Setting up a new restaurant in New York City - Report

The Problem

The Business Problem that I wish to help solve is basically in which neighborhood of New York city should a cook/business man build his new restaurant, given the fact that Manhattan is already a place filled with many many restaurants of several cultures. Thus, the idea of a new establishment in this city would already be huge challenge to undertake and even more so for the business to thrive.

According to an article made by Nick Hines from the Vinepair website (https://vinepair.com/booze-news/new-york-restaurants-eat-at-every-on/) "you can't walk a New City block without passing a restaurant". It even states that "80 percent of restaurants fail within five years", so it would seem very difficult to get a new restaurant business going in this city.

A study of venues in other metropolitan areas around the world will help inform which types of restaurant are less common in New York city, thus improving the possibility of success in a city already filled with so many restaurants.

The Data

Foursquare API will be the chosen API to collect the data related to the venues for each geographical point.

To gather the information about geographical location (postal code, neighborhood, borough), ***Open Data Soft*** API (https://data.opendatasoft.com/pages/home/) was used, which is very simple to use, by simply writing down the country and city that you wish to research.

The chosen cities were New York City (which was already researched upon in the Laboratory for the Capstone section: https://cocl.us/new_york_dataset), Toronto (also researched as a deliverable in this final course: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'), Paris, Berlin and Porto.

These last three were the ones that were researched through Open Data Soft. Paris and Berlin were chosen, since they definitely are known as multicultural cities with a significant diversity of venues and a considerable number of people living in them. Porto is sort of an outlier, but it's an emergent city of Portugal, with lots of tourists coming every year and with evolving throughout time with new businesses being implemented.

The information for Toronto and New York is already in the "correct" format but in the case of the other cities the data is categorized in another format, for example, for Paris:

country code	postal code	place name	admin name1	admin code1	admin name2	admin code2	admin name3	admin code3	latitude	longitude	accuracy	coordinates
FR	38170	Seyssinet-Pari	Auvergne-RhÃ´ne-Alpes	84	IsÃ¨re	38	Arrondissemer	381	45.1768	5.6939	5	45.1768,5.6939
FR	75005	Paris 05	ÃŽle-de-France	11	Paris	75	Paris	751	48.8448	2.3471	5	48.8448,2.3471
FR	75053 CEDEX 01	Paris 01	ÃŽle-de-France	11	Paris	75	Paris	751	48.8592	2.3417	5	48.8592,2.3417
FR	75068 CEDEX 02	Paris 02	ÃŽle-de-France	11	Paris	75	Paris	751	48.8655	2.3426	5	48.8655,2.3426
FR	75078 CEDEX 02	Paris 02	ÃŽle-de-France	11	Paris	75	Paris	751	48.8655	2.3426	5	48.8655,2.3426
FR	75096 CEDEX 02	Paris 02	ÃŽle-de-France	11	Paris	75	Paris	751	48.8655	2.3426	5	48.8655,2.3426
FR	75109 CEDEX 02	Paris 02	ÃŽle-de-France	11	Paris	75	Paris	751	48.8655	2.3426	5	48.8655,2.3426
FR	75111 CEDEX 02	Paris 02	ÃŽle-de-France	11	Paris	75	Paris	751	48.8655	2.3426	5	48.8655,2.3426
FR	75122 CEDEX 03	Paris 03	ÃŽle-de-France	11	Paris	75	Paris	751	48.8637	2.3615	5	48.8637,2.3615
FR	75141 CEDEX 03	Paris 03	ÃŽle-de-France	11	Paris	75	Paris	751	48.8637	2.3615	5	48.8637,2.3615
FR	75144 CEDEX 19	Paris 19	ÃŽle-de-France	11	Paris	75	Paris	751	48.8817	2.3822	5	48.8817,2.3822
FR	75164 CEDEX 19	Paris 19	ÃŽle-de-France	11	Paris	75	Paris	751	48.8817	2.3822	5	48.8817,2.3822
FR	75191 CEDEX 04	Paris 04	ÃŽle-de-France	11	Paris	75	Paris	751	48.8601	2.3507	5	48.8601,2.3507
FR	75208 CEDEX 16	Paris 16	ÃŽle-de-France	11	Paris	75	Paris	751	48.8637	2.2769	5	48.8637,2.2769
FR	75217 CEDEX 16	Paris 16	ÃŽle-de-France	11	Paris	75	Paris	751	48.8637	2.2769	5	48.8637,2.2769
FR	75230 CEDEX 05	Paris 05	ÃŽle-de-France	11	Paris	75	Paris	751	48.8448	2.3471	5	48.8448,2.3471
FR	75239 CEDEX 05	Paris 05	ÃŽle-de-France	11	Paris	75	Paris	751	48.8448	2.3471	5	48.8448,2.3471

For these cases the Neighborhood and Borough were considered to be the same thing, which is the column “Place Name”, and of course the columns with the postal codes and the latitude and longitude geographical coordinates were kept. All the other columns were not used.

For the case of the Berlin data, the Postal Code was used as the Neighborhood name:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
country code	postal code	place name	admin name1	admin coc	admin nar	admin coc	admin nar	admin coc	latitude	longitude	accuracy	coordinates,,	
DE	10179	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5122	13.4164	6	52.5122,13.4164	
DE	10249	Berlin Friedrich	Berlin	BE		0	Berlin, Sta	11000	52.5238	13.4428	6	52.5238,13.4428	
DE	10627	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.508	13.303	6	52.508,13.303	
DE	10715	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4824	13.3289	6	52.4824,13.3289	
DE	10717	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4908	13.3275	6	52.4908,13.3275	
DE	10777	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4975	13.3427	6	52.4975,13.3427	
DE	10779	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4921	13.3395	6	52.4921,13.3395	
DE	12103	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4641	13.3747	6	52.4641,13.3747	
DE	12107	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4312	13.3917	6	52.4312,13.3917	
DE	12459	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4656	13.528	6	52.4656,13.528	
DE	12527	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.3856	13.6339	6	52.3856,13.6339	
DE	12559	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4149	13.6634	6	52.4149,13.6634	
DE	13051	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5815	13.4908	6	52.5815,13.4908	
DE	13129	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5921	13.4579	6	52.5921,13.4579	
DE	13189	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5643	13.4219	6	52.5643,13.4219	
DE	13595	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5116	13.1962	6	52.5116,13.1962	
DE	14053	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5159	13.2387	6	52.5159,13.2387	
DE	14165	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.4175	13.2536	6	52.4175,13.2536	
DE	10369	Berlin	Berlin	BE		0	Berlin, Sta	11000	52.5295	13.4695	6	52.5295,13.4695	

Notice how almost all of the entries have “Berlin” as the place name, so to be able to get the venues for each location, the postal code, which is unique, will be used to distinguish each place from the other.

In the case of Porto’s data, the case is a bit more complicated:

country co	postal coc	place nam	admin nar	admin coc	admin nar	admin coc	admin nar	admin coc	latitude	longitude	accuracy	coordinates
PT	4600-004	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-009	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-026	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-034	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-066	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-108	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-116	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-118	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-137	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-216	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-219	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-232	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-281	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-296	Amarante	Porto	17	Amarante	1301	Amarante		41.2727	-8.0825	4	41.2727,-8.0825
PT	4600-540	Canadelo	Porto	17	Amarante	1301	Canadelo		41.3242	-7.9717	4	41.3242,-7.9717
PT	4600-558	Candemil	Porto	17	Amarante	1301	Candemil Amt		41.2475	-7.9671	4	41.2475,-7.9671
PT	4600-570	PÃ© Redc	Porto	17	Amarante	1301	Carvalho De Rei		41.1742	-7.966	3	41.1742,-7.966
PT	4600-611	Pensais	Porto	17	Amarante	1301	Freixo De Baixo		41.3013	-8.1164	3	41.3013,-8.1164
PT	4600-611	Faia	Porto	17	Amarante	1301	Freixo De Baixo		41.3013	-8.1164	3	41.3013,-8.1164
PT	4600-612	Pardelhas	Porto	17	Amarante	1301	Freixo De Baixo		41.2723	-8.1973	3	41.2723,-8.1973
PT	4600-612	Olivais	Porto	17	Amarante	1301	Freixo De Baixo		41.2723	-8.1973	3	41.2723,-8.1973
PT	4600-612	Mouril	Porto	17	Amarante	1301	Freixo De Baixo		41.2723	-8.1973	3	41.2723,-8.1973
PT	4600-613	Carvalhad	Porto	17	Amarante	1301	Freixo De Baixo		41.2684	-8.1106	3	41.2684,-8.1106

The data found is a lot more scattered throughout the entire Porto district, thus the data is not necessarily centered in the city's core, but comprises a much larger area. But it will prove to be an interesting observation to compare this particular Portuguese city with metropolitan areas from powerful countries.

The Source data for the NY City locations should have the following format (after interpreting the JSON file):

	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

Finally, the data for the venues to be analyzed should have the following format:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

Methodology

First and foremost, the venues of each Manhattan location described in the Data section of this report, were retrieved from Foursquare's API, with the output being a dataframe with following structure:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Astral Fitness & Wellness Center	40.876705	-73.906372	Gym

Then the one-hot coding procedure was done in order to get a matrix with the neighborhoods on the left side and the column names being the type of venue, with 1s marking if the current venue exists for a given location and 0s marking the opposite.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant
0	Marble Hill	0	0	0	0	0	0	0	0	0
1	Marble Hill	0	0	0	0	0	0	0	0	0
2	Marble Hill	0	0	0	0	0	0	0	0	0
3	Marble Hill	0	0	0	0	0	0	0	0	0
4	Marble Hill	0	0	0	0	0	0	0	0	0

Then the data would be structured in the form that will be essential for this case study and will be used for Machine Learning (ML) algorithms:

	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	Hotel	Coffee Shop	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Gourmet Shop	BBQ Joint
1	Carnegie Hill	Coffee Shop	Café	Pizza Place	Yoga Studio	Bookstore	Bakery	French Restaurant	Japanese Restaurant	Italian Restaurant	Bar
2	Central Harlem	African Restaurant	Seafood Restaurant	Bar	American Restaurant	Chinese Restaurant	French Restaurant	Cosmetics Shop	Market	Caribbean Restaurant	Library
3	Chelsea	Coffee Shop	Art Gallery	American Restaurant	Italian Restaurant	Bakery	Seafood Restaurant	French Restaurant	Market	Hotel	Pizza Place
4	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	Salon / Barbershop	Vietnamese Restaurant	Spa	Bubble Tea Shop	Hotpot Restaurant	Ice Cream Shop	Dessert Shop

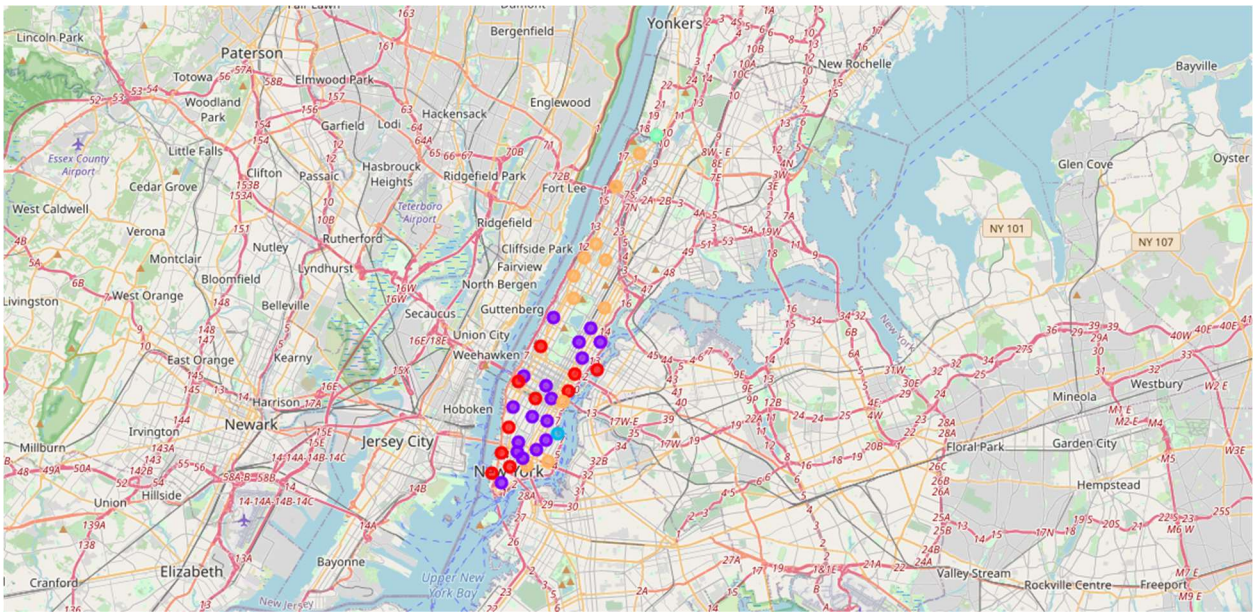
With venues being shown from left to right ordered by their frequency in their corresponding neighborhoods.

Since we don't have label for our data, in other words, we don't have our resulting output "y" for the set of features present in the data, an unsupervised method of ML must be used, which Clustering. This algorithm will ease the process of labeling each individual row of data.

Here, a number of 5 clusters was chosen, given the dimension of the data at hand, and by passing the data through the Clustering technique the following results were obtained:

	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	3	Gym	Coffee Shop	Yoga Studio	Diner	Seafood Restaurant	Sandwich Place	Supplement Shop	Tennis Stadium	Donut Shop	Shopping Mall
1	Manhattan	Chinatown	40.715618	-73.994279	4	Chinese Restaurant	Cocktail Bar	Bakery	Salon / Barbershop	Vietnamese Restaurant	Spa	Bubble Tea Shop	Hotpot Restaurant	Ice Cream Shop	Dessert Shop
2	Manhattan	Washington Heights	40.851903	-73.936900	4	Café	Bakery	Mobile Phone Shop	Chinese Restaurant	Seafood Restaurant	Bank	Tapas Restaurant	Mexican Restaurant	Coffee Shop	Italian Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	4	Mexican Restaurant	Lounge	Restaurant	Café	Frozen Yogurt Shop	Bakery	Spanish Restaurant	Chinese Restaurant	Caribbean Restaurant	American Restaurant
4	Manhattan	Hamilton Heights	40.823604	-73.949688	4	Pizza Place	Coffee Shop	Café	Deli / Bodega	Mexican Restaurant	Bakery	Park	Cocktail Bar	Sandwich Place	Chinese Restaurant

Of course, given the fact that we chose K=5, the obtained Cluster Labels are numbered from 0 to 4 (zero index-based). Then we use Folium package in order to map the obtained labels into their corresponding neighbors throughout New York City, Manhattan.



As you can see, we have obtained 5 separate clusters in 5 different colors, with each of them representing the numbers with the following 10 most common venues:

Cluster 1

	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
11	Roosevelt Island	Park	Hotel	School	Gym	Coffee Shop	Greek Restaurant	Dry Cleaner	Sandwich Place	Liquor Store	Noodle House
13	Lincoln Square	Café	Gym / Fitness Center	Plaza	Theater	Concert Hall	Performing Arts Venue	Wine Shop	American Restaurant	Italian Restaurant	Coffee Shop
21	Tribeca	Park	Italian Restaurant	Café	American Restaurant	Wine Bar	Spa	Coffee Shop	Skate Park	Hotel	Greek Restaurant
24	West Village	Italian Restaurant	New American Restaurant	American Restaurant	Park	Cocktail Bar	French Restaurant	Jazz Club	Coffee Shop	Wine Bar	Theater
28	Battery Park City	Park	Hotel	Coffee Shop	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Gourmet Shop	BBQ Joint
32	Civic Center	Coffee Shop	Cocktail Bar	Hotel	Gym / Fitness Center	Spa	Yoga Studio	Café	French Restaurant	Italian Restaurant	Bakery
33	Midtown South	Korean Restaurant	Hotel	Japanese Restaurant	Burger Joint	Cosmetics Shop	Gym / Fitness Center	Clothing Store	Coffee Shop	Bakery	Pizza Place
34	Sutton Place	Italian Restaurant	Gym / Fitness Center	Furniture / Home Store	Park	Coffee Shop	Gym	Bakery	Thai Restaurant	Beer Bar	Spa
35	Turtle Bay	Coffee Shop	Sushi Restaurant	Italian Restaurant	Wine Bar	Park	Seafood Restaurant	Café	Japanese Restaurant	Deli / Bodega	French Restaurant
39	Hudson Yards	Gym / Fitness Center	Hotel	American Restaurant	Café	Italian Restaurant	Burger Joint	Dog Run	Gym	Park	Coffee Shop

Cluster 2

	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	Italian Restaurant	Bakery	Coffee Shop	Gym / Fitness Center	Yoga Studio	Spa	French Restaurant	Juice Bar	Exhibit	Pizza Place
9	Yorkville	Italian Restaurant	Coffee Shop	Bar	Gym	Sushi Restaurant	Deli / Bodega	Mexican Restaurant	Wine Shop	Pizza Place	Japanese Restaurant
10	Lenox Hill	Coffee Shop	Italian Restaurant	Sushi Restaurant	Pizza Place	Café	Cocktail Bar	Gym / Fitness Center	Gym	Burger Joint	Deli / Bodega
12	Upper West Side	Italian Restaurant	Bar	Wine Bar	Coffee Shop	Mediterranean Restaurant	Bakery	Thai Restaurant	Indian Restaurant	Ice Cream Shop	Middle Eastern Restaurant
14	Clinton	Italian Restaurant	Theater	Gym / Fitness Center	Wine Shop	American Restaurant	Coffee Shop	Sandwich Place	Cocktail Bar	Gym	Hotel
15	Midtown	Coffee Shop	Clothing Store	Hotel	Bakery	Theater	Steakhouse	Bookstore	Pizza Place	Sporting Goods Shop	Japanese Restaurant
16	Murray Hill	Sandwich Place	Japanese Restaurant	Coffee Shop	Burger Joint	Mediterranean Restaurant	Pizza Place	Hotel	American Restaurant	Gym / Fitness Center	Steakhouse
17	Chelsea	Coffee Shop	Art Gallery	American Restaurant	Italian Restaurant	Bakery	Seafood Restaurant	French Restaurant	Market	Hotel	Pizza Place
18	Greenwich Village	Italian Restaurant	Sushi Restaurant	Café	Ice Cream Shop	American Restaurant	Indian Restaurant	French Restaurant	Gym	Coffee Shop	Clothing Store
19	East Village	Bar	Ice Cream Shop	Pizza Place	Mexican Restaurant	Cocktail Bar	Korean Restaurant	Dessert Shop	Wine Bar	Italian Restaurant	Vietnamese Restaurant
22	Little Italy	Bakery	Café	Italian Restaurant	Chinese Restaurant	Ice Cream Shop	Coffee Shop	Salon / Barbershop	Mediterranean Restaurant	Cocktail Bar	Thai Restaurant
23	Soho	Italian Restaurant	Coffee Shop	Clothing Store	Mediterranean Restaurant	Salon / Barbershop	Bakery	Sandwich Place	French Restaurant	American Restaurant	Café
27	Gramercy	Bar	Pizza Place	Bagel Shop	American Restaurant	Italian Restaurant	Coffee Shop	Diner	Thai Restaurant	Playground	Cocktail Bar
29	Financial District	Coffee Shop	Pizza Place	Cocktail Bar	Sandwich Place	Hotel	Bar	Italian Restaurant	Café	Gym	Gym / Fitness Center
30	Carnegie Hill	Coffee Shop	Café	Pizza Place	Yoga Studio	Bookstore	Bakery	French Restaurant	Japanese Restaurant	Italian Restaurant	Bar
31	Noho	Italian Restaurant	Coffee Shop	Pizza Place	Wine Bar	Grocery Store	French Restaurant	Mexican Restaurant	Sandwich Place	Bookstore	Hotel
38	Flatiron	Gym / Fitness Center	Italian Restaurant	New American Restaurant	Cosmetics Shop	Café	Sporting Goods Shop	Vegetarian / Vegan Restaurant	Mediterranean Restaurant	Yoga Studio	Furniture / Home Store

Cluster 3

	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	Park	Playground	Heliport	Fountain	Farmers Market	Skating Rink	Bistro	Gas Station	Baseball Field	Gym / Fitness Center

Cluster 4

	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	Coffee Shop	Yoga Studio	Diner	Seafood Restaurant	Sandwich Place	Supplement Shop	Tennis Stadium	Donut Shop	Shopping Mall

Cluster 5

	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
1	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	Salon / Barbershop	Vietnamese Restaurant	Spa	Bubble Tea Shop	Hotpot Restaurant	Ice Cream Shop	Dessert Shop
2	Washington Heights	Café	Bakery	Mobile Phone Shop	Chinese Restaurant	Seafood Restaurant	Bank	Tapas Restaurant	Mexican Restaurant	Coffee Shop	Italian Restaurant
3	Inwood	Mexican Restaurant	Lounge	Restaurant	Café	Frozen Yogurt Shop	Bakery	Spanish Restaurant	Chinese Restaurant	Caribbean Restaurant	American Restaurant
4	Hamilton Heights	Pizza Place	Coffee Shop	Café	Deli / Bodega	Mexican Restaurant	Bakery	Park	Cocktail Bar	Sandwich Place	Chinese Restaurant
5	Manhattanville	Coffee Shop	Bar	Deli / Bodega	Mexican Restaurant	Seafood Restaurant	Italian Restaurant	Spanish Restaurant	Ramen Restaurant	Dumpling Restaurant	Scenic Lookout
6	Central Harlem	African Restaurant	Seafood Restaurant	Bar	American Restaurant	Chinese Restaurant	French Restaurant	Cosmetics Shop	Market	Caribbean Restaurant	Library
7	East Harlem	Mexican Restaurant	Thai Restaurant	Deli / Bodega	Bakery	Spa	Latin American Restaurant	Sandwich Place	Liquor Store	Coffee Shop	Cocktail Bar
20	Lower East Side	Chinese Restaurant	Cocktail Bar	Japanese Restaurant	Pizza Place	Bakery	Ramen Restaurant	Art Gallery	Coffee Shop	French Restaurant	Mediterranean Restaurant
25	Manhattan Valley	Bar	Yoga Studio	Thai Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Dog Run	Bubble Tea Shop	Café	Caribbean Restaurant
26	Morningside Heights	Park	American Restaurant	Coffee Shop	Bookstore	Burger Joint	Sandwich Place	Deli / Bodega	Supermarket	Salad Place	New American Restaurant
36	Tudor City	Café	Park	Mexican Restaurant	Sushi Restaurant	Deli / Bodega	Greek Restaurant	Garden	Spanish Restaurant	Dog Run	Coffee Shop

Then the approach was the following:

Since we already have our labels through unsupervised learning we can now proceed to supervised learning. We will use here Classification Techniques learned throughout this course, with our features being, of course, the Columns regarding the most common venues, discarding the Neighborhood column.

	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	Gym	Coffee Shop	Yoga Studio	Diner	Seafood Restaurant	Sandwich Place	Supplement Shop	Tennis Stadium	Donut Shop	Shopping Mall
1	Chinese Restaurant	Cocktail Bar	Bakery	Salon / Barbershop	Vietnamese Restaurant	Spa	Bubble Tea Shop	Hotpot Restaurant	Ice Cream Shop	Dessert Shop
2	Café	Bakery	Mobile Phone Shop	Chinese Restaurant	Seafood Restaurant	Bank	Tapas Restaurant	Mexican Restaurant	Coffee Shop	Italian Restaurant
3	Mexican Restaurant	Lounge	Restaurant	Café	Frozen Yogurt Shop	Bakery	Spanish Restaurant	Chinese Restaurant	Caribbean Restaurant	American Restaurant
4	Pizza Place	Coffee Shop	Café	Deli / Bodega	Mexican Restaurant	Bakery	Park	Cocktail Bar	Sandwich Place	Chinese Restaurant
5	Coffee Shop	Bar	Deli / Bodega	Mexican Restaurant	Seafood Restaurant	Italian Restaurant	Spanish Restaurant	Ramen Restaurant	Dumpling Restaurant	Scenic Lookout

Before using this new dataframe as our input variable (y being the cluster labels), we used a method called factorize in order to code each individual venue to a corresponding unique number per column.

	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	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	0	2	2	2	2	2
3	3	3	3	3	2	3	3	3	3	3
4	4	0	4	4	3	3	4	4	4	4
5	5	4	5	5	0	4	3	5	5	5

Then we trained this data for four separate ML Classification algorithms, such as K Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM) and, finally, Logistic Regression.

In the case of KNN, a relation between train data & test data of 80% and 20%, and the model was trained in order to obtain the K which outputs a greater accuracy, which will be discussed in the results section of this report.

This train-test data split was used also, of course, to train the other ML models and their respective accuracies were measured.

The idea of this trained models is to feed a new dataframe that we'll create, which features the "ideal cluster label" in which we want to place our restaurant. Then after obtaining the cluster label for our new entry, we would check the neighborhoods with the resulting label and check which entry would be the closest to our new entry.

The test data comprises of one entry which doesn't feature any restaurant as being at least the most common venue, but with venues that would be great to have right next to the new restaurant.

Test data:

	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	8	11	11	11	12	1	20	7	0	20

The numbers are according to the factorized feature dataframe.

8 – Park

11 – Clothing Store

11 – Plaza

11 – Coffee Shop

12 – American Restaurant

1 – Spa

20 – Dog Run

7 – Liquor Store

0 – Donut Shop

20 - Thai Restaurant

Again we factorize this data and then use as input for the ML models that we dimensioned.

Results

For this new entry, with our models we obtained the following accuracy scores for the 2 main metrics (Jaccard accuracy and F1-score).

ML Techniques Accuracy Scores		
Algorithm	**Jaccard**	**F1-score**
KNN	0.625	0.643
Decision Tree	0.375	0.383
SVM	0.625	0.625
Logistic Regression	0.625	0.611

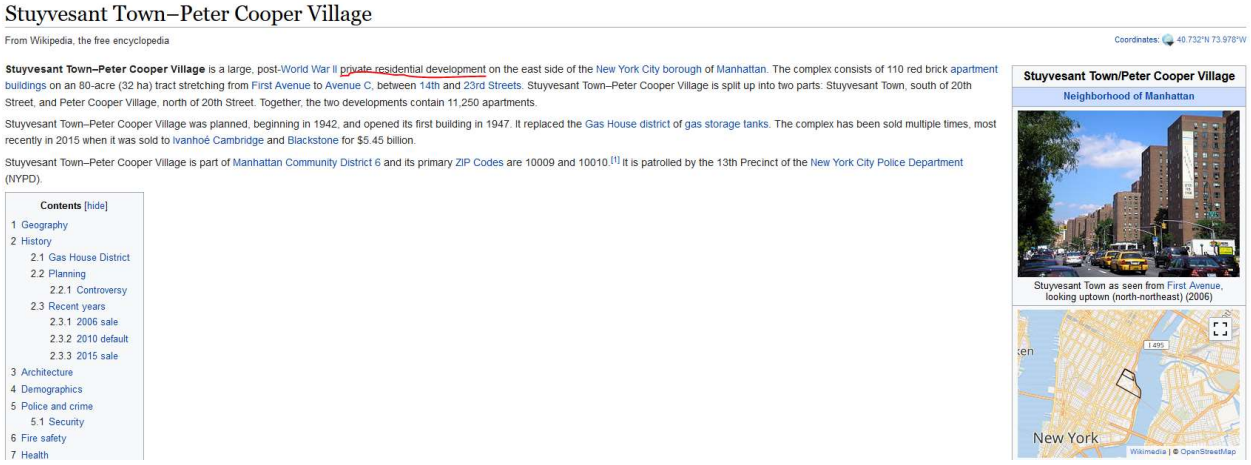
As you can see, the KNN algorithm has the best results overall and was the algorithm chosen for obtaining the label for our new entry.

Passing the dataframe through our KNN trained model we obtained the label 0, which corresponds to **Cluster no. 1**:

	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
11	Roosevelt Island	Park	Hotel	School	Gym	Coffee Shop	Greek Restaurant	Dry Cleaner	Sandwich Place	Liquor Store	Noodle House
13	Lincoln Square	Café	Gym / Fitness Center	Plaza	Theater	Concert Hall	Performing Arts Venue	Wine Shop	American Restaurant	Italian Restaurant	Coffee Shop
21	Tribeca	Park	Italian Restaurant	Café	American Restaurant	Wine Bar	Spa	Coffee Shop	Skate Park	Hotel	Greek Restaurant
24	West Village	Italian Restaurant	New American Restaurant	American Restaurant	Park	Cocktail Bar	French Restaurant	Jazz Club	Coffee Shop	Wine Bar	Theater
28	Battery Park City	Park	Hotel	Coffee Shop	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Gourmet Shop	BBQ Joint
32	Civic Center	Coffee Shop	Cocktail Bar	Hotel	Gym / Fitness Center	Spa	Yoga Studio	Café	French Restaurant	Italian Restaurant	Bakery
33	Midtown South	Korean Restaurant	Hotel	Japanese Restaurant	Burger Joint	Cosmetics Shop	Gym / Fitness Center	Clothing Store	Coffee Shop	Bakery	Pizza Place
34	Sutton Place	Italian Restaurant	Gym / Fitness Center	Furniture / Home Store	Park	Coffee Shop	Gym	Bakery	Thai Restaurant	Beer Bar	Spa
35	Turtle Bay	Coffee Shop	Sushi Restaurant	Italian Restaurant	Wine Bar	Park	Seafood Restaurant	Café	Japanese Restaurant	Deli / Bodega	French Restaurant
39	Hudson Yards	Gym / Fitness Center	Hotel	American Restaurant	Café	Italian Restaurant	Burger Joint	Dog Run	Gym	Park	Coffee Shop

Discussion

Cluster 3, which only has the neighborhood Stuyvesant Town, can be discarded since by just googling location we notice that this is a large private residential development, which are usually not a great ideal place to build your restaurant.



Stuyvesant Town–Peter Cooper Village

From Wikipedia, the free encyclopedia

Coordinates: 40.732°N 73.978°W

Stuyvesant Town–Peter Cooper Village is a large, post-World War II private residential development on the east side of the New York City borough of Manhattan. The complex consists of 110 red brick apartment buildings on an 80-acre (32 ha) tract stretching from First Avenue to Avenue C, between 14th and 23rd Streets. Stuyvesant Town–Peter Cooper Village is split up into two parts: Stuyvesant Town, south of 20th Street, and Peter Cooper Village, north of 20th Street. Together, the two developments contain 11,250 apartments.

Stuyvesant Town–Peter Cooper Village was planned, beginning in 1942, and opened its first building in 1947. It replaced the Gas House district of gas storage tanks. The complex has been sold multiple times, most recently in 2015 when it was sold to Ivanhoé Cambridge and Blackstone for $5.45 billion.

Stuyvesant Town–Peter Cooper Village is part of Manhattan Community District 6 and its primary ZIP Codes are 10009 and 10010.[1] It is patrolled by the 13th Precinct of the New York City Police Department (NYPD).

Contents [hide]

1 Geography
2 History
2.1 Gas House District
2.2 Planning
2.2.1 Controversy
2.3 Recent years
2.3.1 2006 sale
2.3.2 2010 default
2.3.3 2015 sale
3 Architecture
4 Demographics
5 Police and crime
5.1 Security
6 Fire safety
7 Health

Stuyvesant Town/Peter Cooper Village

Neighborhood of Manhattan

Stuyvesant Town as seen from First Avenue, looking uptown (north-northeast) (2006)

New York

Wikimedia | © OpenStreetMap

Cluster 4 can also be discarded, but for a different reason. We only have one entry for this particular cluster which, if the model's chosen label were to be this one, would not be a result giving us great confidence in it.

Doing a statistical analysis for Cluster 5 we obtain the following:

	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
count	11	11	11	11	11	11	11	11	11	11	11
unique	11	8	10	10	11	10	10	10	11	9	11
top	Inwood	Mexican Restaurant	Cocktail Bar	Deli / Bodega	Mexican Restaurant	Seafood Restaurant	Bakery	Spanish Restaurant	Mexican Restaurant	Coffee Shop	Scenic Lookout
freq	1	2	2	2	1	2	2	2	1	2	1

Restaurants are really predominant for this cluster, so it just isn't a very good candidate.

We are left with cluster 1 and 2.

Cluster 2 has the following statistics:

	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
count	17	17	17	17	17	17	17	17	17	17	17
unique	17	6	12	14	17	14	13	12	15	15	14
top	Little Italy	Italian Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Mediterranean Restaurant	Bakery	French Restaurant	Mediterranean Restaurant	Italian Restaurant	Japanese Restaurant
freq	1	7	3	3	1	2	3	4	2	2	2

Which features a great predominance of restaurants with a total number of 20 most common restaurants (not counting the Pizza Places). In contrast with cluster 1, which has the following statistics:

	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
count	10	10	10	10	10	10	10	10	10	10	10
unique	10	6	6	9	8	9	10	9	9	9	9
top	West Village	Park	Hotel	American Restaurant	Park	Coffee Shop	French Restaurant	Café	Coffee Shop	Italian Restaurant	Coffee Shop
freq	1	3	4	2	2	2	1	2	2	2	2

The total number of existing most common restaurants is 5. With the first column (1st Most Common Venue) not even being a restaurant. Even the 2nd one isn't a restaurant, but a Hotel, which is great for tourists who stay at their hotel and look for a place to eat which is close to the hotel.

Given this statistics, we can state with fairly good confidence that we should place our new restaurant in a neighborhood with Cluster 1.

Looking back at Cluster 1's dataframe:

	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
11	Roosevelt Island	Park	Hotel	School	Gym	Coffee Shop	Greek Restaurant	Dry Cleaner	Sandwich Place	Liquor Store	Noodle House
13	Lincoln Square	Café	Gym / Fitness Center	Plaza	Theater	Concert Hall	Performing Arts Venue	Wine Shop	American Restaurant	Italian Restaurant	Coffee Shop
21	Tribeca	Park	Italian Restaurant	Café	American Restaurant	Wine Bar	Spa	Coffee Shop	Skate Park	Hotel	Greek Restaurant
24	West Village	Italian Restaurant	New American Restaurant	American Restaurant	Park	Cocktail Bar	French Restaurant	Jazz Club	Coffee Shop	Wine Bar	Theater
28	Battery Park City	Park	Hotel	Coffee Shop	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Gourmet Shop	BBQ Joint
32	Civic Center	Coffee Shop	Cocktail Bar	Hotel	Gym / Fitness Center	Spa	Yoga Studio	Café	French Restaurant	Italian Restaurant	Bakery
33	Midtown South	Korean Restaurant	Hotel	Japanese Restaurant	Burger Joint	Cosmetics Shop	Gym / Fitness Center	Clothing Store	Coffee Shop	Bakery	Pizza Place
34	Sutton Place	Italian Restaurant	Gym / Fitness Center	Furniture / Home Store	Park	Coffee Shop	Gym	Bakery	Thai Restaurant	Beer Bar	Spa
35	Turtle Bay	Coffee Shop	Sushi Restaurant	Italian Restaurant	Wine Bar	Park	Seafood Restaurant	Café	Japanese Restaurant	Deli / Bodega	French Restaurant
39	Hudson Yards	Gym / Fitness Center	Hotel	American Restaurant	Café	Italian Restaurant	Burger Joint	Dog Run	Gym	Park	Coffee Shop

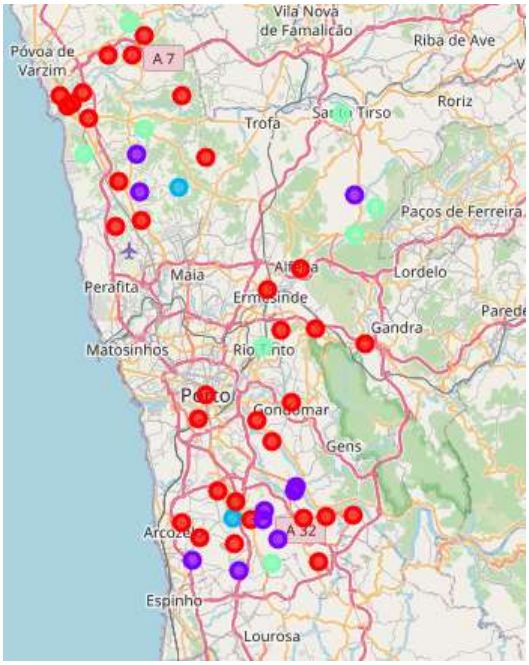
We can already discard neighborhoods Tribeca, West Village, Midtown South, Sutton Place, Turtle Bay and Hudson Yards because they have restaurants at least as their 3rd most common venue. Leaving us with just Roosevelt Island, Lincoln Square, Battery Park City and Civic Center neighborhoods.

	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
11	Roosevelt Island	Park	Hotel	School	Gym	Coffee Shop	Greek Restaurant	Dry Cleaner	Sandwich Place	Liquor Store	Noodle House
13	Lincoln Square	Café	Gym / Fitness Center	Plaza	Theater	Concert Hall	Performing Arts Venue	Wine Shop	American Restaurant	Italian Restaurant	Coffee Shop
21											
24											
28	Battery Park City	Park	Hotel	Coffee Shop	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Gourmet Shop	BBQ Joint
32	Civic Center	Coffee Shop	Cocktail Bar	Hotel	Gym / Fitness Center	Spa	Yoga Studio	Café	French Restaurant	Italian Restaurant	Bakery
33											
34											
35											
39											

Then in order to choose just one neighborhood, we started to analyze from left to right until we found the first neighborhood with a restaurant as its most common venue, which is Roosevelt Island with a Greek Restaurant as its 6th most common venue.

Again, reading from left to right, we get to the "8th Most Common Venue" column, which has, for both Lincoln Square and Civic Center a restaurant. We eliminate these entries and reach the conclusion that our best choice of a neighborhood is **Battery Park City**, featuring virtually no restaurant as its most common venue, featuring entries such as Park and Plaza which are in the single-entry dataframe that we used as test data.

Now which type of restaurant do we build? For this we looked at several cities like Toronto, Paris and Germany and found that they are somewhat similar in terms of venues, since they're all multicultural cities (see more in the jupyter notebooks in the links). But I tried to look into my own country, Portugal, regarding a city that I love and with fantastic cuisine, which is Porto.



Clusters such as the one shown in the following dataframe feature Portuguese Restaurants which would certainly be a differentiator in a city such as New York City, which, at least according to the data gathered in this study, seems to lack.

	Borough	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
38	Porto	1.0	Portuguese Restaurant	Winery	Fish & Chips Shop	Concert Hall	Convenience Store	Creperie	Dessert Shop	Diner	Dumpling Restaurant	Dutch Restaurant
94	Porto	1.0	Historic Site	Portuguese Restaurant	Winery	Fish & Chips Shop	Concert Hall	Convenience Store	Creperie	Dessert Shop	Diner	Dumpling Restaurant
528	Porto	1.0	Portuguese Restaurant	Furniture / Home Store	Winery	Fish & Chips Shop	Concert Hall	Convenience Store	Creperie	Dessert Shop	Diner	Dumpling Restaurant

And, being the Portuguese cuisine a very fine cuisine and Battery Park City a neighborhood lacking in restaurants but with places to take a walk and for tourists to spend their nights, it definitely seems like a match made in heaven!

Conclusion

There a few limitations with this study. Foursquare's API doesn't always return the same results which does not help at all in the analysis, the lack of geographical points to some cities (like Porto for example, in which the entire district of Porto had to be considered and not just the city itself). Also, this approach may not be optimal since the dataframe used as test data was put together from the top of my head, with places that I think should be next to a restaurant.

An American restaurant is in that same dataframe, but I opted for a more realistic approach, since it is difficult to assume that in New York City, a place already filled up to the top with so many restaurants (as discussed in the Introduction section), would have a neighborhood with no restaurants whatsoever. So it is very easy to imply that you'll probably find an American Restaurant in your neighborhood.

The Thai Restaurant as the last one was chosen since it is usual to have restaurants of foreign cuisine in American neighborhood. However, in this study a Portuguese restaurant was not found!

In pretty much very restaurant-focused data, it is easy to assume these options.

The test data set could be, of course, enhanced in order to feature more entries to check the validity of the model according to the results obtained in this work.

And, also, many more cities could have been researched upon in order to place the new restaurant, however New York City, as it was shown, proved to be a great challenge.