Capstone Project: Finding a location

Applied Data Science Capstone by IBM/Coursera

Table of contents

- Introduction
- Data
- Methodology
- Analysis
- Discussion
- Conclusion

Introduction

In this project we will try to find an optimal location to rent a studio apartment in New York City. Specifically, this report will be addressed to **those students who need to rent a room to access their classes at Columbia University in New York City**.

As there are many studio apartments in New York City, we will try to detect places considering 3 factors; crime rate, rent cost and finally, proximity to places of preference. We would also prefer locations as close as possible to the university, assuming the first three conditions are met.

We will use data science to generate some more promising neighborhoods based on this criterion. The advantages of each area will be clearly expressed so that those interested can choose the best possible final location.

Data

Based on definition of our problem, factors that will influence our decision are:

- Number of crimes committed in each county of New York City
- Average rental cost of a studio apartment
- Proximity to places of preference

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

- Number of total crimes per county, for which the database of **Open Data** will be used "NYPD Complaint Data Current (Year To Date)" 'https://data.cityofnewyork.us/api/views/5uac-w243/rows.csv'

- The exact middle asking rent among all rental listings available on **StreetEasy** at any point during the month/quarter/year. In general, median values are more accurate than average values, which may be skewed by price outliers (a few rentals that are extremely expensive or extremely inexpensive).

  https://streeteasy-market-data-download.s3.amazonaws.com/rentals/Studio/medianAskingRent_Studio.zip

- Number of preference places and their type and location in every neighborhood will be obtained using **Foursquare API.**

Methodology

1. Crimes in NYC

1.1. Downloading and Prepping Data

After importing the libraries necessary for reading the codes, the data is downloaded from the website of the New York Police Department and read as a dataframe.

Table 1. NYPD Complaint Data Current (Year To Date) 2017, Source Open Data

CMPLNT_NUM	ADDR_PCT_CD	BORO_NM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM
314773184	48	BRONX	12/31/2019	18:00:00		
289837961	25	MANHATTAN	12/30/2019	20:30:00	12/31/2019	10:00:00
535744284	77	BROOKLYN	12/24/2019	16:55:00	12/24/2019	17:00:00
895678119	52	BRONX	12/30/2019	19:32:00		
299841674	18	MANHATTAN	12/30/2019	15:30:00	12/30/2019	16:50:00
136697381	94	BROOKLYN	12/28/2019	13:00:00	12/29/2019	08:30:00
628084657	69	BROOKLYN	12/22/2019	16:30:00		
497138011	43	BRONX	12/29/2019	17:20:00		

As a result, a table of 461711 rows and 35 columns is obtained, which consists of the following:

1. **CMPLNT_NUM**: Randomly generated persistent ID for each complaint
2. **ADDR_PCT_CD**: The precinct in which the incident occurred
3. **BORO_NM**: The name of the borough in which the incident occurred
4. **CMPLNT_FR_DT**: Exact date of occurrence for the reported event
5. **CMPLNT_FR_TM**: Exact time of occurrence for the reported event
6. **CMPLNT_TO_DT**: Ending date of occurrence for the reported event, if exact time of occurrence is unknown
7. **CMPLNT_TO_TM**: Ending time of occurrence for the reported event, if exact time of occurrence is unknown
8. **CRM_ATPT_CPTD_CD**: Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely

9. **HADEVELOPT**: Name of NYCHA housing development of occurrence, if applicable
10. **HOUSING_PSA**: Development Level Code
11. **JURISDICTION_CODE**: Jurisdiction responsible for incident
12. **JURIS_DESC**: Description of the jurisdiction code
13. **KY_CD**: Three digit offense classification code
14. **LAW_CAT_CD**: Level of offense: felony, misdemeanor, violation
15. **LOC_OF_OCCUR_DESC**: Specific location of occurrence in or around the premises
16. **OFNS_DESC**: Description of offense corresponding with key code
17. **PARKS_NM**: Name of NYC park, playground or greenspace of occurrence, if applicable
18. **PATROL_BORO**: The name of the patrol borough in which the incident occurred
19. **PD_CD**: Three digit internal classification code
20. **PD_DESC**: Description of internal classification corresponding with PD code
21. **PREM_TYP_DESC**: Specific description of premises; grocery store, residence, street, etc.
22. **RPT_DT**: Date event was reported to police
23. **STATION_NAME**: Transit station name
24. **SUSP_AGE_GROUP**: Suspect's Age Group
25. **SUSP_RACE**: Suspect's Race Description
26. **SUSP_SEX**: Suspect's Sex Description
27. **TRANSIT_DISTRICT**: Transit district in which the offense occurred.
28. **VIC_AGE_GROUP**: Victim's Age Group
29. **VIC_RACE**: Victim's Race Description
30. **VIC_SEX**: Victim's Sex Description
31. **X_COORD_CD**: X-coordinate for New York State Plane Coordinate System
32. **Y_COORD_CD**: Y-coordinate for New York State Plane Coordinate System
33. **Latitude**: Midblock Latitude coordinate for Global Coordinate System
34. **Longitude**: Midblock Longitude coordinate for Global Coordinate System
35. **Lat_Lon**: (Latitude,Longitude)

The original data was modified to facilitate the creation of visualizations, removing unnecessary columns, renaming identification and borough columns. In addition, all rows with NaN values were removed. Accordingly, the crime data grouped by borough is presented as follows.

Table 2. Crimes Dataframe

	Id	Latitude	Longitude
Borough			
BRONX	100994	100994	100994
BROOKLYN	132445	132445	132445
MANHATTAN	116352	116352	116352
QUEENS	92575	92575	92575
STATEN ISLAND	19019	19019	19019

1.2. Visualization

First, folium was imported. So, the dataframe consists of 461,385 crimes, which took place in the year 2017.

Let's just work with the first 100 incidents in this dataset and let's visualize where these crimes took place in the city of New York.

We use the default style and we will initialize the zoom level to 12. The latitude of New York City, NY, USA is 40.730610, and the longitude is -73.93524.

To visualize these crimes on the map of New York, we enter the loop code to add them to the map considering the latitude and longitude coordinates.

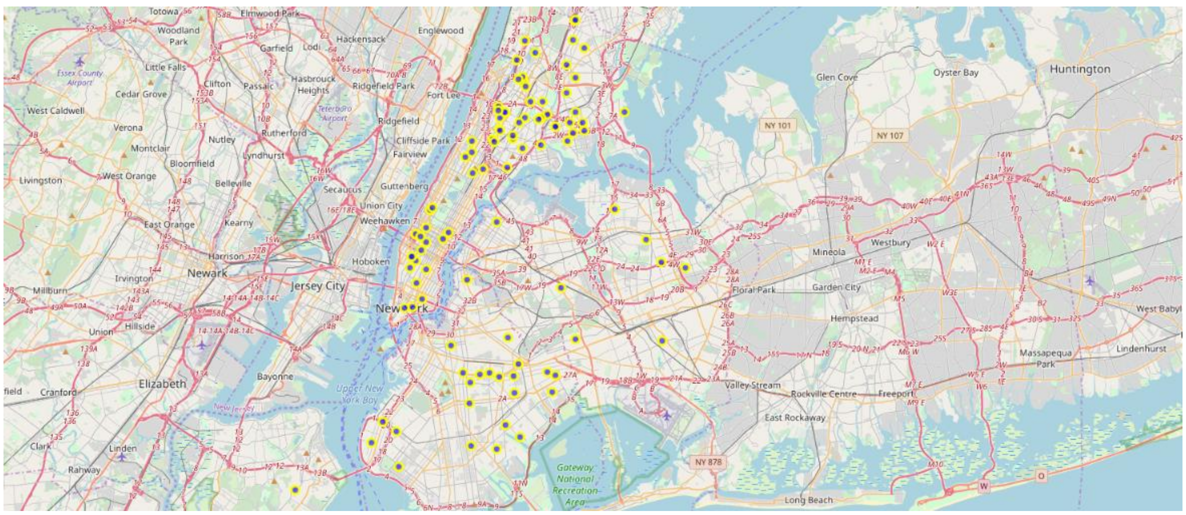

Figure 1. Crimes Map

2. Apartment rental prices

2.1. Downloading and Prepping Data

The data corresponding to the rental price of a studio type apartment was downloaded from the StreetEasy website in csv format and read as dataframe.

Table 3. StreetEasy Dataframe

	areaName	Borough	areaType	2010-01	2010-02	2010-03	2010-04	2010-05	2010-06	2010-07	...	2019-03	2019-04	2019-05	201 06
0	All Downtown	Manhattan	submarket	2350.0	2300.0	2200.0	2263.0	2300.0	2300.0	2290.0	...	2895.0	2900.0	2950.0	295
1	All Midtown	Manhattan	submarket	2000.0	1995.0	1995.0	2000.0	2000.0	2000.0	2050.0	...	2650.0	2650.0	2650.0	269
2	All Upper East Side	Manhattan	submarket	1750.0	1750.0	1750.0	1780.0	1800.0	1750.0	1750.0	...	2150.0	2150.0	2175.0	215

This data contains the average rental value from January 2010 to December 2019. In order to use the most up-to-date and representative data, dataset cleaning consisted of removing the columns from 2010 to 2018.

Next, the columns corresponding to the months of the year 2019 will be renamed with the month format, for example; Jan, Feb, Dec.

Finally, all NaN values were eliminated, grouping the remaining data by borough and adding a new column that calculated the average cost of rent of each borough. Cost prices were rounded in 1 digit.

Table 4. Average monthly cost of rent Dataframe.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Average
Borough													
Bronx	1420.5	1500.0	1462.5	1436.5	1544.0	1500.0	1544.0	1562.5	1575.0	1556.5	1550.0	1550.0	1516.8
Brooklyn	2046.4	2044.8	2064.4	2091.0	2155.4	2150.2	2162.6	2159.2	2166.7	2149.7	2121.1	2160.4	2122.7
Manhattan	2514.7	2551.8	2531.7	2563.5	2579.5	2600.1	2615.3	2599.6	2619.2	2649.6	2638.4	2639.2	2591.9
Queens	1681.6	1687.6	1692.5	1745.4	1747.0	1795.0	1783.7	1783.5	1794.9	1782.0	1768.8	1778.5	1753.4

Analysis

According to the data corresponding to the crime rate in New York, the following results can be obtained. For a better analysis, the crimes are presented in ascending order.

Table 5. Number of crimes by borough

	Number of crimes
Borough	
STATEN ISLAND	19019
QUEENS	92575
BRONX	100994
MANHATTAN	116352
BROOKLYN	132445

What is mentioned in the previous paragraph can be seen graphically in the following pie chart.

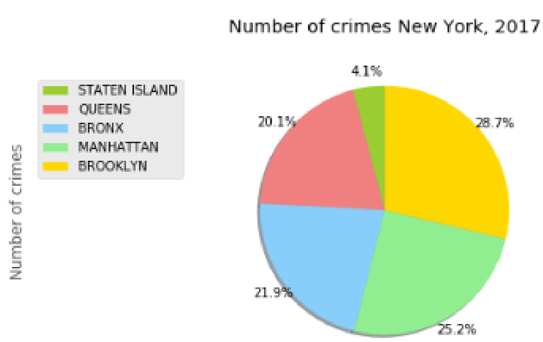

Figure 2. Number of crimes chart

As we already knew, Staten Island presents the lowest amount of crimes, represented by 4.1%, while the largest amount is obtained by Brooklyn with 132445 crimes, represented by 28.7%, which can be seen in the following pie chart.

With this in mind, let's analyze the rental cost of a studio apartment, according to the 2019 data.

The following graph shows the average monthly rental cost of a study in the Bronx, Brooklyn, Manhattan and Queens districts, expressed in USD / month.

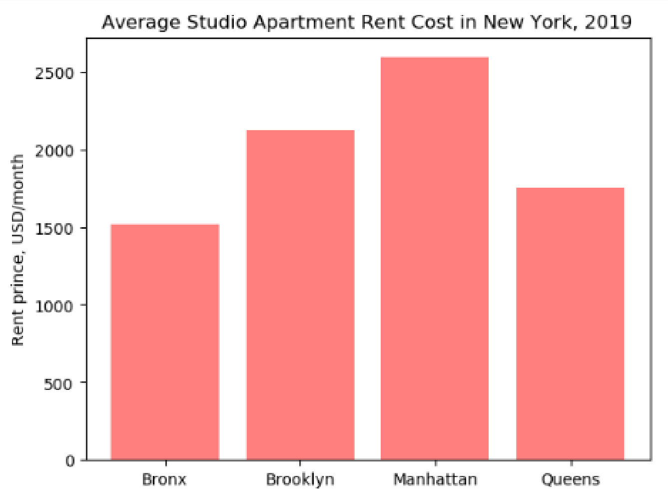

Figure 3. Average Sudio Apartment Rent Cost by borough.

Unfortunately, Staten Island is not in this list of bourghs, so even though it has the lowest crime rate, we have no way of relating the cost of living associated with the monthly rental of a studio apartment in Staten Island.

Considering the crime rate in Brooklyn, and the high costs of Manhattan, we will analyze the leisure activities offered by each alternative using the Foursquare API. Before we get the data and start exploring it, let's download all the dependencies that we will need.

The imported libraries are: json to handle json files and request to handle request.

In addition, geopy was installed to be able to import Nominatim and convert addresses into coordinate values (latitude, longitude)

Finally, for the clustering stage, KMeans was imported from sklearn.cluster

The files are placed on a server (https://cocl.us/new_york_dataset), so we can simply run a wget command and access the data.

Using the open code in the .json file, it is obtained that the attributes for each result are the following:


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
```

```
'annoline1': 'Wakefield',
'annoline2': None,
'annoline3': None,
'annoangle': 0.0,
'borough': 'Bronx',
'bbox': [-73.84720052054902,
 40.89470517661,
 -73.84720052054902,
 40.89470517661]}}
```

Based on the above information, a new dataframe is defined (neighborhoods), assigning as column names 'Borough', 'Neighborhood', 'Latitude' and 'Longitude'.

Using the geopy library, the latitude and longitude values of New York City were obtained. Next, a map of New York was generated, with neighborhoods superimposed on top.

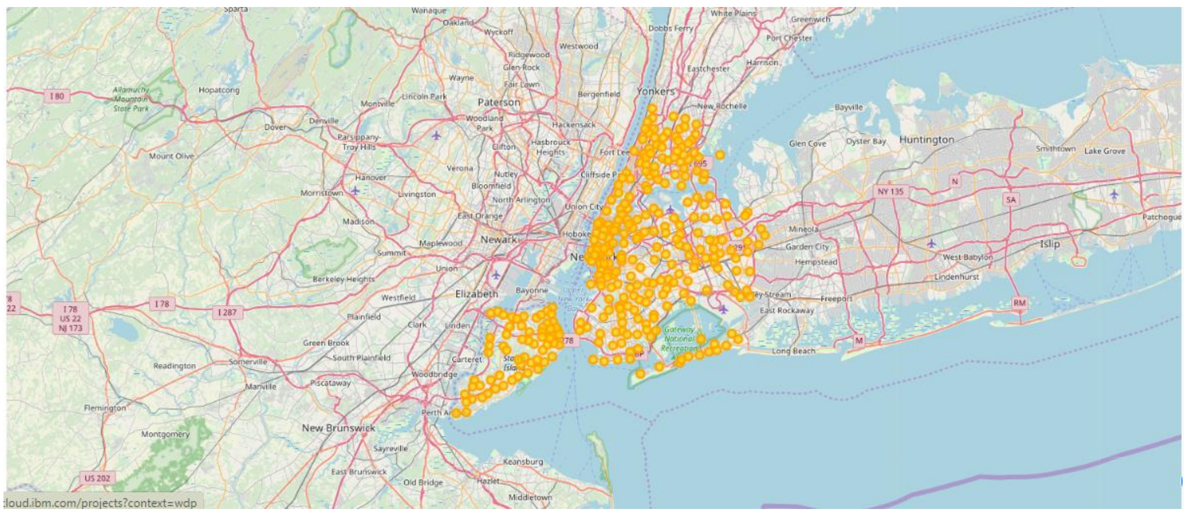

Figure 4. New York Neighborhoods Map

Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them. The limit of number of venues returned by Foursquare API is 200 and the radius is defined in 500. This segmetation, based on neighborhoods, was defined in a new dataframe, resulting in 10244 rows.

This segmetation, based on neighborhoods, was defined in a new dataframe, resulting in 10244 rows, presenting venues such as restaurants, bar, bus stop, pharmacy, among others, those that can be represented in 427 uniques categories.

In order to analyze each neighborhood, the one hot encoding code was used to subsequently group the rows by neighborhood and take the mean of the frequency of occurrence of each category.

Here is an example of the impression of the 5 most common places in each neighborhood

```
----Allerton----
          venue  freq
0      Pizza Place  0.16
1     Deli / Bodega  0.08
2      Supermarket  0.08
```

3 Chinese Restaurant 0.05
4 Department Store 0.05

These values were put in a dataframe, in which the 10 most common venues are presented by Neighborhood:

Table 5. 10 most common venues by neighborhood.

	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
0	Allerton	Pizza Place	Deli / Bodega	Supermarket	Chinese Restaurant	Department Store	Fast Food Restaurant	Martial Arts Dojo	Grocery Store	Gas Station
1	Annadale	Bakery	Pizza Place	Sports Bar	Pharmacy	American Restaurant	Restaurant	Train Station	Diner	English Restauran
2	Arden Heights	Pharmacy	Deli / Bodega	Bus Stop	Coffee Shop	Pizza Place	Women's Store	Field	Event Service	Event Space
3	Arlington	Bus Stop	Deli / Bodega	Boat or Ferry	Grocery Store	Women's Store	Fish Market	Exhibit	Factory	Falafel Restauran
4	Arrochar	Italian Restaurant	Deli / Bodega	Bus Stop	Polish Restaurant	Food Truck	Bagel Shop	Middle Eastern Restaurant	Outdoors & Recreation	Sandwich Place

The neighborhoods were clustered into 10, running k-means.

As results the following 10 clusters are obtained.

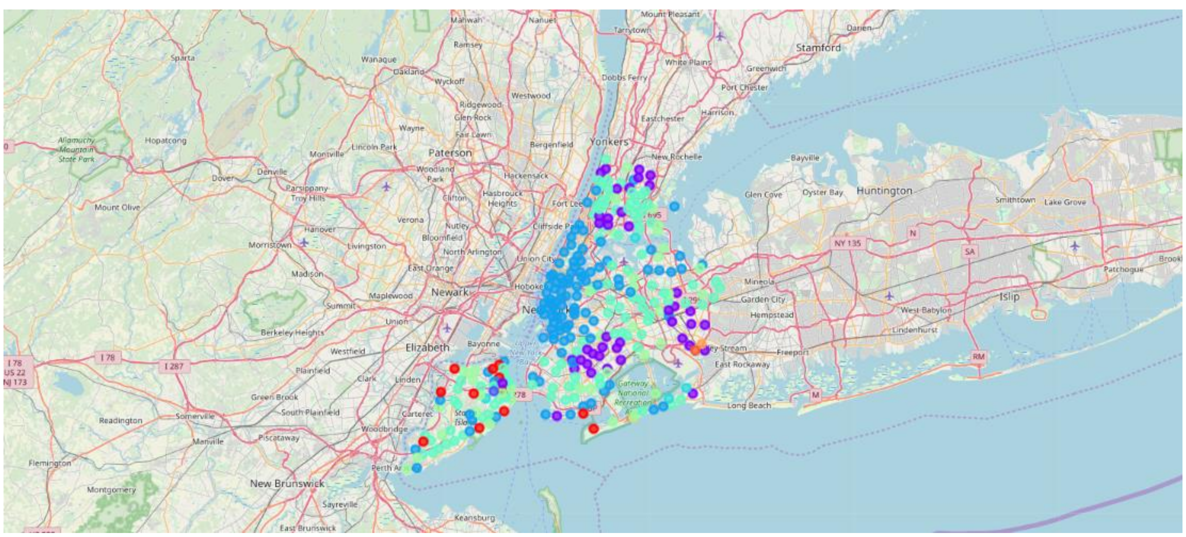

Figure 5. New York Neighborhoods Clusters Map

- **Cluster 1:** Bus Stop
- **Cluster 2:** Caribbean and Chinese Restaurant
- **Cluster 3:** Historic Site
- **Cluster 4:** Bar
- **Cluster 5:** Park
- **Cluster 6:** Pizza

- **Cluster 7:** Italian Restaurant
- **Cluster 8:** Beach
- **Cluster 9:** Caribbean Restaurant
- **Cluster 10:** Deli/Bodega

Discussion

After analyzing the data corresponding to each study factor, the following can be commented:

- The 5 boroughs analyzed show crime rates, of which Staten Island stands out, with an indicator of less than 5%. The rest of the boroughs have a similar percentage of crimes.
- Considering that Staten Island is far from the study point, it can be discarded from the analysis.
- Based on the rental cost of a study department, the lowest costs can be found in Bronx and Queens (1518.6 and 1753.4 USD / month respectively).
- Adding the factor of the availability of places to carry out different activities, it is possible to observe that in Manhattan it would be probable to find one kind of activities (bars), while the rest of the options offer greater type of variety, between food and parks.
- It could be expected that the best decision for a student looking to rent an apartment and attend classes at Columbia University is to live in Brooklyn.

Conclusion

It can be concluded that this method of analysis allows to visualize the existing options when making a decision.

It is important to mention that the quality of the dataset have a very important role in the output information. In these times of rapid changes, a better approximation could be obtained if the information corresponded to at least the last 2 years.

External factors such as personal tastes and particular opinions were not considered in this analysis, so the result will depend on the preferences of each student.