IODS course project

Josephine Dufitinema

Contents

Introduction to Open Data Science - Course Project

About the Course

I heard about the course from an email send by **Prof. Kimmo**, himself, and decided to join. So thanks **Kimmo**. I am so excited to learn about **all things Data Science**.

Expect to learn

I have used **R** and **Rstudio** for about three years now. However, I have not yet tried out **GitHub**. I thought this would be a good chance and refresh my analytics skills.

Link to my GitHub repository

Here is the link to my GitHub **Josephine's GitHub**.

A curious question

As I am a regular R user, I guess things like

- Import and read data
- Data visualization
- Run linear models
- Explore data

```
head(mtcars$disp) # few first
```

```
## [1] 160 160 108 258 360 225
```

```
tail(mtcars$disp) # few last
```

```
## [1] 120.3  95.1 351.0 145.0 301.0 121.0
```

Would be familiar. I was wondering if, during the course, we will also explore writing **R packages**.

Regression and model validation

This chapter focuses on performing and interpreting regression analysis.

Read the students2014 data from my local folder

The data is from an international survey of Approaches to Learning, made possible by Teachers' Academy funding for KV in 2013-2015. The data has been filtered to include the desirable variables for analysis. The original data and variables descriptions can be found **here**.

```
data_analysis <- read.csv("learning2014.csv")
str(data_analysis) # The data structure is data frame
```

```
## 'data.frame':    166 obs. of  7 variables:
##  $ gender  : chr  "F" "M" "F" "M" ...
##  $ Age     : int  53 55 49 53 49 38 50 37 37 42 ...
##  $ Attitude: int  37 31 25 35 37 38 35 29 38 21 ...
##  $ deep    : num  3.58 2.92 3.5 3.5 3.67 ...
##  $ stra    : num  3.38 2.75 3.62 3.12 3.62 ...
##  $ surf    : num  2.58 3.17 2.25 2.25 2.83 ...
##  $ Points  : int  25 12 24 10 22 21 21 31 24 26 ...
```

```
dim(data_analysis) # The data contains 166 observations
```

```
## [1] 166   7
```

```
                   # or rows and 7 variables or columns.
head(data_analysis, n = 3) # Three first rows
```

```
##   gender Age Attitude     deep  stra     surf Points
## 1      F  53       37 3.583333 3.375 2.583333     25
## 2      M  55       31 2.916667 2.750 3.166667     12
## 3      F  49       25 3.500000 3.625 2.250000     24
```

```
tail(data_analysis, n = 3) # Three last rows
```

```
##     gender Age Attitude     deep  stra     surf Points
## 164      F  18       37 3.166667 2.625 3.416667     18
## 165      F  19       36 3.416667 2.625 3.000000     30
## 166      M  21       18 4.083333 3.375 2.666667     19
```

(more chapters to be added similarly as we proceed with the course!)