Data Science Project
Coursera Capstone

Deepak Singh

Report Content

1. Introduction Section :
- The “business problem” to be solved by this project and who may be interested
2. Data Section:
- Describe Data requirements and Sources needed to solve the problem
3. Methodology section:
- Main component of the report - Execute data processing, describe/discuss any exploratory data analysis and/or inferential statistical testing performed, and/or machine learnings used.
4. Results section:
- Discussion of the results and finding of answer
5. Discussion section:
- Discussion of observations noted and any recommendations
6. Conclusion section:
- Answer chosen and conclusions.

1.0 Introduction

Discussion of the business problem and the audience who would be interested in this project.

Description of the Problem and Background

Scenario:

• I am a data scientist residing in Downtown Los Angeles.
• I currently live within walking distance to Downtown metro station and I enjoy many amenities and venues in the area, such as various international cousins restaurants, cafes, food shops and entertainment.
• I have been offered a great opportunity to work for a leader firm in Manhattan, NY. I am very excited and I want to use this opportunity to practice my learning in Coursera in order to answer relevant questions arisen. The key question is

How can I find a convenient and enjoyable place similar to mine now in Los Angeles?
Certainly, I can use available real estate apps and Google but the idea is to use and apply myself the learned tools during the course.
In order to make a comparison and evaluation of the rental options in Manhattan NY, I must set some basis, therefore the apartment in Manhattan must meet the following demands:

1. Apartment must be 2 or 3 bedrooms

2. desired location is near a metro station in the Manhattan area and within 1.0 mile (1.6 km) radius

3. price of rent not exceed 7000 dollars per month

4. Top amenities in the selected neighborhood shall be similar to current residence.

5. Desirable to have venues such as coffee shops, restaurants Asian Thai, wine stores, gym and food.

2.0 Data Section

2.1 Data Requirements

- Geodata for current residence in Singapore with venues established using Foursquare.

- List of Manhattan (MH) neighborhoods with clustered venues established via Foursquare (as Lab).
https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods#Midtown_neighborho

- List of subway metro stations in Manhattan with addresses and geo data (lat,long): https://en.wikipedia.org/wiki/List_of_New_York_City_Subway_stations_in_Manhattan) , (https://www.google.com/_maps/search/manhattan+subway+metro+stations/@40.7837297,-74.1033043,11z/data=!3m1!4b1)

- List of apartments for rent in Manhattan area with information on neighborhood location, add beds, area size, monthly rent price and complemented with geo data via Nomination
http:// www.rentmanhattan.com/index.cfm?page=search&state=results

2.2 Data Sources, Data Processing and Tools used

- Manhattan neighborhoods were obtained from Wikipedia and organized by Neighborhoods Nomination for mapping with Folium.

- List of Subway stations was obtained via Wikipedia, NY Transit web site and Google map,

-List of apartments for rent was consolidated from web-scraping real estate sites for Manhattan. The (latitude, longitude) data was found with algorithm coding and using Nomination.

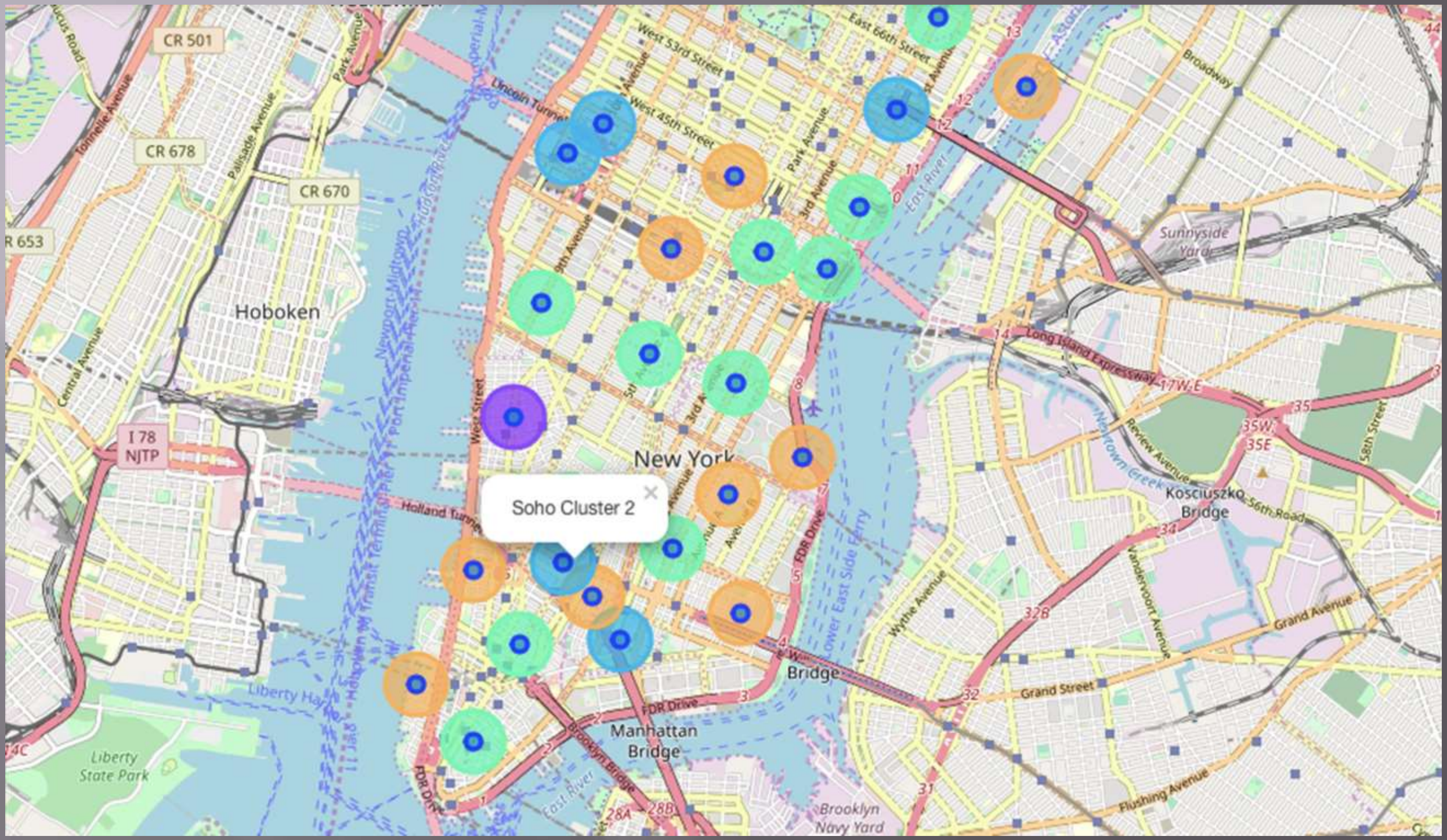
3.0 Methodology

The Strategy to find the answer:
The strategy is based on mapping the described data in section 2.0, in order choice of at least two candidate places for rent. The information will be consol MAP where one can see the details of the apartment, the cluster of venues in and the relative location from a subway station and from work place. A mea will also be provided. The popups on the map items will display rent price, lo cluster of venues applicable.

The Tools:
Web-scraping of sites is used to consolidate data-frame information which files for convenience and to simply the report. Geodata was obtained by coding to use Nominatim to get latitude and longitude of subway stations and also units) the apartments for rent listed. Geopy_distance and Nominatim were used relative distances. Seaborn graphic was used for general statistics on rental dat Maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy

4.0 Execution and Results

Manhattan Map - Neighborhoods and Cluster of V



GeoData Manhattan apts for rent

```
]: mh_rent=pd.read_csv('MH_rent_latlong.csv')
   mh_rent.head()
```

]:

	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
0	West 105th Street	Upper West Side	2.94	5.0	3400	10000	40.799771	-73.966213
1	East 97th Street	Upper East Side	3.57	3.0	2100	7500	40.788585	-73.955277
2	West 105th Street	Upper West Side	1.89	4.0	2800	5300	40.799771	-73.966213
3	CARMINE ST.	West Village	3.03	2.0	1650	5000	40.730523	-74.001873
4	171 W 23RD ST.	Chelsea	3.45	2.0	1450	5000	40.744118	-73.995299

```
]: mh_rent.tail()
```

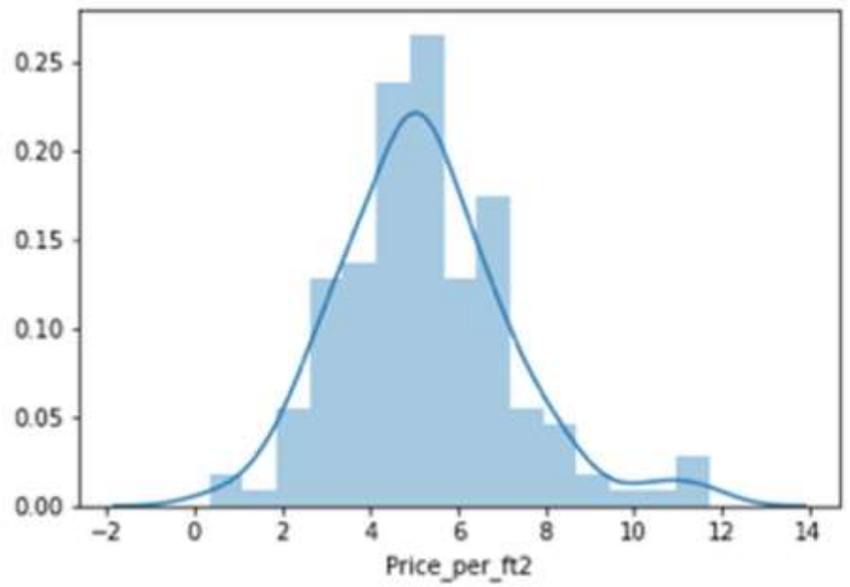
]:

	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
139	200 East 72nd Street	Rental in Lenox Hill	5.15	3.0	1700	8750	40.769465	-73.960339
140	50 Murray Street	No fee rental in Tribeca	7.11	2.0	1223	8700	40.714051	-74.009608
141	300 East 56th Street	No fee rental in Midtown East	3.87	3.0	2100	8118	40.758216	-73.965190
142	1930 Broadway	No fee rental in Central Park West	5.06	2.0	1600	8095	40.772474	-73.981901
143	33 West 9th Street	Rental in Greenwich Village	6.67	2.0	1500	10000	40.733691	-73.997323

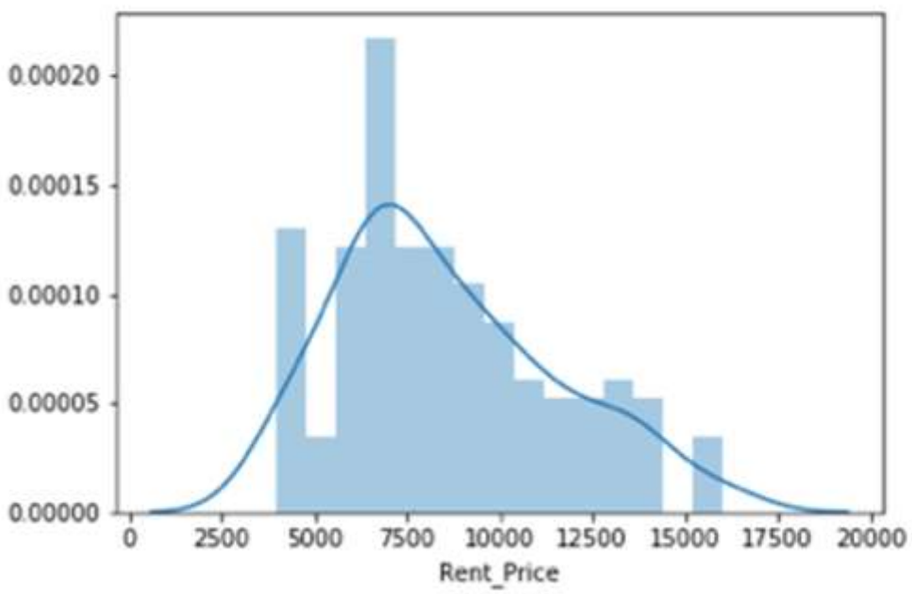
RENTAL PRICE STATISTICS MANHATTAN APARTMENTS

Budget US7000/month is around the mean

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2415fc18>
```

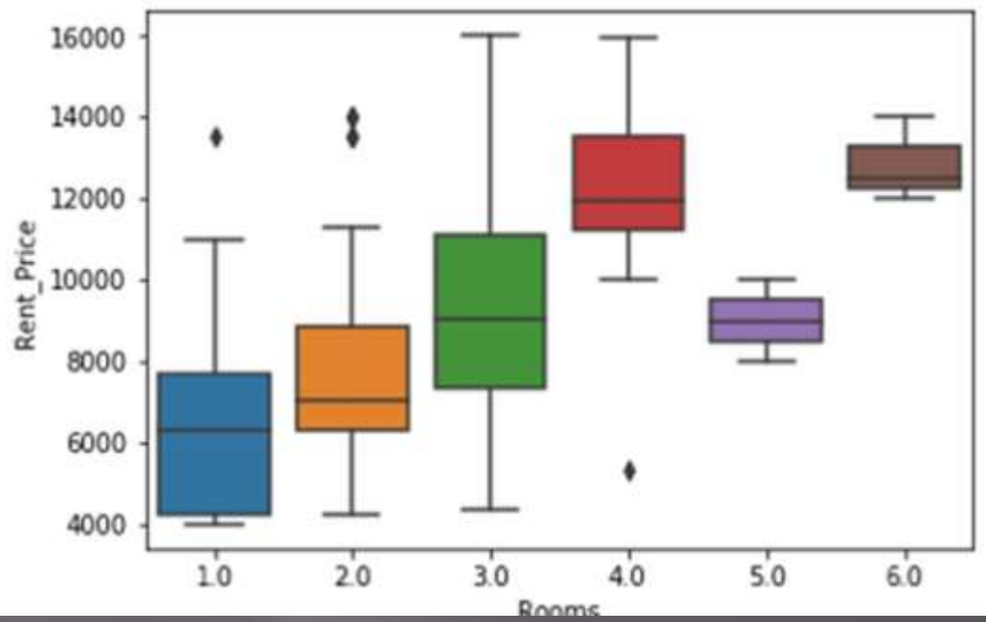



```
<matplotlib.axes._subplots.AxesSubplot at 0x1a25dd8400>
```

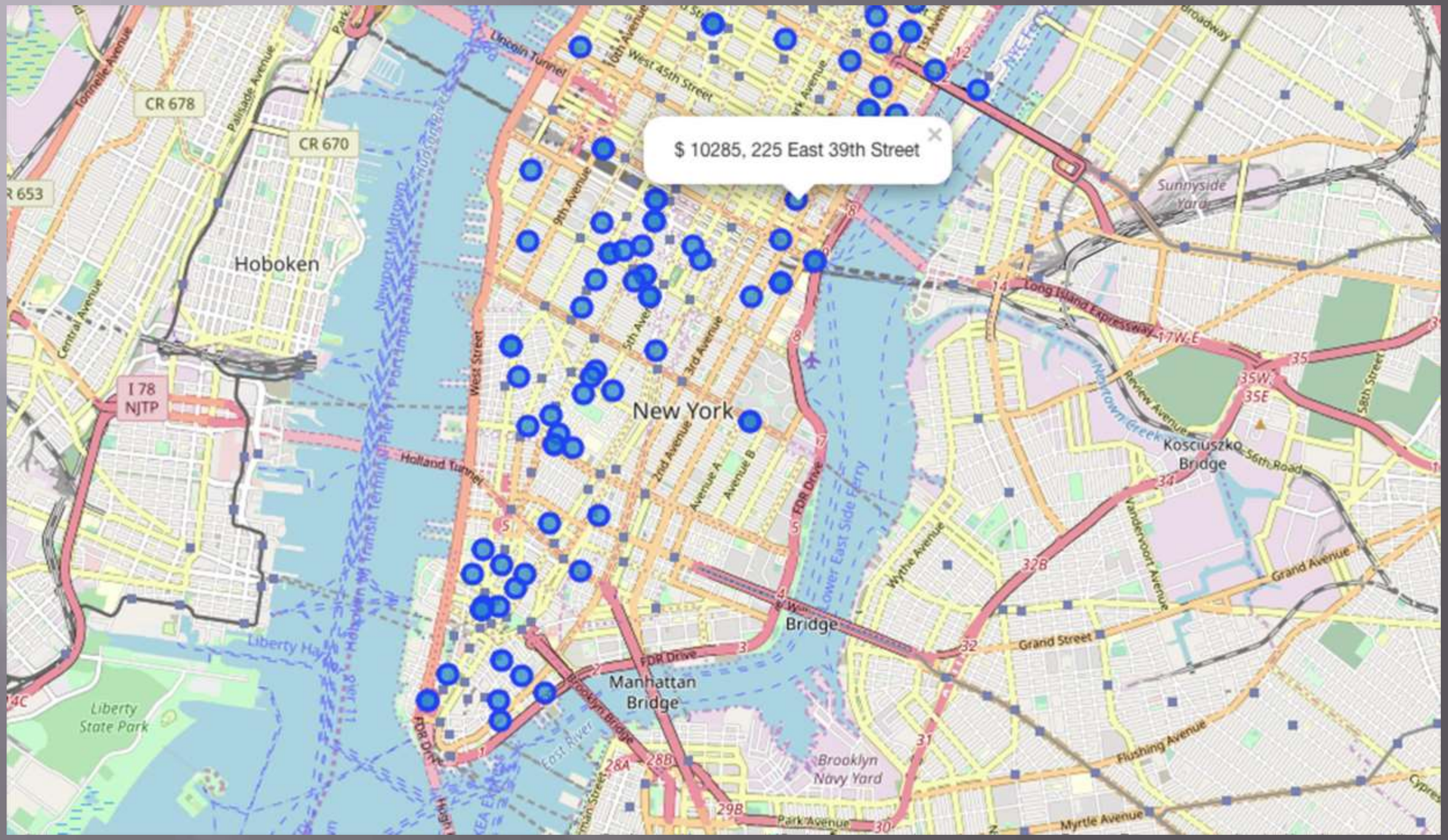



```
sns.boxplot(x='Rooms', y= 'Rent_Price', data=mh_rent)
```

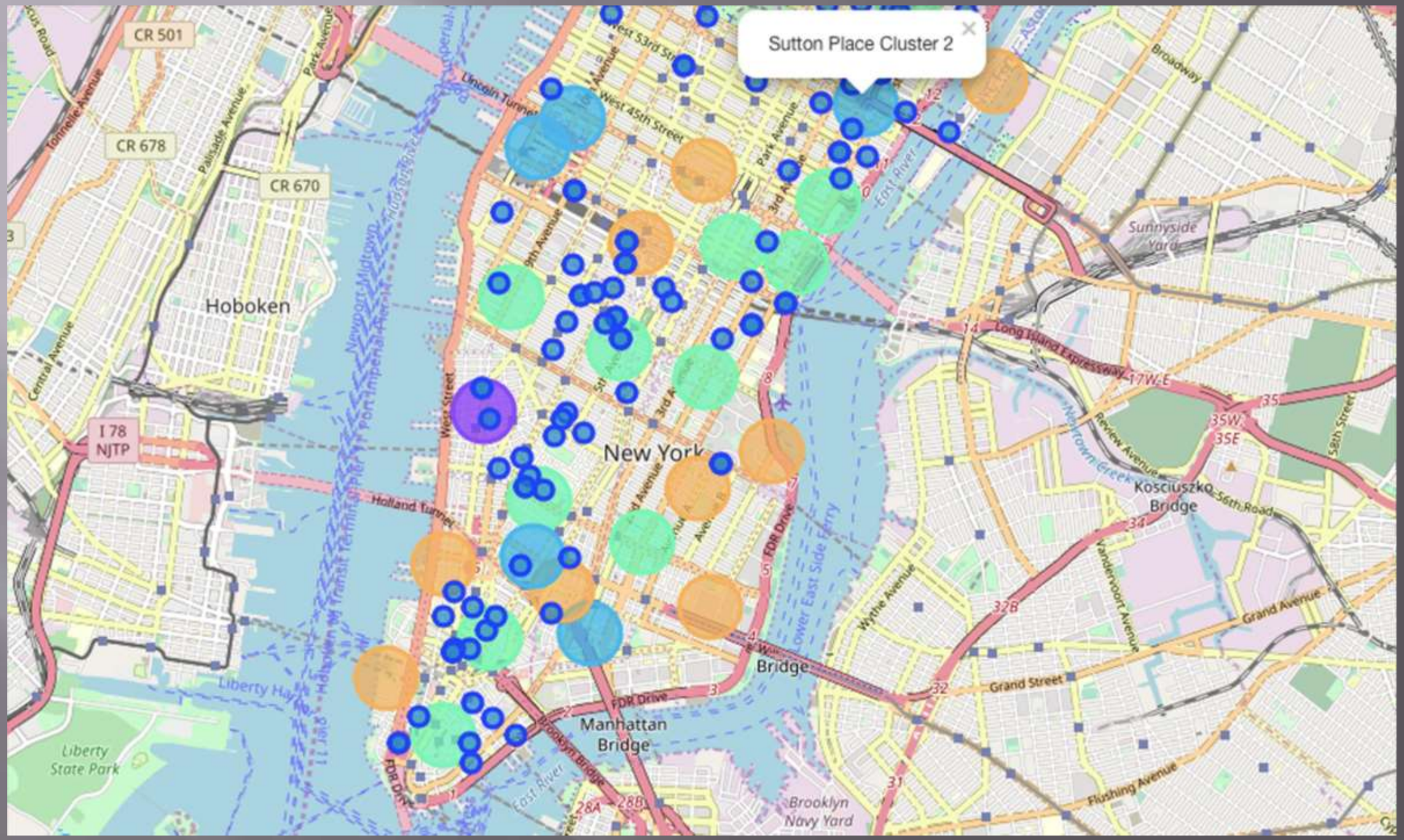
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a25f2a2b0>
```



Apartments for Rent in Manhatten



Manhattan apartments for rent with venue cluster



Venues of cluster 3

```
## kk is the cluster number to explore
kk = 3
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == kk, manhattan_merged.columns[[1] + list(range(5, manhattan_m
```

	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	10th Most Common Venue
3	Inwood	Mexican Restaurant	Lounge	Pizza Place	Café	Wine Bar	Bakery	American Restaurant	Park	Frozen Yogurt Shop	Spanish Restaurant
5	Manhattanville	Deli / Bodega	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	Sushi Restaurant	Beer Garden	Coffee Shop	Falafel Restaurant	Bike Trail	Other Nightlife
10	Lenox Hill	Sushi Restaurant	Italian Restaurant	Coffee Shop	Gym / Fitness Center	Pizza Place	Burger Joint	Deli / Bodega	Gym	Sporting Goods Shop	Thai Restaurant
12	Upper West Side	Italian Restaurant	Bar	Bakery	Vegetarian / Vegan Restaurant	Indian Restaurant	Coffee Shop	Cosmetics Shop	Wine Bar	Mexican Restaurant	Sushi Restaurant
16	Murray Hill	Sandwich Place	Hotel	Japanese Restaurant	Gym / Fitness Center	Coffee Shop	Salon / Barbershop	Burger Joint	French Restaurant	Bar	Italian Restaurant
17	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Nightclub	Theater	Art Gallery	Seafood Restaurant	American Restaurant	Hotel
18	Greenwich Village	Italian Restaurant	Sushi Restaurant	French Restaurant	Clothing Store	Chinese Restaurant	Café	Indian Restaurant	Bakery	Seafood Restaurant	Electronics Store
27	Gramercy	Italian Restaurant	Restaurant	Thrift / Vintage Store	Cocktail Bar	Bagel Shop	Coffee Shop	Pizza Place	Mexican Restaurant	Grocery Store	Wine Shop
29	Financial District	Coffee Shop	Hotel	Gym	Wine Shop	Steakhouse	Bar	Italian Restaurant	Pizza Place	Park	Gym / Fitness Center
31	Noho	Italian Restaurant	French Restaurant	Cocktail Bar	Gift Shop	Bookstore	Grocery Store	Mexican Restaurant	Hotel	Sushi Restaurant	Coffee Shop

Manhattan subway stations geodata

click to scroll output; double click to hide

	sub_station	sub_address	lat	long
0	Dyckman Street Subway Station	170 Nagle Ave, New York, NY 10034, USA	40.861857	-73.924509
1	57 Street Subway Station	New York, NY 10106, USA	40.764250	-73.954525
2	Broad St	New York, NY 10005, USA	40.730862	-73.987156
3	175 Street Station	807 W 177th St, New York, NY 10033, USA	40.847991	-73.939785
4	5 Av and 53 St	New York, NY 10022, USA	40.764250	-73.954525

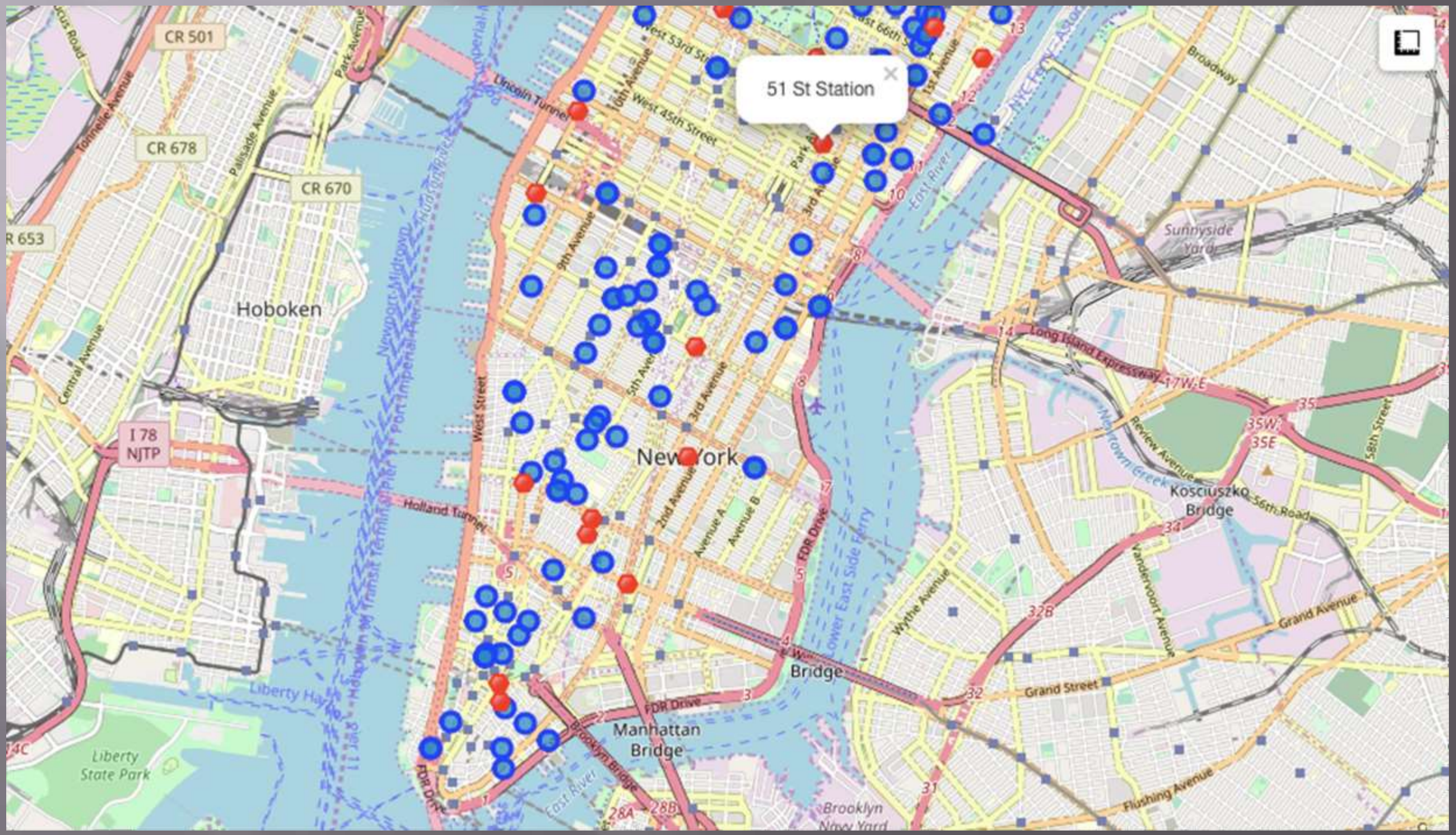
```
# removing duplicate rows and creating new set mhsub1
mhsub1=mh.drop_duplicates(subset=['lat','long'], keep="last").reset_index(drop=True)
mhsub1.shape
```

```
(22, 4)
```

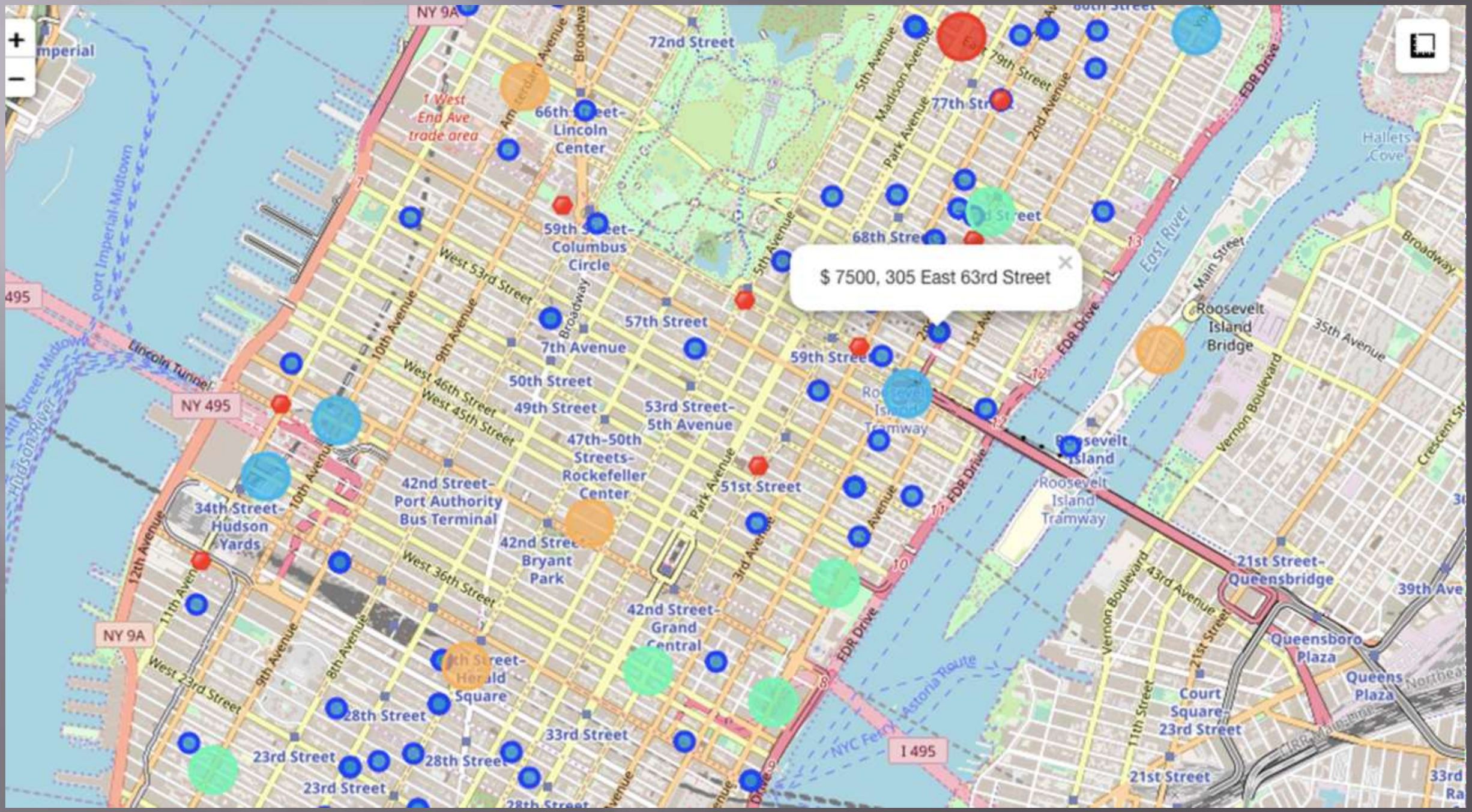
```
mhsub1.tail()
```

	sub_station	sub_address	lat	long
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

Apts for rent (blue) and subway stations (red)



Selected Apartment



Apartment Selection

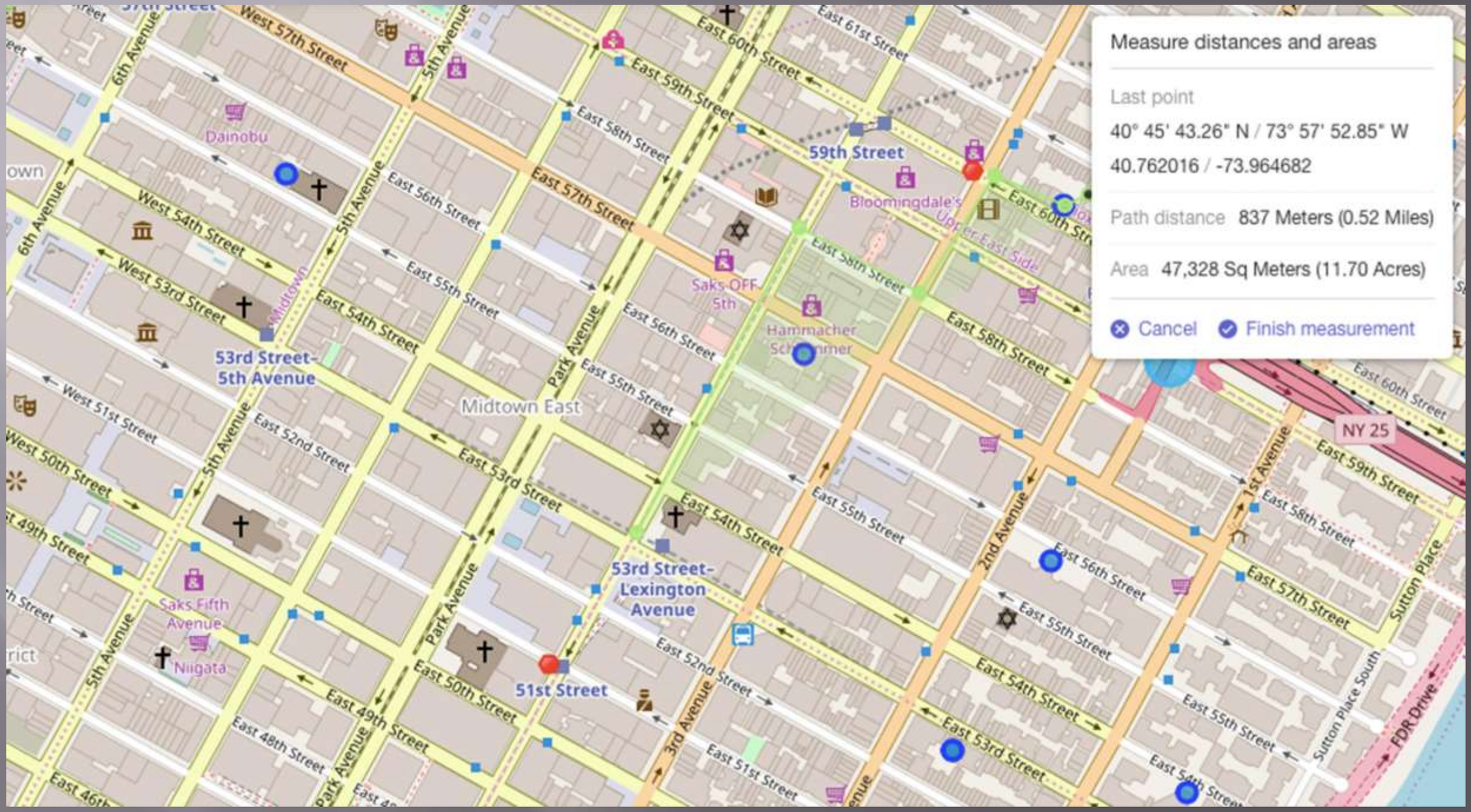
Using the "one map" above, I was able to explore all possibilities since the the information needed for a good decision.

Apartment 1 rent cost is US7500 slightly above the US7000 budget. Apt 1 400 meters from subway station at 59th Street and work place (Park Ave and 600 meters way. I can walk to work place and use subway for other places a this apt are as of Cluster 2 and it is located in a fine district in the East side

Apartment 2 rent cost is US6935, just under the US7000 budget. Apt 2 is from subway station at Fulton Street, but I will have to ride the subway dail possibly 40-60 min ride. Venues for this apt are as of Cluster 3.¶

Based on current Singapore venues, I feel that Cluster 2 type of venues is a resemblance to my current place. That means that APARTMENT 1 is a bett

Neighborhood



Venus in Cluster 2 near future home

```
## kk is the cluster number to explore
kk = 2
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == kk, manhattan_merged.columns[[1] + list(range(5, manhattan_m
```

	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	10th Most Common Venue
0	Marble Hill	Coffee Shop	Discount Store	Yoga Studio	Steakhouse	Supplement Shop	Tennis Stadium	Shoe Store	Gym	Bank	Seafood Restaurant
1	Chinatown	Chinese Restaurant	Cocktail Bar	Dim Sum Restaurant	American Restaurant	Vietnamese Restaurant	Salon / Barbershop	Noodle House	Bakery	Bubble Tea Shop	Ice Cream Shop
6	Central Harlem	African Restaurant	Seafood Restaurant	French Restaurant	American Restaurant	Cosmetics Shop	Chinese Restaurant	Event Space	Liquor Store	Beer Bar	Gym / Fitness Center
9	Yorkville	Coffee Shop	Gym	Bar	Italian Restaurant	Sushi Restaurant	Pizza Place	Mexican Restaurant	Deli / Bodega	Japanese Restaurant	Pub
14	Clinton	Theater	Italian Restaurant	Coffee Shop	American Restaurant	Gym / Fitness Center	Hotel	Wine Shop	Spa	Gym	Indie Theater
23	Soho	Clothing Store	Boutique	Women's Store	Shoe Store	Men's Store	Furniture / Home Store	Italian Restaurant	Mediterranean Restaurant	Art Gallery	Design Studio
26	Morningside Heights	Coffee Shop	American Restaurant	Park	Bookstore	Pizza Place	Sandwich Place	Burger Joint	Café	Deli / Bodega	Tennis Court
34	Sutton Place	Gym / Fitness Center	Italian Restaurant	Furniture / Home Store	Indian Restaurant	Dessert Shop	American Restaurant	Bakery	Juice Bar	Boutique	Sushi Restaurant
39	Hudson Yards	Coffee Shop	Italian Restaurant	Hotel	Theater	American Restaurant	Café	Gym / Fitness Center	Thai Restaurant	Restaurant	Gym

5.0 Discussion

- In general, I am positively impressed with the organization, content and lab works presented the Coursera IBM Certification Course
- I feel this Capstone project presented me a opportunity to practice and apply the Data tools and methodologies learned.
- I have created a good project that I can present example to show my potential.
- I feel I have acquired a good starting point to kick start my career again.

6.0 Conclusions

- I feel rewarded with the efforts and time, I believe this course with all the topics covered is of appreciation.
- This project has shown me a practical application a real situation that has impacting personal and impact using Data Science tools.
- The mapping with Folium is a very powerful te consolidate information and make the analysis a thoroughly and with confidence. I would recommend use in similar situations.