Capstone Project ->Battle of Neighbourhood.

INTRODUCTION:

The goal of the project is to explore the neighbourhood of New York City in order to explore the relationship between real estate value and surrounding venues.The plan comes from the process of family migrating to new city. It s common tha owners,/ agents advertise their properties are close to some time of venue like school, supermarket, restaurants, hospitals and so on, showing the 'conveniece' of the locality to raise the property value.

Does the surrounding venue rise the price of the house and if so , how it is correlated.

The target audience are:
Potential buyer/renter who roughly estimate the value of the house/ rent/lease based on surrounding venuea and average price

#Real Estate Makers and planners who can decide what kind of venues to put around their products to maximize selling price.

#House seller who can optimize their advertisements

DATA DESCRIPTION:

New York city neighbourhood are chosen as the observation target for the following reasons:

#The availability of geo data which can be used to visualize the dataset onto a map.

#The diversity of price between neighbourhood

#The availability of real estate prices.

The dataset will be composed of the two main source:
-CityRealty which provides the neighborhood average prices
-Foursquare API which will provide surrounding venues of given latitude , longitude.

DATA COLLECTION & CLEANING:

-Scrap the CityRealty website for a list of New York City neighborhoods and their corresponding condo average price.
-Find geographic data of the neighbourhoods, both their ceter coordinates and their border.

-For each neighborhood, pass the obtained coordinate to the FourSquare API. The "explore endpoint will return a list of surrounding venues in a pre-defined radius.
-Count the occurrence of each venue type in a neighborhood and then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
-Standardize the average price by removing the mean and scaling to unit variance.

The result of the dataset is a two dimensional dataframe (Diagram : 1)
-Each row represents a neighbourhood
-Each column , except the last one is the occurance of the venue type.
The last column will be standardized average price.

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Animal Shelter	Antiq Shop	...	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	StandardizedAvgPrice
0	Battery Park City	0	0	0	3	0	0		0	1	4	0	1	0	-1.303912
1	Bedford-Stuyvesant	0	0	0	0	0	0		0	1	6	0	0	1	-0.418350
2	Boerum Hill	0	0	0	1	0	0		0	0	2	0	0	2	0.015011
3	Brooklyn Heights	0	0	0	2	0	0		0	1	4	0	0	5	-1.099479
4	Bushwick	0	0	0	1	0	0		0	0	1	0	0	2	-0.587926

Diagram : 1

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to Foursquare API may return different recommended venue at different time.

The number of features are more than number of samples. This may cause issue during analysis. Details and counter measurement will be discussed in the later section.

METHODOLOGY:

It is assumed that real estate price depends on the surrounding venues. Thus regression techniques will be used to analyze the dataset. The regressors will be occurrences of the venue type and the dependent variable will be standardized average prices.

At the end, the regression model will be obtained. Along with the coefficients list which describes how each venue type may be directly or inversely related to the neighborhood's real estate average price around the mean.

1. First insight using visualization:
In order to have a first insight of New York city real estate average price between neighborhoods, there is no better way than visualization.
The medium chosen is Choropleth map, which uses differences in shading or coloring to indicate a property's values or quantity within predefined areas. It is

ideal for showing how differently real estate priced between neighborhoods across the New York city map.
The map (Diagram 2) shows high price in neighborhoods that located around Central Park, Midtown and Lower Manhattan. The price reduces further toward North Manhattan or toward Brooklyn.
Manhattan can be considered the heart of New York city. It's where most businesses, tourist attractions and entertainments located. So, the venue types that can attract many people are expected to have the most positive coefficients in the regression model.

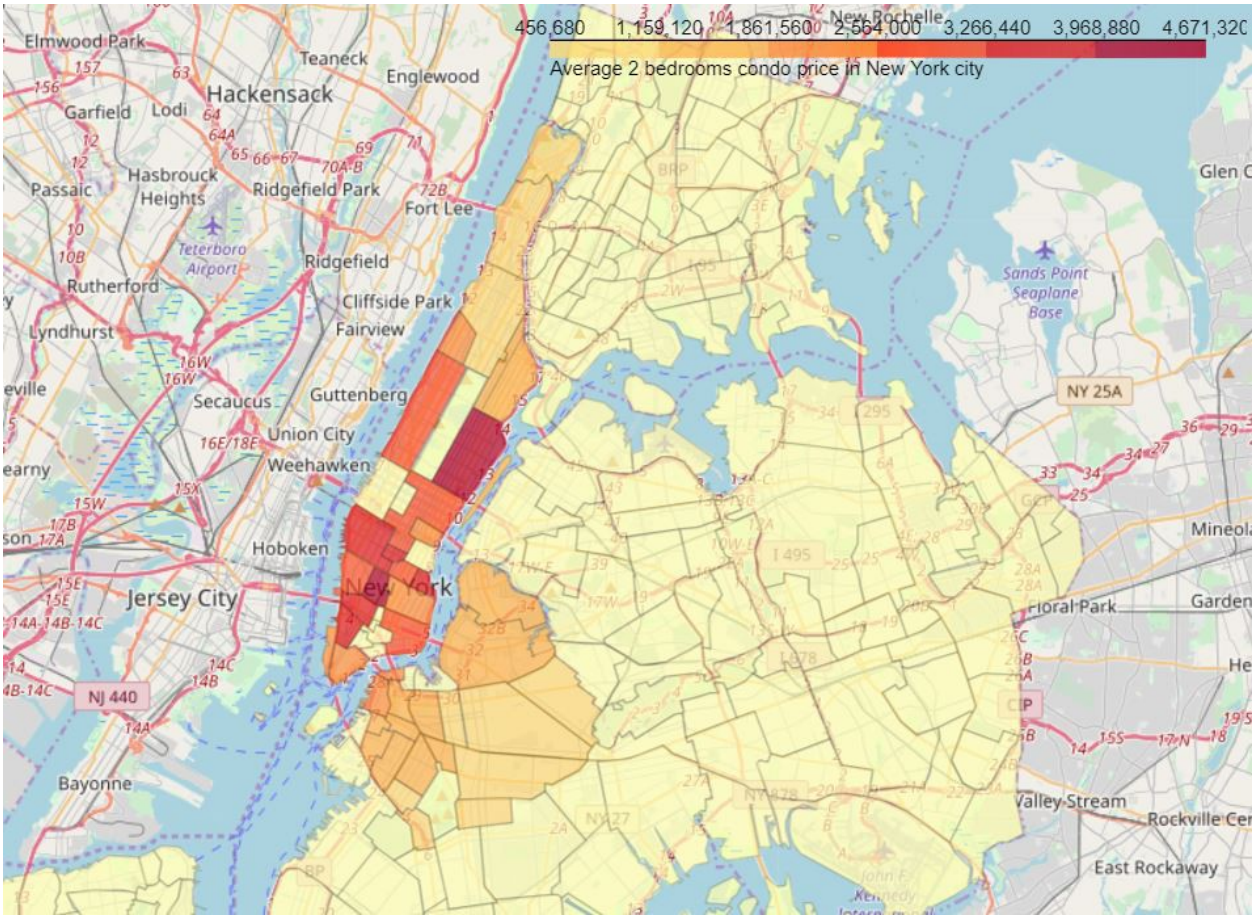

Diagram 2: New York city real estate price spread between neighborhoods

2. LINEAR REGRESSION:

Linear Regression was chosen because it is a simple technique. And by using Sklearn library, implementing the model is quick and easy. Which is perfect to start the analyzing process.

The model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

The result (Diagram 3) doesn't seem very promising. R2 score is small, which means the model may not be suitable for the data.

```
R2-score: 0.273792308888
Mean Squared Error: 0.254179706388
Max positive coefs: [ 0.26348338  0.26213799  0.26213799  0.26213799  0.25818747  0.25818747
  0.25135936  0.24564842  0.23349638  0.22658134]
Venue types with most postive effect: ['Design Studio' 'Train Station' 'Jewish Restaurant' 'Resort' 'Buffet'
 'Cafeteria' 'Colombian Restaurant' 'Dumpling Restaurant' 'Other Nightlife'
 'Botanical Garden']
Max negative coefs: [-0.20813947 -0.20763403 -0.1798399  -0.1798399  -0.1798399  -0.17776278
 -0.17776278 -0.17776278 -0.17776278 -0.17776278]
Venue types with most negative effect: ['Board Shop' 'Gay Bar' 'Supplement Shop' 'Rest Area' 'Lighthouse' 'Office'
 'Flea Market' 'Golf Driving Range' 'Recreation Center'
 'General Entertainment']
Min coefs: [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
Venue types with least effect: ['TV Station' 'Gas Station' 'Pakistani Restaurant' 'Volleyball Court'
 'Hookah Bar' 'Indoor Play Area' 'Laser Tag' 'Christmas Market' 'Cemetery'
 'Mini Golf']
```

Diagram 3 - Linear Regression result

But on the bright side, the coefficient list shows some interest and logical information:

“Studios” and “Eateries” both mean businesses. “Train Station” means ease of transportation. All of which usually increase the value of a location.

“Bar” and “Market” sure are nice to visit sometimes but may not be a suitable neighborhood for family with kids. “Lighthouse” and “Golf” usually located in the rural areas. The demand for such locations is usually low.

“TV station”, “Cemetery”, “Laser Tag”, “Mini Golf” all give value to a limited range of people. “Gas Station” is available everywhere. These types of venue usually are not dicision factor when considering a location.

Back to the model, what seems to be the problem? And what are the possible solutions?

Looking back further to the dataset, its dimensions sizes is clearly unbalanced, only 50 samples, and more than 300 features. Logical steps to take are either collecting more samples or trying to reduce the number of features.

But since there are no other public source available, increasing sample size is not possible at the moment. So, deceasing features is the only option for now.

And that’s why Principal Component Regression is chosen to analyze the dataset in the next part.

3. Principal Component Regression (PCR):

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression. (Wikipedia, n.d.)

PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. As the result, the number of features is reduced while keeping most of the characteristic of the dataset.

Then PCR use Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

Again, R2 score and MSE are used to see how well the model fit the dataset.

```
R2 score: 0.454460324852
MSE: 0.190944155714
```

Figure 4 - PCR scores

The result is promising as it shows improvement over the simple Linear Regression.

As for the coefficient list, the size has been reduced after performing PCA. So, a dot product with eigenvectors is needed to get it back to the original features size.

```
Max positive coefs: [ 0.07212567  0.0696754   0.06052737  0.0582199   0.05228078  0.05222561
  0.04901431  0.04597368  0.04465698  0.04399769]
Venue types with most positive effect: ['Dumpling Restaurant' 'Pilates Studio' 'Design Studio' 'Pie Shop'
 'Southern / Soul Food Restaurant' 'Library' 'Sushi Restaurant' 'Resort'
 'Korean Restaurant' 'Buffet']
Max negative coefs: [-0.05116074 -0.03897274 -0.03710211 -0.03457056 -0.03452567 -0.0345195
 -0.03414522 -0.03304223 -0.03284579 -0.03284275]
Venue types with most negative effect: ['Market' 'Lingerie Store' 'Gay Bar' 'Kosher Restaurant' 'Optical Shop'
 'Food' 'Food Truck' 'Wine Bar' 'Food & Drink Shop' 'Climbing Gym']
Min coefs: [ -8.90366289e-06  -8.90366289e-06   4.09236430e-05  -4.99918920e-05
  -5.87234477e-05   1.27322576e-04   1.27322576e-04   1.27322576e-04
   1.27322576e-04   1.41722883e-04]
Venue types with least effect: ['Christmas Market' 'TV Station' 'Cemetery' 'Event Space'
 'Indoor Play Area' 'Modern European Restaurant' 'Mini Golf'
 'Volleyball Court' 'Molecular Gastronomy Restaurant' 'Community Center']
```

Figure 5 - Coefficient list in original size

The insight is still consistent compared to the Linear Regression's.

IV. RESULTS:

Even though the scores seem to be improved after applying a more sophisticate method, the model is still not suitable for the dataset. Thus, it can't be used to precisely predict a neighborhood average price.

Explanations for the poor model can be:

The real estate price is hard to predict.

The data is incomplete (small sample size, missing deciding factors).

The machine learning techniques are chosen or applied poorly.

But again, on the bright side, the insight, gotten from observing the analysis results, seems consistent and logical. And the insight is business venues that can serve the needs of most normal people usually situated in pricy neighborhoods.

V. DISCUSSION:

The real challenge is constructing the dataset:

Usually the needed data isn't publicly available.

When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.

For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.

Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

VI. CONCLUSION:

It's unfortunately that the analysis couldn't produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Some notes on the analysis result:

This project is done by a web developer who only started self-studying Data Science for 4 months. So please take it with a grain of salt.

The coefficients only show correlation, not causation. So, if your neighborhood average price is low, please don't go destroying the surrounding bars and food trucks. There might be another reason.

Toward the person that went through this project, many thanks for the time and patient.

REFERENCES:

Wikipedia. (n.d.). Principal component regression. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Principal_component_regression