Find best venues categories near metro entrances

IBM Data Science Professional Certificate

Applied Data Science Capstone

Eleonora Picca

Summary

- Business Problem and Stakeholders
- Data Set
- Overall Strategy
- Data Preparation
- Choose best k
- K-means clusters
- Results
- Recommendations
- Conclusion

Business Problem and Stakeholders (1)

New York City is the third most congested city in the world in terms of traffic and the second worst in the United States.

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

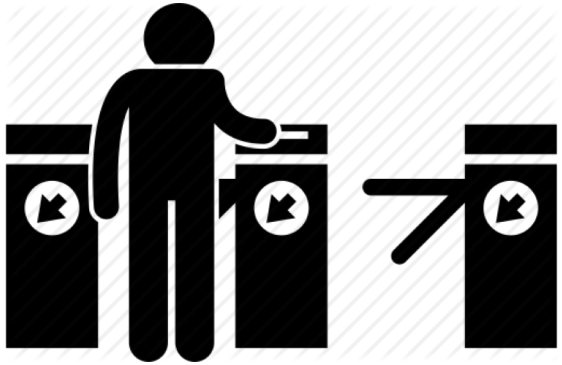
NYC Metropolitan Transportation Authority (MTA) needs a strategy to alleviating traffic.

If more New Yorkers used the subway:

- Less cars would be on the road.
- Gain would raise and it could be spent to pay for a major overhaul of the aging subway .

Business Problem and Stakeholders (2)

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.



Data Set (1)

kaggle

New York City subway entrances, with latitudes and longitudes. A 400 subset is selected.

Use Foursquare API “Explore” to have all the venues near the subway entrances:

- 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.

Data Set (2)

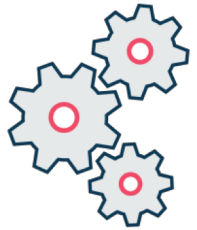
A subset of the final data set.
Venue Category will be used for characterize each venue entrance.

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

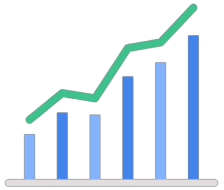
Overall Strategy

In order to create clusters based on venues categories, **K-means algorithm** can be used.

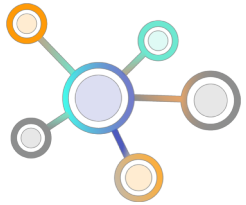
- *Data Preparation*
  - Remove duplicates and transformation on subway entrances.
  - One-Hot encoding and calculate the average number of category types per each venue.



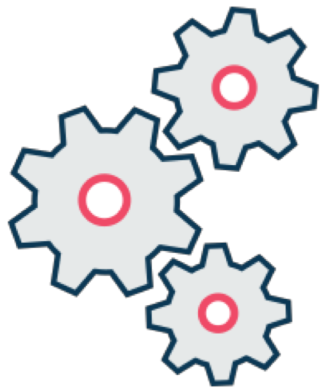
- *Choose best K*
  - Elbow method.
  - Silhouette analysis.



- *K-means clusterization.*



Data Preparation



- Kaggle data set had latitude and longitude in a single field call "the_geom": latitudes and longitudes will have dedicated columns. Furthermore, some rows don't have coordinates information, so they will be dropped.

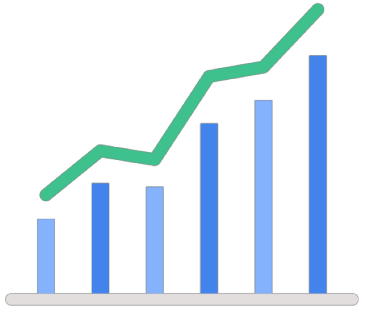
- K-means algorithm doesn't support categorical variables, so one hot encoding needs to be performed on the data. I also performed average of the number of categories per each venue entrance.

- I also normalized data set over the standard deviation (using *StandardScaler*) to help the algorithm to better interpret features with different magnitudes and distributions.

This is the final data used for the k-means:

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

Choose best k (1)



To choose best k I used two methods.

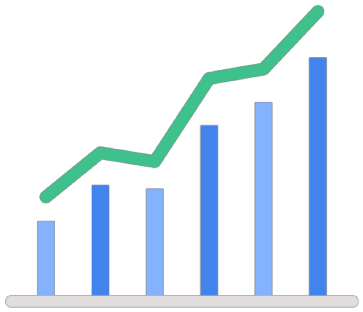
- *The elbow method*.

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.

- *Silhouette analysis*.

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.

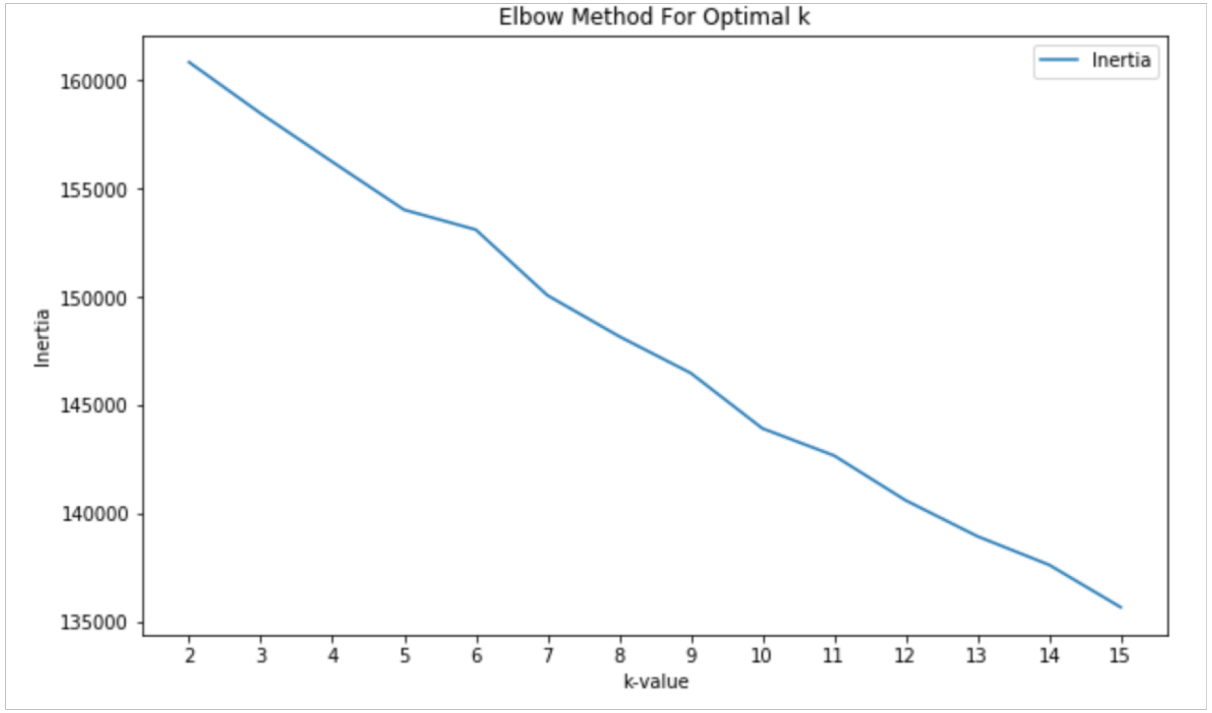
Choose best k (2)

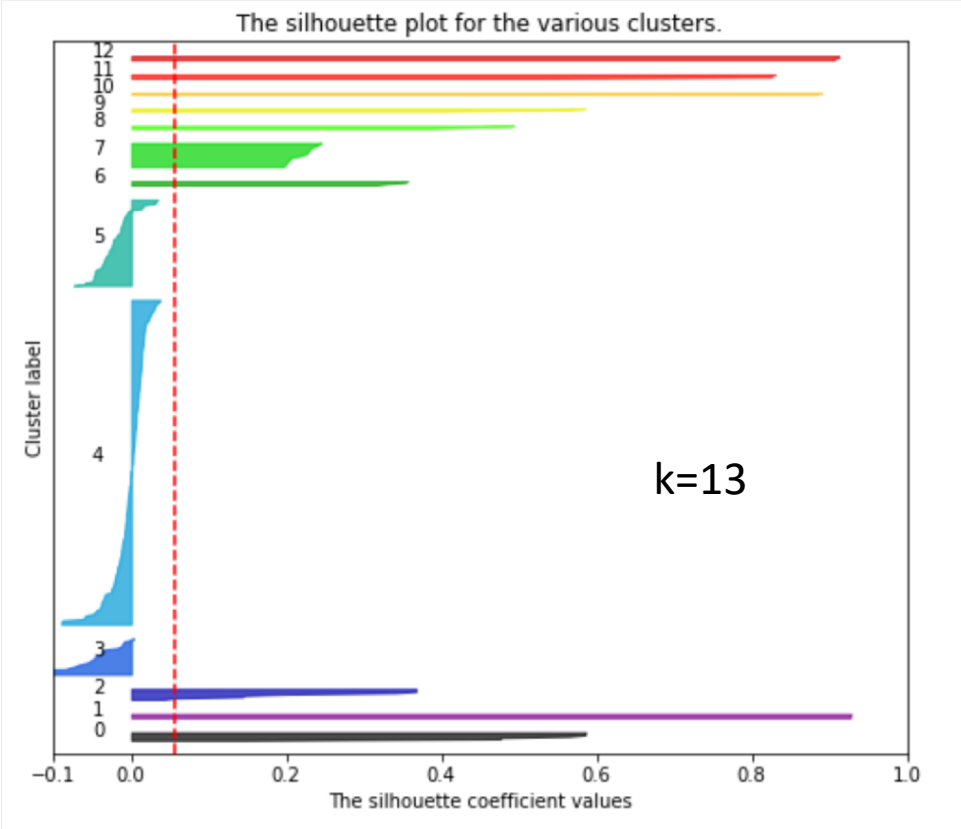


For the elbow method, there is not a clear k to choose, because there is not a clear elbow.
Combining the silhouette I choose:

k=13

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.

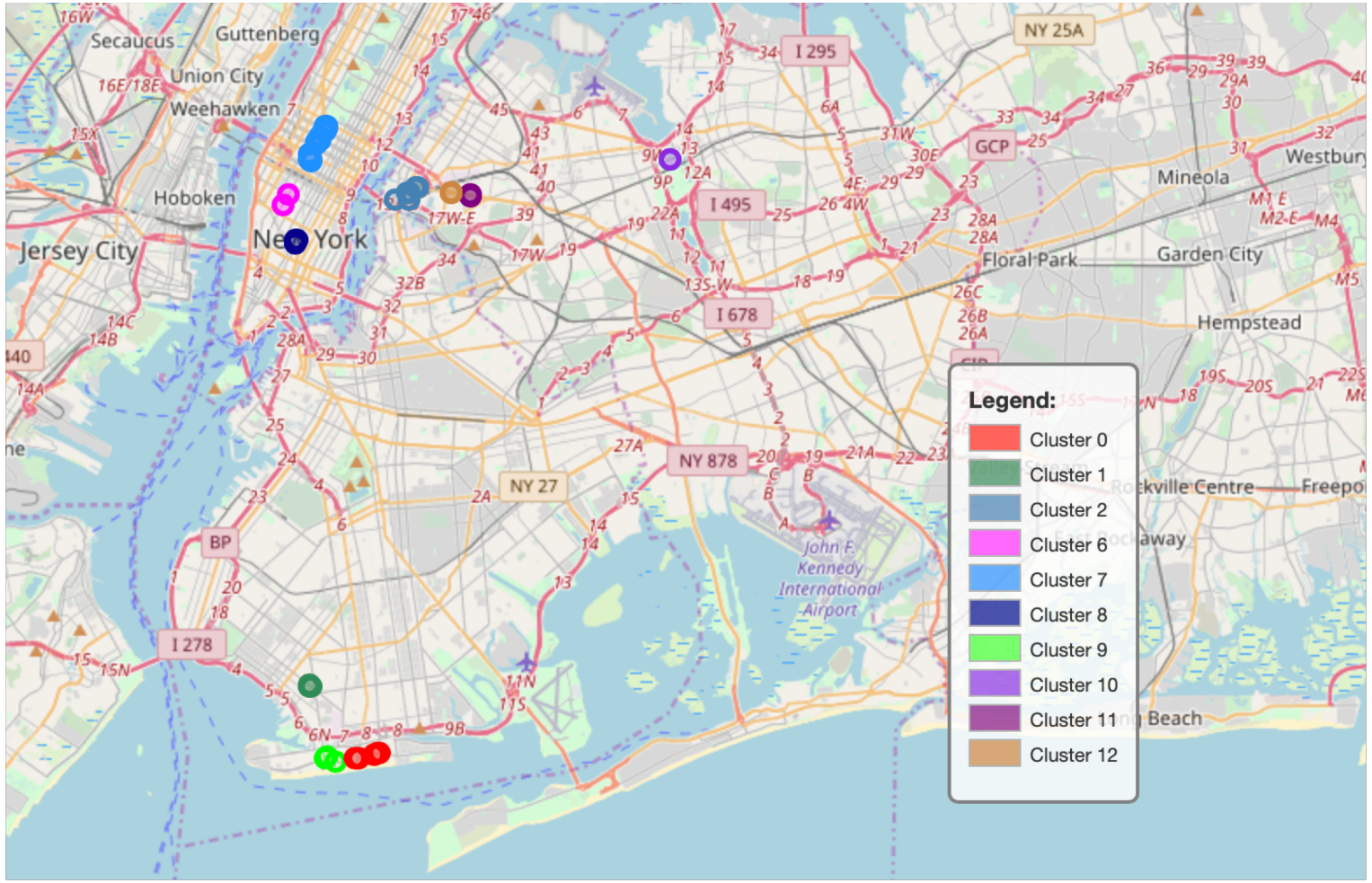




K-means clusters

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
- algorithm= elkan



Results (1)

Here the final clusters' 1st and 2nd most common venue with the relative counts (separated by ;).

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

Results (2)

- 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. AS can be seen from the map, cluster 0 is near Coney Island.

- 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.

Recommendations

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.

Conclusions

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.

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.