DAV Project(CS306)

Assigned By: Prof. Pankaj Kumar

Khanti Rindani(201601412) Samhitha Gundam(201601065)

1 Abstract

The report aims to classify and visualize customer span of Yes bank. The database taken for analysis is a random sample of 1 lakh entries from the total of 10, 594, 391 entries. The fields contained in database are customer_id, current and permanent addresses divided into address line, city and state resulting into total of 9 fields.

2 Data Pre-processing

Data Cleaning:

- Checked for unique customer_id. If the ids are not unique, we need to drop all the duplicate ids.
- Eliminated the rows with dummy entries such as add3, addcity, addstate, dummy or only characters like '.' or'..' or '...'
- Also eliminated the rows where the latitude and longitude of a location cannot be found (Due to non-meaningful addresses or unavailability of both pin-code and address)

Data Augmentation:

- We have created 2 new columns for storing the combined address i.e, address, city and state(both current and permanent)
- Also we have augmented 4 new columns for storing the co-ordinates(latitudes longitude) of the current and permanent address location.

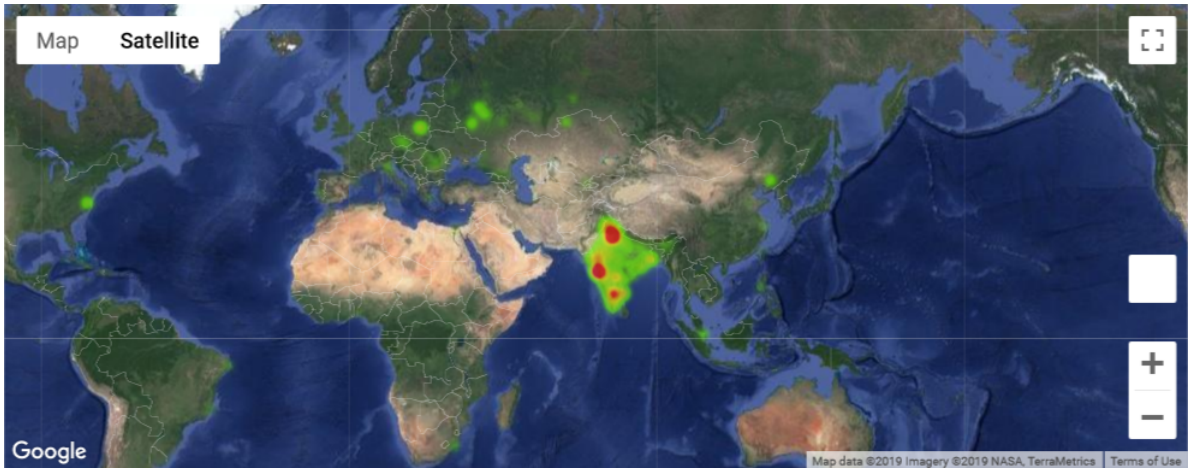

Figure 1: Heat map visualization of current addresses of customers

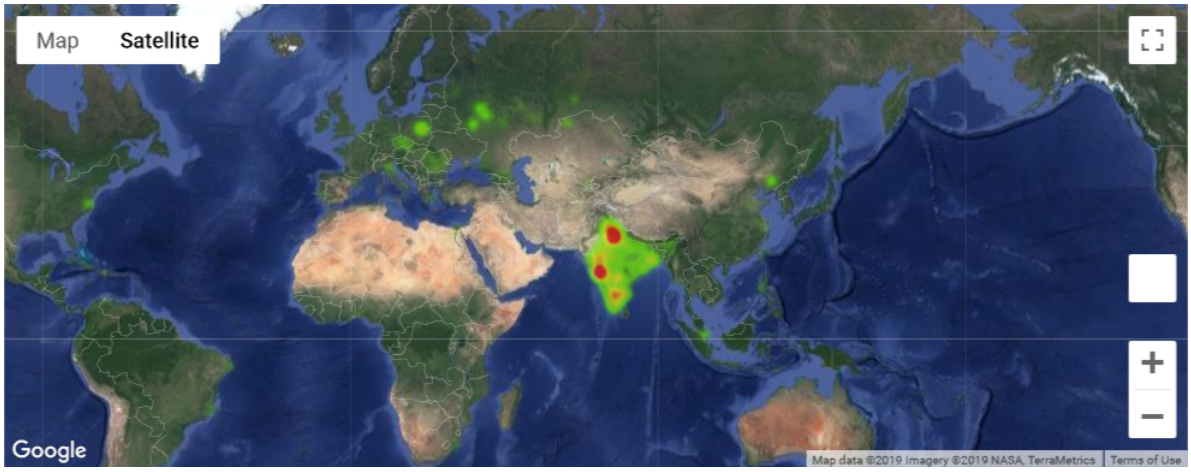

Figure 2: Heat map visualization of permanent addresses of customers

3 Data Visualization

3.1 Problem-1

Problem Statement
Visualize the permanent and current address of customers a heat map on a global world map.

Method:

After completing the data cleaning and data augmentation, we used geopy library. From geopy.geolocators we have used Nominatim to find the coordinates of the location using either pincode or the new column of concatenated address of the customer. On not being able to find any location, the respective row is dropped treated as junk value.

After obtaining the latitudes and longitudes of all the locations, we have plotted the heat map using heapmap_layer feature of the gmaps library for python.

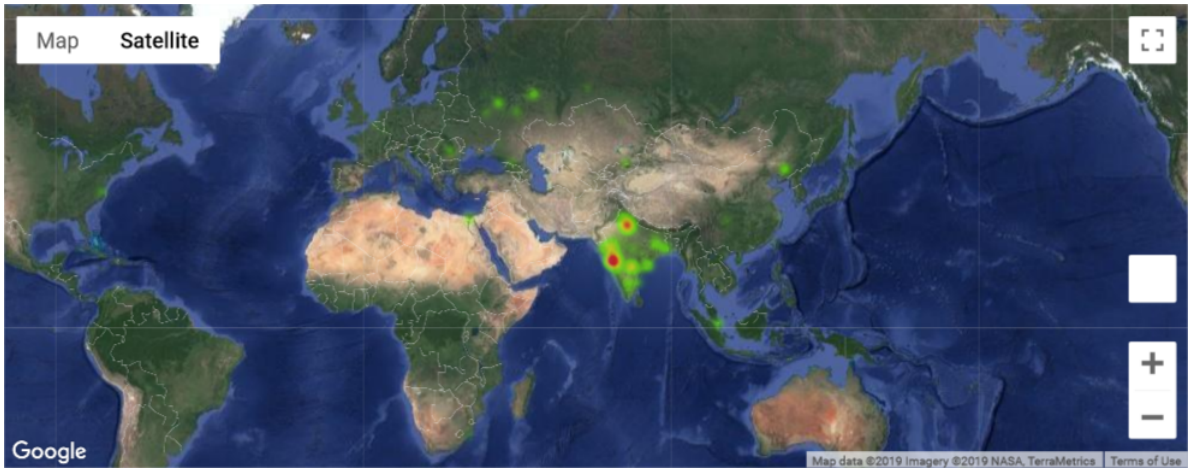

Figure 3: Heat map visualization of current addresses of household customers

3.2 Problem-2

Problem Statement
Classify the customers in the data set as business and household and visualize them accordingly.

Method

After the overall cleaning of data, we also removed the records with empty address fields or having non-meaningful data such as 0,A,C,E,etc. We carried out the task of classification by using the keywords with high frequency obtained from mapper_reducer (word-count) program.

We first removed the ambiguous address which neither contributes to household nor to business. These includes the addresses which have only the city/state name.

We identified household customers with the keywords: SOC, SOCIETY, COLONY, TOWER, BUNGLOW

The identification of business customers was difficult due to a company being landmark for normal address. We searched for keywords: PRIVATE LIMITED, LTD., SHOP, SCHOOL, COLLEGE, TOWER, COMPANY, SHOP, OFFICE, BUSINESS, MARKET, CLINIC, BANK, STATION, STN, POLICESTATION, CORPORATION, MUNICIPALITY, INSTITUTE, STORE, FACTORY, HOTEL, RESTAURANT, RETAIL, MERCHANT, SONS, COOPERATIVE, INDUSTRIAL, MILL, DAIRY with additional constraint of not being as landmark using keywords: NEAR, NR, NR., BESIDE, B/H, BEHIND, OPP, OPP., OPPOSITE, NEXT TO, BEFORE

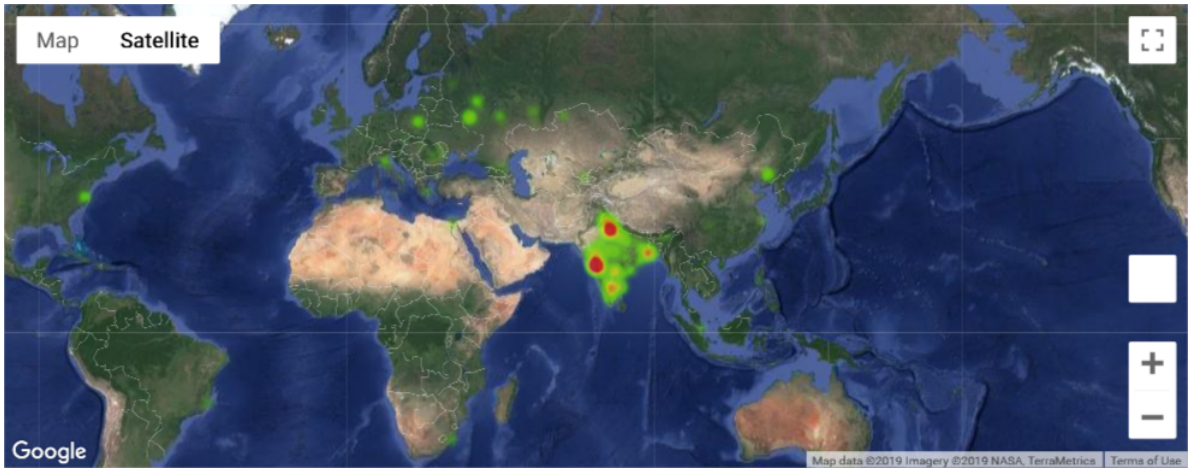

Figure 4: Heat map visualization of current addresses of business customers

4 Conclusion

- The sampled database mostly contains the data about Indian customers along with minority of European customers.
- The visualization reflects that majority of customers are located around the metropolitan cities of India such as Mumbai, Delhi and Banglore.
- The sampled database shows that there are significantly smaller number of household customers than business customers.
- The analysis and visualization carried out helps company (Yes Bank) in launching of new policies and new branches according to the targeted customers.