grade	Crime Frequency
'02139'	C-	A crime occurs every 6 hours 53 minutes (on av...
'02138'	F	A crime occurs every 3 hours 59 minutes (on av...
'94305'	D-	A crime occurs every 4 hours 29 minutes (on av...
'14853'	F	A crime occurs every 1 day 7 hours (on average...
'94720'	F	A crime occurs every 4 hours 59 minutes (on av...

Violent Crime Rate	Property Crime Rate	Frequency in Hours
4.434	23.67	6.883333
8.416	38.95	3.983333
7.027	36.73	4.483333
7.908	53.5	31.000000
17.54	231.6	4.983333

3. Zillow Home Value Index (CSV)
   https://www.zillow.com/research/data/
   This is a CSV file I downloaded from the Zillow Research website. The original dataset contains Zillow monthly housing marketing indicator price from 2000 to 2024. I used 4 numbers that stand for 4 quarters to get an average house price for the last year. After cleaning, the dataset has the shape (26350, 6). The 6 columns are "Zip Code," "2023-06-30," "2023-09-30," "2023-12-31," "2024-03-31," and "Average House Price."

Zip Code	2023-06-30	2023-09-30	2023-12-31	2024-03-31	Average House Price
'77494'	481531.770944	488672.706004	491236.726086	497633.682304	489768.721334
'08701'	542238.662882	560433.407041	579209.518786	588383.424533	567566.253310
'77449'	278227.439797	280436.466336	279896.109545	282060.397358	280155.103259
'11368'	476324.274236	473866.596636	463909.897770	462496.331770	469149.275103
'77084'	273028.966082	275197.791142	274336.157995	275920.544246	274620.864866

Flowchart for Dataset Generation:



I combined three datasets respectively about university's basic information, crime rate, and house price by using the same zip code. In the end, I got a combined dataset that has a shape (1272, 6). The columns are "University," "Rank," "Average Tuition," "Overall Grade," "Violent Crime Rate," "Property Crime Rate," "Frequency in Hours," and "Average House Price."

University	Rank	Average Tuition	Overall Grade	Violent Crime Rate	Property Crime Rate	Frequency in Hours	Average House Price
University of Virginia	33	13.0	B-	1.636	12.61	10.683333	490000.0
Massachusetts Institute of Technology	1	17.0	C-	4.434	23.67	6.883333	945000.0
Stanford University	3	17.0	D-	7.027	36.73	4.483333	3095000.0
Tufts University	45	17.0	D-	2.187	24.03	4.033333	780000.0
Washington University in St.	47	17.0	D-	6.417	59.29	4.016667	250000.0

Challenges:

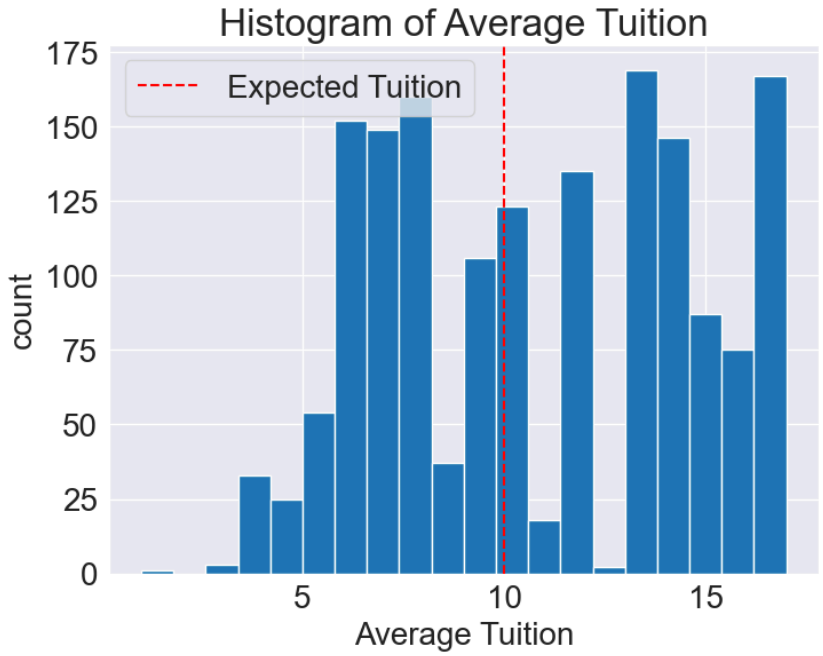
1. One challenge I overcame was the storage of zip codes. As you can see, I added single quote marks before and after the zip code for the separate data files. I did it because I found that no matter what types of data I stored (string or integer), the leading zeros automatically cleared when I read the CSV file in Python. For example, if one zip code should be "02139," when I open the file by Python, it will become "2139." When I searched online, I didn't find others with the same

problems. There might be some bugs on my side. But after many tries, I used a quotation mark to protect my zip codes. In this way, Python won't clear my zeros.

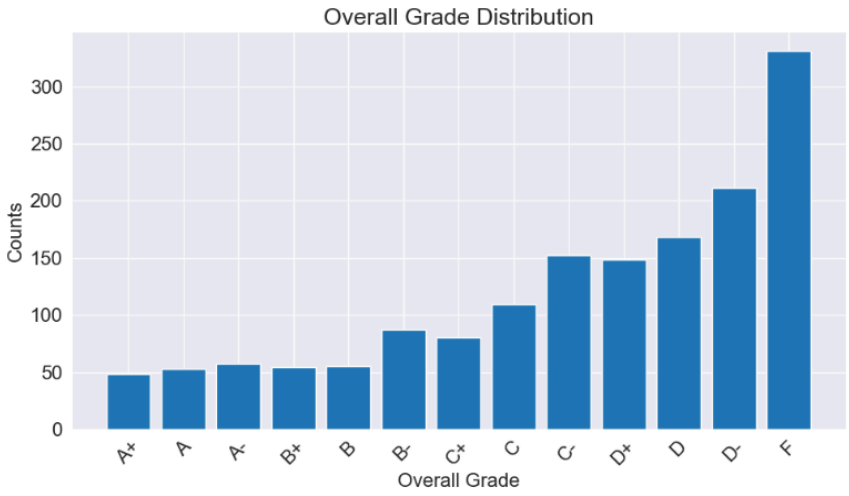
2. Some other challenges come with more complex coding requirements for this project. I searched online for some coding techniques we didn't learn in the class but needed for this project. For example, some of the data obtained online was not directly usable for data analysis, such as the sentence describing the crime rate. I need to re-code this data using a data frame. So I learned "apply" online for this re-coding requirement. Searching online helped a lot with this project

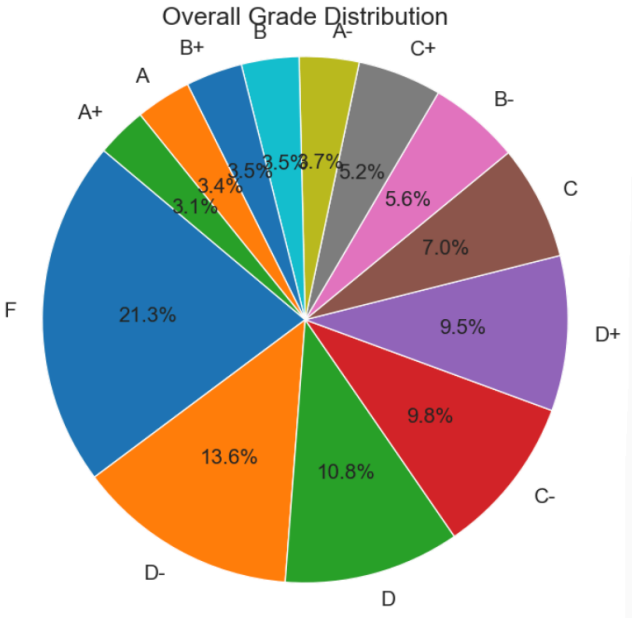
EDA:

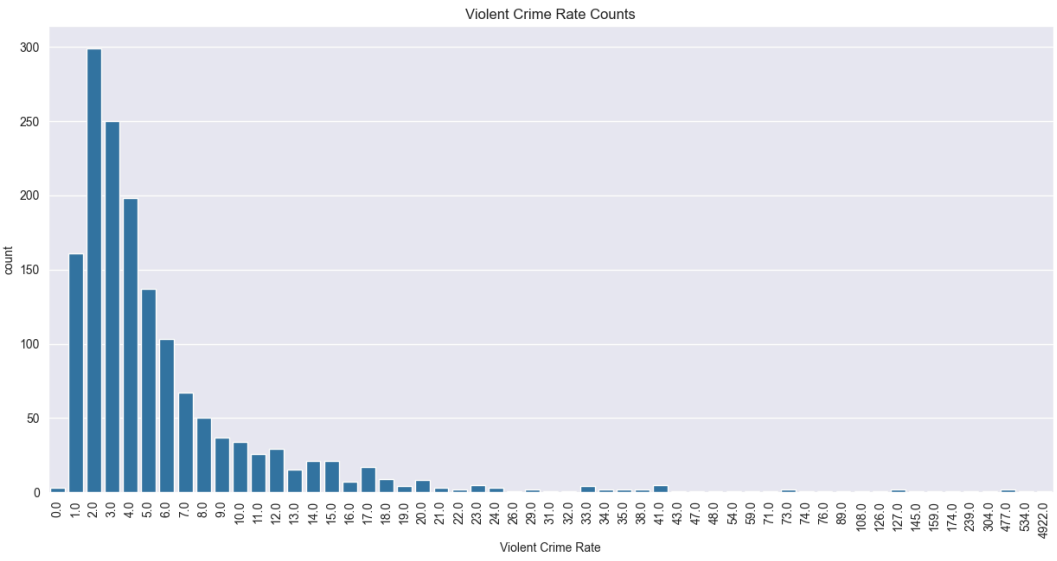
1. The distribution of the average tuition is asymmetrical. The median of the tuition is around 10, which refers to "20000-24999 (US dollars)" (conversion shown in Appendix A). The top 3 most frequent tuition rates are respectively "over 50000," "30000-34999," and "12500-14999" (conversion shown in Appendix A).



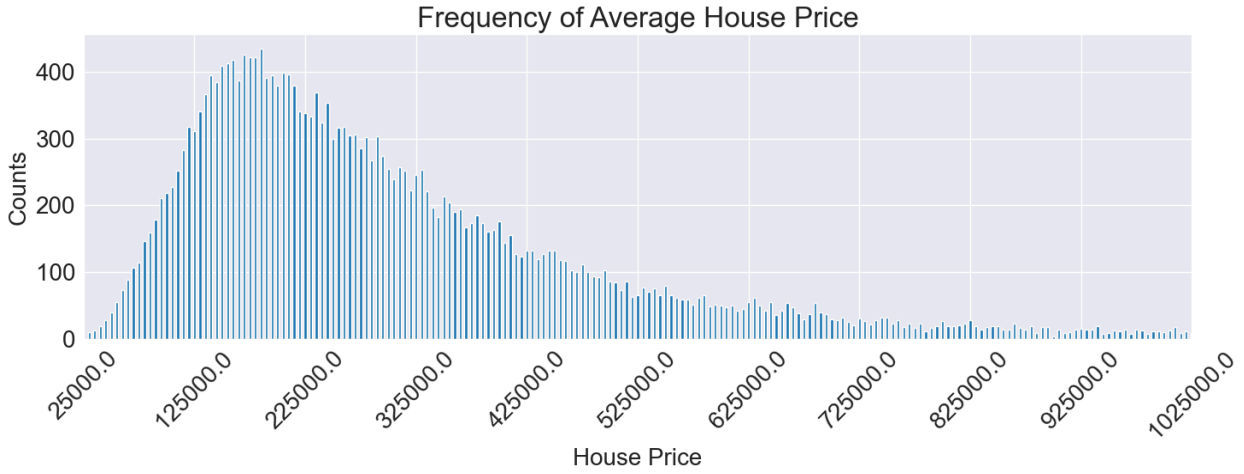
2. The distribution of crime grades is left-skewed, with zip codes with crime grades "F," "D-," or "D" taking up around half of the total number of zip codes. For other crime-related variables, such as violent crime rate (I use it for representative here), the distribution is right-skewed, which means the medians are much lower than the means and the frequencies of lower crime rates are higher.


Overall Grade Distribution
Counts
300
250
200
150
100
50
0
A+
A
A-
B+
B
B-
C+
C
C-
D+
D
D-
F
Overall Grade


Overall Grade Distribution
A+
A
B+
B
A-
C+
B-
C
D+
C-
D
D-
F
3.1%
3.4%
3.5%
3.5%
3.7%
5.2%
5.6%
7.0%
9.5%
9.8%
10.8%
13.6%
21.3%


Violent Crime Rate Counts
count
300
250
200
150
100
50
0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
26.0
29.0
31.0
32.0
33.0
34.0
35.0
38.0
41.0
43.0
47.0
48.0
54.0
59.0
71.0
73.0
74.0
76.0
89.0
108.0
126.0
127.0
145.0
159.0
174.0
239.0
304.0
477.0
534.0
4922.0
Violent Crime Rate

3. The distribution of house price is right skewed as well. While the Q3 is priced at 404099 dollars, the maximan price is 7839761 dollars, which is 18 times larger than the former. Most of the house prices here are lower than the mean of all prices.



	2023-06-30	2023-09-30	2023-12-31	2024-03-31	Average House Price
count	25884.00	25885.00	26234.00	26350.00	25884.00
mean	337522.83	343064.54	343506.56	345050.40	343282.55
std	306073.84	314139.15	316821.32	317023.83	314247.02
min	26461.48	26863.76	26045.50	27900.02	27210.27
25%	171704.26	173654.95	172834.37	173495.84	173514.55
50%	257529.77	261492.24	261477.82	263129.08	261509.12
75%	397784.30	403613.49	404006.38	406553.45	404098.71
max	7855186.58	7933379.77	7859018.83	7711460.45	7839761.41

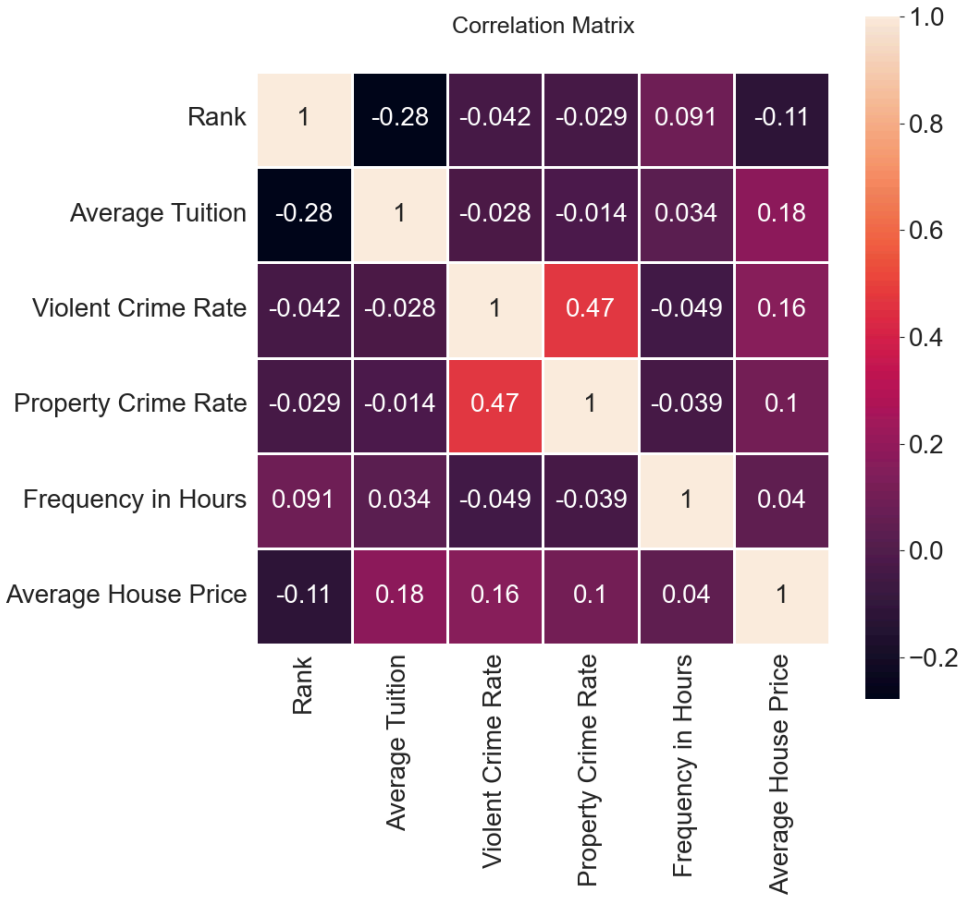
Analysis performed:

First, using my combined dataset, for top 50 universities, I create a list by sorting their “Frequency in Hours,” “Average Tuition,” and “Average House Price” in descending, ascending, and ascending orders. In this way, I can know which are the safest and cheapest top-50-universities in the US. The top 5 universities are UVA, MIT, Stanford, Tufts, and WashU. The tailed 5 universities are Columbia, UPenn, UChi, UIUC, and BU.

University	Rank	Average Tuition	Overall Grade	Violent Crime Rate	Property Crime Rate	Frequency in Hours	Average House Price
University of Virginia	33	13.0	B-	1.636	12.61	10.683333	490000.0
Massachusetts Institute of Technology	1	17.0	C-	4.434	23.67	6.883333	945000.0
Stanford University	3	17.0	D-	7.027	36.73	4.483333	3095000.0
Tufts University	45	17.0	D-	2.187	24.03	4.033333	780000.0
Washington University in St. Louis	47	17.0	D-	6.417	59.29	4.016667	250000.0

University	Rank	Average Tuition	Overall Grade	Violent Crime Rate	Property Crime Rate	Frequency in Hours	Average House Price
Boston University	40	17.0	F	14.880	133.00	2.033333	665000.0
University of Illinois Urbana-Champaign	30	11.5	F	11.430	91.61	1.833333	150000.0
University of Chicago	16	17.0	F	15.720	68.45	1.516667	205000.0
University of Pennsylvania	9	17.0	F	11.410	70.79	1.466667	205000.0
Columbia University in the City of New York	8	17.0	F	5.517	49.25	1.083333	820000.0

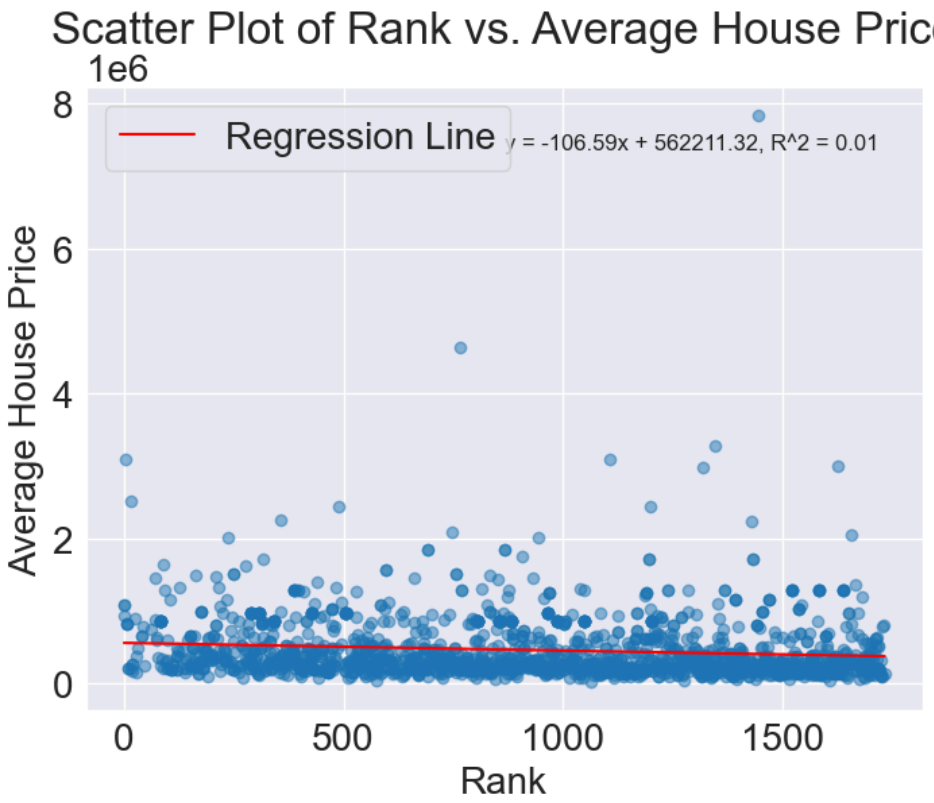
Second, I did correlation analysis of universities' rankings, tuition, crime frequency, and house price around them. The following is a correlation matrix of all variables. Most of the correlation coefficients here are low except that "Violent Crime Rate" and "Property Crime Rate"'s coefficient is moderate.



Third, I looked at Pearson correlations between some variables.

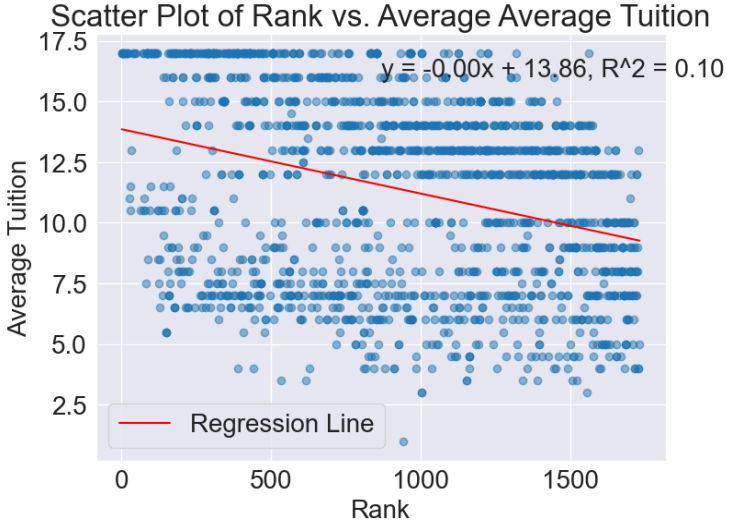
1. Rank & Average House Price
   The p-value is close to zero, so I'm nearly 100% confident that there is a significant correlation between the university's rank and house prices around the university. As shown in the scatter plot, there is a negative correlation between these two variables: for about 1 unit increase in the rank, the house price goes down about 100 dollars, although the R^2 here is very low, which means that this model doesn't really apply to anyone. However, at least it somewhat indicates that the general trend is that these 2 variables are negatively correlated.

```
p-value is 2.07104932706156e-06
There is a significant correlation between 'Rank' and 'Average House Price'.
```



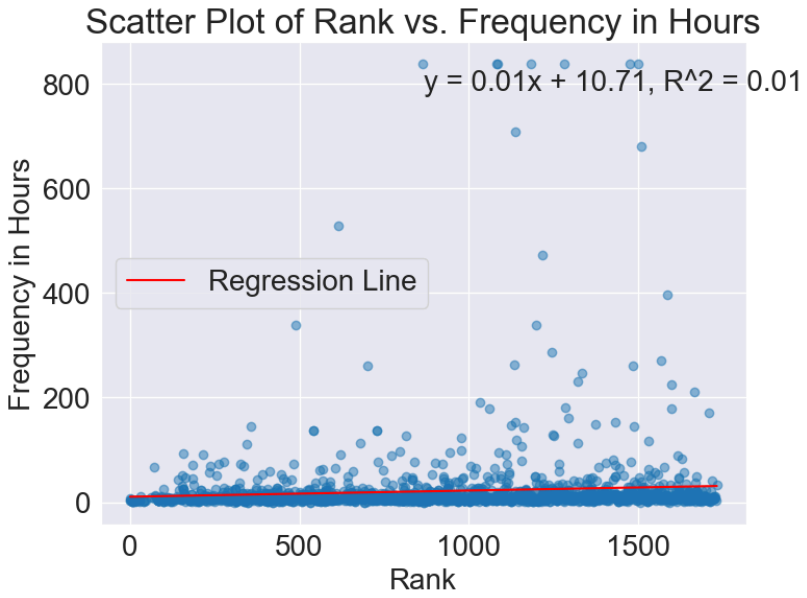
2. Rank & Average Tuition
   The p-value here is close to zero as well, so it's confident that there is a significant association between university's rank and tuition. As shown in the scatter plot, there is a tiny negative correlation between these two variables, and the R^2 is a little bit larger than the last time.

```
p-value is 4.69504950009481e-41
There is a significant correlation between 'Rank' and 'Average Tuition'.
```



3. Rank & Frequency in Hours
   The p-value here is smaller than 0.001, which means that I'm 99% sure there is an association between university rank and the crime rate around it. As shown in the scatter plot, the regression line is upward sloping, which means that the higher the university ranking is, the longer it will be to happen a crime (the less frequent the crime will be), although the R^2 isn't great as well.

```
p-value is 0.0007426251295918862
There is a significant correlation between 'Rank' and 'Frequency in Hours'.
```



4. I then did the Pearson correlations between Rank and Violent Crime Rate, Rank and Property Crime Rate, Average Tuition and Frequency in House. However, none of these peers have statistically significant correlations. It also sheds light on the fact that the overall crime frequency doesn't represent the crime rate of more specific types of crimes.

```
p-value is 0.12099440267498593
There is no significant correlation between 'Rank' and 'Violent Crime Rate'.
```

```
p-value is 0.2826241511505945
There is no significant correlation between 'Rank' and 'Property Crime Rate'.
```

```
p-value is 0.9325754242861386
There is no significant correlation between 'Average Tuition' and 'Frequency in Hours
```

Conclusion:

From the basic analysis, we can conclude that when a university's ranking increases, the house price will decrease, the tuition will decrease, and the crime frequency will

decrease. The generalizability of these results might be limited due to low R^2s. More complex analyses, such as multivariate regression, could be conducted in the future to explore the relationships between these variables further while each other is present in the same model. The other limitation could result from the availability of the sources of data. For example, the university's tuition is coded as an ordinal variable here, which might decrease the precisions of those variables. The future research will be more precise if more detailed variables can be found.

Appendix A:

```
tuition_replacements = {
    'Not reported': 'NA',
    '0-999': 1,
    '1000-2499': 2,
    '2500-4999': 3,
    '5000-7499': 4,
    '7500-9999': 5,
    '10000-12499': 6,
    '12500-14999': 7,
    '15000-17499': 8,
    '17500-19999': 9,
    '20000-24999': 10,
    'over 20000': 11,
    '25000-29999': 12,
    '30000-34999': 13,
    '35000-39999': 14,
    '40000-44999': 15,
    '45000-49999': 16,
    'over 50000': 17
}
```