



## 1 | Disclaimer

I do not usually provide answers to these questions, as I am expecting you to either attend the seminar, or – in case you cannot make it – to contact me for advice or to share notes with another student. But I have nonetheless put this document together, so that you get an idea about my expectation horizon for answering these. Please bear this in mind in weeks to come.

## 2 | Answers

1. Why is it good practice to always use an RScript, rather than simply entering your commands into the console in R?
  - Ensures reproducibility
    - Scripts provide a saved record of all your steps, making it easy to reproduce or rerun your analysis.
  - Improves organization
    - You can structure your code with comments, sections, and logical flow, making complex work manageable.
  - Easier debugging
    - You can edit and re-run parts of your code without retyping everything or losing previous work.
  - Saves time
    - Once your script is written, you can re-run the entire analysis with one click or command—ideal for iterative work.
  - Better collaboration and communication
    - Sharing an R script allows others to understand, verify, or extend your work.
2. Give an example of an unordered factor variable and an ordered one.
  - Unordered factor variable (nominal variable)
    - “fruit” with categories “apple”, “banana”, “orange”
  - Ordered factor variable (ordinal variable)
    - “education\_level” with categories “High School”, “Bachelor”, “Master”, “PhD”

3. Why are descriptive statistics useful?

- Summarize data quickly
  - They condense large datasets into a few key numbers (e.g., mean, median, standard deviation) for quick interpretation.
- Understand data distribution
  - Measures like range, variance, skewness, and kurtosis reveal the shape and spread of your data.
- Identify patterns and trends
  - Descriptive statistics highlight central tendencies and variability, helping detect patterns or irregularities.
- Spot errors or outliers
  - Unusual values in summary statistics can signal data entry mistakes or meaningful anomalies.
- Prepare for further analysis
  - Descriptive stats are a critical first step before performing inferential statistics or building models.

4. Give an example of a situation where you would need to transform a character variable into a factor variable.

- Whenever you need the values of a variable as levels:
  - gender: “male”, “female”
  - month: “January”, “February”, etc.
  - education: “High School”, “Bachelor”, “Master”, “Doctoral”
  - region: “north”, “south”, etc.

5. Why should you always create a new variable, instead of overwriting an existing one?

- Preserve original data: Keeping the original variable intact ensures you can always refer back to the unmodified data if needed.
- Avoid accidental data loss: Overwriting removes the original content, which can't be recovered unless previously saved or backed up.
- Improve reproducibility: Creating new variables helps others (or future you) understand each step of the transformation process.
- Enhance debugging: If something goes wrong, it's easier to trace the issue when intermediate variables are preserved.
- Facilitate comparison: You can compare the original and modified data side-by-side, which helps with validation and analysis.
- Encourage cleaner code: Using new variables reflects clearer logic and makes your code more readable and maintainable.