APPLIED DATA SCIENCE CAPSTONE BY IBM

IBM Data Science Capstone Project

Picking the right location for a new vegan restaurant in New York

Maria Koptseli

27/3/2021

The IBM Data Science Professional Certificate course on Coursera concludes with a Capstone Project. This project is about using data science toolset on a real life problem and demonstrating the creation of value by applying the learned data analysis skills. This report presents this capstone project. The analysis was performed in Python.

1. Introduction

The IBM Data Science Professional certificate course on Coursera concludes with a Capstone Project. The project is about using data science toolset on a real-life case and demonstrating the creation of value by applying the learned data analysis skills. This report presents the capstone project. The analysis was performed in Python.

2. Problem Definition

a. The Problem

Every successful enterprise is the result of an inspiration, continuous hard work, great determination, and aspirations. One must have a strong desire to take an idea from its generation to implementation. Idea generation is certainly a crucial process for businesses seeking to succeed and gain competitive advantage, but collecting invaluable insights to well define this process in order to implement the generated ideas and put them into action is a necessity. For this project, we chose a hypothetical business scenario about a prospective business owner, who is a passionate supporter of veganism and he has decided to invest in a vegan gourmet restaurant in New York, US. Taking into consideration the price level at which the restaurant will operate, his main intent is to find an optimal location in an area of Manhattan, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well.

b. Assumptions and business logic

The assumption behind the analysis is that we can use unsupervised machine learning to create clusters of districts that will provide us with a list of areas for consideration for the restaurant. The intent is to find an optimal location in an area, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well. The prospective owner will obtain the desired data per every neighborhood regarding: which cuisine based restaurants are the least in number per neighborhood, if the area of his choice is lacking of a vegan gourmet restaurant, which type of restaurants are prevalent in a particular area and in the adjoining neighborhoods and so on. The analysis is going to provide him all the essential data to make an informed decision and boost his chance of success.

c. Audience

In our project we are assuming a prospective business owner to whom we are addressing this report, however this analysis could be useful for a group of market investors such as restaurant owners.

3. Data

To perform this analysis, we need the following data:

1. List of the districts of New York City
2. Geo-coordinates of the districts in New York City
3. Top venues of districts

- List of Boroughs of New York
  The first dataset we will be using would contain all the required geographical data about New York City. We would be using 'Borough', 'Neighborhood', 'Latitude', and ‘Longitude’ among all the other data elements present in the data. For our convenience, we would be using the same data set which was provided to us in Week 3 of this course: newyork_data.json
- Geo-coordinates of areas will be obtained with the help of the geo-coder tool in the notebook.
- Top venues data will be obtained from Foursquare through an API.

4. Methodology

a. Use of data

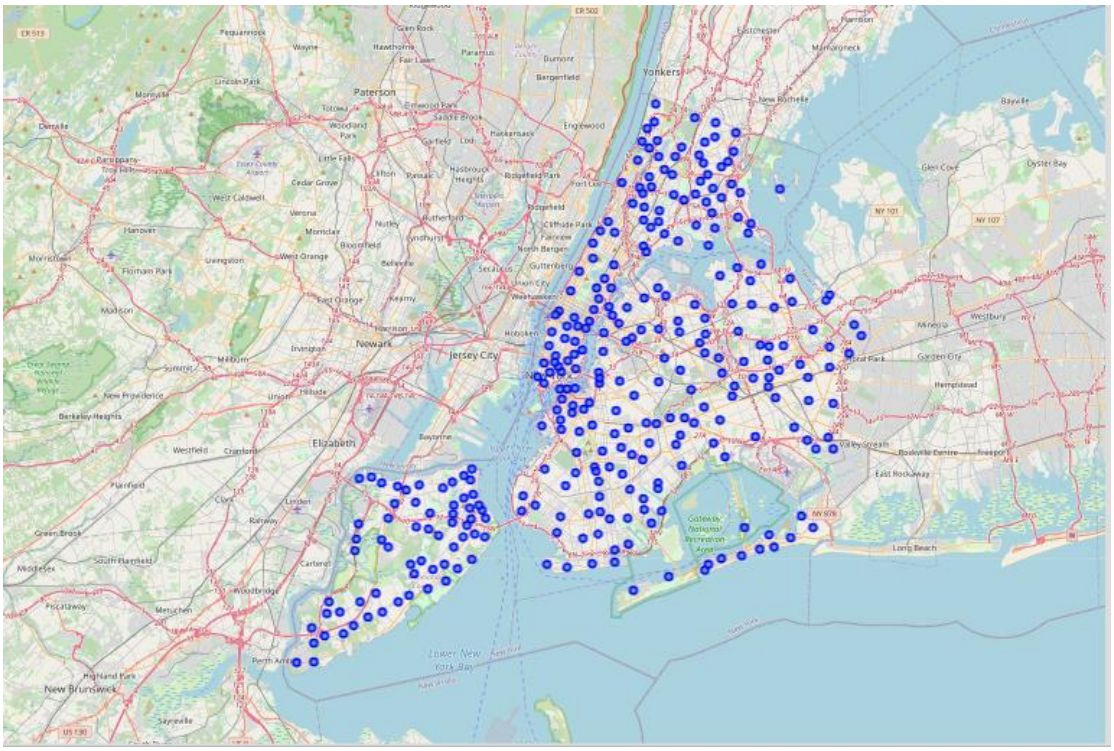
After collecting, framing and exploring the data, we will apply the K-means machine learning technique for creating clusters of neighborhoods of the borough of Manhattan.

b. Analysis

i. Data Preparation and exploration

As a part of data preparation, we start by creating a list of Boroughs & Neighborhoods in New York and add the geo-coordinates of each to this table. That is done by first importing a list of Boroughs and then by using this list and geo-code python library, we add the latitude and longitude coordinates to each. There are 5 boroughs and 306 neighborhoods in New York. In the next step, we create a visual representation of how the 306 neighborhoods are situated in New York. For this, the folium library was used.

Picture 1: visual representation the 306 neighborhoods in New York



As we have already mentioned, our prospective owner is interested in particular in opening a vegan restaurant in the borough of Manhattan, since his main intent is to find an optimal location in an area, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well. In the following table,

we are able to see the results we get by using pandas data frame format regarding the Borough of Manhattan.

Picture 2: Pandas data frame results - Borough of Manhattan

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

In the next step of the analysis, the borough of Manhattan was explored in greater detail. Venues were collected for each neighborhood of Manhattan via Foursquare API. After arranging the data, we have up to 100 venues for each neighborhood. Venues are collected within a radius of 1000 meters from the point of our district of interest coordinates. The following table shows the collected and arranged data for the 10 most common venues from several neighborhoods of the Borough of Manhattan.

Picture 3: Pandas data frame results of most 10 common venues: Neighborhoods in Manhattan

	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	Battery Park City	Park	Coffee Shop	Gym	Clothing Store	Hotel	Memorial Site	Playground	Shopping Mall	BBQ Joint	Boat or Ferry
1	Carnegie Hill	Coffee Shop	Café	Bar	Italian Restaurant	Pizza Place	French Restaurant	Cosmetics Shop	Yoga Studio	Gym	Wine Shop
2	Central Harlem	African Restaurant	Seafood Restaurant	Cosmetics Shop	Bar	French Restaurant	Chinese Restaurant	Gym / Fitness Center	American Restaurant	Park	Cafeteria
3	Chelsea	Coffee Shop	Art Gallery	Bakery	American Restaurant	Italian Restaurant	Wine Shop	Seafood Restaurant	French Restaurant	Ice Cream Shop	Cocktail Bar
4	Chinatown	Chinese Restaurant	Bakery	Dessert Shop	Cocktail Bar	Hotpot Restaurant	Salon / Barbershop	American Restaurant	Optical Shop	Spa	Shanghai Restaurant

ii. Clustering

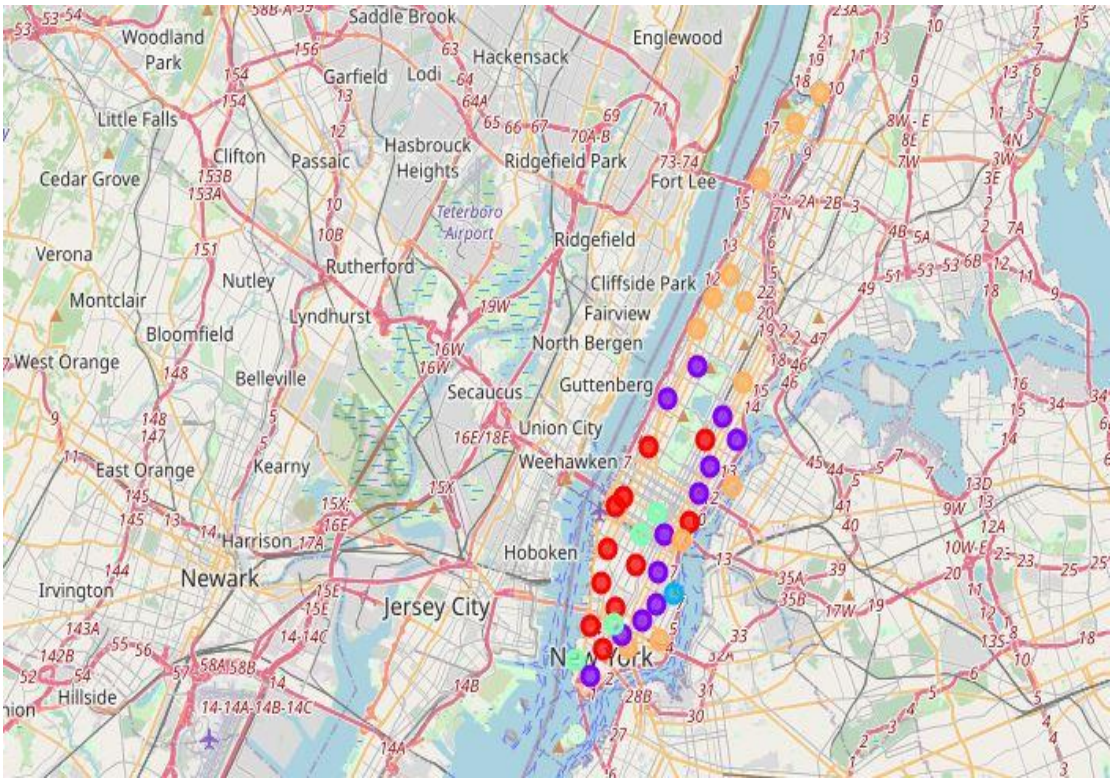
Our dataset is ready, so in the next step we will perform clustering. For this, unsupervised machine learning technique will be used based on K-means. For K-means clustering, we need to decide on the number of clusters that we want to use. We run the K-means clustering algorithm with the parameter of 5 as the number of clusters. When done, we add the cluster labels to the dataset. In the following table we can see an example of the result we get. We can see the cluster as well as the top 10 venues for each neighborhood.

Picture 4: Cluster Labels & 10 most common venues: Neighborhoods in Manhattan

	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 Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	4	Gym	Discount Store	Coffee Shop	Sandwich Place	Yoga Studio	Tennis Stadium	Kids Store	Donut Shop	Diner	Pharmacy
1	Manhattan	Chinatown	40.715618	-73.994279	4	Chinese Restaurant	Bakery	Dessert Shop	Cocktail Bar	Hotpot Restaurant	Salon / Barbershop	American Restaurant	Optical Shop	Spa	Shanghai Restaurant
2	Manhattan	Washington Heights	40.851903	-73.936900	4	Café	Bakery	Mobile Phone Shop	Deli / Bodega	Gym	Latin American Restaurant	Tapas Restaurant	Sandwich Place	Bank	Chinese Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	4	Mexican Restaurant	Café	Restaurant	Lounge	Chinese Restaurant	Pizza Place	Bakery	Park	Wine Bar	Deli / Bodega
4	Manhattan	Hamilton Heights	40.823604	-73.949688	4	Pizza Place	Café	Coffee Shop	Deli / Bodega	Mexican Restaurant	Yoga Studio	Juice Bar	Chinese Restaurant	School	Sandwich Place
5	Manhattan	Manhattanville	40.816934	-73.957385	4	Seafood Restaurant	Coffee Shop	Deli / Bodega	Park	Sushi Restaurant	Italian Restaurant	Mexican Restaurant	Chinese Restaurant	Café	Lounge
6	Manhattan	Central Harlem	40.815976	-73.943211	4	African Restaurant	Seafood Restaurant	Cosmetics Shop	Bar	French Restaurant	Chinese Restaurant	Gym / Fitness Center	American Restaurant	Park	Cafeteria
7	Manhattan	East Harlem	40.792249	-73.944182	4	Mexican Restaurant	Bakery	Latin American Restaurant	Thai Restaurant	Sandwich Place	Steakhouse	Spa	Deli / Bodega	French Restaurant	Dance Studio
8	Manhattan	Upper East Side	40.775639	-73.960508	0	Coffee Shop	Italian Restaurant	Bakery	Gym / Fitness Center	Exhibit	Yoga Studio	Hotel	Pizza Place	American Restaurant	French Restaurant
9	Manhattan	Yorkville	40.775930	-73.947118	1	Italian Restaurant	Bar	Gym	Coffee Shop	Deli / Bodega	Sushi Restaurant	Wine Shop	Japanese Restaurant	Diner	Pub
10	Manhattan	Lenox Hill	40.768113	-73.958860	1	Italian Restaurant	Pizza Place	Sushi Restaurant	Cocktail Bar	Coffee Shop	Burger Joint	Gym / Fitness Center	Café	Gym	Cycle Studio
11	Manhattan	Roosevelt Island	40.762160	-73.949168	4	Park	Restaurant	Residential Building (Apartment / Condo)	Dry Cleaner	Metro Station	Coffee Shop	Soccer Field	Baseball Field	Greek Restaurant	Dog Run
12	Manhattan	Upper West Side	40.787658	-73.977059	1	Italian Restaurant	Mediterranean Restaurant	Bakery	Bar	Wine Bar	Café	Pizza Place	Coffee Shop	Pub	Indian Restaurant

Also, we can visualize the clusters on the map that we created earlier.

Picture 5: Map: Clusters & 10 common venues: Neighborhoods in Manhattan



5. Results

Understanding the Clusters

By looking at the cluster data, we can see that clusters 1 and 2 are the ones that we are the most interested in. These clusters meet the main criteria, which is the intent of our prospective owner to find an optimal location in an area, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well.

1. Cluster 1

In the first cluster we find neighborhoods, where in the first most common venue we can see mainly coffee shops, but in the following next ranks we can see several restaurants and also hotels.

Picture 6: Cluster 1 - results

	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
8	Upper East Side	Coffee Shop	Italian Restaurant	Bakery	Gym / Fitness Center	Exhibit	Yoga Studio	Hotel	Pizza Place	American Restaurant	French Restaurant
13	Lincoln Square	Plaza	Café	Concert Hall	Performing Arts Venue	Theater	American Restaurant	French Restaurant	Bakery	Wine Shop	Indie Movie Theater
14	Clinton	Italian Restaurant	Gym / Fitness Center	Theater	American Restaurant	Coffee Shop	Sandwich Place	Spa	Cocktail Bar	Hotel	Wine Shop
17	Chelsea	Coffee Shop	Art Gallery	Bakery	American Restaurant	Italian Restaurant	Wine Shop	Seafood Restaurant	French Restaurant	Ice Cream Shop	Cocktail Bar
18	Greenwich Village	Italian Restaurant	Clothing Store	Sushi Restaurant	Gym	Coffee Shop	Indian Restaurant	Café	Dessert Shop	Vietnamese Restaurant	French Restaurant
21	Tribeca	Park	American Restaurant	Italian Restaurant	Spa	Wine Bar	Café	Scenic Lookout	French Restaurant	Basketball Court	Greek Restaurant
24	West Village	Italian Restaurant	American Restaurant	New American Restaurant	Cosmetics Shop	Cocktail Bar	Park	Wine Bar	Coffee Shop	French Restaurant	Chinese Restaurant
32	Civic Center	Coffee Shop	Cocktail Bar	Gym / Fitness Center	American Restaurant	French Restaurant	Hotel	Spa	Yoga Studio	Park	Hotel Bar
35	Turtle Bay	Coffee Shop	Italian Restaurant	Sushi Restaurant	Japanese Restaurant	Ramen Restaurant	Café	Hotel	Seafood Restaurant	Park	Deli / Bodega
38	Flatiron	Gym / Fitness Center	Italian Restaurant	American Restaurant	New American Restaurant	Japanese Restaurant	Mediterranean Restaurant	Spa	Coffee Shop	Sporting Goods Shop	Women's Store
39	Hudson Yards	American Restaurant	Gym / Fitness Center	Café	Hotel	Italian Restaurant	Nightclub	Spanish Restaurant	Bar	Thai Restaurant	Gym

2. Cluster 2

In Cluster 2 is we see lots of gastronomy related venues such as coffee shop, pizza place, Italian restaurant, bar, Mexican restaurant, Thai restaurant etc) and we can also several hotels as well.

Picture 7: Cluster 2 - results

	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
9	Yorkville	Italian Restaurant	Bar	Gym	Coffee Shop	Deli / Bodega	Sushi Restaurant	Wine Shop	Japanese Restaurant	Diner	Pub
10	Lenox Hill	Italian Restaurant	Pizza Place	Sushi Restaurant	Cocktail Bar	Coffee Shop	Burger Joint	Gym / Fitness Center	Café	Gym	Cycle Studio
12	Upper West Side	Italian Restaurant	Mediterranean Restaurant	Bakery	Bar	Wine Bar	Café	Pizza Place	Coffee Shop	Pub	Indian Restaurant
16	Murray Hill	Hotel	Coffee Shop	Japanese Restaurant	Sandwich Place	American Restaurant	Bar	Taco Place	Burger Joint	Gym / Fitness Center	Gym
19	East Village	Bar	Mexican Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Speakeasy	Korean Restaurant	Coffee Shop	Ice Cream Shop	Cocktail Bar	Italian Restaurant
22	Little Italy	Bakery	Café	Italian Restaurant	Bubble Tea Shop	Hotel	Pizza Place	Sandwich Place	Mediterranean Restaurant	Ice Cream Shop	Salon / Barbershop
25	Manhattan Valley	Mexican Restaurant	Bar	Thai Restaurant	Pizza Place	Coffee Shop	Yoga Studio	Indian Restaurant	Cuban Restaurant	Park	Cosmetics Shop
27	Gramercy	Bar	Bagel Shop	Italian Restaurant	American Restaurant	Pizza Place	Ice Cream Shop	Mexican Restaurant	Cocktail Bar	Coffee Shop	Thai Restaurant
29	Financial District	Coffee Shop	American Restaurant	Pizza Place	Gym	Cocktail Bar	Bar	Italian Restaurant	Gym / Fitness Center	Café	Salad Place
30	Carnegie Hill	Coffee Shop	Café	Bar	Italian Restaurant	Pizza Place	French Restaurant	Cosmetics Shop	Yoga Studio	Gym	Wine Shop
31	Noho	Italian Restaurant	Coffee Shop	Mexican Restaurant	French Restaurant	Pizza Place	Grocery Store	Sandwich Place	Art Gallery	Bakery	Japanese Restaurant

3. Cluster 3

Cluster 3 contains only one neighborhood. Here we can find the bar category at the top, but behind that, is a park, and next a bakery shop. This is a cluster we are less interested in.

Picture 8: Cluster 3 - results

	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
37	Stuyvesant Town	Bar	Park	Coffee Shop	Boat or Ferry	Pet Service	Baseball Field	Fountain	Harbor / Marina	Cocktail Bar	Heliport

4. Cluster 4

Cluster 4 contains only four neighborhoods. We can mainly see the hotel category, park, clothing store at the top, but behind that, it is about art galleries, clothing stores, and shopping malls. This is also a cluster we are not interested in.

Picture 9: Cluster 4 - results

	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
15	Midtown	Hotel	Coffee Shop	Clothing Store	Theater	American Restaurant	Sporting Goods Shop	Bookstore	Sandwich Place	Café	Bakery
23	Soho	Clothing Store	Italian Restaurant	Boutique	Mediterranean Restaurant	Coffee Shop	Salon / Barbershop	Bakery	Art Gallery	Dessert Shop	Falafel Restaurant
28	Battery Park City	Park	Coffee Shop	Gym	Clothing Store	Hotel	Memorial Site	Playground	Shopping Mall	BBQ Joint	Boat or Ferry
33	Midtown South	Korean Restaurant	Hotel	Hotel Bar	Japanese Restaurant	Cosmetics Shop	Coffee Shop	Café	Salad Place	American Restaurant	Dessert Shop

5. Cluster 5

This cluster contains many neighborhoods but in the first place we can see gym category, park, and restaurants with ethnic menus mostly, such as Chinese, Mexican and African one. We also believe this is a cluster our prospective owner should not focus on.

Picture 10: Cluster 5 - results

	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	Marble Hill	Gym	Discount Store	Coffee Shop	Sandwich Place	Yoga Studio	Tennis Stadium	Kids Store	Donut Shop	Diner	Pharmacy
1	Chinatown	Chinese Restaurant	Bakery	Dessert Shop	Cocktail Bar	Hotpot Restaurant	Salon / Barbershop	American Restaurant	Optical Shop	Spa	Shanghai Restaurant
2	Washington Heights	Café	Bakery	Mobile Phone Shop	Deli / Bodega	Gym	Latin American Restaurant	Tapas Restaurant	Sandwich Place	Bank	Chinese Restaurant
3	Inwood	Mexican Restaurant	Café	Restaurant	Lounge	Chinese Restaurant	Pizza Place	Bakery	Park	Wine Bar	Deli / Bodega
4	Hamilton Heights	Pizza Place	Café	Coffee Shop	Deli / Bodega	Mexican Restaurant	Yoga Studio	Juice Bar	Chinese Restaurant	School	Sandwich Place
5	Manhattanville	Seafood Restaurant	Coffee Shop	Deli / Bodega	Park	Sushi Restaurant	Italian Restaurant	Mexican Restaurant	Chinese Restaurant	Café	Lounge
6	Central Harlem	African Restaurant	Seafood Restaurant	Cosmetics Shop	Bar	French Restaurant	Chinese Restaurant	Gym / Fitness Center	American Restaurant	Park	Cafeteria
7	East Harlem	Mexican Restaurant	Bakery	Latin American Restaurant	Thai Restaurant	Sandwich Place	Steakhouse	Spa	Deli / Bodega	French Restaurant	Dance Studio
11	Roosevelt Island	Park	Restaurant	Residential Building (Apartment / Condo)	Dry Cleaner	Metro Station	Coffee Shop	Soccer Field	Baseball Field	Greek Restaurant	Dog Run
20	Lower East Side	Chinese Restaurant	Bakery	Café	Ramen Restaurant	Art Gallery	Coffee Shop	Latin American Restaurant	Yoga Studio	Sandwich Place	Bubble Tea Shop
26	Morningside Heights	Coffee Shop	Bookstore	Park	American Restaurant	Burger Joint	Café	Seafood Restaurant	Supermarket	Mediterranean Restaurant	Mexican Restaurant
36	Tudor City	Park	Mexican Restaurant	Café	Pizza Place	Coffee Shop	Deli / Bodega	Diner	Garden	Sushi Restaurant	Seafood Restaurant

6. Discussion and Recommendations

With this project, we have covered the main steps of the data analysis process. We started with collecting the data, then proceeding to Exploratory Data Analysis before advancing to application of Machine Learning algorithms (k-means clustering) in our case. While it is a certainly a significant step in the right direction, there are a few lessons learnt which can now be discussed as recommendations and are as follows: based on what we learned about the clusters, we can advise the prospective restaurant owner to consider the

Neighborhoods from cluster 1 or 2 as a potential location for the new restaurant. These are the neighborhoods where gastronomy is well represented and hotels are frequent. These satisfy the two original criteria that the location should be in a gastronomical center and in a location that is easily accessible for tourists and for wealthier local citizens as well.

7. Conclusion

This paper discussed the process of coming up with an answer for a hypothetical though real-life like business problem. The analysis was performed based on the toolset of data science and relied heavily on the use of Python and Python libraries such as Pandas, Folium to mention a few. For analysis, the machine learning technique was used. Given the extensive scope of the subject which seems limitless, we have merely made a dent and there are far bigger horizons for us to conquer. A prospective business owner can use this project to see which neighborhoods are lacking of a vegan restaurant or what type of restaurants are doing well and in which neighborhoods. He can then make an informed decision and to have a better chance of establishing a successful business. The output of the analysis provided a thorough base for the recommendation for the business problem in question.

8. References

The Jupyter notebook of the analysis can be found on GitHub.