The Battle of Neighbourhood final Project Report

Introduction / Problem Description

To open a new restaurant in a particular area my client wants to know what type of food Restaurant are not present in that area so that he can think of opening a Restaurant of that particular cuisine.

This system will analyse local geographical data via Foursquare and recommends restaurant/eatery type (cuisines) for upcoming eateries and restaurants based on one's neighbourhood preference for a better shot at success. It analyses all the eateries and restaurants in every neighbourhood of a city and then creates a list of top 10 spots (Restaurant/Eatery type) in every neighbourhood displayed in percentages of the total restaurants in that particular neighbourhood.

So my client can use this to see which cuisine based restaurants are lacking or what type of Restaurants are doing well (due to their high number) in which neighbourhoods. He can then make an informed decision and fill the void and have a better chance of establishing a successful business.

For example, suppose Anand who is really passionate about food and different cuisines wants to invest his savings in the restaurants business in New York. Since he wants to make sure a steady Return on his Investment, he takes my help of Data Science to analyse the restaurant data of the New York City. He can get the desired data per every neighbourhood for the entire city and see which cuisine based restaurants are the least in number per neighbourhood or he can simply see the same list for his choice of neighbourhood. Suppose he is leaning towards opening a restaurant in the 'Midtown' area, he then uses my analysis to see which type of restaurants are prevalent in that particular area and in the adjoining neighbourhoods. He observes that among all other cuisines, the area of his choice is lacking 'Mughlai' and an 'Indian' joint. So the analysis helps Anand make a decision on opening the Restaurant.

The Required data for the analysis

Datasets required for the Project:

1) **New York Data (Boroughs + Neighbourhoods)**
   a. The first Dataset we will be using would contain all the required geographical data about New York City. Namely, we would be using 'Borough', 'Neighbourhood', 'Latitude', and ‘Longitude’ among all the other data elements present in the data. For convenience, we would be using the same data set which was provided to us in Week 3 of this course (Applied Data Science Capstone)
   https://geo.nyu.edu/catalog/nyu_2451_34572
   We will use the same link like we did to load this data from where it is downloaded and hosted on (https://cocl.us/new_york_dataset)
   b. New York City has a total of 5 boroughs and 306 neighbourhoods. In order to segment the neighbourhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood. The 'Latitude' and 'Longitude' extracted from this dataset will also be pivotal when we use it perform

Clustering using K-Means. All the relevant data is in the features key, which is basically a list of the neighbourhoods. If we dive into the elements of this features key, we will find all of its components.

2) **Four Square City Guide Data (Venues)**

   a. Foursquare City Guide, commonly known as Foursquare, is a local search-and-discovery mobile app which provides search results for its users. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history.

   b. So using the Foursquare API, we can search for specific type of venues or stores around a given location. It is important to remember that for this data, we make a regular call to the API, and if you have a free personal developer account, you can make up to approximately 99 thousand regular calls per day. We can also learn more about a specific venue or store or shop, like their full address, their working hours, and their menu if they have one, and so on. It's also important to remember that for this data, we would need to make a premium call and with the personal developer account, you can make approximately 500 calls per day. Also with the Foursquare API, we can learn more about a specific foursquare user, their full name, and any tips or photos that they have posted about venues and stores. For this data, a regular call to the API would be made. Furthermore, we can explore a given location by finding what popular spots exist in the vicinity of the location, and for this data a regular call to the API would be made. And finally, with the Foursquare API, we can explore trending venues around a given location. These are venues with the highest foot traffic at the time this regular call to the API is made

Exploratory Data Analysis

1. Exploring Datasets

Neighbourhood has a total of 5 boroughs and 306 neighbourhoods. In order to segment the neighbourhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood.

- Extract all the relevant data which are basically a list of Neighbourhoods
- Transfer the extracted data into a Data Frame
- Fill the Data Frame with 'Borough', 'Neighbourhood', 'Latitude', 'Longitude' data
- For display, create a map of NY city with Neighbourhoods superimposed on top
- Segment the neighbourhoods of all 5 boroughs
- Visualize all 5 boroughs with all the neighbourhoods in it (Brooklyn, Manhattan, Queens, Bronx, and Staten Island. Explore all 5 Data Frames which are created for every borough
- Use Foursquare API to explore venues in all the neighbourhoods
- Extract category and clean the json file to produce new Data Frames
- Repeat calling the Foursquare API for all 5 boroughs

2. Explore Neighbourhoods in all 5 Boroughs

- Extract complete neighbourhood list for all 5 boroughs
- Create Data Frames covering the venues of all neighbourhood
- Group all the returned venues by 'Neighbourhood'
- Find out how many unique categories are present in returned 'venues'

3. Analyse Each Neighbourhood of Each Borough

We will analyse each neighbourhood of each borough (Brooklyn, Manhattan, Queens, Bronx and Staten Island) and the final product of this stage would be a Data Frame which top 10 most common Restaurant/Eatery type in each neighbourhood which would be repeated for each of 5 boroughs.

- Create One hot encoding Data Frames for each Borough
- Group rows by 'Neighbourhood'
- Take the mean of the frequency of occurrence of each category
- Manually Selecting (Subletting) Related Features for the Restaurants/Eateries
- Updating the One-hot Encoded DataFrame
- Extracting Top 10 Restaurant/Eatery from 1 Neighbourhood in Brooklyn
- Extracting Top 10 Restaurant/Eatery from 1 Neighbourhood in Manhattan
- Extracting Top 10 Restaurant/Eatery from 1 Neighbourhood in Queens
- Extracting Top 10 Restaurant/Eatery from 1 Neighbourhood in Bronx
- Extracting Top 10 Restaurant/Eatery from 1 Neighbourhood in Staten Island
- Let’s put that into a pandas DataFrame and sort them in descending order

4. Clustering

Now that we have the sorted Data Frames from all 5 boroughs containing all the neighbourhoods, it is time we use clustering. We will be using k-means clustering. We would be using k-means to cluster all of our 5 boroughs:

- Brooklyn
- Manhattan
- Queens
- Bronx
- Staten Island

Why k-means?

K-Means can group data only unsupervised based on the similarity of customers to each other. There are various types of clustering algorithms such as partitioning, hierarchical or density-based clustering. K-Means is a type of partitioning clustering, that is, it divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. So we can say K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances. For example you may use

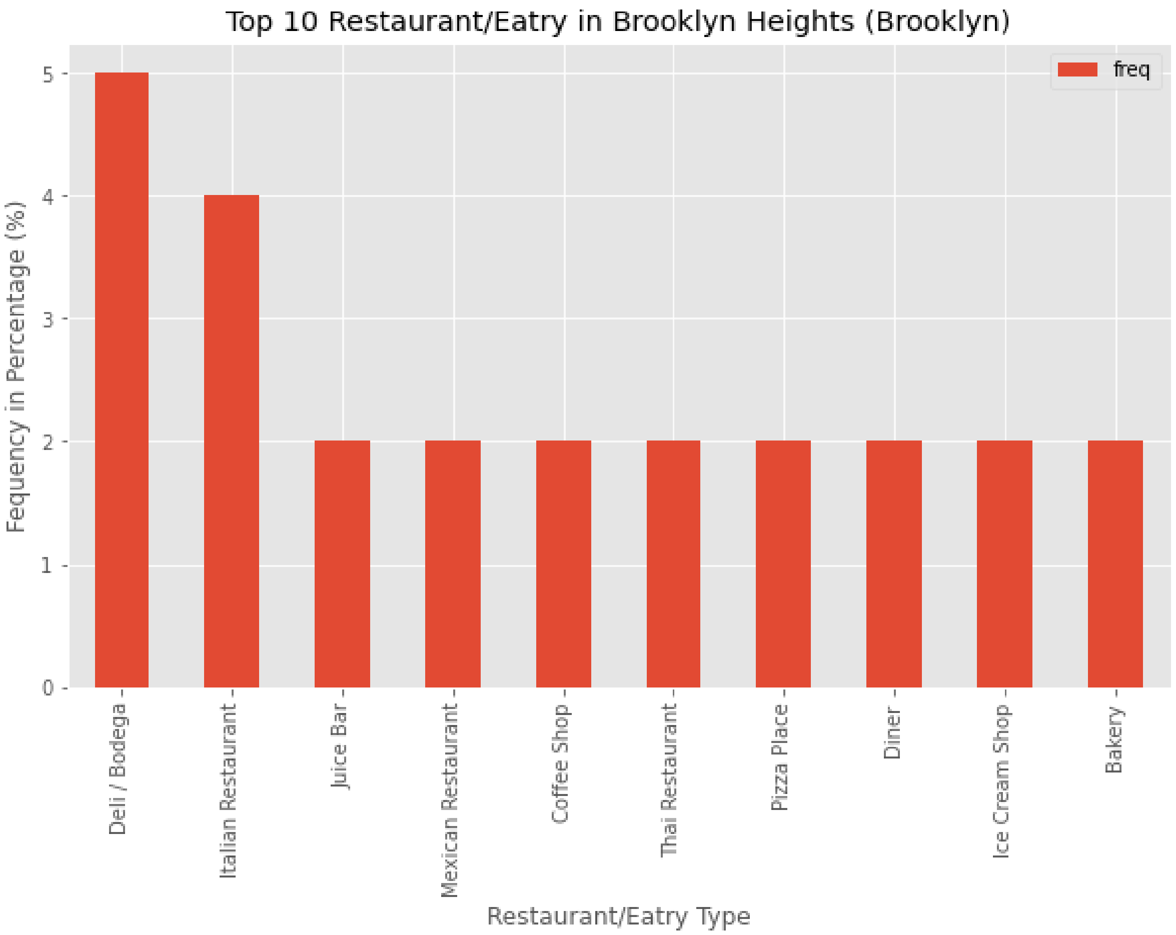
Euclidean distance, Cosine similarity, Average distance, and so on. Indeed, the similarity measure highly controls how the clusters are formed, so it is recommended to understand the domain knowledge of your dataset and datatype of features and then choose the meaningful distance measurement.

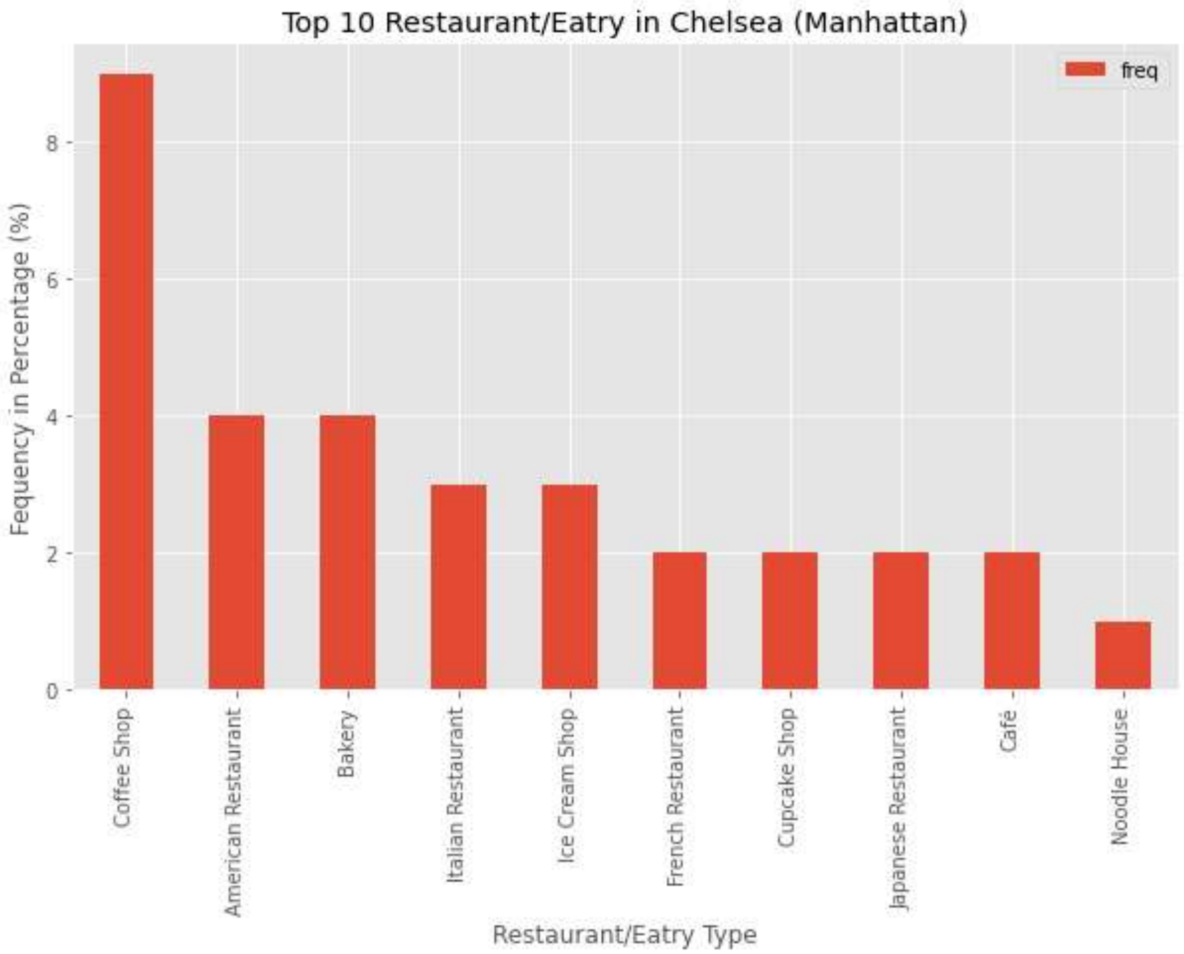
- Create DataFrame that includes the clusters as well as top 10 venues for each neighbourhood.
- Repeat the above step for all our 5 boroughs
- Visualize the resulting clusters

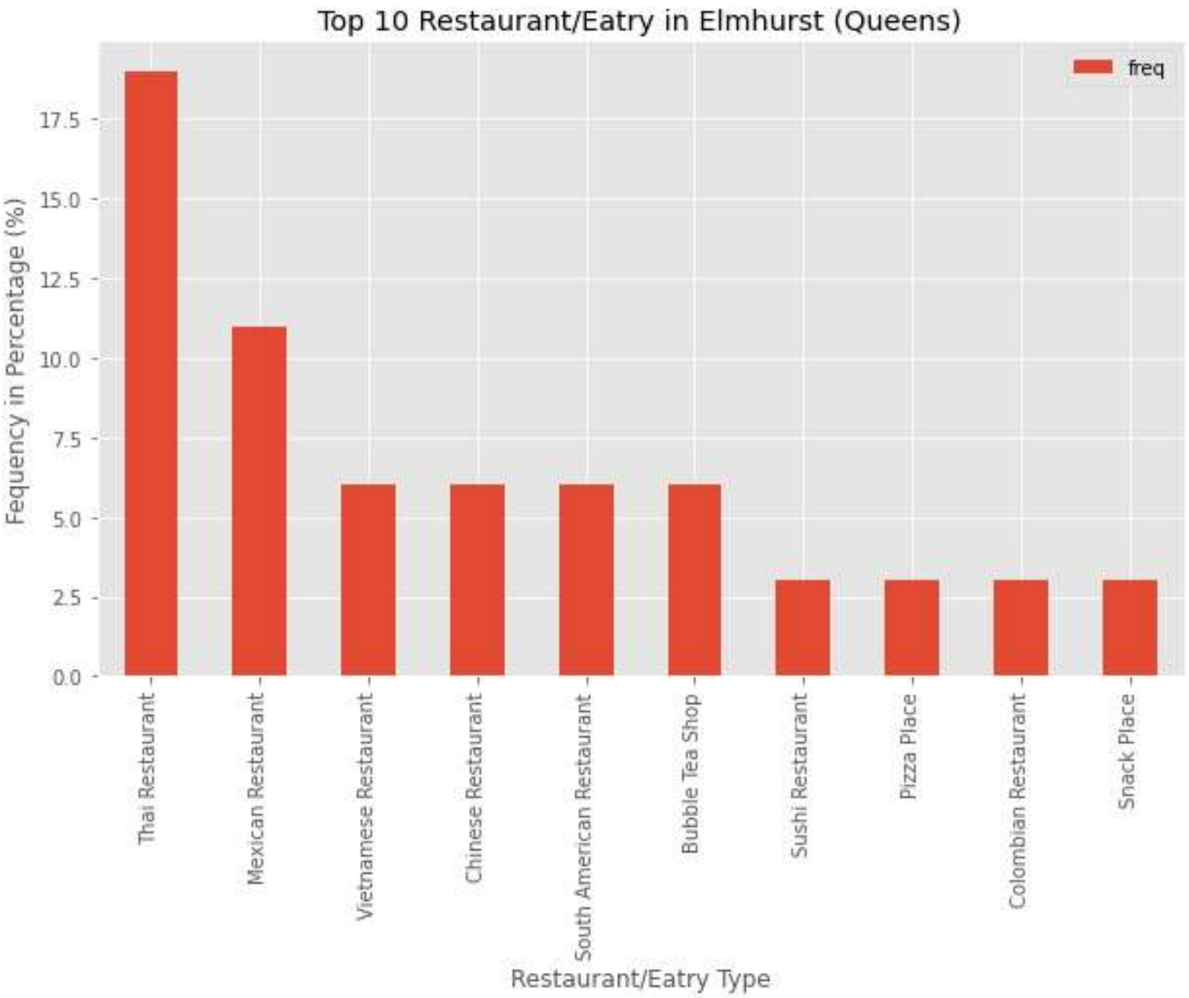
5. Examine clusters

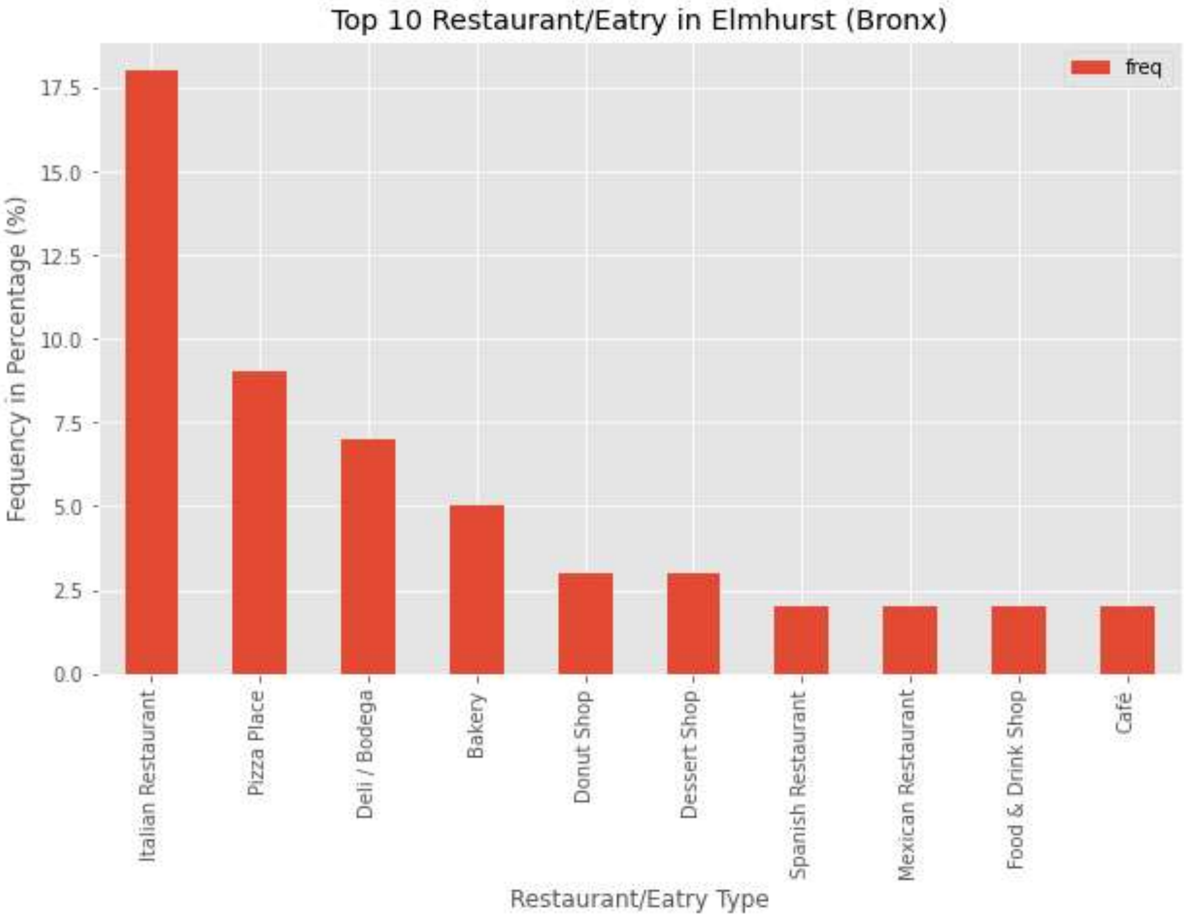
Examine each cluster in each borough and determine the discriminating venue categories that distinguish each cluster

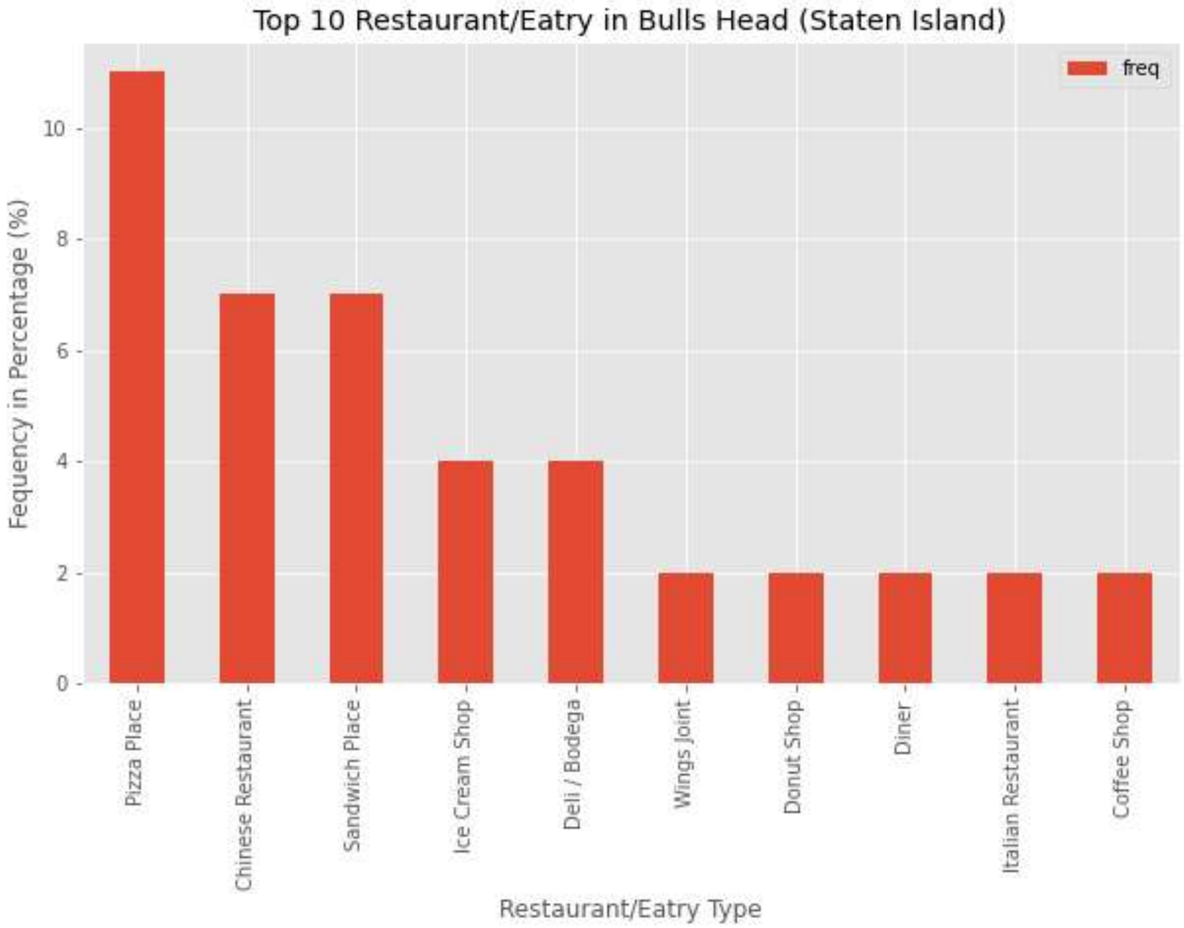
6. Results


Top 10 Restaurant/Eatry in Brooklyn Heights (Brooklyn)
freq
Fequency in Percentage (%)
0
1
2
3
4
5
Deli / Bodega
Italian Restaurant
Juice Bar
Mexican Restaurant
Coffee Shop
Thai Restaurant
Pizza Place
Diner
Ice Cream Shop
Bakery
Restaurant/Eatry Type


Top 10 Restaurant/Eatry in Chelsea (Manhattan)
freq
Fequency in Percentage (%)
0
2
4
6
8
Coffee Shop
American Restaurant
Bakery
Italian Restaurant
Ice Cream Shop
French Restaurant
Cupcake Shop
Japanese Restaurant
Café
Noodle House
Restaurant/Eatry Type


Top 10 Restaurant/Eatry in Elmhurst (Queens)
freq
Fequency in Percentage (%)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Thai Restaurant
Mexican Restaurant
Vietnamese Restaurant
Chinese Restaurant
South American Restaurant
Bubble Tea Shop
Sushi Restaurant
Pizza Place
Colombian Restaurant
Snack Place
Restaurant/Eatry Type


Top 10 Restaurant/Eatry in Elmhurst (Bronx)
freq
Fequency in Percentage (%)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Italian Restaurant
Pizza Place
Deli / Bodega
Bakery
Donut Shop
Dessert Shop
Spanish Restaurant
Mexican Restaurant
Food & Drink Shop
Café
Restaurant/Eatry Type


Top 10 Restaurant/Eatry in Bulls Head (Staten Island)
freq
Fequency in Percentage (%)
0
2
4
6
8
10
Pizza Place
Chinese Restaurant
Sandwich Place
Ice Cream Shop
Deli / Bodega
Wings Joint
Donut Shop
Diner
Italian Restaurant
Coffee Shop
Restaurant/Eatry Type