Comparison of Toronto Neighbourhoods as Rental Locations

Manan Bhati

Researching city neighbourhoods is important to recent arrivals

* Discovering rental value proposition of city neighbourhoods is important to newly arrived people
* Key considerations are average rent for the type and size of housing they want, distance/ commute time from their work location
* Other important factors would be neighbourhood amenities for shopping, healthcare, recreation, education, etc.

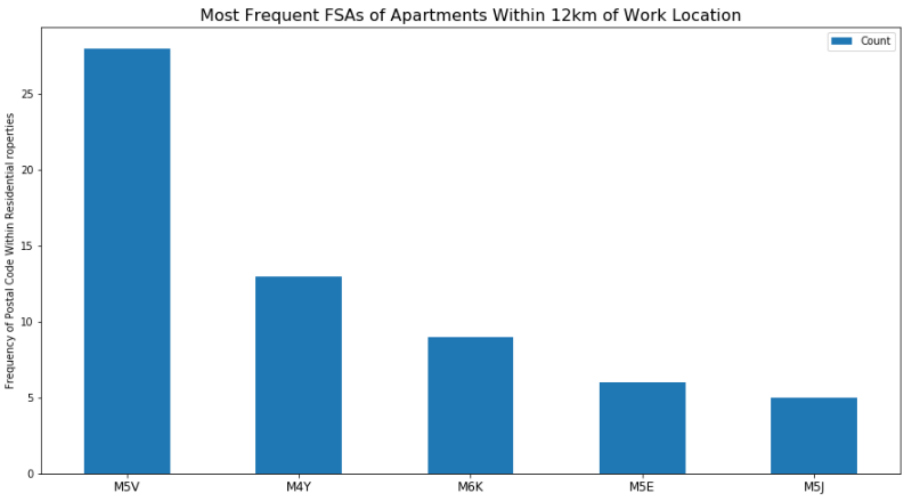
Data acquisition and cleaning

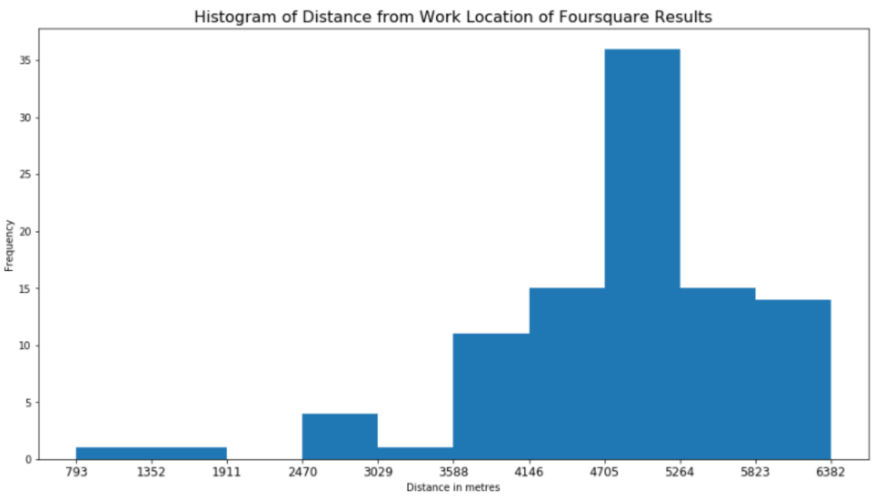
* Name, address, location coordinates, distance and postal code of residential apartments within '*x*' km of work location, from Foursquare location database (100 rows and 8 features)
* Average rent for apartments of different configurations by Toronto neighbourhood, from CMHC, Canada's national housing agency (134 rows and 11 features)
  * Irrelevant rows and columns were dropped
  * Compound neighbourhood names were separated, cleaned up
* Forward Sortation Areas (FSA) (initial 3 characters of postal codes), were extracted
* Final dataset for clustering had 21 unique FSAs, average rent for a 2BR apartment by FSA, average distance from work location, number of apartment buildings by FSA, and FSA location coordinates

Key Data Preparation Challenge

- Toronto [FSAs](#) and official [neighbourhoods](#) are not the same and do not map one-on-one. They are useful proxies for one another, however
  - An FSA may overlap adjacent neighbourhoods and vice-versa
  - Rent statistics maintained by CMHC are aligned with neighbourhoods
  - Location data providers like Foursquare and OpenStreetMap provide postal codes (FSAs)
    - 45 of the 100 residential buildings within 12km of work location, provided by Foursquare, did not have any postal code (FSA) tagged
    - Missing FSAs had to be fetched by providing name and address information to OpenStreetMap database through Nominatim geocoder
  - Neighbourhood names in CMHC table are often aggregates of adjacent neighbourhoods and do not line up squarely with official neighbourhood names
    - Parsed neighbourhood name components were passed on to OpenStreetMap geocoder service to retrieve associated postal codes
  - FSAs are the key to linking rent data table and residential properties data table

Exploratory Data Analysis - I





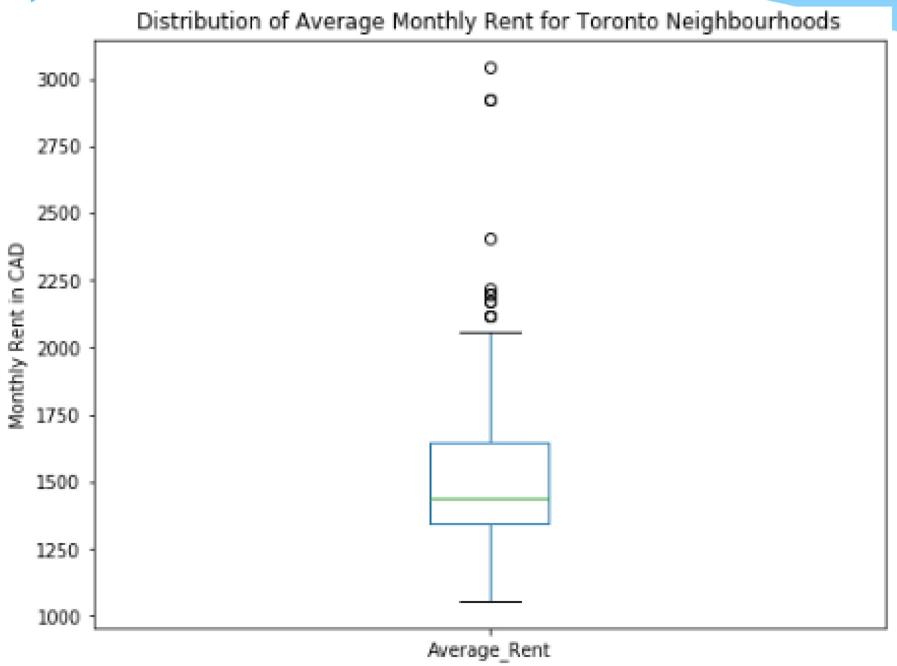
Of 26 FSAs associated with 98 residential buildings,

- M5V is heavily represented (28 times)
- 10 FSAs occur only once, 8 occur twice

Foursquare returns maximum 100 venues:

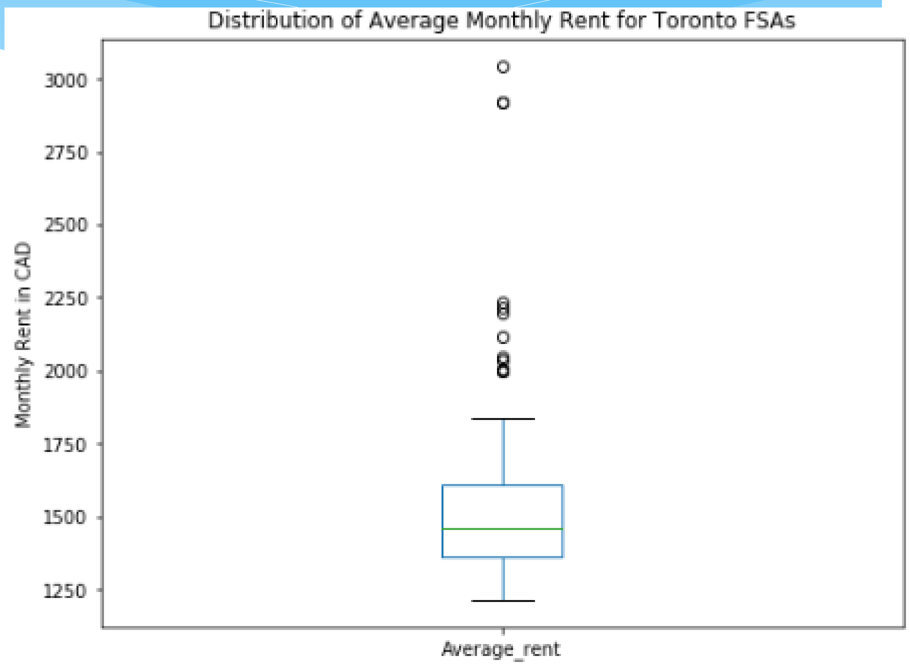
- 100 residential buildings were found within a radius of 6382 metres

Exploratory Data Analysis - II



Rent figures as they exist in CMHC table for 168 neighbourhood names

- Range is from CAD 1,055 to CAD 3,047
- Median = CAD 1,437, Mean = CAD 1,555



Rent figures for 90 FSAs: FSA average rent is mean of rent of neighbourhoods tagged to an FSA

- Range is from CAD 1,213 to CAD 3,047
- Median = CAD 1,460, Mean = CAD 1,573

Final Dataset for Unsupervised ML Modeling

	Postal Code	Neigh_Latitude	Neigh_Longitude	Number of Properties	Average_rent	DistanceFromOffice_y
0	M6P	43.661608	-79.464763	2	2005.000000	1202.000000
1	M6K	43.636847	-79.428191	9	1509.428571	3754.444444
2	M5T	43.653206	-79.400049	2	2004.000000	3837.500000
3	M5V	43.636039	-79.397400	28	2920.000000	4749.392857

- 21 FSAs representing 90 apartment buildings within 12 km of work location, for which average distance and average monthly rent could be mapped
- Average rent for an FSA is the mean of monthly rent figures of all CMHC neighbourhoods partially or fully tagged to that FSA
- Average distance is the mean of distances of all residential apartment buildings tagged to an FSA
- Average rent and average distance from work will be the two features to be used for cluster modeling of FSAs

K-Means clustering for Toronto FSAs

- K-means is an efficient clustering algorithm and quickly converges to a local optimum
- Optimum number of clusters '*k*' can be determined by examining elbow point of a plot of inertia (sum of squared distances of samples to their closest cluster centre) vs. '*k*'
- Setting number of iterations with random centroid initializations to high (>20) mitigates inherent variance in results generated by k-means algorithm
- Both features - Average rent and Distance from work – are standardized
- Our modeling penalizes average rent values more compared to distance from work. Two ways to model this:
  - Assign 80:20 weights to values of average rent and distance from work ✓
  - Derive a composite score as:

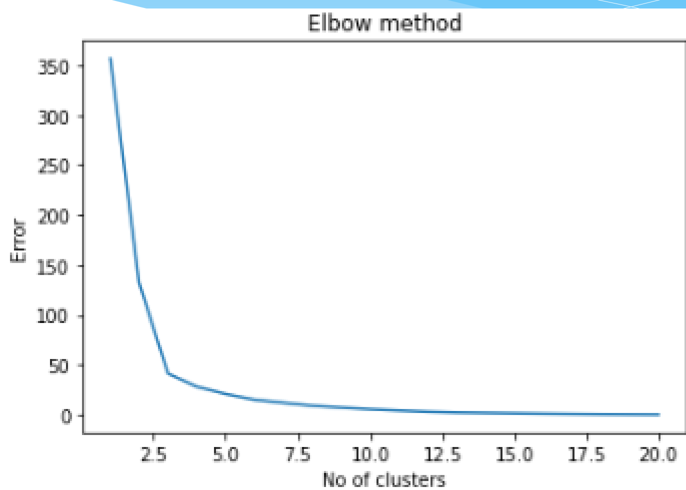
*Composite Score = 0.8 * Standardized Rent Value + 0.2 * Standardized Distance*

and then use a univariate grouping algorithm like Jenks optimization

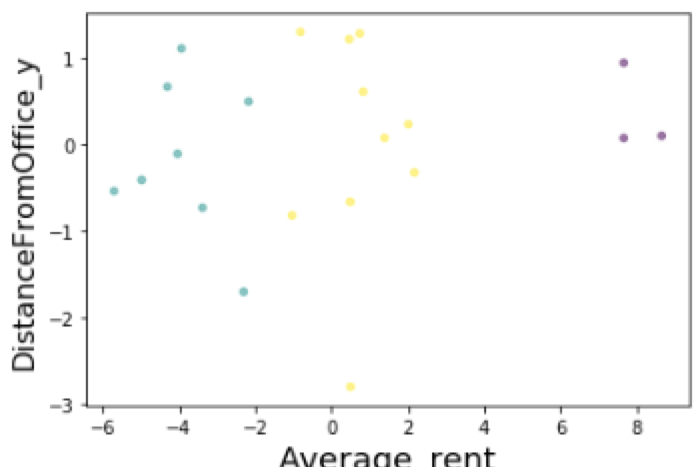
K-Means Results

	Postal Code	Average_rent	DistanceFromOffice_y	Cluster Labels
0	M6P	2005.000000	1202.000000	0
1	M6K	1509.428571	3754.444444	1
2	M5T	2004.000000	3837.500000	0
3	M5V	2920.000000	4749.392857	2
4	M4W	2118.000000	4751.000000	0
5	M5S	3047.000000	4782.000000	2
6	M4Y	1664.000000	5272.076923	1
7	M5G	2198.000000	4946.500000	0
8	M5B	2046.500000	5413.000000	0
9	M5J	2920.000000	5824.800000	2
10	M5E	1439.000000	6030.666667	1
11	M4S	2035.000000	6245.000000	0
12	M5A	1838.000000	6265.500000	0
13	M6R	1647.500000	2560.000000	1
14	M5P	1809.500000	3644.000000	0
15	M8V	1213.000000	3992.000000	1
16	M6L	1305.666667	4151.000000	1

Sample observations with 3 cluster labels (unstandardized)



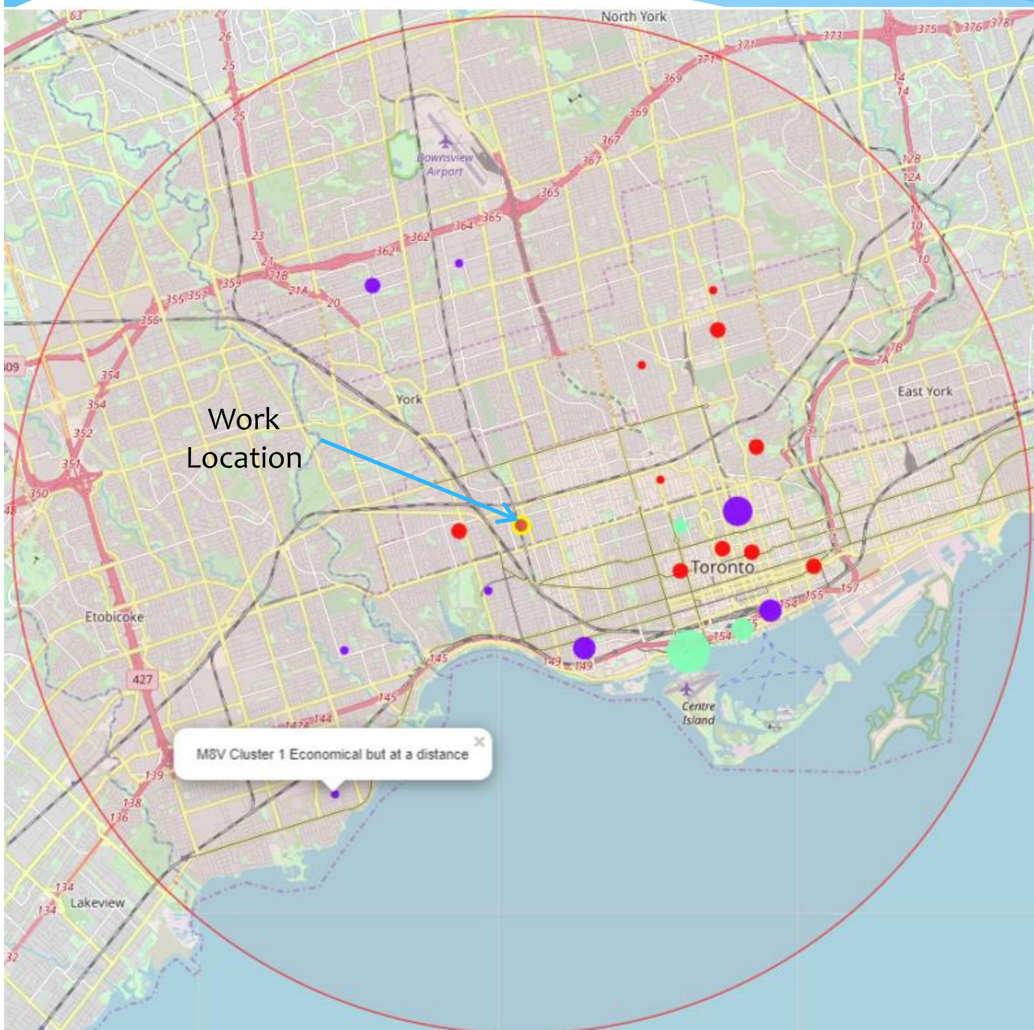
'Elbow' (sharp inflexion point) lies at a value of k=3



Clusters with standardized features:

- *Cluster 2*
- *Cluster 1*
- *Cluster 0*

Cluster Visualization on Toronto Map



Cluster Description	Circle Marker Colour
'Downtown Experience at a Steep Price' (2)	
'Mid-priced Experience in Popular Neighbourhoods' (0)	
'Economical but at some distance' (1)	

Size of FSA location markers is related to the number of residential properties discovered in that FSA

FSA M5V has the most properties (28) of any FSA

Conclusion

- **Aim:** to group Toronto neighbourhoods/ FSAs based on average rent for a 2BR apartment and distance from the target user's work location in Toronto
- **Modeling**: unsupervised learning through k-means clustering. Greater weightage assigned to average rent feature in forming clusters
- **Accuracy** of the clusters is heavily dependent upon the reliability of postal code mappings to venues in location data providers like Foursquare and OpenStreetMap
- **Improvement** avenues: include more data related to the attractiveness of a neighbourhood, e.g. easy neighbourhood amenities for shopping, recreation, etc., number of available residential units, etc.