FINAL ASSIGNMENT – CAPSTONE PROJECT
DECISION MAKING-PROCESS BASED ON CLUSTERING

EL HOUR REDA

FOR COURSERA FOR ACADEMIC PURPOSES

- 04TH DECEMBER 2020 -

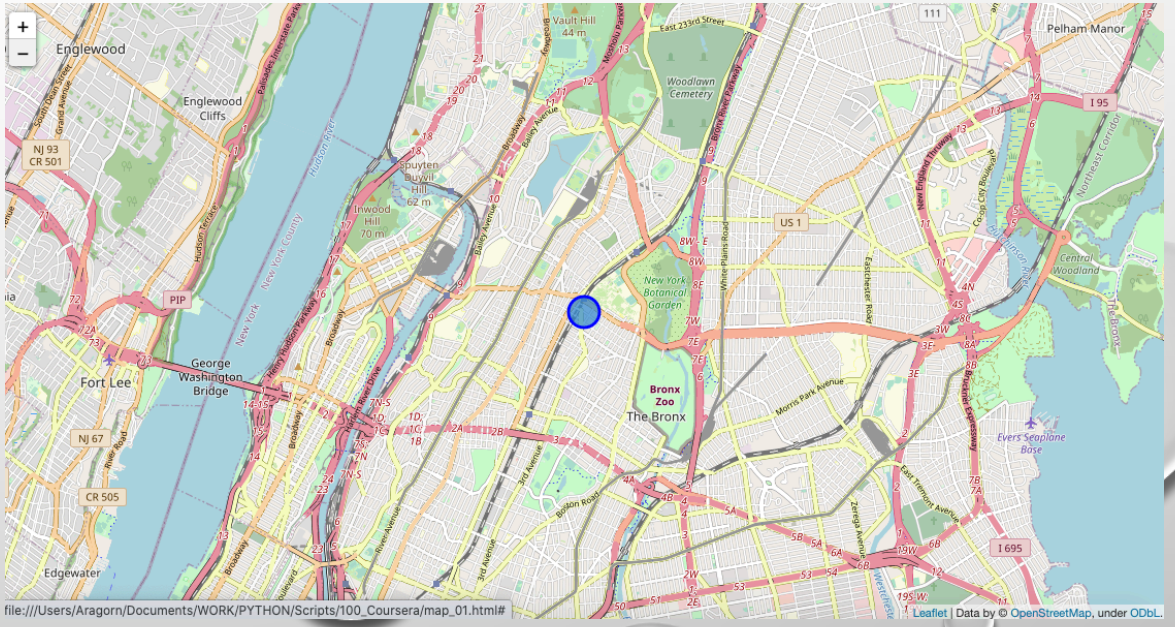
HELPING JOHN ASSES NEIGHBORHOODS IN STATEN ISLAND

John lives and works in 522 E 189th St Belmont, Bronx, New York. He has been offered a new job opportunity to work in Staten Island, NY which is far from his home in his Belmont neighbourhood. John has a preference for bakery shops.

The Objective:
Look for neighbourhoods in Staten Island having similar properties such as John's address in Belmont

The tool:
Unsupervised learning algorithm K-means

The features:
Location data from Foursquare API



DATA ACQUISITION AND PREPARATION (1/2)

- Data from https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN SkillsNetwork/labs/newyork_data.json
- *Data organisation: data has a total of 5 boroughs and 306 neighbourhoods with longitude and latitude coordinates.*
- *Data visualisation achieved using geopy for maps matplotlib for graphs*
- *Data cleaning realized for neighborhoods in Staten Island.*
- *John's address is converted to longitude and latitude using geopy nominatim*

DATA ACQUISITION AND PREPARATION (2/2)

- Foursquare API is called to get informations on each neighbourhood.
- The parameters called are venue categories for each neighbourhood.
- Parameters used in API call: Radius =800 m , LIMIT = 100

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Belmont	40.859823	-73.890671	Primavera Cafe	40.862004	-73.891537	Café
Belmont	40.859823	-73.890671	Pizza Studio - Fordham	40.860749	-73.889815	Pizza Place
Belmont	40.859823	-73.890671	North End Wine and Liquor Store	40.861131	-73.892024	Liquor Store
Belmont	40.859823	-73.890671	Fordham Plaza	40.861440	-73.890910	Plaza
Belmont	40.859823	-73.890671	Tino's Delicatessen	40.855882	-73.887166	Italian Restaurant

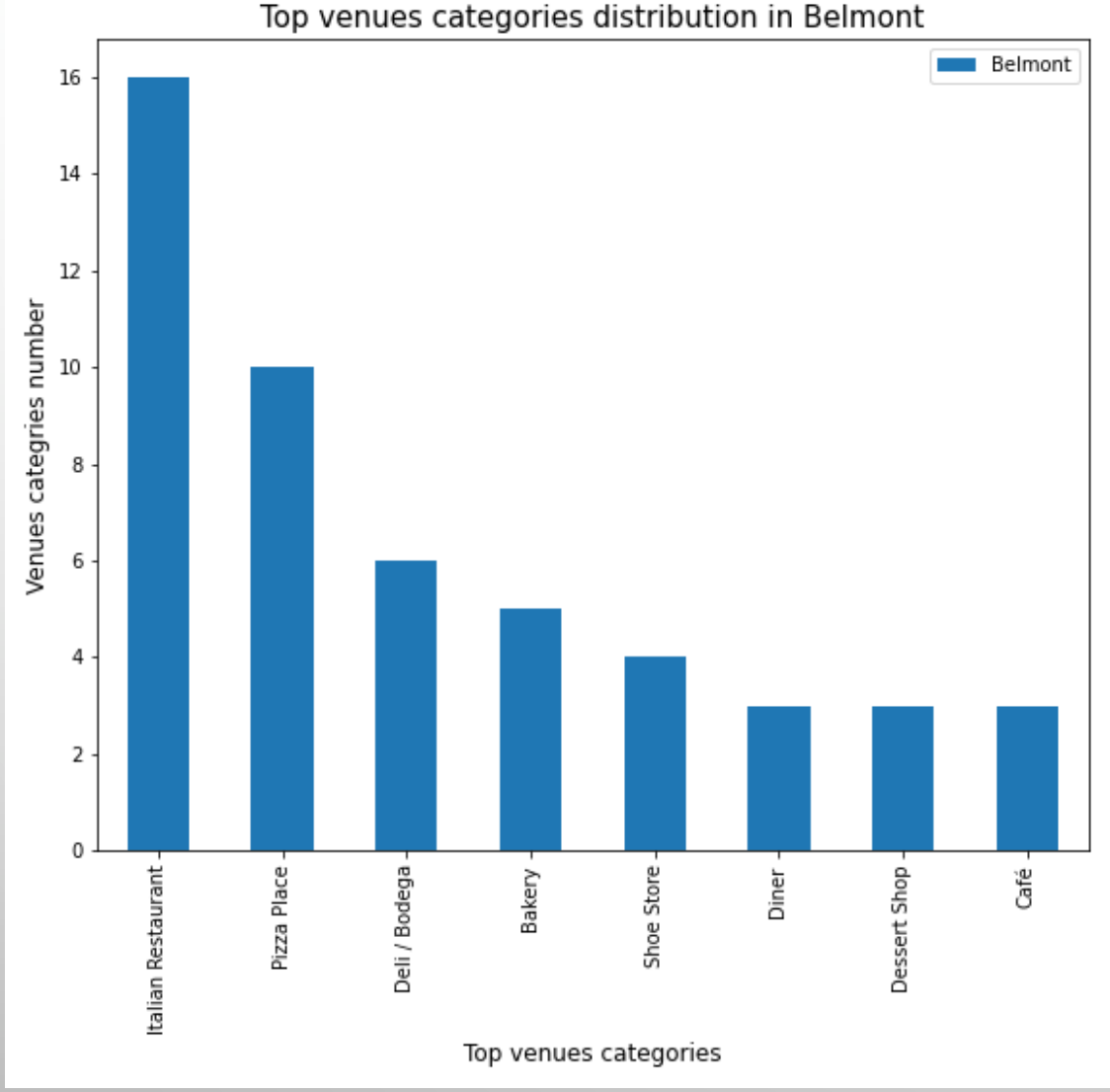
DATA ANALYSIS

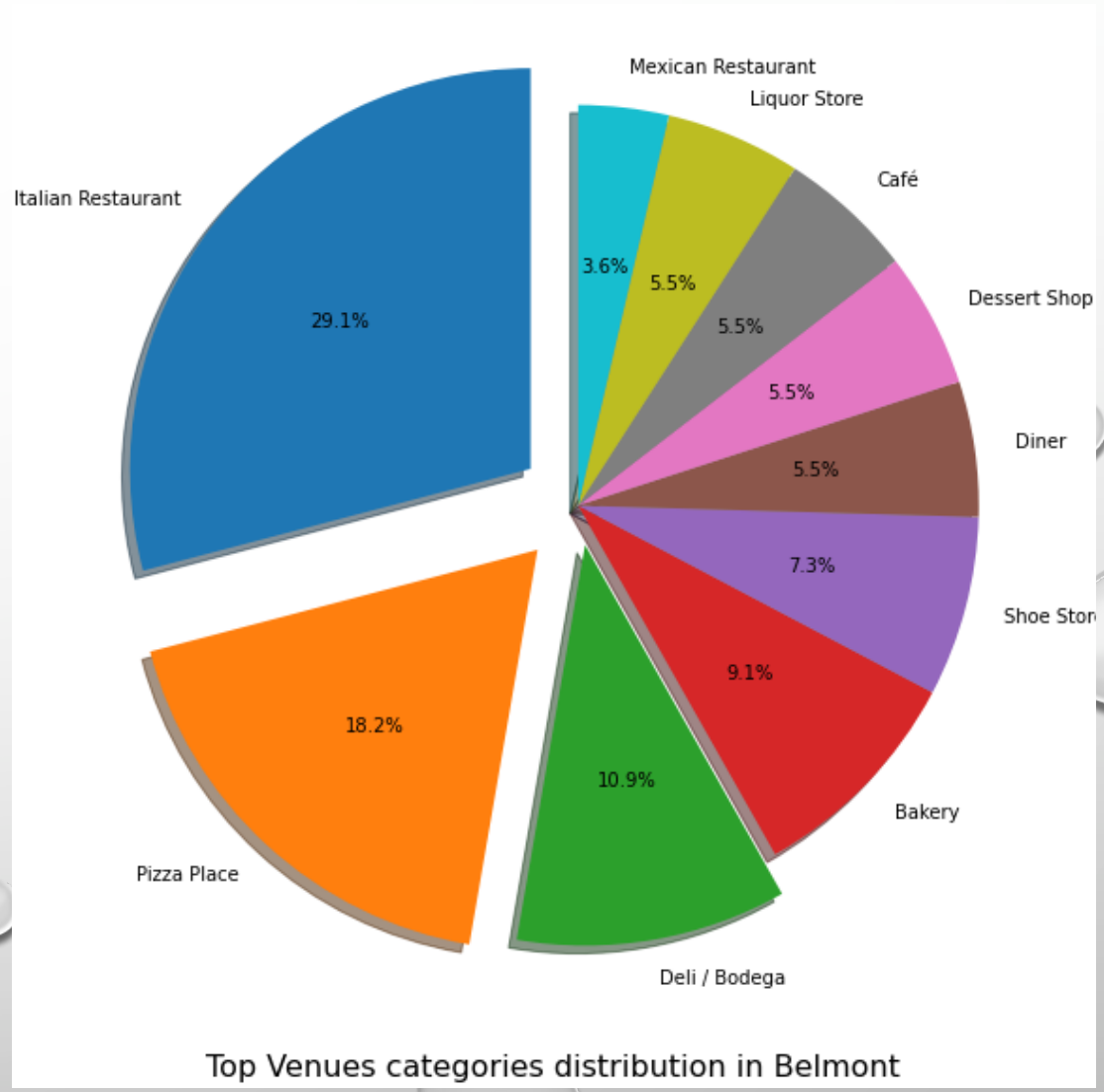
- Venues categories are represented by categorical data. The clustering algorithm isn't directly applicable to categorical variables because Euclidean distance function isn't really meaningful for discrete variables. Hence, one- hot encoding approach is used to convert these categorical data to numerical data.

- We divide each venue categories by the sum of all venues categories in order to have numerical values ranging between 0 and 1 which will allow the algorithm to compare all the neighbourhoods.

- The process is applied to all venues categories of Staten Island.

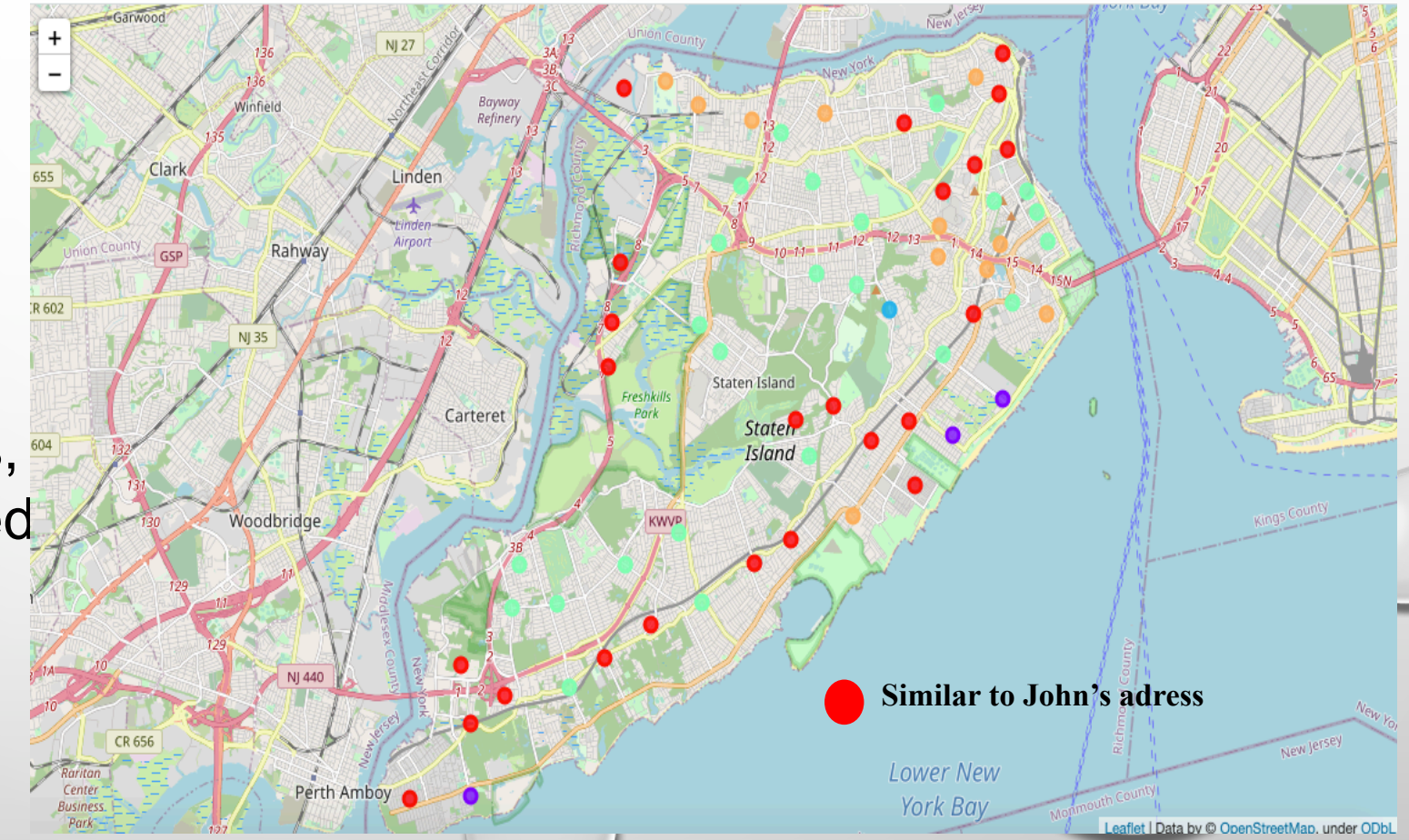
SCALING FEATURE





RESULTS AND DISCUSSION (1/2)

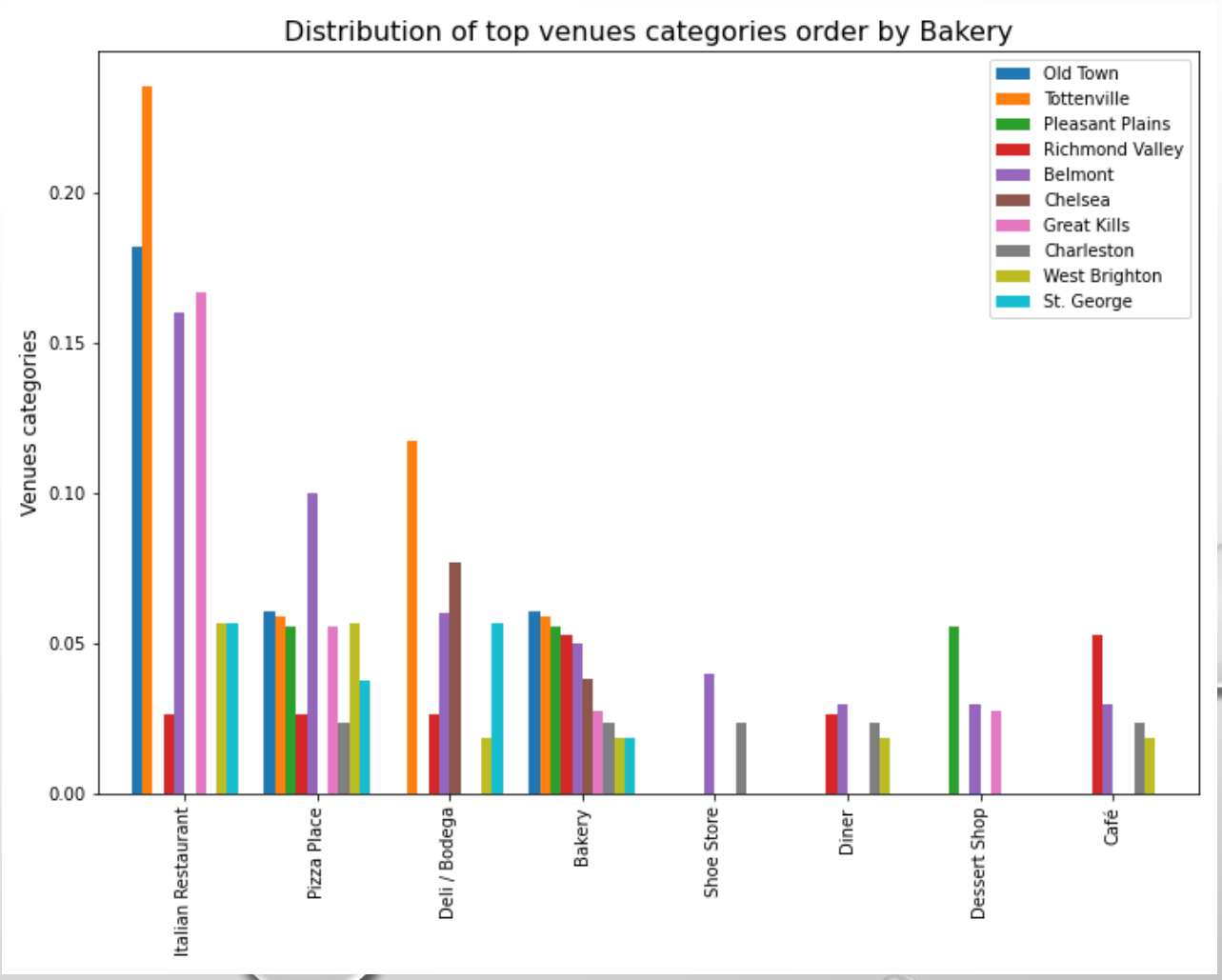
- The results show that there are 24 neighbourhoods in Staten Island similar to John's address in Belmont.
- We know John has a preference for bakery shops, we can plot our results based on this assumption.



RESULTS AND DISCUSSION (2/2)

- Four neighbourhoods having approximately the same distribution of bakeries such as Belmont: Old Town, Tottenville, Pleasant Plains, Richmond Valley.

- Tottenville and Great Kills have a significant distribution of the most common venues in Belmont but the less common venues are not well represented while West Brighton is well distributed between many most common venues of Belmont.



CONCLUSION AND PERSPECTIVES

- We can advise John to look for a new place in Tottenville if he is very interested in finding the four most common venues categories of Belmont (Italian restaurant, Pizza place, Deli/Bodega and bakery). Otherwise, if John is more interested in the overall most common venues categories, we could advise him West Brighton or Richmond valley.

- John knows now that there are similar neighbourhoods in Staten Island and he could accept the new job offer. A further analysis could be improving the decision by narrowing the final list of 24 neighbourhoods by adding features such as proximity to his new work, proximity to airport or rent price.