HW #2: Visualizing FEMA NRI Data

Emily Miller

2026-01-28

Introduction

This analysis explores FEMA National Risk Index (NRI) scores across US counties, comparing California counties to other states. The NRI measures community risk from natural hazards, considering exposure, annual loss, and social vulnerability.

Data Loading and Wrangling

```
# Load required packages
library(tidyverse)
library(ggbeeswarm)

# Read raw NRI data
nri_data <- read_csv("data/National_Risk_Index_Counties_807384124455672111.csv")

# Clean and prepare data for visualization
nri_data_clean <- nri_data %>%
  # Select relevant columns: identifiers and NRI score
  select(1:9, 15) %>%

  # Standardize column names to lowercase with underscores
  janitor::clean_names() %>%

  # Create binary indicator for California vs other states
  mutate(
    is_ca = if_else(state_name_abbreviation == "CA", "California", "Other States"),
    is_ca = factor(is_ca, levels = c("California", "Other States"))
  ) %>%
```

```
  # Calculate state-level summary statistics
  group_by(state_name_abbreviation, is_ca) %>%
  mutate(
    state_median = median(national_risk_index_score_composite, na.rm = TRUE),
    state_mean = mean(national_risk_index_score_composite, na.rm = TRUE),
    n_counties = n()
  ) %>%
  ungroup() %>%

  # Filter to complete cases and 50 US states only (exclude territories)
  filter(
    !is.na(national_risk_index_score_composite),
    state_name_abbreviation %in% state.abb
  )
```

Visualization

```
# Define custom color palettes for highlighting California
ca_colors <- c("California" = "#047C91", "Other States" = "#D3D3D3")
ca_outline <- c("California" = "#025A6B", "Other States" = "gray50")

# Create visualization
nri_data_clean %>%
  # Order states by median risk score (highest to lowest) and reverse to put highest first
  ggplot(aes(
    x = fct_rev(fct_reorder(state_name_abbreviation, state_median)),
    y = national_risk_index_score_composite,
    fill = is_ca,
    color = is_ca
  )) +

  # Add box plots showing distribution quartiles
  geom_boxplot(
    outlier.alpha = 0.4,
    outlier.size = 0.8,
    linewidth = 0.5
  ) +

  # Overlay individual county points to show data density
  geom_quasirandom(
```

```
    alpha = 0.2,
    size = 1.2,
    width = 0.3
  ) +

  # Add descriptive labels
  labs(
    title = "California Has the Highest Median Natural Hazard Risk Among US States",
    subtitle = "California's counties show consistently high risk scores with notable variati
    x = "State (ordered by median risk score, highest to lowest)",
    y = "National Risk Index Score",
    caption = "Data: FEMA National Risk Index (2025 Release)",
    fill = NULL,
    color = NULL
  ) +

  # Apply custom color schemes
  scale_fill_manual(values = ca_colors) +
  scale_color_manual(values = ca_outline, guide = "none") +

  # Set y-axis breaks for readability
  scale_y_continuous(breaks = seq(0, 100, 10)) +

  # Apply minimalist theme with custom styling
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 18, margin = margin(b = 5))
    plot.subtitle = element_text(hjust = 0.5, size = 14, color = "gray30", margin = margin(b
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 12),
    axis.text.y = element_text(size = 12),
    axis.title.x = element_text(size = 12, face = "bold", margin = margin(t = 8)),
    axis.title.y = element_text(size = 12, face = "bold", margin = margin(r = 8)),
    legend.position = "top",
    legend.text = element_text(size = 12),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank(),
    panel.grid.major.y = element_line(color = "gray90", linewidth = 0.3),
    plot.caption = element_text(size = 12, color = "gray50", hjust = 0, margin = margin(t =
  )
```

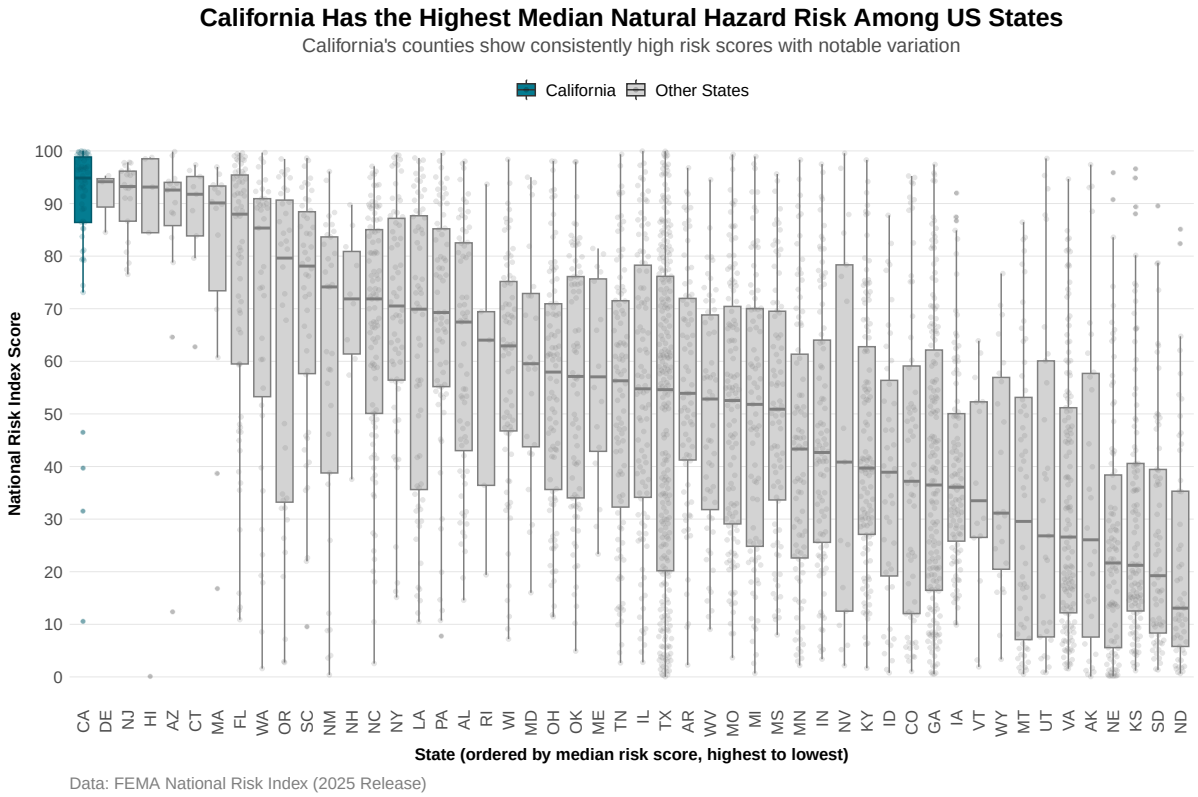

Figure 1

Analysis Questions

1. Variables of Interest

- **State abbreviation** (`state_name_abbreviation`): categorical (nominal), 50 levels representing US states
- **County name** (`county_name`): categorical (nominal), 3000+ levels representing individual counties
- **National Risk Index Score** (`national_risk_index_score_composite`): numeric (continuous), ranging from 0-100
- **California indicator** (`is_ca`): categorical (binary), derived variable with two levels (California, Other States)
- **State median score** (`state_median`): numeric (continuous), derived variable representing median NRI score per state

2. Graphic Form Selection

I chose a **box plot with overlaid individual points** (using `geom_quasirandom`) to answer the question because it effectively shows both summary statistics and individual data points simultaneously. The box plot component reveals median values, quartiles, and outliers for each state, while the overlaid points show the density and distribution of individual counties within each state.

Alternative graphic forms considered:

- **Ridge plot**: Would show distribution shapes clearly but makes it harder to compare medians across many states
- **Violin plot**: Similar to box plot but emphasizes distribution shape; however, with 50 states, the plot becomes cluttered
- **Beeswarm plot alone**: Shows all individual counties but loses the summary statistics that box plots provide
- **Density overlay**: Would cleanly compare California to all other states combined, but loses state-by-state comparison

I chose a box plot with points because it balances data transparency (showing every county as a point) with statistical summary (showing medians and quartiles), while still showing a direct comparison across all 50 states. The combination makes California's position as the highest-risk state clear while preserving nuance about within-state variation.

3. Main Finding

California counties have the highest median natural hazard risk score among all 50 US states, with a median NRI score of approximately 95. California shows some within-state variation, with outlier counties scoring lower, but the majority of California counties cluster in the high-risk range (85-100), which is still substantially higher than other states whose median scores range from approximately 5 to 40.

4. Readability Modifications

To enhance readability, I made the following modifications:

- **Ordered states by median risk score** (highest to lowest) to create a natural ranking and immediately highlight California's position
- **Used contrasting colors**: California in teal (#047C91) with a darker outline (#025A6B), other states in light gray, creating clear visual hierarchy
- **Overlaid semi-transparent points** (alpha = 0.2) on box plots to show individual county data without overwhelming the plot
- **Rotated x-axis labels 90 degrees** to prevent overlap with 50 state abbreviations

- **Removed unnecessary grid lines** (minor grid and vertical major grid) to reduce visual clutter
- **Added consistent y-axis breaks** (every 10 units) for easier value reading
- **Centered title and subtitle** to create a polished, professional appearance
- **Positioned legend at top** to make it easy to locate and when viewing the plot.

5. Future Implementation

I added reference lines from each state's median to the y-axis at first to make it easier to read exact median values but I felt that the lines created visual clutter than clarity when applied to all 50 states. Additionally, I considered adding state name labels directly on the plot for the top 5 highest-risk states to draw more attention to them, but was unsure how to implement this cleanly without obscuring the box plots.