**Name:** Mingcan Li **USC ID:** 1706-6792-66

**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.

Table

Description automatically generated

1. 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:

Table

Description automatically generated

1. 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:

Table

Description automatically generated

**FLOWCHART FOR DATASET GENERATION AND ANALYSIS:**

Diagram

Description automatically generated

**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.Table

Description automatically generated

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.

Chart, scatter chart

Description automatically generated

**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.