IBM Data Science Professional Certification

Final Capstone Project - Exploration and clustering suburbs in Cape Town, South Africa

By Chantel Bok

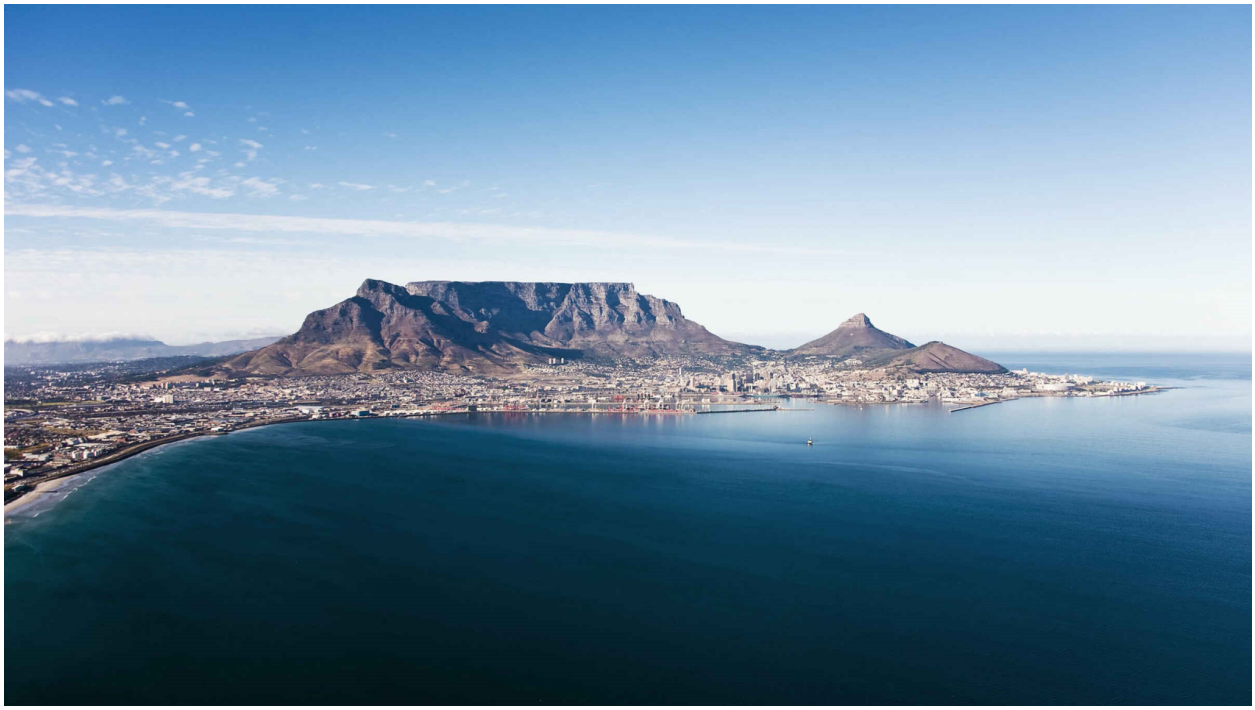

Image source: Picture of Table Mountain

Introduction

This report is a submission for the Final Capstone Project of the IBM Data Science Professional Certification on the Coursera platform. The project aims to demonstrate that the student can use Python and various libraries to perform RESTful API calls, web scraping and parsing of HTML code, data manipulation and exploratory data analysis.

For this project, neighbourhoods in Cape Town, South Africa will be explored in order to visualise where Italian restaurants are situated in relation to suburb locations.

The most common types of food venues located in each suburb can give an idea of the current and/or prior demands by the locals in that suburb (assuming a legitimate business). Therefore suburbs of Cape Town were clustered based on types of food venues located in the area and then visualised to provide additional information.

There are many factors to consider when assessing whether a new business will be feasible at a certain location, however for the purpose of this study, insights will only be related to what categories and quantities of food related venues can be found in the location of interest.

Business Problem

The owners of a popular Italian restaurant in Johannesburg wants to branch out to Cape Town, but is unsure where would be a good location to do so. They would like insight into the locations, categories and quantities of food venues in the area which would assist them in making their decision.

Data Sources

Locations and Postal Codes in Cape Town were scraped from https://www.southafricapostcode.com/location/western-cape/city-of-cape-town

Location data was retrieved by using the Foursquare API. Go to the https://developer.foursquare.com/ link and follow the instructions on the website to gain access to the Foursquare API.

Methodology

Various libraries and modules have been used for the purpose of the assignment. The library name along with a brief description of how can be found in Appendix A. Steps in the analysis process are discussed briefly below.

1) Web Scraping

Locations in Cape Town along with their postal codes were retrieved from the South Africa Postcode website using a web scraping process.

Web scraping is a method used to automatically extract large amounts of data from websites. There are different techniques, but in this project, the requests library was utilised to retrieve webpage contents. The BeautifulSoup library was then used to parse the HTML and find the locations and postal codes.

2) Create Pandas Dataframe

Pandas was chosen to handle data due to it being known for its versatility and ease of use when handling data. After retrieving the postal codes and location names, this data was stored in a pandas dataframe and exported to an Excel CSV file.

3) Geocoding

In order to generate a map of Cape Town and impose the locations, the latitudes and longitude values for each location were retrieved via geocoding. The geocoder library was chosen since it is open source and allows for multiple requests. The geocoder library occasionally returns empty values when requesting, thus a user has to repeat the request in order to obtain the desired value. This was accommodated for through use of a while loop to repeat requests until the desired values were obtained.

Geopy is another tool providing geocoding services. This was used to obtain the latitude and longitude values of Cape Town only for centralising the Folium map.

4) Folium Visualisation

Folium is a simple tool to visualize manipulated data on an interactive leaflet map. The folium library was used to generate a map with each suburb in the dataframe represented by a blue marker.

5) Utilisation of Foursquare API

Venue data was retrieved from the Foursquare database, specifically in the food category.

Venue data was retrieved within 500 m of each of the location coordinates in the dataframe and relevant information from the JSON files were added into a new dataframe containing venue names, venue location coordinates and venue categories (type of food venue).

6) Explore Cape Town Food Venue Data

The dataframe was inspected to understand whether manipulation is required for downstream analysis.

A subset of the dataframe was explored where venue categories were explicitly labelled as “Italian Restaurant” in order to retrieve the amount of Italian Restaurants in the dataset and subsequent mapping of these Italian restaurants based on their location coordinates.

7) Data Preparation and One Hot Encoding based on Venue Category

Suburbs for which less than 3 food venues were retrieved were excluded from the dataframe to assist with more relevant cluster separation. The dataset was further examined to understand venue categories and quantities. Post this further manipulation on dataframes were performed.

One hot encoding was then performed to transform the categorical data into a numerical form for subsequent clustering analysis. This allowed for the estimation of the top 3 venues for each suburb using the mean frequency of occurrence of each category for each location. A limitation here for example is if there were only 3 venues retrieved for a suburb, then the top 3 would not represent any order, but merely display the 3 venues retrieved for a location.

8) Clustering suburbs of Cape Town using KMeans

KMeans clustering is appropriate when working with datasets consisting of categorical data, and it is simple and flexible to implement. KMeans clustering is appropriate when working with datasets consisting of categorical data, and it is simple and flexible to implement. For the clustering analysis, only the venue categories for each suburb are relevant. The location coordinates were therefore excluded from the subset used for clustering analysis.

KMeans clustering was used to attempt grouping suburbs together based on the type of food venues located in the area and their frequency. It was not only the most common venues that were used in the analysis, but rather all the venue categories were considered if the venue category appeared at least 5 times in the dataset.

Prior to KMeans clustering, KElbowVisualizer was implemented to assist in selecting the optimal number of clusters by fitting a KMeans model for a range of k-values. The inflection point on a curve is a good indication that the underlying model fits best at that point. Data that is not clustered well will show a smooth curve and the value of K will be unclear.

Based on the results from the elbow method, the optimal amount of clusters were set and the KMeans model fit to the data.

9) Exploring the Clusters

The relevant data frames were merged along with the corresponding cluster labels. Subsequent dataframe manipulation was done to remove null values and convert cluster labels to integer type. This was necessary in order to visualise the clusters on a map.

The locations along with each cluster label representing a different colour was plotted on a folium map.

Results and Discussion

Post web scraping, a dataframe was successfully created containing 1265 locations along with their postal codes.

Shape of the dataframe is: (1265, 2)
First five rows of the dataframe:

	Location	Postcode
0	Admirals Hill	7798
1	Adriaanse	7490
2	Airlie	7806
3	Airport City	7490
4	Airport Industria	7490

Figure 1: Initial dataframe retrieved post web scraping.

Initially requests to the geocoder API were based on the postal codes for each location, however after further examination it was found that the dataset was inaccurate. Incorrect location coordinates were retrieved due to insufficient data in the geocoder database. The process was repeated using location names which resulted in accurate location data.

	Location	Latitude	Longitude
0	Admirals Hill	-34.069507	18.559225
1	Adriaanse	-33.939550	18.585440
2	Airlie	-34.038710	18.435080
3	Airport City	-33.978170	18.591080
4	Airport Industria	-33.958700	18.589430

Figure 2: Dataframe post retrieving location coordinates.

The location data was subsequently used to create a Folium map with blue markers representing each suburb in the dataset.

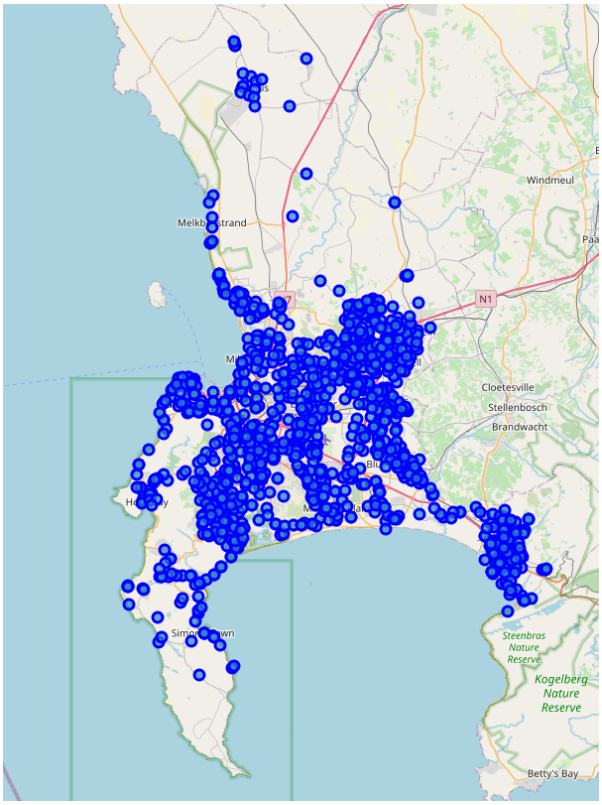

FIgure 3: Map of Suburbs in Cape Town

The FourSquare database provided the additional data required for the clustering analysis. Using the FourSquare API, names of food venues, food venue coordinates and categories were retrieved. This resulted in a dataframe with 11041 rows. The resulting dataframe also contained 134 unique food venue categories.

	Location	Location Latitude	Location Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Castle Rock	-34.23301	18.47351	Black Marlin Seafood Restaurant	-34.230116	18.471356	Seafood Restaurant
1	Millers Point	-34.23142	18.47592	Black Marlin Seafood Restaurant	-34.230116	18.471356	Seafood Restaurant
2	Scarborough	-34.19834	18.37602	Camel Rock	-34.197126	18.375143	Seafood Restaurant
3	Scarborough	-34.19834	18.37602	The Hub Café	-34.196884	18.375010	Coffee Shop
4	Scarborough	-34.19834	18.37602	Foragers Deli & Wholefoods	-34.196880	18.374964	Organic Grocery

Figure 4: Dataframe post collection of food venue data.

Food venue locations were retrieved in a 500 m range and thus there was significant overlap observed in the dataframe. Suburbs that overlap within a 500 m range are likely to be similar to one another, therefore only one location per specific venue was retained for the subsequent analysis. Consequently the dataset was reduced to 3587 rows.

Italian restaurants represent 118 venues in the dataframe.

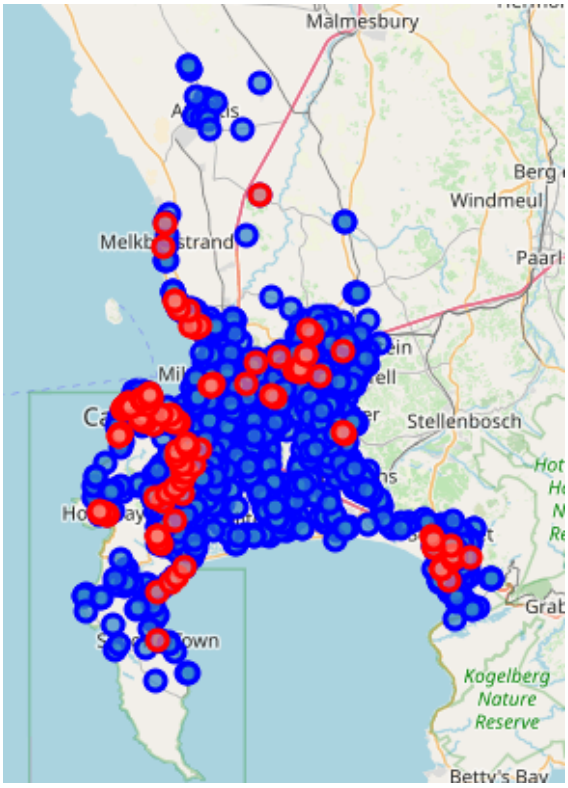

Figure 5: Map with Italian Restaurant locations in Cape Town.

Further examination of the data revealed 254 locations with less than 3 food venues retrieved. These locations were excluded from the clustering analysis and resulted in a dataframe with 3248 rows.

Venue categories were further explored to determine whether amendments should be made to improve the clustering analysis.

It was found that the most commonly occurring food venue category is labelled as 'Coffee Shop' in the Foursquare database and the second most common venue is labelled as 'Café'.

It is assumed in this analysis that 'Coffee Shop' and 'Café' are meant to be under one label and the data frame was manipulated as such. Below are the top 10 most common venue categories, before and after combining the 'Café' and 'Coffee Shop' categories.

	Venue Category	Venues
0	Coffee Shop	366
1	Café	355
2	Restaurant	223
3	Fast Food Restaurant	204
4	Pizza Place	198
5	Seafood Restaurant	123
6	Burger Joint	121
7	Bakery	111
8	Italian Restaurant	111
9	Indian Restaurant	85

	Venue Category	Venues
0	Café	721
1	Restaurant	223
2	Fast Food Restaurant	204
3	Pizza Place	198
4	Seafood Restaurant	123
5	Burger Joint	121
6	Italian Restaurant	111
7	Bakery	111
8	Indian Restaurant	85
9	Fish & Chips Shop	77

Figure 6: Before and after merge of the Coffee Shop and Café label.

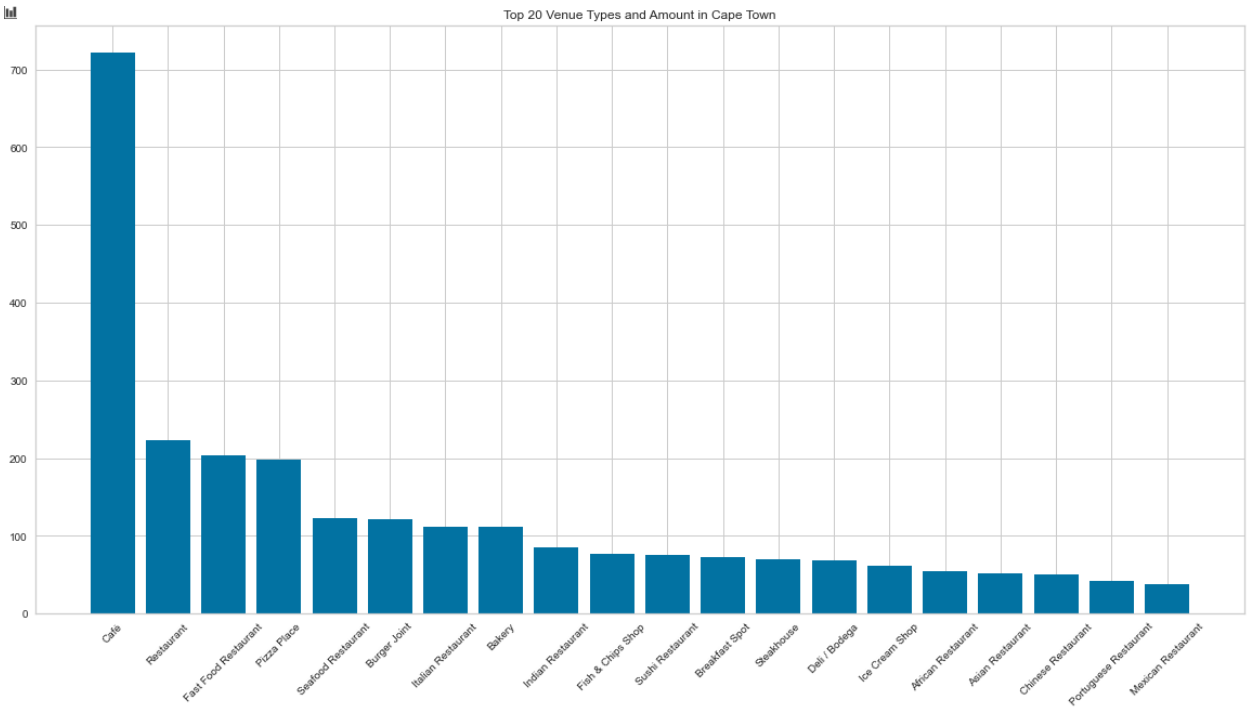

Figure 7: Top 20 Food Venue Types in Cape Town

It is clear that the most common label in the food category of the Foursquare database for Cape Town is 'Café'. Close to three times as many Café's are labelled than the second to fourth most common food venue label.

Unfortunately just over 200 food venues are not specifically categorised, and are labelled as 'Restaurant'. This data was however kept in the data frame since it is still relevant considering that venue categories that are not of 'Restaurant' type (e.g. Bakery) are also present in the data frame.

Venue categories for which there were less than 5 records were removed from the data frame, since their relevance was questionable.

This resulted in a data frame with 3108 rows.

One hot encoding was employed to transform the categorical venue data to numerical form in order to derive mean and frequency of each food venue category for each location - the basis on which the clustering analysis is performed.

From this transformed data, the top 3 most common food venue types for each location in the data frame was retrieved.

	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Albowville	Café	Sushi Restaurant	African Restaurant
1	Amandaglen	Vegetarian / Vegan Restaurant	Food Truck	Steakhouse
2	Annandale	Restaurant	Pizza Place	Breakfast Spot
3	Arauna	Fish & Chips Shop	Fast Food Restaurant	Deli / Bodega
4	Area A1	Fish & Chips Shop	Bakery	Café

Figure 8: Dataframe with top 3 most common venues per location

A data frame from which the 'Location' column was created for the clustering analysis and comprised only the transformed venue category data.

The elbow method was utilised to estimate the optimum amount of clusters to generate when fitting a KMeans model to the data.

An elbow was identified at a cluster count of 4, however the elbow is not very distinct.

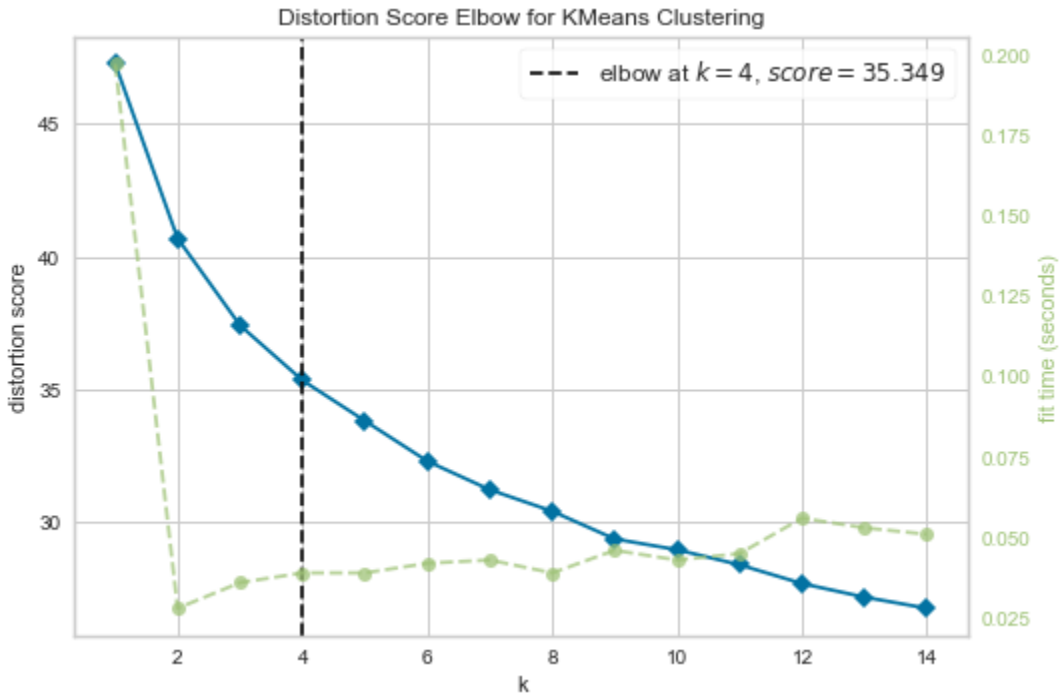

Figure 9: Graph representing results obtained from using the Elbow method.

The KMeans model was fit with a k value of 4, however the lack of a distinct elbow is indicating that the data is not well clustered. Optimisations on the dataset can still be done to determine whether a better fit can be obtained without compromising data integrity.

Post fitting of a KMeans cluster model, a cluster label between 0 and 3 were generated per data point. These labels were merged with previous data frames created to generate the final data frame from which clusters were then visualised on a map.

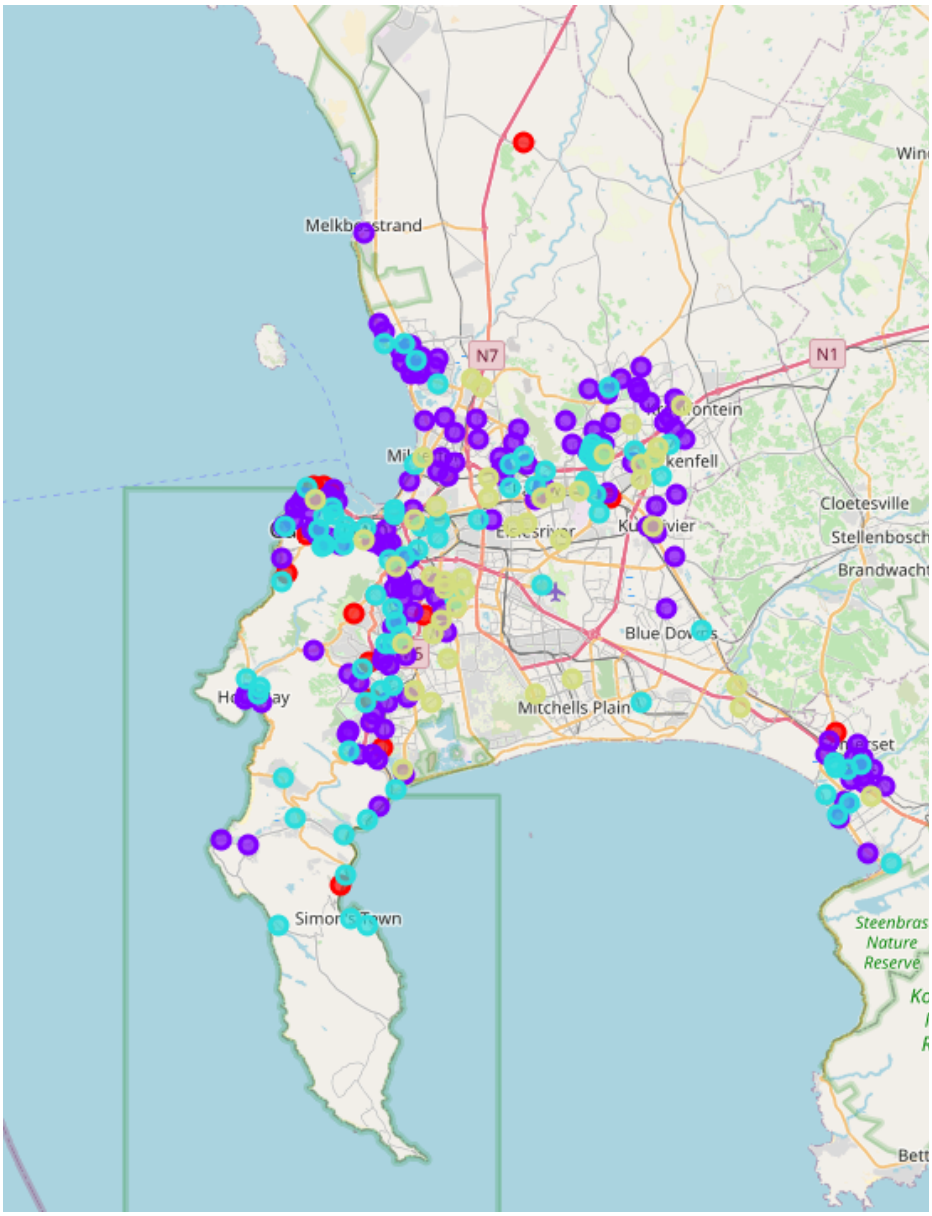

Figure 10: Visualisation of clustered locations across Cape Town

Cluster	Marker Colour	Locations in Cluster
0	Red	14
1	Purple	137
2	Light blue	82
3	Light green	42

Cluster 0

1st Most Common Venue	Location
African Restaurant	4
Ice Cream Shop	1
Restaurant	9

2nd Most Common Venue	Location
Café	2
African Restaurant	1
American Restaurant	1

3rd Most Common Venue	Location
Fast Food Restaurant	2
Italian Restaurant	2
Pizza Place	2

Figure 11: Top 3 counts of 1st, 2nd and 3rd most common venues for locations in Cluster 0.

There are 3 unique food venue categories identified for the most common food venue type in Cluster 0 and 14 locations. The top most common venue category in this cluster is the uncategorised 'Restaurant' category, followed by 'African Restaurant'.

Cluster 1

1st Most Common Venue	Location
Café	41
Pizza Place	18
Restaurant	16

2nd Most Common Venue	Location
Café	18
Pizza Place	16
Italian Restaurant	11

3rd Most Common Venue	Location
Pizza Place	13
Restaurant	9
Italian Restaurant	8

Figure 12: Top 3 counts of 1st, 2nd and 3rd most common venues for locations in Cluster 1.

Cluster 1 has a large range of most common venues due to the cluster containing 137 locations and 29 unique category types for most common food venues. The most common venue category is Café, followed by 'Pizza Place' and 'Restaurant'. Five locations have Italian Restaurant as their most common venue in this cluster, 11 in the second most common venue category and 8 in the 3rd most common venue category.

Cluster 2

1st Most Common Venue	Location
Café	69
Breakfast Spot	2
African Restaurant	1

2nd Most Common Venue	Location
Restaurant	11
African Restaurant	7
Café	7

3rd Most Common Venue	Location
African Restaurant	11
Restaurant	8
Café	6

Figure 13: Top 3 counts of 1st, 2nd and 3rd most common venues for locations in Cluster 2.

In this cluster, the Café venue category stands out, with hardly any other venue categories making a significant contribution. In terms of second most common venues, the uncategorised 'Restaurant' category is most frequent, followed by 'African Restaurant' and Café. 'African Restaurant', 'Restaurant' and 'Café' are again most frequent in the 3rd most common venues.

Cluster 3

1st Most Common Venue	Location
Fast Food Restaurant	19
Bakery	3
Burger Joint	2

2nd Most Common Venue	Location
Fast Food Restaurant	12
Café	7
Pizza Place	5

3rd Most Common Venue	Location
Fast Food Restaurant	8
African Restaurant	6
Restaurant	5

Figure 14: Top 3 counts of 1st, 2nd and 3rd most common venues for locations in Cluster 3.

Cluster 3 clearly has 'Fast Food Restaurants' as their top most common venue. Other food venue categories in this cluster such as 'Burger Joint', 'Fish and Chips Shop' etc. can also be labelled as 'Fast Food Restaurant'.

Overall the cluster analysis gives a good representation of the distribution of food venues and their categories on a map. It does not appear that there is a distinct difference between major areas in Cape Town and the food categories available. The only exception is cluster 3 which appears to display a pattern of being located towards the center of the

map. Since this cluster represents a majority of fast food chains, these locations seem well placed to provide easy access for frequent service delivery.

Conclusion and Recommendations

Web scraping, geocoding and retrieving data from the Foursquare database provided the data for this assignment. From this data, data frames were constructed and manipulated to improve the quality of the data being analysed, and ultimately improve the accuracy of the KMeans clustering algorithm applied on the dataset. Frequent inspection of the data frames along with visualisation tools assisted in understanding the dataset being analysed.

The map of Cape Town, along with the 1265 locations plotted, provided an overview of the suburbs in Cape Town, which may also be potential Italian restaurant locations. Understanding if and where there may be competition is crucial to the success of a new business. The map of the Italian restaurants that are currently in the Foursquare database provided insight. Ideally one would not open an Italian restaurant in a suburb where several Italian restaurants may already be fulfilling demand and success rate mostly depends on capturing market share from the competition.

The KMeans clustering resulted in four clusters of suburbs in Cape Town. Cluster 0 contained the fewest samples and the most common food venues for this cluster was the non-specific type 'Restaurant'. Cluster 1 had the most locations and the most common food venue was of type 'Café'. In this cluster, 'Pizza Place' and 'Italian Restaurant' also feature frequently. It may be worth looking at whether the market is saturated in this cluster. Cluster 2 had the most common food venue of type 'Café', but the second and third most common venues are 'African Restaurant' and the non-specific type 'Restaurant'. Cluster 3 has 'Fast Food' restaurant as the most common food venue and is the only cluster that slightly indicates some sort of geographical pattern when inspecting the map. They seem to be dispersed more centrally on the map.

In conclusion, the analysis provided some insight into how Cape Town suburbs are dispersed and what their differences are in terms of the types and quantities of food venues located within them. It does not provide enough information to base a decision on for opening an Italian restaurant, but can be a valuable tool when combined with a full market study.

Appendix A

Table 1: *Libraries/modules utilised*

Name	**Basic description of use**
Requests	Utilised for making HTTP requests in Python.
html5lib	Utilised for HTML parsing.
BeautifulSoup	Utilised for scraping information from web pages.
time	Python library providing several time-related functions.
random	Python module implementing pseudo-random number generators.
pandas	Utilised for data structuring and analysis in Python.
NumPy	Utilised for performing mathematical operations, particularly when working with arrays and matrix data structures.
geocoder	Utilised to retrieve latitude and longitude information for a particular address / postal code.
geopy	Utilised to retrieve latitude and longitude information per instance.
folium	Utilised for visualisation of manipulated data on an interactive leaflet map.
json	Utilised in handling JSON files.
matplotlib	Utilised for data visualisation.
sklearn	Provides tools for predictive data analysis.
yellowbrick	An extension of the sklearn library providing tools for model selection and hyperparameter tuning