Segmenting and Clustering hottest suburbs in SA eight (8) metropolitans based on their social activities

Comparing these suburbs social activities with their property values for investment.

Prepared by:

Khumeleni Makungo

IBM Data Science
The Battle of Neighbourhoods

Prepared for:

IBM Data Science

Applied Data Science Capstone
Coursera

July 2021

Key Words: SA's 8 Metropolitans, Suburbs, Segmentation, Clustering, Property Prices,

Table of Contents

1. Introduction

1.1 Background to the study

In south Africa, the constitution provides for three categories of municipalities. There are 278 municipalities in South Africa, comprising of eight metropolitans, 44 district and 226 local municipalities. They are focused on growing local economies and providing infrastructure and service. As a results, the metros contain the hottest cities and suburbs in South Africa with City of Johannesburg, City of Cape Town, City of Tshwane and eThekwini as the four biggest metros containing the hottest suburbs in SA. These metro host cities like Johannesburg, Cape Town, Pretoria and Durban respectively.

As a results we find most SA's residents flocking to these metros from their villages (rural areas) to uncover new treasures like, better education, job opportunities and etc. Eventually, most SA's residents end up residing permanently, buying properties and starting businesses in these cities and thus they become populated. When we think of it from a property investment point of view, we expect investor to buy in areas which are not too populated (for security/safety reasons) and the property value are reasonably low, in addition such area should also have a variety of social places nearby. However, it is not always easy to obtain information that will guide investors in this direction.

1.2 Objectives of this study

1.2.1 Problems to be investigated

The eight (8) metropolitans in South Africa host the hottest cities and suburbs, thus most people flock into these cities and suburbs to uncover new treasure. These cities end up being populated, for instance City of Johannesburg is home to over 5million people and is constantly growing its infrastructures to accommodate new residents.

As a results, property demand grows and mostly people buy where the property value is lower and has a potential to grow in value, and also there are social activities diversity. However, this information is not always readily available and a map showing how the hottest suburbs in SA's metropolitans compares in terms of social activities/venues and property values is vital for property investors and people moving into those suburbs.

1.2.2 Purpose of the study

Given the problems we have in terms of finding people flocking to the cities to uncover new treasure. we can create a map and information chart where the property value is shown and each hottest suburbs within the eight metros is clustered according to the social venue density and similarities.

These informative maps are not readily available, they can assist people when deciding where to invest in property taking social activities and property values into considerations. They can also assist municipalities (metros in particular) to advance their social development by comparing themselves with the metros in terms of how they compare, where they can improve in order to attract more property investor

1.3 Scope and Limitations

The study deals only with the hottest suburbs within the eight (8) metros in South Africa. It focuses only on 34 suburbs as classified as the hottest property investment suburbs in SA by property24 and private-property.

The suburbs in other municipalities which are not classified as metros are not included in this study.

Foursquare API is used to retrieve up to 100 venues within a given radius of each suburbs central latitude and longitude coordinates, this means only up to 100 venues can be used to segment and determine each suburbs clusters.

1.4 Plan of development

The rest of the report is set out as follow:

- Data acquisition and pre-processing
- Analysis of each suburb using pandas and matplotlib
- Suburbs segmentation and clustering using K-Means algorithm
- Analyse the results obtained from K-Means clustering
- Draw conclusion and make recommendations based on the study objectives and analyses.

2. Data acquisition

2.1 Data sources

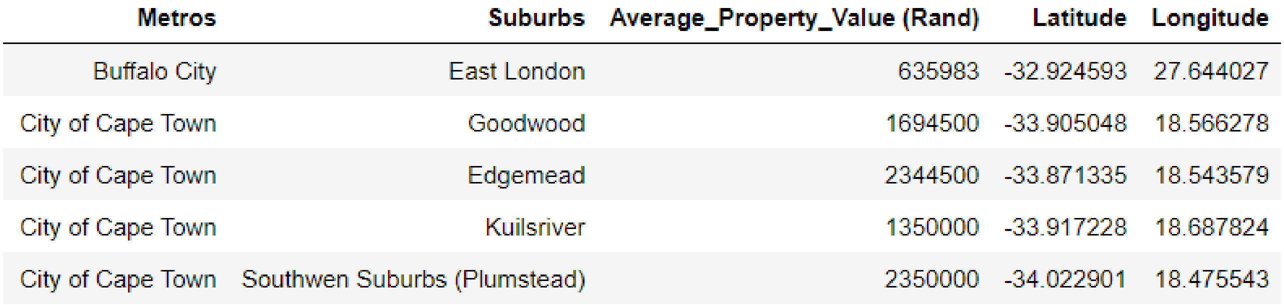
To consider the problem, we can list the data used and its sources as follow:

- The eight (8) metropolitan's average property value was obtain from the national property report, dated 3 March 2021. Then, this was merged with the hottest suburbs' average property prices within these metros in terms of property investment.
- The average property values were retrieved from property24 and private-property websites.
- Used google earth to obtain the central geographical coordinates of these suburbs in these metros
- Foursquare API is used to get the most common venues nearby each of these suburbs up to a radius of 10km, using the central suburbs geographical coordinates obtained from Google maps search engine.

3. Methodology

3.1 Data pre-processing

The data scrapped from property24, private-property websites, google maps search engines and national property report was merged into a dataframe using pandas' library in python.
The dataframe had four columns, containing, metropolitan's name, suburbs name, average property value in that suburb and geographical coordination (Latitude and Longitude). See the figure below



Metros	Suburbs	Average_Property_Value (Rand)	Latitude	Longitude
Buffalo City	East London	635983	-32.924593	27.644027
City of Cape Town	Goodwood	1694500	-33.905048	18.566278
City of Cape Town	Edgemead	2344500	-33.871335	18.543579
City of Cape Town	Kuilsriver	1350000	-33.917228	18.687824
City of Cape Town	Southwen Suburbs (Plumstead)	2350000	-34.022901	18.475543

Figure 1: Dataframe of SA's hottest suburbs within the eight metros

Then foursquare API was used to get the most common venues in these suburbs within a radius of 10km from the central geographical coordinates of each suburbs. Only up to a total limit of 100 venues could be retrieved for each suburb, with 21 out of 34 suburbs reaching the limit of 100 foursquare API venue search. See the figure below

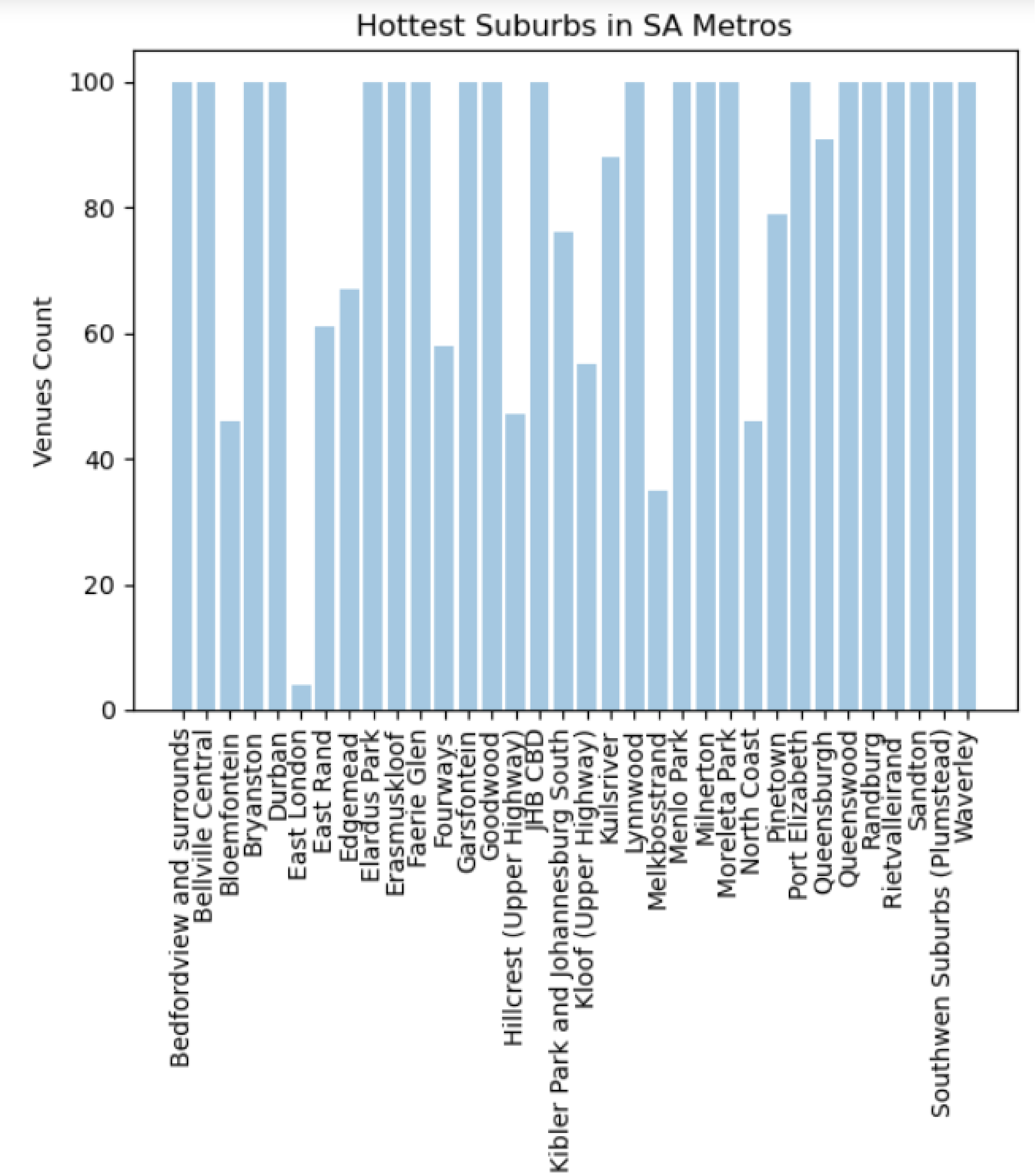
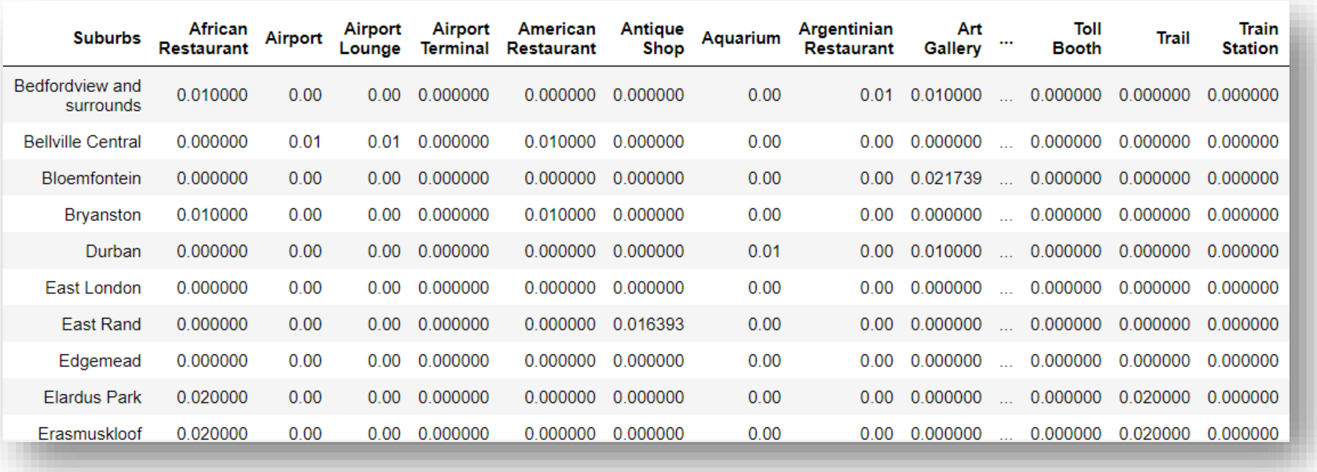


Figure 2: Foursquare API common venue search count within the hottest suburbs

In total, 2853 venues we found and out of these, 170 unique venues categories were found.

3.2 Exploratory Data Analysis by Suburbs

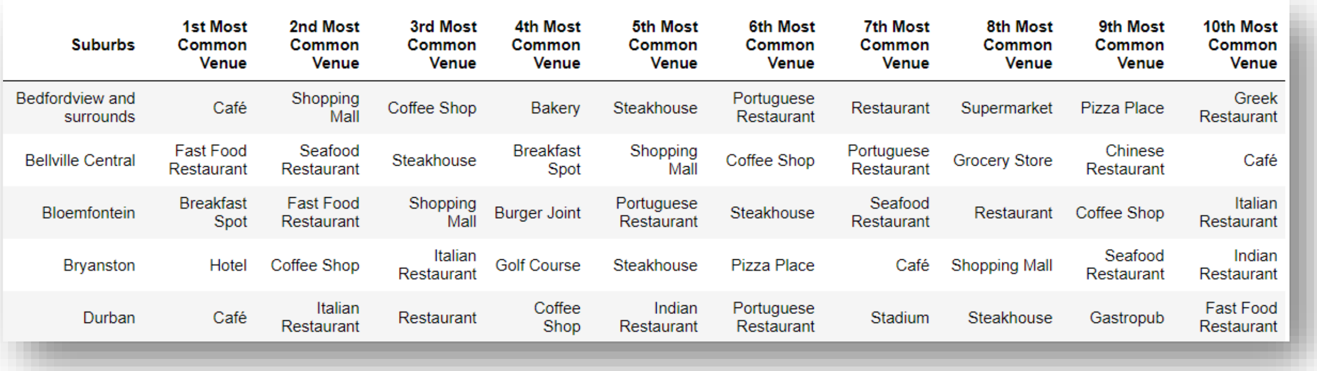
Now that common venues in each suburb have been found, the dataframe was manipulated further by grouping its rows by suburbs and then take the mean of the frequency of occurrence of each venue category. See figure below



Suburbs	African Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Argentinian Restaurant	Art Gallery	...	Toll Booth	Trail	Train Station
Bedfordview and surrounds	0.010000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.01	0.010000	...	0.000000	0.000000	0.000000
Bellville Central	0.000000	0.01	0.01	0.000000	0.010000	0.000000	0.00	0.00	0.000000	...	0.000000	0.000000	0.000000
Bloemfontein	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	0.021739	...	0.000000	0.000000	0.000000
Bryanston	0.010000	0.00	0.00	0.000000	0.010000	0.000000	0.00	0.00	0.000000	...	0.000000	0.000000	0.000000
Durban	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.01	0.00	0.010000	...	0.000000	0.000000	0.000000
East London	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	0.000000	...	0.000000	0.000000	0.000000
East Rand	0.000000	0.00	0.00	0.000000	0.000000	0.016393	0.00	0.00	0.000000	...	0.000000	0.000000	0.000000
Edgemead	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	0.000000	...	0.000000	0.000000	0.000000
Elardus Park	0.020000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	0.000000	...	0.000000	0.020000	0.000000
Erasmuskloof	0.020000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	0.000000	...	0.000000	0.020000	0.000000

Figure 3: Venue category frequency of occurrence within each suburb

From this dataframe, top 10 common venues for each suburb was calculated and added into a dataframe.



Suburbs	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bedfordview and surrounds	Café	Shopping Mall	Coffee Shop	Bakery	Steakhouse	Portuguese Restaurant	Restaurant	Supermarket	Pizza Place	Greek Restaurant
Bellville Central	Fast Food Restaurant	Seafood Restaurant	Steakhouse	Breakfast Spot	Shopping Mall	Coffee Shop	Portuguese Restaurant	Grocery Store	Chinese Restaurant	Café
Bloemfontein	Breakfast Spot	Fast Food Restaurant	Shopping Mall	Burger Joint	Portuguese Restaurant	Steakhouse	Seafood Restaurant	Restaurant	Coffee Shop	Italian Restaurant
Bryanston	Hotel	Coffee Shop	Italian Restaurant	Golf Course	Steakhouse	Pizza Place	Café	Shopping Mall	Seafood Restaurant	Indian Restaurant
Durban	Café	Italian Restaurant	Restaurant	Coffee Shop	Indian Restaurant	Portuguese Restaurant	Stadium	Steakhouse	Gastropub	Fast Food Restaurant

Figure 4: Suburbs vs top 10 common venues

3.3 Suburbs' clustering using K-Means algorithm

K-Means algorithm is one of the most common cluster method of unsupervised learning. It was then used for this study in this project.

First, the Elbow method to find the optimal k value was used and a value of 2 was found. This implied that there are two clusters in the dataset, meaning the 34 suburbs will be segmented into two clusters based on the social venues/activity similarities. See the Elbow results below,

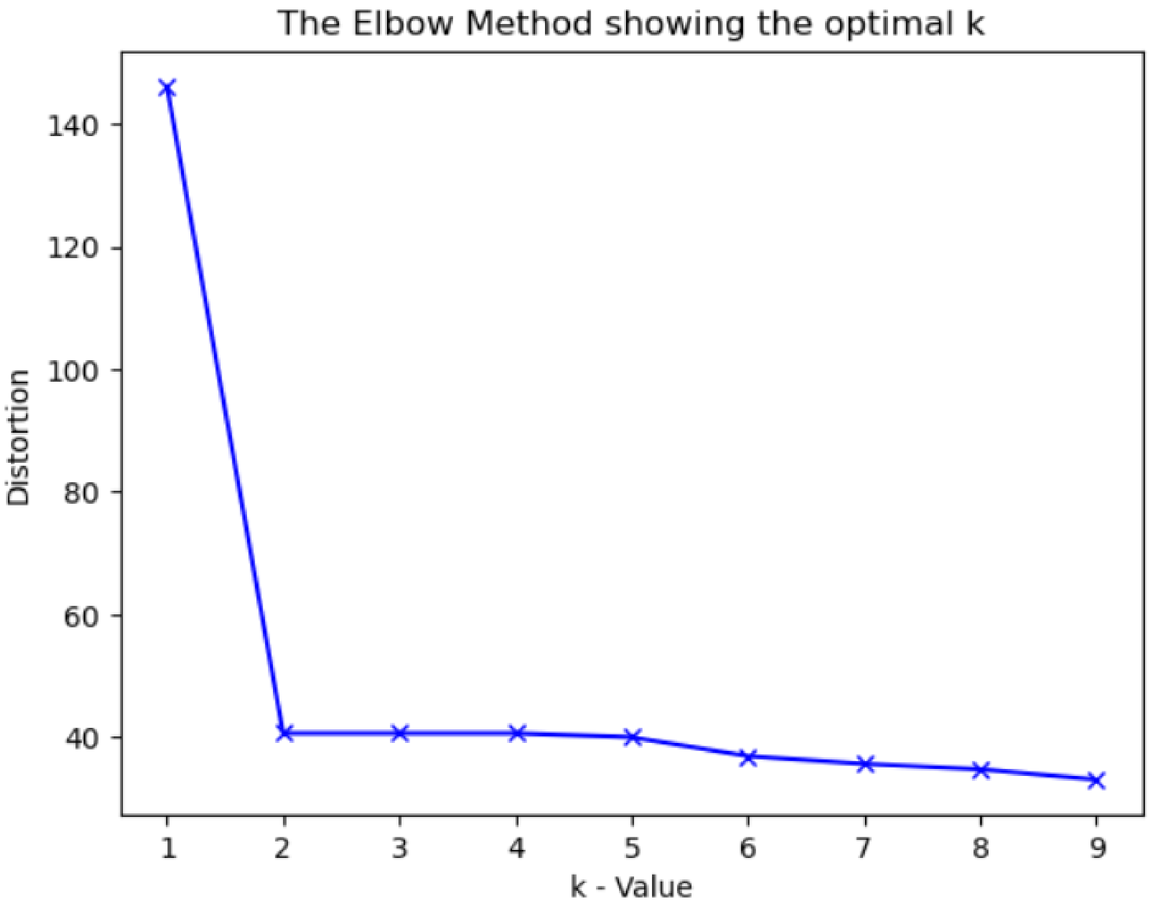


Figure 5: Elbow method showing the optimal k value

The cluster labels were then added to the data frame. Also these two clusters were given names based on the venue category frequency of occurrence, cluster 0, was named "**multiple social venues"** and cluster 1, "**fast food restaurant venues**". Below is the bar chart that accurately demonstrate this

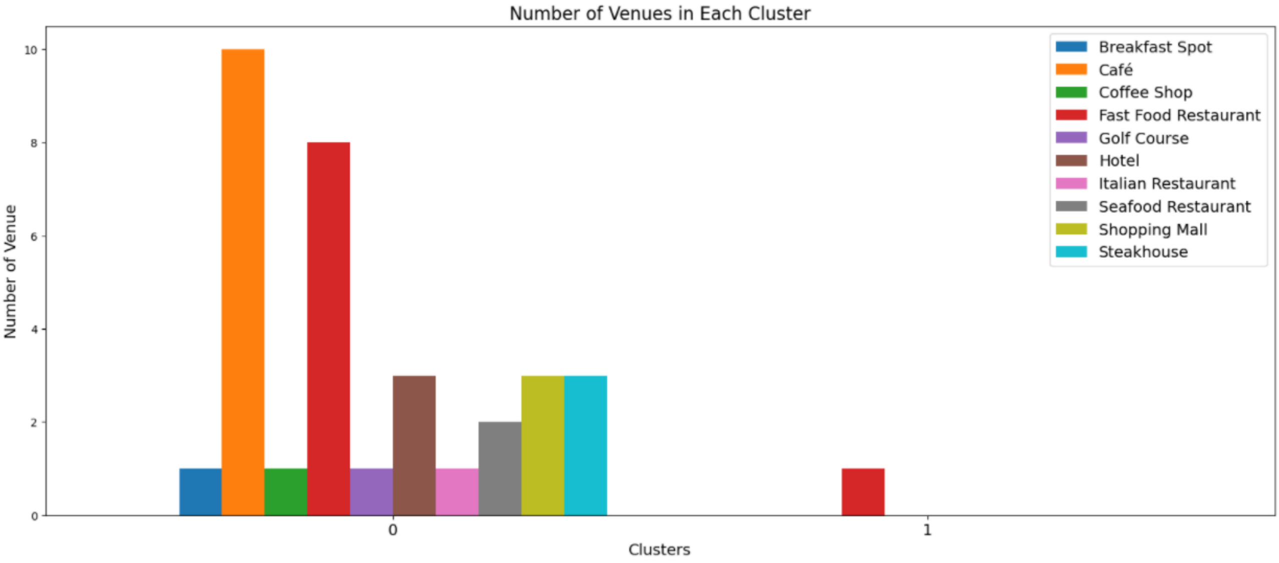
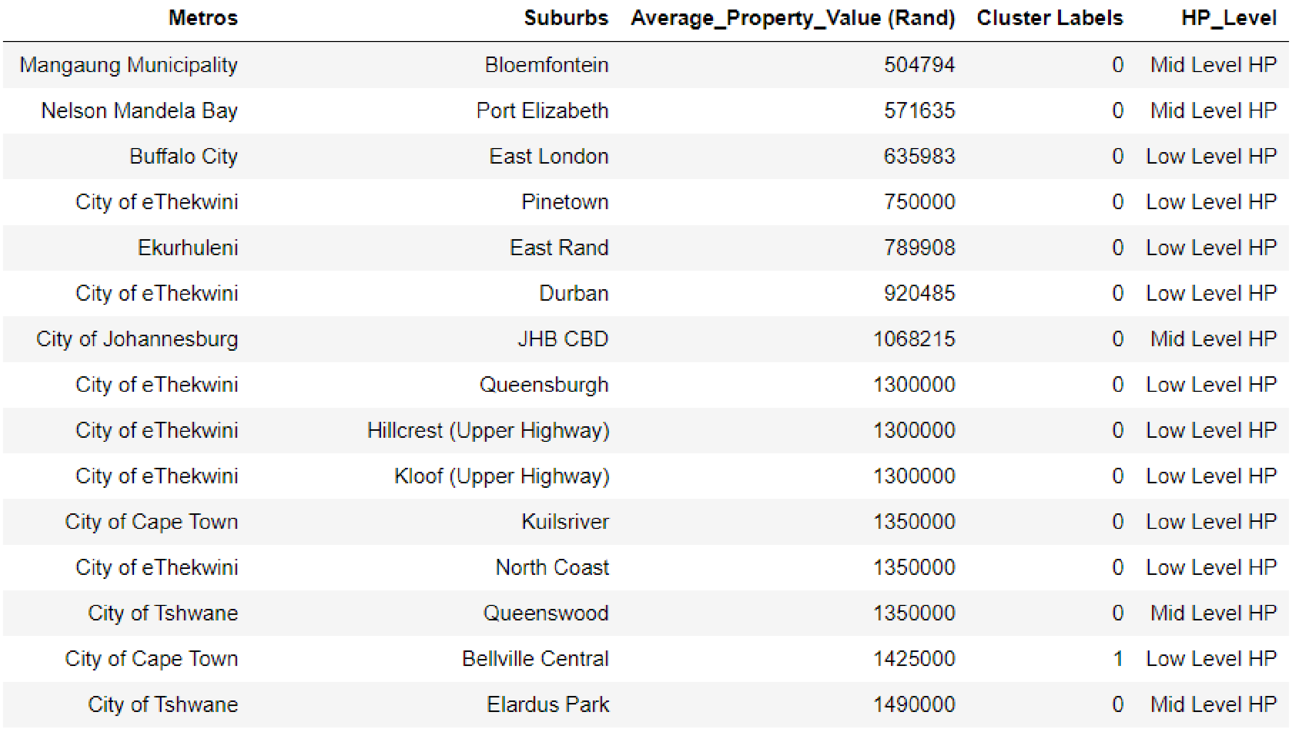


Figure 6: Cluster vs Number of venues

3.4 Housing prices analyses

The housing price from the original dataframe were then analyses and categorized into three categories, namely, low level HP (housing price), mid-level housing price and high-level housing price. This was added into a dataframe together the suburb's cluster number. See the figure below demonstrating this

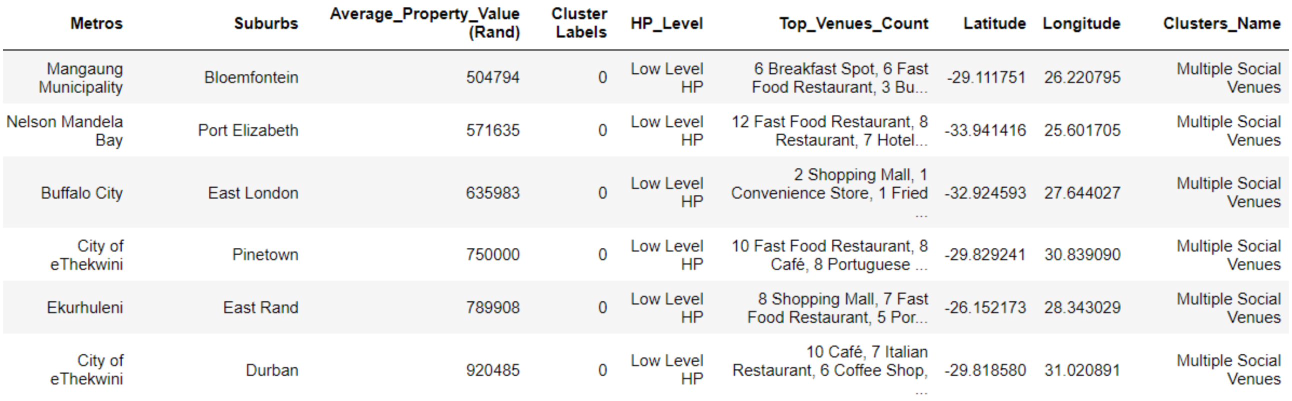


Metros	Suburbs	Average_Property_Value (Rand)	Cluster Labels	HP_Level
Mangaung Municipality	Bloemfontein	504794	0	Mid Level HP
Nelson Mandela Bay	Port Elizabeth	571635	0	Mid Level HP
Buffalo City	East London	635983	0	Low Level HP
City of eThekwini	Pinetown	750000	0	Low Level HP
Ekurhuleni	East Rand	789908	0	Low Level HP
City of eThekwini	Durban	920485	0	Low Level HP
City of Johannesburg	JHB CBD	1068215	0	Mid Level HP
City of eThekwini	Queensburgh	1300000	0	Low Level HP
City of eThekwini	Hillcrest (Upper Highway)	1300000	0	Low Level HP
City of eThekwini	Kloof (Upper Highway)	1300000	0	Low Level HP
City of Cape Town	Kuilsriver	1350000	0	Low Level HP
City of eThekwini	North Coast	1350000	0	Low Level HP
City of Tshwane	Queenswood	1350000	0	Mid Level HP
City of Cape Town	Bellville Central	1425000	1	Low Level HP
City of Tshwane	Elardus Park	1490000	0	Mid Level HP

Figure 7: Housing price vs suburbs clusters

4. Results

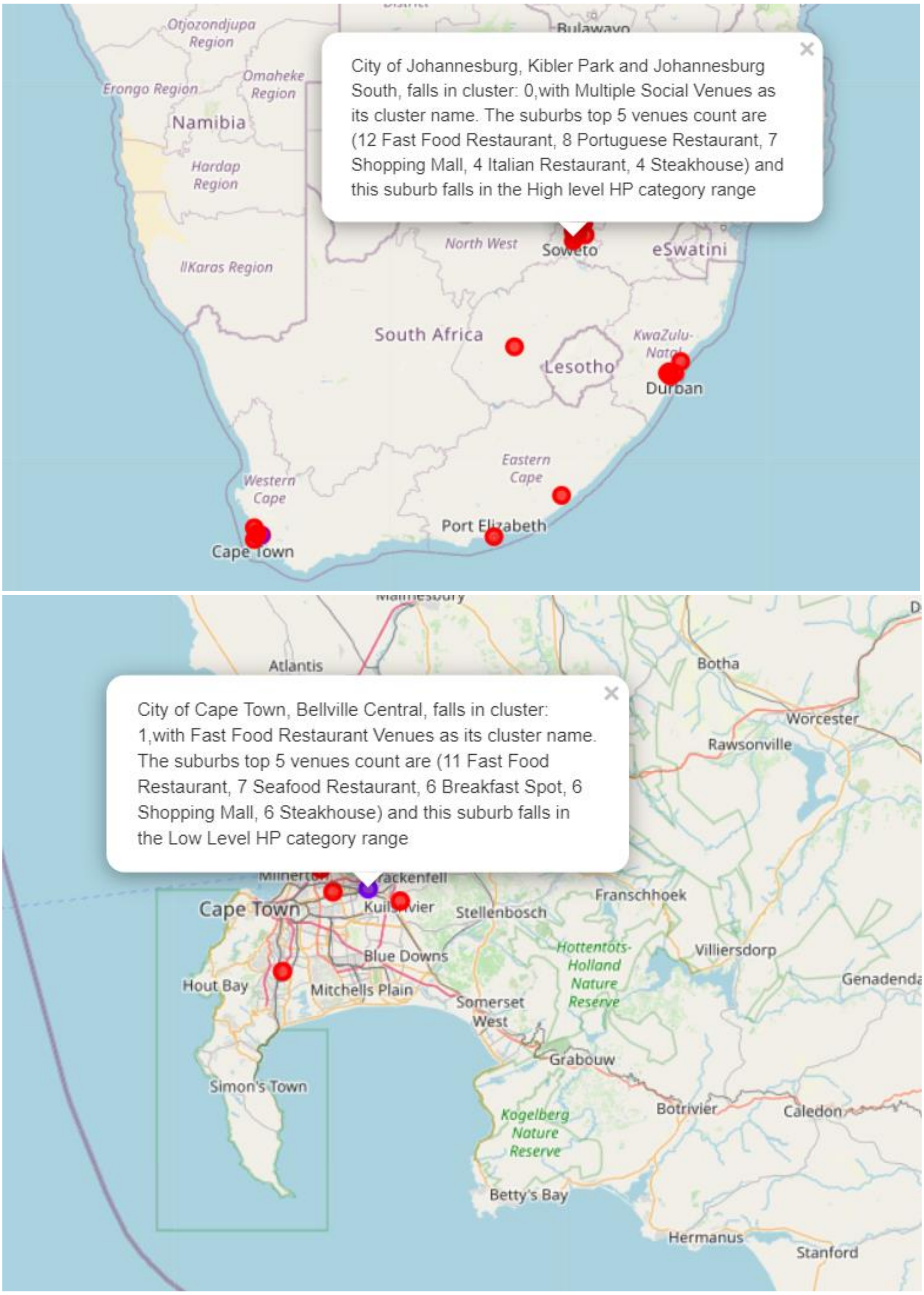
Finally, combined, the metros, suburbs, geographical coordinates, suburb's property value category, top 5 venues count and cluster name into dataframe. See the figure below,



Metros	Suburbs	Average_Property_Value (Rand)	Cluster Labels	HP_Level	Top_Venues_Count	Latitude	Longitude	Clusters_Name
Mangaung Municipality	Bloemfontein	504794	0	Low Level HP	6 Breakfast Spot, 6 Fast Food Restaurant, 3 Bu...	-29.111751	26.220795	Multiple Social Venues
Nelson Mandela Bay	Port Elizabeth	571635	0	Low Level HP	12 Fast Food Restaurant, 8 Restaurant, 7 Hotel...	-33.941416	25.601705	Multiple Social Venues
Buffalo City	East London	635983	0	Low Level HP	2 Shopping Mall, 1 Convenience Store, 1 Fried ...	-32.924593	27.644027	Multiple Social Venues
City of eThekwini	Pinetown	750000	0	Low Level HP	10 Fast Food Restaurant, 8 Café, 8 Portuguese ...	-29.829241	30.839090	Multiple Social Venues
Ekurhuleni	East Rand	789908	0	Low Level HP	8 Shopping Mall, 7 Fast Food Restaurant, 5 Por...	-26.152173	28.343029	Multiple Social Venues
City of eThekwini	Durban	920485	0	Low Level HP	10 Café, 7 Italian Restaurant, 6 Coffee Shop, ...	-29.818580	31.020891	Multiple Social Venues

Figure 8: Final suburbs segmentation and clustering dataframe

The results were then visualized in a map created using folium, and can then be published for SA property investor and municipal managers use. The information displayed on the figure below is now readily available for each suburb within the eight metros in South Africa.



City of Johannesburg, Kibler Park and Johannesburg South, falls in cluster: 0,with Multiple Social Venues as its cluster name. The suburbs top 5 venues count are (12 Fast Food Restaurant, 8 Portuguese Restaurant, 7 Shopping Mall, 4 Italian Restaurant, 4 Steakhouse) and this suburb falls in the High level HP category range

City of Cape Town, Bellville Central, falls in cluster: 1,with Fast Food Restaurant Venues as its cluster name. The suburbs top 5 venues count are (11 Fast Food Restaurant, 7 Seafood Restaurant, 6 Breakfast Spot, 6 Shopping Mall, 6 Steakhouse) and this suburb falls in the Low Level HP category range

Figure 9: Final map of SA's hottest suburbs' property values vs social venues with added vital information

From the results above, one can easily access this information for each suburbs and this can help investors to compare the property prices with social activities within different metros in South Africa.

5. Discussion

As mentioned before, South Africa has 8 metropolitans which host the hottest suburbs in the country. So, these suburbs differ when it comes to average property values but they have very similar social venues or activities. Thus, in this project I wanted to cluster the suburbs in these metros and map them based on the cluster they fall in and also include information like property price /value class, top 5 social activities and other information about the suburb. The main objective was to see if the suburbs are similar based on the social activities and how does the property value/price compares.

I used the K-Means algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 2, meaning there were only two clusters in the dataset. However, only 34 suburbs were used, as classified as the hottest in South Africa when it comes to property investment. For more detailed and accurate guidance, the data set can be expanded and the details of the neighbourhood or street can also be drilled.

I ended the study by visualizing the data and clustering information on the SA_SuburbsMap. In future studies, web or telephone applications can be carried out to direct investors.

6. Conclusions

It can be concluded that South African metros host most of the hottest cities and suburbs, which in turn attract a lot of people to relocate to these areas. As a results, most SA's residents end up residing permanently, buying properties and starting businesses in these cities and thus they become populated.

It can be seen that these cities/suburbs are quite similar in terms of social venues/life, but they have distinct property values for investments. The suburbs with the highest property values falls in the top three metro in South Africa, namely, City of Cape Town, City of Johannesburg and City of Tshwane.

For this reason, SA residents and property investors can achieve better outcomes through their access to the platforms where the property values and social venues information is readily available for suburbs' comparison purposes before buying into.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

7. Recommendations

The model was able to segment the suburbs into two clusters and it only achieved this using common venues for social activities. The results make sense, given the fact that most suburbs tend to have similar social activities to attract more people into investing and some permanently relocating into those suburbs/cities. For more in depth results, more information will need to be used to segment suburbs into different cluster, information like, top class university availability in the metro, nature reserves nearby, some people prefer outdoor activities and nature. This information can further assist investor in comparing suburbs and eventually buying properties in those metros.

Information like quality of transportation networks (roads quality), quality and availability of water and electricity also plays a role in people deciding where to buy property. This information is also vital in segmenting and clustering of these suburbs.