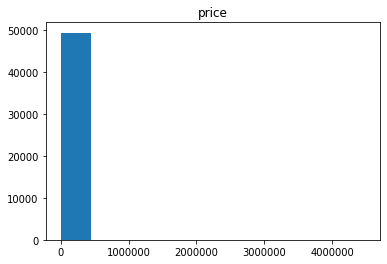
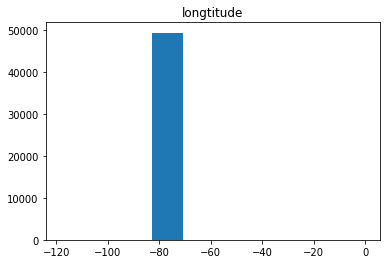
Code repo: <https://github.com/DekaiLin/cmpt459project>

**Exploratory data analysis:**

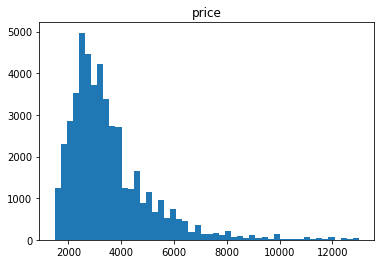
1. The outliers affect the histogram, so we need to drop the heads and tails to avoid the outliers’ effect.

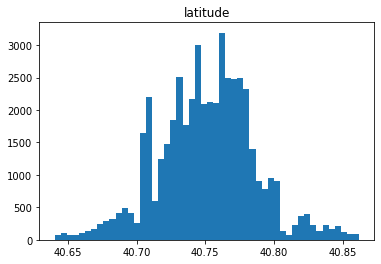


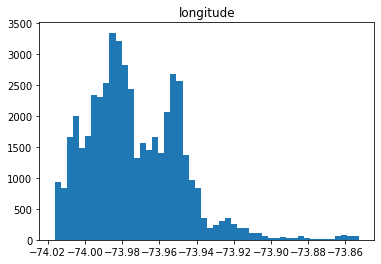




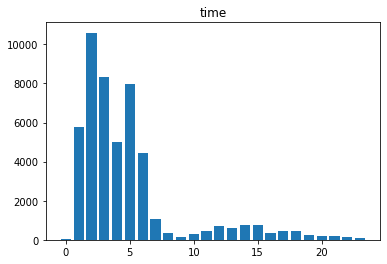
After dropping the outliers:







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



1. 这个题目没看懂，需要问老师确定。

**Dealing with missing values, outliers：**

1. The number of each missing values in each variables is shown in the following chart.

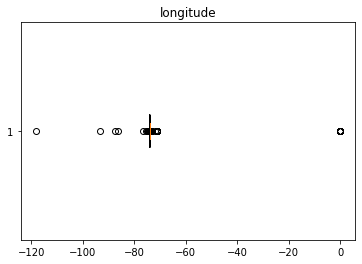
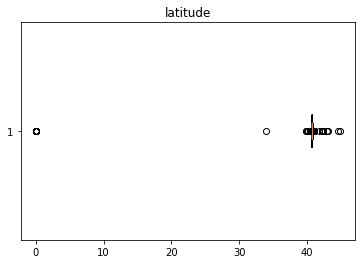
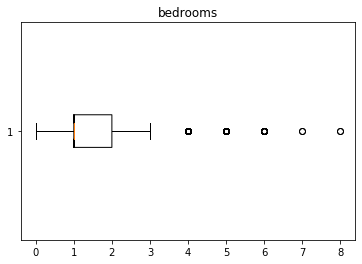
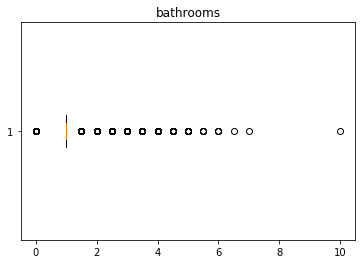
|  |  |  |
| --- | --- | --- |
| Attribute Name | Number of Missing Values | Description |
| bedrooms | 0 |  |
| bathrooms | 0 |  |
| building\_id | 8286 | Building id is 0 |
| created | 0 |  |
| description | 1685 | No description |
| display\_adderss | 137 | No address |
| features | 3218 | No features |
| 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 |
| latitude | 38 |
| longitude | 16 |
| price | 4 |

For interest\_level, all values are among high, medium and low.

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



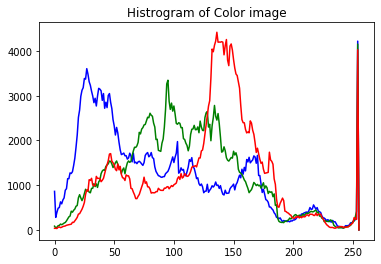
1. Yes

**Feature extraction from images and text:**

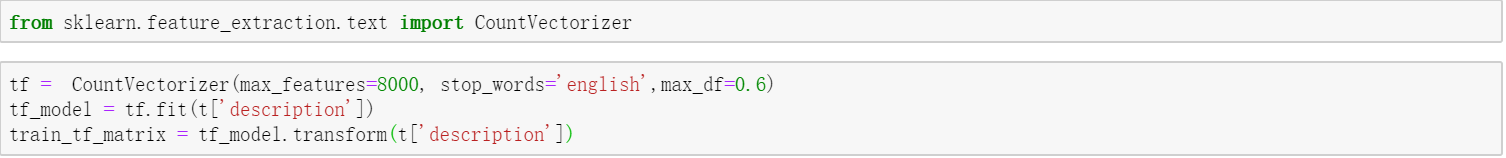
1. For the image data we used histograms of colors to extract features. For example,



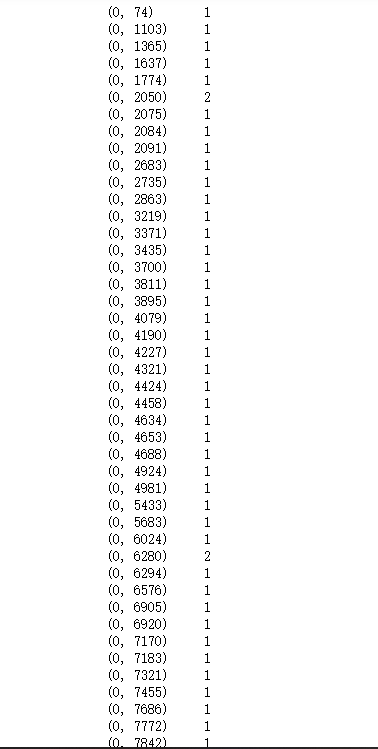
The colors histogram of this picture:



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.



