# 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

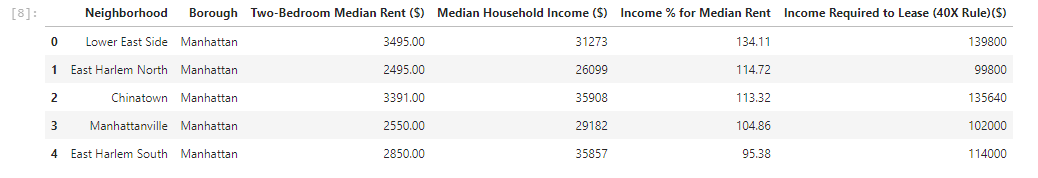
### 2.1 Data Description

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

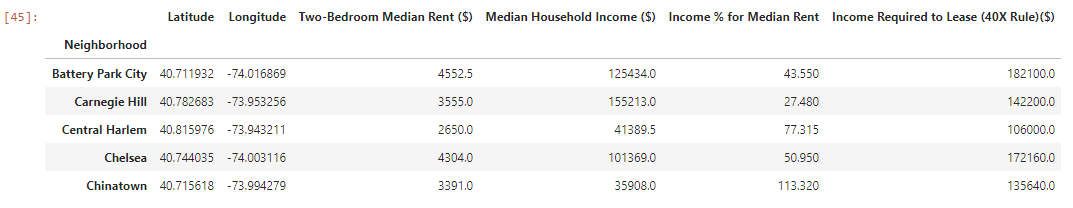
* The table of neighbourhoods in New York from <https://geo.nyu.edu/catalog/nyu_2451_34572>, helpfully converted to a json file on <https://cocl.us/new_york_dataset> and filtered it to Manhattan neighbourhoods only.
* Foursquare API to get restaurant information of any given neighbourhood of Manhattan, 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

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).

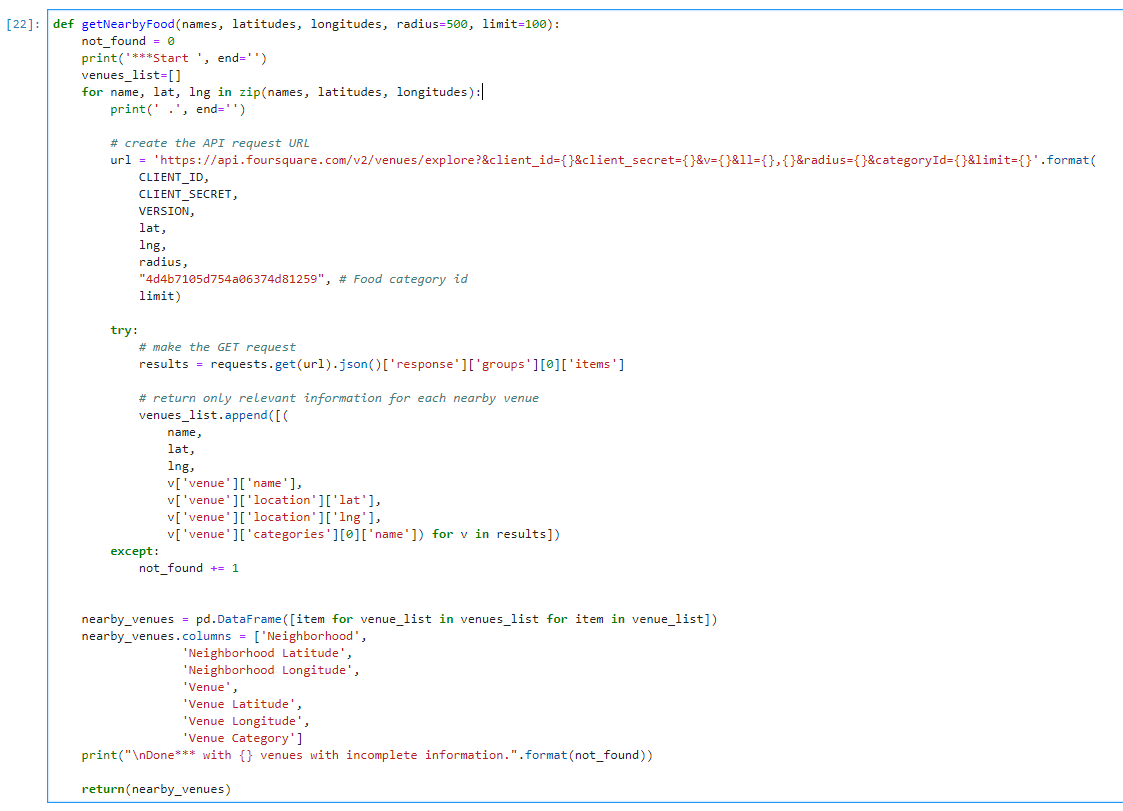


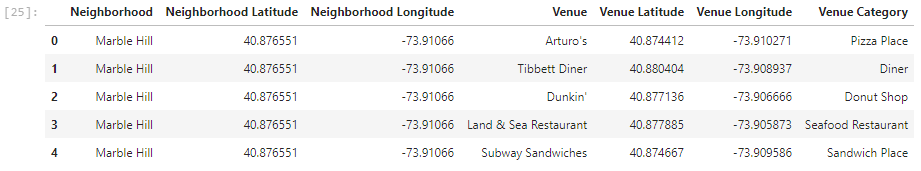
I then joined this data with the coordinates data (sample below).



*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).





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).

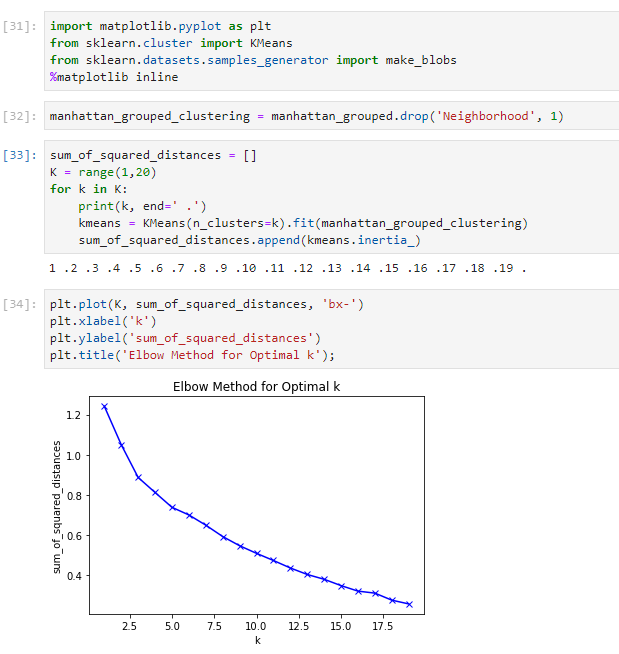


I then chose cluster groups for the data using ‘k-means’. This is an unsupervised machine learning algorithm that creates clusters based on similar points. I used this algorithm to cluster neighbourhoods together with no fixed cluster size.

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’.

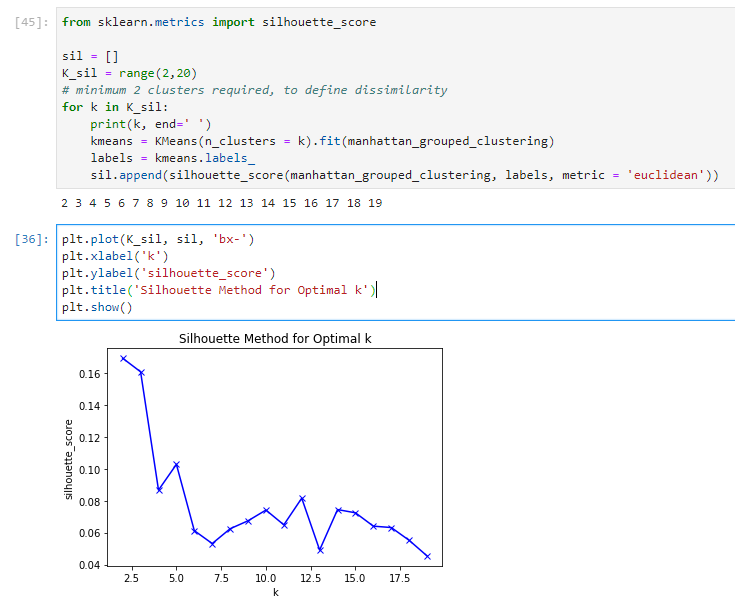
The Elbow Method finds the sum of squared distances of samples to their closest cluster centre for different values of ‘k’. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances (at the elbow).

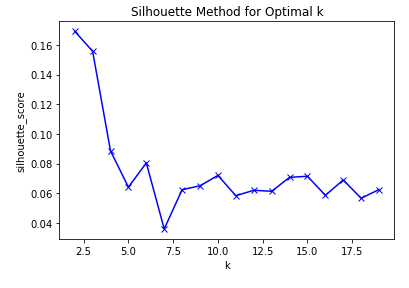
My implementation is below:

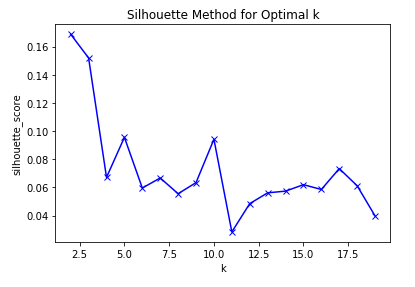


As you can see, there is no defined elbow when using this method over my data set. I moved on to implementing the ‘Silhouette Method’ instead.

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

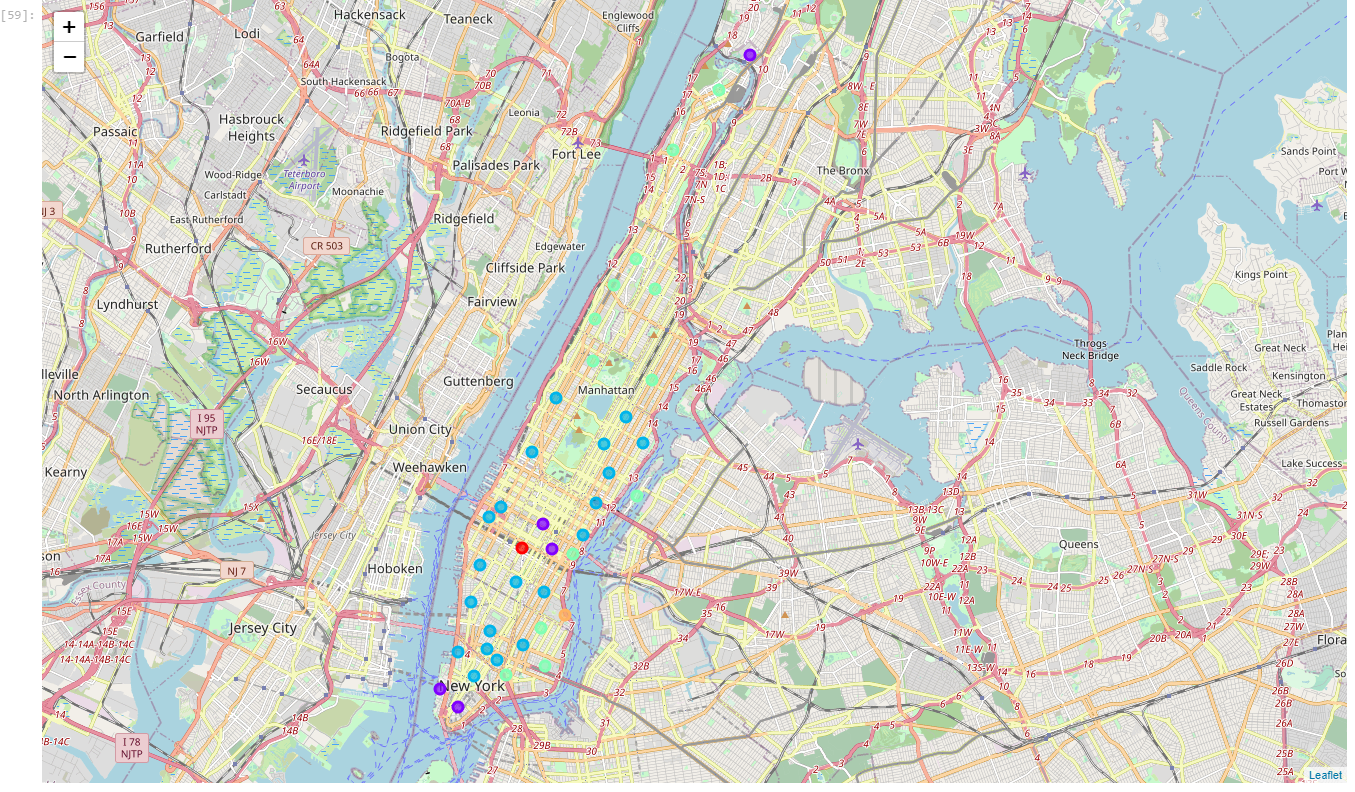


However, as this algorithm uses random starting points (centres) every time it is run, I ran it a few times. Here are a few different results I received:



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.



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

