Finding the More Select Area to Establish the Restaurant

Zhilin Zhang

March 2020

1. Introduction

1.1 Background

It is an artwork to find a proper place for a restaurant. Good location represents many customers and a steady stream of money. We used to determine position by handcraft, but now, with the machine learning technology, we can complete it by computer.

The development of machine learning and geographic technology gives us a new way to understand the business value. How to utilize the machine to find a fitting location is the newest problem.

1.2 Problem

This project aims to support who is attending to open a restaurant in San Francisco. We will compare the space and provide the result.

Where is the best place that they open it? Where can we meet more customers' needs? Where have more potential customers? These all the questions we intend to solve.

2. Data Processing

2.1 Data sources

We got the district data source from Coursera course Data Visualization with Python. Meanwhile, the POI and geographical coordinates data are both got from Foursquare by web crawler.

2.2 Data Processing

a) We loaded the coordinates data to the notebook, which includes 161,996 rows. Dividing the coordinates to Latitude and Longitude, we deleted the userID column. Latitude and Longitude data are necessary for us to cluster the space. The original dataset showed below:

	POI	jw	city	tag
0	19542	37.6163560649,-122.3861503601	San Francisco	Airport
1	13338	37.7615082532,-122.4257665873	San Francisco	Ice Cream Shop
2	9153	37.759688872,-122.4271774292	San Francisco	Park
3	202692	37.78594785,-122.4106178	San Francisco	Hotel
4	14608	37.7826046833,-122.4076080167	San Francisco	Coffee Shop
...	...	...	...	...
161991	51704	37.8077919908,-122.4183046818	San Francisco	Burger Joint
161992	153386	37.8087922168,-122.4102687836	San Francisco	Seafood Restaurant
161993	568307	37.8059864638,-122.4050921202	San Francisco	American Restaurant
161994	561688	37.6144195596,-122.3901003005	San Francisco	Airport
161995	19542	37.6163560649,-122.3861503601	San Francisco	Airport

b) Second, we grouped the same POI data, and recorded the count of the same POI raws. We wanted to get how numerous people arrive at this place within this dataset. The dataset looked below:

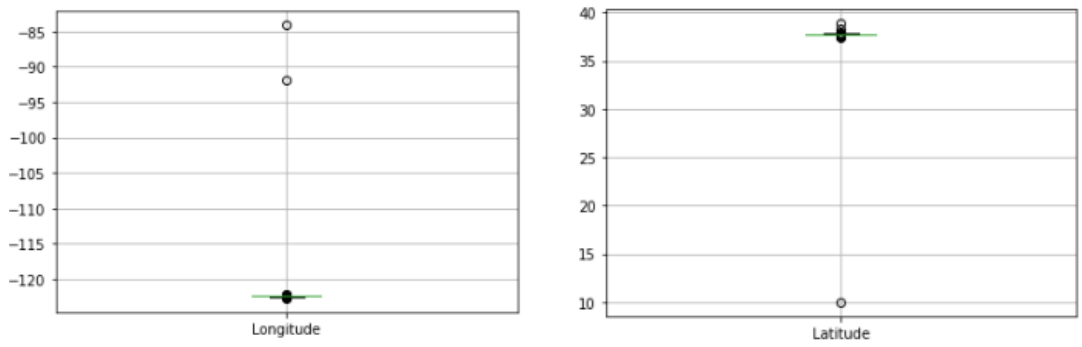
	POI	count	tag	Latitude	Longitude
8	9113	74	Fastfood	37.7946797244	-122.4054139853
11	9127	260	Restaurant	37.8083207137	-122.4157217145
18	9161	113	Restaurant	37.8005003833	-122.4074578167
19	9162	16	Fastfood	37.7936008841	-122.4063527584
26	9583	43	Bar	37.7830445667	-122.46513605
...	...	...	...	...	...
11472	5648694	1	Fastfood	37.808149233	-122.41495409
11476	5663346	1	Fastfood	37.8073239326	-122.4182081223
11480	5701584	1	Restaurant	37.747096467	-122.413270333
11487	5766710	1	Restaurant	37.786207	-122.408196
11505	5915895	1	Bar	37.76757748	-122.406822481

c) Then, searching tags of the data, which is the type of the location, and combine the same tag into the same word, such as University and College are all changed to School. This step purposes to help us analyze what kind of potential customers will have in this area.

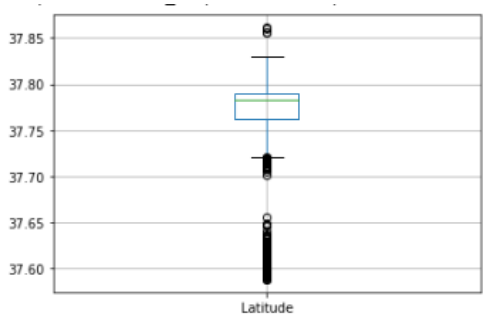
d) We divided the data into two types; one is the place which may have possible customers; another is other restaurants.

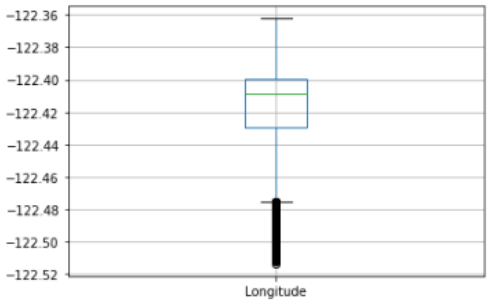
e) Finally, we did the data cleaning. By drawing the boxplot, we could observe visually that there are any coordinates are not in San Francisco. We drop them and arrange the dataset.

The boxplots below show the data without cleaning:



The boxplots below show the data after cleaning:





3. Problem Solving

Cluster

For the cluster part, we practice K-means, which is a popular method, to do cluster. As we require to find the areas, we decide to employ the total map of San Francisco to discover the location.

To cluster the data , as we what to show the result on map, we selected Latitude and Longitude columns as input.

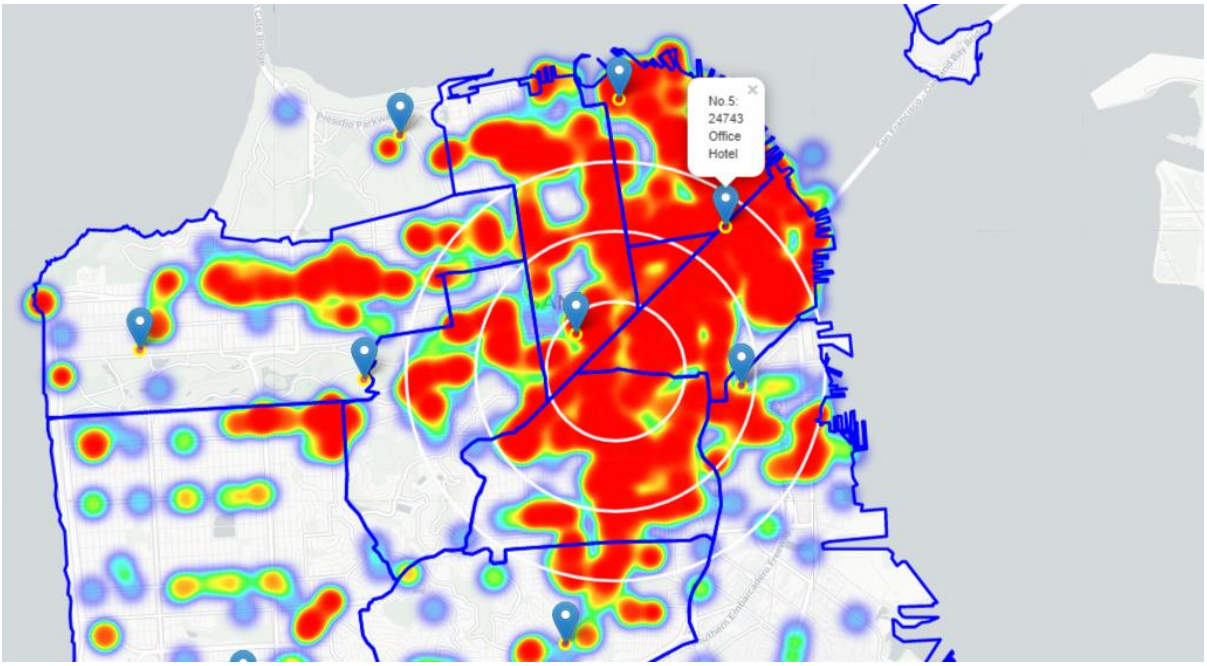
We tried the k from 8 to 15, which is how many clusters we want, and find the best k is 10, which will not divide the airport part to two-part. Then do the cluster.

4. Analyses

For analyses part, we got the table below:

ID	Close Space 1	Close Space 2	Crowed	Latitude	Longitude
1	Office	Sport	6509	37.76807836	-122.39945868
2	Airport	Hotel	11400	37.6129963	-122.39120866
3	Park	Site	751	37.77277806	-122.4975116
4	Park	Sport	4578	37.8050439	-122.41952574
5	Office	Hotel	24743	37.78865967	-122.40204586
6	Sport	Art Event	7549	37.77485188	-122.42653226
7	Park	Art Event	959	37.73481564	-122.42821392
8	Mall	Park	1321	37.72868246	-122.48066291
9	Park	Art Event	2843	37.76898467	-122.46101539
10	Sport	Site	2077	37.80043101	-122.45519854

To analyses each location we select, we draw a map by python folium library. Which should below:



The Picture showed the heat map of how many restaurants and how many people visit, and the blue markers show the area we select.

5. Results and Discussion

Our analysis shows that although there is a significant number of restaurants in San Francisco, there are pockets of low restaurant density reasonably close to the Airport. We select ten areas to judge whether it fits for a restaurant. It can explain that the location near the city center has a considerable number of potential customers but also have plenty of competitors.

Meanwhile, the place nearby the Airport also satisfies the requirement, but it may cost a lot if someone desires to establish a restaurant. For further detail, we may make marketing exploration and find what kind of people like what kind of food. That will benefit us in finding a more accurate location.

6. Conclusion

The project shows the 10 suitable position for establishing a restaurant. The place nearby Airport shows the best result, and place nearby CBD also have a good score.