Transforming Data in R in Preparation for Analysis

In this exercise, you will use R commands to rescale data, detect and deal with missing values, bin numeric data into ranges, convert between data types, combine data sets both vertically and horizontally, and, finally, write your own functions, as well as use some built-in functions.

**1: Querying from data sets**

.

1. Read the weather data set (RStudio Tutorial 1- weather.csv) into memory and rescale the Rainfall variable to the center:

**myData <- read.csv("F:\\My Documents\\Ex11-weather.csv", sep=",", header= TRUE);**

**scale(myData$Rainfall, center=TRUE, scale=FALSE)**

3. Generate z-scores for the same variable:

**scale(myData$Rainfall, center=TRUE, scale=TRUE)**

1. To detect and exclude missing values, create the following vectors and use the is.na operator to check for missing values.

x1 <- c(1, 4, 3, NA, 7)

x2 <- c("a", "B", NA, "NA")

**is.na(x1)**

**is.na(x2)**

5. Create the following vector and calculate its mean. Note the result is

NA. Calculate the mean a second time while excluding the missing values:

y <- c(1,NA,2,3)

**mean(y) # returns NA**

**mean(y, na.rm=TRUE) # returns 2**

1. Replace missing values in the WindSpeed9am attribute of myData, with the mean of that attribute:

**myData$WindSpeed9am[is.na(myData$WindSpeed9am)]**

**myData$WindSpeed9am[is.na(myData$WindSpeed9am)] <-mean(myData$WindSpeed9am,na.rm=TRUE)**

**myData$WindSpeed9am[is.na(myData$WindSpeed9am)]**

7. Load the file SocialData.csv as a variable SocialData.

**SocialData <-read.csv("F:\\MyDocuments\\SocialData.csv")**

1. View the SocialData data set before adding a new variable called **newvar**, which takes the NumberNewFriends attribute and bins it into four bins.

The binning() function is in the rattle package. Load this package into memory first.

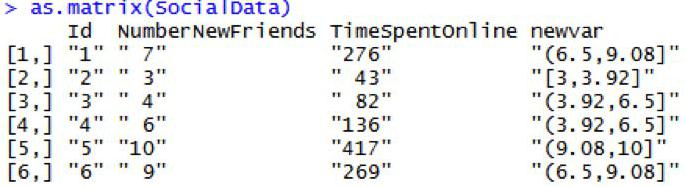
library(rattle);

SocialData

SocialData$newvar <- binning(SocialData $NumberNewFriends, 4, method="quantile", ordered=FALSE)

SocialData

1. Using the as.matrix conversion function, convert the data frame SocialData into a matrix.



1. Set the working directory to My Documents:

**setwd("F:\\My Documents")**

1. Read in the two data sets, data1.csv and data2.csv and bind them together by row.

X <-read.csv("data1.csv",header=T,sep=",")

x2 <- read.csv("data2.csv",header=T,sep=",")

rbind(x,x2)

12. Read in the data set data3.csv and bind it with the data frame x by column.

**x3 <- read.csv("data3.csv", header=T, sep=",") cbind(x,x3)**

1. Generate a sequence of numbers between 6 and 10. Now generate a sequence of numbers between 1 and 8. In one command, check to see if the first vector is contained within the second vector.

**6:10**

**6:10 %in% 1:8**

14. Write and execute a function that prints "Hello World".

**fn <- function(){ print("Hello World")**

**}**

**fn()**

* 1. **"Hello World"**

1. Write a function that accepts a value as an input parameter and returns its squared value.

**square <- function(x){ x2 <- x^2;**

**return(x2)**

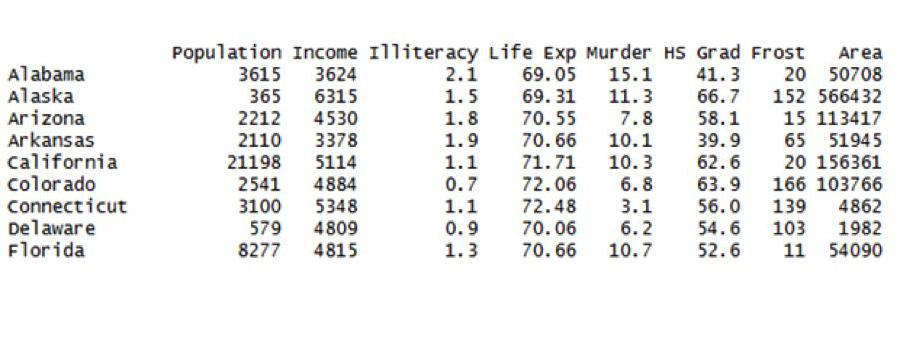
**}**

**square(3)**

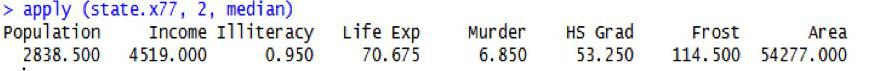
1. Load the datasets package. List the data sets available in this package and view the data in the state.x77 data set in the data viewer:

**library(datasets)**

**data(package="datasets")**

**View(state.x77)**

1. Calculate the median of every column in the state.x77 data set with one command:
2. Perform a principal component analysis (PCA) on the iris data set (hint: exclude the Species attribute).



**myPrinComp <- princomp(iris[,-5])**

1. View the proportion of the total variance explained by each component, using the summary() function.

**summary(myPrinComp)**

20. View the coefficients of the new variables.

**myPrinComp$loadings**