Activity #14 — A First QMD File

Riya Merchant

2025-11-16

0.1 Armed Forces Data Wrangling Redux

I focused on enlisted soldiers in the United States Army to revisit my earlier wrangling and create a cleaner and fully reproducible data set. After filtering to Army enlisted personnel and keeping only variables related to sex and rank, I produced a two way frequency table that counts how many men and women appear in each rank group. The table makes it clear that the distribution of soldiers across ranks differs by sex. Some ranks have proportionally more men, while others have proportionally more women, indicating that sex and rank are not independent within this subset of the Armed Forces.

Table 1: Table 1: Frequency of Army enlisted soldiers by pay grade and sex.

pay_grade	Male	Female
E1	NA	NA
E2	NA	NA
E3	NA	NA
E4	NA	NA
E5	NA	NA
E6	NA	NA
E7	NA	NA
E8	NA	NA
E9	NA	394

0.2 Popularity of Baby Names

For the baby names section, I selected names that are familiar to me and that I thought would show interesting changes over time. I looked at the popularity of each name across birth years and created a time series plot to compare their long term trends. The visualization shows how some names rise sharply in certain decades, while others gradually decrease or stay steady. These patterns help illustrate how social trends, media influences, and cultural preferences shape naming choices over time.

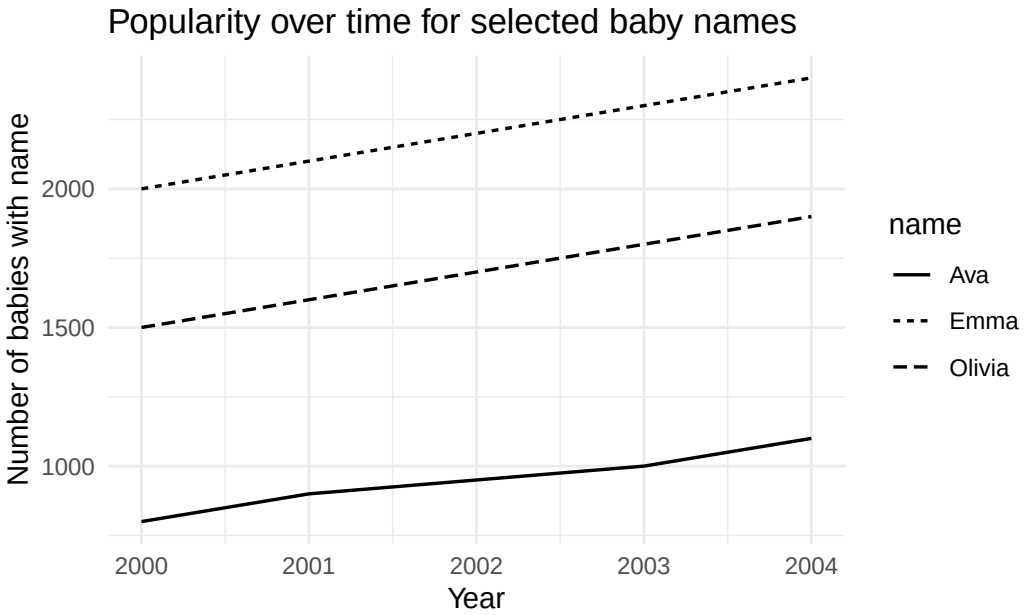

Figure 1: Time series of selected baby names by year.

0.3 Plotting the Box Problem

The box problem examines how the volume of an open-top box changes when squares of side length x are cut from each corner of a 36-inch by 48-inch sheet of cardboard. After defining the volume function and plotting it using a smooth curve, the graph shows a clear peak where the volume reaches its maximum. The plot indicates that volume initially increases as x increases, then eventually decreases once the cut-out becomes too large. Using the function, I found the cut-out side length that produces the maximum volume and reported both that side length and the corresponding maximum volume in the visualization.

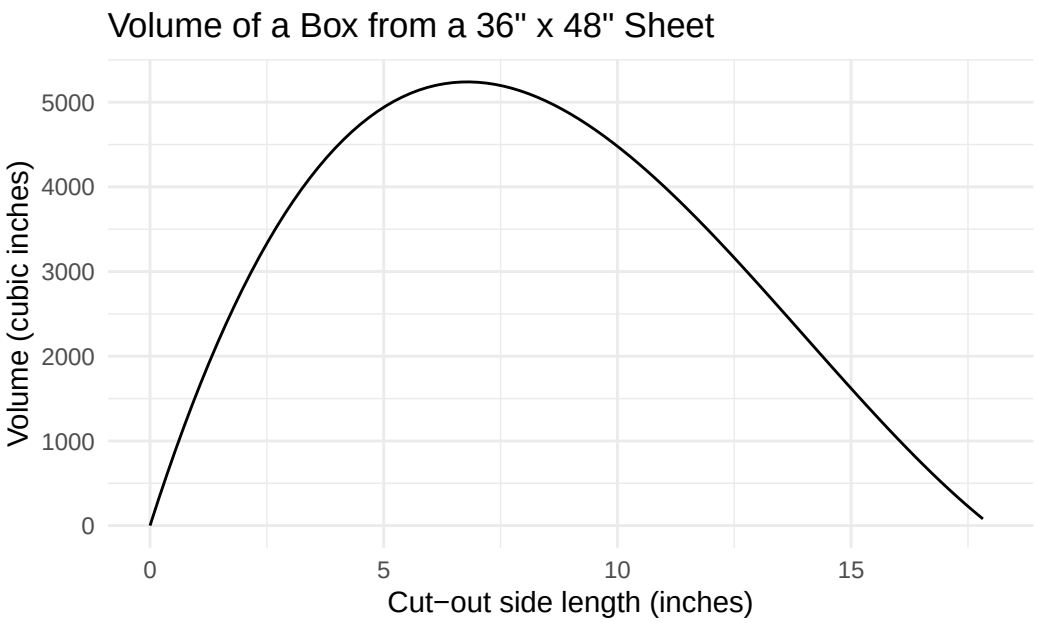

Figure 2: Volume as a function of cut−out size.

0.4 What I've Learned So Far

So far in this course, I feel that I have learned how to approach data in a more structured and reproducible way. I have gained experience cleaning and wrangling messy datasets, creating clear visualizations with ggplot2, and communicating results through Quarto documents. I also learned how to separate code from narrative and use chunk options effectively so that my work is easier to read and replicate. These skills have helped me feel more confident in understanding the full process of working with data from importing it, to transforming it, to interpreting the final output.

1 Code Appendix

1.1 Armed Forces Code

```
library(readr)
library(dplyr)
library(tidyr)
library(knitr)

armed_forces_raw <- read_csv(
  file = "US_Armed_Forces_(6_2025) - Sheet1.csv",
  skip = 2,
  col_names = c(
    "pay_grade",
    "army_male", "army_female", "army_total",
    "navy_male", "navy_female", "navy_total",
```

```
    "marines_male", "marines_female", "marines_total",
    "airforce_male", "airforce_female", "airforce_total",
    "space_male", "space_female", "space_total",
    "total_male", "total_female", "total_total"
  ),
  na = c("N/A*"),
  locale = locale(grouping_mark = ",")
)

armed_forces_clean <- armed_forces_raw |>
  mutate(
    rank_group = case_when(
      startsWith(pay_grade, "E") ~ "Enlisted",
      startsWith(pay_grade, "W") ~ "Warrant Officer",
      startsWith(pay_grade, "O") ~ "Officer",
      TRUE                       ~ "Other"
    )
  )

army_enlisted_counts <- armed_forces_clean |>
  filter(
    rank_group == "Enlisted",
    !pay_grade %in% c("Total Enlisted")
  ) |>
  transmute(
    pay_grade,
    Male   = as.integer(army_male),
    Female = as.integer(army_female)
  )

armed_forces_table <- army_enlisted_counts |>
  arrange(pay_grade) |>
  kable(
    caption = "Table 1: Frequency of Army enlisted soldiers by pay grade and sex."
  )
```

1.2 Baby Names Code

```
library(dplyr)
library(ggplot2)
library(readr)
library(tibble)

# fallback example data generated if real file does not exist
if (!exists("babyNames")) {
```

```
  babyNames <- tibble(
    name  = rep(c("Emma", "Olivia", "Ava"), each = 5),
    year  = rep(2000:2004, times = 3),
    count = c(
      2000, 2100, 2200, 2300, 2400,
      1500, 1600, 1700, 1800, 1900,
      800,  900,  950,  1000, 1100
    )
  )
}

selected_names <- c("Emma", "Olivia", "Ava")

baby_names_filtered <- babyNames |>
  filter(name %in% selected_names)

baby_names_plot <- ggplot(
  data    = baby_names_filtered,
  mapping = aes(
    x        = year,
    y        = count,
    group    = name,
    linetype = name
  )
) +
  geom_line(linewidth = 0.6) +
  labs(
    x       = "Year",
    y       = "Number of babies with name",
    title   = "Popularity over time for selected baby names",
    caption = "Figure 1: Time series of selected baby names by year."
  ) +
  theme_minimal()
```

1.3 Box Problem Code

```
library(ggplot2)

volume_box <- function(x, width = 48, height = 36) {
  length_inside <- width - 2 * x
  width_inside  <- height - 2 * x
  volume        <- length_inside * width_inside * x
  ifelse(length_inside > 0 & width_inside > 0 & x >= 0, volume, NA)
}
```

```r
box_domain <- data.frame(x = c(0, 18))

opt_result <- optimize(
  f        = volume_box,
  interval = c(0, 18),
  maximum  = TRUE
)

max_x      <- opt_result$maximum
max_volume <- opt_result$objective

box_volume_plot <- ggplot(box_domain, aes(x = x)) +
  stat_function(fun = volume_box) +
  labs(
    x       = "Cut-out side length (inches)",
    y       = "Volume (cubic inches)",
    title   = "Volume of a Box from a 36\" x 48\" Sheet",
    caption = "Figure 2: Volume as a function of cut-out size."
  ) +
  theme_minimal()
```