Capstone project

The Battle of Neighbourhoods

Introduction and context

In this (hypothetical) problem a friend of mine has quit his Wall Street job after having worked for a bank for 10+ years. Wanting a change from corporate culture and having a bit of spare money to invest, but unable to leave the City he's looking into opening a Hungarian restaurant in New York. While most of his friends think this is at least risky if not borderline madness, New York is probably the only and best place to open a place with a less known and somewhat less health-conscious cuisine.

New York has more than 8.5 million people, while the greater metropolitan area's population is more than 23 million, with neighborhoods home to an extremely vast majority of places from eateries through hole in the wall places up to 2-3 michelin star fine dining restaurants. According to the New York City Department of Health and statista (https://www.statista.com/statistics/259776/number-of-people-who-went-to-restaurants-in-new-york-by-type/) there were more than 26'000 restaurants in the City in 2017, this gives a glimpse of hope for having yet another obscure place to make ends meet.

The ask was to analyze where it may make sense to open such a restaurant. New York is very diverse with no very obvious concentration of cuisines (apart from Chinatown and some Flushing Meadows districts), so it requires further analysis to see if there is a trend in concentration of small cuisines or maybe even Easter European block in places.

Business problem

There are a few ways to approach this problem, I'll take 3 here and based on the later data analysis it may be possible to pick one (or may not). The approaches are:

1. Chinatown approach

New York's Chinatown is one the largest of its kind with a massive number of prospering restaurants. Obviously not only residents and people of Chinese origin visit these, but is famous among visitors, tourists and in general as well. This approach assumes that if there are areas with concentrated Hungarian and in a broader sense Eastern European restaurants another one can still fit in as people do visit these parts to eat a particular dish. As Hungarian is a small

portion of the city's population with a lesser known cuisine it does seem to make sense to extend the radius with similar cuisines as well. (admittedly with a subjective list)

2. Go against the current

This is the direct opposite - seeing if there are places with no or very limited number of similar restaurants. In other cities this may be a plain bad approach as purely residential areas, suburbs or other industrial districts would not be a good fit, but NYC, especially Manhattan is so packed with restaurants that this may not be an issue there. However, if such area is found, it does make sense to see how many / how concentrated the place is.

3. Go with the flow

This approach will simply look at the areas with the highest concentration of restaurants (assuming it also is proportional to their variety) and will recommend to set up a place where there are already a lot, as people do go there to eat and is a well-established neighbourhood from this point of view.

More specifically during the exercise the following questions will be answered:

- What are the particular areas with high concentration of Hungarian restaurants?
- In a more generic sense, what are the areas with high concentration of Eastern European (Hungarian, Czech, Slovakian, Polish, Romanian) cuisines?
- Which areas do not have Eastern European restaurants?
- What are the areas with the highest concentration of restaurants?

Description of the data and how it'll be used

I will essentially be using the same data sets from the previous week's exercises, as follows:

New York neighbourhood data, latitude, longitude information

- Source: https://cocl.us/new_york_dataset
- Usage: mapping restaurant data including address to borough and neighbourhood for classification

Hungarian and Eastern European restaurants in New York City

- Source: Foursquare API
- Usage: getting the list of Hungarian and Eastern European restaurants in NYC for each neighbourhood

GeoSpacial data

- Source: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm
- Usage: using for neighbourhood boundaries for visualization`

Methodology

For this particular example the methodology is quite straightforward and analyzing the data is not overly complex either. The steps to achieve the desired answers and to draw some conclusions (if possible) are the following:

1. Gather and clean / scrape New York neighbourhood data
    a. Read data from https://cocl.us/new_york_dataset
    b. Read and map geo data (latitude and longitude) to each Neighborhood and borough
2. Read the list of Hungarian and Eastern European restaurants from the Foursquare API
3. Do some exploratory data analysis to see if the scope of the analysis could be or should be changed (i.e.) remove Neighborhoods or find areas to focus on
4. Run a simple analysis on neighbourhoods to see which ones have the highest concentration of restaurants
5. Use the K-Means method to cluster neighbourhoods and get answers for the distribution of Eastern European restaurants.

Reading New York data

In [95]:
```
def get_new_york_data():
    url='https://cocl.us/new_york_dataset'
    resp=requests.get(url).json()
    # all data is present in features label
    features=resp['features']

    # define the dataframe columns
    column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']
    # instantiate the dataframe
    new_york_data = pd.DataFrame(columns=column_names)

    for data in features:
        borough = data['properties']['borough']
        neighborhood_name = data['properties']['name']

        neighborhood_latlon = data['geometry']['coordinates']
        neighborhood_lat = neighborhood_latlon[1]
        neighborhood_lon = neighborhood_latlon[0]

        new_york_data = new_york_data.append({'Borough': borough,
                                              'Neighborhood': neighborhood_name,
                                              'Latitude': neighborhood_lat,
                                              'Longitude': neighborhood_lon}, ignore_index=True)

    return new_york_data
```

In [96]:
```
#reading new york data
new_york_data=get_new_york_data()
```

In [97]:
```
new_york_data.shape
```

Out[97]: (306, 4)

```
In [98]: new_york_data.head()
```

Out[98]:

	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

Once we have all the 306 neighborhoods in the dataframe we can check the Foursquare API for Hungarian and similar (Czech, Polish, Slovakian and Romanian) cuisines and restaurants.

Due to Foursquare APIs responsiveness (or the lack of it) and to save some running time and bandwidth the list of venues is already filtered for related categories. The categories can be picked by simply calling the following url on the Foursquare API:

```
GET https://api.foursquare.com/v2/venues/categories
```

(for more information: https://developer.foursquare.com/docs/api-reference/venues/categories/)

Once the categories are filtered from the json they are used as parameters in the getvenue methods.

Finding Hungarian and similar restaurants

As we are interested not only in the number but also the concentration of said restaurants in New York city we need to read the Foursquare API for all the neighborhoods. Please note that this will issue 306 queries to the API so the sandbox or personal account (and its licence) limits can quickly be reached.

In [12]:
```
#so there are 306 new york neighborhoods, need to look up Hungarian and other central / eastern european restaurants
# reading Hungarian, Romanian, Czech, Slovakian and Polish restaurants in one go rather than having 5 calls to the API.
#(While there is an Eastern European category in Foursquare it is not the one we're looking for)

column_names=['Borough', 'Neighborhood', 'Latitude','Longitutde', 'Hun', 'CEE']
restaurants = pd.DataFrame(columns = column_names)

cee = ["Hungarian Restaurant","Czech Restaurant","Slovak Restaurant", "Polish Restaurant", "Romanian Restaurant"]

#iterating through all neighborhoods, getting venues and adding 2 columns, Hun count and CE count (Hun + all others)

for row in new_york_data.values.tolist():
    Borough, Neighborhood, Latitude, Longitude=row
    venues = get_venues(Latitude,Longitude)
    hun_restaurants=venues[venues['Category']=='Hungarian Restaurant']
    cee_restaurants=venues[venues['Category'].isin(cee)]

    print('Hungarian Resturants in '+Neighborhood+', '+Borough+':'+str(len(hun_restaurants)))
    print('CEE Resturants in '+Neighborhood+', '+Borough+':'+str(len(cee_restaurants)))
    #adding a new row to our restaurants frame
    restaurants = restaurants.append({'Borough': Borough,
                        'Neighborhood': Neighborhood,
                        'Latitude': Latitude,
                        'Longitutde' : Longitude,
                        'Hun' : len(hun_restaurants),
                        'CEE' : len(cee_restaurants)
                                        }, ignore_index=True)
```

```
Hungarian Resturants in Wakefield, Bronx:0
CEE Resturants in Wakefield, Bronx:0
Hungarian Resturants in Co-op City, Bronx:0
CEE Resturants in Co-op City, Bronx:0
Hungarian Resturants in Eastchester, Bronx:0
CEE Resturants in Eastchester, Bronx:0
```

```
CEE Resturants in Bronxdale, Bronx:0
Hungarian Resturants in Allerton, Bronx:0
CEE Resturants in Allerton, Bronx:0
Hungarian Resturants in Kingsbridge Heights, Bronx:0
CEE Resturants in Kingsbridge Heights, Bronx:0
Hungarian Resturants in Erasmus, Brooklyn:0
CEE Resturants in Erasmus, Brooklyn:0
Hungarian Resturants in Hudson Yards, Manhattan:0
CEE Resturants in Hudson Yards, Manhattan:0
Hungarian Resturants in Hammels, Queens:0
CEE Resturants in Hammels, Queens:0
Hungarian Resturants in Bayswater, Queens:0
CEE Resturants in Bayswater, Queens:0
Hungarian Resturants in Queensbridge, Queens:0
CEE Resturants in Queensbridge, Queens:0
Hungarian Resturants in Fox Hills, Staten Island:0
CEE Resturants in Fox Hills, Staten Island:0
```

In [13]:
```
restaurants.shape
```

Out[13]: (306, 6)

In [14]:
```
restaurants.head()
```

Out[14]:

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

The results are appalling, the original assumption that there will be a few Hungarian restaurants many Central / Eastern European in New York turned out to be false. While there may be in New Jersey (i.e. https://www.greenpeargroup.com/locations) there are such places in NYC. Moreover, there are only 9 (out of more than 26'000) restaurants that serve Czech, Polish, Slovakian or Romanian food. Whether this is due to the main immigration concentration of those nations being outside the City or not is an interesting question, but is beyond the scope of the current exercise.

Exploratory data analysis

The outcome of reading all Hungarian and Central / Eastern European restaurants in New York is somewhat surprising. In the given categories only 9 restaurants were found, while there was no dedicated Hungarian restaurant in New York City. This is somewhat contradicting with the assumtions, as there are some places in New Jersey and there seem to be some Hungarian places present in NYC as well. Further analysis, however, showed that those are mainly cateogized as i.e. bakery and others under food. As the original business problem was to analyze if a restaurant is feasible, those are ommited now.

The distribution of Eastern European restaurants does not yield any significant result either, as those are:

- Greenpoint / Brooklyn: 3
- Arrochar / Staten Island: 1
- Blissville / Queens: 1
- Lenox Hill / Manhattan: 1
- Ridgewood / Queens: 1
- Roosevelt Island / Manhattan: 1
- Steinway / Queens: 1

Conclusion 1

Given the very low number of restaurants we can safely state that neither the Chinatown model (flocking to the same type of restaurants) seem to be doable, nor the 'Go against the current', when the owner explicitly chooses a place where no similar restaurants are present are really options.

Conclusion 2

We should recommend a location with a very high density of restaurants so it is likely to get enough visitors - the place is already known and liked for its food selection and variety. For this we will analyze the restaurant density for Manhattan only.

Analysis of Manhattan restaurants

Since there are not enough samples for the dedicated cuisine type we will need to look into Manhattan restaurants and find the neighborhoods with the highest density or the highest number of restaurants as those may provide the highest probability of success for a new place.

The method to find the best neighborhood is based on finding the top 100 venues for each of them and analyze what proportion of those are restaurants, running a simple clustering exercise.

276	Manhattan	Flatiron	40.739673	-73.990947
301	Manhattan	Hudson Yards	40.756658	-74.000111

In [44]:
```
manhattan_venues = getNearbyVenues(names=manhattan_hoods['Neighborhood'],
                                   latitudes=manhattan_hoods['Latitude'],
                                   longitudes=manhattan_hoods['Longitude']
                                  )

manhattan_venues
```

2973	Hudson Yards	40.756658	-74.000111	Playboy Club New York	40.760000	-73.996367	Lounge
2974	Hudson Yards	40.756658	-74.000111	Cachet Boutique Hotel	40.759773	-73.996460	Hotel
2975	Hudson Yards	40.756658	-74.000111	Silver Towers Dog Run	40.760854	-73.999765	Dog Run
2976	Hudson Yards	40.756658	-74.000111	Treadwell	40.759964	-73.996284	Restaurant
2977	Hudson Yards	40.756658	-74.000111	George's	40.757760	-74.000963	Burger Joint
2978	Hudson Yards	40.756658	-74.000111	Big George's Smokehouse	40.757805	-74.001660	BBQ Joint
2979	Hudson Yards	40.756658	-74.000111	Unlimited Biking	40.759560	-74.003975	Athletics & Sports
2980	Hudson Yards	40.756658	-74.000111	NY Waterway 42nd St Bus	40.760050	-74.003379	Bus Station
2981	Hudson Yards	40.756658	-74.000111	Gray Line New York Sightseeing Cruises - Pier 78	40.759721	-74.003982	Harbor / Marina
2982	Hudson Yards	40.756658	-74.000111	Twilight Cruise By Citysightseeing	40.759744	-74.004096	Boat or Ferry
2983	Hudson Yards	40.756658	-74.000111	City Lights Cruises	40.759804	-74.004025	Boat or Ferry

2984 rows × 7 columns

In [ ]:

Next step is to analyze the 2984 venues in 321 unique categories. This latter number is quite large and shows the large diversity of venues in the city.

Running one hot encoding all the venue data:

In [46]:
```
# one hot encoding
manhattan_onehot = pd.get_dummies(manhattan_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
manhattan_onehot['Neighborhood'] = manhattan_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [manhattan_onehot.columns[-1]] + list(manhattan_onehot.columns[:-1])
manhattan_onehot = manhattan_onehot[fixed_columns]

manhattan_onehot.head()
```

Out[46]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Video Store	Vietnamese Restaurant	Volleyball Court	Waterfront	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yo Stu
0	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
1	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
2	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
3	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
4	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	

5 rows × 322 columns

Clustering all the neighborhoods into **3 clusters** (this seems adequate to exclude obviously different hoods and to be able to make a recommendation) shows:

In [57]:
```
# set number of clusters
kclusters = 3

manhattan_grouped_clustering = manhattan_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(manhattan_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

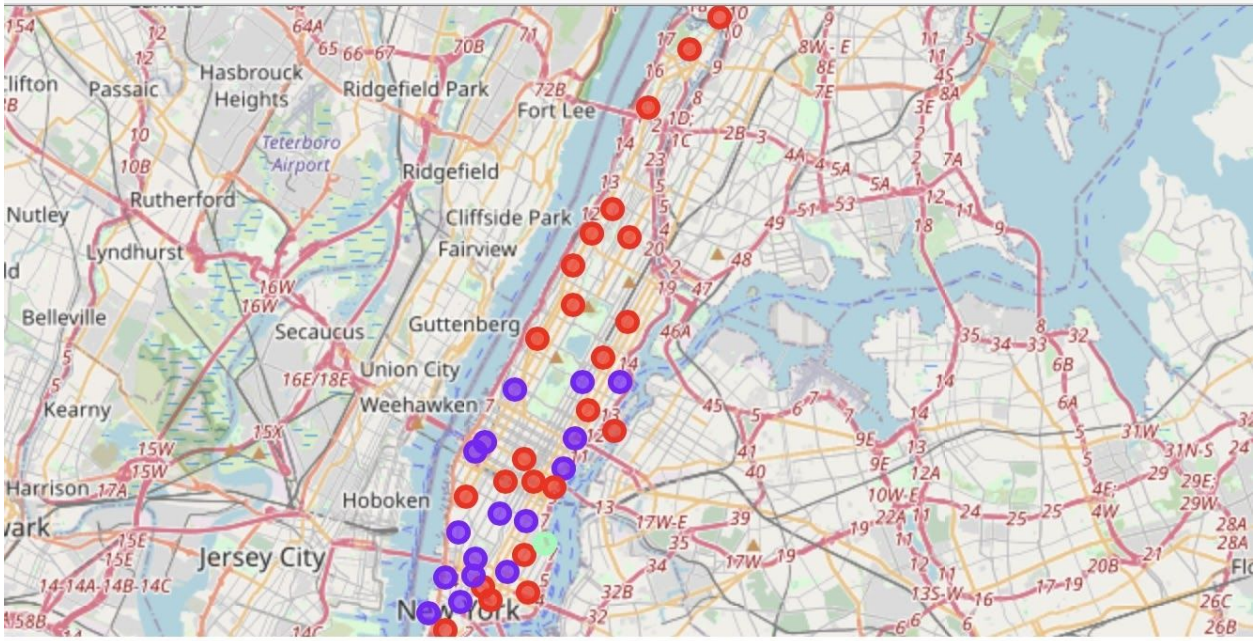
Out[57]:
```
array([1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 1, 2, 1, 1, 0, 1, 1, 0, 0, 1, 1], dtype=int32)
```

In [58]:
```
#merge the dataframes together with the 10 most frequent and the cluster value
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

manhattan_merged = manhattan_hoods

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
manhattan_merged = manhattan_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

manhattan_merged.head()
```

Out[58]:

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Common
6	Manhattan	Marble Hill	40.876551	-73.910660	0	Sandwich Place	Gym	American Restaurant	Coffee Shop	Yoga Studio	Deli / Bodega	Supplement Shop	Steakhouse	Seafood Restaurant	Pizza
100	Manhattan	Chinatown	40.715618	-73.994279	0	Chinese Restaurant	Bakery	Cocktail Bar	Coffee Shop	Spa	American Restaurant	Salon / Barbershop	Optical Shop	Bar	Din Rest
101	Manhattan	Washington Heights	40.851903	-73.936900	0	Café	Bakery	Mobile Phone Shop	Pizza Place	Grocery Store	Chinese Restaurant	Latin American Restaurant	Tapas Restaurant	New American Restaurant	
102	Manhattan	Inwood	40.867684	-73.921210	0	Mexican Restaurant	Café	Bakery	Pizza Place	Lounge	Restaurant	Park	Chinese Restaurant	Deli / Bodega	Am Rest
103	Manhattan	Hamilton	40.823604	-73.949688	0	Pizza Place	Coffee Shop	Mexican	Café	Deli / Bodega	Chinese	Sushi	Cocktail Bar	Yoga Studio	Cari

And visualizing it shows that the clusters are not really representatives either:



Just by looking at it we can see that apart from one cluster (Cluster 2, see later) the only distinction we can make is that Cluster 1 seems to be more concentrated on the lower side of Manhattan (Tribeca, Greenwich, SoHo), while Cluster 0 is more Upper West and FiDi.

Analyzing the clusters in more details shows that we can safely remove cluster 2 as it's more of a harbour:

Cluster 2

```
In [66]: manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 2, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))]]
```

Out[66]:

	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
275	Stuyvesant Town	Boat or Ferry	Park	Bar	Pet Service	Gas Station	Farmers Market	German Restaurant	Gym / Fitness Center	Baseball Field	Harbor / Marina

The main difference between Cluster 0 and Cluster 1 is that Cluster 1 is mainly dominated by Italian restaurants, has more wine bars, theaters, art galleries, performance venues and in

general more upscale types of places. There are a few cafes, gyms and delis, but in general the top 10 types gravitate towards more expensive places.

Cluster 0 shows a somewhat more varied and in general a less upscale picture, with cafes, bakeries, sandwich places, pizza places being way more prominent and being the most frequent.

So as further shown in the next section, the selection of the neighborhood will essentially boil down to the owner's preference of the type of restaurant to be opened. If a more upscale, fine dining type of niche place is the preference then some of the more upscale neighborhoods from Cluster 1 would be the best choice, if he / she thinks about opening a more eatery-type place than some of the Cluster 0 hoods look like a good choice.

Results

The analysis for finding a good place for a new Hungarian restaurant yielded a bit disappointing result, there are not enough similar places in NYC to be able to draw a definite conclusion on that. Further analysis showed - based on clustering - that the recommended area would be based on whether the new place would be more of an eatery type or rather a more upscale restaurant.

The answers for the original questions are:

1. What are the particular areas with high concentration of Hungarian restaurants?

There are no areas where Hungarian restaurants are concentrated. As a matter of fact, the inner city does not really have one.

2. In a more generic sense, what are the areas with high concentration of Eastern European (Hungarian, Czech, Slovakian, Polish, Romanian) cuisines?

Somewhat surprisingly there are only 9 such restaurants in New York City, and the most concentrated neighborhood has 3, which is not significant. Hence decision making based on existing places' concentration is not possible

3. Which areas do not have Eastern European restaurants?

As an inverse to the previous question, there are only 7 neighborhoods with Central European cuisine, so it is not possible to further limit the recommended area from the other almost 300.

4. What are the areas with the highest concentration of restaurants?

A simple clustering showed 2 somewhat different groups of restaurants, one mostly (with some prime exceptions) concentrating a bit on the lower side of Manhattan consisting of more upscale places, the other more scattered toward more residential areas as well, however, having more smaller places like eateries or cafes.

My recommendation would be - based on the above - that if the owner wants to open

a more upscale place than pick one of the following neighborhoods:

- Upper East Side
- Greenwich Village
- Soho
- West Village

For a more **eatery-like place, i.e. cafe, deli** one of the following areas seem to be suitable

- Marble Hill
- Washington Heights
- East Village

Discussion

According to the above analysis and further manual data exploration there is not a single place that I would recommend. From a practical point of view there seems to be some distinction between the 2 clusters, but they are not as significant as i.e. between those and Cluster 3 so it can be somewhat risky to deduct very

New York has an extraordinarily high number of restaurants and while some cuisines (Chinese, Korean and some Indian) are more concentrated, the rest is extremely scattered, probably (speculation though) to accommodate the residents comfort level not to walk too far for a decent dish.

My original assumption to find the place where there are already existing and well-known places failed as there is no concentration, so the next recommendation is to find a small place to open where there are already many existing ones in a well known and frequented area.

The two separate clusters have a thin line between them essentially separating a little cheaper places from the more upscale or trendy ones, so this may be the distinguishing factor.

These somewhat rudimentary recommendations, however, can serve as the basis for the owner to move forward with a business case with anticipated revenue, visits, rents, fees, etc.

To improve the model one could run some further analysis to see if some neighborhoods have some other distinguishing factors (hotels, concert halls, landmarks or other non-food related venues) and position the place accordingly.

Conclusion

This project was a somewhat simplified hypothetical example of a practical data science problem and while the result is not what I was expecting when starting it it still proved to be interesting. Of course, the clustering and some of the hypothesis and data was simplified (i.e. location within an area, proximity to subway, shops, landmarks, etc.), so there are plenty of possible improvements.