Analyzing venues nearby the Airbnb listings in Staten Island, N.Y.

Oscar Aguilar
oe.agur@gmail.com

July 2020

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Since Decemeber 2019, the world has encountered several challenges in the healthcare sector due to the outburst of a novel virus: COVID-19. Furthermore, the impact of this situation has reached several major fields with economic, political and social implications; for instance, tourism. Since the commencement of the lockdown in many countries, different travel restrictions have emerged (locally, regionally and internationally). The latter has affected SMEs and globally recognized companies such as Airbnb (Gössling, 2020).

In late June 2020, Brian Chesky (CEO of Airbnb) publicly expressed: "It took us 12 years to build Airbnb, and we lost almost everything in four to six weeks" (Entrepreneur, 2020). He added that the future of the company is uncertain since the future of travelling will not be the same until there is a vaccine that can fight this pandemic. The company has published drastic statistics using cancellations and availability information of their listings worldwide: after the Wuhan lockdown, their booking activities showed a fall of 57.8% (Hu and Lee, 2020), therefore the company laid off 25% of their employees and raises a debt equivalent to $2 billion in equities (Pavlovska, 2020)

Nowadays, the governments have started to reactivate the economy of their respective countries. Society has to learn to live a new normality with extreme precautions. Chesky said that since people is not getting on airplanes (or crossing borders), there is an urge to promote local travelling. The latter because people are showing a clear preference to travel by car if required. Hence, with some precautions in mind, people could start making small trips to near-local areas in order to support small businesses and the accommodation industry. Some precautions Chesky suggested were asking the owner of the property: when the listing was rented, what is being done to clean it, if it is being rented continuously, and local-or-regional restrictions (Krstic, 2020). In addition, it is preferable to wear masks upon arrival, bring sanitizer and disinfectants, as well as an air purifier, to mention a few.

1.2 Problem

Considering New York, N.Y. as one of the most preferred and concurred cities by tourists in the world, it will be of paramount importance to analyze the venues offer of a smaller-and-near local area such as Staten Island. Using the location of Staten Island's Airbnb listings, it will be possible to find the kind of venues around them. In this way, local people near the zone from will be able to see which kind of businesses and places are around a potential Airbnb accommodation. This will benefit local tourists, Airbnb, listings' owners and local businesses.

2 Data Acquisition and Preprocessing

2.1 Data Sources Description

In order to perform this analysis, two different sources of data will be utilized. In this section of the report, each data set will be described thoroughly. The description includes their respective source, features, and origin.

2.1.1 New York's Airbnb Data Set

This first data set was extracted from Kaggle as a csv file. The data retrieval process started in 2008, with the aid of guests and hosts of Airbnb from N.Y. They contributed by providing information that describes the listings' generalities, availability and metrics. This continued until 2019, and three different version of the data set have been built. This project utilizes the most updated version. The included features are listed in the table below:

Table 1.N.Y. Airbnb Data Set Feature Description

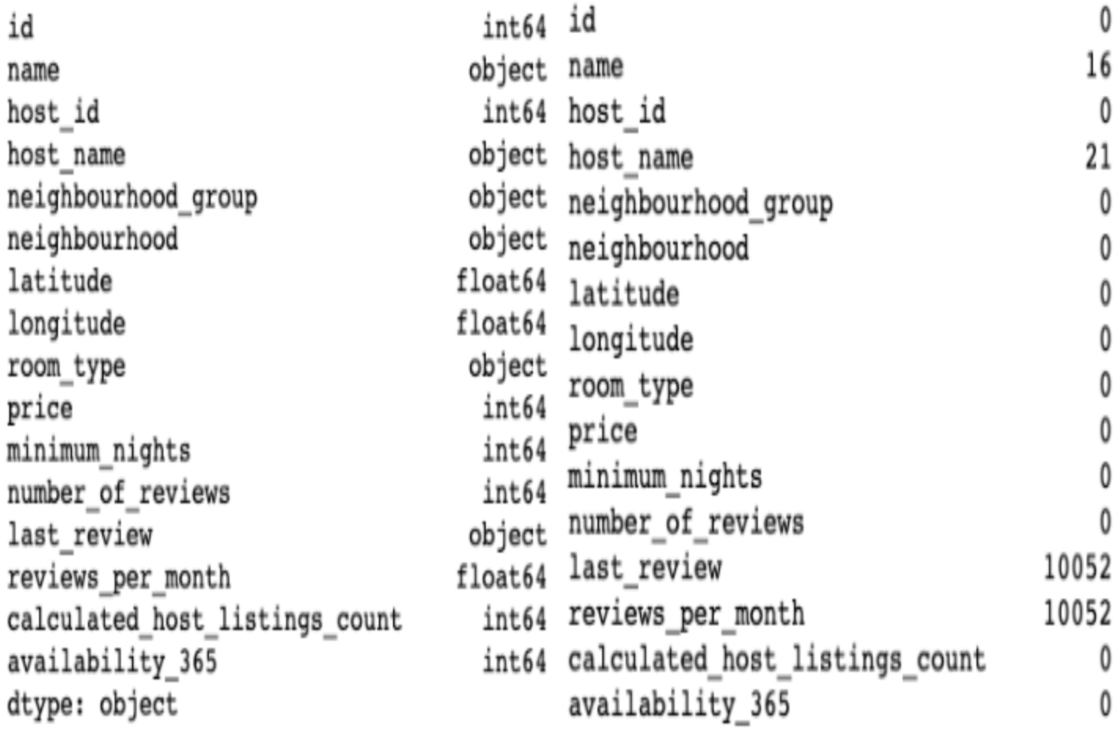
Feature	Type	Description
Id	Integer	Identification number of the listing
Name	Object	Name of the listing
Host_id	Integer	Identification number of the host
Host_name	Object	Name of the host
Neighbourhood_group	Object	Boroughs of N.Y. (e.g. Manhattan, Queens, Staten Island)
Neighbourhood	Object	Neighborhoods within the boroughs
Latitude	Float	Latitude coordinate
Longitude	Float	Longitude coordinate
Room_type	Object	Type of listing: private, shared, etc.
Price	Integer	Price per night
Minimum_nights	Integer	Minimum required number of nights to book
Number_of_reviews	Integer	Total number of reviews of the listing
Last_review	Object	Date of the last review
Reviews_per_month	Float	Average number of reviews the listing gets per month
Calculated_host_listings	Integer	Number of listings that the host owns
Availability_365	Integer	Number of days/year available for booking

2.1.2 Foursquare location data

The second source of information are the location-based services offered by Foursquare which consist of a RESTful service used to request JSON or XML data. Foursquare is a free location discovery app that allows users to find and share information about businesses and attractions in any part of the world. The Foursquare API allows developers to interact directly with their data platform. The different API methods allow developers to retrieve check-ins, venues, categories, tips, menus, among other data. Since these requests require authentication, programmers are required to create an account, which can be upgraded by a monthly fee. An upgraded account has more perks regarding the number of calls that can be done per day, the ability to make premium calls, etc.

2.2 Preprocessing: Data Cleaning

After importing the csv file extracted from Kaggle, the preprocessing consisted of 3 steps: inspecting the type of variable of each feature, checking for missing values, and dropping some columns that were not necessary for the analysis. The functions dtypes, isnull and drop were used respectively. The output of the first 2 steps is shown in the figure 1. On the other hand, due to the high amount of missing values (10,052 out of 48,895) the dropped features were: last_review and reviews per month. Similarly, features that did not contribute to the analysis were deleted, such as: id, host_id, host_name, calculated_host_linsting_count and availiability_365.



```
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
```

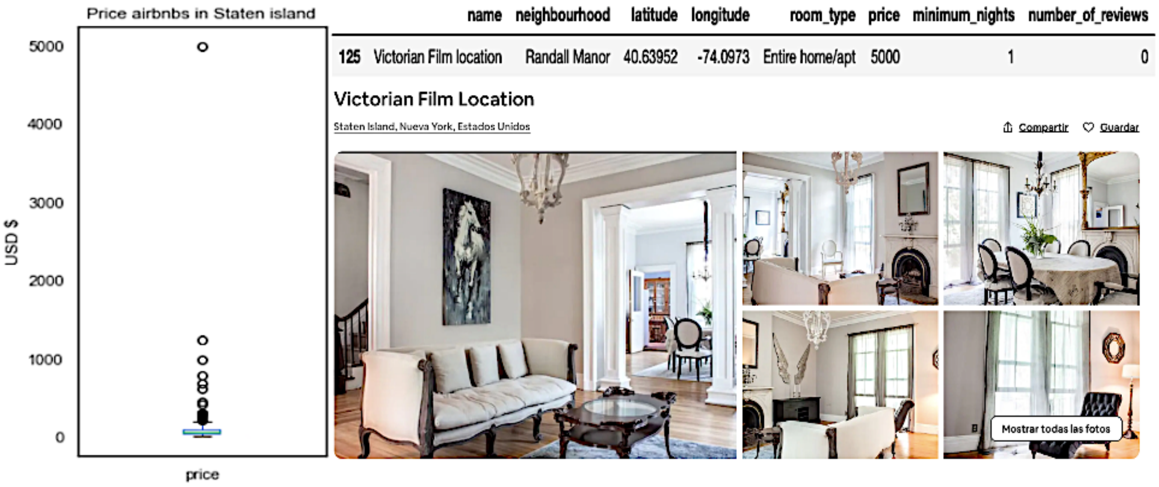
Figure 1. Data Preprocessing: features type and missing values

3 Methodology

In this section of the report, every explanatory and analytical procedure will be enlisted and described. In general terms, a sequential explanation of the followed steps is included. However, any important graphic or numerical result will be included in section 4. Basically, the analysis goes from general to particular information. Techniques such as clustering, and one hot encoding are utilized. The most important libraries that were utilized are: pandas, numpy, scipy, matplotlib, seaborn, folium, geocode, and Sklearn.

Firstly, a broad inspection of the data was made. Comparisons between the types of rooms and their mean price was performed. Similarly, the mean price of the listings per borough was obtained. After analyzing the distribution of prices and the number of listings per borough, Staten Island was the chosen local area to work with (373 listings versus +20,000 in Brooklyn and Manhattan).

Therefore, the data set was modified, and only Staten Island listings were kept. From there, 4 different maps were constructed: one with all the listings, and one per type of room (entire homes/apartments, private rooms, and shared rooms). This helped to visualize the distribution of listings in the area. Then, when the price distribution of the Staten Island's listings was obtained, a very extreme outlier was found; therefore, it was researched on the Airbnb webpage and it was successfully found (figure 2).



	name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
125	Victorian Film location	Randall Manor	40.63952	-74.0973	Entire home/apt	5000	1	0

Figure 2. Extreme outlier in Staten Island

After that, the listings were grouped by neighborhood, and the analysis showed that the neighborhood with more listings was St. George with 48. From there, the Foursquare API was used to retrieve JSON data from the venues of all the neighborhoods. The retrieved information showed that in Staten Island there were around 5,663 venues with 219 unique categories. Next, the top 10 most common venues per neighborhood were found and displayed. This allowed to retrieve the most common venue per neighborhood and construct a graph that shows the category and count of the most common venues in the island.

For the interest of the Airbnb hosts, a clustering algorithm was used to group all the listings of the island in 6 different clusters. After that, the mean price of each cluster was calculated. Similarly, the clusters were visually represented in a map of the island. This allows the hosts to find their listings on the map and compare if what they are charging per night is similar to the price of other listings within their cluster. The number of clusters was obtained with the elbow method.

4 Results

Now, the graphics and important information obtained from the project are going to be displayed in this section. To understand better the finding, the results are going to be splitted into 3 main sections:

a. Broad analysis of the N.Y. data set (including all boroughs)

b. Specific analysis of the chosen local area: Staten Island

c. Clustering of the Staten Island's listings: K-Means algorithm

4.1 Broad Analysis of N.Y. data

The N.Y. Airbnb data set originally contained 16 columns and 48,895 rows (before preprocessing). In this first analysis, two graphs were generated: on the left, the mean price of the listings per borough; and in the right, the mean price of the listings per type of room. It is clearly shown that Manhattan is where listings with the higher mean prices per night are. Also, it is obvious that the most expensive type of listings are entire homes or apartments.

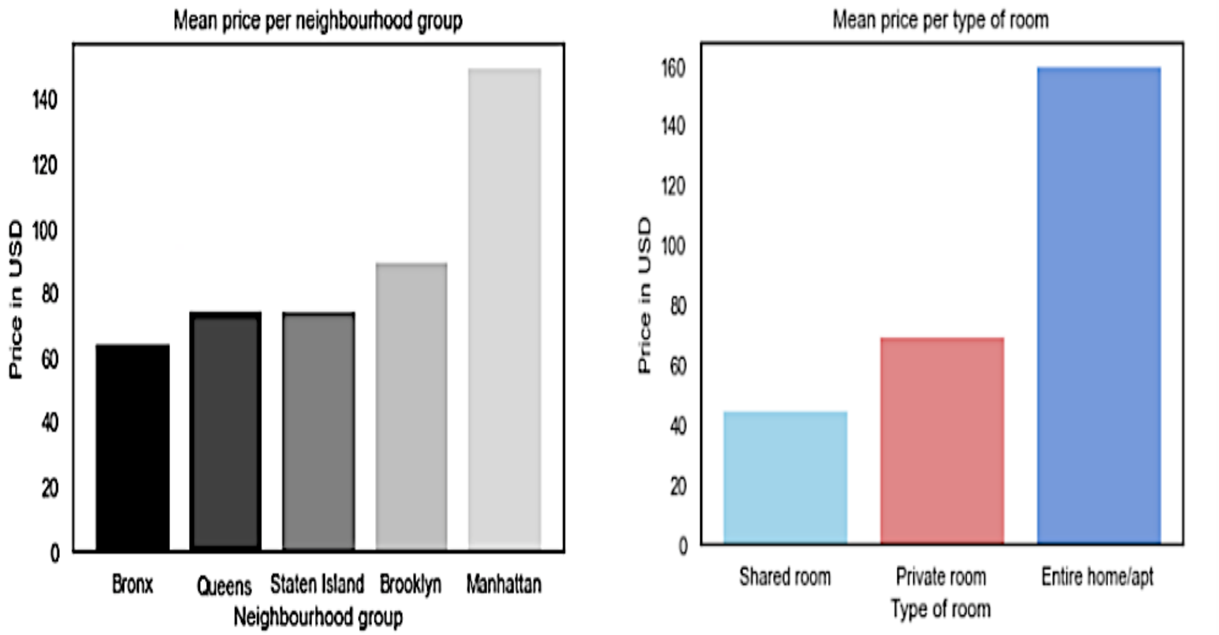


Figure 3. Mean price per borough and per type of listing

To obtain a more insightful look at the listing's prices, the same previous analysis was performed. However, instead of computing the mean prices, the distributions were plotted. In these plots it is shown that Staten Island has a higher spectrum of choses regarding price; since there are listings for less than 250 USD per night, up to 750 USD.

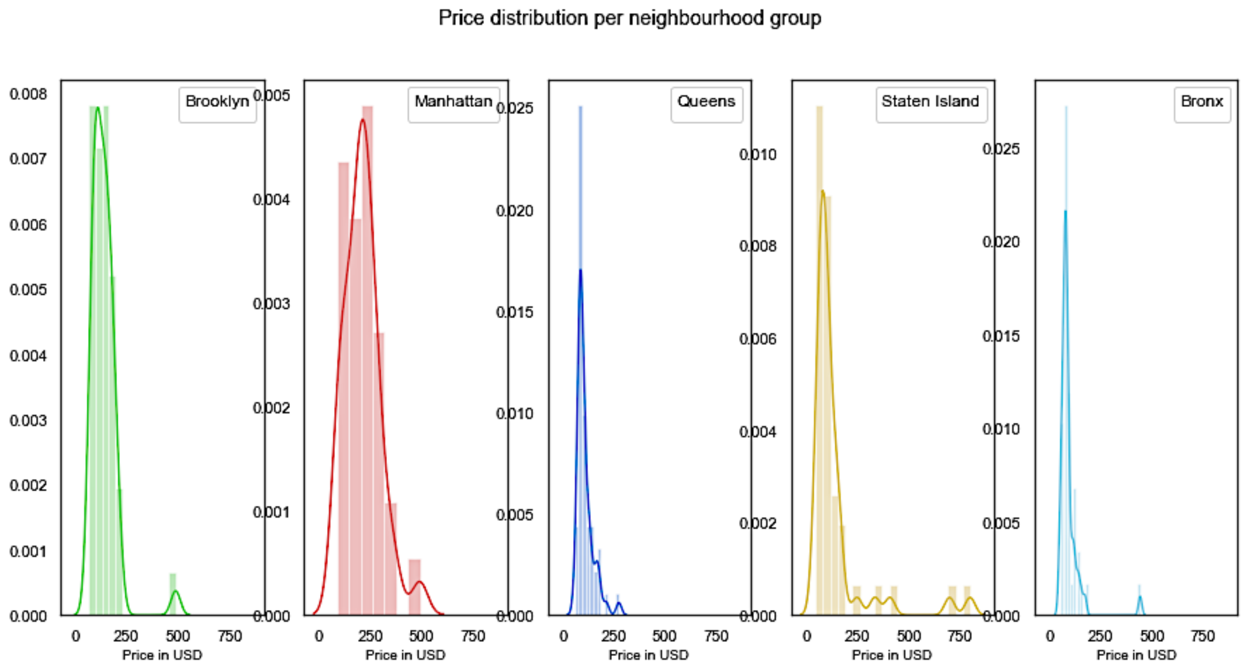


Figure 4. Price distribution by borough

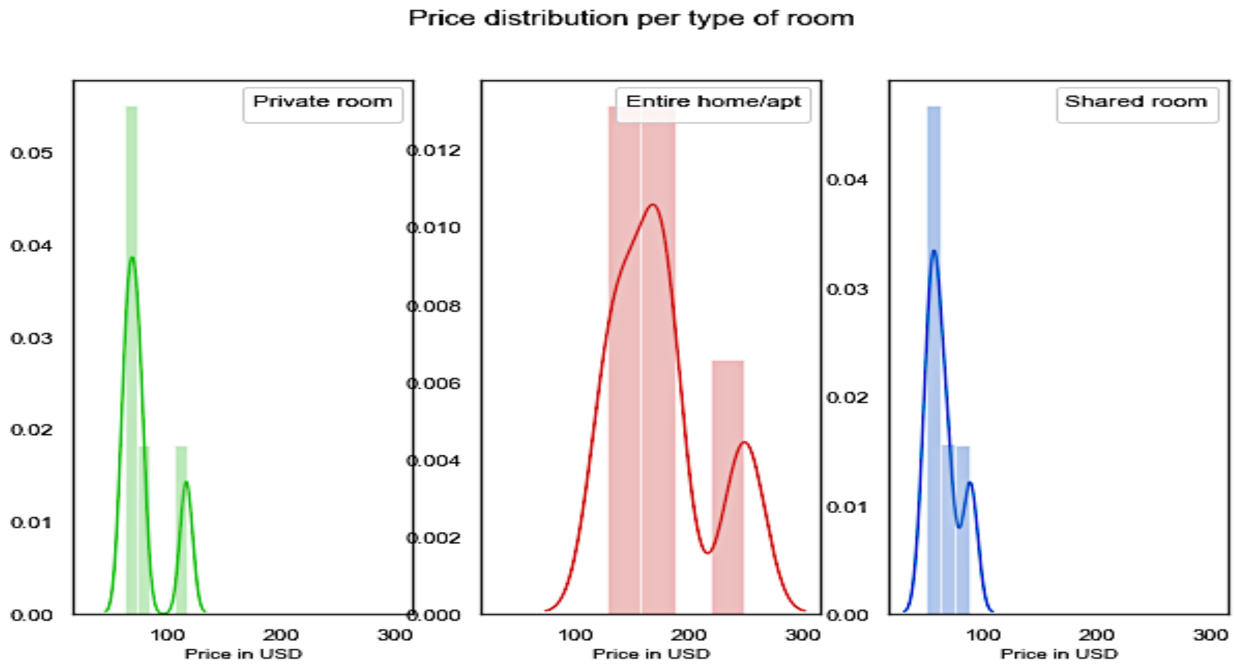


Figure 5. Price distribution by type b listing

All the information that has been established above is summarized in the following two bar charts: one that shows the prices of each type of listing classified by borough, and a second one that shows the count of each type of listing classified by borough as well.

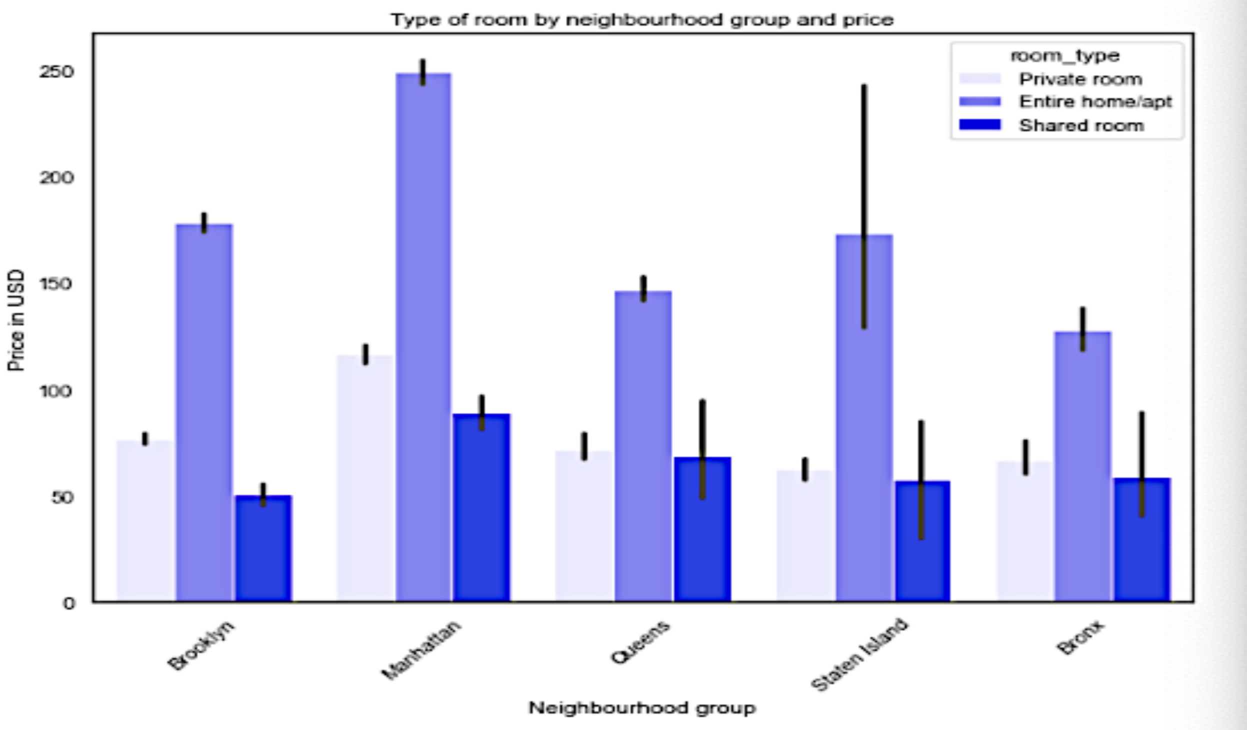


Figure 6. Price of each type of listing per borough

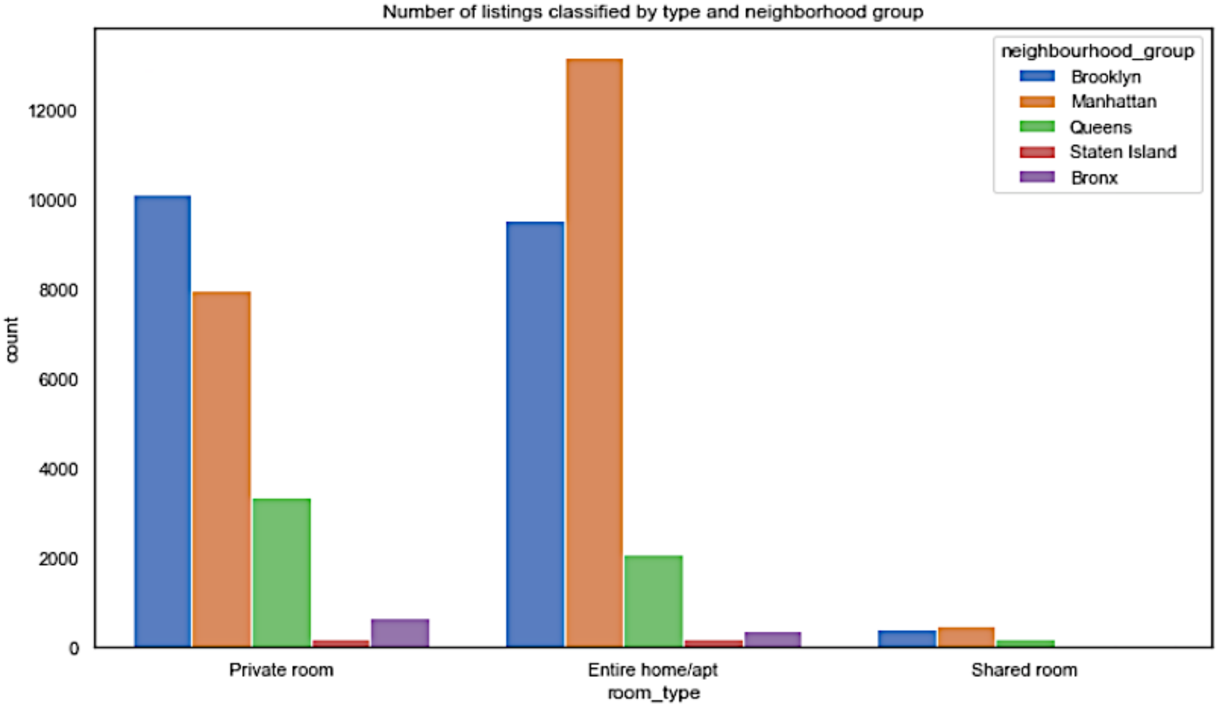


Figure 7. Count of each type of listings per borough

4.2 Staten Island’s Analysis

From figure 7, it is obvious that Staten Island has the lower number of Airbnb listings, which suggests that fits more the description of a local area. The latter is of main interest, because it is important to remember that this project is trying to find the most important venues of a local area near the big and highly visited city of NYC (due to the new normality that COVID-19 has imposed). In figure 8, it is shown the geographical distribution of the listings in the island. Moreover, it shows where are the listings located by type of room. It shows that there are few sharing listings, and plenty of entire homes/apartments and private rooms.

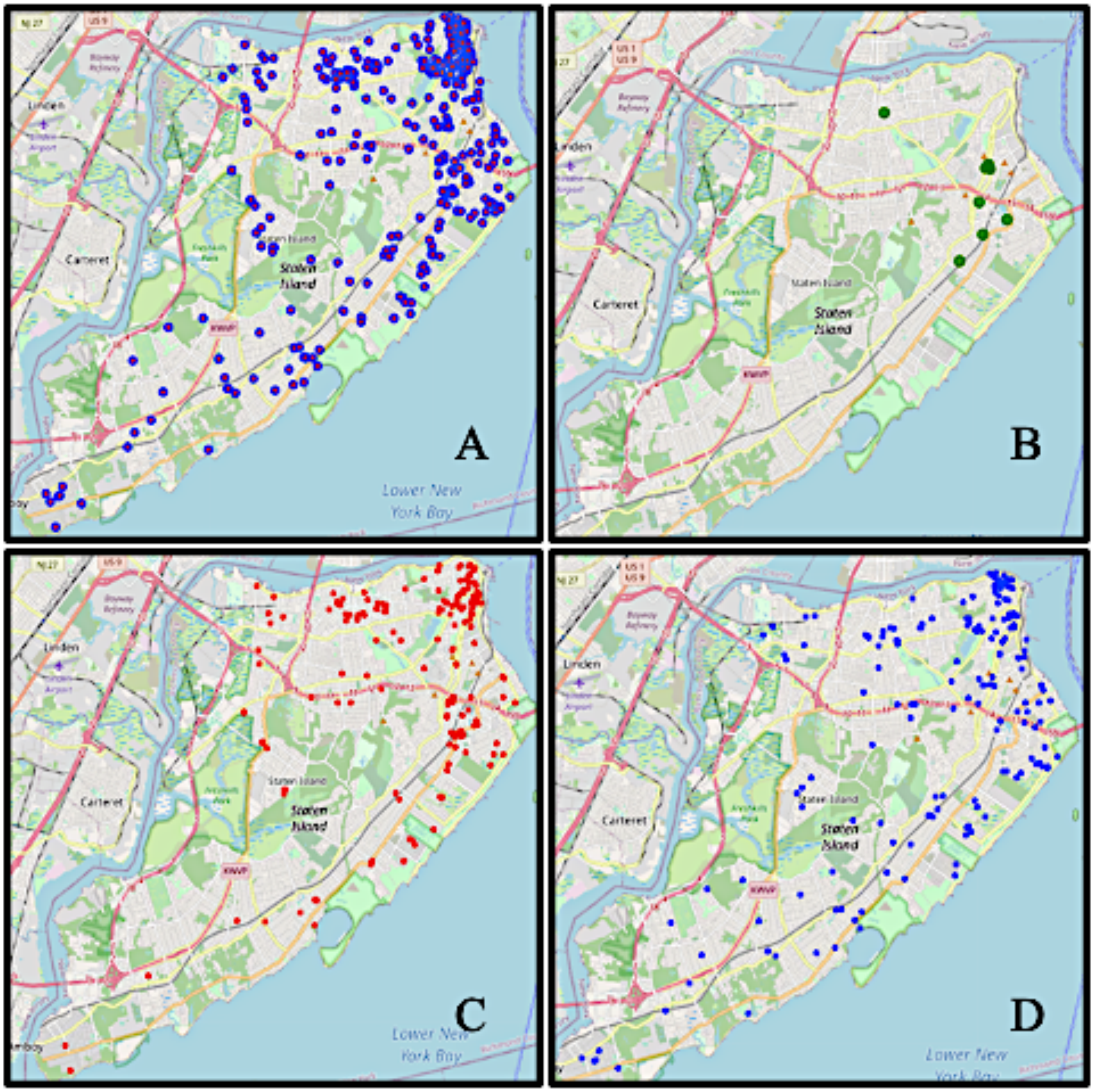
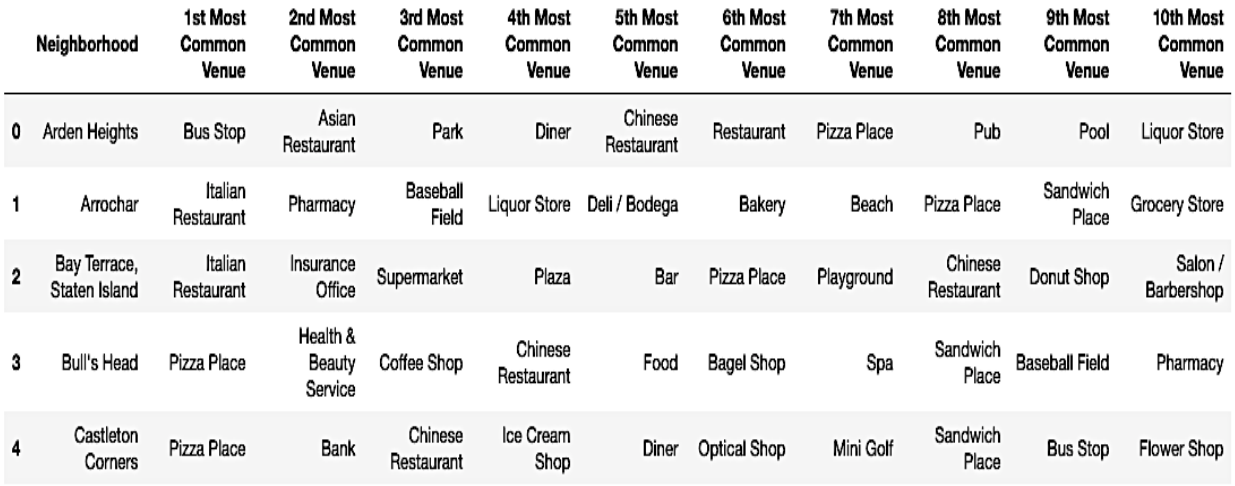


Figure 8. A) Geographical location of listings B) Shared rooms C) Private rooms D) Entire homes/apartments

After obtaining the JSON data from the Foursquare API, using techniques such as one hot encoding, and other technical steps, the top 10 venues per neighborhood were obtained. These neighborhoods are all within Staten Island, and some examples are shown in figure 9.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Arden Heights	Bus Stop	Asian Restaurant	Park	Diner	Chinese Restaurant	Restaurant	Pizza Place	Pub	Pool	Liquor Store
1	Arrochar	Italian Restaurant	Pharmacy	Baseball Field	Liquor Store	Deli / Bodega	Bakery	Beach	Pizza Place	Sandwich Place	Grocery Store
2	Bay Terrace, Staten Island	Italian Restaurant	Insurance Office	Supermarket	Plaza	Bar	Pizza Place	Playground	Chinese Restaurant	Donut Shop	Salon / Barbershop
3	Bull's Head	Pizza Place	Health & Beauty Service	Coffee Shop	Chinese Restaurant	Food	Bagel Shop	Spa	Sandwich Place	Baseball Field	Pharmacy
4	Castleton Corners	Pizza Place	Bank	Chinese Restaurant	Ice Cream Shop	Diner	Optical Shop	Mini Golf	Sandwich Place	Bus Stop	Flower Shop

Figure 9. Top 10 venues in neighborhood within Staten Island

Using the information above, the Top 1 venue of each neighborhood was isolated and graphed in figure 10. This figure shows the venues' categories with their respective count. The most common spots within Staten Island are pizza places, delis/bodegas and American restaurants. Nonetheless, there are interesting venues within the area such as: a sports bar, a lighthouse, a baseball field, a lake, among others.

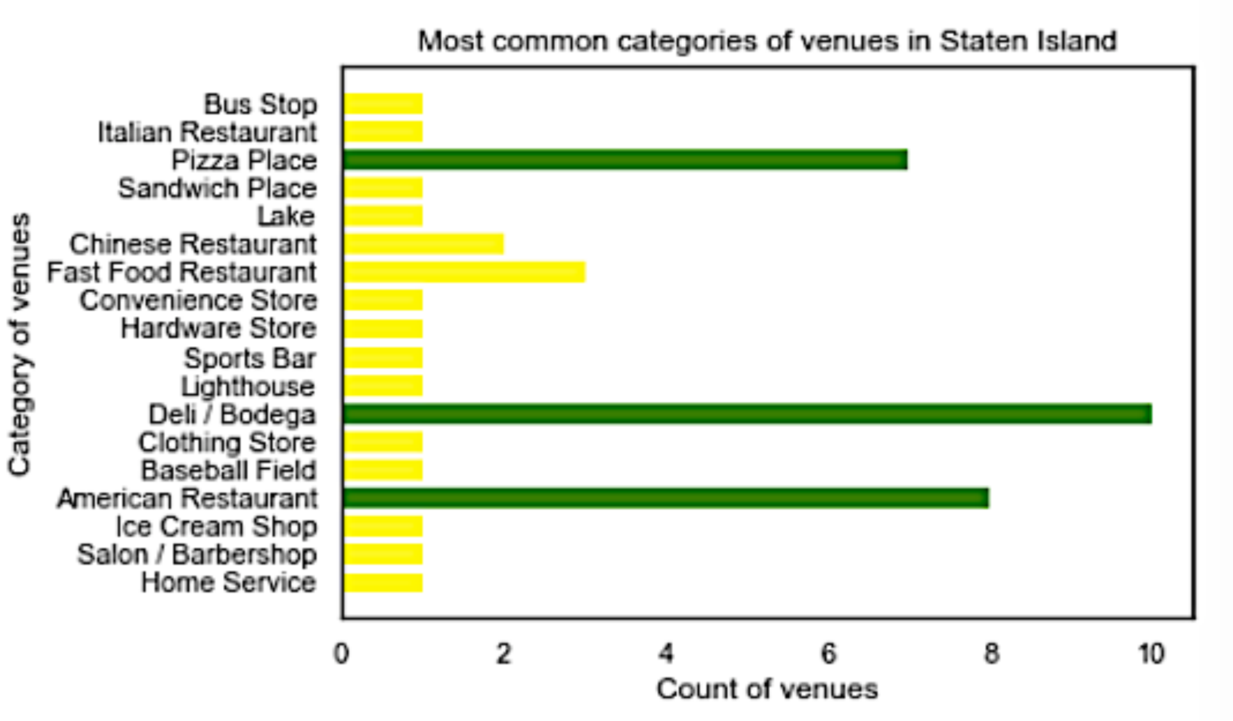


Figure 10. Venues' categories count

4.3 Clustering

For this project, it is also very important to provide the listings' owners with enough financial information to evaluate if their prices per night are within the price range of the listings around them. Therefore, in this final analysis, a clustering technique (K-Means) was used to find the average price of each group of similar listings. The number of clusters was obtained with the elbow method, please see figure 11. Furthermore, the mean price per cluster was included in table 2. Finally, a visual representation of the clusters was plotted with the folium library in figure 12.

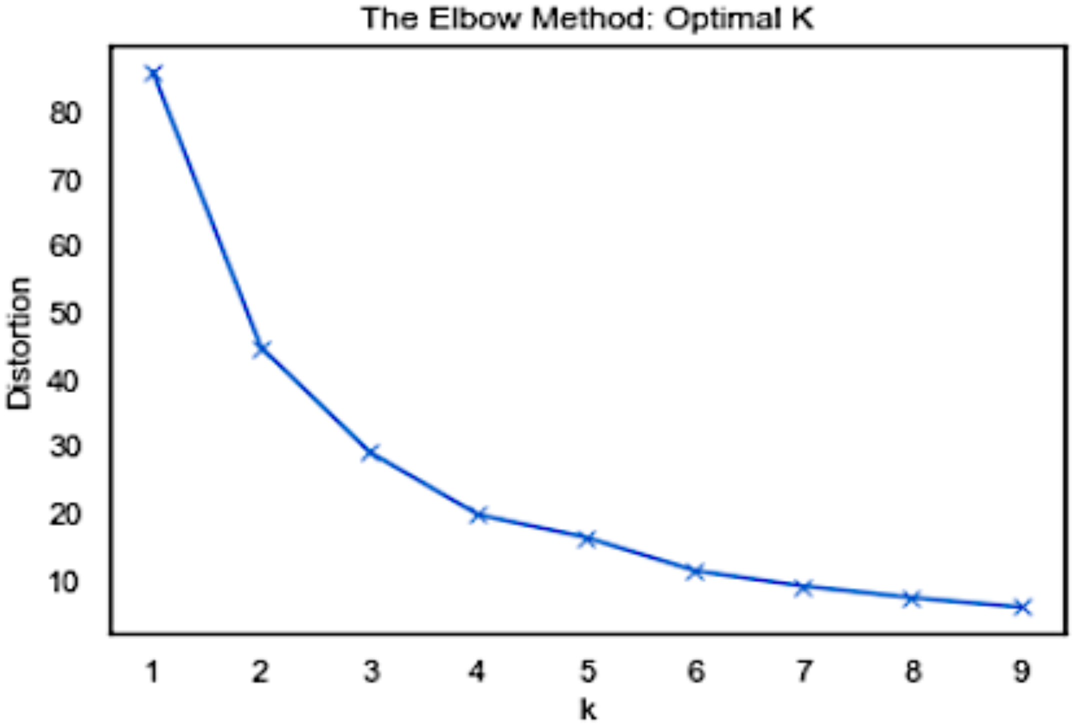


Figure 11. Elbow method to obtain number of clusters

Table 2. Mean price per cluster

Cluster	Mean price in USD
#1	152.12
#2	750.00
#3	348.78
#4	106.72
#5	249.00
#6	71.55

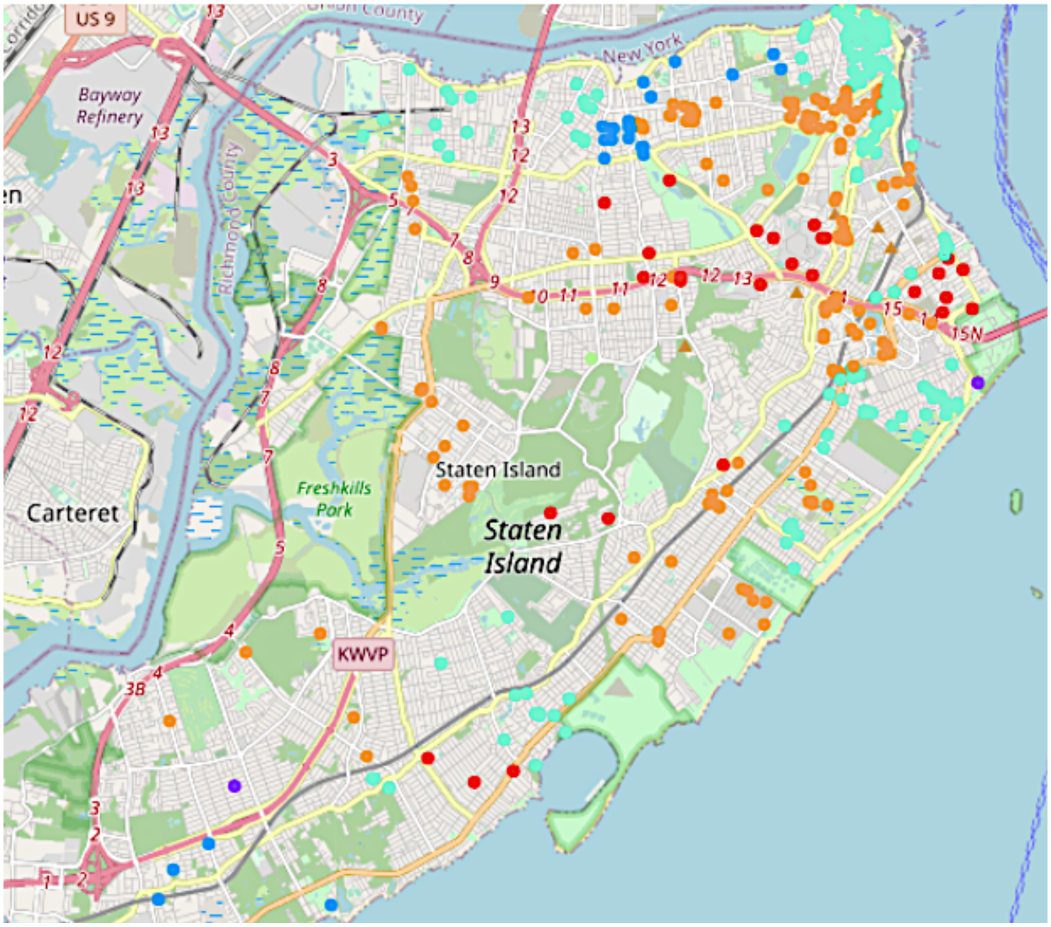


Figure 12. Clusters' geographical representation

5 Discussion

As mentioned above, the main findings of this project were that near New York City, there are some local areas that has a lot to offer to people near the surroundings. In the case of Staten Island, there plenty of restaurants to go to (especially pizza places and American food). On the other hand, there are outdoor activities to do such as baseball, lake visiting, lighthouse exploring, among others. Nonetheless, the top venues were pizza places, delis/bodegas, and American restaurants.

On the other hand, all the listings within Staten Island were clustered into 6 different groups. With this information, the mean price of the listings of each clustered was computed. This information will allow the owners of the listings to analyze if their prices are competitive or not. The latter will also help them to plan accordingly to the new normal era of tourism, which is more local and in shorter periods.

The whole data set was analyzed too. The main findings were not surprising, due to the fact that showed that the most expensive Airbnbs were situated within Manhattan. It also showed that the most expensive type of listings were entire homes and apartments. The price distribution of each type of listing by borough was graphed and showed that Staten Island offered listings with a wide range of prices, which also inspired the project to take the analysis further.

6 Conclusion

While the pandemic of COVID-19 remains among us, the tourism sector will need to readapt their clauses in order to survive. In term of accommodation, some actions will need to be taken. However, one of the most common aspects of this new normality, is that tourism needs to be more local. Tourists will be forced to travel only by car, and the most obvious thought is that they are going to visit near locations.

In this project, under the assumptions mentioned above, a local area near the most important city of the world was analyzed. The Airbnb listings within Staten Island were retrieved and positioned in the map; furthermore, the most common venues were obtained and studied. Similarly, these listings were clustered into 6 different groups and their mean price was computed. The latter will allow the owners of the listings to reassure or restructure their prices, in order to be more competitive.

7 References

Entrepreneur staff (2020). Airbnb CEO: It Took Us 12 Years to Build, and We Lost Almost Everything in 6 Weeks. *Entrepreneur magazine*

Gössling, S., Scott, D., & Hall, C. M. (2020). Pandemics, tourism and global change: a rapid assessment of COVID-19. *Journal of Sustainable Tourism*, 1-20.

Hu, M. R., & Lee, A. D. (2020). Airbnb, COVID-19 Risk and Lockdowns: Local and Global Evidence

Krstic, Z. (2020). Is it safe to Rent an Airbnb? How to Lower COVID-19 Risks in Vacation Homes

Pavlovska, E. (2020). “Airbnb CEO says post-COVID travel will never be the same again” in *New Europe*.