Stat 184, Collection of Learning

Ean Anciso

2025-11-11

1 Armed Forces Data Wrangling Redux

1.1 Visualization

Table 1 is a frequency table focused on the Army branch of the US Armed Forces and displays the distribution of sex within the enlisted ranks. Values are given in both relative and absolute frequency. This table shows how the US Army consists of a majority male. The higher the rank, the more likely it is for there to be less female percentage of that rank.

Table 1: US Armed Forces, Distribution of Sex in Enlisted Army

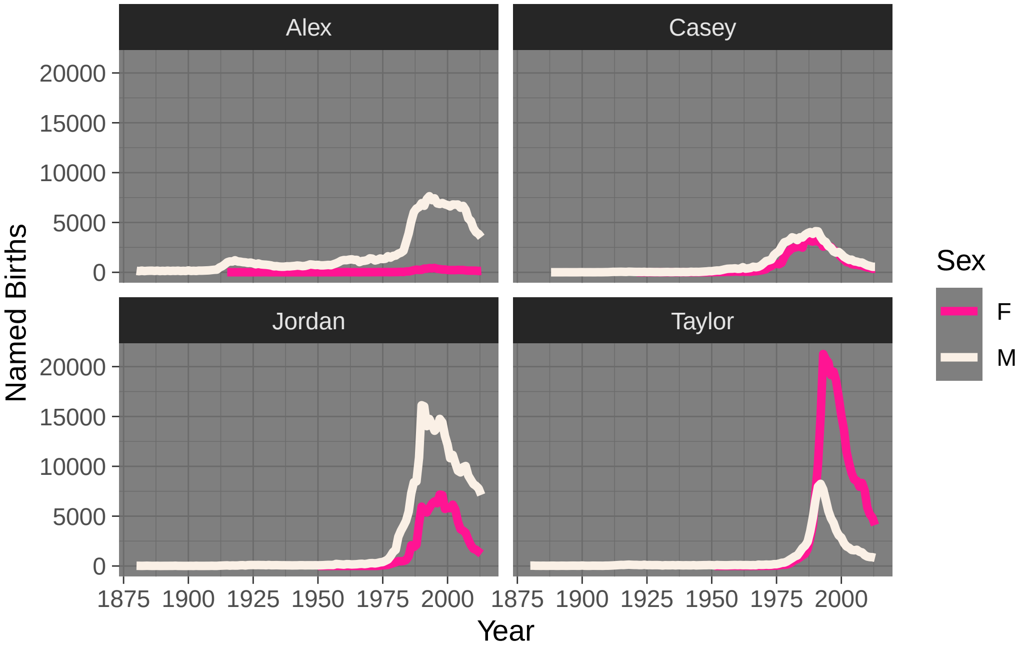
```
 Pay.Grade         Female            Male            Total
        E1  1,326  (0.4%)   7,429  (2.1%)   8,755   (2.5%)
        E2  4,336  (1.2%)  22,338  (6.3%)  26,674   (7.5%)
        E3 10,229  (2.9%)  43,775 (12.3%)  54,004  (15.2%)
        E4 15,143  (4.3%)  79,234 (22.3%)  94,377  (26.6%)
        E5 10,954  (3.1%)  54,803 (15.4%)  65,757  (18.5%)
        E6  7,363  (2.1%)  49,502 (13.9%)  56,865  (16.0%)
        E7  4,410  (1.2%)  30,264  (8.5%)  34,674   (9.8%)
        E8  1,472  (0.4%)   9,482  (2.7%)  10,954   (3.1%)
        E9    394  (0.1%)   2,865  (0.8%)   3,259   (0.9%)
     Total 55,627 (15.7%) 299,692 (84.3%) 355,319 (100.0%)
```

2 Popularity of Baby Names

2.1 Visualization

In Figure 1, four chosen names are plotted for both male and female babies over a period of time. I specifically chose the names Alex, Casey, Jordan, and Taylor because they are all names that are generally accepted for both sexes. This data shows how gender neutral names are effected not only by the time period, but also the sex of the baby.

Figure 1: Popularity of Gender Neutral Baby Names Over Time

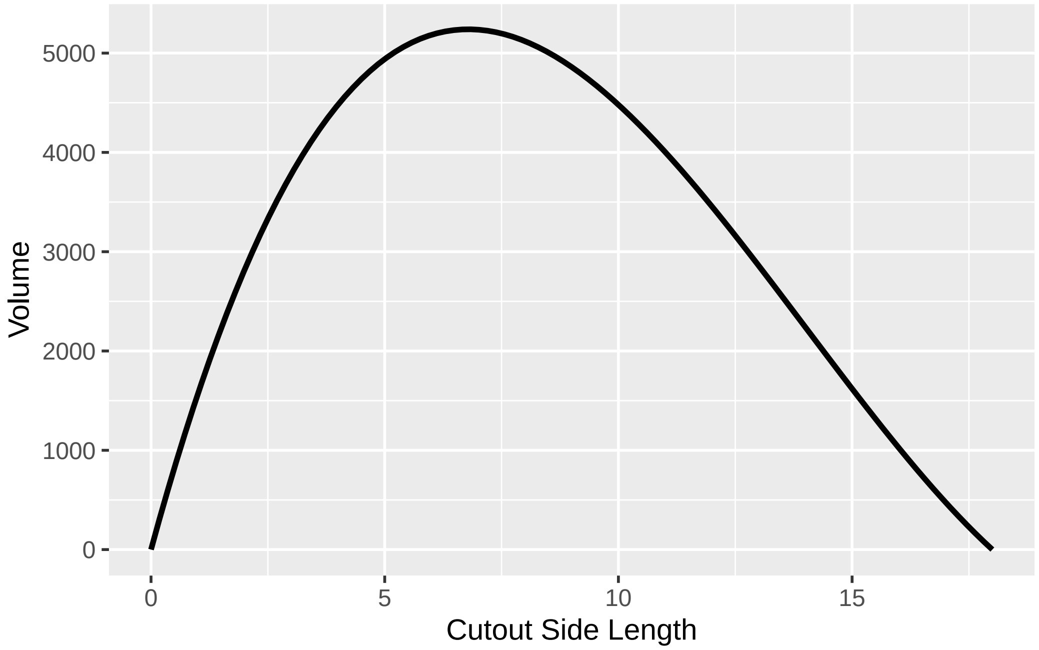


3 Box Problem

3.1 Visualization

Figure 2 deals with the box problem. Specifically, the plot shows the volume function with respect to the length of the cutout side. In the plot, the function peaks right around 6 inches of cutout side length. At this peak the box's volume is just above 5000 cubic inches.

Figure 2: Box Problem, Volume with Respect to Length of Cutout Side



4 What I have Learned

In Stat 184 Fall 2025, I have learned a great amount of methods, techniques, and ideas. In this section I will explore what I have learned. Furthermore, I will organize my exploration by the units in which the class is structured.

4.1 Unit 1

This first unit of the class was where I was introduced to R and R studio. Previously, I was unaware of this language and it's purpose. After becoming familiar with the environment and the types of things I could do with R, we began to learn the proper technique for creating functions. I learned about the importance of planning and identifying the various tools and steps that would lead me to whatever goal. This was put into practice by creating functions for the box problem, Collatz conjecture, and the spinner game.

4.2 Unit 2

The second unit of this class was where I began to really learn about working with data. Specifically, I learned how to obtain data and then structure that data into something workable. I learned about important packages like tidyverse and rvest. I was taught methods of data wrangling and the tools(verbs) I would need. I was able to practice this new knowledge through working with US Armed Forces data and PSU football data.

4.3 Unit 3

The third unit of this class is where I began to understand how to create meaningful visualizations and interpretations of data. I learned about various tables and plotting techniques. Valuable readings from Tufte and Kosslyn provided insight into how data is perceived and how we as creators and consumers can be better. Through this unit I gained an awareness of my audience and how I can employ techniques to serve more people. I learned of important packages like janitor, knitr, kableExtra, and esquisse. I put this newfound knowledge to use in working with diamonds data, busiest-airport data, Palmer penguin data, and more.

4.4 Unit 4

The fourth and final unit of this class. While not finished with this unit, I have already learned valuable information. Largely, I learned about the importance of open science and how I should hold myself to certain standards when creating any work. I gained a familiarity with quarto, similar to tools I have used before, but a great addition still. Most recently, I have gained a familiarity with GitHub. Specifically, I have learned how to create and interact with repositories. Additionally, I learned about proper usage of repos and how I can utilize Git to better my own workflow, as well as to work collaboratively.

# Code Appendix

```r
# Load Packages
library(tidyverse)
library(rvest)
library(googlesheets4)
library(knitr)
library(janitor)
library(dcData)
library(ggplot2)

# Armed Forces Data Wrangling Redux Section

## Data Wrangling and Tidying

### Scrape Data
rank_raw <- read_html("https://neilhatfield.github.io/Stat184_PayGradeRanks.html") %>%
  html_elements(css = "table") %>%
  html_table()

### Extract ranks data frame
rawRanks <- rank_raw[[1]]

### Wrangle Ranks Data
rawRanks[1, 1] <- "Type"
rankHeaders <- rawRanks[1, ] # Extract column Headers
names(rawRanks) <- rankHeaders[1,] # Make headers column names
rawRanks <- rawRanks[-c(1, 26), ] # Discard redundant rows

cleanRanks <- rawRanks %>%
  select(!Type) %>% # Remove extra column
  pivot_longer(
    cols = !`Pay Grade`,
    names_to = "Branch",
    values_to = "Rank"
  ) %>%
  mutate(
    Rank = na_if(x = Rank, y = "--") # Make blank values NA values
  )

### Load Armed Forces Data
gs4_deauth() # Prevent the need to sign in
forcesHeaders <- read_sheet(
  ss = "https://docs.google.com/spreadsheets/d/19xQnI1cBh6Jkw7eP8YQuuicMlVDF7Gr-nXCb5qbwb_E/edi
  col_names = FALSE,
  n_max = 3 # Read first three rows only
)
```

```
rawForces <- read_sheet(
  ss = "https://docs.google.com/spreadsheets/d/19xQnI1cBh6Jkw7eP8YQuuicMlVDF7Gr-nXCb5qbwb_E/edi
  col_names = FALSE,
  skip = 3,
  n_max = 28, # Read all rows before footer
  na = c("N/A*")
)

### Wrangle Armed Forces Data
branchNames <- rep( # Create three copies of each branch
  x = c("Army", "Navy", "Marine Corps", "Air Force", "Space Force", "Total"),
  each = 3
)
combHeaders <- paste( # Combine branch with other headers
  c("", branchNames),
  forcesHeaders[3,],
  sep = "."
)

names(rawForces) <- combHeaders

cleanForces <- rawForces %>%
  rename(Pay.Grade = `.Pay Grade`) %>%
  select(!contains("Total")) %>% # Remove total columns
  filter( # Remove total rows
    !Pay.Grade %in% c(
      "Total Enlisted",
      "Total Warrant Officers",
      "Total Officers",
      "Total"
    )
  ) %>%
  pivot_longer( # Reshape data
    cols = !Pay.Grade,
    names_to = c("Branch", "Sex"),
    names_sep = "\\.",
    values_to = "Frequency"
  )

### Merge Data Frames ----
forcesRanks <- left_join(
  x = cleanForces,
  y = cleanRanks,
  by = join_by(Pay.Grade == `Pay Grade`, Branch == Branch)
)

### Transform Group into Individual ----
```

```r
individualRanks <- forcesRanks %>%
  filter(!is.na(Frequency)) %>% # Remove all cases with missing counts
  uncount(
    weights = Frequency
  )

## Data Visualization

individualRanks %>%
  filter(Branch == "Army", str_starts(Pay.Grade, "E")) %>% # Look at enlisted ranks only
  tabyl(Pay.Grade, Sex) %>%
  adorn_totals("both") %>%
  adorn_percentages("all") %>%
  adorn_pct_formatting(digits = 1) %>%
  adorn_ns(position = "front")

# Baby Names Section

data("BabyNames")

## Chosen Names
names = c("Alex", "Casey", "Jordan", "Taylor") # Gender Neutral Names

## Data Visualization
BabyNames %>% # Filter for chosen names and group by year and sex
  filter(name %in% names) %>%
  group_by(year, sex) %>%
  ggplot(
    aes(
      x = year,
      y = count,
      color = sex
    )
  ) +
  geom_line(size = 1.5) +
  facet_wrap(~name) +
  scale_color_manual( # Contrasting colors for accessibility
    values = c(
      "F" = "deeppink",
      "M" = "linen"
    )
  ) +
  labs(
    x = "Year",
    y = "Named Births",
    color = "Sex",
    alt = "Four graphs show the popularity over time of
    the names Alex, Casey, Jordan, and Taylor by gender"
    # Alt text created using ASU image accessibility tool
  ) +
  theme_dark()
# Box Problem Section

## Box Problem Function
volume_prob <- function(cutoutLength, paperLength = 36, paperWidth = 48) {
  volume = (paperLength - cutoutLength*2)*(paperWidth - cutoutLength*2)*(cutoutLength)
}


## Box Problem Plot

ggplot(
  data.frame(x = c(0, 18)), # Limited by half of smallest side(ie 36 inches)
  aes(
    x = x
  )
) +
  stat_function(fun = volume_prob, size = 1) +
  labs(
    x = "Cutout Side Length",
    y = "Volume",
    alt = "Graph showing volume of a box versus cutout side length,
    with a parabolic curve peaking above 5000 volume units"
    # Alt text created using ASU image accesibility tool
  )
```