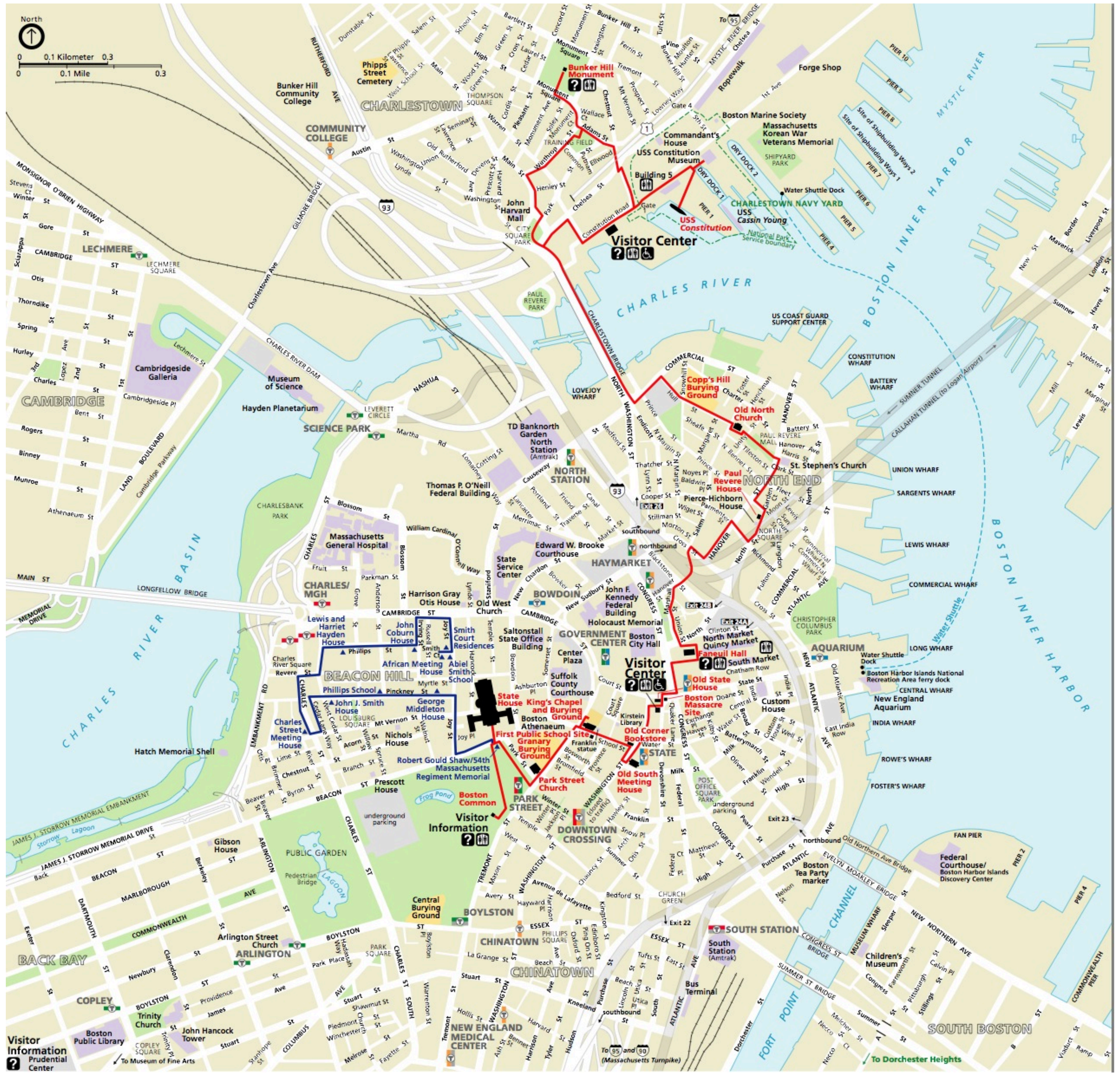
Coursera Data Science Final Project

Density and Specializations of Boston’s Restaurants

By Alexander Hahn - 8/1/2020

Introduction

- We want to discover a good place to build a new Mexican Restaurant in Boston
- We want to find a place with high density of restaurants but a low density of Mexican restaurants



Data

Foursquare and District Data

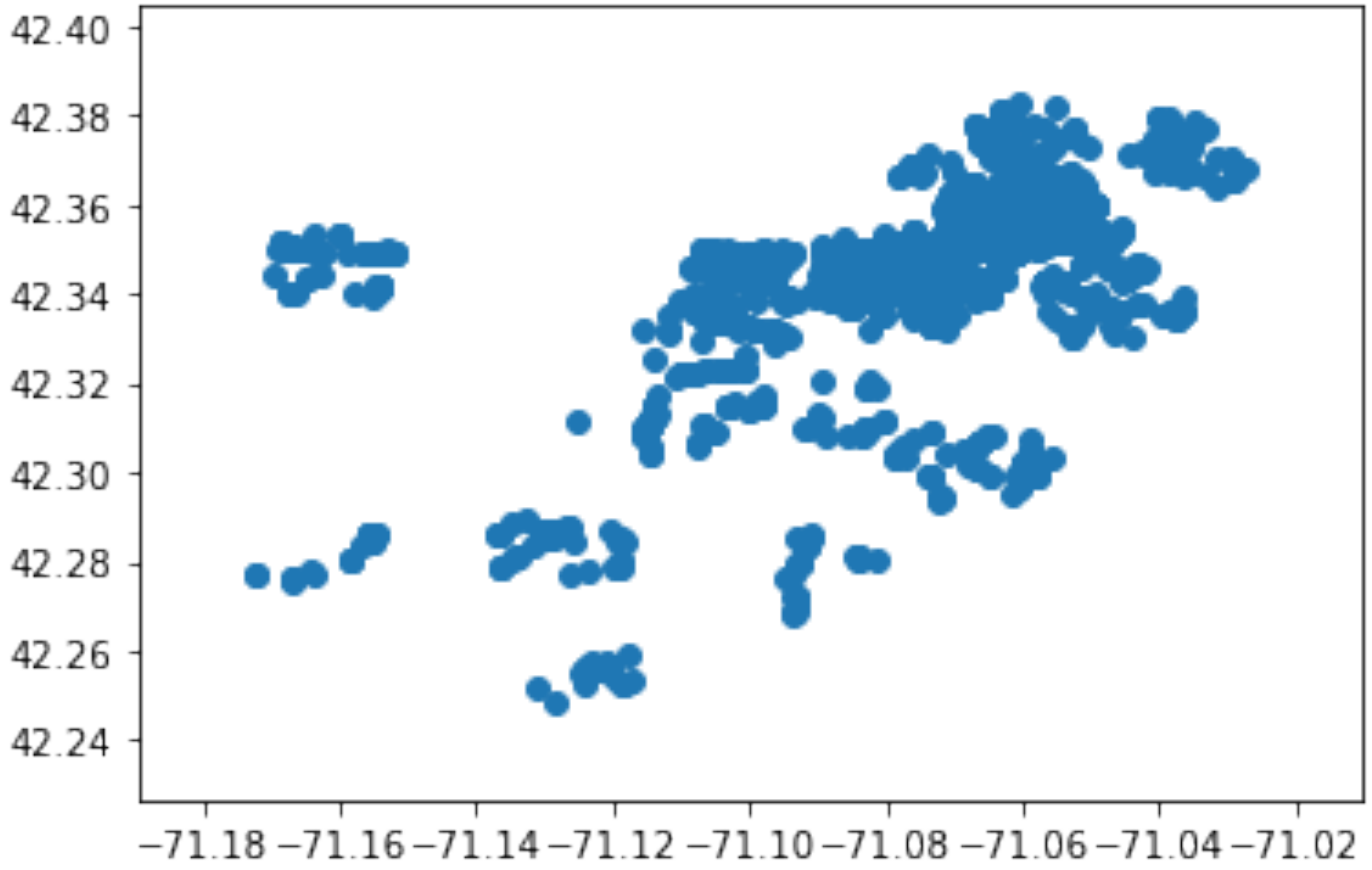
African	Thai	German	Mediterranean
American	BBQ	Greek	Middle eastern
Asian	Burger	Indian	Pizza
Chinese	Burger	Italian	Southern
Japanese	Fish	Latin	Spanish
Korean	French	Mexican	Steak

- We pulled data from an external CSV with Boston District Data
- We use Foursquare to find information on the various venues. Venue name, Restaurant Category and Longitude/Latitude will all be collected.

Exploratory Data Analysis

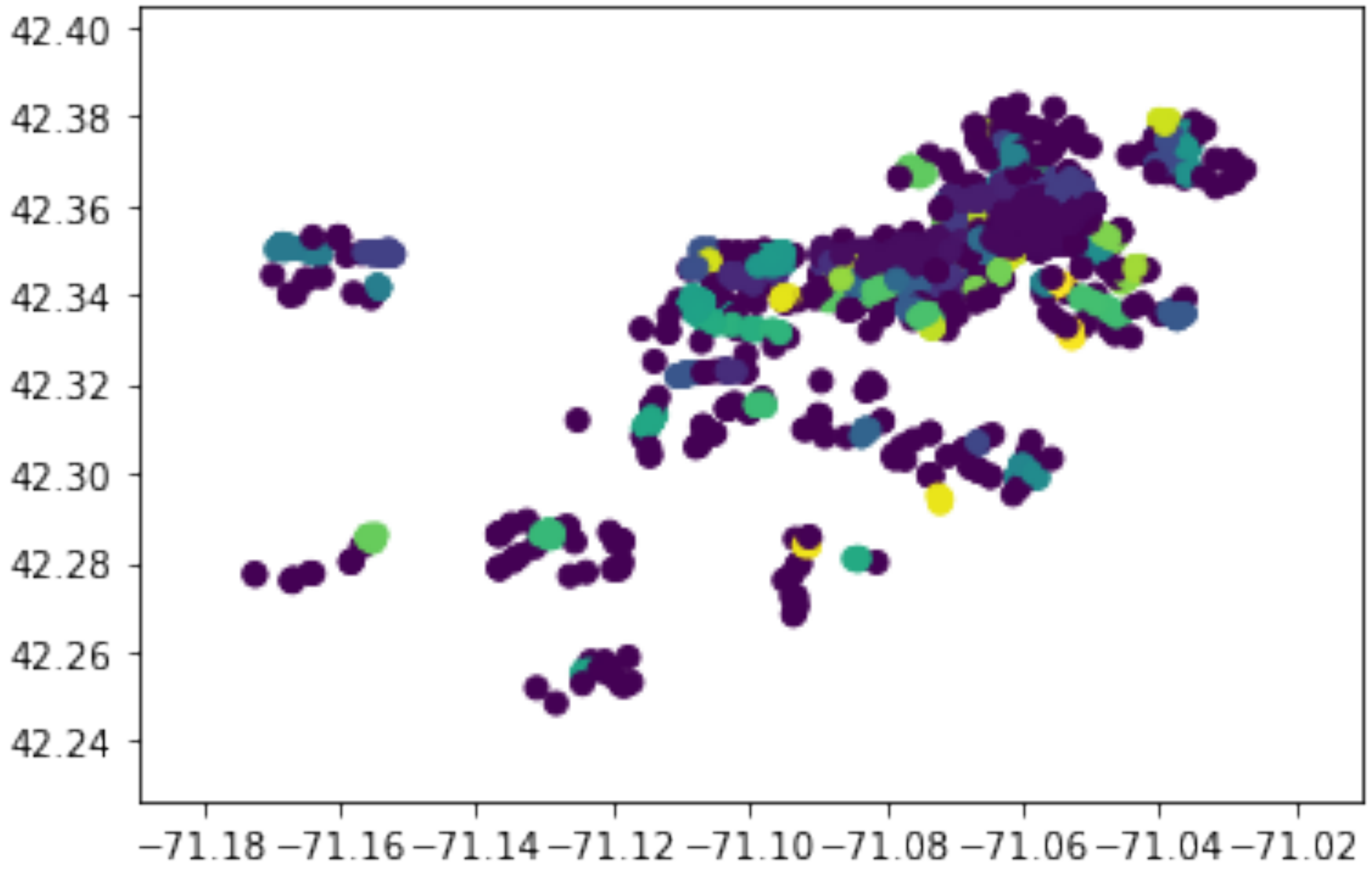
Foursquare location data

- We pulled a total amount of 57200 venue's from foursquare with a net amount of 3956 venues loaded in
- The graph on the right shows the information loaded in from foursquare showing the X/Y location data from each of the venues.



Predictive Modeling

Using DB Scan clustering



- We used DB Scan to auto-generate clusters finding 92 unique clusters using an epsilon of 0.001 and a min cluster value of 5

- Used the formula show below to calculate the value level of each district based on total number of venues and the number of unique instances in each cluster

- Value = (# of Restaurants in Cluster/Total Restaurants) * (# of Unique Restaurants/ Total # of Restaurants in cluster)

Conclusion

- The top candidate we found from the analysis was cluster 1 with a value of 0.0195.
- The next three candidates we see a sharp drop in value but still look promising.
- The 1st cluster has a Longitude and Latitude value of -71.059 and 42.356 respectively.

Cluster	Total Value	Cluster	Total Value
1	0.019473950429944400	6	0.000505816894284269 0
2	0.0032878098128477500	12	0.000505816894284269 0
32	0.0030349013657056100	0	0.000505816894284269 0
8	0.002529084471421350	7	0.000505816894284269 0
16	0.0015174506828528100	73	0.000505816894284269 0
47	0.0010116337885685400	10	0.000505816894284269 0
20	0.0010116337885685400	55	0.00025290844714213500
3	0.0007587253414264040	58	0.00025290844714213500
33	0.0007587253414264040	64	0.00025290844714213500
57	0.0005058168942842690	22	0.00025290844714213500
52	0.0005058168942842690	21	0.00025290844714213500
70	0.0005058168942842690	75	0.00025290844714213500

Future Directions

- Increased the number of samples pulled
- Introduce additional data based on population of the districts and density of people in the area.
- Could pull data in about affluence of each region to see if the area we are looking at could sustain the expense level of the restaurant we have in mind.
- Preform additional analysis combine the cluster data we have above with the sources listed above.