Finding the optimal coastal city for investment between Cape Town and Durban (South Africa)

By: Sihle Mkaza

Project For: Applied Data Science Capstone course (IBM Data Science Professional Certificate)

Table of Contents

1. Introduction

1.1 Background

Cape Town and Durban are two of the most popular cities in South Africa and both are coastal cities[1] [2]. Both cities are business centers, with Cape Town having many tourist attractions such as Table Mountain and Robben Island and with Durban having tourist attractions such as the Durban Botanic Gardens. Many international companies choose Cape Town as an initial entry into the African market and Durban is the first choice for holiday destinations. Setting up a viable business in one of the cities would be a good investment for any entrepreneur or investor.

1.2 Problem

Seeing that both Cape Town and Durban share many similar features such as being costal cities and business hubs, it is hard for investors and/or entrepreneurs to choose a city to start a business in when having to pick between the two. The aim of this report and project is to provide an analysis of both cities and determine which of the cities would be the most optimal one to invest in. This will be accomplished by providing information on what types of businesses (or business categories) are the most popular in each of the cities and suggesting on what business category to invest in and in which city to invest in.

The information produced from this report would be most useful to investors and/or entrepreneurs that have identified Cape Town and Durban as potential business locations and are not sure what kind of business to start yet and in which city.

2. Data

2.1 Data Acquisition

The names of the neighbourhoods in each city is needed for an analyses of businesses categories in each city. No single website with names of neighbourhoods for each city and the associated postal codes exist, so two csv files were manually created with data obtained from multiple websites. The first csv file contains Cape Town neighbourhoods and their associated postal codes [3] and the second csv file contains Durban neighbourhoods and their associated postal codes [4].

The Foursquare API will be used to obtain popular venues (businesses/places) for each of the neighbourhoods in each city using the 'venues/explore' foursquare endpoint [5].

The latitude and longitude of each neighbourhood is needed in order to get popular venues using the foursquare 'venues/explore' endpoint. The GeoPy Nominatim library [6] was used to obtain the latitude and longitude of each of the neighbourhoods in Cape Town and Durban. The retrieved latitude and longitude values were merged with the dataframe containing neighbourhood names.

2.2 Data Cleansing

On some occasions the geopy nominatim module fails to get the latitude and longitude values of a neighbourhood and results in the neighbourhood dataframe containing NaN values where the latitude and longitude values are meant to be. All neighbourhoods with NaN values in the latitude and longitude column were discarded as it would not be possible to get popular venues for them using the foursquare API.

2.3 Feature Selection

Even though each of the neighbourhoods have postal codes, some neighbourhoods (two or more) share the same postal code, for example Esplanade, Essenwood and Glenmore in Durban share the postal code 4001. Postal code is there not very representative of each neighbourhood as two

neighbourhoods that share the same postal code can very in the trends of popular businesses in each neighbourhood [7]. Postal code was therefore discarded from the data.

	Neighbourhood	City	Latitude	Longitude
0	Bakoven	Cape Town	-33.960000	18.382778
1	Bantry Bay	Cape Town	-33.928151	18.378970
2	Camps Bay	Cape Town	-33.954774	18.381852
3	Clifton	Cape Town	-33.935285	18.379070
4	Fresnaye	Cape Town	-33.925194	18.387743

Table 1: first five rows of Cape Town data after feature selection

	Neighbourhood	City	Latitude	Longitude
0	Addington	Durban	-29.868127	31.043306
1	Austerville	Durban	-29.945278	30.980833
2	Avondale Road	Durban	-29.844753	31.009084
3	Bayhead	Durban	-29.891389	30.991389
4	Beach	Durban	-29.861825	31.009909

Table 2: first 5 rows of Durban data after feature selection

3. Methodology

3.1 Mapping neighbourhoods before analysis

Before any machine learning algorithms were used, the neighbourhoods had to be first visualised on a map. Using the folium library, maps for Cape Town and Durban were created using the neighbourhood names and location coordinates. The following maps show the neighbourhoods in Cape Town with blue markers and the neighbourhoods in Durban with purple markers.

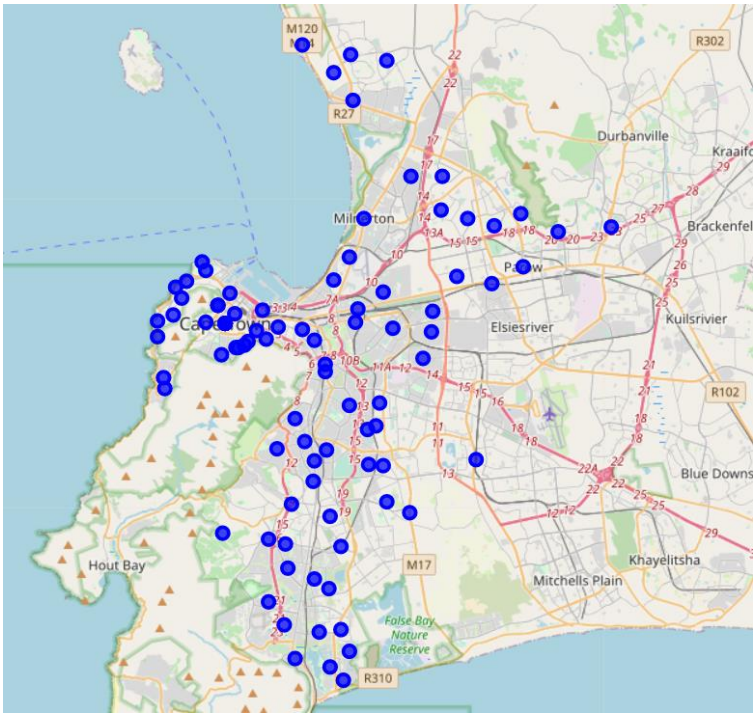

Figure 1: Cape Town before any clustering

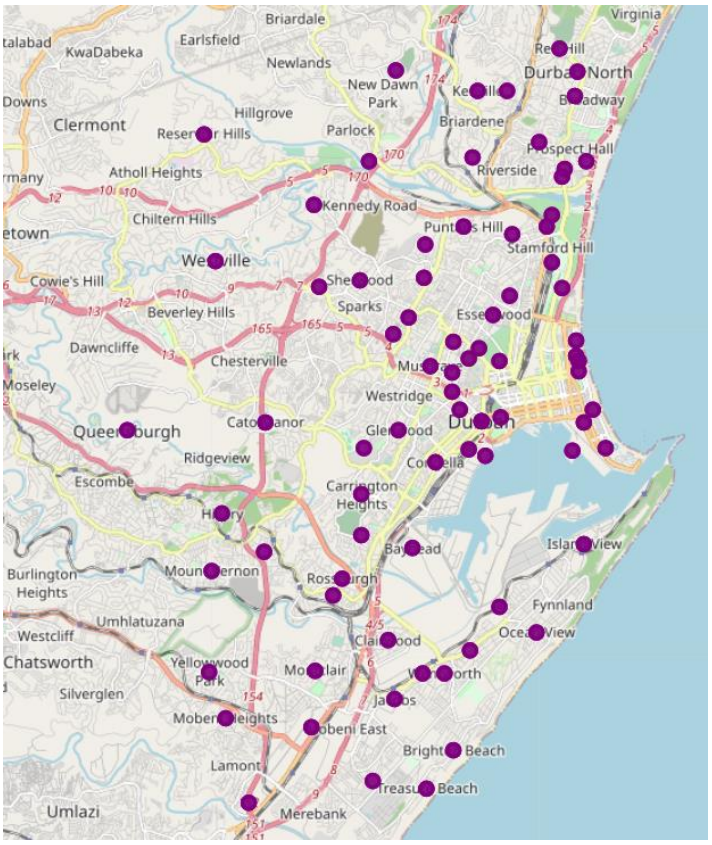

Figure 2: Durban before any clustering

3.2 Venue popularity exploratory analysis

The foursquare API was used to identify popular venues (places) in each of the cities by supplying neighbourhood names and locations to the 'venue/explore' foursquare endpoint. From the retrieved

data, the top 5 popular venues were chosen. 163 unique venue categories were identified for Cape Town while 124 unique categories were found for Durban from the foursquare data. The following tables show the 5 popular venue categories for each city.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Athlone	Jewelry Store	Yoga Studio	Factory	Fountain	Food Service
1	Bakoven	Bed & Breakfast	Hotel	Farm	Fountain	Food Service
2	Bantry Bay	Hotel	Hotel Pool	Dog Run	Scenic Lookout	Bed & Breakfast
3	Bellville	Bar	Bakery	Mexican Restaurant	Café	Other Nightlife
4	Bergvliet	Hockey Field	Breakfast Spot	Yoga Studio	Farm	Fountain
5	Bishopscourt	Hotel	Pub	Event Space	Food Service	Food & Drink Shop
6	Bloubergstrand	Hotel	Café	Seafood Restaurant	Farm	Food Service
7	Bo-Kaap	Café	Burger Joint	Restaurant	Steakhouse	Coffee Shop
8	Bothasig	Convenience Store	Yoga Studio	Factory	Food Service	Food & Drink Shop
9	Brooklyn	Furniture / Home Store	Bus Station	Fast Food Restaurant	Gym / Fitness Center	Factory

Table 3: First 10 rows showing the top 5 popular venues in Cape Town per neighbourhood

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Addington	Seafood Restaurant	Beach	Fast Food Restaurant	Café	Stadium
1	Austerville	Construction & Landscaping	Women's Store	Garden Center	Dessert Shop	Diner
2	Avondale Road	Boxing Gym	Bakery	Café	Garden	Shopping Mall
3	Bayhead	Gas Station	Garden Center	Department Store	Dessert Shop	Diner
4	Bellair	Train Station	Women's Store	Garden Center	Department Store	Dessert Shop
5	Berea	Coffee Shop	Café	Indian Restaurant	Multiplex	Cosmetics Shop
6	Berea Road	Pharmacy	Convenience Store	Breakfast Spot	Café	Sporting Goods Shop
7	Bishopsgate	Portuguese Restaurant	Train Station	Department Store	Playground	Fast Food Restaurant
8	Bluff	Construction & Landscaping	Burger Joint	Women's Store	Deli / Bodega	Dessert Shop
9	Botanic Gardens	Bakery	Garden	Women's Store	Gas Station	Dessert Shop

Table 4: First 10 rows showing the top 5 popular venues in Durban per neighbourhood

3.3 Machine learning and clustering neighbourhoods

The goal of the project is to identify popular venues in each city. To accomplish this, neighbourhoods have to be grouped in clusters and a cluster with the highest count of popularity for a venue chosen. The **k means** algorithm was selected for use in clustering the neighbourhoods.

The popular venue data for each neighbourhood is categorical and in order to use the k means algorithm, numeric data is needed. **One hot encoding** was used to convert the neighbourhood venue data into a more usable format for analysis.

	Neighbourhood	Accessories Store	African Restaurant	American Restaurant	Arcade	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	...	Trail	Train Station	Turkish Restaurant	Vegetarian / Vegan Restaurant	V S
0	Athlone	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
1	Bakoven	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
2	Bantry Bay	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
3	Bellville	0.0	0.0	0.0	0.0	0.0	0.0	0.090909	0.0	0.0	...	0.0	0.0	0.0	0.0	0
4	Bergvliet	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
81	Walmer Estate	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
82	West Beach	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
83	Woodstock	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
84	Wynberg	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0
85	Zonnebloem	0.0	0.0	0.2	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0

86 rows × 164 columns

Table 5: One hot encoding and mean for each venue per neighbourhood in Cape Town

	Neighbourhood	African Restaurant	Aquarium	Arts & Crafts Store	Athletics & Sports	Auto Workshop	BBQ Joint	Bakery	Bar	Baseball Field	...	Theater	Theme Park	Thrift / Vintage Store	Track	Track Stadium	Trai
0	Addington	0.03125	0.03125	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.03125	0.000000	0.0	0.0	0.0
1	Austerville	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
2	Avondale Road	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.090909	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
3	Bayhead	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
4	Bellair	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65	West Ridge	0.00000	0.00000	0.142857	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
66	West Riding	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
67	Windermere	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.037037	0.0	0.0	0.0
68	Windsor Park	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0
69	Yellowwood Park	0.00000	0.00000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.00000	0.000000	0.0	0.0	0.0

70 rows × 125 columns

Table 6: One hot encoding and mean for each venue per neighbourhood in Cape Town

The number of clusters (**k value**) that are needed must be supplied to the k means algorithm before any clustering occurs. To find the best k value the elbow method was used. For this, the sum of squared distances (distortions/inertias) for a range of k values was calculated and plotted for each city. The following figures show the graphs plotting the sum of squared distances for each k value.

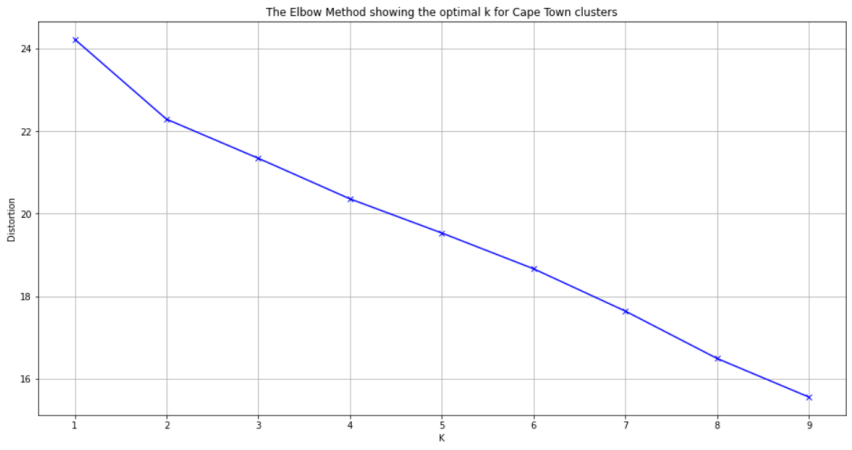

Figure 3: Graph showing distortions vs k values for Cape Town

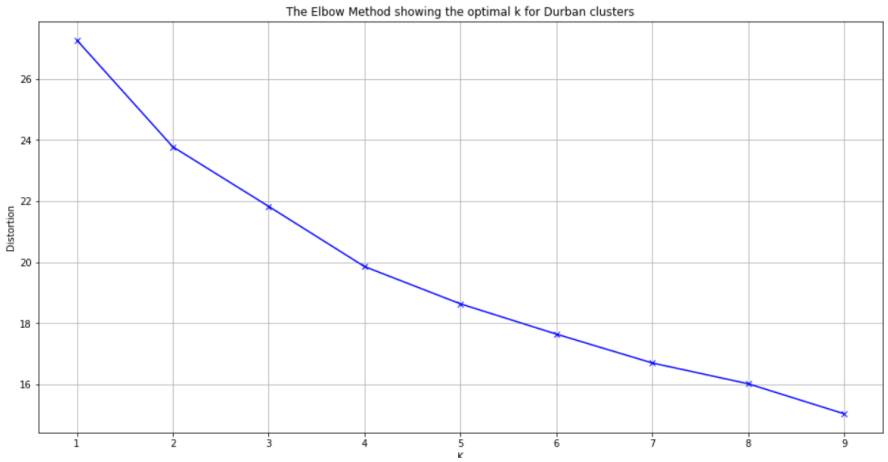

Figure 4: Graph showing distortions vs k values for Durban

From analysing the elbow method graphs above, 7 was chosen for the number of clusters in Cape Town and 7 chosen for the number of clusters in Durban. Using the k means algorithm and 7 as the number of clusters, the following mapped clusters were found.

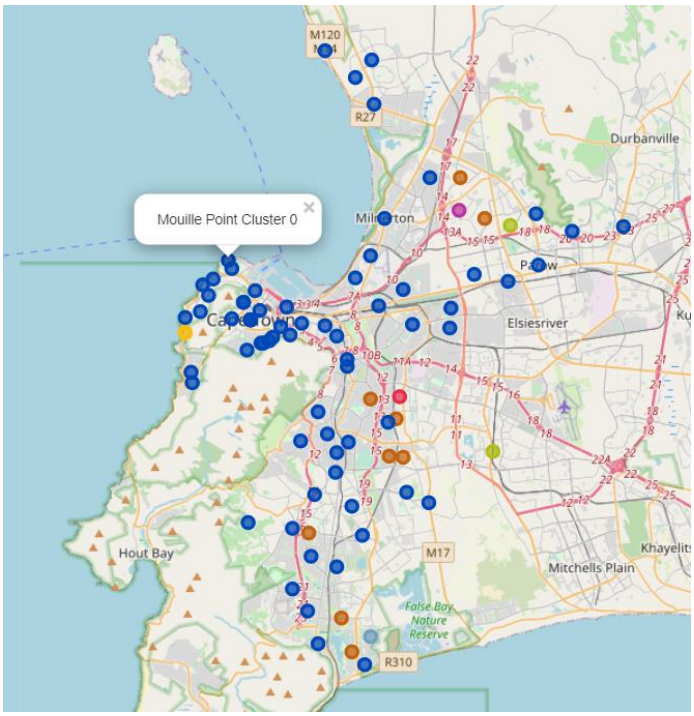

Figure 5: Cape Town venue clusters

Figure 6: Durban venue clusters

3.4 Visualising neighbourhood clusters

7 clusters were created for each city and in each city one of the clusters is spread out more compared to other clusters. The clusters that are the most common (have a high number of popular venues) will be shown in the following graphs.

3.4.1 Popular (large) clusters

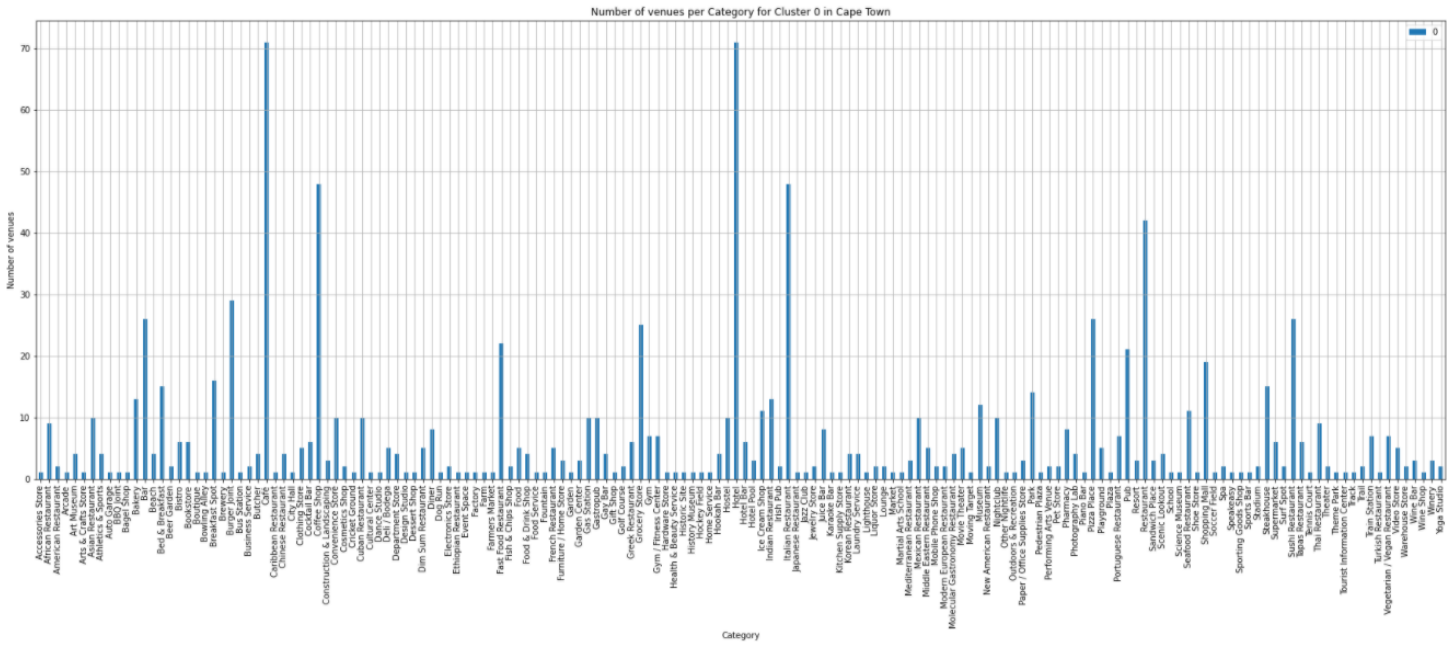

Figure 7: Cluster 0 in Cape Town

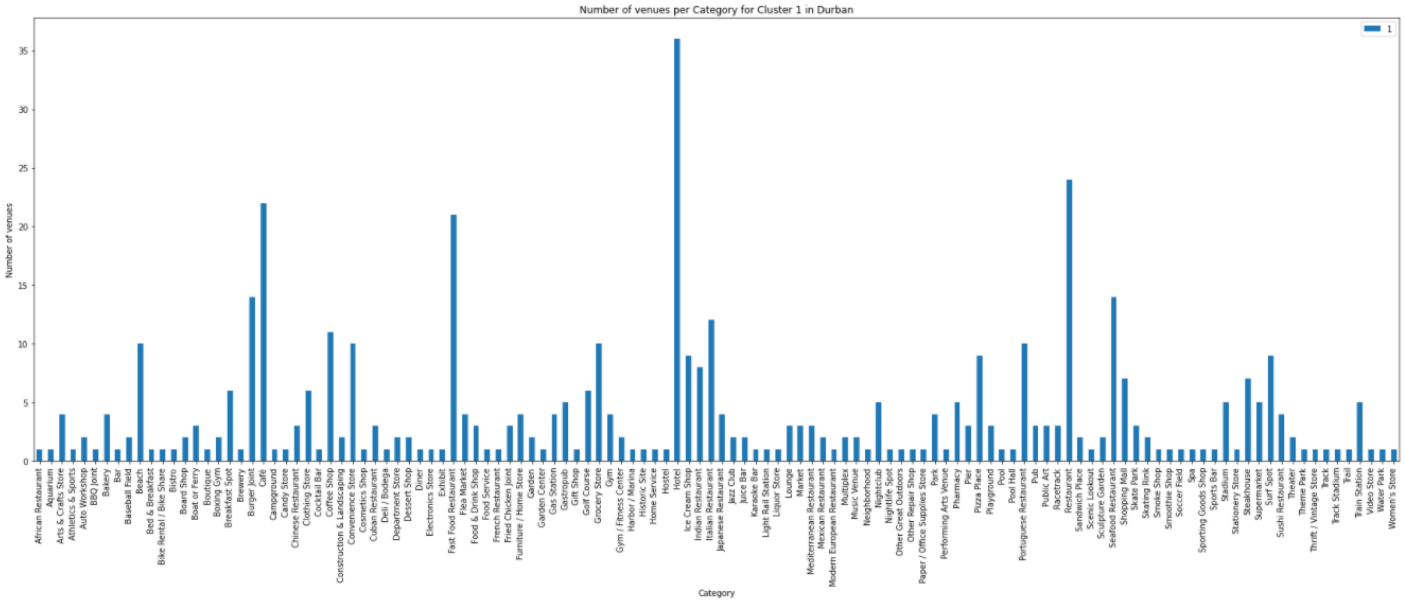

Figure 8: Cluster 1 in Durban

From each cluster, the venue category with the highest number was chosen. From Cape Town cluster 0, there are two venue categories that are the most popular, hotels and cafes with 71 venues each. In Durban cluster one, hotel is the most popular venue category with 36 venues.

3.4.2 smaller clusters

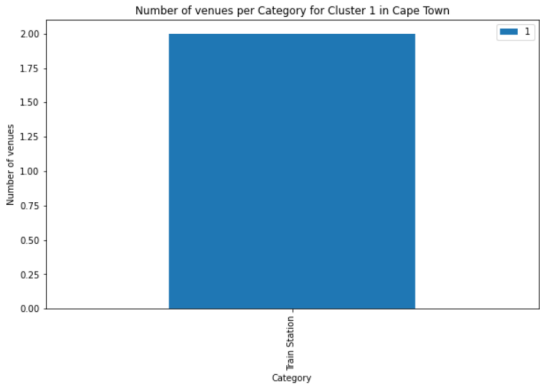

Figure 9: Cape Town cluster 1

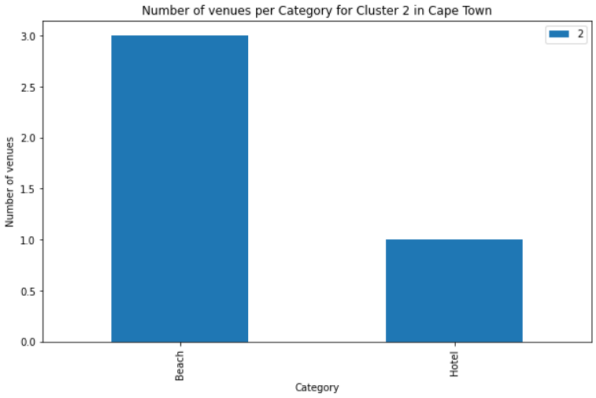

Figure 10: Cape Town cluster 2

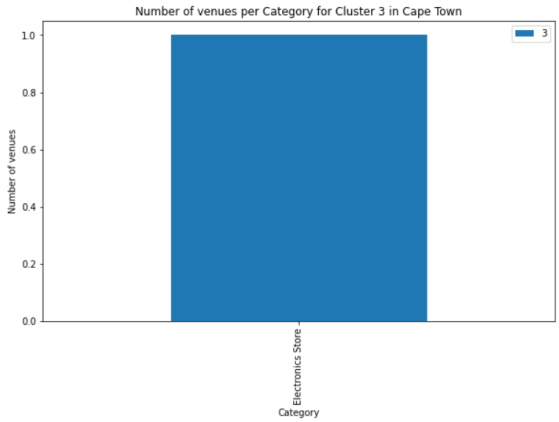

Figure 11: Cape Town cluster 3

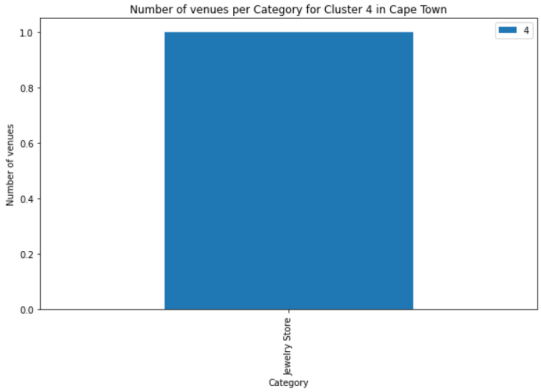

Figure 12: Cape Town cluster 4

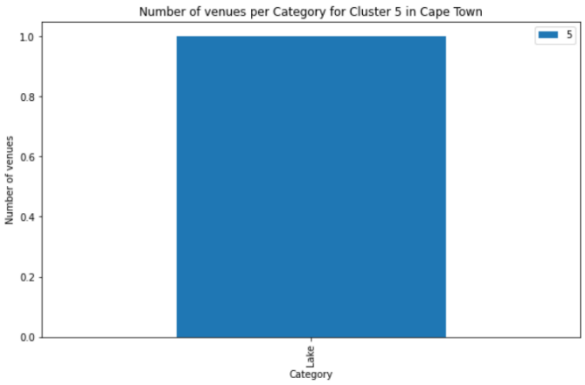

Figure 13: Cape Town cluster 5

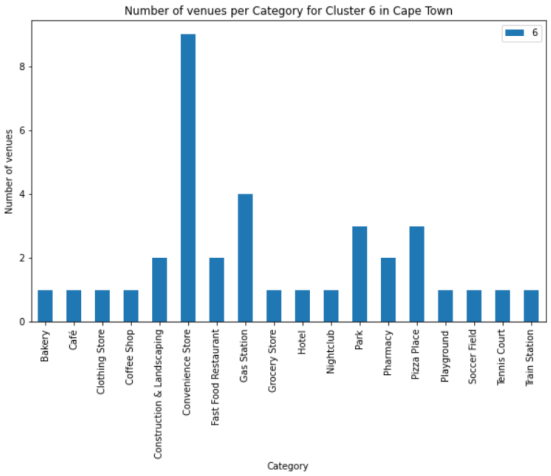

Figure 14: Cape Town cluster 6

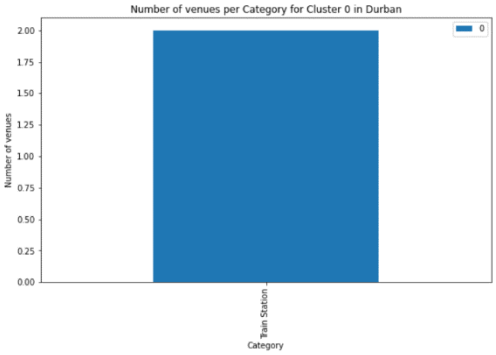

Figure 15: Durban cluster 0

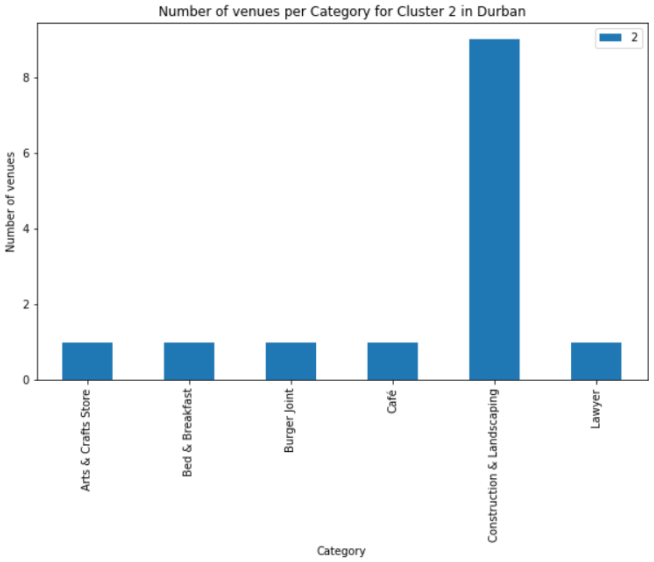

Figure 16: Durban cluster 2

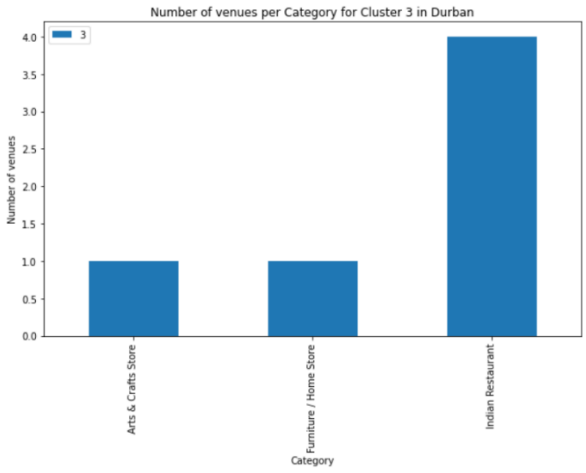

Figure 17: Durban cluster 3

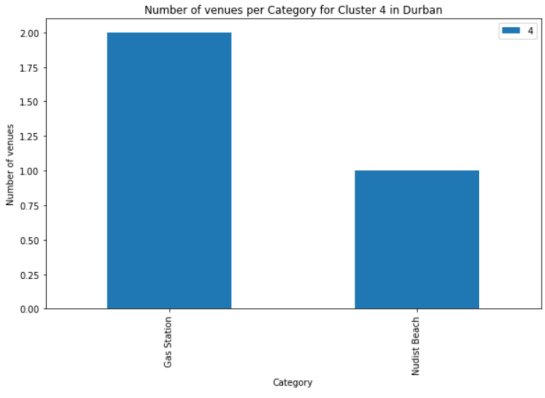

Figure 18: Durban cluster 4

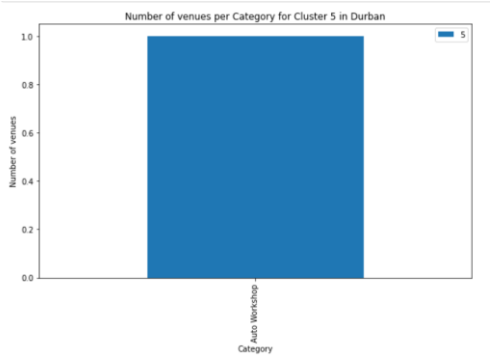

Figure 19: Durban cluster 5

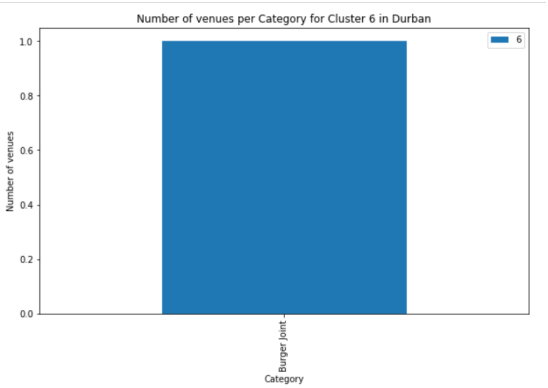

Figure 20: Durban cluster 6

Results

From looking at the maps it can be seen that the k means algorithm created one large cluster for each city and the other clusters were smaller in comparison. These large clusters formed because all of those neighbourhoods share the same popular venues and were grouped together by the algorithm.

From figure 7 it can be seen that hotel and café are the most popular. To make this data more useful, the neighbourhoods that contributed to the popularity of the two venues must be analysed. The following table shows venues that contributed to the popularity of hotels and cafes.

	Neighbourhood	Hotel
0	Bakoven	1
1	Bantry Bay	3
2	Bishopscourt	1
3	Bloubergstrand	1
4	Bo-Kaap	1
5	Camps Bay	7
6	Clovelly	3
7	Darling	3
8	De Waterkant	12
9	Diep River	2
10	Fresnaye	2
11	Gardens	3
12	Green Point	1
13	Kreupelbosch	2
14	Mouille Point	2
15	Ocean View	3
16	Oranjezicht	1
17	Rondebosch East	1
18	Schotsche Kloof	1
19	Sea Point	3
20	Strand	3
21	Sun Valley	3
22	Tamboerskloof	6
23	Three Anchor Bay	3
24	Vredehoek	1
25	Walmer Estate	1
26	Woodstock	1

Table 7: Neighbourhoods that contributed to hotel popularity in Cape Town

	Neighbourhood	Café
0	Bellville	1
1	Bloubergstrand	1
2	Bo-Kaap	5
3	Capri Village	3
4	Claremont	3
5	Clovelly	4
6	Darling	5
7	De Waterkant	10
8	Devil's Peak Estate	2
9	Diep River	1
10	Gardens	4
11	Mouille Point	1
12	Newlands	2
13	Salt River	1
14	Schotsche Kloof	5
15	St James	2
16	Strand	4
17	Sun Valley	4
18	Table View	2
19	Tamboerskloof	3
20	Three Anchor Bay	2
21	University Estate	1
22	Vredehoek	3
23	Woodstock	2

Table 8: Neighbourhoods that contributed to cafe popularity in Cape Town

From table 7 and 8, in Cape Town it can be seen that 27 neighbourhoods contributed to the popularity of hotels while 24 neighbourhoods contributed to the popularity of cafés.

	Neighbourhood	Hotel
0	Berea	1
1	Durban North	1
2	Essenwood	1
3	Glenmore	1
4	Kenneth Gardens	1
5	Marine Parade	6
6	Marlborough Park	6
7	Musgrave Road	1
8	North Beach	5
9	Ocean View	1
10	Pavilion	5
11	Snell Parade	4
12	South Beach	1
13	Windermere	2

Table 9: Neighbourhoods that contributed to hotel popularity in Durban

From table 9, in Durban it can be seen that there are 14 neighbourhoods that contributed to the popularity of hotels.

The k means algorithm also created smaller clusters, as shown in section 3.4.2. These small clusters would not be of interest to investors as the low number of venues per category shows that these venues are not as popular as the venues in cluster 0 of Cape Town (figure 7) and cluster 1 of Durban (figure 8).

Discussion

An observation that can be made from the results and graphs is that cluster 0 of Cape Town and cluster 1 of Durban are much larger than the other clusters in the city's respective maps. This is because the neighbourhoods in those clusters share the same venue popularity. The popularity of a venue is influenced by how many people (foursquare users) check into the venue [8]. So when analysing the data, it has to be taken into account that a venue might seem to not be popular based on foursquare results but actually be popular in the real world as not everyone that visits these

venues is a foursquare user and might not check in to the venue (giving the illusion that a venue is not popular while it actually is).

From the cluster results, only the highest number of venues for a category was considered when choosing a business category. It is also possible to choose a second and third most popular venue category by analysing Cape Town's cluster 0 (figure 7) and Durban's cluster 1 (figure 8).

A recommendation to make from these observations is that it would be useful to use a second API that works similarly to the foursquare API that also allows retrieving popular venues for the same neighbourhoods. The analysis and clustering could be repeated using data from the second API and the results compared to the ones identified in this project. Should they produce similar results then there is a high chance that the popular venues found are indeed popular venues in the cities.

Conclusion

From the results it was identified that in Cape Town hotels and cafes are the most popular with 71 venues each while in Durban the most popular venue are hotels. In Cape Town 27 neighbourhoods contributed to the popularity of hotels while 24 neighbourhoods contributed to the popularity of cafes. From Durban 14 neighbourhoods contributed to the popularity of hotels. The "De Waterkant" neighbourhood in Cape Town has the highest hotel popularity so building near or in that neighbourhood would be beneficial. It can therefore be concluded that hotels should be the category an investor or entrepreneur should choose and start the business in Cape Town as the hotel category is most popular in Cape Town.

References

[1] https://www.southafricavisa.com/cities-to-visit-in-south-africa/

[2] https://www.touropia.com/best-cities-to-visit-in-south-africa/

[3] https://github.com/sihlemkaza/Coursera_Capstone/blob/main/CAPETOWN_CITY.csv

[4] https://github.com/sihlemkaza/Coursera_Capstone/blob/main/DURBAN_CITY.csv

[5] https://developer.foursquare.com/docs/api-reference/venues/explore/

[6] https://geopy.readthedocs.io/en/stable/#nominatim

[7] https://towardsdatascience.com/stop-using-zip-codes-for-geospatial-analysis-ceacb6e80c38

[8] https://silo.tips/download/exploring-venue-popularity-in-foursquare