# Exploring the Relationship of Uber Pickups and Venues of New York City

## 1. Introduction

I explored New York City, segmented and clustered its neighborhoods based on their popular venue categories following the instructions of the course lab. The biggest clustered dataset showed that there were many coffee shops, cafes, restaurants of various cuisines, and bars and clubs. It also had health-related venues such as gyms, yoga studios, cycle studios, etc. The data clearly showed that the city is very diverse.

In a big city like New York City, it is very popular to take taxes or use ride sharing services like Uber and Lyft. The Uber Pickups in New York City is available on kaggle.com. I would like to further explore New York City combining the Foursquare's location data with the Uber Pickups data to see if venue categories are related to Uber usage count. This may discover popular pickup locations and this finding may be helpful for the city to better understand how much traffic increase in certain areas by Uber.

## 2. Data Acquisition and Cleaning

### 2.1 Data Sources

The Uber Pickups in New York City dataset can be found on Kaggle.com [link]. This dataset contains 6 months of Uber pick up geo-locations (latitudes and longitudes) of year 2014. Location information along with venue details will be dynamically obtained by utilizing the Foursquare location APIs.

### 2.2 Data Cleaning

The Uber Pickups data is a zipped file (`Uber-dataset.zip`) and contains 6 CSV files in it. And each CSV file contains pick up data for each month from Apr through Sep 2014. The individual file has 4 attributes (columns): Date/Time, Lat, Lon, and Base. Latitude and longitude data are required to query nearest venues with the Foursquare APIs, but the other columns, Date/Time and Base, are not needed for the analysis. So, I dropped these two unnecessary columns from the dataset and combined all the data from the 6 CSV files and created a table as shown in <Figure 1>. In order to read CSV files in sequence with for loop, I renamed the file names with index number at the end: `uber-raw-data-0.csv` `uber-raw-data-1.csv`, …, `uber-raw-data-5`.

| | Lat | Lon |
|---|---|---|
| 0 | 40.7690 | -73.9549 |
| 1 | 40.7267 | -74.0345 |
| 2 | 40.7316 | -73.9873 |
| 3 | 40.7588 | -73.9776 |
| 4 | 40.7594 | -73.9722 |

<Figure 1>

I created a table with all the location data and there were 4,534,327 rows. I then inserted the entire data into the Locations table in the SQLite database as shown in <Figure 2>. It was not possible to call the Foursquare API for all the locations in the table due to the daily limit, I randomly chose 8,000 locations among them. Utilizing the SQLite database, I was able to keep track of locations that had been used to get nearby venues by calling the search API (https://api.foursquare.com/v2/venues/search).

I collected venues within a radius of 75 meters from each coordinate (latitude and longitude) in the Locations table. I selected 75 meters after testing several shorter distances. The API did not return any nearby venues for many pick-up locations for shorter radiuses. 75 meters was big enough to find more than one nearby venue. Since the API also provides a distance between the pick-up location and the venue, I can later use only the first n venues listed in the order of distance. It required too many days to get nearby venues for all those 8,000 randomly chosen locations due to the daily limit of calling the search API. I spent several days and was able to get 20,156 venues for 4,228 locations. I then put all the collected venues' information such as venue_id, name, latitude, longitude, category and a distance from the pick-up location into another table, Venues, in the database as shown in <Figure 3>.

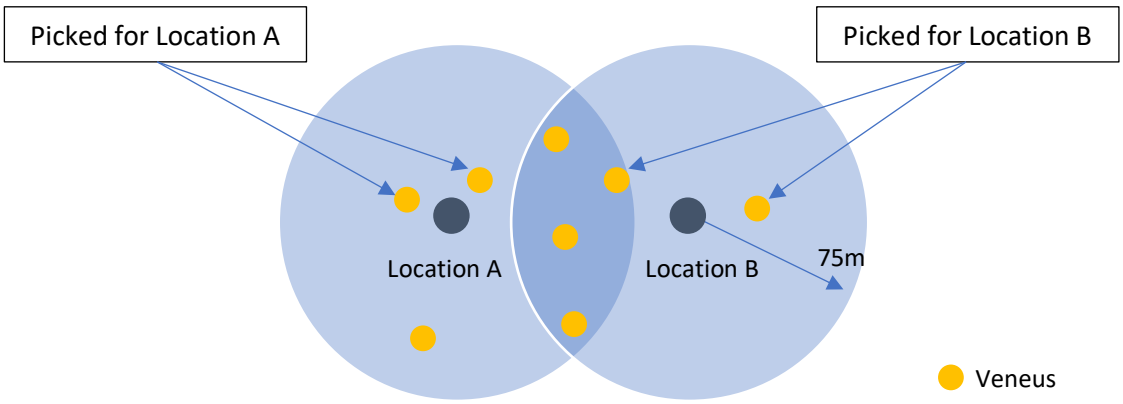
| location_id | latitude | longitude | api_called | picked |
|---|---|---|---|---|
| 40 | 40.7217 | -73.9875 | 1 | 1 |
| 333 | 40.6334 | -73.9954 | 1 | 1 |
| 439 | 40.7211 | -74.0005 | 1 | 1 |
| 773 | 40.7642 | -73.9731 | 1 | 1 |
| 850 | 40.7592 | -73.9777 | 1 | 1 |
| 1087 | 40.7196 | -73.9995 | 1 | 1 |
| 1194 | 40.731 | -73.9796 | 1 | 1 |
| 1364 | 40.7442 | -73.9833 | 1 | 1 |
| 1751 | 40.7606 | -73.9736 | 1 | 1 |

<Figure 2>

| venue_id | name | v_latitude | v_longitude | category | distance | location_id |
|---|---|---|---|---|---|---|
| 50771637e4b0e6ae4f3abd52 | vòDD. New York | 40.7213 | -73.9875 | Boutique | 40 | 40 |
| 4cd5f138aeb1224b77be25ff | Il Laboratorio del Gelato | 40.7222 | -73.987 | Ice Cream Shop | 69 | 40 |
| 4e4963bab3add952d314007e | Grit N Glory | 40.722 | -73.9881 | Boutique | 62 | 40 |
| 4acbe67af964a52044c820e3 | Katz's Delicatessen | 40.7223 | -73.9874 | Sandwich Place | 69 | 40 |
| 49ebc25ff964a52023671fe3 | September Wines & Spirits | 40.7213 | -73.9878 | Wine Shop | 52 | 40 |
| 537b5a29498ec121cf9fa1f4 | Sweet Chick | 40.7218 | -73.9875 | Southern / Soul Food Restaurant | 7 | 40 |
| 536020eb11d2ce653fb711d0 | The Ludlow Hotel | 40.7219 | -73.9872 | Hotel | 30 | 40 |
| 5303f306498ebbca8f7e1158 | Le French Diner | 40.7221 | -73.9881 | French Restaurant | 70 | 40 |
| 56c73a4a498ea736d7388245 | Ludlow Coffee Supply | 40.7217 | -73.9875 | Coffee Shop | 2 | 40 |

<Figure 3>

As the next step, I originally picked only two closely located venues from each pick-up location in order to reduce duplicated venues that may be located in the overlapped areas within radiuses of multiple pick-up locations as shown in <Figure 4>. However, I realized that if I picked only two venues, it means that I can only use two categories as features for k-means clustering algorithm. So, I increased the number to five.




<Figure 4>

This was achieved through a few steps of data processing. First, I got all the venue data from the Venues table in the database and created a Pandas dataframe. As the second step, I sorted the dataset with these two columns, location_id and distance, and got the results as shown in <Figure 5>. Lastly, I grouped the dataset by the location_id column and picked only the top five rows by using head(5) method. The results are shown in <Figure 6>. Note that some locations have less than five nearby venues.

| | location_id | distance | name | category | v_latitude | v_longitude |
|---|---|---|---|---|---|---|
| **0** | 40 | 2 | Ludlow Coffee Supply | Coffee Shop | 40.7217 | -73.9875 |
| **1** | 40 | 7 | Sweet Chick | Southern / Soul Food Restaurant | 40.7218 | -73.9875 |
| **2** | 40 | 10 | Tre Restaurant & Wine Bar | Italian Restaurant | 40.7218 | -73.9875 |
| **3** | 40 | 20 | Van Leeuwen Artisan Ice Cream | Ice Cream Shop | 40.7215 | -73.9874 |
| **4** | 40 | 20 | Assembly New York | Clothing Store | 40.7215 | -73.9876 |
| **...** | ... | ... | ... | ... | ... | ... |
| **6593** | 795683 | 61 | The Algonquin Hotel, Autograph Collection | Hotel | 40.7560 | -73.9823 |
| **6594** | 797814 | 37 | Michael's, The Consignment Shop for Women | Women's Store | 40.7796 | -73.9593 |
| **6595** | 797814 | 49 | Le Pain Quotidien | Café | 40.7797 | -73.9594 |
| **6596** | 797814 | 64 | James Perse | Boutique | 40.7795 | -73.9600 |
| **6597** | 797814 | 65 | Starbucks Reserve | Coffee Shop | 40.7798 | -73.9596 |

<Figure 5>

| | location_id | distance | name | category | v_latitude | v_longitude |
|---|---|---|---|---|---|---|
| **0** | 40 | 2 | Ludlow Coffee Supply | Coffee Shop | 40.7217 | -73.9875 |
| **1** | 40 | 7 | Sweet Chick | Southern / Soul Food Restaurant | 40.7218 | -73.9875 |
| **2** | 40 | 10 | Tre Restaurant & Wine Bar | Italian Restaurant | 40.7218 | -73.9875 |
| **3** | 40 | 20 | Van Leeuwen Artisan Ice Cream | Ice Cream Shop | 40.7215 | -73.9874 |
| **4** | 40 | 20 | Assembly New York | Clothing Store | 40.7215 | -73.9876 |
| **...** | ... | ... | ... | ... | ... | ... |
| **15671** | 1886293 | 33 | Carroll Place | Italian Restaurant | 40.7285 | -73.9998 |
| **15672** | 1886293 | 36 | Le Poisson Rouge | Rock Club | 40.7285 | -74.0000 |
| **15673** | 1886293 | 41 | The Malt House | Bar | 40.7286 | -73.9994 |
| **15679** | 1886632 | 18 | Calliope | Furniture / Home Store | 40.7377 | -74.0073 |
| **15680** | 1886632 | 62 | Maison Margiela | Boutique | 40.7379 | -74.0065 |

<Figure 6>

2.3 Feature Selection

Analyzing the venue data in the Venues table in the database, I found that there are 411 unique venue categories. I had already created a new dataset with only top 5 nearby venues for 4,228 locations among 8,000 randomly picked ones. Some of the locations have less than 5 nearby venues, but most of locations have more than 5 venues. Therefore, using the first top 5 venue categories for each location will be enough to run the k-means clustering algorithm to cluster them into 5 different clusters. As a preparation, I created a new dataset with columns for all the venue categories. And I take means of the frequency of occurrences of each category for every single location. Finally, I picked the first top 5 venue categories for each location as shown in <Figure 7>. This is the final dataset that will be used for clustering.

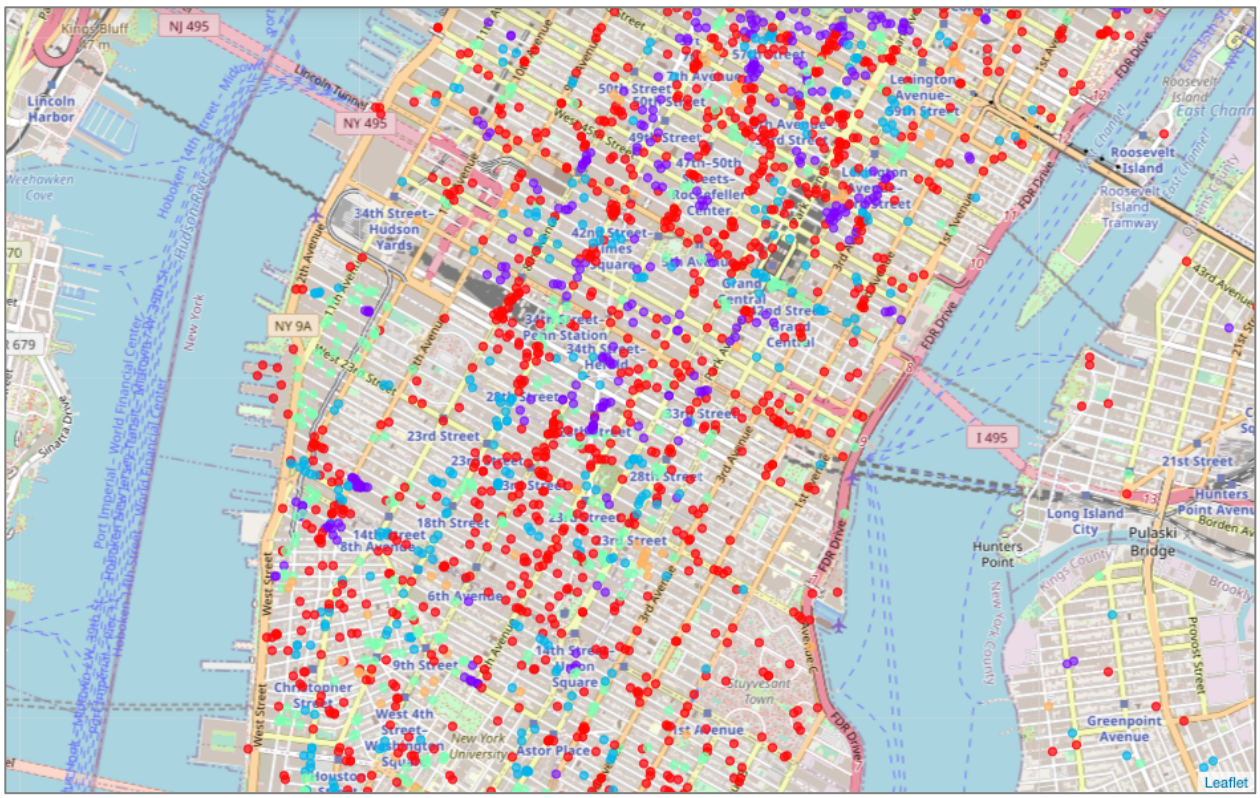
| | location_id | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| **0** | 40 | Clothing Store | Ice Cream Shop | Coffee Shop | Italian Restaurant | Southern / Soul Food Restaurant |
| **1** | 439 | Clothing Store | Vegetarian / Vegan Restaurant | Pop-Up Shop | Miscellaneous Shop | Spa |
| **2** | 773 | Jewelry Store | Plaza | Italian Restaurant | Tea Room | Hotel |
| **3** | 850 | Greek Restaurant | French Restaurant | Toy / Game Store | Outdoor Sculpture | Gym |
| **4** | 1087 | French Restaurant | History Museum | Speakeasy | Hotel | Arts & Crafts Store |
| **...** | ... | ... | ... | ... | ... | ... |
| **3746** | 2426806 | Airport Service | Donut Shop | Food | Falafel Restaurant | Farm |
| **3747** | 2428168 | Ice Cream Shop | Burger Joint | Coffee Shop | Scenic Lookout | Zoo Exhibit |
| **3748** | 2429281 | Café | Harbor / Marina | Flower Shop | Falafel Restaurant | Farm |
| **3749** | 2430145 | Arts & Crafts Store | Men's Store | Women's Store | Pet Store | Boutique |
| **3750** | 2430861 | Italian Restaurant | Bar | Flower Shop | Eye Doctor | Falafel Restaurant |

<Figure 7>

## 3. Data Analysis

3.1 k-clusters = 5

I ran the k-means clustering algorithm with the prepared dataset with k_clusters set to 5 and plotted all the clustered locations on the map as shown in <Figure 8>. As you can see, the clustered points were scattered all around Manhattan. They were not clustered with any shapes that were separated each other. Therefore, it was necessary to carefully examine the data for each cluster.




<Figure 8>

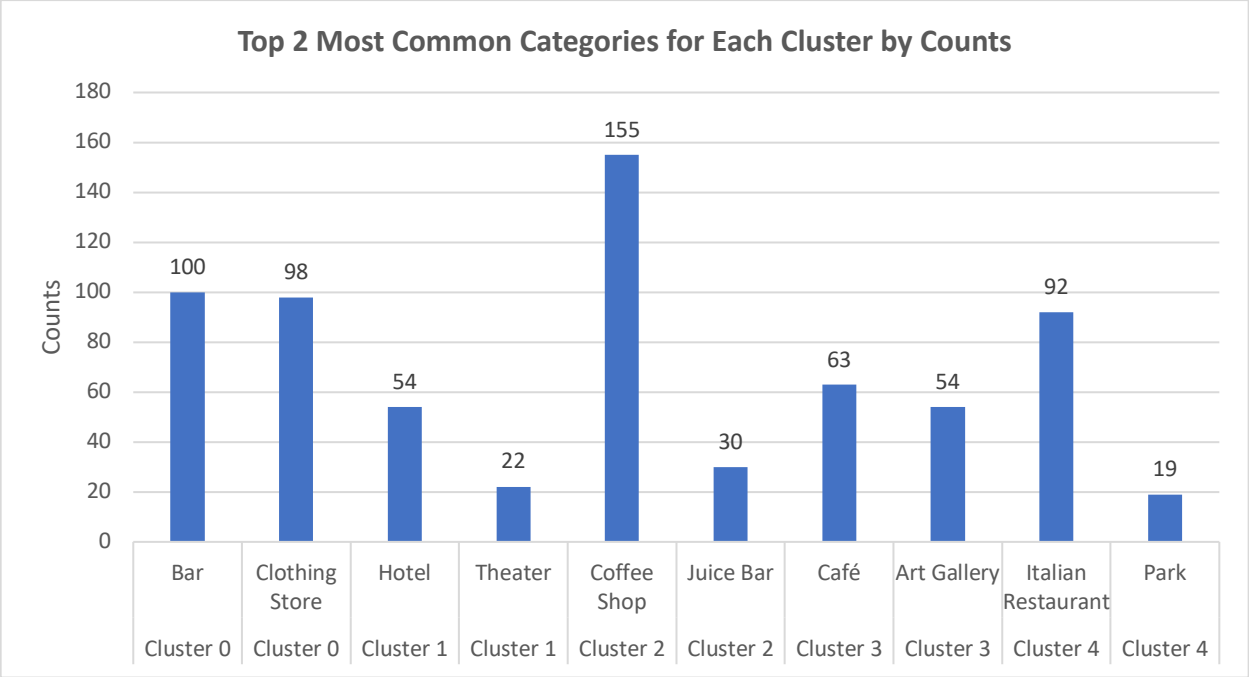
I count the frequency of all the categories for the 1st Most Common Venue for each cluster and created tables that show the top 20 popular venue categories for each cluster. With the sorted and counted categories in these tables, I could easily find what venue categories that each cluster represents. It was easy to anticipate seeing Bar, Hotel, Coffee shop, and Café. However, it was very interesting to find that Clothing Store and Juice Bar were among the top categories. Theater, Art Gallery and Park were listed as popular categories as well. The most interesting finding was that Italian Restaurant was the top in the Cluster 4 even though there were so many different types of restaurants around. Although I had expected Airport and Hotel to be ranked top, I was not able to see them in top 2 for all the clusters.

| NO | Cluster 0 | Counts | Cluster 1 | Counts |
|---|---|---|---|---|
| 1 | Bar | 100 | Hotel | 54 |
| 2 | Clothing Store | 98 | Theater | 22 |
| 3 | Deli / Bodega | 79 | Cocktail Bar | 20 |
| 4 | Sandwich Place | 63 | Coffee Shop | 17 |
| 5 | Juice Bar | 58 | Bar | 15 |
| 6 | Pizza Place | 53 | Mexican Restaurant | 15 |
| 7 | Airport Service | 52 | Steakhouse | 14 |
| 8 | Steakhouse | 51 | Juice Bar | 12 |
| 9 | Bakery | 38 | Jewelry Store | 11 |
| 10 | Yoga Studio | 36 | Deli / Bodega | 11 |
| 11 | Japanese Restaurant | 34 | Dessert Shop | 10 |
| 12 | Furniture / Home Store | 31 | Mediterranean Restaurant | 9 |
| 13 | New American Restaurant | 31 | New American Restaurant | 9 |
| 14 | Theater | 29 | French Restaurant | 9 |
| 15 | Cocktail Bar | 29 | Korean Restaurant | 8 |
| 16 | Spa | 29 | Nightclub | 8 |
| 17 | Jewelry Store | 28 | Jazz Club | 8 |
| 18 | Mexican Restaurant | 27 | Café | 7 |
| 19 | Greek Restaurant | 27 | Lingerie Store | 7 |
| 20 | Indian Restaurant | 26 | Sandwich Place | 7 |

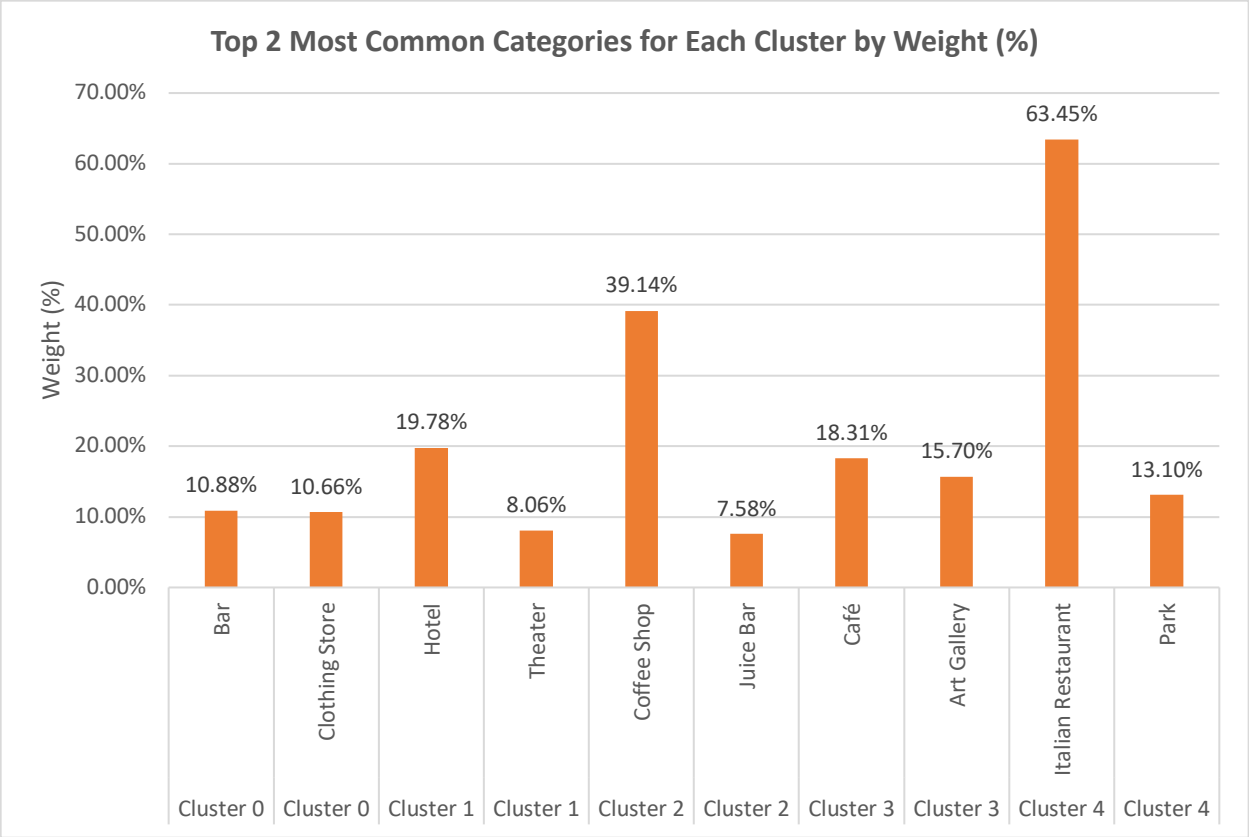
| NO | Cluster 2 | Counts | Cluster 3 | Counts |
|---|---|---|---|---|
| 1 | Coffee Shop | 155 | Café | 63 |
| 2 | Juice Bar | 30 | Art Gallery | 54 |
| 3 | Bar | 21 | American Restaurant | 37 |
| 4 | Cocktail Bar | 20 | Cocktail Bar | 25 |
| 5 | Pizza Place | 17 | New American Restaurant | 18 |
| 6 | Steakhouse | 15 | Bar | 16 |
| 7 | New American Restaurant | 14 | Sandwich Place | 14 |
| 8 | Sandwich Place | 13 | Juice Bar | 14 |
| 9 | Yoga Studio | 12 | Clothing Store | 14 |
| 10 | Theater | 11 | Burger Joint | 10 |
| 11 | Clothing Store | 11 | Steakhouse | 10 |
| 12 | French Restaurant | 10 | Spa | 10 |
| 13 | Furniture / Home Store | 10 | Furniture / Home Store | 10 |
| 14 | Cycle Studio | 9 | Italian Restaurant | 10 |
| 15 | Deli / Bodega | 9 | Indian Restaurant | 7 |
| 16 | Mediterranean Restaurant | 8 | Ice Cream Shop | 7 |
| 17 | Bridal Shop | 8 | Jewelry Store | 7 |
| 18 | Cosmetics Shop | 8 | Deli / Bodega | 6 |
| 19 | Grocery Store | 8 | French Restaurant | 6 |
| 20 | Bakery | 7 | Boutique | 6 |

| NO | Cluster 4 | Counts |
|---|---|---|
| 1 | Italian Restaurant | 92 |
| 2 | Park | 19 |
| 3 | Café | 5 |
| 4 | Steakhouse | 4 |
| 5 | Clothing Store | 4 |
| 6 | Restaurant | 3 |
| 7 | Department Store | 2 |
| 8 | History Museum | 2 |
| 9 | Juice Bar | 2 |
| 10 | Kebab Restaurant | 2 |
| 11 | Mexican Restaurant | 1 |
| 12 | Smoke Shop | 1 |
| 13 | Poke Place | 1 |
| 14 | Pharmacy | 1 |
| 15 | Middle Eastern Restaurant | 1 |
| 16 | Art Gallery | 1 |
| 17 | Bakery | 1 |
| 18 | Gastropub | 1 |
| 19 | French Restaurant | 1 |
| 20 | Dance Studio | 1 |

I collected all the top 2 categories and plotted them by their counts as a bar chart. I also calculated weights of those top categories in their own cluster and plotted them as another bar chart. They are shown in <Figure 9> and <Figure 10>, respectively.



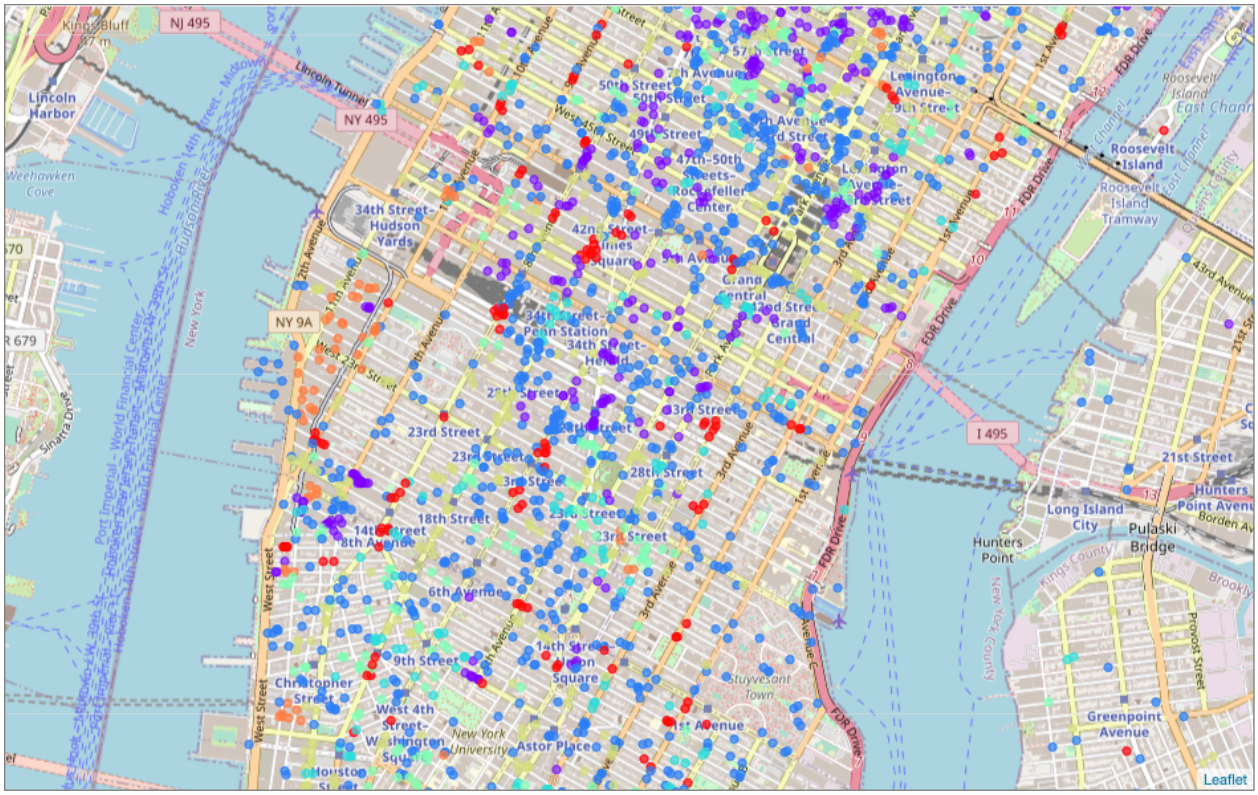

<Figure 9>




<Figure 10>

3.2 k-cluster = 7

Through the first clustering results, I was able to find quite interesting top common categories and was curious to see more clusters. As the second analysis, I increased k_clusters value to 7. After completing clustering, I plotted the clustered locations on the map shown in <Figure 11>.




<Figure 11>

And I count the frequency of the 1st Most Common Venue categories for each cluster and the tables below show the top 20 popular venue categories for 7 clusters.

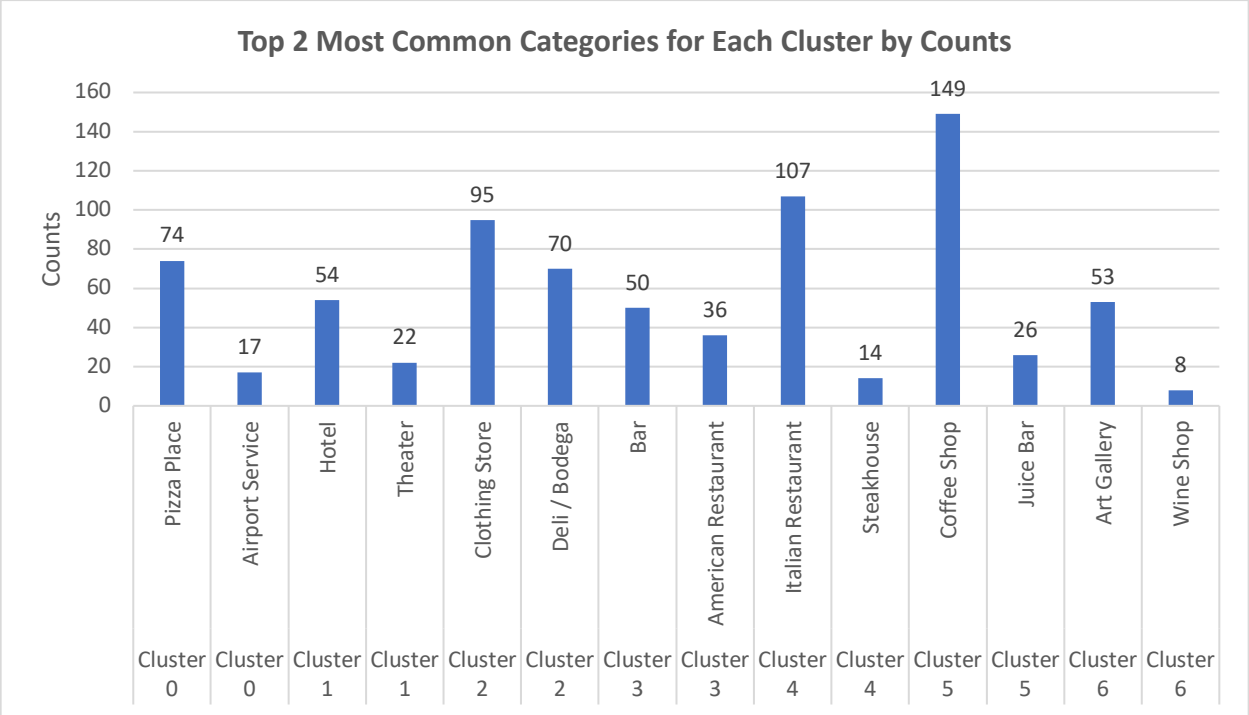
| NO | Cluster 0 | Counts | Cluster 1 | Counts |
|---|---|---|---|---|
| 1 | Pizza Place | 74 | Hotel | 54 |
| 2 | Airport Service | 17 | Theater | 22 |
| 3 | Juice Bar | 14 | Cocktail Bar | 20 |
| 4 | Sandwich Place | 13 | Coffee Shop | 18 |
| 5 | Bar | 12 | Mexican Restaurant | 15 |
| 6 | Ice Cream Shop | 10 | Bar | 14 |
| 7 | Deli / Bodega | 8 | Steakhouse | 14 |
| 8 | Dance Studio | 8 | Juice Bar | 12 |
| 9 | Jewelry Store | 8 | Jewelry Store | 11 |
| 10 | Coffee Shop | 6 | Dessert Shop | 10 |
| 11 | Scenic Lookout | 6 | Deli / Bodega | 10 |
| 12 | Mexican Restaurant | 6 | Mediterranean Restaurant | 9 |
| 13 | Airport Lounge | 6 | New American Restaurant | 9 |
| 14 | Yoga Studio | 5 | French Restaurant | 9 |
| 15 | Market | 5 | Jazz Club | 8 |
| 16 | Mediterranean Restaurant | 5 | Nightclub | 8 |
| 17 | Japanese Restaurant | 4 | Korean Restaurant | 8 |
| 18 | Indian Restaurant | 4 | Café | 8 |
| 19 | Mobile Phone Shop | 4 | Sandwich Place | 7 |
| 20 | Airport Food Court | 4 | Burger Joint | 7 |

| NO | Cluster 2 | Counts | Cluster 3 | Counts |
|---|---|---|---|---|
| 1 | Clothing Store | 95 | Bar | 50 |
| 2 | Deli / Bodega | 70 | American Restaurant | 36 |
| 3 | Sandwich Place | 56 | Cocktail | 11 |
| 4 | Juice Bar | 50 | New American Restaurant | 10 |
| 5 | Café | 46 | Steakhouse | 9 |
| 6 | Bar | 45 | Burger Joint | 7 |
| 7 | Steakhouse | 40 | Sandwich Place | 6 |
| 8 | Airport Service | 35 | Juice Bar | 6 |
| 9 | Furniture / Home Store | 35 | Salad Place | 5 |
| 10 | Bakery | 35 | Bakery | 4 |
| 11 | Park | 32 | Cosmetics Shop | 4 |
| 12 | Yoga Studio | 29 | Café | 4 |
| 13 | New American Restaurant | 29 | Jazz Club | 4 |
| 14 | Spa | 28 | Clothing Store | 4 |
| 15 | Theater | 26 | Mexican Restaurant | 3 |
| 16 | Food Truck | 25 | Jewelry Store | 3 |
| 17 | Cocktail Bar | 25 | French Restaurant | 3 |
| 18 | Indian Restaurant | 24 | Deli / Bodega | 3 |
| 19 | Greek Restaurant | 24 | Historic Site | 3 |
| 20 | Taxi Stand | 23 | Ice Cream Shop | 2 |

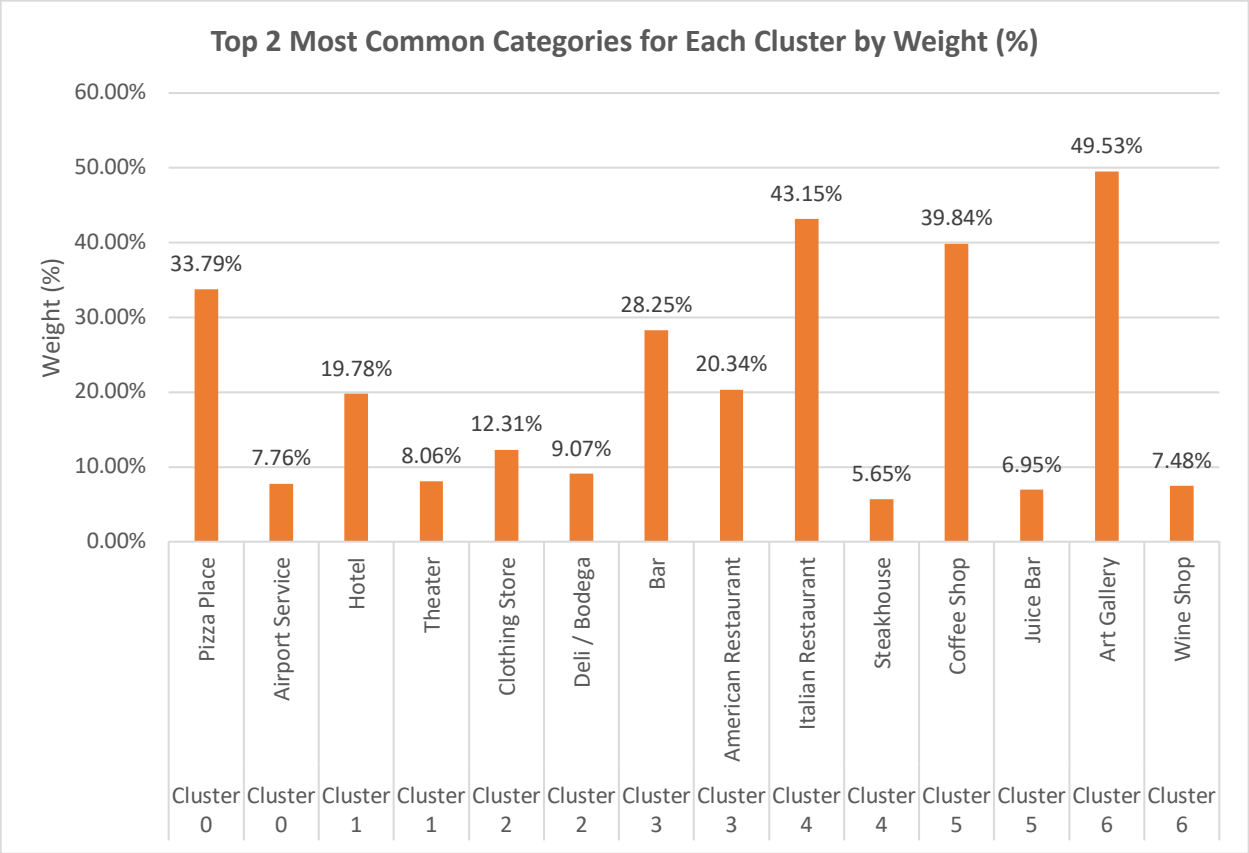
| NO | Cluster 4 | Counts | Cluster 5 | Counts |
|---|---|---|---|---|
| 1 | Italian Restaurant | 107 | Coffee Shop | 149 |
| 2 | Steakhouse | 14 | Juice Bar | 26 |
| 3 | Cocktail Bar | 13 | Cocktail Bar | 20 |
| 4 | French Restaurant | 12 | Bar | 19 |
| 5 | Bar | 10 | New American Restaurant | 18 |
| 6 | Café | 10 | Steakhouse | 15 |
| 7 | Clothing Store | 9 | Sandwich Place | 12 |
| 8 | Japanese Restaurant | 8 | Clothing Store | 11 |
| 9 | Juice Bar | 7 | Theater | 11 |
| 10 | Spa | 7 | Furniture / Home Store | 10 |
| 11 | Cycle Studio | 7 | Yoga Studio | 10 |
| 12 | Burger Joint | 6 | French Restaurant | 10 |
| 13 | Men's Store | 5 | Deli / Bodega | 9 |
| 14 | New American Restaurant | 5 | Cycle Studio | 9 |
| 15 | Yoga Studio | 5 | Bridal Shop | 9 |
| 16 | Vegetarian / Vegan Restaurant | 5 | Cosmetics Shop | 8 |
| 17 | Salon / Barbershop | 5 | Grocery Store | 8 |
| 18 | Thai Restaurant | 5 | Mediterranean Restaurant | 8 |
| 19 | Nail Salon | 4 | Airport Service | 6 |
| 20 | Mexican Restaurant | 4 | Perfume Shop | 6 |

| NO | Cluster 6 | Counts |
|---|---|---|
| 1 | Art Gallery | 53 |
| 2 | Wine Shop | 8 |
| 3 | Clothing Store | 7 |
| 4 | Café | 6 |
| 5 | Boutique | 4 |
| 6 | Asian Restaurant | 3 |
| 7 | Salon / Barbershop | 3 |
| 8 | History Museum | 3 |
| 9 | Chinese Restaurant | 3 |
| 10 | Lounge | 2 |
| 11 | Bar | 2 |
| 12 | Pub | 2 |
| 13 | Nightclub | 2 |
| 14 | Building | 2 |
| 15 | Cocktail Bar | 2 |
| 16 | South American Restaurant | 1 |
| 17 | Poke Place | 1 |
| 18 | Performing Arts Venue | 1 |
| 19 | Pie Shop | 1 |
| 20 | Thrift / Vintage Store | 1 |

As for the first analysis with k_clusters=5, I plotted every top 2 frequent categories in each cluster by their counts and weight in <Figure 12> and <Figure 13>.




<Figure 12>

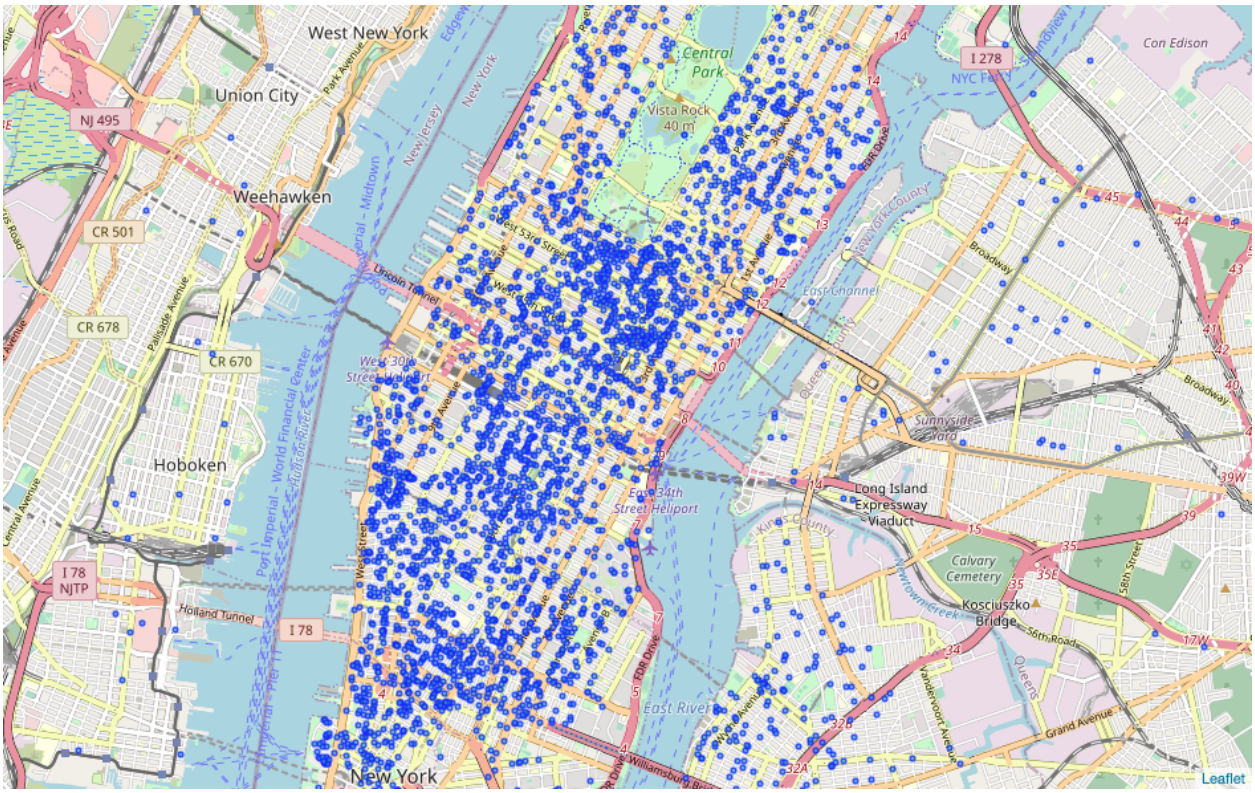



<Figure 13>

By increasing the number of clusters to 7, I could see new categories such as Airport Service, Deli / Bodega, American Restaurant, Steakhouse and Wine Shop.

## 4. Results

As shown in the map below, the Uber pick-up locations that were used to call the Foursquare API were very well distributed in the entire Manhattan area and I had not expected to see any dominant venue categories other than a few categories that I could easily guessed before running the k-means clustering algorithm and carefully examined the results. It was somewhat expected to see categories like Bar, Coffee Shop, Hotel and Theater; however, the other categories such as Italian Restaurant, Clothing Store, Art Gallery, and Wine Shop were very interesting findings.



<Figure 14>

## 5. Conclusion

New York city is very diverse, and a lot of people use car sharing services like Uber. Even though the pick-up locations seemed to be randomly scattered around without any patterns, it was very interesting to see there were very popular venue categories in the clustered locations. I could have found some different results if I could have collected venue information for more pick-up locations. Because of the daily limit of API calls, I was not able to collect more data within a relatively short period of time. However, I collected 20,156 different venues for 4,228 different pick-up locations, and it was still very interesting to see how all those venue categories were clustered into several distinctive clusters. I mainly used Python programming language and its packages such as Pandas, NumPy, matplotlib, folium and sqlite3. In addition, I used Excel, SQLite Studio, and Jupyter Notebook running on my local machine as tools. I realized that there were so many useful libraries and tools to manipulate, analyze and visualize the data. I utilized existing datasets like the Uber pick-ups and Foursquare location data, and successfully proved that there was some level of relationship between Uber pick-up locations and nearly venues by showing the top common venue categories near those locations.

## 6. Future Direction

The Uber pick-ups dataset has the Date/Time column which tells when an Uber passenger was picked up at that location. The Date information can provide what day was the passenger was picked up, and Time information can provide whether it was morning, afternoon, evening, or late night. The Date information can also provide whether it was a weekday or weekend. These can be another set of features that we can use for clustering algorithms. I guess that I would see categories like Night Club and Bar if I have used Weekday vs Weekend, and Day vs Night as features.