DSCI 510 Project Description

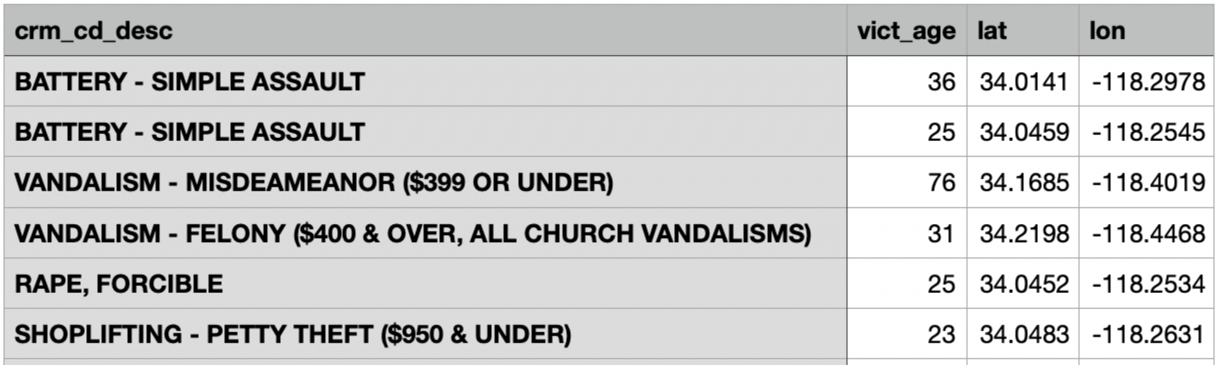
Description: For the final project, I have created a dataset in the form of a csv file called 'dsci_510_dataset.csv'. This dataset is created using three different data sources listed below in the Data Sources Section. Following files: 'dataset_zipcode.csv' and 'crime_dataset.csv' are intermediate csv files that have led to the creation of the final dataset. The three aforementioned datasets are provided in this zip folder and are present in a folder named 'Datasets'. Using these datasets I have analyzed crimes taking place in the several areas (zip codes) of Los Angeles and their correlation with the housing prices in that particular zip code.

Motivation: I find Real Estate to be quite intriguing and before moving to Los Angeles I watched a couple of apartment hunting videos on youtube where apartments in different areas of Los Angeles were covered. One of the aspects that comes into consideration while renting an apartment is the frequency of crime in that neighborhood. Some say that areas, where crimes are less frequent, have high rentals as compared to areas where crimes frequently occur. However, some are of the opinion that rich neighborhoods (where apartments are highly priced), are more likely to be hotspots for crimes. These conflicting opinions motivated me to statistically find out if crime frequency actually affects the rental price or not using the past crime data. Moreover, the datasets that I used also helped me perform a statistical analysis on the age of victims. Additionally, I have also generated a list of safest and most dangerous neighborhoods in LA using the crime frequency.

Data Sources:

1. Lacity Crime Data from 2020 to Present:
(https://www.splitgraph.com/lacity/crime-data-from-2020-to-present-2nrs-mtv8)
Link to fetch json:
"https://data.splitgraph.com:443/lacity/crime-data-from-2020-to-present-2nrs-mtv8/latest/-/rest/crime_data_from_2020_to_present?limit=5000"
This URL is already present in the code and is not to be given as an input by the user. Using the requests library, I have accessed the json from the URL and created crime_dataset.csv file. This dataset consists of 5000 rows and 4 columns. The columns in the dataset are as follows: crm_cd_desc (gives description of crime), lat, lon and vict_age (gives age of victim). I have used the coordinates of latitude and longitude from every row in the dataset to obtain the zip code of the crime location. First few rows of Crime Dataset shown below.

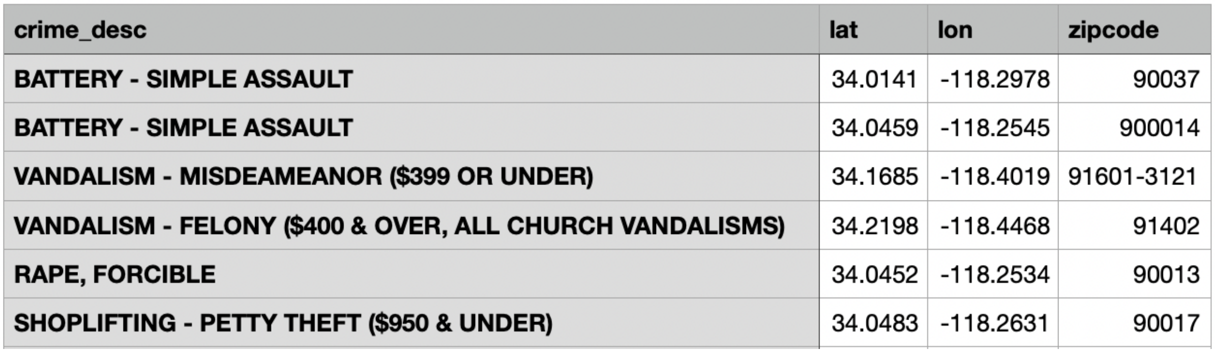
crime_dataset



crm_cd_desc	vict_age	lat	lon
BATTERY - SIMPLE ASSAULT	36	34.0141	-118.2978
BATTERY - SIMPLE ASSAULT	25	34.0459	-118.2545
VANDALISM - MISDEAMEANOR ($399 OR UNDER)	76	34.1685	-118.4019
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	31	34.2198	-118.4468
RAPE, FORCIBLE	25	34.0452	-118.2534
SHOPLIFTING - PETTY THEFT ($950 & UNDER)	23	34.0483	-118.2631

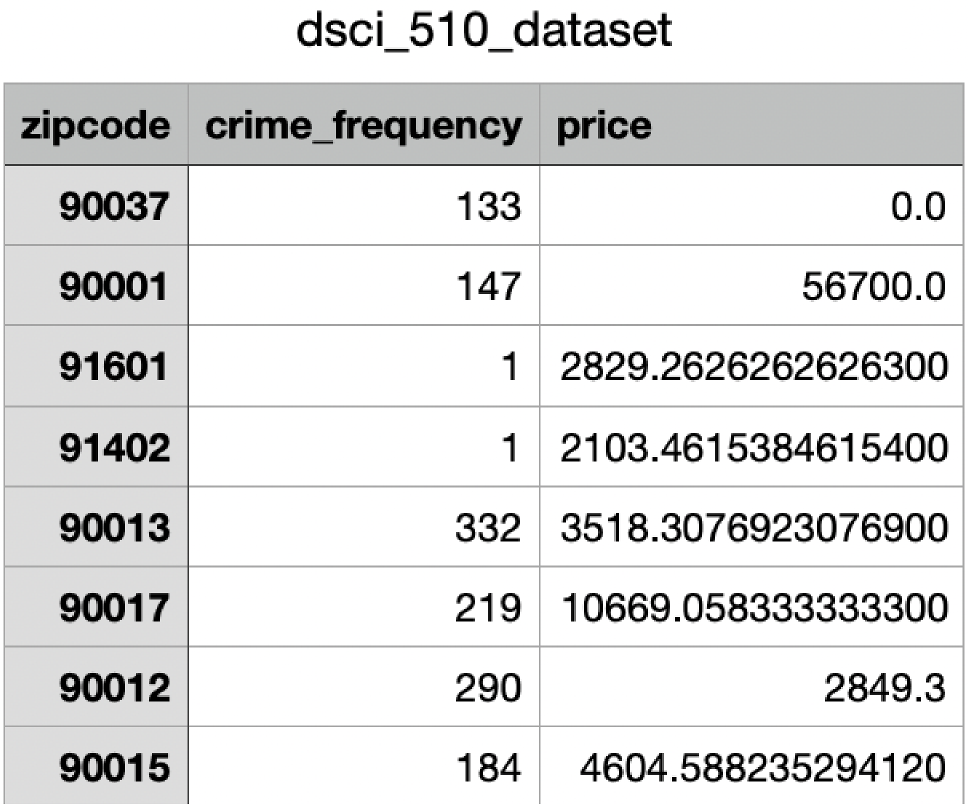
2. Reverse Geocoding service by openmapquest:
(https://towardsdatascience.com/reverse-geocoding-in-python-a915acf29eb6)
Using reverse geocoding, I have obtained zip codes from latitude and longitude coordinates. I have created a second dataset called 'dataset_zipcode'. There are 4 columns: crm_cd_desc (gives description of crime), lat, lon and zip code, in this dataset. Moreover, using this dataset, I have created a dictionary whose keys are zip codes and values are the number of times a zip code is occuring in the dataset. Frequency of zip codes will indicate the frequency of crime in a particular zip code. First few rows of Zipcode Dataset shown below:

dataset_zipcode



crime_desc	lat	lon	zipcode
BATTERY - SIMPLE ASSAULT	34.0141	-118.2978	90037
BATTERY - SIMPLE ASSAULT	34.0459	-118.2545	900014
VANDALISM - MISDEAMEANOR ($399 OR UNDER)	34.1685	-118.4019	91601-3121
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	34.2198	-118.4468	91402
RAPE, FORCIBLE	34.0452	-118.2534	90013
SHOPLIFTING - PETTY THEFT ($950 & UNDER)	34.0483	-118.2631	90017

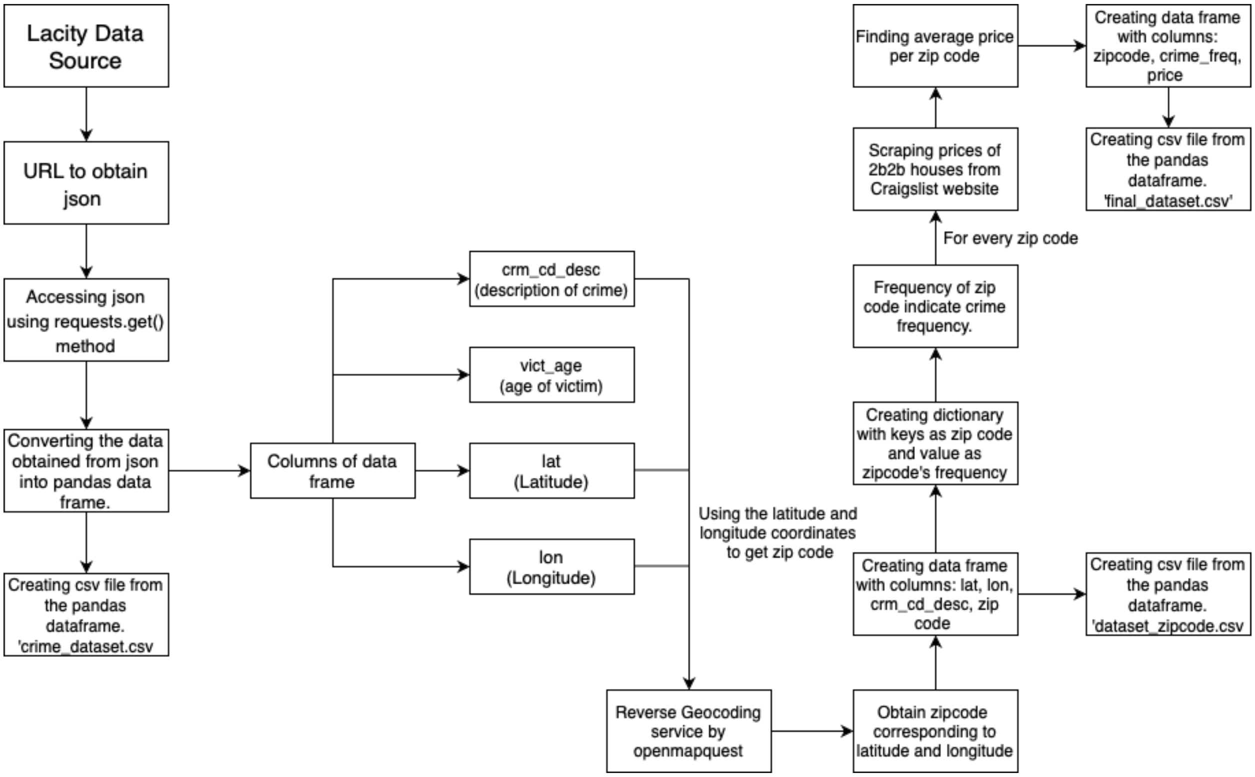
3. Craigslist House Price:
Using the zip codes listed in the above dictionary, I am scraping prices (monthly rent) of all 2 Bedroom and 2 Bathroom apartments (to maintain uniformity) and then finding the average. Furthermore, I am creating the final dataset on which I will be performing analysis in future. This dataset is stored as 'dsci_510_dataset.csv' in the folder 'Datasets'. First few rows of Final Dataset shown below:



dsci_510_dataset

zipcode	crime_frequency	price
90037	133	0.0
90001	147	56700.0
91601	1	2829.2626262626300
91402	1	2103.4615384615400
90013	332	3518.3076923076900
90017	219	10669.058333333300
90012	290	2849.3
90015	184	4604.588235294120

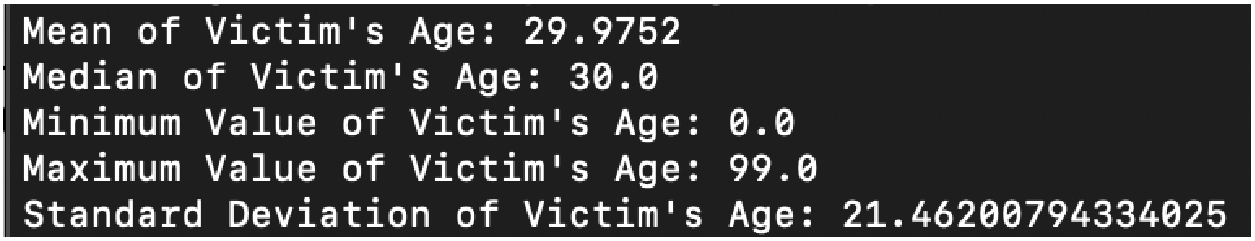
FLOWCHART FOR DATASET GENERATION:


Lacity Data Source
URL to obtain json
Accessing json using requests.get() method
Converting the data obtained from json into pandas data frame.
Creating csv file from the pandas dataframe. 'crime_dataset.csv
Columns of data frame
crm_cd_desc (description of crime)
vict_age (age of victim)
lat (Latitude)
lon (Longitude)
Using the latitude and longitude coordinates to get zip code
Reverse Geocoding service by openmapquest
Obtain zipcode corresponding to latitude and longitude
Creating data frame with columns: lat, lon, crm_cd_desc, zip code
Creating csv file from the pandas dataframe. 'dataset_zipcode.csv'
Creating dictionary with keys as zip code and value as zipcode's frequency
Frequency of zip code indicate crime frequency.
For every zip code
Scraping prices of 2b2b houses from Craigslist website
Finding average price per zip code
Creating data frame with columns: zipcode, crime_freq, price
Creating csv file from the pandas dataframe. 'final_dataset.csv'

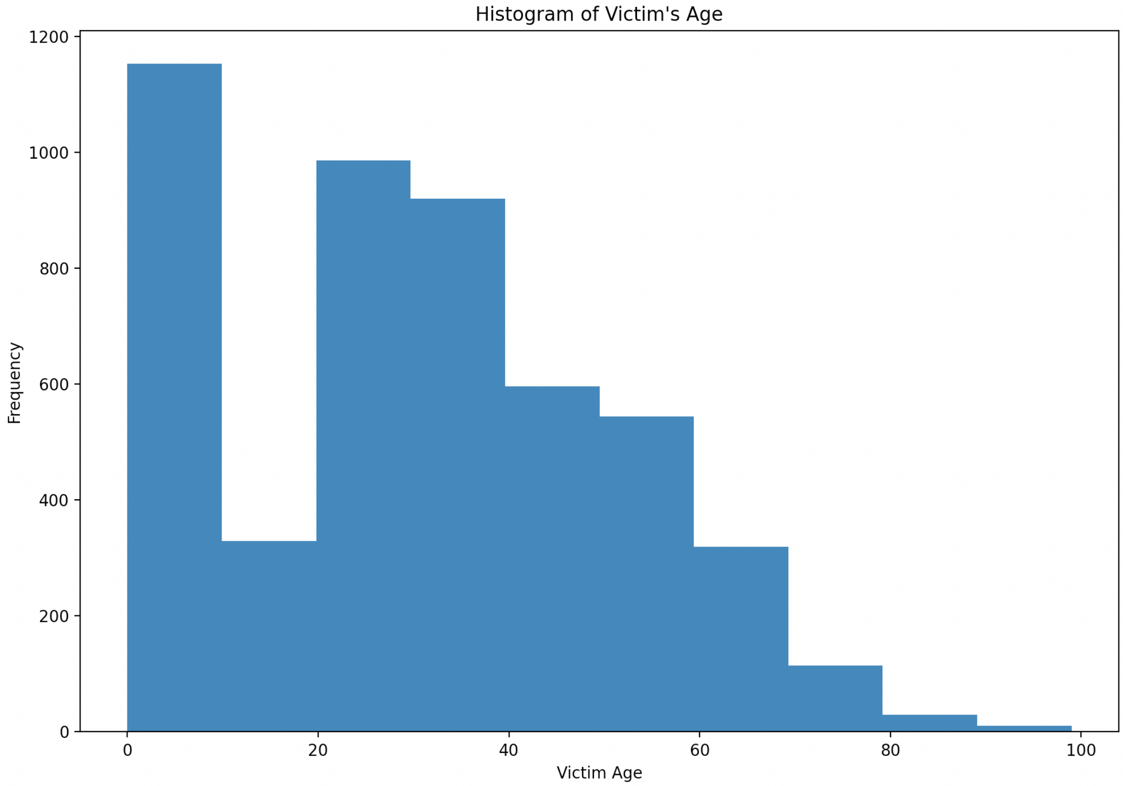
Analysis performed:

Using the first dataset (crime_dataset), I have performed an analysis on victim's age. Crime dataset has a column called vict_age which contains the age of victims associated with a particular crime. I have calculated several statistical measures (like mean, median, standard deviation etc) for the entire column. Moreover, I have plotted a histogram showing which age groups are targeted the most.
From the histogram, it is clearly evident that people belonging to age groups 0-10 and 20-40 are targeted the most.

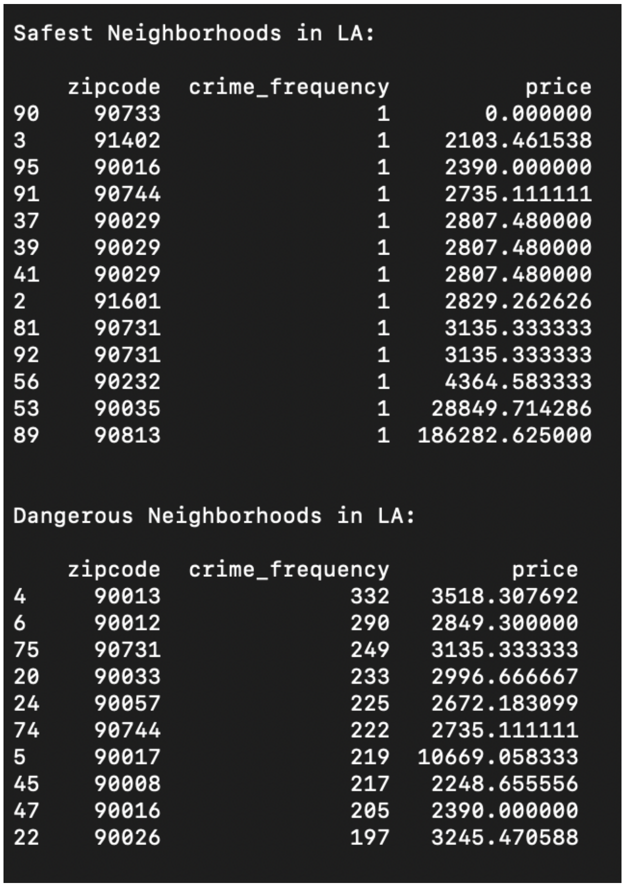
Following are the screenshots of the output obtained.



```
Mean of Victim's Age: 29.9752
Median of Victim's Age: 30.0
Minimum Value of Victim's Age: 0.0
Maximum Value of Victim's Age: 99.0
Standard Deviation of Victim's Age: 21.46200794334025
```



From the second dataset (dataset_zipcode), that gives crime_frequency corresponding to every zip code, I have generated a table of safest neighborhoods (crime frequency = 1) and dangerous neighborhoods (top 10 highest crime frequency). The average price of all 2b2b apartments in that neighborhood, obtained from Craigslist website, is also displayed.



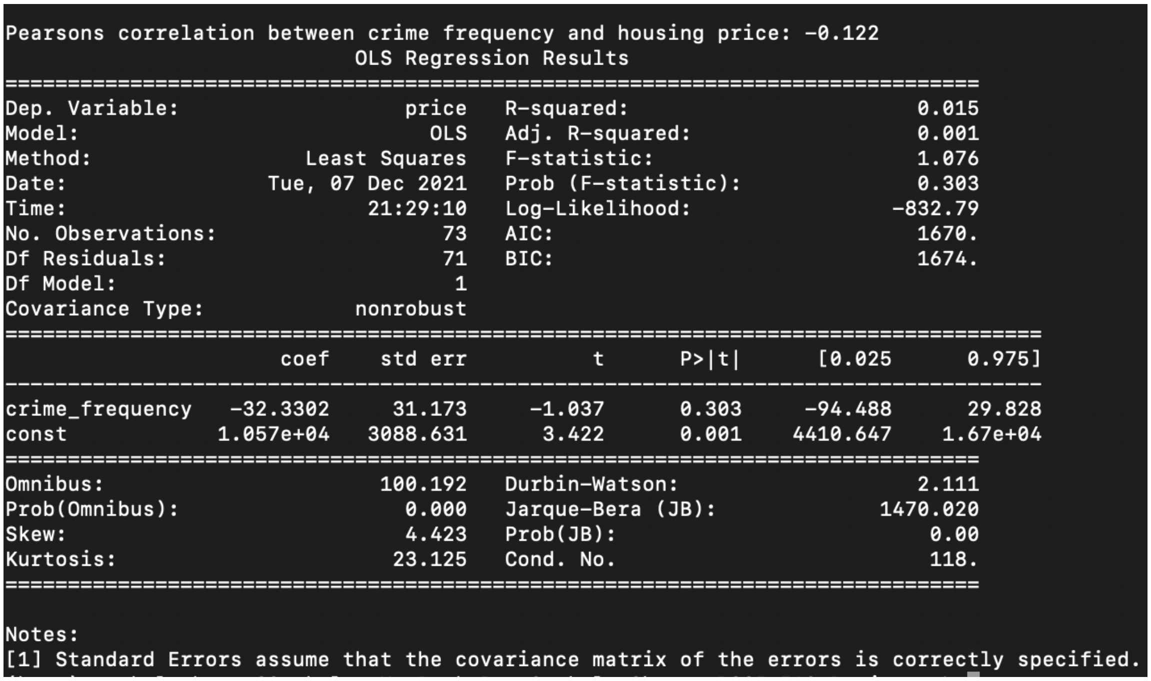
```
Safest Neighborhoods in LA:

    zipcode  crime_frequency          price
90    90733                1       0.000000
3     91402                1    2103.461538
95    90016                1    2390.000000
91    90744                1    2735.111111
37    90029                1    2807.480000
39    90029                1    2807.480000
41    90029                1    2807.480000
2     91601                1    2829.262626
81    90731                1    3135.333333
92    90731                1    3135.333333
56    90232                1    4364.583333
53    90035                1   28849.714286
89    90813                1  186282.625000


Dangerous Neighborhoods in LA:

    zipcode  crime_frequency         price
4     90013              332   3518.307692
6     90012              290   2849.300000
75    90731              249   3135.333333
20    90033              233   2996.666667
24    90057              225   2672.183099
74    90744              222   2735.111111
5     90017              219  10669.058333
45    90008              217   2248.655556
47    90016              205   2390.000000
22    90026              197   3245.470588
```

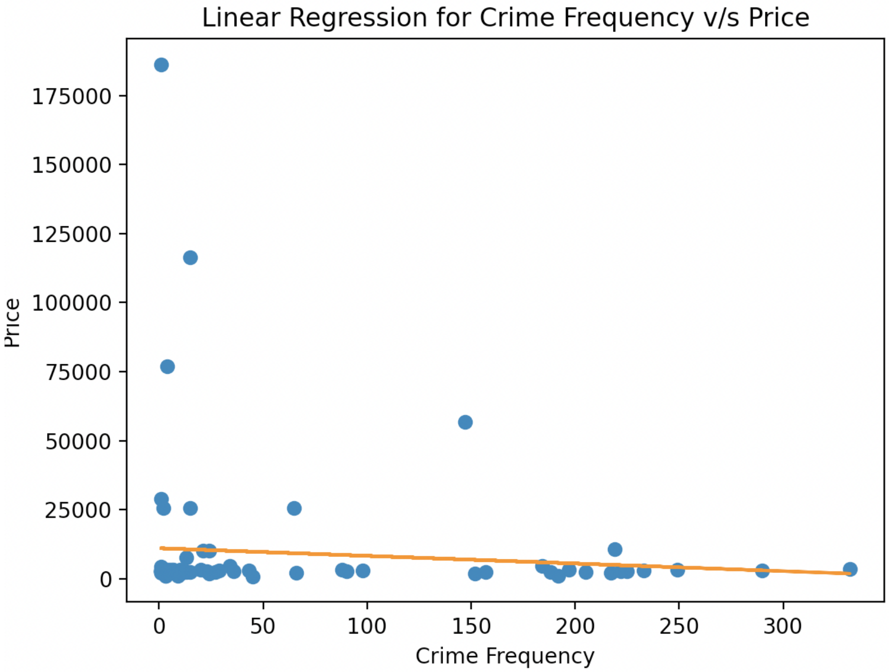
Finally, I have used my third dataset to analyze if there is any correlation between crime frequency and apartment prices (2b2b) or not. Following is the screenshot of the statsmodel output when applied to our data.



```
Pearsons correlation between crime frequency and housing price: -0.122
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.015
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.076
Date:                Tue, 07 Dec 2021   Prob (F-statistic):              0.303
Time:                        21:29:10   Log-Likelihood:                -832.79
No. Observations:                  73   AIC:                             1670.
Df Residuals:                      71   BIC:                             1674.
Df Model:                           1
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
crime_frequency   -32.3302     31.173     -1.037      0.303     -94.488      29.828
const            1.057e+04   3088.631      3.422      0.001    4410.647    1.67e+04
==============================================================================
Omnibus:                      100.192   Durbin-Watson:                   2.111
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1470.020
Skew:                           4.423   Prob(JB):                         0.00
Kurtosis:                      23.125   Cond. No.                         118.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

As seen above, p value for crime_frequency is more than 0.05 (general threshold) which proves that the crime_frequency is statistically insignificant while predicting monthly rental prices of apartments. Applying statsmodel is basically fitting linear regression to data. In the first line of output image, it is shown that pearson’s coefficient for correlation between crime_frequency and housing price is slightly negative (approx -0.122) which slightly does indicate that crime negatively affects housing price (if crime in an area is more, then monthly rentals in the area is less). But the value of -0.1 is not sufficient enough to strongly propose this fact. Below is the screenshot of linear model fit to data.



Conclusion:

From the analysis, it can be concluded that crime frequency in a specific neighborhood does not have a significant impact on the apartment rental price in that neighborhood. There is a slight negative correlation (Pearson Coefficient = -0.122) which does support the fact that areas where crime is frequent will have low priced rental apartments. This conclusion is made based on the limited data available and also considering 2b2b apartments from one particular website (Craigslist LA). The results may vary if more data sources were taken into consideration. This is one way to further extend this project.