CADS Assignment 1 - Library Data: Reading data from Library Using R Studio

Be sure to install packages and load:

install.packages() #

library(dplyr) # There are alternative functions available in base R

library(ggplot2) # There are alternative functions available in base R as well

1. Read the dataset "USArrests" from the built in library datasets (it is available to you once we install packages) by executing the following:

a) data () # Run the code. It will give you a long list of built in datasets.

b) data (USArrests) # Run the code and it brings in the dataset USA ready to be used.

c) head (USArrests) # Gives you the first 6 rows of observations, displaying column names.

d) tail (USArrests) # Gives you the last 6 rows of observations, displaying column names.

e) colnames (USArrests) # Gives you a list of column/variable names.

f) ncol(USArrests) # Give the number of columns; notice that you may include blank spaces or not.

g) nrow(USArrests) # Gives number of columns

h) Summarize the dataset: Hint: Try summary(USArrests)

i) Which State has the largest number of arrests for murder? # You can access a column by using the "$" sign as in USArrests$Murder

j) fivenum(USArrests$Murder) # Gives FIVE-number summary for the column. Other functions available are-sum(), mean(), median(), min(), max(), var(), sd(), range(), IQR(). The command selects a column and does the calculation for that column.

h) Create a new column by adding all arrests for each State, call it TotalArrests

Use rowSum() function to calculate the new variable

Hints will be provided upon request; consider cbind() to add a new column

HInt-Use rbind() to add a new row

Caution: While doing arithmetic on a column is very common, we may want to do arithmetic on all rows but for selected columns.

i) Assign a new name to the updated dataframe. Take a look at the updated data frame by using view(), head(), tail(), colnames() etc.

j) Which State has the largest number of arrests? #Use max(TotalArrests), along with subset() function. I can provide you with the subset function

k) Is there a correlation between UrbanPop and TotalArrests?

l) What is the correlation coefficient?

m) Create a boxplot for each of the numerical variables

n) Find the regression line with x as UrbanPop and y and TotalArrests

2. Read in the data file with the name "diamonds" and answer the following questions:

This is a big dataset (tens of thousands), so be careful when inspecting the dataset to limit yourself to summary(), head(), tail().

NB: the dataset diamonds belong in the library, and so it is available to you.

a) Bring (load) in the dataset.

b) Examine the dataset. Create a summary of the dataset. Hint: summary(diamonds)

c) What are the variables?

d) How many observations?

e) Summarize the dataset.

f) What are the column names?

g) What is the maximum price and which ”carat” has it?

h) What is the minimum price and which “carat” has it?

i) Examine the correlation between x=carat and y=price.

j) Create a scatter plot for x=carat and y=price. Hint use ggplot2

k) Create the Regression line for the relationship above.

Helpful Web Resources

Statology: https://www.statology.org/

Statisticsglobe.com: https://statisticsglobe.com/r-programming-language

R Tutorials: Data to Fish: R Tutorials - Data to Fish