



# **LAB -4**

**Data Cleaning**

***ID :442801472***

***Name: Wed Ghanem***

2024/11/03

## Screenshot:

## Step 1:

The screenshot displays the RStudio interface with the 'banking' dataset loaded into the 'Data\_Frame' environment. The top panel shows the data table with columns Date, Income, and Expenditure. The bottom panel shows the R console output for 'describe(banking)' and 'head(banking)'.

**Environment Panel:**

- Data:**
  - banking:** 18 obs. of 3 variables
  - Data\_Frame:** 3 obs. of 3 variables
    - \$ Training: chr "wed" "Fatima" "mona"
    - \$ Pulse : num 100 150 120
    - \$ Duration: num 60 30 45

**Files Panel:**

- Install:** Name, Description, Ver
- User Library:**
  - askpass: Password Entry Utilities for R, Git, and SSH 1.2
  - backports: Reimplementations of Functions Introduced Since R-3.0.0 1.5
  - base64enc: Tools for base64 encoding 0.1
  - bit: Classes and Methods for Fast Memory-Efficient Boolean Selections 4.0
  - bit64: A S3 Class for Vectors of 64bit Integers 4.0
  - blob: A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2
  - bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown' 0.8
  - cachem: Cache R Objects with Automatic Pruning 1.1
  - callr: Call R from R 3.7
  - cellranger: Translate Spreadsheet Cell Ranges to Rows and Columns 1.1
  - checkmate: Fast and Versatile Argument Checks 2.3
  - cli: Helpers for Developing Command Line Interfaces 3.6
  - clipr: Read and Write from the System Clipboard 0.8
  - colorspace: A Toolbox for Manipulating and Assessing Colors 2.1

**R Console Output:**

```

R 4.4.1 ~ /
The following objects are masked from 'package:base':
  format.pval, units

> describe(banking)
banking
3 variables      18 observations
-----
Date
  n missing distinct      Info      Mean      pMedian      Gmd      .05
18      0         17 0.999 2010-07-14 1.279e+09 3431153 2010-06-02
 .10     .25     .50     .75     .90     .95
2010-06-03 2010-06-05 2010-07-08 2010-08-17 2010-08-24 2010-08-26

lowest : 2010-06-01 2010-06-02 2010-06-03 2010-06-04 2010-06-05
highest: 2010-08-12 2010-08-13 2010-08-14 2010-08-23 2010-08-25 2010-08-26

value      -1443 -2919 -3000 -6340  0 12330 3000  NA
frequency    1    2    2    2    7    1    1    2
proportion 0.056 0.111 0.111 0.111 0.389 0.056 0.056 0.111

> head(banking)
# A tibble: 6 x 3
  Date           Income Expenditure
<dtm>          <chr>    <chr>
1 2010-06-01 00:00:00 2523      0
2 2010-06-02 00:00:00 0         -2919
3 2010-06-03 00:00:00 0         -6340
4 2010-06-04 00:00:00 5803      0
5 2010-06-05 00:00:00 5338      0
6 2010-06-06 00:00:00 0         -3000
  
```

## Step2:

The screenshot shows the RStudio interface with the following components:

- Top Panel (Data Viewer):** Displays a data frame with columns Date, Income, and Expenditure. The data is sorted by Date, showing entries from 2010-06-01 to 2010-08-12.
- Bottom Panel (R Console):** Shows the execution of the following R code:
 

```
> str(banking)
tibble [18 x 3] (S3: tbl_df/tbl/data.frame)
 $ Date      : POSIXct[1:18], format: "2010-06-01" "2010-06-02" "2010-06-03" ...
 $ Income    : chr [1:18] "2523" "0" "0" "5803" ...
 $ Expenditure: chr [1:18] "0" "-2919" "-6340" "0" ...
> tail(banking)
# A tibble: 6 x 3
  Date           Income Expenditure
  <dtm>         <chr>    <chr>
1 2010-08-12 00:00:00 2000      0
2 2010-08-19 00:00:00 0         -1443
3 2010-08-19 00:00:00 NA        12330
4 2010-08-23 00:00:00 4000      NA
5 2010-08-25 00:00:00 NA         NA
6 2010-08-29 00:00:00 NA         3000
> View(banking)
```
- Right Panel (Environment and Help):**
  - Environment:** Shows the data objects 'banking' (18 obs. of 3 variables) and 'Data\_Frame' (3 obs. of 3 variables).
  - Help:** Displays the documentation for the 'dplyr' package version 1.1.4, including links to the DESCRIPTION file, user guides, vignettes, and package news.

The second screenshot shows the RStudio interface after running the following R code:

```
src, summarize

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

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

> library(ggplot2)
> library(dplyr)
> banking %>%
  + length()
[1] 18
```

The right panel shows the same documentation for 'dplyr' version 1.1.4, but with the 'Help Pages' section visible at the bottom, listing the alphabetically ordered help pages: ABCDEFGIJKLMNPQRSTUWV.

The screenshot shows the RStudio interface. On the left, the 'Data Frame' tab displays a table with 18 rows and 3 columns: Date, Income, and Expenditure. The 'Date' column contains dates from 2010-06-06 to 2010-08-29. The 'Income' column contains values from 0 to 5338, with some NA values. The 'Expenditure' column contains values from -3000 to 3000, with some NA values. Below the table, the R console shows the following commands and output:

```
R 4.4.1 ~ /
> banking$Date %>%
+ length()
[1] 18
> banking$Date %>%
+ unique()
[1] "2010-06-01 UTC" "2010-06-02 UTC" "2010-06-03 UTC" "2010-06-04 UTC" "2010-06-05 UTC"
[6] "2010-06-06 UTC" "2010-07-03 UTC" "2010-07-05 UTC" "2010-07-07 UTC" "2010-07-09 UTC"
[11] "2010-08-05 UTC" "2010-08-07 UTC" "2010-08-12 UTC" "2010-08-19 UTC" "2010-08-23 UTC"
[16] "2010-08-25 UTC" "2010-08-29 UTC"
> banking$Date %>%
+ length()
[1] 18
> banking$Date %>%
+ unique() %>%
+ length()
[1] 17
```

On the right, the 'Environment' pane shows the objects 'banking' (18 obs. of 3 variables) and 'Data\_Frame' (3 obs. of 3 variables). Below it, the 'R: A Grammar of Data Manipulation' pane displays the title 'A Grammar of Data Manipulation' and the subtitle 'Documentation for package 'dplyr' version 1.1.4'. It also includes links for 'DESCRIPTION file', 'User guides, package vignettes and other documentation', and 'Package NEWS'.

The screenshot shows the RStudio interface. On the left, the 'Data Frame' tab displays a table with 18 rows and 3 columns: Date, Income, and Expenditure. The 'Date' column contains dates from 2010-06-06 to 2010-08-29. The 'Income' column contains values from 0 to 5338, with some NA values. The 'Expenditure' column contains values from -3000 to 3000, with some NA values. Below the table, the R console shows the following commands and output:

```
R 4.4.1 ~ /
> str(banking)
tibble [18 x 3] (S3: tbl_df/tbl/data.frame)
 $ Date      : POSIXct[1:18], format: "2010-06-01" "2010-06-02" "2010-06-03" ...
 $ Income    : chr [1:18] "2523" "0" "0" "5803" ...
 $ Expenditure: chr [1:18] "0" "-2919" "-6340" "0" ...
```

On the right, the 'Environment' pane shows the objects 'banking' (18 obs. of 3 variables) and 'Data\_Frame' (3 obs. of 3 variables). Below it, the 'R: A Grammar of Data Manipulation' pane displays the title 'A Grammar of Data Manipulation' and the subtitle 'Documentation for package 'dplyr' version 1.1.4'. It also includes links for 'DESCRIPTION file' and 'User guides, package vignettes and other documentation'.