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.

Successful businesses help the economy grow, lower the unemployment, and reduce crime. The multicultural 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.

AUDIENCE

The target audience for this service: Business Entrepreneurs seeking to establish new businesses in the city of Toronto

Having a tool that can help the entrepreneurs to choose the right location for their business will assure long and prosperous existence for such businesses, serving the communities and helping them grow.

PROBLEM SOLUTION

The solution will leverage the 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
- GeoCoder/Google Geolocation APIs
- 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.

GeoCoder/Google Geolocation APIs: converts addresses into geographic coordinates

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

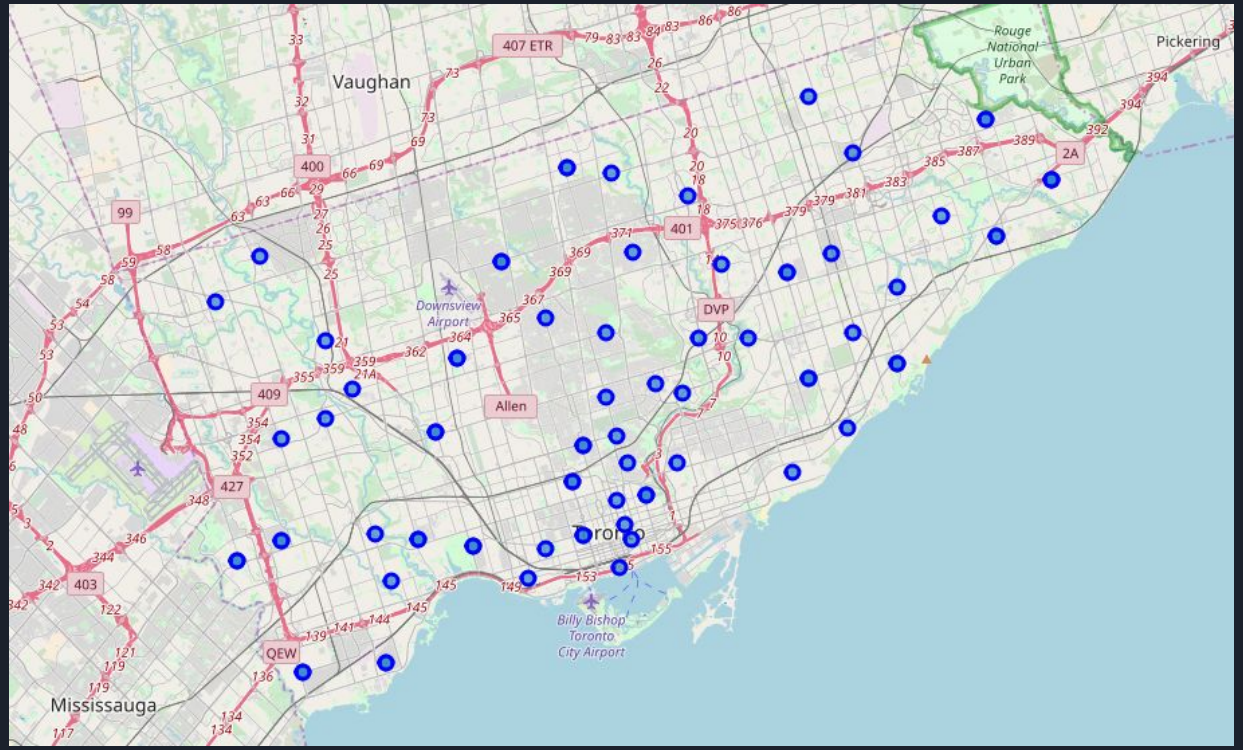
METHODOLOGY

- REPORT - MAIN COMPONENTS
  a. DATA [ACQUISITION/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

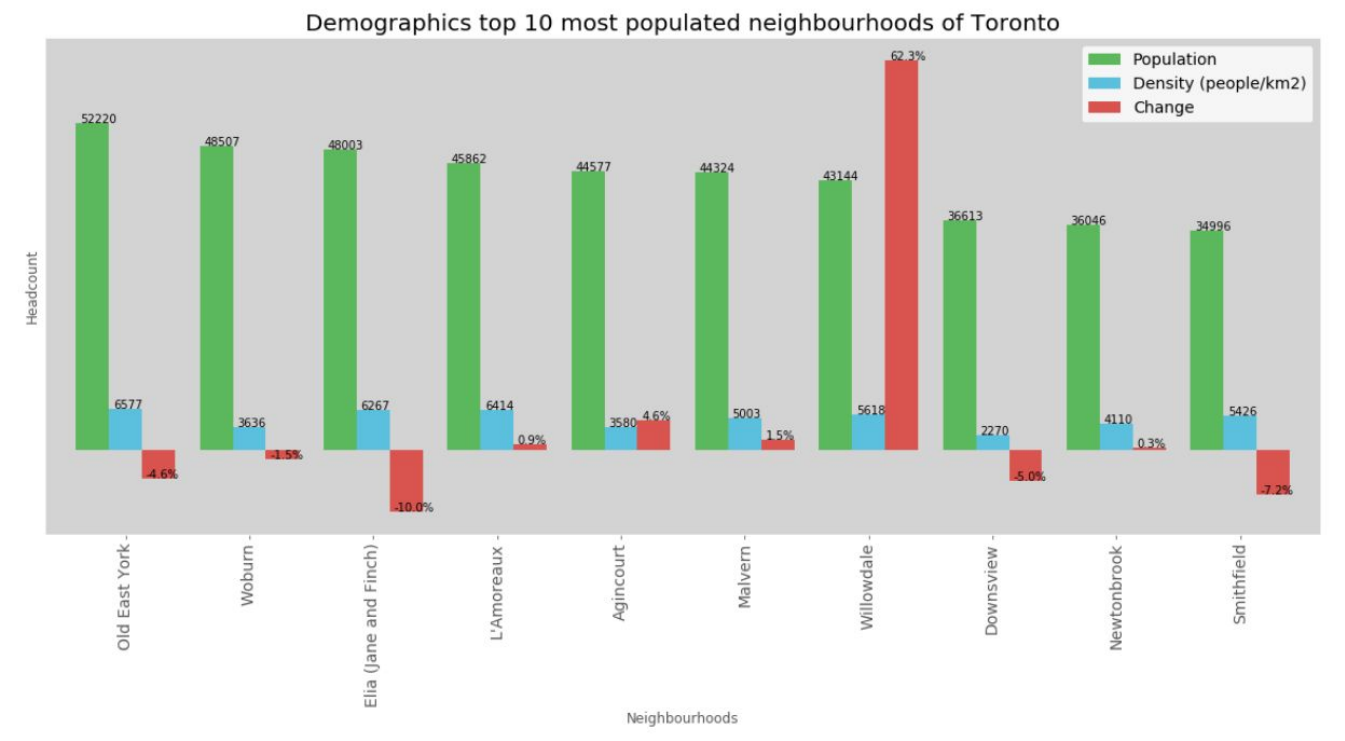
DATA: TORONTO BOROUGHS



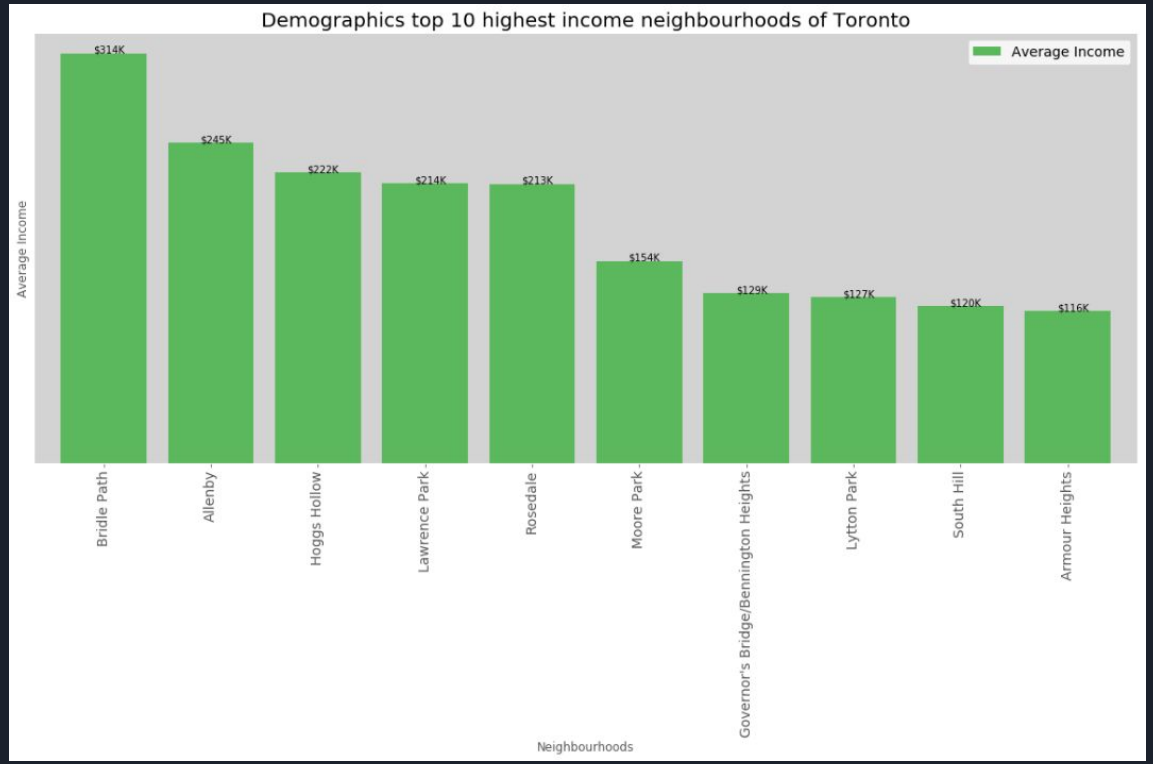
DATA: TORONTO DEMOGRAPHICS [174 records]

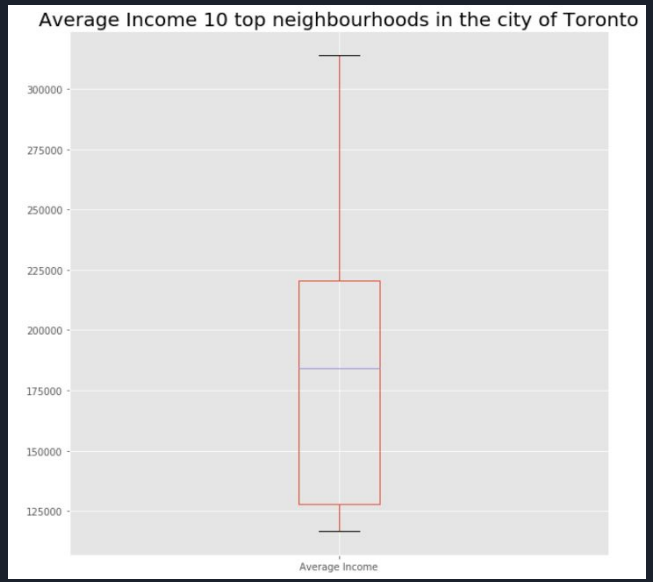
Neighbourhood	Population	Density (people/km2)	Average Income	Percentage	Language
Agincourt	44577	3580	25750	19.3	Cantonese
Alderwood	11656	2360	35239	06.2	Polish
Alexandra Park	4355	13609	19687	17.9	Cantonese
Allenby	2513	4333	245592	01.4	Russian
Amesbury	17318	4934	27546	06.1	Spanish
Armour Heights	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 NEIGHBOURHOODS

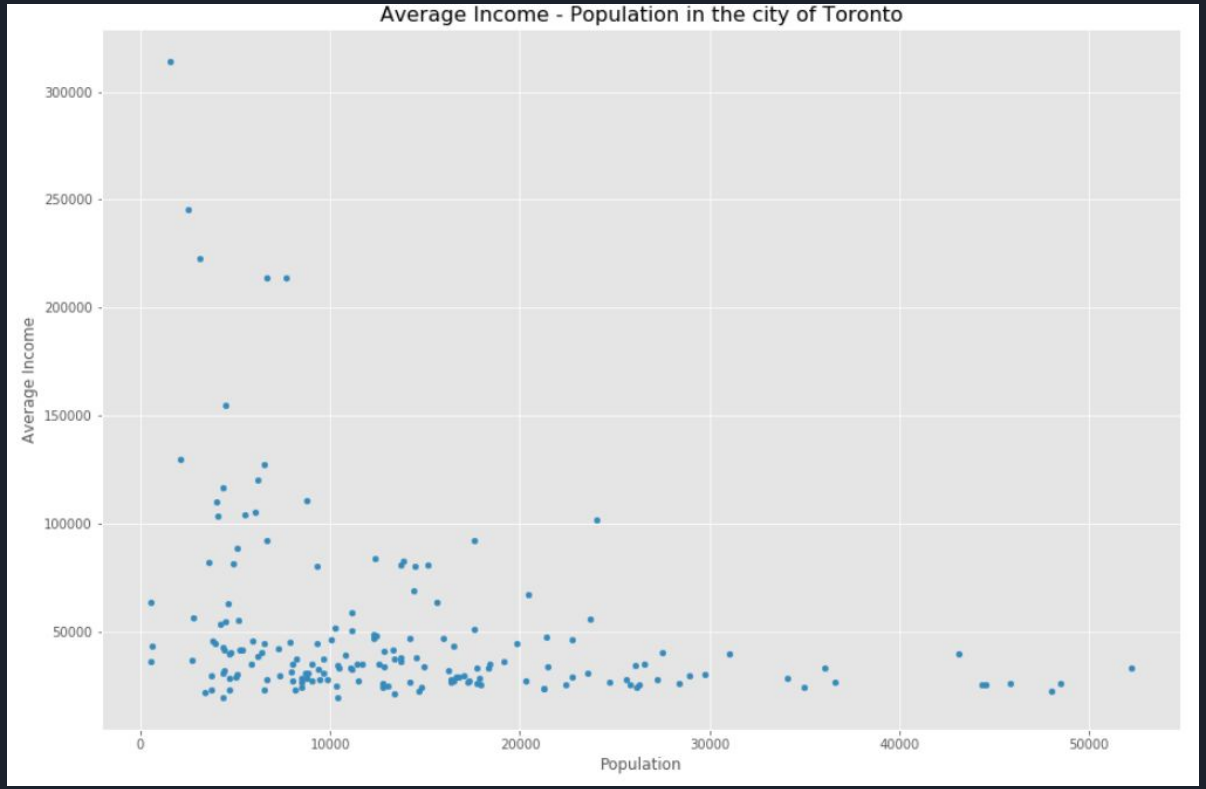


DATA ANALYSIS
HIGHEST INCOME NEIGHBOURHOODS





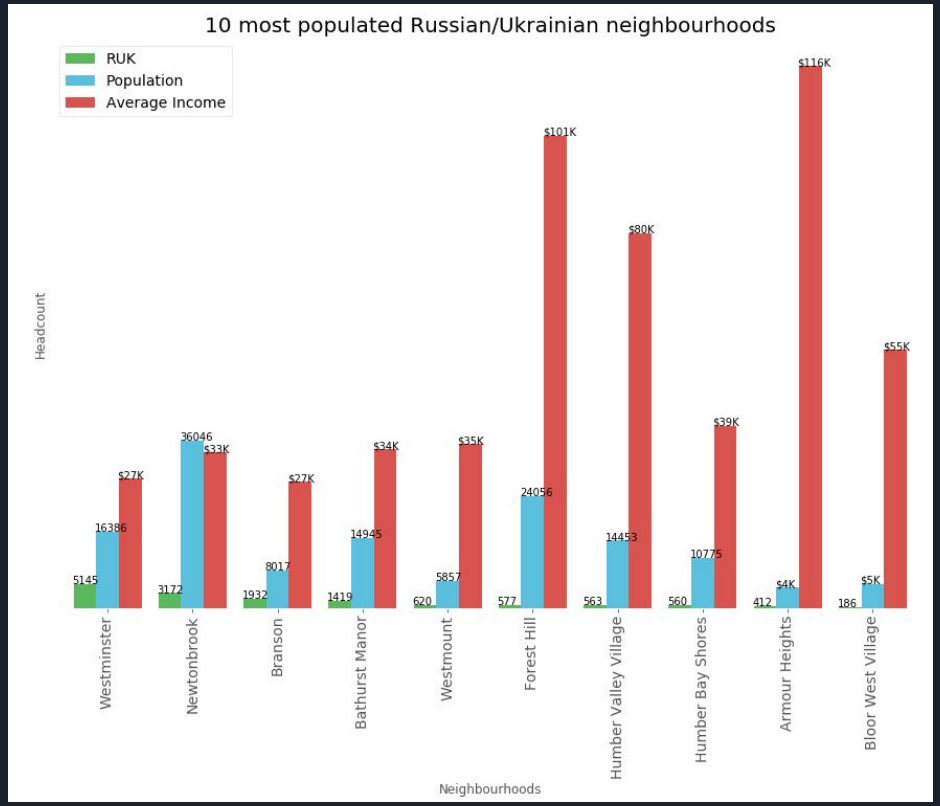
DATA ANALYSIS
AVERAGE INCOME - POPULATION

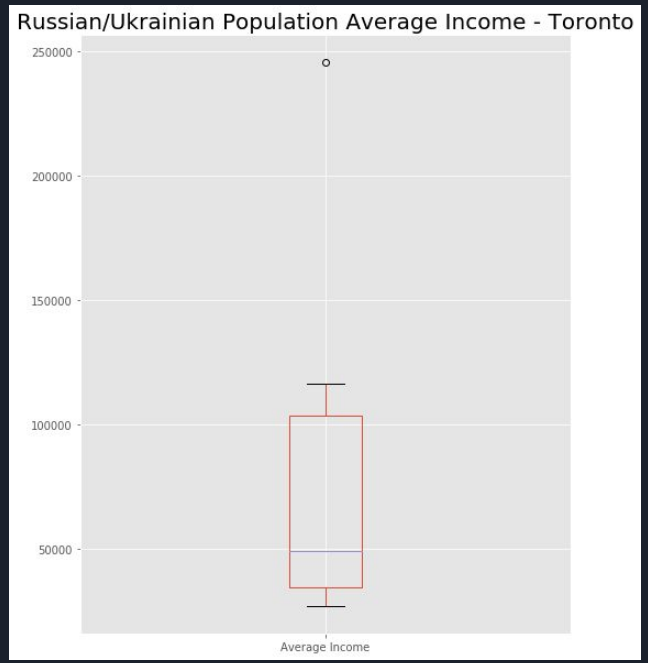


DATA ANALYSIS
ETHNIC COMMUNITY AS BUSINESS TARGET

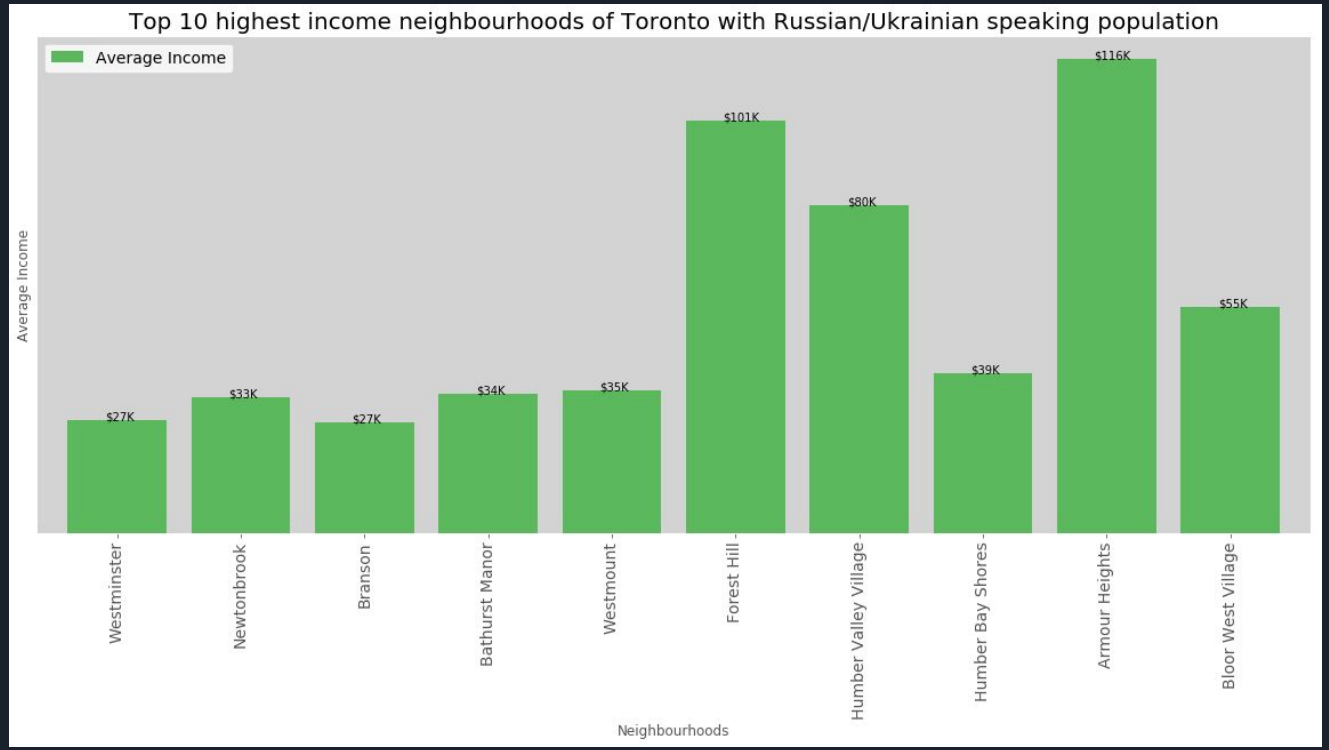
Neighbourhood	Population	Density (people/km2)	Average Income	Percentage	Language	RUK
Newtonbrook	36046	4110	33428	8.8	Russian	3172
Bathurst Manor	14945	3187	34169	9.5	Russian	1419
Westmount	5857	5932	35183	10.6	Ukrainian	620
Humber Bay Shores	10775	7588	39186	5.2	Russian	560
Deer Park	15165	10387	80704	1.1	Russian	166
The Kingsway	8780	3403	110944	1.8	Ukrainian	158
West Deane Park	4395	2063	41582	2.3	Ukrainian	101
Runnymede	4382	5155	42635	2.2	Ukrainian	96

DATA ANALYSIS
IDENTIFYING TOP 10 TARGET GROUP BOROUGHS





DATA ANALYSIS
IDENTIFYING TOP 10 TARGET GROUP BOROUGHS



MACHINE LEARNING

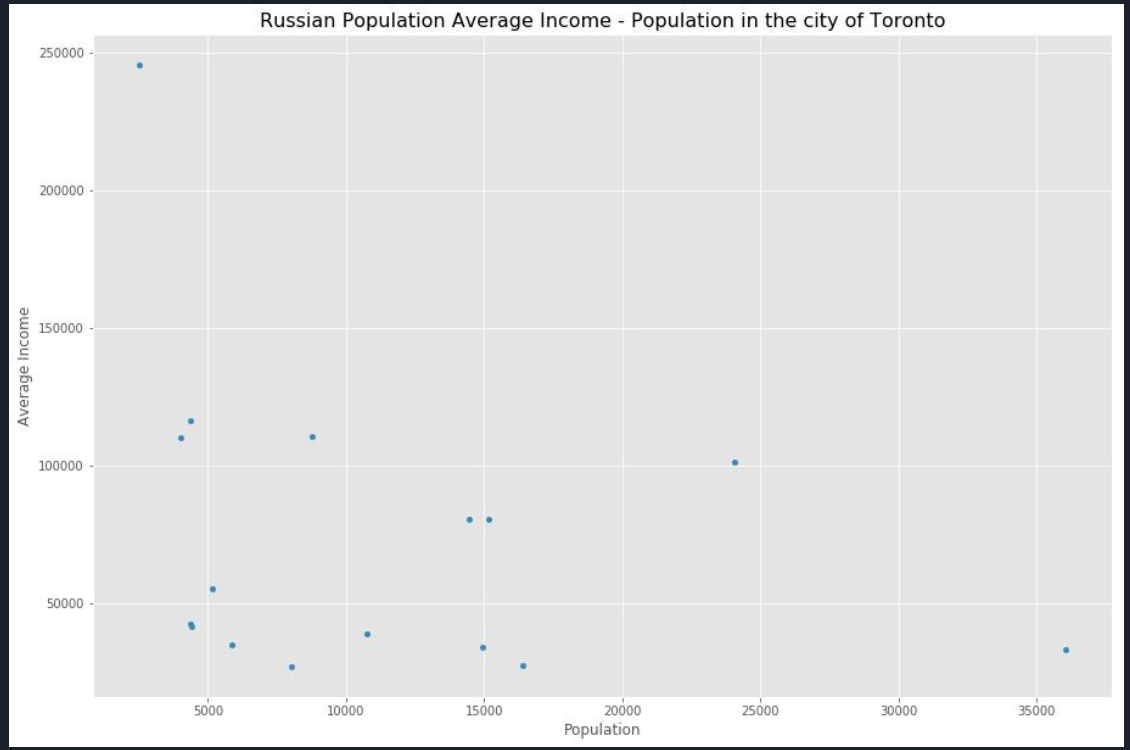
OUR DATA ANALYSIS SHOWS LACK OF PROPER DATA LABELING IN THE DATASETS USED BY THE SOLUTION

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 GOOGLE GEO LOCATIONS AND FOURSQUARE APIS

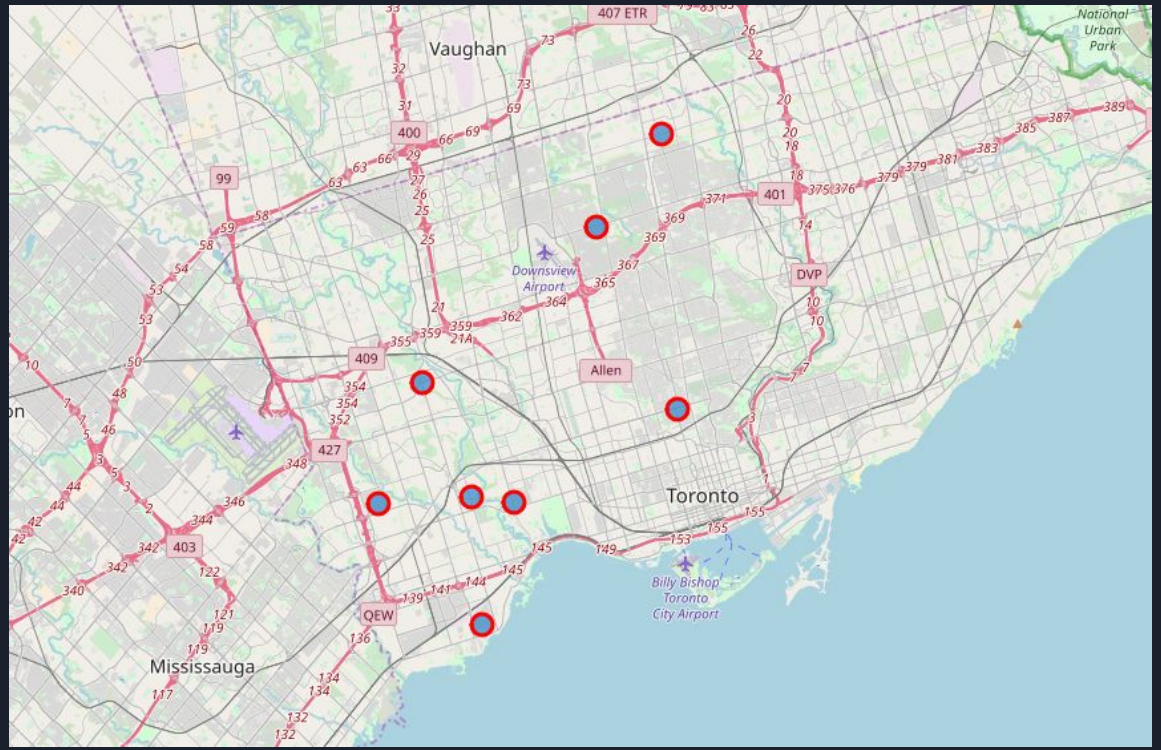
DATA RESULTS
TARGET GROUP POPULATION/INCOME



DATA RESULTS
TARGET GROUP GEO-MAPPING

Postcode	Borough	Neighbourhood	Latitude	Longitude	Population	Density (people/km2)	Average Income	Percentage	Language
M9P	Etobicoke	Westmount	43.696319	-79.532242	5857	5932	35183	10.6	Ukrainian
M3H	North York	Bathurst Manor	43.754328	-79.442259	14945	3187	34169	09.5	Russian
M2M	North York	Newtonbrook	43.789053	-79.408493	36046	4110	33428	08.8	Russian
M8V	Etobicoke	Humber Bay Shores	43.605647	-79.501321	10775	7588	39186	05.2	Russian
M9B	Etobicoke	West Deane Park	43.650943	-79.554724	4395	2063	41582	02.3	Ukrainian
M6S	West Toronto	Runnymede	43.651571	-79.484450	4382	5155	42635	02.2	Ukrainian
M8X	Etobicoke	The Kingsway	43.653654	-79.506944	8780	3403	110944	01.8	Ukrainian
M4V	Central Toronto	Deer Park	43.686412	-79.400049	15165	10387	80704	01.1	Russian

DATA RESULTS
TARGET GROUP GEO-MAPPING



DATA RESULTS
TARGET GROUP BOROUGHS BUSINESS VENUES

- Westmount
- Bathurst Manor
- Newtonbrook
- Humber Bay Shores
- West Deane Park
- Runnymede
- The Kingsway
- Deer Park

Neighborhood	
Bathurst Manor	18
Deer Park	14
Humber Bay Shores	15
Newtonbrook	1
Runnymede	35
The Kingsway	2
West Deane Park	1
Westmount	7

TOTAL NUMBER OF VENUES: 93

TOTAL NUMBER OF UNIQUE CATEGORIES: 54

DATA RESULTS
TARGET GROUP MOST COMMON VENUES

Neighborhood	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
Bathurst Manor	Coffee Shop	Sandwich Place	Deli / Bodega	Ice Cream Shop	Pizza Place	Bridal Shop	Grocery Store	Video Store	Restaurant	Pharmacy
Deer Park	Pub	Coffee Shop	Vietnamese Restaurant	Bagel Shop	Convenience Store	Fried Chicken Joint	Light Rail Station	Pizza Place	American Restaurant	Supermarket
Humber Bay Shores	Coffee Shop	Restaurant	Fish & Chips Shop	Fast Food Restaurant	Gym	Construction & Landscaping	Liquor Store	Mexican Restaurant	Pharmacy	Fried Chicken Joint
Newtonbrook	Gym	Vietnamese Restaurant	Construction & Landscaping	Fried Chicken Joint	French Restaurant	Food	Fish & Chips Shop	Fast Food Restaurant	Falafel Restaurant	Diner
Runnymede	Coffee Shop	Café	Sushi Restaurant	Italian Restaurant	Pizza Place	Gourmet Shop	Falafel Restaurant	Fish & Chips Shop	Food	French Restaurant
The Kingsway	River	Pool	Vietnamese Restaurant	Coffee Shop	Fried Chicken Joint	French Restaurant	Food	Fish & Chips Shop	Fast Food Restaurant	Falafel Restaurant
West Deane Park	Bank	Vietnamese Restaurant	Construction & Landscaping	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Food	Fish & Chips Shop	Fast Food Restaurant	Falafel Restaurant
Westmount	Pizza Place	Playground	Coffee Shop	Intersection	Sandwich Place	Chinese Restaurant	French Restaurant	Food	Fish & Chips Shop	Fast Food Restaurant

DATA RESULTS
TARGET GROUP LEAST COMMON VENUES

Neighborhood	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
Bathurst Manor	American Restaurant	Gastropub	Gym	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
Deer Park	Gourmet Shop	Frozen Yogurt Shop	Gastropub	Video Store	Grocery Store	Gym	Ice Cream Shop	Indie Movie Theater	Intersection	River
Humber Bay Shores	Gourmet Shop	Video Store	Grocery Store	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Pizza Place
Newtonbrook	American Restaurant	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant	Pharmacy
Runnymede	American Restaurant	Frozen Yogurt Shop	Video Store	Grocery Store	Ice Cream Shop	Intersection	Light Rail Station	Liquor Store	Mexican Restaurant	Playground
The Kingsway	American Restaurant	Gym	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
West Deane Park	American Restaurant	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant	Pharmacy
Westmount	American Restaurant	Grocery Store	Gym	Ice Cream Shop	Indie Movie Theater	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant

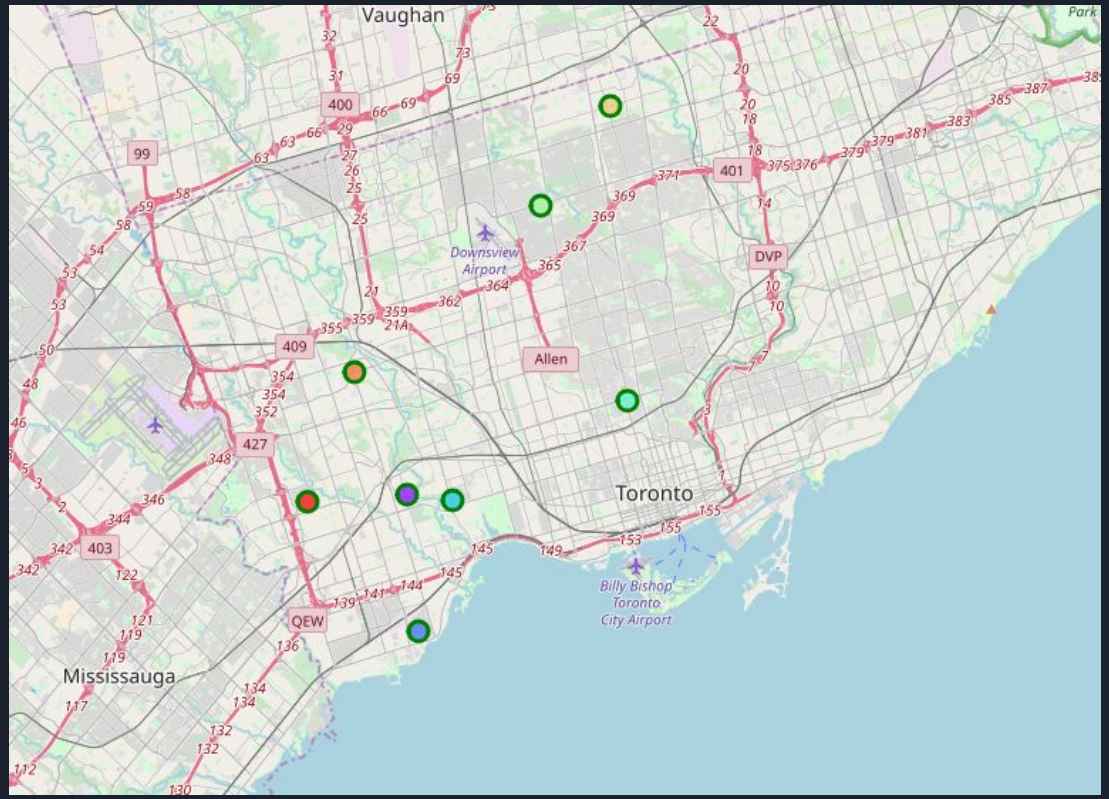
DATA RESULTS: K-MEANS CLUSTERING TARGET GROUP CLUSTERING

Postcode	Borough	Neighbourhood	Latitude	Longitude	Population	Density (people/km2)	Average Income	Percentage	Language	Cluster Labels
M9P	Etobicoke	Westmount	43.696319	-79.532242	5857	5932	35183	10.6	Ukrainian	7
M3H	North York	Bathurst Manor	43.754328	-79.442259	14945	3187	34169	09.5	Russian	5
M2M	North York	Newtonbrook	43.789053	-79.408493	36046	4110	33428	08.8	Russian	6
M8V	Etobicoke	Humber Bay Shores	43.605647	-79.501321	10775	7588	39186	05.2	Russian	2
M9B	Etobicoke	West Deane Park	43.650943	-79.554724	4395	2063	41582	02.3	Ukrainian	0
M6S	West Toronto	Runnymede	43.651571	-79.484450	4382	5155	42635	02.2	Ukrainian	3
M8X	Etobicoke	The Kingsway	43.653654	-79.506944	8780	3403	110944	01.8	Ukrainian	1
M4V	Central Toronto	Deer Park	43.686412	-79.400049	15165	10387	80704	01.1	Russian	4

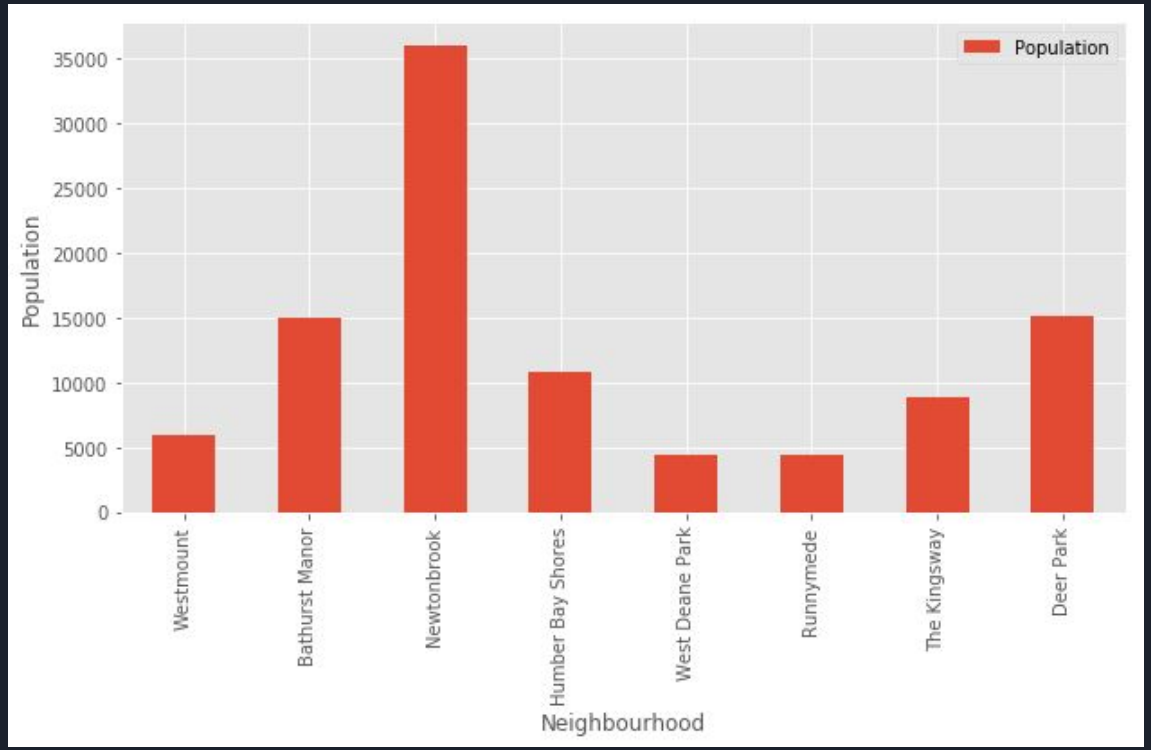
DATA RESULTS: K-MEANS CLUSTERING TARGET GROUP VENUE CLUSTERING

Average Income	Percentage	Language	Cluster Labels	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
35183	10.6	Ukrainian	7	American Restaurant	Grocery Store	Gym	Ice Cream Shop	Indie Movie Theater	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
34169	09.5	Russian	5	American Restaurant	Gastropub	Gym	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
33428	08.8	Russian	6	American Restaurant	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant	Pharmacy
39186	05.2	Russian	2	Gourmet Shop	Video Store	Grocery Store	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Pizza Place
41582	02.3	Ukrainian	0	American Restaurant	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant	Pharmacy
42635	02.2	Ukrainian	3	American Restaurant	Frozen Yogurt Shop	Video Store	Grocery Store	Ice Cream Shop	Intersection	Light Rail Station	Liquor Store	Mexican Restaurant	Playground
110944	01.8	Ukrainian	1	American Restaurant	Gym	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
80704	01.1	Russian	4	Gourmet Shop	Frozen Yogurt Shop	Gastropub	Video Store	Grocery Store	Gym	Ice Cream Shop	Indie Movie Theater	Intersection	River

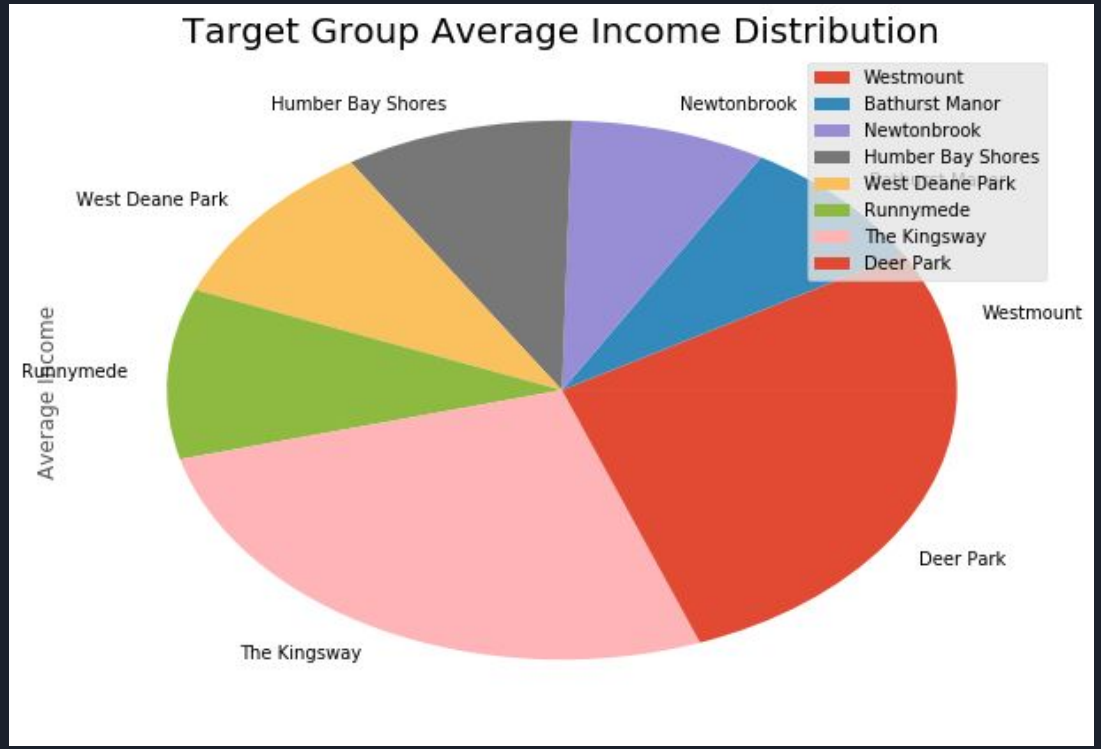
DATA RESULTS
TARGET GROUP CLUSTERS MAP



DATA RESULTS
TARGET GROUP POPULATION/NEIGHB



DATA RESULTS
TARGET GROUP INCOME DISTRIBUTION



DATA ANALYSIS
BUSINESS VENUE RECOMMENDATIONS

Etobicoke-West Deane Park	American Restaurant	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant	Pharmacy
Etobicoke-The Kingsway	American Restaurant	Gym	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
Etobicoke-Humber Bay Shores	Gourmet Shop	Video Store	Grocery Store	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Pizza Place
West Toronto-Runnymede	American Restaurant	Frozen Yogurt Shop	Video Store	Grocery Store	Ice Cream Shop	Intersection	Light Rail Station	Liquor Store	Mexican Restaurant	Playground
Central Toronto-Deer Park	Gourmet Shop	Frozen Yogurt Shop	Gastropub	Video Store	Grocery Store	Gym	Ice Cream Shop	Indie Movie Theater	Intersection	River
North York-Bathurst Manor	American Restaurant	Gastropub	Gym	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant
North York-Newtonbrook	American Restaurant	Ice Cream Shop	Indie Movie Theater	Intersection	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant	Pharmacy
Etobicoke-Westmount	American Restaurant	Grocery Store	Gym	Ice Cream Shop	Indie Movie Theater	Italian Restaurant	Latin American Restaurant	Light Rail Station	Liquor Store	Mexican Restaurant

DISCUSSION

THERE ARE VERY INTERESTING TRENDS SHOWING UP N 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.

THERE IS ADDITIONAL POTENTIAL IN EMPLOYING A RECOMMENDER SYSTEM ALGORITHM TO FURTHER IMPROVE THE RECOMMENDATIONS REPORT BASED ON A NUMBER OF BUSINESS REQUIREMENTS.

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.