Final project report

Data Analysis of office location preferences at M.N. Dastur &Co.

Table of contents

Introduction:

Background to the business model problem choice:

I am a metallurgical engineer at IIT Bombay and having done my internship at a metallurgical consultancy firm named M.N. Dastur I had the opportunity of spending a large amount of time at the organization with its offices ranging throughout the world, mainly concentrated in India.

Being a consultancy company based on the eastern coalfield belt of India they have taken an approach of creating a local community of companies together coming up with the solutions of certain complex metallurgical problems. The main approach is creating independent spaces for the companies to work in unison alongside each other.

On this basis I've decided to focus my project on M.N. Dastur office locations in Kolkata from where the company is headquartered. And thus, on this basis a locational data analysis of the offices for future expansion of the organization has been studied through this project.

Business Problem:

As a reputed consultancy firm in the field of steel market, the company established din 1965 has worked extensively with Indian as well as with many foreign steel plants They are now diversifying their field of interest and thus moving into the oil and gas and synthetic product market and are focusing on dedicating teams with exclusive offices for venturing into the diverse fields.

The question we aim, to answer through this project:

On the assumption that the current 6 offices operating in Kolkata with one being the headquarters are well established and successful with over half a decade of working experience, what factors are common between them and thus linking the offices situated in the city and could thus be informative for setting up of future offices in the city. That is what is a particular type of neighborhood that is suited for the employees of the organization in this region of the country that should be focused on further expansion.

Interest:

The target audience for this project is the board of directors of M.N. Dastur company who are planning on the expansion of the organization in the near future. Using this data, the organization could have a way forward in exploration of future locations for their offices based on the previous successes in the certain neighborhoods in this regional space.

Data Exploration:

The data of the current offices of the M.N. Dastur organization is taken and explored for deriving certain locational features for future expansion

The data will be explored by a clustering of neighborhood characteristics of the current locations of offices of M.N. Dastur In Kolkata by the k means clustering algorithm.

The present Dataset:

- Where are the current offices of the organization spread at a global level, inside India and lastly within the city of Kolkata. We take the data from their current website: http://www.dastur.com/content-59-Dastur-Offices and then perform the following operations:
  - Scrape the city names from the webpage mentioned above using Beautiful Soup
  - Find the latitude and longitude of all the locations using Goopy.
  - Plot the map of the scrapped locations using Folium.

K-Means Clustering:

The *k*-means algorithm searches for a pre-determined number of clusters within an unlabelled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:

- The "cluster centre" is the arithmetic mean of all the points belonging to the cluster.
- Each point is closer to its own cluster centre than to other cluster centres.

Those two assumptions are the basis of the *k*-means model. We will soon dive into exactly *how* the algorithm reaches this solution, but for now let's take a look at a simple dataset and see the *k*-means result.

Foursquare data for each office:

The final data set to be gathered is using Foursquare to provide data for the immediate vicinity for each of the 6 offices of interest in Kolkata.

In this section we find up 100 venues in the 400m closest to the offices listed in Kolkata. The final output is a Data Frame containing these details which can then be used for clustering in the next step of data analysis (out of scope for this week's assignment.

The steps followed are the following:

- Using the latitude and longitude found in the previous section, query Foursquare using the explore API to find up to 100 local popular venues.
- Group the data produced by office name and shape using one hot encoding.
- Find the 10 most common venues for each office and create a final DataFrame with this data for analysis.

Methodology :

For this project we will use the following methodology:

- Data has been collected from the M.N.Dastur website for Kolkata based offices and exact locations added
- The areas around each office will be characterised using data from Foursquare for venue types in a 400m radius around each. This radius was chosen as the “ ped shed " which is deemed as equivalent to 5 mins walk.
- The data will be explored using heatmapping and K-means clustering in order to find similarities between the localities of the various M.N.Dastur offices in Kolkata.

After performing of the heat map, it can be seen that there are hotspots in the southern area of Kolkata namely the areas near Chandni chowk to Rabindra sadan within which we can see the technological hub area of park street and esplanade all the way up to tollygunge.

Clustering :

After clustering done on the basis of k-means algorithm we can see there is majority of cluster 1(label 0) within which majority of the current office location falls

Results:

In this analysis we have explored the locations of M.N.Dastur offices in Kolkata. Using clustering based on venues within walking distance (defined as 400m radius) we have clustered the offices into three groups. Of these groups, only one group has more than one member.

Comparing the three largest clusters (clusters 1,2 and 3 in order) we note the following:

- The largest cluster share a number of features. ATM, Asian restaurants ,BBQ joints, beer garden comes under these popular features.
- Being specific about the type of each area beyond these shared features is difficult but we note a few differences using the top ten venue types for each cluster:
    - Cluster 1 (label 0) has the "core" venue types one might except, but has less variety in other areas i.e., types of variety in restaurants and movie theatres and lounges i.e., it lacks some form of entertainment.

Discussions:

Based on our results we have identified the central- southern part of Kolkata as areas where M.N.Dastur offices appear to have been successful. As noted at the beginning of this project, a key assumption here is that the current M.N.Dastur offices are "successful". This assumption would need to be validated with data from M.N.Dastur highlighting offices which, by their own metric, are defined as successful. Potential metrics to use here could be profitability or high rates of incoming project by clients. Or potentially even other metrics measuring the outcomes desired from such working spaces, potentially harder to quantify, such as small independent organisation success rates, innovations or partnering successes.

In theory, using these results, M.N.Dastur could characterise other areas even in the other metropolitan cities where they already have a base where they might be considering new offices to see if they are similar to those seen here, influencing their decision to create a new office there or not.

One thing noted during the exploration of the data and in particular a heat map of the offices was the higher incidence of offices around key commuter links into and out of the capital. In further work it would make sense to quantify this and add into the clustering methodology.

Conclusion:

We have been able to show that the current offices fall into three main types of neighbourhood based on the types of venues in those neighbourhoods. This analysis might allow the key stakeholder here, of M.N.Dastur themselves, to rate potential locations for future offices.

Key things to explore further would be how to define / subset the more successful offices to better inform the clustering process. Also, to explore the distances from commuter train stations / key roads, distance from the manufacturing and marketing bases of the popular clients in the field of oil and gas where they are aiming to penetrate into which will most likely have impact here also.