Hunting Best Neighborhood for Living in NYC

IBM Applied Data Science Capstone

By: Surya Gurung

Mar 2020

1. Background

New York City (NYC) is the most populous city in the United States (US) with over 8 million population spread over five boroughs, Brooklyn, Queens, Manhattan, Bronx, and Staten Island. It is well known as global capital of finance, media, and immigrant. NYC is also a global leader in entertainment, fashion, tourism, technology, education, arts, sports, politics, research, and many more industries. The World's largest stock exchanges operate from the famous financial center Wall Street. Many of the World's largest financial and media companies are based in NYC.

Time Square, NYC

In brief, NYC is center of opportunities to make the American dream come true. There is no other place like New York to start your living to pursue your American dream no matter where you come from. People around the World migrate here to raise and build better future for their family. Today, you can find the communities representing whole world, in fact more than 800 languages are spoken in NYC. But being the most populous and diverse city in US, it can be very intimidating to find right place to start living in NYC. It may seem easy to find the best place for your family's new life but people might easily get lost in the process and end up in wrong place. It requires careful and thorough research and decision to find the best place to live with happiness.

2. Business Problem

So, the problem every new city dweller or migrant faces is how to find the best neighborhood to rent a new apartment or to own a new house to start a happy and successful life in NYC? The main objective of this capstone is to develop a k-means clustering model using data science methodology and visualize the clusters in the NYC map by using NYC neighborhood datasets along with Foursquare API to help the new city dwellers to select best possible NYC

neighborhood quickly with greater precision. So, this project will greatly simplify the tedious process of hunting the new apartment or house in NYC.

3. Target Audience

Obviously, the new apartment hunters and home owners would be very interested to find the best neighborhoods in NYC to own a home or rent an apartment to start their NYC life successfully. Beside the new city dwellers, the real estate investors, financial institutions, and business investors are also going to be interested to find the best neighborhoods to own the real estate properties.

4. Data Requirements & Sources

To build the machine learning models, first we need the datasets representative of the business problem. For this project, we are going to use following datasets:

- We need dataset containing the neighborhoods name and its geographic locations. New York City neighborhood dataset contains list of neighborhood names and their geographic coordinates in five boroughs, Brooklyn, Bronx, Manhattan, Queens, and Staten Island. This dataset defines the scope of this project which is the neighborhoods in five boroughs of NYC.
- The Foursquare venue dataset doesn't contain the NYC schools information. Because the school information is one of the very important features when searching for apartment or house, I am going to use following NYC schools data sources:
    - NYC elementary school dataset contains list of elementary schools in five boroughs.
    - The middle school dataset contains list of middle schools in NYC.
    - The high school dataset contains the list of high schools in the NYC five boroughs.
    - The performance of middle and elementary schools and
    - The high schools performance .

- The Foursquare venue dataset doesn't contain the crime data but it is critical information when making decision for living in new neighborhood. The NYC crime dataset contains all valid felony, misdemeanor, and violation crimes reported the New York City Police Department in 2019 and the geographic location of the crimes. This dataset doesn't contain the neighborhood name. So, I am going to use Census zip code tabulation file and NYC neighborhood zip codes to add neighborhood names to the NYC crime dataset based on the crime's geographic location.

- We are also going to use the venue data returned by the Foursquare API based on the NYC neighborhoods dataset. This data will be used to perform clustering of the NYC neighborhoods based on the most important venues for living and to raise kids. By clustering the neighborhoods based on the most important venues, it will help to make the neighborhood selection process based on the criterions of individual dweller.

5. Data Collection & Preprocessing:

All of the datasets are retrieved using either the requests python library or pandas read_csv function. The important features are extracted and converted into a pandas dataframe. The NYC neighborhoods dataset is the main dataframe and all other dataframe are going to merge to it. For the consistency, the feature names are going to be lower case with underscore for space.

Four features, *borough, latitude, longitude, and neighborhood* are extracted from the NYC neighborhood dataset and converted into a pandas dataframe. There are 299 rows each representing one neighborhood in one of the five boroughs of NYC. As a note, four neighborhoods in different boroughs share same name and those are Sunnyside, Bay Terrace, Murray Hill, and Chelsea. The latitude and longitude values are going to be used as inputs for Foursquare API to retrieve the venues data around the neighborhoods.

Since Foursquare API doesn't return NYC schools dataset, I am going to create a dataframe by extracting the school number, school name, neighborhood, and students' achievement ratings from the five NYC schools data sources and then merging them together. After merging five dataframe, the school dataframe is going to have 1471 rows with four features, *dbn, school_name, student_rating, and neighborhood*. Then this dataframe is merged with the NYC neighborhood dataset.

Similarly, the NYC crime dataset is retrieved as json file and converted into a dataframe with *crime_id, borough, crime_level, crime_desc, latitude, and longitude* as features. To get the neighborhood name of the crimes, I am retrieving NYC neighborhood zip codes with its' geographic location by merging the US census zip code table with NYC neighborhood zip codes. By using geopy's distance function, I calculated the distance between crime locations and the NYC neighborhood geographic locations to find out the neighborhood names of the crimes. Similarly, this dataset is also merged with the NYC neighborhood dataset.

So far, we have added the school and crime data to the NYC neighborhood dataset. The final piece of data we need add to the main dataset is the Foursquare venue dataset. We are going to use the Foursquare API to get the venues around 500 meters from each neighborhood geographic locations. This API returns many different venues from 10 major categories but we are not going to need all kind of venue data. I am going to selects the venues from the *Outdoors*

& Recreation, Professional & Other Places, Shop & Service, Food, and Arts & Entertainment categories.