# Introduction

This project aims to determine the best location in New York, specifically the borough of Manhattan, for an aspiring restaurateur to open up her new Mexican restaurant. She wants to make sure that she is paying a fair amount for rent, has an ability to stand out from other restaurants in the area, and that there is actually a customer base that can make her business successful. The final result of this project is to determine a geographic grid that gives the highest chance of her business succeeding.

# Data

There are several data sets that will have to be used in order to make this project viable. The first is the location of the Mexican restaurants in Manhattan and their average ratings. This data will be pulled from Foursquare. This will allow the aspiring business owner to determine how much competition is in the area, and more importantly, the quality of the competition. This will allow her to know whether or not there is a realistic chance of building a customer base.

The second source of data needed is the average rent prices for commercial properties in different parts of Manhattan. This will provide some insight on what she can expect one of her main expenses to be, and that is valuable data to know for aspiring business owners. It is currently unknown where this data will come from, and will require some research on the internet to build a data set.

The last data set needed is the average age of residents in different parts of Manhattan. This is based on the hypothesis that the majority of people who eat out at restaurants are young adults. In order for a new restaurant to succeed there has to be a customer base, and customers are more likely to go to restaurants closer to them.

# Methodology

The initial data provided in the spreadsheet includes the population of each neighborhood, the number of adults under 30, the average rent for commercial spaces capable of being restaurants, and the latitude and longitude of the center of each neighborhood. The population data is pulled from arcgis.com (full link is provided in the notebook), the latitude and longitude of each neighborhood are manually entered from Google maps, and the rent is based on locations currently available in New York City found on loopnet.com.

Once the Excel sheet is read into the notebook the data is converted into a Pandas data frame. This makes the data easy to manipulate in Python, which is important for the rest of the project.

The first step is to create a column in the data frame for the percentage of adults under 30 in each neighborhood. This is done by dividing the number of under 30 adults by the total population. After the percentage column is created and filled then the number of adults under 30, and total population columns can be dropped. This nicely limits the amount of columns in the data frame and makes it easier to interpret the data.

The last part of data required for this project is the amount of Mexican restaurants in each New York City neighborhood. This is generated by using the Foursquare API. By iterating through each row in the data frame and using the latitude and longitude of the center of each neighborhood the amount of Mexican restaurants within a kilometer radius is determined and entered into an empty list. A kilometer was chosen as the search radius because the author assumes most adults in a city will be more than willing to walk a kilometer to a restaurant, however this can be easily modified by the user of the code.

The populated list is then converted into a Python dictionary, and then it is added to the Pandas data frame. Minimum amount of competition is desired, and a lower rent is preferred to keep operating costs as low as possible, so these two columns are normalized with respect to the lowest value in each column. The neighborhood with the largest amount of adults under 30 will have the most potential customers, so that column is normalized with respect to the maximum value in the column. A neighborhood score is determined by providing equal weight to each of the three factors. A neighborhood score of 1 is the lowest possible score, and is considered the best potential score.

# Results

According to the data used in this project the top three neighborhoods that should be considered for opening a Mexican restaurant in New York City are Harlem, Stuyvesant Town, and Morningside Heights. These neighborhoods have the lowest Neighborhood Scores, and theoretically provide the best chance of succeeding for the restaurateur. The worst neighborhoods to open a Mexican restaurant are Chelsea, the Upper East Side, and West Village.

# Discussion

While there are three different data sources used in determining which neighborhood is best for opening a restaurant the results could be improved by including several key pieces of data. These include the average income in each neighborhood and survey results showing the propensity that people in each neighborhood eat out. There is likely a correlation between the two unused data sets, and they will likely impact the results as they currently stand.

The author is satisfied with the results of the project despite the data not being as robust as it could be. This is because the logic used to determine the hypothesis makes sense: fewer competitors make it more likely that a new restaurant will attract customers, low rent presents a smaller barrier to enter the market, and a large potential customer base makes it more likely for a restaurant to succeed.

# Conclusion

The goal of this project is to determine the best neighborhood in New York City to open a Mexican restaurant using three different data sets. While the model is not perfect and could use some first-person collected data in the form of a city-wide survey, it provides the user with a rough idea about where they should initially look to open a restaurant. The model is versatile enough where any type of restaurant can be used in the search; it is not limited to Mexican restaurants.

Future changes to the model (other than increasing the amount of data sets used) would be in determining the average rent in each neighborhood. Commercial rent data is not easily available, and is often limited to what is currently on the market. Therefore, the rent listed in the data set is not always for the same size spaces, or spaces with the same amenities/construction level. Getting more standardized rent data would improve the model greatly.