OPINION

Revisiting John Snow's Cholera Map: A Data Visualisation Case Study for Statistical Education

Niamh Mimnagh, Department of Mathematics and Statistics, Maynooth University, Kildare, Ireland.
Email: mailto:niamhmimnagh@gmail.com

Abstract

Data visualisation is a fundamental tool in statistical analysis, enabling the identification of patterns and relationships that might otherwise remain hidden in raw data. One of the most famous historical examples is John Snow's 1854 cholera map, which demonstrated the spatial clustering of cholera cases around a contaminated water pump in London. This study explores how Snow's visualisation can be effectively incorporated into statistics education as an interactive case study. Revisiting Snow’s cholera map in 2025 provides a powerful bridge between foundational epidemiological reasoning and modern statistical practice. It offers students an intuitive, historically grounded pathway into spatial thinking and data visualisation using tools like R. Using R, we outline the steps involved in reproducing Snow's cholera map, demonstrating geospatial data manipulation, visualisation techniques, and spatial analysis. We discuss the pedagogical benefits of historical case studies in statistics courses, emphasising their role in fostering curiosity, critical thinking, and technical proficiency. Additionally, we explore how these methods can be extended beyond epidemiology to applications in public health, urban analytics and environmental science. By integrating historical datasets with modern computational tools, educators can create engaging, hands-on learning experiences that reinforce core statistical principles while illustrating the real-world impact of data analysis.

Keywords: Visualisation, Exploratory Data Analysis, John Snow, Cholera, Spatial Data.

1. Introduction

Data visualisation plays an important role in statistical analysis and decision-making, allowing patterns and relationships to emerge that might otherwise remain hidden in raw data. One of the most famous early examples of discovery driven by visualisation comes from John Snow's 1854 cholera map (Snow, 1856), which is presented in Figure 1, and is widely recognised as a pioneering work in epidemiology and spatial statistics (Tufte and Robins, 1997). By plotting cholera cases on a street map of Soho, London, Snow succeeded in demonstrating a spatial pattern that linked cases to a contaminated water pump, providing critical evidence against the dominant miasma theory of disease transmission (Brody et al., 2000). His analysis contributed to the foundation of modern epidemiology and remains a compelling example of how statistical thinking, when combined with effective visualisation, can drive scientific progress (Friendly and Dennis, 2001).

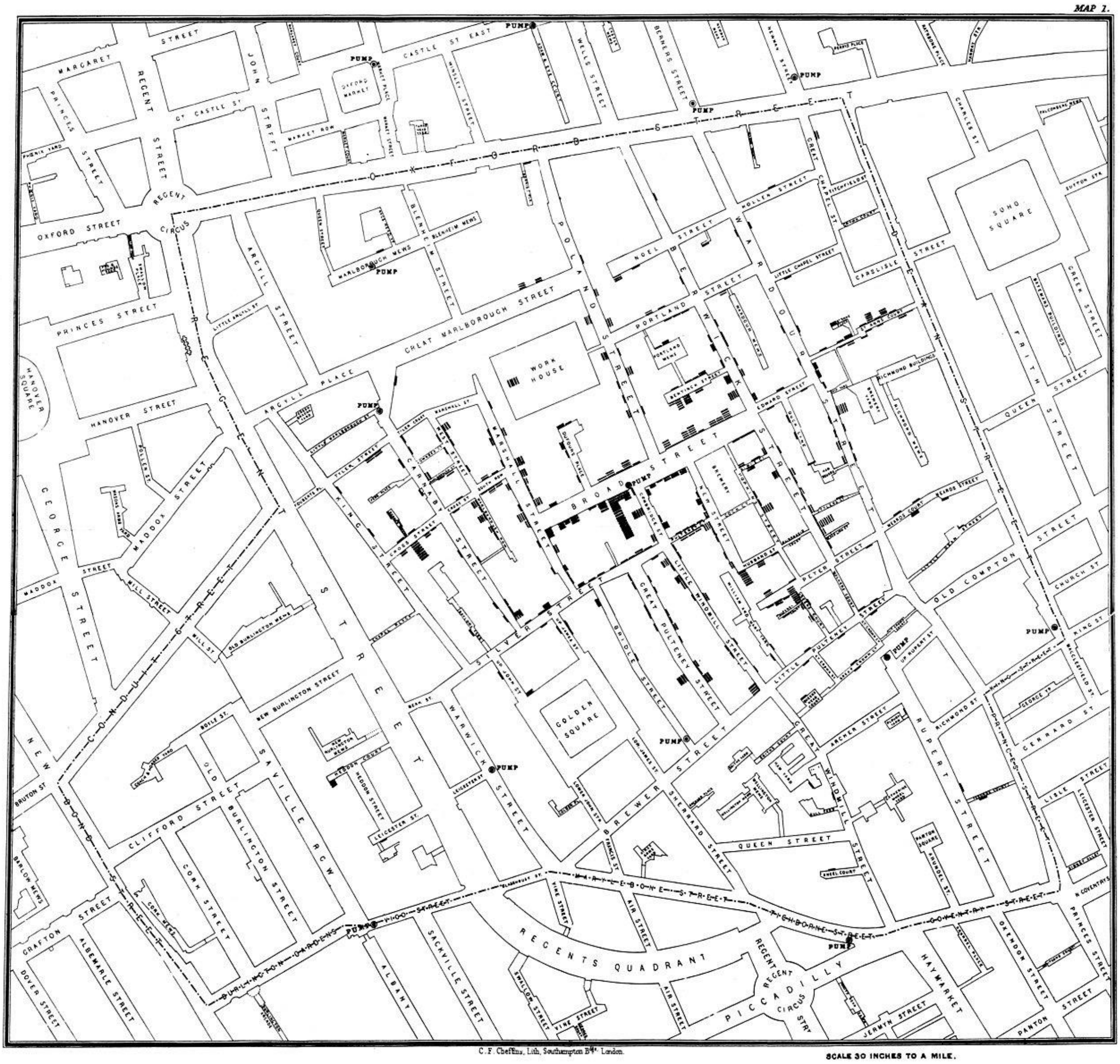

Figure 1: *John Snow's original 1854 cholera map of Soho, London.* This map visualises the spatial distribution of cholera deaths during the outbreak, revealing a clear clustering of cases around the Broad Street pump (centre-left of Broad Street). Water pumps are depicted as solid black circles, while cholera cases are shown as stacked black bars along the streets. The height of each bar corresponds to the number of deaths recorded at that location, enabling a street-level view of the outbreak's intensity.

Beyond its historical significance, John Snow's work serves as an excellent teaching tool for students learning about data visualisation, spatial analysis and statistical reasoning. The process of mapping the data, exploring patterns, and drawing inferences mirrors the methods used by statisticians and data scientists. In an era where computational tools such as R, Python and GIS (Geographic Information System) software enable increasingly sophisticated analyses, revisiting Snow's approach using contemporary methods offers a unique opportunity to engage students with both historical and modern perspectives on data science. Revisiting John Snow's cholera map in 2025 offers students a compelling, historically grounded introduction to spatial data analysis, especially in an era where geospatial epidemiology and storytelling through data are integral to contemporary statistical practice and communication.

In this article, we explore how Snow's visualisation can be incorporated into statistics education as an interactive case study. Using R, we outline the steps involved in reproducing the cholera map, demonstrating how students can use geospatial data manipulation and visualisation techniques to uncover the same insights Snow did over 150 years ago.

Snow's case study provides a foundation for discussing widely used statistical techniques. For instance, bar plots and histograms can illustrate disease counts across different city districts, scatter plots can visualise spatial clustering, and kernel density estimation (KDE) can provide a smooth, continuous surface that highlights areas with higher concentrations of cases and can provide a modern alternative to Snow's hand-drawn map. Regression modelling can extend the case study to infer relationships between environmental factors and disease spread.

We also discuss the pedagogical benefits of using historical case studies in statistics courses, emphasising how they can foster curiosity, critical thinking, and technical proficiency. Finally, we consider how this approach can be extended beyond epidemiology, illustrating broader applications in public health, urban analytics, and environmental science, where similar data visualisation and modelling techniques are vital.

By blending historical datasets with modern computation techniques, we can create engaging, hands-on learning experiences that not only reinforce core statistical principles, but also help students appreciate the real-world impact of data analysis. Snow's cholera map is more than just a historical artifact, it is a powerful example of how data visualisation can reveal insights, shape scientific understanding, and drive evidence-based decision-making, making it an invaluable resource for a statistical education.

To fully realise the educational value of data visualisation, it is important to examine why visualisation plays such a central role in statistics education. Snow’s map is more than historically significant. It embodies key principles of effective data storytelling, spatial reasoning, and analytical clarity. Before delving into the techniques involved in the recreation of the map, we first explore how visualisation fosters statistical thinking and why it remains a foundational skill in the modern data science classroom.

2. The Educational Value of Data Visualisation

Data visualisation is a fundamental tool in statistics education, allowing students to explore patterns, uncover relationships, and communicate insights effectively (Nolan and Perrett, 2016; Wilkerson et al., 2025). As Baumer et al. (2022) describe, it is “one of the most accessible, compelling, and expressive modes to investigate and depict patterns in data.” Unlike raw numerical summaries, visual representations provide an intuitive and immediate understanding of complex datasets. When raw data is translated into visual form, it becomes easier to interpret, allowing observers to quickly identify trends, anomalies, and key insights (Kirk, 2021). Whether through scatter plots, histograms or geospatial maps, visualisation helps students bridge the gap between abstract statistical concepts and real-world interpretation (Chang et al., 2024).

John Snow's 1854 cholera map offers a compelling case study in this regard. By plotting cholera cases spatially, Snow was able to reveal an epidemiological pattern that was not immediately obvious from tabular data alone. This example underscores a key lesson for students: data visualisation is not just about making data look appealing - it is a vital step in the analytical process that can lead to new discoveries and enables more informed decision-making (Tukey et al., 1977; Unwin, 2020; Newburger & Elmqvist, 2023; Bach et al., 2023).

In modern statistics and data science curricula, incorporating historical case studies like Snow's cholera map can serve multiple educational purposes:

1. **Enhancing Statistical Thinking**
   By visualising the spread of cholera cases, students engage in exploratory data analysis, a crucial step in any statistical workflow. This encourages pattern recognition, hypothesis generation and critical thinking, skills that are essential for informed decision-making. It also reinforces the idea that data should be represented visually before any formal modelling begins (Tukey, 1972; Baumer et al., 2022), as unexpected insights may emerge.
2. **Introducing Spatial Data and Geographic Information Systems**
   Traditional statistics courses often focus on numerical and categorical data, but spatial data introduces new dimensions of analysis. Snow's map allows students to work with coordinate-based data, spatial clustering techniques, and geospatial visualisation tools in R (Unwin, 2018). It also serves as an introduction to Geographic Information Systems (GIS) - computer-based tools for capturing, analysing, and visualising spatial data - which are widely used in fields such as epidemiology, urban planning, and environmental science (Cromley and McLafferty, 2011).
3. **Connecting Historical Data to Modern Applications**
   Many of the same principles Snow applied in the 19th Century are used today in public health monitoring, pandemic modelling and risk assessment. Students can draw parallels between Snow's cholera outbreak investigation and modern epidemiological studies, such as COVID-19 hotspot mapping (Franch-Pardo et al., 2020; Zhang and Schwartz, 2020) or air pollution analysis (Brunekreef and Holgate, 2002), which can use interactive features to enhance storytelling (Li, 2024). This connection reinforces the relevance of statistical reasoning beyond the classroom.
4. **Encouraging Computational Skill Development**
   Implementing Snow's visualisation in R introduces students to essential programming concepts, including data wrangling (e.g. dplyr (Wickham, François, et al. 2023) and tidyverse (Wickham, Averick, et al., 2019)), geospatial mapping (e.g. ggplot2 (Wickham, 2016) and sf (Pebesma and Bivand, 2023; Pebesma, 2018)), and reproducible research practices. This hands-on experience helps student build confidence in working with real-world data and computational tools. It demonstrates how open-source software can be used for both historical data analysis and contemporary statistical applications

By integrating historical data visualisation projects into statistics education, instructors can create engaging, interdisciplinary learning experiences that blend historical inquiry, data science and statistical reasoning. John Snow's cholera map is not just an artifact of epidemiological history; it is a timeless example of the power of data visualisation as an investigative and educational tool.

With this pedagogical foundation in place, we turn now to the practical implementation of this case study. Using R, students can bring this historical dataset to life, gaining firsthand experience with geospatial data, visualisation techniques, and computational workflows. The following section outlines how to recreate Snow’s cholera map step by step, highlighting the technical process and the educational opportunities it offers.

3. Recreating John Snow’s Map Using R

Reconstructing John Snow's 1854 cholera map using modern statistical tools provides an engaging hands-on exercise for students learning data visualisation, geospatial analysis and epidemiological modelling. With R, we can replicate Snow's approach by using real historical data, reinforcing key statistical concepts while demonstrating the way computation tools enhance our ability to analyse and visualise spatial data.

In this section, we outline the key steps involved in reproducing Snow's map using R, focusing on data acquisition, spatial visualisation and interpretation.

3.1 Preparing the Data

To recreate the cholera map, we first need data on:

- Cholera cases: the locations where cholera deaths were recorded.
- Water pumps: The suspected sources of contamination.
- Street layout: the geographical context for visualisation.

These datasets are available from historical sources and have been digitised for analysis in R. The SnowData package (Mimnagh, 2025) provides a clean version of Snow's data, which can be loaded as follows:

```
# Load required packages and data
library(ggplot2)
library(sf)
library(dplyr)
library(SnowData)
data(cholera_cases)
data(pump_locations)
```

At this stage, we have two key spatial datasets: one containing the cholera deaths and another with the water pump locations.

3.2 Mapping Cholera Cases and Water Pumps

Next, we visualise the data using ggplot2, overlaying the cholera cases and water pumps onto a simple map.

```
# Plot cholera deaths (red) and water pumps (blue)

ggplot() +
  geom_point(data = cholera_cases, aes(x = Easting, y = Northing),
            color = "red", alpha = 0.6, size = 2) + # Plot deaths as semi-
transparent red circles
  geom_point(data = pump_locations, aes(x = Easting, y = Northing),
            color = "blue", size = 4, shape = 17) + # Plot pumps as blue triangles
  labs(title = "John Snow's 1854 Cholera Map",
       subtitle = "Red circles = Cholera deaths, Blue triangles = Water pumps") +
  theme_minimal()# Use a clean background
```

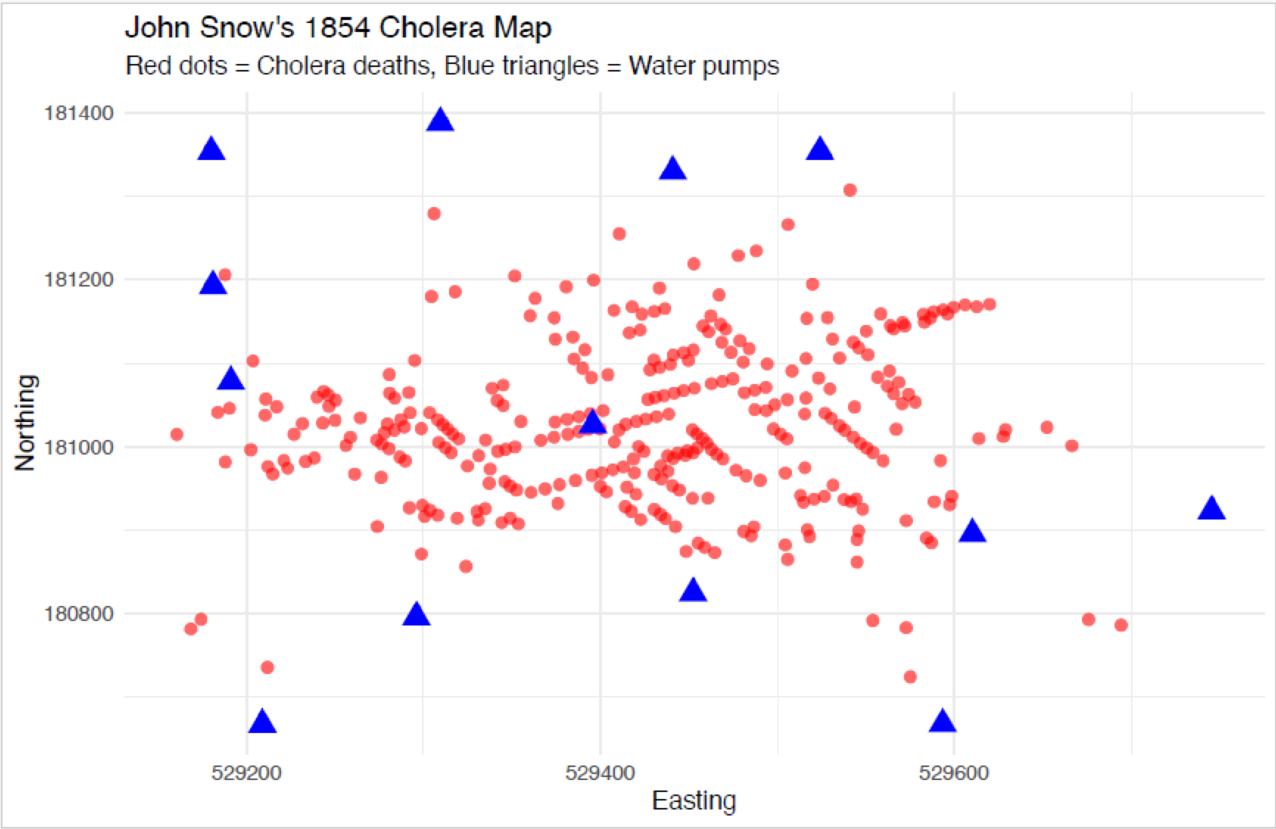

Figure 2: *Cholera cases and water pump locations during the 1854 Broad Street outbreak.* Cholera deaths are plotted as red circles and water pumps as blue triangles using coordinates from the British National Grid (EPSG:27700), with Easting and Northing on the x- and y-axes, respectively. Although no street map is shown, the spatial clustering of cases around the central Broad Street pump is clearly visible, illustrating the geographic concentration of the outbreak.

This initial visualisation highlights the spatial clustering of cholera cases around certain water pumps, reflecting Snow's original analysis.

3.3 Adding Street Layout for Context

To make the visualisation more informative, we can overlay the data onto the original street map. This can be achieved using a digitised version of Snow's map, available as part of the SnowData package, or OpenStreetMap data.

```
library(osmdata)
# SnowData uses the British National Grid (EPSG:27700), so we define the bounding
box using the same projection
soho_bbox_27700 <- st_bbox(c(xmin = 529150.0, ymin = 180720.6,
                              xmax = 529750.9, ymax = 181370.5),
                              crs = st_crs(27700))

# Transform bounding box to WGS84 (EPSG:4326) for OpenStreetMap API request
soho_bbox_4326 <- st_bbox(st_transform(st_as_sfc(soho_bbox_27700), crs = 4326))

# Request street data for roads ("highway") from OpenStreetMap
streets <- opq(bbox = soho_bbox_4326) %>%
```

```
            add_osm_feature(key = "highway") %>%
            osmdata_sf()

# Convert retrieved street data to EPSG:27700 to match Snow dataset
streets <- st_transform(streets$osm_lines, crs = 27700)

# Plot street network, cholera deaths, and pumps in projected space
ggplot() +
      geom_sf(data = streets, color = "gray50", size = 0.5)+ # OSM streets
      geom_point(data = cholera_cases,
                  aes(x = Easting, y = Northing),
                  color = "red", alpha = 0.6) + # Overlay deaths
      geom_point(data = pump_locations,
                  aes(x = Easting, y = Northing),
                  color = "blue",  shape = 17) + # Overlay pumps
      theme(panel.grid = element_blank(), # Clean up background
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      panel.background = element_rect(fill = "white"),
      plot.background = element_rect(fill = "white")) +
coord_sf(crs = 27700, datum = st_crs(27700)) # Ensure correct spatial reference
```

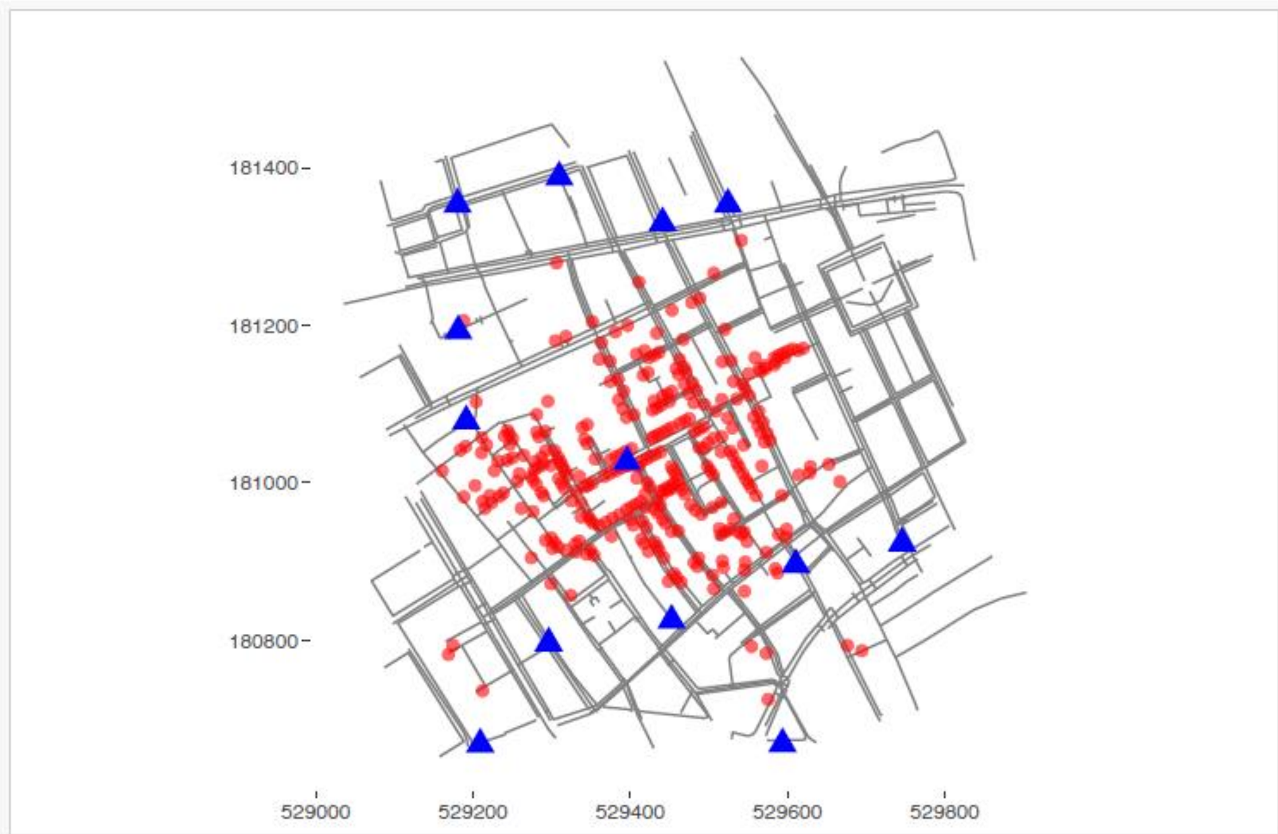

Figure 3: *Cholera cases and water pump locations overlaid on the 1854 Soho street-network.* Red circles represent individual cholera deaths, while blue triangles mark the positions of water pumps. The underlying street map provides geographic context, highlighting how cholera cases cluster near the Broad Street pump. This spatial pattern supports John Snow's hypothesis of a waterborne source of the outbreak.

Alternatively, as a more complicated exercise, students may attempt to accurately recreate John Snow's original 1854 map. Students may begin by loading the georeferenced map to use as a background.

```
map <- load_map() # Load base raster map for background
data(cholera_cases)
data(pump_locations)
library(rasterVis)
library(tidyr)
```

John Snow's original map represents the cholera outbreak in the Soho district of London, showing the locations of cholera deaths as black bars stacked along the streets. To create this plot, students must generate these bars and rotate them accordingly. This can be done using a user-defined function, and Angle and Count columns in the cholera cases dataset, to specify the necessary angle of rotation and the height of the black bars, respectively.

```
# Custom function to build bar-style rectangles, rotated to match the street orientation
rotate_rectangle <- function(x, y, count, angle) {
# Trigonometry functions in R expect radians so we convert from degrees.
 theta <- angle * pi / 180
# Width is fixed, and height is controlled by death count
 hw <- 2      # Fixed width for each bar
 hh <- count  # Bar height scaled by death count

# Define the four corners of a rectangle centred at the origin (0,0)
# The rectangle has a fixed width (hw) and a height (hh) proportional to the number of cholera deaths
corners<-matrix(c(-hw, 0,     # Bottom-left: x = -hw (left of centre), y=0
                  hw, 0,      # Bottom-right: x = +hw (right of centre), y=0
                  hw, hh,     # Top-right: x = +hw, y = hh (height above base)
                  -hw, hh),   # Top-left: x = -hw), y = hh (same as top-right)
                  ncol = 2, byrow = TRUE)

# Define a 2D rotation matrix to rotate the rectangle by a given angle (theta)
# This matrix will rotate points counterclockwise around the origin
rotation_matrix <- matrix(c(cos(theta), -sin(theta),
                            sin(theta),  cos(theta)), ncol = 2)

# Rotate the rectangle corners using matrix multiplication
rotated_corners <- corners %*% rotation_matrix

# Convert rotated corner coordinates into a data frame for plotting
rotated_corners <- data.frame(rotated_corners)
colnames(rotated_corners) <- c("x", "y")

# The rotated rectangle is currently centred at (0,0), we want to shift it to its correct map position (x,y)
rotated_corners$x <- rotated_corners$x + x
rotated_corners$y <- rotated_corners$y + y

return(rotated_corners)
}
# Generate a rotated rectangle for each cholera case
# Height is scaled by Count*2 to match the proportional bar heights in Snow’s original map
rectangles <- cholera_cases %>%
  rowwise() %>%
  mutate(corners = list(rotate_rectangle(Easting, Northing, Count * 2, Angle))) %>%
```

```
  # Expand the list of corners into individual rows for plotting polygons
  unnest(corners) %>%
  ungroup() %>%

  # Assign a unique group ID to each rectangle (used to draw polygons correctly)
  mutate(id = rep(1:nrow(cholera_cases), each = 4))

# Plot map with rotated bars (as in Snow’s original)
gplot(map, maxpixels = 5e6) + # Background raster
  geom_tile(aes(fill = value)) +
  scale_fill_gradient(low = 'black', high = 'white')+
  geom_point(data = pump_locations,
            aes(x = Easting, y = Northing),
            colour = "blue") +
  geom_polygon(data = rectangles,
        aes(x = x, y = y, group = id),
        color = "black") +
  theme(axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        axis.line = element_blank(),
        legend.position="none")
```

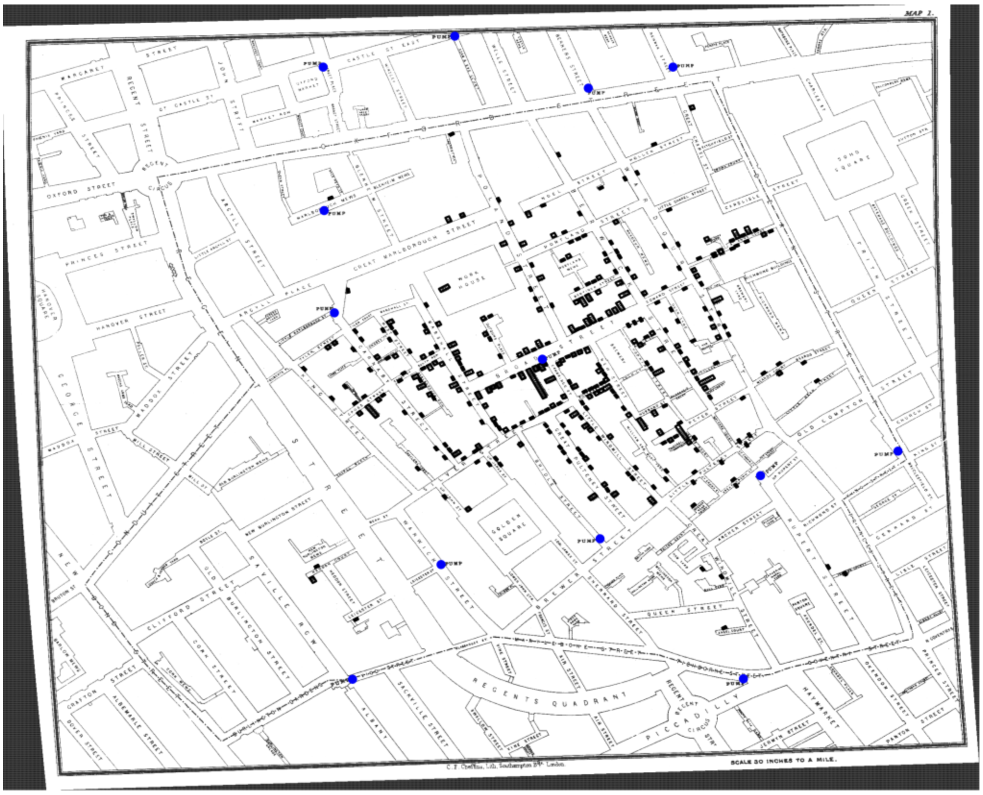

Figure 4: *Reproduction made with R of John Snow’s 1854 cholera map, showing cholera cases and water pumps overlaid on the original street layout of Soho.* Cholera deaths are represented as stacked black bars along the streets, with bar height indicating the number of deaths at each location. Water pumps are shown as blue circles. This historically faithful visualisation mirrors Snow’s original approach and illustrates how spatial clustering of cases near the Broad Street pump provided evidence for his theory of waterborne transmission.

The resulting map is slightly warped due to stretching that occurred during the georeferencing process, in which the original 1854 map was aligned to modern geographic coordinate systems. This distortion arises because historical maps can lack consistent scale, orientation, or reference points, making precise alignment with modern spatial data challenging. Such warping is common when working with historical spatial sources, particularly those that were hand-drawn or created before the standardisation of cartographic projection systems. Despite this, georeferencing enables students to overlay historical and modern spatial data, offering a compelling and context-rich view of Snow's findings while introducing practical challenges frequently encountered in spatial data analysis.

3.4 Interpreting the Results

Once the map is complete, students can explore spatial clustering and discuss key takeaways:

- The high concentration of cholera deaths around a single water pump (Broad Street pump) supports Snow's hypothesis.
- Some deaths occur further away, leading to discussions of outliers, confounding factors and alternative explanations - in interviewing Soho residents, John Snow found that some of those who died from cholera were known to prefer to collect water from the Broad Street pump even though it was not their nearest pump (Brody et al., 2000).
- Additional analytical techniques, such as spatial point pattern analysis, can be introduced to further explore the data.

For example, students could calculate the distance from each cholera case to the nearest water pump to quantify spatial patterns.

```
data(streets)
# Ensure all coordinate columns are numeric
streets <- streets %>% mutate(across(everything(), as.numeric))

# Convert start and end points into LINESTRING geometries
streets <- streets %>%
  rowwise() %>%
  mutate(geometry = st_sfc(st_linestring(matrix(c(start_coord_east,
                              start_coord_north, end_coord_east,
                              end_coord_north), ncol = 2,
                              byrow = TRUE)))) %>%
  ungroup() %>%
  st_as_sf(crs = 27700) # Convert to spatial object with British Grid CRS

# Convert point data for cases and pumps into spatial (sf) format
cholera_cases <- st_as_sf(cholera_cases,
                        coords = c("Easting", "Northing"),
                        crs = 27700)
pump_locations <- st_as_sf(pump_locations,
                        coords = c("Easting", "Northing"),
                        crs = 27700)

# Identify nearest pump to each cholera case
cholera_cases <- cholera_cases %>%
  mutate(nearest_pump = st_nearest_feature(geometry,
                                                pump_locations$geometry,
   dist = as.numeric(min(st_distance(geometry,
                         st_geometry(pump_locations))))))
```

```
# Create a connecting LINESTRING between each death and its nearest pump
connections <- cholera_cases %>%
  rowwise() %>%
  mutate(line = st_sfc(st_cast(st_union(geometry,
  pump_locations$geometry[nearest_pump]), "LINESTRING"))) %>%
  ungroup() %>% st_as_sf()

# Assign color based on pump ID
connections$pump_id<-as.factor(connections$nearest_pump)

# Plot deaths, pumps, and connecting lines
ggplot() +
  geom_sf(data = connections, aes(geometry = line, color = pump_id)) + #
Connections
  geom_sf(data = cholera_cases, color="black", size = 0.5, alpha = 0.8) + #
Deaths
  geom_sf(data = pump_locations, color = "red", size = 3, pch = 8) + # Pumps
  geom_sf(data = streets, alpha = 0.2)+ # Street background
  theme_minimal() +
  labs(color = "Pump ID")
```

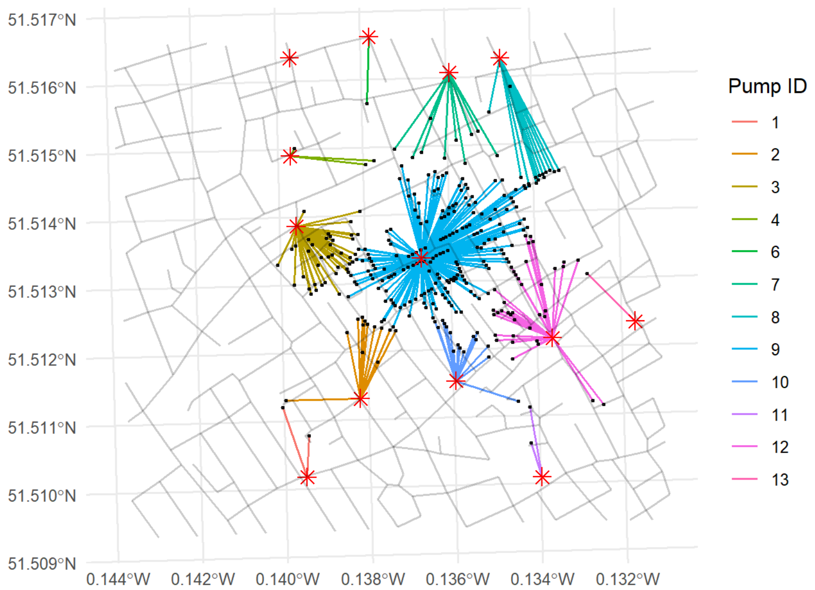

Figure 5: *Cholera cases, water pumps, and spatial connections during the 1854 Soho outbreak.* Black circles represent individual cholera deaths, and red stars indicate water pump locations. Each case is connected by a line to its nearest pump using Euclidean distance, with line colours distinguishing different pumps. Overlaid on the Soho street network, this modern spatial visualisation highlights potential zones of influence and reinforces John Snow's hypothesis that the outbreak was linked to a contaminated water source - particularly the Broad Street pump.

Figure 5 overlays cholera deaths and water pump locations onto the street network and connects each death to its nearest pump using straight lines (Euclidean distance). These connecting lines are colour-coded by pump ID, visually grouping deaths by their closest water source. This highlights clusters of cases associated with specific pumps - notably the dense cluster around the Broad Street pump. By examining which pumps are associated with more deaths, students can explore the spatial distribution of risk. This type of representation also allows discussions about spatial assumptions: although lines show the shortest possible distance "as the crow flies," they do not account for actual street layouts.

The visualisations presented in Figures 4 and 5 offer complementary perspectives on the 1854 cholera outbreak. Figure 4 prioritises historical fidelity by replicating the original visual language of Snow’s map - specifically, the stacked bars representing cholera deaths along streets - preserving the intuitive representation of deaths that Snow himself used to draw causal inferences. This allows students to appreciate the historical context and reasoning behind the original analysis. In contrast, Figure 5 introduces a more modern spatial visualisation by connecting each cholera case to its nearest water pump using Euclidean distance lines. While this approach offers analytical clarity and facilitates computational spatial analysis, it also abstracts away from the lived urban geography - ignoring street layouts and accessibility. Discussing both versions with students can deepen their understanding of how visual representation shapes analytical interpretation. It also highlights an important pedagogical point: modern tools can enhance precision and enable new kinds of analysis, but they may also impose assumptions that diverge from historical or practical realities.

3.5 Extending the Analysis

Once students have recreated Snow's map, they can modify the visualisation and extend their analysis in several ways:

- Comparing different visualisation types (e.g. point maps versus density plots as in Parker and Bahrami (2021)).
- Adding interactivity using shiny or leaflet to explore data dynamically.
- Applying similar methods to modern epidemiological datasets (e.g., COVID-19 case data).
- Performing predictive modelling using machine learning to identify potential hotspots based on simulated data.

These extensions help reinforce the practical applications of spatial statistics in real-world scenarios.

Recreating John Snow's cholera map using R is an effective way to teach students the power of data visualisation in statistics. This exercise introduces key concepts in exploratory data analysis, geospatial statistics and epidemiology, while also building practical computational skills.

4. Pedagogical Considerations and Applications

The recreation of John Snow's cholera map using R offers a rich opportunity for educators to introduce students to key concepts in data visualisation, spatial statistics and epidemiology. However, to maximise the educational impact of this exercise, instructors should consider how best to structure the activity, tailor it to different levels of student experience, and connect it to broader statistical and data science principles.

In this section, I outline practical strategies for incorporating this case study into statistics education, discuss common challenges and solutions, and suggest extensions that can help students deepen their understanding of spatial analysis and epidemiological modelling.

4.1 Integrating the Activity into a Statistics Curriculum

The John Snow cholera case study can be integrated into a variety of statistics and data science courses, including:

- Introductory Statistics: Teaching basic data visualisation, exploratory data analysis, and the importance of graphical methods in discovering patterns.
- Applied Data Science: Demonstrating the power of geospatial analysis and reproducible research using R.
- Public Health and Epidemiology: Connecting statistical reasoning to real-world disease modelling and outbreak investigation.
- History of Statistics: Providing a historical perspective on data-driven decision-making and how early statisticians shaped modern scientific inquiry.

Following the guidance of Cobb (1992), instructors may choose to structure the activity as:

- A guided tutorial, where students follow a step-by-step replication of John Snow's analysis.
- An open-ended project, where students experiment with different visualisation techniques and alternative datasets.
- A discussion-based case study, using Snow's map as a starting point for critical thinking about data ethics, causality and statistical inference.

To make implementation more concrete, Table 1 outlines suggested learning outcomes, classroom activities, and assessment strategies aligned with this case study. These are adaptable to different levels of statistical expertise, from introductory courses to applied data science modules.

Table 4: Suggested Learning Outcomes, Activities, and Assessments for Integrating the John Snow Case Study.

Learning Outcome	Classroom Activity	Assessment Strategy
Interpret and visualise spatial data using R	Recreate Snow’s map using cholera and pump datasets with ggplot and sf	Submit a reproducible report explaining each visualisation step
Explain the role of visualisation in hypothesis generation	Class discussion comparing Figures 4 and 5; prompt: “How can visualisation shape interpretation?”	Reflective essay: “What can we learn from a map?”
Use spatial reasoning to identify clustering and spatial patterns	Calculate nearest-pump connections and discuss outliers and exceptions	Short quiz on spatial measures (e.g., Euclidean vs. real-world distance)
Connect historical analysis to modern public health data	Compare Snow’s outbreak data to COVID-19 hotspot or air pollution maps	Group project: Apply same pipeline to a modern dataset and present findings
Evaluate ethical and practical implications of epidemiological data visualisation	Debate: “Do data visualisations clarify or obscure public understanding of risk?”	Group project: Identify and discuss potential biases, privacy issues and misinterpretations in visualisations

4.2 Common Challenges and How to Address Them

While the John Snow case study is engaging and historically significant, students may encounter several challenges when working with geospatial data in R, including those detailed by Kross and Guo (2019). Below in Table 2 are some potential difficulties and strategies for addressing them.

Table 2: Challenges that may arise in implementing the John Snow case study, and potential solutions

Challenge	Solution
Students unfamiliar with R	Provide starter code and structured walk-throughs to build confidence
Geospatial data concepts are new	Begin with simple visualisations (scatter plots) before introducing spatial mapping tools
Data manipulation is complex	Use tidyverse functions to introduce step-by-step data transformation
Students struggle with interpreting spatial patterns	Use interactive tools (RShiny, leaflet) to allow dynamic exploration of the data
Linking historical insights to modern applications	Assign follow-up tasks using current datasets (e.g., COVID-19 cases, air pollution data)

By anticipating these difficulties, educators can structure the learning process effectively and ensure that students engage with the material in a meaningful way.

4.3 Extending the Case Study for Deeper Learning

Once students have successfully recreated Snow's map, instructors can encourage further exploration through additional activities:

1. Data Interpretation and Statistical Inference
    - Ask students to quantify the clustering of cholera cases using nearest-neighbour distances or spatial autocorrelation methods.
    - Compare different visualisation methods, such as heatmaps, density plots or contour plots, to illustrate how different graphical techniques highlight patterns
2. Apply Spatial Analysis to Modern Data
    - Extend the methods learned in this case study to modern epidemiological data (e.g. COVID-19 cases, influenza outbreaks).
    - Use open datasets (e.g. public health, crime data) to explore spatial clustering in different contexts.
    - Compare Snow's approach to modern machine learning methods for disease mapping and risk prediction.
3. Creating an Interactive Shiny App
    - Students with programming experience can build an interactive Shiny dashboard that allows users to explore data dynamically.
    - By incorporating slider controls, hover-over tooltips, and filtering options, students can experiment with making data visualisations more engaging and accessible.

4. Ethical Discussions on Data-Driven Decision-Making
    - Snow's work ultimately influenced public health policy by leading to improvements in sanitation infrastructure.
    - Ask students to discuss modern ethical considerations in epidemiological modelling, such as data privacy, bias in disease surveillance, and the use of AI for outbreak prediction.

4.4 Reflections on the Educational Value of Historical Case Studies

Using historical examples like John Snow's cholera map provides students with more than just technical skills, it also helps them develop an appreciation for the role of statisticians in shaping scientific knowledge. This case study reinforces several broader lessons.

- The importance of visualisation in statistical discovery: Snow's work illustrates that seeing patterns in data is often the first step in generating hypotheses.
- The power of interdisciplinary thinking: Combining spatial data, medical knowledge and statistical reasoning led to statistical advancement.
- The value of open data and reproducibility: By recreating Snow's map in R, students engage in the same investigative process that epidemiologists and data scientists use today.

Through hands-on exploration, students not only gain computational and analytical skills, but also develop a deeper understanding of how statistics contributes to real-world problem-solving. This blend of technical knowledge, historical context, and critical thinking makes the John Snow cholera map an exceptional teaching tool for modern statistics education.

These pedagogical insights underscore the enduring relevance of Snow’s work in today’s data-driven world. As educators look to the future, the integration of historical case studies with modern analytical tools offers a promising model for teaching statistics in a way that is both rigorous and engaging. In the concluding section, we reflect on the broader implications of this approach and suggest potential directions for further development and adaptation.

5. Conclusions

John Snow's cholera map remains a compelling example of how visualisation and statistical reasoning can uncover hidden patterns and drive change. By recreating and analysing this historical case study, students engage with foundational concepts in spatial statistics and epidemiology while developing computational skills essential for modern data science.

Through this exercise, learners gain experience in data wrangling and geospatial visualisation using R packages, exploratory data analysis and pattern recognition in a real-world dataset, and critical thinking about statistical inference and causality, using a case where evidence led to major public health interventions.

Additionally, this pedagogical approach highlights the interdisciplinary nature of statistical thinking, connecting historical data analysis with contemporary public health and policy applications. The ability to bridge historical insights with modern computational methods fosters a deeper appreciation for the evolving role of statistics in decision-making.

While John Snow's cholera map is an ideal case study for teaching spatial statistics, the underlying principles can be extended to many other datasets. Future implementations of this approach could include:

- Contemporary Public Health

Analysing COVID-19 case distributions, vaccination uptake or air pollution data, and comparing different data visualisation techniques to communicate health risks effectively.

- Environmental and Climate Data

Exploring spatial patterns in climate-related data, such as temperature anomalies or flood occurrences.

- Urban Analytics and Social Science Applications

Using geospatial data to analyse crime rates, or transportation networks and examining spatial clustering in these datasets.

By incorporating these extensions, educators can adapt this historical case study into a versatile framework for teaching modern data science skills.

The use of historical case studies in teaching statistics provides an engaging, tangible way to illustrate core concepts while also highlighting the societal impact of data-driven decision-making. By revisiting Snow's work through a modern computation lens, not do we only honour one of the earliest applications of spatial statistics but also equip students with the skills to tackle today's pressing data analysis challenges.

6. References

Bach, B., Keck, M., Rajabiyazdi, F., Losev, T., Meirelles, I., Dykes, J., ... & Carpendale, S. (2023). Challenges and opportunities in data visualization education: A call to action. *IEEE Transactions on visualization and computer graphics*, *30*(1), 649-660.

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). *Modern data science with R*. Chapman and Hall/CRC.

Brody, H., Rip, M. R., Vinten-Johansen, P., Paneth, N., & Rachman, S., 2000. Map-making and myth-making in Broad Street: The London cholera epidemic, 1854. *The Lancet*, 356 (9223), 64–68.

Brunekreef, B., & Holgate, S. T., 2002. Air pollution and health. *The Lancet*, 360 (9341), 1233–1242.

Chang, H. Y., Chang, Y. J., & Tsai, M. J. (2024). Strategies and difficulties during students' construction of data visualizations. *International Journal of STEM Education*, *11*(1), 11.

Cobb, G., 1992. Teaching statistics. *Heeding the call for change: Suggestions for curricular action*, 22, 3–43.

Cromley, E. K., & McLafferty, S. L., 2011. GIS and public health. *Guilford Press*.

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L., 2020. Spatial analysis and GIS in the study of covid-19. a review. *Science of the total environment*, 739, 140033.

Friendly, M., & Denis, D. J., 2001. *Milestones in the history of thematic cartography, statistical graphics, and data visualization.* [online] Available at: http://www.datavis.ca/milestones [Accessed 07 March 2025]

Kross, S., & Guo, P. J., 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. *Proceedings of the 2019 CHI conference on human factors in computing systems,* 1–14.

Li, M. (2024). Incorporating Data Visualisation into Teaching and Learning. *Mathematics Education Research Group of Australasia*.

Mimnagh, N., 2025. *SnowData: Historical data from John Snow's 1854 cholera outbreak map* [R package version 1.0.0]. Available at: https://CRAN.R-project.org/package=SnowData

Newburger, E., & Elmqvist, N. (2023). Visualization according to statisticians: an interview study on the role of visualization for inferential statistics. *IEEE transactions on visualization and computer graphics*, *30*(1), 230-239.

Nolan, D., & Perrett, J., 2016. Teaching and learning data visualization: Ideas and assignments. *The American Statistician*, 70 (3), 260–269.

Parker, R., & Bahrami, M. (2021, August 3). *John Snow: The birth of epidemiology, data analysis & visualization*. Wolfram Blog. https://blog.wolfram.com/2021/08/03/john-snow-the-birth-of-epidemiology-data-analysis-visualization/

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10 (1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E., and Bivand, R., 2023. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC.

Snow, J., 1856. On the mode of communication of cholera. *Edinburgh medical journal*, 1 (7), 668.

Tufte, E. R., and Robins, D., 1997. *Visual explanations*. Graphics Cheshire, CT.

Tukey, J. W., 1972. Some graphic and semigraphic displays. *Statistical papers in honor of George W. Snedecor*, 5, 293–316.

Tukey, J. W., et al., 1977. Exploratory data analysis (Vol. 2). Springer.

Unwin, A., 2018. *Graphical data analysis with R*. Chapman; Hall/CRC.

Unwin, A., 2020. Why is data visualization important? what is important in data visualization. *Harvard Data Science Review*, 2 (1), 1.

Wickham, H., 2016. *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. Yutani, H., 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4 (43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D., 2023. *Dplyr: A grammar of data manipulation* [R package version 1.1.4] Available at: https://dplyr.tidyverse.org

Wilkerson, M. H., Kim, J., Lee, H. S., Stokes, D. J., & Ferrell, M. (2025). How Teachers Envision Using Data Visualization Discussion Tasks in Classroom Instruction. *International Journal of Science and Mathematics Education*, 1-35.

Zhang, C. H., & Schwartz, G. G., 2020. Spatial disparities in coronavirus incidence and mortality in the United States: An ecological analysis as of May 2020. *The Journal of Rural Health*, 36 (3), 433–445.