Aanand Aggarwal

Traffic Accident Severity Analysis Using US Accidents Dataset

Project Topic: My final project will analyze the factors contributing to the severity of traffic accidents in the United States, using the 'US Accidents Dataset' from Kaggle. With millions of accident records from 2016 to 2023 across 49 states, the dataset offers a detailed view of traffic incidents and related variables. It includes weather conditions, road characteristics, and geographic data. I aim to identify critical variables and patterns that contribute to higher accident severities.

What is Known: Traffic accidents are a major cause of injury, property damage, and loss of life, impacting communities across the globe. Research shows that several key factors influence where and when accidents happen, as well as their severity. For example, accidents are more common during rush hours when traffic is heaviest, and weekends or holidays often see an increase due to higher travel volume (Adeyemi et al., 2021) Weather conditions like rain, snow, and fog also play a major role, reducing visibility and making roads more hazardous (Federal Highway Administration, 2016). Geographically, accidents tend to cluster in urban areas with dense traffic and busy intersections (Insurance Institute for Highway Safety, 2017). Rural areas often experience more severe accidents due to higher speeds and slower emergency response times (Safe Transportation Research and Education Center, 2024). Another common factor is poor road infrastructure, like insufficient signage or a lack of traffic signals (CDC, 2024). While these trends are well documented, analyzing a large-scale dataset allows a deeper dive into how these factors play out across different regions, environments, and timeframes.

Questions to Investigate:

1. What are the patterns in accident severity levels, and are severe accidents more common under specific conditions?
2. How do accidents vary across different times of the day, days of the week, and months of the year?
3. How do environmental factors like temperature, visibility, and weather conditions impact the frequency and severity of accidents?
4. Where are the geographic hotspots for accidents in the United States, and what patterns can be observed across regions, cities, and states?

Data Used: The 'US Accidents Dataset' contains accident records from February 2016 to March 2023 across almost all states.

Although the dataset is comprehensive, there were some challenges that quickly arose. Several key environmental fields, such as Wind_Chill, Precipitation, and Visibility, had significant missing data, which required careful handling during analysis. Additionally, the Weather_Condition variable contained over 140 unique weather descriptions, which required consolidation into broader categories (Rain, Fog, Snow, etc). Further, some rural areas and less populated regions had fewer accident records whereas urban and high-traffic areas showed much denser data. Lastly, the dataset's size (7.7 million records, 3.8GB) required efficient methods to work with.

Despite these challenges, the dataset provided a strong foundation for analyzing accident patterns, severity, and the relationships between geographic and environmental factors.

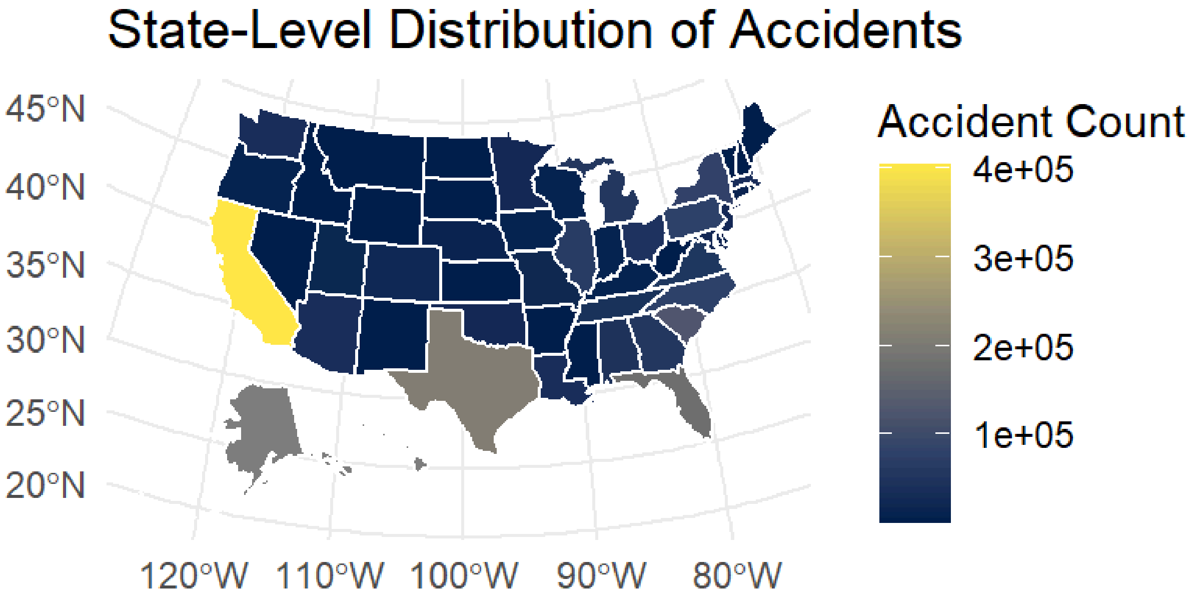
The dataset is publicly available on Kaggle and can be downloaded from US Accidents Dataset.

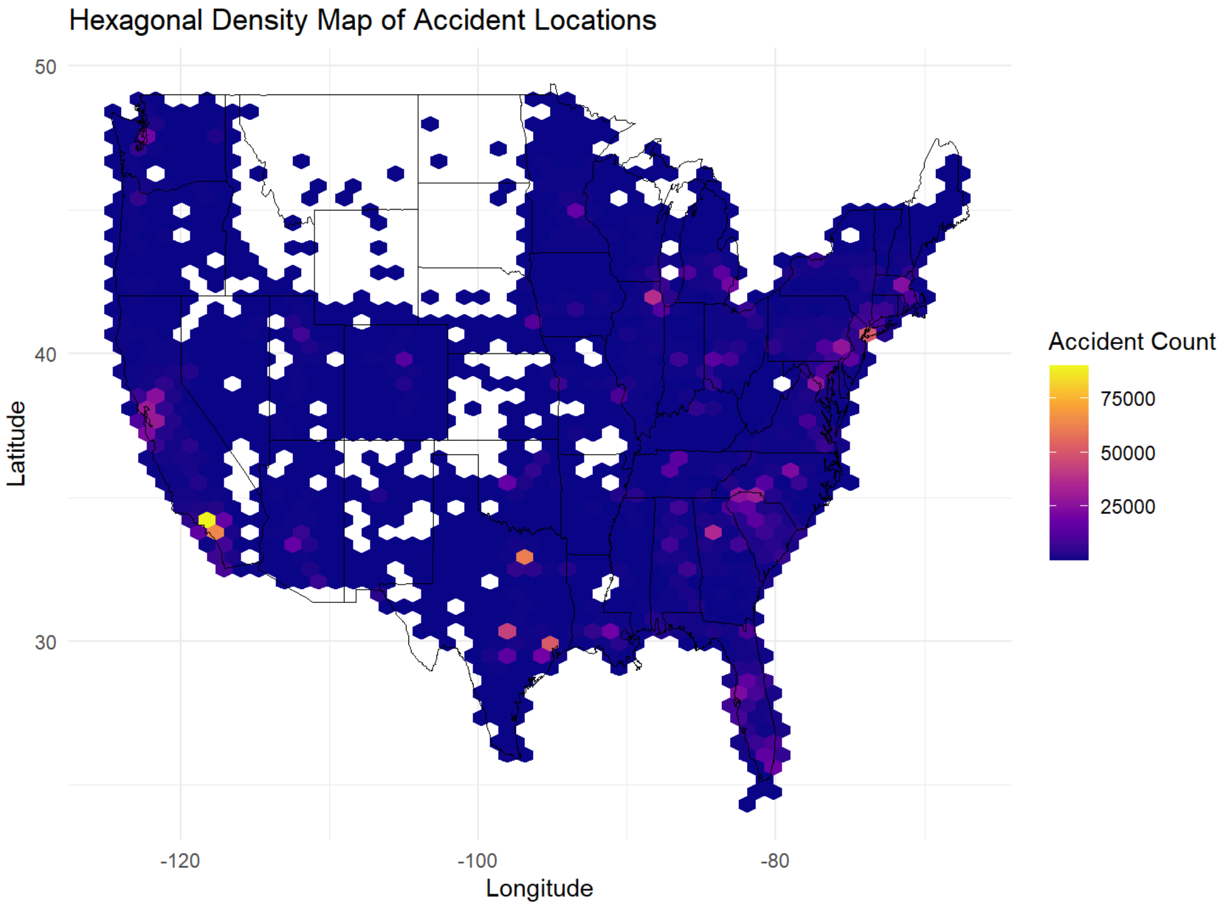
Data Properties: The dataset consists of 7,728,394 records and 46 variables. Data spans from February 2016 to September 2022, covering 49 states, and key fields include:

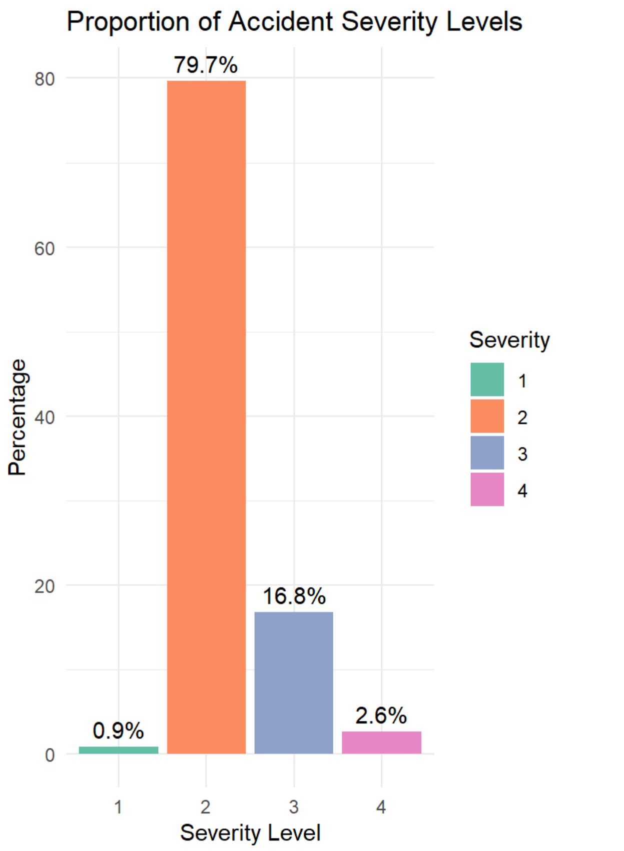
- Accident Details: ID, Severity, Start_Time, End_Time, Description
- Geographic Information: Start_Lat, Start_Lng, City, County, State, Zipcode
- Environmental Conditions: Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Precipitation(in), Weather_Condition
- Road Features: Indicators such as Amenity, Bump, Crossing, Junction, Traffic_Calming, and Traffic_Signal.

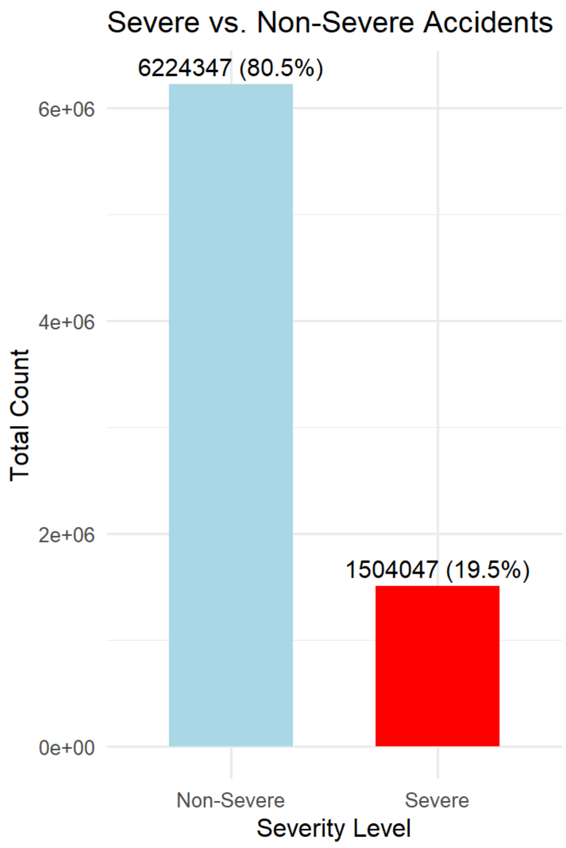
The dataset is mostly clean, but certain variables like Wind_Chill and Precipitation contain many missing values. Additionally, extreme values were observed in fields such as Temperature (ranging from -89°F to 203°F) and WindSpeed (up to 822.8 mph). These had to be carefully considered or ignored during analysis.

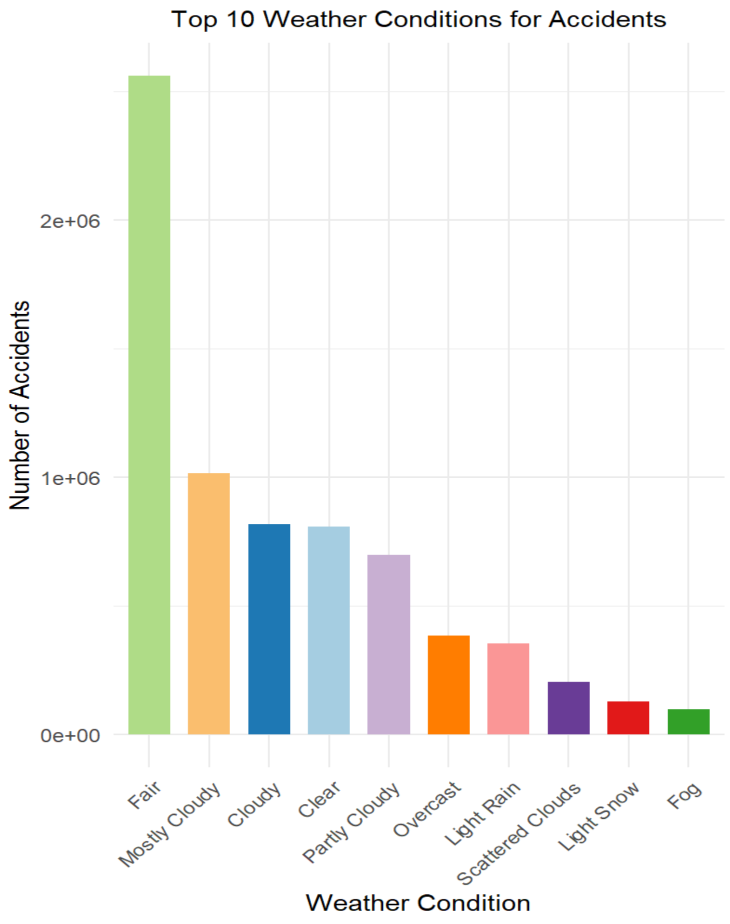
Plots:

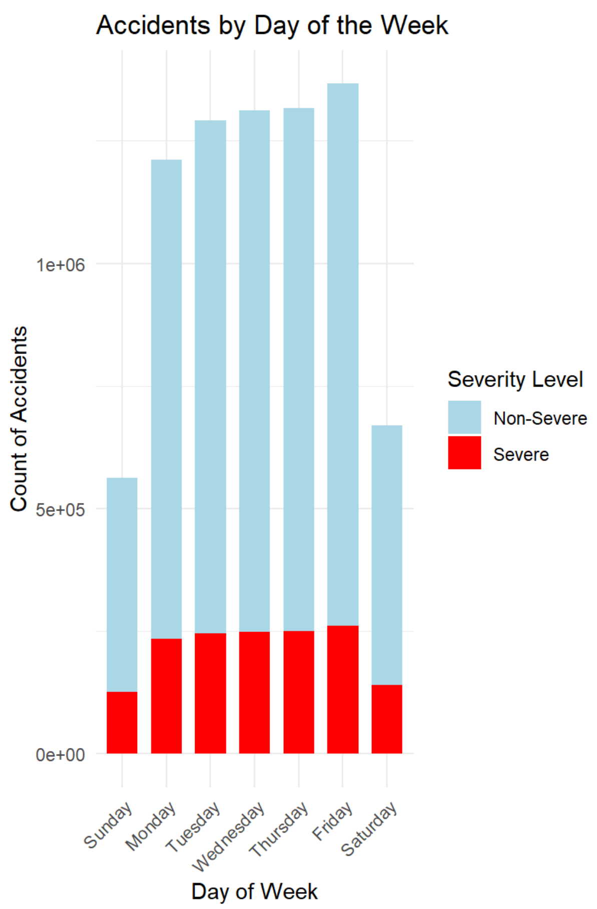

State-Level Distribution of Accidents
45°N
40°N
35°N
30°N
25°N
20°N
120°W
110°W
100°W
90°W
80°W
Accident Count
4e+05
3e+05
2e+05
1e+05


Hexagonal Density Map of Accident Locations
50
40
30
Latitude
-120
-100
-80
Longitude
Accident Count
75000
50000
25000

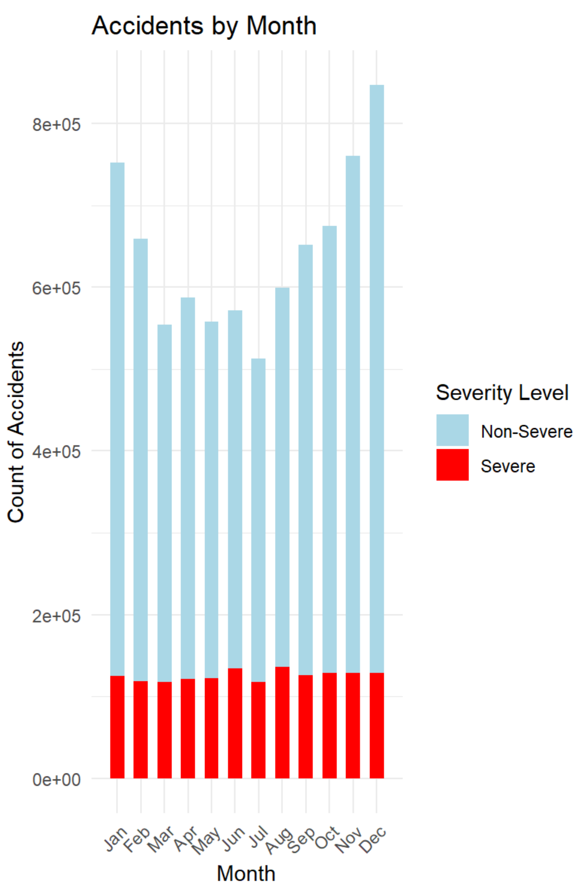

Proportion of Accident Severity Levels


Severe vs. Non-Severe Accidents

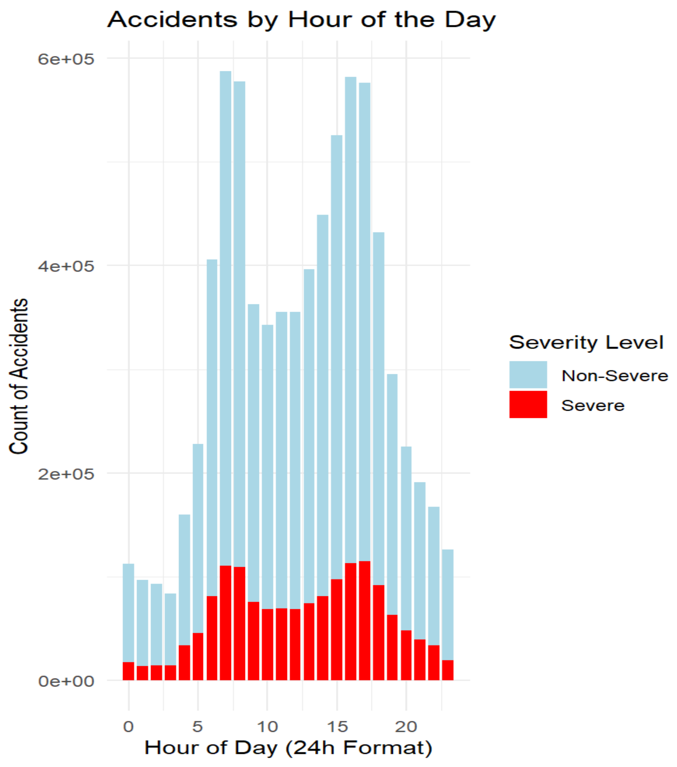

Top 10 Weather Conditions for Accidents



Accidents by Day of the Week



Accidents by Month



Accidents by Hour of the Day

R Code for Plots/Analysis:

```
library(tidyverse)
library(data.table)

data_path <- "C:/Users/aanan/Documents/final/US_Accidents_March23.csv"
data <- fread(data_path, stringsAsFactors = FALSE)

data <- data %>% select(-c(End_Lat, End_Lng))

num_cols <- c("Temperature(F)", "Humidity(%)", "Pressure(in)", "Visibility(mi)", "Wind_Speed(mph)")

for (col in num_cols) {
  data[is.na(get(col)), (col) := median(data[[col]], na.rm = TRUE)]
}

data$Wind_Chill_Status <- ifelse(is.na(data$Wind_Chill.F.), "Not Recorded", "Recorded")
data$Precipitation_Status <- ifelse(is.na(data$Precipitation.in.), "Not Recorded", "Recorded")
data$Weather_Timestamp <- ifelse(is.na(data$Weather_Timestamp), "Unknown", data$Weather_Timestamp)

write.csv(data, "Cleaned_US_Accidents.csv", row.names = FALSE)

library(ggplot2)

severity_counts <- data %>%
    count(Severity) %>%
    mutate(Proportion = n / sum(n) * 100)

ggplot(severity_counts, aes(x = Severity, y = Proportion, fill = Severity)) +
    geom_bar(stat = "identity") +
    scale_fill_brewer(palette = "Set2") +
    labs(title = "Proportion of Accident Severity Levels",
         x = "Severity Level",
         y = "Percentage") +
    geom_text(aes(label = paste0(round(Proportion, 1), "%")), vjust = -0.5) +
    theme_minimal()

data <- data %>%
    mutate(Severity_Level = ifelse(Severity %in% c("3", "4"), "Severe", "Non-Severe"))

severity_summary <- data %>%
    group_by(Severity_Level) %>%
    summarise(Count = n()) %>%
    mutate(Percentage = round((Count / sum(Count)) * 100, 1))

ggplot(severity_summary, aes(x = Severity_Level, y = Count, fill = Severity_Level)) +
    geom_bar(stat = "identity", width = 0.6) +
    geom_text(aes(label = paste0(Count, " (", Percentage, "%)")),
          vjust = -0.5, size = 4) +  # Reduced text size
    scale_fill_manual(values = c("Severe" = "red", "Non-Severe" = "lightblue")) +
    labs(title = "Severe vs. Non-Severe Accidents",
         x = "Severity Level",
         y = "Total Count") +
    theme_minimal(base_size = 12) +
    theme(legend.position = "none")

data$Hour <- lubridate::hour(data$Start_Time)

ggplot(data, aes(x = Hour, fill = Severity_Level)) +
    geom_bar(position = "stack", width = 0.8) +
    scale_fill_manual(values = c("Severe" = "red", "Non-Severe" = "lightblue")) +
    labs(title = "Accidents by Hour of the Day",
         x = "Hour of Day (24h Format)",
         y = "Count of Accidents",
         fill = "Severity Level") +
    theme_minimal()

ggplot(data, aes(x = Day_of_Week, fill = Severity_Level)) +
    geom_bar(position = "stack", width = 0.7) +
    scale_fill_manual(values = c("Severe" = "red", "Non-Severe" = "lightblue")) +
    labs(title = "Accidents by Day of the Week",
         x = "Day of Week",
         y = "Count of Accidents",
         fill = "Severity Level") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

ggplot(data, aes(x = Month, fill = Severity_Level)) +
    geom_bar(position = "stack", width = 0.6) +
    scale_fill_manual(values = c("Severe" = "red", "Non-Severe" = "lightblue")) +
    scale_x_discrete(expand = expansion(mult = c(0.1, 0.1))) +
    labs(title = "Accidents by Month",
         x = "Month",
         y = "Count of Accidents",
         fill = "Severity Level") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

top_weather <- data %>%
    count(Weather_Condition, sort = TRUE) %>%
    top_n(10, n)
```

```
ggplot(top_weather, aes(x = reorder(Weather_Condition, -n), y = n, fill = Weather_Condition)) +
  geom_bar(stat = "identity", width = 0.7) +
  scale_fill_brewer(palette = "Paired") +
  labs(title = "Top 10 Weather Conditions for Accidents",
     x = "Weather Condition",
     y = "Number of Accidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1),
      plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
      legend.position = "none")

library(usmap)
library(hexbin)
library(dplyr)
library(sf)

file_path <- "Cleaned_US_Accidents.csv"
data <- fread(file_path)

data_clean <- data[!is.na(Start_Lat) & !is.na(Start_Lng)]

ggplot(data_clean, aes(x = Start_Lng, y = Start_Lat)) +
  geom_hex(bins = 70) +  # Adjust 'bins' for granularity
  scale_fill_viridis_c(name = "Accident Count", option = "inferno") +
  labs(title = "Hexagonal Density Map of Accidents",
     x = "Longitude",
     y = "Latitude") +
  theme_minimal()

state_summary <- data_clean[, .N, by = State]
setnames(state_summary, "State", "state")

plot_usmap(data = state_summary, values = "N", color = "white") +
  scale_fill_viridis_c(name = "Accident Count", option = "cividis") +
  labs(title = "State-Level Distribution of Accidents") +
  theme_minimal()
```

Reasoning for Plot Selection: I chose a variety of visualizations to highlight trends and distributions across the dataset. The plots for accident severity ("Proportion of Accident Severity Levels" and "Severe vs Non-Severe Accidents") offer clear insights into the overall breakdown of accident severity. These help quantify the extent of severe accidents relative to non-severe incidents.

Next, time-based analysis was shown through the Accidents by Month, Day of the Week, and Hour of the Day plots. These identify patterns in accident frequency over time and address when accidents occur most frequently. This is important for understanding rush hour risks, seasonal variations, and weekday-weekend dynamics.

The geospatial visualizations, including the Hexagonal Density Map and State-Level Accident Distribution, reveal the geographical concentration of accidents. These maps provide an intuitive way to identify accident hotspots and could help with prioritization of road safety measures.

Finally, environmental factors were explored through the Top 10 Weather Conditions for Accidents plot. This highlights the most common weather scenarios under which accidents occur and helps to connect environmental conditions to accident rates.

What I Learned from Plots: The severity distribution plots revealed that the majority of accidents are non-severe, with severe accidents accounting for approximately 19.5% of the total. This highlights the relative rarity of high-severity accidents but also reveals their alarming absolute count. Analysis of time patterns demonstrated clear trends in accident frequency; accidents are most common during weekday rush hours with peaks around 7-8 AM and 4-6 PM.

The Accidents by Day of the Week graph showed elevated counts on weekdays, particularly Wednesday through Friday, which are likely due to higher commuting volumes. Monthly analysis revealed that December sees the highest accident counts, suggesting a potential link to winter weather conditions or increased holiday travel.

The geospatial analysis uncovered major accident hotspots, primarily concentrated in urbanized areas. This aligns with the expectation that higher population density and traffic volume lead to increased accident rates. The State-Level Distribution map further pinpointed states with disproportionately high accident counts, like California and Texas.

From the weather condition analysis, accidents predominantly occurred under “Fair,” “Mostly Cloudy,” and “Clear” conditions, suggesting that good weather does not necessarily equate to safer driving. However, conditions like “Light Rain” and “Fog” were also prominent, so reduced visibility and traction still likely contribute to accident frequency.

Answering Questions: The analysis and visualizations were largely successful in answering the posed questions. By examining temporal patterns, I identified clear trends in accident occurrence by hour, day, and month. This ultimately confirmed the influence of commuting and seasonal factors. The geospatial analysis effectively highlighted accident hotspots and validated my hypothesis that urban areas and certain states see disproportionately higher accident frequencies.

The environmental analysis provided a clearer understanding of how weather conditions impact accident rates. Severe weather events do contribute to accidents, but fair weather conditions still dominate, indicating that overall traffic and road design probably play significant roles in accident occurrence.

The severity distribution analysis addressed how common severe accidents are relative to non-severe ones. Most accidents are non-severe yet the absolute count of severe accidents is significant.

Conclusions: Even though the majority of accidents are non-severe, severe accidents account for a sizable portion, especially during peak traffic hours and in urbanized regions. Time trend analysis highlighted weekday rush hours and December as high-risk periods while geospatial maps identified major hotspots across densely populated states. Weather conditions, even fair or clear, play a critical role in accident frequency which suggests that road density and infrastructure remain significant factors. These findings highlight the need for targeted interventions, such as improved traffic management during peak hours and enhanced safety measures in high-risk regions, to mitigate accident occurrences and severity.

Recommendations: Based on my findings, I would recommend focusing on traffic safety interventions during peak traffic hours (morning and evening rush) and high-risk months, particularly in December. Urban areas and accident hotspots identified in the geospatial analysis should be prioritized for infrastructure improvements like better road design, enhanced traffic signals, and calming measures. Additionally, driver education programs emphasizing safe driving during all weather conditions, especially "fair" weather where complacency may contribute to accidents, could reduce accident rates.

Future Studies: Future research should focus on understanding how driver behavior like distraction or fatigue interacts with environmental and temporal factors to influence accident severity. Comparative studies between urban and rural areas would help identify location-specific risks and tailor intervention strategies. Additionally, analyzing the impact of extreme weather conditions and traffic policies on accident trends would provide actionable insights for safety improvements. Integrating real-time data sources, like live traffic, paves the way for predictive models and early warning systems to proactively reduce accidents.

References:

2024 SafeTREC Traffic Safety Facts: Emergency Medical Services | Safe Transportation Research and Education Center. (2024). Berkeley.edu. https://safetrec.berkeley.edu/2024-safetrec-traffic-safety-facts

Adeyemi, O. J., Arif, A. A., & Paul, R. (2021). Exploring the relationship of rush hour period and fatal and non-fatal crash injuries in the U.S.: A systematic review and meta-analysis. Accident; analysis and prevention, 163, 106462. https://doi.org/10.1016/j.aap.2021.106462

CDC. (2024, July 11). Risk and Protective Factors for Tribal Road Safety. Tribal Road Safety. https://www.cdc.gov/tribal-road-safety/risk-factors/index.html

Federal Highway Administration. (2016). How Do Weather Events Impact Roads? - FHWA Road Weather Management. Dot.gov. https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

Insurance Institute for Highway Safety. (2017). Fatality Facts 2017: Urban/rural comparison. IIHS-HLDI Crash Testing and Highway Safety. https://www.iihs.org/topics/fatality-statistics/detail/urban-rural-comparison

Moosavi, Sobhan. (2023). US Accidents (2016 - 2023) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/199387