African Cities Clustering and Segmentation

Adilson Pacheco

Background

- By 2050 over 2 billion people will be living in urban areas worldwide with most of this growth occurring in Africa and China.
- Africa as an emergent market full of new opportunities is prompting the interest of investors worldwide.
- Five African cities selected in this project are Lagos, Accra, Nairobi, Kampala and Kigali.

Aim

- The aim of this project is to understand which cities in Africa are similar with regards to venues by category.
- Performing a clustering and segmentation analysis to understand similarities between different clusters of cities in Africa could help investors decide whether a particular investment could be profitable for a particular city.

Data Description

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.

	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

Table 1: Dataframe containing venue names and their categories for each city retrieved from the Foursquare API.

Methods

Data Sourcing

- Geocode cities to retrieve geographic coordinates.
- From Foursquare API, retrieve all venue location for each city within a 100000m radius from the city latitude and longitude coordinate.

Data Cleaning

- Group venues category per city to determine their mean frequency of occurrence in each of the cities.

K-means Algorithm

- Remove all the labels from the data (i.e. removing column names and rows containing labels)
- Find the optimal K value for the number of cluster using the "elbow point"
- Run the k-means algorithm with the optimal k value.

Clean Data Results

Foursquare API

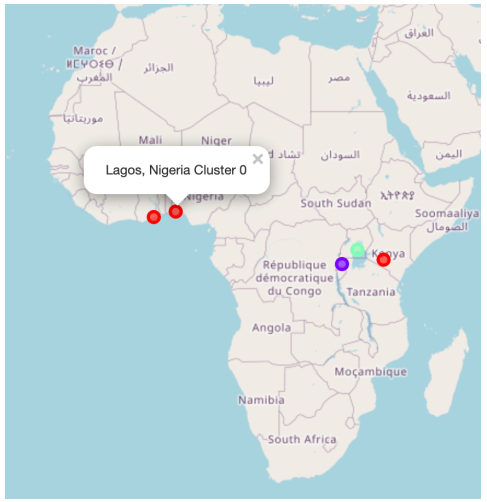
- The cleaned data consists of columns with venue categories and rows with the mean frequency of the occurrence of each category per city.

	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

Table 2: Dataframe containing venue category according to their mean frequency of occurrence in each city.

Predictive Results

- The results suggest that the cities of Lagos, Accra and Nairobi show similar venue categories, all clustered in cluster 1.
- Kampala and Kigali on the other hand are very distinct cities between each other and from the other cities, both clustered in different clusters, cluster 2 and cluster 3 respectively.



Cities	Clusters
Lagos	1
Accra	1
Nairobi	1
Kampala	2
Kigali	3

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.

End