Capstone Project Report

Nhat Le 24/03/2020

I. Business Problem

A businessman wants to open a new Asian restaurant in New York City so he needs a piece of advice for the location. He doesn't want to be in a highly competitive location with other restaurants nearby because he just has enough capital to open a small restaurant. By the way, he requires 2 main points:

- No other Asian restaurant within the same neighborhood.

- The restaurant's density is not too crowded (it means there are 5 restaurants maximum within a neighborhood).

My job as a data scientist will help him to pick the best location that is suitable for his needs. So let's begin!

II. Data Understanding and Preparing

I will use Foursquare data to understand more about the restaurant market in New York City, some data I need is:

- location: to pick where a venue is (lat, long)

- categories: category of the venue (coffee shop, restaurant, Asian restaurant, etc.)

As an outcome, it has more than 10,000 venues in New York City!

```python
newyork_venues = getNearbyVenues(names=neighborhoods['Neighborhood'],latitudes=neighborhoods['Latitude'],longitudes=neighborhoods['Longitude'])

# Check the size of newyork_venues
print(newyork_venues.shape)
newyork_venues.head()

(10273, 7)
```

III. Methodology & Results

The methodology I will do:

- I will find a neighborhood that has Asian restaurants in New York City to drop it.

- I will calculate the total restaurant in each neighborhood then remove which one has more than 5 restaurants.

Finally, I will have a shortlist of suggested locations for him to call to check the leasing price to make the decision of which neighborhood he should open an Asian restaurant.

```
In [118]:  # Asian Restaurants list
           newyork_asian_res = newyork_venues[newyork_venues['Venue Category'] == 'Asian Restaurant']
           print('There are', newyork_asian_res.shape[0],'Asian Restaurants in New York city.')
           newyork_asian_res.head()
```

There are 54 Asian Restaurants in New York city.

Out[118]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
637	Throgs Neck	40.81510925804005	-73.81635002158441	Green Dragon	40.81887820776003	-73.81679306899113	Asian Restaurant
653	Parkchester	40.837937822267286	-73.85600310535783	Neerob	40.83611083931938	-73.85490233193293	Asian Restaurant
676	Parkchester	40.837937822267286	-73.85600310535783	Sing Hing	40.839496652600666	-73.85841341148014	Asian Restaurant
698	Westchester Square	40.8406194964327	-73.84219407604444	Kai Asian Fusion	40.843590837833744	-73.84558937077266	Asian Restaurant
894	Pelham Bay	40.85064140940335	-73.8320737824047	Shangri-La	40.84715972410305	-73.83264181934335	Asian Restaurant

I visualize it in a New York City map:
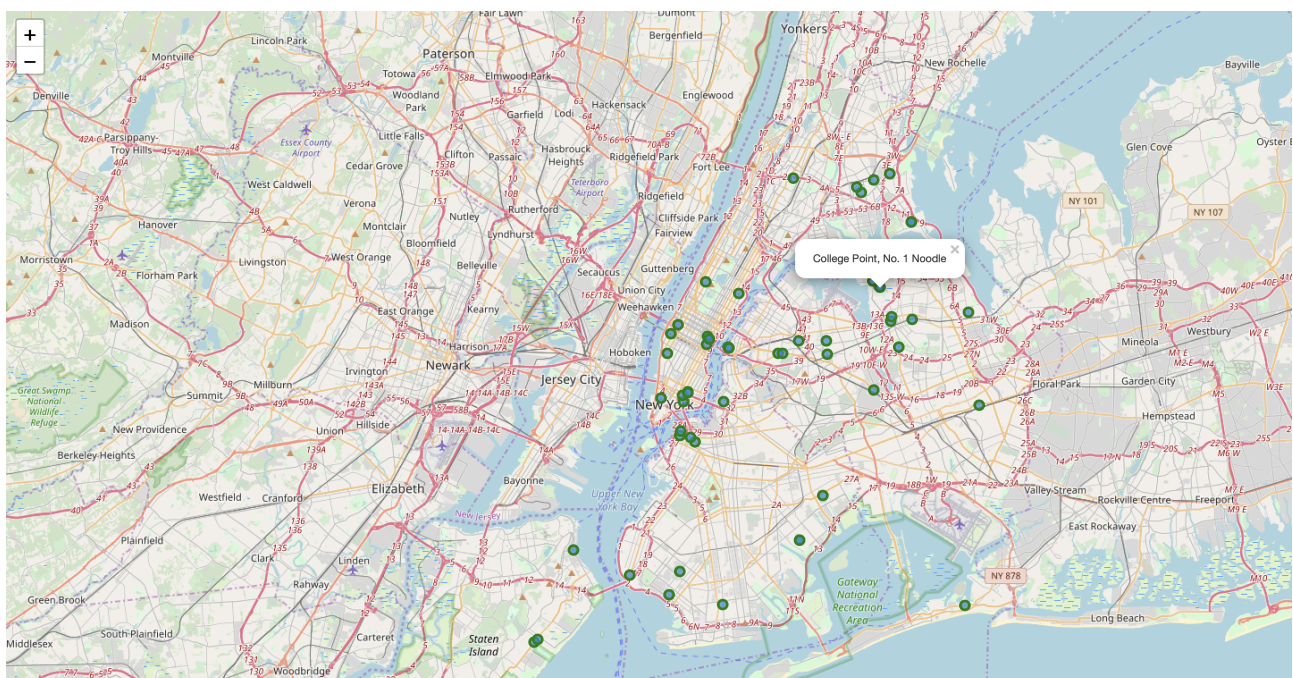
As you can see, the name of the Asian restaurant on the map is №1 Noodle! It sounds Asian right? :) Next, I find and remove all the neighborhood which has more than 5 restaurants by filtering then counting the restaurants in each neighborhood. It's 132. Then I merge 2 tables to check one more time and remove the neighborhood that has an Asian restaurant (there are 7).

Merge 2 dataframe based of restaurant counts max 5 to check if there is a neighborhood has Asian restaurant

```
In [244]: check_result = pd.merge(neigh_restaurant_count, neighborhood_asian, how='inner', on=['Neighborhood'])
          print('There are',check_result.shape[0],'neighborhoods has Asian restaurant')
          check_result

          There are 7 neighborhoods has Asian restaurant

Out[244]:
```

	Neighborhood	Counts	Check
0	New Lots	4	1
1	Mount Hope	3	1
2	Throgs Neck	3	1
3	Canarsie	3	1
4	Hollis	3	1
5	Edgemere	2	1
6	Paerdegat Basin	1	1

Finally, we remove 7 neighborhoods to have the final sort list

I draw a histogram chart to count the number of neighborhood location that has 1 restaurant, 2 restaurants, etc. so that my client will have a better final decision and have a priority to check which tier to go first. I suggest that a neighborhood that has only 1 restaurant is good to try.

Histogram of the number of restaurants in each neighborhood
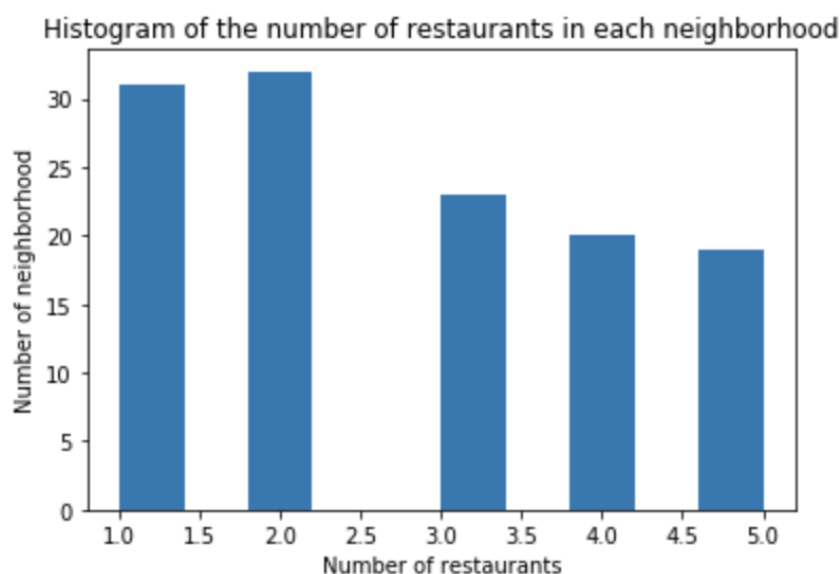
IV. Discussion

I have learned a lot from this course, especially by doing the last project with the real dataset for a real-world problem/demand. I think I can do better in the future with some ideas. I want to do more for the last project but my skills are not good enough to do it such as:

- Convert lat, long to x, y to calculate the distance to make better decisions.

- Using a Foursquare dataset to check the pricing tier or the popularity of the venue.

- Find the pattern based on the popular Asian restaurant to find another location (neighborhood) that will likely be successful.

V. Conclusion

I hope that my result will be helpful to solve a part of the business problem.
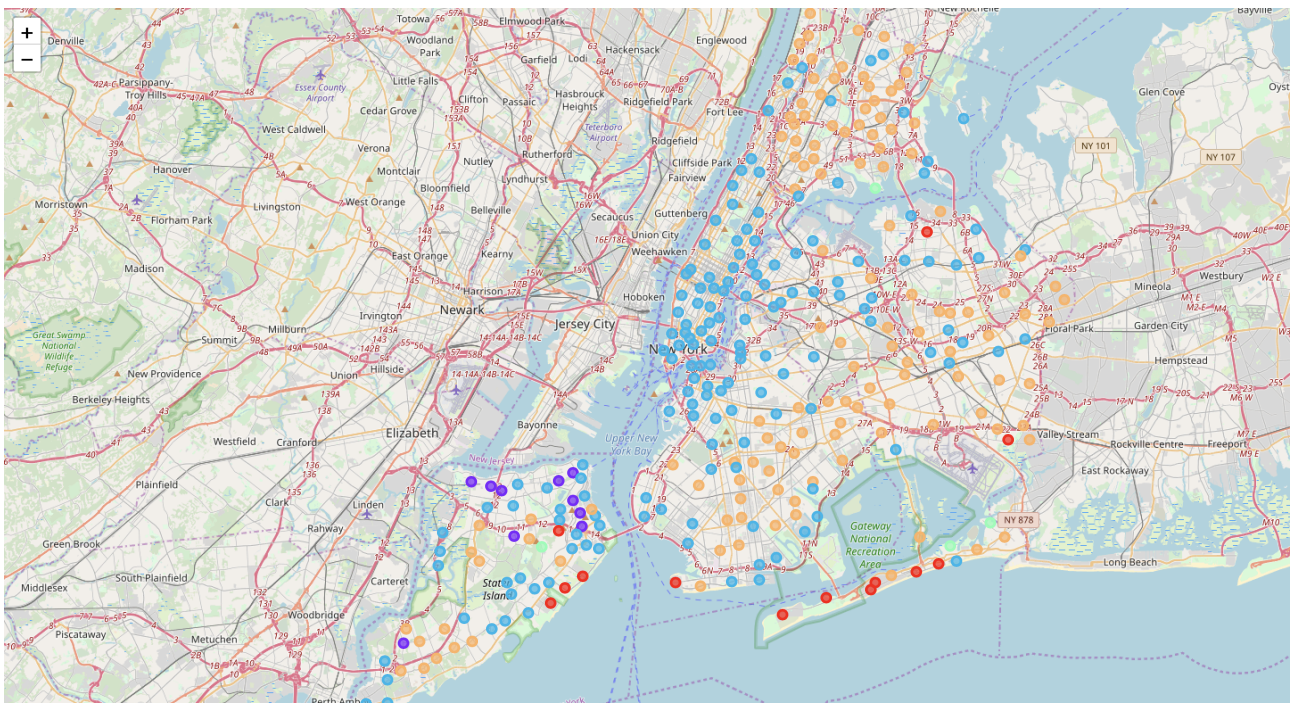
P/s: I use the KNN algorithm to cluster the New York map into 5 clusters based on the top 10 venues of each neighborhood.

Thank you for reading.