HW #2: Visualizing FEMA NRI Data

Ixel M.

```
# Upload necessary librarys
library(tidyverse)
library(janitor)
library(here)
```

Visualization Goal:

How do FEMA National Risk Index scores for counties in California compare to those in other states? This may require some data wrangling.

```
# Load in datafile
nr_index <- read_csv(here("data/National_Risk_Index_Counties_807384124455672111.csv"))
view(nr_index)
```

Data exploration

```
# Quck overview
head(nr_index)
# View column names and types
glimpse(nr_index)
# Summary statistics for all columns
summary(nr_index)

# Check unique values in County Type
unique(nr_index$`County Type`)
table(nr_index$`County Type`)

# Check how many states you have
```

```r
unique(nr_index$`State Name`)
length(unique(nr_index$`State Name`)) # should be > 50 if territories included

# Check data type of Risk Index Score
class(nr_index$`National Risk Index - Score - Composite`)

# Look at distribution of risk scores
summary(nr_index$`National Risk Index - Score - Composite`)
hist(nr_index$`National Risk Index - Score - Composite`)
```

```r
# Count NAs in risk score column
sum(is.na(nr_index$`National Risk Index - Score - Composite`))

nr_index %>% count(`County Type`)
nr_index %>% count(`County Name`)
```

Wrangle data for graph interpitation

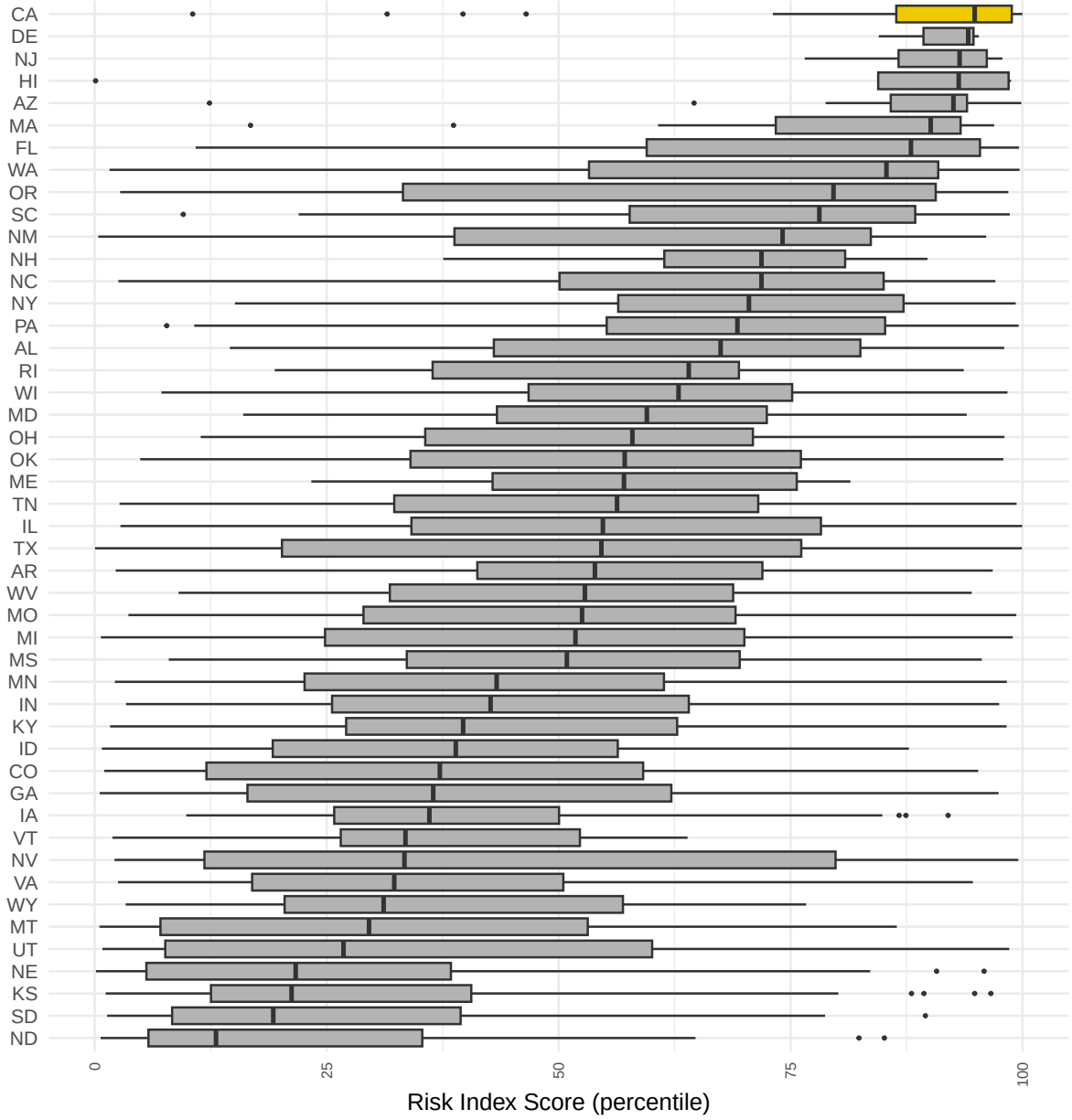
```r
nri_clean <- nr_index %>%
  clean_names() %>%
  filter(county_type == "County") %>% # keep only County rows
  # remove rows with missing risk scores
  filter(!is.na(national_risk_index_score_composite)) %>%
  # select and rename columns
  select(state = state_name,
         state_abbr = state_name_abbreviation,
         county = county_name,
         nri_score = national_risk_index_score_composite) %>%
# Create indicator for highlighting California in the plot
  mutate(is_california = state_abbr == "CA") %>%
# Calculate state median scores to reorder states for plotting
  group_by(state_abbr) %>%
  mutate(state_median = median(nri_score)) %>%
  ungroup() %>%
  mutate(state_abbr = fct_reorder(state_abbr, state_median))

# Check clean data
glimpse(nri_clean)
summary(nri_clean)
```

Distribution of risk scores for California counties compared to other states”

```
ggplot(nri_clean, aes(x = state_abbr, y = nri_score, fill = is_california)) +
  geom_boxplot(outlier.size = 0.5) +
  scale_fill_manual(values = c("TRUE" = "gold2", "FALSE" = "gray70"), guide = "none") +
  labs(
    title = "California Counties Face Higher Natural Hazard Risk than most States",
    subtitle = "Distribution of FEMA National Risk Index scores by state",
    x = NULL,
    y = "Risk Index Score (percentile)",
    caption = "Date: FEMA National Risk Index (2025 Release)",
    alt = "Boxplot showing distribution of FEMA National Risk Index scores for all 50 US
    states, ordered from lowest to highest median risk score.
    California is highlighted in gold and shows one of the highest median risk
    scores with relatively low variability among counties.
    Vermont shows high variability with low overall risk.") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 7)) +
  coord_flip()
```

California Counties Face Higher Natural Hazard Risk than most States

Distribution of FEMA National Risk Index scores by state



Date: FEMA National Risk Index (2025 Release)

Answer some questions

1. What are your variables of interest and what kinds of data (e.g. numeric, categorical, ordered, etc.) are they (a bullet point list is fine)?

- State, County Type, County Name are character (categorical) variables.
- National Risk Index - Score - Composite is a numeric (continuous) variable.

2. How did you decide which type of graphic form was best suited for answering the question? What alternative graphic forms could you have used instead? Why did you settle on this particular graphic form?

First I thought about the question and told myself I need to show distribution. Using the Data Viz decision tree, I identified several options for comparing distributions across multiple categories. I considered a ridge line plot as an alternative, which would show the full density curves for each state. However, I chose a boxplot because it clearly displays key distribution statistic (median, quartiles, and outliers) in a compact format, making it easier to quickly compare California's risk profile against all other state at once.

3. Summarize your main finding in no more than two sentences.

My findings show that California counties have higher natural hazard risk than most counties in other states. California's median risk score ranks among the highest nationally, with relatively low variability, meaning most counties face similar levels of risk. In contrast, Vermont counties show high variability—with counties having very different risk scores—but overall lower risk levels.

4. What modifications did you make to this visualization to make it more easily readable?

I used `coord_flip()` to display state abbreviation horizontally, making state labels and boxplots easier to read. I also created a California indicator variable (`is_california`) to highlight California with a distinct color, drawing immediate attention to the comparison of interest. Additionally, I ordered state by their median risk score to reveal ranking patterns more clearly.

5. Is there anything you wanted to implement, but didn't know how? If so, please describe.

I wanted to implement regional grouping to reduce visual clutter and add a reference line showing the national median,but I didn't know how to approach these modifications.