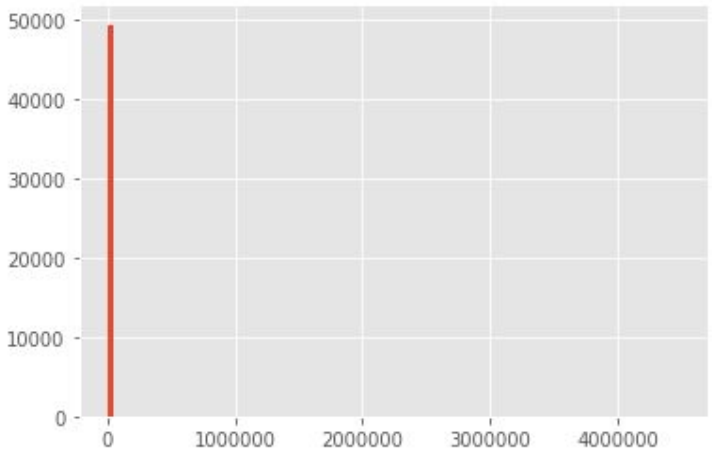
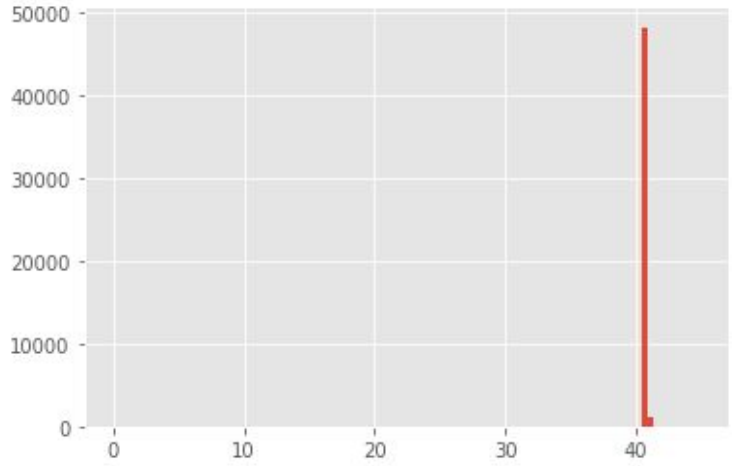
**Course project: Milestone 1**

**Github link: https://github.com/1349884366/cmpt459/tree/master**

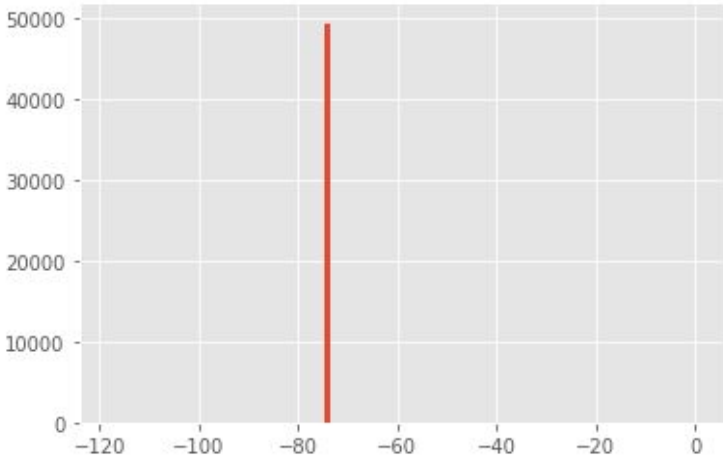
**Exploratory data analysis**



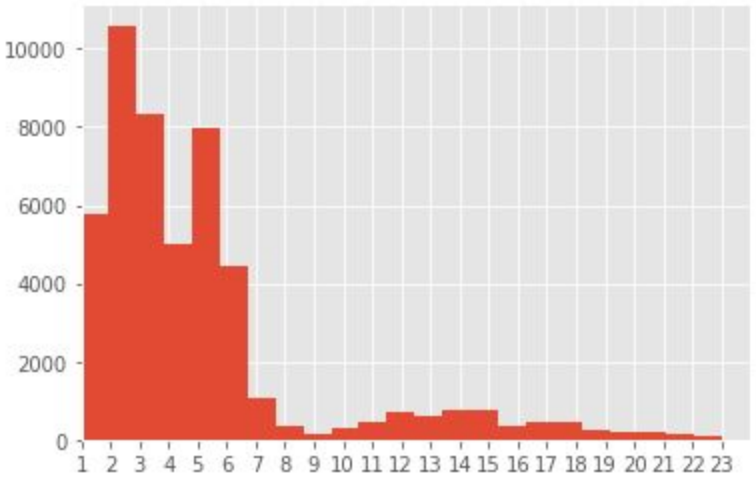

Price Histogram



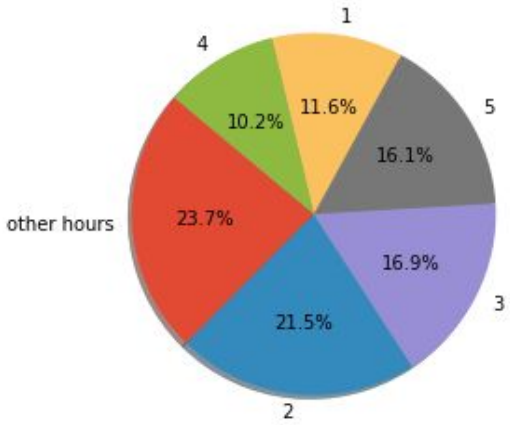

Latitude Histogram




Longitude Histogram




Hour Listing Trend




The busiest hours are 1, 2, 3, 4 and 5 o'clock.

proportion of 5 busiest hours and the other hours

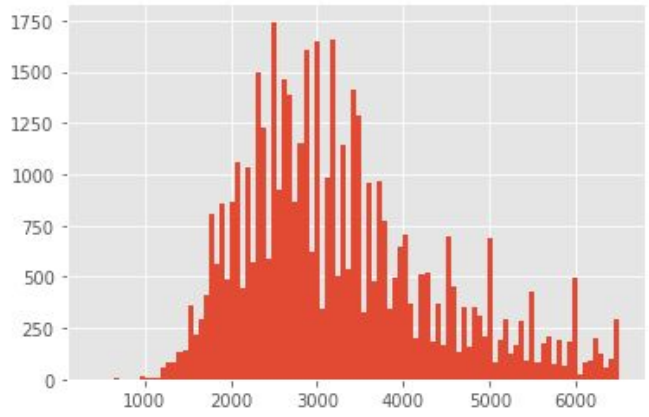
**Dealing with missing values, outliers**

Missing value count

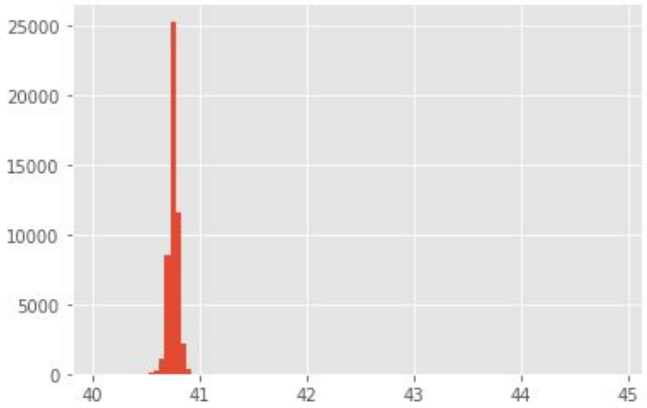
| feature | display_address | description | building_id | latitude | longitude | street_address |
|---|---|---|---|---|---|---|
| 3218 | 135 | 1446 | 8286 | 12 | 12 | 10 |

Outliers count

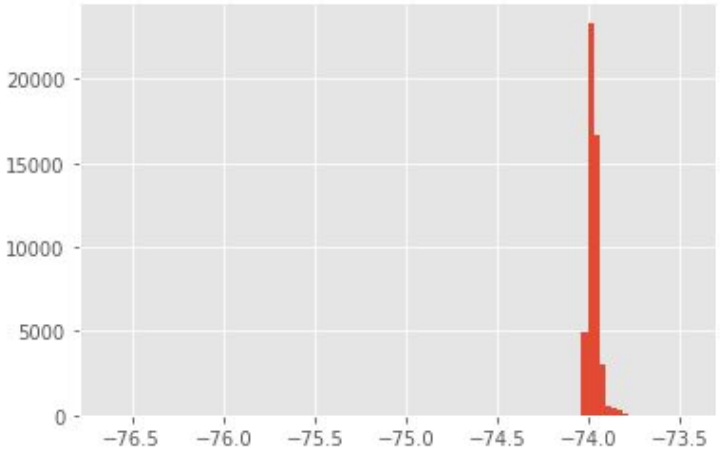
| outlier_price | outlier_latitude | outlier_longitude |
|---|---|---|
| 2788 | 15 | 29 |



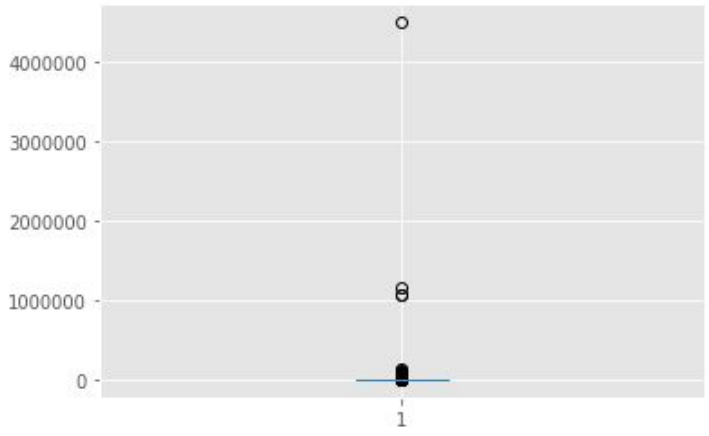

Price Histogram without Outliers



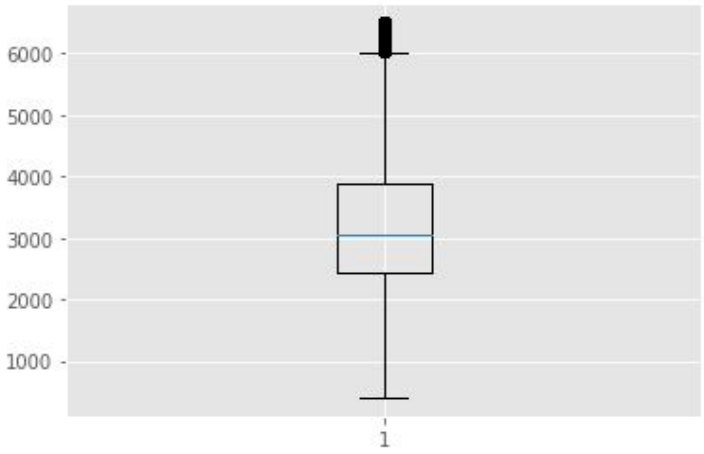

Latitude Histogram without Outliers



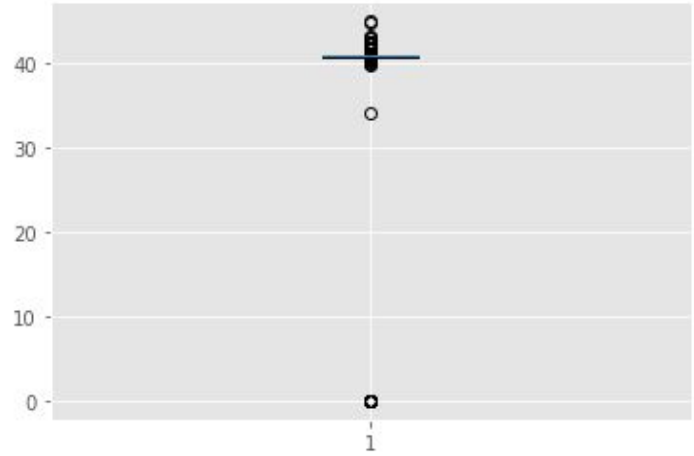

Longitude Histogram without Outliers



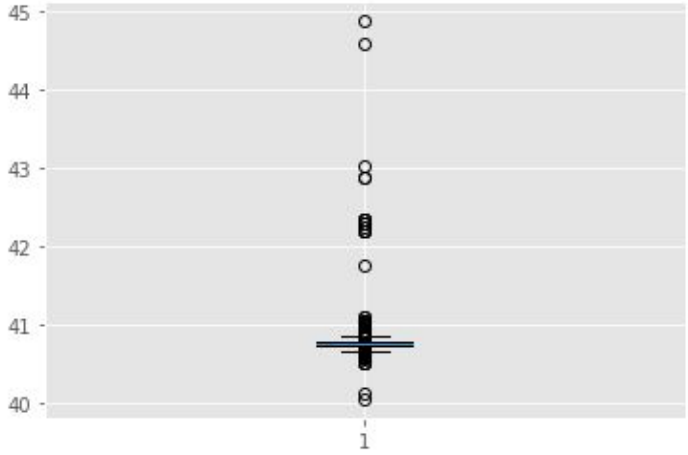

Price Boxplot with Outliers



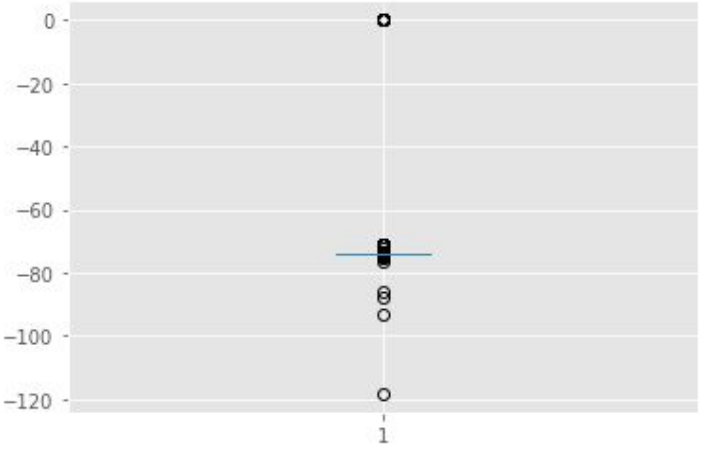

Price Boxplot without Outliers



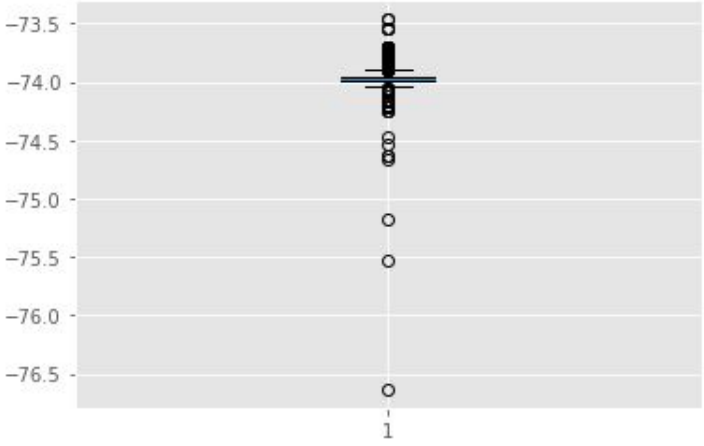

Latitude Boxplot with Outliers




Latitude Boxplot without Outliers




Longitude Histogram with Outliers




Longitude Histogram without Outliers

**Discussion about outliers**: We only remove outliers in price, latitude and longitude. For price, some of them are as much as 90000 and as little as 40. These prices are extreme values which are not realistic in the real world and affect the procedure of exploring the overall trend, within which other prices' values are significantly far from the outliers.
For latitude and longitude, the data was collected around New York but some locations of the listings are far from the latitude and longitude of New York. Thus, these data are considered outliers and should be removed.
For the other attributes, they are all descriptive terms and are not numeric values. They do not create noises when analyzing the overall trend. Thus, these data need not be removed.

**Discussion about missing values:** The missing values can not be safely dropped since this will affect the representativeness of the population on all attributes. As for numerical values, we filled missing values with the mean or mode values. For descriptive values, we marked them 'unknown.'

### Feature extraction from images and text

For the text documents, sklearn.feature_extraction.text library was used to count the numbers and frequencies of vocabularies, and the fetched data was stored dynamically as the program runs. For images, the function feature.hog was used. This function produces the histogram of texture. Due to the large size of the extracted features from images, these data were stored in local desktops as .npy file for future uses.