Armed Forces Data Wrangling Redux

Tianyu Duan

2025-11-15

Introduction

This document (1) Data Wrangling Code for Armed Forces Data from public sources, (2) Visualization for the Armed Forces and create a two-way frequency table to explore the impact of sex and rank in the US Armed Forces. (3) The narrative text in the Armed Forces section provides a brief explanation of the tables.

Data Wrangling Code for Armed Forces Data

```
library(tidyverse)
library(rvest)
library(googlesheets4)

gs4_deauth()  # use public access; no OAuth prompts

sheet_url <- "https://docs.google.com/spreadsheets/d/19xQnI1cBh6Jkw7eP8YQuuicMlVDF7Gr-nXCb5qbwl
rank_url  <- "https://neilhatfield.github.io/Stat184_PayGradeRanks.html"

ranks_html  <- read_html(x = rank_url)
ranks_list  <- ranks_html %>% html_elements(css = "table") %>% html_table()
raw_ranks   <- ranks_list[[1]]

# clean header row and drop footnotes
raw_ranks[1, 1] <- "Type"
names(raw_ranks) <- raw_ranks[1, ]
raw_ranks <- raw_ranks[-c(1, 26), ]

# tidy to long: Pay Grade / Branch / Rank
ranks <- raw_ranks %>%
  select(!Type) %>%
  pivot_longer(cols      = !`Pay Grade`,
               names_to  = "Branch",
               values_to = "Rank") %>%
  mutate(Rank = na_if(Rank, "--"))    # "--" means that branch doesn't use that pay grade
```

```r
forces_hdr <- read_sheet(ss = sheet_url, col_names = FALSE, n_max = 3)
forces_raw <- read_sheet(ss = sheet_url, col_names = FALSE, skip = 3, n_max = 28, col_types = "

# construct proper column names like "Army.Male", "Army.Female", etc.
branch_names <- rep(c("Army","Navy","Marine Corps","Air Force","Space Force","Total"), each = 3
tmp_headers  <- paste(c("", branch_names), forces_hdr[3, ], sep = ".")
names(forces_raw) <- tmp_headers

forces_clean <- forces_raw %>%
  rename(Pay_Grade = `.Pay Grade`) %>%
  select(!contains("Total")) %>%
  filter(!Pay_Grade %in% c("Total Enlisted","Total Warrant Officers","Total Officers","Total"))

forces_long <- forces_clean %>%
  pivot_longer(cols      = !Pay_Grade,
               names_to  = "Branch.Sex",
               values_to = "Frequency") %>%
  separate_wider_delim(cols  = Branch.Sex,
                       delim = ".",
                       names = c("Branch","Sex")) %>%
  mutate(Frequency = na_if(Frequency, "N/A*"),
         Frequency = na_if(Frequency, "--"),
         Frequency = readr::parse_number(Frequency)) %>%
  filter(!is.na(Frequency))

forces_with_rank <- forces_long %>%
  left_join(y = ranks,
            by = join_by(Pay_Grade == `Pay Grade`, Branch == Branch)) %>%
  relocate(Rank, .after = Pay_Grade)

forces_individual <- forces_with_rank %>%
  tidyr::uncount(weights = Frequency) %>%
  select(Branch, Sex, Pay_Grade, Rank)  # keep only the needed individual-level fields

summary_output <- c(
  n_rows_input_long      = nrow(forces_with_rank),
  total_count_expanded   = nrow(forces_individual),
  distinct_pay_grades    = dplyr::n_distinct(forces_individual$Pay_Grade),
  distinct_branches      = dplyr::n_distinct(forces_individual$Branch)
)
print(summary_output)
```

```
   n_rows_input_long total_count_expanded  distinct_pay_grades
                 230              1278162                   24
   distinct_branches
                   5
```

```r
head(forces_with_rank)   # tidy, with counts and rank names
```

```
# A tibble: 6 x 5
  Pay_Grade Rank           Branch       Sex    Frequency
  <chr>     <chr>          <chr>        <chr>      <dbl>
1 E1        Private        Army         Male        7429
2 E1        Private        Army         Female      1326
3 E1        Seaman Recruit Navy         Male        8903
4 E1        Seaman Recruit Navy         Female      3434
5 E1        Private        Marine Corps Male        7849
6 E1        Private        Marine Corps Female       655
```

```r
head(forces_individual)  # individual-level rows required by the rubric
```

```
# A tibble: 6 x 4
  Branch Sex   Pay_Grade Rank
  <chr>  <chr> <chr>     <chr>
1 Army   Male  E1        Private
2 Army   Male  E1        Private
3 Army   Male  E1        Private
4 Army   Male  E1        Private
5 Army   Male  E1        Private
6 Army   Male  E1        Private
```

Visualization for the Armed Forces

```r
library(knitr)

# Choose subgroup
branch_choice <- "Army"
rank_prefix   <- "^E"
group_label   <- "Enlisted"

# Filter and summarize
sub_df <- forces_long %>%
  filter(Branch == branch_choice,
         grepl(rank_prefix, Pay_Grade))

freq_table <- sub_df %>%
  group_by(Sex, Pay_Grade) %>%
  summarise(Frequency = sum(Frequency), .groups = "drop") %>%
  pivot_wider(names_from = Pay_Grade, values_from = Frequency, values_fill = 0)

kable(freq_table,
      caption = paste0("Table 1. Frequency of Sex by Pay Grade (", group_label, "), U.S. ", bra
```

Table 1: Table 1. Frequency of Sex by Pay Grade (Enlisted), U.S. Army

Sex	E1	E2	E3	E4	E5	E6	E7	E8	E9
Female	1326	4336	10229	15143	10954	7363	4410	1472	394
Male	7429	22338	43775	79234	54803	49502	30264	9482	2865

Narrative Text for the Armed Forces Section

Table 1 shows the distribution of male and female enlisted personnel across pay grades in the U.S. Army. At lower ranks (E1–E3), female representation is slightly higher, but it decreases in higher grades (E7–E9). This pattern suggests that "sex and rank are not independent" in this subgroup, as the proportion of women varies with pay grade.