

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

**Submitted by:
Shereena George**

Table of contents

1. [Introduction: Business Problem](#)
2. [Data](#)
 - 2.1. [Initial Datasets](#)
 - 2.2. [Data Cleaning and Feature](#)
3. [Methodology](#)
4. [Analysis](#)
5. [Results and Discussion](#)
6. [Conclusion](#)
7. [References](#)

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. The Hotel Occupancy rate is also high – in 2018 year it was 88%. Visitors prefer short stays that are often over weekends - averaging 2.4 nights.

Problem description

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, Empire State Building, 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.

Success Criteria

The success criteria of this project will be a recommendation with the set of apartments clusters have the best score calculated based on

- Accommodation price with fees
- Location of the accommodation
- Venues in radius of 1000 meters from the accommodation
- Crime rate in radius of 100 meters from the accommodation.

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.

Airbnb New York City apartment listing

<http://data.insideairbnb.com/united-states/ny/new-york-city/2019-12-04/data/listings.csv.gz>

It is available below under a Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license. Initially data contains 50,599 rows and 106 columns with the information about available accommodations are name, borough, neighborhood, price per night, cleaning fee, minimum nights, guest number and so on.

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.

Neighborhood Tabulation Areas

<https://data.cityofnewyork.us/api/geospatial/cpf4-rkhq?method=export&format=GeoJSON>

The dataset contains Multi Polygon GIS data with the coordinates of each NYC neighborhood. We will use these data for the maps and for the mapping of Airbnb neighborhoods because Airbnb has different neighborhoods structure.

Foursquare API data about venues

Food places, museums, galleries, shopping centers, sightseeing attractions, concert halls and so on. We will check Top-50 venues for the Top-100 Manhattan's Airbnb accommodations in radius of 1000 meters.

New York Police Crime Records

<https://data.cityofnewyork.us/api/views/5uac-w243/rows.csv?accessType=DOWNLOAD>

We will use this statistic during our apartment evaluation. Originally it contains 461,711 rows and 35 columns. 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. This dataset contains NYC Precincts column which is not the same as Neighborhood Tabulation Areas. We need to define the NYC Neighborhood name by the latitude/longitude of each crime record from this dataset.

	CMPLNT_NUM	ADDR_PCT_CD	BORO_NM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	CRM_ATPT_CPTD_CD	
0	314773184	48	BRONX	31.12.19	06:00:00 PM	NaN	NaN	COMPLETED	N
1	289837961	25	MANHATTAN	30.12.19	08:30:00 PM	31.12.19	10:00:00 AM	COMPLETED	N
2	535744284	77	BROOKLYN	24.12.19	04:55:00 PM	24.12.19	05:00:00 PM	COMPLETED	N
3	895678119	52	BRONX	30.12.19	07:32:00 PM	NaN	NaN	COMPLETED	N
4	299841674	18	MANHATTAN	30.12.19	03:30:00 PM	30.12.19	04:50:00 PM	COMPLETED	N

Data Cleaning and Feature

Airbnb

We do not need all columns from the original dataset so let's create a subset of the needed columns:

- id - listing identifier
- name - accommodation's name
- last_review - accommodation's last review date
- listing_url - accommodation's URL
- picture_url - accommodation's picture URL
- neighbourhood_group_cleansed - NYC Borough's name. e.g. Manhattan, Bronx. We will use accommodations only from Manhattan
- neighbourhood_cleansed - Airbnb Neighborhood's name, e.g. Hell's Kitchen. These Names are not the same as NYC Neighborhood Tabulation Areas
- review_scores_rating - accommodation's weighted sum of other scores_review_scores_location, review_scores_value
- latitude - accommodation's latitude
- longitude - accommodation's longitude
- property_type - accommodation's type e.g. Entire home/Apt, Private Room. We exclude Hostels and Camper/RV
- room_type - accommodation's room type. We exclude Shared rooms
- accommodates - number of persons allowed. We use this value to calculate price_per_person custom column
- bathrooms - number of bathrooms. Keep it for informative reasons
- bedrooms - number of bedrooms. Keep it for informative reasons
- square_foot - accommodation's size. Keep it for informative reasons
- price - price per night
- security_deposit - security deposit
- cleaning_fee - additional fee. We will use it to calculate full_price per night for the accommodation
- minimum_nights - minimum nights for rent. We use accommodations with 1 or 2 minimum nights
- number_of_reviews_1m - number of reviews for the last month. Keep it for informative reasons
- reviews_per_month - average number of reviews per month. Keep it for informative reasons
- number_of_reviews_all - overall number of reviews. Keep it for informative reasons
- availability_365 - available days/year.

Now we should clean different Prices columns:

- fill in empty values
- convert String to Float, e.g. \$2,100.00 => 2100.00.

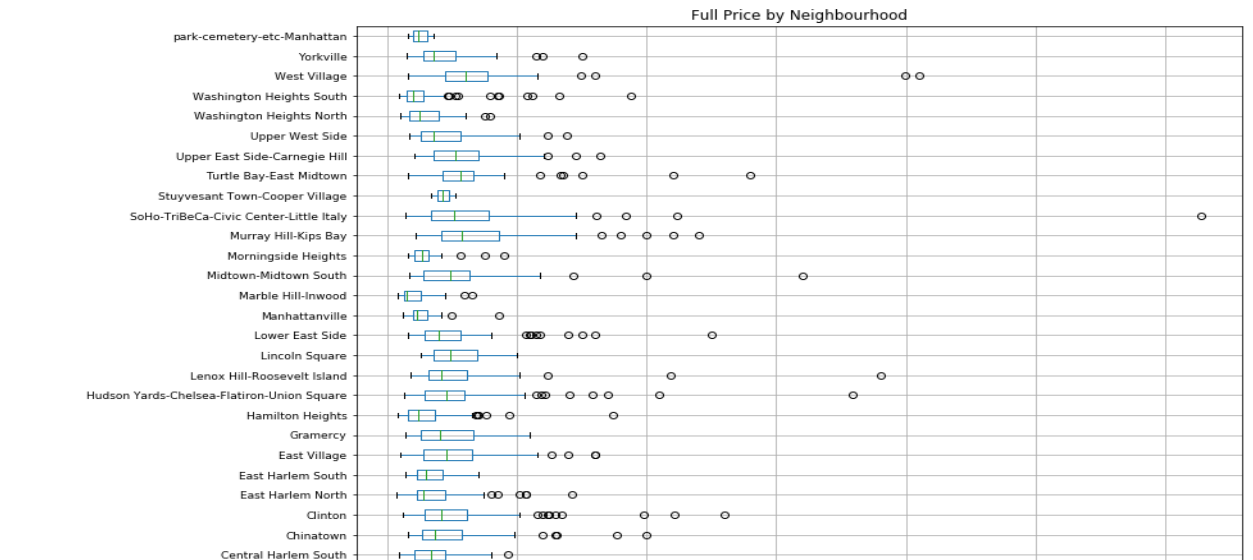
id	name	last_review	listing_url	picture_url	neighbour
5178	Large Furnished Room Near B'way	2019-11-21	https://www.airbnb.com/rooms/5178	https://a0.muscache.com/im/pictures/12065/f070997b_original.jpg?aki_policy=large	Manhattan
7322	Chelsea Perfect by Doti, an AIRBNB Super Host!	2019-11-16	https://www.airbnb.com/rooms/7322	https://a0.muscache.com/im/pictures/23207/33258e91_original.jpg?aki_policy=large	Manhattan
9704	Spacious 1 bedroom in luxe building	2019-11-09	https://www.airbnb.com/rooms/9704	https://a0.muscache.com/im/pictures/38418/569b54fd_original.jpg?aki_policy=large	Manhattan
12192	ENJOY Downtown NYC!	2019-11-13	https://www.airbnb.com/rooms/12192	https://a0.muscache.com/im/pictures/93658190/67480448_original.jpg?aki_policy=large	Manhattan
15711	2 bedroom - Upper East Side.	2019-11-24	https://www.airbnb.com/rooms/15711	https://a0.muscache.com/im/pictures/c444be23-6b18-4094-b2c7-cep811c08c9c_inn?aki_policy=large	Manhattan

Airbnb Features

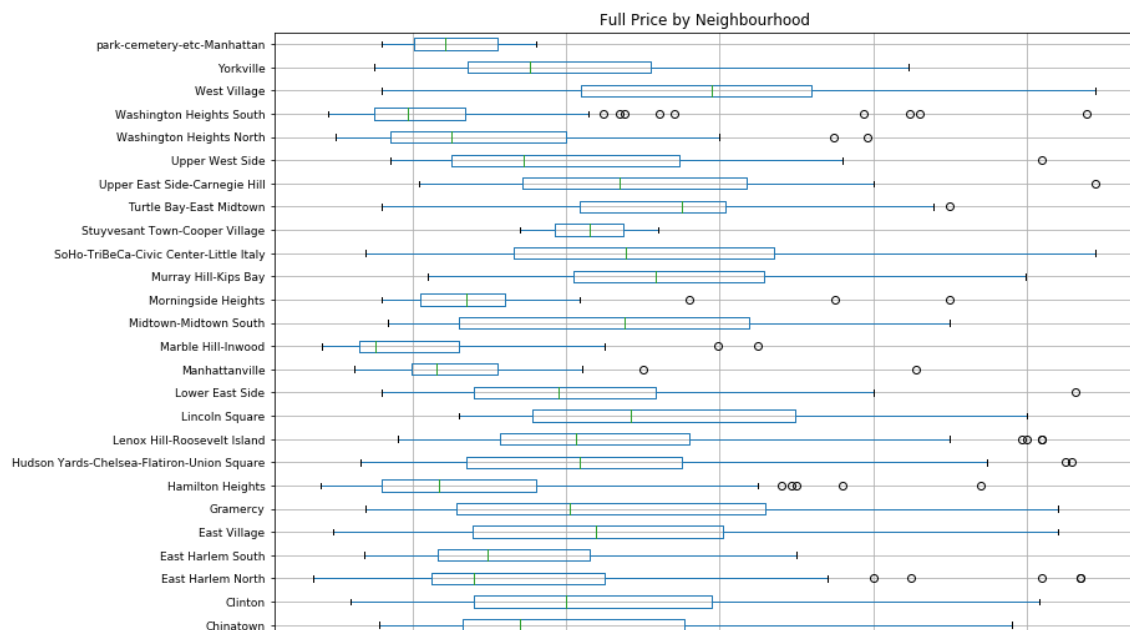
Now we are going to add some new features (columns) to our Airbnb dataset:

- ❖ 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. This calculation takes 25-50 minutes in Python.

Check for outliers for the Full price column:



In the chart above we can see some outliers - observation points that are distant from other observations. In our case they are the indication of variance in the data. Let's remove these outliers via IQR method. After removing the outliers, a new Bar chart looks like this



Neighborhood Tabulation Areas

'Neighborhood Tabulation Areas.geojson' file contains only polygon area coordinates for each Neighborhoods. So, we need to

- remove neighborhoods outside Manhattan
- calculate Centroid points ('latitude', 'longitude') for each Neighborhood.

For used Nominatim service to detect Centroid points for each Neighborhood and then made some manual correction because Nominatim service is not quite accurate.

New York Police Crime Records

We are interested in crimes only for Manhattan and types are 'Felony' or 'Misdemeanor'. After filtering we have 101,086 crimes records for Manhattan in 2019 year. We keep all columns, but it's needed to convert Latitude and Longitude columns from String to Float. 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. And removed 'Not defined' values for tab_area column as they do not belong to Manhattan.

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:

- ❖ select Top-100 Airbnb accommodations based on summary rating, number of crimes and price per person
- ❖ 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

In this section, we will explore the cleansed data and visualize them. Then, we will conduct cluster analysis to try to classify Manhattan's NYC Airbnb Neighborhoods.

Average Price per Person Neighborhoods

Calculate average price_per_person, average crimes rate and number of accommodations for each Airbnb neighborhoods.

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 accommodations
- Washington Heights South - 46.79 USD - 82 accommodations
- Washington Heights North - 54.74 USD - 53 accommodations
- Central Harlem North-Polo Grounds - 57 USD - 132 accommodations
- Manhattanville - 59.75 USD - 25 accommodations

Crime Rate Neighborhoods

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:

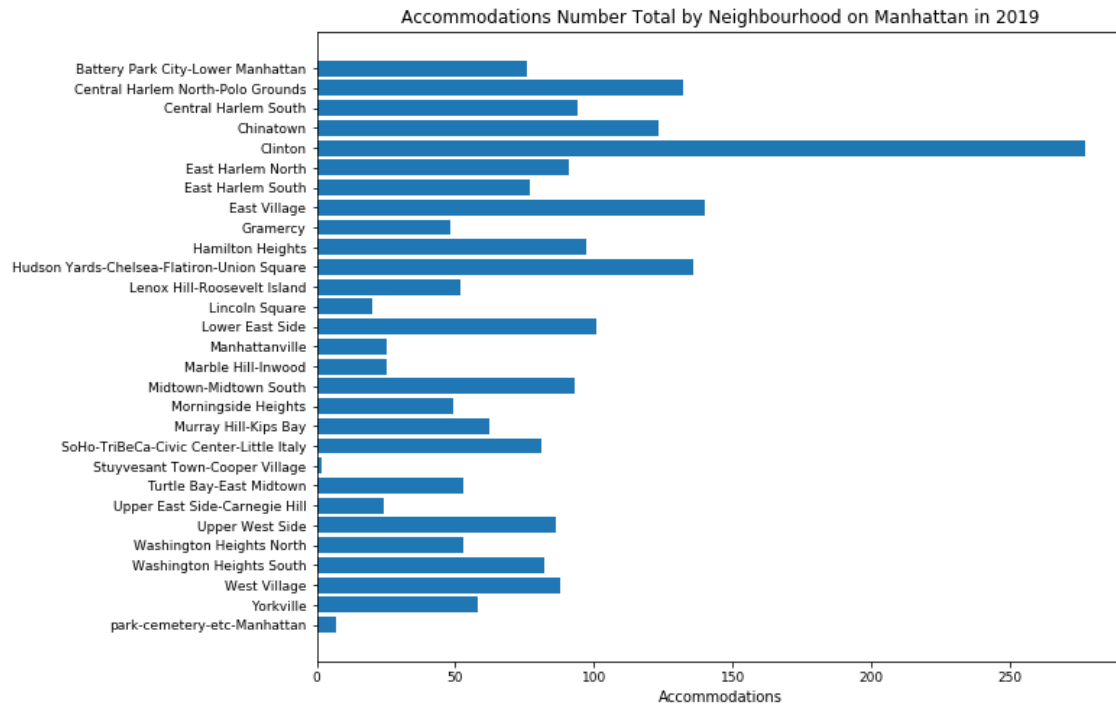
- Stuyvesant Town-Cooper Village - 145
- park-cemetery-etc-Manhattan - 1,213
- Lenox Hill-Roosevelt Island - 1,604
- Manhattanville - 1,832
- Yorkville - 1,898

All data

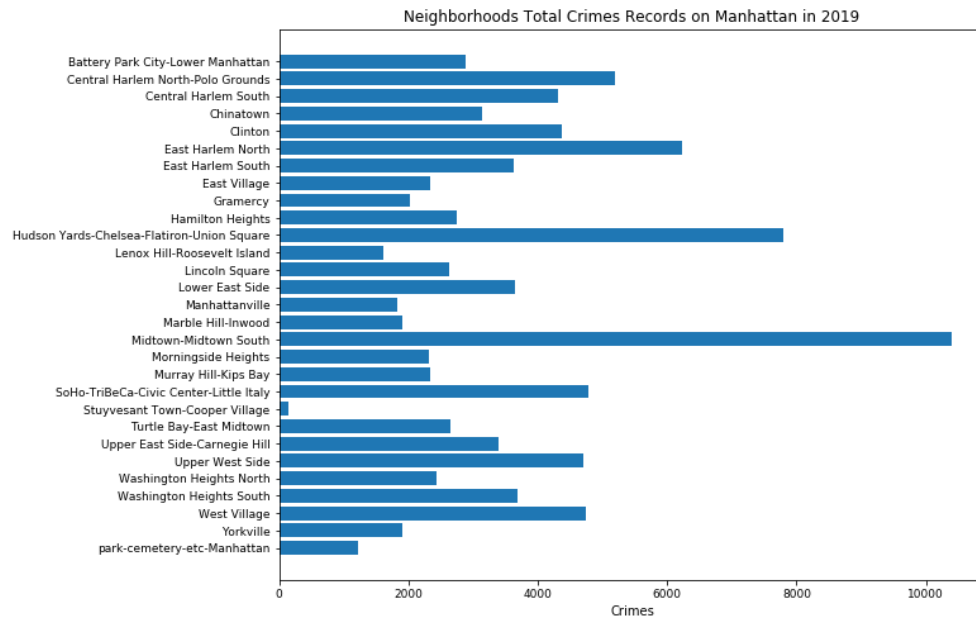
	tab_area	mean_price_per_person	accommodates	mean_crimes
15	Marble Hill-Inwood	45.475238	25	54.440000
25	Washington Heights South	46.793172	82	57.682927
24	Washington Heights North	54.738050	53	50.886792
1	Central Harlem North-Polo Grounds	57.038497	132	73.265152
14	Manhattanville	59.746667	25	70.520000
9	Hamilton Heights	60.069404	97	66.206186
28	park-cemetery-etc-Manhattan	61.404762	7	13.428571
5	East Harlem North	62.275497	91	104.824176
6	East Harlem South	64.950000	77	90.051948
2	Central Harlem South	67.000823	94	87.148936
17	Morningside Heights	71.386054	49	56.204082
23	Upper West Side	78.127907	86	57.813953
16	Midtown-Midtown South	81.491705	93	139.000000
0	Battery Park City-Lower Manhattan	82.133130	76	90.394737
3	Chinatown	82.362703	123	93.081301
4	Clinton	84.308819	277	84.314079
7	East Village	84.842758	140	74.385714

NYC Manhattan's Neighborhoods Analysis Charts

Apartments Total by Neighborhood Chart



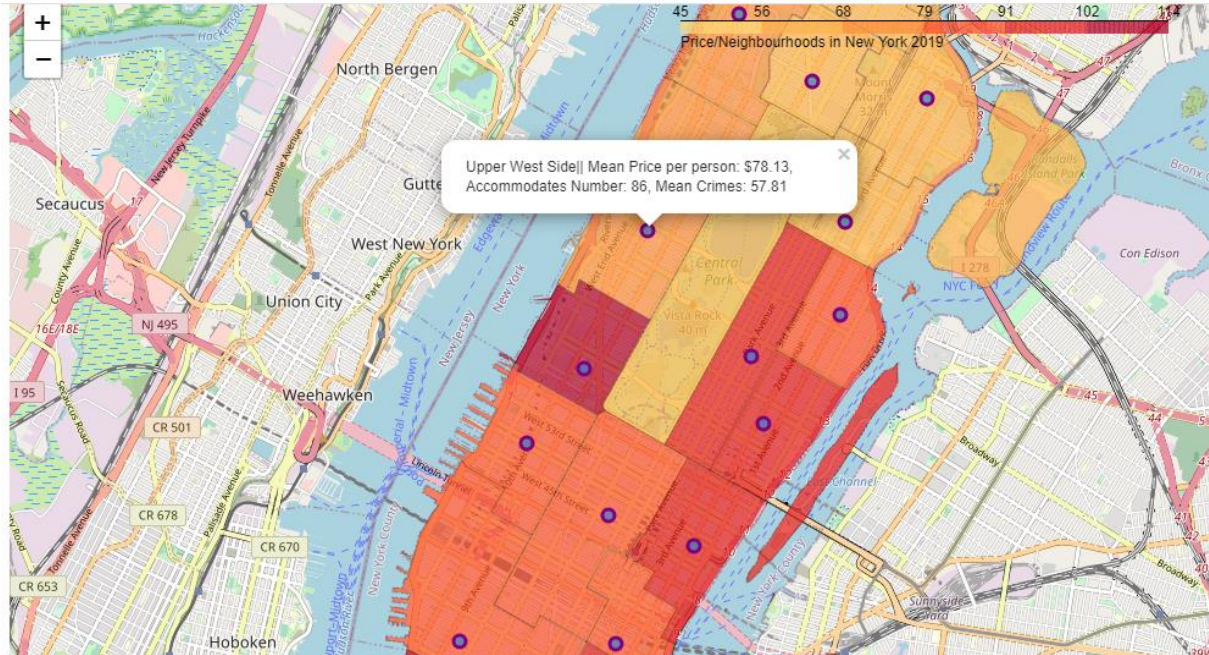
Neighborhoods Crimes Records Chart



NYC Manhattan's Neighborhoods Analysis Maps

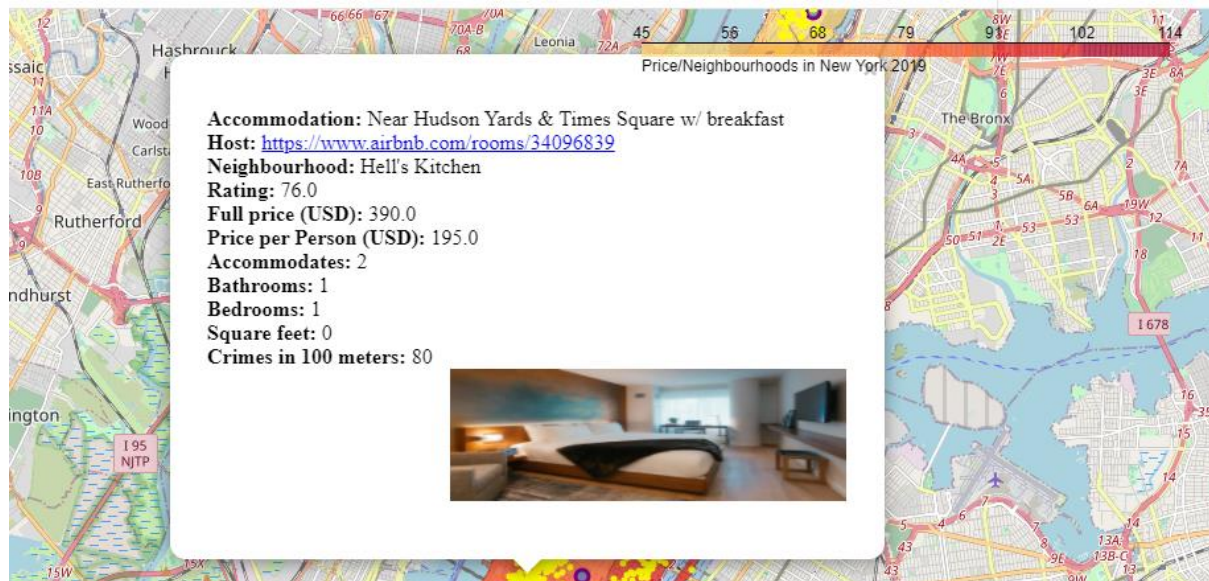
We created several maps to visualize our results and help the user select needed lodges.

NYC Tabulation Area Neighborhoods Average Prices per Person on Manhattan in 2019 Map



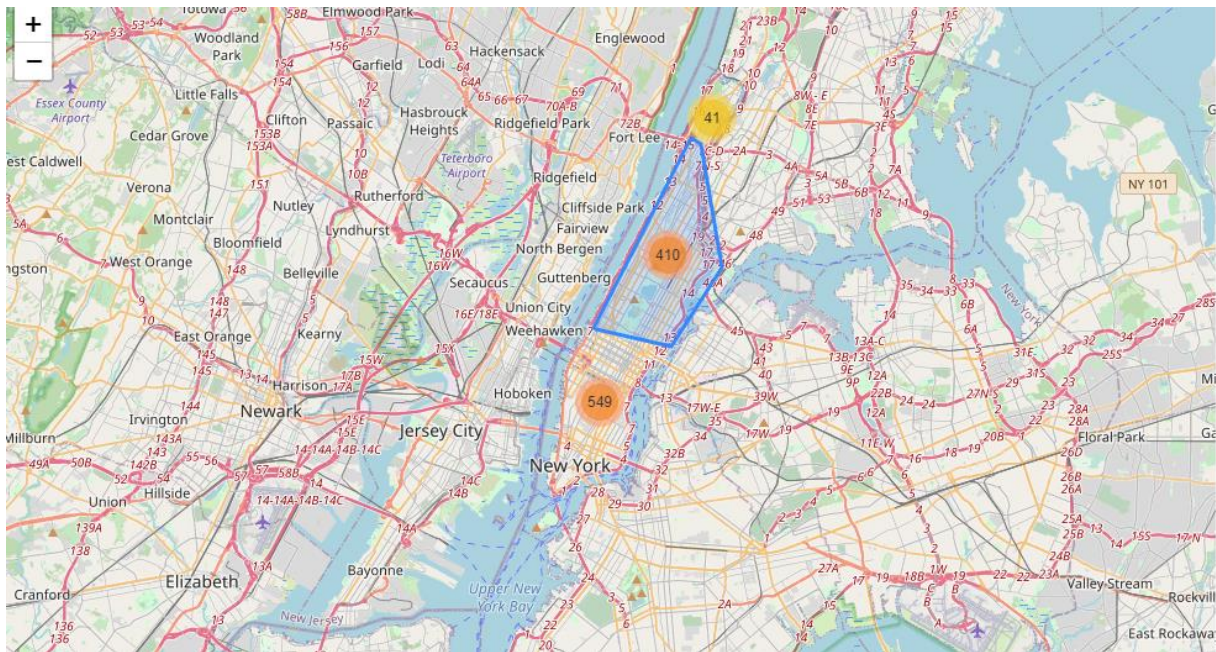
Accommodations Detailed Info Map

We put all our Airbnb accommodations on the Map - so the user can easily review them.



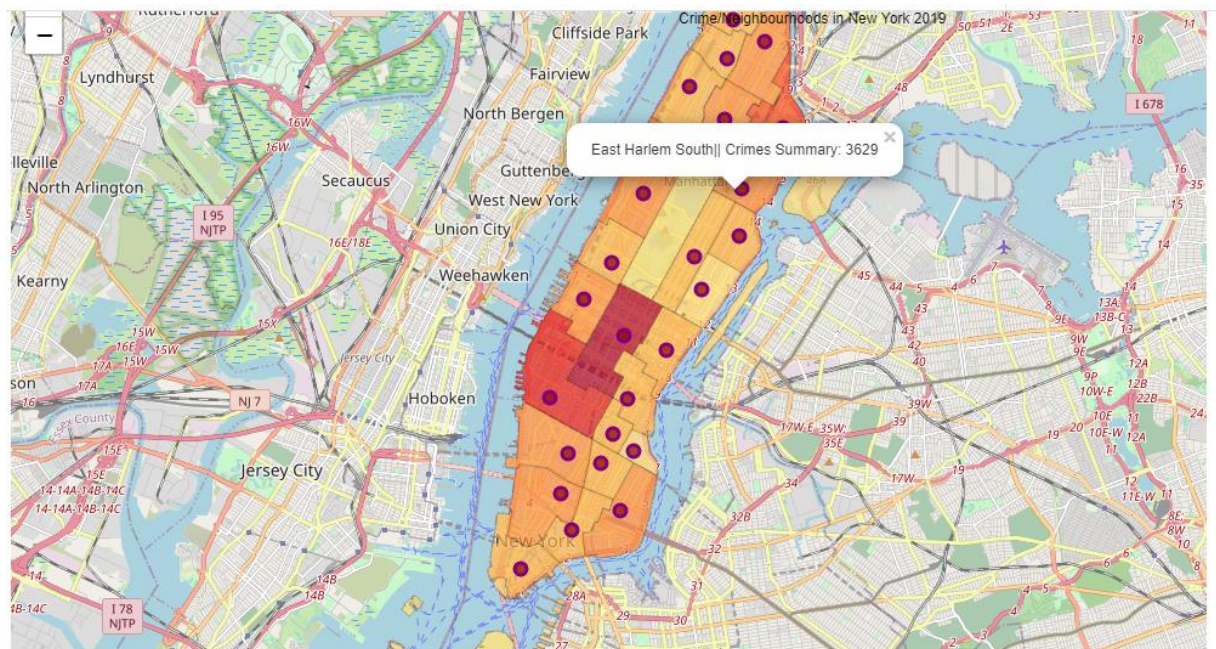
Crimes Cluster Map

Cluster Map of 1000 crimes by Neighborhood from New York Police Crime data set.



Crimes by Neighborhoods Map

In this map we depict the summary number of crimes for each neighborhood on Manhattan for 2019 year.



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.

Let's choose Top 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

	name	tab_area	neighbourhood_cleansed	latitude	longitude	review_scores_rating	property_type	room_type	accommod:
2029	Private Bedroom in Cozy Hamilton Heights Apartment	Hamilton Heights	Harlem	40.82749	-73.94461	100.0	Apartment	Private room	2
1801	Mr. B - Room Apartment in NYC	Washington Heights South	Washington Heights	40.84377	-73.94094	100.0	Apartment	Private room	1
1354	Cozy sanctuary/Big master bedroom/NYC/The Heights	Washington Heights North	Washington Heights	40.84911	-73.93097	100.0	Apartment	Private room	2
1529	Little Safe Haven	Hamilton Heights	Harlem	40.82494	-73.94280	100.0	Apartment	Private room	1
654	One cozy private BR close to the mecca of shopping	Turtle Bay-East Midtown	Midtown	40.76026	-73.96590	100.0	Apartment	Private room	1

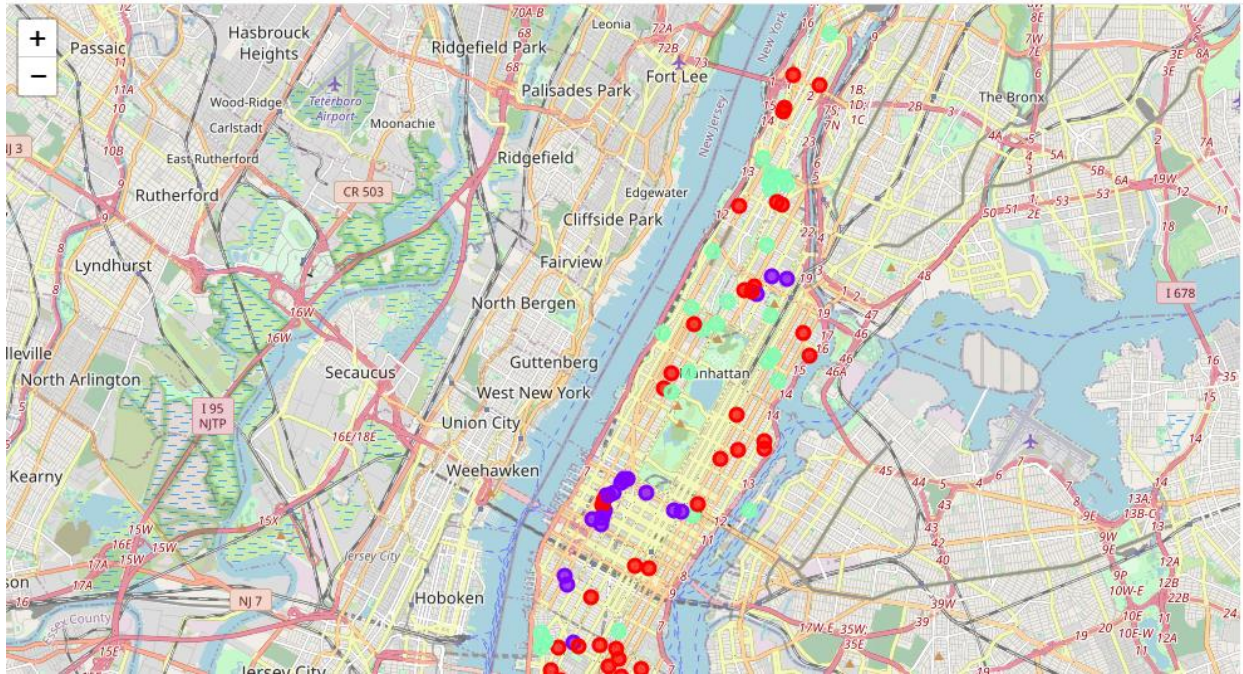
We define our custom Top-Level categories for Venues

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
0	1	**Stylish, Quiet, Centrally Located (9th & 52nd)	Food Place	Entertainment	Shopping	0.65	0.31	0.04
1	2	157-C	Food Place	Sightseeing	Fine Art	0.64	0.2	0.12
2	2	A neat bedroom in a cozy 3-bedroom apartment	Sightseeing	Food Place	Shopping	0.47	0.43	0.03
3	0	Art filled peaceful paradise EV Union Square	Food Place	Entertainment	Sightseeing	0.67	0.13	0.07
4	0	Artsy Parisian Apt in Greenwich Village	Food Place	Entertainment	Sightseeing	0.71	0.19	0.1

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

	review_scores_rating	accommodates	full_price	price_per_person	crimes	bathrooms	bedrooms	Cluster Labels
count	45.0	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.0
mean	100.0	2.133333	214.422222	105.602963	57.577778	1.000000	1.000000	0.0
std	0.0	0.919486	107.150168	41.601995	33.132303	0.213201	0.369274	0.0
min	100.0	1.000000	67.000000	37.000000	3.000000	0.000000	0.000000	0.0
25%	100.0	2.000000	140.000000	75.000000	32.000000	1.000000	1.000000	0.0
50%	100.0	2.000000	191.000000	98.333333	51.000000	1.000000	1.000000	0.0
75%	100.0	2.000000	274.000000	130.000000	69.000000	1.000000	1.000000	0.0
max	100.0	5.000000	510.000000	210.000000	164.000000	2.000000	2.000000	0.0

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

	review_scores_rating	accommodates	full_price	price_per_person	crimes	bathrooms	bedrooms	Cluster Labels
count	20.0	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.0
mean	100.0	2.650000	273.250000	109.702500	92.250000	1.050000	0.950000	1.0
std	0.0	1.089423	99.181851	39.040655	60.596053	0.223607	0.604805	0.0
min	100.0	1.000000	110.000000	55.000000	23.000000	1.000000	0.000000	1.0
25%	100.0	2.000000	171.250000	79.375000	54.500000	1.000000	1.000000	1.0
50%	100.0	2.000000	277.500000	102.500000	72.500000	1.000000	1.000000	1.0
75%	100.0	4.000000	341.750000	138.125000	117.000000	1.000000	1.000000	1.0
max	100.0	5.000000	420.000000	200.000000	257.000000	2.000000	2.000000	1.0

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

	review_scores_rating	accommodates	full_price	price_per_person	crimes	bathrooms	bedrooms	Cluster Labels
count	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.0
mean	99.942857	2.942857	205.542857	76.064286	78.257143	1.200000	1.114286	2.0
std	0.235504	1.814044	130.237869	50.626787	72.560053	0.472789	0.471008	0.0
min	99.000000	1.000000	54.000000	27.000000	4.000000	1.000000	0.000000	2.0
25%	100.000000	2.000000	90.000000	47.250000	36.000000	1.000000	1.000000	2.0
50%	100.000000	2.000000	158.000000	64.000000	63.000000	1.000000	1.000000	2.0
75%	100.000000	4.000000	275.000000	74.500000	89.000000	1.000000	1.000000	2.0
max	100.000000	10.000000	519.000000	259.500000	385.000000	3.000000	3.000000	2.0

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.

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.

Top Neighborhoods Statistics

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

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

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

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.

References

- https://en.wikipedia.org/wiki/Tourism_in_New_York_City
- https://assets.simpleviewinc.com/simpleview/image/upload/v1/clients/newyorkcity/FYI_Hotel_reports_February_2019_8607015b-b32a-4c7f-9fbd-84cd2a93cbe6.pdf
- https://aka.nyc/content/uploads/2017/12/new_york_city_travel_and_tourism_trend_report_2017.pdf