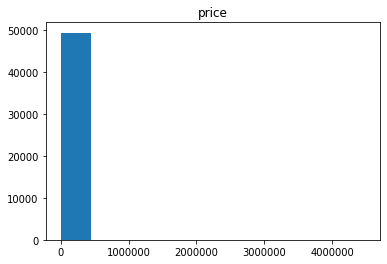
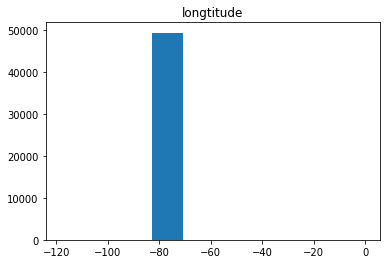
Code repo: <https://github.com/DekaiLin/cmpt459project>

Exploratory data analysis:

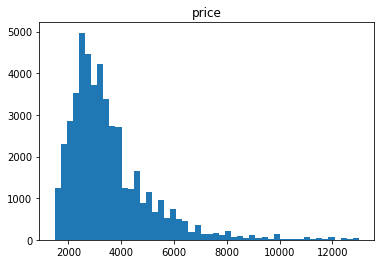
1. The outliers affect the histogram, so we need to drop the heads and tails to avoid the outliers’ effect.

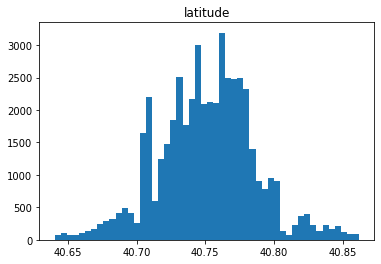


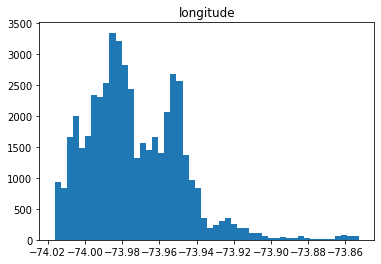




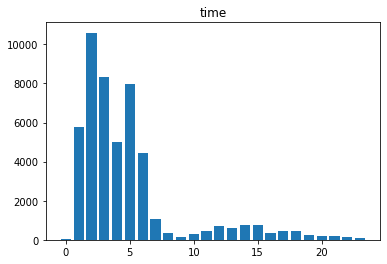
After dropping the outliers:







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



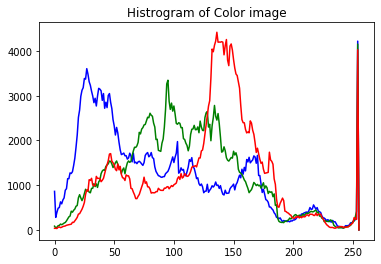
1. 这个题目没看懂，需要问老师确定。

Feature extraction from images and text:

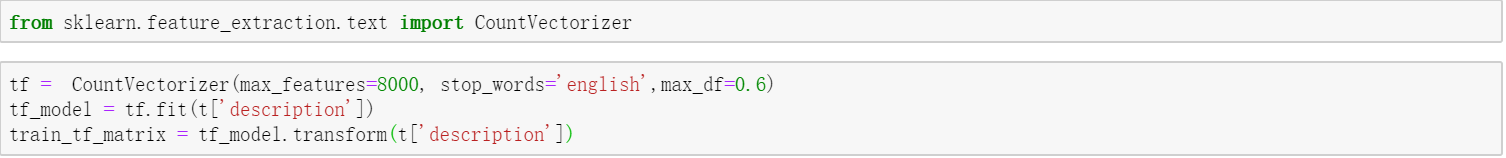
1. For the image data we used histograms of colors to extract features. For example,



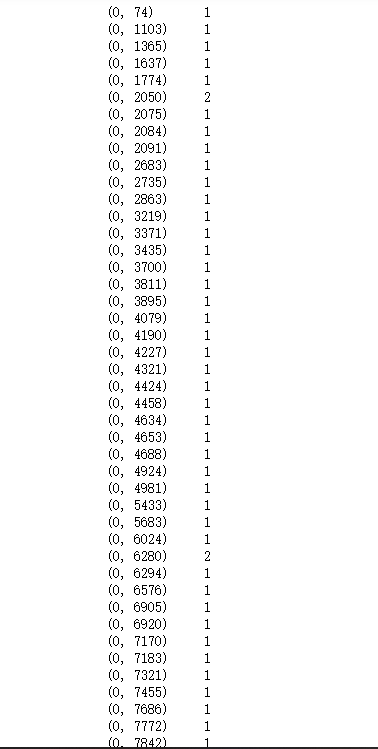
The colors histogram of this picture:



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.



