Rohan Srivastava

Applied Data Science Capstone Project

Battle of the Neighbourhoods

Contents

[Introduction 2](#_Toc47714432)

[Data 2](#_Toc47714433)

[Methodology 2](#_Toc47714434)

[Results 3](#_Toc47714435)

[Discussion 4](#_Toc47714436)

[Conclusion 4](#_Toc47714437)

# Introduction

Truffles Ltd. is a chain of fast-food restaurants, and is planning to open a new branch to cater to the increase in demand. Since COVID-19 has affected the footfall of the restaurant, they want to focus more on home deliveries for the time being and are looking to open the new restaurant in a heavily populated area to compensate for the lack of footfall with an increase in their home deliveries. They primarily operate in New York and are looking to open a new branch in New York itself. This project will help Truffles choose an ideal neighbourhood to set up their restaurants in. Also, it will help other restaurants identify the extent of competition in each neighbourhood which will help them decide on whether or not to expand their current business, and in case they choose to expand their business, which neighbourhood will be the most ideal.

# Data

Census Data of New York will be used to plot areas with different population densities. The census data will also be used to identify any other factors such as education and age to check whether these factors have a major impact on the demand for fast-food restaurants. With the help of Foursquare APIs, the number of fast-food restaurants in each neighbourhood will be identified to understand the extent of competition in the market. The neighbourhoods of New York will be divided into clusters and each cluster will show different venues that can act as alternatives to fast-food restaurants, such as other fancy restaurants or movie theatres.

# Methodology

Before acquiring and exploring the data, the required dependencies were downloaded.

|  |  |
| --- | --- |
| Sl.No | Dependency |
| 1. | Numpy |
| 2. | Pandas |
| 3. | Json |
| 4. | Geocoders |
| 5. | Matplotlib |
| 6. | Folium |
| 7. | Kmeans |

Once the dependencies and libraries were downloaded and installed, the data set was downloaded from the internet. Using the **‘wget’** command the data was accessed and was then loaded with the help of Json. Upon exploring the initial data, it was found that the relevant data, i.e. the co-ordinates of various neighbourhoods based in New York was found in the **‘features’** key.

This data was defined with the help of a new variable which was then converted into an empty data frame with the help of pandas library. This empty data frame was then filled with the required data by running a loop, wherein data was filled one row at a time.

Once the data frame was completed, the co-ordinates of New York were obtained with the help of the geo-locator. With the co-ordinates, a map of New York was created with the help of **‘Folium.’** On the created map, the various neighbourhoods of New York were superimposed on top to highlight the different neighbourhoods and to get an idea of how the city is divided amongst its various neighbourhoods.

With the help of Foursquare APIs, we could identify all the elements such as, Movie Theatres, Fast Food Restaurants, etc. which were offered by each neighbourhood. In order to identify these elements or venues, we defined a function which returned all the required information about each venue such as its location, footfall, etc.

A new data frame was created to accommodate this newfound information. Further exploratory analysis was conducted on this new data frame. All unique values in this data frame were identified. Once this was completed, each neighbourhood was again analysed on an individual level. Through this analysis on an individual level we were able to identify the mean of the frequency of the occurrence of each venue in each neighbourhood. Using this, we were able to pin point the top 5 most common venues in each neighbourhood.

In order to better understand the data, we defined a function to arrange the venues in descending order. A new data frame was again created to display the top 10 venues of each neighbourhood. Now that the required data frame was created, we ran the clustering algorithm of K-means to group venues into 5 clusters. These clusters were then plotted onto a map to be able to identify them easily.

Once the clusters were plotted on a map, each of these clusters were then individually analysed to see what each cluster offered and the footfall in various venues of each cluster.

# Results

Based on the clusters that were obtained, we can see the following results in each cluster. An abstract of each cluster is attached in the report. The full results can be viewed on the Jupyter notebook pushed to the GitHub repository.

**Cluster 1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
| 1. | Inwood | Mexican Restaurant | Café | Restaurant |
| 2. | Hamilton Heights | Pizza Place | Coffee Shop | Café |
| 3. | Manhattanville | Coffee Shop | Chinese Restaurant | Italian Restaurant |

**Cluster 2:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
| 1. | Chinatown | Chinese Restaurant | Cocktail Bar | Dessert Shop |
| 2. | Hamilton Heights | Café | Bakery | Mobile Phone Shop |
| 3. | Manhattanville | African Restaurant | Chinese Restaurant | Cosmetics Shop |

**Cluster 3:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
| 1. | Upper East Side | Italian Restaurant | Coffee Shop | Bakery |
| 2. | Lenox Hill | Sushi Restaurant | Italian Restaurant | Coffee Shop |
| 3. | Midtown | Hotel | Clothing Store | Coffee Shop |

**Cluster 4:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
| 1. | Stuyvesant Town | Park | Bar | Baseball Field |

**Cluster 5:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
| 1. | Marble Hill | Sandwich Place | Coffee Shop | Gym |

# Discussion

Based on the results, we can conclude that in 4 out of the 5 clusters, the most common venue is a restaurant. This shows that there is a huge demand for food in these areas as they are densely populated. Truffles can decide to open their new restaurant in one of these 4 clusters to cater to the neighbourhoods in the chosen cluster. Since there will be a lot of competition from existing restaurants, Truffles will have to compete on the basis of price in order to survive in the market. The only way they can compensate for low prices, is if they choose to serve a very high volume of customers. Clusters 2 and 3 cover the highest number of neighbourhoods amongst all 5 clusters. In order to increase the volume of the customers served, Truffle can choose to open their new restaurant in one of these clusters. They can choose to collaborate with cafes to offer coffee and other products such as bagels and donuts, which cater to the lifestyles of New Yorkers, and this will also help them capture market share from existing restaurants.

# Conclusion

Based on the findings, Truffles can choose to set up their restaurants in either Cluster 2 or Cluster 3. In order to enter the new market, they’ll have to collaborate with one of the most visited venues or imitate their style to absorb market share. As the footfall in restaurants is quite limited because of COVID-19, they can afford to spend more of their budget on better home delivery service rather than focusing on in-store ambience.