

Statistical Graphics in the News, Media, and Scientific Publications: Deceptive Practices and How to Fix Them

Jürgen Symanzik

Utah State University

Logan, UT, USA

juergen.symanzik@usu.edu

<https://www.usu.edu/math/symanzik/>

IX Latin American Conference on Statistical Computing
(LACSC 2025)

Valparaíso, Chile (Virtual)

November 7, 2025

Contents

- Motivation
- “How to Display Data Badly”
- Part I: Misleading Graphics from the Web and Statistical Literature
 - The Bad and the Ugly Graphics from the Covid Years
 - The Bad and the Ugly Graphics from the Statistical Literature
- Part II: Graphics Created with the Help of Artificial Intelligence (AI)
 - AI Graphics Tasks in the Classroom
 - Overall Assessment of LLM Graphics
- Overall Conclusions
- Appendix

Motivation

Motivation

- ❑ As an expert in statistical graphics, it hurts to see bad graphics in the news, on the web, and produced by students.
- ❑ Even worse are bad graphics in science (including those from articles in peer-reviewed statistics journals).
- ❑ As the former chair of a task force of the *Statistical Graphics Section of the American Statistical Association (ASA)*, we reevaluated the winning posters of the annual ASA poster competition for children from kindergarten to grade 12 – and noticed many bad graphics, even among the winners.
- ❑ In the early 2020s, numerous bad graphics related to Covid-19 were published.

Bad Graphics Facts

- Gordon & Finch (2015)* sampled and rated graphics from major statistics and applied science journals.
- They observed: *“nearly 40% of the 97 graphs we sampled were rated as poor, with no striking differences between the applied science and statistics graphs.”*
- They proposed: *“Progress over the next 30 years needs to be supported by changes in software defaults.”*

* Gordon, I. & Finch, S. (2015): Statistician Heal Thyself: Have We Lost the Plot? Journal of Computational and Graphical Statistics, 24(4), 1210-1229,
<https://doi.org/10.1080/10618600.2014.989324>

“How to Display Data Badly”

“How to Display Data Badly”

- Wainer (1997)* proclaimed: “The aim of good data graphics is to display data accurately and clearly. [...]
- Thus, if we wish to display data badly, we have three avenues to follow.
 - A. Don't show much data.
 - B. Show the data inaccurately.
 - C. Obfuscate the data.” [i.e., show the data unclearly]

* Wainer, H. (1997): Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot, Copernicus/Springer, New York, NY.

A. Don't show much data

- Rule 1: Show as little data as possible (minimize the data density).
- Rule 2: Hide what data you do show (minimize the data/ink ratio).

B. Show the data inaccurately

- Rule 3: Ignore the visual metaphor altogether.
- Rule 4: Only order matters.
- Rule 5: Graph data out of context.

C. Obfuscate the data

- Rule 6: Change scales in mid-axis.
- Rule 7: Emphasize the trivial (ignore the important).
- Rule 8: Jiggle the baseline.
- Rule 9: Alabama first!
- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.
- Rule 11: More is murkier: (a) more decimal places and (b) more dimensions.
- Rule 12: If it has been done well in the past, think of a new way to do it.

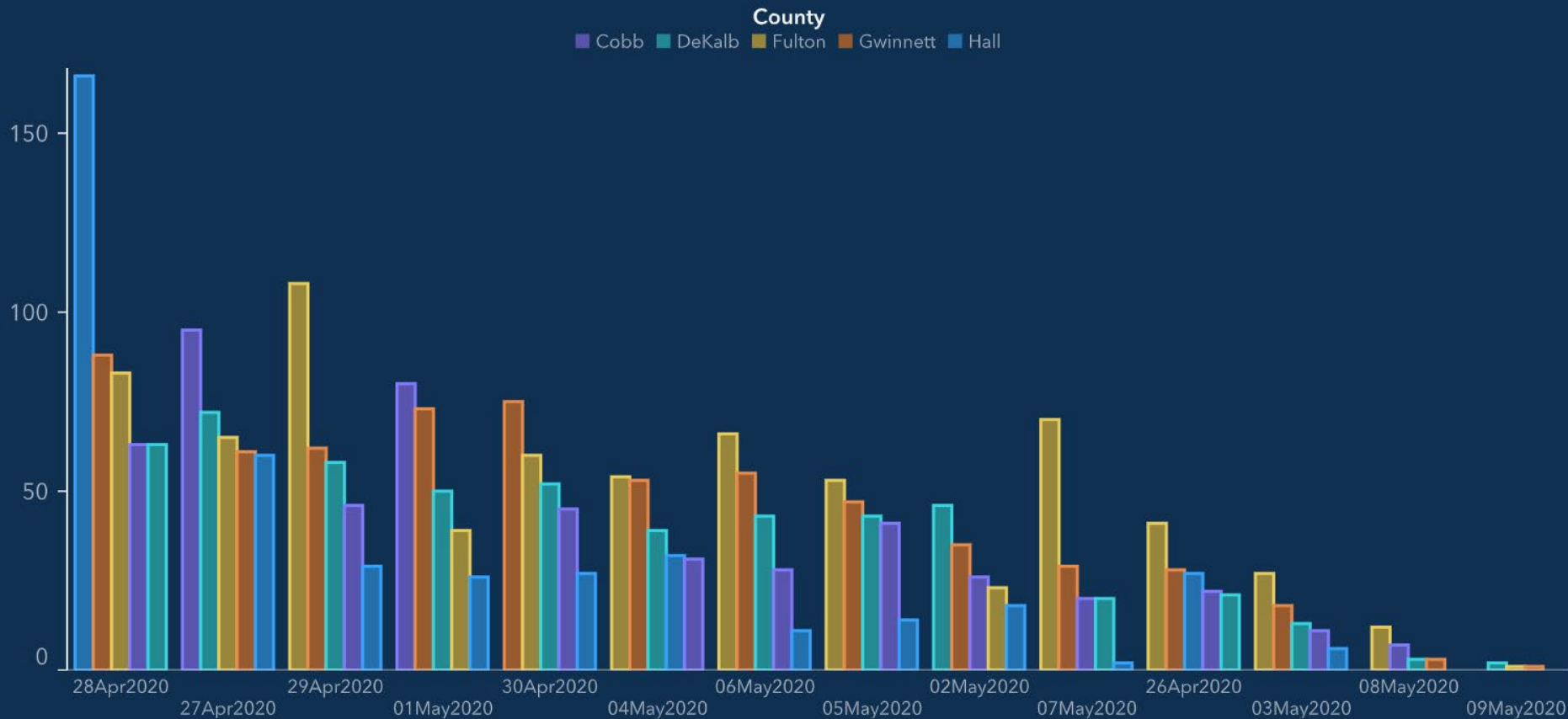
**Part I: Misleading Graphics from
the Web and Statistical Literature**

The Bad and the Ugly Graphics
from the Covid Years

Georgia's Covid-19 Cases

Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

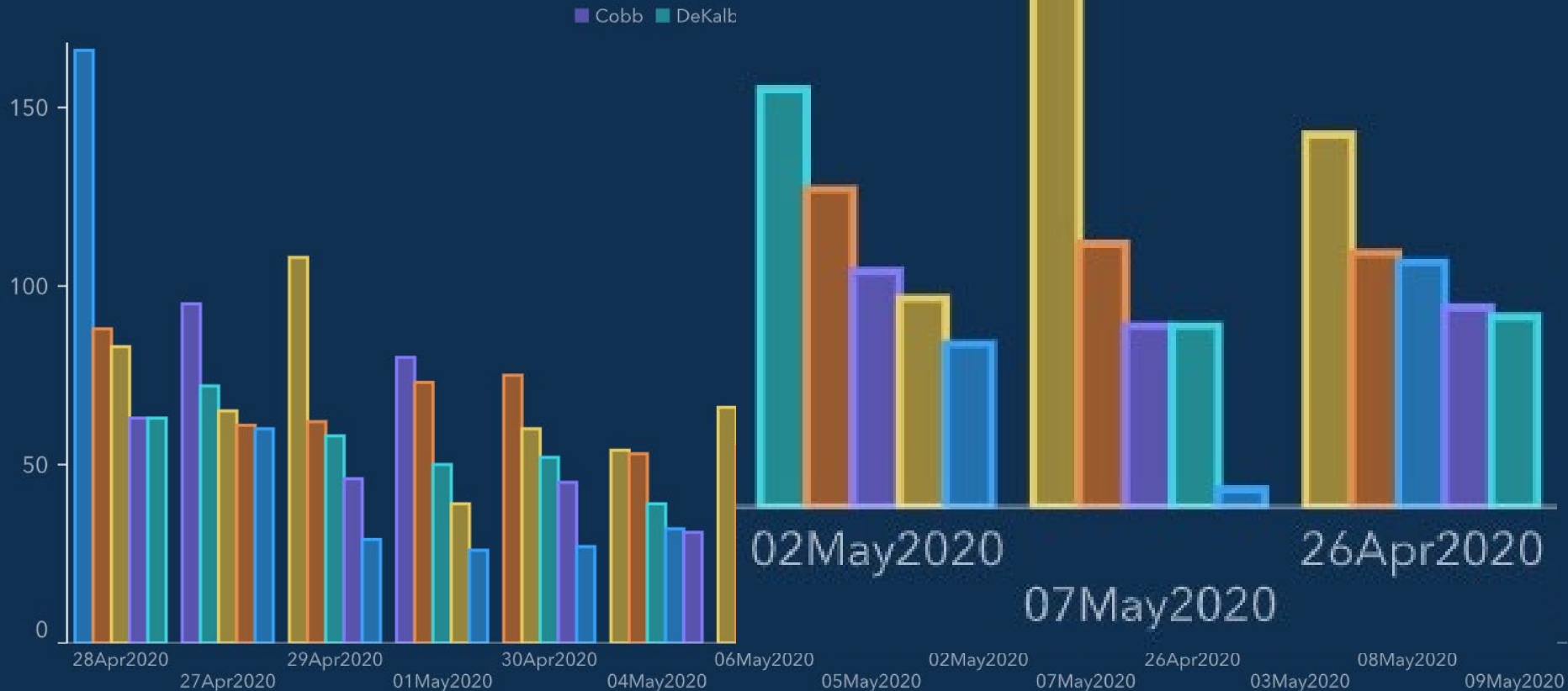


Georgia Department of Public Health (Graph no longer available from source)

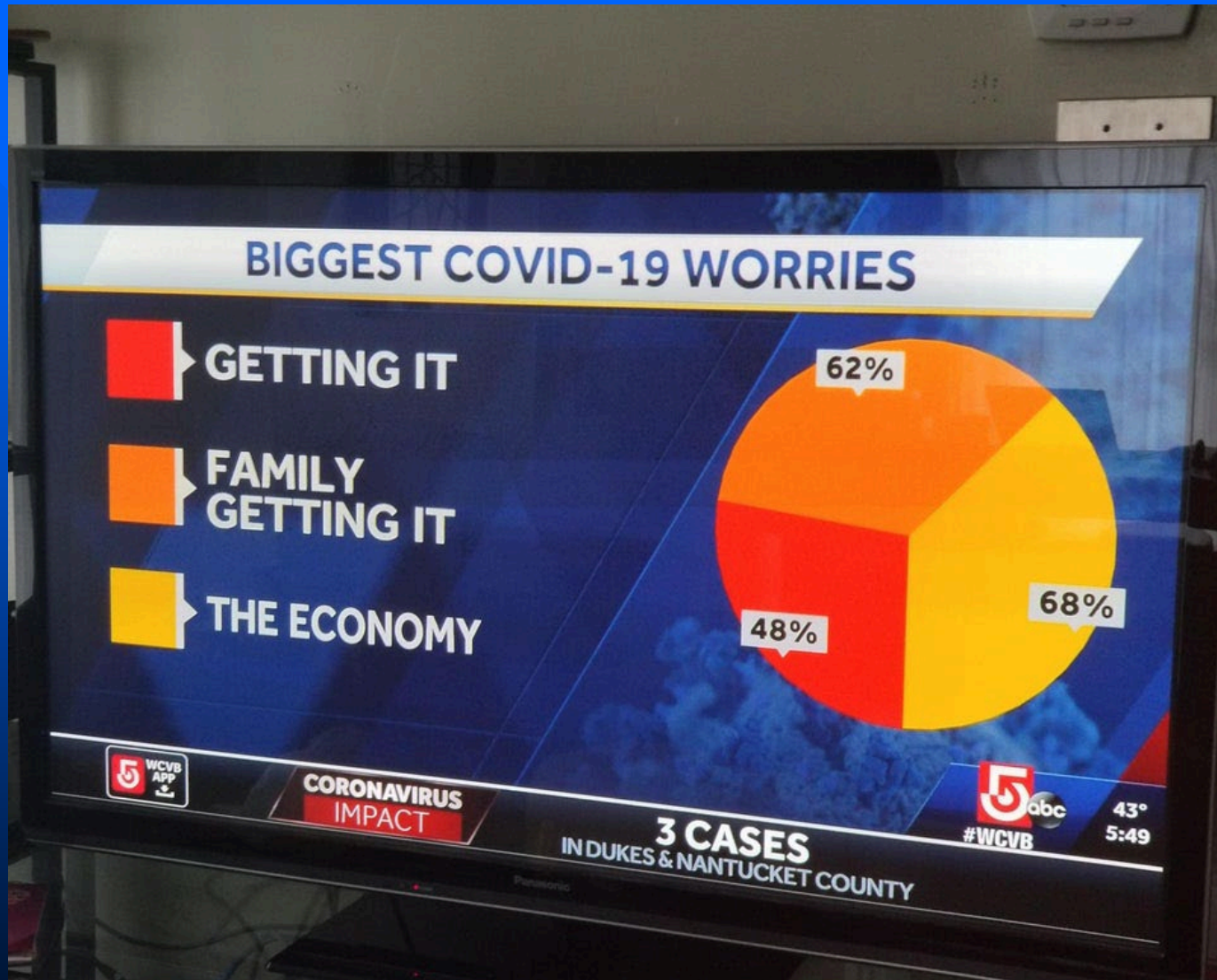
Georgia's Covid-19 Cases – Close-Up

Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over also represents the number of deaths and hospitalizations in €

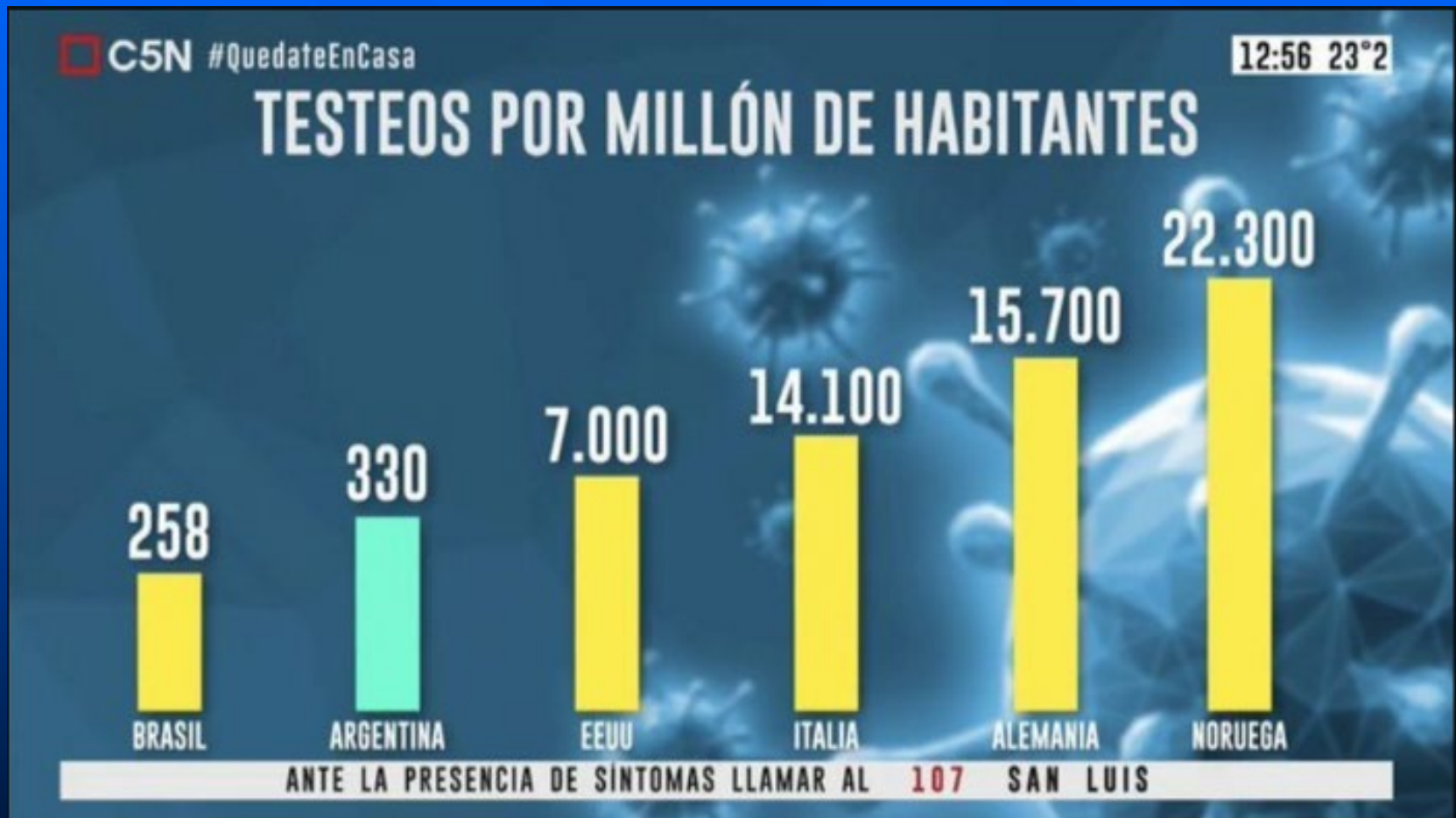


A Pie Chart out of 178%



WCVB Channel 5 Boston (Graph no longer available from source)

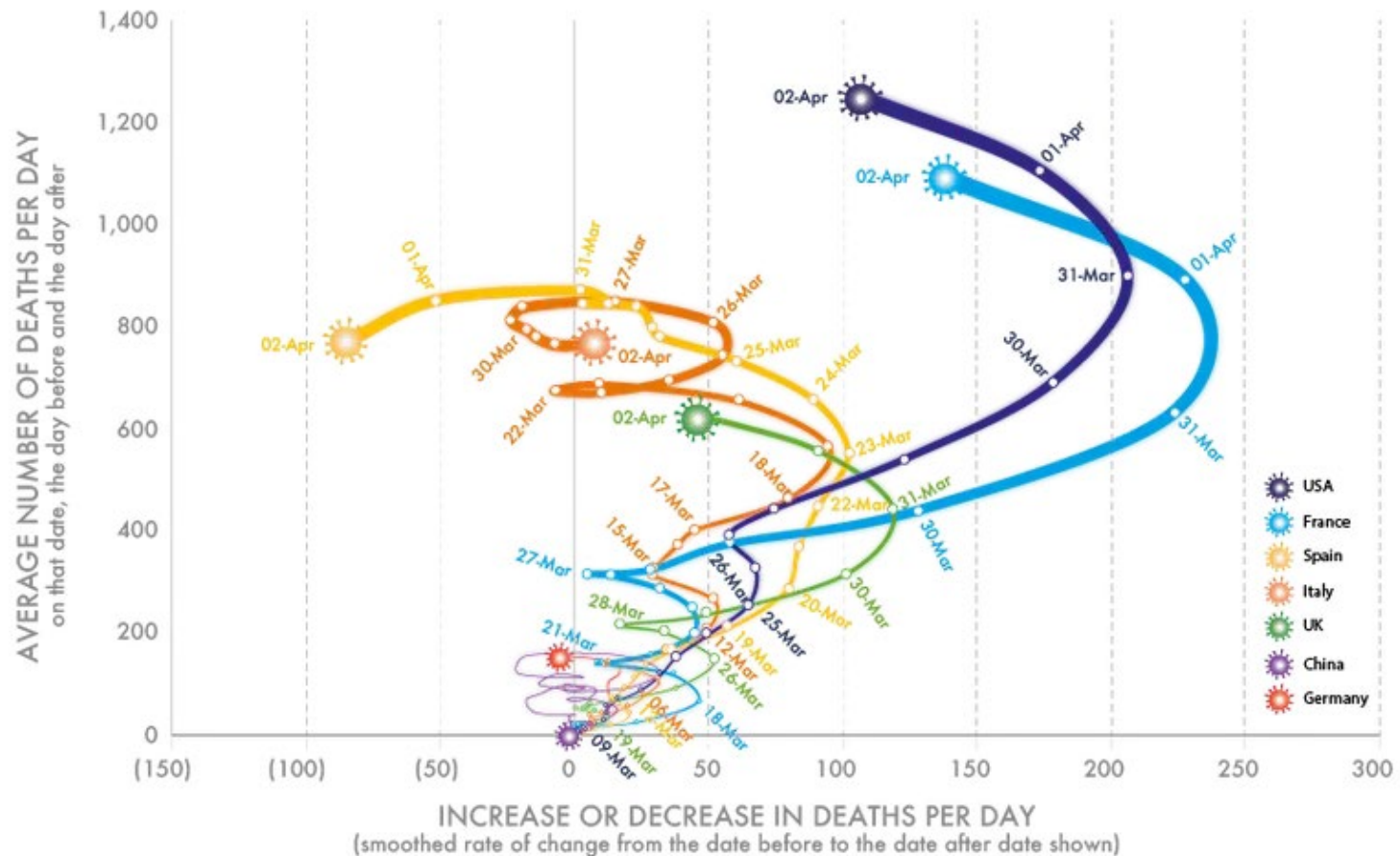
Latin American-European Comparison



Argentinian TV Channel C5N (Graph no longer available from source)

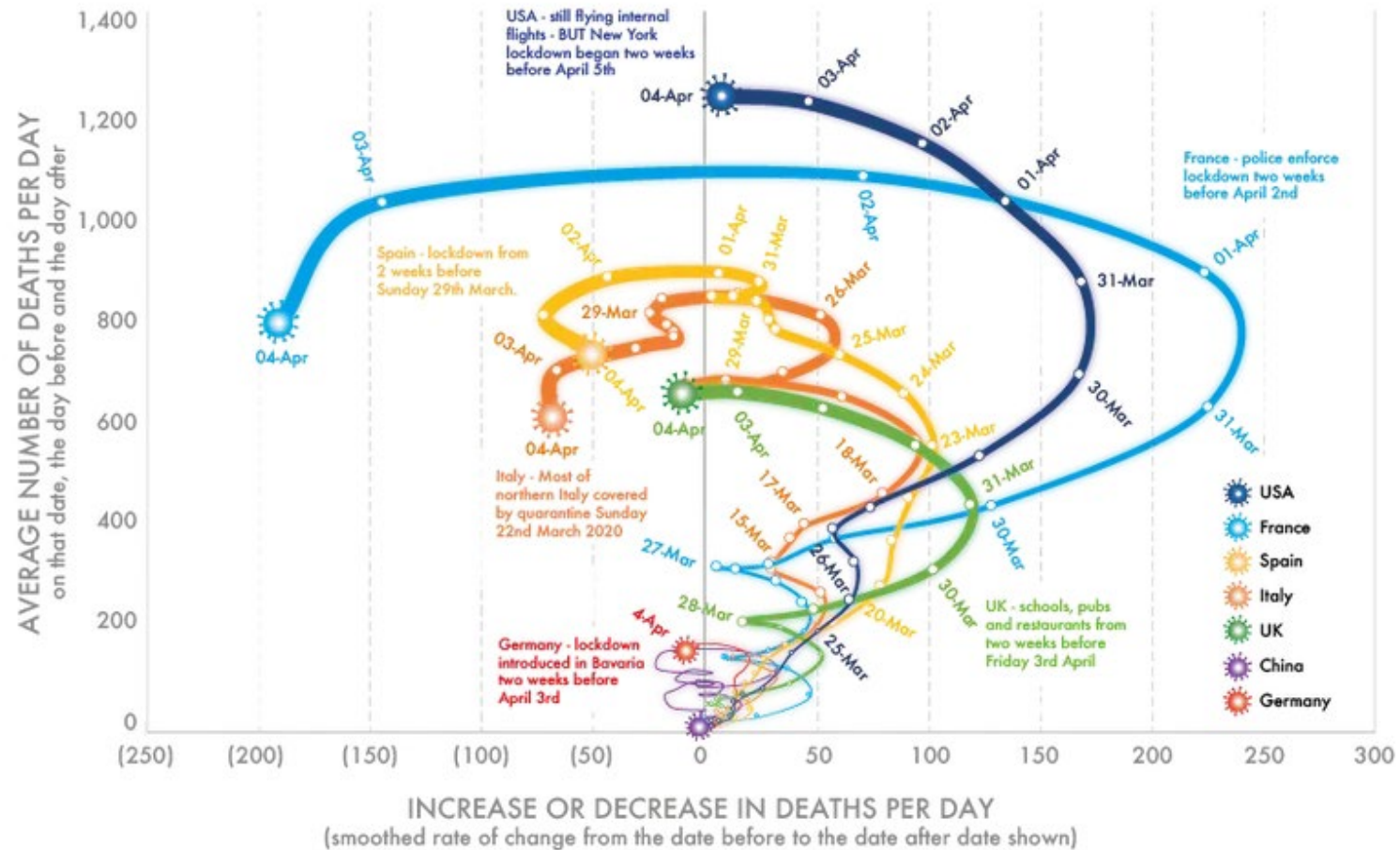
Mortality due to Covid-19 – Graph (1)

Mortality in seven countries attributed to COVID-19 (January 23 to April 2, 2020).



Mortality due to Covid-19 – Graph (2)

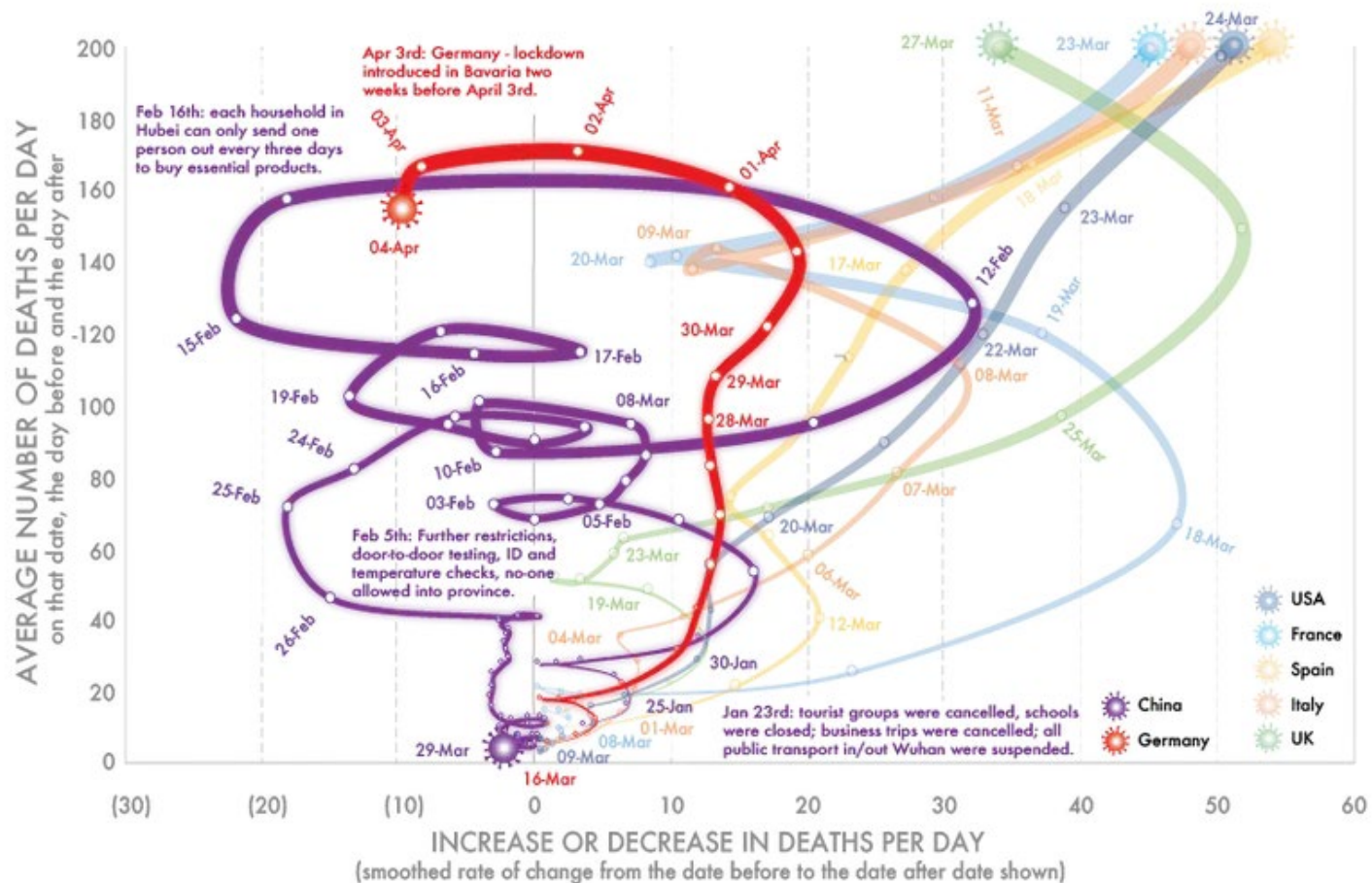
Mortality in seven countries attributed to COVID-19 (January 23 to April 4, 2020).



Danny Doring/Kirsten McClure, Author provided

Mortality due to Covid-19 – Graph (3)

Mortality in China and Germany attributed to COVID-19 (January 23 to April 4, 2020)



Danny Dorling/Kirsten McClure, Author provided

The Bad and the Ugly Graphics from the Statistical Literature

Death by Age Group – Legend Mix Up

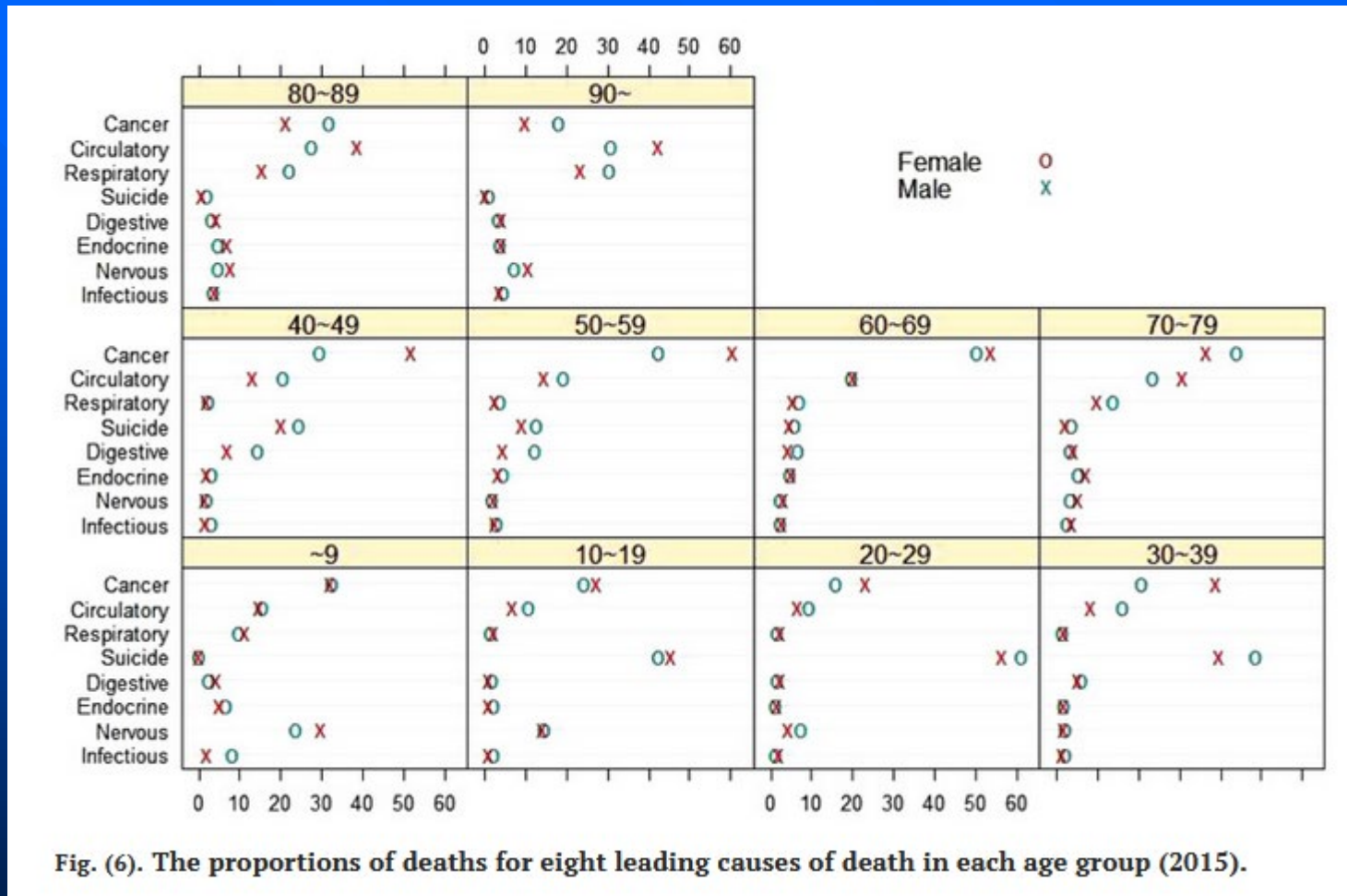



Fig. (6). The proportions of deaths for eight leading causes of death in each age group (2015).

Death by Age Group – Author Identification - Oooops

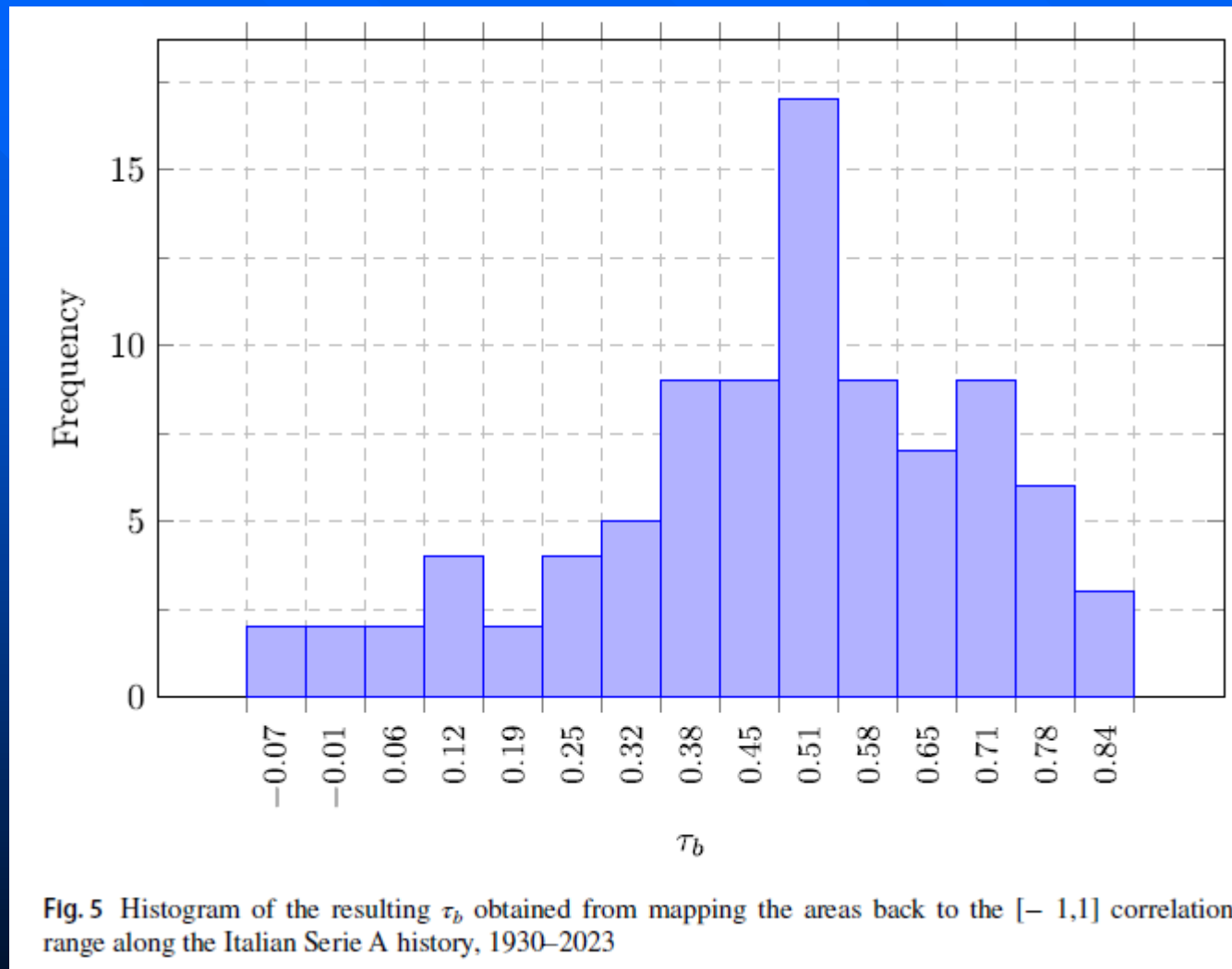
Exploring Mortality Rates for Major Causes of Death in Korea

Hyo Jung Oh¹, Donng Min Yang¹, Chong Hyuck Kim¹, Jae Gyu Jeon², Nam Hyung Jung², Chan Young Kim³,
Jürgen Symanzik⁴, Hyo Won Oh⁵, Akugizibwe Edwin⁶, [...] Jeong Yong Ahn⁶, *  +2 authors [Authors Info & Affiliations](#)

THE OPEN PUBLIC HEALTH JOURNAL • 28 Jan 2019 • RESEARCH ARTICLE • DOI: 10.2174/1874944501912010016



Computational Statistics (COST) (2025)



Computational Statistics & Data Analysis (CSDA) (2025)

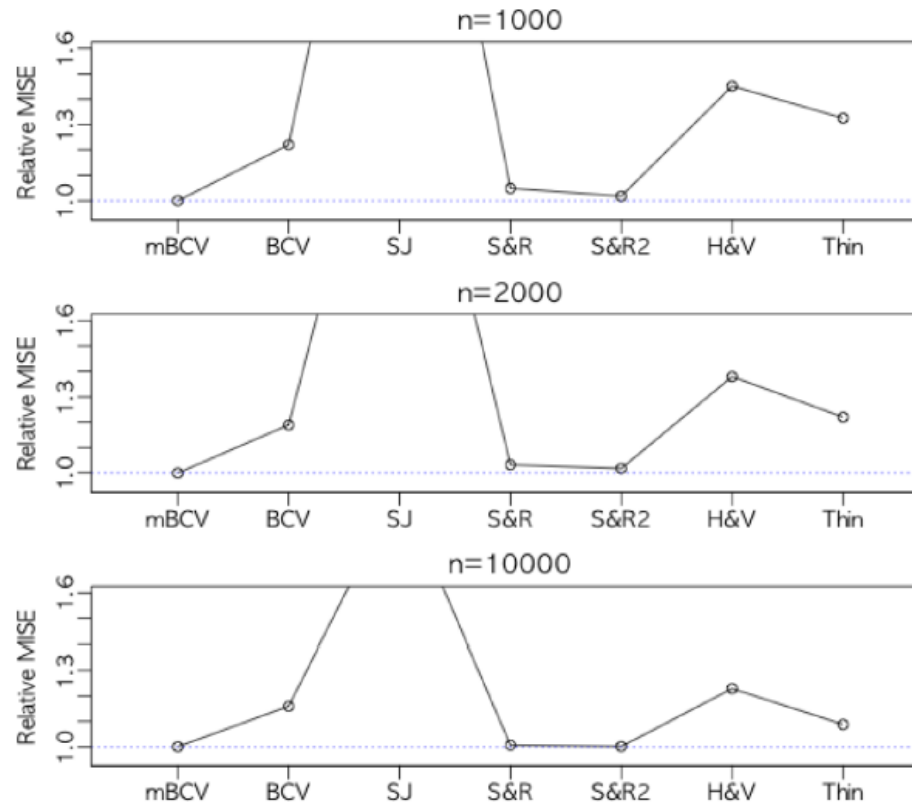


Fig. 2. The relative MISE of kernel density estimates with Metropolis-Hastings samples of different sample sizes ($n = 1000, 2000,$ and 10000) whose target distribution is a normal distribution with a mean of 30 and standard deviation of 20. The relative MISE is the ratio of the empirical MISE of a method to that of mBCV. mBCV denotes the suggested method; BCV, the original BCV; S&R, the original method of Sköld and Roberts (2003); S&R2, modified Sköld and Roberts (2003) with the objective function in Eq. (4.2); H&V, the leave-out CV of Hart and Vieu (1990); and Thin, the thinned sample method. Note that the relative MISEs of the SJ plug-in method are not displayed here, but in Table A.2, whose values (3.30, 2.86, and 2.02, respectively for three settings) are substantially larger than those of other methods. The horizontal line drawn at 1 indicates the relative MISE of mBCV to itself.

Journal of Computational and Graphical Statistics (JCGS) (2025)

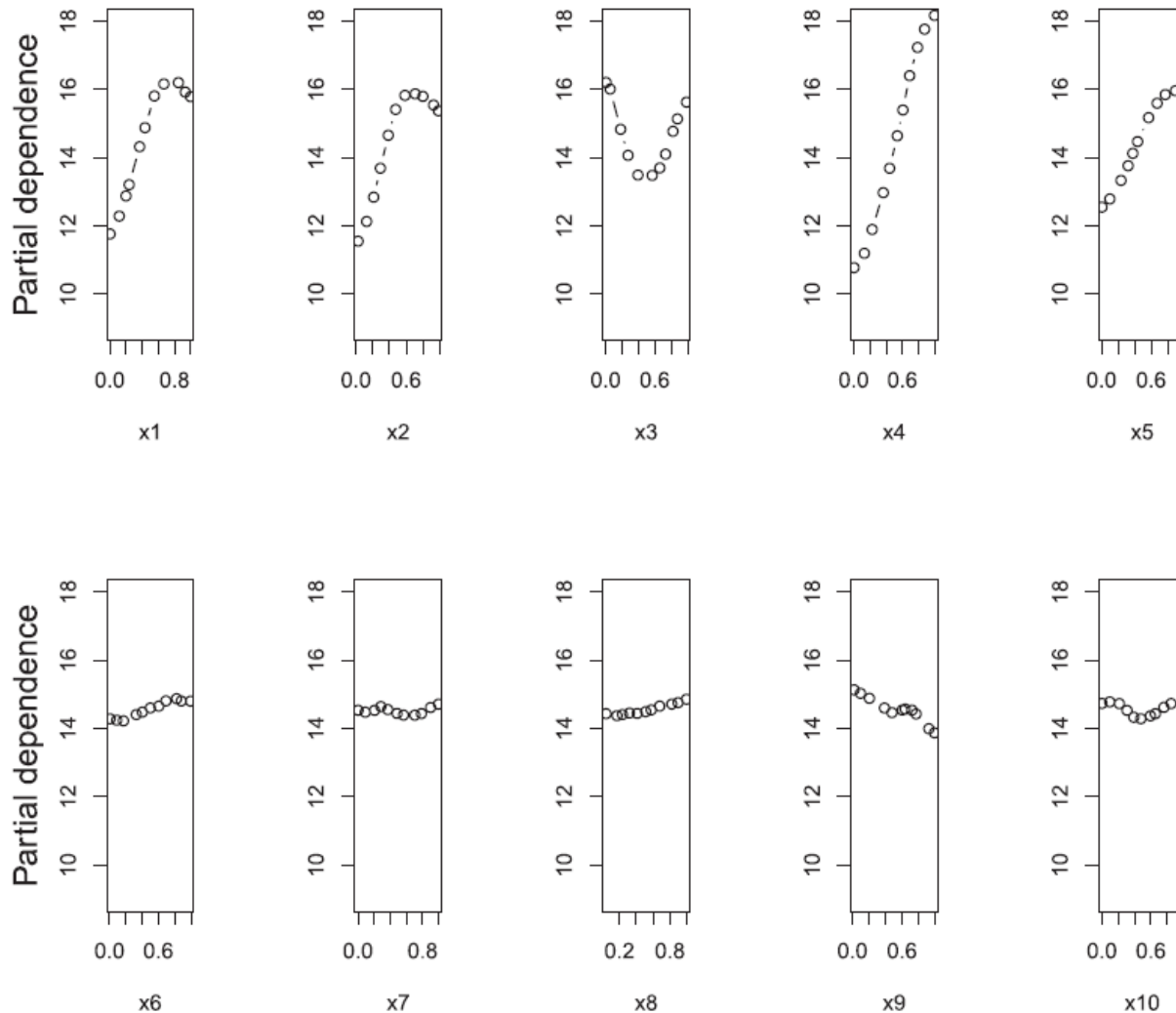


Figure 5. Partial dependence plots for the 10 predictors in Friedman function.

Part II: Graphics Created with the Help of Artificial Intelligence (AI)

AI Tools (Large Language Models, LLMs) in R

- ellmer: Chat with Large Language Models, <https://cran.r-project.org/package=ellmer>
- air: AI Assistant to Write and Understand R Code, <https://cran.r-project.org/package=air>
- chatLLM: A Flexible Interface for 'LLM' API Interactions, <https://cran.r-project.org/package=chatLLM>
- tidyllm: Tidy Integration of Large Language Models, <https://cran.r-project.org/package=tidyllm>
- and more ...

Copilot in RStudio

- GitHub Copilot is an “AI pair programmer that offers autocomplete-style suggestions as you code”. GitHub Copilot is available as an opt-in integration with RStudio. To provide feedback or report bugs, please open a GitHub Issue on the RStudio repository, see <https://docs.posit.co/ide/user/ide/guide/tools/copilot.html>

LLMs in the Classroom

- Some disadvantages of AI tools in R for classroom use:
 - Relatively complex setup with environmental variables, Application Program Interface (API) keys, etc.
 - Although minimal cost to use per “token”, students may not be willing to take the risk of paying on their own.

Freely Accessible LLMs

- Alternatively, some LLMs are freely accessible through web browsers, operating systems, or free accounts.
- While the number of daily prompts typically is limited, this number is usually sufficient for classroom use.
- Students (and I) worked with:
 - ChatGPT / GPT-5: 12 – <https://chatgpt.com/>
 - Google Gemini: 2 – <https://gemini.google.com/app>
 - Claude AI: 1 – <https://claude.ai> (requires login)
 - Yandex Alice: 1 – <https://alice.yandex.ru/>
 - Co-Pilot: 1 (myself) – <https://copilot.microsoft.com>

AI Graphics Tasks in the Classroom

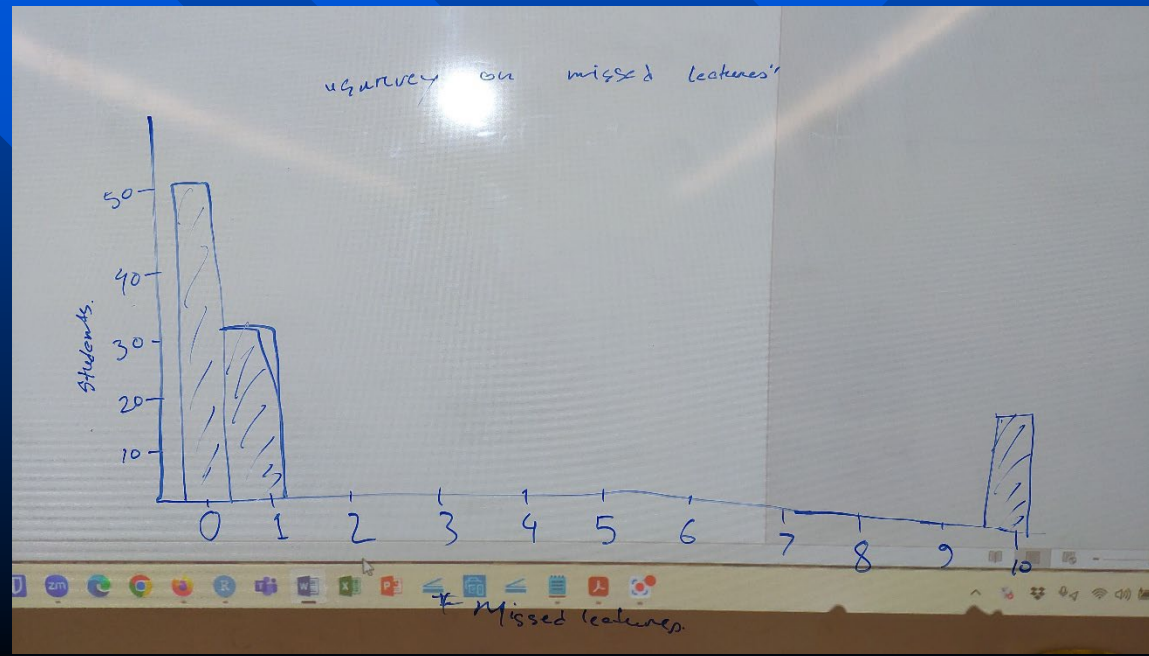
Task 1.1: Sketch a Histogram

Consider the following thought experiment - then look at R and AI results later:

Here is a made-up Stat 2000-like data set: In a 1-credit course with 100 students that has 10 lectures, students were surveyed at the end of the course how many lectures they had missed. Their answers were as follows:

50 x 0 Lectures, 30 x 1 Lecture, 20 x 10 Lectures

How would YOU manually draw a histogram for these observations?



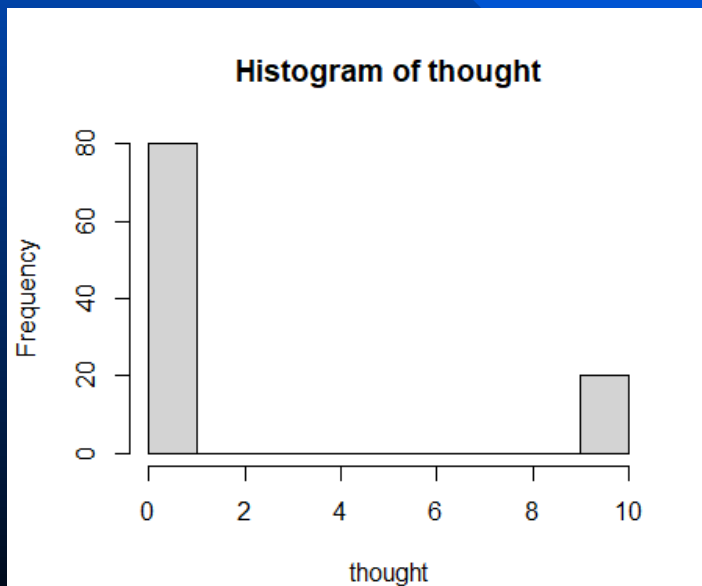
Task 1.2: baseR Histogram

Use baseR to create a histogram that resembles our hand-drawn histogram. As always, you should start with the most basic histogram, then make necessary changes to it. The initial version from baseR most likely will not match the version we have presented in class.

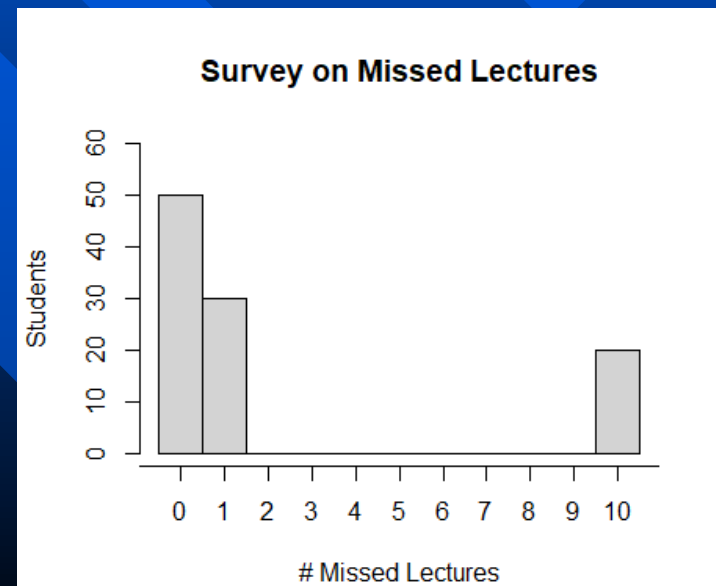
For convenience, here is an R vector to start with:

```
thought <- c(rep(0, 50), rep(1, 30), rep(10, 20))
```

Initial Version



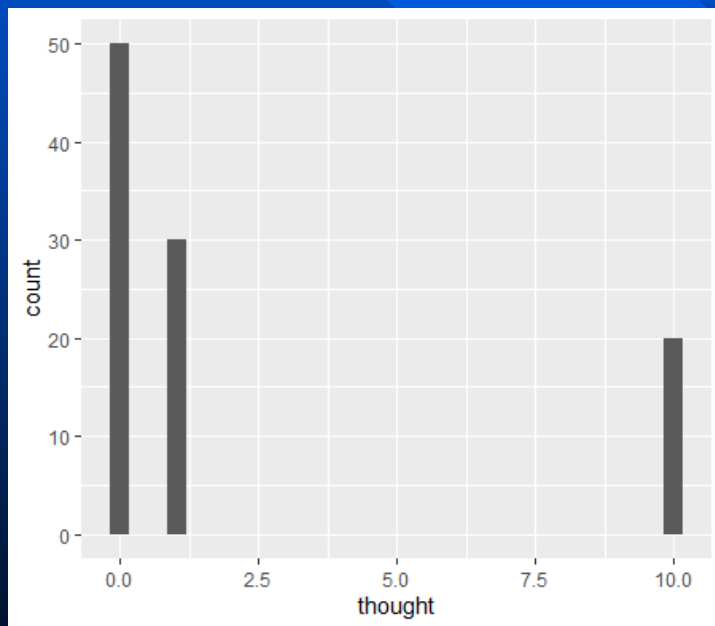
Final Version



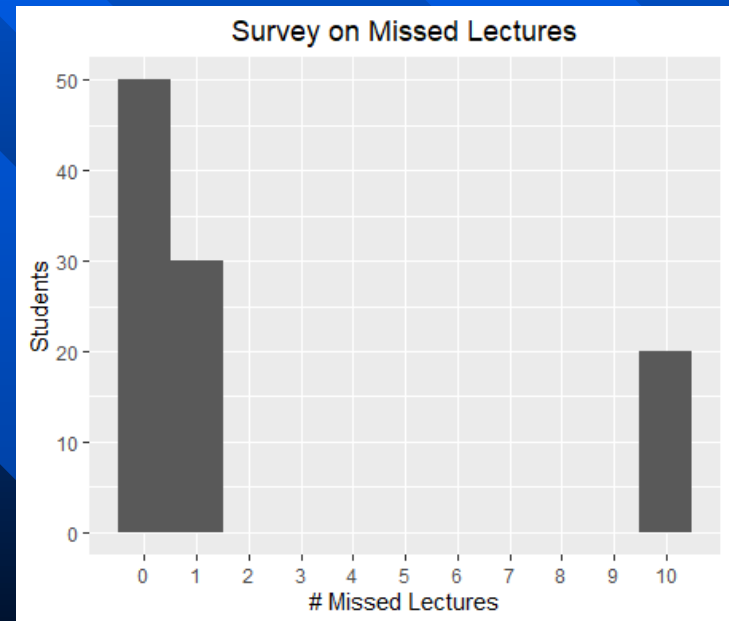
Task 1.3: ggplot2 Histogram

Use ggplot2 to create a histogram that resembles our hand-drawn histogram. As always, you should start with the most basic histogram, then make necessary changes to it. The initial version from ggplot2 most likely will not match the version we have presented in class.

Initial Version



Final Version

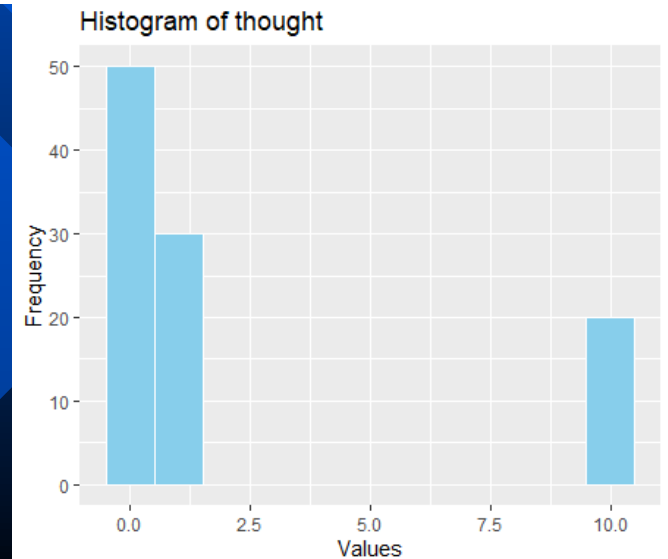
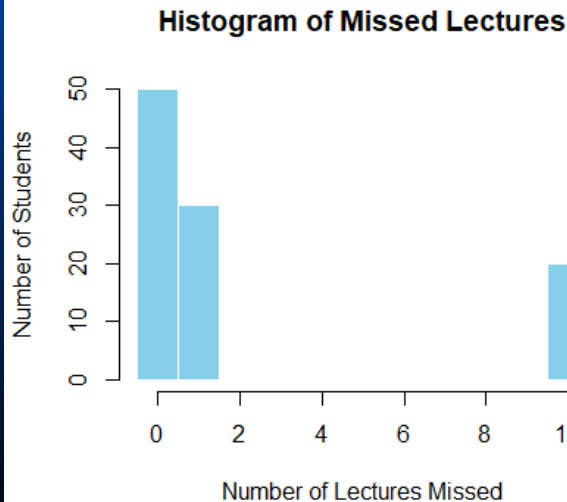
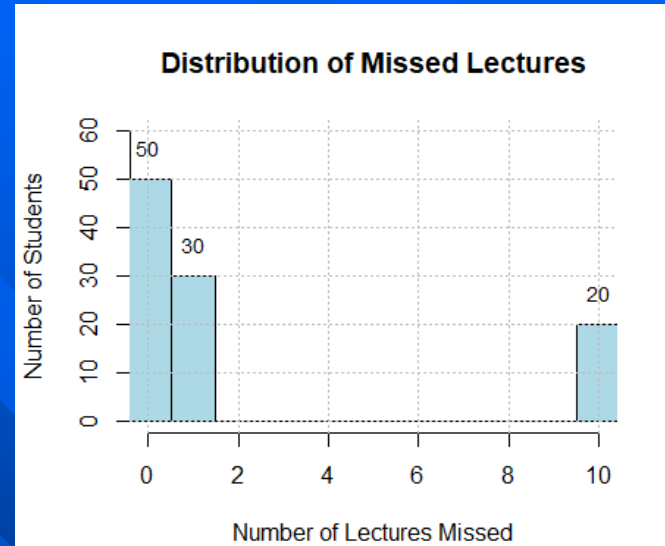
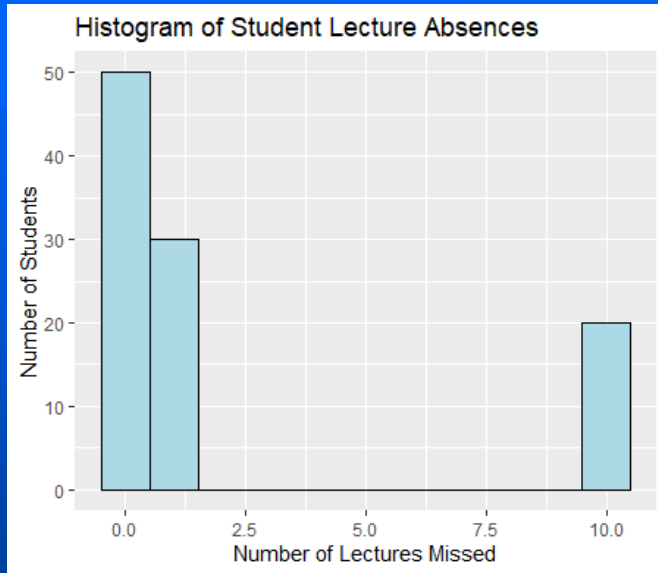


Task 1.4: LLM Histograms

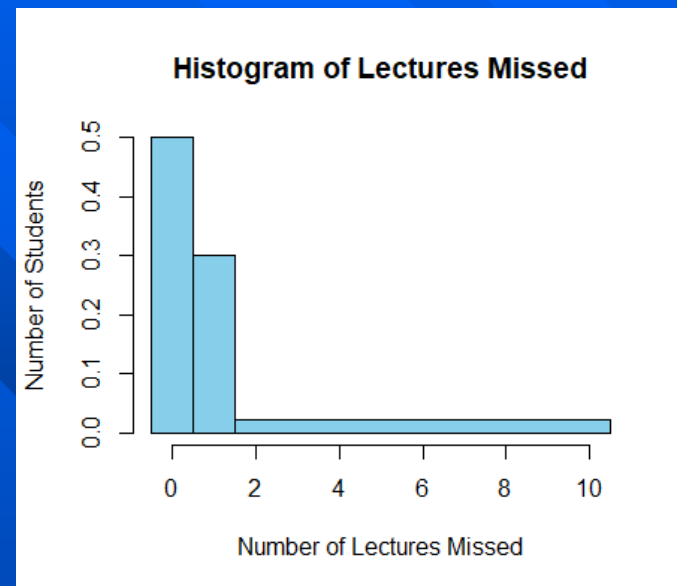
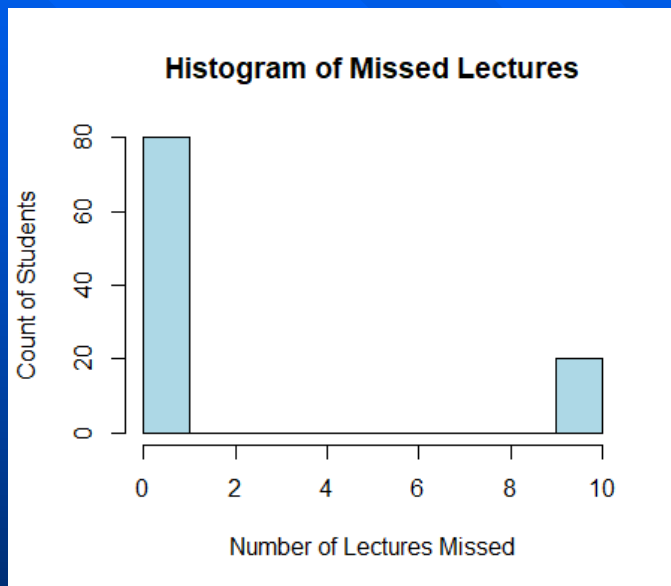
Use an AI tool of your choice to create R code that shows a histogram (either baseR or ggplot2) for the data from Part 1. You have to find a suitable prompt for the AI tool and pass the data via the R vector from Part 2 or via the background story from Part 1 to it.

Indicate which AI tool you used and which prompt you used for the graph construction. Also upload the initial version of the AI-based R code and the resulting graph to Canvas.

Task 1.4: LLM Histograms – OK Ones

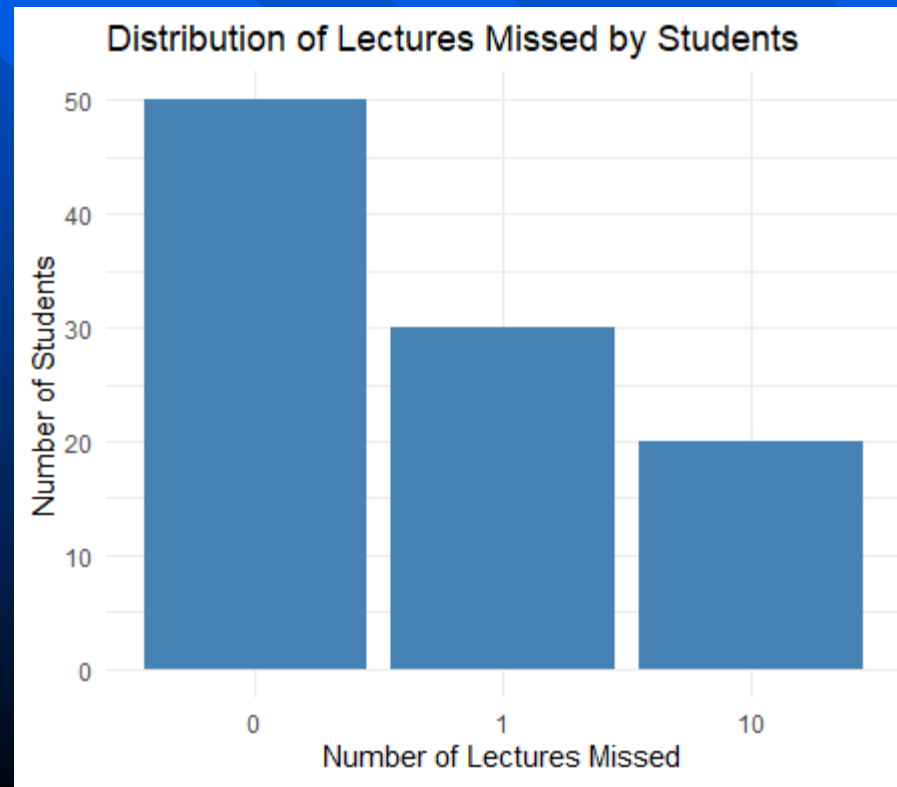


Task 1.4: LLM Histograms – Bad Ones



Task 1.4: LLM Histograms – Terrible One

- Prompt: make a graph in ggplot that uses this data: here is a made-up Stat 2000-like data set: In a 1-credit course with 100 students that has 10 lectures, students were surveyed at the end of the course how many lectures they had missed. Their answers were as follows: 50 x 0 Lectures, 30 x 1 Lecture, 20 x 10 Lectures



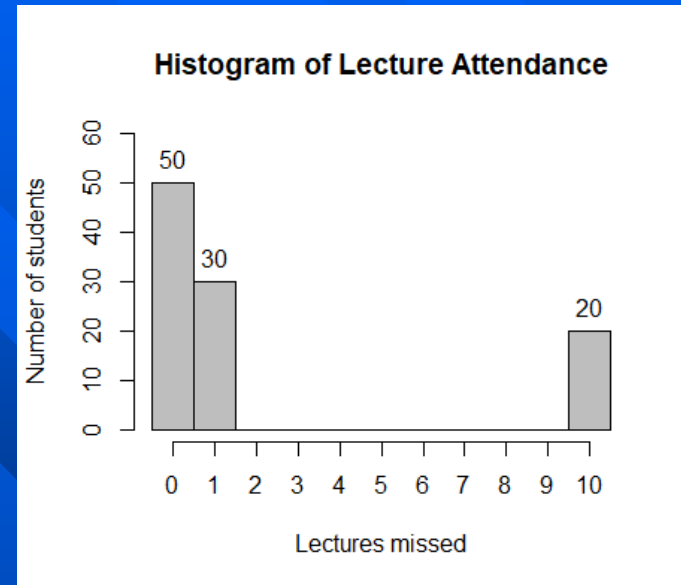
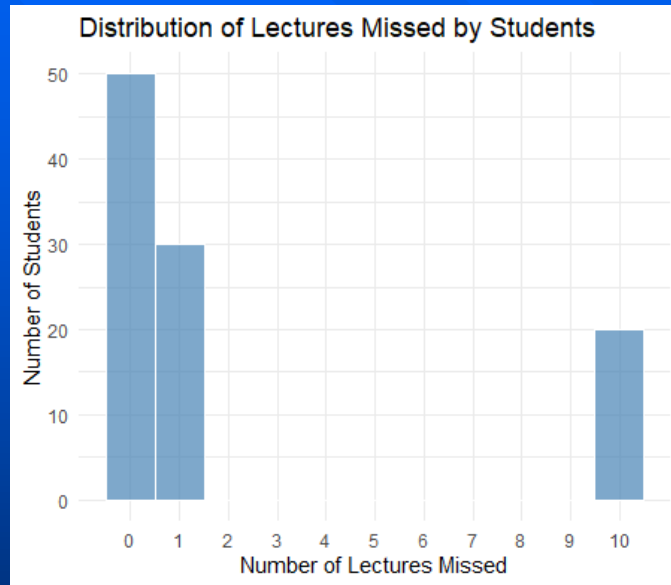
Task 1.5: Refined LLM Histograms

Continue with the initial AI-based R code from Part 4, ideally in the same session. Give the AI tool further instructions how to improve the original R code it created. You are allowed to make up to 3 modification steps, but each step can contain more than just 1 change, e.g., "add a label to the x-axis and extend the y-axis to the maximum" would count as 1 step.

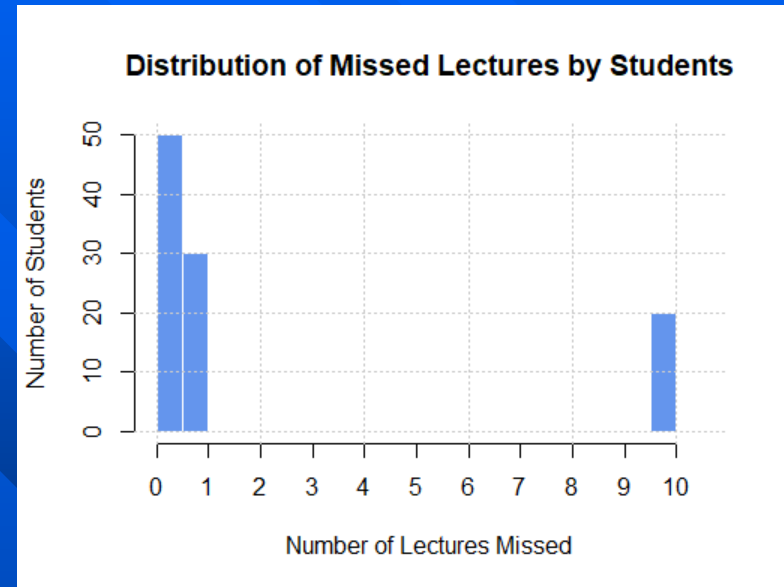
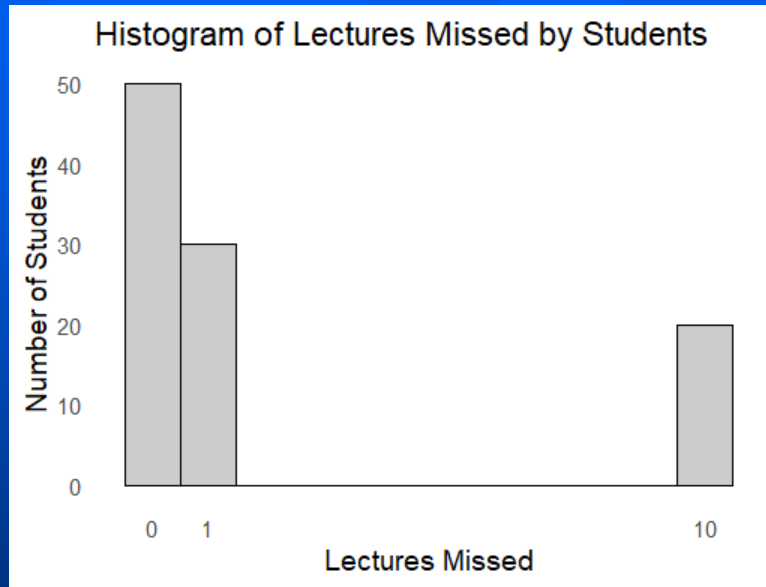
Could you get the AI tool to provide you with R code that resembles our hand-drawn version of the histogram from Part 1? If no, indicate which changes would still be needed.

Also upload the final version of the AI-based R code and the resulting graph to Canvas.

Task 1.5: LLM Histograms – Improved



Task 1.5: LLM Histograms – Still Bad



Task 2: AI Created 3D Pie Charts

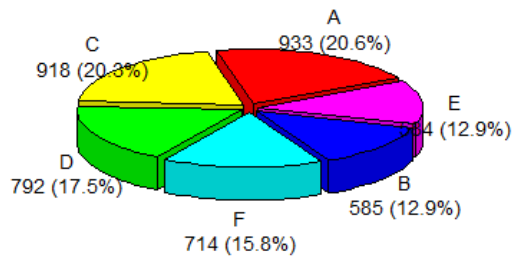
You should never ever (I really mean: **NEVER EVER EVER**) do this in real life - but let's see whether your AI tool is capable of producing R code for a 3D pie chart for you.

Try this with the 6 departments from the UCB admissions data set. The AI tool probably will create such R code (there exists a `pie3D` function in the `plotrix` R package), but I hope it will at least give you a warning about the disadvantage of such a 3D graph. But, let's see...

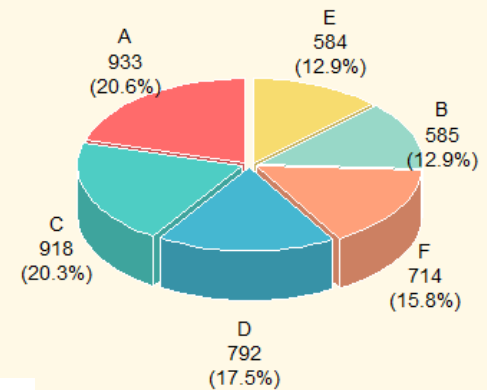
- Student Answers:
 - AI Tool:
 - Initial Prompt:
 - AI Comment (if any):
 - AI-Produced R Code:
 - Resulting Graph:
 - Your Final Comment:

Task 2: AI Created 3D Pie Charts – Bad Ones

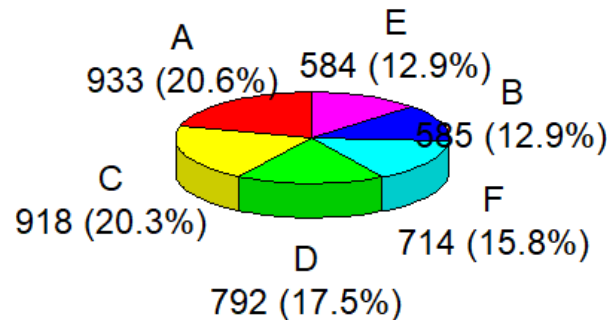
UCB Admissions: Applicants by Department



UCB Admissions by Department



Total Applicants by Department - UCB Admissions



Total Applications: 4,526

Task 2: AI Created 3D Pie Charts – Summary of AI Comments (via AI)

Discussion Summary



This summary is generated by AI and reflects the latest contributions to the discussion. Please note that the output may not always be accurate. Summaries are only visible to instructors.

Topics to focus on (optional)

Which of the AI tools gave a warning about the use of 3D pie charts? And what did they say specifically?

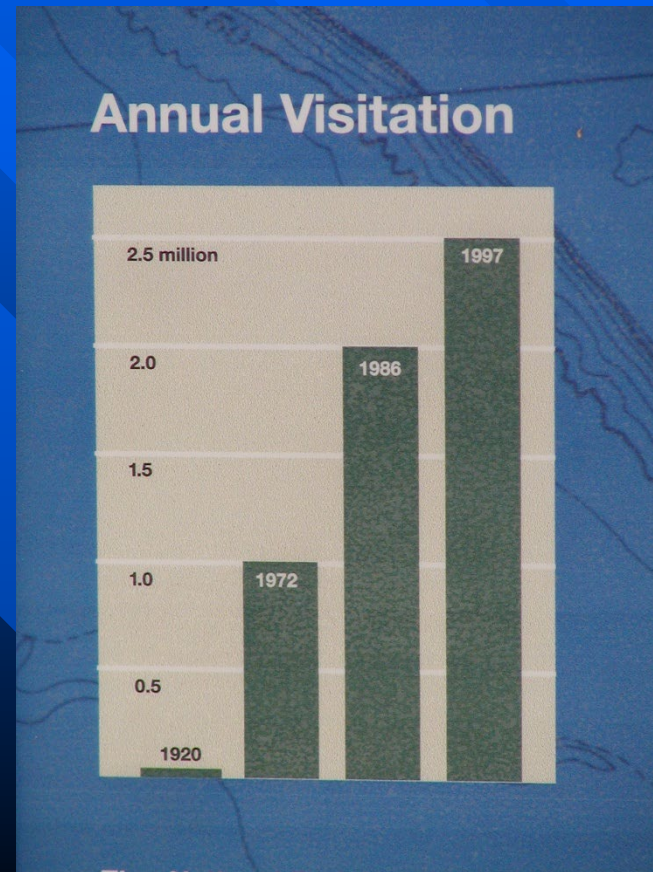
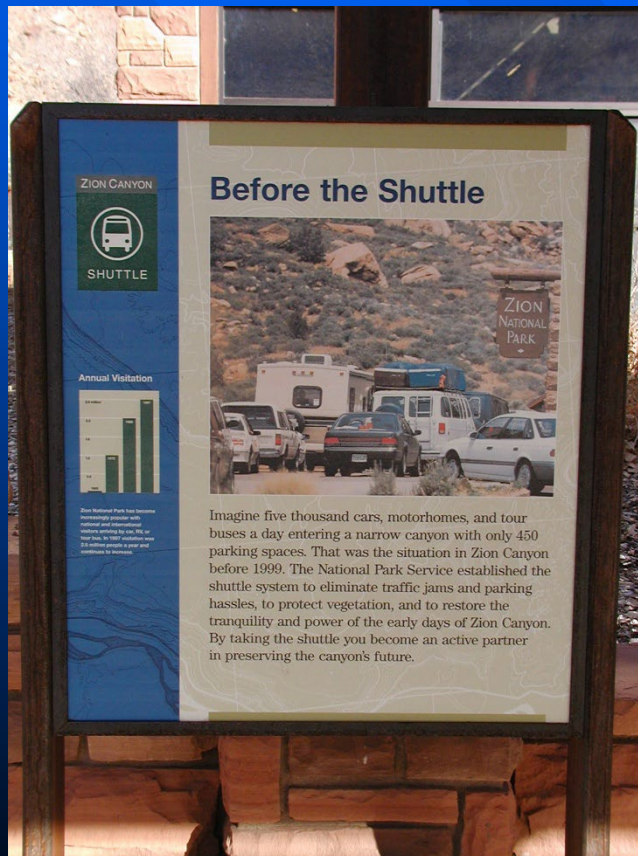
✦ Summarize

Generated Summary (0 / 25)

According to the discussion, only one AI tool, Google Gemini, provided a warning about the use of 3D pie charts. The warning stated that "3D pie charts are generally discouraged in data visualization as the perspective distortion can misrepresent the size of the segments."

Task 3: AI Improvement of a Bad Graphic

Personal Photo: Sign in Zion National Park, UT, Shuttle Parking Lot, December 28, 2002



Task 3: AI Improvement of a Bad Graphic – Step 1 (via Copilot)

Provide R code for a graph for these data vectors that represent the number of visitors (in million) to Zion National Park over time: year <- c(1920, 1972, 1986, 1997) and visits <- c(0.05, 1.0, 2.0, 2.5)

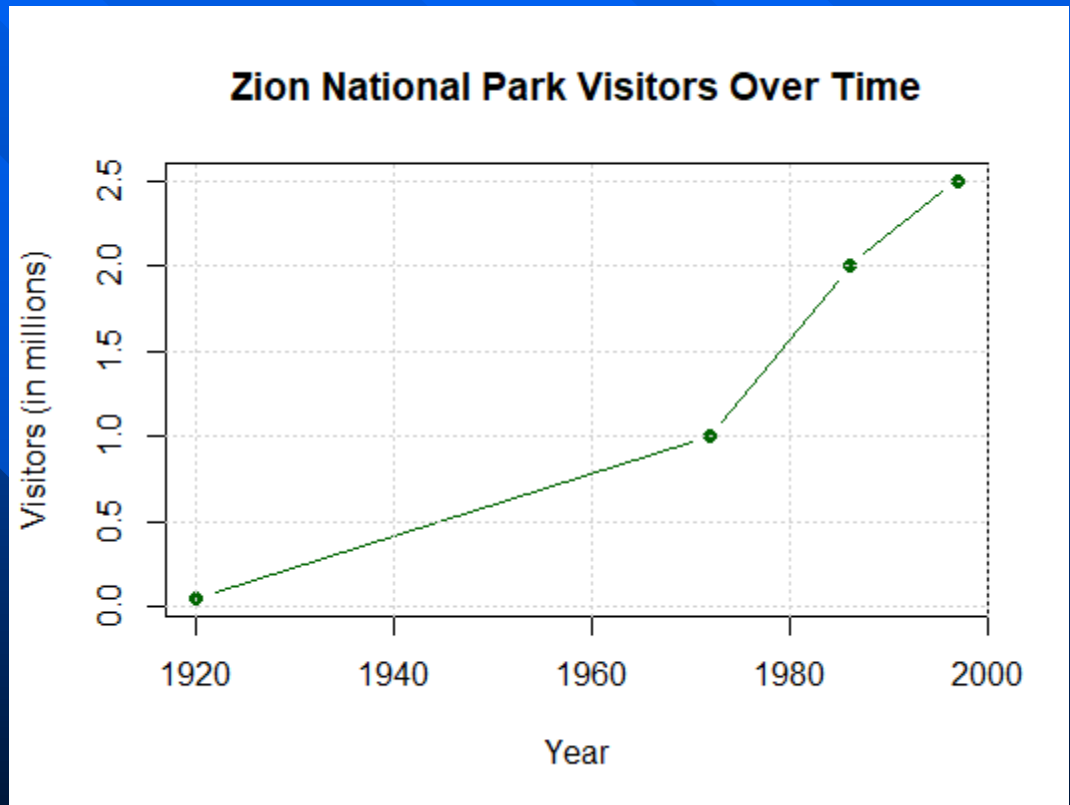
Here's a simple and clear R code snippet to visualize the number of visitors to Zion National Park over time using base R plotting:

```
R ^ Copy

# Data vectors
year <- c(1920, 1972, 1986, 1997)
visits <- c(0.05, 1.0, 2.0, 2.5)

# Create the plot
plot(year, visits,
      type = "b",           # both point:
      pch = 19,            # solid circle
      col = "darkgreen",   # color of points
      xlab = "Year",
      ylab = "Visitors (in millions)",
      main = "Zion National Park Visitors Over Time")

# Optional: add grid for readability
grid()
```



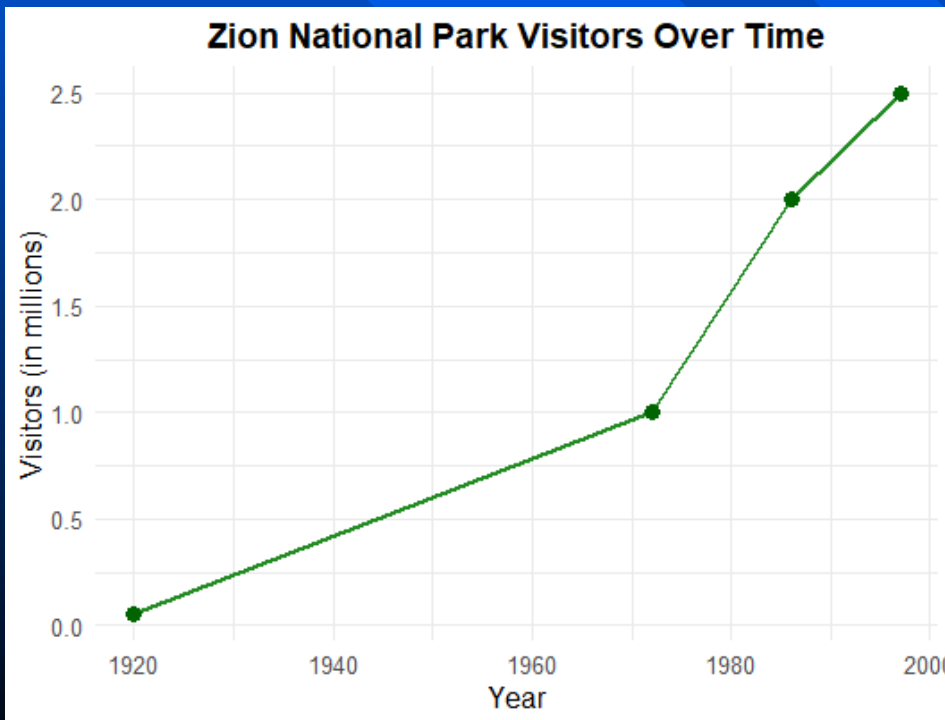
Task 3: AI Improvement of a Bad Graphic –

Step 2

If you'd like a version using `ggplot2` for more customization and polish, I can provide that too. Would you like to explore that?

Yes, also provide a ggplot2 version.

Here's a polished version using `ggplot2`, which offers more control over aesthetics and styling:



```
R ^ Copy

# Load ggplot2
library(ggplot2)

# Create data frame
zion_data <- data.frame(
  year = c(1920, 1972, 1986, 1997),
  visits = c(0.05, 1.0, 2.0, 2.5)
)

# Create the ggplot
ggplot(zion_data, aes(x = year, y = visits))
  geom_line(color = "forestgreen", size = 1)
  geom_point(color = "darkgreen", size = 3)
  labs(
    title = "Zion National Park Visitors Over Time",
    x = "Year",
    y = "Visitors (in millions)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold",
                               size = 12),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10)
  )
```

This version uses a clean minimal theme and centers the title for better readability. If you'd like to add annotations, a log scale, or seasonal styling (like a vintage look for historical data), I'd be happy to help you iterate further.

Task 3: AI Improvement of a Bad Graphic –

Step 3 Challenge

Can you find additional data for more recent years on a web site (if so, then indicate the exact source) and extend the graph with this new data.

Yes! I found detailed annual visitor data for Zion National Park through 2024 from official and reputable sources. Here's how to extend your graph with it.

Sources of Recent Visitor Data

The most complete and up-to-date annual visitation numbers come from:

- **National Parked:** [Zion Visitation Statistics 1919–2024](https://www.nationalparked.com/zion/visitation-statistics)
National Parked
- **Smithsonian Magazine:** Zion had **4,946,592** visitors in **2024**, making it the second most visited national park Smithsonian Maga...

URLs provided:

<https://www.nationalparked.com/zion/visitation-statistics>

<https://www.smithsonianmag.com/smart-news/these-were-the-most-and-least-visited-national-parks-in-2024-180986251/>

Task 3: AI Improvement of a Bad Graphic –

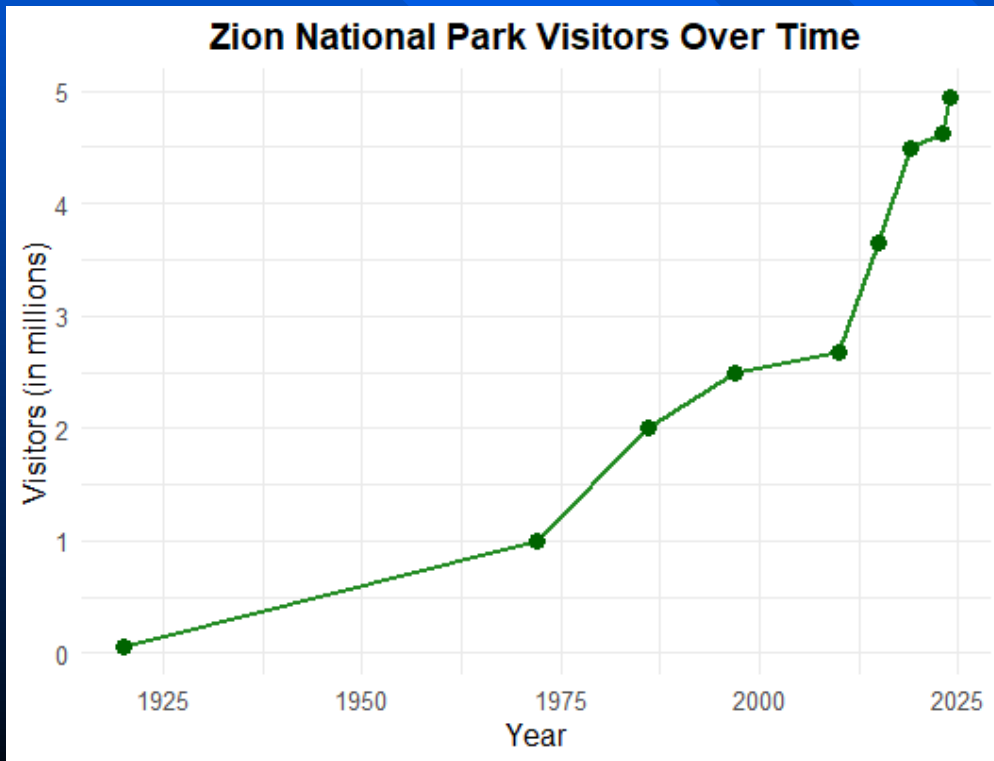
Step 3 Result

Updated Data Vectors

Here's the extended dataset including recent years:

R ^ Copy

```
visitors  
(2, 1986, 1997, 2010, 2015, 2019, 2023, 2024)  
..0, 2.0, 2.5, 2.67, 3.65, 4.49, 4.62, 4.95)
```



Updated ggplot2 Graph

R ^ Copy

```
library(ggplot2)  
  
# Create updated data frame  
zion_data <- data.frame(  
  year = c(1920, 1972, 1986, 1997, 2010, 2015, 2019, 2023, 2024),  
  visits = c(0.05, 1.0, 2.0, 2.5, 2.67, 3.65, 4.49, 4.62, 4.95)  
)  
  
# Create the updated ggplot  
ggplot(zion_data, aes(x = year, y = visits))  
  geom_line(color = "forestgreen", size = 1)  
  geom_point(color = "darkgreen", size = 3)  
  labs(  
    title = "Zion National Park Visitors Over Time",  
    x = "Year",  
    y = "Visitors (in millions)"  
  ) +  
  theme_minimal() +  
  theme(  
    plot.title = element_text(face = "bold", size = 12),  
    axis.title = element_text(size = 12),  
    axis.text = element_text(size = 10)  
  )
```

Task 3: AI Improvement of a Bad Graphic – Step 3 Assessment

```
Updated Data Vectors
Here's the extended dataset including recent years:

R ^ Copy
:tors
'2, 1986, 1997, 2010, 2015, 2019, 2023, 2024)
..0, 2.0, 2.5, 2.67, 3.65, 4.49, 4.62, 4.95)
```

https://www.nationalparked.com/zion/visitation-statistics

Zion National Park

Home » Zion » Visitation Statistics

Zion Visitation By Year

Below is a table showing the annual visitation for every year that the statistic has been kept. In the case of Zion, this runs from 1919 to present. These are the official numbers reported by the National Park Service. Counting methods sometimes change over time, which may account for occasional irregularities.

Year	Visitors
2021	5,039,835
2020	3,591,254
2019	4,488,268
2018	4,320,033
2017	4,504,812
2016	4,295,127
2015	3,648,846

https://www.smithsonianmag.com/smart-news/these-were-the-most-and-least-visited-national-parks-in-2024-1

Smithsonian magazine

Search Shop Newsletters

Subscribe to Smithsonian magazine and get a FREE tote.

The Most Visited National Parks

The United States has 63 national parks. These were the ten most visited in 2024:

1. Great Smoky Mountains National Park (12,191,834 visits)
2. Zion National Park (4,946,592 visits)
3. Grand Canyon National Park (4,919,163 visits)
4. Yellowstone National Park (4,744,353 visits)
5. Rocky Mountain National Park (4,154,349 visits)

Positive: Both URL existed! And most of the numbers existed.

Negative: Where did the value for 2023 come from? Not from these URLs – A hallucination!
And why was 2019, but not 2020 nor 2021 included in the R code?

Overall Assessment of LLM Graphics

Positive Takeaways of LLM Graphics

- ❑ LLMs provide relatively high level R code with comments, good style, and good formatting.
- ❑ Most unexperienced graphics developers or R novices probably would need hours to produce similar graphics in R.
- ❑ Often good titles, axis labels, and ticmark labels.
- ❑ Often infer missing information such as units, dimensions, and so on.

Negative Takeaways of LLM Graphics

- ❑ LLMs often change defaults, in particular colors to “lightblue”, “skyblue”, or “steelblue”, etc. – Why?
- ❑ Often reduce ggplot2 theme to “theme_minimal()” – Why?
- ❑ Small changes in the prompt can result in big changes in the resulting R code / graphic.
- ❑ As always, be careful with LLM hallucinations !!!

Overall Conclusions

Outlook on Future Graphics

- ❑ My hypothesis: We will see many AI created graphics in the future – on the web, in the news, and in scientific publications.
- ❑ Same as in the past, some of these will be good – many will be bad (at least in the near future).
- ❑ AI-created graphics seem to require as much fine-tuning as initially created graphics in baseR or ggplot2.
- ❑ Graphic designers must be knowledgeable enough about good graphics to initiate the necessary fine-tuning steps.
- ❑ Going back to Gordon & Finch (2015), LLMs need to be trained by good graphics only – and not by bad ones!

Appendix

Sources of Bad Graphics on the Web

- <http://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6>
- https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/
- <https://www.buzzfeednews.com/article/katienotopoulos/graphs-that-lied-to-us>
- <https://www.reddit.com/r/dataisugly>
- <https://getdolphins.com/blog/the-worst-graphs-of-2017/>
- <https://getdolphins.com/blog/worst-graphs-of-2018/>
- https://bcuchta.com/wall_of_shame/
- and many more ...

Sources for Constructing Better Graphs

Sources for Constructing Better Graphs: Books & Articles

- Robbins, N. B. (2013). *Creating More Effective Graphs*. Chart House: Ramsey, NJ. (reprinted from Wiley 2005)
- Su, Y-S. (2008). “It’s Easy to Produce Chartjunk using Microsoft Excel 2007, but Hard to Make Good Graphs.” *Computational Statistics and Data Analysis* 52:4594-4601.
- Tufte, E. R. (2001). *The Visual Display of Quantitative Information*, 2nd edition. Graphics Press: Cheshire, CT. (first edition 1983)
- Wainer, H. (2000). *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Psychology Press: London, UK. (reprinted from Copernicus 1997)

Sources for Constructing Better Graphs:

Blogs

- ❑ Camoes, J. “ExcelCharts,” <http://www.excelcharts.com/blog/>
- ❑ Peltier, J. “Peltier Tech Blog— Peltier Tech Excel Charts and Programming Blog,” <http://peltiertech.com/WordPress/>
- ❑ Robbins, N. B. “Effective Graphs,” <http://www.forbes.com/sites/naomirobbins/>
- ❑ Robbins, N. B. (February 28, 2012). “Misleading Graphs: Displaying a Change in One Variable Using Area or Volume,” Forbes.com LLC. <http://www.forbes.com/sites/naomirobbins/2012/02/28/misleading-graphs-displaying-a-change-in-one-variable-using-area-or-volume/>
- ❑ Wyatt, S. (March 31, 2014). “Dishonest Fox Charts: Obamacare Enrollment Edition,” Media Matters for America. <http://mediamatters.org/blog/2014/03/31/dishonest-fox-charts-obamacare-enrollment-editi/198679>

Sources for Constructing Better Graphs: For Teachers

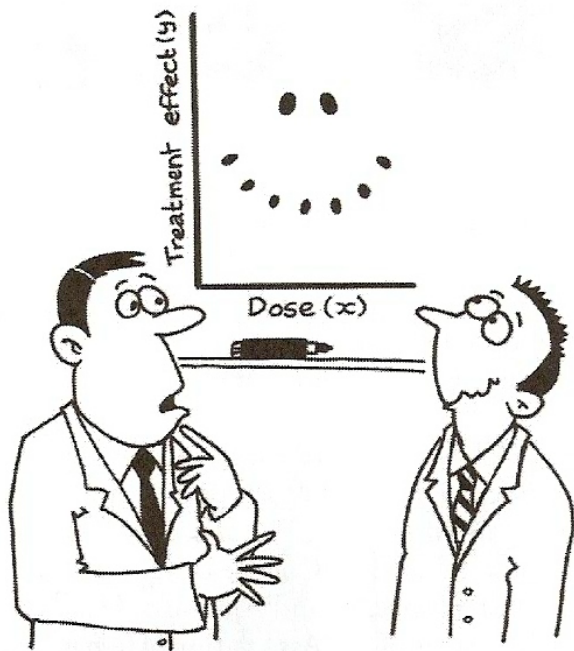
- Heiberger, R. M., Robbins, N. B., Symanzik, J. (2014). “Statistical Graphics Recommendations for the ASA/NCTM Annual Poster Competition and Project Competition”, 2014 JSM Proceedings, American Statistical Association, Alexandria, VA.
- Symanzik, J., Robbins, N. B., Heiberger, R. M. (2014). “Observations from the Winners of the 2013 Statistics Poster Competition — Praise and Future Improvements.” The Statistics Teacher Network, 83:2–5.
<https://ww2.amstat.org/education/stn/pdfs/stn83.pdf>
- Symanzik, J., Robbins, N. B., Heiberger, R. M. (2016): “Observations on the Type and Quality of Graphs Used in the ASA/NCTM Annual Poster Competition during the Years 2013 to 2016”, 2016 JSM Proceedings, American Statistical Association, Alexandria, Virginia.
- <http://www.statlit.org/pdf/2016-Symanzik-Robbins-Heiberger-ASA.pdf>

Acknowledgements

Thanks to my current Fall 2025 students in Statistical Visualization I for their creative way of using LLMs to create good and bad graphics!

Also thanks to many of my former students, my colleagues, and my collaborators who contributed to the collection of good and bad graphs and graph web pages shown in this presentation. I really appreciate obtaining links to bad graphics and bad graphics web pages even when not discussing bad graphics in the classroom or in a committee or task force.

Thank you!



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

A CAUSE-commissioned cartoon that is part of the CAUSEweb collection and available for free noncommercial use by statistics teachers. Cartoon by John Landers ©. Provided by permission.

Questions ???
— or —
send e-mail to:

juergen.symanzik@usu.edu

From: Amstat News, January 2009, p. 25