IMPACT OF ADVERSE WEATHER CONDITIONS ON TRAFFIC ACCIDENTS IN THE UNITED STATES: A 2017-2022 ANALYSIS

DAAN 703 Technical Report

Sai Manigopal Reddy Kanumathireddy, Venkata Aditya Kantipudi

2024-11-20

```r
library(tidyr)
library(knitr)
library(tidyverse)
library(readr)
library(dplyr)
```

INTRODUCTION

Research Questions of Interest

1. How do adverse weather conditions, such as rain, snow, and fog, impact the frequency and severity of traffic accidents across the USA from 2017 to 2022?
2. What additional factors, such as light conditions, driver behavior, and road conditions, interact with adverse weather to influence accident outcomes?

Importance and Relevance

These questions are crucial for identifying safety measures to reduce weather-related accidents. Bad weather impairs visibility and road traction, increasing accident risks. Understanding how factors like vehicle type and driver behavior interact with weather aids in developing effective safety policies and raising public awareness.

Literature Review

Research shows a strong link between adverse weather and traffic accidents.(Xing, 2019) found non-linear impacts of factors like temperature on accident rates. (Becker, 2022) demonstrated that rain, snow, and glare affect crash types differently. (Pińskwar, 2024) highlighted how extreme weather increases driver cognitive load. (Tobin, 2022) emphasized the role of winter conditions like snow and low visibility in fatal crashes. Our study builds on prior research by analyzing recent U.S. data, focusing exclusively on crashes under adverse weather to provide updated insights into weather-related road safety.

DATA DESCRIPTION

Source

The data for this study is sourced from the Fatality Analysis Reporting System (FARS), covering traffic accident data from 2017-2022 and weather data is obtained from `Metostat` Python package.

Table 1: Variable Description Table

Variable	Description	Type
STATENAME	Name of the state where the accident occurred.	Categorical
ST_CASE	Unique identifier for each accident case.	Numeric (int)
PERMVIT	Count of people involved in the accident.	Numeric (int)
COUNTYNAME, CITYNAME	Name of the county and city where the accident occurred.	Categorical
DATE	Date of the accident (Year, Month, Day).	Date
HOUR, MINUTE	Hour and Minute at which the accident occurred.	Numeric (int)
RUR_URBNAME	Indicates if the accident location is rural or urban.	Categorical
LATITUDE, LONGITUDE	Latitude and Longitude of the accident location.	Numeric (int)
LGT_CONDNAME	Light condition at the time of the accident (e.g., Daylight, Dark).	Categorical
WEATHERNAME	Type of adverse weather condition (e.g., Rain, Snow, Fog).	Categorical
FATALS	Number of fatalities resulting from the accident.	Numeric (int)
MAKENAME	Vehicle make involved in the accident (e.g., Ford, Toyota).	Categorical
TRAV_SP	Vehicle travel speed at the time of the accident (mph).	Numeric (double)
DR_DRINK	Indicator of whether the driver was under the influence of alcohol.	Categorical
VSURCONDNAME	Vehicle surface condition (e.g., Wet, Icy, Dry).	Categorical
AGE	Age of the driver involved in the accident.	Numeric (int)
SEXNAME	Gender of the driver (e.g., Male, Female).	Categorical
INJ_SEV	Injury severity level (e.g., Fatal, Serious Injury, Minor Injury).	Categorical
DRIVERRFNAME	Driver-related factors contributing to the crash (e.g., Speeding, Distracted).	Categorical
VISIONNAME	Vision impairment status of the driver.	Categorical
TEMP	Temperature at the time of the accident (°C).	Numeric (double)
DWPT	Dew point temperature, indicating humidity levels (°C).	Numeric (double)
RHUM	Relative humidity percentage (%).	Numeric (int)
PRCP	Precipitation amount (mm).	Numeric (double)
WSPD	Wind speed at the time of the accident (m/s).	Numeric (double)
PRES	Atmospheric pressure (hPa).	Numeric (double)

Data Cleaning Steps

1. Data Consolidation: Loaded all the relevant files and merged them into a single file and and dropped irrelevant columns (e.g., road ownership, work zone, school bus) from FARS accident file.

```r
library(plyr)
# List all CSV files in the folder
file_paths <- list.files(pattern = "^accident\\d{4}\\.csv$")

# Read and combine files
df <- plyr::ldply(file_paths, function(file) {
  # Read each file and select relevant columns
  read.csv(file) %>%
    dplyr::select(
      STATENAME, ST_CASE, PERMVIT, COUNTYNAME, CITYNAME, MONTH, DAY,
      DAY_WEEK, DAY_WEEKNAME, YEAR, HOUR, MINUTE, RUR_URBNAME,
      LATITUDE, LONGITUD, LGT_COND, LGT_CONDNAME, WEATHER, WEATHERNAME, FATALS
    )
})

# Detach plyr to prevent conflicts
detach("package:plyr", unload = TRUE)
```

2. Data Filtering: Filtered the dataset to only include adverse weather conditions and excluded rows with clear or unknown weather conditions.

3. Column Standardization: Converted all the columns to lowercase and properly named all the columns.

4. Field Cleanup: Cleaned fields like `countyname` to remove special characters such as parentheses (`(` and `)`).

```r
# Filtering rows based on weather column
library(plyr)
df<-df%>%
  filter(!(WEATHERNAME %in% c("Clear", "Not Reported","Other","Reported as Unknown")))

# Converting to lower case
colnames(df)<-tolower(colnames(df))

df <- rename(df, c(longitud = "longitude"))

detach("package:plyr", unload = TRUE)
```

```r
df$countyname <- sub("\\s*\\(.*$", "", df$countyname)
# checking for NA values
colSums(is.na(df))
```

```
##    statename      st_case      permvit   countyname     cityname        month
##            0            0            0            0            0            0
##          day     day_week day_weekname         year         hour       minute
##            0            0            0            0            0            0
##  rur_urbname     latitude    longitude     lgt_cond lgt_condname      weather
##            0            0            0            0            0            0
##  weathername       fatals
##            0            0
```

5. Outlier Handling: Identified and replaced outliers in `hour` and `minute` columns with `NA` values. Detected outliers in `latitude` and `longitude` columns and corrected coordinates using `statename` and `countyname` with the **tidygeocoder** package.

```r
# identifying number of outlier rows in hour and minute
outliers <- df %>%
    filter(hour < 0 | hour > 23 | minute < 0 | minute > 59)
nrow(outliers)
```

```
## [1] 353
```

```r
# replacing outlier rows of hour and minute with NA values
df <- df %>%
    mutate(
        hour = ifelse(hour < 0 | hour > 23, NA, hour),
        minute = ifelse(minute < 0 | minute > 59, NA, minute)
    )

# identifying number of outlier rows in latitude and longitude and identifying their location.
outliers <- df %>%
    filter( latitude < -90 | latitude > 90 | longitude < -180 | longitude> 180) %>%
  select(statename,cityname,countyname,latitude,longitude)

#outliers
```

```r
library(dplyr)
library(tidygeocoder)

# Combine location and filter invalid rows
df <- df %>%
  mutate(combined_location = ifelse(
    countyname == "NOT APPLICABLE",
    statename,
    paste(countyname, statename, sep = ", ")
  ))

num_rows = nrow(df)

# creating ID column vector
row_id <- c(1:num_rows)

# binding id column to the data frame
df <- cbind(row_id , df)

# Filter invalid rows (latitude or longitude out of range)
invalid_rows <- df %>%
  filter(latitude < -90 | latitude > 90 | longitude < -180 | longitude > 180)

# Geocode invalid rows
geocoded_rows <- invalid_rows %>%
  geocode(combined_location, method = 'osm', lat = new_latitude, long = new_longitude)
```

6. Dataset Integration: Combined FARS accident data with Metostat weather data by matching location (`latitude`, `longitude`) and time details (hour, minute, and date) to create a comprehensive dataset.

```r
# Merge geocoded rows back to original dataset
df <- df %>%
  left_join(geocoded_rows %>% select(row_id, new_latitude, new_longitude), by = "row_id") %>%
  mutate(
    latitude = coalesce(new_latitude, latitude),
    longitude = coalesce(new_longitude, longitude)
  ) %>%
  select(-new_latitude, -new_longitude)  # Drop temporary columns

# write_csv(df,path="accident2017-2022.csv")
```

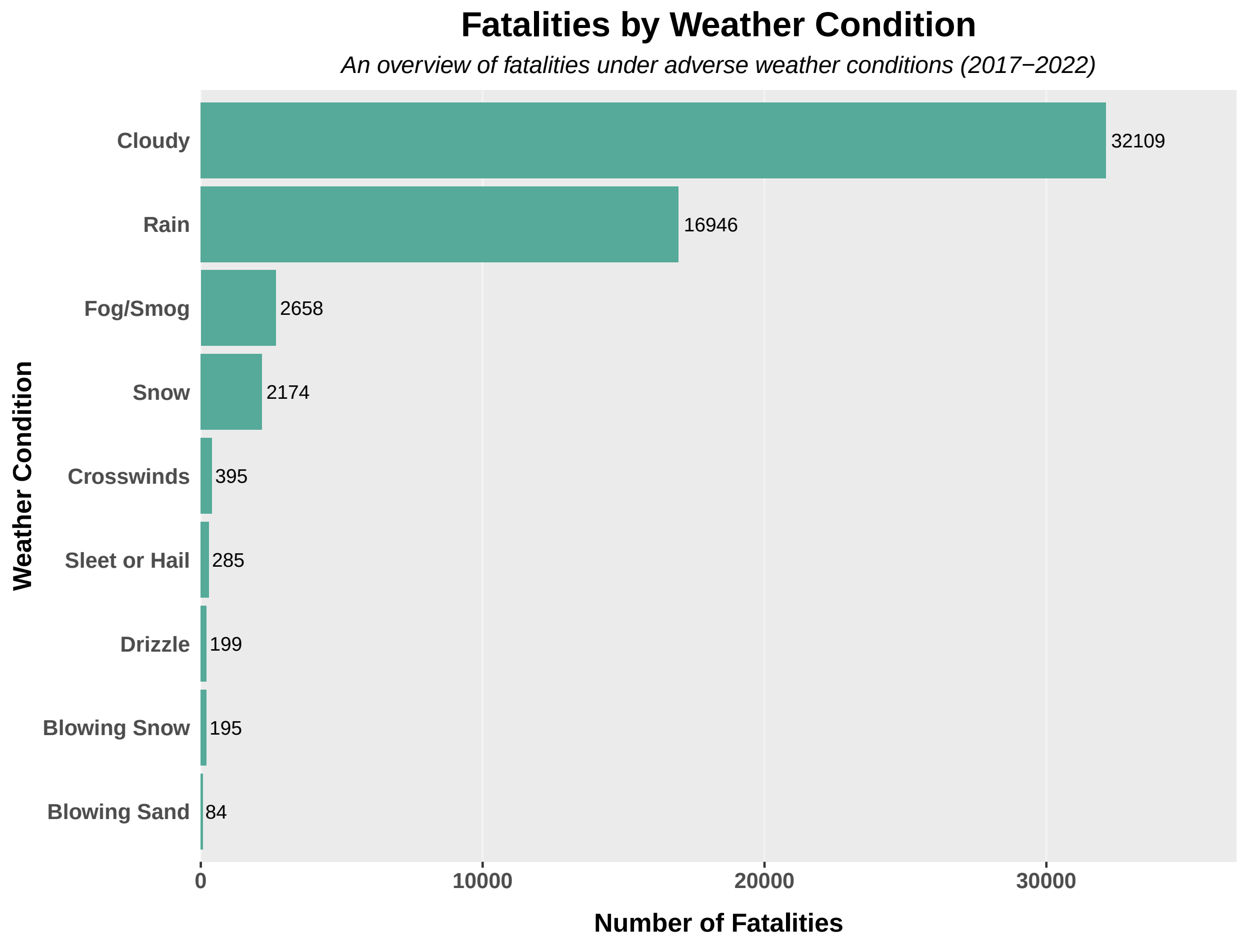
DESCRIPTIONS OF VISUALIZATIONS

```r
library(tidyverse)
library(readr)
library(dplyr)
library(readr)
library(ggplot2)
library(tibble)
library(plotly)
```

```r
df<-read_csv("accident_weather_2017-2022.csv")
df<-df%>%
  filter(weathername!="Unknown")%>%
  select(-time,-snow,-wdir,-wpgt,-tsun,-coco)
weather_fatals<-df%>%
  mutate(weathername =recode(weathername,"Blowing Sand, Soil, Dirt" = "Blowing Sand","Freezing Rain or 
 group_by(weathername)%>%
summarize(total_fatals=sum(fatals,na.rm=TRUE))
```

```r
ggplot(weather_fatals, aes(x = total_fatals, y = reorder(weathername, total_fatals))) +
  geom_text(aes(label = total_fatals), hjust = -0.1, size = 3.2, color = "black") +
  geom_col(fill = "#5A9") +
  labs(
    title = "Fatalities by Weather Condition",
    subtitle = "An overview of fatalities under adverse weather conditions (2017-2022)",
    x = "Number of Fatalities",
    y = "Weather Condition"
  ) +
  scale_x_continuous(expand = expansion(mult = c(0, 0.05)), limits = c(0, 35000)) +
  theme(
    axis.title = element_text(size = 12, face = "bold"),
    axis.title.x = element_text(vjust = -1.5),
    axis.text.y = element_text(size = 10, face = "bold"),
    axis.text.x = element_text(size = 10,face="bold"),
    axis.ticks.y = element_blank(),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 10.5, face = "italic"),
    panel.grid.minor = element_blank(),
    panel.grid.major.y = element_blank(),
```

```
  panel.grid.major.x = element_line(color = "gray95")
)
```



Key Takeaways

This bar chart highlights that Cloudy conditions account for the highest number of fatalities (32,109), followed by Rain (16,946), which together dominate fatality counts. Interestingly, Fog/Smog and Snow also contribute significantly but at much lower levels than cloudy or rainy conditions. Rare conditions like Blowing Sand and Drizzle show minimal fatalities, suggesting lower risk but possibly lower exposure. These findings emphasize the disproportionate fatality risk under common adverse weather conditions like cloudy and rainy weather.

Tools and Techniques

The graphic was created using R and the ggplot2 package, showcasing fatalities by weather condition from 2017-2022. Data manipulation was performed with functions from the dplyr package, including filter, select, mutate, and group_by, to clean, transform, and summarize the data. Specific coding techniques included recoding weather conditions using recode and summarizing fatalities with summarize. The visualization leverages geom_col for a bar chart, with geom_text to label fatalities directly on the bars for clarity. Graphical choices included a custom "#5A9" fill color for the bars, a limited x-axis range for focus, and a reordered y-axis to rank weather conditions by fatalities, enhancing interpretability and visual appeal.

```r
pf<-read_csv("people2017-2022.csv")
```

```r
inj_weather<-df%>%
  left_join(pf,by="caseyear")%>%
  select(weathername,INJ_SEVNAME,INJ_SEV,AGE,SEXNAME,YEAR)
```

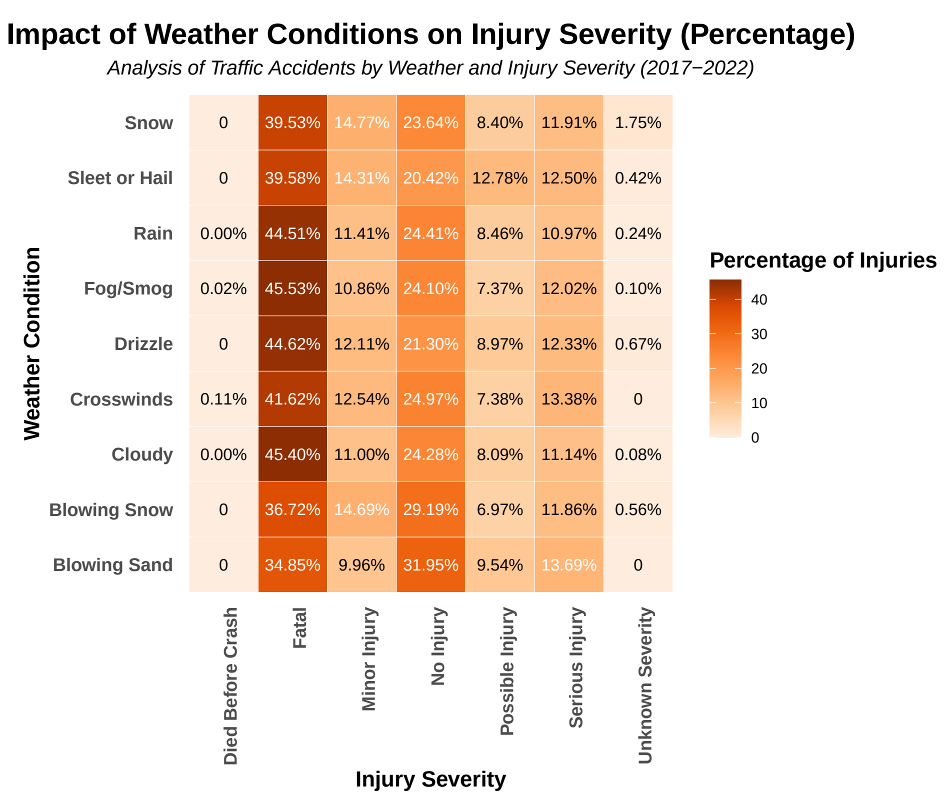
```r
library(tidyr)

inj_weather_summary <- inj_weather %>%
  filter(INJ_SEV != 9) %>%
  mutate(
    weathername = recode(weathername,
      "Blowing Sand, Soil, Dirt" = "Blowing Sand",
      "Freezing Rain or Drizzle" = "Drizzle",
      "Fog, Smog, Smoke" = "Fog/Smog",
      "Severe Crosswinds" = "Crosswinds"
    ),
    INJ_SEVNAME = recode(INJ_SEVNAME,
      "Died Prior to Crash*" = "Died Before Crash",
      "Fatal Injury (K)" = "Fatal",
      "Injured, Severity Unknown" = "Unknown Severity",
      "No Apparent Injury (O)" = "No Injury",
      "Possible Injury (C)" = "Possible Injury",
      "Suspected Minor Injury (B)" = "Minor Injury",
      "Suspected Serious Injury (A)" = "Serious Injury"
    )
  ) %>%
  group_by(weathername) %>%
  mutate(total_count = n()) %>%
  group_by(weathername, INJ_SEVNAME) %>%
  summarize(
    count = n(),
    percentage = (count / first(total_count)) * 100,
    .groups = "drop"
  ) %>%
  complete(weathername, INJ_SEVNAME, fill = list(count = 0, percentage = 0))

plot_injury_weather <- function(data, value_col, value_label, title, fill_name, palette = "Oranges") {
  ggplot(data, aes(x = INJ_SEVNAME, y = weathername, fill = !!sym(value_col))) +
    geom_tile(color = "white") +
    geom_text(
      aes(label = ifelse(!!sym(value_col) == 0, "0", sprintf(value_label, !!sym(value_col)))),
      color = ifelse(data[[value_col]] > max(data[[value_col]]) * 0.3, "white", "black"),
      size = 3
    ) +
    scale_fill_distiller(
      palette = palette,
      direction = 1,
      name = fill_name
    ) +
    labs(
      title = title,
      subtitle = "Analysis of Traffic Accidents by Weather and Injury Severity (2017-2022)",
```

```
      x = "Injury Severity",
      y = "Weather Condition"
    ) +
    coord_fixed(ratio = 0.8) +
    theme_minimal() +
    theme(
      axis.text.x = element_text(angle = 90, hjust = 1, size = 10, face = "bold"),
      axis.text.y = element_text(size = 10, face = "bold"),
      legend.title = element_text(size = 12, face = "bold"),
      legend.text = element_text(size = 8),
      plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
      plot.subtitle = element_text(hjust = 0.5, size = 10.5,face="italic"),
      axis.title.x = element_text(size = 12, face = "bold"),
      axis.title.y = element_text(size = 12, face = "bold"),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank()
    )
}

plot_injury_weather(
  inj_weather_summary,
  "percentage",
  "%.2f%%",
  "Impact of Weather Conditions on Injury Severity (Percentage)",
  "Percentage of Injuries",
  palette = "Oranges"
)
```



Key Takeaways

The heatmap reveals that Fog/Smog has the highest percentage of fatalities (45.53%), followed closely by Rain and Cloudy conditions. Surprisingly, Blowing Sand, Blowing Snow, and Crosswinds show relatively higher percentages of "No Injury" and "Serious Injury" cases, suggesting that these conditions, while less fatal, are still associated with significant accident outcomes. These findings highlight the varying impact of weather conditions on traffic accident severity, with visibility and surface hazards playing a crucial role.

Tools and Techniques

The heatmap was created using R, leveraging the ggplot2 package for its flexible plotting capabilities. Data preparation involved using dplyr for tasks like recoding variables, filtering, grouping, and summarizing data, as well as the tidyr package's complete function to handle missing combinations of weather and injury severity categories. The geom_tile function was used to create the heatmap, with geom_text overlaying the injury percentages directly on the tiles for enhanced readability. A diverging "Oranges" palette was selected to intuitively represent the percentage of injuries, with white or black text dynamically adjusted for contrast.

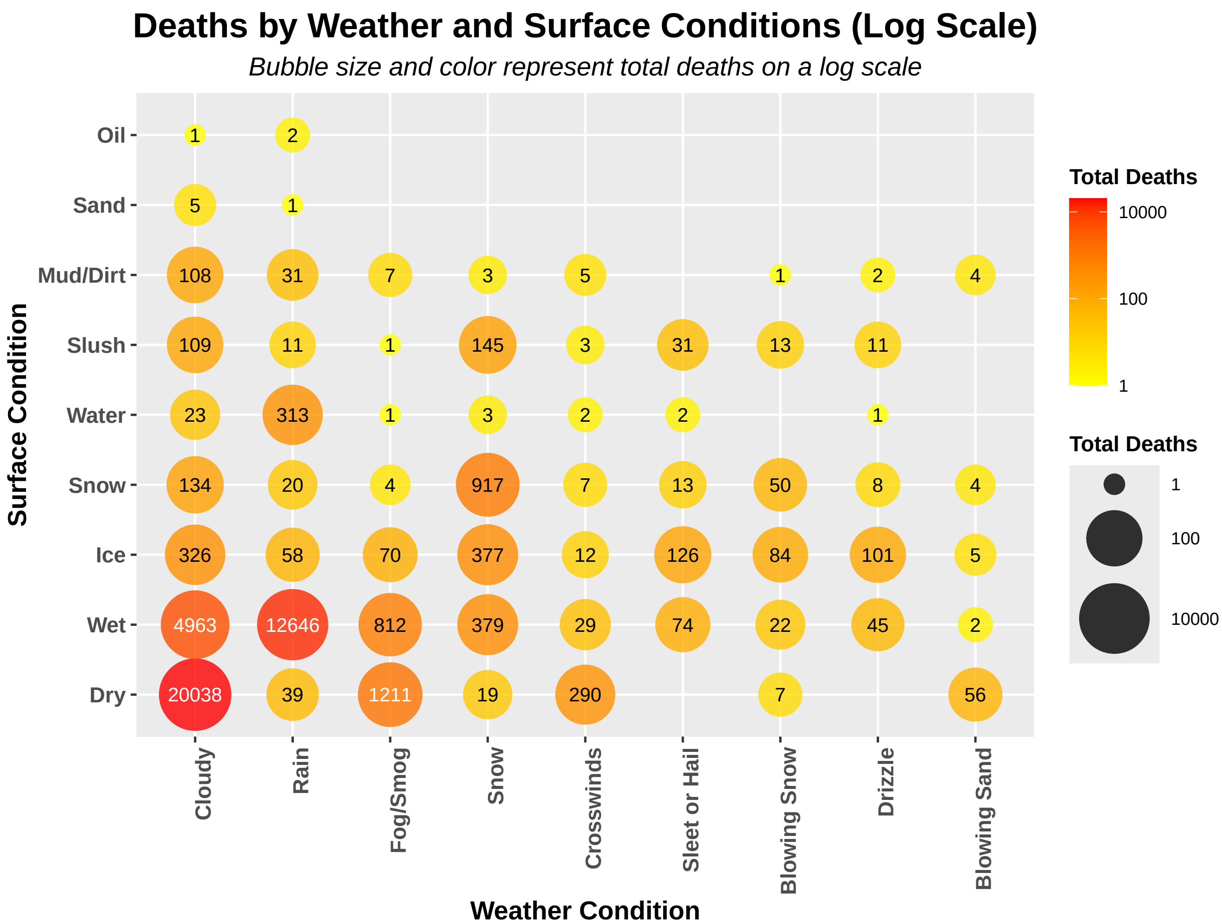
```
vf<-read_csv("vehicle2017-2022.csv")
vf<-vf%>%
  filter(!VSURCONDNAME %in% c("Reported as Unknown","Other","Not Reported","Unknown","Non-Trafficway or

weather_vehicle<-df%>%
  left_join(vf,by="caseyear") %>%
  select(weathername,BODY_TYPNAME,TRAV_SP,VSURCOND,VSURCONDNAME,DEATHS,DR_DRINK,YEAR)

colnames(weather_vehicle)<-tolower(colnames(weather_vehicle))
```

```
bubble_data <- weather_vehicle %>%
  filter(!is.na(vsurcondname)) %>%
  group_by(weathername, vsurcondname) %>%
  summarize(
    deaths = sum(deaths, na.rm = TRUE), .groups = "drop") %>%
  mutate(
    weathername = recode(weathername,
      "Blowing Sand, Soil, Dirt" = "Blowing Sand",
      "Freezing Rain or Drizzle" = "Drizzle",
      "Fog, Smog, Smoke" = "Fog/Smog",
      "Severe Crosswinds" = "Crosswinds"
    ),
    vsurcondname = recode(vsurcondname,
      "Non-Trafficway or Driveway Access" = "Driveway",
      "Water (Standing or Moving)" = "Water",
      "Mud, Dirt or Gravel" = "Mud/Dirt",
      "Reported as Unknown" = "Unknown",
      "Ice/Frost" = "Ice"
    ),
    weathername = reorder(weathername, -deaths),
    vsurcondname = reorder(vsurcondname, -deaths)
  )
ggplot(bubble_data, aes(x = weathername, y = vsurcondname, size = deaths, color = deaths)) +
  geom_point(alpha = 0.8) +
  geom_text(
    aes(label = ifelse(deaths > 0, deaths, ""),size=deaths),
```

```
    size = 3,
    color = ifelse(bubble_data$deaths > 1000, "white", "black"),
    show.legend = FALSE
  ) +
  scale_size(
    range = c(4, 15),
    name = "Total Deaths",
    trans = "log10"
  ) +
  scale_color_gradientn(
    colors = c("yellow", "red"),
    trans = "log10",
    name = "Total Deaths"
  ) +
  labs(
    title = "Deaths by Weather and Surface Conditions (Log Scale)",
    subtitle = "Bubble size and color represent total deaths on a log scale",
    x = "Weather Condition",
    y = "Surface Condition"
  ) +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, size = 10, face = "bold"),
    axis.text.y = element_text(size = 10, face = "bold"),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5, size = 12, face = "italic"),
    legend.title = element_text(size = 10, face = "bold"),
    legend.text = element_text(size = 8),
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold")
  )
```



Key Takeaways

This bubble chart reveals several key insights about fatalities by weather and surface conditions. Dry surfaces under Cloudy conditions have the highest fatalities (20,038), followed by Wet surfaces during Rain (12,646), highlighting the danger of these common conditions. Interestingly, Ice and Snow surfaces under adverse weather, though less frequent, still show significant fatalities, with Snow on snow-covered surfaces contributing notably (917 deaths). Rare conditions like Slush and Mud/Dirt have lower fatalities but demonstrate risk in specific scenarios. This visualization emphasizes the interplay between weather and surface conditions, with wet and dry surfaces being the most hazardous under cloudy and rainy weather.

Tools and Techniques

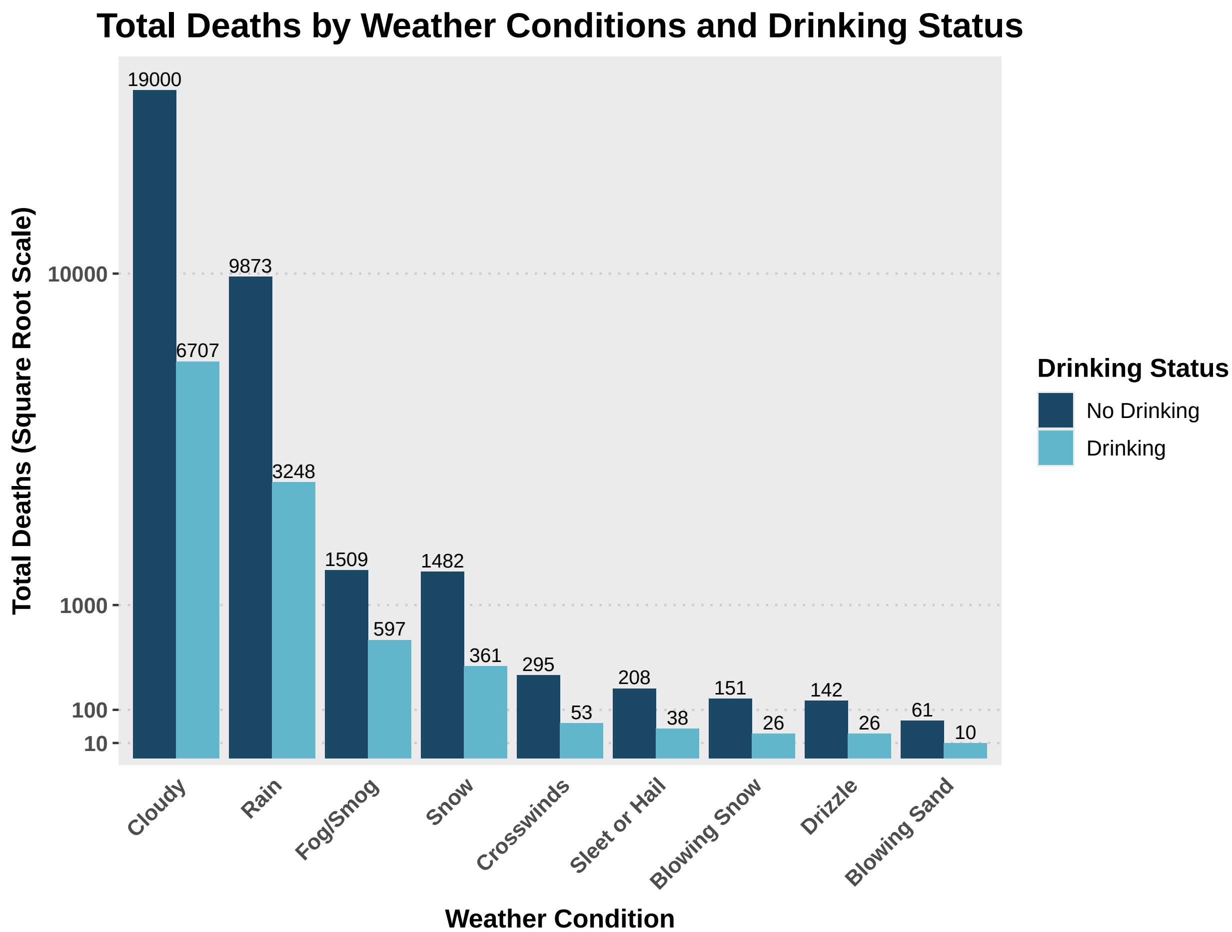
The visualization was created using ggplot2, leveraging geom_point for bubbles and geom_text to label death counts. Data manipulation, such as filtering, grouping, and recoding categories, was performed with dplyr to ensure clarity and relevance. A log scale was applied to bubble size and color using scale_size and scale_color_gradientn, which enabled a consistent representation of the wide range of death counts. Graphical choices included rotating x-axis labels for readability, bold styling for emphasis, and a red-yellow gradient to intuitively convey severity, enhancing both aesthetic appeal and interpretability.

```
bar_data <- weather_vehicle %>%
  filter(!is.na(dr_drink), !is.na(weathername), !is.na(deaths)) %>%
  group_by(weathername, dr_drink) %>%
```

```
  summarize(total_deaths = sum(deaths, na.rm = TRUE), .groups = "drop") %>%
  mutate(
    dr_drink = recode(dr_drink, `0` = "No Drinking", `1` = "Drinking"),
    weathername = recode(weathername,
      "Blowing Sand, Soil, Dirt" = "Blowing Sand",
      "Freezing Rain or Drizzle" = "Drizzle",
      "Fog, Smog, Smoke" = "Fog/Smog",
      "Severe Crosswinds" = "Crosswinds"
    )
  )
ggplot(bar_data %>%
    mutate(weathername = reorder(weathername, -total_deaths),
           dr_drink = factor(dr_drink, levels = c("No Drinking", "Drinking"))), aes(x = weathername, y =
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(
    position = position_dodge(width = 0.9),
    vjust = -0.3,
    size = 3.2
  ) +
  scale_fill_manual(values = c("#1b4965","#62b6cb"), name = "Drinking Status") +
  scale_y_continuous(
    trans = "sqrt",
    name = "Total Deaths (Square Root Scale)",
    breaks = c(10,100, 1000, 10000),
    expand = expansion(mult = c(0.01, 0.05))

  ) +
  labs(
    title = "Total Deaths by Weather Conditions and Drinking Status",
    x = "Weather Condition",
    y = "Total Deaths (Square Root Scale)"
  ) +

  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10, face = "bold"),
    axis.text.y = element_text(size = 10, face = "bold"),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    legend.title = element_text(size = 12, face = "bold"),
    legend.text = element_text(size = 10),
    axis.title.x = element_text(size = 12, face = "bold", vjust = -0.5),
    axis.title.y = element_text(size = 12, face = "bold", vjust = 2),
    panel.grid.major.y = element_line(color = "gray80",linetype="dotted"),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank(),
    axis.ticks.x = element_blank()
  )
```



Key Takeaways

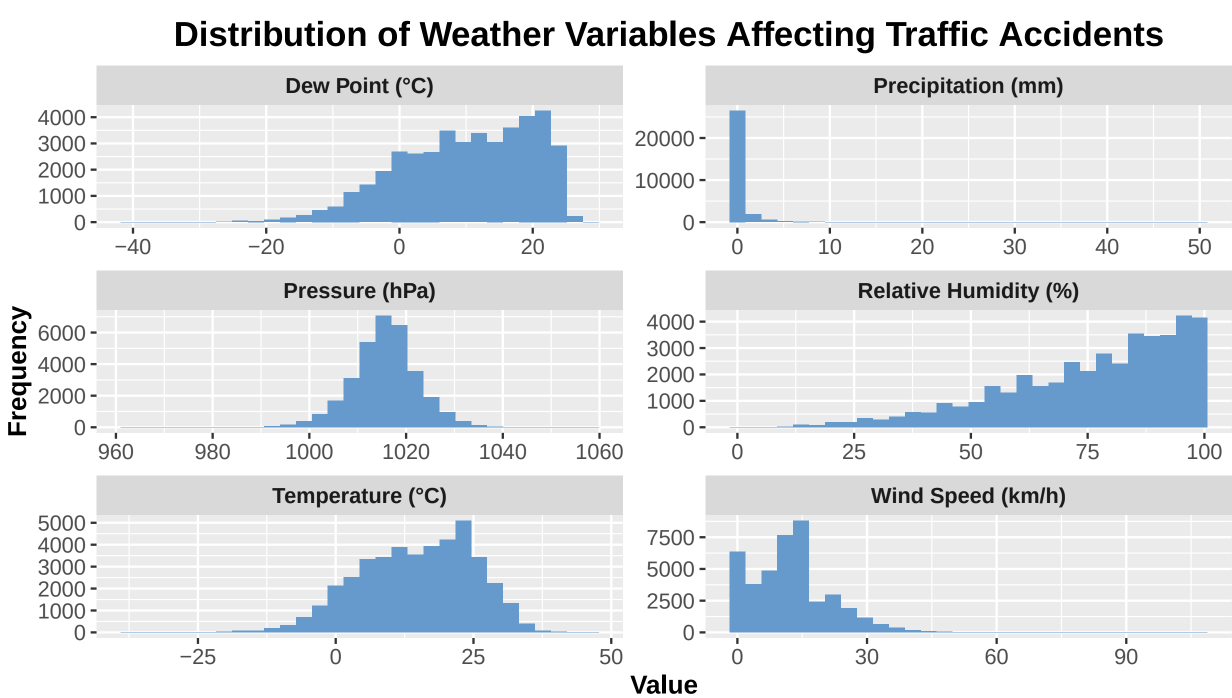
This grouped bar chart illustrates the total deaths under different weather conditions, categorized by drinking status, using a square root scale. Cloudy weather conditions record the highest fatalities, with 19,000 under non-drinking and 6,707 under drinking status, followed by Rain and Fog/Smog. The chart reveals a consistent pattern where non-drinking fatalities outnumber drinking fatalities across conditions suggesting people who drink do not drive mostly. These insights highlight the critical impact of driver sobriety and weather conditions on fatal accidents.

Tools and Techniques

The chart was built using ggplot2 in R, leveraging dplyr for data grouping and summarization by weather condition and drinking status. Bars were created with geom_bar() and separated using position_dodge() for clarity. The square root scale was applied via scale_y_continuous(trans = "sqrt") to manage value distribution. Additional customizations, such as a manually defined color palette and precise label positioning with geom_text(), enhanced readability and visual appeal.

```
df_long <- df %>%
  select(temp, dwpt, rhum, wspd, prcp, pres) %>%
  pivot_longer(
    cols = everything(),
    names_to = "Variable",
```

```
    values_to = "Value"
  )
custom_labels <- c(
  temp = "Temperature (°C)",
  dwpt = "Dew Point (°C)",
  rhum = "Relative Humidity (%)",
  wspd = "Wind Speed (km/h)",
  prcp = "Precipitation (mm)",
  pres = "Pressure (hPa)"
)
ggplot(df_long, aes(x = Value)) +
  geom_histogram(bins = 30, fill = "#6699CC", alpha = 2) +
  facet_wrap(
    ~Variable,
    scales = "free",
    ncol = 2,
    labeller = labeller(Variable = custom_labels)
  ) +
  labs(
    title = " Distribution of Weather Variables Affecting Traffic Accidents",
    x = "Value",
    y = "Frequency"
  ) +
  theme(
    strip.text = element_text(size = 10, face = "bold"),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 12,face="bold"),
    axis.text = element_text(size = 10)
  )
```



Key Takeaways

The histograms reveal distinct patterns for weather variables impacting traffic accidents. Dew Point and Temperature exhibit near-normal distributions, while Precipitation is skewed toward low values, indicating its rarity during accidents. Pressure is concentrated between 1000 and 1025 hPa, reflecting stable atmospheric conditions. Relative Humidity and Wind Speed show skewed distributions, with higher humidity and lower wind speeds being more common during traffic incidents.

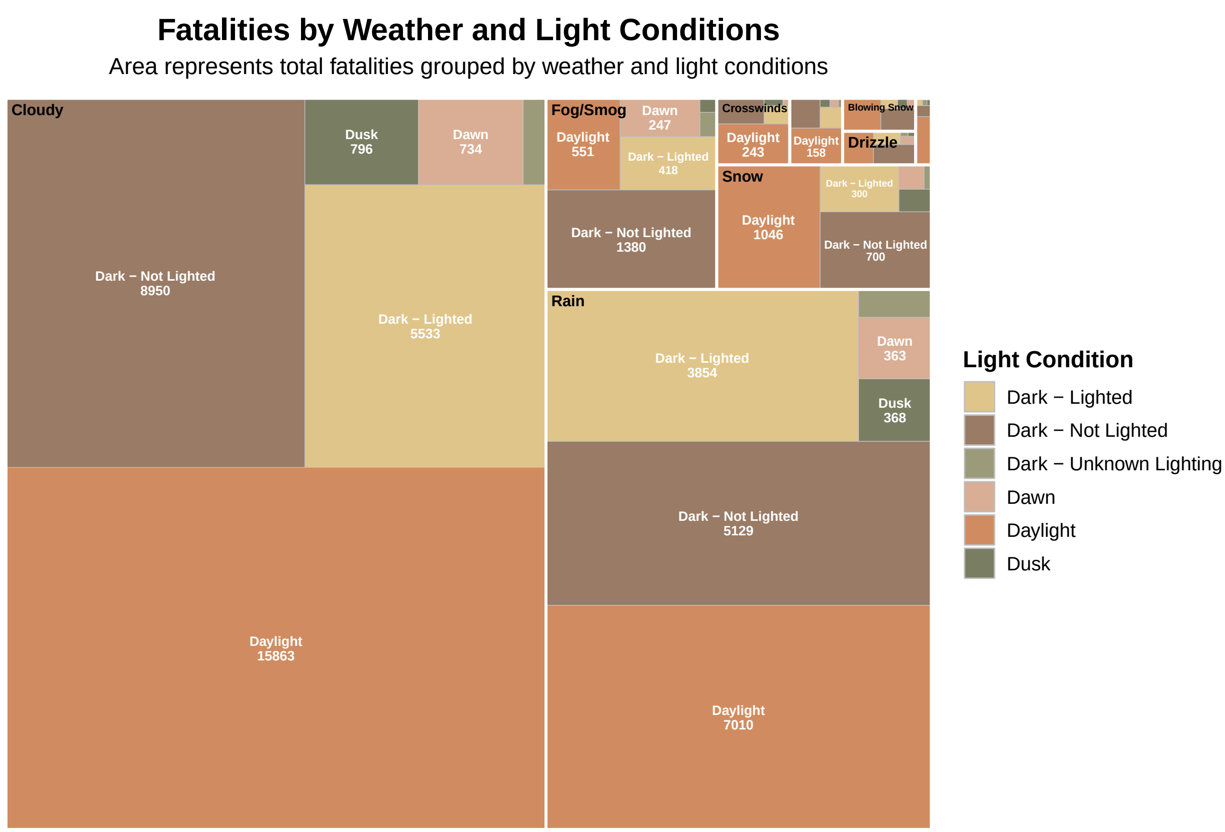
Tools and Techniques

Data was transformed into a long format using pivot_longer to streamline plotting multiple variables. Histograms were created with geom_histogram in ggplot2, a R package, using 30 bins for consistent granularity. Faceting with facet_wrap enabled individual scaling for each variable, improving interpretability.

```r
library(treemapify)
treemap_data <- df %>%
  filter(!is.na(weathername), !is.na(lgt_condname), !is.na(fatals)) %>%
  filter(!lgt_condname %in% c("Other", "Reported as Unknown", "Not Reported", "Unknown")) %>%
  group_by(weathername, lgt_condname) %>%
  summarize(total_fatals = sum(fatals, na.rm = TRUE), .groups = "drop") %>%
  mutate(
    weathername = recode(weathername,
      "Blowing Sand, Soil, Dirt" = "Blowing Sand",
      "Freezing Rain or Drizzle" = "Drizzle",
      "Fog, Smog, Smoke" = "Fog/Smog",
      "Severe Crosswinds" = "Crosswinds"
    )
  )

ggplot(treemap_data, aes(
  area = total_fatals,
  fill = lgt_condname,
  subgroup = weathername
)) +
  geom_treemap() +
  geom_treemap_subgroup_border(color = "white", size = 2) +
  geom_treemap_subgroup_text(
    aes(label = weathername),
    color = "black",
    place = "topleft",
    grow = FALSE,
    size = 8,
    fontface = "bold"
  ) +
  geom_treemap_text(
    aes(label = sprintf("%s\n%d", lgt_condname, total_fatals)),
    color = "white",
    place = "center",
    grow = FALSE,
    size = 7,
    fontface = "bold"
  ) +
  scale_fill_manual(
```

```
    values = c(
      "Daylight" = "#d08c60",
      "Dark - Not Lighted" = "#997b66",
      "Dark - Lighted" = "#e0c58a",
      "Dawn" = "#d9ae94",
      "Dusk" = "#797d62",
      "Dark - Unknown Lighting" = "#9b9b7a"
    ),
    name = "Light Condition"
  ) +
  labs(
    title = "Fatalities by Weather and Light Conditions",
    subtitle = "Area represents total fatalities grouped by weather and light conditions",
    fill = "Light Condition"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 12, margin = margin(b = 10)),
    legend.title = element_text(size = 12, face = "bold"),
    legend.text = element_text(size = 10),
    legend.key.size = unit(1.2, "lines")
  )
```



Key Takeaways

This treemap highlights that fatalities under Cloudy and Rainy weather conditions dominate, with Daylight having the highest number of fatalities under Cloudy (15,563) and Rain (7,010). Interestingly, Dark - Not Lighted conditions contribute significantly under both weather types, especially for cloudy conditions (8,950 fatalities), suggesting poor visibility is a critical factor. Rare conditions like Snow and Fog/Smog show fewer fatalities, but Dark - Not Lighted still plays a notable role in these cases. The data underscores the combined risks of adverse weather and inadequate lighting, particularly under cloudy and rainy conditions.

Tools and Techniques

The treemap was created using the treemapify package, and data was preprocessed using dplyr to group and filter relevant conditions. Custom colors were assigned to light conditions with scale_fill_manual, ensuring clarity and visual distinction. Subgroup borders and labels for weather categories were enhanced with geom_treemap_subgroup_border and geom_treemap_subgroup_text, improving readability. The design prioritizes interpretability through bold text, white subgroup borders, and a balanced color palette.

CONCLUSIONS

Practical Findings

The analysis reveals critical insights into the impact of adverse weather conditions on traffic accidents. Fog and smog conditions have the highest percentage of fatalities, followed closely by rain and cloudy weather, highlighting the significant roles of reduced visibility and traction in accident severity. Cloudy and rainy conditions, despite being common, disproportionately contribute to total fatalities, emphasizing their high risk. Surface conditions further illustrate this trend, with dry and wet surfaces being the most hazardous under adverse weather, while less frequent conditions like ice and snow still result in significant fatalities. Poor lighting, particularly "Dark - Not Lighted" conditions, exacerbates the fatality risk, especially under cloudy and rainy weather. Interestingly, fatalities among non- drinking drivers consistently outnumber those involving drinking, reinforcing the importance of sobriety in reducing fatal accidents. Additionally, the distribution of weather variables shows that while some, like temperature and dew point, follow normal distributions, others, such as precipitation and wind speed, are skewed, reflecting the influence of extreme weather on accident occurrences.

Design Weaknesses

While the study offers valuable insights, it is limited by its focus solely on fatal accidents, excluding minor and non- fatal incidents that could provide critical patterns. The reliance on point-in-time weather data fails to capture rapid changes in conditions, while crucial variables such as traffic density, seasonal patterns, and vehicle safety features remain underexplored. Additionally, the visualizations, though effective, could benefit from interactive and temporal elements to better analyze multidimensional relationships. Future research should include granular time-based analysis, geographical clustering, and multivariate approaches to examine interaction effects more comprehensively.Incorporating variables like road infrastructure quality, driver experience, and emergency response times, along with predictive modeling and robust statistical controls, would enhance the study's applicability and provide actionable insights for improving traffic safety under adverse weather conditions.