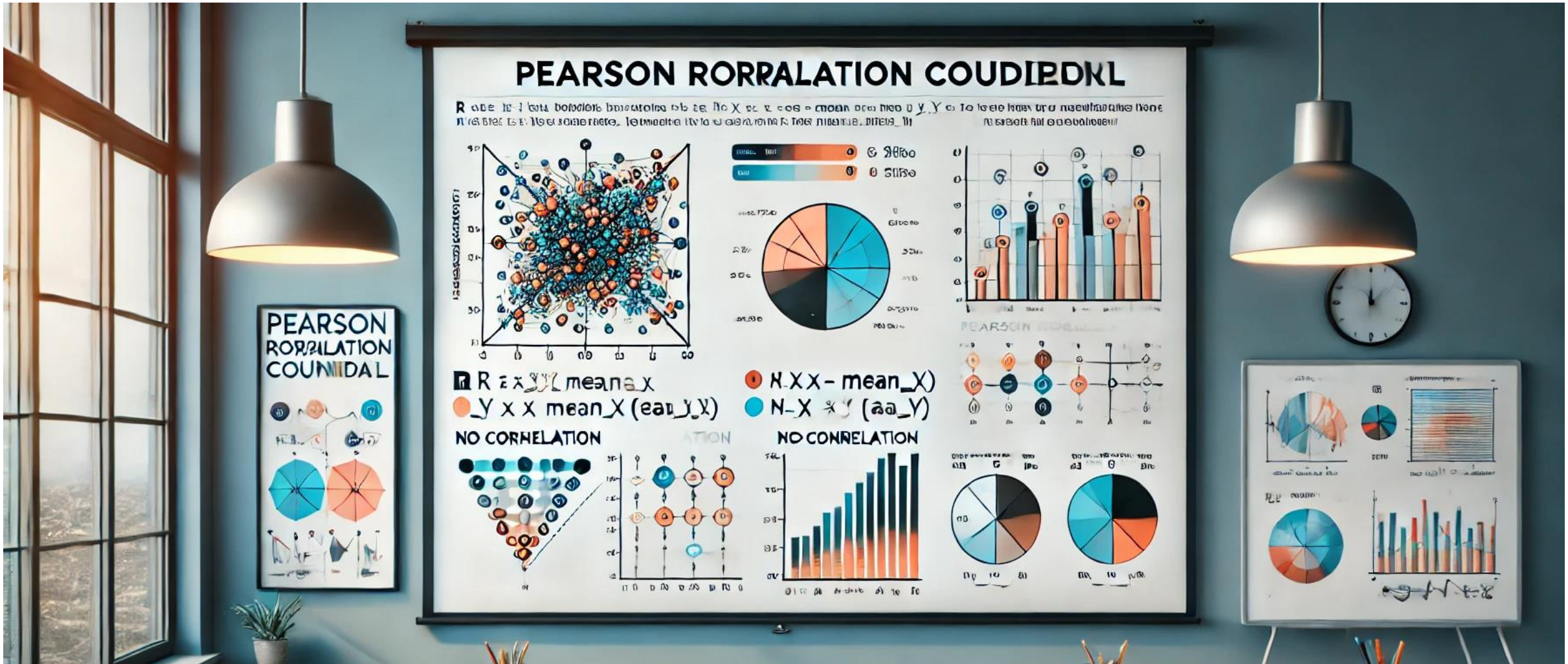
PEARSON CORRELATION COEFFICIENT

(PCC)

TECHNISCHE UNIVERSITÄT DARMSTADT

Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt



Research design

Research design

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

>> the more I learn, the better my grade will be <<

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

>> the more I learn, the better my grade will be <<

positive correlation

Research design

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

"Is there a correlation between the grade and the prep-time for an exam?"

H_1 *>> the more I learn, the better my grade will be <<*

Survey

PCC in R

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

>> the more I learn, the better my grade will be <<

Survey

PCC in R

i	name	grade	prep-time
1			
2			
3			
4			

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

>> the more I learn, the better my grade will be <<

Survey

PCC in R

i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

Data-prep

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

Data-prep

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

sequence number

dependent variable

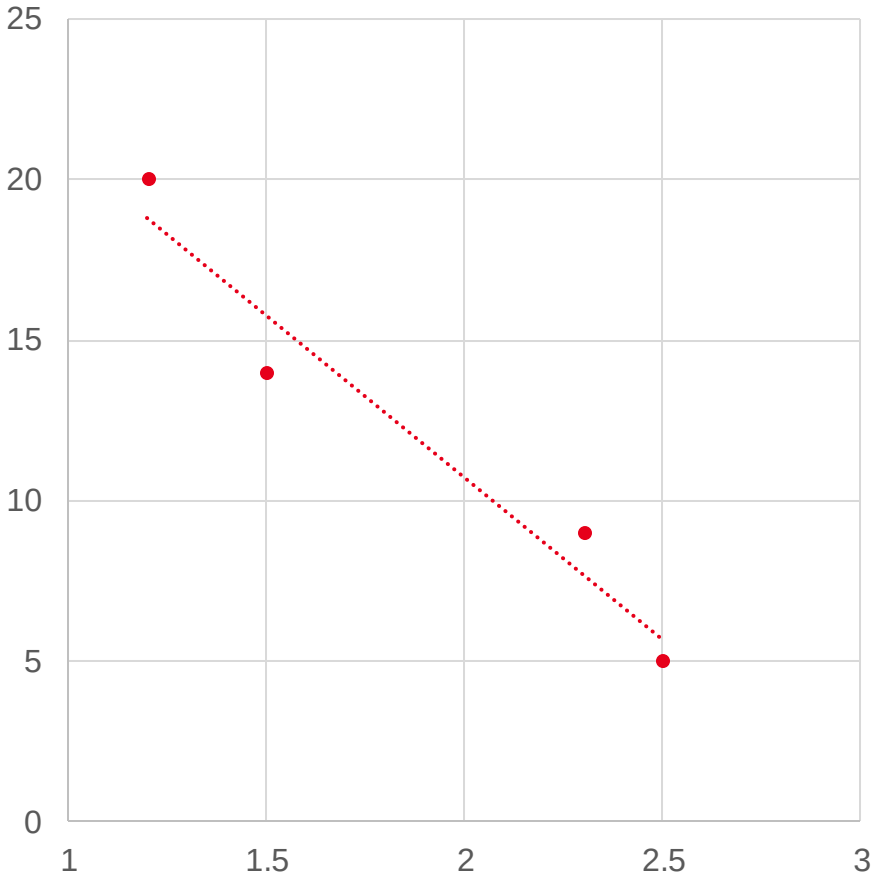
independent variable

i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

Data-prep

i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

independent variable



dependent variable

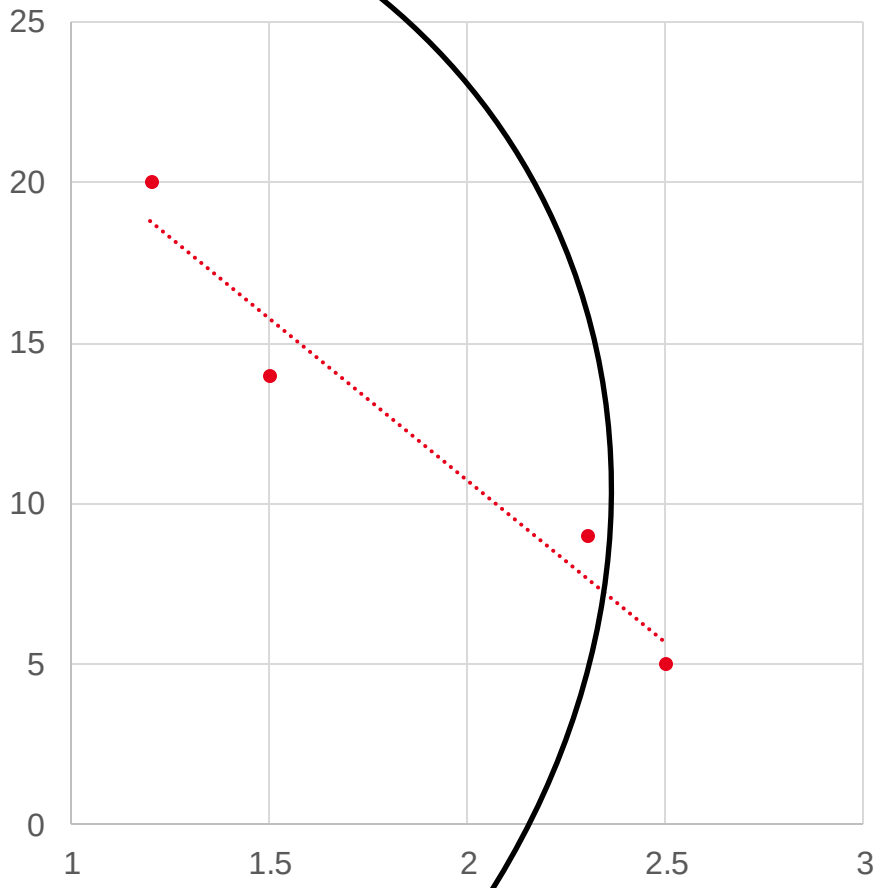
Data-prep

i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

x_i

y_i

independent variable



dependent variable

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

>> the more I learn, the better my grade will be <<

Survey

PCC in R

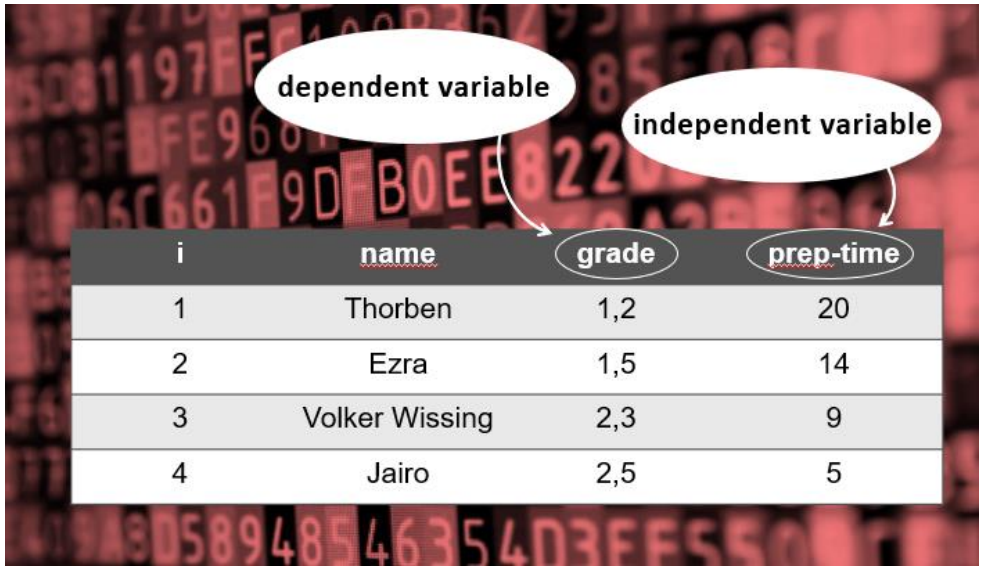
i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

PCC = r

- **Pearson Correlation Coefficient = „r“ is just one of many ways to measure of the degree of linear correlation between two variables (x, y)**
- **popularized by the French physicist and crystallographer Auguste Bravais, 19th century**
- **Determine:**

1) How strong is the correlation?

2) Is the correlation positive or negative?



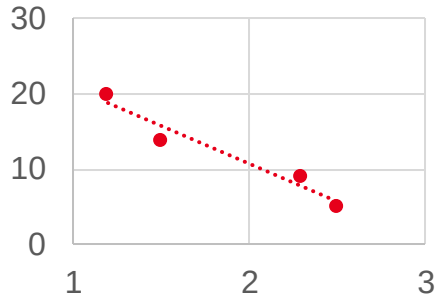
i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

PCC = r

- **1) How strong is the correlation?**

absolut value of r	strength of the correlation
0,0 < 0,1	no correlation
0,1 < 0,3	low correlation
0,3 < 0,5	moderate correlation
0,5 < 0,7	high correlation
0,7 < 1	very high correlation

- **2) Is the correlation positive or negative?**



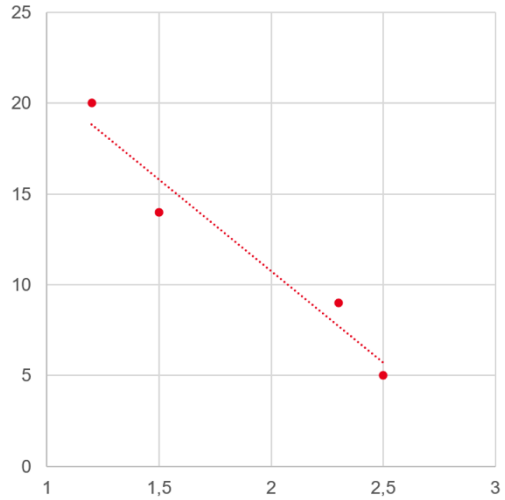
- **positive correlation: the more, the more**
- **negative correlation: the more, the less (exceptions: e.g. grades!)**
- **r ε [-1;+1]**
  +1 = perfect positiv correlation
  - 1 = perfect negativ correlation
- **r = 0, no linear correlation between the variable independently = Null Hypothesis (H_0)**

PCC = r

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

i	name	x_i grade	y_i prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

independent variable



dependent variable

PCC

limitations

- PCC can only identify linear correlation
- quadratic or exponential correlations are not detected
- NO causality – only correlation!!!
- metric scaled data (cardinal scale)

&

biases

PCC

limitations

- **PCC can only identify linear correlation**
- **quadratic or exponential correlations are not detected**
- **NO causality – only correlation!!!**
- **metric scaled data**

&

biases

- **sampling bias:** not representative sample **– solution: random and representative samples**
- **outlier bias:** individual extreme values **– solution: large sample**
- **spurious correlation:** no real correlation **– solution: control of confounding variables**
- **…**

Research design

“Is there a correlation between the grade and the prep-time for an exam?”

>> the more I learn, the better my grade will be <<

Survey

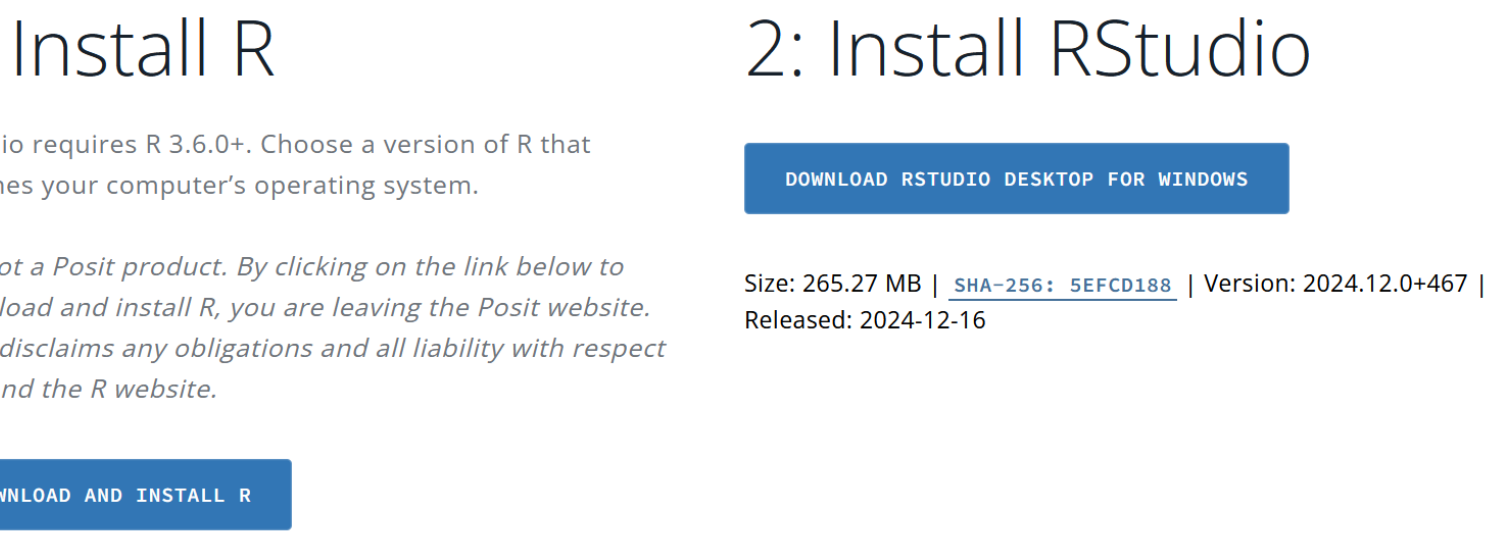
PCC in R

i	name	grade	prep-time
1	Thorben	1,2	20
2	Ezra	1,5	14
3	Volker Wissing	2,3	9
4	Jairo	2,5	5

back-up R

STEP 1 **Installation of R and RStudio**

- https://posit.co/download/rstudio-desktop/



R – installation packages & library

STEP 2 **import Excel file:**

install and download the required packages (if not already installed):

```
install.packages("readxl") # if not installed
install.packages("ggplot2") # for visualization of the correlation
```

STEP 3 **load the libraries:**

```
library(readxl)
library(ggplot2)
```

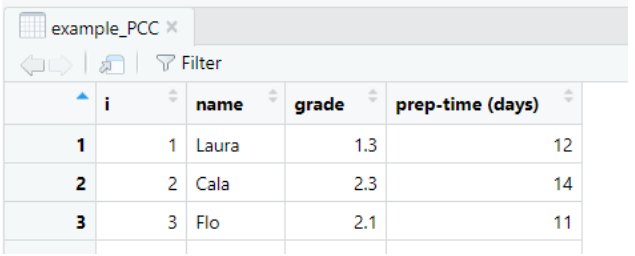
R – import excel file

STEP 4 **to import Excel file:**

```
# import file (replace 'file.xlsx' with the actual file name)
library(readxl)
# name excel-sheet and import the sheet (read_excel)
example_PCC <- read_excel("C:/Users/ralfv/Desktop/_TU Darmstadt MSc/_24_25 WS/ASTRAI/Lecture_Bonus_PCC/example_PCC.xlsx")
# show excel-sheet
View(example_PCC)
# name sheet df
df <- read_excel("C:/Users/ralfv/Desktop/_TU Darmstadt MSc/_24_25 WS/ASTRAI/Lecture_Bonus_PCC/example_PCC.xlsx", sheet = 1)

 # show the first lines to check the data

head(df)
```



	i	name	grade	prep-time (days)
1	1	Laura	1.3	12
2	2	Cala	2.3	14
3	3	Flo	2.1	11

```
# A tibble: 6 × 4
      i name  grade `prep-time (days)`
  <dbl> <chr> <dbl>              <dbl>
1     1 Laura   1.3                 12
2     2 Cala    2.3                 14
3     3 Flo     2.1                 11
4     4 Adian   1.7                 15
5     5 Jairo   2.4                 10
6     6 Sara    2.6                 11
```

R – Pearson correlation

STEP 5 **calculate the PEARSON CORRELATION:**

```
# if non-numeric columns are present, remove them
df_numeric <- df[sapply(df, is.numeric)]

# calculate the correlation
cor_matrix <- cor(df_numeric, method = "pearson", use = "pairwise.complete.obs")

# display the correlation table
print(cor_matrix)
```

```
                          i      grade prep-time (days)
i                 1.0000000  0.2178065       -0.1009091
grade             0.2178065  1.0000000       -0.8205796
prep-time (days) -0.1009091 -0.8205796        1.0000000
```

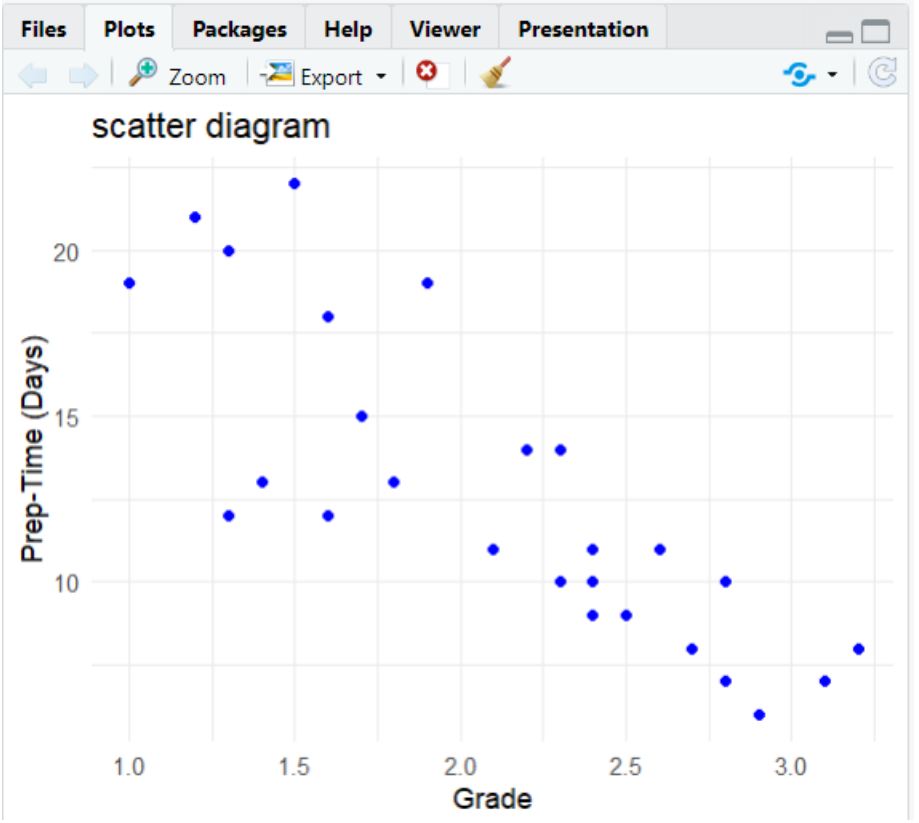
R – plot

STEP 6

visualize a scatter diagram:

```
library(ggplot2)

ggplot(df, aes(x = grade, y = `prep-time (days)`)) +
 # insert points
geom_point(color = "blue") +
# name titel
ggtitle("scatter diagram") +
# name x- and y-axis
xlab("Grade") +
ylab("Prep-Time (Days)") +
theme_minimal()
```

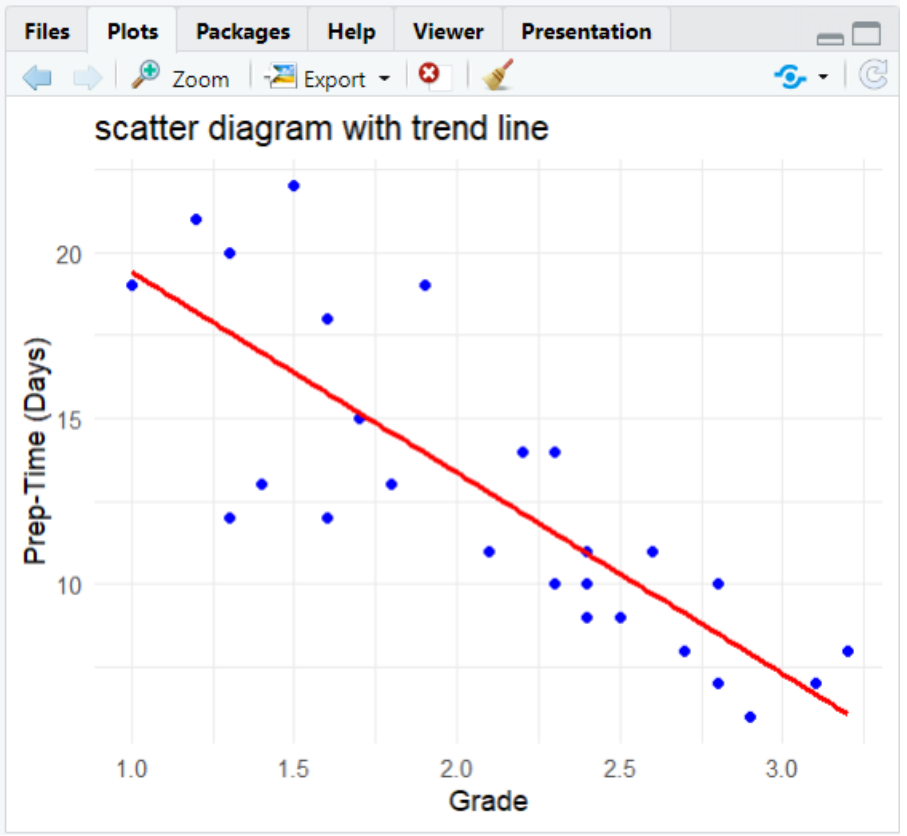


R – plot

STEP 7

visualize a scatter diagram with trend line:

```
ggplot(df, aes(x = grade, y = `prep-time (days)`)) +
 # insert points
geom_point(color = "blue") +
geom_smooth(method = "lm", se = FALSE, color = "red") +
 # linear trend line
ggtitle("scatter diagram with trend line") +
# name x- and y-axis
xlab("grade") +
ylab("prep-time (days)") +
theme_minimal()
```



R – p-value

STEP 8 **calculate p-value, based on Pearson correlation coefficient:**

```
result <- cor.test(df$grade, df$ `prep-time (days)`, method = "pearson")
# show p-value
result$p.value
```

```
[1] 2.84699e-07
```

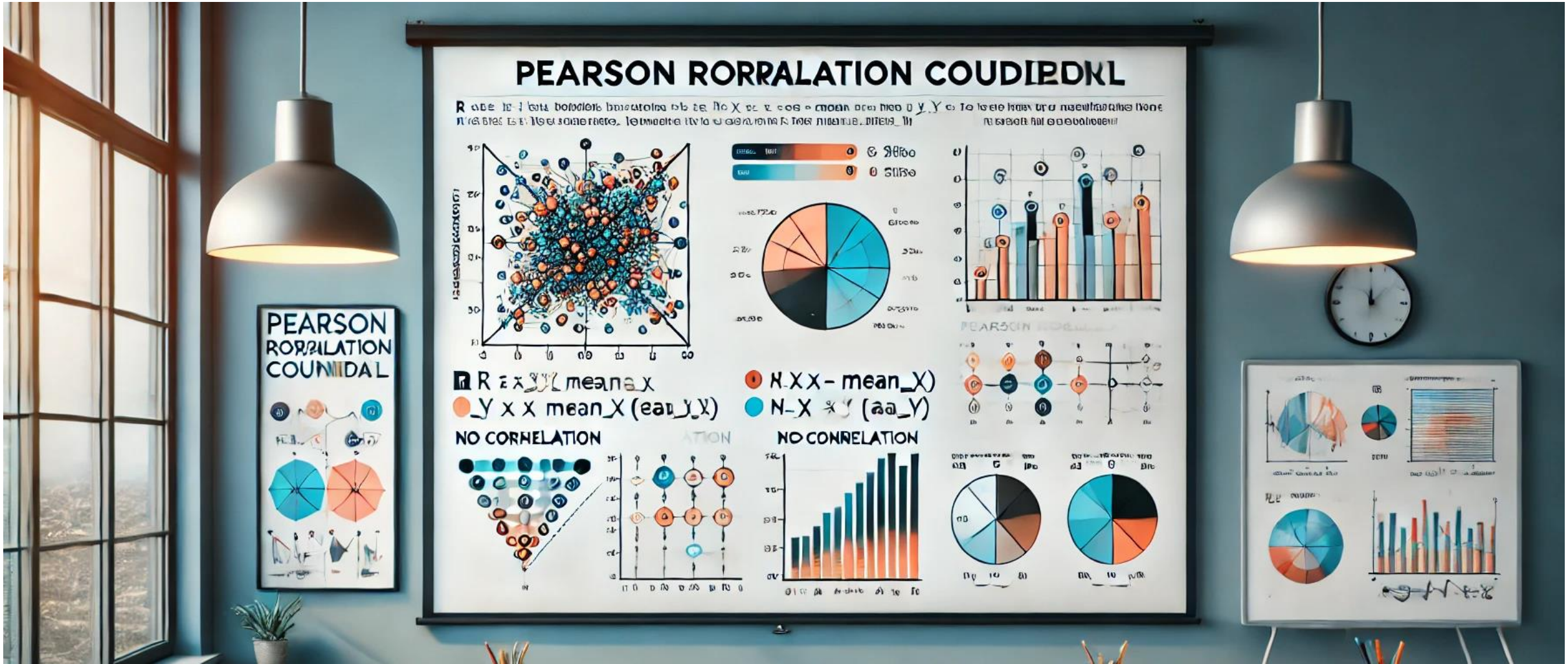
p-value

- indicates **whether the correlation between two variables is statistically significant** or **whether it may have arisen by chance**
- **tests the null hypothesis (H_0)**
- **low p-value (< 0.05):** assumtion that **there is a statistically significant correlation**
- **high p-value (> 0.05):** indicates that the **correlation is not significant**
- **requirement: to test if the correlation coefficient deviates significantly from zero, both variables (x, y) need to be normally distributed variables**

ANY QUESTIONS?

TECHNISCHE UNIVERSITÄT DARMSTADT

Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt



literature

- Becker, T., Herrmann, R., Heumann, C., Pilz, S., Sandor, V., Schäfer, D., & Wellisch, U. (2025). *Stochastische Risikomodellierung und statistische Methoden: Angewandte Stochastik für die aktuarielle Praxis*. Springer-Verlag. S145-153.
- Brückler, F. M., & Brückler, F. M. (2018). Geschichte der Topologie. *Geschichte der Mathematik kompakt: Das Wichtigste aus Analysis, Wahrscheinlichkeitstheorie, angewandter Mathematik, Topologie und Mengenlehre*, S.116.
- Holland, H., Scharnbacher, K., Holland, H., & Scharnbacher, K. (2015). Regressions-und Korrelationsanalyse. *Statistik im Betrieb: Lehrbuch mit praktischen Beispielen*, 193-209.
- Dingil, A. E., Schweizer, J., Rupi, F., & Stasiskiene, Z. (2018). Transport indicator analysis and comparison of 151 urban areas, based on open source data. *European Transport Research Review*, *10*, 1-9.
- Karch, J. D., Perez-Alonso, A. F., & Bergsma, W. P. (2024). Beyond Pearson's correlation: modern nonparametric independence tests for psychological research. *Multivariate Behavioral Research*, *59*(5), 957-977
- Kuckartz, U., Rädiker, S., Ebert, T., Schehl, J., Kuckartz, U., Rädiker, S., ... & Schehl, J. (2010). Korrelation: Zusammenhänge identifizieren. *Statistik: Eine verständliche Einführung*, 189-213.
- Sedgwick, P. (2012). Pearson's correlation coefficient. *Bmj*, *345*.
- Schmuller, J. (2021). *Statistik mit R für Dummies* (2. Aufl.). Wiley-VCH.

back-up

Update R to newest version

1) Open RStudio
2) Go to Tools > Global Options > General
3) Under “R Version” click on Change and select the new R version (4.2.3)
4) Restart RStudio