African Cities Clustering and Segmentation

Adilson Pacheco

Wednesday, 13 May 2020

1 Introduction

1.1 Background/ Problem

In recent year, cities all across the African continent have experienced a significant fast-paced development. According to (DW, 2020), by 2050 over 2 billion people will be living in urban areas worldwide with most of this growth taking place in Africa and Asia. Africa as an emergent market full of new opportunities is prompting the attention of investors across the world.

The aim of this project is to understand similarities between cities in Africa using the distribution of venues by category. Performing clustering and segmentation analysis to understand similarities across different fast-growing cities in Africa could help investors decide which cities to invest a specify venue and whether the investment might be profitable using a case study from another city within the same cluster.

The following 5 most developed cities in Africa (Table 1) were selected to conduct this study.

Table 1: Cities selected for the study according to their respective countries.

Cities	Countries
Lagos	Nigeria
Accra	Ghana
Nairobi	Kenya
Kampala	Uganda
Kigali	Rwanda

2 Data Description

2.1 Foursquare. API

The Places API offers real-time access to Foursquare's global database of rich venue data which includes the venue name, venue category, average ratings, and location. This API was used to retrieve venues from each respective capital city using their latitude and longitude.

3 Methods

3.1 Data Sourcing

The first process to this project started with the sourcing of the geographic coordinates for each city. This is achieved by using the Geopy module which allows to geocode addresses, cities, countries and landmarks. Once all cities are geocoded, we create a dataframe such as Table 2 to store the information.

Table 2: Table showing the geographic coordinates for each capital city and their respective countries.

	City	Latitude	Longitude
0	Lagos, Nigeria	6.455057	3.394179
1	Nairobi, Kenya	-1.283253	36.817245
2	Accra, Ghana	5.560014	-0.205744
3	Kigali, Rwanda	-1.885960	30.129675
4	Kampala, Uganda	0.317714	32.581354

We then visualise the data stored in the dataframe using the Folium library package to confirm whether all coordinates where correctly placed in the right location as shown in Figure 1.

Figure 1: Markers showing the location of the selected capital cities.

Using the Foursquare API, we will retrieve all venue location for each city within a 500m radius from the city latitude and longitude coordinate (Table 3).

Table 3: Table shows venue names and their categories for each city retrieved from the Foursquare API.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lagos, Nigeria	6.455057	3.394179	Sakura Japanese Restaurant	6.427309	3.412219	Japanese Restaurant
1	Lagos, Nigeria	6.455057	3.394179	Freedom Park	6.449065	3.396536	Park
2	Lagos, Nigeria	6.455057	3.394179	Film House Cinema	6.490242	3.357371	Multiplex
3	Lagos, Nigeria	6.455057	3.394179	Muson Centre	6.443333	3.401084	Convention Center
4	Lagos, Nigeria	6.455057	3.394179	Wheatbaker Hotel	6.453605	3.445594	Hotel

3.2 Data Cleaning

These venue categories are grouped per city to determine their mean frequency of occurrence in each of the cities (Table 5). Once the mean frequency of venue category occurrence for each city is known, the k-means algorithm is applied to determine the level of similarity between the cities and cluster them accordingly.

Table 4: Table shows venue category according to their mean frequency occurrence in each city.

	City	African Restaurant	Airport Terminal	American Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	...	Shopping Mall	Spa	Sports Bar	Strip Club	Supermarket	Thai Restaurant	Track	Train Station
0	Accra, Ghana	0.045455	0.000000	0.045455	0.00	0.000000	0.00	0.000000	0.000000	0.000000	...	0.075758	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.015152
1	Kampala, Uganda	0.074627	0.014925	0.014925	0.00	0.000000	0.00	0.000000	0.000000	0.000000	...	0.029851	0.00	0.000000	0.000000	0.000000	0.014925	0.00	0.000000
2	Kigali, Rwanda	0.057692	0.000000	0.000000	0.00	0.000000	0.00	0.019231	0.019231	0.038462	...	0.000000	0.00	0.019231	0.000000	0.000000	0.000000	0.00	0.000000
3	Lagos, Nigeria	0.053333	0.000000	0.000000	0.00	0.013333	0.00	0.013333	0.013333	0.000000	...	0.053333	0.00	0.000000	0.013333	0.013333	0.000000	0.00	0.000000
4	Nairobi, Kenya	0.030000	0.000000	0.000000	0.01	0.020000	0.01	0.000000	0.010000	0.010000	...	0.080000	0.01	0.000000	0.000000	0.040000	0.010000	0.01	0.000000

3.3 K-Means Algorithm

K-means clustering is a type of unsupervised learning, which is used when we have unlabelled data (i.e. data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

The algorithm described above finds the clusters and data set labels for a particular pre-chosen *K*. To find the number of clusters in the data, the user needs to run the *K*-means clustering algorithm for a range of *K* values and compare the results.

One of the metrics that is commonly used to compare results across different values of *K* is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing *K* will *always* decrease this metric, to the extreme of reaching zero when *K* is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of *K* is plotted and the "elbow point," (Figure 2) where the rate of decrease sharply shifts, can be used to roughly determine *K*.

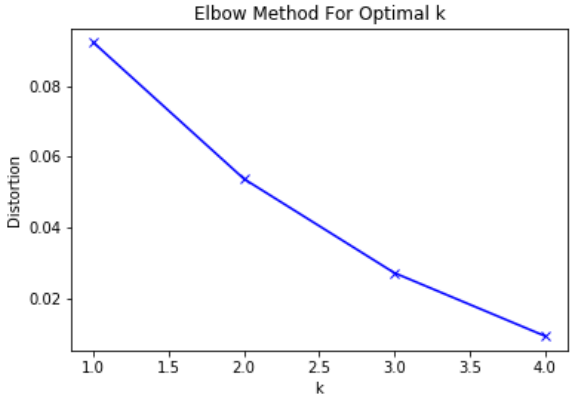

Figure 2: Figure shows the elbow method graph used to determine the optimal value for k.

4 Results

The results shown in Figure 3 and Table 5 suggest that cities such as Lagos, Accra and Nairobi are rather similar when it comes to their venue category. On the other hand, Kampala and Kigali are rather dissimilar between each other and from the other cities, both clustered in different clusters, cluster 2 and cluster 3 respectively.

Figure 3: Figure shows cities clustered according to their venue categories similarity. Marker's colour indicates the cluster.

Table 5: Shows cities according to their respective clusters.

Cities	Countries	Clusters
Lagos	Nigeria	1
Accra	Ghana	1
Nairobi	Kenya	1
Kampala	Uganda	2
Kigali	Rwanda	3

5 Discussion

From Figure 3, we can observe a clear separation between cities in the east and west of Africa. If we have a closer look at cluster 1, we can see that shopping mall occurs amongst frequently among the top 5 venues, which could suggest a potential profitable investment in commercial property for these three cities. Despite been in different clusters, Kampala and Kigali show a high frequency of hotel and resort venues which could suggest that investment towards tourism in those areas could yield profitable returns.

However, due to limitations in data available for these cities, these assumptions may only be partially supported depending on whether the data gathered from the API is in fact representative of the ground truth. If this not the cause, more data from various sources need to be aggregated.

6 Conclusion

Cities in West Africa showed a high mean frequency of shopping mall venues suggesting that investment towards commercial property could yield great return. Cities in East Africa showed a high mean frequency of hotels and resort venues suggesting that investment towards tourism in those areas could yield profitable returns.

Future studies would require comparing these results with other clustering techniques as well as increasing the number of features in the data without increasing the bias of the algorithm.

7 Reference

DW, 2020. *The new African cities of the future | DW | 15.11.2019*. [online] DW.COM. Available at: <https://www.dw.com/en/the-new-african-cities-of-the-future/av-51261472> [Accessed 14 May 2020].