2. DATA ACQUISITION AND CLEANING

The data used for this project comprises of three sources:

1) List of Neighbourhood in Auckland

2) Latitude and Longitude of the desired neighbourhoods.

3) Venue data obtained from Foursquare app

Here the scope of this project is confined to the city of Auckland, New Zealand.

The first data is scraped from a Wikipedia page using the Beautiful Soup library in python. With the help of this library, we can extract the data in the tabular format as shown in the website. (Website- https://en.wikipedia.org/wiki/List_of_suburbs_of_Auckland) This is a list of 64 suburbs in the Auckland metropolitan area, New Zealand, surrounding the Auckland Central Business District. They are broadly grouped into the local government areas that existed from 1989 to 2010.

	Neighborhood
0	Arch Hill
1	Auckland CBD
2	Avondale
3	Balmoral
4	Blackpool

Figure 2.1: List of first five Neighbourhoods after scrapping

Latitude and longitude coordinates of those neighbourhoods are required in order to plot the map and also to get the venue data. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give the latitude and longitude coordinates of the neighbourhoods.

	Neighborhood	Latitude	Longitude
0	Arch Hill	-36.863020	174.748580
1	Auckland CBD	-36.848399	174.764388
2	Avondale	-36.890448	174.687017
3	Balmoral	-36.888200	174.740190
4	Blackpool	-37.051564	174.884397
...	...	...	...
59	Wai o Taiki Bay	-36.868200	174.870190
60	Waterview	-36.879466	174.699364
61	Western Springs	-36.863106	174.720365
62	Westfield	-36.950000	174.850000
63	Westmere	-37.568210	175.140200

64 rows × 3 columns

Figure 2.2: Latitude and Longitude details are added using Geocoder

Venue data, particularly data related to fruits and vegetable stores. We will use this data to perform clustering on the neighbourhoods. Foursquare API is used to get the venue data for the neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide different categories of the venue data, we are particularly interested in the category in order to help us to solve the business problem put forward.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Arch Hill	-36.86302	174.74858	Grey Lynn Park	-36.861524	174.743148	Park
1	Arch Hill	-36.86302	174.74858	Countdown	-36.858375	174.748862	Market
2	Arch Hill	-36.86302	174.74858	Ponsonby Central	-36.856276	174.746169	Shopping Mall
3	Arch Hill	-36.86302	174.74858	El Sizzling Chorizo	-36.856290	174.746131	Argentinian Restaurant
4	Arch Hill	-36.86302	174.74858	Viva Latino! Dance Studios	-36.860666	174.753579	Dance Studio

Figure 2.3 Venue Category added using Foursquare API