IBM Data Science Professional Certification Course

Final Capstone project

Relation between a Real estate prices and its surrounding venues

Table of content:

I. Introduction:

This report is for the final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach can be designed with available data, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

The main goal will be exploring the neighborhoods of New York city in order to extract the correlation between the real estate value and its surrounding venues.

The idea comes from the process of a normal family finding a place to stay after moving to another city. It's common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, etc.; showing the "convenience" of the location in order to raise their house's value.

So, can the surrounding venues affect the price of a house? If so, what types of venues have the most affect, both positively and negatively?

The target audience for this report are:

- Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price.
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements.

II. Data description:

New York city neighborhoods were chosen as the observation target due to the following reasons:

- The availability of real estate prices. Though very limited.
- The diversity of prices between neighborhoods.
- The availability of geo data which can be used to visualize the dataset onto a map.

The type of real estate to be considered is 2-bedroom condo, which is common for most normal nuclear families.

The dataset will be composed from the following two main sources:

- CityRealty which provides the neighborhoods average prices. https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804
- FourSquare API which provides the surrounding venues of a given coordinates.

The process of collecting and clean data:

- Scrap the CityRealty webpage for a list of New York city neighborhoods and their corresponding 2-bedroom condo average price.
- Find the geographic data of the neighborhoods. Both their center coordinates and their border.
- For each neighborhood, pass the obtained coordinates to 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. 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 dataset is a 2 dimensions data frame (Figure 1):

- Each row represents a neighborhood.
- Each column, except the last one, is the occurrence of a venue type. The last column will be the 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

Figure 1 - Final dataset

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time.

The number of features is much bigger than the number of samples. This will cause problem for the analysis process. Detail and counter-measurement will be discussed further in the next section.

III. Methodology:

The assumption is that real estate price is dependent on the surrounding venue. Thus, regression techniques will be used to analyze the dataset. The regressors will be the occurrences of venue types. And the dependent variable will be standardized average prices.

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

Python data science tools will be used to help analyze the data. Completed code can be found here:

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 (Figure 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.

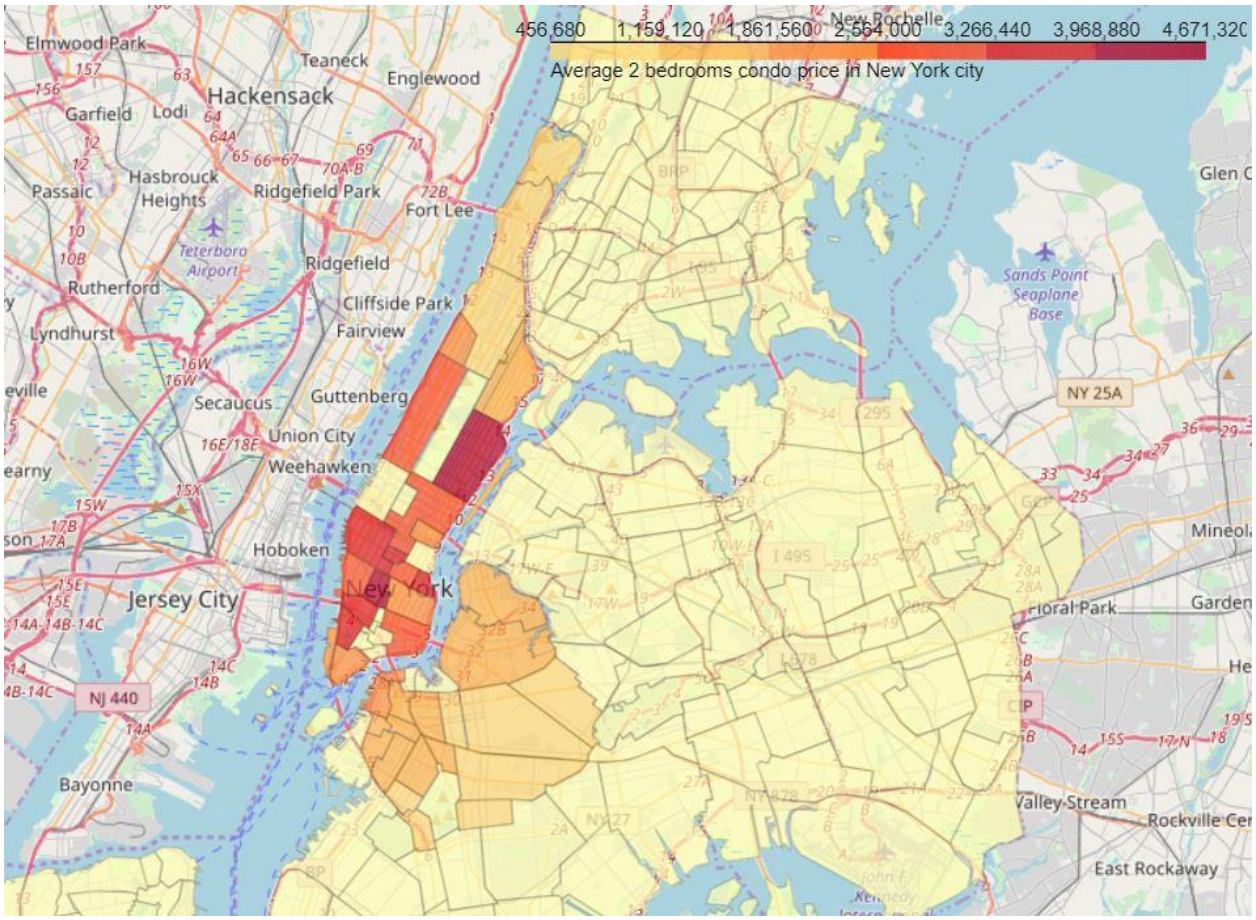

Figure 2 - New York city real estate price spread between neighborhoods

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

```
R2-score: 0.04357374764813149
Mean Squared Error: 0.3652608916407075
Max positive coefs: [0.37170083 0.29611452 0.28806654 0.28806654 0.2826957  0.2520534
 0.2520534  0.2520534  0.2520534  0.22199173]
Venue types with most postive effect: ['General Entertainment' 'Other Nightlife' 'Cafeteria' 'Buffet'
 'Colombian Restaurant' 'Jewish Restaurant' 'Train Station'
 'Persian Restaurant' 'Resort' 'Dumpling Restaurant']
Max negative coefs: [-0.25809931 -0.22186812 -0.22186812 -0.2092001  -0.2092001  -0.2092001
 -0.19697375 -0.18880818 -0.17901003 -0.17901003]
Venue types with most negative effect: ['Board Shop' 'Flea Market' 'Golf Driving Range' 'Street Food Gathering'
 'Other Repair Shop' 'Print Shop' 'Drugstore' 'Street Art'
 'Gluten-free Restaurant' 'Sports Club']
Min coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with least effect: ['Food Stand' 'TV Station' 'State / Provincial Park' 'Hookah Bar' 'Bridge'
 'Factory' 'Molecular Gastronomy Restaurant' 'Shipping Store'
 'Pakistani Restaurant' 'Volleyball Court']
```

Figure 3 - Linear Regression result

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

- “General Entertainment” 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 decision 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.

```
Best n: 49 R2 score: 0.279494010789739
Best n: 49 MSE: 0.25218405564964974
```

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: [0.06363667 0.06263333 0.05825532 0.05714383 0.05082018 0.0508087
 0.04752883 0.04676123 0.04676123 0.04496008]
Venue types with most positive effect: ['Dumpling Restaurant' 'Design Studio' 'Pilates Studio' 'Library'
 'Korean Restaurant' 'Colombian Restaurant'
 'Southern / Soul Food Restaurant' 'Buffet' 'Cafeteria' 'Sushi Restaurant']
Max negative coefs: [-0.05673619 -0.04809685 -0.04443655 -0.04199834 -0.04144816 -0.0400175
 -0.03882273 -0.03803995 -0.0368368  -0.03549397]
Venue types with most negative effect: ['Market' 'Trail' 'Food & Drink Shop' 'Lingerie Store' 'Tapas Restaurant'
 'Garden' 'New American Restaurant' 'Coffee Shop' 'Street Art'
 'Sculpture Garden']
Min coefs: [-2.53016106e-05 -2.53016106e-05 -2.53016106e-05  8.46090084e-05
  1.15505999e-04 -1.51972077e-04  1.86976331e-04  1.86976331e-04
  3.03887206e-04 -3.04732795e-04]
Venue types with least effect: ['Gymnastics Gym' 'Fruit & Vegetable Store' 'Coworking Space'
 'Szechuan Restaurant' 'Vape Store' 'Veterinarian' 'Food Stand'
 'Pakistani Restaurant' 'Food Court' 'Nightclub']
```

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.