

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**

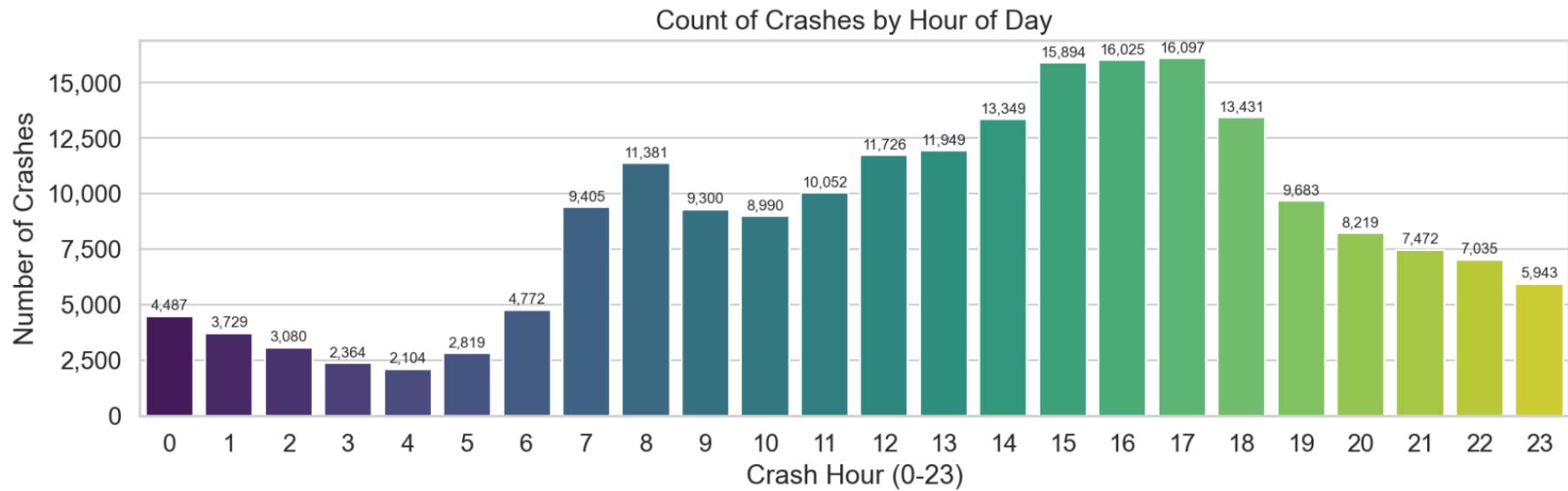
Relevant Features

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

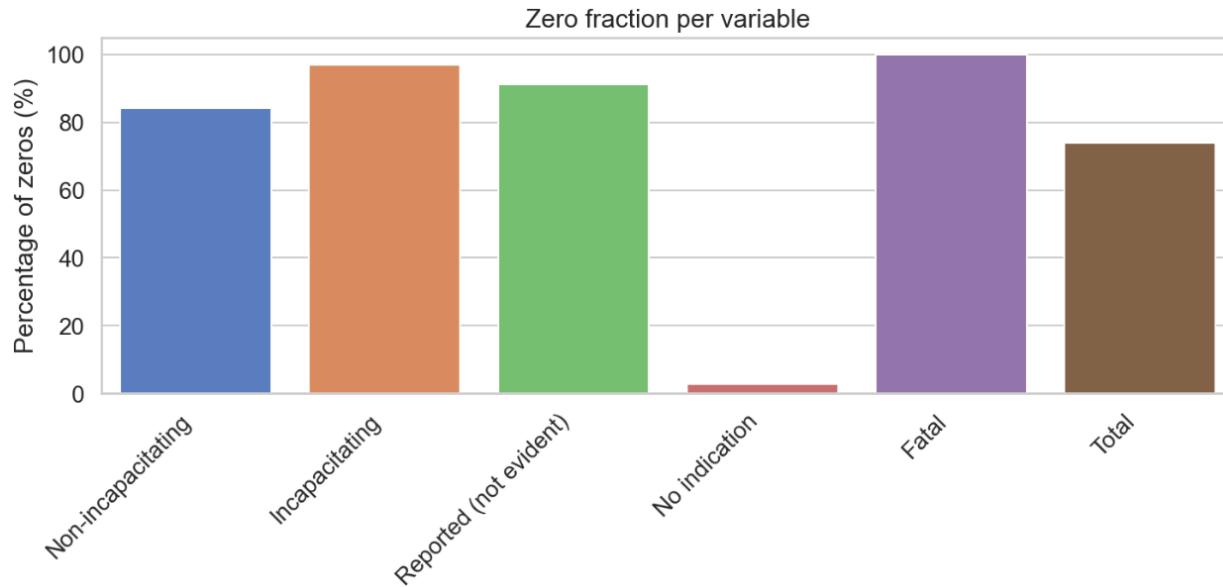
Basic Analysis / Descriptive stats

	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

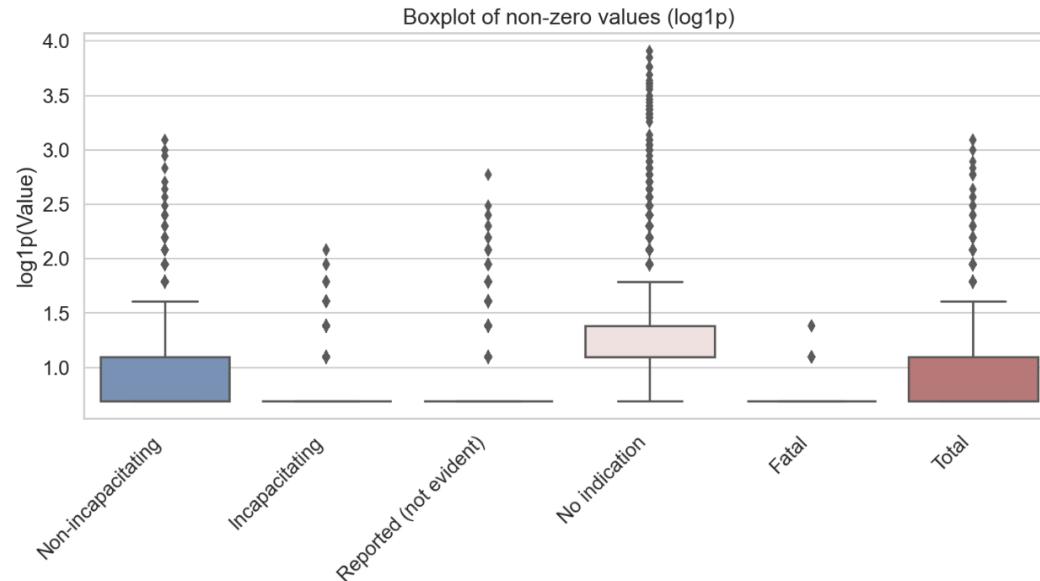
Basic Analysis / Grouped Summaries



Preprocessing / Outliers



Preprocessing / Outliers

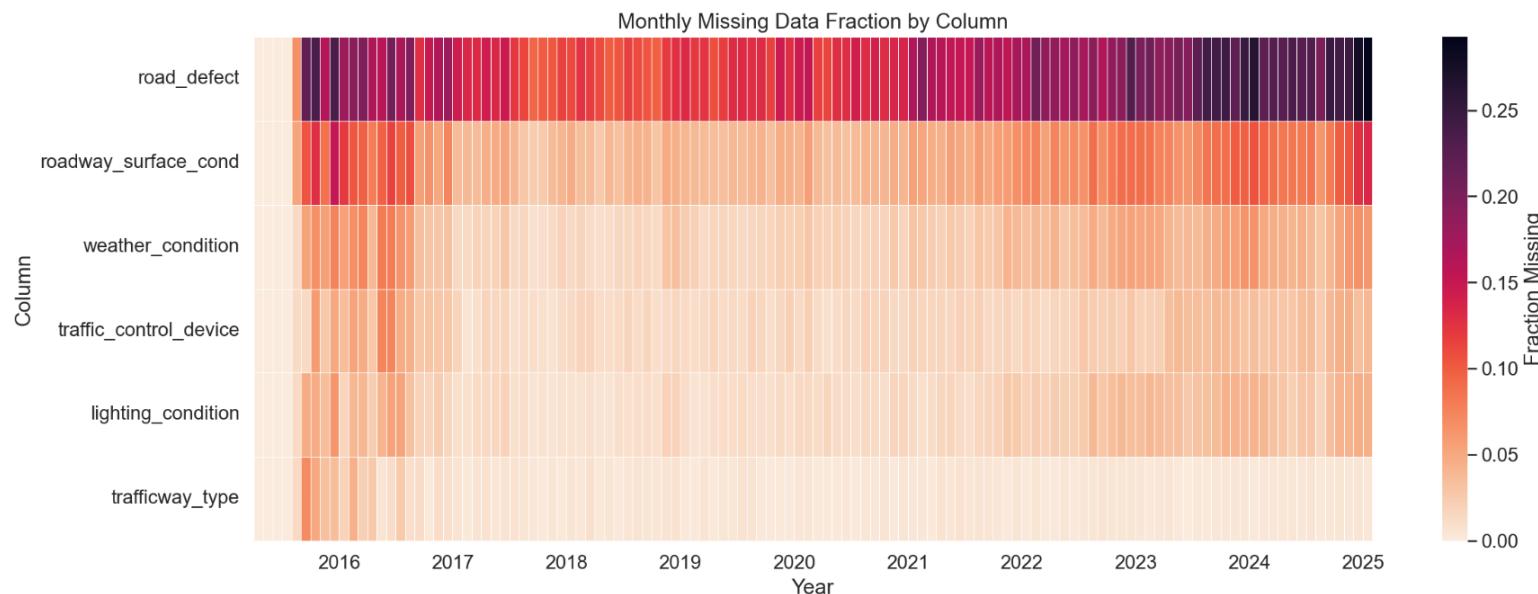


Preprocessing / Missing Values

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

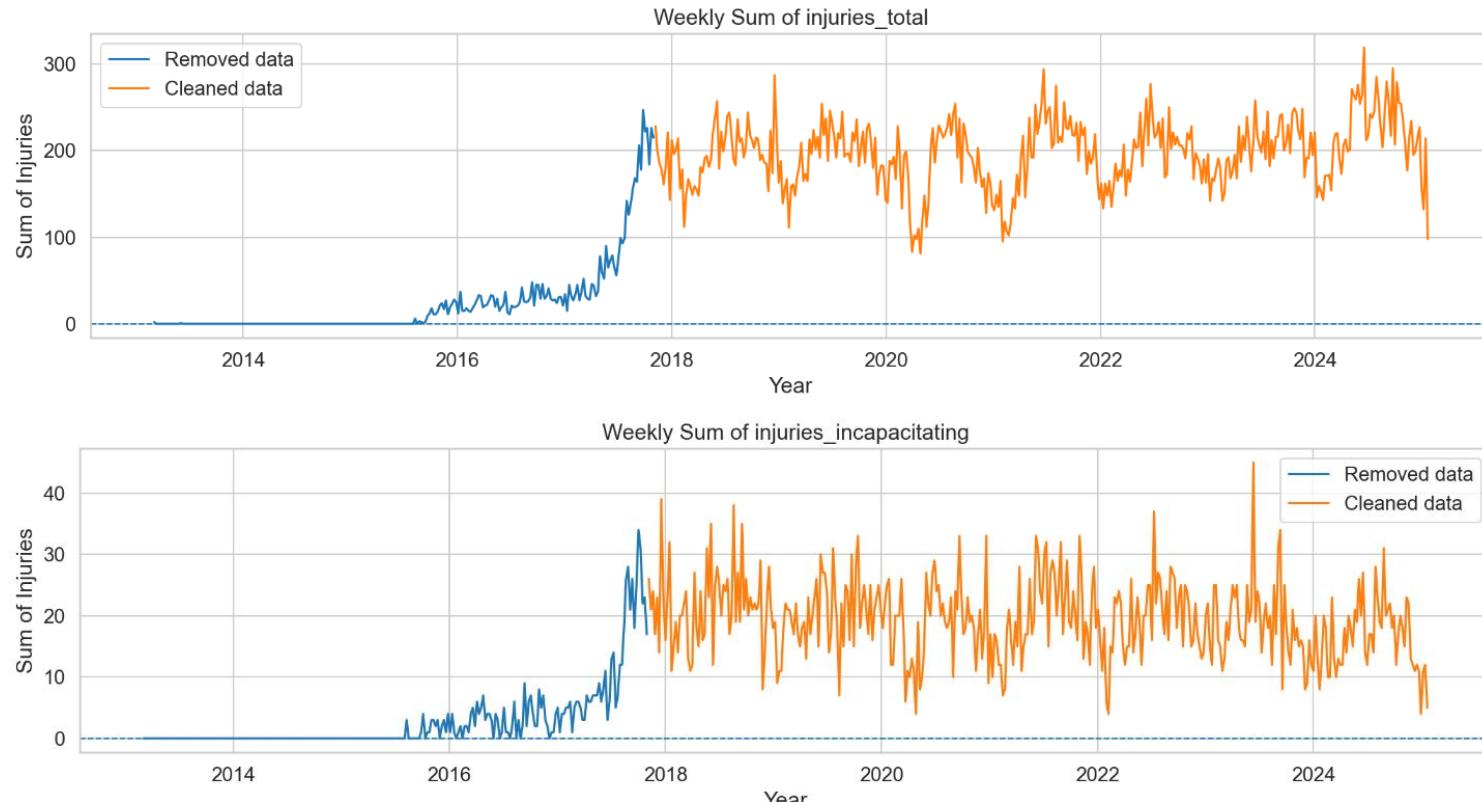
Variable	Count	Percentage
road_defect	34426	16.45
roadway_surface_cond	12509	5.98
weather_condition	6534	3.12
traffic_control_device	4455	2.13
lighting_condition	4336	2.07
trafficway_type	1060	0.51

Preprocessing / Missing Values

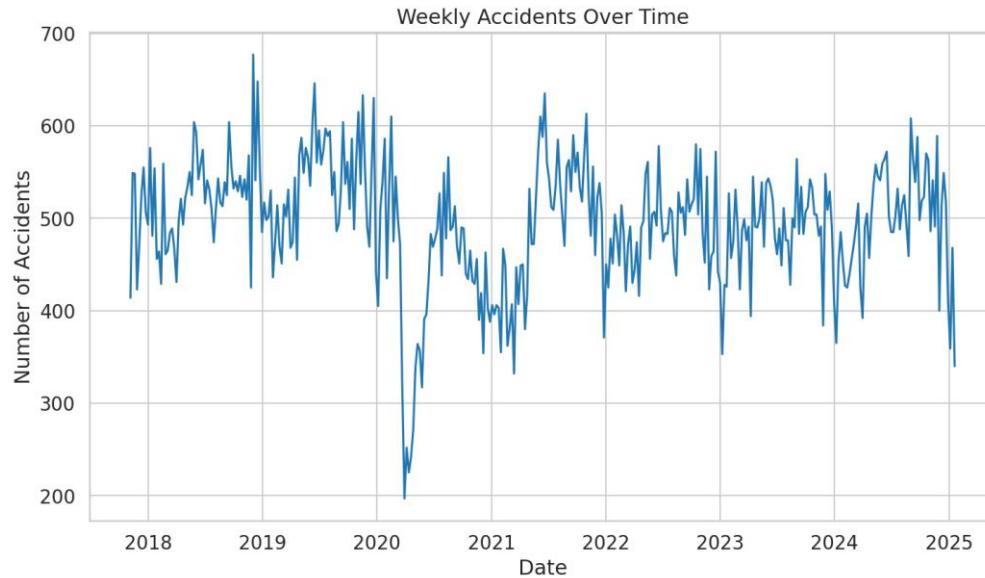


Preprocessing / Missing Values

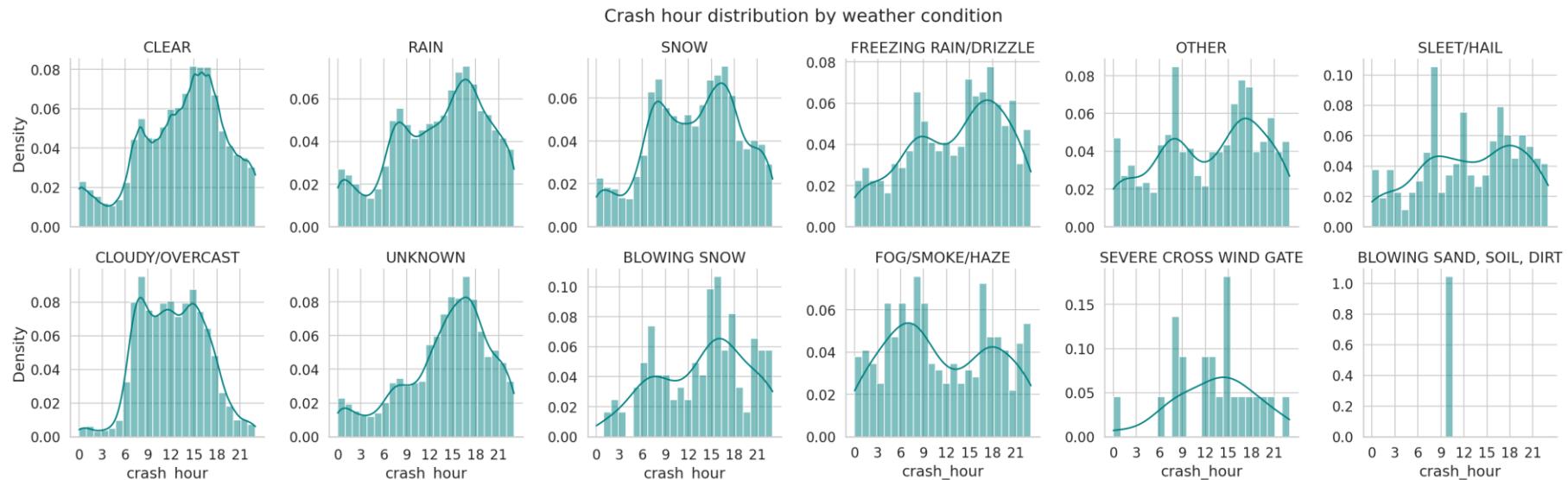




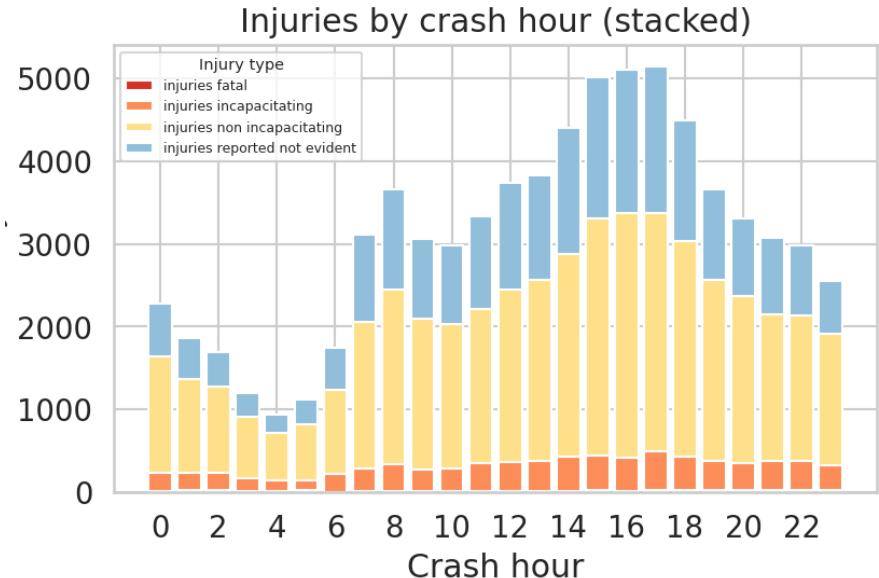
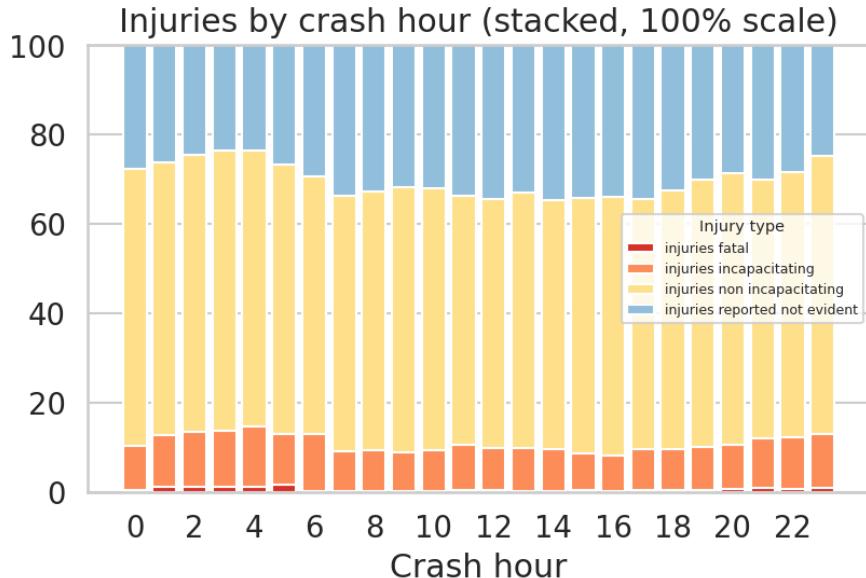
Visualization and Exploratory Analysis: Timeseries



Visualization and Exploratory Analysis: Distribution analysis



Visualization and Exploratory Analysis: Distribution analysis



Visualization and Exploratory Analysis: Correlation

Spearman correlation (rank)

	injuries_fatal	injuries_incapacitating	injuries_non_incapacitating	injuries_reported_not_evident	num_units	crash_hour	crash_day_of_week	crash_month
injuries_fatal	1.00	0.04	0.01	0.00	0.02	-0.00	-0.00	0.00
injuries_incapacitating	0.04	1.00	0.02	-0.01	0.04	-0.00	-0.01	0.00
injuries_non_incapacitating	0.01	0.02	1.00	-0.04	0.09	-0.01	-0.01	0.01
injuries_reported_not_evident	0.00	-0.01	-0.04	1.00	0.07	-0.00	-0.01	0.01
num_units	0.02	0.04	0.09	0.07	1.00	0.02	0.00	0.00
crash_hour	-0.00	-0.00	-0.01	-0.00	0.02	1.00	0.06	0.00
crash_day_of_week	-0.00	-0.01	-0.01	0.00	0.06	0.00	1.00	-0.01
crash_month	0.00	0.00	0.01	0.00	0.00	-0.01	0.00	1.00

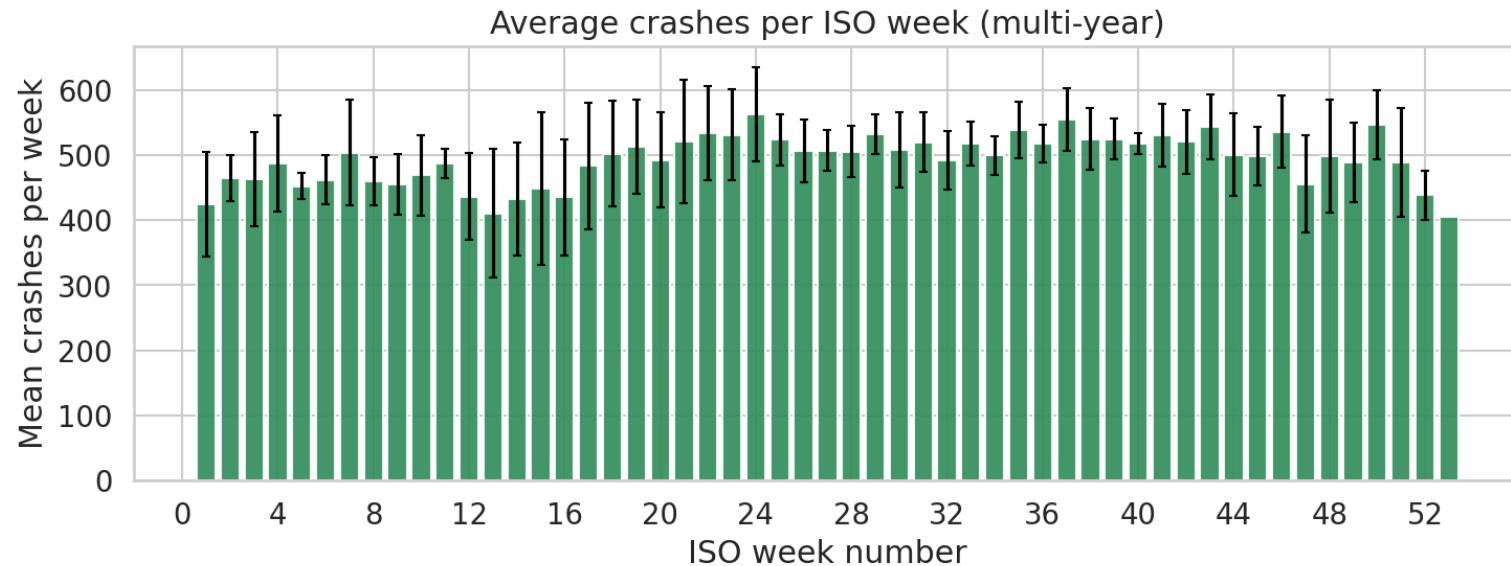
injuries_fatal
injuries_incapacitating
injuries_non_incapacitating
injuries_reported_not_evident
num_units
crash_hour
crash_day_of_week
crash_month

Pearson correlation (linear)

	injuries_fatal	injuries_incapacitating	injuries_non_incapacitating	injuries_reported_not_evident	num_units	crash_hour	crash_day_of_week	crash_month
injuries_fatal	1.00	0.05	0.02	0.01	0.03	-0.00	-0.00	0.00
injuries_incapacitating	0.05	1.00	0.04	-0.00	0.07	-0.01	-0.01	0.00
injuries_non_incapacitating	0.02	0.04	1.00	-0.03	0.12	-0.01	-0.01	0.01
injuries_reported_not_evident	0.01	-0.00	-0.03	1.00	0.08	-0.00	-0.01	0.01
num_units	0.03	0.07	0.12	0.08	1.00	0.02	0.00	0.00
crash_hour	-0.00	-0.01	-0.01	-0.00	0.02	1.00	0.06	0.00
crash_day_of_week	-0.00	-0.01	-0.01	0.00	0.06	0.00	1.00	-0.01
crash_month	0.00	0.00	0.01	0.01	0.00	0.00	-0.01	1.00

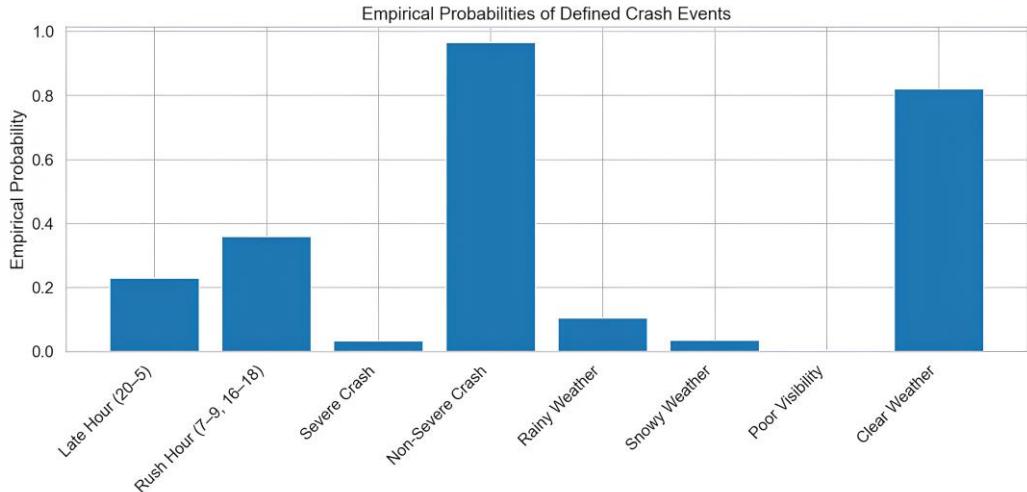
injuries_fatal
injuries_incapacitating
injuries_non_incapacitating
injuries_reported_not_evident
num_units
crash_hour
crash_day_of_week
crash_month

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

		Rush-Hour Crash vs Non-Severe Crash	
		False	True
Rush-Hour Crash	False	4407 (3.7%)	115155 (96.3%)
	True	1931 (2.9%)	65108 (97.1%)
Non-Severe Crash		False	True

		Poor Visibility vs Severe Crash	
		False	True
Poor Visibility	False	179546 (96.6%)	6319 (3.4%)
	True	717 (97.4%)	19 (2.6%)
Severe Crash		False	True

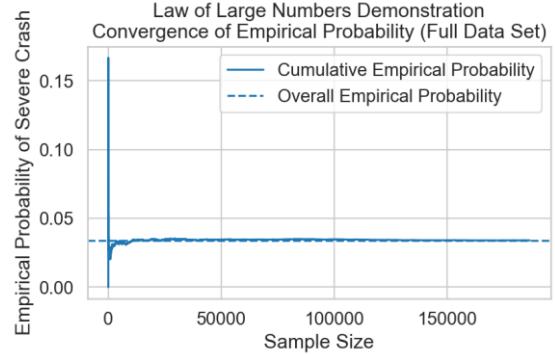
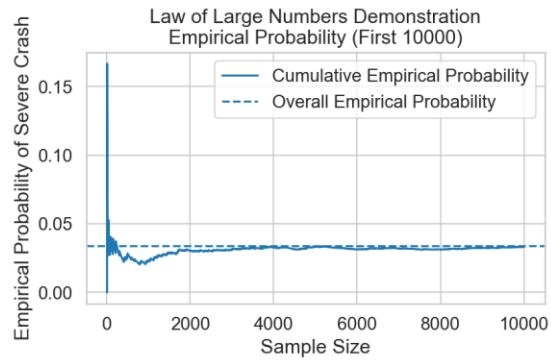
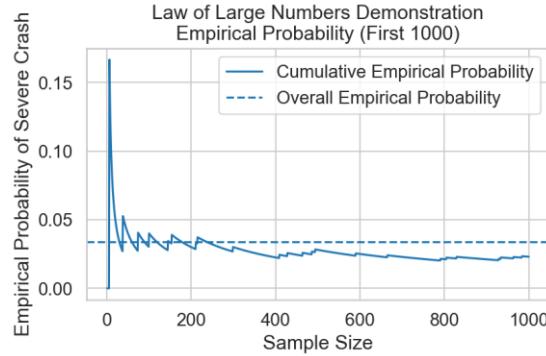
		Late-Hour Crash vs Severe Crash	
		False	True
Late-Hour Crash	False	139608 (97.0%)	4344 (3.0%)
	True	40655 (95.3%)	1994 (4.7%)
Severe Crash		False	True

		Clear Weather vs Non-Severe Crash	
		False	True
Clear Weather	False	931 (2.8%)	32394 (97.2%)
	True	5407 (3.5%)	147869 (96.5%)
Non-Severe Crash		False	True

Probability Analysis: Conditional probability analysis

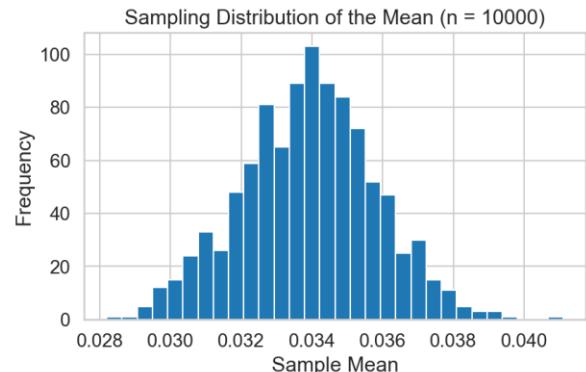
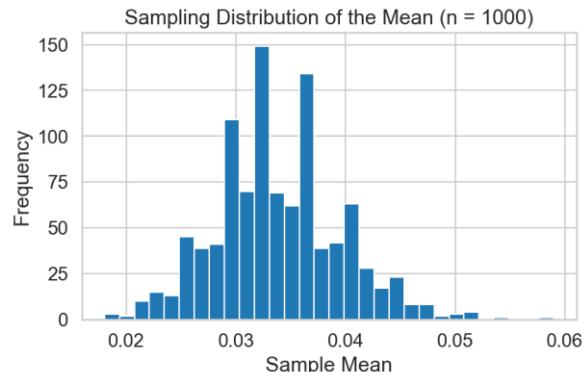
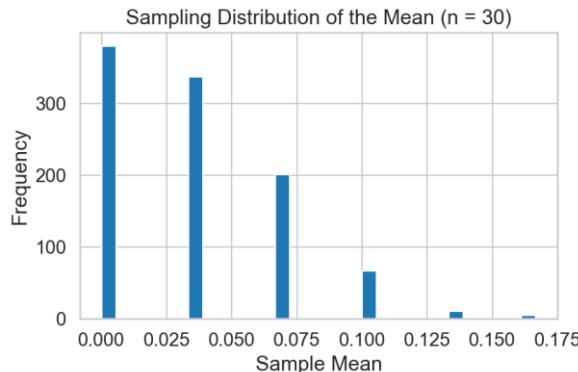
P(Severe Crash)	3,397%
P(Late-Hour Crash)	22,856%
P(Severe Crash Late-Hour Crash)	31,461 %
P(Late-Hour Crash Severe Crash)	31.461%

Statistical Theory Applications: LLN

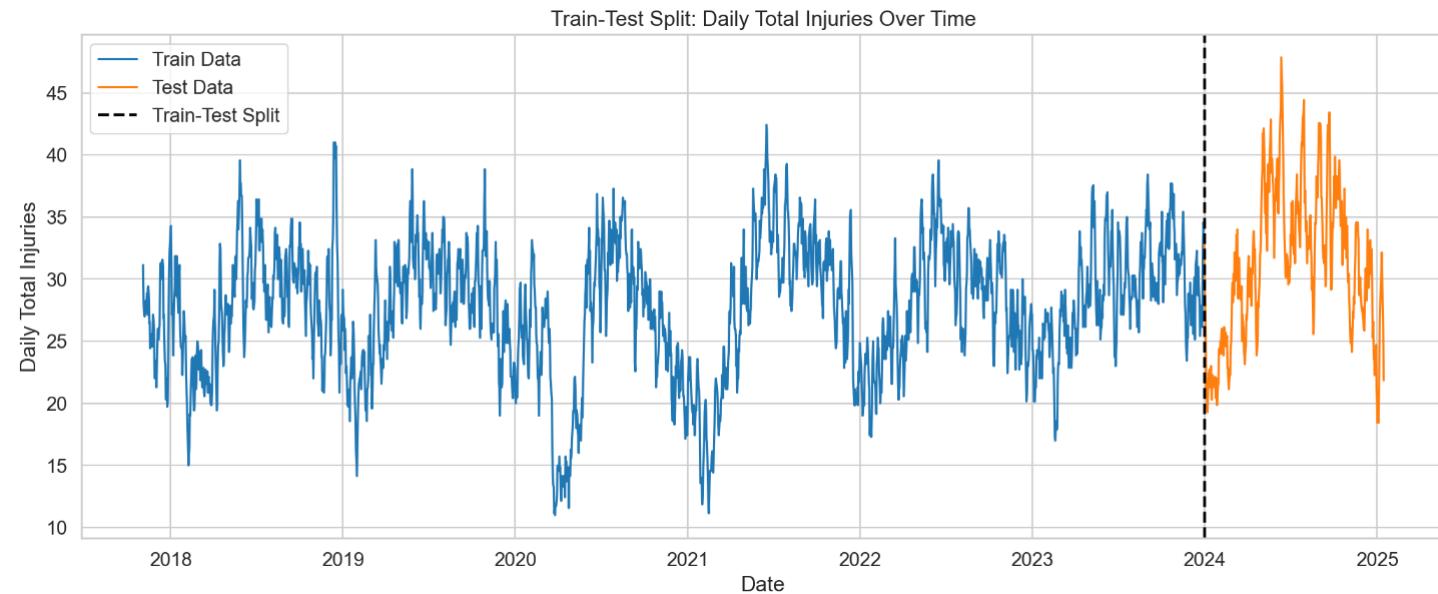


Statistical Theory Applications: CLT

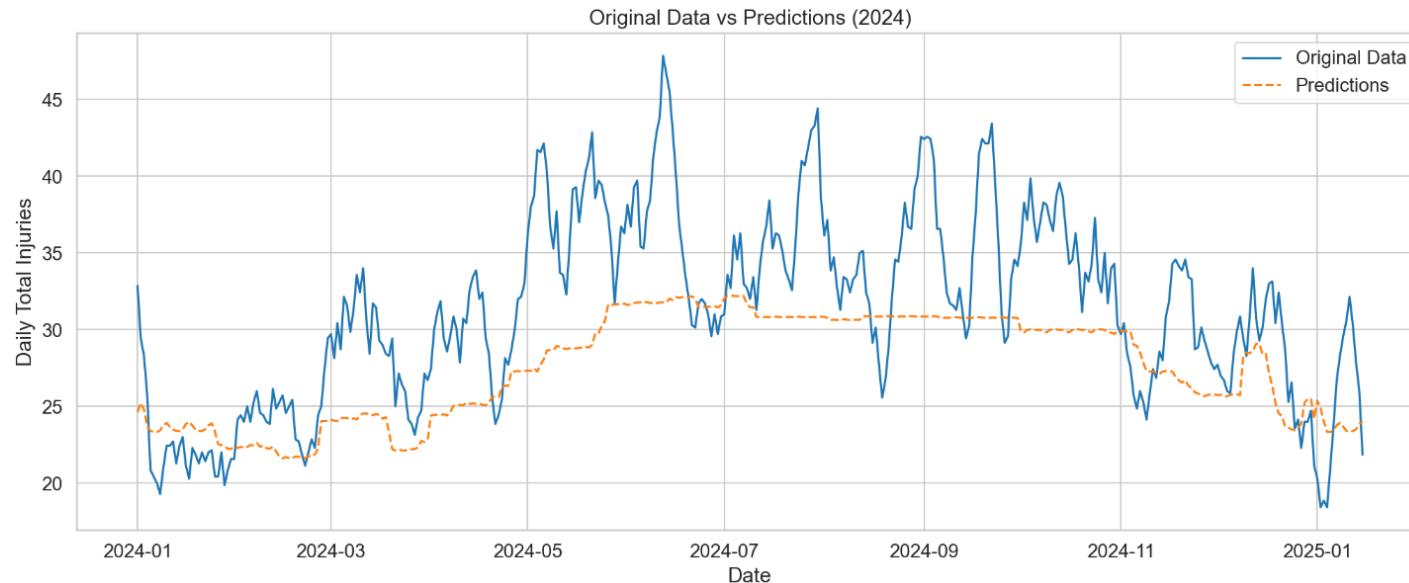
Severe Crash



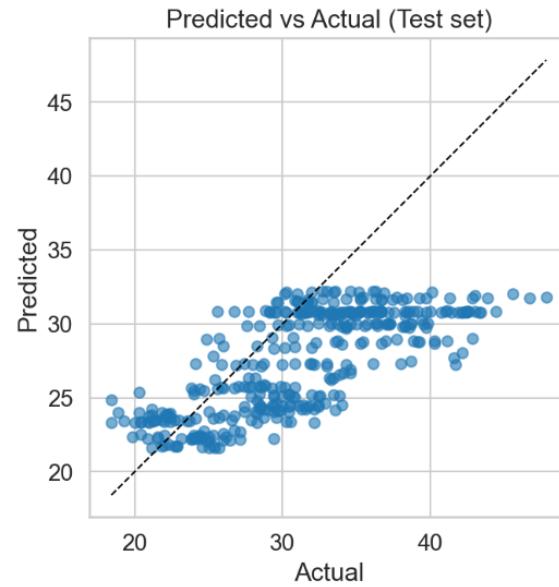
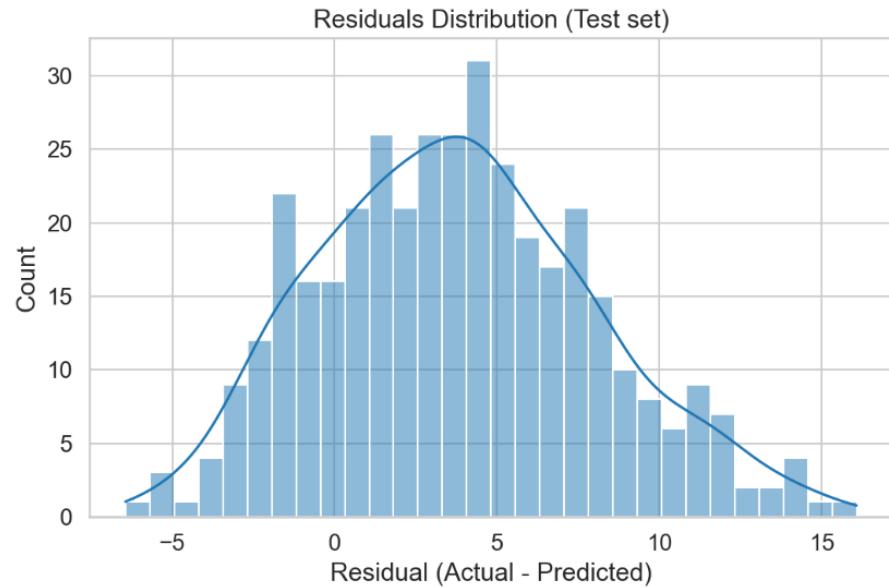
Regression Analysis



Regression Analysis



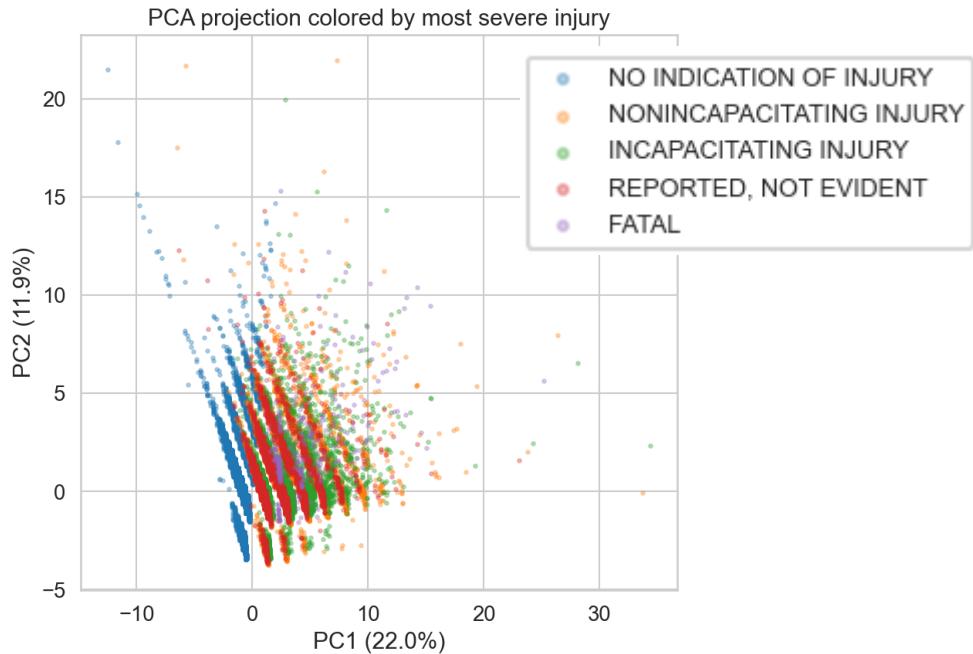
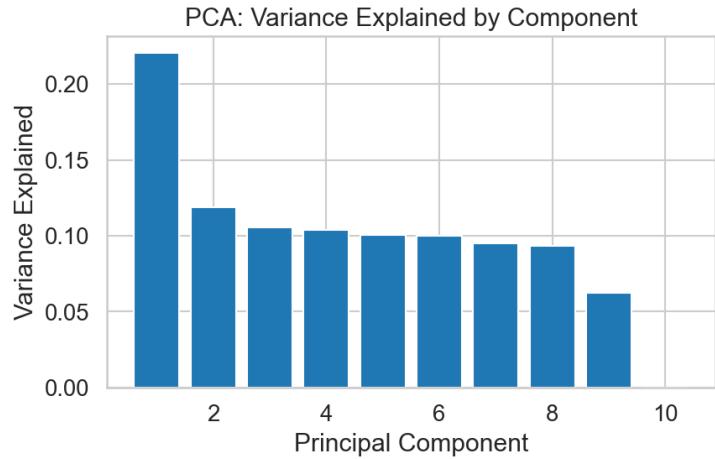
Regression Analysis / Residual Analysis



Regression Analysis / Parameters

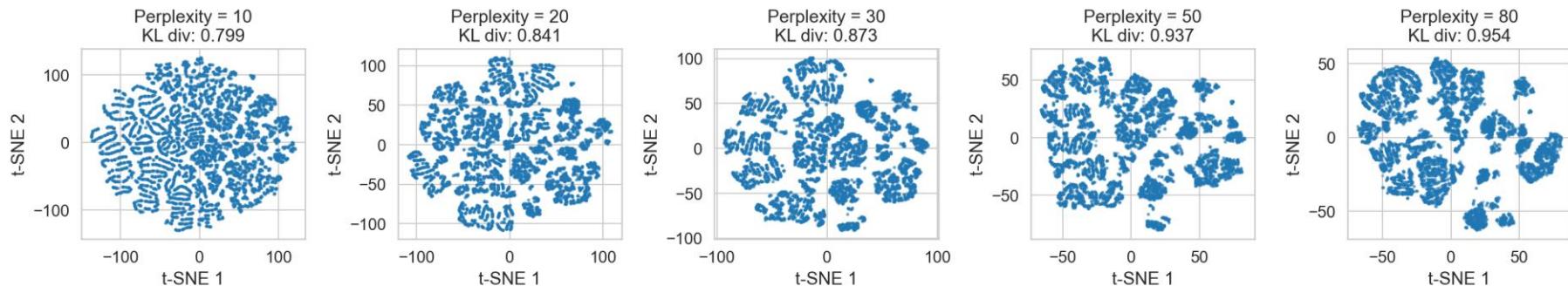
Metric	Value
Test RMSE	5.662
Test MAE	4.533
Test R ²	0.131
Test MAPE	13.57%

Dimensionality Reduction: PCA

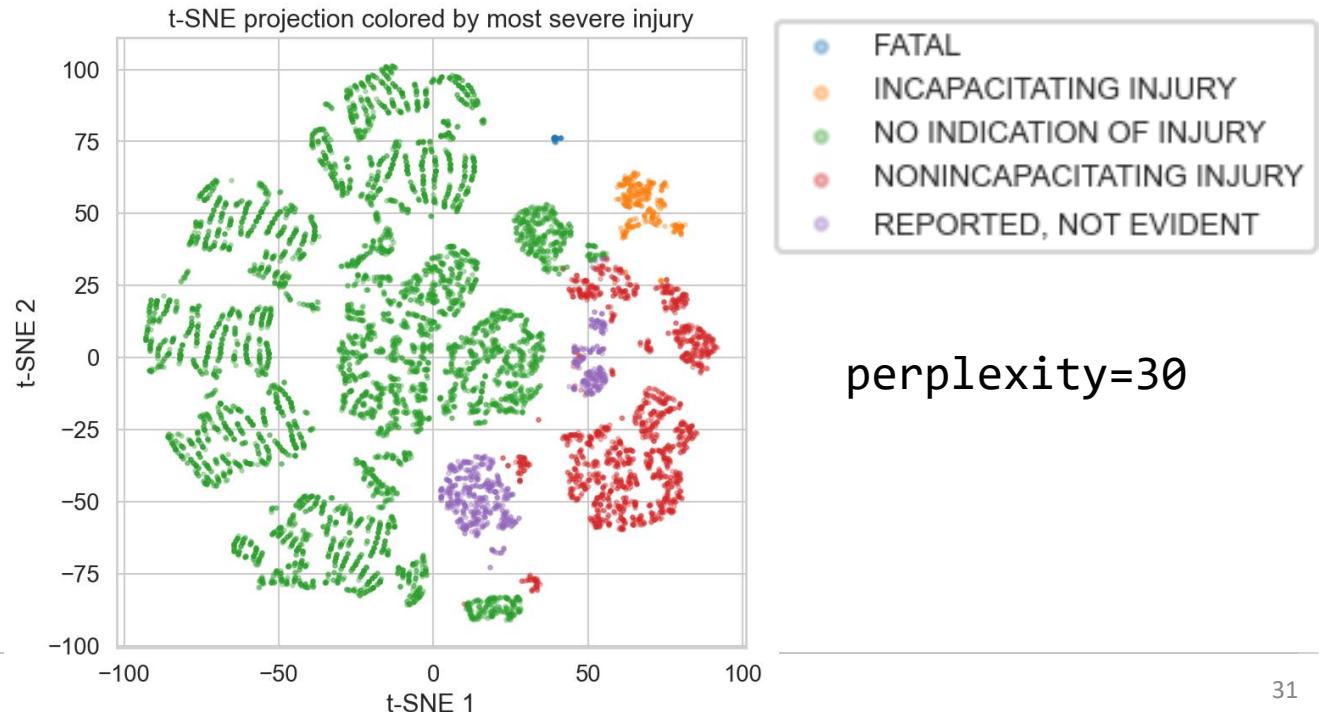


Dimensionality Reduction: t-SNE

Effect of perplexity on t-SNE embedding

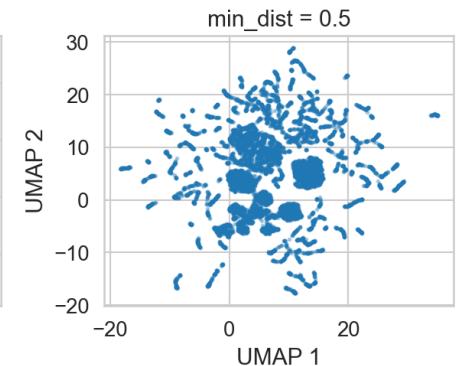
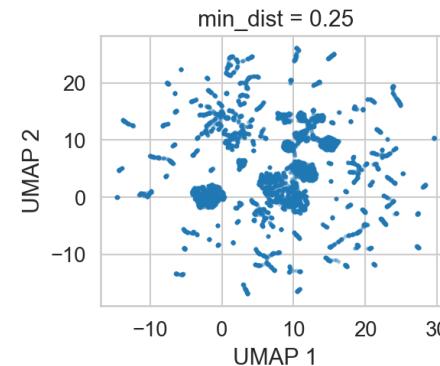
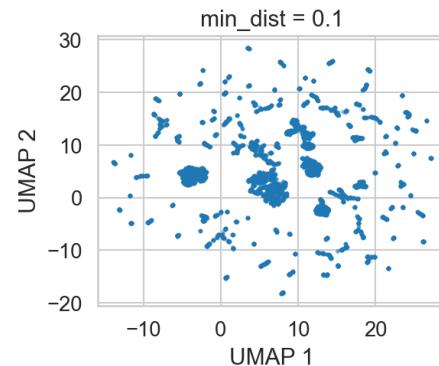
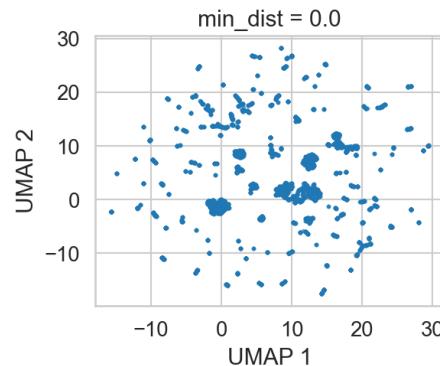


Dimensionality Reduction: t-SNE

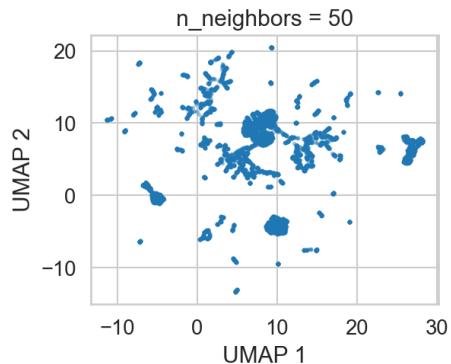
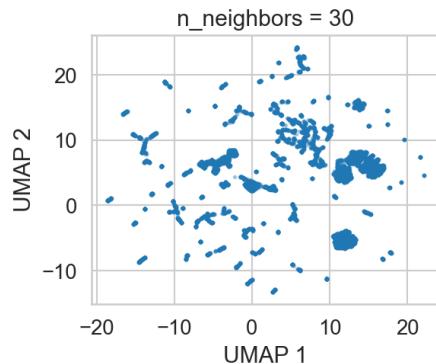
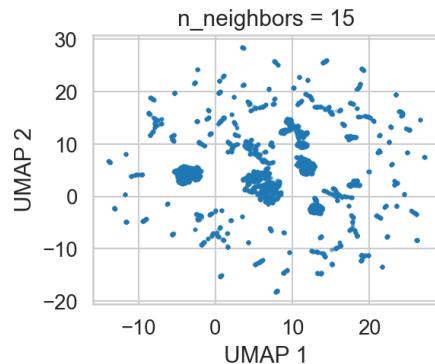
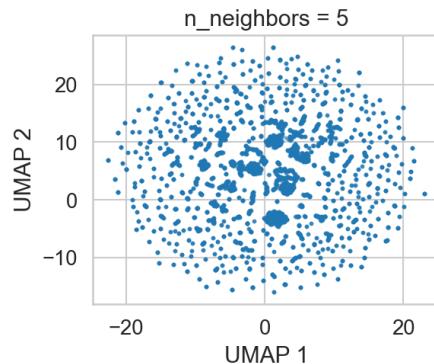


Dimensionality Reduction: UMAP

Effect of min_dist on UMAP embedding (n_neighbors=15)



Dimensionality Reduction: UMAP

Effect of $n_{neighbors}$ on UMAP embedding ($\text{min_dist}=0.1$)

Dimensionality Reduction: UMAP

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

