Activity #14 — A First QMD File

Ritvik Kothapalyam

2025-11-11

Armed Forces Data Wrangling Redux

Introduction: This document presents a comprehensive analysis of data wrangling and visualization techniques applied to United States Armed Forces personnel data. The analysis examines the relationship between gender and rank distribution across military branches. Through systematic data preparation and transformation, the dataset is restructured such that each observation represents an individual service member with associated rank information, thereby facilitating statistical analysis and visualization.

Table 1: Frequency Table of Sex and Rank in the US Army

sex	Enlisted	Officers	Warrant Officers
female	55,627	16,517	1,678
male	299,692	60,345	14,417

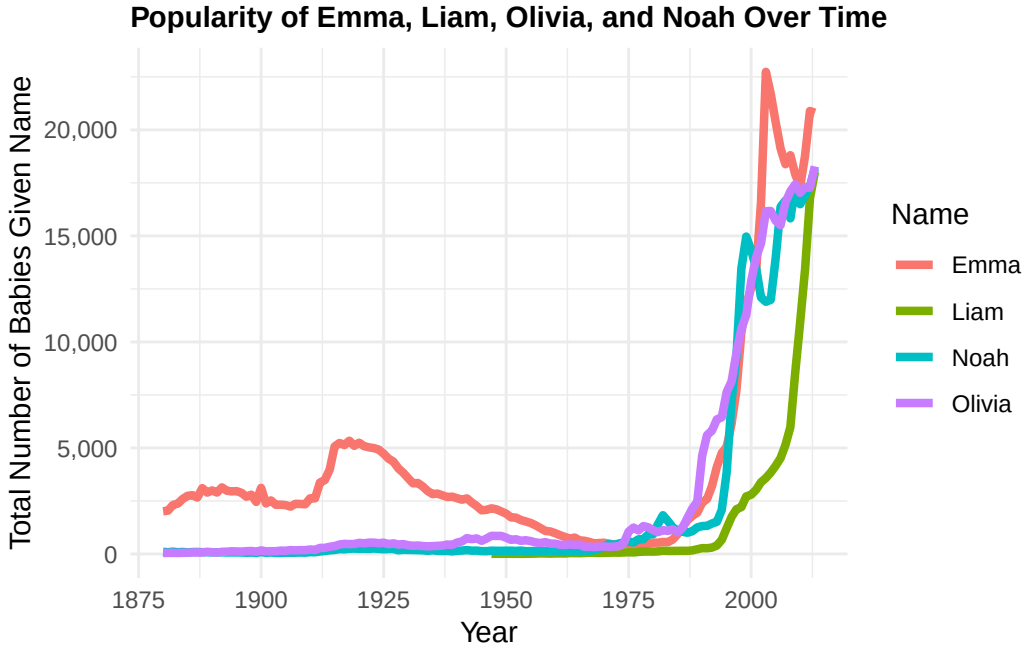
Narrative Text - Armed Forces Data:

The frequency table presented above illustrates the distribution of military personnel by sex and rank within the US Army. The data encompasses three distinct rank categories: Enlisted, Warrant Officers, and Officers, disaggregated by sex (female and male). Several noteworthy patterns emerge from this analysis. Enlisted personnel constitute the majority of the force structure, comprising 299,692 males and 55,627 females. This distribution aligns with conventional military hierarchies, wherein enlisted ranks form the foundational layer of personnel composition. The Warrant Officers category represents the smallest cohort, with 14,417 males and 1,678 females, reflecting the highly specialized and technical expertise required for these positions. The Officers category includes 60,345 males and 16,517 females, representing the command and leadership echelon of the Army. Examination of the relationship between sex and rank reveals that these variables are not statistically independent within the US Army. Female representation varies significantly across rank categories: approximately 15.7% of enlisted

personnel, 10.4% of Warrant Officers, and 21.5% of Officers. The proportional representation of females demonstrates considerable variation across ranks, with the highest concentration among Officers and the lowest among Warrant Officers. This differential distribution across rank categories suggests a substantive association between sex and rank within the Army's organizational structure, warranting further investigation into the underlying factors contributing to these patterns.

Popularity of Baby Names

Visualization - This analysis examines the temporal popularity trends of four contemporary names: Emma, Liam, Olivia, and Noah. These names were selected based on their prominence in recent naming conventions, ranking among the most frequently chosen baby names in the United States. The objective of this investigation is to determine whether their current popularity represents a modern phenomenon or reflects historical naming patterns with deeper longitudinal trends.



Narrative Text - The visualization reveals distinct patterns in the popularity trajectories of these four names. Throughout most of the 20th century, all four names maintained relatively low usage rates; however, they experienced substantial increases beginning in the 1990s and 2000s. Emma and Olivia demonstrated particularly pronounced growth trajectories. Notably, Emma's annual frequency increased from fewer than 5,000 instances in 1980 to over 20,000 by 2014. These trends indicate that the contemporary popularity of these names represents a

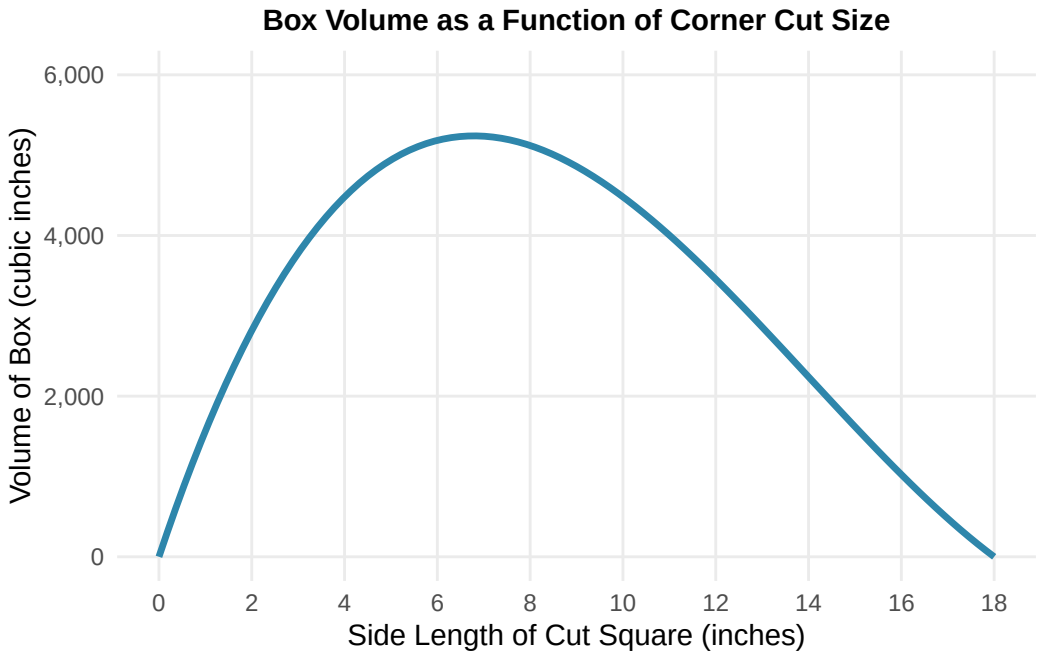
recent phenomenon rather than enduring historical preferences.The observed sharp increases can be attributed to multiple sociocultural factors, including popular culture influences, celebrity naming choices, and evolving preferences toward names that convey traditional aesthetics while retaining relative novelty. The visualization further demonstrates that naming preferences can undergo rapid transformation within a single generation, with these four names transitioning from relative obscurity to widespread adoption within two to three decades.The methodological decision to aggregate counts across gender categories was deliberate and analytically justified. This approach provides a more comprehensive assessment of overall name popularity independent of gender association. While certain names—such as Liam and Noah—are predominantly masculine, and others—such as Emma and Olivia—are primarily feminine, the aggregation of counts enables a more accurate representation of each name's total cultural prevalence and societal impact.

Plotting a Mathematical Function

Mathematical Optimization: The Box Maximization Problem This analysis addresses a classical optimization problem involving the construction of an open-top box from a rectangular sheet of material. The procedure entails removing congruent squares from each corner of the sheet and folding the resulting flaps upward to form the box sides. For this investigation, the initial rectangular sheet measures 36 inches by 48 inches. When squares with side length x are excised from each corner, the resulting box dimensions are defined as follows:

Height: x inches Length: (48 - 2x) inches Width: (36 - 2x) inches

The volume function of the box is expressed as: $V(x) = x(48 - 2x)(36 - 2x)$ The objective of this analysis is to determine the optimal value of x—the side length of the removed squares—that maximizes the volume of the constructed box.



Analysis and Results: The visualization demonstrates several critical insights regarding this optimization problem. The volume function generates a smooth, continuous curve that originates at zero (when x = 0, no box is formed), ascends to a global maximum, and subsequently decreases to zero (when x = 18 inches, the cuts converge at the center, precluding box formation).The maximum volume occurs at approximately x = 6 inches. At this optimal cutting dimension, the box parameters are as follows: Height: 6 inches Length: 48 - 2(6) = 36 inches Width: 36 - 2(6) = 24 inches Maximum Volume: approximately 5,184 cubic inches This result reflects fundamental optimization principles: excessively small cuts produce a shallow box with minimal volume, while excessively large cuts consume disproportionate material, yielding inadequate dimensions for the remaining box structure. The optimal solution achieves equilibrium among all three dimensions, thereby maximizing the enclosed volumetric space.The parabolic symmetry of the curve illustrates the inherent mathematical properties of the cubic volume function. The practical constraints of the problem—where x must satisfy $0 < x < 18$ inches—are clearly delineated in the graphical representation, as the volume approaches zero at both boundary conditions where physical box construction becomes infeasible.

What I have learned

Throughout this course, I have developed proficiency in computational data analysis, statistical programming, and the R programming language. I have acquired comprehensive skills in utilizing the tidyverse ecosystem for data wrangling and transformation, generating sophisticated visualizations through ggplot2, and producing reproducible code that executes seamlessly across different computing environments. Additionally, I have cultivated the ability to integrate clean,

well-documented code with rigorous statistical analysis and meaningful visual insights.The course activities have reinforced my understanding of fundamental principles in data science, particularly the critical importance of organizational structure, comprehensive code annotation, and reproducibility. These foundational practices are essential for ensuring that analytical work is transparent, verifiable, and accessible to the broader scientific community.

Code Appendix

Armed Forces Data

```
# Armed Forces Data:
library(tidyverse)

# Create Army data frame with totals by rank and sex
# Each row represents aggregate counts for a rank-sex combination
army_data <- data.frame(
  branch = "Army",
  rank = c(rep(x = "Enlisted", times = 2),
           rep(x = "Warrant Officers", times = 2),
           rep(x = "Officers", times = 2)),
  sex = rep(x = c("male", "female"), times = 3),
  count = c(299692, 55627, 14417, 1678, 60345, 16517)
)

# Create Navy data frame with totals by rank and sex
navy_data <- data.frame(
  branch = "Navy",
  rank = c(rep(x = "Enlisted", times = 2),
           rep(x = "Warrant Officers", times = 2),
           rep(x = "Officers", times = 2)),
  sex = rep(x = c("male", "female"), times = 3),
  count = c(217304, 59100, 1930, 257, 41876, 12234)
)

# Create Marine Corps data frame with totals by rank and sex
marines_data <- data.frame(
  branch = "Marine Corps",
  rank = c(rep(x = "Enlisted", times = 2),
           rep(x = "Warrant Officers", times = 2),
           rep(x = "Officers", times = 2)),
```

```
  sex = rep(x = c("male", "female"), times = 3),
  count = c(133858, 14795, 2106, 144, 17166, 2132)
)

# Combine all three branch data frames into one
armed_forces <- rbind(army_data, navy_data, marines_data)

# Convert from aggregate to individual-level data
# This expands each row to have 'count' number of individual observations
# The result is a data frame where each row represents one soldier
armed_forces_long <- armed_forces[rep(x = seq_len(nrow(armed_forces)),
                                      times = armed_forces$count), ]

# Visualization: for the Armed Forces For this analysis, I chose to examine the relationship

# Filter the data to include only Army observations
army_subset <- subset(x = armed_forces_long, subset = branch == "Army")

# Create a two-way frequency table of sex and rank
army_freq <- table(army_subset$sex, army_subset$rank)

# Convert the table to a data frame for better formatting
army_table <- as.data.frame.matrix(x = army_freq)

# Add sex as a column instead of row names
army_table <- cbind(sex = rownames(army_table), army_table)
rownames(army_table) <- NULL

# Display the table with formatted numbers and caption
knitr::kable(x = army_table,
             format.args = list(big.mark = ","),
             align = c("l", "r", "r", "r"),
             caption = "Frequency Table of Sex and Rank in the US Army")
```

Table 2: Frequency Table of Sex and Rank in the US Army

sex	Enlisted	Officers	Warrant Officers
female	55,627	16,517	1,678
male	299,692	60,345	14,417

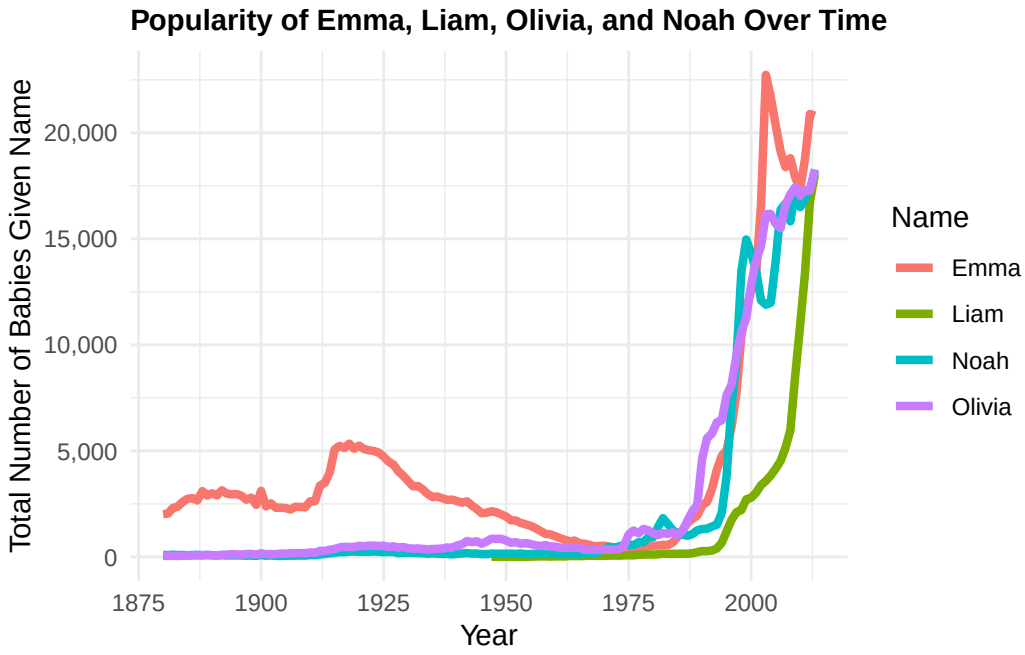
Popularity of Baby Names

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

# Load the BabyNames dataset
data(BabyNames)

# Filter for my selected names: Emma, Liam, Olivia, and Noah
# Group by name and year, then add up counts for males and females
# This combines both sexes since these names can be given to any baby
selected_names <- BabyNames %>%
  filter(name %in% c("Emma", "Liam", "Olivia", "Noah")) %>%
  group_by(name, year) %>%
  summarize(total_count = sum(count), .groups = "drop")

# Create line plot showing how name popularity changed over time
ggplot(selected_names, aes(x = year, y = total_count, color = name)) +
  geom_line(linewidth = 1.5) +
  labs(
    title = "Popularity of Emma, Liam, Olivia, and Noah Over Time",
    x = "Year",
    y = "Total Number of Babies Given Name",
    color = "Name",
    alt = "Line graph showing the popularity of four baby names from 1880 to 2014, with each
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 11, face = "bold"),
    axis.title = element_text(size = 11),
    legend.title = element_text(size = 11),
    legend.position = "right"
  ) +
  scale_y_continuous(labels = scales::comma)
```



Plotting Math Function

```
# Load required package
library(ggplot2)

# Define the volume function for the box problem
# Paper dimensions: 36 inches by 48 inches
# x = side length of square cut from each corner
box_volume <- function(x) {
  length <- 48 - 2*x  # Length after cutting and folding
  width <- 36 - 2*x   # Width after cutting and folding
  height <- x         # Height equals the cut size
  volume <- length * width * height
  return(volume)
}

#| label: fig-box-volume
#| fig-cap: "Volume of the open-top box as a function of the side length of squares cut from
#| fig-width: 8
#| fig-height: 5
```

```r
# Create the plot of the volume function using stat_function
ggplot(data = data.frame(x = c(0, 18)), aes(x = x)) +
  stat_function(
    fun = box_volume,
    linewidth = 1.2,
    color = "#2E86AB"
  ) +
  labs(
    title = "Box Volume as a Function of Corner Cut Size",
    x = "Side Length of Cut Square (inches)",
    y = "Volume of Box (cubic inches)",
    alt = "A curve showing box volume on the y-axis versus cut size on the x-axis. The curve
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 11, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 11),
    panel.grid.minor = element_blank()
  ) +
  scale_y_continuous(labels = scales::comma, limits = c(0, 6000)) +
  scale_x_continuous(breaks = seq(0, 18, 2))
```