Capstone Project – The Battle of Neighbourhoods

Restaurant Location Problem in the city of New York using Foursquare Location Data.

Aparajita Sinha

1. Introduction:

1.1 Background:

New York is the most densely populated culturally diverse major city in the United States. It consists of five boroughs – Brooklyn, Queens, Manhattan, The Bronx and Staten Island, and more than 300 neighbourhoods. It is described as the cultural, financial and media capital of the world, which makes it a great place to start any new business. However, Location is one of the most important factors responsible for the success of any business. Therefore, it is advantageous to set up business in a safe place where there is abundance of customers who can afford your product, and the cost of doing business is optimum.

1.2 Problem:

We are working to find a place to open a new Italian restaurant in the city of New York using data-driven methodology. Data that might contribute in selecting restaurant location are: population density, demographics, purchasing power, crime rates, competitors and property affordability. This project aims to predict in which neighbourhood of New York, we should open our Italian restaurant based on these data.

1.3 Interest:

Obviously, anybody setting up a new business, particularly restaurant business will be interested in knowing the best location for setting up their businesses.

2. Data Acquisition and Cleaning:

2.1 Data Required and their Sources:

Data that might contribute in selecting restaurant location are: population density (as in higher the no. of people residing in your area higher are the chances of sales), demographics, purchasing power (as in whether people in your area can afford your product), crime rates, competitors (this suggests an already existing demand for your product) and property affordability. The data on population density and purchasing power can be found on the Wikipedia page of New York given here: https://en.wikipedia.org/wiki/Demographics_of_New_York_City ; and the data on competitors and nearby businesses can be found using foursquare location data; and the list of neighbourhoods of New York will be used from https://cocl.us/new_york_dataset . However, the data on crime rates and property affordability was not readily available.

2.2 Data Cleaning:

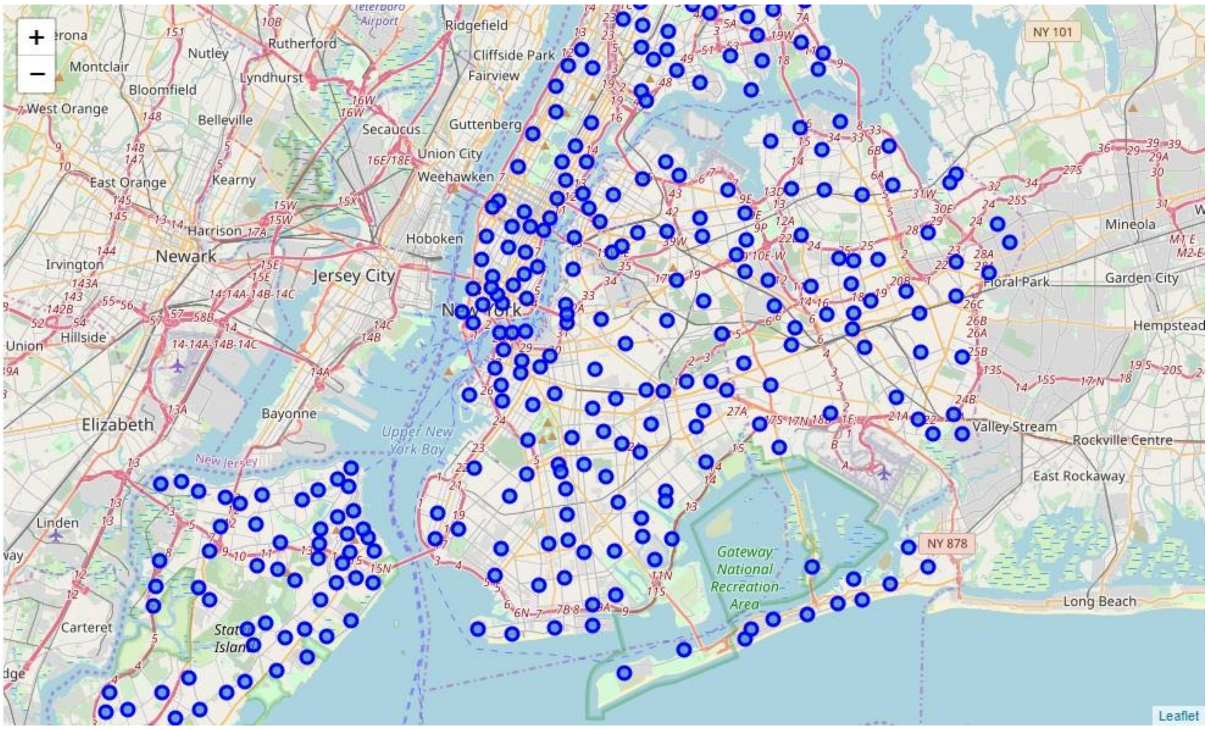
Both the list of neighbourhoods of New York and the foursquare location data were in the form of JSON file. They were transformed into pandas Dataframe for easier processing. Data on population density and per capita income were scrapped from Wikipedia in the form of table which was transformed into Pandas Dataframe.

2.3 Feature Selection:

The population and income dataset had some redundancy like total GDP and per capita GDP with the little difference that total GDP increases with population whereas per capita GDP does not. Also, there were three types of population data: total population, population density (per sq. Miles), and population density (per sq. Km). For our analysis, we chose per capita GDP as income data and population density in S.I. units as our population data.From foursquare location data, we chose venue name, their location and category name for our analysis.

3. Methodology:

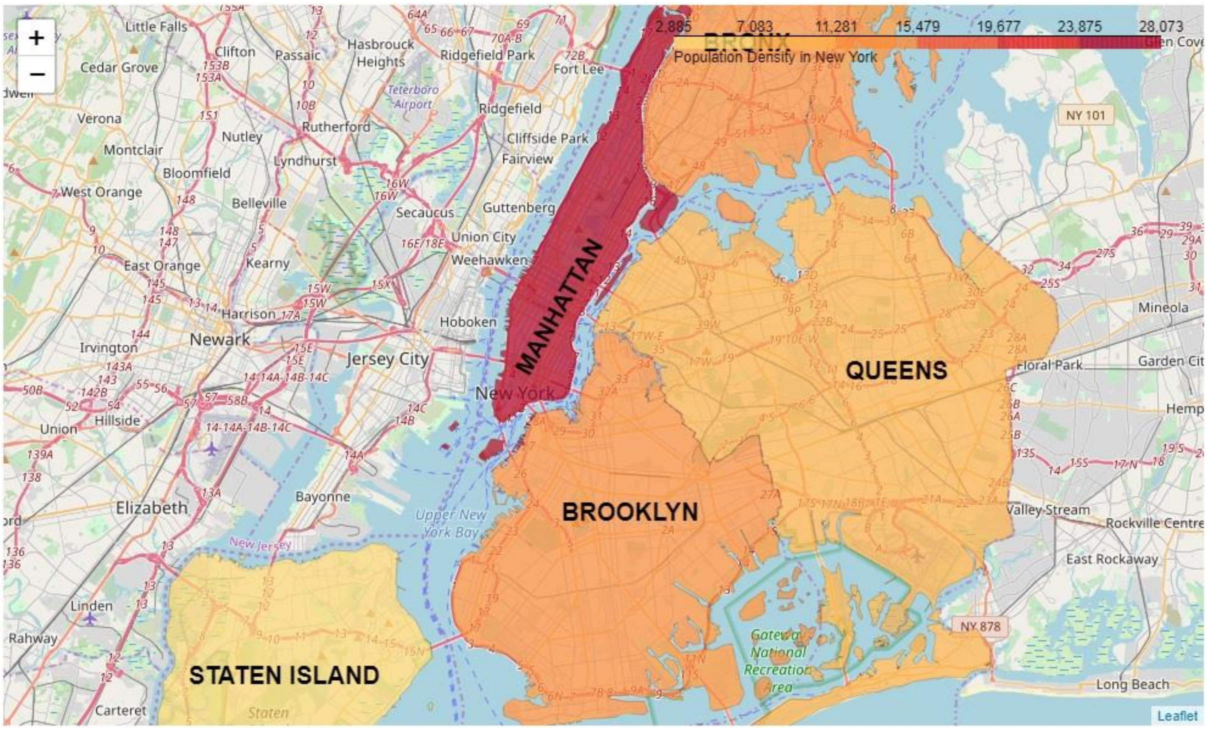
Our New York dataset contains a list of 5 boroughs and 306 neighbourhoods. Since, we are dealing with location problem; we have used folium maps in this project for better visualization of results. Given below, there is a map of New York showing all of their neighbourhoods.



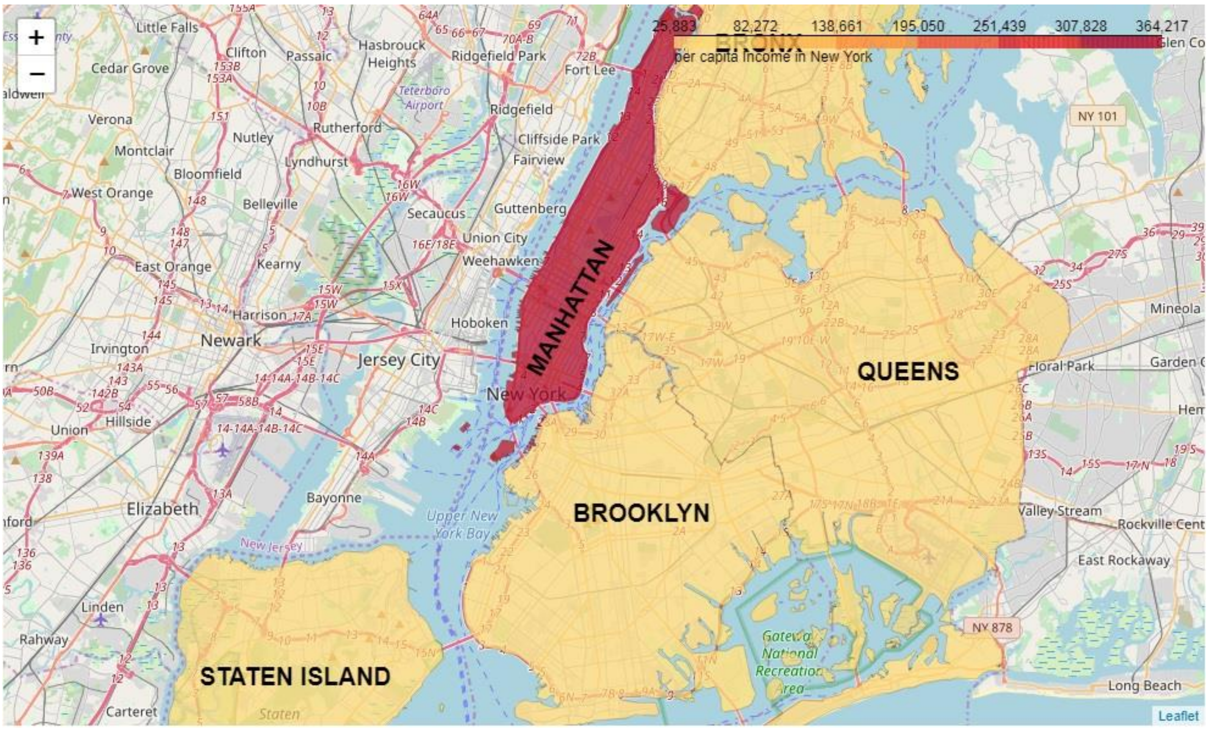
We have to narrow down this list of possible neighbourhoods to a few for our restaurant problem. We tried to achieve this in two parts. At first, we tried to analyse our income and population data which are present for our boroughs. We will use this to narrow down our boroughs. And then, we will use the foursquare location data for the neighbourhoods present in the narrowed down boroughs. Since, it is an unsupervised data problem, we will use clustering algorithm to find our suitable cluster and the list of neighbourhoods in this cluster will be our resulting list of possible locations for our Italian restaurant project.

3.1 Analysis of population density and per capita income of boroughs:

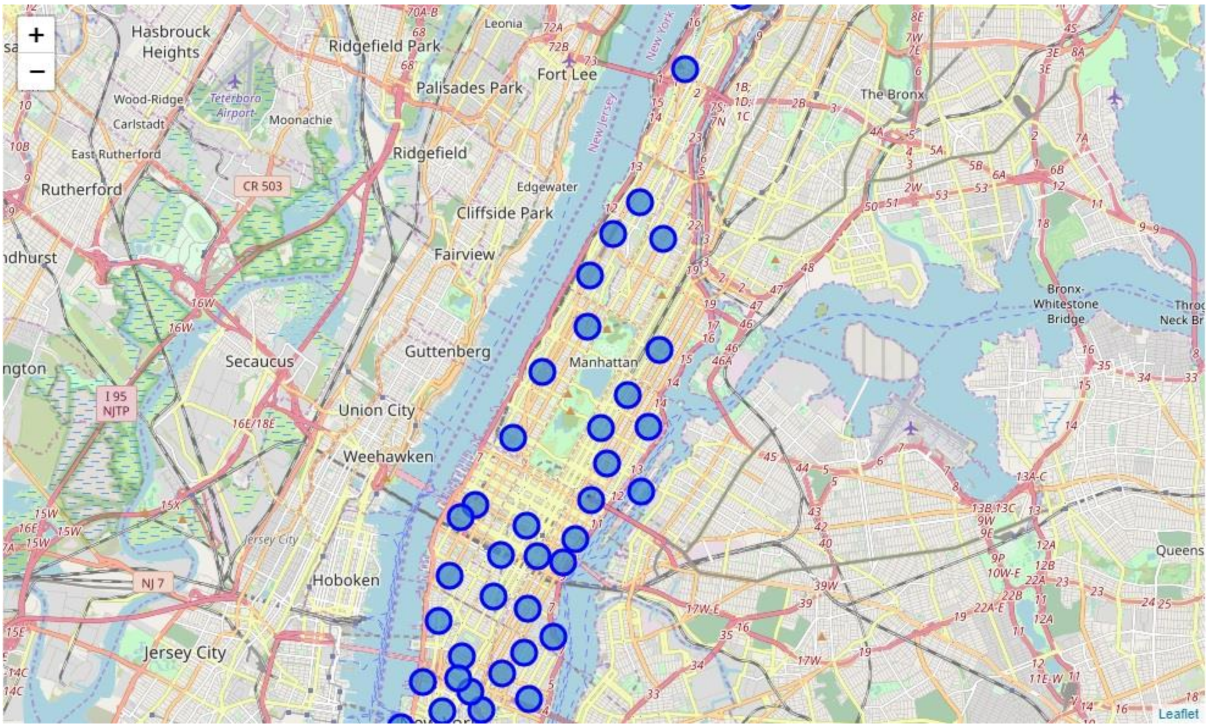
We have used choropleth maps of population density and per capita income of each borough to visually analyse their distribution. Given below, there is a choropleth map of New York showing population density of each borough which clearly shows that Manhattan has the highest population density, approximately, double the population density of other boroughs.



Given below, there is a choropleth map of New York showing per capita income of each borough which clearly shows that Manhattan is leading in per capita income with great margin from others.

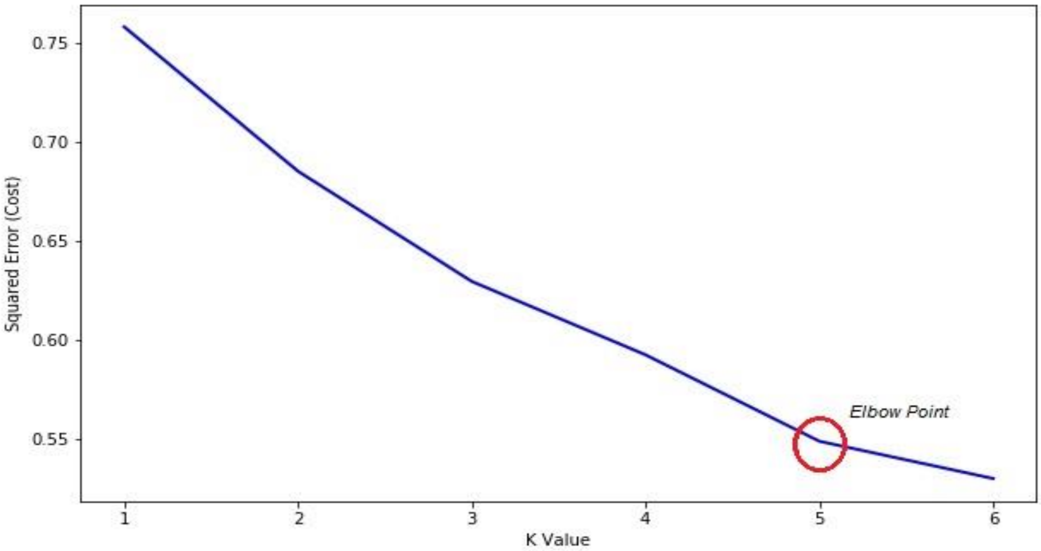


In both these choropleth maps, we have observed that Manhattan has the highest population density showing presence of a high number of consumers as well as highest per capita income showing more purchasing power of people of Manhattan. This suggests that Manhattan is the best place to start our business. Thus, we can narrow down our search to the neighbourhoods present in Manhattan. Given below, there is a map of Manhattan showing all the 40 neighbourhoods present in Manhattan.

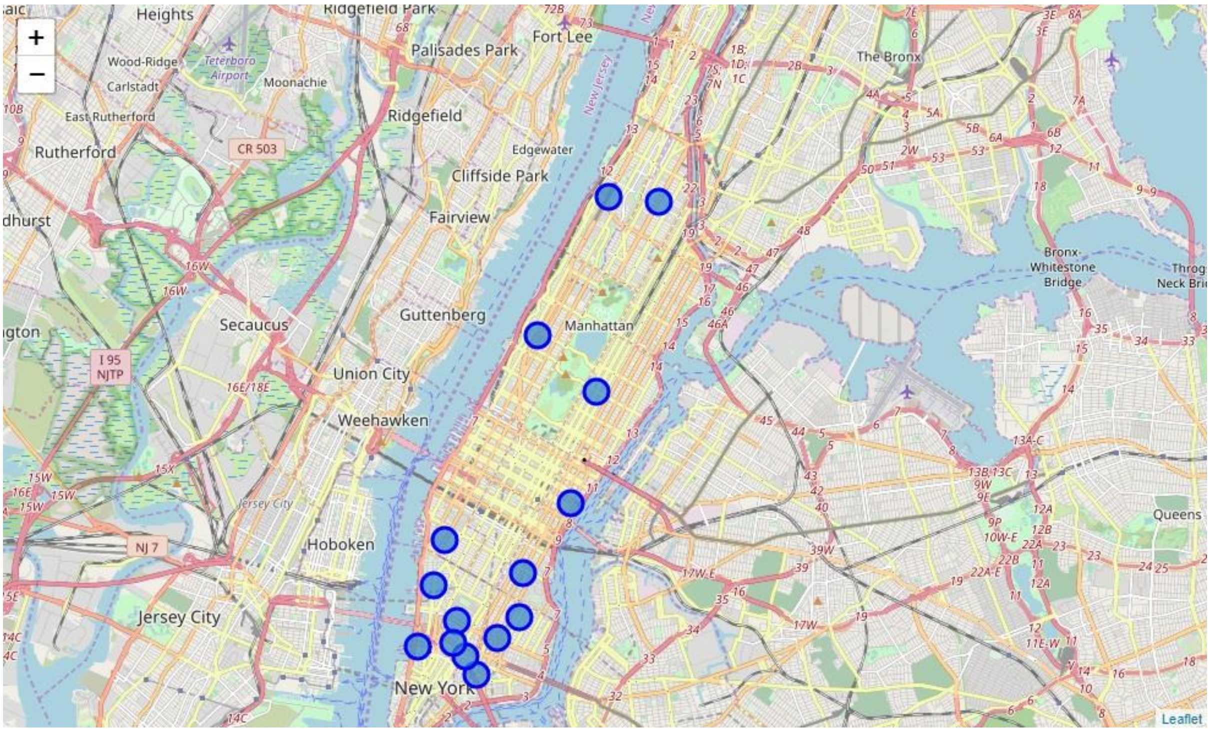


3.2 K-Means Clustering Algorithm:

Since, our problem is an unsupervised learning problem we will use K-means clustering algorithm on foursquare location data for grouping of Manhattan neighbourhoods into different clusters. For running K-means clustering, we need to decide an optimum K- value. We will achieve this by analysing the squared error cost function against k-values through elbow method as shown below.



We can observe that the elbow point gives optimum value at k=5. We have run k-means algorithm for this k-value on foursquare location data to obtain 5 clusters of Manhattan neighbourhoods. We have analyzed these clusters to look for a cluster similar to our need. We noticed that one of the cluster contained mostly restaurants and cafes as most common venues in neighbourhoods. We chose this cluster as it aligns the most with our requirements. This cluster contained 15 neighbourhoods. These neighbourhoods will be our final list of neighbourhoods for this project. The map below shows these neighbourhoods in Manhattan.

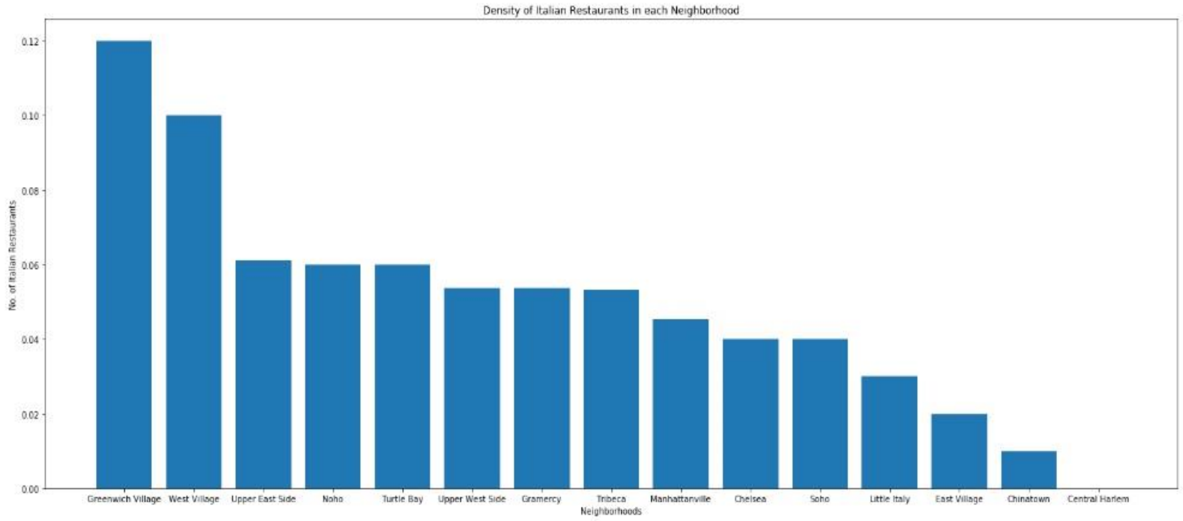


4. Results:

We have narrowed down our list of neighbourhoods from 306 to these 15 neighbourhoods:

- Greenwich Village
- West Village
- Upper East Side
- Noho
- Turtle Bay
- Upper West Side
- Gramercy
- Tribeca
- Manhattanville
- Chelsea
- Soho
- Little Italy
- East Village
- Chinatown
- Central Harlem

For better understanding of results, we have plotted a bar chart of density of Italian Restaurants in these neighbourhoods.



5. Discussion:

In the above bar chart, we can observe that the density of Italian restaurants in the first two neighbourhoods are almost double than the others. This shows greater competition in these neighbourhoods compared to others. We might want to choose our neighbourhood from one with lower competition. However, the last four neighbourhoods have very low density with the last one being anomalously low. We tried to look for a reason for this anomalous behaviour. The last neighbourhood, Central Harlem, is less culturally diverse and less safer which could be a possible explanation for such low density. The same reason was true for all four of last neighbourhoods. Therefore, we might want to exclude these four neighbourhoods from our list of possible neighbourhoods for our business. This leaves us with 9 neighbourhoods falling in the middle with approximately same level of competition. Therefore, we can decide from any of these neighbourhoods depending on availability and other qualitative factors:

- Upper East Side
- Noho
- Turtle Bay
- Upper West Side
- Gramercy
- Tribeca
- Manhattanville
- Chelsea
- Soho

However, if I had to choose I would go for either one among Upper East Side, Upper West Side and Manhattanville. All of these are centrally located with waterfront on one side, and Upper East and West Side has Central Park on the other side. Also these are among the safest neighbourhoods of New York.

6. Conclusion:

In this project, we analysed the population density and per capita income of boroughs through data visualizations like choropleth maps and used it for further analysis. We segmented foursquare location data through K-means clustering algorithm and analysed these clusters to find our interest cluster. We also used data visualization like bar chart to further refine our results.

Our findings can be very useful for anybody starting a new restaurant business. This will help them to decide the location of their businesses. However, there is always room for improvement. Our analysis would have improved if we had other data available to us as well like data on crime rates and property affordability. This would have further enhanced our results.