A Guide for Turkish People in New York

By Afsar Onat Aydinhan

12 January, 2021

1. Introduction/Business Problem

New York is a very popular destination for tourists and Turkish people are no exception. There are so many places to visit and so many things to do in New York. However, there is a downside of having so many choices and that is not being able to decide where to actually go, book a hotel/house! The main aim of this project is to analyze the city of New York and determine locations that might be of interest for Turkish tourists, so they can make an informed decision about where to actually stay and/or go in New York. The stakeholders would be interested in this project, because it might be too hard or time-consuming to do such an analysis by themselves and the results of this analysis can actually improve their vacation in New York. As a Turk myself, I am also motivated and interested about this project and hope to use it to recommend places in New York to my friends.

2.Data

We will scrape our data about New York from the following link: 'https://cocl.us/new_york_dataset'

The columns that we are going to use from the data that we scrape are as follows:

1. Borough
2. Neighborhood
3. Latitude
4. Longitude

So basically, we get a list of boroughs, the neighborhood that they belong to and their geographical coordinates. We will build different maps to visualize our results, therefore the geographical data will be crucial to our analysis.

We also need information about venues in the neighborhoods in order to be able to determine which places may attract Turkish people. We will use the Foursquare API to gather information about the venues.

The columns of data that we are interested in from the Foursquare API is as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Venue Category

Once we merge the datasets described above, we will end up with venues in different neighborhoods in New York city and that will be the main core data that we will use to analyze New York City to find places that can attract Turkish people.

3.Methodology

We first start by importing the python libraries that we are going to use as follows:

```
import pandas as pd
import numpy as np
import requests
import matplotlib.cm as cm
import matplotlib.colors as colors
import folium
from geopy.geocoders import Nominatim

# import k-means for the clustering stage
from sklearn.cluster import KMeans
```

Note that in the original notebook, I have imported the libraries when they are required, not all at the beginning, but they are placed here to make the report tidier.

Data Collection and Feature Selection

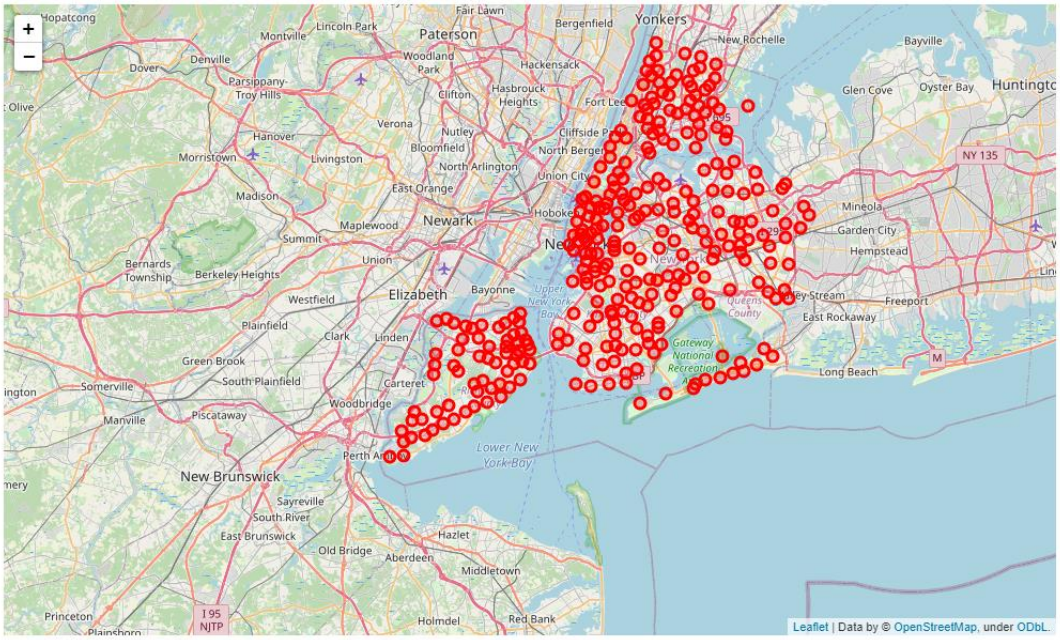
In the data collection part, we begin by collecting the required data for New York. We need data about boroughs, neighborhoods and geographical locations. We collect our data from the following link:'https://cocl.us/new_york_dataset'. The data we get includes extra information that is of no use to us. Therefore, we do feature selection and choose the columns that we need for the analysis.

```
data =requests.get('https://cocl.us/new_york_dataset').json()
data = data['features']
columns = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']

ny_data = pd.DataFrame(columns=columns)

for i in data:
    borough = i['properties']['borough']
    neighborhood = i['properties']['name']
    coordinates = i['geometry']['coordinates']
    longitude = coordinates[0]
    latitude = coordinates[1]

    ny_data = ny_data.append({'Borough': borough,'Neighborhood': neighborhood,'Latitude': latitude,'Longitude': longitude}, ignor
ny_data.head()
```

	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

Data Visualization

Now that we have our initial dataset ready, it is important to use data visualization tools to better understand what is going on with our data. We use the folium library to visualize the location of our neighborhoods. We notice that the dots are denser around the Manhattan region and sparser outside of Manhattan. Manhattan is a very populated area, therefore this was an expected result.



After visualizing the neighbourhoods, we need to determine which of these neighborhoods contain which type of venues. This is where we use our Foursquare API data. The type of data we gather is as follows:

```
print(ny_venues.shape)
ny_venues.head()
```

```
(6120, 5)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Walgreens	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	Donut Shop

One-hot Encoding

We will use one-hot encoding to our venue categories to be able to implement a clustering algorithm. The results we get are as follows:

```
ny_onehot = pd.get_dummies(ny_venues[['Venue Category']], prefix="", prefix_sep="")

ny_onehot['Neighborhood'] = ny_venues['Neighborhood']

fixed_columns = [ny_onehot.columns[-1]] + list(ny_onehot.columns[:-1])
ny_onehot = ny_onehot[fixed_columns]

ny_onehot.head()
```

	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	...	Vietnamese Restaurant	Warehouse Store	Waste Facility	Waterfront
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Two Approaches

We will try to conduct our analysis in two different ways. In the first one, we will use all the avenues and do a clustering to see which neighborhoods are similar to each other. This is not specific to Turkish people, but to a general audience. The point of this analysis is to help Turkish people to determine other places that they may enjoy given they liked a place that they already visited. This portion of the report is to be used after the tourists start enjoying the New York and develop their own tastes regarding New York. The second part will be specifically based on Turkish tourists and hopefully they can use that analysis to determine places that they can like before coming to New York.

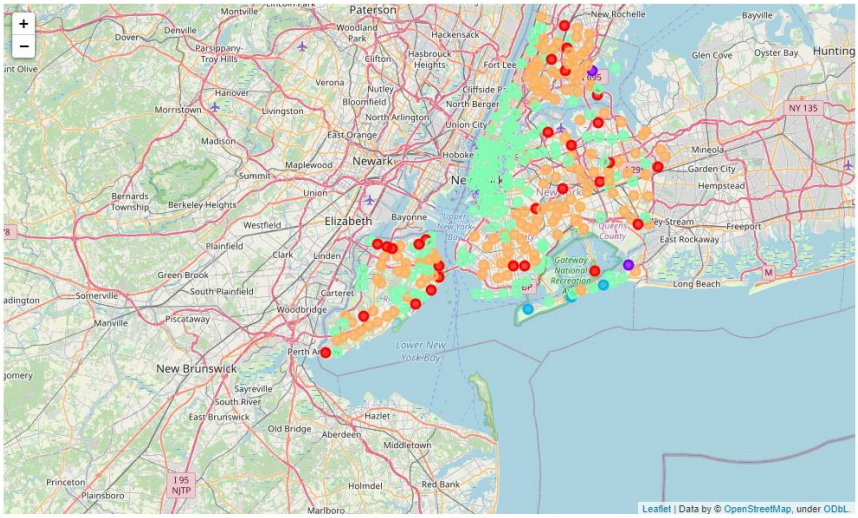
Approach 1 (General):

We first start by finding the top 10 most common venues in our neighborhood. The results that we get are as follows:

	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	Allerton	Pizza Place	Supermarket	Deli / Bodega	Chinese Restaurant	Breakfast Spot	Used Auto Dealership	Spa	Fried Chicken Joint	Fast Food Restaurant	Bike Trail
1	Annadale	Pizza Place	American Restaurant	Bakery	Park	Train Station	Pharmacy	Liquor Store	Restaurant	Diner	Field
2	Arden Heights	Coffee Shop	Deli / Bodega	Lawyer	Bus Stop	Pizza Place	Business Service	Pharmacy	Women's Store	Ethiopian Restaurant	Event Service
3	Arlington	Intersection	Grocery Store	American Restaurant	Deli / Bodega	Bus Stop	Women's Store	Fast Food Restaurant	Ethiopian Restaurant	Event Service	Event Space
4	Arrochar	Bus Stop	Pizza Place	Deli / Bodega	Italian Restaurant	Bagel Shop	Sandwich Place	Athletics & Sports	Middle Eastern Restaurant	Pharmacy	Supermarket

K-mean Clustering

We will use K-means clustering algorithm as our machine learning algorithm. We are going to cluster similar neighborhoods based on their venues. The idea behind this is that a person who enjoys a neighborhood is likely to enjoy another neighborhood that has similar type of venues. Once we conduct our method, we visualize our clusters using the folium library. The result is as follows:



Approach 2(Turkish Tourists):

One drawback of the clustering algorithm above is that we haven't put anything in the algorithm that signifies that the stakeholders are tourists, or Turkish. The following venues would be of interest for Turkish tourists:

'Turkish Restaurant'

'Middle Eastern Restaurant'

'Rental Car Location',

'Tourist Information Center'

'Theme Park'

'Shopping Mall'

'Nightclub'

'Metro Station'

'Bar'

The first two venues are directly related to Turkish people. Turkish people care about their food a lot, and they would certainly enjoy their own cuisine, or middle eastern cuisine in general. The other venues could be used for any other tourists as well, not specifically for Turkish tourists, but they are definitely important for Turkish Tourists as well. So, we proceed to isolate our data to see how the neighborhoods perform in these specific venues. The results for

Let's look at each venue individually to get a better understanding of our data

```
ny_turkish = ny_grouped[['Neighborhood','Turkish Restaurant']]
ny_turkish = ny_turkish[ny_turkish['Turkish Restaurant'] != 0]
ny_turkish
```

	Neighborhood	Turkish Restaurant
9	Bath Beach	0.033333
245	Sheepshead Bay	0.120000
264	Sunnyside Gardens	0.033333
274	Turtle Bay	0.033333
278	Upper West Side	0.033333

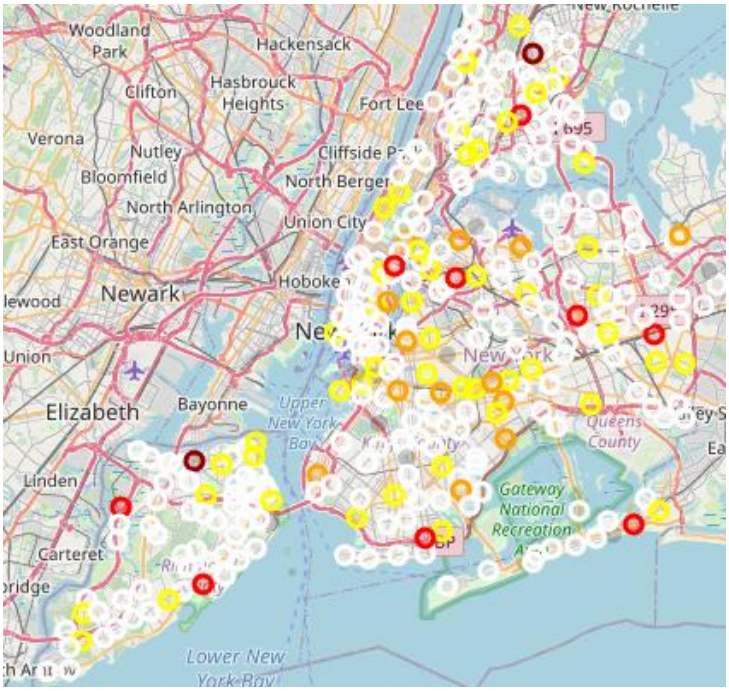
Now we are going to calculate the score of each neighborhood by summing up their normalized venue scores. This is going to give every category in our list an equal weight. It is hard to evaluate the relative importance of each category in the list, so we will stick to equal weight. Depending on the individuals, the weights in our model can be modified. The final score data table we get is as follows:

Sorting based on our score criterion

```
score_ny = ny_interest[["Neighborhood","Score"]]
score_ny.sort_values(by="Score", ascending=False)
```

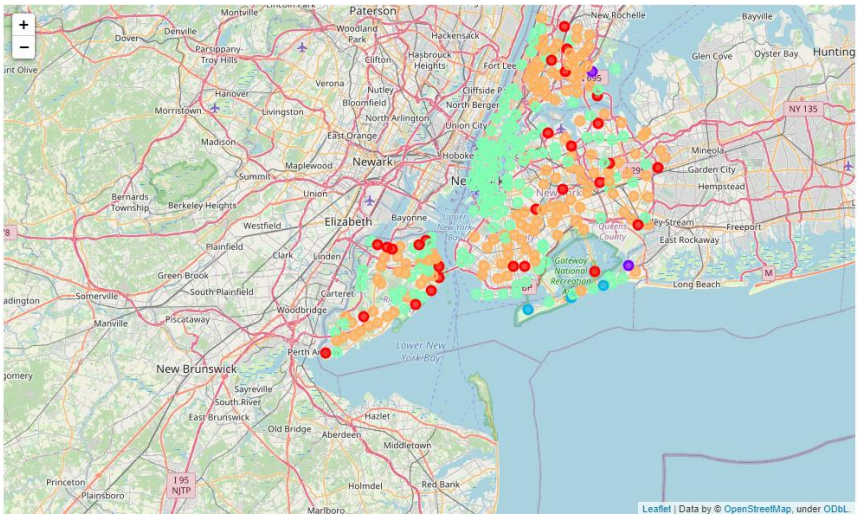
	Neighborhood	Score
291	Williamsbridge	1.800000
215	Port Richmond	1.666667
274	Turtle Bay	1.277778
264	Sunnyside Gardens	1.244444
245	Sheepshead Bay	1.000000
...	...	...
136	Hunters Point	0.000000
137	Hunts Point	0.000000
139	Jackson Heights	0.000000
140	Jamaica Center	0.000000
300	Yorkville	0.000000

The final thing left to do is to visualize our scores. Drawing a new folium map based on our score will be the way to go. The redder the markers, the more attractive the locations are for Turkish tourists.



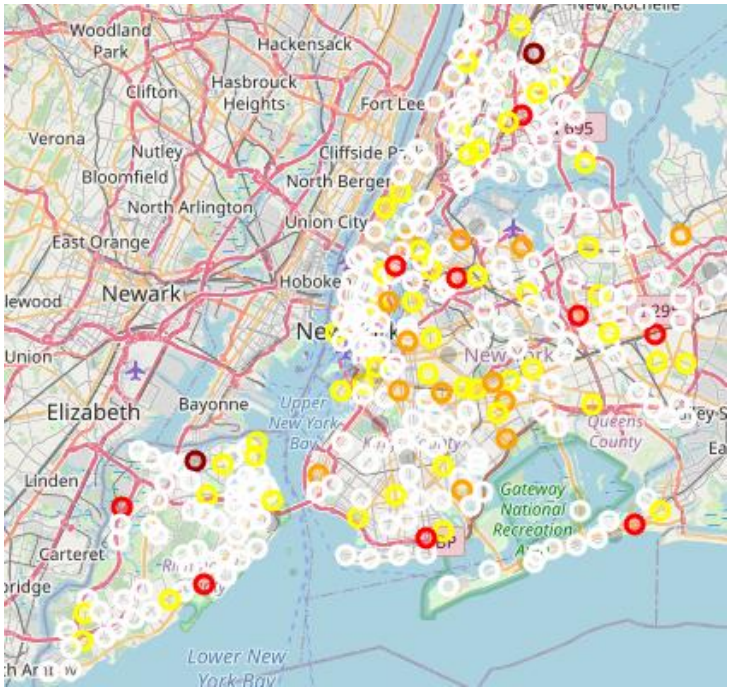
4. Results

Let’s look at our clustering results that we get from our general approach first:



We clearly see that the Manhattan area constitutes a cluster combined with other coastal regions of other parts of New York. This is an expected result since Manhattan is a relatively homogenous region filled with all sorts of entertainment possibilities. Staten Island, Brooklyn and Queens regions seem to be similar to each other in general, mostly orange dots with red dots here and there. Considering how unique Manhattan is in general, it is logical to see that the other regions were not uniquely described with their own cluster, but the clustering was done based on the activities that you can do in those local regions.

Now, we can proceed to the result that we got from the analysis done for the Turkish Tourists. We see that Manhattan and Staten Island is mostly white (not of interest) while the Brooklyn-Queens border is filled with multiple yellow and orange dots surrounded by occasional red dots.



5. Discussion

We see that the two approaches we have used yielded different results. The general approach gave us a general understanding of which neighborhoods are similar. After tourists arrive and visit these places, they can use our analysis to determine locations that they like or dislike depending on their experiences. It seems like Manhattan is a cluster by itself, meaning that if you like Manhattan, you should stay in Manhattan and if you don't, you should visit other parts of New York. Queens, Staten Island and Brooklyn don't seem to be separated by location; they all look similar to each other in terms of our clusters. So you can easily switch between these locations and enjoy venues of your interest. When we look at our second analysis for Turkish tourists, we see that the location that is most suitable for Turkish people according to our criteria, is the Brooklyn-Queens border. The coastal regions and Manhattan in general do not seem to be quite for the Turkish tourists.

6. Conclusion

The main goal of this project was to create a guide for the Turkish people visiting New York. We have used data of neighborhoods and boroughs in New York city and used Foursquare API to find the venues in these locations. Next, we have used the k-means clustering algorithm to create neighborhood clusters that are similar to each other based on the venues that they have. The goal was to determine similar neighborhoods so that a tourist would know where to go next if they enjoy a specific place in New York. Finally, we have done a second analysis for Turkish tourists in general where we used a score-based system to rank neighborhoods, based on venues that we think are important for Turkish people. We have determined that the Brooklyn-Queens border is the optimal location for a Turkish tourist.