Applied Data Science Capstone

Battle of Neighborhoods : Bengaluru, India

Geospatial Agglomeration with Location Data

ARUN P. R.

June 27, 2020

Contents

Chapter 1

Introduction

This project aims to segregrate the neighborhoods in Bengaluru, India based on their similarities/differences and to make use of this data to interpret the concentration of population in these neighborhoods.

1.1 Background

The distribution of population in neighborhoods has always been a topic of interest to various government/private agencies. The census data is normally used by these agencies. However a comparison between similar neighborhoods is lacking in such studies. This project aims to fill this gap by providing useful insights in to how the neighborhoods are segregrated so that the planning for the above cited activities can be done more effectively. As a representative entity, Bengaluru, India is chosen as the target location.

1.2 Business Problem

To segregrate neighborhoods in Bengaluru, India based on location data obtained using Foursquare API and to use this in conjunction with geospatial and population statistics to arrive at useful insights that can aid different agencices to implement their schemes more effectively.

1.3 Description of the problem

The problem consists of following subproblems:

1. Get Data Source for Population/Equivalent for all neighborhoods in Bengaluru, India
2. Get geojson corresponding to neighborhoods in Bengaluru, India
3. Get lattitude, longitude information for all neighborhoods using geocoder
4. Get location data corresponding to all neighborhoods using Foursquare API
5. Clean all data, explore them, extract features
6. Segregate neighborhoods using location data
7. Analyse segregated neighborhoods data in conjuction with population data and interpret the results

1.4 Benefits to Stakeholders

The project will be beneficial to various government/private agencies involved in demographic studies, town planning, resource allocation, planning of development projects, etc. by providing useful insights in to how the neighborhoods are segregrated so that the planning for the above cited activities can be done more effectively.

Chapter 2

Data Acquisiton and Cleaning

Details about proposed data sources and how data is proposed to be extracted from them are detailed in this chapter.

2.1 Data Sources

The following are the data sources used in this project:

1. Data Source for Population/Equivalent for all neighborhoods in Bengaluru, India
   - KARNATAKA STATE ELECTION COMMISSION, Ward Wise Voters Data
   - This data in tabular format is available as a pdf file and contains the total number of voters in each neighborhood and is representative of the population.
2. Get geojson corresponding to neighborhoods in Bengaluru, India
   - BBMP.GeoJSON
   - This dataset is shared under Creative Commons Attribution-ShareAlike 2.5 India license
3. Get lattitude, longitude information for all neighborhoods using geocoder
   - Given that this package can be very unreliable, in case it is not possible to get the geographical coordinates of the neighborhoods using the Geocoder package, this step will have to be performed manually.
4. Get location data corresponding to all neighborhoods using Foursquare API and get venue categories for each neighborhood.
   - With neighborhood names and latitude-longitude information, Foursquare API is used to get location data consisting of upto 100 venues within a 2km radius from given geospatial coordinates.

2.2 Data Cleaning

Corresponding to each dataset, specific cleaning operations will have to be carried out to make it usable to solve the problem.

For example, WARD WISE VOTERS ABSTRACT contains information useful to this project in columns WARD_NAME and TOTAL whereas all other columns are to be dropped. WARD_NAME is of the format WARD_NO followed by WARD_NAME with a space. WARD_NO needs to be removed from this.

2.3 Feature Selection

Feature selection to be carried out before segregating neighborhoods so that we get a good segregation. The features selected in this case are venue categories derived from Foursqare API.

2.4 Approaches in using the data to solve the problem

The neighborhood data is to be segregated using the feature set venue categories provided by Foursquare API. A clustering algorithm needs to be chosen for this and k-Means Clustering Algorithm is used in this project. The optimum number of clusters are to be determined and k-Means clustering is carried out. The chloropleth map of Bengaluru with voters data is prepared and the neighborhoods segregated by clusters are overplotted on this. This data is to be analysed to arrive at insights in to the relation between the clusters and population.