# Improving Road Safety: A United States Road Traffic Accident Breakdown

## 1 Introduction

Road traffic accidents remain one of the leading causes of death and injury around the world with approximately 1.19 million yearly fatalities and 20 to 50 million injuries, including lifelong disabilities[13]. Various intricate factors cause traffic accidents, the most notable being passenger behaviour and environmental conditions. Components of road safety are difficult, if not impossible, to effectively control. The complexity of the subject has made road safety a high-priority issue for many nations.

Most countries handle traffic law at the national level. This is not exactly the case for the United States as it is divided into states. Each state is responsible for its vehicle code, which involves standardised federal, common traffic, and state-specific laws [7]. For example, New Hampshire does not require front-seat passengers to use seatbelts, whilst all other states have primary or secondary enforcement[5]. This adds a layer of complexity to the problem.

In 2022, the United States experienced 5.93 million police-reported traffic accidents, of which 42,514 resulted in fatalities and 2.38 million in injuries[2]. Apart from the high social toll, a large economic cost is involved. For most countries, 3% of their total GDP is spent on road accidents yearly[13]. In 2019, the United States spent a staggering $340 billion of which 66% was caused by passenger behaviour. Seatbelt use prevented $93 billion in injury-related costs, and it is still not mandatory in every state[1]. The government also spends its budget on maintaining road safety. In 2020, $1 billion was budgeted for the National Highway Traffic Safety Administration. In 2024, $1 billion in grants was provided by the Biden-Harris Administration to ensure community road safety[11][10].

A less-discussed topic is the disruption of traffic flow after an accident. A huge amount of unreported costs are associated with time spent in traffic. The accumulation in traffic affects commuters and first responders, including medical services and law enforcement. This is a serious issue that impacts both the social and economic sides of traffic accidents.

## 2 Objectives

By analysing the impacts of traffic accidents in the United States, this report aims to provide a better understanding of the factors involved. In addition, this study will provide recommendations for improving road safety.

- **Impacts:**
  - **Accident Context:** Analyse infrastructure through points of interest highlighting areas of risk. Investigate temporal trends, breaking down periods of higher risk. Explore the relationships between accidents and contributing factors.
  - **Accident Response:** Examine traffic descriptions identifying commonalities. Analyse traffic severity including relationships and trends.
- **Recommendations:**
  - **Infrastructure and Law:** States and the Federal Government can read the analysis to establish or retract laws and to improve city planning to ensure road safety.
  - **Driver Awareness:** Drivers can obtain information on road safety to better understand risk factors.

- **Accident Reporting:** Using key findings, a better traffic description system could be implemented to ensure that services and commuters receive better traffic descriptions.

# 3 Methodology

## 3.1 Data Sourcing

The original dataset from Kaggle, [Original][9][8] spans nearly 8,000,000 records with 46 different attributes covering infrastructure, weather, time-related information, and geographic information. For this study, a sampled version of that dataset was used [Sampled][9][8]. From this, a further random sample was taken providing 100,000 records with 46 attributes. This dataset is ideal to analyse due to the diverse array of information.

## 3.2 Schema Creation

The schema designed in Figure 1 was created to ensure scalability and deep analysis. To guarantee a diverse analysis, the schema features 5 tables with 52 attributes, including primary and foreign keys, text, real, integer, datetime, and boolean data types. Some values are allowed to be *NULL*. The schema is as follows:

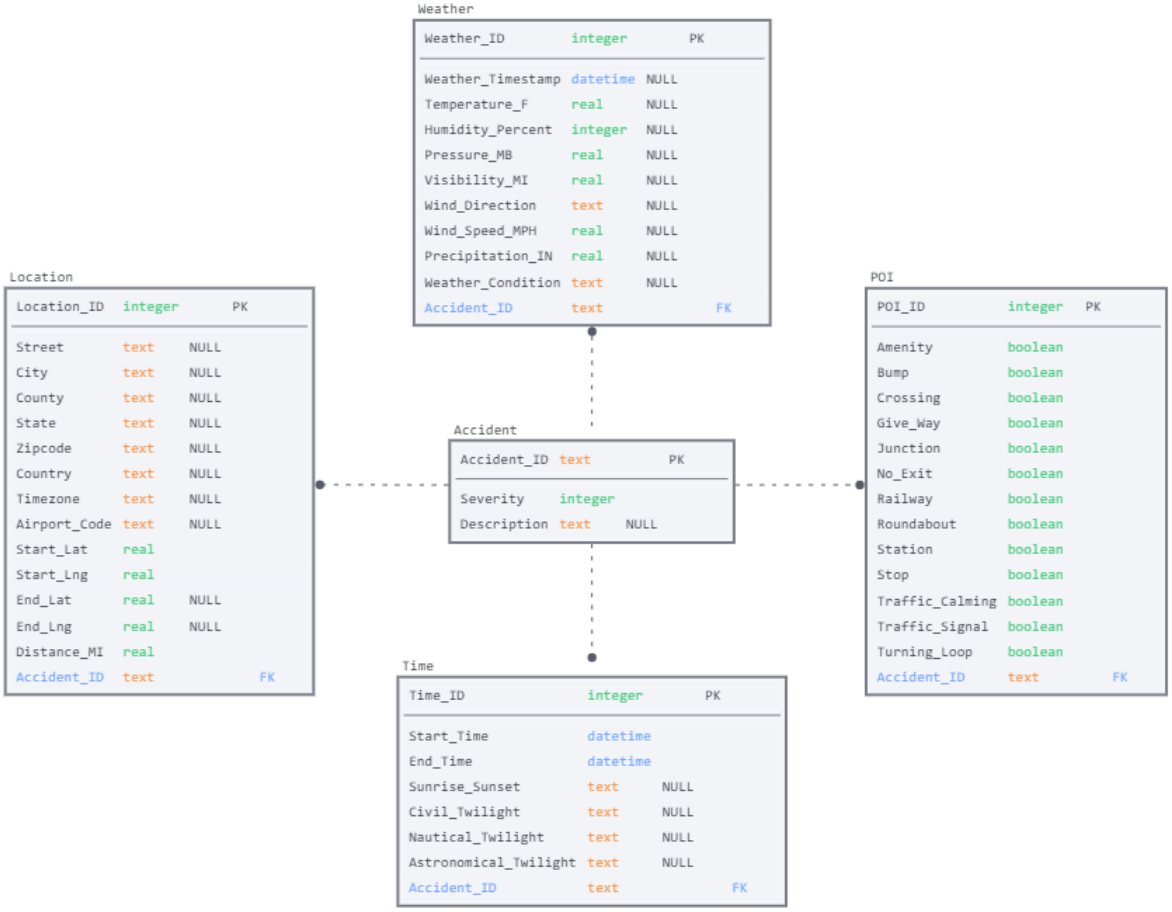
- **Accident table:** The central table containing information for each accident including: `Accident_ID` (primary key), `Severity` (integer), `Description` (text). Every other table includes `Accident_ID` (foreign key)

- **POI table:** A infrastructure table which lists notable points of interest in the area of the accident. Important features include: `POI_ID` (primary key), `Amenity` (boolean), `Crossing` (boolean), `Junction` (boolean), `Roundabout` (boolean), `Stop` (boolean), `Traffic_Calming` (boolean), `Traffic_Signal` (boolean).

- **Weather table:** A table containing environmental factors in the form of numeric data, text, and dates. Important features include: `Weather_ID` (primary key), `Temperature_F` (real), `Humidity_Percent` (integer), `Pressure_MB` (real), `Visibility_MI` (real), `Precipitation_IN` (real), `Weather_Condition` (text).

- **Time table:** A table consisting of timestamps and light conditions. It features: `Time_ID` (primary key), `Start_Time` (datetime), `End_Time` (datetime), `Sunrise_Sunset` (text), `Civil_Twilight` (text), `Nautical_Twilight` (text), `Astronomical_Twilight` (text).

- **Location table:** A table with geographical data. Important features include: `Location_ID` (primary key), `City` (text), `Start_Lat` (real), `Start_Lng` (real), `End_Lat` (real), `End_Lng` (real), `Distance_MI` (real). The country column has been included to ensure future scalability even though all current entries are ’US’.



Weather

| Weather_ID | integer | PK |
|---|---|---|
| Weather_Timestamp | datetime | NULL |
| Temperature_F | real | NULL |
| Humidity_Percent | integer | NULL |
| Pressure_MB | real | NULL |
| Visibility_MI | real | NULL |
| Wind_Direction | text | NULL |
| Wind_Speed_MPH | real | NULL |
| Precipitation_IN | real | NULL |
| Weather_Condition | text | NULL |
| Accident_ID | text | FK |

Location

| Location_ID | integer | PK |
|---|---|---|
| Street | text | NULL |
| City | text | NULL |
| County | text | NULL |
| State | text | NULL |
| Zipcode | text | NULL |
| Country | text | NULL |
| Timezone | text | NULL |
| Airport_Code | text | NULL |
| Start_Lat | real | |
| Start_Lng | real | |
| End_Lat | real | NULL |
| End_Lng | real | NULL |
| Distance_MI | real | |
| Accident_ID | text | FK |

Accident

| Accident_ID | text | PK |
|---|---|---|
| Severity | integer | |
| Description | text | NULL |

POI

| POI_ID | integer | PK |
|---|---|---|
| Amenity | boolean | |
| Bump | boolean | |
| Crossing | boolean | |
| Give_Way | boolean | |
| Junction | boolean | |
| No_Exit | boolean | |
| Railway | boolean | |
| Roundabout | boolean | |
| Station | boolean | |
| Stop | boolean | |
| Traffic_Calming | boolean | |
| Traffic_Signal | boolean | |
| Turning_Loop | boolean | |
| Accident_ID | text | FK |

Time

| Time_ID | integer | PK |
|---|---|---|
| Start_Time | datetime | |
| End_Time | datetime | |
| Sunrise_Sunset | text | NULL |
| Civil_Twilight | text | NULL |
| Nautical_Twilight | text | NULL |
| Astronomical_Twilight | text | NULL |
| Accident_ID | text | FK |

Figure 1: Relational database schema diagram

## 3.3 Data Preprocessing

Preprocessing involves cleaning the data to ensure it matches the designated database schema. Three key steps were performed to completely clean the data, as well as some minor changes. The column Source was removed because it did not provide useful information.

### 3.3.1 Identifying Missing Values

As seen in Figure 2, there are a total of 19 columns with missing values. Columns with missing values less than 10% will have rows removed while the other columns will be examined individually. Of the 19 columns, 15 were removed and 4 remained for individual examination.

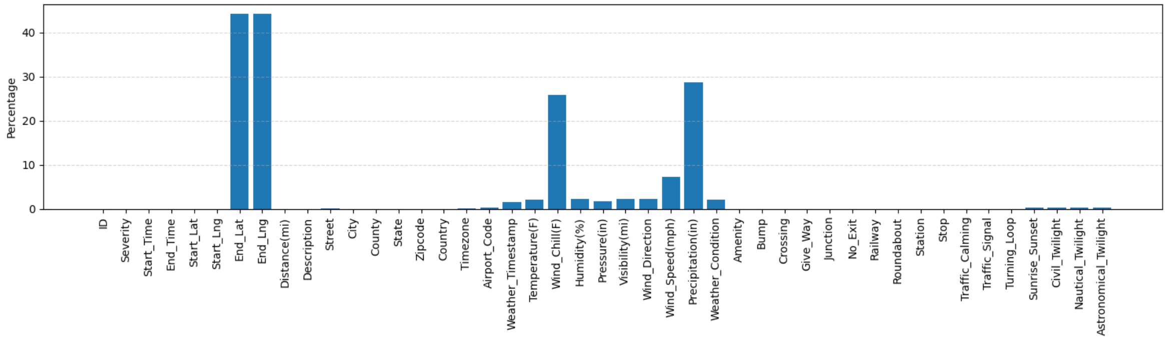


Figure 2: Percentages of missing values

The column `Wind_Chill(F)` had around 20% missing values after the initial removal of the rows. After observing the columns of the dataset, a noticeable pattern is observed, many `Wind_Chill(F)` values are the same as `Temperature(F)` values. This makes this column redundant, meaning it will be removed.

The column `Precipitation(in)` has around 24% missing values. Precipitation is an important statistic in road traffic accidents as it can be a contributing factor. A potential idea would involve filling the missing values with 0 entries symbolising no rain, however, some values in this column are already 0, with entries of `Weather_Conditions` as 'light rain'. The assumption is that precipitation is small and rounded to 0. Missing values will have to be left as *NaN* and imported into the database as *NULL*.

The columns `End_Lat`, `End_Lng` will be filled using the column `Distance(mi)`. A distance of 0 is shown to have the same End_Lng/Lat as start_Lng/Lat, meaning values can be copied over. Using this method, missing values have gone from 42% to 6%. The rest of the values are not critical and will be left as *NaN*.

### 3.3.2 Handling Outliers

The next course of action is to clean numeric values. A new column `difference` was created to plot the contrast between start and end times. Figure 3 and Figure 4, illustrate outliers in 7 of the 8 boxplots.

The outliers in columns `Temperature(F), Pressure_MB, Wind_Speed(mph), difference` are extreme values and were removed using the interquartile range method. Visibility outliers were only removed if they exceeded 11 miles, below this threshold, weather conditions like fog can be observed. Almost all precipitation values are at 0 inches, suggesting rain is rare or incorrectly measured. Values over 1 inch were removed because of their scarcity. Some distance outliers were removed during the removal process for other columns. The rest of the values were left untouched as they can provide valuable insights into traffic flow.

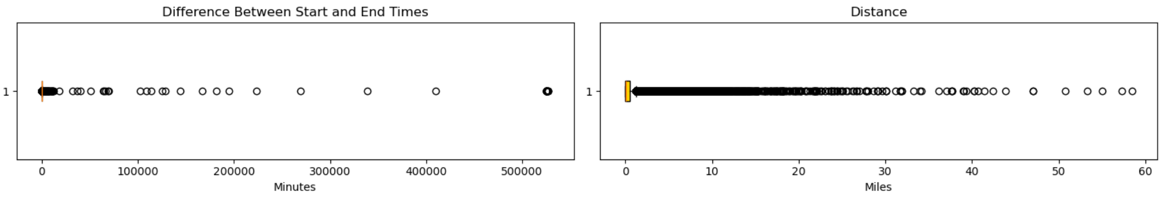


Figure 3: Boxplot of time differences and distances in accidents

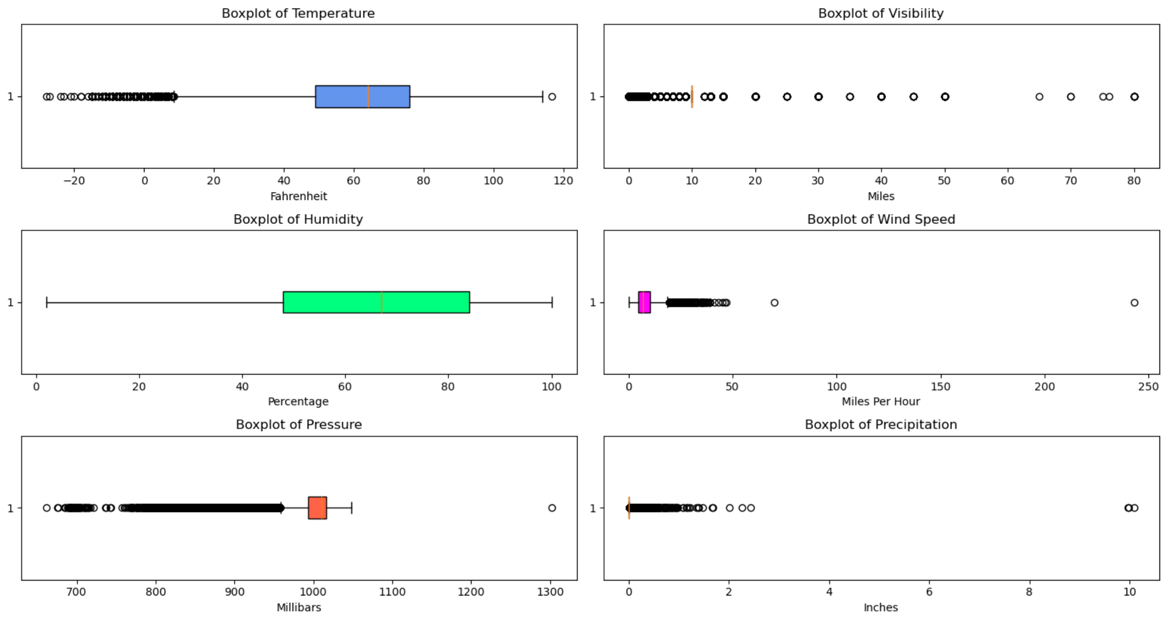


Figure 4: Boxplots of weather values

### 3.3.3 Attributes

Several attribute checks were required to improve clarity in the analysis. Pressure units were changed from inches to millibars. Columns Weather_Condition, Wind_Direction were standardised, for example, 'T-storm' to 'Thunderstorm'. Datetime columns were standardised removing any trailing digits past the second mark. Furthermore, 4 new columns were created to house the IDs for the primary keys.

## 3.4 Exploratory Data Analysis

To complete a clear analysis meeting the outlined objectives, a series of graphs were used to establish trends and relationships between variables in the data. These included:

- **Categorical Analysis:** Pie and bar charts were used to evaluate the spread of categorical data to find the dominant variables. This involved POI attributes, Weather_Conditions, State, Severity, and Description.
- **Numerical Analysis:** Histograms were applied to analyse the distribution of numerical values. This involved Weather attributes, Start/End_Times, and Distance_MI.
- **Temporal Analysis:** Using line and bar charts, variables were analysed for trends over time, including, frequencies over hours, days, months, and years. This involved Start/End_Times, Severity, Distance_MI, and total accidents.
- **Correlation Analysis:** Correlation heatmaps and scatter plots were used to identify relationships between Start/End_Times, Severity, Distance_MI.

# 4 Discussion of Key Findings

The pie charts in Figure 5 show the presence of a point of interest (POI) near the accident. There are three prominent POIs: crossings, traffic signals, and junctions. They often appear together, and running red lights is a high contributor to accidents[4]. Stations, amenities, and railways are congested areas with high levels of traffic, meaning accidents are less frequent due to lower speeds. Stops and give-ways seem to be associated with fewer accidents, possibly due to being less frequent. Traffic calming methods are excellent at reducing speed and therefore cause fewer accidents. Roundabouts are generally safer, much slower and compared to a junction, have 4 times lower conflict points. (A 4-way roundabout has only 8 vehicle conflict points compared to the 32 of a 4-way junction [14].) Turning loops, not shown in the figure, are at 0%, possibly because they are extremely rare.

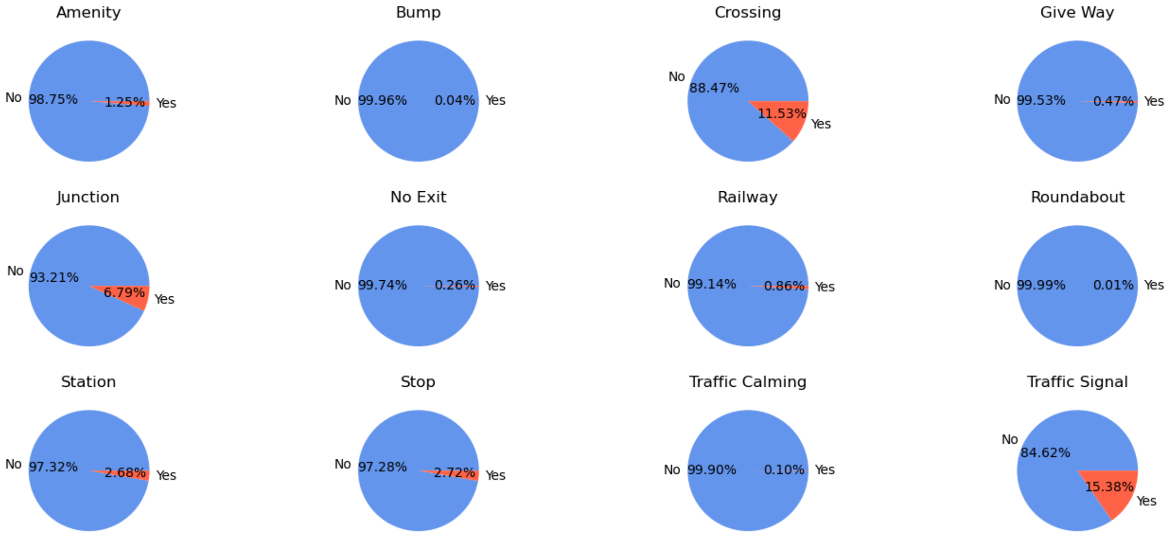


Figure 5: Points of interest

In Figure 6 the five most observed weather conditions are regular and cause no driving impairment. Weather like rain and fog are less common yet appear at high rates suggesting they are more significant contributors to accidents, possibly due to impairments like reduced visibility.

Temperature shows a negative skew with most temperatures around 63F, with extremes of 20F and 100F. Extremely low temperatures affect the battery of a vehicle while high temperatures cause overheating potentially leading to accidents. Humidity has little effect on vehicle performance, however can affect the driver by causing increased condensation impairing visibility. Wind speed does not affect vehicle or driver performance until 39mph, at which point breaking branches can be hazardous due to shock or car damage. [Wind Chart]

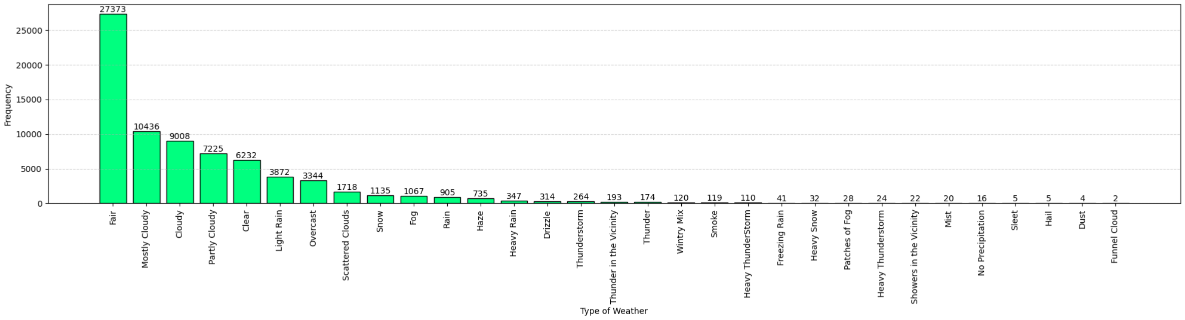


Figure 6: Described Weather Conditions

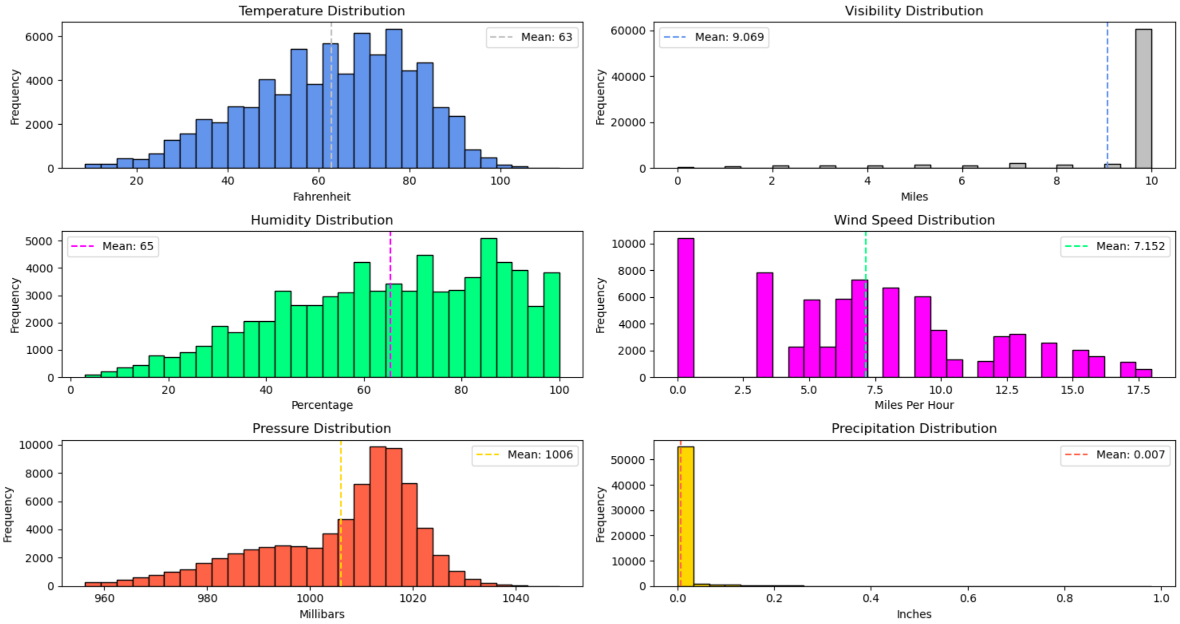


Figure 7: Distribution of weather values

In Figure 8, 48 states are shown with a majority in California, Florida, and Texas. The three states are the largest in the US[3]. As of 2025, California has no speed cameras and Texas has neither speed nor red light cameras. Alongside the population, this could be the reason why accident frequency is high, although Florida has both cameras[6].

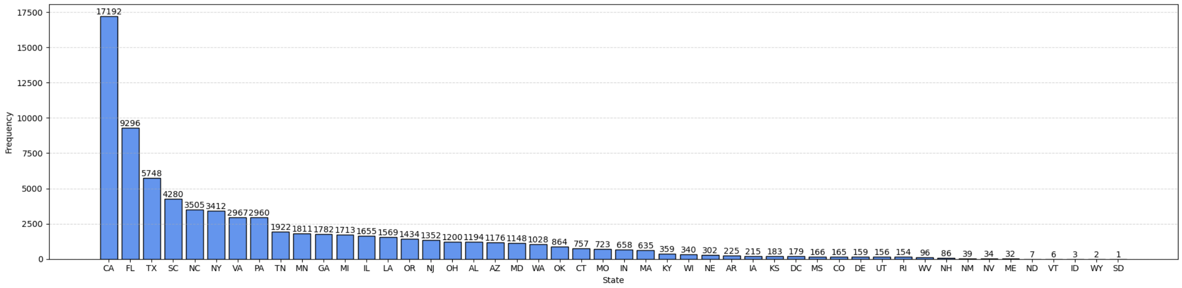


Figure 8: Accidents by state

The temporal analysis in Figure 9 illustrates clear trends. In the hourly graph, accident frequency is highest during both rush hours with a slight decline during the day. The lowest frequency occurs from 20 to 3, as less traffic is around. It is important to note that night driving is still dangerous due to hazardous factors such as visibility and driver fatigue.

In the daily graph, there is a gradual increase from the 1st to the 23rd, followed by a sharp drop. The decline from the 30th to the 31st is due to only half the months having 31 days.

Accident frequency is lowest in July, possibly due to reduced vehicle usage during summer weather, especially in California and Florida. From July, there is a sharp increase in the fall to winter months followed by a decline thereafter. High frequencies during winter can be associated with dangerous driving conditions such as icy roads, resulting in more accidents.

In Figure 10 there is a slow increase in accidents per month. The same monthly pattern can be seen as accidents increase during fall and winter. During COVID-19, a sharp decline occurred as more people started working from home and less travel was necessary. However, over the months after the decline, the frequency jumped to nearly double what it was before. It can be speculated that people drove more recklessly over this time[12].

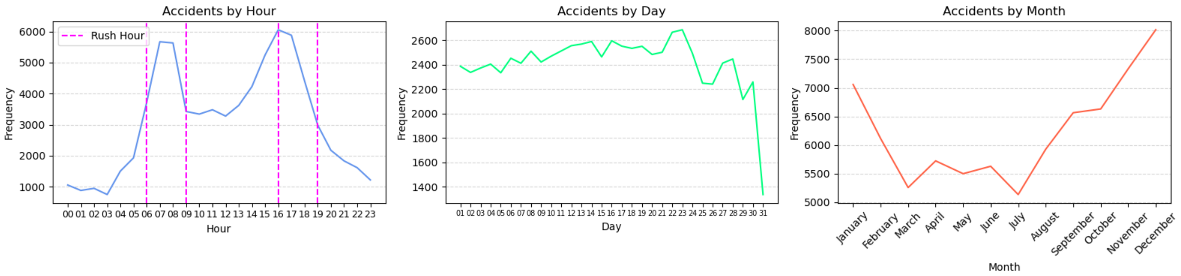


Figure 9: Grouped accident frequency

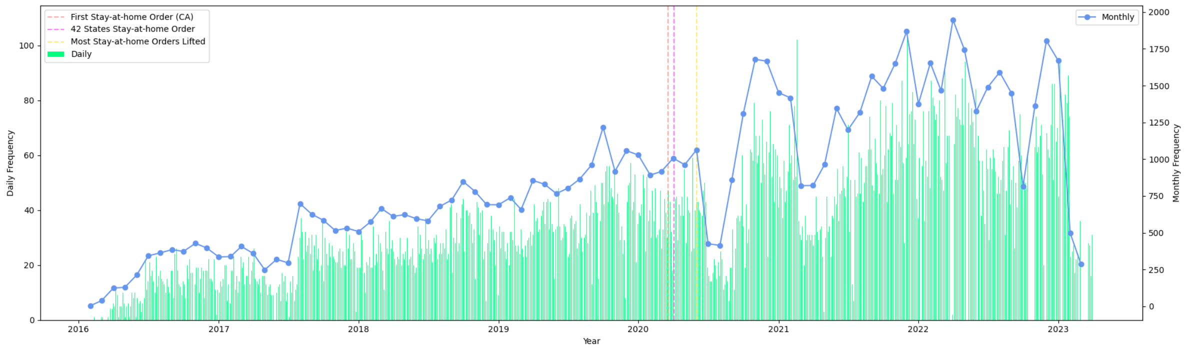


Figure 10: Accidents frequency from 2016-2023

In Figure 11, time to settle traffic flow is positively skewed with most values around 25 to 75 minutes. The mean is considerably higher than the median, associating higher values. During the year, the time follows a similar pattern to monthly frequency.

The majority of traffic flow is classified at severity 2, with some severity 3 and few in 1 and 4. Monthly severity seems to hover around the mean with the winter months hovering below the mean, suggesting an inverse relationship to frequency and time.

The extent of the road affected is positively skewed with most values around 0. The mean 0.476 miles suggests most accidents result in small disruptions. Higher values can suggest long stretches of road without diversion such as highways. Monthly, the extent seems to follow a similar monthly pattern to frequency, but with added seasonal effects. In June/July people will leave for the summer holiday increasing the amount of vehicles on these long stretches of road.

As seen in Figure 12, there is a negative correlation between severity and time and a slight positive correlation between distance and severity. This is unusual and creates ambiguity as to what severity is based on. The positive relationship between distance and time is to be expected. In the scatter graph, most severity 3 values are under 50 minutes with a wide stretch between miles. Severity 2 seems to dominate below 10 miles through all waiting times. Severity 1 is seen only at the start and severity 4 seems to be randomly scattered.

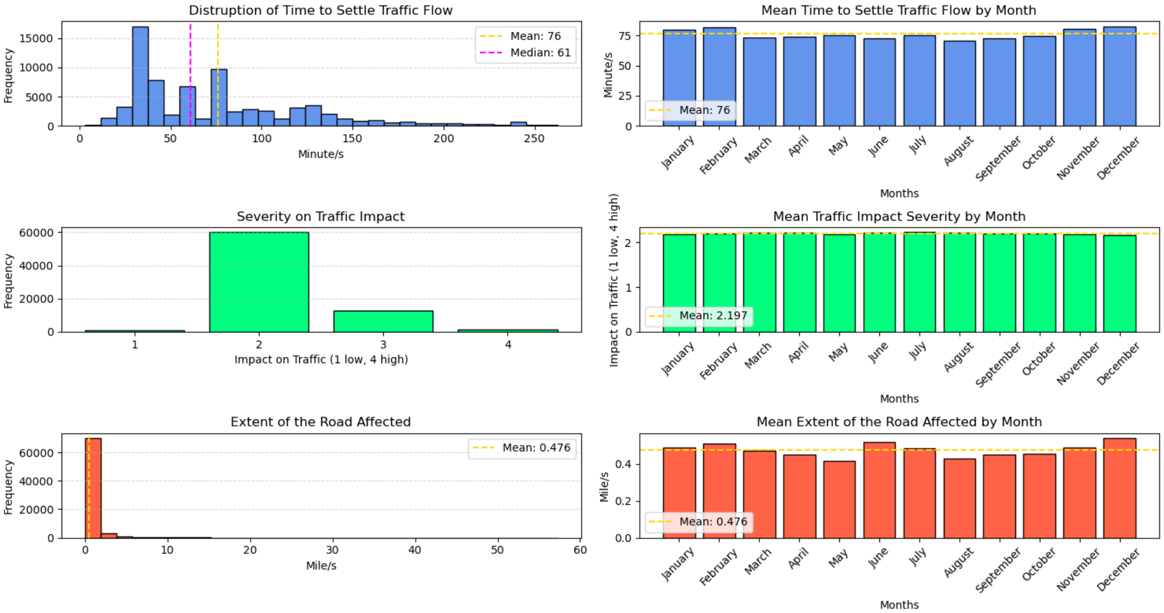


Figure 11: Distributions of traffic elements

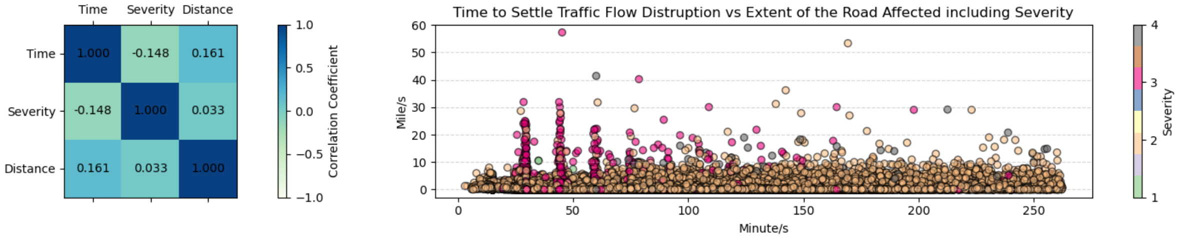


Figure 12: Relationships of traffic elements

The keyword analysis in Figure 13 is used to find common phrases in descriptions. 'Accident on' and 'Incident on' are general statements to report the location of an accident. There are many words describing particular traffic settings such as 'lane blocked' but most descriptions do not have them. Overall, the description system is not standardised hence most likely inefficient.

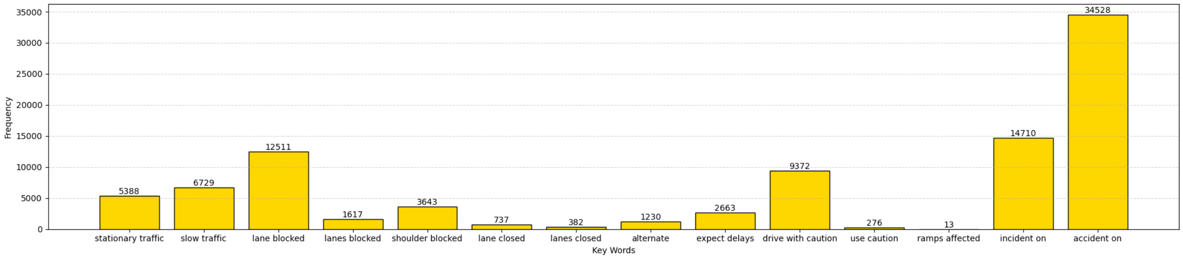


Figure 13: Key words in the traffic description

## 5 Conclusions and Recommendations

Infrastructure analysis indicates areas with crossings, traffic signals and junctions are consistent with higher accident frequencies. The temporal analysis reveals that most accidents occur during rush hours, when traffic volume is high, however, nighttime frequency stays relatively high for the decreased traffic. Winter months show the most accidents, despite this, weather analysis indicates that clear and cloudy weather dominates the accident count, suggesting drivers should exercise more caution during average weather.

Across the 7-year dataset with 76 thousand samples and a mean waiting time of 76 minutes, a combined 11 years' worth of cumulative waiting time occurs. The unsampled dataset contains nearly 8 million samples, which would mean cumulative waiting time is multiplied by at least 100. This is a perplexing amount that highlights the costs of traffic accidents, including loss of productivity but also delayed emergency response times.

Some of the recommendations that need to be taken into consideration are:

- **Infrastructure and Law:** Prioritising measures of speed reduction such as roundabouts and speed bumps should result in fewer accidents. States can lower speed limits in areas of high congestion but must reach a balance with potential increased traffic.
- **Driver Awareness:** Drivers must care about areas of higher risk such as junctions. They must show more caution around weather that is perceived to have little risk as accident frequency is high.
- **Accident reporting:** Accidents are inevitable, so a better traffic description system can ensure that both emergency services and commuters better understand what is happening. A system emphasising the location followed by the traffic situation could result in clearer communication and response times.

Code: https://colab.research.google.com/drive/1bp_ZtDQdVBzLU_S8XNIUDm-ycqeC1GFR?usp=sharing

## References

[1] National Highway Traffic Safety Administration. Nhtsa: Traffic crashes cost america $340 billion in 2019. https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019, 2023.

[2] National Highway Traffic Safety Administration. Overview of motor vehicle traffic crashes in 2022. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813560, 2024.

[3] United States Census Bureau. U.s. and world population clock. https://www.census.gov/popclock/, 2025.

[4] Insurance Institute for Highway Safety and Highway Loss Data Institute. During covid-19, road fatalities increased and transit ridership dipped. https://www.iihs.org/topics/red-light-running, 2024.

[5] Insurance Institute for Highway Safety and Highway Loss Data Institute. Seat belts. https://www.iihs.org/topics/seat-belts, 2024.

[6] Insurance Institute for Highway Safety and Highway Loss Data Institute. Safety camera laws. https://www.iihs.org/topics/red-light-running/safety-camera-laws, 2025.

[7] Hannah Hilst. State traffic laws. https://www.findlaw.com/traffic/traffic-tickets/state-traffic-laws, 2023.

[8] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset, 2019.

[9] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, page 33–42. ACM, November 2019.

[10] U.S. Department of Transportation. Budget estimates. https://www.transportation.gov/sites/dot.gov/files/2021-05/NHTSA-FY-2022-Congressional-Justification.pdf, 2022.

[11] U.S. Department of Transportation. Investing in america: Biden-harris administration announces more than $1 billion in grants for 350+ communities to make local roads safer. https://www.transportation.gov/briefing-room/investing-america-biden-harris-administration-announces-more-1-billion-grants-350-0, 2024.

[12] U.S. Government Accountability Office. During covid-19, road fatalities increased and transit ridership dipped. https://www.gao.gov/blog/during-covid-19-road-fatalities-increased-and-transit-ridership-dipped, 2022.

[13] World Health Organization. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, 2023.

[14] Elle Riorson. Do roundabouts really help reduce car accidents? https://somethingaboutorange.com/do-roundabouts-really-help-reduce-car-accidents/, 2024.