Coursera Capstone Project

for

IBM Data Science Professional Certificate

“The Choice for Housing”

Analysis of preferred apartments in New York City

By: Joseph Oisamoje

December, 2018

Table of content:

I. Introduction:

1. Description of the Problem:

Relocating to a big city anywhere in the world is no small endeavor and for a middle aged family this could be very daunting. New York City the big apple as she's fondly called is one of the premier cities in the world and is known as the finance center and a tourist hot-spot in North America and as such the real estate cost are some of the priciest. A quick web search shows that real estate price can vary by a large margin from neighborhood to neighborhood. As an example, a 2-bedrooms condo in Central Park West, Upper West Side can cost up to $4.91 million on average; while in Inwood, Upper Manhattan, just 30 minutes away cost about $498 thousands.

What features of a neighborhood affects the price of real estate to such extend? One hypothesis is that the surrounding venues can be a decision factor.
Surely anyone, who has attempted to find an accommodation for rent or buy, has seen advertisements such as: This condo is located near the subway station, malls, supermarkets, dinners, etc. And it's likely that the price will be higher than others with locations not as "convenient".
Can the venues surrounding an accommodation effect its price? And what kind of venues can affect the most? And by what margin?

2. The question to answer:

This project will try to explore the neighborhoods of New York City to see:

- If the surrounding areas/landmarks can affect the price of real estate.
- The kind of surrounding areas/landmarks, and to what extent affects the price.
- Can we use the surrounding areas/landmarks to estimate the value of an accommodation over the average price of one area? And to what degree of accuracy?

The result can be useful for home buyers/renters, who can roughly estimate the value of a target house over the average in a given area. It can also be of help to realtors as valuable data that shows where projects can be cited for long-term revenue/profitability

II. Description of the data:

The main data used for this project will be from two sources:

- The average price by neighborhoods in New York City culled from a real estate portal. (CityRealty)
- The venues in each neighborhood from location services. (FourSquare API)

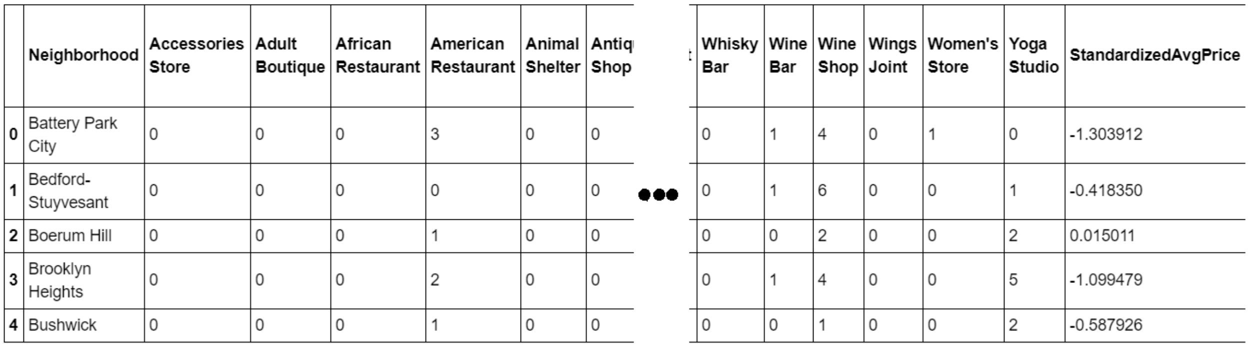
*Note: This project will only consider the average price of 2-bedrooms apartments, which is a common type of real estate among normal families.

1. Data collection process:

- The average price will be scrapped from the CityRealty website.
- For each neighborhood I'll use the Geocoder Python library to get its coordinate.
- For each neighborhood's coordinate, I'll use the FourSquare API to get the surrounding venues.
- Count the occurrences of each venue type and attach that information to each neighborhood.

The output of the data collecting process will be a 2 dimensional table:

- Each row represents a neighborhood.
- Each column will be the count of one type of venue in that neighborhood.
- The last column will be the average 2-bedroom apartment price of that neighborhood.



	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: https://github.com/lethien/coursera-ibm-ds-capstone/blob/master/Capstone_Analyze.ipynb

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.

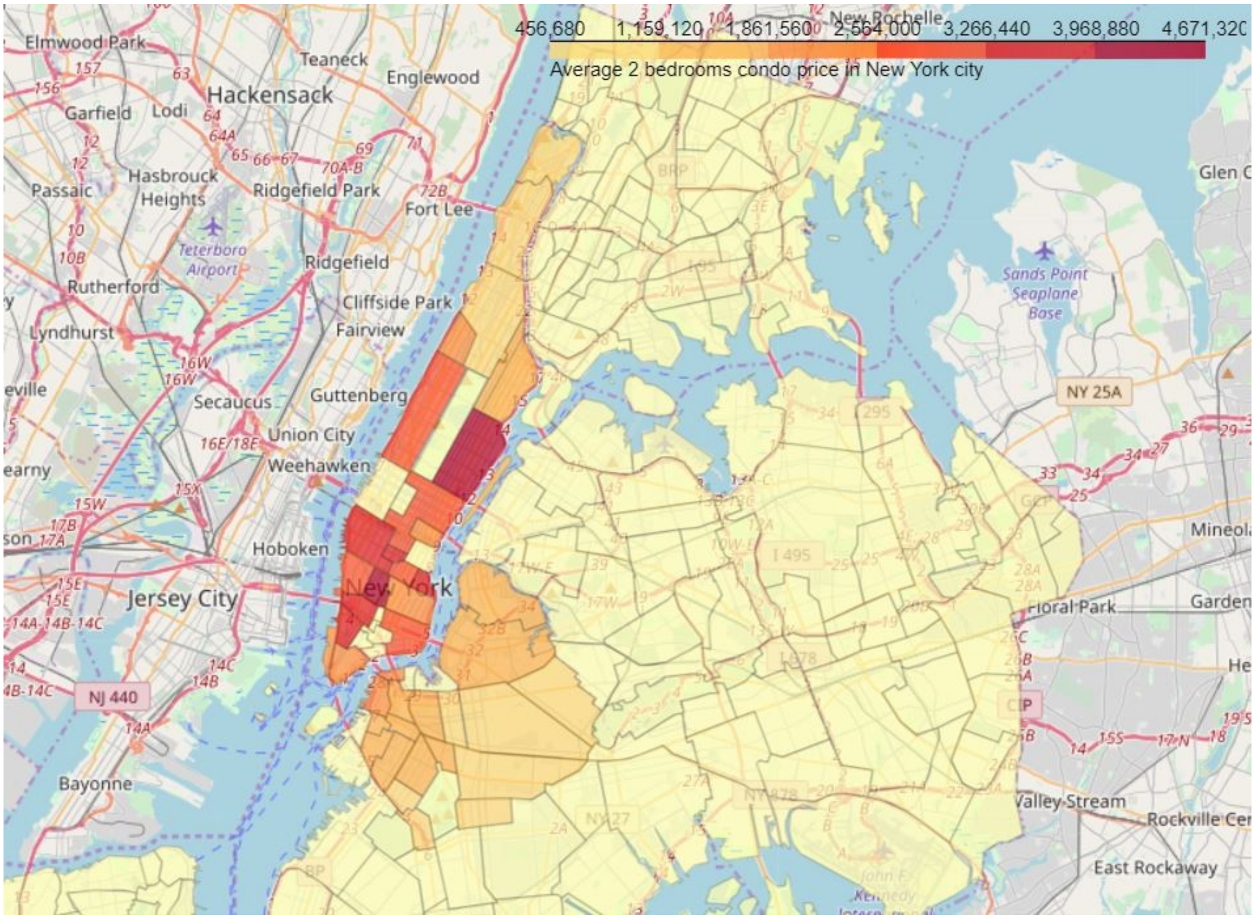


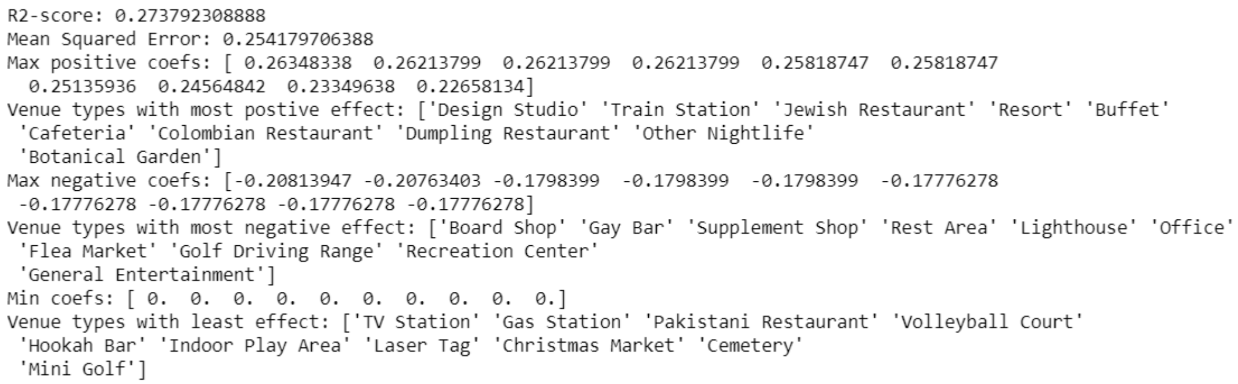
Figure 2 - New York city real estate price spread between neighborhoods

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: 0.273792308888
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']
```

Figure 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.

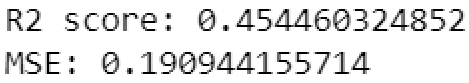
3. Principal Component Regression (PCR):

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression. [CITATION Wik \l 1033]

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.

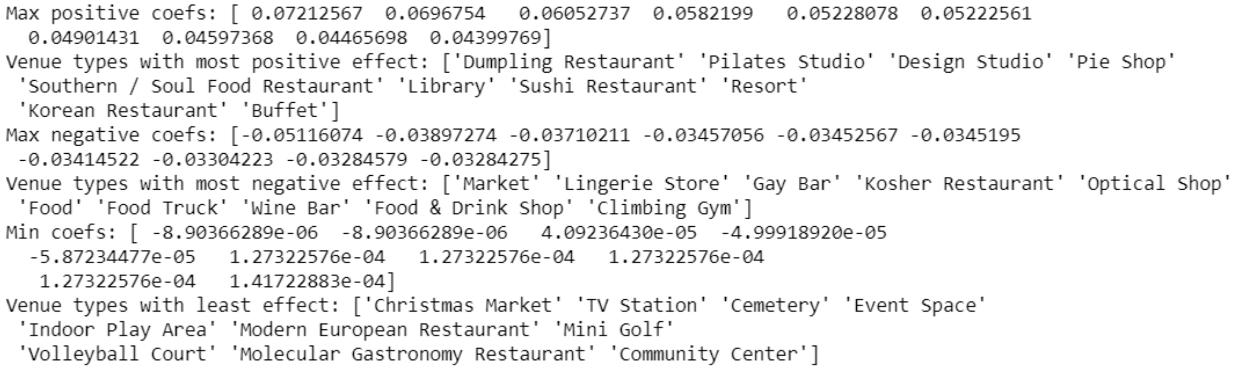


```
R2 score: 0.454460324852
MSE: 0.190944155714
```

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:

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 are 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:

A larger dataset would be needed for analysis to produce a precise model that may show stronger coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

References:

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

Table of Figures: