STAT 632: Regression Analysis

Predicting Road Accident Severity in the US (2016-2023) : A Multiple Linear Regression and Random Forest Analysis

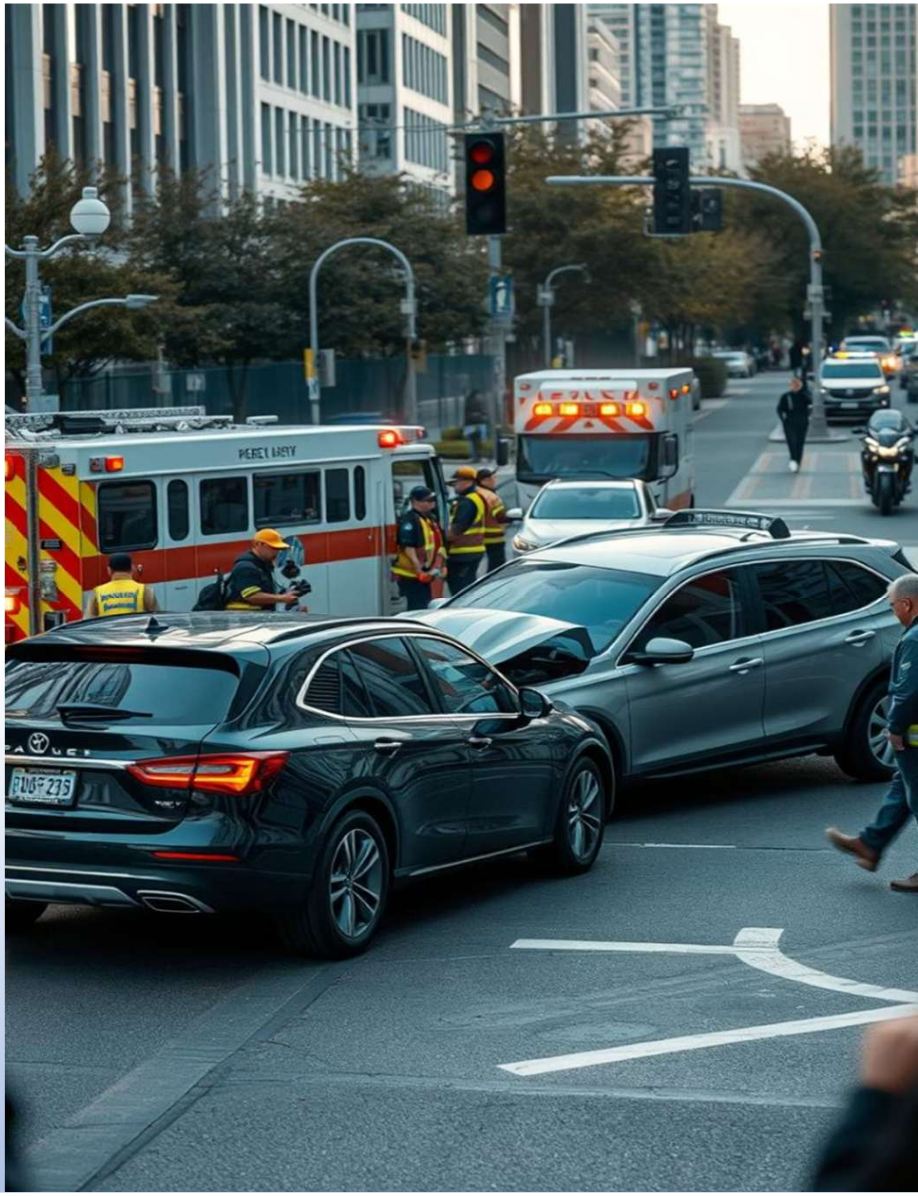
Presented by:
Shrusthi H C
Shruti Gajre
Namitha Kuthani

Instructor :
Dr. Joshua Kerr

Last updated: May 6th, 2025

Agenda

1. Introduction & Motivation
2. Project Goal & Research Questions
3. Dataset Overview
4. Data Preprocessing & EDA
5. Modeling Approaches
    - Multiple Linear Regression (Full vs. Reduced Model)
    - Logistic Regression
    - Random Forest Classification
6. Model Evaluation
7. Key Results & Interpretation
8. Conclusion & Future Work



Project Goal & Research Question:

Project Goal

To analyze patterns in road accidents across the United States from 2016 to 2023 using regression models. The aim is to **predict accident severity** and identify the most influential factors contributing to severe outcomes.

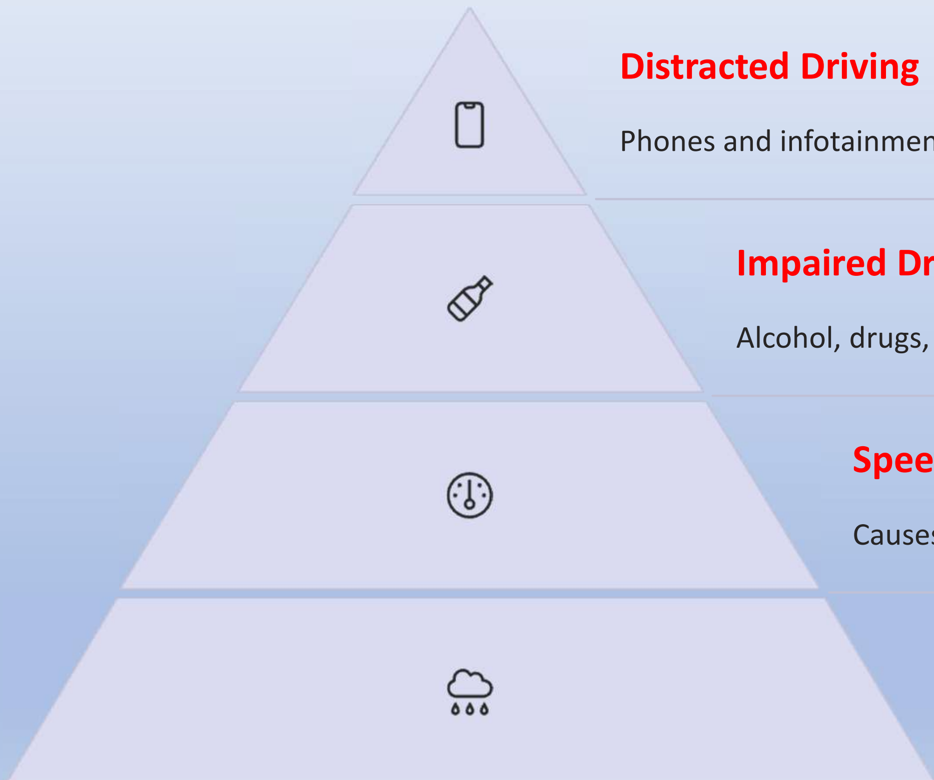
Research Questions

- What are the key predictors of road accident severity in the US?
- How do environmental and temporal factors (e.g., weather, visibility, time of day) influence the severity of road accidents?
- Can Multi-linear regression and Random forest models accurately classify and predict accident severity?

Dataset Overview

- Source: Kaggle – U.S. Accidents (2016-2023)
- Data size: 7.7 Million entries, 46 total columns
- Sample Size: 50,000 rows
- Target Variable: Severity (1 to 4)
- Key Predictors: Weather: temperature, rain, fog, State etc.
- Road Features: Bump, junction, traffic signal etc..
- Time: Peak hour, weekend

Common Causes of Car Accidents



Distracted Driving

Phones and infotainment systems divert attention

Impaired Driving

Alcohol, drugs, and fatigue impair judgment

Speeding

Causes 29% of US traffic deaths

Weather Conditions

Rain, snow, and fog reduce visibility and control

Multiple Linear Regression

MLR Full Model Results:

```
Call:
lm(formula = Severity ~ Distance + Temperature + Humidity + Visibility +
    Wind_Speed + Pressure + Precipitation + Weather_Simple +
    Rush_Hour + Weekend + Is_Daytime + Traffic_Signal_Flag +
    Road_Features + State, data = accident_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4874 -0.1975 -0.1052 -0.0492  2.1603

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.8178716  0.0828328  21.946  < 2e-16 ***
Distance                0.0021842  0.0006904   3.164 0.001559 **
Temperature             0.0009606  0.0001185   8.109 5.18e-16 ***
Humidity                0.0002248  0.0001035   2.172 0.029839 *
Visibility             -0.0005650  0.0009724  -0.581 0.561267
Wind_Speed              0.0006393  0.0003521   1.816 0.069426 .
Pressure                0.0108640  0.0027150   4.001 6.30e-05 ***
Precipitation           0.0399863  0.0161080   2.482 0.013053 *
Weather_SimpleCloudy    0.0183983  0.0039791   4.624 3.77e-06 ***
Weather_SimpleFreezing  0.2788720  0.0598741   4.658 3.20e-06 ***
Weather_SimpleRainy     0.0761174  0.0077852   9.777  < 2e-16 ***
Weather_SimpleSnowy     0.0389131  0.0130372   2.985 0.002839 **
Rush_Hour              -0.0064177  0.0037293  -1.721 0.085277 .
Weekend                 0.0393409  0.0045397   8.666  < 2e-16 ***
Is_Daytime             -0.0085665  0.0042549  -2.013 0.044085 *
Traffic_Signal_Flag    -0.1048751  0.0053955 -19.438  < 2e-16 ***
Road_Features          -0.0417544  0.0044502  -9.383  < 2e-16 ***
StateAR                 0.0731888  0.0315940   2.317 0.020532 *
StateAZ                -0.1388048  0.0191841  -7.235 4.69e-13 ***
StateCA                -0.1242984  0.0146645  -8.476  < 2e-16 ***
```

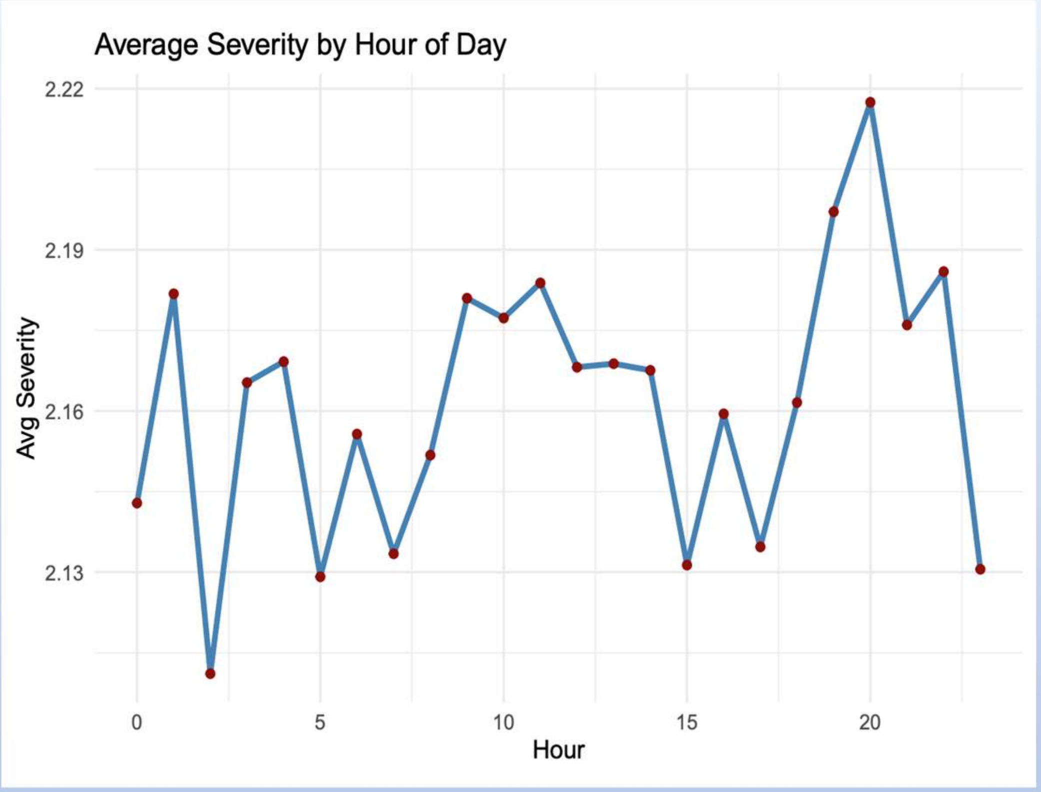
- *Adjusted R^2 = **0.060** → Model explains ~6% of variance in severity*
- *F-statistic = **21.08**, $p < 0.05$ → Model is statistically significant*
- *Residual standard error = 0.4379 (indicates moderate prediction error)*

Reduced model Results

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.9660903  0.0297308  66.130  < 2e-16 ***
## Distance               -0.0031206  0.0026210  -1.191 0.233822
## Temperature             0.0011373  0.0002505   4.540 5.68e-06 ***
## Humidity                0.0004753  0.0002080   2.286 0.022294 *
## Wind_Speed              0.0001377  0.0007332   0.188 0.851003
## Weather_SimpleCloudy    0.0090643  0.0081249   1.116 0.264603
## Weather_SimpleFreezing  0.1208959  0.1103757   1.095 0.273399
## Weather_SimpleRainy     0.0786488  0.0147898   5.318 1.07e-07 ***
## Weekend                 0.0376573  0.0092539   4.069 4.74e-05 ***
## Traffic_Signal_Flag    -0.1089508  0.0113773  -9.576  < 2e-16 ***
## Road_Features          -0.0393543  0.0090081  -4.369 1.26e-05 ***
## StateCA                 0.0279006  0.0211715   1.318 0.187580
## StateCO                 0.3702451  0.0341941  10.828  < 2e-16 ***
## StateFL                 0.0146901  0.0227501   0.646 0.518474
## StateGA                 0.3315951  0.0301842  10.986  < 2e-16 ***
## StateIA                 0.3494500  0.0569610   6.135 8.76e-10 ***
## StateIL                 0.3486123  0.0299254  11.649  < 2e-16 ***
## StateIN                 0.3549783  0.0405513   8.754  < 2e-16 ***
## StateKY                 0.2894248  0.0494526   5.853 4.95e-09 ***
## StateLA                 0.0257535  0.0287378   0.896 0.370188
## StateMA                 0.2094755  0.0439204   4.769 1.87e-06 ***
## StateME                 0.0631046  0.1460898   0.432 0.665779
## StateMI                 0.1395208  0.0311093   4.485 7.36e-06 ***
## StateMN                 0.0424529  0.0289212   1.468 0.142160
## StateMO                 0.3225876  0.0366880   8.793  < 2e-16 ***
## StateMT                 0.0164094  0.0543041   0.302 0.762522
## StateNC                 0.0953699  0.0247546   3.853 0.000117 ***
## StateNH                 0.0964838  0.1154056   0.836 0.403147
## StateNM                 0.3646688  0.0749817   4.863 1.17e-06 ***
## StateOH                 0.1875112  0.0328704   5.705 1.19e-08 ***
## StateOK                 0.0117528  0.0367951   0.319 0.749419
## StateOR                 0.0354479  0.0281422   1.260 0.207835
## StateRI                 0.3924769  0.0658177   5.963 2.54e-09 ***
## StateSC                 0.0036847  0.0242074   0.152 0.879021
## StateWA                 0.2502880  0.0333549   7.504 6.59e-14 ***
## StateWI                 0.3794455  0.0471177   8.053 8.75e-16 ***
## StateWV                 0.0079154  0.0687329   0.115 0.908318
## StateWY                 0.0705615  0.1461711   0.483 0.629294
```

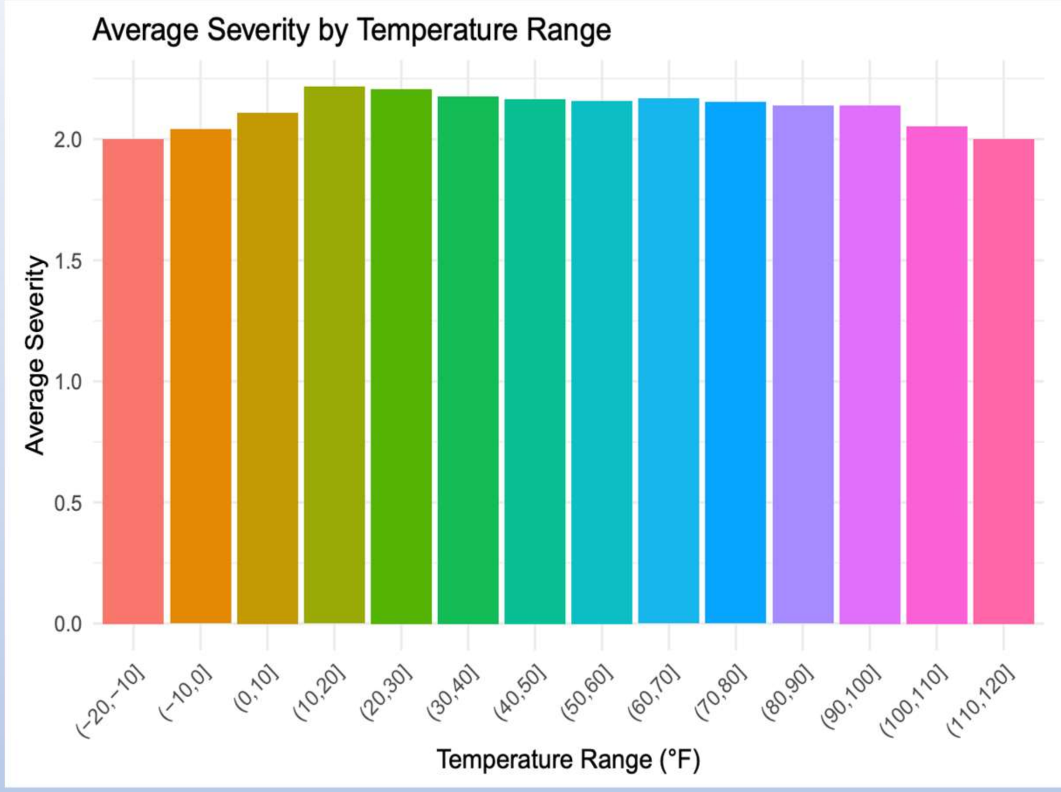
- *The **reduced model has better explanatory power** (Adjusted R^2 = 7.8% vs. 6.0%)*
- *It also has **lower residual error**, indicating tighter model fit*
- *Unnecessary or insignificant predictors were removed — improving **model simplicity and interpretability***
- *Focuses only on **important weather types and states***
- ***Avoids multicollinearity**, as shown by improved VIF scores*

Trend of average severity by hour:



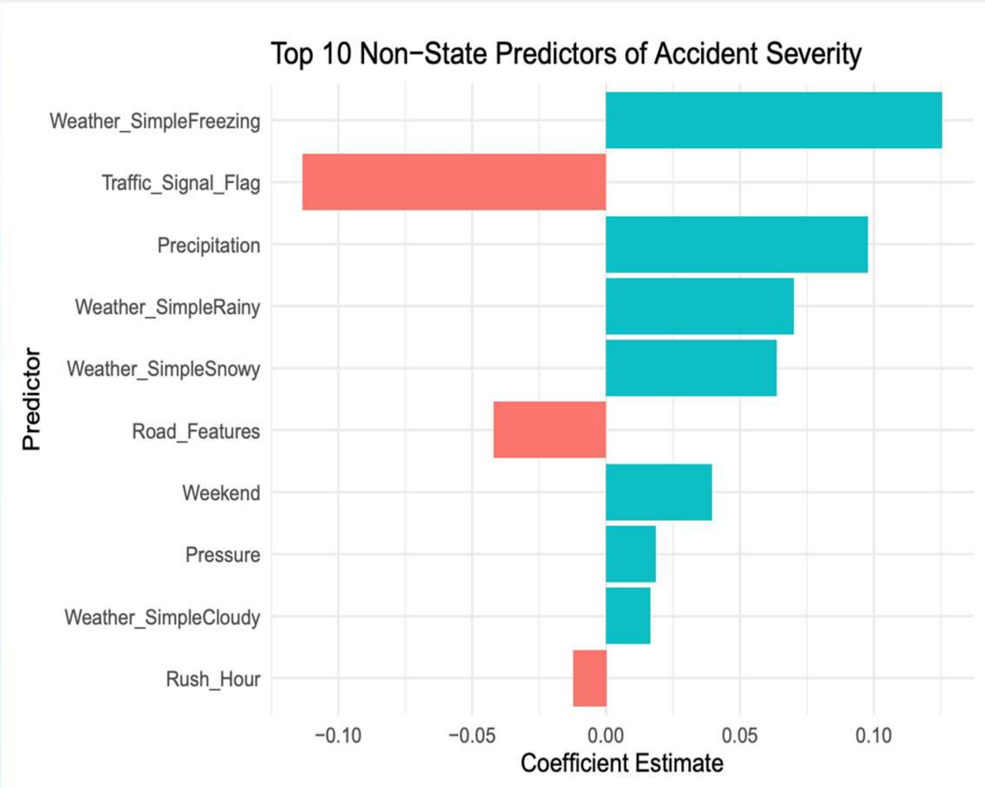
- Shows **average accident severity** for each hour (0 to 23)
- Severity fluctuates throughout the day, mostly ranging between **2.13 and 2.22**
- **Highest average severity occurs between 7 PM and 9 PM**
- **Lowest severity appears around early morning (2–4 AM) and mid-afternoon (3–5 PM)**
- No clear peak during traditional **rush hours** (7–9 AM and 4–6 PM)

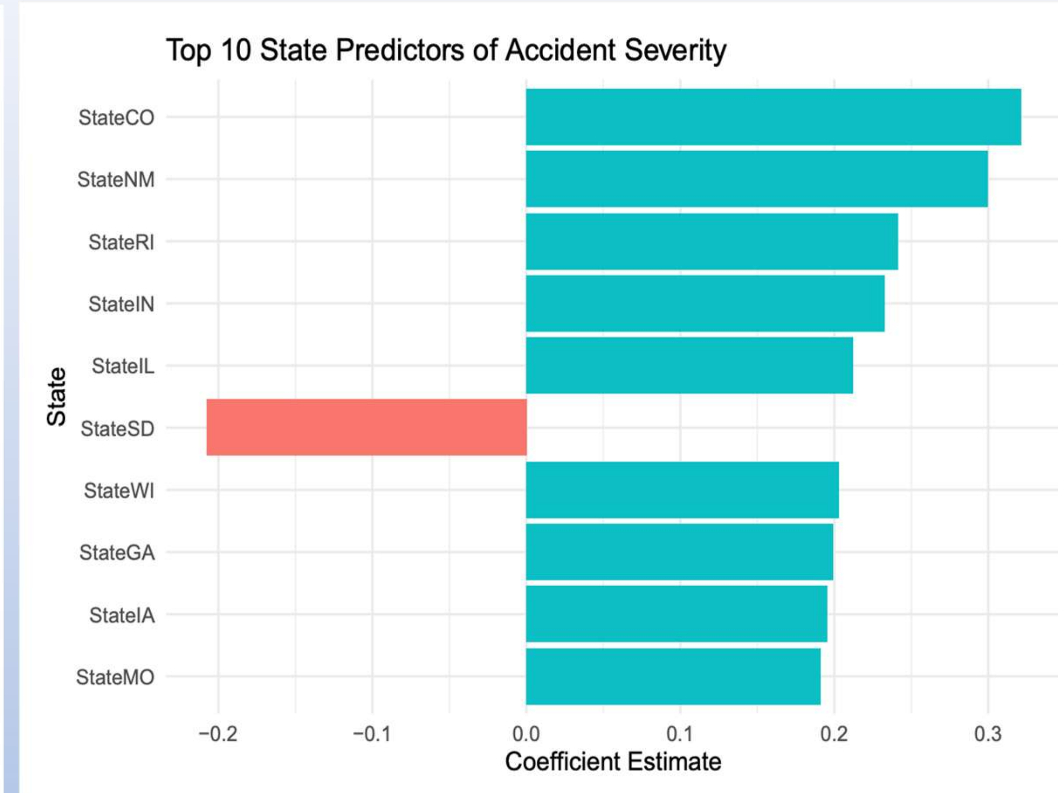
Average Severity by Temperature Range:



- Severity peaks in the **20°F to 30°F** range
- Temperatures **below freezing and above 100°F** are associated with **lower average severity**
- Overall, **mid-range temperatures (20–60°F)** show the highest average severity
- Severity gradually declines after ~60°F

Bar plot of model coefficients:

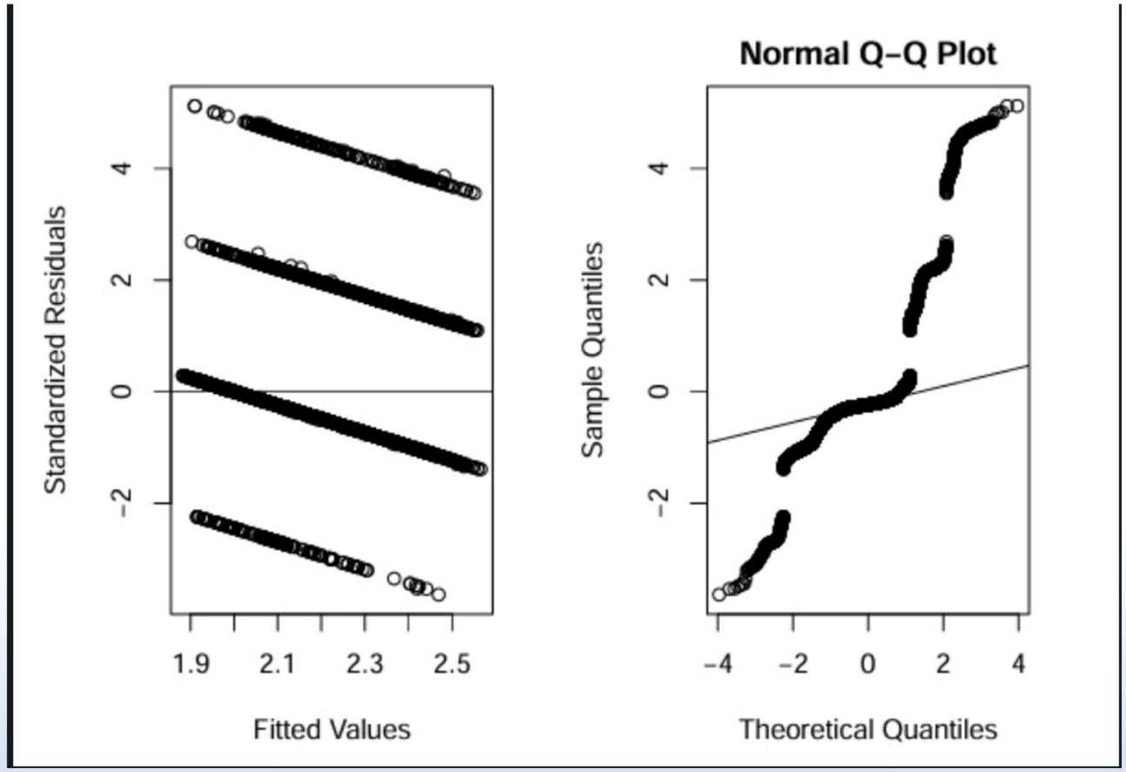




- **Weather Freezing** and **Precipitation** have the strongest positive effects on severity
- **Traffic Signal Flag** and **Road Features** show **negative effects**, meaning their presence reduces severity
- **Rainy** and **Snowy** weather both increase severity, with Rainy having a stronger impact
- **Weekend** accidents are slightly more severe on average
- **Pressure** and **Cloudy conditions** also contribute, but to a lesser extent
- **Rush Hour** appears to slightly reduce severity (possibly due to congestion reducing speed)

- **Colorado (CO)** shows the **highest positive effect** on accident severity
- **New Mexico (NM), Rhode Island (RI), Indiana (IN), and Illinois (IL)** also show significantly higher severity
- **South Dakota (SD)** is the only state in this list with a **negative effect**, indicating lower average severity
- Differences reflect **regional driving behavior, infrastructure, enforcement**, or reporting standards
- State effects capture variability not explained by weather or road features alone

Diagnostics for reduced model



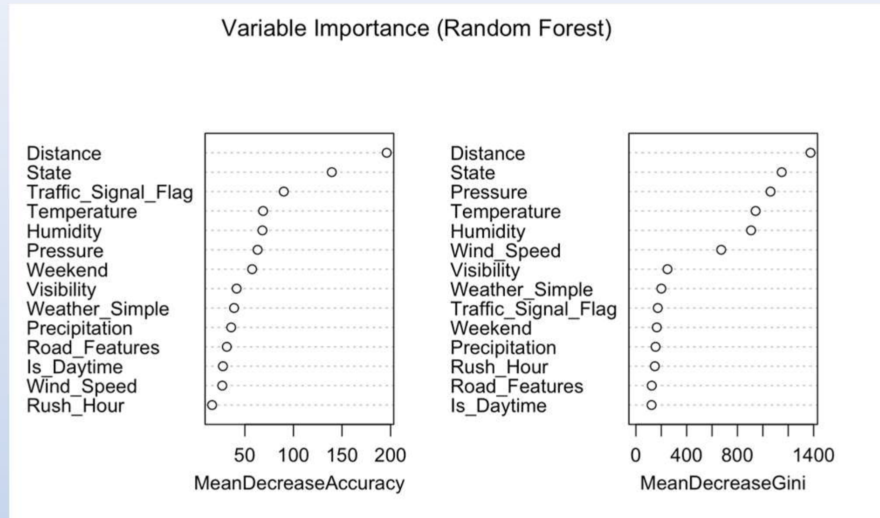
- Residuals vs. Fitted Plot
- Residuals form **distinct horizontal bands**, indicating discrete response levels (Severity = 1 to 4)
- No obvious **fan shape** or severe heteroscedasticity
- Pattern reflects the **ordinal nature** of the response variable rather than violation of assumptions

- Q-Q Plot
- Deviations from the diagonal line at both ends
- Suggests **non-normality** in residuals
- Caused by the **discrete structure** of the response, not by poor model fit

Random Forest model & Variable importance from Random Forest

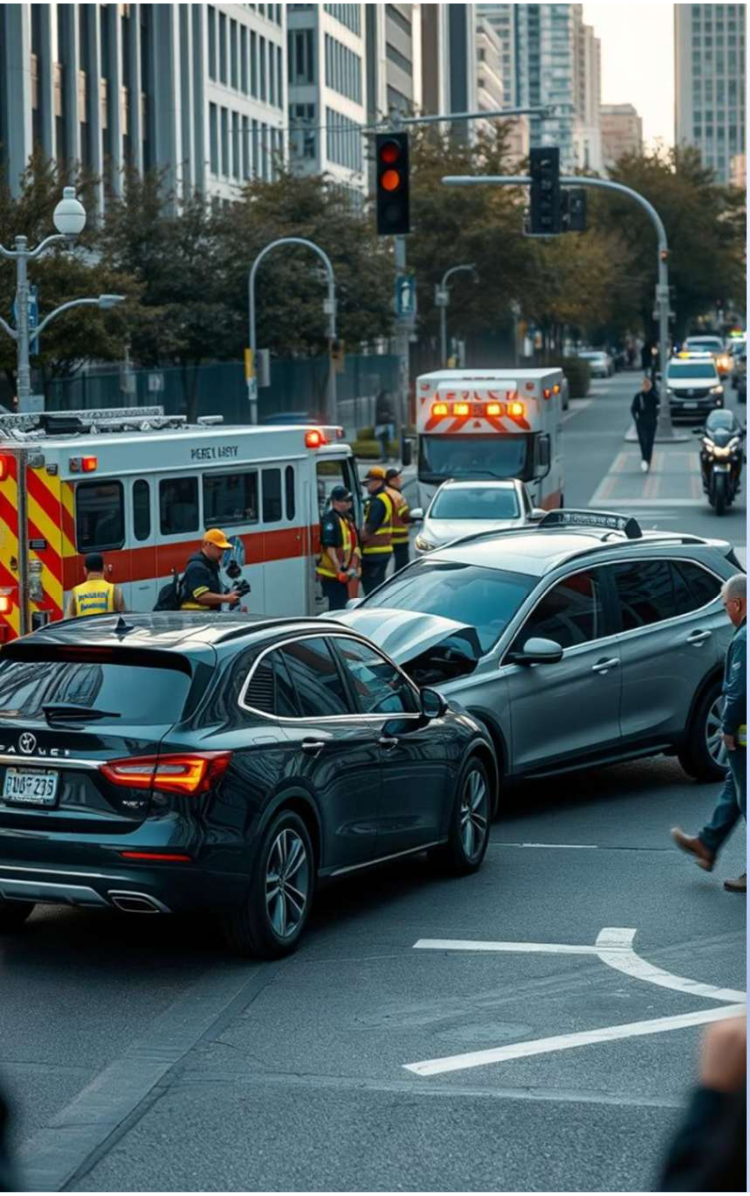
```
Call:
 randomForest(formula = Severity ~ Distance + Temperature + Humidity +      Visibility + Wind_Speed + Pressure + Precipitation
+ Weather_Simple +      Rush_Hour + Weekend + Is_Daytime + Traffic_Signal_Flag +      Road_Features + State, data = rf_data,
ntree = 800, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 800
No. of variables tried at each split: 3

        OOB estimate of  error rate: 14.48%
Confusion matrix:
   1     2   3 4 class.error
1 23   382  14 0  0.94510740
2  8 27659 380 6  0.01404484
3  0  3232 741 0  0.81349106
4  0   784   6 4  0.99496222
[1] "Overall Accuracy: 85.52 %"
```

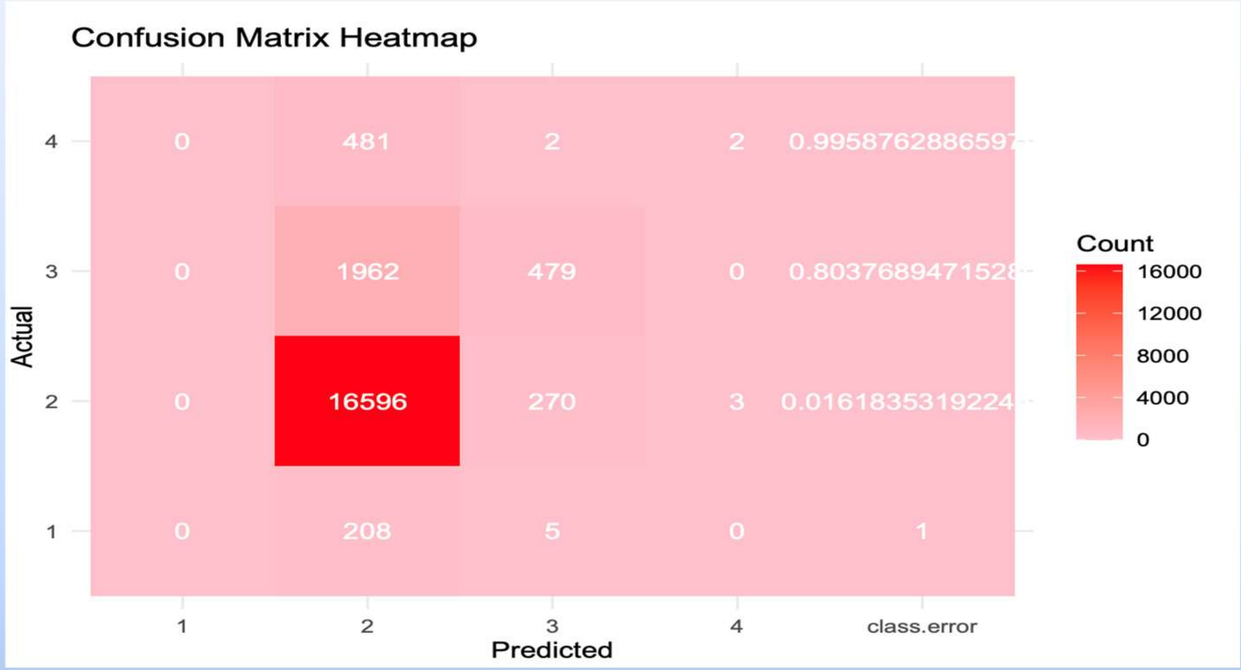


"Overall Accuracy: 85.52 %"

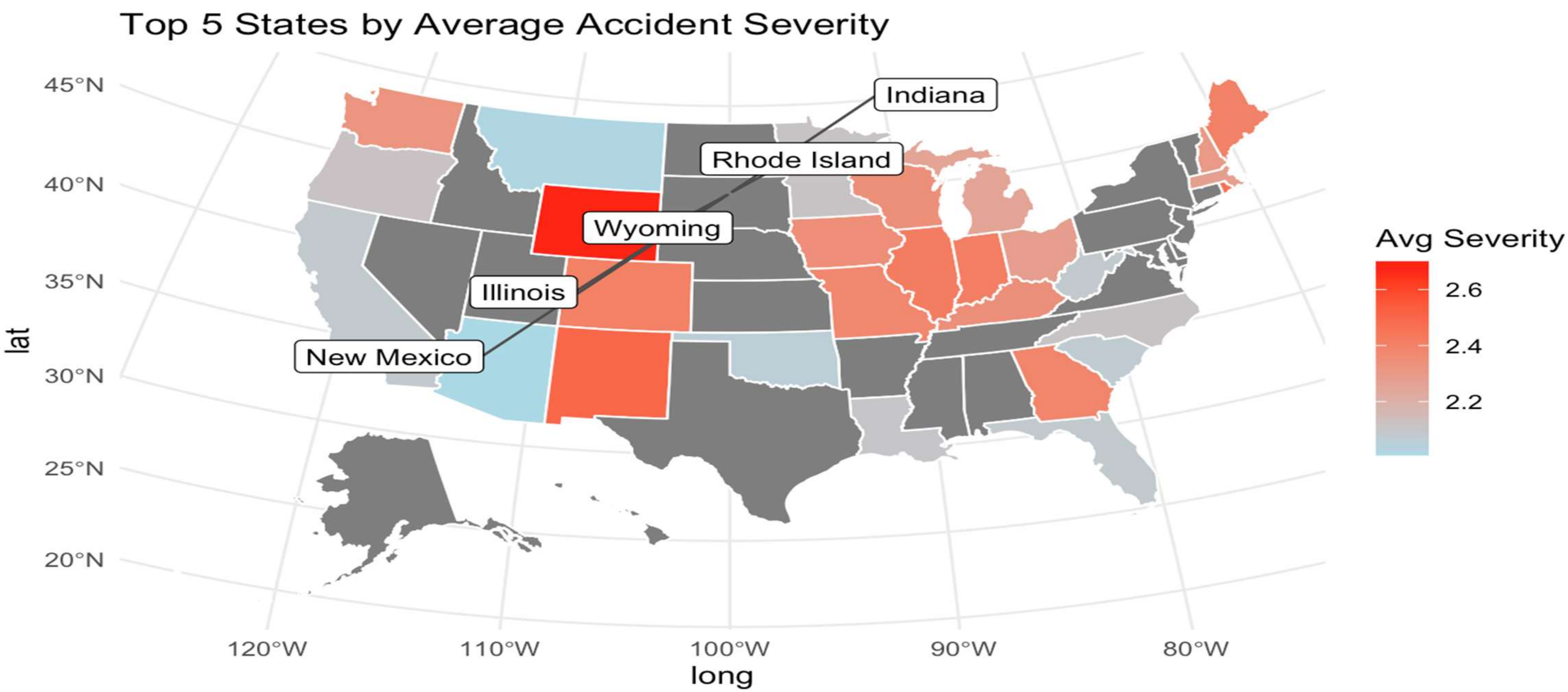
- Top predictors: Distance, State, Traffic Signal, Pressure
- Distance is the most important feature across both metrics
- State also highly influential — reflects geographic variation in severity
- Traffic Signal Flag has a clear role: accidents with signals tend to be less severe
- Weather and visibility play a moderate role
- Peak Hour and Road Features are less important in the model



Confusion matrix heatmap (Random Forest)



- The confusion matrix shows that the Random Forest model is highly accurate for the majority class (Severity = 2), but performs poorly on rare classes like Severity = 1 and 4.
- This imbalance skews predictions and suggests the need for class rebalancing (e.g., SMOTE, downsampling, or weighting) to improve generalization to all severity levels.



- Highlighted states: **Wyoming, Illinois, New Mexico, Rhode Island, Indiana**
- These states have the **highest average severity** across all crashes in the dataset
- Severity ranges from **~2.4 to 2.65**, with red indicating more severe crashes
- Patterns suggest possible **regional effects** beyond weather and road conditions
- Could reflect differences in **driving behavior, road design, enforcement, or reporting practices**

Conclusion:

- Linear models helped us interpret relationships, while random forest gave us better accuracy.
- For future work, we’d like to improve prediction for rare severity levels and include more time-based seasonal features.

THANK YOU
DRIVE SAFE.

"Every accident is a lesson. Let's learn and prevent."