Data Exploraton for NOAA's Storm Events Database in Puerto Rico (1996-2025)

Melannie Moreno Rolón

2026-02-03

Data Exploration: NOAA's Storm Events Database

This work involves exploring the database, which has been filtered within the database to include only the following event types:

- Flash Flood

- Heavy Rain
- Waterspout

- Flood
- Coastal Flood
- Drought
- Hurricane (Typhoon)
- Tropical Storm
- Tropical Depression

- Excessive Heat

```
# Load packages
library(tidyverse)
library(janitor)
library(here)
```

Data Loading

```
# Save all data files
storm_files <- list.files(
  path = "C:/Documents/MEDS/EDS240 - data visualization/eds240-infographic/data",
  pattern = "\\.csv$",
  full.names = TRUE
)

# Read in all data files and merge as one dataframe
storm_data <- storm_files %>%
  map_dfr(
    ~ read_csv(.x, col_types = cols(.default = col_character()))
  )
```

Clean Data

```
# Clean data
# Keep only the events of interest
storm_events <- c("Tropical Depression", "Tropical Storm",
                  "Hurricane (Typhoon)", "Heavy Rain",
                  "Flood", "Flash Flood", "Coastal Flood",
                  "Drought", "Excessive Heat", "Waterspout")

storm_clean <- storm_data %>%
  # Convert variables to numeric
  mutate(across(.cols = c(DEATHS_DIRECT, DEATHS_INDIRECT, INJURIES_DIRECT, INJURIES_INDIRECT

  # Standardize time strings
  mutate(
    BEGIN_TIME = str_pad(BEGIN_TIME, width = 4, side = "left",
                         pad = "0"),
    END_TIME = str_pad(END_TIME,   width = 4, side = "left",
                         pad = "0")
  ) %>%

  # Convert dates and derive time features
  mutate(BEGIN_DATE = mdy(BEGIN_DATE),
         END_DATE = mdy(END_DATE),
```

```
        YEAR = year(BEGIN_DATE),
        BEGIN_HOUR = as.integer(substr(BEGIN_TIME, 1, 2)),
        END_HOUR = as.integer(substr(END_TIME, 1, 2))) %>%

  # Select only the variables that will be used for analysis
  select(EVENT_ID, EPISODE_ID,
        STATE_ABBR, CZ_NAME_STR, CZ_TYPE, CZ_FIPS,
        EVENT_TYPE, MAGNITUDE, MAGNITUDE_TYPE,
        BEGIN_DATE, END_DATE, YEAR, BEGIN_HOUR, END_HOUR,
        DEATHS_DIRECT, DEATHS_INDIRECT,
        INJURIES_DIRECT, INJURIES_INDIRECT,
        DAMAGE_PROPERTY_NUM, DAMAGE_CROPS_NUM,
        SOURCE, FLOOD_CAUSE
        ) %>%

  # Standardize variable names
  clean_names() %>%

  # Filter for events of interest
  filter(event_type %in% storm_events)
```

Data Wrangling and Plotting

Figure 1. Frequency of Storm Events in Puerto Rico from 1996 to 2025.

```
# Plot 1
storm_clean %>%
  # Count all distinct events
  count(event_type) %>%
  # Plot a bar chart for event type frequencies
  ggplot(aes(x = reorder(event_type, n), y = n)) +
  geom_col(na.rm = TRUE, fill = "royalblue") +

  # Switch axes
  coord_flip() +

  # Label chart
  labs(
    x = NULL,
    y = "Number of reported events",
```

```
    title = "Storm event frequency in Puerto Rico (1996-2025)"
  ) +

  # Set the chart's theme
  theme_classic() +

  # Remove spacing between axes and data
  scale_x_discrete(expand = c(0, NA)) +
  scale_y_continuous(expand = c(0, NA))
```

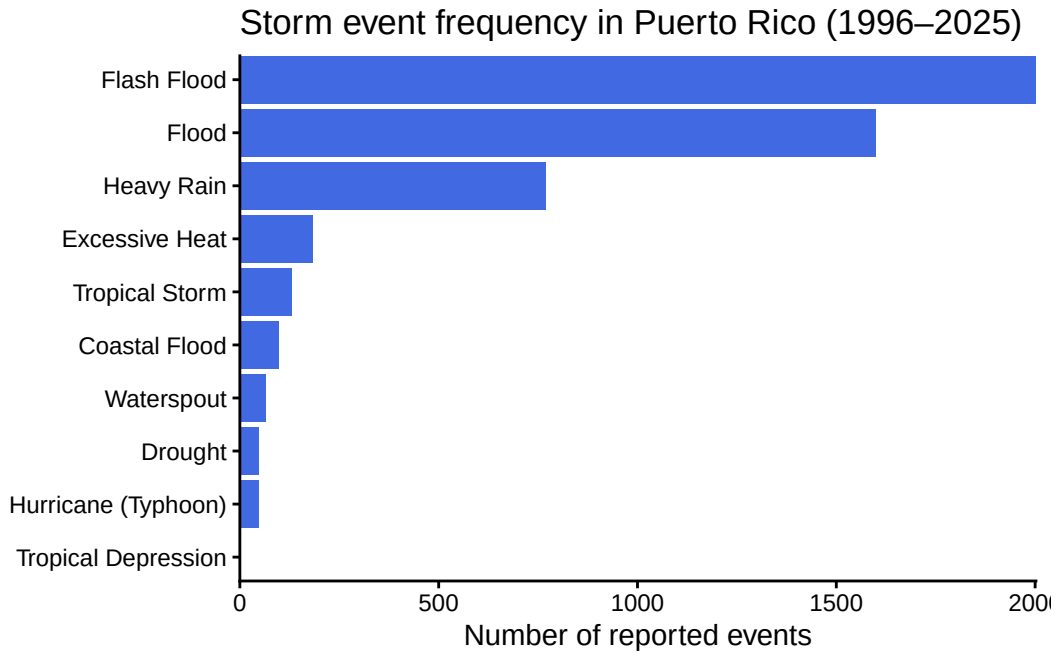


Figure 2. Distribution of top 6 most frequent storm events in PR over time.

```
# Plot 2
# Create dataframe for only the top 6 most frequent event types
top_events <- storm_clean %>%
  count(event_type) %>%
  slice_max(n, n = 6) %>%
  pull(event_type)

storm_clean %>%
  # Filter for only the top 6 events in the cleaned dataframe
  filter(event_type %in% top_events) %>%
  # Count storm event types for each year
```

```
  count(year, event_type) %>%
  ggplot(aes(year, n, color = event_type)) +
  geom_line(linewidth = 0.8) +
  # Set aspect ratio
  coord_fixed(ratio = 1/10) +
  # Label chart
  labs(
    title = "Most frequent storm events over time",
    y = "Number of events",
    x = NULL,
    color = "Event Type"
  ) +
  # Set the chart theme
  theme_classic()
```

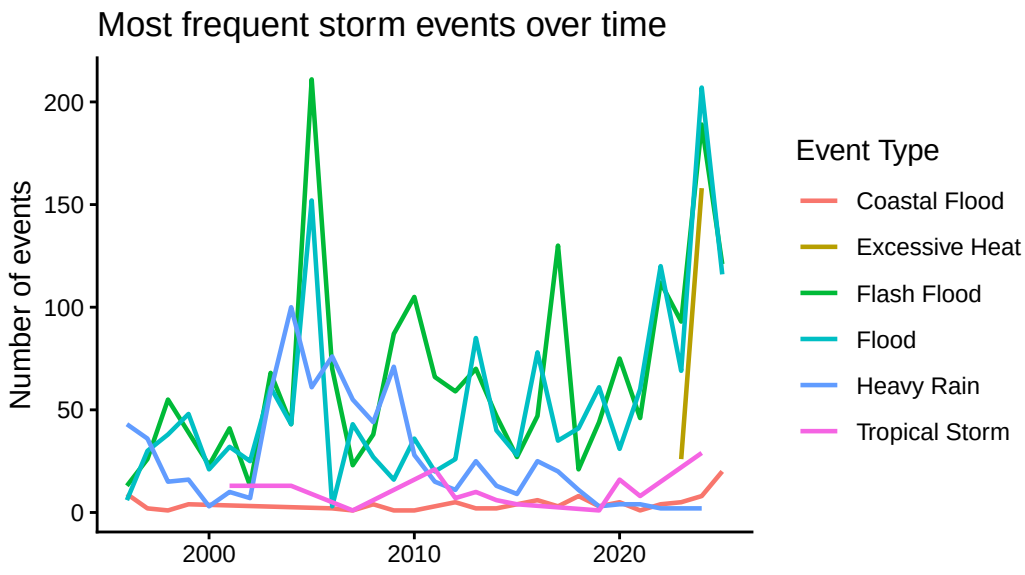


Figure 3. Comparing how often storms occur vs. how damaging they are

```
# Plot 3
storm_clean %>%
  group_by(event_type) %>%
  summarise(
    # Count the number of events per event type
    events = n(),
```

```
    # Calculate the median property damage
    median_damage = median(damage_property_num, na.rm = TRUE),
    # Keep only property damage numbers that are not zero
    median_damage_nz = median(damage_property_num[damage_property_num > 0], na.rm = TRUE),
    .groups = "drop"
  ) %>%
  # Filter out any nonzero values
  filter(!is.na(median_damage_nz)) %>%
  ggplot(aes(x = events, y = median_damage_nz, label = event_type)) +
  geom_point() +
  # Set log scale and add a dollar format for values in the y-axis
  scale_y_log10(labels = scales::dollar_format()) +
  # Label all point on the chart
  ggrepel::geom_text_repel(size = 3, max.overlaps = 20) +
  # Label the chart
  labs(
    x = "Number of events",
    y = "Median property damage per event"
  ) +
  # Add chart theme
  theme_classic()
```

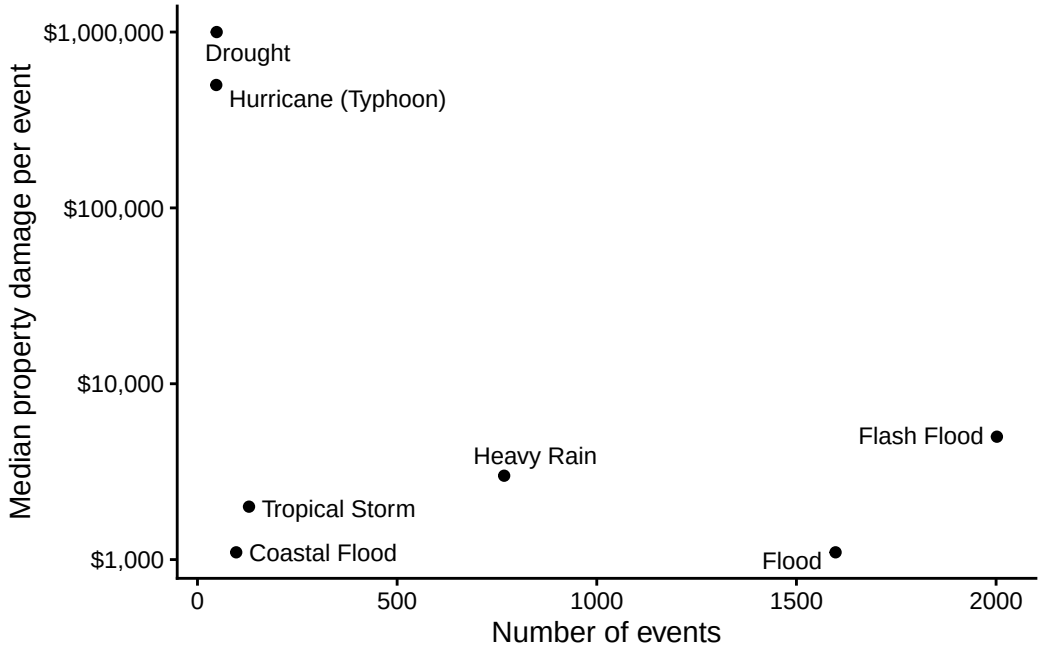


Figure 4. Comparing storm frequency and median damage

```r
# Plot 4
# Create dataframe to keep only events where median property damage values are greater than 0
storm_summary <- storm_clean %>%
  group_by(event_type) %>%
  summarise(
    events = n(),
    median_damage = median(damage_property_num[damage_property_num > 0], na.rm = TRUE),
    .groups = "drop"
  ) %>%
  filter(!is.na(median_damage))

# Convert data to long format
storm_long <- storm_summary %>%
  pivot_longer(
    cols = c(events, median_damage),
    names_to = "metric",
    values_to = "value"
  )

ggplot(storm_long, aes(x = value, y = event_type, color = metric)) +
  geom_point(size = 3) +
  geom_line(aes(group = event_type), color = "grey50") +
  # Set log scale on x-axis
  scale_x_log10(labels = scales::comma_format()) +
  labs(
    x = "Value (log)",
    y = NULL,
    title = "Storm frequency vs. damage",
    subtitle = "Puerto Rico, 1996-2025"
  ) +
  theme_classic()
```

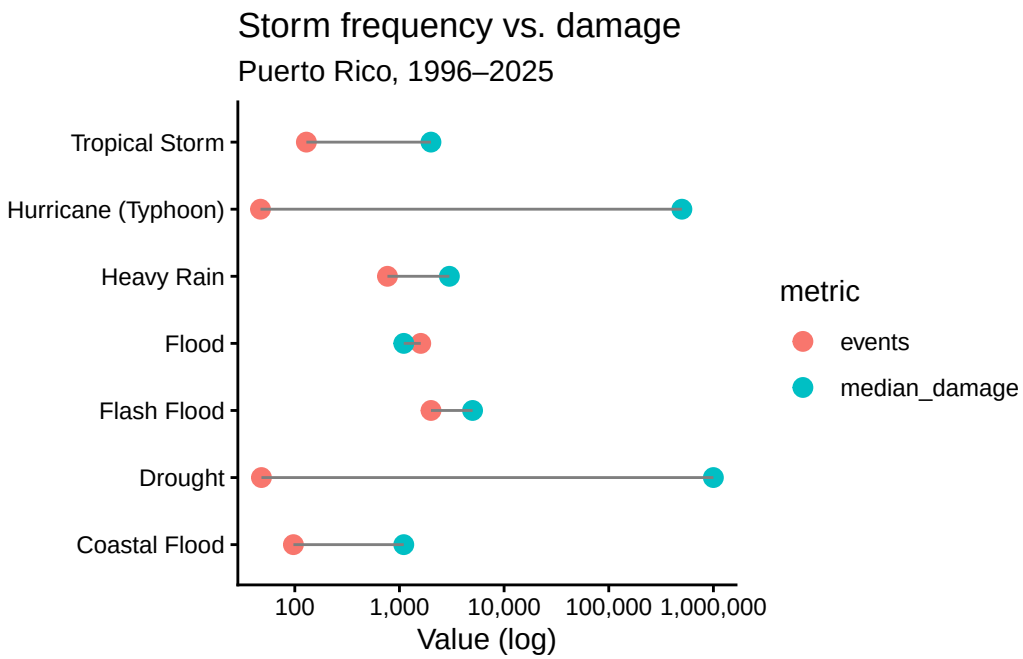
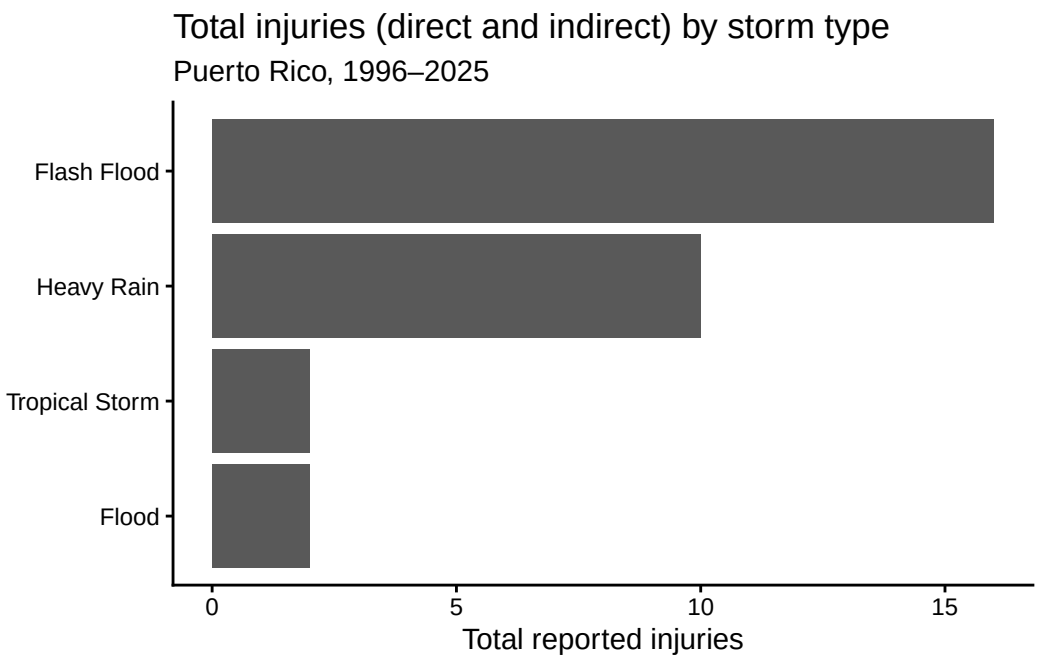


Figure 5. Injuries by storm type

```
storm_clean %>%
  # Find the total number of injuries
  mutate(total_injuries = injuries_direct + injuries_indirect) %>%
  # Find total injuries grouped by storm type
  group_by(event_type) %>%
  summarise(
    total_injuries = sum(total_injuries, na.rm = TRUE),
    events = n(),
    .groups = "drop"
  ) %>%
  # Keep only non-zero values
  filter(total_injuries > 0) %>%
  ggplot(aes(x = reorder(event_type, total_injuries), y = total_injuries)) +
  geom_col() +
  # Flip axes
  coord_flip() +
  # Set chart labels
  labs(
    x = NULL,
    y = "Total reported injuries",
    title = "Total injuries (direct and indirect) by storm type",
```

```
    subtitle = "Puerto Rico, 1996-2025"
  ) +
  # Set chart theme
  theme_classic()
```



V. Answer questions

1. What have you learned about your data? Have any potentially interesting patterns emerged? Point to specific visualizations that you created as you describe your findings.

After conducting an exploratory analysis of the NOAA Storm Events data for Puerto Rico from 1996 through 2025, I found differences between how often hazards occur and how damaging they are (Figure 3 and Figure 4). Flood-related events dominate the dataset by frequency, while hurricanes and droughts occur less often. The graphs comparing event counts to median property damage such as hurricanes and droughts tend to cause substantially higher damage per event than more frequent hazards. In contrast, more common events like floods and heavy rain are associated with lower damage. These patterns highlight that storm impacts in Puerto Rico are uneven and hazard-specific rather than uniform across event types. The total injuries plot (Figure 5) showed that injuries are relatively rare overall and concentrated in a small number of event types.

2. In FPM #1, you outlined some questions that you wanted to answer using these data. Have you made any strides towards answering those questions? If yes, how so? If no, what next steps do you need to take (e.g. I need to create X plot type, I still need to track down Y data, I need to restructure existing data so that you can visualize it in Z ways, etc.)? Have any new questions emerged?

The exploratory work has helped me make real progress toward answering my original questions about how storm hazards in Puerto Rico differ in how often they occur and how severe their impacts are. Through summary statistics and visualizations, I was able to clearly separate hazards that occur frequently but tend to have lower per-event impacts from hazards that occur less often but are associated with much higher damage when they do occur, especially using paired comparisons of event counts and median property damage per event. Injuries were harder to compare across storm types because the data are sparse and highly concentrated in just a few hazards, which showed me that injuries are not widespread across all events. Based on this, I am considering creating a waffle plot for the final infographic to more clearly communicate how different storm types contribute to the overall composition of reported events. Working through these analyses also raised new questions for me, including how changes in reporting practices may affect event counts over time, whether grouping hazards into broader categories would improve clarity, and whether impacts are better communicated on a per-event basis or as cumulative totals.

3. What challenges do you foresee encountering with your data? These can be data wrangling and / or visualization challenges.

I expect to encounter a few challenges to continue coming up as I work with these data. Both the damage and injury variables are highly zero-inflated, which makes it harder to use standard summary statistics. Event counts also vary a lot across storm types, so some hazards have enough data to support comparisons while others do not, which limits how broadly I can compare distributions without being misleading. I also ran into challenges when comparing metrics with very different units and scales, such as event counts and dollar amounts, and had to think carefully about how to design visuals so they are interpretable rather than confusing. Finally, I found that some plots were useful for exploring the data but not appropriate for final presentation, so a key challenge moving forward will be deciding which visualizations to refine for the infographic and which to leave as exploratory work.