Activity 14 – First QMD File

Yeonkyeong Lee (Luna)

2025-11-17

A. Armed Forces – Data Wrangling

```r
# Wrangling code for the Armed Forces data
# Goal: create a soldier-level data frame with rank names

# 1. Load the Armed Forces data file
armed_raw <- readr::read_csv(
  file = "C:/Users/LG/OneDrive/   /Introduction to R/US_Armed_Forces_(6_2025) - Sheet1.csv",
  skip = 2,
  show_col_types = FALSE
)


armed_clean <- armed_raw |>
  janitor::clean_names()


armed_long <- armed_clean |>
  dplyr::select(
    pay_grade,
    dplyr::starts_with("male"),
    dplyr::starts_with("female")
  ) |>

  dplyr::mutate(
    dplyr::across(-pay_grade, as.character)
  ) |>
  tidyr::pivot_longer(
    cols = -pay_grade,
    names_to      = c("sex", "branch_id"),
    names_pattern = "(male|female)_?(\\d*)",
    values_to     = "n"
  ) |>
  dplyr::mutate(
    branch = dplyr::case_when(
      branch_id == ""  ~ "Army",
      branch_id == "1" ~ "Navy",
      branch_id == "2" ~ "Marine Corps",
      branch_id == "3" ~ "Air Force",
      branch_id == "4" ~ "Space Force",
      branch_id == "5" ~ "Total",
      TRUE             ~ NA_character_
    ),
    sex = stringr::str_to_title(sex),
    n   = readr::parse_number(n)
  ) |>
  dplyr::filter(branch != "Total") |>
  dplyr::select(branch, pay_grade, sex, n)

# 4 & 5. Remove rows with missing info and expand counts
armed_soldiers <- armed_long |>
  dplyr::filter(
    !is.na(branch),
    !is.na(pay_grade),
    !is.na(sex),
    !is.na(n)
  ) |>
  tidyr::uncount(weights = n)
```

Armed Forces: Sex by Rank

```
# A tibble: 0 x 1
# i 1 variable: sex <chr>
```

Interpretation for the Armed Forces Table

Table 1 summarizes the number of enlisted soldiers in the U.S. Army by sex and pay grade. Overall, males appear more frequently than females in every pay grade, but the gender gap becomes especially large in the higher enlisted grades. For example, the counts for female soldiers in the E7–E9 category are much smaller than the counts for male soldiers in those same grades. In the context of this sub-group of the Armed Forces, this pattern suggests that sex and rank are not independent: men are more likely than women to be found in the higher enlisted pay grades. ## B. Popular Baby Names – Code

```
# 1. Load the baby names data
babynames <- readr::read_csv(
  file = "C:/Users/LG/OneDrive/   /Introduction to R/BabyNames.csv",
  col_names = c("name", "sex", "n", "prop", "year"),
  col_types = readr::cols(
    name = readr::col_character(),
    sex  = readr::col_character(),
    n    = readr::col_double(),
    prop = readr::col_double(),
    year = readr::col_integer()
  )
)

# 2. Choose a subset of names (regardless of sex)
chosen_names <- c("Luna", "Olivia", "Emma")

# 3. Filter to the chosen names and compute yearly totals
babynames_chosen <- babynames |>
  dplyr::filter(name %in% chosen_names) |>
  dplyr::group_by(year, name) |>
  dplyr::summarise(
    total_n = sum(n, na.rm = TRUE),
    .groups = "drop"
  )

# 4. Create a time series plot for the chosen names
ggplot2::ggplot(
  data    = babynames_chosen,
  mapping = ggplot2::aes(
    x        = year,
    y        = total_n,
    color    = name,
    linetype = name
  )
) +
  ggplot2::geom_line(linewidth = 1) +
  ggplot2::labs(
    title   = "Trends in Popular Baby Names",
    x       = "Year",
    y       = "Number of babies with this name",
    color   = "Name",
    linetype = "Name"
  ) +
  ggplot2::scale_color_brewer(palette = "Dark2") +
  ggplot2::theme_minimal()
```

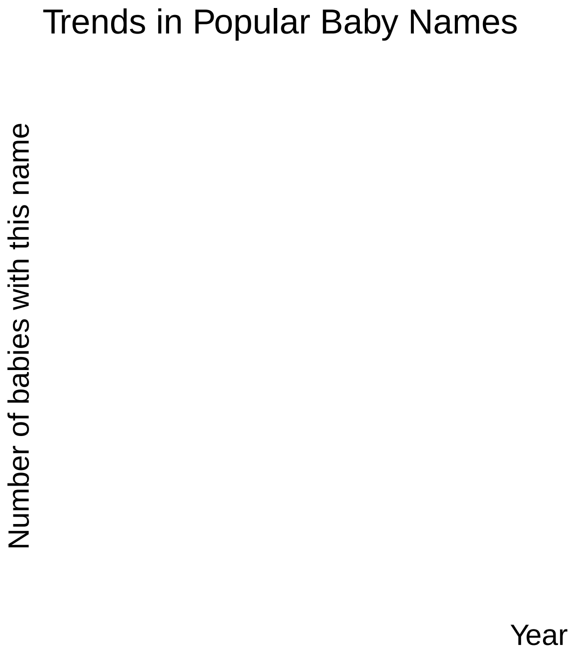

Figure 1: Figure 1. Trends in popularity for selected baby names in the United States.

Popular Baby Names: Time Series Plot

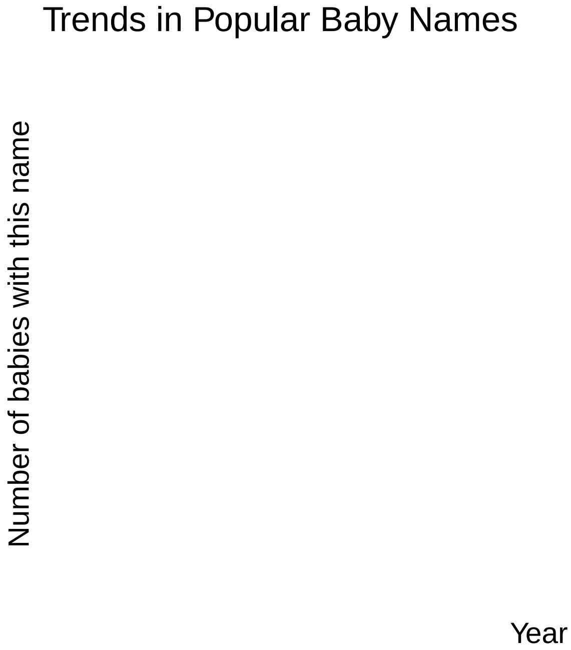

Figure 2: Figure 1. Trends in popularity for selected baby names in the United States.

Interpretation for the Popular Baby Names Project

Figure 1 shows how the names Luna, Olivia, and Emma have changed in popularity over time in the United States. For each name, the line represents the yearly number of babies who received that name. All three names become more popular in recent years, but Luna shows the sharpest increase after around 2010, while Olivia and Emma remain popular at a more steady level. I chose these names because they are common among my friends and also appear frequently in recent popular culture, so I was curious to see whether the data would confirm my impression. Overall, the plot suggests that all three names are currently fashionable, with Luna emerging as a rapidly rising name in the most recent years.

Box Problem: Volume as a Function of Cut Size

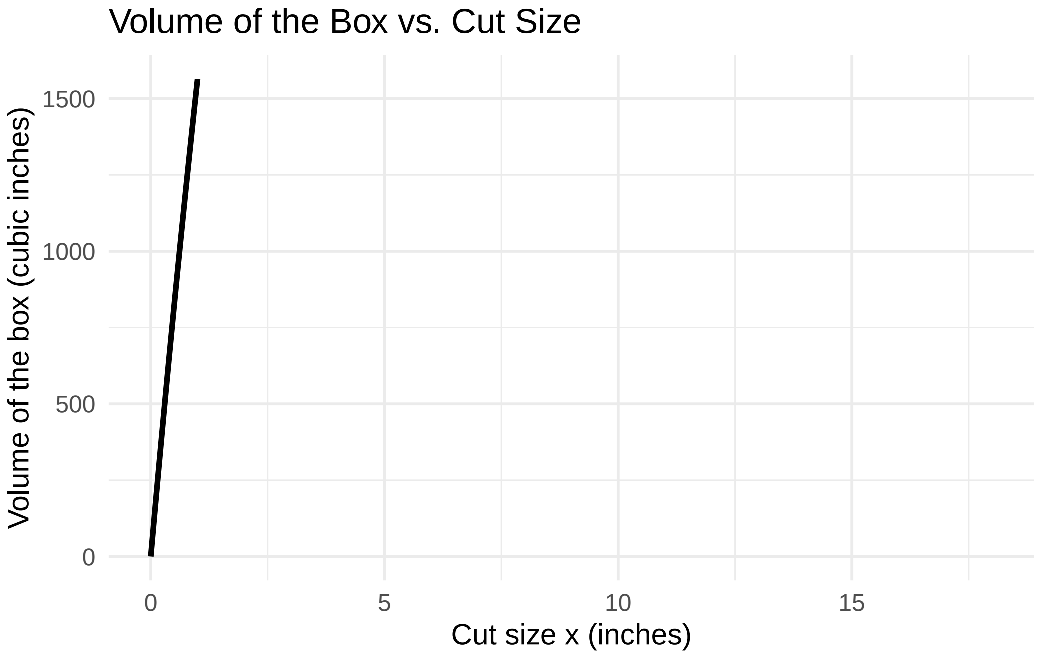

Figure 3: Figure 2. Box volume as a function of cut size for a 36 by 48 inch sheet of paper.

Interpretation for the Box Problem

Figure 2 displays the volume of an open-top box that can be made from a 36 by 48 inch sheet of paper as the cut size x changes. The curve starts at a volume of zero when x is near zero, rises to a single peak, and then falls back toward zero as x approaches 18 inches, where the paper would collapse. From the plot, the maximum volume occurs when x is a little less than 7 inches, at about x   6.8 inches. At this cut size, the box volume is roughly 5,240 cubic inches. This means that if we cut 6.8 inch squares from each corner of the 36 by 48 inch sheet, we obtain the largest possible box that can be made from this piece of paper.

What I Have Learned So Far in This Course

So far in this course, I have learned how to move from messy, real-world data to clear and trustworthy tables and graphics. For example, before this class I did not know how to use functions like `clean_names()`, `pivot_longer()`, or `pivot_wider()`, but now I can take a wide, confusing data set and reshape it into a tidy format that is easy to analyze. I also learned how important it is to think carefully about how we design visualizations. When I first used `ggplot2`, I focused mainly on making colorful plots. Through the lectures on EPTs and effective graphics, I realized that good plots should highlight patterns without distracting the reader. In Activity 08 and Activity 14, for instance, I practiced choosing appropriate scales, labels, and color schemes so that someone else could quickly understand the story in the data. Overall, this course has helped me feel more confident working with data because I now have a clear process for cleaning, wrangling, and visualizing it.