Report :

The Battle of the Neighbourhoods

Introduction & Business Problem :

Background:

New York, the financial capital of USA, is the most populous city in the United States. It has become a global hub of business and commerce by attracting many different players into the market by providing business opportunities and business friendly environment. The market in New York city is highly competitive, thus, any new business venture or expansion needs to be analysed carefully. The insights derived from the current analysis will give good understanding of business environment which help in strategically targeting the market.

A restaurant or an eatery, is a business that prepares and serves food and drinks to customers in return for money. Restaurants vary greatly in appearance and offerings, including a wide variety of cuisines and service models ranging from inexpensive fast food restaurants and cafeterias, to mid-priced family restaurants, to high-priced luxury establishments. The City of New York is famous for its excellent cuisine. It's food culture includes an array of international cuisines influenced by the city's immigrant history. So it is evident that to survive in such competitive market it is very important to plan strategically. Various factors need to be studied in order to decide on the Location such as, Population, Cuisine served / Menu of the competitors and markets, etc,.

Objective:

The objective is to Identifying the best neighbourhood of New York city to start a restaurant business. This would interest to the people looking to start a new restaurant in New York city.

Data Sources:

- Latitude and longitude coordinates: https://geo.nyu.edu/catalog/nyu_2451_34572

- Farmers Markets: https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets/8vwk-6iz2
- GrowNYC's Fresh Food Box Program data: https://www.grownyc.org/freshfoodbox
- New York Population: https://en.wikipedia.org/wiki/New_York_City
- New York Economy: https://en.wikipedia.org/wiki/Economy_of_New_York_City
- New York Portal: https://en.wikipedia.org/wiki/Portal:New_York_City

Data Preparation

Data Tools used in the analysis are listed below:

- IBM Watson® Studio: to prepare data and build models at scale for this project
- Pandas: To use for data manipulation and analysis
- BeautifulSoup: to create parse trees that is helpful to extract the data
- GitHub: to share project notebooks using Git repository hosting service
- Geopy: to get geological location by address name
- folium: to visualize the distribution pattern
- Foursquare API: to get the most common venues of given Borough of New York City.
- Foursquare API: to get the venues' record of given venues of New York City.

Data Science Workflow

- **Outline the initial data that is required:**
  - Boroughs data for New York including names, location data if available, and any other details required.
- **Obtain the Data:**
  - Research and find suitable sources for the district data for New York.
  - Access and explore the data to determine if it can be manipulated for our purposes.
- **Initial Data Wrangling and Cleaning:**
  - Clean the data and convert to a useable form as a dataframe.

Data Analysis and Location Data:

- Foursquare location data will be leveraged to explore or compare boroughs around New York.
- Data manipulation and analysis to derive subsets of the initial data.

- **Visualization:**
  - Analysis and plotting visualizations.
  - Data visualization using various mapping libraries.
- **Discussion and Conclusions:**
  - Recommendations and results based on the data analysis.
  - Discussion of any limitations and how the results can be used, and any conclusions that can be drawn.

Methodology

Exploratory Data Analysis

New York 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. In this project first part is clustering of Manhattan and Brooklyn and second part is clustering of Bronx, Queens and Staten Island.

New York city Geographical Coordinates

In this, I have loaded the data from newyork_data.json file. and transformed the data of nested python dictionaries into pandas dataframe. This dataframe contains the geographical coordinates of New York city neighbourhoods which was used to get Venues data from Fouresquare. Map of New York city with neighbourhoods superimposed on top was created using geopy and folium libraries.

New York Population

Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

Web scrapping of Population data from wikipedia page using BeautifulSoup was done and extracted the population details of New York.

Segmenting and Clustering Neighbourhoods - Brooklyn and Manhattan

New York city geographical coordinates data has be utilized as input for the Foursquare API, that has been leveraged to provision venues information for each neighbourhood. We used the Foursquare API data to explore neighbourhoods in New York City. Later the list of restaurants will be extracted and enlisted.