Did you know that California holds the record for the highest number of car accidents in the entire United States?

Have you ever wondered which severity level is most likely in the event of a car accident?

Well, statistically speaking, there is a high probability of it being classified as a Level 2 severity.

Contents

Introduction

US Accidents dataset is available on **Kaggle**. It contains information about **traffic accidents** that occurred in the United States from February 2016 to December 2020. The dataset is in CSV format and contains **47 columns** and **2845342 rows**. It is a relatively large dataset. The data includes a variety of features such as the location of the accident, the severity of the accident, the weather conditions at the time of the accident, and the number of people involved.

Basic Statistics

Raw Counts

Name	Value
Rows	2,845,342
Columns	47
Discrete columns	33
Continuous columns	14
All missing columns	0
Missing observations	3,193,068
Complete Rows	947,262
Total observations	133,731,074
Memory allocation	1.7 Gb

Data Overview

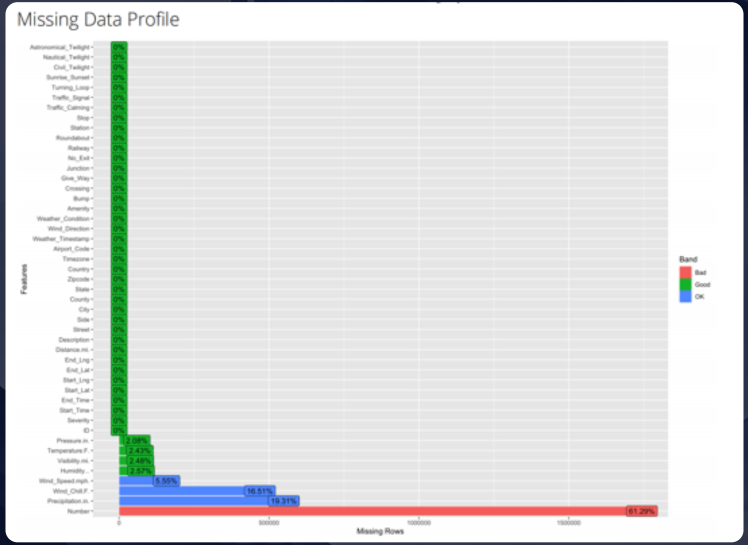
Variables	Description
ID	This is a unique identifier of the accident record.
Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay
Start_Time/End_Time	Shows start/ end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow
Start_Lat/Start_Lng/End_Lat/End_Lng	Shows longitude / latitude in GPS coordinate of the start point/ end point
Distance(mi)	The length of the road extent affected by the accident.
Description	Shows a human provided description of the accident.
Number	Shows the street number in address field.
Street	Shows the street name in address field.
Side	Shows the relative side of the street (Right/Left) in address field.
City	Shows the city in address field.
County	Shows the county in address field.
State	Shows the state in address field.
Zipcode	Shows the zipcode in address field.
Country	Shows the country in address field.
Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
Temperature(F)	Shows the temperature (in Fahrenheit).
Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
Humidity(%)	Shows the humidity (in percentage).
Pressure(in)	Shows the air pressure (in inches).
Visibility(mi)	Shows visibility (in miles).
Wind_Direction	Shows wind direction.
Wind_Speed(mph)	Shows precipitation amount in inches, if there is any.
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
Amenity	A POI annotation which indicates presence of amenity in a nearby location.
Amenity/Bump/Bump/Crossing/Give_Way/Junction/No_Exit/Railway/Roundabout /Station/Stop/Traffic_Calming/Traffic_Signal/Turning_Loop	A POI annotation which indicates presence of the sign in a nearby location.
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.

Data Cleaning:

- **Remove some columns**

We removed some columns that were irrelevant to our analysis. These columns were *"ID"*, *"description"*, *"Timezone"*, *"Airport_Code"*, *"Weather_Timestamp"*, and *"Wind_Direction"*

- **Clean Missing Values**



Before **After**

Data Cleaning:

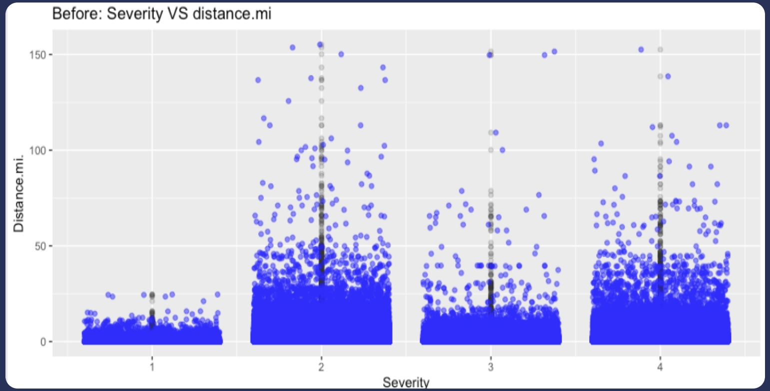
- **Checking for Duplication**

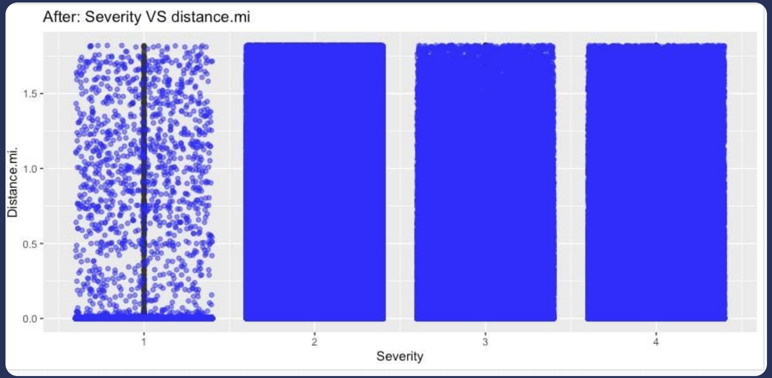
We check and remove any duplicated rows from the dataset.

- **Transfer bool_columns to 0,1**

Some columns were Boolean variables represented as strings "True" and "False". We converted these variables to binary variables.

Data Cleaning: Quantile and Outlier Removal





Before **After**

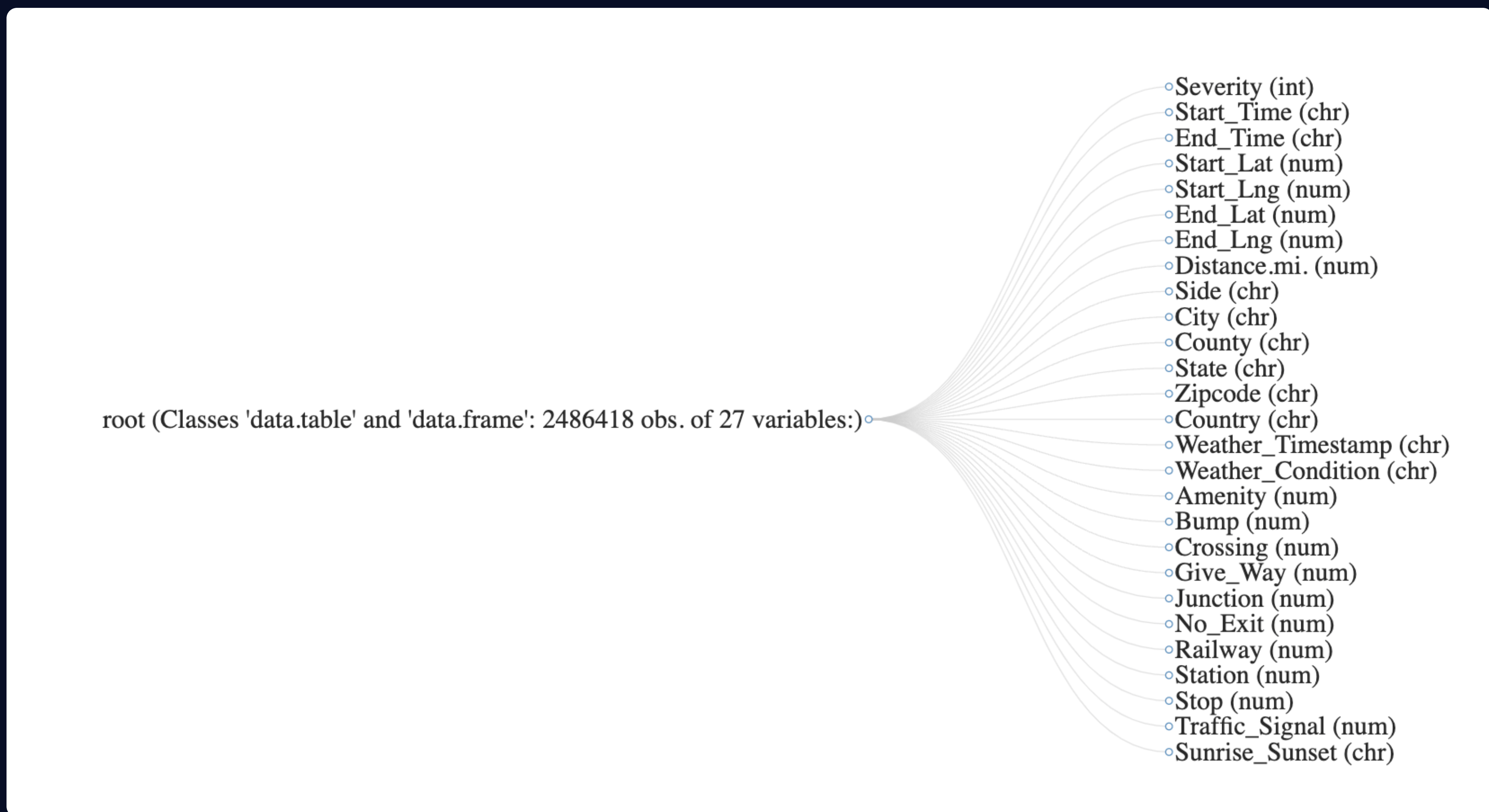
We calculated the IQR and used it to identify the upper and lower limits. Any values outside these limits were considered outliers and removed using the subset function.

Cleaned Data

Basic Statistics

Raw Counts

Name	Value
Rows	2,486,418
Columns	27
Discrete columns	11
Continuous columns	16
All missing columns	0
Missing observations	0
Complete Rows	2,486,418
Total observations	67,133,286
Memory allocation	852.7 Mb



These two charts depict the cleaned dataset which will be utilized for future analysis purposes.

Business Questions

1

***Location* and *place* Analysis.**

- Which U.S. states and cities have the highest car accident rates?

2

***Date* and *Time* Analysis.**

- When are car accidents more likely to occur in the United States?
- How has the number of accidents changed over time?

3

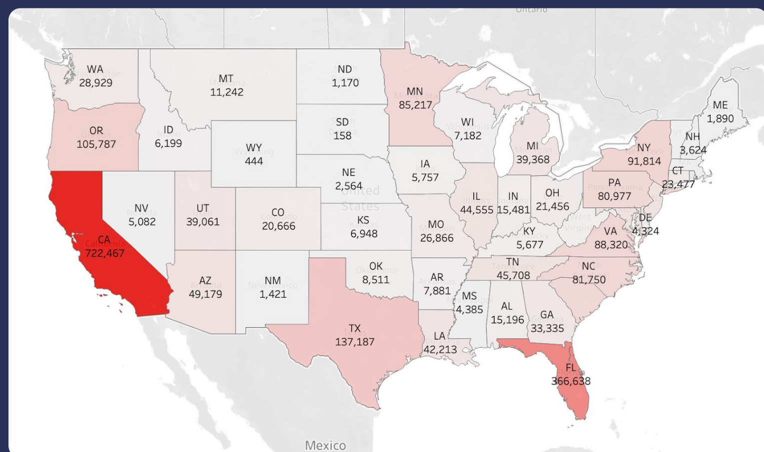
***External Factors* Analysis.**

- What are the factors that affect the severity of accidents?
- What is the most common type of accident in the United States?
- Do the POI annotations affect the accidents?

Location and place Analysis

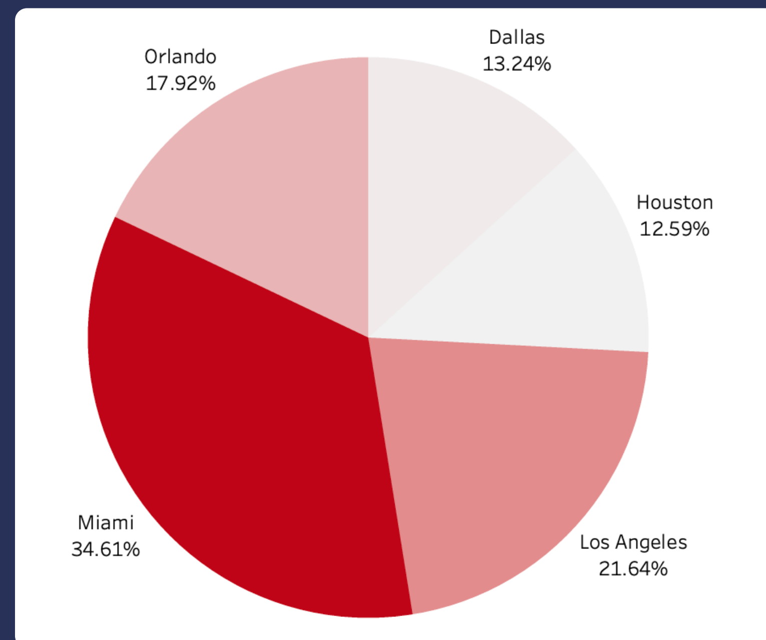
Accidents by State

California has the highest rate of car accidents, followed by Florida.



Accidents by City

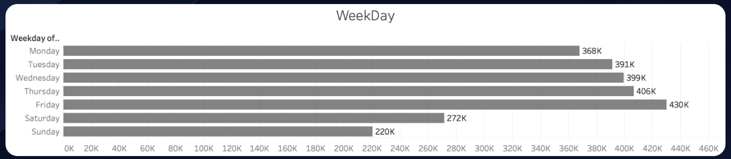
Miami, Los Angeles, Dallas, Houston, and Orlando are the top five cities with higher accident rates.



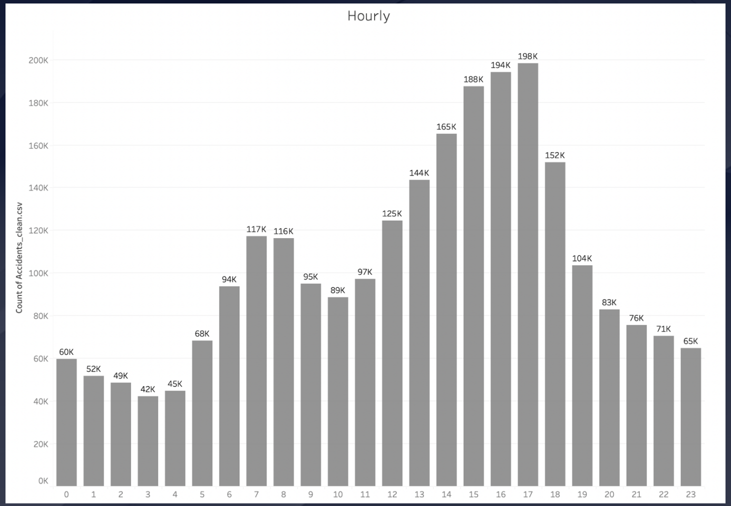
Date and Time Analysis

1 Accidents by Day and Time

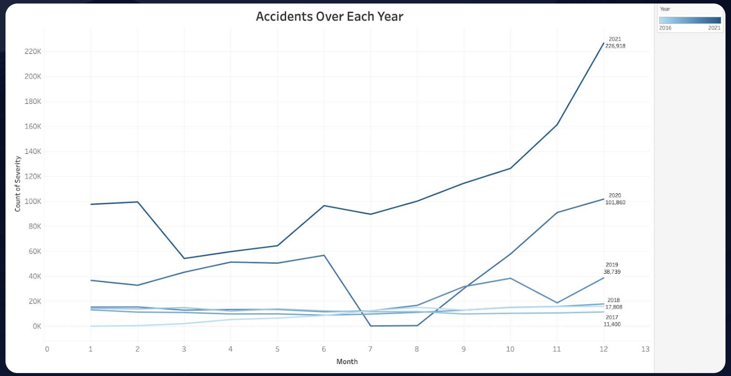
There are obvious peaks on weekdays, Monday to Friday, while weekends are significantly lower than weekdays.



Car accidents are most likely to occur during the morning and evening rush hour, from 6-9 am and 3-6 pm.



2 Accident by Year

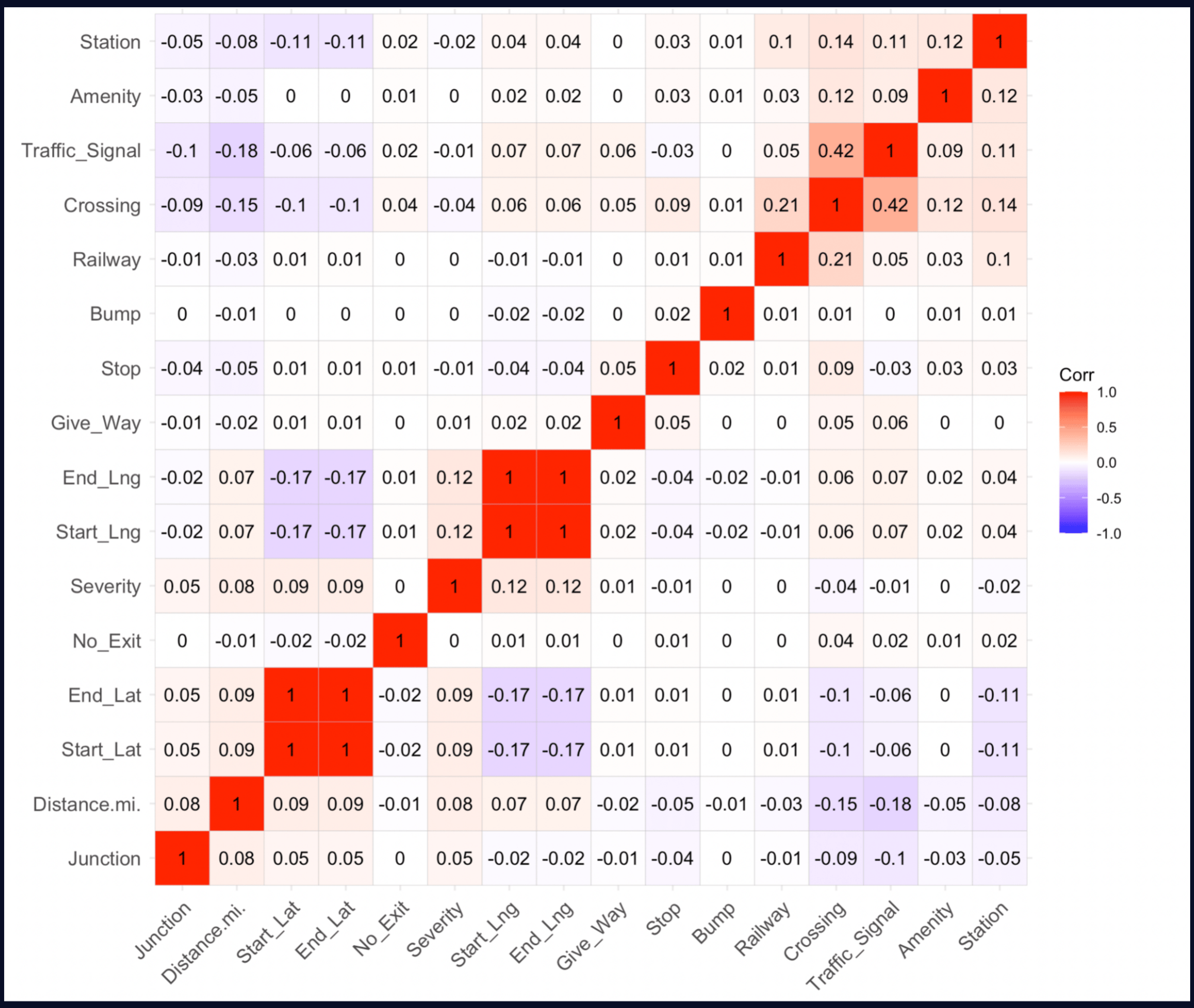


There has been a steady increase in the number of accidents since 2016, with a peak in 2020. The line graph also shows a particularly sharp increase in 2021.

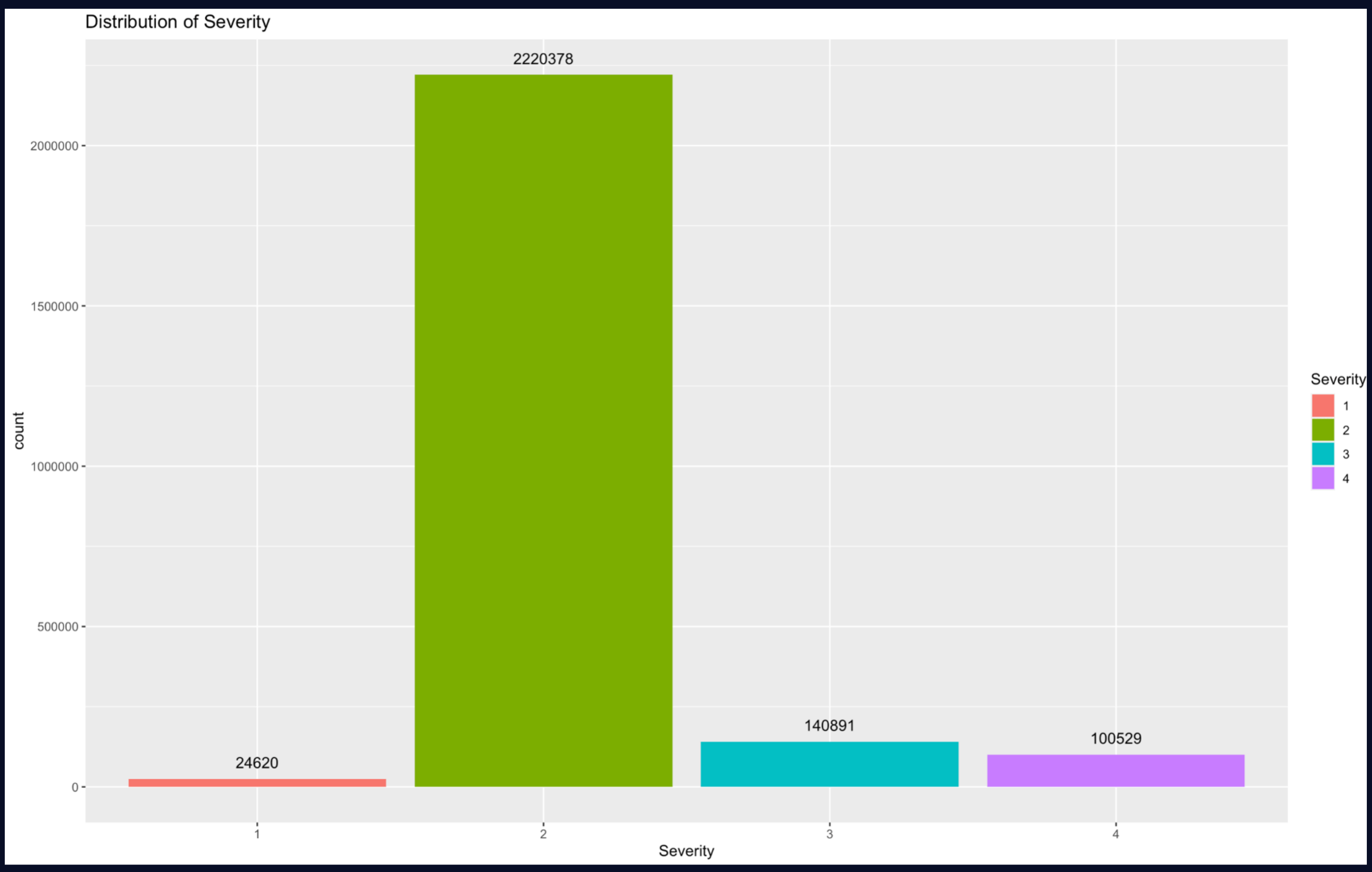
The number of accidents in 2021 has risen significantly compared to previous years, reaching a new peak of over 1.2 million accidents.

External Factors Analysis

What are the factors that affect the severity of accidents?

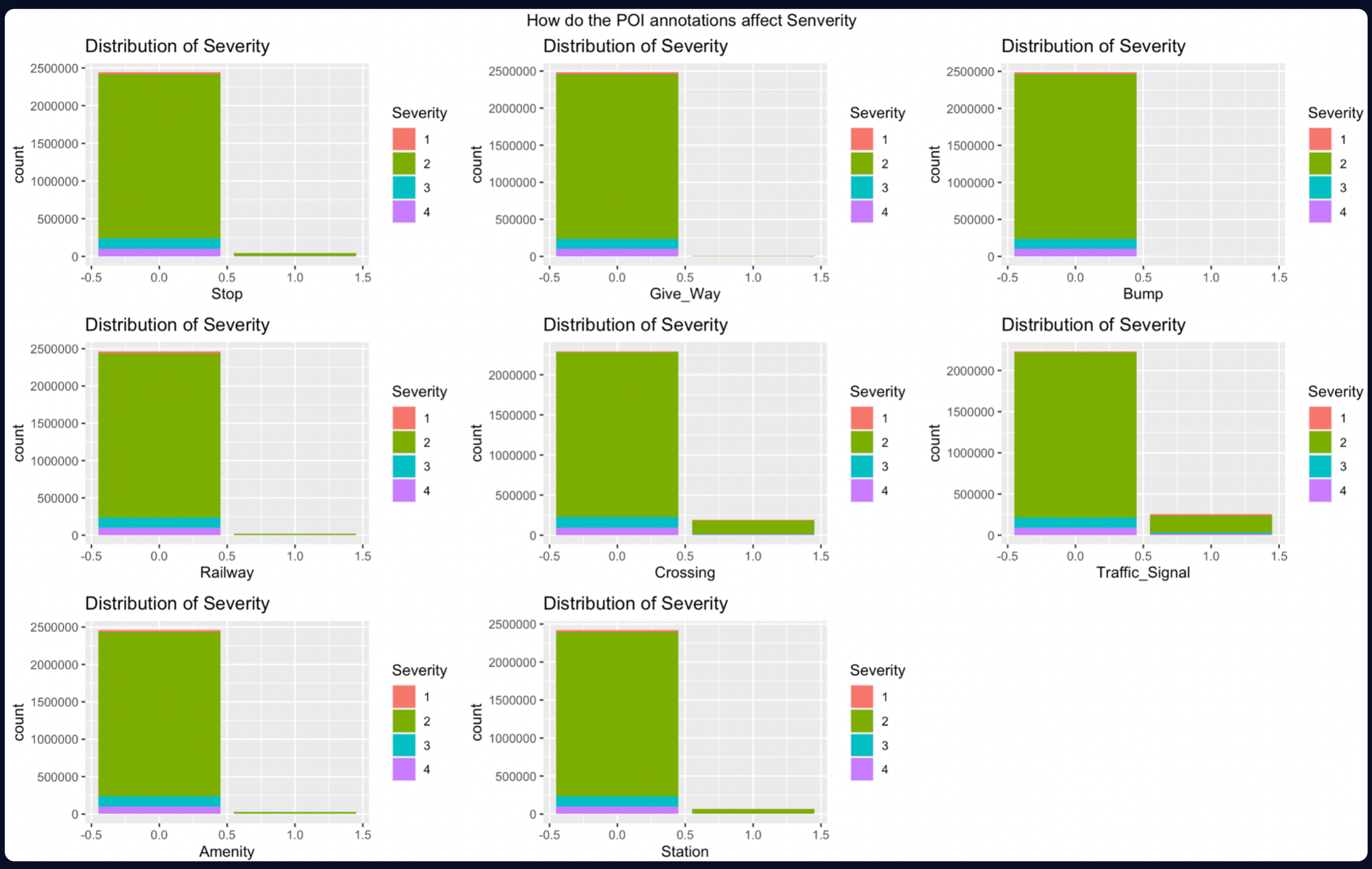


What is the most common type of accident in the United States?



The bar chart illustrates the distribution of Severity levels of accidents on a scale of 1 to 4, with 1 indicating the least impact on traffic. From the graph we can know the majority of accidents have severity level of 2. And least accidents have severity level of 1.

Do the POI annotations affect the accidents?



The bar chart describe the relationship between different types of POI annotations and the Severity of accidents. One notable finding from the chart is that a significant number of accidents occur due to the absence of POI annotations.

Logistic Model

```
> print(sorted_anova_results)
          Variable  Test       P_Value
1        Start_Lat ANOVA  0.000000e+00
2        Start_Lng ANOVA  0.000000e+00
3          End_Lat ANOVA  0.000000e+00
4          End_Lng ANOVA  0.000000e+00
5      Distance.mi. ANOVA  0.000000e+00
7    Temperature.F. ANOVA  0.000000e+00
8     Wind_Chill.F. ANOVA  0.000000e+00
9       Humidity... ANOVA 5.968430e-283
10     Pressure.in. ANOVA  6.577355e-63
11   Visibility.mi. ANOVA  1.516848e-32
6           Number ANOVA  1.770483e-05
12  Wind_Speed.mph. ANOVA  1.263008e-04
13 Precipitation.in. ANOVA  3.235678e-03
```

```
> print(sorted_chi_square_results)
                Variable       Test       P_Value
1             Start_Time Chi-Square  0.000000e+00
2               End_Time Chi-Square  0.000000e+00
3                 Street Chi-Square  0.000000e+00
5                   City Chi-Square  0.000000e+00
6                 County Chi-Square  0.000000e+00
7                  State Chi-Square  0.000000e+00
8                Zipcode Chi-Square  0.000000e+00
9      Weather_Condition Chi-Square  0.000000e+00
12              Crossing Chi-Square  0.000000e+00
14              Junction Chi-Square  0.000000e+00
21        Traffic_Signal Chi-Square  0.000000e+00
22        Sunrise_Sunset Chi-Square  0.000000e+00
23        Civil_Twilight Chi-Square  0.000000e+00
24     Nautical_Twilight Chi-Square  0.000000e+00
25 Astronomical_Twilight Chi-Square  0.000000e+00
15               No_Exit Chi-Square 1.752183e-105
16               Railway Chi-Square 2.737578e-101
18               Station Chi-Square  6.521627e-65
4                   Side Chi-Square  4.460940e-63
10               Amenity Chi-Square  3.131718e-51
13              Give_Way Chi-Square  2.191934e-07
19                  Stop Chi-Square  8.913192e-06
11                  Bump Chi-Square  1.835286e-05
20       Traffic_Calming Chi-Square  3.478238e-03
17            Roundabout Chi-Square  4.643541e-01
```

We split **80%** data as **train** data, and **20%** as **test** data. then, we use the **ANOVA** and **Chi-square** test to sorted the variables. The smallest p-values, indicating a higher level of statistical significance. We eliminated the variables with a p-value equal to 0 to form the logistic model.

```
> summary(logistic.m1)

Call:
glm(formula = Severity ~ Pressure.in. + Temperature.F. + Wind_Chill.F. +
    No_Exit + Junction + Railway + Sunrise_Sunset + Crossing +
    Amenity, family = "binomial", data = Accidents_clean_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6587   0.0580   0.0984   0.1144   0.9888

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.722253   2.154746  -3.120  0.00181 **
Pressure.in.         0.428471   0.073396   5.838 5.29e-09 ***
Temperature.F.      -0.003816   0.078980  -0.048  0.96146
Wind_Chill.F.       -0.006974   0.074373  -0.094  0.92529
No_Exit1            -0.969889   1.068415  -0.908  0.36399
Junction1           -2.516264   0.789020  -3.189  0.00143 **
Railway1            14.003424 410.677966   0.034  0.97280
Sunrise_SunsetNight  1.157331   0.357019   3.242  0.00119 **
Crossing1           -2.210056   0.232054  -9.524  < 2e-16 ***
Amenity1            -0.816544   0.420479  -1.942  0.05214 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 932.59  on 7999  degrees of freedom
Residual deviance: 776.55  on 7990  degrees of freedom
AIC: 796.55

Number of Fisher Scoring iterations: 16
```

Severity ~ Pressure.in. + Temperature.F. + Wind_Chill.F. + No_Exit + Junction + Railway + Sunrise_Sunset + Crossing + Amenity

```
> summary(logistic.m2)

Call:
glm(formula = Severity ~ Pressure.in. + Temperature.F. + Junction +
    Railway + Sunrise_Sunset + Crossing + Amenity, family = "binomial",
    data = Accidents_clean_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6597   0.0581   0.0986   0.1146   0.9897

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.61705    2.09371  -3.160  0.00158 **
Pressure.in.         0.42588    0.07315   5.822 5.82e-09 ***
Temperature.F.      -0.01120    0.00703  -1.593  0.11110
Junction1           -2.51109    0.78881  -3.183  0.00146 **
Railway1            14.01413  410.58961   0.034  0.97277
Sunrise_SunsetNight  1.15409    0.35656   3.237  0.00121 **
Crossing1           -2.21802    0.23163  -9.576  < 2e-16 ***
Amenity1            -0.82375    0.42003  -1.961  0.04986 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 932.59  on 7999  degrees of freedom
Residual deviance: 777.21  on 7992  degrees of freedom
AIC: 793.21

Number of Fisher Scoring iterations: 16
```

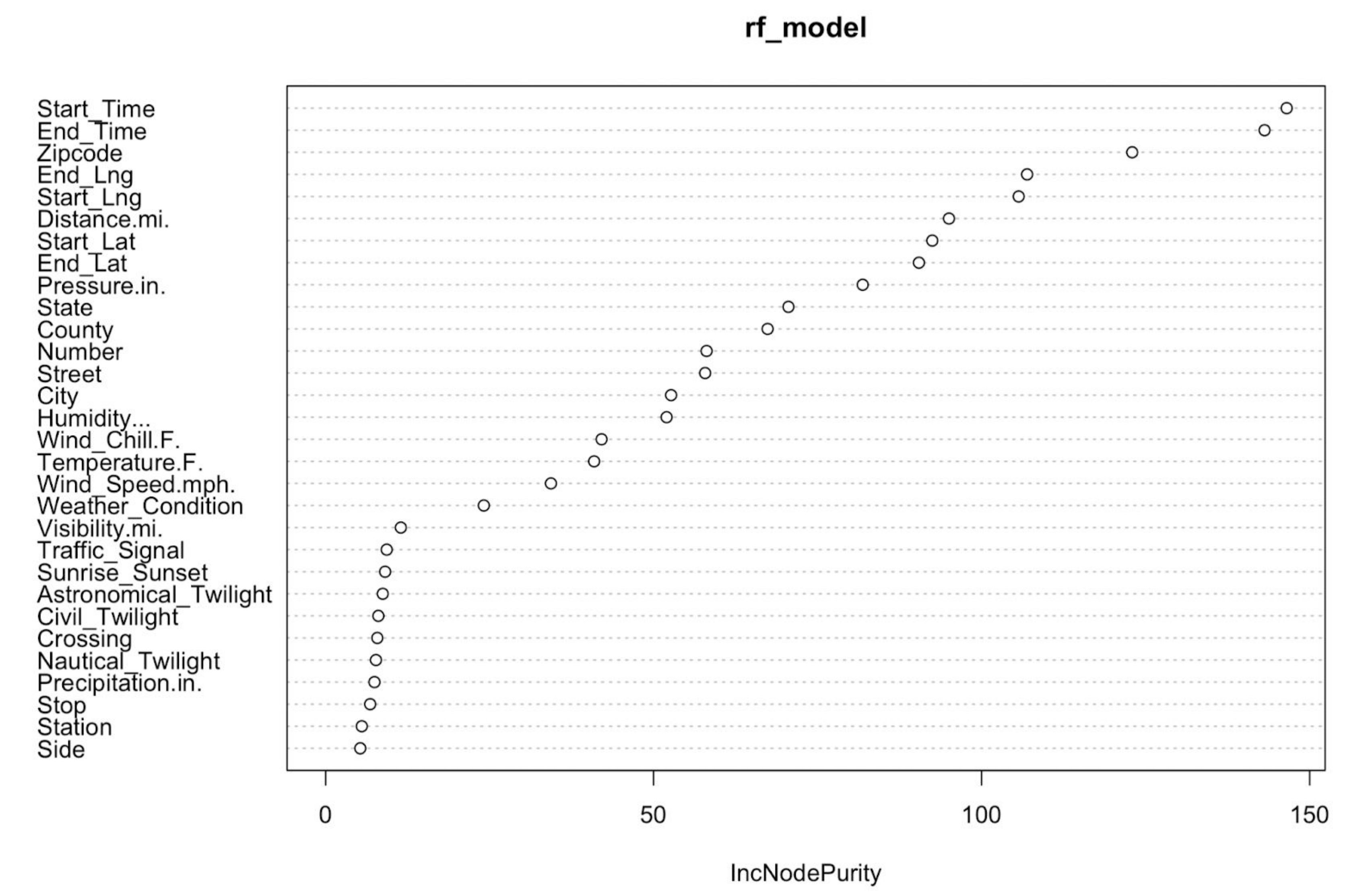
We use the stepwise method to adjust the model, and get the new model **Severity ~ Pressure.in. + Temperature.F. + Junction + Railway + Sunrise_Sunset + Crossing + Amenity**

The AIC value decrease from 796.55 to 793.21.

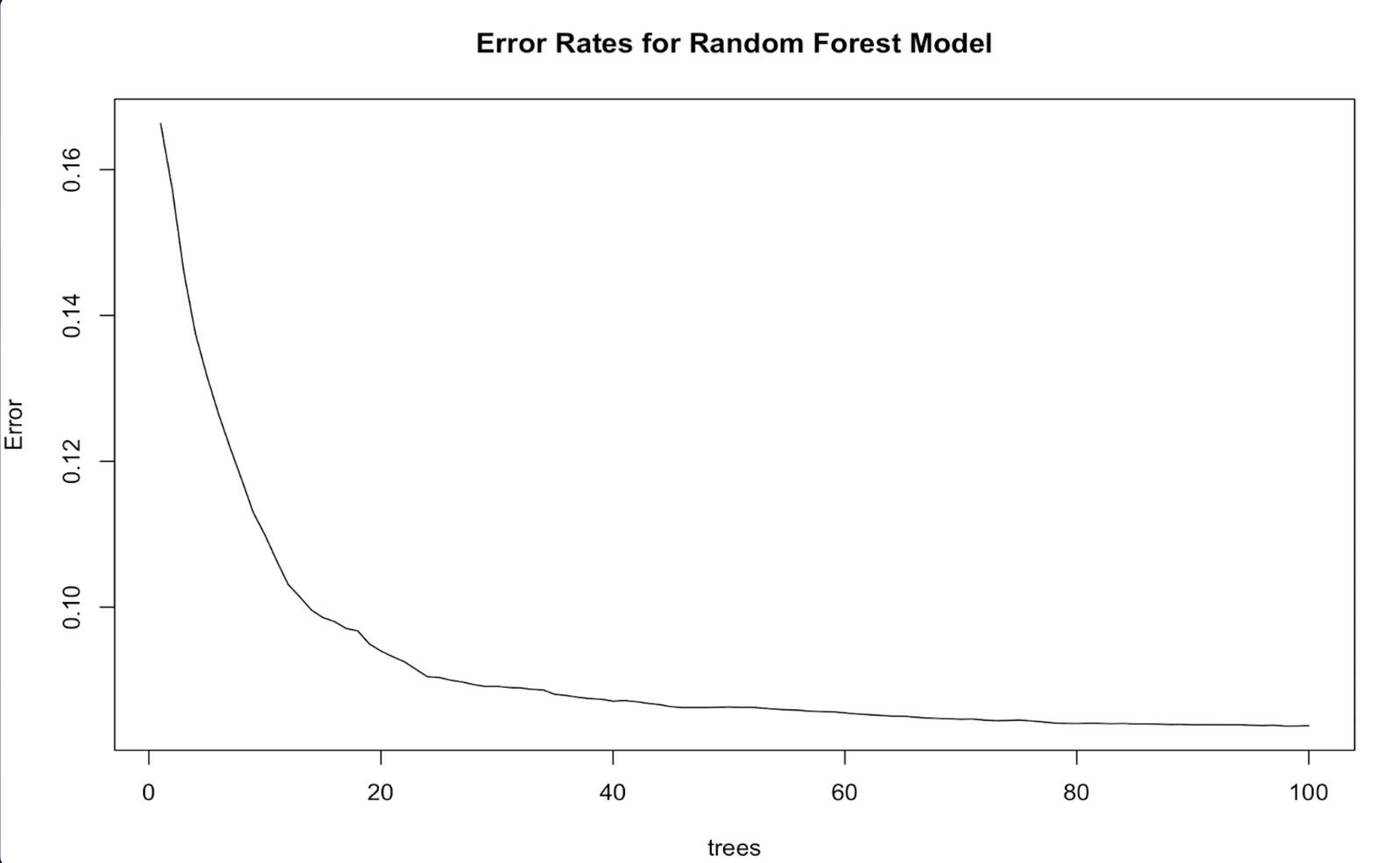
Random Forest Model

```
> print(rf_model)

Call:
 randomForest(formula = Severity ~ ., data = Accidents_clean_train_sampled,      ntree = 100)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 13

          Mean of squared residuals: 0.08373159
                    % Var explained: 26.73
```

The model was trained with the "**Severity**" column as the **target** variable and all other columns as predictors. We used 100 trees in our random forest model.



Plot of the variable importance

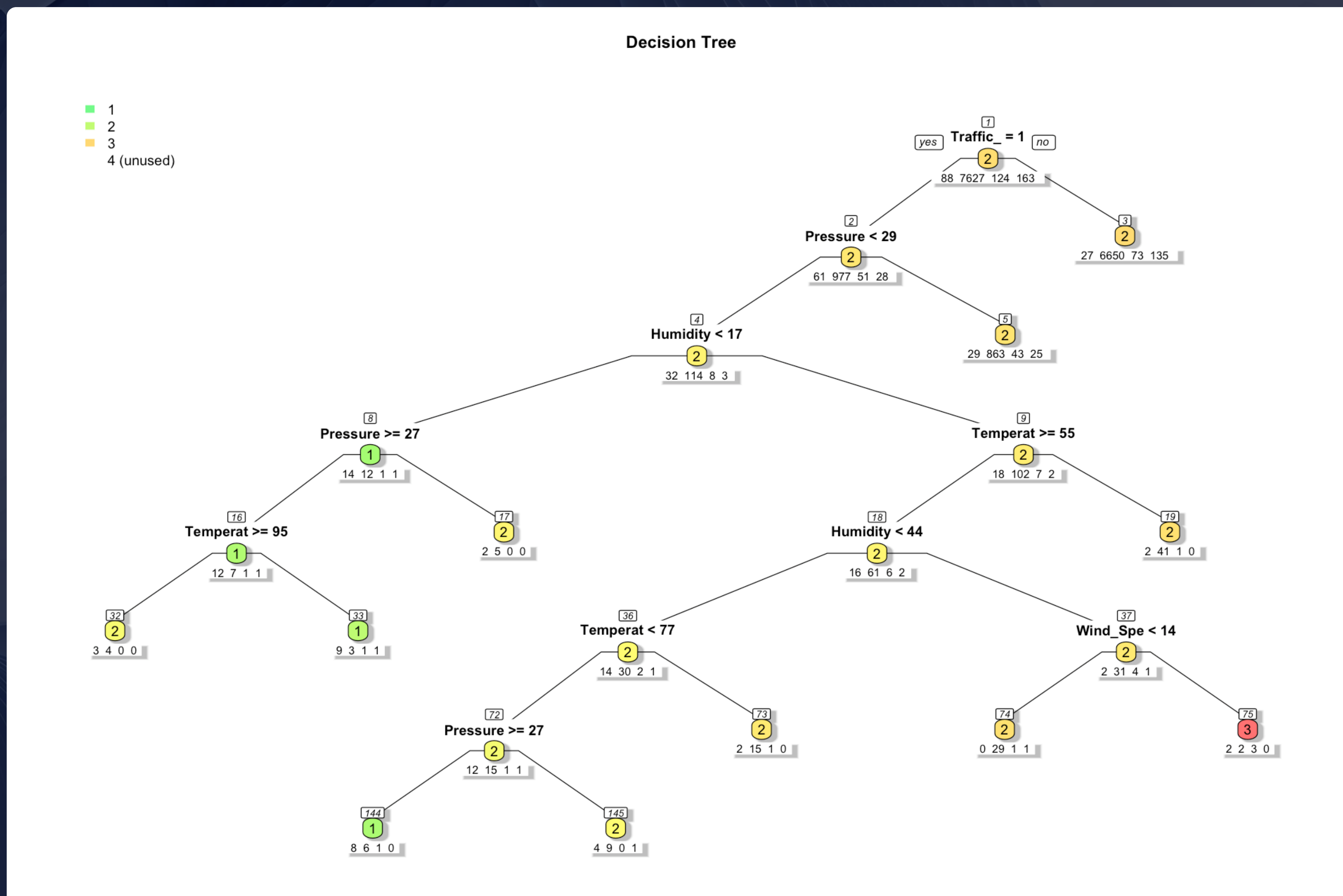


As the number of trees in the model increases, the error rate goes down, but the improvement becomes less significant after around **50** trees.

The error rate is **lowest** when the number of trees is around **80-90**. It suggests that adding more trees to the model beyond around **80-90** does not significantly improve performance.

Decision Tree

Decision trees are a popular machine learning algorithm used for classification and regression tasks. In this data set, the decision tree can be used to predict the severity of an accident based on the various characteristics available in the data set.



tree_model<- rpart(Severity ~ Temperature.F.+ Junction + Wind_Chill.F. + Wind_Speed.mph. + Visibility.mi. + Humidity... + Pressure.in. + Stop + Crossing + Traffic_Signal + Station + Amenity + Railway, data = Accidents_clean_train, method = "class", minsplit = 20, cp = 0.001)

Modeling & Performance Evaluation

Model	Accuracy Score	AUC	Sensitivity (Class 2)	Confusion Matrix
Logistic Regression	1.3%	0.61	0	[[26, 1905, 28, 41], [0, 0, 0, 0], [0, 0, 0, 0], [0, 0, 0, 0]]
Random Forest	88%	0.66	0.92	[[1, 1, 0, 0], [20, 1747, 22, 25], [1, 17, 7, 2], [0, 141, 1, 13]]
Decision Tree	95%	0.69	0.99	[[1, 47, 0, 0], [0, 1905, 0, 0], [0, 28, 0, 0], [0, 41, 0, 0]]

🏆 **Decision Tree Model**

Conclusion

Our project showcased the importance of data cleaning, EDA, and the application of various machine learning models in analyzing the US car accidents dataset. By combining these steps, we gained valuable insights into the factors influencing accident severity and identified potential strategies to mitigate their impact. Our findings contribute to the broader understanding of road safety and provide actionable recommendations for improving accident prevention measures.