Final Report: The best place for residential income property in Toronto (from family perspective)

By Veronika K. Zaharidova June 15, 2020

Problem Description and Background

According the Canadian statistics[1] Toronto has an estimated population of over 2.8 million in 2016, which makes it the 4th most populous city in North America and the most populous city in Canada. It is also the largest urban and metro area, with a population density of 4,149.5 people per square kilometer. According to the 2006 Census, foreign-born people have been 45.7 % of the population of Toronto[2], which convert it in the second-highest percentage of foreign-born residents of all world cities after Miami. The data related to the population and its diversity made me believe that Toronto attracts a lot a people from all over the world for business, tourism, etc. and the property business and the residential investment properties can be pretty rewarding if the right choice is made.

Also, the expectations of possible financial crisis, due to the Coronavirus would pull the properties` prices down and would cause an increase in the alternative to standard market investments (as investments in income properties).

The first step and one of the key factors that has to be considered when shopping for an income property is the review of the neighborhood's livability and facilities. After that few key factors has to be considered as well: the neighborhood vacancy rate, the local selling prices, and the average rent in the area in order to determine the financial feasibility.

In order to find the best place for residential investment properties, an analysis of the neighborhoods in Toronto will be made. The primary goal of this project will be to evaluate and determine which neighborhood/s would be more appropriate for investment in property. As they are a lot of types of potential tenants and the neighborhood in which the property would be bought will determine the types of tenants the perspective of a family/couple as a potential tenants will be taken (based on the assumptions that families or couples are generally better tenants than singles because they are more likely to be financially stable and pay the rent regularly).

Another assumption taken into consideration is that the home values can be affected by the 'energy' of the neighborhood in which they are located. And form the family/couple perspective the affects will be limited to: more green spaces, parks, restaurants, theaters; less breaks and enters, less nightclub and bars, more child care spaces, the financial status of the neighbors (debt risk score and income after taxes), total business establishments, total local jobs. Also, the home prices by neighborhoods will be revised, in order to help the potential investors (buyers).

This analysis is meant to be helpful for potential property investors or families trying to find the right neighborhood in Toronto, where to move/live. According to one of the newest real estate forecasts 'prices are still trending upward, but Coronavirus containment efforts pull prices down. It is likely that prices will be lower in 2021.'[3]so probably more people will take advantage of this report.

[1] https://www12.statcan.gc.ca/census-recensement/2011/as-sa/fogs-spg/Facts-csd-eng.cfm?Lang=eng&GK=CSD&GC=3520005

[2] https://www12.statcan.gc.ca/english/census06/analysis/immcit/charts/chart4.htm

[3] 05.05.2020; https://www.mortgagesandbox.com/toronto-real-estate-forecast

Data Description

The data that will be used to make a 'profile' of the neighborhoods is the data used during the assignment form week 3 (the Foursquare location data will be used in order to find the neighborhoods with less nightclub and bars, with more parks, restaurants, theaters etc.) combined with more data from this Canadian open data portal: https://open.toronto.ca/catalogue/?search=neighbourhood&sort=score%20desc.

The information related to child care spaces, debt risk score, household income after taxes, total business establishments, total local jobs, home prices by neighborhoods will be extracted from the open data portal and properly handled, arranged and converted into pandas data frame and combined with the data form Foursquare in order to help us find the 'right' neighborhood or group of neighborhoods for rental property form family perspective.

The following data has been used:

Open Data
Catalogue: https://open.toronto.ca/catalogue/?search=neighbourhood&sort=score%20desc;
Toronto Neighbourhoods: https://open.toronto.ca/dataset/neighbourhoods/ (the data related to neighborhoods names and latitude and longitude will be used);
Wellbeing Toronto - Economics: https://open.toronto.ca/dataset/wellbeing-toronto-economics/ (the information about businesses, child care spaces, debt risk score, home prices and local employment by neighborhood);
Wellbeing Toronto - Demographics: https://open.toronto.ca/dataset/wellbeing-toronto-demographics/ (information about total population, population by age (0-14 and 15-24 years);
Neighborhood Profiles: https://open.toronto.ca/dataset/neighbourhood-profiles/ (the data by neighborhoods related to Average household size, Couples with children, Total income by household, Private households by tenure (owner, renter));
Wellbeing Toronto - Safety: https://open.toronto.ca/dataset/wellbeing-toronto-safety/ (the data by neighborhoods related to breaks and enters);
Wellbeing Toronto - Environment: https://open.toronto.ca/dataset/wellbeing-toronto-environment/ (the data by neighborhoods related to green spaces);
Foursquare Developers Access to venue data: https://foursquare.com/ (the data related to nightclub and bars, parks, restaurants, theaters, etc.);

Table 1: Data description

PROVENANCE	SHORT_NAME	DESCRIPTION
Statistics Canada Census	Child 0-14	Calculated as a sum of the 0-4 years, 5-9 years and 10-14 years age cohorts.
Statistics Canada Census	Youth 15-24	Calculated as a sum of the 15-19 years and 20-24 years age cohorts.
Toronto City Planning	Local Employment	For Reference Period 2011: City of Toronto, City Planning, Policy and Research, Research and Information unit. Toronto Employment Survey, 2011 data. Total local employment (jobs), persons aged 15+ years.
City of Toronto, Economic Development Culture & Tourism	Businesses	For Reference Period 2011: City of Toronto, City Planning, Policy and Research, Research and Information unit. Toronto Employment Survey, 2011 data. Total number of licensed business establishments.
City of Toronto, Children's Services	Child Care Spaces	The number of spaces for children from birth to senior kindergarten and the number of spaces for school age children grade 1 and up. CONTEXT: A quality licensed child care program provides a safe environment for children to learn and socialize through play and exploration under the guidance of qualified Early Childhood Educators. It is a necessary factor in the economic well-being of families because it allows parents to participate in the workforce or attend school which prepares them for the paid workforce so they can provide their families with food, clothing and shelter. LIMITATIONS: Approximately 1/3 of families choose child care outside their neighborhood. Some neighborhoods tend to have more spaces as commuters bring their children to centers near their work or along their commuting route. For Reference Period 2011: City of Toronto, Children's Services. Licensed child care spaces in 2011. Context and limitations as above.
For Reference Period 2011: TransUnion 2012 via the Community Data Program.	Debt Risk Score	Methodological Notes: The Risk Score is a proprietary index value provided by TransUnion Canada that indicates the likelihood of missing three consecutive loan payments. Low-value scores (<707) indicate a High Risk of missing 3 consecutive loan payments; High-value scores (769+) indicate a low risk. These risk scores are calculated for non-mortgage consumer debt such as lines of credit, credit cards, automobile loans and installment loans. TransUnion data is provided by postal code and covers approximately 92% of all Canadians with credit files. For privacy reasons, postal codes with fewer than 15 credit files are suppressed. TransUnion dataset provided by the Community Data Program (www.communitydata.ca).
Realosophy.com.	Home Prices	Average price for residential real estate sales during the period 2011-2012, in Canadian dollars. Data collated by Realosophy.com.

Toronto Police Service	Break & Enters	Data Source: ECRIME Database, CIPS Database (* Drug Charges are based on CIPS Cases where the Status is either Closed or Approved.) as of 2009.07.24. Qualifiers: The Toronto Police Service (TPS) Crime Information Analysis Unit does not make any warranty, representation or guarantee as to the content, accuracy or completeness of any of the statistics provided in this report. The aggregated data provided in this report is based ONLY on the records that were geocoded. The statistics represented in this report should never be used to attempt to reconstruct Uniform Crime Reporting numbers, other TPS official numbers or to correlate events to a specific location. Please refer to the Publications section of the TPS Website at www.torontopolice.on.ca/publications/ for official statistics.
City of Toronto, Parks Forestry & Recreation	Green Spaces	For Reference Period 2011: City of Toronto, Parks Forestry & Recreation, Jan 2012 data. Extracted from GCC SDE geospatial repository (layer TCL3.UPARK) in March 2013. Total land area (in square kilometers) designated as parkland or green space (including utility corridors and utility areas such as soccer fields).
Census Profile 98-316-X2016001	Average household size	
Census Profile 98-316-X2016001	Couples with children	
Census Profile 98-316-X2016001	Average after-tax income of households in 2015 ($)	
Census Profile 98-316-X2016001	Owner	Total - Private households by tenure - 25% sample data
Census Profile 98-316-X2016001	Renter	As number of renters as part of Total - Private households by tenure - 25% sample data

Methodology

The analysis begins with research of all variables found in Toronto`s open portal related to neighborhoods and their profiles and Initial data analysis (focused on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed). After that, based on some assumptions only the relevant data has been added and past trough descriptive statistics (using the functions .corr() and .describe()) in order to find out the relationships between different values and reduce the number of the data to be used. The Pearson Correlation Coefficient and P-value has been calculated for some of the categories, in order to deeper the investigation. Scatterplots are used their visualization. When the initial dataset has been defined, the analysis has spread in 3 main stages (parts):

- 1st stage: Checking the data important for investors in order to find out the clusters of neighborhoods good enough for investment in residential property for rent.

- 2nd stage: Finding the clusters of neighborhoods that are the best from family perspective (the potential tenants) – neighborhoods with more families with kids, green spaces and child care spaces.

- 3rd stage: From the 140 neighborhoods at the beginning of the study only the neighborhoods that met the criteria above have been clustered based on the Foursquare data (with less pubs and bars and more gardens, garden centers and playgrounds in order to finalize this research.

Each stage begins with exploratory data analysis – finding the five number summary and visualized by boxplots in order to find out the mean, standard deviation, minimum, maximum and the quartiles and for better understanding of the data. The distribution of the data set has been standardized by StandardScaler function from the sklearn library (making the values lie in the same range). After that, the Elbow method has been used in order to find the optimal k (number of clusters) for the k-mean clustering unsupervised machine learning algorithms. Plots (line) have been used for visualization of the distortions and inertias. The following step on each stage was using the k-mean clustering algorithm to find the clusters, finding the mean of each variable, choosing the suitable ones and using folium to visualize them on the map of Toronto.

The most important libraries used in the study are pandas, json, folium, matplotlib, scikit-learn, numpy, etc.

Results

In the initial data analysis, the data form Toronto`s open data portal was loaded, transformed into pandas dataframe and handled properly. As there was a huge statistical dada in each table, related to different categories (economy, demography, environment, etc.), after review of the dada description, only the meaningful data was used for further analysis. The irrelevant data has been identified and expunged from further consideration. As no missing data was considered important and systematical in the features that were used, no actions of exclusion of missing data were taken. After the preprocessing part, the transformation of the data was made by merging the data form different tables and the new data frame the **actual output data** (after the initial transformation) for the study looked like:

Neighbourhood	Businesses	Child Care Spaces	Debt Risk Score	Home Prices	Local Employment	Neighbourhood Id	Average household size	Couples with children	Average after-tax income of households in 2015 ($)	Private households by tenure - 25% sample data)	Child 0-14	Youth 15-24	Break & Enters	
West Humber-Clairville	2463	195	719	317508	58271	1	3.20	4490	426156	3275	5960	5400	175	2.0
Mount Olive-Silverstone-Jamestown	271	60	687	251119	3244	2	3.32	4365	360648	5455	7665	5015	61	1.0
Thistletown-Beaumond Heights	217	25	718	414216	1311	3	3.09	1310	140050	1245	1815	1350	36	0.9
Rexdale-Kipling	144	75	721	392271	1178	4	2.69	1300	134305	1685	1590	1315	32	0.2
Elms-Old Rexdale	67	60	692	233832	903	5	2.93	1085	123119	1470	2110	1380	25	0.7

The correlation between the categories of the output data has been calculated:

```
In [32]: #cheking the statistical info and correlation between the data
         df_merge3.corr()
```

	Businesses	Child Care Spaces	Debt Risk Score	Home Prices	Local Employment	Neighbourhood Id	Average household size	Couples with children	Average after-tax income of households in 2015 ($)	Renter (Total - Private households by tenure - 25% sample data)	Child 0-14	Youth 15-24	Break & Enters	Green Spaces
sinesses	1.000000	0.140211	0.005309	-0.105364	0.880345	0.021246	-0.205106	0.240902	0.233775	0.572459	0.165388	0.492639	0.628717	0.112223
ild Care Spaces	0.140211	1.000000	0.243536	0.045469	0.007367	-0.030855	0.019704	0.388271	0.310785	0.159968	0.325207	0.315077	0.329445	0.131324
ebt Risk Score	0.005309	0.243536	1.000000	0.625700	0.045077	0.001592	-0.139360	0.037327	0.480309	-0.164009	-0.232329	-0.148715	0.011723	0.003773
e Prices	-0.105364	0.045469	0.625700	1.000000	-0.069937	-0.031553	-0.247828	-0.223786	0.617351	-0.159242	-0.309772	-0.284765	-0.015540	-0.072070
Local loyment	0.880345	0.007367	0.045077	-0.069937	1.000000	-0.032609	-0.239622	0.123170	0.177289	0.530400	0.055128	0.366426	0.435583	0.090642
ourhood Id	0.021246	-0.030855	0.001592	-0.031553	-0.032609	1.000000	-0.012570	0.123946	0.025649	0.038748	0.089951	0.139166	0.057479	0.060945

The result, was:

- strong correlation between:

  Businesses and Local employment (~ 0,88);
  Couples with children and Child (~ 0,909) and Youth (~ 0,876);
  Youth and Child (~ 0,845);

- moderately strong correlation between:

  Home prices and Debt Risk (~ 0,625);
  Home prices and Average Income (~ 0,61);
  Income and Debt Risk (~ 0,617);
  Break and Enters and Youth (~ 0,688);
  Break and Enters and Businesses (~ 0,63);

- moderate correlation between:

  Renters and Businesses (~ 0,572);
  Number of renters and Local employment jobs (~ 0,53);
  Renters and Average household size (~ -0,52);
  Renters and Youth (~ 0,475);

Further checks have been made (calculating the Pearson Correlation Coefficient and P-value for some of the categories) and visualizing them, using matplotlib.

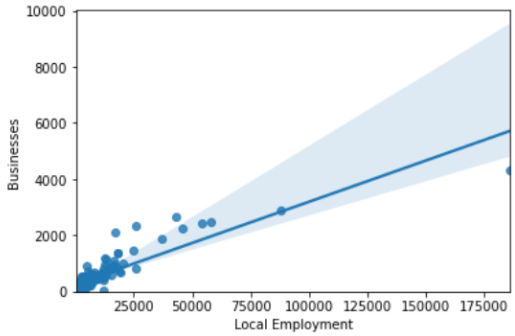
- calculating the Pearson Correlation Coefficient and P-value for Businesses and Local employment (~ 0,88);

```
In [34]:
#Let`s calculate the Pearson Correlation Coefficient and P-value of 'Businesses' and 'Renters'.
pearson_coef, p_value = stats.pearsonr(df_merge3['Businesses'], df_merge3['Local Employment'])
print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)

The Pearson Correlation Coefficient is 0.8803449209449341  with a P-value of P = 1.5321538125140373e-46
```

```
In [36]:
# Let`s visualize the correlation (Plotting the correlation)
sns.regplot(x='Local Employment', y= 'Businesses', data=df_merge3)
plt.ylim(0,)

Out[36]: (0.0, 10006.701630520374)
```

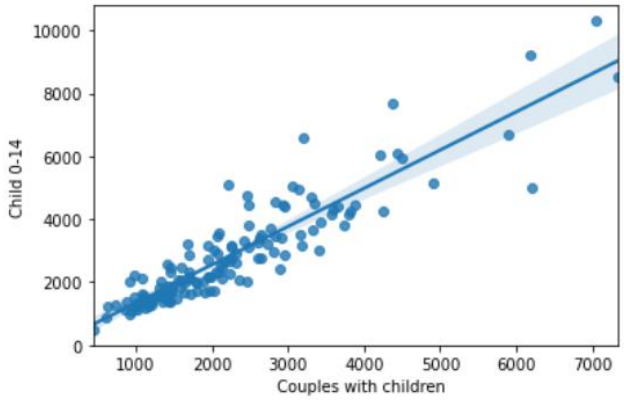


Since the p-value is $<$ 0.001, the correlation between both categories is statistically significant, and the linear relationship is quite strong (~0.88). So, only the local employments jobs will be used as further variable. And the conclusions made for the local employments can be considered strongly relevant to businesses (as well).

- calculating the Pearson Correlation Coefficient and P-value of 'Couples with children' and 'Child 0-4':

```
In [38]: # Let`s visualize the correlation (Plotting the correlation)
         sns.regplot(x='  Couples with children', y= 'Child 0-14', data=df_merge3)
         plt.ylim(0,)
```

```
Out[38]: (0.0, 10787.751882342092)
```



As result, since the p-value is $<$ 0.001, the correlation between both categories is statistically significant, and the linear relationship is strong (~0.91), only the first variable (couples with children will be used for further analysis (in order to reduce the volume of the data and to facilitate the analysis).

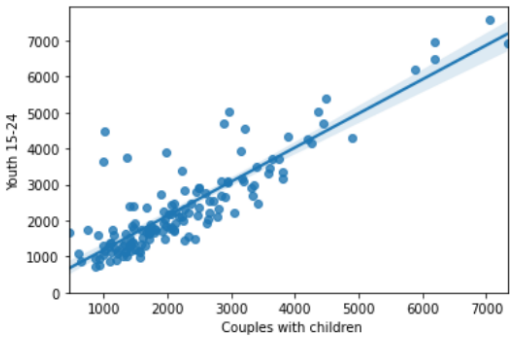
- calculating the Pearson Correlation Coefficient and P-value of 'Couples with children' and 'Youth 15-24':

```
# Let`s calculate the Pearson Correlation Coefficient and P-value of 'Couples with children' and 'Youth 15-24'.
pearson_coef, p_value = stats.pearsonr(df_merge3[ '  Couples with children'], df_merge3['Youth 15-24'])
print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)
```

```
The Pearson Correlation Coefficient is 0.8756675966827706  with a P-value of P = 1.8281427985670998e-45
```

```
# Let`s visualize the correlation (Plotting the correlation)
sns.regplot(x='  Couples with children', y= 'Youth 15-24', data=df_merge3)
plt.ylim(0,)
```

```
40]: (0.0, 7939.412966766886)
```



In conclusion, since the p-value is $<$ 0.001, the correlation between both categories is statistically significant, and the linear relationship is quite strong (~0.88), only the couples with children will be used for further analysis.

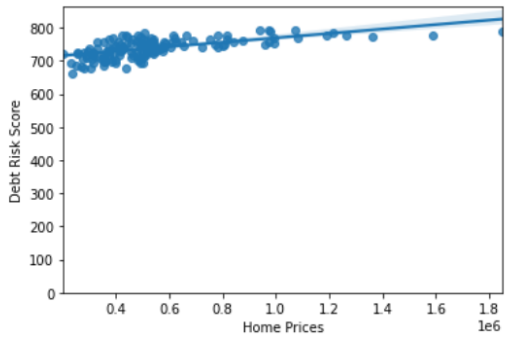
- Calculating the Pearson Correlation Coefficient and P-value of 'Home Prices' and 'Debt Risk Score'.

```
# Let's calculate the Pearson Correlation Coefficient and P-value of 'Home Prices' and 'Debt Risk Score'.
pearson_coef, p_value = stats.pearsonr(df_merge3['Home Prices'], df_merge3['Debt Risk Score'])
print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)
```

```
The Pearson Correlation Coefficient is 0.6256998028579022  with a P-value of P = 1.3935401603796514e-16
```

```
# Let`s visualize the correlation (Plotting the correlation)
sns.regplot(x='Home Prices', y= 'Debt Risk Score', data=df_merge3)
plt.ylim(0,)
```

```
42]: (0.0, 862.832922234318)
```



Since the p-value is < 0.001, the correlation between both categories is statistically significant, and the linear relationship isn`t very strong (~0.63), but the relation between both categories that is interesting and can be used.

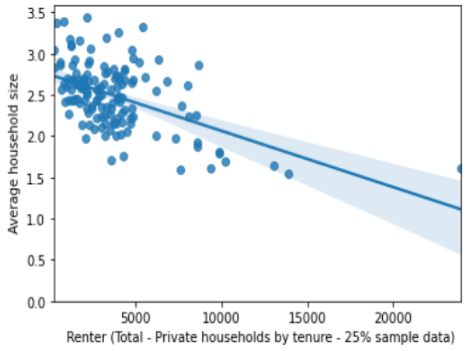
- Checking the Pearson Correlation Coefficient and P-value of Renters and Average household size

```
# Let`s check the  Pearson Correlation Coefficient and P-value of Renters and Average household size
pearson_coef, p_value = stats.pearsonr(df_merge3['  Renter (Total - Private households by tenure - 25% sample data)'], df_me
print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)
```

```
The Pearson Correlation Coefficient is -0.5201209320563664  with a P-value of P = 4.5164218002695466e-11
```

```
# Let`s visualize the correlation (Plotting the correlation)
sns.regplot(x='  Renter (Total - Private households by tenure - 25% sample data)', y=' Average household size', data=df_merg
plt.ylim(0,)
```

```
]: (0.0, 3.583911790702737)
```



In conclusion, the p-value is $<$ 0.001, the correlation between both categories is statistically significant, but the negative linear relationship isn`t very strong (~ -0.52).

Based on the correlations above, the output data for the analysis has been reduced to: Local Employment, Debt Risk Score, Home Prices, Average after-tax income of household, Renter (Total - Private households by tenure - 25% sample data), Break & Enters, Couples with children, Child Care Spaces and Green Spaces. This data has been separated in 2 parts (for stage 1and 2 respectively): Local Employment, Debt Risk Score, Home Prices, Average after-tax income of household, Renter (Total - Private households by tenure - 25% sample data), Break & Enters and Couples with children, Child Care Spaces and Green Spaces.

After the preparation and separation of the data, the next step was segmenting and clustering the neighborhoods, 1st form investors perspective and 2nd form family perspective.

Segmenting and clustering the neighborhoods in Toronto – in 3 different stages

The segmentation helps me divide the Toronto's neighborhoods into groups with similar characteristics. The entire process includes 3 stages:

- 1st stage: Finding the clusters of neighborhoods suitable for buying a residential investment property for rent (the investors perspective).

- 2nd stage: Finding the best clusters of neighborhoods from family perspective (the potential tenants) – neighborhoods with more families with kids, green spaces and child care spaces.

- 3rd stage: Clustering the Toronto`s neighborhoods based on the Foursquare data (with less pubs and bars and more gardens, garden centers and playgrounds in order to finalize this research.

1st stage

The key purpose of the study of stage 1 is to help the potential property investor in their strategy of targeting these specific groups of profitable neighborhoods. In order to segment the 140 neighborhoods, only the data related to debt risk, local employment jobs, average income, house prices and breaks and enters is been used. The key question that has to be answered is "Which neighborhood/s is/are the best for residential property investment?".

Here the data is used to show the potential investors, if the neighborhood is good enough to invest. The assumption is that a neighborhood that is good for investment in residential property must have more financially independent people (as potential renters) with bigger incomes and less debts. Also, a good area for such investment is a neighborhood with more business establishments and local employment jobs, so the clusters, where the local employment is near or above the average (9409.35) and the debt risk is below the average. The information about the level of home prices, won`t be used to determine the suitable neighborhoods, because this decision will be part of the concrete investor. Also, as the safety of the neighborhoods is important from potential investors and potential tenants, the neighborhood/s with less beaks and enters will be preferred. breaks and enters was used to find the neighborhoods with less crimes of this type. The main reason here is to show the potential investors the most secure neighborhood and also, the place where the insurance price would (hypothetically) be smaller, because this affects directly their potential winnings (rent > mortgages+ tax and fees+insurance +additional costs). Other question, waiting for answer is "where are the neighborhoods whit more renters?". The number of renters can be used from future investors, to figure out where the number of potential tenants is bigger. Also, the more renters in the neighborhood, may result in bigger future competition, so the clusters with moderate to high number of renters will be considered potentially good for residential property investment.

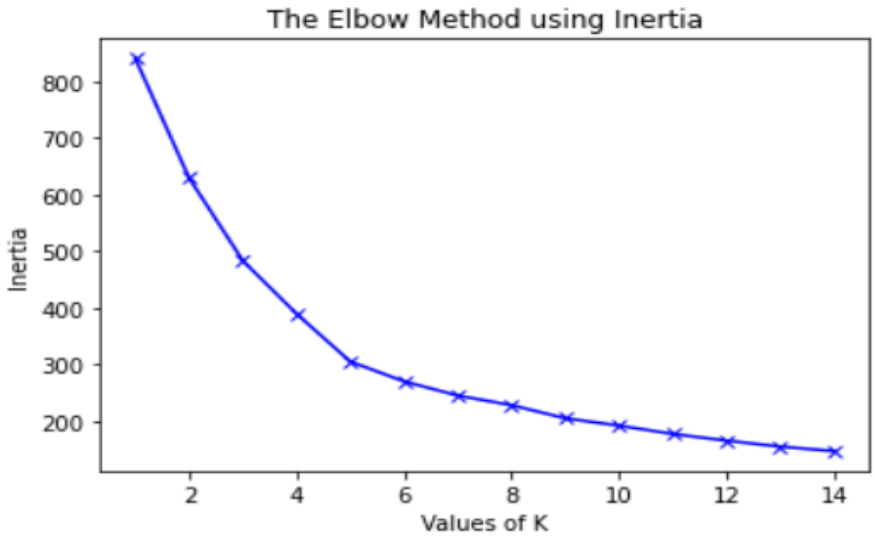
After transforming the data only the part related to income, local employments, debt risk score, home prices, renters and breaks was used to figure out the basic statistical information (mean, standard deviation, quartiles, etc.) for each category – using descriptive model/ function - .describe*()*.

	Debt Risk Score	Home Prices	Local Employment	Average after-tax income of households in 2015 ($)	Renter (Total - Private households by tenure - 25% sample data)	Break & Enters
count	140.000000	1.400000e+02	140.0000	1.400000e+02	140.000000	140.00000
mean	739.157143	5.481934e+05	9409.3500	3.512761e+05	3755.678571	75.25000
std	28.626162	2.676674e+05	19125.3383	2.309379e+05	3008.979559	42.15595
min	661.000000	2.041040e+05	438.0000	1.022590e+05	280.000000	12.00000
25%	720.500000	3.749645e+05	2069.5000	1.953375e+05	1946.250000	44.75000
50%	741.000000	4.912100e+05	4052.5000	2.915495e+05	3232.500000	63.00000
75%	759.000000	5.902160e+05	10127.0000	4.305408e+05	4338.750000	99.00000
max	793.000000	1.849084e+06	185891.0000	1.413132e+06	23930.000000	219.00000

As result, we can determine the mean of each variable, that should be used further, when finding the proper clusters of neighborhoods.

And at the next step, the Elbow method was used to find the optimal k (number of clusters) and the machine learning algorithm k-means clustering was used in order to find the clusters and to choose the best ones for potential investment.

As result of the usage of the Elbow method, the optimal k is near 5, as shown below:



So, the k-means clustering algorithm was run with k=5 and the summary of the data in each cluster looked like:

```
df_segm.groupby('Clus_data').mean()
```

]:

Clus_data	Debt Risk Score	Home Prices	Local Employment	Average after-tax income of households in 2015 ($)	Renter (Total - Private households by tenure - 25% sample data)	Break & Enters
0	757.000000	7.843599e+05	4548.000000	2.859371e+05	2660.625000	54.812500
1	725.987179	4.227271e+05	5771.320513	2.191211e+05	3085.320513	61.910256
2	781.000000	1.478962e+06	8623.750000	1.201593e+06	2212.500000	74.750000
3	745.093750	4.676418e+05	21438.531250	5.014442e+05	6364.218750	113.250000
4	777.600000	1.034422e+06	7385.000000	6.659635e+05	3006.500000	90.600000

After grouping the data by clusters and calculating the mean, the neighborhoods that are good enough for investment can be determined.

By descending home prices, the clusters are: 2,4,0,3,1.
By ascending debt risk score the clusters' order is: 1,3,0,4,2. We can figure out that the neighborhoods with smaller home prices (cluster 1), are same as the neighborhoods with smaller debt risk score and vice versa.

By descending average households' income, the order of the clusters is: 2,4,3,0,1. So in the neighborhoods from clusters 2,3,4 the average households' income is more than the average (3.512761e+05). Also, interesting fact is that the neighborhoods where the income is the least (cluster 1), the home prices and the debt risk are the smallest.
By descending number of renters, the clusters' order is: 3,1,4,0,2. Interesting result, is that the neighborhoods (cluster 2) with the highest prices, match the ones with the least number of renters. So, these neighborhoods are preferred form buyers of their own home (where to live).
By descending local employment jobs, the clusters are in this line: 3,2,4,1,0.
By ascending number of breaks and enters the clusters are in this configuration: 0,1,2,4,3.

In the results, we can see the cluster 0 includes the neighborhoods with the highest local employments and businesses and the income is above the average. The neighborhoods are the safest (B&E = 54.8) and there aren`t a lot of renters. As disadvantages that should be considered by the potential investors, we can determine the higher (of the average) debt risk and the higher (than the average) home price.

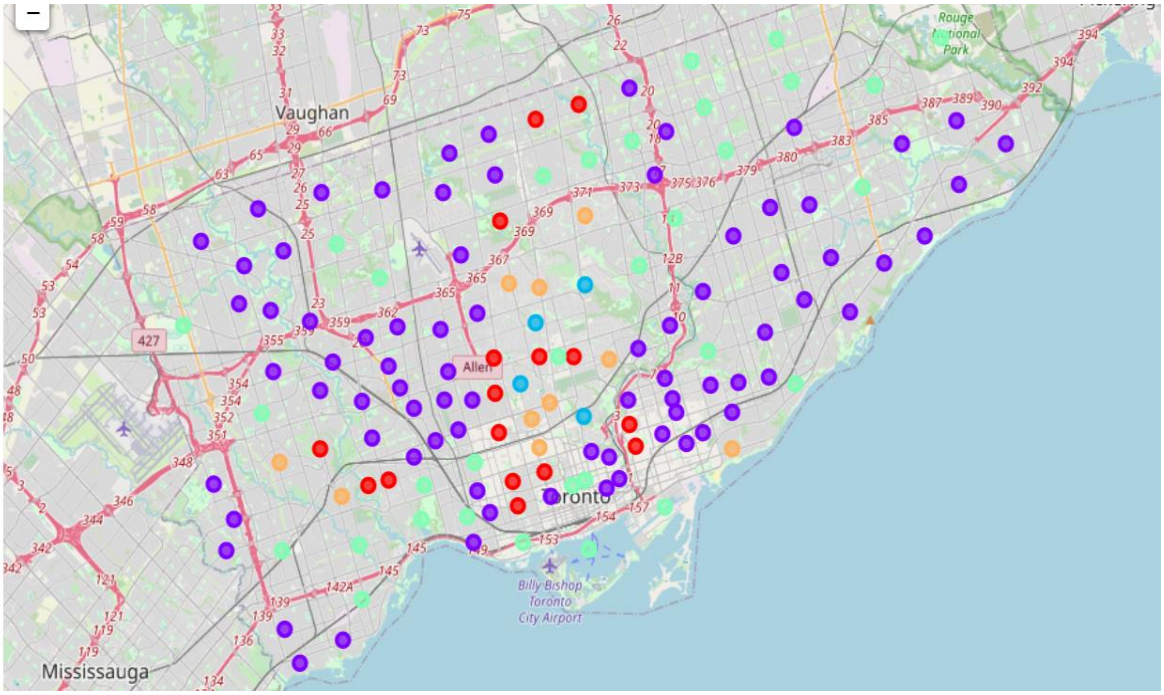
The neighborhoods in cluster 1 have lower (than the average) price, the smallest debt risk and are pretty safe (B&E less than the average). The basic disadvantages are the bigger number of renters (more competition), the lowest level of income and less local employments (and businesses). The neighborhoods form this clusters are mainly the neighborhoods used by renters, because the properties are cheap. The bigger number of renters, mean bigger competition when finding potential tenants. The income of the households is the lowest, so the potential renters could be considered insolvent. The number of local employments jobs and businesses in the neighborhoods of the cluster that is below the average means that the potential renters should travel on a daily base from home to their work places (offices). The basic advantages are the safety of the neighborhoods, the debt risk which is below the average and (possibly) low prices (smaller investment is needed). But, the disadvantages (not enough places where the potential renters to work, the least size of the income and the potential competition when finding the tenants) are more than the advantages and are more important, so the neighborhoods in the cluster (1) would be excluded at this stage.

The neighborhoods in cluster 2 are with the highest income and the least number of renters and local employment (and businesses) above the average (advantages). The disadvantages are the highest prices (and initial investment) and the biggest debt risk. According the break and enters, they are almost on the average.

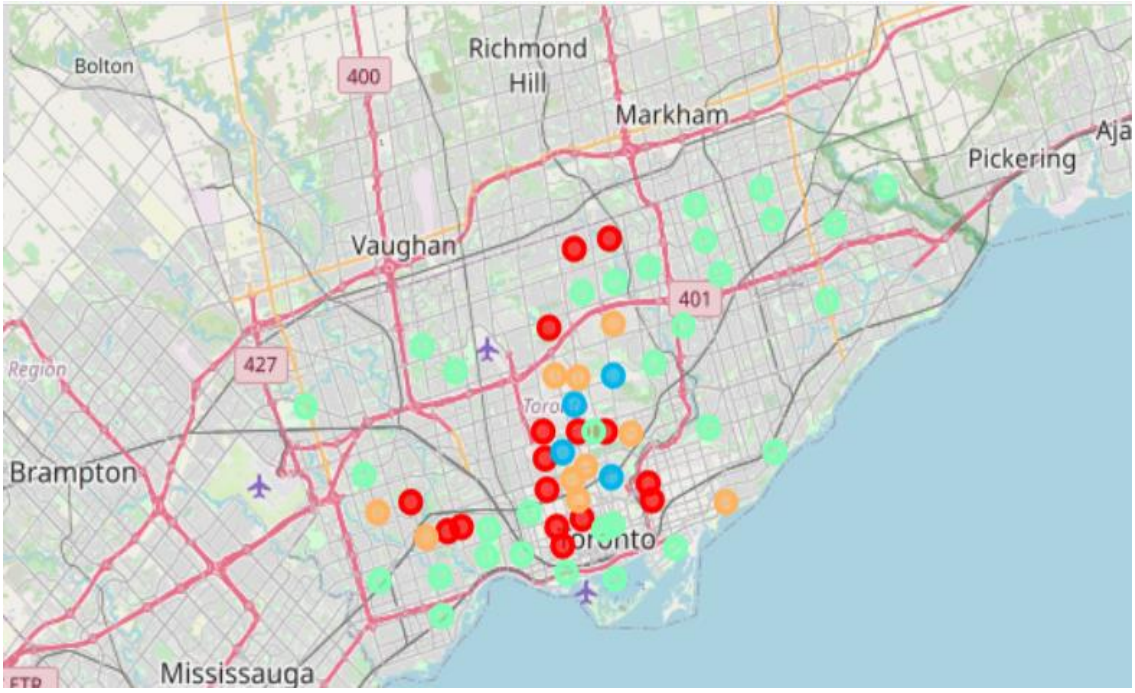
The cluster 3, includes neighborhoods where the number of the local employment jobs is the bigger. Other advantages are the income above the average and the home prices are below the medium price level. These neighborhoods could be interesting for investors, looking for smaller amount of the investment. The disadvantages are: pretty big number of renters (bigger potential competition), debt risk bigger than the average and potential high price of the insurance (due to the biggest number of breaks and enters).

The last cluster (4) includes neighborhoods where the number of renters is a bit less than the average, the debt risk score is low and the average income of the households is above the average level. The main disadvantages are break and enters above the average (95>75), local employments below the average level, and home price higher than the average.

Based on the results, the neighborhoods in cluster 1 have been dropped. And after the visualization the Toronto's map with the clusters has been transformed from:



into this:



2nd stage: Segmenting the neighborhoods from family perspective

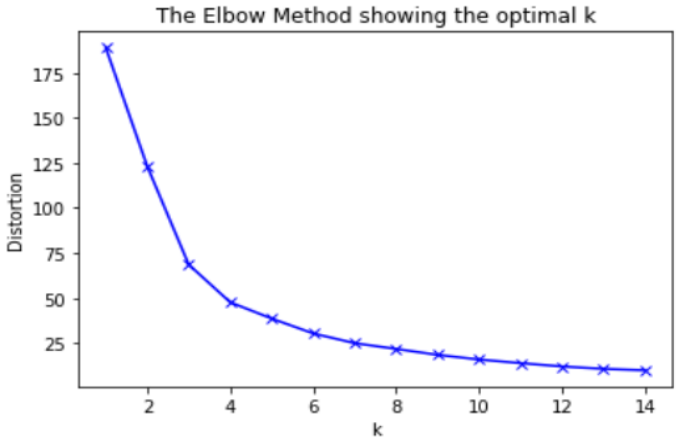
At this stage, the data related to couples with kids, green spaces and child care spaces in each neighborhood has been used. The initial data set was (.head()):

	Neighbourhood	Child Care Spaces	Couples with children	Green Spaces
0	West Humber-Clairville	195	4490	2.078836
1	Willowridge-Martingrove-Richview	165	2905	0.348688
2	Edenbridge-Humber Valley	30	1980	0.544804
3	Princess-Rosethorn	367	1805	0.220795
4	Islington-City Centre West	204	4900	0.510250

Basic statistics were calculated (in order to understand better the data):

	Child Care Spaces	Couples with children	Green Spaces	Clus_data
count	63.000000	63.000000	63.000000	63.000000
mean	154.857143	2696.984127	0.778333	1.492063
std	91.093865	1537.804662	1.830720	1.189646
min	0.000000	450.000000	0.005830	0.000000
25%	81.500000	1530.000000	0.123605	0.000000
50%	162.000000	2325.000000	0.348688	2.000000
75%	213.000000	3410.000000	0.802522	2.000000
max	441.000000	7330.000000	14.271455	3.000000

The next step was normalizing the data and finding the optimal k (using the Elbow method once more).



Finding the clusters (k=4) using the K-mean clustering algorithm and summarizing the data in each cluster:

Clus_data	Child Care Spaces	Couples with children	Green Spaces
0	151.476190	2418.095238	0.764156
1	168.000000	6531.000000	3.414688
2	119.590909	1312.045455	0.257601
3	206.933333	3840.666667	0.683134

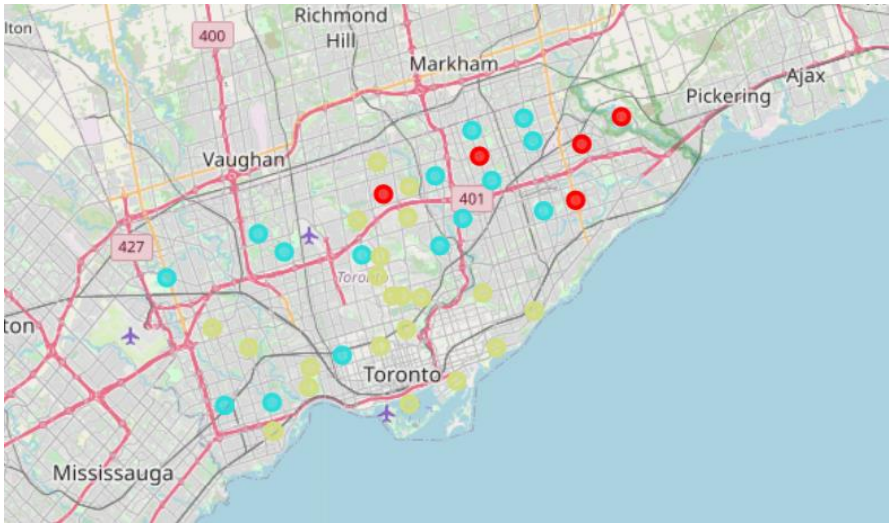
In cluster 0, the number of couples with children (respectively the kids and the young people) and the child care spaces are under the average (2418< 2696 and 151.5 < 154). Only the variable related to the number of the green spaces is above the average.

In cluster 1, the number of the couples with children (kids and young people) and the green spaces are the biggest and the child care spaces are above the average.

In cluster 2 we observe the smallest number of child care spaces, of couples with kids and the least green space. So, the neighborhoods belonging to the cluster, would be removed as not preferable of the potential tenants.

Cluster 3, represents the places with the biggest number of the child care spaces and the green spaces and the couple with children are above the average.

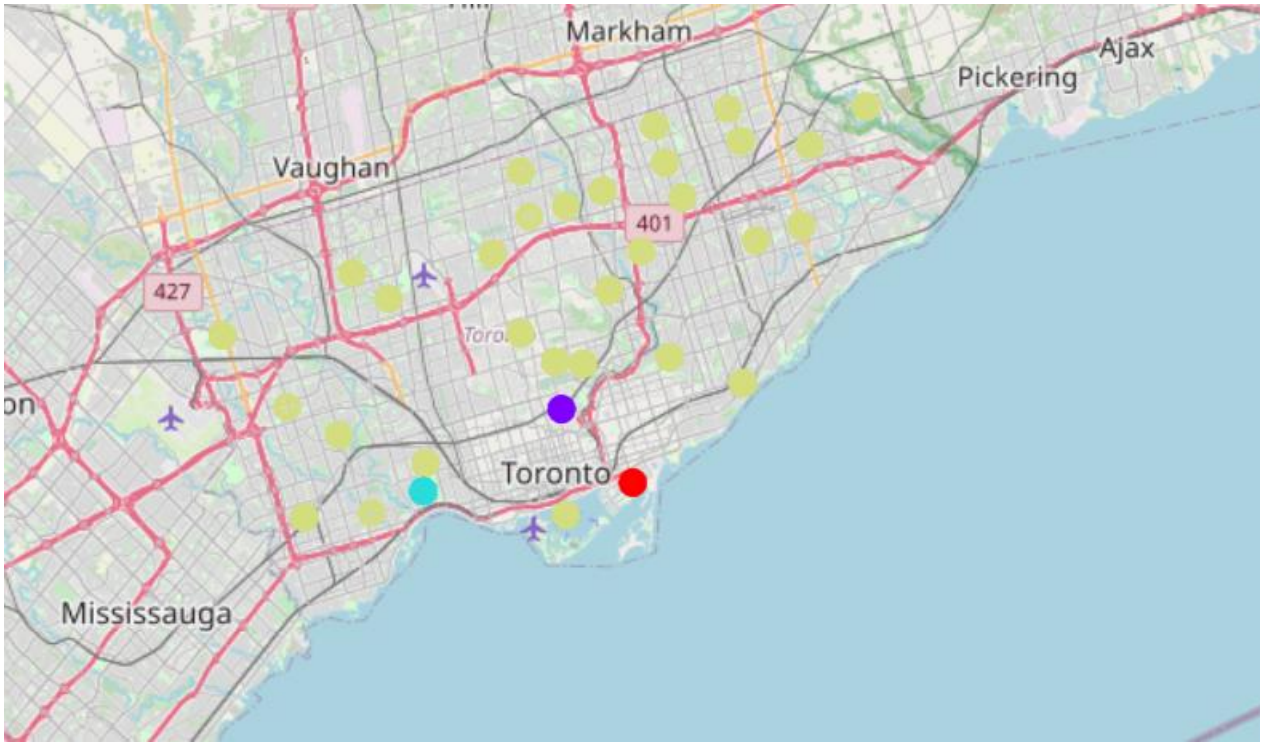
Visualizing the clusters considered good enough from family perspective (clusters 0,1,3), using folium, leaded to:



3rd stage – exploring the reduced number of neighborhoods using Foursquare

At this stage, the Toronto`s neighborhoods have been clustered based on the Foursquare data. The neighborhoods with less pubs and bars and more gardens, garden centers and playgrounds have been found in order to finalize this research.

As first step, I have set my Foursquare credentials. After that the top 100 venues in radius of 500 meters in the area of L'Amoreaux have been found. The algorithm has been repeated for the rest of the neighborhoods. After converting the json data into pandas dataframe, finding the unique categories and calculating some basic statistics, the neighborhoods with Bars and Pubs have been excluded form the list of the potential neighborhoods. The final result after applying the K-means clustering algorithm (k=4, here we didn`t use the Elbow method, because the number of the clusters was obvious – only 3 different neighborhoods found) included 40 neighborhoods and only 3 of them had garden, garden center and playground (respectively one in each one of the neighborhoods) according the Foursquare data. The visualization of the very last clusters looked like:

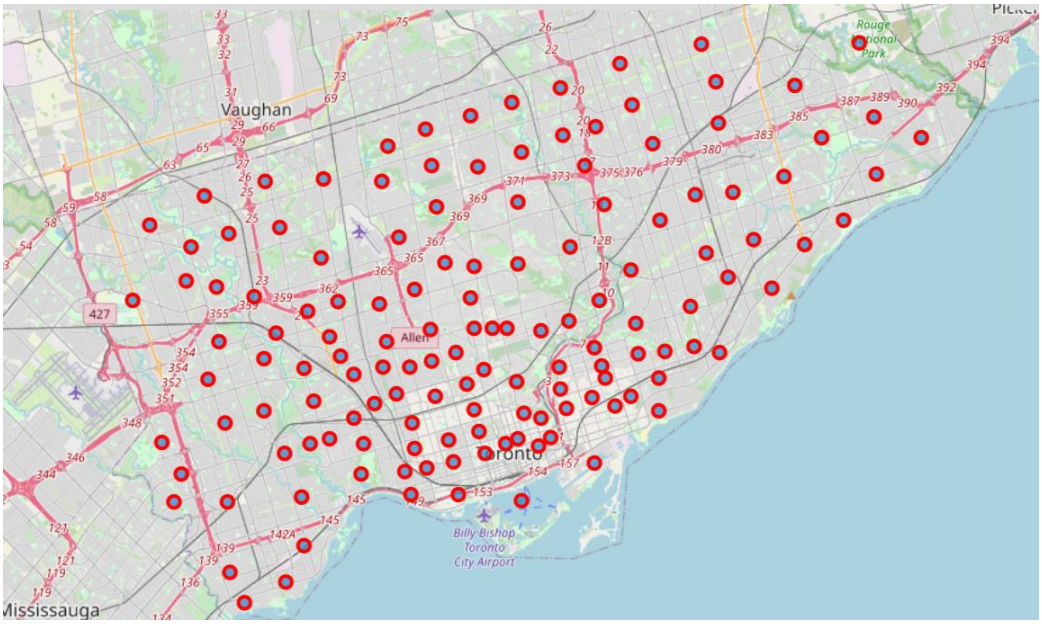


Discussion section

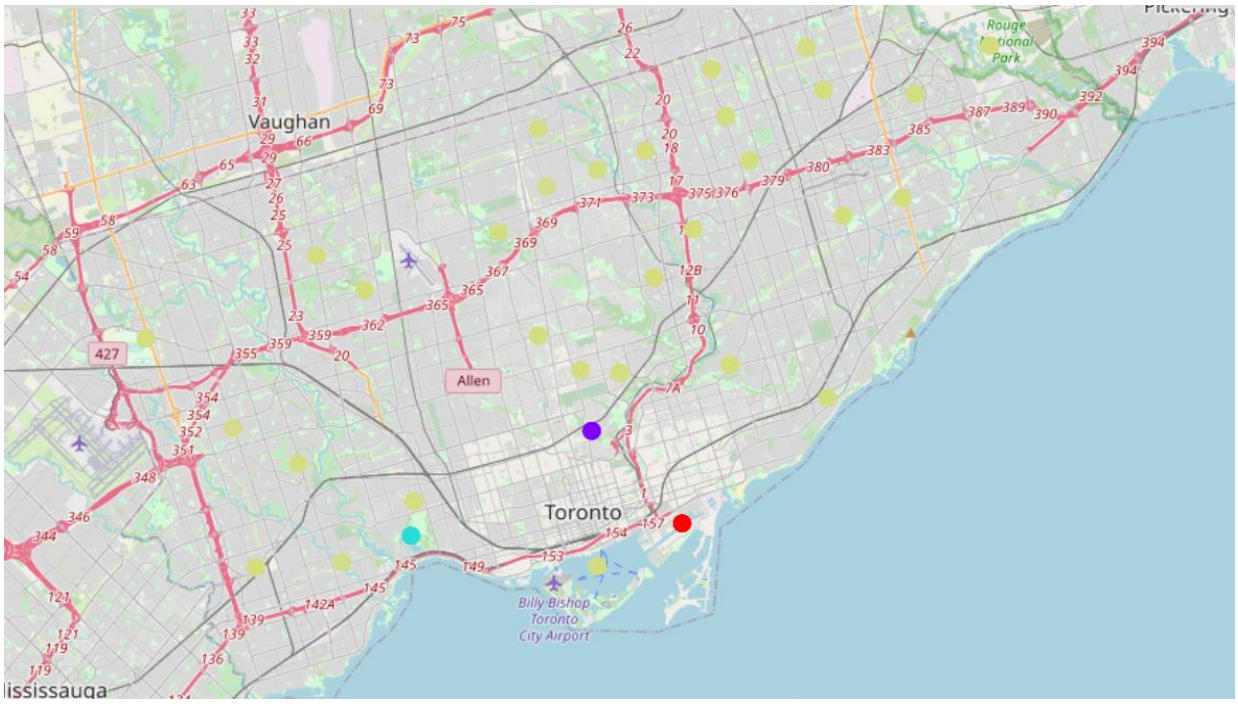
The basic goal in the first stage was to find the neighborhoods good for investments in residential property based on debt risk score, income of the households, local employments, number of renters, average home prices and breaks and enters. Based on the data (which is not that actual) the neighborhoods from cluster 1 has been excluded and the number of the results has been reduced from 140 to 62. The neighborhoods excluded (on average base) because the income of the potential renters was the smallest, but the number of the renters was quite above the average and the potential competition when finding tenants would be quite strong. Also, there weren`t enough businesses and local employments jobs which meant that the future renters would travel to their offices on a daily base. Basically, the neighborhoods in this cluster were the neighborhoods where the mass of the properties were used for renting. When investing the logical thing is to find a location with less competition and to thing out of the box, the neighborhoods from cluster 1 were excluded.

On the second stage, the data related to the 62 neighborhoods was upgraded by variables considered important from family perspective. The data related to number of couples with children, green spaces and child care spaces in the neighborhoods was added. And the result of the k-mean clustering algorithm, helped me reduce even more the number of the potential neighborhoods where the potential investors could put their money (with less risk because they would find potential stable tenants in long terms easier). The cluster 2, which contained neighborhoods without enough green spaces, child care spaces, and the least couples with kids, was excluded as not showing satisfactory results. And the number of the neighborhoods was reduced once more with 21 neighborhoods.

At the final stage the data form Foursquare, was used to help the investors when piking the right place for investment in residential property for rent. The neighborhoods with bars and pubs, were excluded and the places with playgrounds, gardens and garden centers were used for the last cluster. And after all the reductions, the initial map of the neighborhoods in Toronto



was converted to this map showing the best neighborhoods for investment in residential property to rent:



Conclusion section

In conclusion, at the 1st stage the 140 neighborhoods in Toronto has been reduced initially with 78 neighborhoods which has been considered worst place for investment in residential property to rent, due to the bigger competition, the lowest average income of the future tenants and the lack of businesses and local employment jobs.

At the 2nd stage, the neighborhoods were reduced once more, excluding the ones without satisfactory amount of green spaces, kids and young people, couple with children and child care spaces. The number of the neighborhoods, considered good from family perspective was 42.

At the 3rd stage, from these neighborhoods 7 more were dropped, due to having bars and pubs according the Foursquare database.

The final list of neighborhoods where the investors can put their money less risky contained 33 neighborhoods (Agincourt North, Banbury-Don Mills, Bayview Village, Bendale, Birchcliffe-Cliffside, Don Valley Village, Downsview-Roding-CFB , Edenbridge-Humber Valley, Glenfield-Jane Heights, High Park North, High Park-Swansea, Islington-City Centre West, L'Amoreaux, Lansing-Westgate, Lawrence Park South, Leaside-Bennington, Malvern, Millike, Mount Pleasant East, Newtonbrook East, O'Connor-Parkview, Parkwoods-Donalda, Rosedale-Moore Park, Rouge, South Riverdale, Steeles, Stonegate-Queensway, Tam O'Shanter-Sullivan, Waterfront Communities-The Island, West Humber-Clairville, Willowdale East, Willowridge-Martingrove-Richview, Woburn).