Applied Data Science Capstone

By IBM

Opening a new mall in the city of Bangalore

The Battle of Neighbourhood

REPORT

By Shoban Dinesh

1. Introduction

The objective of this capstone project is to analyse and select the best location in the city of Bangalore to open a new shopping mall. Using data science techniques, the project aims to provide a solution to answer the business question: where would you recommend a property developer to open a new shopping mall?

The project will use data visualization techniques to enable the stakeholder visualize the locations of the neighbourhoods geographically. And also use machine learning to group neighbourhoods in a way to allow a property developer can a better choice in picking a neighbourhood to open a new mall.

2. Data

The following data were required to solve the problem:

- List of all neighbourhood in the city of Bangalore
- The latitude and longitude coordinates of each of the neighbourhood.
- Venue data and data related to the shopping mall in these neighbourhood.

Sources and Methods:

The list of all the neighbourhood was taken from the Wikipedia page(https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore) were all the neighbourhoods were mentioned.

I used web scraping techniques to extract data which the website, with the help of Python's libraries namely Beautiful Soup and requests.

The latitudes and longitudes of the individual neighbourhoods were obtained by using the Python Geocoder package.

And for the venue data of each neighbourhood, I used Foursquare API. Foursquare API will provide many categories of the venue data, but the category I needed to solve the business problem was 'shopping mall' data particularly.

3. Methodology section

The project will be deal with many data science skills. Firstly, web scraping. We will extract the list of all neighbourhoods in the city of Bangalore using the Python library Beautiful Soup.

Secondly, we will Python’s Geocoder library to get the geographical coordinates of each of the neighbourhood.

We then plot these coordinates on the map of Bangalore to help the stakeholder visualize the locations geographically. This is done using the library called Folium.

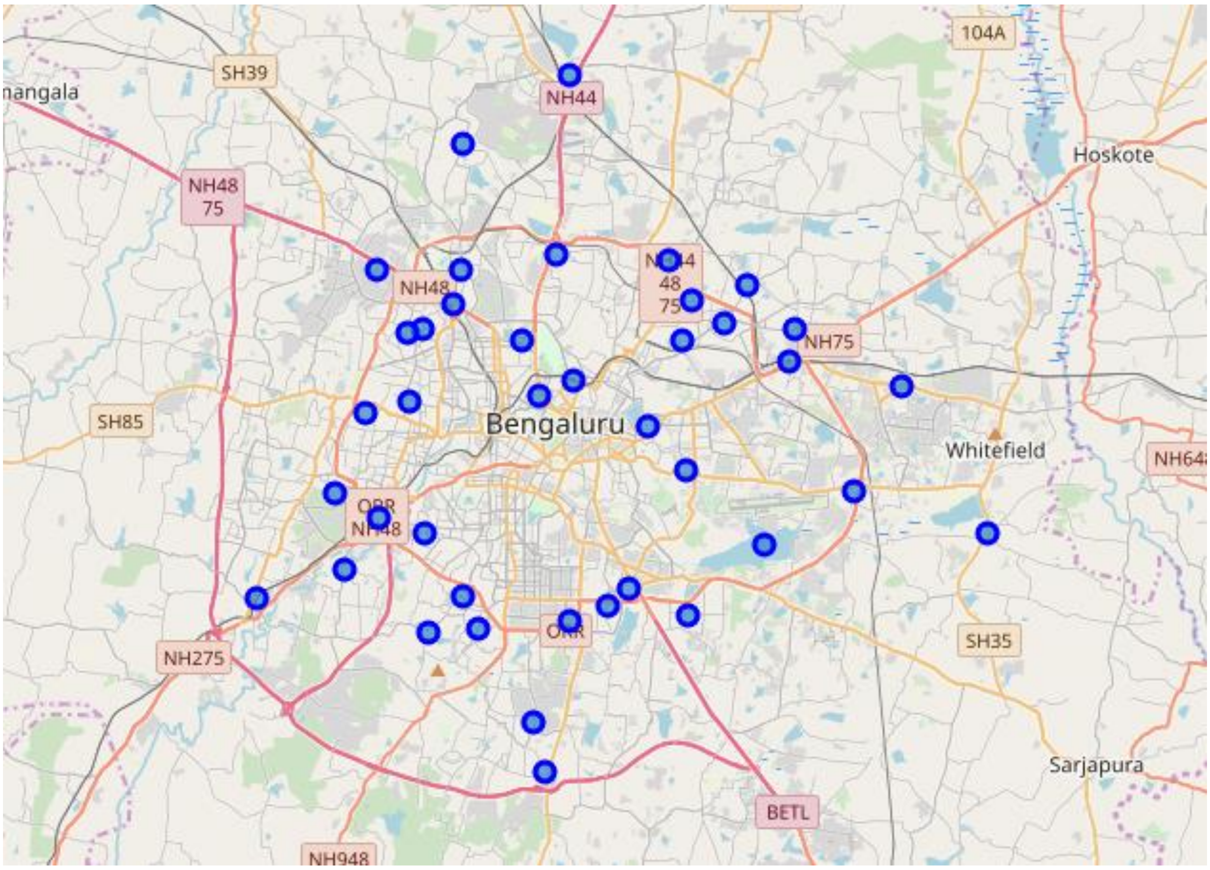
We use the Foursquare API to access the venue data. Using the Foursquare ID and client secret, I create a URL to access the desired data. We got Venue name and the category of each venue in the neighbourhood.

Now that we are concerned only with the shopping mall data, we did one hot encoding and found the mean for each of the individual neighbourhoods, which of the frequency of occurrence.

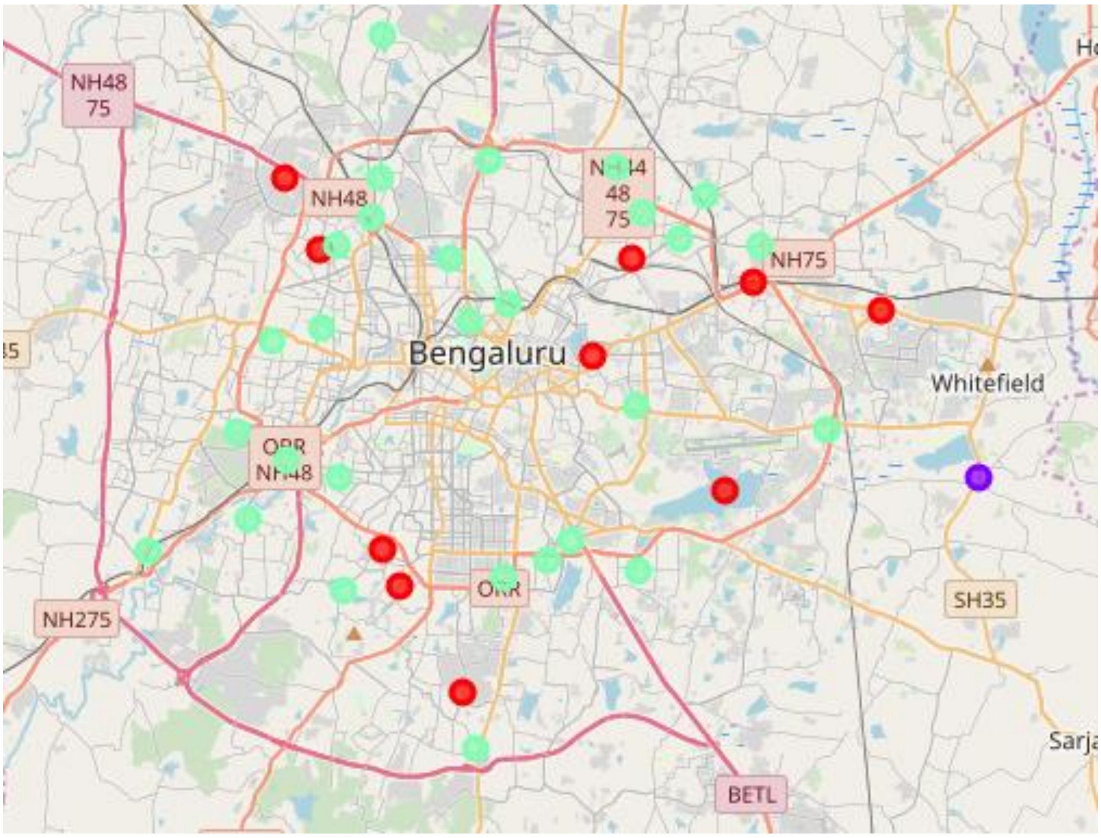
This data was then ready to act as an input to our Machine Learning algorithm, K-Means clustering. The number of clusters was decided to be 3, which represents the higher concentration of malls, fewer and the least concentration of malls.

We then plot these clusters on the map, with different colours. Which helps the stakeholder, in visualizing the neighbourhood with the highest concentration, fewer and the least concentration of malls. This will enable him to take to better decision. He would consider the neighbourhood will relative low concentration of malls to built mall to low competition in the market.

4. Results



This is the result of the visualization of the geographical coordinates of neighbourhoods in the city of Bangalore.



K-Means clustering algorithm categorized the neighbourhoods into 3 categories based on the frequency of occurrence:

- Cluster 0: with high concentration of malls in the neighbourhood
- Cluster 1: with moderate concentration
- Cluster 2: with the lowest concentration.

Cluster 2 represents in the map in mint green colour, Cluster 1 in blue, Cluster 0 in red.

5. Discussion

The maps displayed in the results section show the geographical locations of neighbourhood with high, moderate and low concentration of malls.

The red points are neighbourhoods with high concentration of malls, blue is with moderate concentration and the mint green colour points are with the low concentration of malls.

A property developer looking at the result would like to consider the places with low concentration of malls as these places will have low competition.

The red points with high concentration of malls would be already have high competition. On the other hand, if the property developer with a unique selling point can consider the neighbourhood in mint green colour.

6. Conclusion

This project deals with several skills like web scraping, using an API, data cleaning and wrangling, data visualization and machine learning.

In an attempt to answer the business problem statement, the project does give insights to the stakeholders about the different types of neighbourhoods he/she can consider to open a new mall, based on the number of existing malls in those neighbourhoods.

Post clustering the neighbourhoods into 3 categories, the neighbourhoods are plotted on the map of Bangalore, which gives the stakeholder a visual insight about the geographical locations.