CS591L1 Report

New York House Renting Advice

Lingzhou Ao

Overview

New York is a popular city among students and visitors, so every year lots of people would like to choose this city to settle down. For these people who are new to New York, they really need some advice for renting house. So, I collect some datasets to analyze and give each house on Zillow a rate to represent its value.

Data will tell the truth and it is more reliable than agents.

Data Collection

In this project, I use 5 datasets from NYC Open Data, which are New York Subway, New York Crime, New York School, New York Hospital, New York Stores. And for house dataset, I implement a web crawler to obtain from Zillow by using web Driver to simulate human action and Beautiful Soup to parse website.

Data Transformations

Original Dataset	Transformation Description	New Dataset
New York Subway, New York Crime, New York School, New York Hospital, New York Stores, New York House	For each house, I calculate the number of subway station, crime, school, hospital and store, and then I use these as the house attributes	New York Houses Attributes
New York Houses Attributes	By using Z-score method, I normalize all the attributes I get before.	New York Norm Houses

New York Norm Houses	I use all attributes to generate a new rate, and calculate the correlation coefficient and p value between each attribute and new rate	Correlation
New York Houses, New York Norm Houses	I use all normalized attributes except the crime to generate a new rate for each house, then I use it and house's coordinates to do the K-means cluster.	Cluster

Data Normalization

The original rate of a house is based on its price and I will calculate the number of subway stations, crimes, schools, hospitals and stores near a house to generate the new rate.
Here I use Z-score method to normalize all factors.

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation.

Correlation coefficient and P value

The formula to calculate the Correlation coefficient between each factor and the new rate is following:

$$corr\left(N_{factor}, N_{rate}\right) = \frac{cov\left(N_{factor}, N_{rate}\right)}{std\left(N_{factor}\right) - std\left(N_{rate}\right)}$$

The result of Correlation coefficient and P value is following. I find out that the factor crime is not important because its P value is significantly high and its Correlation coefficient is small enough to ignore. In future research, I will discard this factor.

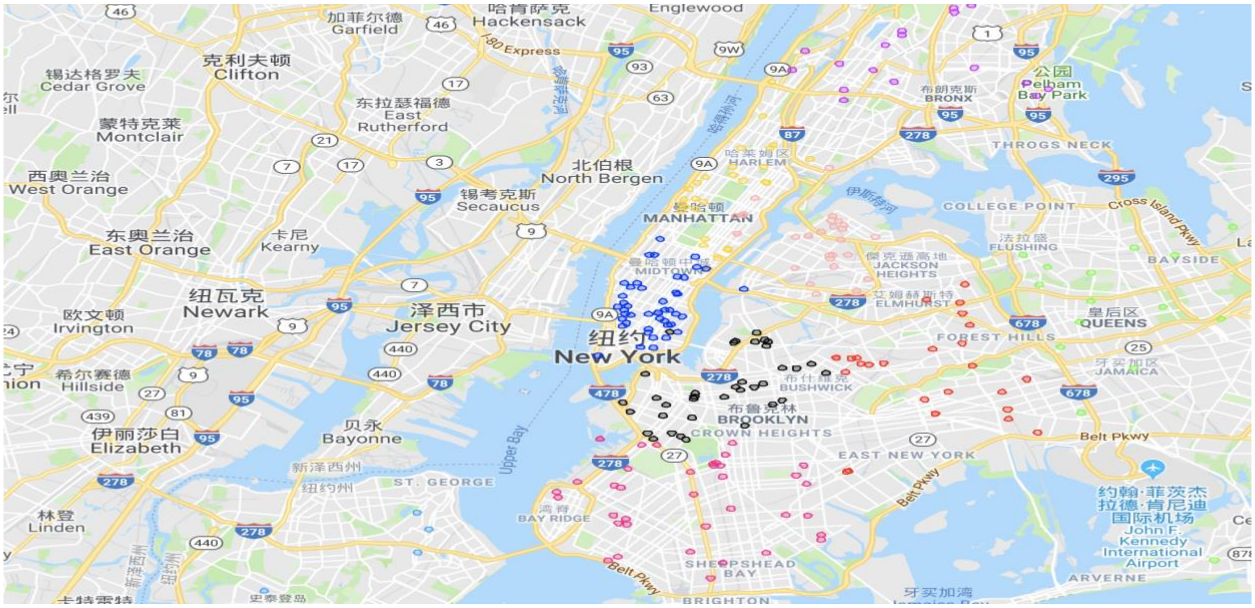
Factor	Correlation coefficient	P Value
Crime	0.13095148010181989	0.038539876761523276
Subway	0.81151775292531803	8.4120129244698685e-60
School	0.8775738524696437	3.827419488666782e-81
Hospital	0.63057488987461152	4.0315722807958251e-29
Store	0.78933469677175383	1.8418759267572758e-54

K-means Clustering

I discard the crime factor and generate the new rate of houses. And then I use it and coordinates of houses for K-means Algorithm.

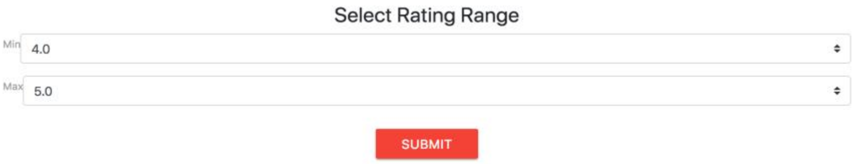
By analyzing the error, I choose to use 10 as the number of clusters.

I use the gmplot package instead of matplot so I can plot all the results on a real map and as the following figure shows, I think the clusters are quite reasonable based on the information we have known about New York.



Visualization

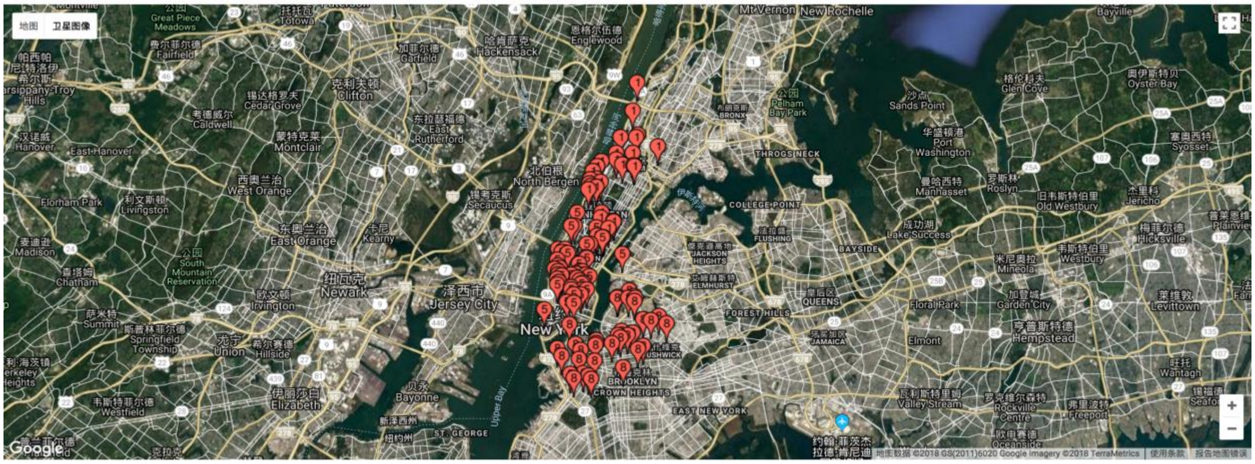
For Visualization, I use JavaScript and Django to implement a simple website that allows users to select the rating range they want and the results will be shown on the Google Map. Also, if user moves the Mouse cursor to the marker, it will show some details of this house and when user clicks it, it will direct to Zillow to show more details about the house.

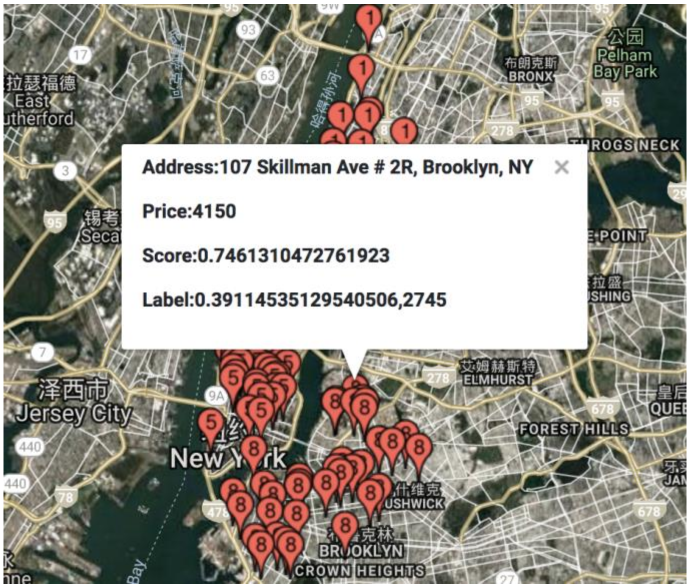


Select Rating Range

Min 4.0

Max 5.0





Address:107 Skillman Ave # 2R, Brooklyn, NY

Price:4150

Score:0.7461310472761923

Label:0.39114535129540506,2745

Future Work

In the future, I will gather more house data cause I think the data from Zillow is a little bit not enough. And I will do more preprocessing to avoid some invalid data. Also, I will try to find some more interesting factor that will influence the evaluation of a house.