Data Vizualization Portfolio

Kanit Mann

Reflection

This portfolio's five visualizations demonstrate my mastery of the course criteria. Throughout the semester, your feedback was instrumental in helping me identify and strengthen several key areas where I was initially weak, particularly regarding professional polish, effective communication, and a deeper understanding of the data itself.

Initially, I struggled with professional presentation, specifically criteria #13 (data artifacts) and #8 (table formatting). Early assignments featured processing artifacts like underscores in labels, and your feedback, a stark reminder of how such details are perceived in a professional context, was a critical turning point. This newfound attention to detail is evident in visualizations 1 and 3, which are free of such errors. Similarly, after receiving extensive feedback on a previous table, visualization 1 now demonstrates mastery of proper formatting, incorporating column spanners, correct alignment, and clear labels to create a polished, effective summary.

My approach to effective communication also evolved significantly. I learned to write punchy, insightful captions (criteria #15) that prioritize key takeaways for a general audience, a clear improvement from my earlier, methodology-focused descriptions (visualizations 2 and 4). My design decisions (criteria #11) are now guided by accessibility; visualization 3, for example, uses a colorblind-accessible palette, while all visualizations feature readable font sizes to reduce cognitive load. This is linked to a deeper understanding of the data itself (criteria #10). Where I once failed to consult data dictionaries or made misleading axis choices, I now make deliberate, honest design decisions grounded in the data's context, as seen in my handling of the Apollo asteroid data (visualization 3).

Based on this demonstrated growth, I believe I have earned an A in this course. I have met the assignment requirements, and my portfolio demonstrates mastery of nearly all course criteria, aligning with the A-level standard. More importantly, it documents a clear evolution from basic competency to creating professional, publication-ready work. I have internalized your feedback to eliminate data artifacts, write effective captions, implement accessibility standards, and make honest design choices, proving the mastery and dramatic improvement expected for a top grade.

Contents

Global Health Metrics Comparison

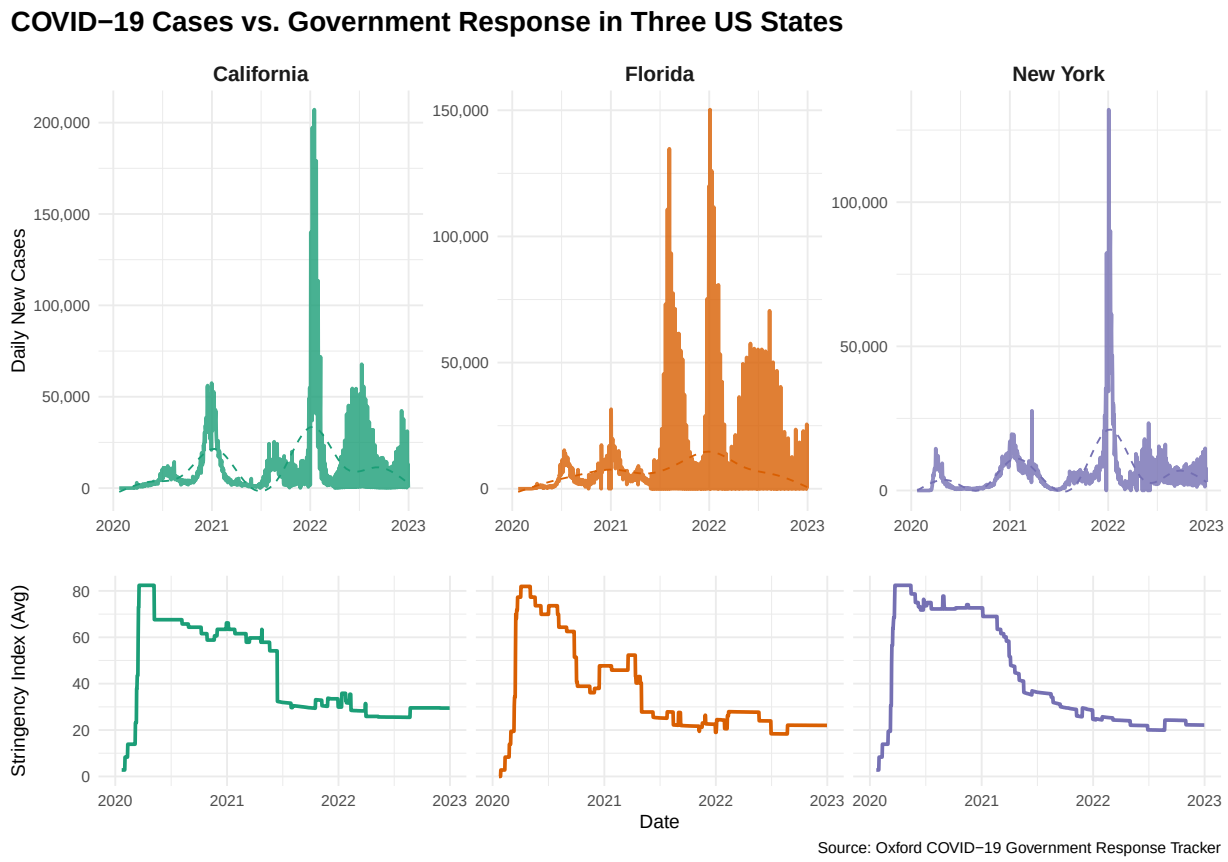
Data for the year 2022

Country	Health Indicators			Performance Ranks		
	Life Expectancy	Infant Mortality	Death Rate (per 1k)	Life Exp.	Infant Mort.	Death Rate
Japan	84.0	1.8	12.9	1	1	9
Australia	83.2	3.2	7.3	2	3	2
United Kingdom	81.0	3.9	9.5	3	4	6
Germany	80.6	3.1	12.7	4	2	8
China	78.2	4.8	7.4	5	5	3
United States	77.4	5.5	9.8	6	6	7
Brazil	74.9	12.6	7.5	7	7	4
India	71.7	25.6	6.6	8	9	1
South Africa	65.5	24.5	9.4	9	8	5

Visualization 1 - Table

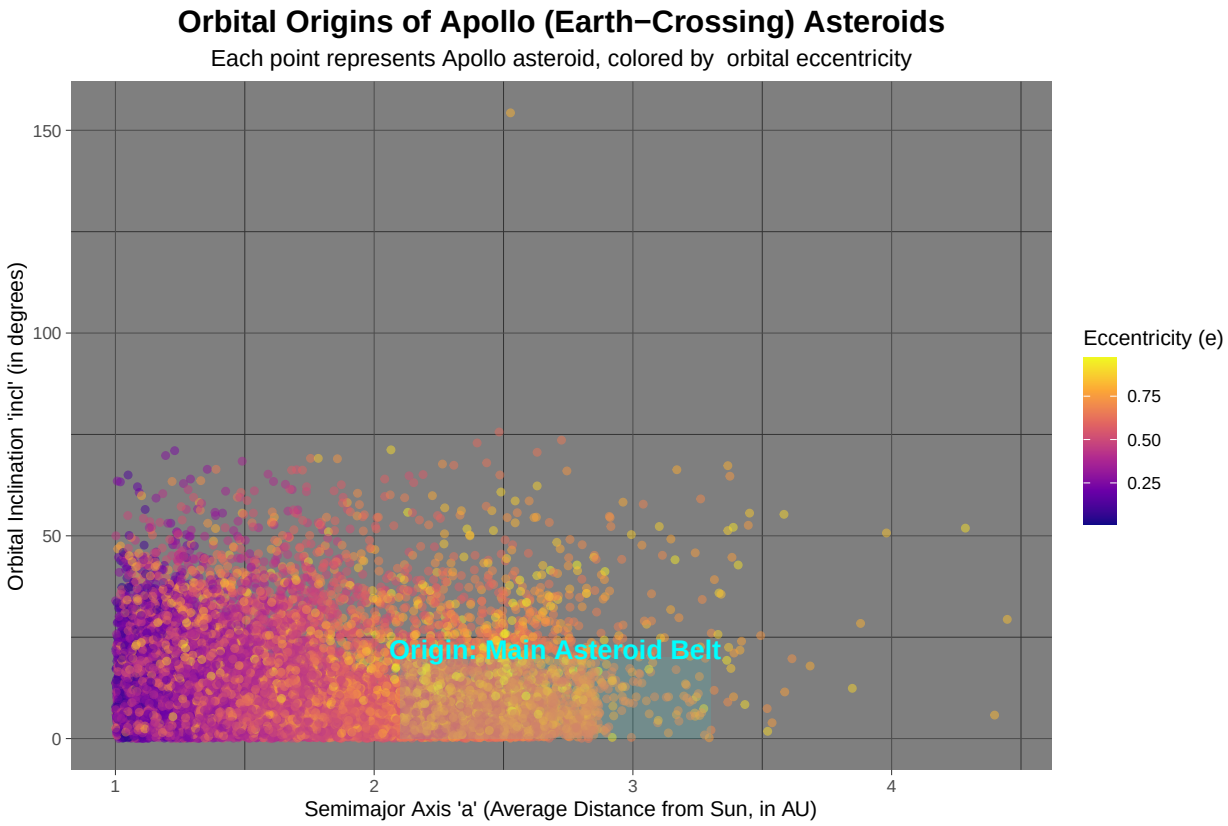
Caption: A comparison of key health indicators and performance ranks for ten selected countries in 2022. Japan leads the group in life expectancy, while other metrics vary significantly. Data sourced from *World Bank Development Indicators*. Ranks are color-coded for visual comparison, where green indicates a better outcome.

Visualization 2 - Stacked Line Plots



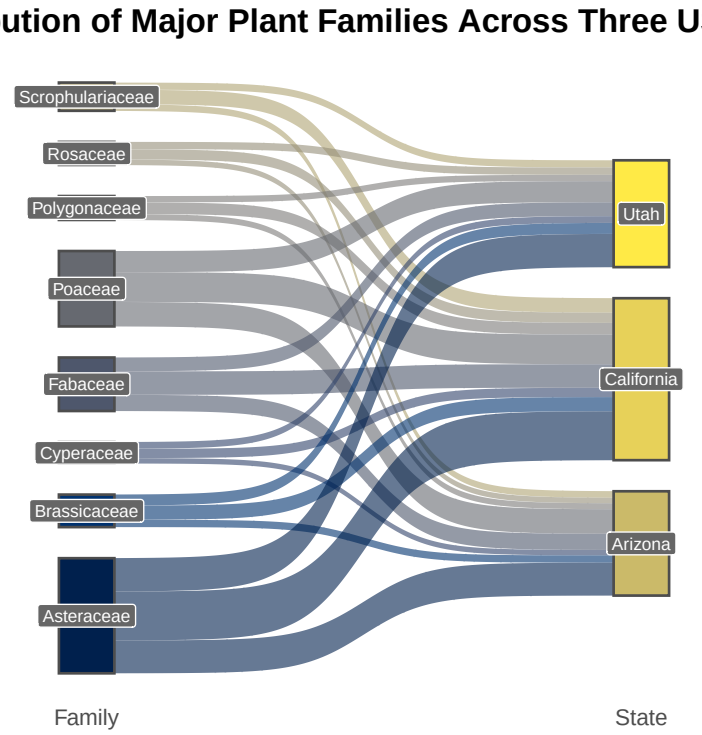
Caption: Did the Government become proactive, and increase stringency levels during COVID-19 waves? From **this dataset**, we can see the daily new COVID-19 cases (top) and government stringency levels (bottom) for New York, California, and Florida. This stacked view clearly shows how policy stringency (a measure of lockdown severity) rises in response to major waves of infection. It is clear from the data that government stringency levels increased at the beginning of the pandemic (2020), and we see spike in number of cases as soon as this strictness is lifted in mid 2021. Hence, the data supports that the government stringency levels did help control COVID-19 levels amongst other factors.

Visualization 3 - Scatter Plot



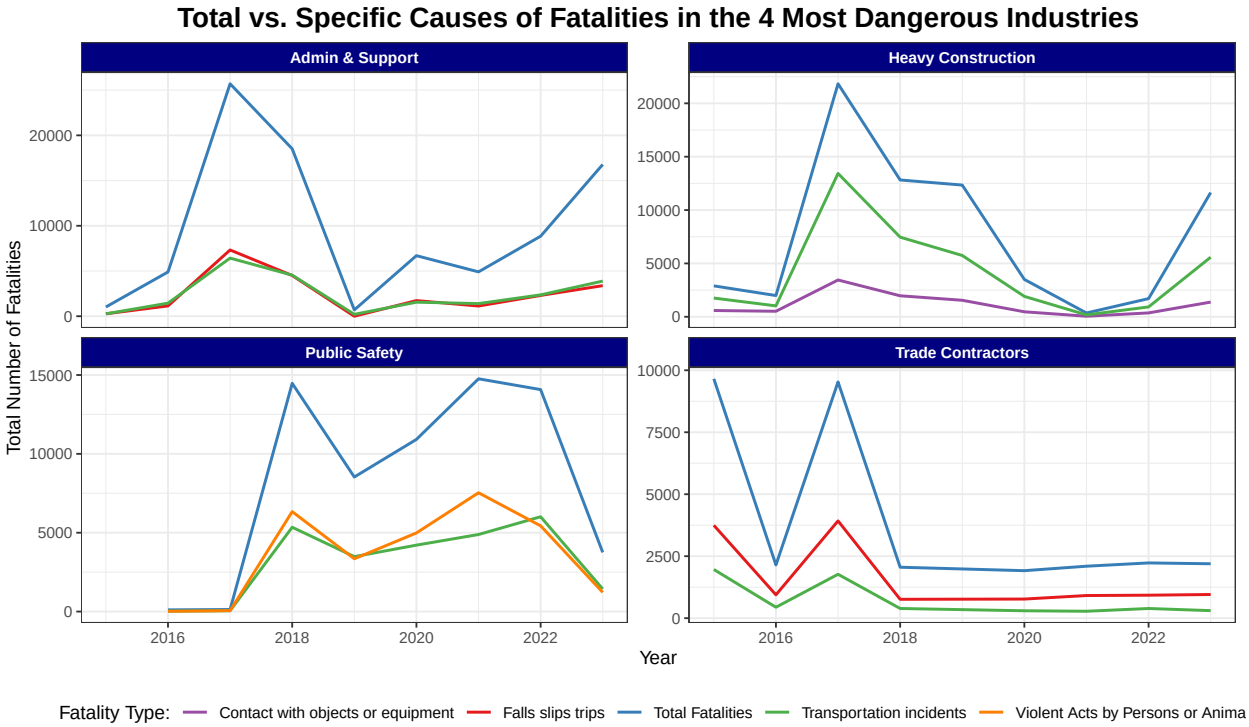
Caption:What are the orbital origins of Apollo (Earth-crossing) asteroids, and how do their orbital properties (distance, inclination, and shape) relate to each other? This scatter plot visualizes the orbital distribution of known Apollo asteroids. While many of these objects originate in the Main Asteroid Belt (the dense cluster from 2.1 to 3.3 AU), this dataset is a specific sample of asteroids whose orbits bring them close to the Sun and Earth.

Visualization 4 - Sankey Diagram



Caption: Are the most common plant families distributed equally across California, Arizona, and Utah, or do they show regional preferences? This Sankey diagram reveals the distribution patterns of major plant families across three distinct Western states. It shows that most families are not distributed equally and have clear regional preferences. For example, Asteraceae (the sunflower family) is highly dominant in California and Arizona but has a much smaller presence in Utah. In contrast, Poaceae (the grass family) is more evenly distributed across all three states, indicating it is more of a generalist.

Visualization 5 - Facetted Line Plot



Caption: How have the trends in total occupational fatalities differed across the most dangerous major industry groups? This series of line charts displays the annual trend in total occupational fatalities, along with Total Fatalities, faceted by the top nine industry groups with the highest incident rates. Each panel shows a separate industry, allowing for a direct comparison of their safety trends over time. We can see that, the major type of injury varies by the type of job that industry looks over.

My Code:

```
library(dplyr)
library(gridExtra)
library(stringr)
library(ggplot2)
library(ggsankey)
library(tidyverse)
library(gt)
library(janitor)
library(RColorBrewer)
library(lubridate)
library(patchwork)
```

```
health <- read_csv("./health.csv", skip = 4)
target_countries <- c("United States", "India", "Japan", "Germany", "Brazil",
                      "South Africa", "China", "Russia", "Australia", "United Kingdom")

metrics <- tibble(
  metric = c("Life Expectancy", "Infant Mortality", "Death Rate"),
  indicator_code = c("SP.DYN.LE00.IN",   # Life expectancy at birth, total (years)
                     "SP.DYN.IMRT.IN",   # Infant mortality rate (per 1,000 live births)
                     "SP.DYN.CDRT.IN")   # Death rate, crude (per 1,000 people)
)

filtered <- health %>%
  filter(`Country Name` %in% target_countries,
       `Indicator Code` %in% metrics$indicator_code) %>%
  select(`Country Name`, `Indicator Code`, `2022`) %>%
  pivot_wider(names_from = `Indicator Code`, values_from = `2022`) %>%
  rename(`Life Expectancy` = `SP.DYN.LE00.IN`,
         `Infant Mortality` = `SP.DYN.IMRT.IN`,
         `Death Rate` = `SP.DYN.CDRT.IN`)

ranked <- filtered %>%
  mutate(Rank_Life_Expectancy = rank(-`Life Expectancy`, na.last = "keep"),
         Rank_Infant_Mortality = rank(`Infant Mortality`, na.last = "keep"),
         Rank_Death_Rate = rank(`Death Rate`, na.last = "keep"))

ranked %>%
  arrange(Rank_Life_Expectancy) %>%
  gt() %>%
  tab_header(
    title = md("**Global Health Metrics Comparison**"),
    subtitle = "Data for the year 2022"
  ) %>%
  tab_spanner(
    label = md("**Health Indicators**"),
    columns = c(`Life Expectancy`, `Infant Mortality`, `Death Rate`)
  ) %>%
  tab_spanner(
    label = md("**Performance Ranks**"),
    columns = c(Rank_Life_Expectancy, Rank_Infant_Mortality, Rank_Death_Rate)
```

```
  ) %>%
  fmt_number(
    columns = c(`Life Expectancy`, `Infant Mortality`, `Death Rate`),
    decimals = 1
  ) %>%
  data_color(
    columns = Rank_Life_Expectancy,
    colors = scales::col_numeric(
      palette = c("tomato", "yellow", "green"),
      domain = c(10, 1)
    )
  ) %>%
  data_color(
    columns = c(Rank_Infant_Mortality, Rank_Death_Rate),
    colors = scales::col_numeric(
      palette = c("green", "yellow", "tomato"),
      domain = c(1, 10)
    )
  ) %>%
  # Cleaning up column labels
  cols_label(
    `Country Name` = "Country",
    `Death Rate` = "Death Rate (per 1k)",
    Rank_Life_Expectancy = "Life Exp.",
    Rank_Infant_Mortality = "Infant Mort.",
    Rank_Death_Rate = "Death Rate"
  ) %>%
  cols_width(
    `Country Name` ~ px(140),
    everything() ~ px(85)
  ) %>%
  opt_row_striping() %>%

  opt_horizontal_padding(scale = 0.5) %>%
  tab_options(
    table.font.size = px(14),

    heading.title.font.size = px(24),
    table.border.top.style = "hidden",
    table.border.bottom.style = "hidden"
  )
```

```
covid_data <- read_csv("./OxCGRT_USA_latest.csv")

states_of_interest <- c("US_NY", "US_CA", "US_FL")

plot_data <- covid_data %>%
  filter(RegionCode %in% states_of_interest) %>%
  mutate(Date = ymd(Date)) %>%
  arrange(RegionCode, Date) %>%
  group_by(RegionCode) %>%
  mutate(DailyCases = ConfirmedCases - lag(ConfirmedCases, default = 0)) %>%
  mutate(DailyCases = ifelse(DailyCases < 0, 0, DailyCases)) %>%
```

```
  select(Date, RegionCode, DailyCases, StringencyIndex_Average) %>%
  ungroup() %>%
  mutate(State = case_when(
    RegionCode == "US_NY" ~ "New York",
    RegionCode == "US_CA" ~ "California",
    RegionCode == "US_FL" ~ "Florida"
  ))

p_cases <- ggplot(plot_data, aes(
  x = Date,
  y = DailyCases,
  color = State)) +
  geom_line(alpha = 0.8, size = 1) +
  geom_smooth(se = FALSE, linetype = "dashed", size = 0.5) +
  facet_wrap(~ State, scales = "free_y") +
  scale_y_continuous(labels = scales::comma) +
  scale_color_brewer(palette = "Dark2") +
  labs(y = "Daily New Cases", x = "") +
  theme_minimal() +
  theme(legend.position = "none", strip.text = element_text(face = "bold", size = 12))

p_stringency <- ggplot(plot_data,
                       aes(
                         x = Date,
                         y = StringencyIndex_Average,
                         color = State)) +
  geom_line(size = 1) +
  facet_wrap(~ State) +
  scale_color_brewer(palette = "Dark2") +
  labs(
    y = "Stringency Index (Avg)",
    x = "Date",
    caption = "Source: Oxford COVID-19 Government Response Tracker"
  ) +
  theme_minimal() +
  theme(legend.position = "none", strip.text = element_blank())

p_cases / p_stringency +
  plot_layout(heights = c(2, 1)) +
  plot_annotation(
    title = "COVID-19 Cases vs. Government Response in Three US States",
  ) &
  theme(plot.title = element_text(size = 16, face = "bold"),
        plot.subtitle = element_text(size = 12))
```

```
asteroids_data <- read_csv("./Apollo Minor Planets Approaches.csv")

asteroids_cleaned <- asteroids_data %>%
  clean_names() %>%
  transmute(
    designation = designation,
    a = as.numeric(a),           # Semimajor axis
    incl = as.numeric(incl),     # Inclination
```

```
    e = as.numeric(e)            # Eccentricity
  ) %>%
  filter(!is.na(a), !is.na(incl), !is.na(e)) %>%
  filter(a < 5)

ggplot(asteroids_cleaned, aes(x = a, y = incl, color = e)) +
  geom_point(alpha = 0.5, size = 1.5) +
  scale_color_viridis_c(name = "Eccentricity (e)", option = "plasma") +

  labs(
    title = "Orbital Origins of Apollo (Earth-Crossing) Asteroids",
    subtitle = "Each point represents Apollo asteroid, colored by  orbital eccentricity",
    x = "Semimajor Axis 'a' (Average Distance from Sun, in AU)",
    y = "Orbital Inclination 'incl' (in degrees)"
  ) +

  theme_dark() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5, size = 12),
    panel.grid.major = element_line(color = "gray30"),
    panel.grid.minor = element_line(color = "gray20")
  ) +

  annotate("rect", xmin = 2.1, xmax = 3.3, ymin = 0, ymax = 20,
           alpha = 0.1, fill = "cyan") +
  annotate("text", x = 2.7, y = 22, label = "Origin: Main Asteroid Belt",
           color = "cyan", size = 5, fontface = "bold")
```

```
fatalities_data <- read_csv("./Dangerous Jobs.csv")

fatalities_renamed <- fatalities_data %>%
  clean_names() %>%
  filter(major_group != "NA" & !is.na(cause)) %>%
  mutate(
    major_group = case_when(
      major_group == "Administrative And Support Services"    ~ "Admin & Support",
      major_group == "Heavy And Civil Engineering Construction" ~ "Heavy Construction",
      major_group == "Justice, Public Order, And Safety Activities" ~ "Public Safety",
      major_group == "Specialty Trade Contractors"            ~ "Trade Contractors",
      major_group == "Transit And Ground Passenger Transportation" ~ "Passenger Transit",
      major_group == "Support Activities For Transportation"    ~ "Transport Support",
      major_group == "Waste Management And Remediation Services" ~ "Waste Management",
      TRUE                                                ~ major_group
    ),
    cause = if_else(cause == "Total.Fatalities", "Total Fatalities", cause)
  )

top_industries <- fatalities_renamed %>%
  group_by(major_group) %>%
  summarise(total = sum(fatalities, na.rm = TRUE)) %>%
  slice_max(order_by = total, n = 4)
```

```r
total_fatalities_summary <- fatalities_renamed %>%
  filter(major_group %in% top_industries$major_group) %>%
  group_by(major_group, year) %>%
  summarise(total_fatalities = sum(fatalities, na.rm = TRUE), .groups = 'drop')

top_causes_list <- fatalities_renamed %>%
  filter(major_group %in% top_industries$major_group) %>%
  group_by(major_group, cause) %>%
  summarise(cause_total = sum(fatalities, na.rm = TRUE), .groups = 'drop') %>%
  group_by(major_group) %>%
  slice_max(order_by = cause_total, n = 3)

top_causes_summary <- fatalities_renamed %>%
  semi_join(top_causes_list, by = c("major_group", "cause")) %>%
  group_by(major_group, year, cause) %>%
  summarise(total_fatalities = sum(fatalities, na.rm = TRUE), .groups = 'drop')

cause_names <- unique(top_causes_summary$cause)
cause_colors <- brewer.pal(length(cause_names), "Set1")
full_palette <- c("Total" = "gray30", setNames(cause_colors, cause_names))

ggplot() +
  # Layer 2: The Top 3 Causes
  geom_line(data = top_causes_summary,
            aes(x = year, y = total_fatalities, color = cause),
            linewidth = 0.8) +

  facet_wrap(~ major_group, scales = "free_y") +

  scale_color_manual(name = "Fatality Type:", values = full_palette) +

  labs(
    title = "Total vs. Specific Causes of Fatalities in the 4 Most Dangerous Industries",
    x = "Year",
    y = "Total Number of Fatalities"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5),
    strip.background = element_rect(fill = "navy"),
    strip.text = element_text(color = "white", face = "bold"),
    legend.position = "bottom"
  )
```