Data Summarization in R

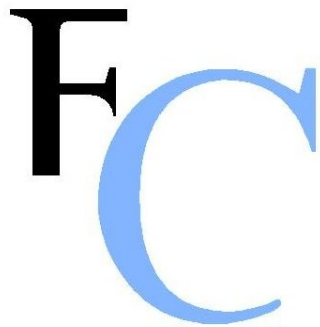
Solutions to Hands On Exercises

L. Torgo

ltorgo@fc.up.pt

Departamento de Ciência de Computadores / Faculdade de Ciências
Universidade do Porto

Feb, 2021



Hands on Summarization - the algae data set

Concerning the algae data set answer the following question:

1. Which season has more water samples? solution
2. What is the average value of `a5`? solution
3. What is the average value of `NO3`? solution
4. Check if there are unusually high values of `a2` and show the respective water samples. solution
5. Obtain a summary of the basic descriptive statistics of `a1` and `a4`, for each season of the year. solution
6. Try to obtain a table with the seasons ordered by decreasing average value of `NO3`. Hint: explore the capabilities of the function `aggregate()` that has similar objectives as the function `by()`. Also explore the function `order()`. solution

Solution to exercise 1

- Which season has more water samples?

```
data(algae,package="DMwR2")
table(algae$season)

##
## autumn spring summer winter
##     40     53     45     62
```

```
counts <- table(algae$season)
names(counts)[which.max(counts)]

## [1] "winter"
```

Solution to exercise 1 - cont.

- Which season has more water samples?

```
data(algae,package="DMwR2")
library(dplyr)
group_by(algae,season) %>% tally() %>% arrange(desc(n)) %>% slice(1)

## # A tibble: 1 x 2
##   season     n
##   <fct>  <int>
## 1 winter    62
```

Solution to exercise 1 - cont.

- Which season has more water samples?

Even simpler solution!

```
library(dplyr)
library(DMwR2)
data(algae)
centralValue(algae$season)

## [1] "winter"
```

Go Back

Solution to exercise 2

- What is the average value of `a5`?

```
mean(algae$a5)

## [1] 5.0645
```

```
summarize(algae,avgA5=mean(a5))

##     avgA5
## 1 5.0645
```

Go Back

Solution to exercise 2

- What is the average value of `a5`?

```
mean(algae$a5)

## [1] 5.0645
```

```
summarize(algae,avgA5=mean(a5))

##    avgA5
## 1 5.0645
```

Go Back

Solution to exercise 3

- What is the average value of `NO3`?

```
mean(algae$NO3)

## [1] NA
```

```
mean(algae$NO3,na.rm=TRUE)

## [1] 3.282389
```

```
summarize(algae,avgA5=mean(NO3,na.rm=TRUE))

##      avgA5
## 1 3.282389
```

Go Back

Solution to exercise 3

- What is the average value of `NO3`?

```
mean(algae$NO3)

## [1] NA
```

```
mean(algae$NO3,na.rm=TRUE)

## [1] 3.282389
```

```
summarize(algae,avgA5=mean(NO3,na.rm=TRUE))

##      avgA5
## 1 3.282389
```

Go Back

Solution to exercise 3

- What is the average value of `NO3`?

```
mean(algae$NO3)

## [1] NA
```

```
mean(algae$NO3,na.rm=TRUE)

## [1] 3.282389
```

```
summarize(algae,avgA5=mean(NO3,na.rm=TRUE))

##      avgA5
## 1 3.282389
```

Go Back

Solution to exercise 4

- Check if there are unusually high values of `a2` and show the respective water samples.

```
boxplot.stats(algae$a2)$stats
```

```
## [1]  0.00  0.00  3.00 11.45 28.20
```

```
subset(algae,a2 > 28.2)
```

```
##     season   size  speed mxPH mnO2     Cl   NO3       NH4    oPO4     PO4
## 3   autumn  small medium 8.10 11.4 40.020 5.330  346.667 125.667 187.057
## 4   spring  small medium 8.07  4.8 77.364 2.302   98.182  61.182 138.700
## 35  winter  small medium 8.27  7.8 29.200 0.050 6400.000   7.400  23.000
## 39  winter  small medium 8.30  8.9 20.625 3.414  228.750 196.620 253.250
## 43  winter  small   high 8.30  7.7 50.000 8.543   76.000 264.900 344.600
## 97  winter medium    low 9.10  5.4 61.050 0.308  105.556 104.222 239.000
## 110 summer medium   high 8.16 11.1 32.056 5.694  461.875  71.000 132.546
## 127 winter medium medium 9.10 11.6 31.091 5.099  246.364  55.000 284.000
## 129 summer medium medium 8.30 10.0 30.125 3.726  102.500  75.875 177.625
## 137 autumn medium medium 8.10 11.7 35.660 5.130   46.500  49.000  88.500
##       Chla  a1   a2   a3  a4  a5   a6   a7
## 3   15.600 3.3 53.6  1.9 0.0 0.0  0.0  9.7
## 4    1.400 3.1 41.0 18.9 0.0 1.4  0.0  1.4
## 35   0.900 5.3 40.7  3.3 0.0 0.0  0.0  1.9
## 39  12.320 2.0 38.5  4.1 2.2 0.0  0.0 10.2
## 43  22.500 0.0 40.9  7.5 0.0 2.4  1.5  0.0
## 97  72.478 3.6 31.9  2.4 0.0 0.0  0.0  2.2
## 110 15.000 3.6 38.8  0.0 0.0 1.2  0.0  2.4
## 127 88.255 0.0 36.6  4.1 0.0 1.2 16.7  6.1
```

Solution to exercise 4 - cont.

- Check if there are unusually high values of `a2` and show the respective water samples.

```
isout <- function(x) {
    q13 <- quantile(x,probs=c(0.25,0.75))
    iq <- q13[2]-q13[1]
    x < q13[1]-1.5*iq | x > q13[2]+1.5*iq
}
filter(algae,isout(a2))
```

```
##    season   size  speed mxPH mnO2     Cl   NO3      NH4    oPO4     PO4
## 1  autumn  small medium 8.10 11.4 40.020 5.330  346.667 125.667 187.057
## 2  spring  small medium 8.07  4.8 77.364 2.302   98.182  61.182 138.700
## 3  winter  small medium 8.27  7.8 29.200 0.050 6400.000   7.400  23.000
## 4  winter  small medium 8.30  8.9 20.625 3.414  228.750 196.620 253.250
## 5  winter  small   high 8.30  7.7 50.000 8.543   76.000 264.900 344.600
## 6  winter medium    low 9.10  5.4 61.050 0.308  105.556 104.222 239.000
## 7  summer medium   high 8.16 11.1 32.056 5.694  461.875  71.000 132.546
## 8  winter medium medium 9.10 11.6 31.091 5.099  246.364  55.000 284.000
## 9  summer medium medium 8.30 10.0 30.125 3.726  102.500  75.875 177.625
## 10 autumn medium medium 8.10 11.7 35.660 5.130   46.500  49.000  88.500
##      Chla  a1   a2   a3  a4  a5   a6   a7
## 1  15.600 3.3 53.6  1.9 0.0 0.0  0.0  9.7
## 2   1.400 3.1 41.0 18.9 0.0 1.4  0.0  1.4
## 3   0.900 5.3 40.7  3.3 0.0 0.0  0.0  1.9
## 4  12.320 2.0 38.5  4.1 2.2 0.0  0.0 10.2
## 5  22.500 0.0 40.9  7.5 0.0 2.4  1.5  0.0
## 6  72.478 3.6 31.9  2.4 0.0 0.0  0.0  2.2
## 7  15.000 3.6 38.8  0.0 0.0 1.2  0.0  2.4
```

Solution to exercise 5

- Obtain a summary of the basic descriptive statistics of `a1` and `a4`, for each season of the year.

```
by(algae[,c("a1","a4")],algae$season,summary)
```

```
## algae$season: autumn
##        a1               a4
##  Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 2.65   1st Qu.: 0.000
##  Median : 8.50   Median : 0.000
##  Mean   :17.75   Mean   : 1.133
##  3rd Qu.:23.98   3rd Qu.: 1.200
##  Max.   :86.60   Max.   :11.500
## ------------------------------------------------------------
## algae$season: spring
##        a1               a4
##  Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 1.70   1st Qu.: 0.000
##  Median : 4.10   Median : 1.000
##  Mean   :16.65   Mean   : 3.087
##  3rd Qu.:20.30   3rd Qu.: 3.200
##  Max.   :89.80   Max.   :44.600
...
...
```

Solution to exercise 6

- Try to obtain a table with the seasons ordered by decreasing average value of `NO3`. Hint: explore the capabilities of the function `aggregate()` that has similar objectives as the function `by()`. Also explore the function `order()`.

```
tab <- aggregate(algae$NO3,list(algae$season),mean,na.rm=TRUE)
colnames(tab) <- c("season","avgNO3")
tab[order(tab$avgNO3,decreasing=TRUE),]

##   season   avgNO3
## 1 autumn 4.496025
## 3 summer 3.237523
## 4 winter 3.212443
## 2 spring 2.484189
```

Solution to exercise 6 - cont.

- Try to obtain a table with the seasons ordered by decreasing average value of `NO3`. Hint: explore the capabilities of the function `aggregate()` that has similar objectives as the function `by()`. Also explore the function `order()`.

```
group_by(algae,season) %>%
    summarise(avgNO3=mean(NO3,na.rm=TRUE)) %>%
    arrange(desc(avgNO3))

## # A tibble: 4 × 2
##   season   avgNO3
##   <fctr>    <dbl>
## 1 autumn 4.496025
## 2 summer 3.237523
## 3 winter 3.212443
## 4 spring 2.484189
```

Go Back