**Capstone Project - The Battle of the Neighborhoods - New York City**

Vartika Tyagi

May 31, 2020

1. **Introduction**

**Business problem**

In this project I have tried to find an optimal location for a restaurant. Particularly, this report is targeted to stakeholders interested in opening an **Italian restaurant** in **New York City**, USA. As there are many restaurants in New York City, tried to detect **locations that are not already crowded with restaurants**. I was specially interested in **areas with no Italian restaurants in vicinity** and preferred locations **as close to city center as possible**, assuming that first two conditions are met, also, used my data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area is clearly expressed and hence the best possible final location can be chosen by stakeholders.

1. **Data and its Sources**

Based on definition of the problem, factors that influenced the decision are:

* number of existing restaurants in the neighborhood (any type of restaurant)
* number of and distance to Italian restaurants in the neighborhood, if any
* distance of neighborhood from city center

In order to define the neighborhoods, a regularly spaced grid of locations, centered around city center is used.

Following are the data sources that were needed to extract the required information:

* centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using Google Maps API reverse geocoding
* number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API
* coordinate of New York City center will be obtained using Google Maps API geocoding

## **Methodology**

In this project I have directed my efforts on detecting areas of New York City that have low restaurant density, particularly those with low number of Italian restaurants, also, limited the analysis to area ~5km around city center.

Initially, collected the required **data: location and category of every restaurant within 5km from city center and identified Italian restaurants**as per the Foursquare categorization.

Secondly, the calculation and exploration of **restaurant density** across different areas of New York city was done. **Heatmaps** are used to identify a few promising areas close to center with low number of restaurants in general and also, no Italian restaurants in the vicinity and focused the attention on those areas.

In the final step, focused on most promising areas and within those, created **clusters of locations that meet some basic requirements** established in discussion with stakeholders. Considered locations with **no more than two restaurants in radius of 250 meters,** and wanted locations **without Italian restaurants in radius of 400 meters.** Presented the map of all such locations but also created clusters using **k-means clustering,** of those locations to identify general zones, neighborhoods, addresses which are the starting point for final street level exploration and search for optimal venue location by stakeholders.

**Exploratory data analysis: K-means**

Clustering refers to the splitting of the data into a set of groups based on the underlying characteristics or patterns in the data. One of the popular clustering algorithms is known as the **‘k-means clustering’**, which splits the data into a set of clusters based on the distances between each data point and the center location of each cluster.

In this project, when a clear indication of zones with low number of restaurants in vicinity and no Italian restaurants nearby is obtained, I **clustered** those locations to create **centers of zones containing good locations**. These zones, their centers and addresses are the final result of the analysis.

For this, firstly I set the number of clusters and fitted the K-means by assigning the random state, then the cluster centers were assigned using the longitude and latitude.

Finally, the clusters represented groupings of most of the candidate locations and cluster centers were placed nicely in the middle of the zones 'rich' with location candidates. Addresses of those cluster centers were a good starting point for exploring the neighborhoods to find the best possible location based on neighborhood specifics. Then finally these zones on a city map were shown without heatmap using shaded areas to indicate these clusters.

## **Results and Discussion**

The analysis showed that although there was a great number of restaurants in New York City, there were packets of low restaurant density fairly close to city center. Highest concentration of restaurants was detected south and north-west from New York City, so I focused my attention to areas south of South Street( Manhattan) and north-west of Brooklyn Street circuit.

I firstly created a dense grid of location candidates (spaced 100m appart) those locations were then filtered so that those with more than two restaurants in radius of 250m and those with an Italian restaurant closer than 400m were removed.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is 15 zones containing largest number of potential new restaurant locations based on number of and distance to existing venues - both restaurants in general and Italian restaurants particularly. This, of course, does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to Berlin center but not crowded with existing restaurants (particularly Italian) - it is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

## **Conclusion**

The purpose of this project was to identify New York City areas close to center with low number of restaurants and particularly the Italian restaurants in order to aid stakeholders in narrowing down the search for optimal location for a new Italian restaurant. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis (South Street( Manhattan) and Brooklyn Street circuit) and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest which contained greatest number of potential location and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location, levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.