Name: Mingcan Li

DSCI 510 HW5 Project Description

Description: I produced a dataset in the form of a csv file named 'avg price.csv' for the final project. The Data Sources Section lists two separate data sources that were used to construct this dataset. The intermediate csv files ‘my file.csv' and 'zipcode.csv' are what lead to the production of the final dataset. The three datasets are included in this zip package. I investigated traffic collisions in various locations (zip codes) of Los Angeles and their link with home values in that zip code using these data.

Motivation: When I was an undergraduate, I studied the relationship between housing prices and price level changes from the perspective of time series. Housing prices are so fascinating to me that before I came to Los Angeles, I compared housing prices in various areas for renting. For me, the level of housing prices depends on the happiness of living here, and the happiness level is determined by things frequently experienced in daily life, such as the convenience of travel, the completeness of the facilities and the safety of the house. So, I decided to analyze the housing price and the convenience of transportation in this project. If there is a lot of traffic around the house, it will bring a lot of noise to the house, and how should the traffic flow be measured? I choose to use the number of traffic accidents to quantify the traffic flow. Large traffic flow means a greater probability of traffic accidents, and traffic accidents also mean traffic jams, and traffic jams will also affect the convenience of traffic. Therefore, this project conducts a correlation analysis on housing prices and the number of traffic accidents.

Data Sources:

1. Traffic Collision Data from 2010 to Present (https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w)

Link to download csv:

https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w/data

This csv is already downloaded and written into codes. I cleaned and filtered the data of the csv, and only kept the latitude and longitude as the key data and wrote it into my_file.csv. Because I just retrieved the traffic collision data from Jan 1, 2022, to Feb 28, 2022, this dataset consists of only 2550+ rows and 2 columns. The columns in the data set are as follows: the first column store the latitude of the collision location and the second column store the longitude of the collision location. Every row in the dataset to obtain the zip code of the collision location. First few rows of my_file.csv shown below.

myfile

34.0817	-118.317
34.1808	-118.4443
34.1802	-118.5076
33.9265	-118.2652
34.0703	-118.268
34.0488	-118.2817
33.7159	-118.3047
34.069	-118.2868
34.1017	-118.3123

2. Microsoft Bing Maps API

(http://dev.virtualearth.net/REST/v1/Locations/47.64054,-122.12934?&key=AtwP2GlmjNkx5aHfJwdZkQa4WJZl2w3WjAYbDmshUmmaEbHhOny1ZdKzQUBH7KR2)

Using reverse geocoding, from Microsoft Bing Maps, I have retrieved the zip codes from latitude and longitude in my_file.csv. I have created the second dataset called ‘zipcode’. There is only one column which is zip code for all the location in my_file.csv. Moreover, I used this dataset to create a dictionary, which recorded the top 10 most frequency of collision and the zip codes of happened area. Here are few rows of zipcode.csv:

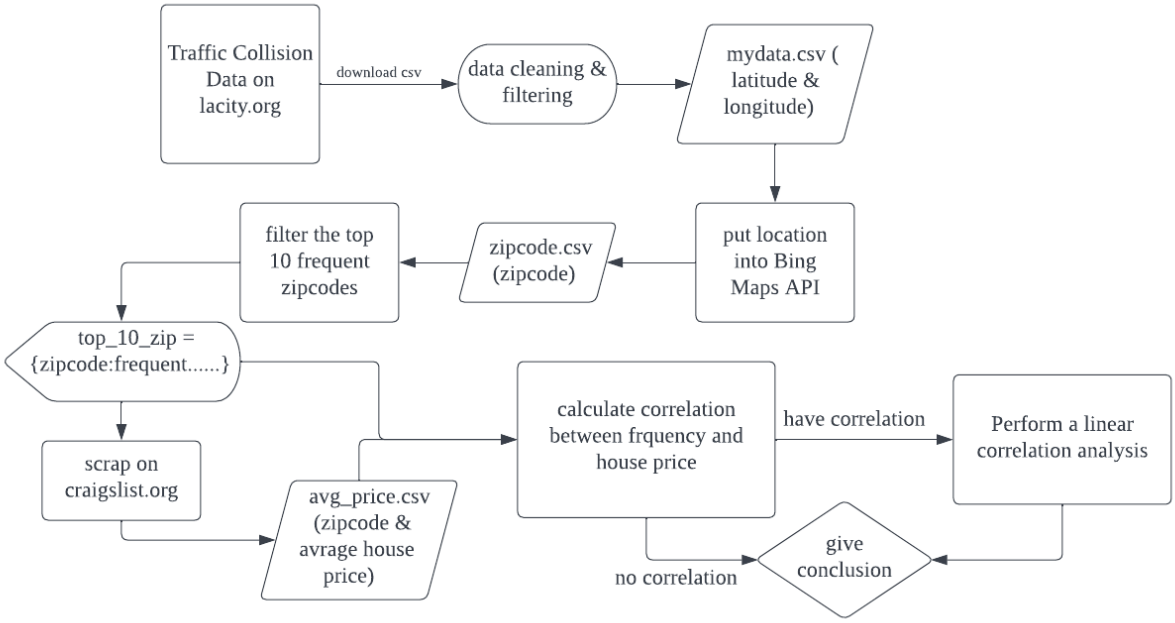
zipcod
90004
91401
91316
90061
90026
90006

3. Craigslist House Price

I'm scraping monthly rent costs of all 2 Bedroom and 2 Bathroom apartments using the zip codes in the above dictionary, then calculating the average. Furthermore, I saved them in a separate dataset called 'avg price,' which contains two columns: the first is zip codes, and the second is the average price of a 2b2b flat in that location. A few rows from avg price.csv are as follows:

avg_price	
90044	**2858.4979079497900**
90003	3080.638655462190
90011	3527.675
90037	3450.2916666666700
91331	2180.9583333333300
91342	2254.196581196580
91335	2808.185758513930

FLOWCHART FOR DATASET GENERATION AND ANALYSIS:



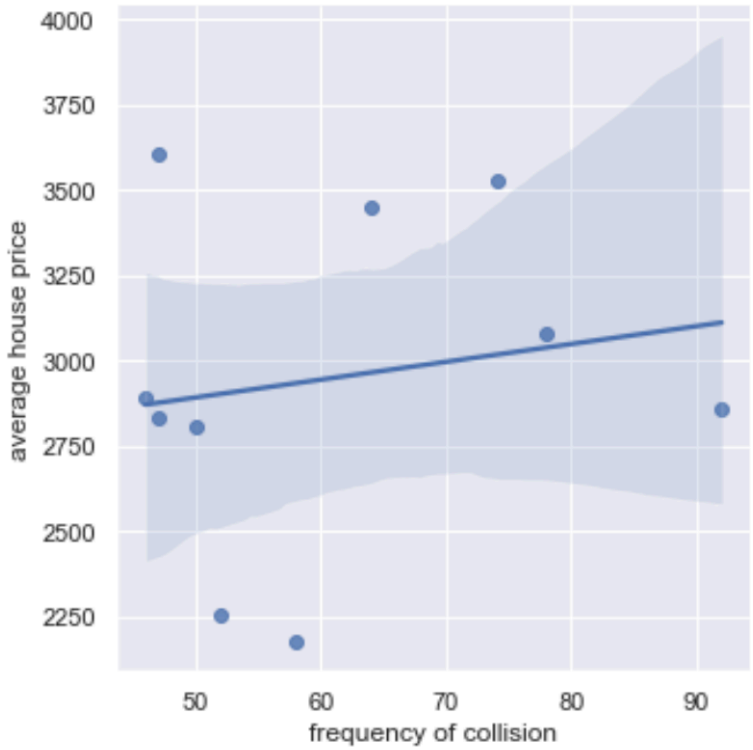
Analysis performed:

I estimated the correlation coefficient of traffic collision and apartment prices using the final dataset (avg price.csv) and the top 10 zip dictionary (2b2b). The outcome of statsmodel when applied to our data is seen in the image below.

```
The correlation coefficient for traffic collisions and house prices is 0.10938019148126638
/Users/mingcanli/opt/anaconda3/lib/python3.9/site-packages/scipy/stats/stats.py:1541: UserWarning: kurtosistest only valid for n>=
20 ... continuing anyway, n=10
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.012
Model:                            OLS   Adj. R-squared:                 -0.112
Method:                 Least Squares   F-statistic:                   0.09687
Date:                Tue, 10 May 2022   Prob (F-statistic):              0.764
Time:                        22:11:59   Log-Likelihood:                -41.222
No. Observations:                  10   AIC:                             86.44
Df Residuals:                       8   BIC:                             87.05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         51.3678     30.761      1.670      0.133     -19.568     122.304
x1             0.0032      0.010      0.311      0.764      -0.021       0.027
==============================================================================
Omnibus:                        2.468   Durbin-Watson:                   0.220
Prob(Omnibus):                  0.291   Jarque-Bera (JB):                1.260
Skew:                           0.855   Prob(JB):                        0.533
Kurtosis:                       2.680   Cond. No.                     1.72e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

As seen above, the p value for traffic collision frequency is more than 0.05, indicating that the frequency of traffic collisions is statistically insignificant in predicting monthly rental pricing of 2b2b flats. Fitting linear regression to data is what statsmodel is all about. The Pearson's coefficient for the connection between traffic collision frequency and apartment rent prices is somewhat positive (approx. 0.11) in the first line of the output graphic, indicating that traffic collisions have a positive impact on housing costs. However, 0.11 is insufficient to strongly suggest this truth. A snapshot of the linear model fit to data is shown below.



Conclusion:

From the analysis, it can be concluded that traffic collision frequency in a specific neighborhood does not have a significant impact on the apartment rental price in that neighborhood. There is a slight positive correlation (Person coefficient = 0.11) which does support the fact that areas where traffic collision is frequent will have high priced rental apartments. This conclusion is made based on the limited data available and considering 2b2b apartments from one particular website (Craigslist LA). The results may vary if more data sources were taken into consideration. At the same time, there may also be more

interference factors brought about by the increase in data, resulting in a less significant correlation. In the future, we may increase the number of samples to conduct further research on the correlation between the two.