1.Introduction

In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it? Similarly, if a contractor is trying to start their own business, where would you recommend that they setup their office?

1.1 Problem discussion

This report would like to select Manhattan to expand this problem. So far we need to select a place to open a restaurant in the Manhattan. Restaurant has many different kinds. It could be a small take out food shop, and it could also be a luxtry sushi place. The different restaurant has different investment and

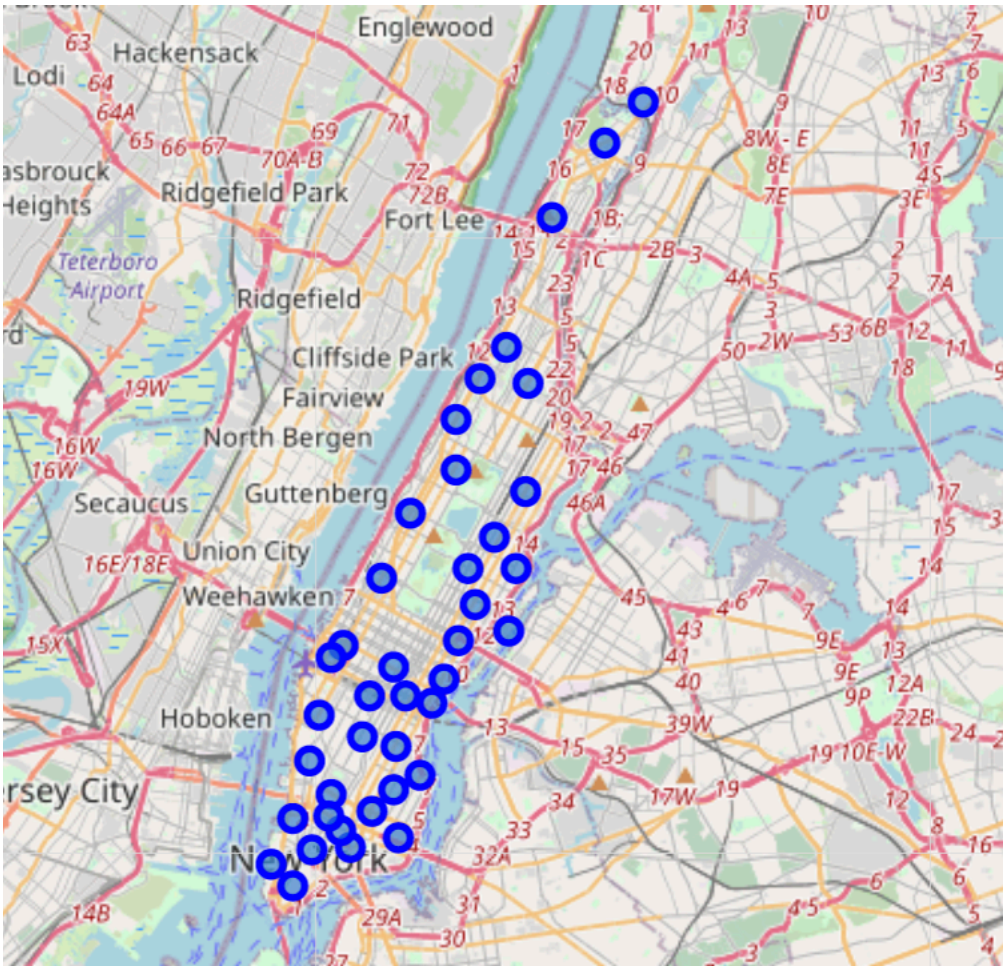

Figure 1: Plot of all the neighbourhood in manhattan area

different incomes. If someone want to open a steak house, it is not a good idea to open it in chinatown. Nevertheless, investing a chinsese restaurant at the

chinatown might be a very good choise. Even if we do not catagory the restaurant kind, there are also many factors to select a place to open a general restaurant. How detailed the problem could be solved depends on how much data and time we have to approach this probelm.

1.2 Map Ploting

We focus on the neighbourhood in the Manhattan. Each neighbourhood represents a possible selection for our restaurant opening place. We need to first plot the neighbourhood distribution of the Manhattan. Implementing the folium to plot the neighbourhood in the Manhattan.

1.3 Problem definition

The above figure shows the neighbourhood distribution for the Manhattan. The first problem could be definded as "If someone want to open a general restaurant in the Manhattan, which neighbourhood shown above is the best choise for him?"
Then if the data and time is allowed, the deeper investigation could be into a problem that "If someone wants to open a specific restaurant in the Manhattan, and he tells you what kind of restaurant he would like to open and how much money he can invest, which neighbourhood shown above is the best choice for him?"

2 Data Selection

2.1 Data Source

The data contains the neibourhood's location data and the venue data. The location data is download from IBM developer skill learning website the location is from http://cocl.us/Geospatial_data. The venue data is request from the foursquare.

2.2 Data Selection

The data contains three levels. The first level is the location of each neighbourhood. The second level is the venues belong to each neighbourhood, and the third level is the properties of each venue.

We will focusing on the venue which belong to the food catagory. If the time is allowed, we will separate the restaurant into detailed catagory to analyze. First, lets look at the first level data.

2.2.1 First level data

The first five rows of the first level data is like this.

	Neighbourhood	Borough	Latitude	Longitude
0	Marble Hill	Manhattan	40.876551	-73.910660
1	Chinatown	Manhattan	40.715618	-73.994279
2	Washington Heights	Manhattan	40.851903	-73.936900
3	Inwood	Manhattan	40.867684	-73.921210
4	Hamilton Heights	Manhattan	40.823604	-73.949688

2.2.2 Second level data

The second level data includes the venues list in the Neighbourhood. Using the explore endpoint to request the food-category venues information from foursquare. The first 5 rows of the second level data is shown below. It has 2803 rows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	id
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place	4b4429abf964a52037f225e3
1	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner	4b79cc46f964a520c5122fe3
2	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop	4b5357adf964a520319827e3
3	Marble Hill	40.876551	-73.91066	Land & Sea Restaurant	40.877885	-73.905873	Seafood Restaurant	4b9c9c6af964a520b27236e3
4	Marble Hill	40.876551	-73.91066	Subway	40.874667	-73.909586	Sandwich Place	4b4f7b65f964a5205a0827e3

The second level data also contains a full venue list. It not only contains the restaurant list but also contains other venues such as gym and gas stations. The first 5 rows of the second level data is shown below. It has 3442 rows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	id
0	Marble Hill	40.876551	-73.910660	Arturo's	40.874412	-73.910271	Pizza Place	4b4429abf964a52037f225e3
1	Marble Hill	40.876551	-73.910660	Bikram Yoga	40.876844	-73.906204	Yoga Studio	4baf59e8f964a520a6f93be3
2	Marble Hill	40.876551	-73.910660	Tibbett Diner	40.880404	-73.908937	Diner	4b79cc46f964a520c5122fe3

3	Marble Hill	40.876551	-73.910660	Dunkin'	40.877136	-73.906666	Donut Shop	4b5357adf964a520319827e3
4	Marble Hill	40.876551	-73.910660	Astral Fitness & Wellness Center	40.876705	-73.906372	Gym	4cf6ae55d3a8a1cd71a9d243

2.2.3 Third level data

The third level data comes from the foursquare API, by using premium request. Each line corresponds to the same line in second level restaurant data. The first 5 rows of the third level data is shown below. It has 2803 rows.

	Tier	Rating	TipNumber	Likes	Catagory Number	Catagory
0	2	9.0	16	18	1	[Caribbean Restaurant]
1	1	6.3	2	2	1	[Pizza Place]
2	1	0.0	4	1	1	[Chinese Restaurant]
3	2	0.0	2	1	1	[Caribbean Restaurant]
4	2	9.0	16	18	1	[Caribbean Restaurant]

Clean the third level data

It is possible that we asked a venue for detailed properties but the response message lacks some data such as rating. We put zero for the lacking integer and float format data. The category column will always have non-void data.

3 Analysis

3.1 General data visualization

After obtaining the three levels of data, one need to obtain some basic statistics to analyze. The first and the most cardinal index of a neighbourhood if it is good to open a restaurant is the existing restaurant numbers. Figure 2 shows the restaurant distribution on the manhattan neighbourhoods.

Not only the food venue numbers, but the total venue number is important too. The total venue numbers reflects the economical development level for the certain neighbourhood. Figure 3 shows the total venue distribution on the manhattan neighbourhoods. Although some old neighbourhood kept the high level of this index, but it still a reasonable index for reference.

Figure 2: Food Venue numbers comparison between all the neighbourhood in manhattan(red for large number of food venues, blue for small number of food venues)

The food venue portion which reflects if this neighbourhood is a restaurant oriented business circle. Figure 4 shows the food venue portion on the manhattan neighbourhoods. The index indicates if the neighbourhood is suit for restaurant. Opening a restaurant into an industry area is not a good idea to accumulate money

Figure 3: Total Venue numbers comparison between all the neighbourhood in manhattan(red for large number of venues, blue for small number of venues)

Figure 4: Food Venue portion comparison between all the neighbourhood in manhattan(red for large portion of food venue, blue for small portion of food venue)

3.2 Third Level data Statistics

The last section investigates the general statistics of the neighbourhood. This section, we used the data collected from each food venues(restaurants) in the neighbourhood to see what kind of restaurant is in each neighbourhood. The first index is the average like number of each food venues. This number indicates the customer rate in the restaurant to some extent. The most important factor for a restaurant to success is if it has enough customer number. Although each restaurant has its own factor such as the flavour and decoration to attract the customers, this factor can more or less indicates the location factor on the customer number. The figure 5 shows this index.

The next index is the price of the food. The higher price indicates the more profits on single customer. This is not always true, but also to some extent it can reflects. The figure 6 provides the average price tier level of the restaurant venue in each

Figure 5: Average like numbers for food venue in the neighbourhood. Red for large like number, blue for small like number.

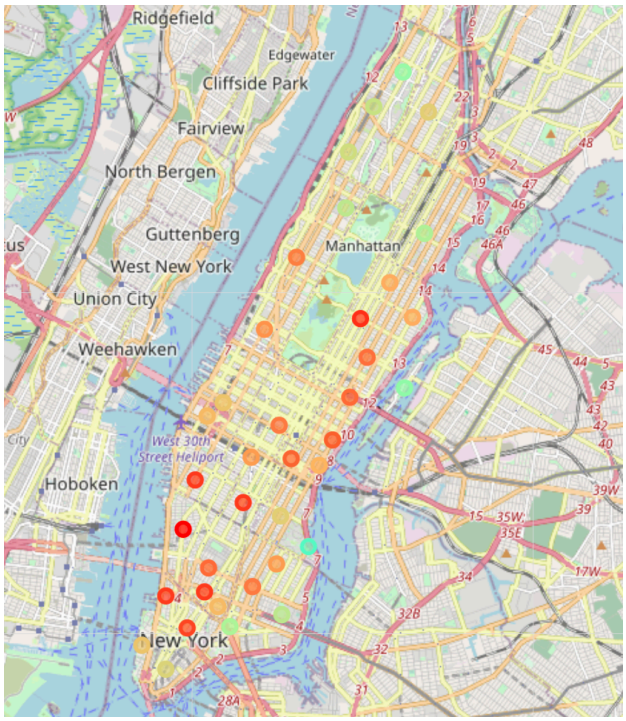

Figure 6: Average price level for food venue in the neighbourhood. Red for higher price, blue for lower price.

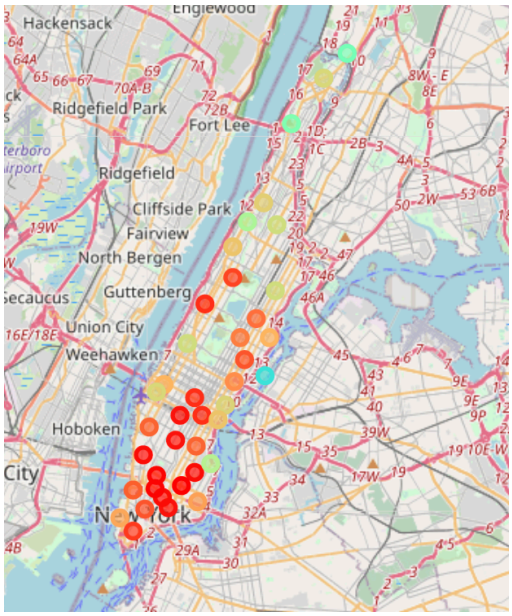

Figure 7: Average rating numbers for food venue in the neighbourhood. Red for higher rate, blue for lower rate.

neighbourhood. The red circle represents the higher average price while the blue colour circle represents the lower average price.

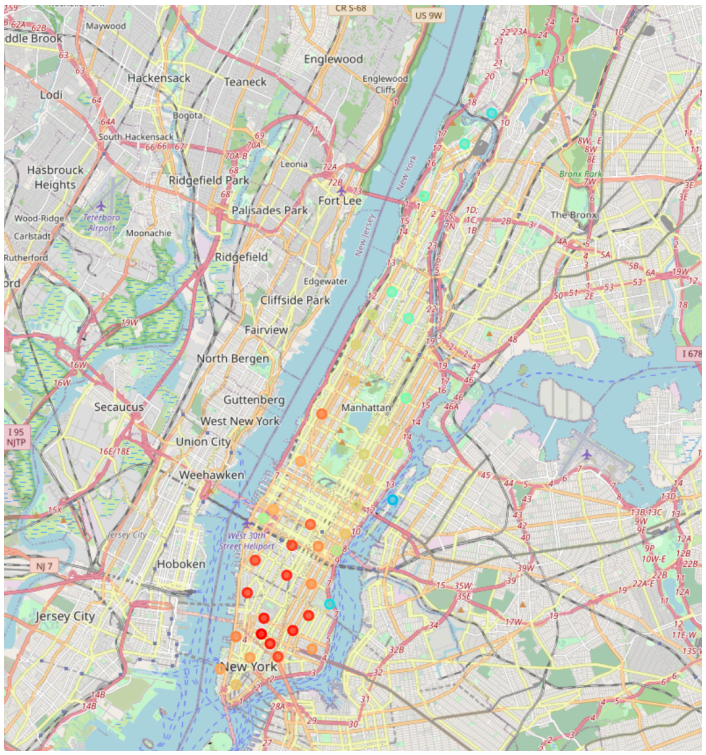

Figure 8: Average tip numbers for food venue in the neighbourhood. Red for higher tip number, blue for lower tip number.

The next index is the average rating of each restaurant in the neighbourhood. This index reflects the returning customer rate to some extent. Figure 7 shows this index. The red circle mean the better average rate the food venues has in this neighbourhood. While the blue circle means the average rate of the food venues is bad.

The next index is the average tip numbers of the food venues. This index can also reflect the average customer numbers. It can also reflect the customer feedback on the food venue. Figure 8 provides this index. The red circles represent the better tip rates while the blue circles represent the bad average tip rates.

4 Modelling

4.1 The overall relative matrix.

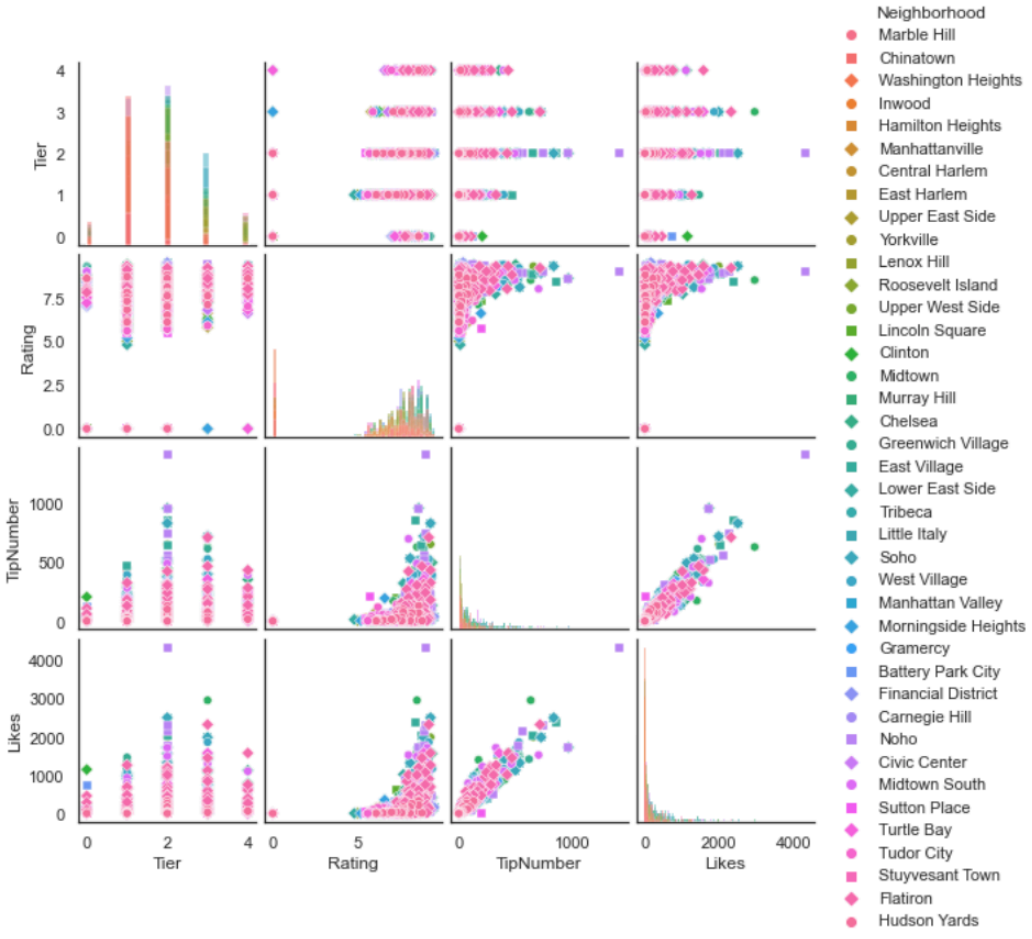

Figure 9: Correlation plot for 4 third level indexes for each individual food venue, grouped by their neighbourhood

First we plot the four third level average variables which we discussed in the last section by each venue to see if they are related to each other or not(Figure 9). It is able to see that the tip number is highly related to the total number of likes.

Here we first normalize the average indexes into 0 to 1. We used the min-max normalization method.

$$V_{normal} = \frac{v - v_{min}}{v_{max} - v_{min}}$$

We found the total number of likes and the tip number has a very large distribution range. They are not uniformly distributed but congested into the area near zero. We consider normalizing these two variables into a log scale.

$$V_{normal} = \frac{ln(v) - ln(v_{min})}{ln(v_{max}) - ln(v_{min})}$$

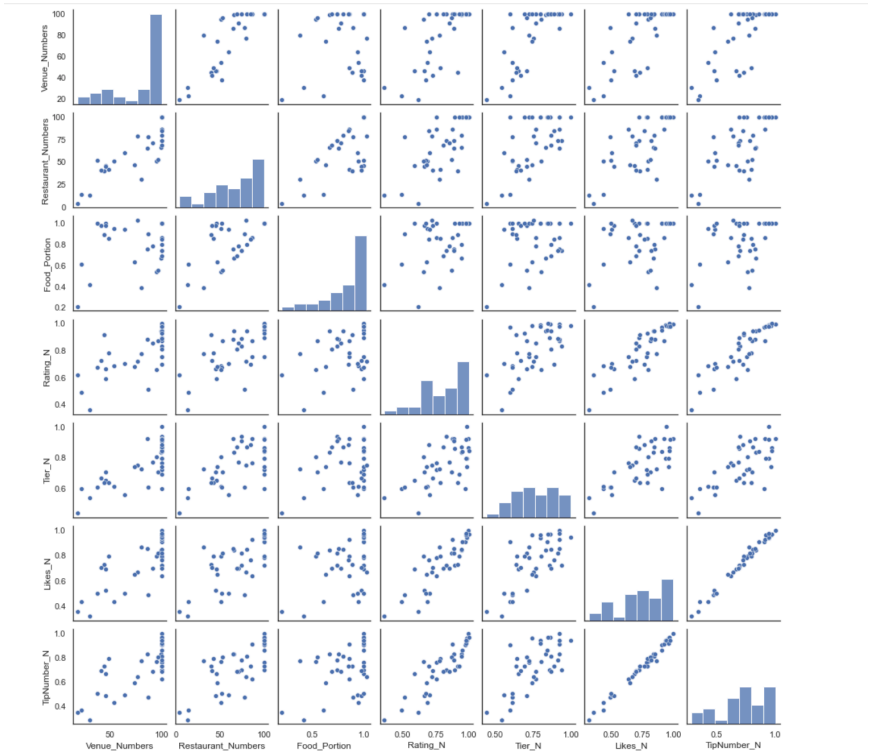

Figure 10: Correlation Matrix plot of all the statistics indexes in the neighbourhood

Then we plot the relative matrix for each variable we showed in last section. Figure 10 records the relative plot for these variables. What we found is that the average number of likes is highly related to the total number of tips from the relative matrix plot. We

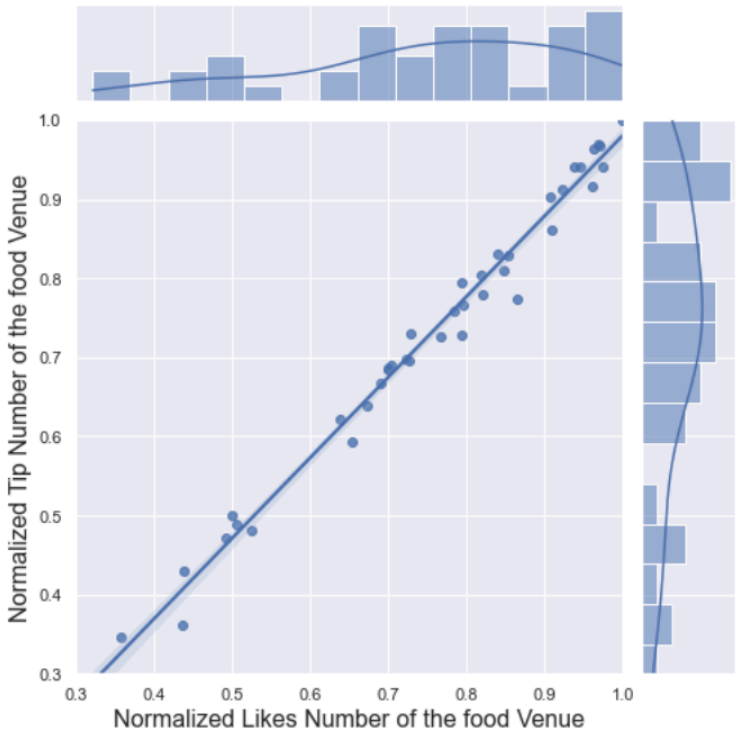

Figure 11: Correlation plot for normalized average like numbers for each venue in the neighbourhood and normalized average tip numbers for each venue in the neighbourhood

further plot this two indexes into Figure 11. The plot shows the highly relativity of these two variables.

This means we can combine the two indexes into one variable by

$$V_{new} = \frac{Like_N + TipNumber_N}{2}$$

4.2 Setting up an index to evaluate the goodness of open a restaurant in certain neighbourhood

we would like to set up an index to criticize the value of setting up a restaurant in a cirtain neighbourhood. We have two models to set up the certain index.

4.2.1 First Model

All the Normalized indexes are used to criticize the restaurant values. The overall index is a linear combination of all the normalized indexes.

$$Q = (k1 * Food\_protion + k2 * Rating_N + k3 * Tier_N + k4 * TipNumber_N + k5 * Likes_N)/(k1 + k2 + k3 + k4 + k5)$$

The key problem is that how to determine k1-k5. If one of the k value is significantly smaller than all the others, it can be considered as zero. If one of the k value is significantly larger than all the others, all the other k can be considered as zero. So they should be in the same scale. let's say that they all range from 1 to 10.

Since the number of likes is highly related to the tip numbers, k4=k5

$$Q = (k1 * Food\_protion + k2 * Rating_N + k3 * Tier_N + k4 * (TipNumber_N + Likes)/2)/(k1 + k2 + k3 + k4)$$

First lets see what happen if all the k values are equal.

Neighborhood	Q
West Village	0.981814
Soho	0.980519
Flatiron	0.964482
Greenwich Village	0.956989
Noho	0.953841
East Village	0.931072
Midtown South	0.928539
Little Italy	0.921789
Midtown	0.918968
Chelsea	0.893608
Chinatown	0.868909
Lenox Hill	0.865935
Murray Hill	0.861305

Tribeca	0.851168
Financial District	0.844792
Civic Center	0.835544
Upper West Side	0.820544
Clinton	0.812141
Upper East Side	0.805524
Turtle Bay	0.795313
Manhattan Valley	0.793329
Tudor City	0.788750
Carnegie Hill	0.774784
Lower East Side	0.768959
Morningside Heights	0.766267
Sutton Place	0.765001
Yorkville	0.753416
Gramercy	0.735289
Manhattanville	0.730460
Central Harlem	0.712062
East Harlem	0.697255
Lincoln Square	0.695368
Battery Park City	0.677209
Hamilton Heights	0.673921
Hudson Yards	0.670152
Inwood	0.670122
Washington Heights	0.626096
Marble Hill	0.524416
Roosevelt Island	0.406718
Stuyvesant Town	0.405523

We can see that the West Village, Soho and Flatiron are the top 3 choice for opening a restaurant. Then we should consider is some indexes more important than others? The most important index for the restaurant is the numbers of customers. The customer number can be reflected as the numbers of likes. The more people like the restaurant, the more customer will visit the restaurant in the unit time.

Lets set up another model that k4=2 and k1=k2=k3=1. In this situation, what the ranking of our choice?

Neighborhood	Q
Soho	0.984415
West Village	0.974078
Flatiron	0.963210
Greenwich Village	0.959383
Noho	0.956983
East Village	0.932743
Midtown South	0.930651

Little Italy	0.930091
Midtown	0.912249
Chelsea	0.898432
Chinatown	0.876058
Murray Hill	0.854827
Tribeca	0.849161
Lenox Hill	0.834919
Civic Center	0.828494
Financial District	0.828147
Upper West Side	0.823577
Clinton	0.803882
Upper East Side	0.786531
Turtle Bay	0.785626
Manhattan Valley	0.780411
Lower East Side	0.773973
Tudor City	0.762014
Carnegie Hill	0.758433
Morningside Heights	0.752365
Gramercy	0.750473
Sutton Place	0.750325
Yorkville	0.728701
Manhattanville	0.720202
Lincoln Square	0.712404
Battery Park City	0.705747
Central Harlem	0.670237
Hudson Yards	0.660829
East Harlem	0.657895
Hamilton Heights	0.638545
Inwood	0.622909
Washington Heights	0.597275
Marble Hill	0.499336
Stuyvesant Town	0.394852
Roosevelt Island	0.386044

We see that the top 3 choice are still Soho, West Village and Flatiron. But in this criterion, Soho becomes the best choice and the West Village lost the top 1 position.

But how about increase the k4 to 3 and keep all other k at 1. We illustrated this rank below.

Neighborhood	Q
Soho	0.987012
West Village	0.968921
Flatiron	0.962362
Greenwich Village	0.960979
Noho	0.959078
Little Italy	0.935625
East Village	0.933857
Midtown South	0.932059
Midtown	0.907770
Chelsea	0.901647
Chinatown	0.880824
Murray Hill	0.850509
Tribeca	0.847824
Upper West Side	0.825599
Civic Center	0.823793
Financial District	0.817051
Lenox Hill	0.814241
Clinton	0.798376
Turtle Bay	0.779168
Lower East Side	0.777316
Upper East Side	0.773870
Manhattan Valley	0.771799
Gramercy	0.760596
Carnegie Hill	0.747532
Tudor City	0.744190
Morningside Heights	0.743097
Sutton Place	0.740540
Battery Park City	0.724773
Lincoln Square	0.723761
Manhattanville	0.713364
Yorkville	0.712224
Hudson Yards	0.654613
Central Harlem	0.642354

East Harlem	0.631655
Hamilton Heights	0.614962
Inwood	0.591433
Washington Heights	0.578062
Marble Hill	0.482615
Stuyvesant Town	0.387738
Roosevelt Island	0.372261

We could see that the top 3 neighbourhood choices are not change comparing to the k4=2. That means we could control the k values with the ratio between 1 and 2. Do not need to control any k values significantly larger than all other k than double values.
Now we have to think about a sequence for the importance of these 4 values. Here we think the most important index is still the numbers of like and the numbers of tips. The second important value is tiers. The third important value is ratings and the least important is the portion of the food in the total venues.
Owing to this consideration, we set k1:k2:k3:k4=1:2:3:4. In this setting, the values of opening a restaurant can be listed as below.

Neighborhood	Q
Soho	0.976926
West Village	0.974078
Flatiron	0.955280
Greenwich Village	0.945727
Noho	0.941565
East Village	0.912479
Midtown South	0.910387
Chelsea	0.904635
Little Italy	0.904100
Midtown	0.896390
Tribeca	0.865752
Murray Hill	0.861832
Civic Center	0.845576
Upper West Side	0.840344
Chinatown	0.835970
Lenox Hill	0.820822
Upper East Side	0.806590
Financial District	0.797310
Turtle Bay	0.789250
Clinton	0.776129
Sutton Place	0.768154
Gramercy	0.765916

Carnegie Hill	0.757348
Manhattan Valley	0.755399
Lower East Side	0.752240
Battery Park City	0.739471
Lincoln Square	0.739051
Tudor City	0.734696
Morningside Heights	0.721362
Yorkville	0.719537
Manhattanville	0.685324
Hudson Yards	0.671361
Central Harlem	0.642896
East Harlem	0.618891
Hamilton Heights	0.600596
Inwood	0.589793
Washington Heights	0.568616
Marble Hill	0.498252
Stuyvesant Town	0.417852
Roosevelt Island	0.398327

We could say that the best choice of opening a restaurant in the Manhattan area is at Soho neighbourhood. The corresponding map plot of hot neighbourhood for opening a restaurant is below(Figure 12).

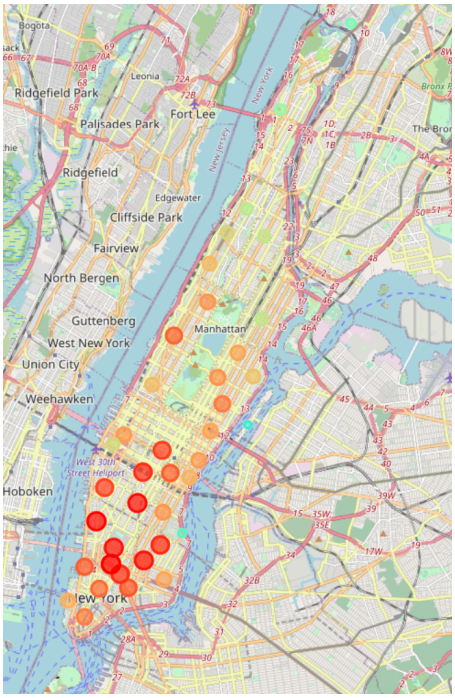

Figure 12: The suggested place to open restaurant from the first model. Red place for more recommendation; blue place for less recommendation.

4.2.1 Second Model

The model above added some personal thinking into the model, which is the k values selection. To eliminate this bios, the new index is set up. This index using the multiple of each normalized variable values.

$$Q = (Food\_protion \times Rating_N \times Tier_N \times (TipNumber_N + Likes)/2)$$

The following chart provides the quality of opening restaurant in this index.

Neighborhood	Q
West Village	0.928157
Soho	0.922302
Flatiron	0.863710
Greenwich Village	0.832922
Noho	0.820061
East Village	0.739700
Midtown South	0.732098
Midtown	0.707035
Little Italy	0.701354
Chelsea	0.636509
Murray Hill	0.545949
Lenox Hill	0.545829
Chinatown	0.526744
Tribeca	0.519175
Financial District	0.487730
Civic Center	0.480643
Upper West Side	0.439491
Clinton	0.420608
Upper East Side	0.410965
Turtle Bay	0.392789
Manhattan Valley	0.379397
Tudor City	0.364865
Carnegie Hill	0.356320
Lower East Side	0.341767
Sutton Place	0.335674
Morningside Heights	0.328933
Yorkville	0.314418
Gramercy	0.275708

Manhattanville	0.261593
Central Harlem	0.229644
Lincoln Square	0.222424
East Harlem	0.207149
Hudson Yards	0.199839
Hamilton Heights	0.182828
Battery Park City	0.178840
Inwood	0.173147
Washington Heights	0.135977
Marble Hill	0.071456
Roosevelt Island	0.024966
Stuyvesant Town	0.020213

The benefit of this index is that, all the factors has been considerate. If one of the factor is significantly lower than the average value, the index would become very small. It is also a reflection of Cannikin Law: only a neighbourhood with all the considerate factors in high values can be evaluated as a good selection of opening a restaurant. Figure 13 plots the map of hot neighbourhood to open restaurant in manhattan area within this model.

5 Conclusion

The current work established two evaluation models to solve the problem that which neighbourhood is the best choice to open a restaurant in the manhattan area. The models do not specify the food kind and the price level of the restaurant but only provide a general area value for each neighbourhood on this problem. The higher value a neighbourhood gain, the more likely to success within running a restaurant in this neighbourhood. The first model think that the Soho is the best neighbourhood to open. The second model think the West Village is the best choice. The second model is more

conservative comparing to the first model. If one want to minimize the risk of failure, the West Village would be the first selection.

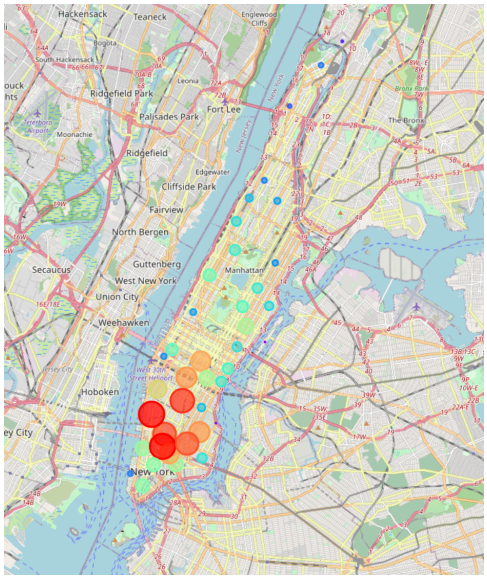

Figure 13: The suggested place to open restaurant from the second model. Red place for more recommendation; blue place for less recommendation.

6 Future Work

The future work could be on specifying the model details and let the model fit for all kind of restaurant. It can provide different options for different investors by inputing the specific restaurant food category and restaurant price level.