Applied Data Science Capstone

Finding the best Belgian restaurant location in NYC

By Brecht Beertens | 13 January 2021

Table of Contents

1. Introduction & Business Problem

Starting a new business is a big risk. However, there are a lot of things that we can do to mitigate that risk. One of these things is using data science to find out the prime location to set up shop. If you pick a location with high foot traffic, affordable rent and that's quite safe to boot, you can limit the risk involved.

In this report, we'll be analysing New York City to find the optimal location for a Belgian restaurant. In order to find the best possible neighbourhood, we'd need to consult multiple data sources. As New York City is enormous, it's more efficient to start looking at the city on a borough-level, rather than a neighbourhood-level. As such, this report will include a first, brief analysis of on borough-level to determine the best borough, before we drill down further and conduct a full analysis on each neighbourhood in the selected borough.

2. Data

2.1 Borough-level data

When setting up a business, it's crucial that the surrounding area is lively and willing to spend. As such, we'll need to compare each borough based on **GDP** and **total population**. This data is available on Wikipedia (https://en.wikipedia.org/wiki/Boroughs_of_New_York_City) in a table format. We can scrape this data into a dataframe to use it for further analysis.

However, these aren't the only important metrics to compare. A new business benefits from a safe environment, which means we also have to take crime rates into account. We can get a fragmented overview of the **number of crimes per borough** through the NYPD-website (https://www1.nyc.gov/site/nypd/stats/crime-statistics/borough-and-precinct-crime-stats.page). The NYPD only lists these per borough, so we'll have to compile the Excel-files ourselves before we are able to conduct a full analysis.

We can then use this to calculate the number of crimes per capita and split the crimes between misdemeanours and felonies. Ideally, we will find a borough that has a high GDP & population, while having a relatively low crime rate.

2.2 Neighbourhood-level data

Once the ideal borough has been found, we can zoom in and have a look at the best neighbourhood in that borough. There are a few important factors that will influence the decision.

First of all, we need to have a look at the **other businesses in the neighbourhood**. We could either look for a neighbourhood with a high density of restaurants of multiple cuisines, as this would be a neighbourhood that's frequented by foodies who are looking to try something new.

On the other hand, we could also look for a neighbourhood that has a very small number of restaurants so that the competition is rather limited. However, we have to keep in mind that having an existing Belgian restaurant in a neighbourhood would make business harder than it has to be. All of this data will be obtained through the Foursquare API (https://api.foursquare.com)

Another thing to keep in mind is the average housing costs in the neighbourhood. Streeteasy gives an overview of the average housing cost per neighbourhood. You can consult their data here: https://streeteasy.com/blog/data-dashboard/

It is worth noting that these are not the prices for retail spaces, which are generally more expensive. However, as there was no trustworthy commercial real-estate data readily available, this data will give us an indication of how expensive it would be to set up our business in that neighbourhood.

3. Boroughs - Methodology

3.1 General demographic data

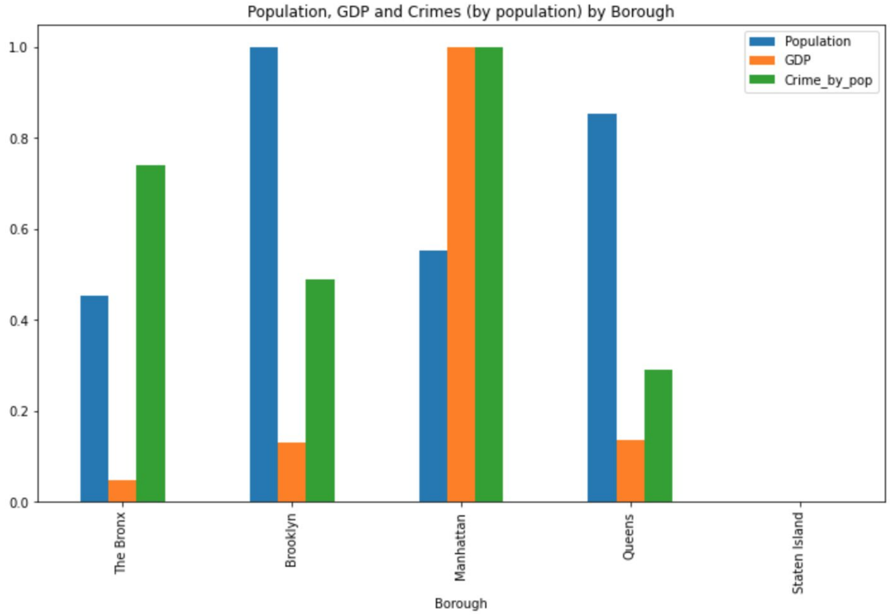
Let's have a look at the boroughs of New York City first. Most boroughs have over 30 neighbourhoods, so it would be wise to select a suitable borough before we look at the neighbourhoods. If we were to look at every neighbourhood at once, the free Foursquare API wouldn't be sufficient.

First of all, I had a look at the crimes (per capita), population and GDP (per capita) of each area. This data was gathered by scraping the Wikipedia article "Boroughs of New York City" with Beautiful Soup and then converting it into a dataframe.

I applied min-max feature scaling to this data to make it easier to compare the metrics in a single graph. This means that my highest values are equal to 1, my lowest values are equal to 0 and everything else is proportionally distributed between these points.

You can achieve this form of normalisation by applying the following equation to your dataset.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

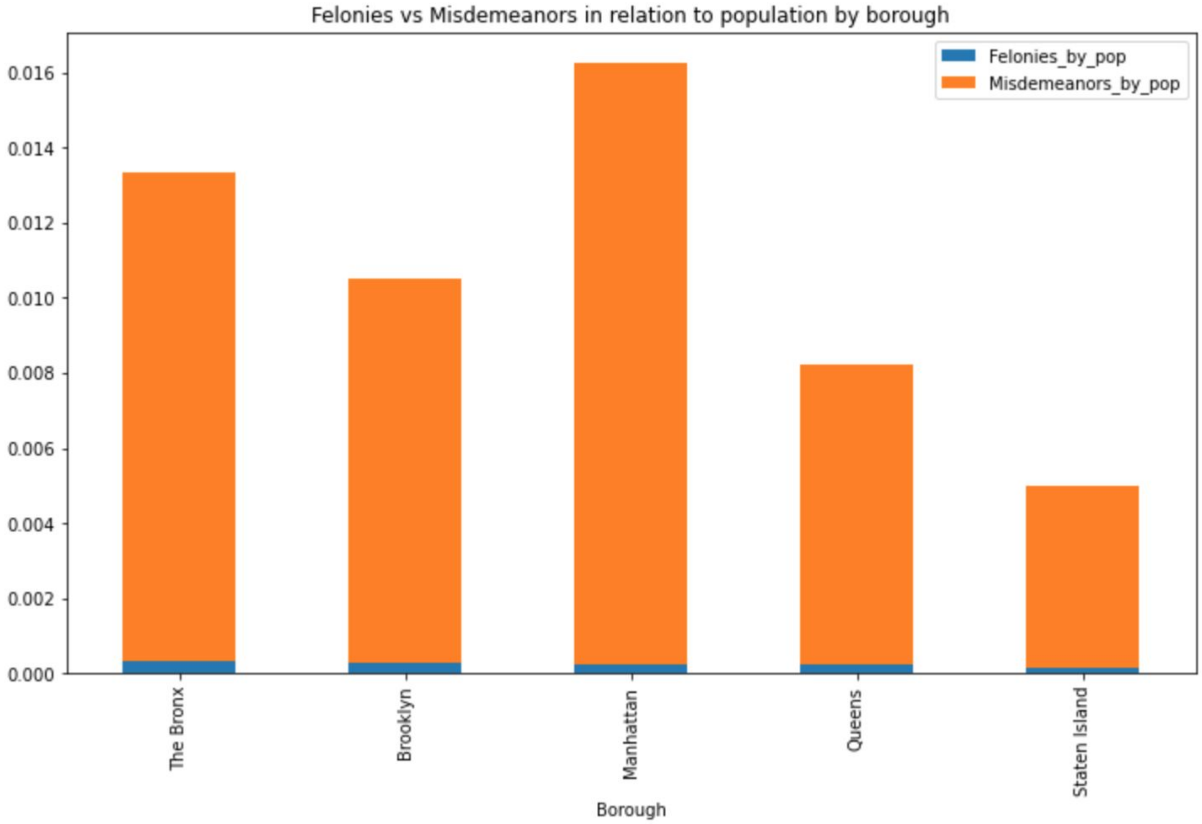


3.2 A more in-depth look at crime in each borough

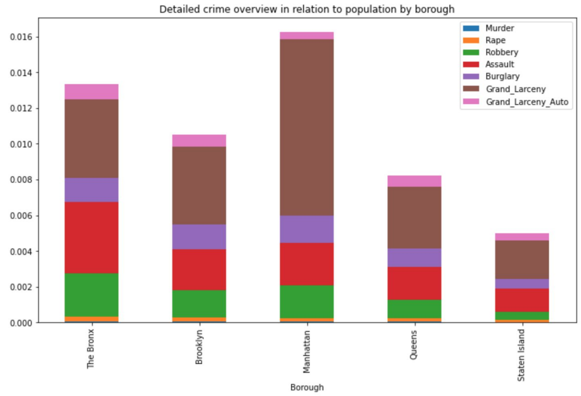
It's important to not just look at crime as a whole. We need to identify which crimes are more or less common in each borough. After all, if borough A had 30 crimes and borough B had 50, but 15 out of 30 were murder in borough A, while only 10 out of 50 in borough B were murder, I'd much rather live in borough B.

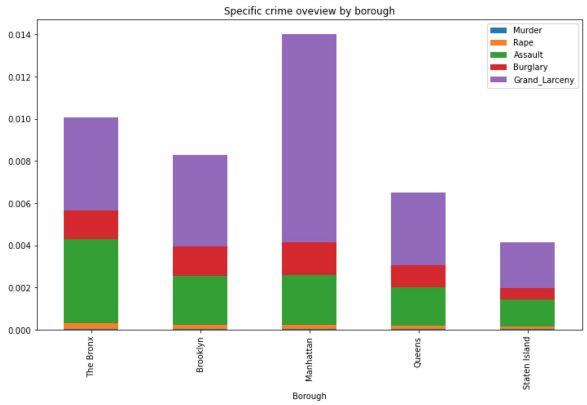
In order to identify more general trends, I downloaded each Excel-file on the NYPD's website. I then copy-pasted the needed columns into one file and exported it as a .csv. This .csv-file became the basis of my dataframe. After that, I added a few calculated fields, mostly by dividing the number of crimes by the total number of inhabitants of that borough.

I created a graph plotting the number of felonies and the number of misdemeanours against each other. The felonies included in my dataset were murder and rape, while the included misdemeanours were assault, burglary, grand larceny, grand larceny auto and robbery.



While this graph will help me identify high-over trends, I also need more detailed graphs. Here I decided to make two different graphs. The first graph (left) shows an overview of crime by borough for each specific type in my dataset, while the second graph (right) highlights a few select crimes.





Our selected crimes for the graph on the right and the reasons for selecting them are:

- Murder, Assault & Rape: The more these crimes happen in the area, the more residents will be afraid to go out at night. This means you'll get less local customers which will hurt your bottom line.
- Burglary & Grand Larceny: As a new business, you want to limit the risk of being the victim of these crimes. You spent a lot of money to decorate your new restaurant. When they steal your valuable possessions after you've invested most of your money into your new business, chances are it'll ruin you financially.

4. Boroughs - Results

As you can see, Staten Island has the lowest population, the lowest GDP and the lowest crime rate. Brooklyn clearly houses the highest number of potential customers, but Queens isn't too far off while being much safer. Manhattan, surprisingly, came out as the borough with the highest crime rate, while The Bronx is supposed to be quite notorious for its high crime rate.

When we have a more in-depth look at crime though, we notice that the number of felonies per capita is rather consistent across the boroughs. The Bronx is slightly ahead with its 0.000331 felonies per capita, where most other boroughs are in the area of 0.00023. The number of misdemeanours per capita, however, clearly peaks in Manhattan and less so in The Bronx.

While examining the graph with the specific crime data, we see that the distribution of the different types of crimes is rather consistent, save for two instances. The number of assaults in The Bronx and the number of grand larcenies in Manhattan. If we zoom in a bit further to filter out the 'less important' crimes, we see that Manhattan & The Bronx still come out on top.

5. Boroughs - Discussion

If we look at the general demographic data, we can already create a shortlist of boroughs that are good contenders: Manhattan, Brooklyn & Queens. Staten Island is discarded due to its low population and GDP per capita, while The Bronx is discarded due to its higher crime rate.

While Manhattan has more crime than The Bronx, it's enormous GDP per capita still got it on the shortlist. However, when we see the high number of grand larcenies taking place in the area, we can rule it out as well. We have to take into account that the GDP of this area is likely inflated due to the presence of Wall Street and is not representative of the borough as a whole.

When we compare Queens & Brooklyn, they are quit evenly matched. In the end, I'd prefer Queens over Brooklyn. While the population there is lower than in Queens, it's also much safer and has a lower rate of assault, which is good for business.

6. Boroughs - Conclusion

As Queens is safer, I would much prefer this borough for a first business. Due to Brooklyn's higher population, but also its higher risk of being the victim of crime, it is still a great pick if you're already established and want to open up a new location. However, considering that we're starting from scratch in this case, Queens is the optimal borough to start our restaurant in.

7. Neighbourhoods - Methodology

7.1 Gathering the basic neighbourhood-level data

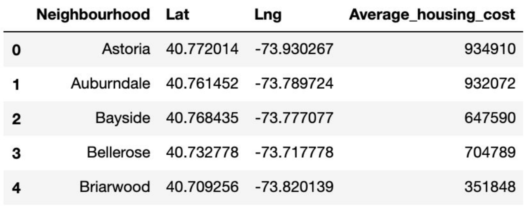
First of all, I downloaded the zip-file from Streeteasy and removed some unnecessary columns in Excel, after which I exported it as a .csv containing the name of the neighbourhood and the average housing cost.

However, with just this data, we cannot run any queries through the Foursquare API or visualise the data through Folium. We need to add the coordinates as well. This was done by looping my list of neighbourhoods through geopy's geolocator.geocode.

While doing so, I realised that it's quite important to further specify the location in your query. When I checked my API query results, I found out that 'Northwestern Queens' had suddenly returned venues like 'Oslo Airport'. I did this by running my query like this, with 'inputNeigh' being the neighbourhood's name:

```
location = geolocator.geocode("{}, Queens, New York,
USA".format(inputNeigh))
```

I ended up losing 2 neighbourhoods from my initial list, as geopy couldn't pinpoint their coordinates. These were Jamaica Hills and Central Queens. Jamaica Hills is rather small and mostly residential, while Central Queens is a collection of other neighbourhoods like Forest Hills. Since the smaller neighbourhoods were all still present, I decided to continue the analysis without these neighbourhoods.

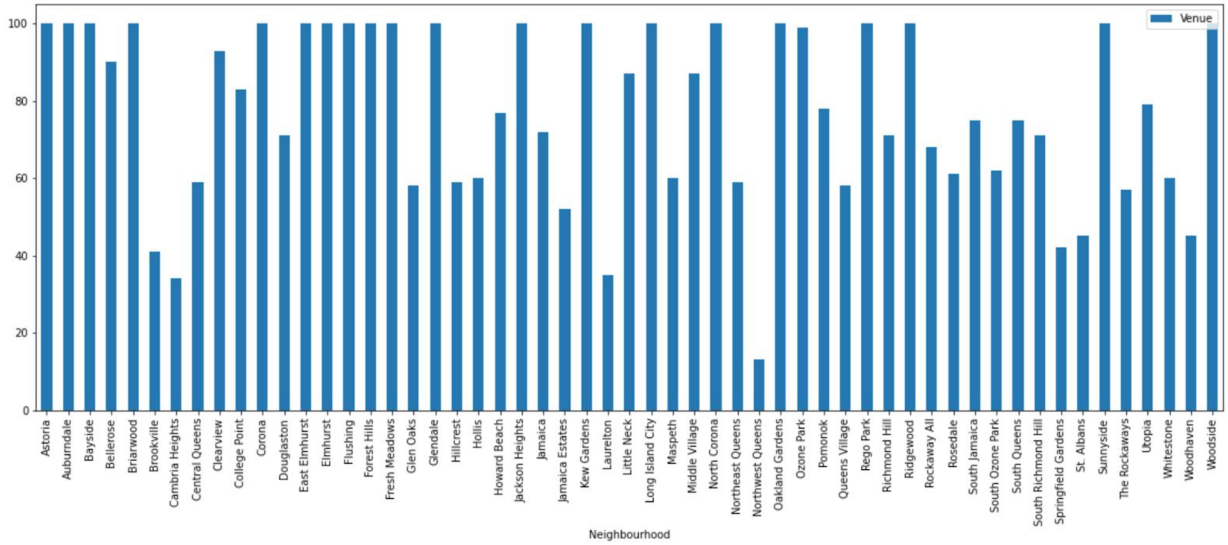


	Neighbourhood	Lat	Lng	Average_housing_cost
0	Astoria	40.772014	-73.930267	934910
1	Auburndale	40.761452	-73.789724	932072
2	Bayside	40.768435	-73.777077	647590
3	Bellerose	40.732778	-73.717778	704789
4	Briarwood	40.709256	-73.820139	351848

7.2 Adding venue data through Foursquare & calculations

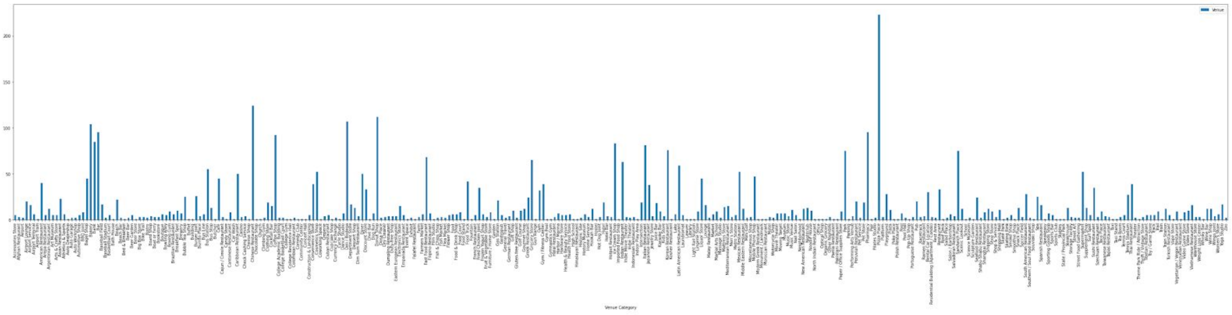
Once the basic dataframe was made, I used the neighbourhood names, latitude and longitude values to run queries through Foursquare's API. My queries returned the "Venue", "Venue latitude", "Venue longitude" and "Venue category" for 4236 venues, belonging to 322 unique categories.

I then plotted out a graph showing the total number of retrieved venues by neighbourhood. This would allow me to make sure I had enough data to continue my analysis. Afterwards, the total number of retrieved venues was also added to the dataframe.



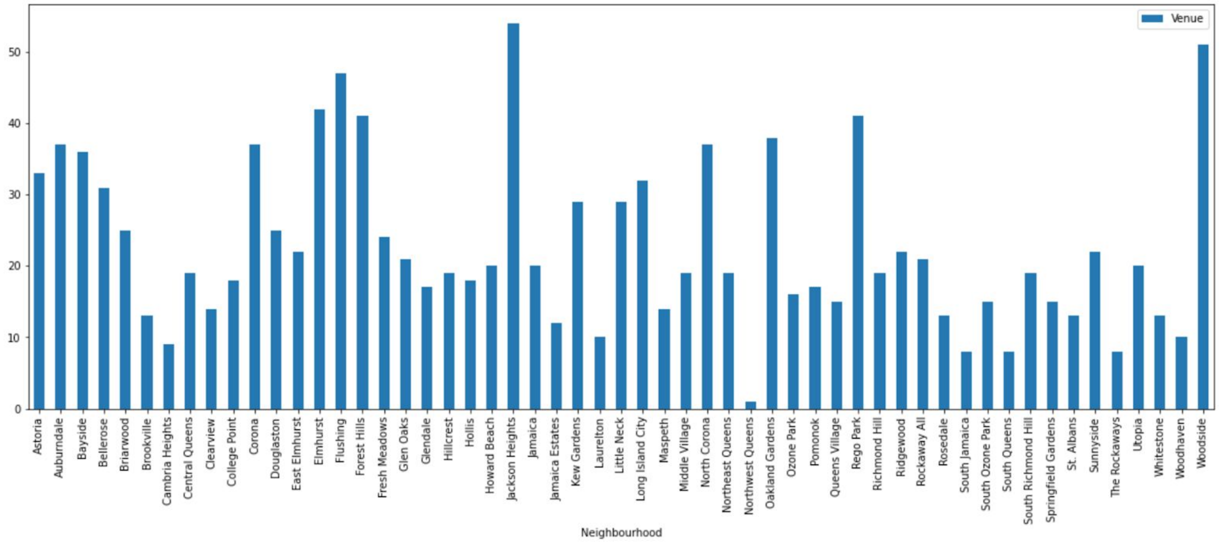
While some neighbourhoods definitely returned less venues than others, I couldn't increase my radius much further than 1.25km out of fear of clipping another neighbourhood when I run the query on a smaller neighbourhood.

I also plotted out the distribution of venues across all neighbourhoods to see if there were any venues that were more common than others.

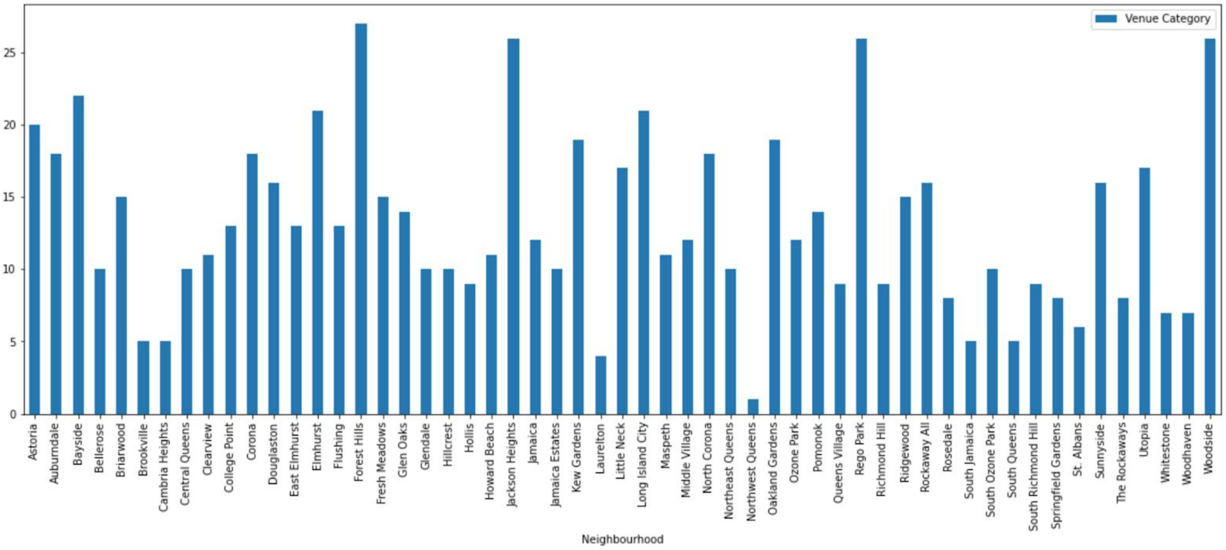


Next to this, I decided to calculate the total number of restaurants in the neighbourhoods as well. I did this by filtering my venues dataframe on a simple condition. When the Venue Category contains the words 'restaurant', 'joint' or 'food', I would include them. 'Restaurant' speaks for itself, while joint was commonly used for burger or fried chicken places. By adding the word food as well, I'd also include food trucks.

I plotted this data into the following graph, as well as added it to my dataframe as "Total_restaurants". Based on the "Total_restaurants" and "Total_venues" I had available, I also calculated the "Restaurant_rate" for each neighbourhood by dividing the total number of restaurants by the total number of venues.



The final interesting number I added in this step was "Unique_restaurant_types", by counting the unique entries in the "Venue categories" column of my filtered dataframe containing only restaurants. This data will help us determine the neighbourhood with the most diverse cuisines and is visualised in the graph below.

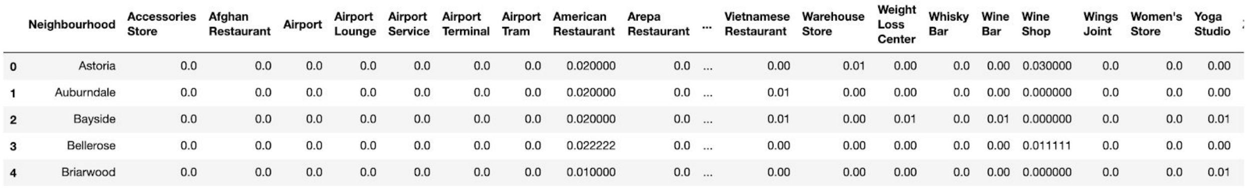


7.3 Preparing the data for machine learning

If we're going to use machine learning with the current data set, there's one small problem. The venue categories are all strings instead of integers or floats. To change this, I used a process called 'one hot encoding'. This will create a separate column for each string and assign either a 1 or a 0 in the column.

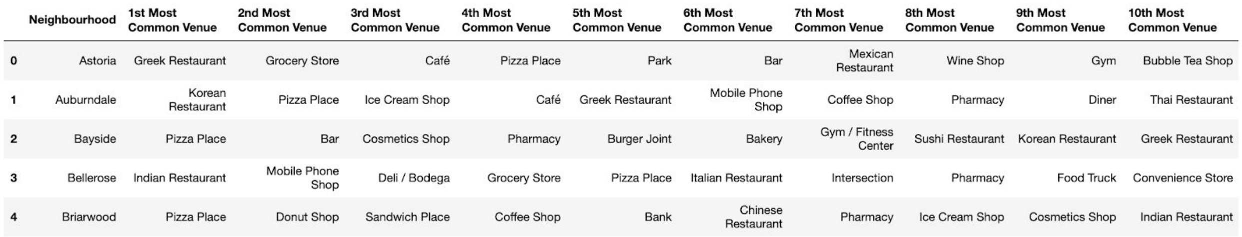
For example, if we were to use one hot encoding on a column named 'meal_type' that could either contain 'cold' or 'warm', it would create 2 columns: 'cold' and 'warm'. If a row's value was 'cold', the cold-column would contain 1 while the warm-column would contain 0.

Once my entire dataframe was encoded, I grouped the results by neighbourhood and got the average value of the 1's and 0's for each column. The higher the resulting value was, the more often it occurred in that neighbourhood. This resulted in a dataframe of 323 columns, that looked like this:



	Neighbourhood	Accessories Store	Afghan Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	Airport Tram	American Restaurant	Arepa Restaurant	...	Vietnamese Restaurant	Warehouse Store	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Astoria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0	...	0.00	0.01	0.00	0.0	0.00	0.030000	0.0	0.0	0.00
1	Auburndale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0	...	0.01	0.00	0.00	0.0	0.00	0.000000	0.0	0.0	0.00
2	Bayside	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0	...	0.01	0.00	0.01	0.0	0.01	0.000000	0.0	0.0	0.01
3	Bellerose	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.022222	0.0	...	0.00	0.00	0.00	0.0	0.00	0.011111	0.0	0.0	0.00
4	Briarwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.010000	0.0	...	0.00	0.00	0.00	0.0	0.00	0.000000	0.0	0.0	0.01

This dataframe let me create an overview of the top 10 most popular venues by neighbourhood. I would simply have to sort my columns in a descending order for a specific neighbourhood, get the first column names 2 up until and including 11 and add that to a new dataframe. I then used a loop to do this for each neighbourhood, resulting in this dataframe, which would be used more in a later phase.



	Neighbourhood	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	Astoria	Greek Restaurant	Grocery Store	Café	Pizza Place	Park	Bar	Mexican Restaurant	Wine Shop	Gym	Bubble Tea Shop
1	Auburndale	Korean Restaurant	Pizza Place	Ice Cream Shop	Café	Greek Restaurant	Mobile Phone Shop	Coffee Shop	Pharmacy	Diner	Thai Restaurant
2	Bayside	Pizza Place	Bar	Cosmetics Shop	Pharmacy	Burger Joint	Bakery	Gym / Fitness Center	Sushi Restaurant	Korean Restaurant	Greek Restaurant
3	Bellerose	Indian Restaurant	Mobile Phone Shop	Deli / Bodega	Grocery Store	Pizza Place	Italian Restaurant	Intersection	Pharmacy	Food Truck	Convenience Store
4	Briarwood	Pizza Place	Donut Shop	Sandwich Place	Coffee Shop	Bank	Chinese Restaurant	Pharmacy	Ice Cream Shop	Cosmetics Shop	Indian Restaurant

7.4 Why kmeans clustering?

Kmeans clustering is an algorithm that excels at identifying clusters in a dataset. It's a form of unsupervised learning, which means it doesn't need some pre-labeled data to be trained on. It'll spot the trends in the data on its own. Creating clusters will let me dentify similar neighbourhoods which will help me filter out uninteresting neighbourhoods.

Kmeans will split the data into 'k' clusters, k being a number we have to define. Each cluster starts as what's called a 'centroid' or a point that will become the centre of its cluster. The distance from each data point to the centroids is calculated and the data point is assigned to the centroid it's closest to.

The centroid is then moved to the mean location of all of its data points. This process keeps on repeating until the centroids no longer move. This is the point where the clustering is finalised.

7.5 Using kmeans clustering on the dataset

7.5.1 Calculating the optimal k-value

Before the clustering can start, I need a certain number of clusters. One way to calculate this is by using what's called the Elbow Method. In this method we calculate the 'distortion' or Sum of Squared Errors (SSE) that occurs for multiple k values.

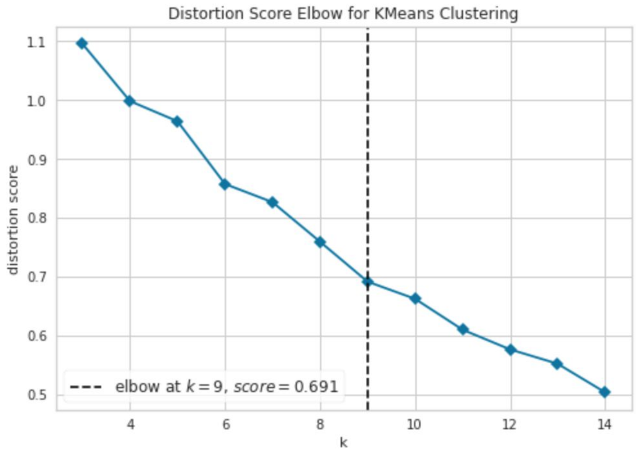
We get this by subtracting the centroid value from the data point's value, squaring our result and then adding this value for each datapoint.

$$\sum\left(Data\ X_i\ -\ Centroid\ X\right)^2 + \left(Data\ Y_i\ -\ Centroid\ Y\right)^2 \ldots$$

When we plot the distortion for each k-value, you'll often see an angle in your graph. This is what's called an 'elbow'. This elbow is the breaking point where any extra clusters would divide your data more than necessary and cause overfitting.

To generate my plot I used Yellowbrick's KElbowVisualizer. This can be done with only a few lines of code and is much faster than doing it manually. You can find the code and resulting graph below.

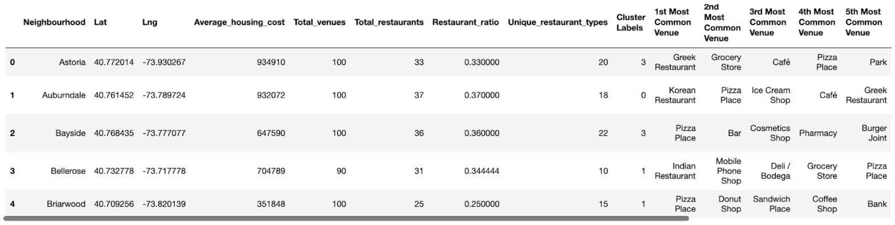
```
# Step 1: Import the needed resources to use Yellowbrick's Elbow method
!pip install yellowbrick
from yellowbrick.cluster import KElbowVisualizer

# Step 2: Set the correct model, kmeans in this case
model = KMeans()

# Step 3: Call the visualiser and define a range of k values that you want
to test
visualizer = KElbowVisualizer(model, k=(3,15), timings=False)

# Step 4: Fit the data and show the graph
visualizer.fit(Queens_grouped_kmeans)
visualizer.show()
```



7.5.2 Generating our clusters

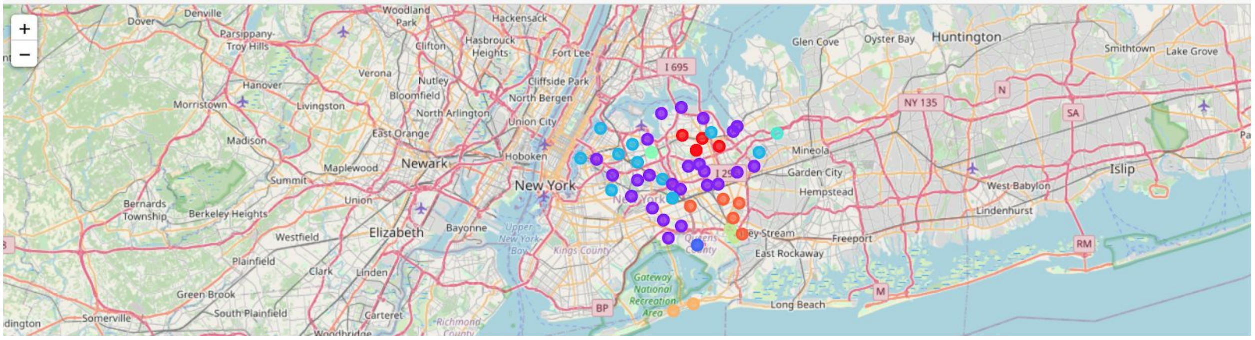
Now that the optimal k has been calculated, it's time to generate the 9 clusters using scikitlearn's kmeans algorithm.

Once the clustering was done, I added the cluster labels & my most common venue categories to my initial dataframe, resulting in the following dataframe.



	Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Astoria	40.772014	-73.930267	934910	100	33	0.330000	20	3	Greek Restaurant	Grocery Store	Café	Pizza Place	Park
1	Auburndale	40.761452	-73.789724	932072	100	37	0.370000	18	0	Korean Restaurant	Pizza Place	Ice Cream Shop	Café	Greek Restaurant
2	Bayside	40.768435	-73.777077	647590	100	36	0.360000	22	3	Pizza Place	Bar	Cosmetics Shop	Pharmacy	Burger Joint
3	Bellerose	40.732778	-73.717778	704789	90	31	0.344444	10	1	Indian Restaurant	Mobile Phone Shop	Deli / Bodega	Grocery Store	Pizza Place
4	Briarwood	40.709256	-73.820139	351848	100	25	0.250000	15	1	Pizza Place	Donut Shop	Sandwich Place	Coffee Shop	Bank

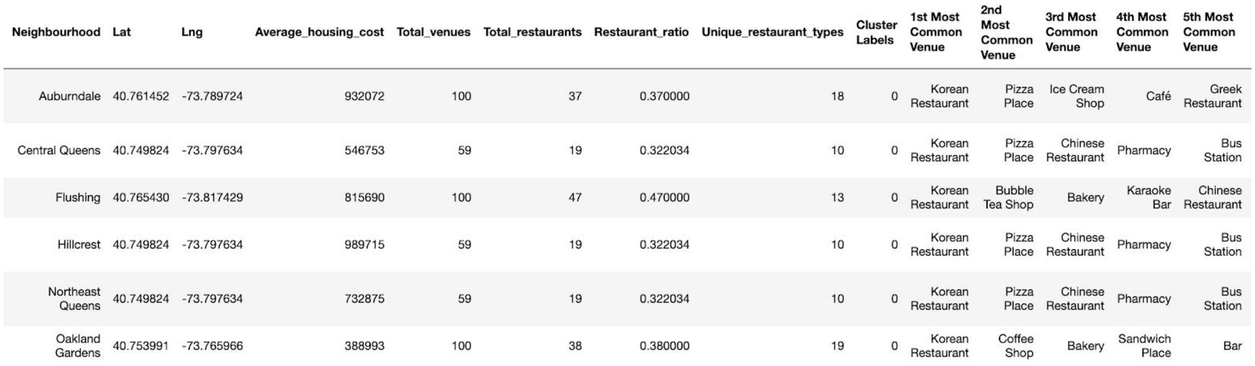
When this data is visualised on a Folium map, we see the following result



7.6 Viewing the clusters

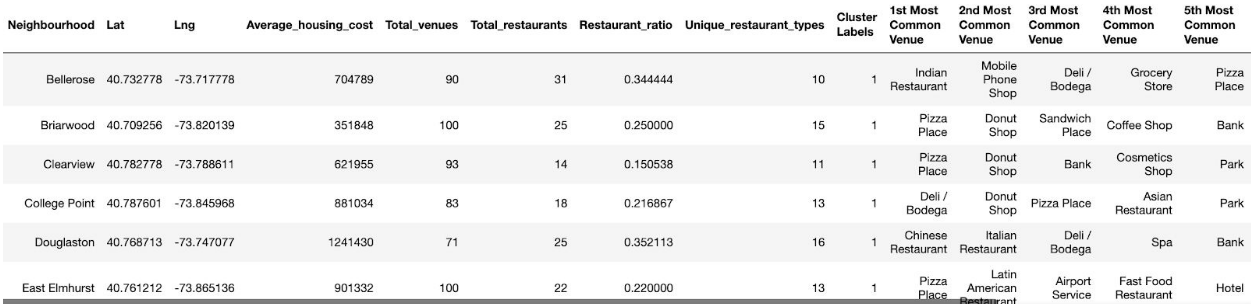
When we create a single dataframe for analysis inside each cluster, we see the following:

Cluster 1



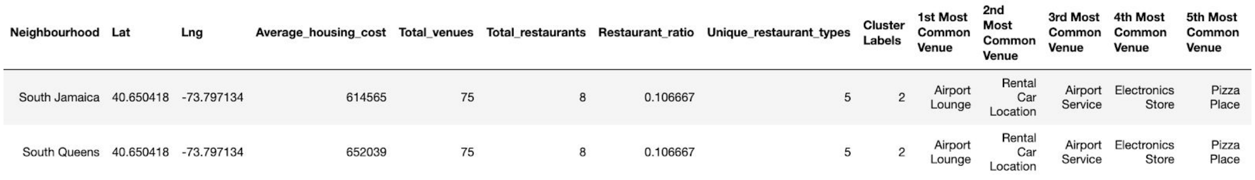
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Auburndale	40.761452	-73.789724	932072	100	37	0.370000	18	0	Korean Restaurant	Pizza Place	Ice Cream Shop	Café	Greek Restaurant
Central Queens	40.749824	-73.797634	546753	59	19	0.322034	10	0	Korean Restaurant	Pizza Place	Chinese Restaurant	Pharmacy	Bus Station
Flushing	40.765430	-73.817429	815690	100	47	0.470000	13	0	Korean Restaurant	Bubble Tea Shop	Bakery	Karaoke Bar	Chinese Restaurant
Hillcrest	40.749824	-73.797634	989715	59	19	0.322034	10	0	Korean Restaurant	Pizza Place	Chinese Restaurant	Pharmacy	Bus Station
Northeast Queens	40.749824	-73.797634	732875	59	19	0.322034	10	0	Korean Restaurant	Pizza Place	Chinese Restaurant	Pharmacy	Bus Station
Oakland Gardens	40.753991	-73.765966	388993	100	38	0.380000	19	0	Korean Restaurant	Coffee Shop	Bakery	Sandwich Place	Bar

Cluster 2



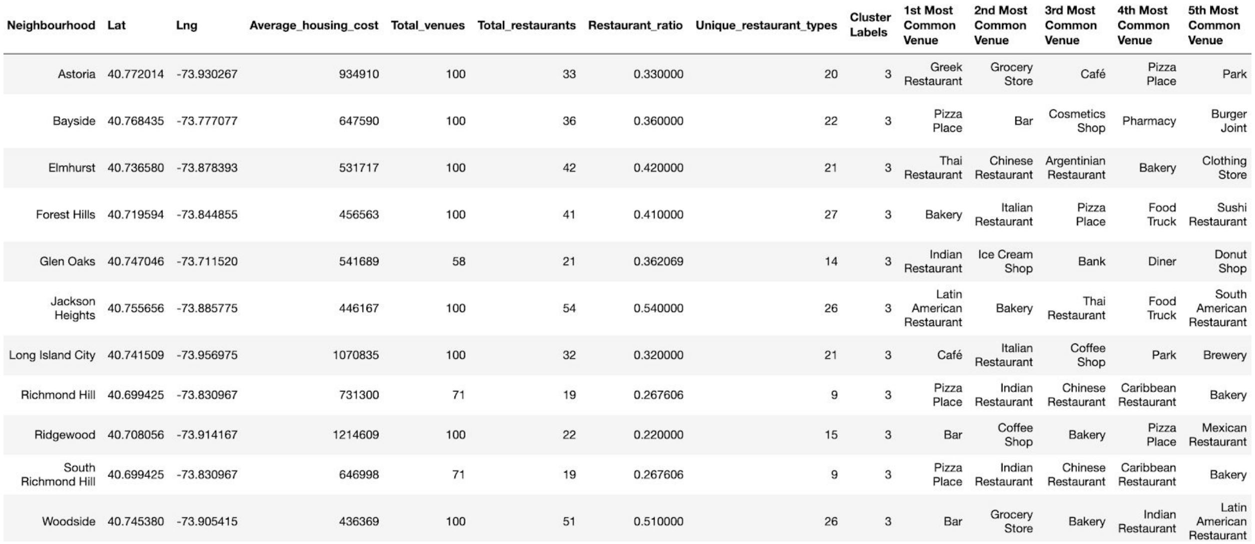
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bellerose	40.732778	-73.717778	704789	90	31	0.344444	10	1	Indian Restaurant	Mobile Phone Shop	Deli / Bodega	Grocery Store	Pizza Place
Briarwood	40.709256	-73.820139	351848	100	25	0.250000	15	1	Pizza Place	Donut Shop	Sandwich Place	Coffee Shop	Bank
Clearview	40.782778	-73.788611	621955	93	14	0.150538	11	1	Pizza Place	Donut Shop	Bank	Cosmetics Shop	Park
College Point	40.787601	-73.845968	881034	83	18	0.216867	13	1	Deli / Bodega	Donut Shop	Pizza Place	Asian Restaurant	Park
Douglaston	40.768713	-73.747077	1241430	71	25	0.352113	16	1	Chinese Restaurant	Italian Restaurant	Deli / Bodega	Spa	Bank
East Elmhurst	40.761212	-73.865136	901332	100	22	0.220000	13	1	Pizza Place	Latin American Restaurant	Airport Service	Fast Food Restaurant	Hotel

Cluster 3



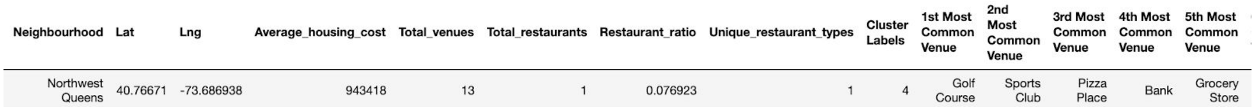
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
South Jamaica	40.650418	-73.797134	614565	75	8	0.106667	5	2	Airport Lounge	Rental Car Location	Airport Service	Electronics Store	Pizza Place
South Queens	40.650418	-73.797134	652039	75	8	0.106667	5	2	Airport Lounge	Rental Car Location	Airport Service	Electronics Store	Pizza Place

Cluster 4



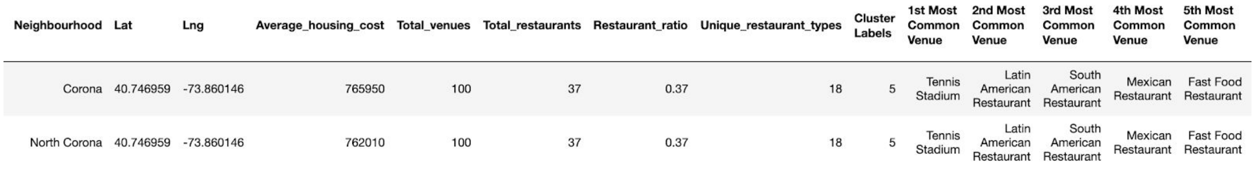
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Astoria	40.772014	-73.930267	934910	100	33	0.330000	20	3	Greek Restaurant	Grocery Store	Café	Pizza Place	Park
Bayside	40.768435	-73.777077	647590	100	36	0.360000	22	3	Pizza Place	Bar	Cosmetics Shop	Pharmacy	Burger Joint
Elmhurst	40.736580	-73.878393	531717	100	42	0.420000	21	3	Thai Restaurant	Chinese Restaurant	Argentinian Restaurant	Bakery	Clothing Store
Forest Hills	40.719594	-73.844855	456563	100	41	0.410000	27	3	Bakery	Italian Restaurant	Pizza Place	Food Truck	Sushi Restaurant
Glen Oaks	40.747046	-73.711520	541689	58	21	0.362069	14	3	Indian Restaurant	Ice Cream Shop	Bank	Diner	Donut Shop
Jackson Heights	40.755656	-73.885775	446167	100	54	0.540000	26	3	Latin American Restaurant	Bakery	Thai Restaurant	Food Truck	South American Restaurant
Long Island City	40.741509	-73.956975	1070835	100	32	0.320000	21	3	Café	Italian Restaurant	Coffee Shop	Park	Brewery
Richmond Hill	40.699425	-73.830967	731300	71	19	0.267606	9	3	Pizza Place	Indian Restaurant	Chinese Restaurant	Caribbean Restaurant	Bakery
Ridgewood	40.708056	-73.914167	1214609	100	22	0.220000	15	3	Bar	Coffee Shop	Bakery	Pizza Place	Mexican Restaurant
South Richmond Hill	40.699425	-73.830967	646998	71	19	0.267606	9	3	Pizza Place	Indian Restaurant	Chinese Restaurant	Caribbean Restaurant	Bakery
Woodside	40.745380	-73.905415	436369	100	51	0.510000	26	3	Bar	Grocery Store	Bakery	Indian Restaurant	Latin American Restaurant

Cluster 5



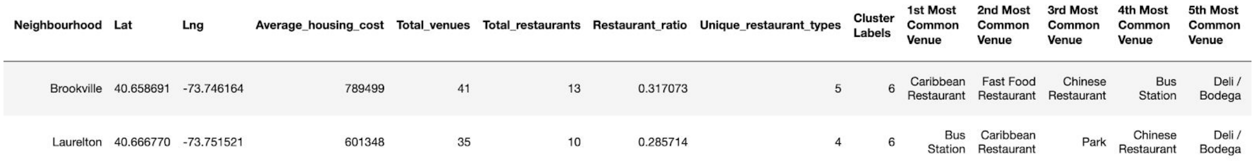
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Northwest Queens	40.76671	-73.686938	943418	13	1	0.076923	1	4	Golf Course	Sports Club	Pizza Place	Bank	Grocery Store

Cluster 6



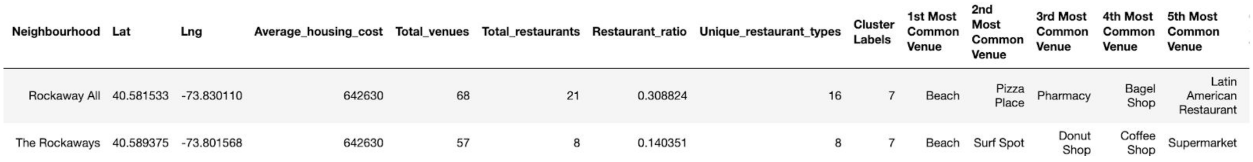
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Corona	40.746959	-73.860146	765950	100	37	0.37	18	5	Tennis Stadium	Latin American Restaurant	South American Restaurant	Mexican Restaurant	Fast Food Restaurant
North Corona	40.746959	-73.860146	762010	100	37	0.37	18	5	Tennis Stadium	Latin American Restaurant	South American Restaurant	Mexican Restaurant	Fast Food Restaurant

Cluster 7



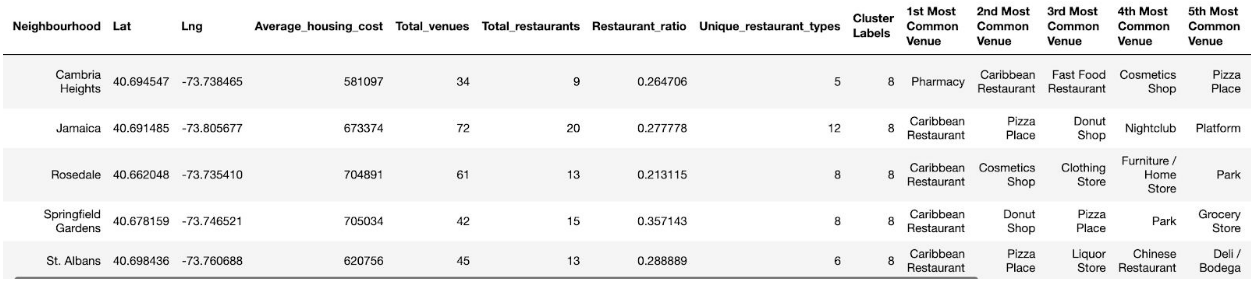
Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Brookville	40.658691	-73.746164	789499	41	13	0.317073	5	6	Caribbean Restaurant	Fast Food Restaurant	Chinese Restaurant	Bus Station	Deli / Bodega
Laurelton	40.666770	-73.751521	601348	35	10	0.285714	4	6	Bus Station	Caribbean Restaurant	Park	Chinese Restaurant	Deli / Bodega

Cluster 8



Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Rockaway All	40.581533	-73.830110	642630	68	21	0.308824	16	7	Beach	Pizza Place	Pharmacy	Bagel Shop	Latin American Restaurant
The Rockaways	40.589375	-73.801568	642630	57	8	0.140351	8	7	Beach	Surf Spot	Donut Shop	Coffee Shop	Supermarket

Cluster 9



Neighbourhood	Lat	Lng	Average_housing_cost	Total_venues	Total_restaurants	Restaurant_ratio	Unique_restaurant_types	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Cambria Heights	40.694547	-73.738465	581097	34	9	0.264706	5	8	Pharmacy	Caribbean Restaurant	Fast Food Restaurant	Cosmetics Shop	Pizza Place
Jamaica	40.691485	-73.805677	673374	72	20	0.277778	12	8	Caribbean Restaurant	Pizza Place	Donut Shop	Nightclub	Platform
Rosedale	40.662048	-73.735410	704891	61	13	0.213115	8	8	Caribbean Restaurant	Cosmetics Shop	Clothing Store	Furniture / Home Store	Park
Springfield Gardens	40.678159	-73.746521	705034	42	15	0.357143	8	8	Caribbean Restaurant	Donut Shop	Pizza Place	Park	Grocery Store
St. Albans	40.698436	-73.760688	620756	45	13	0.288889	6	8	Caribbean Restaurant	Pizza Place	Liquor Store	Chinese Restaurant	Deli / Bodega

8. Neighbourhoods - Preliminary results & discussion of the clusters

If we look at our clusters, they have clear trends.

- **Cluster 1:** These are predominantly Asian neighbourhoods, with their high number of Korean & Chinese restaurants, bubble tea shops, karaoke places…
- **Cluster 2:** This cluster has the least focus. It contains some restaurants, but also parks, banks and even train stations.
- **Cluster 3:** These neighbourhoods are situated in the vicinity of the airport and don't have many restaurants.
- **Cluster 4:** Greek, mexican, chinese, korean, burgers, bakeries, pizza... Even Middle Eastern cuisine. It has a wide variety of different restaurants.
- **Cluster 5:** This cluster can be summed up in 2 words: golf courses.
- **Cluster 6:** Where cluster 1 was the Asian neighbourhood, this cluster contains the Latino neighbourhoods. Mexican, Argentinian, South American… It has it all.
- **Cluster 7:** While this cluster focuses on restaurants, these restaurants mainly serve Caribbean food, Chinese food and fried chicken.
- **Cluster 8:** The neighbourhoods in this cluster are all rather close to the beach.

- Cluster 9: This cluster is quite similar to cluster 7, with a focus on Caribbean food. However, it also features a lot of non-restaurant venues.

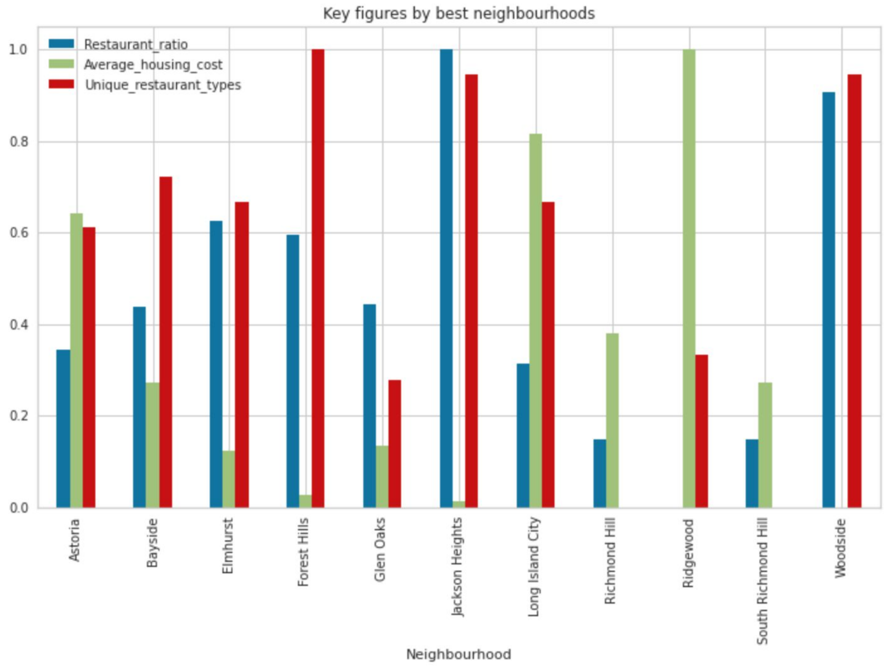
Considering that we want to set up our restaurant in a neighbourhood with a wide variety of cuisines, there's only one real option here. Cluster 4 offers a wide variety of cuisines and seems to contain great neighbourhoods with high restaurant ratios. For the next part of my methodology section, I'll zoom in a bit more on the neighbourhoods of cluster 4 to identify the best neighbourhood from the cluster.

9. Neighbourhoods - Methodology, part 2

9.1 Looking at the known metrics for each neighbourhood in the cluster

I've already gathered the needed data in the previous methodology segment. For clarity's sake I decided to filter out all neighbourhoods that didn't belong to cluster 4 and left out all columns that were except: Restaurant_ratio, Average_housing_cost and Unique_restaurant_types.

This data was then normalised through the min-max scaling method I used in section 3.1. The result was the following graph



10. Neighbourhoods - Results

I'd like to evaluate the 11 neighbourhoods of cluster 4 based on the 3 criteria in the graph.

When we look at the **restaurant ratio**, our top 5 neighbourhoods are:

1. Jackson Heights
2. Woodside
3. Elmhurst
4. Forest Hills
5. Glen Oaks

When we look at the **average housing costs**, our top 5 neighbourhoods are:

1. Woodside
2. Jackson Heights
3. Forest Hills
4. Elmhurst
5. Glen Oaks

When we look at the **unique restaurant types**, our top 5 neighbourhoods are:

1. Forest Hills
2. Jackson Heights = Woodside (tied)
3. Bayside
4. Long Island City

11. Neighbourhoods - Discussion, with a little extra

11.1 Discussing the neighbourhood results

Out of these top 5's, it's quite easy to generate a shortlist. Forest Hills, Jackson Heights & Woodside are the only neighbourhoods that are present in all 3.

I'd argue that Jackson Heights & Woodside would be a better pick than Forest Hills. The only time Forest Hills outperforms the other two neighbourhoods, is when we look at unique restaurant types. Even then, in this category, the difference between our 3 best neighbourhoods is quite small.

You can't go wrong with either Jackson Heights or Woodside. Although I would personally prefer Jackson Heights due to it's higher restaurant ratio and thus a better restaurant density. Real estate is only 2.25% more expensive than it is in Woodside, while the number of unique restaurant types is the exact same at 26. But that's not all!

When we have an in depth look at the exact type of restaurants in each neighbour hood we see the following.

Jackson Heights: Italian Restaurant, Restaurant, Latin American Restaurant, Peruvian Restaurant, Thai Restaurant, Empanada Restaurant, Argentinian Restaurant, Paella Restaurant, Cajun / Creole Restaurant, South American Restaurant, Indian Restaurant, Food Truck, Cuban Restaurant, Spanish Restaurant, Mexican Restaurant, Tibetan Restaurant, Arepa Restaurant, Burger Joint, Asian Restaurant, Colombian Restaurant, Vegetarian / Vegan Restaurant, Mediterranean Restaurant, Dumpling Restaurant, Vietnamese Restaurant, Greek Restaurant, Salvadoran Restaurant

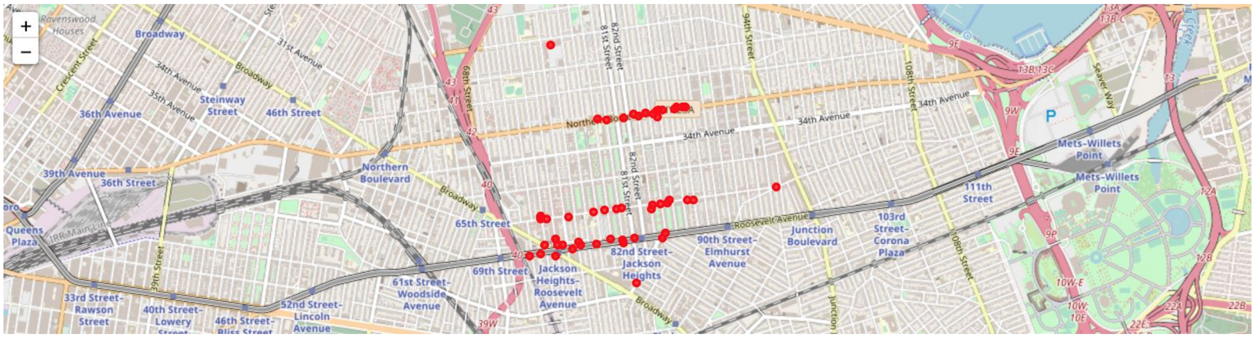
> These are mainly South American cuisines, with some Asian and even European mixed in. However, nothing is similar to Belgian!

Woodside: Chinese Restaurant, American Restaurant, Fast Food Restaurant, Thai Restaurant, Mexican Restaurant, Latin American Restaurant, Burger Joint, Himalayan Restaurant, Sushi Restaurant, Food Truck, Japanese Restaurant, Filipino Restaurant, Arepa Restaurant, Korean Restaurant, Tibetan Restaurant, French Restaurant, Indian Restaurant, Turkish Restaurant, Italian Restaurant, Fried Chicken Joint, Vietnamese Restaurant, Southern / Soul Food Restaurant, Health Food Store, Vegetarian / Vegan Restaurant, Peruvian Restaurant, Dumpling Restaurant

Here we see almost the opposite of Jackson Heights. Woodside has mostly Asian cuisines, with some South American and European. However, there's already a French restaurant, which is quite similar to Belgian.

11.2 Can we go even further with our current data?

Of course we can! We have our venue locations… So what if we were to plot out the location of each restaurant on a Folium map?



As you can see on the map, our restaurants are concentrated around 2 points. Those must be the culinary centres of Jackson Heights! But where oh where are they located? 2 restaurants at the centre of these clusters are 'Dela Mora' and 'Lali Guras Restaurant'.

Using geopy's geolocator.reverse function, we can look up the address based on the coordinates of the restaurants. The result? 84-19, Northern Boulevard & 76-02, 37th Road.

12. Neighbourhoods - Conclusion

All-in-all Queens has multiple neighbourhoods that you can't really go wrong with. Jackson Heights, Woodside and Forest Hills are all great options. However, when you look at the data, it does seem that Jackson Heights is the best location for a Belgian restaurant. More specifically, the areas around 84-19, Northern Boulevard & 76-02, 37th Road.

This is an area in a borough with a low crime rate, a high population with a good GDP per capita, where housing prices are on the low end and that's sprawling with restaurants of different cuisines. And luckily, all those different cuisines don't include Belgian or French! This means that it combines the perfect combination of traits to be a suitable location for a Belgian restaurant in New York City.