Battle of Neighborhoods:

New York City & Airbnb Listings

Luz Ortega

Applied Data Science Capstone Project Report

IBM Data Science Professional Certificate Specialization

January 2019

Note

This project report is for the Applied Data Science Capstone Course, a course part of the IBM Data Science Professional Certificate Specialization, where students learn about location data, location data providers, as well as Python libraries to manipulate, explore, analyze, and visualize data. The problem and analysis approach for this project can be chosen by the student, however, it is required to leverage the Foursquare location data to solve the problem.

Introduction

Nowadays, with the increasing smart devices penetration worldwide, location data (i.e., geographical position in terms of latitude and longitude) have become easier to obtain and it is used by multiple companies/apps to build better mobile experiences. Examples of such companies are: UBER, Lyft, Starbucks, Mindbody, Facebook, just to name a few. Another company that uses location data is Airbnb, a community-driven hospitality company, which lets people rent out their properties or spare room to guests. In other words, Airbnb offers their users someone's home to stay, instead of the conventional lodging option of a hotel. One of the most popular destination in the world is New York City. In March of 2018, Mayor Bill de Blasio announced that the city welcomed an estimated 62.8 million visitors in 2017, 49.7 million corresponded to domestic visitors, while 13.1 million corresponded to international visitors. It is known that New York City has one of the most active hotel development pipeline in the United States. According to the announcement, even with the 4,000 new rooms added to the city's inventory in 2017 (bringing the total to about 116,500 rooms) room's demand remained strong. In 2017, New York City sold record 36.4 million hotel room nights. Development of the tourist sector, especially in the topic of accommodations, remains important for the big apple.

The main goal of this project is to explore and analyze data of New York City neighborhoods and Inside Airbnb to recommend Airbnb locations based on the surrounding venues of the rental places and listing data. There are three main stakeholders interested in this type of information:

- Travelers/Renters: Knowing the category of the surrounding venues of Airbnb locations can help travelers decide where to stay. For example, if a traveler is interested in Broadway shows, that person might want to stay in a neighborhood where the surrounding venues include theaters and perhaps bars, in case he/she wants to enjoy some cocktails before or after the show. On the other hand, if a traveler is into fitness and exercising, that person might want to stay in a location surrounded by common venues such as gyms, parks, juice bars, or yoga studios.
- Home-owners interested or already renting out their properties: It is important to identify and profile your potential travelers. By knowing what type of travelers might be interested in certain locations, home-owners can create and improve marketing strategies (e.g., pictures, recommended locations nearby the rental place, rental fees etc.) to promote their property. This information can also help home-owners maximize their vacation rental listing description to make the most impact on guests.
- Real estate investors: People that want to invest in short-term vacation rentals need to know what places are profitable, how many listings are in certain neighborhood, what is the average night fee for properties in the different neighborhoods, what are the common surrounding venues of a potential place, etc.

Note that both home-owners and real estate investors are probably interested in knowing how many listings are in specific neighborhoods, where are the listing located, what are the popular surrounding venues for specific listing, are the travelers short-term or long-term visitors, among many other questions.

Data

This section provides a description of the data used for this project. There are three datasets:

- **New York City Neighborhood Names**: This dataset was created as a guide to New York City's neighborhoods that appear on the web resource, "New York: A City of Neighborhoods" and it is available online at the New York University Spatial Data Repository website (https://geo.nyu.edu/catalog/nyu_2451_34572). The dataset has all 5 boroughs and 306 neighborhoods as well as the latitude and longitude coordinates of each neighborhood. This dataset will be used to plot the neighborhoods where Airbnb listings are being offered, i.e., this dataset along with geographical, plotting, and mapping libraries are used to visualize the neighborhoods on a map as shown in Figure 1. This dataset is a collection of nested Python dictionaries. This data will be transformed into a *pandas* dataframe, with the following columns: 'Borough', 'Neighborhood', 'Latitude', and 'Longitude'. The resulting dataframe has 306 rows and 4 columns.

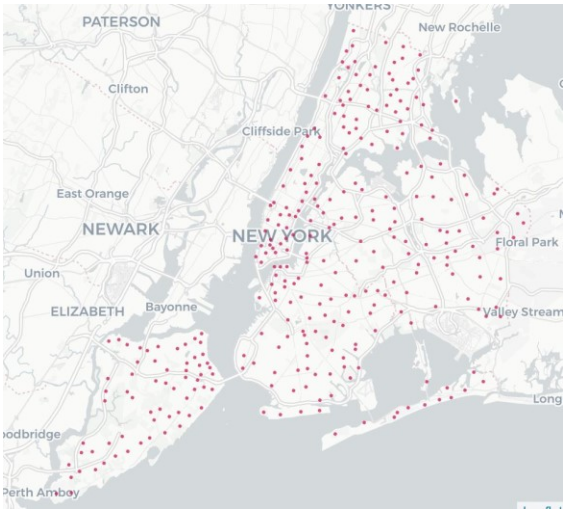

Figure 1. Neighborhoods of NYC

- **Inside Airbnb**: Inside Airbnb is an independent, non-commercial, open source data tool that allows users to explore Airbnb data. The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The dataset for this project corresponds to Airbnb listings of New York City and it was downloaded from http://insideairbnb.com/get-the-data.html and compiled on December 06, 2018. The keys of the dataframe (i.e., columns) include listing ID, neighborhood, borough, price, rating scores, geographical coordinates of the listings, room type, and minimum nights, among others. For the purpose of this project, only listings meeting these requirements were considered:
    - Listings must accommodate minimum two people
    - Room type must be 'Entire home/apt'
    - No minimum nights requirements, i.e., guests can book the rental place for just one night.

  After cleaning the data, the final dataset looks like the one in Figure 2. Data exploration using this dataset let the users find out information such as, what neighborhoods have more listing, what borough is the most expensive, what is the average price for listings in an specific neighborhood, or what listings have the highest or least rating scores?. It is important to note that this dataset doesn't provide geographical coordinates for each neighborhood, only provides them for the listings. However, the neighborhood coordinates can be obtained from the New York City Neighborhood Names dataset explained above in case users want to visualize what neighborhoods offer Airbnb rental places.

	ID	Neighborhood	Borough	Latitude	Longitude	Price	RS_rating	RS_accuracy	RS_cleanliness	RS_checkin	RS_communication	RS_location	RS_value
0	2595	Midtown	Manhattan	40.753621	-73.983774	225	95.0	9.0	9.0	10.0	10.0	10.0	9.0
1	3831	Clinton Hill	Brooklyn	40.685138	-73.959757	89	90.0	9.0	9.0	10.0	10.0	9.0	9.0
2	5238	Chinatown	Manhattan	40.713444	-73.990375	150	93.0	9.0	9.0	10.0	10.0	9.0	9.0
3	16595	Williamsburg	Brooklyn	40.709330	-73.967918	270	93.0	10.0	9.0	10.0	10.0	9.0	9.0

Figure 2. Dataframe created using the Data from InsideAirbnb

- Foursquare Location Data: Foursquare API is used to explore the Airbnb listings and segment them. Foursquare data can provide the trending surrounding venues of a given pair of coordinates, in this case the pair of coordinates corresponds to an Airbnb rental place. This data will be used to create a dataframe, where each row represents a rental place, and most of the columns represent the occurrence of a venue type nearby such rental place. This dataframe can give us insights about the location of the rental place, including trending venues, type of venues, or even what listings share common nearby places. Note that to use this data, it is necessary to create a Foursquare developers account. Once the account is created, you will obtain a Client ID and a Client secret, these credentials are needed to make data requests.

Methodology

This section describes the exploratory analysis, statistical analysis, and machine learning techniques used for this project. This project was developed using Jupyter Notebook and Python as the programming language. The following Python packages/libraries were used:

- NumPy: Library to handle data in a vectorized manner
- Pandas: Library for data analysis
- JSON: Library to handle JSON files
- Geopy: Library to obtain geographical coordinates
- Requests: Library to handle requests
- Matplotlib: Plotting library
- Scipy: Library for scientific computing
- Sklearn: Machine learning library
- Folium: Map rendering library

Exploratory Analysis

The project implementation started with data acquisition, which is the process of loading and reading data into Python. For this project, there were three main sources of data, as mentioned in the Data Section.

- The first dataset did not need any data preprocessing, besides transforming the dataset (which was a collection of nested Python dictionaries) into a pandas dataframe. As part of the dataset exploratory analysis, the author used the groupby function on this dataframe to determine how many neighborhoods are in each borough, as shown in Figure 3. The results were also plotted as shown in Figure 4. To visualize the neighborhoods of this dataset, a map of NYC was created with neighborhoods superimposed on top as shown in Figure 5.

```
# We can use groupby to check how many neighborhoods are in each borough
neighborhoods.groupby('Borough').size()
```

```
Borough
Bronx            52
Brooklyn         70
Manhattan        40
Queens           81
Staten Island    63
dtype: int64
```

Figure 3. Groupby Function: Number of Neighborhoods per Borough

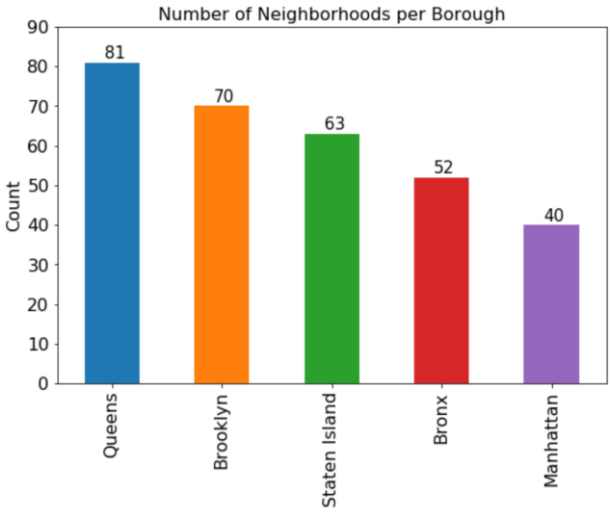

Figure 4.Plot: Number of Neighborhoods per Borough

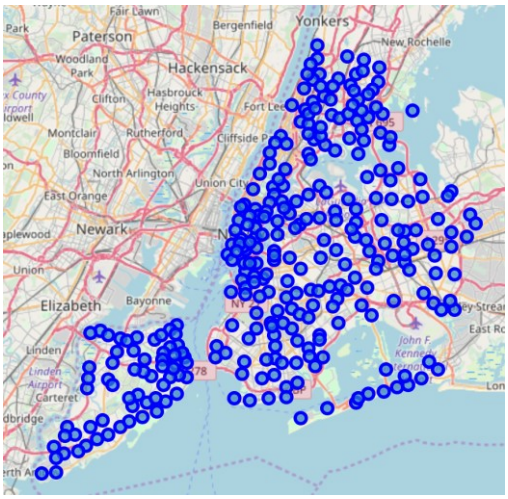

Figure 5. Neighborhoods of NYC

- The Inside Airbnb data needed to be preprocessed. Some of the data cleaning process' tasks were: identification and handling of missing values, data formatting, among others. After the dataframe was cleaned, the groupby method was used to determine the amount of listings in each borough as shown in Figure 6. The unique function was used to determine the number of neighborhoods where there were listings. There were listings in 166 neighborhoods. The number of neighborhoods where there are listings is shown in Figure 7.

```
print(listings.groupby('Borough').size())
```

```
Borough
Bronx              65
Brooklyn         1181
Manhattan        1922
Queens            494
Staten Island      22
dtype: int64
```

Figure 6. Number of listings per Borough

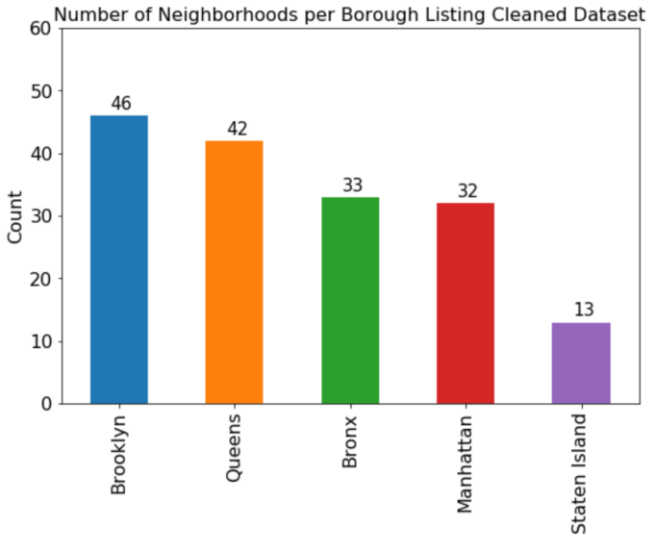

Figure 7. Number of Neighborhoods where there are listings

- Note that the Inside Airbnb dataset has the latitude and longitude for each listing. However, it doesn't provide the coordinates for each neighborhood, so in order to visualize in a map the

neighborhoods contained in this dataset the author needed to verify the names to match the NYC Map dataset and then create new dataframe merging information from both datasets of the listings dataset, there were 21 neighborhoods that required to be modified since they were not in the NYC Map dataset. After this cleaning was performed, the neighborhoods where there are rental place were mapped as shown in Figure 8.

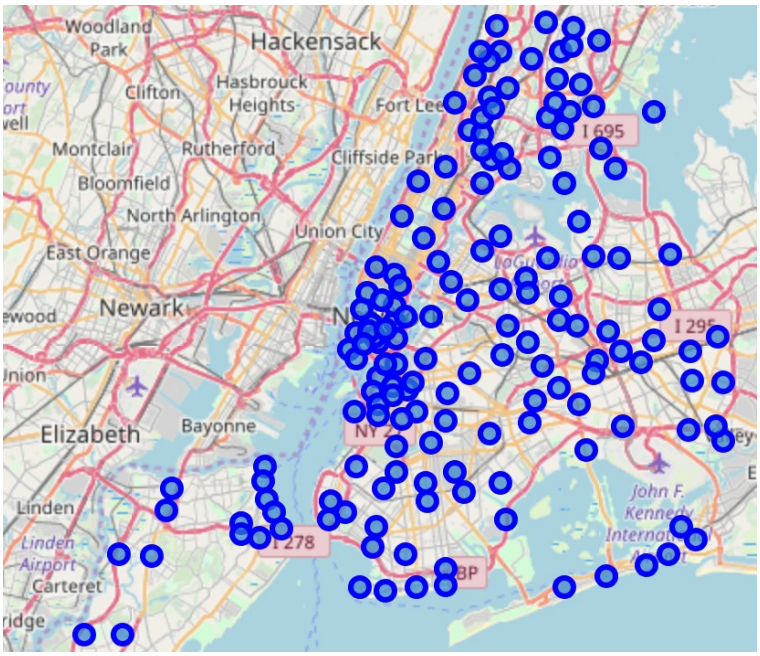

Figure 8. Neighborhoods where listings are located.

- For the purpose of this project, only the listings located in Staten Island were considered for further analysis, segmentation, and clustering. Note that the same analysis can be applied to the other boroughs. Based on the exploratory analysis of the listings dataframe, it was already known that there were 22 listings and 13 neighborhoods in the borough of Staten Island. Those neighborhoods where the listings were available were replotted, this time only including the ones of Staten Island as shown in Figure 9. The listings were replotted too, but this time only the ones of Staten Island. As shown in Figure 10.

Figure 9. Staten Island Neighborhoods

Figure 10. Listings in Staten Island

- The Foursquare data was used to obtain the most common venues nearby the listings. A dataframe was created including the listings, listings' coordinates, venues, venues' coordinates, and venue category. The groupby method along with the count function were used on this dataframe to check how many venues were returned for each listing/rental place.

Statistics

As part of the statistical analysis, we used the mean function to determine which borough has the highest/lowest listing average price as shown in Figure 9. Then an analysis of variance, better known as ANOVA, was performed to test whether there was a significant difference between

the means of the prices for the different boroughs. ANOVA results are presented in the next section. Descriptive statistics were also obtained by using the describe method on the listings dataframe. Results of this method are shown in the Results section.

	Borough	Price
0	Bronx	105.707692
1	Brooklyn	197.159187
2	Manhattan	235.511967
3	Queens	134.125506
4	Staten Island	106.454545

Figure 11. Average Price of listings per borough

Machine Learning

- In order to implement machine learning techniques, a one hot encoded was performed. One hot encoding is a process by which categorical variables, in this case the venue types, are converted into a form that could be provided to machine learning algorithms to do a better job in prediction. After the hot encoding, the 5 most common venues were printed.
- Clustering is an unsupervised machine learning technique that can group data based on data similarity, for this project we performed a listing segmentation, i.e., partition each listing into groups of listings that have similar characteristics. In this case the similar characteristics are based on the surrounding venues.
    - K-means is a partitioned based clustering technique that can group data only unsupervised based on the similarity of instances to each other. This technique divides the data into k non-overlapping subsets, i.e., clusters, without any cluster internal structure. The objects within a cluster are very similar. For this project we created clusters to group similar listings.

Results

- ANOVA returns two parameters:
    - F-test score: ANOVA assumes the means of all groups (in this case each borough), calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A large score means there is a larger difference between the means.
    - P-value: this parameter tells the statistical significance of the calculated score value.

```
# from scipy import stats
borough_anova=listings[['Borough','Price']].groupby(['Borough'])
f_val, p_val = stats.f_oneway(borough_anova.get_group('Brooklyn')['Pr
print( "ANOVA results: F=", f_val, ", P =", p_val)
```

```
ANOVA results: F= 11.440862923819846 , P = 3.1391644920354877e-09
```

Figure 12. Anova Results

The ANOVA results show that the average price is significantly different depending on the Borough.

- Some of the insights from the descriptive statistics of the listings dataframe shown in Figure 13 are:
    - The mean price of the listings is $206.56
    - The standard deviation of the prices is very large indicating that the data points are spread far from the mean. This makes sense when having a listing with a price of $10,000 and one with a price of $10.00

	Latitude	Longitude	Price	RS_rating	RS_accuracy	RS_cleanliness	RS_checkin	RS_communication	RS_location	RS_value
count	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000
mean	40.730767	-73.952822	206.560803	93.519815	9.577090	9.293974	9.717155	9.755429	9.554832	9.328719
std	0.053310	0.052056	338.631421	7.445258	0.777624	0.982970	0.658963	0.610489	0.714446	0.788342
min	40.532648	-74.210166	10.000000	20.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
25%	40.693197	-73.987185	105.000000	91.000000	9.000000	9.000000	10.000000	10.000000	9.000000	9.000000
50%	40.729379	-73.960638	150.000000	95.000000	10.000000	10.000000	10.000000	10.000000	10.000000	9.000000
75%	40.763832	-73.939648	218.250000	100.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
max	40.903895	-73.728375	10000.000000	100.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000

Figure 13. Descriptive statistics of the listings dataframe

- Describe method of the categorical variables, i.e., Neighborhood and Borough

	ID	Neighborhood	Borough
count	3684	3684	3684
unique	3684	166	5
top	16426099	Williamsburg	Manhattan
freq	1	236	1922

Figure 14. Describe method of the categorical variables of listings dataframe

- Further exploration was done only for the listings located in Staten Island. The total of unique venue categories was 114. There were listings with as few as two venues nearby and as much as 46. A dataframe of the listings and 3 most common venues was created. The describe method was used on this dataframe. The 1st most common venues surrounding listings are pizza place and bus stop surrounding 4 listings, sandwich places are surround 2 listings, as shown in the figure below.

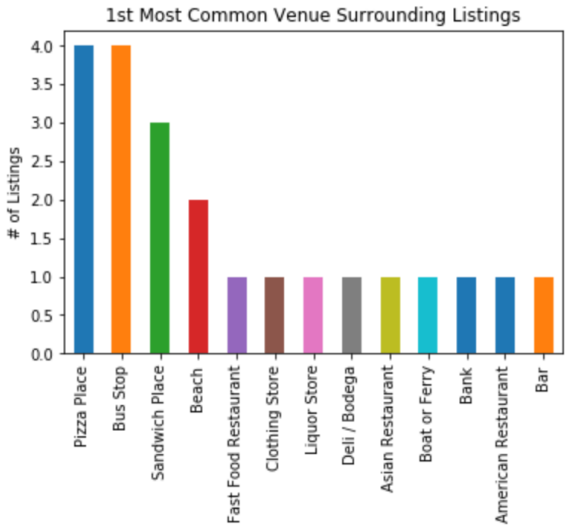

Figure 15. 1st Most Common Venues Surrounding Listing in Staten Island

- Cluster Listings: The author ran k-means to cluster the listings into 5 clusters. The resulting clusters are shown in figure 13. 15 of the 22 listings were assigned to cluster 1, 2 listings were assigned to cluster 2, 1 listing was assigned to cluster 3, 3 listings we assigned to cluster 4, and 1 listing was assigned to cluster 5. Most of the listings of cluster 1 have 'Bus Stop', 'Pizza Place' or 'Sandwich Place' as the 1st most common surrounding venue.

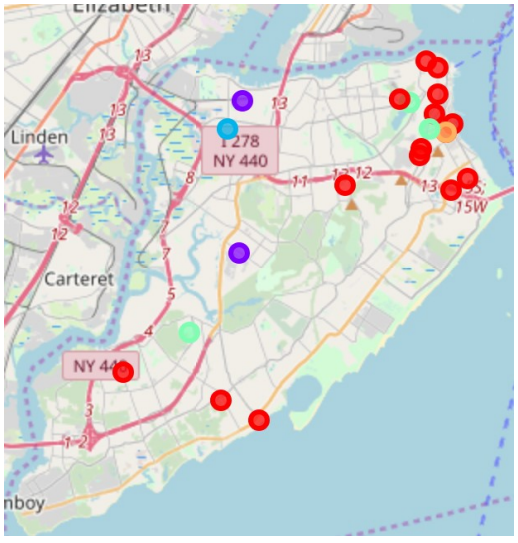

Figure 16. Clusters of Listings in Staten Island

Discussion

- The most consuming part of this project was the construction of the Inside Airbnb final dataframe. When merging or combining information from multiple datasets, usually one will find inconsistencies, such as different formats, different names, just to name a few.
- Foursquare data is obtained through API calls, therefore different results will be obtained if the code is run at different times. The results presented in this project were obtained at 5:20 pm ET. It might be interesting to expand this project and analyze the trending venues at different times during the day. Another possibility is to compare the trending venues during weekdays compared to those during the weekends.
- K-means is a good algorithm for medium to large datasets. There is not a right answer for what the minimum sample size to conduct cluster analysis is. However, the number of listings for Staten Island is only 22 and it seems is not large. The borough of Staten Island was chosen by the author for further exploration because of a personal linkage.
- The ANOVA results show there is a statistical significance difference among the mean prices of the listings in the different boroughs. Further analysis can be performed and test the difference between each pair of boroughs, e.g., is there a significance different in the mean price of the listings in Queens compared to Brooklyn?
- This project can be expanded and apply the same approach to the other 4 boroughs of the city and compare what the trending venues are in each borough.

Conclusion

This capstone project is a good way to end the IBM Data Science Certification. The development of this project forced the author to apply multiple data science concepts and techniques learned throughout the courses of the specialization, including but not limited to the following: data science methodology, data visualization, data wrangling, data analysis, machine learning, and Python (including a vast variety of libraries and packages).

Overall, this project covers the process of coming up with a problem or idea, using data to understand it and get insights of that particular idea, in this case rental places in NYC. Exploratory analysis, statistical analysis, as well as machine learning techniques were used for this project. Despite the fact that the borough that was chosen for clustering segmentation was not large enough, the project still provides a meaningful and useful approach to perform segmentation (clustering) of rental places based on trending surrounding venues.