Capstone Report

Ian Carver

6/19/2021

Introduction

Cities like New York are very high density, which can introduce problems for prospective restaurant owners due to an abundance of competition. To open a business, they should know what type of restaurant they should open and in what area. Moreover, as their business begins to expand, they will likely want to branch out into new neighborhoods in the same city.

If an area has a lot of restaurants, people will gravitate to that area when looking for new restaurants, this is similar to the way malls work. However, restaurants must differentiate themselves from other businesses that are nearby, so they are not in direct competition.

Data

1. Location Data for New York
    a. This data, obtained from the IBM skills network, will allow me to group the restaurants into neighborhoods and calculate the density of neighborhoods.
    b.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

2. Restaurant Data
    a. This data, obtained from the Foursquare API, will allow me to determine the density of restaurants in a neighborhood, as well as calculate the least common cuisines in the neighborhood.
    b.

Venue	Venue Latitude	Venue Longitude	Venue Category
Roselle Desserts	43.653447	-79.362017	Bakery
Tandem Coffee	43.653559	-79.361809	Coffee Shop
Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center

3. Cuisine frequency

a. This data, calculated using the count of categories by neighborhood, will allow me to determine which types of cuisine are not common in a neighborhood.

b.

	Neighborhood	American Restaurant	Bagel Shop	Bakery	Burger Joint	Café	Chinese Restaurant	Deli / Bodega	Diner	Donut Shop	...	Food	Food Truck	Italian Restaurant
0	Allerton	0.047619	0.000000	0.142857	0.0	0.0	0.142857	0.190476	0.000000	0.047619	...	0.095238	0.000000	0.000000
1	Annadale	0.166667	0.000000	0.083333	0.0	0.0	0.000000	0.083333	0.083333	0.000000	...	0.083333	0.000000	0.000000
2	Arden Heights	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.500000	0.000000	0.000000	...	0.000000	0.000000	0.000000
3	Arlington	0.250000	0.000000	0.000000	0.0	0.0	0.000000	0.250000	0.000000	0.000000	...	0.000000	0.000000	0.000000

4. Restaurant density
    a. This data, calculated using the scaled number of restaurants per neighborhood, will allow me to determine how many restaurants already exist in a particular neighborhood.
    b.

	Density
Neighborhood	
Allerton	0.277778
Annadale	0.152778
Arden Heights	0.013889
Arlington	0.041667
Arrochar	0.138889

5. Combined Dataframe
    a. This data, created by merging the previous two dataframes, is what will be fed into the KNN algorithm to group neighborhoods by density and cuisine deficits.
    b.

	Neighborhood	Density	American Restaurant	Bagel Shop	Bakery	Burger Joint	Café	Chinese Restaurant	Deli / Bodega	Diner	...	Food	Food Truck	Italian Restaurant
0	Allerton	0.277778	0.047619	0.000000	0.142857	0.0	0.0	0.142857	0.190476	0.000000	...	0.095238	0.000000	0.000000
1	Annadale	0.152778	0.166667	0.000000	0.083333	0.0	0.0	0.000000	0.083333	0.083333	...	0.083333	0.000000	0.000000
2	Arden Heights	0.013889	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.500000	0.000000	...	0.000000	0.000000	0.000000
3	Arlington	0.041667	0.250000	0.000000	0.000000	0.0	0.0	0.000000	0.250000	0.000000	...	0.000000	0.000000	0.000000
4	Arrochar	0.138889	0.000000	0.181818	0.000000	0.0	0.0	0.000000	0.181818	0.000000	...	0.000000	0.090909	0.181818

6. Final Dataframe
    a. This data, created using the cluster labels from the KNN algorithm and by calculating the lowest cuisine frequency values from the previous step, will be used to determine which clusters have the most demand for food, using density, and what kinds of cuisine they still do not have.

b.

	Neighborhood	Density	Borough	Latitude	Longitude	Cluster Labels	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Least Common Venue
0	Allerton	0.277778	Bronx	40.865788	-73.859319	0.0	Thai Restaurant	Bagel Shop	Sandwich Place	Burger Joint	Café
1	Annadale	0.152778	Staten Island	40.538114	-74.178549	0.0	Fast Food Restaurant	Sandwich Place	Mexican Restaurant	Japanese Restaurant	Italian Restaurant
2	Arden Heights	0.013889	Staten Island	40.549286	-74.185887	2.0	American Restaurant	Sandwich Place	Restaurant	Mexican Restaurant	Japanese Restaurant
3	Arlington	0.041667	Staten Island	40.635325	-74.165104	0.0	Thai Restaurant	Bagel Shop	Bakery	Burger Joint	Café
4	Arrochar	0.138889	Staten Island	40.596313	-74.067124	0.0	American Restaurant	Restaurant	Mexican Restaurant	Japanese Restaurant	Food

Methodology

1. Data Selection –
    a. I chose New York city as the location for this project because it is more walkable than many other cities, meaning that customers are more likely to choose restaurants that are closer in proximity to them, in opposition to cities in the west, like Phoenix, where customers are more accustomed to driving longer distances to get food.
    b. In order to get recommendations that are more useful, I filtered the restaurant data to be in the top 50 most common types of restaurants. This ensured that restaurant categories that only had a low number of occurrences did not dominate the model.
2. KNN model – I chose to use the KNN model to group similar neighborhoods together in terms of what the demand was for restaurants in general, and by what types of restaurants they already had. This has the added benefit of giving several possible starting locations for a restaurant, as well as additional neighborhoods to expand into once the first restaurant is established.
3. Feature Selection/refinement
    a. Density – This feature is meant to measure how many restaurants already exist in a particular neighborhood, which is an indication of demand for food in that area. The feature was calculated by adding the number of venues together in each neighborhood and then Min-Max scaling between 0 and 1.
    b. Restaurant Categories – This feature is meant to measure how many restaurants of each type there are in each neighborhood. These features were also scaled using a Min-Max scaler in order to create more distance between the clusters.
    c.

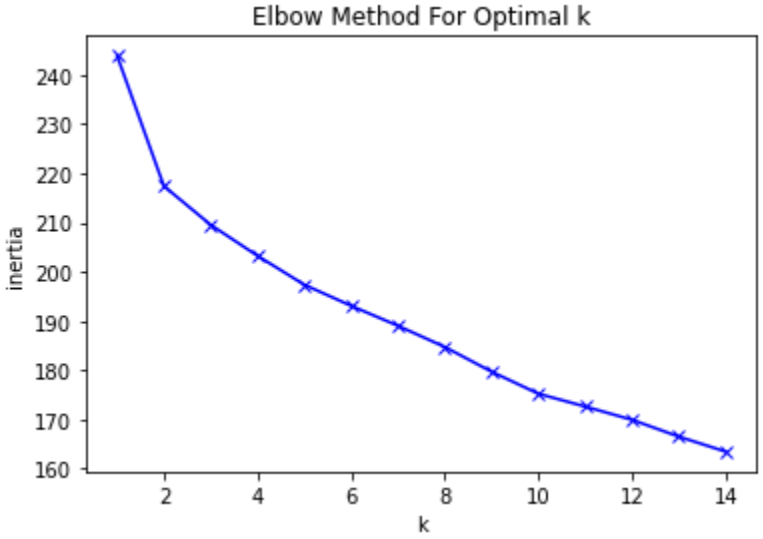
	Density	American Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Breakfast Spot	Burger Joint	Café	Caribbean Restaurant
count	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000
mean	0.263349	0.034918	0.037515	0.007115	0.052168	0.101185	0.055063	0.016725	0.067820	0.029714
std	0.257589	0.083555	0.108753	0.060034	0.110117	0.130504	0.146161	0.065258	0.123831	0.102981
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.064516	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.172043	0.000000	0.000000	0.000000	0.000000	0.053763	0.000000	0.000000	0.000000	0.000000
75%	0.376344	0.041667	0.000000	0.000000	0.058824	0.178571	0.000000	0.015625	0.106383	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

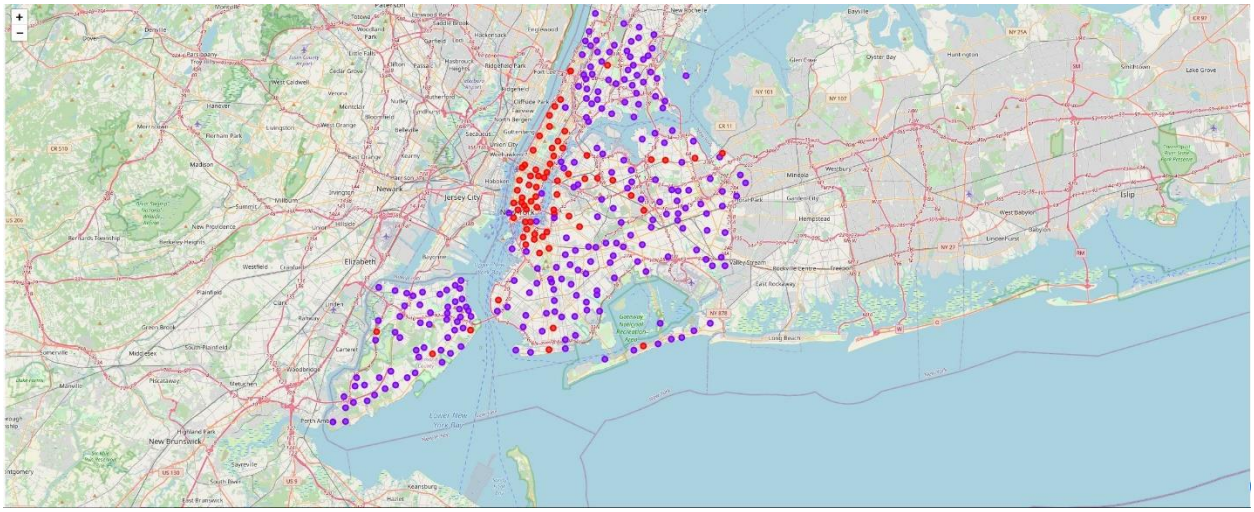
4. After the neighborhoods were clustered, it is easy to derive conclusions from the data.

a. I chose a cluster based upon the highest average density.
b. I calculated how often each restaurant occurred in the top 5 least common types of restaurants and then chose a restaurant that had the highest number of occurrences.
c. I sorted the cluster by highest density and looked for the first occurrences of the selected restaurant.

Results

The result of the KNN algorithm was two optimal clusters. I expected to be able to divide the neighborhoods into more clusters, but even after filtering and scaling the data, the result from the elbow method was still the same.





The two clusters seem to divide the city into Manhattan and then the rest of NYC, which makes sense when considering how dense Manhattan is.

Cluster 1:

Neighborhood	Neighborhood	Density	Borough	Latitude	Longitude	Cluster Labels	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Lea Commc Venu
Arrochar	Arrochar	0.161290	Staten Island	40.596313	-74.067124	0.0	American Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Korean Restaurant	Latin American Restaurant
Astoria	Astoria	0.838710	Queens	40.768509	-73.915654	0.0	American Restaurant	Tapas Restaurant	Taco Place	Steakhouse	Spanish Restaurant
Bay Ridge	Bay Ridge	0.795699	Brooklyn	40.625801	-74.030621	0.0	Food Truck	Gastropub	Fried Chicken Joint	French Restaurant	Salad Place
Bayside	Bayside	0.569892	Queens	40.766041	-73.774274	0.0	Salad Place	Restaurant	Fast Food Restaurant	Falafel Restaurant	Eastern European Restaurant
Belmont	Belmont	0.870968	Bronx	40.857277	-73.888452	0.0	Vietnamese Restaurant	Noodle House	New American Restaurant	Middle Eastern Restaurant	Snack Place
...	...	...	...	...	...	...	...	...	...	...	...
Upper West Side	Upper West Side	0.537634	Manhattan	40.787658	-73.977059	0.0	Food Truck	Spanish Restaurant	Steakhouse	Mexican Restaurant	Gastropub
Washington Heights	Washington Heights	0.752688	Manhattan	40.851903	-73.936900	0.0	Vietnamese Restaurant	Japanese Restaurant	Vegetarian / Vegan Restaurant	Middle Eastern Restaurant	Greek Restaurant
West Village	West Village	0.989247	Manhattan	40.734434	-74.006180	0.0	Vietnamese Restaurant	Southern / Soul Food Restaurant	South American Restaurant	Snack Place	Salad Place
Woodside	Woodside	0.655914	Queens	40.746349	-73.901842	0.0	Italian Restaurant	Greek Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant	Mediterranea Restaurant
Yorkville	Yorkville	0.870968	Manhattan	40.775930	-73.947118	0.0	Fast Food Restaurant	Middle Eastern Restaurant	Salad Place	Latin American Restaurant	Korean Restaurant

Cluster 2:

Neighborhood	Neighborhood	Density	Borough	Latitude	Longitude	Cluster Labels	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Lea Comm Ven
Allerton	Allerton	0.247312	Bronx	40.865788	-73.859319	1.0	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Latin American Restaurant	Mediterrane Restaurant
Annadale	Annadale	0.118280	Staten Island	40.538114	-74.178549	1.0	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Latin American Restaurant	Mediterrane Restaurant
Arden Heights	Arden Heights	0.010753	Staten Island	40.549286	-74.185887	1.0	American Restaurant	Korean Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant
Arlington	Arlington	0.043011	Staten Island	40.635325	-74.165104	1.0	Italian Restaurant	Korean Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant
Arverne	Arverne	0.064516	Queens	40.589144	-73.791992	1.0	American Restaurant	Korean Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant
...	...	...	...	...	...	...	...	...	...	...	...
Windsor Terrace	Windsor Terrace	0.301075	Brooklyn	40.656946	-73.980073	1.0	Vietnamese Restaurant	Ramen Restaurant	Restaurant	Mediterranean Restaurant	Latin American Restaurant
Wingate	Wingate	0.161290	Brooklyn	40.660947	-73.937187	1.0	American Restaurant	Korean Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant
Woodhaven	Woodhaven	0.215054	Queens	40.689887	-73.858110	1.0	American Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Korean Restaurant	Mediterrane Restaurant
Woodlawn	Woodlawn	0.161290	Bronx	40.898273	-73.867315	1.0	Vietnamese Restaurant	Steakhouse	Vegetarian / Vegan Restaurant	Japanese Restaurant	Korean Restaurant
Woodrow	Woodrow	0.096774	Staten Island	40.541968	-74.205246	1.0	American Restaurant	Japanese Restaurant	Korean Restaurant	Latin American Restaurant	Mediterrane Restaurant

Discussion

Looking at the Categories from the Foursquare API, I noticed that not all the categories are optimal for this project.

Deli / Bodega	719
Pizza Place	719
Chinese Restaurant	485
Italian Restaurant	463
Bakery	353
Café	299
Sandwich Place	294
Mexican Restaurant	287
American Restaurant	254
Donut Shop	223
Sushi Restaurant	189
Food Truck	175
Restaurant	167
Bagel Shop	156
Fast Food Restaurant	156
Diner	145
Japanese Restaurant	141
Burger Joint	132
Thai Restaurant	128
Fried Chicken Joint	119
Seafood Restaurant	118
Food	116
Korean Restaurant	112
Indian Restaurant	109
Latin American Restaurant	108
Spanish Restaurant	108
Asian Restaurant	102
Caribbean Restaurant	101
French Restaurant	93
Mediterranean Restaurant	74
Breakfast Spot	67
Greek Restaurant	64
Steakhouse	60
Vegetarian / Vegan Restaurant	58
Vietnamese Restaurant	57
Taco Place	56
New American Restaurant	55
Middle Eastern Restaurant	53
Salad Place	44
Ramen Restaurant	35
South American Restaurant	32
BBQ Joint	32
Falafel Restaurant	30
Noodle House	28
Southern / Soul Food Restaurant	28
Gastropub	27
Cuban Restaurant	26
Eastern European Restaurant	24
Tapas Restaurant	24
Snack Place	24

A couple of examples of this are “Food” and “Snack Place” which do not adequately describe the type of food being sold. Additionally, categories like “Fast Food” may include restaurants which are not in direct competition with one another. Categories like “Mexican Restaurant”, “Latin American Restaurant”, and “Taco Place” seem like they should be categorized together, rather than separately.

As far as the end result of the project, deriving one possible solution to the problem was fairly easy. I simply selected the higher-density cluster, found out what types of restaurants occurred most often in the top 5 least common for the neighborhoods in that cluster, and then found several neighborhoods that would be good to start in based upon their current demand for food.

```
cluster1 = ny_merged[ny_merged['Cluster Labels'] == 0]
cluster2 = ny_merged[ny_merged['Cluster Labels'] == 1]
print(cluster1.describe())
print(cluster2.describe())
```

```
         Density   Latitude  Longitude  Cluster Labels
count  70.000000  70.000000  70.000000            70.0
mean    0.638095  40.726893 -73.961133             0.0
std     0.237562   0.059715   0.066896             0.0
min     0.161290  40.572572 -74.189560             0.0
25%     0.438172  40.703203 -73.994155             0.0
50%     0.645161  40.732317 -73.969531             0.0
75%     0.862903  40.763165 -73.947760             0.0
max     1.000000  40.857277 -73.738898             0.0
          Density    Latitude   Longitude  Cluster Labels
count  227.000000  227.000000  227.000000           227.0
mean     0.153190   40.696471  -73.941442             1.0
std      0.125691    0.104418    0.132612             0.0
min      0.000000   40.505334  -74.246569             1.0
25%      0.053763   40.613198  -74.049328             1.0
50%      0.129032   40.675211  -73.912585             1.0
75%      0.231183   40.782067  -73.847614             1.0
max      0.688172   40.908543  -73.708847             1.0
```

	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Least Common Venue	Totals
Vietnamese Restaurant	26.0	0.0	0.0	0.0	0.0	26.0
Vegetarian / Vegan Restaurant	0.0	7.0	9.0	5.0	3.0	24.0
Korean Restaurant	0.0	0.0	6.0	7.0	7.0	20.0
Latin American Restaurant	0.0	4.0	4.0	7.0	4.0	19.0
Greek Restaurant	1.0	6.0	4.0	3.0	3.0	17.0

	Neighborhood	Density	Borough	Latitude	Longitude	Cluster Labels	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Least Common Venue
174	Midtown South	1.000000	Manhattan	40.748510	-73.988713	0.0	Vietnamese Restaurant	Seafood Restaurant	Noodle House	Middle Eastern Restaurant	Mexican Restaurant
150	Lenox Hill	1.000000	Manhattan	40.768113	-73.958860	0.0	Vietnamese Restaurant	Snack Place	Food	Fast Food Restaurant	Falafel Restaurant

Interestingly enough, the top two categories in this solution were Vietnamese and Vegetarian/Vegan restaurants, which aren't uncommon in the dataset overall, with New York having 57 and 58 of those types of restaurants, respectively.

One odd thing about the low-density cluster was that the least common restaurants seemed to be American and Italian, which are some of the most common restaurants overall. I think this is because these areas have so few restaurants overall.

Conclusion

In conclusion, I think that this project was successful. If a prospective restaurant owner took a more naive approach to the problem they might simply choose their first location based upon which neighborhoods have the highest density and their type of restaurant based off of what that neighborhood lacks. This solution might be optimal in the short term, but in the long term they would likely run into issues if they wanted to expand in the future.

I think if I was to repeat this type of project, I might use restaurant data from a different source in order to eliminate some of the non-descriptive and repetitive categories mentioned earlier. This would likely improve the result of the algorithm.

Additionally, repeating this project on a different scale in order to control for things like population density or geographic area. Perhaps limiting the geographic area to a single borough of New York, or a downtown area in another city would be the easiest way to achieve this.