

# Project 4- Visualization

YENUS IBRAHIM AYALEW

January 16, 2026 | 01:20:34 | CET

---

```
# LOAD THE PACKAGES YOU ARE USING IN THIS CODE CHUNK
library(dplyr)
library(ggplot2)
#library(RColorBrewer)
```

---

## Task 1 - Principles of good data visualization

Over at Our World in Data you will find a chart illustrating child mortality vs. health expenditure, 2000 to 2019, across countries.

Download the data and reproduce the plot as closely as possible using only the 2019 data (i.e. the bubble scatter plot that you see when you move the slider to the right) and log scales. Your plot does not have to be interactive and the colors don't have to exactly match those from the original plot as long as your plot remains well readable and transports the same information as the original plot.

```
# -----
# TASK 1 - Child Mortality vs Health Expenditure (2019)
# -----
```

```
# Load necessary libraries
library(tidyverse)

# Load the 2019 dataset
child_mort_data <- read_csv(
  "child-mortality-vs-health-expenditure Logarithmic scale/child-mortality-vs-health-expenditure 2019.csv")

# Preview the column names to identify the population column
colnames(child_mort_data)
```

```
[1] "Entity"
[2] "Code"
[3] "Year"
[4] "Child mortality rate of children aged under five years, per 100 live births"
[5] "Current health expenditure per capita, PPP (current international $)"
[6] "Population (historical)"
[7] "World regions according to OWID"
```

```

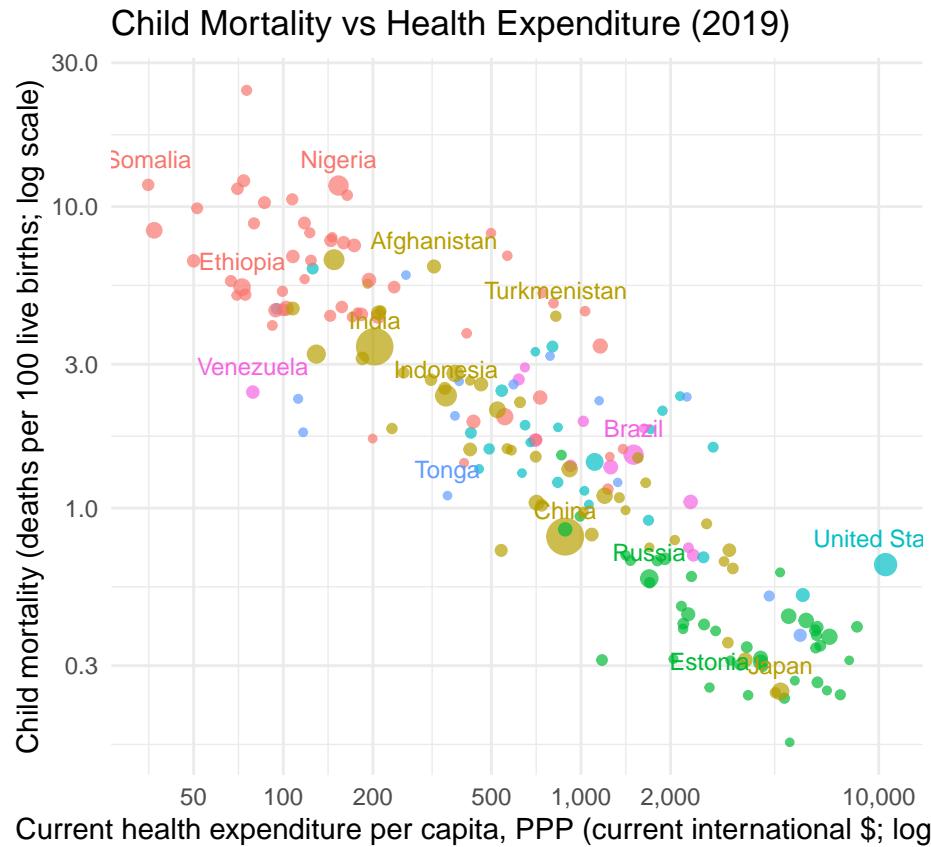
[8] "time...8"
[9] "time...9"
[10] "time...10"
[11] "time...11"

# Suppose the population column is called "Population (historical)" - adjust if different
child_mort_data <- child_mort_data %>%
  mutate(Population_billions = `Population (historical)` / 1e9)

# Countries to label on the plot
highlight_countries <- c("China", "India", "United States", "Indonesia", "Nigeria",
                         "Brazil", "Russia", "Japan", "Ethiopia", "Afghanistan",
                         "Estonia", "Somalia", "Tonga", "Turkmenistan", "Venezuela")

# Plot: Bubble scatter plot with log scales
ggplot(child_mort_data,
       aes(
         x = `Current health expenditure per capita, PPP (current international $)`,
         y = `Child mortality rate of children aged under five years, per 100 live births`,
         color = `World regions according to OWID`,
         size = Population_billions
       )) +
  geom_point(alpha = 0.7) +
  scale_x_log10(
    breaks = c(50, 100, 200, 500, 1000, 2000, 10000),
    labels = c("50", "100", "200", "500", "1,000", "2,000", "10,000")
  ) +
  scale_y_log10() +
  geom_text(
    data = child_mort_data %>% filter(Entity %in% highlight_countries),
    aes(label = Entity),
    vjust = -1,
    size = 3
  ) +
  labs(
    title = "Child Mortality vs Health Expenditure (2019)",
    x = "Current health expenditure per capita, PPP (current international $; log scale)",
    y = "Child mortality (deaths per 100 live births; log scale)",
    color = "WHO Region",
    size = "Population (billions)"
  ) +
  theme_minimal() +
  theme(legend.position = "right")

```



\*\*\* ### Task 2 - Principles of good data visualization

On slide 78 of the lecture slides (“Dos and”Don’ts”) you find a linked list of 20 statements expressing principles of good data visualization. Follow the links to learn more about them. Then, come up with another principle of good data visualization **that is not listed on the slide** and illustrate it following the instructions below:

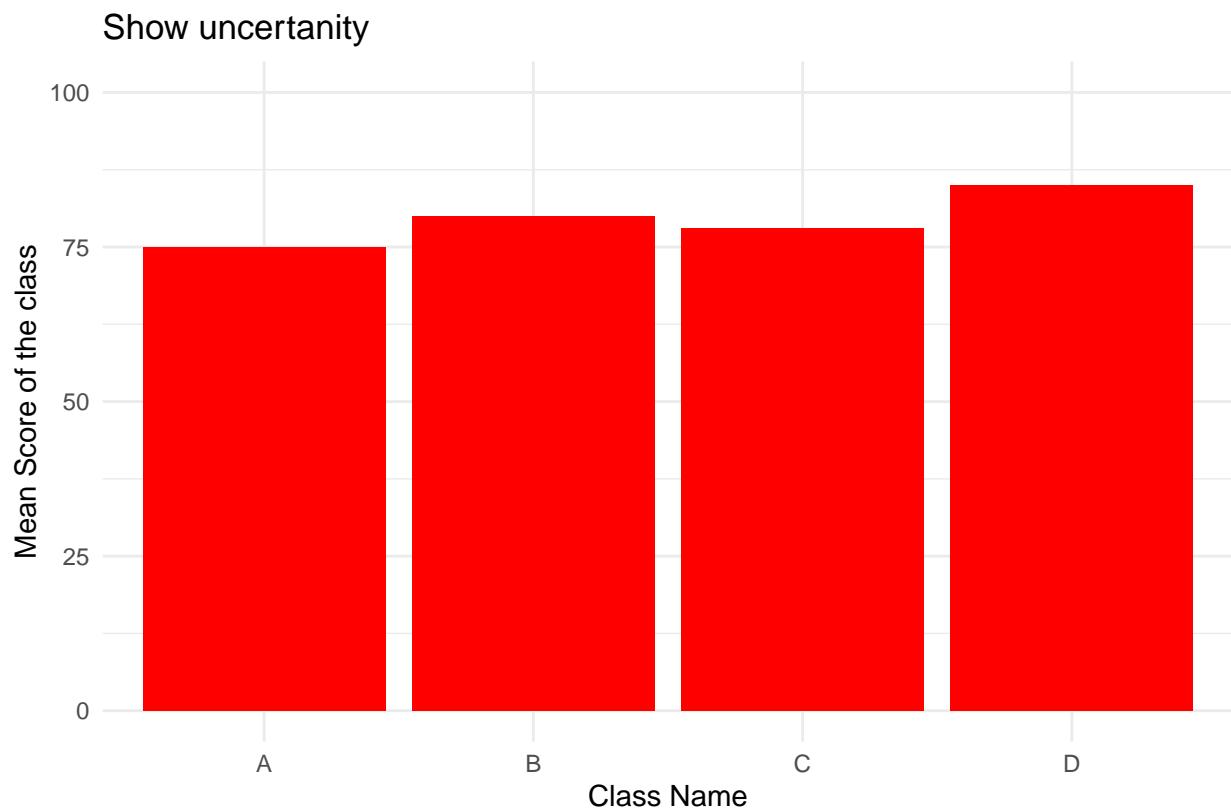
- Create a two-panel plot. The left panel shows a poorly designed plot (e.g., a 3D plot), the right panel shows a well-designed alternative using the same data. You are free to use whatever data you want to make your point.
- The title of the plot should be the name of the principle, e.g. “**Don’t go 3D.**”
- A note embedded in the bottom of the plot should explain, in a few sentences, the principle illustrated in the plot and how the right is an improved over the left version.
- Embed the plot in your .Rmd but also provide it as a .png in your submission repo.

```
# In the book by Claus O. Wilke's Fundamentals of Data Visualization, chapter 16, he stated how uncertainty
# if the data is visualized without showing uncertainty we tend to interpret it as a precise representation
#
# Sample data of 4 classes with mean and SD
df <- data.frame(
  Class = factor(c("A", "B", "C", "D")),
  Mean_score = c(75, 80, 78, 85),
  SD = c(5, 3, 6, 4)
)
```

```

# The bad plot which show only mean:
p1 <- ggplot(df, aes(x = Class, y = Mean_score)) +
  geom_col(fill = "red") +
  ylim(0, 100) +
  ggtitle("Show uncertainty") +
  theme_minimal() +
  labs(y = "Mean Score of the class",
       x = "Class Name",
       caption = "Left: Only mean values are shown.")
p1

```



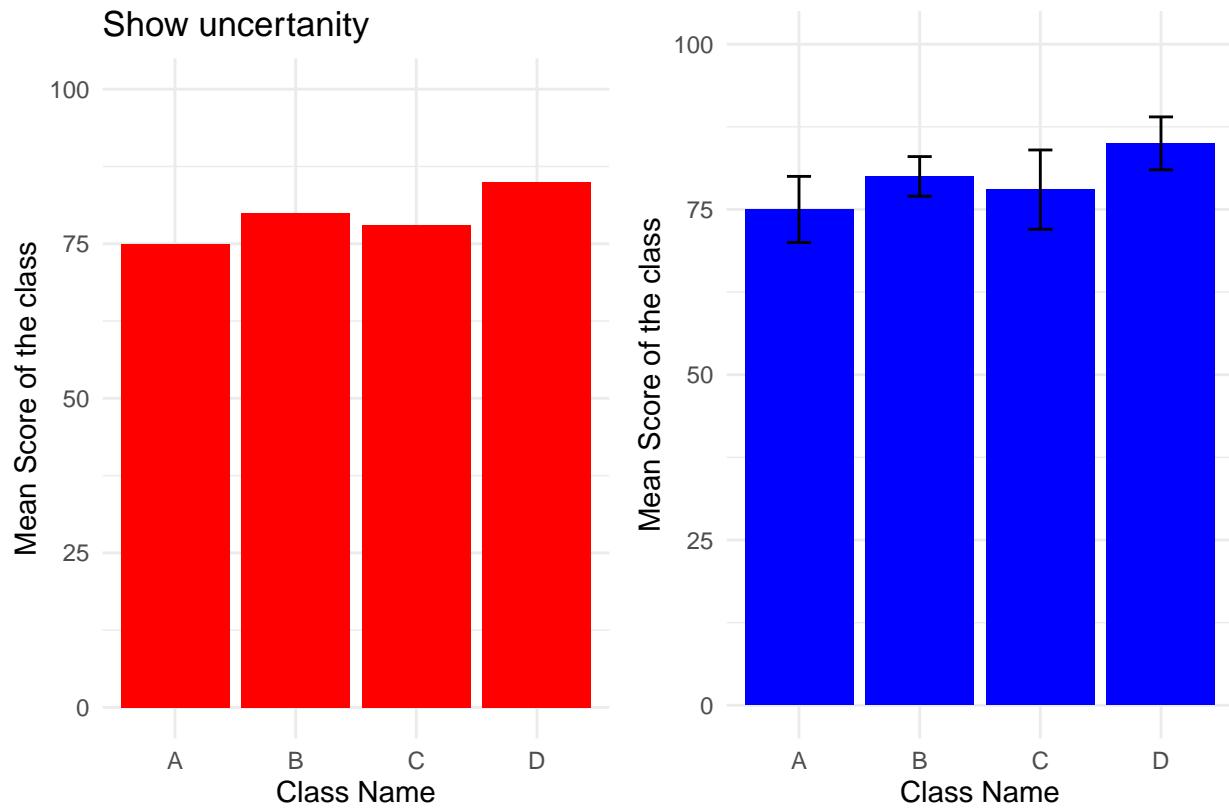
```

# Good plot: mean scores with error bars
p2 <- ggplot(df, aes(x = Class, y = Mean_score)) +
  geom_col(fill = "blue") +
  geom_errorbar(aes(ymin = Mean_score - SD, ymax = Mean_score + SD), width = 0.2, color = "black") +
  ylim(0, 100) +
  theme_minimal() +
  labs(y = "Mean Score of the class",
       x = "Class Name",
       caption = "Right: Error bars show standard deviation, variability across students.")

# Combine plots side by side
# In order to join the two plots, Grid extra packages is needed.
#install.packages("gridExtra")

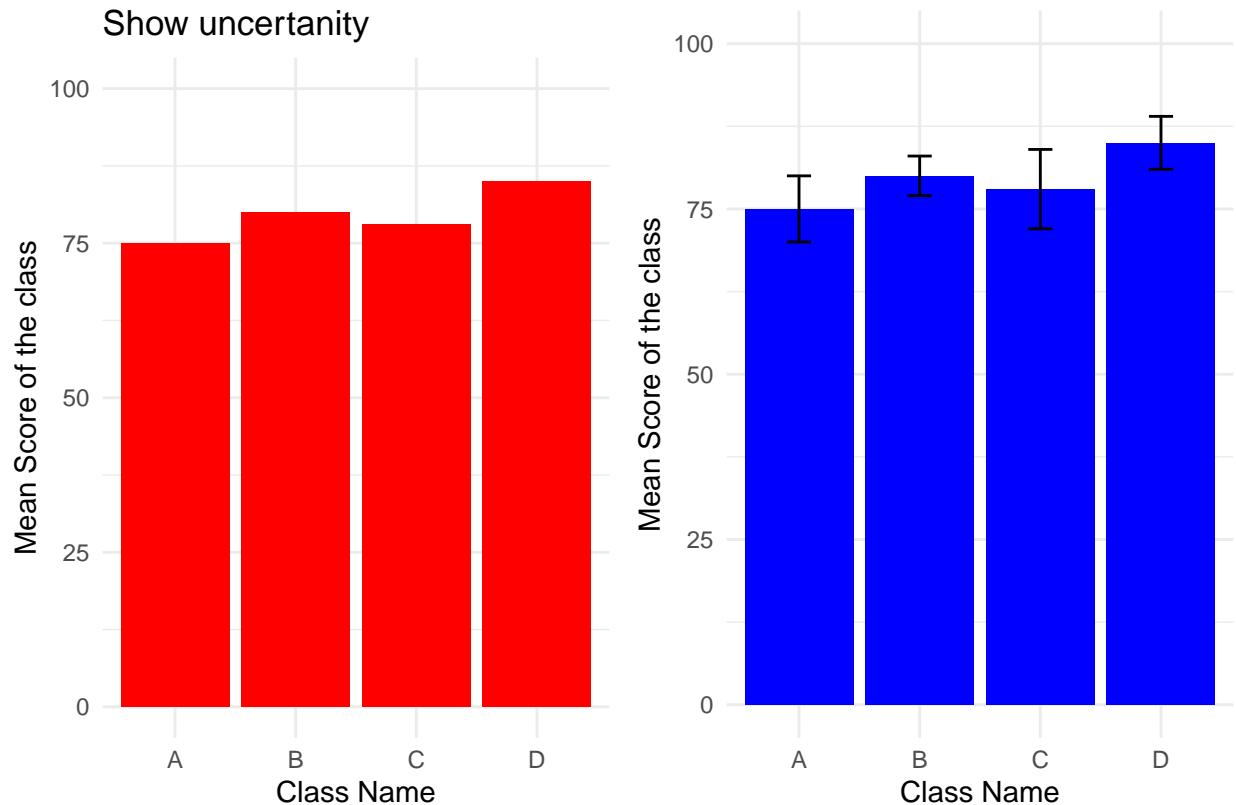
```

```
library(gridExtra)  
grid.arrange(p1, p2, ncol = 2)
```



```
# PNG  
ggsave("show_variability_error_bars.png", grid.arrange(p1, p2, ncol = 2), width = 10, height = 5)
```

### Show uncertainty



Left: Only mean values are shown. Right: Error bars show standard deviation, variability across students.