**Predicting Motor Vehicle Accident Severity Within Individual States**

Contents

[**Executive Summary 3**](#_heading=h.1ci93xb)

[**Section 1: Introduction 3**](#_heading=h.3whwml4)

[**1.1**](#_heading=h.2bn6wsx) **WHY ACCIDENT SEVERITY MATTERS** 3

[**1.2**](#_heading=h.qsh70q) **OBJECTIVE** 5

[**Section 2: Preliminary Analysis 6**](#_heading=h.3as4poj)

[**2.1**](#_heading=h.49x2ik5) **DATABASE OVERVIEW** 6

[*2.1.1*](#_heading=h.2p2csry) *PREDICTORS 6*

[*2.1.2*](#_heading=h.147n2zr) *OVERVIEW BY STATES 7*

[*2.1.3*](#_heading=h.3o7alnk) *OVERVIEW BY TIME 7*

[**2.2**](#_heading=h.23ckvvd) **EXPLORATORY DATA ANALYSIS** 8

[*2.2.1*](#_heading=h.ihv636) *SEVERITY BY STATE 8*

[*2.2.2*](#_heading=h.32hioqz) *SEVERITY BY TIME OF THE DAY 9*

[*2.2.3*](#_heading=h.1hmsyys) *SEVERITY BY WEATHER CONDITIONS 10*

[*2.2.4*](#_heading=h.41mghml) *SEVERITY BY ACCIDENT SITE RELATED INFORMATION 12*

[*2.2.5*](#_heading=h.2grqrue) *CORRELATIONS BETWEEN PREDICTORS 12*

[**Section 3: Analysis and Results 14**](#_heading=h.vx1227)

[**3.1**](#_heading=h.3fwokq0) **DATA SELECTION AND TRANSFORMATION** 14

[**3.2**](#_heading=h.1v1yuxt) **METHODOLOGY** 14

[*3.2.1*](#_heading=h.4f1mdlm) *MODEL SELECTION: GENERALIZED LOGISTIC MIXED-EFFECTS MODEL 14*

[*3.2.2*](#_heading=h.2u6wntf) *L1-PENALIZED ESTIMATION FOR FEATURE SELECTION 15*

[*3.2.3*](#_heading=h.5xe4v2a4ukfx) *MODEL BUILDING: TOP-DOWN PROCESS 15*

[*3.2.4*](#_heading=h.xeiiaaygrd0b) *COMPARISON: MIXED-EFFECT MODELS VS GLM MODELS 16*

[**3.3**](#_heading=h.19c6y18) **RESULTS** 17

[**Section 4: Conclusion and Discussion 20**](#_heading=h.ibfjdrssr9a7)

[**4.1**](#_heading=h.3tbugp1) **CONCLUSION** 20

[**4.2**](#_heading=h.5xvbmk1fr65n) **DISCUSSION OF LIMITATIONS** 20

[**4.3**](#_heading=h.xhsnw159eix9) **FURTHER DISCUSSION** 22

[*4.3.1*](#_heading=h.8a79gtbhlwve) *POTENTIAL IMPROVEMENTS 22*

[*4.3.2*](#_heading=h.gerrqyd2525y) *MODEL COMPARISON BETWEEN RANDOM FOREST AND LOGISTIC REGRESSION 22*

[**Section 5: References 27**](#_heading=h.ne7cvmbwbzd)

[**Appendix A: Supplementary Graphs 28**](#_heading=h.28h4qwu)

[**Appendix B: R Code 34**](#_heading=h.nmf14n)

[**Appendix C: Python Code 38**](#_heading=h.37m2jsg)

Predicting Motor Vehicle Accident Severity Within Individual States

Applying Recent Data to Predict Accident Severity for Selected States

# Executive Summary

The trend of accident severity has a significant influence on the insurance cost and rate making strategies. Hence, to help insurance companies better underwrite a claim or predict the potential loss, in this report, we will introduce a framework of mixed-effect models that predict the accident severity given certain conditions. We use the continuous gathered data for the most recent years, to predict the accident severity of selected states in the United States with different geographical and weather conditions. It provides insights on the approach of using historical data to predict the severity of the accidents .

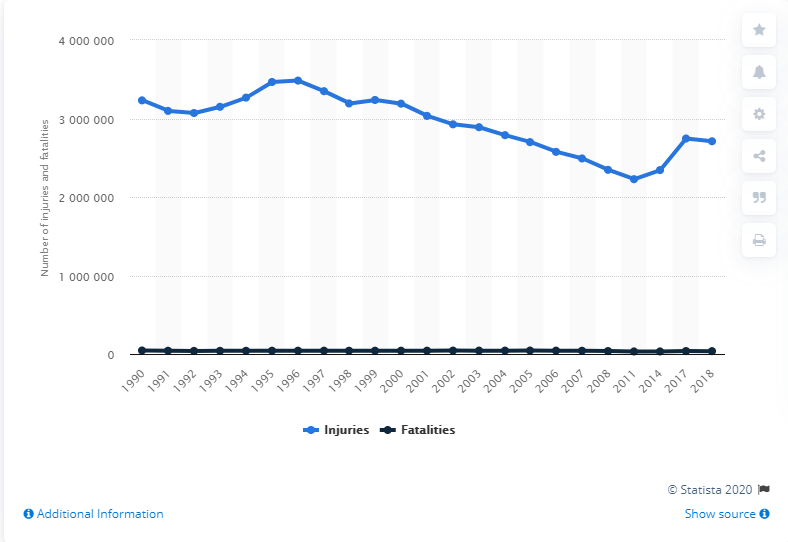
# Section 1: Introduction

## **1.1 WHY ACCIDENT SEVERITY MATTERS**

When analyzing the damage of motor vehicle accidents across a period of time, insurance companies generally can look at two factors: frequency and severity. The frequency of an event is how often an event occurs in a given period of time, while the severity of an accident is the loss size of an individual accident. In 2018, 36,560 fatal motor vehicle crashes occurred in the United States, resulting in $55 billion in medical costs, loss of income, and other immeasurable costs such as the burden on the victims’ friends and family (State-Specific Costs of Motor Vehicle Crash Deaths). However, while the number of fatal accidents was large, it was only a small proportion of the total number of traffic accidents. In 2018, there were 4.8 million motor vehicle accidents that occurred in total while 2.7 million accidents resulted in injuries (Wagner). Therefore, it is clear that high severity car crashes occur at a far lower rate than low severity accidents.

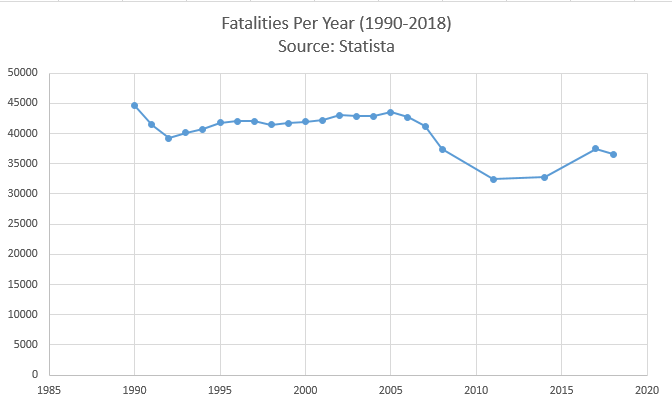
**Figure 1.1**

NUMBER OF ROAD TRAFFIC-RELATED INJURIES AND FATALITIES IN THE U.S. FROM 1990 TO 2018 (WAGNER)



**Figure 1.2**

NUMBER OF ROAD TRAFFIC-RELATED FATALITIES IN THE U.S. FROM 1990 TO 2018 (DATA BY STATISTA)



From Figure 1.1, we can see that the number of fatalities is completely shadowed by the number of injuries. In general, the number of accidents has remained relatively constant throughout time, with a slight negative trend. In Figure 1.2, it is clear that the trend of fatalities per year follows a similar pattern as the number of motor vehicle injuries overall. However, despite that many accidents don’t involve a fatality, the costs associated with an accident involving fatality is far greater than that of purely an accident resulting in injury.

**Figure 1.3**

AVERAGE ECONOMIC COST BY INJURY SEVERITY OR CRASH, 2018 (GUIDE TO CALCULATING COSTS)

**Figure 1.4:**

AVERAGE COMPREHENSIVE COST BY INJURY SEVERITY, 2018 (GUIDE TO CALCULATING COSTS)

Figure 1.3 displays the economic costs associated with injury severity in 2018. It displays the costs associated with death, disability, non-incapacitating visible injuries (labeled by ‘Evident’) possible injuries that aren’t visible (labeled by ‘Possible’), and property damage without injury. The economic costs associated with fatality is $1.66 million, completely overshadowing the economic costs associated with any sort of non-fatal injury. This pattern holds when we look at the comprehensive costs, which also include the value of decrease in quality of life as well as increased health risks, where the comprehensive cost of death is $10.86 million, the cost of disability is $1.187 million, while the costs of visible or nonvisible injuries are only $327,000 and $151,000, respectively.

Therefore, it is evident that an important distinction to look for when determining accident severity is differentiating death/disability and smaller injuries that don’t result in either. Doing so results in a classification problem to determine the following: What factors result in an accident having high severity (Death/Disability) or low severity (Visible/Possible injury). While low severity accidents are far more frequent, the monetary cost of high severity accidents can potentially more than offset the low frequency at which they occur.

## **1.2 OBJECTIVE**

The goal of this project is to predict whether or not an accident will be severe by using a logistic mixed-effects model, with States as a level. We will use various factors that were available at the time of the accident, such as weather attributes, period of day attributes, and address attributes to aid us. It will also provide insights on how to apply historical data into predicting accident severity.

# Section 2: Preliminary Analysis

## **2.1 DATABASE OVERVIEW**

The motor vehicle accident dataset contains 3.5 million accident records, covering 49 states of the United States, collected by Sobhan Moosavi[[1]](#footnote-1). The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

### 2.1.1 PREDICTORS

Each record in the database has 49 features. The features are described as below.

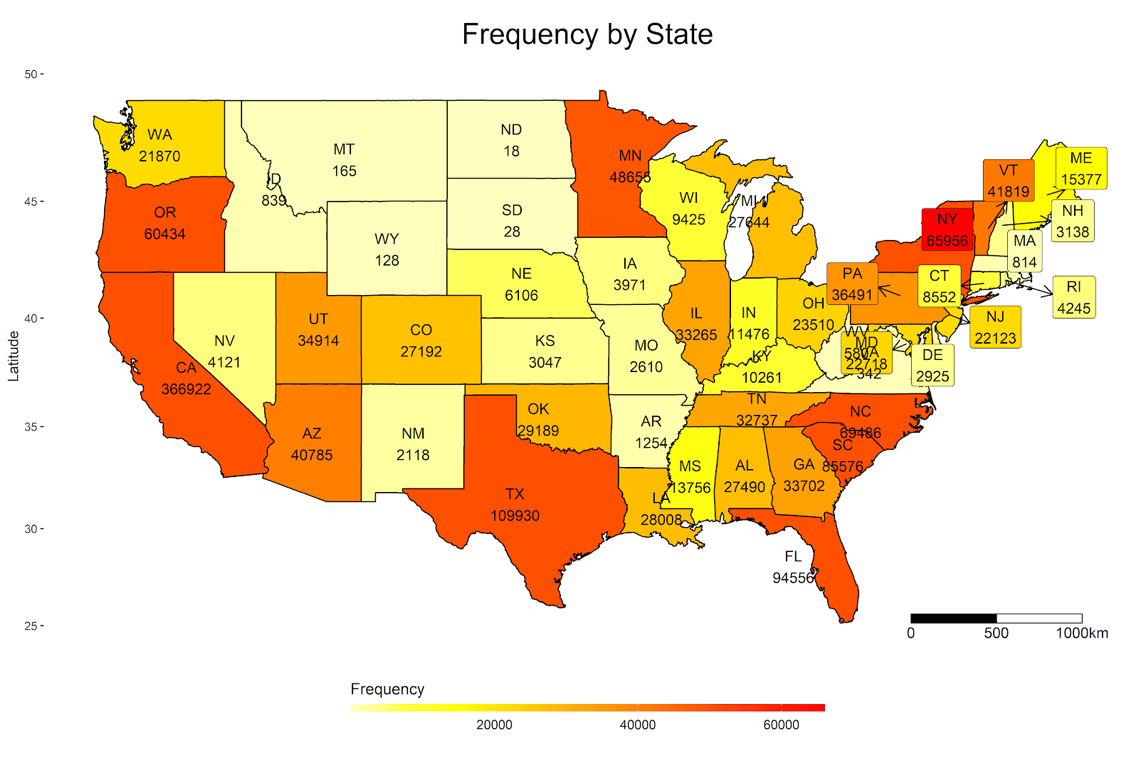
* Severity: A integer scale from 1 to 4, indicating the severity of the accident.
* Descriptive information: A set of predictors including unique ID, detailed description (if available), accident’s information from the TMC[[2]](#footnote-2) and accident reported sources (usually an API[[3]](#footnote-3) such as MapQuest).
* Time related Information: Seven predictors including the time zones the accidents resulted in, the durations of the accidents (if any traffic delay was created), the start times of the accidents and the period of the day the accident happens depending on four different definitions of twilight periods.
* Geographical Information: A set of 13 predictors including specific coordinates of the accident (both longitude and latitude), the specific address of the accident (from state level to the street number), the zip code of the area and the nearest weather station code.
* Site specific Information: A set of 14 predictors that indicates if the accident site has certain traffic amenity (such as traffic bump, traffic signals etc.) and the road condition (such as if there is a junction, railway and etc.).
* Weather related information: A set of 10 predictors indicate the weather conditions when the accidents happened, including humidity, air pressure, temperature, visibility and etc.

### 2.1.2 OVERVIEW BY STATES

The data records are from 49 states[[4]](#footnote-4), all states in the U.S. excluding Alaska and Hawaii. However, the data is not balanced across states. As shown in figure 2.1, in the states such as North Dakota or South Dakota, the data amount is surprisingly small for the most recent year and a half. There are 16 states with less than 5000 accident reports for the past year. Meanwhile, California has the most accident counts , 366,922 which has tripled the state with second most accident counts, Texas with a count of 109,930. In the next session, we will look closer into the distribution of different severities across different regions.

**Figure 2.1:**

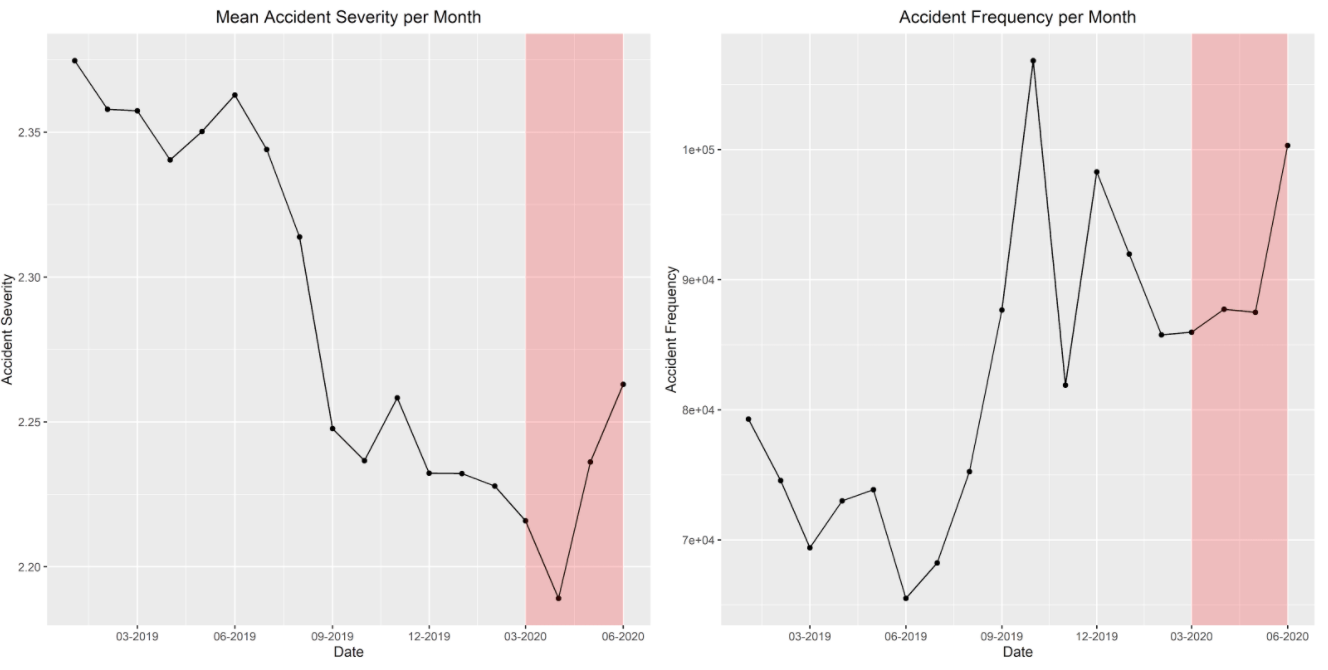
FREQUENCY OF THE ACCIDENT BY STATES FROM 2019 JAN TO 2020 JUN



### 2.1.3 OVERVIEW BY TIME

For the most recent year and a half, we can see that the frequency of the accident recorded is increasing along the timeline. More data is recorded for the most recent months. However, the severity decreases over time. The red shade highlights the period after COVID-19 pandemic strikes U.S.. The impact COVID given the short period of time is difficult to conclude.

**Figure 2.2:**

ACCIDENT RECORDED BY TIME FROM 2019 JAN TO 2020 JUN

## **2.2 EXPLORATORY DATA ANALYSIS**

Before building the model for prediction, it is necessary to check the underlying distributions of the severity against other predictors. In this session, we will analyze the severity using two different scales in order to visualize the result.

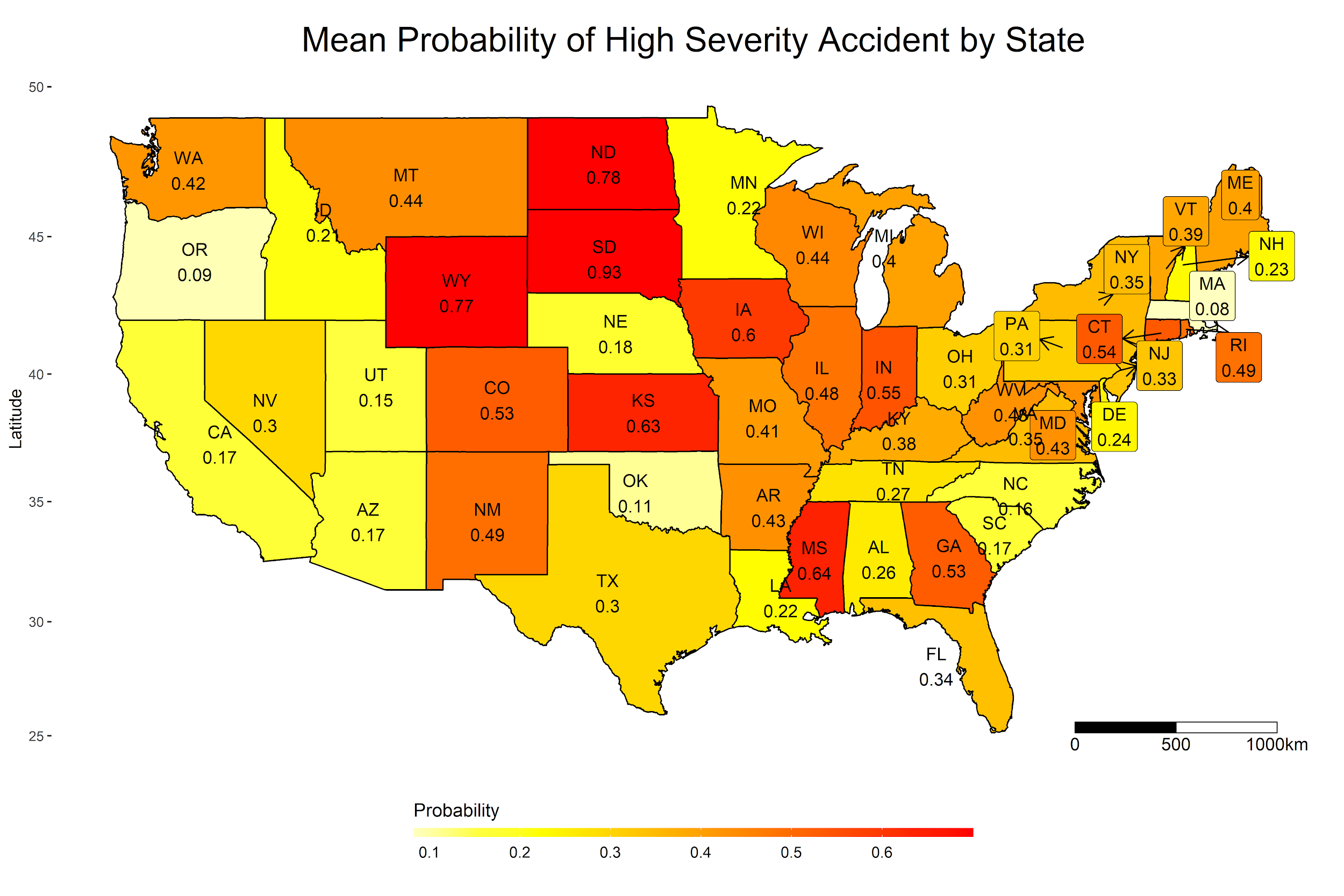
* Original Scale: From 1 to 4, which presents
  + 1: Chance of Property Damage, at least causing traffic delays.
  + 2: Chance of minor injuries
  + 3: Chance of serious injuries
  + 4: Chance of fatality
* Binary Scale: As discussed in Section 1, the financial impact magnitude between accidents involving death and disability was far greater than accidents without. Therefore we categorized the severity into Low and high levels:
  + Low Severity: Original scale 1 or 2, which means the accidents with no serious injuries.
  + High Severity: Original scale 3 or 4, which means at least serious injuries with possible case of death.

### 2.2.1 SEVERITY BY STATE

As we already checked the frequencies, we look at the proportion of high severity accidents to the low severity accidents inside each state. As shown in Figure 2.3, the proportion varies across states a lot. There exists a strong negative correlation between the proportion of severe accidents with frequency of the accidents. For the states with more than 70% proportion of high severe accidents, the states all have less than 5000 reported cases. For the states with more than 10,000 reported cases, the proportion ranges from 0.09 to 0.64 which still varies but on a smaller scale.

**Figure 2.3:**

PROPORTION OF THE HIGH SEVERITY ACCIDENTS BY STATES FROM 2019 JAN TO 2020 JUN



The correlation between the proportion and the frequency can be explained by the data collection method of the database. The reports are based on the reports from public traffic services. In states with fewer people, or mostly countryside, people tend to only report the serious accidents, and the low severity cases also tend to less likely affect the traffic, given there is a lower traffic on average.

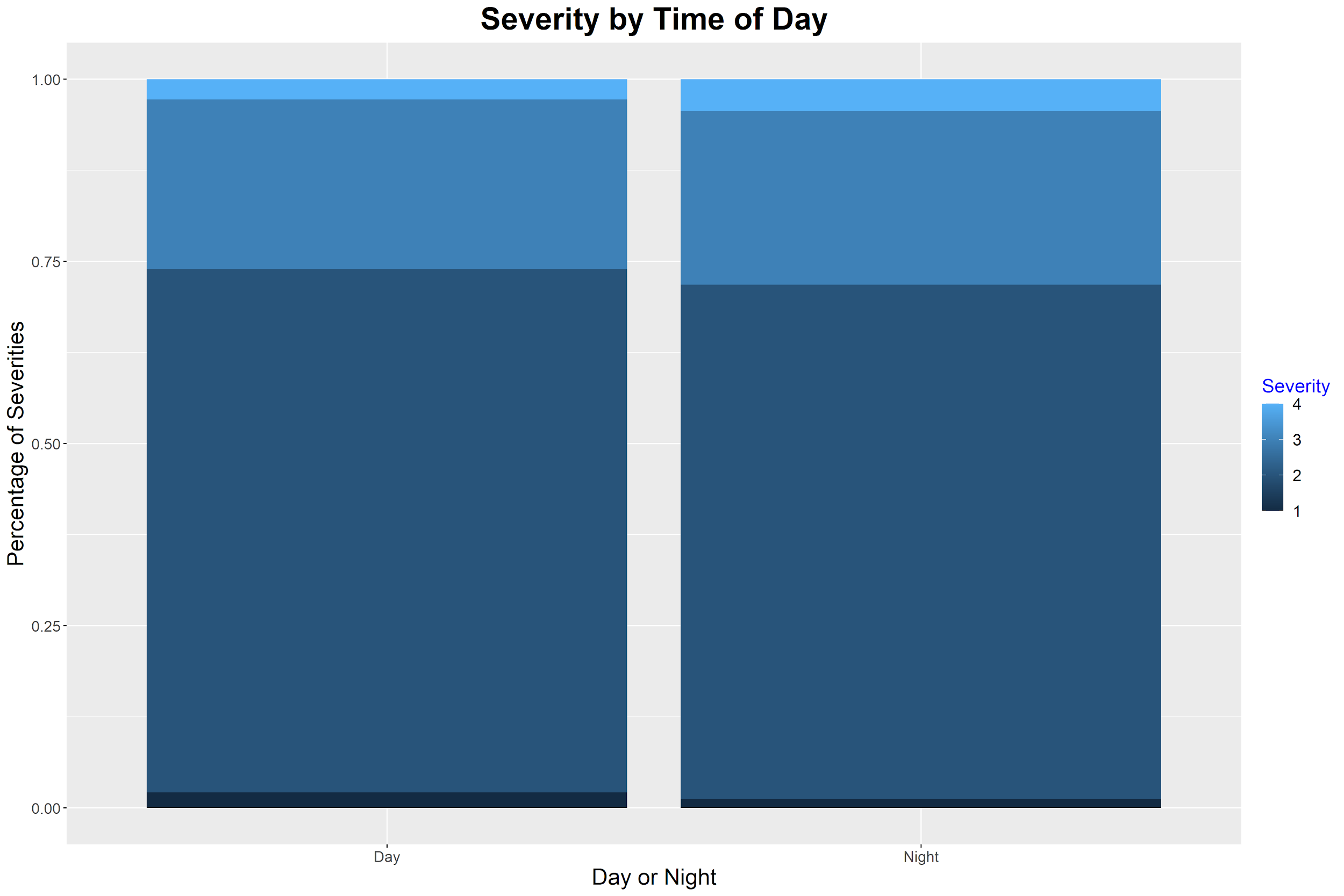
### 2.2.2 SEVERITY BY TIME OF THE DAY

In the database, the day period is divided into four different ways: Sunrise/Sunset, Civil Twilight, Nautical Twilight and Astronomical Twilight[[5]](#footnote-5). In our report, we choose to use the Sunrise/Sunset to determine the period of the day, because the definition is the most straight forward and we can only use one definition to decide the period of a day.

According to figure 2.4, the proportional amount of the accidents does not vary significantly with the time of a day. Nevertheless, there is still a higher proportion of high level severity accidents that happen during the night. Moreover, the graph also suggests that there is a significantly more amount of data with severity level of 2 than all the others.

**Figure 2.4:**

PERCENTAGE OF ACCIDENT WITH TIME OF A DAY

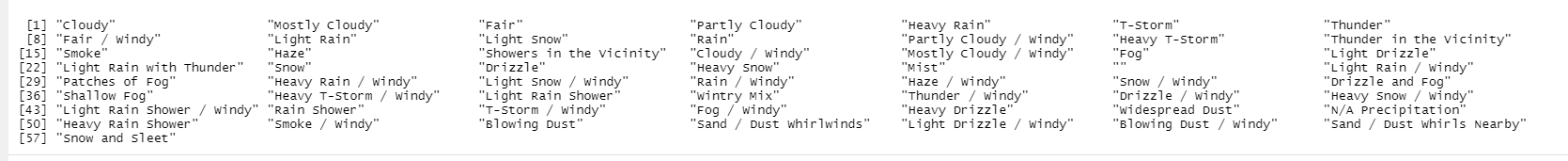


### 2.2.3 SEVERITY BY WEATHER CONDITIONS

In the database, there are 10 predictors related to weathers including temperature, wind speed, humidity, precipitation etc. There exists two categorical features: wind direction and weather conditions. However, as shown in figure 2.5, the weather is poorly categorized with many overlaps between the categories. Thus we will only consider the weather conditions that are continuously recorded. In this session, we will showcase the analysis of the correlations between the severity of accidents and the temperature.

**Figure 2.5:**

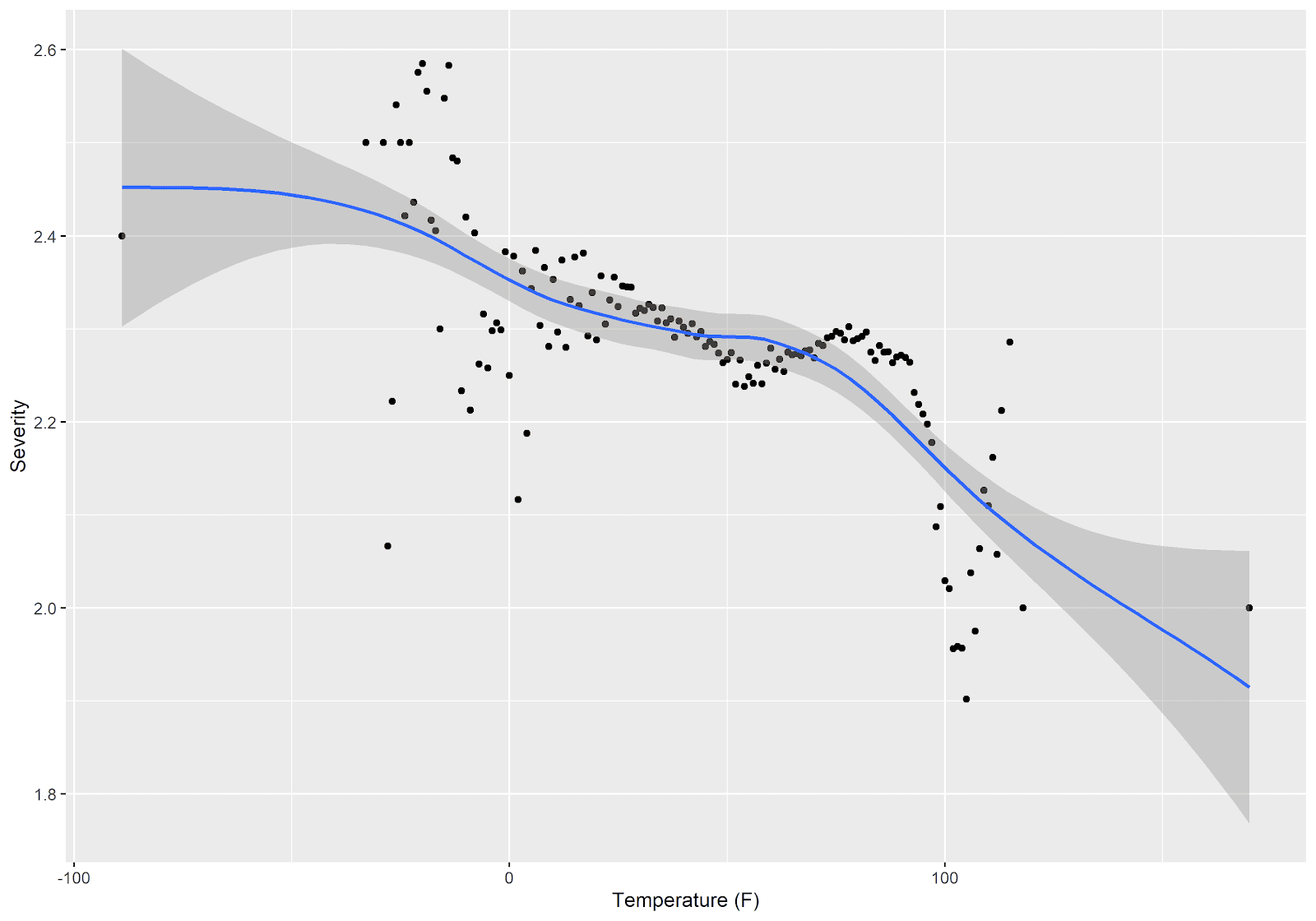
CATEGORIES OF THE VARIABLE NAMED WEATHER CONDITION



First we check the average severity that happened for each temperature. Overall there is a non-linear negative correlation between the two factors. When the temperature goes to the extreme values from both sides, the severity tends to increase. This pattern illustrated that the extreme temperature usually correlated in a higher chance of high severity accident.

**Figure 2.6:**

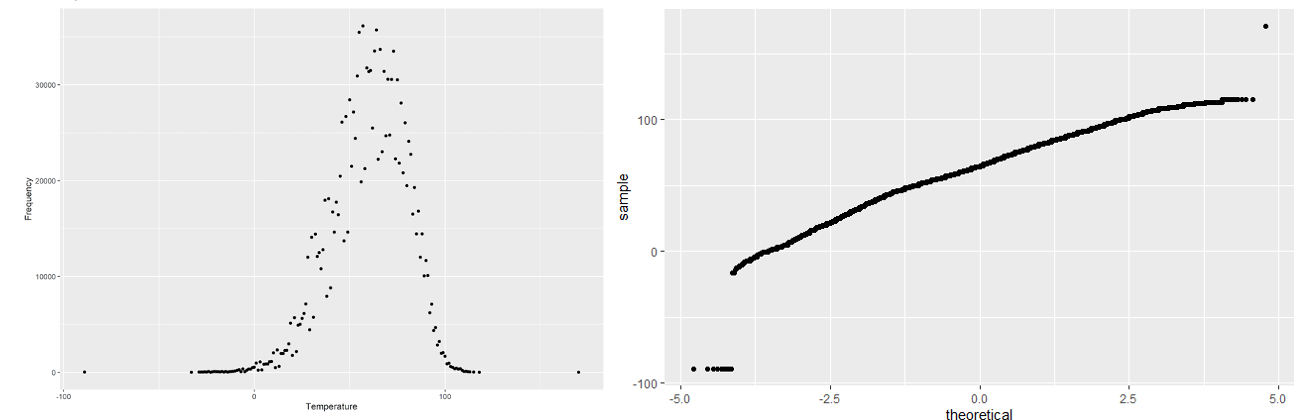
MEAN SEVERITY OF DIFFERENT TEMPERATURES



Because the temperature is measured continuously, it is necessary to also look at the distribution of the temperature over all. As shown in Figure 2.7, the temperature is slightly skewed to the left with an almost normal distribution. However, there exists outliers of the extreme temperatures on both sides, which may be due to recording errors.

**Figure 2.7:**

FREQUENCY OF THE TEMPERATURES



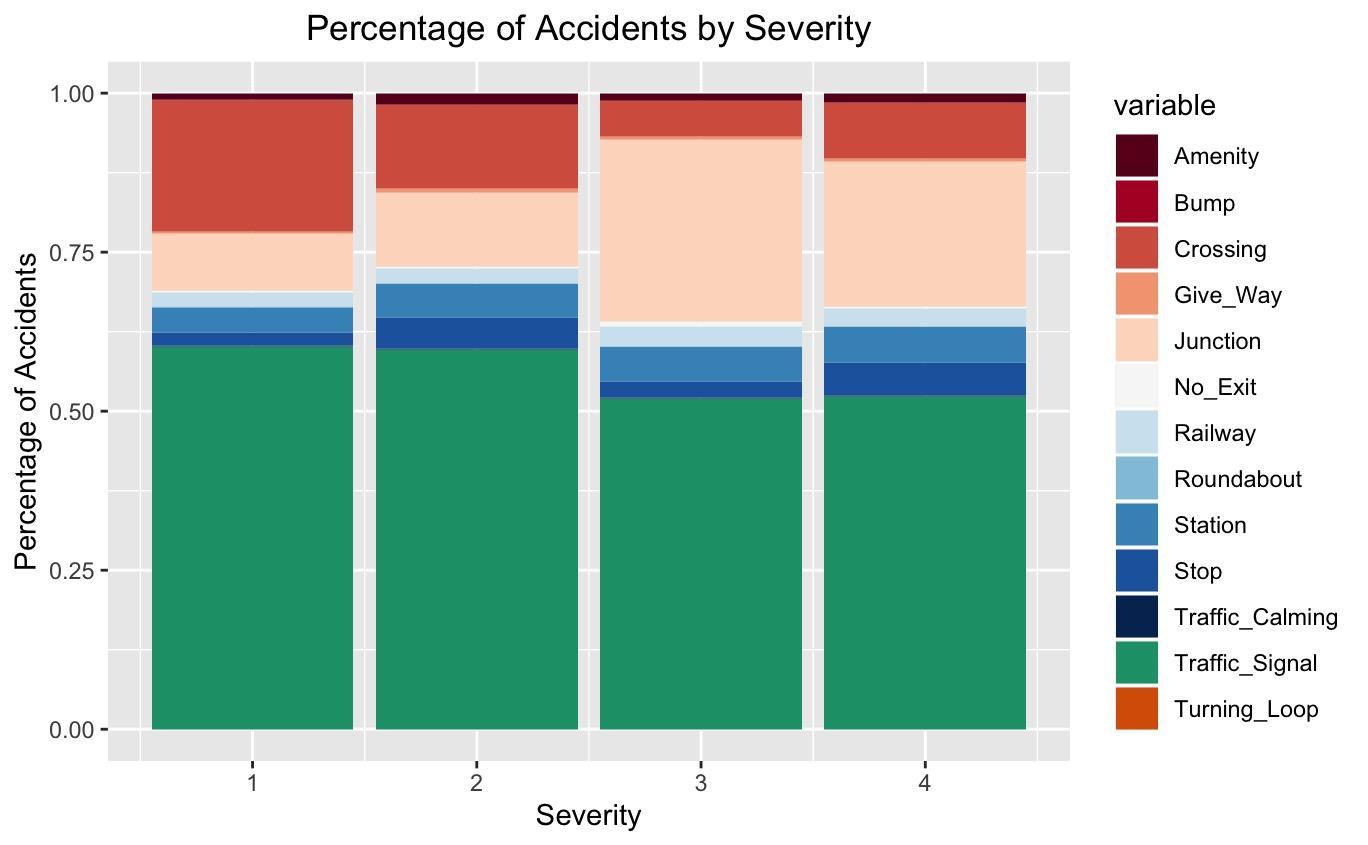
Similar analysis was performed on all of the predictors. Among all the predictors, precipitation, humidity and wind speed also showed positive correlations with the severity. Please see the Appendix for corresponding plots. It is important to note not all predictors follow a normal distribution. In the case of the abnormality, in the model section below we will discuss the use of transformation method in order to fit the model.

### 2.2.4 SEVERITY BY ACCIDENT SITE RELATED INFORMATION

The database contains 14 predictors of site specific information, such as junction, crossroad, stop sign, etc.. In this report, we use dummy variables for all the predictors, and plot the proportion of the feature presented at the site for different severity of the accidents. Most patterns tend to be random distributed across severity. However, for high level severities, there tends to be more junctions and less crossings and traffic signals presented on the site.

**Figure 2.8:**

PERCENTAGE OF ACCIDENTS BY SEVERITY

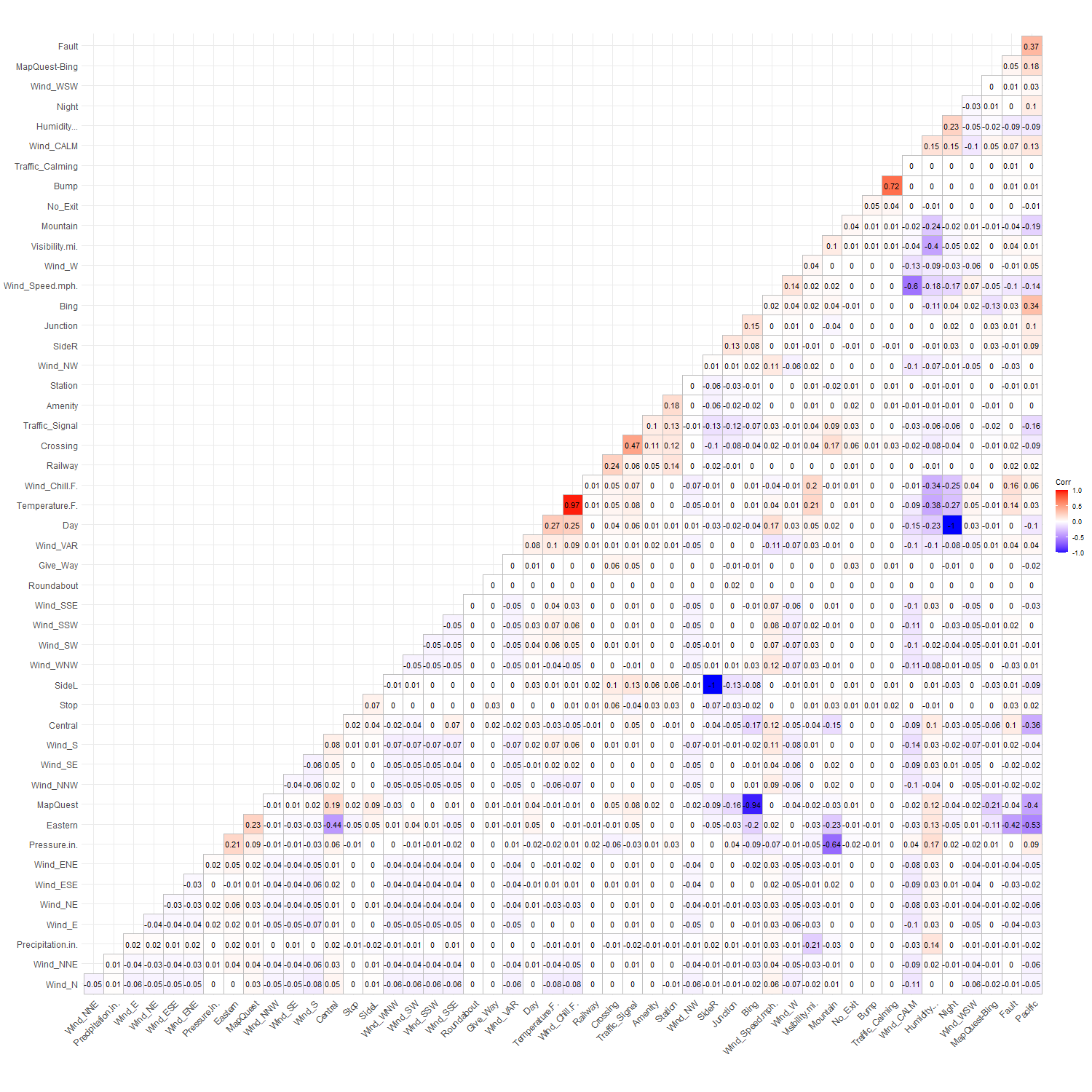


### 2.2.5 CORRELATIONS BETWEEN PREDICTORS

For all the predictors available in the dataset, we plot a correlation heat map according to the Pearson correlation coefficients to check potential covariate variables. Note that categorical variables are transformed into dummy variables which lead to the perfect correlations.

From figure 2.9 we observe that temperature and wind chill is highly correlated, since wind chill will always lead to the low temperature but not the other way around. The bump is highly correlated with traffic calming as well since sometimes traffic calming is identified as a bump. All the correlated variables will be discussed in Session 3 while building the model.

**Figure 2.9:**

CORRELATION HEATMAP FOR PREDICTORS

# Section 3: Analysis and Results

## **3.1 DATA SELECTION AND TRANSFORMATION**

In this report, due to different considerations, we selected our data as below:

* We selected the data from 2019 January to 2020 June, which can represent the most recent information.
* We chose to only use the top ten states with the most accident counts in our model, due to lack of computational power for a large dataset. The resulting dataset contains 962,303 records of accidents across ten states in the U.S.. Please see the Appendix for the states we chose. Even though this posted limitation on our results, the model can be extended to all states with the same methodology, and we will discuss it further in Section 4.
* We use the binary scale of severity as our response variable: low severity and high severity which are defined as in the Section 2.2.
* All the predictors we chose to use are observable at the time of accident, such as the side of the street, the temperature, the air pressure, wind conditions, the objects/landmarks in the near vicinity of the accident, and time of the day.
* In the data, no information about medical costs, degree of injury to the victim, insurance premium increases, and other post-accident data was available.
* We deleted the predictors which are neither categorical nor quantitative, such as description of the accident or street numbers.
* All the binary categorical variables are transformed into dummy variables for further uses.
* There are two categorical variables with multiple levels: wind direction and weather conditions. As mentioned in Section 2.2, the weather conditions are vaguely defined. Meanwhile, the wind direction shows no correlation with the response variable (correlation of 0.005). Thus, both categorical variables were dropped.
* Some of the quantitative variables such as wind speed and pressure, are abnormally distributed (See Appendix). We transformed these variables using a Box-Cox transformation with optimal parameters.
* Between covariate variables that had some degree of correlation, such as wind chill and temperature, we only kept the one variable which is more general, such as temperature in this case.

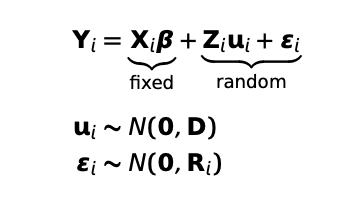
## **3.2 METHODOLOGY**

### 3.2.1 MODEL SELECTION: GENERALIZED LOGISTIC MIXED-EFFECTS MODEL

Because we determined that there was dependence in the probability of a high severity accident within each state through a chi-square test of independence (p < 0.01 for the 10 selected states), we decided that each state should be its own cluster, as not separating states into clusters would violate the assumption of independent observations in an ordinary least squares regression. Therefore, we decided to use a mixed-effects model to predict accident severity, with Level 1 being Individual accident and Level 2 being State. Since we decided to predict whether an accident had a high severity or a low severity, and this only has two classes, we decided to use a generalized logistic mixed-effects model**,** with a Binomial(logit) link function.

**Figure 3.1:**

GENERAL LINEAR MIXED MODEL FORMULA



A mixed-effect model allows us to preserve the fixed-effects coefficients, and also add random effects that can reflect different conditions at the State level. Additionally, it allows us to handle unbalanced data between clusters. Since each state has a vastly different population, such as California vs Alabama, the clusters are unbalanced. A mixed-effects model handles this well as groups with less data will automatically get shrunk towards the overall population mean. Finally, this mixed-effects model allows us to model a non-continuous predictor, as it can be generalized to a logistic regression.

### 3.2.2 L1-PENALIZED ESTIMATION FOR FEATURE SELECTION

Generalized linear mixed models are restricted to few covariates, because the presence of many predictors yields unstable estimates. We could include an L 1-penalty term that enforces variable selection and shrinkage simultaneously to fit generalized linear mixed models. Because our dataset contains a large number of potentially influential explanatory variables. We can use a gradient descent algorithm to maximize the penalized log-likelihood yielding models with reduced complexity. However, the result of feature selection returns with a fairly small optimal parameter which leads to that all the features are retained. Thus, further model selection is introduced in the next session.

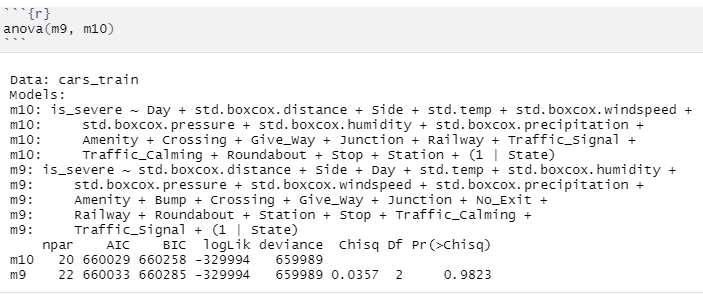
### 3.2.3 MODEL BUILDING: TOP-DOWN PROCESS

We used a top-down model building process in order to obtain the model with the best predictive power, by both p-values to determine significance of variables before removing them from the full-model and then using a likelihood ratio test on the full-model and the reduced model to test the significance of those variables. We began by including all potential features we get from previous sessions to predict Is Severe: Is Day, Distance, Side, Temperature, Wind Speed, Pressure, Humidity, Precipitation, Amenity, Crossing, Give Way, Junction, Railway, Traffic Signal, Traffic Calming, Roundabout, No Exit, and Traffic Bump. We then determined that some of the traffic-specific features, such as Traffic Calming and Traffic Bump were very highly correlated as determined in the correlation heatmap Figure 2.9.

For example, in our first step we test the model with hypothesis:

**Figure 3.2:**

LIKELIHOOD RATIO TEST RESULT



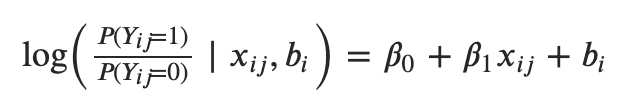
By fitting the model with and without “bump”, the likelihood ratio test returns a p-value of 0.98, which leads us to conclude that we should not reject the null hypothesis and that the reduced model performs better.

Similarly we repeated the process for multiple times, until there is no significant p-value from the likelihood ratio test for any reduced models. The resulting final model is discussed in the next session.

### 3.2.4 COMPARISON: MIXED-EFFECT MODELS VS GLM MODELS

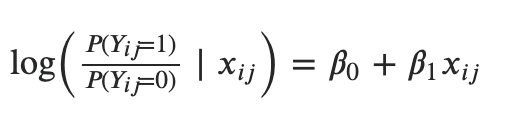
We also built GLM models to compare with GLMER models. The GLMER model concludes mixed effects, so the biggest difference between GLMER and GLM in our case is that GLM considers the various conditions among different states when an accident occurs .

When modeling GLMER, the odds ratio of severity of accident is different based on 𝑏i when occurring in different states. We modeled the expected conditions on the fixed design matrix and random effects among different states.



When modeling GLM, we didn’t account for the random effects and only focused on the fixed design matrix. The model can be seen as ‘nation-wise’ motor vehicle accident severity.

We can find that our estimates based on mixed-effect models differ from the ones based on GLM models.

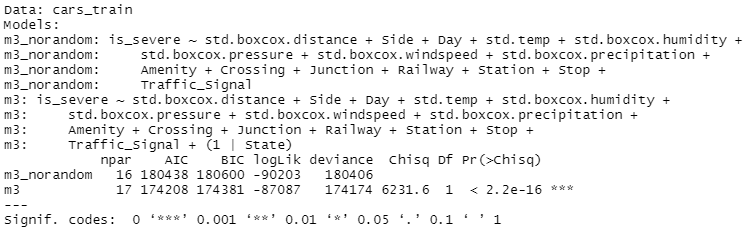


The results were compared using a likelihood ratio test between two models. Since the logistic model with no mixed effects eliminated all the random parameters, it was considered the reduced model. Note that the random effects should be compared using REML instead of maximum likelihood, since it is the variance of random effects we are eliminating.

According to the result in Figure 3.1, the P-value was small and we were able to conclude that the mixed-model provides a better prediction than the logistic model with no random effects.

**Figure 3.3:**

COMPARISON BETWEEN LOGISTIC MODEL WITH AND WITHOUT MIXED EFFECTS



## **3.3 RESULTS**

Following the methods discussed in the previous session, the model with all predictors involved without doing any top-down feature selection are discussed first.

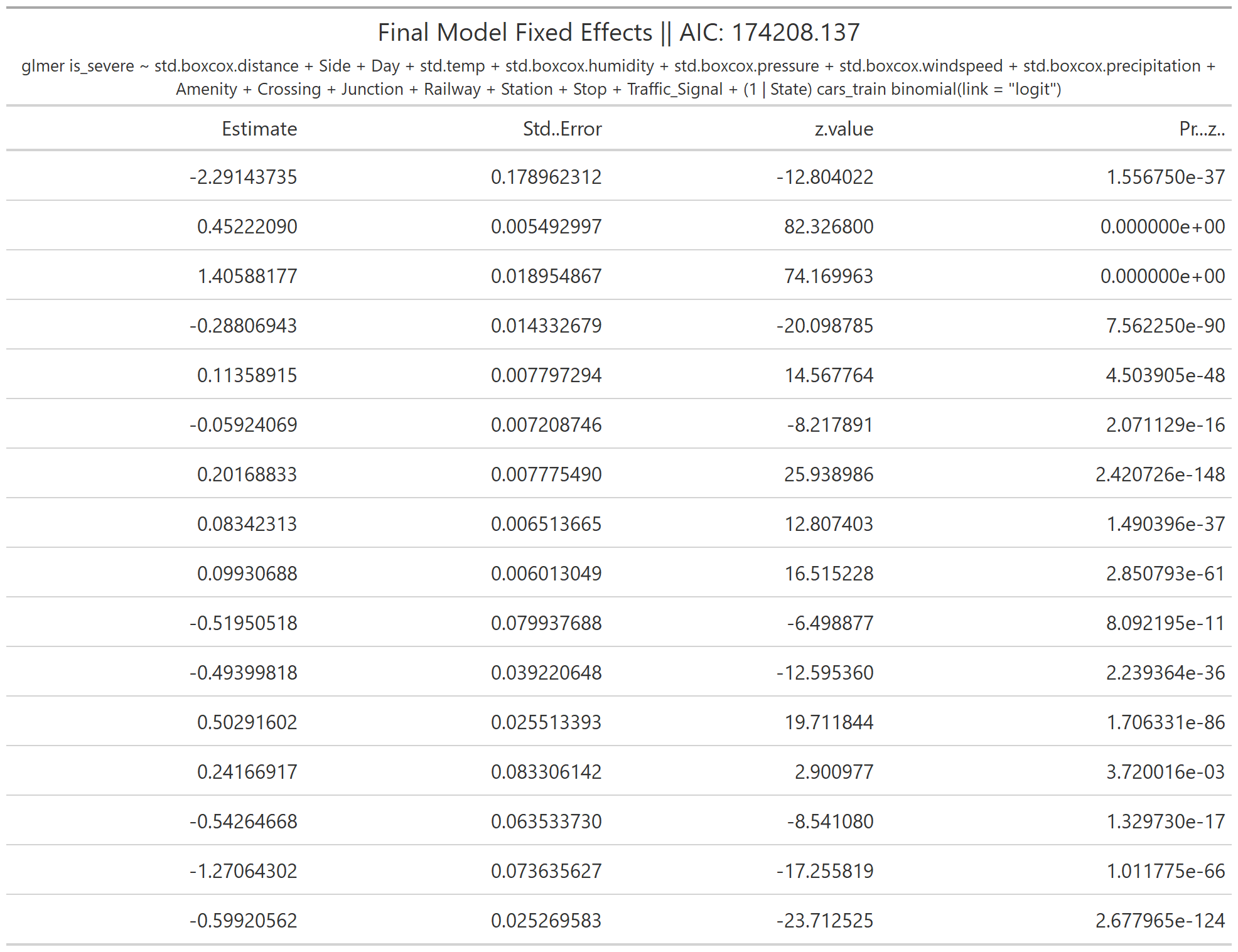
According to the resulting coefficients, there were several parameters we chose to drop following a top-down feature selection process.

Our full model showed that there were 3 features that were not significant, and we decided to drop some of the features considering the multicollinearity effect as mentioned in the previous session, so we got our adjusted further model. However, After testing for each feature, we discovered that the traffic\_calming was still not significant, as well as Roundabout. Therefore, we attempted to drop these 2 features for a more efficient result. Thus, we arrived at a better model. The specific result of full and intermediate models are included in Appendix A.

In the final model, we got rid of the Traffic\_Calming and the Roundabout variables. After doing some further analysis, there existed fewer observations in our model, as well as the AIC-scores was 174208, which was accepted.

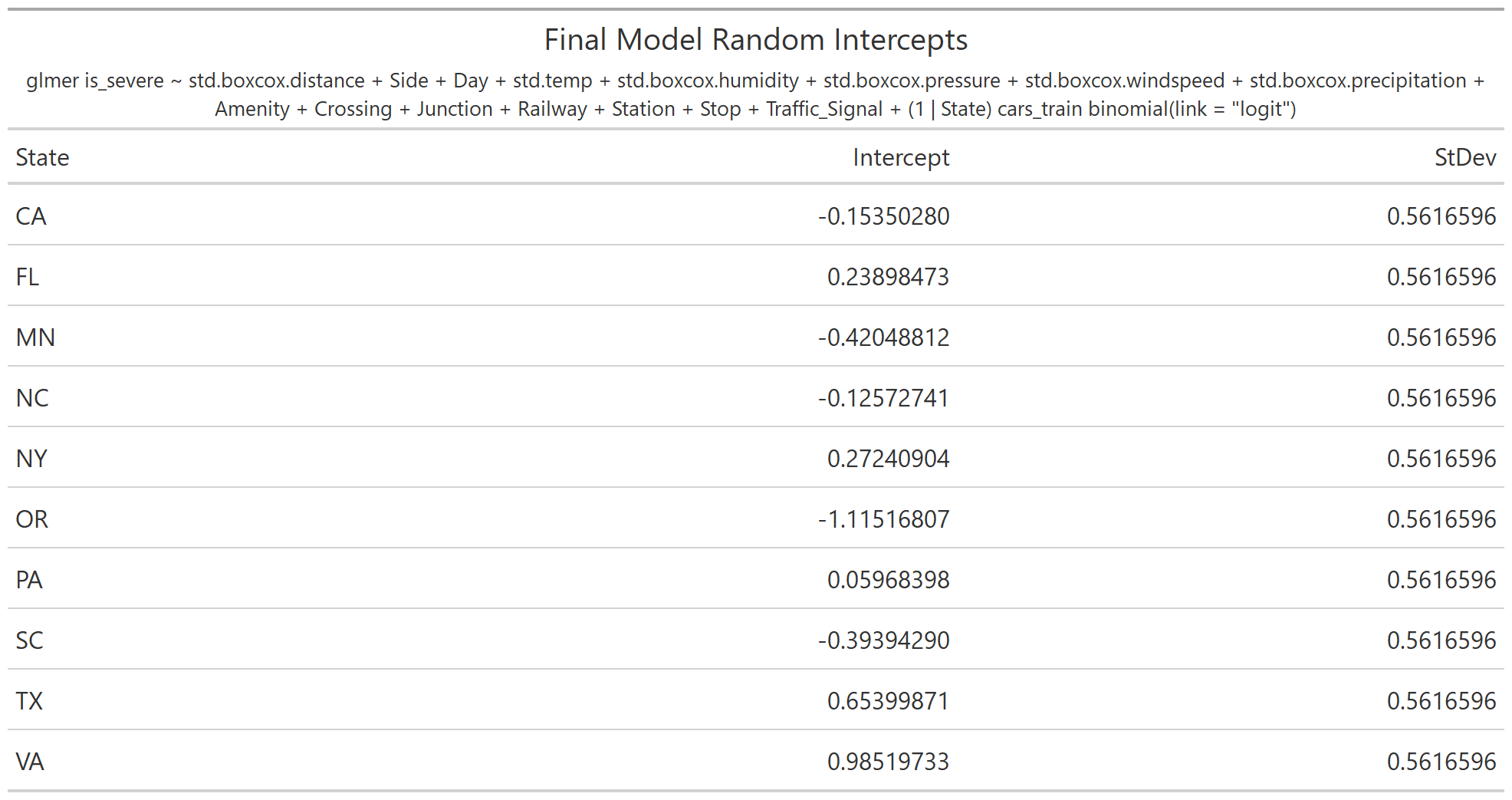
**Figure 3.4:**

FINAL MODEL RESULT



**Figure 3.4:**

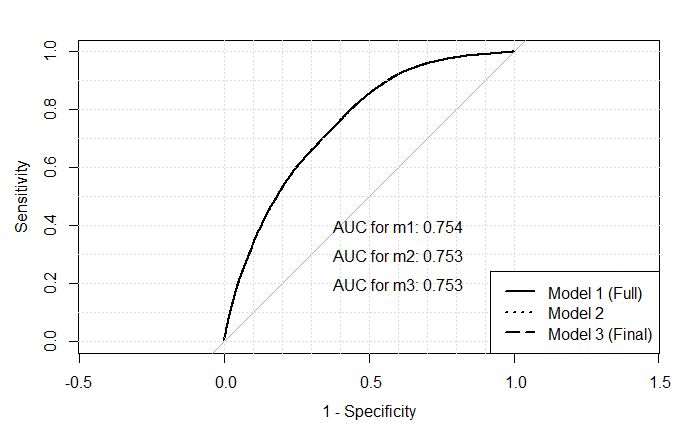
FINAL MODEL RANDOM INTERCEPTION



As the in figure 3.5 showed, our full model had a fair score of 0.754. After our analysis, our AUC decreased to 0.753. While the AUC was better in the Full model in comparison to the Final model, we built the model using AIC. The full model contained 20 variables, while the final model contained 15 variables, which shows that the final model, while resulting in a slightly lower AUC, is far more parsimonious. Since we were using raw data, which were disturbed by other factors, these factors may not appear in the collected data. Therefore, we needed more features to facilitate more accurate feature engineering.

**Figure 3.5:**

ROC CURVES

.

# Section 4: Conclusion and Discussion

## **4.1 CONCLUSION**

In order to predict whether or not an accident will be severe, we chose various factors that were available at the time of the accident, including weather conditions, time of the event, road conditions etc. Because the financial cost of an accident is directly tied to severity, we categorized the accident severity into 2 levels: low severity (no serious injuries), and high severity (serious injuries with possible death). We began exploring the data by first plotting a map of the probability of high severity by state and a line graph of the severity levels by time, and then used EDA to help us identify potential problems. After a thorough overview of our data, we decided to use a generalized logistic linear mixed model with two levels (Accident and State) as our main method to predict severity.

After doing selections and transformations to our original dataset, we used the top-down model building process to fit the data in a full model with state effects. After that, we dropped the variables that may cause multicollinearity, as well as the features that were not significant according to the summary of the model. After several steps of feature selection, we finalized our model to only include 15 features and the state effects. Although the AIC for the final model was slightly higher than the adjusted model (174208.137 for the final and 174208.113 for the adjusted model), we selected the model with one less independent variable. Also, according to the BIC criteria (174380.8 for the final model and 174391.0 for the adjusted model), our final model provided a better fit.

We used the final model to predict the severity. Our model provides a better specificity with 0.7997, which provides us that it can correctly predict 80% of the results that are categorized as low severity accidents. A relative lower sensitivity with 55% may result in prediction that leads to sudden loss for the insurance companies. In the next session we will also discuss other machine learning processes that yield similar results. Such a result may be caused by our unbalanced data structure of the response variable. Thus, boosting method and importance of cross validation will also be discussed in the next session.

## **4.2 DISCUSSION OF LIMITATIONS**

* The data selection:

The current model only contains the top 10 states with the most accident counts. As we discussed in Section 3, running the full database with 1.5 million data points with 40+ features each is very computationally expensive. However, limiting the selection of states certainly inhibits the scope of our results as we only considered part of the data. To correctly predict the accident severity in a national range, it is possible to fit the entire dataset on all the states, which may result in different random intercepts for the model and provide a more generalized result.

* Lack of data in some states:

Figure 2.3, indicated that some midwest states, such as North Dakota, Wyoming, Montana, and West Virginia, all appeared to have higher severities than others. This situation could be caused by lack of data. The high severity was likely to be caused by low accident counts since the population is so low in those states.

Meanwhile, Figure 2.1 also indicated that those midwest states had very low accident frequencies. When we try to analyze the whole U.S. accidents dataset by utilizing each state, it could lead to high volatility and inaccurate results. So from whatever aspect we analyze, the lack of data for those states significantly influenced our countrywide analysis, which caused us to drop certain states, limiting the scope of our predictive power.

* Too many assumptions for our model:

We have seen that linear mixed models offer a powerful and flexible tool for analysis of clustered and longitudinal data. In this case, we desired to use each state as an individual cluster in order to maximize the effectiveness of a linear mixed model. However, a potential disadvantage of linear mixed models is that more distributional assumptions need to be made. Moreover, usually approximations have to be used to estimate certain parameters of the model. As a consequence, conclusions depend on more assumptions, increasing the risk of misspecifying the model and hence biased parameter estimates.

* Multicollinearity between features/lack of domain knowledge of some features:

Our feature correlation matrix and full model showed that we had many correlated features in our research process. For example, the Bump and the Traffic\_Calming variables had a correlation score of 0.72, while Day and Wind\_Chill.F had a correlation score of 0.97. After doing some more research in the variables, we discovered that Traffic\_Calming is a feature in the road used to slow down traffic, while a speed bump serves a similar purpose. Therefore, we needed to do some more research into the individual variables and fully understand their influence in order to get rid of the multicollinearity effect and obtain a meaningful result.

* Inefficiency of the Linear Mixed Model runtime:

While a linear mixed model offers a powerful way to analyze clustered data, a tradeoff of this is the fact that more processing power is required in order to achieve a result. This isn’t meaningful for smaller datasets, but is far more impactful in larger datasets, such as the one used in this study. This limitation means that we cannot use certain feature selection tools, such as automatic stepwise selection, in order to determine which features we wish to include in the final model.

## **4.3 FURTHER DISCUSSION**

### 4.3.1 POTENTIAL IMPROVEMENTS

In the future studies on the objectives of predicting accident severity, the following directions can be considered to improve the performance of the overall model.

* Given the state and specific situations, we can predict the severity of an accident.
* Analyze the severity of an accident, or the frequency within a certain time period, or consider other significant predictors could help insurance companies to underwrite and ratemaking their current insurances.
* We can run the same analysis, with a longer time frame (going back to February 2016) or with all states, given access to more powerful processing power.
* We divided the data by time into two periods: 2019.1 to 2020.3 is pre COVID-19 time and 2020.3 - 2020.6 is post COVID-19 time. In the future, we can analyze the data as a time series with pre/post COVID, and/or with seasonality effects.
* In our model, we did not use any measures to estimate the effects of overfitting the model. Cross-Validation can be applied while training the data to minimize the bias introduced by overfitting the data.
* The Generalized Linear Mixed Model as linear predictor contains random effects in addition to the usual fixed effects, but would be estimated as a one step regression rather than Expectation Maximisation model. For further studies, we could expand to a generalized linear mixed model if take the severity as categorical variable.

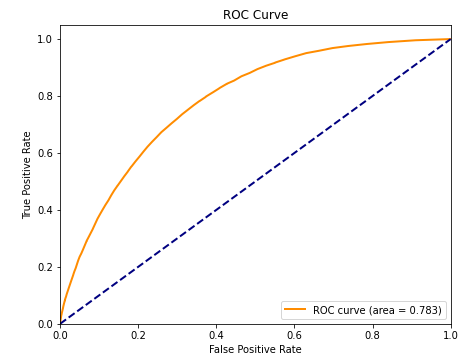
### 4.3.2 MODEL COMPARISON BETWEEN RANDOM FOREST AND LOGISTIC REGRESSION

We also tried using other methods including some machine learning methods to do the classification to predict the result.

The machine learning approach we mainly compared with is Random forest, which is an ensemble learning method for classification, that operates by constructing many decision trees at training time and usually outputting the class that is the mode of the all classes. We performed the same analysis using Random Forest on the same dataset, and obtained the following ROC curve.

**Figure 4.1:**

ROC CURVE FOR RANDOM FOREST



The Area Under the Curve (AUC) is 0.783, while it was only 0.753 for our GLMM. This is because our data, as we determined, was not very linear, which makes the random forest algorithm quite robust. Additionally, our data as we stated earlier was quite unbalanced across states. These characteristics make a Random Forest a powerful tool to optimize predictive accuracy.

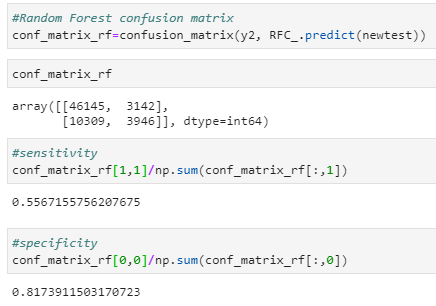
However, due to the fact that a Random Forest handles unbalanced data better than linear models, this may cause potential overfitting. Some states, such as California, had an almost 1/3 population size, which may overpower other states. In that sense, machine learning processes with boosting strategies such as XGboost may provide better outcomes.

For higher dimensional models using random forest, the parameter tuning is also important as well as test the significance of the features. In our report, we did not estimate any of the parameters based on our data, which may result in overfitting the model as well.

One disadvantage of the Random Forest algorithm in this situation is that the model itself is less interpretable than a GLMM, and we cannot as easily determine the relationships between different states and their probabilities of severity. Also the influence of each predictor is hard to access and extract from the model.

**Figure 4.2:**

RANDOM FOREST CONFUSION MATRIX

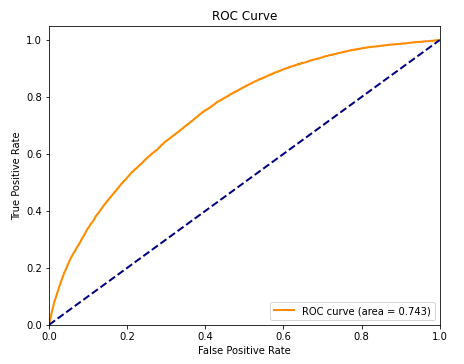
  
Using the Random Forest, we obtained a sensitivity of 0.557, and a sensitivity of 0.817. The sensitivity was lower than the GLMM, but the sensitivity was higher. This means that using a random forest, we have a lower chance of achieving a true positive when compared to a false positive, but a lower chance of achieving a false negative when compared to a true negative. This may be detrimental to the overall model, since predicting a false positive in this situation means that the insurance company’s predictions of high severity accidents would be 55.7% correct, which is less than that of the GLMM. In general, it is more beneficial for an insurance company to predict high severity accidents properly than it is low severity accidents.

To compare with we will also discuss the method we used involving Logistic Regression

. Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. Like the Random Forest, we performed the same analysis using Logistic Regression on the same dataset, and obtained the following ROC curve.

**Figure 4.3:**

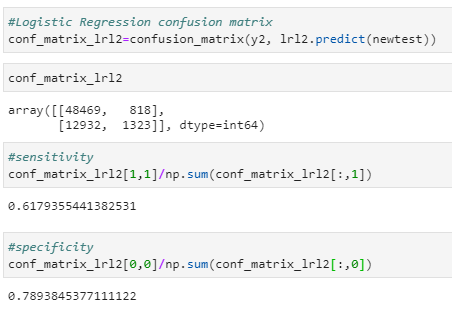
LOGISTIC REGRESSION ROC CURVE



The Area Under the Curve (AUC) was 0.743, while it was 0.753 for our GLMM, when running the Logistic Regression with the same features on the same dataset, treating State as a categorical feature instead of a Level. This decrease in predictive power may be due to not capturing the random baseline effect of each state as well as the Logistic Linear Mixed Model.

**Figure 4.4:**

LOGSITIC REGRESSION CONFUSION MATRIX

  
Using the logistic regression, we obtained a sensitivity of 0.618, and a sensitivity of 0.789. The sensitivity was higher than the GLMM, but the sensitivity was lower. This means that using logistic regression, we have a higher chance of achieving a true positive when compared to a false positive, but a higher chance of achieving a false negative when compared to a true negative. This may be beneficial to the overall model, since predicting a false positive in this situation means that the insurance company’s predictions of high severity accidents would be 61.8% correct, which is an improvement over the GLMM.

# Section 5: References

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. [“A Countrywide Traffic Accident Dataset.”](https://arxiv.org/abs/1906.05409), arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. [“Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.”](https://arxiv.org/abs/1909.09638) In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[A. Groll](https://www.semanticscholar.org/author/A.-Groll/115638162), [G. Tutz](https://www.semanticscholar.org/author/G.-Tutz/67049458). “Variable selection for generalized linear mixed models by L1-penalized estimation” Statistics and Computing,2014

“Fatality Facts 2018: State by State.” *IIHS*, Dec. 2019, www.iihs.org/topics/fatality-statistics/detail/state-by-state

“Guide to Calculating Costs - Data Details.” *Injury Facts*, 27 May 2020, injuryfacts.nsc.org/all-injuries/costs/guide-to-calculating-costs/data-details/.

Lowy, Joan. “Traffic Accidents in the U.S. Cost $871 Billion a Year, Federal Study Finds.” *PBS*, Public Broadcasting Service, 29 May 2014, www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds.

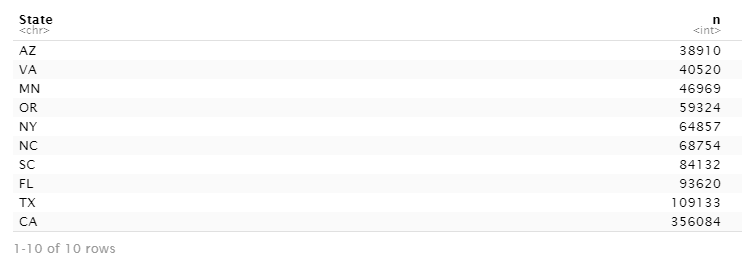
“State-Specific Costs of Motor Vehicle Crash Deaths.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 5 Nov. 2020, www.cdc.gov/transportationsafety/statecosts/index.html.

Wagner, I. “U.S. Property Damage Traffic Crashes.” *Statista*, 10 Aug. 2020, www.statista.com/statistics/192103/vehicles-involved-in-us-traffic-crashes-with-property-damages/.

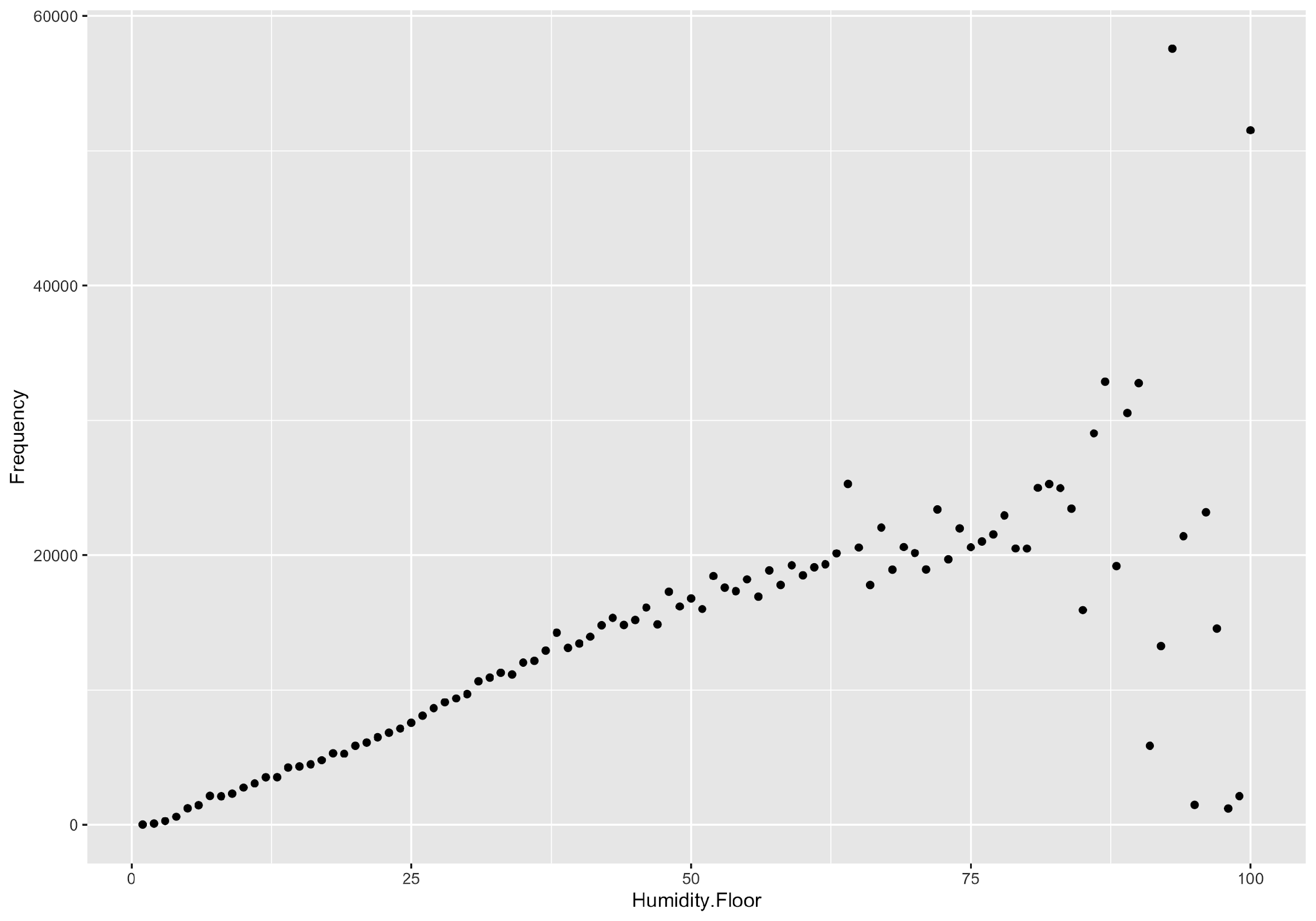
Wagner, I. “U.S. Traffic-Related Injuries and Fatalities.” *Statista*, 21 July 2020, www.statista.com/statistics/191900/road-traffic-related-injuries-and-fatalities-in-the-us-since-1988/.

# Appendix A: Supplementary Graphs

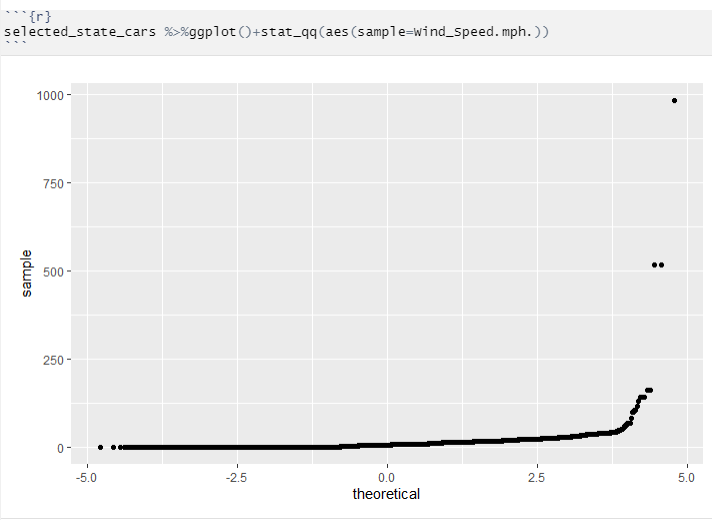
**Figure 1.** SELECTED TOP-10 STATES



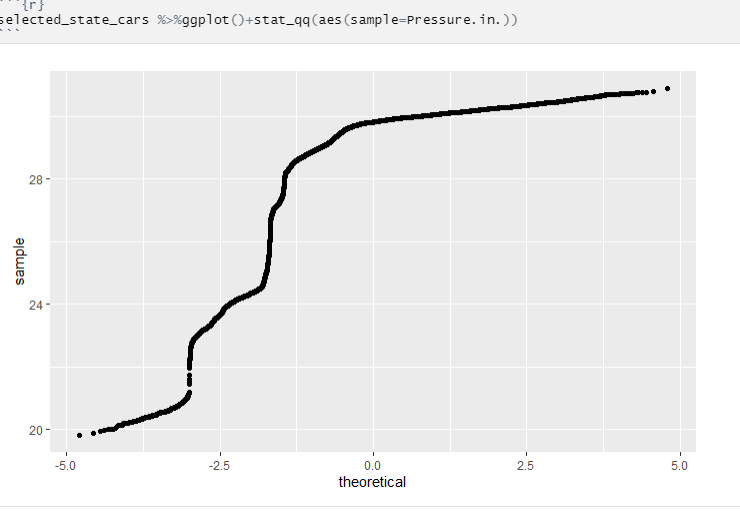
**Figure 2.** HUMIDITY PLOT AGAINST NORMALITY



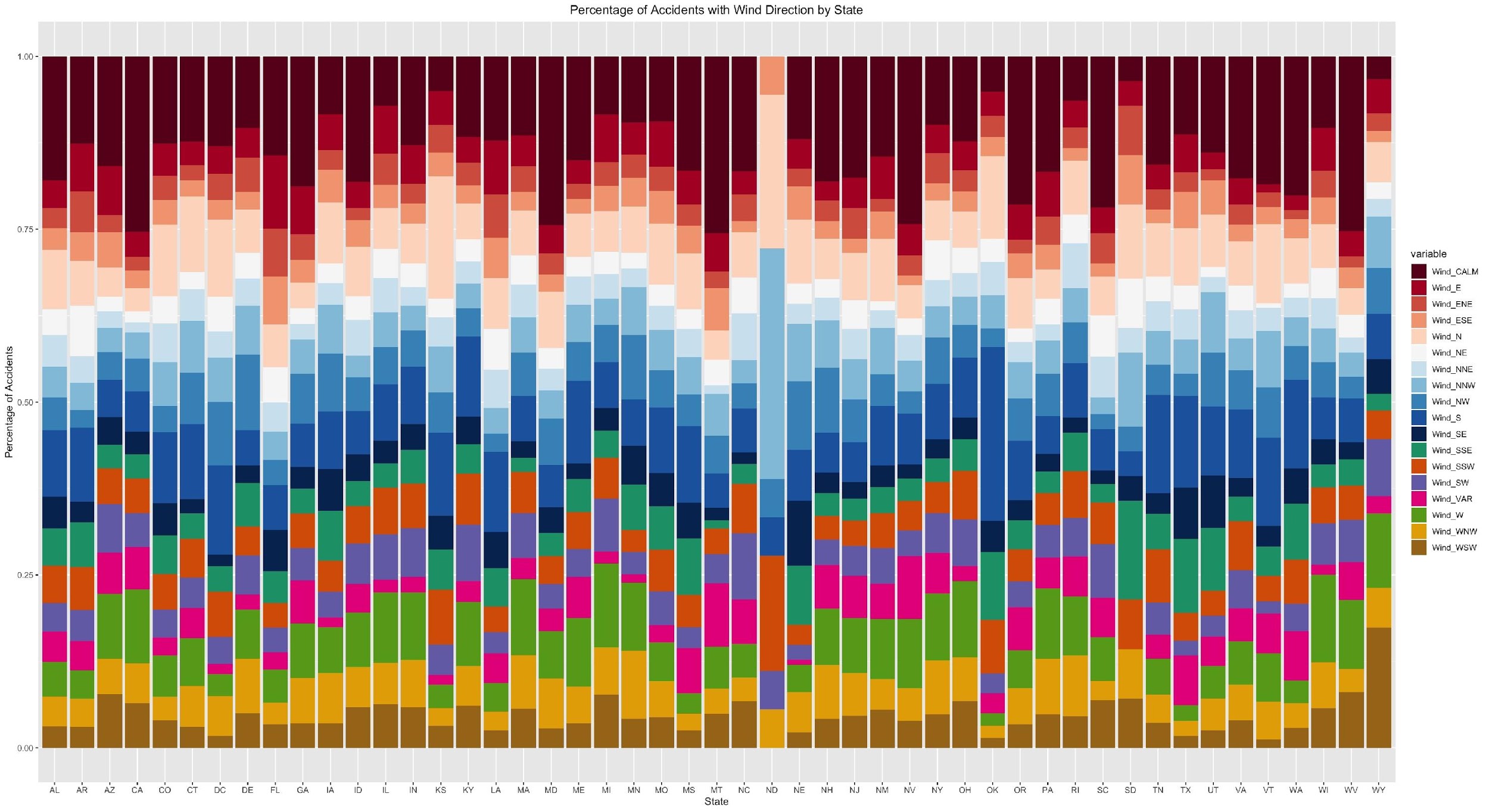
**Figure 3.** WIND SPEED PLOT AGAINST NORMALITY

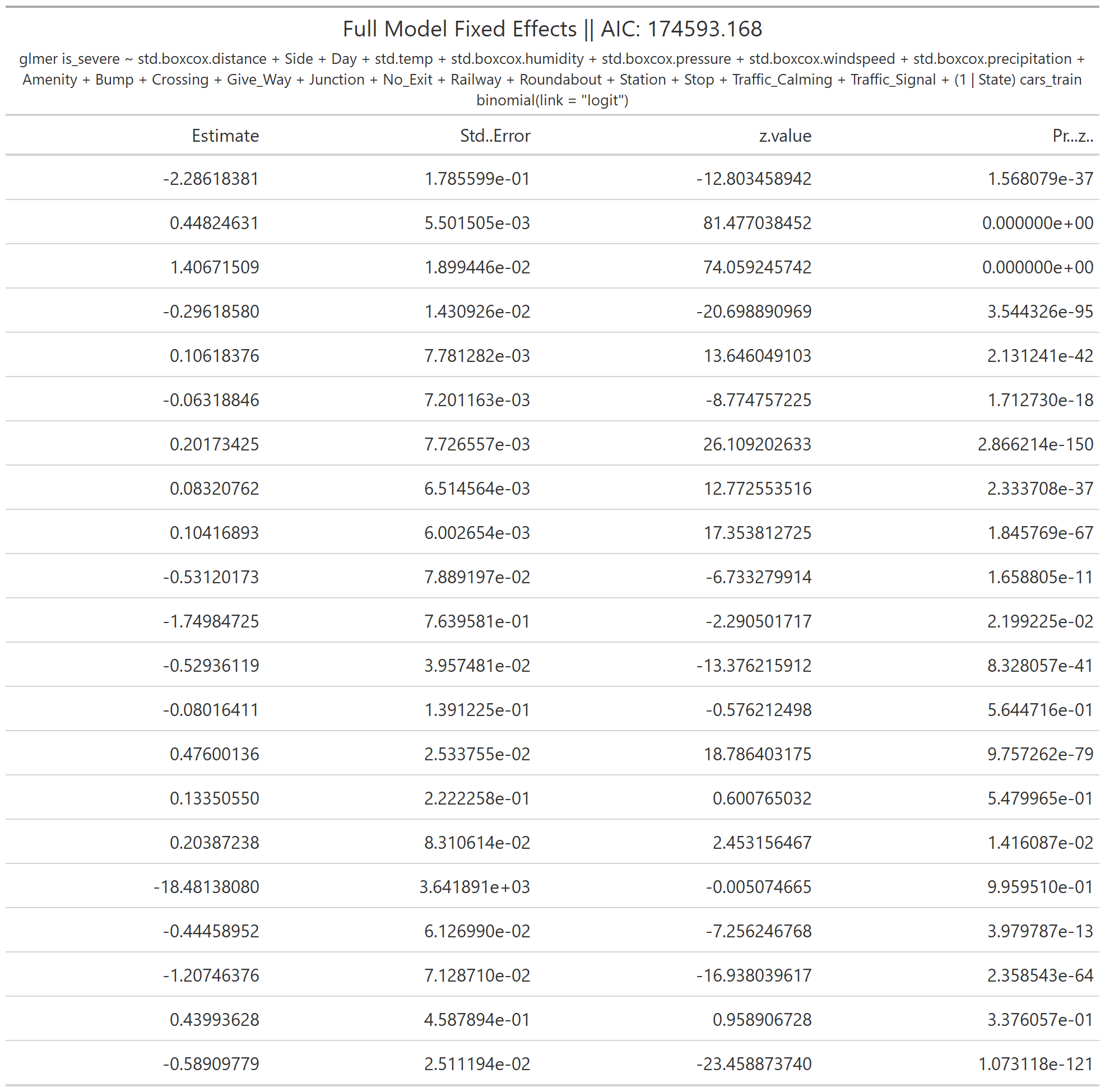


**Figure 4.** PRESSURE PLOT AGAINST NORMALITY

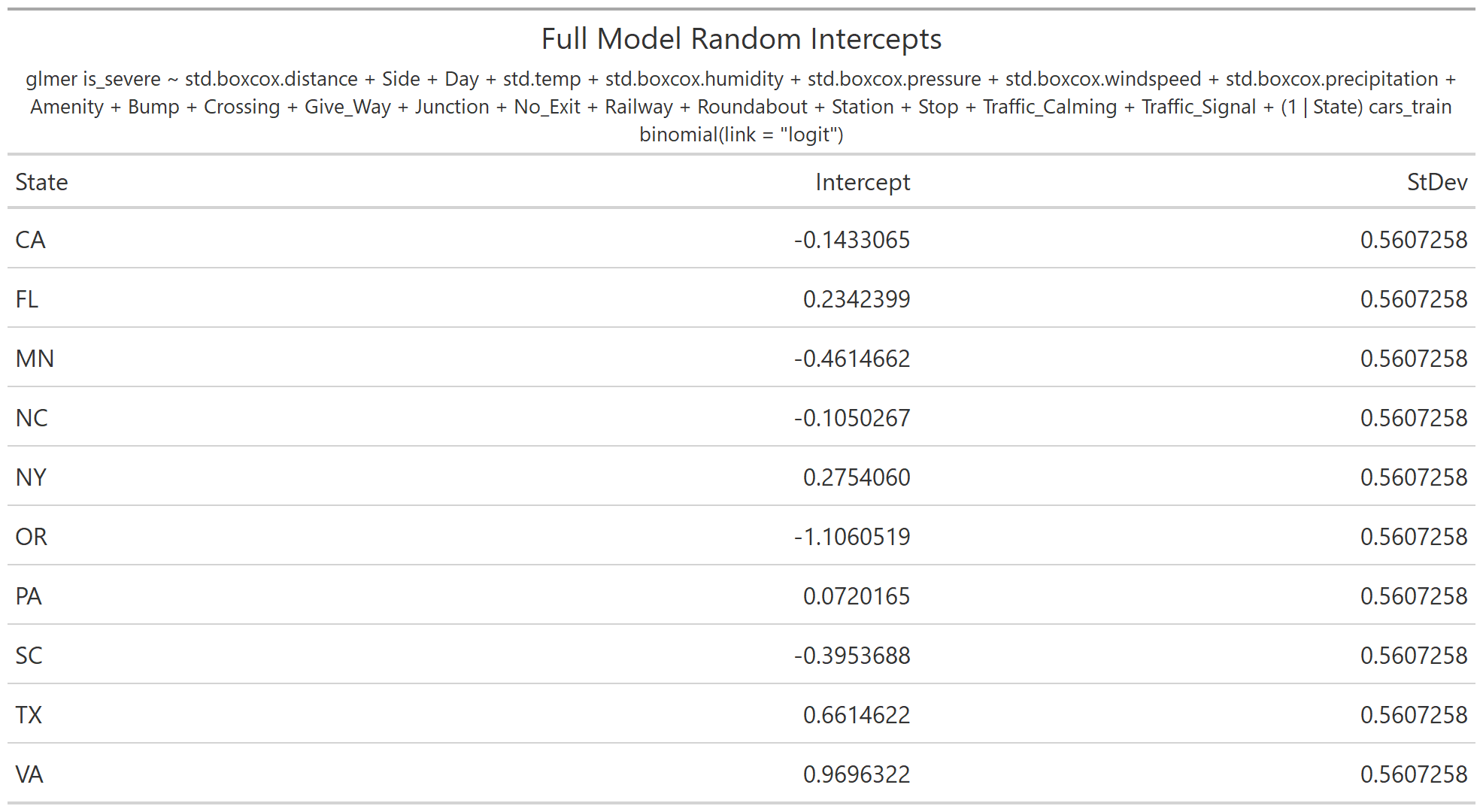


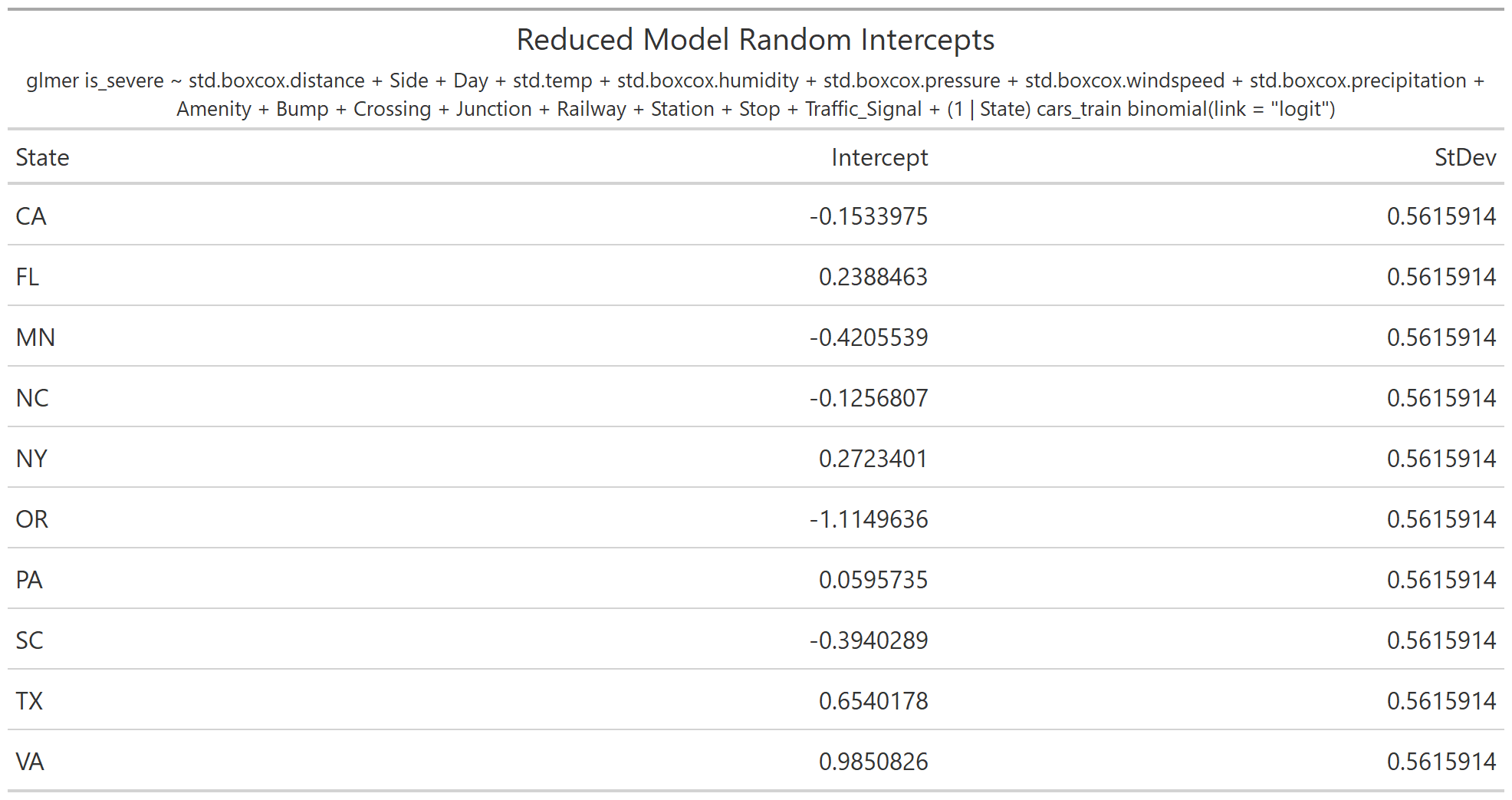
**Figure 5.** PERCENTAGE OF ACCIDENTS WITH WIND DIRECTIONS



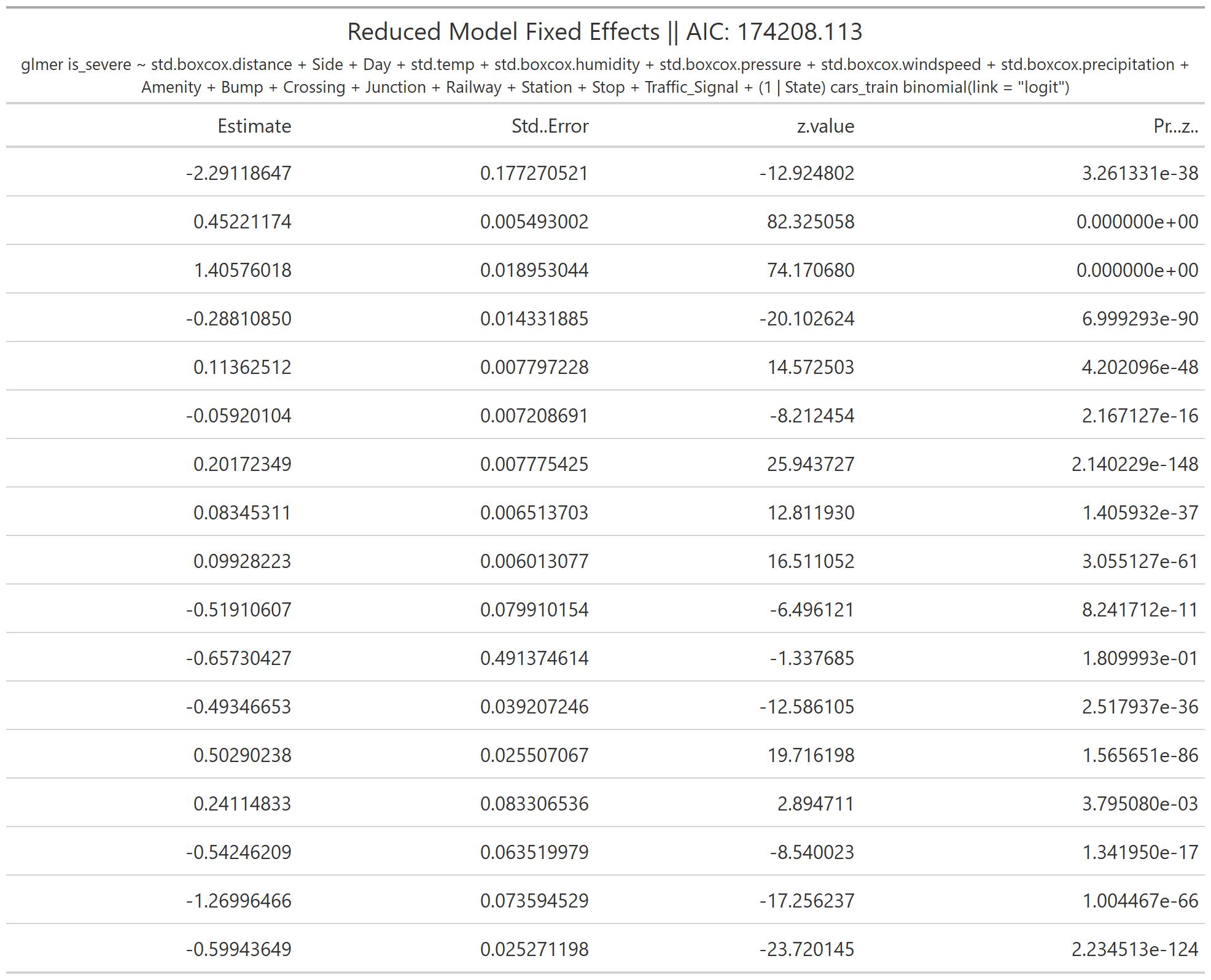
**Figure 6:**  MODEL RESULT WITH ALL POSSIBLE PREDICTORS

**Figure 7:** RANDOM INTERCEPT WITH ALL POSSIBLE PREDICTORS



**Figure 8:** RANDOM INTERCEPT WITH THREE PREDICTOR DELETED

**Figure 9:**  MODEL RESULT WITH SEVERAL NONSIGNIFICANT PREDICTOR DELETED



# Appendix B: R Code

library(tidyverse)

library(lme4)

library(pROC)

cars\_train <- read.csv('carstrain\_2.csv')

cars\_test <- read.csv('carstest\_2.csv')

m1 <- glmer(is\_severe~std.boxcox.distance+Side+Day+std.temp+std.boxcox.humidity+std.boxcox.pressure+std.boxcox.windspeed+std.boxcox.precipitation+Amenity+Bump+Crossing+Give\_Way+Junction+No\_Exit+Railway+Roundabout+Station+Stop+Traffic\_Calming+Traffic\_Signal+(1|State), data = cars\_train, family = binomial(link = "logit"))

m2 <- glmer(is\_severe~std.boxcox.distance+Side+Day+std.temp+std.boxcox.humidity+std.boxcox.pressure+std.boxcox.windspeed+std.boxcox.precipitation+Amenity+Bump+Crossing+Junction+Railway+Station+Stop+Traffic\_Signal+(1|State), data = cars\_train, family = binomial(link = "logit"))

anova(m1, m2)

m3 <- glmer(is\_severe~std.boxcox.distance+Side+Day+std.temp+std.boxcox.humidity+std.boxcox.pressure+std.boxcox.windspeed+std.boxcox.precipitation+Amenity+Crossing+Junction+Railway+Station+Stop+Traffic\_Signal+(1|State), data = cars\_train, family = binomial(link = "logit"))

anova(m2, m3)

#Prediction accuracy of final model

sum(cars\_test$is\_severe== ifelse(predict(m3, cars\_test, type='response')>=0.5, 1, 0))/nrow(cars\_test)

m3\_norandom <- glm(is\_severe~std.boxcox.distance+Side+Day+std.temp+std.boxcox.humidity+std.boxcox.pressure+std.boxcox.windspeed+std.boxcox.precipitation+Amenity+Crossing+Junction+Railway+Station+Stop+Traffic\_Signal+State, data = cars\_train, family = binomial(link = "logit"))

anova(m3, m3\_norandom)

#Prediction accuracy of logistic regression

sum(cars\_test$is\_severe== ifelse(predict(m3\_norandom, cars\_test, type='response')>=0.5, 1, 0))/nrow(cars\_test)

library(gt)

summary\_m1 <- summary(m1)

coef\_m1 <- data.frame(State = rownames(ranef(m1)$State), Intercept=unlist(ranef(m1)[['State']]), StDev = attributes(VarCorr(m1)$State)$stddev)

summary\_m1$coefficients %>% data.frame() %>% gt()%>%

tab\_header(

title = glue::glue("Full Model Fixed Effects || AIC: {round(summary\_m1$AICtab[1], 3)}"),

subtitle = glue::glue("{summary\_m1$call}")

) %>% gtsave('m1\_fix.png')

coef\_m1 %>% gt() %>%

tab\_header(

title = "Full Model Random Intercepts",

subtitle = glue::glue("{summary\_m1$call}")

) %>% gtsave('m1\_random.png')

summary\_m2 <- summary(m2)

coef\_m2 <- data.frame(State = rownames(ranef(m2)$State), Intercept=unlist(ranef(m2)[['State']]), StDev = attributes(VarCorr(m2)$State)$stddev)

summary\_m2$coefficients %>% data.frame() %>% gt()%>%

tab\_header(

title = glue::glue("Reduced Model Fixed Effects || AIC: {round(summary\_m2$AICtab[1], 3)}"),

subtitle = glue::glue("{summary\_m2$call}")

) %>% gtsave('m2\_fix.png')

coef\_m2 %>% gt() %>%

tab\_header(

title = "Reduced Model Random Intercepts",

subtitle = glue::glue("{summary\_m2$call}")

) %>% gtsave('m2\_random.png')

summary\_m3 <- summary(m3)

coef\_m3 <- data.frame(State = rownames(ranef(m3)$State), Intercept=unlist(ranef(m3)[['State']]), StDev = attributes(VarCorr(m3)$State)$stddev)

summary\_m3$coefficients %>% data.frame() %>% gt()%>%

tab\_header(

title = glue::glue("Final Model Fixed Effects || AIC: {round(summary\_m3$AICtab[1], 3)}"),

subtitle = glue::glue("{summary\_m3$call}")

) %>% gtsave('m3\_fix.png')

coef\_m3 %>% gt() %>%

tab\_header(

title = "Final Model Random Intercepts",

subtitle = glue::glue("{summary\_m3$call}")

) %>% gtsave('m3\_random.png')

## ROC Curves

test\_prob\_m1 <- predict(m1, newdata =cars\_test, type='response')

test\_roc\_m1 = roc(cars\_test$is\_severe ~ test\_prob\_m1)

test\_prob\_m2 <- predict(m2, newdata =cars\_test, type='response')

test\_roc\_m2 = roc(cars\_test$is\_severe ~ test\_prob\_m2)

test\_prob\_m3 <- predict(m3, newdata =cars\_test, type='response')

test\_roc\_m3 = roc(cars\_test$is\_severe ~ test\_prob\_m3)

test\_prob\_m3\_norandom <- predict(m3\_norandom, newdata =cars\_test, type='response')

test\_roc\_m3\_norandom = roc(cars\_test$is\_severe ~ test\_prob\_m3\_norandom)

plot(test\_roc\_m1, print.auc=FALSE, col="black", lty=1, lwd=2, legacy.axes = TRUE, print.auc.y=.35, grid=TRUE)

plot(test\_roc\_m2, print.auc=FALSE, col="black", lty=3, lwd=2, legacy.axes = TRUE, print.auc.y=.3, grid=TRUE, add=TRUE)

plot(test\_roc\_m3, print.auc=FALSE, col="black", lty=5, lwd=2, legacy.axes = TRUE, print.auc.y=.3, grid=TRUE, add=TRUE)

legend("bottomright",

legend=c("Model 1 (Full)","Model 2", "Model 3 (Final)"),

col=c("black", "black", "black"),

lty=c(1,3,5),

lwd=c(2,2,2))

text(0.4, 0.4, paste("AUC for m1:", round(test\_roc\_m1$auc, 3)))

text(0.4, 0.30, paste("AUC for m2:", round(test\_roc\_m2$auc, 3)))

text(0.4, 0.20, paste("AUC for m3:", round(test\_roc\_m3$auc, 3)))

## Sensitivity and Specificity

conf.matrix <- table(cars\_test$is\_severe,ifelse(test\_prob\_m3>=0.5, 1, 0))

conf.matrix

### Sensitivity

conf.matrix[2,2]/sum(conf.matrix[,2])

### Specificity

conf.matrix[1,1]/sum(conf.matrix[,1])

## Pearson Residual Plot

plot(m3)

## Random Intercepts

m3\_df <- data.frame(State=rownames(coef(m3)$State), Intercept=coef(m3)$State[,1], sd=as.numeric(attributes(VarCorr(m3)$State)$stddev))

pd <- position\_dodge(0.78)

ggplot(m3\_df, aes(x=State, y = Intercept, group = State)) +

#draws the means

geom\_point(position=pd) +

#draws the CI error bars

geom\_errorbar(aes(ymin=Intercept-2\*sd, ymax=Intercept+2\*sd,

color=State), width=.1, position=pd)

# Appendix C: Python Code

import numpy as np

import pandas as pd

traind=pd.read\_csv(r'carstrain\_2.csv')

train=traind.drop(columns='Unnamed: 0')

testd=pd.read\_csv(r'carstest\_2.csv')

test=testd.drop(columns='Unnamed: 0')

train=train.dropna(axis=0)

train.index=range(train.shape[0])

from sklearn.preprocessing import OneHotEncoder

enc=OneHotEncoder(categories='auto').fit(train.loc[:,'State'].values.reshape(-1,1))

result=OneHotEncoder(categories='auto').fit\_transform(train.loc[:,'State'].values.reshape(-1,1))

result1=result.toarray()

newtrain=pd.concat([train,pd.DataFrame(result1)],axis=1)

newtrain.columns=['is\_severe', 'State', 'std.boxcox.distance', 'Side', 'Day',

'std.temp', 'std.boxcox.humidity', 'std.boxcox.pressure',

'std.boxcox.windspeed', 'std.boxcox.precipitation', 'Amenity',

'Bump', 'Crossing', 'Give\_Way', 'Junction', 'No\_Exit', 'Railway',

'Roundabout', 'Station', 'Stop', 'Traffic\_Calming','Traffic\_Signal', 'Turning\_Loop','x0\_AZ', 'x0\_CA', 'x0\_FL', 'x0\_MN', 'x0\_NC', 'x0\_NY', 'x0\_OR',

'x0\_SC', 'x0\_TX', 'x0\_VA']

resulttest=OneHotEncoder(categories='auto').fit\_transform(test.loc[:,'State'].values.reshape(-1,1))

resultt=resulttest.toarray()

newtest=pd.concat([test,pd.DataFrame(resultt)],axis=1)

newtest.columns=['is\_severe', 'State', 'std.boxcox.distance', 'Side', 'Day',

'std.temp', 'std.boxcox.humidity', 'std.boxcox.pressure',

'std.boxcox.windspeed', 'std.boxcox.precipitation', 'Amenity',

'Bump', 'Crossing', 'Give\_Way', 'Junction', 'No\_Exit', 'Railway',

'Roundabout', 'Station', 'Stop', 'Traffic\_Calming','Traffic\_Signal', 'Turning\_Loop','x0\_AZ', 'x0\_CA', 'x0\_FL', 'x0\_MN', 'x0\_NC', 'x0\_NY', 'x0\_OR',

'x0\_SC', 'x0\_TX', 'x0\_VA']

from sklearn.preprocessing import OrdinalEncoder

newtrain.iloc[:,0]=OrdinalEncoder().fit\_transform(newtrain.iloc[:,0].values.reshape(-1,1))

newtest.iloc[:,0]=OrdinalEncoder().fit\_transform(newtest.iloc[:,0].values.reshape(-1,1))

from sklearn.feature\_selection import SelectFromModel

from sklearn.ensemble import RandomForestClassifier as RFC

y1=newtrain.iloc[:,0].values

newtrain=newtrain.drop(columns='is\_severe')

y2=newtest.iloc[:,0].values

newtest=newtest.drop(columns='is\_severe')

newtrain.iloc[:,1]=OrdinalEncoder().fit\_transform(newtrain.iloc[:,1].values.reshape(-1,1))

newtest.iloc[:,1]=OrdinalEncoder().fit\_transform(newtest.iloc[:,1].values.reshape(-1,1))

newtrain=newtrain.drop(columns='State')

newtrain.Side=np.where(newtrain.Side == 'R', 1, 0)

newtest=newtest.drop(columns='State')

newtest.Side=np.where(newtest.Side == 'R', 1, 0)

RFC\_=RFC(n\_estimators=100,random\_state=0)

x\_embedded=SelectFromModel(RFC\_).fit\_transform(newtrain,y1)

tenbed=newtest.iloc[:,[0,1,3,4,5,6]]

from sklearn.model\_selection import cross\_val\_score

from sklearn.tree import DecisionTreeClassifier

import matplotlib.pyplot as plt

clf=DecisionTreeClassifier(random\_state=0)

clf=clf.fit(x\_embedded,y1)

RFC\_=RFC\_.fit(x\_embedded,y1)

score\_c=clf.score(tenbed,y2)

score\_r=RFC\_.score(tenbed,y2)

print("single tree:{}".format(score\_c),

"random forest:{}".format(score\_r))

clf=DecisionTreeClassifier(random\_state=0)

clf=clf.fit(newtrain,y1)

RFC\_=RFC\_.fit(newtrain,y1)

score\_c=clf.score(newtest,y2)

score\_r=RFC\_.score(newtest,y2)

print("single tree:{}".format(score\_c),

"random forest:{}".format(score\_r))

from sklearn.linear\_model import LogisticRegression as LR

from sklearn.metrics import accuracy\_score

lrl2=LR(penalty="l2",solver="liblinear",C=0.5,max\_iter=1000)

lrl2=lrl2.fit(newtrain,y1)

lrl2.coef\_

accuracy\_score(lrl2.predict(newtest),y2)

y\_score = RFC\_.predict\_proba(newtest)

y\_score\_lrl2 = lrl2.predict\_proba(newtest)

from sklearn.metrics import roc\_curve, auc

import matplotlib.pyplot as plt

#ROC

fpr, tpr, thresholds = roc\_curve(y2, y\_score[:, 1])

roc\_auc = auc(fpr, tpr)

def drawRoc(roc\_auc,fpr,tpr):

plt.subplots(figsize=(7, 5.5))

plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.3f)' % roc\_auc)

plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('ROC Curve')

plt.legend(loc="lower right")

plt.show()

drawRoc(roc\_auc, fpr, tpr)

from sklearn.metrics import confusion\_matrix

#Random Forest confusion matrix

conf\_matrix\_rf=confusion\_matrix(y2, RFC\_.predict(newtest))

conf\_matrix\_rf

#sensitivity

conf\_matrix\_rf[1,1]/np.sum(conf\_matrix\_rf[:,1])

#specificity

conf\_matrix\_rf[0,0]/np.sum(conf\_matrix\_rf[:,0])

#Logistic Regression confusion matrix

conf\_matrix\_lrl2=confusion\_matrix(y2, lrl2.predict(newtest))

conf\_matrix\_lrl2

#sensitivity

conf\_matrix\_lrl2[1,1]/np.sum(conf\_matrix\_lrl2[:,1])

#specificity

conf\_matrix\_lrl2[0,0]/np.sum(conf\_matrix\_lrl2[:,0])

1. A data scientist at Lyft who publishes the dataset for Research purposes. [↑](#footnote-ref-1)
2. [Traffic Message Channel (TMC)](https://wiki.openstreetmap.org/wiki/TMC/Event_Code_List) code which provides more detailed description of the event. [↑](#footnote-ref-2)
3. The Traffic Service provides real time traffic information related to markets, incidents, and flow. [↑](#footnote-ref-3)
4. The District of Columbia also counts as one separate state. [↑](#footnote-ref-4)
5. Twilight definition: https://en.wikipedia.org/wiki/Twilight [↑](#footnote-ref-5)