

AI Tools for R

Day 1 - Introduction to Data Analysis with R

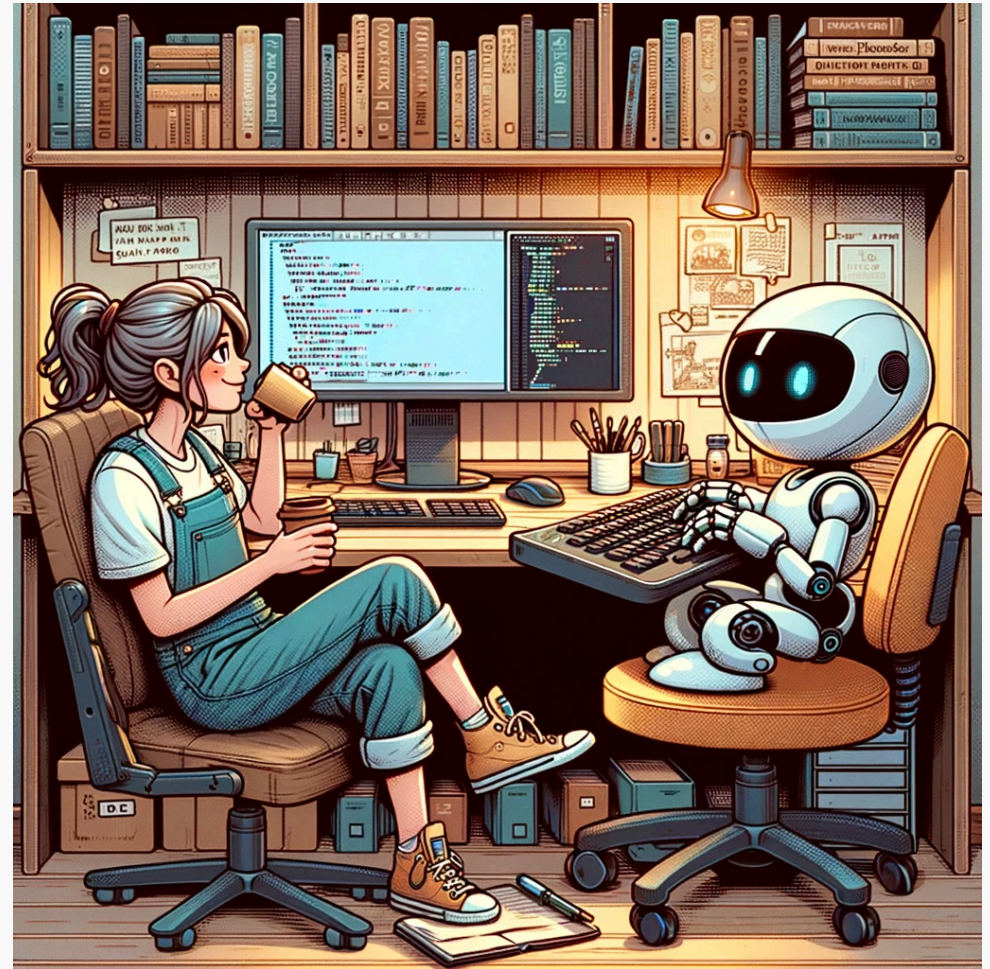
Selina Baldauf

Freie Universität Berlin - Theoretical Ecology

October 11, 2024

Motivation

- AI tools assist programmers with
 - Coding
 - Debugging
 - Learning
 - ...
- Higher productivity and efficiency
- More motivation
- But careful: You still need to understand what's going on!



Overview of tools

- Browser-based chat bots ([ChatGPT](#), [Bard](#), ...)
 - General-purpose
- Data-analysis tools ([Julius AI](#), [RTutor](#), ...)
 - Upload data and ask questions about it
 - Can also execute code
 - Chat with your data
- Integrated AI tools ([GitHub Copilot](#), [Codium AI](#), ...)
 - Integrated directly in programming environment
 - Real-time suggestions, chat, debugging, ...

Find the tools that best fit your workflow!

R Tutor

- <https://rtutor.ai/>
- Free browser tool
- Upload data and ask questions about it
- Use the demo data

The screenshot shows the R Tutor web interface. The top navigation bar includes 'RTutor', 'Home', 'EDA', 'Report', and 'More'. The main interface is divided into two panels. The left panel, titled '1. Select Dataset', '2. Modify Data Fields', and '3. Send Request', shows the 'mpg (examples)' dataset selected. The right panel, titled 'Selected Dataset', shows a table of 10 rows of data with 11 columns: maker, model, dis, year, cylinder, transmission, drive, city, highway, fuel, and type. The table is sorted by 'maker' and 'model'. The 'maker' column has values 'audi' and 'audi'. The 'model' column has values 'a4' and 'a4 quattro'. The 'dis' column has values 1.8, 2, 2.8, and 3.1. The 'year' column has values 1999 and 2008. The 'cylinder' column has values 4 and 6. The 'transmission' column has values 'auto(l5)', 'manual(m5)', and 'manual(m6)'. The 'drive' column has values 'f' and '4'. The 'city' column has values 16, 18, 20, and 21. The 'highway' column has values 25, 26, 27, 28, 29, and 31. The 'fuel' column has values 'p' and 'p'. The 'type' column has values 'compact' and 'compact'.

Quick start

RTutor
Data Science for All

Selected Dataset
234 rows X 11 columns

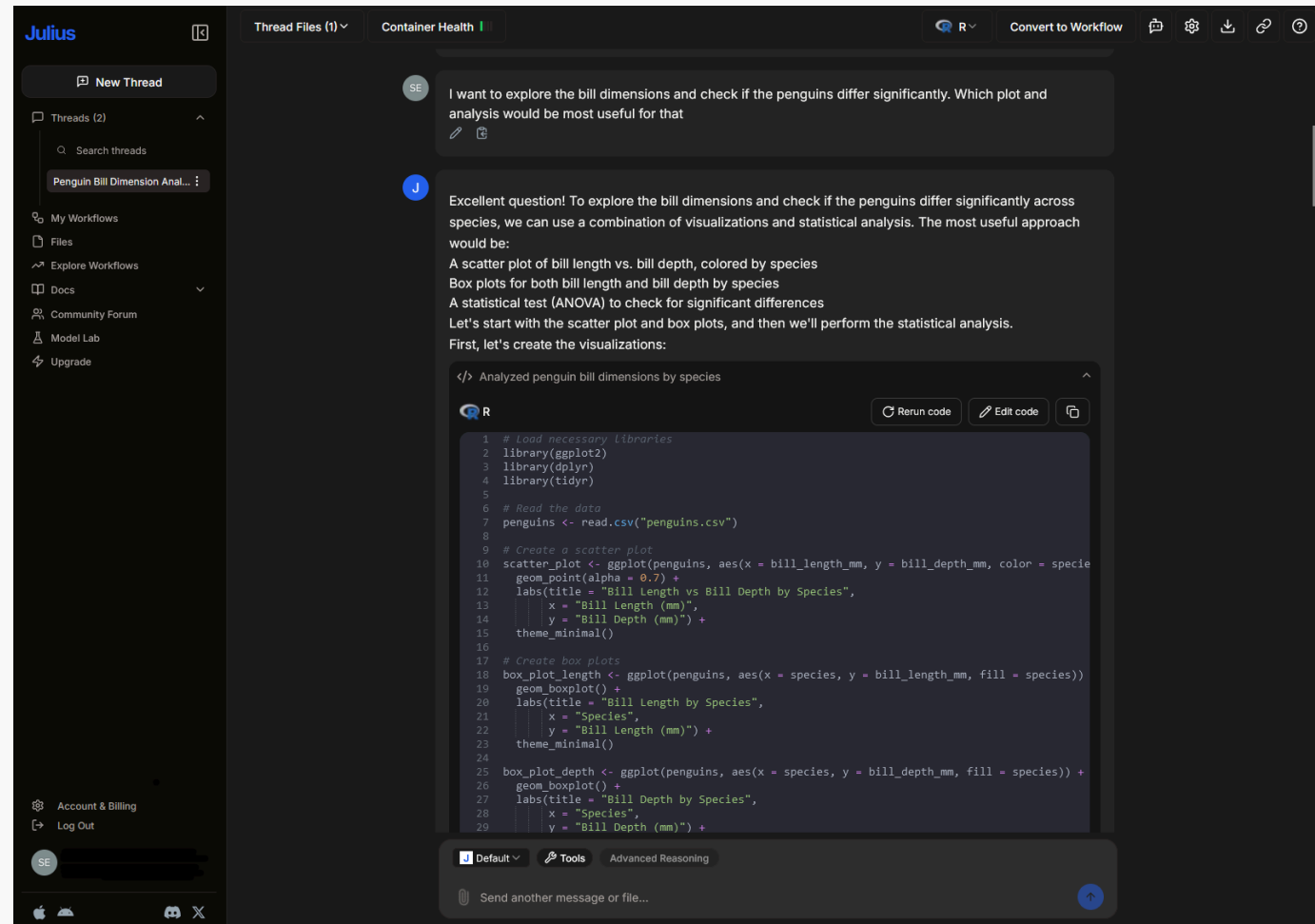
Search:

maker	model	dis	year	cylinder	transmission	drive	city	highway	fuel	type
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2	2008	4	manual(m6)	4	20	28	p	compact

Previous 1 2 3 4 5 ... 24 Next

Julius AI

- <https://julius.ai/>
- Try for free
- Basic plan ~20€ per months (-50% academic discount)
- Upload data and ask questions about it



Github Copilot

- <https://github.com/features/copilot>
- Model based on GPT-4 and OpenAI's Codex
 - Specifically trained on source code
- Basic idea: Integrate directly into R Studio (or other IDEs)
- Works best for well-represented languages (Python, JS, ...)

How to get GitHub Copilot

See [this website](#) for step-by-step guide and more information.

It's really easy, but you need:

- GitHub Account
- Active GH Copilot subscription (10\$ per month)
 - Get it for free as an academic with an educational account
- IDE that supports Copilot
 - Full support: Visual Studio (Code), Vim, Neovim, JetBrains IDEs (e.g. PyCharm)
 - Limited support: RStudio, ?

GH Copilot: Inline code suggestions

- Copilot tries to predict what you want to do next
- Suggestions are based on the context
 - Previous code
 - Comments
 - Variable and function names
 - ...

```
fibonacci.R > fibonacci
1  fibonacci <- function(n) {
2    if (n == 0) {
3      return(0)
4    } else if (n == 1) {
5      return(1)
6    } else {
7      return(fibonacci(n - 1) + fibonacci(n - 2))
8    }
9  }
```


Get better suggestions

- **Provide context**
 - Open other files
 - Add top level comments explaining the purpose of the script
 - Name variables and functions properly
 - Copy-paste sample code and delete it later
- **Be consistent**
 - “Garbage in, garbage out”
 - Have a nice and consistent coding style

Nice side effect of using Copilot: More good-practice coding

Chat with GH copilot in R Studio

- Available through the `chattr` package
- Chat with Copilot in the sidebar
- Also supports other LLMs (e.g. GPT4o, ...)

Concerns to consider

- Privacy
 - Chose whether your prompts and suggestions will be used by Github (**Github -> Seetings -> Copilot -> Policies**)
 - Check privacy guidelines before you upload data
- Plagiarism
 - Block suggestions matching public code (**Github -> Seetings -> Copilot -> Policies**)
- Ethical concerns
 - For-profit tool trained on open-source
- Environmental concerns
 - Water and energy usage

Usage guidelines

- No definite guidelines, but see examples [listed here](#)
- **Responsibility**
 - You are responsible for your scientific output
 - Stay critical, double-check
- **Transparency**
 - Make clear for which tasks you used which AI
- **Know relevant guidelines**
 - Journals
 - Your university
- **Still understand what is happening!**

