DATA101_FinalProjectPaper

Khoa & Eurika

2025-11-12

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.2.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.4.1 --
## v broom        1.0.10     v rsample      1.3.1
## v dials        1.4.2      v tailor       0.1.0
## v infer        1.0.9      v tune         2.0.1
## v modeldata    1.5.1      v workflows    1.3.0
## v parsnip      1.3.3      v workflowsets 1.1.1
## v recipes      1.3.1      v yardstick    1.3.2
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```
library(tidyverse)
df <- read_csv("FinalProject.csv")
```

```
## Rows: 3000 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (3): event_id, country, event_type
## dbl  (16): year, month, severity, duration_days, affected_population, deaths...
## date  (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Section 1 - Introduction

Having lived in disaster prone countries, we've had first hand experience of the economic damage natural disasters bring along when occurring. To resolve our curiosity of the scale of damage different disasters cause, our research paper investigates the question: Which groups of natural disasters cause the greatest economic damage? To address this question, we used a public dataset from Kaggle that compiled worldwide records of natural disasters along with country-level economic damage estimates. The dataset includes information on disaster type and severity, duration, number of people affected, fatalities, and economic impact, allowing us to compare the economic consequences of different categories of natural disasters.

Section 2 - Data

The dataset used in this project contains information on natural disasters and their economic impacts across multiple countries and years, and it was likely compiled from sources such as the EM-DAT, World Bank economic reports, and national disaster response agencies. In total, the dataset consists of approximately 3,000 observations and 22 variables that capture the characteristics, impacts, and responses associated with these disasters. Each row in the dataset represents a recorded natural disaster event along with its corresponding economic impact, while each variable describes a specific characteristic of that event. Key variables used in this analysis include the disaster type (group), duration, severity level, and estimated economic damage. There are several potential limitations to the data: Economic impact figures may not be adjusted for inflation, exchange rates, or differences in population size, GDP, etc. across countries and years. Additionally, underreporting may occur, particularly in developing countries, which could lead to underestimated damage values. Finally, some disaster types, such as floods or earthquakes, are more frequently recorded than others, creating an uneven representation that could bias results and make certain disaster categories appear more or less economically damaging simply due to their higher number of recorded events.

Section 3 - Analytic Framework

The key variables used in this analysis are:

- economic_impact_million_usd (the estimated total economic loss): the response variable which is a numerical / continuous variables.
- event_group (the categorized version of disasters types): the primary explanatory variable which is a categorical variable.
- duration_days (how long the event lasted): the secondary explanatory variable, which is a continuous variable (acts like a major confounding variable).
- severity_level (how severe the disaster was): the last explanatory variable, which is a categorical variable as well (sort of acts like a confounding variable).

```
df <- df %>%
  mutate(
    event_group = case_when(
      event_type %in% c("Drought", "Wildfire", "Volcanic Eruption", "Heatwave") ~ "Heat Events",
      event_type %in% c("Tsunami", "Flood") ~ "Water Events",
      event_type %in% c("Earthquake", "Landslide") ~ "Earth Movement Events",
      event_type %in% "Cold Wave" ~ "Cold Events",
      event_type %in% c("Hurricane", "Tornado", "Hailstorm") ~ "Storm Events"
    ))
```

Because there are many different types of disasters, the visualizations become hard to interpret, so we group them into five main categories: Heat Events, Water Events, Earth Movement Events, Cold Events, and Storm Events.

```
df <- df %>%
  mutate(
    severity_level = case_when(
```

```
      severity %in% 1:2 ~ "Minor",
      severity %in% 3:4 ~ "Moderate",
      severity %in% 5:6 ~ "Severe",
      severity %in% 7:8 ~ "Very Severe",
      severity %in% 9:10 ~ "Catastrophic"
    ))
```

The 'severity' variable only takes interger values from 1 to 10 to represent the ordered levels of disaster intensity. This confirms that the variable is discrete and ordinal rather than continuous. In other words, each number represents a distinct category of intensity, not a numeric scale with equal intervals. Therefore, we grouped them into five ordered categories: Minor, Moderate, Severe, Very Severe, and Catastrophic.

In most of the plots we've used a log10 transformation for economic impact because the raw values are extremely right-skewed, with many small events and a few disasters that cause hundreds of millions in losses. The log10 scale reduces the influence of these extreme outliers and spreads out the smaller values, allowing the distributions, and trends to become visually interpretable. Importantly, the transformation does not change the ordering of events or the underlying relationships because it simply rescales the data so that meaningful patterns can emerge.

Section 4 - Results

```
cat_summary <- df %>%
  count(event_group, severity_level) %>%
  pivot_wider(
    names_from = severity_level,
    values_from = n,
    values_fill = 0
  )

cat_summary
```

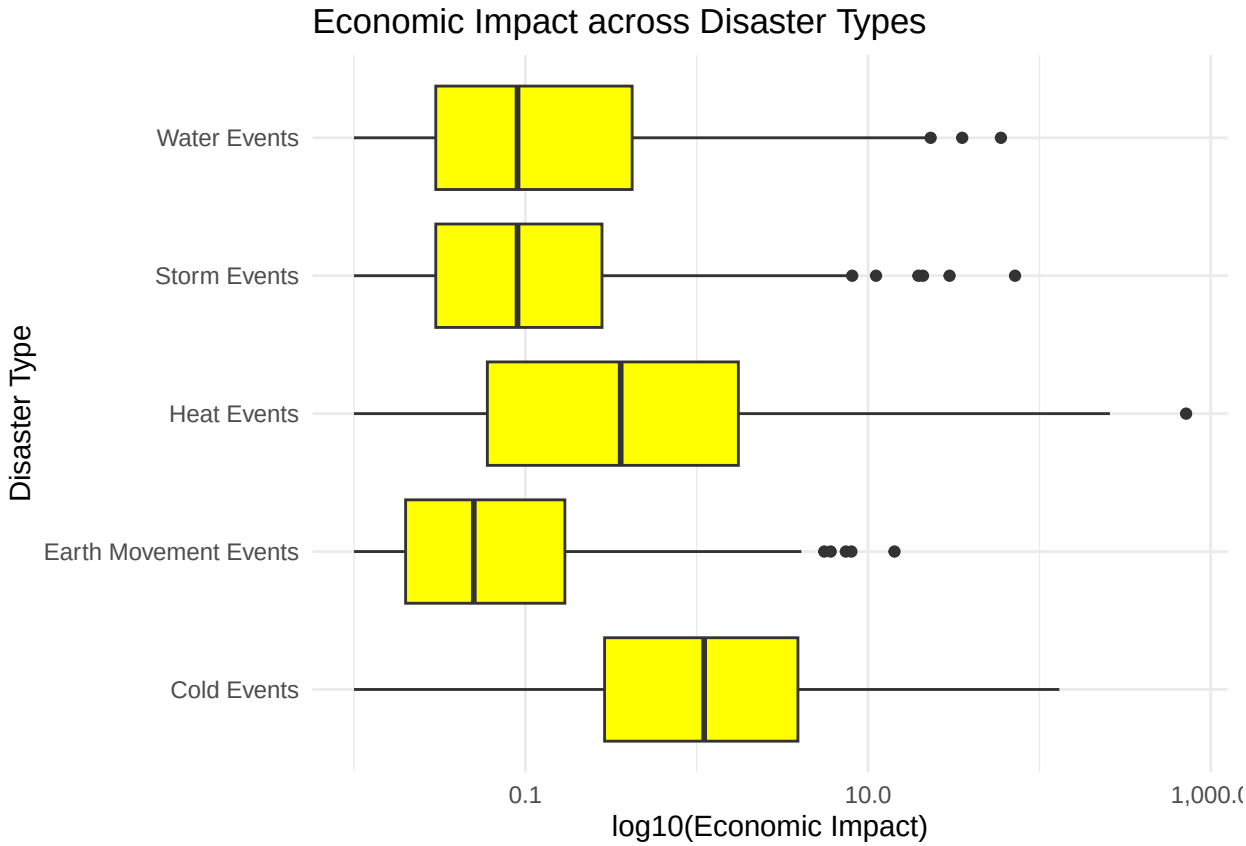
```
## # A tibble: 5 x 6
##   event_group           Catastrophic Minor Moderate Severe `Very Severe`
##   <chr>                        <int> <int>    <int>  <int>         <int>
## 1 Cold Events                      1    37       84     79            37
## 2 Earth Movement Events            2   204      184    131            41
## 3 Heat Events                      6   266      335    268           120
## 4 Storm Events                     2   223      241    201            70
## 5 Water Events                     3   196      136     94            39
```

The table shows general distribution of the 5 event group categories across the severity levels. Across all groups, Minor and Moderate events make up the majority of disasters, while Catastrophic events are extremely rare in every category. Heat Events and Storm Events are the most common overall, while Cold Events and Water Events appear less frequently. Unlike our expectation of disasters like water events or storm events being the most severe type, we noticed heat events had the greatest occurrence of being the most severe/ catastrophic.

```
ggplot(df, aes(x = event_group, y = economic_impact_million_usd)) +
  geom_boxplot(fill = "yellow") +
  scale_y_log10(labels = scales::comma) +
  coord_flip()+
  theme_minimal() +
  labs(title = "Economic Impact across Disaster Types",
       x = "Disaster Type",
       y = "log10(Economic Impact)")
```

```
## Warning in scale_y_log10(labels = scales::comma): log-10 transformation
## introduced infinite values.
```

```
## Warning: Removed 432 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



The boxplot shows the distribution of log-transformed economic losses across disaster groups. Cold events have the highest median log-impact values, indicating they may cause the greatest financial damage. Heat Events might be the most serious disaster because they also contain some really high outliers. Earth Movement Events have relatively lower impacts.

```
df_lm <- lm(economic_impact_million_usd ~ event_group + duration_days + severity_level, data = df)
summary(df_lm)
```

```
##
## Call:
## lm(formula = economic_impact_million_usd ~ event_group + duration_days +
##     severity_level, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18.60  -1.84  -0.36   0.56 706.24
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.14342    4.28490   0.267    0.790
## event_groupEarth Movement Events 0.27499    1.34758   0.204    0.838
## event_groupHeat Events           1.23749    1.15591   1.071    0.284
```

```
## event_groupStorm Events           0.20626    1.29499   0.159    0.873
## event_groupWater Events           0.48765    1.36211   0.358    0.720
## duration_days                     0.16678    0.02309   7.222 6.45e-13 ***
## severity_levelMinor              -2.59417    4.15569  -0.624    0.533
## severity_levelModerate           -1.92819    4.15377  -0.464    0.643
## severity_levelSevere             -0.46646    4.16184  -0.112    0.911
## severity_levelVery Severe         2.91403    4.21754   0.691    0.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.43 on 2990 degrees of freedom
## Multiple R-squared:  0.04156,    Adjusted R-squared:  0.03868
## F-statistic: 14.41 on 9 and 2990 DF,  p-value: < 2.2e-16
```

The results show that duration is the only strong and statistically significant predictor, with each additional day increasing expected economic losses by about 0.17 million USD ($p < 0.001$).

Heat Events have the largest positive coefficient, suggesting they tend to cause higher losses than Cold Events, but the estimate is not statistically significant, and the same is true for Earth Movement, Storm, and Water Events($p > 0.05$).

The Severity level coefficients are also not important because they all have large standard errors and p>0.05. This indicates that severity differences do not translate into reliably different economic impacts when duration is also accounted for.

The model explains only about 4% of the variation in economic impact, which is expected given how unpredictable and heterogeneous real disasters are.

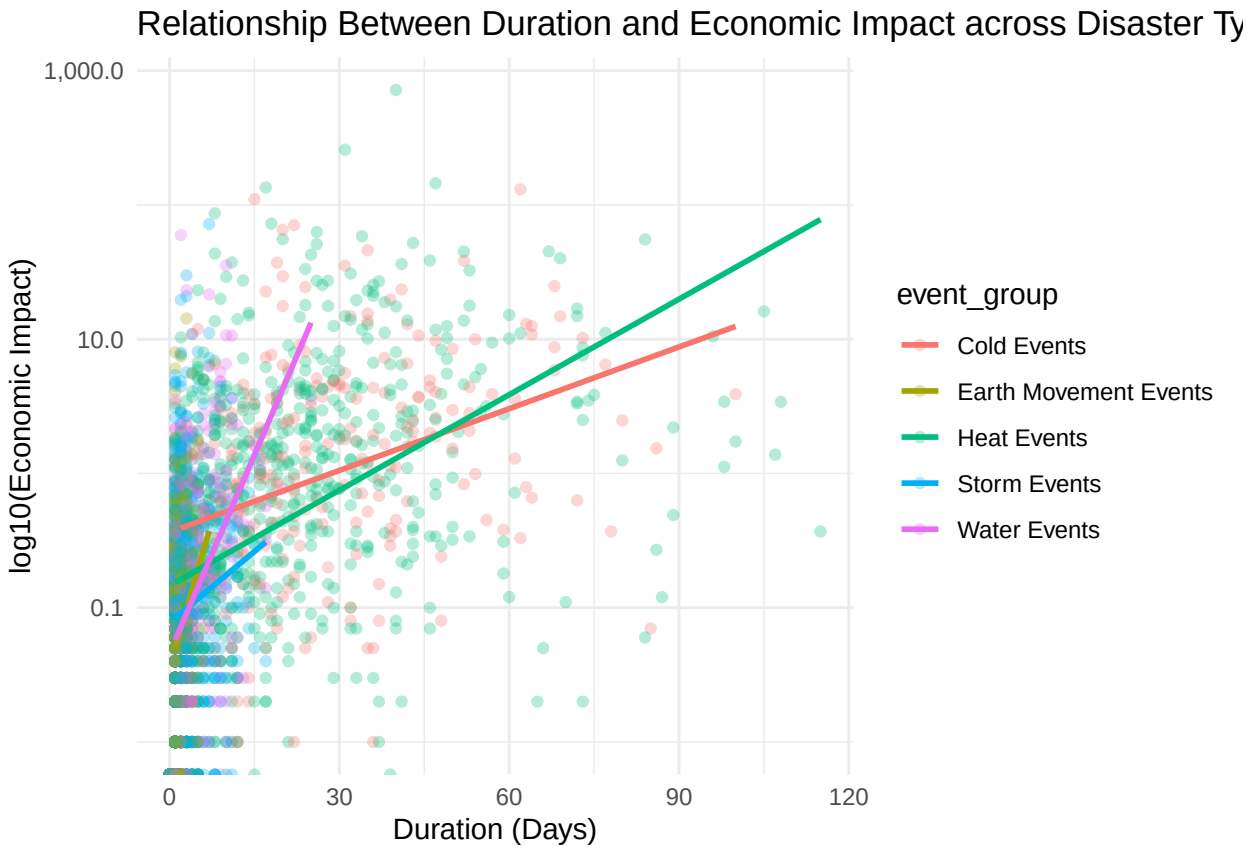
We can conclude that the duration of a disaster is far more important than its severity or event type in predicting economic damage.

```
ggplot(df, aes(x = duration_days, y = economic_impact_million_usd, color = event_group)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_y_log10(labels = scales::comma) +
  theme_minimal() +
  labs(
    title = "Relationship Between Duration and Economic Impact across Disaster Types",
    x = "Duration (Days)",
    y = "log10(Economic Impact)"
  )
```

```
## Warning in scale_y_log10(labels = scales::comma): log-10 transformation introduced infinite values.
## log-10 transformation introduced infinite values.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 432 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

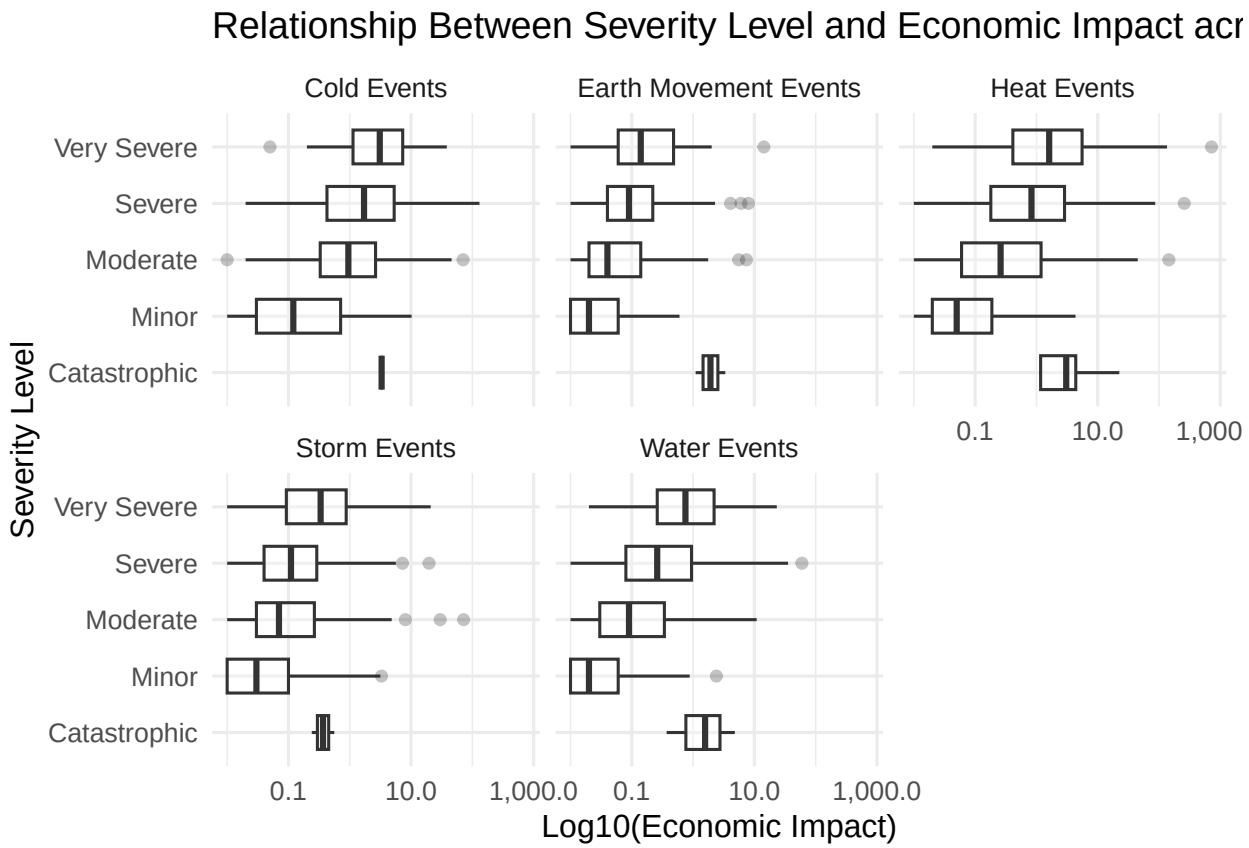


This scatterplot shows the relationship between disaster duration and the log-transformed economic impact for each event group. Across all groups, the fitted lines slope upward, indicating a clear positive relationship: as duration increases, economic losses increase, regardless of the type of disaster. Heat Events and Cold Events show the steepest slopes, which means that these categories experience the fastest growth in economic damage as they last longer. Storm Events and Water Events have flatter slopes. Earth Movement Events tend to last in short days.

```
ggplot(df, aes(x = severity_level, y = economic_impact_million_usd)) +
  geom_boxplot(alpha = 0.3, width = 0.6) +
  facet_wrap(~ event_group) +
  coord_flip()+
  scale_y_log10(labels = scales::comma)+
  theme_minimal(base_size = 12) +
  labs(
    title = "Relationship Between Severity Level and Economic Impact across Disaster Types",
    x = "Severity Level",
    y = "Log10(Economic Impact)"
  )
```

```
## Warning in scale_y_log10(labels = scales::comma): log-10 transformation
## introduced infinite values.
```

```
## Warning: Removed 432 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



This faceted plot shows how economic impact varies across severity levels within each of the five disaster groups. Across all groups, the pattern is consistent: higher severity levels tend to be associated with higher economic losses. Heat Events and Water Events show the strongest upward trend, with Severe and Very Severe events producing much larger impacts than Minor ones. Catastrophic events appear very rarely in the dataset, so their boxplots are based on only a few observations and should not be interpreted too heavily.

Section 5 - Implications

The findings suggest that how long a disaster lasts matters more than its type or severity level when predicting economic damage. While severity and event type show descriptive patterns, they do not provide reliable predictive power once duration is accounted for. These results highlight the importance of focusing on disaster persistence when assessing economic risk.

The findings of this study have broader implications for both policymakers and the general public. By identifying which types of natural disasters cause the greatest economic damage, the results can help governments make more informed decisions about how disaster prevention, preparedness, and recovery resources should be allocated, particularly for long-term planning. A clearer understanding of the true economic scale of different disasters also allows policymakers to prioritize investments in infrastructure, early warning systems, and resilience strategies. In addition, this research helps the general public develop a better understanding of which natural disasters pose the greatest threat to economic stability.

However, some issues remain unresolved. Although we can estimate it, duration days is still hard to accurately guess in order to make a strong prediction about the economic impact. The only way to improve this is to record more and more data regarding duration of each disaster, increasing sample size which increases confidence. There could be possible bias from grouping the disasters into broader categories, which may hide important differences within each type. We can simply solve this by including more granular disaster categories.

Citations

https://www.mdpi.com/2073-4433/12/11/1448

https://www.kaggle.com/datasets/uom190346a/global-climate-events-and-economic-impact-dataset