Data Science Capstone Project

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

Sarim Hassan
5-11-2020

Table of Contents

1 Introduction

1.1 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.

1.2 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

2 Data

2.1 Data Requirements and Sources:

There are numerous types of data that would be required for this project:

- List of neighborhoods and their geographical coordinates
    - Potential Sources: New York University (Catalog of NYC Neighborhoods)
- Location data of Subway stations in New York City
    - Potential Sources: MTA, Wikipedia etc.
- Information on venues in Neighborhoods and category information
    - Potential Sources: Foursquare API
- Hotel information including price, amenities, and location information
    - Potential Source: Hotel Aggregating sites like Trivago, Google etc.

2.2 Data Pre-Processing:

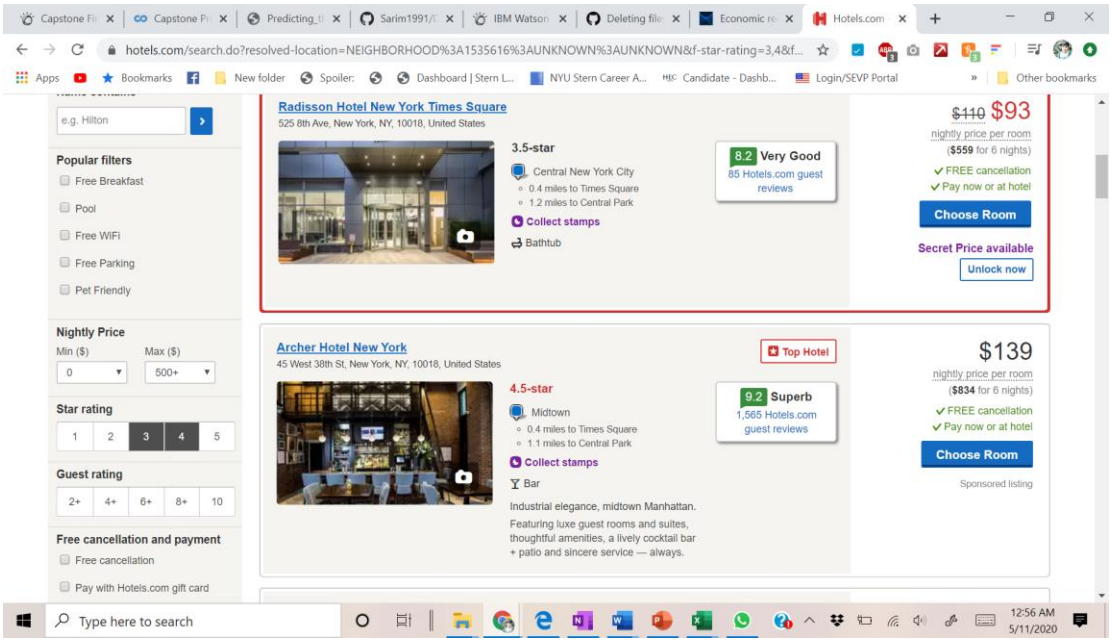
- List of neighborhoods was obtained from the NYU catalog and the same file used in the course lab with the requisite latitude and longitude information was used
- The list of subway stations was obtained from developer data from MTA
- The hotels data was generated by going to Hotels.com, entering the 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 due to difficulty installing the local packages. 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

3 Methodology

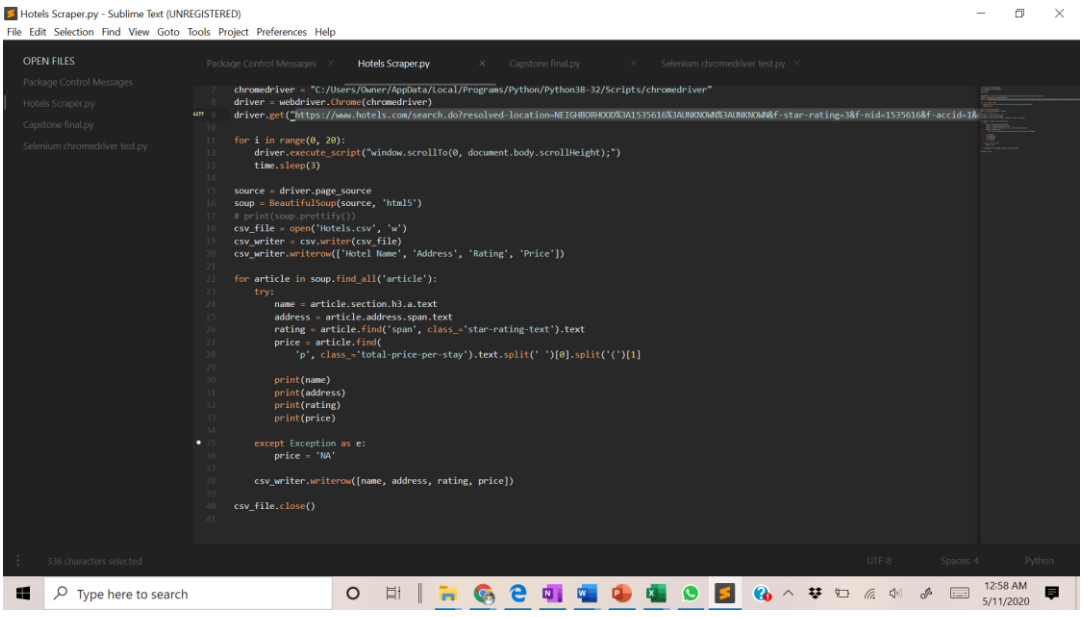
3.1 Data Visualization - Hotels

The hotels data will be 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

Beautiful soup and nominatim were used to scrape the websites for hotels that met the requisite criteria. The filters pre-applied were to get only 3 and 4 star hotels for the date range may 21 -may 27.



Hotels.com uses an infinite page scroll architecture. To scrape data from this source, a headless browser and selenium is required. I faced difficulty running this code in the cloud-based Watson studio environment. Therefore, I installed sublime text on my local machine and ran the code on a local client.



3.2 Data Visualization - Subways

The subway data will be 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

3.3 Manhattan Neighborhood Cluster Analysis

The 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.

The manifestation in code:

To differentiate neighborhoods based on these criteria, the following process was adopted.

- From the location data of venues obtained for various neighborhoods from the Foursquare API, venue categories with high cultural significance were identified. The relative proportion of venue categories were determined using a one-hot encoding and analyzing the mean frequencies of the venue categories relative to the dataset. The cultural square was then calculated as the sum of these normalized mean frequencies only for the subset of venue categories that were determined to be culturally significant

  The following venue categories were assigned to the cultural categories dictionary

  cultural_*categories* *= ['Art Gallery','Art Museum','Auditorium', 'Basketball Stadium', 'Concert Hall', 'Event Space', 'Exhibit', 'Historic Site', 'History Museum', ' Indie Movie Theater', 'Indie Theater', 'Jazz Club', 'Library', 'Memorial Site', 'Monument / Landmark', 'Museum', 'Opera House', 'Outdoor Sculpture', 'Park', 'Performing Arts Venue', 'Public Art', 'Plaza', 'Tennis Stadium', 'Theater', 'Tourist Information Center']*

- From the location data of venues obtained for various neighborhoods from the Foursquare API, venue categories with aligned food options were identified (Eg. Indian Restaurants, North Indian Restaurants etc). The relative proportion of restaurants were determined using a one-hot encoding and analyzing the mean frequencies of the venue categories relative to the dataset. The cultural square was then calculated as the sum of these normalized mean frequencies only for the subset of venue categories that were determined to be culturally significant

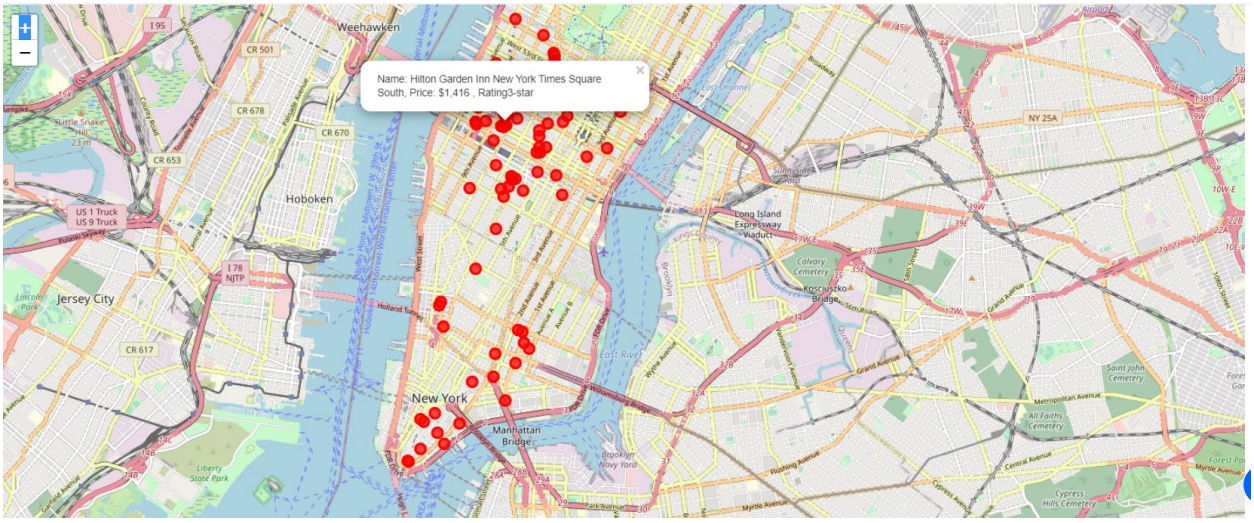
  The following venue categories were assigned to the preferred food categories list

  preferred_food_categories *= ['Indian Restaurant', 'North Indian Restaurant', 'Middle Eastern Restaurant']*

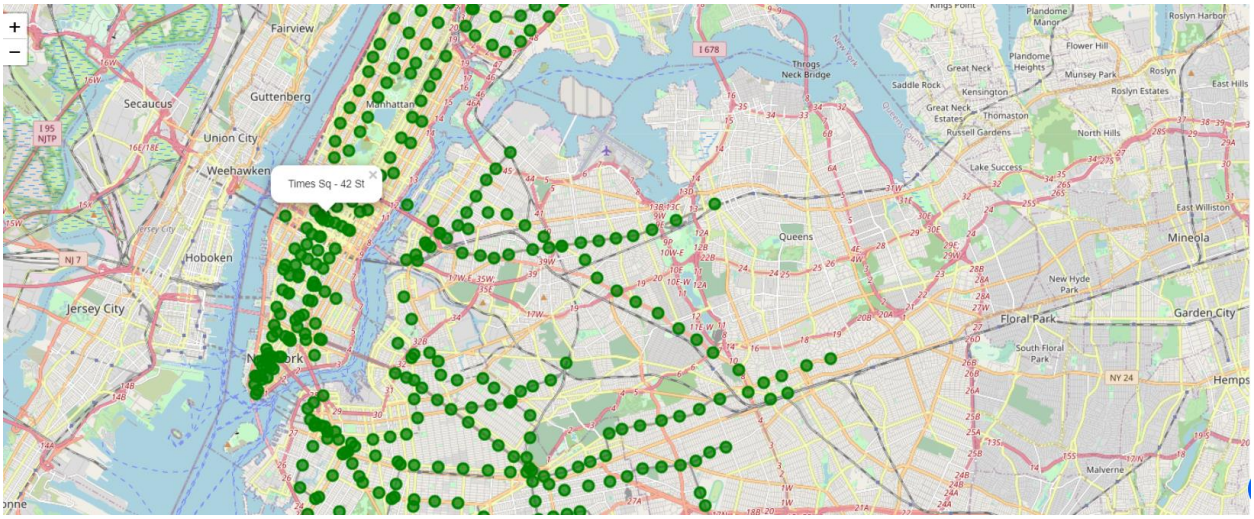
4 Results

4.1 Data Visualization results

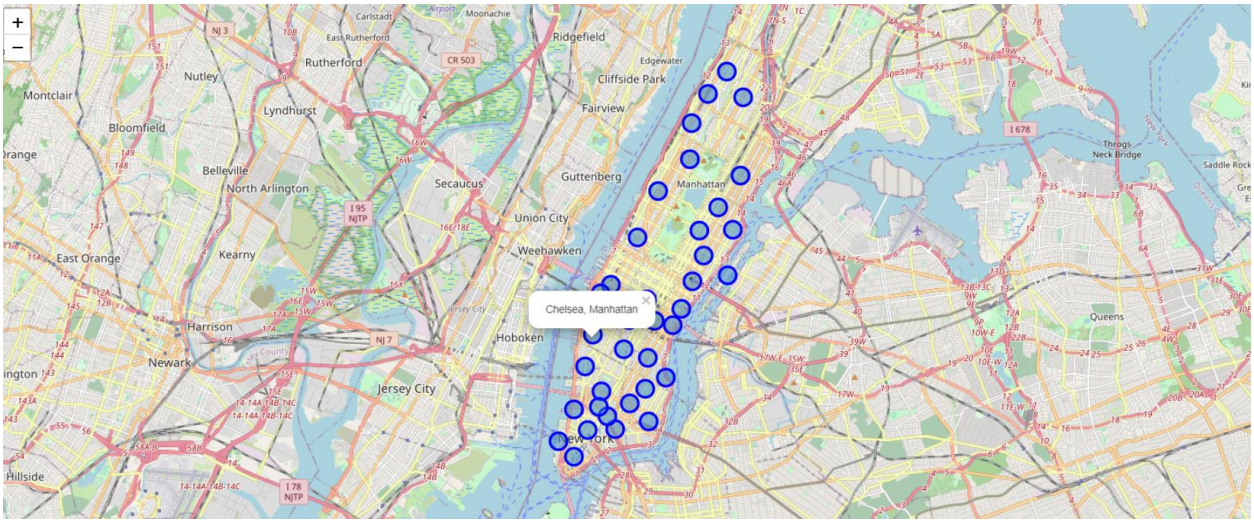
Folium map of hotels visualized with price and star rating information



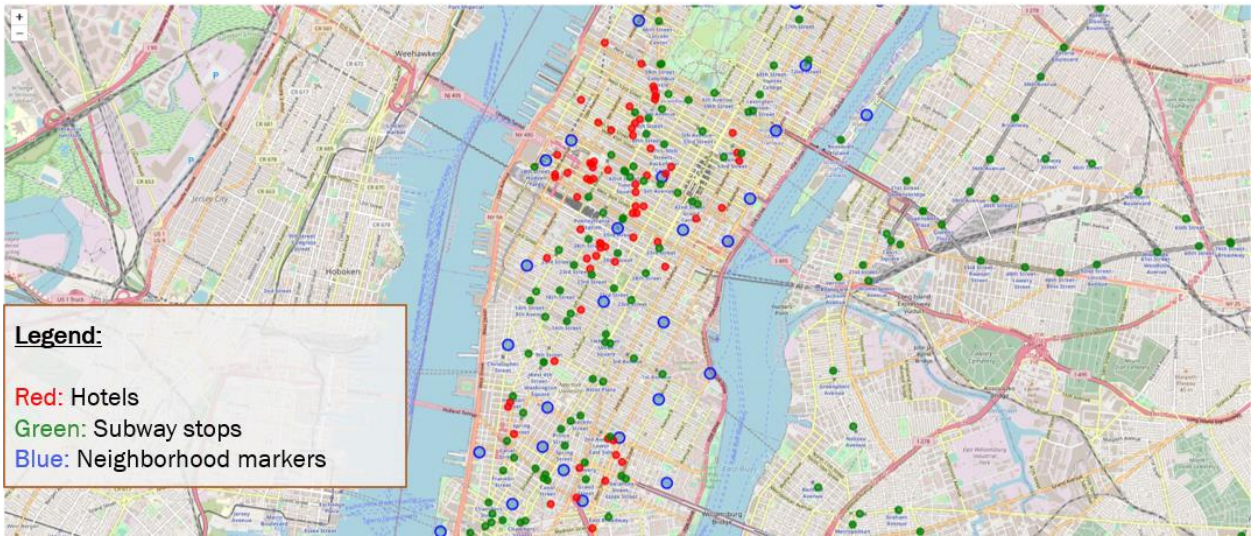
Folium map of subways visualized with stop names



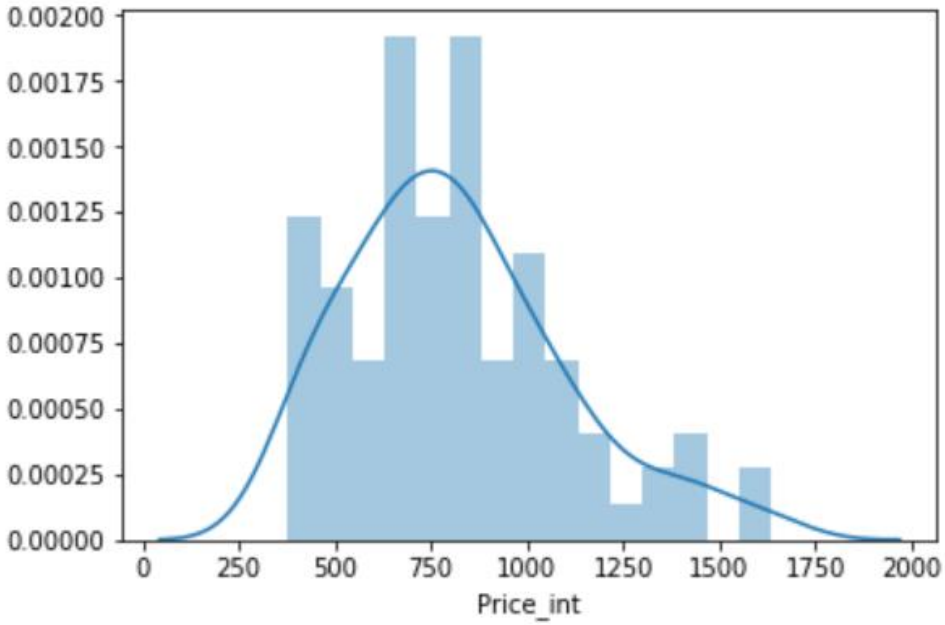
Folium map of neighborhoods visualized

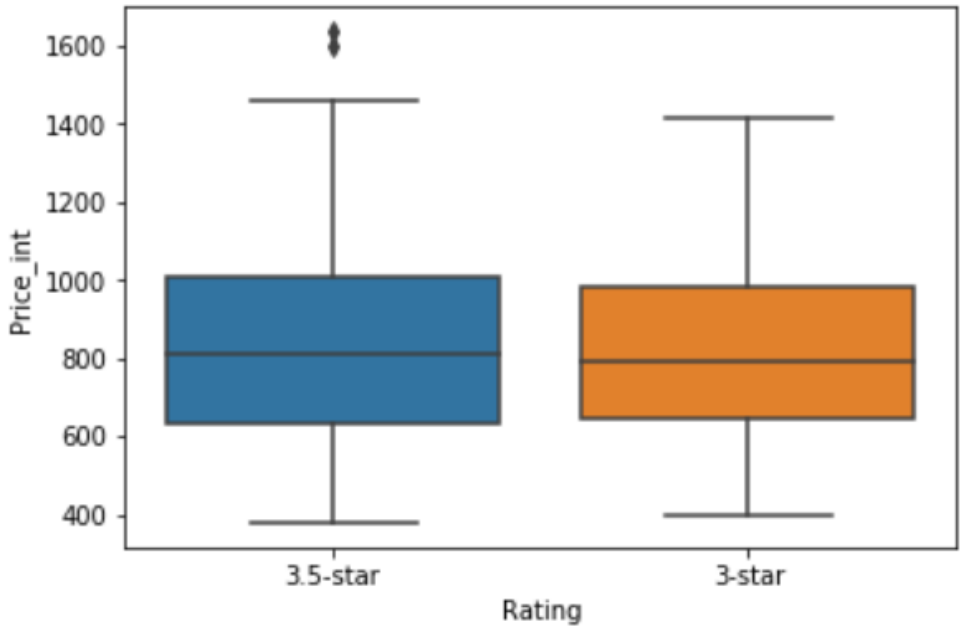


Consolidated map showing the hotels, neighborhoods and subway stops before clustering



4.2 Exploratory Data Analysis Results

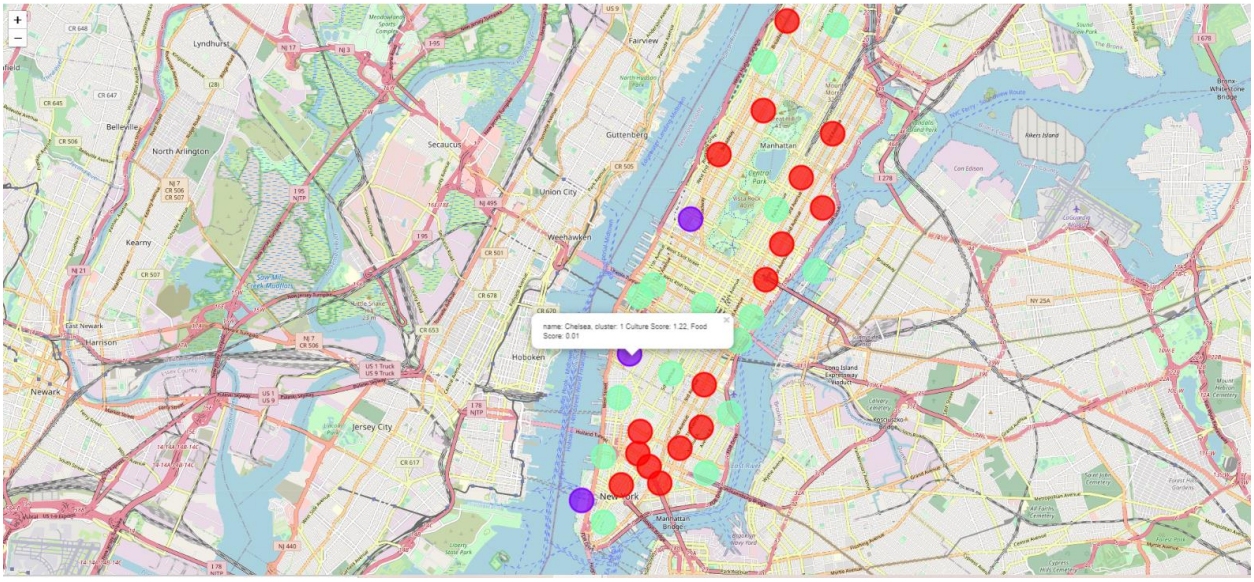




- Our budget of USD 1000 falls in the higher range of the histogram.
- The average price for 3 and 3.5 star rated hotels in our dataset is about the same

4.3 Clustering Analysis Results

Visualization of Clusters and examining the cultural and preferred food category scores



	Cluster No	Avg. Culture Score	Avg. Food Score
0	1	0.04	0.01
1	2	0.23	0.0
2	3	0.12	0.01

Cluster Mapping

Red Cluster

Purple Cluster

Green Cluster

- 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
- It would be prudent as a next step to look for suitable neighborhoods tat belong to the green cluster

Cluster 1 Deep Dive – Red Cluster

	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	40.792240	-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	0.026641	0.026641
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 Deep Dive – Purple Cluster

	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 Deep Dive – Green Cluster

	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	-73.960608	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	-73.980690	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	40.739573	-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

5 Discussion

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.

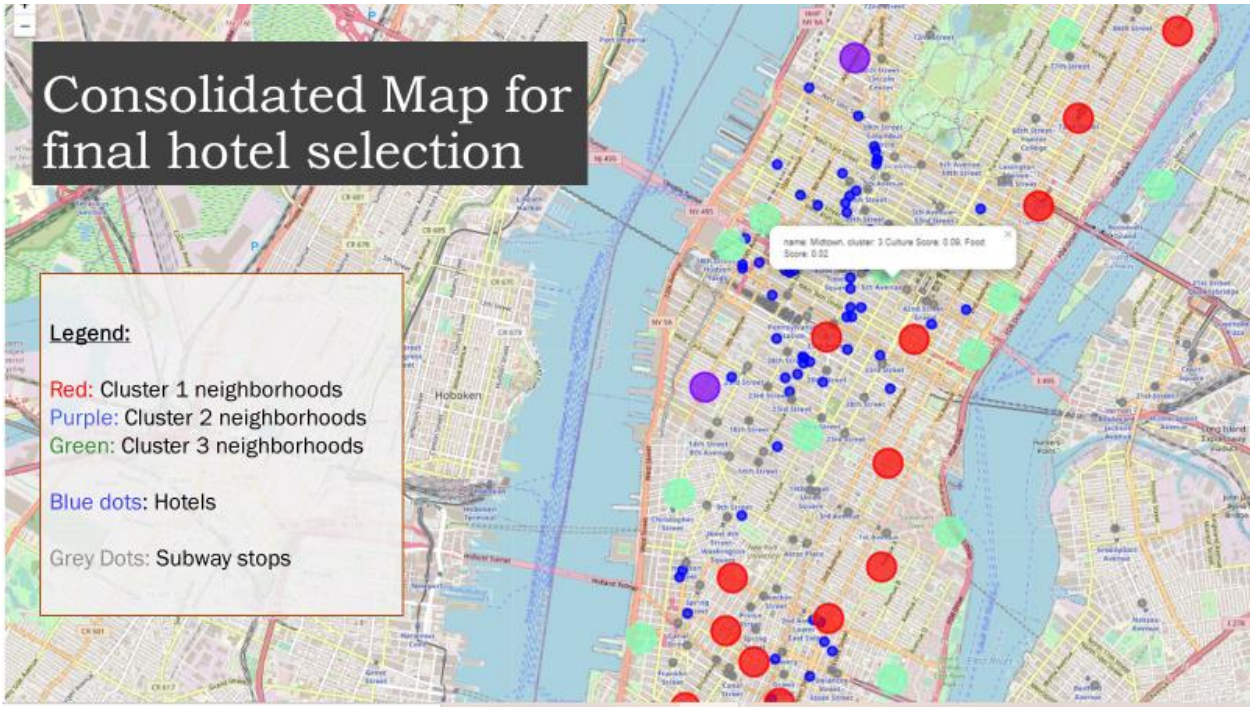
Within the green cluster neighborhoods an analysis of the cultural and preferred food category scores suggests that potential neighborhoods for consideration are:

1. West Village
2. Morning Side Heights
3. Midtown

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.

6 Conclusion



A consolidated map was plotted to help facilitate the final decision-making taking into consideration the proximity to well-connected subway stops.

The courtyard by Marriott was selected as the final choice. It is only 112 meters from a very prominent nearby subway stop.

It also offers good prospects for sight-seeing in close proximity and has a view of the Empire State Building from the rooms.

The hotel also has excellent reviews on Booking.com, Hotels.com and other hotel aggregating websites that praise its location, and the quality of hospitality offered.

I feel comfortable selecting this hotel for my parents' visit and it was a rewarding process to use data science to make my decision