Exploring, Segmenting, and Clustering Venues near Boston, Massachusetts

1. Introduction

1.1 Background

The restaurant industry constantly strives to update, adapt, and provide new opportunities to attract potential customers. In recent years, "fusion-style" restaurants have combined various styles of cuisine in order to appeal to a broader audience, as well as to provide fresh ideas when compared to other localized offerings. This capstone project seeks to explore various ways to incorporate new restaurants while comparing baseline data in the greater Boston area (Suffolk County, Massachusetts).

1.2 Problem

This is the Capstone Project will analyze venues in the county of Suffolk, Massachusetts (Including Boston and the surrounding towns). This data will optimize the location for the opening of a new fusion restaurant, containing two popular cuisine styles. The first half of the project will provide background and data collection for opening a fusion restaurant in the greater Boston area of Suffolk County. The project will use location data to see where the top 2 styles of restaurants would be fused together and be successful in that neighborhood. The second half of the project will employ data processing for recommending the location of a new fusion restaurant based on data of Suffolk County and Boston neighborhoods, popular venues in that area, and top existing restaurants in the area.

2. Data Acquisition and Cleaning

2.1 Data Sources

In this project, the notebook, and tools contained within it, will convert addresses into their equivalent latitude and longitude values. The data for all cities and towns in the state of Massachusetts is provided at the following link: https://geo.nyu.edu/download/file/harvard-mgisgeonamx2-geojson.json.

This data is provided by the NYU Spatial Data Repository, named "Massachusetts Geographic Place Names: Civic Features."

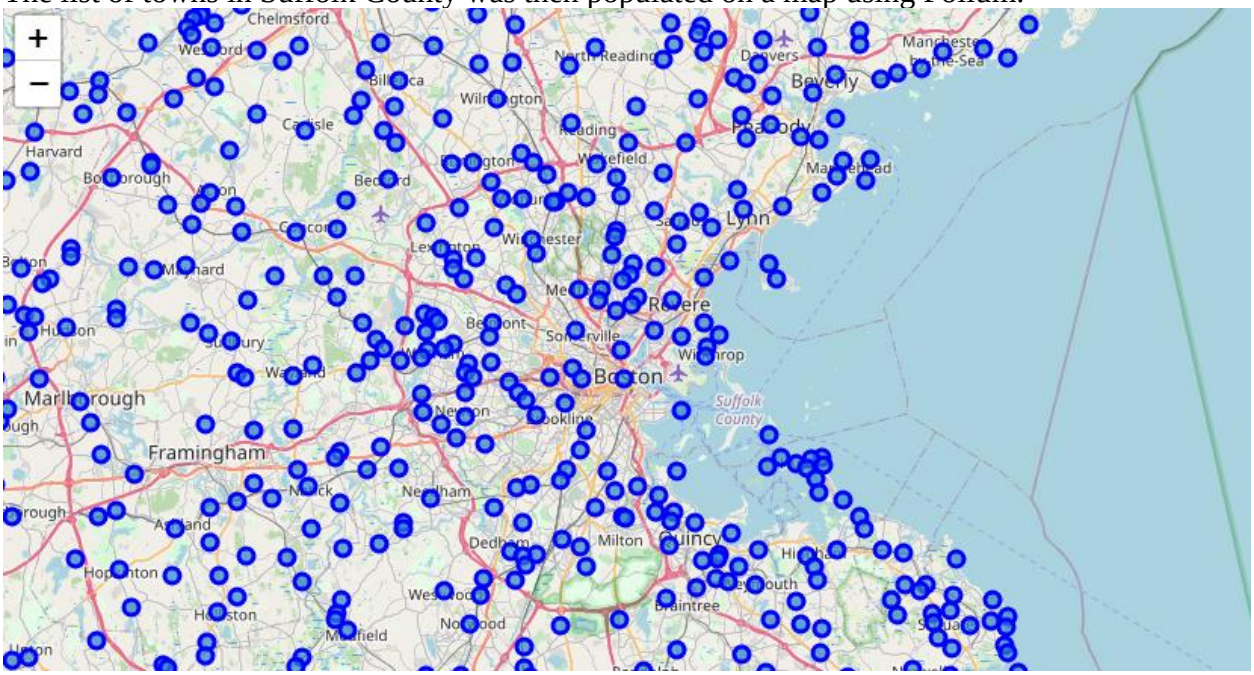
Foursquare API will provide data to explore towns and neighborhoods in Suffolk County and the greater Boston areas. Using the explore function will determine the most common venue restaurant style categories in each neighborhood. This feature will group the neighborhoods into clusters. The k-means clustering algorithm will enable the project to complete this task. Finally, the project will employ the Folium Library to visualize the towns in Suffolk County and their emerging clusters.

2.2 Data Cleaning

The location data provided from the NYU Spatial Data Repository contained the names of towns and counties for all of the state of Massachusetts. In order to focus the project, the data was cleansed in order to sort for towns and neighborhoods specifically in Suffolk County, which includes Boston. This provides a dataframe of the towns and neighborhoods from the most highly populated areas in the greater Boston metropolitan area.

	COUNTY	Neighborhood	Latitude	Longitude
0	25025	POINT OF PINES	42.437468	-70.965568
1	25025	BEACHMONT	42.395601	-70.990215
2	25025	REVERE	42.411107	-71.018667
3	25025	CHELSEA	42.391430	-71.035140
4	25025	ORIENT HEIGHTS	42.387261	-71.009795

The list of towns in Suffolk County was then populated on a map using Folium.



2.3 Feature Selection

These towns were used to pull data from Foursquare, which was used to provide the list of venues in each of the towns of Suffolk County.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
ABERDEEN	92	92	92	92	92	92
ALLSTON	100	100	100	100	100	100
ASHMONT	28	28	28	28	28	28
BEACHMONT	27	27	27	27	27	27
BELLEVUE	39	39	39	39	39	39
BOSTON	100	100	100	100	100	100
BRIGHTON	79	79	79	79	79	79
CHARLESTOWN	82	82	82	82	82	82
CHELSEA	51	51	51	51	51	51
DORCHESTER	18	18	18	18	18	18
FAIRMOUNT	30	30	30	30	30	30
FANEUIL	66	66	66	66	66	66
FOREST HILLS	28	28	28	28	28	28

Finally, this provided the starting point for the exploratory data analysis.

3. Exploratory Data Analysis

3.1 Analysis and Grouping of Each Neighborhood

The first step for data analysis was to group each neighborhood by taking the mean of the frequency of occurrence of each category (category equaling type of restaurant).

	Neighborhood	ATM	Afghan Restaurant	African Restaurant	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop	Automotive Shop
0	ABERDEEN	0.000000	0.00	0.00000	0.010870	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	ALLSTON	0.000000	0.01	0.00000	0.000000	0.000000	0.000000	0.010000	0.020000	0.000000	0.000000
2	ASHMONT	0.000000	0.00	0.00000	0.035714	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	BEACHMONT	0.000000	0.00	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	BELLEVUE	0.000000	0.00	0.00000	0.051282	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	BOSTON	0.000000	0.00	0.00000	0.010000	0.000000	0.000000	0.010000	0.010000	0.000000	0.000000
6	BRIGHTON	0.000000	0.00	0.00000	0.012658	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	CHARLESTOWN	0.000000	0.00	0.00000	0.024390	0.012195	0.000000	0.000000	0.012195	0.012195	0.000000
8	CHELSEA	0.019608	0.00	0.00000	0.039216	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	DORCHESTER	0.000000	0.00	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10	FAIRMOUNT	0.000000	0.00	0.00000	0.066667	0.000000	0.000000	0.000000	0.033333	0.000000	0.000000
11	FANEUIL	0.000000	0.00	0.00000	0.000000	0.000000	0.015152	0.000000	0.000000	0.000000	0.000000

Next, each neighborhood was filtered for its top five most common venues.

```
----ABERDEEN----
                venue  freq
0         Pizza Place  0.08
1               Café  0.07
2         Coffee Shop  0.04
3  Convenience Store  0.04
4             Bakery  0.04


----ALLSTON----
                venue  freq
0         Coffee Shop  0.06
1  Korean Restaurant  0.05
2    Thai Restaurant  0.04
3             Bakery  0.04
4         Pizza Place  0.03
```

Next, the project used this data to populate a dataframe with the top ten most popular venues, organized by town.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ABERDEEN	Pizza Place	Café	Bakery	Coffee Shop	Convenience Store	Bank	Mexican Restaurant	Donut Shop	Bus Station	Sushi Restaurant
1	ALLSTON	Coffee Shop	Korean Restaurant	Bakery	Thai Restaurant	Bubble Tea Shop	Rental Car Location	Chinese Restaurant	Pizza Place	Seafood Restaurant	Sushi Restaurant
2	ASHMONT	Grocery Store	Metro Station	Park	Farmers Market	Breakfast Spot	Mexican Restaurant	Pizza Place	Speakeasy	Caribbean Restaurant	Fast Food Restaurant
3	BEACHMONT	Liquor Store	Food Truck	Park	Sandwich Place	Gas Station	Mattress Store	Gym	Metro Station	Supermarket	Italian Restaurant
4	BELLEVUE	Home Service	Thai Restaurant	American Restaurant	Park	Mediterranean Restaurant	Gym	Grocery Store	Liquor Store	Locksmith	Convenience Store

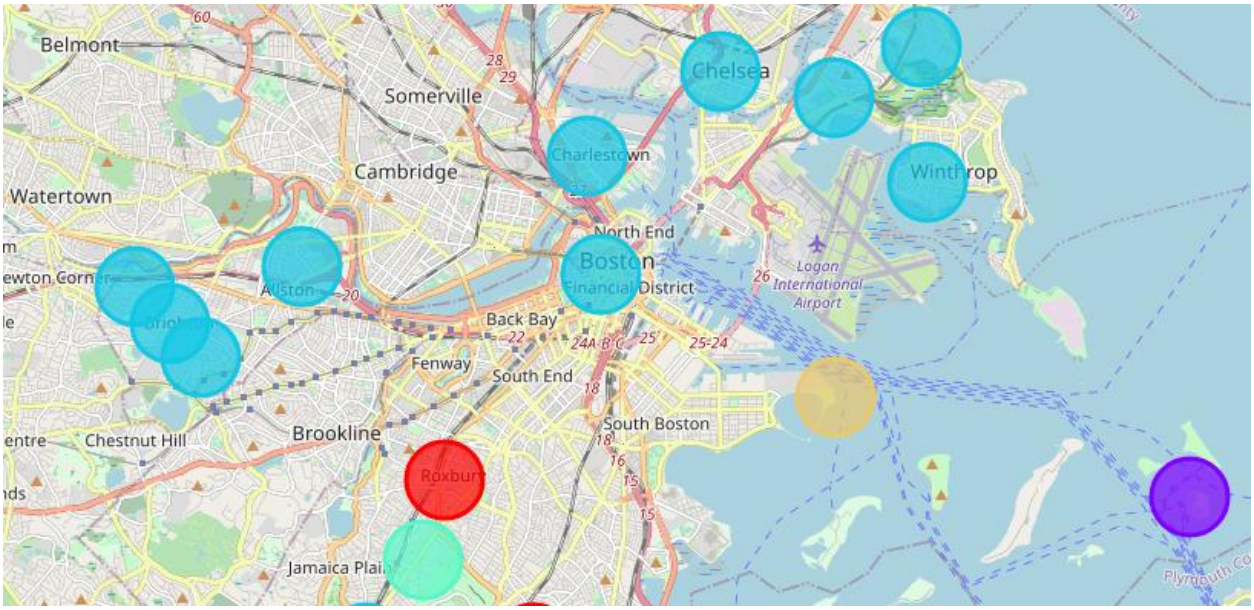
4. Clustering Neighborhoods (Machine Learning Algorithm)

4.1 Run K-means Clustering

K-means clustering is a Machine Learning Algorithm that is an unsupervised learning algorithm. It finds similarities among the data set to group the entries in to similar clusters. In this project, K-means clustering was used with 8 clusters. The results of the top row are provided below.

	COUNTY	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	25025	POINT OF PINES	42.437468	-70.965568	2	Beach	Restaurant	River	Business Service	Zoo Exhibit	Fast Food Restaurant	Frozen Yogurt Shop
1	25025	BEACHMONT	42.395601	-70.990215	3	Liquor Store	Food Truck	Park	Sandwich Place	Gas Station	Mattress Store	Gym
2	25025	REVERE	42.411107	-71.018667	3	Pharmacy	Pizza Place	Bank	Donut Shop	Shopping Mall	Chinese Restaurant	Greek Restaurant
3	25025	CHELSEA	42.391430	-71.035140	3	Hotel	Donut Shop	Grocery Store	Mexican Restaurant	Food	Harbor / Marina	Fast Food Restaurant
4	25025	ORIENT HEIGHTS	42.387261	-71.009795	3	Sandwich Place	Harbor / Marina	Cosmetics Shop	Baseball Field	Skating Rink	Food Truck	Circus

The cluster each town is assigned to can be found in the column labeled “Cluster Labels.” This describes the results of the K-means clustering algorithm following completion of the process. The results in this dataframe are used to create the final visualization of the data.



More detailed information about which towns belong in which clusters based on similarities are further provided below.

Cluster 4

```
BostonSuffolkCounty_merged.loc[BostonSuffolkCounty_merged['Cluster Labels'] == 3, BostonSuffolkCounty_merged.columns
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	
1	BEACHMONT	Liquor Store	Food Truck	Park	Sandwich Place	Gas Station	Mattress Store	Gym	Metro Station	Supermarket	
2	REVERE	Pharmacy	Pizza Place	Bank	Donut Shop	Shopping Mall	Chinese Restaurant	Greek Restaurant	Mexican Restaurant	Smoke Shop	Sl
3	CHELSEA	Hotel	Donut Shop	Grocery Store	Mexican Restaurant	Food	Harbor / Marina	Fast Food Restaurant	Pizza Place	Bank	
4	ORIENT HEIGHTS	Sandwich Place	Harbor / Marina	Cosmetics Shop	Baseball Field	Skating Rink	Food Truck	Circus	Mexican Restaurant	Pharmacy	C
5	CHARLESTOWN	Park	Café	Gastropub	Bar	Donut Shop	Pizza Place	Coffee Shop	Pub	Sandwich Place	
6	WINTHROP	Deli / Bodega	Park	Dance Studio	Bank	Pizza Place	Pharmacy	Restaurant	Construction & Landscaping	Chinese Restaurant	

4.2 Observations and Recommendations

Each cluster grouping provides the towns and popular venues. Cluster 4, which is provided above, contained the most diverse group of towns and restaurant choices. It also contained the largest group of towns, indicating that it is a robust economic zone for the restaurant business. Based on this data, it can be seen that the best location for a restaurant is among the neighborhoods in cluster four. This also provides the best opportunity to introduce a fusion-style restaurant, as various cuisine styles currently exist independent of each other.

5. Conclusions

In conclusion, the data shows that Cluster 4 contains the most variety of venues. This includes Boston and the surrounding towns. From the data set, it can be seen that the neighborhood of Allston has the most variety of restaurants within the Cluster 4. This indicates that a fusion style restaurant would have a good chance of success here, as there are many different tastes already in the area. The top two highest rated venues currently are a Korean Restaurant and a Thai Restaurant. In addition, further down the list are a Chinese Restaurant, a Seafood restaurant, a Pizza Restaurant, and a Sushi Restaurant. In order to maximize the uniqueness of my proposed Fusion style restaurant, I would recommend the bottom two venues from the list (Pizza and Sushi) for a fusion opportunity. This would allow them to combine two separate styles, and potentially increase popularity by drawing different groups of customers.