Data Analysis
Final Assignment

Team Circuit Synergy

Lorenz Buchinger & Jeremia Baumgartner & Tim Zwölfer

Dataset Description

- Traffic Accidents (Kaggle)
- Number of observations (rows): **209306**
- First entry: **1st of January 2016**
- Last entry: **31st of December 2025**
- Number of features (columns): **24**

Relevant Features

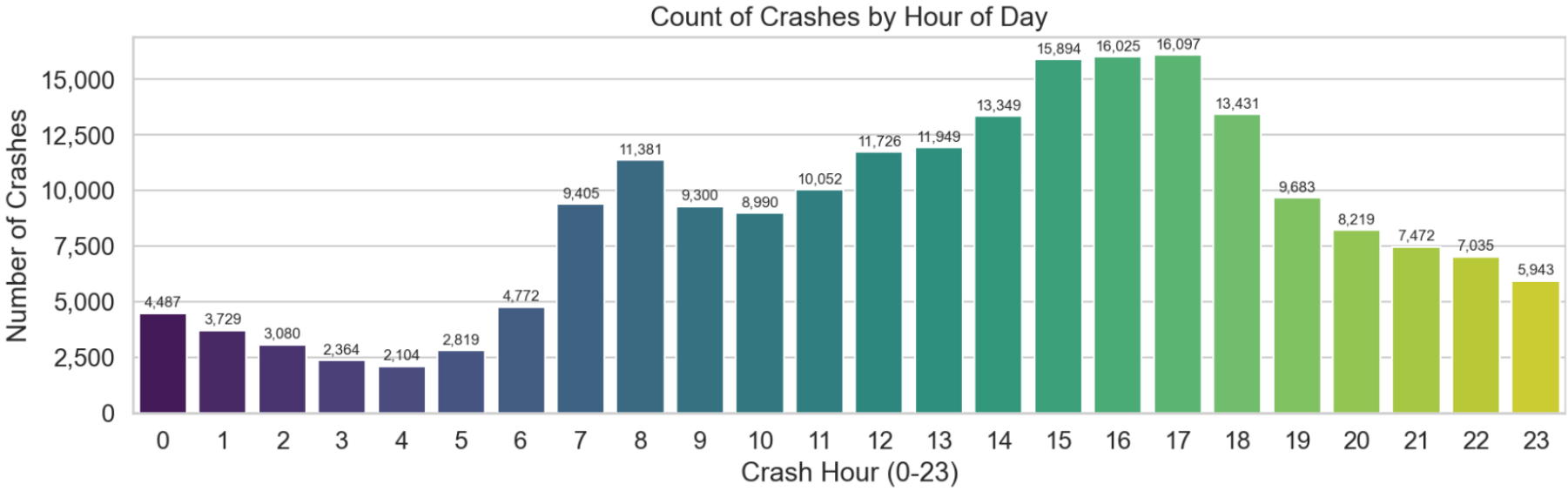
- crash_date
- weather_condition
- num_units
- Injuries
  - Total
  - Fatal
  - Incapacitating
  - Non Incapacitating
  - Reported but not visibly evident
  - No indication of injury

Basic Analysis / Descriptive stats

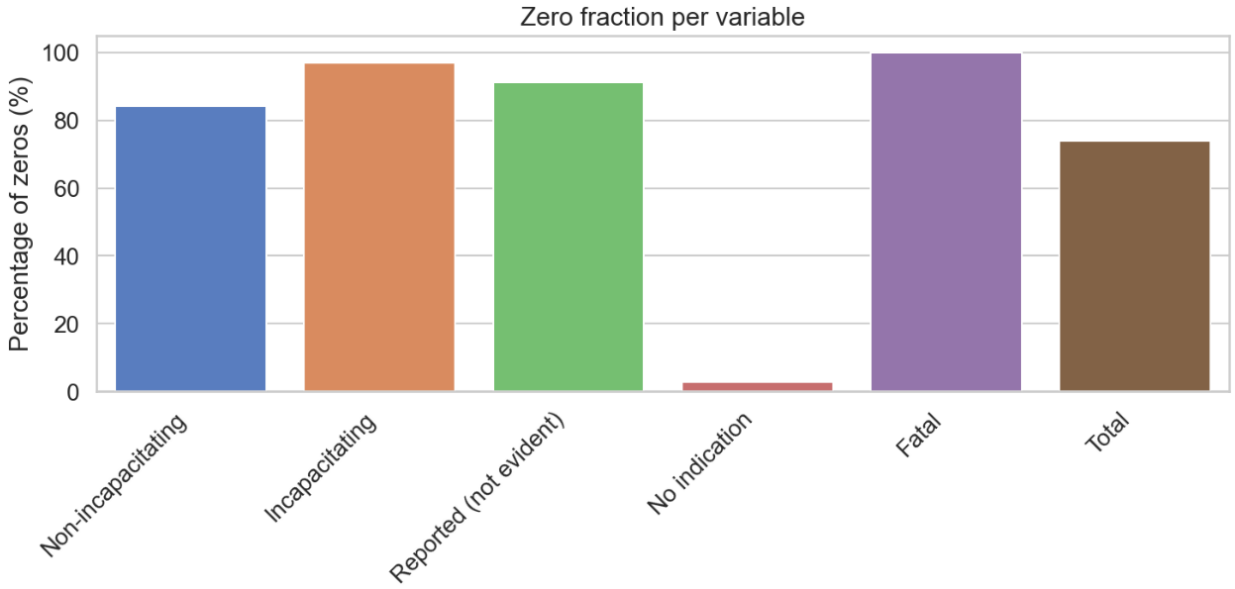
	mean	std	min	25%	50%	75%	max
num_units	2.06	0.40	1.00	2.00	2.00	2.00	11.00
injuries_total	0.38	0.80	0.00	0.00	0.00	1.00	21.00
injuries_fatal	0.00	0.05	0.00	0.00	0.00	0.00	3.00
injuries_incapacitating	0.04	0.23	0.00	0.00	0.00	0.00	7.00
injuries_non_incapacitating	0.22	0.61	0.00	0.00	0.00	0.00	21.00
injuries_reported_not_evident	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

FH JOANNEUM
University of Applied Sciences

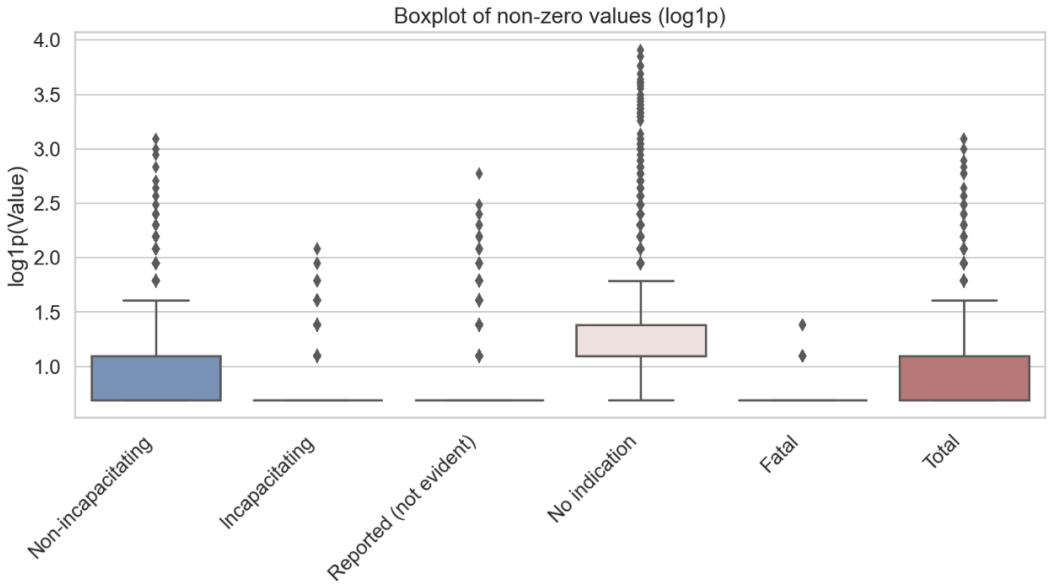
Basic Analysis / Grouped Summaries



Preprocessing / Outliers



Preprocessing / Outliers

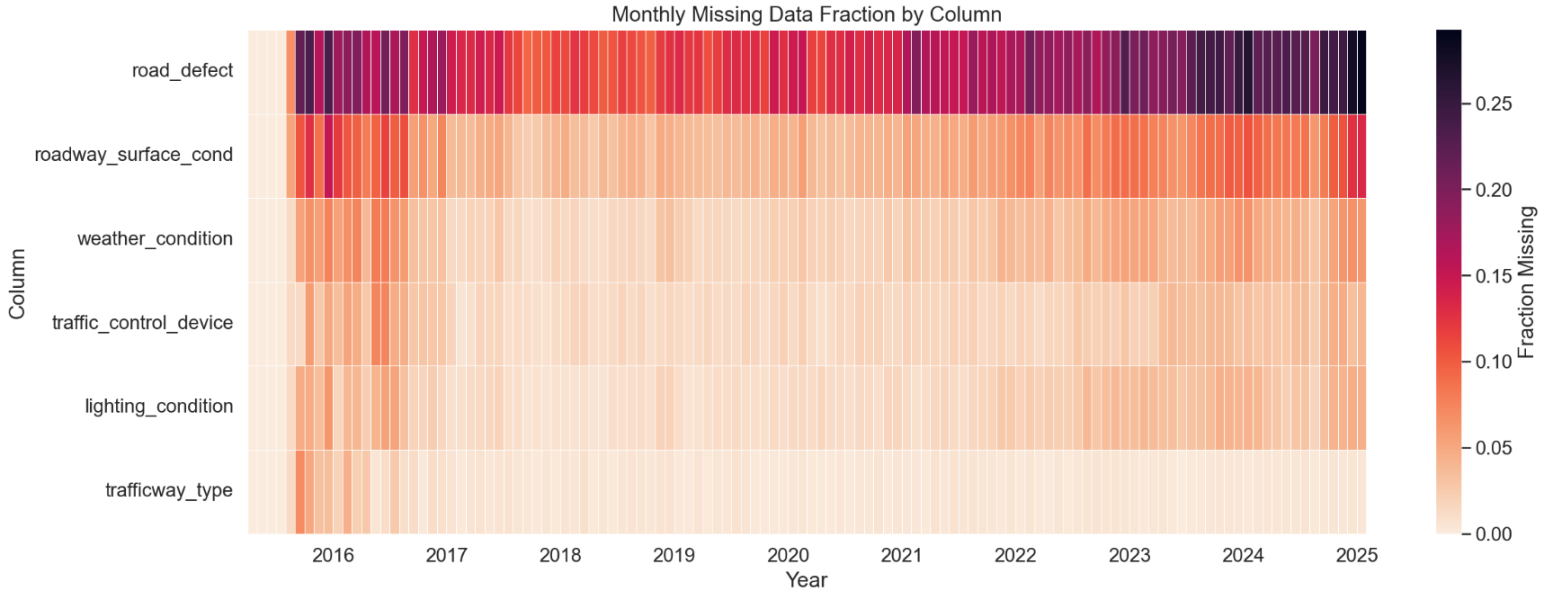


Preprocessing / Missing Values

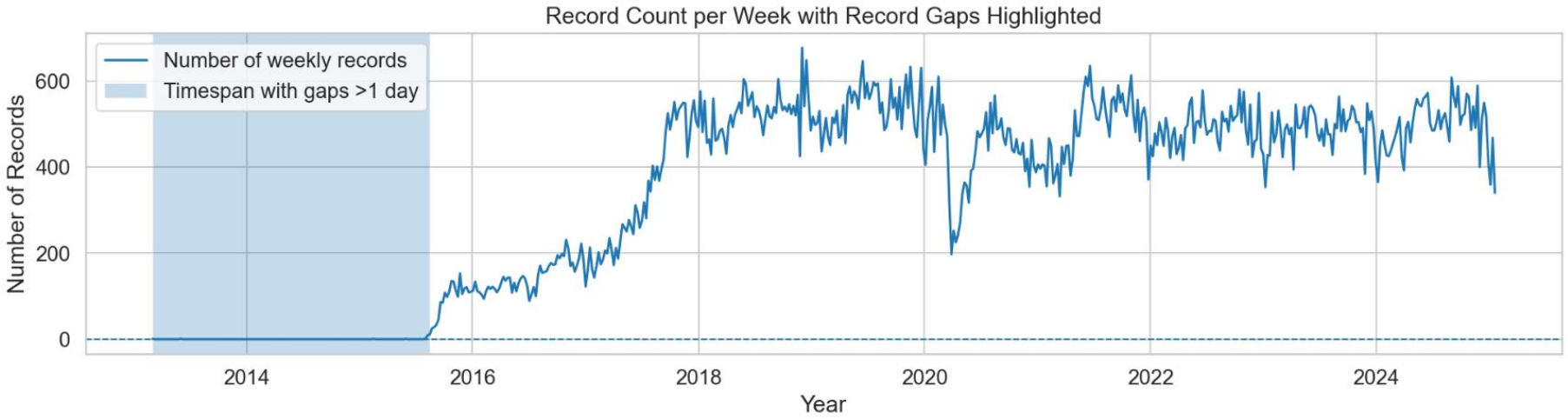
- Encoded with “UNKNOWN”
- 1.26% (63320) of values missing
- -> Replaced with nan

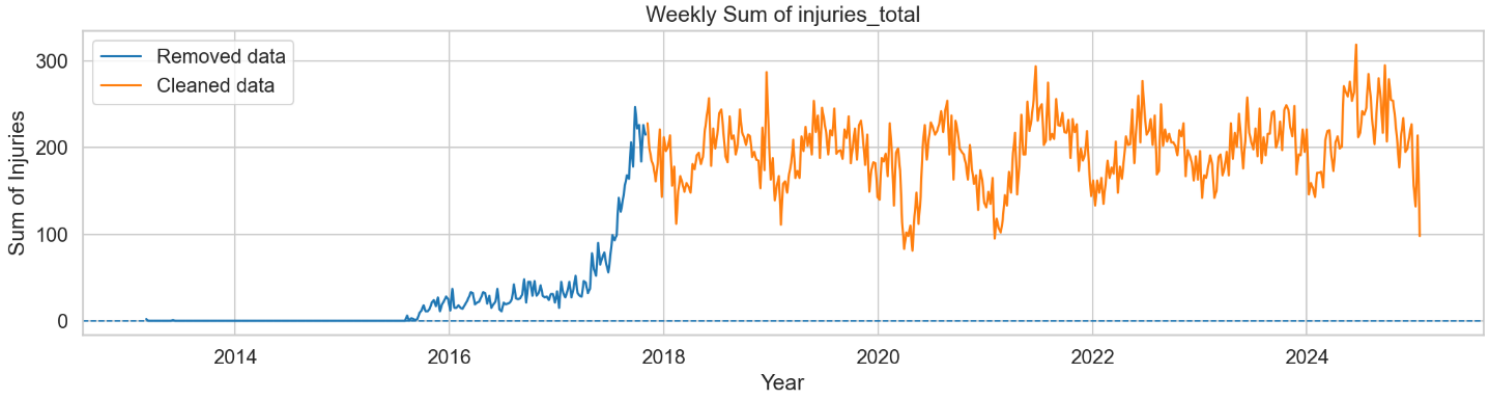
Variable	Count	Percentage
road_defect	34426	16.45
roadway_surface_cond	12509	5.98
weather_condition	6534	3.12
traffic_control_device	4455	2.13
lighting_condition	4336	2.07
trafficway_type	1060	0.51

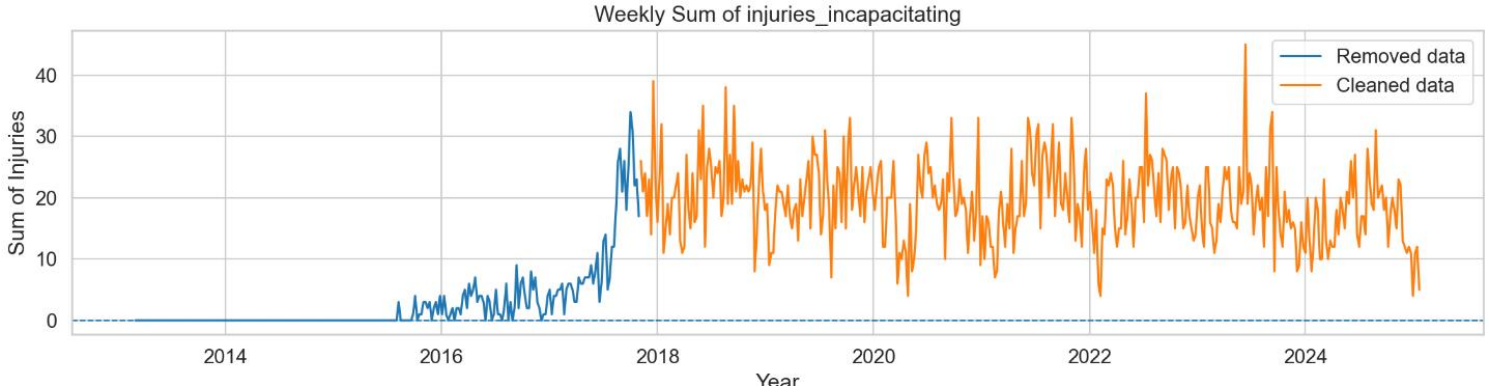
Preprocessing / Missing Values



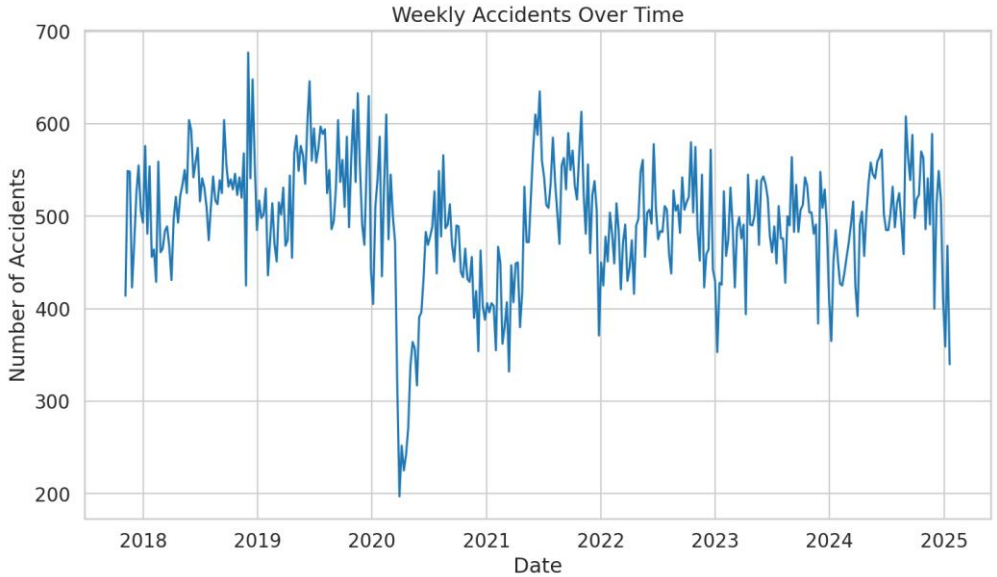
Preprocessing / Missing Values




Weekly Sum of injuries_total
Removed data
Cleaned data
Sum of Injuries
300
200
100
0
2014
2016
2018
2020
2022
2024
Year

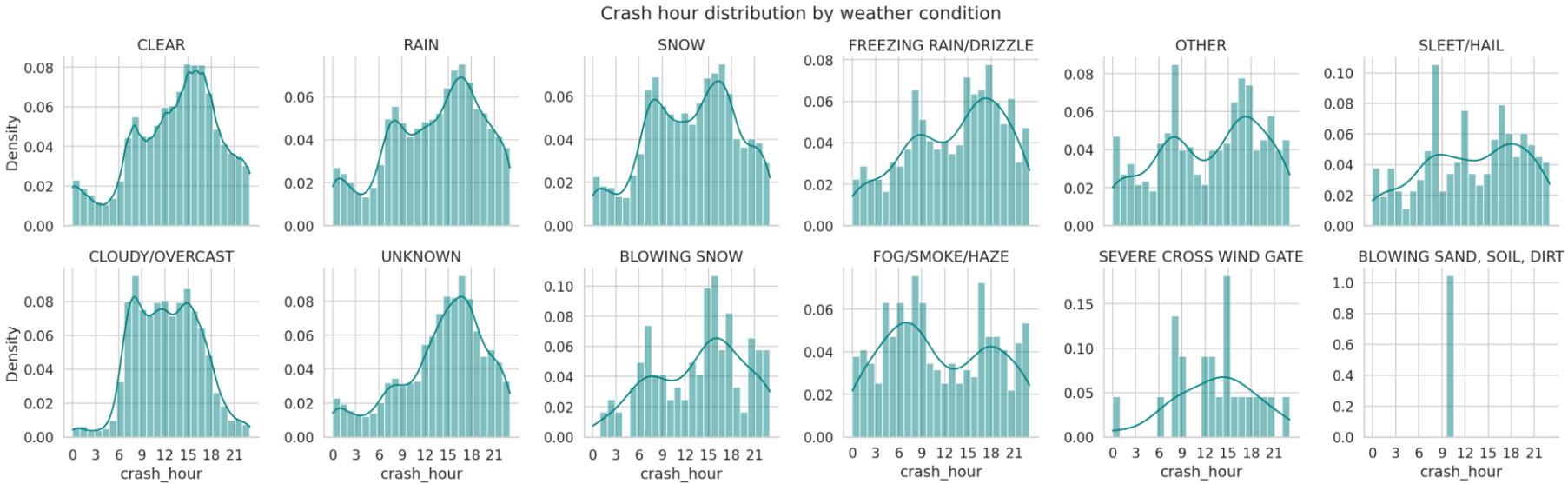

Weekly Sum of injuries_incapacitating
Removed data
Cleaned data
Sum of Injuries
40
30
20
10
0
2014
2016
2018
2020
2022
2024
Year

Visualization and Exploratory Analysis: Timeseries



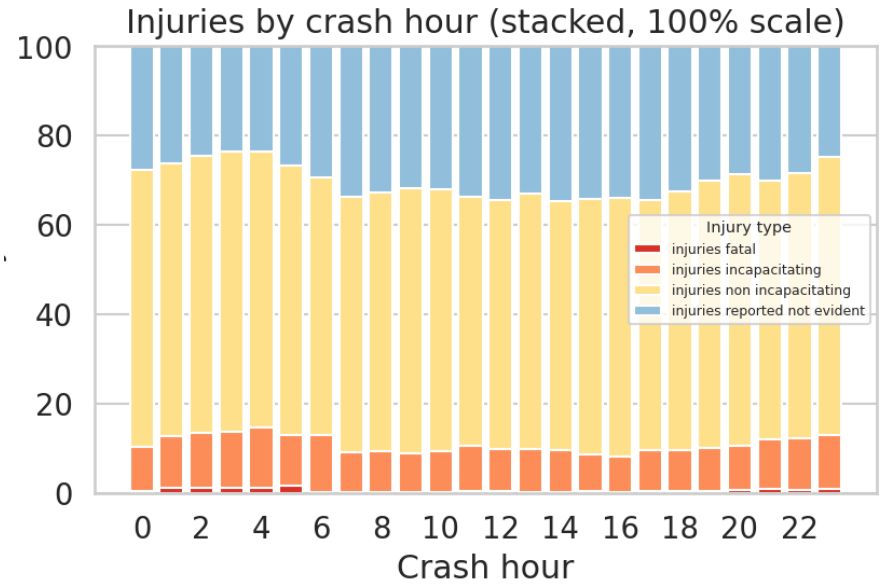
FH JOANNEUM
University of Applied Sciences

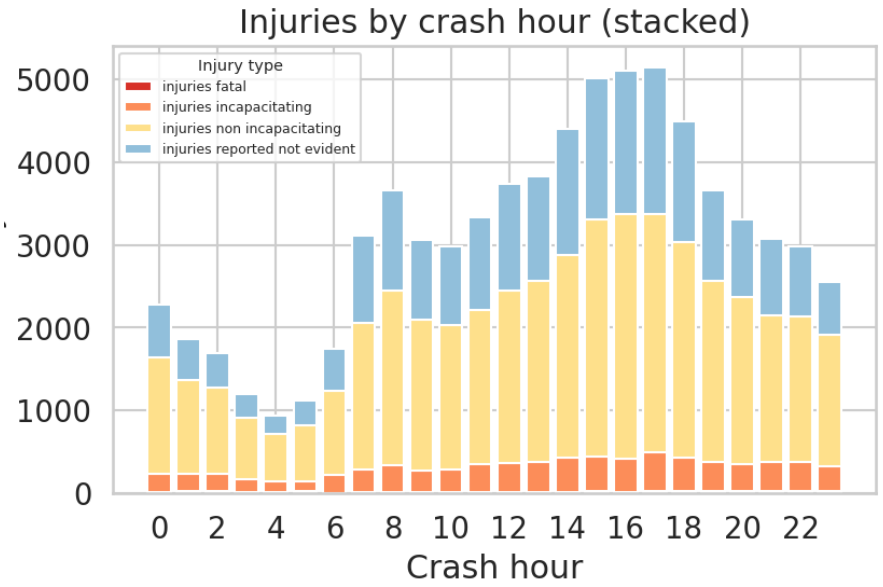
Visualization and Exploratory Analysis: Distribution analysis



FH | JOANNEUM
University of Applied Sciences

Visualization and Exploratory Analysis: Distribution analysis

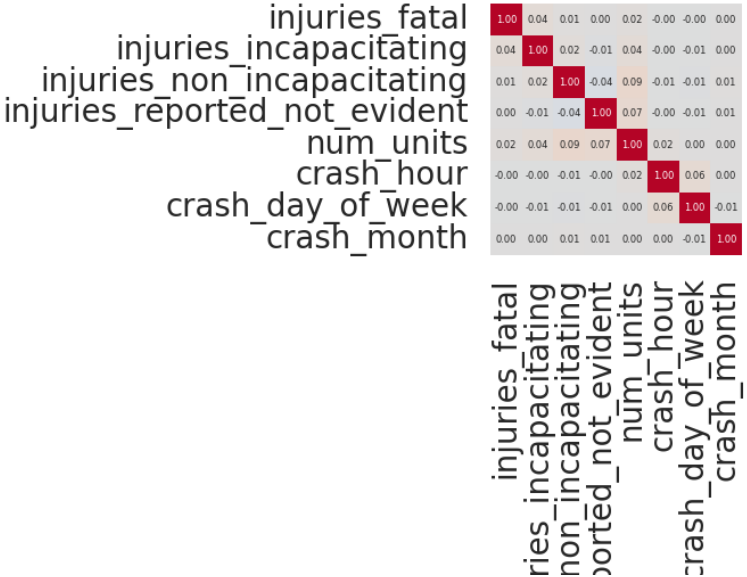




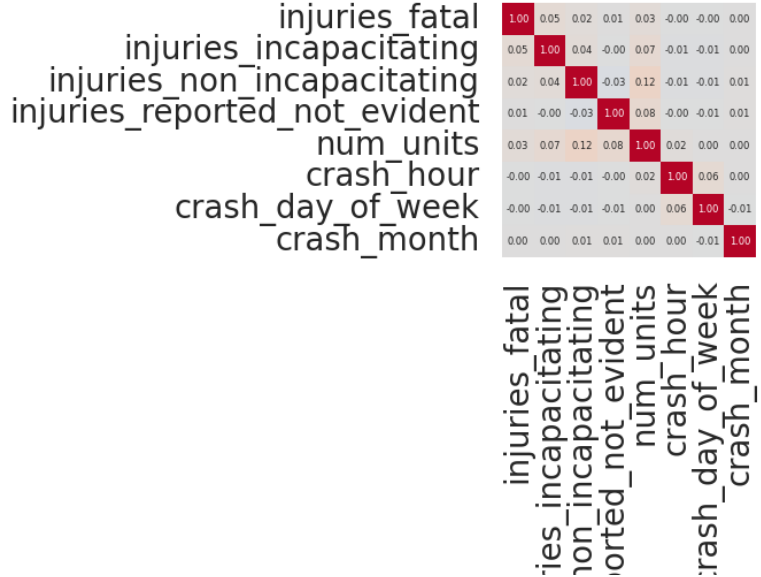
FH JOANNEUM
University of Applied Sciences

Visualization and Exploratory Analysis: Correlation

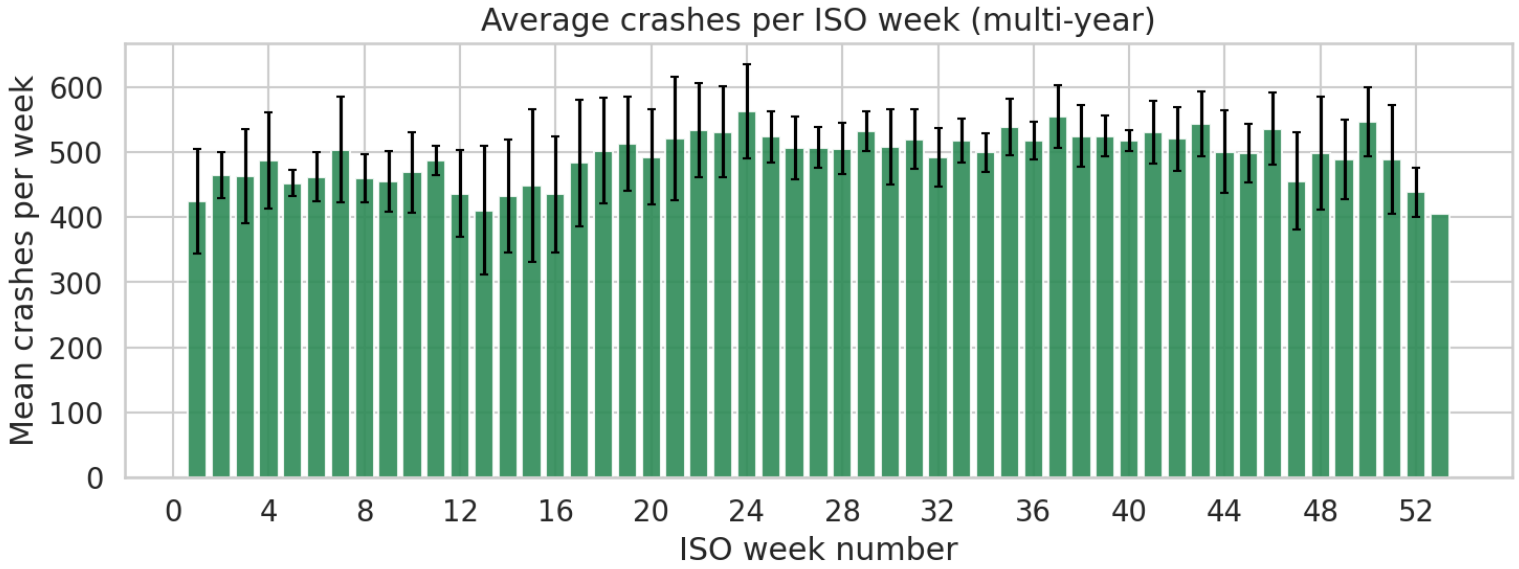
Spearman correlation (rank)



Pearson correlation (linear)

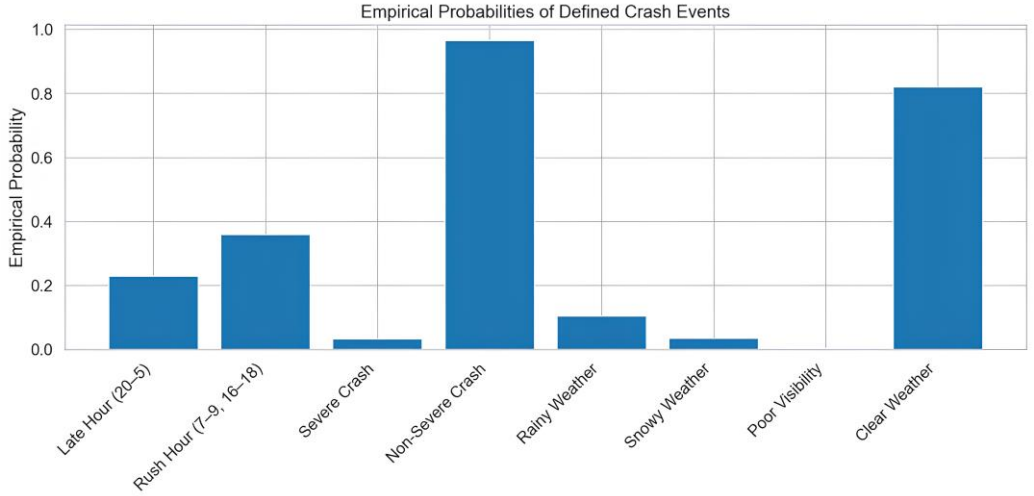


Visualization and Exploratory Analysis: Correlation

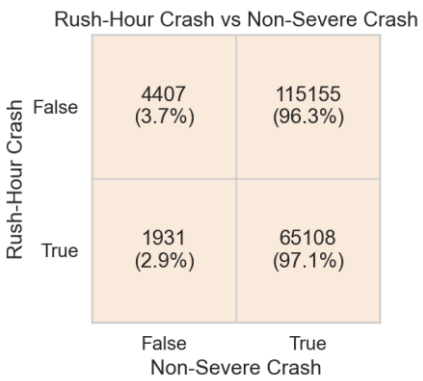


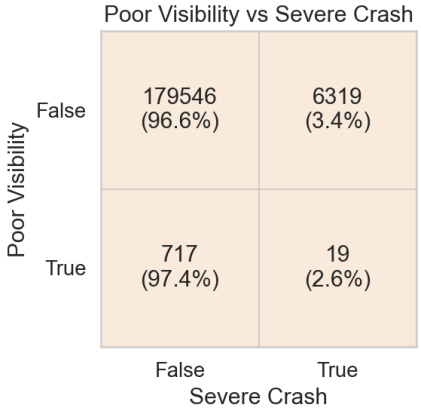
Probability Analysis: Probability Estimation

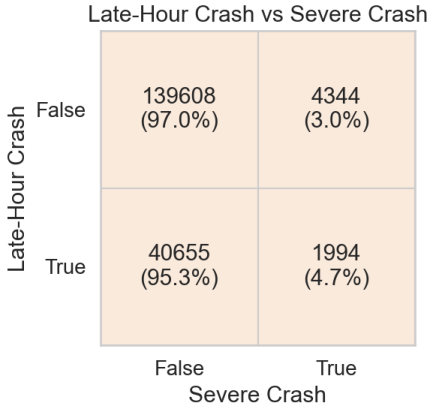
- Late Hour Crash
- Rush-Hour Crash
- Severe Crash
- Clear Weather Crash

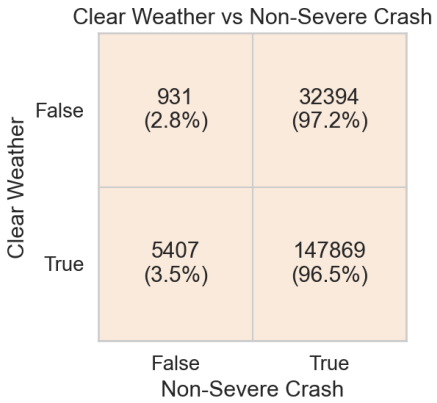


Probability Analysis: Cross-tabulation analysis





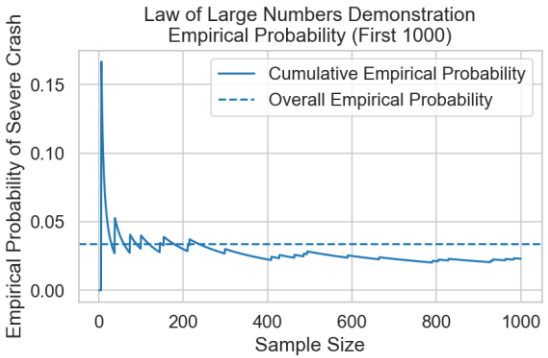


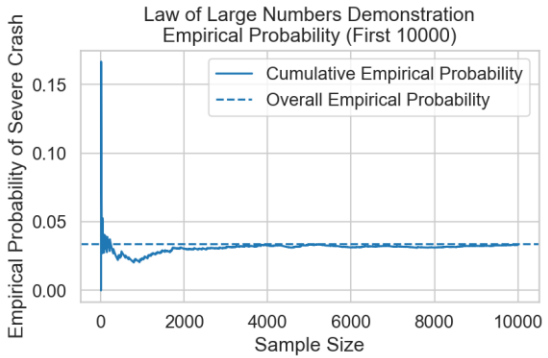


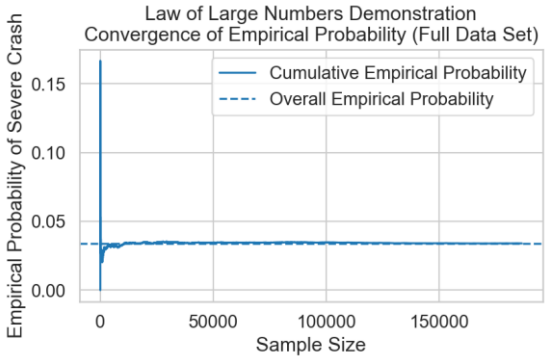
Probability Analysis: Conditional probability analysis

P(Severe Crash)	3,397%
P(Late-Hour Crash)	22,856%
P(Severe Crash | Late-Hour Crash)	31,461 %
P(Late-Hour Crash | Severe Crash)	31.461%

Statistical Theory Applications: LLN

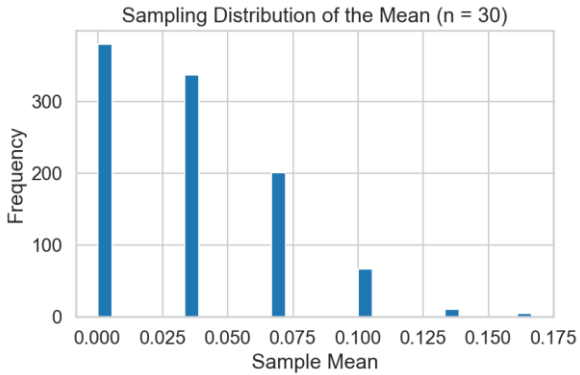


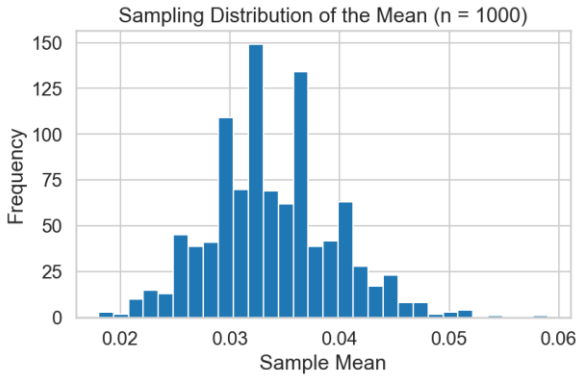


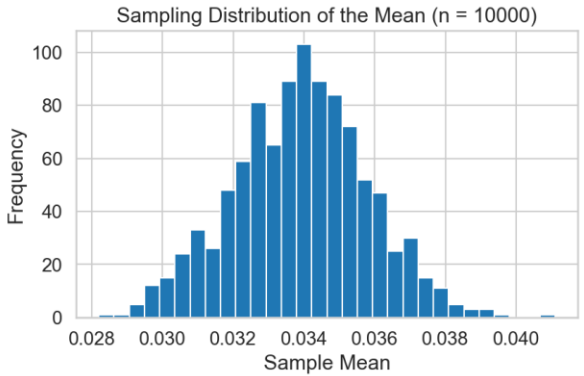


Statistical Theory Applications: CLT

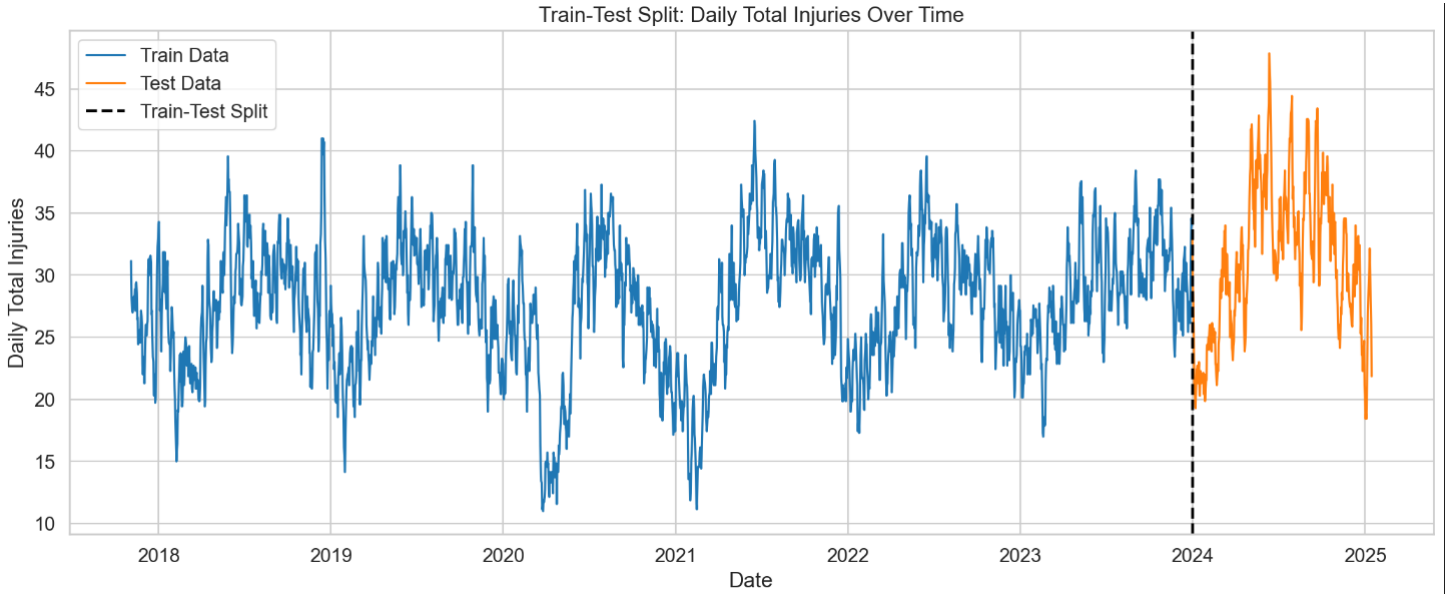
Severe Crash



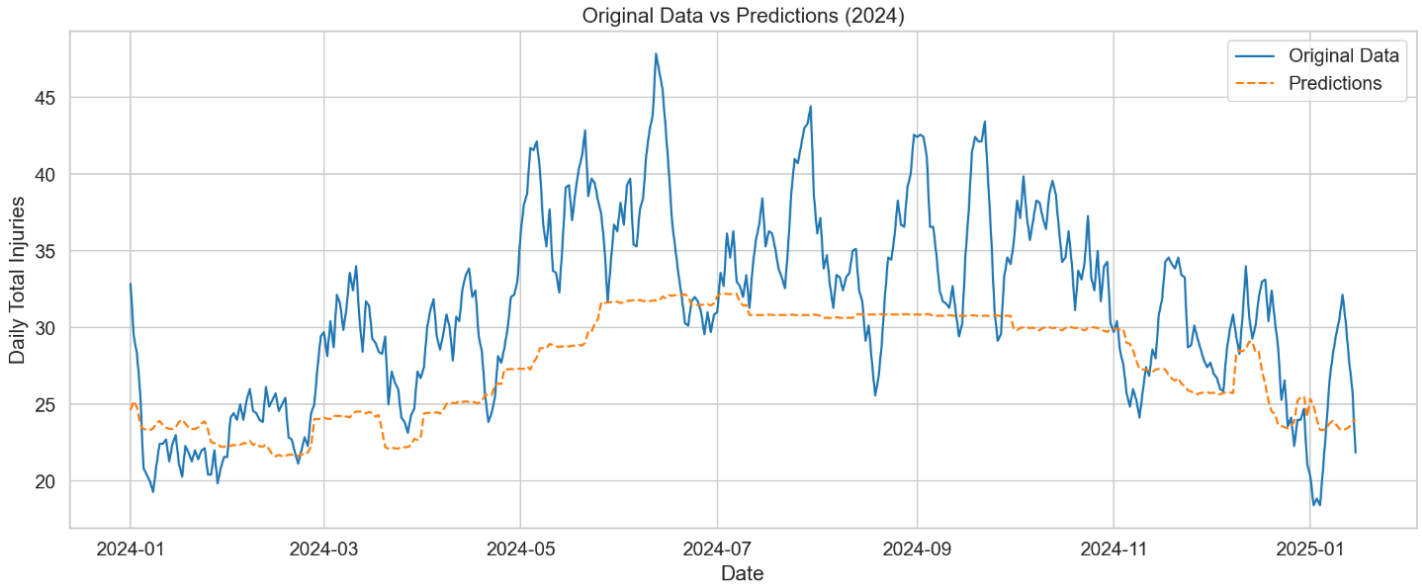




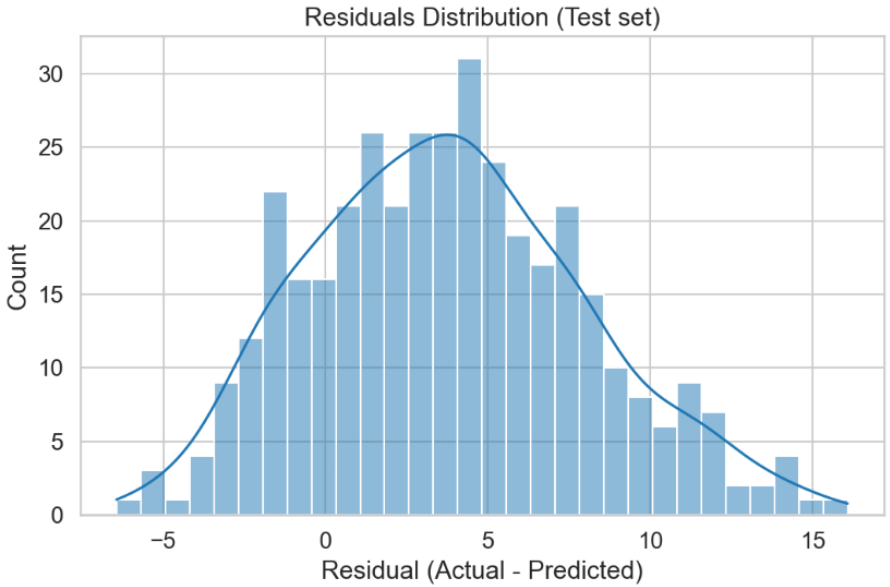
Regression Analysis

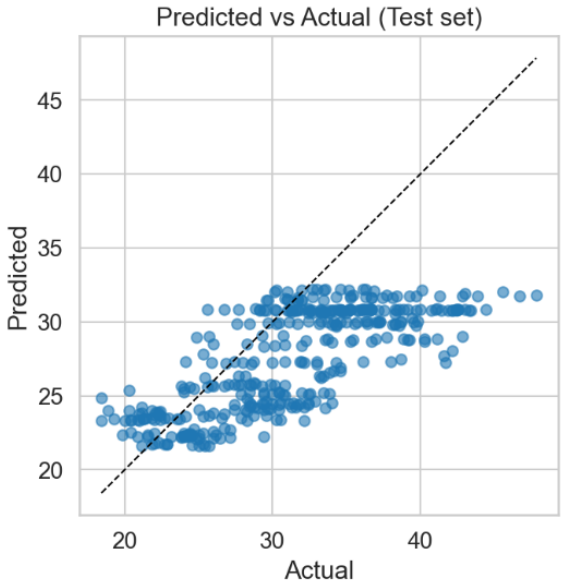


Regression Analysis



Regression Analysis / Residual Analysis

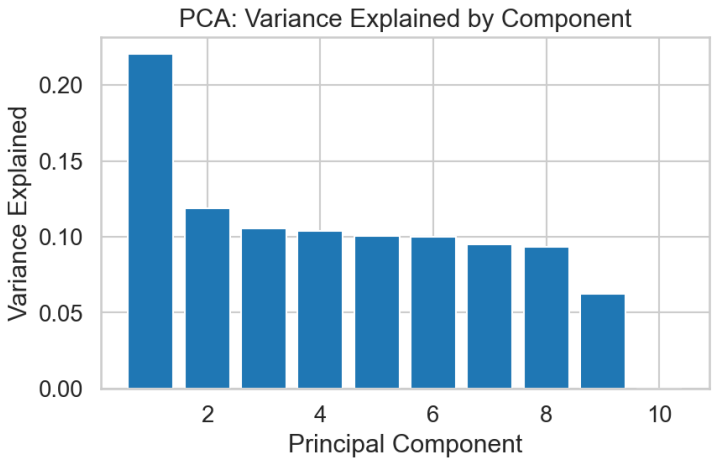


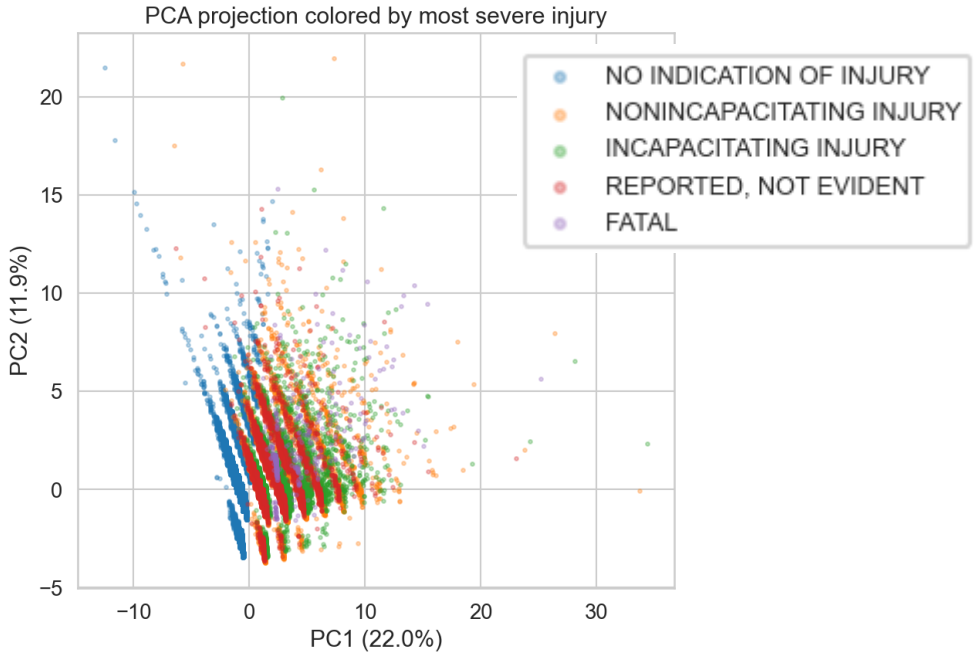


Regression Analysis / Parameters

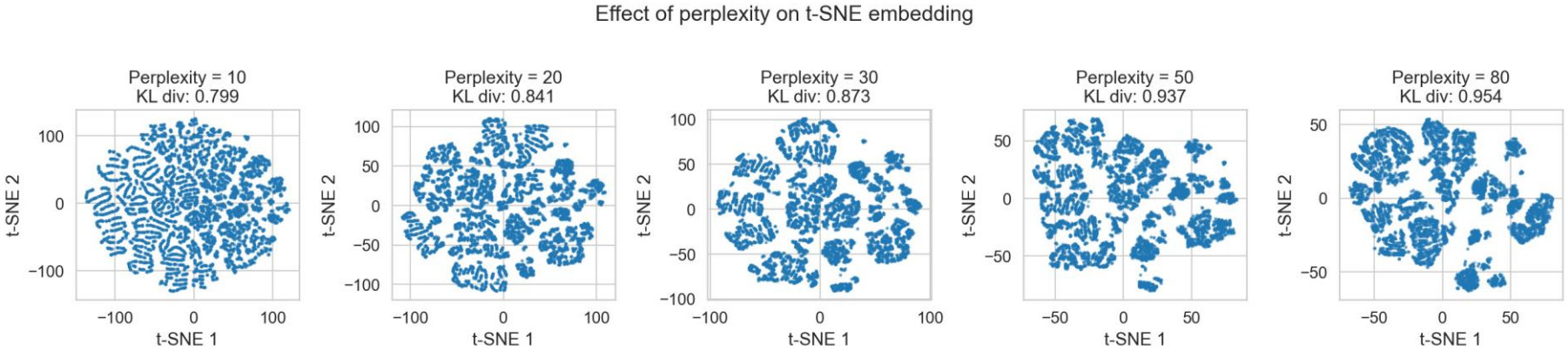
Metric	Value
Test RMSE	5.662
Test MAE	4.533
Test R^2	0.131
Test MAPE	13.57%

Dimensionality Reduction: PCA

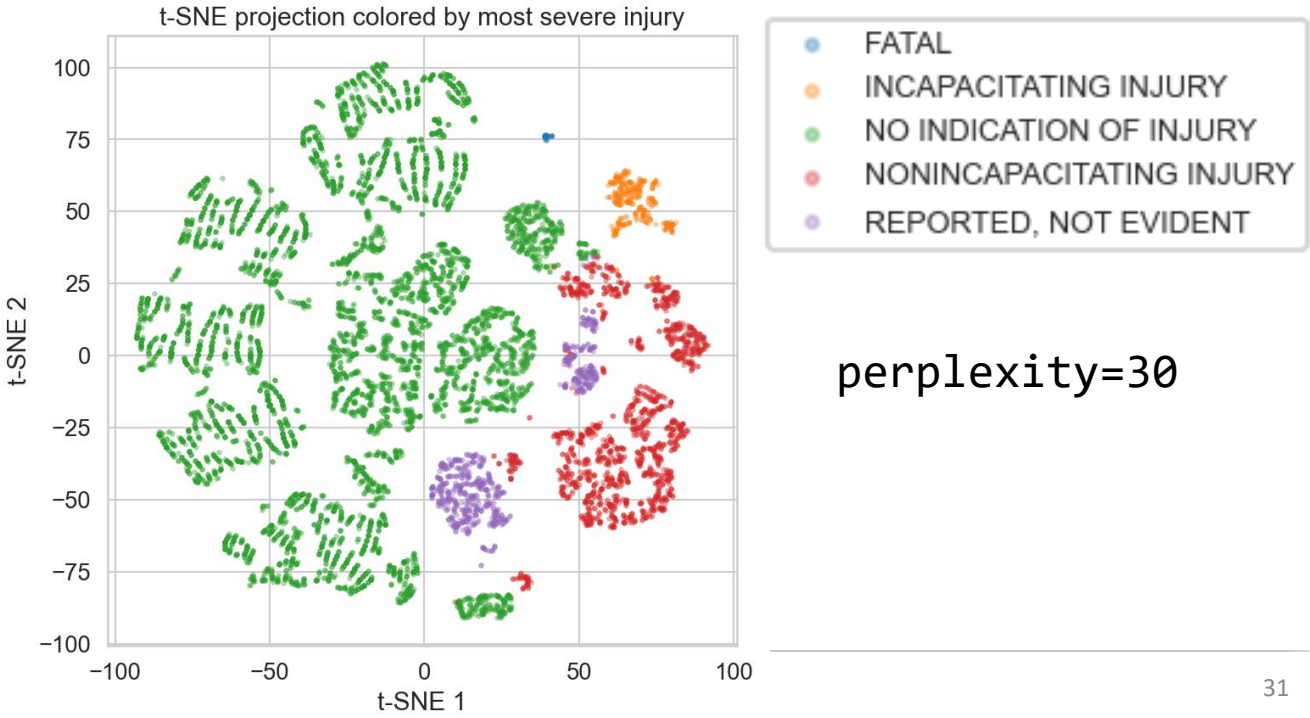




Dimensionality Reduction: t-SNE

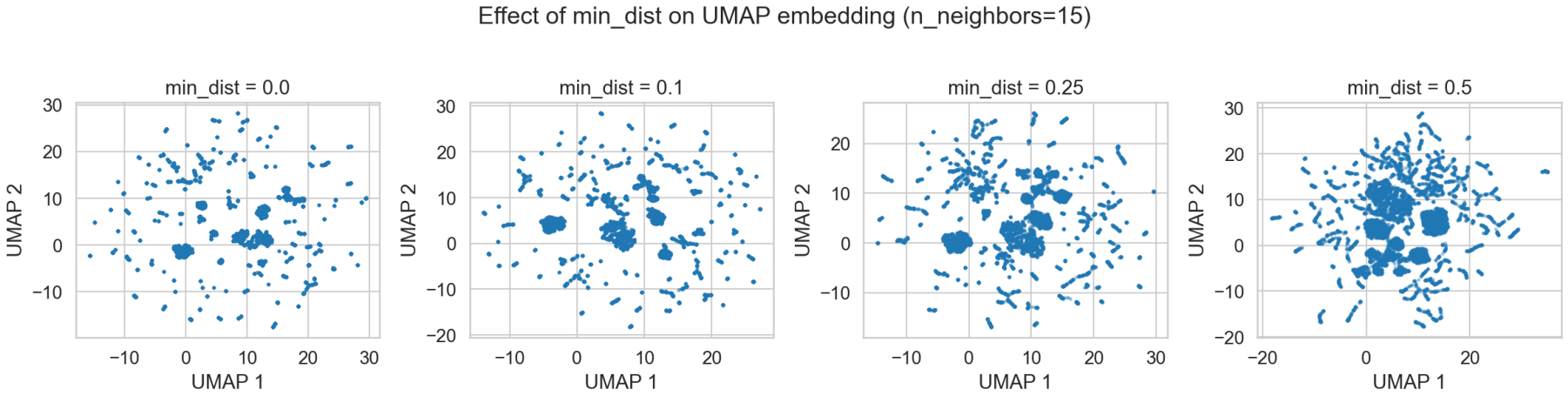


Dimensionality Reduction: t-SNE

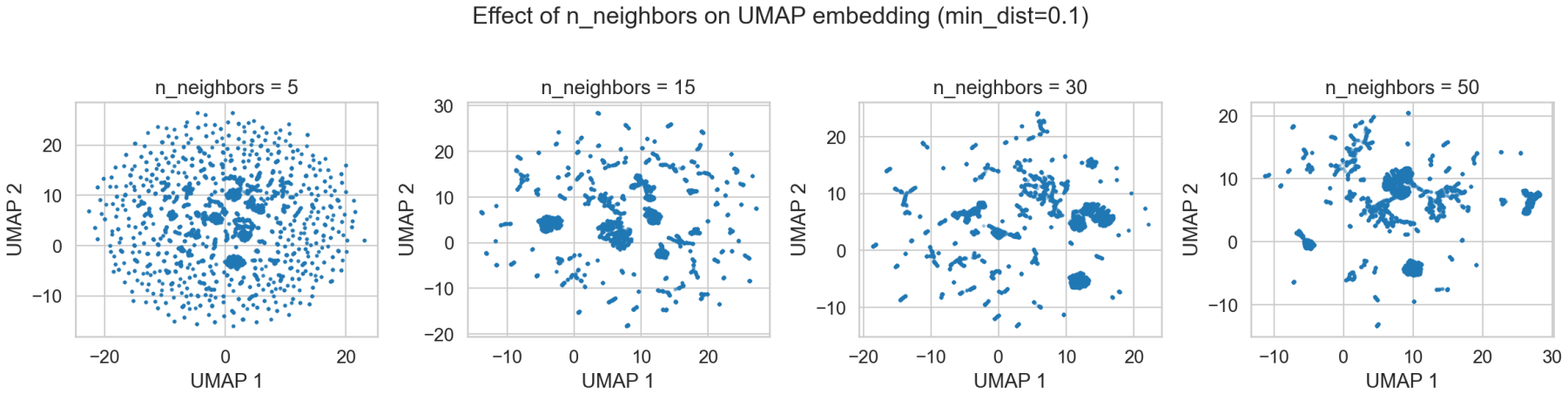


perplexity=30

Dimensionality Reduction: UMAP



Dimensionality Reduction: UMAP



Dimensionality Reduction: UMAP

```
n_neighbors = 15
   min_dist = 0.1
          n = 1000
```