Data Manipulation in R

Solutions to Hands On Exercises

L. Torgo

ltorgo@dal.ca

Faculty of Computer Science / Institute for Big Data Analytics
Dalhousie University

May, 2021

Hands On Data Manipulation with dplyr

Package **mlbench** (an extra package that you need to install) contains several data sets (from UCI repository). Load the Ozone data set and check its help page to understand what this data is about. Answer these questions:

1. Create a data frame table with the data for easier manipulation solution
2. What is the average Humidity per Month? solution
3. Show the information of Mondays solution
4. For each combination of *Month* and *Day of the Week* obtain the maximum and minimum temperature at the *Sand* location solution

Solution to Exercise 1

- Create a data frame table with the data for easier manipulation

```
data(Ozone,package="mlbench")
colnames(Ozone) <- c("Month","DayMonth","DayWeek","mxOZ","Press",
                     "WindSp","Hum","TempSand","TempEl","InvHeig",
                     "PressGrad","InvTemp","Vis")
library(dplyr)
oz <- as_tibble(Ozone)
```

Go Back

Solution to Exercise 2

- What is the average Humidity per Month?

```
group_by(oz,Month) %>% summarise(avgH=mean(Hum,na.rm=TRUE))
```

```
## # A tibble: 12 x 2
##    Month  avgH
##    <fct> <dbl>
##  1 1      37.4
##  2 2      56.9
##  3 3      49.8
##  4 4      58.2
##  5 5      70.9
##  6 6      63.3
##  7 7      72.8
##  8 8      67.6
##  9 9      73.6
## 10 10     62.7
## 11 11     50.1
## 12 12     36.1
```

Go Back

Solution to Exercise 3

- Show the information of Mondays

```
filter(oz,DayWeek==1)
```

```
## # A tibble: 52 x 13
##    Month DayMonth DayWeek  mxOZ Press
##    <fct> <fct>    <fct>   <dbl> <dbl>
##  1 1     5        1           5  5760
##  2 1     12       1           6  5720
##  3 1     19       1           4  5680
##  4 1     26       1           5  5780
##  5 2     2        1          12  5770
##  6 2     9        1           3  5490
##  7 2     16       1           7  5730
##  8 2     23       1           4  5690
##  9 3     1        1           2  5550
## 10 3     8        1           7  5580
## # ... with 42 more rows, and 8 more
## #   variables: WindSp <dbl>, Hum <dbl>,
## #   TempSand <dbl>, TempEl <dbl>,
## #   InvHeig <dbl>, PressGrad <dbl>,
## #   InvTemp <dbl>, Vis <dbl>
```

Solution to Exercise 4

- For each combination of *Month* and *Day of the Week* obtain the maximum and minimum temperature at the *Sand* location

```
group_by(oz,Month,DayWeek) %>%
    summarize(mnTempSd=min(TempSand,na.rm=TRUE), mxTempSd=max(TempSand,na.rm=TRUE))

## `summarise()` has grouped output by 'Month'.  You can override using the `.groups` argument.

## # A tibble: 84 x 4
## # Groups:   Month [12]
##    Month DayWeek mnTempSd mxTempSd
##    <fct> <fct>      <dbl>    <dbl>
##  1 1     1             51       56
##  2 1     2             35       59
##  3 1     3             45       59
##  4 1     4             55       64
##  5 1     5             38       69
##  6 1     6             40       64
##  7 1     7             45       67
##  8 2     1             37       63
##  9 2     2             41       54
## 10 2     3             36       60
## # ... with 74 more rows
```

Go Back