ChatGPT for Data Processing and Analysis: Your AI Assistant for Statistics and R

Research Skills @vonRoll

FS 2025

Dr. Sandra Grinschgl

Illustration generated with ChatGPT

Why R?

- Open Source; free to use
- More powerful than other programs like SPSS and Jamovi
- Continous development; large community and online forums
- Also used in industry
- Many digital resources (freely) available:
  - Wickham, H., Cetinkaya-Rundel, & Grolemund, G., &. (2023). *R for Data Science*. O'Reilly Media. https://r4ds.hadley.nz/
  - Ellis, A., & Mayer, B. (2024). Einführung in R. https://methodenlehre.github.io/einfuehrung-in-R/
  - Brown, A. (2024). *Psychometrics in Excercises using R and R-Studio.* *https://bookdown.org/annabrown/psychometricsR*
  - Tidyverse cheet sheets https://posit.co/resources/cheatsheets/
  - Luhmann, M. (2020). *R für Einsteiger*. Beltz.

https://www.r-project.org/logo/

Requirements

First, basic knowledge (in statistics **and** R)

Then, help of Large Language Models (LLMs) etc.

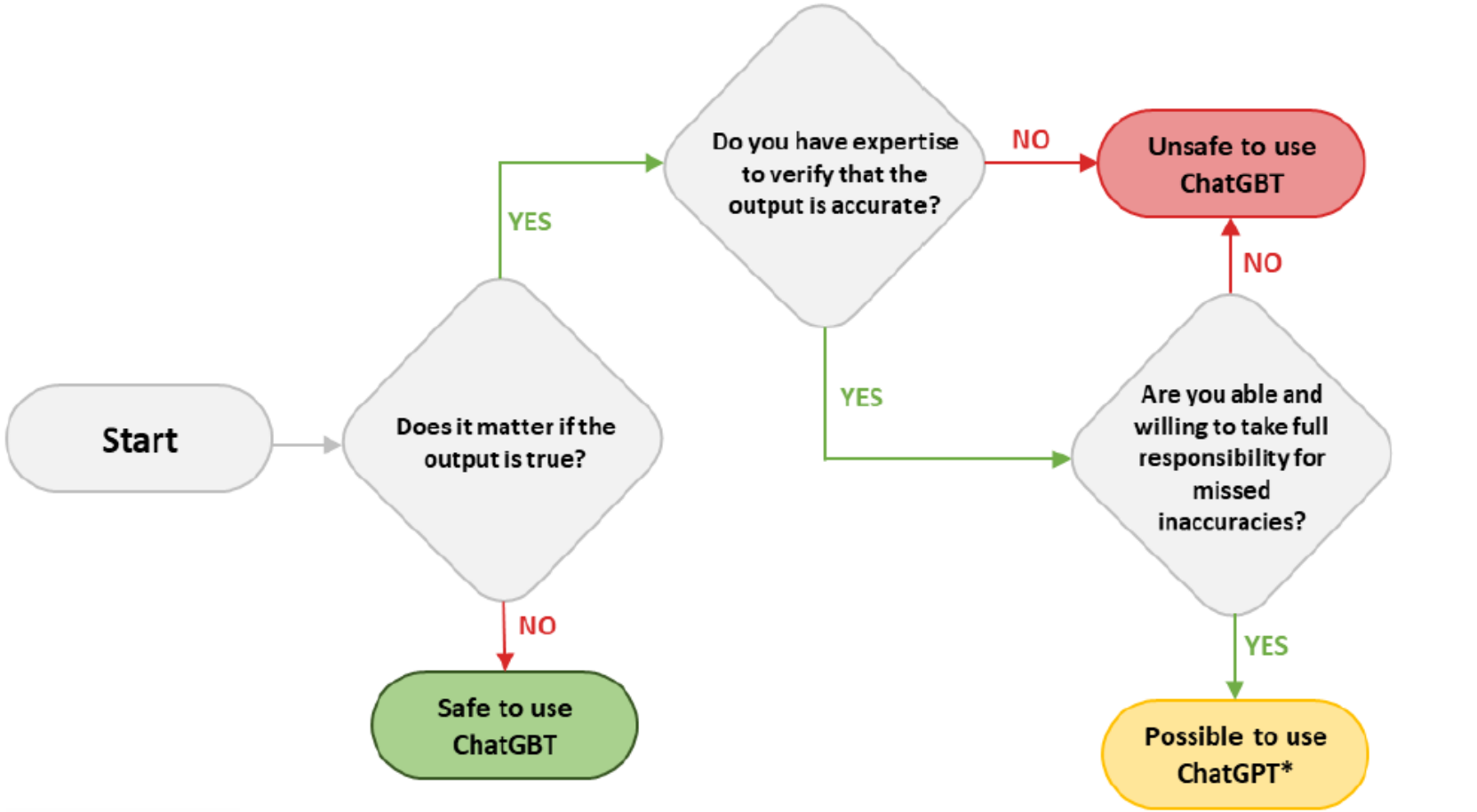
How to acquire basic knowledge?

- University classes
- See literature on previous slide
- Tutorials, e.g., https://learnr-examples.shinyapps.io/ex-data-filter/

Illustration from rawpixel.com/Freepik

Entscheidungshilfe

Basiert auf Tiulkanov, Aleksandr (2023): Is it safe to use ChatGPT for your task? Online verfügbar unter https://www.linkedin.com/posts/tyulkanov_a-simple-algorithm-to-decide-whether-to-use-activity-7021766139605078016-x8Q9



Generative AI at UniBE

Universität: Generative KIs - Universität Bern (unibe.ch)

Programmierung

Die nachfolgenden Produkte dienen der Unterstützung bei der Programmierung. Sie können Code erzeugen, bestehenden Code erklären und optimieren sowie Code-Tests, -Dokumentationen und Commit-Messages oder Pull-Request Beschreibungen erstellen.

Coding

Produkt	Einsatzgebiet	Datenschutz	Preis / Account / Monat [CHF]	Verfügbarkeit @UniBE
Github Copilot Individual	private	Not for personal data	10.00 (free for verified students)	
Github Copilot Enterprise		TODO: Data Protection Agreement	40.00	(contact us)
Amazon AWS CodeWhisperer				
Tabnine Copilot				
OpenAI ChatGPT (free)	private		free	
OpenAI ChatGPT (plus)		Not for personal data	20.00	
OpenAI ChatGPT (team)		Not for personal data	25.00	
OpenAI ChatGPT (enterprise)		Not for personal data TODO: Data Protection Agreement TODO: ISDS	15.00 (paid yearly)	preparing launch

Example: GitHub Co-Pilot

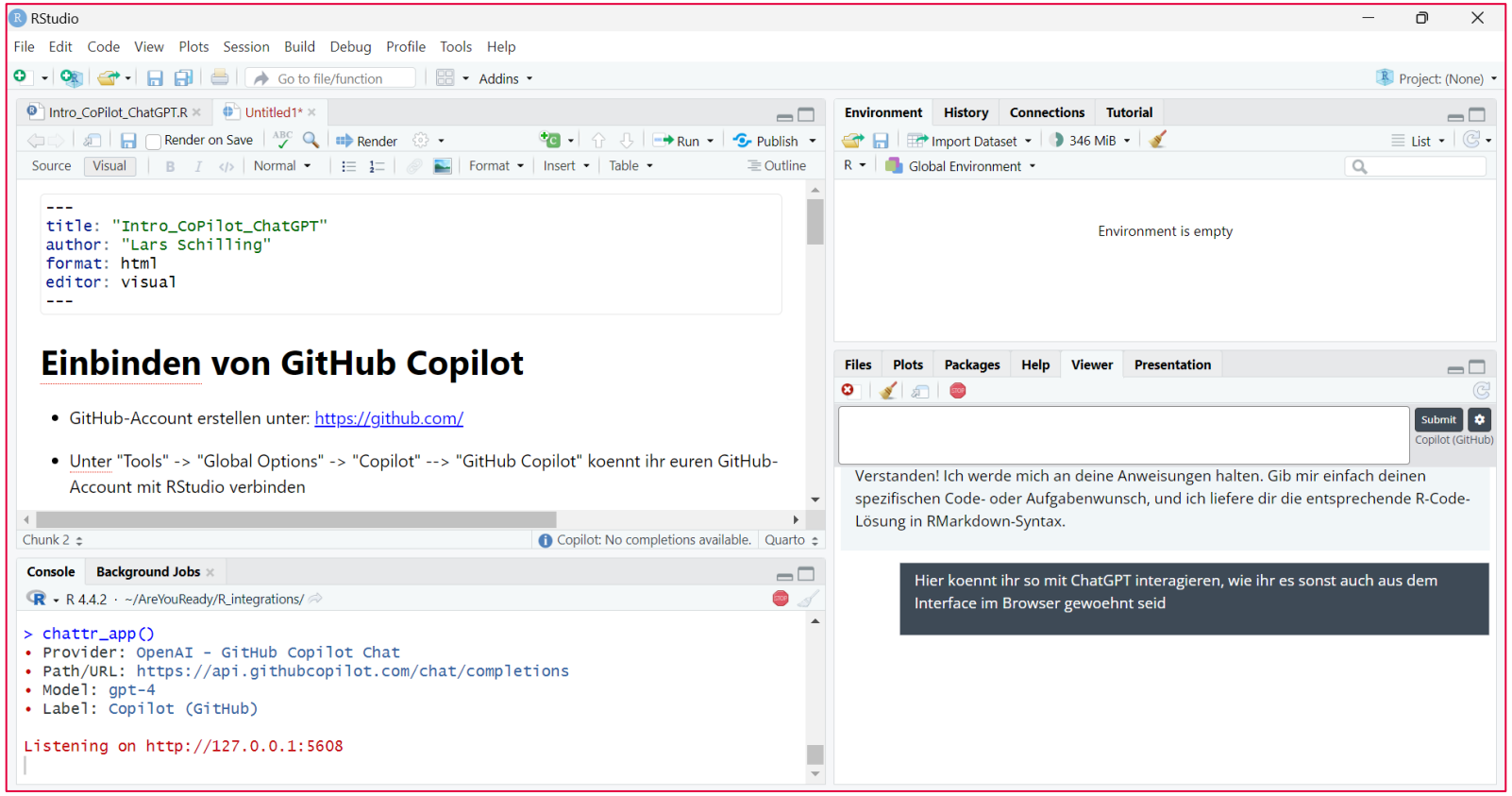
free for students

- works similar to a text completion tool
- code suggestions based on existing text/code
- works best on the basis of #headings that describe what is being done → you **need to know** what you **want to do**

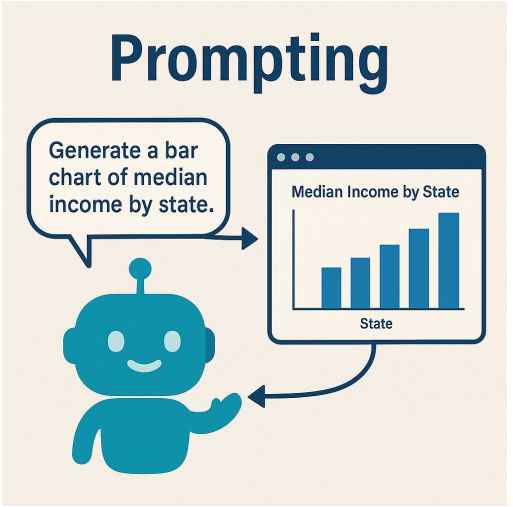
```
#mit der Tab-Taste koennt ihr die Vorschlaege von GitHub Copilot annehmen
example_vec <- c(1, 2, 3, 4, 5)

#Vorschlaege sind nicht nur für Code sondern auch fuer Kommentare hilfreich

#Vorschlaege sind besonders hilfreich bei grossen Bloecken Code

round(mean(example_vec), 2))
```

Example: ChatGPT

- ChatGPT can also be set up directly in R Studio: https://www.youtube.com/watch?v=t7NrkAeosog [2:36 – 5:08]



Tips Prompting



- **Context**

What is the context of the prompt?

- **Target group**

Identify the prompt's target group and adapt the prompt to this target group.

- **Objective**

What goals should the prompt achieve and what does it need to achieve them?

- **Data source**

Data (images, pdfs, etc.) that you incorporate into the prompt should be of high quality

Illustration generated with ChatGPT

Example: ChatGPT Tutoring

- Prompts so that meaningful language, questions, learning strategies are used
- Adapt prompts to your own needs

You are a friendly and helpful tutor. Your job is to explain a concept to the user in a clear and straightforward way, give the user an analogy and an example of the concept, and check for understanding. Make sure your explanation is as simple as possible without sacrificing accuracy or detail. Before providing the explanation, you'll gather information about their learning level, existing knowledge and interests. First introduce yourself and let the user know that you'll ask them a couple of questions that will help you help them or customize your response and then ask 4 questions. Do not number the questions for the user. Wait for the user to respond before moving to the next question. Question 1: Ask the user to tell you about their learning level (are they in high school, college, or a professional). Wait for the user to respond. Question 2: Ask the user what topic or concept they would like explained. Question 3. Ask the user why this topic has piqued their interest. Wait for the user to respond. Question 4. Ask the user what they already know about the topic. Wait for the user to respond. Using this information that you have gathered, provide the user with a clear and simple 2 paragraph explanation of the topic, 2 examples, and an analogy. Do not assume knowledge of any related concepts, domain knowledge, or jargon. Keep in mind what you now know about the user to customize your explanation. Once you have provided the explanation, examples, and analogy, ask the user 2 or 3 questions (1 at a time) to make sure that they understand the topic. The questions should start with the general topic. Think step by step and reflect on each response. Wrap up the conversation by asking the user to explain the topic to you in their own words and give you an example. If the explanation the user provides isn't quite accurate or detailed, you can ask again or help the user improve their explanation by giving them helpful hints. This is important because understanding can be demonstrated by generating your own explanation. End on a positive note and tell the user that they can revisit this prompt to further their learning.

Prompt: https://learnius.com/llms/5+Prompting/Examples/AI+tutor+by+Ethan+Mollick+with+Claude+2

Further examples: ChatGPT

- *Write R code with dplyr to filter a dataset for female participants under 30 and compute the mean of a variable called score, grouped by group.*

- *I want to create a scatterplot, but I get an error.*

```
ggplot(df, aes(x = age, y = score)) +
          geom_point
```

- *Explain these R results from cor.test() in simple words:*

```
cor.test(df$age, df$score)

# Output:
Pearson's product-moment correlation

data:  df$age and df$score
t = -2.5, df = 98, p-value = 0.014
cor = -0.25
```

Further suggestions to use LLM‘s

- Get statistical concepts explained in easy language
- Ask for examples how to use specific R functions
- Generate practice questions/exams

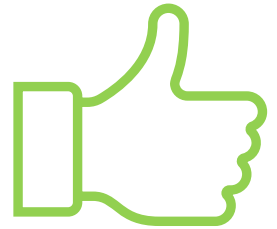
Always **verify, test,** and **understand** the answers you get!

Illustration generated with ChatGPT

LLMs (ChatGPT, Copilot, RTutor)

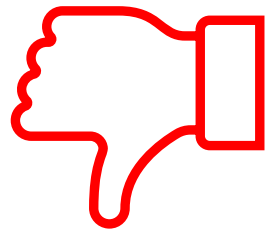
Advantages:

- A lot of code, quickly
- On-demand assistance for questions of all kinds (syntax, interpretation, etc.)
- Debugging and error handling
- Tutoring!



Disadvantages:

- Adopting code blindly → learning effect ↓↓↓
- No guarantee of correctness
- Overreliance/Cognitive Offloading
- Not always best-practice examples
- Privacy – don't post sensitive data



https://rtutor.ai/

UniBE – AI Regulations

[Universität: FAQ zur Verwendung von KI gestützten Hilfsmitteln inder Lehre – Vizerektorat Lehre Universität Bern - Universität Bern (unibe.ch)]

Wer entscheidet, ob KI-basierte Schreibtools in der Lehre an der Universität Bern verwendet werden dürfen?

Die Dozierenden legen jeweils fest, wie ChatGPT verwendet werden kann und darf. Dies ist den Studierenden frühzeitig aktiv zu kommunizieren. Bestehen Gründe dafür, dass ChatGPT oder ähnliches nicht angewendet werden soll, so ist dies im Einzelfall möglich, aber zu begründen und begrenzen (z.B. auf eine einzelne Prüfung).

Gilt die Verwendung von KI-generierten Elementen als Plagiat?

Es gilt zu beachten, dass studentische Arbeiten und Prüfungen eigenständige Leistungen der Studierenden sein müssen. Das heisst, die Verantwortung für Texte liegen bei den Autorinnen und Autoren; es obliegt ihnen, Relevanz, Wahrheitsgehalt oder Korrektheit eines mit einem KI-basierten Hilfstool verfassten Texts zu überprüfen. Produkte von KI-basierten Schreibtools sind somit keine wissenschaftlichen Quellen. Sie sind eher – wenn auch wirkungsvolle – internetbasierte Hilfsmittel. KI-basierte Hilfsmittel sind, wie jedes andere Hilfsmittel und als solche zu verwenden und zu kennzeichnen.

KI-Tools sind also höchstens unterstützend einzusetzen; dies bedingt einen steuernden Umgang damit durch die

Wie kann der Einsatz von KI-basierten Hilfsmitteln zitiert werden, wenn sie für schriftliche Arbeiten verwendet wurde?

Es ist wichtig, dass Studierende sich bewusst sind, dass der Einsatz von KI-basierten Schreibtools – wie bei anderen Hilfsmitteln auch – angegeben werden muss. Die Art und Weise der Angabe sollte in Absprache zwischen Lehrenden und Studierenden festgelegt werden. Mögliche Angaben umfassen den allgemeinen Einsatz des Tools, die spezifischen Funktionen (wie z.B. Strukturierungshilfen beim Schreiben einer Arbeit oder sprachliche und formale Überarbeitung) und sogar konkrete Angaben mit spezifischen Verweisen im Text, ähnlich wie bei Zitationsregeln. Als Entscheidungshilfe kann der Leitfaden «Aus KI zitieren» der Universität Basel genutzt werden.

Gibt es eine Selbständigkeitserklärung inklusive Verwendung für KI-basierte Hilfsmittel in der Lehre?

Der Rechtsdienst der Universität Bern hat Vorschläge erarbeitet, wie Selbstständigkeitserklärungen bei schriftlichen Arbeiten angepasst werden können. Dies betrifft die Fälle, in denen der Einsatz KI-gestützter Schreibtools erlaubt und nicht erlaubt sind.

Further R Ressources

R-specific search engine:
https://rseek.org/

- https://rweekly.org/ → Updates in R
- https://cran.r-project.org/web/packages/available_packages_by_name.html → information on packages
- Stack Overflow → Forum

Created with chatGPT

Google

- Search queries with “R” and package names if applicable
- In English
- Googling error messages

→ In English, change R to English

`Sys.setenv(LANGUAGE = "en")`

→ Program (R), function/package + error message

- Google search could be automated with the “errorist” package
- Information about the operating system can also be helpful

→ `sessionInfo()`

Reproducible data processing and analyses

- Comment steps of data processing and analyses → many formating options when using **Markdown** or **Quarto**
- Use meaningful variable and script names → **snake_case** recommended

camelCase

kebab-case

snake_case

https://juniortoexpert.com/de/namenskonvention-fur-variablen/

- Script(s) should document all steps from raw data to analyses

If files should be run in a particular order, prefix them with numbers. If it seems likely you'll have more than 10 files, left pad with zero:

```
00_download.R
01_explore.R
...
09_model.R
10_visualize.R
```

- Load all required packages at the beginning of a script
- Relative paths (e.g. data/datafile.csv) are better than absolute paths (Users/grinschs/Documents/r4ds/data/datafile.csv)

https://style.tidyverse.org/files.html

Reproducible data processing and analyses

- Open Science Practices:
  - Preregistration or Data Analyses Plan via www.osf.io
  - Open Data
  - Open Code

Codebook
see e.g., https://doi.org/10.1027/1015-5759/a000620

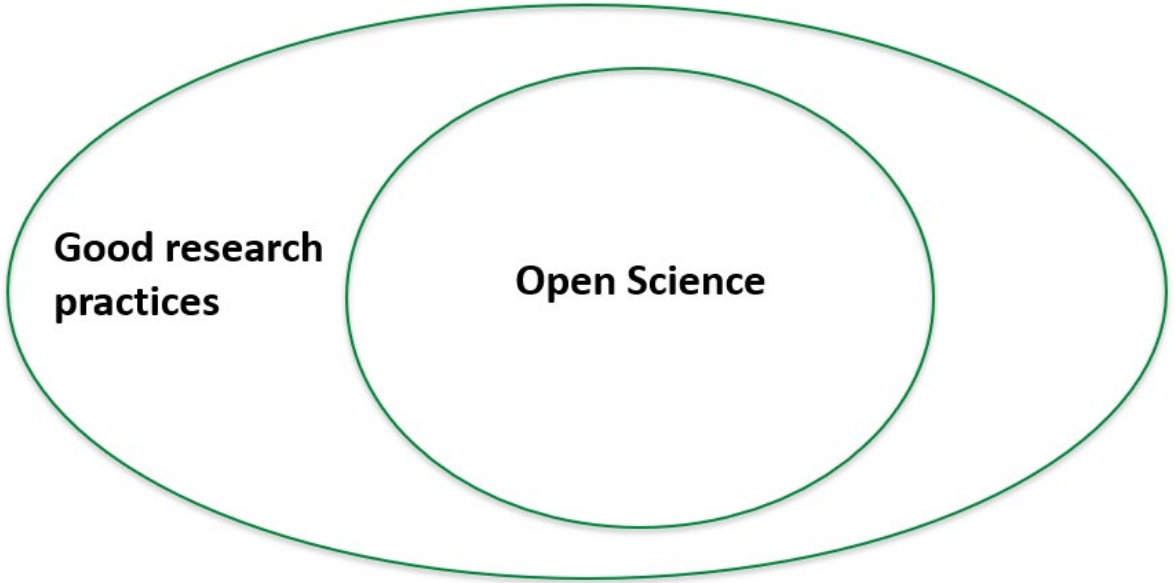


Figure: Stefan et al. (2023)

Final Tips

- Use LLMs as a “sparring partner,” not a calculator
- Validate with your own knowledge
- Good prompts heavily improve the output
- Make your data analyses as reproducible as possible

Contact: sandra.grinschgl@unibe.ch

Illustration from storyset/Freepik

Illustration von felicities/Freepik

Github CoPilot - Installation

- Prerequisite is an account and a CoPilot subscription (free for students of the University of Bern) on Github. Github is a cloud service from Microsoft that specializes in sharing and collaborating on code: https://github.com/

- Here is a guide to setting up Github CoPilot in RStudio: https://www.youtube.com/watch?v=t7NrkAeosog [1:40 – 2:36]

- Both the integration of CoPilot and ChatGPT are also described in the R Skript "Intro_CoPilot_ChatGPT.qmd"