R for Data Science

Center for Health Data Science

June 2025

Who are we?

Center for Health Data Science

- Data Science Research Groups
- HDS Sandbox
- SUND DataLab

https://heads.ku.dk/

Thilde Terkelsen, Ph.D.

Helene Wegener

SUND DataLab

- Courses & Workshops
- Consultations
- Commission Research & Supervision
- Events, Seminars

What will I Learn?

Program

	Day 1	Day 2	Day 3
08:30-09:00	Installation Issues + Coffee	Optional Q&A + Coffee	Optional Q&A + Coffee
09:00-09:15	Introduction to course		
09:15-10:00	Presentation 1: Base R	Functions, for-loops,	
10:00-10:45	Exercise 1: Base R		
10:45-11:00	Break	Break	Break
11:00-12:00	Presentation 2: Data Wrangling		
12:00-13:00	Lunch	Lunch	Lunch
13:00-14:30	Exercise 2: Data Wrangling		
14:30-14:45	Break	Break	Project work
14:45-15:15	Presentation 3: Advanced ggplot2		
15:15-16:00	Exercise 3: Advanced ggplot2		
			Wrap up
16:00	See you tomorrow!	See you tomorrow!	Bye Bye!

Program

+ 2*15 minut break

Day 1		Day 2		Day 3	
08:30-09:00	Installation Issues + Coffee	08:30-09:00	Optional Q&A + Coffee	08:30-09:00	Optional Q&A + Coffee
09:00-09:15	Introduction to course	09:00-09:15	Functions, for-loops, if-else statements	09:00-09:15	
09:15-10:00	Presentation 1: Data Clean-Up in Base R and Tidyverse	09:15-10:00		09:15-10:00	
10:00-11:15	Exercise 1: Data Clean-Up	10:00-10:45		10:00-10:45	
11:15-12:00	Presentation 2: Advanced Data Wrangling	11:00-12:00		11:00-12:00	
12:00-13:00	Lunch	12:00-13:00	Lunch	12:00-13:00	Lunch
13:00-14:45	Exercise 2: Advanced Data Wrangling	13:00-14:30		13:00-16:00	Project work
14:45-15:15	Presentation 3: Advanced ggplot2?	14:45-15:15			
15:15-16:00	Exercise 3: Advanced ggplot2?	15:15-16:00			
16:00	See you tomorrow!	16:00	See you tomorrow!	16:00	Wrap up + Bye Bye!

Part 0
R script and Quarto

R script

- Flat script
  - Submit to HPC
- Comment script and build structure using #
- Source

Quarto

- Markdown-based
- Render to get a nice report in html or website
- Headers and text

R script

Quarto

```
##########################################
# R for Data Science - How to R script #
# Author: DataLab HeaDS                  #
# Date: 8 Novmeber 2024                  #
##########################################

##################################################
################## Load Packages #################
##################################################
library(tidyverse)
library(readxl)

##################################################
################### Load Data ####################
##################################################
diabetes <- read_excel('~/Desktop/DataLab/R4DataScience/data/diabetes_toy.xlsx')

##################################################
################## Inspect Data ##################
##################################################

# Check dimensions of data
dim(diabetes)

# Check structure of data
str(diabetes)

# Check for NA's in each column
colSums(is.na(diabetes))

##################################################
############ Exploratory Data Analysis ###########
##################################################

# Plot distribution of BMI
diabetes %>%
  ggplot(aes(x = BMI)) +
  geom_histogram(bins = 10)
```

```
---
title: "R for Data Science - How to R script"
format: html
author: DataLab HeaDS
editor: visual
---
```

Load Packages

```{r}
library(tidyverse)
library(readxl)
```

Load Data

```{r}
diabetes <- read_excel('~/Desktop/DataLab/R4DataScience/data/diabetes_toy.xlsx')
```

Inspect Data

Check dimensions of data

```{r}
dim(diabetes)
```

Check structure of data

```{r}
str(diabetes)
```

Check for NA's in each column

```{r}
colSums(is.na(diabetes))
```

Exploratory Data Analysis

Plot distribution of BMI

```{r}
diabetes %>%
  ggplot(aes(x = BMI)) +
  geom_histogram(bins = 10)
```

Render to html

R for Data Science - How to Quarto

AUTHOR
DataLab HeaDS

Load Packages

```
library(tidyverse)
library(readxl)
```

Load Data

```
diabetes <- read_excel('~/Desktop/DataLab/R4DataScience/data/diabet
```

Inspect Data

Check dimensions of data

```
dim(diabetes)
```

```
[1] 100   8
```

Check structure of data

```
str(diabetes)
```

```
tibble [100 × 8] (S3: tbl_df/tbl/data.frame)
 $ ID                : chr [1:100] "ID_035" "ID_020" "ID_090" "ID_
 $ Fasting_Blood_Sugar  : num [1:100] 157 146 128 134 157 101 144 14
 $ Post_Meal_Blood_Sugar: num [1:100] 204 178 177 206 206 250 204 18
 $ HbA1c             : num [1:100] 5.8 6.3 6.6 8 6.9 7.4 6.9 5.9 
 $ Age               : num [1:100] 36 70 63 49 56 39 59 43 57 56 .
 $ Sex               : chr [1:100] "Male" "Male" "Male" "Male" ...
 $ BMI               : num [1:100] 33.7 26.5 26 22.9 25.7 24.4 25.
 $ Blood_Pressure    : num [1:100] 122 120 137 126 115 121 116 11
```

Check for NA's in each column

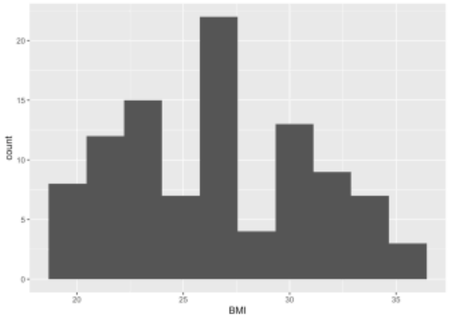
```
colSums(is.na(diabetes))
```

```
           ID  Fasting_Blood_Sugar Post_Meal_Blood_Sugar 
            0                     0                     0 
        HbA1c                   Age                   Sex 
            0                     0                     0 
          BMI        Blood_Pressure 
            0                     0 
```

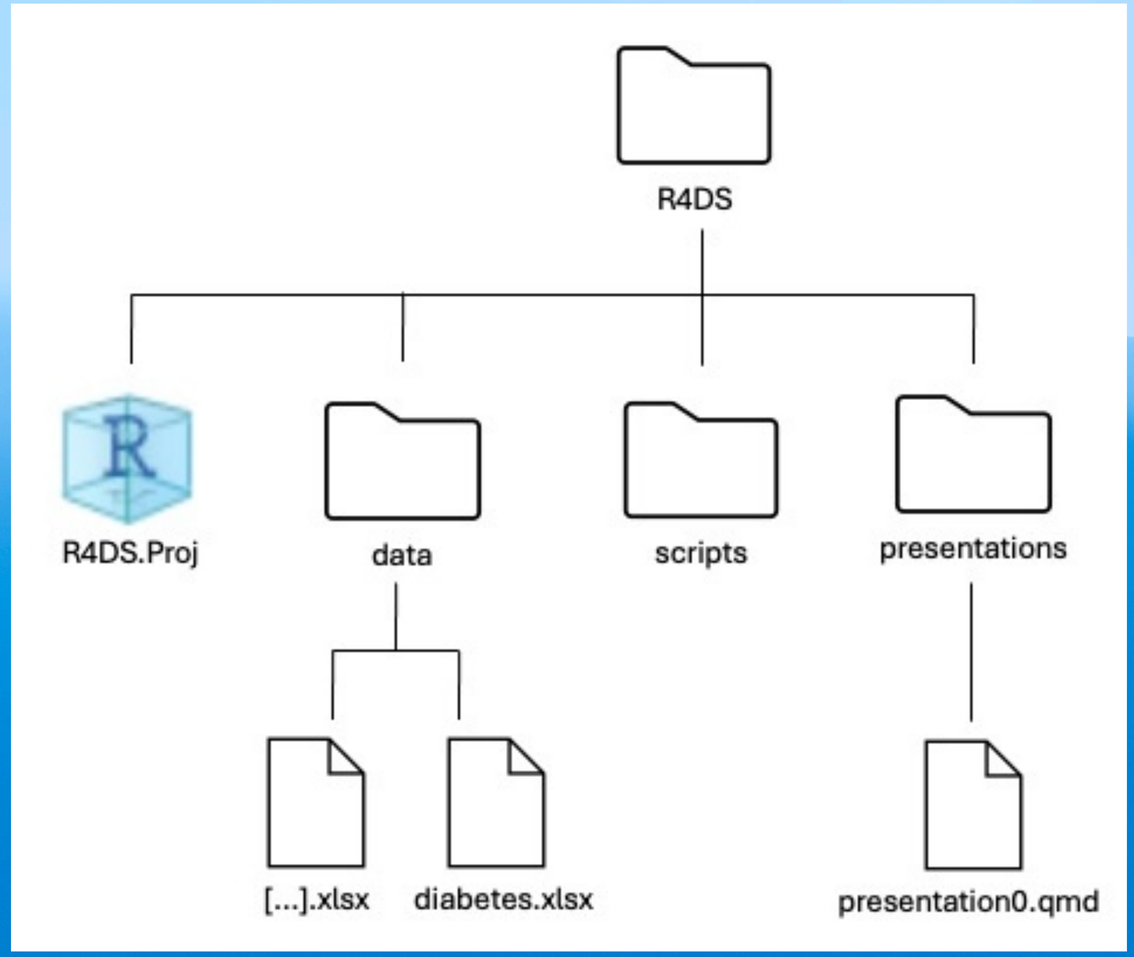
Exploratory Data Analysis

Plot distribution of BMI

```
diabetes %>%
  ggplot(aes(x = BMI)) +
  geom_histogram(bins = 10)
```



R project


R4DS
R4DS.Proj
data
scripts
presentations
[...].xlsx
diabetes.xlsx
presentation0.qmd

Exercise 0
R script and Quarto

Part 1
Base R and Tidyverse

Base R

- Basic library that is pre-installed in R
- Supports older versions of R
- Efficient for small tasks
- Complex workflows are doable, but syntax can be convoluted
  - Specify data frame multiple times

Tidyverse

- A collection of packages such as
  - dplyr (data manipulation, %>%)
  - ggplot2 (visualization)
  - tidyr (reshaping)
  - readr (reading data)
  - stringr (string manipulation)
- Moderen and intuitive syntax, even for complex workflows
- Assumes tidy data (each column is a variable, and each row is an observation)

Cheat Sheet - Base R and Tidyverse

	Base R	Tidyverse
Select columns	`df[, c("col1", "col2")]`	`df %>% select(col1, col2)`
Access one column	`df$col1` `df[["col1"]]`	`df %>% pull(col1)`
Add new column	`df$col_new <- list_new`	`df <- df %>%` `mutate(col_new = list_new)`
Filter rows	`df[df$col1 < 10 ,]`	`df %>% filter(col1 < 10)`
Remove rows with NA's	`df[complete.cases(df_baseR), ]`	`df %>% drop_na()`

Exercise 1
Base R and Tidyverse

Part 2
Advanced Tidyverse

Long and wide format

```
tree_long <- tree_wide %>%
  pivot_longer(cols = starts_with("Site"),
               names_to = "Site",
               values_to = "Average diameter (cm)")
```

Type species	Site A	Site B	Site C	Site D
Acer rubrum	15	8	30	27
Quercus alba	29	17	14	42
Pinus teada	10	19	25	23

```
tree_wide <- tree_long %>%
  pivot_wider(names_from = Site,
              values_from = `Average diameter (cm)`)
```

Tree species	Site	Average diameter (cm)
Acer rubrum	Site A	15
Acer rubrum	Site B	8
Acer rubrum	Site C	30
Acer rubrum	Site D	27
Quercus alba	Site A	29
Quercus alba	Site B	17
Quercus alba	Site C	14
Quercus alba	Site D	42
Pinus teada	Site A	10
Pinus teada	Site B	19
Pinus teada	Site C	25
Pinus teada	Site D	23

Nesting

```
tree_long_nested <- tree_long %>%
  group_by(`Type species`) %>%
  nest(Data = c(Site, `Average diameter (cm)`)) %>%
  ungroup()
```

Tree species	Site	Average diameter (cm)
Acer rubrum	Site A	15
Acer rubrum	Site B	8
Acer rubrum	Site C	30
Acer rubrum	Site D	27
Quercus alba	Site A	29
Quercus alba	Site B	17
Quercus alba	Site C	14
Quercus alba	Site D	42
Pinus teada	Site A	10
Pinus teada	Site B	19
Pinus teada	Site C	25
Pinus teada	Site D	23

Type species	Data
Acer rubrum	<tibble>
Quercus alba	<tibble>
Pinus teada	<tibble>

```
tree_long_nested %>%
  filter(`Type species` == 'Quercus alba') %>%
  pull(Data)
```

Site	Average diameter (cm)
Site A	29
Site B	17
Site C	14
Site D	42

Exercise 2
Advanced Tidyverse

See you tomorrow!

Part 3
Functions

Functions

If-statements

Loops