CS306 Project

Data analysis and visualization for Yes Bank customers

—

Samhitha Gundam -201601065
Khanti Rindani -201601412

Introduction

- The database delivers the information about customers of Yes Bank world-wide providing their addresses.
- Database used for analysis consists of randomly chosen 100,000 records from total of 10, 594, 391 records.
- For reducing the sample size, uniform random numbers are used in order to avoid biasing.
- 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.

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 or non-meaningful entries.
- 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).

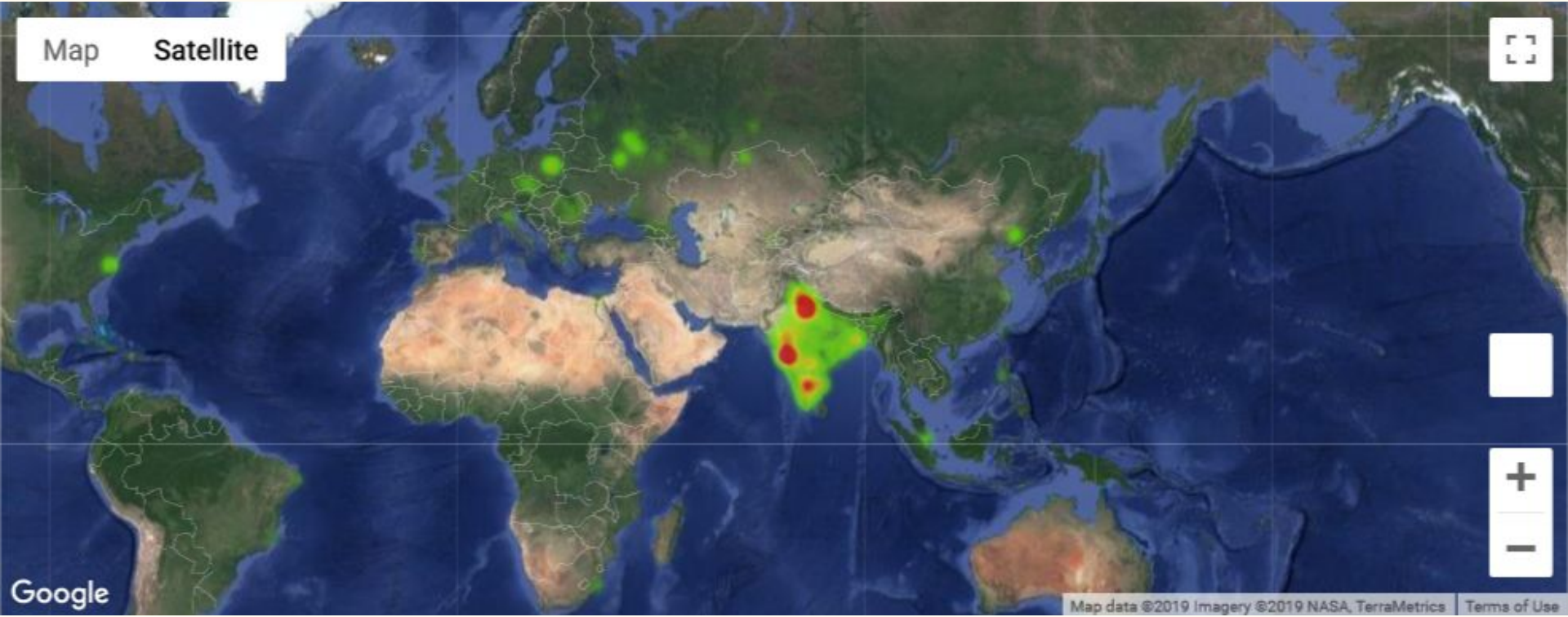
Data Augmentation

- Created two new columns for storing the combined address i.e, address, city and state(both current and permanent)
- Also we have augmented four new columns for storing the coordinates(latitudes longitude) of the current and permanent address location.
- Checked for multiple countries having same pin-code. If found, we need to augment a new column consisting of the country name for each record.

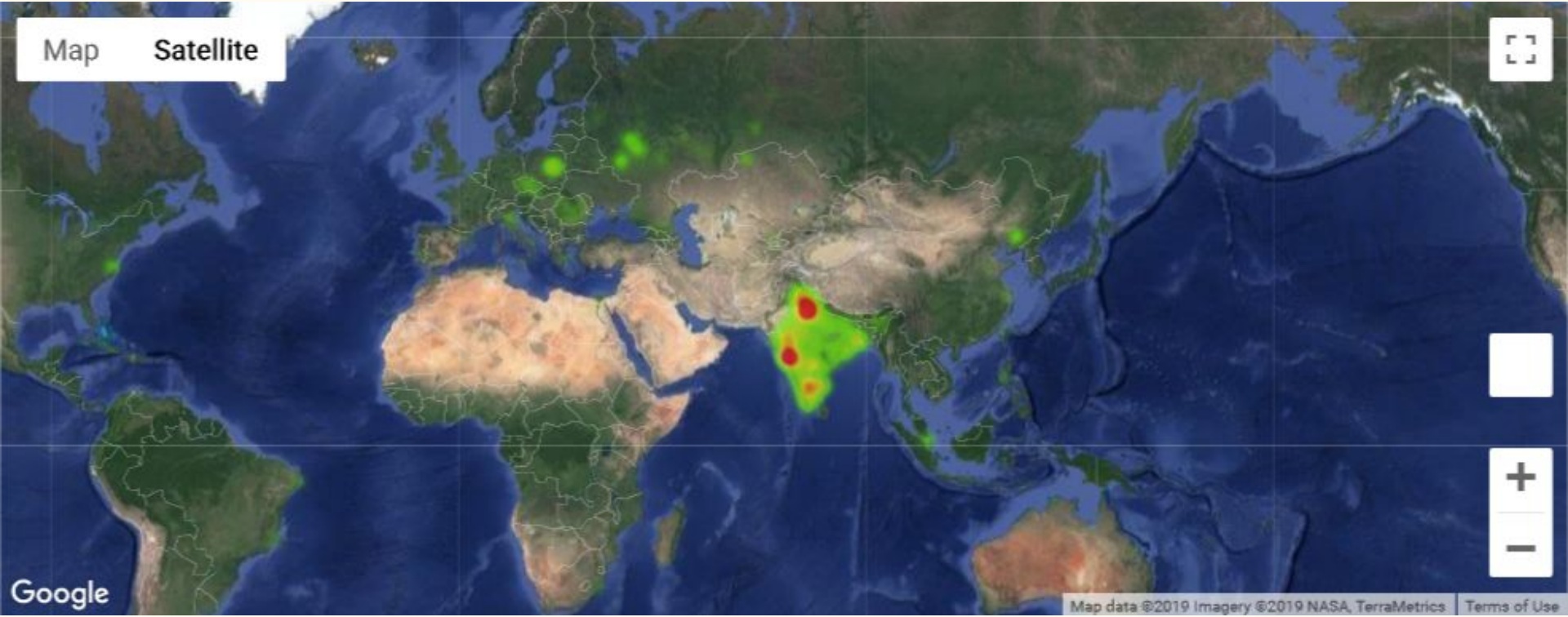
Visualizing the permanent and current address of customers a heat map on a global world map:

- To find precise latitude and longitude, we have used entire address rather than just city name and pincode also wherever the address contained null or non-meaningful entries.
- From geopy.geolocators library, 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 heatmap layer feature of the gmaps library for python.

Heatmap for Current addresses



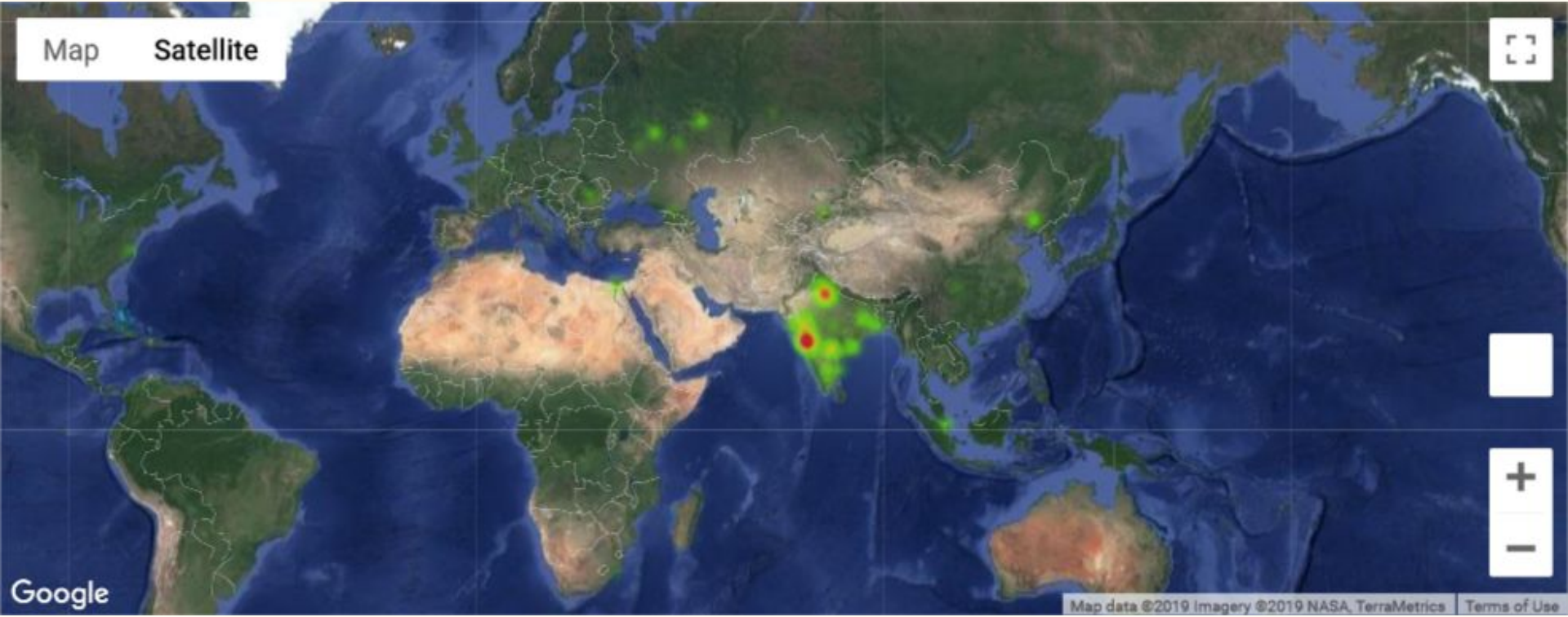
Heatmap for Permanent addresses



Classifying the customers in the data set as business and household and visualize them accordingly.

- After the overall cleaning of data, we also removed the records with empty address fields or having non-meaningful data.
- 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 and business customers with their appropriate keywords.
- The identification of business customers was difficult due to a company being landmark for normal address.

Heatmap for household customers



Heatmap for business customers

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 Bangalore.
- 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.

Thank You