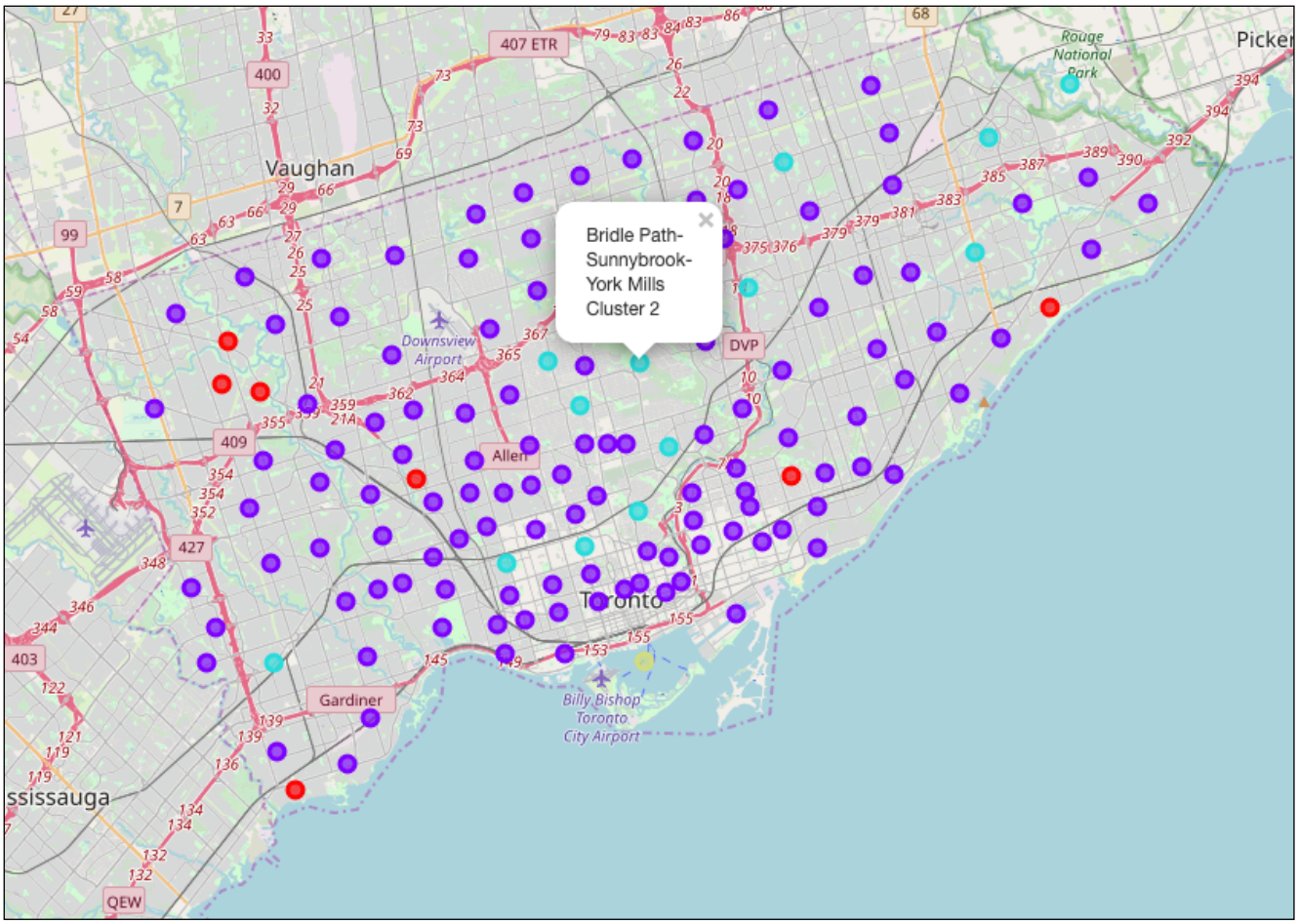
The Battle of Neighborhoods

-Finding the best location for a business

Andrew Zou - July 29, 2019



1. Introduction

1.1 The Background

In a big International city like Toronto, how do you find the best location for a business of Children and Youth Learning Center?

First, let us have big picture about city of Toronto.

"The strength of Toronto is that we have such a diversified economy."

— Eva Pyatt, director of business services for the city's economic development and culture division

City of Toronto's strong economy is a natural draw for entrepreneurs. There are following economic factors make this city becoming a most desired business destination.

1. Ontario has a low net debt-to-GDP ratio, which helps keep taxes down. Early in 2010, the international professional auditing firm KPMG, based in the Netherlands, assessed 41 large cities worldwide on their general tax competitiveness, including corporate income taxes and statutory labor costs. It ranked Toronto fifth, above New York, Los Angeles, and London.
2. There are many local business incentives available,. They include resource conservation and energy efficiency incentives offered through the city's better buildings and funding for early stage businesses from the province.
3. Population of Toronto are highly educated; 64 percent of people aged 25 to 64 have a post-secondary degree. The city's population is also intensely diverse. Half of the 5.1 million people in the greater metropolitan area were born elsewhere.
4. City has advanced transportation system, which includes subways, highways connect to US highways, and international airport.

1.2 The Business Problem

A small investor team planning to open a Youth and Children Learning Center in City of Toronto, Ontario, Canada. Being that Toronto is the most populated city in Canada, and continually ranks as an important global city based on a high quality of education, the choice to expand business into the neighbor was an easy selection for the investor team. However, with limited information of the Toronto market, this team needs us to assist in the selection of which area of Toronto will facilitate their Learning Center.

This team would like to find a location which meet following criteria:

(1) A neighborhood has above average population of Children, Youth.

(2) The household after tax income is above the average.

(3) Above average population has college degree in this neighborhood

(4) Business venues is in high demand in this neighborhood

With these criteria given by the investor team, based on previous success in other markets, the objective is to locate and recommend to the investors, the target audience, which neighborhood(s) of Toronto will be the best choice to operate their **Youth and Child Learning Center**. The information gained will assist in choosing the right location by providing data about the population of each neighborhood, in addition to other established venues present in these areas.

2. Data acquisition and cleaning

2.1 Data Sources

2.1.1 City of Toronto Neighborhood Profiles for providing an overview of the neighborhoods in Toronto (https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/).

2.1.2 City of Toronto Open Data Catalogue : The Census of Population is held across Canada every five years (the last being in 2016), and collects data about age and sex, families and households, language, immigration and internal migration, ethnocultural diversity, Aboriginal peoples, housing, education, income, and labor. City of Toronto Neighborhood Profiles use this Census data to provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighborhood. The profiles present selected highlights from the data, but these accompanying data files provide the full data set assembled for each neighborhood (https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#8c732154-5012-9afe-d0cd-ba3ffc813d5a).

2.1.3 Here is the Data File in CSV format (https://www.toronto.ca/ext/open_data/catalog/data_set_files/2016_neighbourhood_profiles.csv). Each data point in this file is presented for the City's 140 neighborhoods, as well as for the City of Toronto as a whole. The data is sourced from several Census tables released by Statistics Canada. The general Census Profile is the main source table for this data, but other Census tables have also been used to provide additional information.

2.1.4 Foursquare API (https://developer.foursquare.com/) to collect information on other venues/competitors in the neighborhoods of Toronto.

2.2 Data cleaning

2.2.1 Preparing the Census DataSet

The original massive Census data file provides 2,383 rows with 145 columns. In order to solve our problem with the essential data. A subset of the Census dataset was created with only extracted entries as following,

(1) Neighborhood code

(2) Children Population

(3) Youth Population

(4) Household Income After Tax

(5) Population holding Bachelor's degree

Here is the view of the dataset sample.

	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village	Bayview Woods-Steeles	Bedford Park-Nortown	Beechborough Greenbroo
0	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	42	34	76	52	49	39	11
9	Population	Age characteristics	Census Profile 98-316-X2016001	Children (0-14 years)	398,135	3,840	3,075	1,760	2,360	3,605	2,325	1,695	2,415	1,515	4,555	1,12
10	Population	Age characteristics	Census Profile 98-316-X2016001	Youth (15-24 years)	340,270	3,705	3,360	1,235	3,750	2,730	1,940	6,860	2,505	1,635	3,210	85
1026	Income	Income of households in 2015	Census Profile 98-316-X2016001	Average after-tax income of households in 20...	81,495	427,037	278,390	168,602	792,507	493,486	251,583	352,218	354,894	253,036	720,203	109,88
1709	Education	Highest certificate, diploma or degree	Census Profile 98-316-X2016001	Bachelor's degree	534610	4380	4210	1660	9135	6500	3075	8370	6055	3060	5655	44

Now, it is time to drop certain columns which are not used for data analysis work, then transpose the dataset as following view to have it arranged with particular neighborhood as the first column.

	Neighbourhood	Neighbourhood_Code	Children	Youth	Avg_Household_Income	College_Educated
0	Agincourt North	129	3,840	3,705	427,037	4380
1	Agincourt South-Malvern West	128	3,075	3,360	278,390	4210
2	Alderwood	20	1,760	1,235	168,602	1660
3	Annex	95	2,360	3,750	792,507	9135
4	Banbury-Don Mills	42	3,605	2,730	493,486	6500

At this point, one more thing need to be done is convert all the string format digits to integer type, then, we can say the Census dataset is ready to for data processing and analyzing.

2.2.2 Prepare for geographic information of 140 Neighborhoods

The original data loaded from a CSV data file. After we drop certain columns, the dataset sample looks like as following. Keep in mind, these Latitude and Longitude are critical information for us to apply the Folium Map and FourSquare API calls

	AREA_SHORT_CODE	AREA_NAME	LATITUDE	LONGITUDE
63	1	West Humber-Clairville (1)	43.716180	-79.596356
20	2	Mount Olive-Silverstone-Jamestown (2)	43.746868	-79.587259
56	3	Thistletown-Beaumond Heights (3)	43.737988	-79.563491
40	4	Rexdale-Kipling (4)	43.723725	-79.566228
112	5	Elms-Old Rexdale (5)	43.721519	-79.548983

2.2.3 Merge DataSets

Merge these above two Census dataset and Neighborhoods Geographic dataset into one dataset for data analysis work as below.

Neighbourhood_Code	Neighbourhood	Children	Youth	Avg_Household_Income	College_Educated	LATITUDE	LONGITUDE
129	Agincourt North	3840	3705	427037	4380	43.805441	-79.266712
128	Agincourt South-Malvern West	3075	3360	278390	4210	43.788658	-79.265612
20	Alderwood	1760	1235	168602	1660	43.604937	-79.541611
95	Annex	2360	3750	792507	9135	43.671585	-79.404001
42	Banbury-Don Mills	3605	2730	493486	6500	43.737657	-79.349718

2.3 Summary

Every time while we go the raw data source, we have to apply the Data Science data manipulation skills to extract information, correct data, reformat data, in order to have the data ready for data analysis manipulations.

3. Exploratory Data Analysis

Before we move on let us ask one question. Who are the targeted customers? The customers are families who have children or youth in schools, who need develop certain learning skills, or participate some interest subjects, such as help children or youth to develop complex math skills, or reading, writing, or practical science thru discovering experiments etc., Also, these families have enough financial resource to support their children or youth to go to our Learning Center, and the parents of the household have a will to spend money for the children and youth. Identify our customers through the data process and analysis is one of our goals.

3.1 Census Data Analysis

Put it into our mind again, as mentioned in the earlier, following are business requirements to meet.

(1) A neighborhood has above average population of Children, Youth.
(2) The household after tax income is above the average.
(3) Above average population has college degree in this neighborhood.
(4) Business venues is in high demand in this neighborhood.

Before we start to analyze the data, we need ask ourself following question.

(1) Is there any existing pattern among these data?
(2) What type of correlation could be used to represent the pattern?
(3) Is there a mathematical model hidden inside the data?

To answer these three questions, the first thing to do is using Python matplotlib package to create data visualization plots.

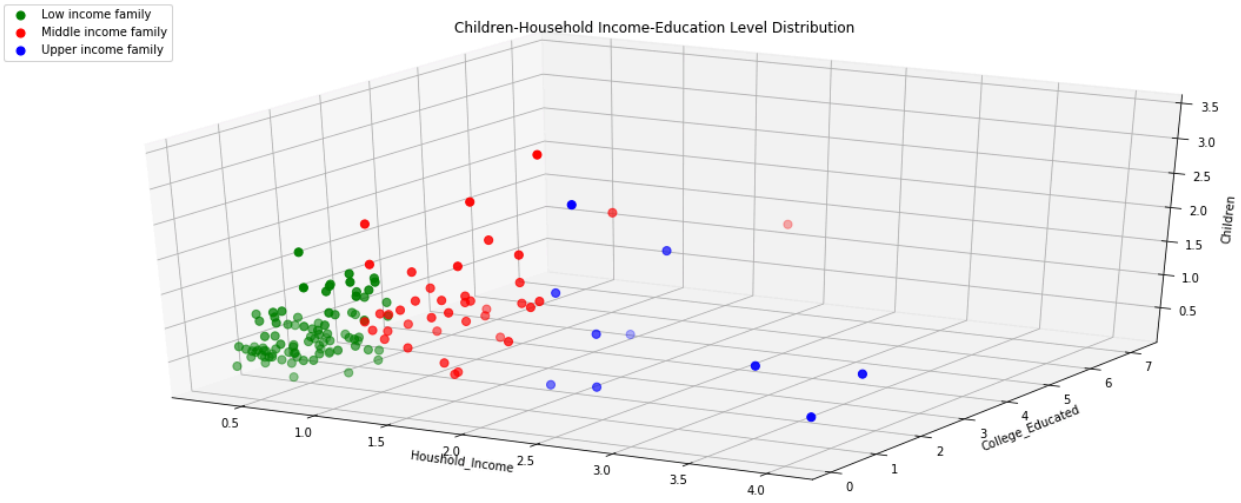
3.1.1 Normalize DataSet

We normalize the Census dataset by divided by mean number of each column. Here is the sample view of the normalized dataset.

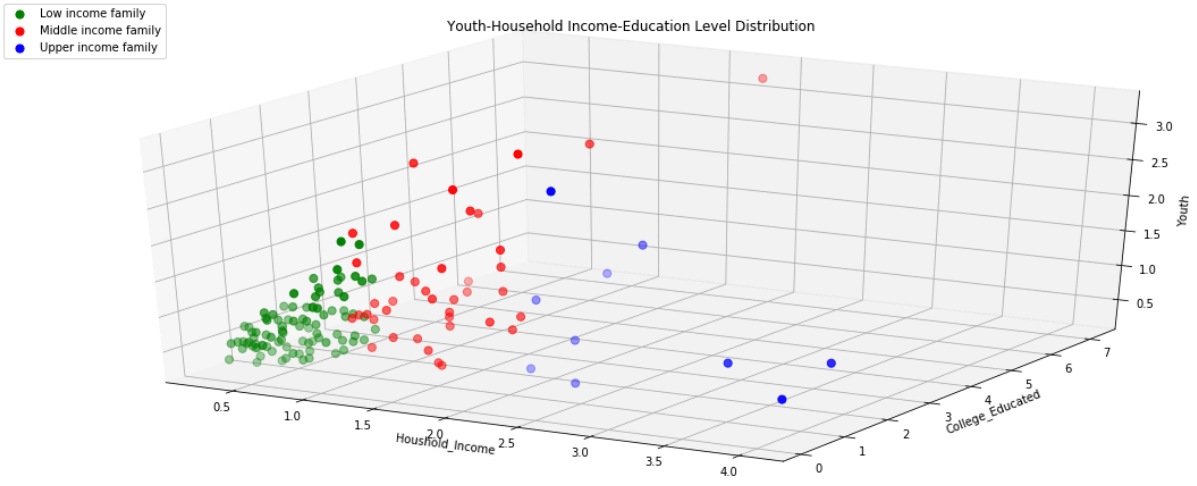
	Neighbourhood	Neighbourhood_Code	Children_Weight	Youth_Weight	Household_Income_Weight	College_Educated_Weight
0	Agincourt North	129	1.350211	1.524064	1.215673	1.146897
1	Agincourt South-Malvern West	128	1.081224	1.382147	0.792511	1.102383
2	Alderwood	20	0.618847	0.508021	0.479970	0.434669
3	Annex	95	0.829817	1.542575	2.256080	2.391987
4	Banbury-Don Mills	42	1.267581	1.122995	1.404838	1.702016

Based on the business common sense 101, we got to identify the dataset with questions as following. First, do people have the financial power for purchasing service products? Second, what is the demand for people to use the service products? Third, what is level of the urgency to use the service products? We definitely have no answer at the first glance of the dataset.

3.1.2 Observe the whole dataset distribution for Children and Youth Population
3.1.2.1 We created a 3D plot to observe the distribution of family with Children, Household income, people with college degree. In this plot the different color of dots represents the family financial wealth.



3.1.2.2 We created a 3D plot to observe the distribution of family with Youth, Household income, people with college degree. In this plot the different color of dots represents the family financial wealth.



3.1.2.3 From above two plots, we do understand the population distribution with the household income and education level. The middle class families have higher Children, Youth population compare to lower income families or upper level income families. However, we still not see the correlations between them.

3.1.3 Observe data pattern thru Scatter Plots

3.1.3.1 In common sense, we know the more financial resource the family have they more demand they have for education service products. We would like to see if this is true in our dataset. Here are the plots. Figure 1,

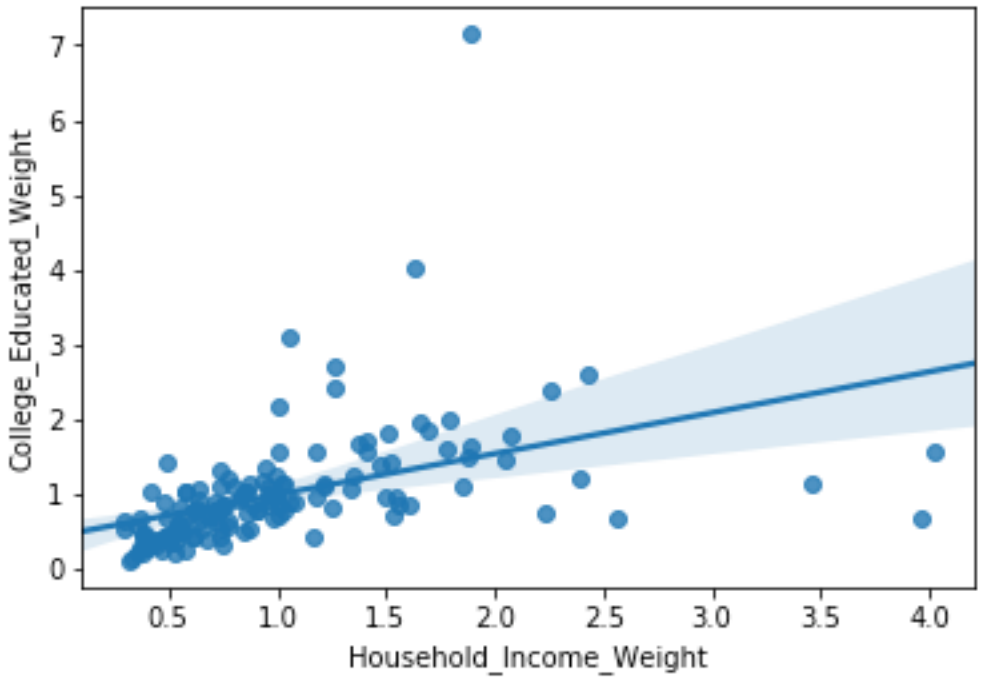


Figure 2

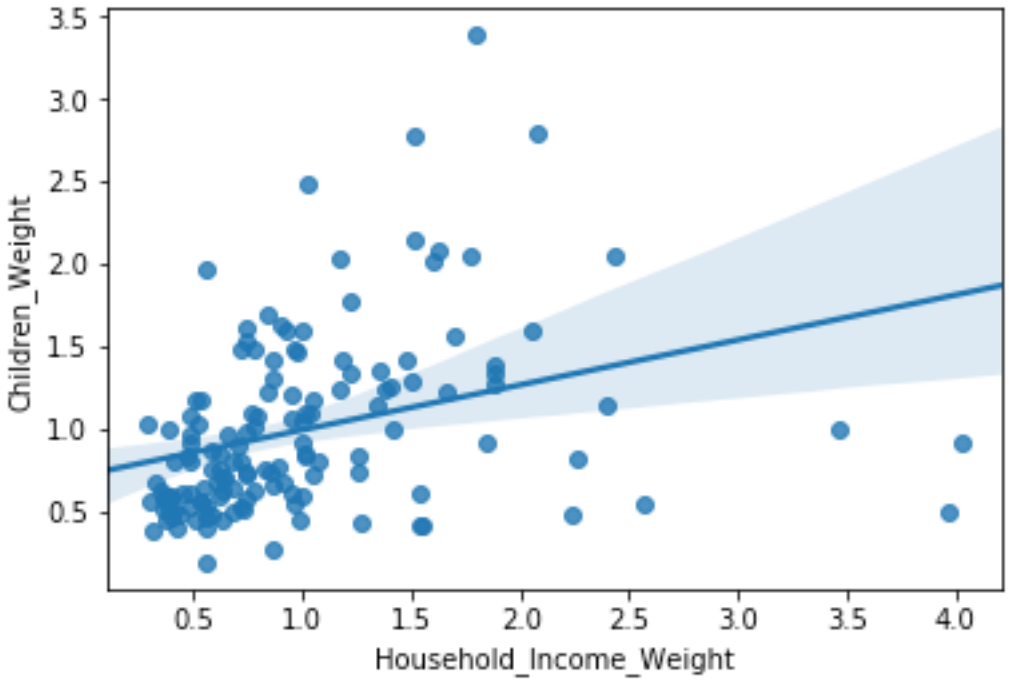


Figure 3

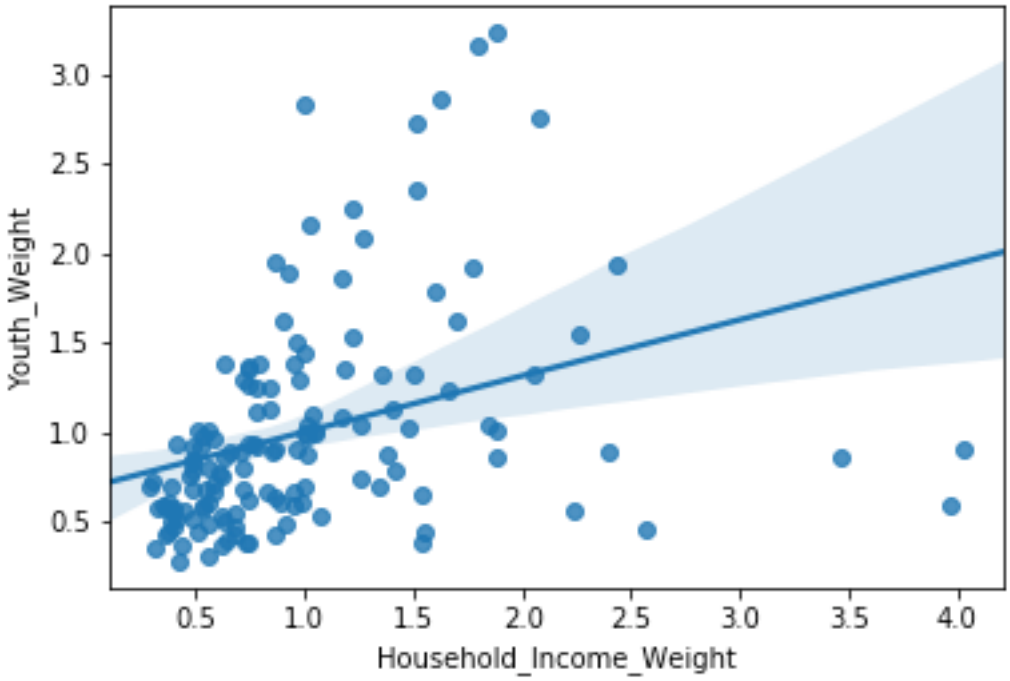
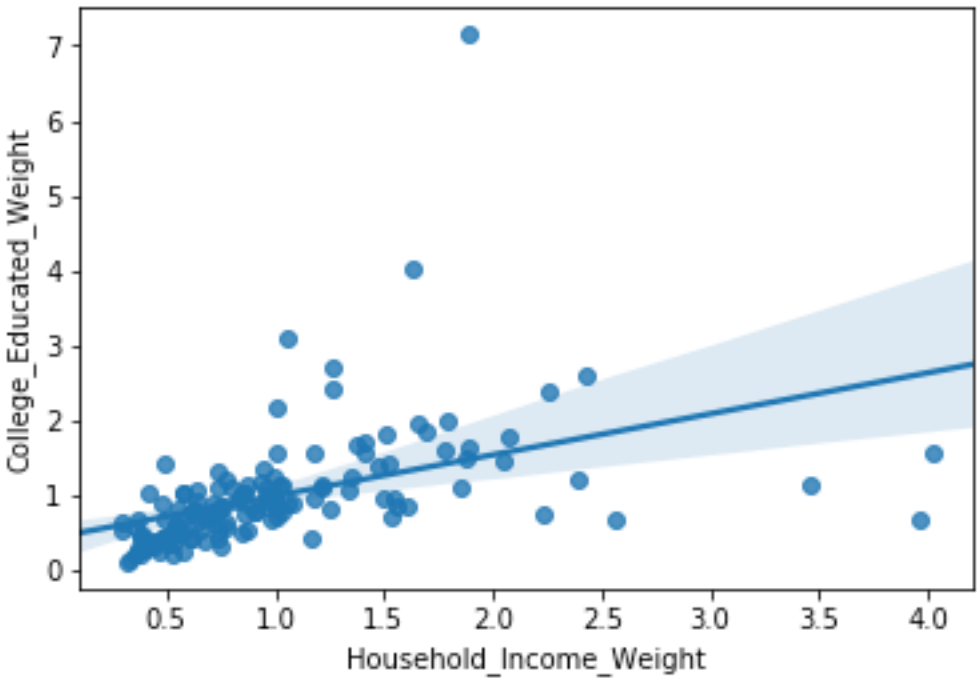


Figure 4



3.1.3.2 Summary of the observation

Through observing above plots, we saw the correlation in our data elements. This positive correlation helps to identify our customer group as we expected in the beginning.

The positive relation among the above plots proved our concern to choose the right customers based on the census dataset. Based this proof, we could build a mathematical model as our hypothesis to select the right location as the business location.

Hold on, before we move on to build our mathematical model, we seems remembered there is another factor is needed to achieve our goal, that is the neighborhood venues. Because, we definitely not like to be the only venues in the targeted neighborhood. Business is a social community activity, even though we do not like have another Learning Center in the same location plays as our competitor, but we do like too see more venues in the same location. That means something to us.

3.1.4 Neighborhood Venues Analysis

we understand the existing Venues status is a significant factor for business activities, for instance the volume of customer flow, the real estate cost, transportation service benefits etc. Fortunately, the Fousquare API is the tool that helps us to collective the corresponding venues for each particular location. Remember we have the geographic information, the Latitude and Longitude dataset can be used for our research. Here is a sample result dataset after we collected the Venus for all 140 neighborhoods.

Neighbourhood_Code	Neighbourhood	Children	Youth	Avg_Household_Income	College_Educated	Venue_Count
129	Agincourt North	3840	3705	427037	4380	40
128	Agincourt South-Malvern West	3075	3360	278390	4210	100
20	Alderwood	1760	1235	168602	1660	100
95	Annex	2360	3750	792507	9135	100
42	Banbury-Don Mills	3605	2730	493486	6500	100

3.1.5 Summary

Good work, at this point, we collected all data that could be used for our research work, such as children and youth population, household income, college degree population, local venues. Next, let us move on to build our mathematical model with all these criterial information.

4. The Mathematical model

It is time to build a mathematical model to let math works for our problem. In order to create a good quality mathematical model, we still need to manipulating the existing dataset.

4.1 Normalize the dataset

Because the data in the cleaned dataset have all kind of ranges, ignorer to have better understand of the data, we use standard normalization method to manipulate them, which every data element is divided by mean of the current row. Here is the sample of the dataset after normalization.

Neighbourhood_Code	Children_Weight	Youth_Weight	Household_Income_Weight	College_Educated_Weight	Venue_Weight
129	1.350228	1.524109	1.215673	1.146897	0.507292
128	1.081237	1.382188	0.792510	1.102383	1.268231
20	0.618854	0.508036	0.479970	0.434669	1.268231
95	0.829828	1.542620	2.256080	2.391987	1.268231
42	1.267597	1.123028	1.404838	1.702016	1.268231

4.2 The Mathematic Equation

Before the mathematic equation could be born, let us do some rational critical thinking of the business and the data items.

For example, one family have children, who get the information about our Learning Center, which we have the best program to help their children develop very complex math skills, that could help their children to achieve outstanding academic excellence. But first question in their parents mind, is how much does it cost? Second question is, if the family have the money, will the parents be happy to send their children to this program? The answer may depends the parents education level, higher degree educated parents might have the higher intention to push their children, compare to the rest of the parents may not have the opportunity for college education. Is a location good for have a Learning Center? The answer depends on two factors, first the Children and Youth population level, second, how is the business environment here, cheap labor, cheap real estate cost, cheap energy cost etc., There is one number could disclose this information is the local venues, the higher number of venues, the better location for business. Let us list our items in the above thinking.

(1) The Family Income Level

(2) The Parents Education Level

(3) The Children and Youth Population Level

(4) The Venues Level

If we our thinking in a mathematical way, we can say the determination of running the business depends on variables as below.
Under even situation for 100 as the score, every item could be weighted as 25. However, not all of these items are even. Let us set the mean of the score as 25, the highest priority one add extra 10 points to be 35, the second highest priority on add 5 points to be 30

(1) Very important item, the household income — x_1 ,
(2) Second important item, the parents education level — x_2,
(3) Third important item, the Children population level — x_3,
(4) Fourth important item. the Youth Population level — x_4,
(5) Fifth important item, the Venues Level — x_5

Here we will have a mathematic equation as below,

$$\mathbf{f(x_1, x_2, x_3, x_4, x_5) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5}$$

This is the mathematic equation that is used to set up the desired cluster label values. The higher the label value, the better the location for business.

4.3 Calculate the Complex Score
Regarding the mathematic equation, we calculated the complex score for each data entry. Here is the sample view of the clustered dataset.

Neighbourhood_Code	Children_Weight	Youth_Weight	Household_Income_Weight	College_Educated_Weight	Venue_Weight	Complex_Score
77	1.283420	3.225105	1.885505	7.143231	1.268231	3
51	2.081601	2.854876	1.628790	4.041634	0.481928	2
98	0.923007	0.898833	4.022852	1.573710	1.268231	2
131	2.798910	2.756148	2.075729	1.771406	0.342422	2
132	2.781329	2.723239	1.517900	1.431003	0.659480	2

4.4 Summary
We create a mathematical model to label each data entry with a label value, the range is [0, 3]. The higher the label value represents a better location for business.

5. Machine Learning

5.1 Build machine learning data model

In section 4, we build a mathematical model shaping the dataset. We can use Python Machine Learning packages to build a machine learning data model with the RandomForest Classifier algorithm. Here is how to build the machine learning model.

1. Splitting the dataset into two parts of train set and test set.
2. Training the RandomForestClassifier model
3. Save the model to file and deliver it to investor team

Here is the code snippets about creating, and training the data model

```
X = df_ML_model[['Children_Weight', 'Youth_Weight', 'Household_Income_Weight', 'College_Educated_Weight', 'Venue_Weight']].values
y = df_ML_model["Complex_Score"]
```

```
from sklearn.model_selection import train_test_split
X_trainset, X_testset, y_trainset, y_testset = train_test_split(X, y, test_size=0.3, random_state=3)
#Import Random Forest Model
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=100)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(X_trainset, y_trainset)

y_pred=clf.predict(X_testset)
```

```
print(y_pred)
```

```
[1 1 2 1 1 1 1 1 2 0 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1]
```

5.2 Discussion of this machine learning model

This model has its limitation to recognize the label value as 3 location, because there is only one data entry was calculated as label 3 in the existing data set. In the future, we need to make improvement on it to get more label 3 data entries to train this model. Under this limitation, if any predict value reach value 2, we can concluded it as a highly recommended business location.

5.3 Delivery trained data model

After the machine learning data model was trained, we can delivery it as a production model to this investor team. The team may use it to predict the rest of cities like Toronto in Canada. Let us take a look some of the test cases in the code snip below.

The code snippets below demonstrate how to save the trained data model, reload it to run other prediction.

Save the trained model for production purpose (delivery)

```
import pickle

filename_model = '/tmp/finalized_model.sav'
pickle.dump(clf, open(filename_model, 'wb'))
```

Load this model to run prediction test

```
[ ]  loaded_model = pickle.load(open(filename_model, 'rb'))
```

Test with a good location case

The label is 2, means a highly recommended location for business.

```
[ ]  print(loaded_model.predict([[1.0, 1.2, 3, 2, 1.5]]))
```

```
[2]
```

Try a worst case

The label is 1, means a location not good for business.

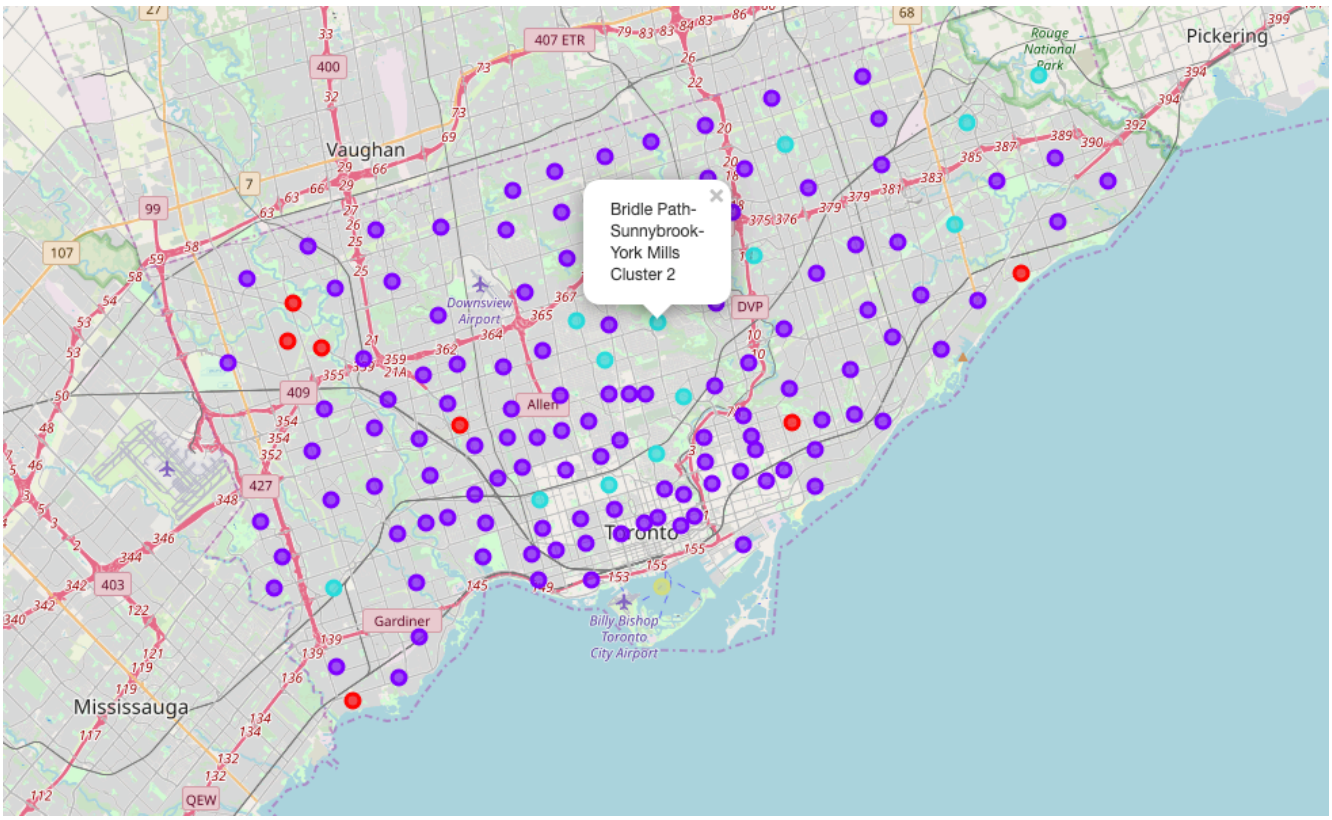
```
[ ]  print(loaded_model.predict([[0.1, 0.2, 0.8, 0.7, 0.5]]))
```

```
[1]
```

6.Business Analysis

6.1 Visualization through Folium map

After the 140 neighborhoods were classified by the mathematical model in last section. We can create a Folium map to present these classified area in the Toronto map. Here is the view of the map.



In the above map the red color dot tells us these are areas were labeled as 0 value locations. Value of 0 means it is definitely not a location for business. The purple color dot means these area were label as 1 value locations, which also tells us these are the areas not good for business. We one yellow dot location was label as value 3, which it is the best location for business. The rest of blue color dot areas are highly recommended locations for business.

6.2 Business analysis

With the labeled areas with yellow color and blue color, we double check these perfect or highly recommended locations having competitions in the same area. Fortunately, there are

no existing competitors in these area. Naturally, these are the location we would suggest to operate the Learning Center business. Let us review these locations in the map again.

6.3 The Final List

Through all the research works had been done, we finally achieve our goal, the final list. Here it is.

- **Waterfront Communities-The Island**
- **Annex**
- **Bedford Park-Nortown**
- **Dovercourt-Wallace Emerson-Junction**
- **Rosedale-Moore Park**
- **Lawrence Park South**
- **Leaside-Bennington**
- **L'Amoreaux**
- **Woburn**

7. Discussions

After review the whole research process again, we realize there are certain places limited our works. If we could do the following, a better quality final list, and trained machine learning model will be delivered.

1. We took Census data as major data resource. There are business data source could be help us , such as local real estate data, local school data etc.
2. The Foursquare API version could limit our view of the newer competitors in these areas.
3. The mathematical model need to be optimized to generate more layers of the classified labels.