# INTRODUCTION

**New York** is the **most populous city** in the U.S. Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass. New York had approximately 65.2 million visitors, comprising 51.6 million domestic and 13.5 million international visitors in 2018. Thus, business opportunities are vast in the city and so is the competition. Many big players have been attracted to this city in all areas of business. As is it a highly developed city, the cost of doing the business is huge and risk is high. Thus, any new business venture or expansion needs to be analyzed and studied carefully. This will ensure good understanding of the business environment and strategical planning to reduce the risk factor and increase returns.

# BUSINESS PROBLEM

As a business and tourist hub, the city is famous for its cuisine as well. From street food to high end Michelin star restaurants, you will find everything in New York City. Various well-known chefs have their restaurants there. Eat this, not that placed New York City as one of the top cities for food.

In 2015, NYC started building public WIFI hubs in locations where pay phones used to exist. There are plans to have over 7500 WIFI hotspots installed. Hotspots are to have up to 1GBps speed internet.

The objective is to locate and determine if a restaurant’s proximity to a public WIFI hotspot has influence on the rating of that restaurant. The **target audience** interested in this problem is a group of small business owners who are interested in opening a restaurant in Manhattan.

# DATA ACQUISITION

This demonstration will make use of the following data sources:

*New York Wifi Venue Data Set (Kaggle)*

Data contains location of wifi hotspots with latitude and longitude values

*New York Top Venue Recommendations from FourSquare API*

(FourSquare website: www.foursquare.com)

I will be using the FourSquare API to explore venues located near NYC WIFI hotspots. The Foursquare explore function will be used to get venues located in the proximity of WIFI hotspots. Once these venues are obtained, I will be using the FourSquare API to get the rating of each venue. The resulting dataset will have the following columns:

* Venue ID
* Venue Name
* Coordinates : Latitude and Longitude
* Category Name
* Distance from WIFI point
* Rating of Venue

From these columns I will be making plots and showing stats on the relationship between distance from WIFI hotspot and rating of the venue.

# METHODOLOGY

NY WIFI hotspot data was downloaded from Kaggle. This data set contains NY city public WIFI hotspots located throughout multiple boroughs. Because the data set was very large, I narrowed it down to the Borough of Manhattan. The WIFI hotspot dataset contains many columns but the following are the most useful:

* type of hotspot
* borough of hotspot
* longitude of hotspot
* latitude of the hotspot

Once the list was narrowed down to these columns, and rows within the borough of Manhattan, I retrieved the nearest venues to these hotspot locations using the FourSquare API. Because WIFI hotspot locations are densely populated compared to neighborhoods data, I limited the results to 10 venues per WIFI hotspot.

After calling the FourSquare API per WIFI venue and concatenating the results to the original WIFI hotspot list, I saved the results to a .csv file. The reason for this is twofold:

1. I wanted to narrow down the results to just restaurants and food places and it was easier to do this in Excel than in Python
2. I wanted to use Excel to calculate the distance of the venue to the WIFI hotspot using a formula:

*=ACOS(COS(RADIANS(90-C2)) \*COS(RADIANS(90-G2)) +SIN(RADIANS(90-C2)) \*SIN(RADIANS(90-G2)) \*COS(RADIANS(D2-H2))) \*6371000*

Where

*C2 = Latitude of Wifi Location*

*D2 = Longitude of Wifi Location*

*G2 = Latitude of Venue Location*

*H2 = Longitude of Venue Location*

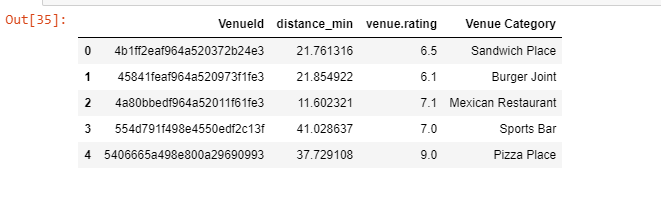
Below is what the resulting dataframe looks like:



The next step is to get the rating of each Venue located near a hotspot. This was achieved by creating a function that returns a dataframe consisting for Venue ID, Venue Name, and Venue Rating. Below is a sample of the resulting dataframe.

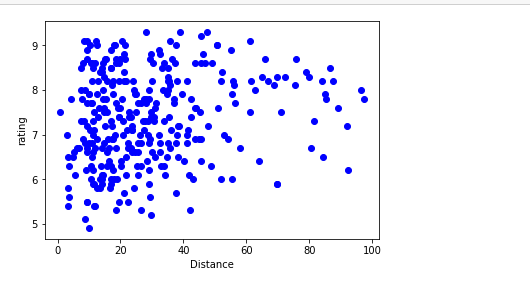


Next I merged the output of the ratings dataframe with the dataframe containing WIFI hotspot locations and nearby venues. Because two separate hotspots could be in the vicinity of a single venue, I de-duped the results by taking the hotspot that was closest to the venue. Below is a sample of the resulting dataframe:

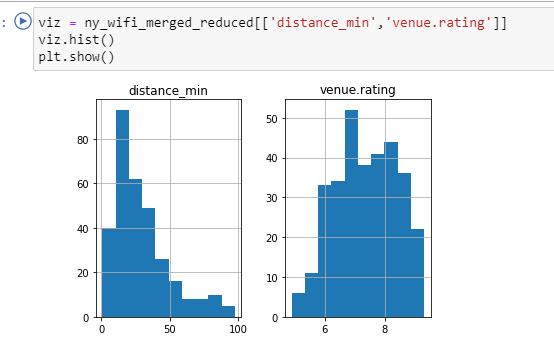


## Analysis and Modeling Approach

The first attempt was to show if there is any kind of relationship between rating and distance from public WIFI point. Below is a scatter plot of distance vs. rating:



The relationship between the two variables is not apparent from visual inspection. Also taking a histogram of venue ratings and distances from the nearest WIFI point yielded the following results:



Based on these results, I decided to change the approach to a cluster analysis.

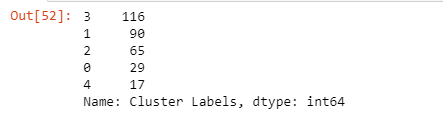
## Clustering

I decided to change the approach to this business problem by clustering the venues using venue category, rating, and distance. Recall that the objective

First, I encoded the venue categories by using the Pandas get dummies function to convert the categories into Boolean values. Each category becomes a dataframe column where a particular Venue ID is assigned a value of 1 if that Venue ID falls into category. Below is a sample of the resulting dataframe.

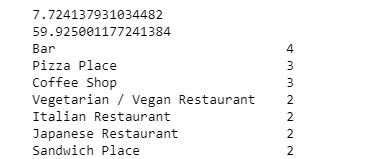


Then I added the ratings and distances back to the dataframe. I ran a K-means cluster analysis using a cluster size of 5. The result of the cluster analysis has the following amount of venues in each cluster.

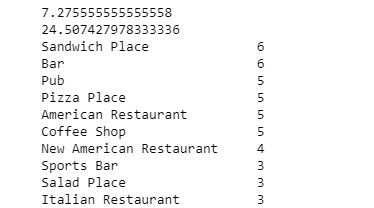


Next I examined each of the clusters looking at the mean rating, mean distance to venue and the number of venue categories in each cluster.

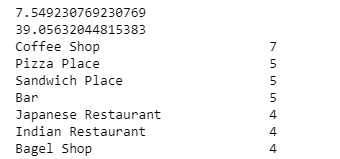
#### Cluster 1



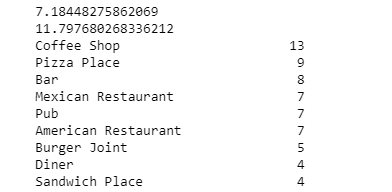
#### Cluster 2



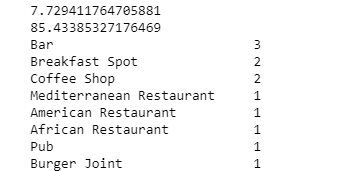
#### Cluster 3



#### Cluster 4



#### Cluster 5



# Discussion and Conclusion

On this notebook, analysis of the best location and type of restaurant to open in Manhattan has been presented. Here are some of the findings:

* Based on rating of restaurants, types of restaurants, and proximity of restaurant to a NYC WIFI hotspot, I determined what recommendation to make to business on type and location of restaurant.
* Using FourSquare API, I have collected a good amount of data regarding the restaurants located near a NYC WIFI hotspot in the area.
* The sourcing from FourSquare had its limitations though. There are many restaurant venues located near a NYC WIFI hotspot, but only a subset of these venues actually have a rating. Therefore the recommendation may have changed if I had access to more data.
* Based on the data available, and the generated clusters from the data, I recommend that the business stakeholders open a **Sandwich shop** located within **50m** of a NYC WIFI hotspot.