The Battle of Neighborhoods: Searching for Maximum Return for Rental Properties in New York City

A. Introduction

1. Background

New York City is the most populous city in the United States with a population of 8,253,213 and a density of 27,274.3 people per square mile. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports, and is the most photographed city in the world. With a description like that, who wouldn't want to live there?

2. Business Problem

A real estate investment firm is looking to begin purchasing properties in New York City to be used as rental properties. However, they have had mixed luck with real estate agencies leading them to what they think are the best opportunities across the five boroughs. The investment firm now would like an analysis to indicate where the most desirable rental properties are located for the lowest price.

- As many unique venues within walking distance (because owning a car in NYC is too expensive).
- What does the crime rate look like by borough and neighborhood?
- Rental property median price.

Therefore, the project goal is to figure out the best locations to rent in New York City to maximize the return on investment to have access to the most unique venue types within walking distance.

3. Target Audience

This study is of interest to **everyone** who may have a curiosity in relocating to or moving within New York City to maximize their return for rent paid. The project will also be of interest to **business owners** seeking to locate their business in an area of little competition or high diversity. **Real estate investors** will additionally be interested who may wonder how data science can provide insight into neighborhoods that will maximize their investment.

B. Data Description

1. Data Requirements and Collection

For this project we will need historical rental prices for properties in New York and historical crime data for incidents in each of the boroughs. We can also leverage Foursquare Location data to compare neighborhoods with venue locations and their respective ratings. The following are data sources that were used for this project:

- **Zillow Observed Rent Index (ZORI):** The most up-to-date median rental prices for all U.S. cities segmented by zip code.
- **New York City Borough and Neighborhoods:** JSON data containing the 5 boroughs and 306 neighborhoods classified as New York City with latitude and longitude coordinates.

- **NYPD Arrests and Crime Data:** Combined data of arrests, shootings, and complaints with geo locations retrieved from NYC OpenData. Each crime file was cleaned and merged into a singular crime database file.
- **Foursquare Venue Data:** The most popular venues of a given neighborhood in New York City. This information is stored inside **Foursquare Location Data**, and we will use the **Foursquare API** to access it.

2. Data Cleaning and Extraction

- The first dataset is a CSV file retrieved from Zillow containing 1743 rows and 94 columns. The data is collected each month with the median rental price separated by zip code and month for the entire United States. We will focus on the rows containing only zip codes within New York City.
- The second dataset is a JSON file containing geolocation coordinates matching all of the zip codes to the corresponding neighborhoods and boroughs.
- The third datasets are three separate CSV files retrieved from NYC OpenData containing all crimes within New York City. They will be combined into a singular file and grouped using geolocations.
- The fourth dataset is stored within Foursquare Location Data and will be accessed through the API. We will utilize postal coordinates to retrieve venues, categories, and their ratings. We will then use this data to cluster the unique categories to each of the neighborhoods to find the best mix with minimal distance between them.