IBM Data Science Capstone
Conducting Data Analysis on Restaurant Venues near Suffolk County, MA
Brian Barry

1. Introduction

a. Background
Starting a restaurant remains fraught with failure, with the majority of these eatery ventures ending in ruin. Prior to purchasing the real estate, establishing the menu, setting prices, a potential restaurantor must understand their competition. This includes restaurant types, density of geographic locations, popularity and opportunities. This capstone will look to explore data while incorporating restaurant data from the Foursquare API to help answers these questions in the Suffolk, MA area. This sort of data manipulation would be beneficial for anyone looking to enter the restaurant business within the Suffolk, MA area, or a chain/franchise looking to expound their reach within the same area.

b. Problem
This capstone will examine restaurant venues in Suffolk, MA. The purpose of this analysis will be to optimize location and type of restaurant based on the collating off data for multiple purposes. The first have of the project will focus on data acquisition, pulling from geospatial based JSON, and foursquare APIs, creating usable data frames for the second half of the project. This is where the data frames will be broken into neighborhoods, with corresponding popularity of basic restaurant venues. From here, k-means clustering will be used for populating density, resulting in sufficient info to make the appropriate restaurant decision.

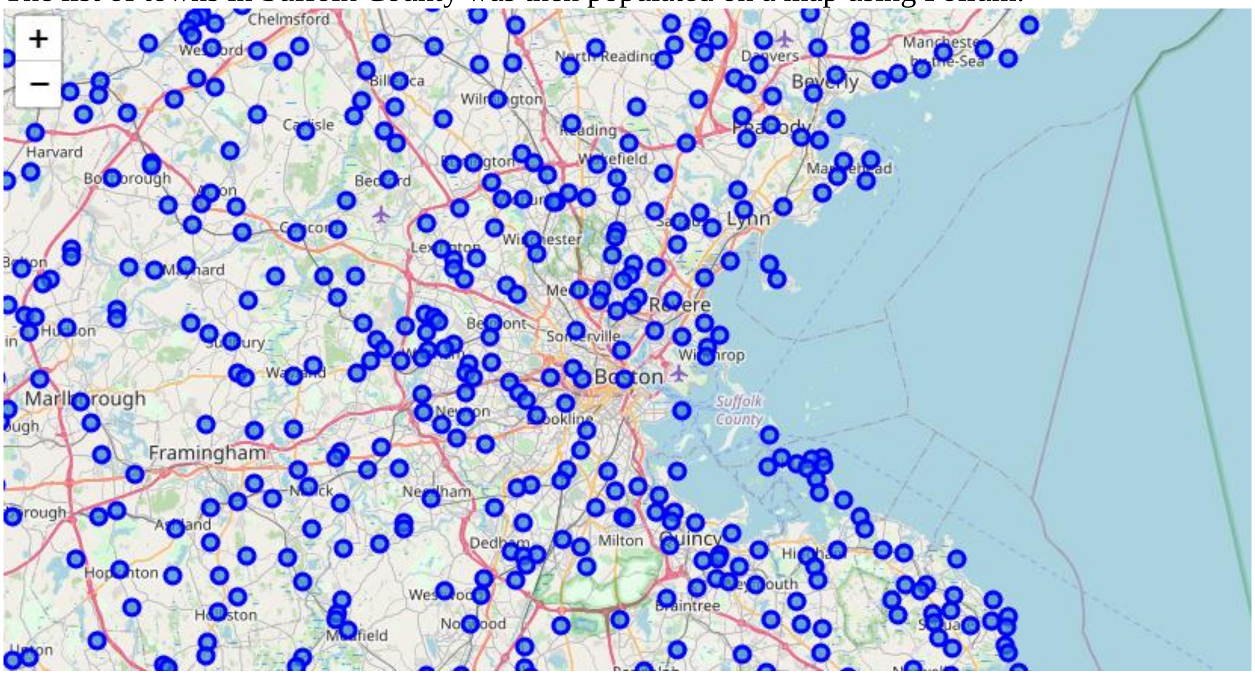
2. Data Acquisition, Management and Cleaning.

a. Data Source
This project pulled geographical data, converted into latitude and longitude from 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."
The Foursquare API will be used to provide data regarding venues within Suffol County. The explore function, providing feature groups will eventually lead to cluster analysis via k-means. Finally the folium library will be used for geographical representation.

b. Data Management and Cleaning
The data from the aforementioned sources was cleansed and combined into a useful data frame format. Using pandas, each neighborhood with the data set was eliminated with the exception of Suffolk County locations. They were additionally combined with their geographical coordinates, then cross referenced with the associated venue by popularity.

	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.



c. Feature and Venue Selection

These towns were used to set parameters within the explore function of the Foursquare API, segmented for the venues in Suffolk, MA

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

Thus began the starting point for further data analysis.

3. Data Analysis

a. Data Grouping: Venue Popularity by Neighborhood
Initial actions involved taking the average of the frequency within each category.

	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

The following step involved filtering for the top five restaurants by neighborhood.

```
----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
```

Finally, a DF with the ten most popular venues was created.

	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

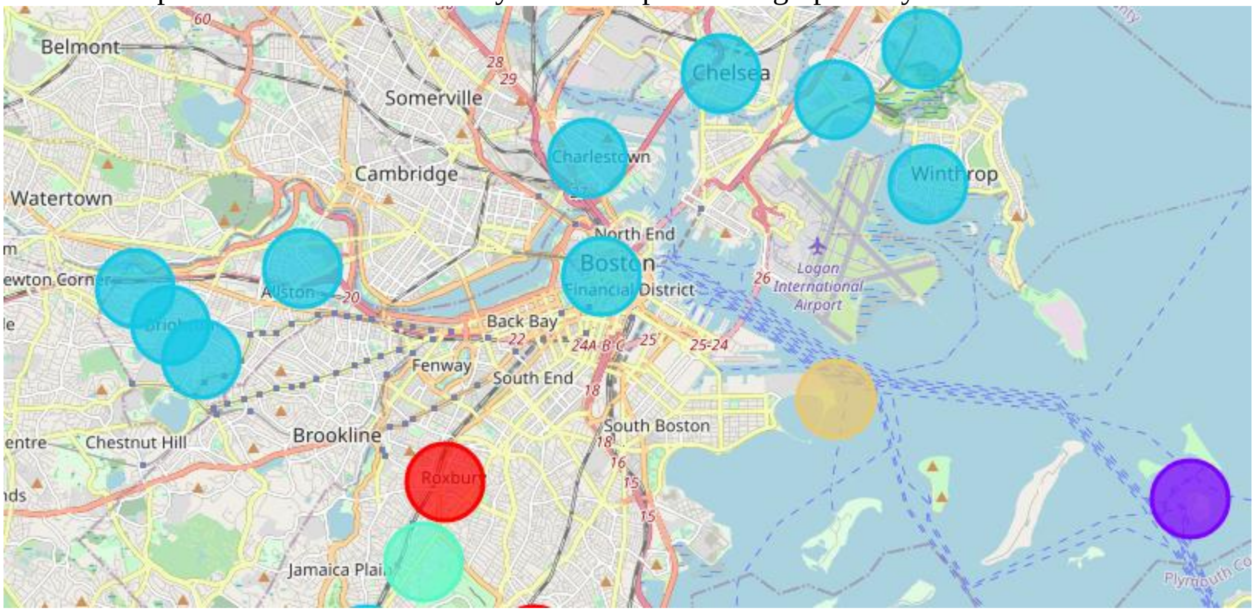
4. Clustering Data: Unsupervised Algorithm

a. K-Means

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	River	Beach	Restaurant	Business Service	Zoo Exhibit	Flower Shop	Fruit & Vegetable Store
1	25025	BEACHMONT	42.395601	-70.990215	7	Liquor Store	Park	Sandwich Place	Food Truck	Metro Station	Italian Restaurant	Beach
2	25025	REVERE	42.411107	-71.018667	1	Pharmacy	Bank	Pizza Place	Skating Rink	Café	Sandwich Place	Chinese Restaurant
3	25025	CHELSEA	42.391430	-71.035140	1	Hotel	Pizza Place	Donut Shop	Mexican Restaurant	Train Station	Bank	American Restaurant
4	25025	ORIENT HEIGHTS	42.387261	-71.009795	1	Sandwich Place	Pizza Place	Harbor / Marina	Pool Hall	Baseball Field	Café	Skating Rink
5	25025	CHARLESTOWN	42.377601	-71.065068	1	Park	Pizza Place	Bar	Café	Gastropub	Donut Shop	Pub
6	25025	WINTHROP	42.373326	-70.988690	1	Pharmacy	Park	Dance Studio	Bank	Deli / Bodega	Construction & Landscaping	Restaurant

Each cluster represents the associated density of venues per neighborhood, optimized via SSE, and used to provide the follow on analysis. It is represented graphically below.



More detailed analysis and recommendations can be drawn from the associated cluster data frames.

Cluster 2

```
BostonSuffolkCounty_merged.loc[BostonSuffolkCounty_merged['Cluster Labels'] == 1, BostonSuffolkCounty_merged.columns[[1] + li
```

	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
2	REVERE	Pharmacy	Bank	Pizza Place	Skating Rink	Café	Sandwich Place	Chinese Restaurant	Construction & Landscaping	Convenience Store	Plaza
3	CHELSEA	Hotel	Pizza Place	Donut Shop	Mexican Restaurant	Train Station	Bank	American Restaurant	Harbor / Marina	Spanish Restaurant	Discount Store
4	ORIENT HEIGHTS	Sandwich Place	Pizza Place	Harbor / Marina	Pool Hall	Baseball Field	Café	Skating Rink	Circus	Mexican Restaurant	Coffee Shop
5	CHARLESTOWN	Park	Pizza Place	Bar	Café	Gastropub	Donut Shop	Pub	Playground	Sandwich Place	National Park
6	WINTHROP	Pharmacy	Park	Dance Studio	Bank	Deli / Bodega	Construction & Landscaping	Restaurant	Chinese Restaurant	Pizza Place	Gift Shop
8	BOSTON	Coffee	Historic	Park	Italian	Bakery	Seafood	Sandwich	Restaurant	Hotel	Salad Place

b. Observations and Recommendations

- Cluster 2, an area with a preponderance corresponding to the greater Boston area, was by far the most data rich cluster.
- This includes not only the most amount of restaurants, but the largest diversity of restaurants as well.
- This likely corresponds with population density in these areas
- The most common restaurants were: Pizza Places, Café, and Asian themed restaurants

5. Conclusions

- Due to the density of cluster 2, it likely over-saturated with a customer base and will make starting a restaurant there more difficult.
- Therefore, due to the popularity of Pizza Places, it is recommended to place one in a non-cluster 2 location.
- Conversely, due to geographic analysis, the restaurants in the non-dense areas seem to be more upscale establishments, therefore this approach also seems effective.
- Finally, examining those less frequently occurring venues may offer future development.