Report

on

Exploratory Analysis of Geolocational Data

By

Kunal Atram

(205119049)

MCA Third Year

2021-2022

Department of Computer Applications

National Institute of Technology,

Tiruchirappalli

Introduction:

This project involves the use of K-Means Clustering to find the best accommodation for students in Pune (or any other city of our choice) by classifying accommodation for incoming students on the basis of their preferences on amenities, budget and proximity to the location.

In the fast-moving, effort-intense environment that the average person inhabits, It's a frequent occurrence that one is too tired to fix oneself a home-cooked meal. And of course, even if one gets home-cooked meals every day, it is not unusual to want to go out for a good meal every once in a while for social/recreational purposes. Either way, it's a commonly understood idea that regardless of where one lives, the food one eats is an important aspect of the lifestyle one leads.

Now, imagine a scenario where a person has newly moved into a new location. They already have certain preferences, certain tastes. It would save both the student and the food providers a lot of hassle if the student lived close to their preferred outlets. Convenience means better sales, and saved time for the customer.

Food delivery apps aside, managers of restaurant chains and hotels can also leverage this information. For example, if a manager of a restaurant already knows the demographic of his current customers, they'd ideally want to open at a location where this demographic is at its highest concentration, ensuring short commute times to the location and more customers served. If potential hotel locations are being evaluated, a site that caters to a wide variety of tastes would be ideal, since one would want every guest to have something to their liking.

Project Stages:

- Fetch Datasets (Data Collection)
- Clean the Datasets to prepare them for analysis. (Data Cleaning via Pandas)
- Visualise the data using boxplots. (Using Matplotlib /Seaborn /Pandas)
- Fetch Geolocational Data from the Foursquare API. (REST APIs)
- Use K-Means Clustering to cluster the locations (Using ScikitLearn)
- Present findings on a map. (Using Folium/Seaborn)

Fetch data Data Exploration and Visualisation:

First load the dataset and then there are around 70 parameters. Not all of them are relevant. By extracting the most relevant features into a pandas dataframe.Then we have to clean the data. (process of Extracting the features, and dealing with different kinds of values as well as NaN values)

After fetching the datset we have to explore and visualize the data. A good way to do this is by visualising the data via graphs. Graphs help us quickly get a sense of the data, and are a much more user-friendly way of understanding data as compared to reading thousands of rows of data. A good graph to look at distributed groups is a boxplot. It can tell us at glance where the population is concentrated, and how the outliers compare to the average object in the group. plot a boxplot on the dataframe. We should be looking at things like how much each person exercises on average (these people will need gyms), whether they are vegetarian

/ non vegetarian (we should account for both), and income (this affects the lifestyle of the person significantly.)

Run KMeans Clustering on the data:

K Means Clustering will help us group locations based on the amenities located around them. For example, a location with a high amount of shops nearby will be labeled "Amenity Rich" while a location with less amenities will be labeled "Amenity Poor". Similar locations will be grouped (clustered) together. (Check the references for a more formal explanation!). On the data we have to run the KMeans Clustering Algorithm and figure out the best value for K, which we will use later.

Get Geolocational Data from Foursquare API:

Now that we know the best K value for our population, we need to get geolocational data from the Foursquare API to find people some accommodation. By creating a free foursquare account we get our API credentials and then we have to set up. Set up our query in such a way that we can check for residential locations in a fixed radius around a point of our choosing. For example, we can pick (18.5204, 73.8567) these are the coordinates of Pune City. After Hitting the endpoint, we have to parse the response data into a usable dataframe. There will be a lot of information that we don't need, so we have to apply the same data cleaning principles that we have used for previous dataset to get a workable dataframe.

Plot the clustered locations on a map:

Now it's time to run K Means clustering on the data and plot the results on a map. For that we have to run K Means clustering on the dataset we have prepared previously with the optimal K value we found. Now after having the results, it's time to visualise them. Using Folium, we have to plot our results on a map of the world, centered on the location we have selected.