Concept Note

Data Cleaning:

Pipe delimited file would be imported using Pandas. NULLS would be handled as per instructions. (More info in the Queries Section) To identify if data is missing we would use the structure of the input as mentioned in the dataset. For eg: Pin Code can be an alpha-numeric string of any length.

Algorithm Description:

We need to identify countries, given the city, state, and pin code. We then would cluster the data into spatial groups to reduce the time complexity of the data provided to plot the city heatmap faster. Grouping can be done at multiple levels (City, State or Pincode). Address types of permanent and current addresses can be used to classify and plot a heat map of the world. We use the unique customer IDs to mark different customers if required.

Latitude and Longitudes can be extracted using the Google Maps Geocoding API given the address details which we can extract using the dataset provided. We will generate the heat map using an open source plugin called gmaps which is available on Github. The project would be implemented on an iPython Notebook to show the code and the output together.

Challenges:

1. Improving rendering time of the heatmap at city level where there would be many different addresses that are close together.

Queries:

1. Do households and offices need to be distinguished from each other?
2. Should we use the street address or the pin code(or both) for generating latitudes and longitudes?
3. Can we use the pin code to group together two or more addresses for the city level heatmap?
4. If the current address is missing, do we ignore it or consider the permanent address?