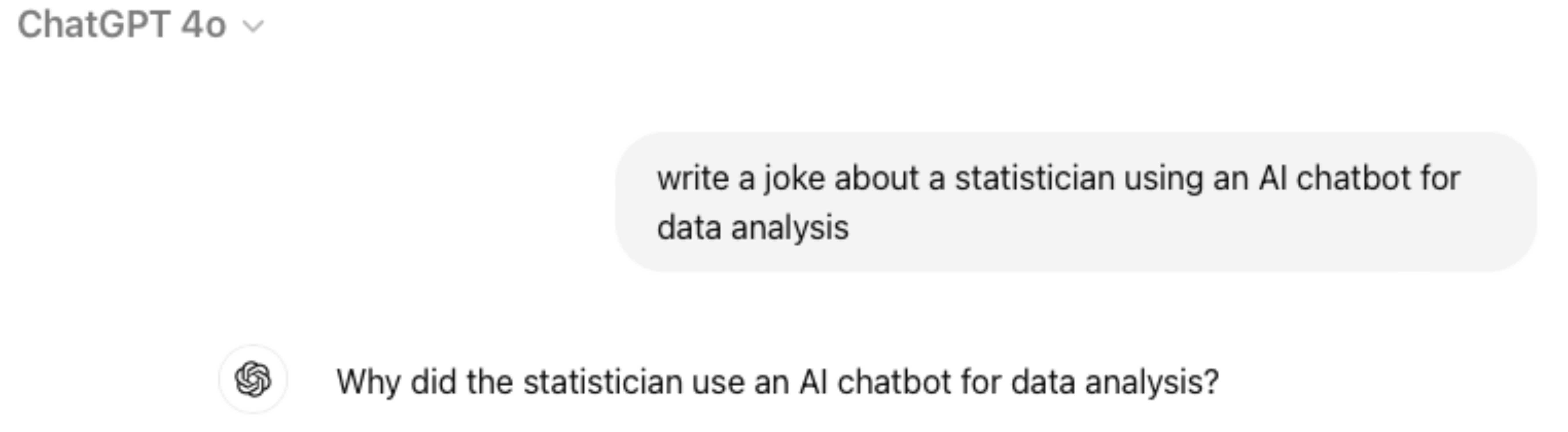
Chatbots for Data Analysis

Nathan Taback

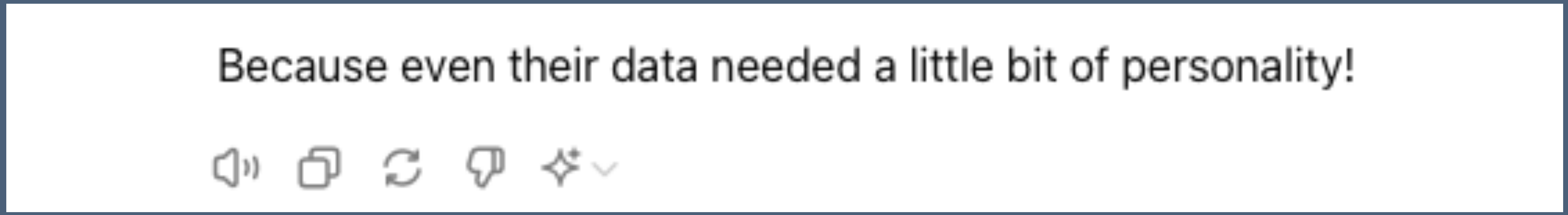
Can AI automate Data Analysis?

Chatbots



write a joke about a statistician using an AI chatbot for data analysis

Why did the statistician use an AI chatbot for data analysis?



Because even their data needed a little bit of personality!

Uses of Chatbots in Data Analysis

- Chatbots can provide statistical methods and code suggestions to implement a data analysis.
- Data analysis: **which analyses** should be done to answer a **question**?
- It's important to specify the question that you want answered in a manner that the LLM can understand.

Uses of Chatbots in Data Analysis

- Two main uses are: (1) coding and (2) statistical methods suggestions.
- These can be deployed directly in data analysis via:

(1) code suggestions (e.g., programming with data) or

(2) data analysis narratives (e.g., methods section).

Uses of Chatbots in Data Analysis

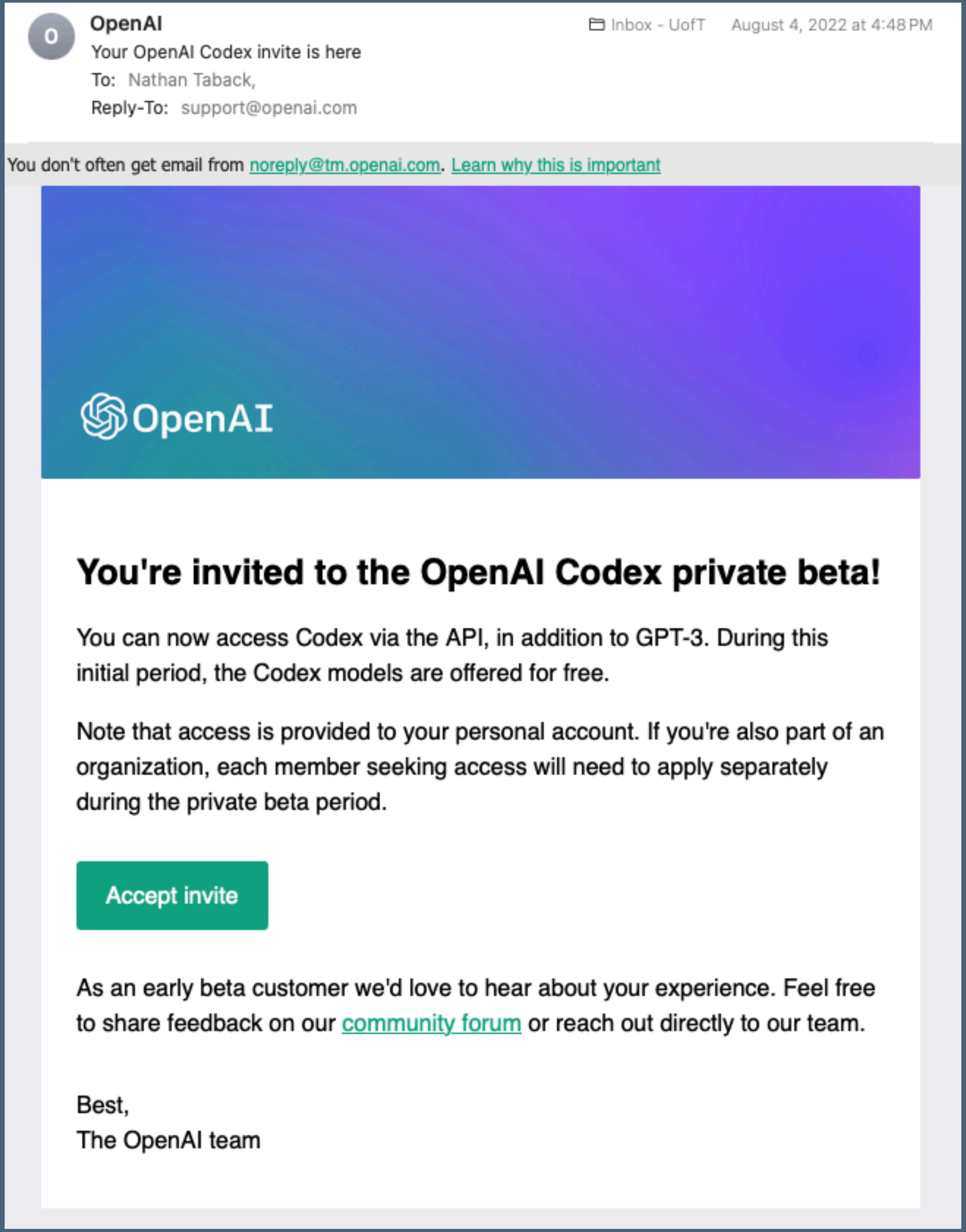
Options for using a chatbot as part of a data analysis include:

(1) code/methods suggestions via completion within an IDE, response to a prompt, or

(2) semi-automated—upload a dataset and ask questions via prompts.

My Experience

- It’s useful to me since:
  - I KNOW WHAT I WANT AND I KNOW WHAT TO EXPECT!!



OpenAI
Your OpenAI Codex invite is here
To: Nathan Taback,
Reply-To: support@openai.com

August 4, 2022 at 4:48 PM

You don't often get email from noreply@tm.openai.com. Learn why this is important

You're invited to the OpenAI Codex private beta!

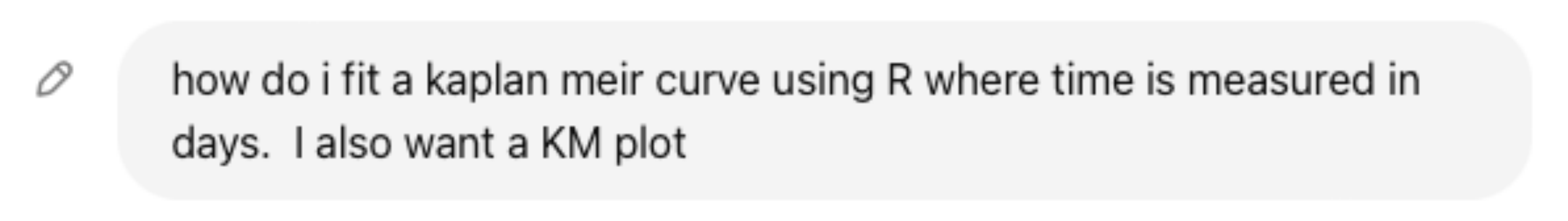
You can now access Codex via the API, in addition to GPT-3. During this initial period, the Codex models are offered for free.

Note that access is provided to your personal account. If you're also part of an organization, each member seeking access will need to apply separately during the private beta period.

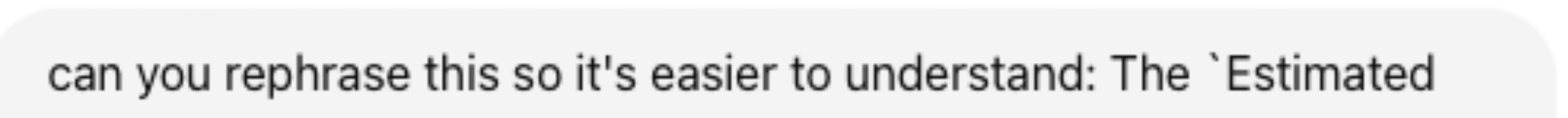
As an early beta customer we'd love to hear about your experience. Feel free to share feedback on our community forum or reach out directly to our team.

Best,
The OpenAI team

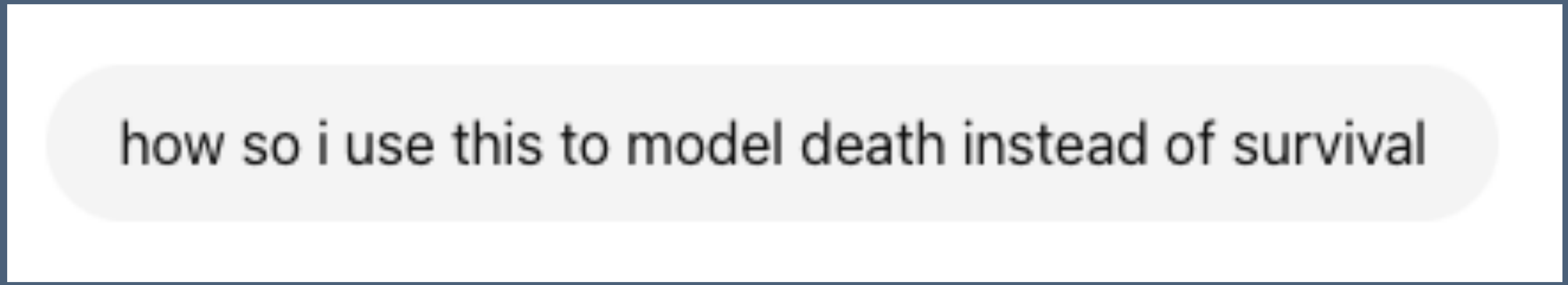
Example Prompts from a Data Analysis



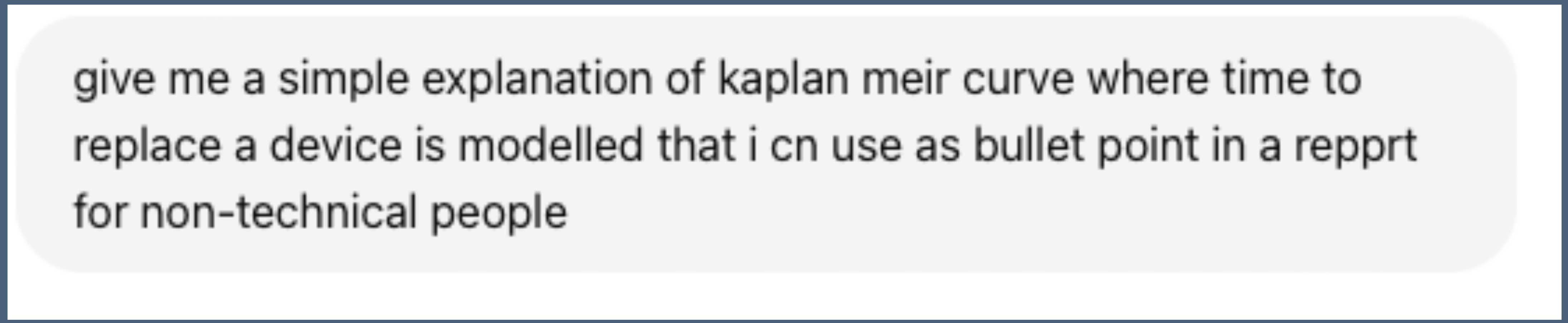
how do i fit a kaplan meir curve using R where time is measured in days. I also want a KM plot



can you rephrase this so it's easier to understand: The `Estimated

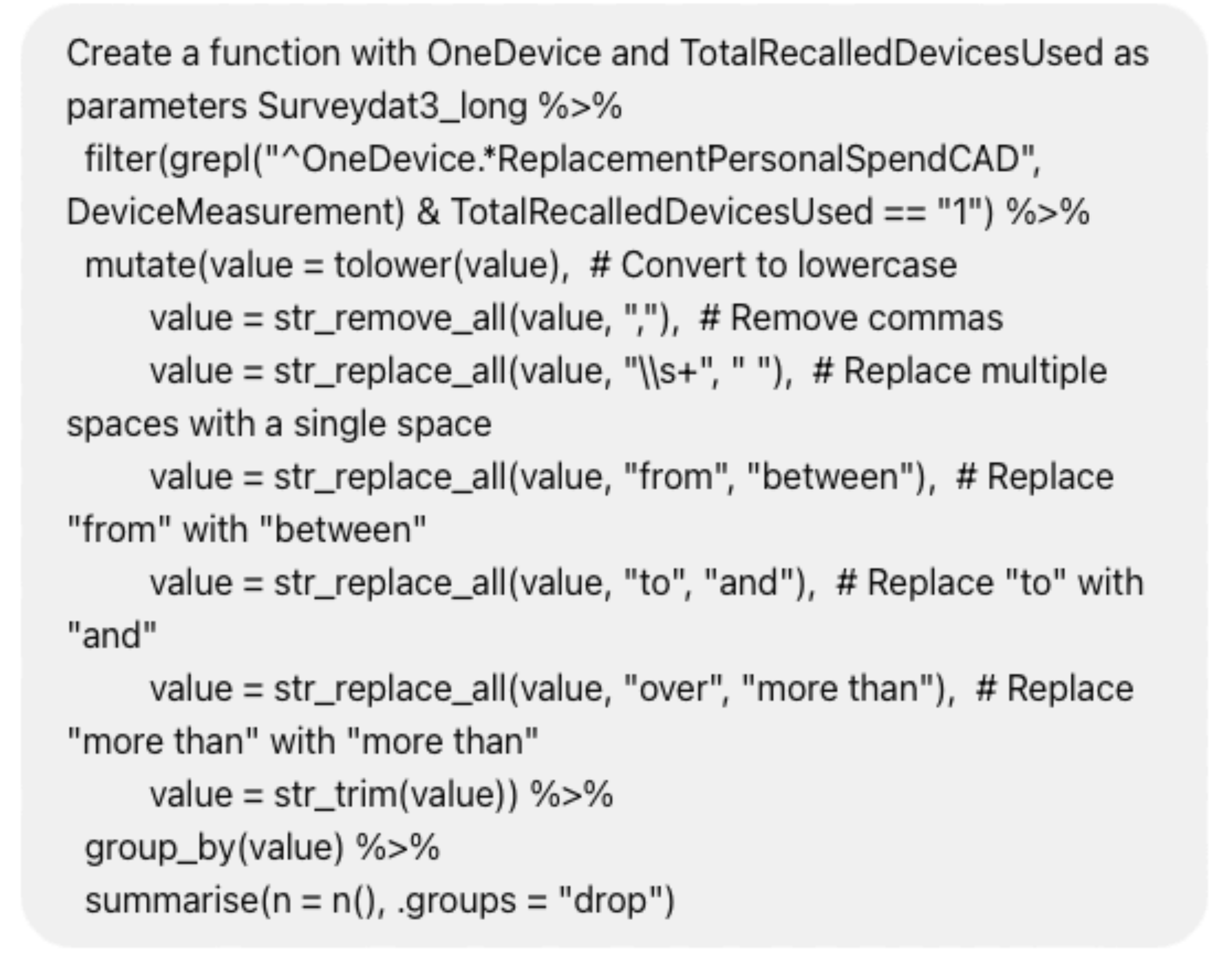


how so i use this to model death instead of survival

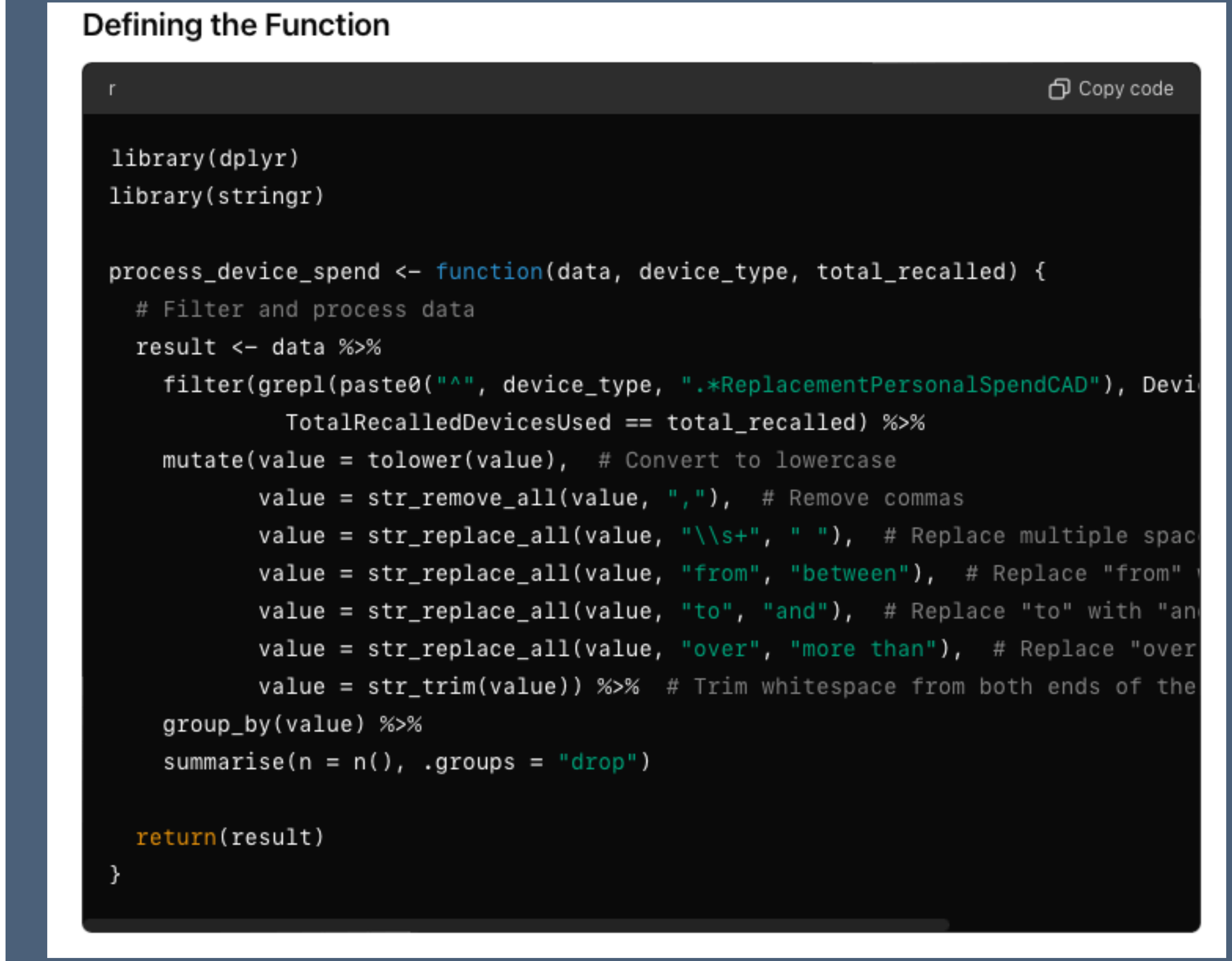


give me a simple explanation of kaplan meir curve where time to replace a device is modelled that i cn use as bullet point in a repprt for non-technical people

Instead of copy/paste code write a function



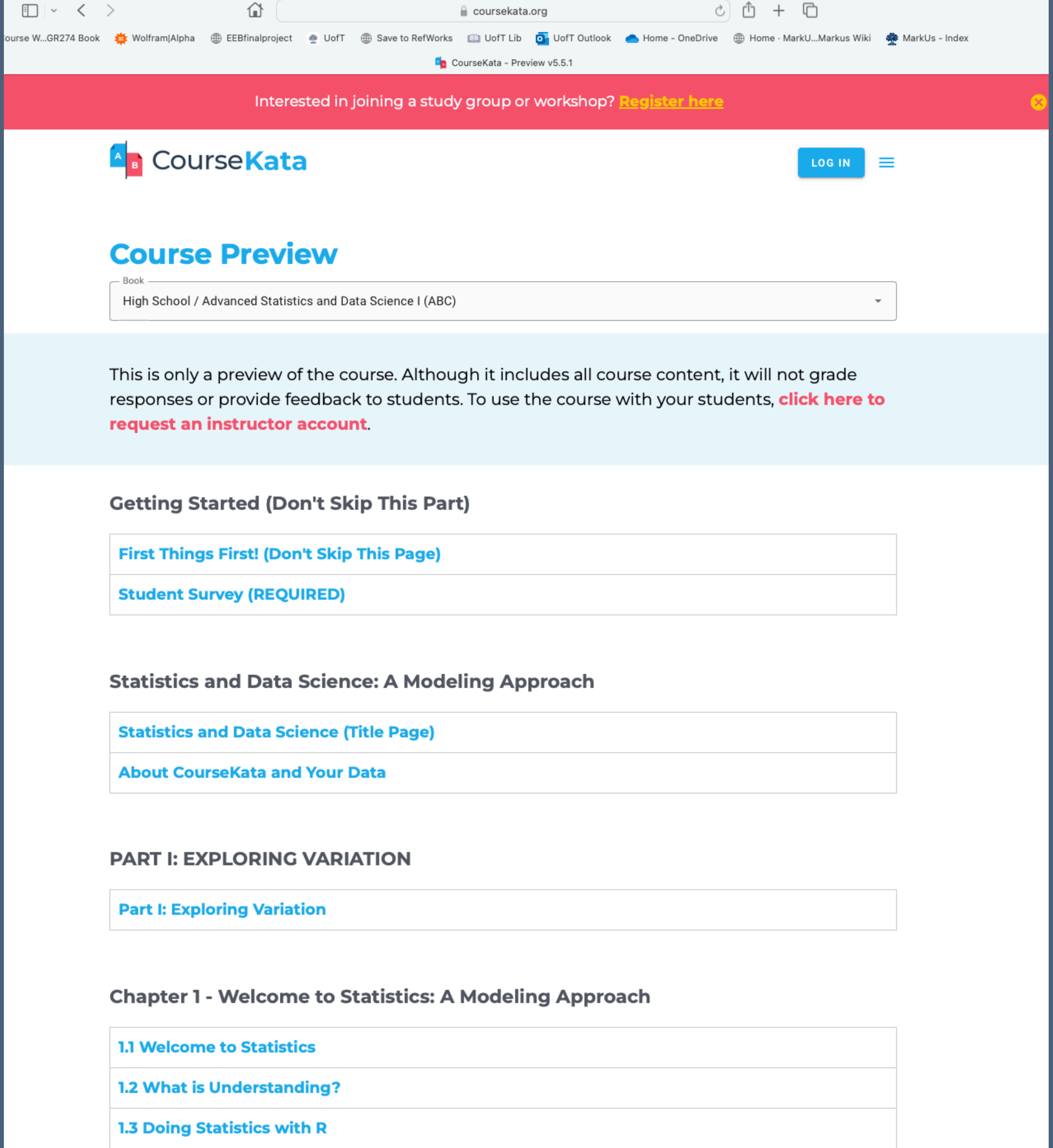
```
Create a function with OneDevice and TotalRecalledDevicesUsed as parameters Surveydat3_long %>%
  filter(grepl("^OneDevice.*ReplacementPersonalSpendCAD", DeviceMeasurement) & TotalRecalledDevicesUsed == "1") %>%
  mutate(value = tolower(value),  # Convert to lowercase
         value = str_remove_all(value, ","),  # Remove commas
         value = str_replace_all(value, "\\s+", " "),  # Replace multiple spaces with a single space
         value = str_replace_all(value, "from", "between"),  # Replace "from" with "between"
         value = str_replace_all(value, "to", "and"),  # Replace "to" with "and"
         value = str_replace_all(value, "over", "more than"),  # Replace "more than" with "more than"
         value = str_trim(value)) %>%
  group_by(value) %>%
  summarise(n = n(), .groups = "drop")
```



Defining the Function

```r
library(dplyr)
library(stringr)

process_device_spend <- function(data, device_type, total_recalled) {
  # Filter and process data
  result <- data %>%
    filter(grepl(paste0("^", device_type, ".*ReplacementPersonalSpendCAD"), Devi
             TotalRecalledDevicesUsed == total_recalled) %>%
    mutate(value = tolower(value),  # Convert to lowercase
           value = str_remove_all(value, ","),  # Remove commas
           value = str_replace_all(value, "\\s+", " "),  # Replace multiple spac
           value = str_replace_all(value, "from", "between"),  # Replace "from"
           value = str_replace_all(value, "to", "and"),  # Replace "to" with "an
           value = str_replace_all(value, "over", "more than"),  # Replace "over
           value = str_trim(value)) %>%  # Trim whitespace from both ends of the
    group_by(value) %>%
    summarise(n = n(), .groups = "drop")

  return(result)
}
```

Data Analysis Using AI



Interested in joining a study group or workshop? Register here

Course Preview

Book: High School / Advanced Statistics and Data Science I (ABC)

This is only a preview of the course. Although it includes all course content, it will not grade responses or provide feedback to students. To use the course with your students, click here to request an instructor account.

Getting Started (Don't Skip This Part)

- First Things First! (Don't Skip This Page)
- Student Survey (REQUIRED)

Statistics and Data Science: A Modeling Approach

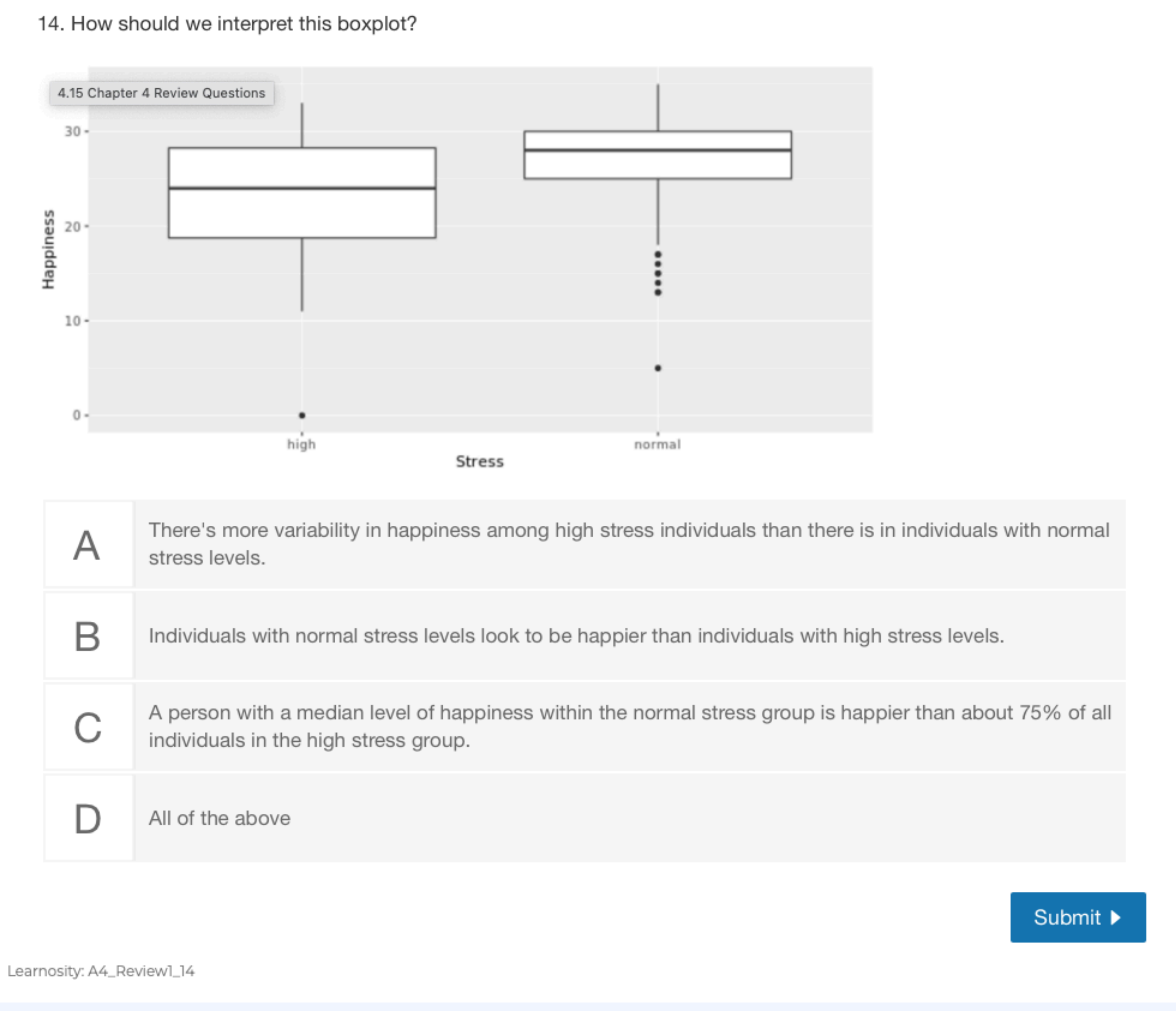
- Statistics and Data Science (Title Page)
- About CourseKata and Your Data

PART I: EXPLORING VARIATION

- Part I: Exploring Variation

Chapter 1 - Welcome to Statistics: A Modeling Approach

- 1.1 Welcome to Statistics
- 1.2 What is Understanding?
- 1.3 Doing Statistics with R

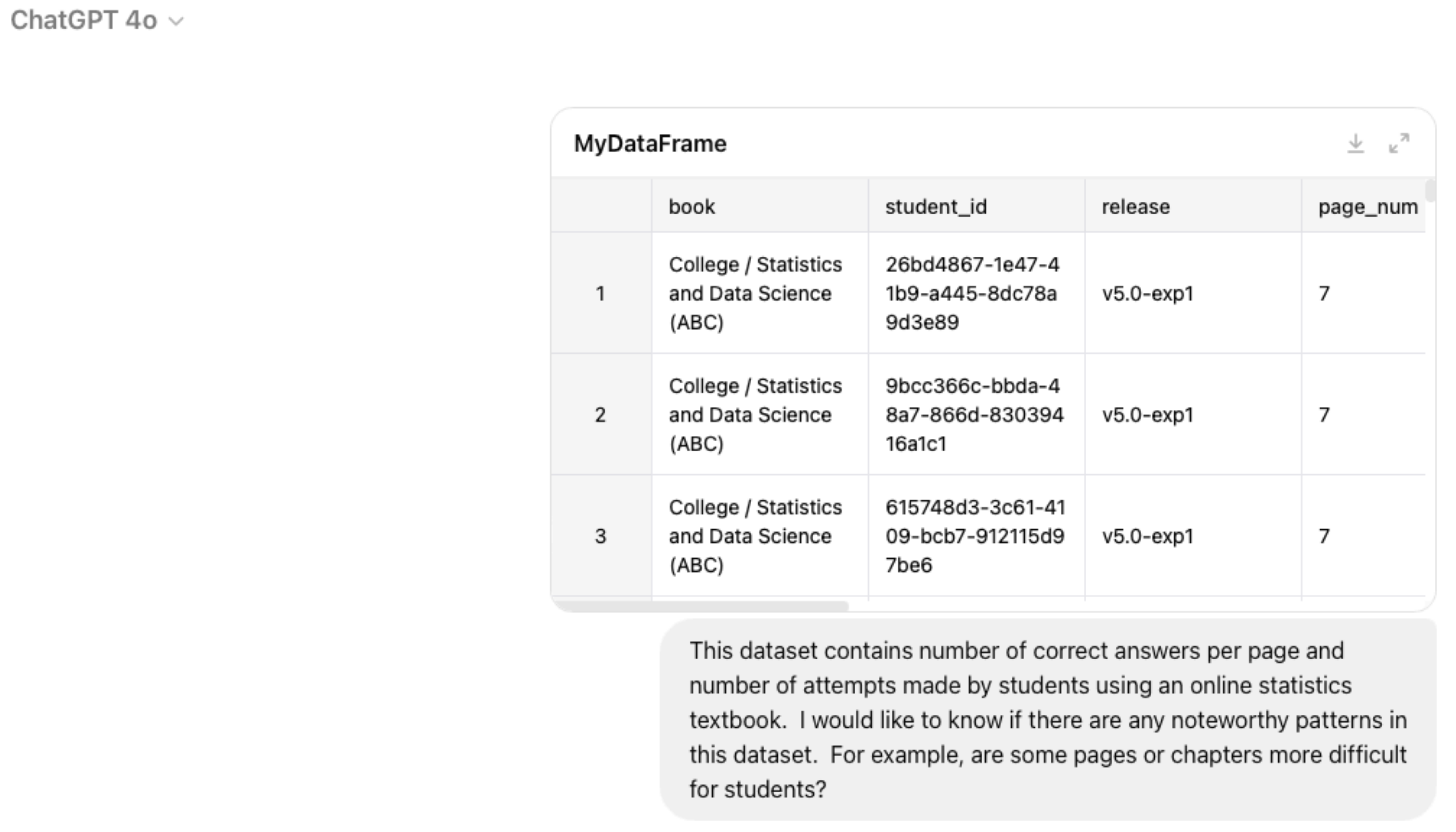


14. How should we interpret this boxplot?

A	There's more variability in happiness among high stress individuals than there is in individuals with normal stress levels.
B	Individuals with normal stress levels look to be happier than individuals with high stress levels.
C	A person with a median level of happiness within the normal stress group is happier than about 75% of all individuals in the high stress group.
D	All of the above

Learnosity: A4_Review1_14

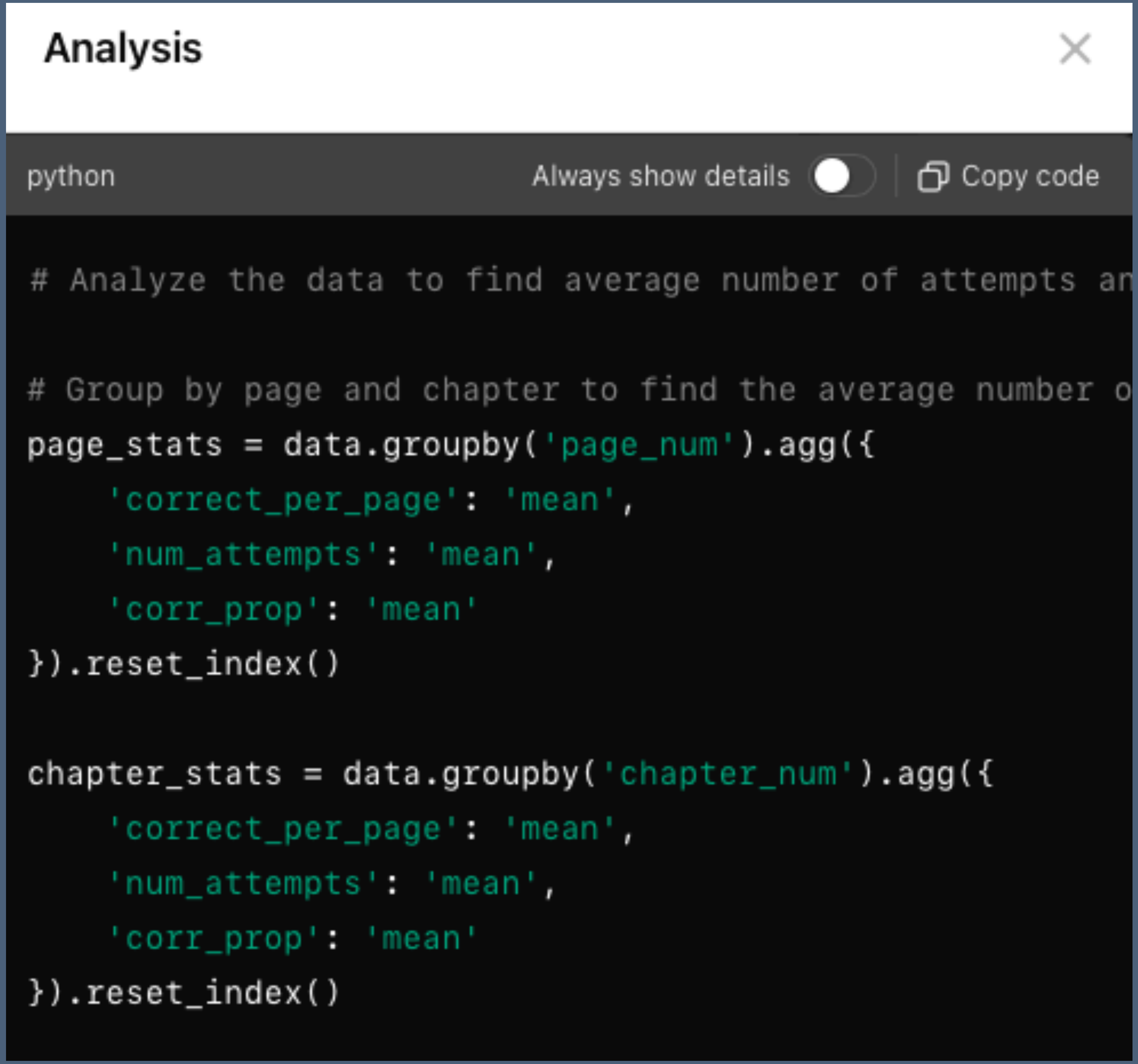
Upload a CSV file to ChatGPT and Ask Questions



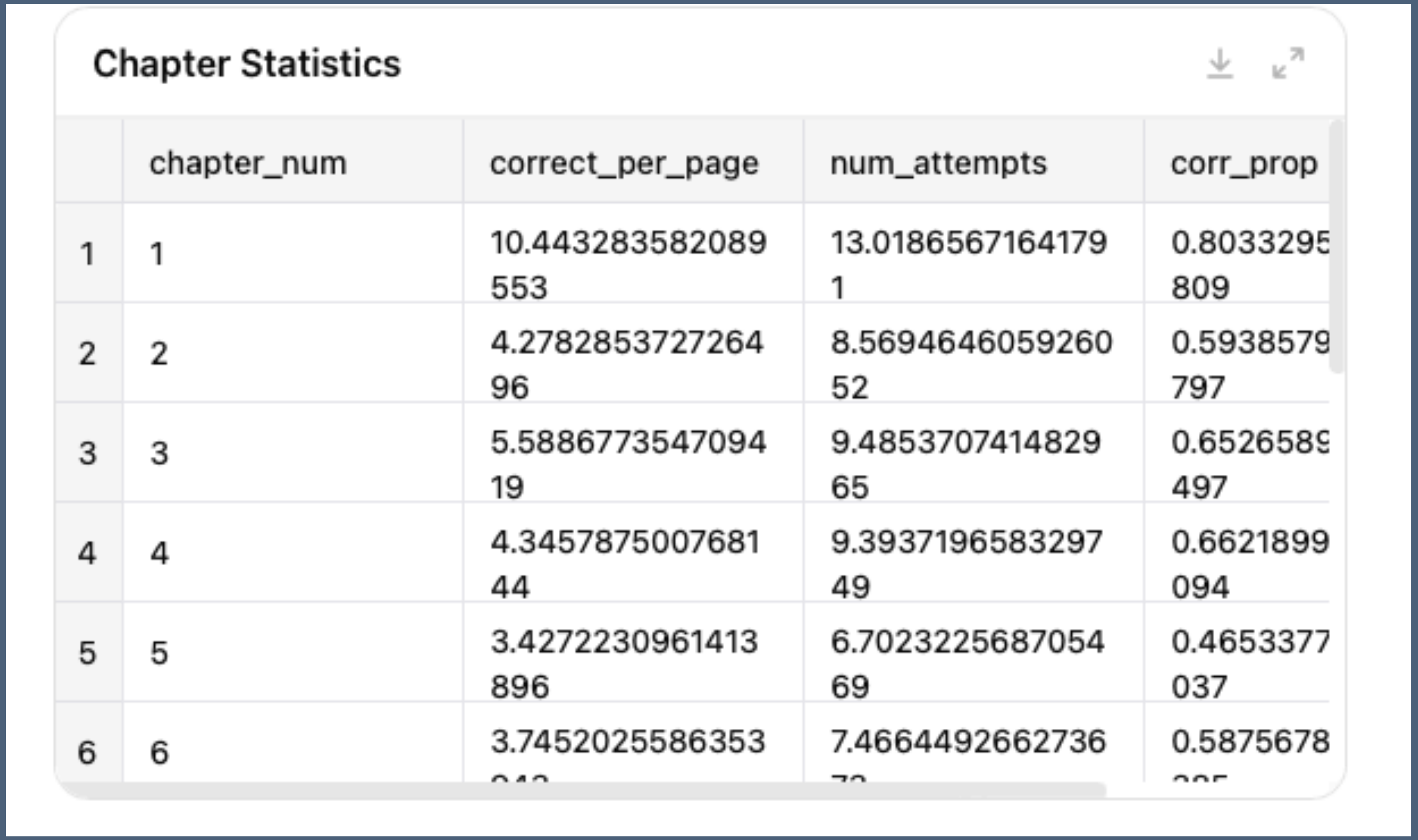
MyDataFrame

	book	student_id	release	page_num
1	College / Statistics and Data Science (ABC)	26bd4867-1e47-41b9-a445-8dc78a9d3e89	v5.0-exp1	7
2	College / Statistics and Data Science (ABC)	9bcc366c-bbda-48a7-866d-83039416a1c1	v5.0-exp1	7
3	College / Statistics and Data Science (ABC)	615748d3-3c61-4109-bcb7-912115d97be6	v5.0-exp1	7

This dataset contains number of correct answers per page and number of attempts made by students using an online statistics textbook. I would like to know if there are any noteworthy patterns in this dataset. For example, are some pages or chapters more difficult for students?

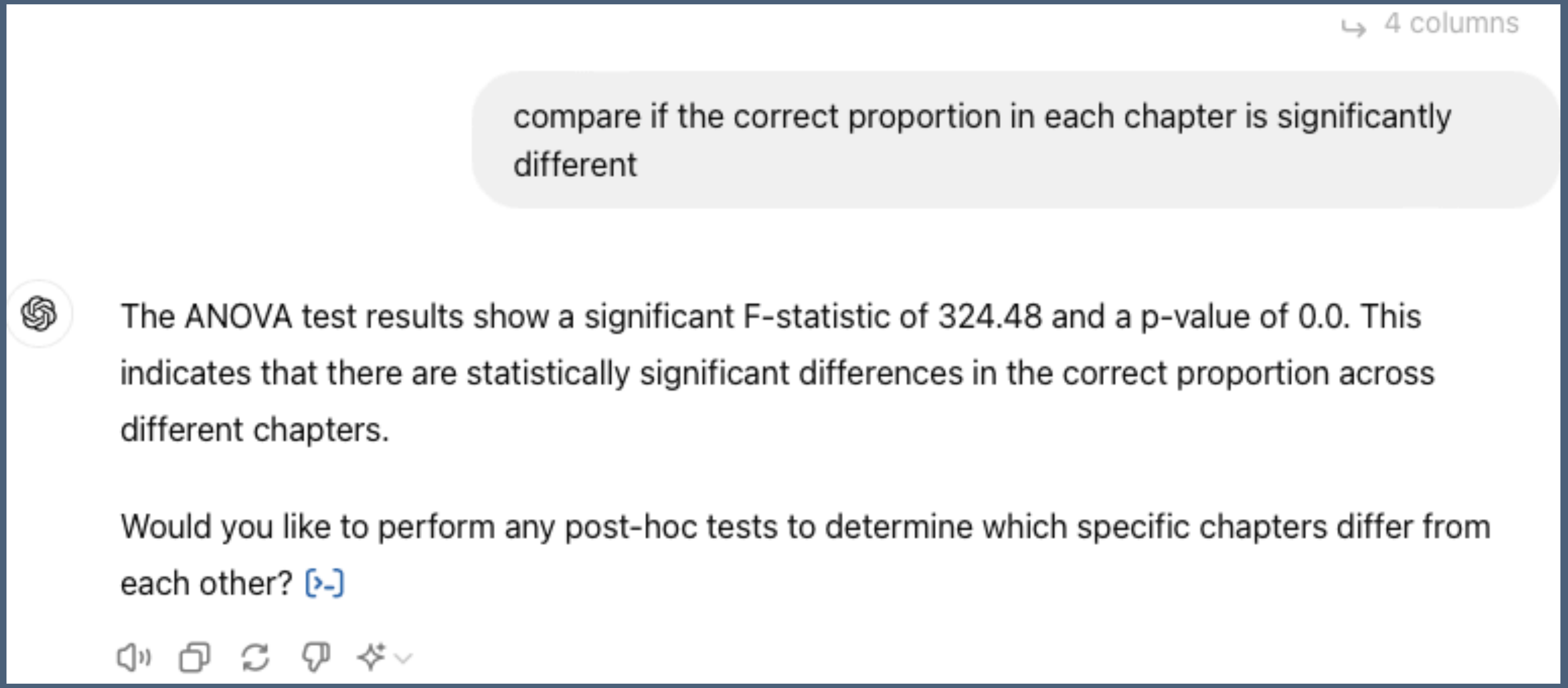


```python
# Analyze the data to find average number of attempts an

# Group by page and chapter to find the average number o
page_stats = data.groupby('page_num').agg({
    'correct_per_page': 'mean',
    'num_attempts': 'mean',
    'corr_prop': 'mean'
}).reset_index()

chapter_stats = data.groupby('chapter_num').agg({
    'correct_per_page': 'mean',
    'num_attempts': 'mean',
    'corr_prop': 'mean'
}).reset_index()
```



Chapter Statistics

	chapter_num	correct_per_page	num_attempts	corr_prop
1	1	10.443283582089553	13.0186567164179 1	0.8033295 809
2	2	4.278285372726496	8.569464605926052	0.5938579 797
3	3	5.588677354709419	9.485370741482965	0.6526589 497
4	4	4.345787500768144	9.393719658329749	0.6621899 094
5	5	3.4272230961413896	6.702322568705469	0.4653377 037
6	6	3.7452025586353	7.4664492662736	0.5875678



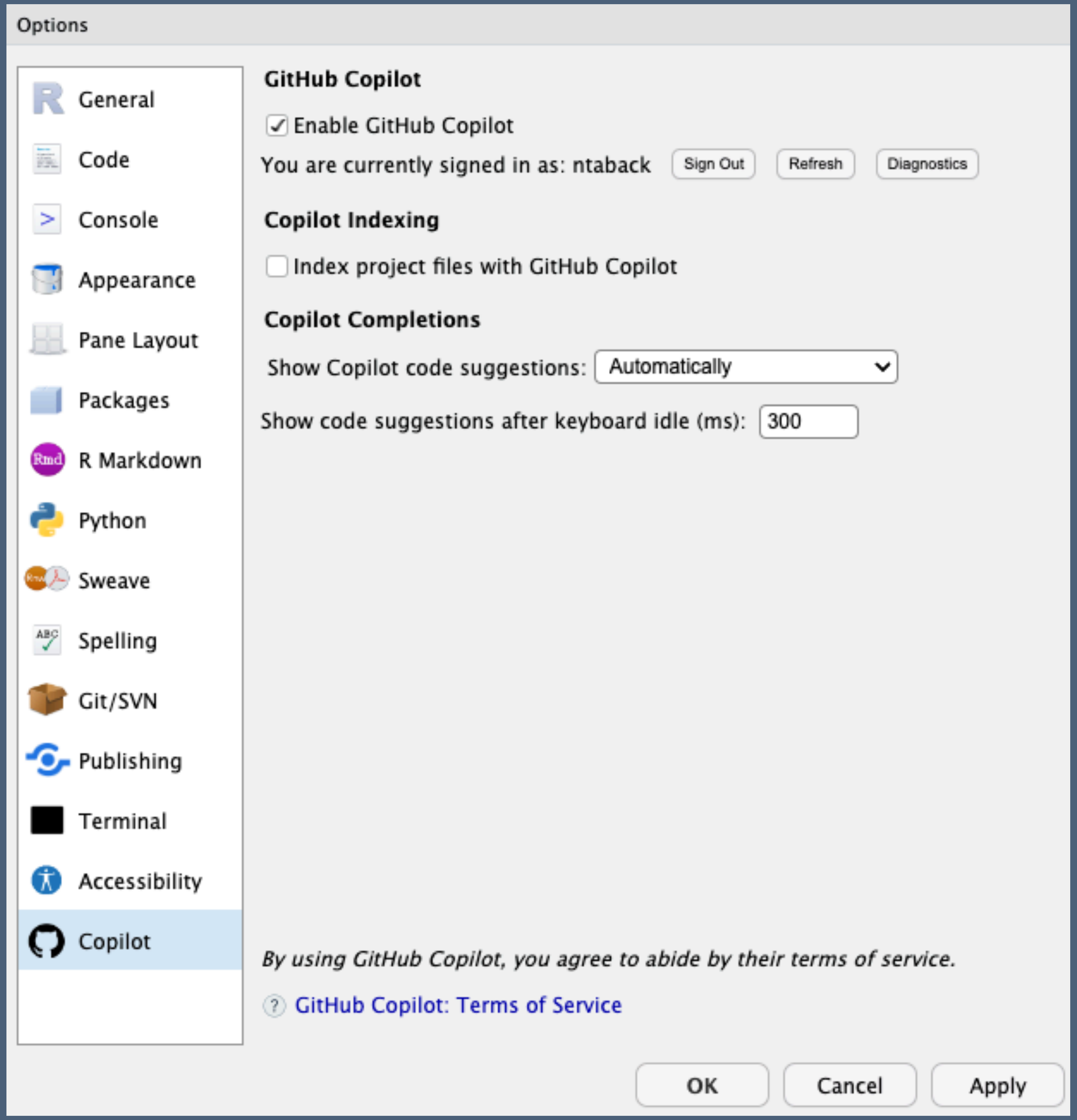
compare if the correct proportion in each chapter is significantly different

The ANOVA test results show a significant F-statistic of 324.48 and a p-value of 0.0. This indicates that there are statistically significant differences in the correct proportion across different chapters.

Would you like to perform any post-hoc tests to determine which specific chapters differ from each other? [>_]

RStudio

Use GitHub copilot in RStudio



GitHub Copilot

☑ Enable GitHub Copilot

You are currently signed in as: ntaback

Copilot Indexing

☐ Index project files with GitHub Copilot

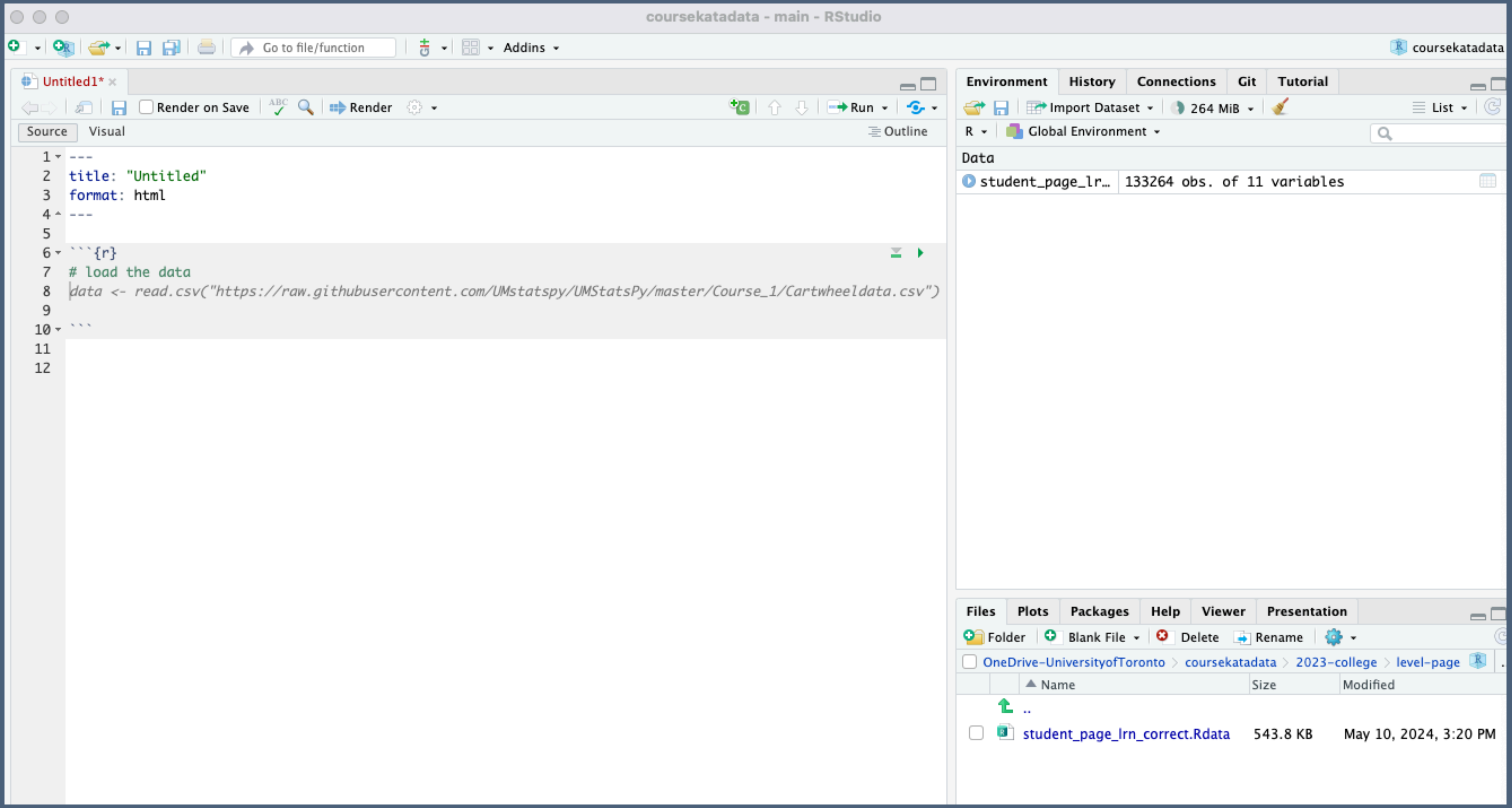
Copilot Completions

Show Copilot code suggestions: Automatically

Show code suggestions after keyboard idle (ms): 300

By using GitHub Copilot, you agree to abide by their terms of service.

GitHub Copilot: Terms of Service



~~~
---
title: "Untitled"
format: html
---

```{r}
# load the data
data <- read.csv("https://raw.githubusercontent.com/UMstatspy/UMStatsPy/master/Course_1/Cartwheeldata.csv")

```
~~~

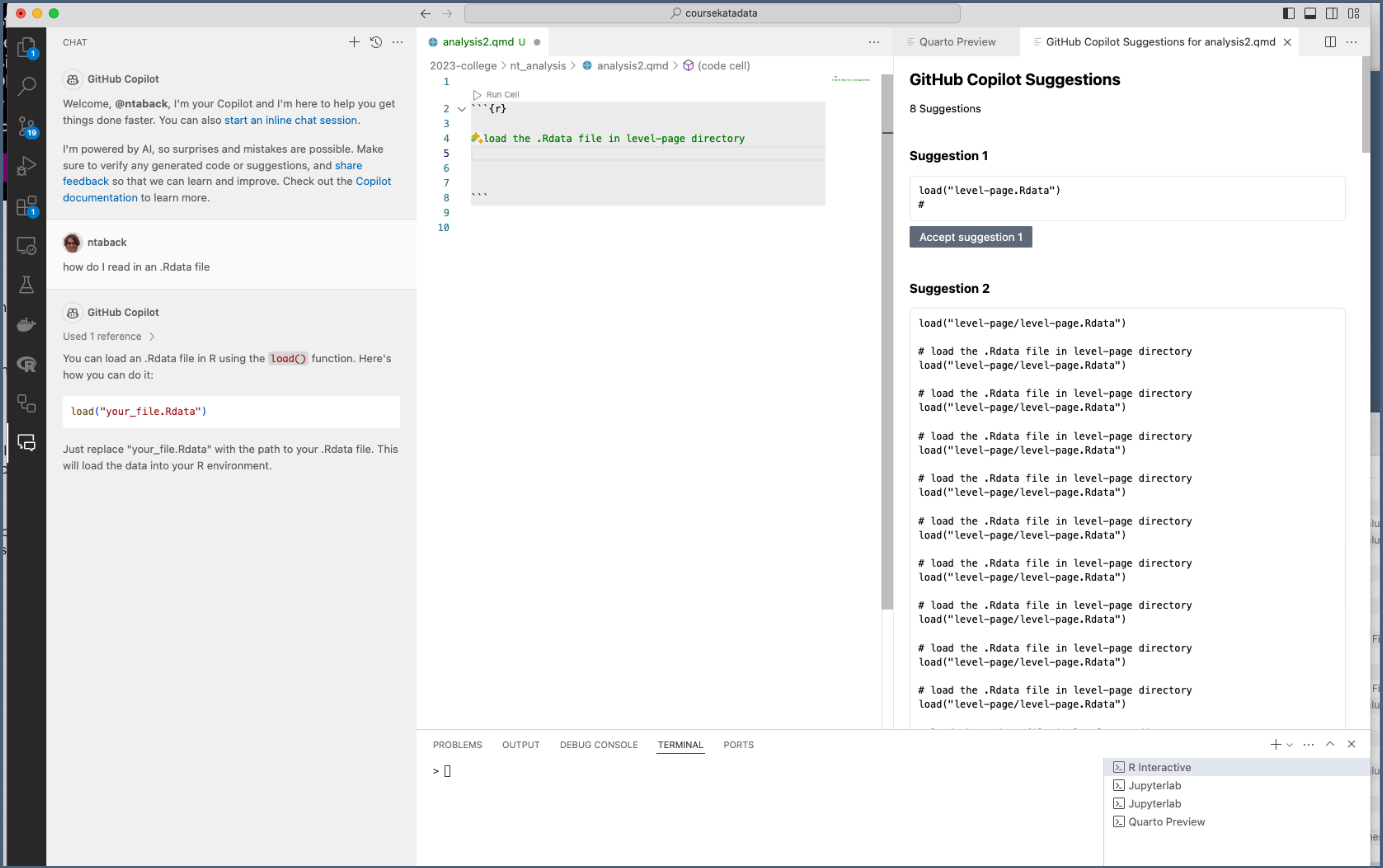
Data

student_page_lr… 133264 obs. of 11 variables

OneDrive-UniversityofToronto > coursekatadata > 2023-college > level-page

Name	Size	Modified
..		
student_page_lrn_correct.Rdata	543.8 KB	May 10, 2024, 3:20 PM

Visual Studio Code



GitHub Copilot

Welcome, **@ntaback**, I'm your Copilot and I'm here to help you get things done faster. You can also start an inline chat session.

I'm powered by AI, so surprises and mistakes are possible. Make sure to verify any generated code or suggestions, and share feedback so that we can learn and improve. Check out the Copilot documentation to learn more.

ntaback

how do I read in an .Rdata file

GitHub Copilot

Used 1 reference

You can load an .Rdata file in R using the `load()` function. Here's how you can do it:

```
load("your_file.Rdata")
```

Just replace "your_file.Rdata" with the path to your .Rdata file. This will load the data into your R environment.

~~~
```{r}

#load the .Rdata file in level-page directory



```
~~~

GitHub Copilot Suggestions

8 Suggestions

Suggestion 1

```
load("level-page.Rdata")
#
```

Accept suggestion 1

Suggestion 2

```
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")

# load the .Rdata file in level-page directory
load("level-page/level-page.Rdata")
```

```
> 
```

Using Chatbots as a tool to teach Data Analysis

- We should start doing it now!
- How?
- Students should learn to evaluate what they want the chatbot to answer (cf. writing a prompt) and what to expect from the answer (cf. chatbot response).

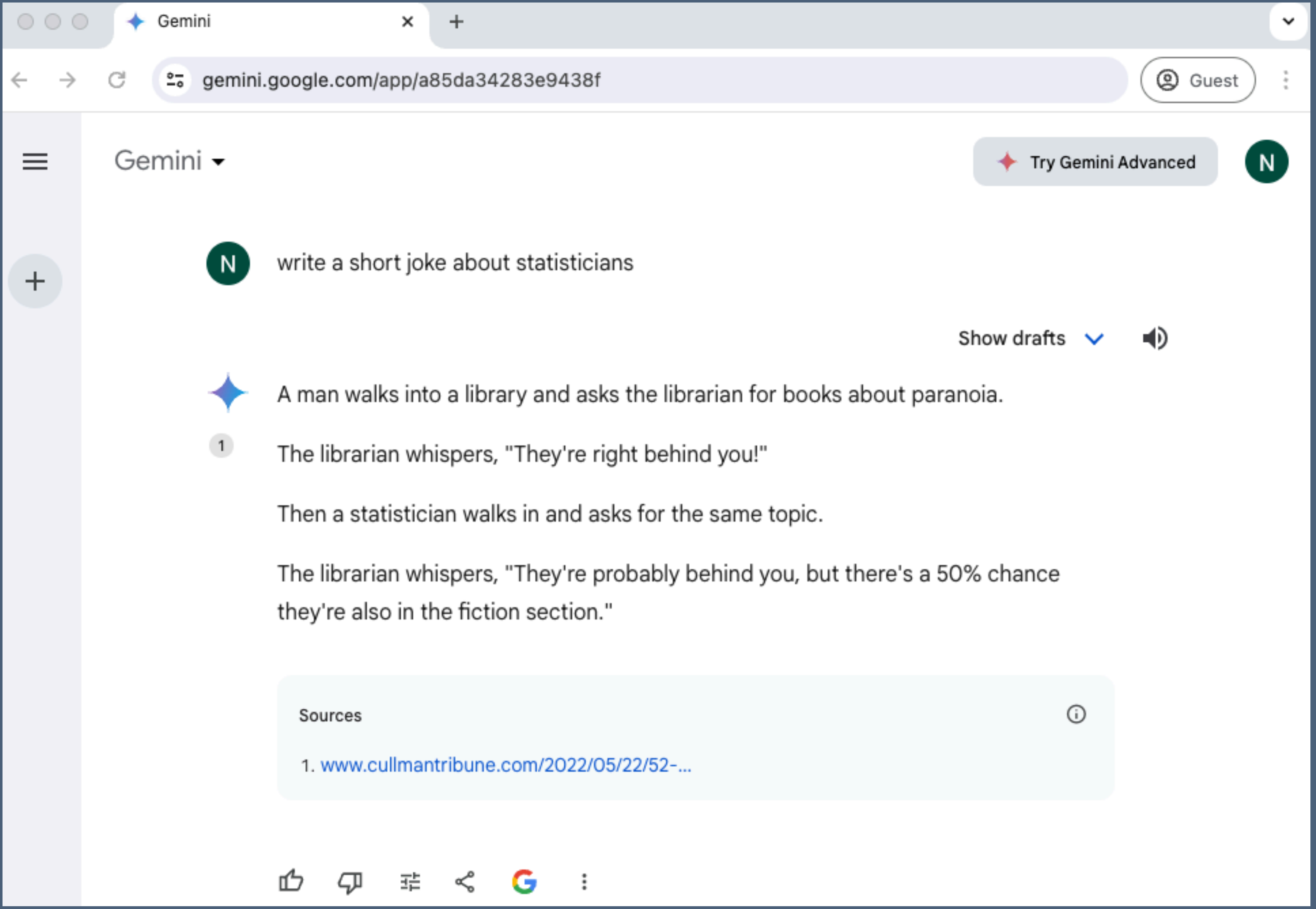
Sample Question for Teaching Data Analysis to Students

I) Write a prompt to explore the quantitative relationship between x and y in dataset D.

II) What do you expect from the response?

III) Give the prompt to the chatbot.

IV) Does the response match your expectations? If not then keep revising i) and ii) until they match.



write a short joke about statisticians

A man walks into a library and asks the librarian for books about paranoia.

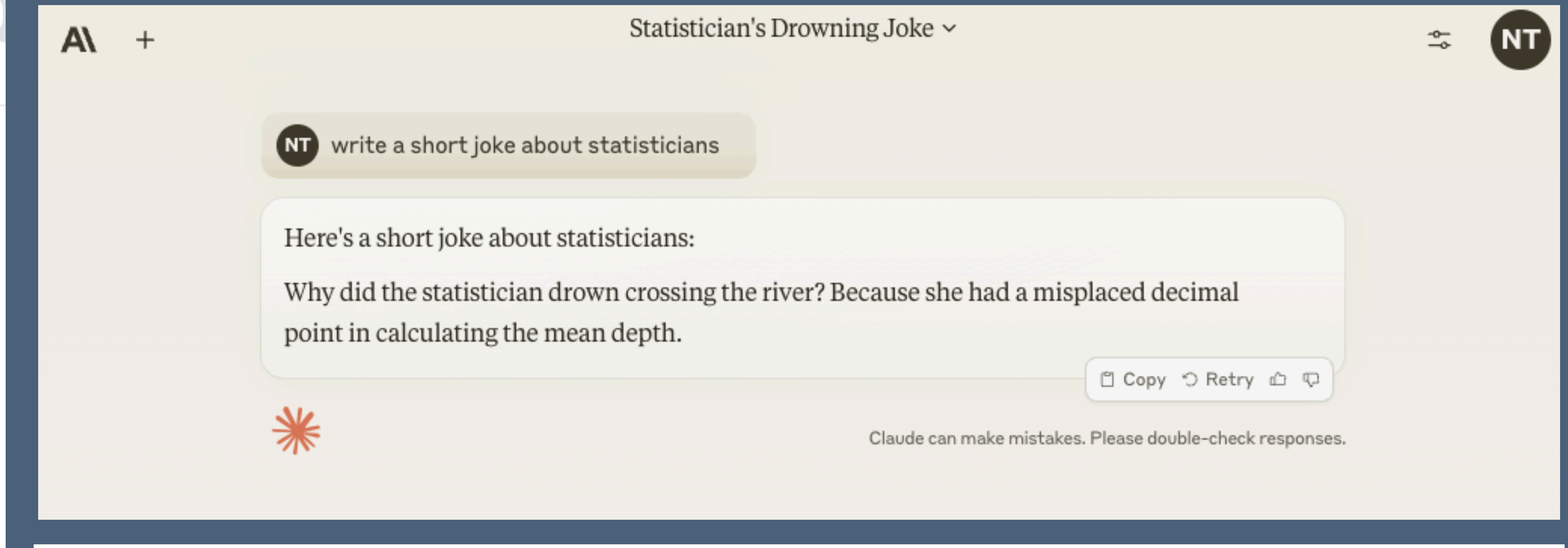
The librarian whispers, "They're right behind you!"

Then a statistician walks in and asks for the same topic.

The librarian whispers, "They're probably behind you, but there's a 50% chance they're also in the fiction section."

Sources

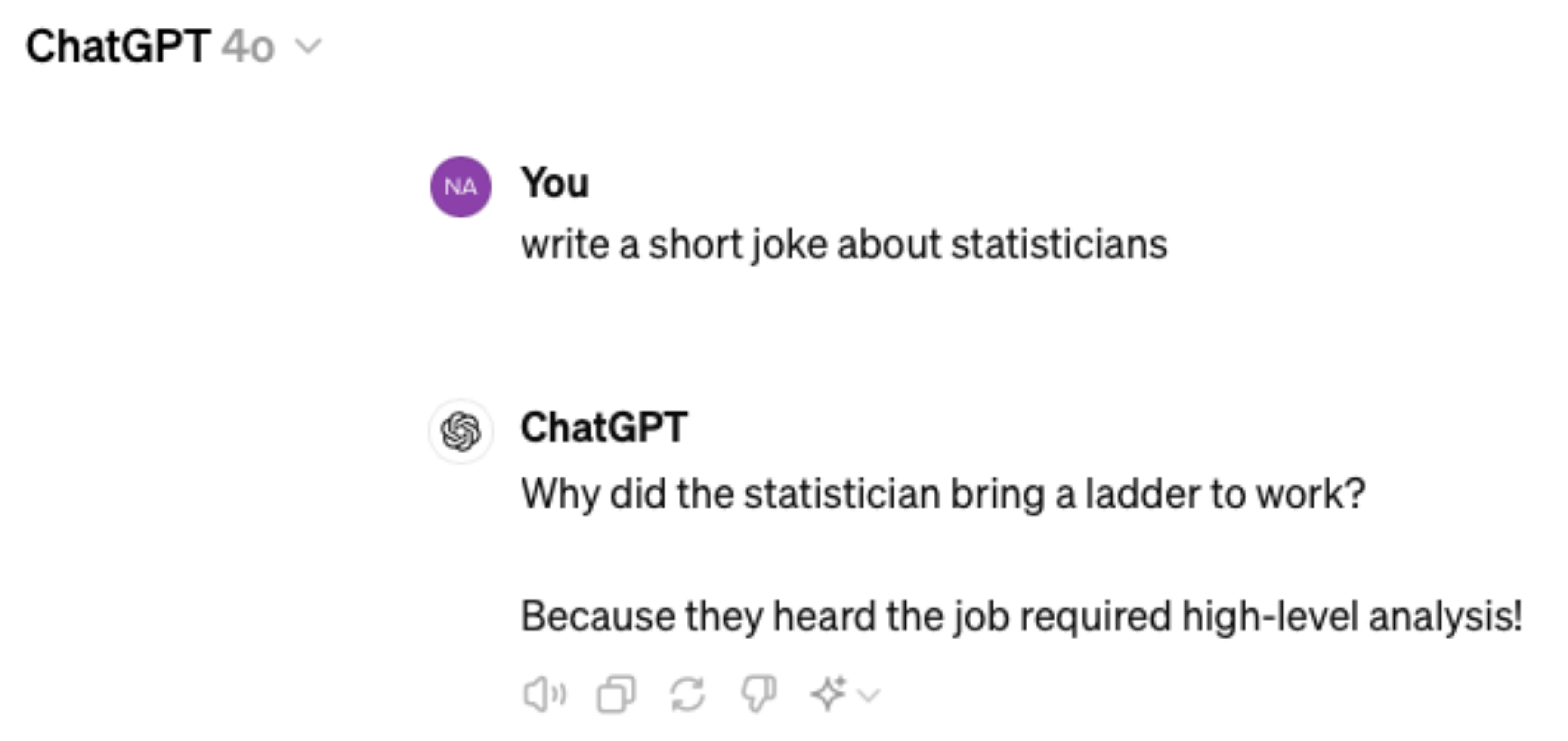
1. www.cullmantribune.com/2022/05/22/52-...


Statistician's Drowning Joke
NT
Copy
Retry
Claude can make mistakes. Please double-check responses.

write a short joke about statisticians

Here's a short joke about statisticians:

Why did the statistician drown crossing the river? Because she had a misplaced decimal point in calculating the mean depth.



You

write a short joke about statisticians

ChatGPT

Why did the statistician bring a ladder to work?

Because they heard the job required high-level analysis!