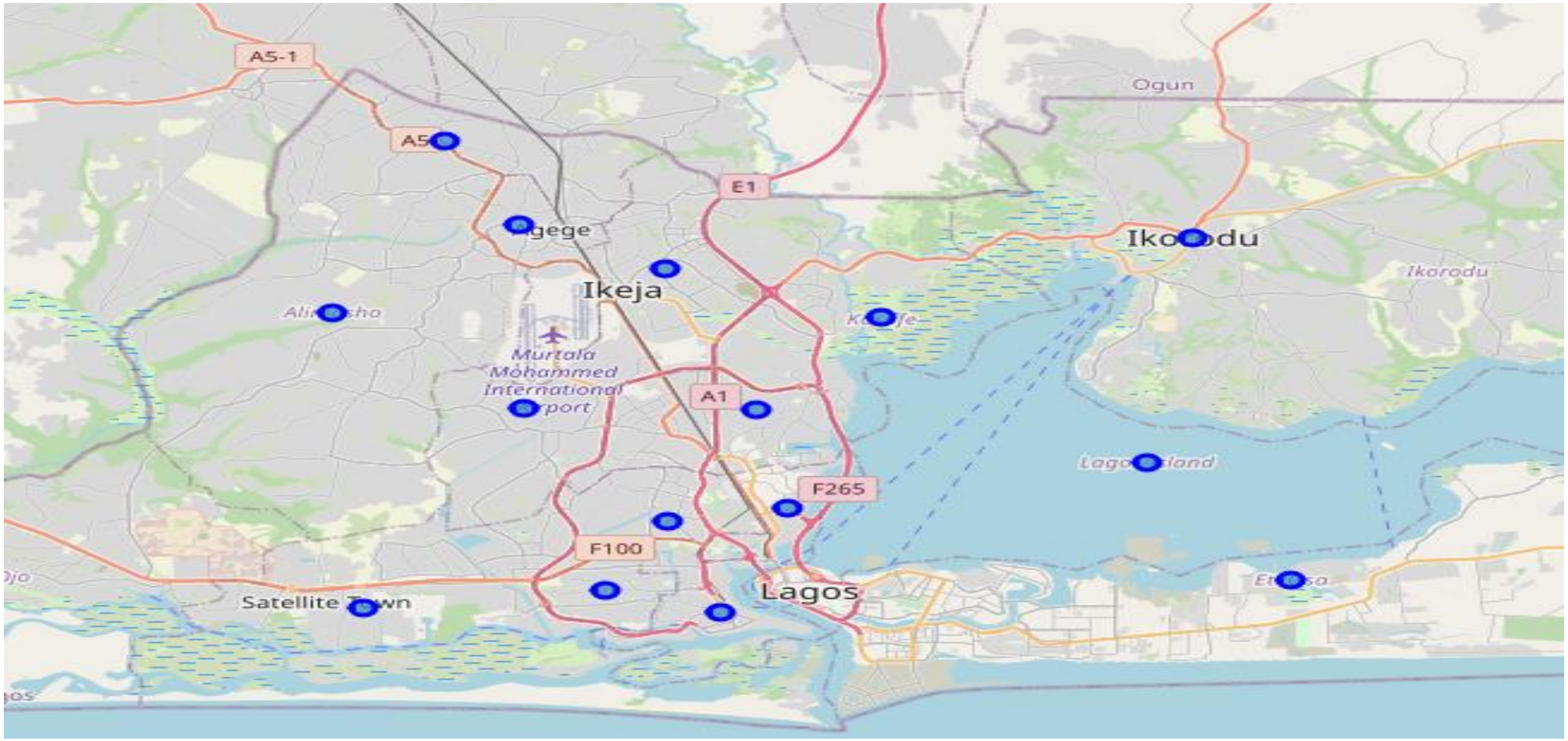
The Battle of Neighbourhoods (Lagos, Nigeria)



Applied Data Science Capstone by IBM on Coursera

(Adebayo AMODU)

Introduction: Problem Background

- This project deals with discussing the neighbourhoods of **Lagos city, The commercial capital of Nigeria** (The Most populous black nation) and the commercial hub of West Africa. It is the 4th wealthiest city in Africa, fourth place behind Johannesburg, Cairo and Cape Town.
- This project would help people planning to start-up business in the city find the most ideal location.

Data Requirements

- The Dataset is the Wikipedia page of Lagos state
- https://en.wikipedia.org/wiki/List_of_Lagos_State_local_government_areas_by_population
- We shall explore Lagos city through its respective Local Government Areas (LGA) or Boroughs. The above link is a web page that shows the respective boroughs in Lagos State and each population figure. It's a public Wikipedia data page.
- In order to obtain the venue details in each neighbourhood Foursquare API is used.
- https://foursquare.com/

Data Requirements

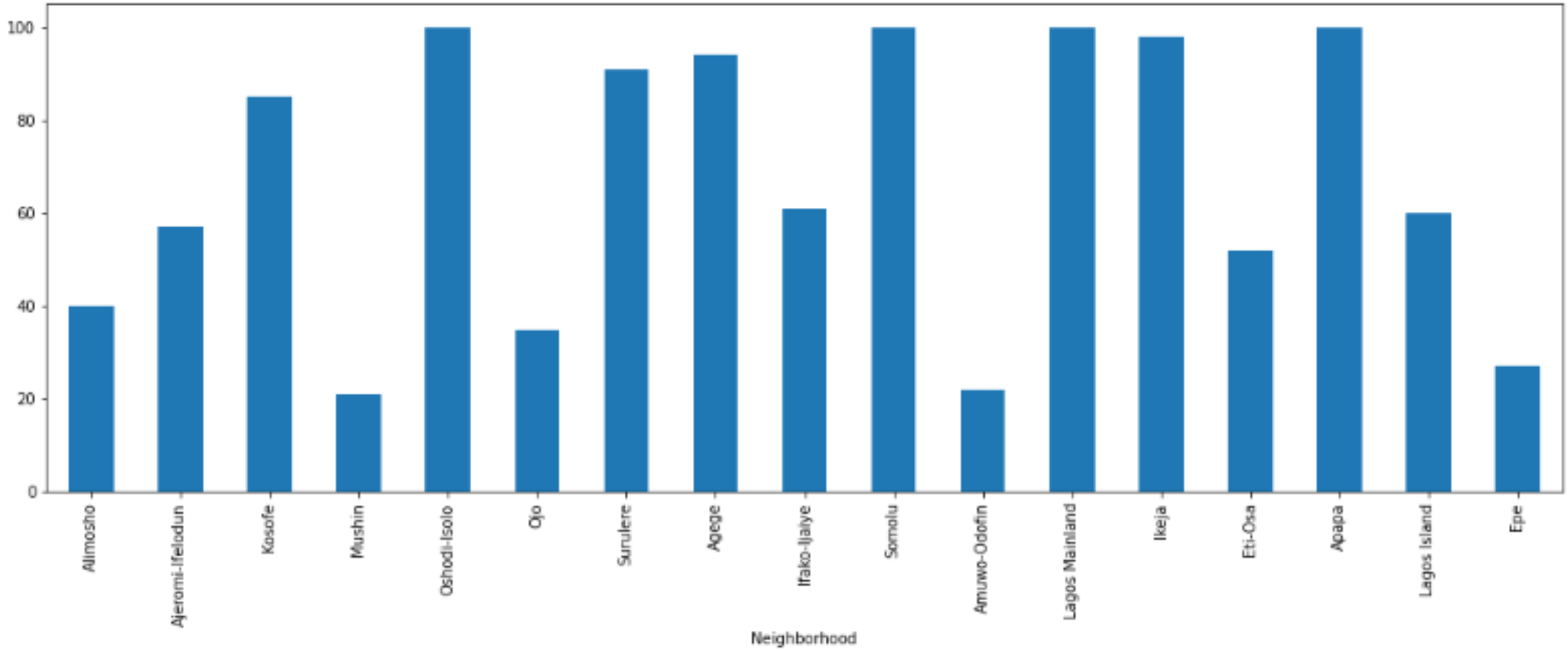
- A total of 1144 venues data have been obtained from Foursquare. The resultant venues dataset, (shown in Fig) is used for the analysis process.
- The following steps are taken to analysed the data:
  - We scrape the web page using the beautiful soup library
  - Follow by using Foursquare API calls to retrieve geolocation data and then fetch the text data using the requests library
  - We than convert it from JSON to Pandas data frame using the json_normalize module
  - We shall use the folium library to render the maps and plot these via The Matplotlib library. Then we shall explore respective boroughs and analyse each area as a location for Start-up based on the aforementioned parameters.
  - After this, we shall select our top location.

Methodology

- Now, with the neighbourhood's data we also have the most popular venues in each neighbourhood obtained using Foursquare API. A total of 1144 venues have been obtained in the whole city and 135 unique categories. But as seen we have multiple neighbourhoods with less than 10 venues returned. In order to create a good analysis let's consider only the neighbourhoods with more than 10 venues.
- We can perform one hot encoding on the obtained data set and use it find the 10 most common venue category in each neighbourhood. Then clustering can be performed on the dataset. Here K - Nearest Neighbor clustering technique have been used. To find the optimal number of clusters silhouette score metric technique is used.
- The clusters obtained can be analysed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

Analysis

- Looking into the dataset we found that there are neighbourhoods with less than 20 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighbourhoods from which 20 or more than 10 venues were obtained.



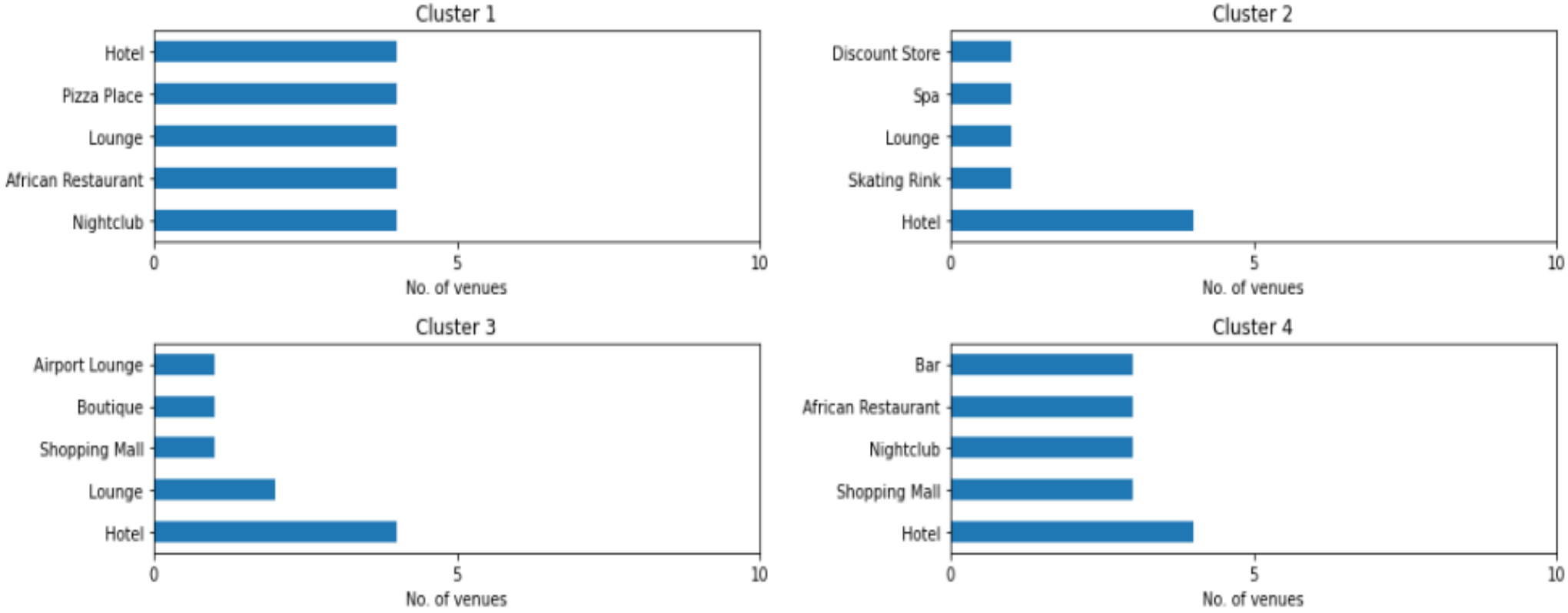
Analysis

- One hot encoding is performed on the filtered data to obtain the venue categories in each neighbourhood. Then group the data by neighbourhood and take the mean value of the frequency of occurrence of each category.
- This is used to obtain the top 10 most common venues in each neighbourhood i.e. the 10 venues with the highest mean of frequency of occurrence.
- The resultant dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters.
- For optimal result we used the best value for K.

Results and Discussion

- Using the clusters and the top venue categories let's visualize the top 5 venue category in each Cluster for comparison.

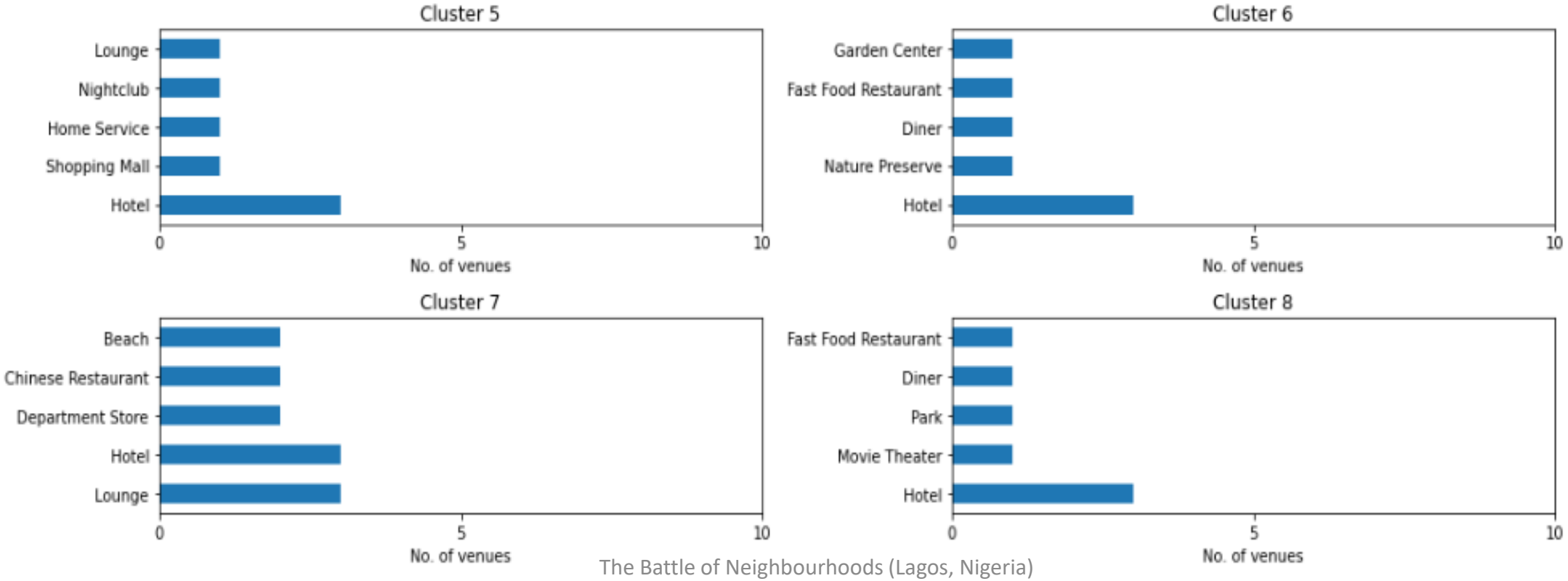
<u>Clusters 1 to 4:</u>



Results and Discussion

- Using the clusters and the top venue categories let's visualize the top 5 venue category in each Cluster for comparison.

Clusters 5 to 8:



Results and Discussion

- This plot can be used to suggest valuable information to Business persons. Let's discuss a few examples considering they would like to start the following category of business.

1. Ice Cream Shop

The only neighbourhoods cluster with the highest number of ice cream shop is 4 while the rest has few or none. Hence opening one in others will be a likely good move.

2. Pizza Place

Just like ice cream shop, only neighbourhoods cluster 3 highest number of Pizza Place while the rest has few or none.

Results and Discussion

3. African Restaurant

Since Lagos is one of the biggest cities in Africa, once would expert all neighbourhoods cluster to have good number of African restaurants, but base on this analysis only cluster 2 and 7 have.

Word Cloud

The most common or frequent venues in all of Lagos State

CONCLUSION

- Purpose of this project was to analyze the neighborhoods of Lagos, Nigeria and create a clustering model to suggest personals places to start a new business based on the category.
- The neighborhoods data was obtained from an Wikipedia and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 10 venues returned. In order to build a good model, we filtered out these locations. The remaining locations were used to create a clustering model.
- The best number of clusters was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster.
- A few examples for the applications that the clusters can be used for have also been discussed. Stakeholders can use this project to decide the location for the particular type of business.

DRAWBACKS

- A major drawback of this project was that the Foursquare API returned only few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.

Thank you!