Is life in Toronto linear or nonlinear?

1 Introduction

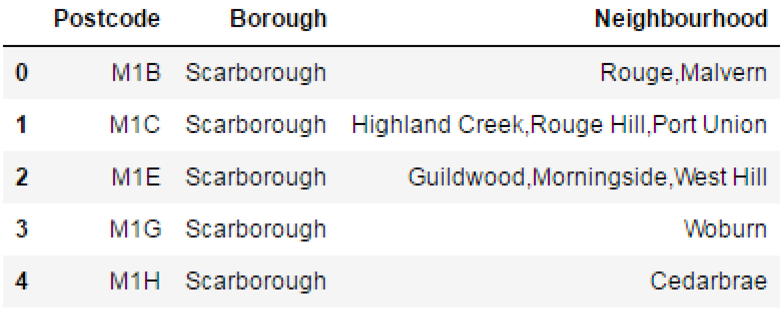
When people are hungry, they will go to a restaurant, when people are thirsty, they will find a coffee shop, and when people need new clothes, they will go to store. These venues are reflections of people’s activities. The locations of venues are not random. For example, a food joint usually picks a location where most people visit, because the more people pass by, the more likely they pick up some food in this food joint.

And an interesting question is, from a geography point of view, are the venues located linearly or nonlinearly? Using locations of venues as a proxy, we can take a look at the life pattern of people.

In this project, we will explore the locations of venues of Toronto. And answer the question, do venues in Toronto located linearly or nonlinearly, which type of venue is located linearly or nonlinearly.

2 Data

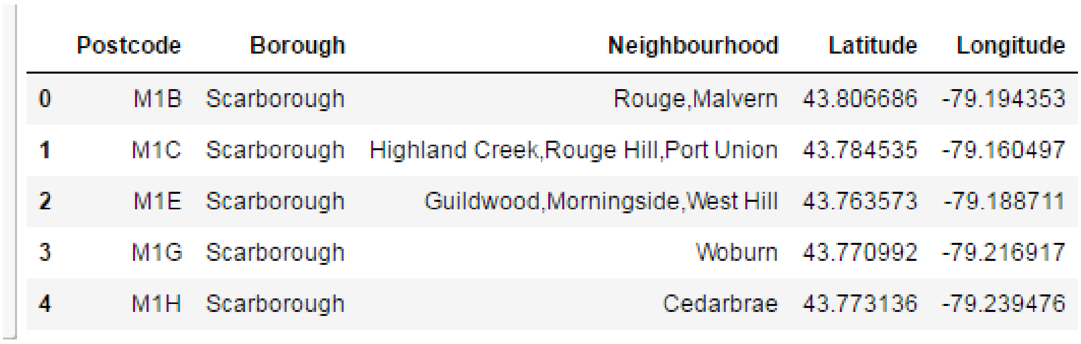
The data used in this project are acquired through Wikipedia and Foursquare. First the postcodes, Borough and Neighborhood information is downloaded from Wikipedia page. After removing not assigned borough and combine neighborhood in the same borough, we get 103 neighborhoods.



	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

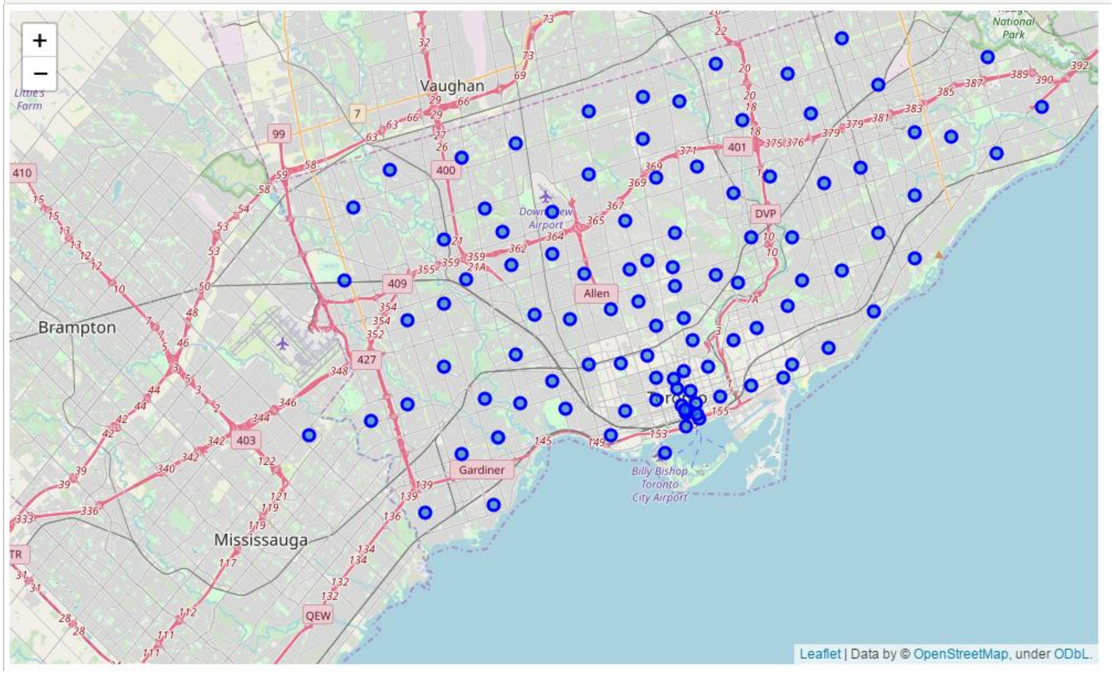
Figure 1 DataFrame of postcode, Borough and Neighborhood

Then for the locations of the neighborhood, we use latitude and longitude to represent the physical location of the neighborhood. The latitude and longitude could be achieved through geocoder, however the codes to get coordinates from geocoder always fail because max retries exceed with url. In this project we use the coordinate data directly from the webpage provided in week3 of Applied Data Science Capstone.



	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

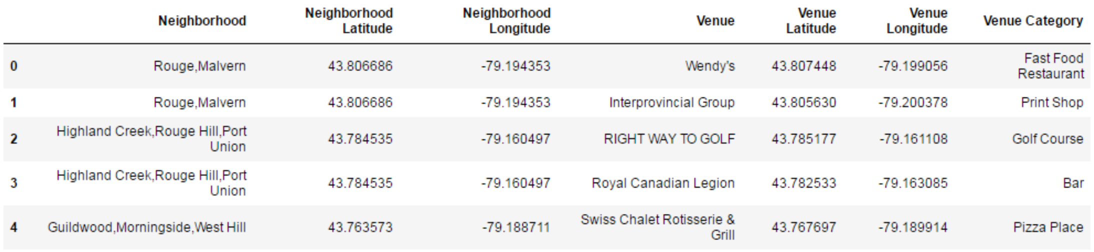
Figure 2 DataFrame of postcode, Borough and Neighborhood, with coordinates



Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 3 map of Toronto where neightborhoods are marked

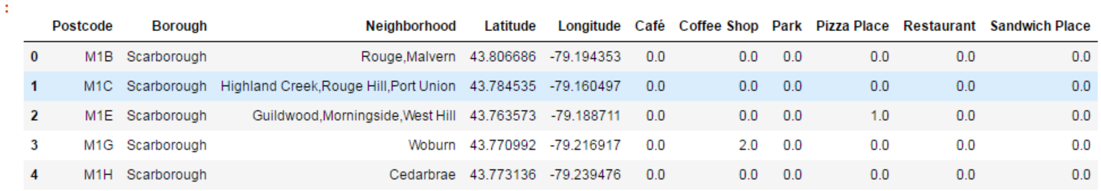
Once we have the coordinates of the neighborhoods, venues in each neighborhood can be accessed using Foursquare API. The data we get contains the type of the venue, and the venue's coordinates. When accessing the venue data, we limit the maximum number of venues to be 100. For all the neighborhoods we get 2255 venue information. And these 2255 venues belong to 274 unique venue types.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge,Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge,Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	RIGHT WAY TO GOLF	43.785177	-79.161108	Golf Course
3	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Guildwood,Morningside,West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

Figure 4 sample of DataFrame containing venue information

The venues are grouped by their types and their neighborhoods. The dataframe looks like this:



:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Café	Coffee Shop	Park	Pizza Place	Restaurant	Sandwich Place
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	0.0	0.0	0.0	0.0	0.0	0.0
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	0.0	0.0	0.0	0.0	0.0	0.0
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	0.0	0.0	0.0	1.0	0.0	0.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	0.0	2.0	0.0	0.0	0.0	0.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5 sample of DataFrame where venue types are processed by one-hot method

Because regression analysis is used in this project, we need to filter out some venue types that do not have enough sample. In this project we choose the venue types that are not zero In at least 30 neighborhoods. Also before we feed the data into the model, for each venue type selected, the neighborhoods that do not have venues of this type are removed.

3 Methodology

In this project regression is used to explore the relations between number of venues and coordinates of neighborhoods. In detail, linear regression and nonlinear regression models will be used.

3.1 Linear model

According to the lesson in Data Analysis with Python, a linear model takes the following form:

$$\mathrm{Yhat} = \mathrm{a} + \mathrm{bX}$$

In the above formula, Yhat is the predicted value for dependent variable. In this project, it is the number of certain type of venues in the neighborhood. X is independent variable. And in this project, it is the latitude and longitude of neighborhood.

For example, generate a data set which follows the relatioin:

$$\mathrm{Yhat} = 4 + 3\mathrm{X} + \mathrm{e}$$

In the above formula, e is a random series follows normal distribution with mean 0 and variance 10. In a regression plot, the linear model looks like the following figure.

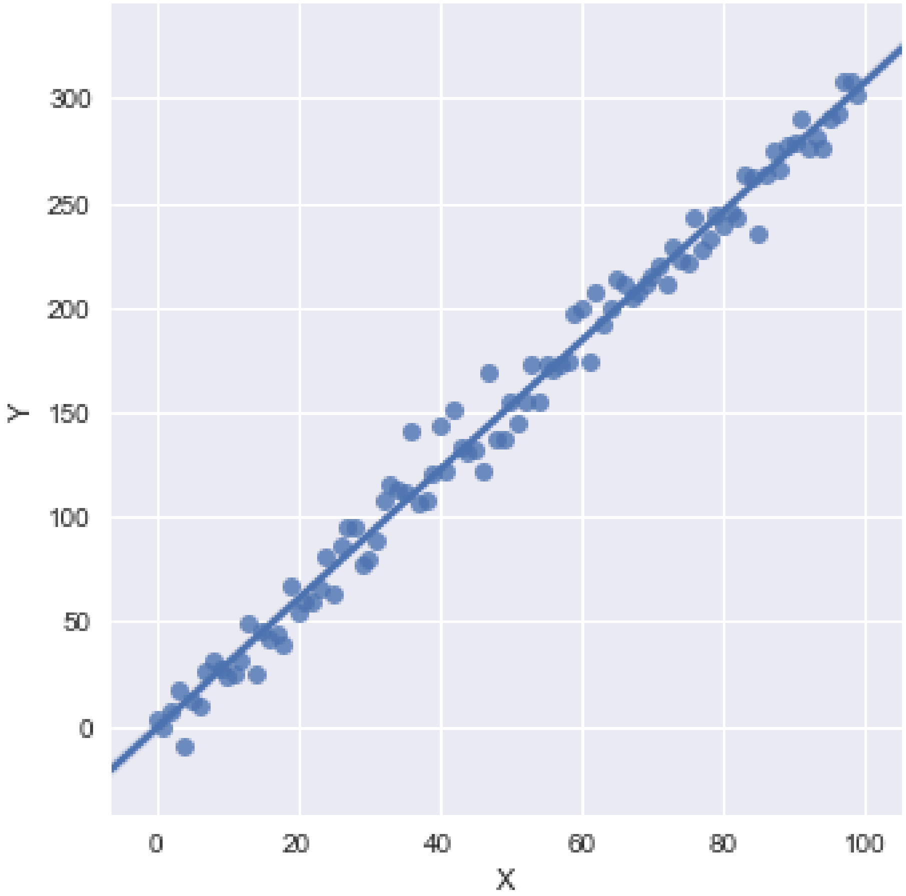


Figure 6 exmple of linear model

3.2 Nonlinear model

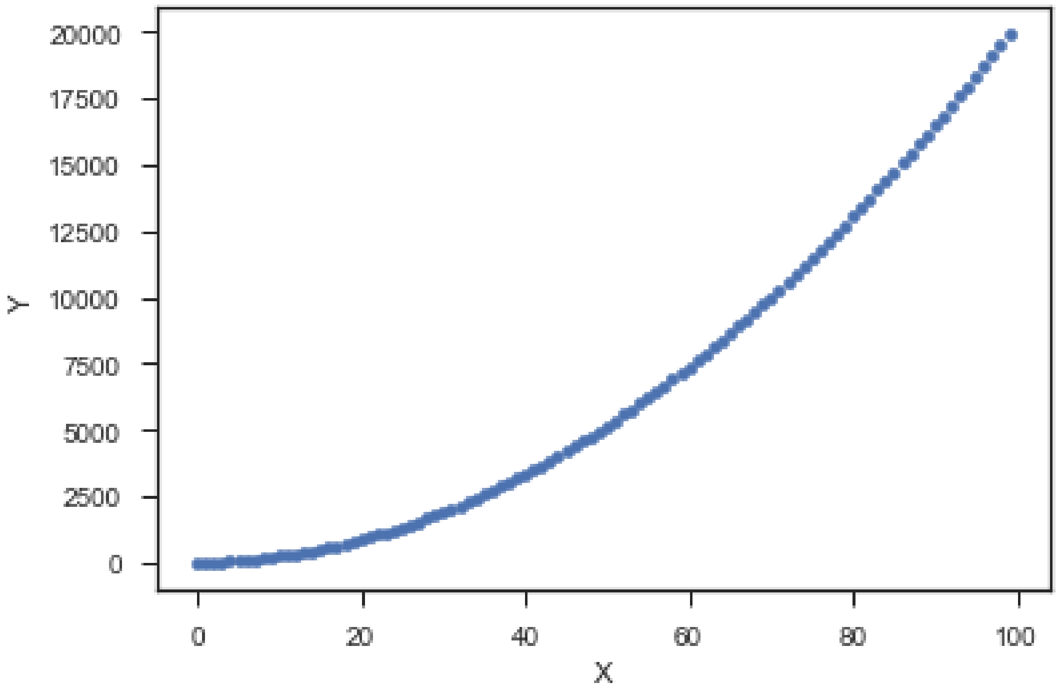
There are many ways to model nonlinear relationships between dependent variable and independent variables. In this project, a linear regression model containing the second order of independent variable will be used:

$$\mathrm{Yhat} = \mathrm{a} + \mathrm{bX} + \mathrm{cX}^2$$

And the Yhat and X have the same meaning as in section 3.1 Linear model. For example, for a data set following the relation:

$$\mathrm{Yhat} = 4 + 3\mathrm{X} + 2\mathrm{X}^2 + \mathrm{e}$$

The figure looks like:



3.3 Model evaluation

To evaluate how good the model fits the data, a measure is needed. Among the measures r squared is chosen. R squared is also called the coefficient of determination. It is the most frequently used in Statistics to determine the accuracy of a model. The meaning of r squared is a number indicating how close the data is to the fitted regression line.

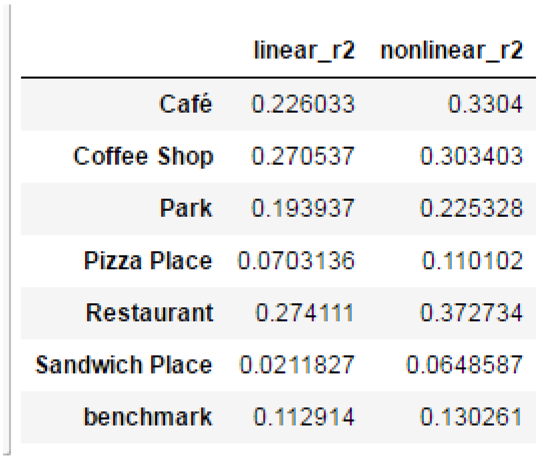
The formula for calculating r square is:

$$R^2 = 1 - \frac{\sum(\mathrm{Y} - \mathrm{Yhat})^2}{\sum(\mathrm{Y} - \mathrm{Ymean})^2}$$

In the above formula, Ymean is the average of Y, and Yhat is the predicted value of Y. The larger the r squared, the better the model fit the data. And the range of r squared is between 0 and 1.

4 Results

Before linear model and nonlinear model are fitted to the data. A benchmark is needed. Because the geography of the city has a high impact on locations of venues. If the city is a thin rectangular shape, then the distribution of the locations of venues will most likely be rectangular shaped. All of our results are based on the comparison to a benchmark. The benchmark chosen here is the linear model and nonlinear model fitted to all the data in the project. It is a regression on venues ignoring venue types. And the results are shown in the following table



	linear_r2	nonlinear_r2
Café	0.226033	0.3304
Coffee Shop	0.270537	0.303403
Park	0.193937	0.225328
Pizza Place	0.0703136	0.110102
Restaurant	0.274111	0.372734
Sandwich Place	0.0211827	0.0648587
benchmark	0.112914	0.130261

Table 1 regression results

The above table contains two columns. The first column is the r squares of the linear model for each type of venue, for example the model for type Cafe is :

# of venues of type Cafe in the neighborhood = a+b1*latitude+b2*longitude

b1 and b2 are the coefficients of coordinates, and a is the intercept of the model

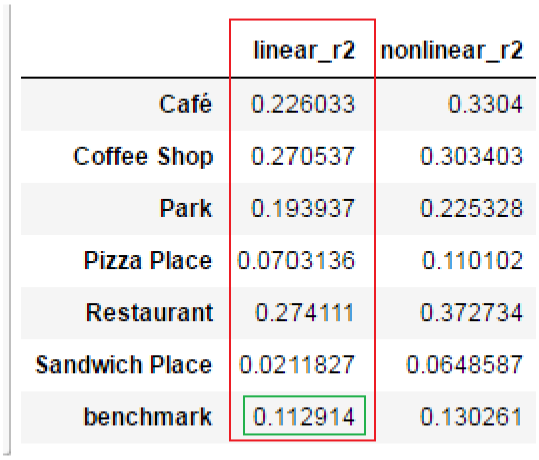
And the second column is the r squared of nonlinear model. For example, the model for type Café is:

# of venues of type Cafe in the neighborhood =
a+b1*latitude+b2*longitude+c1*latitude^2+c2*longitude^2

b1 and b2 are the coefficients of first order coordinates, c1 and c2 are the coefficients of second order coordinates and a is the intercept of the model.

4.1 Linear results

First let’s take a look at the linear regression model results.



	linear_r2	nonlinear_r2
Café	0.226033	0.3304
Coffee Shop	0.270537	0.303403
Park	0.193937	0.225328
Pizza Place	0.0703136	0.110102
Restaurant	0.274111	0.372734
Sandwich Place	0.0211827	0.0648587
benchmark	0.112914	0.130261

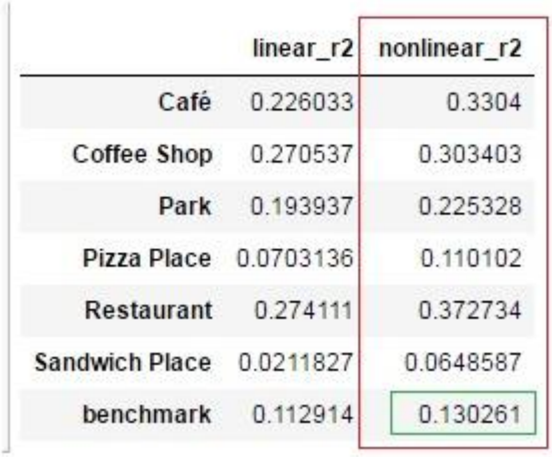
Table 2 linear regression results

In the above table, the r squared of linear model is highlighted by the red rectangle, and the benchmark is highlighted by the green rectangle. The coefficients and intercepts are not listed because we only want to know if a certain type of venue is located linearly, the position of the line is irrelevant.

The average of linear r squared is 0.176 (excluding the benchmark), and the standard deviation is 0.097. We can see the average 0.176 is much higher than the benchmark r squared. This means comparing to the locations of different types of venues as a whole, single type of venues is more linear. And we can also see that the standard variance is more than half of the average r squared, this shows the difference between r squared of different types of venues is large. Indeed the highest r squared is 0.274 for type Restaurant. And the lowest r squared is 0.021 for type Sandwich Place.

4.2 Nonlinear results

The Nonlinear results are tabulated in the following table:



	linear_r2	nonlinear_r2
Café	0.226033	0.3304
Coffee Shop	0.270537	0.303403
Park	0.193937	0.225328
Pizza Place	0.0703136	0.110102
Restaurant	0.274111	0.372734
Sandwich Place	0.0211827	0.0648587
benchmark	0.112914	0.130261

Same as linear results, the results for nonlinear r squared are highlighted in the red rectangle. And the benchmark is highlighted in the green rectangle. Also same as the linear model, we only show the r squared because the actual shape is not the concern of the project, we only want to know how good the locations of venues follow a nonlinear model.

The average r squared of nonlinear model (excluding benchmark) is 0.234 and the standard deviation is 0.114. For the nonlinear models, we can also observe that modeling venues according to their types separately yield higher r squared than treating them as a whole. This phenomenon is more obvious than linear model. Because the average r squared is more than twice of the benchmark. And the highest r squared is as high as 0.373, nearly 200% more than the benchmark.

5 Discussion

According to the r squared results shown in section 4 Results, we can see that if we group venues by different types, both linear models and nonlinear models have higher r squared. Moreover, nonlinear model tends to have higher r squared then linear model. These observations show that life in Toronto is indeed nonlinear. And the distributions of venues of the same type fit nonlinear model better than a linear model. At a glance, the venues in Toronto seem randomly located, but after separating venues according to venue type, certain pattern can be discovered.

Of all the venue types studied, restaurants show highest r squared both in linear model and nonlinear model, which means when opening new restaurants, location is very important. Moreover, there are patterns for existing restaurants, therefore a careful study of existing restaurant is highly recommended when opening new ones.

On the other hand, sandwich place has the lowest r squared for both linear model and nonlinear model. This result shows that for sandwich place, the locations are more likely to be random.

Final observation from the results is that the average r squared of nonlinear model, 0.234, is higher than 0.176, which is the average of linear model r squared.

6 Conclusion

In this project, we use linear and nonlinear models to study the venues of Toronto. By separating the venues into different groups based on their types. We find that despite the venues as a whole do not show any linear or nonlinear patterns, the distribution of certain type of venues fit linear and nonlinear model better. For example, restaurants have a much higher r squared for both linear and nonlinear models, comparing to benchmark. This result indicates that when choosing location for a new restaurant, location is very important and worth further studying existing restaurant location patterns.

And by comparing linear and nonlinear model results, we find that the average r squared of nonlinear model is higher than that of linear model. And to answer the question of this project, life in Toronto is indeed nonlinear.

7 Reference

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
'https://cocl.us/Geospatial_data
Forsquare API
Model Development, Data Analysis with Python, courser lab