Recommending the Best Location for Opening an Italian Restaurant

Deniz Aytemiz

12/20/2020

1. Introduction

1.1 Background

Opening a restaurant in a city can be challenging especially if the person is new to the city. There are many factors to consider when deciding the location. For instance the competition around the area, safety of the area, whether the renting prices are too expensive or affordable, in addition to popularity of the area. All these issues can change the preference on the location. However all locations will have their pros and cons in terms of criteria and no location will be perfect. Therefore segmenting the locations based on their properties will illustrate the person what he/she will prefer over what.

1.2 Problem

Data on number of Italian restaurants around the area, the popularity of the area, safety of the area, the housing prices of the area and the restaurant rankings around the area will give a description of the areas. The areas are determined by postal codes. This project aims to segment distinct postal codes areas based on the properties discussed above. In this project; I examined the areas of Boston city for a person considering opening an Italian restaurant.

2. Data acquisition and cleaning

2.1 Data Sources& Data Cleaning

Datasets are acquired from Foursquare and www.boston.gov. From Foursquare we get the list of venues and their corresponding postal codes. From Boston's government website, dataset on crime rates and housing prices are retrieved. From crime rates dataset, the corresponding total number of crimes in a district is hold, other columns are dropped. The districts with NaN or "external" values are also dropped. From housing prices dataset the

average of the housing prices corresponding to postal code/zip code is hold and other columns are dropped. Some postal code areas did not hold information on the housing prices; therefore I filled that data with the overall average. These informations are merged with dataset from the foursquare dataset later on. They are added as columns to dataset as "Number of Crimes in the area" and "Average Housing Prices in the Area".

From Foursquare I retrieved a dataset with latitude and longitude values of Boston. Since the grouping of the dataset is mainly done by postal codes, the rows with NaN value of postal codes are dropped.

2.2 Feature Selection

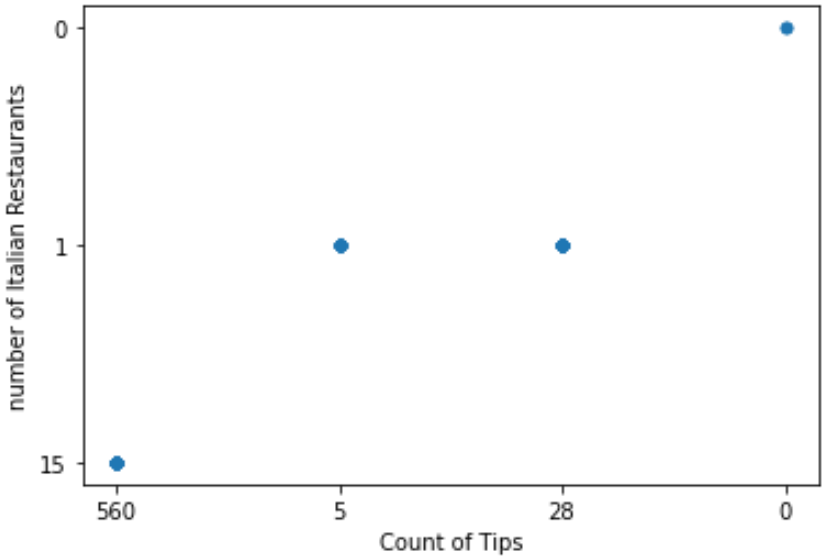
The features selected are, average housing prices of the area, number of crimes in the area, number of Italian Restaurants around the area and average ratings of the Italian Restaurants around the area. Average housing prices and number of crimes are retrieved from the datasets from Boston government's website. Number of Italian Restaurants in the area is found from the dataset retrieved from Foursquare by groupby() and valuecounts() methods. For average ratings of the Italian restaurants, using the venue ids I have made another API request and stored the ratings in variables. The number of tips is selected as a feature for an indication of the area popularity. Since trending venues request returned no trending venues, I have decided to use number of tips to see the most spoken and commented venues and which area they belong.

3. Explanatory Data Analysis

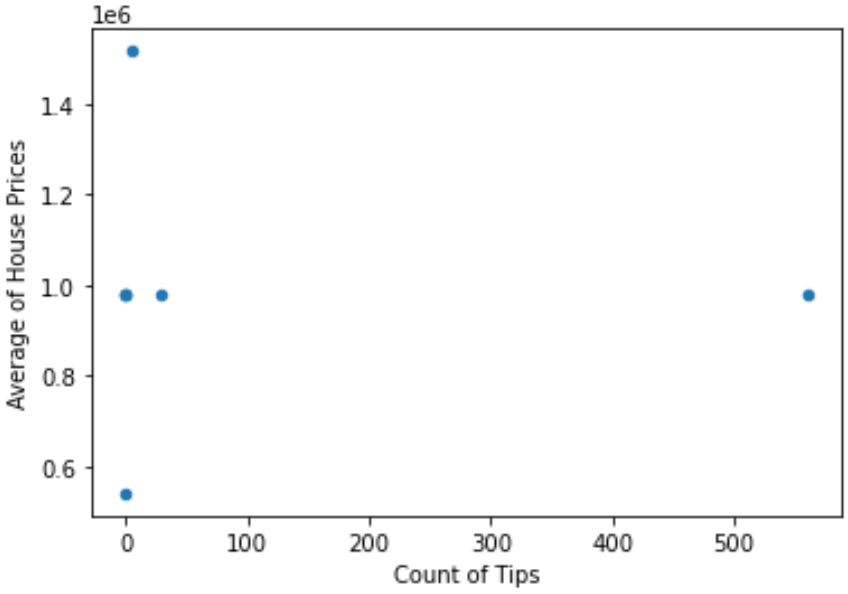
The plots below aims to detect correlation between two features. The final version of the dataframe is given below. This dataframe will later on modeling and the labels for each postal code will be added.

	postalCode	Average Ranking of Italian Restaurant	Count of Tips	Average of House Prices	Number of Crime Incidents	Number of Italian Restaurants
0	02113	8.2	560	981968.24	0	15
1	02108	8.3	5	1516433.33	0	1
2	02109	8.6	28	981968.24	0	1
3	02114	0.0	0	981968.24	13	0
4	02110	0.0	0	541200.00	0	0
5	02111	0.0	0	981968.24	0	0
6	02201-1001	0.0	0	981968.24	0	0

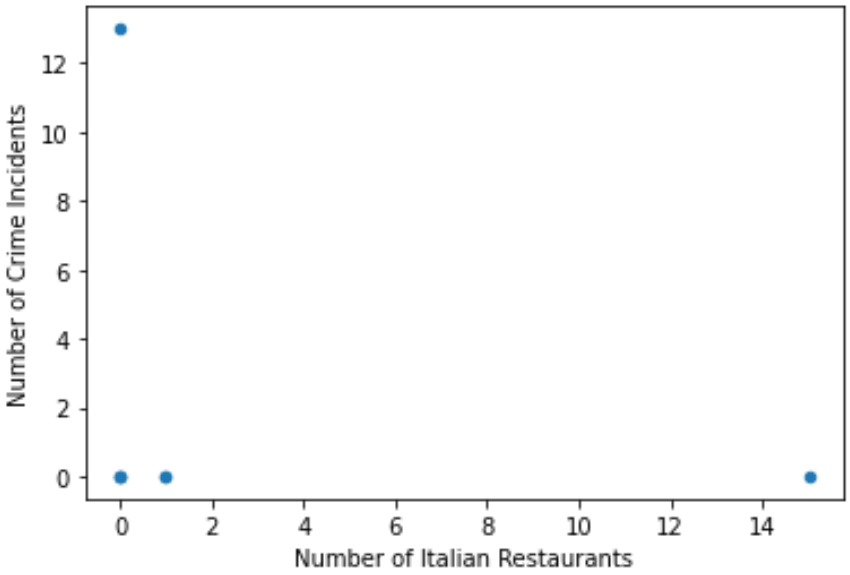
There is a positive correlation between count of tips and number of Italian Restaurants. This is intuitive. Since the number of restaurants increases the possibility for a tip increases.



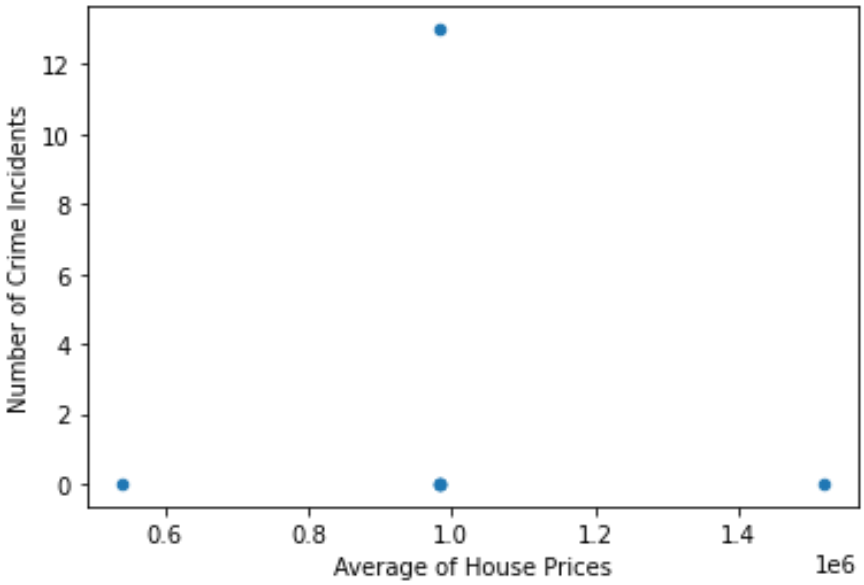
No correlation between average house prices and count of tips is observed.



There is no correlation between crime rates and number of Italian Restaurants. There is just one specific location with high crime rate.

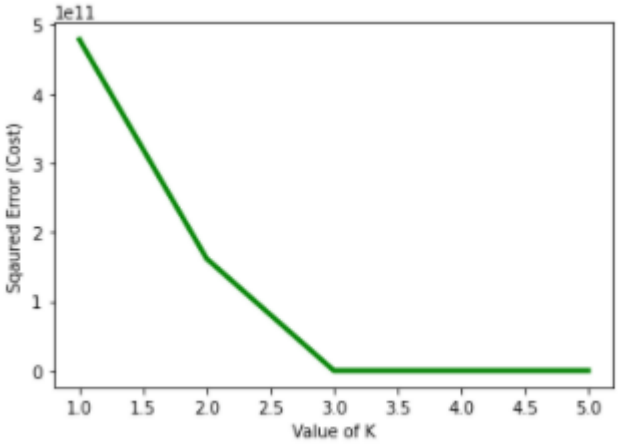


Again there is no correlation between number of crime rates and average house prices.



4. Model

Since there is no labeling or specific category, unsupervised machine techniques must be used. I have used k-means clustering as the model. Depending on the labels assigned to each postal code area, the properties of the areas could be defined. For selecting the best optimal k value, the elbow point must be identified. Below figure gives the graph of k-value with respect to square error cost.



The optimal k value is found to be 3.

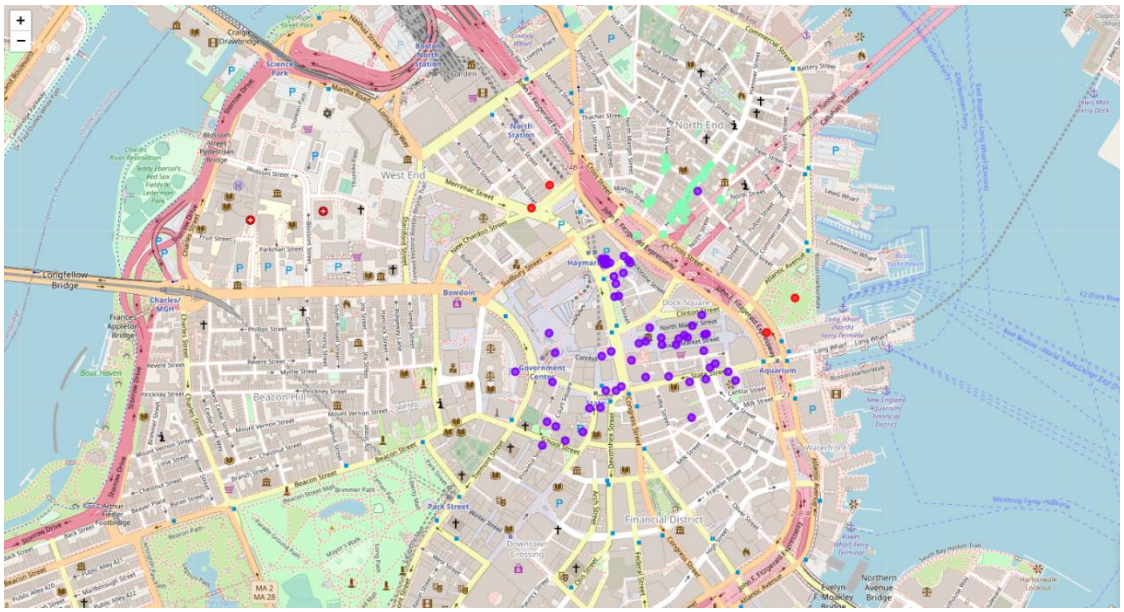
5. Results

Labels	Average Ranking of Italian Restaurant	Count of Tips	Average of House Prices	Number of Crime Incidents	Number of Italian Restaurants
0	0.00	0.0	871776.180	3.25	0.0
1	8.45	16.5	1249200.785	0.00	1.0
2	8.20	560.0	981968.240	0.00	15.0

From the dataframe above, we can gain an insight about the labels. Label 0 can be interpreted as, low ranking, low popularity, cheaper house prices, high rates of crime with very few Italian Restaurants. Label 1 indicates places with, very high rankings, average popularity, expensive housing prices, minimum crime incidents and few Italian Restaurants. Label 2 indicates popular areas with high rankings, average housing prices, low rates of crime and many Italian Restaurants. In order to see which postal code area corresponds to which label; we add column label to dataframe. The result is shown below.

	postalCode	Average Ranking of Italian Restaurant	Count of Tips	Average of House Prices	Number of Crime Incidents	Number of Italian Restaurants	Labels
0	02113	8.2	560	981968.24	0	15	2
1	02108	8.3	5	1516433.33	0	1	1
2	02109	8.6	28	981968.24	0	1	1
3	02114	0.0	0	981968.24	13	0	0
4	02110	0.0	0	541200.00	0	0	0
5	02111	0.0	0	981968.24	0	0	0
6	02201-1001	0.0	0	981968.24	0	0	0

The corresponding map after clustering is shown below figure.



6. Conclusion

In this project, I have tackled the problem of opening an Italian Restaurant in Boston. For this purpose I have used three datasets. From the housing prices data set I have found the average housing prices corresponding to the postal code. From crime incident dataset by replacing districts with postal codes the total number of crime incidents for postal codes are saved. Even though the datasets retrieved from the boston.gov website was large, due to Foursquare's limit, I could retrieve 100 rows. Therefore some postal code areas had little information in other datasets. If the Foursquare limit was increased or the data query was based on postal code, better analysis can be made.

After collecting all the information on a single dataframe with proper data cleaning, the feature set is determined and normalization is done. Then k-means clustering is used as the model for segmenting the areas based on the feature matrix given. To find the optimal value of k, square error mean is calculated for each k. Based on the labeling we could identify each postal code area according to their properties and visualize the clustering through a map.