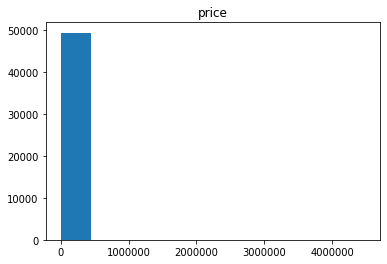
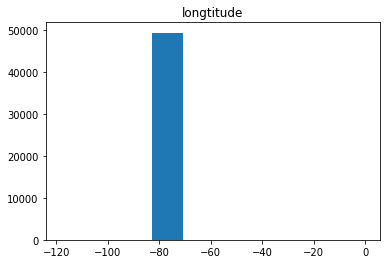
Code repo: <https://github.com/DekaiLin/cmpt459project>

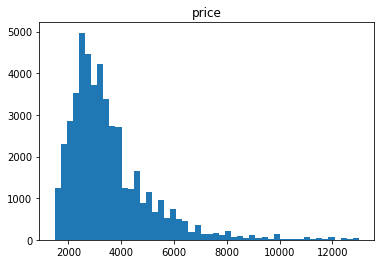
1. **Exploratory data analysis:**
2. The outliers affect the histogram, so we need to drop the heads and tails to avoid the outliers’ effect.

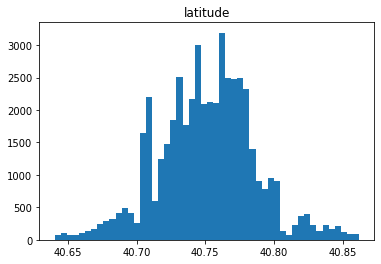


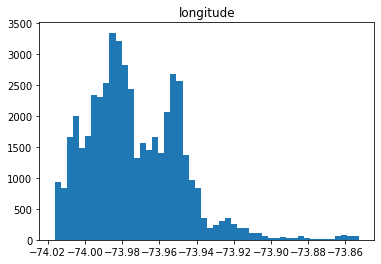




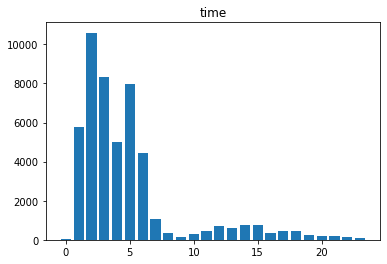
After dropping the outliers:



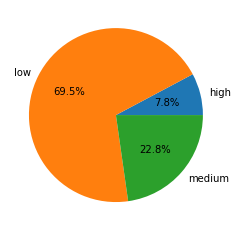




1. Top 5 busiest hours of postings {2, 3, 5, 1, 4}



1. Proportion of target variable values:



1. **Dealing with missing values, outliers：**
2. The number of each missing values in each variables is shown in the following chart.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Number of Missing Values | Description |
| bedrooms | 0 |  |
| bathrooms | 0 |  |
| building\_id | 8286 | Building id is 0 |
| created | 0 |  |
| description | 1685 | No description |
| display\_adderss | 137 | No address |
| features | 3218 | No features |
| latitude | 12 | Latitude is 0 |
| listing\_id | 0 |  |
| longitude | 12 | Longitude is 0 |
| manager\_id | 0 |  |
| photos | 3615 | No photos |
| price | 0 |  |
| street\_address | 10 | No address |
| Interest\_level | 0 |  |

1. The number of outliers in some variables is shown in the following chart.

|  |  |
| --- | --- |
| Attribute Name | Number of Outliers |
| bedrooms | 0 |
| bathrooms | 0 (313 training data are 0 bathrooms) |
| latitude | 38 |
| longitude | 16 |
| price | 4 |

The latitude and longitude outliers can be removed because the data is got from New York. So the actual latitude and longitude can be manually obtained by the address.

For the bathrooms, 0 bathrooms seem to be a outlier. This need to be checked by human. So these values can not be removed.

For the price, the outlier values are unrealistic, so remove them.

For other not numerical attributes, they are not comparable, so outlier detection is not meaningful for them.

1. Building id can use default value (such as “unknown”) to fill in the missing value because this attribute has little effect to classification.

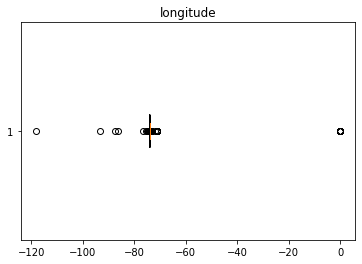
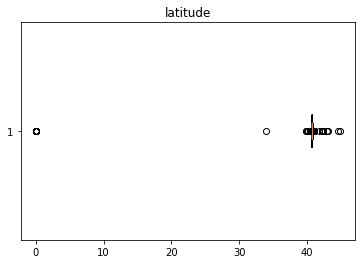
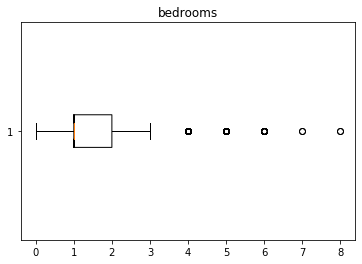
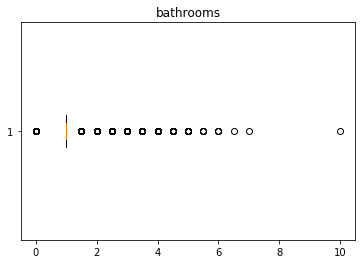
The missing value of description and features attribute can be supplemented by cross-references. If both of them are missing, the features can be extracted from photos.

For the latitude and longitude, we can replace the missing value by the New York’s latitude and longitude.

For the display address and street address, we can get the address by latitude and longitude.

For the photos, we can only drop the missing value because we have no way to impute the missing photos from other attirbutes.

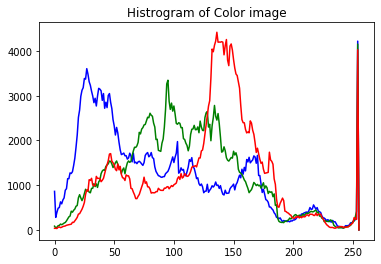
**Plot of the outliers:**



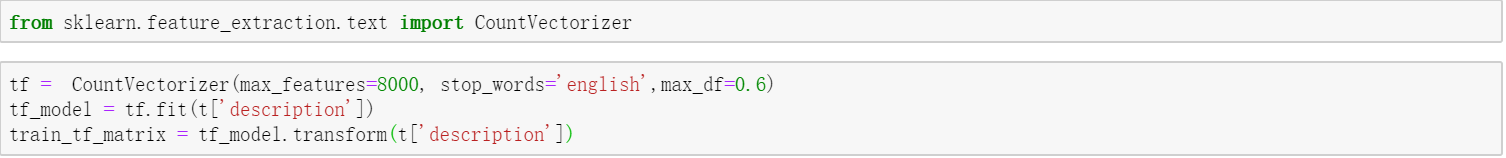
1. **Feature extraction from images and text:**
2. For the image data we used histograms of colors to extract features. For example,



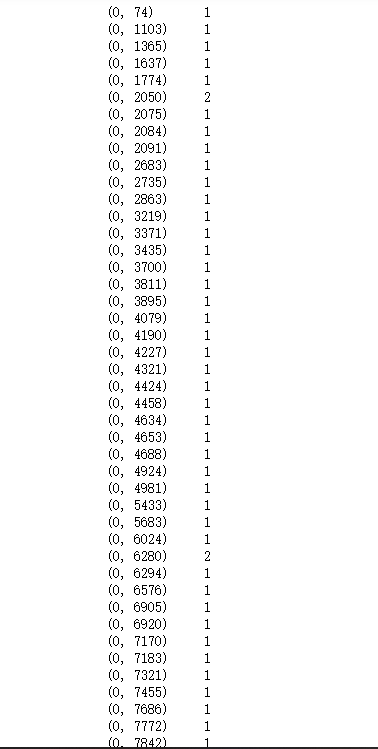
The colors histogram of this picture:



1. For the text data we used term frequencies as features. For example, we used 8000-dimension vector to map description text in term space. We drop the English stop words and some very frequent terms (like ‘<br> which is not in the English stop word list) by maximum document frequency = 0.6



This is the corresponding document vector for a sample training data.



Also, we find out the most frequent terms in the ‘description’ and ‘features’ attributes by the word cloud representation.



