# What is the best location for a coffee stall in Manhattan, NYC?

## Introduction

### Background

This report aims to make a recommendation for the best location for a coffee stall in Manhattan for entrepreneurs and investors. The best location is one in which a coffee stall is considered more likely to create a profit and provide a sustainable income relative to other locations. Competition and potential demand are assumed to be the most important factors for predicting profitability. Competition can be measured quantitatively by the counting the number of coffee venues within a sampling radius. Potential demand can be measured quantitatively by the foot traffic at a sampling point. Competitiveness of a coffee shop would also be affected by the average rent of a location, but this is assumed to be of limited importance for a mobile coffee stall.

### Problem

We assume the factors that best predict profitability are foot traffic and competition. The best location in a relative sense is that with a relatively low number of coffee shops but a high volume of foot traffic. This project aims to predict the best location for a coffee shop within Manhattan. We wish to consider factors in a relative context and deliver a coffee stall location recommendation based on the results.

### Interest

New York City has over 3000 Coffee Shops, a high proportion of these are in Manhattan, making it an already competitive location. A prediction on the best location for a coffee stall would be useful to entrepreneurs or investors in Manhattan looking to create a competitive, profitable, and sustainable business decision.

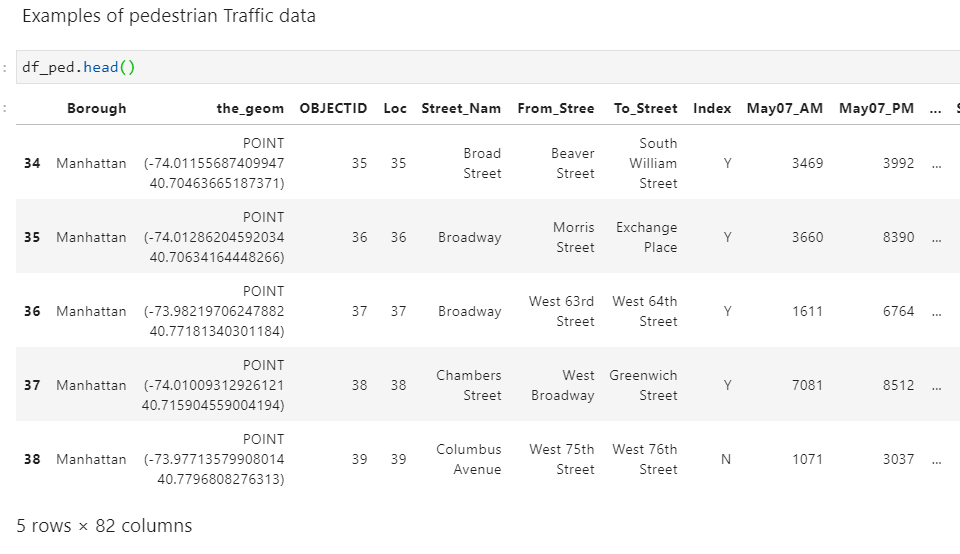
## Data acquisition and cleaning

### Data sources

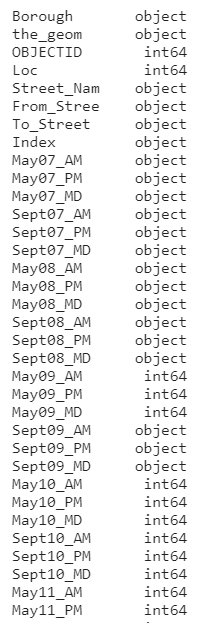
Three datasets will be used:

1. Pedestrian Traffic data from NYC Open Data available [here](https://data.cityofnewyork.us/api/views/cqsj-cfgu/rows.csv?accessType=DOWNLOAD&bom=true&format=true).

This dataset runs from 2007 to 2019. Measurements of foot traffic are taken at sampling points over New York City. Manhattan foot traffic data are identifiable by the column *“Borough”.* Foot traffic data is sampled biannually in May and September at morning, midday, and afternoon.

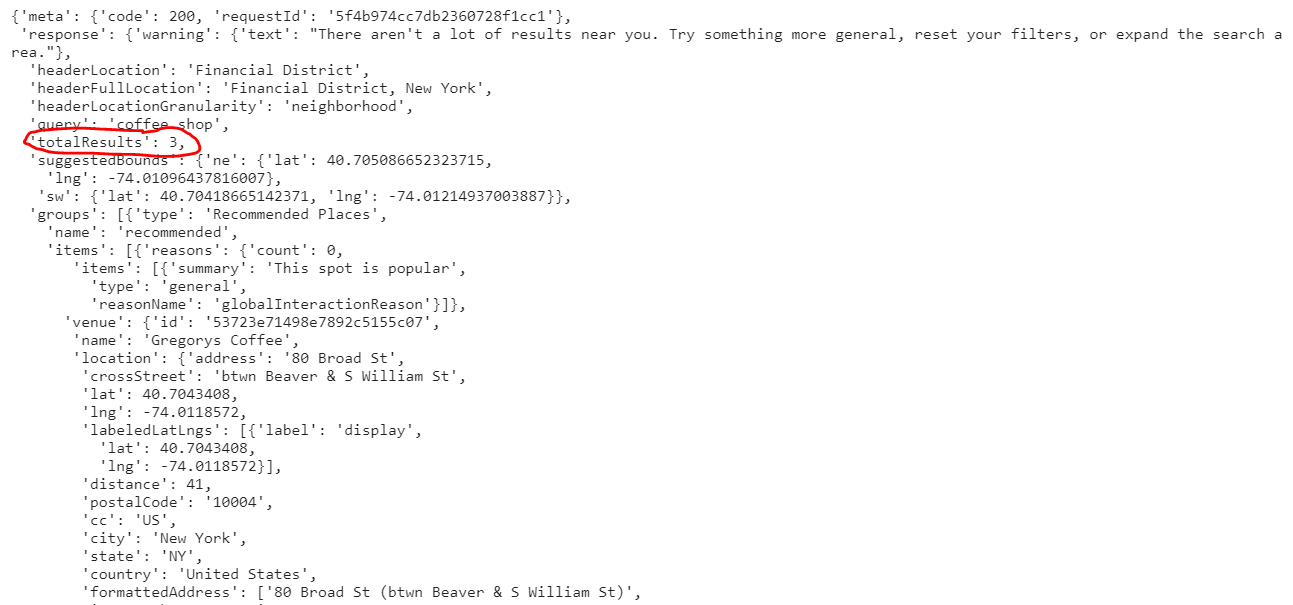


The foot traffic dataset has a variety of data types. *“Borough”, “the\_geom”* (longitude, latitude coordinate) are both object datatype. The time series data varies between object and integer, reflecting inconsistent use of thousand separators.



1. The number of coffee shop venues for each sampling location from Foursquare API.

The Foursquare API venue explore call will be used to find the number of coffee shops within a foot traffic sample point. The call will return a JSON file. The JSON file contains an element *“totalResults”* which provides a count of venues returned.

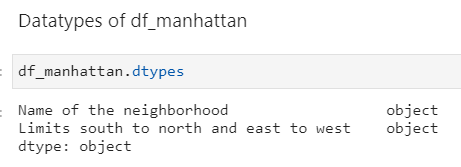


The count of coffee venues will be added as a column to the foot traffic dataset.

1. List of Manhattan neighbourhoods from Wikipedia available [here](https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods#:~:text=The%20following%20approximate%20definitions%20are,34th%20Street%20and%2059th%20Street.).

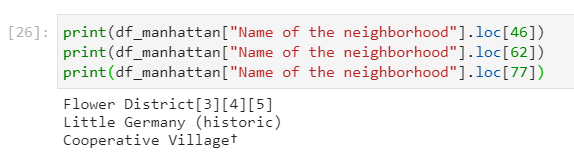
This data will be used to compliment the New York Pedestrian Foot Traffic Data, assigning a neighbourhood based on the latitude and longitude value of a foot traffic sampling point.

Both columns are of the object data type.

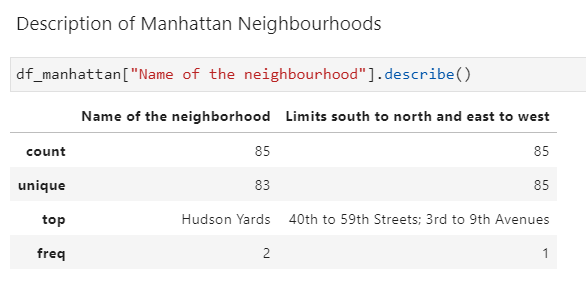


The GeoPy Library will be used to assign an neighbourhood name to each latitude and longitude coordinate value in the Pedestrian Traffic Data using the *“Reverse”* function. If a matching neighbourhood address cannot be assigned, one will be assigned manually.

The relevant column is *“Name of neighbourhood”*. There are some complications with scraping this data in that it will contain bracketed elements, footnotes, and daggers which will interfere with the process of assigning a neighbourhood name to each foot traffic sample point.



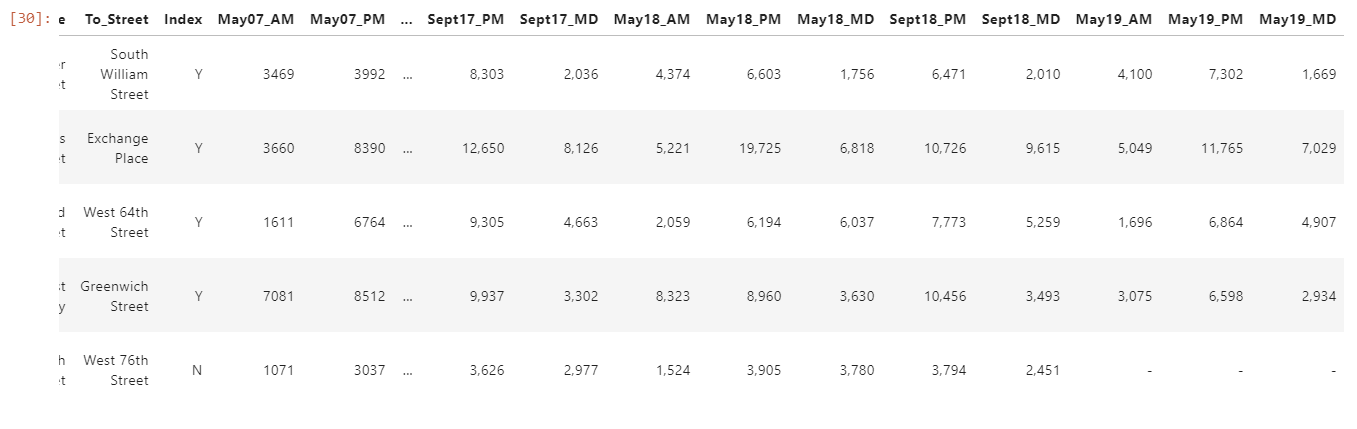
There also appears to be duplicate entries for *“Name of neighbourhood”,* which may be erroneous. One example is *“Midtown”*. One of the duplicates of Midtown is likely a region of Midtown (e.g. Midtown South).



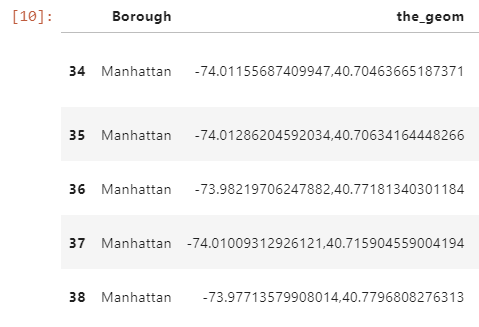
### Data cleaning

1. Pedestrian Traffic data from NYC Open Data.

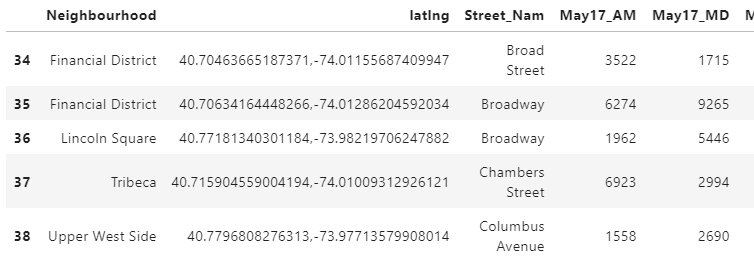
The dataset contains data from New York City. Only data with the Borough value *“Manhattan”* will be used. For some Manhattan sample points, 2019 data is missing. To develop a recent and consistent dataset, we will use data from 2017 to 2018.



The coordinate values in the *“the\_geom”* column is in the format *“POINT (longititude, latitude)”*. *“POINT”* and the brackets were removed and the space between longitude and latitude replaced with a comma.

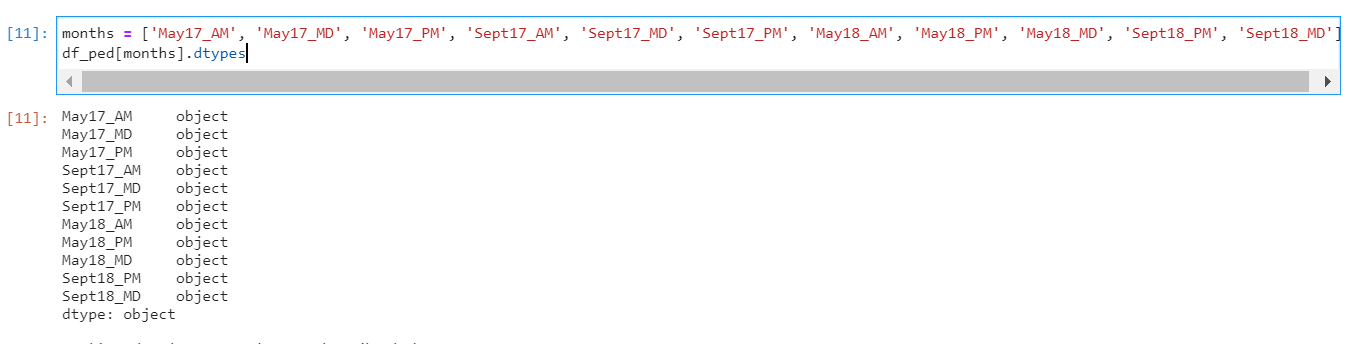


An additional column *“latlng”* was also added which reversed to longitude, latitude value in a latitude, longitude value.



This was reversed into a latitude longitude value so that it can be plotted on a map and have a Neighbourhood name assigned to it using the GeoPy library.

Time series data on foot traffic volume is in an object format, containing thousand separators in the form of a comma.



The thousand separators were removed, and the data type changed to integer.

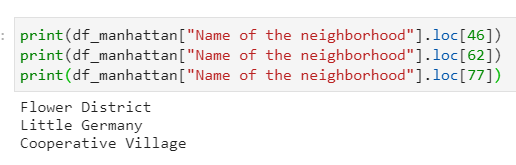


1. Foursquare JSON Data

This JSON file required no cleaning, as an element *“totalResults”* returned by the venue explore call will return a count of the number of venues within a 100m radius of a foot traffic sample point.

1. List of Manhattan neighbourhoods from Wikipedia

The neighbourhood names contain bracketed names, square bracketed footnotes, daggers. These were removed.

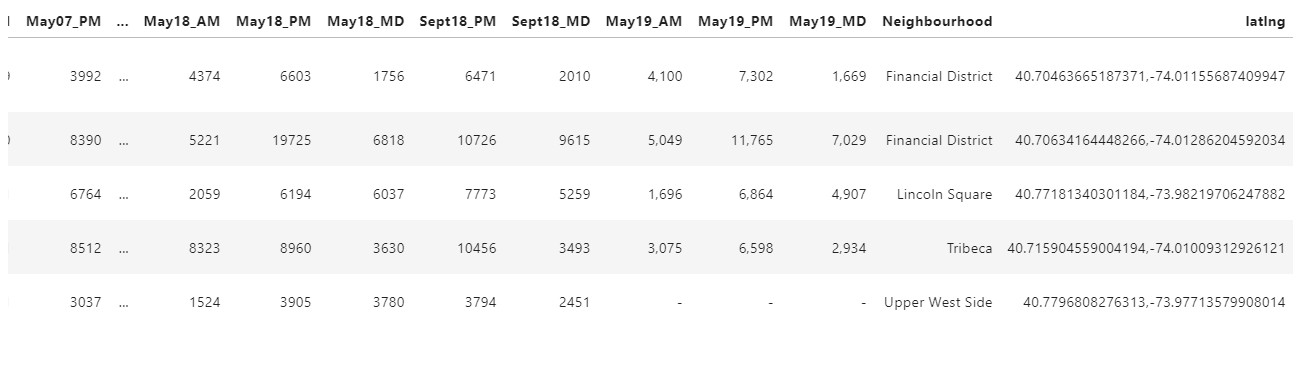


There are two entries for *“Midtown”.* One of the GeoPy *“Reverse”* function calls returned an address for a foot traffic sampling point called which was *“Midtown South”* which matched no existing value in the list of Manhattan neighbourhoods. It is assumed that the duplicate Midtown value is *“Midtown South”* from the address data returned by GeoPy. The sampling point that returned the error was manually assigned the neighbourhood *“Midtown South”.*

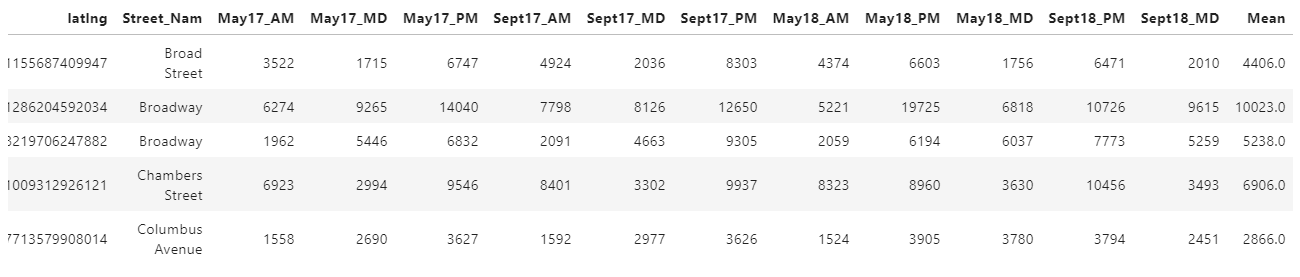


## Methodology

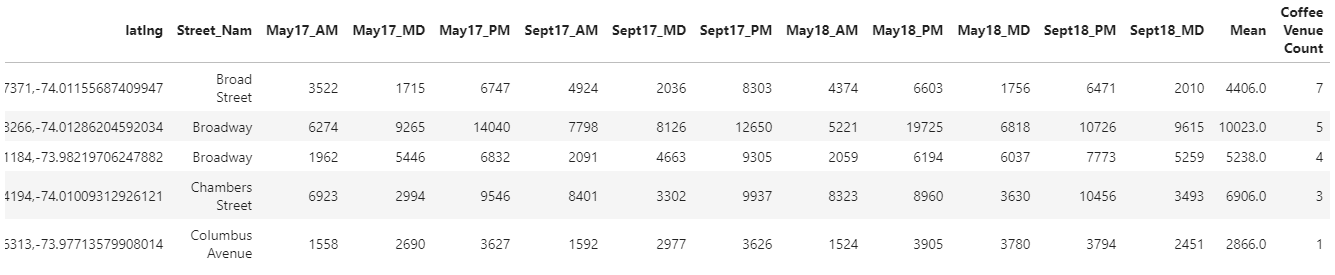
Each sample point in foot traffic dataset was assigned a Manhattan Neighbourhood name according to its latitude and longitude coordinate based on column *“the\_geom”.*



A column *“Mean”* was added, which is the row average of 2017 and 2018 foot traffic data for each sampling point.



Each sample point in the foot traffic data was also assigned a “Coffee Venue Count” value using a Foursquare API call for coffee shop venues within 100m of the sampling point. It was assumed that coffee shops with in 100m are a competitor business and are at an equal level of competitiveness.



Because we wish to compare foot traffic sampling points in a relative sense, min-max normalisation was applied to this data. Min-max normalisation was used to return a value between one and 0.

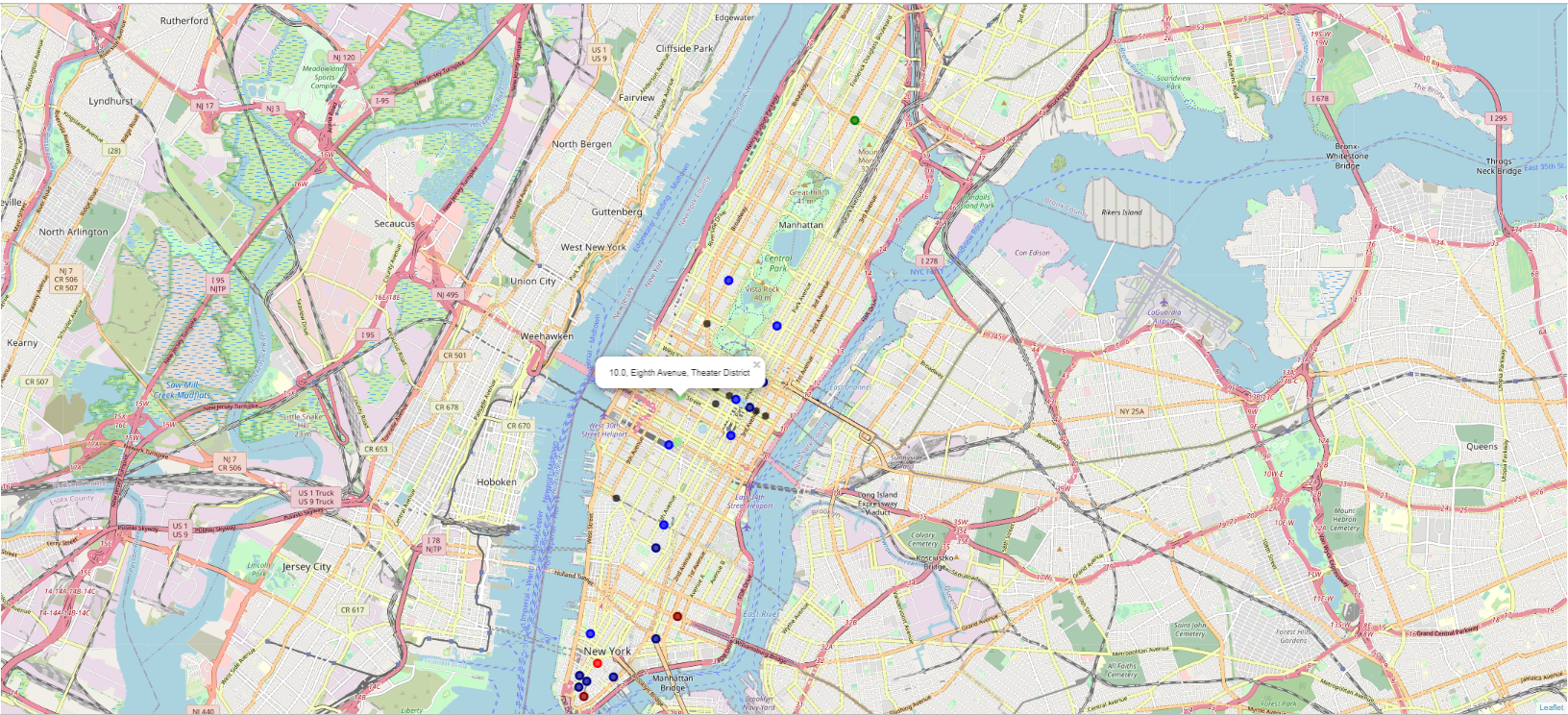
This produced two columns *“CVC Normalised”* and *“Mean Normalised”.*

The difference between *“Mean Normalised”* and *“CVC Normalised”* was calculated and this value was again normalised to produce a column “*Score Normalised”.* The foot traffic sample point with the greatest difference between normalised mean foot traffic and normalised number of coffee venues is considered the best location for a coffee stall in Manhattan.

A map was produced with colour coded markers that reflect the foot traffic sampling point’s suitability as a coffee stall venue. Dark red markers indicate a relatively unsuitable area with either relatively high competition, relatively low foot traffic or some relatively uncompetitive combination of the two. Light green indicates a highly suitable area with either low competition, relatively high foot traffic or some relatively competitive combination between these two factors.

## Results

The results were plotted on the map below:



We assume that scores of 7.5 and above the most relatively competitive suitably locations:

The score for relative competitiveness ranges from dark red (uncompetitive) to light green (most competitive).

The best locations in descending relative competitiveness are:



## Discussion

Of the locations listed above, those that scored over 7.5 are considered the most competitive out of the Manhattan foot traffic sampling locations.

The best locations in descending relative competitiveness are:

1. Eighth Avenue in Theatre District
2. Fifth Avenue in Midtown
3. West 34th Avenue in Garment District
4. West 125th Street in Harlem

Foot traffic sampling locations were concentrated along Midtown and lower Manhattan. Upper Manhattan had relatively few foot traffic sampling locations, which scored relatively high on competitiveness. Upper Manhattan may have been underestimated in the number of competitive locations for a coffee stall. Relatively competitive locations appear to be concentrated around midtown Manhattan inland from a river edge.

## Conclusions

Our analysis aimed to produce a recommendation for the best location to set up a coffee stall in the foot traffic sampling locations available in Manhattan. Our analysis factored in potential demand from foot traffic and the potential competition from existing coffee shop establishments. Normalised data for foot traffic and competition were used to calculate an overall score which was used to grade each foot traffic sampling location against each other. We assumed that any coffee shop establishment within 100m radius of a location was equally competitive and that all locations legally permit coffee stalls.

From this analysis, a map and data frame were created which found that the following Manhattan foot traffic sample locations scoring over 7.5 are likely to be the most competitive in a relative sense.

1. Eighth Avenue in Theatre District
2. Fifth Avenue in Midtown
3. West 34th Avenue in Garment District
4. West 125th Street in Harlem

These conclusions are useful in a relative context, in comparing the competitiveness of a foot traffic sampling point in Manhattan to another. There is no guarantee that the recommended locations for a coffee stall will produce a profit or be sustainable business ventures. The analysis only suggests which foot traffic sampling location may be more competitive than another.

## Future Directions

Foot traffic data is publicly available for other boroughs of New York such as Brooklyn, The Bronx, and Staten Island. The same analysis could be expanded using the data available and methodology described above. An expanded analysis could compare foot traffic sampling points within other neighbourhoods or New York City as a whole.

Foot traffic sampling points could be taken at a greater density to provide more potential coffee stall locations.

Profitability data for coffee shops are unlikely to be publicly available, so developing a predictive model would be difficult without further gathering of data. The lifespan for a venue is likely to be correlated with profitability or sustainability, but no dataset could be found by the author. If data of some quantitative measure of coffee venue success were available, a predictive model for coffee stall profitability could be developed, different profitability factors identified, and their relative importance quantified.

With a predictive model, assumptions of a 100m competitive radius could also be challenged or justified, but this could conceivably be limited by the limits of a Foursquare venue explore call. Only 100 coffee venues at maximum could be returned for any radius. At a higher radius, the mean number of coffee venues returned may be close to 100.

The assumptions of legal permission for coffee stalls being available at all locations could also be challenged or confirmed with further knowledge of New York City, Borough or Neighbourhood law. Foot traffic sample locations that fall within neighbourhoods with restrictions on coffee stalls could be removed from the results.