IBM/Coursera Capstone Project
Dec 2019
M.Sc. Jacobo Perez

Moving to New York City - IBM Applied Data Science

Introduction

This project helps customers to analyze moving destinations in NYC. Client wants to live nearby a subway station, so we obtain all stations and their coordinates and find out how expensive are their surroundings for the client to narrow the search. Within their selected cost cluster, we find which subway stations surroundings has minimum required venues, such as gym/restaurants/etc. We eliminate stations not fulfilling those minimum requirements and perform cluster analysis on the stations that do, based on venues that are desirable to the client, such as café, parks, sushi, etc. Client may also specify subway lines and boroughs to be eliminated from proposals.

This project may be of interest for clients when identifying reallocation areas. Global companies that offer temporary corporate housing to personnel may be interested on determining where to acquire properties. This work could also be included as a component of a real estate broker website to enhance customer experience, enabling advanced neighborhood selection.

Data description

Subway Stations: csv file from mta website.
http://web.mta.info/developers/data/nyct/subway/Stations.csv
Relevant columns: Stop Name, GTFS Latitude, GTFS Longitude, Daytime Routes, Borough.

Station ID	Complex ID	GTFS Stop ID	Division	Line	Stop Name	Borough	Daytime Routes	Structure	Lat	Lng	North Direction Label	South Direction Label
1	1	R01	BMT	Astoria	Astoria - Ditmars Blvd	Q	N W	Elevated	40.775036	-73.912034	NaN	Manhattan
2	2	R03	BMT	Astoria	Astoria Blvd	Q	N W	Elevated	40.770258	-73.917843	Ditmars Blvd	Manhattan
3	3	R04	BMT	Astoria	30 Av	Q	N W	Elevated	40.766779	-73.921479	Astoria - Ditmars Blvd	Manhattan
4	4	R05	BMT	Astoria	Broadway	Q	N W	Elevated	40.761820	-73.925508	Astoria - Ditmars Blvd	Manhattan
5	5	R06	BMT	Astoria	36 Av	Q	N W	Elevated	40.756804	-73.929575	Astoria - Ditmars Blvd	Manhattan

Venues, their prices and categories: Foursquare API.

Data usage: the process obtains New York City subway station and their coordinates. Then, for each station, it searches restaurants per price tag (1 to 4). It determines price distribution per station and performs cluster analysis on all stations per price level. After the client selects desired price level, the process obtains all venues of remaining subway stations, cleans data by grouping category synonyms (such as pharmacy and drugstore), filters venues by desired categories and performs a second cluster analysis, this time by desired categories.

Methodology

The process obtains New York City subway station and their coordinates. Then, for each station, it searches restaurants per price tag (1 to 4), so it is required to perform four queries per station. Given Foursquare license limitation, it is important to persist obtained information in csv files to avoid querying repeatedly for the same stations.

Price distribution extract

Station	Lat	Lng	Price1	Price2	Price3	Price4
Astoria_Q_Astoria - Ditmars Blvd	40.775036	-73.912034	0.508772	0.438596	0.017544	0.035088
Astoria_Q_Astoria Blvd	40.770258	-73.917843	0.250000	0.750000	0.000000	0.000000
Astoria_Q_30 Av	40.766779	-73.921479	0.656250	0.312500	0.000000	0.031250
Astoria_Q_Broadway	40.761820	-73.925508	0.500000	0.500000	0.000000	0.000000
Astoria_Q_36 Av	40.756804	-73.929575	0.545455	0.393939	0.060606	0.000000

The process then determines price distribution per station and performs K-means cluster analysis on all stations per price level. We selected two clusters (0 to 1) to differentiate affordable and expensive neighborhoods.

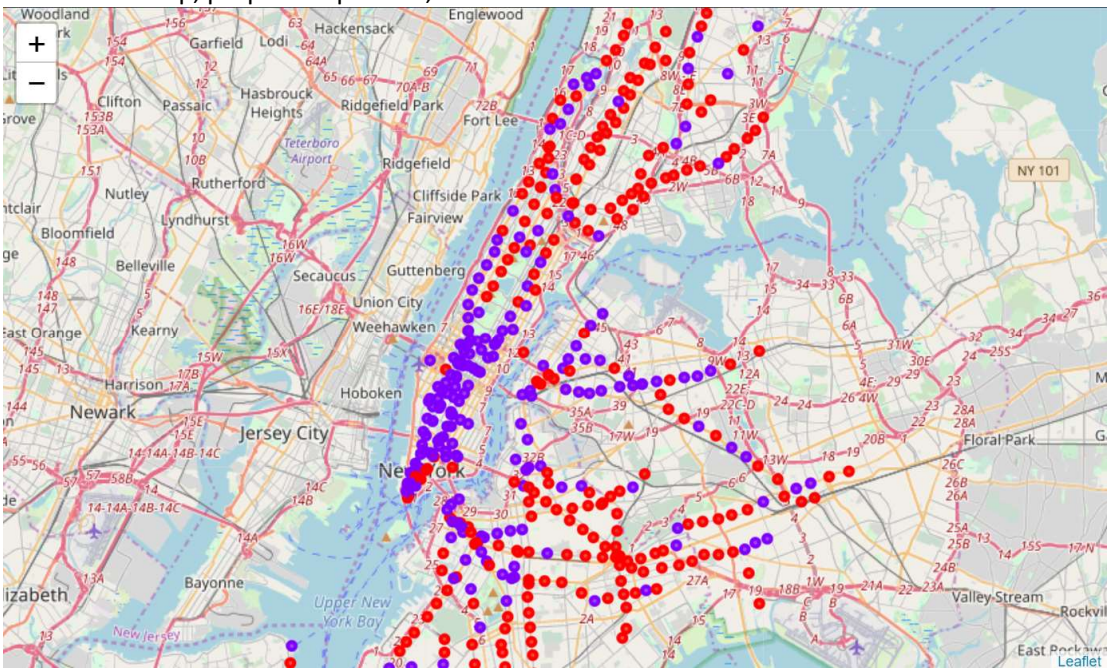
Price clustering extract, 0 being affordable and 1 being expensive.

Cluster	Station	Lat	Lng	Price1	Price2	Price3	Price4
1	Astoria_Q_Astoria - Ditmars Blvd	40.775036	-73.912034	0.508772	0.438596	0.017544	0.035088
1	Astoria_Q_Astoria Blvd	40.770258	-73.917843	0.250000	0.750000	0.000000	0.000000
0	Astoria_Q_30 Av	40.766779	-73.921479	0.656250	0.312500	0.000000	0.031250

Cluster means

Cluster	Price1 mean	Price2 mean	Price3 mean	Price4 mean
0	0.774299	0.198741	0.020302	0.006658
1	0.437000	0.430411	0.099300	0.033290

NYC cluster map, purple = expensive, red = affordable

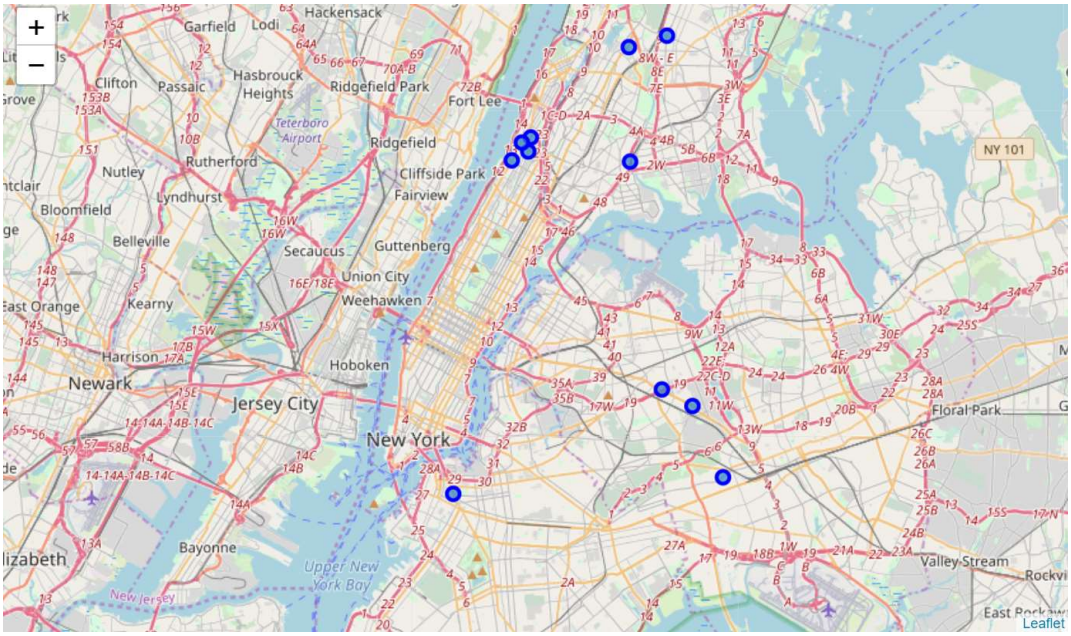


After the client selects the desired price level, the process eliminates undesired stations and obtains all venues of remaining ones, cleans data by grouping category synonyms (such as pharmacy and drugstore).

	Station	StationLat	StationLng	VenueLat	VenueLng	Venue	Category
0	Broadway_Bk_Court St	40.6941	-73.991777	Heatwise	40.693450	-73.991788	Yoga Studio
1	Broadway_Bk_Court St	40.6941	-73.991777	Brooklyn Historical Society	40.694942	-73.992333	History Museum
2	Broadway_Bk_Court St	40.6941	-73.991777	SoulCycle Brooklyn Heights	40.692253	-73.991042	Cycle Studio
3	Broadway_Bk_Court St	40.6941	-73.991777	Xtend Barre Brooklyn Heights	40.693599	-73.992376	Gym / Fitness Center
4	Broadway_Bk_Court St	40.6941	-73.991777	Orangetheory Fitness	40.693967	-73.991519	Gym

Then the process filters venues by desired categories, arriving to the finalist neighborhoods.

Station	Lat	Lng	Price1	Price2	Price3	Price4
Broadway - Brighton_Bk_DeKalb Av	40.690635	-73.981824	0.634615	0.307692	0.057692	0.0
Jamaica_Q_111 St	40.697418	-73.836345	0.714286	0.285714	0.000000	0.0
8th Av - Fulton St_M_163 St - Amsterdam Av	40.836013	-73.939892	0.714286	0.285714	0.000000	0.0
8th Av - Fulton St_M_155 St	40.830518	-73.941514	0.666667	0.333333	0.000000	0.0
Rockaway_Q_Beach 105 St	40.583209	-73.827559	1.000000	0.000000	0.000000	0.0
Concourse_Bx_Bedford Park Blvd	40.873244	-73.887138	0.809524	0.190476	0.000000	0.0
Queens Blvd_Q_67 Av	40.726523	-73.852719	0.653846	0.307692	0.038462	0.0
Queens Blvd_Q_Woodhaven Blvd	40.733106	-73.869229	0.684211	0.263158	0.052632	0.0
Broadway - 7Av_M_157 St	40.834041	-73.944890	0.785714	0.214286	0.000000	0.0
Broadway - 7Av_M_145 St	40.826551	-73.950360	0.761905	0.238095	0.000000	0.0
Pelham_Bx_Whitlock Av	40.826525	-73.886283	0.666667	0.333333	0.000000	0.0
Lenox - White Plains Rd_Bx_Gun Hill Rd	40.877850	-73.866256	0.888889	0.000000	0.111111	0.0



The process determines the distribution of desired categories.

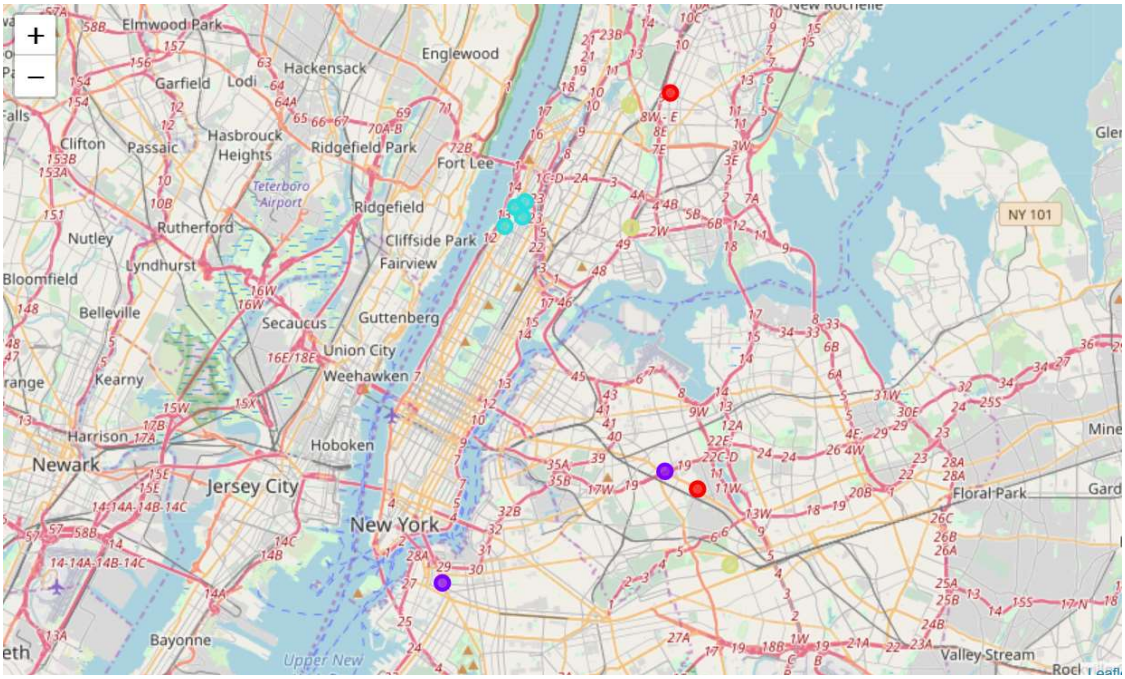
	Station	Café	Dance Studio	Deli / Bodega	Gym	Japanese Restaurant	Laundromat	Market	Park	Pharmacy	Spa
0	8th Av - Fulton St_M_155 St	0.277778	0.055556	0.277778	0.055556	0.000000	0.111111	0.111111	0.111111	0.000000	0.000000
1	8th Av - Fulton St_M_163 St - Amsterdam Av	0.357143	0.000000	0.214286	0.071429	0.000000	0.071429	0.071429	0.142857	0.000000	0.071429
2	Broadway - 7Av_M_145 St	0.291667	0.041667	0.250000	0.083333	0.125000	0.041667	0.041667	0.125000	0.000000	0.000000
3	Broadway - 7Av_M_157 St	0.187500	0.062500	0.312500	0.062500	0.000000	0.062500	0.062500	0.250000	0.000000	0.000000
4	Broadway - Brighton_Bk_DeKalb Av	0.384615	0.153846	0.000000	0.076923	0.076923	0.076923	0.076923	0.076923	0.000000	0.076923
5	Concourse_Bx_Bedford Park Blvd	0.117647	0.000000	0.117647	0.117647	0.000000	0.058824	0.235294	0.176471	0.176471	0.000000
6	Jamaica_Q_111 St	0.000000	0.000000	0.375000	0.125000	0.000000	0.125000	0.125000	0.125000	0.125000	0.000000
7	Lenox - White Plains Rd_Bx_Gun Hill Rd	0.000000	0.000000	0.000000	0.333333	0.000000	0.111111	0.222222	0.000000	0.222222	0.111111
8	Pelham_Bx_Whitlock Av	0.000000	0.000000	0.111111	0.111111	0.000000	0.111111	0.111111	0.222222	0.333333	0.000000
9	Queens Blvd_Q_67 Av	0.050000	0.000000	0.050000	0.150000	0.200000	0.100000	0.100000	0.150000	0.100000	0.100000
10	Queens Blvd_Q_Woodhaven Blvd	0.375000	0.000000	0.125000	0.125000	0.187500	0.062500	0.062500	0.000000	0.062500	0.000000
11	Rockaway_Q_Beach 105 St	0.000000	0.000000	0.250000	0.125000	0.000000	0.125000	0.250000	0.000000	0.250000	0.000000

We obtain the most common venue categories per subway station for supporting client neighborhood selection.

	Station	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
0	8th Av - Fulton St_M_155 St	Deli / Bodega	Café	Park	Market	Laundromat	Gym	Dance Studio
1	8th Av - Fulton St_M_163 St - Amsterdam Av	Café	Deli / Bodega	Park	Spa	Market	Laundromat	Gym
2	Broadway - 7Av_M_145 St	Café	Deli / Bodega	Park	Japanese Restaurant	Gym	Market	Laundromat
3	Broadway - 7Av_M_157 St	Deli / Bodega	Park	Café	Market	Laundromat	Gym	Dance Studio
4	Broadway - Brighton_Bk_DeKalb Av	Café	Dance Studio	Spa	Park	Market	Laundromat	Japanese Restaurant
5	Concourse_Bx_Bedford Park Blvd	Market	Pharmacy	Park	Gym	Deli / Bodega	Café	Laundromat
6	Jamaica_Q_111 St	Deli / Bodega	Pharmacy	Park	Market	Laundromat	Gym	Spa
7	Lenox - White Plains Rd_Bx_Gun Hill Rd	Gym	Pharmacy	Market	Spa	Laundromat	Park	Japanese Restaurant
8	Pelham_Bx_Whitlock Av	Pharmacy	Park	Market	Laundromat	Gym	Deli / Bodega	Spa
9	Queens Blvd_Q_67 Av	Japanese Restaurant	Park	Gym	Spa	Pharmacy	Market	Laundromat
10	Queens Blvd_Q_Woodhaven Blvd	Café	Japanese Restaurant	Gym	Deli / Bodega	Pharmacy	Market	Laundromat
11	Rockaway_Q_Beach 105 St	Pharmacy	Market	Deli / Bodega	Laundromat	Gym	Spa	Park

We perform a second K-means cluster analysis, this time by desired categories. We chose four clusters as results were more intuitive to interpret compared to other choices.

Station	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	Cluster
Lenox - White Plains Rd_Bx_Gun Hill Rd	Gym	Pharmacy	Market	Spa	Laundromat	Park	Japanese Restaurant	0
Queens Blvd_Q_67 Av	Japanese Restaurant	Park	Gym	Spa	Pharmacy	Market	Laundromat	0
Broadway - Brighton_Bk_DeKalb Av	Café	Dance Studio	Spa	Park	Market	Laundromat	Japanese Restaurant	1
Queens Blvd_Q_Woodhaven Blvd	Café	Japanese Restaurant	Gym	Deli / Bodega	Pharmacy	Market	Laundromat	1
8th Av - Fulton St_M_155 St	Deli / Bodega	Café	Park	Market	Laundromat	Gym	Dance Studio	2
8th Av - Fulton St_M_163 St - Amsterdam Av	Café	Deli / Bodega	Park	Spa	Market	Laundromat	Gym	2
Broadway - 7Av_M_145 St	Café	Deli / Bodega	Park	Japanese Restaurant	Gym	Market	Laundromat	2
Broadway - 7Av_M_157 St	Deli / Bodega	Park	Café	Market	Laundromat	Gym	Dance Studio	2
Concourse_Bx_Bedford Park Blvd	Market	Pharmacy	Park	Gym	Deli / Bodega	Café	Laundromat	3
Jamaica_Q_111 St	Deli / Bodega	Pharmacy	Park	Market	Laundromat	Gym	Spa	3
Pelham_Bx_Whitlock Av	Pharmacy	Park	Market	Laundromat	Gym	Deli / Bodega	Spa	3
Rockaway_Q_Beach 105 St	Pharmacy	Market	Deli / Bodega	Laundromat	Gym	Spa	Park	3



Results

As expected, most expensive neighborhoods are located in Manhattan. There are affordable areas elsewhere. Cluster analysis was very helpful for identifying areas with similar venues, in this particular case:

```
Cluster 0: Gym, spa.
Cluster 1: Café.
Cluster 2: Deli, café, park.
Cluster 3: Pharmacy, market, park.
```

This program is fully customizable to different client venue needs, and could be implanted as a functionality inside real estate websites.

Process performance was fast; the only limitation is Foursquare license, so the full venue set was obtained querying in different days. For commercial implementation we would require a professional license.

Further developments / recommendations

For making this process more robust, we identified the following opportunity areas:

- Obtain available housing prices for sale and rental.
- Obtain criminality statistics per neighborhood.
- Obtain commuting estimates from each subway station to work at rush hours by different transportation means.

Conclusions

- Machine learning techniques, in this case cluster analysis, is very helpful for identifying data trends.
- Traditional programming would involve setting custom business rules, which may miss relevant data relationships.
- Developing models in Python is very productive, as the code is readable and powerful, as there are plenty of functions and libraries in place for the programmer to use, so we can focus in the functionality.
- Foursquare API is fast and very useful, although our license was very limited.
- Pandas dataframes are very powerful, fulfilling all data functionalities we can think of, even obtaining tables from the internet.
- Obtaining data and its preparation is hard, as information sometimes is not available or requires pre-processing.