Capstone Project - The Battle of Neighborhoods

TOPIC: OPENING A NEW FRUIT & VEGETABLE STORE IN AUCKLAND, NEW ZEALAND

AINU JOY

1. INTRODUCTION

1.1 BACKGROUND

Auckland is a large metropolitan city in the North Island of New Zealand. The most populous urban area in the country, Auckland has an urban population of about 1,467,800 (June 2019), which is 29.9 percent of New Zealand's population. Auckland ranked third in a survey of the quality of life of 215 major cities of the world (2015 data).

Low fruit and vegetable intake are a risk factor for global mortality. Globally, approximately 1.7 million (2.8 percent) of deaths worldwide are attributed to low fruit and vegetable consumption.

1.2 BUSINESS PROBLEM

The objective of this project is to analyse and find out the best locations in the city of Auckland, New Zealand to open a new Fruits and Vegetable store. Data science methodologies, statistical and machine learning techniques like clustering are being used in this project to provide solutions to answer the business question: “If an investor is planning to open a new Fruits and Vegetable store in the city of Auckland, New Zealand, where would you recommend them to open it?”

1.3 TARGET AUDIENCE OF THIS PROJECT

This project is particularly useful to the investors who would like to open or invest in Vegetable and Fruit stores in the city of Auckland, New Zealand. It was found that Food and vegetable industry have no major players with a market share of greater than 5% in New Zealand, also they eat 1,800 tonnes of fruit and vegetables per day. A recent survey by the 5 + A Day Charitable Trust found that New Zealanders are rated among the highest consumers of fruit and vegetables globally. A 2015 joint study between Auckland, Otago and Oxford Universities investigated the potential impact of price subsidies on fruit and vegetables in New Zealand. It found that a 20 percent subsidy on fruit and vegetables, with the resulting impact on affordability, could prevent or postpone around 560 deaths a year. So being one of the busiest cities in New Zealand, Auckland is the best city to start with the project.

5 + A Day is a charitable trust set up to encourage New Zealanders to eat five or more servings of colourful, fresh fruit and vegetables every day. In New Zealand, almost one-third of adults are obese, with a further 35 percent being overweight. Considering these factors this is the right time to attain a monopoly in Fruits and Vegetable market in the Auckland city as the consumption increases with growing population and based on the dietary recommendations.

2. DATA ACQUISITION AND CLEANING

The data used for this project comprises of three sources:

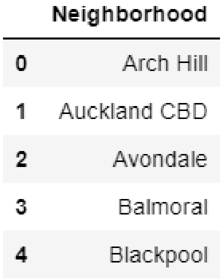
1) List of Neighbourhood in Auckland

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.

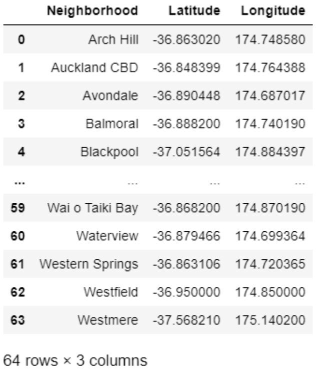
The first data is scraped from a Wikipedia page using the Beautiful Soup library in python. With the help of this library, we can extract the data in the tabular format as shown in the website. (Website- https://en.wikipedia.org/wiki/List_of_suburbs_of_Auckland) This is a list of 64 suburbs in the Auckland 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.

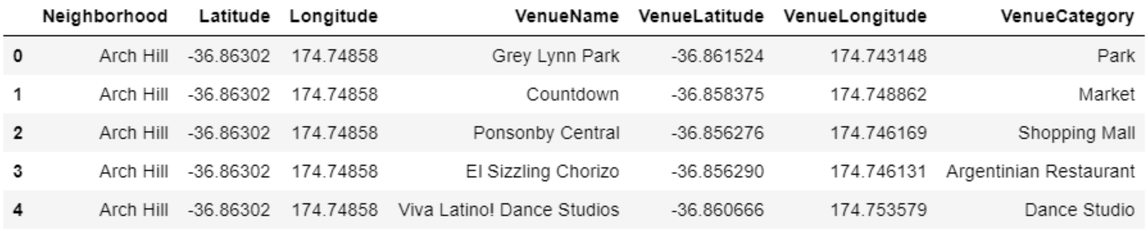


	Neighborhood	Latitude	Longitude
0	Arch Hill	-36.863020	174.748580
1	Auckland CBD	-36.848399	174.764388
2	Avondale	-36.890448	174.687017
3	Balmoral	-36.888200	174.740190
4	Blackpool	-37.051564	174.884397
...	...	...	...
59	Wai o Taiki Bay	-36.868200	174.870190
60	Waterview	-36.879466	174.699364
61	Western 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 Details added using Foursquare API

3. METHODOLOGY

3.1 EXPLANATORY DATA ANALYSIS

3.1.1 Web Scrapping using BeautifulSoup

Web Scraping is the process of downloading data from websites and extracting valuable information from that data. BeautifulSoup is a web scraping library which is best used for small projects. Here in order to obtain the list of neighbourhoods in the city of Auckland, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_suburbs_of_Auckland) Web scraping is done using Python requests and BeautifulSoup packages to extract the list of neighbourhood's data.

3.1.2 Assigning latitude and longitude using Geocoder

Since only a list of data can be obtained from the above method there is a need to get the geographical coordinates in the form of latitude and longitude to use in Foursquare API. So Geocoder package is used to convert address into geographical coordinates in the form of latitude and longitude. After this the neighbourhood data is populated into a pandas Data Frame and is visualized on a map using Folium package. This is done in order to perform a check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Auckland.

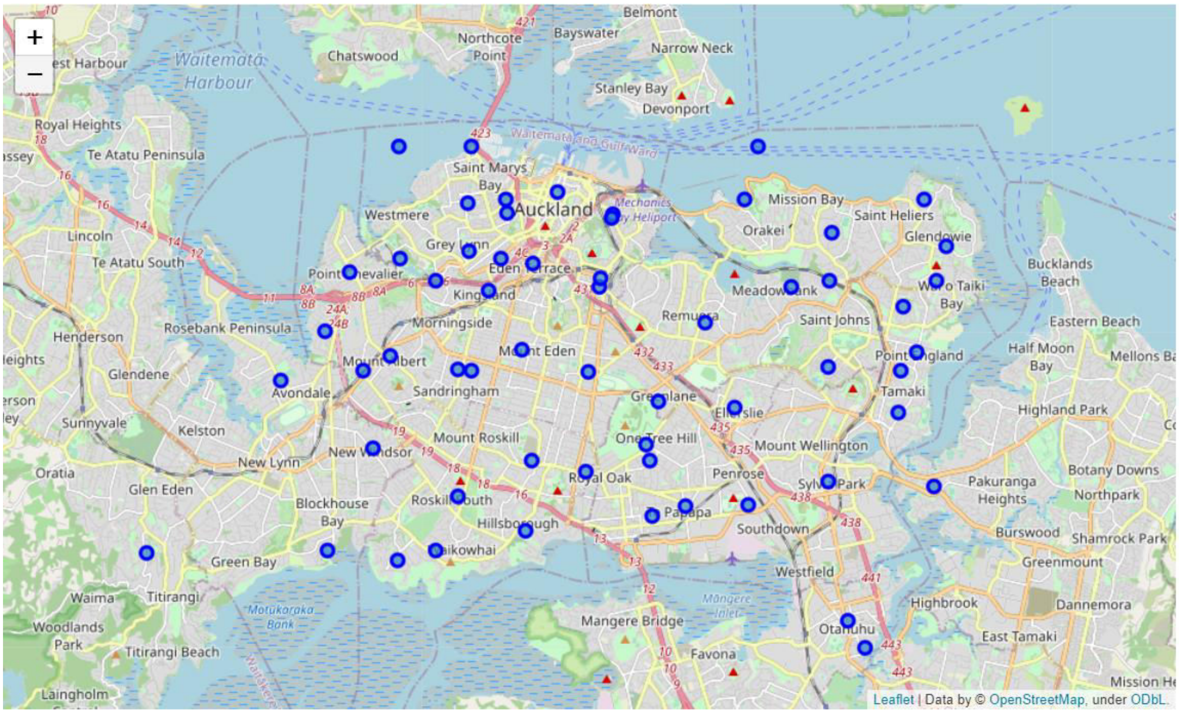


Figure 3.1 Map of Auckland using latitude and longitude values

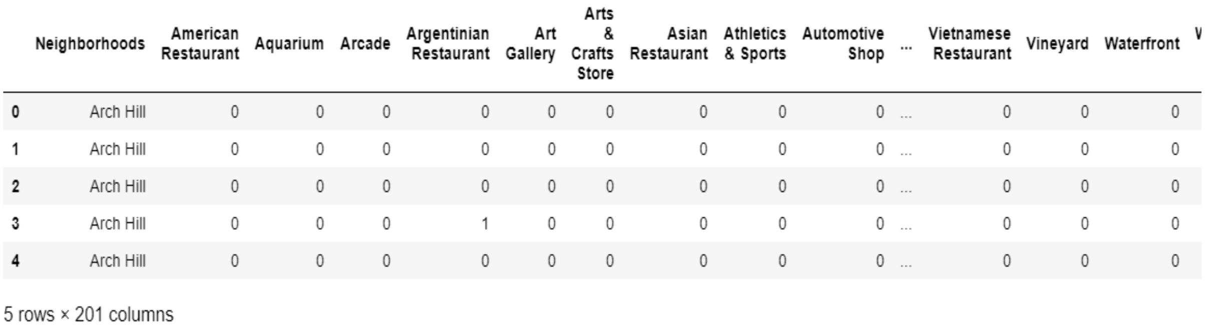
3.1.3 Foursquare API

Now Foursquare API is used to get the top 100 venues that are within a radius of 2000 meters. For that a Foursquare Developer Account has to registered in order to obtain the Foursquare ID and Foursquare secret key. API calls can be made to Foursquare by passing the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format where the venue name, venue category, venue latitude and longitude are extracted. Next data is analysed in each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, data for clustering is also prepared. Since analysing the “Food and vegetable store” data, we will filter the “Food and vegetable store” as venue category for the neighbourhoods.

Lastly, the clustering on the data is performed by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. The neighbourhoods are clustered into 3 clusters based on their frequency of occurrence for “Food and vegetable store”. The results allow to identify which neighbourhoods have higher concentration of Food and vegetable stores while which neighbourhoods have fewer number of Food and vegetable stores. Based on the occurrence of Food and vegetable stores in different neighbourhoods, it answers the question as to which neighbourhoods are most suitable to open new Food and vegetable stores.

3.2 MODELLING

One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighbourhood and the mean of the venues are calculated. K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size.



	Neighborhoods	American Restaurant	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	...	Vietnamese Restaurant	Vineyard	Waterfront	V
0	Arch Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	
1	Arch Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	
2	Arch Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	
3	Arch Hill	0	0	0	1	0	0	0	0	0	...	0	0	0	
4	Arch Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	

5 rows × 201 columns

4. RESULTS

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Fruits and Vegetable store":

Cluster 2: Neighbourhoods with moderate number of Food and vegetable stores

Cluster 0: Neighbourhoods with low number to no existence of Food and vegetable stores

Cluster 1: Neighbourhoods with high concentration of Food and vegetable stores

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 2 in purple colour, and cluster 2 in mint green colour.

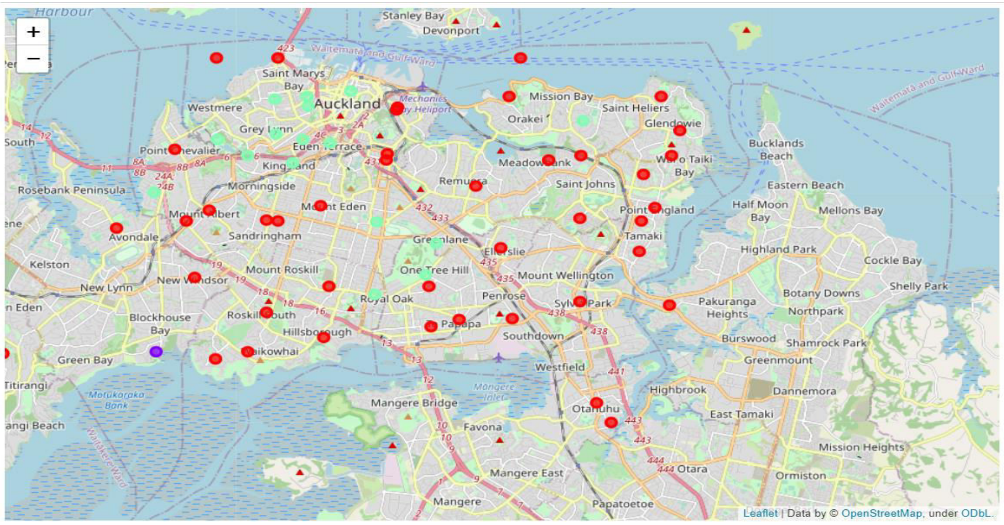
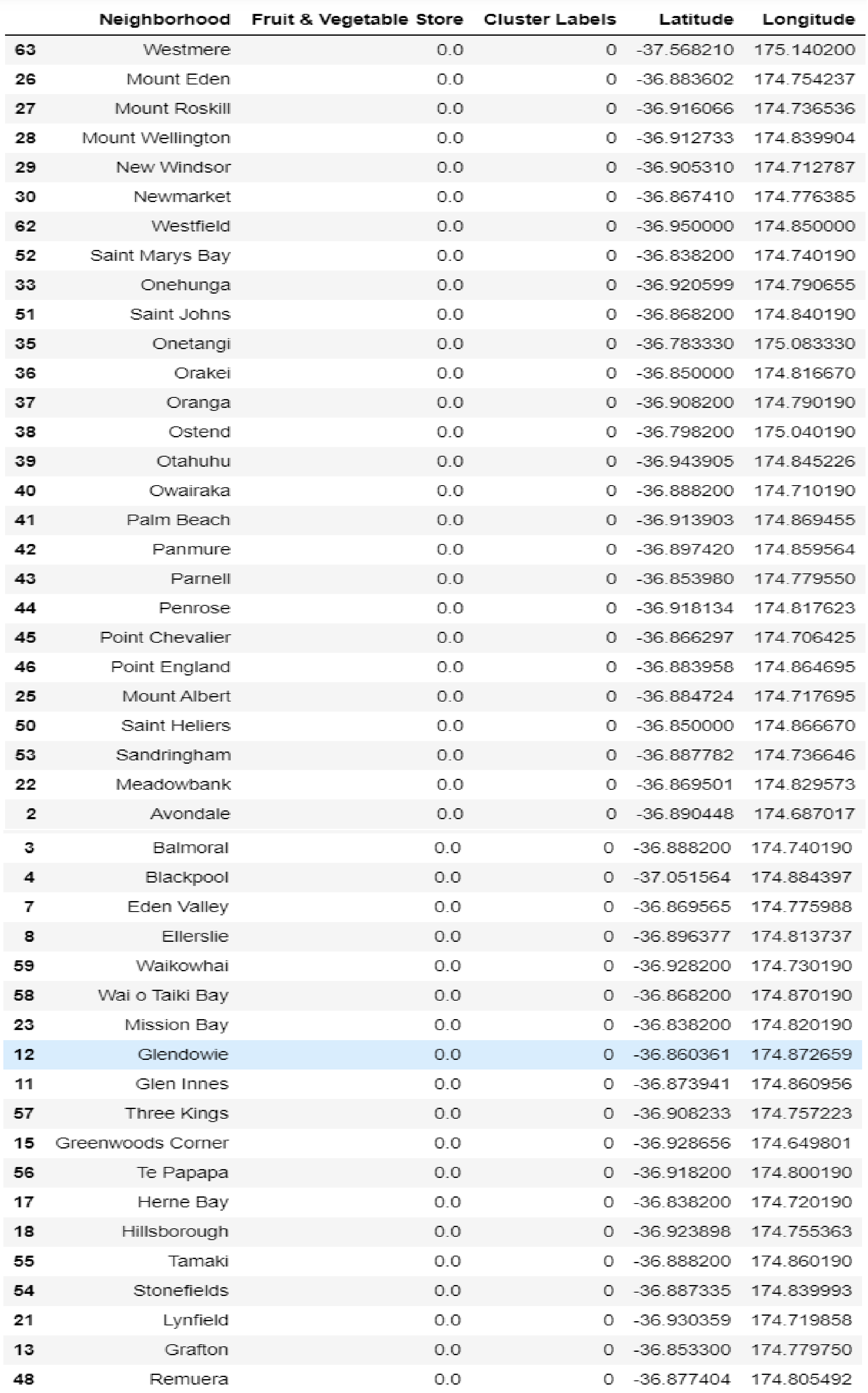
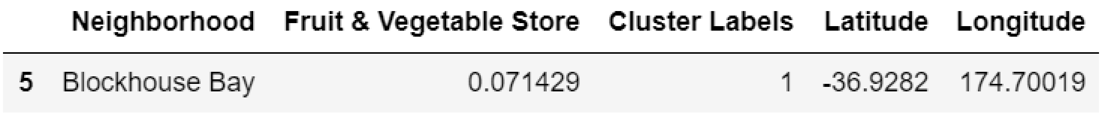


Figure 4.1 Clustered groups



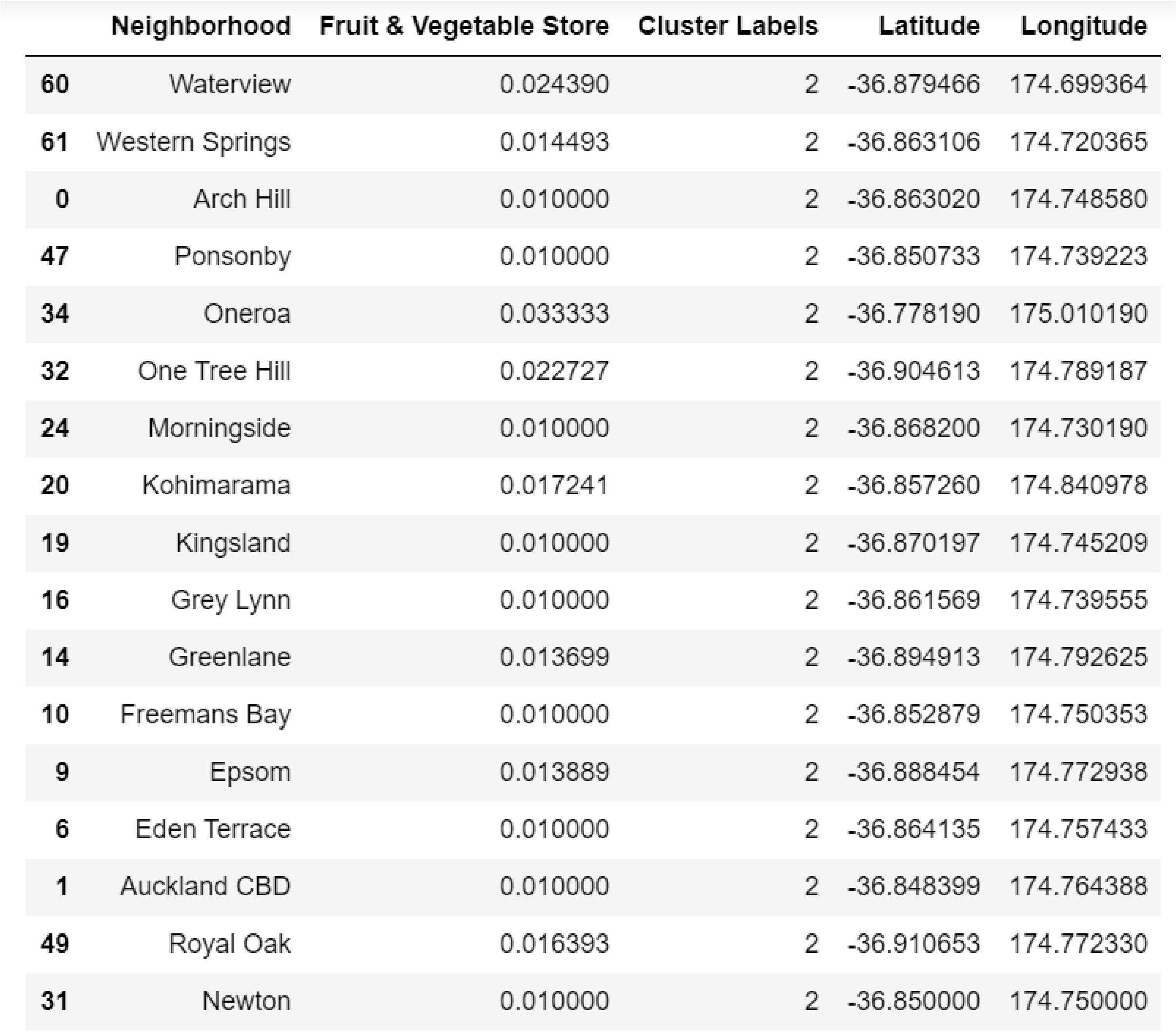
	Neighborhood	Fruit & Vegetable Store	Cluster Labels	Latitude	Longitude
63	Westmere	0.0	0	-37.568210	175.140200
26	Mount Eden	0.0	0	-36.883602	174.754237
27	Mount Roskill	0.0	0	-36.916066	174.736536
28	Mount Wellington	0.0	0	-36.912733	174.839904
29	New Windsor	0.0	0	-36.905310	174.712787
30	Newmarket	0.0	0	-36.867410	174.776385
62	Westfield	0.0	0	-36.950000	174.850000
52	Saint Marys Bay	0.0	0	-36.838200	174.740190
33	Onehunga	0.0	0	-36.920599	174.790655
51	Saint Johns	0.0	0	-36.868200	174.840190
35	Onetangi	0.0	0	-36.783330	175.083330
36	Orakei	0.0	0	-36.850000	174.816670
37	Oranga	0.0	0	-36.908200	174.790190
38	Ostend	0.0	0	-36.798200	175.040190
39	Otahuhu	0.0	0	-36.943905	174.845226
40	Owairaka	0.0	0	-36.888200	174.710190
41	Palm Beach	0.0	0	-36.913903	174.869455
42	Panmure	0.0	0	-36.897420	174.859564
43	Parnell	0.0	0	-36.853980	174.779550
44	Penrose	0.0	0	-36.918134	174.817623
45	Point Chevalier	0.0	0	-36.866297	174.706425
46	Point England	0.0	0	-36.883958	174.864695
25	Mount Albert	0.0	0	-36.884724	174.717695
50	Saint Heliers	0.0	0	-36.850000	174.866670
53	Sandringham	0.0	0	-36.887782	174.736646
22	Meadowbank	0.0	0	-36.869501	174.829573
2	Avondale	0.0	0	-36.890448	174.687017
3	Balmoral	0.0	0	-36.888200	174.740190
4	Blackpool	0.0	0	-37.051564	174.884397
7	Eden Valley	0.0	0	-36.869565	174.775988
8	Ellerslie	0.0	0	-36.896377	174.813737
59	Waikowhai	0.0	0	-36.928200	174.730190
58	Wai o Taiki Bay	0.0	0	-36.868200	174.870190
23	Mission Bay	0.0	0	-36.838200	174.820190
12	Glendowie	0.0	0	-36.860361	174.872659
11	Glen Innes	0.0	0	-36.873941	174.860956
57	Three Kings	0.0	0	-36.908233	174.757223
15	Greenwoods Corner	0.0	0	-36.928656	174.649801
56	Te Papapa	0.0	0	-36.918200	174.800190
17	Herne Bay	0.0	0	-36.838200	174.720190
18	Hillsborough	0.0	0	-36.923898	174.755363
55	Tamaki	0.0	0	-36.888200	174.860190
54	Stonefields	0.0	0	-36.887335	174.839993
21	Lynfield	0.0	0	-36.930359	174.719858
13	Grafton	0.0	0	-36.853300	174.779750
48	Remuera	0.0	0	-36.877404	174.805492

Figure 4.2 Cluster 0



	Neighborhood	Fruit & Vegetable Store	Cluster Labels	Latitude	Longitude
5	Blockhouse Bay	0.071429	1	-36.9282	174.70019

Figure 4.3 Cluster 1



	Neighborhood	Fruit & Vegetable Store	Cluster Labels	Latitude	Longitude
60	Waterview	0.024390	2	-36.879466	174.699364
61	Western Springs	0.014493	2	-36.863106	174.720365
0	Arch Hill	0.010000	2	-36.863020	174.748580
47	Ponsonby	0.010000	2	-36.850733	174.739223
34	Oneroa	0.033333	2	-36.778190	175.010190
32	One Tree Hill	0.022727	2	-36.904613	174.789187
24	Morningside	0.010000	2	-36.868200	174.730190
20	Kohimarama	0.017241	2	-36.857260	174.840978
19	Kingsland	0.010000	2	-36.870197	174.745209
16	Grey Lynn	0.010000	2	-36.861569	174.739555
14	Greenlane	0.013699	2	-36.894913	174.792625
10	Freemans Bay	0.010000	2	-36.852879	174.750353
9	Epsom	0.013889	2	-36.888454	174.772938
6	Eden Terrace	0.010000	2	-36.864135	174.757433
1	Auckland CBD	0.010000	2	-36.848399	174.764388
49	Royal Oak	0.016393	2	-36.910653	174.772330
31	Newton	0.010000	2	-36.850000	174.750000

Figure 4.4 Cluster 3

5. DISCUSSION

In the analysis neighbourhood with the highest number of fruit and vegetable store is in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to totally no fruit and vegetable store in the neighbourhoods. This represents a great opportunity and high potential areas to open new fruit and vegetable store as there is very little to no competition from existing ones. Meanwhile, fruit and vegetable stores in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of fruit and vegetable stores. Therefore, this project recommends investors to capitalize on these findings to open new fruit and vegetable store in neighbourhoods in cluster 0 with little to no competition. Investors with unique selling propositions to stand out from the competition can also open new fruits and vegetable stores in neighbourhoods in cluster 2 with moderate competition. Lastly, investors are advised to avoid neighbourhoods in cluster 1 which already have high concentration of fruit and vegetable store and suffering from intense competition.

6. CONCLUSION

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. investors regarding the best locations to open a new fruit and vegetable store. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new fruit and vegetable store. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new fruit and vegetable store.