IBM – COURSERA

Data Science Specialization

Capstone Project – Final Report

Name: Loi Dinh

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 real estate 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.