1. INTRODUCTION

The dataset being utilized by us is titled "US Accidents (2016 - 2023)" and is accessible on Kaggle. Providing a detailed view of traffic accidents across the United States from 2016 to 2023, this dataset contains extensive information, including attributes such as location, weather conditions, and accident severity.

Research Question:

The research question we aim to address in this study is: "What are the key factors influencing the severity of traffic accidents in the United States, and how do these factors vary across different regions and time periods?"
Curiosity about the relevance of our research question led us to consider its importance. Our research has the potential to contribute to the development of more effective accident prevention strategies. By pinpointing the key factors contributing to severe accidents, authorities can implement targeted measures to reduce their occurrence. Gaining insights into the relationship between various factors and accident severity is a step towards creating safer road environments for all.

Motivation:

This research question is crucial because traffic accidents significantly impact public safety, healthcare, and the economy. In 2020, nearly 39,000 lives were lost in the U.S. due to traffic crashes. Understanding what makes accidents more severe is vital for reducing this toll. Analyzing the "US Accidents (2016 - 2023)" dataset, we aimed to uncover patterns and factors influencing accident severity. Our goal is to contribute to making roads safer and lessening the devastating impact of traffic accidents on people's lives.

2. DATASET DESCRIPTION

The "US Accidents (2016 - 2023)" dataset is a comprehensive collection of information on traffic accidents that occurred in the United States over the period from 2016 to 2023. Each unit in this dataset corresponds to a single traffic accident, and the dataset contains various attributes describing different aspects of each incident.

Key Variables:

ID: A unique identifier for each accident record.
Country: Indicates the country where the city, where the accident took place, is located.
Source: The source of the accident report (e.g., Bing, MapQuest).
Severity: A categorical variable indicating the severity of the accident (ranging from 1 for low severity to 4 for high severity).
Start and End Time: Timestamps indicating when the accident occurred and when it was cleared.
Distance: The distance covered by the accident.
Description: A narrative description of the accident.
Weather Conditions: Information about weather conditions at the time of the accident.
Road Conditions: Details about the state of the road surface.
Visibility: Visibility conditions during the accident.

The response variable in the context of our dataset would be the severity of the accidents. In statistical terms, the **severity of the accident** would be the response variable. It is the variable that is expected to change based on the influence of other variables, such as weather conditions, road type, time of day, and other factors.
The key explanatory variables in the context of our dataset would include various factors that are expected to influence the severity of traffic accidents. These variables can encompass a range of variables such as **weather conditions**, **road features**, t**ime-related factors**, and other relevant attributes.
Key insight: We printed a table of the percentage of accidents in the top 5 weather conditions. These results indicate a possible relationship between weather variables and accidents. This motivated us to explore this relationship and the possible relationship of accidents with the road conditions as well.

weather_condition	percentage
<chr>	<dbl>
Fair	38.912510
Mostly Cloudy	12.791397
Cloudy	12.257472
Partly Cloudy	8.784146
Clear	6.109037

3. ANALYSIS

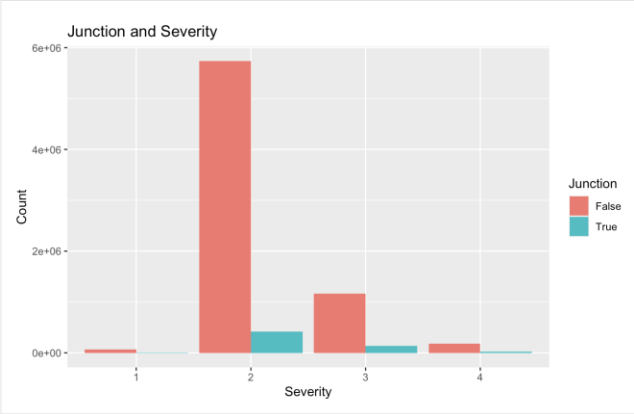
Analysis 1: Chi Squared test

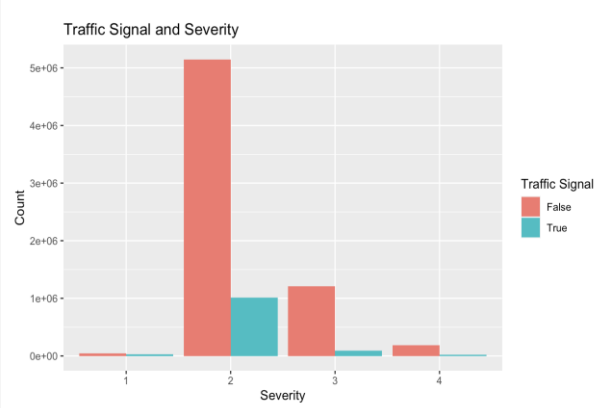
Methodology:

We employed the Chi-square test to explore the relationship between road features and accident severity, dividing the analysis into two parts: Part A for the entire United States and Part B specifically for California. California was chosen due to its lower data imbalance and fewer null values. Three variables were selected for each part for Part A: Junction & Bump and for Part B: Traffic Signal and No exit, showcasing a diverse range of statistical analyses. After dropping rows with null values, visualizations were created for each variable, and Chi-square tests were conducted on subsets focusing on accident severities 3 and 4, as well as 1 and 4. This approach allowed us to uncover potential associations between road features and accident severity in a comprehensive manner.

Conclusion:

In analyzing the relationship between severity and road features, visualizations alone were insufficient, prompting the use of the Chi-squared test. For the entire United States dataset, the analysis revealed a significant relationship between the presence of a junction and accident severity ($p < 0.05$ rejecting H_0), as well as a similar relationship with the presence of a bump ($p < 0.05$). In the specific case of California, there was no discernible relationship between the absence of an exit and severity ($p > 0.05$) Thus lack of an exit has no effect on severity of an accident. However, a significant relationship was observed between the presence of a traffic signal and accident severity ($p < 0.05$).





Analysis 2: Analysis of Accident Severity Across States Using Road Extent Affected by Accidents across Different States:

Assumptions check: The data are created randomly and are independent. Since the sample size is Large according to CLT distribution is normal.

In our analysis to assess the severity of road accidents across different states, we focused on the variable Distance.mi., representing the length of road extent affected by each accident in miles. This continuous variable offers a more nuanced understanding of accident severity compared to categorical variables like 'Severity'.

Methodology:

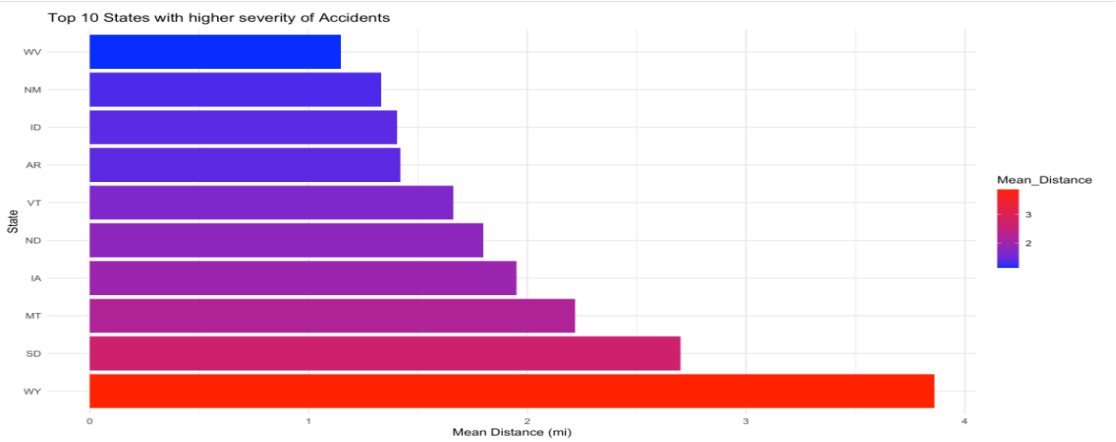
We calculated the 95% confidence intervals for the mean distance affected by accidents for each state. This approach allows us to understand the range within which the true mean distance is likely to fall, with a 95% level of confidence. To incorporate the varying number of accidents across states, we calculated a weighted mean of all upper confidence intervals. The weights were derived from the proportion of total accidents occurring in each state, thereby giving more representation to states with a higher frequency of accidents. We then compared each state's upper confidence interval against this overall weighted mean. States with an upper confidence interval exceeding the overall weighted mean were identified as having significantly higher accident severity.

Conclusion:

The analysis identified the top 10 states where accidents are, on average, more severe in terms of the affected road extent. These states have upper confidence intervals for the mean accident distance that are higher than the weighted average across all states. This result suggests that in these states, accidents tend to affect a longer stretch of road, potentially

indicating higher severity or impact. The findings provide valuable insights for policymakers and transportation authorities to prioritize safety measures and resources in states with higher severity accidents.

Graph:



<u>Analysis 3:</u> Analysis of differences in the proportion of severity level 2 in states with the national proportion

The analysis focused on severity proportions at level 2 in different states and compared them with the national severity proportion. We specifically examined accidents falling within the severity scale of 2, considering them to be the most common occurrences in terms of severity. This approach aimed to better understand whether there is any geographical impact or infrastructural lag related to the prevalence of these common accidents.

Methodology:

Assumptions Check: The samples were randomly created and are independent. The sample size is sufficiently large (np>5, nq>5). In this study, we extracted traffic accident data from the "US_Accidents_March23.csv" file and conducted an initial exploration. Missing data in the 'State' and 'Severity' columns were addressed. Descriptive statistics were then calculated, summarizing severity levels across states. The proportion of severity level 2 As this is the most common severity in all the states) incidents was computed for each state, along with the overall population proportion. Z-scores were calculated to assess the statistical significance of state-level variations. Results were visualized through a histogram, and hypothesis testing was performed to determine if severity proportions significantly differed from the national average.

Conclusion:

The analysis of severity proportions in different states reveals distinct patterns in comparison to the national average. The top 5 states with severity level 2 incidents significantly lower than the national proportion include Georgia (GA), Illinois (IL), Colorado (CO), Missouri (MO), and Ohio (OH). These states exhibit z-scores with large negative values, indicating a substantial deviation from the national norm. Conversely, the top 5 states with severity level 2 incidents significantly higher than the national proportion consists of North Dakota (ND), South Carolina (SC), Oregon (OR), Florida (FL), and Montana (MT). These states demonstrate z-scores with large positive values, signifying a substantial increase in severity proportions compared to the national average. These findings warrant further investigation into the contributing factors. Public health interventions and safety measures may need to be tailored to address the specific challenges faced by states with proportions deviating significantly from the national average.

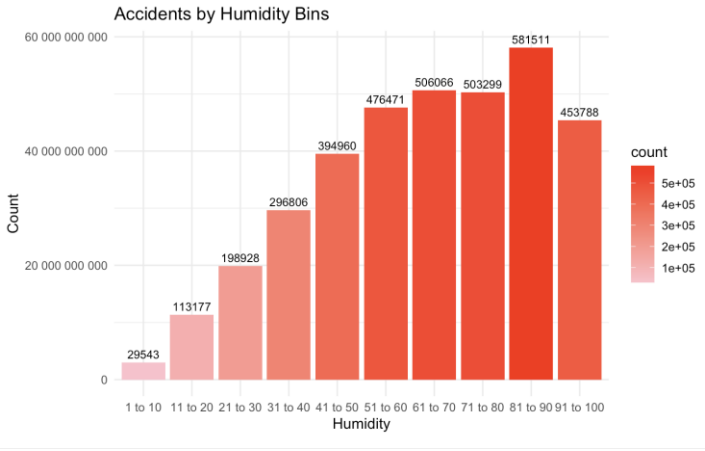
The top 5 states have a proportion of severity 2 less than the national proportion.

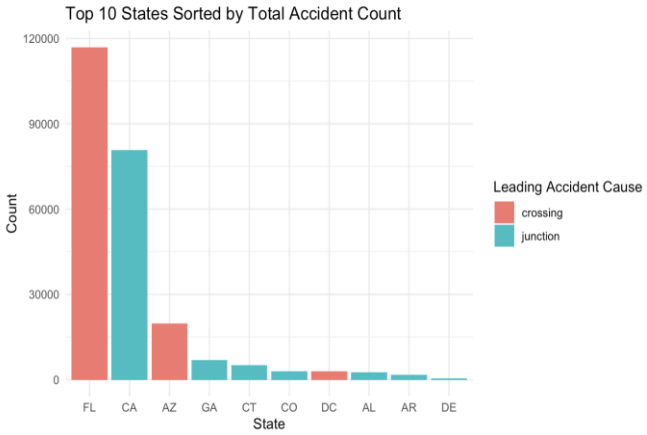
State	proportion_of _severity	n	Standard Error	z_score	p_value
GA	0.5566612	169234	0.0012075901	-198.750377	0.00E+00
IL	0.623072	168958	0.0011789868	-147.243524	0.00E+00
CO	0.6173186	90885	0.001612231	-111.244351	0.00E+00
MO	0.6306791	77323	0.0017356087	-95.638537	0.00E+00
OH	0.6719976	118115	0.0013660602	-91.264314	0.00E+00

The Top 5 states have proportion of severity 2 more than the national proportion.

State	proportion_of _severity	n	Standard Error	z_score	p_value
ND	0.9908231	3487	0.0016148087	120.232741	0.00E+00
SC	0.8647522	382557	0.0005529208	123.131579	0.00E+00
OR	0.9056718	179660	0.0006895733	158.071154	0.00E+00
FL	0.8587842	880192	0.0003711889	167.338117	0.00E+00
MT	0.9788742	28496	0.0008518788	213.884958	0.00E+00

<u>Analysis 4:</u> Leading cause of accidents in first 10 states sorted by accident count and analysis of road accidents based on weather conditions:

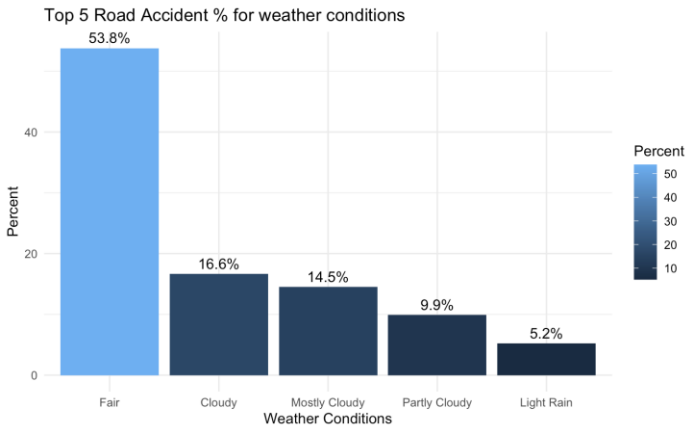


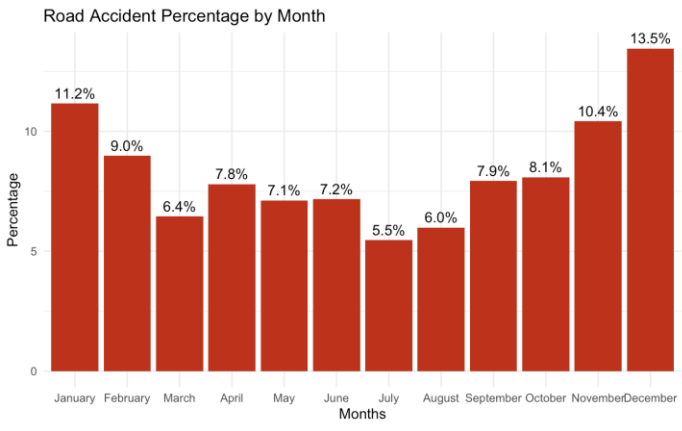


Methodology:

Analyzing the impact of various independent variables on accident rates in each state, we investigated correlations with factors such as distance (indicating accident severity), weather conditions (e.g., wind speed, visibility), and road conditions (e.g., bumps, junctions). Handling missing values by omitting corresponding rows, we identified leading causes by grouping the dataset by state and selecting the cause with the highest count, considering columns such as "bump," "crossing," "give_way," "junction," and "no_exit." Among the first 10 states, "crossing" emerged as the leading cause in 3 states, while "junction" accounted for 7 states. Notably, Florida experienced the highest accident count, followed by California and Arizona.

Further exploration involved examining accident counts across humidity bins, grouping temperature (F) into 10-degree intervals. The analysis revealed that higher temperatures were associated with an increased number of accidents, with the peak occurring in the temperature range of 81F to 90F.



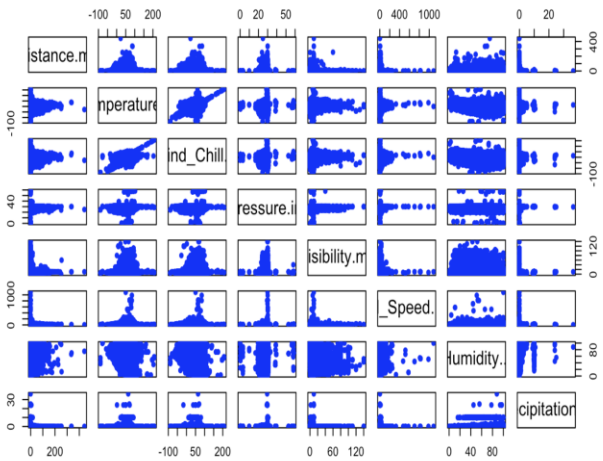


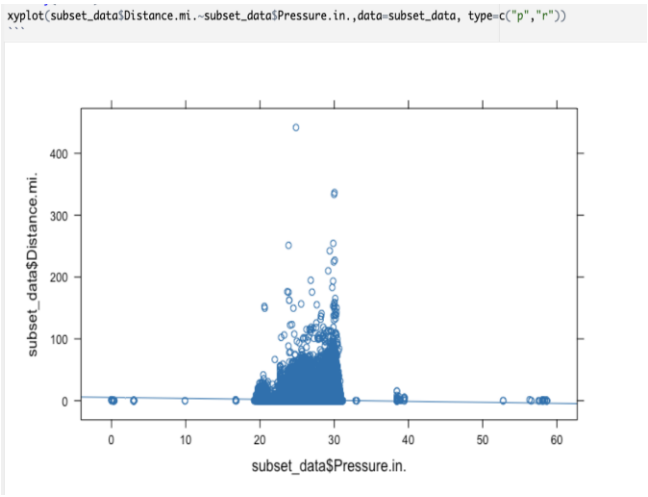
Additionally, we found that most accidents (53.8%) happened under fair weather conditions, followed by cloudy (16.6%) and mostly cloudy (14.5%). Furthermore, December recorded the highest number of accidents, with January and November following suit.

Conclusion:

There is a positive correlation between temperature and accidents, with the peak occurring in the temperature range of 81F to 90F. This suggests that drivers may be more likely to make mistakes or take risks in hot weather.
The months with the highest number of accidents are December, January, and November. This could be due to a number of factors, such as increased holiday traffic, harsher winter weather conditions, or shorter days.
The percentage of accidents that occur under fair weather conditions is 53.8%, while the remaining 46.2% of accidents occur under other conditions. This suggests that weather may play a role in accident rates, but it is not the only factor.

Analysis 5: Impact of weather conditions on the severity of the accident.
Methodology:
We have continuous weather variables: Temperature, Wind_Chill, Pressure, Visibility, Wind_speed, Humidity and Precipitation. We explored the relationship between distance impacted by the accident in miles (continuous variable) and the weather variables. I used the pairs method to see if distance showed any relationship between the weather variables.
As it is visible in the screenshot, there is no visible relationship. To investigate further, we have tried to fit the SLR model. We first preprocessed the data by filling the NaN values in categorical variables by their mode and filled the NaN values in the weather variables, by their median grouped by Astronomical_Twilight, City, Junction, Crossing. We then checked the correlations of distance with each of the weather variables, the correlation levels were very low. We then selected the variable with the highest absolute correlation coefficient(pressure) and fit SLR. The multiple R square value is very low, which is in line with the unchecked assumptions for SLR:
1)linearity- this condition is not satisfied as there does not seem to be any linear relation of distance with pressure
2)constant variance and Independence: There is a pattern in the scatterplot of residuals, the spread is not random. There seem to be more residuals on the positive side of x=0, showing that the expected value of residuals is not zero.
3)Normality:
The qq plot is far from linear. Therefore, there is no normality in the spread of residuals. This condition is not satisfied
Conclusion: all four assumptions are not satisfied, which in line with the low multiple R square value of the fitted model, hence we cannot use SLR.





The other approaches adopted:

- Fitting MLR model: very low R^2 value- not a good fit
- scale the values with log scaling and standard normal and check the values of correlation of weather variables with distance variable (tried both one by one): still very low correlation coefficients.
- Fitting Polynomial model: very low R^2 value- not a good fit

Conclusion:
We can conclude that the continuous weather variables are not sufficient on their own to predict the distance impacted by the accident.
Choice of model: "Severity" response variable is ordered and has values "1","2","3","4" based on how severe the accident was. We want to analyze the relationship between severity of the accident and our continuous weather variables. Therefore, we will have to use an ordinal regression model.
Cumulative link model (CLM) is a powerful model for such data since observations are treated correctly as categorical, the ordered nature is exploited, and the flexible regression framework allows for in-depth analyses.

Assumptions:

1)Independence of Observations:

Observations in your dataset should be independent of each other. This means that the occurrence or value of the response variable for one observation should not be influenced by or correlated with the occurrence or value of the response variable for any other observation.

2)Proportional Odds Assumption:

The proportional odds assumption, also known as the parallel regression assumption, states that the effect of a predictor variable on the odds of being in a higher response category is constant across all levels of the response variable.

Since we have a large dataset with observations that are collected randomly in different areas and includes a vast number of accidents with different levels of severity, we assume these conditions are satisfied.

Results:

```
subset_data$Severity ~ subset_data$Temperature.F. + subset_data$Wind_Chill.F. +
subset_data$Pressure.in. + subset_data$Visibility.mi. + subset_data$Wind_Speed.mph. +
subset_data$Humidity... + subset_data$Precipitation.in.
data:     subset_data


Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
subset_data$Temperature.F.     1.528e-02  1.175e-04  129.99   <2e-16 ***
subset_data$Wind_Chill.F.     -1.818e-02  1.123e-04 -161.87   <2e-16 ***
subset_data$Pressure.in.       1.613e-01  1.077e-03  149.87   <2e-16 ***
subset_data$Visibility.mi.     7.369e-03  3.619e-04   20.36   <2e-16 ***
subset_data$Wind_Speed.mph.    1.812e-02  1.779e-04  101.85   <2e-16 ***
subset_data$Humidity...        2.674e-03  4.663e-05   57.34   <2e-16 ***
subset_data$Precipitation.in.  1.780e-01  7.775e-03   22.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  0.24919    0.03150   7.911
2|3  6.45398    0.03156 204.467
3|4  8.64781    0.03166 273.175
```

Interpretation of results:

```
Interpretation of results:
subset_data$Temperature.F.:

Estimate: 0.01528
Interpretation: For a one-unit increase in temperature, the log-odds of moving to a higher severity category increase by approximately 0.01528.
Z value: 129.99 (large and positive)
Significance: Highly significant (p < 0.001)
Interpretation: Temperature is positively associated with higher severity.
subset_data$Wind_Chill.F.:

Estimate: -0.01818
Interpretation: For a one-unit increase in wind chill, the log-odds of moving to a higher severity category decrease by approximately 0.01818.
Z value: -161.87 (extremely large and negative)
Significance: Highly significant (p < 0.001)
Interpretation: Wind chill is negatively associated with higher severity.
subset_data$Pressure.in.:

Estimate: 0.1613
Interpretation: For a one-unit increase in pressure, the log-odds of moving to a higher severity category increase by approximately 0.1613.
Z value: 149.87 (large and positive)
Significance: Highly significant (p < 0.001)
Interpretation: Pressure is positively associated with higher severity.
subset_data$Visibility.mi.:

Estimate: 0.007369
Interpretation: For a one-unit increase in visibility, the log-odds of moving to a higher severity category increase by approximately 0.007369.
Z value: 20.36 (large and positive)
Significance: Highly significant (p < 0.001)
Interpretation: Visibility is positively associated with higher severity.
subset_data$Wind_Speed.mph.:

Estimate: 0.01812
Interpretation: For a one-unit increase in wind speed, the log-odds of moving to a higher severity category increase by approximately 0.01812.
Z value: 101.85 (large and positive)
Significance: Highly significant (p < 0.001)
Interpretation: Wind speed is positively associated with higher severity.
subset_data$Humidity...:

Estimate: 0.002674
Interpretation: For a one-unit increase in humidity, the log-odds of moving to a higher severity category increase by approximately 0.002674.
Z value: 57.34 (large and positive)
Significance: Highly significant (p < 0.001)
Interpretation: Humidity is positively associated with higher severity.
Precipitation.in.:

Estimate: 0.178
Interpretation: For a one-unit increase in precipitation, the log-odds of moving to a higher severity category increase by approximately 0.178.
Z value: 22.89 (large and positive)
Significance: Highly significant (p < 0.001)
Interpretation: Precipitation is positively associated with higher severity.
```

Conclusion:

The continuous weather variables have statistically significant relationships with Severity response variable.