Capstone Project for Data Science course

The Final Assignment

Introduction and Business problem

- This project **deals** with analysing the neighborhoods of Edmonton, the capital city of Alberta, Canada.
- The **objective** of this project was to analyse the neighbourhoods in Edmonton and divide them into different clusters based on the popular venues at each neighbourhood, by using data science methodology and machine learning techniques like clustering.
- The **target audience** of this project is small business owners who want to set up their businesses like restaurants, coffee shops, beauty and health stores, etc.

Data requirements

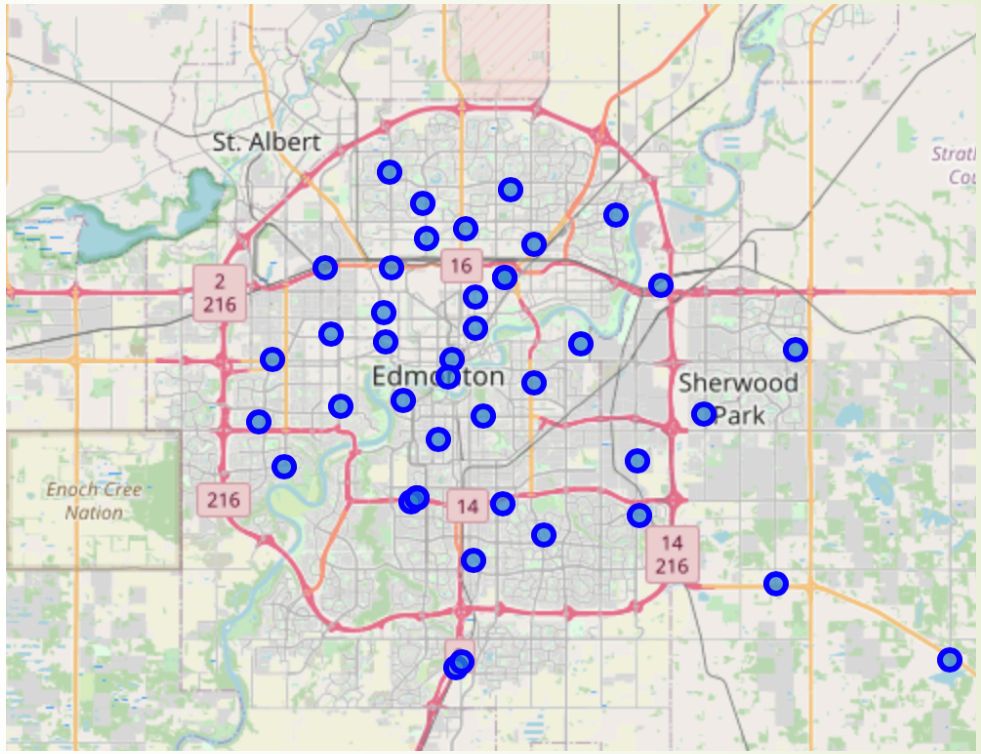
- A dataset with Borough name, Neighborhoods, Post Code and GPS coordinates was taken from Wikipedia website
- The Foursquare API was used to access the venues in the neighborhoods
- Then neighborhoods were clustered based on their venues using Data Science Techniques, namely k-means algorithm
- The optimal number of clusters was obtained using silhouette coefficient

Data requirements continue

- Folium library was used to visualize the clusters on the map of Edmonton
- These clusters were analysed to help small businesses select a suitable location for their need like restaurants, hotels, shopping malls
- The pre-processing and data cleaning involved scraping data of Alberta region with all neighborhoods, then extracting directly neighborhoods of Edmonton and scraping venue items

Methodology

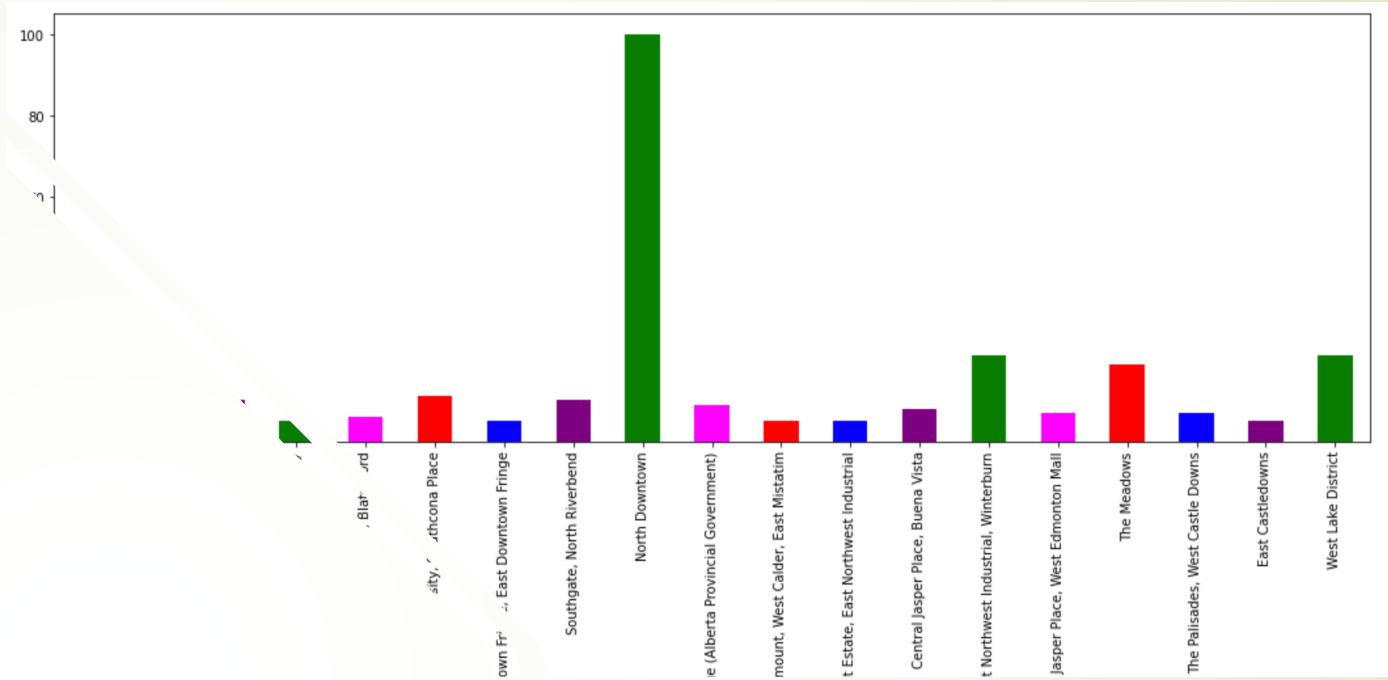
- In total:
  - 38 neighborhoods were obtained
  - 309 venues and 123 unique categories have been identified
- Due to multiple neighborhoods with less than 5 venues were returned, I considered to use only the neighborhoods with more then 5 venues for a better analysis.
- Using one hot encoding I found the 10 most common venue categories in each neighborhood.
  - Then KNN clustering technique have been used. To find the optimal number of clusters silhouette coefficient was counted.
- The clusters can be analysed to suggest stakeholders suitable locations based on the category.



Edmonton map with neighborhoods

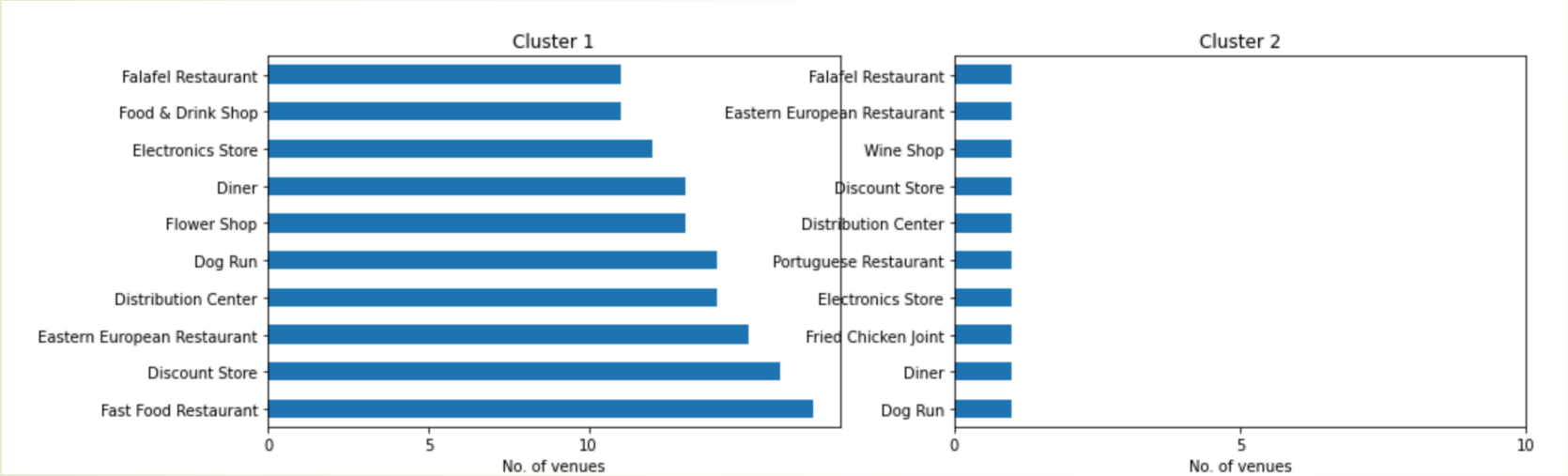
Analysis

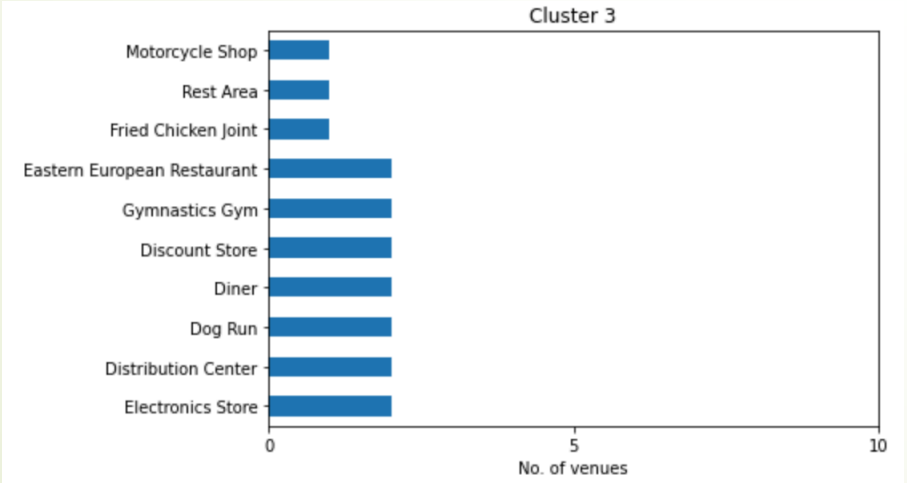
- Neighborhoods with more than 5 venues
- As clustering technique, K-Nearest Neighbor (KNN) was applied
- To find the best value for K, the silhouette coefficient was counted



		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
	Construction & Landscaping	Grocery Store	Fried Chicken Joint	Diner	Discount Store	Distribution Center	Dog Run	Eastern European Restaurant	Electronics Store
	Water Park	Trail	Liquor Store	Electronics Store	Falafel Restaurant	Food Truck	Food & Drink Shop	Flower Shop	Fast Food Restaurant
	Sushi Restaurant	Pizza Place	Convenience Store	Salad Place	Bakery	Café	Fast Food Restaurant	Flower Shop	Falafel Restaurant
	Warehouse Store	Casino	Liquor Store	Wine Shop	Falafel Restaurant	Food Truck	Food & Drink Shop	Flower Shop	Fast Food Restaurant	Electronics Store
	Plaza	Recreation Center	Liquor Store	Skating Rink	Food Truck	Department Store	Diner	Discount Store	Distribution Center	Dog Run

Results and Discussion





- The optimal number of clusters was 3
- Top 10 venues in each cluster were analyzed
- This plot can be used to suggest valuable information to stakeholders

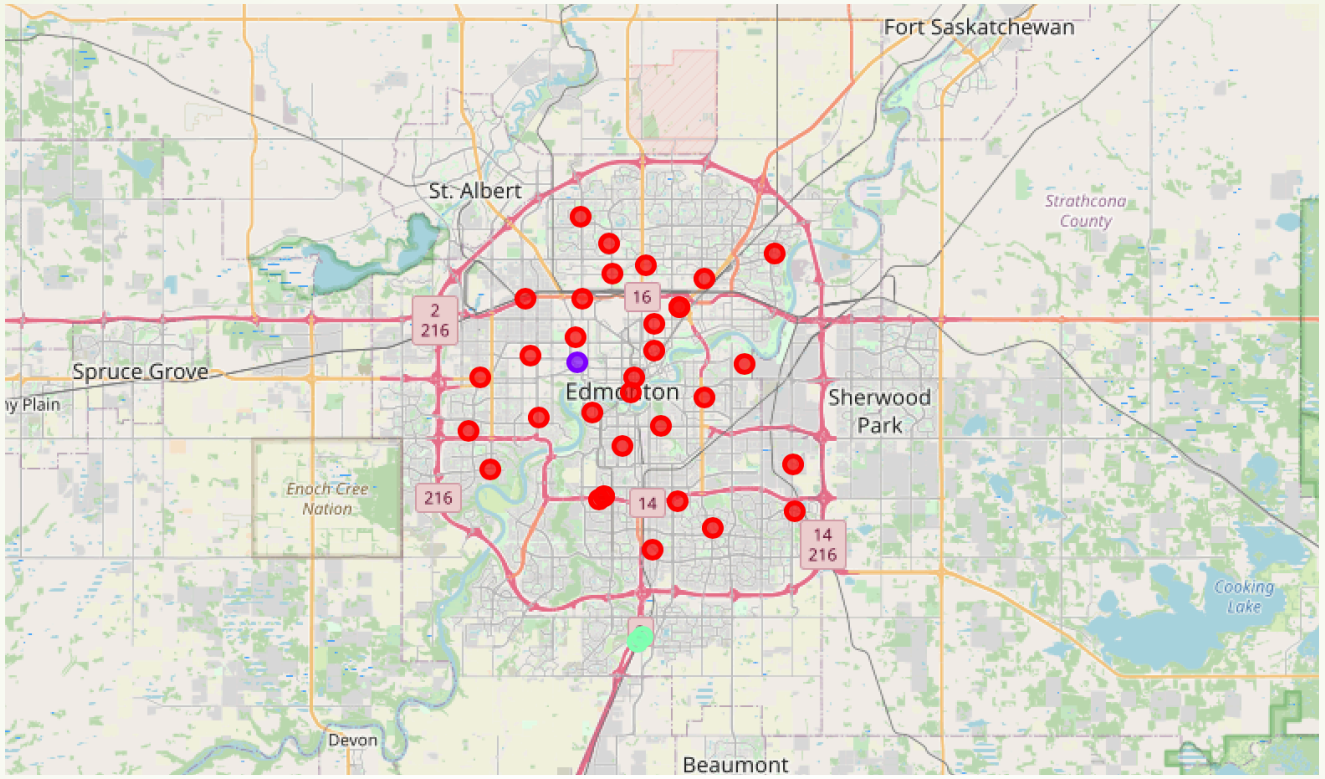
Results and Discussion continue

Let's discuss few examples:

The cluster 1 is situated within the central part of the city, and is saturated with fast food restaurants and some other type restaurants, hence opening one here is not the best choice. I could suggest that this cluster don't need more restaurants, but hotel/hostel, health & beauty shop, or bakery could be a good idea for business.

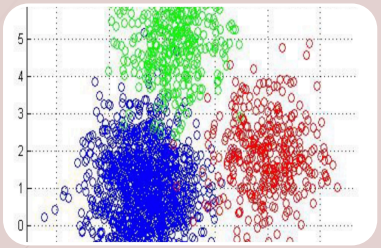
On the cotrary, cluster 3 is situated outside the central part of the city and has one fast food restaurant. Suggestion could be a coffee shop or one more, but another type fast food restaurant. If a deeper analysis could be done, due to location a major project could be suggested, as large shopping mall, indoor climbing park or amusement park for children etc.

Cluster 2 is the smallest one and is located in central part of the city. The most popular venues there are food and wine, so may be cluster 2 is a good place for a bar, pub or disco.



Map of Edmonton with three clusters which can be used to suggest the best location of a new business

Future directions



Better data sources to create a better decision model using KNN clustering

Due to limited API calls, paid account could be used to bypass limitations

For better analysis other factors could be included such as population in each neighborhood and income of residences, the concentration of commercial buildings that could influence the location decision

Conclusion

- I have used different python packages to extract and scrap data from Wikipedia and to visualise it, also Foursquare API was used to explore the venues in neighborhoods
- Final section of the project was to group venues in clusters and define the characteristics for that particular cluster. I have found three clusters:
  - to of which are situated in the central part of the Edmonton,
  - the third cluster is located outside
- A few examples have been discussed
- A map showing the clusters have been provided
- So, all this information can be used by stakeholders to decide the location for the particular type of business

Thank you!