# Analysis of Tourist Accommodation in Manhattan, New York

## Table of Contents



# 1. Introduction

## 1.1 Background

New York City, NYC, is the most populated city in the USA and largest metropolitan area in the world. NYC has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. Many districts and landmarks in New York City are well known, including three of the world's ten most visited tourist attractions in 2013; a record 62.8 million tourists visited in 2017. Several sources have ranked New York the most photographed city in the world.

Manhattan is the most densely populated of the five boroughs of New York City and serves as the city's economic and administrative center. Manhattan has been described as the cultural, financial, media, and entertainment capital of the world. Manhattan hosts three of the world's 10 most-visited tourist attractions in 2013: Times Square, Central Park, and Grand Central Terminal. As one of world’s greatest tourists’ destination, tourism is vital to Manhattan's economy, and the landmarks of Manhattan are the focus of New York City's tourists. As of June 2016, Manhattan had nearly 91,500 hotel rooms, a 26% increase from 2010.

Airbnb, Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. Airbnb provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities. Guests can search for lodging using filters such as lodging type, dates, location, and price. Guests have the ability to search for specific types of homes, such as bed and breakfasts, unique homes, and vacation homes.

It would help to guide a visitor looking for an accommodation and a AirBnB property owner to set right price if we can characterize or cluster the Manhattan neighborhoods in terms of

1. AirBnB properties density and prices
2. Places of interests within the neighborhood

## 1.2 Problem Statement

With this exercise, we shall try to address below problems.
• Is there any co-relation between places of interests and AirBnB property values? If yes, can we describe it.
• How to choose the most suitable and convenient location as per the purpose of visit?
• Can we characterize the Manhattan neighborhoods in terms of the places of interests and AirBnB properties density and prices?
• Can we come up with any guidelines for setting the right price for a property while being listed on AirBnB?

# 2 Data

## 2.1 Data Sources

• **Manhattan, NY Geo Data**
This dataset holds longitude and latitude co-ordinates for all the neighborhoods in Manhattan, NY. Most of the data is available in a json file at https://cocl.us/new_york_dataset (https://cocl.us/new_york_dataset) , though it is missing few neighborhoods which are part of our AirBnB dataset. For the missing ones, I used geocoder api to get the co-ordinates. As we get the data, I have created a csv file, NY_MH_NEIGHBORHOODS.csv , which can be used for subsequent executions. This saves execution time by removing the outgoing network calls for geocoder api, and json file and parsing of json file for required data.

• **Manhattan, NY Four Square Data**
This dataset holds data related to venues of interests in neighborhoods. These venues can be business or financial centers, exhibitions, recreational places, transit facilities, recreational places, sports, restaurants, fastfoods, gourmet joints, and so on. Again, I created a file, NY_FOURSQUARE_DATA.csv, which holds collected data for subsequent executions. This helps to reduce redundant calls to Four Square API as these calls and processing involved is time consuming and costly. I shall use a reference _value*cat.csv* (https://github.com/K2Prasanna/Coursera_Capstone/blob/master/resources/venue_cat_ny.csv (https://github.com/K2Prasanna/Coursera_Capstone/blob/master/resources/venue_cat_ny.csv)) to map Four Square Venue categories to fewer borader categories.

Features to be used for FourSquare API response for each venue are:
Neighboorhood Name
Venue Name
Venue Category
Venue Location (Longitude and Latitude)
Classification: Broader classification for the venue category.

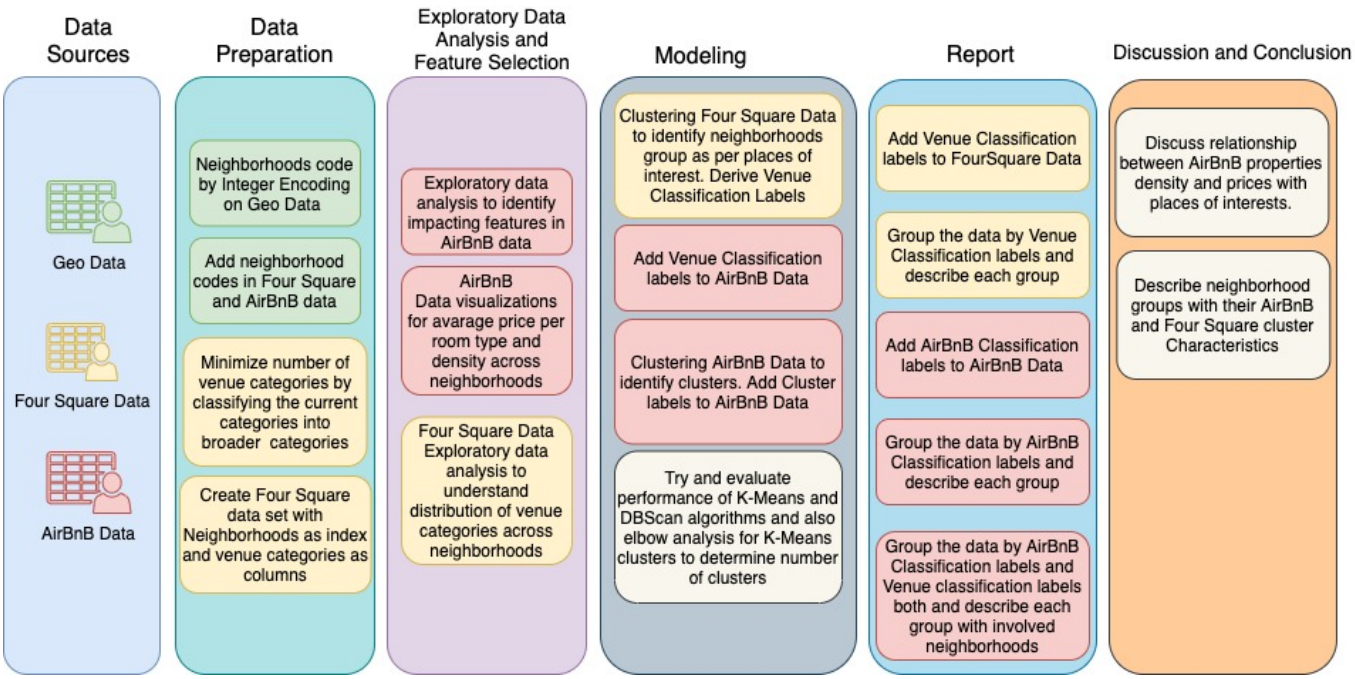
• **Manhattan, NY AirBnB Data**

There is a dataset, AB_NYC_2019.csv, available on Kaggle which has AirBnb properties data for NY. Link to the dataset is https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/downloads/new-york-city-airbnb-open-data.zip/3 (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/downloads/new-york-city-airbnb-open-data.zip/3). This dataset contains listing of AirBnB properties with their below features:

Owner Name
Borough
Neighborhood
Price per night
Minimum nights
Availability over an year
Number of reviews
Last review Date
Reviews per month
Calculated host listing count

## 2.2 Solution Approach

The section outlines how the data will be processed and modeled to achieve the required analysis. Below diagram depicts the stages involved and high level description of tasks to be performed in each step.



**Description**
I shall start with **Data preparation**, where I shall encode neighborhood names to integers and update all the datasets to use the numerical values for neighborhoods. I shall also use the geo data co-ordinates for FourSquare API to get venues in a neighborhoods. Then I shall work to minimize the number of venue categories as Four Square API will return 220+ categories. This almost nullifies the clustering as most of them will be having zero values for many of the neighborhoods. So we consolidate the categories in a broader category. For that I have created a reference as _value*cat.csv*. Then I transform the dataset so that all neighborhoods are as index and each category as columns, with number of venues in that category and neighborhood as values.

As part of **Exploratory data analysis**, I shall analyze the AirBnB dataset to understand what all the features have any relationship with the price and select those for further steps, and also around the preliminary analysis about the density and average prices across neighborhoods. I shall try to plot the average prices and count values for each neighborhoods to get any inference around their distribution. I shall perform similar analysis for venues/foursquare data.

In **Modelling**, I shall perform clustering first on venues data and come up with numerical cluster labels for each neighborhoods. The we use the venue cluster labels as additional data in AirBnB dataset and perform clustering. I shall try to compare the clustering algorithms and their suitability for the exercise in both cases.

As part of **Report** stage, I shall describe both types of clusters (Four Square Venues and AirBnB) independently and combined. I shall visualize the clusters over folium map of Manhattan. In **Discussion and Conclusion** section, I shall try to explain co-relation between places of interests and AirBnB property values and describe nighborhoods in terms these clusters. I shall discuss the observations from the exercise, and how the report can help for AirBnB users and tourists.

# 3. References

https://www.wikipedia.org/ (https://www.wikipedia.org/)
https://www.kaggle.com/ (https://www.kaggle.com/)
https://cocl.us/new_york_dataset (https://cocl.us/new_york_dataset)