Data Handling: Import, Cleaning and Visualisation

Exercise to lecture 4: csv and arrays

Dr. Aurélien Sallin

Working with a data frame

Import data

Have a look at the file `financial_data.txt` using your favorite text editor. What do you notice?

Import the table using the `read.csv()` function in your environment. Make sure you have the right path to access the .txt document. What does this parser do? Explore the data.frame. What is its structure? What are its dimensions?

Partial solution:

```
# Set correct path!

# Import
financial_data <- read.csv("financial_data.txt", sep = ":")
financial_data
```

```
##     Firm Year Revenue    Profit Category
## 1  FirmA 2018    3462 1327.3730     Tech
## 2  FirmB 2018    3510 1114.8687  Finance
## 3  FirmC 2018    3226 1089.2809     Tech
## 4  FirmD 2018    1525  328.8874  Finance
## 5  FirmE 2018    1194  189.6615   Health
## 6  FirmA 2019    3985 1933.5606     Tech
## 7  FirmB 2019    2841 1309.4726   Health
## 8  FirmC 2019    2141  805.6200     Tech
## 9  FirmD 2019    4370 1827.4770     Tech
## 10 FirmE 2019    2252  247.3720  Finance
## 11 FirmA 2020    2267  659.9654     Tech
## 12 FirmB 2020    2037  821.6928   Health
## 13 FirmC 2020    4445  829.2733     Tech
## 14 FirmD 2020    1664  378.1813     Tech
## 15 FirmE 2020    3649  702.9810  Finance
## 16 FirmA 2021    2626  412.5971     Tech
## 17 FirmB 2021    4839 1286.2959   Health
## 18 FirmC 2021    3756  997.1794     Tech
## 19 FirmD 2021    2010  497.5517  Finance
## 20 FirmE 2021    2114  340.3073     Tech
## 21 FirmA 2022    1952  303.5798  Finance
## 22 FirmB 2022    1347  260.2588   Health
## 23 FirmC 2022    2016  577.3521  Finance
## 24 FirmD 2022    4720  974.1563   Health
## 25 FirmE 2022    3012 1334.7108   Health
```

Variable creation

Create a new variable “costs”, which is the revenue - profit. [There are many ways to create a variable in a data frame, which we’ll learn later in the course. Here, use the `$` index.]

Factor variable

Which variable is (should be) a factor? Recode this variable as a factor. What are the levels? Should we have the variable `Firm` as a factor?

Nests

Split your data using the factor variable into three data frames that are contained in a list. Compute the mean profit for each data frame. Hint: use the function `split` .

Advanced: map (not exam relevant)

Do the same as the exercise above using the `map` function. Install the packages `tidyr` , `dplyr` , and `purrr` .

```
# Or (advanced!) with a nested tibble and map
library(tidyr)
library(dplyr)
library(purrr)

tibble_financial_data <- financial_data |>
  group_by(Category) |>
  nest()

map(tibble_financial_data$data, ~mean(.$Profit))
```

```
## [[1]]
## [1] 963.7105
##
## [[2]]
## [1] 538.9418
##
## [[3]]
## [1] 882.3213
```