Data Science Report On Ideal Locations To Establish a Restaurant within The City of New York

Steve Tambo

February 9th, 2021

1. Introduction

1.1 Background

New York City is the most populous city in the United States and the largest metropolitan area in the world by urban landmass with almost 20 million people in it's metropolitan statistical area. It has been described as the cultural, financial, and media capital of the world, significantly influencing entertainment, research, technology, education, politics, tourism, art, fashion, and sports. All of this provides for great opportunity as far as business ventures go especially for restauranteurs with New York being the country's leading restaurant city. Given the diverse nature of its residents, one must be critical in deciding where to locate their business so they can maximize on profits. This can be done through carefully analyzing the distribution of existing venues within the city and clustering them into distinct sections which can be very useful for new companies to make their mark in the land of opportunity.

1.2 Problem Definition

Piacci's is an Italian owned restaurant franchise that has proven successful over the years since its inception. The establishment was formed in Toronto, Canada in the year 2009 and they specialize in gourmet Italian delicacies. Given their current success they wish to expand to other areas in America more specifically within New York.

Data about different venues in New York can be accessed via the FourSquare API. This data shows the names , locations and ratings of these venues as assessed by customers who have already visited them. Having spoken to numerous opportunity assessors they have been advised to seek out the services of a data scientist to access and analyze this data so they may determine what would be the ideal locations to set up new restaurants within the city through the use of clustering algorithms.

2. Data Acquisition and Preparation

2.1 Data Sources

To begin the analysis a data set of all the boroughs and neighborhoods in New York showing their geographic locations in terms of latitude and longitude was needed. The dataset on New York provided in the skills lab was the one I used throughout the analysis.

The second dataset was obtained by connecting to the FourSquare API and downloading data on all the venues around New York from their servers.

2.2 Data Preparation

The first dataset containing information about New York was in JSON format . I had to use the pandas library on Jupyter notebook to open this JSON file and extract the relevant information and place it in a pandas dataframe. The attributes I was looking for were the name of each borough, the names of the neighborhoods within that borough as well as their geographic locations. All of this was stored within the features key of the 'newyork_data' dictionary. After looping through and extracting the data there were 5 boroughs and 306 neighborhoods identified in total. This was all stored in a data frame named 'neighborhoods'.

After preparing the New York 'neighborhoods' dataset it was time to connect to the FourSquare API and get the information about venues. First an API request was made by sending a url containing my client ID and secret as well as the location for the venues I was interested in . I sent a get request to pull all the relevant venue data from the server and used a predefined function to loop through the venues data and store the relevant features in a new pandas dataframe named 'newyork_venues'. The attributes I was interested in were the neighborhood name , latitude and longitude. In relation to the venue I was interested in the name, latitude , longitude and category of each venue within a 500m radius of each neighborhood. There were 10034 venues in total divided across 429 unique venue categories.