Finding an optimal location to open a Coffee Shop in Toronto.

This report is part of the capstone project for the IBM Data Science Professional Certificate course. From defining a business problem, finding the appropriate data, prepare and analyze them to get some useful insight necessary to provide an answer to the business problem, to finally prepare the present report and a presentation, we are asked for this capstone project to walk through all of the main steps of a data science project.

Table of Contents

Problem Description

We will consider the problem of finding an optimal location for a new Coffee Shop in the city of Toronto.

We will need to find an appropriate location for the owner of a future new Coffee Shop. Finding a suitable location is of paramount importance, as it is one of the most important element to ensure the economic viability of the project. This is even more important in a big city like Toronto, where according to the chosen area there could be a big demand, but also a lot of competition for costumers.

Upon discussion with the owner, we identified the characteristics of what could be an interesting location for his project. Such a location would be in a neighborhood with a **high number of potential costumers**, and with **low competition**, meaning a number of already implanted Coffee Shops as low as possible.

Most of the work will be devoted to find a way to identify neighborhoods with a high number of potential costumers. This will be done using the locations of **other** venues.
We will start from the assumption that already installed and successful Coffee Shops are in interesting locations, and that such locations can be characterized in terms of the kind of venues present in the vicinity. Therefore, we will start by identifying the characteristics of the Neighborhoods where successful Coffee Shops are implanted. Then, we will use this information to locate other similar areas, *independently of the presence* of already installed Coffee Shops.
Moreover, we are asked to asses whether the economic level of the neighborhood has an influence on the number of potential costumers. If this was not the case, a **rich neighborhood** would be preferred by the owner, as he wishes to open a high-end Coffee Shop, which in his opinion would be more profitable.

Based on these findings, we will create an evaluation algorithm applicable to any neighborhood, which takes into account the characteristics of the neighborhood in terms of venues present as well as potentially the average income, but penalizes the presence of other Coffee Shops. Such evaluation algorithm will ultimately provide the owner with an ordered list of the most suitable areas where to open his new Coffee Shop.

Data

Introduction : Data description

We will need different geographical and socio-economic data for the city of Toronto, as well as data on the locations of installed venues.

The list of the Neighborhoods of Toronto can be found on the **Wikipedia** web page:
http://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M .
Instead of Neighborhoods, we will rather use **Postal Codes Areas**, which produces a slightly coarser grid, but is necessary for compatibility reasons with the dataset on socio-economic data. The geographical coordinates of each Neighborhood or Postal Code Area will be obtained using *Google Maps API geocoding*.

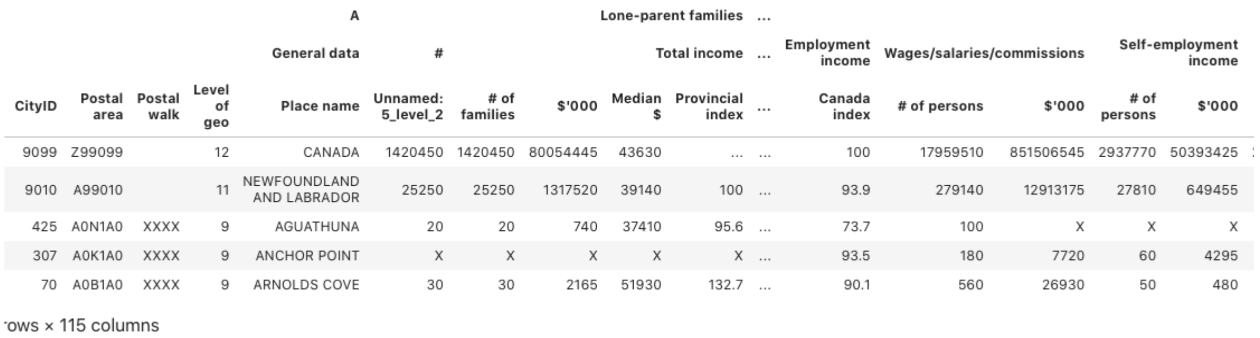
We will also collect data on the average income in each Postal Code Area from the **official page of the Canadian Government**, at:
https://open.canada.ca/en .
These data will have to be cleaned and prepared, then merged with the geographical data described above.

Finally, we will get data on venues present in or near each Neighborhood using **Foursquare's database**.

Retrieve and clean income data

We will start by retrieving the income data from the Canadian Government web page. The latest data are from 2016. However, as we will only use it to compare the relative average income between Postal Areas, we can reasonably assume that in the area of interest (probably located near the center of Toronto), changes have not been so important as to radically change the picture.

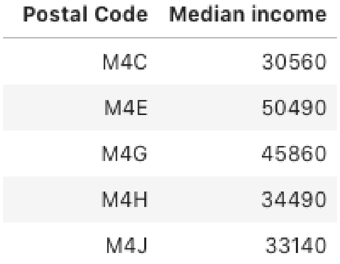
Unfortunately, the full database contains fat too much information for our needs, which makes it hard to read:



				A					Lone-parent families	...					
				General data	#				Total income	...	Employment income		Wages/salaries/commissions		Self-employment income
CityID	Postal area	Postal walk	Level of geo	Place name	Unnamed: 5_level_2	# of families	$'000	Median $	Provincial index	...	Canada index	# of persons	$'000	# of persons	$'000
9099	Z99099		12	CANADA	1420450	1420450	80054445	43630	...	...	100	17959510	851506545	2937770	50393425
9010	A99010		11	NEWFOUNDLAND AND LABRADOR	25250	25250	1317520	39140	100	...	93.9	279140	12913175	27810	649455
425	A0N1A0	XXXX	9	AGUATHUNA	20	20	740	37410	95.6	...	73.7	100	X	X	X
307	A0K1A0	XXXX	9	ANCHOR POINT	X	X	X	X	X	...	93.5	180	7720	60	4295
70	A0B1A0	XXXX	9	ARNOLDS COVE	30	30	2165	51930	132.7	...	90.1	560	26930	50	480

ows × 115 columns

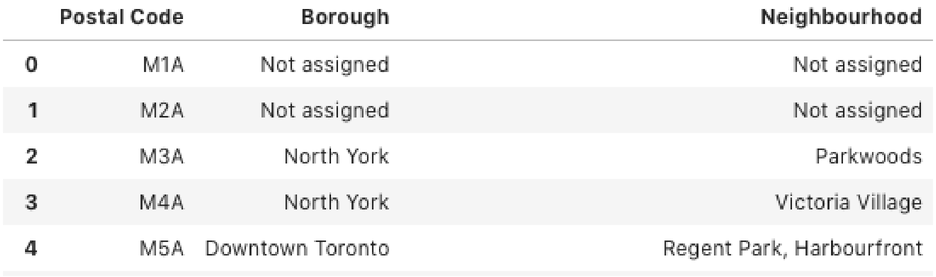
After downloading, loading it into a DataFrame, and analyzing it, we realize that we need to perform some cleaning operations: we select the column concerning the "*Total income*" per person (irrespectively to their family status), drop headers and two column levels, and slice the data to select only the data concerning the neighborhoods of Toronto (excluding the general data for the whole city and region). This is a snapshot of the cleaned data:



Postal Code	Median income
M4C	30560
M4E	50490
M4G	45860
M4H	34490
M4J	33140

Retrieve and clean geographical data

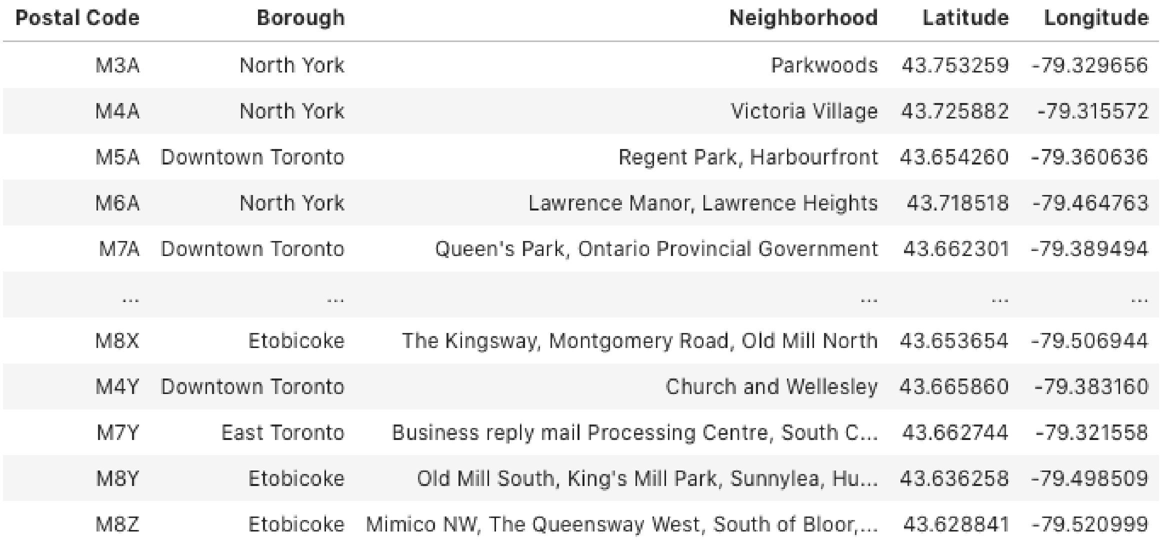
We use the Python method *read_html* together with the *BeautifulSoup* library to retrieve the list of Postal Areas and Neighborhood names for the city of Toronto from the Wikipedia web page:



	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

We rename the columns in a more meaningful way and clean the DataFrame dropping the rows with "*Not assigned*" Borough, then perform a **Data integrity check.** We verify that we have 103 different Postal Codes, and 103 rows in the DataFrame, meaning that no Postal Code is listed twice. We also check that there are no "*Not assigned*" Neighborhoods left in the DataFrame.

Then we use *Google Maps geocoding API* to retrieve the geographical coordinates of the center of each Postal Area, and merge them with the data obtained from the Wikipedia web page. Another data integrity check ensures that each postal code area has now its own coordinates. We obtain the following cleaned data set:

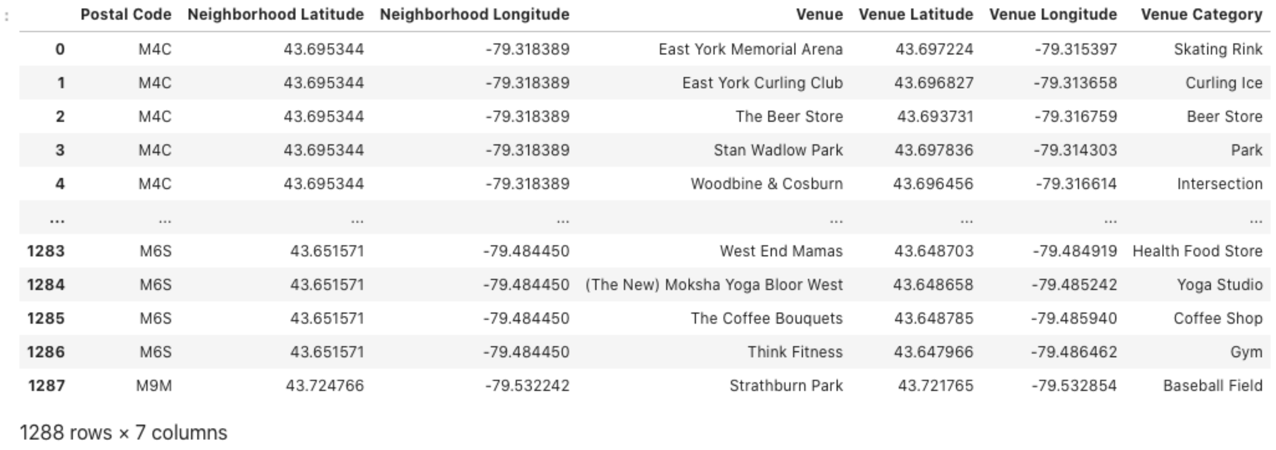


Postal Code	Borough	Neighborhood	Latitude	Longitude
M3A	North York	Parkwoods	43.753259	-79.329656
M4A	North York	Victoria Village	43.725882	-79.315572
M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...
M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

Finally, the income and geographical data sets are merged.

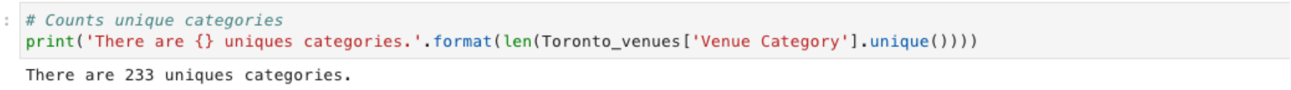
Retrieve and clean venue data

Using Foursquare's API we retrieve the most popular venues in or near every Neighborhood. First, we prepare the credentials for the API calls, then we define a function that retrieves the venues located within 500m of the center of each given Postal Code Area, and finally apply it to all of the Postal Code Areas in Toronto. This is a snapshot of what we obtain: a large dataset with 1288 venues belonging to 233 different categories



	Postal Code	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M4C	43.695344	-79.318389	East York Memorial Arena	43.697224	-79.315397	Skating Rink
1	M4C	43.695344	-79.318389	East York Curling Club	43.696827	-79.313658	Curling Ice
2	M4C	43.695344	-79.318389	The Beer Store	43.693731	-79.316759	Beer Store
3	M4C	43.695344	-79.318389	Stan Wadlow Park	43.697836	-79.314303	Park
4	M4C	43.695344	-79.318389	Woodbine & Cosburn	43.696456	-79.316614	Intersection
...	...	...	...	...	...	...	...
1283	M6S	43.651571	-79.484450	West End Mamas	43.648703	-79.484919	Health Food Store
1284	M6S	43.651571	-79.484450	(The New) Moksha Yoga Bloor West	43.648658	-79.485242	Yoga Studio
1285	M6S	43.651571	-79.484450	The Coffee Bouquets	43.648785	-79.485940	Coffee Shop
1286	M6S	43.651571	-79.484450	Think Fitness	43.647966	-79.486462	Gym
1287	M9M	43.724766	-79.532242	Strathburn Park	43.721765	-79.532854	Baseball Field

1288 rows × 7 columns



```
# Counts unique categories
print('There are {} uniques categories.'.format(len(Toronto_venues['Venue Category'].unique())))
```

```
There are 233 uniques categories.
```

Methodology and Data Analysis

In this section we are going to produce a model which is going to classify the neighborhoods of the city of Toronto according to the suitability of the area to install a new Coffee Shop. This will be done taking into account different elements.

We assume that the Coffee Shops already present and successful in the city are placed in good locations. Therefore, we start by trying to understand how are the neighborhoods in which Coffee shops are already present. This will not only be done in a qualitative way (which venues are present, economic level of the area, etc.), but in a **quantitative** way.
This is done using **correlations**. For example, finding a high correlation between the number of Hotels and Coffee Shops in the same area will indicate that hotels could be used to pin-point good locations, while a very low correlation between the number of parks and Coffee Shops indicates that parks are not useful to locate interesting areas.

The main question we ask here is: how much each aspect (presence, kind and number of venues, as well as economic level of residents) influence in a positive or negative way the environment (from the point of view of a possible implantation of a new Coffee Shop)?
This will allow us to rate every area, giving it a note on how suitable it is to implant a new Coffee Shop.

We will need to perform different steps.

First, concerning the venues. The classification used in Foursquare is not very useful for us, as many categories count only very few venues, which would not be enough to produce significant statistics. However, many different categories actually group very similar kinds of venues. Therefore, we can re-classify the different venues in more meaningful groups, each counting a higher venues. Those new groups may become indicators of the presence of high numbers of

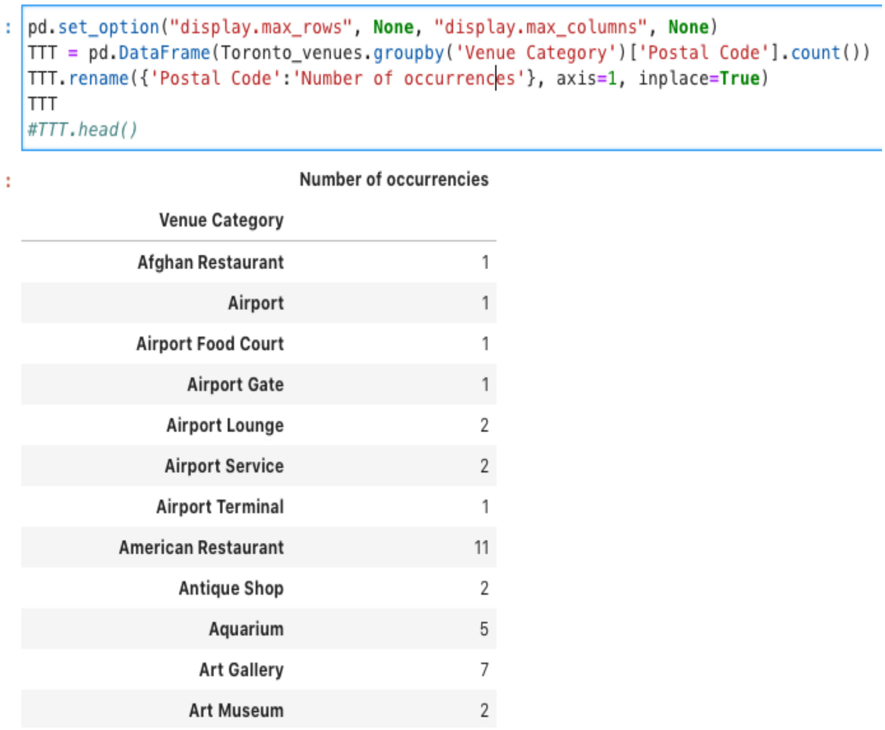
potential costumers! To assert if this is the case, we look for correlations between the number of venues in those groups and the number of already implanted Coffee Shops.
Correlations will be searched also between the total number of venues, the mean income and the number of Coffee Shops.

Then, elements that will show a high correlation with the number of Coffee Shops will be retained to construct the **rating algorithm**. For each area, we will multiply the correlation coefficient of every element with the relative frequency of that element in the area, to obtain the contribution of the specific element to the rate of the area. For example: if Hotels have a high correlation, and in the area in question there are a lot of them, the area will have a higher grade. However, a high number of already installed Coffee Shops will be considered a negative element (competition). We will proceed in a similar way to include in our analysis also the total number of venues and the mean income.

Finally, the algorithm will rate every area based on the characteristics that can indicate or influence an easy and fruitful installation of a new Coffee Shop, and produce an ordered list of the most suitable areas (each with its respective grade) for the owner to chose from.

Data Preparation

First, we inspect the different venue's categories, as reported in the Foursquare database. We list all of them with their respective total number of occurrences within the city of Toronto, and analyze the resulting list. Here's a snapshot of this very long list:



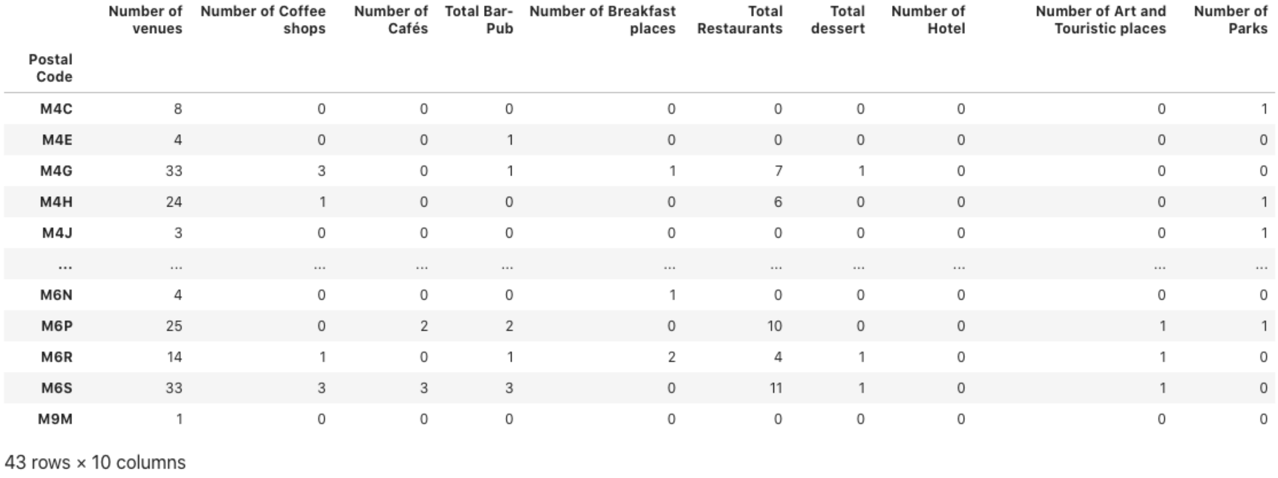
```
pd.set_option("display.max_rows", None, "display.max_columns", None)
TTT = pd.DataFrame(Toronto_venues.groupby('Venue Category')['Postal Code'].count())
TTT.rename({'Postal Code':'Number of occurrences'}, axis=1, inplace=True)
TTT
#TTT.head()
```

	Number of occurrencies
Venue Category	
Afghan Restaurant	1
Airport	1
Airport Food Court	1
Airport Gate	1
Airport Lounge	2
Airport Service	2
Airport Terminal	1
American Restaurant	11
Antique Shop	2
Aquarium	5
Art Gallery	7
Art Museum	2

We produce the following new groups, containing venues in the different Foursquare's categories, as in the following table:

New categories	Forusquare's original categories
Coffee Shop	Coffee Shop
Café	Café
Breakfast places	Breakfast Spot
Eating places	All kinds of “Restaurant”, Bistro, Creperie, Dinner P lace, Steakhouse, Noodle House, all kind of “Joint” food places, Bagel Shops, Fish & Chips.
Dessert places	Dessert Shop, Ice Cream Shop, Frozen Yogurt Shop
Bar or Bup	Bar, Beer Bar, Cocktail Bar, Gay Bar, Hookah Bar, Hotel Bar, Juice Bar, Sake Bar, Sports Bar, Wine Bar, Irish Pub, Pub
Hotel	Hotel
Park	Park, Playground
Art and Touristic Places	Aquarium, Event Space, Fountain, Historic Site, Indie Movie Theater, Jazz Club, Monument, Movie Theater, Scenic Lookout, Sculpture Garden, Theater, all kinds of “Museums”, all kinds of Art or performing “Venues” or “Spaces”.

We obtain the following dataset:



	Number of venues	Number of Coffee shops	Number of Cafés	Total Bar-Pub	Number of Breakfast places	Total Restaurants	Total dessert	Number of Hotel	Number of Art and Touristic places	Number of Parks
Postal Code										
M4C	8	0	0	0	0	0	0	0	0	1
M4E	4	0	0	1	0	0	0	0	0	0
M4G	33	3	0	1	1	7	1	0	0	0
M4H	24	1	0	0	0	6	0	0	0	1
M4J	3	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...
M6N	4	0	0	0	1	0	0	0	0	0
M6P	25	0	2	2	0	10	0	0	1	1
M6R	14	1	0	1	2	4	1	0	1	0
M6S	33	3	3	3	0	11	1	0	1	0
M9M	1	0	0	0	0	0	0	0	0	0

43 rows × 10 columns

Now we have to compute in each neighborhood the percentage of the contribution of the different categories of venues to the total number of venues. To do so, we divide each row of the above matrix (number of each kind of venue in each neighborhood) by the total number of venues in that neighborhood. We shall use here percentages rather than the actual number of venues, as the total number of venues will be taken into account in the algorithm as a different element.

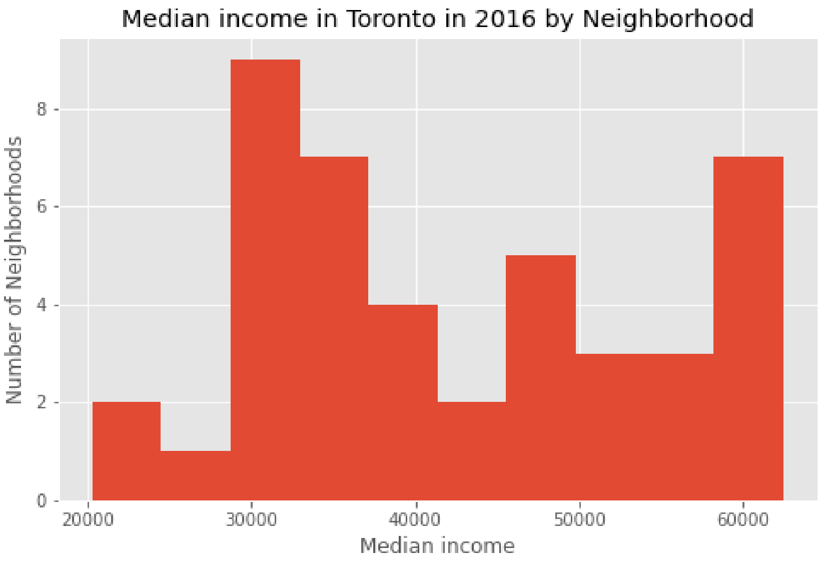
Then we proceed to count the total number of venues in each area.

Finally, we prepare for each area a list of the 7 most common venues. This is not used in the algorithm, but will be useful for the presentation of the result to the stakeholder, as it can help him get a better grasp of the kind of neighborhoods we are considering and suggesting.

Data Analysis: Income and Venues

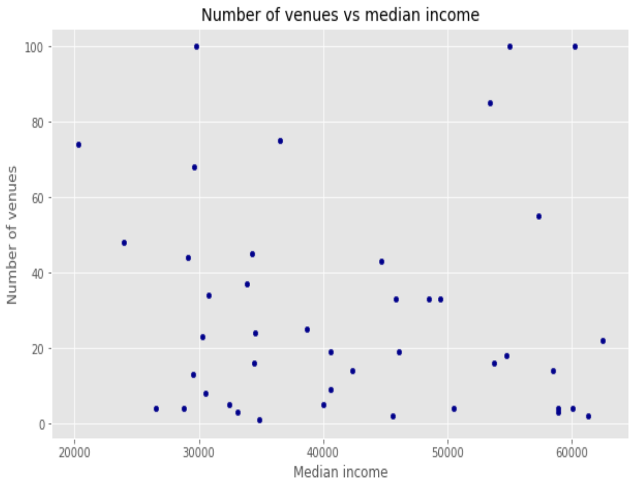
Now that the data are prepared and cleaned, we can start to analyze them to look for interesting correlations.

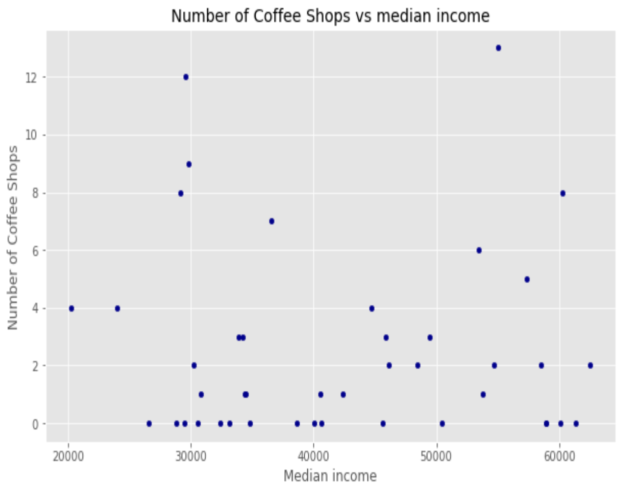
We start by looking a the income data, showing the distribution of the median income by area.

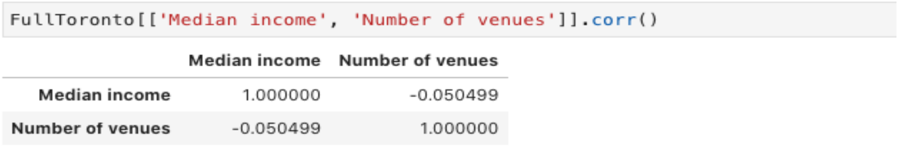


Most areas register an average income between 30.000 and 40.000, but there are quite a few showing a higher average income, between 50.000 and 60.000.

Plotting the average income against the number of venues (each point represents a neighborhood) there does not seem to be a clear relation between income and number of venues. Computing explicitly the correlation, we check that the two values are uncorrelated. The same holds for the median income and the number of Coffee Shops.

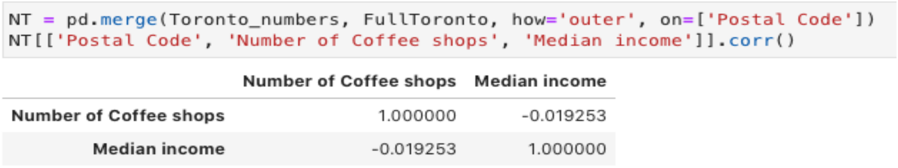






```
FullToronto[['Median income', 'Number of venues']].corr()
```

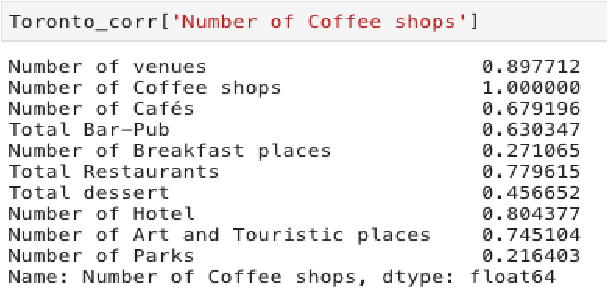
	Median income	Number of venues
Median income	1.000000	-0.050499
Number of venues	-0.050499	1.000000



```
NT = pd.merge(Toronto_numbers, FullToronto, how='outer', on=['Postal Code'])
NT[['Postal Code', 'Number of Coffee shops', 'Median income']].corr()
```

	Number of Coffee shops	Median income
Number of Coffee shops	1.000000	-0.019253
Median income	-0.019253	1.000000

Next, we check the correlations between the categories of venues we created. In particular, we are interested in the correlation between the number of Coffee Shops and the other variables.



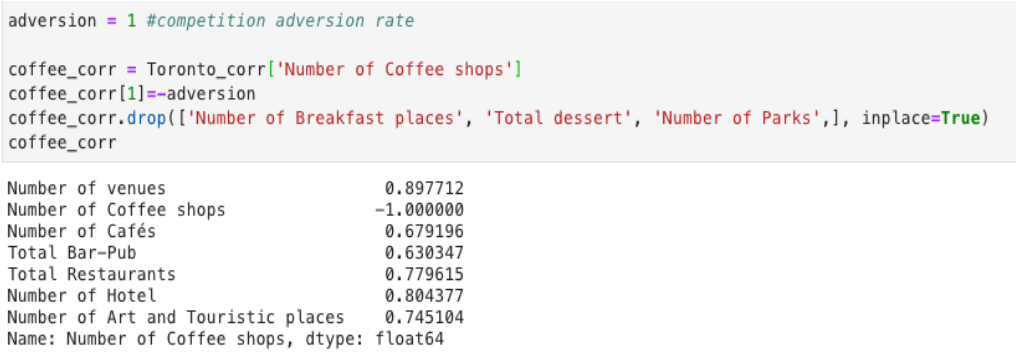
```
Toronto_corr['Number of Coffee shops']
```

```
Number of venues                      0.897712
Number of Coffee shops                1.000000
Number of Cafés                       0.679196
Total Bar-Pub                         0.630347
Number of Breakfast places            0.271065
Total Restaurants                     0.779615
Total dessert                         0.456652
Number of Hotel                       0.804377
Number of Art and Touristic places    0.745104
Number of Parks                       0.216403
Name: Number of Coffee shops, dtype: float64
```

Neighborhood ratings

We are now in possession of all of the ingredients needed to produce a grade for every area.

Looking at the correlations between the number of Coffee Shops and other venues, we see that the number of Coffee shops has a relevant (more than 0.6) positive correlation with the *total number of venues*, the *number of Cafés*, *Bars/Pubs* and especially *Restaurants*, *Hotels* and *Art-related attractions*. We can infer that a high number of those venues in a Neighborhood should indicate an area where there could be a demand for Coffee shops and where our venue could find an interesting flow of costumers.



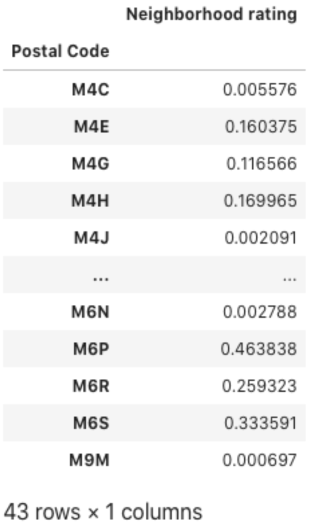
```
adversion = 1 #competition adversion rate

coffee_corr = Toronto_corr['Number of Coffee shops']
coffee_corr[1]=-adversion
coffee_corr.drop(['Number of Breakfast places', 'Total dessert', 'Number of Parks',], inplace=True)
coffee_corr
```

```
Number of venues                      0.897712
Number of Coffee shops               -1.000000
Number of Cafés                       0.679196
Total Bar-Pub                         0.630347
Total Restaurants                     0.779615
Number of Hotel                       0.804377
Number of Art and Touristic places    0.745104
Name: Number of Coffee shops, dtype: float64
```

However, the most suitable Neighborhood would also have a low number of already implanted Coffee shops. This is modeled introducing the **competition aversion index**, which can be tuned to fit the owner aversion to a strong competition in the proximity of his new venue. The higher the index value, the more fearful of competition the owner is. A value of zero means indifference to the presence of other Coffee Shops, and a negative value would mean that the owner actually considers beneficial the presence of other Coffee Shops in the vicinity. This index will substitute the correlation index for Coffee Shops.

Based on the above discussion, we will produce ratings for the neighborhoods of Toronto, based on the presence of venues indicating an area suitable for the installation of a Coffee shop and the number of Coffee shops already installed. We shall weight the relative number of venues in every Neighborhood and the relative total number of venues with the correlations found, using a negative value, the *competition aversion index* we introduced above, instead of 1 for Coffee shops.
Here, the aversion index was set to 1.

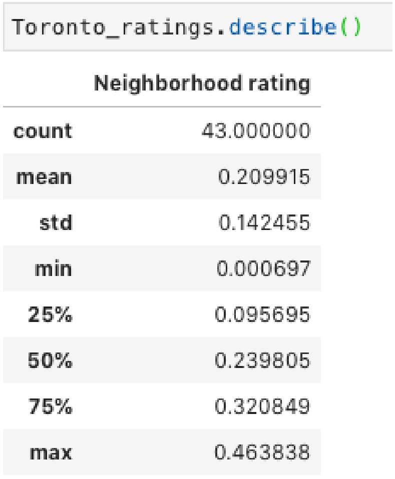
Here is a snapshot of the list of grades provided by our method for every postal area in Toronto.



	Neighborhood rating
Postal Code	
M4C	0.005576
M4E	0.160375
M4G	0.116566
M4H	0.169965
M4J	0.002091
...	...
M6N	0.002788
M6P	0.463838
M6R	0.259323
M6S	0.333591
M9M	0.000697

43 rows × 1 columns

We can check their basic statistical characteristics:



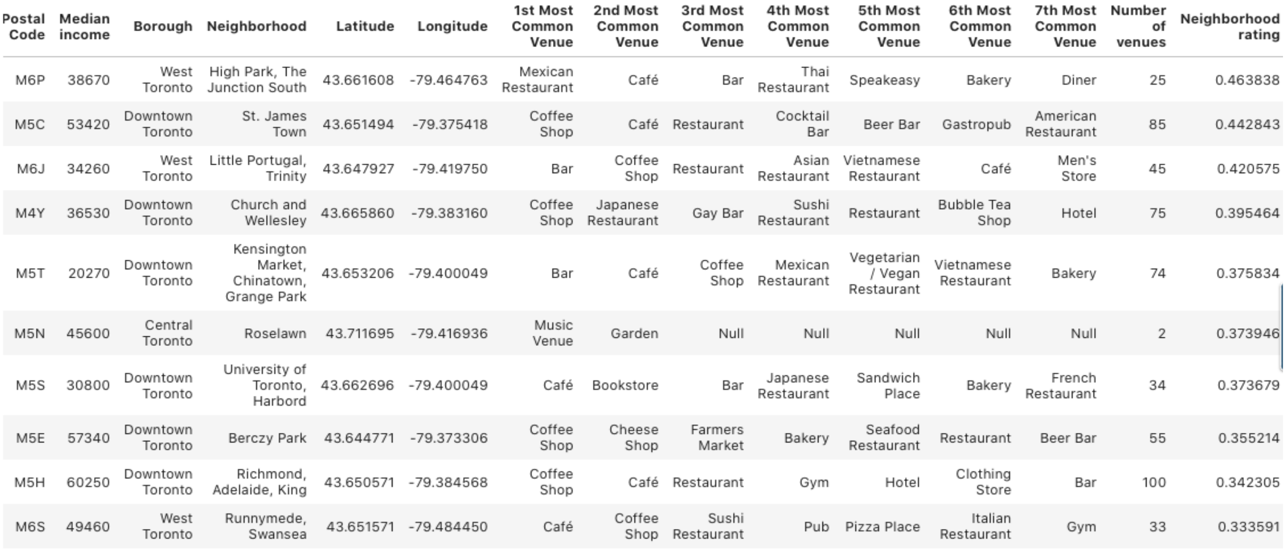
```
Toronto_ratings.describe()
```

	Neighborhood rating
count	43.000000
mean	0.209915
std	0.142455
min	0.000697
25%	0.095695
50%	0.239805
75%	0.320849
max	0.463838

Results

Our analysis shows that the median income of the area most likely will not influence the flow of costumers. Therefore, we did not include this parameter into our model. However, even if the median income of an area does not influence its score, we added a column into the table below so that the owner can take this element into account when making his decision.

We have produced a way to sort the neighborhoods in Toronto according to various criteria, and which can be tuned to take into account the level of aversion to competition of the owner. To ease the owner's choice, we present the results together with the list, for every area, of the most common venues, the average income, total number of venues and the names of the neighborhoods in the Postal Area.



Postal Code	Median income	Borough	Neighborhood	Latitude	Longitude	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	Number of venues	Neighborhood rating
M6P	38670	West Toronto	High Park, The Junction South	43.661608	-79.464763	Mexican Restaurant	Café	Bar	Thai Restaurant	Speakeasy	Bakery	Diner	25	0.463838
M5C	53420	Downtown Toronto	St. James Town	43.651494	-79.375418	Coffee Shop	Café	Restaurant	Cocktail Bar	Beer Bar	Gastropub	American Restaurant	85	0.442843
M6J	34260	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	Bar	Coffee Shop	Restaurant	Asian Restaurant	Vietnamese Restaurant	Café	Men's Store	45	0.420575
M4Y	36530	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	Coffee Shop	Japanese Restaurant	Gay Bar	Sushi Restaurant	Restaurant	Bubble Tea Shop	Hotel	75	0.395464
M5T	20270	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	Bar	Café	Coffee Shop	Mexican Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Bakery	74	0.375834
M5N	45600	Central Toronto	Roselawn	43.711695	-79.416936	Music Venue	Garden	Null	Null	Null	Null	Null	2	0.373946
M5S	30800	Downtown Toronto	University of Toronto, Harbord	43.662696	-79.400049	Café	Bookstore	Bar	Japanese Restaurant	Sandwich Place	Bakery	French Restaurant	34	0.373679
M5E	57340	Downtown Toronto	Berczy Park	43.644771	-79.373306	Coffee Shop	Cheese Shop	Farmers Market	Bakery	Seafood Restaurant	Restaurant	Beer Bar	55	0.355214
M5H	60250	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	Coffee Shop	Café	Restaurant	Gym	Hotel	Clothing Store	Bar	100	0.342305
M6S	49460	West Toronto	Runnymede, Swansea	43.651571	-79.484450	Café	Coffee Shop	Sushi Restaurant	Pub	Pizza Place	Italian Restaurant	Gym	33	0.333591

Discussion

We considered an average reasonable competition aversion index of 1. The ratings vary from virtually 0 to almost 0.5, with a quarter ratings of more than 0.32.

In view of this results, we can select 3 interesting candidate areas where the owner could try to set his new business: one in Downtown Toronto (*St James Town*), and interestingly two in West Toronto (*High Park* and *Little Portugal - Trinity*).

The area in Downtown Toronto is certainly more active (as much as 85 venues present in Foursquare's database) and is a wealthier area, which is a positive point for the owner. However, he should expect a more fierce competition there, as he can see that Coffee Shops are already the most common venue in the area.

On the other hand, the two spots in West Toronto are reasonably active, and even though less wealthy, they have fewer Coffee Shops already implanted.

Conclusion

Leveraging on the use of a model based on the study of correlations between different kinds of venues, this report provides a comprehensive approach to choosing the best spot where to open a new Coffee Shop in Toronto, taking into account socio-economic elements, the presence of different venues which indicates a suitable environment for a Coffee Shop, and Coffee Shops already present. It also allows for the stakeholder to tune in his own level of aversion to competition.

The flexibility of this approach makes it is easy to generalize to other venues and cities. Moreover, since the model can learn which are the most suitable areas from other kinds of venues, it has the ability to single out suitable areas for the new venue that may not have been identified by the market yet (no or very few venues of that kind already installed)!

As a last remark, we observe that in order to improve even more the statistical relevance of the correlations used and therefore the overall accuracy of the model, the model could be trained over a larger set of data. This can be done selecting a number of cities within the same “cultural region” (the same state, country or geographical region for example) and computing the correlations between the number of Coffee Shops and other venues in all of those cities.