Activity 14

Cameryn Lockett

2025-11-21

Optional Installation

#install.packages("remotes")

#remotes::install_github("mdbeckman/dcData")

Armed Forces Data Wrangling Redux

Data Wrangling Code for US Armed Forces Dataset

Visualization for US Armed Forces Dataset

	E1	E2	E3	E4	E5	E6	E7	E8	E9
Female	2681	3603	7493	11855	16254	9580	3098	912	323
Male	9051	10969	23430	39241	57238	45749	18026	6500	2518

Narrative Text for US Armed Forces Dataset

The two-way frequency table shows the number of male and female enlisted personnel in the U.S. Navy across pay grades E1 through E9. Each row represents the total number of men or women serving at each rank level. The table demonstrates that men consistently outnumber women in every pay grade with an increasing gap at higher ranks. Therefore, sex and rank are not independent of each other among Navy enlisted personnel, as the number of female enlisted personnel steadily decreases with increasing rank.

Popularity of Baby Names

Code for the Popular Baby Names Project

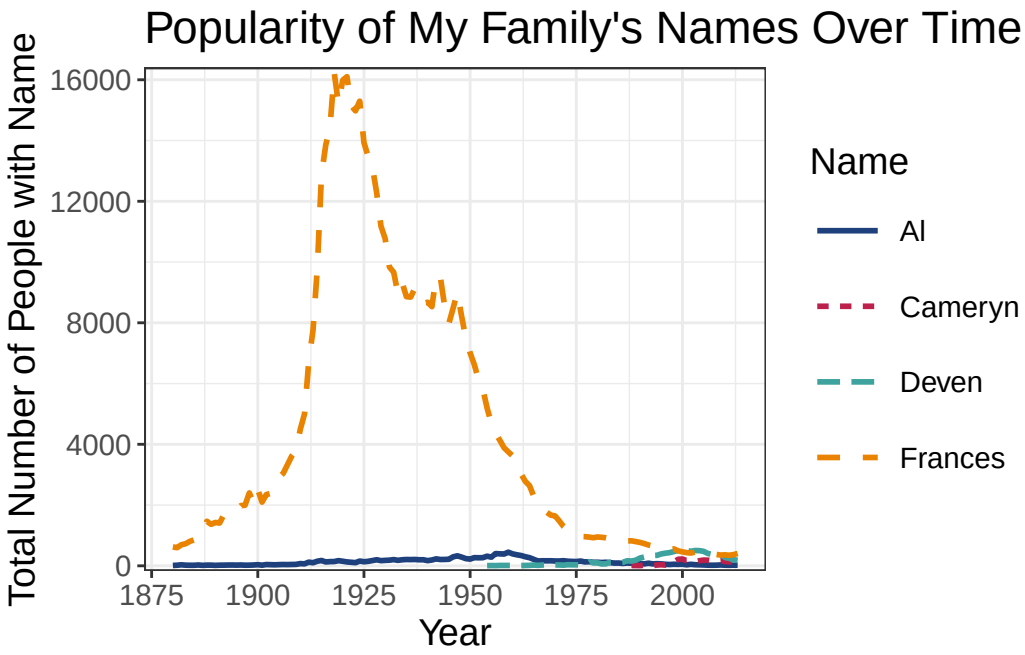

Figure 1: Figure 1. Popularity of selected baby names over time.

Narrative Text for Popular Baby Names Project

The image is a line graph titled "Popularity of My Family's Names Over Time." The x-axis represents the year, ranging from 1875 to 2013, and the y-axis indicates the total number of people with the name, ranging from 0 to 16,000. The graph features four colored lines representing different names: Al (solid blue line), Cameryn (dotted red line), Deven (dashed green line), and Frances (dashed orange line). Frances shows a significant peak around 1925, reaching just over 16,000 before gradually declining. Al, Cameryn, and Deven maintain a relatively stable and low popularity throughout the years. I chose these names because they are a part of my family which makes it more meaninful and interesting to me.

Plotting a Mathematical Function

Code for the Box Problem

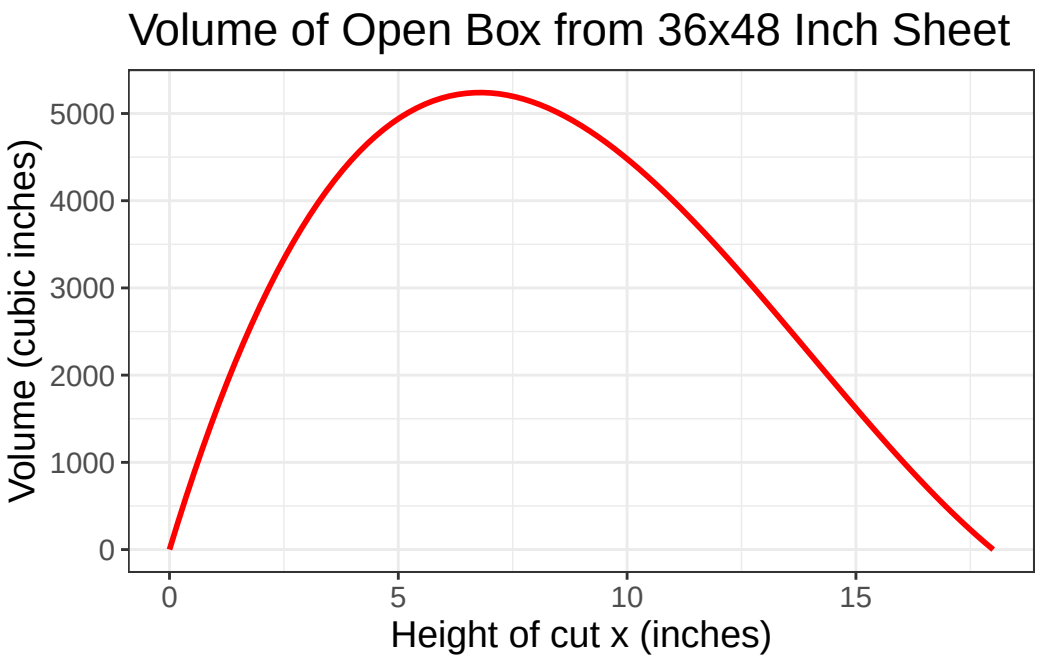

Figure 2: Figure 2. Volume of an open-top box from a 36x48 inch sheet of paper.

Narrative Text for the Box Problem

The figure shows a line plot of the volume of an open-top box as a function of the height of the cut, x, for a 36x48 inch sheet of paper. The x-axis ranges from 0 to 18 inches and represents the height of the square cut at each corner. The y-axis represents the resulting volume of the box in cubic inches. The plot rises sharply as x increases from 0, reaches a clear maximum, and then gradually declines toward 0. The maximum volume of 5,240 cubic inches occurs when the height of the cut is approximately 6.79 inches. The plot clearly demonstrates how box volume changes with cut size.

What I've Learned So Far

Throughout this course, I have learned the essential skills for data wrangling, visualization, and analysis. I have gained experience using ggplot2 to create effective visualization such as line plots and frequency tables. I have learned how to clean and polish my code in order to

create reproducible code containing comments and proper variable names. I have also learned how to properly interpret data visualizations and explain what they reveal about the data.

Code Appendix

==========

1. U.S. Armed Forces

gs4_deauth() # ensures public Google Sheets can be read

Read headers (first 3 rows of the sheet)

forcesHeaders <- read_sheet(

ss = “https://docs.google.com/spreadsheets/d/1cn4i0-ymB1ZytWXCwsJiq6fZ9PhGLUvbMBHlzqG4bwo/edit?

col_names = FALSE, n_max = 3

)

Read main data (rows 4-31, dropping footer rows)

rawForces <- read_sheet(

ss = “https://docs.google.com/spreadsheets/d/1cn4i0-ymB1ZytWXCwsJiq6fZ9PhGLUvbMBHlzqG4bwo/edit?

col_names = FALSE,

skip = 3,

n_max = 28,

col_types = “c”

)

Wrangle Armed Forces Data —-

branchNames <- rep(c(“Army”,“Navy”,“Marine Corps”,“Air Force”,“Space Force”,“Total”), each=3)

tempHeaders <- paste(c(“ “, branchNames), forcesHeaders[3,], sep =””) names(rawForces) <- tempHeaders

Clean and reshape data

cleanForces <- rawForces %>%

```
rename(Pay.Grade = `.Pay Grade`) %>%
dplyr::select(!contains("Total")) %>% # remove total columns
filter( # remove total rows
Pay.Grade != "Total Enlisted",
Pay.Grade != "Total Warrant Officers",
Pay.Grade != "Total Officers",
Pay.Grade != "Total"
) %>%
pivot_longer( # reshape into tidy format
cols = !Pay.Grade,
names_to = "Branch.Sex",
values_to = "Frequency"
) %>%
separate_wider_delim( # separate branch and sex
cols = Branch.Sex,
delim = ".",
names = c("Branch","Sex")
) %>%
mutate(
Frequency = na_if(Frequency, "N/A*"), # convert N/A* to NA
Frequency = parse_number(Frequency) # convert counts to numbers
)
#Case = group of soldiers
forces_group <- cleanForces
#Case = individual soldier
forces_individual <- cleanForces %>%
#Filter to include navy enlisted only
filter(!is.na(Frequency)) %>%
uncount(weights = Frequency)
```

#Create two-way frequency table

navy_enlisted <- forces_individual %>%

filter(Branch == “Navy”, grepl(“^E”, Pay.Grade))

navy_enlisted_table <- table(navy_enlisted$Sex, navy_enlisted$Pay.Grade)

2. Popular Baby Names

#Wrangle the BabyNames

family <- c(“Cameryn”,“Deven”,“Frances”,“Al”)

subsetNames <- BabyNames %>%

filter(name %in% family) %>%

group_by(name, year) %>%

summarize(total = sum(count), .groups = “drop”)

psuPalette <- c(“#1E407C”,“#BC204B”,“#3EA39E”,“#E98300”, “#999999”,“#AC8DCE”,“#F2665E”,“#99C

babyNamesPlot <- ggplot(subsetNames, aes(x=year, y=total, color=name, linetype=name))
+

geom_line(linewidth=1) + #thicker lines

labs(title=“Popularity of My Family’s Names Over Time”,

x=“Year”,

y=“Total Number of People with Name”,

color=“Name”,

linetype=“Name”

)+

scale_y_continuous(expand=expansion(mult=0.01)) + #sets the expansion to 1% of the
range

scale_color_manual(values=psuPalette)+

theme_bw() +

theme(text=element_text(size=14),

legend.key.size=unit(1,“cm”))

3. Box Problem

volume.box <- function(x) { vol_box <- (36 - 2*x*)(48 - 2*x*)(x) }

#Plot the volume as a function of x using ggplot2

ggplot(data = data.frame(x = c(0, 36/2)), aes(x=x)) +

stat_function(fun = volume.box, color = “red”, size = 1) +

labs(

title = “Volume of Open Box from 36x48 Inch Sheet”,

x = “Height of cut x (inches)”, #horizontal axis label

y = “Volume (cubic inches)”) + #vertical axis label

theme_bw() +

theme(text = element_text(size = 14)

)

max_result <- optimize(f = volume.box, interval = c(0, 36/2), maximum = TRUE)

max_x <- max_result$maximum

max_volume <- max_result$objective

max_x

max_volume