# Predicting where to place a New Restaurant in Manhattan

Roopene Hariharan

13-June-2020

## 1. Introduction

### 1.1 Background

Everyone around the world craves for taste and loves good food. This is the reason why restaurants and café run more everywhere in the world. There are people who travels only to taste good food. Therefore, it is important for every restaurant to be in such a place that more people visit them because each restaurant and café is different by itself. So, we got to be precise on recommending a neighborhood for launching a new restaurant. Our major assumption here after analyzing the initial dataset is that Manhattan has lot of restaurant, so we got to locate the similar neighborhoods with the greatest number of restaurants and suggest that as the best place.

**Target Audience:**

The target audience for this analysis will be the one who is searching for a best place in Manhattan to start their new restaurant.

### 1.2 Problem

Data we will be analyzing might include Boroughs, their Neighborhoods, and list of Venue categories around each neighborhood. With that we would be able to find the Borough with the greatest number of restaurants in their top 10 most common venues. This will be our recommendation to launch a New Restaurant.

### 1.3 Interest

Entrepreneurs will be interested in knowing the neighborhoods which has the pattern for having the Restaurants as their top common venues. Also, considering similar Neighborhoods with same neighborhoods where there are a smaller number of Restaurants, this will be the place where a Restaurant is likely to have better reception and less competition.

## 2. Data Acquisition and Data Cleaning

### 2.1 Data Sources

Our initial Borough and Neighborhood details for entire New York City can we be taken from the link here. Then we can use the Four Square API service to extract the list of venue categories around each neighborhood. Geocoder package can be used to extract the Latitude and Longitude data of each Neighborhood we are taken from the above Link

### 2.2 Data Cleaning

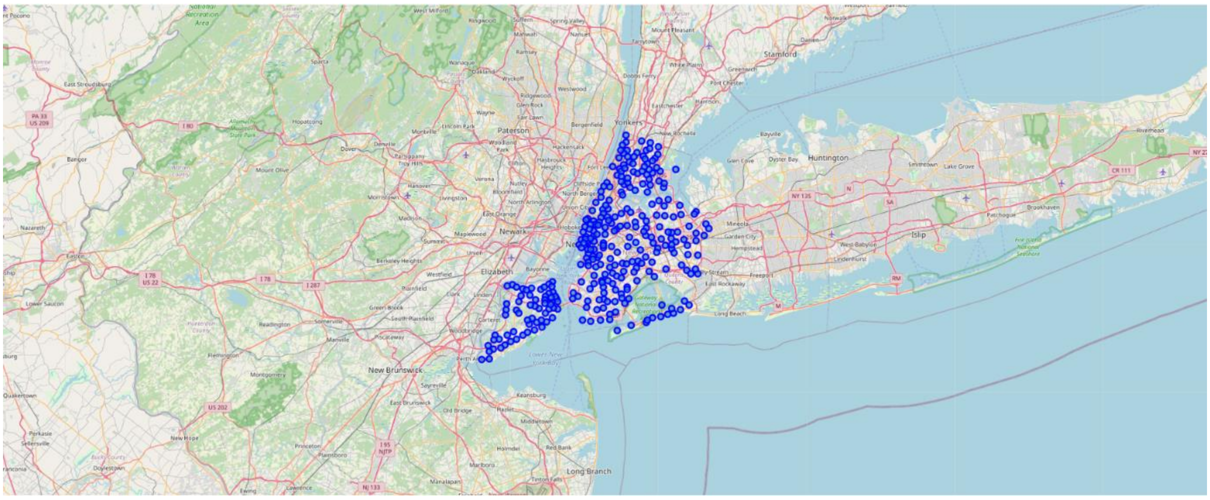
The data we got from the above link is a Json file so as a first step we need to transform that into a Pandas Dataframe. This is done by looping through the Json file by specifying the data we need like Borough, Neighborhood, Latitude and Longitude which will be my Columns of the dataframe.

So, now we have a dataframe with 5 boroughs and 306 neighborhoods with their latitude and longitude details.

To identify pattern between these neighborhoods we need to split them into Clusters, once we find the clusters between these neighborhoods, we need to extract each clusters in to different individual DataFrame for us to analyze the Restaurants accumulation in each neighborhood.

## 3. Exploratory Data Analysis

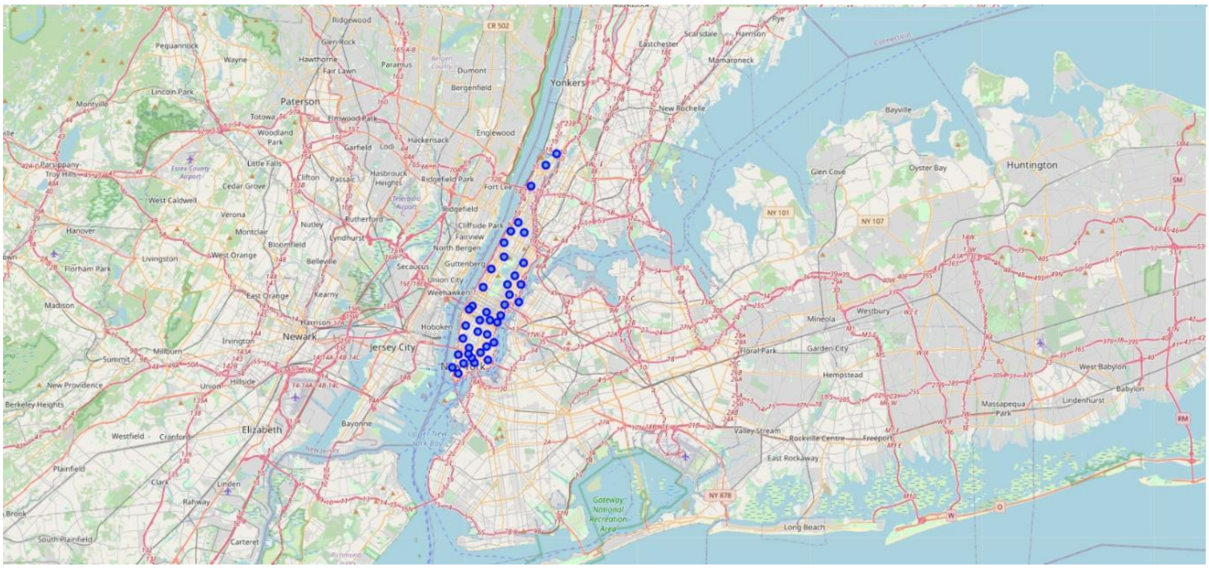
As an initial analysis we create a map to view the neighborhoods plotted in it using folium.



*Neighborhood Plots of New York City*

We further need to get the Venues situated in each of Neighborhoods and see if these neighborhoods are similar with respect to the venues situated.

For our analysis, we need to take only the Manhattan Borough. So, we extracted the Manhattan borough alone from the entire New York data and saved it as a separate dataframe with the list of latitudes and longitudes of Manhattan Neighborhoods. As a result, we got 40 neighborhoods for which we plot a map to see any visual cluster.



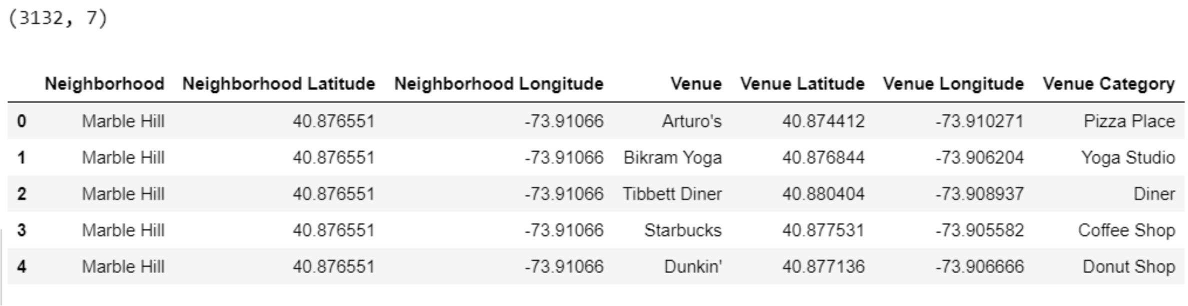
*Neighborhood Plots of Manhattan*

**Using foursquare API:**

Now, we use the foursquare API to analyse the venues of the first neighborhood in our data frame. The input provided will be the foursquare credentials, Neighborhood latitude, longitude, radius and limit of the venues that needs to be extracted.

The result will be a json file from which we will extract the Name, Category, Latitude, Longitude of the Venues in the specified radius.

After this we will do the same for the entire Manhattan data, which returns 3132 venues.

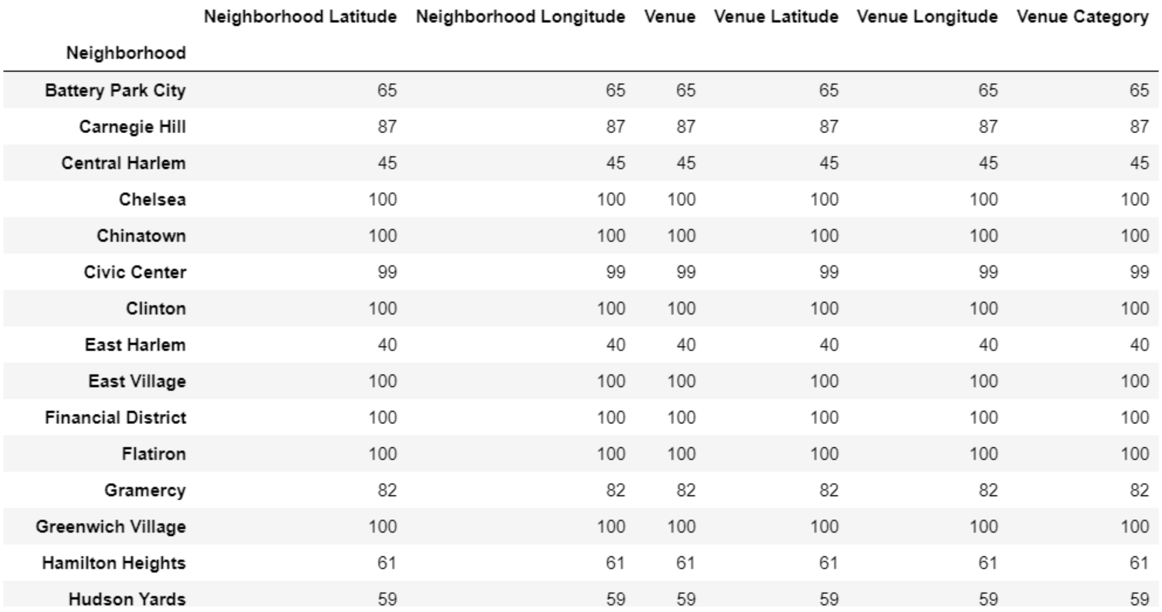


```
(3132, 7)
```

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Marble Hill | 40.876551 | -73.91066 | Arturo's | 40.874412 | -73.910271 | Pizza Place |
| **1** | Marble Hill | 40.876551 | -73.91066 | Bikram Yoga | 40.876844 | -73.906204 | Yoga Studio |
| **2** | Marble Hill | 40.876551 | -73.91066 | Tibbett Diner | 40.880404 | -73.908937 | Diner |
| **3** | Marble Hill | 40.876551 | -73.91066 | Starbucks | 40.877531 | -73.905582 | Coffee Shop |
| **4** | Marble Hill | 40.876551 | -73.91066 | Dunkin' | 40.877136 | -73.906666 | Donut Shop |

*Sample of entire Manhattan Venues*

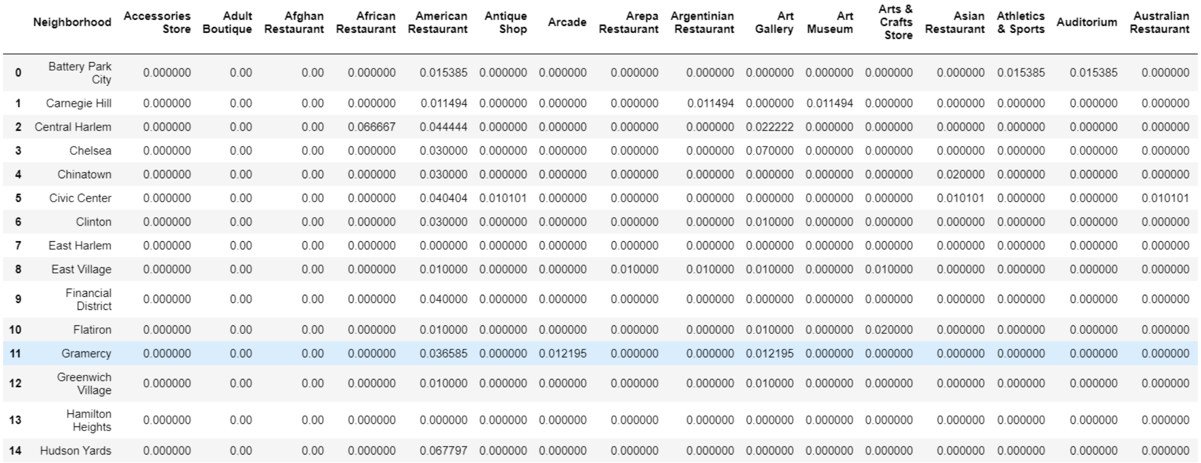
Just for our understanding, lets analyse the count of venues in each neighborhood.



| | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| **Neighborhood** | | | | | | |
| **Battery Park City** | 65 | 65 | 65 | 65 | 65 | 65 |
| **Carnegie Hill** | 87 | 87 | 87 | 87 | 87 | 87 |
| **Central Harlem** | 45 | 45 | 45 | 45 | 45 | 45 |
| **Chelsea** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Chinatown** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Civic Center** | 99 | 99 | 99 | 99 | 99 | 99 |
| **Clinton** | 100 | 100 | 100 | 100 | 100 | 100 |
| **East Harlem** | 40 | 40 | 40 | 40 | 40 | 40 |
| **East Village** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Financial District** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Flatiron** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Gramercy** | 82 | 82 | 82 | 82 | 82 | 82 |
| **Greenwich Village** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Hamilton Heights** | 61 | 61 | 61 | 61 | 61 | 61 |
| **Hudson Yards** | 59 | 59 | 59 | 59 | 59 | 59 |

*Venue Count of each neighborhood*

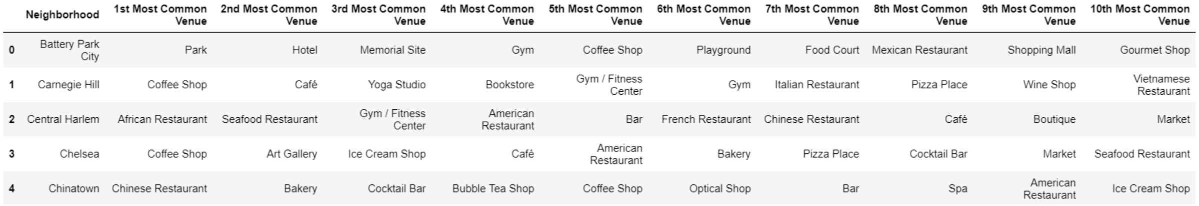
Now it’s time is get a unique category list and group then based on Neighborhood and hotspot. Below is the sample of dataframe with mean of the frequency of occurrence of each catergory.



| | Neighborhood | Accessories Store | Adult Boutique | Afghan Restaurant | African Restaurant | American Restaurant | Antique Shop | Arcade | Arepa Restaurant | Argentinian Restaurant | Art Gallery | Art Museum | Arts & Crafts Store | Asian Restaurant | Athletics & Sports | Auditorium | Australian Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.015385 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.015385 | 0.015385 | 0.000000 |
| 1 | Carnegie Hill | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.011494 | 0.000000 | 0.000000 | 0.000000 | 0.011494 | 0.000000 | 0.011494 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 2 | Central Harlem | 0.000000 | 0.00 | 0.00 | 0.066667 | 0.044444 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.022222 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 3 | Chelsea | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.030000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.070000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | Chinatown | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.030000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.020000 | 0.000000 | 0.000000 | 0.000000 |
| 5 | Civic Center | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.040404 | 0.010101 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.010101 | 0.000000 | 0.000000 | 0.010101 |
| 6 | Clinton | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.030000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 7 | East Harlem | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 8 | East Village | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0.010000 | 0.010000 | 0.010000 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 9 | Financial District | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.040000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 10 | Flatiron | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.010000 | 0.000000 | 0.020000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 11 | Gramercy | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.036585 | 0.000000 | 0.012195 | 0.000000 | 0.000000 | 0.012195 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 12 | Greenwich Village | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 13 | Hamilton Heights | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 14 | Hudson Yards | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.067797 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |

*Sample of the table for Venue Frequencies*

Now its time to find out the top 10 venues of each neighborhood by sorting the venues in descending order



| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | Park | Hotel | Memorial Site | Gym | Coffee Shop | Playground | Food Court | Mexican Restaurant | Shopping Mall | Gourmet Shop |
| 1 | Carnegie Hill | Coffee Shop | Café | Yoga Studio | Bookstore | Gym / Fitness Center | Gym | Italian Restaurant | Pizza Place | Wine Shop | Vietnamese Restaurant |
| 2 | Central Harlem | African Restaurant | Seafood Restaurant | Gym / Fitness Center | American Restaurant | Bar | French Restaurant | Chinese Restaurant | Café | Boutique | Market |
| 3 | Chelsea | Coffee Shop | Art Gallery | Ice Cream Shop | Café | American Restaurant | Bakery | Pizza Place | Cocktail Bar | Market | Seafood Restaurant |
| 4 | Chinatown | Chinese Restaurant | Bakery | Cocktail Bar | Bubble Tea Shop | Coffee Shop | Optical Shop | Bar | Spa | American Restaurant | Ice Cream Shop |

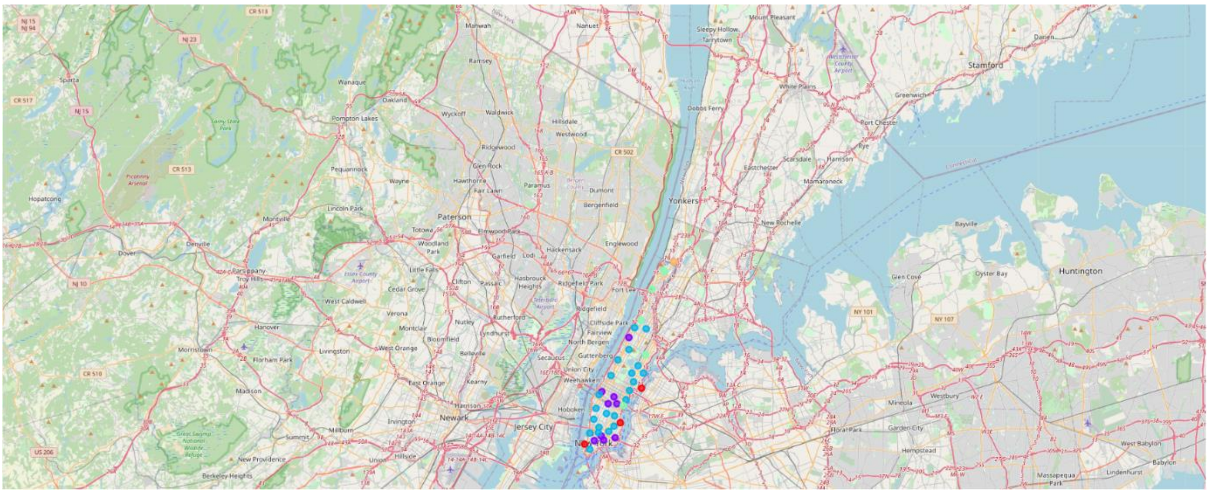
*Sample of top venues of each neighborhood*

# 4. Classification Modelling

## 4.1 k-Means Clustering

For our classification of the Neighborhoods which share the similar venues and similar lifestyle, we chose k-Means classification. k-Means clustering is a method of vector quantization, originally from signal processing that aims to partition n observations into k number clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Which will create number of clusters, each having list of similar neighborhoods.

we plan to specify the number of clusters to be 5 as initial setup, which seems to be optimum or the Manhattan data. My random state for the k-Means would be 0. This would give me a picture of which neighborhoods share the same pattern of lifestyles.

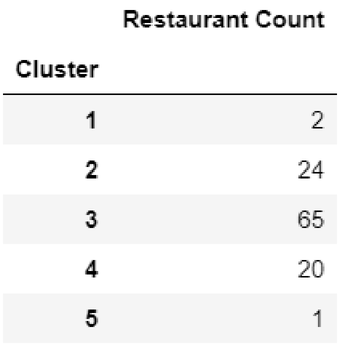


*Clusters of neighborhoods with different Lifestyle*

This plot of the Clusters gives us a clear picture that when we travel from the south its mostly of same pattern of venues that is same set of lifestyles. In between there comes different lifestyle also in minimum number.

Once these clusters are determined, we placed them in different Dataframes i.e. 5 Dataframes each has different clusters.

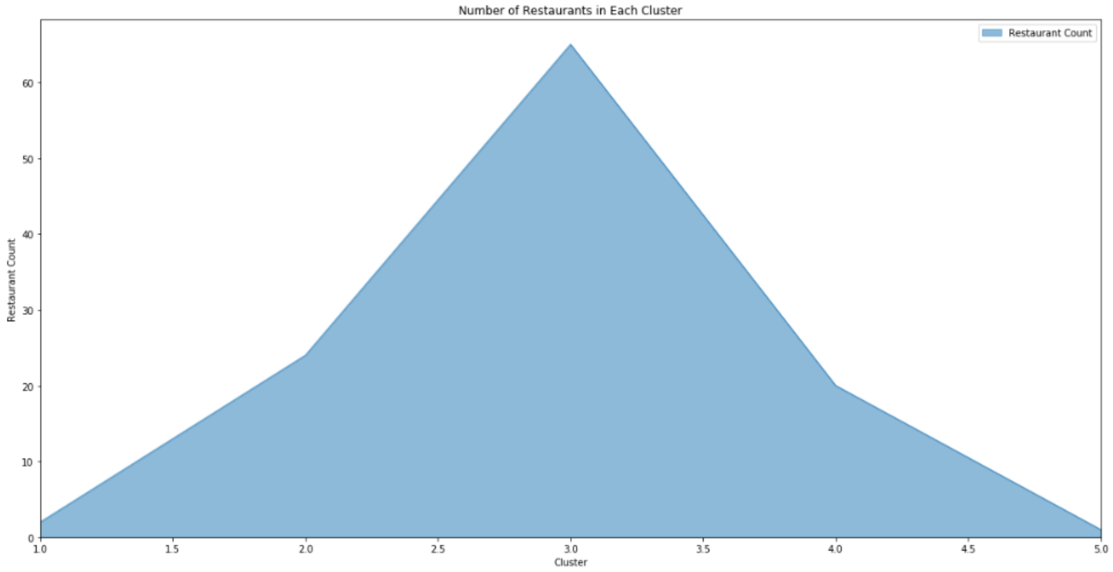
To find which neighborhood has the greatest number of Restaurants in these clusters, we need to find the venues with the category string as “Restaurant” and find the total count for each cluster. These will in-turn give us a Dataframe with Clusters mapped with the number of Restaurants. Most importantly we are referring to the Dataframe we created with top 10 common venues.



| | Restaurant Count |
|---|---|
| **Cluster** | |
| **1** | 2 |
| **2** | 24 |
| **3** | 65 |
| **4** | 20 |
| **5** | 1 |

*Clusters with Restaurant Count*

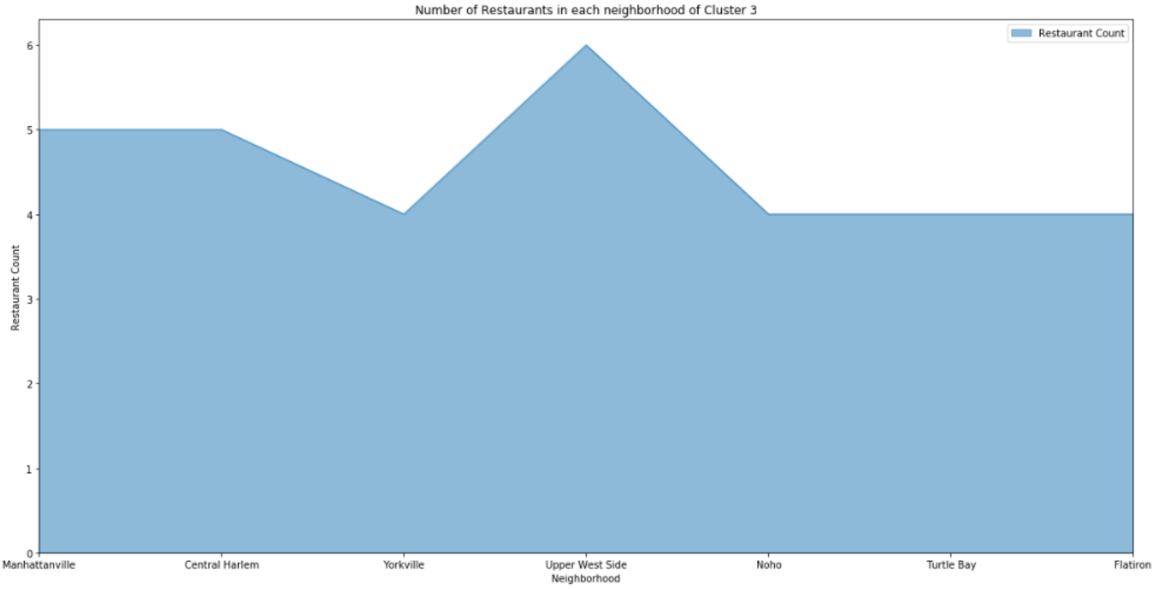
While doing this analysis, we found that cluster 5 has higher number of Restaurant counts which in turn was plotted in a histogram graph for better understanding/visualisation.



*No. of Restaurants in each Cluster*

From this it is found that cluster 3 has a higher number of restaurants. Now it is time to find out the correct neighborhood in cluster 3.

So, we found the restaurant count of each neighborhood in cluster 3 and plotted a histogram graph.



*No. of Restaurants in each neighborhood of cluster 3*

## 4.2 Insights

To get a better view of the story in this DataFrame we extracted, we plotted them using histogram plot.

- **Insight 1** – The first histogram gives a clear picture that only in Cluster 3 we have higher number of restaurants with completely different lifestyle in it.

- **Insight 2** – In the second histogram, we come to know that "Upper West Side" has the highest restaurant count compared to other places
- **Insight 3** – Also, regarding the other neighborhoods with same lifestyle, there are nearly less number of restaurants.

## 5. Conclusion

In this study, we analyzed the Borough, Neighborhoods across New York City with respect to their Venues pattern to identify similar lifestyles. We found varies patterns across the city and tried to determine restaurants count for digging the hidden insights that would help us to recommend a neighborhood for launching a new restaurant. So, considering the Insights derived, we have derived some recommendations, these would be functioning across the business strategy of the entrepreneur

**Strategy 1** – Wanting to Launch a Restaurant in a Neighborhood where majority of venues are Restaurants, considering they would catch the same growth as other venues

For this strategy, Insight 1,2 would help to decide as it explains the neighborhoods with most number of restaurants is cluster 3. So my recommendation would be to launch a new restaurant in Cluster 3 more across all boroughs, but specifically on Upper West Side, as they have the most number of restaurants as the top 10 common venues

**Strategy 2 -** Wanting to Launch a Restaurant in a Neighborhood where people likely to visit Restaurant but they have moderate number of Restaurants

For this strategy, Insight 2 and 3 would help to decide as it explains the neighborhood with opportunity to have a greater number of restaurants but have moderate restaurants. So, my recommendation would be to launch a new restaurant in Cluster 3 in Sutton Place can be a good place to start a New Restaurant.

## 6. Future directions

As we have the Venue Category as my most trusted data feature to determine the recommendation of the neighborhood for launching new restaurant. In future, we can also avail other data that Four Square API provides, that is we can also include User reviews and how frequent users visit these restaurants, to derive a rating of these restaurants or identifying most suitable neighborhood which has more number of visitors to the restaurants.