Business Understanding

A business owner wishes to establish the best place to open a new Indian Restaurant in New York City. The aim is to determine if there is a certain clustering of Indian restaurants and if there are any potential gaps in areas where there could be football. However they are that if there are many other similar restaurants in the region, customers may not be so inclined to visit.

Data Sources

For this K-clustering will be used to analyse the various regions of the city with regards to Indian Restaurants. The data will be obtained from foursquare api and the maps as a .json files from https://cocl.us/new_york_dataset and https://maps.princeton.edu/catalog/nyu-2451-34561

Methodology

The map data was downloaded and sorted into columns based on Borough, Neighborhood, Latitude, and Longitude. The restaurant data (location, name etc) was obtained from foursquare api with the condition that the categoryId was for Indian restaurants within a search radius of 750m for each neighborhood. The data set was cleaned to remove duplicate values and sorted by the Neighborhood Column alphabetically. The distribution of restaurants and neighborhood locations was then plotted with folium.

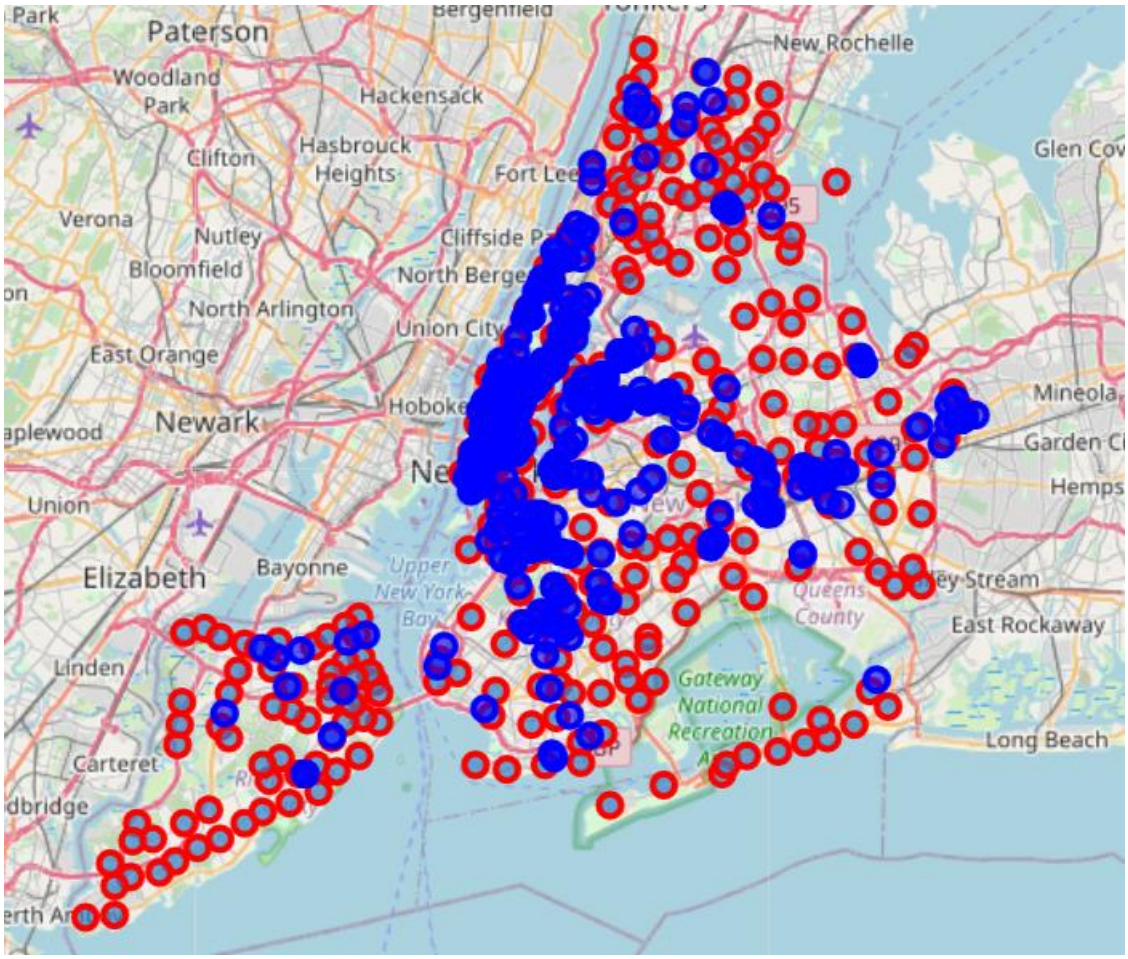

Fig1. Distribution of Indian restaurants in NYC. Restaurant markers are given in blue and neighborhood locations in red.

The data was then grouped by Neighbourhood and the frequency count for each neighbourhood calculated and plotted as a bar chart in Fig2.

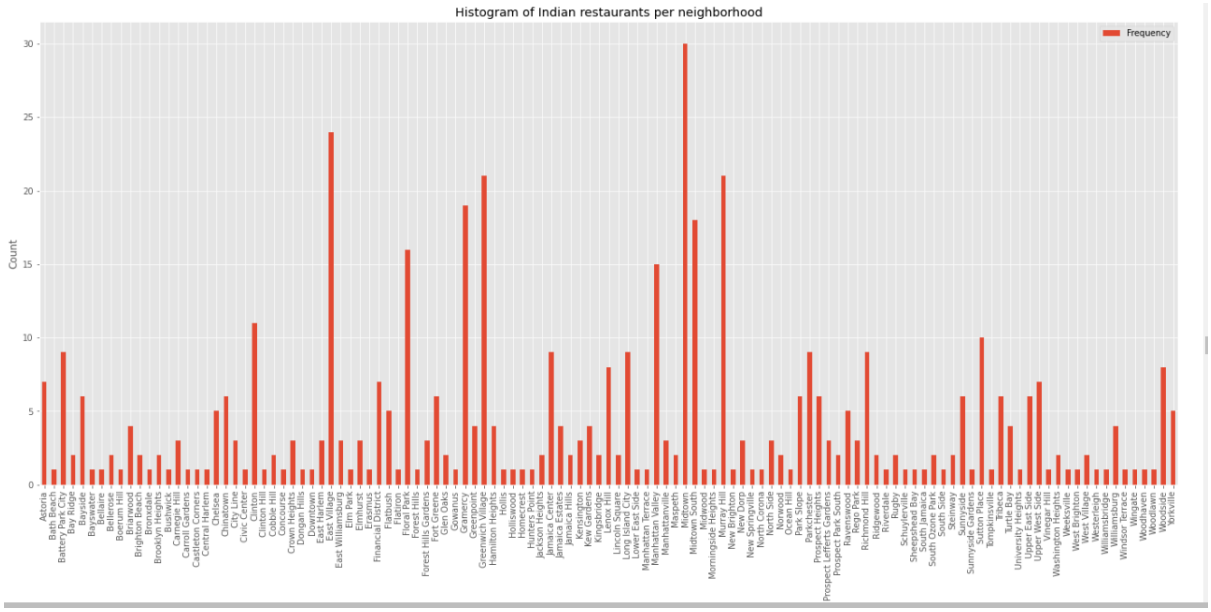

Fig2. Bar chart showing the frequency distribution of Indian restarants for each neighborhood

To visualise this data in a better way, polygon data for each neighborhood was obtained and a Cholorpleth map was plotted to show the frequcy distribution as a heat map for each neighborhood arcross the whole city and to potetially identify patterns by eye.

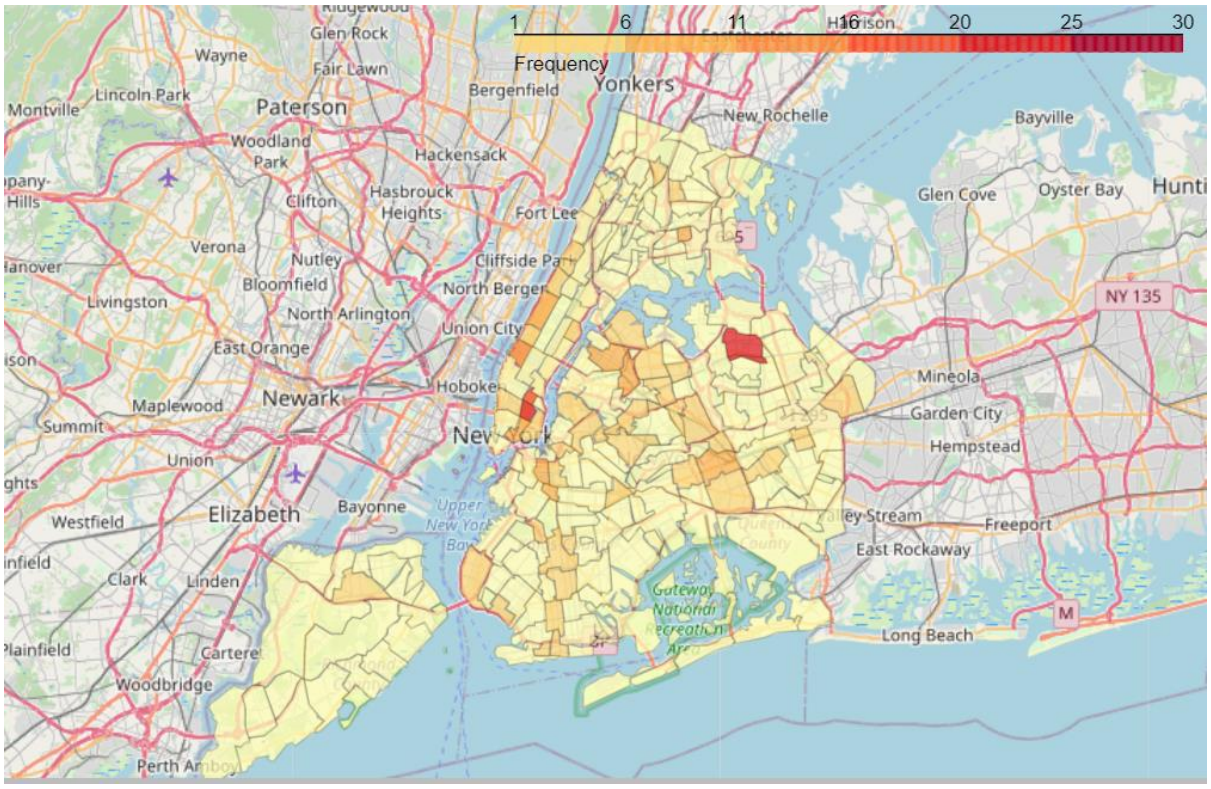

Fig3. Heat distribution on the frequency of Indian restaurant for each neighborhood

K-means clustering was then used to identify any local patterns in the data and establish any postential gaps

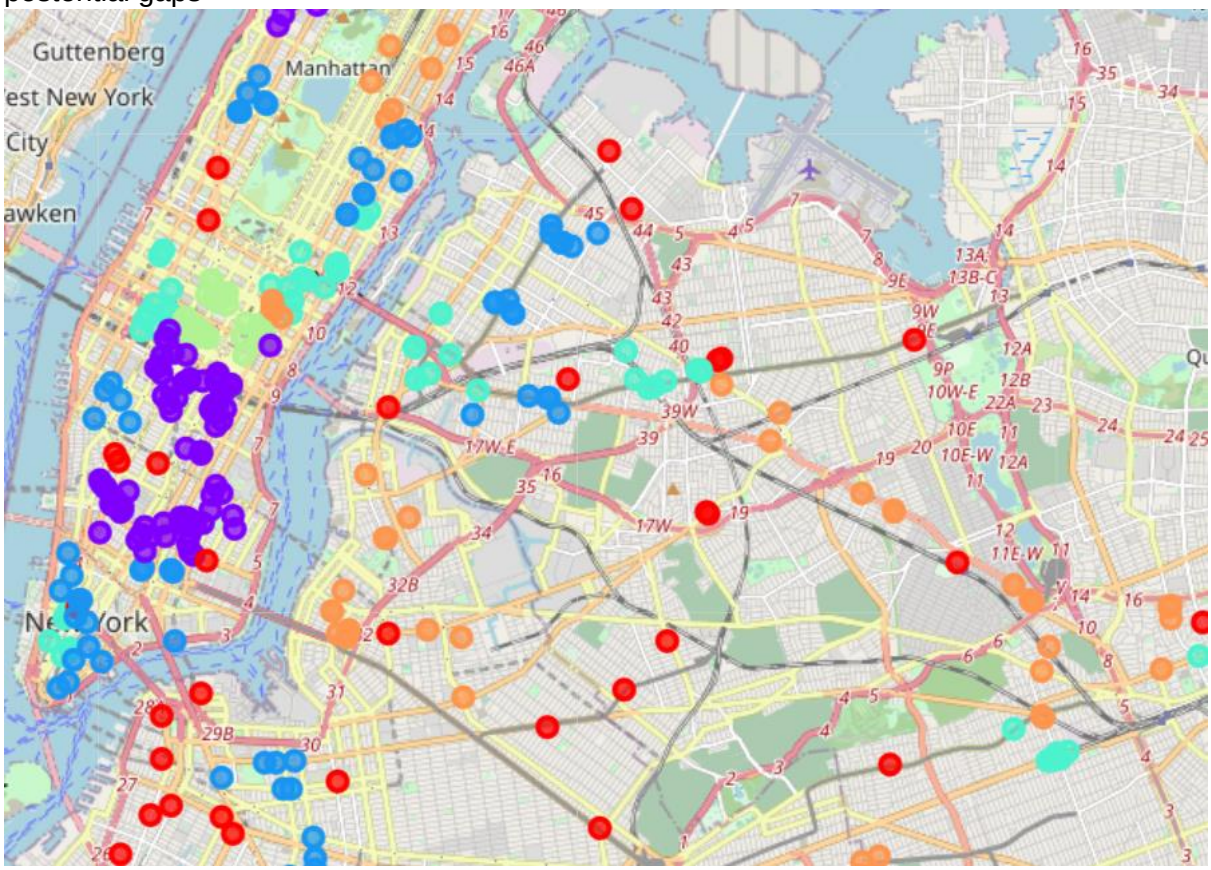

Fig 4. K-means plot of the data. Regions on a local scale where there are fewer restarants can be identified

Red gives the indication that there are less restaurants in the vicinity and purple a high density of nearby restaurants with the other colours falling between. On a local scale, this would be useful to identify any regions where there may be less restaurants, but not too few that there would be no customers at all.

To study this data in a different way, dBScan was then used to identify any further patterns, and regions of different clustering can be seen.

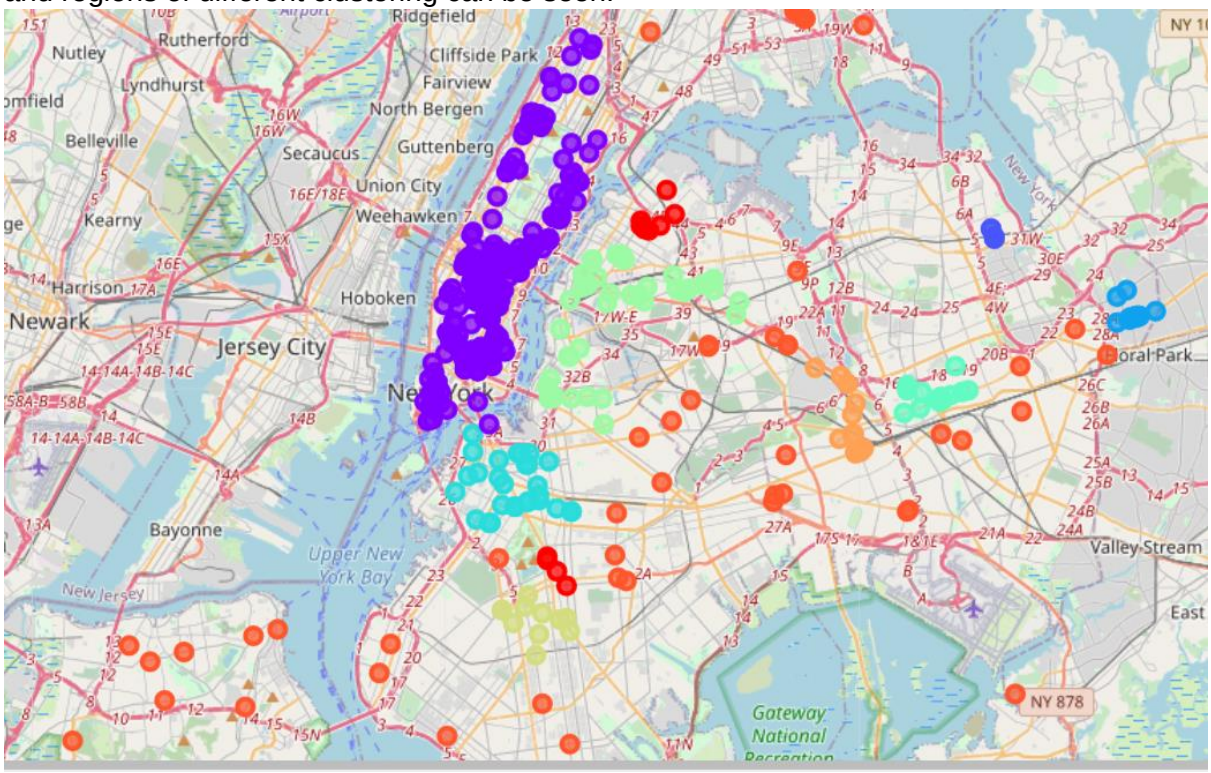

Fig 5. dBScan plot of the data. More advanced clustering patterns can be seen (ellipse=0.2)

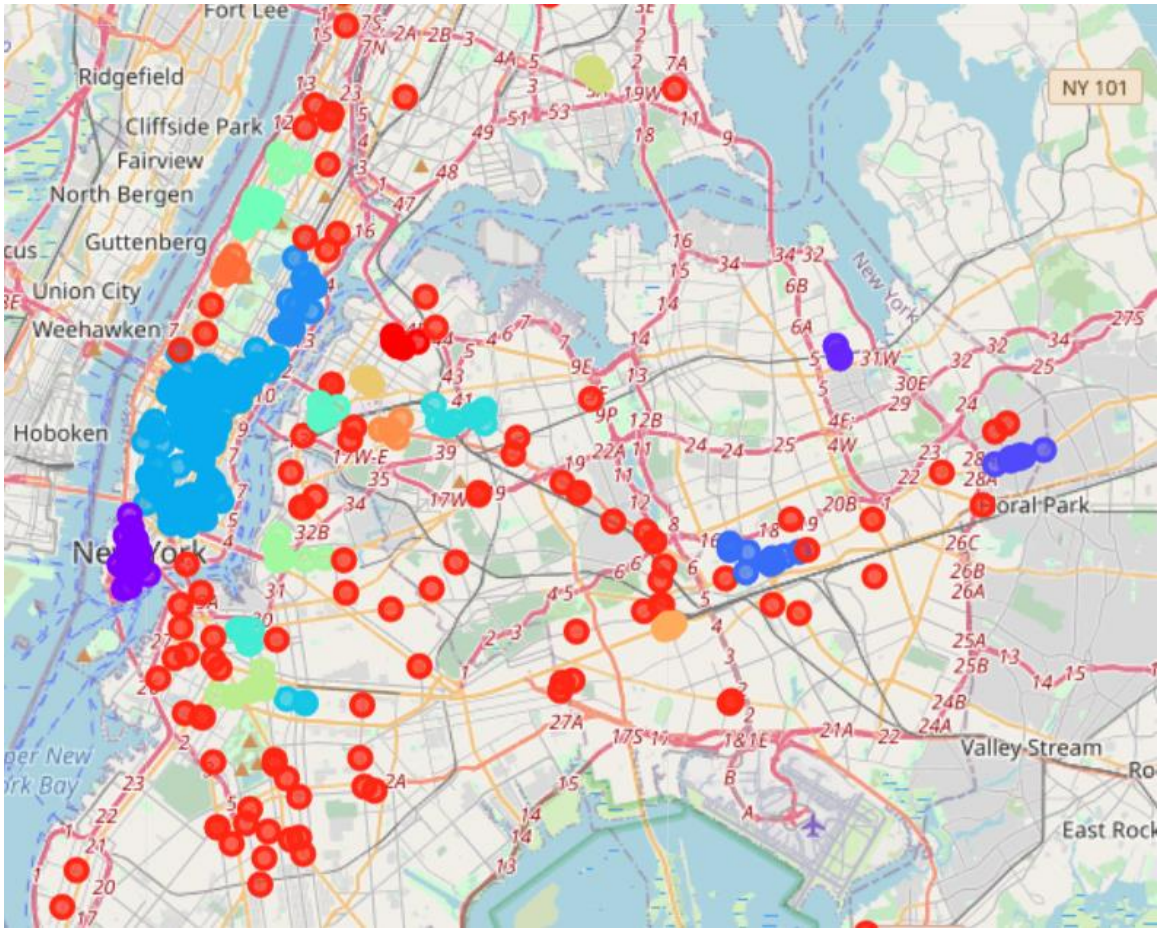

Fig 6. dBScan plot of the data. More advanced clustering patterns can be seen (ellipse=0.1)

Results and discussion

From the data it can clearly be seen that Manhattan has the highest density of restaurants, from the bar chart, Midtown, East village, Greenwich Village and Murray Hill all have more than twenty Indian restaurants in them. The choropleth map gives an indication on a broader scale on the distribution of restaurants on a visual scale. The k-means clustering show patterns on a local scale, and larger scale. It is clear that while some regions do have a high density of Indian restaurants in the vicinity, there are also sparse intercluster areas where perhaps it is not favourable to open a business. However on an intracluster bases games can be identified for example in Fig 6,and 7 one can see some potential areas.

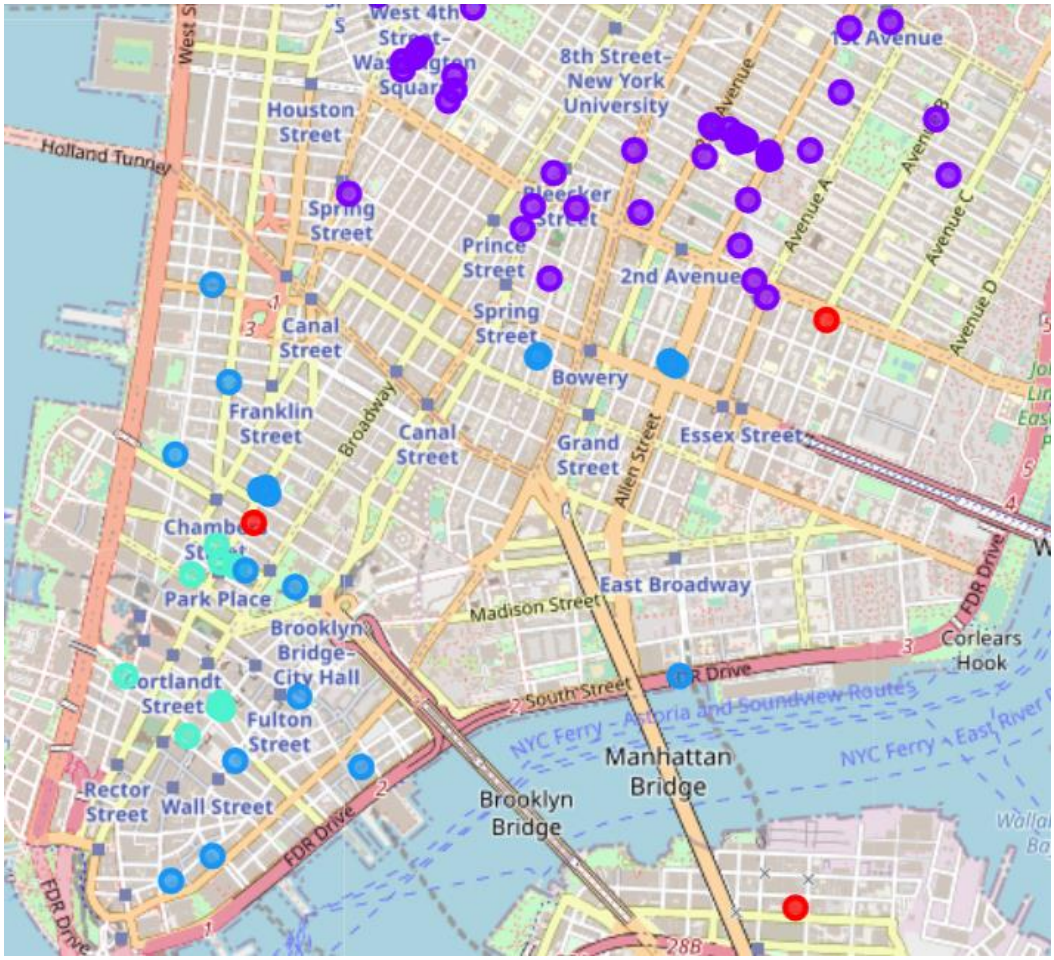

Fig 7. K-means plot zoomed in on lower Manhattan. Although the purple area is very dense, there are intercluster gaps

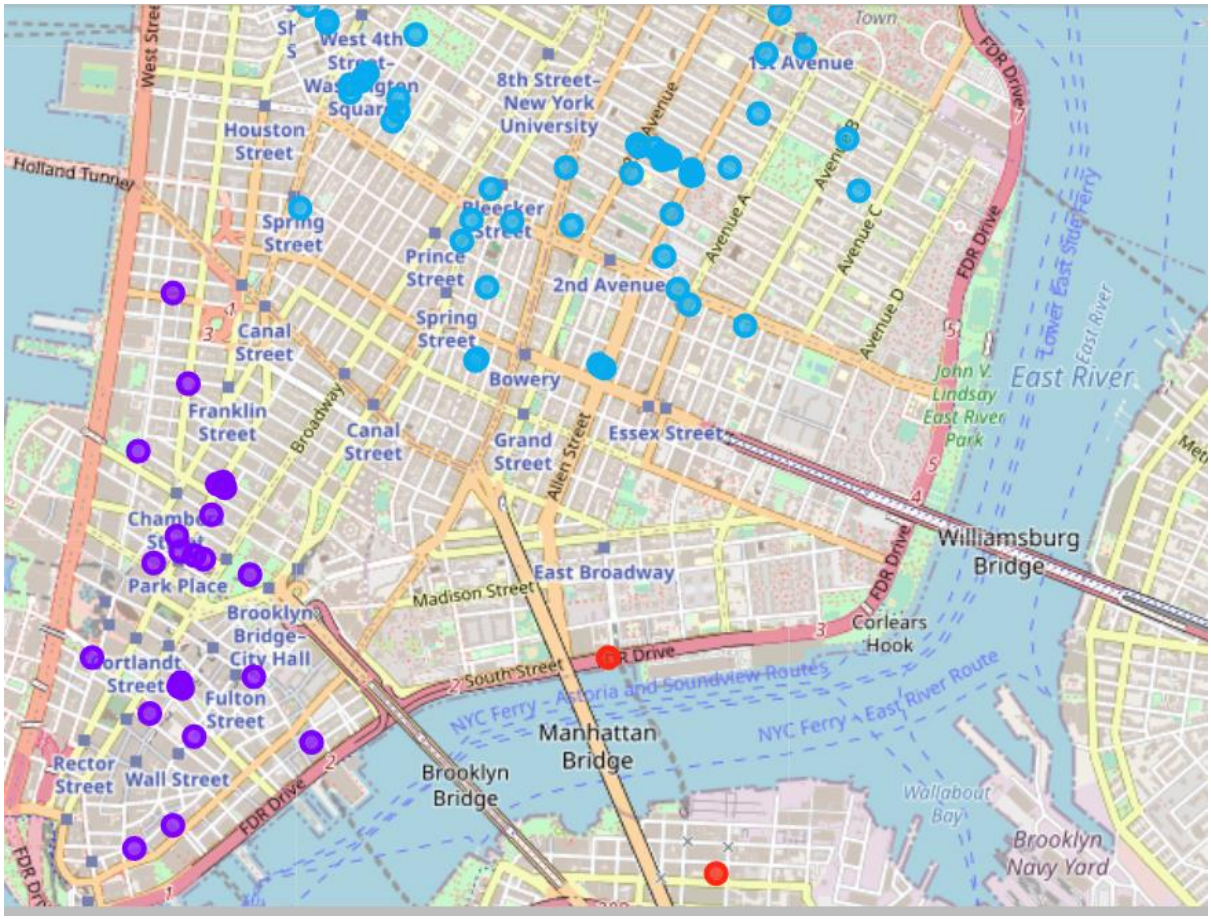

Fig 8. Same as Fig7. But for dBScan. The two clusters are seen separated

Conclusion

This exercise is useful in determining the distribution of restaurants and identifying the best and worst places to set up based on nearest neighbours However, it is clearly just an initial step and more rigorous analysis of the quality of restaurants, population density in a given area and prices would of course all have to be considered but are beyond the scope of this exercise.