Coursera Capstone:

IBM Applied Data Science Project

Situating a New Shopping Mall in Lagos, Nigeria

1 Introduction:

Shopping malls have taken more and more place in our economic and social life and have also taken the place of traditional bazaars. Shopping malls are not only places to shop; they also have become places for having fun and spending time. Consumers spend a considerable time in shopping malls because they host a lot of stores and activities that attract consumers' attention and offer a lot of benefits. Shopping malls in Lagos, Nigeria have become the most attractive places to shop and hang out. Lagos consumers do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services.

Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Lagos and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem:

The objective of this capstone project is to analyze and select the best locations in Lagos, Nigeria to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question*:*

"In Lagos, Nigeria, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?"

"Where are the most promising neighborhoods with least competition and most likely business success?"

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in Lagos, Nigeria. This report seeks to clarify the most promising neighborhoods with least competition and most likely business success.

2 Data:

To solve the problem, we will need the following data:

Area data of Lagos, Nigeria. This defines the scope of this project which is confined to the city of Lagos, Nigeria, the capital of the country of Malaysia in West Africa+.

Spatial data of the Lagos. Latitude and longitude coordinates of those areas, this is required in order to plot the map and also to get the venue data.

Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the areas.

Sources of data and methods to extract them:

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_Lagos_State_local_government_areas_by_population) contains a list of local government arears in Lagos Nigeria, with a total of 20 LGAs. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of pandas packages. Then we will get the geographical coordinates of the areas using Python Geocoder package which will give us the latitude and longitude coordinates of the areas. After that, we will use Foursquare API to get the venue data for those areas. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

3 Methodology:

Firstly, we need to get the list of areas in Lagos state. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_Lagos_State_local_government_areas_by_population). We will do web scraping using Python ***pandas*** package to extract the list of areas data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful ***geocoder*** package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the areas in a map using ***folium*** package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by geocoder are correctly plotted in the areas of Lagos. Next, we will use ***Foursquare*** API to get the top 100 venues that are within a radius of 4000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the areas in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each areas and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each areas by grouping the rows by areas and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the areas. Lastly, we will perform clustering on the data by using *k-means* clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the areas into 3 clusters based on their frequency of occurrence for "Shopping Mall". The results will allow us to identify which areas have higher concentration of shopping malls and which areas have fewer number of shopping malls. Based on the occurrence of shopping malls in different areas, it will help us to answer the question as to which areas are most suitable to open new shopping malls.

4 Results:

The results from the k-means clustering show that we can categorize the areas into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Areas with high number of shopping malls
- Cluster 1: Areas with low number to no existence of shopping malls
- Cluster 2: Areas with moderate concentration of shopping malls

The results of the clustering are visualized in the map below in the notebook project.

5 Discussion:

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Lagos city, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to no shopping mall in the areas. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 0 are likely experiencing intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in areas in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in areas in cluster 1 with moderate competition. Lastly, property developers are advised to avoid areas in cluster 0 which already have high concentration of shopping malls and experiencing intense competition.

6 Limitations and Suggestions for Future Research:

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

7 Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The areas in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.