homework-03

<https://github.com/Zach-Tech/ENVS-193DS_homework-03.git>

# Load required libraries  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(here)

## here() starts at D:/Git/ENVS-193DS\_homework-03/ENVS-193DS\_homework-03

library(gt)  
library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(readxl)

## Problem 1a. Data Summarizing (5 points)

To compare my response variable—**number of steps per day**—between different **days of the week**, I could calculate the **mean number of steps** for each day. This would allow me to identify patterns, such as whether I am more active on weekdays compared to weekends.

This comparison is informative because my schedule varies during the week. For example, I have back-to-back classes on Monday and Wednesday, which usually requires more walking, whereas I tend to stay home on Sundays, resulting in fewer steps.

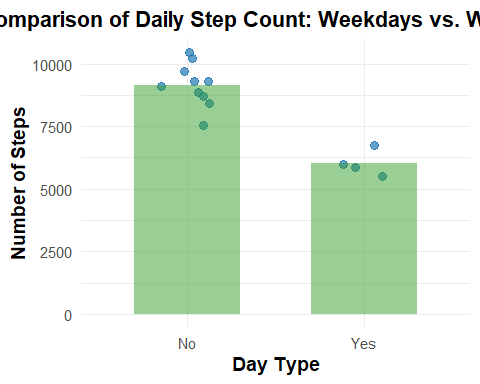
# Load the data

data <- read\_csv(here("data", "personal\_data.csv"))

## Rows: 14 Columns: 6  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): date, day, is\_weekend, mood  
## dbl (2): steps, hours\_sleep  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Problem 1b. Visualization (10 points)

# Summarize: calculate mean steps by weekend vs weekday  
library(dplyr)  
library(ggplot2)  
  
steps\_summary <- data %>%  
 group\_by(is\_weekend) %>%  
 summarize(mean\_steps = mean(steps), .groups = "drop")  
  
# Plot: points for individual observations + bar for mean  
ggplot(data, aes(x = is\_weekend, y = steps)) +  
 geom\_jitter(width = 0.15, size = 3, alpha = 0.7, color = "#1f78b4") + # individual data  
 geom\_col(data = steps\_summary, aes(x = is\_weekend, y = mean\_steps),   
 fill = "#33a02c", alpha = 0.5, width = 0.6) + # mean bars  
 labs(  
 title = "Comparison of Daily Step Count: Weekdays vs. Weekends",  
 x = "Day Type",  
 y = "Number of Steps"  
 ) +  
 theme\_minimal(base\_size = 14) +  
 theme(  
 plot.title = element\_text(face = "bold", hjust = 0.5),  
 axis.title.x = element\_text(face = "bold"),  
 axis.title.y = element\_text(face = "bold")  
 )

 ## Figure 1. Daily step counts separated by weekdays and weekends. Each point represents the number of steps taken on a specific day, while the bars represent the average (mean) step count for each category. The data show that, on average, more steps were taken on weekdays than weekends, likely due to a more structured schedule including classes and commuting.

## Problem 1d. Table Presentation (10 points)  
  
library(gt)  
  
# Create table from summary  
steps\_summary %>%  
 mutate(mean\_steps = round(mean\_steps, 1)) %>% # round to 1 decimal  
 gt() %>%  
 tab\_header(  
 title = "Average Daily Step Count",  
 subtitle = "Comparison of Mean Steps on Weekdays vs. Weekends"  
 ) %>%  
 cols\_label(  
 is\_weekend = "Day Type",  
 mean\_steps = "Mean Number of Steps"  
 ) %>%  
 tab\_options(  
 table.font.size = 14,  
 heading.title.font.size = 16,  
 heading.subtitle.font.size = 14  
 )

Table 1: Average Daily Step Count

Comparison of Mean Steps on Weekdays vs. Weekends

| Day Type | Mean Number of Steps |
| --- | --- |
| No | 9160.0 |
| Yes | 6032.5 |

## Problem 2a. Affective Visualization Description

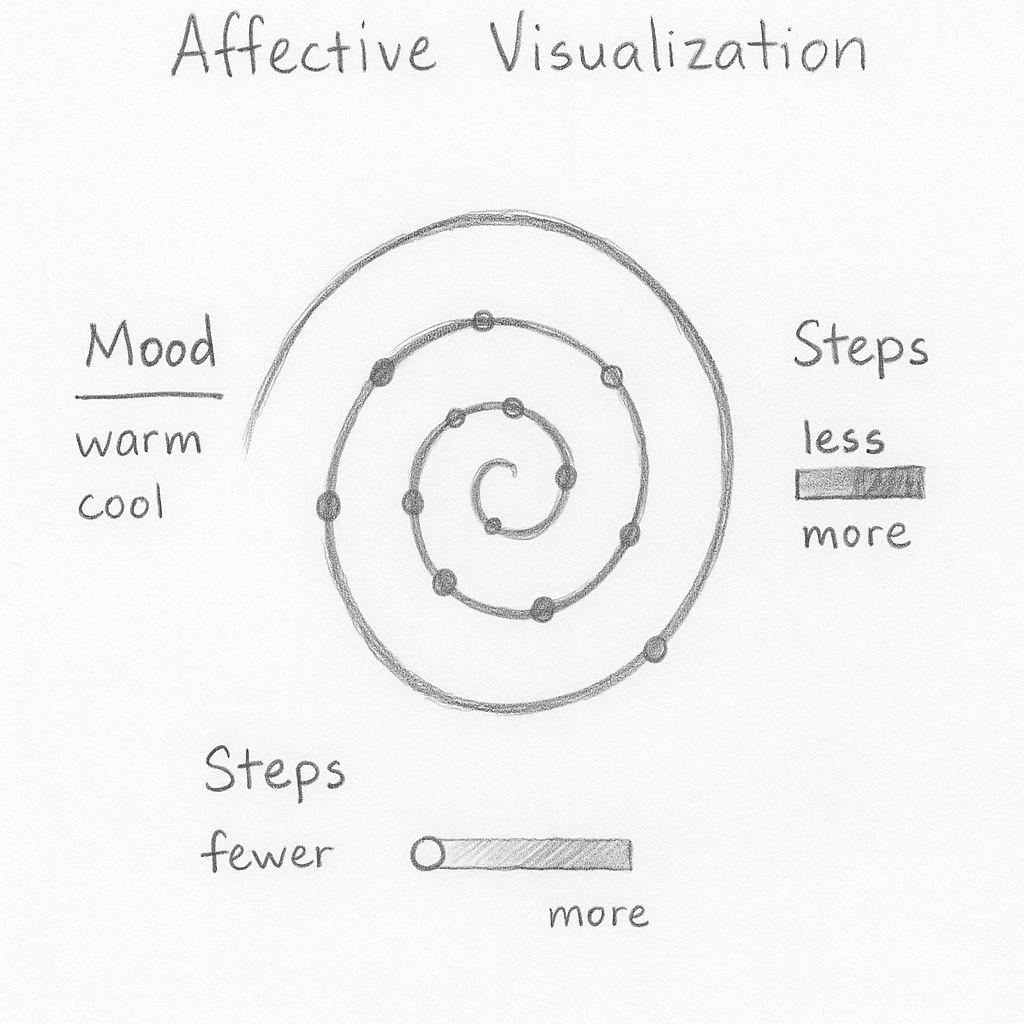
An affective visualization of my personal data could take the form of a color-coded spiral that maps my daily mood using warm and cool tones. Each loop in the spiral would represent one week, with dots or segments sized by how many steps I took that day and shaded by how many hours I slept. The center would start with the earliest week and spiral outward with each passing day. I might use red for low moods and blue or green for more positive moods to evoke an emotional response. This form of visualization would help viewers feel the rhythm of my week—the fluctuations in energy, rest, and emotion—rather than just read it from a chart.

##Code for Dataset Generation

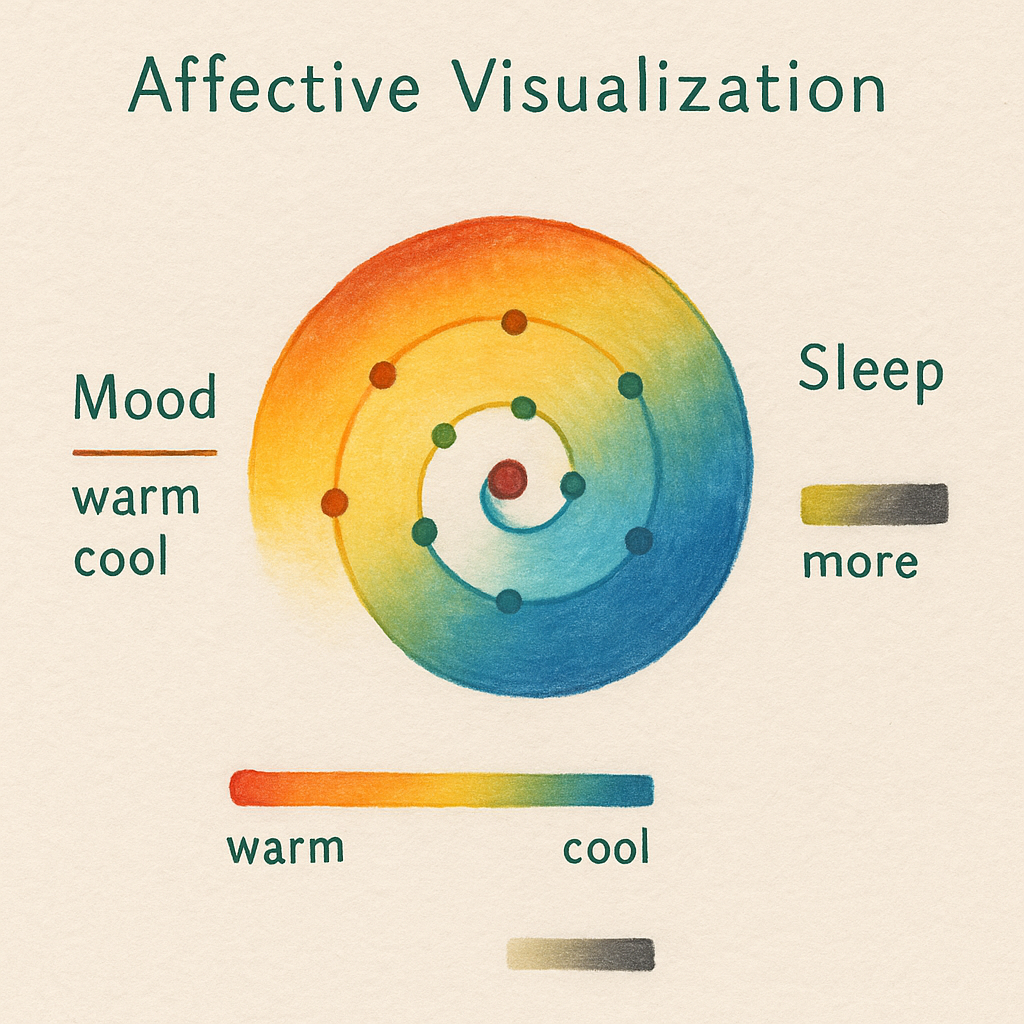
library(tibble)  
library(readr)  
library(here)  
  
data <- tibble(  
 date = as.Date("2025-05-12") + 0:13,  
 steps = c(6543, 7021, 8120, 4402, 6051, 10032, 9211,  
 7421, 5120, 8301, 6003, 4555, 11032, 10567),  
 sleep\_hours = c(6.5, 7.2, 6.8, 5.9, 6.0, 8.0, 7.5,  
 6.4, 5.5, 7.1, 6.0, 6.3, 8.5, 7.8),  
 mood = c(3, 4, 4, 2, 3, 5, 4, 3, 2, 4, 3, 2, 5, 5),  
 is\_weekend = weekdays(as.Date("2025-05-12") + 0:13) %in% c("Saturday", "Sunday")  
)  
  
# Save to your data folder  
write\_csv(data, here("data", "personal2\_data.csv"))

### Problem 2b. Sketch of Affective Visualization

Here is a sketch of my affective data visualization idea:

 ### Problem 2c. Draft of Affective Visualization

Here is a draft of my affective data visualization idea:

 ### Problem 2c. Artist Statement

This piece visualizes the relationship between my mood, physical activity (steps), and sleep patterns over a week. Each layer of the spiral represents a day, with colors shifting from warm to cool to reflect mood changes, dot size indicating step count, and background shading illustrating sleep duration.

I was particularly influenced by Stefanie Posavec and Giorgia Lupi’s Dear Data project, as well as Jill Pelto’s environmental watercolor work, both of which blend personal data with artistic expression to evoke emotion and insight.

The form of my work is a digitally crafted spiral visualization styled to resemble a watercolor painting, created using generative design tools.

To create this piece, I first analyzed trends in my personal data and then designed a visual metaphor in the form of a spiral, incorporating color, shape, and layout to communicate affective qualities like energy, restfulness, and mood.

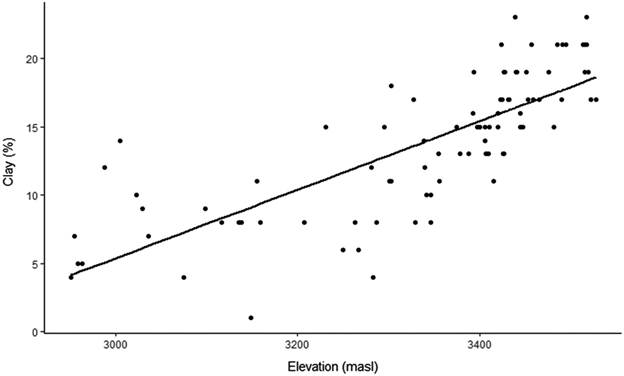
## Problem 3. Statistical critique

In the paper by Caulfield et al. (2020), the authors use a variety of statistical tests including ANOVA (Analysis of Variance), Kruskal-Wallis, and Wilcoxon rank-sum tests. These are used to assess differences in agroecological metrics—such as species richness, biomass, or clay content—across different land-use types and management practices. ANOVA was used when assumptions of normality and equal variance were met, while the non-parametric Kruskal-Wallis and Wilcoxon tests were applied when these assumptions were violated. These tests help the authors evaluate whether environmental, land management, and land-use variables interact to significantly affect ecological patterns in agroecosystems.ical critique

The x-axis of the figure represents Elevation (meters above sea level), while the y-axis shows Clay content as a percentage (%). The figure uses a scatterplot format with a fitted line that illustrates the trend. The main message of the figure is to show a positive correlation between elevation and clay content—clay content increases with higher elevation.

The figure does a good job communicating this result. The trend line adds clarity to the relationship, and the data points are not overly cluttered. However, the figure could be improved with clearer axis labels (e.g., including units explicitly) and perhaps a statistical annotation such as an R² value or p-value to help interpret the strength of the relationship.

Here is a draft of my affective data visualization idea:



## 3b. Visual clarity

The authors visually represented their statistics clearly in the figure. The x- and y-axes are logically positioned and labeled, showing elevation (meters above sea level) and clay content (%), respectively. The scatterplot displays both the raw data points and a fitted regression line, which effectively communicates the positive relationship between elevation and clay content. However, the figure could be improved by including summary statistics such as R² or a p-value to quantify the strength and significance of the trend. Additionally, clearer axis titles with units and a legend explaining the line would enhance interpretation further.

## 3c. Aesthetic clarity

The figure handles visual clutter well, with a clean scatterplot layout and an appropriate number of data points that do not overwhelm the viewer. The <data:ink> ratio is high—most visual elements directly convey meaningful information, such as the data points and fitted line. There are no unnecessary gridlines or distracting colors, though adding subtle visual emphasis (e.g., a contrasting line color or a minimal legend) could improve clarity further without adding clutter.

## 3d. Recommendations

To improve the clarity and interpretability of the figure, I would recommend the following enhancements:

Add Summary Statistics: Include the R² value and p-value directly on the scatterplot. This would allow viewers to immediately assess the strength and statistical significance of the relationship between elevation and clay content. These metrics are standard in regression visualizations and support transparency in data interpretation.

Clarify Axis Labels: While the current axes show the variables, they should include full units (e.g., “Elevation (m above sea level)” and “Clay Content (%)”) in a more prominent, bold font to improve readability and reduce ambiguity for readers unfamiliar with the variables.

Include a Legend: Add a small legend explaining the meaning of the regression line (e.g., “Fitted Linear Trend”) and perhaps distinguish between observed data points and predicted values if both are shown. This will help viewers who may not be statistically trained understand the purpose of each visual element.

Improve Visual Design: Consider enhancing the contrast and color scheme for accessibility—for example, using colorblind-friendly hues and slightly increasing the marker size or transparency (alpha) for overlapping data points to make dense areas easier to read.

Remove Redundant Gridlines: If gridlines are overly dominant or distracting, reduce their opacity or remove them entirely. This can help draw focus to the data points and regression line instead of the background.