# Pizza Place in Manhattan

Sam Lake

9th March 2020

## 1. Introduction

### 1.1 Background and Problem

Pizza restaurants are ubiquitous in New York City. There are chains, stalls and local pizzerias everywhere you look. With 80% of restaurants that open there closing within 5 years, choosing the right location is key to having a profitable eatery. But in a city that is swimming in pizza where is the best location for a debut restauranteur to open a pizza place in Manhattan.

### 1.2 Interest

Entrepreneurs and debut restauranteurs looking to open a pizza restaurant in Manhattan would be interested in this report. Also, anyone interested in understanding the pizza restaurant business and their locations a little better.

## 2. Data

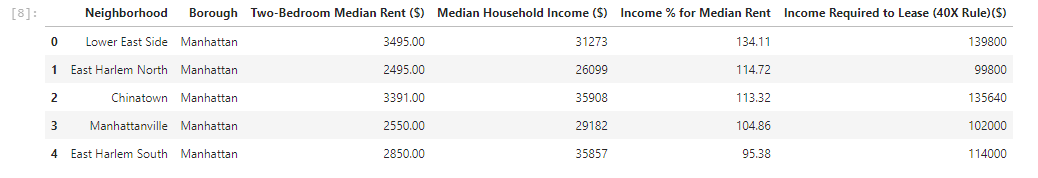
### 2.1 Data Description

I have used the following set of data in this report:

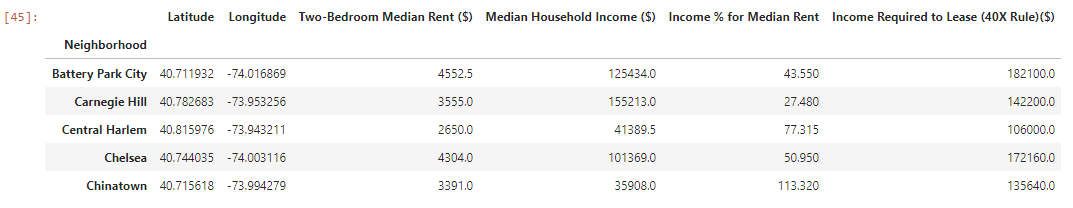
* The table of neighbourhoods in New York from <https://geo.nyu.edu/catalog/nyu_2451_34572>, helpfully converted to a json file on <https://cocl.us/new_york_dataset> and filtered it to Manhattan neighbourhoods only.
* Foursquare API to get restaurant information of any given neighbourhood of Manhattan, New York.
* “RentHop New York Two Bedroom Median Rent Affordability” table from <https://www.renthop.com/study/assets/new-york-city-cost-of-living-2017/nyc-2br-median-rent-and-income-table.html>, converted into a CSV file and again filtered to only Manhattan neighbourhoods.

### 2.2 Data and Methodology

I started by manipulating the RentHop New York Two Bedroom Median Rent Affordability data to divide up areas of Manhattan by economic status (sample shown below).

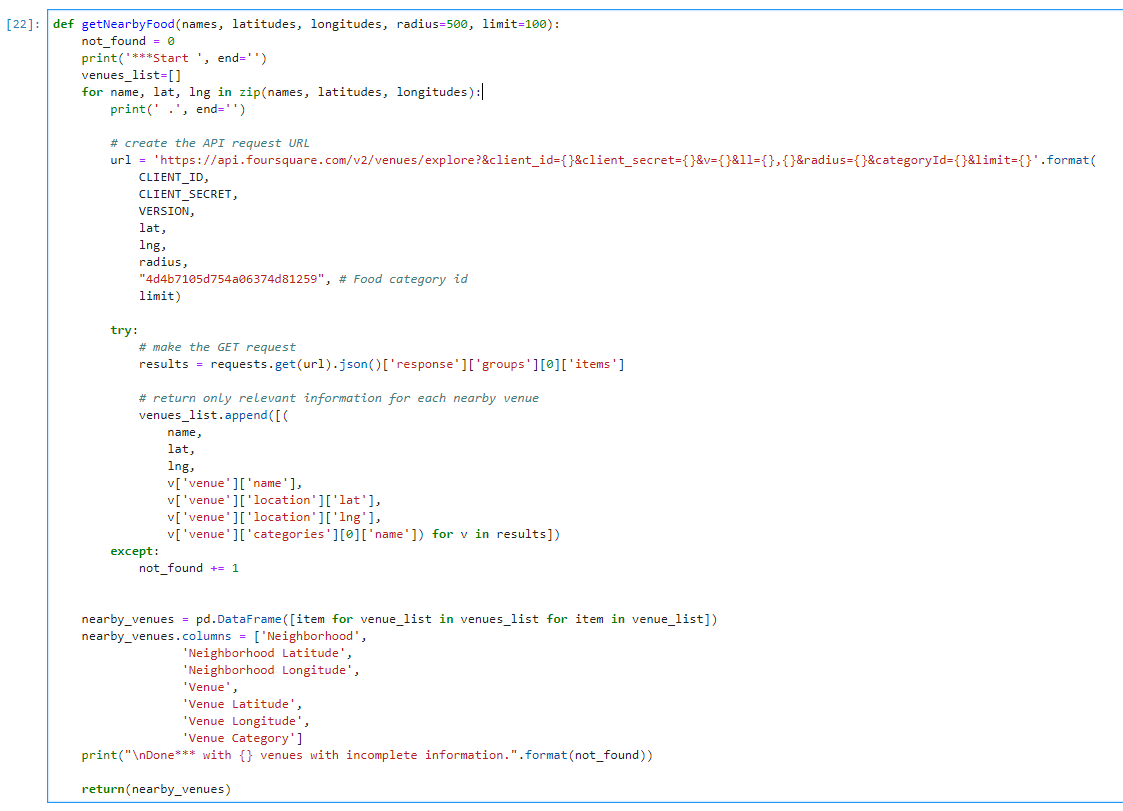


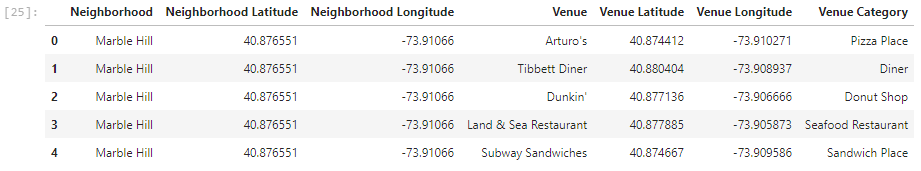
I then joined this data with the coordinates data (sample below).



*I have omitted the neighbourhoods where there was no rent data.*

Then I defined my Foursquare credentials and defined a function for only extracting food venues from the Foursquare API and ran the function to retrieve the venues into a dataframe (sample shown below).





## 3. Visualization and Data Exploration

## 

This dataframe can be transformed using the ‘one hot encoding’ function of the ‘pandas’ library. One hot encoding converts the categorical variables (which are ‘Venue Category’) into a form that Machine Learning algorithms need in order to give better predictions.



Next, I grouped this dataframe by ‘Neighborhood’ and then defined a function that returns the most common food venue types (sample below).

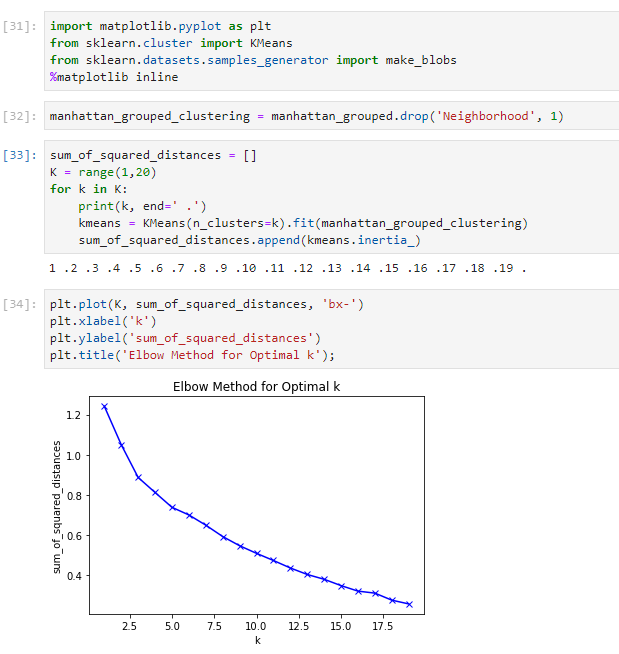


I then chose cluster groups for the data using ‘k-means’. This is an unsupervised machine learning algorithm that creates clusters based on similar points. I used this algorithm to cluster neighbourhoods together with no fixed cluster size.

When using ‘k-means’, it is important to find out the optimal number of clusters (i.e. k). The most popular methods for this are ‘The Elbow Method’ and ‘The Silhouette Method’.

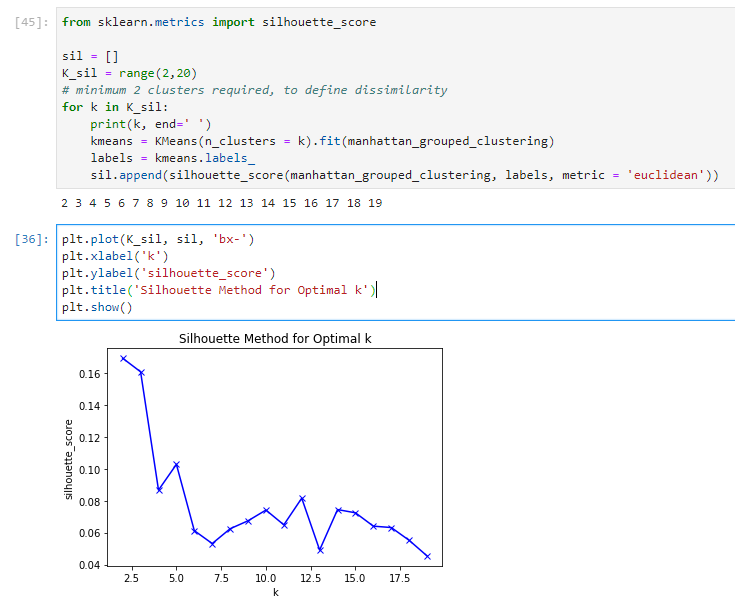
The Elbow Method finds the sum of squared distances of samples to their closest cluster centre for different values of ‘k’. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances (at the elbow).

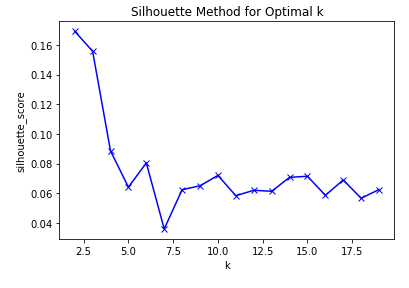
My implementation is below:

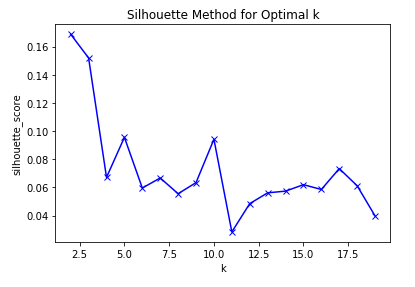


As you can see, there is no defined elbow when using this method over my data set. I moved on to implementing the ‘Silhouette Method’ instead.

As quoted in Wikipedia — “The Silhouette Method measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).”

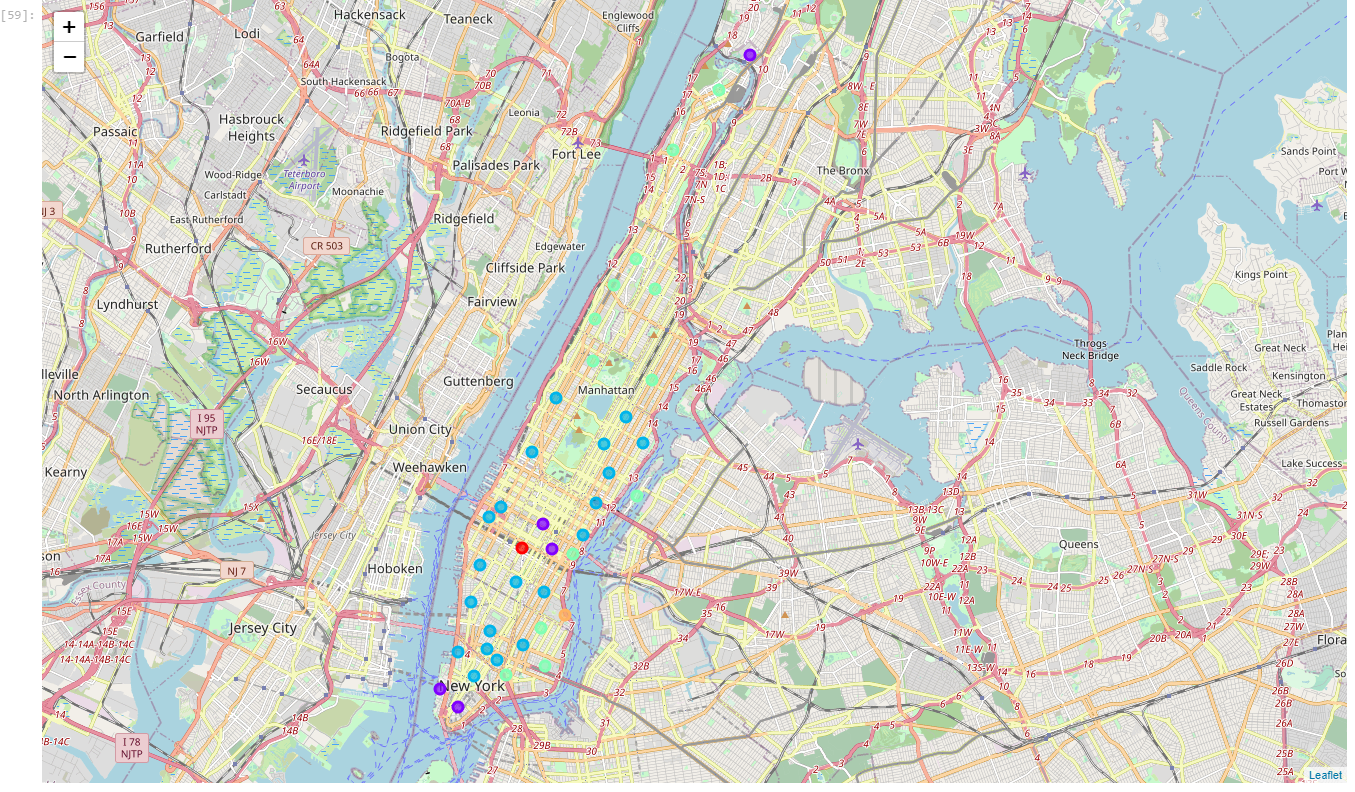


However, as this algorithm uses random starting points (centres) every time it is run, I ran it a few times. Here are a few different results I received:



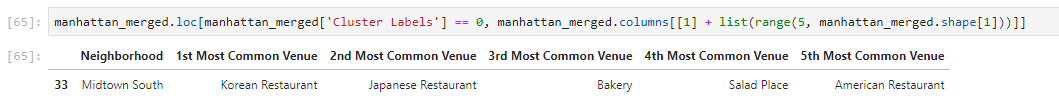
As there seemed to be a consistency around 5 and 10 and ten number of clusters will cluster the neighbourhoods very specifically, the number of clusters (i.e. ‘k’) chosen was 5.

I then generated the following map using the Folium package. The locations of the neighbourhoods and which cluster they belong to is now clear.



## 4. Results

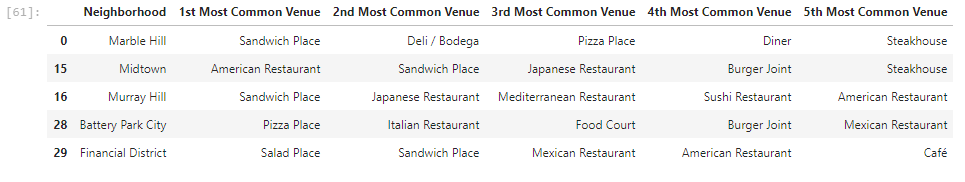
### 4.1 Cluster 0



In a cluster all on its own, Midtown South’s most common venues consist of Asian cuisine. Pizza Place doesn’t even make the top five.

Cluster 0 is a ‘Korean/Japanese Restaurant’ cluster.

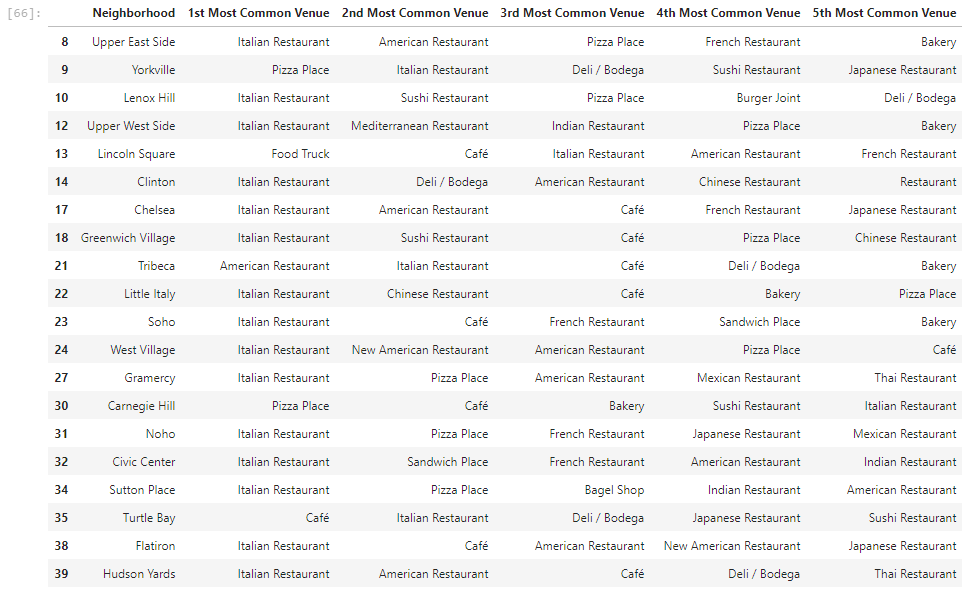
### 4.2 Cluster 1



Cluster 1, although one neighbourhood has ‘Pizza Place’ as its most common venue, is not the most relevant cluster.

Cluster 1 is a predominantly a ‘Sandwich Place’ cluster.

### 4.3 Cluster 2



‘Italian Restaurant’ holds massive accountability for this cluster with 15 occurrences followed by ‘Pizza Place’ with 2 occurrences in ‘1st Most Common Venue’ across different neighborhoods. This is arguably the most relevant cluster as over half of the neighbourhoods had ‘Pizza Place’ in their top five most common venues and there being a big crossover between Italian restaurants and pizza places (i.e. pizzas being Italian).

Cluster 2 is an ‘Italian Restaurant’ cluster.

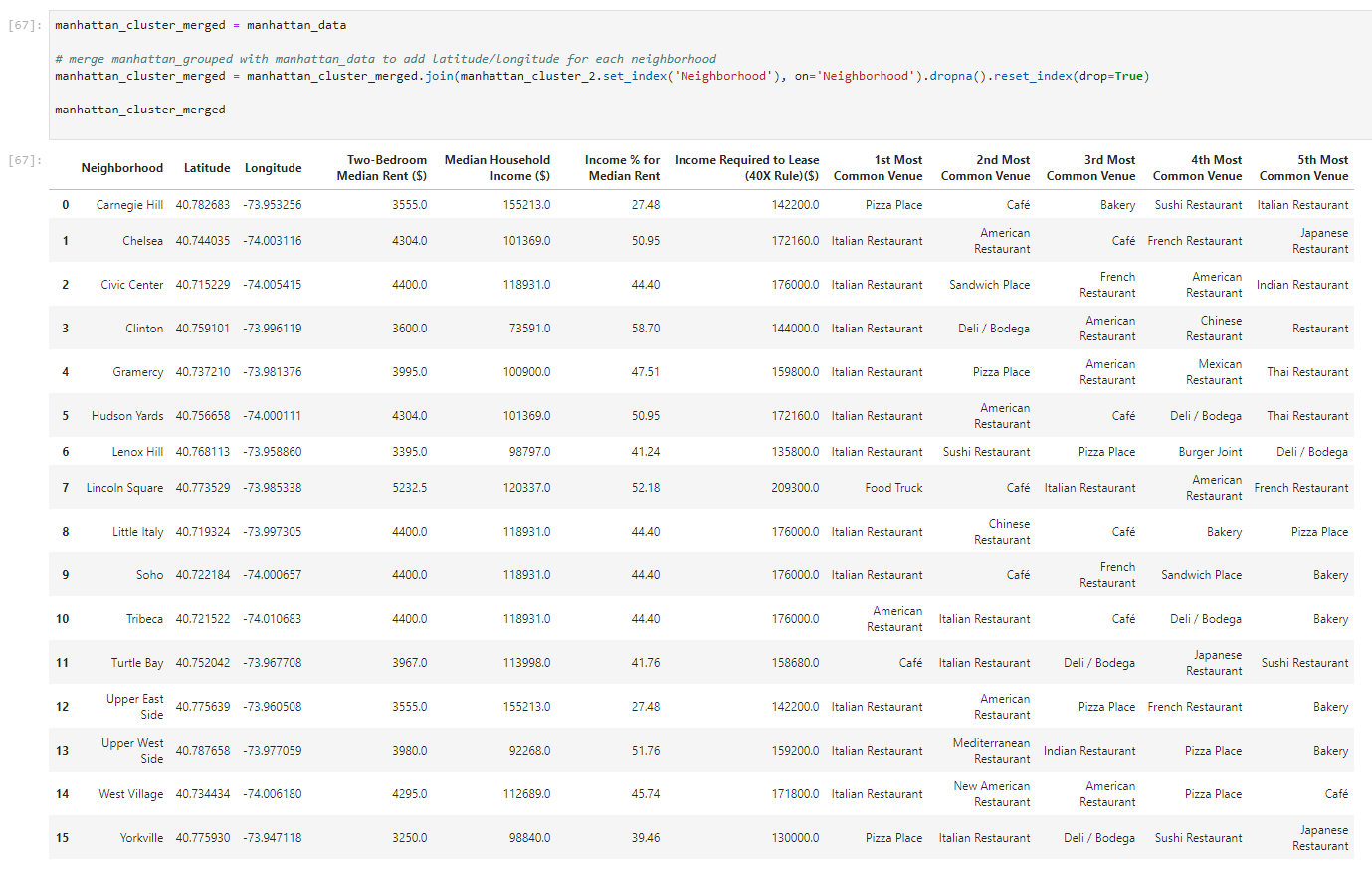
### 4.4 Cluster 3



This final cluster, while having more occurrences of ‘Pizza Place’ as the 1st most common venue than the cluster 2, is mainly a ‘Deli/Bodega’ cluster.

## 5. Discussion

After analysing the clusters, I chose the cluster most relevant to this report (the cluster where ‘Pizza Place’ and ‘Italian Restaurant’ is high up the common venues list). I then joined this table to my RentHop New York Two Bedroom Median Rent Affordability data.



From this table we can see that “Pizza Place” is the most common food venue for Carnegie Hill and that the Median Household Income is $1,552,130 and the Income % for Median Rent in that neighbourhood is 27.48%. Of the neighbourhoods in this cluster, we can assume that Carnegie Hill residents have the most disposable income of all the Manhattan neighbourhoods that like to frequent Pizza places. Therefore, it is the best place to open a Pizza place in Manhattan according to this data.

## 6. Conclusion

Using the k-means clustering algorithm I have managed to cluster the neighbourhoods of Manhattan visually explore the location data to suggest the best neighbourhood to open a pizza place.

This report did not consider tourist numbers and destinations or neighbourhoods with no rent data (e.g. Financial District etc.), focusing mainly on the median income and assumed disposable income of the residents of a neighbourhood. I would have improved this report by investigating more into restaurants in tourist hotspots and looking into the profits of individual restaurants.