R for Everyone: An Interactive, Ethics-Aware Data Science Micro-Curriculum for Middle School

Tene Ariyo

December 09, 2025

Introduction and Motivation

Data has become a native language of civic life and work. People scroll past charts on social media, navigate dashboards, and encounter numerical claims about health, climate, and finance every week. To participate effectively and build strong data literacy, they need guided practice with the full statistical problem-solving process: posing investigable questions, collecting or considering data, analyzing patterns and variability, and interpreting results with appropriate attention to uncertainty and audience.

Why Middle school?

We target middle school because it is the moment when many students begin owning or regularly using personal devices (e.g., iPhones, tablets, laptops), which means they already search, click, and interact with data daily. Meeting them at this stage lets us turn informal scrolling into intentional inquiry asking good questions, collecting or considering data responsibly, and making first plots they can explain.Yet access to formal data science learning opportunities remains uneven, and many middle school curricula still treat statistics as a short end-of-year unit rather than a core problem-solving lens (Bargagliotti et al. 2020).

National reports and coalitions such as Data Science 4 Everyone argue that data literacy is becoming a foundational civic competency, comparable to reading and algebra, and that students should experience tools and tasks that resemble those used by practicing data scientists, not just hand-computed summaries. At the same time, research shows that data literacy outcomes are falling fastest in these middle school grades: between 2019 and 2022, 8th-grade scores on NAEP's Data Analysis, Statistics, and Probability category declined by about 10 points, roughly a full grade level outpacing drops in other math areas (Drozda 2023).

The Guidelines for Assessment and Instruction in Statistics Education (GAISE II) emphasize a four-step statistical problem-solving process, **formulating questions, collecting or considering data, analyzing data, and interpreting results**, and recommend that

instruction center authentic questions and data sources (Bargagliotti et al. 2020). Complementary resources such as youcubed's "Data Big Ideas" provide grade-level guidance for how these ideas can unfold in the middle grades, including attention to variability, fairness, and interpretation of visual displays (YouCubed, n.d.).

This capstone is situated at the intersection of these recommendations and the emerging literature on **embedding ethics in mathematics and statistics education**. Instead of treating ethical considerations as stand-alone lessons, recent work suggests integrating short, recurring "micro-prompts" that ask students to reflect on bias, missing data, and who is affected by decisions based on data (Chiodo, Müller, and Shah 2025). My goal is to design a small set of learnr-based lessons that implement this approach in R, at a level appropriate for middle school students and their teachers.

Curriculum Design and Methods

Framing the four-lesson arc

The four-lesson sequence is explicitly organized around the GAISE problem-solving cycle:

1. **Formulate questions**
2. **Collect or consider data**
3. **Analyze data**
4. **Interpret results**

Using the youcubed "Data Big Ideas" for grades 6–7 as a guide, I mapped these phases onto a four-lesson arc that can fit into a short unit:

- **Lesson 1 – Formulate Statistical Investigative Questions**
  Focus: distinguishing investigable vs. non-investigable questions, identifying variables, and beginning to think about sampling and bias.
- **Lesson 2 – Collect & Consider Data**
  Focus: examining a small, structured dataset; understanding rows and columns; recognizing variability; and running first R commands (`nrow()`, `head()`, simple plots).
- **Lesson 3 – Analyze Patterns in Data**
  Focus: creating and interpreting graphical displays (bar charts or boxplots), comparing groups, and quantifying simple differences.
- **Lesson 4 – Interpret and Communicate Results**
  Focus: synthesizing findings, articulating limitations, and reflecting on how data and ethics intersect in drawing conclusions.

The aim is to make R feel approachable to middle school students while keeping the tasks anchored in authentic question-posing and interpretation, rather than in syntax for its own sake.

Data and Instructional Artifacts

In this project, "data" appears in several layers:

1. **Contextual data** from reports and coalitions (e.g., NAEP results, Data Science 4 Everyone, GAISE) that motivate the need for early data science experiences.
2. **Simulated, instructional data** that I generated inside the learnr tutorials for instructional purposes, embedded directly inside the learnr tutorials (e.g., a small reading-time survey in Lesson 2).
3. **Validation data**, which will eventually include teacher feedback collected via a short Google Forms rubric.

Contextual data

I draw on existing reports and position papers primarily as background rather than as objects of original statistical analysis. These documents provide:

- Evidence that data literacy gaps exist and have equity implications.
- Conceptual frameworks for structuring a coherent progression of ideas across grades.
- Recommendations for integrating technology and real-world datasets.

Rather than re-analyzing those large datasets, I use them to justify the design choices in my four-lesson sequence and to anchor the learning goals for each lesson.

Instructional data: the Lesson 2 reading survey

Lesson 2 introduces students to small, manageable datasets that they can explore within R. For example, a simplified reading survey dataset records how many minutes each student read the previous evening and which device they typically use for reading or homework.

In the learnr tutorial, this dataset is created behind the scenes so that students can immediately focus on interpreting rows, columns, and variability. A simplified version of the dataset creation appears below:

```
set.seed(123)

survey <- tibble(
```

```
  student_id   = 1:12,
  grade        = rep(7, 12),
  minutes_read = c(20, 35, 0, 15, 40, 30, 10, 25, 5, 50, 15, 45),
  device       = c(
    "phone", "tablet", "phone", "phone",
    "tablet", "other", "other", "phone",
    "tablet", "other", "phone", "tablet"
  )
)
survey
```

```
## # A tibble: 12 x 4
##    student_id grade minutes_read device
##         <int> <dbl>        <dbl> <chr>
##  1          1     7           20 phone
##  2          2     7           35 tablet
##  3          3     7            0 phone
##  4          4     7           15 phone
##  5          5     7           40 tablet
##  6          6     7           30 other
##  7          7     7           10 other
##  8          8     7           25 phone
##  9          9     7            5 tablet
## 10         10     7           50 other
## 11         11     7           15 phone
## 12         12     7           45 tablet
```

Students are then guided through basic "first look" operations:

```
nrow(survey)
```

```
## [1] 12
```

```
min(survey$minutes_read)
```

```
## [1] 0
```

```
max(survey$minutes_read)
```

```
## [1] 50
```

These operations are not just technical; they prepare students to reason about **variability** ("Why do the reading minutes differ across students?") and to confront **ethical questions** about who is represented in the data and who is missing.

In lesson 3, a simple plot of reading minutes by device type serves as a springboard for both statistical and ethical interpretation:

```
ggplot(survey, aes(x = device, y = minutes_read)) +
  geom_boxplot() +
  labs(
    title = "Minutes Read by Device Type",
    x = "Device used for reading or homework",
    y = "Minutes read last night"
  )
```

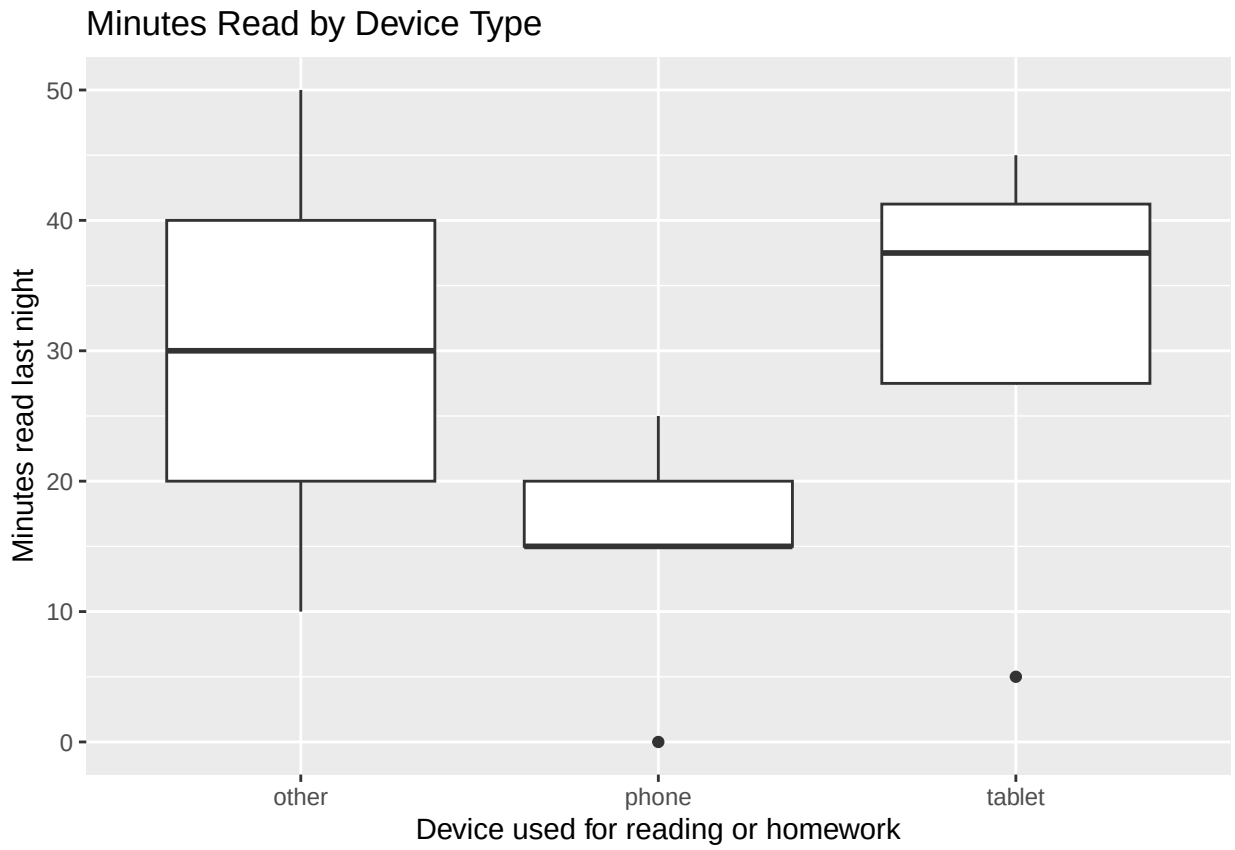

Figure 1: Boxplots of minutes read by device type for the simulated reading survey (n = 12). In this sample, the median minutes read is highest for tablets and lowest for phones, and the tablet and phone groups show less variability than laptops and paper.

Planned teacher validation data

To support iterative refinement, I also designed a short **Teacher Validation Rubric** delivered through Google Forms. The rubric asks teachers (or prospective teacher partners) to rate each lesson on clarity, grade-level appropriateness, alignment with the GAISE framework, and usefulness of the embedded ethics prompts. Open-ended items invite suggestions for improvement and concerns about classroom implementation. *Link to the form: https://forms.gle/5UdDv4Q3WCDaYhu37*

At the time of this draft, I have not yet collected classroom teacher responses. Instead, I have relied on ongoing feedback from my capstone advisor, who has helped refine lesson goals, wording, and pacing. In future work, responses to the rubric will function as implementation data, complementing the design analysis in this write-up.

Writing and iterating on prompts

I developed the lesson prompts in several passes:

1. **Initial drafting from frameworks**
   I began by translating the GAISE steps and youcubed big ideas into concrete middle school tasks. For Lesson 1, this meant building a set of example questions that varied along dimensions such as specificity, measurability, and scope.

2. **Embedding ethics micro-prompts**
   Using the *Teaching Resources for Embedding Ethics in Mathematics* as a guide, I aimed to include **short, recurring touches** on ethics rather than one large "ethics day". This meant adding brief reflective questions in context, such as asking students who might be left out of a survey or how device access could shape the data.

3. **Feedback and refinement**
   Throughout the process, I shared lesson drafts with my capstone advisor, who provided feedback on pacing, clarity, and the cognitive load imposed by R syntax. Teacher feedback via the validation rubric is still an ongoing process, and I plan to incorporate those perspectives in future iterations of both the lessons and this write-up.

Analysis of lessons

In this section, I analyze all four lesssons as designed artifacts, focusing on their alignment with the GAISE framework and the ethics-embedding goals.

Lesson 1: Formulating questions and noticing bias

Lesson 1 asks students to distinguish between questions that can be investigated with data and those that cannot. For instance, students might be given a list such as:

- "What's the best candy"
- "Which candy is the most popular in our class"
- "Is chocolate good"

They then choose the answer they think is **investigable with data**. They are then asked to rewrite this vague question, "Do people read a lot?", to make it investigable. This directly supports the GAISE emphasis on formulating clear, measurable questions.

An embedded ethics micro-prompt invites students to reflect on sampling bias:

> If you only survey friends sitting near you, what sampling bias might that introduce, and how could you reduce it? Write 2–3 sentences.

This prompt encourages students to notice how their own social networks and habits might limit the data they collect. Rather than asking for a single correct answer, the prompt invites argumentation: students must articulate a potential bias (e.g., only surveying people with similar interests or schedules) and propose an improvement (e.g., random sampling, making sure different groups are represented). This shift—from finding "the right answer" to **arguing a position using ideas about bias and representation**—is directly inspired by the ethics resource's recommendations for short, reflective tasks.

Screenshots from Lesson 1 can illustrate these elements in context:

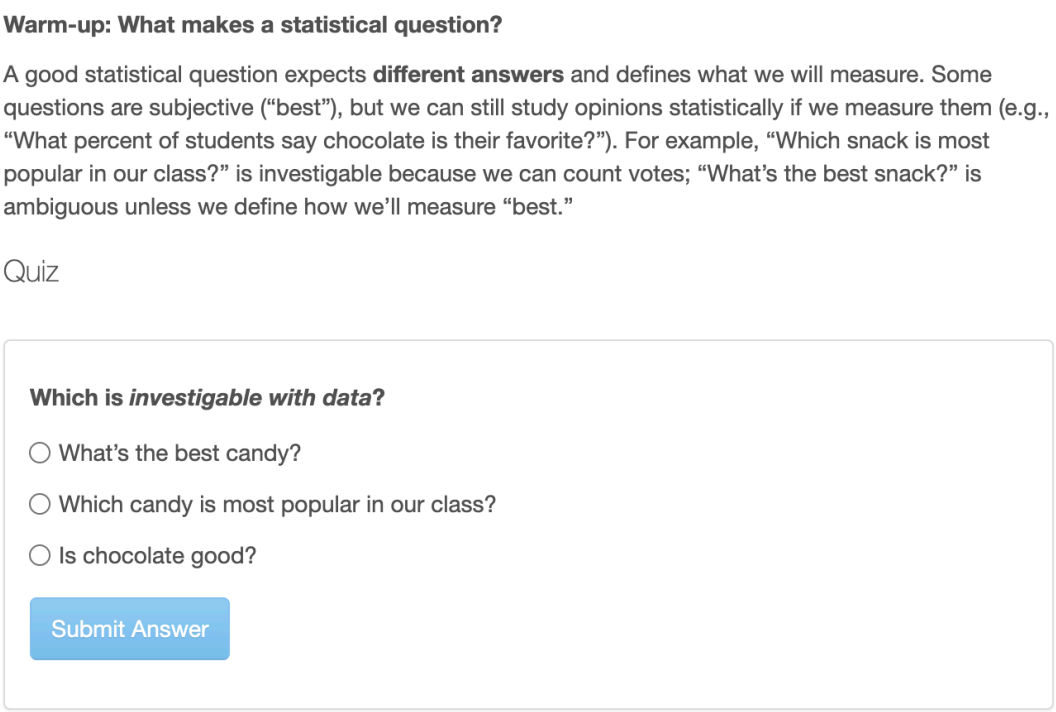

Figure 2: Screenshot of the investigable vs. non-investigable question check from Lesson 1.

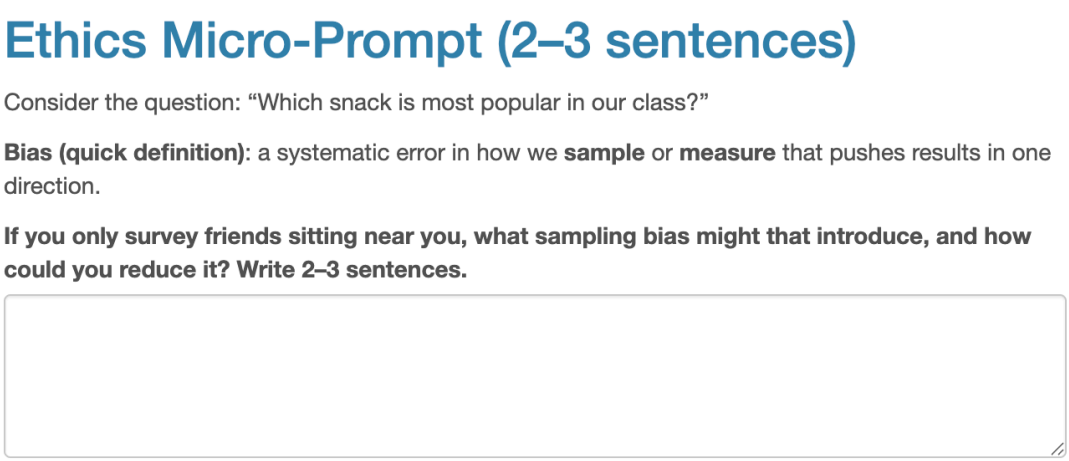

Figure 3: Screenshot of the sampling-bias ethics prompt and free-response box from Lesson 1.

Lesson 2: Considering data and variability

Lesson 2 moves from questions to data. Students are given the `survey` dataset shown earlier and are asked to interpret it:

The `survey` dataset is a small, simulated 12-row dataset created for the tutorial; it does not include any real student data.

- What does each **row** represent?
- What does each **column** represent?
- Which variables are numeric, and which are categorical?

Students then use R code to gain a first sense of the data:

```
min(survey$minutes_read)
```

```
## [1] 0
```

```
max(survey$minutes_read)
```

```
## [1] 50
```

In narrative form, the tutorial explains that even if all students answer the same question, their responses naturally differ. This **variability** is a central big idea:

> Some students read 0 minutes, some read 15–25 minutes, and some read 40–50 minutes. This natural difference across students is called variability.

An interactive exercise asks students to compute or interpret the range of reading minutes, reinforcing the idea that data spread is meaningful rather than just noise.

Ethically, Lesson 2 invites students to discuss questions such as:

- How might the sample or the way we asked the question introduce bias?
- How could we make the data collection more fair?

These short prompts are designed to surface issues of access, equity, and representation without requiring a full separate ethics lecture.

Lesson 3 – Analyze Patterns

Lesson 3 asks students to move from simply "looking at" data to systematically describing patterns and comparing groups, which GAISE highlights as a key middle-grades milestone in the statistical problem-solving cycle (Bargagliotti et al. 2020). Building on the Lesson 2 reading-time survey, students now use R to construct visual and numeric summaries of a small, simulated sample of twelve seventh-grade students (minutes read the previous night and device used for reading or homework). This keeps the cognitive load on interpreting structure rather than on wrestling with a large dataset, while still mirroring the kinds of data stories students encounter in real life (YouCubed, n.d.).

The tutorial begins with a warm-up scenario in which a teacher claims that "students who use tablets tend to read for more minutes than students who use phones." Students are asked to decide whether this is a claim they could reasonably check with the Lesson 2 survey data and, through a short quiz, to distinguish between descriptive comparisons ("in this sample…") and over-strong causal statements ("tablets cause students to read more"). This explicitly rehearses the difference between **describing** and **explaining** that runs through the GAISE framework (Bargagliotti et al. 2020).

Students then compare device groups using a boxplot of reading minutes:

```
survey %>%
  ggplot(aes(x = device, y = minutes_read)) +
  geom_boxplot() +
  labs(
    title = "Minutes Read by Device Type",
    x = "Device used for reading or homework",
    y = "Minutes read last night"
  )
```

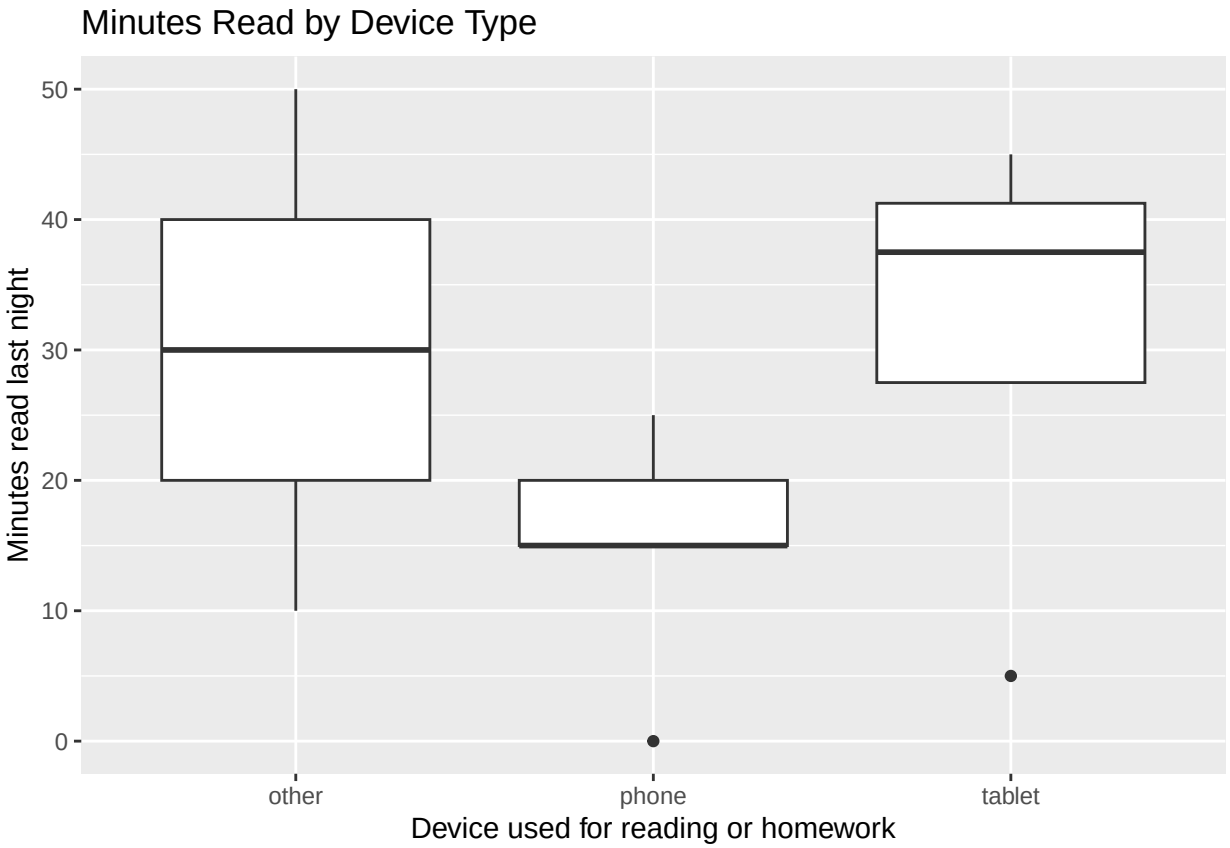

Figure 4: Reused boxplots of minutes read by device type in Lesson 3 (simulated reading survey, n = 12). Students use this graph to compare device groups and practice making cautious, sample-based statements.

In the narrative of the tutorial, I slow down to explain how to read each part of the boxplot—the median, the "middle chunk" of values, and the whiskers—as concrete representations of variability and typical values for each device group. Students answer multiple-choice questions that ask them to decide, for example, whether the tablet group "tends to" read more than the phone group, or whether there is too much overlap to make a strong statement. The focus is on using the graph to support modest, sample-bounded claims, not on memorizing formal definitions.

Next, Lesson 3 introduces a simple grouped summary table that reports the mean reading minutes and sample size for each device category:

```
survey %>%
  group_by(device) %>%
  summarise(
    mean_minutes = mean(minutes_read),
    n = n()
  )
```

```
## # A tibble: 3 x 3
##   device mean_minutes     n
##   <chr>         <dbl> <int>
## 1 other          30       3
## 2 phone          15       5
## 3 tablet         31.2     4
```

Students are guided to read this table in words ("In this sample, tablet users read an average of _____ minutes, based on _____ students") and to connect the numeric summaries back to the boxplot shape. A follow-up quiz contrasts a careful interpretation ("In this sample, tablet users *on average* read more minutes than phone users") with several tempting but incorrect options that erase variability or over-generalize to "all students everywhere." This directly supports youcubed's emphasis on students using precise, sample-aware language when talking about data patterns (YouCubed, n.d.).

The lesson closes with a brief ethics micro-prompt that asks who is—and is not—represented in this twelve-student dataset and what would be fair uses of such limited information. Students are invited to discuss whether it would be reasonable to use this small sample to set school-wide device policies, connecting back to broader concerns about equity and over-claiming from narrow data (Drozda 2023; Chiodo, Müller, and Shah 2025). By the end of Lesson 3, the goal is for students to see boxplots and means as tools for making cautious, evidence-based comparisons between groups, not as magic answers that automatically justify big decisions.

Lesson 4 – Interpret and Communicate

Lesson 4 is the capstone of the sequence and shifts the focus from *making* graphs to *talking* about what graphs and tables show. In GAISE terms, it centers the final "interpret results"

phase of the problem-solving cycle and explicitly foregrounds communication, audience, and limitations (Bargagliotti et al. 2020). Pedagogically, it also responds to youcubed's call for students to write and speak about data in their own words, rather than only calculating numbers (YouCubed, n.d.).

To keep the ideas fresh while staying age-appropriate, Lesson 4 introduces a new but similarly sized dataset about sleep and devices. The data are again simulated within the tutorial (rather than collected from real students) and represent twelve seventh-grade students, with variables for hours of sleep last night, whether they keep a device in their bedroom, and whether they play on a sports team. The dataset is created in an R chunk so that teachers can see and modify the underlying structure:

This `sleep_data` tibble is a small, simulated classroom-sized dataset designed for the lesson; no real student data are used.

```
sleep_data <- tibble(
  student_id    = 1:12,
  grade         = rep(7, 12),
  hours_sleep   = c(8.5, 6.0, 7.0, 5.5, 9.0, 6.5, 7.5, 8.0, 5.0, 8.5, 6.0, 7.0),
  device_bedroom = c(
    "yes", "yes", "no", "yes",
    "no", "yes", "no", "no",
    "yes", "no", "yes", "no"
  ),
  sports_team = c(
    "yes", "no", "yes", "no",
    "yes", "no", "yes", "no",
    "no", "yes", "no", "yes"
  )
)
sleep_data
```

```
## # A tibble: 12 x 5
##    student_id grade hours_sleep device_bedroom sports_team
##         <int> <dbl>       <dbl> <chr>          <chr>
##  1          1     7         8.5 yes            yes
##  2          2     7         6   yes            no
##  3          3     7         7   no             yes
##  4          4     7         5.5 yes            no
##  5          5     7         9   no             yes
##  6          6     7         6.5 yes            no
##  7          7     7         7.5 no             yes
##  8          8     7         8   no             no
##  9          9     7         5   yes            no
## 10         10     7         8.5 no             yes
```

```
## 11         11     7         6   yes            no
## 12         12     7         7   no             yes
```

Students first meet this dataset in plain language, with each column explained in terms of the story ("hours they slept last night," "whether they keep a device in their bedroom," and so on). This mirrors classroom practices in which teachers might build a small, local dataset together on the board, but here it is fully reproducible inside R.

The heart of Lesson 4 is a carefully scaffolded progression from reading a prepared visualization to critiquing and then writing data stories:

1. **A question and a graph.**
   Students are shown a boxplot of `hours_sleep` by `device_bedroom` along with a small summary table of mean hours of sleep for the two groups. They practice making sense of a graph that someone else has prepared, articulating what the graph shows about typical sleep patterns for students with and without devices in their bedrooms.

2. **A model data story.**
   The tutorial presents a short "model" blurb that weaves together the graph and table, names the sample (twelve seventh-grade students), reports the direction and rough size of the difference, and explicitly states that we would need more data before making school-wide claims. The text is broken down sentence by sentence—question, evidence, and limitation—so that students can see how a strong, honest data story is constructed.

3. **Critiquing and choosing blurbs.**
   Through a multiple-choice quiz, students compare several alternative blurbs. Some over-state the findings ("always sleep less" or "schools should ban devices from bedrooms"), while others are vague or omit key details. Students select the most careful option and receive feedback explaining why the others go too far or fail to cite evidence. This directly targets common misconceptions about what small, observational datasets can support.

4. **Guided writing practice.**
   Finally, students draft their own short data story using sentence starters and prompts ("In this sample…", "On average…", "We would need more data to…"). They are encouraged to mention the sample size, describe the pattern in both words and numbers, and acknowledge the observational nature of the data. Teachers can have students share their blurbs aloud or in writing, reinforcing communication skills alongside statistical reasoning.

The lesson ends with an ethics micro-prompt asking students to imagine a principal who wants to use this twelve-student sleep dataset to set a school-wide rule about devices in bedrooms. Students discuss whether that would be fair, what additional information would be needed, and whose experiences might be missing. This framing reinforces the larger theme of the project: data stories are never purely technical, but always embedded in decisions about people, power, and fairness (Drozda 2023; Chiodo, Müller, and Shah 2025). By the

end of Lesson 4, students have had multiple chances to read, critique, and write about data, completing the full GAISE cycle from question to interpretation.

Ethical Challenges and Responses

The ethics component of this project is documented in detail below. I summarize the main challenges I encountered and the design choices I made to address them, drawing on the *Teaching Resources for Embedding Ethics in Mathematics* and related literature.

1. Data Privacy and Consent

A major ethical consideration in this project was ensuring that any data used, especially if it involved students, teachers, or minors, respected privacy and consent standards. Because my *R for Everyone* curriculum is designed for middle school classrooms, I needed to make sure no identifiable or sensitive information from students would ever be stored, shared, or analyzed.

How I addressed it:

- Used only simulated or publicly available datasets rather than collecting student-level data.
- Designed the learnr lessons to process data locally in RStudio or RStudio Cloud on the user's device, without exporting responses externally.
- Removed any automatic data-logging features from quizzes or free-response questions to avoid saving identifiable information.

In classroom use, teachers may still want to review or archive student thinking. To support this without compromising privacy, the curriculum is designed so that any persistent storage of responses happens outside the R tutorials in a teacher-controlled environment (for example, a Google Form or learning management system). Students would copy written reflections or upload screenshots to that space, and any identifying information would be governed by the school's existing consent and data policies. The curriculum itself does not collect or transmit student responses.

2. Informed Participation

Since this project may later involve classroom pilots, I also considered what it would mean for teachers and students to participate voluntarily and knowingly.

How I addressed it:

- Created a teacher validation rubric for expert review rather than using live student data in this phase.

- Included a short consent statement in the teacher feedback form clarifying that participation is voluntary, anonymous, and for educational research purposes only.

3. Algorithmic Bias and Representation

Teaching data science to younger audiences raises the risk of unintentionally reinforcing biased or incomplete views of data. For instance, datasets that overrepresent certain regions or demographics could lead students to draw skewed conclusions.

How I addressed it:

- Added discussion prompts encouraging students to question data sources, collection methods, and who might be excluded.

- Included a short "ethics micro-prompt" in each lesson, for example: *"Whose data is missing here, and how might that affect the story this graph tells?"*

4. Accessibility and Digital Equity

Because not all schools or students have the same access to technology, the project needed to be designed with varying levels of connectivity and device availability in mind.

How I addressed it:

- Structured the learnr lessons so they can be run offline in RStudio or RStudio Cloud using minimal computing resources.

- Provided printable summary worksheets and screenshots for teachers with limited device access.

- Planned for open-source release (e.g., via shinyapps.io) so that all materials remain free to use and adapt.

5. Pedagogical Transparency

Finally, I wanted to ensure that the project models ethical data literacy itself: students should learn not only *how* to analyze data but *why* ethical reflection matters.

How I addressed it:

- Embedded reflection checkpoints after each coding task, asking students to consider reliability, limitations, and implications of their analyses.

- Used plain-language explanations to make ethics approachable rather than abstract or punitive.

- Documented all datasets, their origins, and transformations in reproducible R Markdown cells.

Summary

Throughout this capstone, ethical considerations guided both the design (how data are collected and represented) and the delivery (how learners interact with R safely and thoughtfully). The overarching goal was not just to avoid harm, but to teach ethical awareness as a core part of data science practice.

Conclusion and Future Work

This capstone project demonstrates that it is possible to create **digestible, ethics-aware data science lessons for middle school students** using R and learnr. By grounding the four-lesson sequence in the GAISE statistical problem-solving framework and the youcubed data big ideas, and by incorporating recurring ethics micro-prompts inspired by recent work on embedding ethics in mathematics, the project offers one concrete model for integrating tool use, statistical thinking, and ethical reflection.

The analysis in this write-up focuses on design choices and alignment with established frameworks rather than on classroom outcome data, reflecting the current stage of the project.

Future work will involve:

- Collecting teacher feedback through the validation rubric and possibly through semi-structured interviews.
- Iterating on lesson content, pacing, and ethics prompts in response to that feedback.
- Exploring opportunities to pilot the lessons with local middle schools or after-school programs and to study how students engage with both the statistical and ethical dimensions.

Ultimately, the project aims not only to provide a workable set of materials, but also to contribute to the broader conversation about what it means to teach data science **for everyone**, in ways that are technically accessible, intellectually honest, and ethically attentive.

Attached below are the links to the Lessons

- **Lesson 1 tutorial link:** *https://tene.shinyapps.io/Lesson1/*
- **Lesson 2 tutorial link:** *https://tene.shinyapps.io/Lesson2/*
- **Lesson 3 tutorial link:** *https://tene.shinyapps.io/Lesson3/*
- **Lesson 4 tutorial link:** *https://tene.shinyapps.io/Lesson4/*

References

Bargagliotti, Anna, Christine Franklin, Pip Arnold, Rob Gould, Sheri Johnson, Leticia Perez, and Denise A. Spangler. 2020. *Pre-k–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*. Alexandria, VA: American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaiseiiprek-12_full.pdf.

Chiodo, Maurice, Dennis Müller, and Rehan Shah. 2025. "Teaching Resources for Embedding Ethics in Mathematics: Exercises, Projects, and Handouts." arXiv. https://doi.org/10.48550/arXiv.2310.08467.

Drozda, Zarek. 2023. "Data Science Is Vital to Student Success. So Why Are Outcomes Going Down?" Data Science 4 Everyone. https://d2c47c0f-cda1-420e-ae77-f1279b4ac7ea.usrfiles.com/ugd/d2c47c_067273b2bef041c7ae08cab6c7a3be8c.pdf.

Education, Stanford Graduate School of. 2021. "How to Teach Data Science in k–12 Schools? Stanford-Led Team Launches 'Big Ideas'." October 13, 2021. https://ed.stanford.edu/news/how-teach-data-science-k-12-schools-stanford-led-team-launches-big-ideas.

qmd-lab. 2025. "Closeread: A Quarto Format for Scrollytelling." 2025. https://closeread.dev/guide/.

Schloerke, Barret, Garrick Aden-Buie, RStudio, and PBC Posit Software. 2025. *Learnr: Interactive Tutorials for r*. https://rstudio.github.io/learnr/.

Xie, Yihui. 2024. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://bookdown.org/yihui/bookdown/.

YouCubed. n.d. "Data Science k–10 Big Ideas." https://www.youcubed.org/data-big-ideas/.