COURSERA FINAL CAPSTONE PROJECT

COURSERA IBM DATA SCIENCE CERTIFICATION

Aradhya Mathur

June 2020

REPORT CONTENT AND PRESENTATION OUTLINE

1.Introduction

-The "Business problem" to be solved by this project and interested audience

2.Data section

- Data requirements and data sources needed to investigate the problem

3. Methodology

- Main technical component of the report- execution of data processing techniques, exploratory data analysis and machine learning techniques used

4. Results

- Discussion of results

5. Discussion

-Observations leading to conclusion

6. Conclusion

-Final decision

1.0 INTRODUCTION

1.1 Scenario and Background

I currently live in Riverside Quay, Southbank, Melbourne, Australia within walking distance to the central business district, train stations and food amenities, shopping malls and festivals. I have an offer to move to Manhattan New York and would like to do a cost benefit analysis to see if I can afford to maintain the same lifestyle/location with the offered salary.

2. Problem statement to resolve

Tofind an apartment with minimum of 2 bedrooms, price of Maximum US$7000 per month located within 1.5 kilometers of subway along with great food amenities

3. Interested Audience

I believe this project is interesting for any expat deciding to migrate to the united states and would like to leverage tools such as foursquare and data science to make an informed data driven decision. The project is replicable for other cities and having a background in data science is recommended.

2.DATA SECTION

- **2.1 Data Requirements**
  - Geodata for current residence in Southbank with venues established using Foursquare
  -List of Manhattan (MH) neighbourhoods with clustered venues established via Foursquare (as in Course Lab). https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods#Midtown_neighborhoods
  - 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, address, number of beds, area size, monthly rent price and complemented with geo data via Nominatim. http://www.rentmanhattan.com/index.cfm?page=search&state=results https://www.nestpick.com/search?city=new-
  - Place to work in Manhattan (Park Avenue and 53rd St) for reference

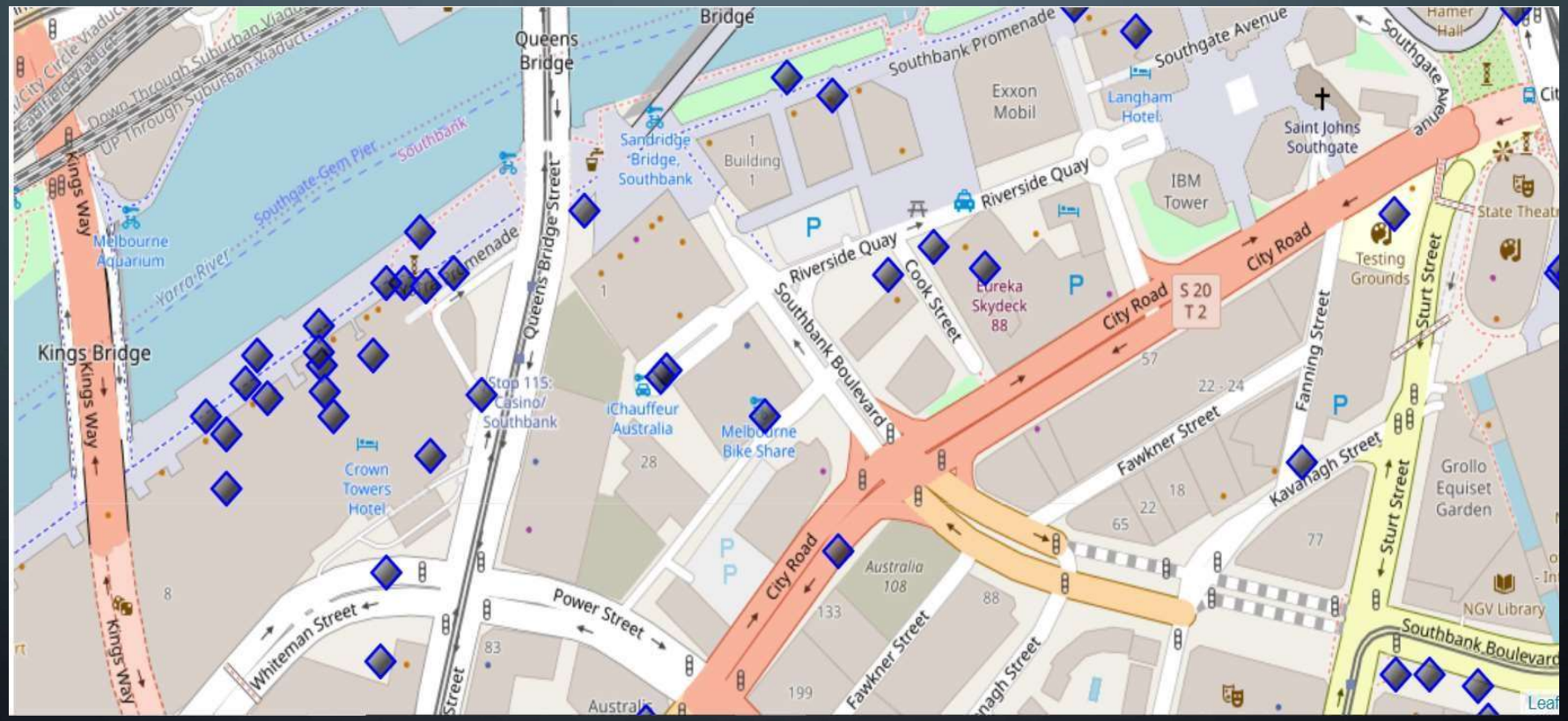
  2.2 Data Sources, Data Processing and Tools used
  - Southbank data and map is to be created with use of Nominatim , Foursquare and Folium mapping
  -Manhattan neighborhoods were obtained from Wikipedia and organized by Neighborhoods with geodata via Nominatim for mapping with Folium.
  - List of Subway 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 MH. The geolocation (lat,long) data was found with algorithm coding and using Nominatim.
  - Folium map was the basis of mapping with various features to consolidate all data in ONE map where one can visualize all details needed to make a selection of apartment

3.0 METHODOLOGY

- The Strategy to find the answer:
  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 workplace. A measurement tool icon will also be provided. The popups 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 popups labels allow quick identification of location, price and feature, thus making the selection very easy

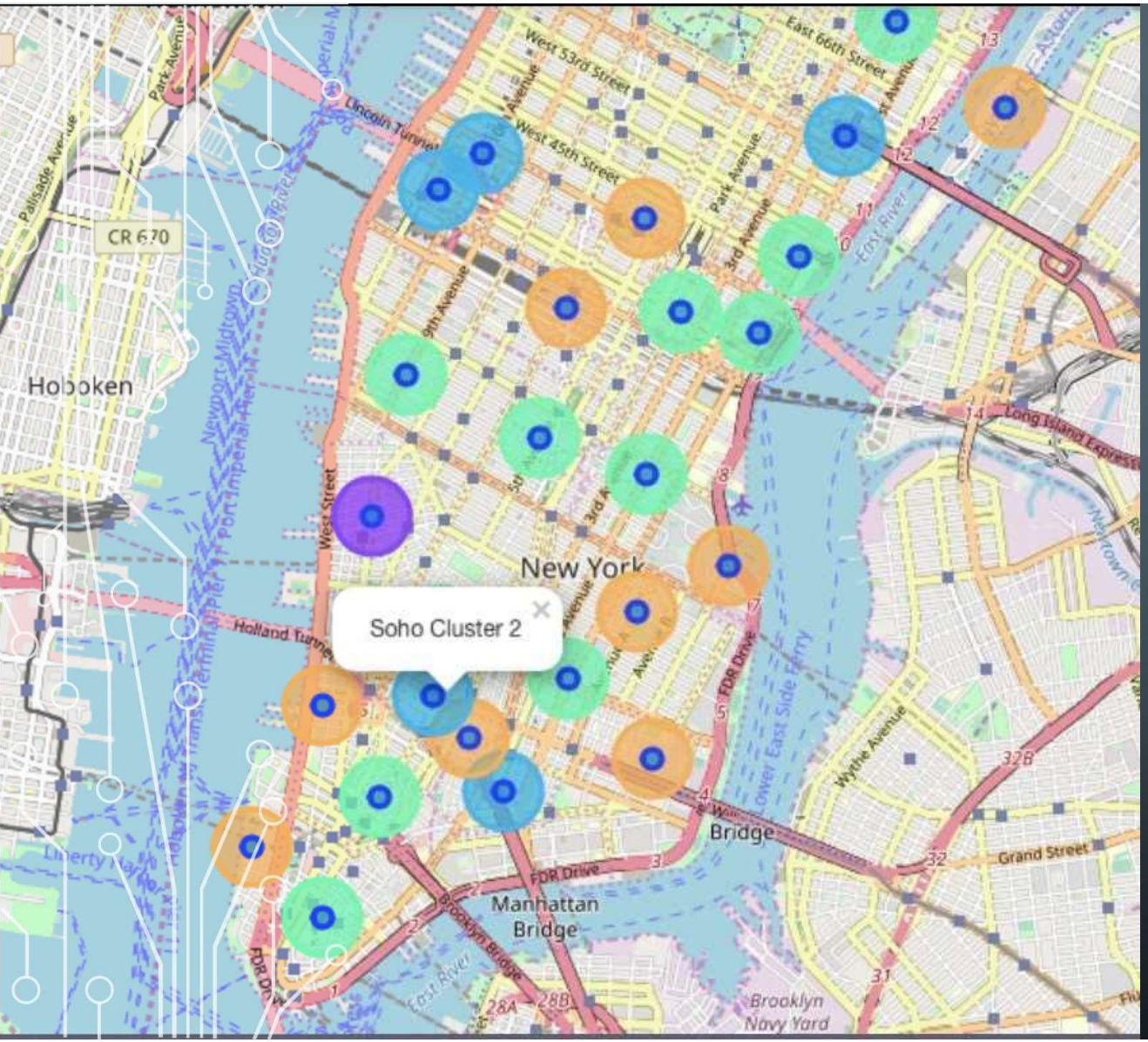
4.0 EXECUTION AND RESULTS

Current Neighborhood in Southbank Melbourne



VENUES AROUND NEIGHBORHOOD IN SOUTHBANK MELBOURNE

	name	categories	lat	lng
0	Southbank Promenade	Pedestrian Plaza	-37.819959	144.965467
1	Ponyfish Island	Bar	-37.819918	144.965021
2	Yarra River	River	-37.819684	144.965115
3	Eureka Skydeck 88	Scenic Lookout	-37.821589	144.964594
4	The Langham	Hotel	-37.820370	144.965710
5	Soho Melbourne	Italian Restaurant	-37.820609	144.963152
6	ENA greek street food	Greek Restaurant	-37.819897	144.966001
7	Waterfront Seafood-Bar-Grill	Seafood Restaurant	-37.820029	144.965557
8	Pure South	Australian Restaurant	-37.820232	144.965259
9	Broad Bean Organic Grocer	Grocery Store	-37.822588	144.966912



MANHATTAN MAP- NEIGHBORHOODS AND CLUSTER OF VENUES

```
h_rent=pd.read_csv('MH_rent_latlong.csv')
h_rent.head()
```

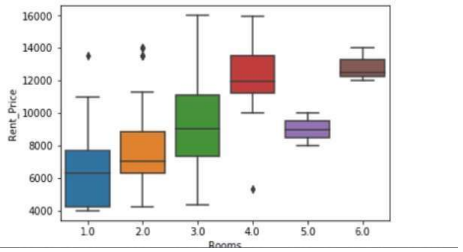
	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
0	West 105th Street	Upper West Side	2.94	5.0	3400	10000	40.799771	-73.966213
1	East 97th Street	Upper East Side	3.57	3.0	2100	7500	40.788585	-73.955277
2	West 105th Street	Upper West Side	1.89	4.0	2800	5300	40.799771	-73.966213
3	CARMINE ST.	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

```
h_rent.tail()
```

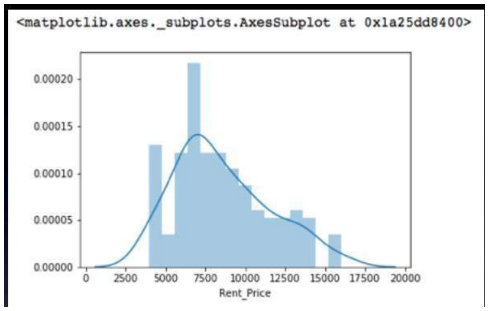
	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	
139	200 East 72nd Street	Rental in Lenox Hill	5.15	3.0	1700	8750	40
140	50 Murray Street	No fee rental in Tribeca	7.11	2.0	1223	8700	40
141	300 East 56th Street	No fee rental in Midtown East	3.87	3.0	2100	8118	40
142	1930 Broadway	No fee rental in Central Park West	5.06	2.0	1600	8095	40
143	33 West 9th Street	Rental in Greenwich Village	6.67	2.0	1500	10000	40

GEODATA MANHATTAN APS FOR RENT

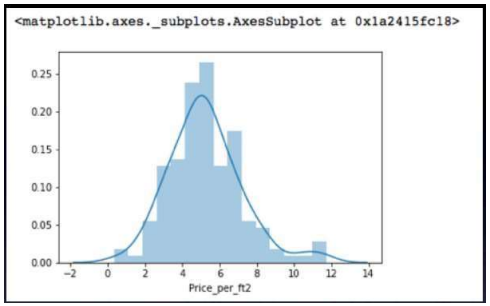
```
sns.boxplot(x='Rooms', y= 'Rent_Price', data=mh_rent)
<matplotlib.axes._subplots.AxesSubplot at 0x1a25f2a2b0>
```



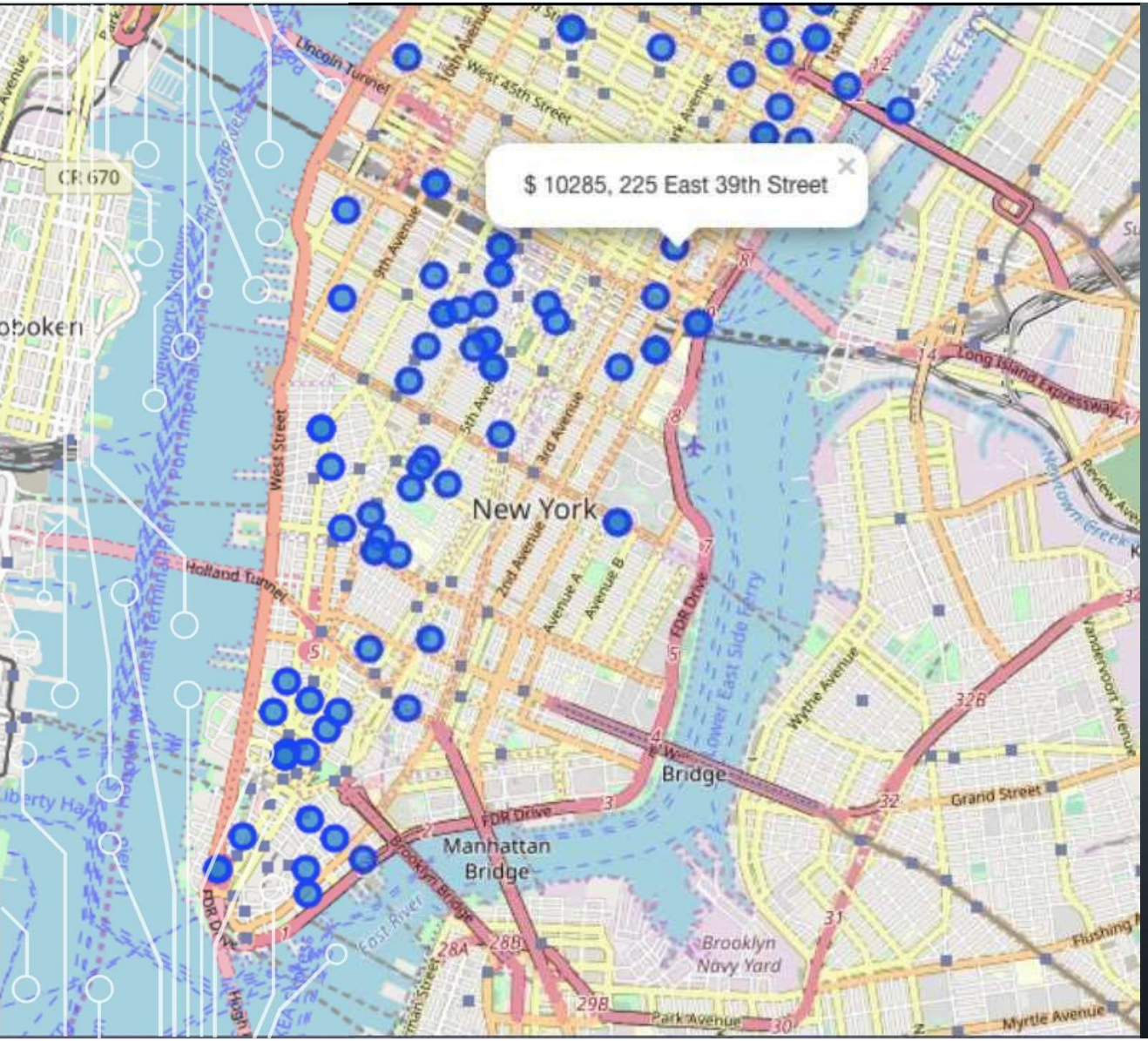

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a25dd8400>
```



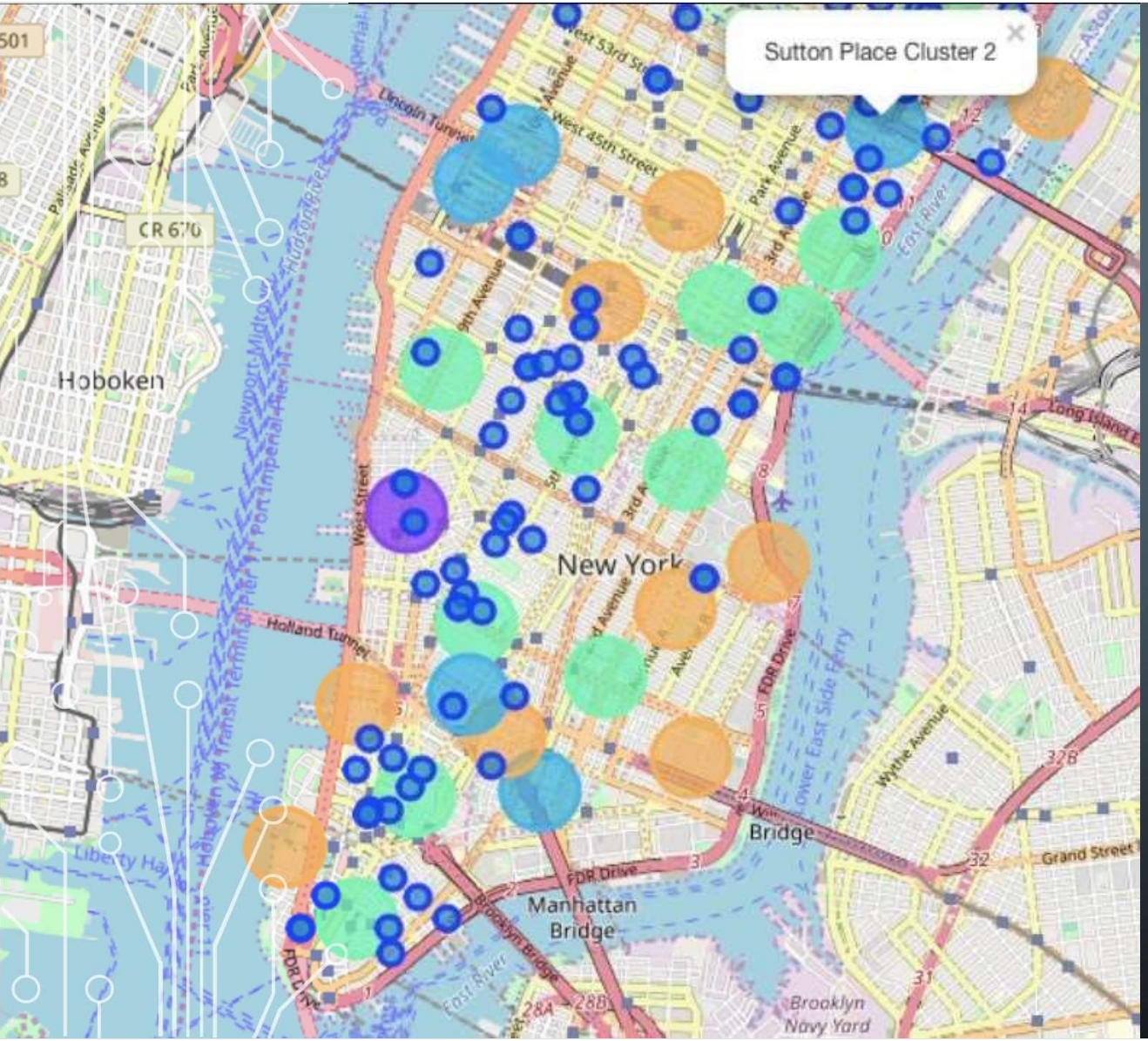

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2415fc18>
```



RENTAL PRICE STATISTICS MH APARTMENTS
RENTAL BUDGET MEANS IS AROUND $7000 USD



APARTMENTS FOR RENT IN MH



MH APARTMENTS FOR RENT WITH VENUE CLUSTERS

```
k is the cluster number to explore
 3
attan_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
Inwood	Mexican Restaurant	Lounge	Pizza Place	Café	Wine Bar	Bakery	American Restaurant	Park	Frozen Yogurt Shop	Spanish Restaurant
Manhattanville	Deli / Bodega	Re							Bike Trail	Other Nightlife
Lenox Hill	Sushi Restaurant	Re							Sporting Goods Shop	Thai Restaurant
Upper West Side	Italian Restaurant								Mexican Restaurant	Sushi Restaurant
Murray Hill	Sandwich Place								Bar	Italian Restaurant
Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Nightclub	Theater	Art Gallery	Seafood Restaurant	American Restaurant	Hotel
Greenwich Village	Italian Restaurant	Sushi Restaurant	French Restaurant	Clothing Store	Chinese Restaurant	Café	Indian Restaurant	Bakery	Seafood Restaurant	Electronics Store
Gramercy	Italian Restaurant	Restaurant	Thrift / Vintage Store	Cocktail Bar	Bagel Shop	Coffee Shop	Pizza Place	Mexican Restaurant	Grocery Store	Wine Shop
Financial District	Coffee Shop	Hotel	Gym	Wine Shop	Steakhouse	Bar	Italian Restaurant	Pizza Place	Park	Gym / Fitness Center
Noho	Italian Restaurant	French Restaurant	Cocktail Bar	Gift Shop	Bookstore	Grocery Store	Mexican Restaurant	Hotel	Sushi Restaurant	Coffee Shop

VENUES OF CLUSTER 3

MH SUBWAY STATION DATA

click to scroll output; double click to hide

		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 mhsubl
mhsubl=mh.drop_duplicates(subset=['lat','long'], keep="last").reset_index(drop=True)
mhsubl.shape
```

```
(22, 4)
```

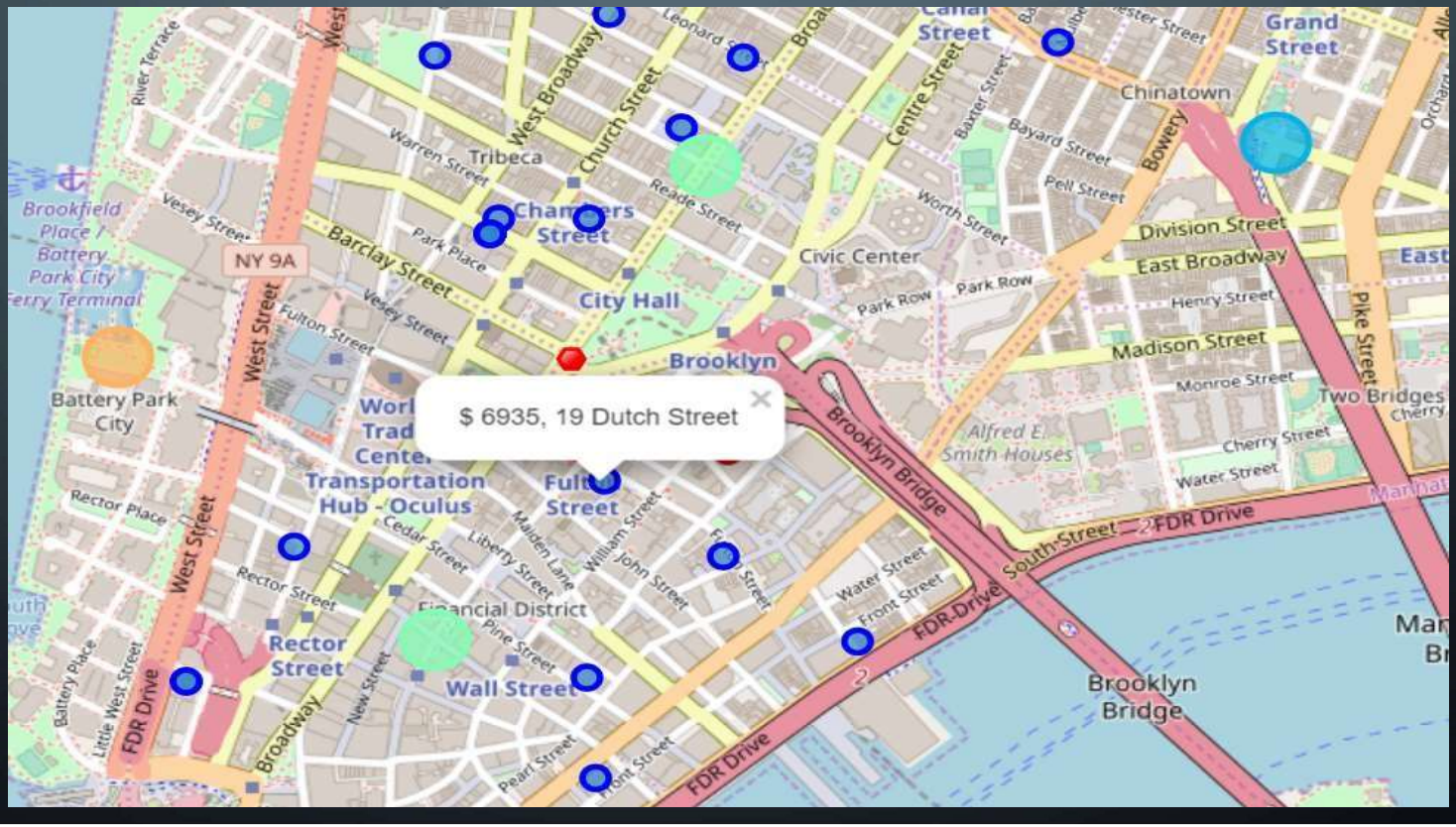
```
mhsubl.tail()
```

	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


51 St Station
BLUE IS APT FOR RENT-RED INDICATES STATIONS
CR 501
CR 678
CR 670
Hoboke
I 78
NJTP
Liberty State Park
Manhattan Bridge
Bridge
FDR Drive
Brooklyn Bridge
Grand Street
Grand Avenue
Kosciuszko Bridge
Sunnyside Yard
Broadway
Flushing Avenue
Brooklyn
West 53rd Street
West 45th Street
Lincoln Tunnel
East River
Lower East Side
Wythe Avenue
Vandervoort Avenue
58th Street
56th Road

SELECTED APARTMENT!

The ONE consolidated map shows all information for decision:
Apartments address, price, neighbourhood, cluster of venues and subway station nearby.
Blue dots=apts , Red dots=Subway station, Bubbles=Cluster of Venues



APARTMENT SELECTION

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

Apartment 1 rent cost is US7500 slightly above the US7000 budget. Apt 1 is located 400 meters from subway station at 59th Street and work place (Park Ave and 53rd) is another 600 meters way. I can walk to work place and use subway for other places around. 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 US6935, just under the US7000 budget. Apt 2 is located 60 meters from subway station at Fulton Street, but I will have to ride the subway daily to work , possibly 40-60 min ride. Venues for this apt are as of Cluster 3.

Based on current Southbank 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

I believe that convenience and location both matter a lot. Having to spend $7000 USD per month considering that I currently pay 2000 USD a month in Southbank and enjoying life means I should stay in Melbourne . 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 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 creating examples of practical cases

6.CONCLUSION

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

Final feedback on the overal 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 possiblities.

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 for use in similar situations.

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