Statistical Tests in R: Descriptive Analysis

Kursus R: Pengenalan dan Praktikal (Sesi 3)

Mohd Azmi Bin Suliman

azmi.suliman@moh.gov.my

Pusat Penyelidikan Penyakit Tak Berjangkit, Institut Kesihatan Umum

Friday, 04 October 2024

Statistical Tests in R: Descriptive Analysis

Descriptive Analysis

- Descriptive analysis refers to summarising and describing the main features of a dataset.
- Count and percentage for categorical data.
- Mean, median, standard deviation, and range for numerical data.
- Data visualisation is also part of descriptive analysis.

Descriptive Analysis

- There are many ways to perform descriptive analysis in R.
- We can use the base function, or commonly, people use the tidyverse package for descriptive analysis.
- In this session, I will introduce `summarytools` and `gtsummary` packages for descriptive analysis.
  - → Easier to use
  - → More informative.
  - → Nicer outcome (especially when rendering the Quarto document!)

Let’s Try!

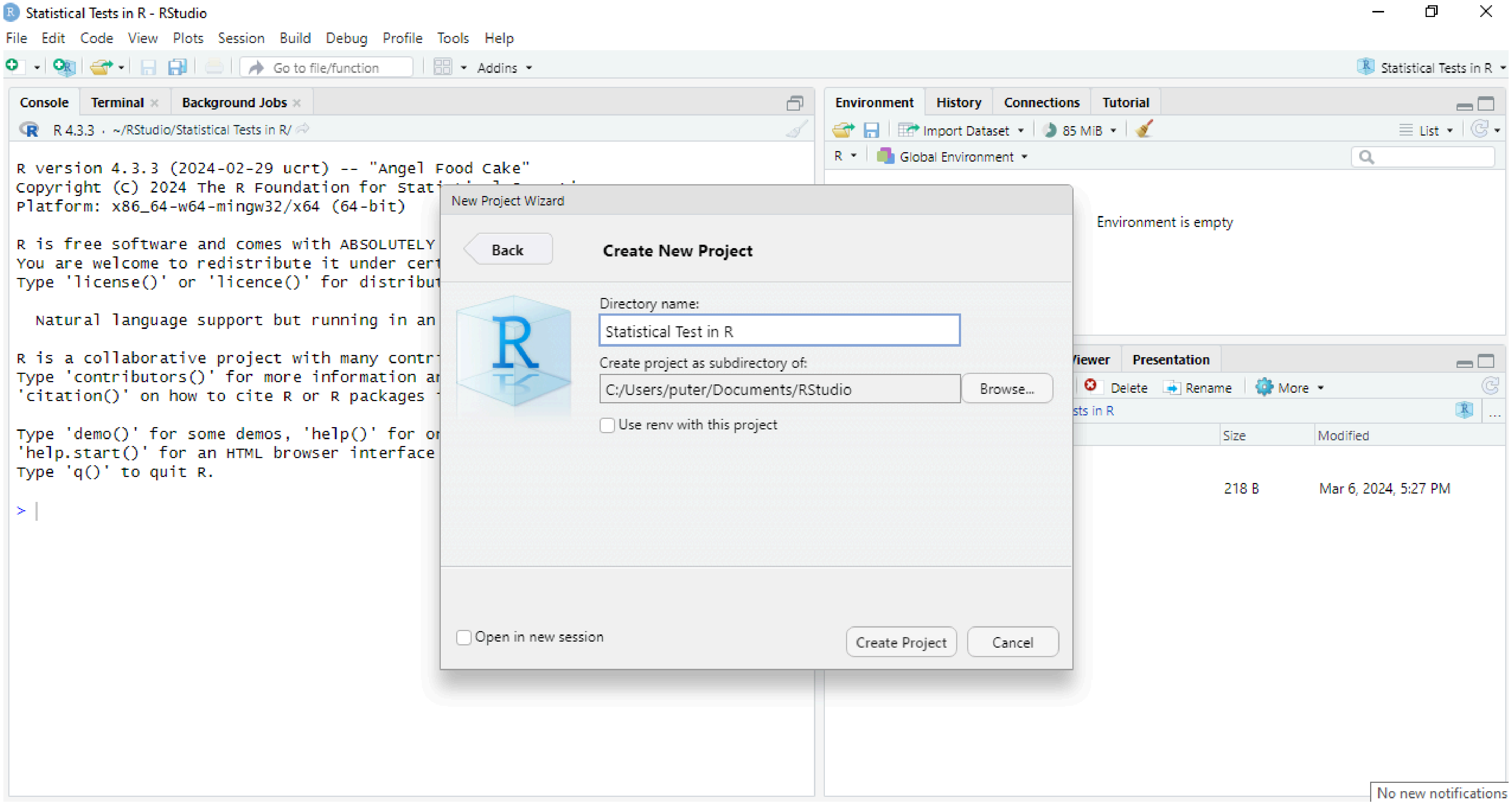
Setup your project & quarto document.

Create New Project

- Remember last week? RStudio allows for project management.
  - → Project as a ‘container’ for our work.

1. Open RStudio.
2. Create a new project.
   - `File` > `New Project` > `New Directory` > `New Project`
3. Set the name and directory.
   - Name: `Statistical Tests in R`
   - Directory: `.../Documents/RStudio`
4. Click `Create Project`

Create New Project



```
R version 4.3.3 (2024-02-29 ucrt) -- "Angel Food Cake"
Copyright (C) 2024 The R Foundation for Stati
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY
You are welcome to redistribute it under cert
Type 'license()' or 'licence()' for distribut

  Natural language support but running in an

R is a collaborative project with many contr
Type 'contributors()' for more information an
'citation()' on how to cite R or R packages

Type 'demo()' for some demos, 'help()' for or
'help.start()' for an HTML browser interface
Type 'q()' to quit R.

> 
```

Directory name: Statistical Test in R

Create project as subdirectory of: C:/Users/puter/Documents/RStudio

☐ Use renv with this project

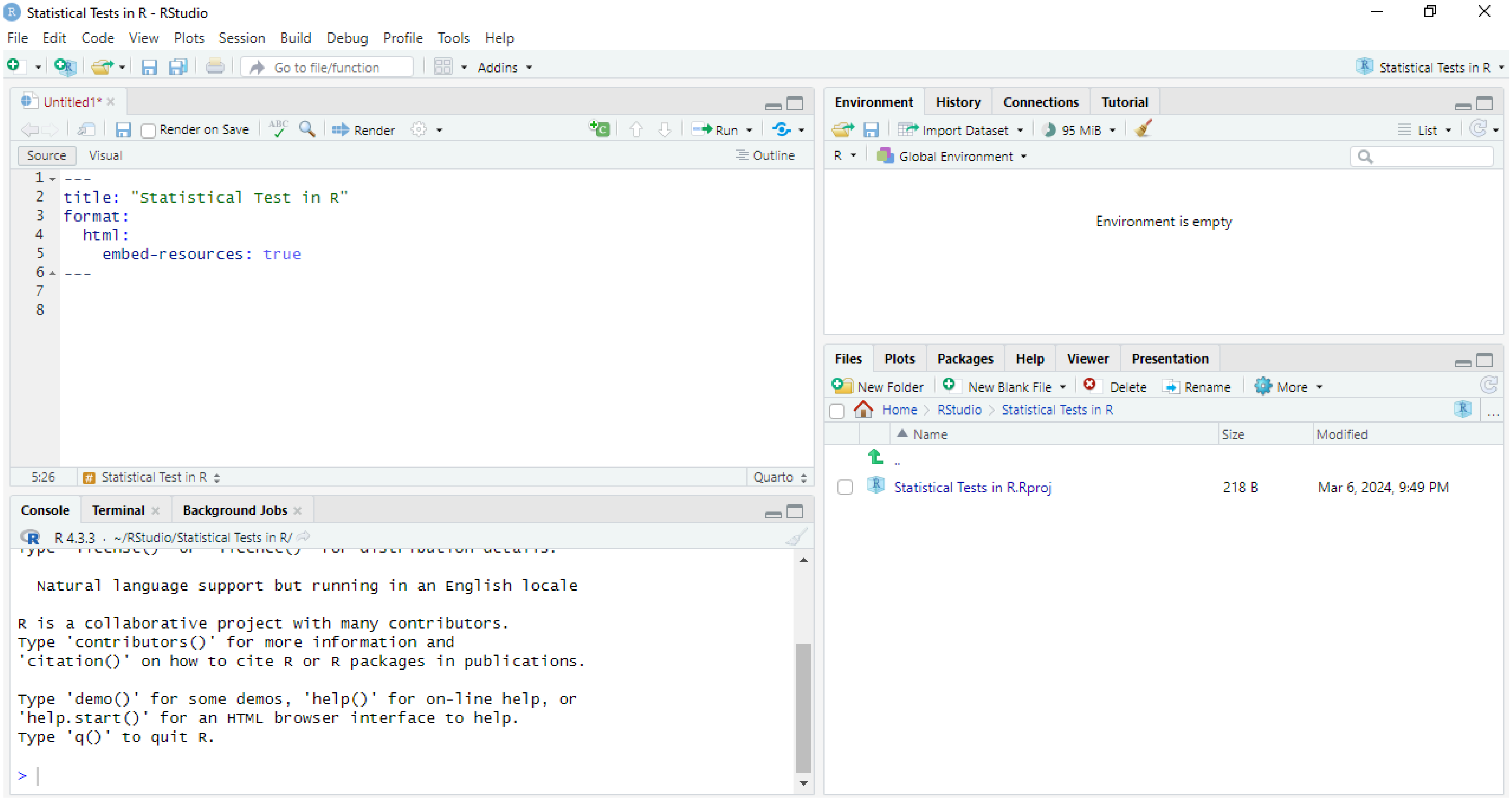
☐ Open in new session

Create New Quarto Document

Quarto as R Notebook.

1. Create a new Quarto document.
   - `File` > `New File` > `Quarto Document...`
2. Set the title
   - Title: `Statistical Tests in R`
   - Untick `Use the visual editor.`
3. Click `Create Empty Document`
4. Edit the YAML
   - Add the `embed-resources: true` parameter.

Create a New Quarto Document



```
---
title: "Statistical Test in R"
format:
  html:
    embed-resources: true
---
```

```
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Name	Size	Modified
Statistical Tests in R.Rproj	218 B	Mar 6, 2024, 9:49 PM

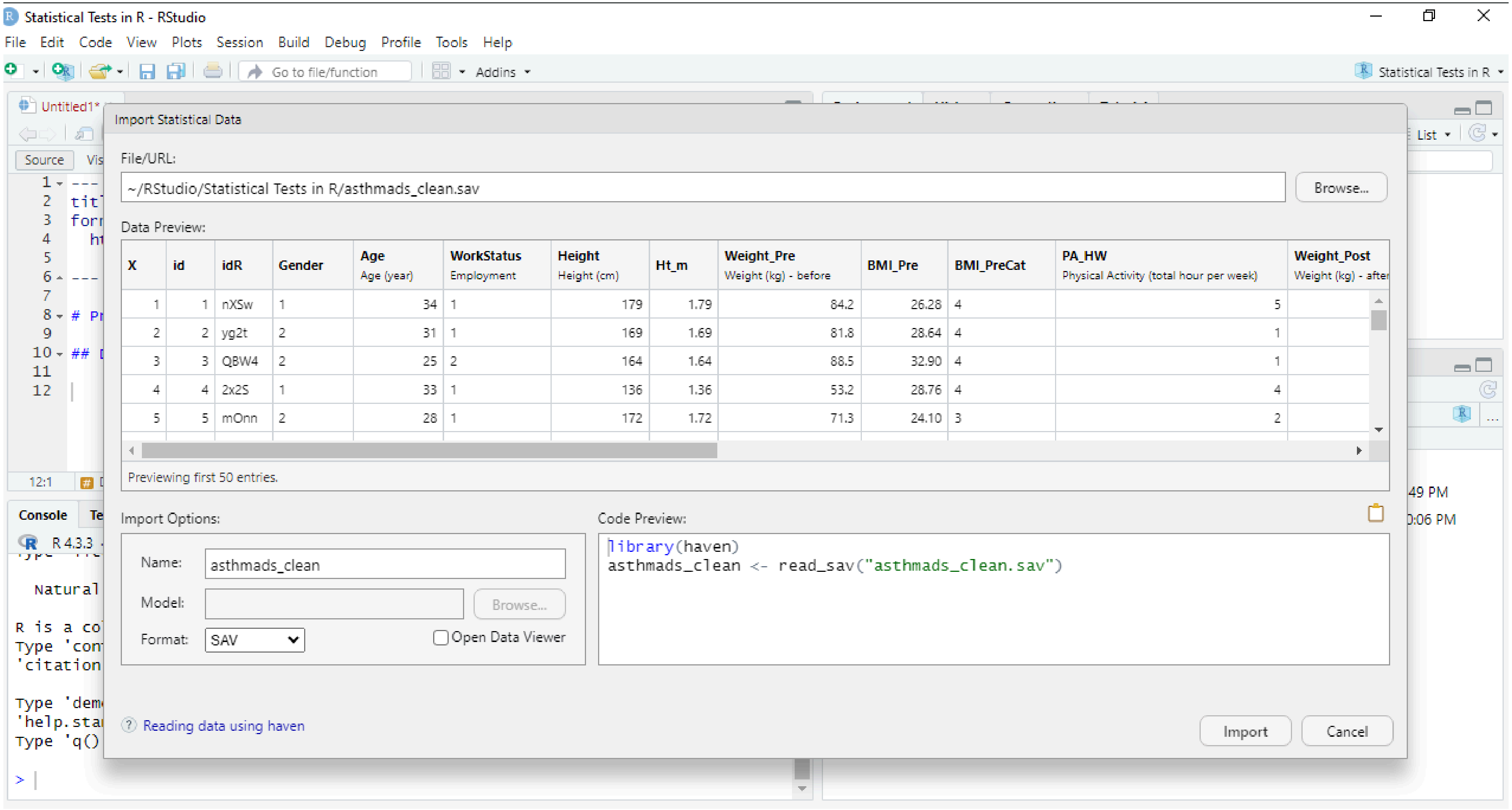
Import Dataset

1. Copy the `asthmads_clean.sav` dataset into the working directory.
   - Download the dataset from the Google Drive folder.
   - Open the folder to which the dataset was downloaded.
   - Copy the dataset.
   - Open your working directory.
   - `File` pane > ⚙ `More` > `Show Folder in New Window`
   - Paste the dataset into the working directory.
2. In the Quarto document, add a new level 1 header & level 2 header
   - `# Preparation`
   - `## Data Import`

Import Dataset

3. Import the dataset. We will use `asthmads_clean.sav` dataset
    - `File` > `Import Dataset` > `From SPSS...`
    - Select the dataset `asthmads_clean.sav`
    - Click `Open`
    - Untick `Open Data Viewer`
    - Click the clipboard symbol 📋

Import Dataset



File/URL: ~/RStudio/Statistical Tests in R/asthmads_clean.sav

Data Preview:

X	id	idR	Gender	Age Age (year)	WorkStatus Employment	Height Height (cm)	Ht_m	Weight_Pre Weight (kg) - before	BMI_Pre	BMI_PreCat	PA_HW Physical Activity (total hour per week)	Weight_Post Weight (kg) - after
1	1	nXSw	1	34	1	179	1.79	84.2	26.28	4	5	
2	2	yg2t	2	31	1	169	1.69	81.8	28.64	4	1	
3	3	QBW4	2	25	2	164	1.64	88.5	32.90	4	1	
4	4	2x2S	1	33	1	136	1.36	53.2	28.76	4	4	
5	5	mOnn	2	28	1	172	1.72	71.3	24.10	3	2	

Previewing first 50 entries.

Import Options:

Name: asthmads_clean

Model:

Format: SAV

☐ Open Data Viewer

Code Preview:

```
library(haven)
asthmads_clean <- read_sav("asthmads_clean.sav")
```

Reading data using haven

Import Dataset

4. Paste the code into the r code chunk
5. Use the `as_factor` function to read the label and apply to the dataset
    - We can use %>% pipe operator to use `as_factor` function
    - Don’t forget to load the `tidyverse` package first

```r
library(tidyverse)
library(haven)
asthmads_clean <- read_sav("Dataset/asthmads_clean.sav") %>%
  as_factor()
```

Import Dataset

```r
asthmads_clean
```

```
# A tibble: 150 × 24
       X    id idR   Gender   Age WorkStatus Height  Ht_m Weight_Pre BMI_Pre
   <dbl> <dbl> <chr> <fct>  <dbl> <fct>       <dbl> <dbl>      <dbl>   <dbl>
 1     1     1 nXSw  Female    34 Unemployed    179  1.79       84.2    26.3
 2     2     2 yg2t  Male      31 Unemployed    169  1.69       81.8    28.6
 3     3     3 QBW4  Male      25 Employed      164  1.64       88.5    32.9
 4     4     4 2x2S  Female    33 Unemployed    136  1.36       53.2    28.8
 5     5     5 mOnn  Male      28 Unemployed    172  1.72       71.3    24.1
 6     6     6 D3sl  Male      33 Unemployed    178  1.78       87.3    27.6
 7     7     7 le6j  Female    31 Unemployed    140  1.4        48.8    24.9
 8     8     8 r3gC  Female    34 Employed      140  1.4        49.1    25.0
 9     9     9 3Tyt  Male      31 Employed      171  1.71       60.1    20.6
10    10    10 cmKF  Male      28 Employed      163  1.63       93.1    35.0
# ℹ 140 more rows
# ℹ 14 more variables: BMI_PreCat <fct>, PA_HW <dbl>, Weight_Post <dbl>,
#   BMI_Post <dbl>, BMI_PostCat <fct>, Tx1 <fct>, Tx2 <fct>, PEFR_Pre <dbl>,
#   PEFR_Post <dbl>, PEFR_Diff <dbl>, SxWheeze_Pre <fct>, SxWheeze_Post <fct>,
#   PS_Pre <dbl>, PS_Post <dbl>
```

Let’s Try!

Descriptive Analysis with gtsummary

Descriptive Analysis with `gtsummary`

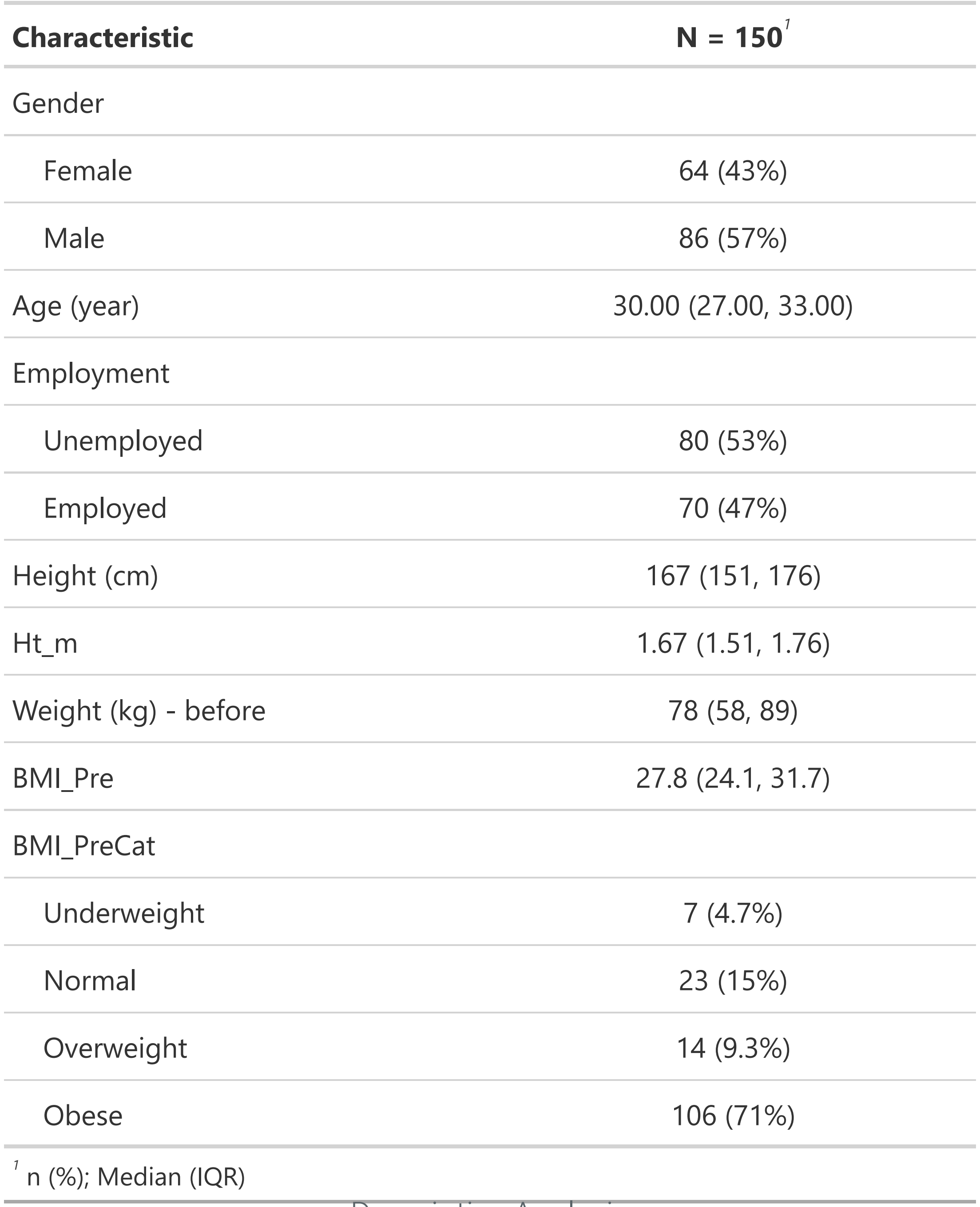
- The primary function of Quarto is for publishing.
- There are various package that can help us to explore our data
- However, if we want to publish the document, we need a better-looking table.
- The `gtsummary` package is an excellent package for this purpose.

```r
library(gtsummary)
```

Descriptive Analysis with `gtsummary`

- using the `tbl_summary` function to summarise the data.
- Note: *The variable column (Characteristic) will automatically read the variable's label, if available.*

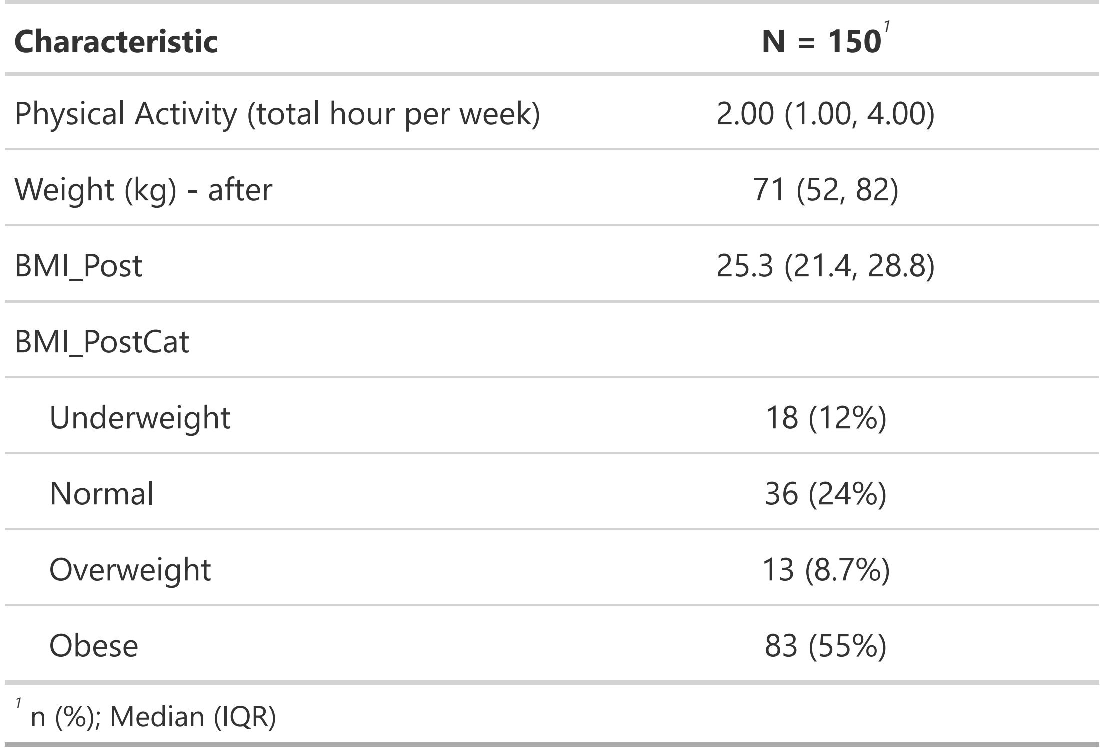
```r
asthmads_clean %>%
  select(Gender:BMI_PostCat) %>%
  tbl_summary()
```



Characteristic	N = 150[1]
Gender	
Female	64 (43%)
Male	86 (57%)
Age (year)	30.00 (27.00, 33.00)
Employment	
Unemployed	80 (53%)
Employed	70 (47%)
Height (cm)	167 (151, 176)
Ht_m	1.67 (1.51, 1.76)
Weight (kg) - before	78 (58, 89)
BMI_Pre	27.8 (24.1, 31.7)
BMI_PreCat	
Underweight	7 (4.7%)
Normal	23 (15%)
Overweight	14 (9.3%)
Obese	106 (71%)

[1] n (%); Median (IQR)

Descriptive Analysis



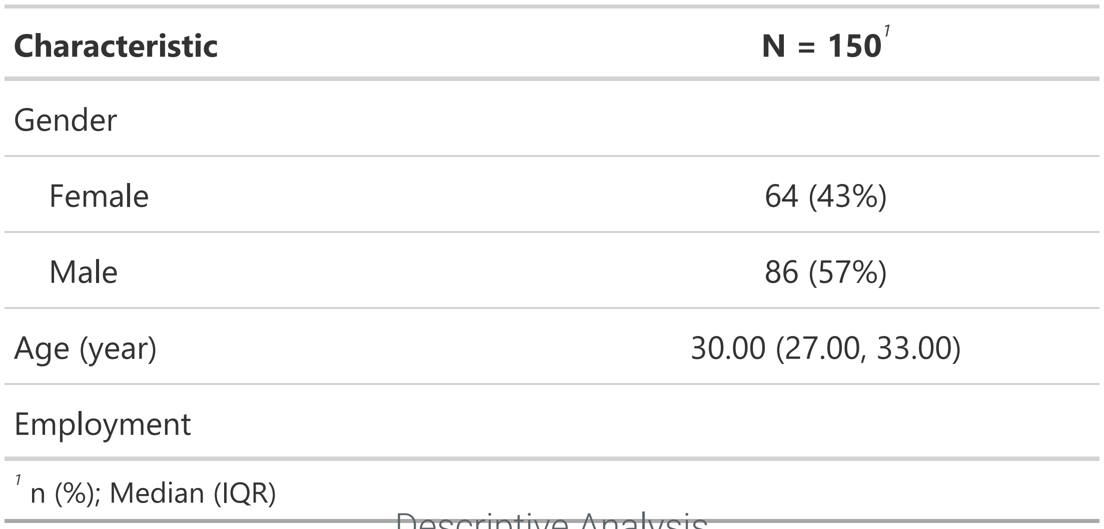
Characteristic	**N = 150**[1]
Physical Activity (total hour per week)	2.00 (1.00, 4.00)
Weight (kg) - after	71 (52, 82)
BMI_Post	25.3 (21.4, 28.8)
BMI_PostCat	
Underweight	18 (12%)
Normal	36 (24%)
Overweight	13 (8.7%)
Obese	83 (55%)

[1] n (%); Median (IQR)

Descriptive Analysis with `gtsummary`

- we can customise it.
  - → change the variable label using the `label =` parameter

```r
asthmads_clean %>%
  select(Gender:BMI_PostCat) %>%
  tbl_summary(label = list(Ht_m = "Height (m)",
                           BMI_Pre = "BMI (Pre)",
                           BMI_PreCat = "BMI Category (Pre)",
                           BMI_Post = "BMI (Post)",
                           BMI_PostCat = "BMI Category (Post)"))
```



Characteristic	**N = 150**[1]
Gender	
Female	64 (43%)
Male	86 (57%)
Age (year)	30.00 (27.00, 33.00)
Employment	

[1] n (%); Median (IQR)

Characteristic	N = 150[1]
Unemployed	80 (53%)
Employed	70 (47%)
Height (cm)	167 (151, 176)
Height (m)	1.67 (1.51, 1.76)
Weight (kg) - before	78 (58, 89)
BMI (Pre)	27.8 (24.1, 31.7)
BMI Category (Pre)	
Underweight	7 (4.7%)
Normal	23 (15%)
Overweight	14 (9.3%)
Obese	106 (71%)
Physical Activity (total hour per week)	2.00 (1.00, 4.00)
Weight (kg) - after	71 (52, 82)
BMI (Post)	25.3 (21.4, 28.8)
BMI Category (Post)	
Underweight	18 (12%)
Normal	36 (24%)
Overweight	13 (8.7%)

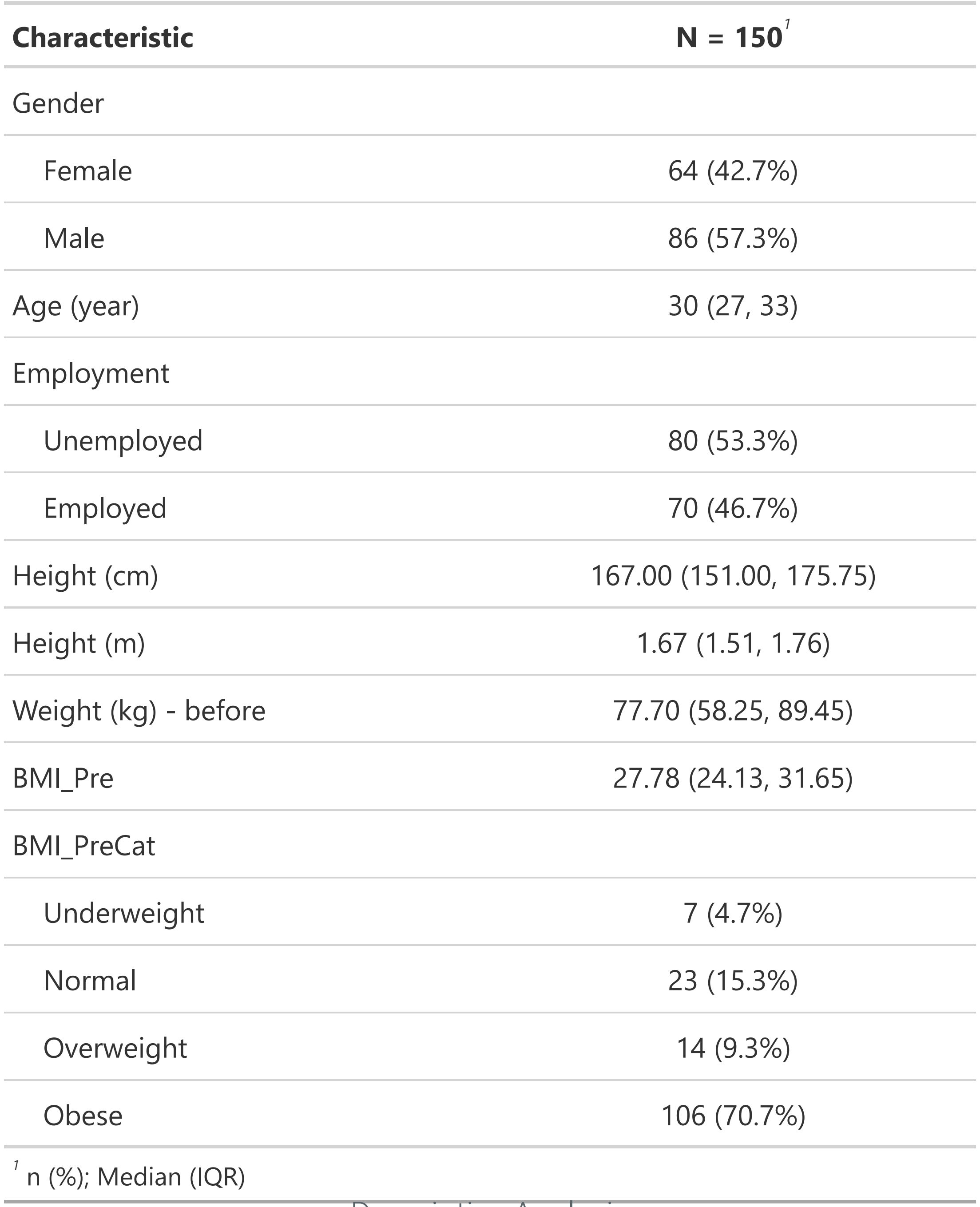
[1] n (%); Median (IQR)

Descriptive Analysis with `gtsummary`

- we can customise it.

  → change the digit using the `digits =` parameter

```
asthmads_clean %>%
  select(Gender:BMI_PostCat) %>%
  tbl_summary(label = list(Ht_m = "Height (m)"),
              digits = list(all_continuous() ~ 2,
                            all_categorical() ~ c(0,1),
                            Age ~ 0))
```



Characteristic	**N = 150**[1]
Gender	
Female	64 (42.7%)
Male	86 (57.3%)
Age (year)	30 (27, 33)
Employment	
Unemployed	80 (53.3%)
Employed	70 (46.7%)
Height (cm)	167.00 (151.00, 175.75)
Height (m)	1.67 (1.51, 1.76)
Weight (kg) - before	77.70 (58.25, 89.45)
BMI_Pre	27.78 (24.13, 31.65)
BMI_PreCat	
Underweight	7 (4.7%)
Normal	23 (15.3%)
Overweight	14 (9.3%)
Obese	106 (70.7%)

[1] n (%); Median (IQR)

Descriptive Analysis

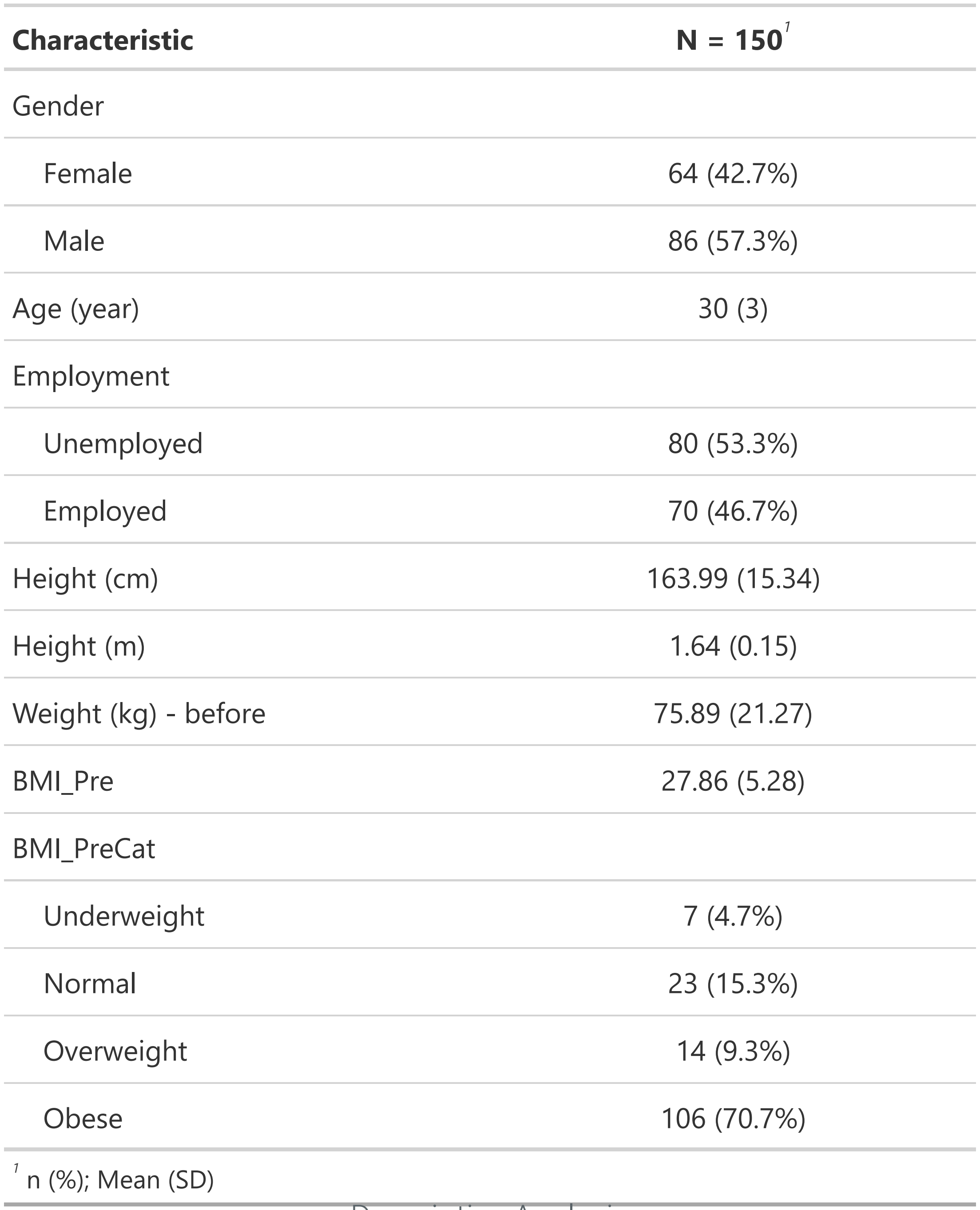
Characteristic	N = 150[1]
Physical Activity (total hour per week)	2.00 (1.00, 4.00)
Weight (kg) - after	70.75 (52.08, 82.08)
BMI_Post	25.31 (21.39, 28.77)
BMI_PostCat	
Underweight	18 (12.0%)
Normal	36 (24.0%)
Overweight	13 (8.7%)
Obese	83 (55.3%)

[1] n (%); Median (IQR)

Descriptive Analysis with `gtsummary`

- we can customise it.
  - → by default, the numerical variables are reported in Median (IQR)
  - → change it to Mean (SD) using the `stat =` parameter

```r
asthmads_clean %>%
  select(Gender:BMI_PostCat) %>%
  tbl_summary(label = list(Ht_m = "Height (m)"),
              digits = list(all_continuous() ~ 2,
                            all_categorical() ~ c(0,1),
                            Age ~ 0),
              stat = list(all_continuous() ~ "{mean} ({sd})"))
```



Characteristic	N = 150[1]
Gender	
Female	64 (42.7%)
Male	86 (57.3%)
Age (year)	30 (3)
Employment	
Unemployed	80 (53.3%)
Employed	70 (46.7%)
Height (cm)	163.99 (15.34)
Height (m)	1.64 (0.15)
Weight (kg) - before	75.89 (21.27)
BMI_Pre	27.86 (5.28)
BMI_PreCat	
Underweight	7 (4.7%)
Normal	23 (15.3%)
Overweight	14 (9.3%)
Obese	106 (70.7%)

[1] n (%); Mean (SD)

Descriptive Analysis

Characteristic	N = 150[1]
Physical Activity (total hour per week)	2.81 (2.27)
Weight (kg) - after	68.55 (20.46)
BMI_Post	25.14 (5.27)
BMI_PostCat	
Underweight	18 (12.0%)
Normal	36 (24.0%)
Overweight	13 (8.7%)
Obese	83 (55.3%)

[1] n (%); Mean (SD)

Grouped Summaries with `gtsummary`

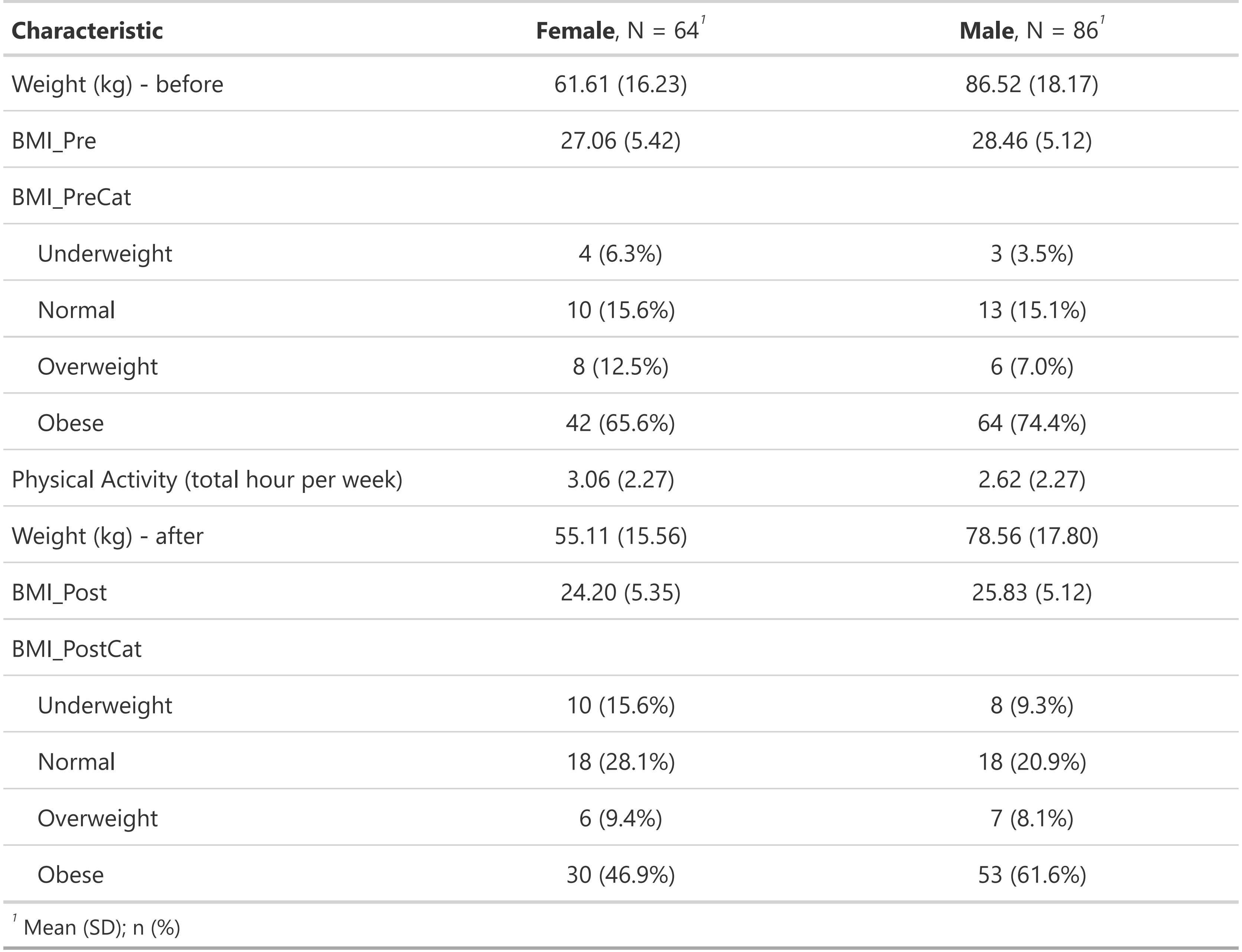
Grouped Summaries with gtsummary

- We can also perform grouped summaries using the by = parameter.

```r
asthmads_clean %>%
  select(Gender:BMI_PostCat) %>%
  tbl_summary(by = Gender,
              label = list(Ht_m = "Height (m)"),
              digits = list(all_continuous() ~ 2,
                            all_categorical() ~ c(0,1),
                            Age ~ 0),
              stat = list(all_continuous() ~ "{mean} ({sd})"))
```

Characteristic	**Female**, N = 64[1]	**Male**, N = 86[1]
Age (year)	31 (3)	30 (3)
Employment		
Unemployed	47 (73.4%)	33 (38.4%)
Employed	17 (26.6%)	53 (61.6%)
Height (cm)	150.39 (10.98)	174.10 (8.98)
Height (m)	1.50 (0.11)	1.74 (0.09)

[1] Mean (SD); n (%)



Characteristic	**Female**, N = 64[1]	**Male**, N = 86[1]
Weight (kg) - before	61.61 (16.23)	86.52 (18.17)
BMI_Pre	27.06 (5.42)	28.46 (5.12)
BMI_PreCat		
Underweight	4 (6.3%)	3 (3.5%)
Normal	10 (15.6%)	13 (15.1%)
Overweight	8 (12.5%)	6 (7.0%)
Obese	42 (65.6%)	64 (74.4%)
Physical Activity (total hour per week)	3.06 (2.27)	2.62 (2.27)
Weight (kg) - after	55.11 (15.56)	78.56 (17.80)
BMI_Post	24.20 (5.35)	25.83 (5.12)
BMI_PostCat		
Underweight	10 (15.6%)	8 (9.3%)
Normal	18 (28.1%)	18 (20.9%)
Overweight	6 (9.4%)	7 (8.1%)
Obese	30 (46.9%)	53 (61.6%)

[1] Mean (SD); n (%)