Data Exploration

Hazim Fitri

2024-12-29

Contents

Import data into R

.txt file

```
read.table('Rclass.txt', sep=';')
```

```
##             V1
## 1 Name\tWeight
## 2      Ali\t60
## 3      Abu\t65
## 4    Ahmad\t70
```

.csv file

```
#read.csv('Rclass.csv', sheet='Rxls')
```

.xlsx file

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.4.2
```

```
#read_excel('Rclass.xlsx')
```

Clipboard

```
#read.table(file='clipboard', header=T)
```

Export data from R

.txt file

```
mydata = data.frame()
#fix(mydata)
mydata
```

```
## data frame with 0 columns and 0 rows
```

```
write.table(mydata, file='Rdata.txt', col.names=T, row.names=F)
getwd()
```

```
## [1] "C:/Users/hazim/OneDrive - Universiti Kebangsaan Malaysia/Math-Stat/09. Data Exploration with R"
```

.csv file

```
write.csv(mydata, file='Rdata.csv', col.names=T, row.names = F)
```

```
## Warning in write.csv(mydata, file = "Rdata.csv", col.names = T, row.names = F):
## attempt to set 'col.names' ignored
```

Clipboard

```
write.table(mydata, col.names=T, row.names = F, file=)
```

```
## ""
```

Descriptive Statistics

1. Central tendencies
2. Quartiles
3. Variation
4. Shape

Central Tendencies

```
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

```
mean(trees$Girth)
```

```
## [1] 13.24839
```

```
rowMeans(trees)
```

```
##  [1] 29.53333 27.96667 27.33333 32.96667 36.83333 37.83333 30.86667 34.73333
##  [9] 37.90000 35.36667 38.16667 36.13333 36.26667 34.00000 35.36667 36.36667
## [17] 43.90000 42.23333 36.80000 34.23333 42.16667 41.96667 41.60000 42.10000
## [25] 45.30000 51.23333 51.73333 52.06667 49.83333 49.66667 61.53333
```

```
colMeans(trees)
```

```
##    Girth   Height   Volume
## 13.24839 76.00000 30.17097
```

```
median(trees$Girth)
```

```
## [1] 12.9
```

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.4.2
```

```
Mode(trees$Girth)
```

```
## [1] 11.0 11.4 12.9 18.0
## attr(,"freq")
## [1] 2
```

```r
x <- c(1,2,2,2,3,4,5,6,7,8,9)
table(x)
```

```
## x
## 1 2 3 4 5 6 7 8 9
## 1 3 1 1 1 1 1 1 1
```

```r
Mode(x)
```

```
## [1] 2
## attr(,"freq")
## [1] 3
```

Quartile

```r
# by default, it'll give the 0, 0.25, 0.5, 0.75, 100
quantile(trees$Girth,c(0.25, 0.75, 0.99))
```

```
##   25%   75%   99%
## 11.05 15.25 19.82
```

```r
summary(trees$Girth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.30   11.05   12.90   13.25   15.25   20.60
```

Measures of Variation / Scale of Parameters

```r
max(trees$Girth) - min(trees$Girth)
```

```
## [1] 12.3
```

```r
range(trees$Girth)
```

```
## [1]  8.3 20.6
```

```r
IQR(trees$Girth)
```

```
## [1] 4.2
```

Variance

It'll provide the sample variance and not the population variance

```r
#var()
#sd() # This will also provide the sample standard deviation
```

Shape

Skewness

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.4.2
```

Kurtosis

Exercise

```
batting = read.table('batting.history.txt', header = T)
colnames(batting)
```

```
##  [1] "Year"   "Tms"    "N.Bat"  "BatAge" "R"      "G"      "PA"     "AB"
##  [9] "H"      "X2B"    "X3B"    "HR"     "RBI"    "SB"     "CS"     "BB"
## [17] "SO"     "BA"     "OBP"    "SLG"    "OPS"    "TB"     "GDP"    "HBP"
## [25] "SH"     "SF"     "IBB"
```

```
BATTING = data.frame(
  Year = batting$Year,
  Tms = batting$Tms,
  N.Bat = batting$N.Bat,
  BatAge = batting$BatAge,
  Avg_PA = mean(batting$PA),
  Avg_AB = mean(batting$AB),
  Avg_H = mean(batting$H)
)

write.csv(BATTING, 'BATTING.csv', row.names = F)
```

```
new = rowMeans(batting[,7:9])
battingnew = data.frame(batting[,1:4], new)
write.csv(battingnew, file='BATTING2.csv', col.names=T, row.names=F)
```

```
## Warning in write.csv(battingnew, file = "BATTING2.csv", col.names = T,
## row.names = F): attempt to set 'col.names' ignored
```

Data Wrangling

```
#head(trees,2)
#tail(trees,2)
#tail(tees, -5) # All observation except the last 5
```

```
which(mtcars$cyl  == 6)
```

```
## [1]  1  2  4  6 10 11 30
```

```
mtcars[which(mtcars$cyl  == c(6,8)),]
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
```

```r
subset(mtcars, subset = mpg<20, select=c(mpg, disp))
```

```
##                    mpg  disp
## Hornet Sportabout   18.7 360.0
## Valiant             18.1 225.0
## Duster 360          14.3 360.0
## Merc 280            19.2 167.6
## Merc 280C           17.8 167.6
## Merc 450SE          16.4 275.8
## Merc 450SL          17.3 275.8
## Merc 450SLC         15.2 275.8
## Cadillac Fleetwood  10.4 472.0
## Lincoln Continental 10.4 460.0
## Chrysler Imperial   14.7 440.0
## Dodge Challenger    15.5 318.0
## AMC Javelin         15.2 304.0
## Camaro Z28          13.3 350.0
## Pontiac Firebird    19.2 400.0
## Ford Pantera L      15.8 351.0
## Ferrari Dino        19.7 145.0
## Maserati Bora       15.0 301.0
```

```r
sort(mtcars$mpg, decreasing = T)
```

```
##  [1] 33.9 32.4 30.4 30.4 27.3 26.0 24.4 22.8 22.8 21.5 21.4 21.4 21.0 21.0 19.7
## [16] 19.2 19.2 18.7 18.1 17.8 17.3 16.4 15.8 15.5 15.2 15.2 15.0 14.7 14.3 13.3
## [31] 10.4 10.4
```

```r
order(mtcars$mpg)
```

```
##  [1] 15 16 24  7 17 31 14 23 22 29 12 13 11  6  5 10 25 30  1  2  4 32 21  3  9
## [26]  8 27 26 19 28 18 20
```

```r
mons <- c("March","April","January","November","January","September",
"October","September", "November","August","January",
"November", "November", "February","May","August","July",
"December","August","August","September","November",
"February","April")
table(mons)
```

```
## mons
##     April    August  December  February   January      July     March       May
##         2         4         1         2         3         1         1         1
##  November   October September
##         5         1         3
```

```r
mons <- factor(mons,levels=c("January","February","March",
"April","May","June","July", "August","September", "October",
"November", "December"), ordered=TRUE)
table(mons)
```

```
## mons
##   January  February     March     April       May      June      July    August
##         3         2         1         2         1         0         1         4
## September   October  November  December
##         3         1         5         1
```

```r
names(table(mons))
```

```
##  [1] "January"   "February"  "March"     "April"     "May"       "June"
##  [7] "July"      "August"    "September" "October"   "November"  "December"
```

```
lapply(trees, FUN=median)
```

```
## $Girth
## [1] 12.9
##
## $Height
## [1] 76
##
## $Volume
## [1] 24.2
```

```
do.call(rbind, lapply(trees, FUN=median))
```

```
##        [,1]
## Girth  12.9
## Height 76.0
## Volume 24.2
```

```
mine<-c("a","b","c")
paste(mine, "!")
```

```
## [1] "a !" "b !" "c !"
```

```
paste0(mine, "!")
```

```
## [1] "a!" "b!" "c!"
```

```
paste(mine, 1:3)
```

```
## [1] "a 1" "b 2" "c 3"
```

```
paste(mine, 1:3, sep=',')
```

```
## [1] "a,1" "b,2" "c,3"
```

```
paste0(mine, 1:3, collapse=',')
```

```
## [1] "a1,b2,c3"
```

Exercise

Question 2

```
length(mtcars$cyl)
```

```
## [1] 32
```

```
table(factor(mtcars$cyl))/32*100
```

```
##
##      4      6      8
## 34.375 21.875 43.750
```

```r
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```r
data(mtcars)

variables <- c("mpg", "disp", "hp", "drat", "wt", "qsec")

iqr_values <- sapply(mtcars[variables], IQR)

print(iqr_values)
```

```
##       mpg      disp        hp      drat        wt      qsec
##   7.37500 205.17500  83.50000   0.84000   1.02875   2.00750
```

```r
data(mtcars)

mtcars$mpg_factor <- cut(mtcars$mpg,
                         breaks = quantile(mtcars$mpg, probs = seq(0, 1, 0.25)),
                         include.lowest = TRUE,
                         labels = c('Low', "Medium", "High", "Very High"))

mtcars$disp_factor <- cut(mtcars$disp,
                          breaks = quantile(mtcars$disp, probs = seq(0, 1, 0.25)),
                          include.lowest = TRUE,
                          labels = c("S", "M", "L", "XL"))

factor_combination <- table(mtcars$mpg_factor, mtcars$disp_factor)

print(factor_combination)
```

```
##
##             S M L XL
##   Low       0 0 3  5
##   Medium    0 2 4  3
##   High      2 5 1  0
##   Very High 6 1 0  0
```

Question 3

```r
animal = data.frame(Farm = c('MO', 'MO', 'MO', 'MO', 'LN', 'SE', 'QM'),
           Month = c('11', '07', '07', NA, '09', '09', '11'),
           Year = c('00', '00', '01', NA, '03', '03', '02'),
           Sex = c(1,2,2,2,1,2,2),
           LengthClass = c(1,1,1,1,1,1,1),
           LengthCT = c(75, 85, 91.6, 95, NA, 105.5, 106),
           Ecervi = c(0,0,0,NA,0,0,0),
           Tb = c(0,0,1,NA,0,0,0))

animal
```

```
##   Farm Month Year Sex LengthClass LengthCT Ecervi Tb
## 1   MO    11   00   1           1     75.0      0  0
## 2   MO    07   00   2           1     85.0      0  0
## 3   MO    07   01   2           1     91.6      0  1
## 4   MO  <NA> <NA>   2           1     95.0     NA NA
## 5   LN    09   03   1           1       NA      0  0
## 6   SE    09   03   2           1    105.5      0  0
## 7   QM    11   02   2           1    106.0      0  0
```

```r
mean(animal$LengthCT, na.rm = T)
```

```
## [1] 93.01667
```

```r
nrow(subset(animal, animal$Sex==1))
```

```
## [1] 2
```

```r
nrow(subset(animal, animal$Tb==1))
```

```
## [1] 1
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
animal %>% mutate(Sqrt = sqrt(animal$LengthCT))
```

```
##   Farm Month Year Sex LengthClass LengthCT Ecervi Tb      Sqrt
## 1   MO    11   00   1           1     75.0      0  0  8.660254
## 2   MO    07   00   2           1     85.0      0  0  9.219544
## 3   MO    07   01   2           1     91.6      0  1  9.570789
## 4   MO  <NA> <NA>   2           1     95.0     NA NA  9.746794
## 5   LN    09   03   1           1       NA      0  0        NA
## 6   SE    09   03   2           1    105.5      0  0 10.271319
## 7   QM    11   02   2           1    106.0      0  0 10.295630
```

```r
mean_animal = mean(animal$LengthCT, na.rm=T)
mae = mean(abs(animal$LengthCT - mean_animal), na.rm = T)
mae
```

```
## [1] 9.15
```