COLLEGE OF ENGINEERING AND COMPUTER SCIENCE

VINUNIVERSITY

[COMP4010] Data Visualization

Project 1 Write-up

Bachelor of Science in Engineering and Computer Science

Prepared by

Ha Phuong Thao	**SIS ID: 202000242**
Nguyen Duy Anh Quan	**SIS ID: 202000284**
Khau Lien Kiet	**SIS ID: 202000119**

Under the supervision of

Supervisor's Name: Prof. Le Duy Dung

1. Introduction

Data description: The dataset under consideration is a comprehensive collection of tornado occurrences in the United States from 1950 to 2022. It includes various attributes such as the date and time of each tornado, its geographical coordinates, magnitude, length, width, as well as the resulting injuries, fatalities, and property loss. This dataset, compiled from the National Weather Service and other meteorological sources, serves as a rich historical record of tornado activities and their impacts.

Variable	Type	Description
om	integer	**Tornado number,** ID for the tornado in this year
yr	integer	**Year,** 1950-2022
mo	integer	**Month,** Jan (1) – Dec (12)
st	character	**State,** two-letter postal abbreviation for the state (DC = Washington, PR = Puerto Rico, VI = Virgin Islands)
mag	integer	**Magnitude,** on the F scale (EF beginning in 2007). Some of the values are estimated (recorded in *fc* variable)
wid	double	**Width,** in yards
len	double	**Length,** in miles
fc	logical	**Was the *mag* column estimated?,** True/False
fat	integer	**Number of fatalities,** when summing for state totals, use *sn* == 1
sn	integer	**State number for this row,** 1 = row contains the entire track information for this state 0 = means there is at least one more entry for this state for this tornado
loss	double	**Estimated property loss information in dollars.** Prior to 1996, values were grouped into ranges. The reported number for such years is the maximum of its range

Reason: We chose this dataset due to its detailed recording of natural disasters, specifically tornadoes, which are frequent and impactful weather events in the United States. Understanding the patterns and effects of tornadoes is crucial for improving safety measures, emergency responses, and public awareness. Additionally, this dataset offers a unique opportunity to explore geographical, temporal, and physical aspects of tornadoes, which can lead to valuable insights for science and society.

2. **Question 1**
    a. Description

"What are the overall characteristics of US tornadoes over times?"

With the aforementioned question, we aim to understand the overall characteristics of tornadoes in the United States over time. The question could be further broken down into smaller aspects:

- *Mini Question 1:* What are the monthly patterns of US tornadoes?
- *Mini Question 2:* How has the distribution of the number of tornados, magnitude, width and length of tornadoes changed annually?

The specific variables of interest are the date of the tornado occurrence (date), the starting longitude and latitude (slon, slat) which define the geographical initiation point of the tornado, magnitude (mag) which quantifies the intensity, width (wid) and length (len) which describe the physical size, and the year (yr) which track changes over time.

For the first sub-question, we can analyze the seasonal distribution of tornado occurrences to identify peak months for tornado activity. It also assists in emergency preparedness and resource allocation by pinpointing times of heightened risk.

For the second sub-question, it can reveal whether tornado occurrences are becoming frequent and indicate possible reasons for such trends, such as climate change or natural variability in weather patterns. By examining the changes in physical dimensions of tornadoes, such as their width and length, we can assess variations in the destructive potential of tornadoes over time.

    b. Approach

Mini Question 1:

This plot uses a heat map faceted by month, where each facet represents different months of the year, displaying the geographic distribution of tornado occurrences across the US. Heat maps are excellent for showing the density and distribution of events across geographical areas. By assigning colors based on the frequency of tornadoes in each region, the map visually communicates areas of high and low activity. Faceting the heat map by month allows for easy comparison across the year, highlighting seasonal trends and enabling viewers to quickly discern which months and which regions experience the highest tornado activity.

Mini Question 2:

This plot analyzes changes in tornado characteristics over time. We first create a new data table summarizing yearly data, aggregate the data by year ('yr'), and calculate the number of tornados, the average magnitude ('mag'), the median width ('wid'), and the median length ('len') for each year. This can be done using data aggregation and summarization functions in Python (Pandas) or R (dplyr).

The reason for using median width and length is to reduce the effect of outliers on the dataset. For magnitude, average was used because the data is contained within a small range, thereby, we do not have to worry about the effect of outliers.

Because the variables we are using have different measurement units and different ranges, we plot them separately, but we give them the same y axis for easier comparison and in the end, combine all of them together.

We also add 3 notable years with significant tornado events to illustrate how such events would contribute to the overall statistics. Finally, we ensure that the graph has clear labels, a legend, and differentiated line styles or colors for each attribute to aid interpretation and comparison.

c. <u>Analysis</u>

Mini Question 1:

```
```{r}
bg <- "white"

# Create a US map
df_state <- map_data("state")

# Create a df to use for plotting different months
df_base <- tornados %>%
  mutate(
    month = format(date, "%b"),
    month = factor(month, levels = c("Jan", "Apr", "Jul", "Oct", "Feb",
"May", "Aug", "Nov", "Mar", "Jun", "Sep", "Dec")),
  )

# Create a dataframe for text position
df_base_labs <- df_base |>
  distinct(month) |>
  mutate(
    x = mean(range(df_state$long)),
    y = max(df_state$lat) + 2
  )

distribution_map <- ggplot(data = df_state) +
```
```

```
  geom_polygon(aes(long, lat, group = group), colour = "black", fill = NA,
linewidth = 0.1) +
  geom_density2d_filled(data = df_base, aes(slon, slat), alpha = 0.6,
contour_var = "count") +
  geom_text(aes(x, y, label = month), df_base_labs, family = main_font, size
= 4, colour = "black") +
  scale_fill_manual(values = pal_4, na.value = NA) +
  facet_wrap(~month, nrow = 3) +
  coord_map(clip = "off") +
  theme_void() +
  labs(title = "US Tornado Patterns by Month") +
  theme(
    text = element_text(family = main_font, size = 48, lineheight = 0.3,
colour = "black"),
    plot.background = element_rect(fill = bg, colour = bg),
    plot.caption = element_markdown(colour = "black", hjust = 0.5, margin =
margin(t = 20)),
    plot.margin = margin(b = 20, t = 50, r = 50, l = 50),
    legend.position = "none",
    strip.text = element_blank(),
    plot.title = element_text(face = "bold", size = 15, hjust = 0.5, margin =
margin(b = 20, unit = "pt"))
  )

distribution_map
```

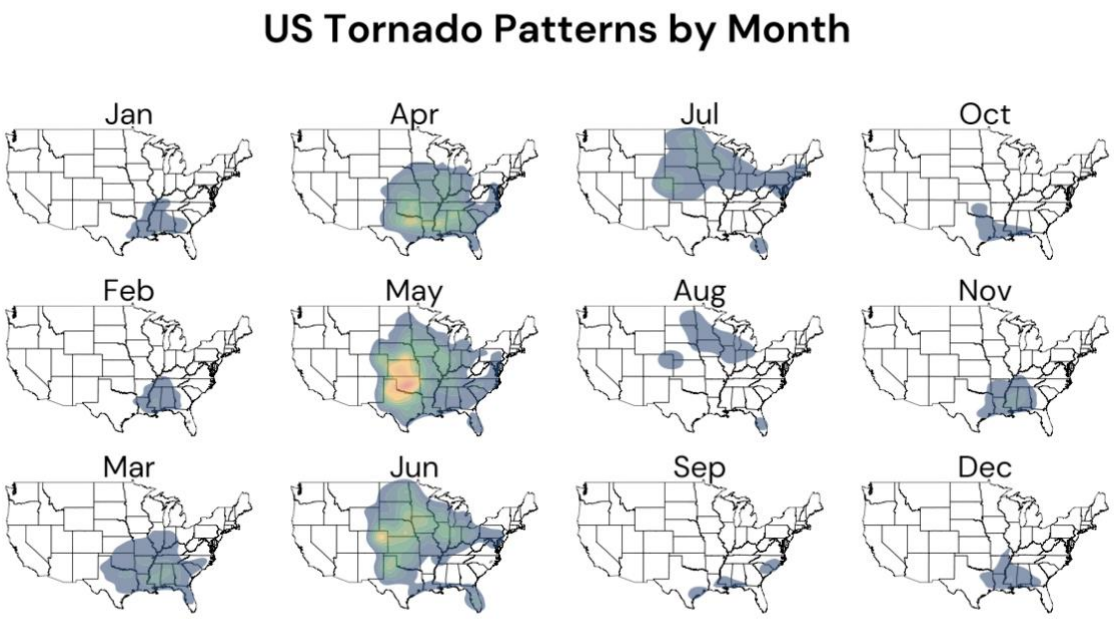

Figure 1

Mini Question 2:

```r
data_1b <- data.frame(tornados)

data_1b$mag <- as.numeric(as.character(data_1b$mag))

yearly_count <- data_1b %>%
  group_by(yr) %>%
  summarise(TotalTornadoes = n())

# Define a function to calculate mode
get_mode <- function(v) {
  uniqv <- unique(na.omit(v))
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Summarizing the data by year
yearly_summary <- data_1b %>%
  group_by(yr) %>%
  summarise(
    AverageMag = mean(mag, na.rm = TRUE),
    AverageWidth = median(wid, na.rm = TRUE),
    AverageLength = median(len, na.rm = TRUE)
  )

special_year <- c(1974, 2011, 2020)

plot_occurrences <- ggplot(yearly_count, aes(x = yr, y = TotalTornadoes)) +
  geom_line(size = 1, color = pal[2]) +
  geom_vline(xintercept = special_year, linetype = 2) +
  annotate("text", x = 2011, y = 470, size = 1.5, label = "2011 Super
Outbreak", angle = 90, vjust = -0.5) +
  annotate("text", x = 1974, y = 475, size = 1.5, label = "1974 Super
Outbreak", angle = 90, vjust = -0.5) +
  annotate("text", x = 2020, y = 620, size = 1.5, label = "2020 Easter
Tornado Outbreak", angle = 90, vjust = -0.5) +
  labs(title = "Total Yearly Tornado Occurrences",
       x = "Year",
       y = "Number of Tornadoes") +
  theme_1d()

# Creating three separate plots
plot_mag <- ggplot(yearly_summary, aes(x = yr, y = AverageMag)) +
  geom_line(size = 1, color = pal[4]) +
  geom_vline(xintercept = special_year, linetype = 2) +
  labs(title = "Annual Average Tornado Magnitude",
       x = "Year",
       y = "Average Magnitude") +
  theme_1d()

plot_width <- ggplot(yearly_summary, aes(x = yr, y = AverageWidth)) +
  geom_line(size = 1, color = pal[6]) +
    geom_vline(xintercept = special_year, linetype = 2) +
  labs(title = "Annual Median Tornado Width",
       x = "Year",
```

```
       y = "Average Width (yards)") +
  theme_1d()

plot_length <- ggplot(yearly_summary, aes(x = yr, y = AverageLength)) +
  geom_line(size = 0.8, color = pal[8]) +
  geom_vline(xintercept = special_year, linetype = 2) +
  labs(title = "Annual Median Tornado Length",
       x = "Year",
       y = "Average Length (miles)") +
  theme_1d()

plot_occurrences <- plot_occurrences + theme(plot.margin = margin(1, 3, 3, 1,
"mm"))
plot_mag <- plot_mag + theme(plot.margin = margin(1, 1, 3, 3, "mm"))
plot_width <- plot_width + theme(plot.margin = margin(3, 3, 1, 1, "mm"))
plot_length <- plot_length + theme(plot.margin = margin(3, 1, 1, 3, "mm"))

combined_plot <- grid.arrange(
  plot_occurrences, plot_mag, plot_width, plot_length,
  ncol = 2
)
```

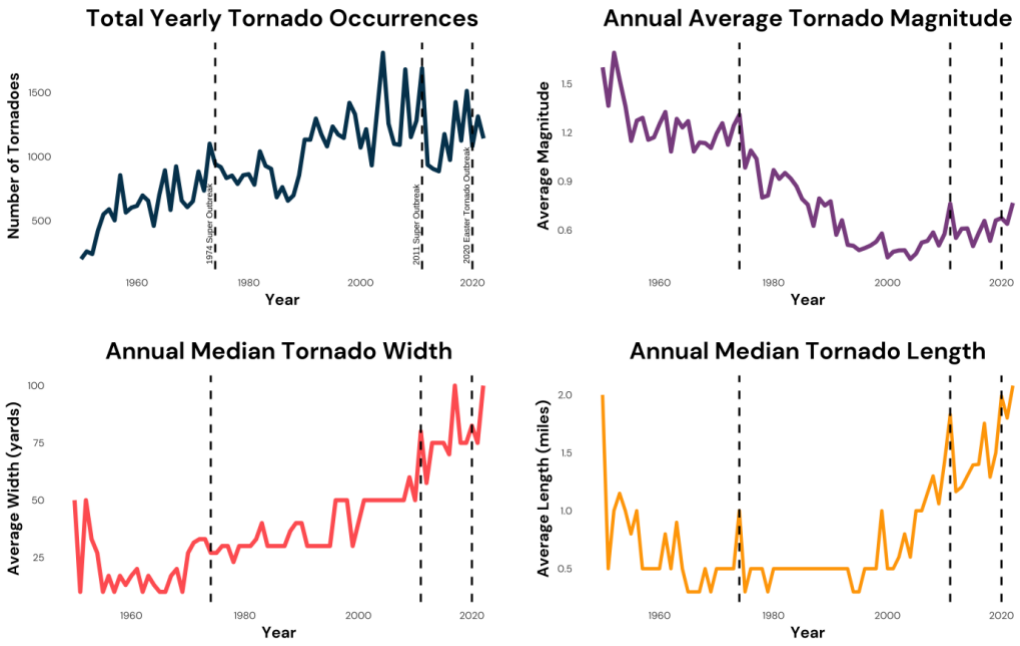

Figure 2

d. Discussion

Mini Question 1:

From the graph, it is abundantly clear that there is a clear indication of seasonal variability in tornado activity. Tornado occurrences in the United States are most frequent during the spring to early summer, spanning March through July. In these months, the mix of warm, humid air from the Gulf of Mexico and cool, dry air from the north results in an unstable

atmosphere. This clash of air masses often gives rise to intense thunderstorms and the genesis of tornadoes.

The central part of the country is the most affected region - often referred to as Tornado Alley, with some activity extending into the Southeast. The highest intensity is observed in states like Texas, Oklahoma, Kansas, and Nebraska. The central and southern parts of the country, often referred to as Tornado Alley, see a high concentration of tornadoes, especially from April to June. There's also notable activity in the Southeast, including Mississippi, Alabama, and Georgia, which may indicate the presence of "Dixie Alley," another region known for significant tornado occurrences. However, tornados can take place in any place, at any time during the year.

Mini Question 2:

There is an overall increase in the amount of yearly tornado occurrences, along with a decrease in the average tornado magnitude. This suggests that there are more lower magnitude tornadoes happening.

The median for tornado width and tornado length has also been steadily increasing. Along with the information from the previous analysis, it could be said that these are tornadoes with smaller magnitude but with a bigger coverage area.

It is interesting to see how big tornado events would affect the overall of the year that such events occur in. For example, in 2011, because of the Super Outbreak, we can notice a big spike in all 4 of the graphs.

3. **Question 2**
    a. <u>Description</u>

"What are the impacts of the tornados?"

The pursuit of understanding the impacts of tornadoes is essential since it could address both the human toll and economic ramifications. Tornadoes, one of nature's most violent phenomena, can lead to facilities, injuries, and significant property loss. The dataset's specific variables, such as state - "st", fatalities - "fat", property loss – "loss", tornado magnitude – "mag", and path dimensions as "len" and "wid" are critical for us to plot a comprehensive picture of these impact. With the aforementioned big questions, we broke it down into smaller aspects:

- *Mini Question 1:* How is each state affected by the tornadoes?
- *Mini Question 2:* What are the economic consequences of tornadoes in terms of property loss?

Regarding the interest of the question, it originates from an incentive to mitigate the adverse effects of tornadoes through informed policymaking and effective emergency response. For the first sub-question, by analyzing state-specific data on tornado occurrences, associated fatalities and injuries, we could identify patterns and prioritize areas for intervention. On the second sub-question, through examining the economic consequences relative to tornado characteristics, we would develop better construction standards and disaster preparedness measures indeed. Hence, our analysis not only aids in immediate response strategies but also in long-term solution against future tornado events.

b. Approach

Mini Question 1:

In terms of the first sub-question, our approach is to plot in the bubble chart to assess the consequences of tornados on each state, associated with the fatalities and injuries. The number of fatalities "fat" and the number of injuries "inj" would be placed on the x-axis and y-axis, respectively. While the bubble size would represent the number of tornados, the color of each bubble illustrates the range of number of tornados (i.e. 0-999, 1000-1999, etc.). The reason why to use this plot is to let the viewer easily understand the overall trend of how the number of tornados could impact on the fatalities and injuries on each state. In addition, the use of color coding and size differentiation facilitates a quick grasp of complex data relationships. When it comes to the visual presentation, it is ensured that the chart would include clear labeling, a coherent color scheme to differentiate tornadoes from each state, which is noted in a legend.

Mini Question 2:

In the second sub-question, the approach is to see if there is significant fluctuations in the number of tornados falling in a range of loss value. If the number of tornados with a huge loss value increases over the year, it may indicate that there are more destructive tornados occurring. By employing a grouped bar chart, one can discern both the fluctuations in the number of tornadoes across different loss categories and the potential increase in more financially destructive tornadoes over the years. These visual tools can aid in identifying long-term trends and assessing the shifting economic impact of tornadoes. To minimize cognitive load, the number of tornados is aggregated over decades. The loss groups are:

- Losses between 1 to 10 million US dollars.
- Losses between 10 to 100 million US dollars.
- Losses over 100 million US dollars.

c. Analysis

Mini Question 1:

```r
# Select the used data for plot 2.1
combined_data <- tornados %>%
  select(st, om, fat, inj) %>%
  group_by(st) %>%
  summarise(number_of_tornados = n(),
            number_of_injuries = sum(inj, na.rm = TRUE),
            number_of_fatalities = sum(fat, na.rm = TRUE),
            fatality_rate = number_of_fatalities / number_of_tornados) %>%
  filter(fatality_rate > 0.1, st != "AL") |>
  arrange(desc(fatality_rate))

# Add the column to inform the number of tornados range
combined_data <- combined_data %>%
  mutate(tornado_color_factor = case_when(
    number_of_tornados < 1000  ~ "0-999",
    number_of_tornados >= 1000 & number_of_tornados < 2000 ~ "1000-1999",
    number_of_tornados >= 2000 & number_of_tornados < 3000 ~ "2000-2999",
    number_of_tornados >= 3000 & number_of_tornados < 4000 ~ "3000-3999",
    TRUE ~ "4000+"
  ))

# Choose the color for the tornado range
colors <- c("0-999" = pal[10], "1000-1999" = pal[9], "2000-2999" = pal[7],
"3000-3999" = pal[2], "4000+" = pal[6])

# Create the bubble chart
bubble_chart <- ggplot(combined_data, aes(x = number_of_fatalities, y =
number_of_injuries, size = number_of_tornados, color = tornado_color_factor))
+
  geom_point(alpha = 0.9) +
  scale_x_continuous(expand = c(0,0), limits = c(0,600), breaks =
seq(0,600,100)) +
  scale_y_continuous(expand = c(0,0), limits = c(0,9500), breaks =
seq(0,9500,2000)) +
  scale_size_continuous(name = "Number of tornados", range = c(3,12)) +
  scale_color_manual(values = colors,
                     labels = c("0-999" = "0-999", "1000-1999" = "1000-1999",
"2000-2999" = "2000-2999", "3000-3999" = "3000-3999"),
                     name = "Tornado Range") +
  geom_label_repel(aes(label = st, color = "black"), size = 3, nudge_x = 1,
nudge_y = 1, na.rm = TRUE, show.legend = FALSE) +
  labs(x = "Number of Fatalities", y = "Number of Injuries", title = "Number
of Fatalities vs. Number of Injuries") +
  guides(color = guide_legend(override.aes = list(size = 4))) +
  theme_minimal() +
  theme_lb() +
  theme (
    axis.text = element_text(color = "black"),
    panel.grid.major = element_line(color = "#E5E5E5"),
    panel.grid.minor = element_line(color = "#E5E5E5")
```

```
  )

# Display the plot
bubble_chart
```

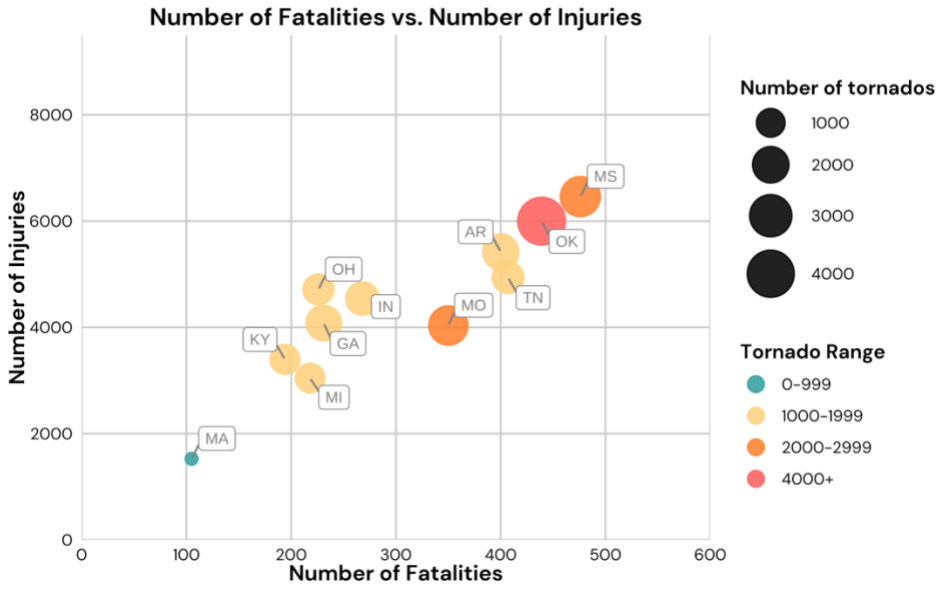

Figure 3

Mini Question 2:

```{r}
# Create a new variable for decade
new_df <- tornados %>%
  mutate(decade = case_when(
    yr >= 2020 & yr <= 2022 ~ "2020-2022",
    TRUE ~ paste0(substr(yr, 1, 3), "0s")
  )) %>%
  select(decade, loss) %>%
  group_by(decade) %>%
  summarise(
    loss_6 = sum(loss >= 1e6 & loss < 1e7, na.rm = TRUE),
    loss_7 = sum(loss >= 1e7 & loss < 1e8, na.rm = TRUE),
    loss_8 = sum(loss >= 1e8, na.rm = TRUE)
  )

# Display the resulting dataframe
new_df

# Crete long df
```

```r
long_df <- new_df %>%
  pivot_longer(
    cols = -decade, # Selects all columns except 'decade'
    names_to = "category",
    values_to = "count"
  )

colors <- c(
  "loss_6" = pal[9],
  "loss_7" = pal[7],
  "loss_8" = pal[6]
)

long_df

# Create bar chart
economic_impact <- ggplot(data = long_df, aes(x = decade, y = count, fill =
category)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = count), position = position_dodge(width = 0.9), vjust
= -0.5, size = 2.5) +
  labs(
    x = "Decade", y = "No. of Tornados",
    title = "Number of US tornados by loss range (in US dollar)"
  ) +
  scale_fill_manual(
    values = colors,
    labels = c(
      "loss_6" = "Loss 1-10 million dollar",
      "loss_7" = "Loss 10-100 million dollar",
      "loss_8" = "Loss over 100 million dollar"
    ),
    name = "Type of tornados"
  ) +
  scale_y_continuous(
    expand = expansion(mult = c(0, 0)),
    limits = c(0, 700)
  ) +
  theme_1b() +
  theme(
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank()
  ) +
  annotate("rect", xmin = 7.5, xmax = 8.5, ymin = -Inf, ymax = Inf, alpha =
0.1) +
  annotate("text", x = 8, y = 600, label = "Ongoing Decade", size = 4, angle
= 90, hjust = 1)
economic_impact
```

```

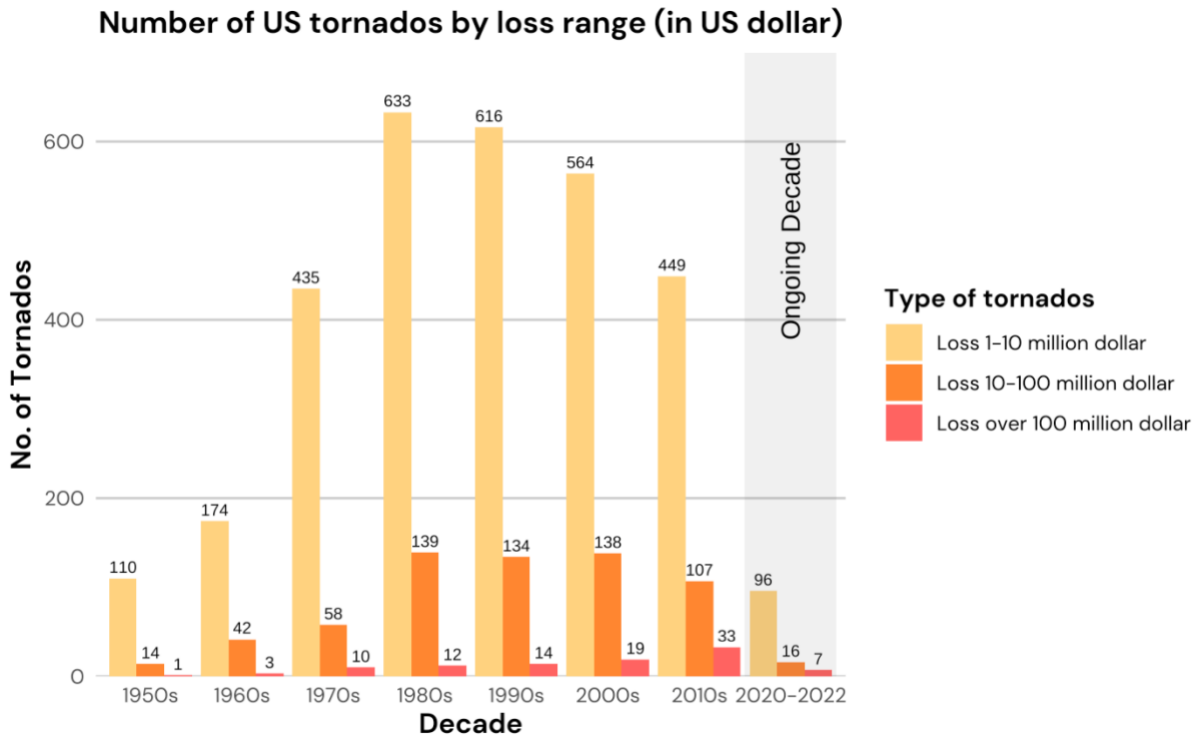


Figure 4

d. Discussion

*Mini Question 1:*

The analysis of the “Number of Fatalities vs. Number of Injuries” bubble chart reveals a relationship between tornado occurrences and the human impact on fatalities and injuries. The bubble sizes correlate with the number of tornadoes, while their position indicates the toll on human life. From the plot, the states as “MS”, “AR”, “OK”, with larger bubbles, suggest not only a higher frequency of tornados but also a significant number of injuries and fatalities.

Comparing this to reality, it is expected that a similar pattern where areas frequently hit by hurricanes experience leads to a greater human loss. The data visualization for this scenario does imply what we have aware of. Through this chart, it is recommended to provide in resource planning to minimize future losses. The absence of trends in some states may suggest effective disaster management or prepared measures. An example is the state Missouri - “MO”. In the graph, although it has occurred more hurricanes than Arkansas - “AR”, their fatalities and injuries take account for the less number since they have provided a safety protocol for citizen to prevent from the tornados indeed.
```

Mini Question 2:

The bar graph provides a striking visual representation of the economic toll from tornadoes in the United States, sorted by the extent of financial loss and categorized by decade. There's a clear upward trend in the occurrence of tornadoes causing losses from $1-10 million, particularly from the 1950s to the 1980s. However, the 1990s show a general decline in tornado numbers across all financial categories, except for the costliest ones, which continue to slightly rise in the later decade. This pattern could suggest either (1) a rise in the destructive power of tornadoes or (2) a rise in the value of the assets they impact. Nonetheless, interpretations for the 2020s should be approached with caution due to the ongoing nature of the decade. Furthermore, the absence of inflation adjustment for the dollar values might distort the actual trend in financial impact over time. Understanding these economic impacts might require a detailed state-by-state analysis, considering the varying local values and densities of properties and infrastructure.

4. Conclusion

In this comprehensive analysis of tornado occurrences, we have discovered valuable insights into the dynamics of these natural phenomena to help us understand 2 big aspects: the overall characteristics of US tornadoes over time and their economically and geographical impacts.

In terms of tornado frequency, our visualizations indicate that tornadoes often occur in spring and early summer, due to favorable weather conditions, but they can also occur at any time of the year. In addition, the most notorious region for tornadoes is "Tornado Alley," which includes parts of Texas, Oklahoma, Kansas, Nebraska, Colorado, and South Dakota. This area is particularly prone due to the collision of cold dry air from Canada with warm moist air from the Gulf of Mexico. Over the years, there is a gradual increase in tornado occurrences, as well as length and width, while the average magnitude trend depicted a steady decrease, emphasizing the prevalence of lower magnitude tornadoes with a bigger size. This finding suggests the need for continued monitoring and research to better understand the factors driving changes in tornado intensity and their impact on affected regions.

Furthermore, with the distribution of tornados, we have highlighted the increasing impacts of tornados across the US. Tornadoes causing losses of $1-10 million surged from the 1950s to the 1980s but declined in the 1990s, except for the costliest ones which slightly increased. In addition, although there is a relationship between tornado occurrences and the human impact on fatalities and injuries, measurements of these metrics differ for different states, even though some states have similar tornadoes frequencies.

Given the insights, a multifaceted solution should be implemented. Enhancing monitoring and early warning systems can significantly improve preparedness, while incorporating tornado risk assessments into urban planning and building codes can mitigate

damage in vulnerable regions. Research into the links between climate change and tornado activity is crucial for future planning. Additionally, economic and humanitarian assistance programs can support recovery efforts, and inter-state collaboration can streamline response and recovery processes. Public education campaigns and community engagement are vital in raising awareness and ensuring safety. Together, these strategies can help manage the risks associated with tornadoes, reducing their impact on communities across the United States.

In conclusion, the insights gained from our study have significant implications for disaster management and risk mitigation efforts. By leveraging data-driven approaches and advanced analytics, we can enhance our understanding of tornado behavior and develop strategies to minimize the impact of destructive tornadoes on communities and infrastructure.

5. Reference

https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-05-16/readme.md