ENVS 193DS - Homework 2

Marissa Maki Gill

2025-04-20

Set up

```
library(tidyverse)
library(janitor)

sbpm <- read_csv("sbpm.csv")

library(readr)
My_Data <- read_csv("ENVS193_MyData - Sheet1.csv", skip = 1)
```

Problem 1. Burrowing owl abundance

a.

[The type of data collected was discrete, as the observations could only take whole integer values. This was the case because we were interested in the number of burrowing owls that returned to the reserve, and distinct separate values needed to be represented for each owl returning.]

b.

```
counts <- c(0,0,1,1,2,2,3,4)
  sd <- sd(counts)
#sd = [1.41 owls]
```

[A better description of variability among the burrowing owl count in this problem, is standard deviation. Standard deviation measures the spread of individual data points (returning owls) in a sample population, whereas if we were to use standard error it would describe reliability in the mean estimate of the population.]

c.

```
counts <- c(0,0,1,1,2,2,3,4)
n <- length(counts)
se_owl <- sd(counts)/ sqrt(n)

#SE = [0.5 owls]
```

[While standard deviation is a better measure of variance among this sample, standard error best describes the uncertainty in the burrowing owl count. The standard error will

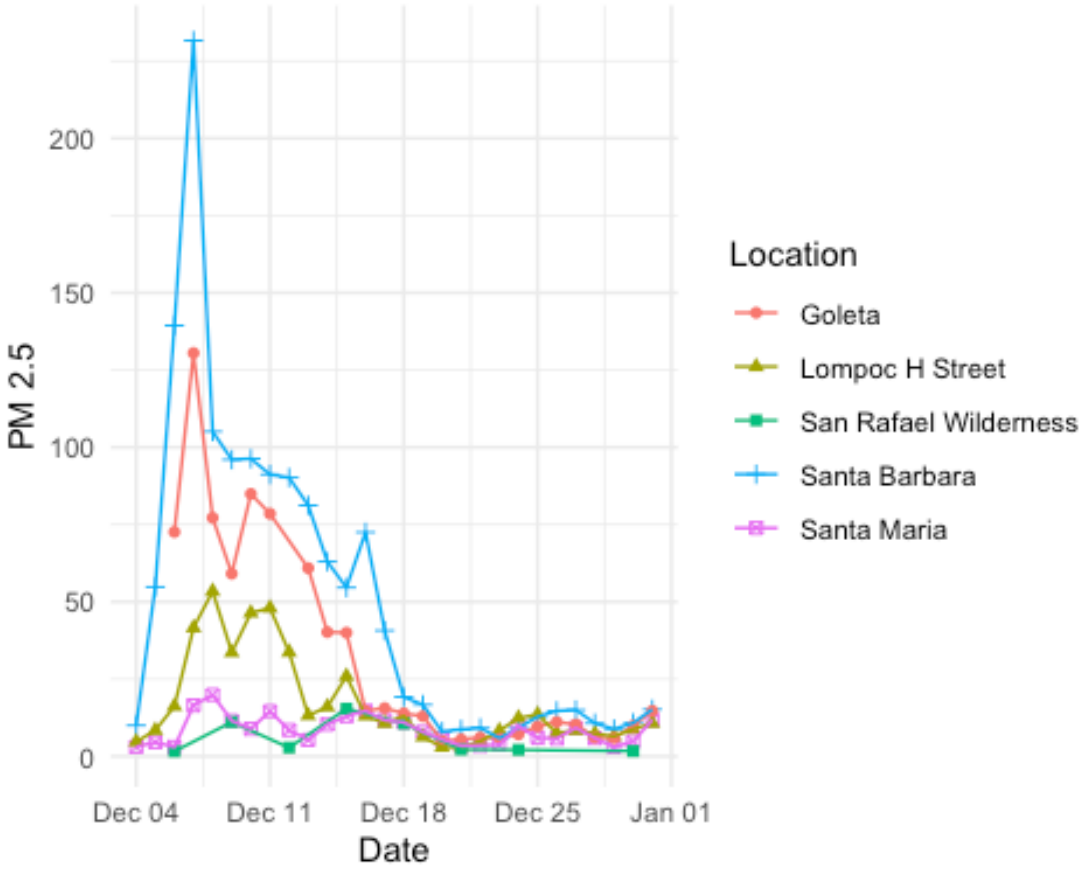
estimate the precision of the sample mean's representation of the true average number of owls observed.]

Problem 2. Fire and particulate matter

a.

```
ggplot(data = sbpm,
       aes(x = date,
           y = pm2_5,
           color = local_site_name,
           shape = local_site_name)) +
  geom_point() +
  geom_line() +

  labs(x = "Date",
       y = "PM 2.5",
       color = "Location",
       shape = "Location") +
  theme_minimal()
```



b.

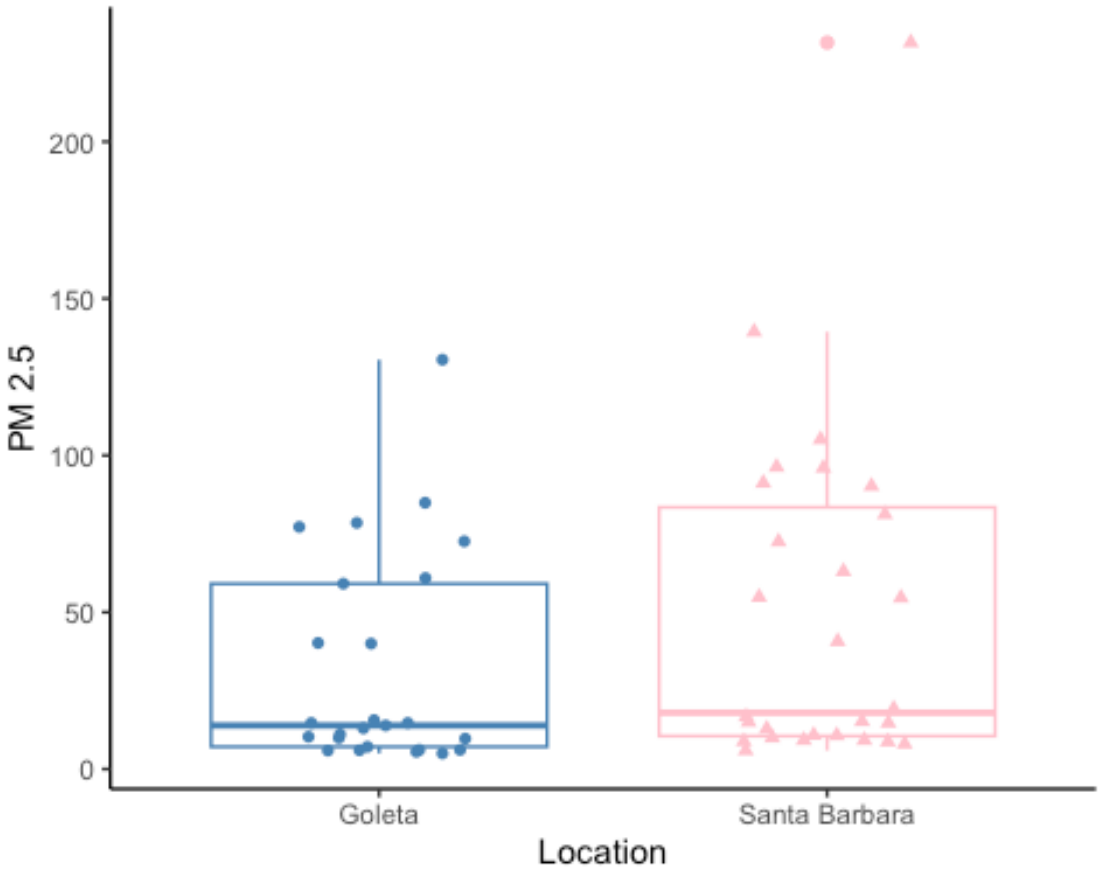
```
gol_sb <- sbpm |>
  filter(local_site_name == "Goleta" | local_site_name == "Santa Barbara") #
filter for both sites
```

c.

[The null hypothesis suggests that the PM 2.5 levels were the same in both Goleta and Santa Barbara from the start of the fire until it was contained, while the alternative hypothesis states that the levels differed throughout that time at these two locations.]

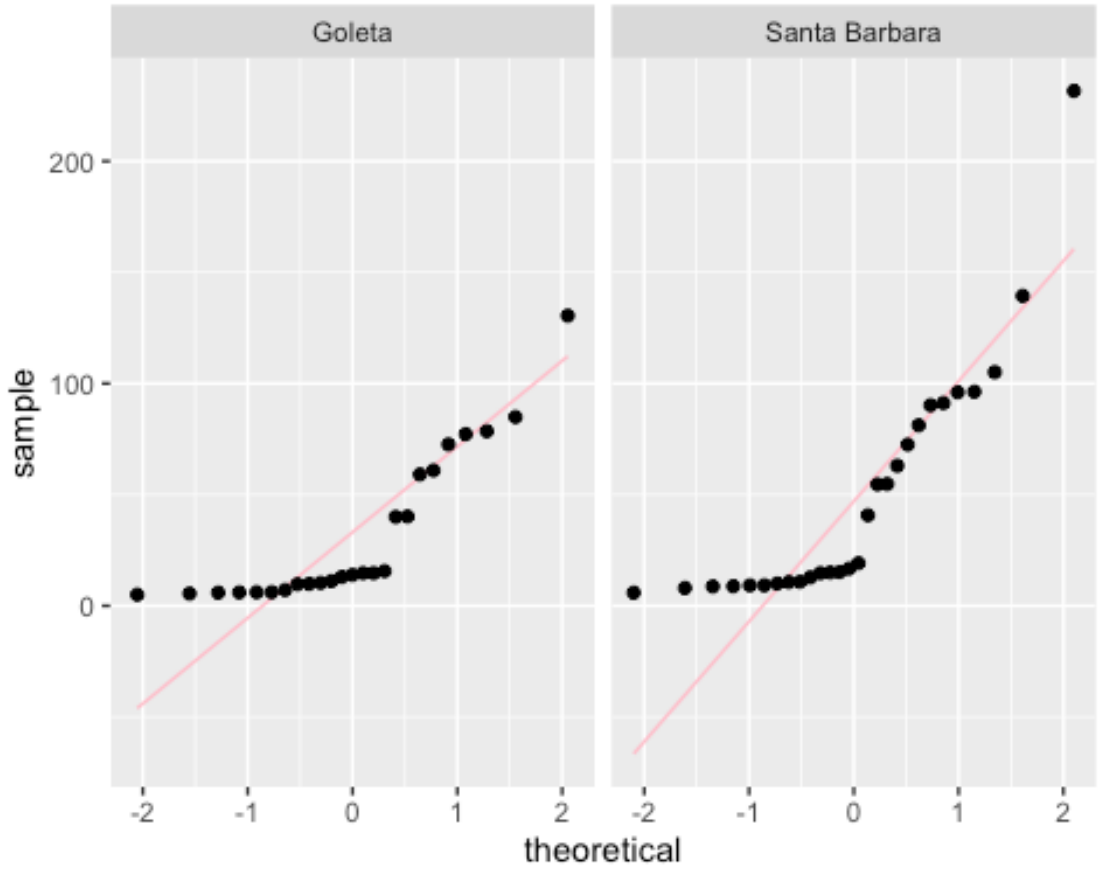
d.

```
ggplot(data = gol_sb,
       aes(x = local_site_name,
           y = pm2_5,
           color = local_site_name,
           shape = local_site_name)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, #  points jitter horizontally
              height = 0) + # points don't jitter vertically
  scale_color_manual(values = c("Goleta" = "steelblue", "Santa Barbara" = "pi
nk")) + #custom colors
  scale_shape_manual(values = c("Goleta" = 16, "Santa Barbara" = 17)) + #cust
om shapes
  labs(x = "Location", # labeling the x-axis
       y = "PM 2.5") + # labeling the y-axis
  theme_classic() +
  theme(legend.position = "none") # remove legend
```



e.

```
ggplot(data = gol_sb,
       aes(sample = pm2_5)) + # y-axis
  # QQ reference line
  geom_qq_line(color = "pink") +
  # QQ plot
  geom_qq() + # adds data points
  # creating "facets"
  facet_wrap(~local_site_name) # create separate QQ plots for sites
```



f.

[Among these two sites, the PM 2.5 variable is not normally distributed, as the QQ points in the plot above do not follow a strait line or cluster.]

g.

```
# calculate variances
PM25_var <- gol_sb |>
  group_by(local_site_name) |>
  summarize(variance = var(pm2_5))

PM25_var

# A tibble: 2 × 2
  local_site_name variance
  <chr>              <dbl>
1 Goleta             1177.
2 Santa Barbara      2800.

#[2.38]

# doing F test of equal variances
```

```
var.test(
  pm2_5 ~ local_site_name,
  data = gol_sb)


    F test to compare two variances

data:  pm2_5 by local_site_name
F = 0.42044, num df = 24, denom df = 27, p-value = 0.03524
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1915806 0.9401427
sample estimates:
ratio of variances
         0.4204419
```

[Using an F test of equal variances, we determined that variances were not equal and the ratio of variances was 0.42, while the p-value was 0.03524, less than the 0.05 significance level. This determines that the variances of Goleta and Santa Barbara are significantly different and not equal.]

h.

```
t.test(
  pm2_5 ~ local_site_name, # response variable ~ grouping variable
  var.equal = TRUE, # argument for equal/unequal variances
  data = gol_sb
)


    Two Sample t-test

data:  pm2_5 by local_site_name
t = -1.43, df = 51, p-value = 0.1588
alternative hypothesis: true difference in means between group Goleta and gro
up Santa Barbara is not equal to 0
95 percent confidence interval:
 -42.684941   7.172084
sample estimates:
       mean in group Goleta mean in group Santa Barbara
                   31.94000                    49.69643
```

i.

[A t-test would have been appropriate to test the hypothesis from part A because we are comparing the mean PM2.5 levels between two completely independent groups to identify statistical significance throughout the duration of the fire. Normality was evaluated using a QQ plot model for each site identifying if the points followed a linear patter, and homogeneity of variance was observed through a variance test comparing the variance

between two groups. The variable was not normally distributed, as observed by the QQ plot, but a t-test was still justified because the sample sizes were relatively equal moderating violations of normality according to the Central Limit Theorem.]

j.

[To compare the mean PM2.5 levels among 28 observations in Santa Barbara, and 25 observations in Goleta, a two-sample t-test was performed. This test resulted in a t-statistic of -1.43, with 51 degrees of freedom. The produced p-value was 0.1588 which is much larger than a significant level of p = 0.05, suggesting no difference in mean PM2.5 levels during the fire event between Santa Barbara and Goleta.]

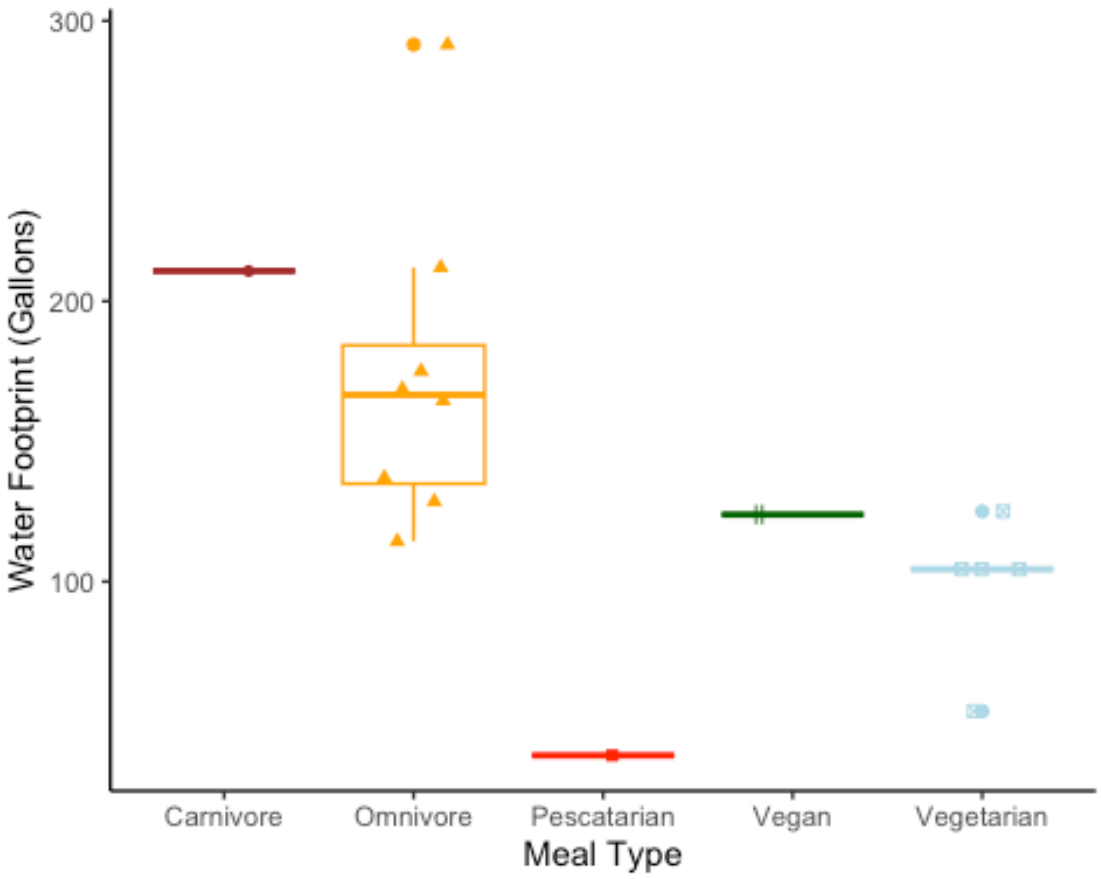
Problem 3. Personal data

```
library(dplyr)
library(ggplot2)
```

a.

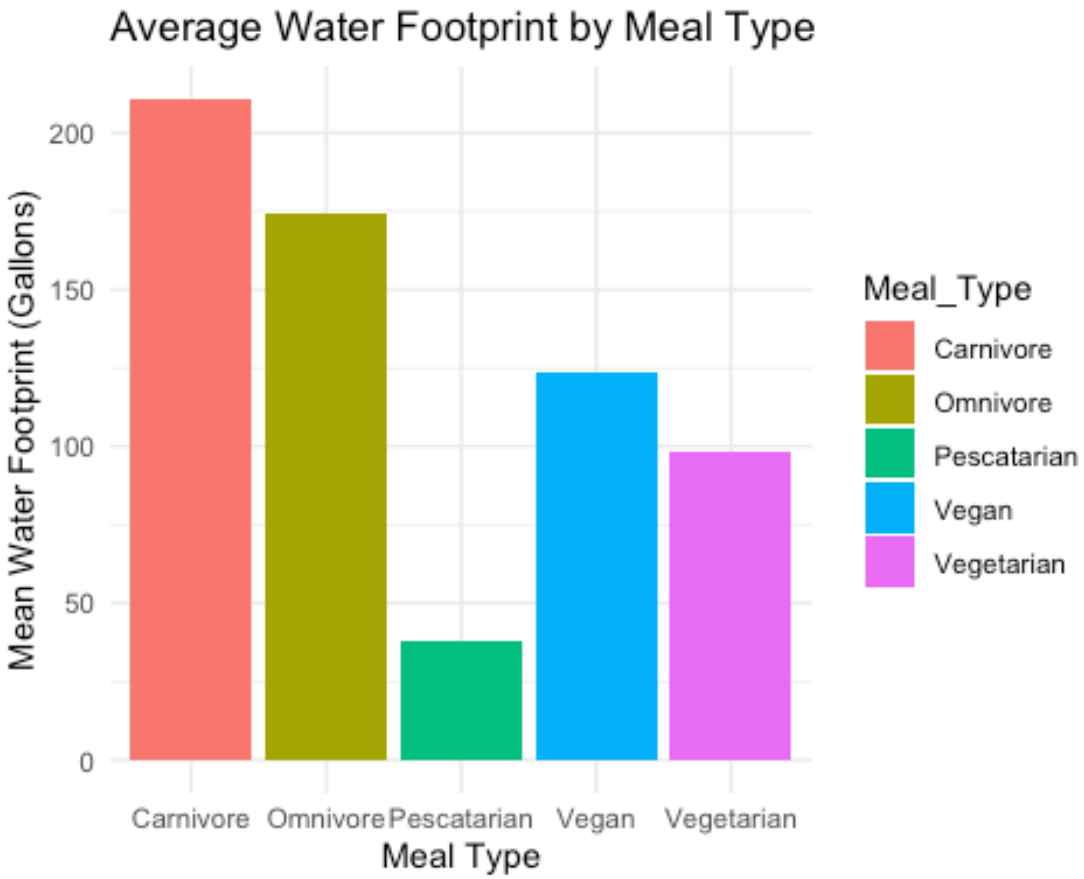
```
ggplot(data = My_Data,
       aes(x = Meal_Type,
           y = Water_Footprint,
           color = Meal_Type,
           shape = Meal_Type)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, #  points jitter horizontally
              height = 0) + # points don't jitter vertically
  scale_color_manual(values = c("Vegan" = "darkgreen", "Omnivore" = "orange",
"Vegetarian" = "lightblue", "Carnivore" = "brown", "Pescatarian" = "red")) +

  labs(x = "Meal Type", # labeling the x-axis
       y = "Water Footprint (Gallons)") + # labeling the y-axis
  theme_classic() +
  theme(legend.position = "none") # remove legend
```



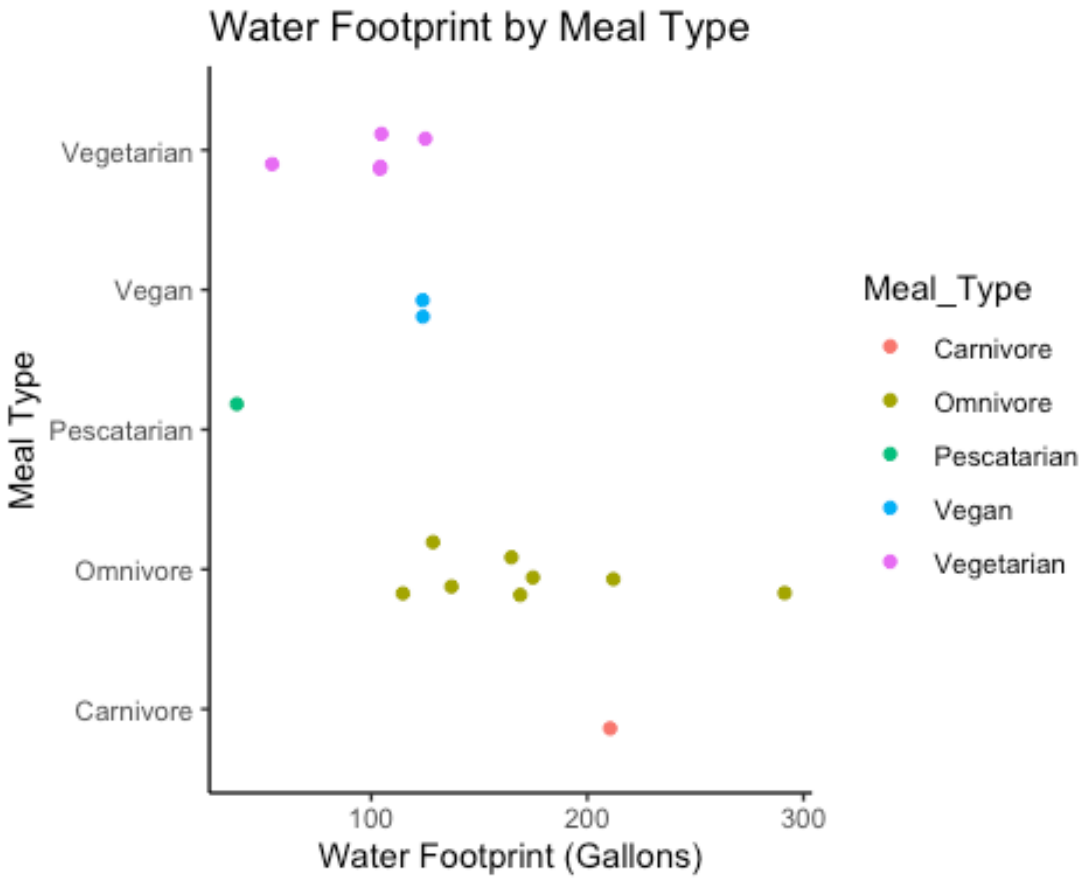

```r
# Summarize data
summary_data <- My_Data |>
  group_by(Meal_Type) |>
  summarise(
    Mean_Footprint = mean(Water_Footprint, na.rm = TRUE),
    se = sd(Water_Footprint, na.rm = TRUE) / sqrt(n())
  )

# Histogram
ggplot(summary_data,
       aes(x = Meal_Type,
           y = Mean_Footprint, fill = Meal_Type)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Water Footprint by Meal Type",
       x = "Meal Type",
       y = "Mean Water Footprint (Gallons)") +
  theme_minimal()
```



b.

```r
ggplot(My_Data, aes(x = Water_Footprint, y = Meal_Type, color = Meal_Type)) +
  geom_jitter(height = 0.2) +
  labs(
    x = "Water Footprint (Gallons)",
    y = "Meal Type",
    title = "Water Footprint by Meal Type"
  ) +
  theme_classic()
```



c.

[So far while analyzing my data I can already see trends in how meal type is affecting my water footprint. Meals with meat included are highest in water production, while vegan and vegetarian meals are offering a smaller daily footprint. Once i collect more data I expect these insights to stay mostly the same, but I may consider adding more description into the meal type and add a type of meat category because beef is less water efficient than chicken. From my collected data I also feel as though I need another continuous variable because I struggled to come up with a graph for question b.]

d.

[The process from getting my data from my spreadsheet collection into R was very smooth. The way I organized my data in a "long" format, and the amount of rows seem to be working well. The only challenge I had when importing the csv was during the wrangling stage. I found it difficult to remove the default heading but after a few Google searches I was able to use the "skip = 1" command. I also had to change the names for the labels because they were not being recognized within R with spaces.]

Problem 4. Statistical critique

a.

[This quarter I am taking a class about rivers along with other environmental and ecological related coursework. When presented this assignment, I was interested in finding a paper that bridged this class' methods and skills with what I am learning in ENVS 144. I found the paper "Assessment of stream quality and health risk in a subtropical Turkey river system: A combined approach using statistical analysis and water quality index" and after reading the methods and intro found that the researchers approach on determining stream water quality fit my interests perfectly.]

b.

[The authors of this article pose to answer the question of how domestic pollution and agricultural activities effect water quality in the Turnasuyu Basin, in Turkey. They aim to asses the spatial and temporal water quality of the stream to evalulate suitability and risk for human use.]

c.

[The statistical tests this paper used to find an answer to their research were Analysis of Variance [(Response: temperature, pH, DO and other water quality parameters), (Predictor: season, sampling station)] and the Kruskal-Wallis test [(Response: salinity, NO3-N, etc), (Predictor: Season or Station)].]

d.

[The One-Way ANOVA is used to test for significant differences in water quality parameters over stations and different seasons, with a p-value of <0.05 would be considered signficant. The Kruskal-Wallis test was used when the data collected was nonparametric, not normally distributed, to test the similar parameters as the ANOVA, differences in water quality, and the same p-value measuring significance.]

e.

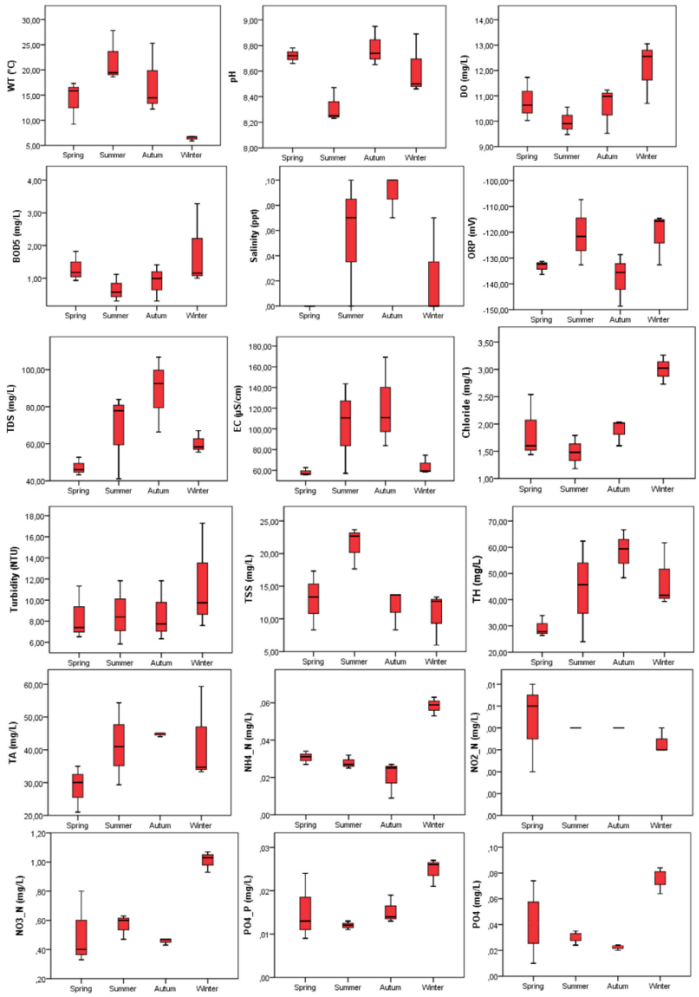


Fig. 2. Boxplot graph for hydrochemical parameters at different seasons of the stream.

f.

[The figure depicts boxplot graphs of hydrochemical parameters, assessing seasonal variation and data normality. The x-axis is season, while the y-axis is varying chemical contents in the tested waters.]