Coursera Capstone Project

The Battle of Neighborhoods – Report

Table of Content

1 Introduction

Description of the problem and a discussion of the background:

New York is one of the cultural and financial centers of the USA. Furthermore, it is a flagship for the "American Way of Life", which is visited by millions of people every year. For this reason, New York is of crucial importance for tourism and for the worldwide reputation of the USA.

In the year 2018 New York City welcomed a new record of over 65 million visitors. According to analysis company Smith Travel Research the hotel occupancy rate of New York City rose to 87.3 percent, which was more than comparable cities like Paris and Berlin (both around 75 percent).

Currently there are over 115,530 hotel rooms in over 630 hotels in the five boroughs of New York City. Most of these rooms, around 80 percent, are in Manhattan. Since 2010, the New York City hotel market had a 42 percent growth in new hotel rooms. Most of this growth has happened in areas outside of Manhattan, recently creating well-established hotel districts in areas of Brooklyn and Queens.

In this environment I am the Business Analyst of a hotel company looking for a location for a new hotel in New York City.

Business Problem:

As seen in the problem description, New York City continues to be a worthwhile environment to build and operate a new Hotel. Since most hotels are in Manhattan and the costs for a new building there are very high, the goal is to analyze the possibilities for a hotel in Brooklyn. Legal problems and non-available construction site do not play a role at this point of the planning yet.

For tourists and businesspeople there are a number of requirements for a hotel. The following demands will be the most important drivers for a successful hotel:

- Proximity to Manhattan
- Access to public transportation
- Presence of services and amenities like restaurants and cultural sights in neighborhood
- Significant office or commercial markets
- Existing number of hotels in neighborhood

Furthermore, the Real Estate prices continue to play an important role for the success of a new hotel and will be part of the analysis.

Ultimately, the Business Problem is in which neighborhood a new hotel would have the greatest success. Also, what neighborhood is the worst for a new hotel and what category would be the best? Therefore, we will look at the demands for a successful hotel and analyze which neighborhood offers the best prerequisites for a new hotel. Interested Audience would be Hotel Operators and Investors but also other businesses that benefit/depend from hotels and want to plan for the future.

2 Data

Description of the Data:

The following information is required to answer the issues of the problem:

- List of neighborhoods of Brooklyn with their geodata (latitud and longitud)
- List real estate prices for each neighborhood of Brooklyn
- Proximity to the Center of Manhattan for each neighborhood of Brooklyn
- Venues like restaurants and cultural sights for each neighborhood of Brooklyn

Therefore we need the follwing data:

- New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude.
    - Data source: https://cocl.us/new_york_dataset
- Restaurants, hotels, Subway stations and cultural sights.
    - Data source: Fousquare API
- GeoSpace data to get the New york Borough boundaries that will help us visualize choropleth map.
    - https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm
- New York Real Estate Prices by neighborhood.
    - https://streeteasy.com/blog/q1-2019-market-reports/

How the data will be used to solve the problem

The approach will be as follows:

- Collect the New York city data from https://cocl.us/new_york_dataset
- Use Foursquare and geopy data to find and map venues for all Brooklyn neighborhoods and clustered in groups
- Filter out all venues that are important for tourists and businesspeople
- create a map that shows the average rental price for all Brooklyn neighborhoods
- Cluster neighborhoods by venues
- Rank the Brooklyn neighborhoods by best usage for a new hotel

The data will allow to answer the key questions to make a decision:

- What is the cost for each Brooklyn neighborhood?
- Which Brooklyn neighborhood has restaurants and cultural sights?
- Are there tradeoffs between Proximity to the center of Manhattan and price and location?

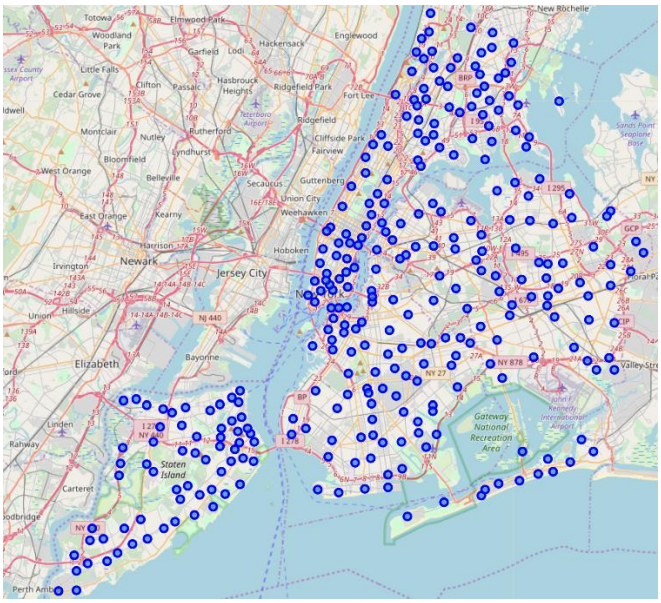
Example of the raw Dataset:

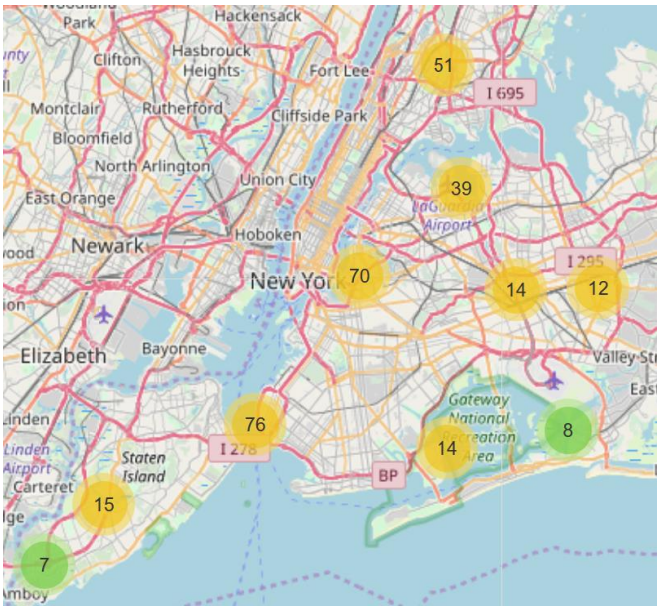
Borough	Neighborhood	Latitude	Longitude
Brooklyn	Coney Island	40.574293	-73.988683
Brooklyn	Manhattan Beach	40.577914	-73.943537
Brooklyn	Sea Gate	40.576375	-74.007873
Brooklyn	Brighton Beach	40.576825	-73.965094
Brooklyn	Sheepshead Bay	40.586890	-73.943186

The dataframe has 5 boroughs and 306 neighborhoods.

- The borough Bronx has 52 neighborhoods.
- The borough Manhattan has 40 neighborhoods.
- The borough Brooklyn has 70 neighborhoods.
- The borough Queens has 81 neighborhoods.
- The borough Staten Island has 63 neighborhoods.

Overview-Maps over the neighborhoods of New York (left: each neighborhood, right: mapped):





3 Methodology

This section represents the main component of the report where the data is gathered, prepared for analysis. The tools described are used here and the Notebook cells indicates the execution of steps.

The analysis and the strategy:

The strategy is based on mapping the above described data, in order to facilitate the choice of a place for a hotel. The choice is made based on the demands imposed: location near a Manhattan Center, price and venues close to the neighborhood. This visual approach and maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy.

First Step is to calculate the distance between each neighborhood and the Center of Manhattan which is the place all tourists and businessmen want to visit. An analysis of tripadvisor shows that the Empire State building is the most important and most visited building in Manhattan and be the center of Manhattan in our Example.

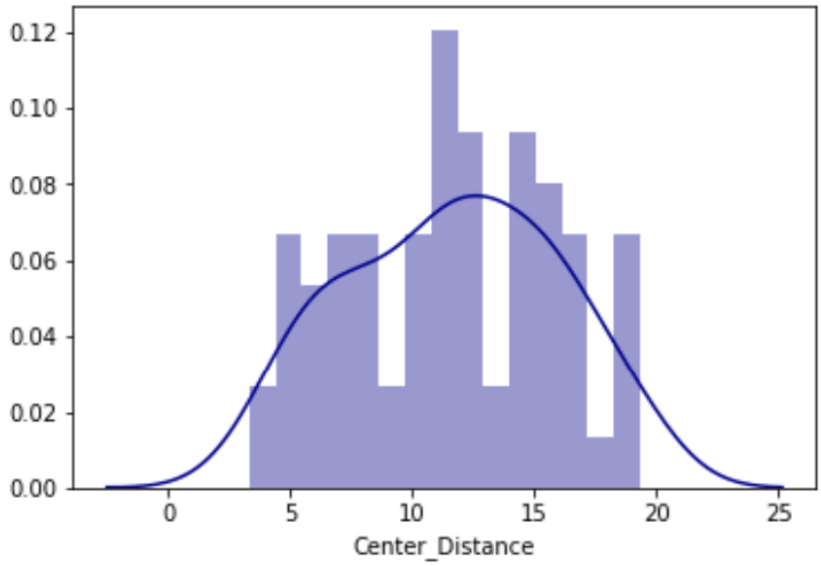
Dataset example with Distance to Center of Manhattan and Real Estate Prices:

Borough	Neighborhood	Latitude	Longitude	Center_Distance	Median Recorded Sales Price
Brooklyn	Bay Ridge	40.625801	-74.030621	14.138050	615000.0
Brooklyn	Bensonhurst	40.611009	-73.995180	15.281355	1140000.0
Brooklyn	Sunset Park	40.645103	-74.010316	11.661878	1100000.0
Brooklyn	Greenpoint	40.730201	-73.954241	3.337298	1609143.0
Brooklyn	Gravesend	40.595260	-73.973471	17.040144	830000.0

Info: From now on we use only the Borough Brooklyn!

Data Analysis

Histogram for the Distance of the neighborhoods to the Center of New York for Brooklyn:

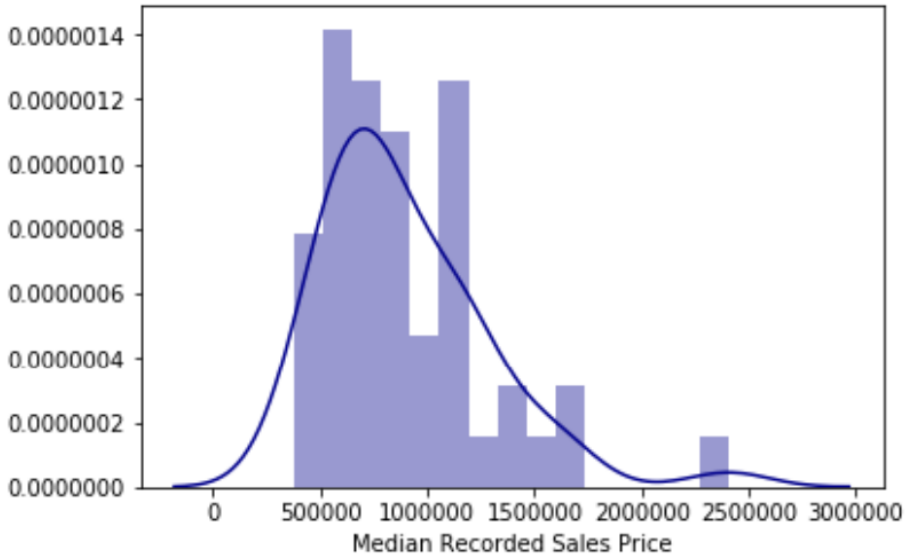


Descriptive Statistics of Center_Distance:

```
count    48.000000
mean     11.521195
std       4.631740
min       3.337298
25%       7.574486
50%      11.723215
75%      15.147785
max      19.339022
Name: Center_Distance, dtype: float64
```

It is obvious that the Distances to the Center are divers in Brookyn. Some hoods are only 3.3 km away of the Center and others are 19 km away of the city. For our analysis the hoods close to the City are more important since tourist look at the distance to Manhattan first.

Histogram for the Median Sales Prices of Brooklyn neighborhoods:

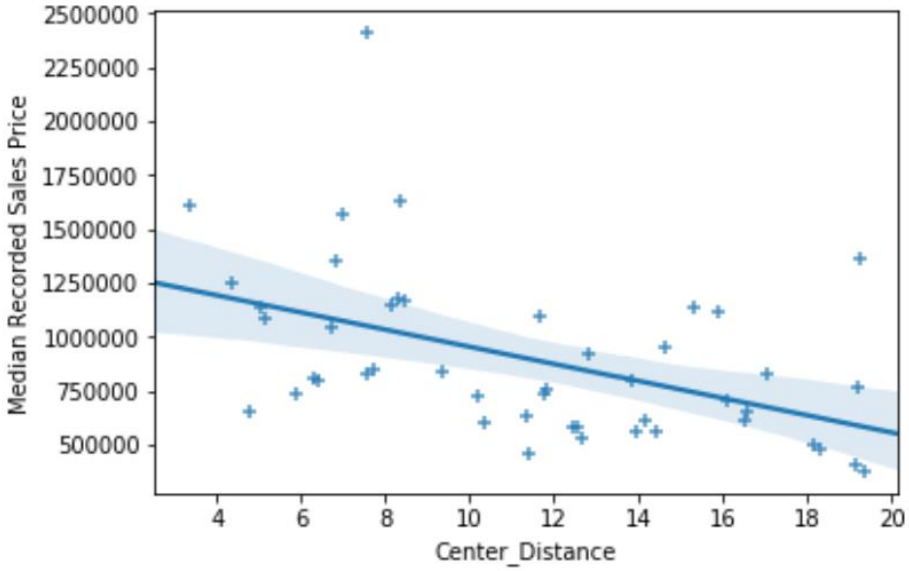


New York is expensive. So is Brooklyn. The median recorded sales prices go from 450 thousand dollars to around 3 Billion dollars. For our analysis the neighborhood with lower prices are more important since high costs for buying the hotel will reduce the profit of the hotel.

Descriptive Statistics of Sales Prices:

```
count    4.700000e+01
mean     8.985091e+05
std      3.910539e+05
min      3.800000e+05
25%      6.150000e+05
50%      8.000000e+05
75%      1.129240e+06
max      2.406250e+06
Name: Median Recorded Sales Price, dtype: float64
```

Correlation between the Sales Price and the Distance to the Center:



It becomes clear, that there is a negative Correlation between the Sales Price and the Distance to the Center. The further away the neighborhoods are, the cheaper the Sales Prices will be. So there is a decision between a hotel that is close to the Center and a hotel that is cheaper.

Analysis and Cluster of the Venues for each neighborhood

We analyze both boroughs neighborhoods through one hot encoding (giving '1' if a venue category is there, and '0' in case of venue category is not there). On the basis of one hot encoding, we calculate mean of the frequency of occurrence of each category and picked top ten venues on that basis for each neighborhood. It means the top venues are showing the foot traffic or the more visited places.

For the Analysis we use a Limit of 20 venues per neighborhood and a Radius of 500m around the center of each neighborhood.

All in all there are 1227 vanues in Brooklyn and 221 categories are unique.

Example for the Dataframe with Vanue Categories:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar
Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
Bay Ridge	40.625801	-74.030621	Ho' Brah Taco Joint	40.622960	-74.031371	Taco Place

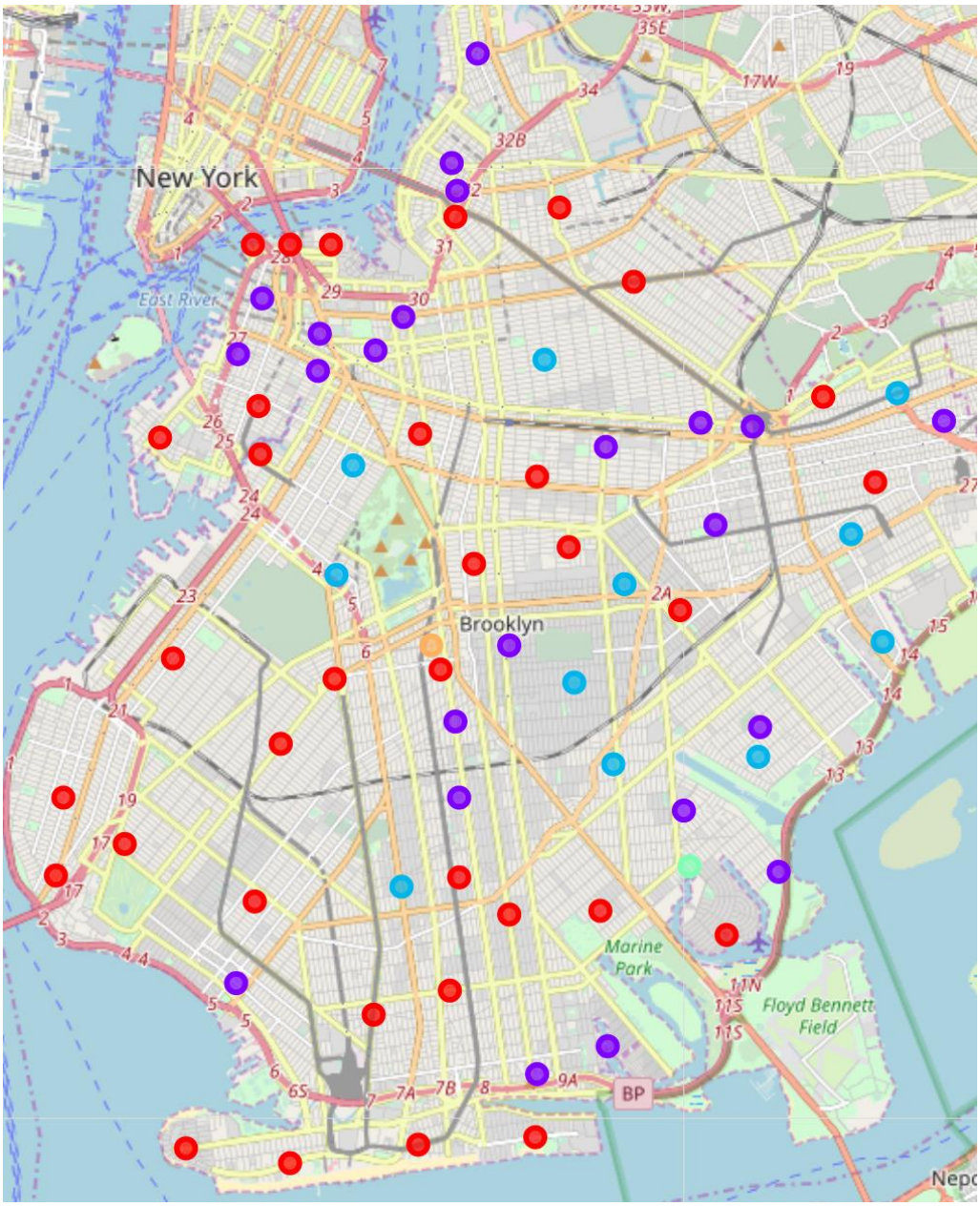
Example for a neighborhood (Bath Beach) top 5 most common venues:

```
----Bath Beach----
                venue  freq
0  Italian Restaurant  0.10
1     Bubble Tea Shop  0.10
2          Restaurant  0.05
3        Dessert Shop  0.05
4   German Restaurant  0.05
```

Furthermore we can analyze the most common words for the venue category of the Brooklyn neighborhood data:

It is clear that Restaurant is the most common name of the vanues. Other common words are Shop, Bar, Café, and different variations of the word restaurant. So is becomes clear, that Brooklyn is a place where you can get food almost everywhere within 500 meters.

For the clustering we use 5 clusters which are presented in the following map:



4 Results

All in all we got 5 different clusters. The count of the Clusters is:

Cluster Labels	count
0	34
1	23
2	11
3	1
4	1

Most common Clusters are Cluster 0 and Cluster 1. Number 3 and 4 are only one time available.

Cluster 0 is the tourist Cluster. We have Museums, a lot of bars and a lot of different restaurants. Cluster 1 is the living Place. There are a lot of everyday locations like Cafes and Sport Studios. Cluster 2 is the Commercial Area with Pharmacies, Shops and Stores. Cluster 3 is a restaurant cluster and Cluster 4 is a shopping Cluster.

In the final Database we have the clusters, the Center Distance and the Median Sales Prices:

	Neighborhood	Cluster Labels	Borough	Latitude	Longitude	Center_Distance	Median Recorded Sales Price
0	Bay Ridge	0	Brooklyn	40.625801	-74.030621	14.138050	615000.0
1	Bensonhurst	0	Brooklyn	40.611009	-73.995180	15.281355	1140000.0
2	Sunset Park	0	Brooklyn	40.645103	-74.010316	11.661878	1100000.0
3	Greenpoint	1	Brooklyn	40.730201	-73.954241	3.337298	1609143.0
4	Gravesend	0	Brooklyn	40.595260	-73.973471	17.040144	830000.0

So we are looking for a Neighborhood in Cluster 1 with a short distance to the center and a low Sales Price. If we only keep neighorhoods with a center distance of less than 7 km and a sales price of less than 1000000 $ than there are 4 neighborhoods left. The best of those are Brooklyn Heights and South Brooklyn.

Dataframe of the last remaining neighborhoods:

	Neighborhood	Cluster Labels	Borough	Latitude	Longitude	Center_Distance	Median Recorded Sales Price
17	Brooklyn Heights	1	Brooklyn	40.695864	-73.993782	5.877482	735000.0
34	Clinton Hill	1	Brooklyn	40.693229	-73.967843	6.311790	807500.0
36	Downtown	1	Brooklyn	40.690844	-73.983463	6.397338	800000.0
43	South Side	1	Brooklyn	40.710861	-73.958001	4.781324	660000.0

5 Discussion

As a Data Analyst I would recommend the hotel company to build a hotel in Brooklyn Heights or in Brooklyn South because of the right Cluster Category for tourists the Distance to the Center and the comparable low sales prices.

Because of the Limitation of the Data, the analysis is only incomplete. Further research would be necessary to get a bigger picture of the hotel market in Brooklyn and in overall New York.

6 Conclusion

Even though Data was limited we found two great neighborhoods for a potential hotel.

Clear evidence were found, that the prices increase with a lower distance to the Center of Manhattan.

All in all the Capstone Project was interesting and fun to play with the data.