Flags

GP SINGH

February 6, 2016

### The dataset Flags is chosen from UCI Machine Learning Repository.

Reading the data from repository and creating a data frame

y<- "https://archive.ics.uci.edu/ml/machine-learning-databases/flags/flag.data"  
  
x <- read.table(y, header = FALSE, sep = ",")

Assigning names to the dataset attributes

names(x) <- c("country", "landmass", "zone", "area", "population", "language",   
 "religion", "bars", "stripes", "colors", "red", "green", "blue",  
 "gold", "white", "black", "orange","mainhue", "circles", "crosses",   
 "saltires", "quarters", "sunstars", "crescents", "triangle", "icon",  
 "animate", "text", "topleft", "botright")

Create a subset of dataset with value of landmass chosen 1, 2 and 3. There might be different scenarios for selection we can choose for the dataset, I am focussing on the landmasses N.America, S.America and Europe

N.America = 1 S.America = 2 Europe = 3

set1 <- subset(x, landmass == 1 | landmass == 2 | landmass == 3)

Set 1 consists of all the information about the N. American, S.American and Europe landmasses. Set 1 has 30 attributes same as the original dataset. The dataset is furthered filtered by their attributes. The information about the flags and their description is omitted. The land dataset is oriented towards languages, population and religion of theses three landmasses. Again according to the analytical requirements, we can select the attributes we wish to consider for analysis.

land <- set1[c(1:83), 1:7]  
str(land)

## 'data.frame': 83 obs. of 7 variables:  
## $ country : Factor w/ 194 levels "Afghanistan",..: 2 5 7 8 9 10 12 13 16 17 ...  
## $ landmass : int 3 3 1 1 2 2 3 1 1 3 ...  
## $ zone : int 1 1 4 4 3 3 1 4 4 1 ...  
## $ area : int 29 0 0 0 2777 2777 84 19 0 31 ...  
## $ population: int 3 0 0 0 28 28 8 0 0 10 ...  
## $ language : int 6 6 1 1 2 2 4 1 1 6 ...  
## $ religion : int 6 0 1 1 0 0 0 1 1 0 ...

Dimension of our new dataset

dim(land)

## [1] 83 7

Frequency of data distribution among three landmasses.

table(land$landmass)

##   
## 1 2 3   
## 31 17 35

Subsetting and statistic

NAm <- subset(land, landmass==1)  
  
  
#population in round millions for N.America landmass  
mean(NAm$population)

## [1] 12.29032