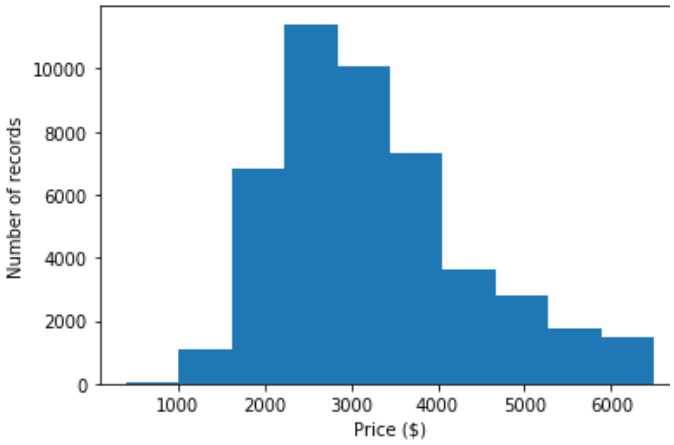
CMPT 459 – Milestone 1

Insoo Rhee
Jin Young Kim
David Sun

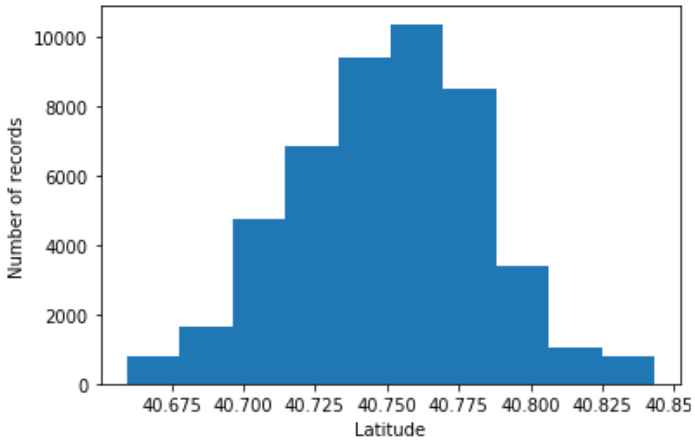
Exploratory Data Analysis

In our exploratory data analysis, we investigated the value ranges corresponding to high numbers of records by creating histograms. The attributes investigated are "price", "latitude", and "longitude", which are important determinants of a listing's popularity. The appropriate range of the independent variables was determined by an outlier analysis.

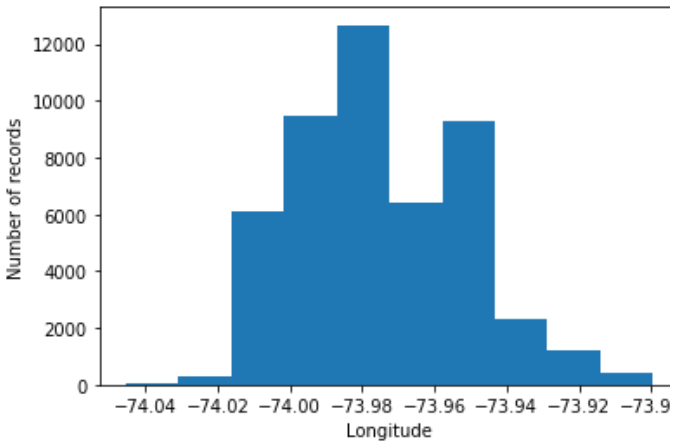
1. Price



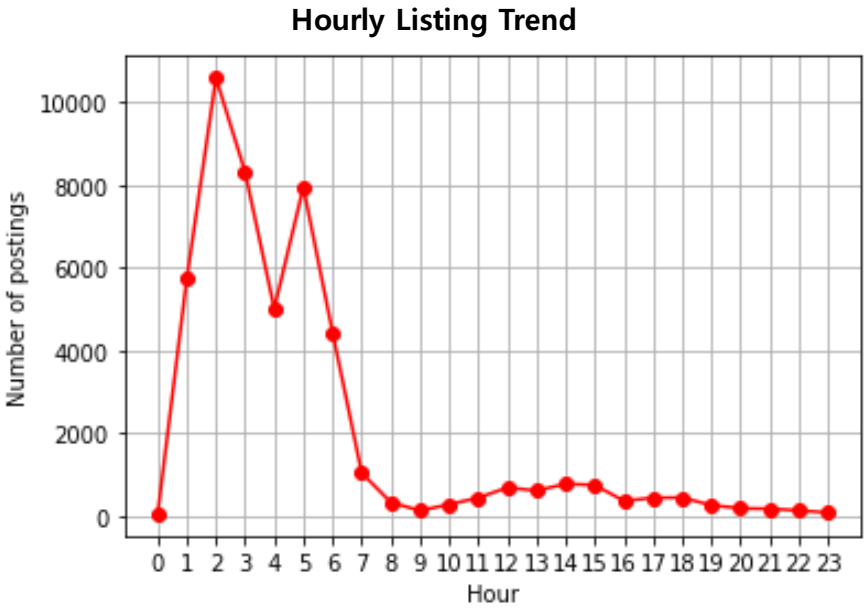
2. Latitude



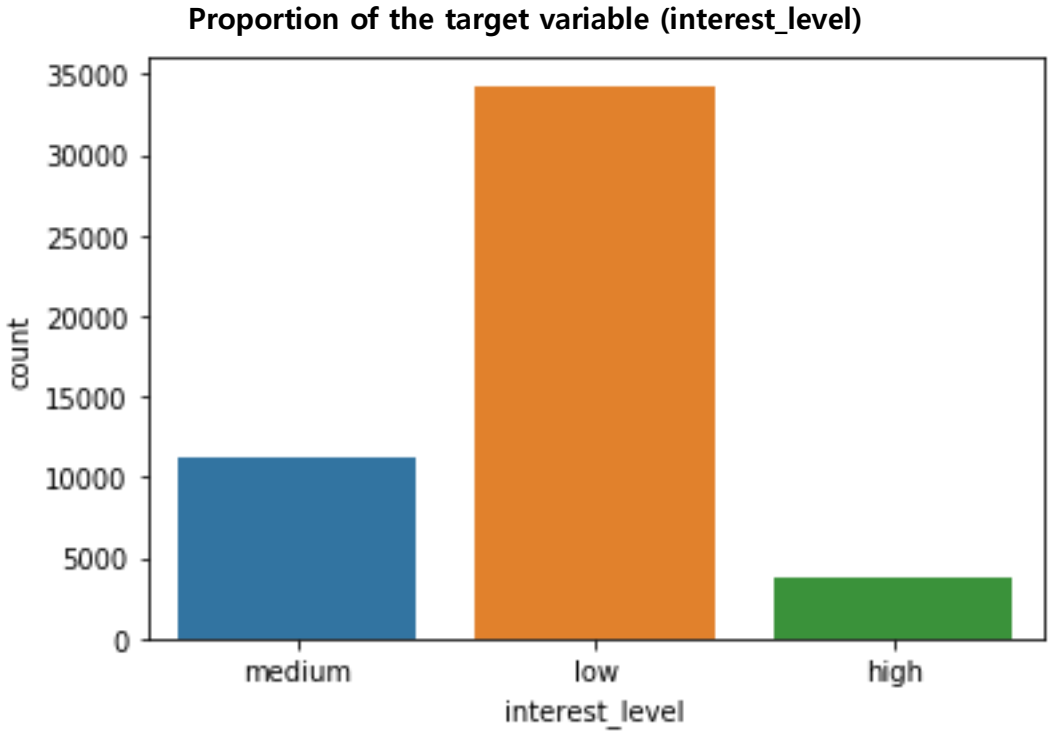
3. Longitude



Shown below is a graph of an hourly listing trend. The top five busiest hours of posting are from 1 a.m. to 5 a.m.



Below is a histogram showing the proportion of different interest levels. Most postings (~34000 posts) have low popularity or interest levels.



Missing Values

Since missing values varied in forms between the attributes (e.g., missing values are represented as an empty list for "features"), we first identified these forms for each attribute. This identification was done by listing the records by their unique values for the given attribute, and then by either manually looking through the list or further sorting it in ascending order to detect any missing value. To ensure no missing values were overlooked, a further step was performed in which all possible forms of a missing value (None, " ", [ ], and 0) were searched in each column. The following results were obtained:

Table 1. Missing value count for each attribute.

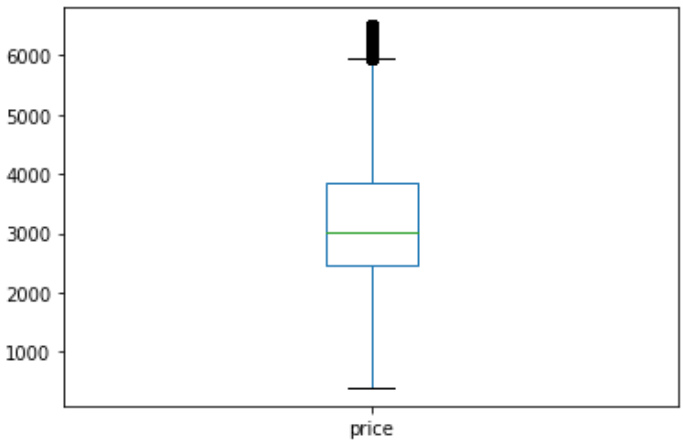
Attribute	Missing value type	Missing value count
Bathrooms	-	0
Bedrooms	-	0
Building_id	Integer 0	8,286
Created	-	0
Description	Empty string " "	3,332
Display_address	-	0
Features	Empty list [ ]	3218
Latitude	-	0
Listing_id	-	0
Longitude	-	0
Manager_id	-	0
Photos	Empty list [ ]	3,615
Price	-	0
Street_address	-	0
Interest level	-	0

Missing values were detected for "building_id", "description", "features", and "photos". The missing values for these attributes cannot be imputed as they are nominal (building_id) or qualitative (rest). Given the large dataset size of ~50,000 records, it is safe to drop the missing values as we will still have a large enough sample to data to produce statistically meaningful results.

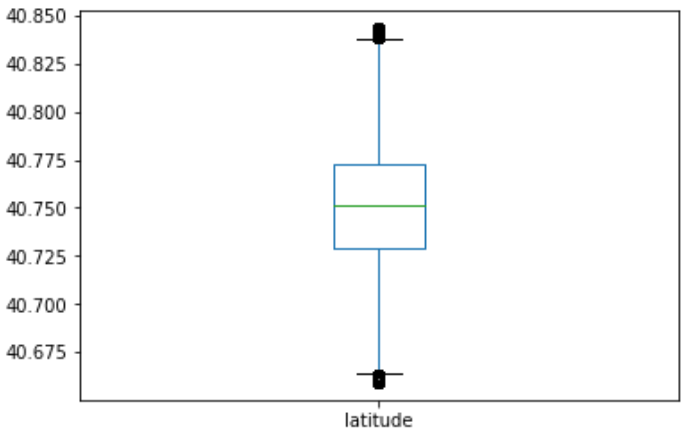
Outliers

Outliers were determined using the Interquartile Range Rule. The rule determines the upper and lower limits for outliers by adding IQR*1.5 to the third quartile(Q3) and subtracting IQR*1.5 from the first quartile (Q1), where IQR = interquartile range = Q3-Q1. The upper and lower limits are represented by the ends of the whiskers coming out of the box. Any data points that lie outside the whiskers are considered outliers. Below are boxplots of quantitative attributes: price, latitude, longitude, bathrooms, and bedrooms.

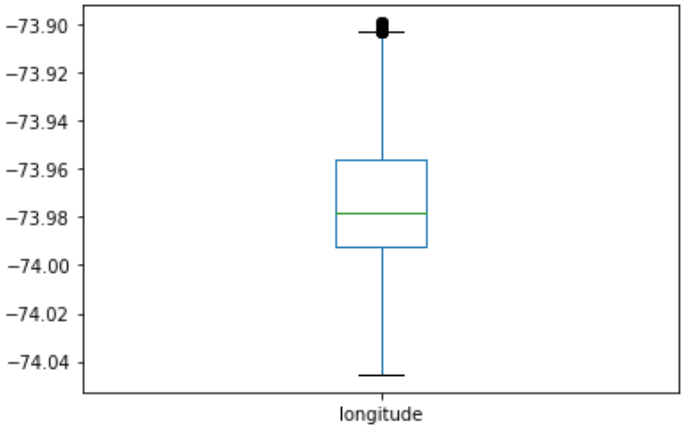
1. Price – Outlier count: 2,964



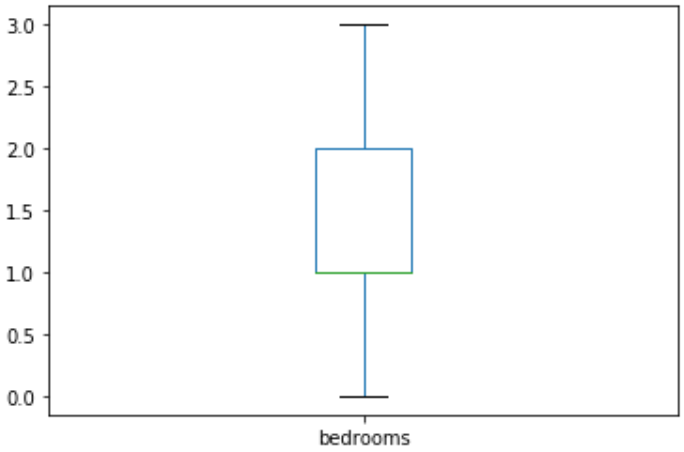
2. Latitude – Outlier count: 1,948



3. Longitude – Outlier count: 1,102



4. Bedrooms – Outlier count: 2,226



5. Bathrooms – Outlier count: 9,940

Outlier detection is not meaningful for all the other attributes including features, street_address, and description because the attributes are nominal or qualitative data that are highly variable in terms of their values. Attributes such as features and description are highly variable qualitative data with little to no common values between records. For instance, it is rare to find multiple listings having the exact same description.

Extract Features from the Text Data

We formed a matrix containing the counts of meaningful words in our text data from the "features" and "descriptions" columns using the nltk library. We pre-processed the text data by filtering out commonly used words (e.g., "the", "a", "an", and "in"). This resulted in a set of meaningful words from the records. The following is an example set resulting from pre-processing 100 data points:

{'access', 'allowed', 'balcony', 'building', 'cats', 'center', 'common', 'concierge', 'construction', 'deck', 'dining', 'dishwasher', 'dogs', 'doorman', 'duplex', 'elevator', 'exclusive', 'fee', 'fireplace', 'fitness', 'floors', 'furnished', 'garage', 'garden/patio', 'hardwood', 'high', 'internet', 'laundry', 'live-in', 'loft', 'lowrise', 'new', 'on-site', 'outdoor', 'parking', 'playroom', 'pool', 'pre-war', 'prewar', 'private', 'private-balcony', 'publicoutdoor', 'reduced', 'roof', 'room', 'simplex', 'space', 'speed', 'super', 'superintendent', 'swimming', 'terrace', 'unit', 'wheelchair', 'wifi'}

Unlike "features", whose records are a list of words, descriptions are freely written sentences, thus using a wider range of words. The following is an example a set of words from 'descriptions' from 10 data points.

{'#', "''", "'ll", "'m", '**flex', '**mid', '-', '/dryer', '032-568-9993', '06/26/16', '064-692-8838email', '1', '1.5', '12', '132', '14ft', '``', abundance', 'abundant', 'ample', 'another', 'antique', 'apartment', apartmentenjoy', 'bars', bath', 'bathroom', 'bathroom-', 'beakexclusive', 'beautiful', 'bed', 'bedroom', bedroomfind', 'bedrooms', 'big', 'bike', 'blocks', 'blue', 'bond', 'br', 'brand', 'case', 'ceilings', 'center', 'centra', 'central', 'check', 'cheese', 'chef', 'city', 'client', 'closet', 'closet/storage', 'closets', 'coffee', 'combines', 'come', 'companion', 'complete', 'concierge', 'confident', 'contact', convenience', 'convenient', 'converted', 'cook', 'corporate', 'counter', 'counters', 'cream', 'create', 'cycle', 'days', 'dazs', 'deal', 'decorative', 'deep', 'dellarocco', 'derrick', 'designed', 'designs', 'different', 'dining', 'dishwasher', 'dishwasher-', 'distance', 'dog', 'door', 'doorman', 'doormanelevatornewly', 'dozens', 'dramatic', 'duplex', 'e', 'east', 'eat', 'eat-in-kitchen', 'edan', 'effective', 'elevator', 'email', 'enough', 'entire', 'epicenter', 'equal', 'forest', 'free', 'freshly', fridge', 'full', 'fully', 'g', 'garage', 'gas', 'ginger', 'good', 'home.call/text', 'hookah', 'hot', 'hour', 'living', 'lobby', 'lobby.', 'local', 'located', 'location', 'lot', 'lots', 'lounge', 'nearby', 'need', 'neighborhood', 'net', new', 'newly', 'nightlife', 'nordstrom', 'noteworthy', 'oak', 'occasions', 'offering', 'offers', 'office', 'old', 'omane', 'washer/dryer', 'water', 'website_redacted', 'well', 'whole', 'williamsburg', 'windows', 'wine', 'within', 'wonderful', 'wooden', 'write', 'yard', 'ymca', 'yoga', 'york'}

Image Processing

The image data are jpg images of listed rooms. We extracted histogram distributions of the RGB channel, brightness (Y value), and visual words/features from the images.

Histogram of RGB channel

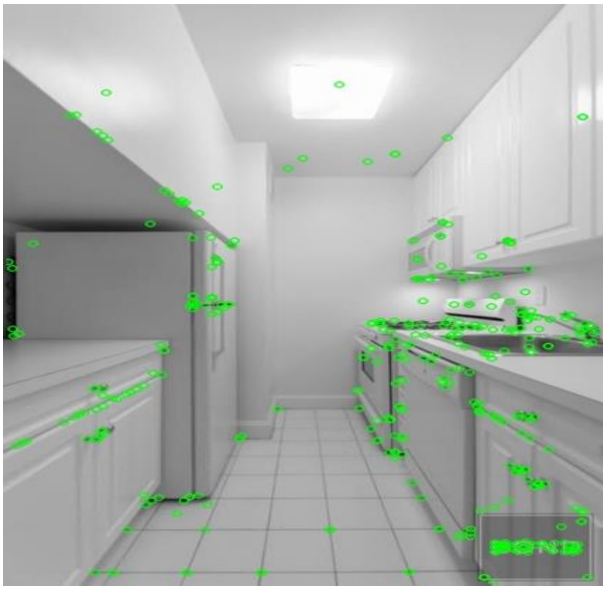
Using RGB information, we can identify patterns in the images that allow us to predict listing prices. For example, a room with wooden floor (wood color) might have a higher price than one with concrete floor (grey color).

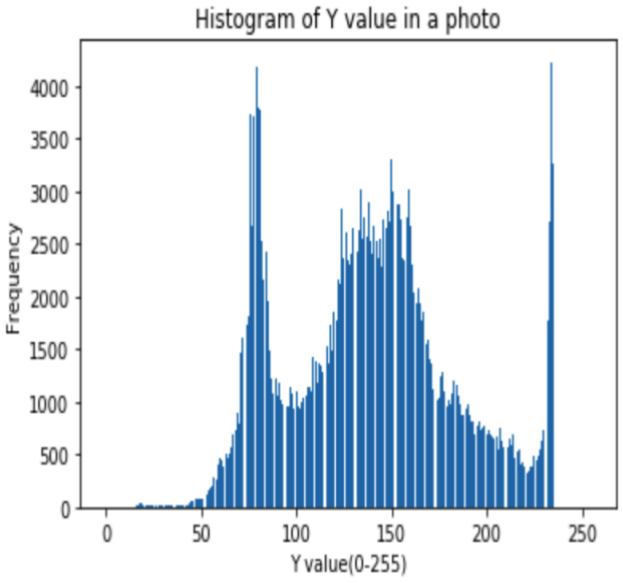
Brightness

We convert the RGB color space into the YUV color space to obtain the Y-value (brightness) of each pixel. Brightness carries information regarding room location and orientation (a brighter room might suggest better sunlight/better lighting) and may suggest a better price.

Visual Words

We used OpenCV's Bag of Visual Words library to extract features from images: it processes image data and returns an array of numbers for each detected feature (green circles in the image below). In the next milestone, clustering analysis will be performed on all green circles to generate "Visual Words". Then the histogram of these "Visual Words" can help predict listing prices. For instance, stairs may suggest that a room is double-decked and hence more expensive than single-decked rooms.





(E.g Features extracted using OpenCV)

Appendix

Fig. 1 Frequency of words extracted from ‘features’ from the first 100 data points

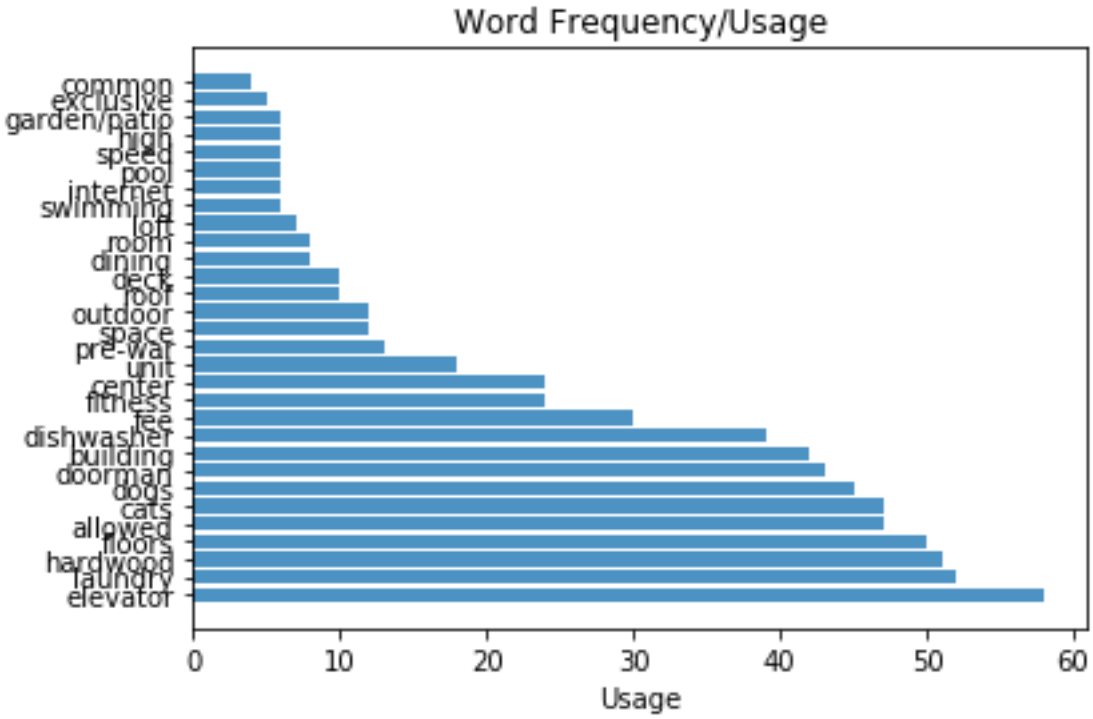


Fig. 2 Frequency of words extracted from ‘descriptions’ from the first 50 points

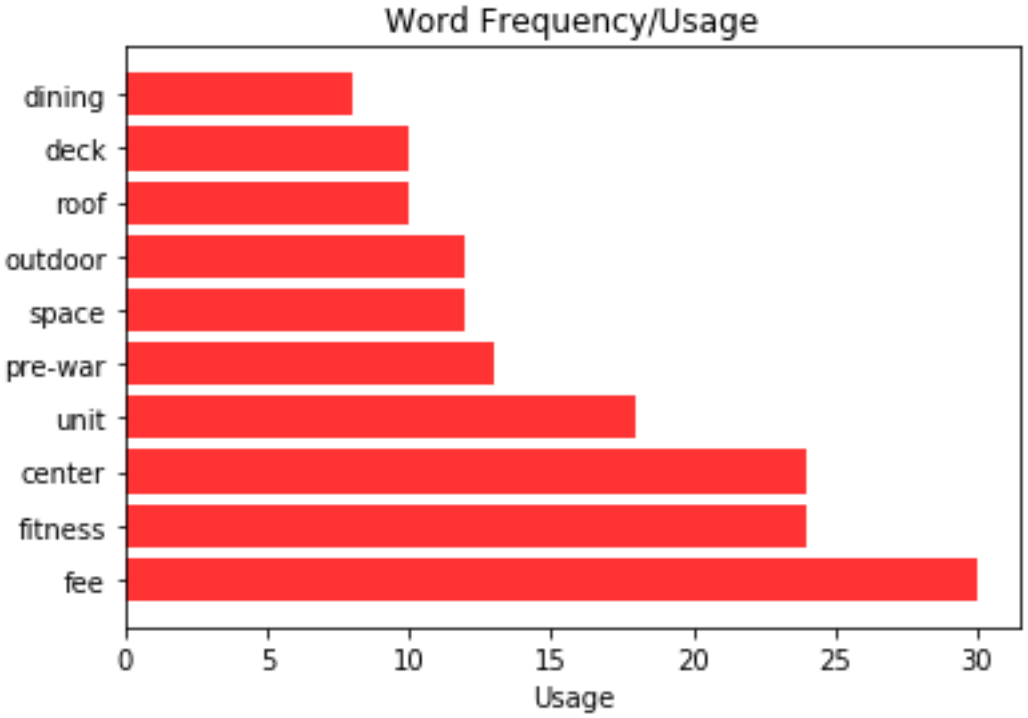


Fig. 3 Locations of minimum and maximum of longitude and latitude after removing outliers