Using Data Science to select a hotel for my parents’ visit

A DATA SCIENCE CAPSTONE PROJECT BY SARIM HASSAN

Table of Contents

Introduction

SCENARIO SET-UP AND PROBLEM DEFINITION

Data Planning

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Methodology & Results

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Discussion

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Problem Resolution

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Introduction

SCENARIO SET-UP AND PROBLEM DEFINITION

Scenario and Introduction:

I will be graduating during the week of May 21 from New York University. I am inviting my family to spend a week in New York City to attend the graduation ceremony and spend some time travelling and exploring the city. My family likes to explore cultural sites within the city. I intend to use location-based data and data science techniques to select the best possible hotel for my duration that addresses their travelling and comfort needs.

Problem Definition:

- The hotel selected should have the following characteristics
- The total cost of stay for a week should be less than the budget of USD 1000
- The hotel should at least have a rating of 3 stars
- The hotel should be in a neighborhood located close to a cluster of culturally significant tourist attractions like museums, art galleries, auditoriums etc.
- The hotel should be located within walking distance to a subway station
- The hotel should not be more than 1 hour away from the closest airport
- The hotel should be in a neighborhood with a high proportion of Indian restaurants and halal food options

Data Planning

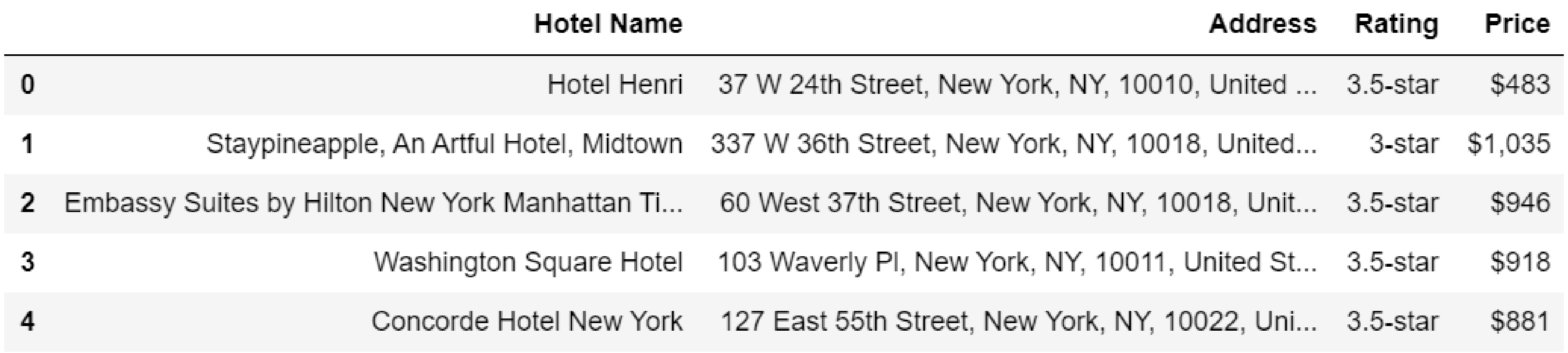
DATA SOURCES, PRE-PROCESSING AND PROCESSING

Data Sources and Pre-Processing

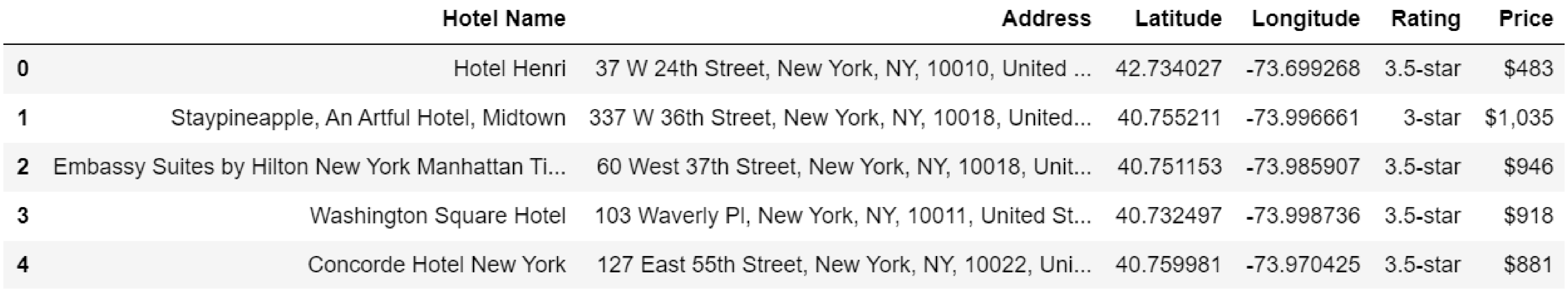
- List of neighborhoods was obtained from the NYU catalog and the same file used in the course lab with the requisite latitude and longitude information was used
- The list of subway stations was obtained from developer data from MTA
- The hotels data was generated by going to Hotels.com, entering the criteria for the dates, the amenities, and the start-rating of hotels under consideration. The BeautifulSoup package was then used in conjunction with the selenium and chromedriver packages to scrape the data as the page was had an infinite scrolling layout which was difficult to scrape from the IBM Watson studio environment. The code was then run on my local machine and the output csv file is used as the starting point in this code base.
- Location data for hotels and neighborhoods was obtained using the nominatim package

Hotels Data Pre-processing

Hotels data was first scraped from Hotels.com using BeautifulSoup and geo co-ordinates added using GeoPy



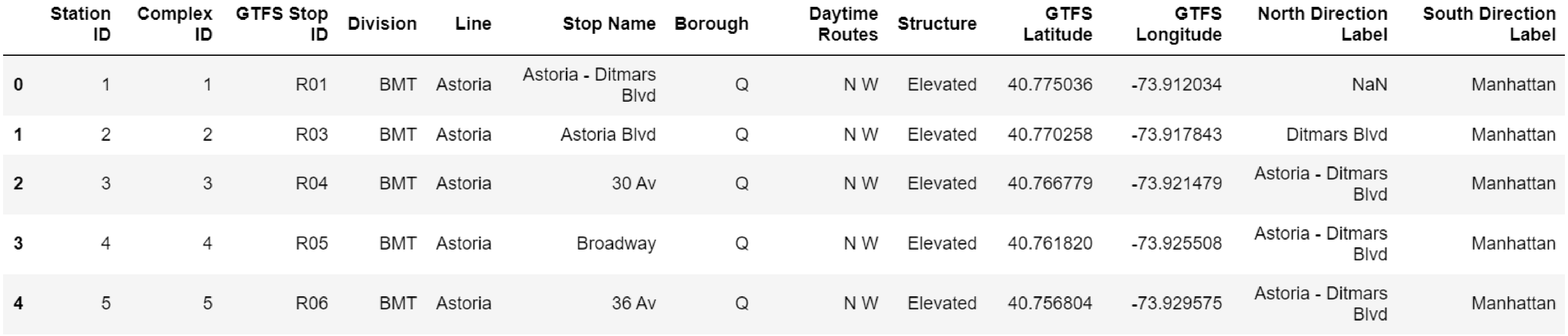
	Hotel Name	Address	Rating	Price
0	Hotel Henri	37 W 24th Street, New York, NY, 10010, United ...	3.5-star	$483
1	Staypineapple, An Artful Hotel, Midtown	337 W 36th Street, New York, NY, 10018, United...	3-star	$1,035
2	Embassy Suites by Hilton New York Manhattan Ti...	60 West 37th Street, New York, NY, 10018, Unit...	3.5-star	$946
3	Washington Square Hotel	103 Waverly Pl, New York, NY, 10011, United St...	3.5-star	$918
4	Concorde Hotel New York	127 East 55th Street, New York, NY, 10022, Uni...	3.5-star	$881



	Hotel Name	Address	Latitude	Longitude	Rating	Price
0	Hotel Henri	37 W 24th Street, New York, NY, 10010, United ...	42.734027	-73.699268	3.5-star	$483
1	Staypineapple, An Artful Hotel, Midtown	337 W 36th Street, New York, NY, 10018, United...	40.755211	-73.996661	3-star	$1,035
2	Embassy Suites by Hilton New York Manhattan Ti...	60 West 37th Street, New York, NY, 10018, Unit...	40.751153	-73.985907	3.5-star	$946
3	Washington Square Hotel	103 Waverly Pl, New York, NY, 10011, United St...	40.732497	-73.998736	3.5-star	$918
4	Concorde Hotel New York	127 East 55th Street, New York, NY, 10022, Uni...	40.759981	-73.970425	3.5-star	$881

Subway Data Pre-processing

The developer data from the MTA dataset had Latitude and Longitude information already. So this dataset was ready to use in its current condition for our purposes



	Station ID	Complex ID	GTFS Stop ID	Division	Line	Stop Name	Borough	Daytime Routes	Structure	GTFS Latitude	GTFS Longitude	North Direction Label	South Direction Label
0	1	1	R01	BMT	Astoria	Astoria - Ditmars Blvd	Q	N W	Elevated	40.775036	-73.912034	NaN	Manhattan
1	2	2	R03	BMT	Astoria	Astoria Blvd	Q	N W	Elevated	40.770258	-73.917843	Ditmars Blvd	Manhattan
2	3	3	R04	BMT	Astoria	30 Av	Q	N W	Elevated	40.766779	-73.921479	Astoria - Ditmars Blvd	Manhattan
3	4	4	R05	BMT	Astoria	Broadway	Q	N W	Elevated	40.761820	-73.925508	Astoria - Ditmars Blvd	Manhattan
4	5	5	R06	BMT	Astoria	36 Av	Q	N W	Elevated	40.756804	-73.929575	Astoria - Ditmars Blvd	Manhattan

Methodology & Results

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Data Visualization

Hotels Data

The hotels data was analyzed to visualize the geographical spread of the hotels and quickly view characteristics such as price and star rating on the pop-ups on the folium map

Techniques used: Data Visualization, Web Scraping

Subway Data

The subway data was plotted on the map concurrently with hotels and neighborhood information to aid in the final process of selecting the right hotel. The hotel must be close to subways and different lines

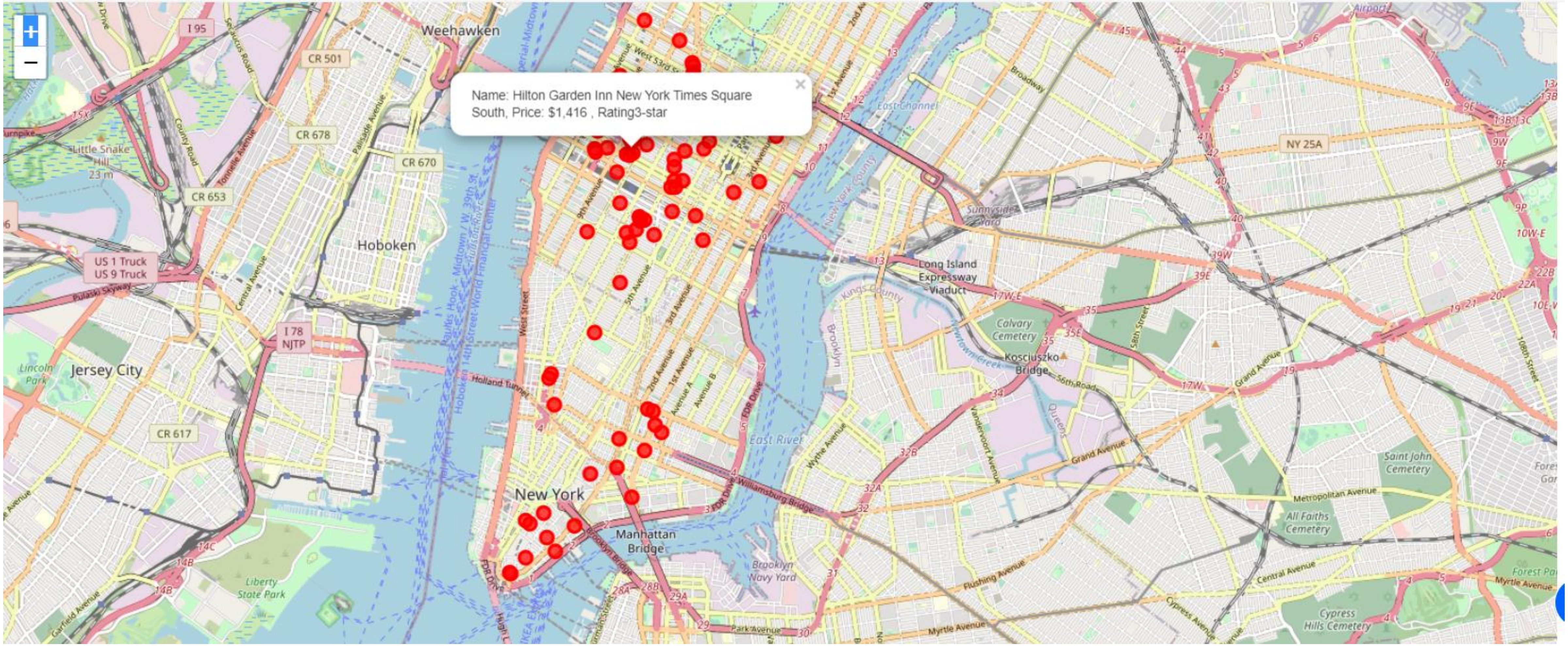
Techniques used: Data Visualization, Web Scraping

Neighborhoods Data

The filtered Manhattan neighborhoods data was plotted on a folium map and then a consolidated map was created to show the neighborhoods, hotels and subway stations before the clustering analysis

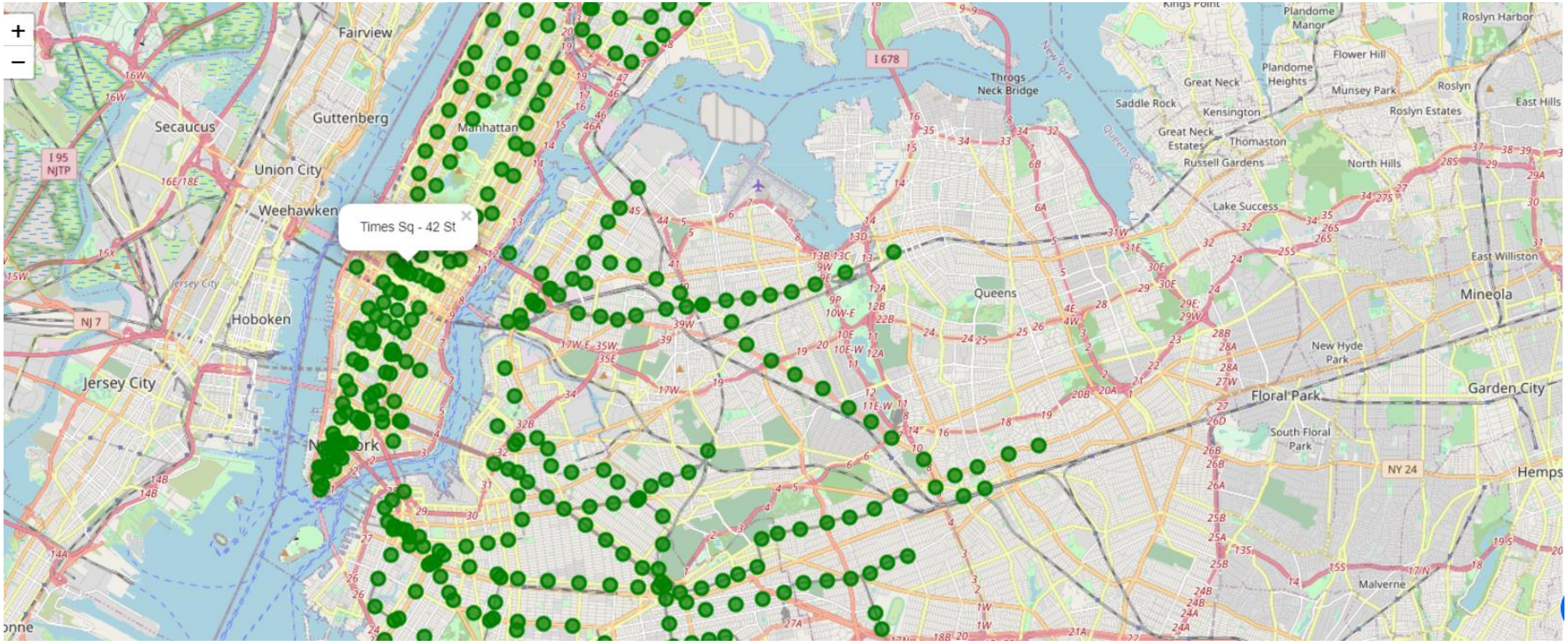
Techniques used: Data Visualization, Web Scraping

Hotels in Manhattan w/ Price information

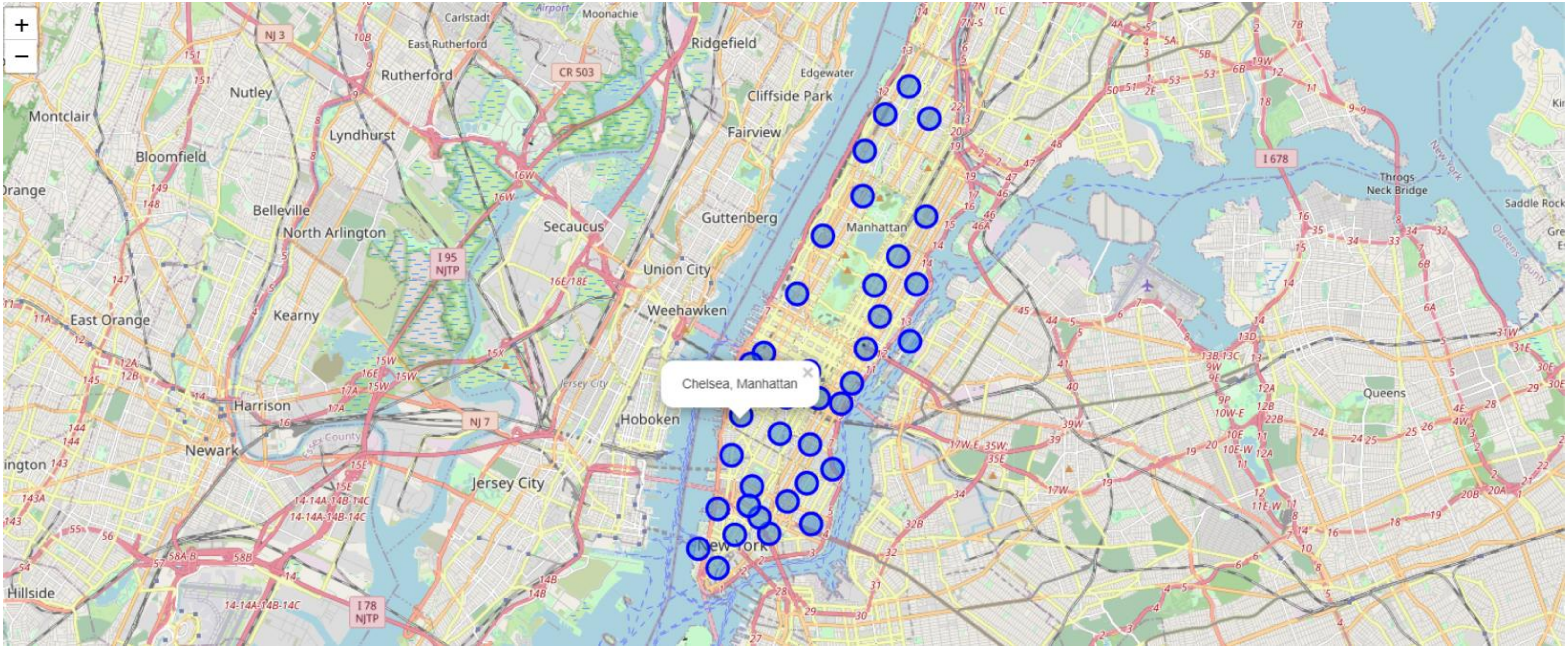


Name: Hilton Garden Inn New York Times Square South, Price: $1,416 , Rating3-star

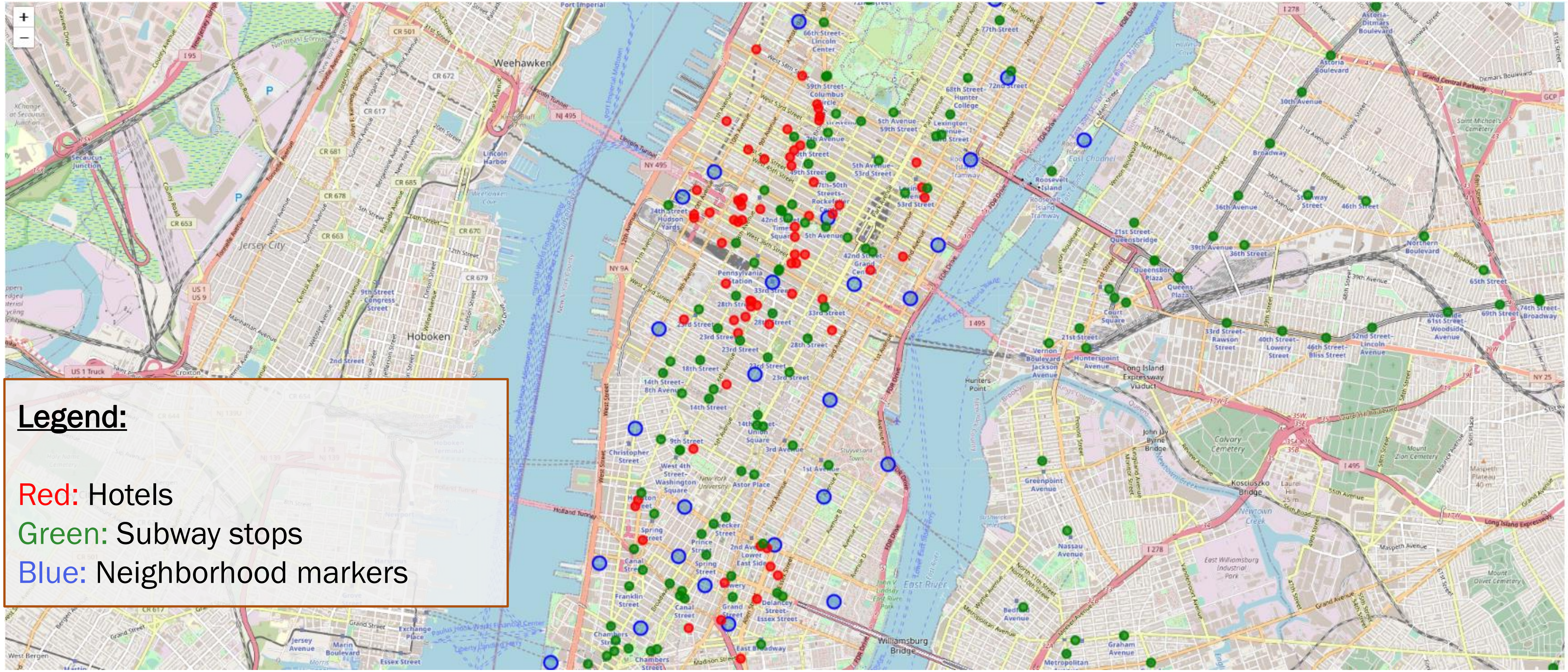
Manhattan Subway Map



Manhattan Neighborhoods Visualized



Consolidated Map



Legend:

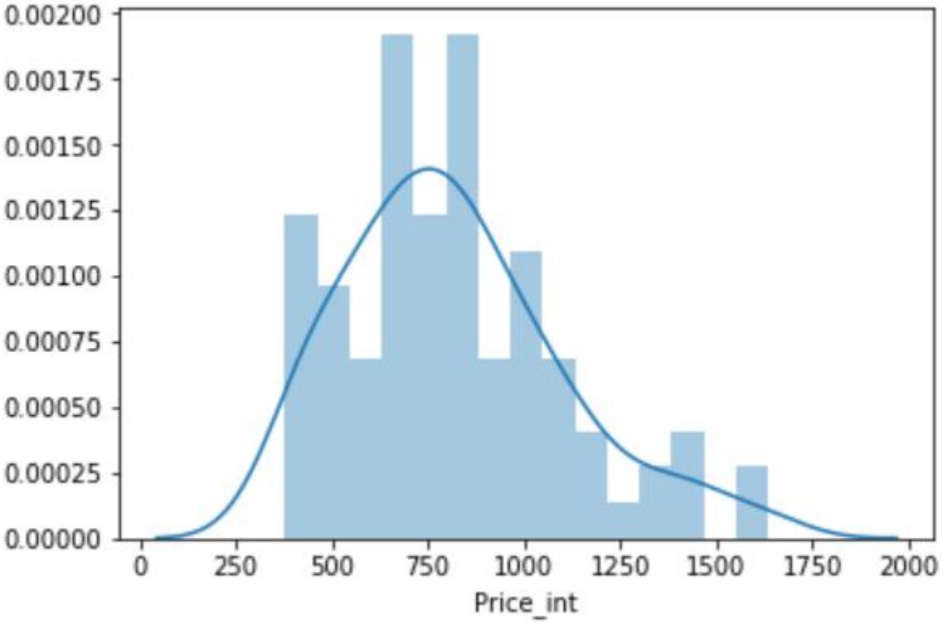
Red: Hotels
Green: Subway stops
Blue: Neighborhood markers

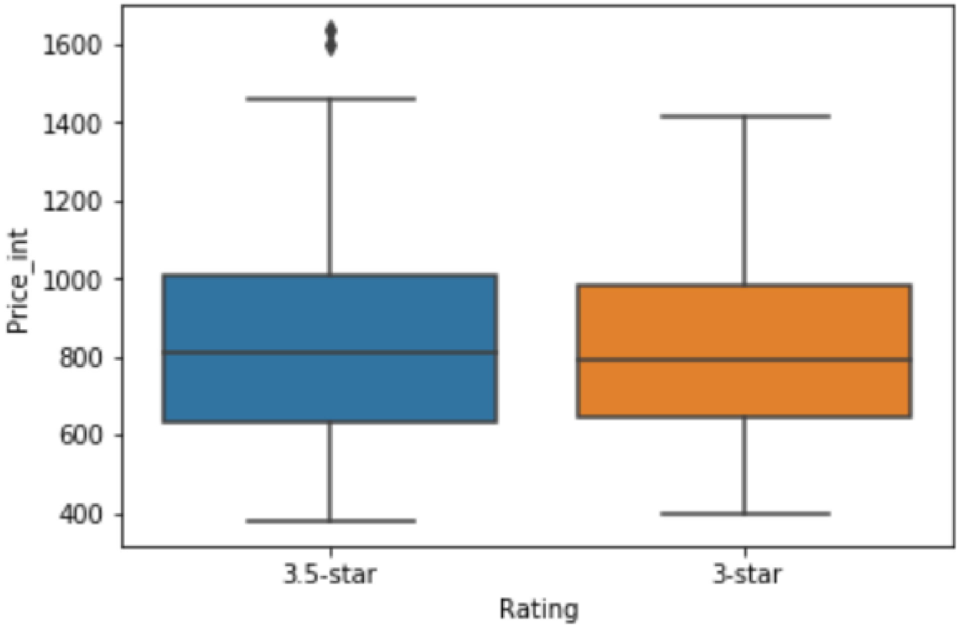
Exploratory Analysis

Hotels Data

- *Analyzing distribution of prices in Manhattan hotels for the date ranges specified in the problem statement*
- *Analyzing distribution of prices and the relationship with star rating*

Hotel Price Statistics





- Our budget of USD 1000 falls in the higher range of the histogram.
- The average price for 3 and 3.5 star rated hotels in our dataset is about the same

Clustering Analysis

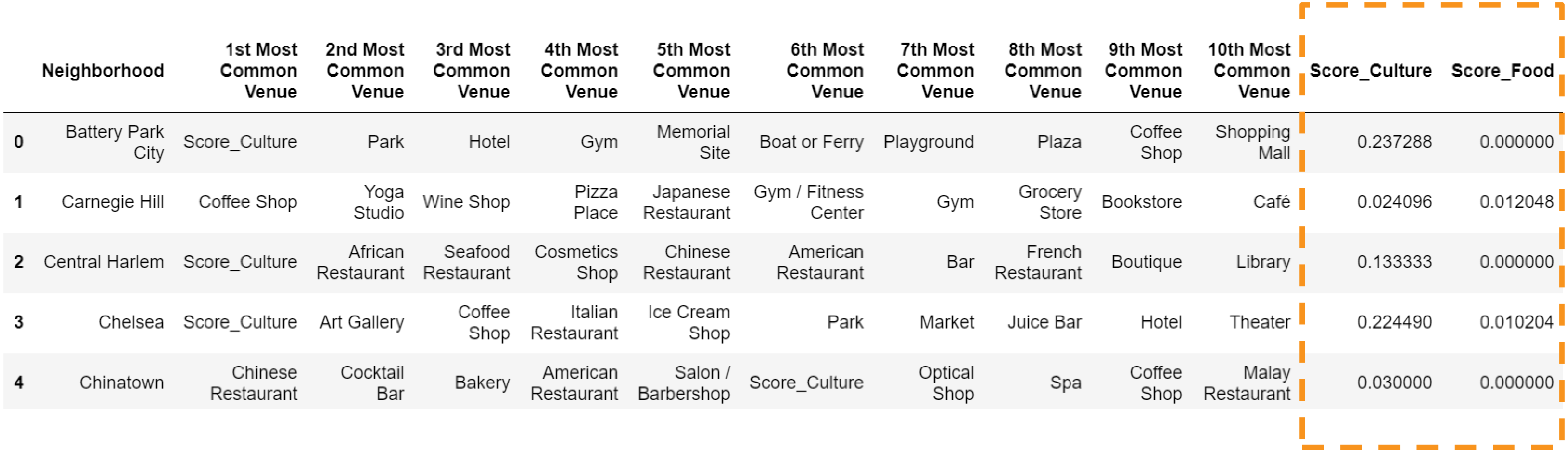
Manhattan Neighborhood Cluster Analysis – Concept

The Manhattan neighborhoods will be analyzed to identify clusters of neighborhoods on the basis of two major considerations:

- **Proliferation of cultural venues:** The neighborhood for the duration of stay should be in a location which is close to a high number of spots that are steeped in the city's history and culture. It would also include numerous categories of venues frequented by tourists like monuments, landmarks, tourist information centres etc.

- **Presence of preferred food cuisine options:** My parents like to eat most of their meals while travelling in a familiar cuisine. It is important that the food locations have Halal restaurants, so they encounter less restrictions while ordering food.

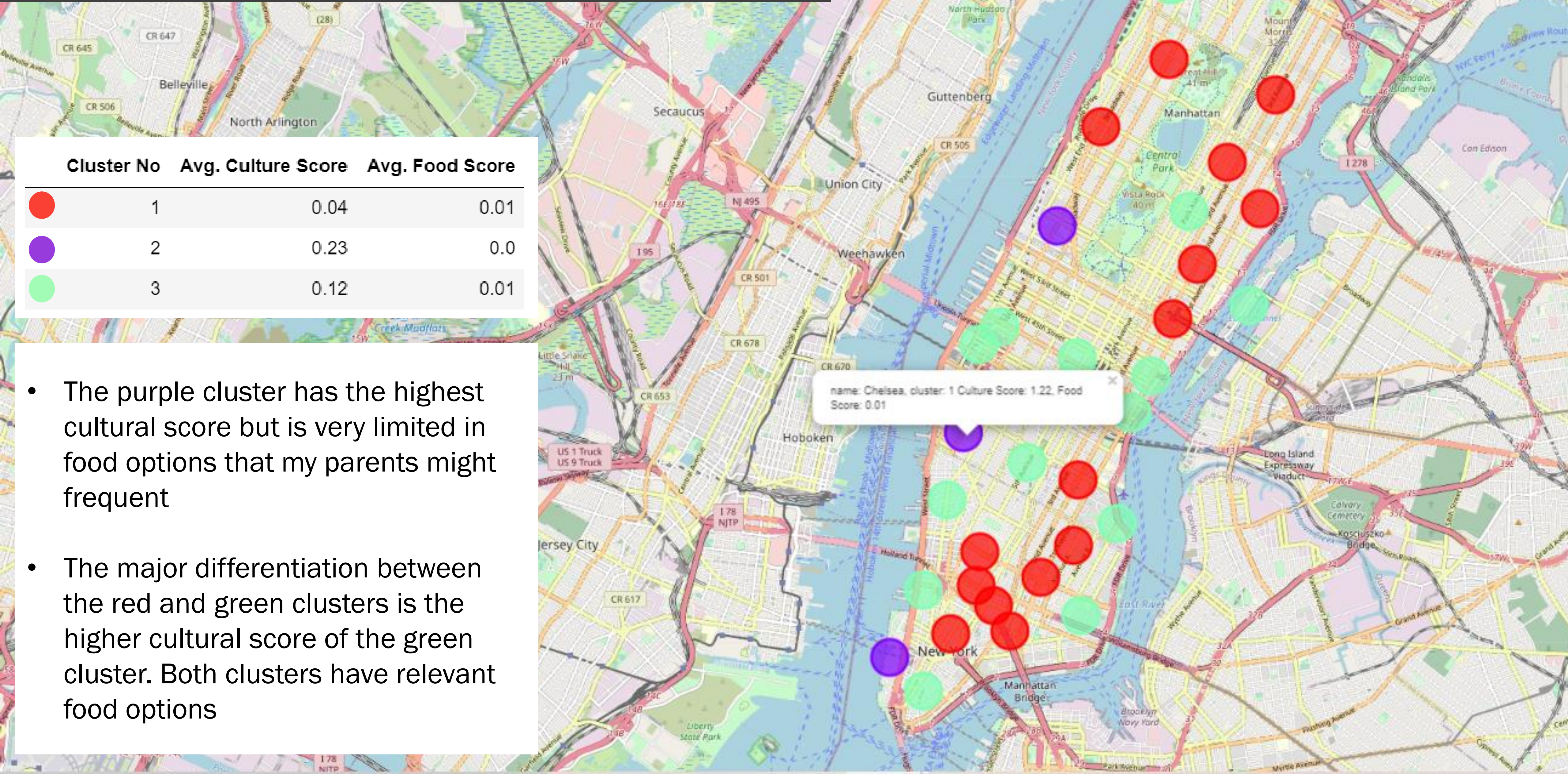
Preparing dataframe for clustering



	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	Score_Culture	Score_Food
0	Battery Park City	Score_Culture	Park	Hotel	Gym	Memorial Site	Boat or Ferry	Playground	Plaza	Coffee Shop	Shopping Mall	0.237288	0.000000
1	Carnegie Hill	Coffee Shop	Yoga Studio	Wine Shop	Pizza Place	Japanese Restaurant	Gym / Fitness Center	Gym	Grocery Store	Bookstore	Café	0.024096	0.012048
2	Central Harlem	Score_Culture	African Restaurant	Seafood Restaurant	Cosmetics Shop	Chinese Restaurant	American Restaurant	Bar	French Restaurant	Boutique	Library	0.133333	0.000000
3	Chelsea	Score_Culture	Art Gallery	Coffee Shop	Italian Restaurant	Ice Cream Shop	Park	Market	Juice Bar	Hotel	Theater	0.224490	0.010204
4	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	American Restaurant	Salon / Barbershop	Score_Culture	Optical Shop	Spa	Coffee Shop	Malay Restaurant	0.030000	0.000000

- The culture and food affinity scores were calculated for each neighborhood based on frequency of occurrence of relevant venue categories.
- A separate view was also created to look at top 10 venue categories for each neighborhood to analyze clusters later

Clustering Results



name: Chelsea, cluster: 1 Culture Score: 1.22, Food Score: 0.01

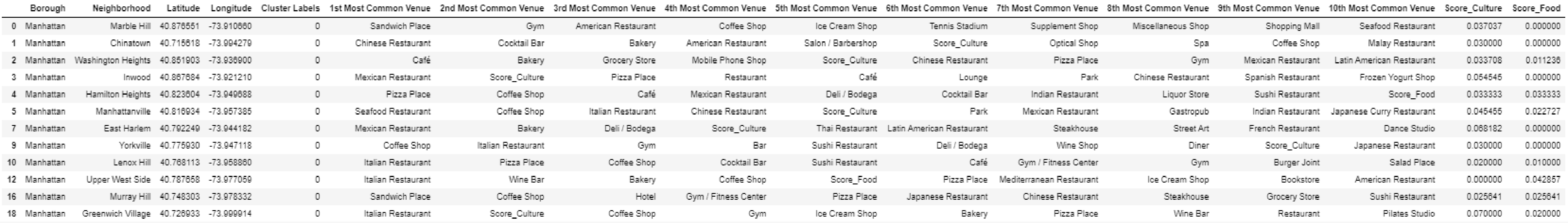
	Cluster No	Avg. Culture Score	Avg. Food Score
🔴	1	0.04	0.01
🟣	2	0.23	0.0
🟢	3	0.12	0.01

- The purple cluster has the highest cultural score but is very limited in food options that my parents might frequent

- The major differentiation between the red and green clusters is the higher cultural score of the green cluster. Both clusters have relevant food options

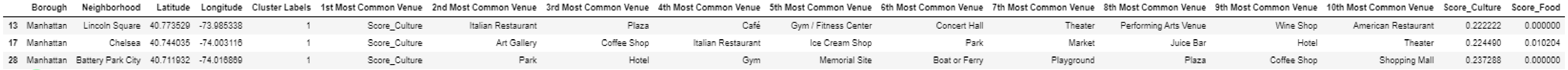
Cluster Deep Dives

● Cluster 1



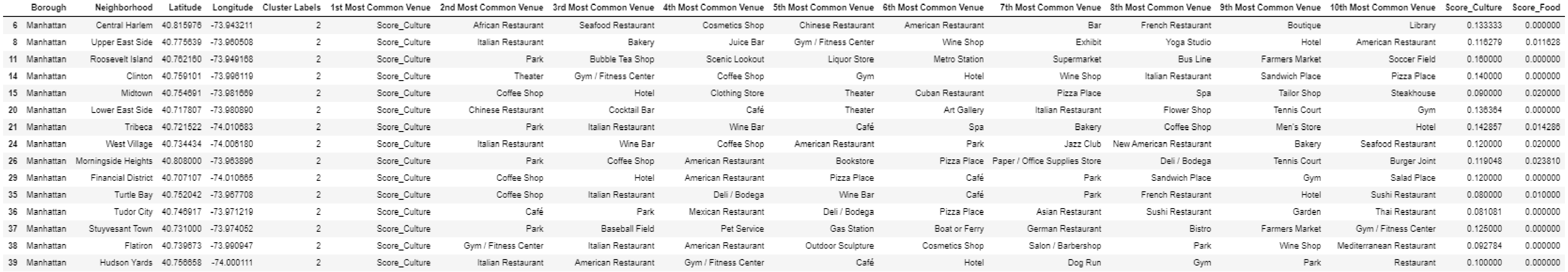
	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Score_Culture	Score_Food
0	Manhattan	Marble Hill	40.876551	-73.910660	0	Sandwich Place	Gym	American Restaurant	Coffee Shop	Ice Cream Shop	Tennis Stadium	Supplement Shop	Miscellaneous Shop	Shopping Mall	Seafood Restaurant	0.037037	0.000000
1	Manhattan	Chinatown	40.715618	-73.994279	0	Chinese Restaurant	Cocktail Bar	Bakery	American Restaurant	Salon / Barbershop	Score_Culture	Optical Shop	Spa	Coffee Shop	Malay Restaurant	0.030000	0.000000
2	Manhattan	Washington Heights	40.851903	-73.936900	0	Café	Bakery	Grocery Store	Mobile Phone Shop	Score_Culture	Chinese Restaurant	Pizza Place	Gym	Mexican Restaurant	Latin American Restaurant	0.033708	0.011236
3	Manhattan	Inwood	40.867684	-73.921210	0	Mexican Restaurant	Score_Culture	Pizza Place	Restaurant	Café	Lounge	Park	Chinese Restaurant	Spanish Restaurant	Frozen Yogurt Shop	0.054545	0.000000
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0	Pizza Place	Coffee Shop	Café	Mexican Restaurant	Deli / Bodega	Cocktail Bar	Indian Restaurant	Liquor Store	Sushi Restaurant	Score_Food	0.033333	0.033333
5	Manhattan	Manhattanville	40.816934	-73.957385	0	Seafood Restaurant	Coffee Shop	Italian Restaurant	Chinese Restaurant	Score_Culture	Park	Mexican Restaurant	Gastropub	Indian Restaurant	Japanese Curry Restaurant	0.045455	0.022727
7	Manhattan	East Harlem	40.792249	-73.944182	0	Mexican Restaurant	Bakery	Deli / Bodega	Score_Culture	Thai Restaurant	Latin American Restaurant	Steakhouse	Street Art	French Restaurant	Dance Studio	0.068182	0.000000
9	Manhattan	Yorkville	40.775930	-73.947118	0	Coffee Shop	Italian Restaurant	Gym	Bar	Sushi Restaurant	Deli / Bodega	Wine Shop	Diner	Score_Culture	Japanese Restaurant	0.030000	0.000000
10	Manhattan	Lenox Hill	40.768113	-73.958860	0	Italian Restaurant	Pizza Place	Coffee Shop	Cocktail Bar	Sushi Restaurant	Café	Gym / Fitness Center	Gym	Burger Joint	Salad Place	0.020000	0.010000
12	Manhattan	Upper West Side	40.787658	-73.977059	0	Italian Restaurant	Wine Bar	Bakery	Coffee Shop	Score_Food	Pizza Place	Mediterranean Restaurant	Ice Cream Shop	Bookstore	American Restaurant	0.000000	0.042857
16	Manhattan	Murray Hill	40.748303	-73.978332	0	Sandwich Place	Coffee Shop	Hotel	Gym / Fitness Center	Pizza Place	Japanese Restaurant	Chinese Restaurant	Steakhouse	Grocery Store	Sushi Restaurant	0.025641	0.025641
18	Manhattan	Greenwich Village	40.726933	-73.999914	0	Italian Restaurant	Score_Culture	Coffee Shop	Gym	Ice Cream Shop	Bakery	Pizza Place	Wine Bar	Restaurant	Pilates Studio	0.070000	0.020000

● Cluster 2



	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	Score_Culture	Score_Food
13	Manhattan	Lincoln Square	40.773529	-73.985338	1	Score_Culture	Italian Restaurant	Plaza	Café	Gym / Fitness Center	Concert Hall	Theater	Performing Arts Venue	Wine Shop	American Restaurant	0.222222	0.000000
17	Manhattan	Chelsea	40.744035	-74.003116	1	Score_Culture	Art Gallery	Coffee Shop	Italian Restaurant	Ice Cream Shop	Park	Market	Juice Bar	Hotel	Theater	0.224490	0.010204
28	Manhattan	Battery Park City	40.711932	-74.016869	1	Score_Culture	Park	Hotel	Gym	Memorial Site	Boat or Ferry	Playground	Plaza	Coffee Shop	Shopping Mall	0.237288	0.000000

● Cluster 3



	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	Score_Culture	Score_Food
6	Manhattan	Central Harlem	40.815976	-73.943211	2	Score_Culture	African Restaurant	Seafood Restaurant	Cosmetics Shop	Chinese Restaurant	American Restaurant	Bar	French Restaurant	Boutique	Library	0.133333	0.000000
8	Manhattan	Upper East Side	40.775639	-73.960508	2	Score_Culture	Italian Restaurant	Bakery	Juice Bar	Gym / Fitness Center	Wine Shop	Exhibit	Yoga Studio	Hotel	American Restaurant	0.116279	0.011628
11	Manhattan	Roosevelt Island	40.762160	-73.949168	2	Score_Culture	Park	Bubble Tea Shop	Scenic Lookout	Liquor Store	Metro Station	Supermarket	Bus Line	Farmers Market	Soccer Field	0.160000	0.000000
14	Manhattan	Clinton	40.759101	-73.996119	2	Score_Culture	Theater	Gym / Fitness Center	Coffee Shop	Gym	Hotel	Wine Shop	Italian Restaurant	Sandwich Place	Pizza Place	0.140000	0.000000
15	Manhattan	Midtown	40.754691	-73.981669	2	Score_Culture	Coffee Shop	Hotel	Clothing Store	Theater	Cuban Restaurant	Pizza Place	Spa	Tailor Shop	Steakhouse	0.090000	0.020000
20	Manhattan	Lower East Side	40.717807	-73.980890	2	Score_Culture	Chinese Restaurant	Cocktail Bar	Café	Theater	Art Gallery	Italian Restaurant	Flower Shop	Tennis Court	Gym	0.136364	0.000000
21	Manhattan	Tribeca	40.721522	-74.010683	2	Score_Culture	Park	Italian Restaurant	Wine Bar	Café	Spa	Bakery	Coffee Shop	Men's Store	Hotel	0.142857	0.014286
24	Manhattan	West Village	40.734434	-74.006180	2	Score_Culture	Italian Restaurant	Wine Bar	Coffee Shop	American Restaurant	Park	Jazz Club	New American Restaurant	Bakery	Seafood Restaurant	0.120000	0.020000
26	Manhattan	Morningside Heights	40.808000	-73.963896	2	Score_Culture	Park	Coffee Shop	American Restaurant	Bookstore	Pizza Place	Paper / Office Supplies Store	Deli / Bodega	Tennis Court	Burger Joint	0.119048	0.023810
29	Manhattan	Financial District	40.707107	-74.010665	2	Score_Culture	Coffee Shop	Hotel	American Restaurant	Pizza Place	Café	Park	Sandwich Place	Gym	Salad Place	0.120000	0.000000
35	Manhattan	Turtle Bay	40.752042	-73.967708	2	Score_Culture	Coffee Shop	Italian Restaurant	Deli / Bodega	Wine Bar	Café	Park	French Restaurant	Hotel	Sushi Restaurant	0.080000	0.010000
36	Manhattan	Tudor City	40.746917	-73.971219	2	Score_Culture	Café	Park	Mexican Restaurant	Deli / Bodega	Pizza Place	Asian Restaurant	Sushi Restaurant	Garden	Thai Restaurant	0.081081	0.000000
37	Manhattan	Stuyvesant Town	40.731000	-73.974052	2	Score_Culture	Park	Baseball Field	Pet Service	Gas Station	Boat or Ferry	German Restaurant	Bistro	Farmers Market	Gym / Fitness Center	0.125000	0.000000
38	Manhattan	Flatiron	40.739673	-73.990947	2	Score_Culture	Gym / Fitness Center	Italian Restaurant	American Restaurant	Outdoor Sculpture	Cosmetics Shop	Salon / Barbershop	Park	Wine Shop	Mediterranean Restaurant	0.092784	0.000000
39	Manhattan	Hudson Yards	40.756658	-74.000111	2	Score_Culture	Italian Restaurant	American Restaurant	Gym / Fitness Center	Café	Hotel	Dog Run	Gym	Park	Restaurant	0.100000	0.000000

Discussion

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Summary of findings on decision criteria

Hotel Budget

The prices for hotels in Manhattan seem to be at a low-point as the budget we set falls at the higher end of the distribution of prices for 3 star and up hotels in manhattan for the dates selected. This is likely due to the reduced traffic due to the Covid-19 pandemic.

Neighborhood Selection

From our clustering analysis, we see that the green cluster offers the right mix of venue categories to cater to the cultural and food requirements. Using the deep dive of clusters, we can select neighborhoods like Midtown which score high on both metrics and look for a suitable hotel.

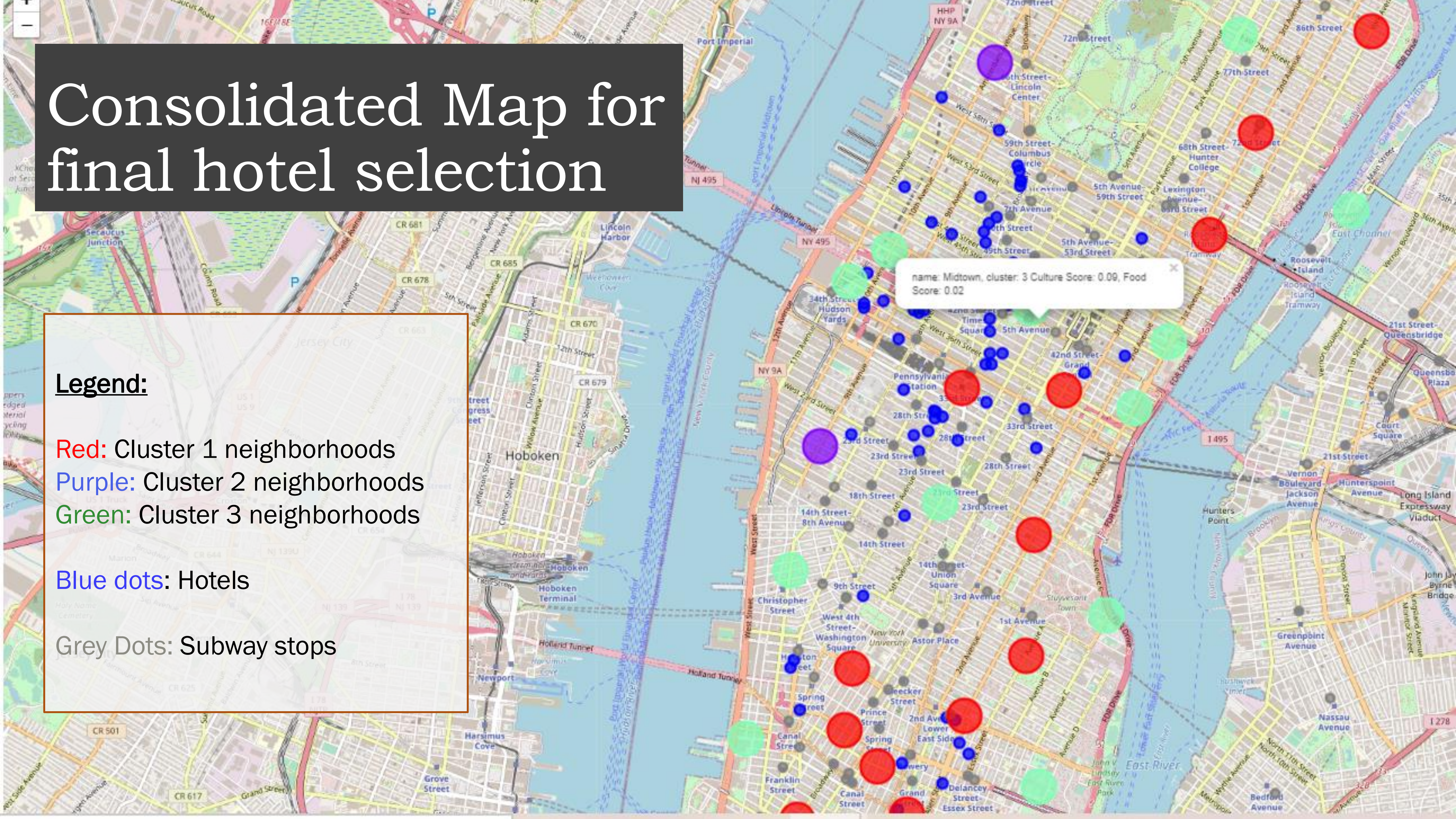
Final hotel selection

To select the final hotel, we should ensure that it is well connected to transport networks and also look at user reviews to ensure we made a good choice

Problem Resolution

DATA SOURCES, PRE-PROCESSING AND PROCESSING

Consolidated Map for final hotel selection



name: Midtown, cluster: 3 Culture Score: 0.09, Food Score: 0.02

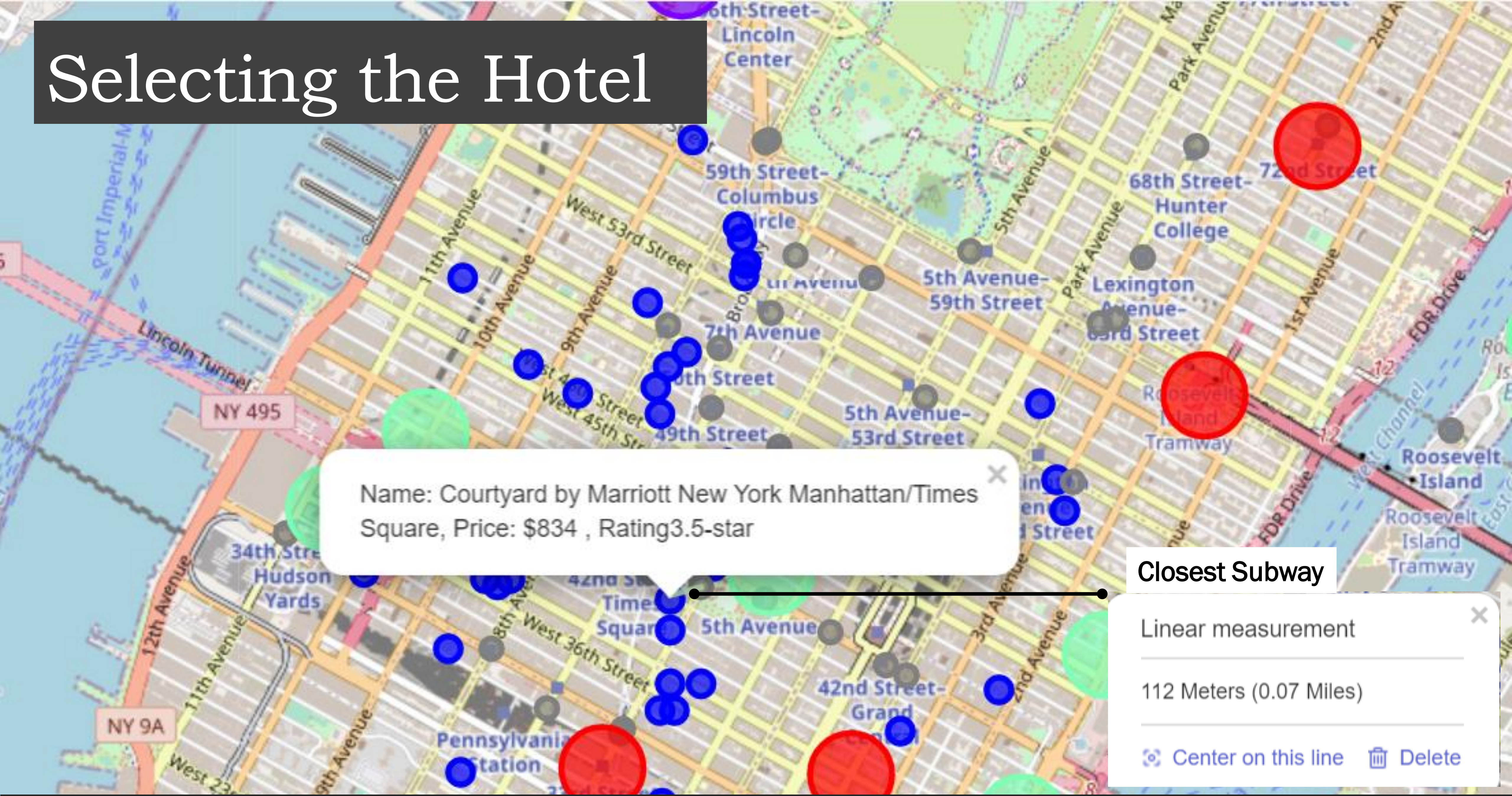
Legend:

Red: Cluster 1 neighborhoods
Purple: Cluster 2 neighborhoods
Green: Cluster 3 neighborhoods

Blue dots: Hotels

Grey Dots: Subway stops

Selecting the Hotel



Name: Courtyard by Marriott New York Manhattan/Times Square, Price: $834 , Rating3.5-star

Closest Subway

Linear measurement

112 Meters (0.07 Miles)

Center on this line Delete

Checking out the hotel and reviews

"The location and value for the money was great. It was safe, clean, and right in the heart of the action. The staff was all very friendly and helpful. I would consider staying here again in the future."

A **Andrew**
United States of America

"Great front desk staff! We could see the empire state building from our window"

M **Matthew**
United States of America

"Location.
Location.
Good standard.
Park just right outside / opposite of the street"

S **Stefan**
United States of America