APPLIED DATA SCIENCE

Correlation between a neighborhood real estate price and its surrounding venues

Contents

I. Introduction

This report is for the final course of Applied Data Science. The problem and the analysis approach are leveraging the Foursquare location data to explore and compare neighbourhoods or cities of choice and to come up with a question that you can use the Foursquare location data to solve.

The main goal will be exploring the neighbourhoods of New York City to extract the correlation between the real estate value and its surrounding venues.

The idea comes from the process of an ordinary family finding a place to stay after moving to another city. Commonly, the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, showing the “convenience” of the location to raise their house’s value.

So the question is, can the surrounding venues affect the price of a house? If so, what types of places have the most effect, 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.
- Houses sellers who can optimize their advertisements.

II. Data Description

New York City neighbourhoods were chosen as the observation target due to the following reasons:

- The availability of real estate prices.
- The diversity of prices between neighbourhoods.
- The availability of geodata which can be used to visualize the dataset onto a map.

The type of real estate to be considered is a 2-bedroom flat, which is typical for most normal ordinary families.

The dataset will be composed of the following two primary sources:

- CityRealty which provides the neighbourhoods average prices.
- FourSquare API which provides the surrounding venues of a given coordinate.

The process of collecting and clean data:

- Scrap the CityRealty webpage for a list of New York City neighbourhoods and their corresponding 2-bedroom flat average price.
- Find the geographic data of the neighborhoods. Both their centre coordinates and their border.
- For each community, 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 neighbourhood. Then apply one-hot encoding to turn each venue type into a column with their appearance as the value.
- Standardize the average price by removing the mean and scaling to unit variance.

The resulting dataset is a two dimensions data frame :

- Each row represents a neighbourhood.

- 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

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

The number of features is much more significant than the number of samples. The number of elements might cause a problem for the analysis process.

III. Methodology

The assumption is that the 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. Moreover, the dependent variable will be standardized, average prices.

In the end, a regression model will be obtained. With a coefficients list which describes each venue type may be related to the increase and decrease of a neighbourhood's real estate average price.

Python data science tools will be used to help analyze the data. Completed code can be found here: https://github.com/SlyUndying/Coursera_Capstone/blob/master/Capstone%20Project.ipynb

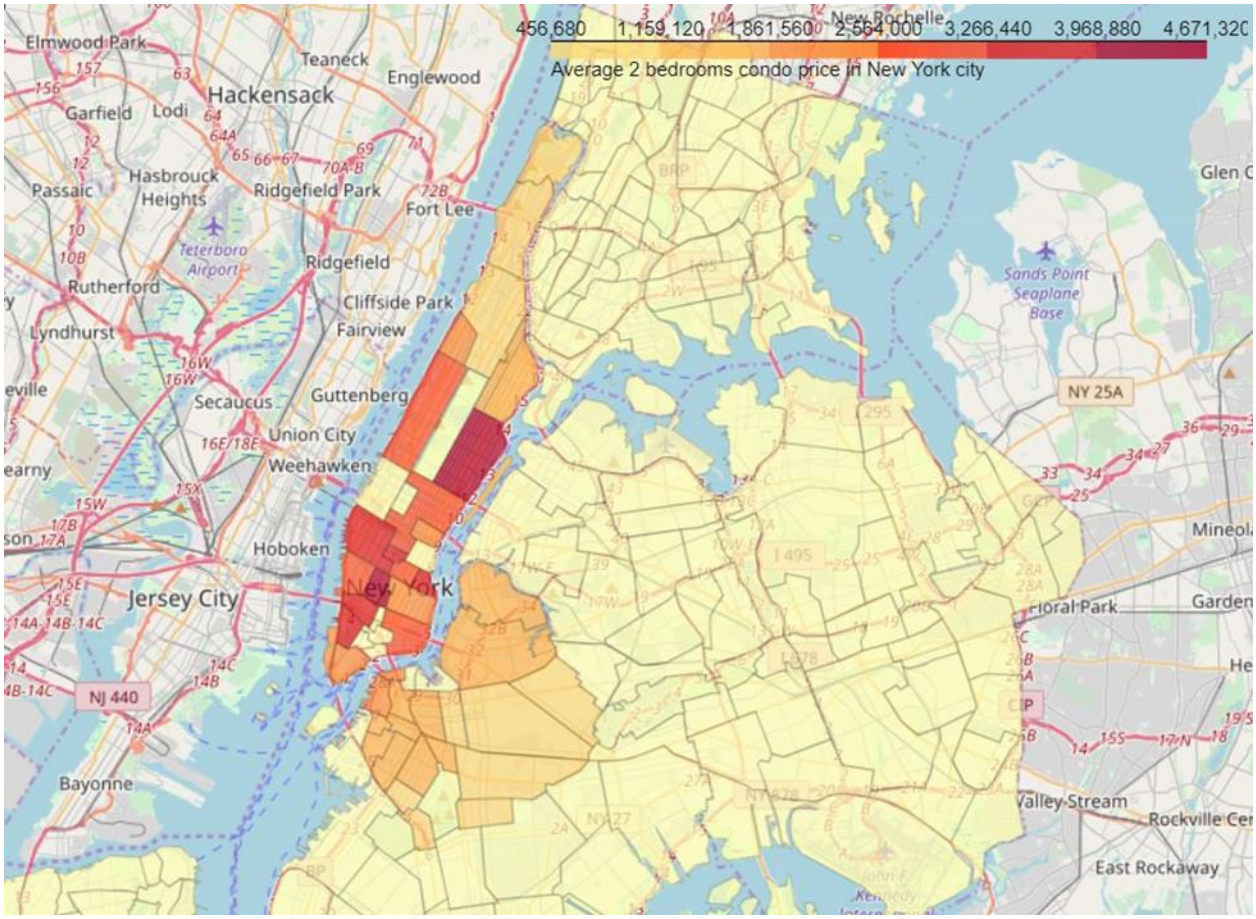
1. First insight using visualization:

The first insight of New York City real estate average price between neighbourhoods, there is no better way than visualization.

The medium chosen is Choropleth map, which uses differences in shading or colouring to indicate a property's values or quantity within predefined areas. It is ideal for showing how differently real estate priced between neighbourhoods across the New York city map.

The map shows a high price in neighbourhoods that located around Central Park, Midtown and Lower Manhattan. The price reduces further toward North Manhattan or 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.



2. Linear Regression:

Linear Regression was chosen because it is a simple technique. Moreover, 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 doesn't seem very satisfying. 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']
```

The coefficient list shows some interest and relevant 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 but may not be a suitable neighbourhood for the family with kids. “Lighthouse” and “Golf” usually located in rural areas. The demand for such locations is generally 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 a decisive factor when considering a location.

Back to the dataset, its dimensions sizes are 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.

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

Moreover, 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 only as of the combination of Principal Component Analysis (PCA) with Linear Regression.

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 a result, the number of features is reduced while keeping most of the characteristic of the dataset.

Then PCR uses Linear Regression on the converted set to return a coefficient list, just like in traditional 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
```

The result is more satisfied 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 size of the original feature.

```
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']
```

The insight is still consistent compared to the Linear Regression.

IV. Result

Even though the result seems to be improved after applying a more sophisticated method, the model is still not suitable for the dataset. It can't be used to predict a neighborhood average price precisely.

Explanations for the reduced model can be:

- The real estate price is hard to predict.
- The data is incomplete (small sample size, missing deciding factors).
- Machine learning techniques are chosen or applied poorly.

The insight, gotten from observing the analysis results, seems consistent and logical. Moreover, the idea is business venues that can serve the needs of most ordinary people, usually situated in pricy neighbourhoods.

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. Moreover, lots of efforts are required to check, research and change the data before the merge.
- For data obtained through API calls, different results are returned with a different set of parameters and different point of time. Many trial and error runs are required to get the optimal result.
- After the dataset has been constructed, lots of research and analysis are needed to decide whether the data should be kept as is or be transformed by normalization or standardization.

It can be considered the most critical process in the whole data science pipeline, which can affect the most on the result.

Choosing a suitable technique to construct the model is also a rewarding process. As this report shows that, by applying a different method, the result can be improved.

VI. Conclusion

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

Doing this project helps to practice every topic in the specialization, and 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. Moreover, as this report shows, there are undoubtedly many things to dive in.