Classifying New York City Subway Stations

Ryan Albertson
Coursera – Applied Data Science Capstone
December 2019

1. Introduction

Every day, roughly five million people ride the New York City subway system, which expands 665 miles across the five boroughs of the city. The system is currently comprised of 473 subway stations. I will be employing data from the *Foursquare API* to learn more about each of these stations. Specifically, I will pull venue data to understand which types of establishments are prominent around each station. From this information, the traffic of each station can be estimated. For example, a station with primarily university venues nearby will be busier on weekdays and not so much on the weekends. Whereas an area that has many entertainment venues will likely experience a traffic spike during the weekends. This information will be useful for city planners, because the classification of the subway stations will provide a good estimation of human traffic within a radius of each station. I will also use k-means clustering to see if the clusters have any correlation to the venue categories of each station.

2. Data

The dataset was taken from *NYC Open Data* (https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49). The only features needed from the dataset are the subway station names, and the matching geographical coordinates. There is a total of 473 subway stations, each with a pair of latitude and longitude coordinates. From the Foursquare API, the names of the ten main categories of venues were first obtained. Then all venues within a given radius of each subway station were retrieved. With this data, the venues surrounding each station will be ranked from first most common to fifth most common based on their count. These classifications will be saved to every station.

3. Exploratory Data Analysis

A. Plotting Subway Stations on a Map

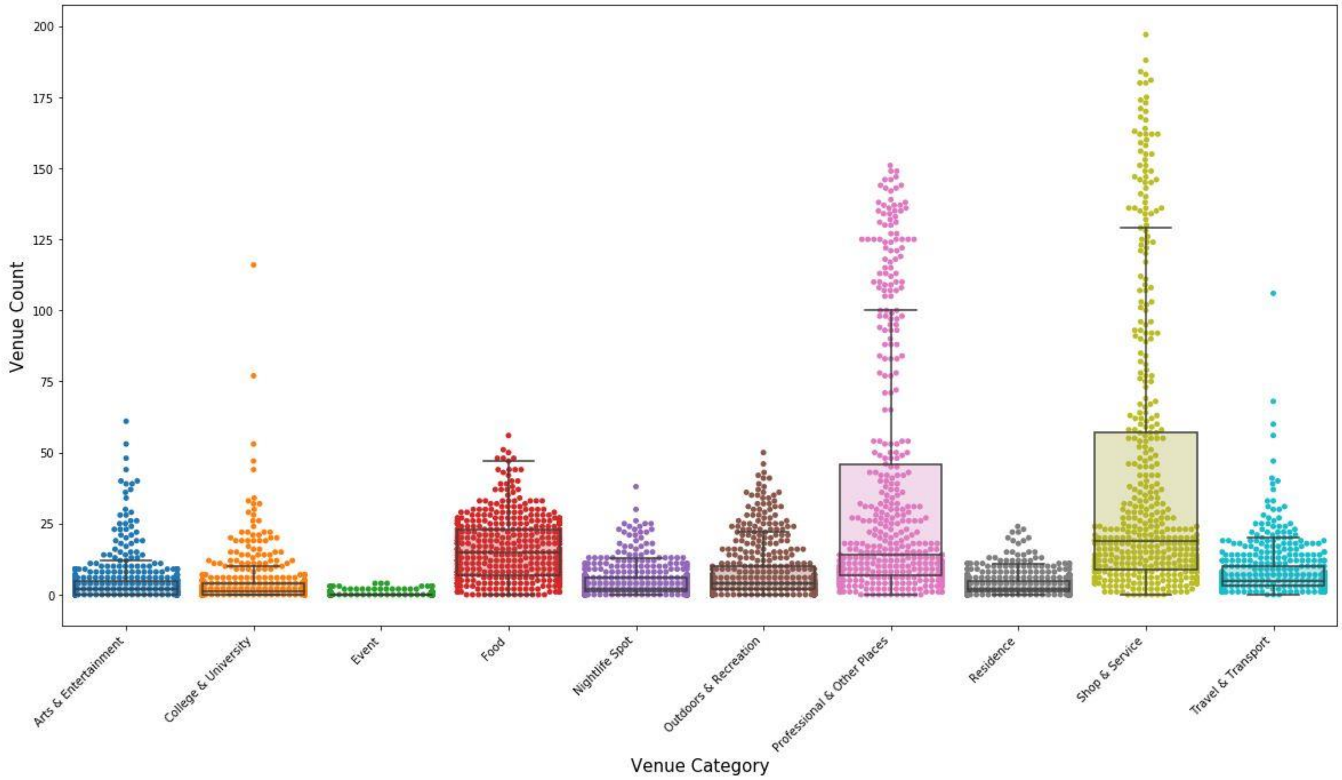
The first thing I did with the dataset was map each subway station. This gave me an idea as to the geographical scope of all the stations. I could also see the size of the search radius, making sure that the size wasn't overlapping with other search radiuses while also being sufficiently bound itself.

B. Finding Venues for Each Subway Station

I retrieved the ten main categories of venues from the Foursquare API. After which I found all venues within a 200-meter radius of each station. I placed each of these venues into one of the ten categories and added on to the dataset such that every station would have all of its venues categorized. Next, I checked which subway stations have the most total venues within 200-meters. I juxtaposed the top twenty list with ridership data from (http://web.mta.info/nyct/facts/ridership/ridership_sub.htm). There was a weak correlation between the count of venues and ridership of stations. Therefore, the number of venues near a subway station can't be used to predict the station's busyness.

C. Visualizing Distribution of Venue Category Counts for Each Station

I created a boxplot with a swarm plot overlaid as to show how many stations had each amount of venues in each category.



D. Creating Top 5 Most Common Venue Categories for Each Station

From the pool of venues surrounding each subway station, I counted the top 5 most common venue categories and added them to each station in the data frame. This arrangement was used to mark each station with a color that represented the venue category for that station. Five maps were made, ranking the first most common venue category to the fifth most common venue category.

E. Using K-Means Clustering

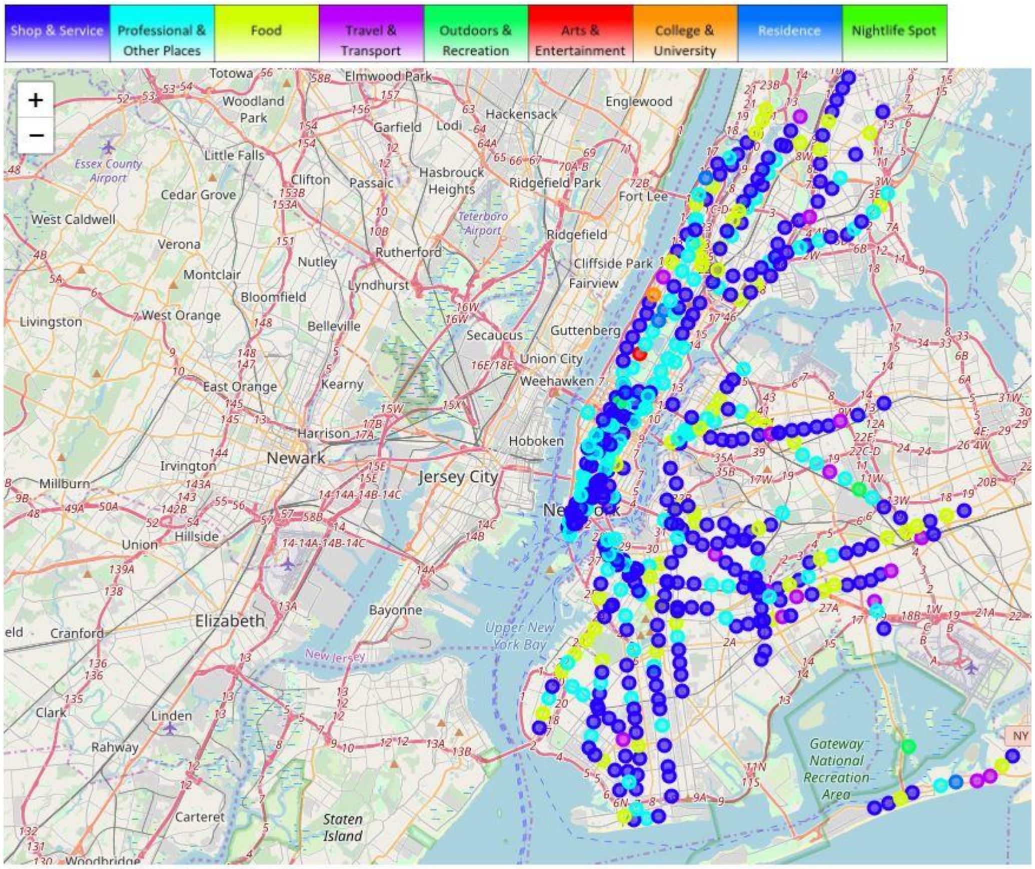
The K-means clustering method was used to distribute subway stations into several clusters. The criteria for clustering was the Euclidean distance between the stations. The K-means algorithm minimizes the distance between stations within the same cluster while maximizing the distance between the stations in different clusters. The resulting clusters can then be analyzed to determine if there are any similarities within them, besides distance.

4. Results

A.) Most Common Venue Categories for Each Subway Station

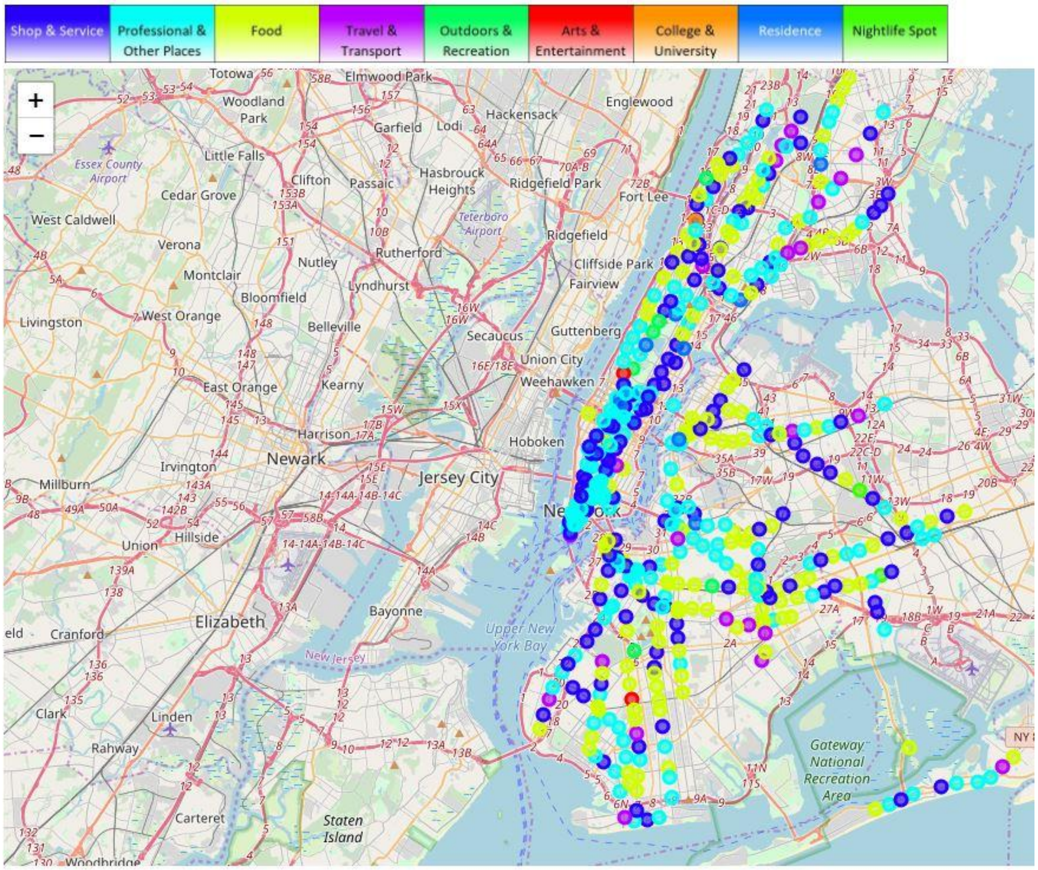
The top five most common venue categories for each subway station in New York, along with an analysis of them is as follows.

1. 1st Most Common Venue Categories



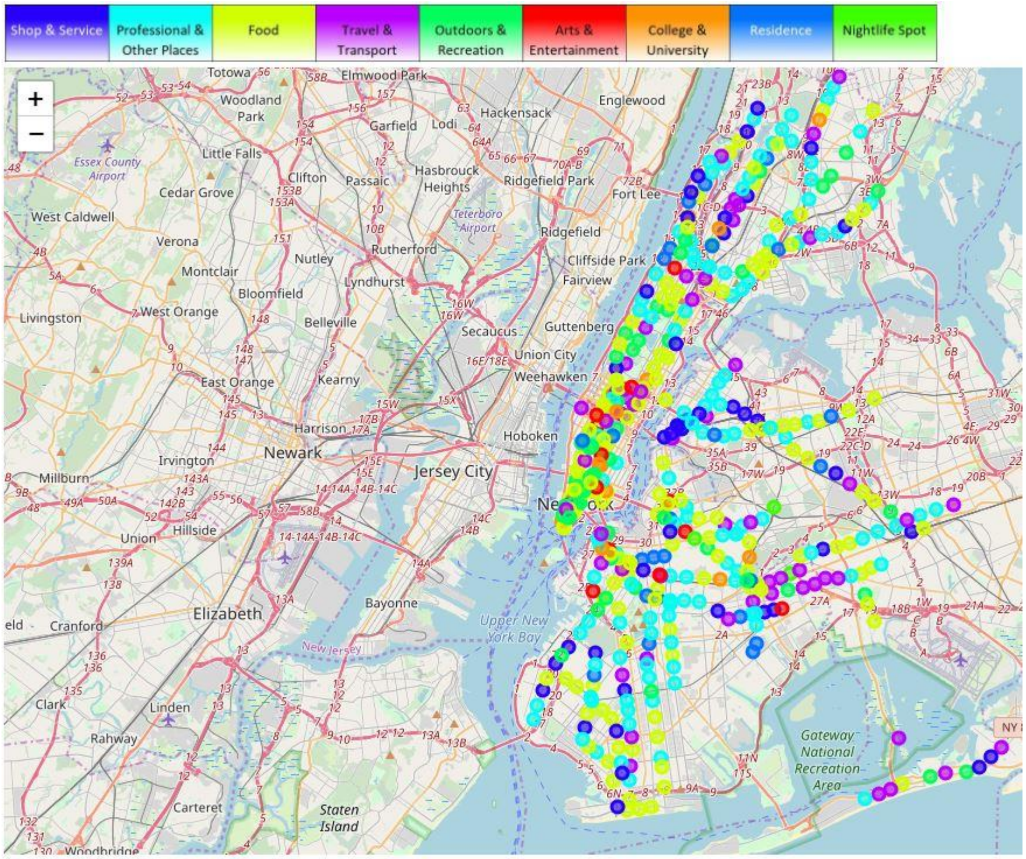
This map illustrates that the overwhelming majority of New York City venues are under the category of *Shop & Service* or *Professional & Other Places*. Some examples from these categories include stores, markets, banks, offices, municipal buildings, and hospitals, via (https://developer.foursquare.com/docs/resources/categories). Venues categorized as *Professional & Other Places* should be the busiest during weekdays. During these times, many people will be commuting to them for work. The *Shop & Service* venues will likely be crowded every day of the week. The count of venues around those stations may be a more significant way to determine the traffic congestion of the stations and their surrounding area.

2. 2nd Most Common Venue Categories



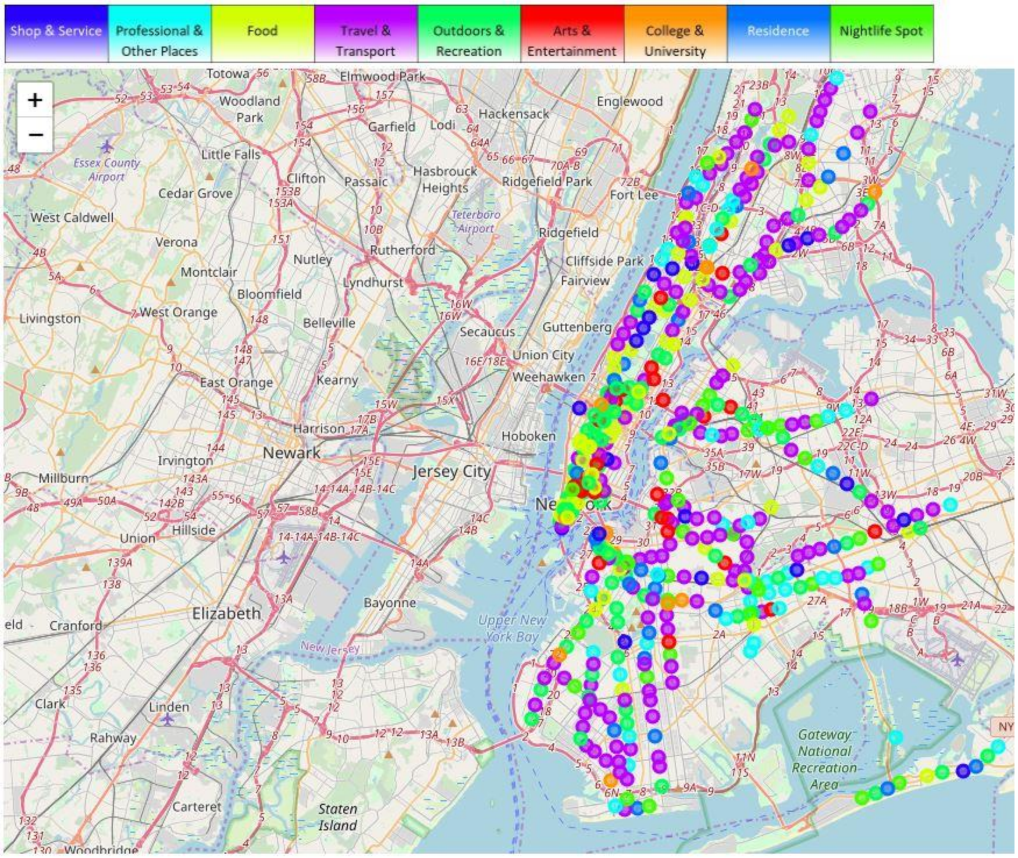
Now that the scope is dialed to second most common venues, many food venues appear. This isn't surprising since New York is known to have many places to eat. This indicates that there are still many food venues across the city, there simply aren't as many as there are other types of venues. The busyness of stations attributed to these food venues is presumably constant across days of the week. During the weekdays, many small places, such as coffee shops, food trucks, and fast food restaurants will be busy when workers grab lunch or something to eat before or after work. Then on weekends, the sit-down restaurants will have their peak sales. So the variety of food venues is what makes the category difficult to use as a metric to predict subway station congestion.

3. 3rd Most Common Venue Categories



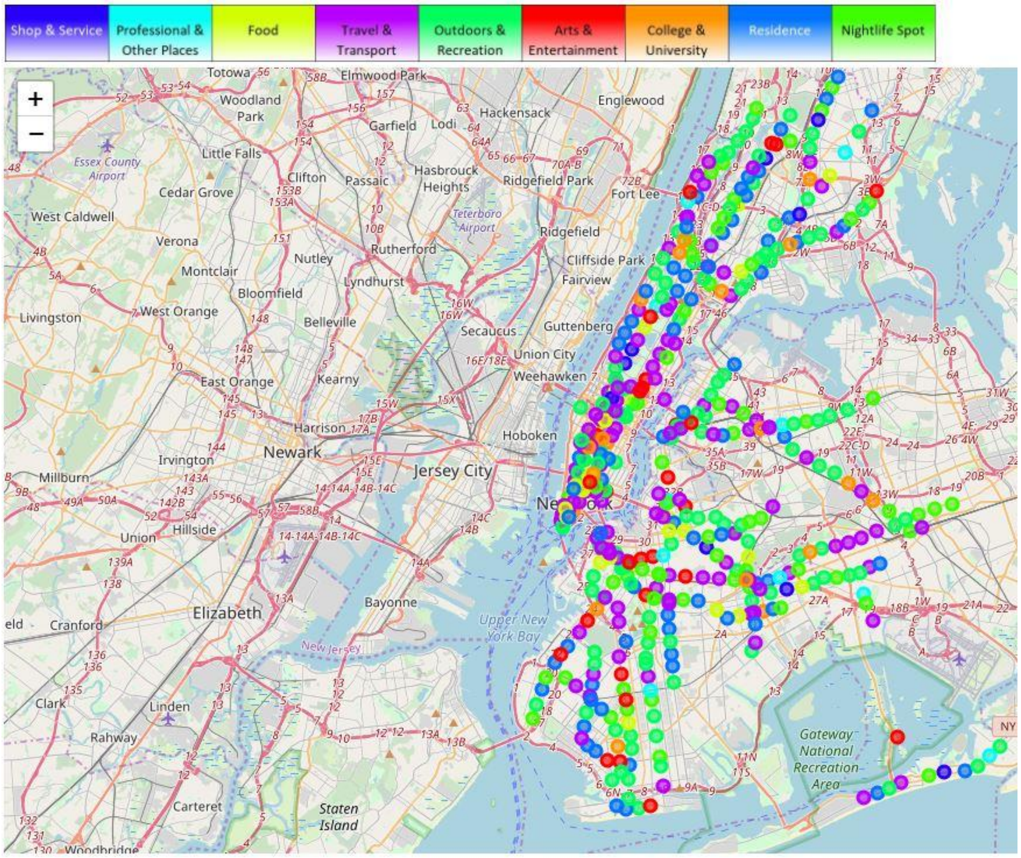
When the third most common venues are revealed, many food venues still appear. It seems that for most of the stations, if food wasn't the second most common, then it's the third most common. Before getting to the third most common venues, there were only a handful of visible food venues in lower and middle Manhattan. This implies that there is such an abundance of professional and shopping venues there that they swamp out the food venues, which are certainly packed into that area, nonetheless.

4. 4th Most Common Venue Categories



Now many *Travel & Transport* venues take the stage. This is another category that will spike traffic in areas during weekdays. Examples of venues include taxis, train stations, bus stops, boats, and airports. The commute to work for millions of people will ensure that these stations are crowded as people get to or from these transportation venues. *Outdoors & Recreation* markers also erupt throughout the map of NYC. These could be places ranging from a golf course to a hiking trail or even a national park. People will be drawn to these during weekends primarily, which is when they have the free time to do so.

5. 5th Most Common Venue Categories

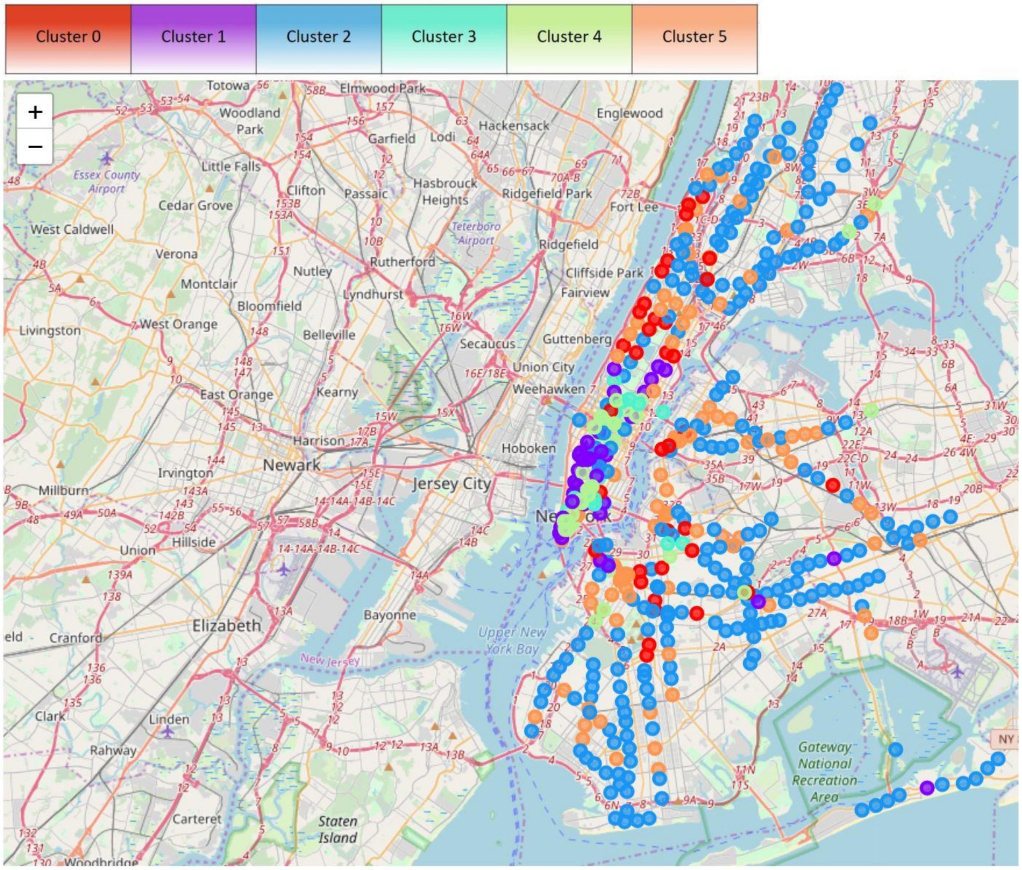


These are the fifth most common venues, so they aren't as significant as the previous ones. Nonetheless, they will help in determining the traffic of subway stations. The underlying *Residence* venues appear here, they are scattered across the city. During weekdays people will commute to and from work using the subway station closest to their residence. On the weekends, people generally won't commute as much so the correlation with station congestion may be negative. Following the same logic, busy during the week and not-so-much on weekends, are the *College & University* venues that arise in a few places. Finally, there are decent amount of *Art & Entertainment* venues that show themselves. These are places that people will enjoy to visit on weekends usually.

B.) K-Means Clustering of Subway Stations

The Silhouette Method was used to determine which number of clusters to use. The result was six, so clusters 0 through 5 were allocated. After running the algorithm, all 473 subway stations were put into a cluster.

This is a geographic representation of the clusters.



The clusters that appear to distinctly represent some area of NYC are cluster 1, 3, and 4. All three are centered around Lower Manhattan, thus it's not a surprise to learn that the primary makeup of these clusters is *Shop & Service* and *Professional & Other Places* venues. Comparing these three clusters with the other three; 1, 3, and 4 barely have instances of venue categories besides the two mentioned. Cluster 2 is the polar opposite. It is made up of a collection of all nine venue categories and as illustrated above, it stretches across the four boroughs. After using a few different values for the number of clusters, I didn't find any clusters that had significant venue category attributes. It seems the relation between the first, third, and fourth clusters is the best that can be found while using a k-value between 3 and 9.

5. Discussion

After relating the categories of venues with each of the subway stations, some meaningful inferences can be made. Specifically with the first most common venue categories, city planners can use this data for whichever projects may relate to human traffic in New York City. By associating the categories with times of congestion, this data can be useful for predictions. Another step can be taken to retrieve the sub-categories of venues and join them to each station accordingly. This would reduce the number of data points for each variable of the data, but in return provide more specific information. The biggest challenge will be to establish the magnitudes of the first most common through fifth most common venue categories. How significant is the first most common, then how much less significant is the second most common, etc.

Moving on to the discussion about the k-means clustering. Using Euclidean distance as the classifier of the k-means algorithm and a k-value of six, three clusters of interest were discovered. Looking at the first and second most common venue categories for all three, each had almost exclusively *Shop & Service* and *Professional & Other* venues. Seeing as these three clusters are also centered geographically around Lower Manhattan, they could be a good candidate for further analysis.

6. Conclusion

In this project, I analyzed New York City subway station geospatial data along with venue data from the Foursquare API. A search radius around each station was defined, with which all venues within were found. Then 10 main categories for the venues were defined and each of the previously discovered venues was classified as one of them. The top five categories based on count were calculated for every station. This provides good insight towards accounting for human traffic in and around the subway stations. The amount of venues and their category can be used to set thresholds for the amount of congestion in areas of New York. City planners, for example, could use this when planning constructions or events in areas near the subway stations. The k-means clustering returned three clusters that share common attributes of distance and venue categories. This could be a promising exploration if someone were to further analyze the clusters.