COURSERA FINAL CAPSTONE PROJECT

COURSERA IBM DATASCIENCE CERTIFICATION

NAEEM SHAIKH APRIL 2020

REPORTCONTENT 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

- 2. Methodology

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

- 3. Results

  Discussion of results

- Discussion
  Observations leading to conclusion

- Conclusion -Final decision

1. 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. Problemstatementto resolve
- Tofind an apartment with minimumof 2 bedrooms,price of MaximumUS$7000 per monthlocated within 1.5 kilometers of subway along with great food amenities
- Interested Audience
- I believe this project is interesting for any expat deciding to migrate to the united states and would liketoleveragetoolssuchasfoursquareanddatasciencetomakeaninformeddata driven decision. The project is replicable for other cities and having a background in data science is recommended.

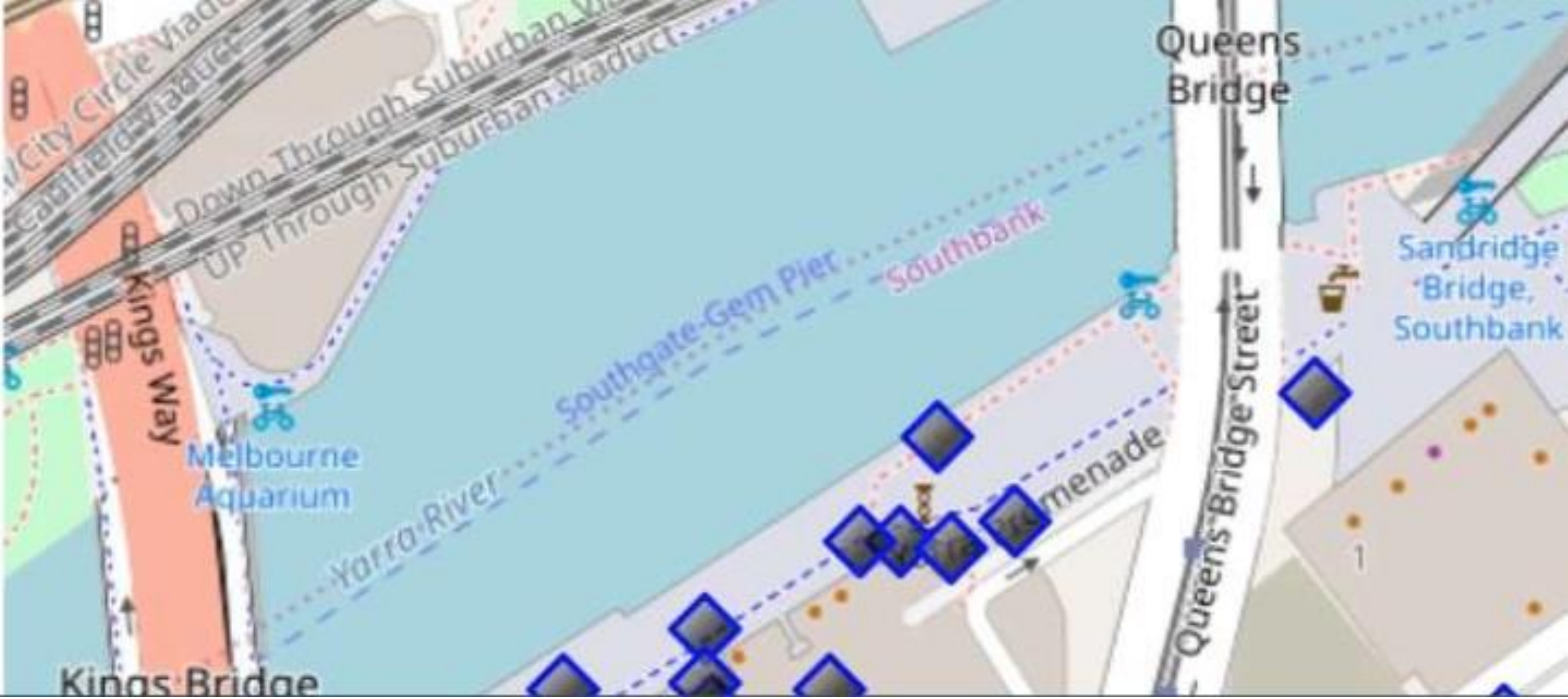
2.DATA SECTION

- **2.1 Data Requirements**
- Geodata for current residence in Southbank with venues established using Foursquare
  List of Manhattan (MH)neighbor-hoods 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 Avenueand 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. METHODOLOGY

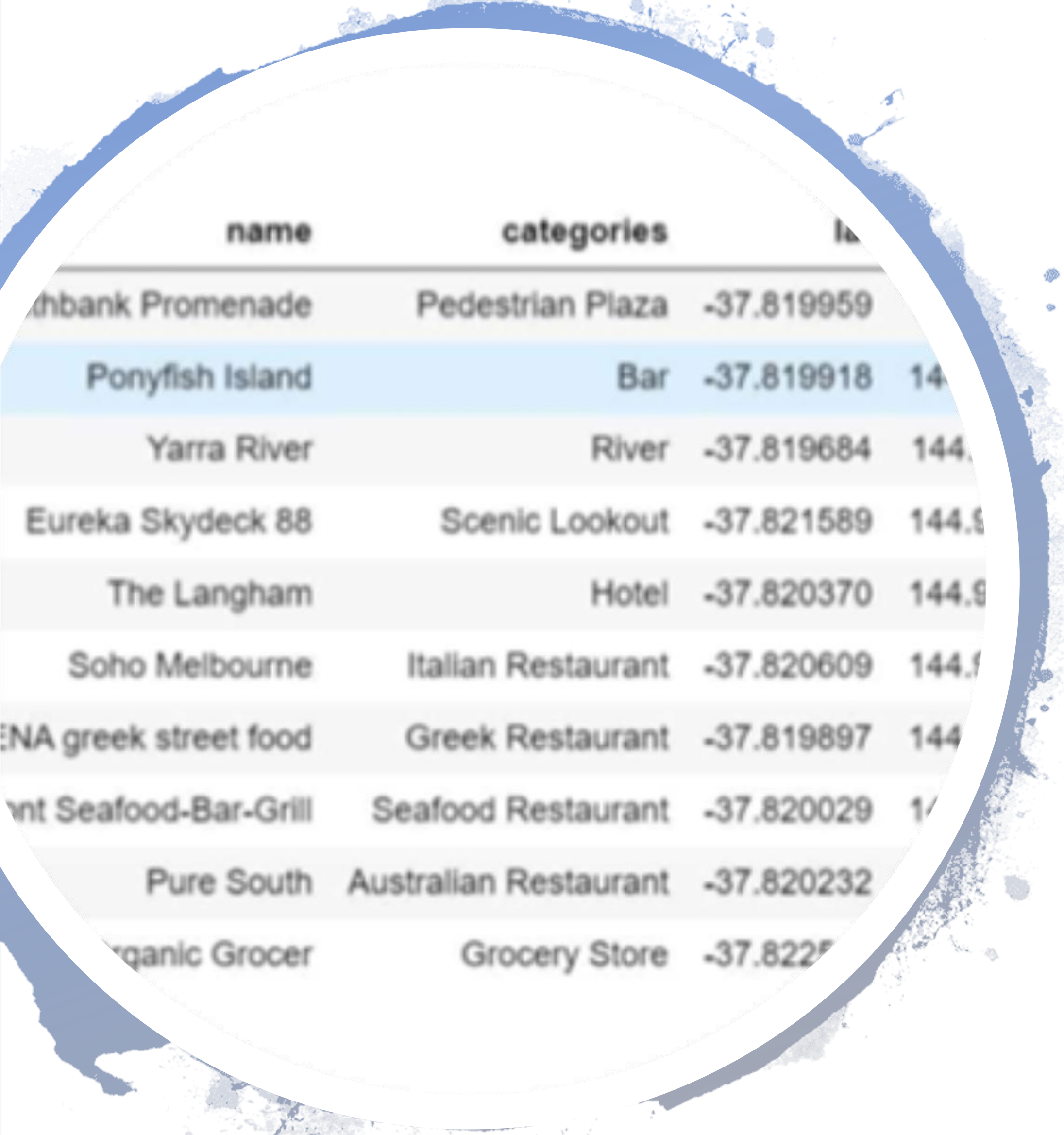
- The Strategy to find the answer:
  The strategy is based on mapping the described data in section 2.0, in order to
- facilitate the choice of at least two candidate places for rent. Theinformation will be consolidated in ONEMAPwhere one canseethe details of the apartment, the cluster of venuesin the neighborhood and the relative location from a subway station and from workplace. A measurementtool icon will also be provided. The popups onthe 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 wasobtained by coding a program to useNominatim 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. EXECUTIONAND RESULTS

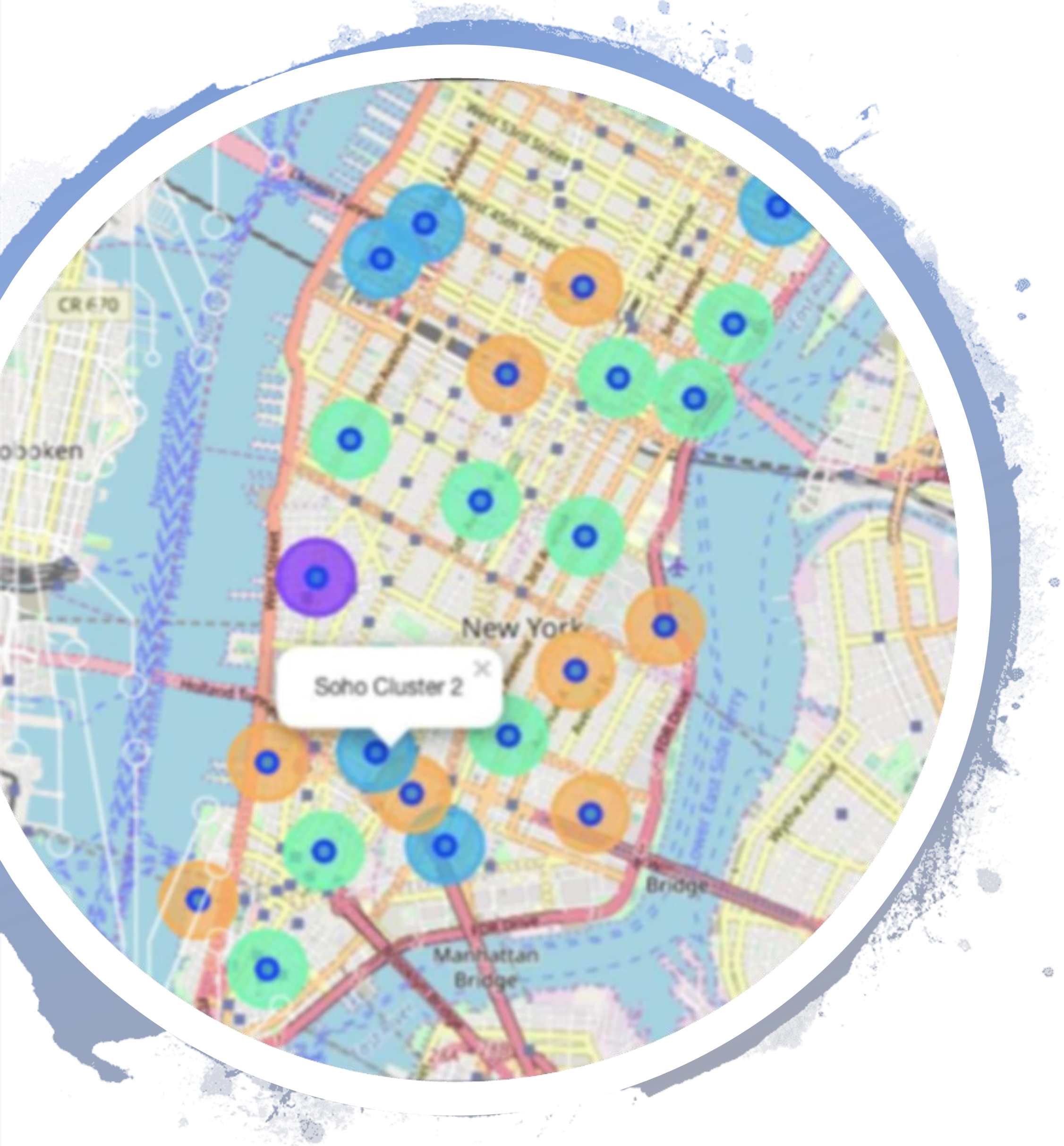


Current Neighborhood in Southbank Melbourne

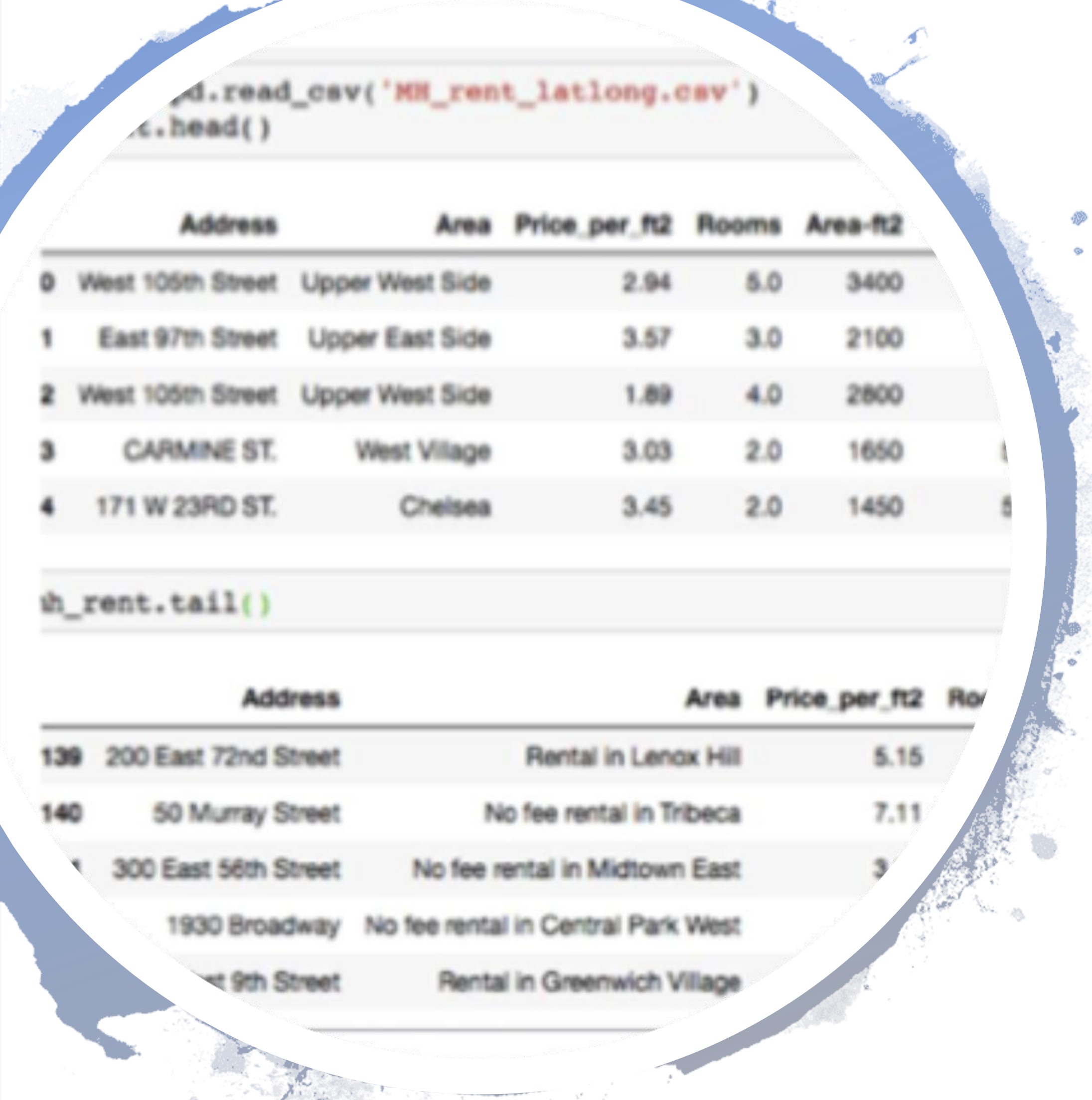
VENUES AROUND NEIGHBORHOOD IN SOUTHBANK MELBOURNE



name	categories	la	
thbank Promenade	Pedestrian Plaza	-37.819959	
Ponyfish Island	Bar	-37.819918	14
Yarra River	River	-37.819684	144.
Eureka Skydeck 88	Scenic Lookout	-37.821589	144.9
The Langham	Hotel	-37.820370	144.9
Soho Melbourne	Italian Restaurant	-37.820609	144.
NA greek street food	Greek Restaurant	-37.819897	144
nt Seafood-Bar-Grill	Seafood Restaurant	-37.820029	14
Pure South	Australian Restaurant	-37.820232	
rganic Grocer	Grocery Store	-37.822	



MANHATTAN MAP- NEIGHBORHOOD'S AND CLUSTEROF VENUES



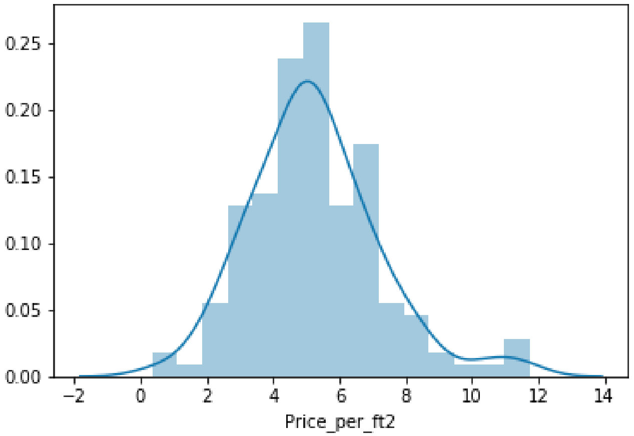
```
d.read_csv('MH_rent_latlong.csv')
.head()
```

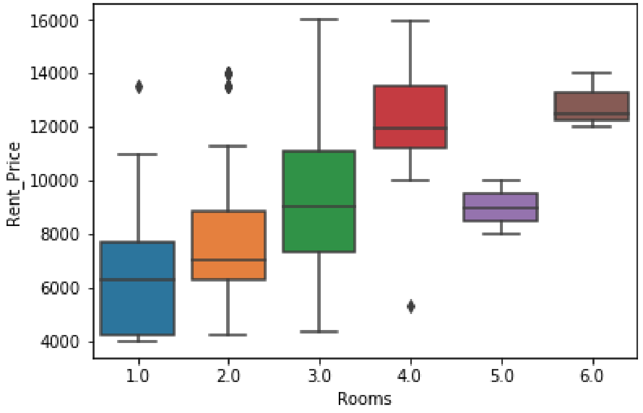
	Address	Area	Price_per_ft2	Rooms	Area-ft2
0	West 105th Street	Upper West Side	2.94	5.0	3400
1	East 97th Street	Upper East Side	3.57	3.0	2100
2	West 105th Street	Upper West Side	1.89	4.0	2800
3	CARMINE ST.	West Village	3.03	2.0	1650
4	171 W 23RD ST.	Chelsea	3.45	2.0	1450

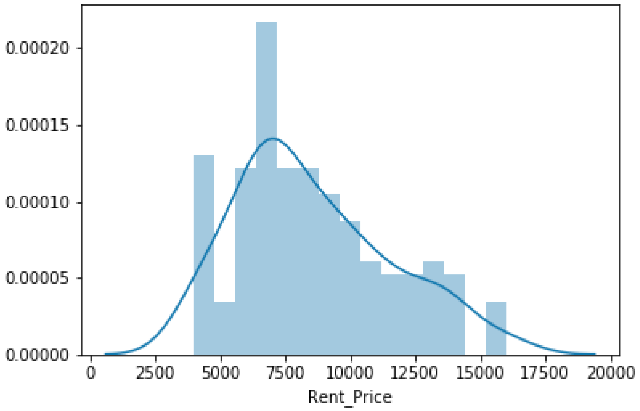
```
h_rent.tail()
```

	Address	Area	Price_per_ft2	Ro
139	200 East 72nd Street	Rental in Lenox Hill	5.15	
140	50 Murray Street	No fee rental in Tribeca	7.11	
	300 East 56th Street	No fee rental in Midtown East	3	
	1930 Broadway	No fee rental in Central Park West		
	t 9th Street	Rental in Greenwich Village		

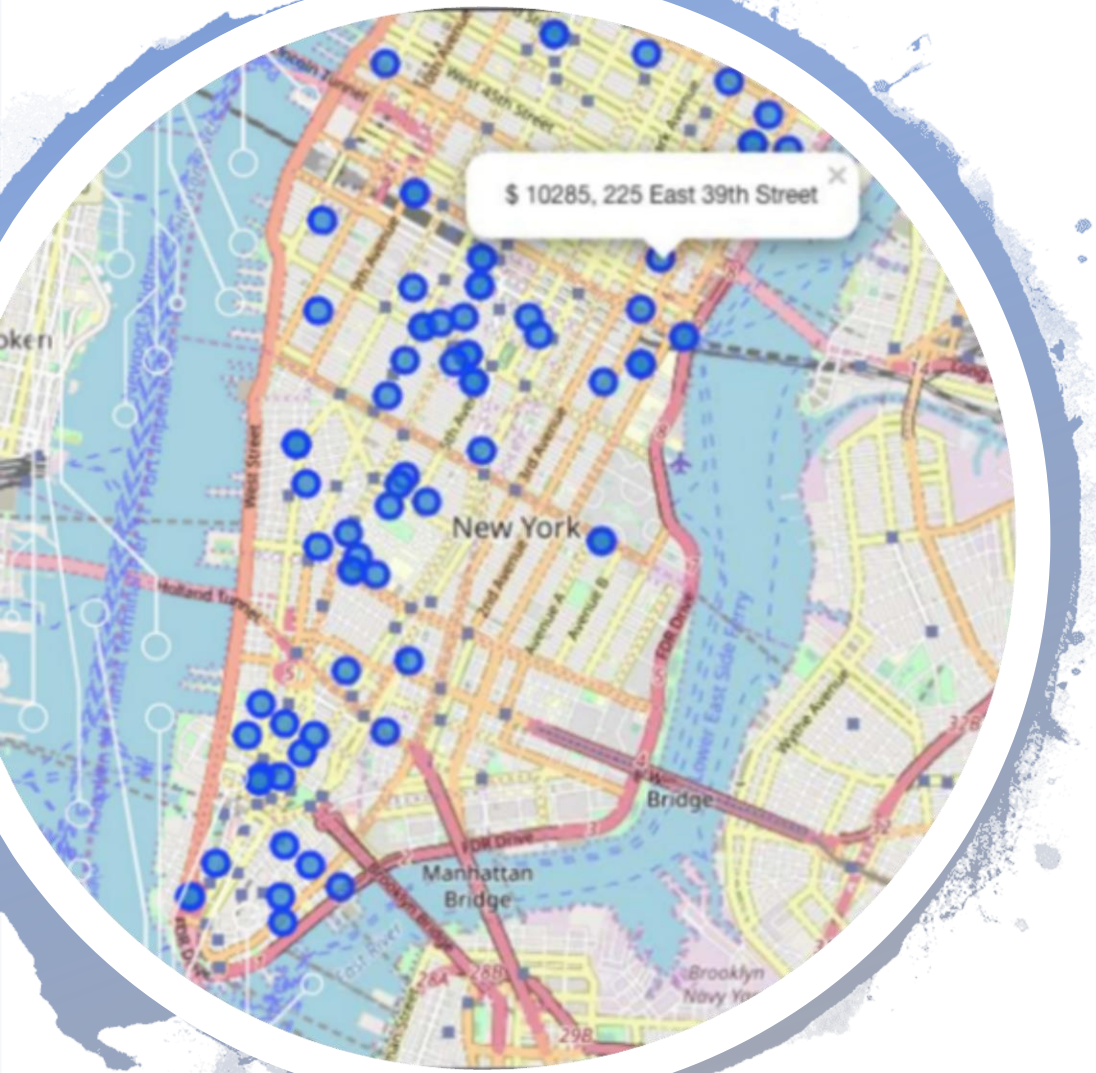
GEO DATA MANHATTAN APS FOR RENT



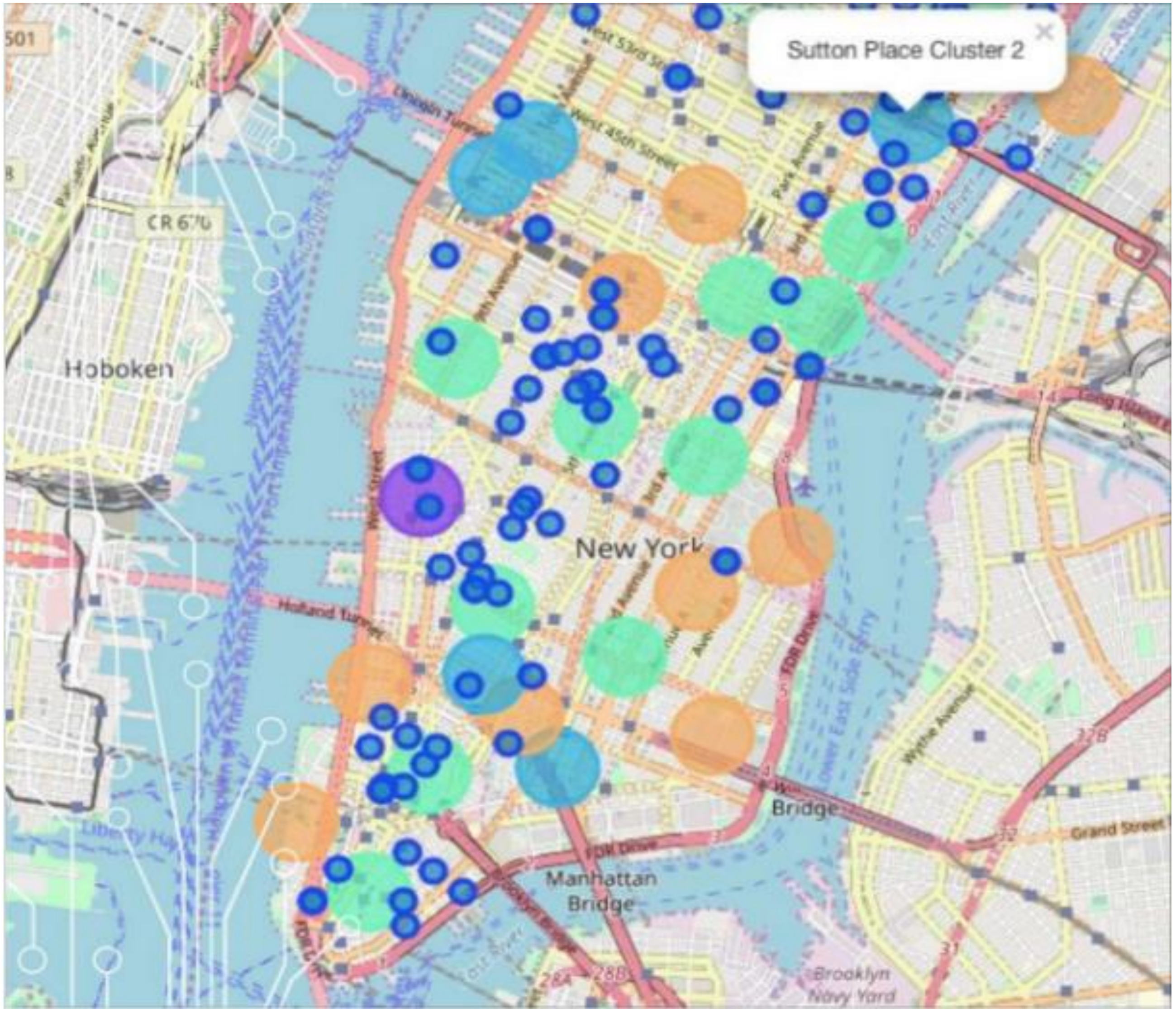




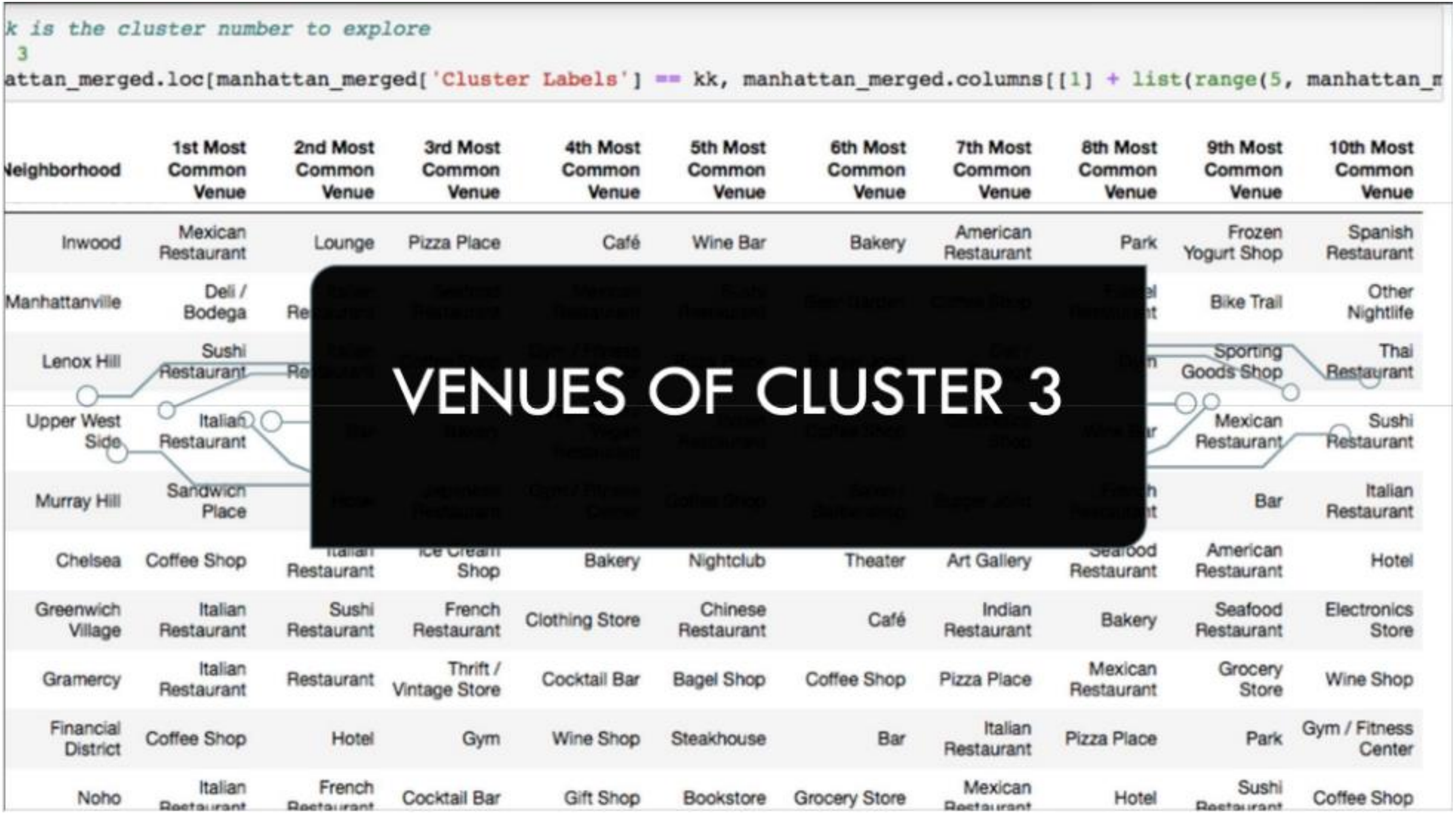
RENTAL PRICE STATISTICS MH APARTMENTS RENTAL BUDGET MEANS IS AROUND $7000USD



APARTMENTS FOR RENT IN MH



MH APARTMENTS FOR RENT WITH VENUE CLUSTERS

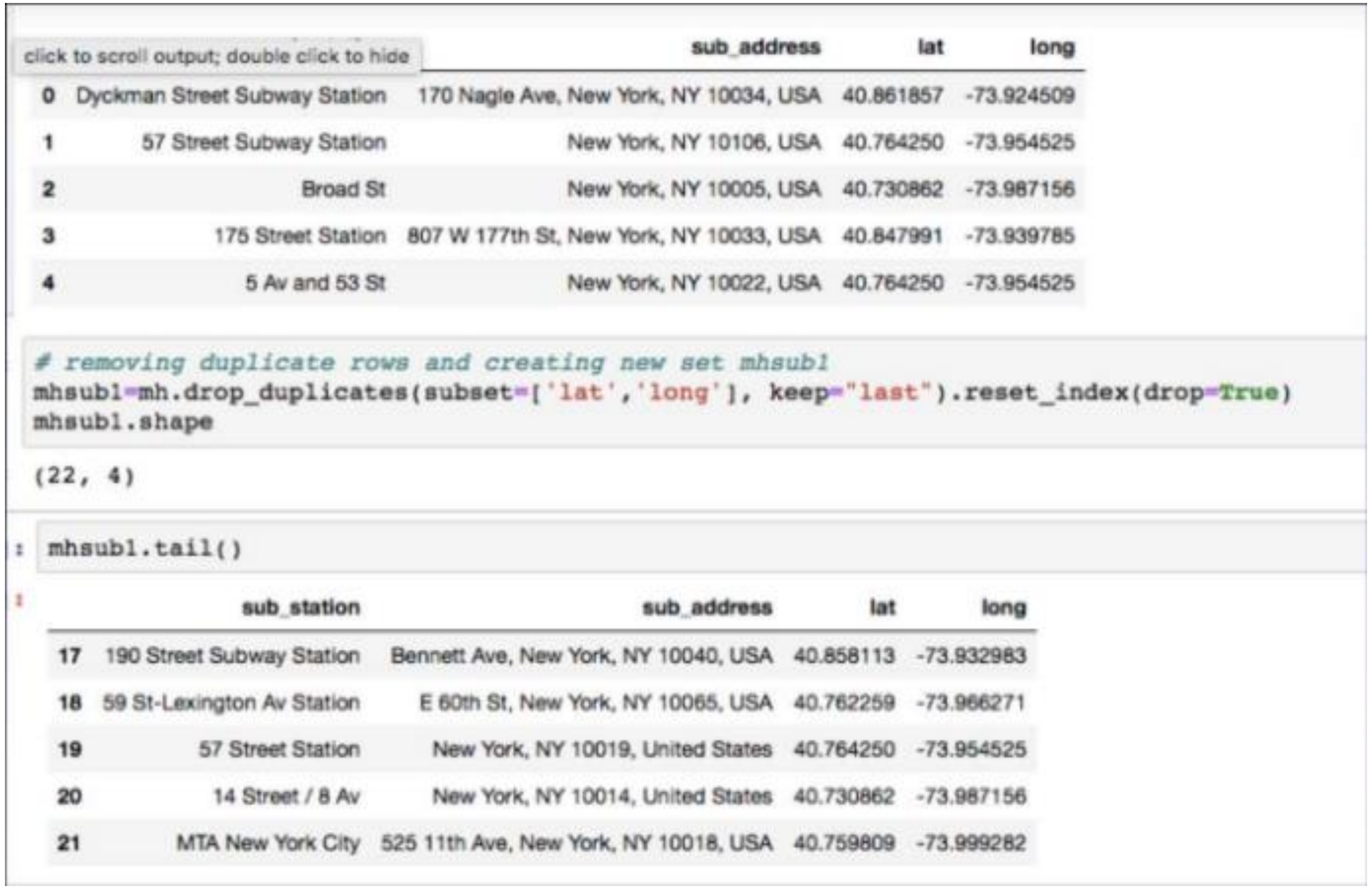


VENUES OF CLUSTER 3

```
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	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	Sushi Restaurant	Beer Garden	Coffee Shop	Falafel Restaurant	Bike Trail	Other Nightlife
Lenox Hill	Sushi Restaurant	Italian Restaurant	Coffee Shop	Gym / Fitness Center	Pizza Place	Burger Joint	Deli / Bodega	Gym	Sporting Goods Shop	Thai Restaurant
Upper West Side	Italian Restaurant	Bar	Bakery	Vegetarian / Vegan Restaurant	Indian Restaurant	Coffee Shop	Ice Cream Shop	Wine Bar	Mexican Restaurant	Sushi Restaurant
Murray Hill	Sandwich Place	Hotel	Japanese Restaurant	Gym / Fitness Center	Coffee Shop	Salon / Barbershop	Burger Joint	French Restaurant	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

MH SUBWAY STATION DATA



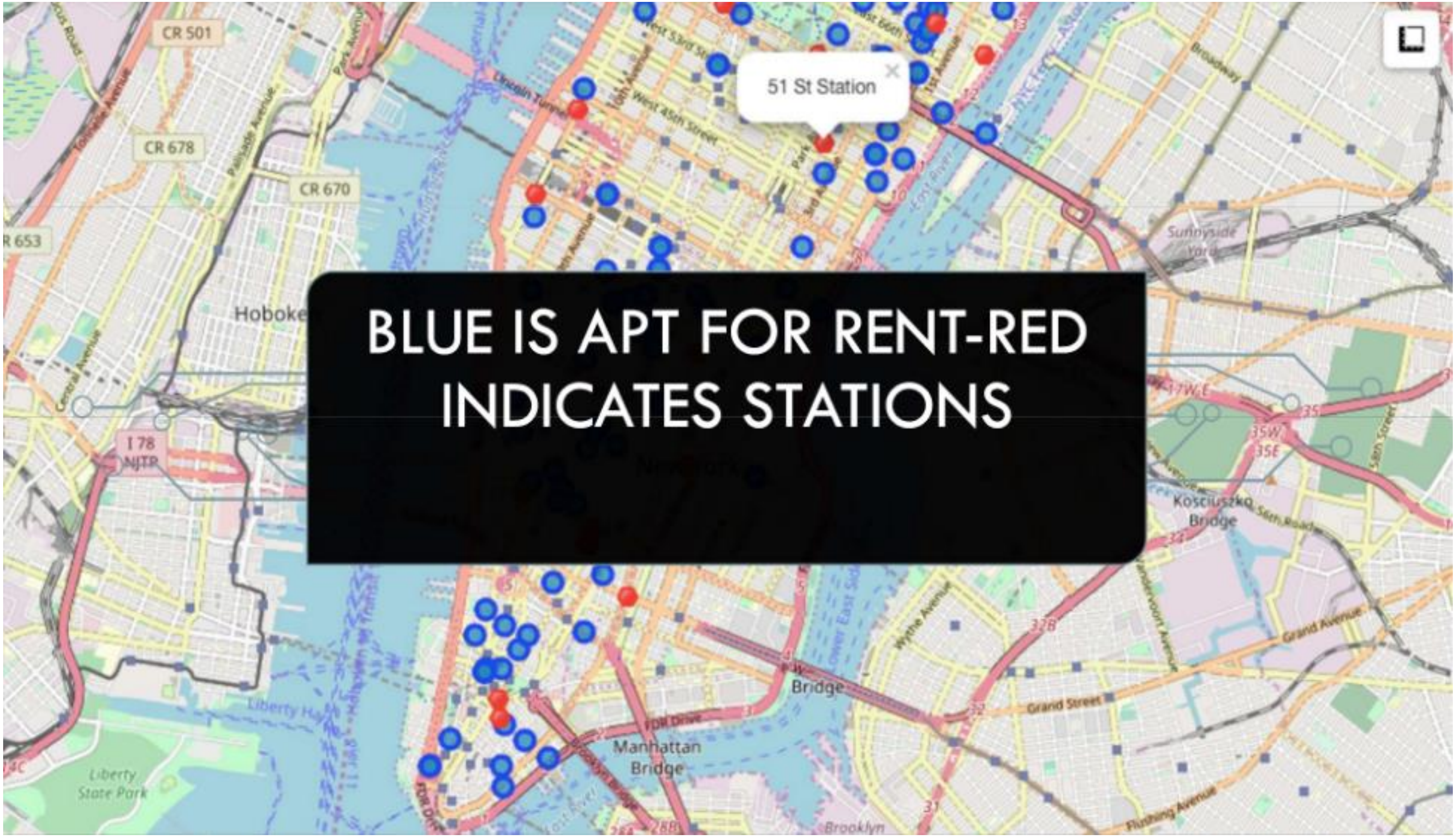
		sub_address	lat	long
0	Dyckman Street Subway Station	170 Nagle Ave, New York, NY 10034, USA	40.861857	-73.924509
1	57 Street Subway Station	New York, NY 10106, USA	40.764250	-73.954525
2	Broad St	New York, NY 10005, USA	40.730862	-73.987156
3	175 Street Station	807 W 177th St, New York, NY 10033, USA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



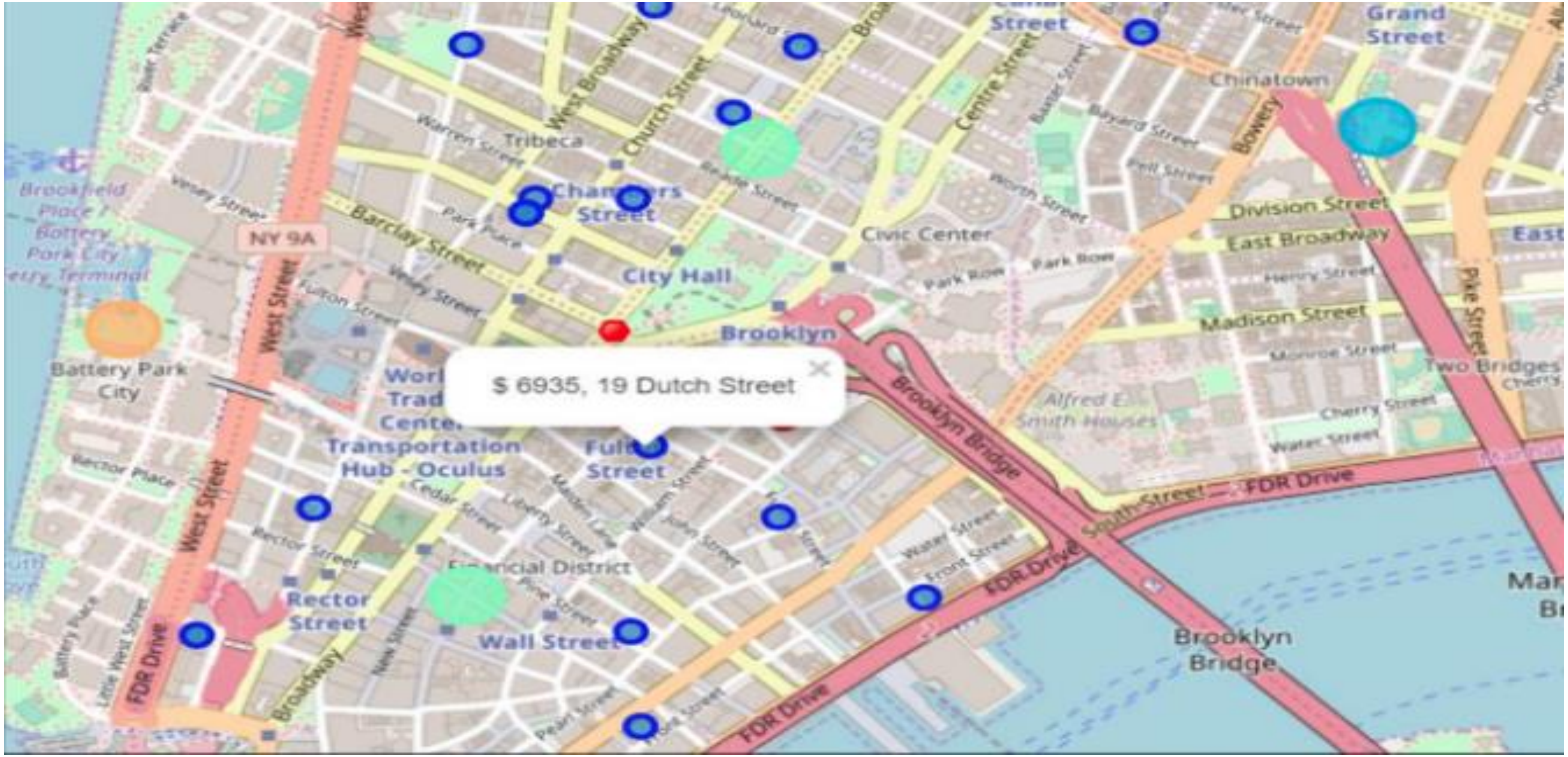
BLUE IS APT FOR RENT-RED INDICATES STATIONS

SELECTED APARTMENT

The ONE consolidated map shows all information for decision:

Apartments address, price, neighbour-hood, cluster of venues and subway station nearby.

Blue dots=apts, Red dots=subway station, Bubles = 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 in 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 organization, 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 have decided to move to the US and stay in Melbourne considering the prices. I will explore Los Angeles for future career opportunities and run the same cost benefit analysis to make an informed data driven decision.

- Final feedback on the overall data science course
- I am very happy to be able to complete the 9 course specialization within couple of months.
- The mapping with Folium is a very powerful technique to consolidates 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 outstanding course.