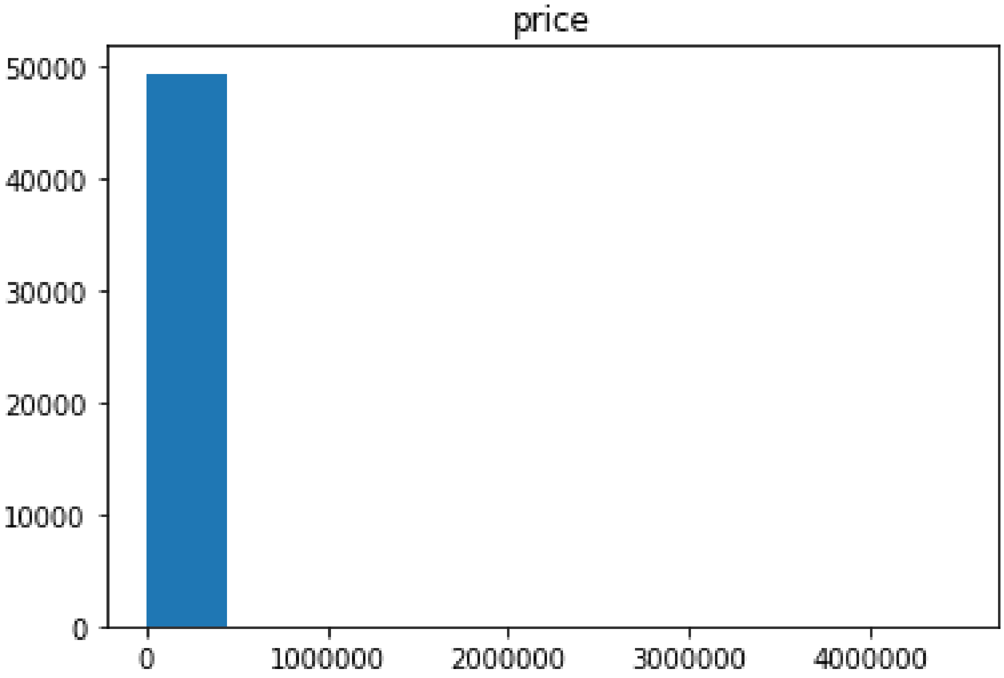
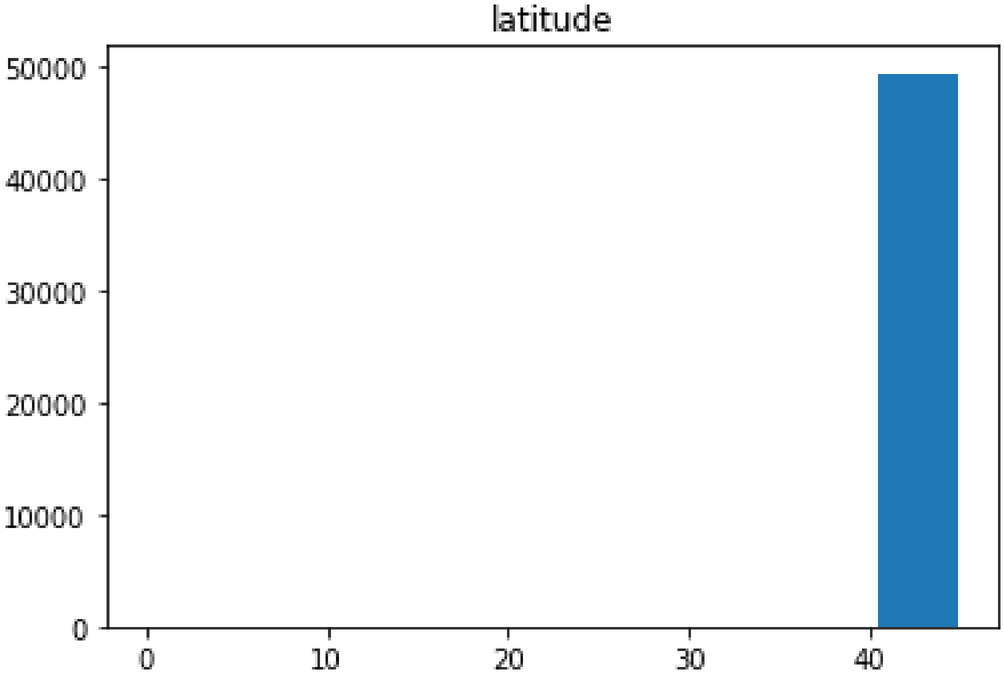
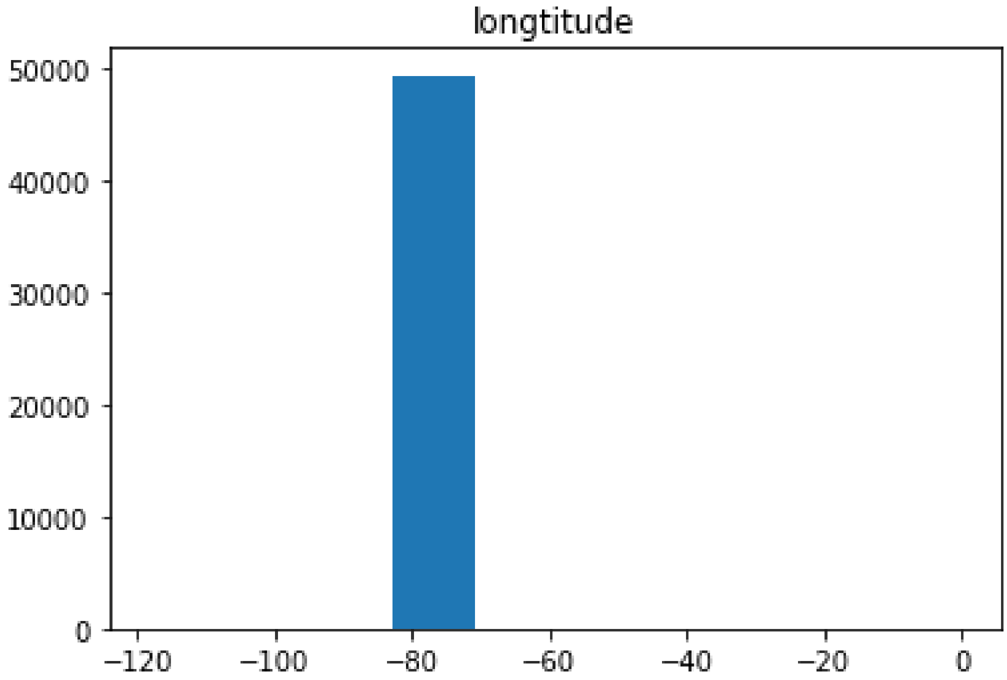
Code repo: https://github.com/DekaiLin/cmpt459project

A. Exploratory data analysis:

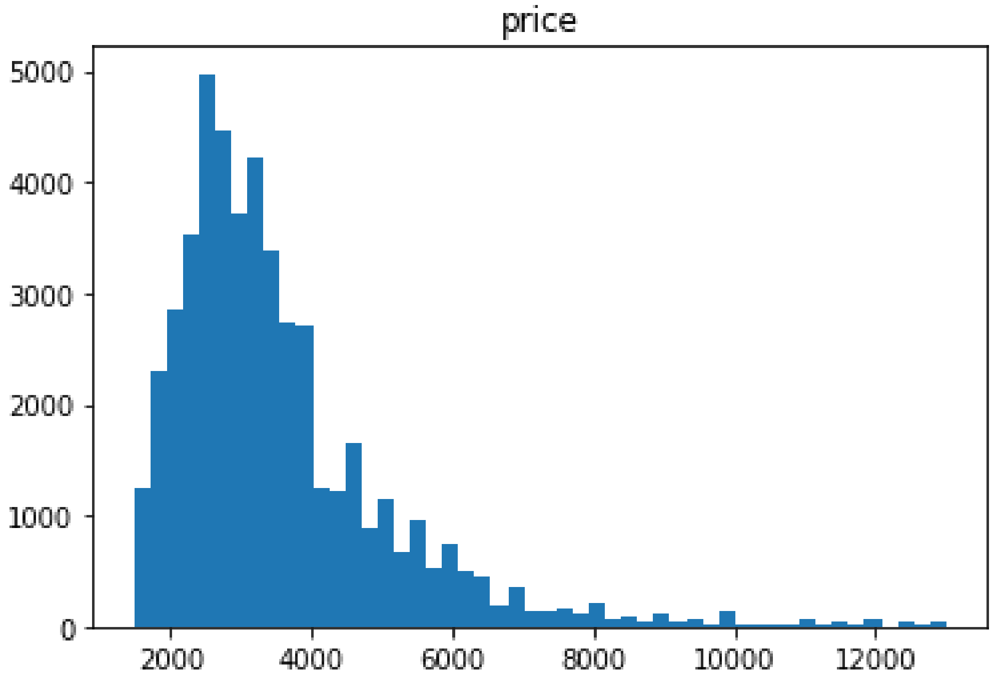
1. The outliers affect the histogram, so we need to drop the heads and tails to avoid the outliers' effect.

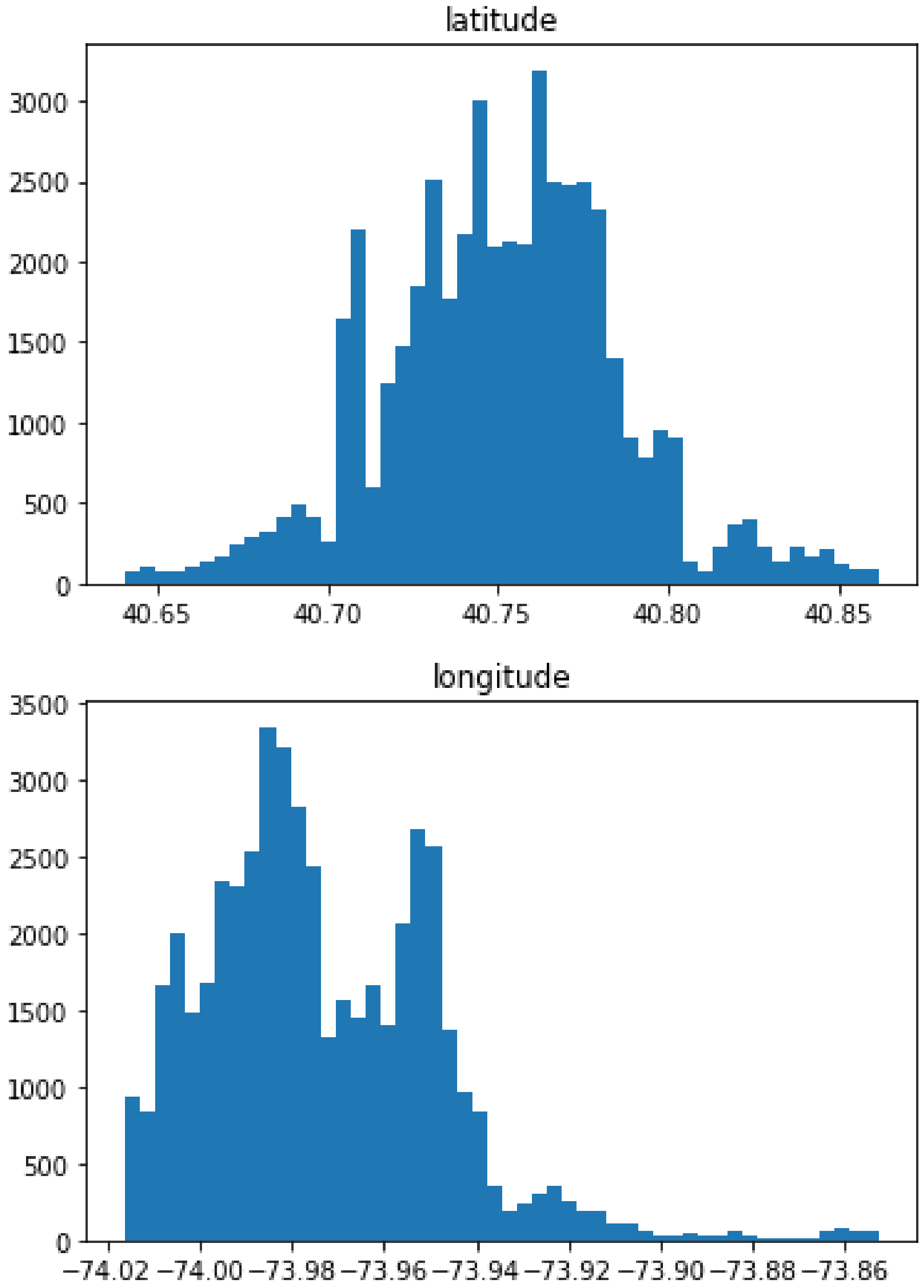




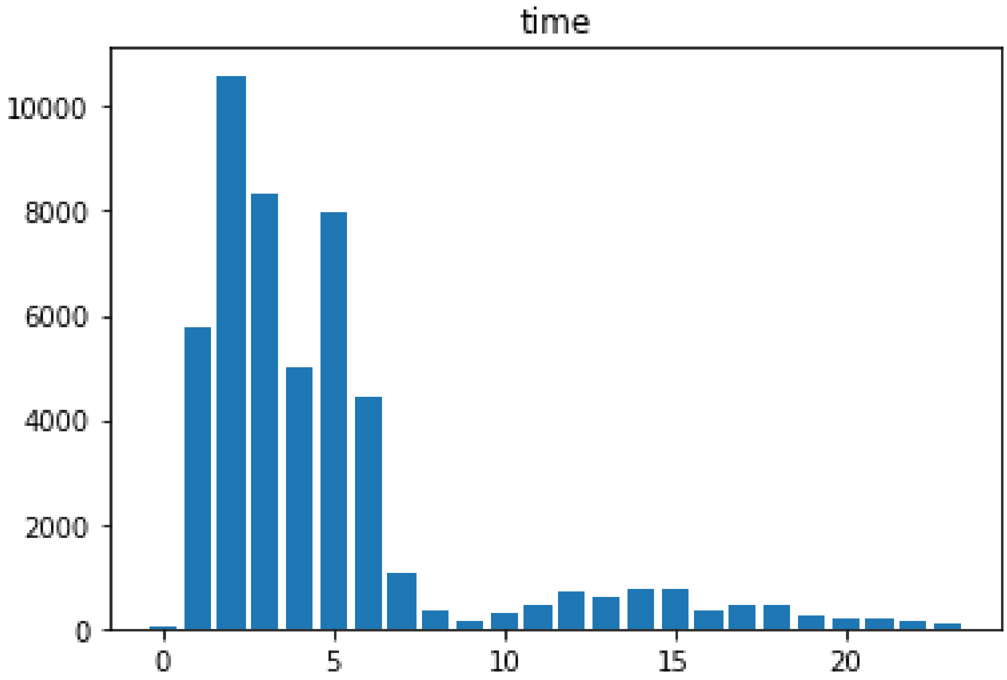


After dropping the outliers:

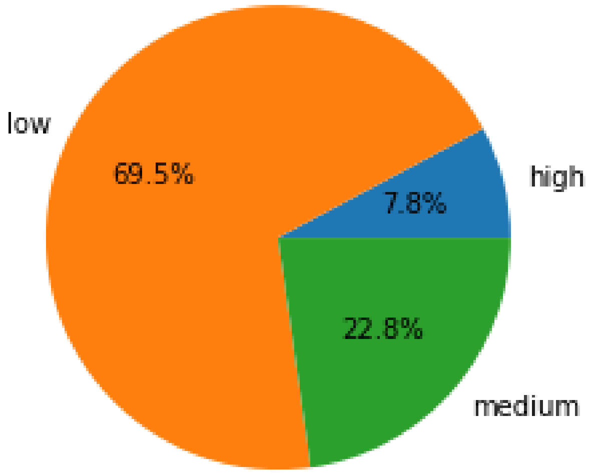



latitude
3000
2500
2000
1500
1000
500
0
40.65
40.70
40.75
40.80
40.85
longitude
3500
3000
2500
2000
1500
1000
500
0
−74.02
−74.00
−73.98
−73.96
−73.94
−73.92
−73.90
−73.88
−73.86

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



3. Proportion of target variable values:



B. 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	

2. 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.

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

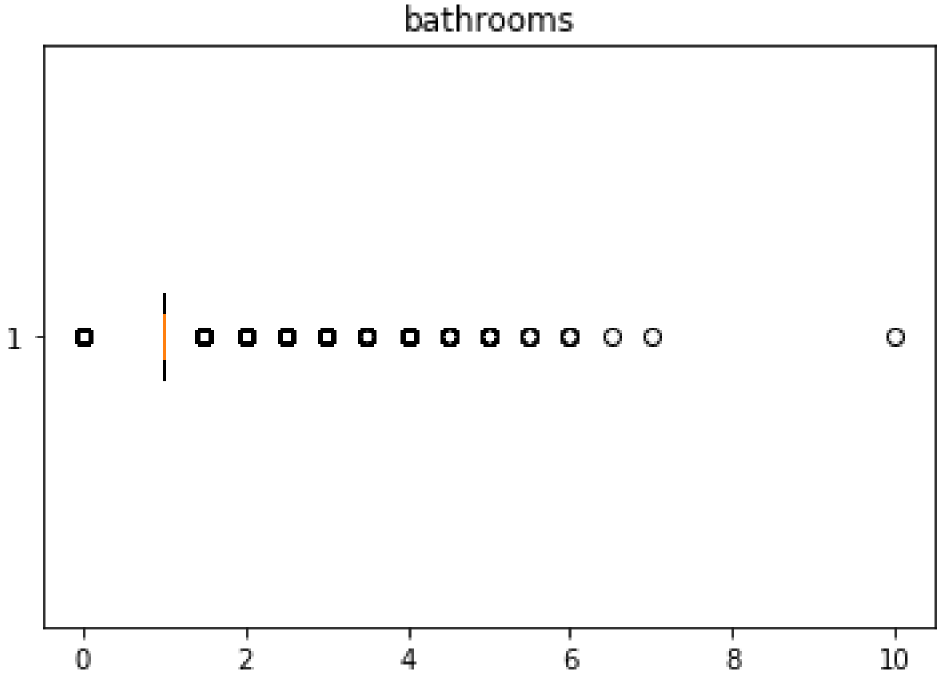
   The missing value of description and features attribute can be supplemented by cross-references. If both of them are missing, the features can be extracted from photos.

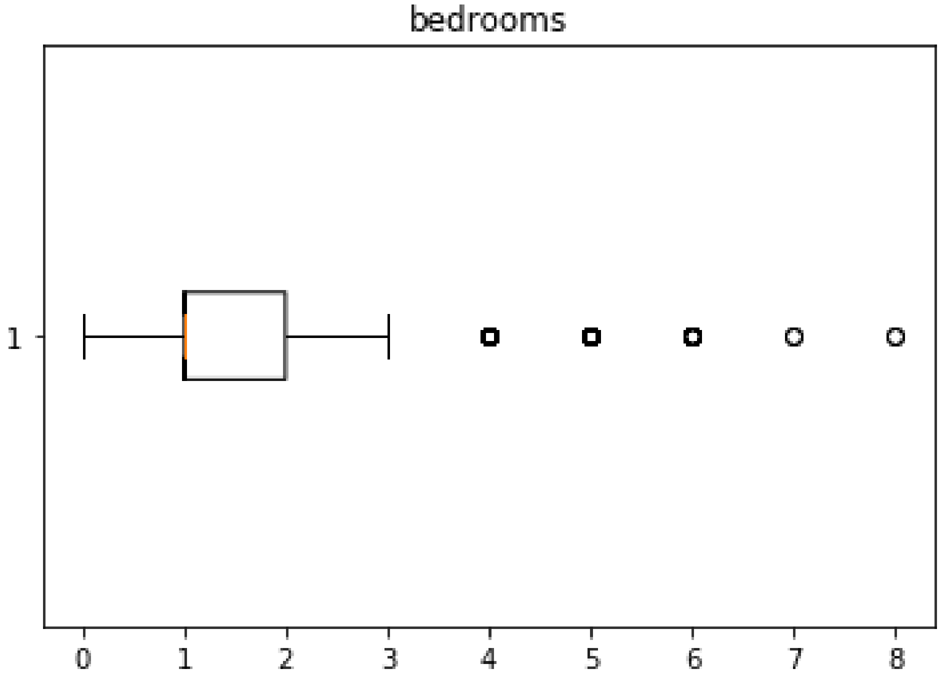
   For the latitude and longitude, we can replace the missing value by the New York’s latitude and longitude.

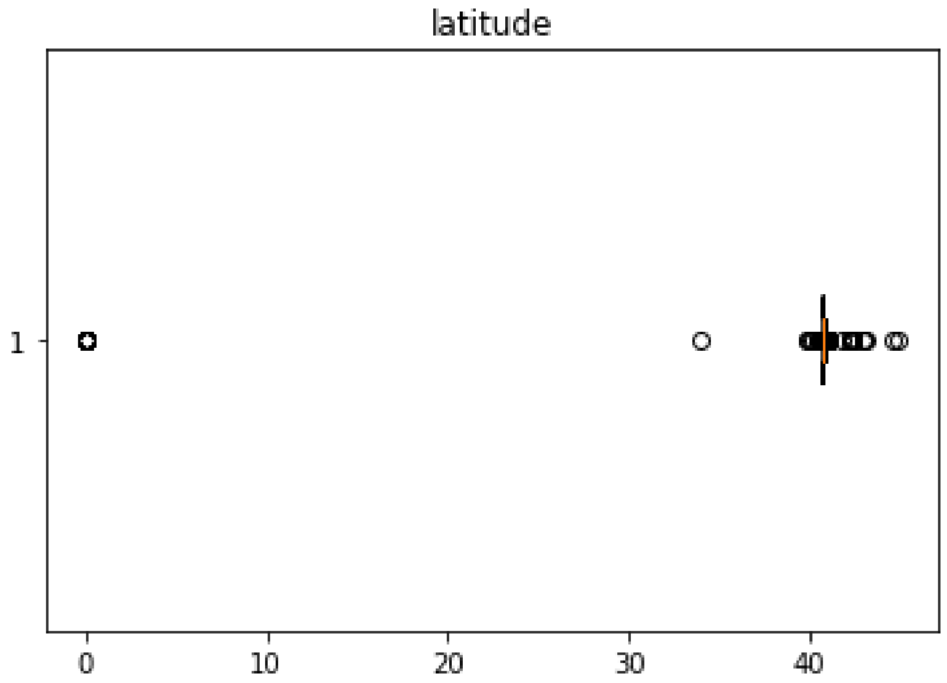
For the display address and street address, we can get the address by latitude and longitude.

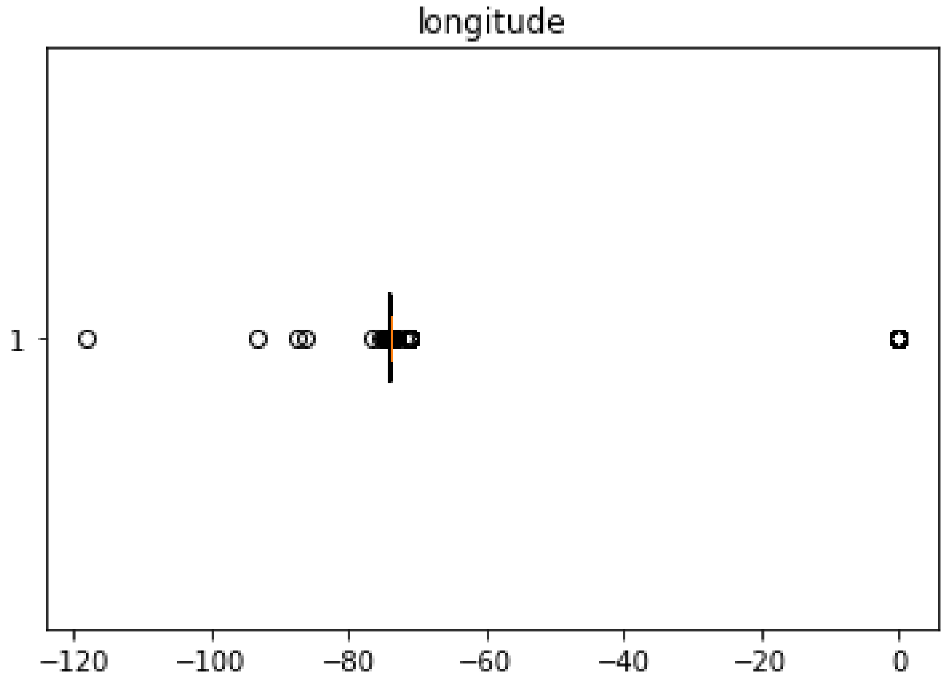
For the photos, we can only drop the missing value because we have no way to impute the missing photos from other attirbutes.

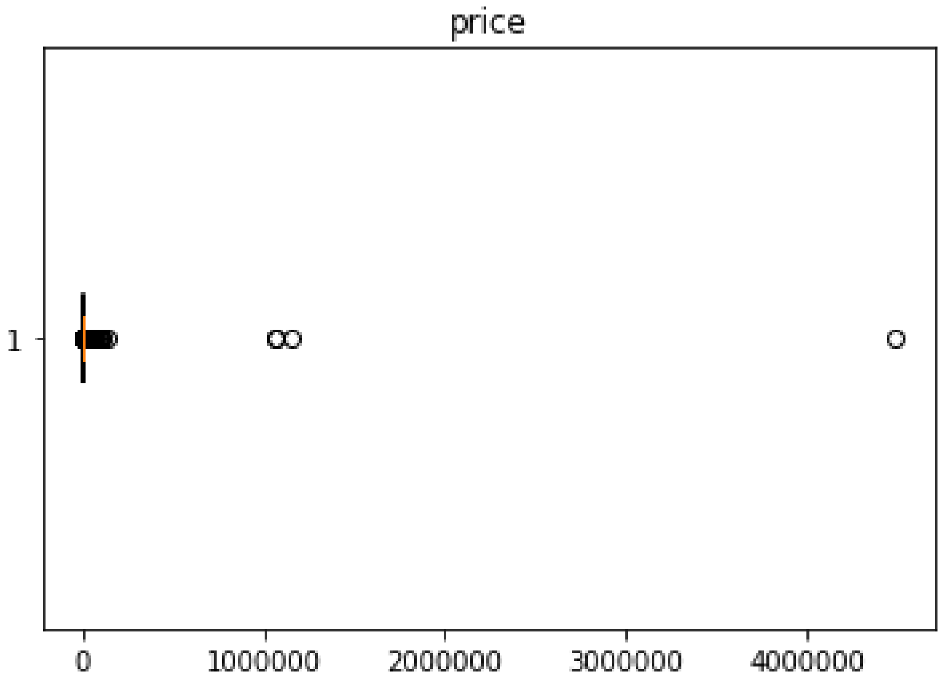
Plot of the outliers:






latitude
1
0
10
20
30
40

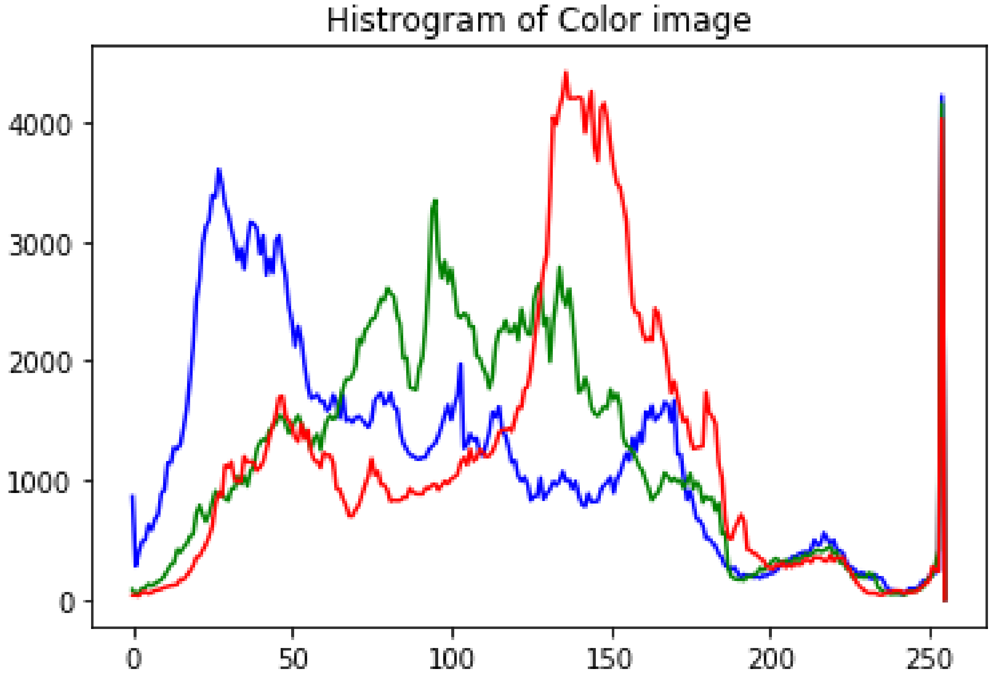

longitude
1
−120
−100
−80
−60
−40
−20
0


price
1
0
1000000
2000000
3000000
4000000

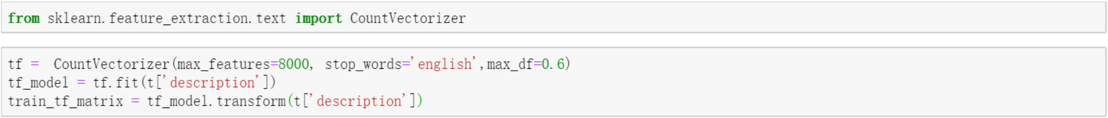
C. Feature extraction from images and text:

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

The colors histogram of this picture:



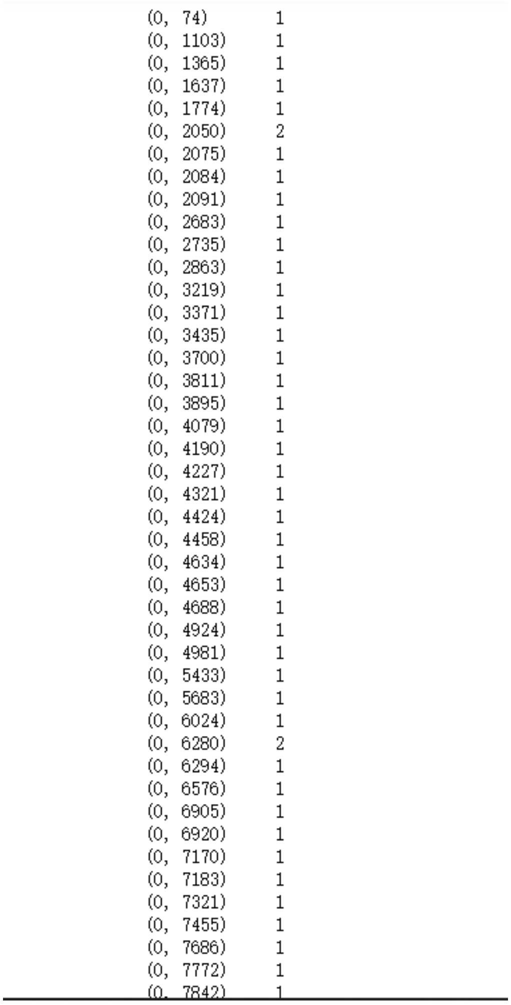
2. For the text data we used term frequencies as features. For example, we used 8000-dimension vector to map description text in term space. We drop the English stop words and some very frequent terms (like '<br> which is not in the English stop word list) by maximum document frequency = 0.6



```
from sklearn.feature_extraction.text import CountVectorizer
```

```
tf =  CountVectorizer(max_features=8000, stop_words='english',max_df=0.6)
tf_model = tf.fit(t['description'])
train_tf_matrix = tf_model.transform(t['description'])
```

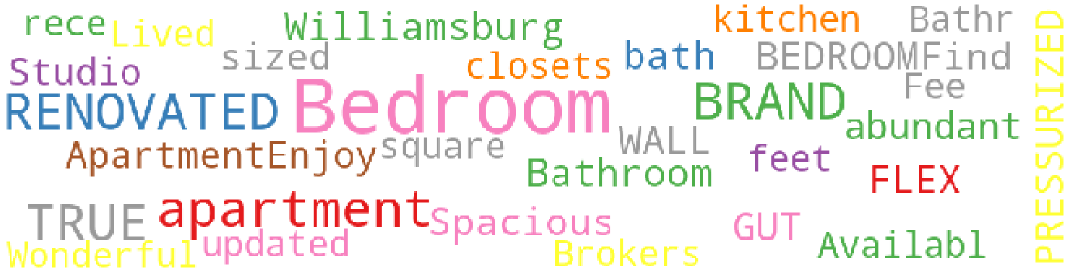
This is the corresponding document vector for a sample training data.



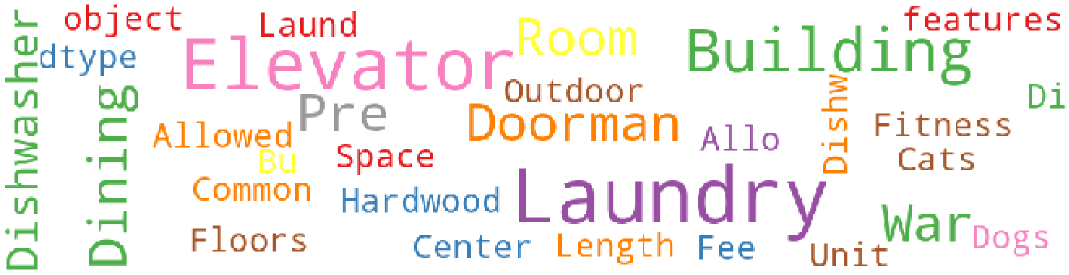
```
  (0, 74)	1
  (0, 1103)	1
  (0, 1365)	1
  (0, 1637)	1
  (0, 1774)	1
  (0, 2050)	2
  (0, 2075)	1
  (0, 2084)	1
  (0, 2091)	1
  (0, 2683)	1
  (0, 2735)	1
  (0, 2863)	1
  (0, 3219)	1
  (0, 3371)	1
  (0, 3435)	1
  (0, 3700)	1
  (0, 3811)	1
  (0, 3895)	1
  (0, 4079)	1
  (0, 4190)	1
  (0, 4227)	1
  (0, 4321)	1
  (0, 4424)	1
  (0, 4458)	1
  (0, 4634)	1
  (0, 4653)	1
  (0, 4688)	1
  (0, 4924)	1
  (0, 4981)	1
  (0, 5433)	1
  (0, 5683)	1
  (0, 6024)	1
  (0, 6280)	2
  (0, 6294)	1
  (0, 6576)	1
  (0, 6905)	1
  (0, 6920)	1
  (0, 7170)	1
  (0, 7183)	1
  (0, 7321)	1
  (0, 7455)	1
  (0, 7686)	1
  (0, 7772)	1
  (0, 7842)	1
```

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

Top 30 frequent terms in DESCRIPTION attribute



Top 30 frequent terms in FEATURES attribute


object
Laund
features
dtype
Elevator
Room
Building
Dishwasher
Dining
Outdoor
Di
Pre
Doorman
Allowed
Allo
Dishw
Fitness
Bu
Space
Cats
Common
Hardwood
Laundry
War
Floors
Center
Length
Fee
Dogs
Unit