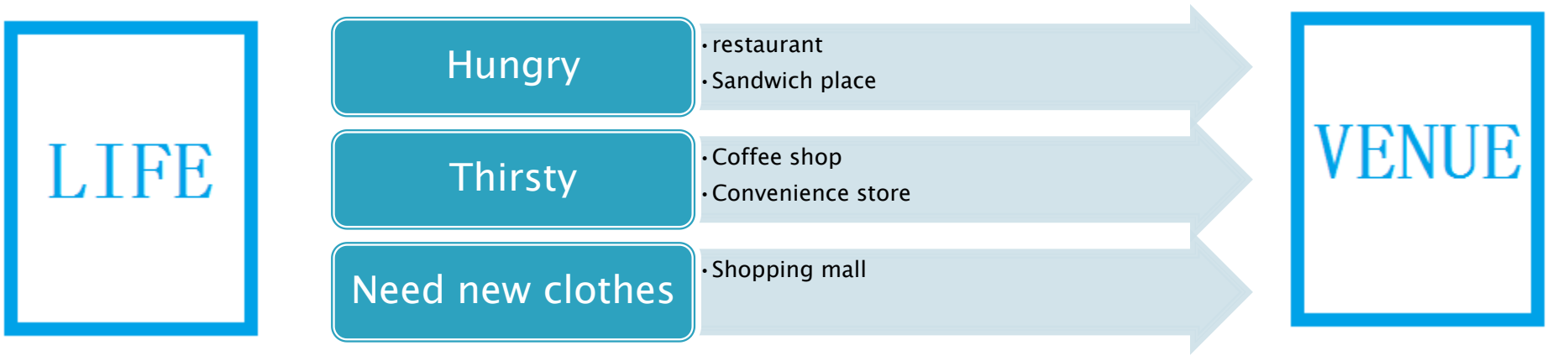
Is life in Toronto linear or nonlinear?

Ding Hao
2019-9-9
project for Applied Data Science Capstone

Part one: Introduction



These venues are reflections of people's activities

Part one: Introduction

- From a geography point of view, are the venues located linearly or nonlinearly?

- 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.

Part two: Data

- Data from Wikipedia

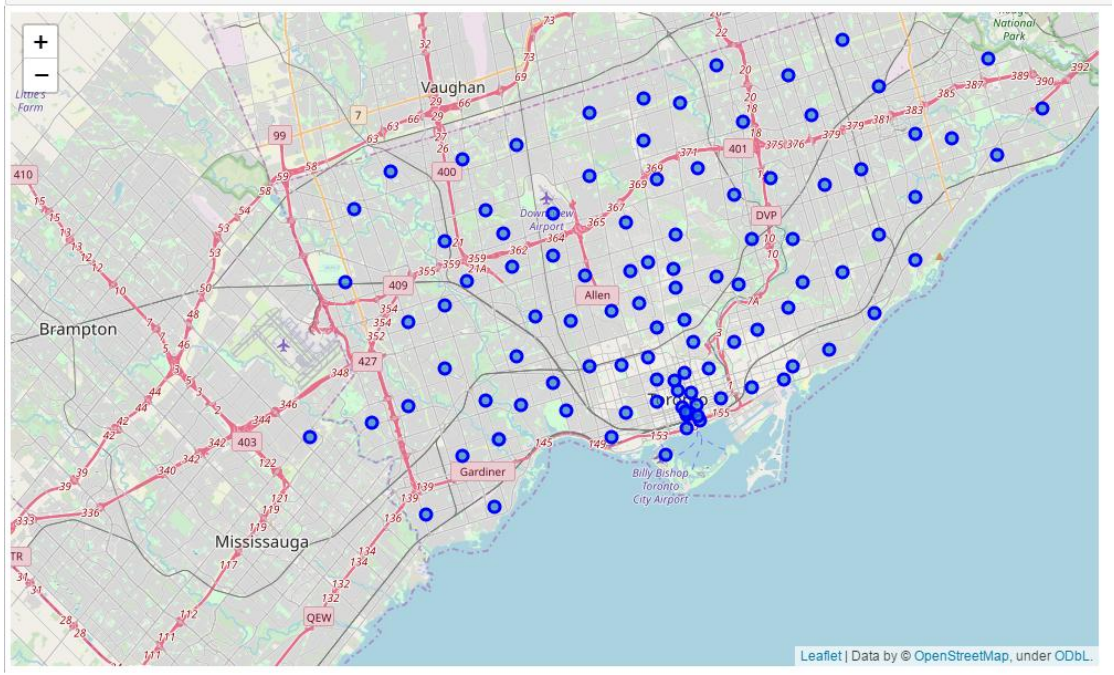
	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

- And geocoder

	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

Part two: Data

- With the help of Foursquare



	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

Part two: Data

- Venues are grouped by their types and their neighborhoods

:

	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

- Choose the venue types that are not zero In at least 30 neighborhoods. Nans are removed before any calculation.

Part three: Methodology

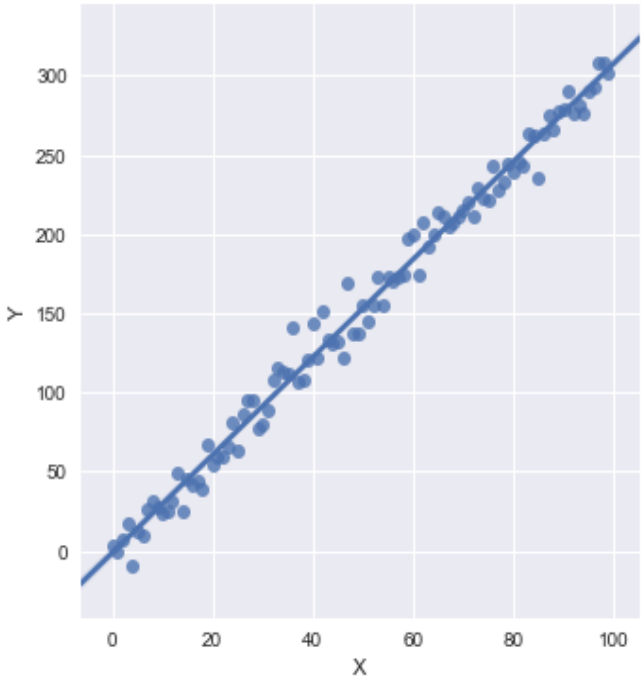
- Linear model
  - Yhat – predicted dependent variable. # of certain type of venues in the neighborhood.
  - X – independent variable. Latitude and longitude of neighborhood.

$$Yhat = a + bX$$

Part three: Methodology

- Linear model example

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



Part three: Methodology

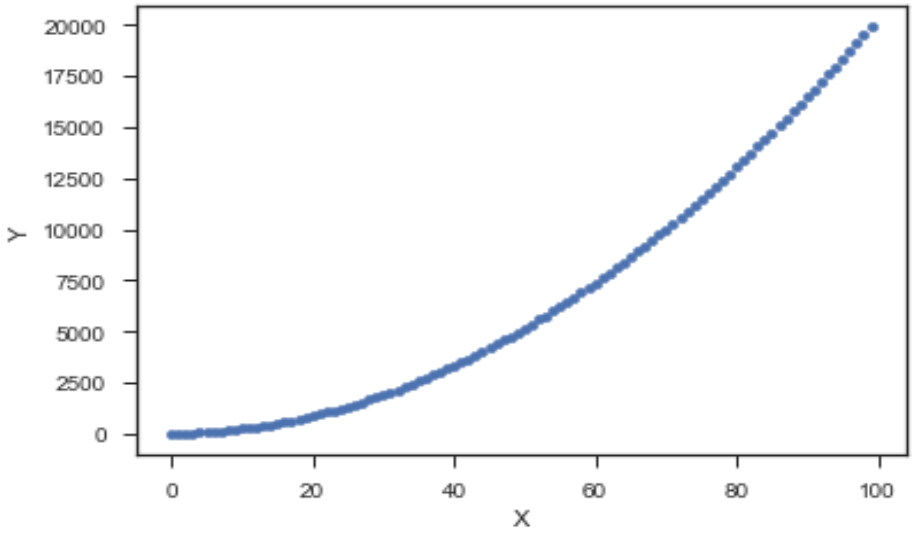
- Nonlinear model
  - Yhat and X have the same meanings as in linear model

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

Part three: Methodology

- Nonlinear model example

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



Part three: Methodology

- Model evaluation

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

- R squared, also called the coefficient of determination.
- One of the most frequently used in Statistics to determine the accuracy of a model.
- indicating how close the data is to the fitted regression line
- The larger the r squared, the better the model fit the data.
- And the range of r squared is between 0 and 1.

Part four: Results

- Benchmark
  - the linear model and nonlinear model fitted to all the data in the project
  - the geography of the city has a high impact on locations of venues.
    - city is a thin rectangular shape → locations of venues will most likely be rectangular shaped.
  - All of our results are based on the comparison to a benchmark.

Part four: Results

- Overall 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

- Column1: r squared of the linear model
  - 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
- Column2 : 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.

Part four: Results

- Linear 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

  - average: 0.176, std: 0.097
  - single type of venues is more linear than benchmark
  - std is large comparing to average. 0.274 (Restaurant) vs 0.021(Sandwich Place)

Part four: Results

- Nonlinear 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

  - average: 0.234, std: 0.114
  - single type of venues fit nonlinear model better than benchmark
  - 0.373 (Restaurant) vs 0.13(benchmark)

Part five: Discussion

- group venues by different types, both linear models and nonlinear models have higher r squared than benchmark
- nonlinear model tends to have higher r squared then linear model
- restaurants show highest r squared both in linear model and nonlinear model, which means when opening new restaurants, location is very important
- 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.

Part six: conclusion

- The distribution of certain type of venues fit linear and nonlinear model better
- Life in Toronto is more likely to be nonlinear than linear

Part seven: 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