Understanding the Development of Students' Multivariate Statistical Thinking in a Data Visualization Course

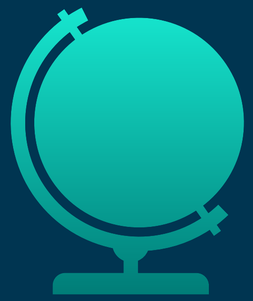
Chelsey A. Legacy
Quantitative Methods in Education
University of Minnesota

July 12, 2022

Outline

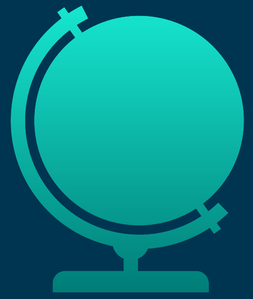
- Background
- Methods
- Results
- Discussion & Conclusions
- Acknowledgements
- Questions

Background



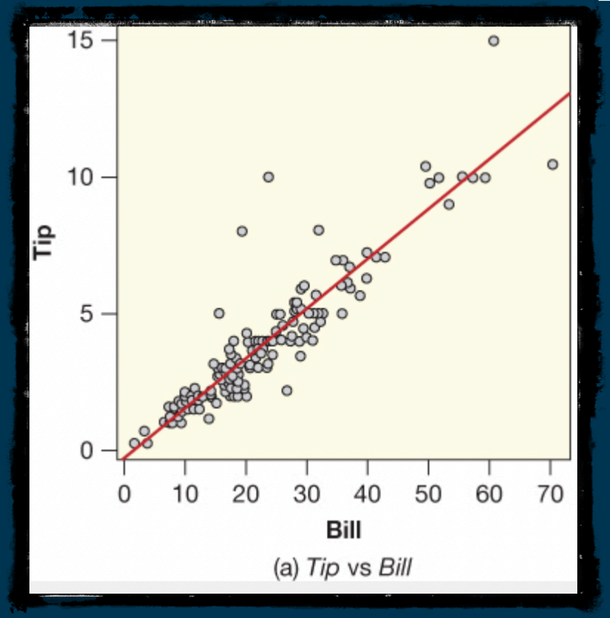
- Data is collected all around us
  - Large data sets with many different variables are being collected
- Students not often introduced to datasets with more than two variables

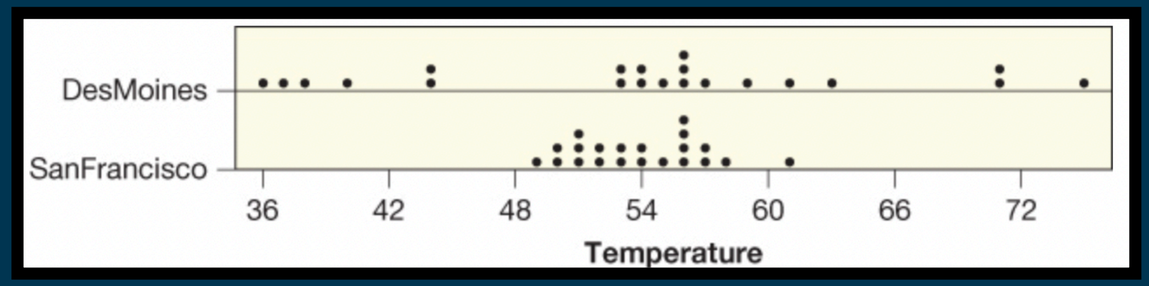
Considerations



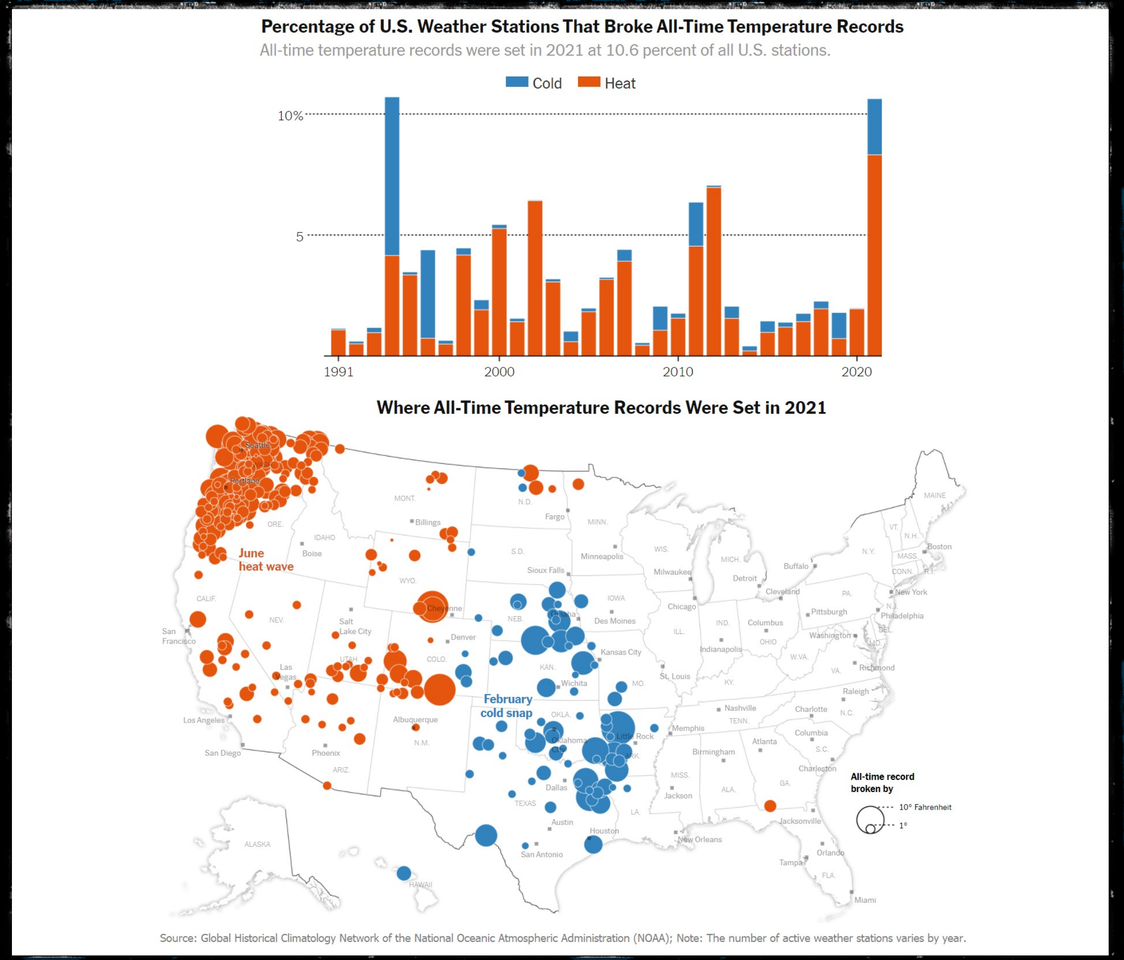
- Looking at relationships between two variables discounts possible effects of other variables
- Media often depicts visualizations with many variables

For example…



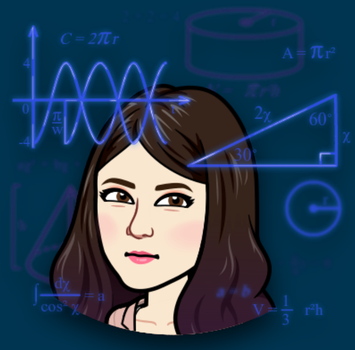


NYT What's Going On in This Graph?



Unlocking the Power of Data Textbook

Multivariate Thinking



GAISE Recommendation:

- “Teach statistical thinking….Give students experience with multivariable thinking.” (GAISE College Report ASA Revision Committee, 2016, p. 3)

More specifically:

> “[m]ultivariable thinkers can employ an intuitive sense of concepts such as confounding, mediation, association, interaction, and causality to create a more complete understanding of relationships in their data” (Adams, Baller, Jonas, Joseph, & Cummiskey, 2021, p. S125).

Literature

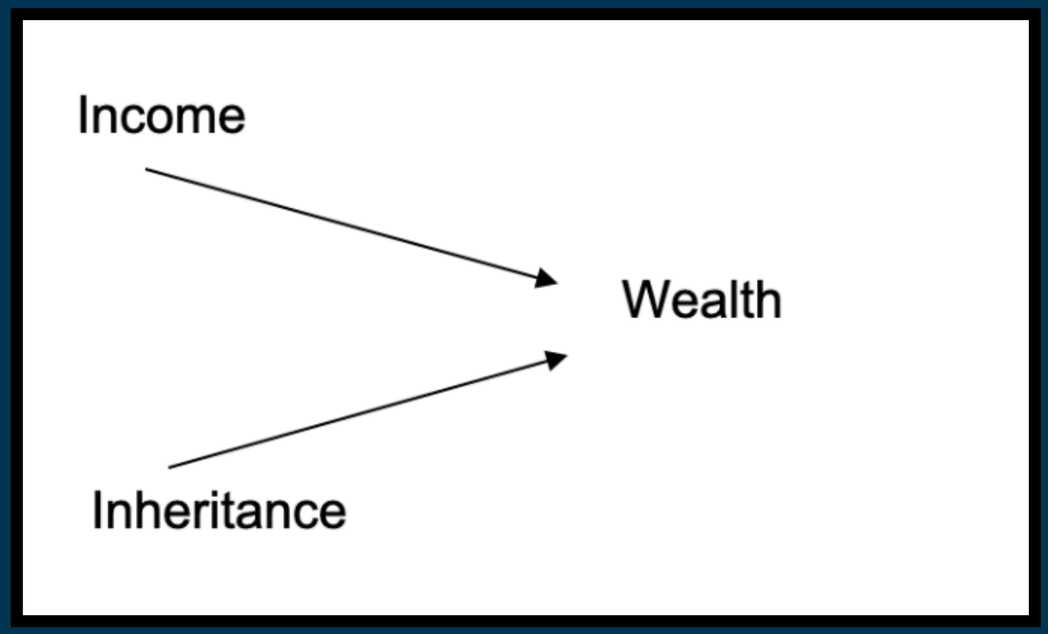
Challenge	Literature
Difficulties with Respect to Conclusions	Abdelhadi, 2016; Casparo & Grulich, 2019; Gil and Gibbs, 2017 Kuhn, 2007; Kuhn 2008; Kuhn et al., 2015; Ridgway et al., 2007
Difficulties with Respect to Size of Effect	Kuhn et al., 2015; Ridgway et al., 2007
Difficulties with Respect to Context	Abdelhadi, 2016; Kuhn, 2007; Kuhn et al., 2015

Causal Reasoning

- Inclination to make causal claims with variables (e.g. Ahn, Kalish, Medin, & Gelman, 1995; Gopnik & Schulz, 2007; Kahneman, 2013)
- Typically we teach “correlation is not causation”
  - Random assignment allows for causation
- Not the whole picture
  - correlation might indicate a relationship worthy of further study (Pearl & Mackenzie, 2018)

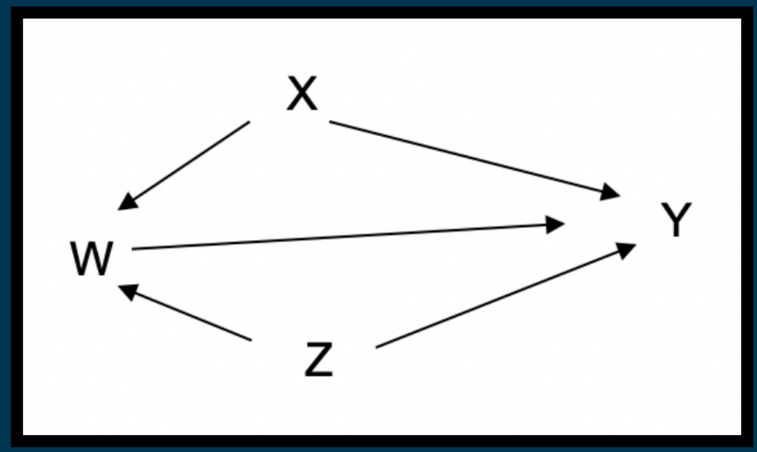
Developing Causal Reasoning

- Provide a framework for considering causal relationships
  - Directed - Acyclic Graphs (DAGs)



Pearl's DAGs

- But only determine causality with advanced methods
  - Not likely to be covered in introductory statistics courses
- Suggest hypothetical causal variables
- Graphically explore those relationships



☑ Background

Methods

Research Questions

1. How does students' ***multivariate thinking develop as they take part in a series of activities*** designed to introduce and promote reasoning with multiple variables? How do student responses to questions requiring multivariate thinking change throughout the semester?

2. What ***challenges*** surrounding multivariate thinking ***persist*** after taking part in the intervention? Do any ***new challenges*** emerge after the completion of these activities?

Student Learning Outcomes -> Materials

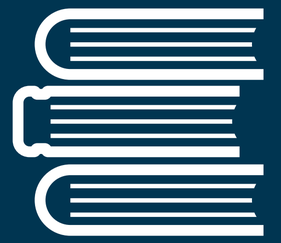
1. Create graphs displaying the relationships among three or four variables in one plot
2. Explain the relationships among three to four variables using graphs
3. Identify data as observational
4. State the limitations in making causal claims with observational data
5. Create DAGs to guide analysis of relationships among variables
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables
7. Develop a hypothesis about variables not investigated and their relation to an outcome variable

Revisions

10 in class Activities

3 Assignments

Methods: Data Collection



Course Information

- Fall 2021 - Hybrid
- EPSY 1261: Visualization Course
  - 2 Sections (Section 1: n = 18; Section 2: n = 25)

Data Collected

- Collected Assignments from all students
- Observed 1 student in each section
- Cognitive Interview with 3 students about the final assignment

Methods: Analysis

Research Question 1: Development of MVT

- Analysis of in class student observations
- Qualitative Analysis of % correct for each learning outcome across assignments

Research Question 2: Challenges of MVT

- Analysis of in class student observations
- Qualitative Analysis of themes from assignments
- Analysis of cognitive interviews with students about final assignment

☑ Background

☑ Methods

Results

Results from Observations

Jordan

- Learned code for creating multivariate plots in R
- Reasoning about visualizations continually challenging
- They often had correct interpretations about the graph but used only outside knowledge to support their answers rather than evidence from the plot.
  - For example:

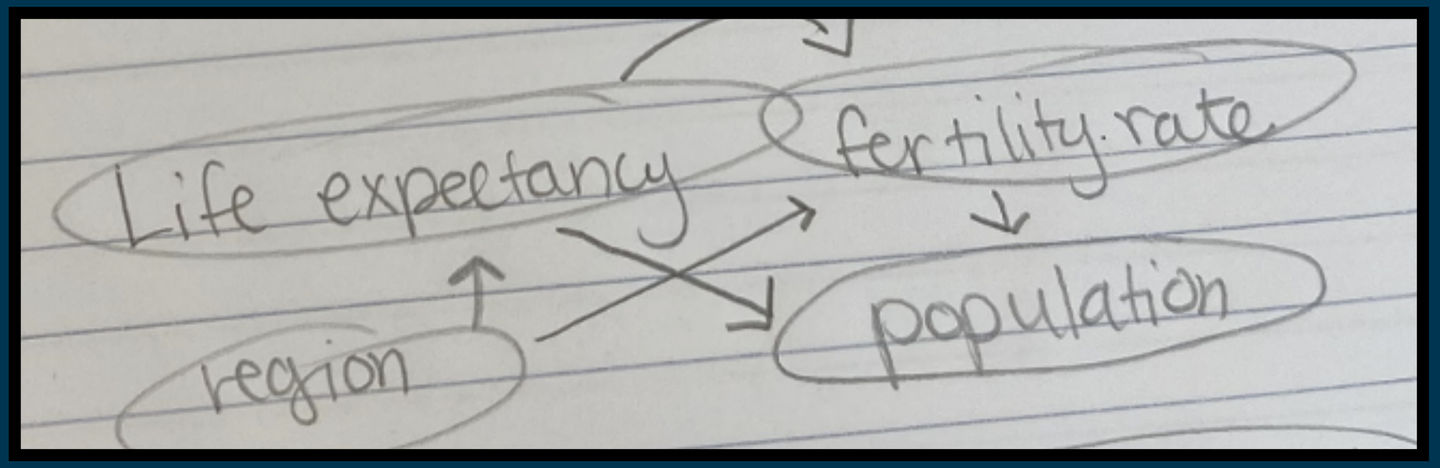
> "region impacts life expectancy because the resources you have in those regions are big determiners in what happens in your life, and thus how long you'll live."

Results from Observations

Jordan

- Unsure why multiple variables were needed to answer research questions
- Occasionally did not update DAG based on visualization

> "I'm not seeing much of a difference between population size and life expectancy for Africa, or really very many of the other countries."

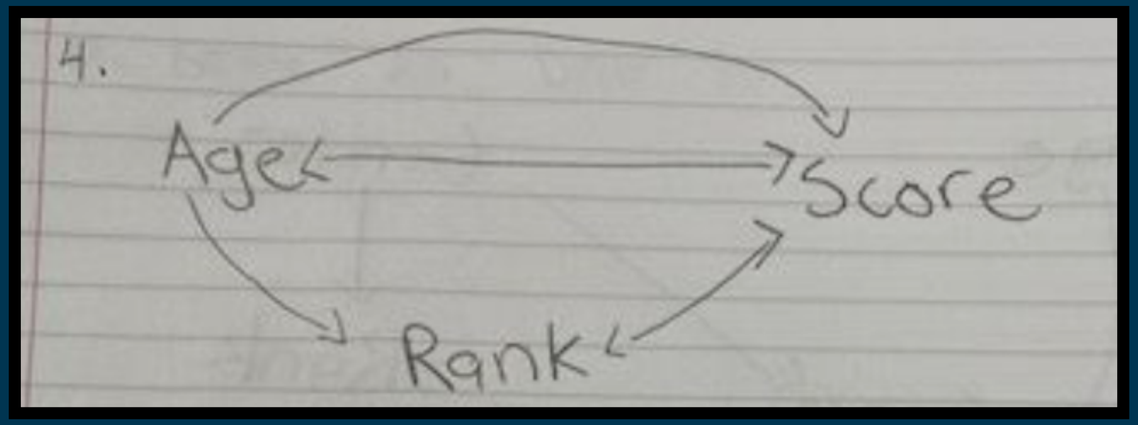


Results from Observations

Kennedy

- They came a long way in improving their graphing skills, creating various scatterplots with creative aesthetics.
- But their proposed DAGs still were not always aligned with what was depicted in the plots and occasionally depicted causal relationships that were not possible.

"based on what I was feeling"



Results from Assignments

Learning Outcome	% Correct for HW 1 (n=38)	% Correct for HW 2 (n=37)	% Correct for HW 3 (n=33)
1. Create graphs displaying the relationships among three or four variables in one plot	50%	91.9%	93.8%
2. Explain the relationships among three to four variables using graphs	50%	73%	21.2%
3. Identify data as observational	26.3%	29.7%	66.7%
4. Explain the limitations in making causal claims with observational data	NA	NA	18.2%
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	71%	73%	66.7%
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	NA	67.6%	69.7%
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	94.7%	91.9%	97%

Results from Assignments

Learning Outcome	% Correct for HW 1 (n=38)	% Correct for HW 2 (n=37)	% Correct for HW 3 (n=33)
1. Create graphs displaying the relationships among three or four variables in one plot	50%	91.9%	93.8%
2. Explain the relationships among three to four variables using graphs	50%	73%	21.2%
3. Identify data as observational	26.3%	29.7%	66.7%
4. Explain the limitations in making causal claims with observational data	NA	NA	18.2%
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	71%	73%	66.7%
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	NA	67.6%	69.7%
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	94.7%	91.9%	97%

Results from Assignments

Learning Outcome	% Correct for HW 1 (n=38)	% Correct for HW 2 (n=37)	% Correct for HW 3 (n=33)
1. Create graphs displaying the relationships among three or four variables in one plot	50%	91.9%	93.8%
2. Explain the relationships among three to four variables using graphs	50%	73%	21.2%
3. Identify data as observational	26.3%	29.7%	66.7%
4. Explain the limitations in making causal claims with observational data	NA	NA	18.2%
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	71%	73%	66.7%
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	NA	67.6%	69.7%
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	94.7%	91.9%	97%

Results from Assignments

Learning Outcome	% Correct for HW 1 (n=38)	% Correct for HW 2 (n=37)	% Correct for HW 3 (n=33)
1. Create graphs displaying the relationships among three or four variables in one plot	50%	91.9%	93.8%
2. Explain the relationships among three to four variables using graphs	50%	73%	21.2%
3. Identify data as observational	26.3%	29.7%	66.7%
4. Explain the limitations in making causal claims with observational data	NA	NA	18.2%
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	71%	73%	66.7%
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	NA	67.6%	69.7%
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	94.7%	91.9%	97%

Results from Assignments

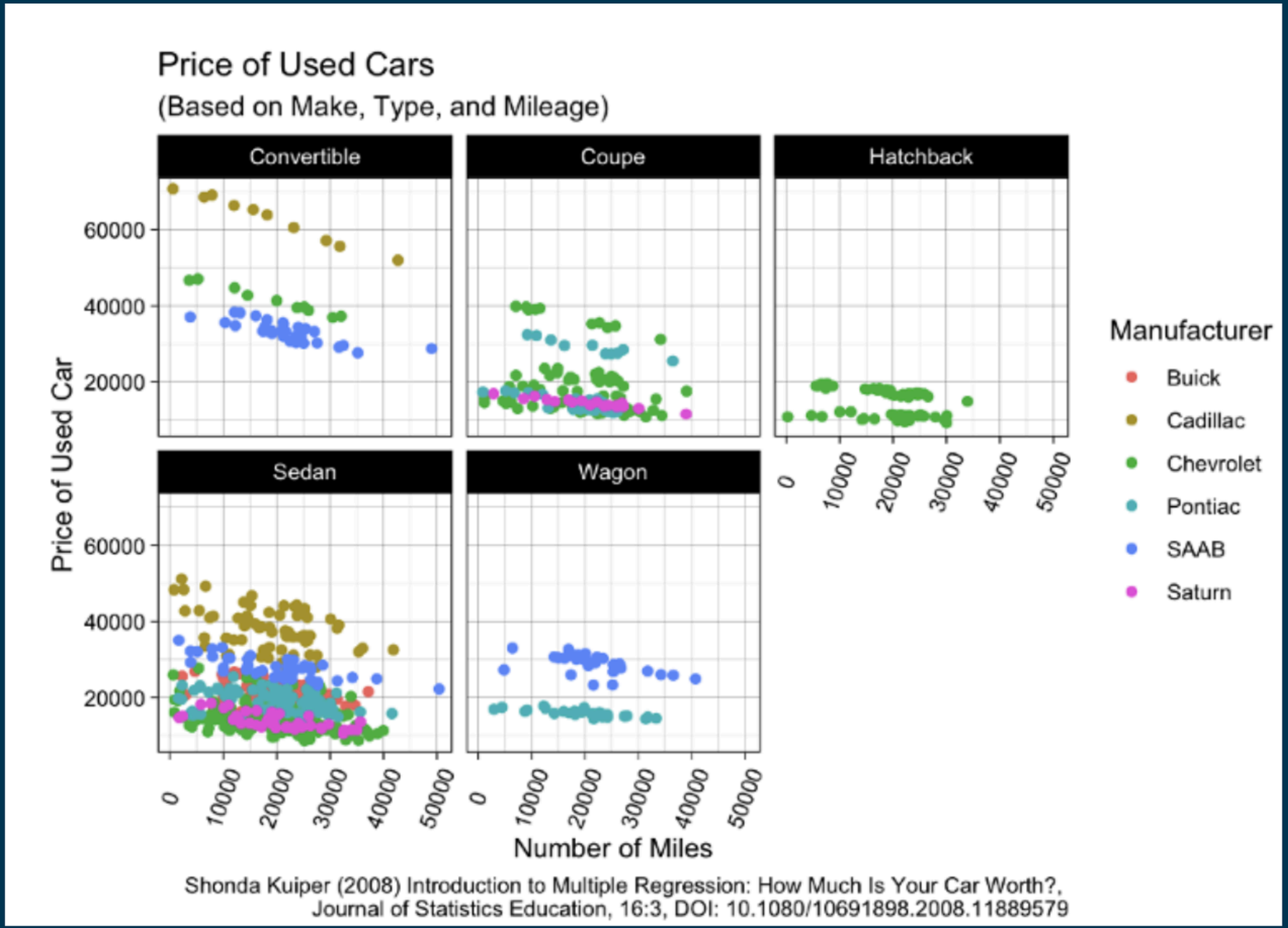
Learning Outcome	% Correct for HW 1 (n=38)	% Correct for HW 2 (n=37)	% Correct for HW 3 (n=33)
1. Create graphs displaying the relationships among three or four variables in one plot	50%	91.9%	93.8%
2. Explain the relationships among three to four variables using graphs	50%	73%	21.2%
3. Identify data as observational	26.3%	29.7%	66.7%
4. Explain the limitations in making causal claims with observational data	NA	NA	18.2%
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	71%	73%	66.7%
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	NA	67.6%	69.7%
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	94.7%	91.9%	97%

Results from Cognitive Interviews

- Interviewed about final Cars Assignment
- Created multivariate graphs
  - Created less insightful plots when told to create something new
- Trouble identifying observational data
- Easily created DAGs despite “little knowledge about cars”

Question #15

Your friend found a GM made vehicle from 2005 **for $40,000. Under what conditions** would you recommend they buy it? Explain your answer below using evidence from your final plot and DAG.



Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

☑ Background

☑ Methods

☑ Results

Discussion & Conclusions

Answering Research Question 1

1. How does students' **multivariate thinking develop as they take part in a series of activities** designed to introduce and promote reasoning with multiple variables? How do student responses to questions requiring multivariate thinking change throughout the semester?

    - Improved in creating graphs with multiple variables
    - Difficulty describing all relationships in a plot
    - Difficulty identifying observational data & knowing when to make causal claims
    - Consistently able to create/update DAGs

Answering Research Question 2

2. What **challenges** surrounding multivariate thinking **persist** after taking part in the intervention? Do any **new challenges** emerge after the completion of these activities?

   - Still difficult to consider multiple variables
   - Answering questions about the visualizations was more difficult than anticipated
   - Understanding why we need to look at multiple variables not clear until too late
   - Context is always a challenge

Limitations

- Hybrid structure of the course/COVID
- Coding in R could be viewed as a barrier
- Contexts of datasets

Implications for Teaching

- DAGs were useful
  - Could have been more useful with more discussion
- Be thoughtful about technology when teaching MVT
- Provide students with the time and scaffolding to think about the data itself

Future Work

- More formal exploration of the learning trajectory outlined here
  - What are the SLOs here?
- What about other graphs (like network graphs)?
- Pearl’s DAGs to identify variables to stratify on
- When to teach Simpson’s Paradox? - sooner!

Acknowledgements

- Uncountably infinite “thanks" to….
  - My advisors: Bob delMas & Andy Zieffler
  - Committee members: Erin Baldinger & Sashank Varma
  - EPSY 1261 Instructor: Suzanne Loch
  - Statistics education cohort: Vimal Rao, Jonathan Brown, Regina Lisinker
  - Students in EPSY 1261 in Fall 2021 for participating

- ☑ Background
- ☑ Methods
- ☑ Results
- ☑ Discussion & Future Work
- ☑ Acknowledgements

Questions?

References

Adams, B., Baller, D., Jonas, B., Joseph, A.-C., & Cummiskey, K. (2021). Computational skills for multivariable thinking in introductory statistics. Journal of Statistics and Data Science Education, 29, S123?S131. http://doi.org/10.1080/10691898.2020.1852139

GAISE College Report ASA Revision Committee, “Guidelines for Assessment and Instruction in Statistics Education College Report 2016,” http://www.amstat.org/education/gaise.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covaria- tion versus mechanism information in causal attribution. Cognition, 54(3), 299-352. http://doi.org/10.1016/0010-0277(94)00640-7

Gopnik, A., & Schulz, L. (Eds.). (2007). Causal learning: Psychology, philosophy, and computation. Oxford ; New York: Oxford University Press.

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical edu- cation: Looking back, looking forward. The American Statistician, 69(2), 138-145. http://doi.org/10.1080/00031305.2015.1032435

Kahneman, D. (2013). Thinking, fast and slow (1st pbk. ed). New York: Farrar, Straus; Giroux.

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2020). *Statistics: Unlocking the power of data*. John Wiley & Sons.

Network, T. L. (2022, March 10). *What's going on in this graph? | extreme temperatures*. The New York Times. Retrieved July 4, 2022, from https://www.nytimes.com/2022/03/10/learning/whats-going-on-in-this-graph-march-16-2022.html

Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect (First edition). Basic Books.