IBM Data Science Capstone Project, The Battle of New York’s Boroughs

1. Introduction.

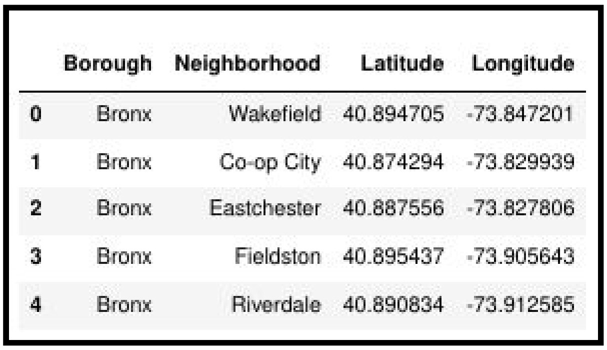
New York’s Boroughs analysis and comparison.
The objective of the project is to compare and establish differences, if they exist, between the five boroughs that form New York, the Bronx, Manhattan, Brooklyn, Queens, and Staten Island, focussing on the distribution and clustering of venues.

Background.
Urban development and the distribution of important city landmarks reveal the history of the city and its social, cultural, and economic development, and from its current composition it’s possible to define differences within and between its geographical and political divisions. From its conception as a world capital, the unique composition of New York’s boroughs presents interesting insights into multiculturalism and approach to urban development, and potential clues to it future and priorities.

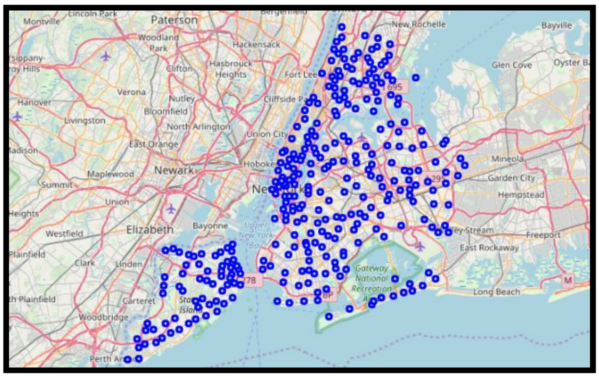
This type of analysis may be helpful to private enterprises if they want to establish the basis for a more specific and complete analysis on the possibilities for new franchises, a new brand, or service.The public sector can also gain insights into the commercial, and cultural distribution of the city, and define long term projects of urban development in function of its priorities and budget constraints.

2. Data.

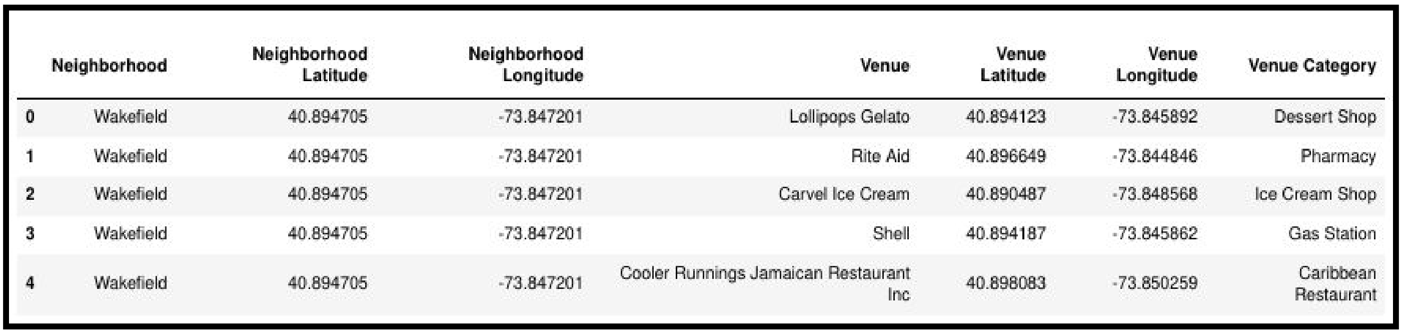
The data that will be used to analyse New York’s boroughs comes from the Spatial Data Repository, provided by the New York University Data Services, the data set presents the geographical distribution of New York’s neighborhoods as well as it political divisions in latitude and longitude forms. It will help to establish the base of the geographical organization of New York, and of the data that will be built on top of it.



	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



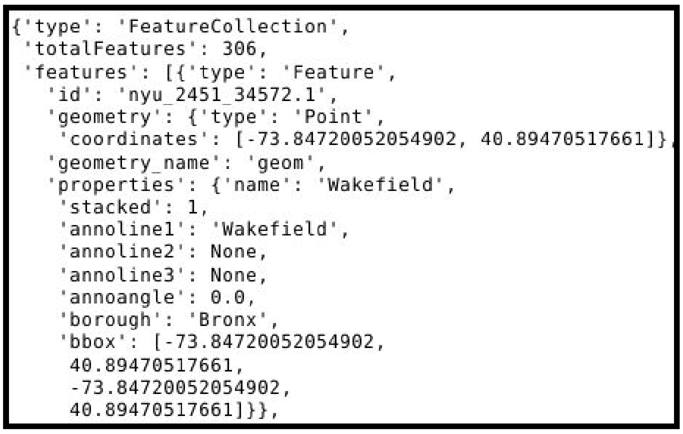
The second data set to be used is the venues information provided by Foursquare, presenting the geographical description of the explored venues around the centers of the neighborhoods in New York City. With it, we will be able to define the characteristics of the venues surrounding the geographical centers of the neighborhoods pointed in the previous dataset, the Foursquare dataset includes the coordinates of the venues as well as their categories, names, address, and country.



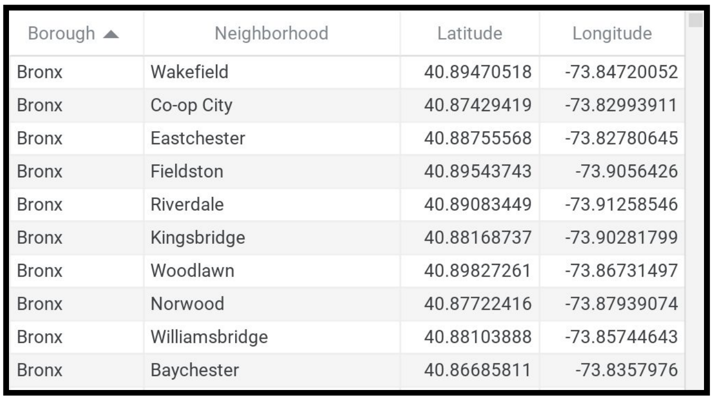
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
4	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant

3. **Exploratory data analysis.**

The first step into building the data set is to extract the geographical characteristic from the New York's neighborhood data set supplied by the New York University Data Services, the names of the five boroughs, its 306 neighborhoods, and matching coordinates are found under the features key, and presented under the *.json* format. The following tables present the raw format and the dataframe cleaned with the base information.



```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'type': 'Feature',
   'id': 'nyu_2451_34572.1',
   'geometry': {'type': 'Point',
    'coordinates': [-73.84720052054902, 40.89470517661]},
   'geometry_name': 'geom',
   'properties': {'name': 'Wakefield',
    'stacked': 1,
    'annoline1': 'Wakefield',
    'annoline2': None,
    'annoline3': None,
    'annoangle': 0.0,
    'borough': 'Bronx',
    'bbox': [-73.84720052054902,
     40.89470517661,
     -73.84720052054902,
     40.89470517661]}},
```

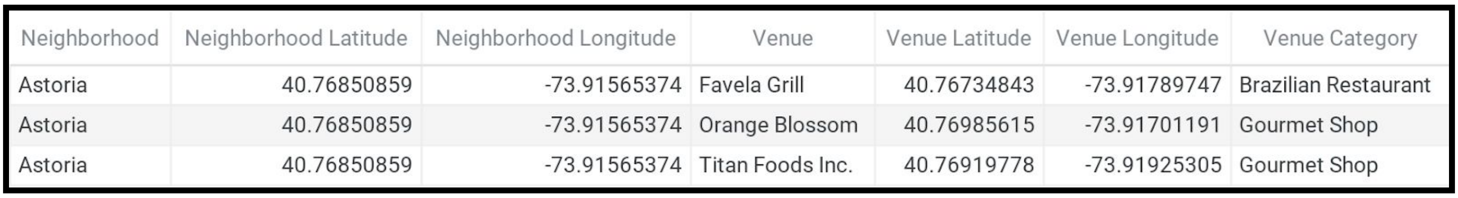


Borough ▲	Neighborhood	Latitude	Longitude
Bronx	Wakefield	40.89470518	-73.84720052
Bronx	Co-op City	40.87429419	-73.82993911
Bronx	Eastchester	40.88755568	-73.82780645
Bronx	Fieldston	40.89543743	-73.9056426
Bronx	Riverdale	40.89083449	-73.91258546
Bronx	Kingsbridge	40.88168737	-73.90281799
Bronx	Woodlawn	40.89827261	-73.86731497
Bronx	Norwood	40.87722416	-73.87939074
Bronx	Williamsbridge	40.88103888	-73.85744643
Bronx	Baychester	40.86685811	-73.8357976

The next step in the data processing stage is to split the aforementioned dataset into five sets each for every borough of the city (Bronx, Manhattan, Brooklyn, Queens, and Staten Island), with the purpose to analyze them independently and for ease of comparison, resulting in each of the boroughs mentioned having 52, 40, 70, 81, 73 neighborhoods respectively.

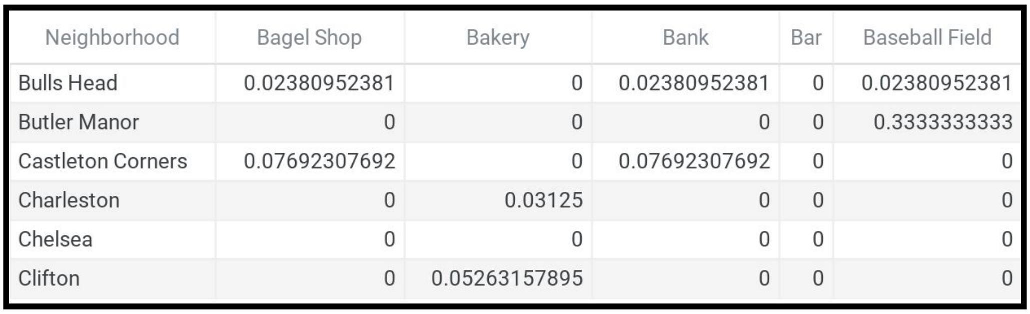
Using the Foursquare API, the next step is to explore for nearby venues in a radius of 500 meters from the geographical center of each neighborhood in their respective borough, the response from the Foursquare API creates a similar formatted output from which we extract the name of the venue, its latitude and longitude, and category associated with it.

The next step is to create five datasets corresponding to each borough, joining the neighborhood data indicated earlier with the venues data requested from Foursquare, and the already mentioned filtering process is applied to the venue data is used to extract the features indicated earlier. The result is a data set with 1230 venue entries for Bronx, 3309 for Manhattan, 2777 for Brooklyn, 2134 for Queens, and 832 for Staten Island, all the above have the structure shown below. It can also be noted that 172, 337, 280, 274, and 182 unique categories of venues were found in Bronx, Manhattan, Brooklyn, Queens, and Staten Island respectively.



Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Astoria	40.76850859	-73.91565374	Favela Grill	40.76734843	-73.91789747	Brazilian Restaurant
Astoria	40.76850859	-73.91565374	Orange Blossom	40.76985615	-73.91701191	Gourmet Shop
Astoria	40.76850859	-73.91565374	Titan Foods Inc.	40.76919778	-73.91925305	Gourmet Shop

With the objective to establish clusters of venues within each borough the last step of the data processing operation is to convert the categorical variables i.e. the category of all the venues into dummy variables that indicate the existence of a type of venue, and calculate its frequency for each neighborhood in their respective borough, this will allow us to use the k-means algorithm to group the neighborhoods according to the type, and quantity, of venues that that they have in common.

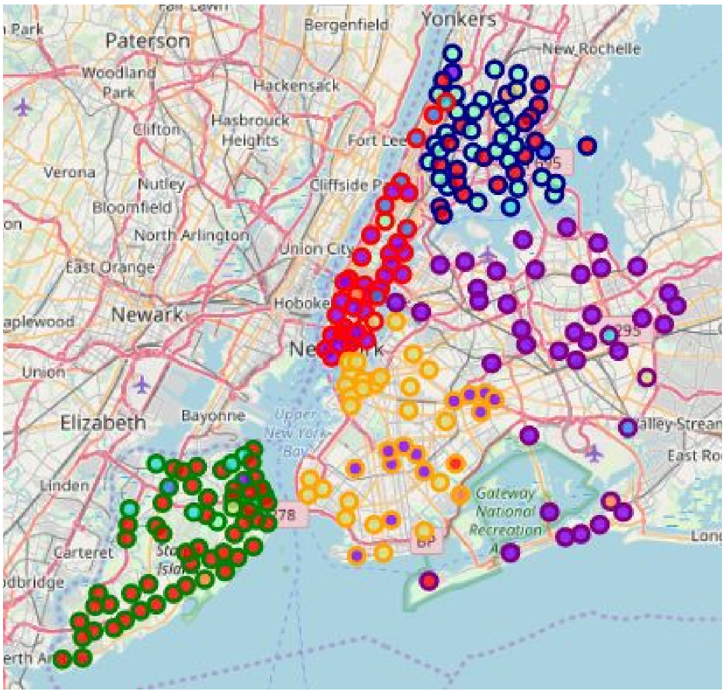


Neighborhood	Bagel Shop	Bakery	Bank	Bar	Baseball Field
Bulls Head	0.02380952381	0	0.02380952381	0	0.02380952381
Butler Manor	0	0	0	0	0.3333333333
Castleton Corners	0.07692307692	0	0.07692307692	0	0
Charleston	0	0.03125	0	0	0
Chelsea	0	0	0	0	0
Clifton	0	0.05263157895	0	0	0

The result is a set of five distinct data frames with the mean frequency for each type of venue found in the neighborhood.

4. Results.

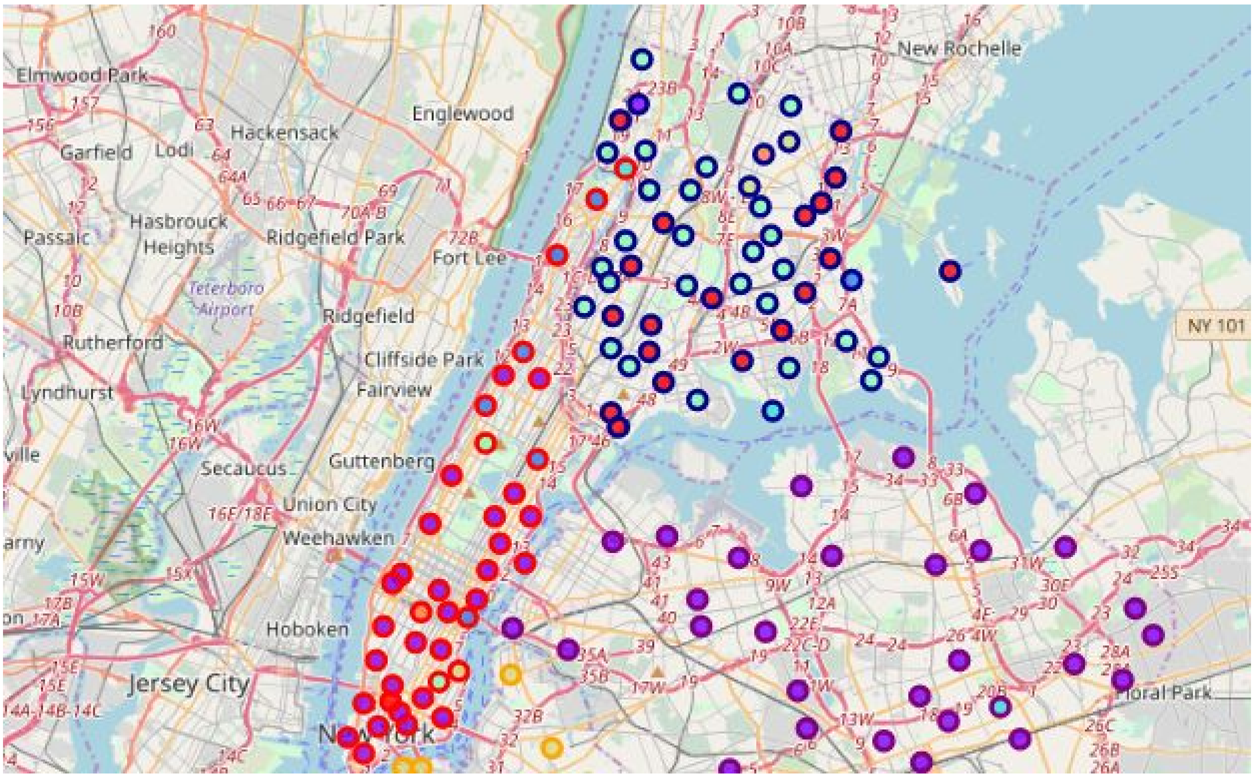
Using the K-means clustering algorithm on the five boroughs of New York, and defining seven as the maximum number of clusters to be identified for each borough, we will analyse and compare the boroughs to each other in terms of its differences and similarities.



Following some general observations on the characteristics of the boroughs and differences between each other, we will approach each borough independently and highlight its characteristics as well as particularities.

A characteristic, one that is more prevalent in each borough, is the amount of neighborhoods that fall in a reduced number of clusters, for each borough only two or three clusters have the majority of neighborhoods and the differences between the first and second biggest cluster are significant, only in the case of the Bronx and Brooklyn the difference between the two bigger clusters is small. One common observation between and within the boroughs, and clusters identified, is the prevalence of food service venues as the identifying factor for all the neighborhoods, and transport services, financial services and public amenities as the differentiating factors between clusters in each borough.

Bronx. Presents three main clusters, with the biggest, almost double the size as the second, 32 and 11 respectively. One identifying factor is the prevalence of banks as a type of identifying factor for the biggest cluster and bus stations for the second one. One element that differentiates these clusters is the absence, in both of them, of american restaurant type venue, in its place foods from the Caribbean and Latin America.

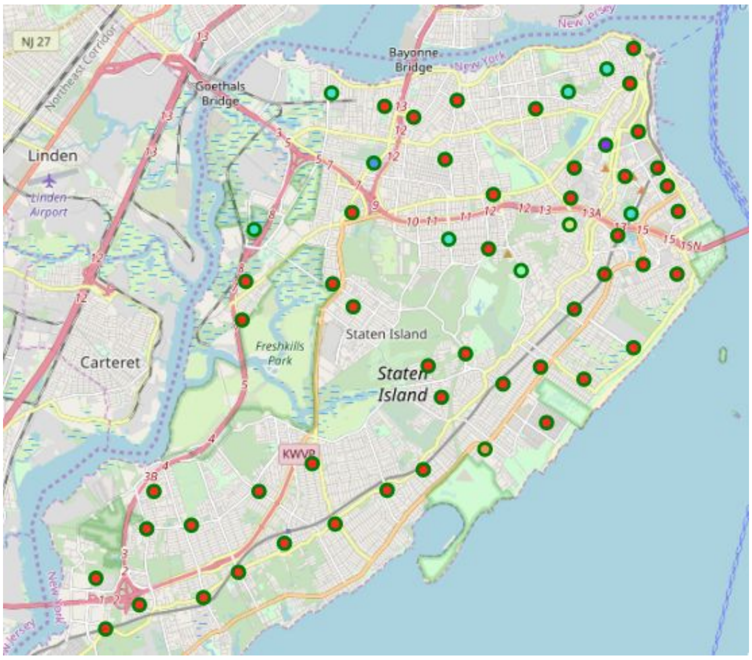


Manhattan. Most of the neighborhoods fall into one cluster with more than 20 neighborhoods, one important characteristic of the big cluster is that its most common venues are associated with food services. The second biggest also presents food a ubiquitous, and its defining factor is the existence of parks as a more common venue and mexican restaurants, in contrast with the italian and american restaurants that prevail in the bigger one.

Brooklyn. In Brooklyn exist two main clusters with similar number of neighbours, differentiated by the frequency of pharmacy type venues, coffee shops and restaurants present in each one, in general terms the main clusters are very similar in the broad categories that define each group of neighborhoods, Broolyn is the borough that presents a more homogeneous distribution in public services.

Queens. Queens is characterized by two main clusters, the biggest one accounting for more than 20 neighborhoods, although both clusters have a chinese restaurant as one of the most common venues, its differentiating factor is in the prevalence of bus stations in the second one, the frequency of the food venues is similar.

Staten Island. Its biggest cluster has more than 30 neighborhood associated with it, the second one has close to 15, the interesting aspect of this set of clusters is the similitude between them analysing its most common venues, the differences are in the number and type of financial services presented, one of the identifying factors of the bigger cluster is the existence of bus stations as one of the most common locations presented in the smaller of the two. Between boroughs Staten Island is the first to have a common venu Bus Stop



5. Discussion.

The biggest take away from the clustering of neighborhoods and comparison between New York's Boroughs is the existence of public amenities as one of the most common differentiator between clusters from a public interest perspective. The fact that public transportation is a differentiator in some cases may suggest a discrepancy in the offer of these services on the boroughs, and depending on the cluster that this category falls into, the service may be characterized by a surplus in its offer, or, from a demand perspective, there is a lack in the reach of the service and urban necessities may not be satisfied with the constant change in the infrastructure of the city, and shifting urban transformation.

Further studies may include a comparison between the geographical distribution of the clusters and the differentiating venues, in specific public services, some analysis may come from a purely efficiency public transportation approach, and the interconnectivity between the private and public sector may be an interesting focus.

6. Conclusion.

From its conception as a world capital, the unique composition of New York's boroughs presents interesting insights into multiculturalism and approach to urban development, and potential clues to it future and priorities.

New York's boroughs present a unique mixture of cultures and ethnicities that is presented throughout the culinary venues, its prevalence and diversity in origin establish the city as a welcoming and transformative capital not only for the United States but for the world, each borough has its own unique identity and its changing circumstances may present a challenge in the public sphere, and the expansion of public services is essential to its success.

7. Data sources.

1. 2014 New York City Neighborhood Names. New York Spatial Data Repository.
2. 2010 New York City Boroughs. New York Spatial Data Repository.
3. Foursquare API.