DATA REPORT

TORONTO ENTREPRENEURIAL

Applied Data Science - Capstone

CONTENT

INTRODUCTION - WE INTRODUCE THE PROBLEM

DATA - WHERE WE GET THE DATA

DATA ANALYSIS - IDENTIFY SIGNIFICANT STATS INDICATORS

METHODOLOGY - ROAD MAP TO SOLVING THE PROBLEM

MACHINE LEARNING - WHAT ML ALGORITHMS WE USE

DATA RESULTS - SHARE DATA FINDINGS

DISCUSSION - SHARE INVESTIGATING FINDINGS

CONCLUSION - FINAL THOUGHTS

INTRODUCTION

The city of Toronto has approached our company to help them develop a service that helps the entrepreneurs who want to establish new businesses in the city of Toronto select an ideal business location based on the ethnic communities they want to be a part of.

This service will help find an ideal location for a new business based on such factors as business venue, population density in the area, the demographics in the area, average income, proximity to other business venues.

PROBLEM STATEMENT

Business success/failure depends on a vast spectrum of economics and demographics factors. Entrepreneurs may want to find an optimal venue and geographic location for their new business venture. Such an optimal venue/place selection process has to consider various indicators that may deliver long and prosperous existence for any new business.

The city of Toronto wants to offer such an online service where the entrepreneurs can receive all the necessary information that will help them in picking the location for their new ventures based on their desire to support a specific ethnic community of Toronto.

PROPOSED SOLUTION

The solution can be provided using Foursquare location data as well as the demographics open dataset available from Wikipedia.

In order to advise the entrepreneurs on a good location, we will consider the density (frequency) of similar business venues in various parts of Toronto that cater to preferred ethnic area/neighbourhoods, average income, population, population density, population growth rate, spoken languages in the same area.

DATA SOURCES

To solve the problem our service will rely on open datasets generated from the following sources:

- Wikipedia
  - Toronto Boroughs/Neighbourhoods
    - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  - Canada Census - Toronto Demographics
    - https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
- Foursquare APIs

DATA SOURCES

Toronto Boroughs/Neighbourhoods: a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario

Canada Census - Toronto Demographics: a list of demographic data on each Toronto neighbourhood as taken from the Canadian Census.

Foursquare APIs: offers rich location-based experiences and enables access to millions of up to date business venues, tips, photos and many other helpful tips

METHODOLOGY

- REPORT - MAIN COMPONENTS
  a. DATA [POST-PROCESSING/SUMMARY] - in order to perform statistical inference, and apply the machine learning algorithms, the data must be acquired and pre-processed based on the rules derived from the preliminary data analysis
  b. DATA ANALYSIS - Identify the significant informational indicators to use in inferential statistics and machine learning algorithm [Unsupervised: K-Means]
  c. DATA ANALYSIS - Statistical Validation: The datasets underwent statistical analysis and cross referencing in order to determine the data validity and proper distribution, mean and standard deviations, outlier identification.
  d. MACHINE LEARNING - UNSUPERVISED MACHINE LEARNING K-MEANS: In order to cluster various regions of the city based on the business analysis requirements the solution utilizes the unsupervised machine learning algorithm K-MEANS
  e. DATA RESULTS - Present the finding to the stakeholders
  f. DISCUSSION - discuss data investigative findings based on the results
  g. CONCLUSION - report conclusions

DATA: TORONTO BOROUGHS [212 records]

Postcode	Borough	Neighbourhood	Latitude	Longitude
M3A	North York	Parkwoods	43.753259	-79.329656
M4A	North York	Victoria Village	43.725882	-79.315572
M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
M6A	North York	Lawrence Heights	43.718518	-79.464763
M6A	North York	Lawrence Manor	43.718518	-79.464763
M7A	Queen's Park	Not assigned	43.662301	-79.389494
M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
M1B	Scarborough	Rouge	43.806686	-79.194353
M1B	Scarborough	Malvern	43.806686	-79.194353

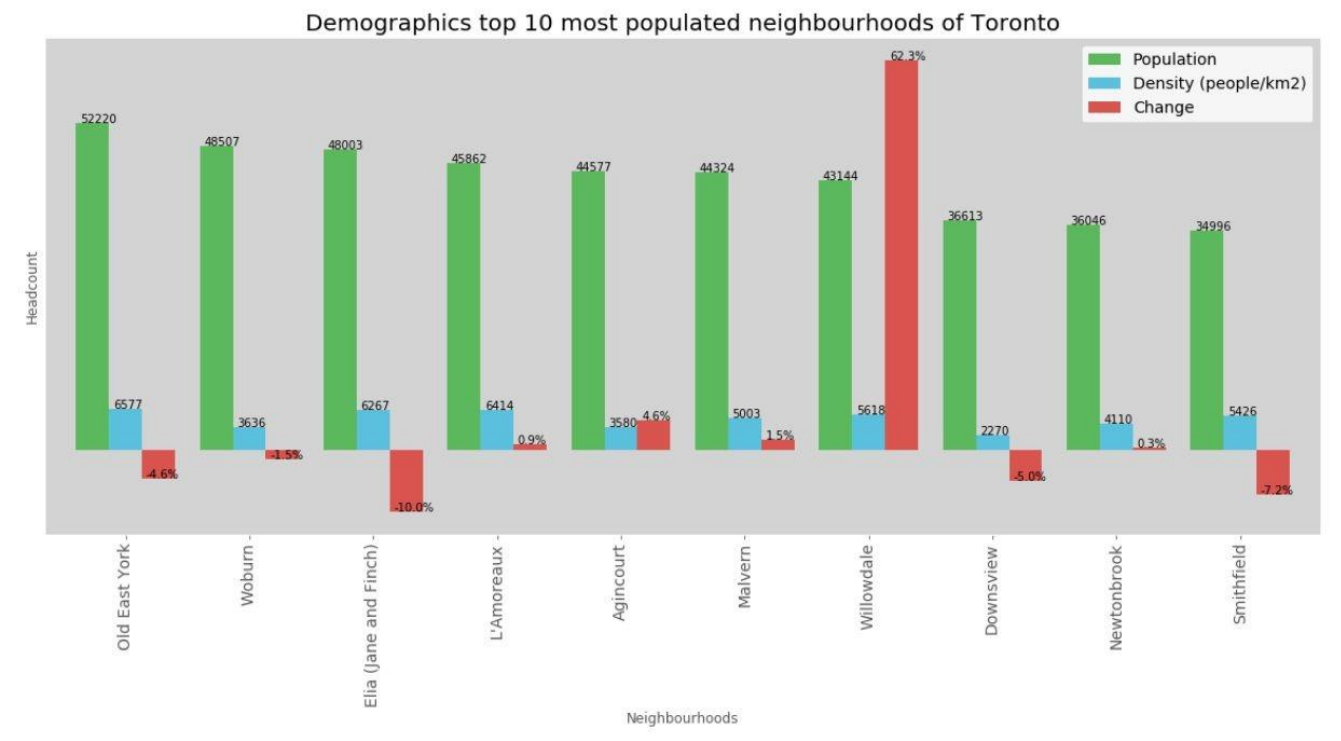
DATA: TORONTO BOROUGHS

DATA: TORONTO DEMOGRAPHICS [174 records]

Neighbourhood	Population	Density (people/km2)	Average Income	Percentage	Language
Agincourt	44577	3580	25750	19.3	Cantonese
Alderwood	11656	2360	35239	06.2	Polish
Alexandra Park	4355	13609	19687	17.9	Cantonese
Allenby	2513	4333	245592	01.4	Russian
Amesbury	17318	4934	27546	06.1	Spanish
Armour Heights	4384	1914	116651	09.4	Russian
Banbury	6641	2442	92319	05.1	Unspecified Chinese
Bathurst Manor	14945	3187	34169	09.5	Russian
Bay Street Corridor	4787	43518	40598	09.6	Mandarin
Bayview Village	12280	2966	46752	08.4	Cantonese

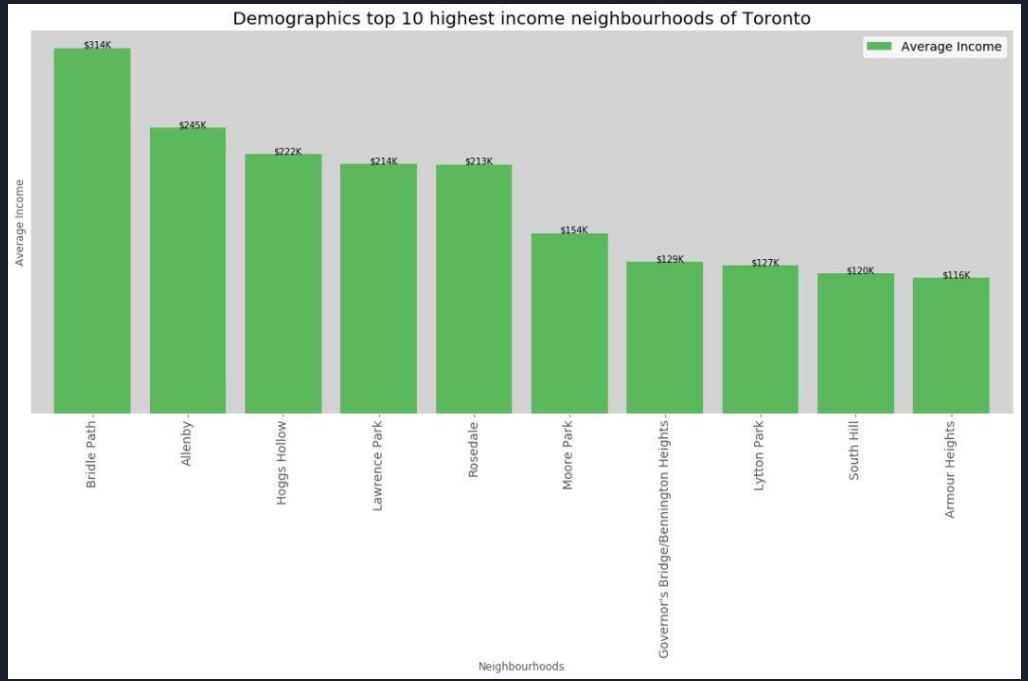
DATA ANALYSIS

MOST POPULATED NEIGHBOURHOOD



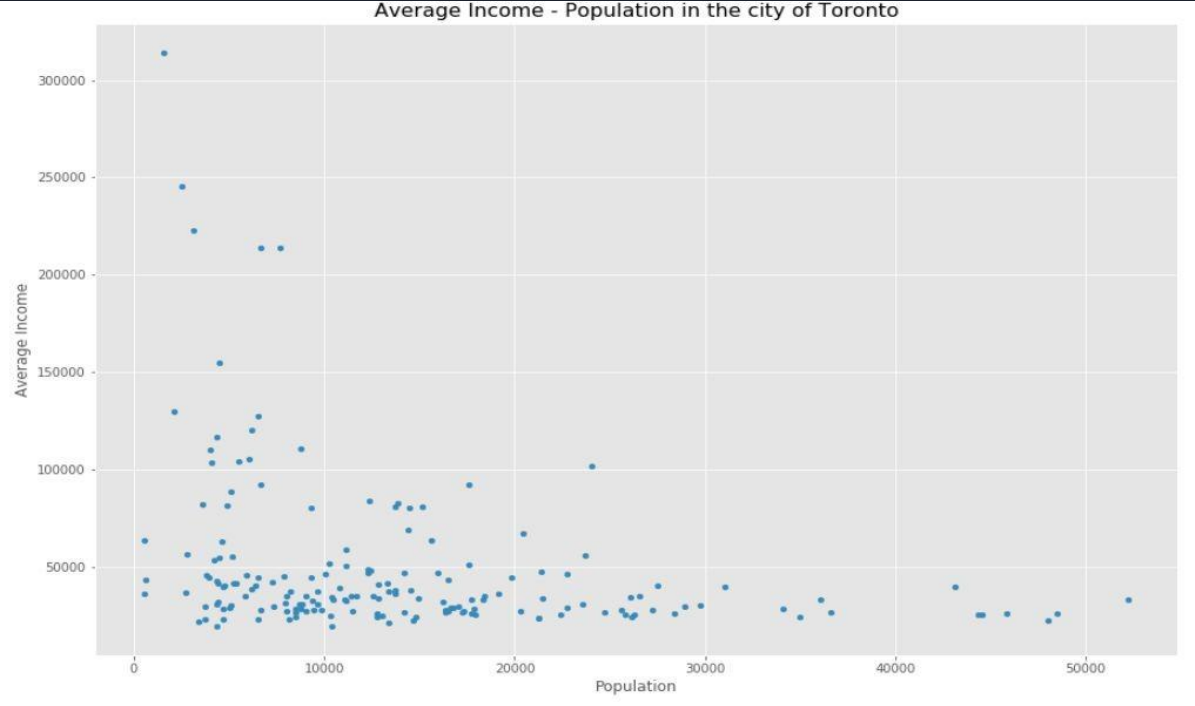
DATA ANALYSIS

HIGHEST INCOME NEIGHBOURHOODS



DATA ANALYSIS

AVERAGE INCOME - POPULATION



DATA ANALYSIS

ETHNIC COMMUNITY AS BUSINESS TARGET

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Population	Density (people/km2)	Average Income	Language
0	M5T	Downtown Toronto	Grange Park	43.653206	-79.400049	9007	10793	35277	Chinese
1	M3A	North York	Parkwoods	43.753259	-79.329656	26533	5349	34811	Chinese
2	M5B	Downtown Toronto	Garden District	43.657162	-79.378937	8240	15846	37614	Chinese
3	M4X	Downtown Toronto	Cabbagetown	43.667967	-79.367675	11120	7943	50398	Chinese
4	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	7672	2821	213941	Chinese
5	M1N	Scarborough	Birch Cliff	43.692657	-79.264848	12266	3525	48965	Chinese

MACHINE LEARNING

OUR DATA ANALYSIS SHOWS LACK OF PROPER DATA LABELING IN THE DATASETS USED BY THE SOLUTION

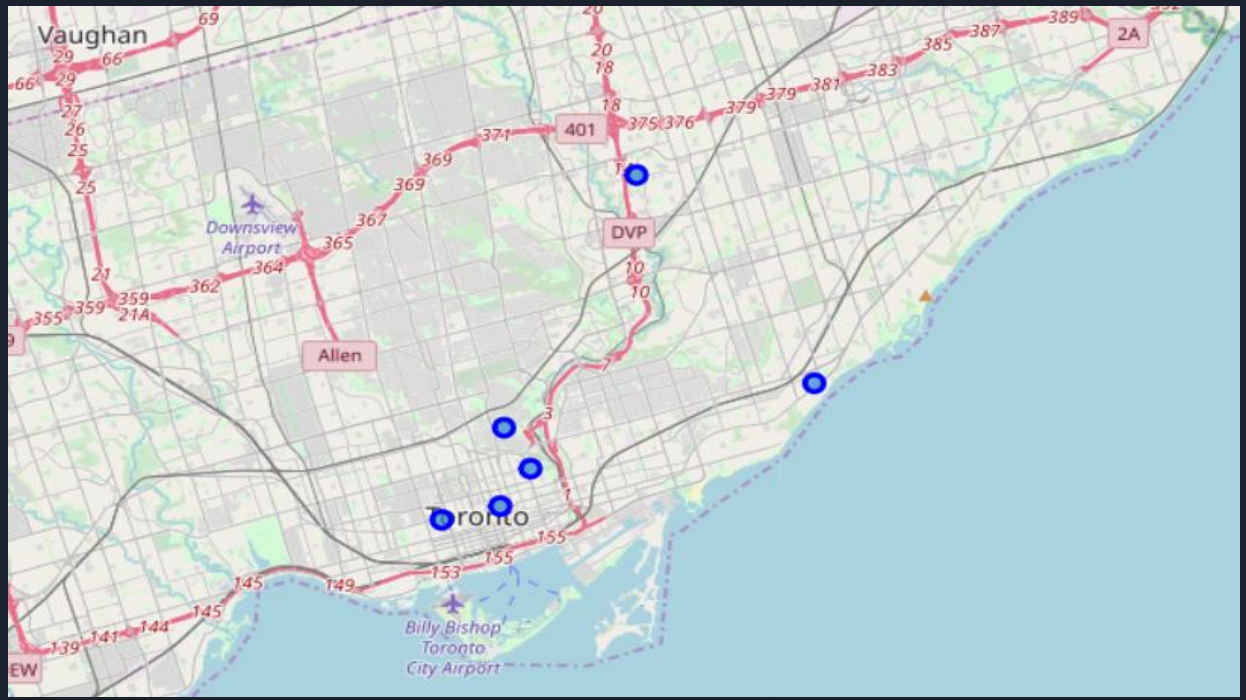
BASED ON THE DATA ANALYSIS AND THE SOLUTION REQUIREMENTS WE SUGGEST USING AN UNSUPERVISED MACHINE LEARNING APPROACH

WE SUGGEST USING K-MEANS UNSUPERVISED MACHINE LEARNING ALGORITHM TO IDENTIFY GEO CLUSTERS IN THE CITY OF TORONTO THAT ARE MOST SUITABLE FOR OPENING NEW SMALL BUSINESSES IN THE CITY OF TORONTO

IN ORDER TO PERFORM ACCURATE GEO CLUSTERING OUR ALGORITHM RELIES ON FOURSQUARE API'S

DATA RESULTS

Target Group Geo-Mapping



DATA RESULTS

TARGET GROUP BUSINESS VENUE

Neighbourhood	Average_Income
Birch Cliff	4
Cabbagetown	48
Garden District	100
Grange Park	100
Parkwoods	4
Rosedale	4

DATA RESULTS

TARGET GROUP MOST COMMON VENUES

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Birch Cliff	College Stadium	General Entertainment	Café	Skating Rink	Dumpling Restaurant	Comic Shop	Concert Hall	Convenience Store	Cosmetics Shop	Deli / Bodega
1	Cabbagetown	Restaurant	Coffee Shop	Pizza Place	Bakery	Italian Restaurant	Park	Café	Convenience Store	Pub	Market
2	Garden District	Coffee Shop	Clothing Store	Café	Cosmetics Shop	Middle Eastern Restaurant	Italian Restaurant	Plaza	Restaurant	Pizza Place	Diner
3	Grange Park	Bar	Café	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Coffee Shop	Bakery	Mexican Restaurant	Chinese Restaurant	Burger Joint
4	Parkwoods	Fast Food Restaurant	Park	Food & Drink Shop	Bus Stop	Farmers Market	Concert Hall	Convenience Store	Cosmetics Shop	Deli / Bodega	Department Store
5	Rosedale	Park	Playground	Trail	Dumpling Restaurant	Comfort Food Restaurant	Comic Shop	Concert Hall	Convenience Store	Cosmetics Shop	Deli / Bodega

DATA RESULTS

TARGET GROUP LEAST COMMON VENUES

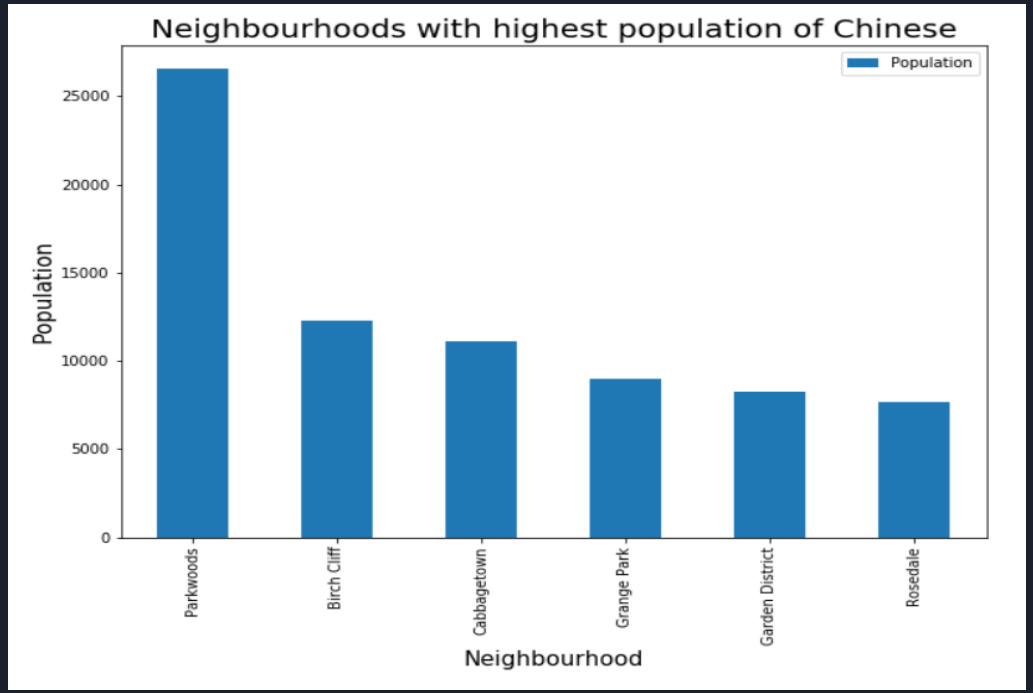
	Neighbourhood	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Least Common Venue	6th Least Common Venue	7th Least Common Venue	8th Least Common Venue	9th Least Common Venue	10th Least Common Venue
0	Birch Cliff	American Restaurant	Other Great Outdoors	Organic Grocery	Office	Noodle House	Music Venue	Movie Theater	Modern European Restaurant	Miscellaneous Shop	Middle Eastern Restaurant
1	Cabbagetown	American Restaurant	Modern European Restaurant	Miscellaneous Shop	Middle Eastern Restaurant	Mexican Restaurant	Martial Arts Dojo	Lounge	Lingerie Store	Lake	Juice Bar
2	Garden District	Dim Sum Restaurant	Martial Arts Dojo	Poutine Place	Doner Restaurant	Donut Shop	Dumpling Restaurant	Farmers Market	Playground	Filipino Restaurant	Fish & Chips Shop
3	Grange Park	American Restaurant	Movie Theater	Modern European Restaurant	Miscellaneous Shop	Middle Eastern Restaurant	Market	Lounge	Lingerie Store	Lake	Music Venue
4	Parkwoods	American Restaurant	Other Great Outdoors	Organic Grocery	Office	Noodle House	Music Venue	Movie Theater	Modern European Restaurant	Miscellaneous Shop	Middle Eastern Restaurant
5	Rosedale	American Restaurant	Organic Grocery	Office	Noodle House	Music Venue	Movie Theater	Modern European Restaurant	Miscellaneous Shop	Middle Eastern Restaurant	Mexican Restaurant

DATA RESULTS: K-MEANS CLUSTERING TARGET GROUP CLUSTERING

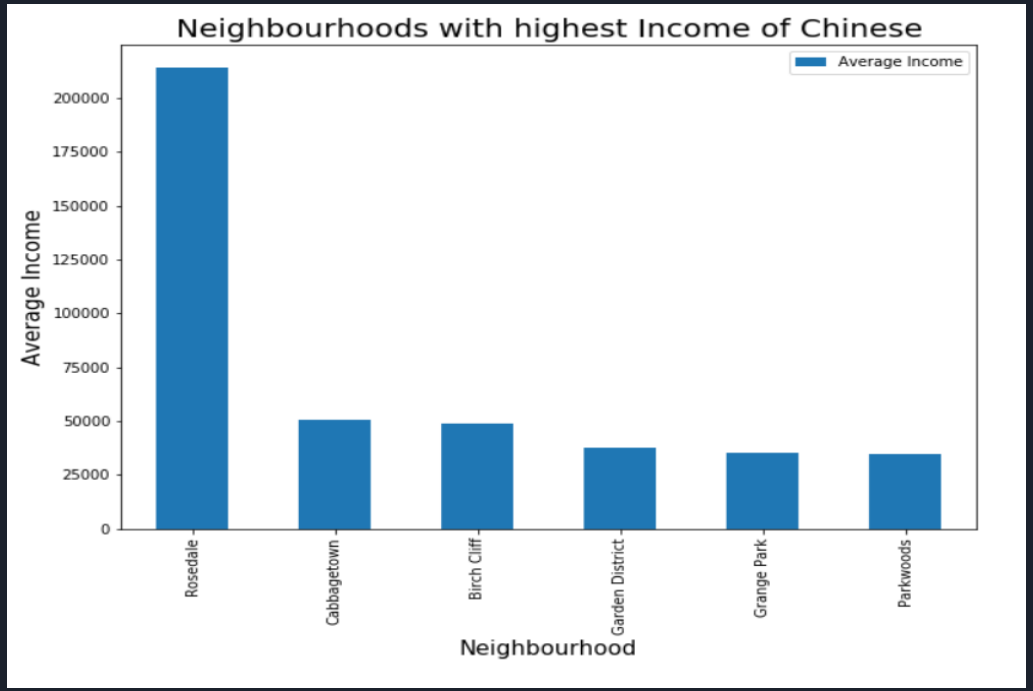
	Postcode	Borough	Neighbourhood	Latitude	Longitude	Population	Density (people/km2)	Average Income	Language	Percentage of people speaking language	Cluster Labels
0	M5T	Downtown Toronto	Grange Park	43.653206	-79.400049	9007	10793	35277	Chinese	14.8	1
1	M3A	North York	Parkwoods	43.753259	-79.329656	26533	5349	34811	Chinese	3.4	0
2	M5B	Downtown Toronto	Garden District	43.657162	-79.378937	8240	15846	37614	Chinese	3.0	5
3	M4X	Downtown Toronto	Cabbagetown	43.667967	-79.367675	11120	7943	50398	Chinese	1.6	4
4	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	7672	2821	213941	Chinese	1.0	2
5	M1N	Scarborough	Birch Cliff	43.692657	-79.264848	12266	3525	48965	Chinese	0.9	3

DATA RESULTS

TARGET GROUP POPULATION DISTRIBUTION



TARGET GROUP INCOME DISTRIBUTION



DISCUSSION

THERE ARE VERY INTERESTING TRENDS SHOWING UP IN THE DATA ANALYSIS THAT SUGGEST THAT IT IS POSSIBLE TO RECOMMEND NEW LOCATIONS TO BUSINESSES THAT WANT TO EXPAND OR NEW BUSINESSES LOOKING FOR THE FIRST LOCATION.THERE ARE MULTIPLE STATISTICAL METHODOLOGIES THAT CAN BE EMPLOYED TO FORMULATE A SOUND BUSINESS HYPOTHESIS. SUCH FORMULATED HYPOTHESIS DO REQUIRE VALIDATION VIA GATHERING AND PROCESSING THE SUPPORTING EVIDENCE.

SUCH SUPPORTING EVIDENCE CAN BE PRODUCED BY EMPLOYING ONE OR MORE MACHINE LEARNING ALGORITHMS.

CONCLUSION

GIVEN ENOUGH RELEVANT DATA IT IS POSSIBLE TO GENERATE SUFFICIENT AMOUNT OF SUPPORTING EVIDENCE IN ORDER TO RECOMMEND WITH A HIGH LEVEL OF PRECISION GEO LOCATIONS FOR NEW OR GROWING BUSINESSES.

THE CURRENT PROJECT DEMONSTRATES THAT A NEW LOCATION CAN BE SELECTED BASED ON A LIST OF INDICATORS DERIVED VIA INFERENTIAL STATISTICS AND THE RESULTS PROCESSED WITH K-MEANS CLUSTERING MACHINE LEARNING ALGORITHM.

BEING A BUSINESS WITH CLOSE TIES TO VARIOUS ETHNIC COMMUNITIES IN TORONTO WE DEFINITELY CONCUR THAT THE FINDINGS PRESENTED IN THIS REPORT HAVE STRONG MERITS.