1 Introduction

1.1 Problem Statement and Context for Readers

Car crashes, also referred to as traffic accidents, are events in which vehicles collide with other vehicles, objects, or pedestrians. Such incidents can result in property damage, physical injuries, and, in severe cases, loss of life. Because of their significant social and economic impact, car crashes are a major public concern. As a result, roadway safety research has become an important field of study. This field focuses on understanding the factors that contribute to traffic accidents and their outcomes.

1.2 Motivation

Identifying the factors that influence car accidents is a central goal in road safety research. Traffic accidents rarely result from a single cause; rather, they typically arise from the interaction of multiple factors, which can lead to severe consequences. Studying these factors is essential for understanding how crashes occur and for developing effective prevention and mitigation strategies.

This study aims to explore how different factors influence crash severity, using an EDA approach and ML models. By analyzing both individual variables and the relationships between them, the goal is not only to assess the impact of each factor alone but also to uncover how these factors interact and jointly contribute to accident outcomes.

1.3 Research Questions

In this study, I analyze a dataset covering various aspects of car crashes in the United States. The main research questions addressed are as follows:

1. **Accident Hotspots:** Which locations experience the highest frequency of traffic collisions? Identifying such areas can help highlight road types or regions that may require targeted safety improvements.

2. **Time Patterns:** Are accidents more frequent at specific times of day, on certain days of the week, or during particular months of the year? Understanding temporal patterns can support more effective traffic management and prevention efforts.

3. **Collision Causes:** What are the most commonly reported causes of collisions in the dataset? Identifying key factors such as speeding, distracted driving, or mechanical failures can help inform preventive policies and interventions.

4. **Environmental Impacts:** How do weather conditions and road environments influence the frequency and severity of accidents? Insights in this area can contribute to better road maintenance and preparedness for adverse conditions.

5. **Vehicle Conditions:** Are certain types of vehicles, such as passenger cars or trucks, more frequently involved in accidents? This analysis can help guide safety regulations and targeted measures for different vehicle categories.

2 Data

2.1 Technical Description

The dataset I analyze consists of 168,727 rows and 41 columns. One known issue with this dataset is that it includes unverified collision records, which may introduce bias or inaccuracies when analyzing and interpreting the variables. During the initial data exploration, I also **identified inconsistencies** between the **ACRS Report Type** and **Injury Severity** columns. For instance, some incidents are labeled as *Injury Crash* in the ACRS Report Type, while the corresponding Injury Severity is recorded as *No Apparent Injury*. Due to these inconsistencies, **I chose to rely on the Injury Severity variable**, as it provides more specific and reliable information about crash outcomes.

Additionally, the "**Vehicle Year"** column **contains abnormal values** such as **0**, **9999 and 2025** (The Crash Date/Time is from 2015 to 2024) (**MCAR**).
And the "**Speed Limit**" column contains **0** while the "**Parked Vehicle**" column says **No.**
These values are likely the result of data entry errors or missing information. I therefore treated these abnormal entries as missing values in the analysis.

2.2 Research plan

As outlined in the Introduction chapter, I plan to investigate five main research questions in this study:

- **Accident Hotspots:** To identify locations with the highest frequency of traffic collisions, I will use the *Route Type* and *Cross-Street Type* variables to pinpoint specific road types or areas where accidents occur more frequently.

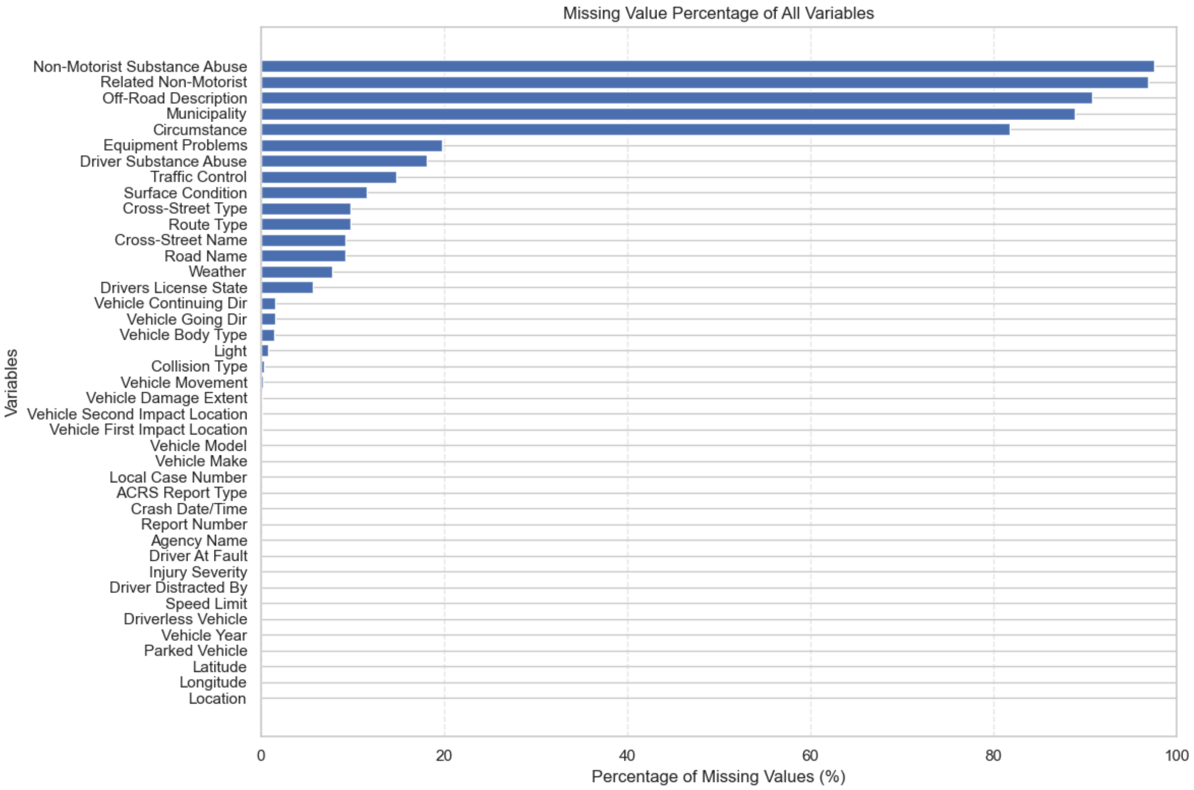
- **Time Patterns:** I will analyze the *Crash Date/Time* variable to examine temporal patterns in traffic accidents. This variable provides detailed information on the date, time, and year of each collision.

- **Collision Causes:** To understand the common causes of collisions and their relationship to crash severity, I will examine variables such as *Driver Substance Abuse*, *Driver Distracted By*, *Vehicle Movement*, and *Equipment Problems*.

- **Environmental Impact:** The influence of external conditions on traffic accidents will be assessed using variables including *Weather*, *Surface Condition*, *Light*, *Traffic Control*, and *Speed Limit*. I will analyze how these factors relate to injury severity and vehicle damage extent.

- **Vehicle Condition:** To determine whether certain types of vehicles are more frequently involved in accidents, I will use the *Vehicle Body Type*, *Vehicle Year*, and *Vehicle Make* variables. These variables will help assess patterns related to vehicle involvement and damage extent.

2.3 Missing value analysis



The graph indicates the following patterns:

Approximately 95% of the observations in the “*Non Motorist Substance Abuse”* and “*Related Non Motorist”* variables are missing, **indicating that** information on non-motorist involvement is **largely unavailable in the dataset**.

The “*Off Road Description*”, “*Municipality*”, and “*Circumstance”* variables exhibit around 80% missing values, which substantially limits their usefulness for reliable analysis.

In contrast, roughly half of the variables **contain no missing values**, providing a solid foundation for further analysis without the need for additional data imputation.

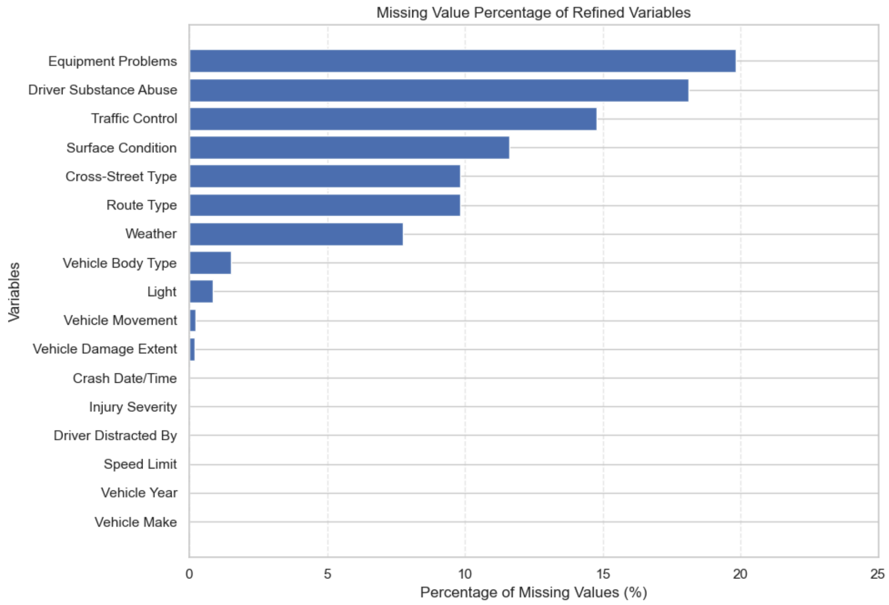
Next, I **refine the dataset** to focus only on variables that are relevant to my research questions. This refinement involves excluding three types of columns:

A. **Identifier columns**, such as “*Report Number”* and “*Local CaseNumber*”. While these fields are useful for tracking individual cases, they **do not provide meaningful information** for the analyses in this study.

B. **Columns with a high proportion of missing values**, such as “*Off Road Description”* and “*Municipality*”. Because these variables contain substantial missing data, **they contribute limited usable information** and may weaken the reliability of the results.

C. **Variables not directly related to the research questions**, such as “*Vehicle Going Dir”* and “*Vehicle First Impact Location”*. **Since these fields are outside the scope** of the questions I investigate, **I removed them** to keep the analysis focused.

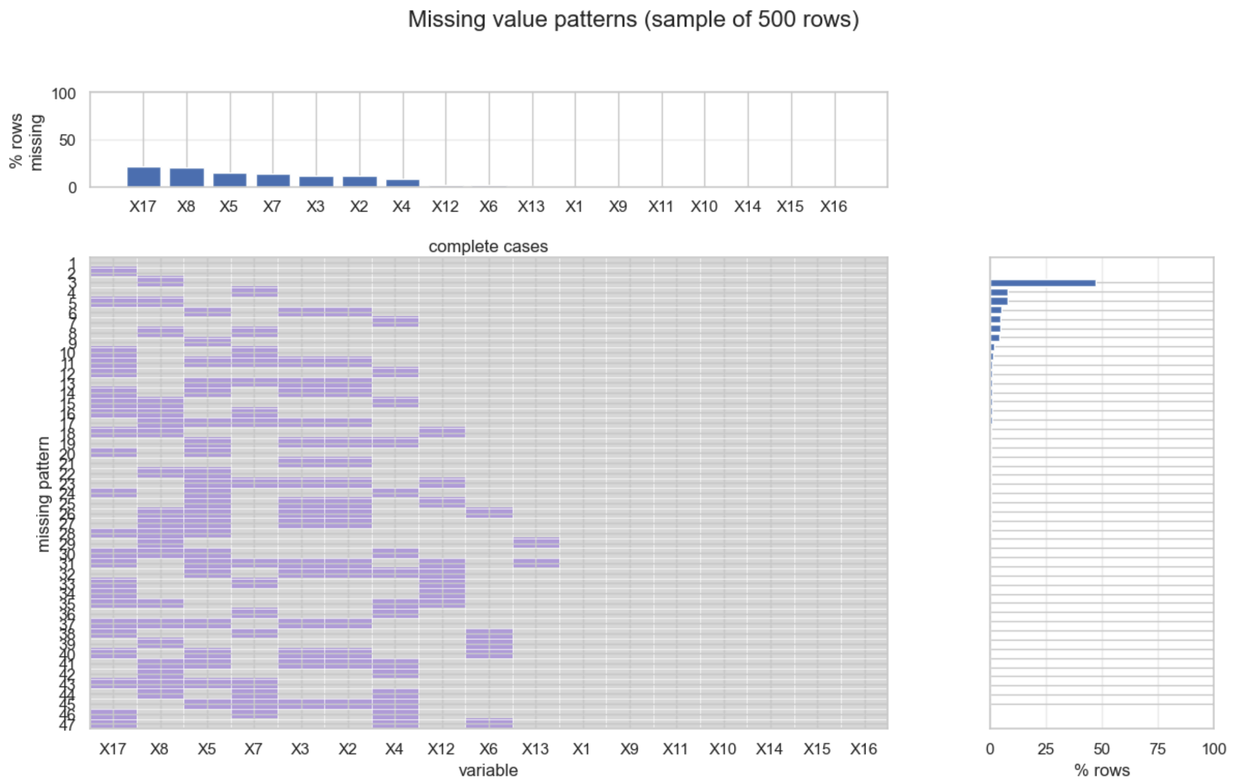
To assess missing data, I use bar charts to visualize the percentage of missing values across variables, and **a missingness plot** (the “plot missing” function) to check for systematic patterns **in where missing values occur**.



From the bar chart, I observe that all variables of interest have **less than 20% missing values**. Among them, “*Equipment Problems”* and “*Driver Substance Abuse”* show relatively higher levels of missingness, each exceeding 15%. Overall, most variables in the refined dataset **contain largely complete data**, making them suitable for further analysis.

To enhance the visual analysis of **missing value patterns**, I selected a random subset of 500 rows and mapped the variable names to numerical labels. Using this sample allows for a clearer visualization of the structure and **distribution of missing values**, making it easier to identify potential patterns or trends in the data.

X1	Crash Date Time	X2	Route Type
X3	Cross Street Type	X4	Weather
X5	Surface Condition	X6	Light
X7	Traffic Control	X8	Drive Substance Abuse
X9	Injury Severity	X10	Driver Distracted By
X11	Vehicle Damage Extent	X12	Vehicle Body type
X13	Vehicle Movement	X14	Speed Limit
X15	Vehicle Year	X16	Vehicle Make
X17	Equipment Problems		



From the graph, I make the following observations:

Approximately 50% of the sampled records contain complete data across all selected variables.

About 5% of the sample is missing values only for “*Equipment Problems”*.

Another 5% of the sample is missing values only for “*Driver Substance Abuse”*.

Roughly 5% of the sample is missing values only for “*Traffic Control”*.

Around 2% of the sample is missing values only for “*Weather”*.

Approximately 2% of the sample is missing values for both “*Equipment Problems”* and “*Driver Substance Abuse”*.

About 2% of the sample is missing values simultaneously for “*Route Type”*, “*Cross Street Type”*, and “*Surface Condition”*.

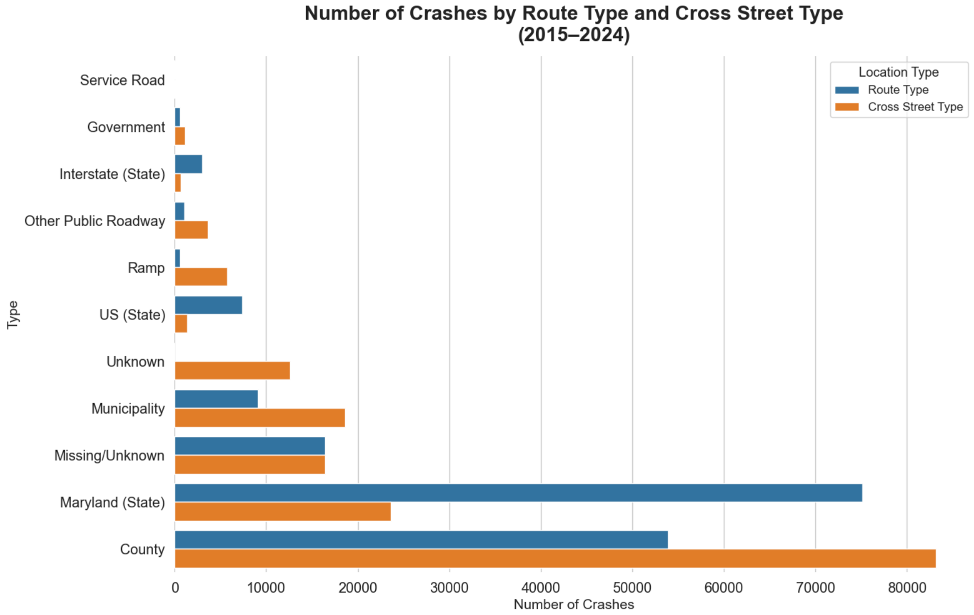
No other distinct or recurring patterns of missing data are evident.

These patterns are **consistent** with earlier observations, as “*Equipment Problems”*, “*Driver Substance Abuse”*, and “*Traffic Control* “ are among the variables with the highest rates of missing values (**MNAR**). In addition, “*Route Type”*, “*Cross Street Type”*, and “*Surface Condition”* are all **related** to road conditions and are therefore more likely to be missing together **(MAR)**.

3 Results

3.1 Accident Hotspots

To examine the **spatial patterns** of traffic collisions, I focus on identifying route types that are **associated with a higher frequency of accidents**. The bar chart shows the number of traffic collisions in Montgomery County from 2015 to 2024, categorized by route type and cross-street type.



From the graph, I observe that **County roads and Maryland State roads** account for the **highest number of crashes**, both as route types and at cross-street intersections. In contrast, road types such as US (State) and Interstate (State) **show fewer crashes overall,** with incidents at cross streets consistently lower than those occurring along the routes themselves. Categories including Ramp, Other Public Roadway, Government, and Service Road **are associated with the lowest number** of crashes and appear to be comparatively safer.

However, **these findings do not necessarily imply** that County roads and Maryland State roads are inherently **more dangerous**. **The dataset is limited to a single county** in Maryland, which naturally increases the representation of these road types and may bias the results. Identifying truly high-risk roads would require additional context, such as traffic volume, exposure rates, and road usage patterns. **This analysis therefore serves as an initial snapshot** that highlights areas for further investigation rather than definitive conclusions about road safety.

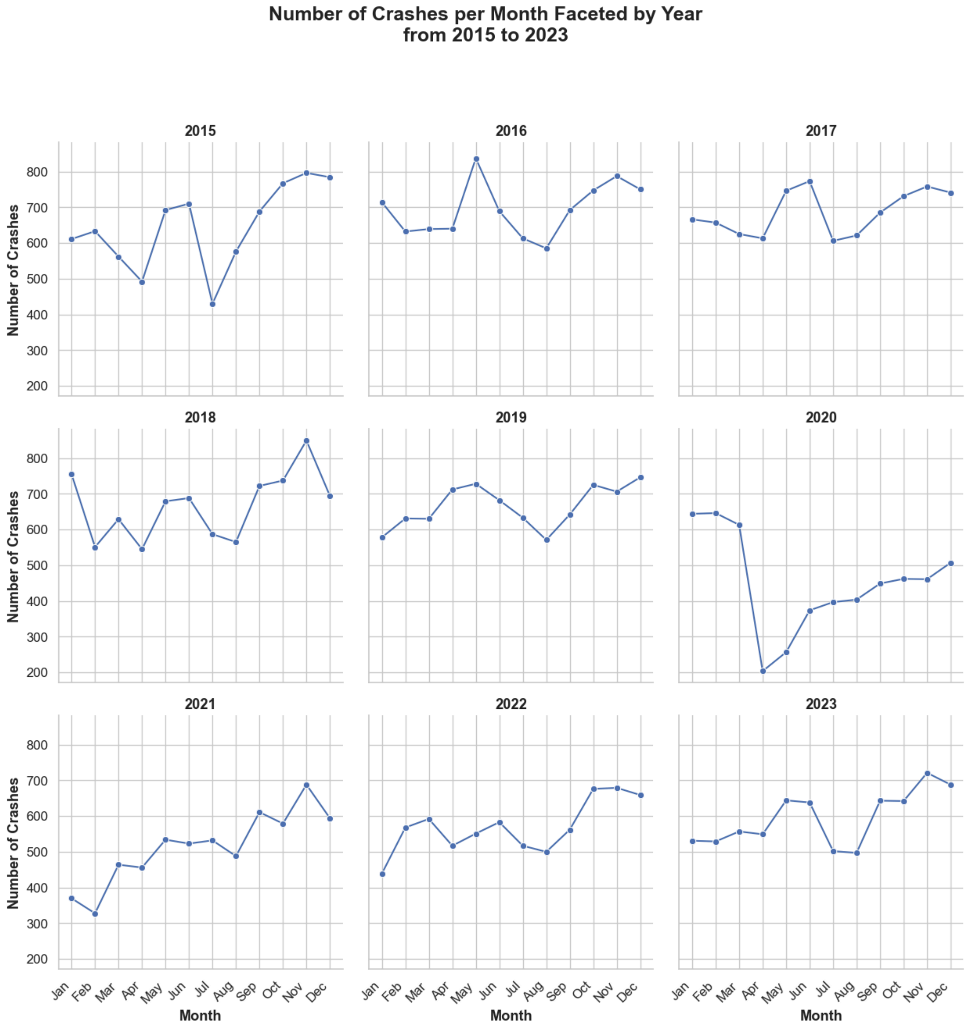
3.2 Time Patterns

Understanding the temporal patterns of road traffic accidents is essential for developing effective safety measures and informed policy decisions. By examining how crash frequencies vary across different times of the day, days of the week, and months of the year, it is possible to identify high-risk periods and gain insight into the factors that may contribute to these patterns.

In this analysis, **I examine traffic collision patterns** in Montgomery County, Maryland at yearly, monthly, weekly, and hourly time scales.

3.2.1 Yearly & Seasonal Patterns

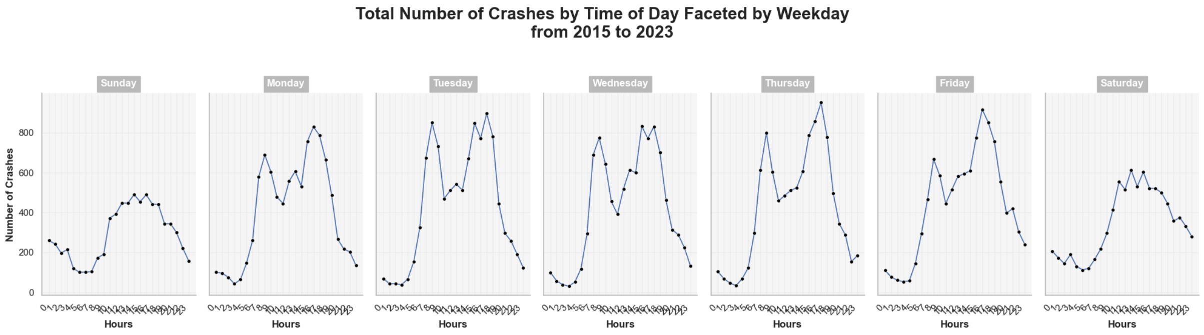
The line graph below presents a detailed temporal analysis of traffic collisions from 2015 to 2023, **highlighting several key patterns in the data.**



A notable decline in the number of crashes is observed in 2020. This deviation **aligns with the onset of the COVID-19 pandemic**, a period marked by lockdowns and widespread restrictions on movement. The substantial reduction in road traffic during this time is likely a key factor behind the decrease in crashes, as stay-at-home orders and the shift to remote work significantly reduced commuting and non-essential travel.

In addition, the graph reveals **recurring peaks** in crash frequency during the months of May and October across multiple years. These increases are likely driven by a combination of factors. In May, higher traffic volumes associated with warmer weather and increased travel, including major holidays such as Memorial Day in the United States, may contribute to a rise in collisions. The October peak may reflect the combined effects of the end of summer travel, the start of the school year, and more challenging driving conditions related to seasonal changes, such as reduced daylight and the presence of fallen leaves on road surfaces and foggier conditions. Overall, these patterns suggest that seasonal behaviors and external events play a significant role in **shaping traffic collision trends.**

3.2.2 Weekly & Time Patterns



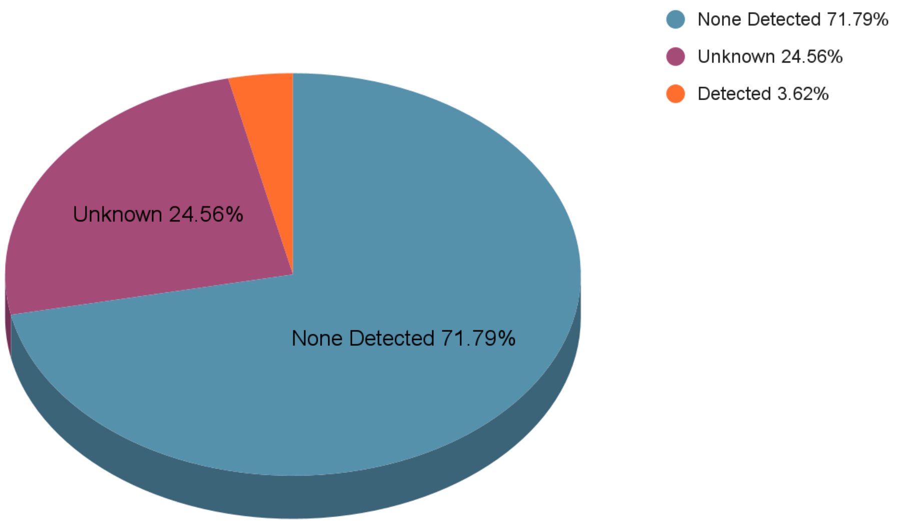
The line graph shows the total number of crashes per hour, faceted per weekday, for the period 2015 to 2023. A clear pattern is that **weekdays have a higher crash frequency** than weekends, which likely reflects higher traffic volume during the workweek due to commuting.

On weekdays, two pronounced peaks appear around **8 AM** and **5 PM**, aligning with typical morning and evening rush hours. In contrast, weekend crash counts peak mainly between **12 PM and 3 PM**, which may correspond to higher midday and afternoon travel as people run errands or go out socially. Overall, these results highlight peak commuting hours on weekdays as periods of elevated crash risk, **suggesting that targeted safety measures and traffic management during these times could help reduce collisions**.

3.3 Collision Causes

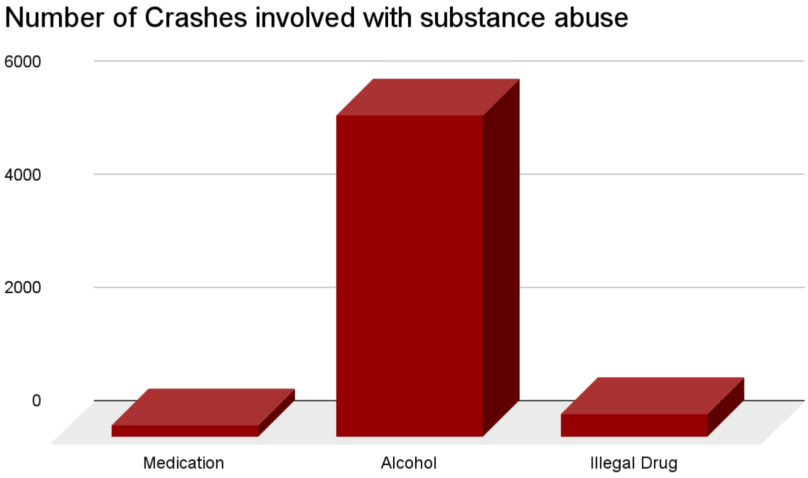
3.3.1 Driver Substance Abuse

Pie chart 1: Percentage of Traffic Collision Involved Driver Substance Abuse



Driver substance abuse is one potential **contributing factor** to traffic collisions. As shown in the table above, the majority of crash cases involve drivers for whom no substance use was detected. **Only a relatively small proportion** of collisions are **associated with confirmed substance abuse.** To further examine these cases, I analyzed the crashes involving detected substance abuse using a stacked bar graph

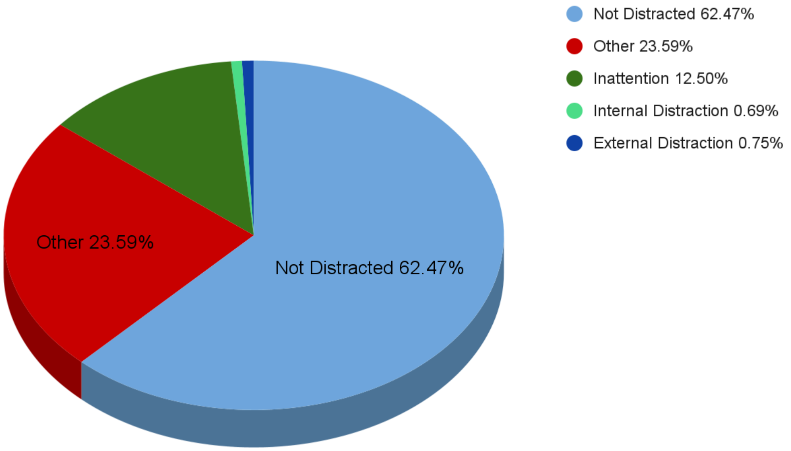
The three main categories of prohibited substances detected among drivers involved in traffic collisions are **alcohol**, **illegal drugs**, and medications that impair driving ability. The graph clearly shows that **alcohol** is the most frequently detected substance in traffic crashes, **far exceeding the other categories.**



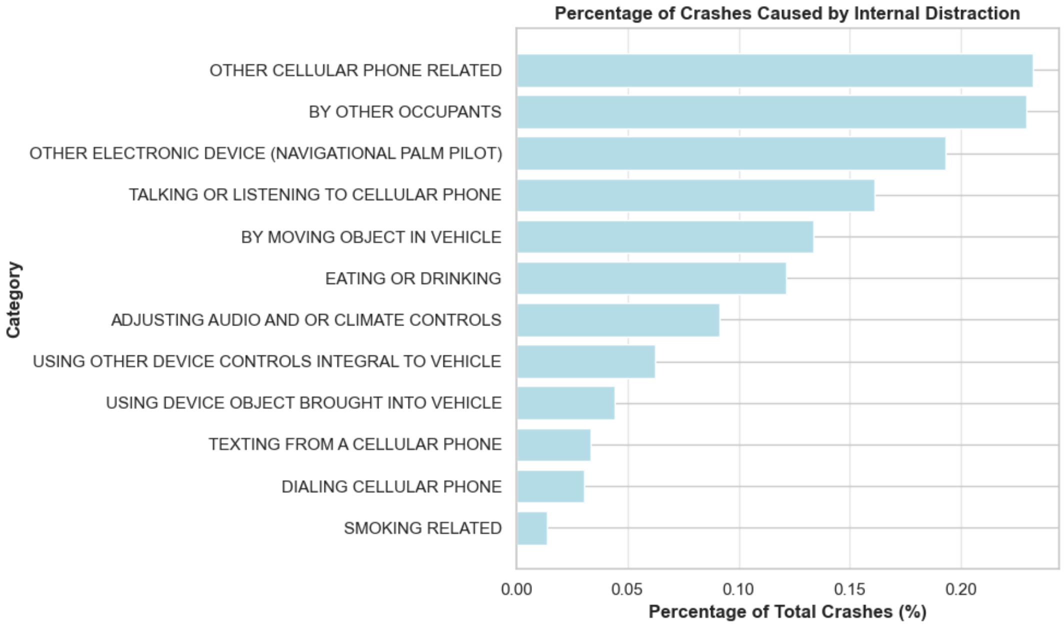
3.3.2 Driver Distraction

Driver distraction is another well-recognized contributing factor to traffic collisions. Distractions can take many forms and are commonly categorized as either internal or external, **depending on their source.** Internal distractions originate within the vehicle and include activities such as talking on or using a mobile phone, texting, eating, or adjusting vehicle controls. External distractions, in contrast, arise from outside the vehicle and may involve attention drawn to surrounding events, objects, or the actions of other road users.

Pie chart 2: Percentage of Traffic Collision Involved Driver Distraction:



According to the data presented in Pie chart 2, approximately **12.5% of traffic collisions involve driver inattention**, including cases where the driver was reported as “looking but did not see” or was “inattentive or lost in thought.” **In contrast**, internal and external distractions each account for only about 1.45% of collisions. **This suggests that general inattention, rather than specific identifiable distractions, is the more prevalent form of driver-related distraction in reported traffic crashes.**



The accompanying graph further breaks down internal distractions and shows that interactions with other occupants in the vehicle are the most frequently reported **internal distraction** leading to crashes.

Conclusion: **To reduce the risks** associated with internal distractions, drivers should **minimize engagement** in activities involving **other occupants** and **avoid using mobile phones** while driving.

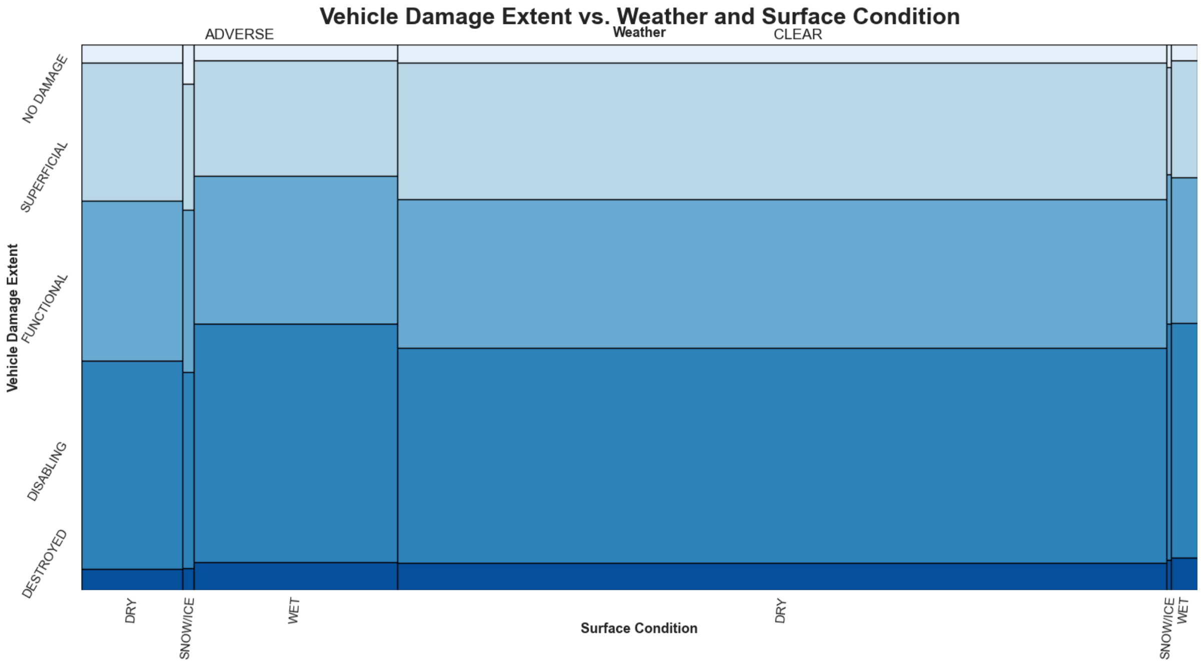
3.4 Environmental Impact

In this section, I use two complementary tools to examine how multiple variables interact and potentially influence one another. The first tool is the **mosaic plot**, which provides a visual representation of the relationships between categorical variables. Mosaic plots allow us to assess **whether different factors appear to be associated** by showing how categories are proportionally divided. If the variables were completely independent, the mosaic plot would display evenly sized segments across categories.

The second tool is the **chi-square test of independence**, a statistical method used to formally evaluate whether the observed relationships between variables are statistically significant. Together, these visual and analytical approaches offer a clearer understanding of the complex interactions among multiple variables.

3.4.1 Weather Condition and Surface Condition

For clearer visualization, we grouped weather and surface conditions into broader categories. I classified **weather** into two levels: **adverse** conditions (such as fog and snow) and **clear** conditions (combining clear and cloudy skies). I also grouped **road surface condition** into four categories: **dry**, **snow/ice**, **wet**, and **other**. Because the “other” category appeared relatively infrequently, I excluded it from the final figure to keep the comparison focused and easier to interpret.



The mosaic plot reveals several notable patterns in the relationship between weather conditions, road surface conditions, and vehicle damage severity. Across both clear and adverse weather conditions, **wet road surfaces are associated with a higher proportion of severe, disabling crashes compared to dry surfaces**, indicating that reduced traction plays a significant role in crash severity.

Under **clear weather**, road surfaces affected by **snow or ice exhibit the highest proportion of severe crashes**. This suggests that drivers may underestimate risk when visibility is good, leading to less cautious driving behavior despite hazardous surface conditions.

In contrast, during **adverse weather**, a somewhat counterintuitive pattern emerges: **dry road surfaces are linked to more crashes than snow or ice**. One possible explanation is that drivers tend to exercise greater caution on visibly hazardous surfaces such as snow or ice, whereas dry roads during adverse weather may create a false sense of security.

The most concerning finding is that the **combination of adverse weather and wet road conditions results in the highest overall severity of vehicle damage** among all examined scenarios. This highlights the compounding effect of poor weather and reduced road traction, emphasizing the importance of targeted safety measures during such conditions.

3.4.4 Chi-Square Test

```
              Dependent        Independent   p_value  Cramers_V
0        Injury Severity        Speed Limit  0.000000      0.0701
1        Injury Severity  Surface Condition  0.000000      0.0523
2        Injury Severity    Traffic Control  0.000000      0.0438
3        Injury Severity              Light  0.000000      0.0218
4        Injury Severity            Weather  0.000082      0.0189
5  Vehicle Damage Extent  Surface Condition  0.000000      0.0940
```

```
6  Vehicle Damage Extent              Light  0.000000     0.0665
7  Vehicle Damage Extent        Speed Limit  0.000000     0.1072
8  Vehicle Damage Extent    Traffic Control  0.000000     0.0599
9  Vehicle Damage Extent            Weather  0.000000     0.0532

   Min_expected  Significant(p<0.01)
0          0.00                 True
1          0.00                 True
2          0.01                 True
3          0.10                 True
4          0.00                 True
5          0.00                 True
6          0.07                 True
7          0.00                 True
8          0.01                 True
9          0.00                 True
```

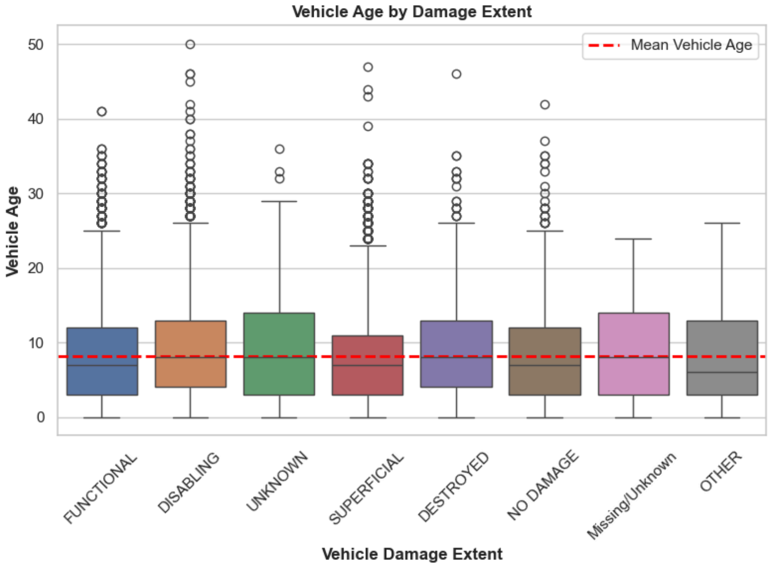
- Most p-values are extremely small (statistically significant), but Cramer's V is mostly in the very weak to weak range (~0.019 to ~0.107).
- Interpretation should emphasize effect size and practical relevance, not significance alone.
- Because `Min_expected` is often below 5, treat chi-square significance as directional evidence and validate with aggregation or robust alternatives when needed.

3.5 Vehicle Condition

In this section, we examine the relationship between collision severity and key vehicular characteristics, focusing specifically on the age and size of the vehicles involved.

3.5.1 Vehicle Age

The boxplot indicates that vehicle age distributions across all damage extent categories are **right-skewed**, with the median age in each category lying to the left of the red line representing the overall mean vehicle age. Compared to other groups, the **no damage** and **superficial damage** categories exhibit lower median vehicle ages, suggesting that newer vehicles are more prevalent in crashes that result in little or no damage.



The **functional** and **disabling** damage categories contain the largest number of vehicles and display similar age distributions, with median ages close to the overall mean. In contrast, vehicles classified as **destroyed** show a wider range of ages, with a substantial proportion of vehicles older than the average. This broader spread suggests that **older vehicles** may be

more susceptible to severe damage when involved in a crash, although additional factors beyond vehicle age are likely to influence this outcome.

Age by damage extent and tested distributional differences (Kruskal-Wallis):

Kruskal H=280.383, p=1.842e-59, $\boldsymbol{\varepsilon}^2 = 0.0044$

Kruskal H = 280.383:

The H-statistic measures how much the ranks of your groups differ from what we would expect if all groups were identical. The Magnitude: 280.383 is a **very large value**.
It indicates a **massive discrepancy between the Damage Extent groups**.

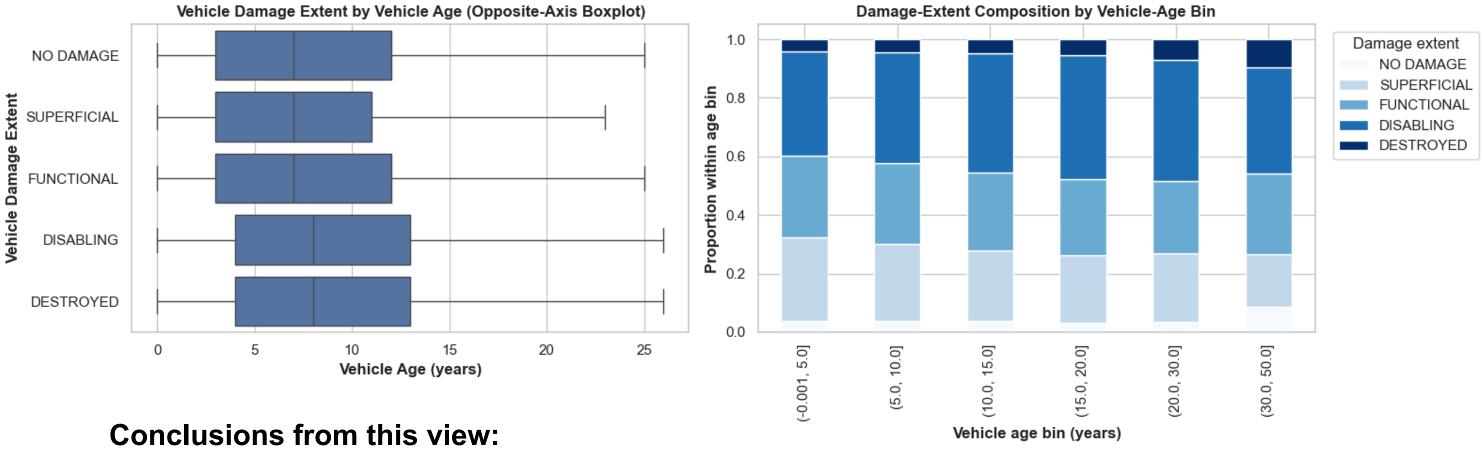
p = 1.842e-59:

Since this is **way below the standard 0.05 cutoff**, this result is **highly statistically significant**.
It means that there is essentially zero chance that the differences we are seeing between the groups happened by random fluke. I can confidently **reject the "null hypothesis"** that all groups are the same.

$\boldsymbol{\varepsilon}^2 = 0.0044$ (The Effect Size):

This is Epsilon Squared, while H and p tells us the groups are different, $\boldsymbol{\varepsilon}^2$ tells us how much they differ. Because the value of 0.0044 is very small, even though the p-value says the difference is "real," **the actual magnitude of that difference is tiny.** In fact, the “Vehicle Damage Extent” **only explains about 0.44%** of the variance in the ranked “Vehicle Age” values.

3.5.2 Additional View: Damage Extent by Vehicle Age (Opposite Orientation):



Conclusions from this view:

- Older age bins tend to show a **modest shift toward higher-severity** damage categories.
- The shift is visible but not large, which is consistent with earlier small-effect results (Kruskal `epsilon^2’).
- Vehicle age provides **directional signals**, but should be interpreted jointly with speed, environment, and traffic-context variables.

4 Predictive Modeling and Validation

4.1 Modeling objective

Main goals:

- Estimate factors associated with severe injury risk.
- Quantify uncertainty using confidence intervals and p-values.
- Evaluate generalization performance on unseen data.

4.2 Feature Engineering

What was done:

- Built cleaned numeric predictors (`VehicleAge`, `SpeedLimitClean`) plus missingness indicators.
- Derived grouped categorical context features (weather, surface, light, traffic, route, weekday, hour-bin).
- **Defined a binary severe-injury target** and applied rare-level collapsing.

How to read related outputs:

- Row count and prevalence summarize sample size and class imbalance.
- Class-balance plot and severe-rate trend plot indicate the rarity and base-rate behavior of the target.

1. After preprocessing and target filtering, the severe-injury modeling sample contains **65,842** records.

2. Severe prevalence %: 0.91 - **The positive class is rare** (about 599 severe cases).

How to interpret this:

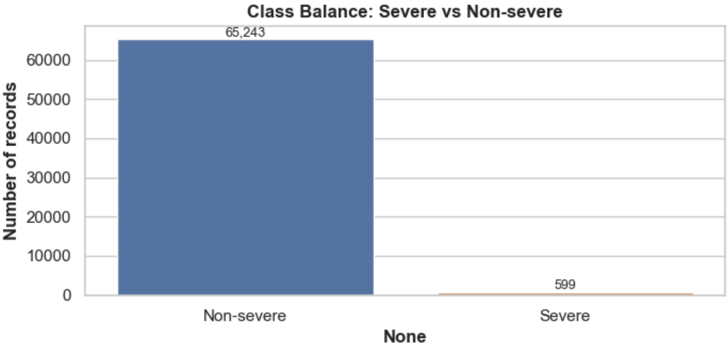
- Severe injury is a rare-event target.
- High accuracy can be misleading if most predictions are non-severe.
- Prefer PR-AUC, calibration, and precision-recall threshold analysis for model selection.

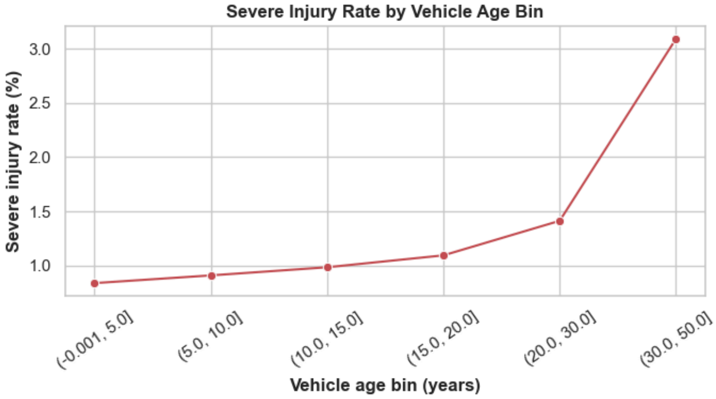
Mini conclusion:
- Class imbalance is a core modeling constraint and must drive metric and threshold choices.

Why This Feature Engineering Matters

- `Vehicle Year` may contain invalid values (`0`, `9999`, future years). These are converted to missing before deriving `VehicleAge`.

- Missingness indicators (`VehicleAgeMissing`, `SpeedMissing`) preserve potential signal instead of silently discarding incomplete records.
- Rare categorical levels are collapsed to `Other` to reduce instability in coefficient estimates.
- Severe injury target definition: `SUSPECTED SERIOUS INJURY` or `FATAL INJURY`.

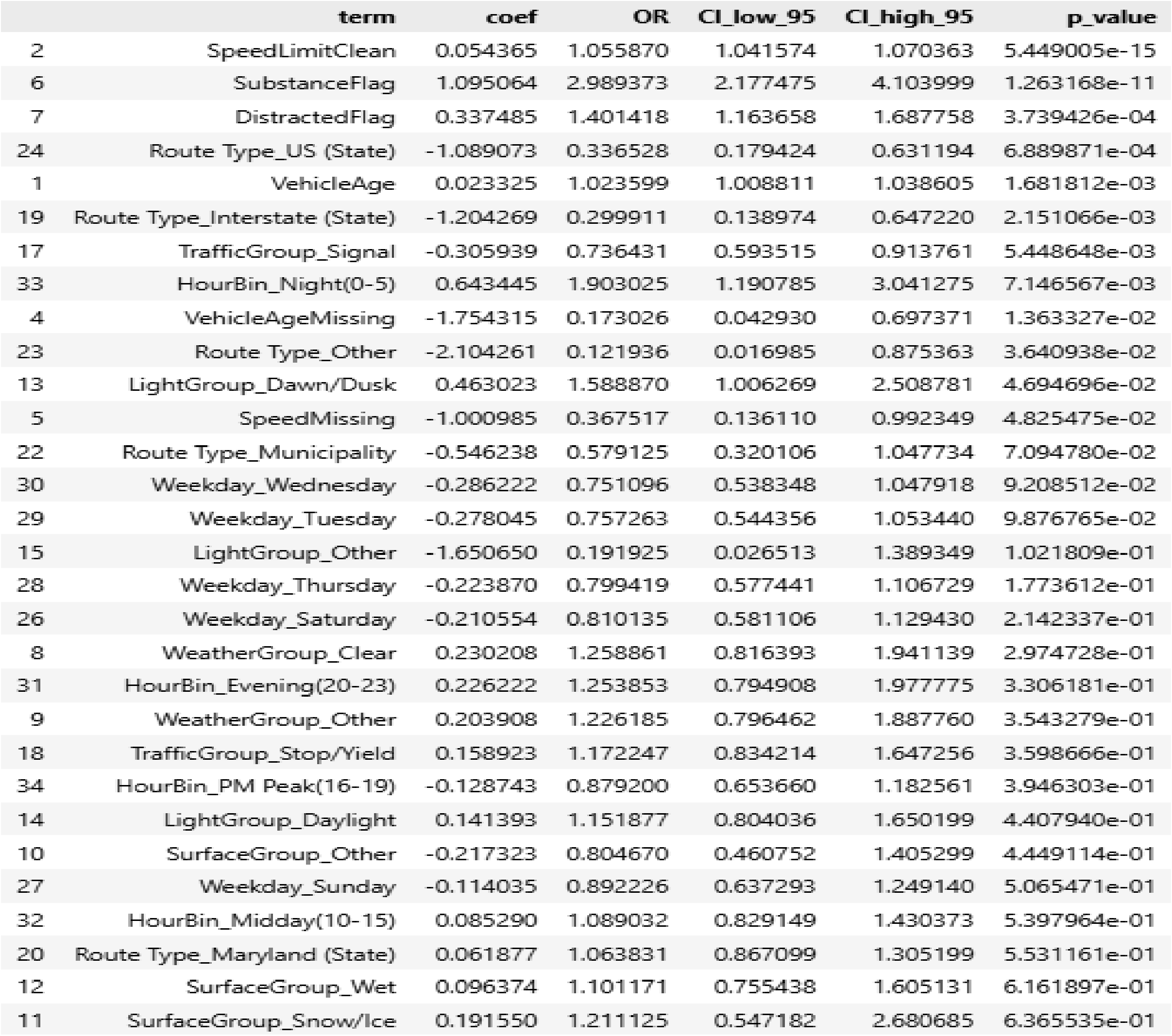

Class Balance: Severe vs Non-severe


Severe Injury Rate by Vehicle Age Bin

4.3 Logistic Regression for Severe Injury Risk

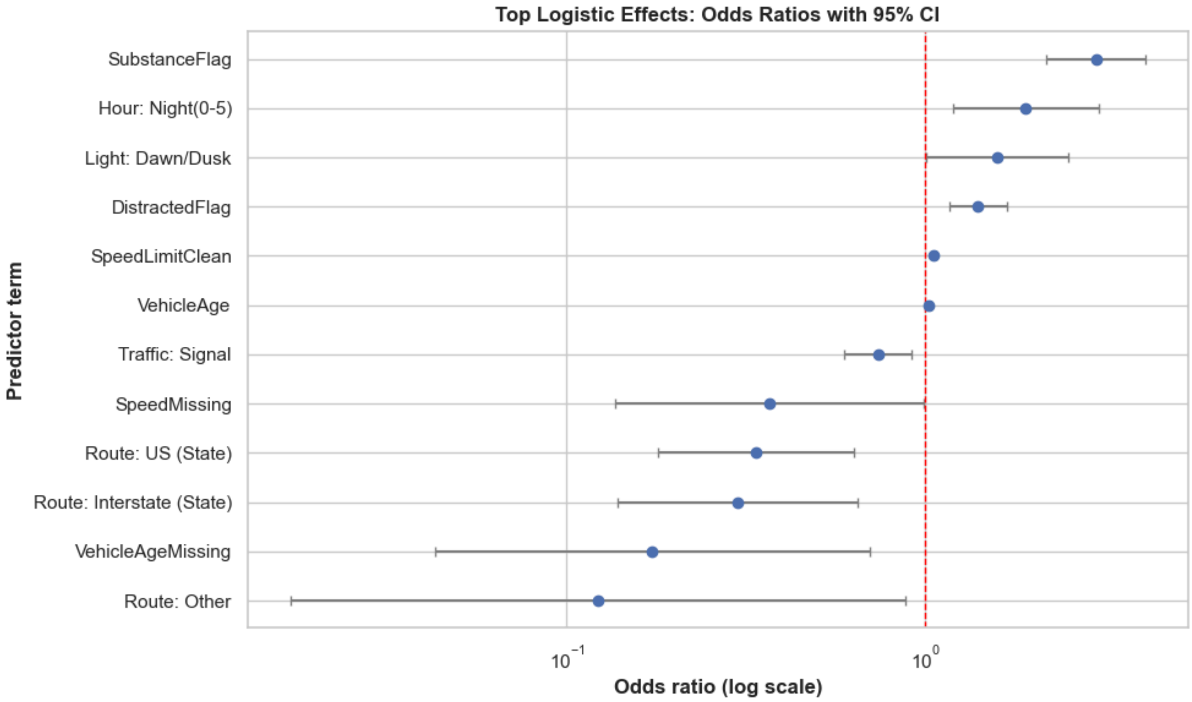
What was done:

- Fit a Binomial GLM for inferential interpretation.
- Reported coefficients, odds ratios, confidence intervals, and p-values.
- Computed holdout metrics plus calibration diagnostics.



	term	coef	OR	CI_low_95	CI_high_95	p_value
2	SpeedLimitClean	0.054365	1.055870	1.041574	1.070363	5.449005e-15
6	SubstanceFlag	1.095064	2.989373	2.177475	4.103999	1.263168e-11
7	DistractedFlag	0.337485	1.401418	1.163658	1.687758	3.739426e-04
24	Route Type_US (State)	-1.089073	0.336528	0.179424	0.631194	6.889871e-04
1	VehicleAge	0.023325	1.023599	1.008811	1.038605	1.681812e-03
19	Route Type_Interstate (State)	-1.204269	0.299911	0.138974	0.647220	2.151066e-03
17	TrafficGroup_Signal	-0.305939	0.736431	0.593515	0.913761	5.448648e-03
33	HourBin_Night(0-5)	0.643445	1.903025	1.190785	3.041275	7.146567e-03
4	VehicleAgeMissing	-1.754315	0.173026	0.042930	0.697371	1.363327e-02
23	Route Type_Other	-2.104261	0.121936	0.016985	0.875363	3.640938e-02
13	LightGroup_Dawn/Dusk	0.463023	1.588870	1.006269	2.508781	4.694696e-02
5	SpeedMissing	-1.000985	0.367517	0.136110	0.992349	4.825475e-02
22	Route Type_Municipality	-0.546238	0.579125	0.320106	1.047734	7.094780e-02
30	Weekday_Wednesday	-0.286222	0.751096	0.538348	1.047918	9.208512e-02
29	Weekday_Tuesday	-0.278045	0.757263	0.544356	1.053440	9.876765e-02
15	LightGroup_Other	-1.650650	0.191925	0.026513	1.389349	1.021809e-01
28	Weekday_Thursday	-0.223870	0.799419	0.577441	1.106729	1.773612e-01
26	Weekday_Saturday	-0.210554	0.810135	0.581106	1.129430	2.142337e-01
8	WeatherGroup_Clear	0.230208	1.258861	0.816393	1.941139	2.974728e-01
31	HourBin_Evening(20-23)	0.226222	1.253853	0.794908	1.977775	3.306181e-01
9	WeatherGroup_Other	0.203908	1.226185	0.796462	1.887760	3.543279e-01
18	TrafficGroup_Stop/Yield	0.158923	1.172247	0.834214	1.647256	3.598666e-01
34	HourBin_PM Peak(16-19)	-0.128743	0.879200	0.653660	1.182561	3.946303e-01
14	LightGroup_Daylight	0.141393	1.151877	0.804036	1.650199	4.407940e-01
10	SurfaceGroup_Other	-0.217323	0.804670	0.460752	1.405299	4.449114e-01
27	Weekday_Sunday	-0.114035	0.892226	0.637293	1.249140	5.065471e-01
32	HourBin_Midday(10-15)	0.085290	1.089032	0.829149	1.430373	5.397964e-01
20	Route Type_Maryland (State)	0.061877	1.063831	0.867099	1.305199	5.531161e-01
12	SurfaceGroup_Wet	0.096374	1.101171	0.755438	1.605131	6.161897e-01
11	SurfaceGroup_Snow/Ice	0.191550	1.211125	0.547182	2.680685	6.365535e-01

- The model identifies directionally interpretable associations, but predictive discrimination remains limited and should be interpreted with caution in operational use.



Model Summary: Top Predictors of Outcome

A multivariate **logistic regression analysis** was conducted to identify the key factors associated with the outcome variable. The results, visualized as Odds Ratios (OR) with 95% Confidence Intervals (CI), **indicate several statistically significant predictors**:

Primary Risk Factors: Substance use (SubstanceFlag) emerged as the **strongest predictor**, significantly increasing the odds of the outcome compared to the reference group. Other significant risk factors included nighttime driving (Hour: Night (0-5)), distracted driving (DistractedFlag), and dawn/dusk light conditions.

Protective Factors: The presence of a traffic signal (Traffic: Signal) and travel on Interstate or State routes were found to be **significant protective factors**, associated with a **reduction** in the odds of the outcome.

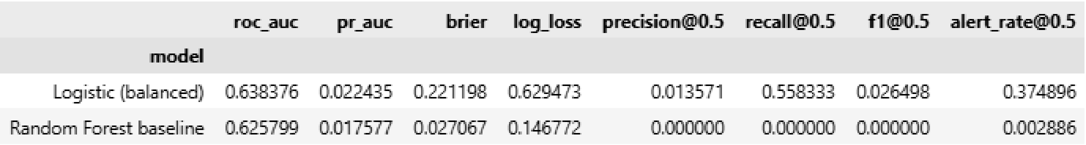
Non-Significant Predictors: Several variables, including VehicleAge and SpeedLimitClean, were found to be **statistically non-significant**, as their 95% confidence intervals included the null value of 1.0.

Model Limitations: While the model identifies directionally interpretable associations, the high level of uncertainty for certain categories (e.g., Route: Other) and **limited predictive discrimination** suggest that findings should be applied with caution in operational settings.

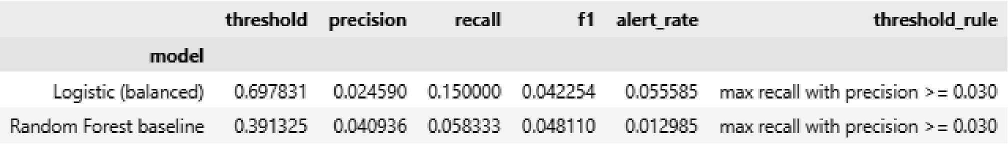
4.4 Model Comparison: Logistic and RF (Validation Protocol)

What was done:

- Trained logistic and random-forest pipelines with identical preprocessing.
- Used a comparable threshold-selection rule (precision floor with recall maximization).
- Reported both threshold-free and threshold-specific metrics.



	roc_auc	pr_auc	brier	log_loss	precision@0.5	recall@0.5	f1@0.5	alert_rate@0.5
model								
Logistic (balanced)	0.638376	0.022435	0.221198	0.629473	0.013571	0.558333	0.026498	0.374896
Random Forest baseline	0.625799	0.017577	0.027067	0.146772	0.000000	0.000000	0.000000	0.002886



	threshold	precision	recall	f1	alert_rate	threshold_rule
model						
Logistic (balanced)	0.697831	0.024590	0.150000	0.042254	0.055585	max recall with precision >= 0.030
Random Forest baseline	0.391325	0.040936	0.058333	0.048110	0.012985	max recall with precision >= 0.030

Model Performance Summary

The evaluation compares a Logistic Regression model (using balanced class weights) against a Random Forest baseline for the classification of a **highly imbalanced target.**

1. Overall Discriminative Power

- **Logistic Regression outperformed the Random Forest** across most key metrics, achieving a higher ROC-AUC (0.638 vs. 0.625) and a superior PR-AUC (0.022 vs. 0.017).
- The higher PR-AUC indicates that the Logistic model is approximately **27% more effective** at **identifying the minority class** than the baseline.

2. Default Threshold Performance (@0.5)

- At the default 0.5 decision threshold, the Logistic model yields a **Recall of 0.558**, **successfully identifying 55.8% of positive cases.**
- However, the **precision** is extremely low (0.013), suggesting a very **high false-positive rate**. This is typical for models using balanced class weights on severely imbalanced data.

3. Optimized Threshold Performance

- To make the model practically useful, the threshold was tuned using the rule: **Maximize Recall while maintaining a minimum Precision of 3% (0.03).**
- Logistic Regression: To meet the 3% precision requirement, the threshold was adjusted to 0.697. At this level, the model maintains a Recall of 15% with an Alert Rate of 5.5%.
- Random Forest: Under the same constraint, the baseline **only achieved a Recall of 5.8%**, confirming that the **Logistic model is the more robust choice** for this specific task.

Key Conclusions

- Model Selection: **The Logistic Regression** (balanced) model **is the superior** performer for identifying severe cases compared to the Random Forest baseline.
- Class Imbalance: The low PR-AUC and Precision scores highlight the extreme difficulty of the classification task, likely due to a high degree of overlap between classes or a very low positive class prevalence.
- Operational Trade-off: **Using the tuned threshold significantly reduces** "alert fatigue" (lowering the alert rate to 5.5%), though it **requires sacrificing significant recall** (dropping from 55% to 15%).

5. Conclusion

Direct answers to the research questions:

1. **Hotspots**: Crash frequency is concentrated in specific road-type contexts.

2. **Time patterns**: Crash counts vary by month, weekday, and hour, with clear temporal structure.

3. **Collision causes**: Substance, distraction, and movement variables show measurable but heterogeneous associations with outcomes.

4. **Environmental impact**: Weather, surface, light, traffic control, and speed context are statistically associated with injury/damage outcomes.

5. **Vehicle condition**: Vehicle age and size contribute descriptive signal, but effect sizes are small when tested directly.

Integrated statistical conclusion:

- Many tests are statistically significant due large sample size.
- Effect-size evidence (Cramer's V and epsilon^2) suggests mostly weak individual associations.
- Severe-injury prediction remains difficult in this dataset because prevalence is very low (0.91%).

Practical conclusion:

- Results are **best used for** risk stratification, prioritization, and hypothesis generation.
- They **should not** be interpreted as causal effects or deterministic decision rules.

Limitations:

- Crash-only observational data without exposure denominator.
- Non-trivial missing/unknown categories and potential reporting bias.