Coursera Capstone Project

Title: Comparison of living costs between Netherlands and New York

Angeliki Kalamara

15th of July 2020

1. Introduction

1.1 Scenario and background:

I am a data scientist currently residing in Utrecht, Netherlands. I currently live within walking distance to Downtown "Telok Ayer MRT metro station" therefore I have access to good public transportation to work. Likewise, I enjoy many amenities in the neighborhood , such as international cuisine restaurants, cafes, food shops and entertainment. I have been offered a great opportunity to work in Manhattan, NY. Although I am very excited about it, I am a bit stressed toward the process to secure a comparable place to live in Manhattan. Therefore, I decided to apply the learned skills during the Coursera course to explore ways to make sure my decision is factual and rewarding. Of course, there are alternatives to achieve the answer using available Google and Social media tools, but it is rewarding doing it myself with learned tools.

1.2 Problem to be resolved

The challenge to resolve is being able to find a rental apartment unit in Manhattan NY that offers similar characteristics and benefits to my current situation. Therefore, in order to set a basis for comparison, I want to find a rental unit subject to the following conditions:

Apartment with min 2 bedrooms with monthly rent not to exceed 7000 US dollars/month Unit located within walking distance (<=1.0 mile, 1.6 km) from a subway metro station in Manhattan Area with amenities and venues similar to the ones described for current location (See item 2.1)

1.3 Targeted Audience for this project

I believe this is a relevant project for a person or entity considering moving to a major city in Europe, US or Asia, since the approach and methodologies used here are applicable in all cases. The use of FourSquare data and mapping techniques combined with data analysis will help resolve the key questions arisen. Lastly, this project is a good practical case toward the development of Data Science skills

2. Data Section

2.1 Data of Current Situation (current residence place)

I Currently reside in the neighborhood of 'Kanaleneiland' in Utrecht. I use Foursquare to identify the venues around the area of residence which are then shown in the Singapore map shown in methodology and execution in section 3.0 . It serves as a reference for comparison with the desired future location in Manhattan NY.

2.2 Data required to resolve the problem

In order to make a good choice of a similar apartment in Manhattan NY, the following data is required:

- List/Information on neighborhoods from Manhattan with their Geodata (latitude and longitude).
- List/Information about the subway metro stations in Manhattan with geodata.
- Listed apartments for rent in Manhattan area with descriptions (how many beds, price, location, address) Venues and amenities in the Manhattan neighborhoods (e.g. top 10)

2.3 Data sources and data manipulation

The list of Manhattan neighborhoods is worked out during Lab skills exercise during the final course of the IBM Professional Certificate in Data Science. A .csv file was created which will be read in order to create a dataframe and its mapping. The .csv file **"mh_neigh_data.csv"** has the following below data structure. The file will be directly read to the Jupyter Notebook for convenience and space savings. The clustering of neighborhoods and mapping will be shown however. An algorithm was used to determine the geodata from Nominatim. Nominatim is a tool to search OSM data by name and address (geocoding) and to generate synthetic addresses of OSM points (reverse geocoding).

As an example, the .csv file **"mh_neigh_data.csv"** contains:

	Borough	Neighborhood	Latitude	Longitude
35	Manhattan	Turtle Bay	40.752042	-73.967708
36	Manhattan	Tudor City	40.746917	-73.971219
37	Manhattan	Stuyvesant Town	40.731000	-73.974052
38	Manhattan	Flatiron	40.739673	-73.990947

A list of Manhattan subway metro stops was compiled in Numbers (Apple excel) and it was complemeted with wikipedia data (https://en.wikipedia.org/wiki/List_of_New_York_City_Subway_stations_in_Manhattan) and information from NY Transit authority and Google maps (https://www.google.com/maps/search/manhattan+subway+metro+stations/@40.7837297,-74.1033043,11z/data=!3m1!4b1) for a final consolidated list of subway stops names and their address. The geolocation was obtained via an algorythm using Nominatim. Details will be shown in the execution of methodolody in section 3.0.

As an example, the subway .csv file **"mh_neigh_data.csv"** contains:

17 190 Street Subway Station Bennett Ave, New York, NY 10040, USA 40.858113 -73.932983

18 59 St-Lexington Av Station E 60th St, New York, NY 10065, USA 40.762259 -73.966271

19 57 Street Station New York, NY 10019, United States 40.764250 -73.954525

20 14 Street / 8 Av New York, NY 10014, United States 40.730862 -73.987156

21 MTA New York City 525 11th Ave, New York, NY 10018, USA 40.759809 -73.999282

A list of places for rent was collected by web-browsing real estate companies in Manhattan :
http://www.rentmanhattan.com/index.cfm?page=search&state=results
https://www.nestpick.com/search?city=new-york&page=1&order=relevance&district=manhattan&gclid=CjwKCAiAjNjgBRAgEiwAGLIf2hkP3A-cPxjZYkURqQEswQK2jKQEpv_MvKcrIhRWRzNkc_r-fGi0IxoCA7cQAvD_BwE&type=apartment&display=list
https://www.realtor.com/apartments/Manhattan_NY

A .csv file was compiled with the rental place that indicated: areas of Manhattan, address, number of beds, area and monthly rental price. The csv file **'nnnn.csv'** had the following below structure. An algorithm was used to create all the geodata using Nominatim, as shown in section 3.0. The actual algorithm coding may be shown in 'markdown' mode because it takes time to run. With the use of geolocator = Nominatim() , it was possible to determine the latitude and longitude for the subway metro locations as well as for the geodata for each rental place listed. The loop algorithms used are shown in the execution of data in section 3.0 "Great_circle" function from geolocator was used to calculate distances between two points , as in the case to calculate average rent price for units around each subway station and at 1.6 km radius. Foursquare is used to find the avenues in Manhattan neighborhoods in general and a cluster is created to later be able to search for the venues depending on the location shown.

2.4 How the data will be used to solve the problem

The data will be used as follows: Use Foursquare and geopy data to map top 10 venues for all Manhattan neighborhoods and clustered in groups (as per Course LAB) Use foursquare and geopy data to map the location of subway metro stations , separately and on top of the above clustered map in order to be able to identify the venues and amenities near each metro station, or explore each subway location separately Use Foursquare and geopy data to map the location of rental places, in some form, linked to the subway locations. create a map that depicts, for instance, the average rental price per square ft, around a radius of 1.0 mile (1.6 km) around each subway station - or a similar metrics. I will be able to quickly point to the popups to know the relative price per subway area. Addresses from rental locations will be converted to geodata(lat, long) using Geopy-distance and Nominatim. Data will be searched in open data sources if available, from real estate sites if open to reading, libraries or other government agencies such as Metro New York MTA, etc.

3. Methodology

The strategy is based on mapping the described data in section 2.0, in order to facilitate the choice of at least two candidate places for rent. The information will be consolidated in ONE MAP where one can see the details of the apartment, the cluster of venues in the neighborhood and the relative location from a subway station and from the workplace. A measurement tool icon will also be provided. The pop ups on the map items will display rent price, location and cluster of venues applicable. The Tools: Web-scraping of sites is used to consolidate data-frame information which was saved as csv files for convenience and to simply the report.

Geodata was obtained by coding a program to use Nominatim to get latitude and longitude of subway stations and also for each of (144 units) the apartments for rent listed. Geopy_distance and Nominatim were used to establish relative distances. Seaborn graphic was used for general statistics on rental data. Maps with pop ups labels allow quick identification of location, price and feature, thus making the selection very easy.

3.1 Process steps and strategy to resolve the problem

The strategy is based on mapping the above described data in section 2.0, in order to facilitate the choice of at least two candidate places for rent. The choice is made based on the demands imposed : location near a subway, rental price and similar venues to Singapore. This visual approach and maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy.

The processing of these DATA and its mapping will allow to answer the key questions to make a decision:

- What is the cost of available rental places that meet the demands?
- What is the cost of rent around a mile radius from each subway metro station?
- What is the area of Manhattan with best rental pricing that meets criteria established?
- What is the distance from workplace (Park Ave and 53 rd St) and the tentative future rental home?
- What are the venues of the two best places to live? How do the prices compare?

- How are venues distributed among Manhattan neighborhoods and around metro stations?
- Are there tradeoffs between size and price and location?
- Any other interesting statistical data findings of the real estate and overall data.

3.2 Data Science Methods, machine learning, mapping tools and exploratory data analysis

Firstly, the "`folium`" Python package was deployed to visualize the map of my current location in Utrecht with venues near residence. As can be seen from the code in the notebook, there are only 4 main attractions near where I live, as I am located in a very quiet neighborhood around 2 kilometers outside the city center.

Following that, to gain insight regarding the Manhattan neighborhoods, cluster neighborhood data was produced with Foursquare during course lab work. This was achieved using the *k-means* unsupervised clustering algorithm.
A .csv file named **"mh_neigh_data.csv"** was produced containing the neighborhoods around the 40 Boroughs. Now, the csv file is just read for convenience and consolidation of the report.
Subsequently, the map of Manhattan neighborhoods with top 10 clustered venues was created.

Now, it is time to dig into Manhattan rental prices. Several Manhattan real estate webs were webscrapped to collect rental data, as mentioned in section 2.0 . The result was summarized in a .csv file for direct reading names **"MH_flats_price.csv"**, in order to consolidate the proces. The initial data for 144 apartments did not have the latitude and longitude data (NaN) but the information was established in the following cell using an algorithm and Nominatim.

Obtained geodata (latitude,longitude) for each rental place in Manhattan with Nominatim Data was stored in a .csv file for simplification report purposes and saving code processing time in future.

Finally, for visualization purposes the `seaborn` Python library was deployed.

4. Results

4.1 Maps

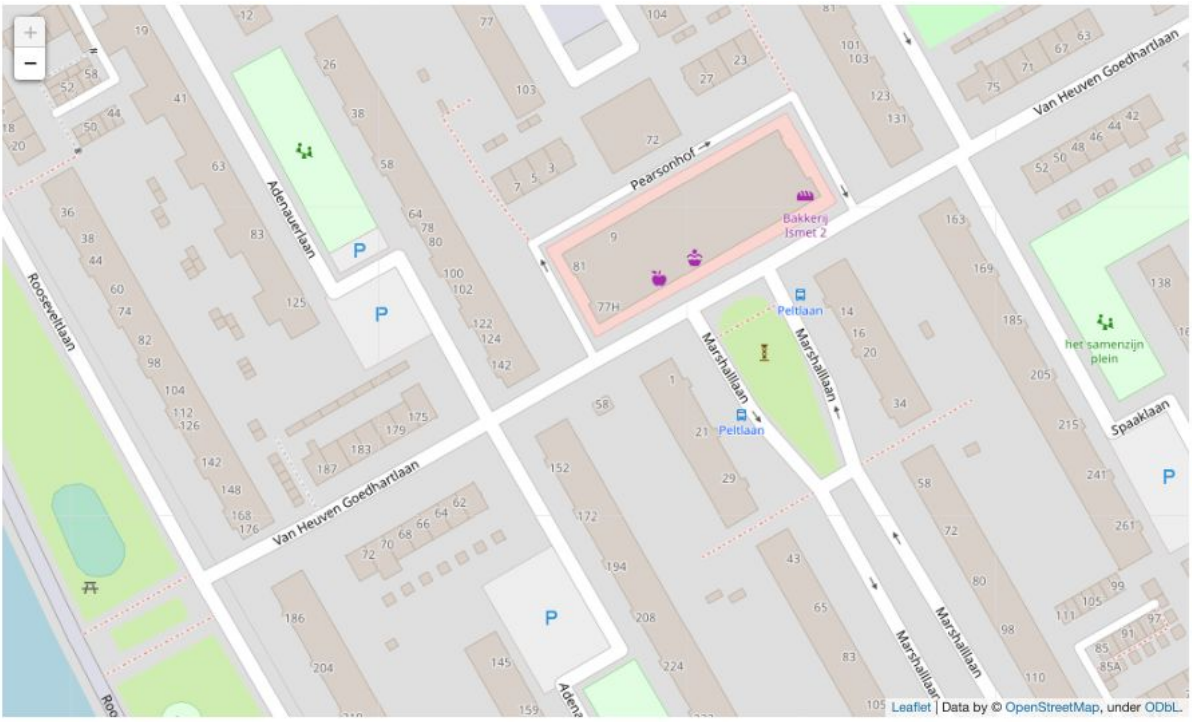


Figure 1: Current map in Utrecht, Netherlands

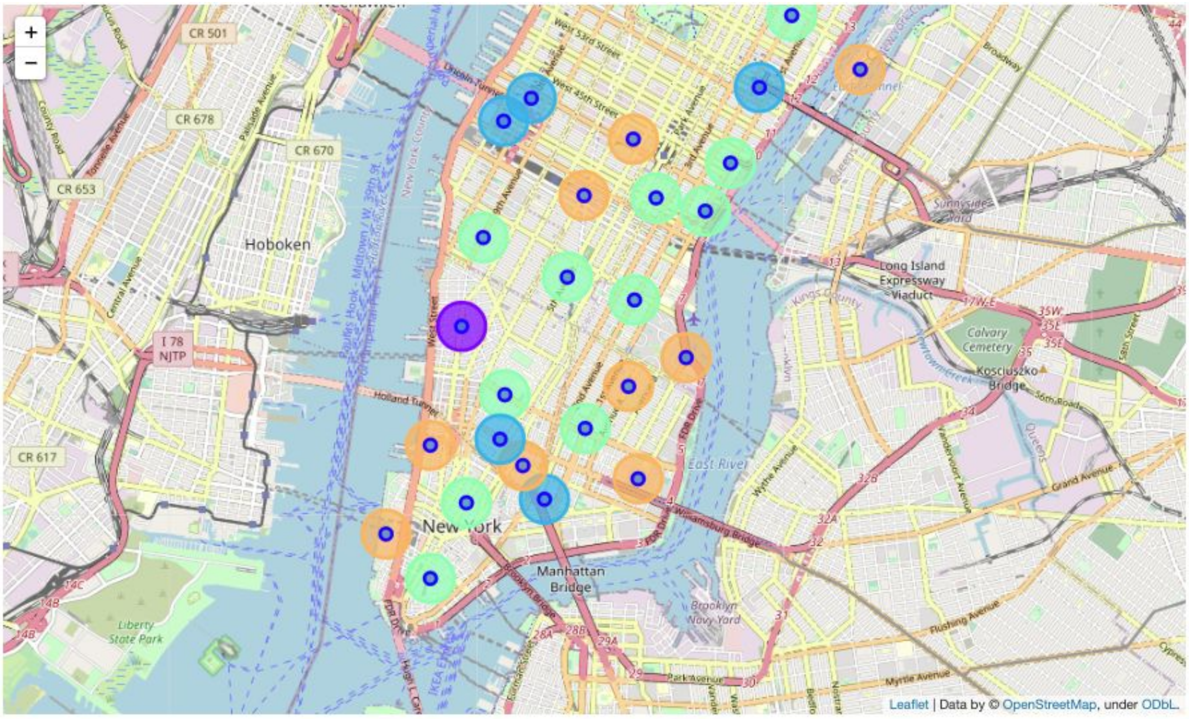


Figure 2: Map of Manhattan neighborhoods with top 10 clustered venues.

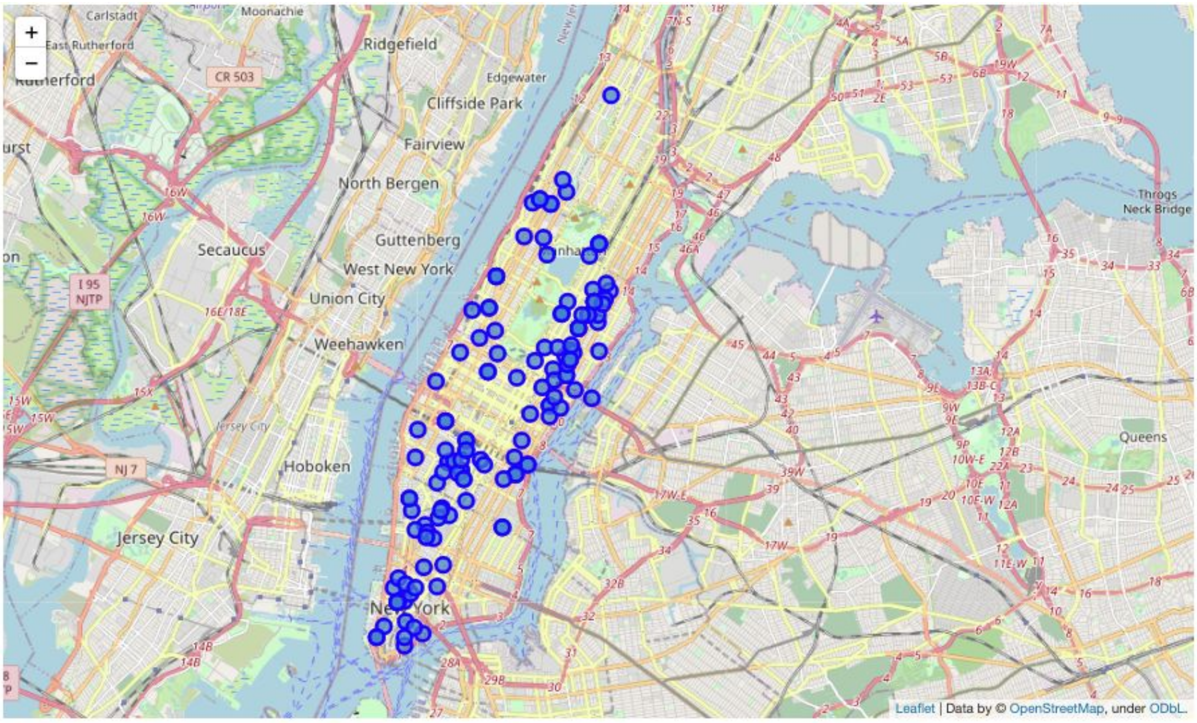


Figure 3: Map of rental apartments in Manhattan

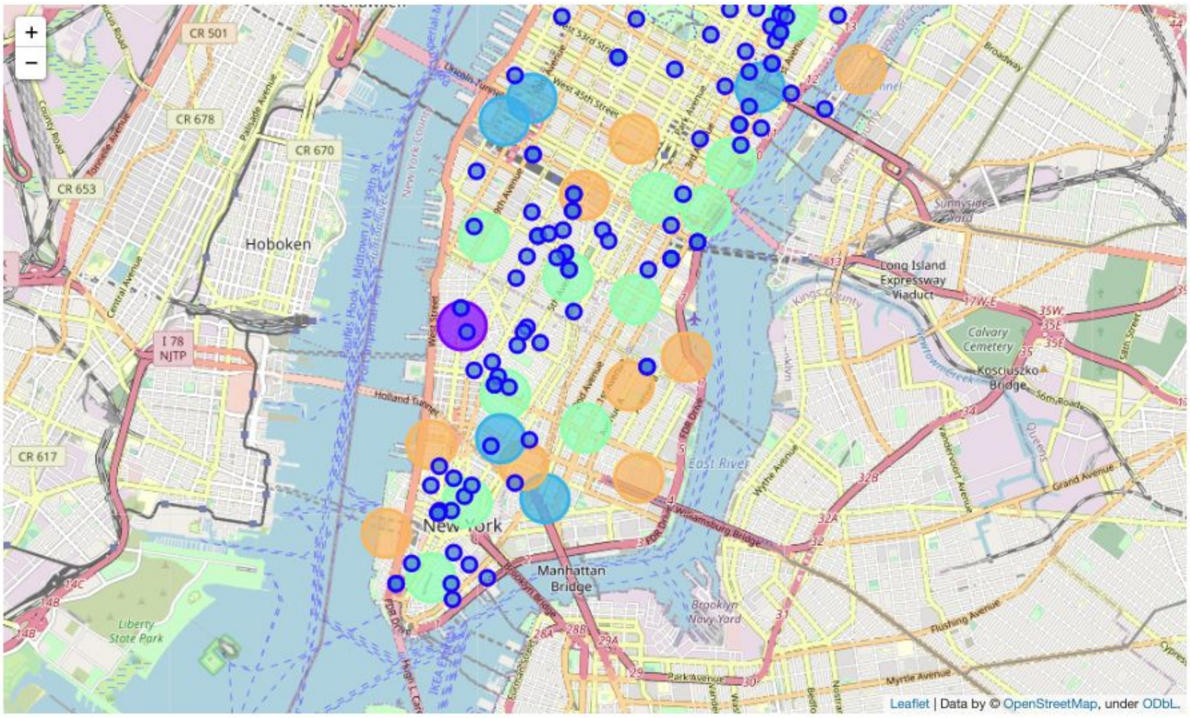


Figure 4: Map of Manhattan showing the places for rent and the cluster of venues

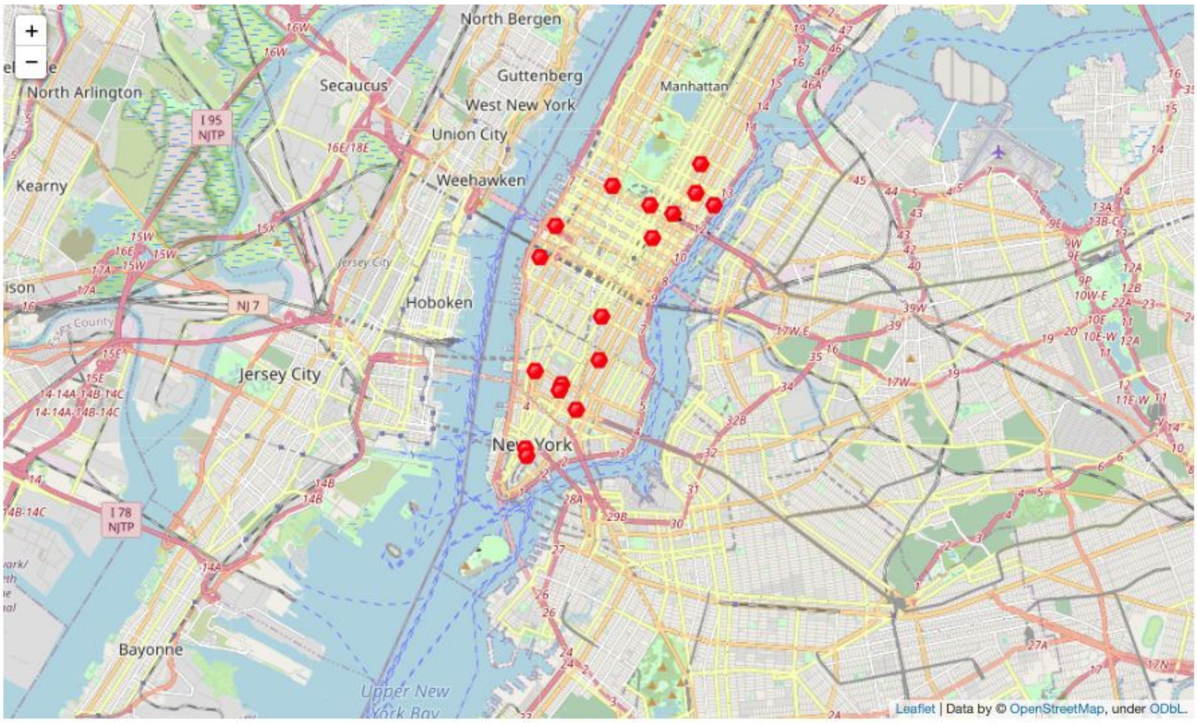


Figure 5: Map of Manhattan showing the location of subway stations.

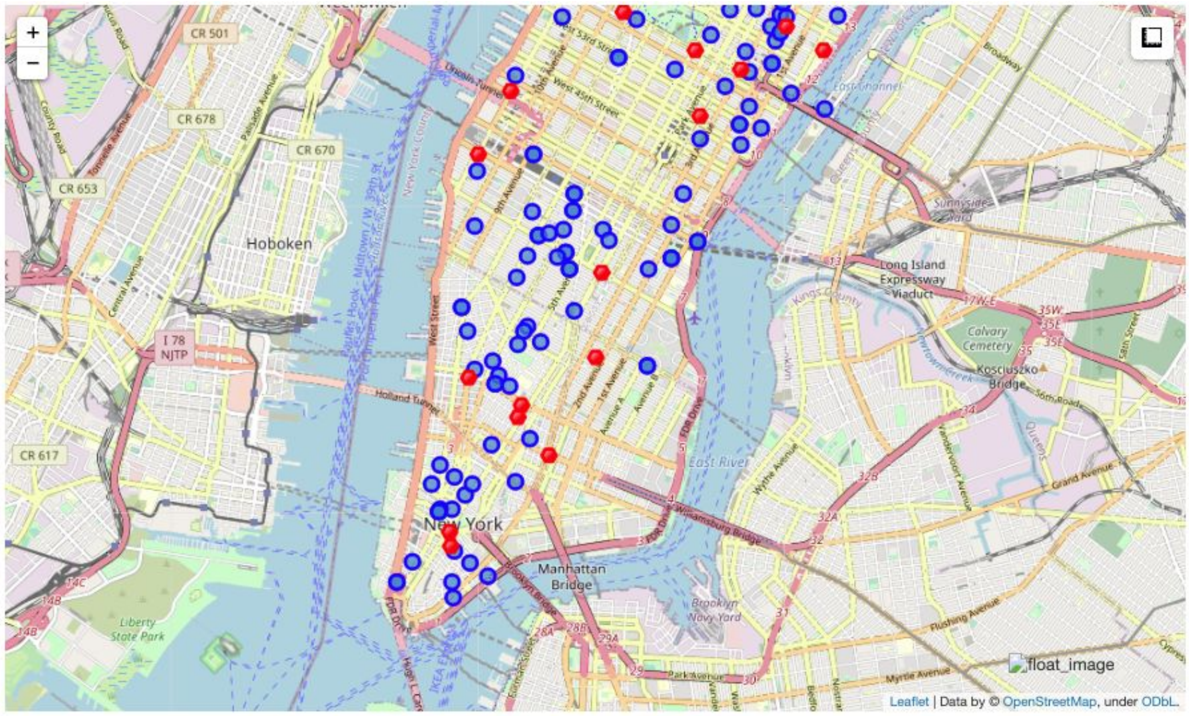
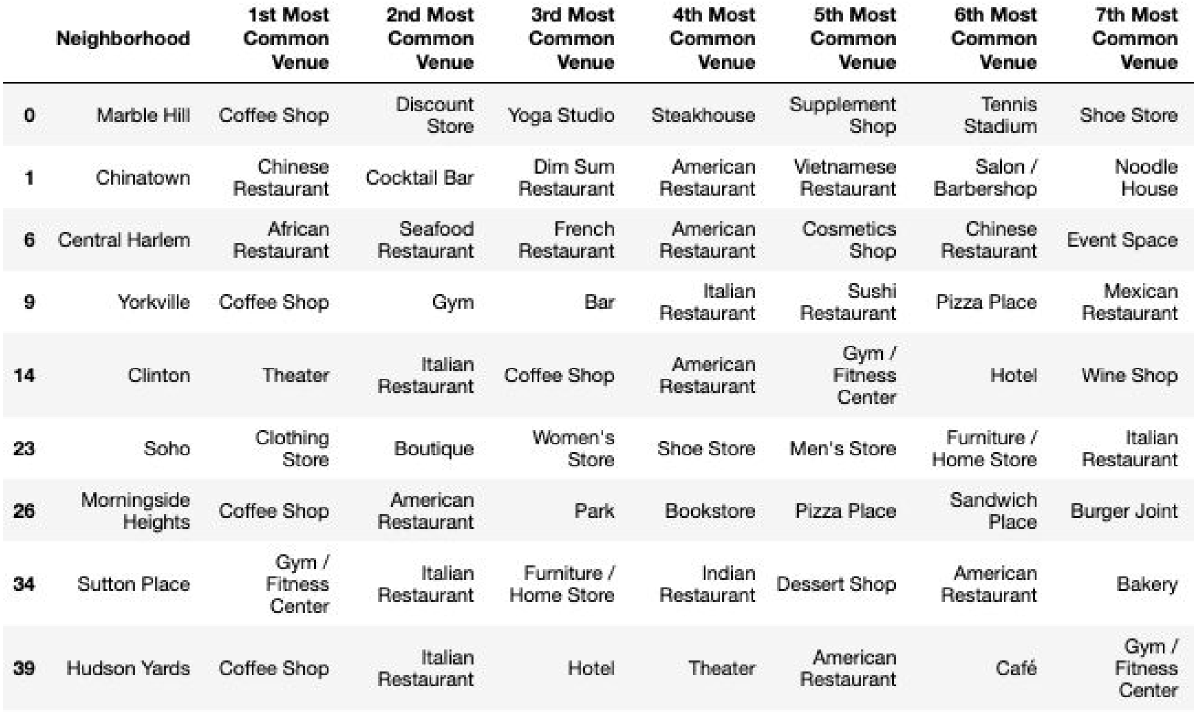


Figure 6: Map of Manhattan with rental places, subway locations and a cluster of venues. Red dots are Subway stations. Blue dots are apartments available for rent, Bubbles are the clusters of venues.

4.2 Examine cluster findings

After examining several cluster data, I concluded that cluster #2 resembles closer the Utrecht place, therefore providing guidance as to where to look for the future apartment.



	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
0	Marble Hill	Coffee Shop	Discount Store	Yoga Studio	Steakhouse	Supplement Shop	Tennis Stadium	Shoe Store
1	Chinatown	Chinese Restaurant	Cocktail Bar	Dim Sum Restaurant	American Restaurant	Vietnamese Restaurant	Salon / Barbershop	Noodle House
6	Central Harlem	African Restaurant	Seafood Restaurant	French Restaurant	American Restaurant	Cosmetics Shop	Chinese Restaurant	Event Space
9	Yorkville	Coffee Shop	Gym	Bar	Italian Restaurant	Sushi Restaurant	Pizza Place	Mexican Restaurant
14	Clinton	Theater	Italian Restaurant	Coffee Shop	American Restaurant	Gym / Fitness Center	Hotel	Wine Shop
23	Soho	Clothing Store	Boutique	Women's Store	Shoe Store	Men's Store	Furniture / Home Store	Italian Restaurant
26	Morningside Heights	Coffee Shop	American Restaurant	Park	Bookstore	Pizza Place	Sandwich Place	Burger Joint
34	Sutton Place	Gym / Fitness Center	Italian Restaurant	Furniture / Home Store	Indian Restaurant	Dessert Shop	American Restaurant	Bakery
39	Hudson Yards	Coffee Shop	Italian Restaurant	Hotel	Theater	American Restaurant	Café	Gym / Fitness Center

Table 1: Venues of Cluster #2 in Manhattan.

4.3 Manhattan apartment rental price statistics

Below there are figures and plots that give an overview of the statistics related to the analysis conducted.

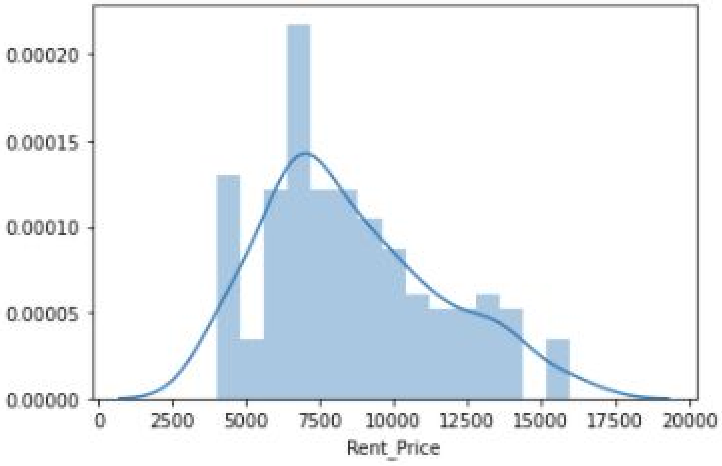


Figure 8: Histogram of Manhattan rental prices.

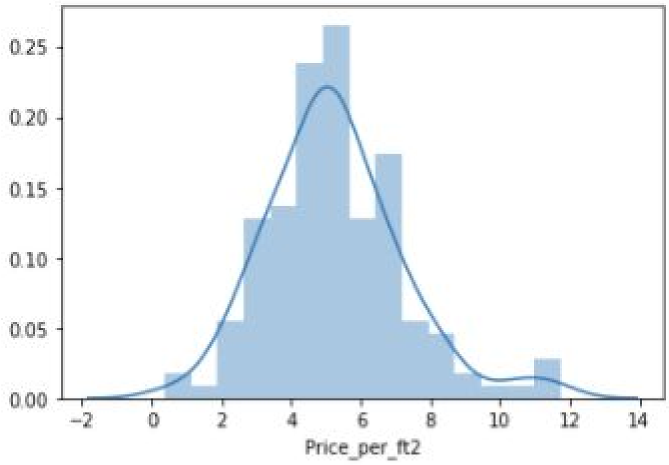


Figure 9: Histogram of prices per square feet.

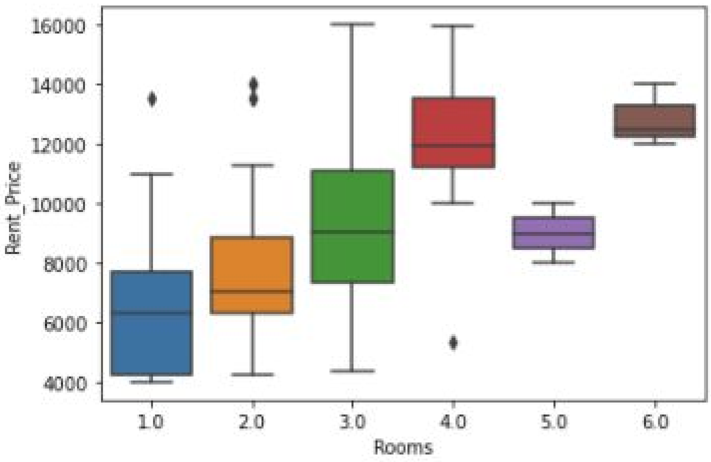
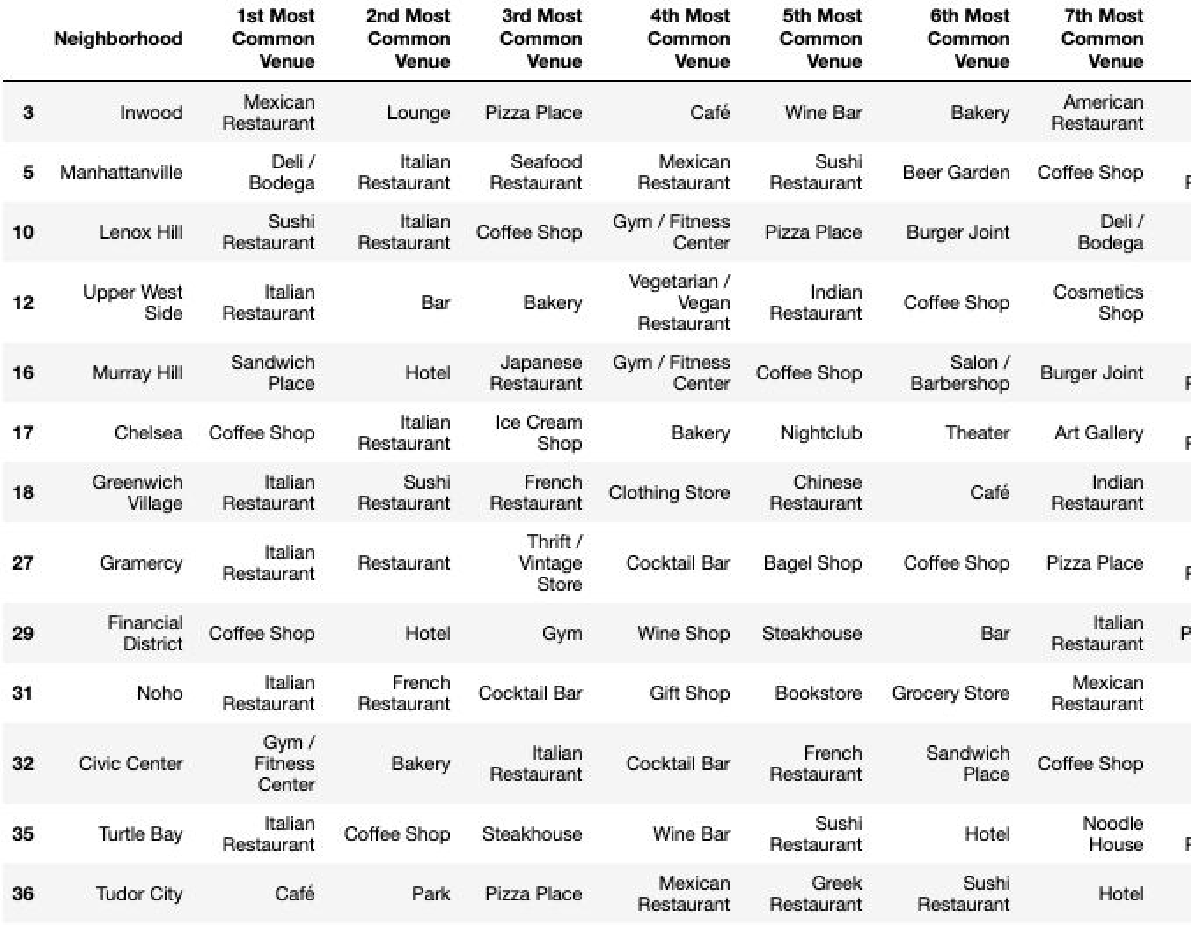


Figure 10: Boxplot of the rental price according to the number of rooms in an apartment



	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	
3	Inwood	Mexican Restaurant	Lounge	Pizza Place	Café	Wine Bar	Bakery	American Restaurant	
5	Manhattanville	Deli / Bodega	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	Sushi Restaurant	Beer Garden	Coffee Shop	F
10	Lenox Hill	Sushi Restaurant	Italian Restaurant	Coffee Shop	Gym / Fitness Center	Pizza Place	Burger Joint	Deli / Bodega	
12	Upper West Side	Italian Restaurant	Bar	Bakery	Vegetarian / Vegan Restaurant	Indian Restaurant	Coffee Shop	Cosmetics Shop	
16	Murray Hill	Sandwich Place	Hotel	Japanese Restaurant	Gym / Fitness Center	Coffee Shop	Salon / Barbershop	Burger Joint	F
17	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Nightclub	Theater	Art Gallery	F
18	Greenwich Village	Italian Restaurant	Sushi Restaurant	French Restaurant	Clothing Store	Chinese Restaurant	Café	Indian Restaurant	
27	Gramercy	Italian Restaurant	Restaurant	Thrift / Vintage Store	Cocktail Bar	Bagel Shop	Coffee Shop	Pizza Place	F
29	Financial District	Coffee Shop	Hotel	Gym	Wine Shop	Steakhouse	Bar	Italian Restaurant	P
31	Noho	Italian Restaurant	French Restaurant	Cocktail Bar	Gift Shop	Bookstore	Grocery Store	Mexican Restaurant	
32	Civic Center	Gym / Fitness Center	Bakery	Italian Restaurant	Cocktail Bar	French Restaurant	Sandwich Place	Coffee Shop	
35	Turtle Bay	Italian Restaurant	Coffee Shop	Steakhouse	Wine Bar	Sushi Restaurant	Hotel	Noodle House	F
36	Tudor City	Café	Park	Pizza Place	Mexican Restaurant	Greek Restaurant	Sushi Restaurant	Hotel	

Table 2: Venues for Apartment 1 - Cluster #3.

To reach a conclusion I should consolidate all the required information to make the apartment selection in one map that will include rental places along with the respective price of US Dollars per month, subway locations and a cluster of venues.

The consolidated map (Figure 6) was used to explore options.

After examining, I have chosen two locations that meet the requirements which will assess to make a choice.

1. Apartment 1: 305 East 63rd Street in the Sutton Place Neighborhood and near 'subway 59th Street' station, Cluster # 2 Monthly rent : 7500 Dollars
2. Apartment 2: 19 Dutch Street in the Financial District Neighborhood and near 'Fulton Street Subway' station, Cluster # 3 Monthly rent : 6935 Dollars

Apartment 1 rent cost is 7,500 US Dollars, which is slightly above the median of 7000 US Dollars. Apartment 1 is located 400 meters from subway station at 59th Street and the workplace (Park Ave and 53rd) is another 600 meters away. I can walk to work place and use subway for other purposes. Venues for this apt are as of Cluster 2 and it is located in a fine district in the East side of Manhattan.

Apartment 2 rent cost is 6,935 US Dollars, just under the 000 US Dollars median. Apartment 2 is located 60 meters from subway station at Fulton Street, but I will have to ride the subway daily to work , possibly a 40-60 min ride. Venues for this apartment areas of Cluster #3.

Based on current Utrecht venues, I feel that Cluster #3 type of venues is a closer resemblance to my current place. That means that APARTMENT 2 is a better choice and cheaper which means I can use it for other expenses. However, there is the issue of transport.

5. Discussion

5.1.Elaboration and discussion on any observations and/or recommendations for improvement

I believe that convenience and location both matter a lot. Having to spend $ 7000 for rent is very high considering I am paying closer to 2,000 US dollars a month in Utrecht and enjoying life. I believe my income should be enough to justify rent of 30-35%. However the US opportunity is closer to 50% of the total, meaning that I am better off staying in Melbourne and looking for another opportunity.

In terms of the Coursera course: In general, I am very impressed with the overall organisation, content and lab works presented during the Coursera IBM Certification Course. It helped me learn a variety of data science tools with my zero previous knowledge of coding.

I feel this Capstone project presented me a great opportunity to practice and apply the Data Science tools and methodologies learned. I have created a good project that I can present as an example to show my potential.

I feel I have acquired a good starting point to become a professional Data Scientist and I will continue exploring to create examples of practical cases.

6. Conclusion

6.1.Desicison taken and Report Conclusion

I decided not to move to the US and stay in Melbourne considering the prices. I will explore Los Angeles for the future career opportunities and run the same cost benefit analysis to make an informed data driven decision.

6. 2 Final feedback on the overall data science course

I am very happy to be able to complete the 9 course specialisation in 6 months with on and off time and money spent. While not in the data science area career wise, this will not help me manage data scientists in the team better and align expectations with possibilities. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision thoroughly and with confidence. I would recommend it for use in similar situations.

Thank you for reviewing my work and thanks to the IBM/Coursera community for this course!