Data Analysis Project: Armed Forces, Baby Names, and Optimization

Rupin R ReddyReddy

2025-11-01

Armed Forces Data Wrangling Redux

Table 1: Distribution of Army Officers by Sex and Rank

sex	O1	O2	O3	O4	O5	O6	O7	O8	O9
Female	2400	3006	6053	3044	1531	452	18	8	5
Male	7122	9650	20986	12350	6939	3161	100	80	46

Looking at Table 1, you can see how Army Officers are distributed across different ranks by sex. Male officers way outnumber female officers at every rank level, especially at the lower ranks like O1 and O2. If sex and rank were independent, we'd expect to see similar proportions of males to females at each rank, but that's not what's happening here. At O1, females are about 25% of the total (2,400 out of 9,522), but by O6 they're only 13% (452 out of 3,613). So the proportion of female officers actually decreases as rank goes up, which shows that sex and rank aren't independent - there's definitely a relationship between them.

Popularity of Baby Names

Figure 1 shows how popular the names Emma and Liam have been over time. I picked these two because they're super popular now but have really different histories. Emma's been pretty consistently used since 1880, with a big spike in the early 1900s and then another huge spike starting around 2000. Liam, though, was basically nonexistent until the 90s and then just exploded in popularity. I used a line plot because it's the best way to show trends over time, and I went with different line styles (solid vs dashed) instead of colors so it would be easier to read for people who are colorblind. It's pretty cool how Emma maintained popularity for over a century while Liam went from nowhere to one of the top names in just like 25 years.

Figure 1: Figure 1: Popularity of the Names Emma and Liam Over Time (1880-2015)

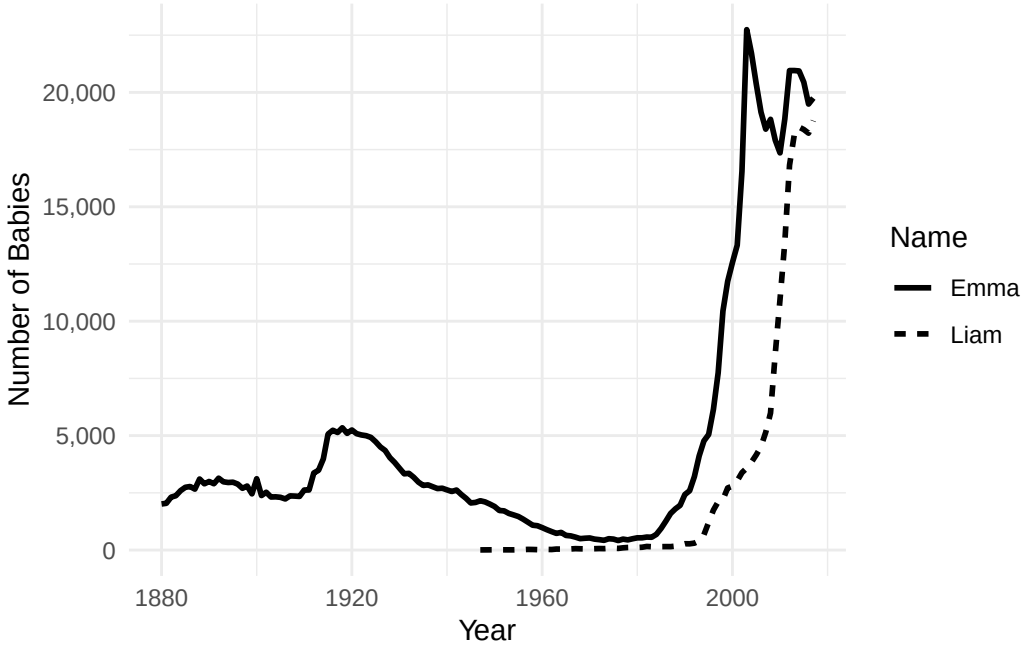

Figure 2: Figure 2: Volume of Box as a Function of Corner Cut Size (36×48 inch paper)

Plotting a Mathematical Function (Box Problem)

Figure 2 shows the volume of the box based on how big of a square we cut from each corner. The maximum volume happens at x = 4.85 inches, where you get 7,033 cubic inches. The graph makes sense when you think about it - if you don't cut much, the box is super shallow and doesn't hold much. But if you cut too much, the base gets tiny and the volume drops again. At x = 0 there's no box, and at x = 18 the whole thing collapses because you've cut away the entire width. The curve peaks right in the middle where there's a balance between height and base size, which is exactly what you'd expect from the math.

What You Feel You Have Learned So Far

Throughout this course, I've learned a lot about working with actual data and getting it cleaned up properly using dplyr and other tidyverse tools. Before taking this class, I honestly didn't realize how much work goes into just preparing data before you can even start analyzing it. I've also gotten way better at making visualizations with ggplot2, and now I actually think about things like making sure my plots are readable for everyone, including people who might be colorblind. Probably the most important thing I've learned is to think more carefully about statistical concepts like independence between variables - now instead of just looking at numbers, I try to understand what they actually mean and whether the patterns I'm seeing are real or just random.

Appendix: Code

Armed Forces Data Wrangling

```
library(tidyverse)

# This is the data from the DOD spreadsheet we used in class
# Just focusing on Army officers (O1 through O9)
armed_forces_data = tibble(
  rank = c("O1", "O2", "O3", "O4", "O5", "O6", "O7", "O8", "O9"),
  Male = c(7122, 9650, 20986, 12350, 6939, 3161, 100, 80, 46),
  Female = c(2400, 3006, 6053, 3044, 1531, 452, 18, 8, 5)
)

# pivot_longer changes it from wide to long format
# then uncount expands it so each person is a row
armed_forces_filtered = armed_forces_data |>
  pivot_longer(cols = c(Male, Female), names_to = "sex", values_to = "count") |>
  uncount(count)

# now make the frequency table
# count gives me the frequencies, then pivot_wider makes it look like a normal table
freq_table = armed_forces_filtered |>
  count(sex, rank) |>
  pivot_wider(names_from = rank, values_from = n, values_fill = 0)

knitr::kable(freq_table)
```

Baby Names Visualization

```
library(babynames)

# filter for Emma and Liam
# group_by and summarize adds up the male and female counts for each name
baby_data = babynames |>
  filter(name %in% c("Emma", "Liam")) |>
  group_by(year, name) |>
  summarize(total = sum(n), .groups = "drop")

# make the time series plot
# using linetype = name so it's accessible for colorblind people
ggplot(baby_data, aes(x = year, y = total, linetype = name)) +
  geom_line(linewidth = 1) +
  labs(x = "Year", y = "Number of Babies", linetype = "Name") +
```

```
  theme_minimal() +
  scale_y_continuous(labels = scales::comma)
```

Box Problem Function and Plot

```
# define the volume function
# x is the size of the cut, then the dimensions are 36-2x and 48-2x
box_volume = function(x) {
  x * (36 - 2*x) * (48 - 2*x)
}

# plot using stat_function
# xlim goes from 0 to 18 because you can't cut more than half of 36
ggplot(data = NULL) +
  stat_function(fun = box_volume, xlim = c(0, 18), linewidth = 1) +
  labs(x = "Size of Cut (inches)", y = "Volume (cubic inches)") +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma)
```