Battle of the Neighbourhoods

A Coursera Capstone project

By Ashwin Thirumala Kumara

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:
    - **What would be the "new normal", the anticipated new baselines for living in Toronto?**
    - **If they were to get a job in Toronto, which neighbourhoods should they prefer to stay in?**
    - **What are some correlations in the data they should be aware of?**
- How would a person evaluate these questions?

Data Sources

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

- **Foursquare** for Toronto Venues data- This was used to inform the venues in each Toronto neighbourhood.
- **Wellbeing Toronto**'s NHS Demographics Indicators, 2010.
- **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)
- **Toronto GeoJSON** from [https://github.com/jasonicarter/toronto-geojson], to help generate the Toronto choropleth maps.

Data Sources

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

Methodology

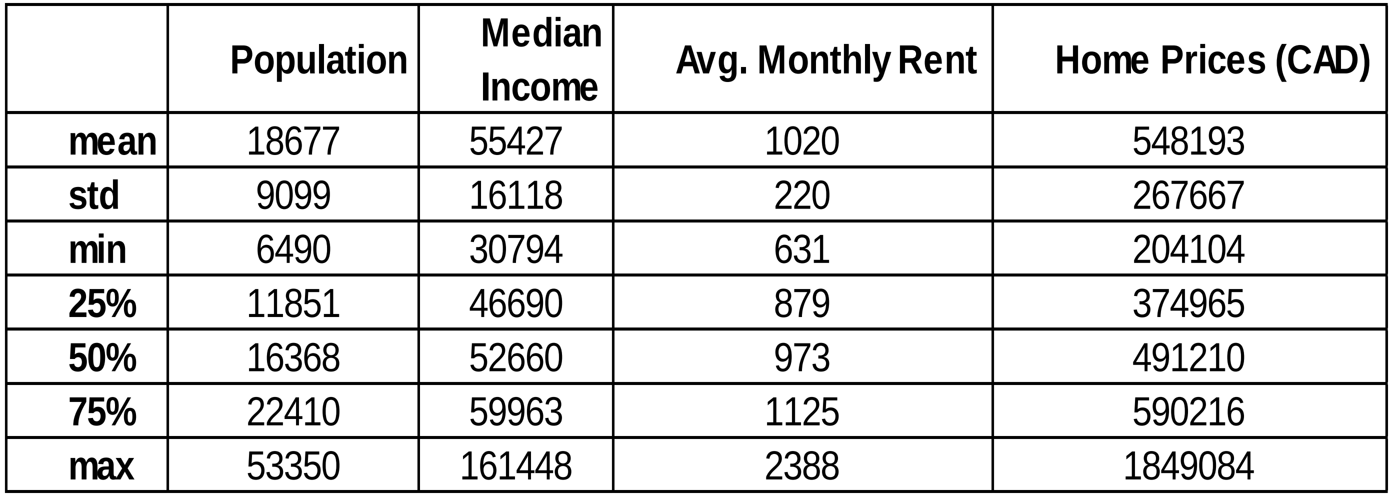
The following methodology was adopted:

- Step-1- Generating the dataframe by joining data sources.
- Step-2- Utilizing Foursquare API to include Venues data. With the merging of Venue-based data to our baseline dataframe, it is ready to be clustered upon, but before that, we conduct a preliminary exploratory data analysis.
- Step-3- Exploratory data analyses, inferential statistics- elaborated in "Results"
- Step-4- Clustering on the basis of "significant" parameters using k-means clustering
- Step-5- Mapping the clusters: A function was defined to generate a choropleth map of Toronto.

Results

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:



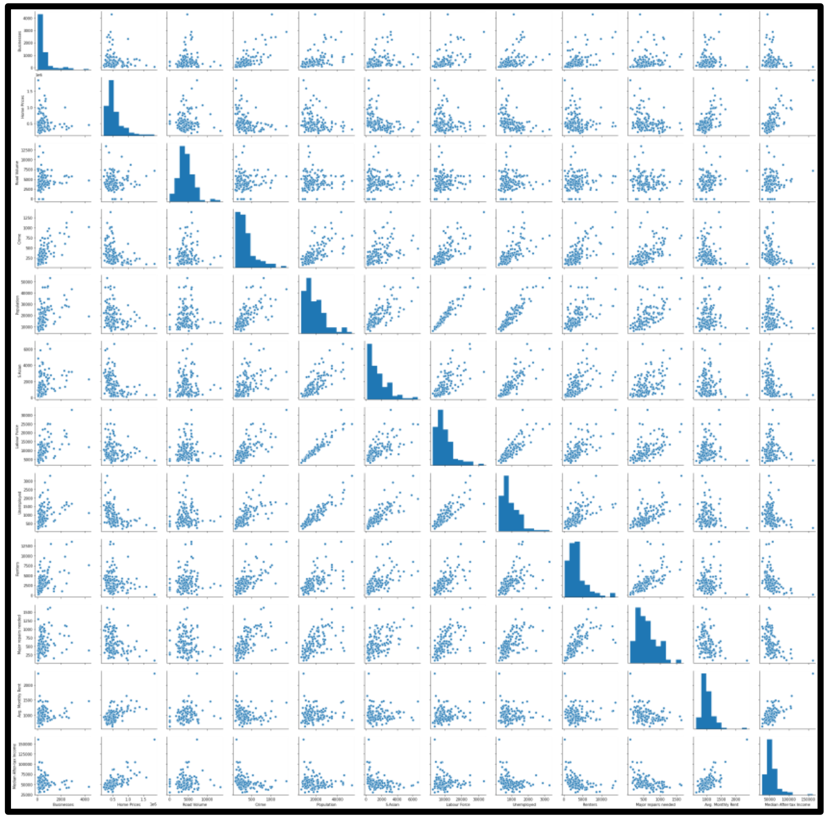
	Population	Median Income	Avg. Monthly Rent	Home Prices (CAD)
mean	18677	55427	1020	548193
std	9099	16118	220	267667
min	6490	30794	631	204104
25%	11851	46690	879	374965
50%	16368	52660	973	491210
75%	22410	59963	1125	590216
max	53350	161448	2388	1849084

Results

1. What would be the "new normal", the anticipated new baselines for living in Toronto?

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.

- While it needs to be zoomed to be read properly, the pairwise plot yields the pictorial visualization of the relationship between variables.

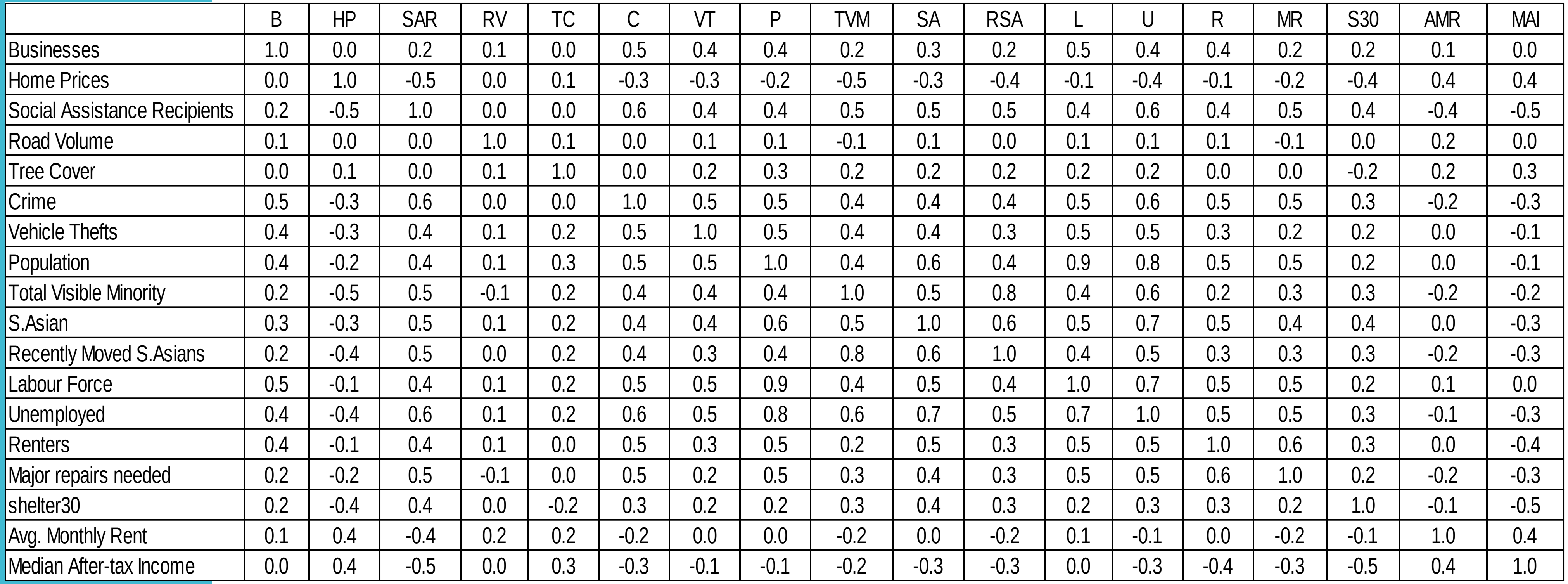


Broadly, the variables are either linearly or inversely related to each other in a predictable or justifiable manner.

Results

1. What would be the "new normal", the anticipated new baselines for living in Toronto?

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

Results

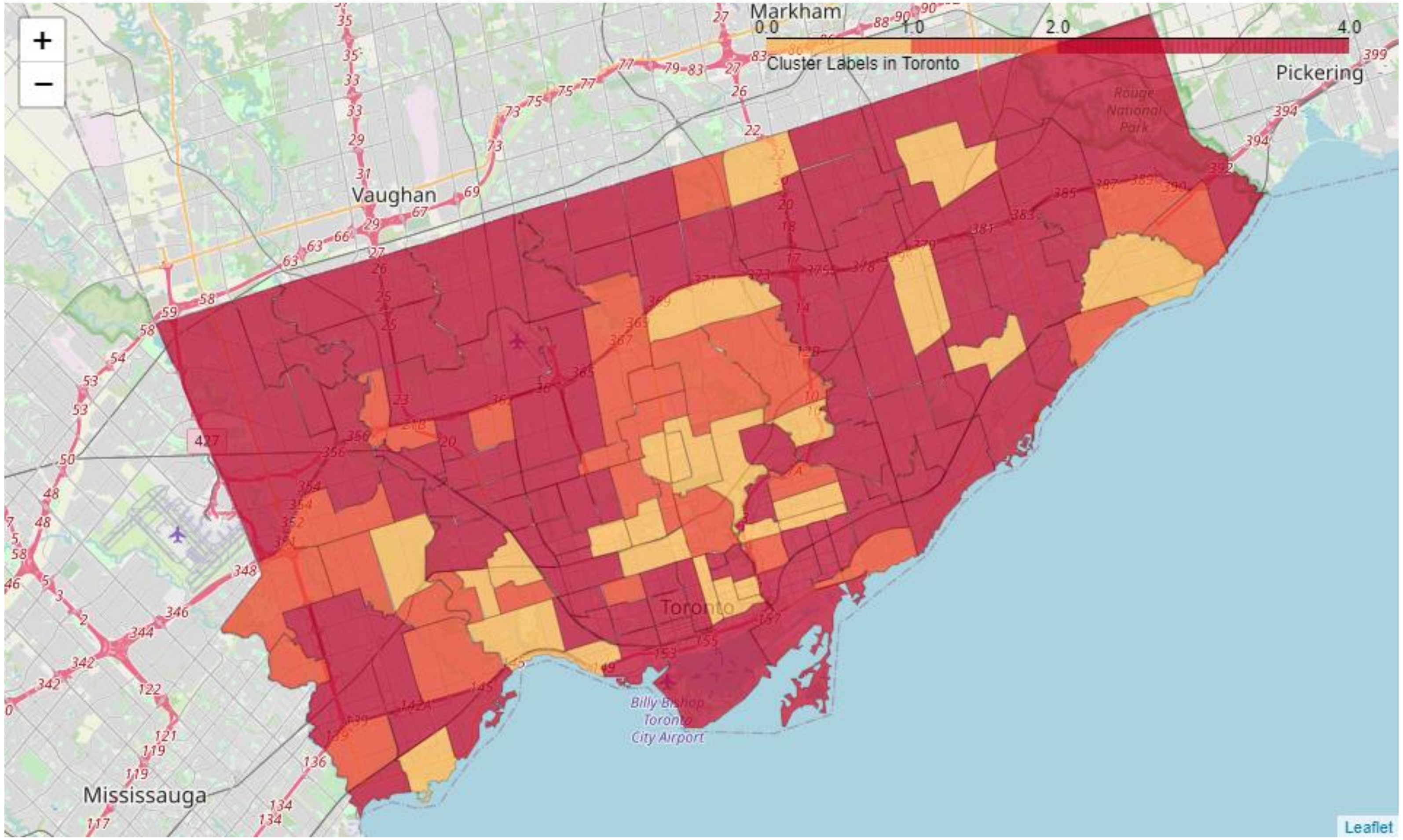
1. What would be the "new normal", the anticipated new baselines for living in Toronto?

- 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, etc.

Results

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 preferable neighbourhoods showing in a lighter tones.



Results

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.

Other relationships are detailed in the pairwise plot.

Discussion

A data science project is as good as the data sources and methodology of analysis that generate it:

- Geolocator data for which Lat-Long pairs are not mapped could not be used to generate Venues data from Foursquare. This can be fixed by manually fixing Lat-Long co-ordinates, but was not done in the present study.
- From exploratory data analysis, 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, impact may be low.

Thank you