IBM – COURSERA

Data Science Specialization

Capstone Project – Final Report

Name: Loi Dinh

Contents

I. Introduction:

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. We all know that Airbnb, Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking.

This paper was designed following the requirement of the project - using Foursquare API. But overall, the problem and the analysis approach are decided by own learner.

In this paper, the main goal is to explore the neighborhoods of New York City by using Foursquare location data. After that, the dataset from New York Airbnb is used to extract the correlation between the price of real estate and the venue surrounding it.

The question is when you travel or arrive New York, is this/that place worth with your money. You will have the "convenience" of the location, where you will stay, and you will do whatever you like without worrying about the utilities and services around you. In particular, for the price you think it is worth it.

So, you will think about the problems, that is whether the surrounding venues affect the price to rent a house. If so, what types of venues have the most affect, and which way it will go, positive or negative?

Finally, what kinds of stakeholders for this paper:

- Visitors can roughly estimate the value of a house based on the surrounding venues and the average price.

- Owner of the real estate who can decide what price is appropriate for their products, in order to match the capability and demand of customers.

II. Data preparing and cleaning:

New York City neighborhoods were collected as the observation and it meet criterias:

- The availability of the rental room prices, the coordinates are spread throughout New York, based on data provided on Airbnb and the frequency of reviews.
- The diversity of prices between neighborhoods, but we can see the prices of room types in a neighborhood do not differ much.
- The availability of data from dataset which can be used to visualize the dataset onto a map.

The type of room to be considered is “Entire home/apt”, as data show the difference between the type of room in one neighborhood. And the most popular type of room available in the dataset is the Entire home/apartment with 25,409 rooms available on Airbnb.

The dataset will be prepared and downloaded from two main sources:

- Kaggle provides some datasets from users, which are prepared advance. Here, the dataset were prepared from Airbnb public database.

  https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

- FourSquare API which provides the surrounding venues of a given coordinates.

The process of collecting and cleaning data:

- Airbnb New York city is available, so I will use “neighborhood”, “type_room” and “price” columns to create a list of neighborhood for analyzing.
- Find the geographic data of the neighborhoods.
- 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. Here, a radius of 3 kilometers was chosen because of the purpose of the rental room. The paper assumes that visitors/travelers rent the room for sightseeing and explore the place they come, and they will choose the optimal place to stay. But overall, a radius of 3 kilometers is not an optimal choice, and lack of sample checking procedure. It is the shortcomings of this project since requests which send to FourSquare API to return a list of surrounding venues are limited.
- 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 of dataset is a 2-dimension 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.

```
(198, 387)
```

	neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	...	Yoga Studio	Zoo	Zoo Exhibit	standardizedavgprice
0	Allerton	100	100	100	...	100	100	100	-0.536540
1	Arden Heights	100	100	100	...	100	100	100	-1.076476
2	Arrochar	100	100	100	...	100	100	100	0.227976
3	Arverne	53	53	53	...	53	53	53	0.628662
4	Astoria	100	100	100	...	100	100	100	-0.316986

Figure 1

The dataset has 198 samples and more than 350 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 the rental house 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.

In 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 rental house 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/garantua/data_science/blob/master/Capstone%20Project%20-%20The%20Battle%20of%20Neighborhoods%20(Week%202).ipynb

1. First insight using visualization:

In order to have a first insight of New York City rental house 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 Randall Manor, Sea Gate, Tribeca and Woodrow. The price reduces further toward Queens 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.

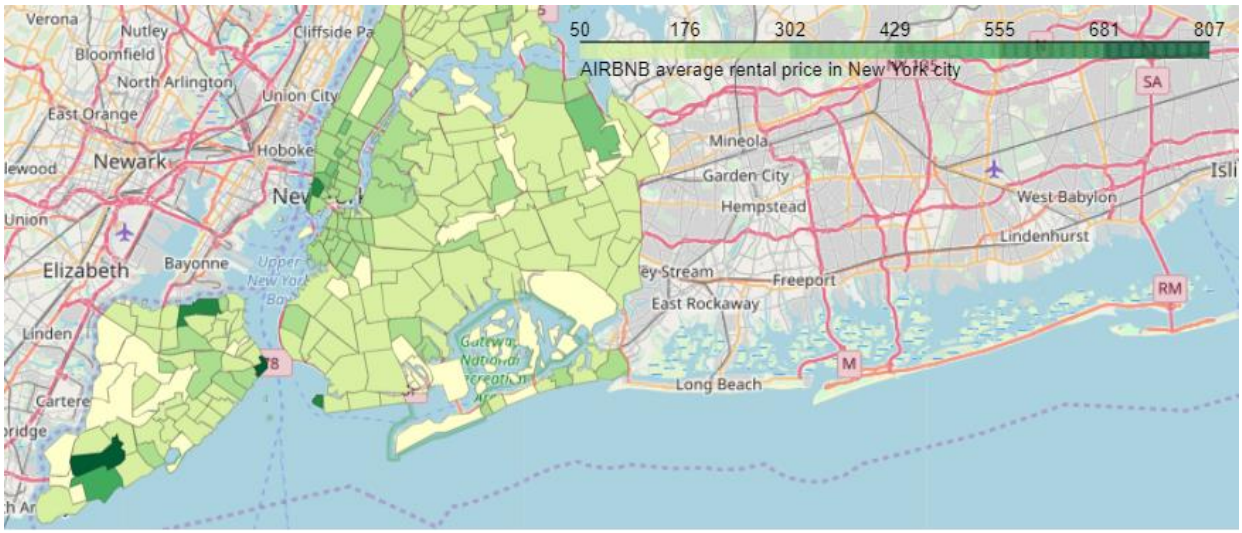

Figure 2

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 negative, which means the model can be arbitrarily worse, and may not be suitable for the data.

```
R2-score: -0.6086489461095095
Mean Squared Error: 0.7764183634103509
Max positive coefs: [ 1.44578633e+11  1.08323604e+11  6.49462195e+10  6.31917700e+10
  3.78704492e+08  3.78704492e+08  3.78704492e+08  3.78704492e+08
  3.78704492e+08 -6.68264262e+07]
Venue types with most postive effect: ['American Restaurant' 'Adult Boutique' 'Airport Lounge'
 'Afghan Restaurant' 'Zoo Exhibit' 'Yoga Studio' 'Yemeni Restaurant'
 "Women's Store" 'Zoo' 'Arepa Restaurant']
Max negative coefs: [-1.75955697e+11 -9.36070680e+10 -6.11345052e+10 -2.73770481e+10
 -6.68264262e+07 -6.68264262e+07 -6.68264262e+07 -6.68264262e+07
 -6.68264262e+07 -6.68264262e+07]
Venue types with most negative effect: ['Airport Food Court' 'Accessories Store' 'African Restaurant'
 'Airport Service' 'Argentinian Restaurant' 'Arcade' 'Arts & Crafts Store'
 'Pet Store' 'Pet Service' 'Pet Café']
Min coefs: [-66826426.16344266 -66826426.16414401 -66826426.16419604
 -66826426.16424242 -66826426.16432274 -66826426.16435441
 -66826426.16435441 -66826426.16435441 -66826426.16435441
 -66826426.16435441]
Venue types with least effect: ['Arepa Restaurant' 'Art Gallery' 'Animal Shelter' 'Art Museum' 'Aquarium'
 'Persian Restaurant' 'Performing Arts Venue' 'Park'
 'Paper / Office Supplies Store' 'Pakistani Restaurant']
```

Figure 3

On the other side, the coefficient list shows some conflict information:

- "Restaurant" and "Zoo Exhibit" both mean businesses for sightseeing. "Airport Lounge" means ease of transportation. All of which usually increase the value of a location.

- "Pet Store", "Pet Service" and "Pet Café" surely are not nice to visit. "But there is still a conflict at the airport and the restaurant. So this really is worth pondering.

- "Gallery", "Museum", "Aquarium", "Park" must be valuable with its surrounding location for tourists.

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 198 samples, and more than 350 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: 0.015303921704279522
MSE: 0.4752659798248535
```

Figure 4

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.

The insight is not 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 rental average price.

Explanations for the poor model can be:

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

V. Discussion:

The real challenge is constructing the dataset-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.

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.

In some ways, there is no correlation between the rental house average price and the surrounding venues. Maybe in a certain paper, we will learn about the correlation between the rental price and the price of the real estate itself. And have a conclusion on these things.