Activity 14 - Visualizations and Data Wrangling Redux

Tyler Gilbert

2025-11-21

Table of contents

0.1 Armed Forces Data Wrangling

Using the data from the previous activities, I created the table requested where every case is a soldier. Then as described made it ordered by rank. Then I created the frequency table to show progress from when we first worked with this data.

```
# A tibble: 1 x 2
  service total_personnel
  <chr>             <dbl>
1 <NA>            5112648
```

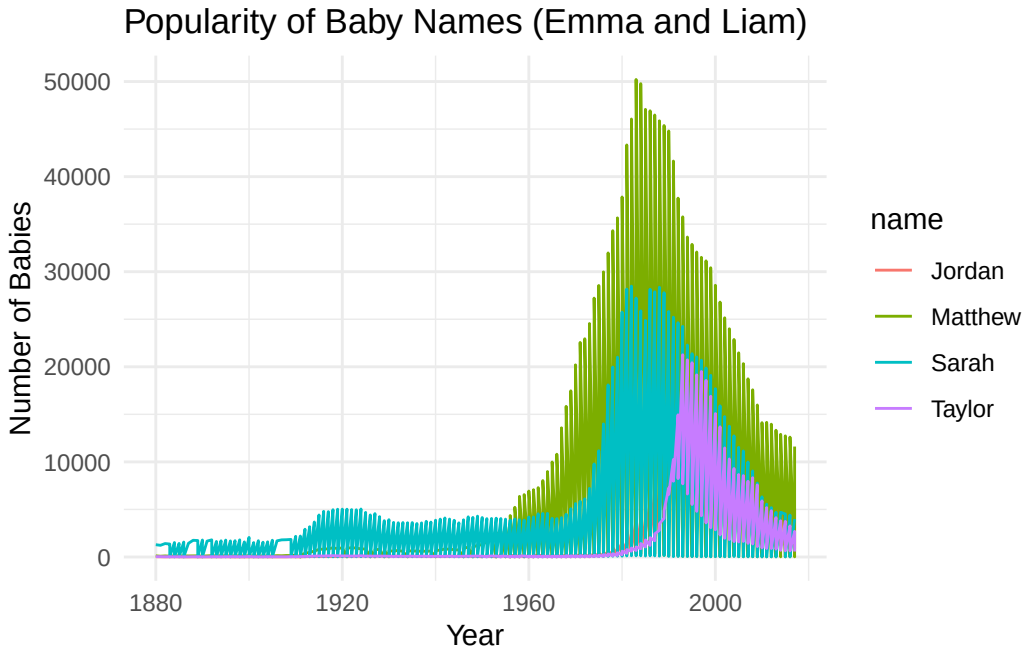
```
$maximum
[1] 5.000008

$objective
[1] 2000
```

```
# A tibble: 432 x 4
   pay_grade service gender count
   <chr>     <chr>   <chr>  <dbl>
 1 E1        <NA>    <NA>    7429
 2 E1        <NA>    <NA>    1326
 3 E1        <NA>    <NA>    8755
 4 E1        <NA>    <NA>    8903
 5 E1        <NA>    <NA>    3434
 6 E1        <NA>    <NA>   12337
 7 E1        <NA>    <NA>    7849
 8 E1        <NA>    <NA>     655
 9 E1        <NA>    <NA>    8504
10 E1        <NA>    <NA>    8537
# i 422 more rows
```

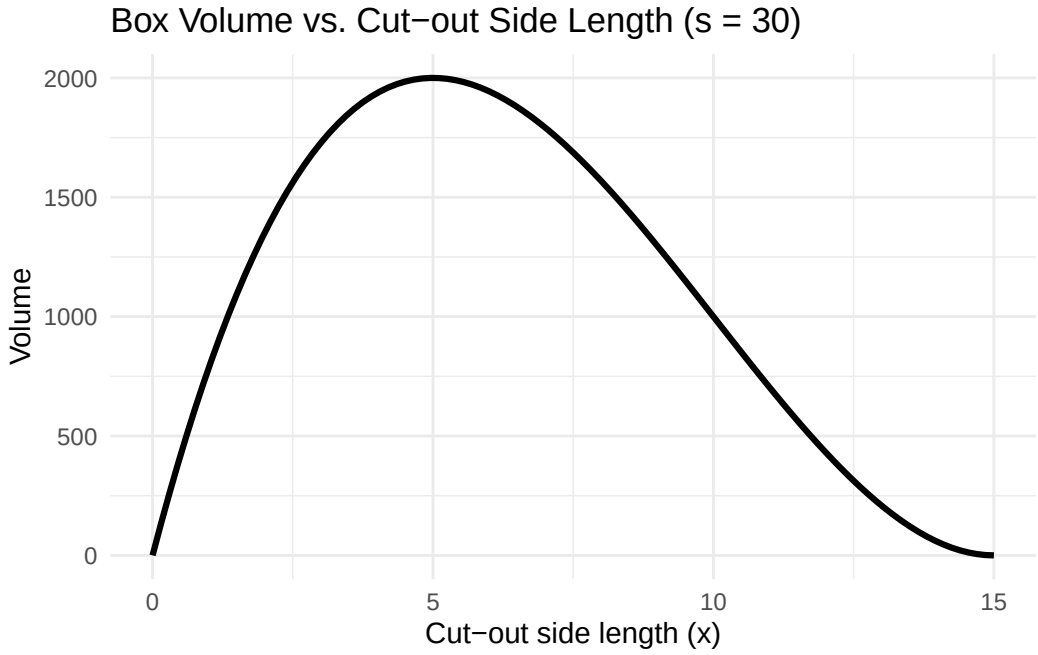
0.2 Popularity of Baby Names

This visualization shows trends in the names Matthew, Sarah, Taylor, and Jordan over time. The names I chose were because I have family or friends with these names, or they have just been names I like and know are fairly common.



0.3 Box Plot Function

The plot below shows the volume of an open-top box and how it changes with the side length of the square cut-out. As the cut size increases, volume first increases and then decreases, showing an optimal point for maximum volume. The data shows a slight skew to the right, given the current side length chosen.



0.4 What I've Learned Thus Far

I feel that I have learned how to think more like a coder. Taking other classes involving coding, none of them have focused on the process of it or shown full in-depth examples of how to plan it out and code what we need to.

```
# ---- Armed Forces Data Wrangling ----

library(googlesheets4)
library(dplyr)
library(tidyr)
library(stringr)
library(readr)
library(ggplot2)
```

```r
# Reading the file
url <- "https://docs.google.com/spreadsheets/d/19xQnI1cBh6Jkw7eP8YQuuicMlVDF7Gr-nXCb5qbwb_E/e

armed_raw <- read_sheet(url, col_types = "c")

#Clean the names of the data
armed_raw <- janitor::clean_names(armed_raw)


# Rename the first column
colnames(armed_raw)[1] <- "pay_grade"

# ---- Reshape (pivot) ----
armedForces_table <- armed_raw %>%
# Only keep the rows with the ranks
  filter(str_detect(pay_grade, "^[EOW][0-9]+")) %>%
  pivot_longer(
    cols = -pay_grade,
    names_to = c("service", "gender"),
    names_pattern = "(.*)_(.*)",
    values_to = "count"
  ) %>%
  mutate(
    count = str_replace_all(count, ",", ""),
    count = as.numeric(count)
  )

freq_by_service <- armedForces_table %>%
  group_by(service) %>%
  summarise(total_personnel = sum(count, na.rm = TRUE)) %>%
  arrange(desc(total_personnel))

freq_by_service
```

```
# A tibble: 1 x 2
  service total_personnel
  <chr>             <dbl>
1 <NA>            5112648
```

```r
# ---- Baby Names Visualization ----
```

```r
library(babynames)

baby_names_plot <- babynames %>%
filter(name %in% c("Matthew", "Sarah", "Taylor", "Jordan")) %>%
ggplot(aes(x = year, y = n, color = name)) +
geom_line() +
labs(title = "Popularity of Baby Names (Emma and Liam)",
x = "Year", y = "Number of Babies") +
theme_minimal()

# ---- Box Problem Plot ----

s <- 30   # sheet side length
volume <- function(x) {
  (s - 2 * x)^2 * x    # use 2 * x, and parentheses around (s - 2*x)
}

# Plot on the valid domain 0 <= x <= s/2
box_plot <- ggplot(data.frame(x = c(0, s/2)), aes(x = x)) +
  stat_function(fun = volume, geom = "line", linewidth = 1.1) +
  labs(title = paste0("Box Volume vs. Cut-out Side Length (s = ", s, ")"),
       x = "Cut-out side length (x)",
       y = "Volume") +
  theme_minimal()


opt <- optimize(volume, interval = c(0, s/2), maximum = TRUE)
opt  # opt$maximum is the x that maximizes volume, opt$objective is the max volume
```

```
$maximum
[1] 5.000008

$objective
[1] 2000
```