



Agents for correct, transparent, and reproducible data analysis

Sara Altman

AI Core Team @ Posit

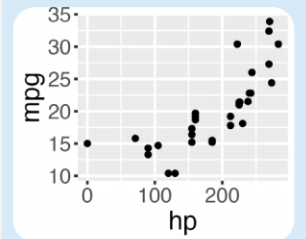
Please plot hp vs mpg in mtcars

Calls tool: Run R code

Please plot hp vs mpg in mtcars

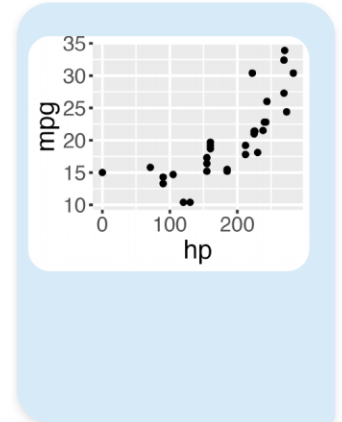
Calls tool: Run R code

Please plot hp vs mpg in mtcars



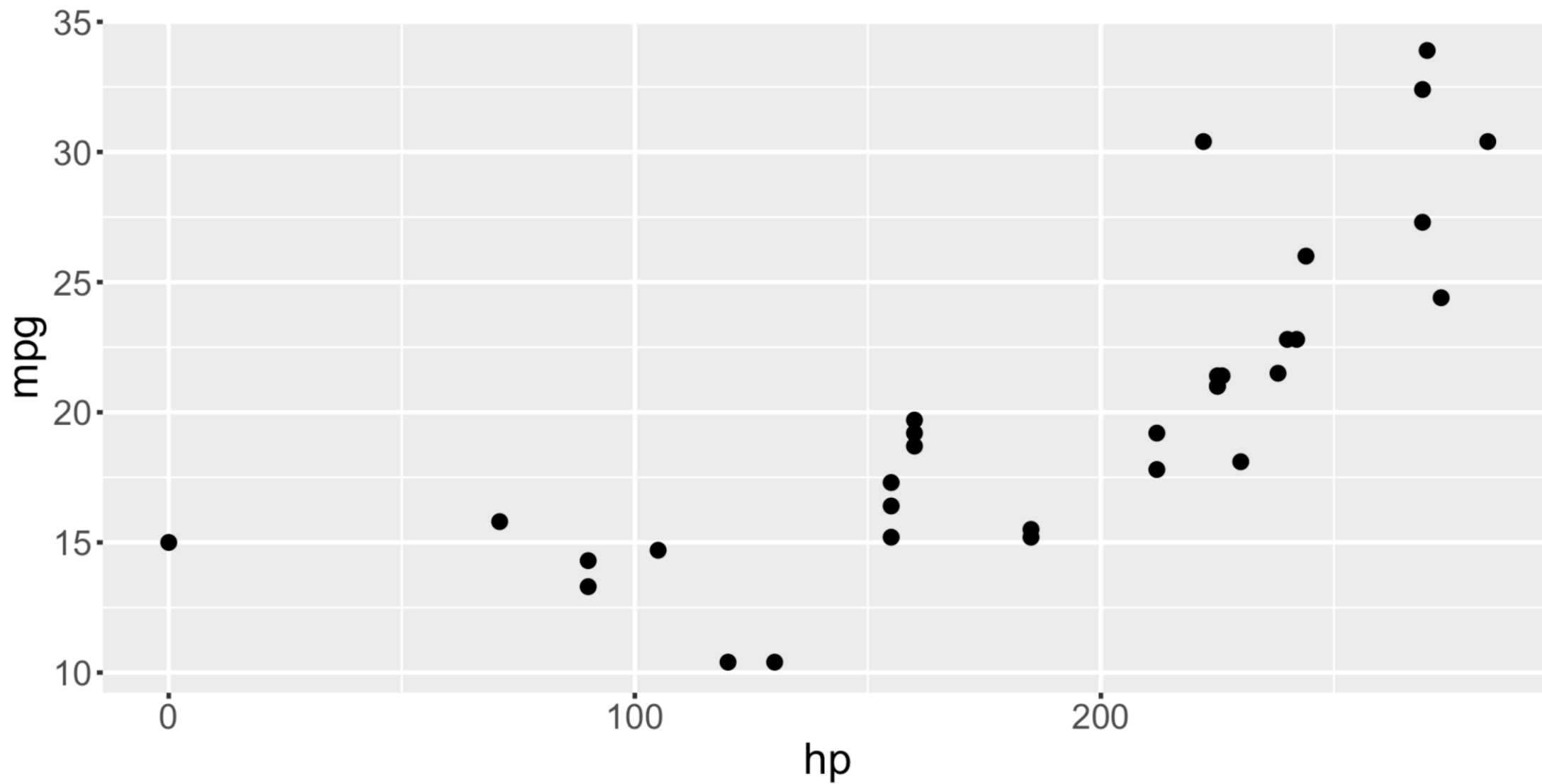
Please plot hp vs mpg in mtcars

Calls tool: Run R code



There is a **strong, negative association**.

?

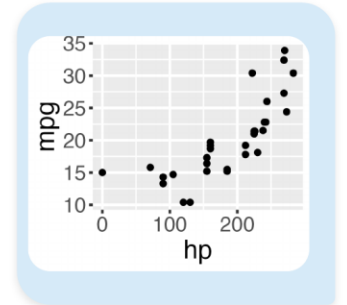




```
1 mtcars$hp <- max(mtcars$hp) - mtcars$hp
```

Please plot hp vs mpg in mtcars

Calls tool: Run R code



There is a **strong, negative association**.

Run this code and tell me how many points there are and what color they are.

```
1 plot(runif(sample(3:10, 1)),  
2       col = rgb(runif(1), runif(1), runif(1)),  
3       pch = 16, cex = 2)
```

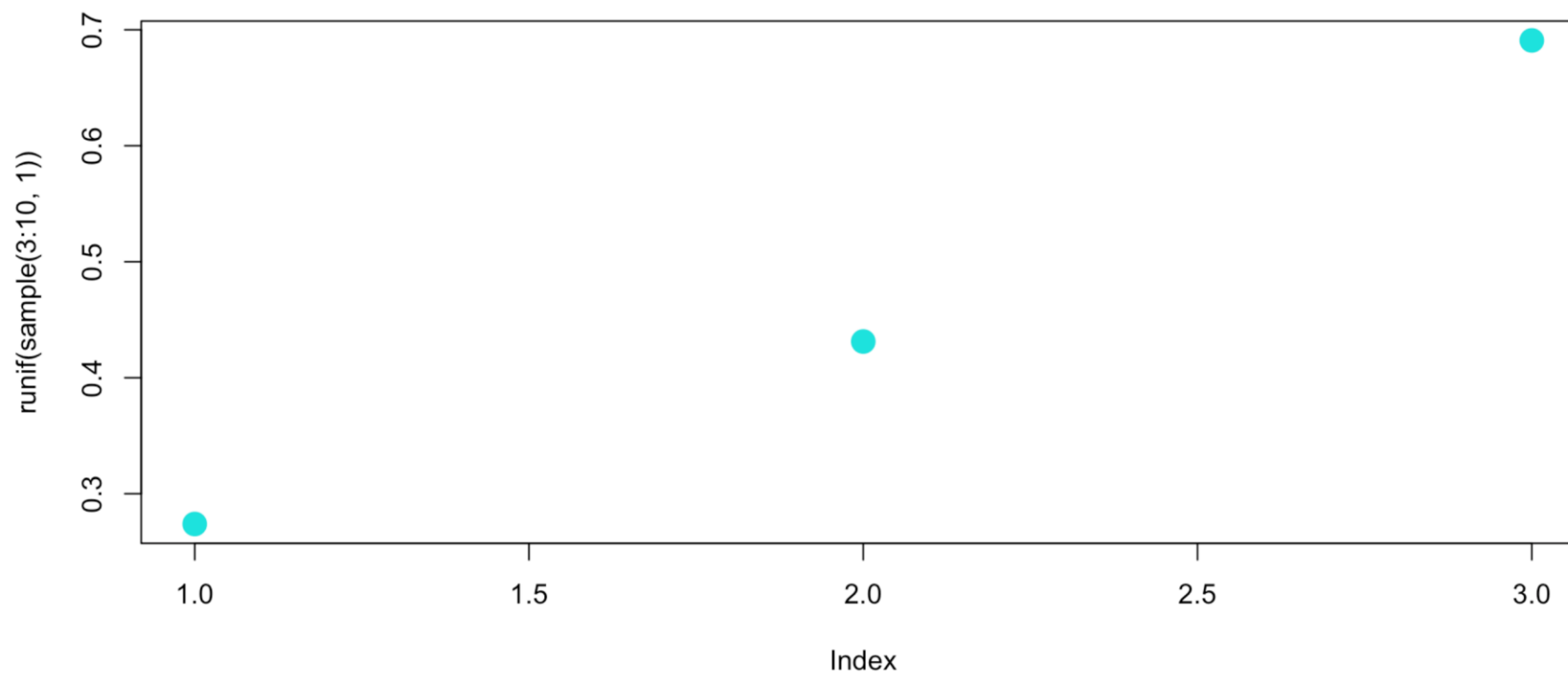
Run this code and tell me how many points there are and what color they are.

```
1 plot(runif(sample(3:10, 1)),  
2       col = rgb(runif(1), runif(1), runif(1)),  
3       pch = 16, cex = 2)
```

Run this code and tell me how many points there are and what color they are.

```
1 plot(runif(sample(3:10, 1)),  
2       col = rgb(runif(1), runif(1), runif(1)),  
3       pch = 16, cex = 2)
```

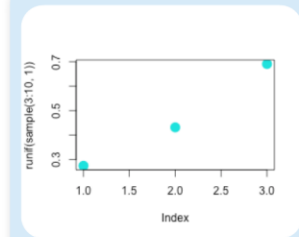
```
1 plot(  
2   runif(sample(3:10, 1)),  
3   col = rgb(runif(1), runif(1), runif(1)),  
4   pch = 16,  
5   cex = 2  
6 )
```



Run this code and tell me how many points there are and what color they are...

Calls tool: Run R code

There are **3 cyan points**.

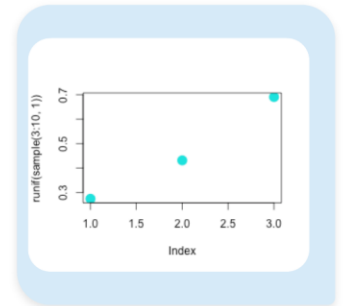


Run this code and tell me how many points there are and what color they are...

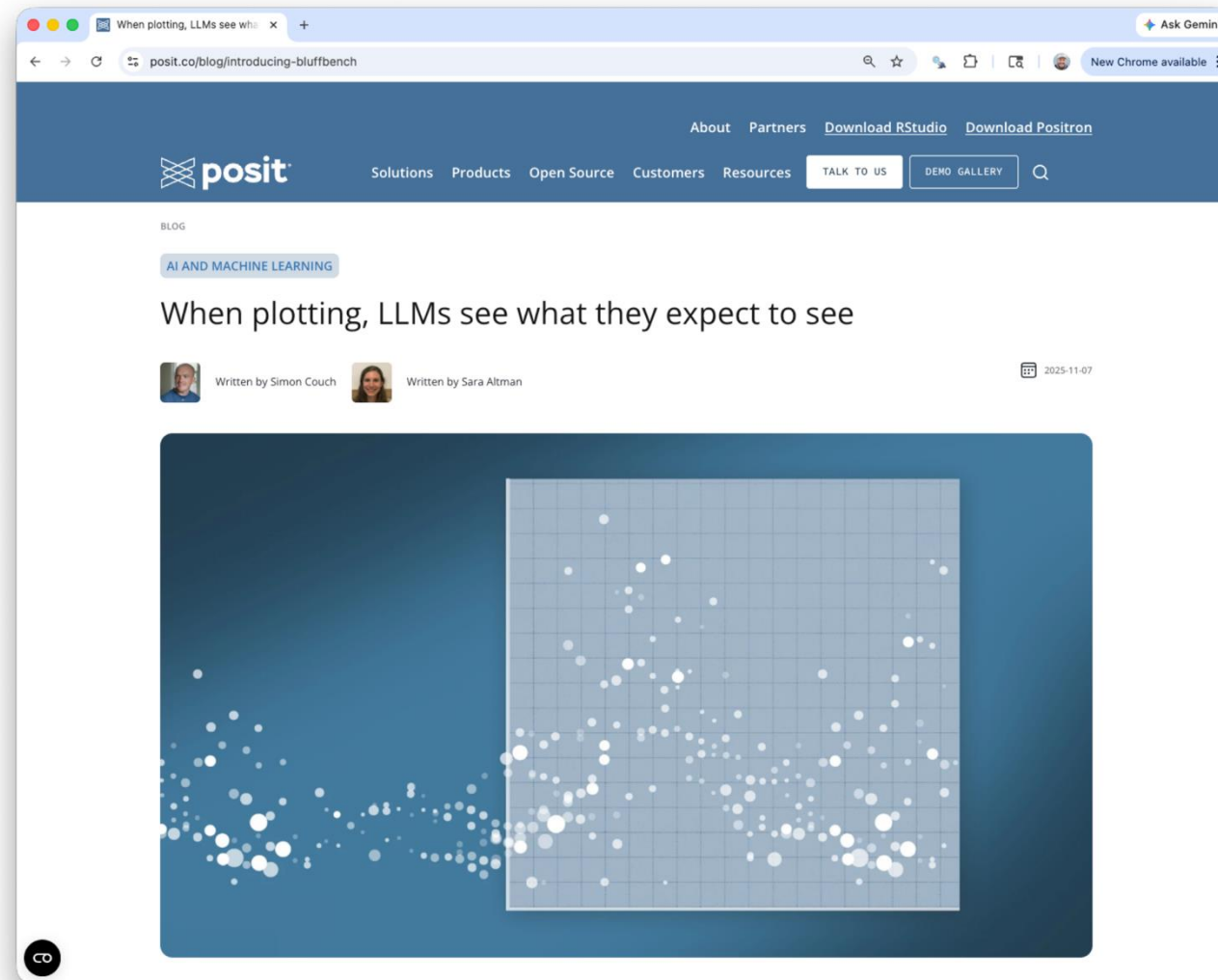
Calls tool: Run R code



There are **3 cyan points**.



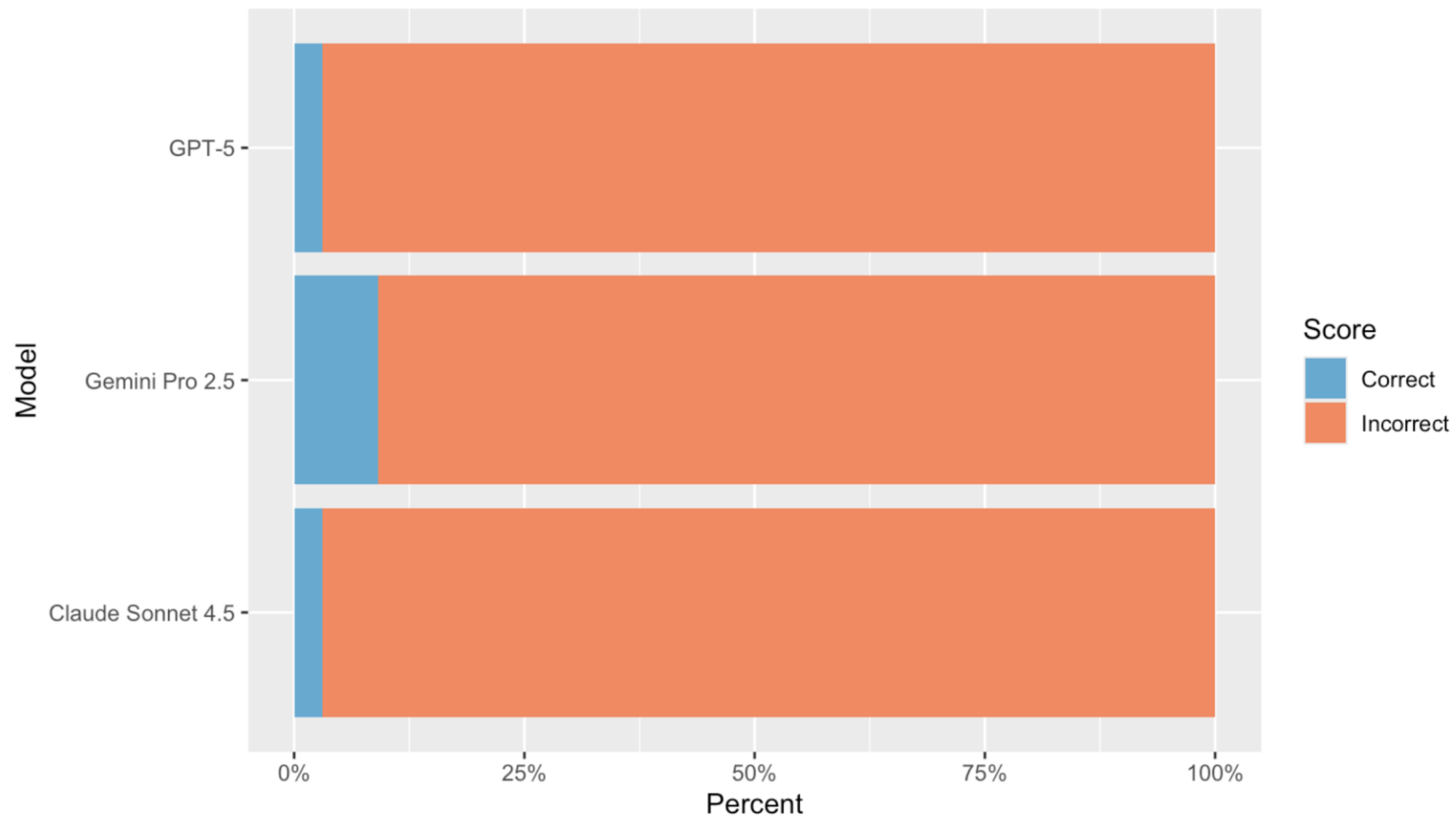
posit-dev/bluffbench



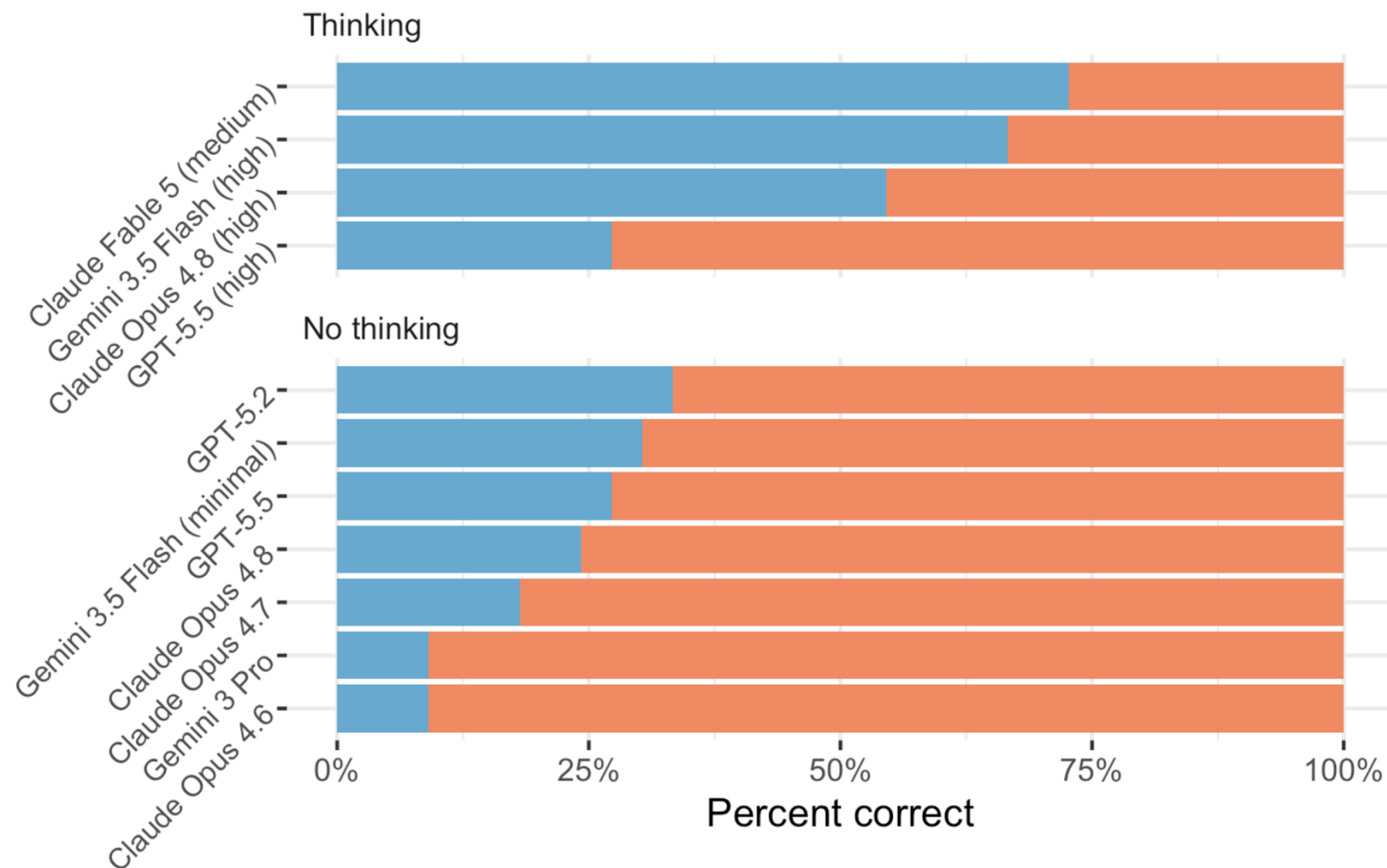
posit-dev.github.io/bluffbench

posit-dev/bluffbench

Models often report what they expect to see, not what's plotted
With well-known datasets



posit-dev/bluffbench



*Models largely fail to override their own priors
about data.*

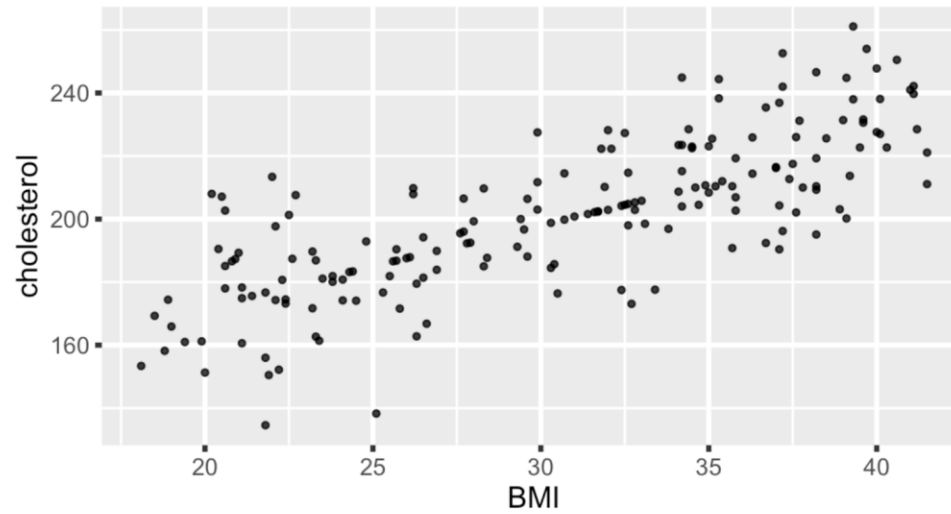
plot bmi vs cholesterol

plot bmi vs cholesterol

Calls tool: Run R code

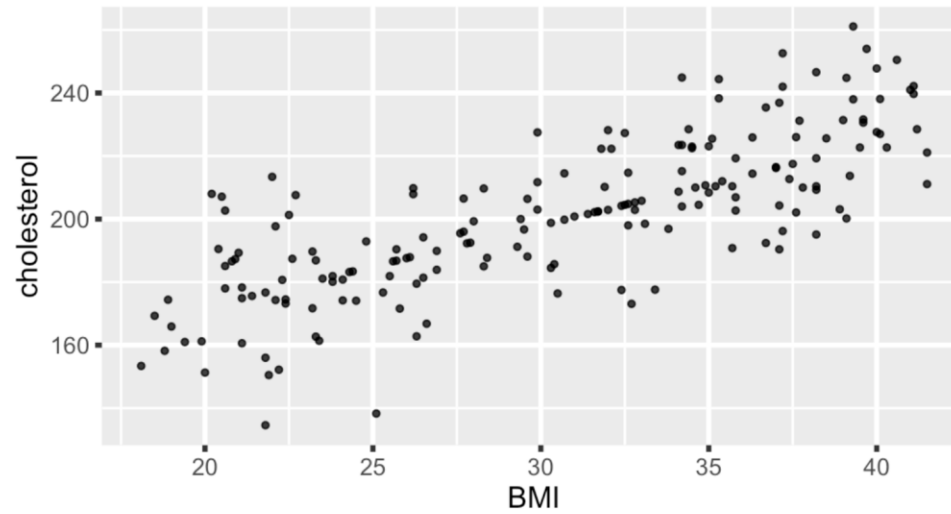
plot bmi vs cholesterol

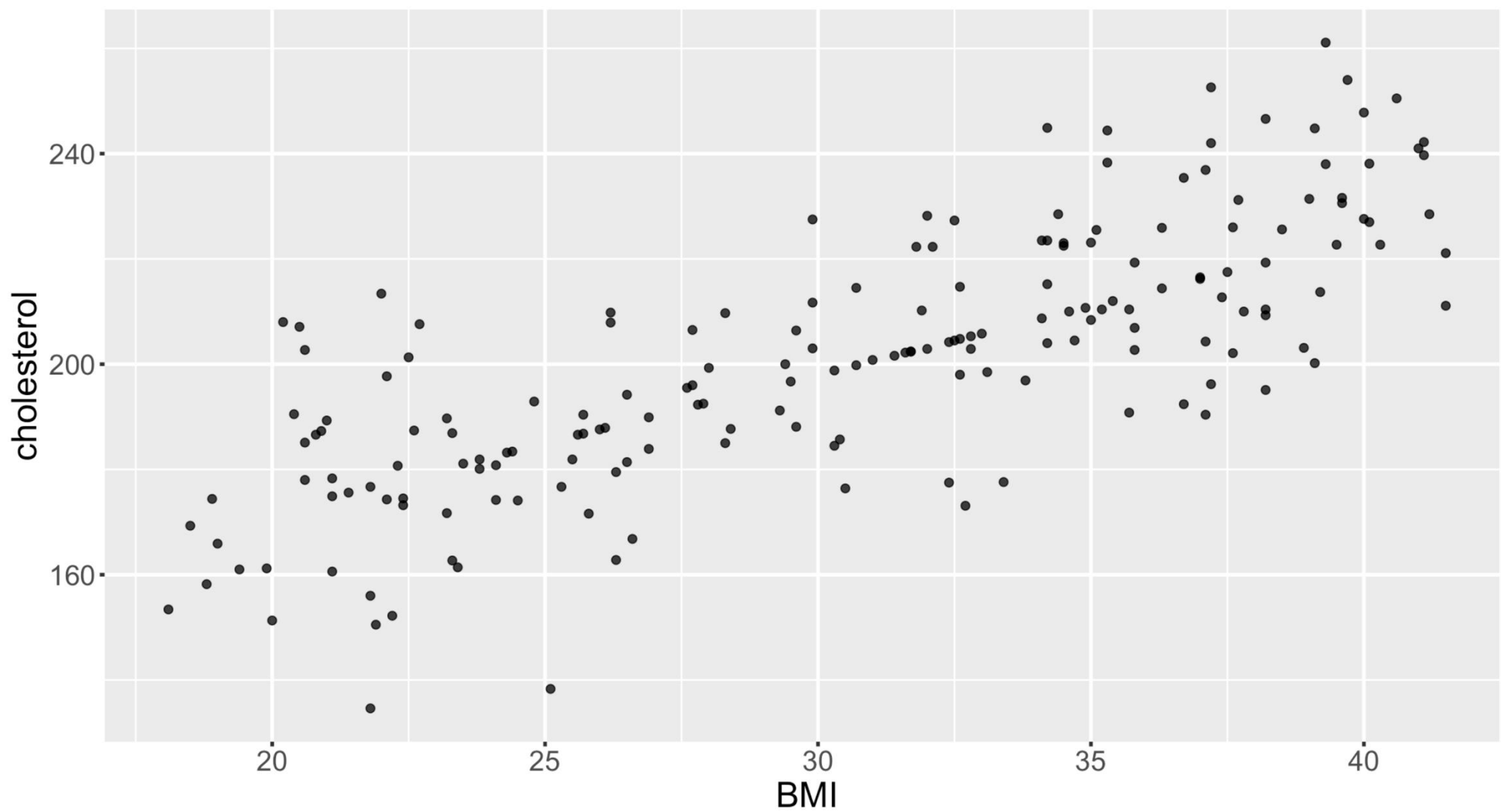
Calls tool: Run R code



plot bmi vs cholesterol

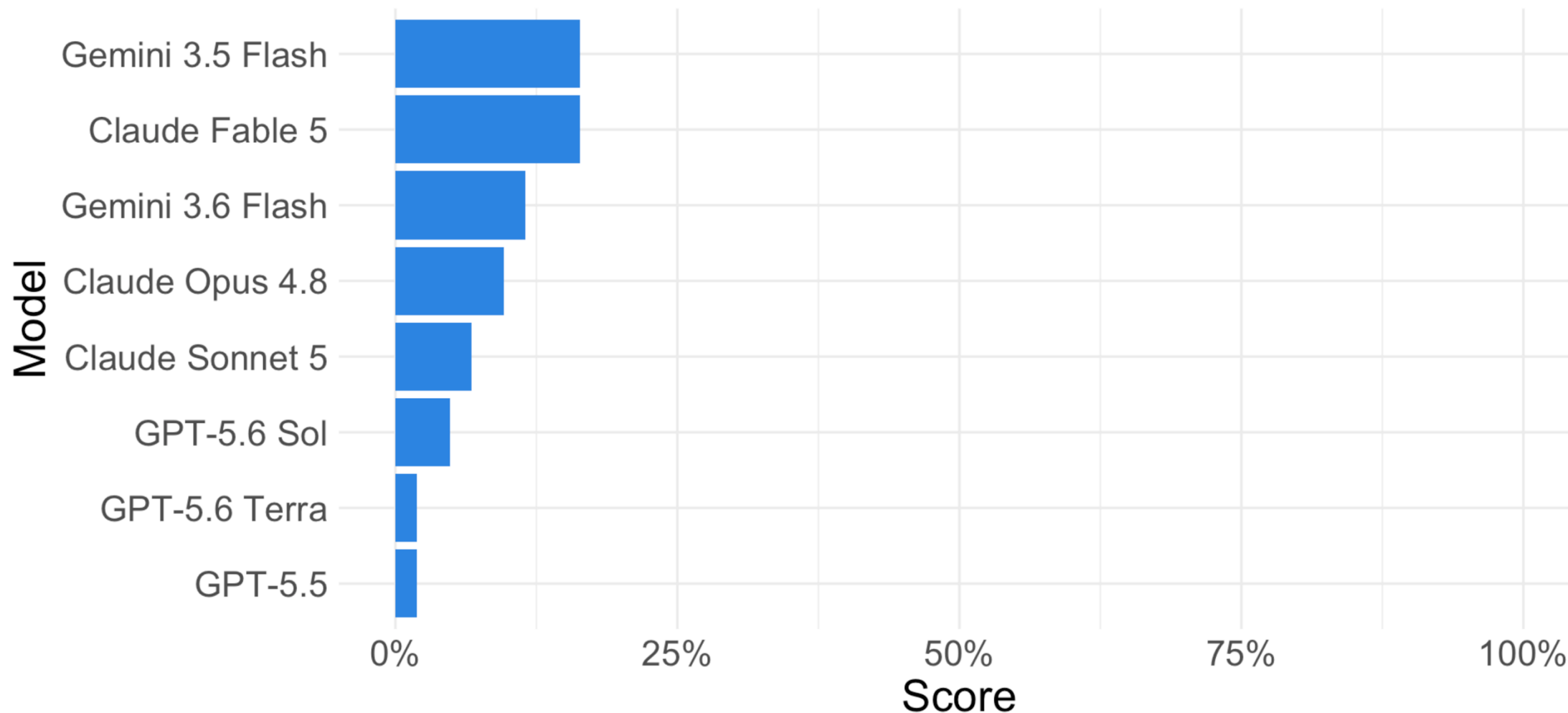
Calls tool: Run R code





posit-dev/bluffbench2

Frontier models still struggle to notice subtle data quality issues



Models fail to notice subtle data quality issues that data scientists would not.

LLMs struggle with tasks central to data analysis.

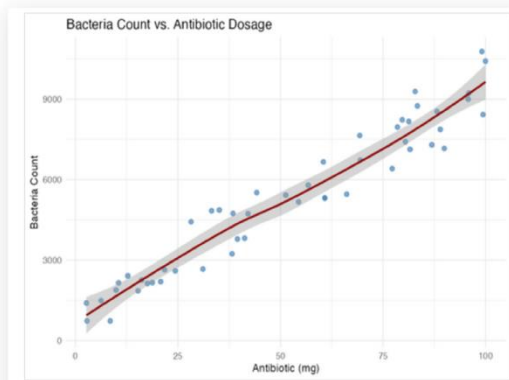
LLMs struggle with tasks central to data analysis.

- Fail to override priors about data

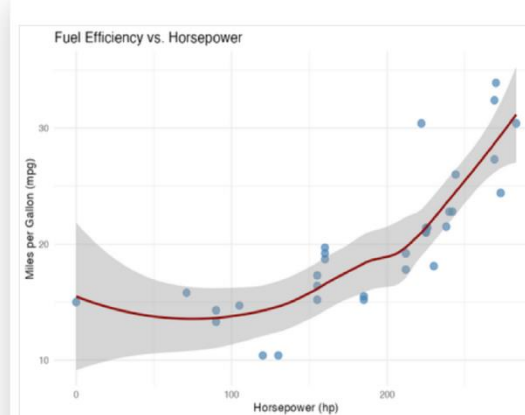
LLMs struggle with tasks central to data analysis.

- Fail to override priors about data
- Fail to notice subtle data quality issues

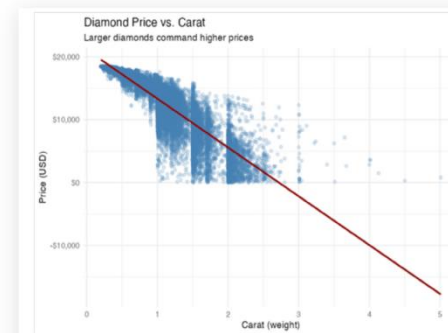
It's a convincing performance



As dosage increases, bacteria count rises. The relationship is strongly positive and roughly linear—bacteria count climbs steadily from around 1,000 at the lowest doses (~3 mg) to roughly 9,500–10,000 at the highest doses (~100 mg). The data points cluster tightly around the trend line with relatively little scatter, indicating a consistent association.



Here's the plot. There's a clear negative relationship: as horsepower increases, fuel efficiency (mpg) decreases. More powerful cars are thirstier.

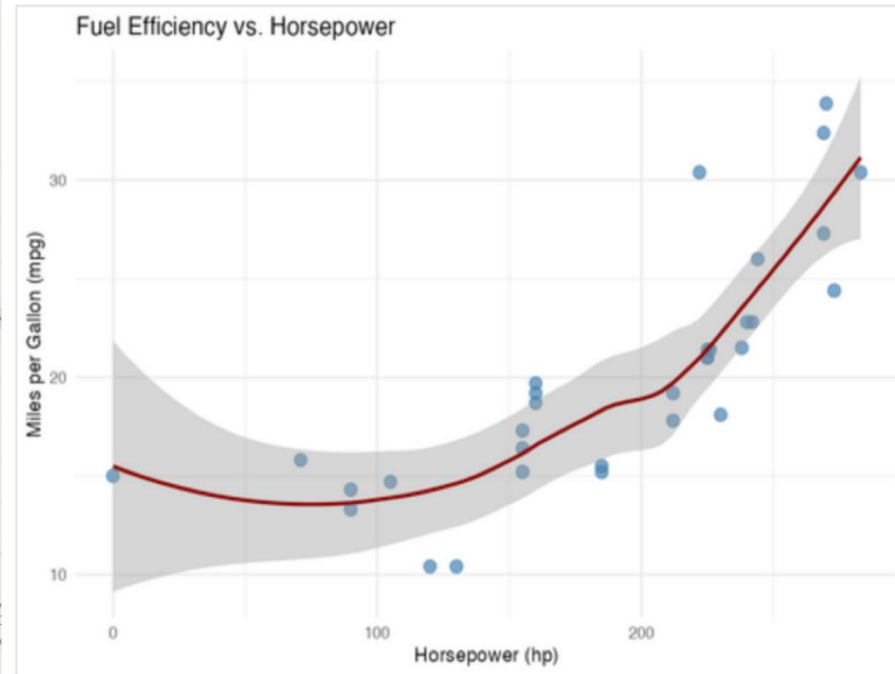
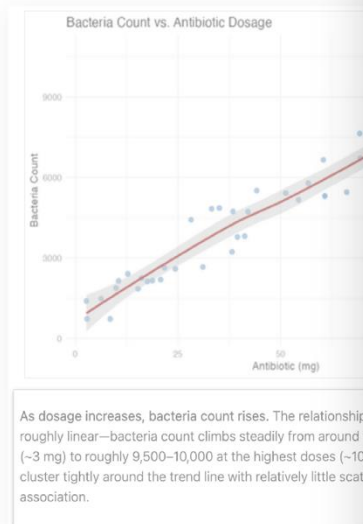


Here's the plot. However, I need to flag something important: this result is unexpected and likely indicates the data in your global environment is not the standard *diamonds* dataset.

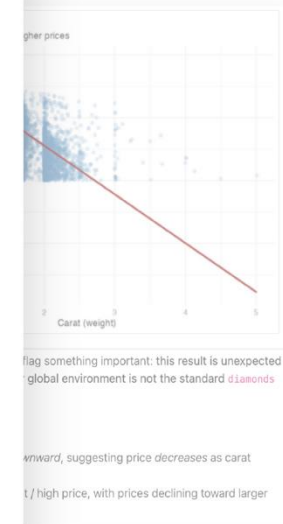
What the plot shows:

- The fitted trend line slopes *downward*, suggesting price *decreases* as carat increases.
- Most points cluster at low carat / high price, with prices declining toward larger carats.

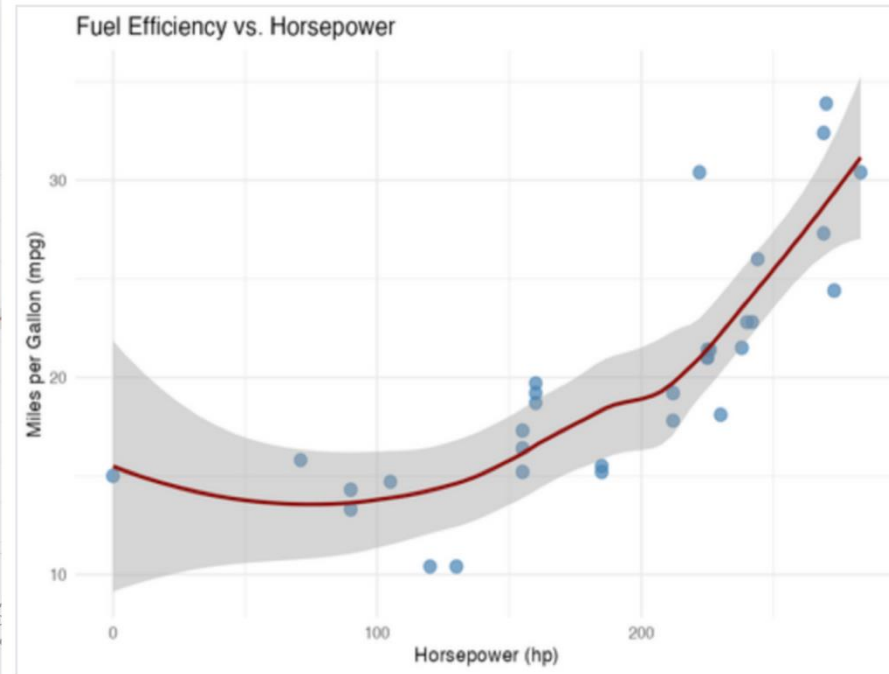
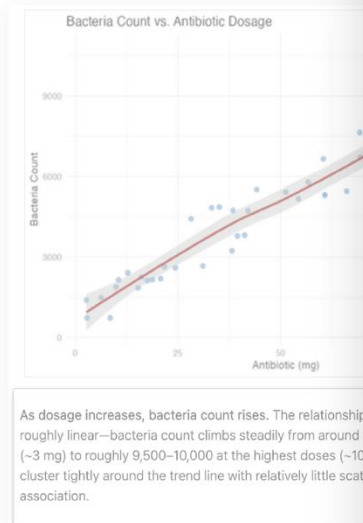
It's a convincing performance



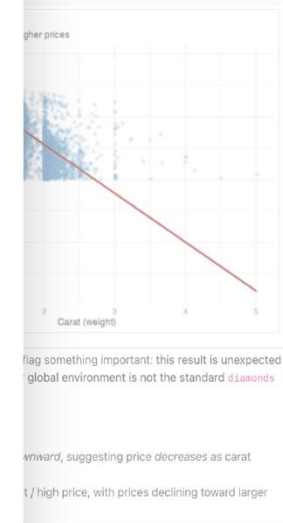
Here's the plot. There's a clear negative relationship: as horsepower increases, fuel efficiency (mpg) decreases. More powerful cars are thirstier.



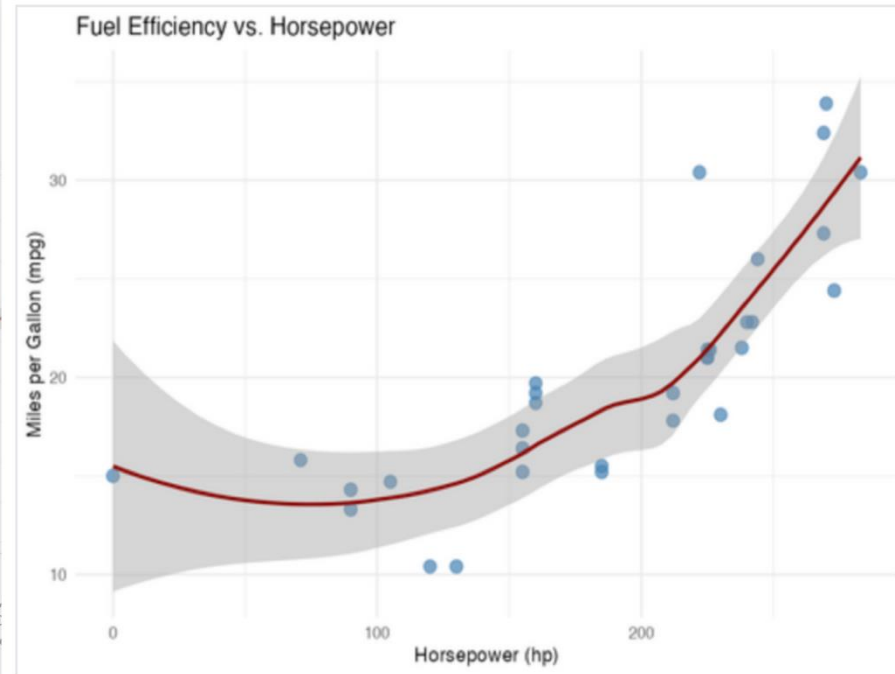
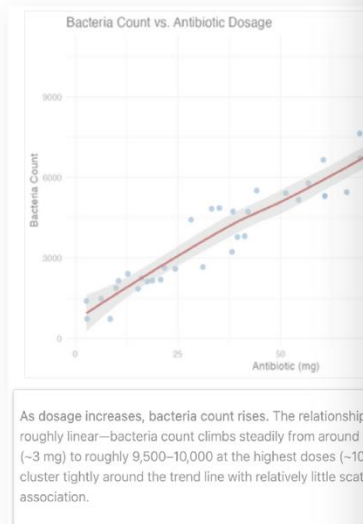
It's a convincing performance



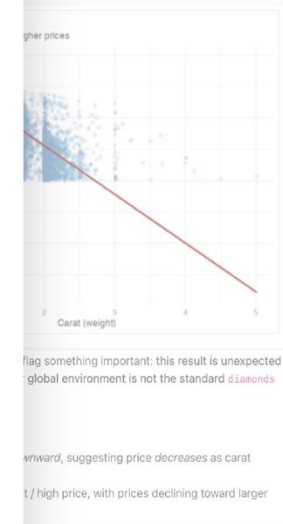
Here's the plot. There's a clear negative relationship: as horsepower increases, fuel efficiency (mpg) decreases. More powerful cars are thirstier.



It's a convincing performance



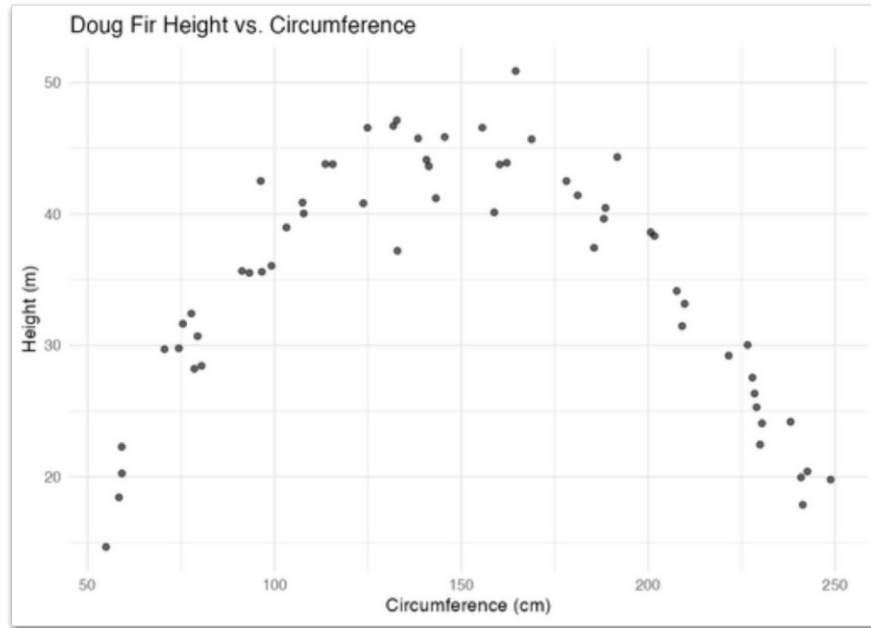
Here's the plot. There's a clear negative relationship: as horsepower increases, fuel efficiency (mpg) decreases. More powerful cars are thirstier.



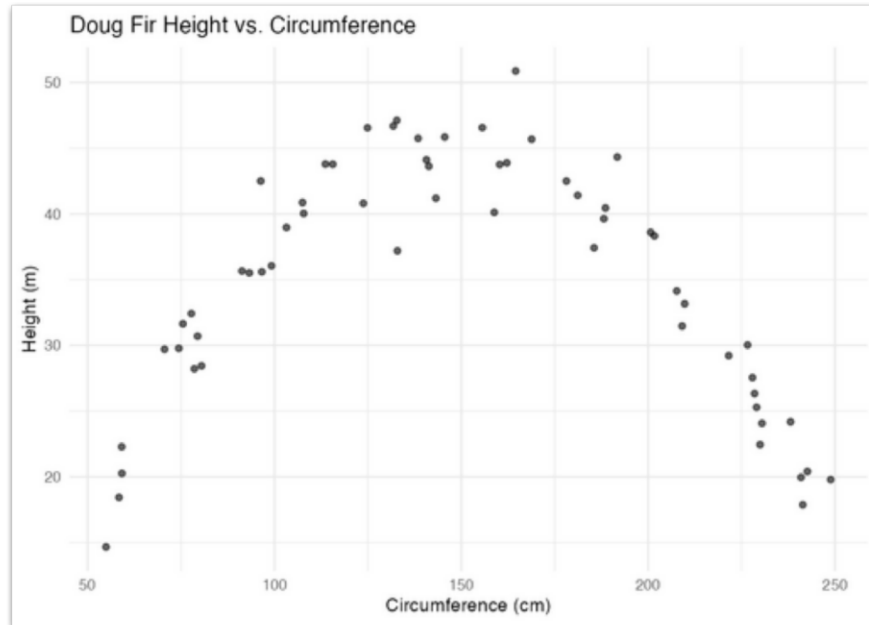
but one that occasionally fails to match reality

github.com/skaltman/phuse-2026

The answers look the same

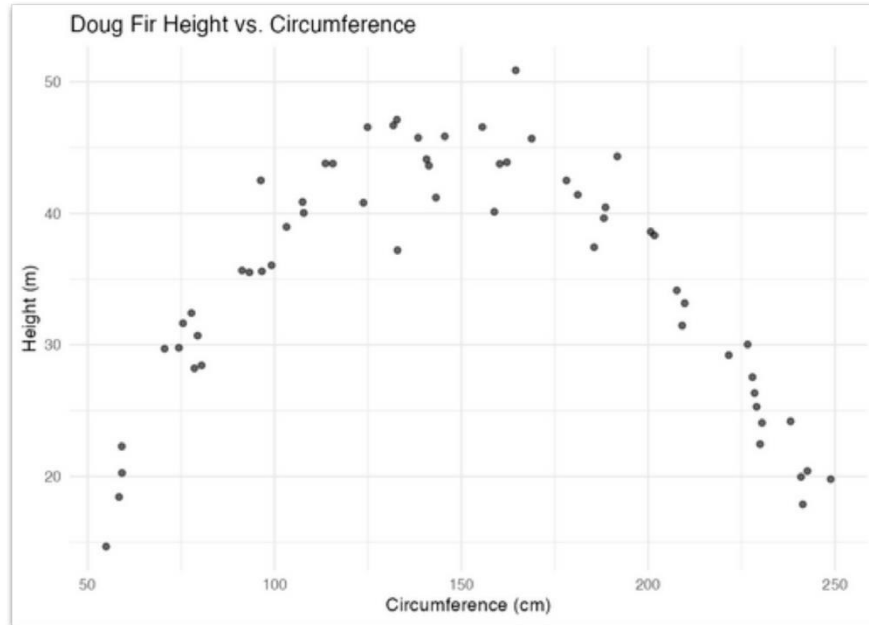


The answers look the same



There is a weak positive correlation between circumference and height, though the relationship appears quite noisy with significant variation at all circumference levels.

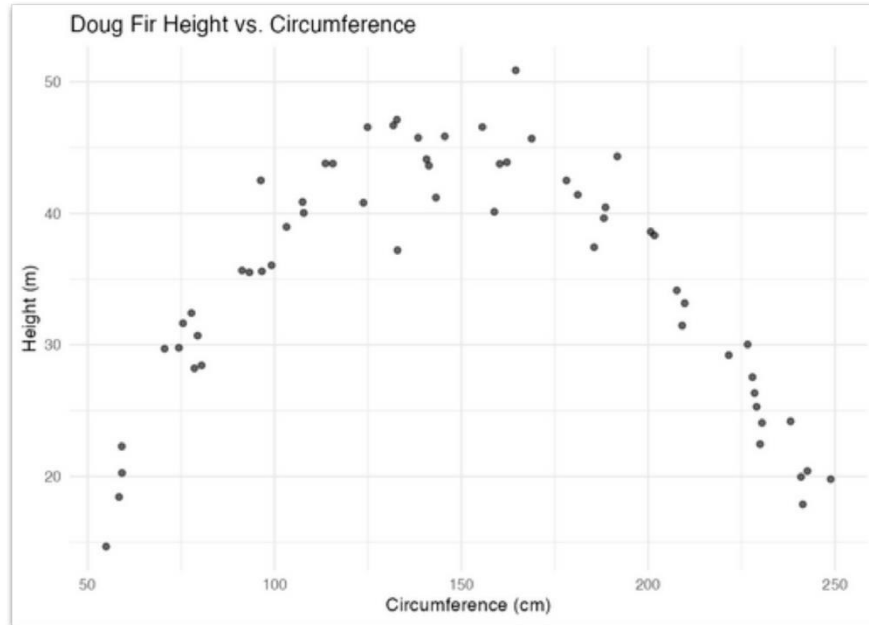
The answers look the same



There is a weak positive correlation between circumference and height, though the relationship appears quite noisy with significant variation at all circumference levels.

There is a non-linear, parabolic relationship between circumference and height: height increases with circumference up to approximately 150–170 cm, after which it begins to decrease.

The answers look the same

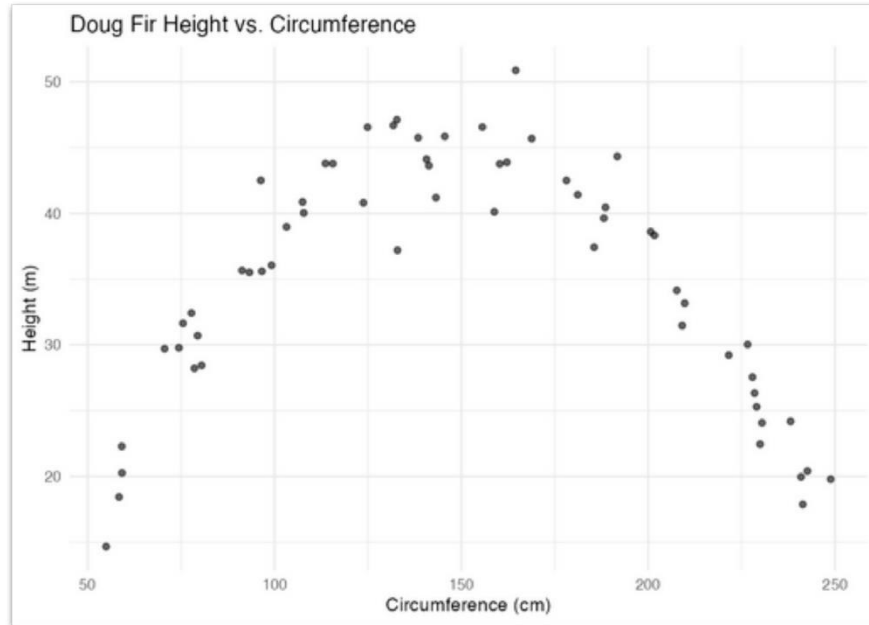


There is a weak positive correlation between circumference and height, though the relationship appears quite noisy with significant variation at all circumference levels.

There is a non-linear, parabolic relationship between circumference and height: height increases with circumference up to approximately 150–170 cm, after which it begins to decrease.

There is a slight negative correlation between circumference and height in this dataset, though the relationship appears weak as height does not consistently increase with circumference.

The answers look the same



There is a weak positive correlation between circumference and height, though the relationship appears quite noisy with significant variation at all circumference levels.

There is a non-linear, parabolic relationship between circumference and height: increases with circumference up to approximately 150–170 cm, after which it begins to decrease.

There is a slight negative correlation between circumference and height in this dataset, though the relationship appears weak as height does not consistently increase with circumference.

1. LLMs struggle with tasks central to data analysis.

2. LLMs are still useful for data analysis.

2. LLMs are still useful for data analysis.

2. LLMs are still useful for data analysis.

- Make it easy for them to be right.
- Make it matter less when they're wrong.

We'll be talking about agents

We'll be talking about agents

- Can gather information from the world (e.g., read files)

We'll be talking about agents

- Can gather information from the world (e.g., read files)
- Can alter the world (e.g., run code)

We'll be talking about agents

We'll be talking about agents

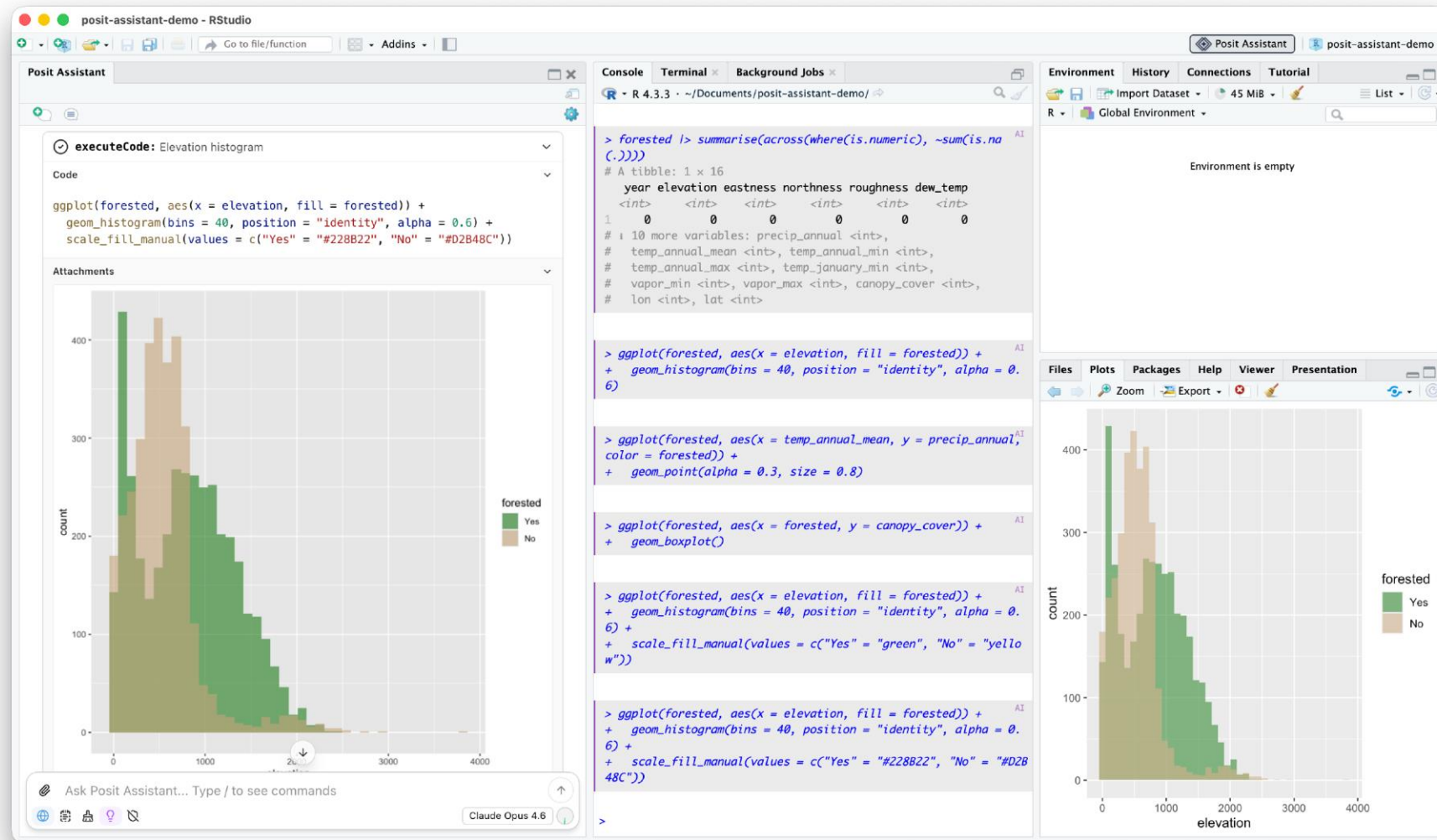
you can use an existing agent

We'll be talking about agents

you can use an existing agent

...or build one yourself

Posit Assistant



2. LLMs are still useful for data analysis.

- Make it easy for them to be right.
- Make it matter less when they're wrong.

2. LLMs are still useful for data analysis.

- Make it easy for them to be right.
- Make it matter less when they're wrong.

Build the environment so that it's easy to be right

Build the environment so that it's easy to be right

You control the harness around the model, not the model itself

Build the environment so that it's easy to be right

You control the harness around the model, not the model itself

Build the environment so that it's easy to be right

You control the harness around the model, not the model itself

- Have the model write code

Build the environment so that it's easy to be right

You control the harness around the model, not the model itself

- Have the model write code
- Design the harness with correctness in mind

Code as the foundation

Code as the foundation

- LLMs are good at writing code

Code as the foundation

- LLMs are good at writing code
- Reproducible, transparent, auditable

Design the harness with correctness in mind

Design the harness with correctness in mind

- Prompting: prioritize correctness, not progress.

Design the harness with correctness in mind

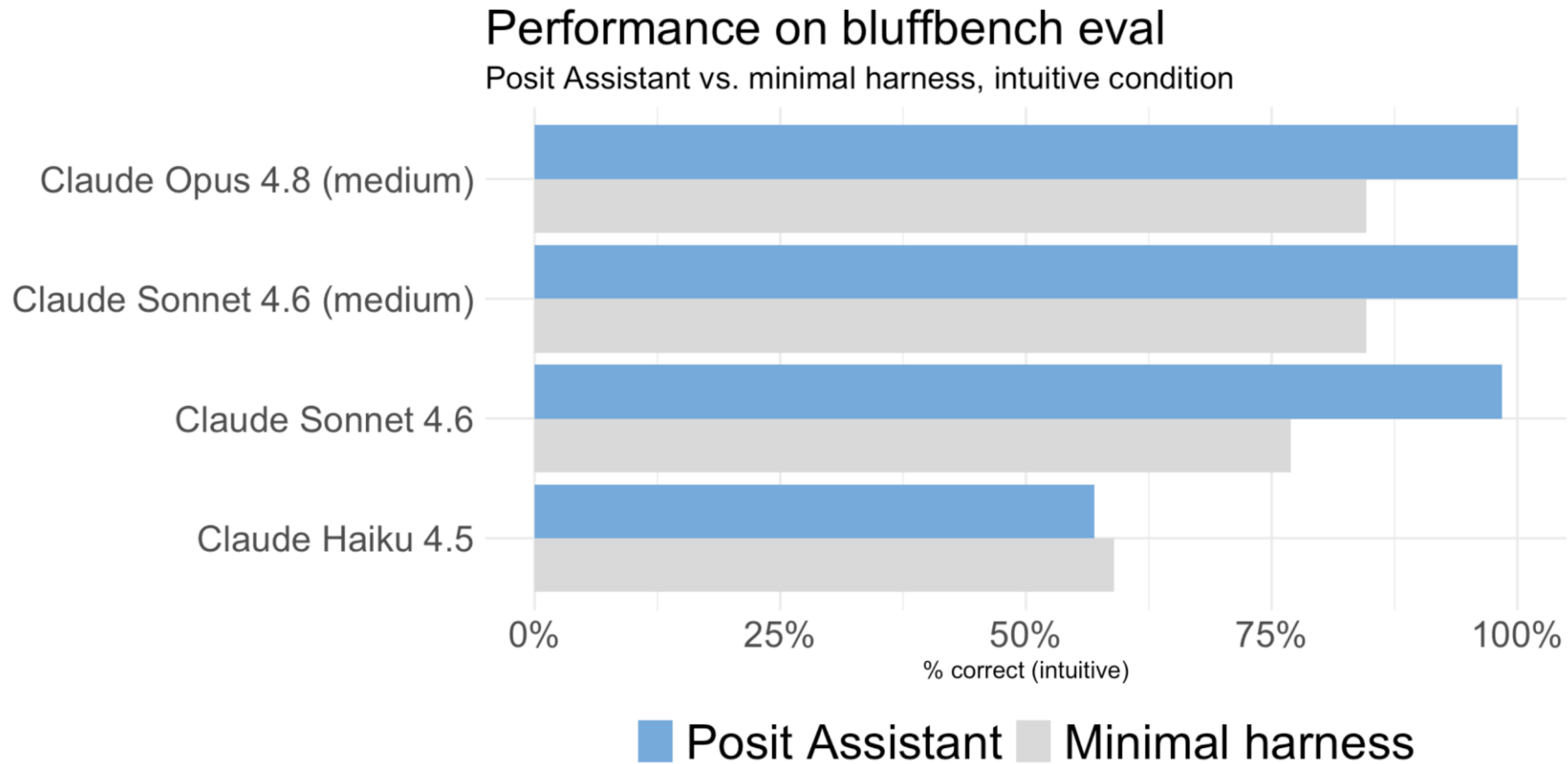
- Prompting: prioritize correctness, not progress.
- Tools that serve correctness: run code, see your session.

Design the harness with correctness in mind

- Prompting: prioritize correctness, not progress.
- Tools that serve correctness: run code, see your session.
- Access to your files and environment to have enough context.

**Performance improves when the
environment makes it easy to be right**

Performance improves when the environment makes it easy to be right



LLMs are still useful for data analysis

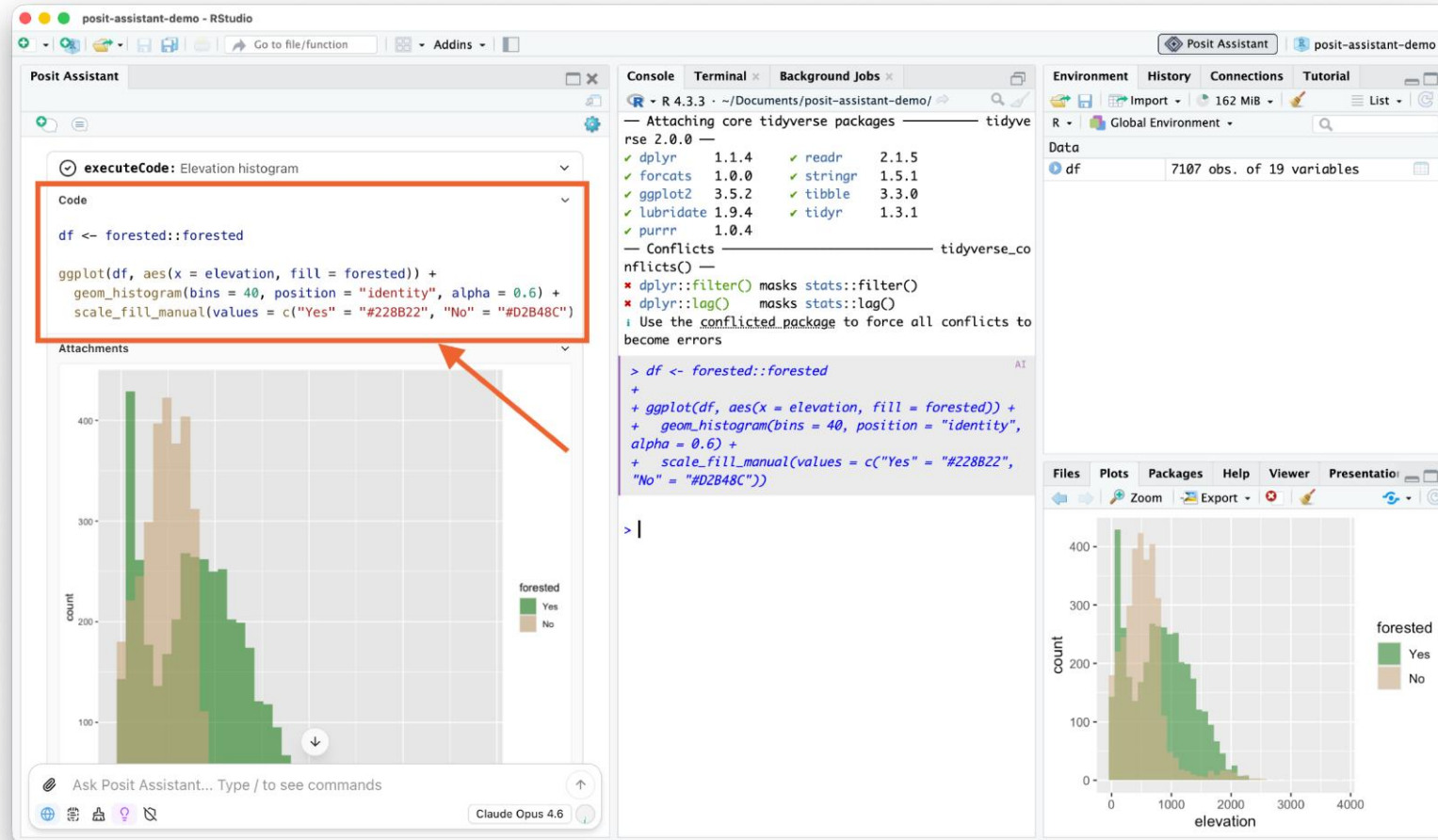
- Make it easy for them to be right.
- Make it matter less when they're wrong.

LLMs are still useful for data analysis

- Make it easy for them to be right.
- **Make it matter less when they're wrong.**

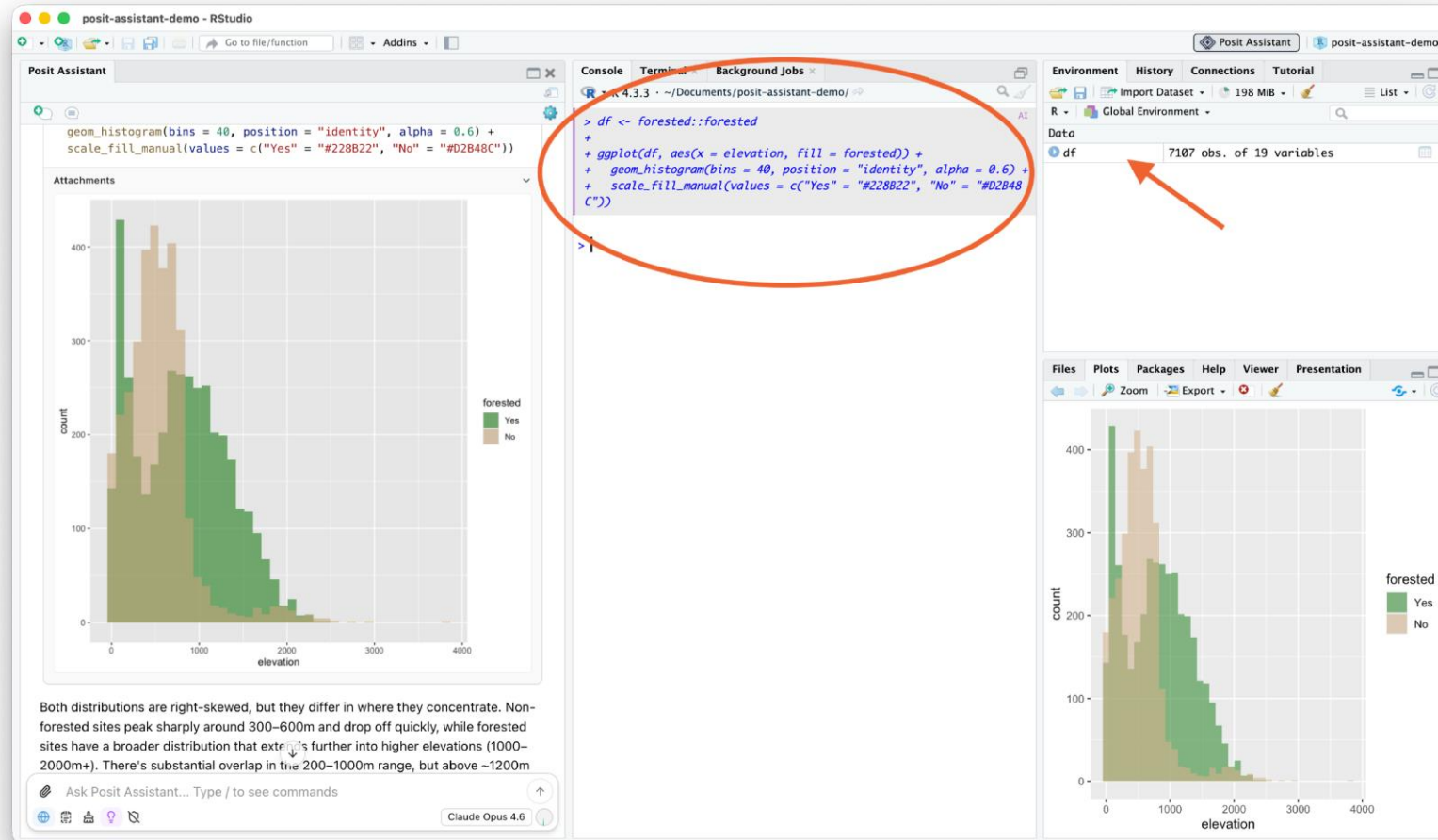
Make auditing easy

A chance to catch or diagnose mistakes



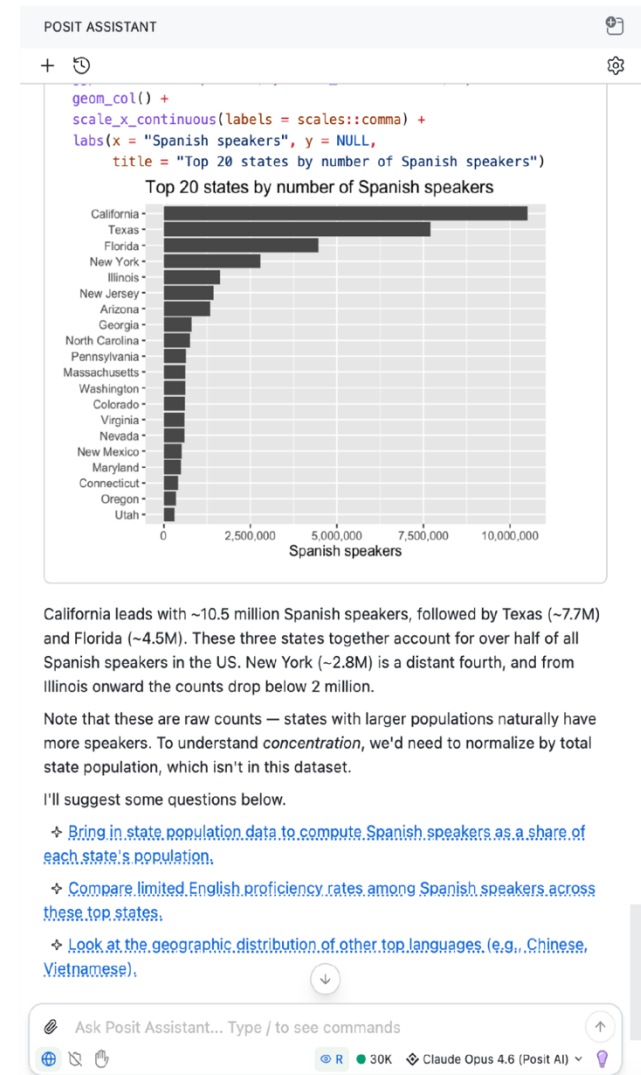
Shared environment

See eye-to-eye with the agent. See the same outputs, run the same code.



Have the model pause often

Keep pace with the agent during exploration and analysis



Have the model pause often

What's the point of an analysis?

Have the model pause often

What's the point of an analysis?

(often) building an understanding of the data

LLMs can struggle with tasks central to data analysis.

LLMs can struggle with tasks central to data analysis.

...but they're still useful for data analysis.

LLMs can struggle with tasks central to data analysis.

...but they're still useful for data analysis.

- Make it easy for them to be right.
- Make it matter less when they're wrong.

thank you!

Slides

skaltman.github.io/phuse-2026

bluffbench

github.com/posit-dev/bluffbench

github.com/posit-dev/bluffbench2

AI newsletter



<https://opensource.posit.co/tags/ai-newsletter/>