Activity #14 — A First QMD File

Kirubel Tamrat

2025-11-12

Armed Forces Data Wrangling (Activities #08 and #10)

Armed Forces Data Wrangling (Activities #08 and #10)

For this section, I used the Active Duty dataset to revisit my wrangling work from Activities #08 and #10. I focused specifically on Army Officers, since this subgroup has enough observations to meaningfully compare sex and rank. Pay grades beginning with "O" represent commissioned officers, so I filtered the dataset to include only those rows.

The frequency table below shows how many male and female Army officers there are at each pay grade. By comparing the counts across O1 through O10, we can start to understand whether sex and rank appear to be independent. If sex and rank were independent, we would expect the distribution of men and women to look similar across all officer ranks.

Table 1. Army Officers by Sex and Pay Grade

```
# A tibble: 1 x 12
  branch sex      O1   O10    O2    O3    O4    O5    O6    O7    O8    O9
  <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Army   Male   7122    11  9550 20986 12350  6939  3161   100    80    46
```

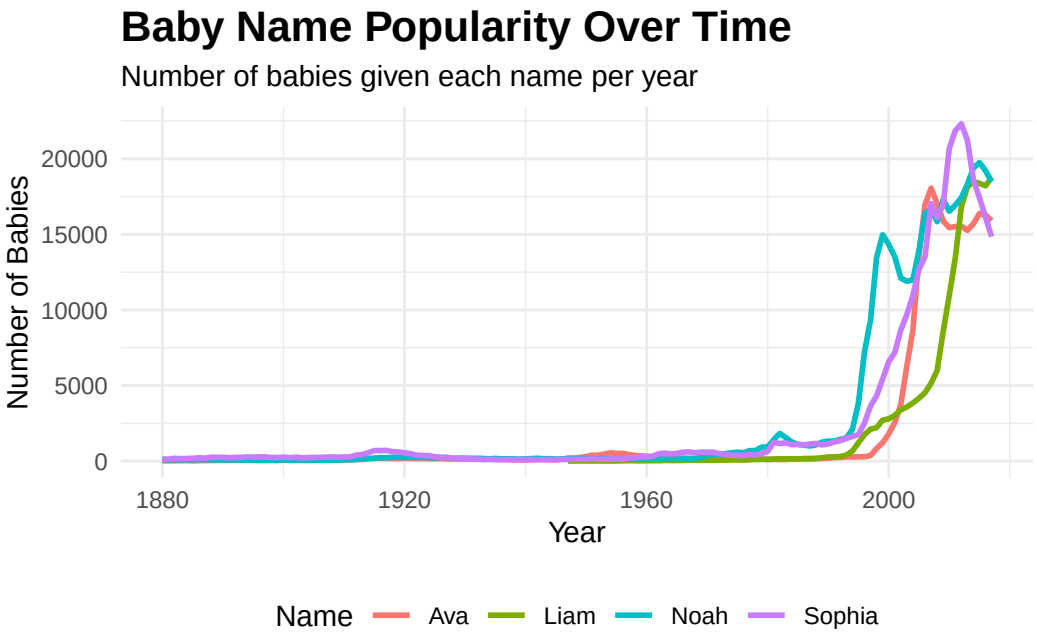
What we Learned from This Table

From the table, male officers significantly outnumber female officers at every pay grade. The size of this difference is especially noticeable at the higher officer ranks, where the counts for women drop much more sharply than the counts for men. This suggests that sex and rank are not independent within Army Officers. Instead, the proportion of women becomes smaller as rank increases, which indicates a structural imbalance across officer levels.

Popularity of Baby Names (Activity #13)

For this section, I recreated my visualization from Activity #13 using the babynames dataset. I chose the names Ava, Liam, Sophia, and Noah because they represent both genders. These names make trends easier to see because they have long histories in the data and strong rises and falls in popularity.

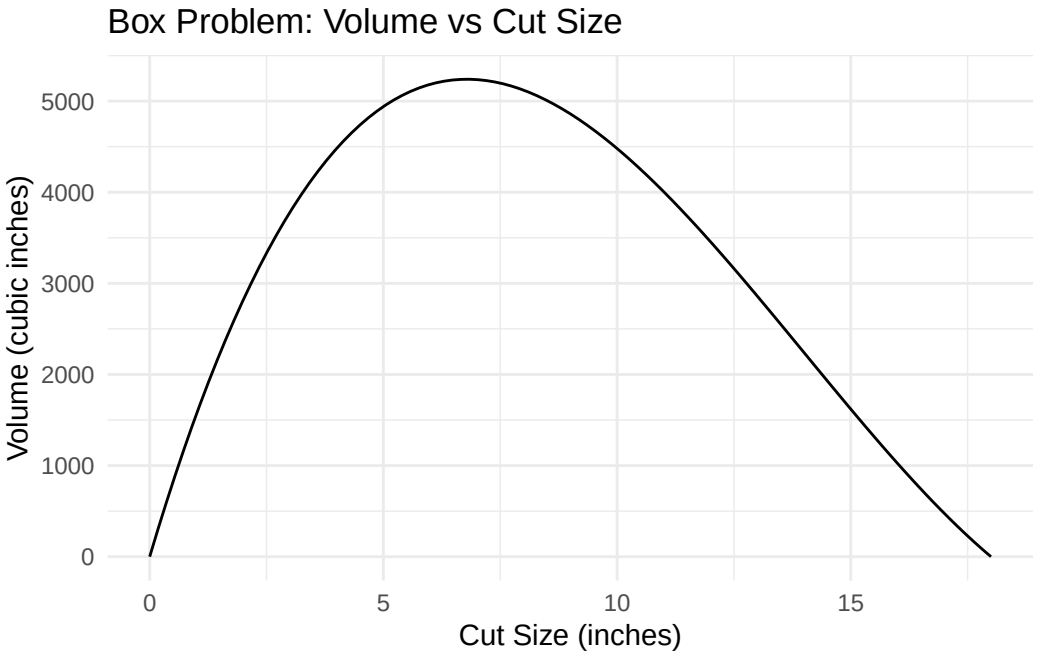
The plot below shows how often each name was given each year. This makes it easy to compare long-term patterns—such as how Liam and Noah have become extremely popular in recent years, while Ava and Sophia also experienced sharp upward trends during the 2000s.



Plotting a Mathematical Function (Activity #04)

In this section, I returned to the Box Problem and used the new dimensions of a 36 inch by 48 inch sheet of paper. The function below represents the volume of an open-top box formed by cutting out a square of side length (x) from each corner and folding up the sides.

The graph helps visualize how the volume changes as the cut size increases. At first, the volume rises as the box gets deeper, but after a certain point the cutouts become too large, reducing the base area and causing the volume to decrease again. This creates a clear peak in the graph where the volume is maximized.



What We Learn from the Graph

From the curve, it is clear that the volume reaches a maximum at a specific cut size between 1 and 2 inches. This cut size produces the tallest possible box that also maintains a wide enough base to maximize volume. The shape of the graph reflects the trade-off between depth and base size, and the peak marks the optimal balance for the 36×48 sheet.

What I Feel I've Learned So Far

Throughout the course so far, I feel like I've learned how to approach data problems in a much more structured and confident way. I've gotten better at cleaning messy datasets, reshaping data, and understanding why each wrangling step matters. I've also learned how to create clear visualizations that actually help explain the data rather than just display it.

Another big thing I've learned is how to write reproducible code. Before this class, I didn't think much about whether someone else could run my code, but now I'm more aware of naming things clearly, avoiding manual fixes, and organizing my work so that it makes sense to others. Overall, I feel like I'm becoming more confident working with data and more thoughtful about how I communicate the results.

Code Appendix

Armed Forces Data Wrangling Code (Activities #08 and #10)

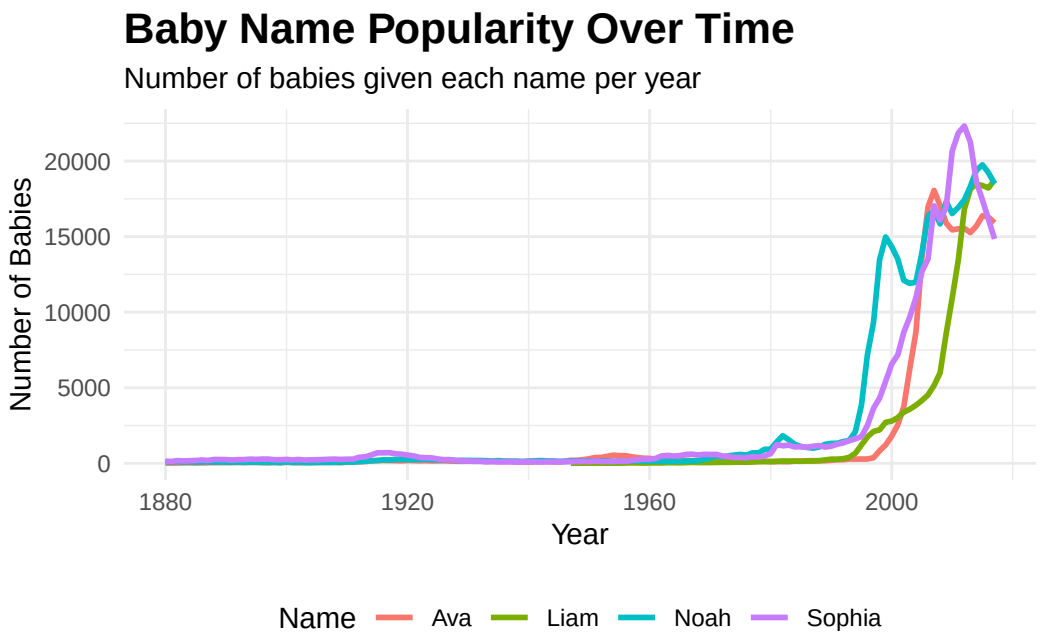
```
# Armed Forces Data Wrangling

library(tidyverse)
library(janitor)

# Load raw file
raw <- read_csv("US_Armed_Forces_(6_2025) - Sheet1.csv",
                col_names = FALSE,
                show_col_types = FALSE)

# Remove first row
raw2 <- raw[-1, ]

# Extract branch and sex headers
branch_row <- raw2[1, ] |> unlist() |> as.character()
sex_row    <- raw2[2, ] |> unlist() |> as.character()

# Combine headers into branch_sex names
combined_headers <- paste(branch_row, sex_row, sep = "_")

# Clean names
cleaned_names <- make_clean_names(combined_headers)

# The first column is the pay grade
cleaned_names[1] <- "pay_grade"

# Apply names and drop header rows
dat <- raw2[-c(1, 2), ]
names(dat) <- cleaned_names

# Keep only male/female columns (drop totals)
dat2 <- dat %>% select(pay_grade, matches("male$|female$"))

# Reshape wide → long
long <- dat2 %>%
  pivot_longer(
    cols = -pay_grade,
    names_to = c("branch", "sex"),
    names_pattern = "^(.*)_(male|female)$",
    values_to = "count"
  ) %>%
  mutate(
```

```
    branch = str_to_title(str_replace_all(branch, "_", " ")),
    sex    = str_to_title(sex),
    count = parse_number(count, na = c("N/A", "N/A*", "NA*", "NA")),
  )

# Grouped dataset used for table and analysis
df_grouped <- long %>%
  arrange(branch, sex, pay_grade)

# Army Officers frequency table
army_table <- df_grouped %>%
  filter(branch == "Army", str_starts(pay_grade, "O")) %>%
  pivot_wider(
    names_from = pay_grade,
    values_from = count,
    values_fill = 0
  )
```

Popularity of Baby Names (Activity #13)

```
# Baby Names Visualization

library(babynames)
library(dplyr)
library(ggplot2)

# Names selected for analysis
chosen_names <- c("Ava", "Liam", "Sophia", "Noah")

# Filter and summarize
names_df <- babynames %>%
  filter(name %in% chosen_names) %>%
  group_by(name, year) %>%
  summarize(total = sum(n), .groups = "drop")

# Line plot
ggplot(names_df, aes(x = year, y = total, color = name)) +
  geom_line(linewidth = 1) +
  labs(
    title = "Baby Name Popularity Over Time",
    subtitle = "Number of babies given each name per year",
    x = "Year",
    y = "Number of Babies",
```

```
    color = "Name"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    legend.position = "bottom"
  )
```



Box Problem Function + Plot (Activity #04)

```
# Box Problem for a 36in x 48in sheet

library(ggplot2)

# Volume function: x = cut size
V <- function(x) {
  x * (48 - 2*x) * (36 - 2*x)
}

# Plot of the volume function
ggplot(data.frame(x = c(0, 18)), aes(x)) +
  stat_function(fun = V) +
  labs(
    title = "Box Problem: Volume vs Cut Size",
    x = "Cut Size (inches)",
```

```
    y = "Volume (cubic inches)"
  ) +
  theme_minimal()
```