The Battle of Neighborhoods: Find the best place to stay in New York City

Applied Data Science Capstone Project

March 05, 2020

Created by: Siarhei Vasiaichau - gmlvsv@gmail.com

Introduction: Business Problem

Background

According the latest *NYC&Company* release New York City welcomed about 65.2 million tourists in 2018 year - 51.6 million domestic and 13.5 million international visitors. And these numbers are continuously increasing from year to year.

New York City has the largest selection of lodging choices in the country – from the hostels to the luxury hotels. The prices vary from 100$ till several thousand dollars with average price 292 USD per night.

Problem Description and Project Goal

In New York City there are almost 300 hotels with over 75,000 hotel rooms and Airbnb has more than 50,000 apartment listings in New York City in 2018 year - it can be hard to find the right fit or know how much you will get with your money.

In this project we will try to find the most optimal neighborhoods on Manhattan where a tourist can rent an accommodation via Airbnb service and have a pleasant stay in NYC and a possibility to attend the most visited attractions like Central Park, Times Square and so on.

Target Audience

This investigation would interest New York City's visitors who prefers short stays (from 1 night) and wants to select the best neighborhoods on Manhattan, New York.

Data

Initial datasets

In our investigation we will use the free and public available datasets.
We will try to evaluate available Airbnb 2019-year accommodations on Manhattan, New York and define the most reasonable apartments sets (clusters) for the visitors.

Based on definition of our problem, we suppose that factors that will help us are:

- accommodation's average price per person by the neighborhood;
- number of tourist attractions near the accommodation;
- number of crimes nearby the accommodation.

Following data sources are needed for our project:

- New York City apartment listing from the Inside Airbnb site;
- New York Neighborhoods Tabulation Areas – official NYC neighborhoods names and coordinates;
- Foursquare API to extract data about venues - food places, museums, galleries, shopping centers, and so on;
- New York Crime data records for 2019 year.

Data

Data Cleaning

Airbnb New York City apartment listing

For our project records were filtered as

- Borough - Manhattan, New York only;
- Number of reviews >= 10;
- Availability >= 10 days/year;
- Last Scraped/Reviewed later than 2019-10-01;
- Minimum nights >= 1;
- Excluded Hostels and Camper/RV;
- Excluded Shared rooms.

After filter was applied, we have 2,356 accommodations in our dataset.

New York Police Crime Records

We filter this dataset by

- Borough – *Manhattan, New York* only;
- Crime type – *FELONY* and *MISDEMEANOR*.

After filtering we have 101,086 crimes records for Manhattan in 2019 year.

Data

Feature Engineering

Airbnb

We add some new features (columns) to our Airbnb dataset**:**

- **full_price** - *price* + *cleaning_fee*. Airbnb *price* column could be misleading because it does not include mandatory cleaning fee price;
- **price_per_person** - (price + cleaning_fee)/accommodates;
- **tab_area** from *New York Area Tabulation Name* dataset to our *Airbnb* data set because Neighborhoods' names are quite different in these data sets. We use custom *define_tab_area* function which returns *New York Area Tabulation Name* for each Airbnb accommodation's latitude/longitude pair;
- **crimes -** calculate the number of crimes in radius of 100 meters from each accommodation.

New York Police Crime Records

We added **tab_area** column (New York Area Tabulation Name) to NYC Manhattan Crimes data set because we need to display **Crime Rate** Information on the New York Area Tabulation Map.

Methodology

In this project we are trying to detect Manhattan's Neighborhoods that have accommodations for rent with positive reviews, reasonable prices, low number of crimes and tourists' attractions nearby.

In the first step we have collected the following data:

- Airbnb Accommodations with their NYC Tabulation Area (official neighborhood names);
- Airbnb Accommodation's number of crimes nearby;
- Defined NYC Tabulation Area (official neighborhood name) for each Manhattan's crime case.

The second step in our analysis will be a calculation and exploration different neighborhoods of Manhattan. We will explore the following characteristics:

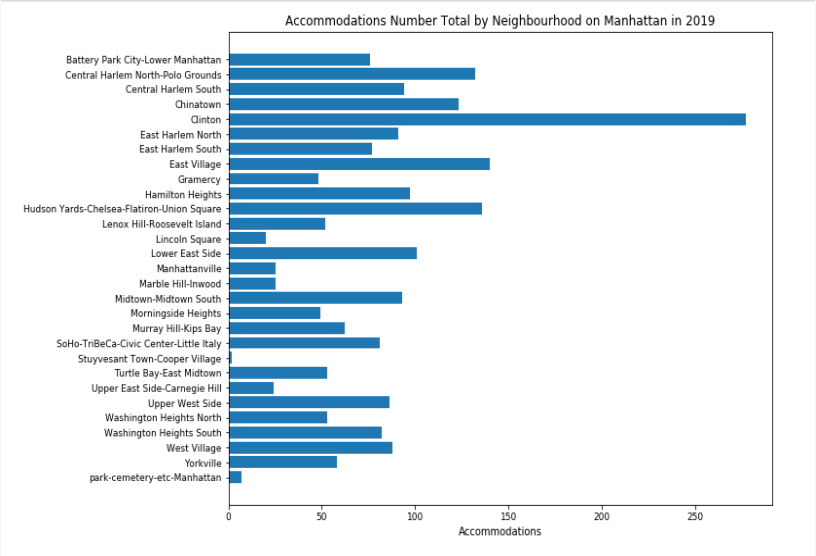
- number of crimes in the area;
- average price per person;
- number of accommodations available.

In the third and final step we will

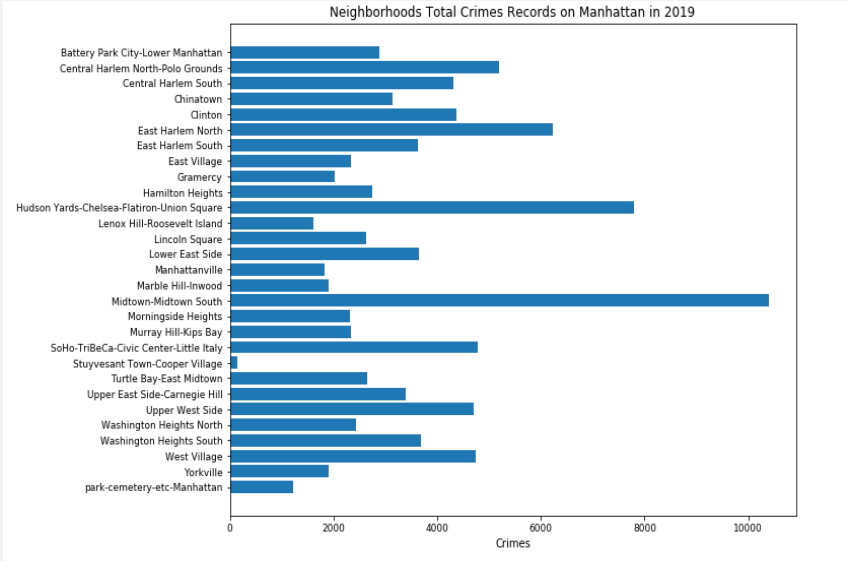
- select Top-100 Airbnb accommodations based on summary rating, number of crimes and price per person, and
- invoke Foursquare API to find Top accommodations' nearby venues
- create and investigate clusters (using k-means clustering) for our accommodations to make some recommendations to our tourists.

Analysis

Apartments Total by Neighborhood Chart

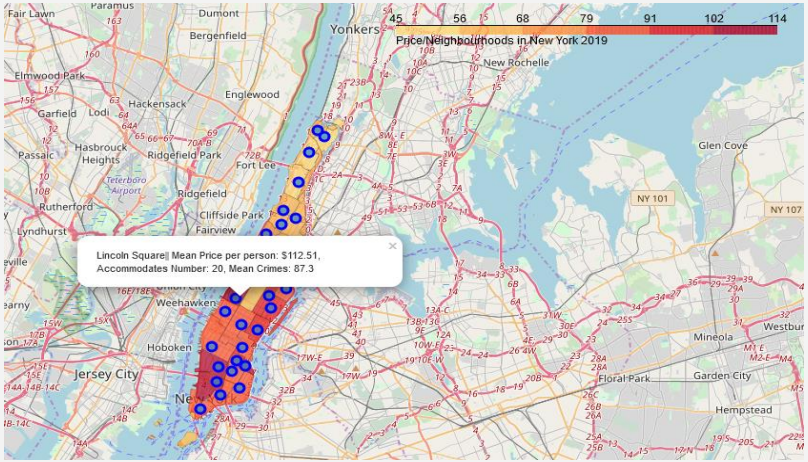


Neighborhoods Crimes Records Chart

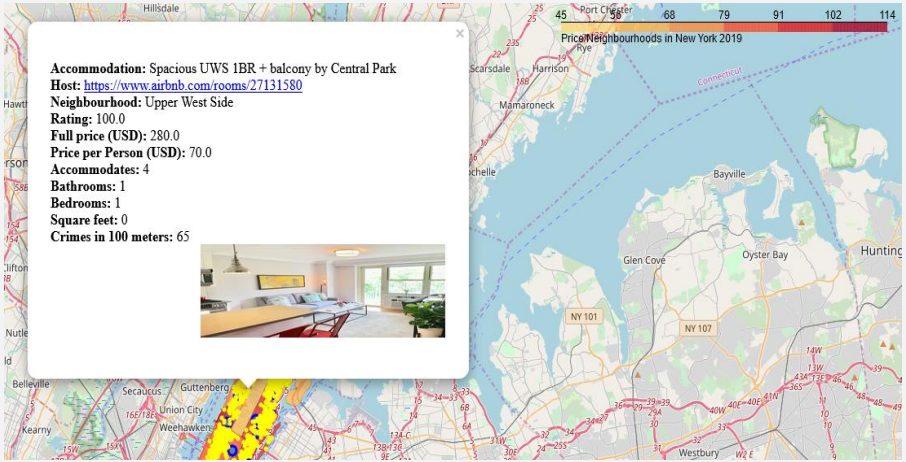


Analysis

NYC Tabulation Area Neighborhoods Average Prices per Person



Accommodations Detailed Info Map



Top-5 Neighborhoods with **Highest average *Price per Person*** in 2019 year:

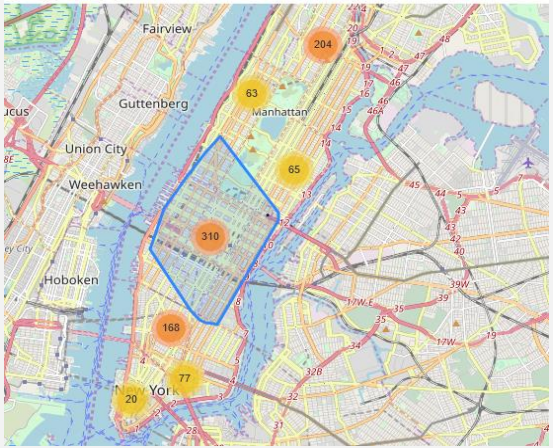
- West Village - 112.85 USD - 88 accommodations
- Lincoln Square - 112.51 USD - 20 accommodations
- Stuyvesant Town-Cooper Village - 107.5 USD - 2 accommodations
- SoHo-TriBeCa-Civic Center-Little Italy - 105.38 USD - 81 accommodations
- Upper East Side-Carnegie Hill - 96.98 USD - 24 accommodations

Top-5 Neighborhoods with **Lowest average *Price per Person*** in 2019 year:

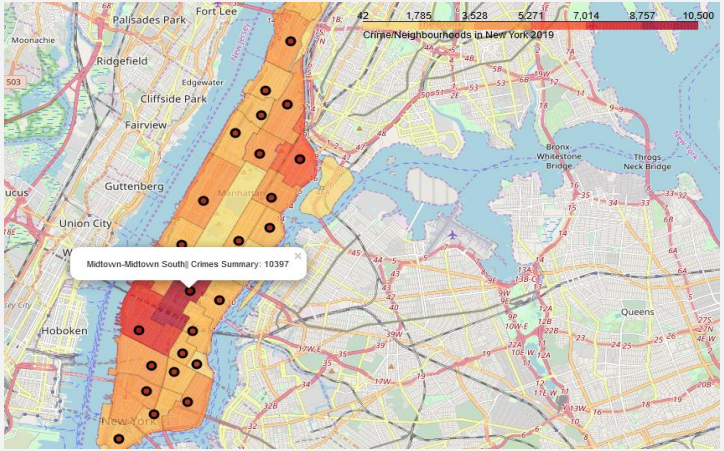
- Marble Hill-Inwood - 45.48 USD - 25 acc.
- Washington Heights South - 46.79 USD - 82 acc.
- Washington Heights North - 54.74 USD - 53 acc.
- Central Harlem North-Polo Grounds - 57 USD - 132 acc.
- Manhattanville - 59.75 USD - 25 acc.

Analysis

Crimes Cluster Map



Summary Crimes by Neighborhoods Map



Top-5 Neighborhoods with the **Highest Crime level** in 2019 year:

- Midtown-Midtown South - 10,397
- Hudson Yards-Chelsea-Flatiron-Union Square - 7,788
- East Harlem North - 6,221
- Central Harlem North-Polo Grounds - 5,186
- SoHo-TriBeCa-Civic Center-Little Italy - 4,789

Top-5 Neighborhoods with the **Lowest Crime level** in 2019 year:

- Stuyvesant Town-Cooper Village - 145
- park-cemetery-etc-Manhattan - - 1,213
- Lenox Hill-Roosevelt Island - 1,604
- Manhattanville - 1,832
- Yorkville - 1,898

Analysis

Foursquare API Neighborhoods Analysis

Because of the Foursquare API limitations for free usage lets analyze Top-100 Accommodations from the Airbnb data set.
We define Top-3 Venue Categories for each accommodation in radius of 1000 meters.
Then we will try to define the 3 clusters for these accommodations.

Select Top-100 Airbnb accommodations by

- **review_scores_rating** - overall accommodations rating - from maximum 100 to lower values;
- **full_price** - from lower price to higher;
- **price_per_person** - from lower price to higher;
- **crimes** - from lower number to higher;

We define our custom Top-Level categories for Venues

```
fine_art_cat = ['Art','Arts','Museum', 'Library','Exhibit','Gallery']
eat_place_cat = ['Restaurant','Steakhouse']
shopping_cat = ['Shopping Mall','Market','Boutique']
outdoor_cat = ['Sculpture Garden','Scenic Lookout','Roof Deck','Outdoor Sculpture','Monument / Landmark',
               'Memorial Site','Lighthouse','Historic Site','Harbor / Marina','Fountain','Event Space','Bridge',
               'Waterfront','Church','Building','Garden','Historic Site','Lake','Park',
               'Pier','Rest Area','River','Synagogue','Field']
entertainment_cat = ['Nightclub','Circus','Club', 'Stadium', 'Karaoke Bar', 'Pub','Theater','Opera', 'Concert', 'Zoo']

#Join all categories' values in one
tourists_categories = fine_art_cat + eat_place_cat + shopping_cat + outdoor_cat +entertainment_cat
```

Analysis

Foursquare API Neighborhoods Analysis

Calculate the Top-3 Venues Categories for each accommodation.

Then run k-means to cluster the neighborhood into 3 clusters.

Cluster Labels	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	1st Most Common Venue Share	2nd Most Common Venue Share	3rd Most Common Venue Share
1	**Stylish, Quiet, Centrally Located (9th & 52nd)	Food Place	Entertainment	Fine Art	0.62	0.35	0.04
2	157-C	Food Place	Sightseeing	Fine Art	0.52	0.29	0.14
2	A neat bedroom in a cozy 3-bedroom apartment	Food Place	Sightseeing	Shopping	0.48	0.41	0.04
0	Art filled peaceful paradise EV Union Square	Food Place	Sightseeing	Shopping	0.71	0.07	0.07
0	Artsy Parisian Apt in Greenwich Village	Food Place	Entertainment	Sightseeing	0.65	0.26	0.09

Analysis

Foursquare API Neighborhoods Analysis

Now, we can examine each cluster and determine our custom venue categories that distinguish each cluster.

Cluster 0 – Mix (red dots) characteristics:

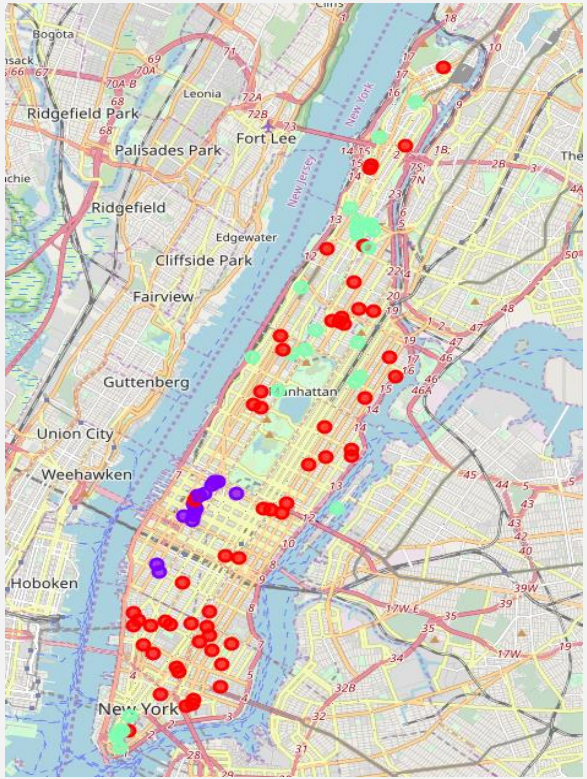
- average *price_per_person;*
- average *crimes* rate;
- second top Common Venue Category has a Mix of all kind of Categories;
- contains 58% from all top accommodations.

Cluster 1 – Entertainment (blue dots) characteristics:

- highest average *price_per_person* among all clusters;
- highest average *crimes* rate among all clusters;
- *Entertainment* is 1st and the 2nd Top Common Venue Categories;
- contains 15% from all top accommodations.

Cluster 2 – Sightseeing (light-green dots) characteristics:

- lowest average *price_per_person;*
- lowest *crimes* rate among all clusters;
- *Sightseeing* is the second top Common Venue Category;
- contains 27% from all top accommodations.



Results and Discussion

During the analysis, three clusters were defined.
All clusters have a 'Food Place' category as the First Common Venues. This is what we have in common among our clusters.

But they are distinguished by the other characteristics as

- average **Price per person;**
- average **Crimes Rate;**
- the second Common Venues;
- number of available Airbnb accommodations;
- neighborhoods location.

Cluster 0 – Mix is the most generic cluster - it has a

- average price_per_person - $110;
- average crimes rate - 67 (but very varying - depends on the neighborhood, from 3 to 385 crime cases in radius of 100 meters from the accommodation);
- mix of all Venue Categories (Fine Arts, Shopping, Entertainment);
- contains 58% from all accommodations selected from analysis (Top-100 Airbnb accommodations);
- spreads almost on all Manhattan's areas.

Results and Discussion

Cluster 1 - Entertainment is the smallest cluster with the following qualities (Nightclub, Stadium, Pub, Theater, Concert and so on):

- highest average *price_per_person* among all clusters - $111;
- highest average *crimes* rate among all clusters – 102;
- *Entertainment* is 1st and the 2nd Top Common Venue Categories;
- contains 15% from all top accommodations (Top-100 Airbnb accommodations);
- spreads on *Chelsea, Hell's Kitchen, and Midtown* Airbnb's Neighborhoods.

Cluster 2 - Sightseeing is the cheapest one with many Sightseeing attractions nearby (Monument/Landmark, Memorial Site, Historic Site, Lake, Park, Pier, and so on)

- lowest average price_per_person - $59;
- lowest crimes rate among all clusters – 65;
- *Sightseeing* is the second top Common Venue Category;
- contains 27% from all top accommodations (Top-100 Airbnb accommodations);
- spreads on *East Harlem, Financial District, Harlem, Inwood, Morningside Heights, Roosevelt Island, Upper West Side, Washington Heights, West Village.*

We identified three clusters from which a visitor could choose an appropriate accommodation based on his/her preferences or needs.

Limitations

- We limited our investigation by Manhattan Borough only;
- Foursquare free account has a limitation of 950 calls/day so maybe it's worth to upgrade our free account to analyze Top-1000 Airbnb accommodations instead of Top-100.

Conclusion

To conclude, the basic data analysis was performed to identify Manhattan's Neighborhoods clusters for a short stay visit.
During the analysis, we cleansed and investigated Manhattan Neighborhoods' datasets, found some statistical characteristics and visualize them.
The aim of this project is to help Manhattan visitors select the Airbnb neighborhoods where to stay based on the most common venues, price policy, and safety characteristics:

- if a person is interested in **entertainment** (Nightlife, Pubs, Concerts, Movies) we recommend paying attention for accommodations from the *Cluster 1 - Entertainment*: *Chelsea, Hell's Kitchen, and Midtown* Airbnb's Neighborhoods. But the person should take into the consideration the high prices and crime rate for this location;
- if a person is looking for a neighborhood with **lower prices** and nice views nearby, we recommend looking at *Cluster 2 - Sightseeing*: *Chelsea, Hell's Kitchen, and Midtown Airbnb's Neighborhoods*;
- if a person **does not have any preferences** - investigate proposals from *Cluster 0 - Mix*. It has average prices and spreads over almost all Manhattan's neighborhood.

Areas of improvement

- We could include the other NYC Boroughs - The Bronx, Brooklyn, Queens, and Staten Island;
- We also could utilize other services like Google API to find nearby Venues;
- We have not analyzed the Hotels. It's very big chunk but we have not found any fresh public data sets about hotels accommodations with rating.