Capstone project- The Battle of Neighbourhoods

Submitted by- Ashwin Thirumala Kumara

Table of Contents

Introduction

Suppose that a person of South Asian origin wishes to immigrate to Toronto, Canada. Being from a different culture, expectations and baseline requirements for lifestyle differ widely from those in Toronto. Three questions occur to the immigrant's mind:

1. What would be the "new normal", the anticipated new baselines for living in Toronto?
2. If they were to get a job in Toronto, which neighbourhoods should they prefer to stay in?
3. What are some correlations in the data they should be aware of?

How would a person evaluate these questions? Fortunately, there is a plenty of data made available by the following sources:

Data sources

We will address the sources for data pertaining to each question:

1. **Foursquare** for Toronto Venues data- This was used to inform the venues in each Toronto neighbourhood.
2. **Wellbeing Toronto**'s NHS Demographics Indicators, 2010.
3. **Wellbeing Toronto**'s 2011 data from the **Open Data Catalogue, City of Toronto** for the following data:
   - Economic data- No. of Businesses, Home Prices (CAD), Social Assistance Recipients (nos.),
   - Traffic data- Road Volume (nos.)
   - Environment data- Tree Cover
   - Safety data- Total Major Crimes, Vehicle Thefts (nos.)
   - Demographics data- Population, Total Visible Minority, S.Asian, Recently Moved S.Asians, No. in Labour Force, Unemployed, Renters, Major repairs needed, (All in nos.), shelter30 (% of owner households spending 30% or more of household total income on shelter costs), Avg. Monthly Rent (CAD), Median After-tax Income (CAD)
4. **Toronto GeoJSON** from [https://github.com/jasonicarter/toronto-geojson], to help generate the Toronto choropleth maps.

The Wellbeing Toronto data, which serves as our baseline, is described as below:

Source	Data selected	Unit
Wellbeing Toronto- Economics (2011)	'Businesses',	Nos.
	'Home Prices',	CAD
	'Social Assistance Recipients'	Nos.
Wellbeing Toronto- Transportation (2011)	'Road Volume'	Nos.
Wellbeing Toronto- Environment (2011)	'Tree Cover’	Sqm.
Wellbeing Toronto- Safety (2011)	'Total Major Crime Incidents'	Nos.
	'Vehicle Thefts’	Nos.
Wellbeing Toronto NHS Demographics (2010)	'Population'	Nos.
	'Total Visible Minority'	Nos.
	'S. Asians'	Nos.
	'Recently Moved S.Asians'	Nos.
	'Labour Force'	Nos.
	'Unemployed'	Nos.
	'Renters'	Nos.
	'Major repairs needed'	Nos.
	'shelter30'	%
	'Avg. Monthly Rent'	CAD
	'Median After-tax Income'	CAD

Of this data,

- Median After-tax income seemed only ~7,000 CAD less than Median pre-tax income, whose reason is unknown.

Data cleaning

This section lists the data cleaning undertaken:

- All column types were recast to numeric from object types, in order to enable descriptive statistics on them.
- Neighbourhood no. 93 spelling was corrected to “'Dovercourt-Wallace Emerson-Junction”.
- In order to enable compatibility with the GeoJSON data source, Neighbourhood 17 was renamed to 'Mimico (includes Humber Bay Shores) (17)'.
- In order to enable compatibility with the GeoJSON data source, Neighbourhood 59 was renamed to 'Danforth East York (59)'.
- No other important data cleaning actions were necessary.

Methodology

The following methodology was taken:

- Step-1- Generating the dataframe by joining data sources. The generation of the dataframe is detailed in the notebook, with data sources and necessary data cleaning actions listed above.
- Step-2- Utilizing Foursquare API to include Venues data. This is fully described in the Notebook. With the merging of Venue-based data to our baseline dataframe, our database of neighbourhoods is loaded with venue data, based on a filter chosen

as applicable for South Asians. It is ready to be clustered upon, but before that, we conduct a preliminary exploratory data analysis.

- Step-3- Exploratory data analyses, inferential statistics: The results of these are elaborated in the “Results” section.
- Step-4- Clustering on the basis of “significant” parameters using k-means clustering:

In order to conduct clustering on the neighbourhoods, we need to identify parameters of significance to the immigrant. For our project, we select 5 parameters upon which to cluster all Toronto neighbourhoods into 4 clusters. The parameters chosen for this purpose are:

1. No. of Crimes, as an immigrant would prefer to stay in a neighbourhood with lower incidence of crime.
2. No. of South Asians, as a South Asian immigrant may find it conducive to live in a neighbourhood with more members of their community.
3. Avg. Monthly rent, as anyone would prefer to stay in an area with lower rent (and Toronto is unfortunately well-known for high rent),
4. shelter30, Which was selected as a sample measure of the amount towards shelter an owner in Toronto neighbourhood would spend, and
5. Venue of interest to the Immigrant. Based on Foursquare API response of venues at Toronto, out of a total of 271 venue categories, Indian/Pakistani restaurants are selected as sample venues of interest to a South Asian. This selection can be reworked on a case-basis for specific customers.

In the clustering activity, we use the dataframe with venue details created. We select the neighbourhoods significant to our optimization routine, as done above, one-hot encode them and use it in our mapping next step.

For clustering, we use k-means clustering we select 4 as our number of clusters. The clustering activity was carried out after normalizing the data to ensure efficient clustering. Hence, the clustering activity yielded the below results for clusters:

```
Cluster No. 3 (Count= 61)
Crime : 358.1311475409836
S. Asians : 1231.72131147541
Avg. Monthly Rent : 920.5737704918033
Venue Count : 0.0
Shelter30 : 28.34426229508197
----------------------------------------------
Cluster No. 2 (Count= 24)
Crime : 521.5416666666666
S. Asians : 3654.5833333333335
Avg. Monthly Rent : 1081.9583333333333
Venue Count : 0.125
Shelter30 : 34.083333333333336
----------------------------------------------
Cluster No. 1 (Count= 29)
Crime : 195.27586206896552
S. Asians : 702.5862068965517
Avg. Monthly Rent : 1196.0689655172414
Venue Count : 0.034482758620689655
```

```
Shelter30 : 19.95862068965517
---------------------------------------------
Cluster No. 0 (Count= 26)
Crime : 351.2307692307692
S. Asians : 1281.923076923077
Avg. Monthly Rent : 998.5769230769231
Venue Count : 1.0384615384615385
Shelter30 : 26.684615384615384
---------------------------------------------
```

The numbers of neighbourhoods in each cluster is comparable. Because the clustering was done on the above-selected parameters, cluster elements will be similar to each other rather than different. Thus, the statistics above enables comparison between clusters:

- Cluster 3 is least favourable with 0 mean Venues.
- Cluster 2 is less favourable than 0 or 1 due to higher Crime.
- With both overall lesser Rent and higher S. Asians to amalgamate into the community, Cluster 0 is chosen the most favourable.

This prompts a convenient ordering: 0 > 1 > 2 > 3

Note: It is true that both Crime and Shelter30 are higher for Cluster 0. This prompts a certain level of caution to be maintained in living in these areas.

- Step-5- Mapping the clusters: A function was defined to generate a choropleth map of Toronto. This is detailed fully in the notebook.

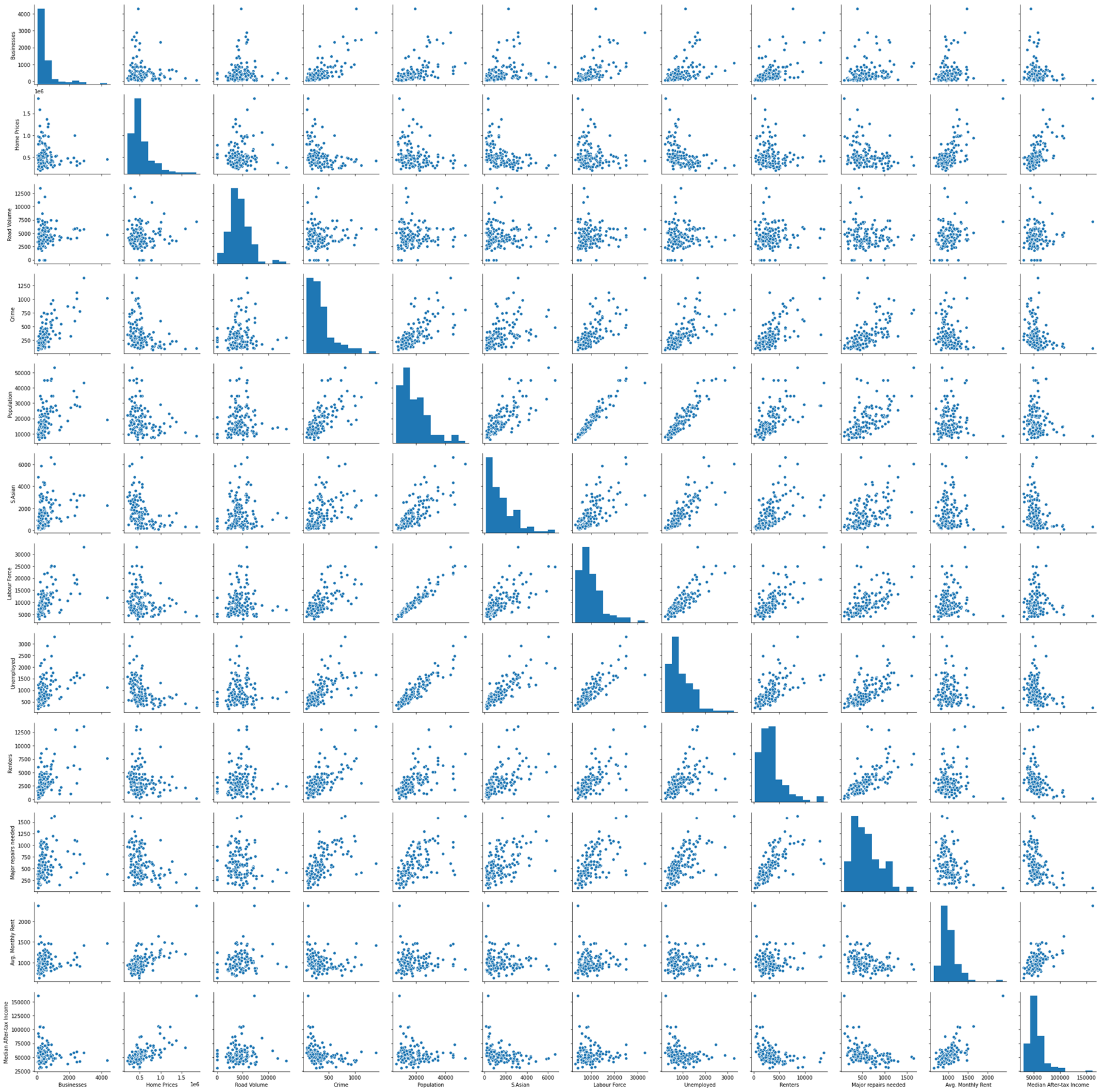
Results ¶

Three questions were posed at the Introduction as follows:

1. **What would be the "new normal", the anticipated new baselines for living in Toronto?** The descriptive statistics on some basic neighbourhood parameters (including **Population, Income, Avg. Rent**) were conducted, and some of these are reproduced in the table below:

	Population	Median Income	Avg. Monthly Rent	Home Prices (CAD)
mean	18677	55427	1020	548193
std	9099	16118	220	267667
min	6490	30794	631	204104
25%	11851	46690	879	374965
50%	16368	52660	973	491210
75%	22410	59963	1125	590216
max	53350	161448	2388	1849084

The pairwise plot below shows in great visual detail the density of points around the mean values, and enhance our idea of expected values for these:



The pairwise plot needs to be zoomed to be read properly- each variable is plotted in a scatter-plot against each other. The diagonal plots indicate the histograms of each variable. Variable names are listed along the axes.

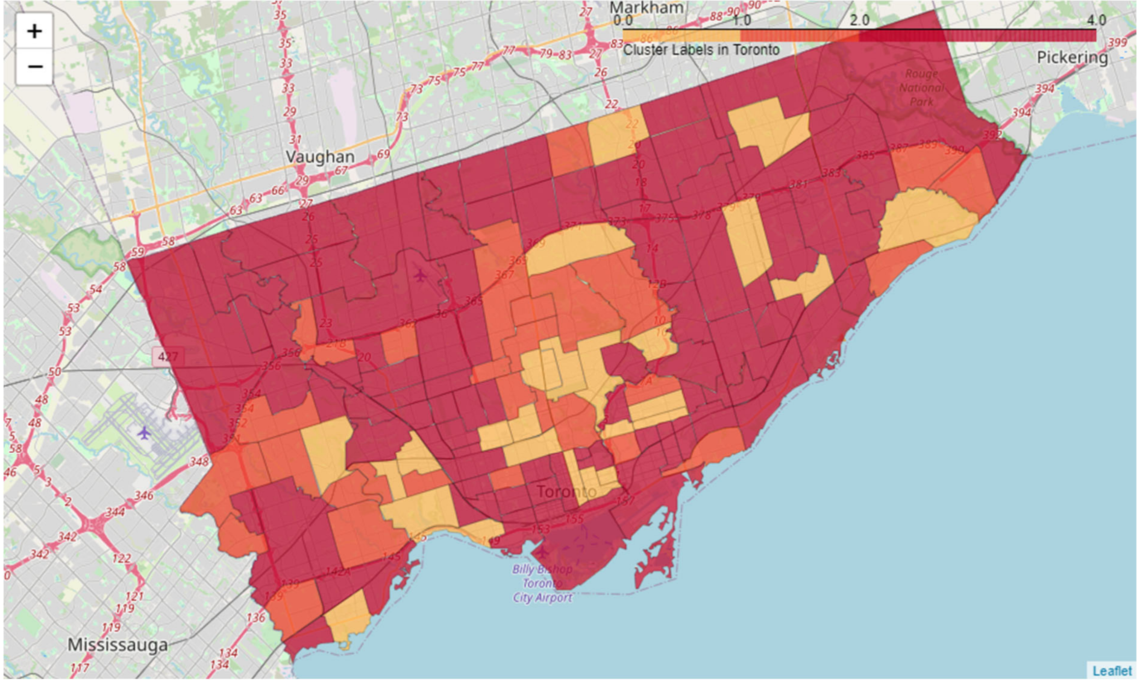
The pairwise plot yields the pictorial visualization of the relationship between variables. Broadly, the variables are either linearly or inversely related to each other in a predictable or justifiable manner.

To support the pairwise plot with numbers, the **pairwise correlation** table is generated:

	B	HP	SAR	RV	TC	C	VT	P	TVM	SA	RSA	L	U	R	MR	S30	AMR	MAI
Businesses	1.0	0.0	0.2	0.1	0.0	0.5	0.4	0.4	0.2	0.3	0.2	0.5	0.4	0.4	0.2	0.2	0.1	0.0
Home Prices	0.0	1.0	-0.5	0.0	0.1	-0.3	-0.3	-0.2	-0.5	-0.3	-0.4	-0.1	-0.4	-0.1	-0.2	-0.4	0.4	0.4
Social Assistance Recipients	0.2	-0.5	1.0	0.0	0.0	0.6	0.4	0.4	0.5	0.5	0.5	0.4	0.6	0.4	0.5	0.4	-0.4	-0.5
Road Volume	0.1	0.0	0.0	1.0	0.1	0.0	0.1	0.1	-0.1	0.1	0.0	0.1	0.1	0.1	-0.1	0.0	0.2	0.0
Tree Cover	0.0	0.1	0.0	0.1	1.0	0.0	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.0	0.0	-0.2	0.2	0.3
Crime	0.5	-0.3	0.6	0.0	0.0	1.0	0.5	0.5	0.4	0.4	0.4	0.5	0.6	0.5	0.5	0.3	-0.2	-0.3
Vehicle Thefts	0.4	-0.3	0.4	0.1	0.2	0.5	1.0	0.5	0.4	0.4	0.3	0.5	0.5	0.3	0.2	0.2	0.0	-0.1
Population	0.4	-0.2	0.4	0.1	0.3	0.5	0.5	1.0	0.4	0.6	0.4	0.9	0.8	0.5	0.5	0.2	0.0	-0.1
Total Visible Minority	0.2	-0.5	0.5	-0.1	0.2	0.4	0.4	0.4	1.0	0.5	0.8	0.4	0.6	0.2	0.3	0.3	-0.2	-0.2
S.Asian	0.3	-0.3	0.5	0.1	0.2	0.4	0.4	0.6	0.5	1.0	0.6	0.5	0.7	0.5	0.4	0.4	0.0	-0.3
Recently Moved S.Asians	0.2	-0.4	0.5	0.0	0.2	0.4	0.3	0.4	0.8	0.6	1.0	0.4	0.5	0.3	0.3	0.3	-0.2	-0.3
Labour Force	0.5	-0.1	0.4	0.1	0.2	0.5	0.5	0.9	0.4	0.5	0.4	1.0	0.7	0.5	0.5	0.2	0.1	0.0
Unemployed	0.4	-0.4	0.6	0.1	0.2	0.6	0.5	0.8	0.6	0.7	0.5	0.7	1.0	0.5	0.5	0.3	-0.1	-0.3
Renters	0.4	-0.1	0.4	0.1	0.0	0.5	0.3	0.5	0.2	0.5	0.3	0.5	0.5	1.0	0.6	0.3	0.0	-0.4
Major repairs needed	0.2	-0.2	0.5	-0.1	0.0	0.5	0.2	0.5	0.3	0.4	0.3	0.5	0.5	0.6	1.0	0.2	-0.2	-0.3
shelter30	0.2	-0.4	0.4	0.0	-0.2	0.3	0.2	0.2	0.3	0.4	0.3	0.2	0.3	0.3	0.2	1.0	-0.1	-0.5
Avg. Monthly Rent	0.1	0.4	-0.4	0.2	0.2	-0.2	0.0	0.0	-0.2	0.0	-0.2	0.1	-0.1	0.0	-0.2	-0.1	1.0	0.4
Median After-tax Income	0.0	0.4	-0.5	0.0	0.3	-0.3	-0.1	-0.1	-0.2	-0.3	-0.3	0.0	-0.3	-0.4	-0.3	-0.5	0.4	1.0

With this, we have the inter-relationships between different variables of interest and better understand how expensive it would be, what salary should we expect to target while in Toronto, how crowded it will be, what the distribution of all the studied variables are in different neighbourhoods, and quick access to all this information.

2. **If they were to get a job in Toronto, which neighbourhoods should they prefer to stay in?** A choropleth map of Toronto neighbourhoods was generated, with more favourable areas showing in lighter colour.



In the map generated, the lighter-shaded areas are those to be preferred by a South Asian immigrant, with progressively darker-shaded areas indicating unfavourability, as based on the clustering above.

The details of how this map was obtained are described in the above “Clustering” section of the report, and attached notebook.

3. **What are some correlations in the data they should be aware of?** From studying the pairwise plot and pairwise correlations, some correlations are listed as below:

    - Home prices and neighbourhood population are negatively correlated. Living in less-crowded areas comes with a premium!
    - Broad positive correlation between Neighbourhood population, Crime, 'Major repairs needed' with inverse correlation to median household income.
    - As far as living in neighbourhoods goes, Higher the income, higher the rent.

Discussion

A data science project is as good as the data sources and methodology of analysis that generate it. The following are the key areas of discussion in this study:

- Geolocator data for which Lat-Long pairs are not mapped could not be used to generate Venues data from Foursquare. In this study, we considered 5 parameters important to a South Asian immigrant- No. of South Asians in a neighbourhood, Crime, Avg. Monthly Rent, Shelter30, and No. of Venues. Therefore, some neighbourhoods would be preferentially clustered into a less favourable cluster. This can be fixed by manually fixing Lat-Long co-ordinates, but is not done in the present study.
- The 5 parameters chosen were Crime, S. Asians count, Avg. Monthly rent, Venue count, Shelter30 may be varied on a need-basis with further model-tuning.
- From exploratory data analysis in this project, it becomes clear that most variables are related in approximately linear or inverse manners. With more data and more segmentation within the data, deeper connections in data may be unearthed.
- The data is from 2010-2011, which makes it dated. However, the data is used to compare between neighbourhoods, which may not lead to very different outcomes with studying latest because of the magnitude of these differences.
- The functions developed in this notebook are easy to adapt to other analyses.

Conclusion

The capstone project was undertaken to answer the questions raised in the Introduction, with sharing results and discussion on how the study may be made more effective.

Acknowledgement

The author acknowledges IBM, Coursera and my many peers I interacted with in this online course for an effective learning program.