

# Applied Data Science Capstone

## The Battle of Neighborhoods

# 1 Introduction

- Attempt to give protentional stakeholders a better description of the Toronto, Canada neighborhoods.
- For the purpose of recognizing which area/neighborhood is the best for opening a new restaurant.
- Utilized by using data mining technique k-mean clustering in Python.

## 2 Data description

- Information about the areas in Toronto were extracted from different sources.
- Each entry consist of:
  - Postal code - unique identifier
  - Borough name
  - Neighborhood name
  - Latitude
  - Longitude
- Total of 103 locations.

	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494



## 2 Data description

- Next, using the Foursquare API, information about nearby food venues were obtained and saved in a data frame together with neighborhood information.
- Venues were then grouped in a data frame and relative frequencies computed.

----Agincourt----

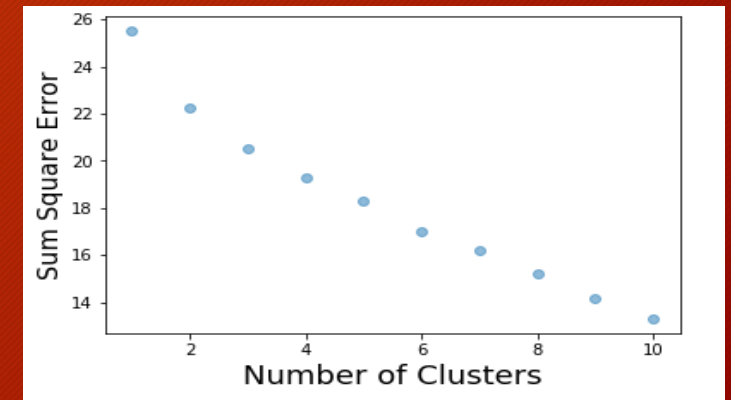
	venue	freq
0	Latin American Restaurant	0.25
1	Sandwich Place	0.25
2	Chinese Restaurant	0.25
3	Breakfast Spot	0.25
4	Afghan Restaurant	0.00

----Adelaide, King, Richmond--

	venue	freq
0	Restaurant	0.08
1	Café	0.08
2	Sandwich Place	0.06
3	Asian Restaurant	0.06
4	American Restaurant	0.05

# 3 Methodology

- $K$ -means clustering algorithm was used to separate our data.
- Number of clusters was determined by performing clustering for multiple values of  $k$  and computing sum of squares error (SSE).
- Elbow method was then used to determine the final number of clusters.
- We used  $k = 3$ , three clusters, for our analysis.

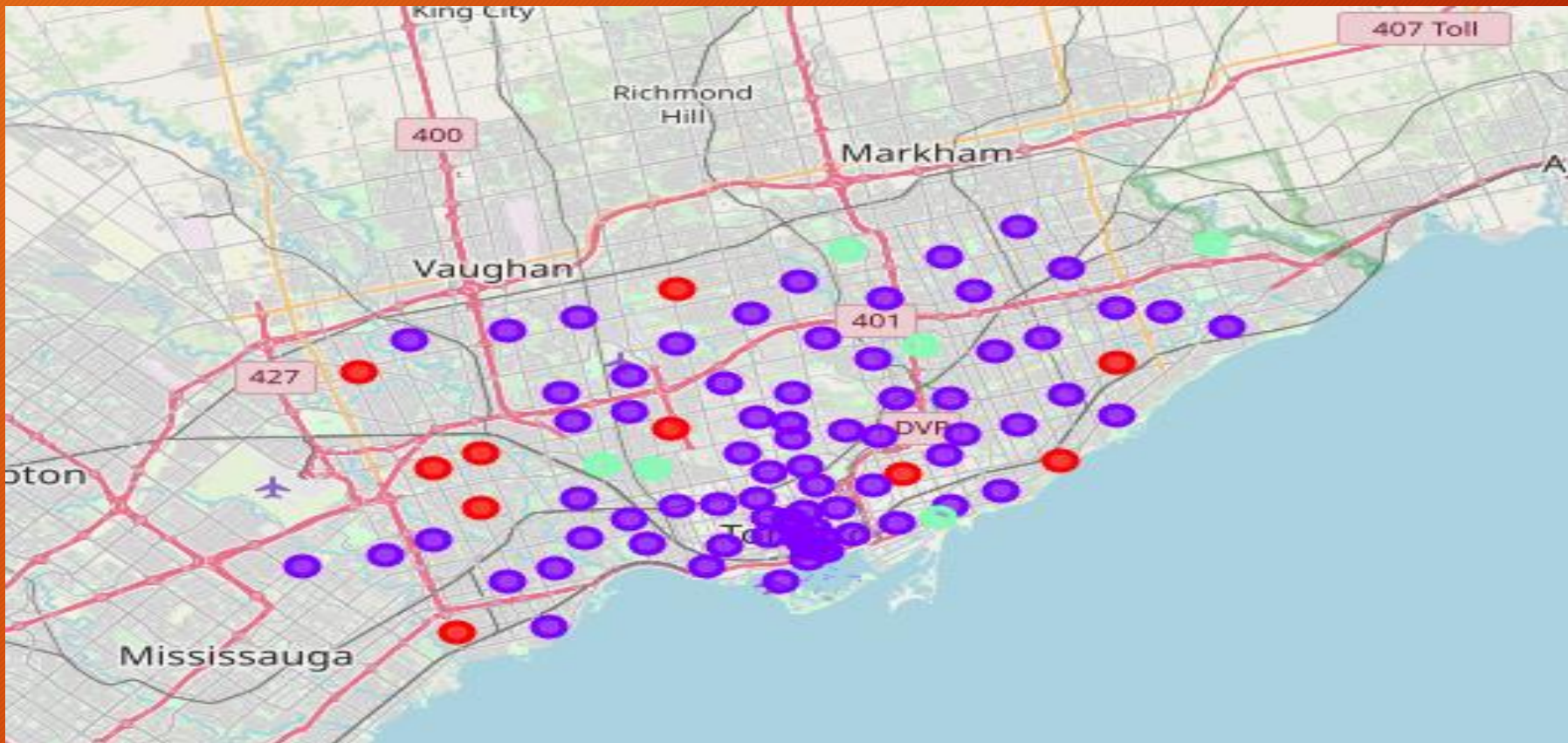


# 4 Results

- Most of the data were assigned to the cluster 1, precisely 83% of locations. Other two clusters 0 and 2 represent 10.5% and 6.5%, respectively.
- We examined the separate clusters and determine the discriminating venue categories that distinguish each cluster:
  - Cluster 0: „Pizza and Fast Food venues” cluster,
  - Cluster 1: „Premium location and high-end places“ cluster,
  - Cluster 2: „Ethnic restaurants” cluster.
- We created a map of Toronto with the marked neighborhoods allocated to their clusters.



# Map of Toronto with clustered neighborhoods



## 5 Discussion

- Different clustering algorithms could be used for this study and could potentially yield different results due to the differences between the neighborhoods.
- Different number of clusters could be used. However, using more clusters seemed unnecessary since when 4 or more clusters were used some clusters only consisted of one location.



## 6 Conclusion

- We were able to cluster Toronto's neighborhoods and derive some characteristics of their food venues.
- Since cluster 1 consisted of majority of locations, further analysis of these neighborhoods would be beneficial for protentional investors in these areas.