Battle of Neighborhoods: New York City & Airbnb Listings

Applied Data Science Capstone Project

Luz Ortega

Project Description

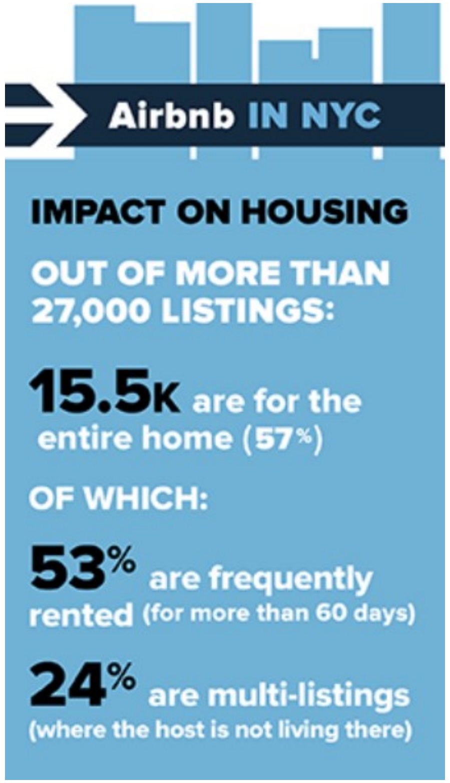
- Nowadays, with the increasing smart devices penetration worldwide, location data have become easier to obtain and it is used by multiple companies/apps to build better mobile experiences

- Airbnb is a community-driven hospitality company, which lets people rent out their properties or spare room to guests. 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. Some of the 2017 stats include:
  - **the city welcomed an estimated 62.8 million visitors in**
  - **49.7 million domestic visitors**
  - **13.1 international visitors**
  - **4,000 new rooms added to the city's inventory**
  - **Sold record of 36.4 million hotel room nights**

Project Objective

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 their surrounding venues.



Airbnb IN NYC

IMPACT ON HOUSING

OUT OF MORE THAN 27,000 LISTINGS:

15.5K are for the entire home (57%)

OF WHICH:

53% are frequently rented (for more than 60 days)

24% are multi-listings (where the host is not living there)

Datasets

New York City Neighborhood Names

Latitude and Longitude of NYC Neighborhoods

Inside Airbnb

Adding data to the debate

INDEPENDENT, NON-COMMERCIAL, OPEN SOURCE DATA TOOL

How is Airbnb really being used in and affecting your neighborhood?

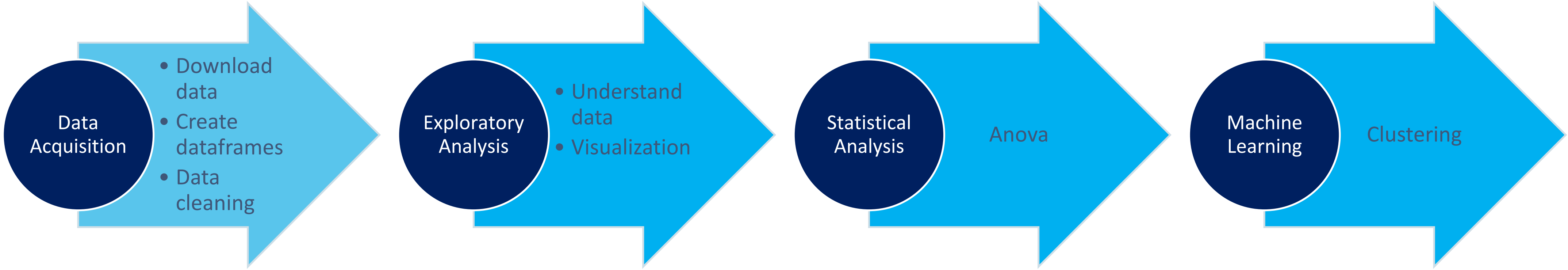
Inside Airbnb

Airbnb listings, including information such as their location, price, rating, and others

Foursquare

Location data that provides trending venues of a given pair of coordinates

Methodology



Data Acquisition

- Download data
- Create dataframes
- Data cleaning

Exploratory Analysis

- Understand data
- Visualization

Statistical Analysis

Anova

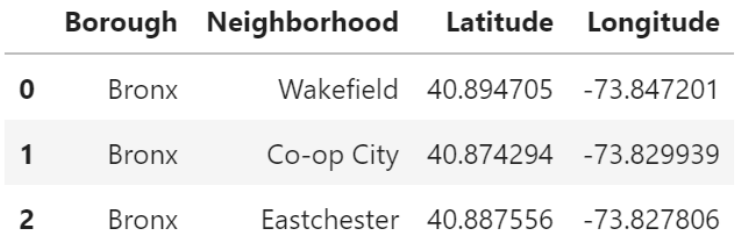
Machine Learning

Clustering

Methodology

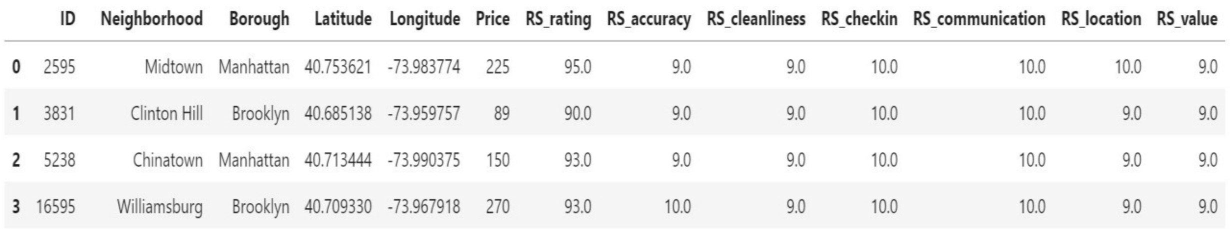
Data Acquisition

- Dataframe created using the NYC Neighborhoods dataset



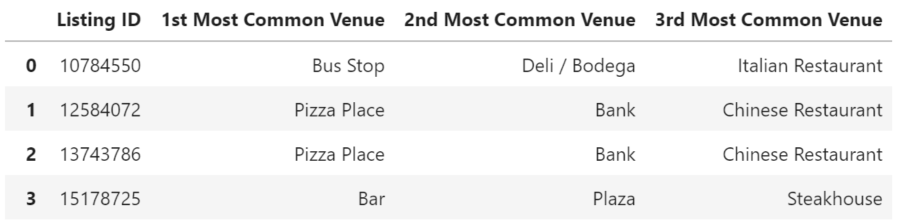
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806

- Dataframe created using the Inside Airbnb dataset



	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

- Dataframe created using Foursquare data and Inside Airbnb data

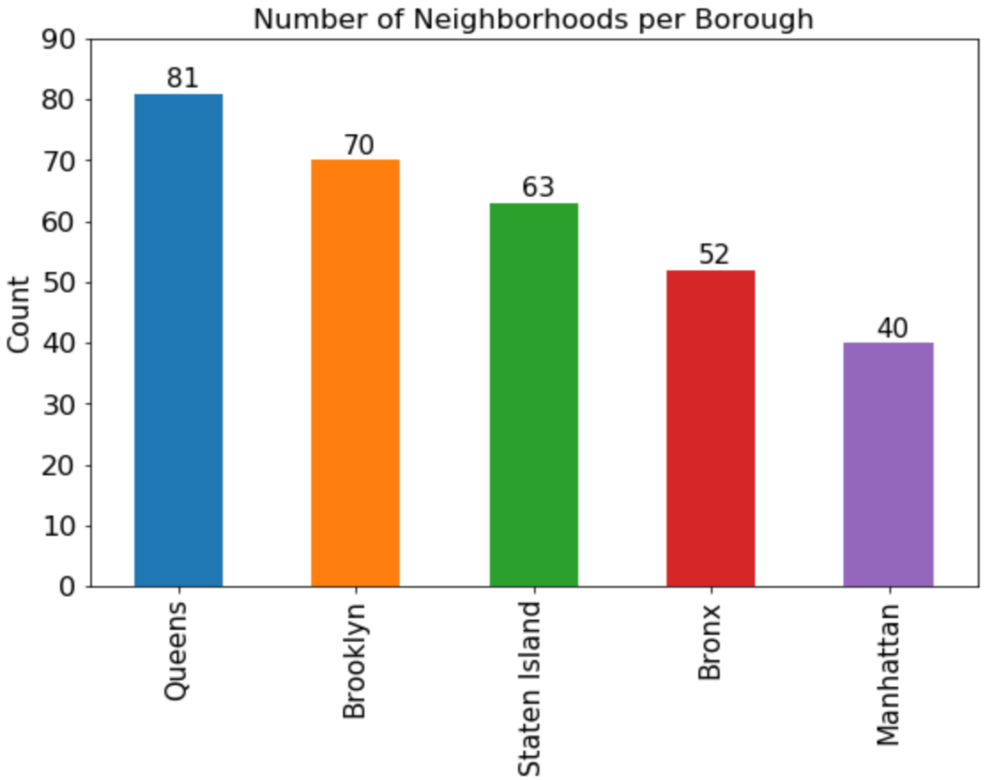


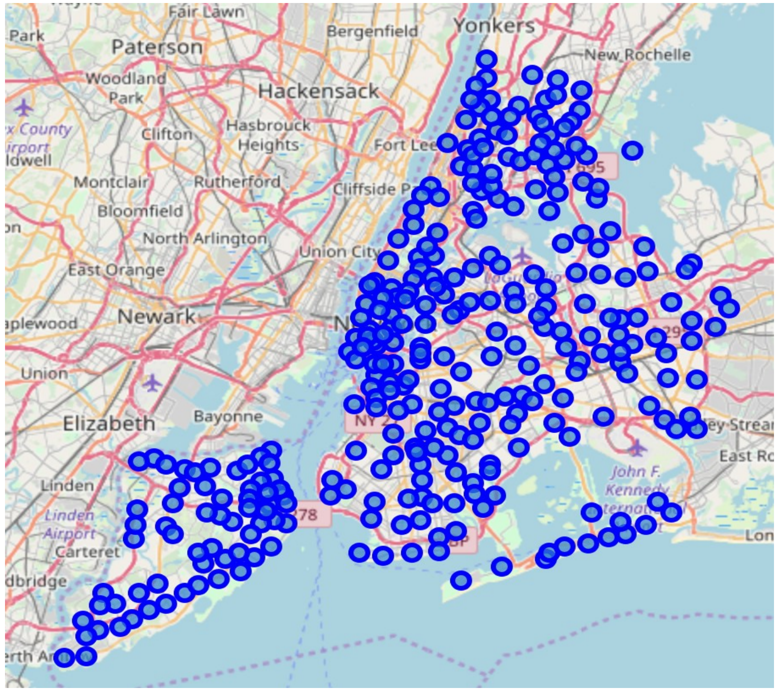
	Listing ID	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	10784550	Bus Stop	Deli / Bodega	Italian Restaurant
1	12584072	Pizza Place	Bank	Chinese Restaurant
2	13743786	Pizza Place	Bank	Chinese Restaurant
3	15178725	Bar	Plaza	Steakhouse

Methodology

Exploratory Analysis

- Source: dataframe created using the NYC Neighborhoods dataset



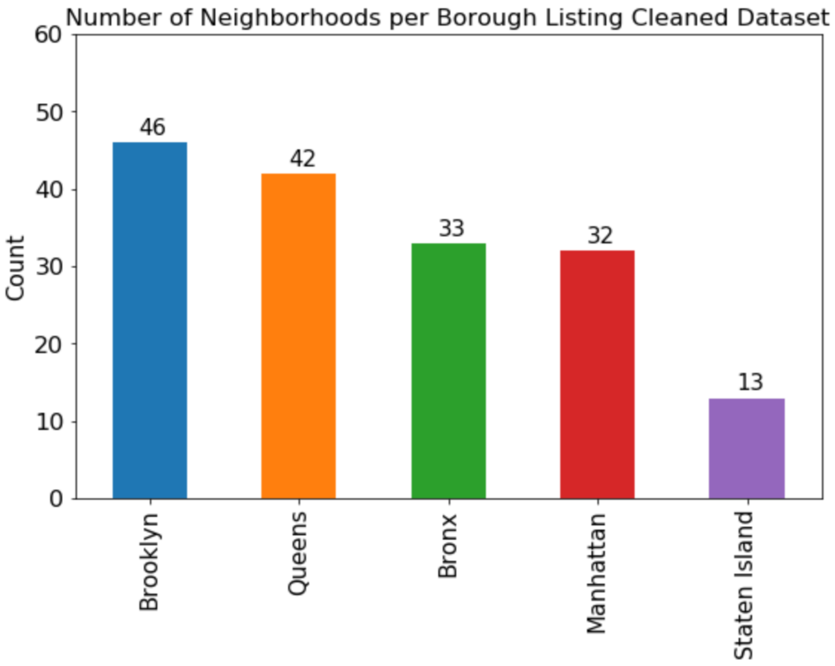


Neighborhoods of NYC

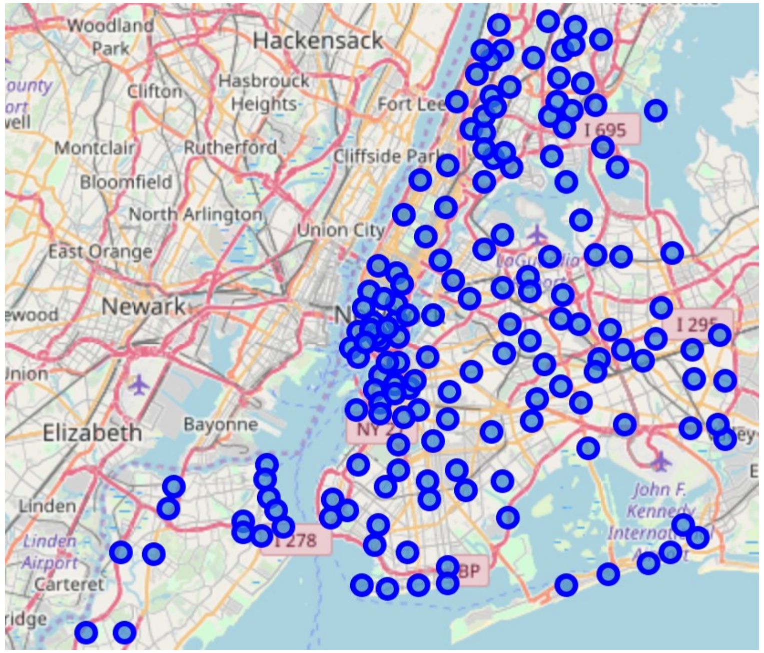
Methodology

Exploratory Analysis

- Source: dataframe created using the Inside Airbnb dataset



Note: Number of neighborhoods where there are rental places available

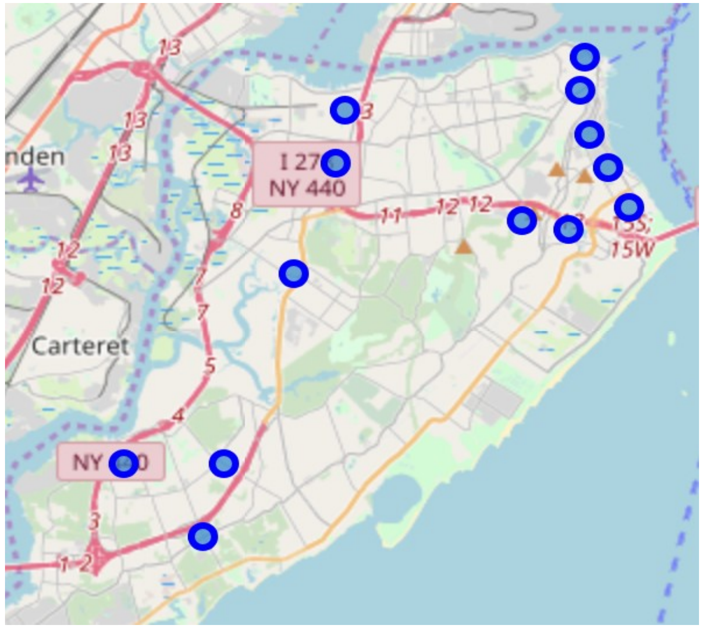


Neighborhoods where listings are located

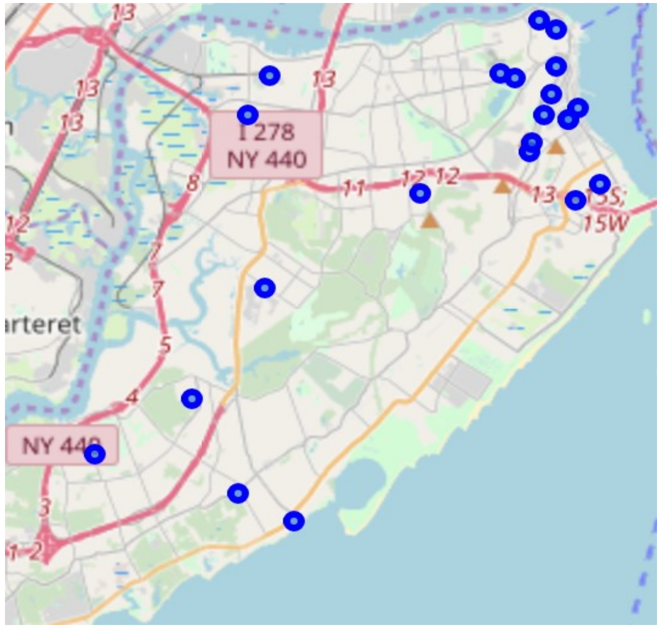
Methodology

Exploratory Analysis

- Staten Island



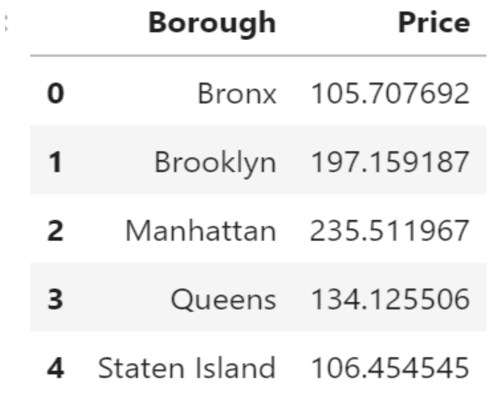
Staten Island Neighborhoods



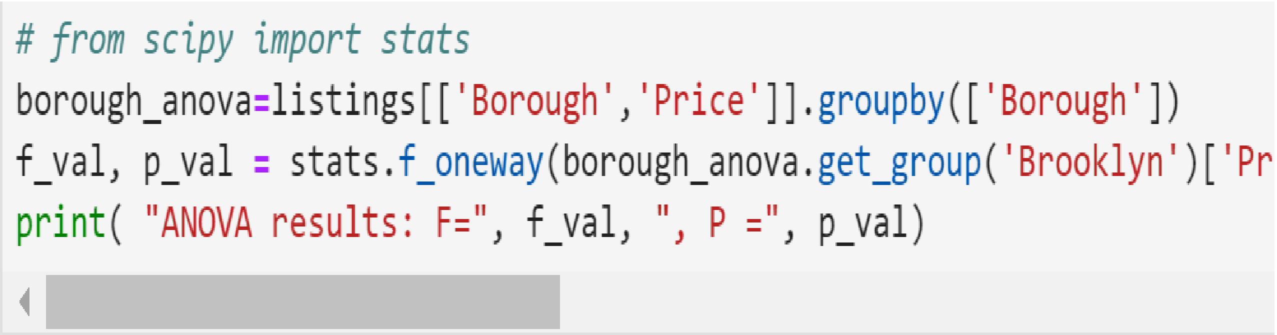
Listings in Staten Island

Results

- ANOVA to test whether there was a significant difference between the mean price of the listings in each Borough.



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

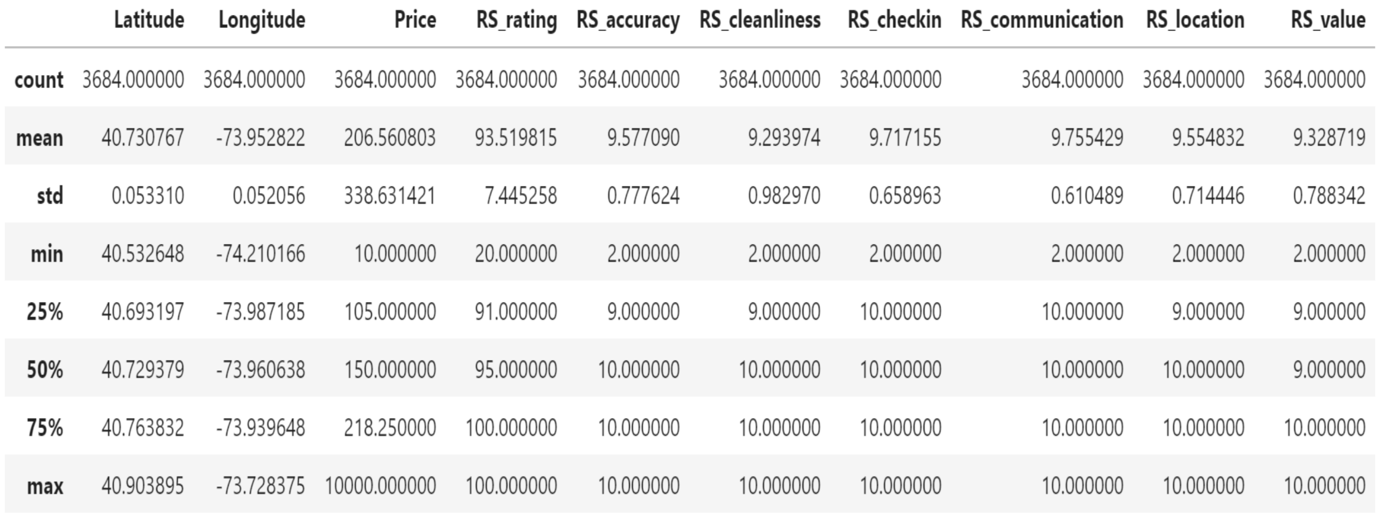


```
# from scipy import stats
borough_anova=listings[['Borough','Price']].groupby(['Borough'])
f_val, p_val = stats.f_oneway(borough_anova.get_group('Brooklyn')['Pr
print( "ANOVA results: F=", f_val, ", P =", p_val)
```

Results

Statistical Analysis

- Descriptive statistics of the listings dataframe

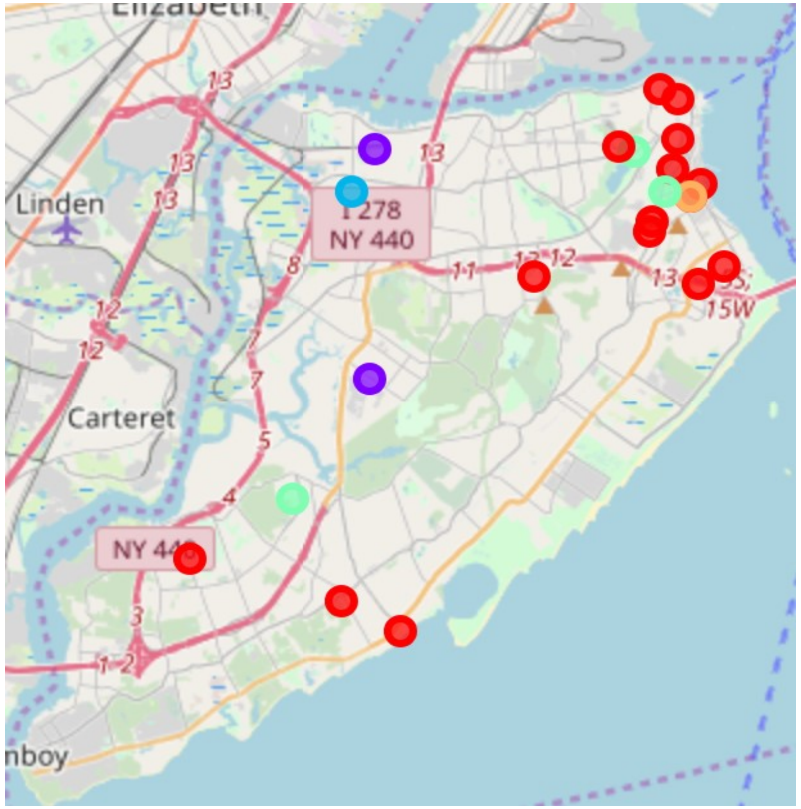


	Latitude	Longitude	Price	RS_rating	RS_accuracy	RS_cleanliness	RS_checkin	RS_communication	RS_location	RS_value
count	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000	3684.000000
mean	40.730767	-73.952822	206.560803	93.519815	9.577090	9.293974	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

Results

Machine Learning

- Clustering:
  - k-means to cluster the listings into 5 clusters
  - 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
  - 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



Resulting Clusters

Conclusion

- The most consuming part of this project was the construction of the Inside Airbnb final dataframe.

- Foursquare data is obtained through API calls, therefore different results will be obtained if the code is run at different times.

- The ANOVA results show there is a statistical significance difference among the mean prices of the listings in the different boroughs.

- 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.

- 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

Thank you!

Luz Ortega