CAPSTONE PROJECT - LUCAS P. CARLINI

CLASSIFYING NEIGHBORHOODS OF TORONTO AND NEW YORK

'THIS IS SO NEW YORK'

- Everyone has heard the following phrase: 'This is so New York!!'. Or maybe saying it about another famous city in the world. Maybe you heard it when someone just saw something that remembers this person of New York. Of course, this can happen to almost everything! However, there is some quite unique places in New York City, isn´t? But, actually, are there characteristics of New York that distinguish it from other cities? Like Toronto, perhaps?

CAPSTONE PROJECT

- We are going to try to answer that question by using some Foursquare data, such as data about restaurants, coffee places, etc! We are going to use some machine learning algorithms trying to extract the characteristics that defines it city in order to classify each Neighbourhood as belonging to Toronto or New York. If our models succeed to do so, there are characteristics that defines each city, and, actually, it is correct to say 'This is so New York!!'. In the other hand, if our models fails to classify each Neighbourhood, we can say that both cities are very similar!

DATA

- We used data from New York and Toronto! Venues data from each city were fetched using the Foursquare API!

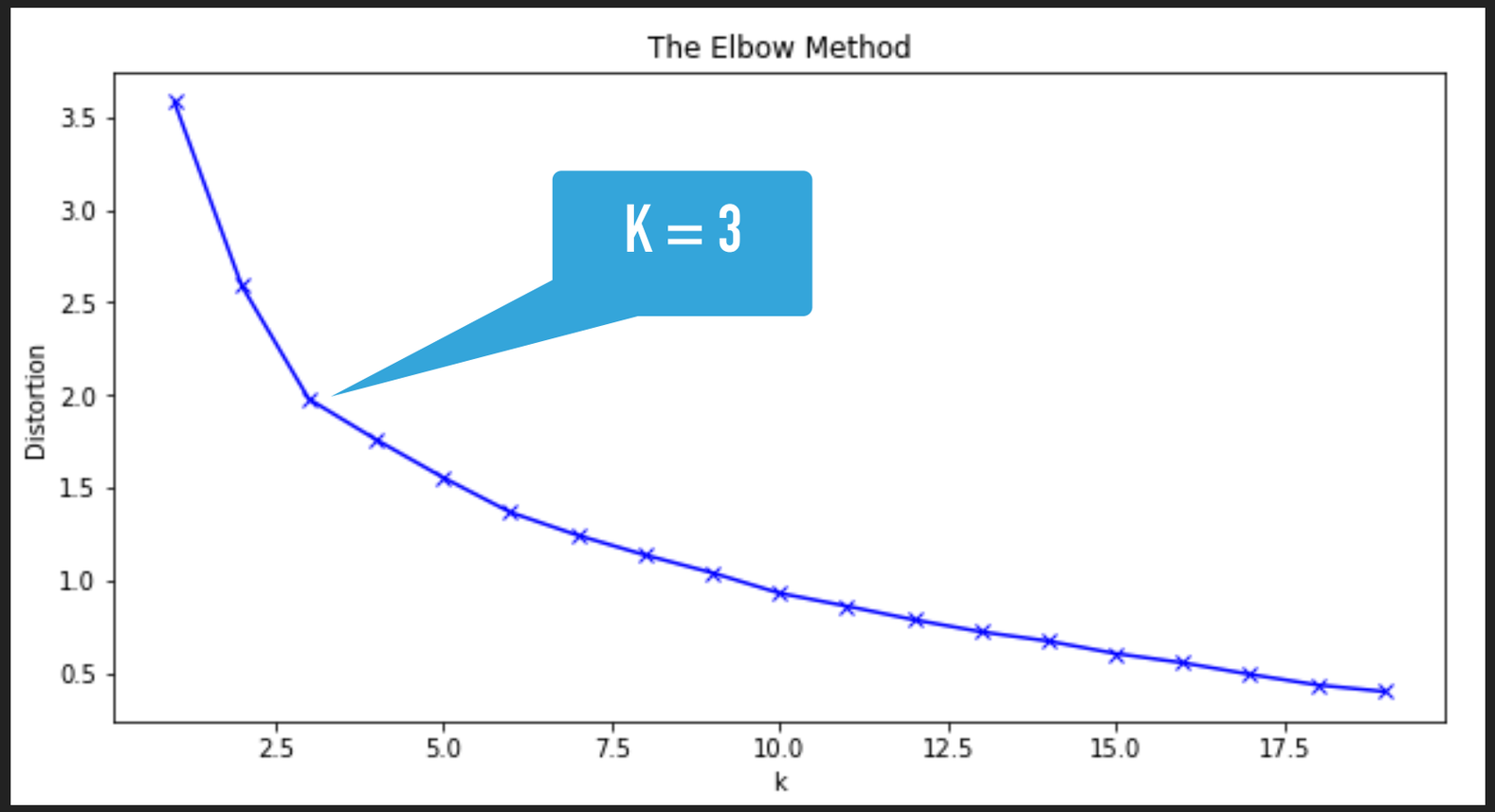
	Borough	Venue	Venue Latitude	Venue Longitude	Venue Category	Neighborhood	Latitude	Longitude	Target
0	East Toronto	Glen Manor Ravine	43.676821	-79.293942	Trail	The Beaches	43.676357	-79.293031	0
1	East Toronto	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store	The Beaches	43.676357	-79.293031	0
2	East Toronto	Grover Pub and Grub	43.679181	-79.297215	Pub	The Beaches	43.676357	-79.293031	0
3	East Toronto	Glen Stewart Ravine	43.676300	-79.294784	Other Great Outdoors	The Beaches	43.676357	-79.293031	0
4	East Toronto	Upper Beaches	43.680563	-79.292869	Neighborhood_Cat	The Beaches	43.676357	-79.293031	0

	Neighborhood	Target	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	...
0	Adelaide, Richmond, King	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	Berczy Park	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	Business Reply Mail Processing Centre 969 Eastern	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	Central Bay Street	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	Christie	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

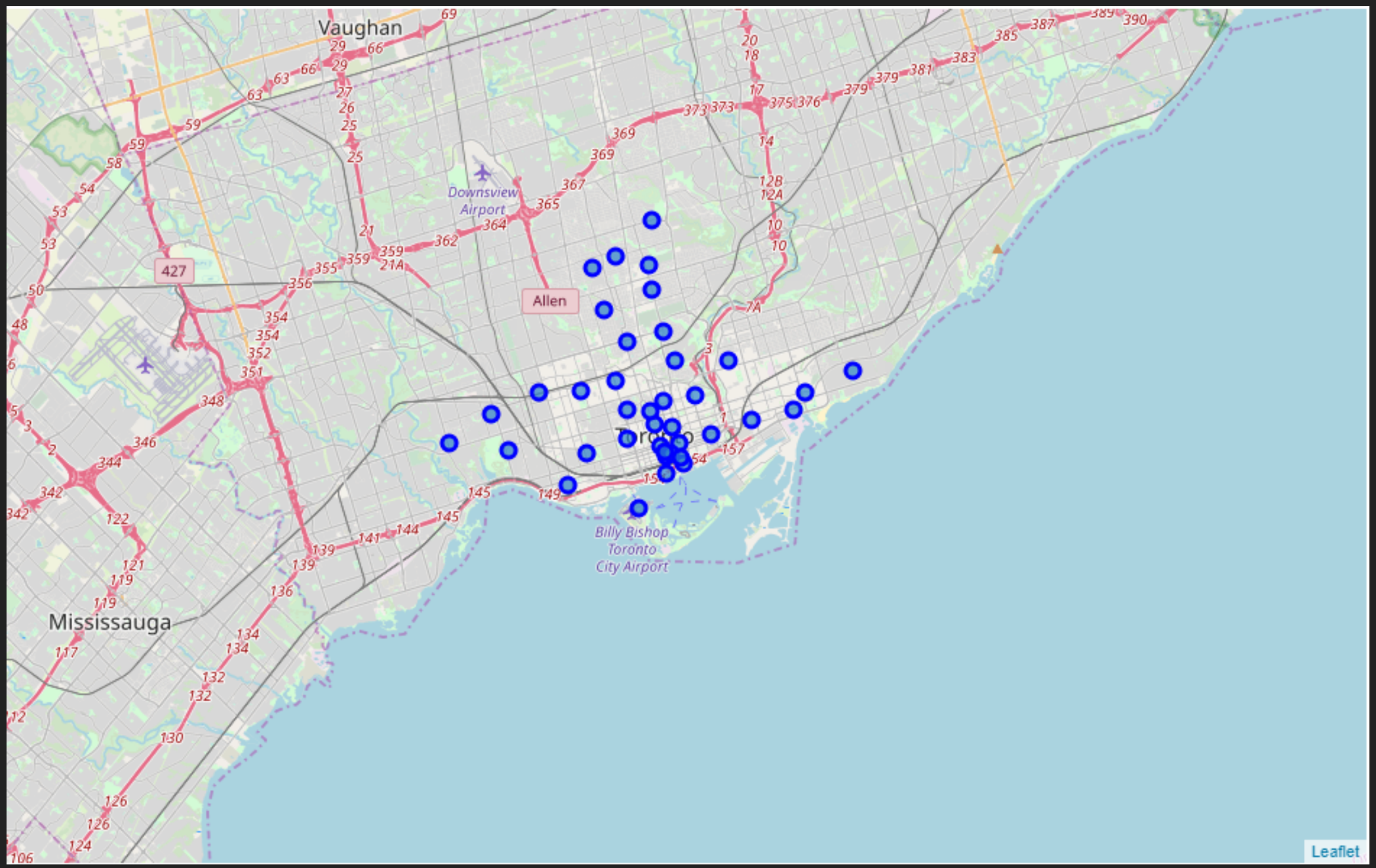
METHODOLOGY

K-MEANS METHOD

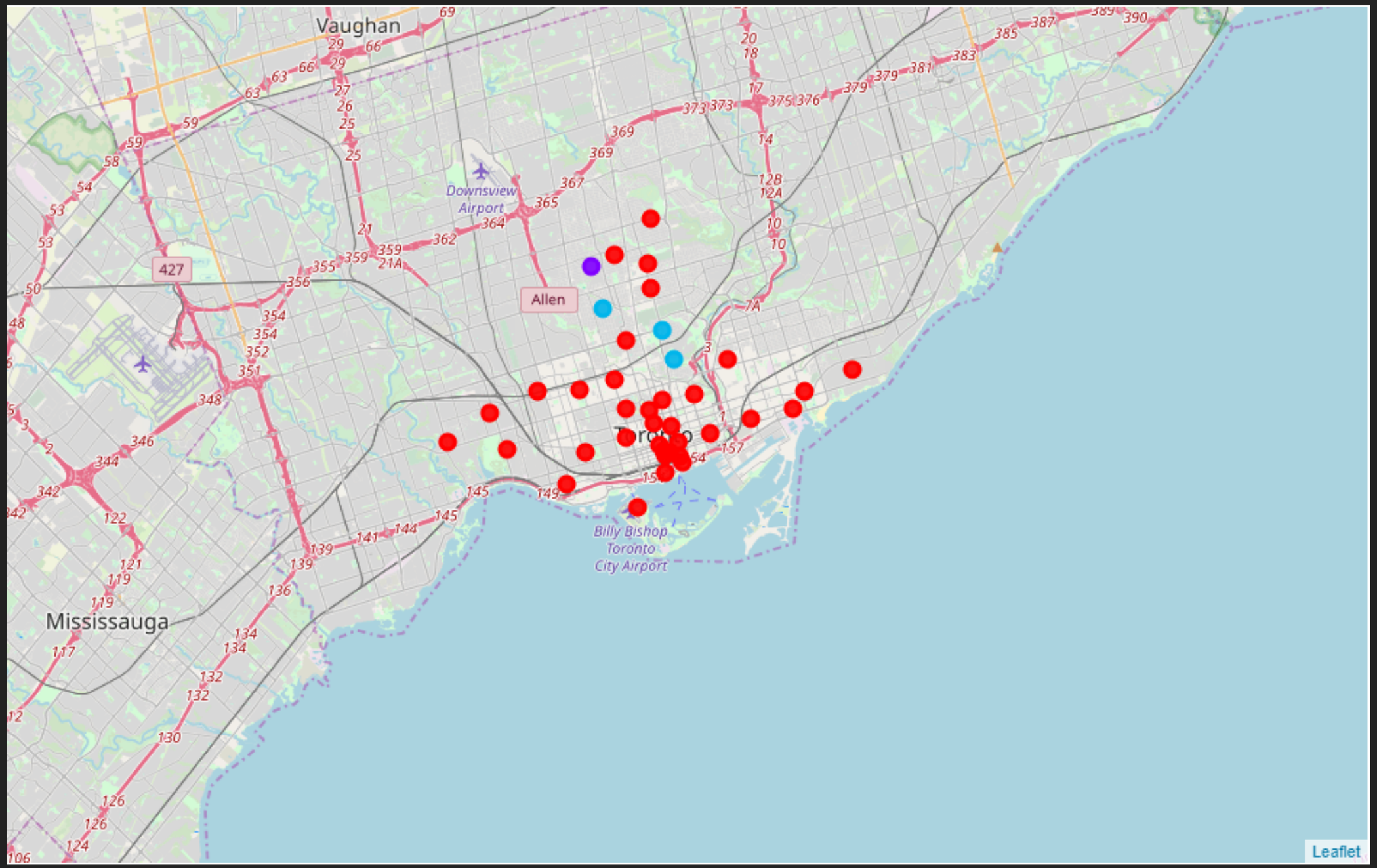
- We used k-Means to cluster the neighborhoods of Toronto based on the frequency of kinds of venues in each one! To evaluate the optimal number of clusters, we applied The Elbow Method.



TORONTO NEIGHBORHOODS

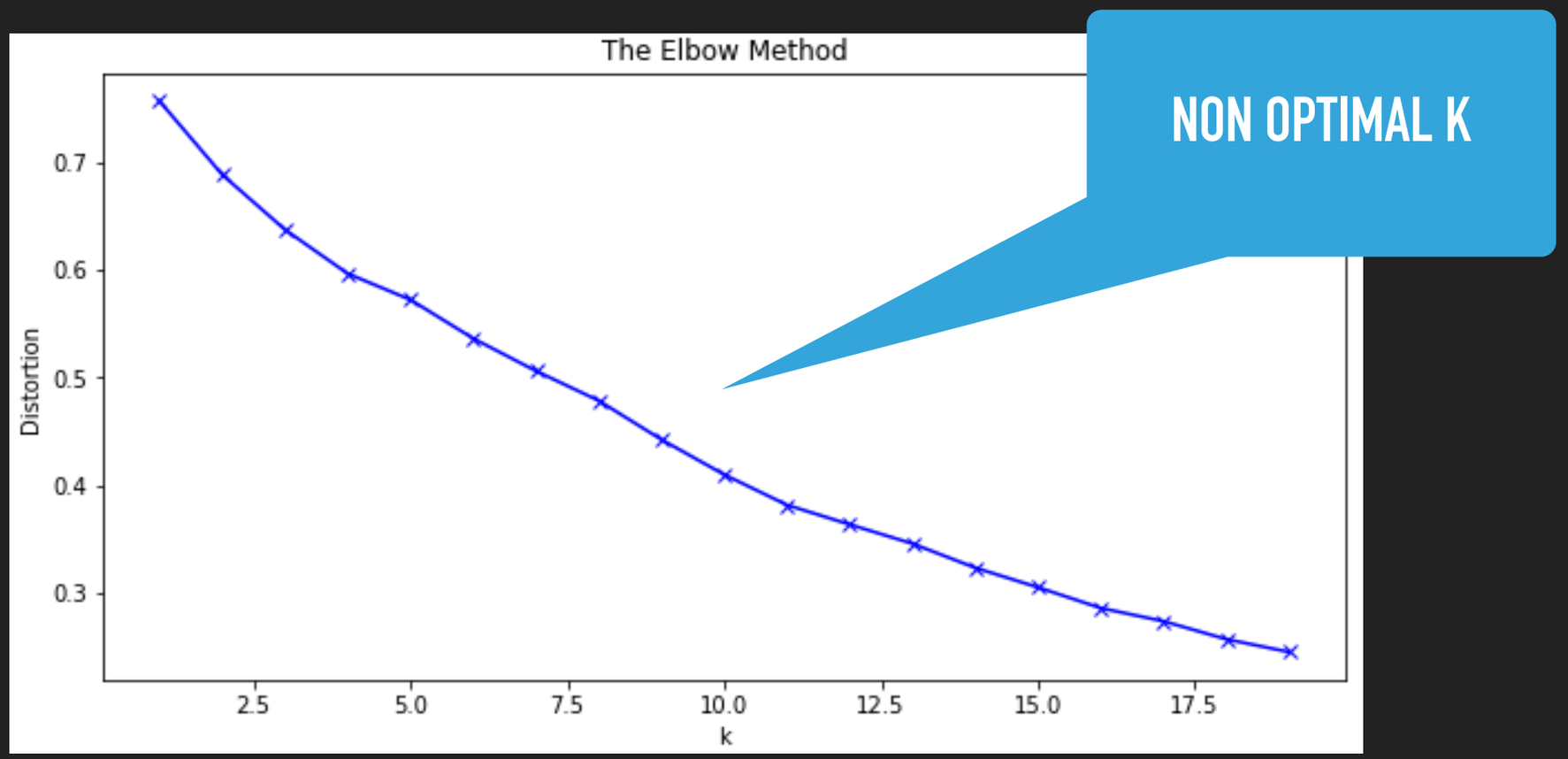


TORONTO NEIGHBORHOODS CLUSTERED BY K-MEANS

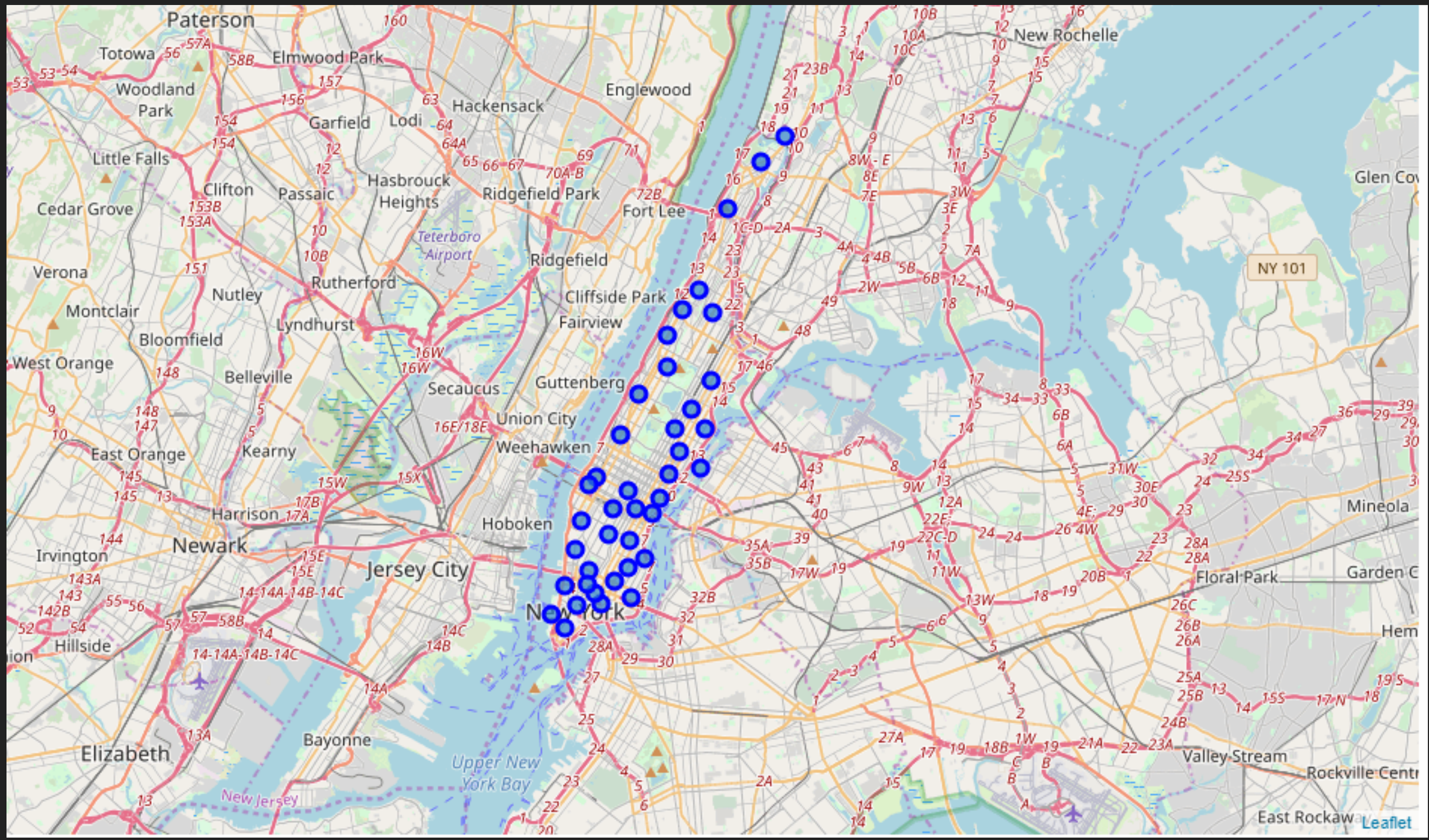


K-MEANS METHOD

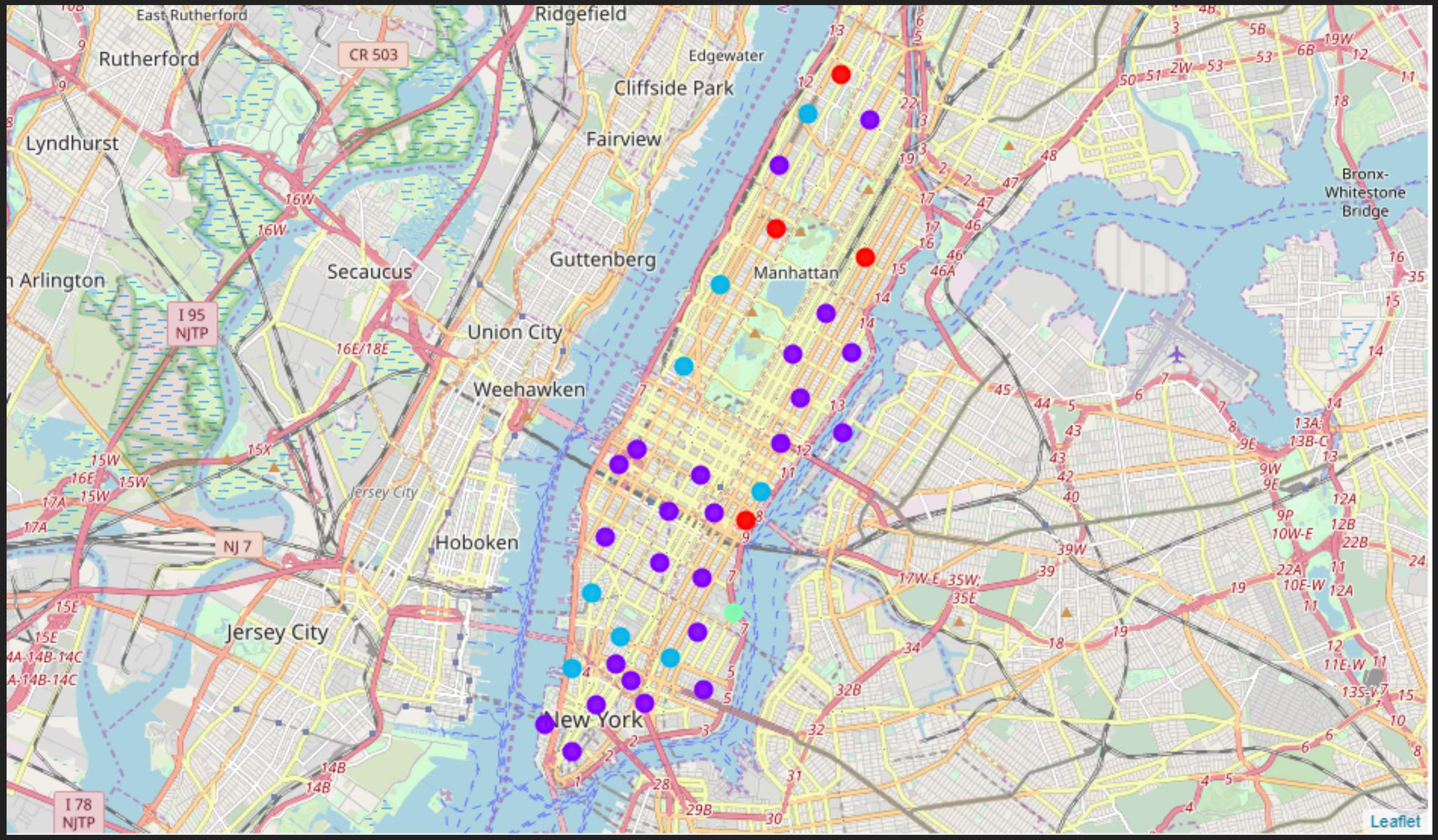
- Similarly to Toronto data, we applied The Elbow Method to New York data! However, non optimal k was found. We, then, used k = 5 to cluster the data!



NEW YORK NEIGHBORHOODS



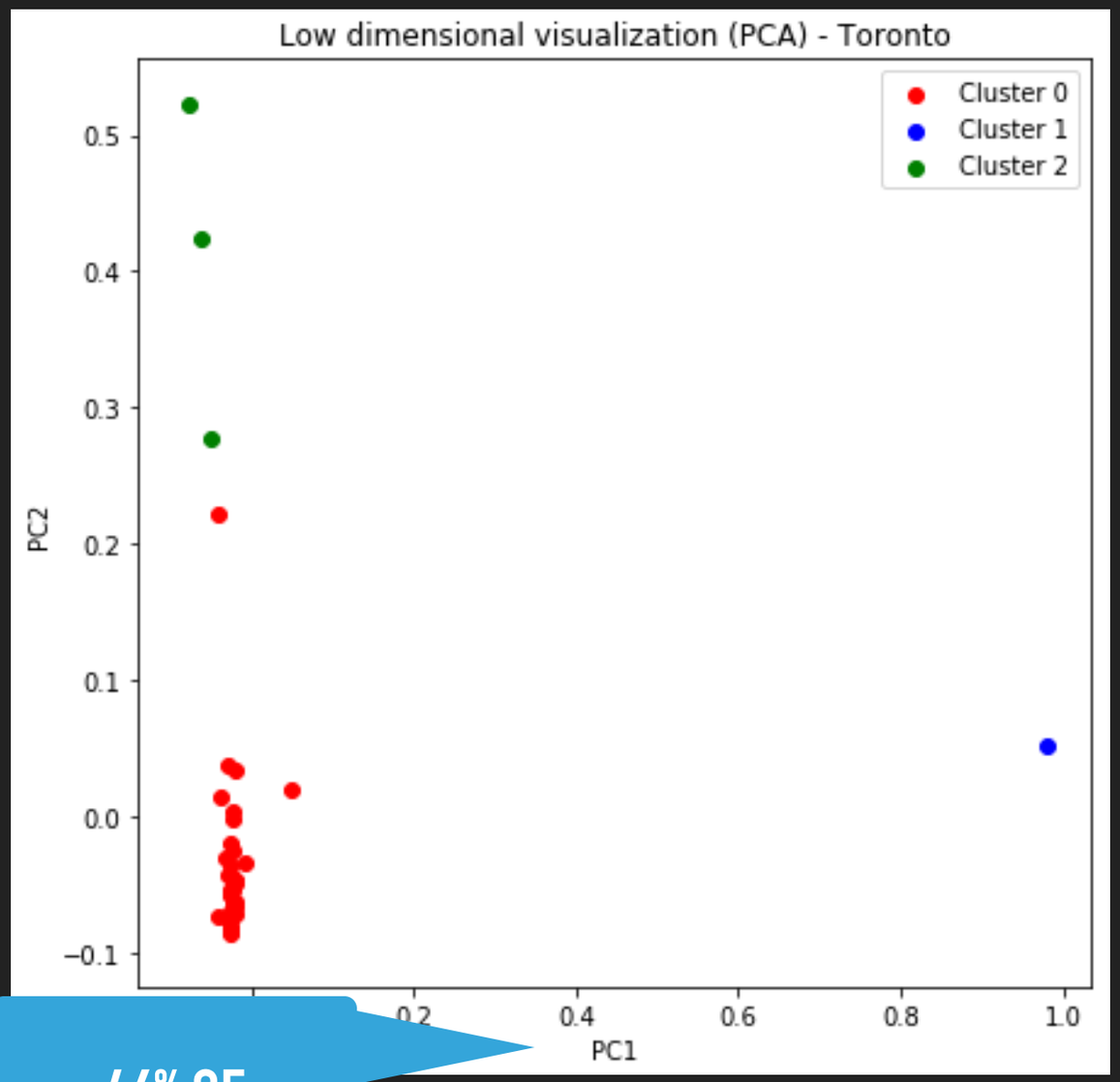
NEW YORK NEIGHBORHOODS CLUSTERED BY K-MEANS

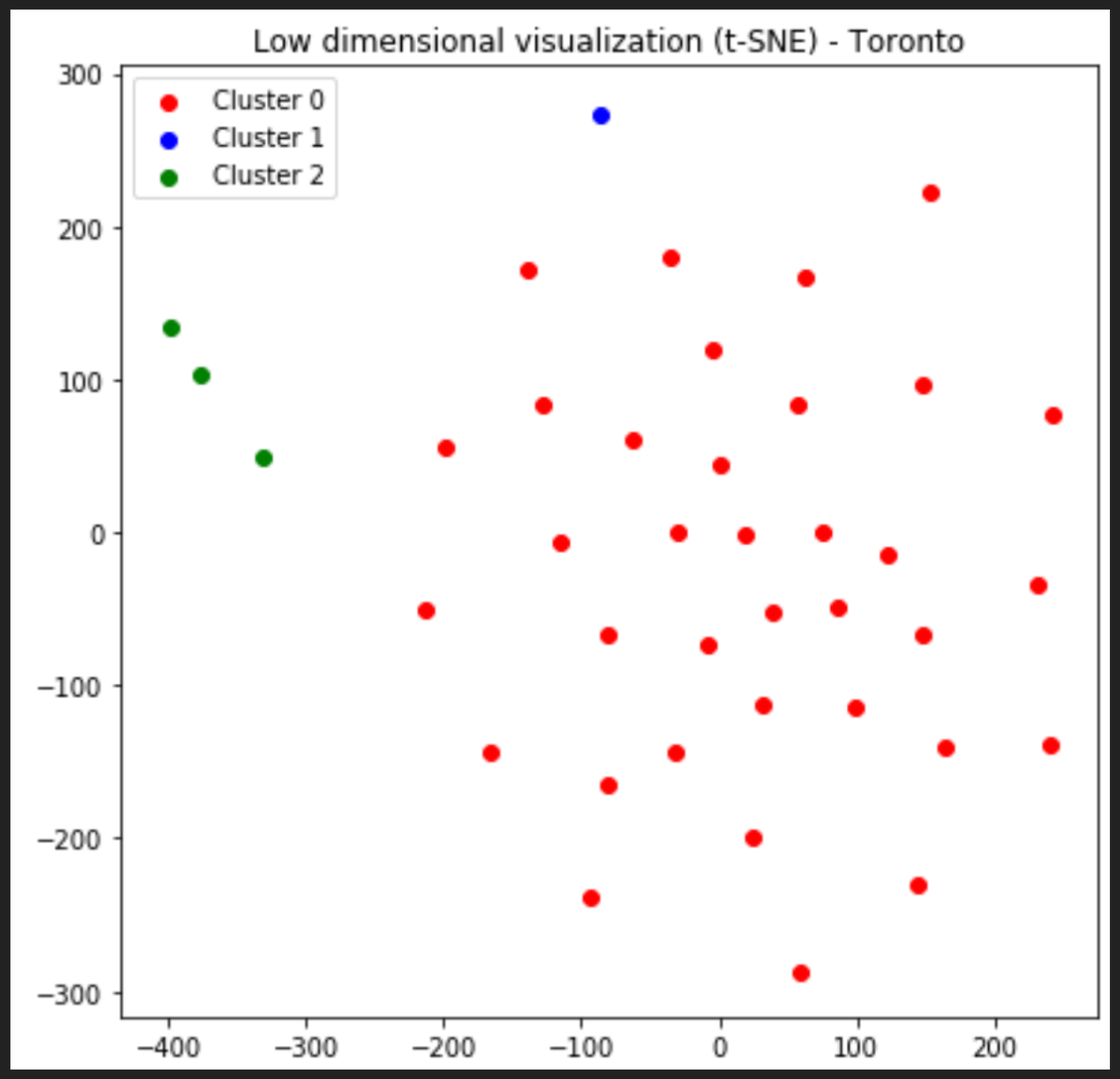


HIGH DIMENSIONAL DATA VISUALIZATION

- We will use the already clustered neighborhoods of each city with a couple of unsupervised statistical approaches, PCA and t-SNE, to visualise these high dimensional data in a two dimensional plot in order to see how the clusters and data are related in the original high dimensional space.

TORONTO NEIGHBORHOOD DATA

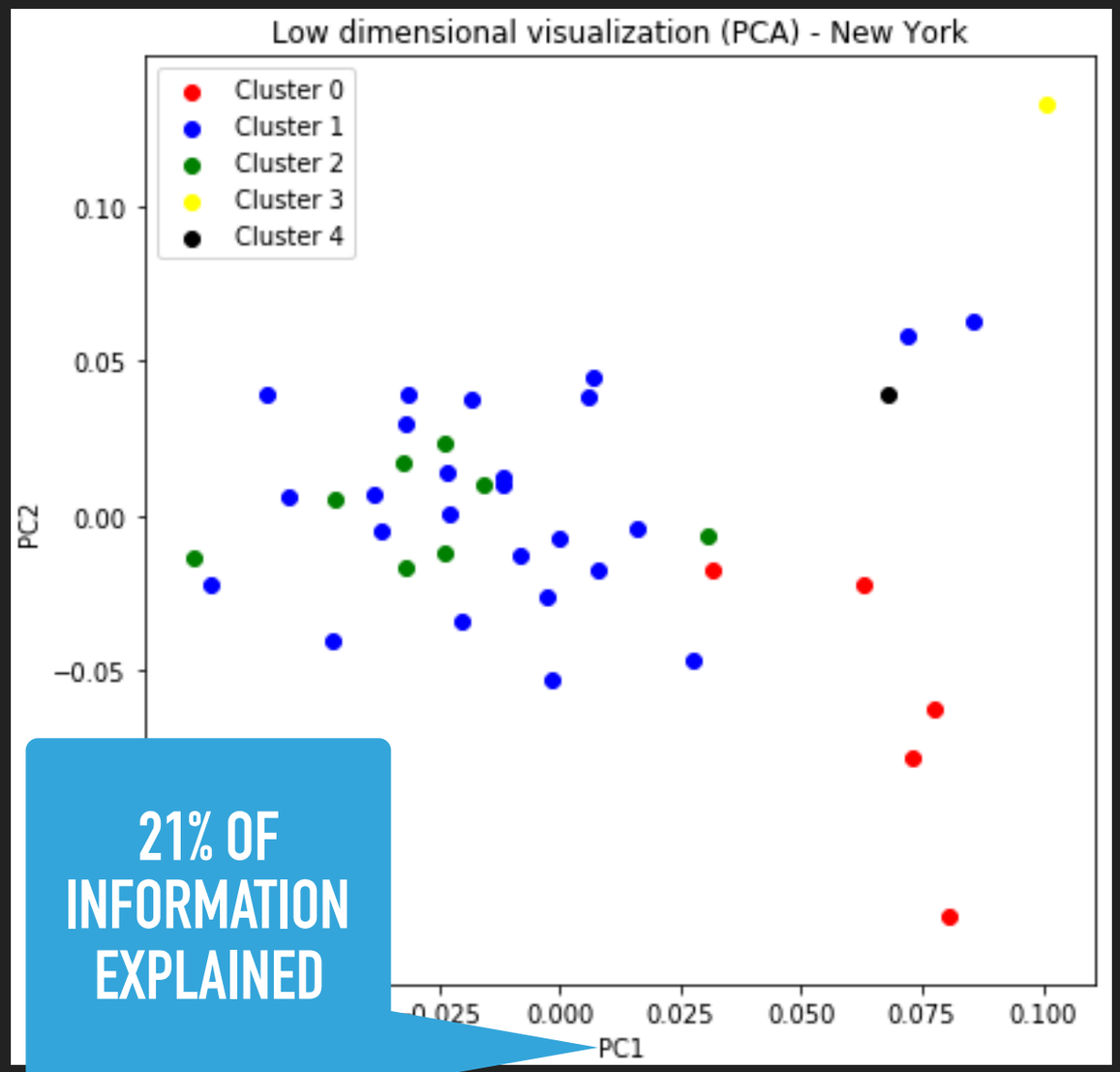




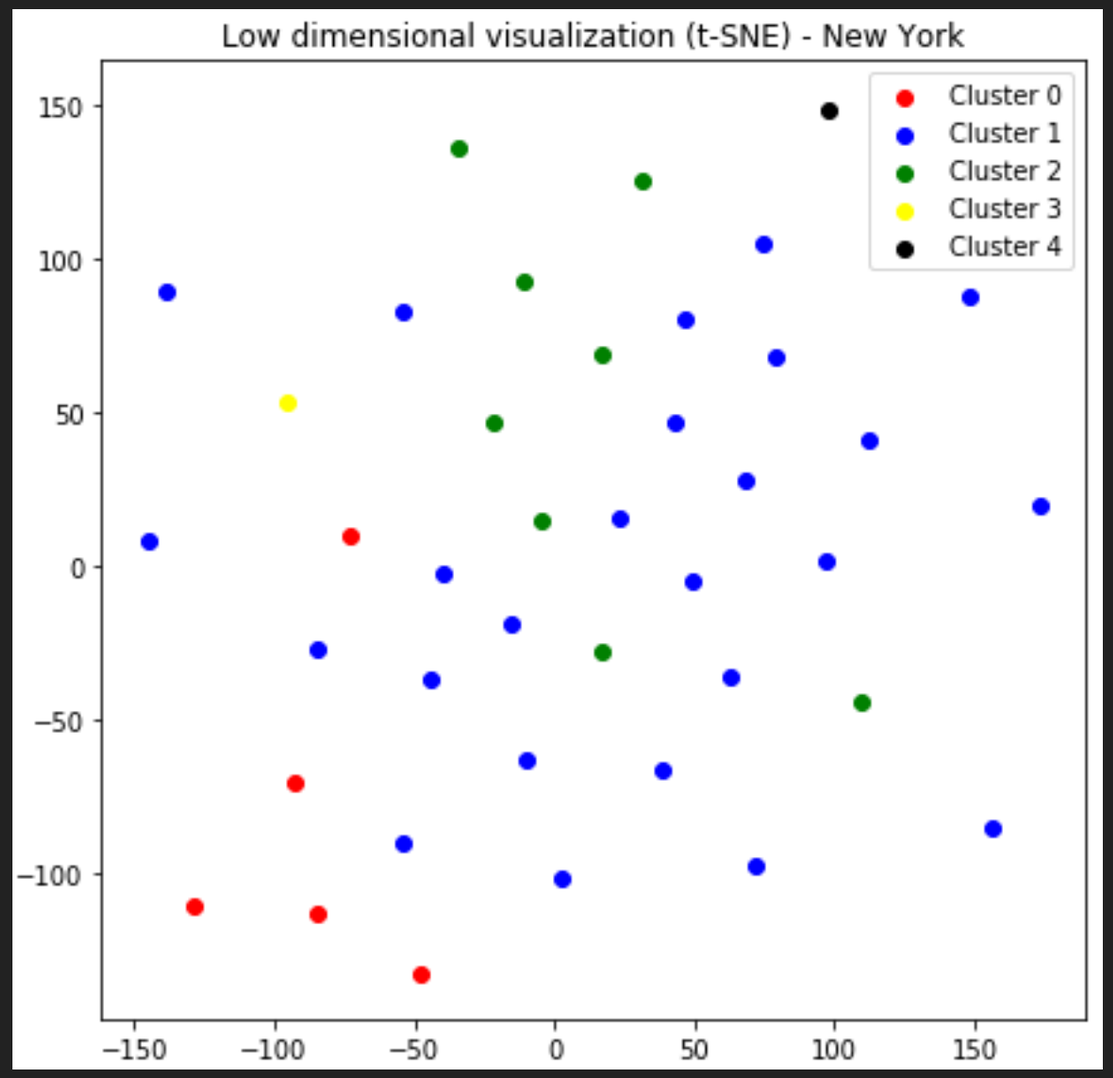
46% OF INFORMATION EXPLAINED

Well defined clusters! It means that k-Means performed well!

NEW YORK NEIGHBORHOOD DATA



21% OF INFORMATION EXPLAINED



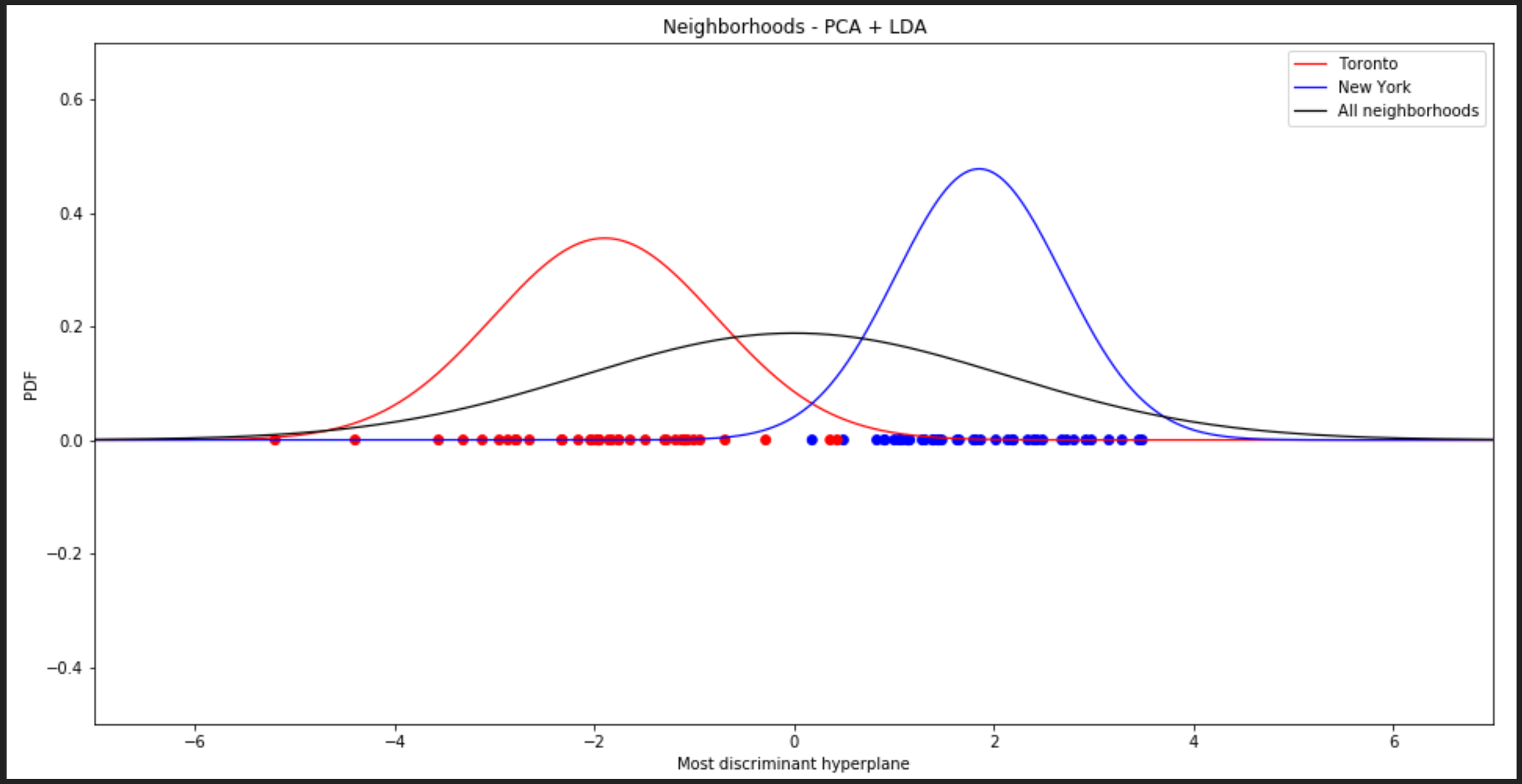
Poorly defined clusters! Since the Elbow Method failed, that was expected!

MACHINE LEARNING MODELING

- We will use the PCA + LDA approach to verify if we are facing a linearly solved problem! This method obtains the direction of discrimination of sample groups! In other words, neighborhoods samples of New York to one side, and Toronto neighborhoods samples to the other!

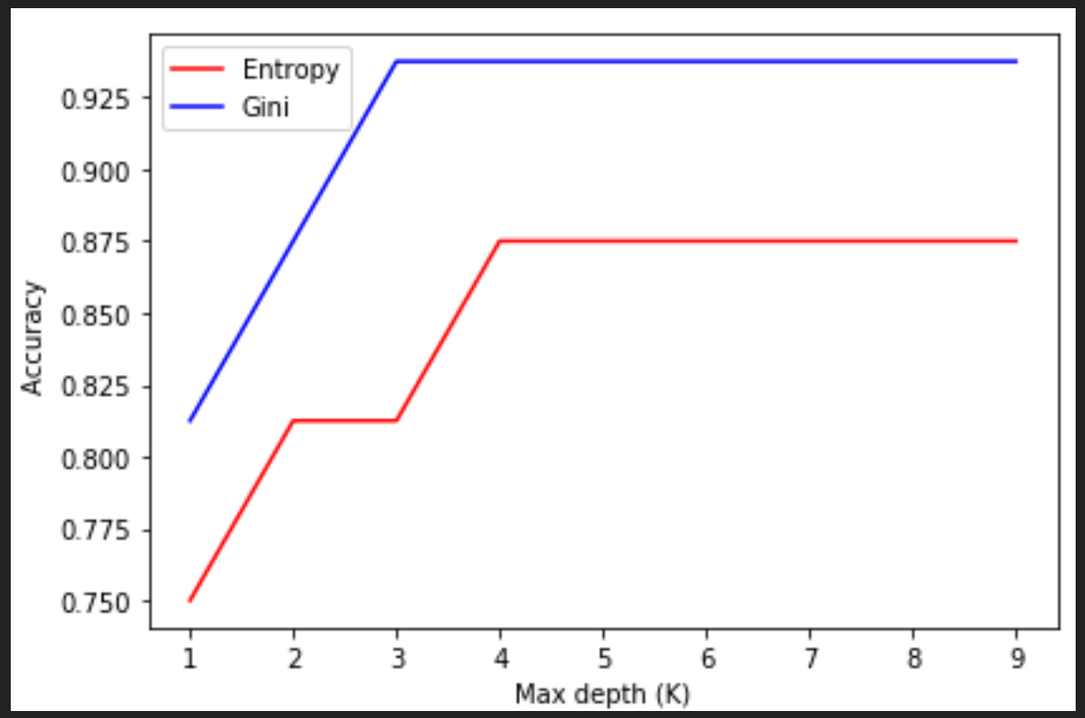
- Then, we use Decision Trees to classify each neighborhood as belonging to New York or Toronto!

PCA + LDA METHOD



We can see that our sample groups are far apart! That means that our samples are linearly discriminant, and, therefore, can be classified by commonly used algorithms such as the k-NN!

DECISION TREE

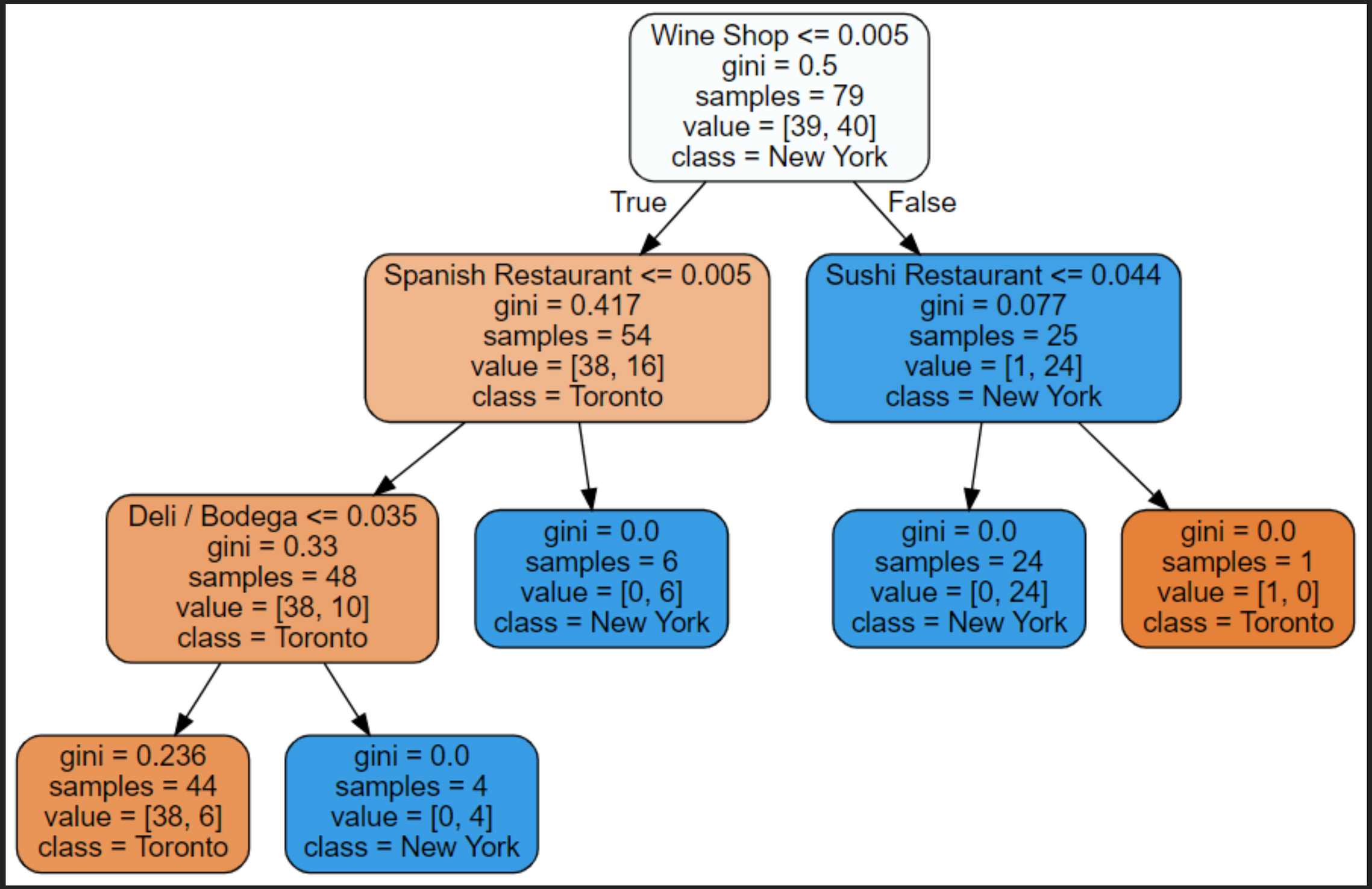


JACCARD SCORE = 0.9375
F1 SCORE = 0.9377

The best entropy accuracy was 87.5 % with max_depth = 4.

The best gini accuracy was 93.75 % with max_depth = 3.

DECISION TREE



CONCLUSIONS

- After this massive analysis of Toronto and New York neighborhood data, are we able to say "This is so New York"? Yes, we can say that! Analysing our results, New York seems to have a higher frequency of wine shops in its neighborhoods than Toronto! Moreover, Sushi Restaurants are not so popular in New York neighborhoods, while Deli / Bodegas shops are more frequent on it than it is in Toronto!
- As future work, we propose the use of data of more cities around the world in order to investigate the differences and similarities among each one. Moreover, we suggest the use of DBSCAN (or other clustering algorithm) instead of k-Means, since the latter presented some instabilities when evaluating New York neighborhood clusters! Lastly, we propose the use of different features to classify each city, such as number of non-natives, imigrations per year and economic indicators.