# Data Analysis Final Assignment Report

Team: Circuit Synergy
Lorenz Buchinger & *Jeremia Baumgartner* & *Tim Zwölfer*

## 1 Contributions

- Lorenz Buchinger:
  - Data Preprocessing and Data Quality
  - Regression and Predictive Modeling
- Jeremia Baumgartner:
  - Visualization and Exploratory Analysis
  - Dimensionality Reduction and Statistical Tests
- Tim Zwölfer:
  - Probability and Event Analysis
  - Statistical Theory Applications

## 2 Dataset Description

- **Dataset name and source**: Traffic Accidents dataset from Kaggle
- **Time span**:
  - **First entry**: 1st of January 2016
  - **Last entry**: 31st of December 2025
- **Frequency**: Not constant, one entry per crash
- **Key variables analyzed**: various variables indicating injury severity, time variables
- **Size and structure**:
  - **Number of observations (rows)**: 209306
  - **Number of features (columns)**: 24
- **Missing data summary**: The dataset seems to be incomplete until November of 2017.
- **Known limitations**: The geographic location of the crashes is unknown.

# 3 Task 1. Data Preprocessing and Basic Analysis

## 3.1 Basic statistical analysis using pandas

**Descriptive stats for key variables:**

| | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| num_units | 2.06 | 0.40 | 1.00 | 2.00 | 2.00 | 2.00 | 11.00 |
| injuries_total | 0.38 | 0.80 | 0.00 | 0.00 | 0.00 | 1.00 | 21.00 |
| injuries_fatal | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 3.00 |
| injuries_incapacitating | 0.04 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 7.00 |
| injuries_non_incapacitating | 0.22 | 0.61 | 0.00 | 0.00 | 0.00 | 0.00 | 21.00 |
| injuries_reported_not_evident | 0.12 | 0.45 | 0.00 | 0.00 | 0.00 | 0.00 | 15.00 |
| injuries_no_indication | 2.24 | 1.24 | 0.00 | 2.00 | 2.00 | 3.00 | 49.00 |

**Grouped summaries:**

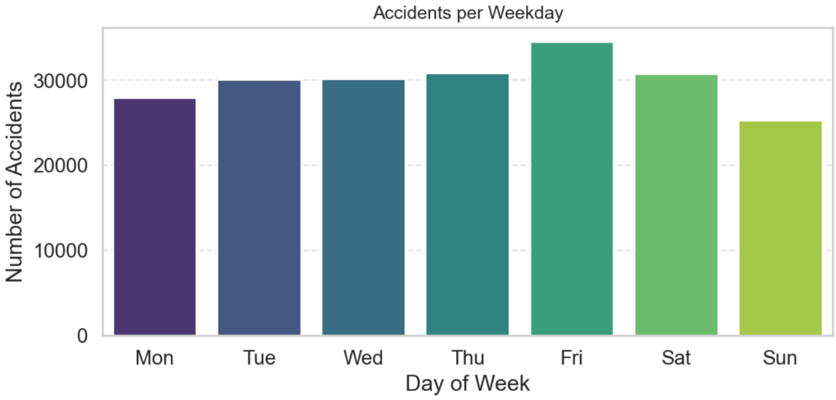


Figure 1: Accidents per Weekday

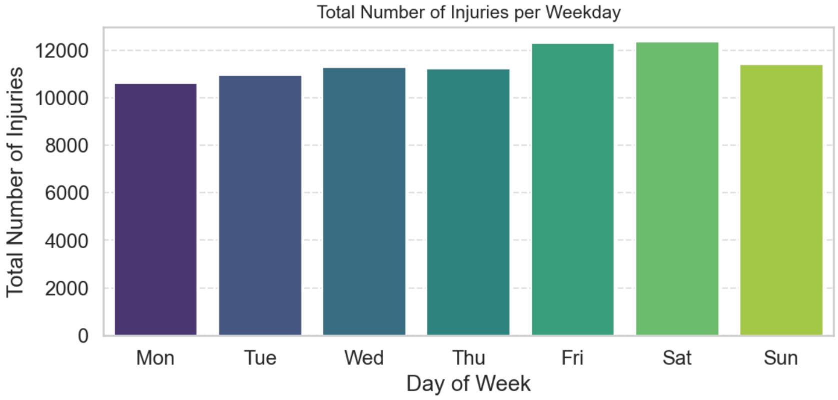


Figure 2: Total Number of Injuries per Weekday

When looking at figures 1 and 2 one can see that the probability of an accident is highest on Fridays. The most accidents with injuries happen on Fridays and Saturdays. Furthermore, the probability for an accident to occur and for injuries to be caused by accidents rises from Monday to Friday.

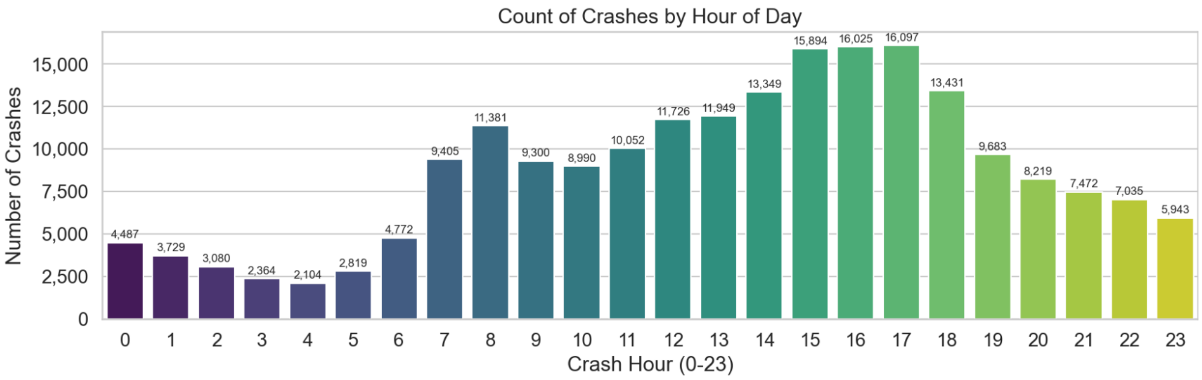


Figure 3: Count of Crashes per Hour of Day

When looking at figure 3 one can see that most crashes occure during rush-hours.

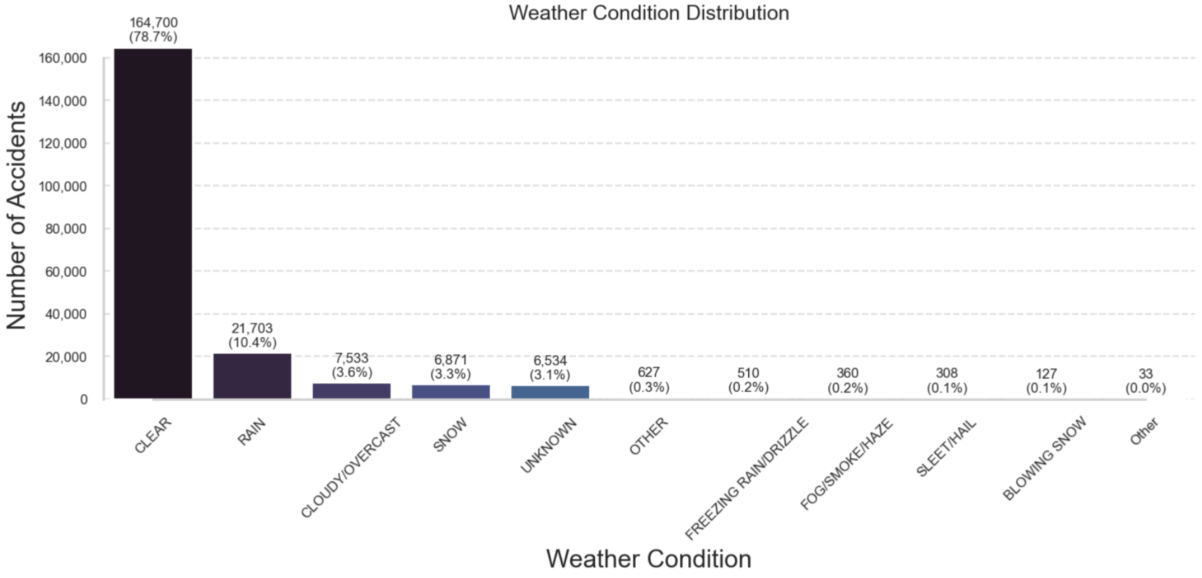


Figure 4: Number of Accidents per Weather Condition

No correlation between bad weather and the likelihood of an accident can be found. The reason for this is that probably the weather is usually clear where the measurements were taken.

## 3.2 Original data quality analysis including visualization

**Missingness patterns:**

Missing values were encoded with the string ”UNKNOWN”. These values were replaced with nan values for easier processing. Only 6 of the 24 features had missing values. A summary can be seen in the following table:

| Variable | Count | Percentage |
|---|---|---|
| road_defect | 34426 | 16.45 |
| roadway_surface_cond | 12509 | 5.98 |
| weather_condition | 6534 | 3.12 |
| traffic_control_device | 4455 | 2.13 |
| lighting_condition | 4336 | 2.07 |
| trafficway_type | 1060 | 0.51 |

The time-distribution of the missing values can be seen in the heatmap in figure 5.

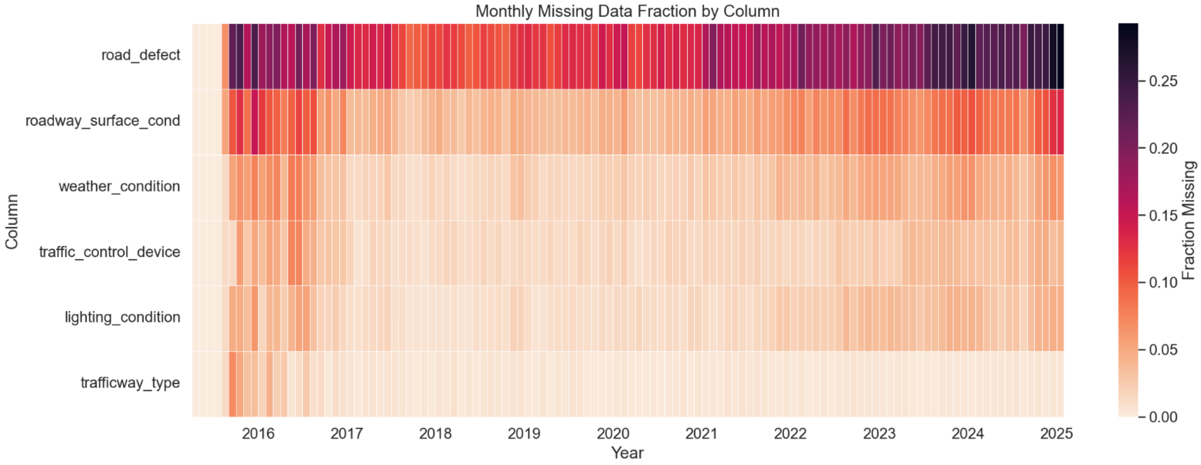


Figure 5: Heatmap of missing values over time. Resolution is 1 week.

**Outliers and suspicious values:**

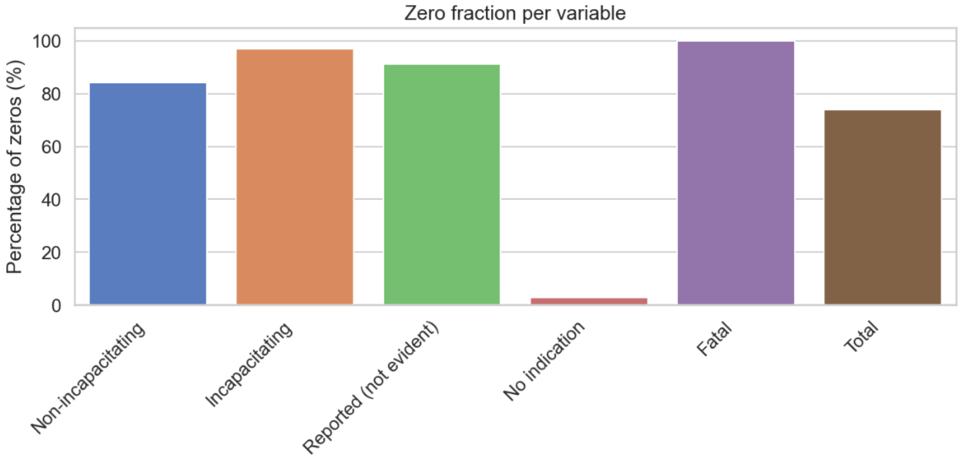


Figure 6: Percentage of type of injury not occuring in accident.

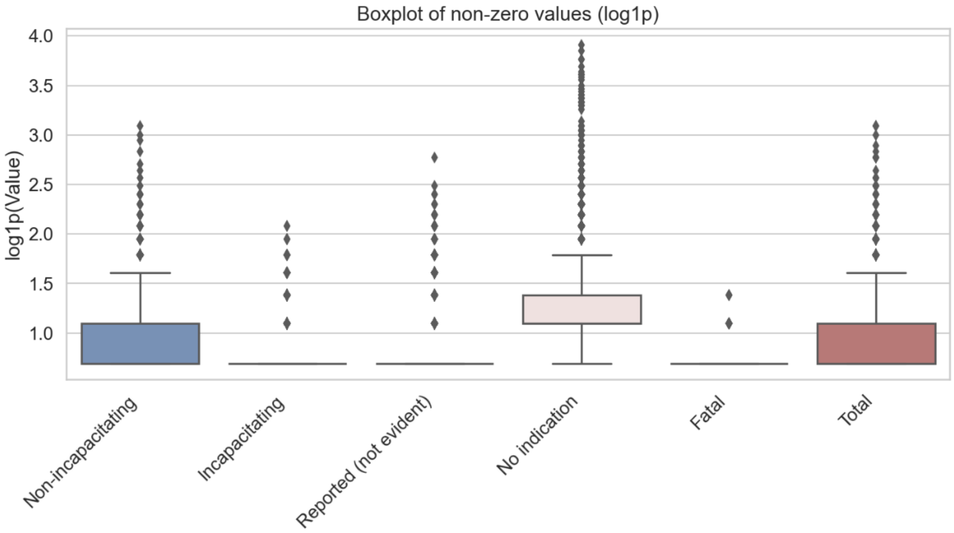


Figure 7: Logarithmically scaled boxplots of non-zero values of injuries.

Since most accidents happen without causing any injuries, basically all injuries are outliers. The percentage of zero-values per injury per accident can be seen in figure 6. Logarithmically scaled box-plots of the non-zero values of injury occurrences can be seen in figure 7.

**Consistency checks:**
All variables were check for consistency. No inconsistencies were found.

## 3.3 Data preprocessing

For cleaning all ”UNKNOWN” strings were replaced with nan values. When observing the the recorded accidents per week it became evident that the records before November of 2017 are not complete. The highlighted time span in figure 8 has gaps between entries of more than 1 day.

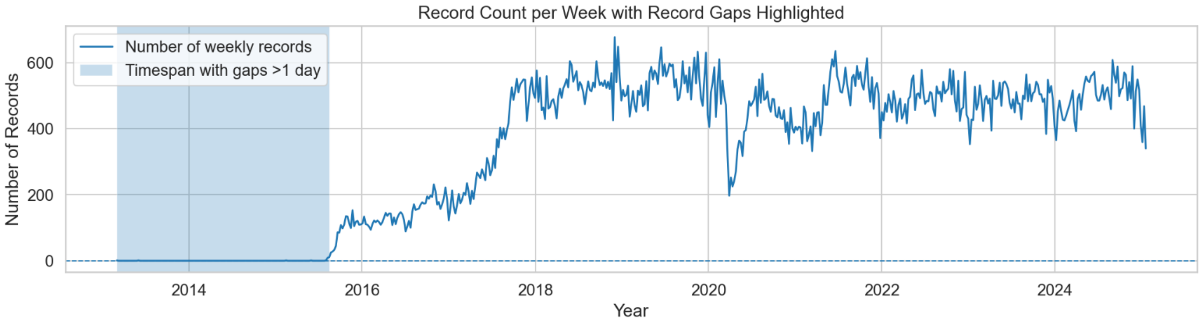


Figure 8: Logarithmically scaled boxplots of non-zero values of injuries.

The decision was made to remove all records before November of 2017. A drop of records can also be seen during the time of the corona virus pandemic (begin of 2020 to mid of 2021). No possibility to correct this anomaly was found.

### 3.4 Preprocessed vs original data visual analysis

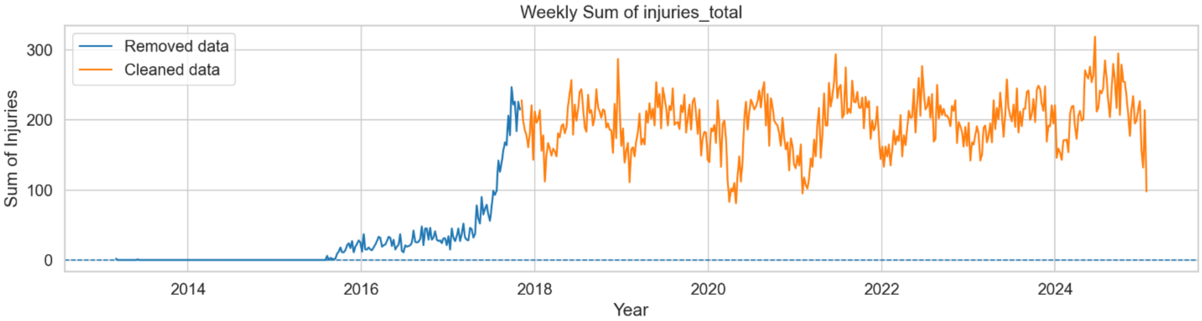


Figure 9: Removed vs. cleaned data of the injuries_total feature.

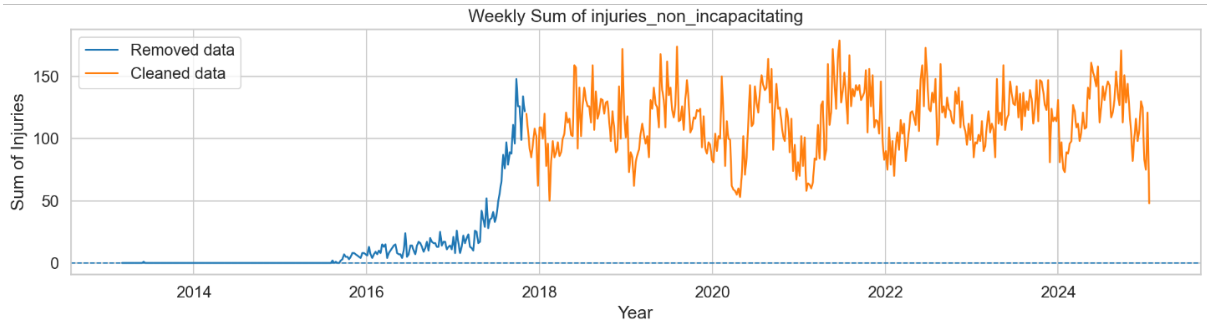


Figure 10: Removed vs. cleaned data of the injuries_non_incapacitating feature.

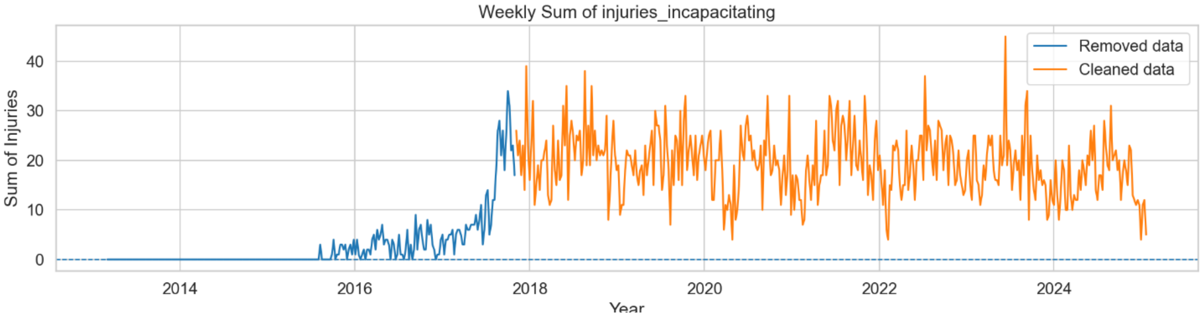


Figure 11: Removed vs. cleaned data of the injuries_incapacitating feature.

# 4 Task 2. Visualization and Exploratory Analysis

## 4.1 Time series visualizations

- Plot of main variable(s) over time:

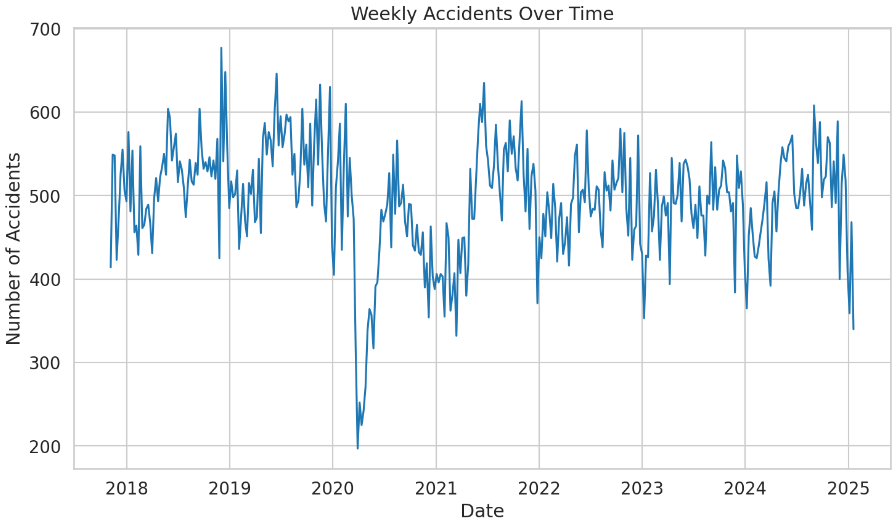


Figure 12: Time series visualizations of weekly accidents

- Annotations for notable events or pattern shifts (if applicable): In the first halve of 2020 there is a big dip in accidents caused by covid lockdowns. Every new year starts with a dip, possibly due to low traffic on holidays.

## 4.2 Distribution analysis with histograms

- Histograms for key numeric variables:

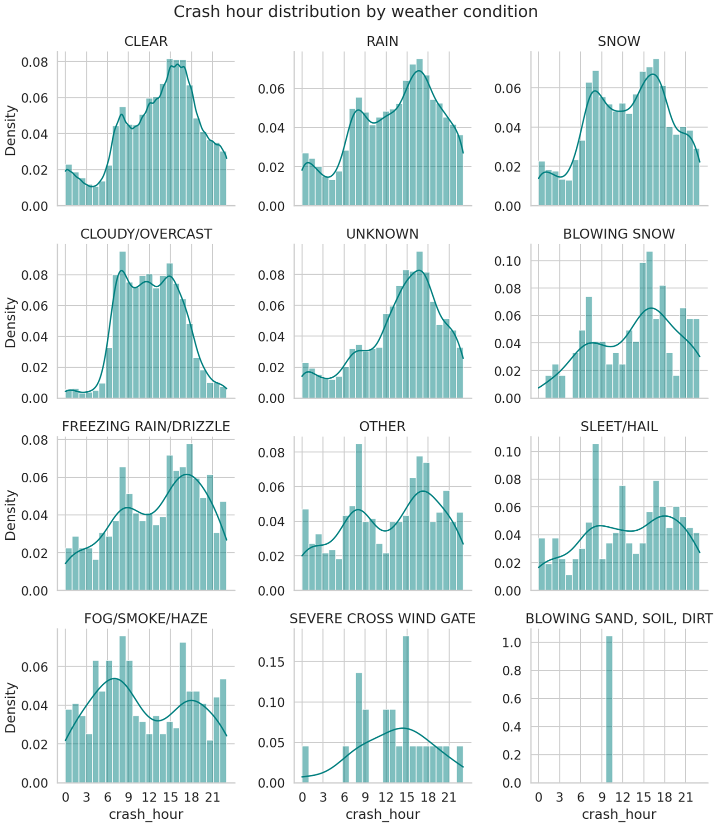
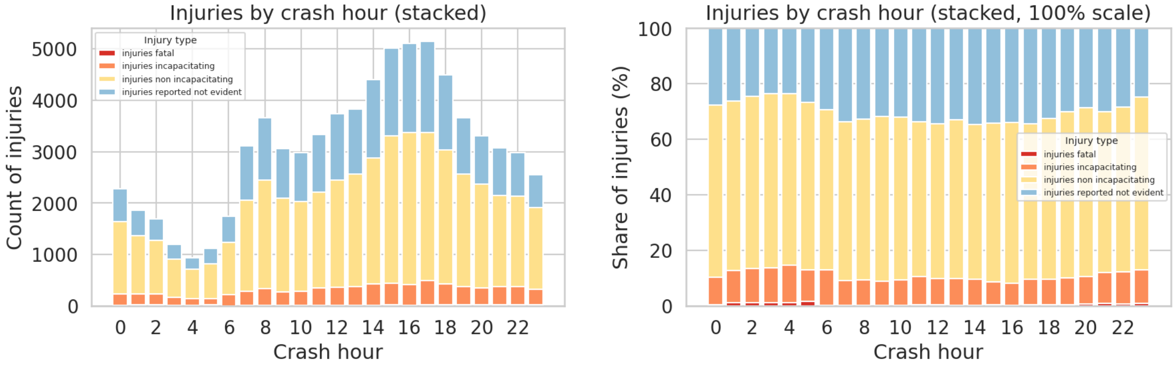


Figure 13: Density plot of crash hour with different weather conditions



(a) absolute count of injuries

(b) relative injuries per hour

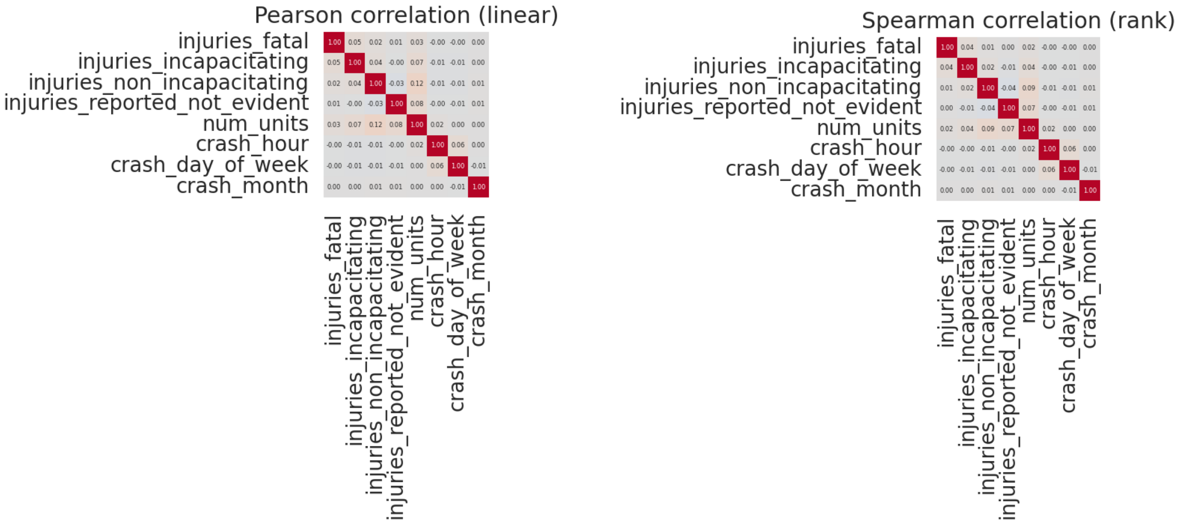
Figure 14: Histogram of the crash hour stacked by injury type

- Notes on skewness, heavy tails, multi-modality: Most plots have two peaks and are skewed to the right. Some show a higher density in the night (and a flatter curve) than others (e.g.: Fog).
  Fatal and incapacitating happen more during night time.

### 4.3 Correlation analysis and heatmaps

- Correlation type used (Pearson or Spearman) and why: Both types were tried but didn't get good results.

- Heatmap and top correlated pairs with short interpretation:



(a) pearson

(b) spearman

Figure 15: Correlation plots

### 4.4 Daily pattern analysis

- Aggregation method (weekly means, day-of-week):

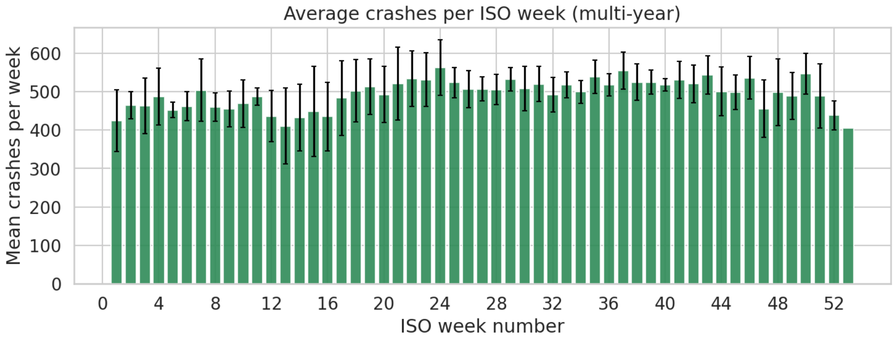


Figure 16: Weekly seseanality plot

### 4.5 Summary of observed patterns, similar to True/False questions

*Write short, testable statements and answer them based on evidence. Example format below.*

- Crash hour distribution differs by weather condition: **True**. Evidence: figure 13
- Fatal and incapacitating injuries are relatively more in night time. **True**. Evidence: figure 14
- The numerical columns correlate. **False**. Evidence: 15
- In the Chrisms holydays the number of accidents is reduced. **True**. Evidence: figure 16

## 5 Task 3. Probability Analysis

### 5.1 Threshold-based probability estimation

- **Event 1: Late Hour Crash (20:00–05:00)**
  *Relevance:* Fatigue, poor visibility, alcohol.
  *Empirical Probability:* 22.856%
- **Event 2: Rush-Hour Crash (07:00–09:00, 16:00–18:00)**
  *Relevance:* High traffic, stop-and-go, time pressure.
  *Empirical Probability:* 35.926%
- **Event 3: Severe Crash**
  *Relevance:* Fatal or incapacitating injuries.
  *Empirical Probability:* 3.397%
- **Event 4: Non-Severe Crash**
  *Relevance:* Most crashes are minor.
  *Empirical Probability:* 96.603% *Note:* Complement of Event 3; together their probabilities sum to 100%.
- **Event 5: Rainy Weather Crash**
  *Relevance:* Reduced friction and visibility.
  *Empirical Probability:* 10.532%
- **Event 6: Snowy Weather Crash**
  *Relevance:* Low friction, impaired control.
  *Empirical Probability:* 3.552%

- **Event 7: Poor Visibility Weather Crash**
  *Relevance:* Limited hazard detection.
  *Empirical Probability:* 0.394%
- **Event 8: Clear Weather Crash**
  *Relevance:* Baseline driving conditions.
  *Empirical Probability:* 82.141%

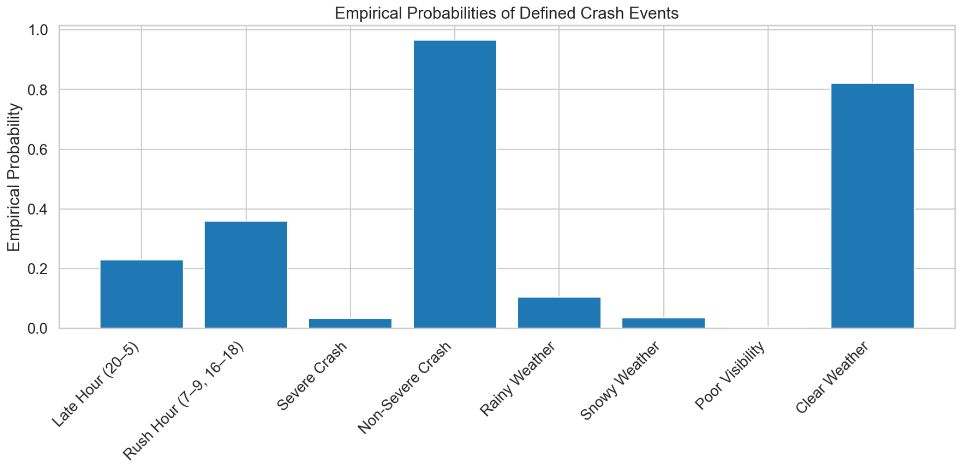


Figure 17: Bar plot of empirical probabilities for selected crash events.

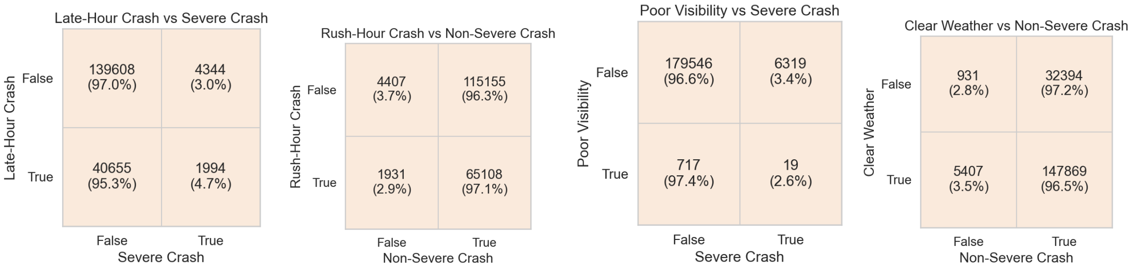
## 5.2 Cross-tabulation analysis

**Late Hour vs Severe Crash** Late-hour crashes have a higher proportion of severe outcomes (4.7%) than non-late-hour crashes (3.0%), confirming increased risk at night.

**Rush Hour vs Non-Severe Crash** Most rush-hour crashes are non-severe (97.1%), supporting the idea that congestion reduces severity despite higher frequency.

**Poor Visibility vs Severe Crash** Severe crashes under poor visibility are rare (2.6%), slightly below the overall severe crash rate, likely due to caution and low frequency.

**Clear Weather vs Non-Severe Crash** Clear conditions show mostly non-severe crashes (96.5%), serving as a baseline.



(a) Late Hour vs Severe

(b) Rush Hour vs Non-Severe

(c) Poor Visibility vs Severe

(d) Clear Weather vs Non-Severe

Figure 18: Cross-tabulation results for selected crash-related categorical variables.

### 5.3 Conditional probability analysis

**Late-Hour and Severe Crash** $P(\text{Severe} \mid \text{Late Hour}) = 4.675\% > P(\text{Severe}) = 3.397\%$, showing higher risk at night. $P(\text{Late Hour} \mid \text{Severe}) = 31.461\%$.

**Poor Visibility and Severe Crash** $P(\text{Severe} \mid \text{Poor Visibility}) = 2.582\% < P(\text{Severe})$, and $P(\text{Poor Visibility} \mid \text{Severe}) = 0.3\%$, reflecting low incidence and cautious driving.

**Rush-Hour and Non-Severe Crash** $P(\text{Non-Severe} \mid \text{Rush Hour}) = 97.120\% > P(\text{Non-Severe})$, with $P(\text{Rush Hour} \mid \text{Non-Severe}) = 36.118\%$, indicating congestion reduces severity.

### 5.4 Summary and limitations

**Thresholds** Crash occurrence depends on time and weather; higher frequency does not imply higher severity. Severe crashes remain rare.

**Cross-tabulation** Late-hour crashes are more severe; rush-hour crashes are mostly non-severe. Clear weather is baseline; poor visibility has minor effect.

**Conditional probabilities** Late-hour increases severe risk; rush-hour favors non-severe; poor visibility has limited impact.

**Limitations** Assumes independent crashes, accurate binary definitions, and correct weather/lighting. Severe crashes are rare, thresholds are conventional, and factors like speed, alcohol, and traffic volume are not controlled.

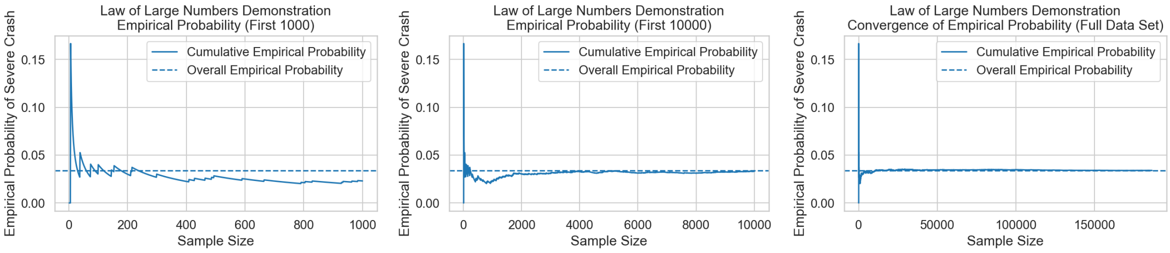
## 6 Task 4. Statistical Theory Applications

### 6.1 Law of Large Numbers (LLN) demonstration

**Variable** The severe-crash indicator was used (1 = severe, 0 = non-severe). This Bernoulli variable allows the sample mean to directly represent the empirical probability of a severe crash.

**Experiment** Observations were randomly shuffled. For increasing sample sizes $n$, the cumulative mean was computed and compared to the overall empirical probability.

**Observations** Small samples show large fluctuations, which may suggest misleading probabilities. As $n$ grows, the sample mean stabilizes. Around 10,000 observations, the empirical probability converges to the full-sample value, illustrating the Law of Large Numbers.



(a) First 1000 obs. (b) First 10000 obs. (c) Full dataset

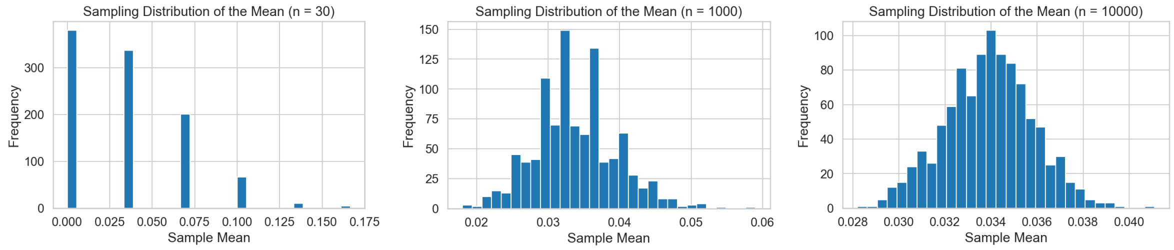
Figure 19: Convergence of the cumulative probability of severe crashes.

### 6.2 Central Limit Theorem (CLT) application

**Sampling procedure** Bootstrap sampling with replacement was performed using $n = 30, 1000, 10000$. For each $n$, 1,000 samples were drawn, and the sample mean computed.

**Sampling distributions** Each distribution represents the mean of sampled severe-crash indicators. Small $n$ yields wide, irregular distributions; larger $n$ produces narrower, more symmetric distributions.

**Interpretation** Despite the binary and skewed nature of the data, sample means approach a normal distribution as $n$ increases, consistent with the Central Limit Theorem. Deviations at small $n$ result from the low incidence of severe crashes.



(a) $n = 30$ (b) $n = 1000$ (c) $n = 10000$

Figure 20: Sampling distributions of the mean for increasing sample sizes. The distributions narrow and approach a normal shape with larger $n$.

### 6.3 Result interpretation and sanity checks

**LLN** The cumulative mean of severe crashes converges to the overall empirical probability as $n$ increases, confirming the Law of Large Numbers. Early fluctuations show that small samples can produce unstable probability estimates.

**CLT** Sampling distributions of the mean narrow and become symmetric as $n$ increases, approaching normality. Deviations at small $n$ are expected due to the low frequency of severe crashes and binary data.

**Sanity checks** - Observations were shuffled to avoid order effects. - Multiple $n$ values verified stable convergence. - Large numbers of samples ensured reliable sampling distributions.

**Potential violations** Dependence between observations, misclassified crash severity, or insufficient sample sizes could invalidate LLN and CLT conclusions.

## 7 Task 5. Regression Analysis

### 7.1 Linear or Polynomial model selection

A XGBoost (Extreme Gradient Boosting) regressor was chosen instead of a linear or polynomial model. XGBoost automatically captures complex non-linear patterns and feature interactions (like how weekend injury rates vary by season), making it far more powerful than linear or polynomial models.

The target feature was selected to be the daily total injuries. This feature had to be created by summing up all injuries per day. The features ”dayofweek”, ”day_of_year” and ”quarter” were derived from ”crash_date” and used as predictors. To smoothen the transition of the target a rolling average filter with a window size of 7 days was applied.

Data from November of 2017 to January 2024 were used as training data. The predictions were made for the whole year of 2024.

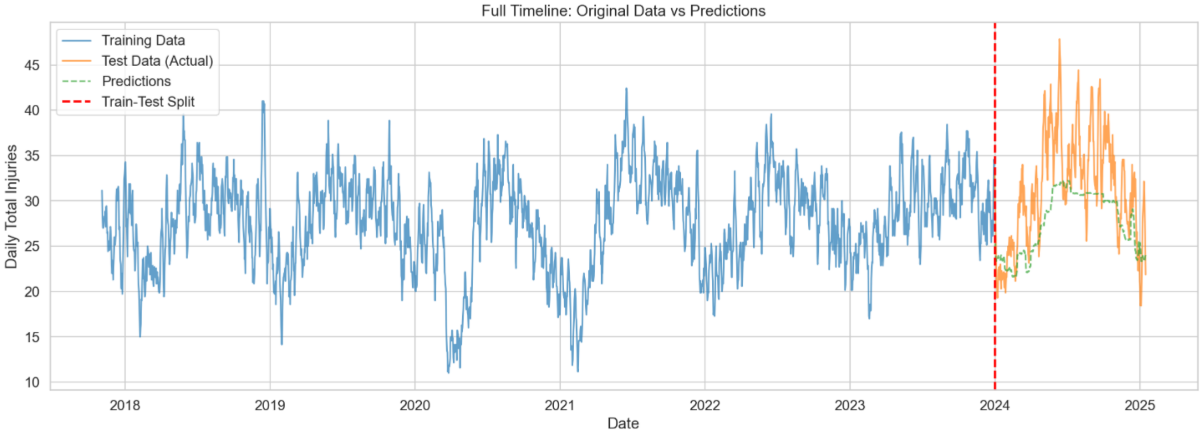


Figure 21: Training data, test data and predicted data.

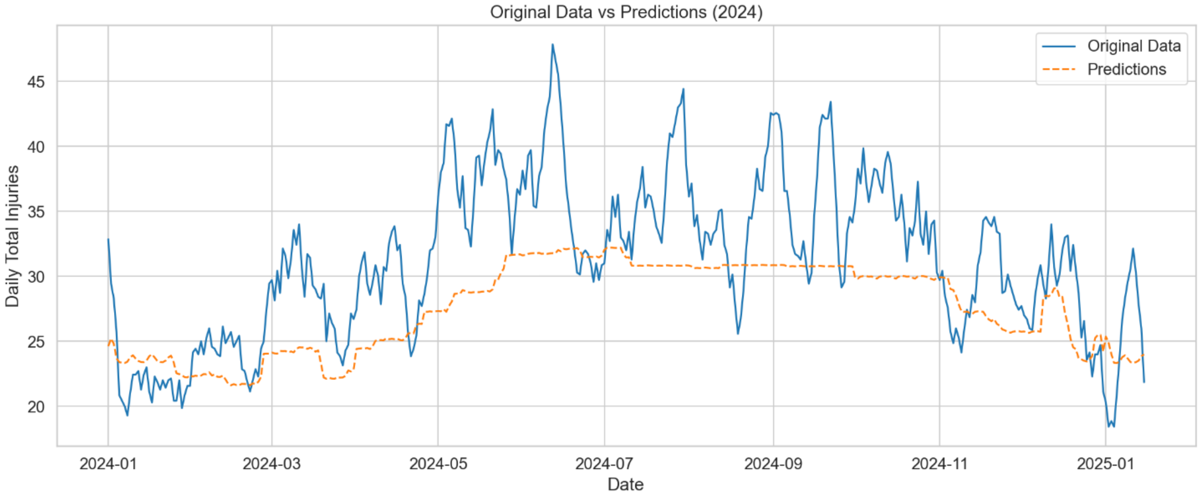


Figure 22: Test data and predicted data for 2024.

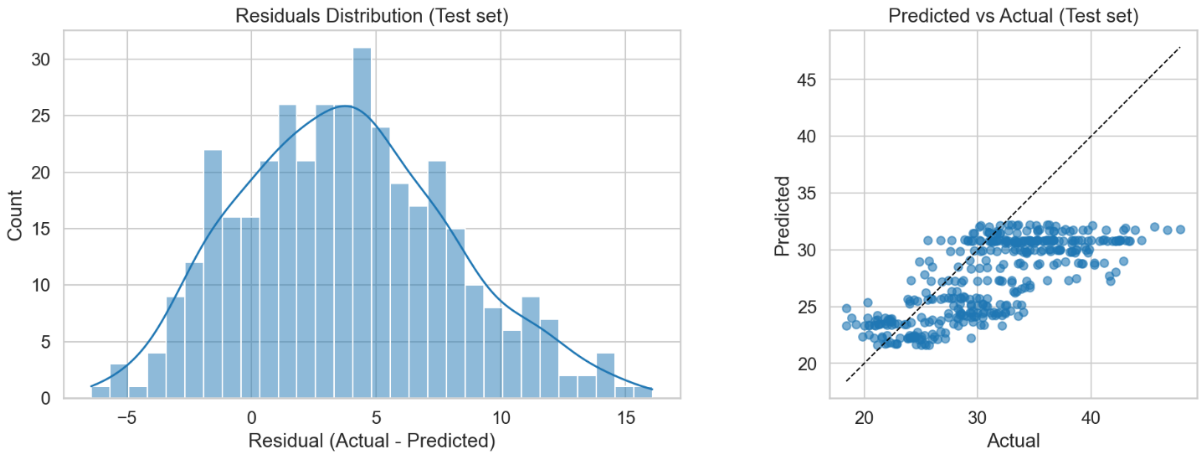


Figure 23: Residual analysis

The model was able to roughly predict the values. There are visible plateaus in the predicted data. The reason for that is that not enough features were supplied to the model.

The predicted values are also smaller than the actual values. The reason for this is that in the year of 2024 more injuries occurred than in the previous years the model was trained on.

Possible improvements are that more features should be used to describe the fluctuations. Also the input data could be pre-processed better. For example outliers could be removed.

| **Metric** | **Value** |
|---|---|
| Test RMSE | 5.662 |
| Test MAE | 4.533 |
| Test $R^2$ | 0.131 |
| Test MAPE | 13.57% |

Table 1: Performance Metrics

# 8 Task 6. Dimensionality Reduction and Statistical Tests

## 8.1 PCA

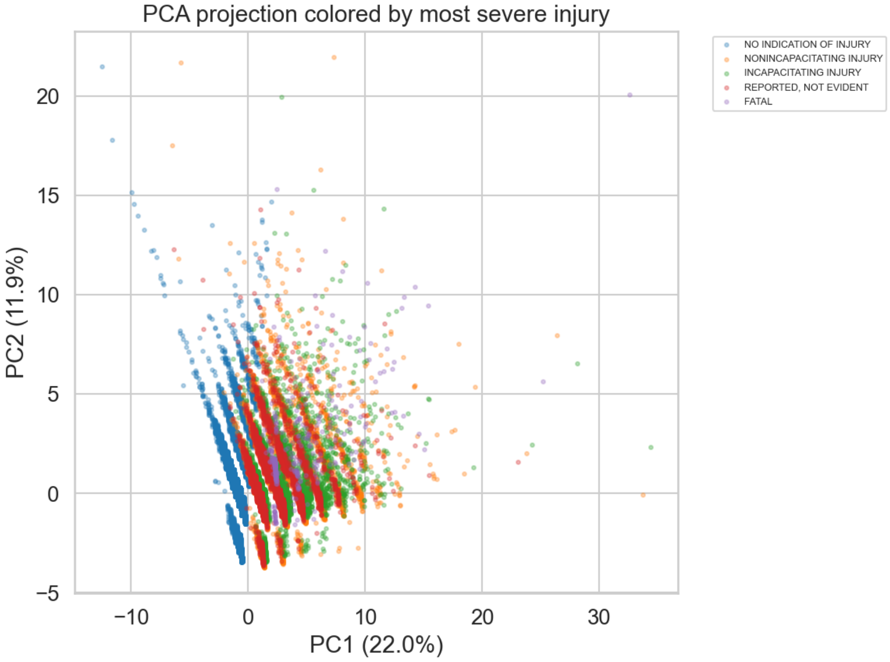


Figure 24: PCA Plot colored by injury type

## 8.2 t-SNE

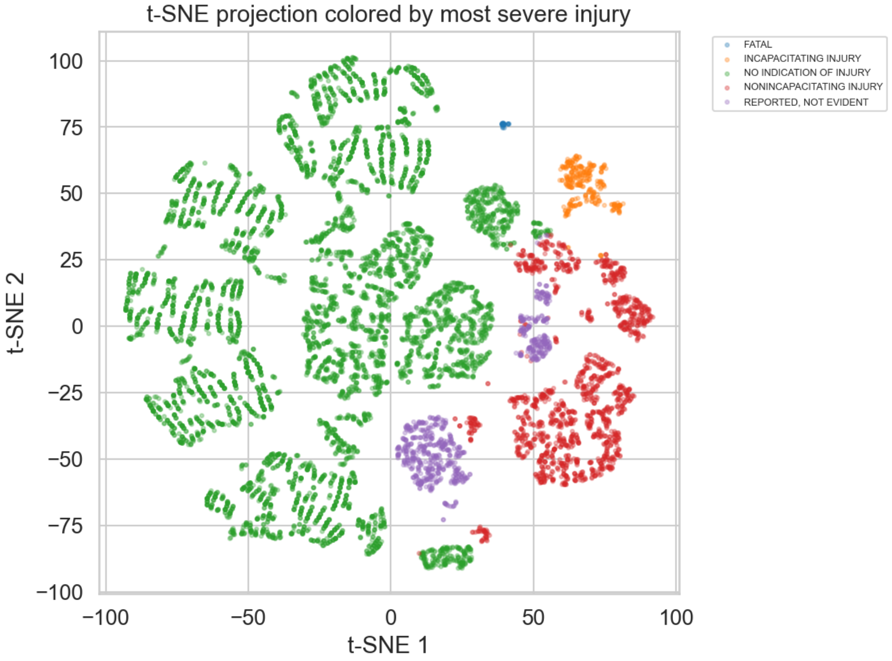


Figure 25: t-SNE plot colored by injury type

The t-SNE plot shows that injury severity in traffic crashes is not random: severe injuries and fatalities occur in specific, repeatable crash scenarios, while no-injury crashes occur across a wide variety of conditions.

### 8.3 UMAP

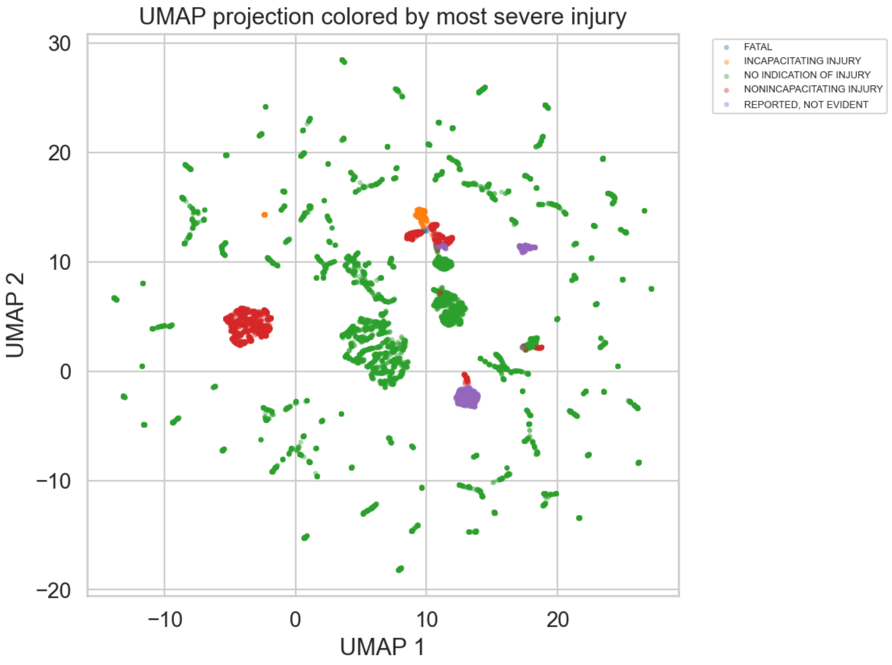


Figure 26: UMAP plot colored by injury type

## 9 Key Findings and Conclusions

- Main findings from preprocessing and EDA: Missing data was found primarily in road_defect (16.45%), roadway_surface_cond (5.98%), and weather_condition (3.12%), with all ”UNKNOWN” values converted to NaN for processing. Records before November 2017 were excluded due to incomplete data collection, which showed gaps of more than one day between entries. There is a notable drop in accident records occurred during the COVID-19 pandemic (early 2020 to mid-2021), though no method was found to correct this anomaly.

- Main findings from probability tasks: Crash frequency varies strongly with time and traffic conditions, while severe crashes remain rare. Late-hour crashes show an increased conditional severity risk, whereas rush-hour crashes are mostly non-severe. Weather effects, especially poor visibility, have limited impact on severity due to low occurrence.

- Main findings from LLN and CLT: The cumulative mean of severe crashes stabilizes with increasing sample size, confirming the Law of Large Numbers. Sampling distributions of the mean become narrower and more symmetric as sample size grows, approaching a normal shape, illustrating the Central Limit Theorem. Small samples show high variability and deviations due to rare events and the binary nature of the data.

- Main findings from regression: While the model learned some patterns and manages to predict the mean value of injuries relatively well, it is not able to predict short time fluctuations caused by other factors like weather condition. To improve the prediction more features could be used in the training.

## 10 Reproducibility Notes

- Exact dataset source link and version or download date:
  - Dataset: *Traffic Accidents*
  - Source: `https://www.kaggle.com/datasets/oktayrdeki/traffic-accidents`
  - Downloaded on: *22.01.2026*
- Python version:
  - Python 3.13.5
- Key libraries used and versions:
  - numpy 2.1.3
  - pandas 2.2.3
  - matplotlib 3.10.0
  - seaborn 0.13.2
  - xgboost 3.1.3
  - scikit 1.8.0