R-code-Topic-01\_introduction.R

duter

2021-02-01

# R code Class 01  
  
## read the data  
## you'll first need to make your 'working directory' to the folder you've stored the data.  
## You can do this from the menu bar, or with a command like  
## setwd("/Users/.../data")  
  
###########################################  
#### advertising data set  
##########################################  
ad = read.csv("Advertising.csv", stringsAsFactors = T)  
predict(  
 lm(Sales ~ TV + Radio + Newspaper, data = ad),  
 data.frame(  
 TV = 100,  
 Radio = 50,  
 Newspaper = 25  
 ),  
 interval = "confidence"  
)

## fit lwr upr  
## 1 16.91592 16.36224 17.46959

predict(lm(Sales ~ TV + Radio + Newspaper, data = ad),  
 data.frame(  
 TV = 100,  
 Radio = 50,  
 Newspaper = 25  
 ))

## 1   
## 16.91592

# ad <- na.omit(ad)  
  
#########################################################################  
getwd() # this will print your working directory

## [1] "D:/duter/Documents/JHU MSIS/Spring 2021/Data Analytics/Data\_Analytics/Week 1"

trucks <- read.csv("pickup.csv", stringsAsFactors = T)  
# trucks <- na.omit(trucks)  
  
## explore the data set  
  
nrow(trucks) # sample size

## [1] 46

# subsetting  
head(trucks)

## year miles price make  
## 1 2008 17638 14995 GMC  
## 2 2003 174000 8500 Dodge  
## 3 2001 1500 9998 Dodge  
## 4 2007 22422 23950 GMC  
## 5 2007 34815 19980 GMC  
## 6 1997 167000 5000 GMC

tail(trucks)

## year miles price make  
## 41 2001 84652 7485 GMC  
## 42 1998 166000 4500 GMC  
## 43 1994 89000 3800 GMC  
## 44 1993 90000 1900 GMC  
## 45 1989 186000 1500 GMC  
## 46 2000 125000 2500 GMC

trucks[1,] # the first observation

## year miles price make  
## 1 2008 17638 14995 GMC

trucks[1:10,] # the first 10 observations

## year miles price make  
## 1 2008 17638 14995 GMC  
## 2 2003 174000 8500 Dodge  
## 3 2001 1500 9998 Dodge  
## 4 2007 22422 23950 GMC  
## 5 2007 34815 19980 GMC  
## 6 1997 167000 5000 GMC  
## 7 1999 142000 2800 Dodge  
## 8 2003 86000 7900 Dodge  
## 9 2002 115000 6700 Dodge  
## 10 2000 207799 4500 GMC

trucks[, 1] # the first variable (year)

## [1] 2008 2003 2001 2007 2007 1997 1999 2003 2002 2000 2003 1996 1995 1998 2003  
## [16] 2005 2004 1990 2006 1995 1996 2005 2002 1999 1998 2002 2006 1996 1999 1978  
## [31] 2002 1996 1995 2004 2002 1996 1995 2006 1999 1995 2001 1998 1994 1993 1989  
## [46] 2000

trucks$year # same thing

## [1] 2008 2003 2001 2007 2007 1997 1999 2003 2002 2000 2003 1996 1995 1998 2003  
## [16] 2005 2004 1990 2006 1995 1996 2005 2002 1999 1998 2002 2006 1996 1999 1978  
## [31] 2002 1996 1995 2004 2002 1996 1995 2006 1999 1995 2001 1998 1994 1993 1989  
## [46] 2000

trucks[, 'year'] # same thing again

## [1] 2008 2003 2001 2007 2007 1997 1999 2003 2002 2000 2003 1996 1995 1998 2003  
## [16] 2005 2004 1990 2006 1995 1996 2005 2002 1999 1998 2002 2006 1996 1999 1978  
## [31] 2002 1996 1995 2004 2002 1996 1995 2006 1999 1995 2001 1998 1994 1993 1989  
## [46] 2000

# summary of each variable  
summary(trucks)

## year miles price make   
## Min. :1978 Min. : 1500 Min. : 1200 Dodge:10   
## 1st Qu.:1996 1st Qu.: 70958 1st Qu.: 4099 Ford :12   
## Median :2000 Median : 96800 Median : 5625 GMC :24   
## Mean :1999 Mean :101233 Mean : 7910   
## 3rd Qu.:2003 3rd Qu.:130375 3rd Qu.: 9725   
## Max. :2008 Max. :215000 Max. :23950

# logic subsetting  
trucks[trucks$miles > 200000,]

## year miles price make  
## 10 2000 207799 4500 GMC  
## 18 1990 215000 1200 GMC

## the make variable is special  
class(trucks$make) # it is a factor (categorical)

## [1] "factor"

levels(trucks$make) # with 3 levels

## [1] "Dodge" "Ford" "GMC"

trucks$make[1:2] # the first two obs are GMC and Dodge

## [1] GMC Dodge  
## Levels: Dodge Ford GMC

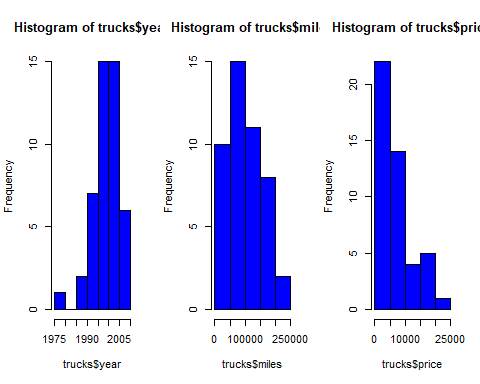
as.numeric(trucks$make[1:2]) # which R calls levels 3 and 1

## [1] 3 1

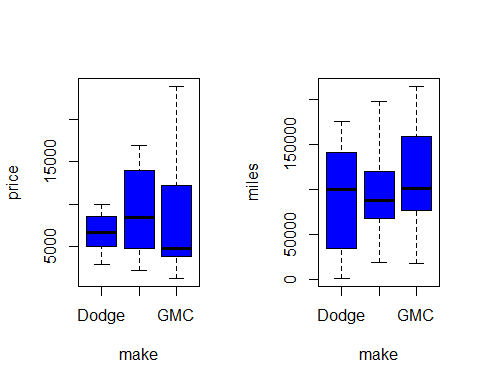
## produce some plots  
? hist()

## starting httpd help server ... done

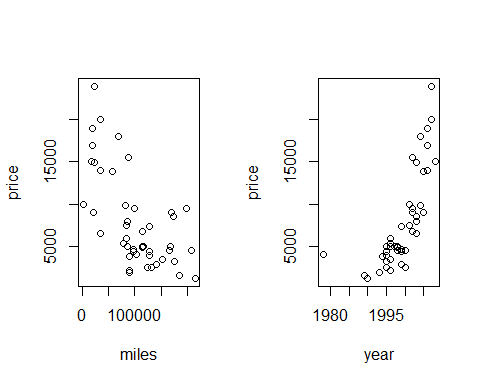
par(mfrow = c(1, 3))  
hist(trucks$year, col = "blue") ## a histogram  
hist(trucks$miles, col = "blue")  
hist(trucks$price, col = "blue")



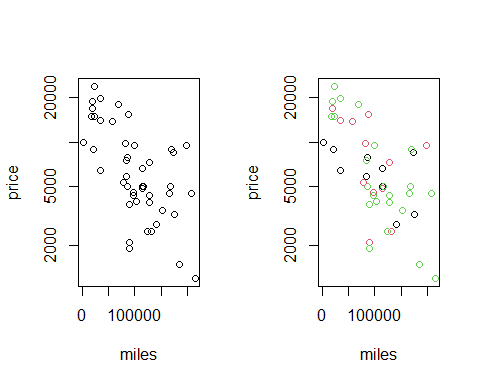
par(mfrow = c(1, 2))  
plot(price ~ make, data = trucks, col = "blue") ## a boxplot  
plot(miles ~ make, data = trucks, col = "blue")



par(mfrow = c(1, 2))  
plot(price ~ miles, data = trucks) ## simple scatterplot  
plot(price ~ year, data = trucks) ## simple scatterplot



plot(price ~ miles, data = trucks, log = "y") ## price on log scale  
plot(price ~ miles,  
 data = trucks,  
 log = "y",  
 col = trucks$make) ## in color



## add a legend (colors 1,2,3 are black,red,green)  
par(mfrow = c(1, 2))  
plot(price ~ miles, data = trucks, col = trucks$make) ## simple scatterplot  
legend("topright",  
 fill = 1:3,  
 legend = levels(trucks$make))  
  
plot(price ~ year, data = trucks, col = trucks$make) ## simple scatterplot  
legend("bottomleft",  
 fill = 1:3,  
 legend = levels(trucks$make))  
  
  
###########################################################################  
## Alumni Donation Data Set  
getwd()

## [1] "D:/duter/Documents/JHU MSIS/Spring 2021/Data Analytics/Data\_Analytics/Week 1"

dir()

## [1] "Advertising.csv"   
## [2] "Auto.csv"   
## [3] "contribution.csv"   
## [4] "Du.Minghao.HW1.docx"   
## [5] "pickup.csv"   
## [6] "R-code-Topic-00.docx"   
## [7] "R-code-Topic-00.pdf"   
## [8] "R-code-Topic-00.R"   
## [9] "R-code-Topic-01\_introduction.docx"   
## [10] "R-code-Topic-01\_introduction.R"   
## [11] "R-code-Topic-01\_introduction.spin.R"   
## [12] "R-code-Topic-01\_introduction.spin.Rmd"

don <- read.csv("contribution.csv", stringsAsFactors = T)  
head(don)

## Gender Class.Year Marital.Status Major Next.Degree FY04Giving  
## 1 M 1957 M History LLB 2500  
## 2 M 1957 M Physics MS 5000  
## 3 F 1957 M Music NONE 5000  
## 4 M 1957 M History NONE 0  
## 5 M 1957 M Biology MD 1000  
## 6 F 1957 M Mathematics NONE 0  
## FY03Giving FY02Giving FY01Giving FY00Giving AttendenceEvent  
## 1 2500 1400 12060 12000 1  
## 2 5000 5000 5000 10000 1  
## 3 5000 5000 5000 10000 1  
## 4 5100 200 200 0 1  
## 5 1000 1000 1005 1000 1  
## 6 0 0 0 0 0

class(don)

## [1] "data.frame"

summary(don)

## Gender Class.Year Marital.Status Major Next.Degree   
## F:615 Min. :1957 D: 78 History :135 NONE :460   
## M:615 1st Qu.:1967 M:711 English :123 MA :135   
## Median :1987 S:428 Biology :115 JD :103   
## Mean :1981 W: 13 Economics : 87 PHD :100   
## 3rd Qu.:1997 Political Science: 81 NDA : 72   
## Max. :1997 Psychology : 79 MS : 65   
## (Other) :610 (Other):295   
## FY04Giving FY03Giving FY02Giving FY01Giving   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0   
## Median : 0.0 Median : 0.0 Median : 0.0 Median : 0.0   
## Mean : 159.4 Mean : 241.5 Mean : 133.5 Mean : 276.5   
## 3rd Qu.: 75.0 3rd Qu.: 75.0 3rd Qu.: 51.0 3rd Qu.: 75.0   
## Max. :14655.2 Max. :58785.5 Max. :11187.3 Max. :161370.1   
##   
## FY00Giving AttendenceEvent   
## Min. : 0.0 Min. :0.0000   
## 1st Qu.: 0.0 1st Qu.:0.0000   
## Median : 0.0 Median :1.0000   
## Mean : 169.2 Mean :0.6041   
## 3rd Qu.: 60.0 3rd Qu.:1.0000   
## Max. :21000.0 Max. :1.0000   
##

attach(don)  
class(don$Marital.Status)

## [1] "factor"

levels(don$Marital.Status)

## [1] "D" "M" "S" "W"

levels(don$Next.Degree)

## [1] "AA" "BA" "BAE" "BD" "BFA" "BN" "BS" "BSE2" "BSN" "DC"   
## [11] "DDS" "DMD" "DO" "DO2" "DP" "JD" "LLB" "LLD" "MA" "MA2"   
## [21] "MAE" "MALS" "MAT" "MBA" "MCP" "MD" "MD2" "ME" "ME2" "MFA"   
## [31] "MHA" "ML" "MLS" "MM" "MPA" "MPH" "MS" "MSE" "MSM" "MSW"   
## [41] "NDA" "NONE" "PHD" "STM" "TC" "UBDS" "UDDS" "UMD" "UMDS" "UNKD"

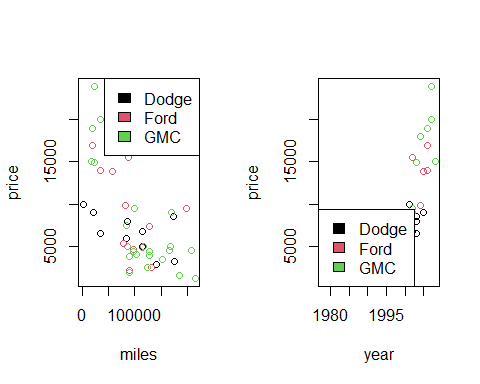
class(don$Next.Degree)

## [1] "factor"

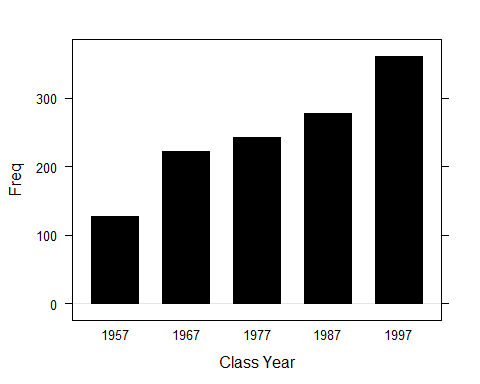
tail(don)

## Gender Class.Year Marital.Status Major Next.Degree FY04Giving  
## 1225 F 1997 S Music PHD 0  
## 1226 M 1997 M Art NONE 0  
## 1227 F 1997 M Art NDA 0  
## 1228 F 1997 S Spanish NDA 10  
## 1229 F 1997 S Spanish NONE 25  
## 1230 M 1997 S Economics NDA 158  
## FY03Giving FY02Giving FY01Giving FY00Giving AttendenceEvent  
## 1225 0 10 0 5 1  
## 1226 0 0 0 0 0  
## 1227 0 0 0 0 0  
## 1228 20 20 20 20 0  
## 1229 0 0 10 0 0  
## 1230 0 50 20 50 1

# fix(don)  
  
library(lattice)



barchart(  
 table(Class.Year),  
 horizontal = F,  
 xlab = "Class Year",  
 col = "black"  
)



don$TGiving = don[, 'FY00Giving'] + don[, 'FY01Giving'] + don[, 'FY02Giving'] + don[, 'FY03Giving'] + don[, 'FY04Giving']  
head(don$TGiving)

## [1] 30460 30000 30000 5500 5005 0

quantile(don$TGiving, probs = seq(0.95, 1, 0.01))

## 95% 96% 97% 98% 99% 100%   
## 2277.50 3133.56 5000.00 7000.00 16442.14 171870.06

head(don)

## Gender Class.Year Marital.Status Major Next.Degree FY04Giving  
## 1 M 1957 M History LLB 2500  
## 2 M 1957 M Physics MS 5000  
## 3 F 1957 M Music NONE 5000  
## 4 M 1957 M History NONE 0  
## 5 M 1957 M Biology MD 1000  
## 6 F 1957 M Mathematics NONE 0  
## FY03Giving FY02Giving FY01Giving FY00Giving AttendenceEvent TGiving  
## 1 2500 1400 12060 12000 1 30460  
## 2 5000 5000 5000 10000 1 30000  
## 3 5000 5000 5000 10000 1 30000  
## 4 5100 200 200 0 1 5500  
## 5 1000 1000 1005 1000 1 5005  
## 6 0 0 0 0 0 0

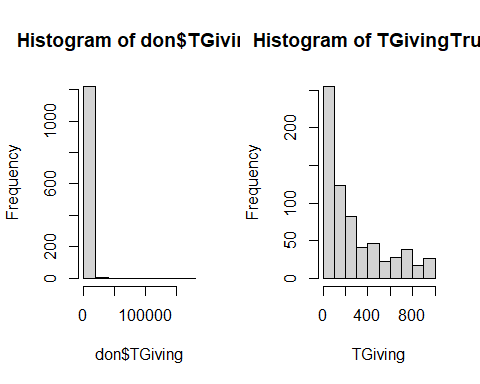
quantile(don$TGiving, probs = seq(0.95, 1, 0.01))

## 95% 96% 97% 98% 99% 100%   
## 2277.50 3133.56 5000.00 7000.00 16442.14 171870.06

sum(don$TGiving)

## [1] 1205454

hist(don$TGiving)  
ff1 = don$TGiving[don$TGiving != 0]  
ff2 = ff1[ff1 <= 1000]  
hist(ff2,  
 main = paste("Histogram of TGivingTrunc"),  
 xlab = "TGiving")



ff2[1:5]

## [1] 500 500 785 500 250

ff1[1:5]

## [1] 30460 30000 30000 5500 5005

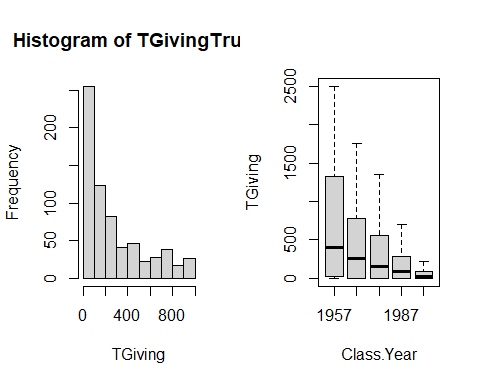
ff1 = don$TGiving[don$TGiving != 0]  
ff1[1:5]

## [1] 30460 30000 30000 5500 5005

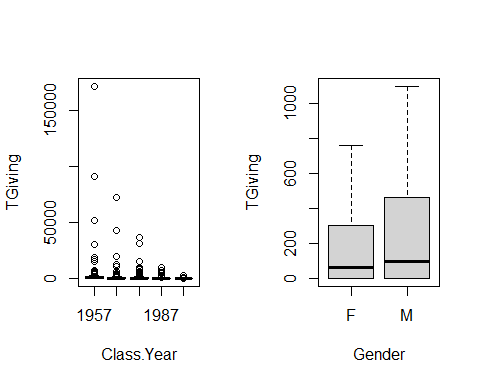
ff2 = ff1[ff1 <= 1000]  
ff2[1:5]

## [1] 500 500 785 500 250

hist(ff2,  
 main = paste("Histogram of TGivingTrunc"),  
 xlab = "TGiving")  
boxplot(TGiving ~ Class.Year, data = don, outline = FALSE)



boxplot(TGiving ~ Class.Year, data = don, outline = T)  
boxplot(TGiving ~ Gender, data = don, outline = F)



boxplot(TGiving ~ Marital.Status, data = don, outline = F)

