IBM Applied Data Science Capstone

Building a New Mall In Riyadh, Saudi Arabia

(Report)

By:

Marwan Akeel

Introduction

Shopping malls are among the favorite places for hanging out, shopping, dining at restaurants watching movies, and walking specially in hot areas like Riyadh where people are searching for cool places. There are many activities to do as couples, families, or for kids. Malls are huge, spread over a wide area in order to have many different stores and enough area for activities and walk. Retailers also search for crowd gathering malls specially that are in central locations. Malls increases with the increase in population, so constructing malls became a big business for real estate developers who are looking for a stable and guaranteed source of income. One of the important keys to the success of malls is their location.

This Project is done in the purpose of practising data science methodology.

Business Problem

The business question we are trying to answer in this assignment is: What is the best location to build a new mall in Riyadh, Saudi Arabia?

We have to analyze the data and select the best locations using data science methodology and machine learning such as clustering.

The Target Audience

This Project targeted the real estate developers how are interested in constructing new mall(s) in Riyadh.

Data Types

In this project we need three types of data:

1. Riyadh Neighborhood names. 100 names were used.

2. Neighborhood latitudes and longitudes.
3. Neighborhood venues (maximum 100 avenues/neighborhood).

Data Sources

We downloaded Riyadh neighborhood names from Wikipedia, which come within an html code (https://en.wikipedia.org/wiki/Riyadh). It is not up to date but still appropriate.

The data were extracted using web scraping techniques (Python request) and beautifulsoup package. We got the latitude and longitude of the Riyadh city using the geocoder package. The neighborhoods latitude and longitude were collected using (https://www.coordinatesfinder.com) and google map. Then, we used the foursquare API to obtain the venues of each neighborhood. Foursquare is a famous database that store geographical data containing millions of places from around the world. Foursquare database is not well covering the venues in Riyadh and Saudi Arabia. The venues returned by the foursquare are only the international famous brands and the most local well known brands locations. These data is sufficient for practising. The data returned by foursquare is the venue's name, district ,tips, location, and others.

Methodology

- We Started the work by downloading the neighborhood names from wikipedia. There are 101 neighborhood names. We used requests library for that. These names are included in list in html file. We used BeautifulSoup package to extract the neighborhood names.
- Riyadh latitude and longitude were found by using Geocoders, which convert an address into latitude and longitude values.
- The list of neighborhood names were passed manualy to coordinatesfinder.com which returned the latitude and longitude of the neigborhood. Some of these latitudes and longitudes were mismatched the reality. It was corrected using google map.
- All the latitudes and longitudes were saved into an Excel file. The file was imported.

- K-means clustering technique was used to group similar neighborhoods. We were interested only in malls number and the total venues in each neighborhood. More venues means that the neighborhood tends to be a commercial area. So the training run using these two features.
- Exploring were done using folium package.

Results

After training our model using the number of shopping malls and venues features, clustering output result falls in 5 categories:

Cluster 0 - Low shopping malls, High number of venues
Cluster 1 - Low shoppinng malls, Low number of venues
Cluster 2 - High number of shopping malls, High number of venues
Cluster 3 - Moderate number of shopping malls, Low number of venues
Cluster 4 - Below moderate number of shopping Malls, Moderate number of venues

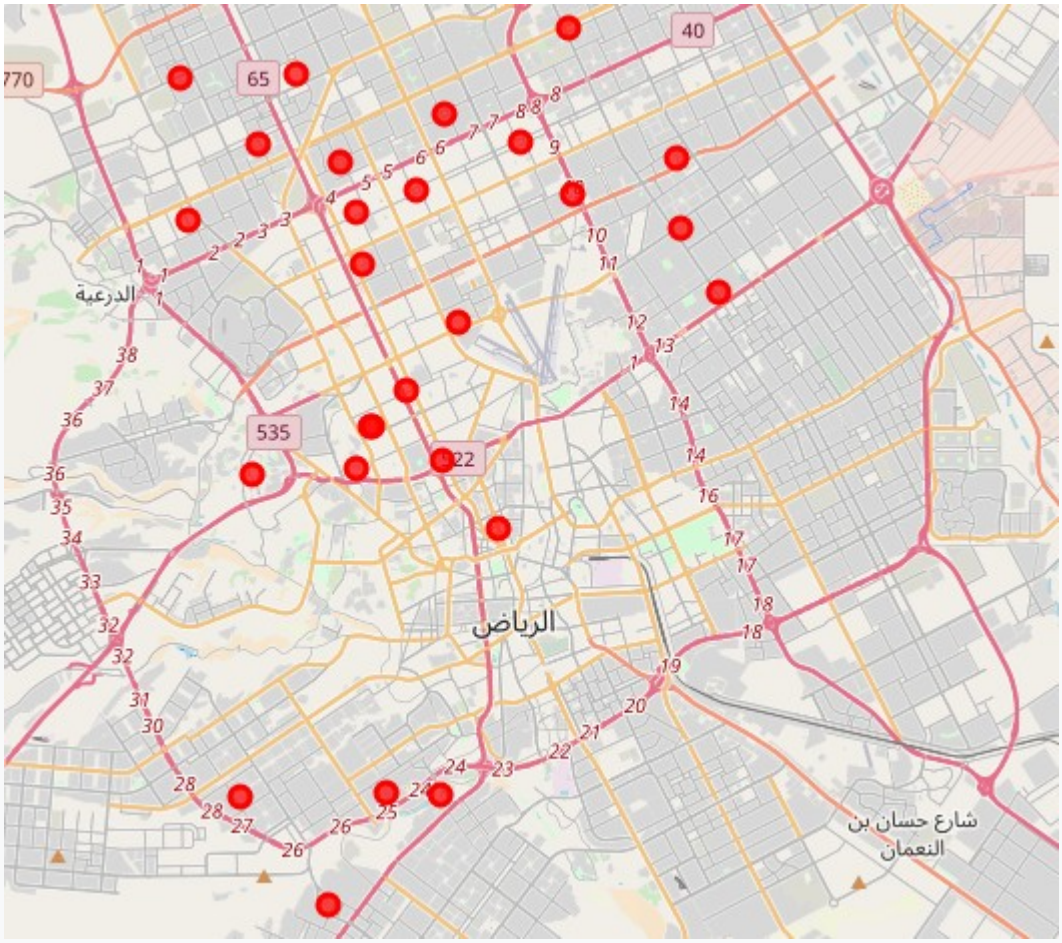
The following map is the visualization of the clusters. Red represents cluster 0, purple represents cluster 1, light blue represents cluster 2, light green represents cluster 3, light brown represents cluster 4.

We take the large number of venues in a neighborhood as a sign of possibly a commercial/highly-populated area, and vise versa.

Discussion

As we can see in the map, the neighborhoods which has the highest number of both shopping malls and venues are located at the center of the city or the old part (cluster 3 – light green). The areas that have low number of shopping malls and venues are the neighborhoods that have low population or located at the city end (cluster 1 – purple). New shopping malls can be built in a neighborhood that has high

number of venues but low number of shopping malls. These areas can be found in cluster 0 (red). Below is the visualization map to the candidates neighborhoods.



Conclusion

Training our data using clustering technique has resulted in 5 categories. It is recommended to build a new shopping mall in a neighborhood that has high number of venues and low number of shopping mall. These neighborhoods are found in cluster 0.