Chagas_Health_Centers

Mughundhan Chandrasekar

7/07/2017

INTRODUCTION

- People often report to local “health posts”, that their house is infested. We can install n (5-10) stations for the people to report.
- The dataset holds information pertaining to the houses like the geospatial coordinates, predicted probabiity of being infested etc.
- **AIM**: Build R shiny app, that will accept the number of stations as input (depends upon the budget, entered by the user) and renders a map with spatial distribution of the health posts at optimal locations.
- In this report, I have clustered the regions and idnetifed the optimal locations for installing the health facility using **K-Means Clustering technique**.

NOTE: The data munging and feature engineering operations (along with sample visualizations) involved 500+ lines of code in R. As this project is worked for the client based out in Arequipa (Peru), the code is hidden intentionally and only a sample of the dataset is used here to make sure that the anonymity is preserved.

1. Creating an Environment

- Involves loading the appropriate libraries
- Load the dataset into the working environment

```
rm(list=ls())
library(lubridate) # for csv files
library(leaflet)   # maps
library(dplyr)     # for piping purpose %>%
library(sp)
library(rgdal)
library(geosphere)
library(dismo)
library(rgeos)
library(fields)

#library(lpSolve)   # fir linear programming in R
setwd("/Users/Mughundhan/DataScienceIntern/Chagas")
fdata <- read.csv("fdata.csv")
```

2. Analyzing the Dataset

- Let us have a deeper look at the dataset in-order to gain more insights.

```
## 'data.frame':    642 obs. of  20 variables:
##  $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ V1                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ UNICODE              : Factor w/ 598 levels "1.10.38.100",..: 1 2 3 4 5 6 6
7 8 9 ...
##  $ USER_NAME            : Factor w/ 8 levels "CC","CCP_1V",..: 6 4 4 4 4 4 4 4
4 4 ...
##  $ GROUP_NAME           : Factor w/ 5 levels "MINISTERIO_DE_SALUD",..: 1 3 3 3
3 3 3 3 3 3 ...
##  $ DATA_ACTION          : Factor w/ 2 levels "INSPECTION_NEW",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ CARACT_PREDIO        : Factor w/ 3 levels "DES","LP","casa_regular": 1 3 3 3
3 3 3 3 3 3 ...
##  $ STATUS_INSPECCION    : Factor w/ 5 levels "C","R","V","entrevista",..: 1 4 1
1 4 5 3 4 5 5 ...
##  $ TEST_DATA            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DATETIME             : Factor w/ 548 levels "01/11/17 19:57",..: 183 437 436
435 434 441 433 432 431 430 ...
##  $ PREDICTED_PROBAB     : num  9.23e-05 5.84e-11 1.39e-10 2.41e-11 8.48e-10 ...
##  $ PREDICTED_PROBAB_MEAN: num  0.0142 0.0144 0.0144 0.0144 0.0144 ...
##  $ PREDICTED_COLOR      : Factor w/ 6 levels "#BD0026","#F03B20",..: 1 5 4 5 3
2 2 3 4 3 ...
##  $ LATITUDE             : num  -16.4 -16.4 -16.4 -16.4 -16.4 ...
##  $ LONGITUDE            : num  -71.5 -71.5 -71.5 -71.5 -71.5 ...
##  $ DATETIME1            : Factor w/ 548 levels "01/06/17 15:59",..: 31 384 383
382 381 396 380 379 378 377 ...
##  $ LOCAL_TIME           : Factor w/ 548 levels "0001-06-17 16:49:36",..: 31 384
383 382 381 396 380 379 378 377 ...
##  $ LOCAL_DATETIME_new   : Factor w/ 548 levels "01/17/11 08:47 PM",..: 483 377 3
76 375 374 381 373 372 371 370 ...
##  $ week                 : Factor w/ 5 levels "Friday","Monday",..: 1 1 1 1 1 2
1 1 1 1 ...
##  $ date                 : Factor w/ 44 levels "2017-01-11","2017-01-16",..: 38
29 29 29 29 30 29 29 29 29 ...
```

3. Data Munging

- Let us have a look at the missing values in each column
- Remove unnecessary columns
- Take a subset to work on that

```
##                     X                    V1               UNICODE
##                     0                     0                     0
##             USER_NAME            GROUP_NAME           DATA_ACTION
##                     0                     0                     0
##         CARACT_PREDIO     STATUS_INSPECCION             TEST_DATA
##                     0                     3                     0
##              DATETIME      PREDICTED_PROBAB PREDICTED_PROBAB_MEAN
##                     0                   101                   101
##        PREDICTED_COLOR              LATITUDE             LONGITUDE
##                   170                     0                     0
##              DATETIME1           LOCAL_TIME    LOCAL_DATETIME_new
##                     0                     0                     0
##                  week                  date
##                     0                     0
```

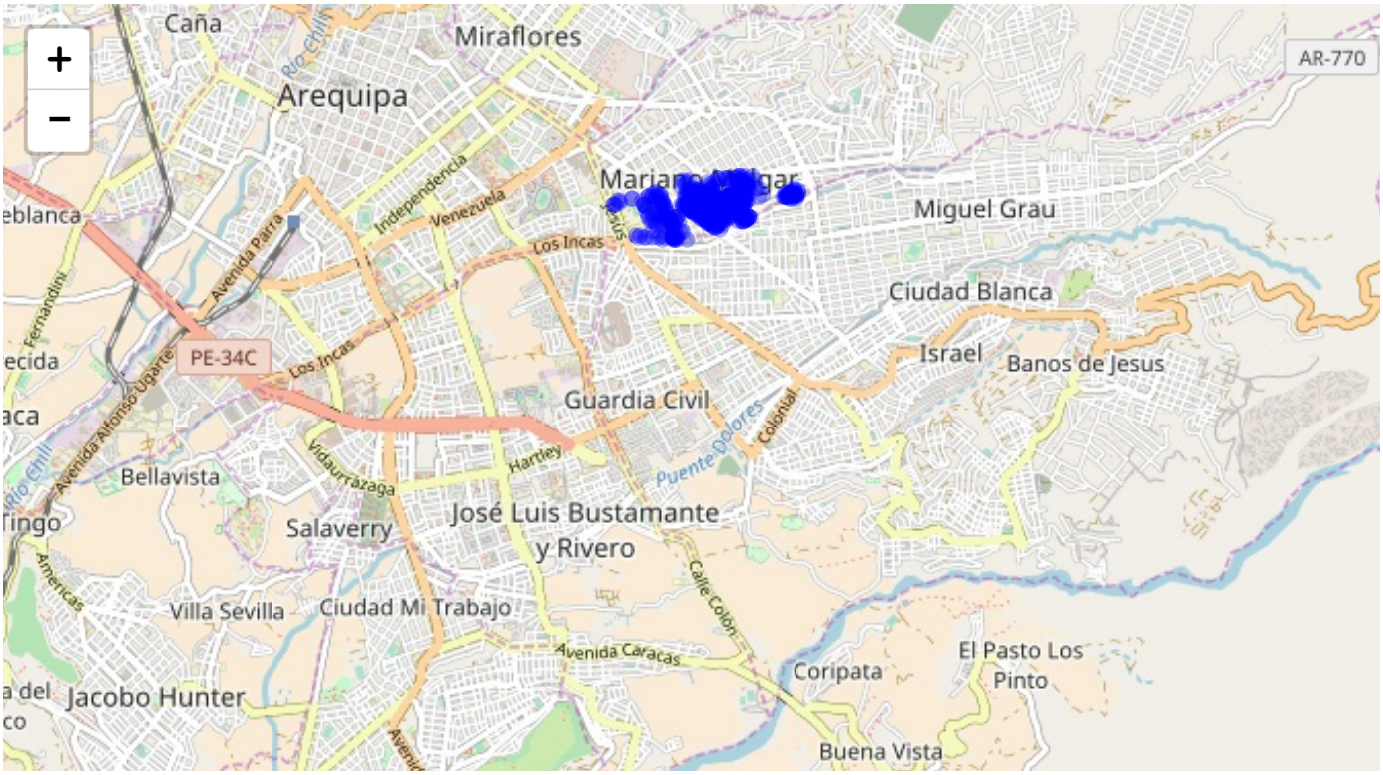
```
##
## FALSE  TRUE
## 12465   375
```

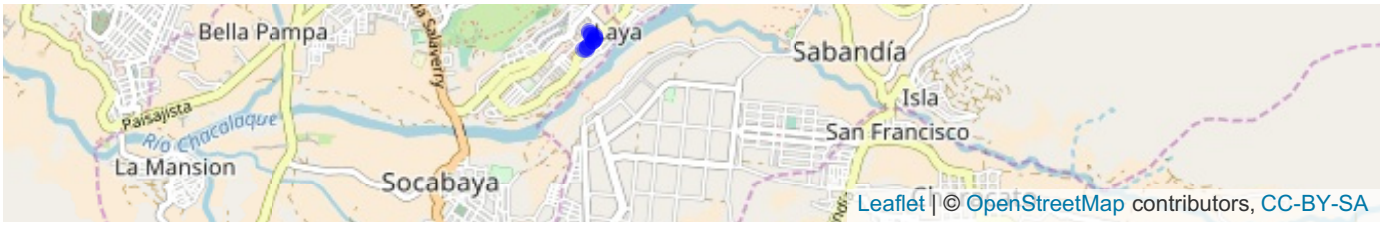
- We need only Geocoordinates (latitude, longitude) and the Unicode. Let us remove all other fields.

```
##   id  LATITUDE LONGITUDE
## 1  1 -16.41048 -71.50998
## 2  2 -16.40770 -71.50583
## 3  3 -16.40775 -71.50592
## 4  4 -16.40776 -71.50611
## 5  5 -16.40783 -71.50617
## 6  6 -16.40788 -71.50632
```

3. Data Visualization

- Let us plot the coordinates on google maps using leaflet and visualize their geospatial data spread.
- Interactive map: Try to zoom-in and zoom-out. Navigate within the allocated window.

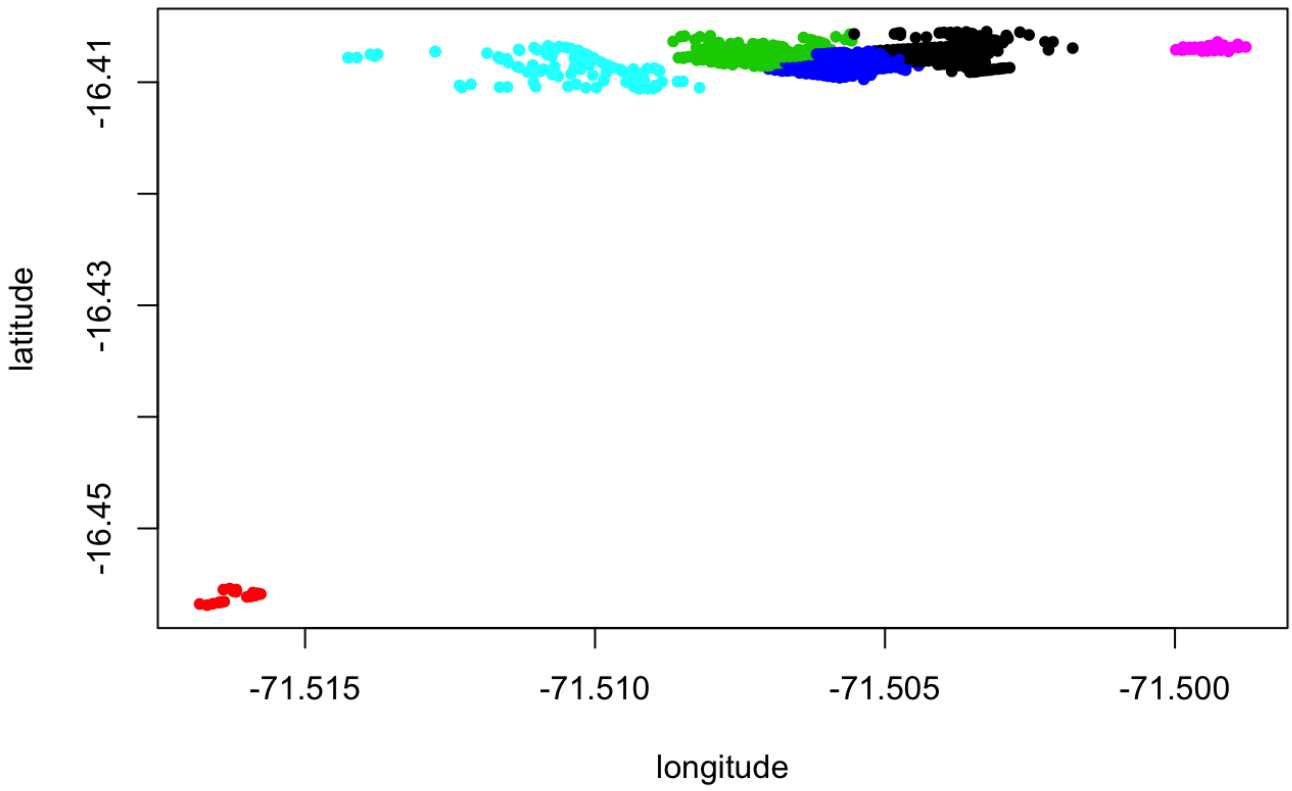




4. Data Clustering & Visualization

- Now let us set the number of clusters or the health facilities that needs to be installed.
- Split the data points based on their geo-coordinates and assign them to each cluster.

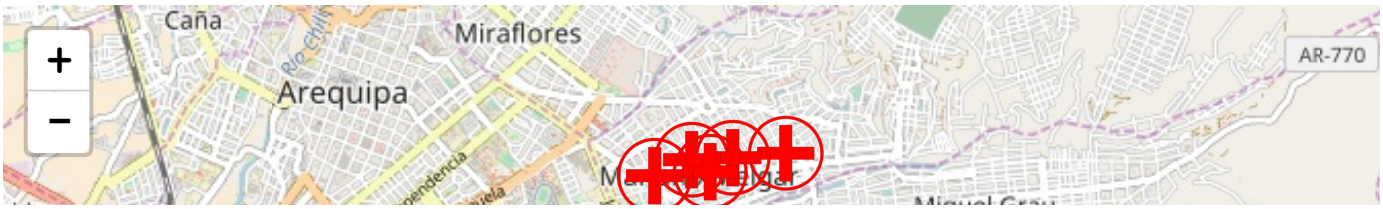
```
latitude<-fdata$LATITUDE
longitude<-fdata$LONGITUDE

km <- kmeans(cbind(latitude, longitude), centers = 6)
plot(longitude, latitude, col = km$cluster, pch = 20)
```

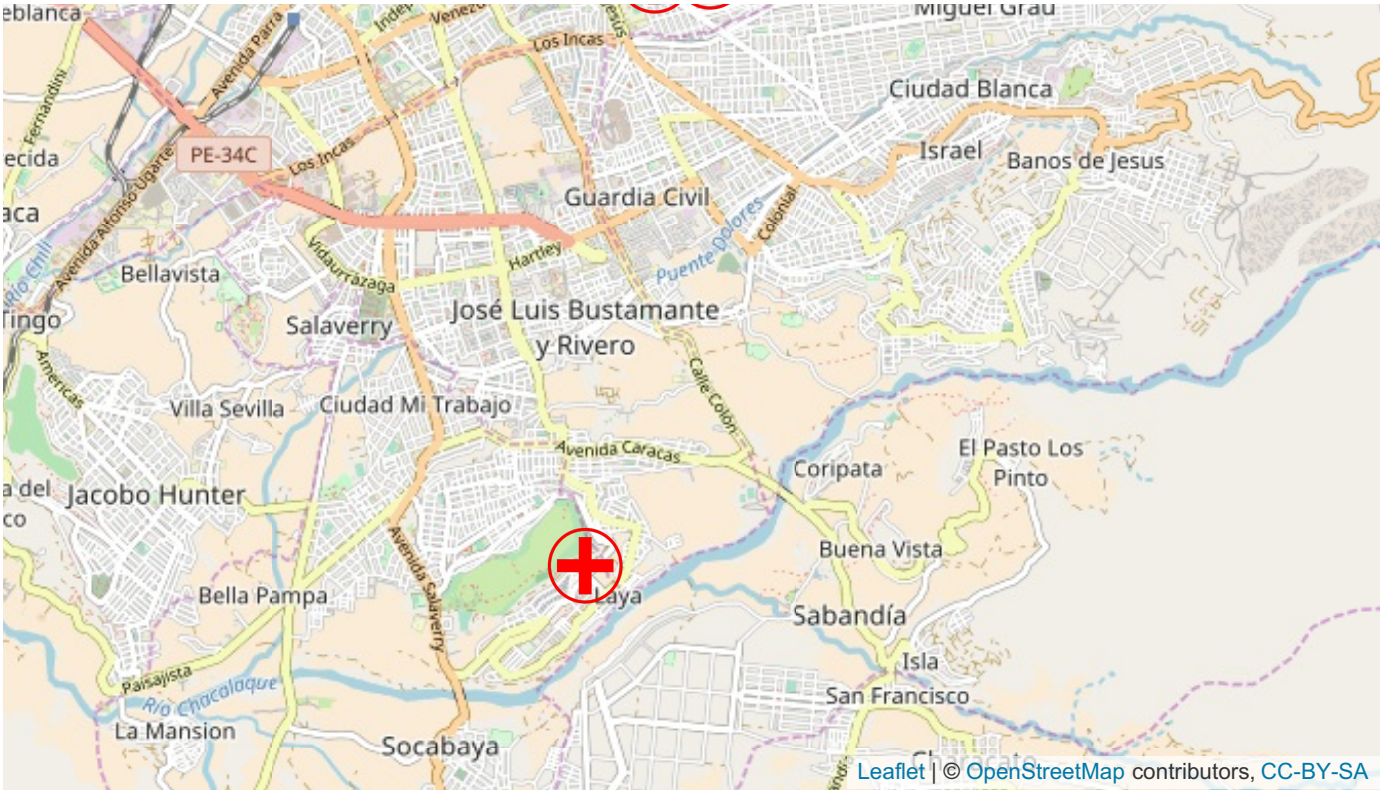


5. Data Centers

5.1. Identify the Health Centers

- Identify the optimal point in each clustered region to have a health facility



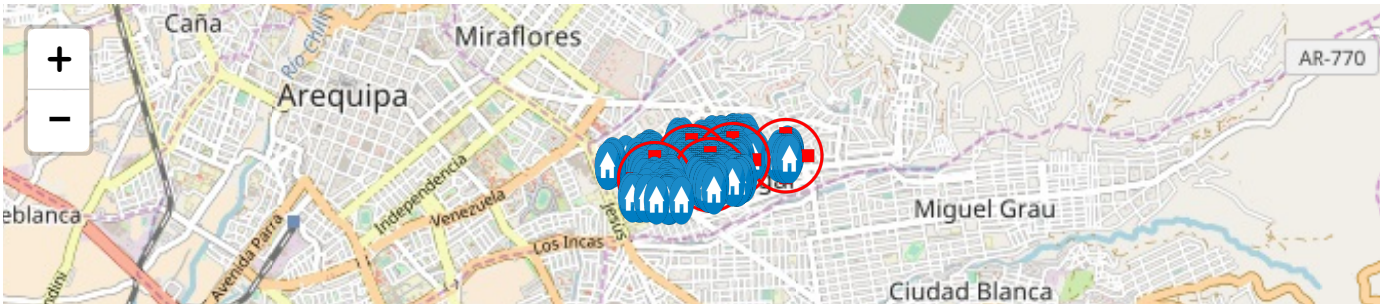


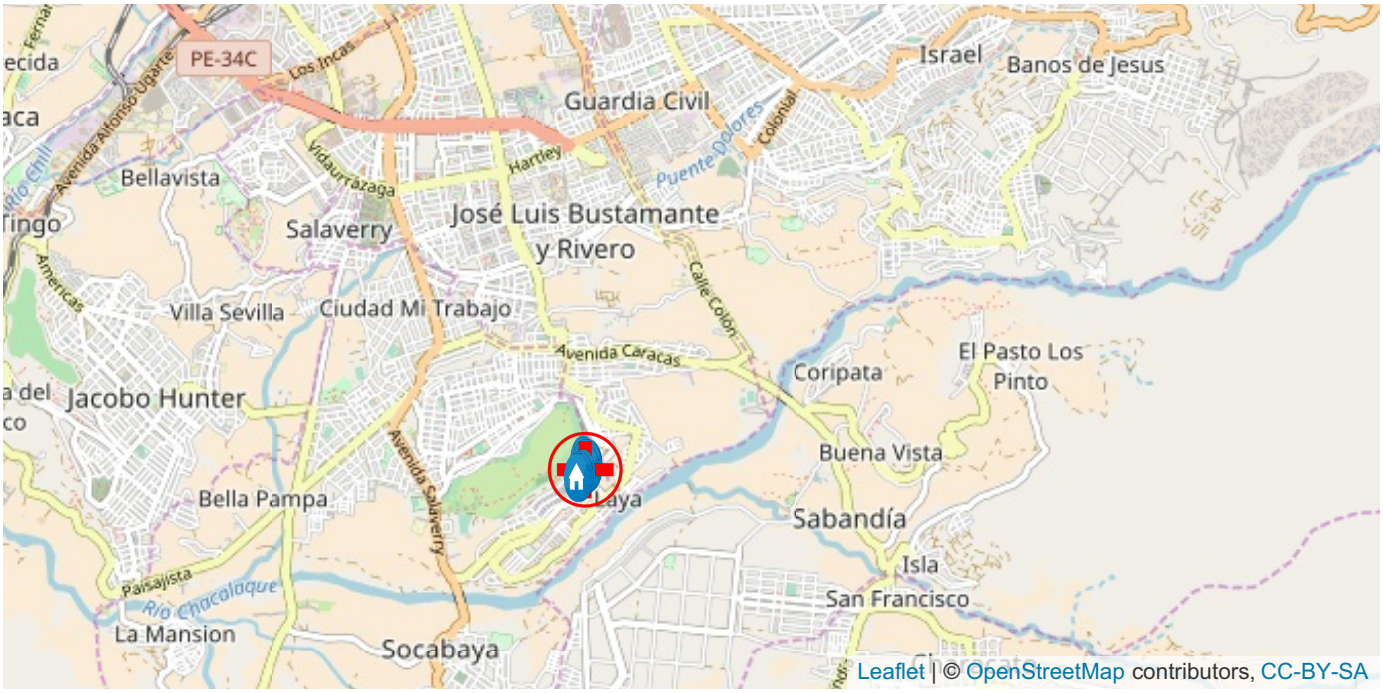
5.2. Identify the Health Centers among other houses

- Assign the nearby houses to the most optimal centroid or the health facility
- Zoom-in and check the pop-up to get an idea about the type of facility, location coordinates of each facility / house.

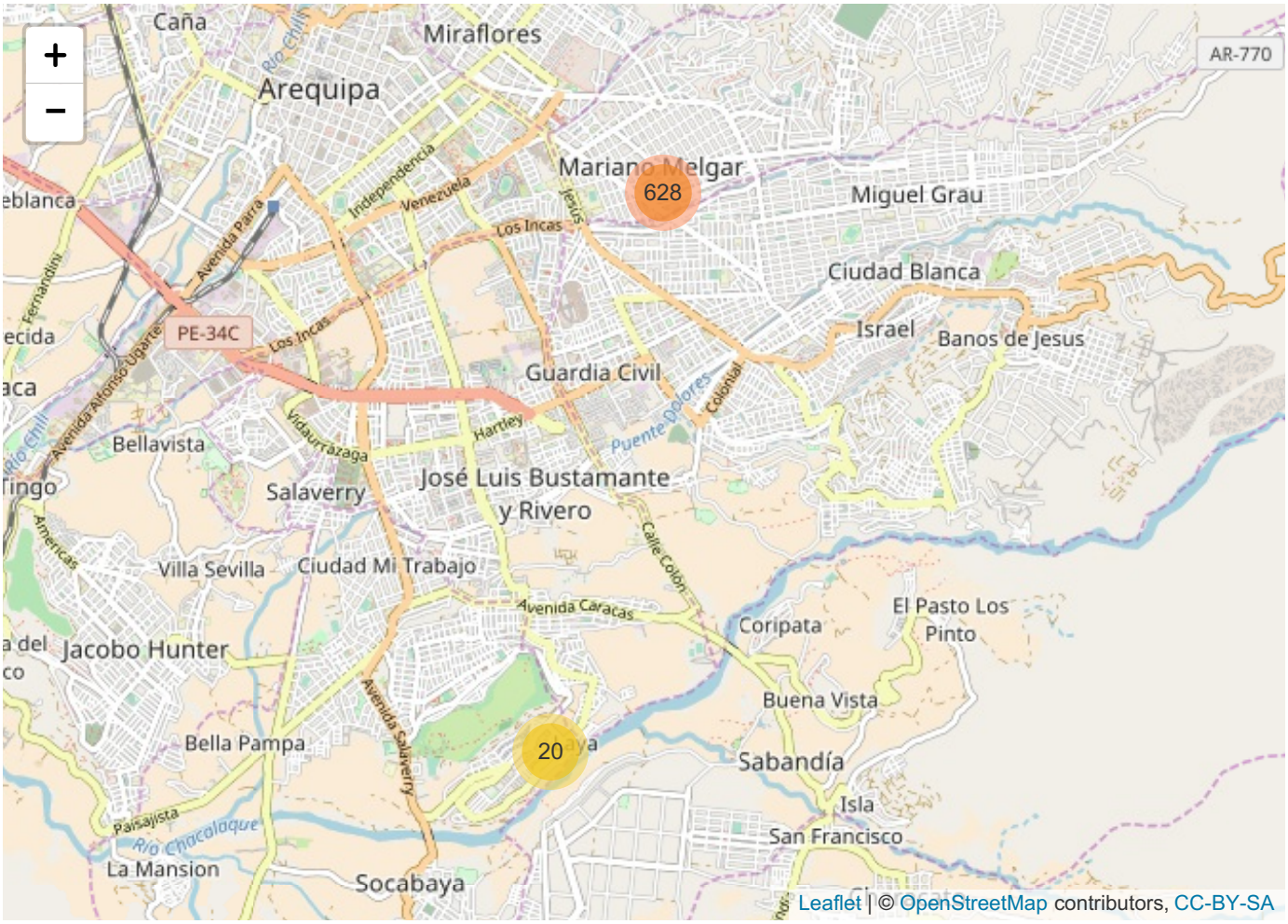
```
##   id  LATITUDE LONGITUDE km$cluster
## 1  1 -16.41048 -71.50998          5
## 2  2 -16.40770 -71.50583          4
## 3  3 -16.40775 -71.50592          4
## 4  4 -16.40776 -71.50611          4
## 5  5 -16.40783 -71.50617          4
## 6  6 -16.40788 -71.50632          4
```

```
##           id  LATITUDE LONGITUDE Cluster_No              Type
## 639      639 -16.45664 -71.51649          2 Residence / House
## 640      640 -16.45675 -71.51659          2 Residence / House
## 641      641 -16.45688 -71.51669          2 Residence / House
## 642      642 -16.45677 -71.51682          2 Residence / House
## 1100 Cluster -16.40727 -71.50378          1   Health Facility
## 2100 Cluster -16.45609 -71.51621          2   Health Facility
## 3100 Cluster -16.40736 -71.50721          3   Health Facility
## 4100 Cluster -16.40844 -71.50578          4   Health Facility
## 5100 Cluster -16.40883 -71.51043          5   Health Facility
## 643  Cluster -16.40694 -71.49940          6   Health Facility
```





- **Final Visualization: The interactive map shown below is a clustered version of the previously shown map.**
- Click the numbered bubbles repeatedly, until the cluster splits to sub groups and the logo is displayed.
- Click the logo to retreive the information pertaining to the health facility or the residence's geocoordinates.



6. RESULT

The optimal locations for installing the health facilities are displayed:


Rinconada
Granjeros Santa Marta
Las Torres de San Isidro
Pueblo Viejo
Villa Fortaleza
Yarabamba
Mollebaya
La Aguada
Giriguaya
Virgen de Chapi
Las Caseras
AR-782
AR-118
Señor de Los