IBM Data Science Professional Certificate

Applied Data Science Capstone

New York City Metropolitan Transportation Authority (MTA):

Find best venues categories near metro entrances

Eleonora Picca

02 - 2019

Contents

Chapter 1

Business Problem and stakeholders

1.1 Business Problem and Stakeholders

New York City is the third most congested city in the world in terms of traffic and the second worst in the United States. In 2017, New York drivers averaged 91 peak hours stuck in traffic and spent 13 percent of their time sitting in congestion, with 11 percent of that being attributed to daytime traffic.

The reasons behind traffic problems are various, but one of them is that lots of New Yorkers choose to go around city by car.

NYC Metropolitan Transportation Authority (MTA) needs a strategy to alleviating traffic. Furthermore, if more people use public transportation, gain would raise and it could be spent to pay for a major overhaul of the aging subway.

One way to do this is to make New Yorkers use the subway instead of other road-kind transportation.

One solution could be to make agreements with venues near metro entrances in order to give special offers to New Yorkers who use the subway. What is proposed and presented in this report is how to choose the venues categories:

- Venues has to be near the metro entrance, in this case we chose a radius of 500 m.
- Best venues are the one more popular, because they have more appeal.

Chapter 2

Data Sets

In this chapter will be described data sets used for the analysis.

2.1 Kaggle Metro Entrances

The first data set used is the list of metro entrance in NYC (that can be found here: https://www.kaggle.com/parichartpanichpol/nyc-subway-entrances).

Data set is composed by the following columns in tab. (2.1):

ColumnName	*Description*	*Example*
OBJECTID	Unique ID	1734
URL	URL to the subway line information	http://web.mta.info/nyct/service
NAME	Entrance name	Birchall Ave and Sagamore St at NW corner
the geom	Longitude and Latitude	POINT (-73.86835600032798 40.84916900104506)
LINE	Line number	2-5

Table 2.1: Original Kaggle data set.

These information will be used to identify the location of the entrances. Latitudes and longitudes will be used for calling the Foursquare APIs.

- For limited budget and computational resources, only the first 400 entrances will be used for the analysis.
- Data preparation will be performed and latitudes and longitudes will have dedicated columns. Some rows don't have coordinates information, so they will be dropped.
- We have some duplicates on NAME column. Duplicates will be removed (with the take first technique).

Original Column	*Final Column*
NAME	Subway Entrance
the geom	Latitude
the geom	Longitude

Table 2.2: Initial and final columns of Kaggle data set.

	OBJECTID	URL	NAME	the_geom	LINE
0	1734	http://web.mta.info/nyct/service/	Birchall Ave & Sagamore St at NW corner	POINT (-73.86835600032798 40.84916900104506)	2-5
1	1735	http://web.mta.info/nyct/service/	Birchall Ave & Sagamore St at NE corner	POINT (-73.86821300022677 40.84912800131844)	2-5
2	1736	http://web.mta.info/nyct/service/	Morris Park Ave & 180th St at NW corner	POINT (-73.87349900050798 40.84122300105249)	2-5
3	1737	http://web.mta.info/nyct/service/	Morris Park Ave & 180th St at NW corner	POINT (-73.8728919997833 40.84145300067447)	2-5
4	1738	http://web.mta.info/nyct/service/	Boston Rd & 178th St at SW corner	POINT (-73.87962300013866 40.84081500075867)	2-5

(a)

	Subway Entrance	Latitude	Longitude
0	Birchall Ave & Sagamore St at NW corner	40.849169	-73.868356
1	Birchall Ave & Sagamore St at NE corner	40.849128	-73.868213
2	Morris Park Ave & 180th St at NW corner	40.841223	-73.873499
4	Boston Rd & 178th St at SW corner	40.840815	-73.879623
5	Boston Rd & E Tremont Ave at NW corner	40.840434	-73.880005

(b)

Figure 2.1: Kaggle dataset before (a) and after (b) data transformation.

The columns used will be the one in table (2.2).
In fig. (2.1) an example of initial and final Kaggle dataset.
Having the coordinates of the metro entrance, in fig. (2.2) the map of the 400 subway entrances selected.

2.2 Foursquare APIs

In addition to the subway entrances locations, we will use the Foursquare explore API in order to extract a list of recommended venues near the given location. The relevant parameters of the API are:

- Radius: Radius to search within, in meters. We will use 500 m.
- Limit: Number of results to return. We will use 100.
- Latitude and Longitude: We will used the ones in the Kaggle dataset.

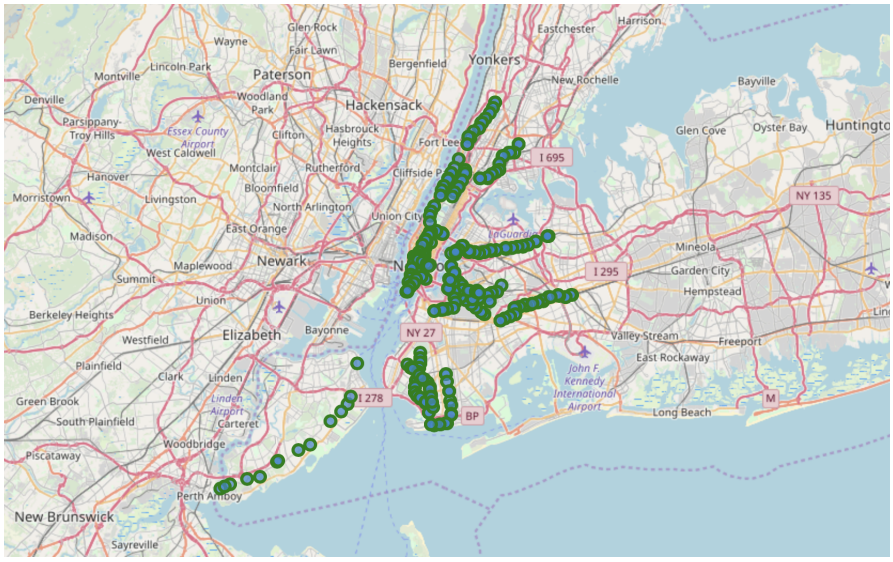


Figure 2.2: 400 subway entrances selected from Kaggle data set.

Subway Entrance	Subway Entrance Latitude	Subway Entrance Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Birchall Ave & Sagamore St at NW corner	40.849169	-73.868356	Bronx Park East	40.849164	-73.868453	Park
Birchall Ave & Sagamore St at NW corner	40.849169	-73.868356	Park Billiards	40.850970	-73.867792	Pool Hall
Birchall Ave & Sagamore St at NW corner	40.849169	-73.868356	Morris Park Pizza	40.844962	-73.867606	Pizza Place
Birchall Ave & Sagamore St at NW corner	40.849169	-73.868356	New Morris Deli	40.846529	-73.863874	Deli / Bodega
Birchall Ave & Sagamore St at NW corner	40.849169	-73.868356	Cafe Colonial	40.852495	-73.867654	Spanish Restaurant

Figure 2.3: Kaggle + Foursquare dataset with the needed information.

The API returns a list of venues with several information, but we will use:

- Venues categories.
- Venues latitudes and longitudes.

In fig. (2.3) an example of the data extracted using the explore Foursquare APIs, with only seven venues per entrance.

Combining the two data sets we will have a list of 100 recommended venues within 500 meters for each of the 400 subway entrances and we will find the 10 most common categories per each entrance (as shown in fig. (2.4)).

We will use these venues categories to create clusters of entrances in order to identify the most common venues categories.

Subway Entrance	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
103rd St & Roosevelt Ave at NE corner	Latin American Restaurant	Mexican Restaurant	Deli / Bodega	Pizza Place	Pharmacy	Food Truck	South American Restaurant	Bank	Gym	Coffee Shop
103rd St & Roosevelt Ave at SE corner	Latin American Restaurant	Mexican Restaurant	Deli / Bodega	Pizza Place	Pharmacy	Food Truck	South American Restaurant	Bank	Gym	Coffee Shop
104th St & Jamaica Ave at NE corner	Peruvian Restaurant	Deli / Bodega	Chinese Restaurant	Liquor Store	Bakery	Moving Target	Lounge	Bookstore	Park	Bagel Shop
104th St & Jamaica Ave at SE corner	Peruvian Restaurant	Deli / Bodega	Chinese Restaurant	Pharmacy	Convenience Store	Mexican Restaurant	Metro Station	Park	Bar	Market
10th Ave & 207th St at NW corner	Lounge	Pizza Place	Café	Grocery Store	Mexican Restaurant	Juice Bar	Nightclub	Food Truck	Bank	Bakery

Figure 2.4: Final dataset with 10 most common venues.

Chapter 3

Methodology

In this chapter I will described the methodology I used fro the analysis.

After performing additional data transformation, I used K-means for clusterization of the subway entrances based on the categories of the nearby venues, choosing the best K with the elbow method.

I decided to use K-means to cluster data because it is an unsupervised algorithm, and I don't have labeled data.

3.1 K-means data preparation

Data sets used and data preparation are described in chapter (2).

Starting from the data set with the 10 most common venues as in fig. (2.4), we performed one hot enconding and mean grouped by the venue entrance, because the k-means algorithm doesn't support categorical variables. I also normalized data set over the standard deviation (using *StandardScaler*) to help the algorithm to better interpret features with different magnitudes and distributions. In fig. (3.1) the final data set used to run the algorithm.

Accessories Store	Adult Boutique	African Restaurant	Airport Lounge	Airport Tram	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	...	Watch Shop	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop
0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	0.0	0.0	0.0	0.028571	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.028571	0.014286

Figure 3.1: Example of a few records from the final data set used for the K-means algorithm.

3.2 Choose k: elbow method and silhouette analysis

In the K-means algorithm, the number of clusters has to be chosen. With increasing the number of clusters, the distance of centroids to data points will always reduce. This means increasing K will always decrease the error.

Two methods I used to choose the best k were the elbow method and the silhouette analysis.

- In the elbow method the value of the metric is plotted as a function of K and the elbow point is determined, i.e. where the rate of decrease sharply shifts. It is the right K for clustering. For the elbow method I used as the metric the inertia that is the sum of distances of samples to their closest centroid.
- The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

In fig. (3.2) the result obtained for the two analysis. For the elbow method, there is not a clear k to choose, because there is not a clear elbow. Combining the silhouette I choose:

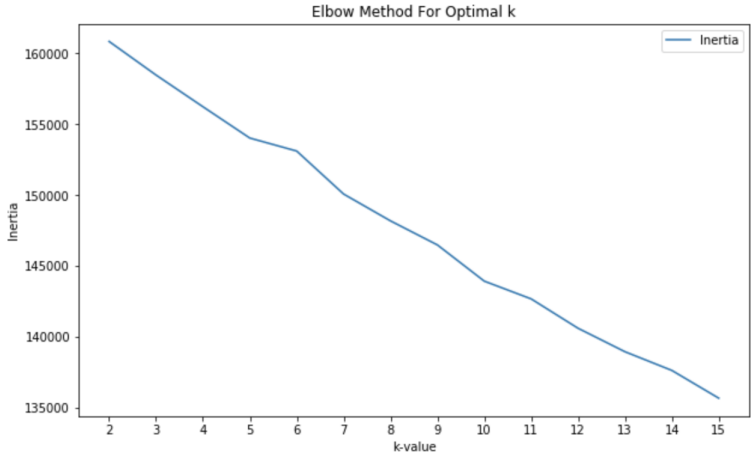
$$k = 13 \tag{3.1}$$

because in the silhouette analysis it was the k with the major number of non negative clusters. As can be seen, cluster 5, 4 and 3 have negative silhouette, so I decided not to consider them in the results sections.

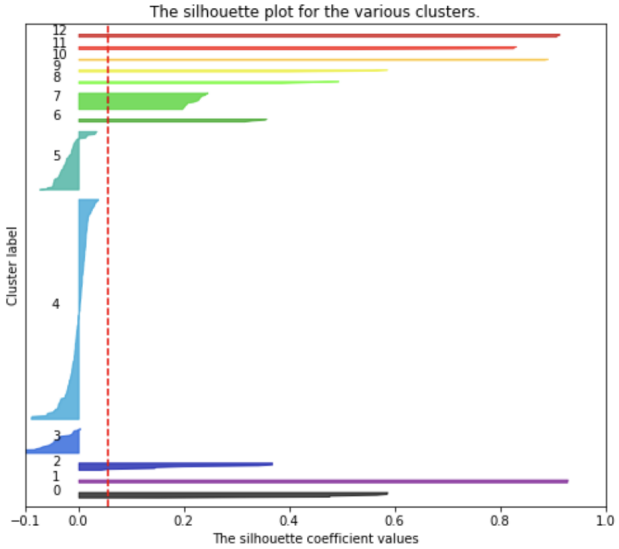
3.3 Metro entrance clusterization based on Venues Categories: K-means

Using 9 clusters, I used the *sklearn.cluster.KMeans* library to fit the data, with the following parameters:

- init= k-means++
- n_init= 20
- n_clusters= 13
- random_state= 10

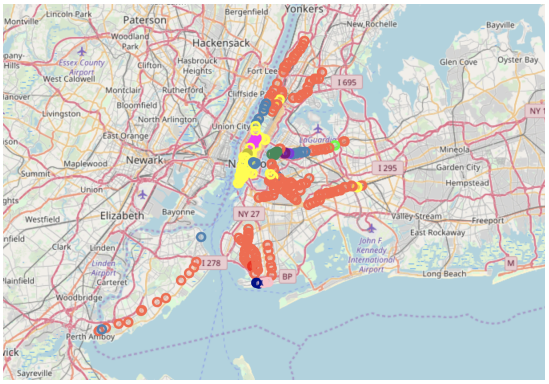


(a)



(b)

Figure 3.2: Elbow method using inertia (a) and silhouette for k=13 (b).



(a)

(b)

Figure 3.3: Map of the subway entrances splitted in the final 13 clusters (a) and the final clusters considered (b).

- algorithm = elkan

in fig. (3.3) the map of the clusters obtained. As mentioned in sec. (3.2), I discarded clusters 3,4,5. In chapter (4) I will discuss the results.

Chapter 4

Results and recommendations

In this chapter I will comment the final results obtained and make suggestions based on them.

4.1 Subway Entrance cluster characterizations

I extracted the most common venue types, as in tab. (4.1). By looking at the venues, the clusters seem characterized:

- By looking at the map in fig. (3.3(b)), clusters 0 and 9 are both near the beach (Coney Island). Cluster 0 has as most common venues restaurant and beach, while cluster 9 has theme park. It seems reasonable, also by double check on the internet about the area.
- Cluster 2 has lots of coffee shops and Italian restaurants.
- Cluster 6 is the fitness cluster: it has lots of gyms and Yoga Studio.
- Cluster 7 has theaters and hotels. By looking at the map, it is in Broadway near the Theater District.
- Cluster 8 has lots of Japanese restaurants.
- Cluster 10 is near the Arthur Ashe Stadium where the tennis US Open takes place, in fact it has tennis stadium has principal attraction.
- Cluster 11 has pizza places.
- Cluster 12 is the sweet cluster: donut shops and bakeries.

Cluster	*Venue Position*	*Venue Type*	*Counts*	*Total*
0	1st Most Common Venue	Restaurant; Beach	3; 3	7
0	2nd Most Common Venue	Gourmet Shop; Eastern European Restaurant	3; 2	7
1	1st Most Common Venue	Cantonese Restaurant	4	4
1	2nd Most Common Venue	Sushi Restaurant	4	4
2	1st Most Common Venue	Coffee Shop; Italian Restaurant	4; 4	9
2	2nd Most Common Venue	Cafe; Donut Shop	5; 1	9
6	1st Most Common Venue	Yoga Studio; Gym / Fitness Center	2; 2	4
6	2nd Most Common Venue	Gym; Yoga Studio	2; 1	4
7	1st Most Common Venue	Theater; Hotel	15; 4	19
7	2nd Most Common Venue	Hotel; Coffee Shop	8; 4	19
8	1st Most Common Venue	Japanese Restaurant	3	3
8	2nd Most Common Venue	Spa; Coffee Shop	1; 1	3
9	1st Most Common Venue	Theme Park Ride / Attraction	3	3
9	2nd Most Common Venue	Theme Park	3	3
10	1st Most Common Venue	Tennis Stadium	2	2
10	2nd Most Common Venue	Bar; Burger Joint	1; 1	2
11	1st Most Common Venue	Pizza Place	4	4
11	2nd Most Common Venue	Bar; Coffee Shop	2; 2	4
12	1st Most Common Venue	Donut Shop	4	4
12	2nd Most Common Venue	Bakery	4	4

Table 4.1: Final cluster 1st and 2nd most common venue with the relative counts (separated by ;).

4.2 MTA Recommendations

Given results in sec. (4.1), we have characterizations of a subset of NYC subway entrances based on what types of venues are around. So for different clusters, MTA could make agreements with those king of venues in order to make New Yorkers use the subway.

Some suggestions:

- 10% off theatre tickets near the Theatre District.
- 20% off on a ticket at the US Open or sale on a burger.
- A free beer on the beach or a free attraction at the Theme Park in Coney Island.
- Make arrangements with gyms on the 7th avenue for sales to the subway users.

It is obvious that commercial agreements have to be taken into account.

Chapter 5

Conclusion

From the k-means algorithm, the result obtained is not so accurate (I discarded a lots of data from the final results). This could be because the matrix used is a sparse matrix (as can be seen in fig. (3.1)). There could be different strategies to take into account:

- Reduce the redundant categories as extracted from Foursquare (for example, *Bar* and *Coffee Shop* could be one category
- Use a different clustering algorithm more suitable for sparse matrix.

The results obtained could be used the MTA, the only problem is the commercial agreements that should be discussed for the sales.