Exercise 3

No Fear of Numbers: Introduction to Quantitative Data Analysis in R

Dr. Susanne de Vogel, Data Science Center, University of Bremen

2 December 2025

Research Question: Is there a Life beyond the PhD?

In this workshop, we will work with one coherent example throughout all four parts.We focus on two outcomes of interest:

- `bpsy01`: overall life satisfaction (10 point scale)
- `blcd06`: Children (yes/no)

We will examine how these outcomes are related to different aspects of doctoral researchers' lives and backgrounds:

PhD and Work conditions

- `adbi01`: Status of the doctorate
- `adbi15`: Discipline
- `bdcd17`: Emotional Support during PhD
- `bwdr12`: Perceived Scientific Pressure
- `bemp81`: Monthly Gross income

Attitudes and Well-Being

- `bldc12`: Satisfaction with work-life balance
- `apar15b`: Relationship with parents
- `bpsy05`: Self-Efficacy

Demographics

- `adem01`: Gender
- `adem02`: Age in years
- `apar10`: Highest vocational degree of parents
- `adem03`: Country of birth
- `blcd01`: Relationship status

Bivariate Analysis

In the previous exercise, we focused on **univariate analysis** looking at one variable at a time. In this sheet, we move one step further to **bivariate analysis**: we look at **pairs of variables** to explore how they are related.

In the context of our research story, our main outcomes of interest are:

- general life satisfaction (`bpsy01`), and
- having children (yes/no, `blcd06`).

Before we build regression models, we first explore **how these outcomes relate to other key variables** (e.g. gender, age).

In this exercise, you will learn to:

- analyse **categorical × categorical** relationships using cross-tabulations, row/column percentages, chi-squared tests, and grouped bar charts,
- analyse **categorical × metric** relationships using group means, standard deviations, boxplots, t-tests and one-way ANOVA,
- analyse **metric × metric** relationships using Pearson's correlation, significance tests, and scatterplots with optional trend lines,
- practice interpreting the numerical results and the plots in simple words (e.g. "There is / is not an association between …").
- export your results as Excel-Sheets or images.

2.1. Set-up: Load packages and dataset

1. Install `tidyverse`, `haven`, `janitor`, `sjmisc`, `skimr`,`scales` and `writexl` if not already installed and load these packages.

Solution:

```
# Install tidyverse
if (!requireNamespace("tidyverse", quietly = TRUE))
  install.packages("tidyverse")

# Install haven
if (!requireNamespace("haven", quietly = TRUE))
  install.packages("haven")

# Install readxl
if (!requireNamespace("janitor", quietly = TRUE))
  install.packages("janitor")

# Install skimr
if (!requireNamespace("skimr", quietly = TRUE))
  install.packages("skimr")

# Install scales
if (!requireNamespace("scales", quietly = TRUE))
  install.packages("scales")

# Install sjmisc
if (!requireNamespace("sjmisc", quietly = TRUE))
```

```
  install.packages("sjmisc")

# Install writexl
if (!requireNamespace("writexl", quietly = TRUE))
  install.packages("writexl")

## load packages
library(tidyverse)
library(haven)
library(sjmisc)
library(skimr)
library(janitor)
library(scales)
library(writexl)
```

2. For this exercise, we will use the file *03_qa_bi_data.sav* in the *exercises* folder. Define the *exercise* folder as your working directory with the function `setwd("path")`.

3. Choose `mydata3` as a name for your data frame and import it with the following structure: `chosen_name <- read_sav("filename")`.

Solution:

```
# set working directory
setwd(
  "C:/Users/Susanne/Nextcloud/share_DSC/003_Trainings/001_Workshops/2025/2025-12-02_SdV_Data_Analysis/Materials/E
  )

# Load the Nacaps data set and name this data frame "mydata3"
mydata3 <- read_sav("03_qa_bi_data.sav")
```

On the right side, in the Environment pane, the data frame `mydata3` should now appear under **Data**.

2.2. Categorical x categorical: Children (Yes/No) by Gender

In this part we look at the relationship between two categorical variables:

- Outcome (dependent variable): `blcd06` – children (yes/no)
- Predictor (independent variable): `adem01` – gender

To explore this, we will create a **contingency table** (gender × children), compute row percentages and column percentages and discuss which is more useful when children yes/no is our outcome, visualize the relationship with a clustered bar chart and run a **chi-squared test of independence**.

1. First convert `blcd06` and `adem01` into *factors* with labels.

Use `mutate()` and `as_factor()` to transform both variables into a factor variable. Save the result again as `mydata3`.

Solution:

```
mydata3 <- mydata3 %>%
  mutate(
    adem01 = as_factor(adem01),   # gender with labels
    blcd06 = as_factor(blcd06)    # children yes/no with labels
  )
```

2. Create a *contingency table* for `blcd01` and `adem01` using `tabyl()` with row and column percentages.

Run the plain code with absolute frequencies, including only cases without `NA`, like this:

```
# cross tabulation children and gender
tab_counts <- mydata3 %>% # store results in new data frame tab_counts
  filter(!is.na(adem01), !is.na(blcd06)) %>% # include only cases without NA
  tabyl(adem01, blcd06) # basic cross tabulation (absolute values)

tab_counts
```

```
##  adem01 yes  no
##  female 206 872
##    male 195 800
##   other  21 101
```

2a. To calculate row percentages (within each gender, how many have / do not have children?), include a last function `adorn_percentages("row")` in the pipe and safe the results in a new data frame `tab_row`.

Solution:

```
# Row percentages: within each gender, how many have / do not have children?
tab_row <- mydata3 %>%              # store results in new data frame tab_row
  filter(!is.na(adem01), !is.na(blcd06)) %>% # include only cases without NA
  tabyl(adem01, blcd06) %>%         # adem01 in row, blcd06 in column
  adorn_percentages("row") %>%      # row-wise percentages
  adorn_pct_formatting(digits = 1)  # format as % with 1 decimal

tab_row
```

```
##  adem01   yes    no
##  female 19.1% 80.9%
##    male 19.6% 80.4%
##   other 17.2% 82.8%
```

Within the group of females, 17.9 % have children, 18.3 % of the males, and within diverse people, only 15.9% have children.

2b. To calculate column percentages (Column percentages: within each children-category, gender distribution), include a last function `adorn_percentages("col")` in the pipe and save the results in a new data frame `tab_col`.

Solution:

```
# Column percentages: within each children-category, gender distribution
tab_col <- mydata3 %>%              # store results in new data frame tab_col
  filter(!is.na(adem01), !is.na(blcd06)) %>% # include only cases without NA
  tabyl(adem01, blcd06) %>%         # adem01 in row, blcd06 in column
  adorn_percentages("col") %>%      # column-wise percentages
  adorn_pct_formatting(digits = 1)  # format as % within 1 decimal

tab_col
```

```
##  adem01   yes    no
##  female 48.8% 49.2%
##    male 46.2% 45.1%
##   other  5.0%  5.7%
```

Within the group of respondents with children, 48.7 % are female, 46.1 % are male, and 5.0 people are diverse.

2c. Which one is more useful when children yes/no is our outcome of interest?

Solution:

If children yes/no is our outcome, we are usually interested in *"What share of people in each gender group have children?"*. Then **row percentages (condition on gender)** are more informative.

3. Use the data frame `tab_row` (which already stores your crosstab with row percentages) and export it as an Excel file named `tab_children_by_gender.xlsx` using `write_xlsx()` in your current working directory.

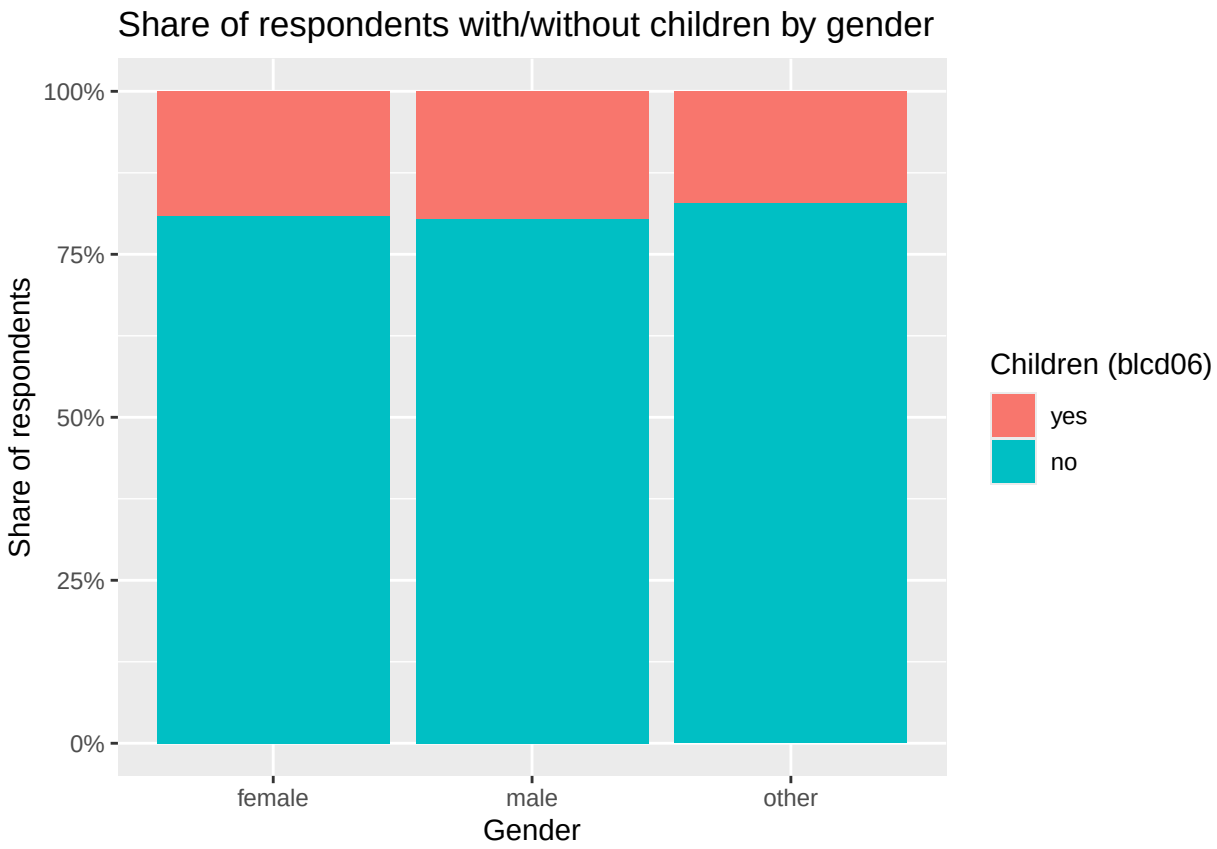
The code should look like this:

```
write_xlsx(tab_row, "tab_children_by_gender.xlsx")
```

4. Create a *clustered bar chart* to visualize the share of respondents with/without children in each gender groups.

The code should look like this:

```
mydata3 %>%                                          # take data frame mydata3
  filter(!is.na(adem01), !is.na(blcd06)) %>%         # keep only cases without NAs
  ggplot(aes(x = adem01, fill = blcd06)) +           # put adem01 on x-axis, colour bars by
   ↪  children yes/no
  geom_bar(position = "fill") +                      # draw stacked bars, each bar scaled to 100%
  scale_y_continuous(labels = percent) +             # show y-axis as %
  labs(                                              # labels
    x = "Gender",
    y = "Share of respondents",
    fill = "Children (blcd06)",
    title = "Share of respondents with/without children by gender"
  )
```



What do we see regarding the gender differences in parenthood?

Solution:

The graph makes it clear that there are hardly any gender differences in having children.

5. Export the figure to your working directory as a PNG file named "barchart_children_by_gender.png" using `ggsave()`.

The code looks like this:

```
ggsave(filename = "barchart_children_by_gender.png")
```

```
## Saving 6.5 x 4.5 in image
```

6. To check whether the gender differences in parenthood are random, calculate the Chi-Square test.

To do so, you can use the simple cross-tabulation data frame `tab_counts` from earlier with the function `chisq.test(tab_chi)`. What does it tell us?

Solution:

```
# chi-sqare test of independence of children and gender
chisq.test(tab_counts)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  tab_counts
## X-squared = 0.4164, df = 2, p-value = 0.812
```

Because the p-value is much larger than 0.05, we **do not reject the null hypothesis of independence**. There is no statistical evidence that having children (yes/no) is related to gender in this sample.

2.3. Categorial x Metric: Overall life satisfaction x children and gender

We now look at how overall life satisfaction (`blcd01`) differs between groups:

- Children (yes/no) – two groups → independent samples t-test
- Gender – more than two groups → one-way ANOVA

For each comparison we will, compute group means and standard deviations, visualize the differences with boxplots and run the appropriate statistical test.

3.2.1 Life satisfaction by children

1. Compare the mean life satisfaction between respondents with and without children.

To do this, use a `%>%` pipe using the `group_by(blcd06)` function and calculate central tendency measures (mean, sd) and distributions conveniently using the `skim()` function‘.

Note: Filter first with `filter(!is.na(bpsy01), !is.na(blcd06)) %>%` to include only valid categories in the calculations.

The code should look like this:

```
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(blcd06)) %>% # keep cases without missings only
  group_by(blcd06) %>%      # group by children yes/no
  skim(bpsy01)              # skim for life satisfaction by group
```

Table 1: Data summary

Name	Piped data
Number of rows	2192
Number of columns	17
Column type frequency:	
numeric	1
Group variables	blcd06

Variable type: numeric

skim_variable	blcd06	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bpsy01	yes	0	1	7.49	1.74	0	7	8	9	10	
bpsy01	no	0	1	6.93	1.93	0	6	7	8	10	

How does the average life satisfaction differ between groups?

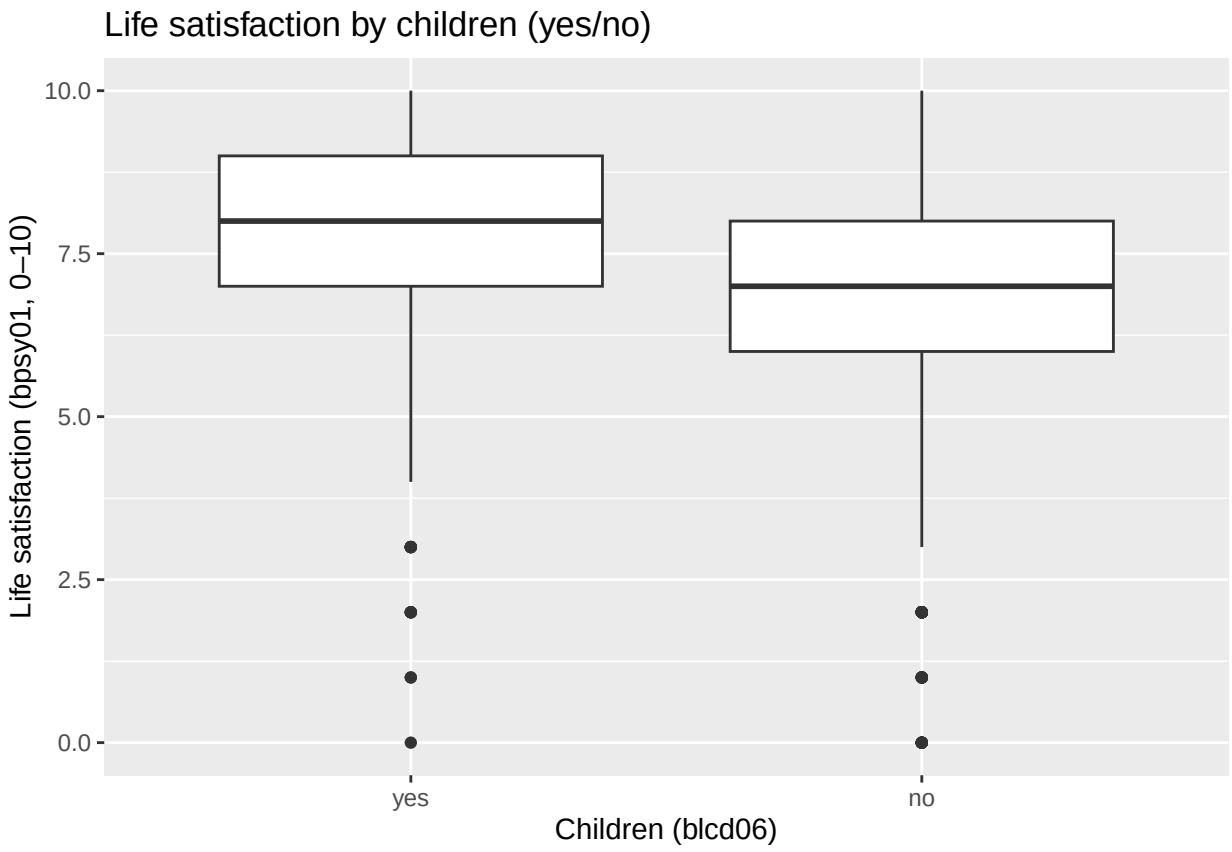
Solution:

On average, respondents with children report a higher overall life satisfaction that childless respondents.

2. Create *grouped boxplots* to visualize the differences in overall life satisfaction by respondents with/without children.

The code should look like this:

```
mydata3 %>%                                  # start with the dataset mydata3
  filter(!is.na(bpsy01), !is.na(blcd06)) %>%  # keep only cases without missings
  ggplot(aes(x = blcd06, y = bpsy01)) +       # map children yes/no to x-axis, life satisfaction to
   ↪  y-axis
  geom_boxplot() +                           # one boxplot for each children group
  labs(                                      # labels
    x = "Children (blcd06)",
    y = "Life satisfaction (bpsy01, 0-10)",
    title = "Life satisfaction by children (yes/no)"
  )
```



What do you see?

Solution:

The difference in life satisfaction between parents and childless respondents are also visibly strong.

3. Test whether the mean life satisfaction differs between people with and without children using the *t-test*.

The code should look like this:

```r
t_test_children <- mydata3 %>%                    # store results in new data frame t_test_children
  filter(!is.na(bpsy01), !is.na(blcd06)) %>%      # drop missings
  t.test(bpsy01 ~ blcd06, data = .)               # t-test: blcd01 by children

t_test_children
```

```
##
##  Welch Two Sample t-test
##
## data:  bpsy01 by blcd06
## t = 5.7798, df = 683.41, p-value = 1.137e-08
## alternative hypothesis: true difference in means between group yes and group no is not equal to 0
## 95 percent confidence interval:
##  0.3676599 0.7459630
## sample estimates:
## mean in group yes  mean in group no
##          7.486874          6.930062
```

What does the t-test tell us?

Solution:

- Test statistic: t = 5.78, df 683
- p-value = 1.14 × 10 , which is much smaller than 0.05
- 95% CI for the mean difference (yes – no) is [0.37, 0.75] → the whole interval is above 0

We **reject the null hypothesis** that the mean of `bpsy01` is the same for people with and without children.

Respondents with children have, on average, higher scores on overall life satisfaction (about 0.4–0.7 points higher) than respondents without children, and this **difference is statistically significant**.

3.2.2 Life satisfaction x Gender

1. Compare the mean life satisfaction between respondents of different gender.

To do this, use a `%>%` pipe using the `group_by(adem01)` function and calculate central tendency measures (mean, sd) and distributions conveniently using the `skim()` function‘.

Note: Filter first with `filter(!is.na(bpsy01), !is.na(adem01)) %>%` to include only valid categories in the calculations.

How does the average life satisfaction differ between groups?

Solution:

```r
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(adem01)) %>% # keep cases without missings only
  group_by(adem01) %>%      # group by gender
  skim(bpsy01)              # skim for life satisfaction by group
```

Table 3: Data summary

Name	Piped data
Number of rows	2196

Number of columns	17
Column type frequency:	
numeric	1
Group variables	adem01

Variable type: numeric

skim_variable	adem01	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bpsy01	female	0	1	7.04	1.89	0	6	7	8	10	
bpsy01	male	0	1	7.04	1.94	0	6	7	8	10	
bpsy01	other	0	1	6.98	1.82	1	6	7	8	10	

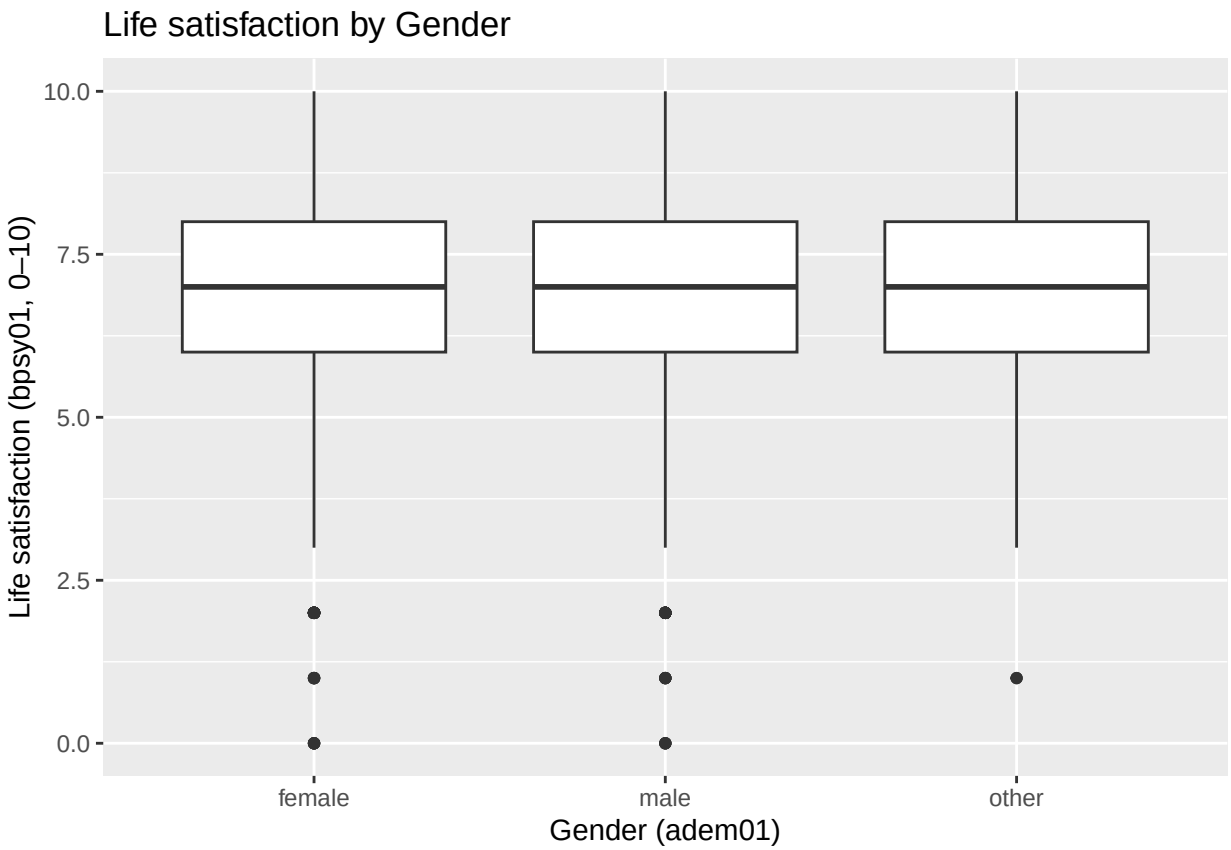
On average, respondents of different gender groups report similar overall life satisfaction (no gender differences).

2. Create *grouped boxplots* to visualize the differences in overall life satisfaction by gender groups.

What do you see?

Solution:

```
mydata3 %>%                                    # start with the dataset mydata3
  filter(!is.na(bpsy01), !is.na(adem01)) %>%   # keep only cases without missings
  ggplot(aes(x = adem01, y = bpsy01)) +        # map gender to x-axis, life satisfaction to y-axis
  geom_boxplot() +                             # one boxplot for each gender group
  labs(                                        # labels
    x = "Gender (adem01)",
    y = "Life satisfaction (bpsy01, 0-10)",
    title = "Life satisfaction by Gender"
  )
```



The boxplots by gender groups suggest no differences in average overall life satisfaction.

3. Test whether the mean life satisfaction differs between people with and without children using the one-way *ANOVA* test.

The code should look like this:

```
anova_gender <- mydata3 %>%                    # store results in new data frame anova_gender
  filter(!is.na(bpsy01), !is.na(adem01)) %>%   # keep cases without missings only
  aov(bpsy01 ~ adem01, data = .)               # ANOVA model: life satisfaction ~ gender

summary(anova_gender)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## adem01         2      0   0.249   0.069  0.934
## Residuals   2193   7962   3.631
```

Is there a significant gender difference in overall life satisfaction?

Solution:

For gender (`adem01`) ANOVA gives $F(2, 2193) = 0.069$, $p = 0.934$.

The p-value is much larger than 0.05, so the test statistic is very small compared to what we would expect under the null hypothesis.

We **do not reject the null hypothesis** that mean life satisfaction is the same across gender groups. In this sample, there is **no statistical evidence** that life satisfaction differs by gender.

2.4. Metric x Metric: Life Satisfaction and Age

We look at the relationship between overall life satisfaction (`bpsy01`) and gross income in Euro `bemp81`:

We will calculate the **Pearson correlation coefficient**, test if it is statistically significant and visualize the relationship with a scatterplot.

1. First, calculate the Pearson correlation coefficient between life satisfaction `bpsy01` and monthly gross income in Euro `bemp81`.

Filter out missing values and use `summarise()` with `r_pearson = cor(bpsy01, bemp81)`.

The code should look like this:

```
# calculate pearsons r
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(bemp81)) %>%          # drop missings
  summarise(
    r_pearson = cor(bpsy01, bemp81)                   # Pearson's r
  )
```

```
## # A tibble: 1 x 1
##   r_pearson
##       <dbl>
## 1     0.149
```

How would you interpret the outcome?

Solution:

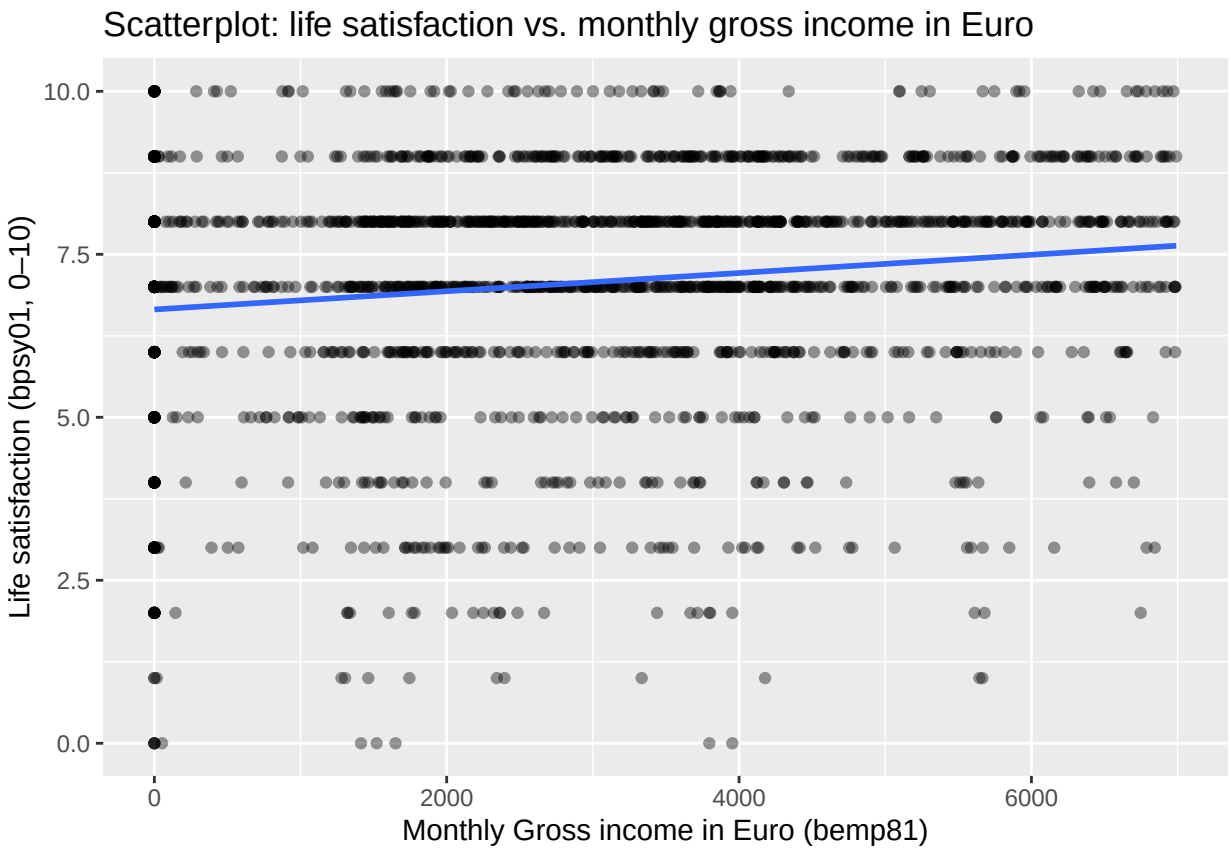
The correlation between life satisfaction (`bpsy01`) and monthly gross income (`bemp81`) is positive but small (r = 0.149). This means that higher income is weakly associated with higher overall life satisfaction.

2. Draw a scatterplot to visualize the relationship between both variables.

The code should look like this:

```
mydata3 %>%                                       # start with the dataset mydata3
  filter(!is.na(bpsy01), !is.na(bemp81)) %>%      # keep only cases without missings
  ggplot(aes(x = bemp81, y = bpsy01)) +           # age on x-axis, life satisfaction on y-axis
  geom_point(alpha = 0.4) +                       # draw one point per person (slightly transparent)
  geom_smooth(method = "lm", se = FALSE) +        # draw regression line (linear model, no CI band)
  labs(
    x = "Monthly Gross income in Euro (bemp81)",      # x-axis label
    y = "Life satisfaction (bpsy01, 0-10)",           # y-axis label
    title = "Scatterplot: life satisfaction vs. monthly gross income in Euro"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



What does the scatterplot imply about the relationship between overall life satisfaction and monthly gross income?

Solution:

The scatterplot shows a weak positive trend: higher incomes tend to be associated with slightly higher life satisfaction, but the pattern is very diffuse. Most points are widely scattered, indicating substantial individual variation at all income levels.

Overall, income and life satisfaction show only a small linear relationship, which aligns with the low correlation (r 0.15).

3. Finally, apply a significance test.

Use the filtered data again and run `cor.test()` to obtain the correlation, a p-value, and a confidence interval to test whether the correlation is significantly different from zero.

The code should look like this:

```
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(bemp81)) %>%          # drops cases with missings
  cor.test(                                           # run cor.test on the filtered data
    ~ bpsy01 + bemp81,                                # formula: two numeric variables
    data   = .,                                       # use the piped data frame
    method = "pearson"                                # apply pearson method
  )
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  bpsy01 and bemp81
## t = 6.7151, df = 1992, p-value = 2.443e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1055740 0.1914279
## sample estimates:
##       cor
## 0.1487813
```

What is your conclusion?

Solution:

You've got r = 0.1488 (Pearson correlation), t = 6.72, df = 1992, p = 2.44e-11, 95% CI for r: 0.1056 to 0.1914.

The **correlation is positive, but very small** (people with higher income tend to report slightly higher life satisfaction, but the effect is weak).

The **p-value is extremely small** ($p < 0.001$), so we reject the null hypothesis that the true correlation is 0. The confidence interval does not include 0.

There is a **statistically significant but very weak** linear association between income and life satisfaction in this sample.

Take Home checklist: Bivariate analysis in R

Step	Question / task	Useful functions / tools
1	What type of variables are you combining (cat × cat, cat × metric, metric × metric)?	Check variable types with `str()`, `summary()`, `skimr::skim()`.
2	For **categorical × categorical**: Is there an association between the two variables?	Contingency tables with `tabyl(var1, var2)`, row/column % with `adorn_percentages()`, `adorn_pct_formatting()`, chi-squared test with `chisq.test()`, grouped bar chart with `ggplot(aes(x = var1, fill = var2)) + geom_bar(position = "fill")`.
3	For **categorical × metric**: Do group means of the metric outcome differ?	Grouped descriptives with `group_by(cat) %>% skim()`, boxplots with `geom_boxplot()`, t-test with `t.test(outcome ~ cat)`, one-way ANOVA with `aov(outcome ~ cat)`.
4	For **metric × metric**: Is there a linear relationship between the two variables?	Pearson correlation with `cor()`, significance test with `cor.test()`, scatterplot with `ggplot(aes(x = xvar, y = yvar)) + geom_point()` (optional `+ geom_smooth(method = "lm", se = FALSE)`).
5	Are the numerical results and the plots telling a consistent story?	Compare effect size (difference in means, correlation) with p-values, confidence intervals and the visual patterns in boxplots / bar charts / scatterplots.