IBM Data Science Capstone - Report

1. Introduction

My Capstone project deals in some way the the globalisation of work forces. These times people have become quite flexible and it is common to go to a different country to start a new job. **But how do you find the right neighbourhood in a city you do not know?**

Well, you certainly know the neighbourhoods in the city in which you are currently living. You know which mix of **features** (shopping, restaurants, entertainment, parks, public transport, education...) you are looking for. Thus you can provide **a point of reference** to find the right place somewhere else.

For simplicity I have chosen Toronto as the (from week 3 of the course) known city of interest. A neighbourhood in Zurich, Switzerland, will be the point of reference. Obviously both choices (origin & destination) are exchangeable. Thus my analysis can give some guidance for anyone who has to move to a new city.

Problem: *Find feasible neighbourhoods in Toronto based on your preferences?*

2. Data

2.1. Data Sources

The analysis will make use of different data sources:

- Foursquare-API venue data to derive the neighborhood's features
- List of neighbourhoods and housing prices from Toronto's 'Open Data Portal': https://open.toronto.ca/
- Crime statistics using Toronto's 'Police Service Open Data Portal': http://data.torontopolice.on.ca/pages/open-data

2.2 Data Wrangling

For the files from open data portals provided by Toronto authorities the data wrangling process was rather straightforward. Luckily all 3 files were available as csv, included the same number of neighborhoods and used the same numeric code to index them. Consequently most of the work came down to merging the 3 files on the index and dropping columns afterwards.

There was one additional step with respect to the crime statistic. Four different types of crime were given in total numbers. I decided to sum them up and then divide by neighbourhood population to derive a figure illustrating *crimes per capita*.

Things turned out to be more complicated for the data derived by the Foursquare-API. Generally I found it challenging to extract information from json-files (but I learned a lot with respect to nested lists and dictionaries!).

A special and high hurdle was the use of Foursquare's main categories. There are only 10 of them. Unfortunately they are deeply nested in the *prefix* within the json-file (...frankly I looked around but there was no table to be found mapping the sub categories to the main categories)...

```
'categories': [{'id': '4bf58dd8d48988d1ca941735',
  'name': 'Pizza Place',
  'pluralName': 'Pizza Places',
  'shortName': 'Pizza',
  'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/pizza_'
```

...and the path to access the piece of information turned out to be a bit complicated:

```
v['venue']['categories'][0]['icon']['prefix'].split('/')[-2]) for v in results])
```

3. Methodology

3.1. General Steps

(A) Perform a preselection of neighbourhoods in Toronto based on crime rates and housing prices

(B) Derive the characteristics of the Zurich neighbourhood (point of reference) with help of Foursquare-API

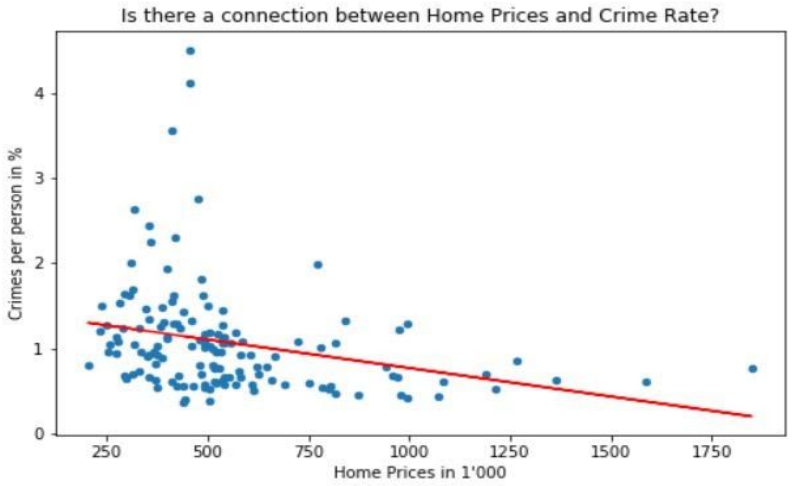
(C) Similarly build a table of feature vectors (columns) for the Toronto neighbourhoods using Foursquare-API

(D) Define and add variations of the Zurich neighbourhood (eg more 'nightlife')

(E) Use k-means clustering to find similar (with respect to the point of reference) neighbourhoods in Toronto

3.2 Exploratory Data Analysis

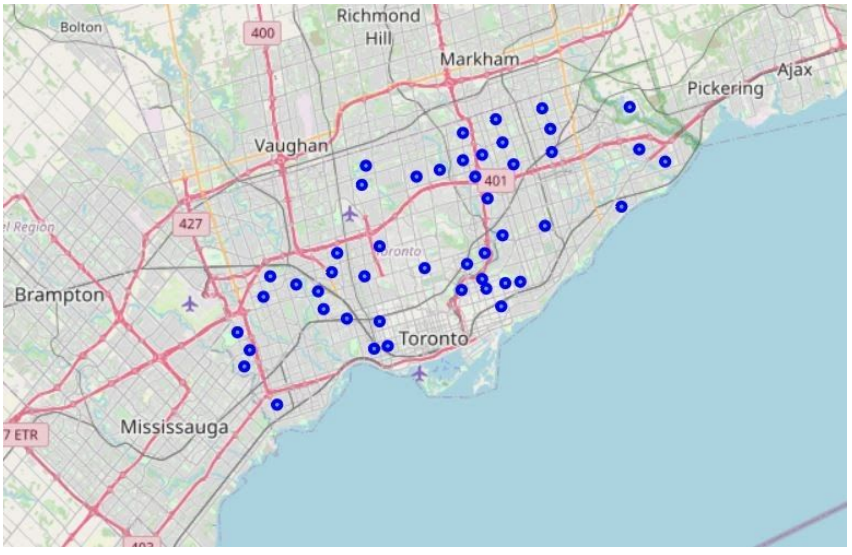
(A) The *preselection* process - ie reducing the number of neighbourhoods - is done with same stats and (always helpful) visualisations.



The scatter plot points to the area on the left-hand-side below the regression line and with the help of the standard stats table...

	crime_rate	Home Prices
count	140.000000	140.000000
mean	1.069860	548.193407
std	0.643798	267.667427
min	0.365512	204.104000
25%	0.652040	374.964500
50%	0.958526	491.210000
75%	1.242825	590.216000
max	4.504400	1849.084000

...the thresholds can be chosen as (below) 1% for crime rates and (below) 600'000 for home prices. The set of neighbourhoods is reduced from 140 to 47!



(D) Defining the *variations of the Zurich neighbourhood* was rather done by hand.

	Neighbourhood	arts	building	food	nightlife	parks	shops	travel
0	Zurich	0.056818	0.022727	0.488636	0.056818	0.056818	0.204545	0.113636
1	Zurich_night	0.030000	0.020000	0.400000	0.200000	0.050000	0.200000	0.100000
2	Zurich_shop	0.050000	0.020000	0.350000	0.050000	0.050000	0.400000	0.080000
3	Zurich_art	0.200000	0.020000	0.300000	0.050000	0.150000	0.180000	0.100000

With the corresponding stats for the Toronto neighbourhoods and a bit of personal taste the new vectors were derived.
Note that there were only 7 out of 10 main categories to be found in the data set of our Zurich point of reference. But the main categories 'event' and 'education' were as well rarely found in the Toronto data set. Thus omitting them seemed to be a reasonable choice.

(E) Finally the k-means algorithm was chosen to cluster the 4 Zurich and 47 Toronto neighbourhoods. The number of clusters were defined as 4, since (ideally) one of each Zurich neighbourhood should be found in a different cluster. In the end cluster 2 was not populated...

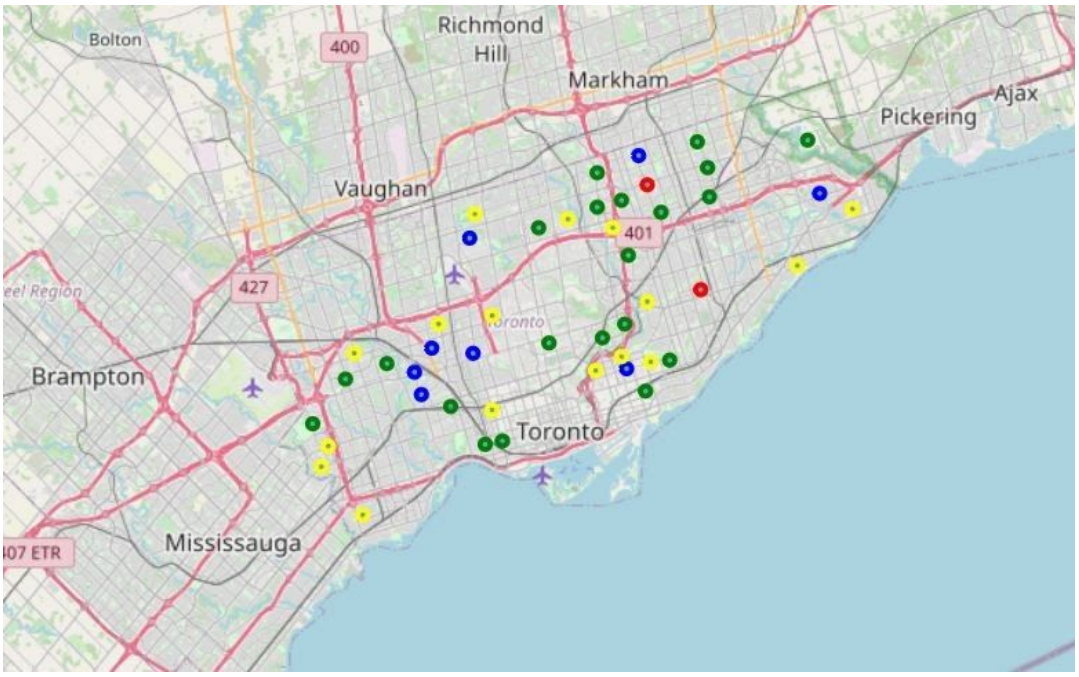
	Neighbourhood	Cluster Labels	arts	building	food	nightlife	parks	shops	travel
0	Zurich	0	0.056818	0.022727	0.488636	0.056818	0.056818	0.204545	0.113636
1	Zurich_night	0	0.030000	0.020000	0.400000	0.200000	0.050000	0.200000	0.100000
2	Zurich_shop	3	0.050000	0.020000	0.350000	0.050000	0.050000	0.400000	0.080000
3	Zurich_art	1	0.200000	0.020000	0.300000	0.050000	0.150000	0.180000	0.100000

...and the distribution of the 51 points into clusters might be something to improve on!

Cluster Labels	Neighbourhood	arts	building	food	nightlife	parks	shops	travel
0	23	23	23	23	23	23	23	23
1	9	9	9	9	9	9	9	9
2	2	2	2	2	2	2	2	2
3	17	17	17	17	17	17	17	17

4. Result

...a picture is worth a thousand words!



The **blue** (label = 0) cluster includes neighbourhoods which are balanced in food, shops and parks. Those neighbourhoods might be nice if you opt for a bit more of nature. The **green** (label = 1) cluster is closest to the original Zurich area, tilted a bit to food and nightlife. The **red** (label = 2) cluster consists mainly of parks. Finally the **yellow** (label = 3) cluster is not so different in characteristics from the blue one, except there are more shops instead of parks.

5. Discussion & Conclusion

Even though there is a clear red thread to the analysis it is obviously somewhat superficial. There are a few points in the process where one might want to dig deeper. Personally I think of questions like:

- Should I differentiate by the type of crime (assault, auto theft, robbery....)?
- Are trends for home prices and crimes more meaningful than the 2019 snapshot?
- Both (home prices and crimes) are predominantly important aspects for a (new) citizen. Nonetheless, could they be integrated in a cluster analysis in a meaningful way?
- How could a more objective or even an automated process to derive variations of your point of reference look like?
- And finally, knowing that my point of reference is in a cluster: How do I get the 'distances' of the other cluster points to it?

Ok, the outset of the analysis was to get a better understanding of the new city and narrow choices down. This is done in a meaningful way. But the neighbourhoods might look very different from your expectations when you visit for the first time. To conclude:

Never forget, there is data science and there is the real live!