Coursera IBM Data Science Professional Certificate, Capstone Project:

Geospatial Analysis of Poverty and Wealth Distributions in Manhattan Island, New York City

Sunil Pakiri

21 June 2020

1 Introduction / Business Problem

New York city is often described as the financial capital of the world. It is also one of the cities in the United States with the highest median household income. Within New York city itself, Manhattan Island is not only the richest borough, also the most famous. Manhattan is synonymous with New York City and is described as the cultural, financial and entertainment capital of the world. Manhattan has been the setting for numerous movies, books and TV shows and hosts the world's largest stock exchanges.

However, as with all cities, Manhattan is a place with high income inequality. Understanding the income and social inequalities and how they relate geographically would aid in better addressing the inequality concerns.

Non-Governmental Organisations (NGOs) are often on the frontlines of the battle for addressing social ills. However, the majority of the NGOs and aid organisations are not well funded. This means that the funding available must be put to the most effective use possible.

Income inequality and poverty are often fought on many fronts, such as minimum wage laws, better access to facilities, access to job opportunities, etc. It is also found that the poverty and income inequality is geographically segregated ("the wrong side of the tracks" problem). With this in mind, the neighbourhoods of Manhattan Island were studied to determine the geospatial distributions of wealth, quality of life and poverty.

The similarity of rich and poor neighbourhoods in terms of facilities and venues around each neighbourhood was also studied. The Foursquare API data was used to determine which venues/facilities are present in each neighbourhood and how they correlate to income levels.

This study aims to understand the differentiating factors within these neighbourhoods and provide guidance to NGOs to more effectively address the ills of inequality and poverty, with the limited funds available.

2 Data

Due to the nature of the study, various data sources were required. The primary data sources used were as follows:

- Geographic Data (GeoJson and Shapefiles)
- Census Data – Income & Population Data
- Foursquare Venues Data

2.1 Geographic Data

The neighbourhood names and neighbourhood boundaries for Manhattan Island are not officially defined. This is especially true for small neighbourhoods (some consisting of a single block) which undergo changes occasionally. Therefore, using the common neighbourhood names proved difficult, especially with regard to mapping economic data to the neighbourhoods.

During the 2010 Census, the neighbourhoods were divided into Census Tracts. Census Tracts are statistical areas built from census block-groups that are designed to have an ideal size of 4,000 residents, with range of 1,200 to 8,000. The 2010 Census data was directly mapped to the geographical Census Tracts. This ensured a complete dataset that was used to determine the geospatial distributions of various economic and quality of life indicators.

However, due to the small sample sizes of the Census Tracts and the large number of tracts overall (2168 for all of New York City), it was decided to use an aggregated dataset for the neighbourhood clustering study. The aggregation chosen was the New York City Community Districts. Community Districts are mandated by the city charter to review and monitor quality of life issues for New York City neighbourhoods. Each Community District comprises of several neighbourhoods grouped together. Also, since they are directly mandated to review and monitor quality of life issues in New York City neighbourhoods, Community Districts are the ideal aggregation of the neighbourhoods for this study, since results obtained herein can be actioned by the Community Districts.

The geographic data used was obtained from the New York University Spatial Data Repository. Two sets of Geojson data was used from this repository, viz. **2010 New York City Community Districts**[1] Geojson data and the **2010 New York City Census Tracts** Geojson data.[2]

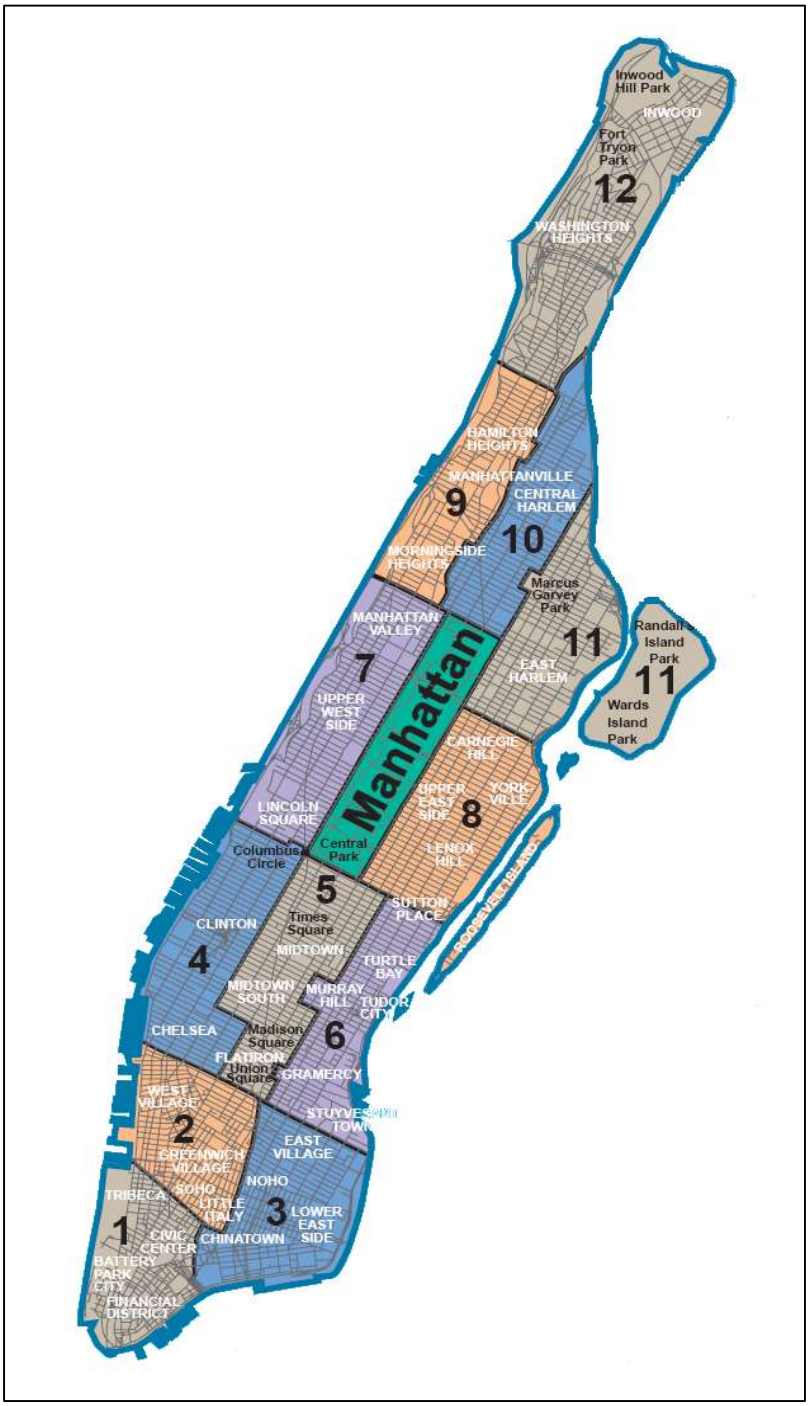

Figure 1: Community Districts of Manhattan Island

[1] https://geo.nyu.edu/catalog/nyu-2451-34159
[2] https://geo.nyu.edu/catalog/nyu-2451-34505

2.2 Census Data

Socio-economic data was obtained from the **DATA2GO**[3] datasets developed by the **Measure of America**[4] initiative by the **Social Science Research Council**[5]. The data was downloaded in Excel format. The excel spreadsheets had to be cleaned prior to being imported into the Jupyter Notebook. This was due to formatting and other issues. The imported datasets were cleaned and filtered for the Manhattan data. This dataset was used to determine the economic and social indicators for the various areas in Manhattan. Choropleth maps showing how the various indicators are distributed were drawn.

2.3 Foursquare API Data

The Foursquare API was used to get the most common venues of each Community District in Manhattan. This data was used in conjunction with the socio-economic data to determine the similarities between high income neighbourhoods and low-income neighbourhoods.

[3] https://www.data2go.nyc
[4] http://measureofamerica.org/
[5] https://www.ssrc.org/

3 Methodology

3.1 Exploratory Data Analysis

3.1.1 GeoJson Data

The GeoJson data for the Census Tracts as well as the Community Districts was plotted using Folium. The Community Districts were overlaid onto the Census Tracts. This showed how the individual Census Tracts conformed to the various Community Districts. The coordinates for Manhattan Island was obtained using Nominatim and geolocator.

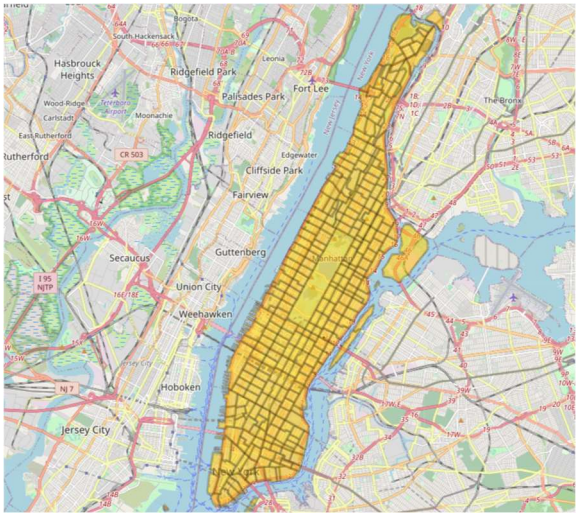

Figure 2: Manhattan Island Census Tracts

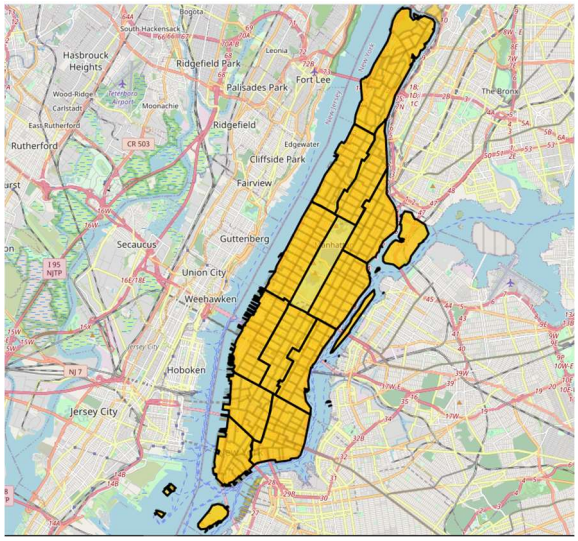

Figure 3: Community Districts Overlaid onto Census Tracts

A problem encountered during the exploratory data analysis was that the Community Districts did not have defined midpoints or locations. The location data for each Community District was important since without it, we would not be able to obtain venues data using the Foursquare API.

The midpoints were calculated using the Geopandas library. The shapefile for the Community Districts was imported into the notebook using Geopandas and the centroids for each of the polygons was calculated. The latitude and longitude data was then obtained by converting the geometry point data using the Coordinate Reference System (CRS).

The midpoints were added to the Community Districts map, thus completing the geographical data analysis.

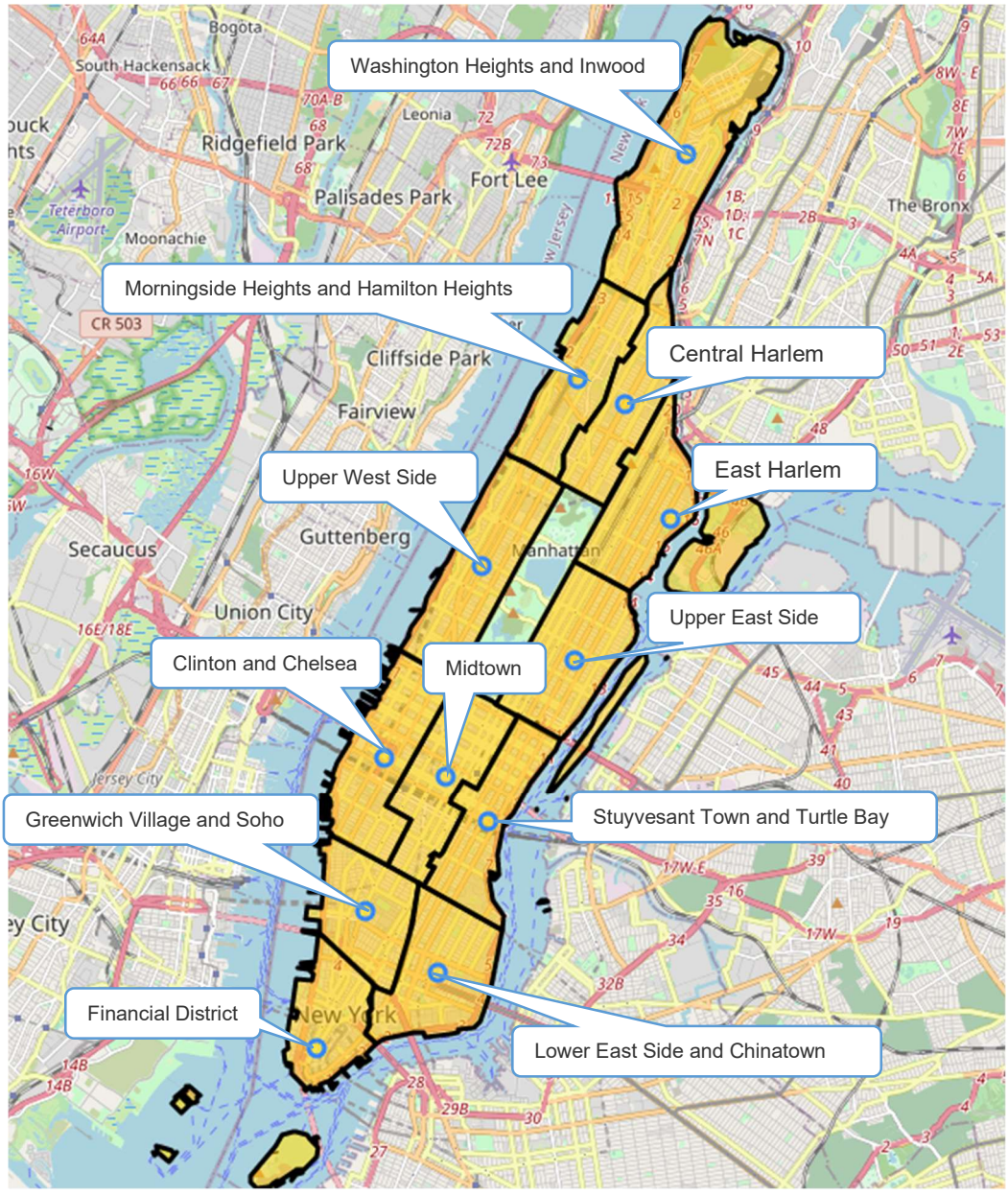

Figure 4: Community Districts Names and Locations

3.1.2 Census Data

The large volume of extremely detailed census data meant that the it was not feasible to analyse the entire dataset. The data was subsetted and into the following categories:

- Wealth
- Education
- Demographics
- Housing

The key indicators in each category was further analysed (e.g. Median Household Income in the Wealth category). Choropleth diagrams were generated using the Census Tracts and overlaid by Community Districts.

Prior to analysing the geospatial distributions of the Census Data, correlations between Median Household Income and various other features was determined. Table 1 below shows the correlations between various features and Median Household Income. From the table it is evident that one of the best predictors of success is education, with "Having completed at least a bachelors degree" highly positively correlated to Median Household Income. It should also be noted that "Did not complete High School" is highly negatively correlated to Median Household Income.

Table 1: Correlation Coefficients for Features against Median Household Income

Features Correlated against Median Household Income	Correlation Coefficient
Did Not Complete High School (% of adults 25+)	-0.69
Single Mother With Children (% of households)	-0.57
Completed High School or High School and Some College (% of adults 25+)	-0.55
Single Father With Children (% of households)	-0.27
Gini Coefficient of Income Inequality	-0.15
Married With Children (% of households)	0.12
Prime Age Adults (% of total population ages 25-54)	0.29
Married (% ages 15+)	0.39
White (% of total population)	0.56
Completed at Least High School (% of adults 25+)	0.69
Completed at Least Bachelor's Degree (% of adults 25+)	0.77
Median Household Income ($)	1.00

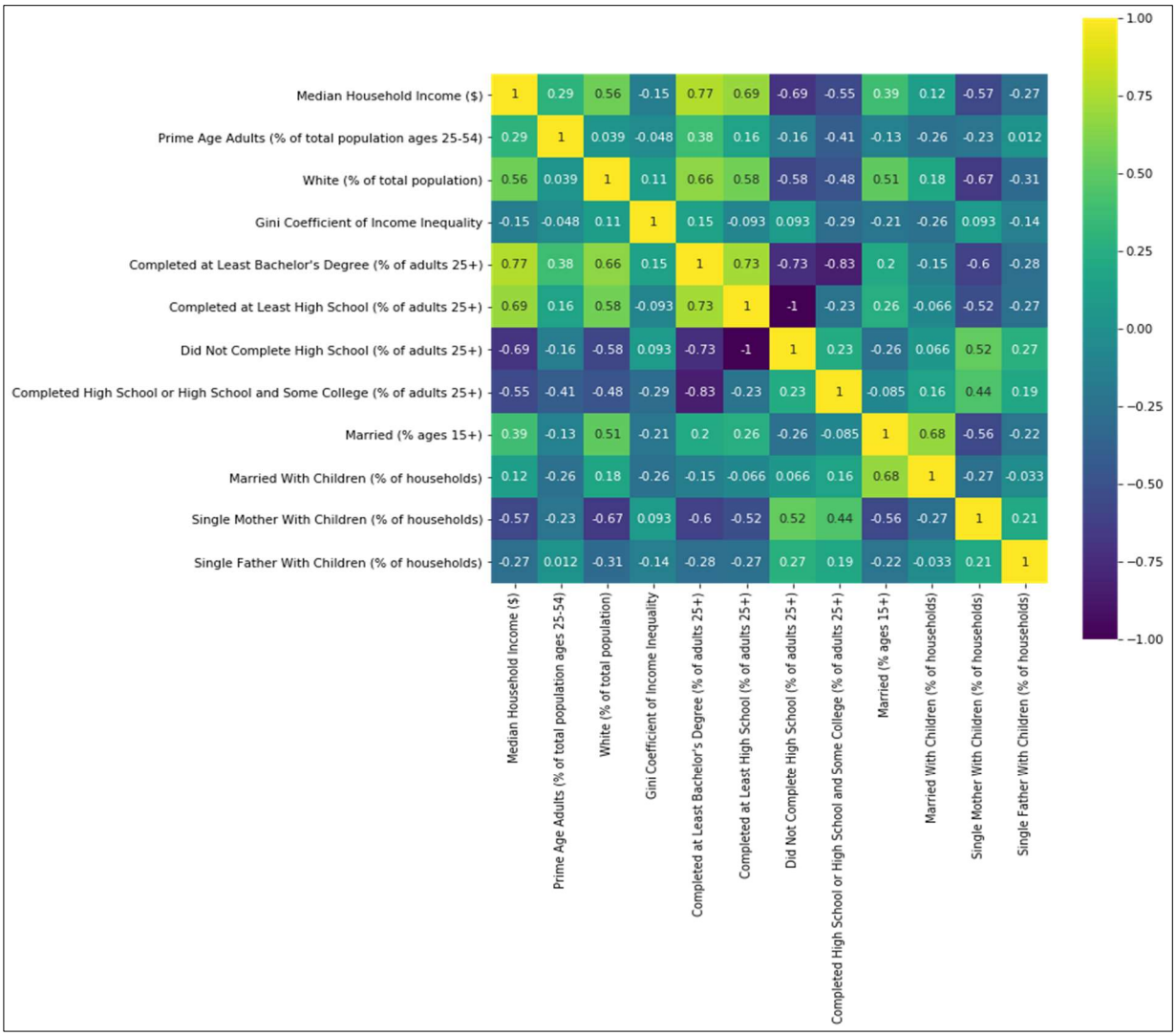

Figure 5: Correlation Matrix of Selected Census Data

3.1.3 Foursquare API Data

The midpoint location of each Community District was used as the locations for the Foursquare API calls. Venues within a 500-metre radius of each Community District were obtained. The data was cleaned to obtain all the unique Venue Categories which were then one-hot encoded to prepare for the clustering analysis. K-means clustering was used to cluster the data, using 5 clusters. The clusters were then plotted onto the Community Districts map and showed that the wealthy neighbourhoods were in fact dissimilar to the poorer neighbourhoods in terms of the different types of venues.

4 Results

4.1 Geospatial Results

4.1.1 Wealth Distribution

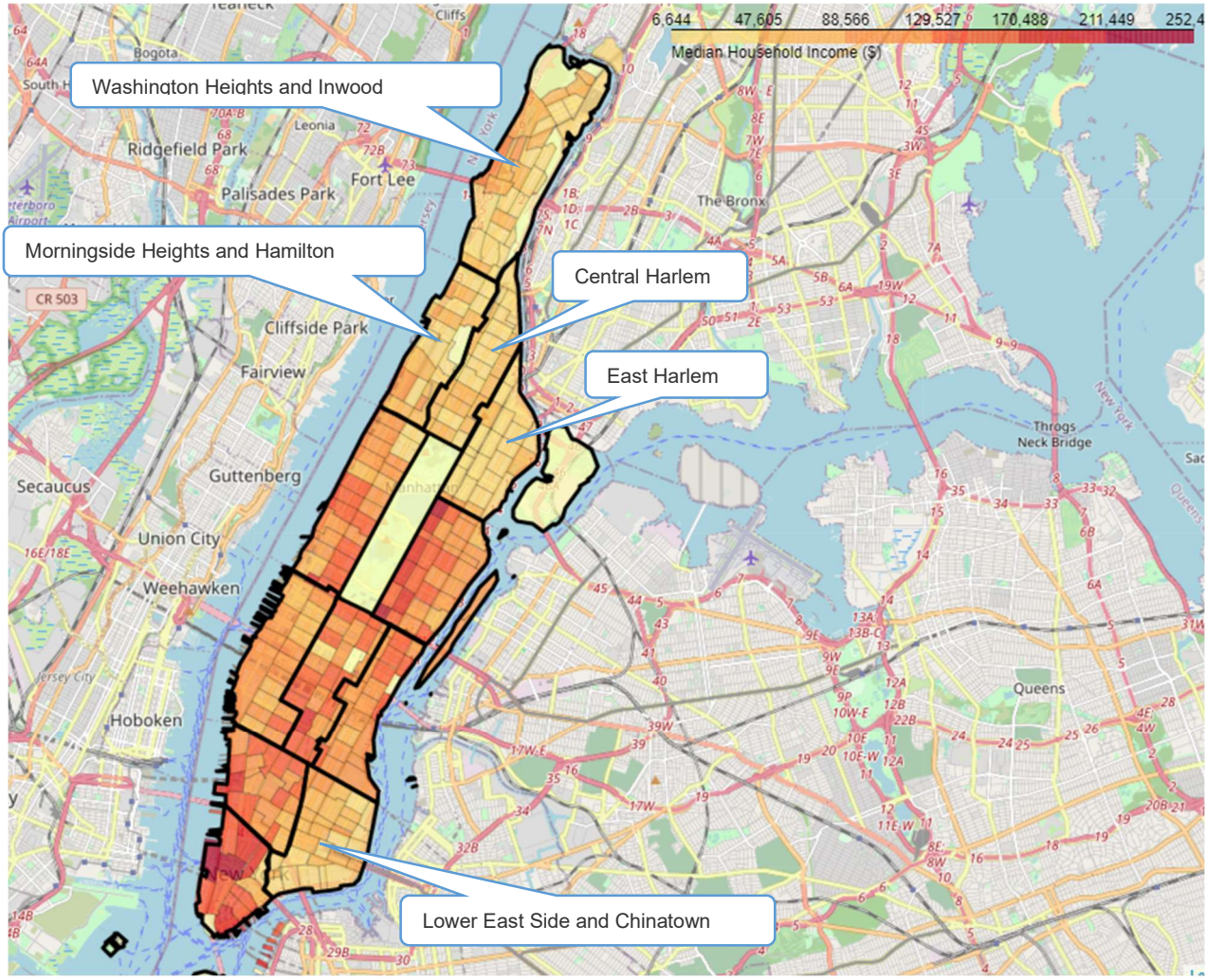

Figure 6: Median Household Income ($)

From the above figure, it is evident that the income distribution is skewed to the Southern parts of Manhattan, with the lower income areas clustered towards the Northern areas (Washington Heights and Inwood, Morningside Heights and Hamilton Heights, Central Harlem, East Harlem), with the exception being “Lower East Side and Chinatown”.

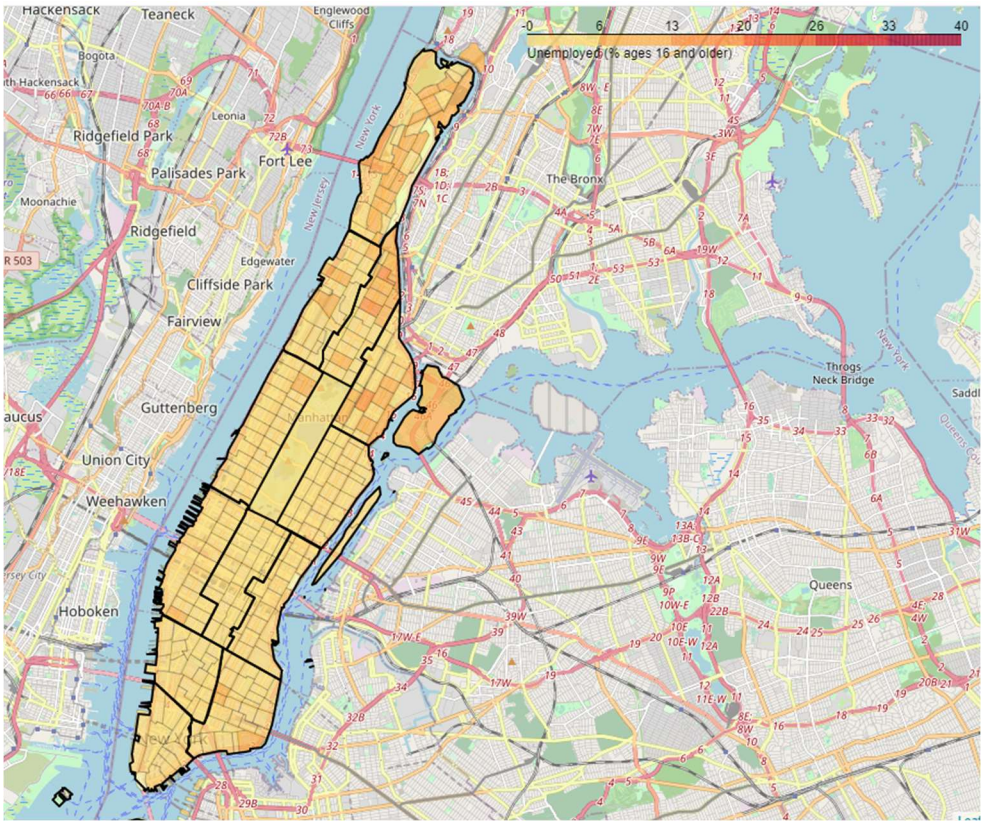

Figure 7: Unemployment Rate

The unemployment rate follows the same geospatial distribution as the Median Household Income, with the highest unemployment rates being in the northern part of Manhattan.

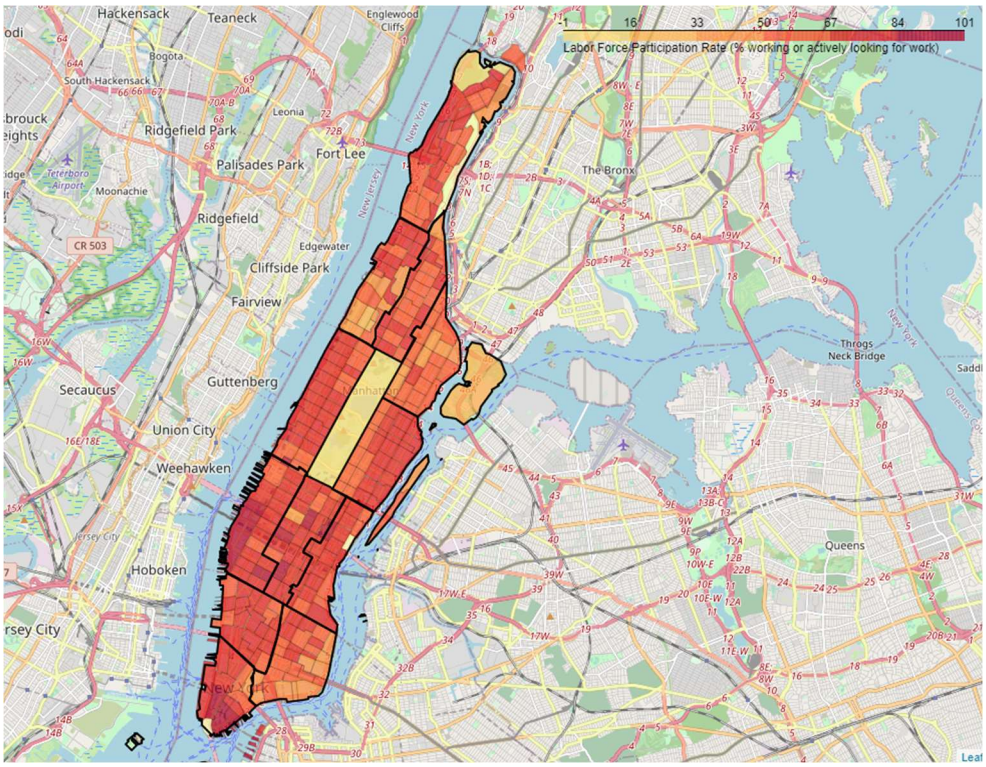

Figure 8: Labour Force Participation Rate

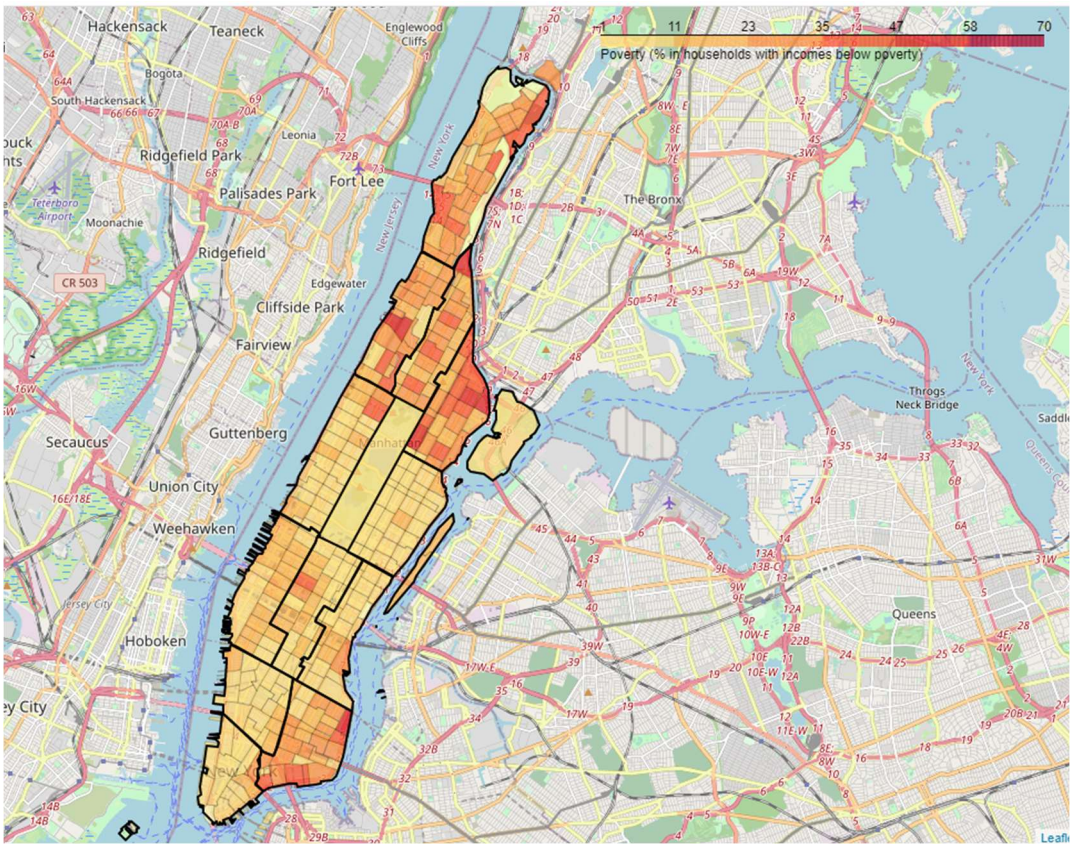

Figure 9: Poverty Rate

The poverty rate follows the same distribution as the Income Distribution. However, it should be noted that there are pockets of poverty within the wealthy Community Districts as well. This could be due to older rent-controlled buildings being within those areas.

4.1.2 Quality of Life

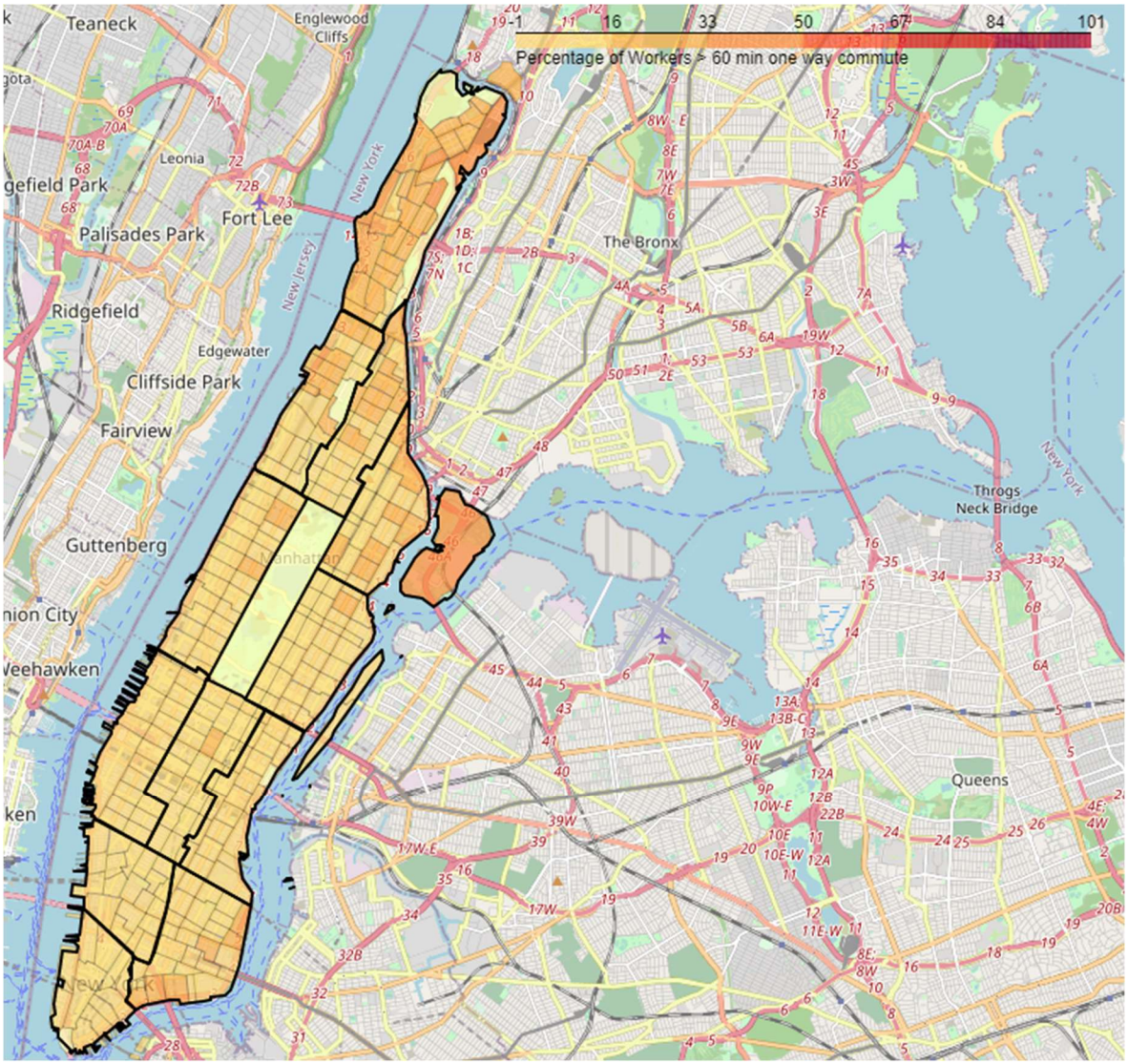

Figure 10: Percentage of workers travelling more than 60 min one way

One of the indicators of quality of life is the average commute to work times. Figure 10: Percentage of workers travelling more than 60 min one way shows that the commute times are also greater in the lower income areas.

4.1.3 Education

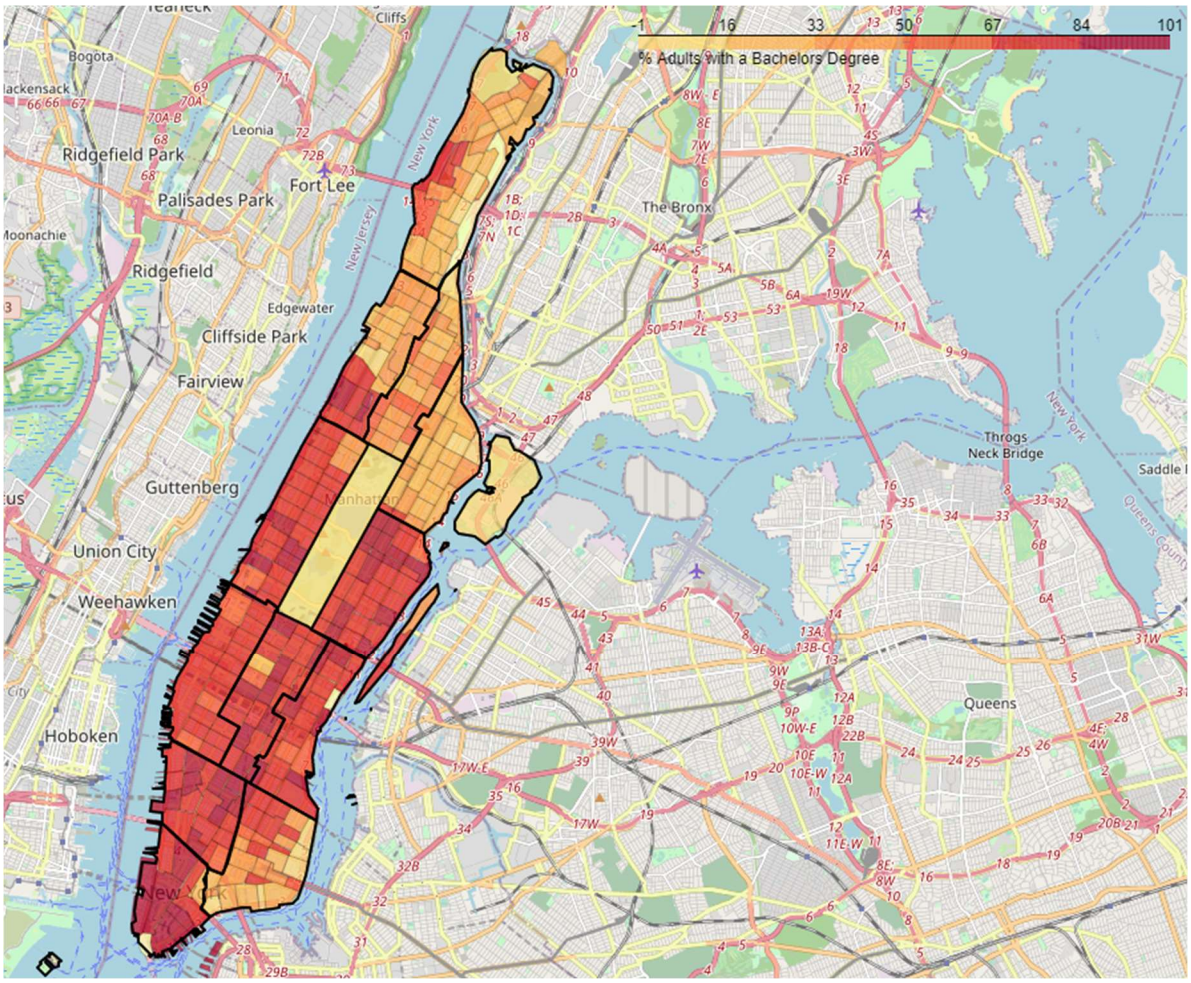

Figure 11: Percentage Adults with a Bachelors Degree

As discussed in the correlation analysis, it is evident in Figure 11 that the level of education impacts the earning potential of the individual. The low income areas have a significantly lower amount of individuals with a tertiary education.

4.2 Cluster Analysis

Cluster analysis of the Community Districts showed that the low income and high-income areas are dissimilar in terms of venues available. It is also important to note that the number of venues returned in the lower income areas are lower than 50% of the venues requested. This means that the lower income areas cannot sustain a large number of venues, due to the poor economic conditions in the areas.

Table 2: Clusters and Total Venues returned from Foursquare

Community District	Cluster	Number of Venues
Clinton and Chelsea	0	84
Financial District	0	100
Midtown	0	100
East Harlem	1	65
Greenwich Village and Soho	1	100
Lower East Side and Chinatown	1	100
Stuyvesant Town and Turtle Bay	1	100
Upper East Side	1	100
Upper West Side	1	95
Washington Heights and Inwood	2	39
Morningside Heights and Hamilton Heights	3	39
Central Harlem	4	46

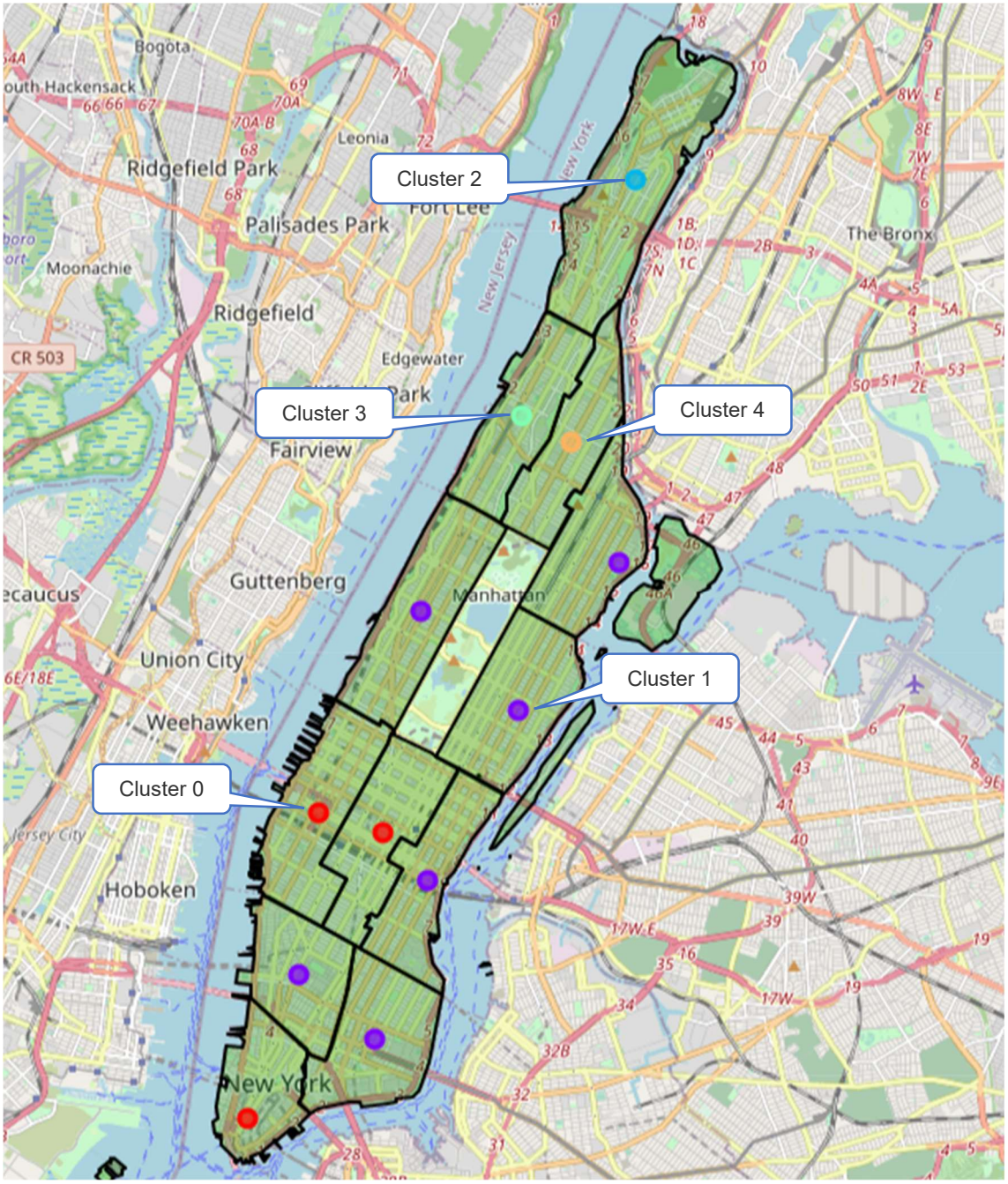

Figure 12: Community District Clusters

4.2.1 Cluster 0

Cluster 0 has hotels as the top most common venue, indicating that it is the tourist/business clusters.

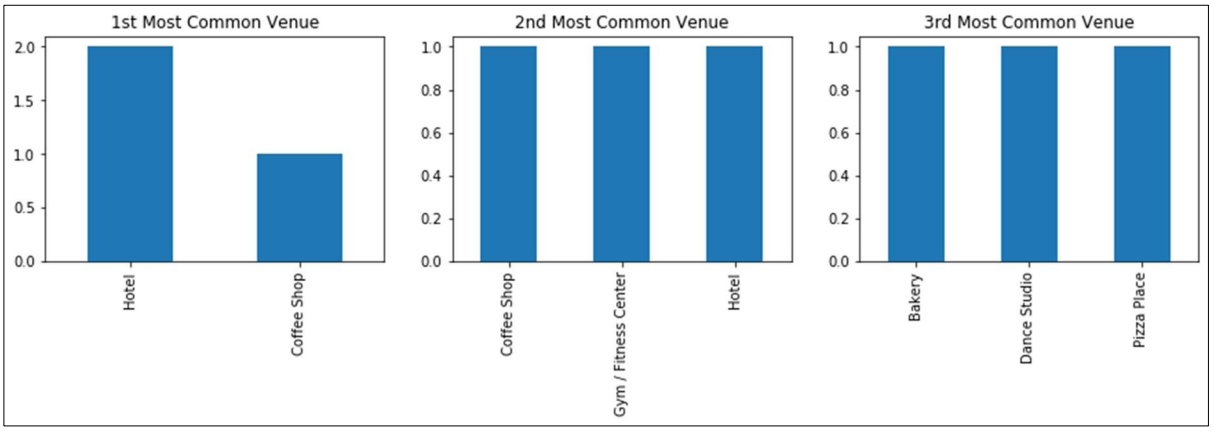

Figure 13: Most Common Venues in Cluster 0

4.2.2 Cluster 1

Cluster 1 has primarily restaurants and bars as its most common venues, indicating that these are the entertainment areas.

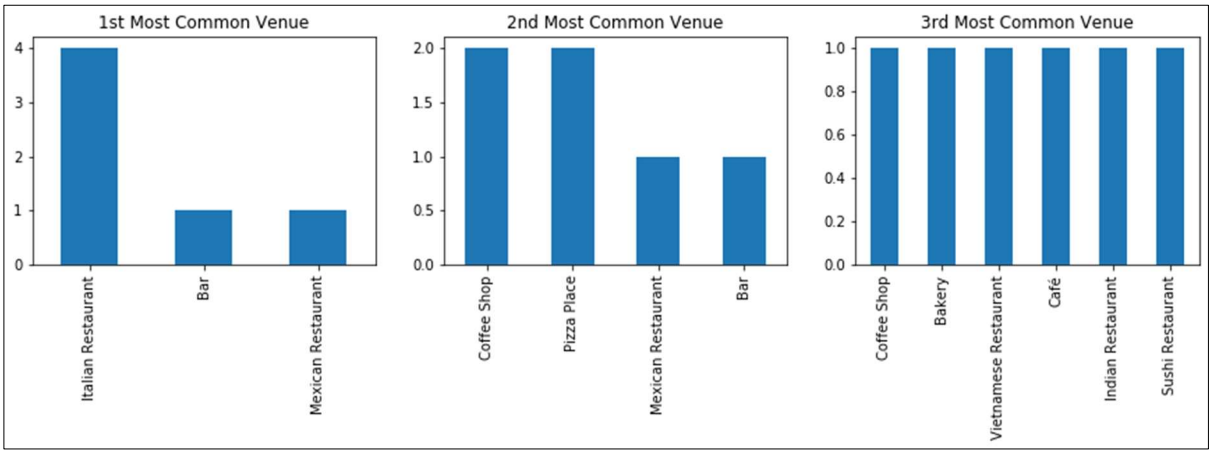

Figure 14:Most Common Venues in Cluster 1

4.2.3 Cluster 2

Cluster 2 is the low-income area. This is highlighted by the fact that the most common restaurant is a pizza place, whilst the 3rd most common is a bar.

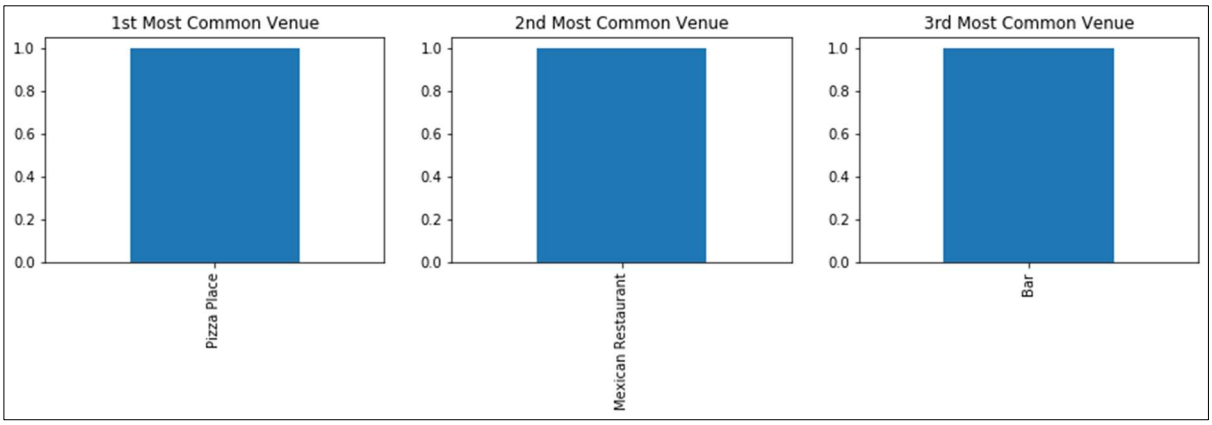

Figure 15:Most Common Venues in Cluster 2

4.2.4 Cluster 3

Cluster 3 is also in the lower income area; however, it is the only cluster where a supermarket is in the top 3 most common venue. This indicates that it is more of a residential area than a business area.

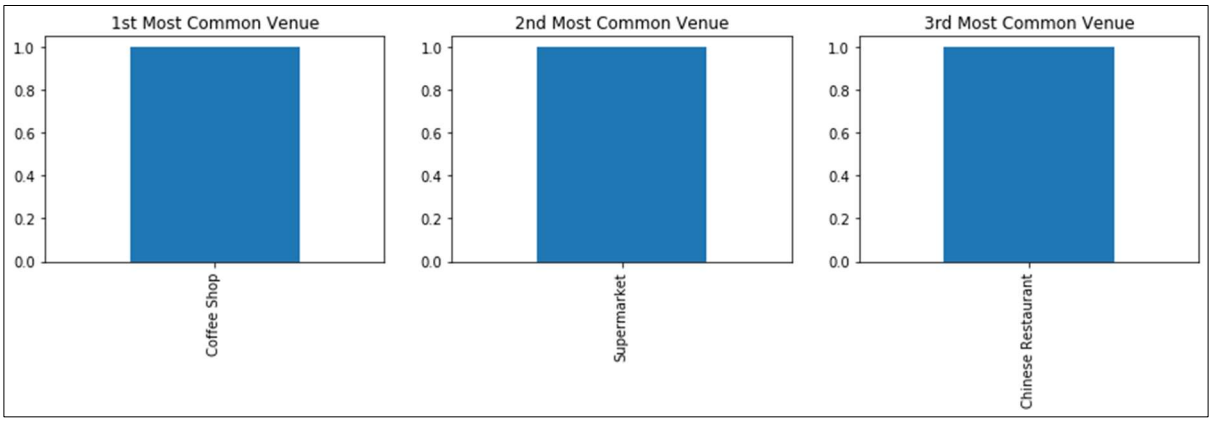

Figure 16:Most Common Venues in Cluster 3

4.2.5 Cluster 4

The thing that sets Cluster 4 apart from the others is the most common venue is an African restaurant. This ties in with the demographics of the area of Central Harlem, being majority African-American.

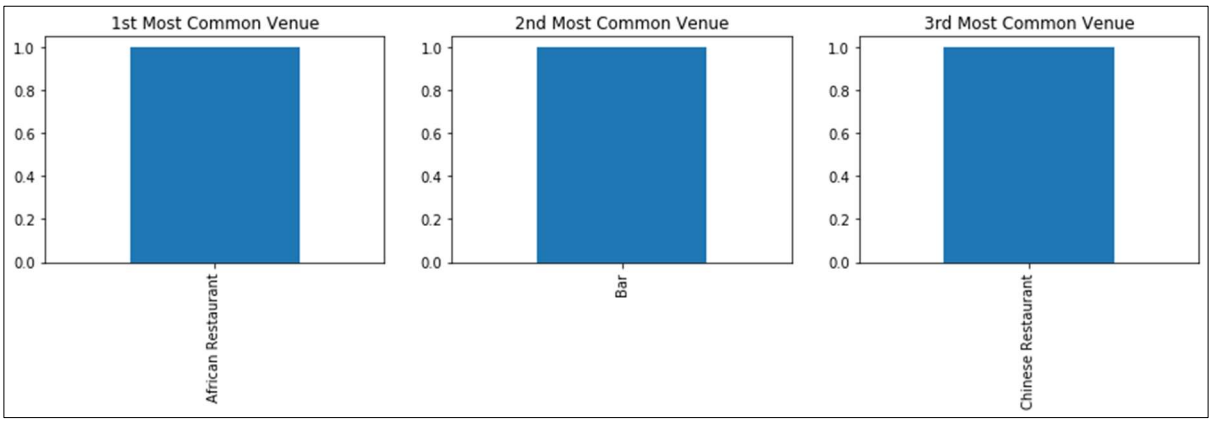

Figure 17:Most Common Venues in Cluster 4

5 Discussion

The geospatial analysis of the Community Districts in Manhattan yielded interesting insights into the wealth distribution within Manhattan. It was observed that the lower income neighbourhoods were situated in the northern parts of the island. However, even within the wealthy areas, there were Census Tracts experiencing poverty. This could be due to rent controlled apartments still being used or expensive buildings reserving a portion of the apartments for lower income groups.

Correlation analysis showed that the best indicator of success is education. Those that obtained a Bachelor's Degree or higher were more likely to earn a higher income. The analysis also showed the importance of finishing high school, with a large negative correlation between the number of people that didn't finish and income.

The commuting times were greater for people living in the low-income areas. This affects quality of life, since they have less free time for other pursuits (furthering their education, family time, etc.). This makes overcoming poverty greater, as we have seen that education correlates strongly with income levels.

The cluster analysis showed that the lower income areas were not similar to the higher income areas. This can be attributed to the low number of venues that are available in the lower income areas. This shows that there isn't sufficient economic activity in the low-income areas to sustain a large number of businesses.

6 Conclusion & Recommendations

The geospatial analysis of poverty and wealth distributions in Manhattan Island showed that there are inequalities that need to be addressed. It is recommended to NGOs looking to improve the socio-economic conditions of the poorer areas, that they:

- Focus on education as a primary method to eradicate poverty.
- Incentivise businesspeople to establish businesses within the areas so that the unemployment rates and commute times can be reduced.
- Look to foster entrepreneurship within those areas (thereby increasing the number of venues).