Business Report: Car Crash Survival Analysis and Prediction

Table Of Contents

1.Problem Statement

The Department of Road Transport has observed a 15% year-over-year increase in car crashes in urban areas. The objective of this analysis is to build a machine learning model that can predict the likelihood of survival in a car crash based on historical data. This will help identify key safety factors and recommend improvements for regulatory and design standards.

2. Objective

- Analyze historical car crash data.
- Identify significant factors contributing to crash survival.
- Build machine learning models to predict survival.
- Evaluate and tune models to improve performance.
- Derive actionable business insights from data patterns.

3. Dataset Overview

The dataset consists of 11,217 car crash records with 12 features such as crash speed, occupant age, airbag deployment, seatbelt usage, and outcome (`deceased`).

. First 5 Rows Of Data

	caseid	speed_range	weight	seatbelt	frontal_impact	sex	age_of_occ	year_of_acc	model_year	airbag	occ_role	deceased
0	02:13:02	55+ km/h	27.07800	none	1	m	32	1997	1987	unavail	driver	yes
1	02:17:01	25-39 km/h	89.62700	belted	0	f	54	1997	1994	nodeploy	driver	yes
2	0.138206019	55+ km/h	27.07800	belted	1	m	67	1997	1992	unavail	driver	yes
3	0.138206019	55+ km/h	27.07800	belted	1	f	64	1997	1992	unavail	pass	yes
4	04:58:01	55+ km/h	13.37400	none	1	m	23	1997	1986	unavail	driver	yes

.Last 5 Rows Of Data

	caseid	speed_range	weight	seatbelt	frontal_impact	sex	age_of_occ	year_of_acc	model_year	airbag	occ_role	deceased
11212	82:107:1	25-39 km/h	3179.68800	belted	1	m	17	2002	1985	unavail	driver	no
11213	82:108:2	10-24 km/h	71.22800	belted	1	m	54	2002	2002	nodeploy	driver	no
11214	82:110:1	10-24 km/h	10.47400	belted	1	f	27	2002	1990	deploy	driver	no
11215	82:110:2	25-39 km/h	10.47400	belted	1	f	18	2002	1999	deploy	driver	no
11216	82:110:2	25-39 km/h	10.47400	belted	1	m	17	2002	1999	deploy	pass	no

.Columns Of Data Set

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 11217 entries, 0 to 11216

Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	caseid	11217 non-null	object
1	speed_range	11217 non-null	object
2	weight	11217 non-null	float64
3	seatbelt	11217 non-null	object
4	frontal_impact	11217 non-null	int64
5	sex	11217 non-null	object

6 age_of_occ 11217 non-null int64

7 year_of_acc 11217 non-null int64

8 model_year 11217 non-null int64

9 airbag 11217 non-null object

10 occ_role 11217 non-null object

11 deceased 11217 non-null object

dtypes: float64(1), int64(4), object(7)

memory usage: 1.0+ MB

Data Types:

- object (7 columns): Categorical/text data (e.g., caseid, seatbelt, sex, deceased).
- int64 (4 columns): Integer values (e.g., age_of_occ, model_year).
- float64 (1 column): Continuous numeric (weight).

Statistical summary of the dataset

	weight	frontal_impact	age_of_occ	year_of_acc	model_year
count	11217.00000	11217.00000	11217.00000	11217.00000	11217.00000
mean	431.40531	0.64402	37.42765	2001.10324	1994.17794
std	1406.20294	0.47883	18.19243	1.05681	5.65870
min	0.00000	0.00000	16.00000	1997.00000	1953.00000
25%	28.29200	0.00000	22.00000	2001.00000	1991.00000
50%	82.19500	1.00000	33.00000	2001.00000	1995.00000
75%	324.05600	1.00000	48.00000	2002.00000	1999.00000
max	31694.04000	1.00000	97.00000	2002.00000	2003.00000

- `weight`: Extremely skewed with outliers — mean =431, but max = 31,694. Most values are much lower (75% under 324).

- `frontal_impact`: Binary feature (0 or 1). About 64% of crashes involved a frontal impact (mean = 0.64).

- `age_of_occ`: Occupants' age ranges from 16 to 97, average age ~37. Most are between 22 and 48.

- `year_of_acc`: Crashes mostly occurred between 1997 and 2002.

- `model_year`: Vehicle model years range from 1953 to 2003. Newer vehicles likely offer better safety features.

Duplicate Values and Missing Values

caseid	**0**
speed_range	**0**
weight	**0**
seatbelt	**0**
frontal_impact	**0**
sex	**0**

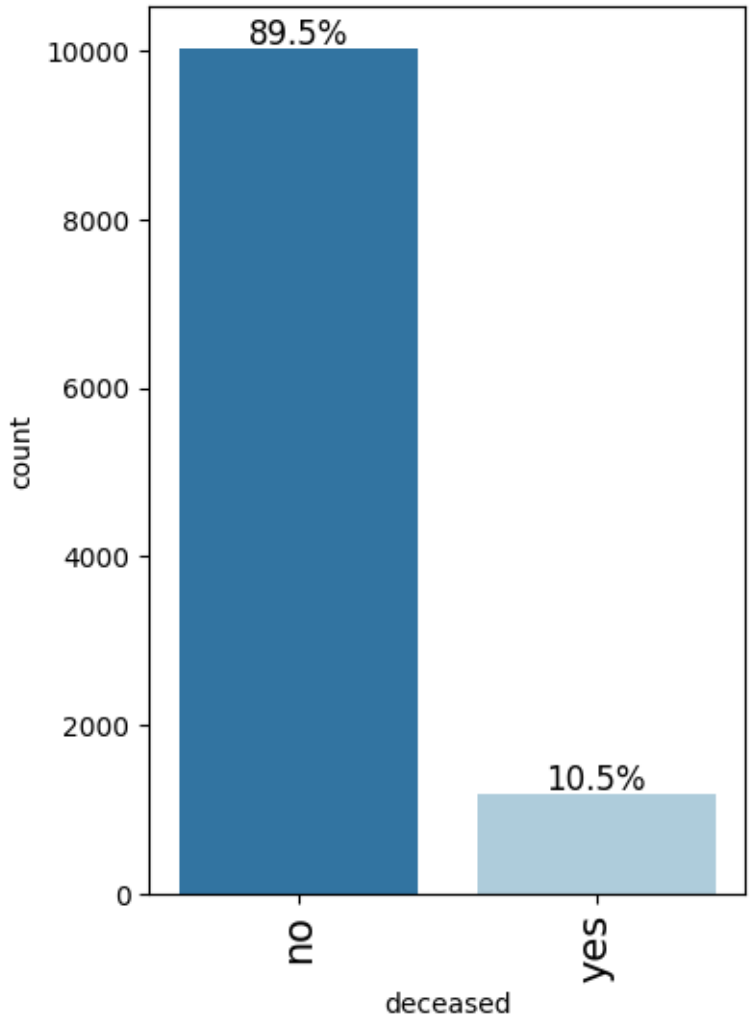
age_of_occ	**0**
year_of_acc	**0**
model_year	**0**
airbag	**0**
occ_role	**0**
deceased	**0**

There are no duplicate and missing values in this data set

4.Exploratory Data Analysis (EDA)

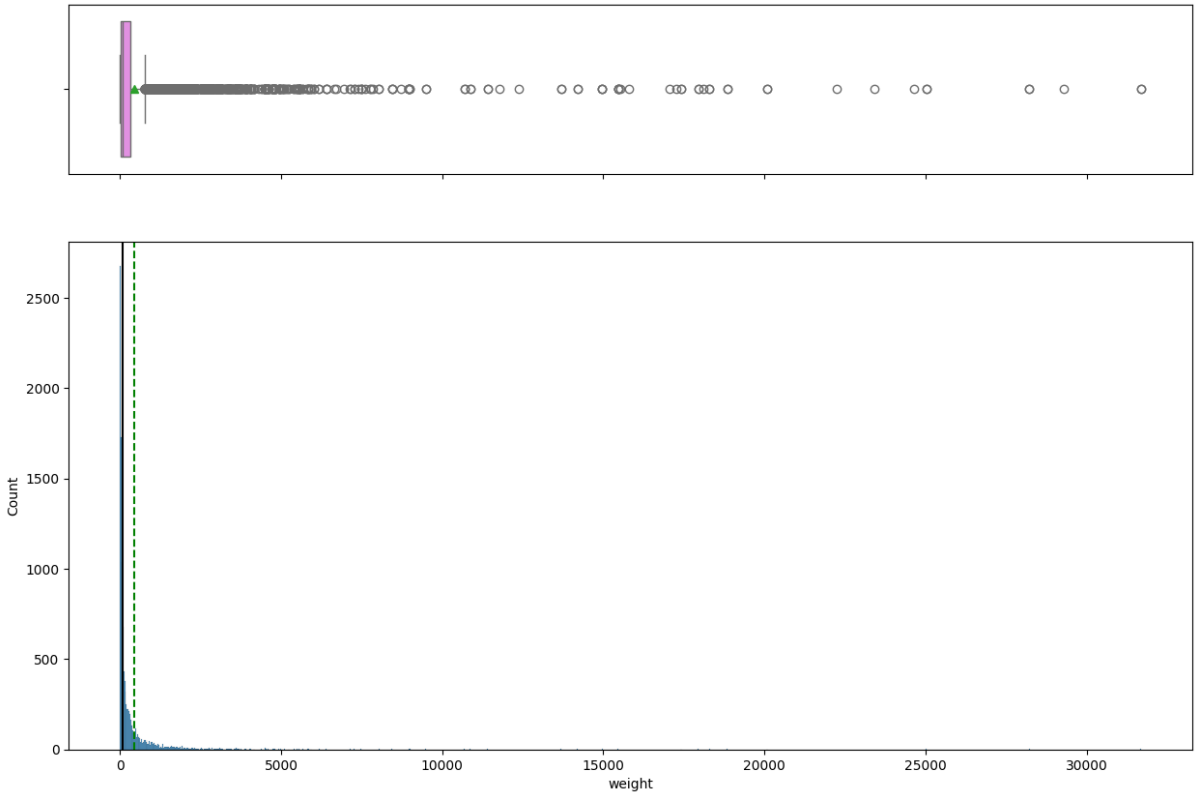
1.Univariate Analysis

Observations On Deceased



There are less deceased under car crash by 10.5%

Observations On weight:



Top Plot (Boxplot): Highlights extreme **outliers** in the weight variable. Most data points are tightly clustered on the lower end, while a few extremely high values stretch the axis, confirming right-skewed distribution.

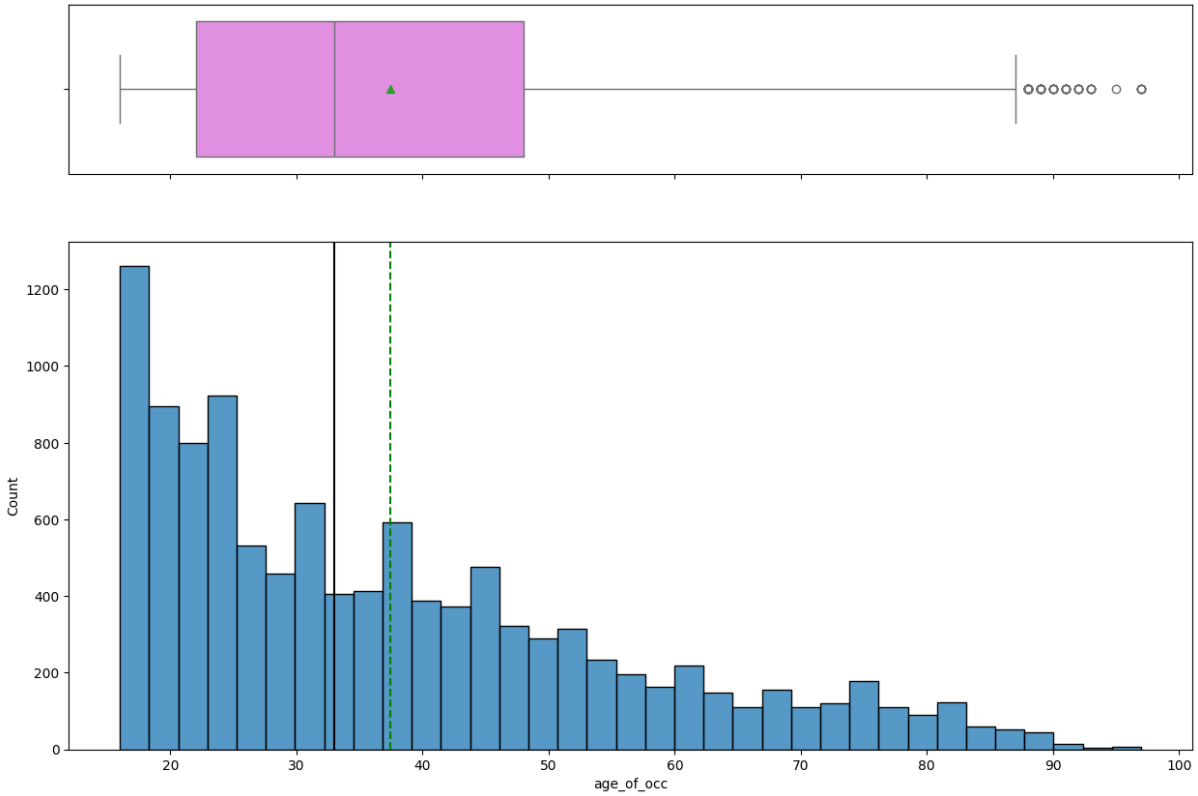
Bottom Plot (Histogram): Shows a **heavily right-skewed distribution**. The majority of vehicle weights are very low (peaking near 0), while a small number of values extend beyond 5,000 and even up to 30,000.

The presence of **extreme outliers** in weight could distort model performance.

Consider applying a **log transformation** or **capping extreme values** to stabilize the distribution before modeling.

The weight column has some **very large values** that are **not common**. These extreme values can **confuse the model**, so we may need to **remove them or fix them** before training.

Observations on age of occ:

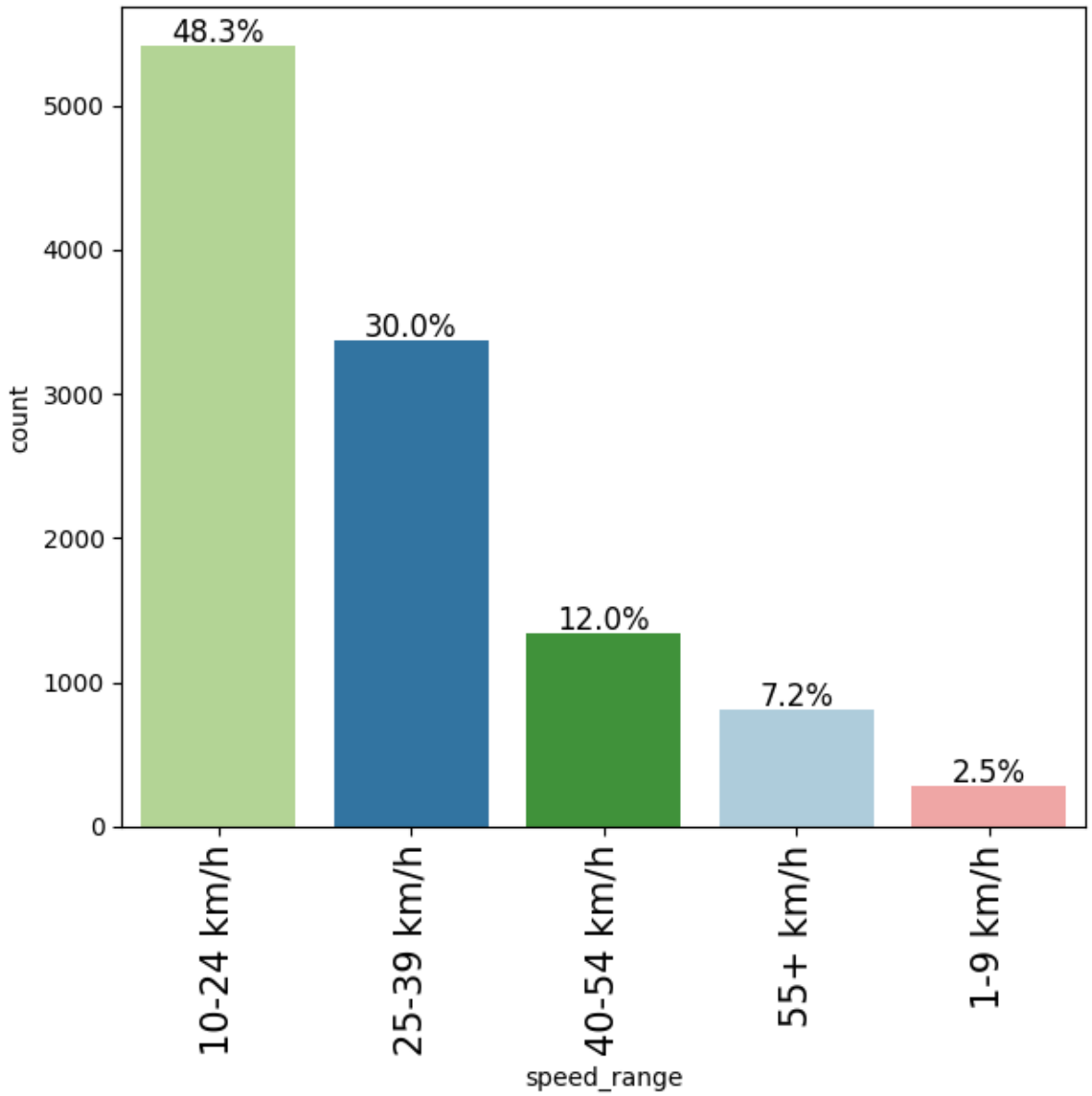


Top (Boxplot): Most occupants are between ~20 and ~60 years old. A few people above ~80 are considered outliers.

Bottom (Histogram): Most car crash occupants are young, especially in the 15–30 age range. As age increases, the number of people involved in crashes decreases**.**

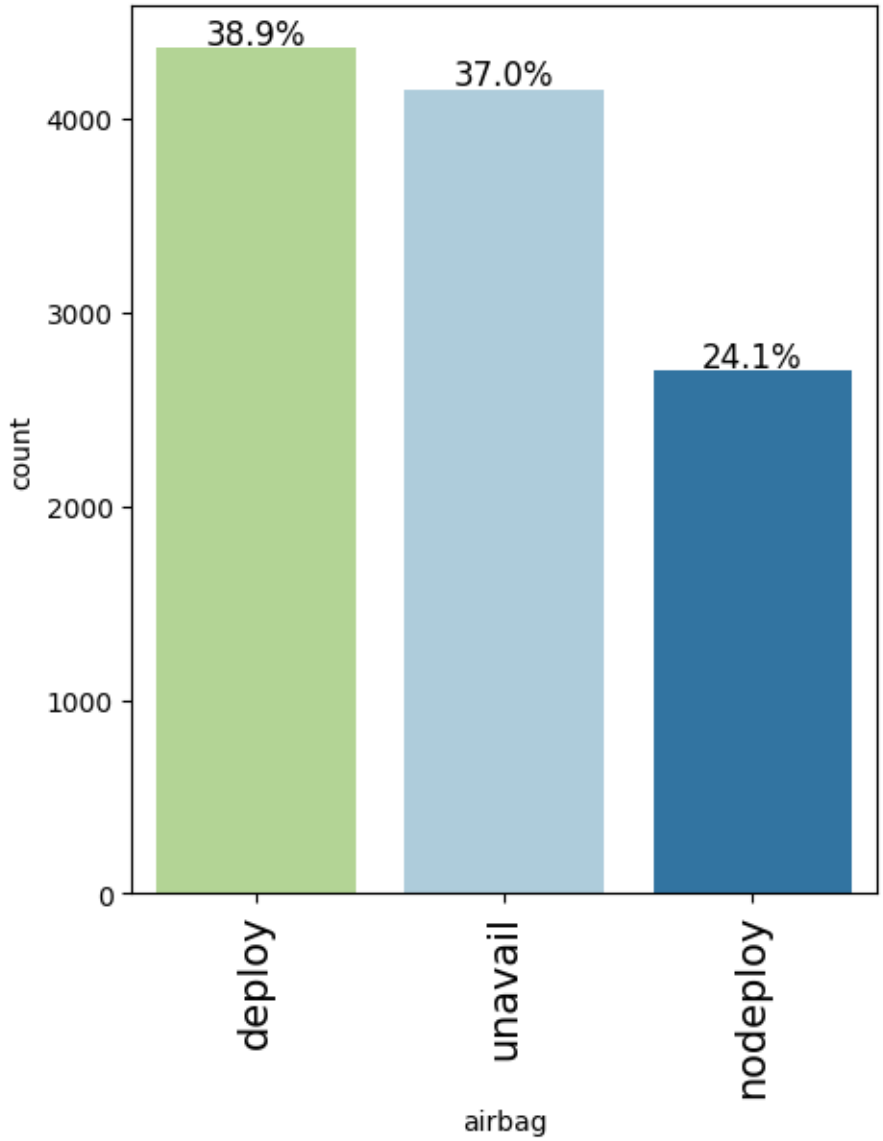
Car crashes mostly involve younger individuals, but the full age range spans from 16 to 97. A few very old ages appear as outliers and may need special attention during analysis.

Observations On Speed Range:



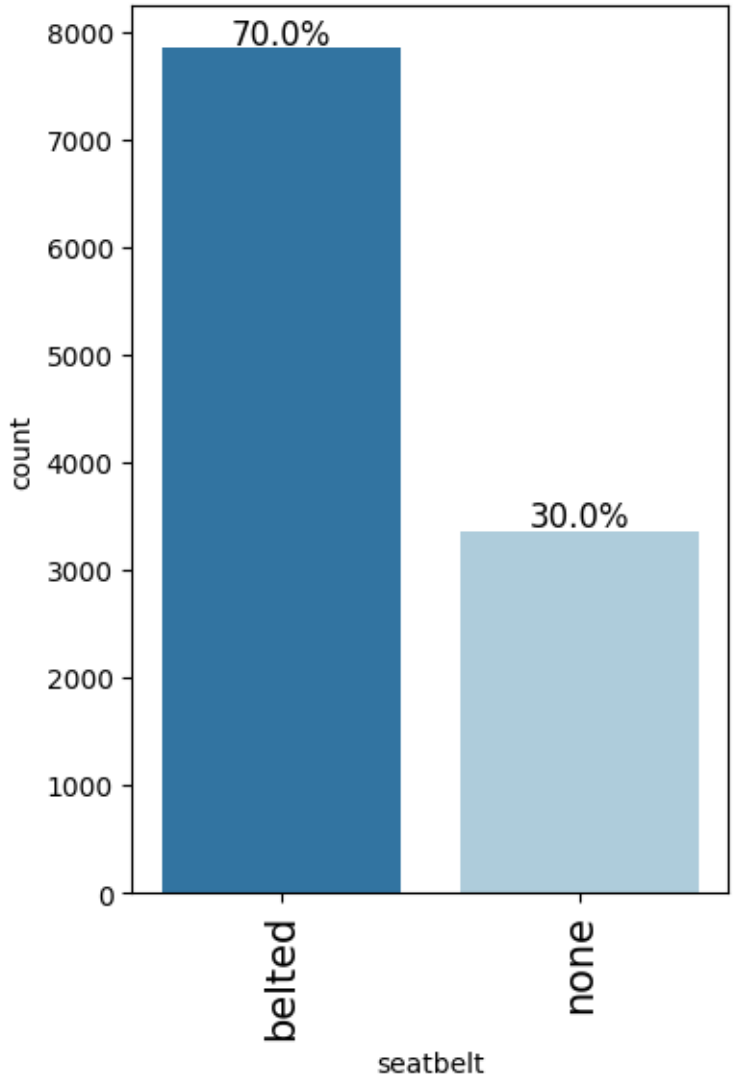
The speed range of 10-24 km/h is high comparing to the 40-50km/h and 55+ km/h (i.e) 48.3%

Observations on airbag:



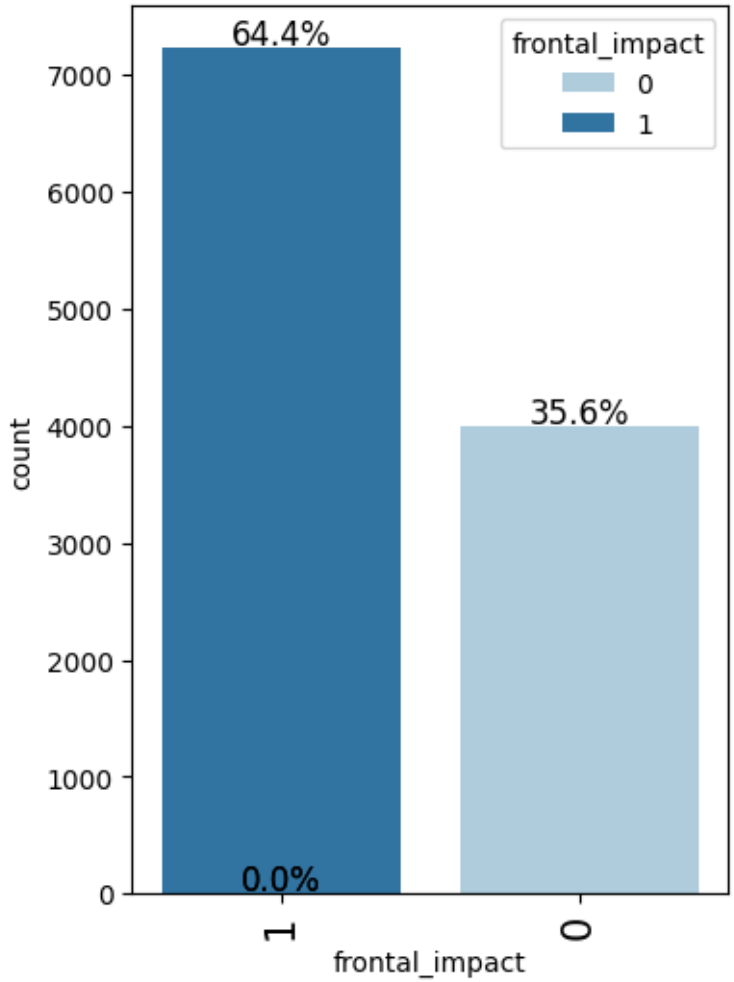
The airbags were deployed in car crashes are more than unavailed and nodeploy by 38.9%. where as no deploy were 24.1%.

Observations On Seatbelt:



The cars which were have a seatbelt are 70% and no seatbelts are 30%.

Observations On Frontal Impact:

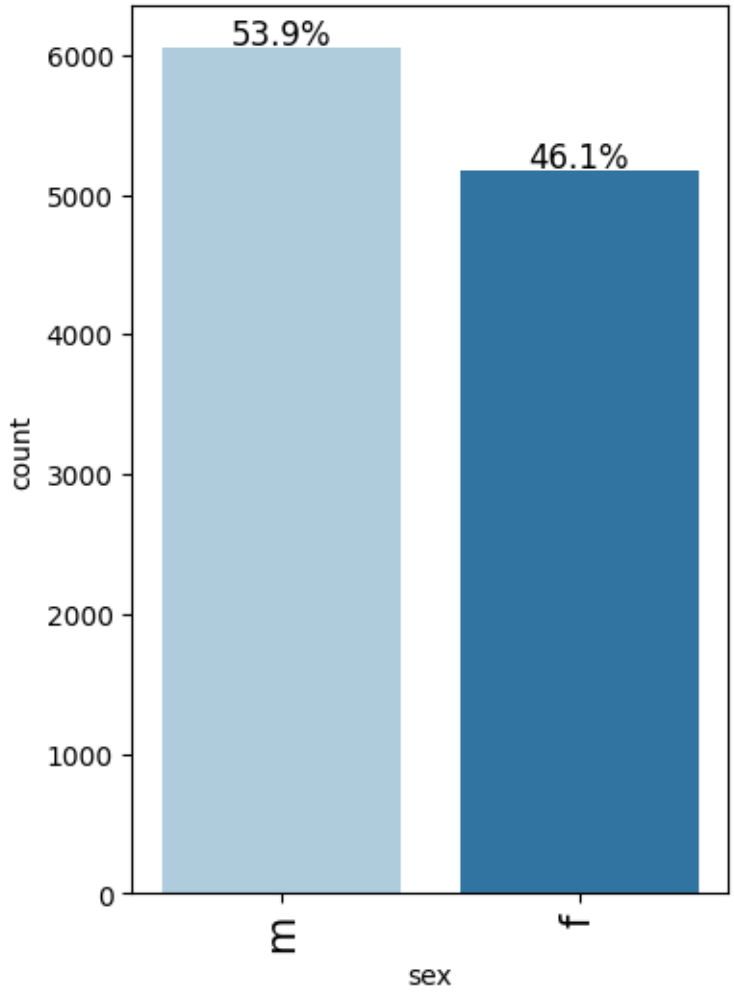


64.4% of crashes involved a frontal impact (value 1)

35.6% did not involve a frontal impact (value 0)

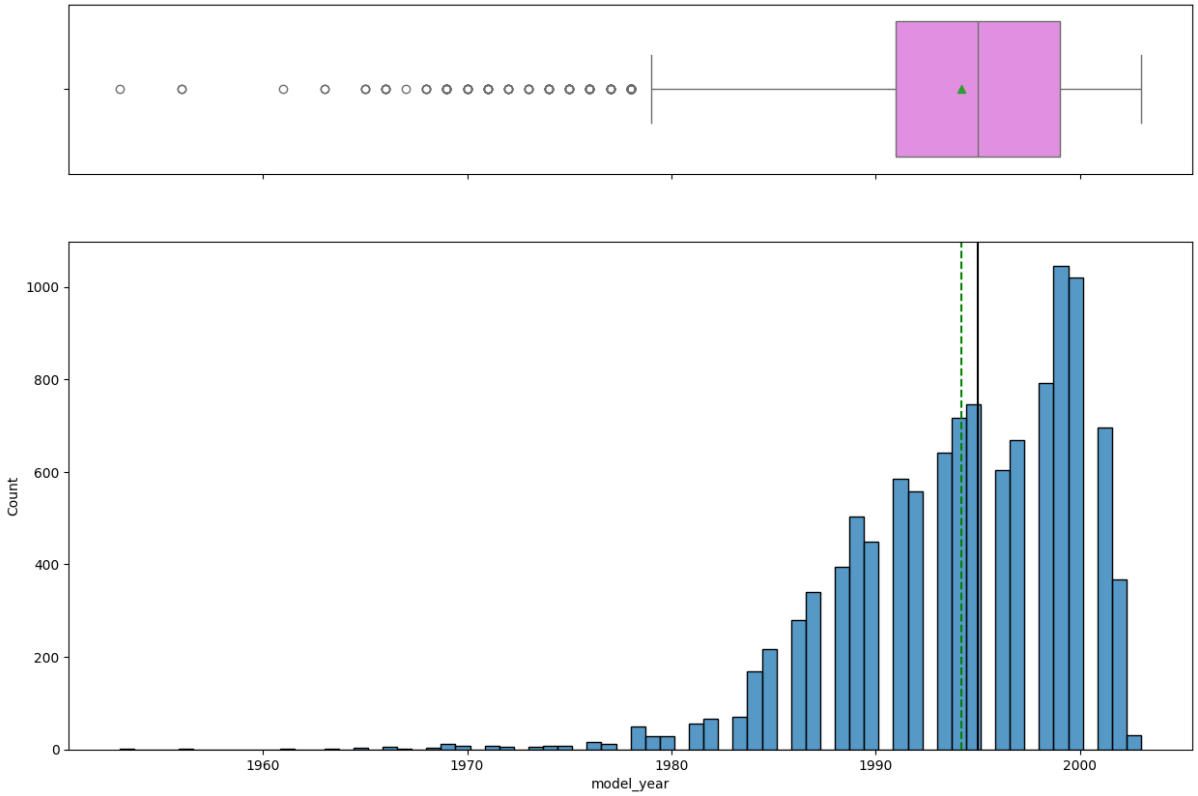
Frontal impacts are more common, making up nearly two-thirds of all crash cases. This suggests frontal collisions are a major crash type and should be a focus for safety improvements.

Observations On Sex:



The male percentage is more than the women who drove the cars by 53.9%.

Observations On Model Year:

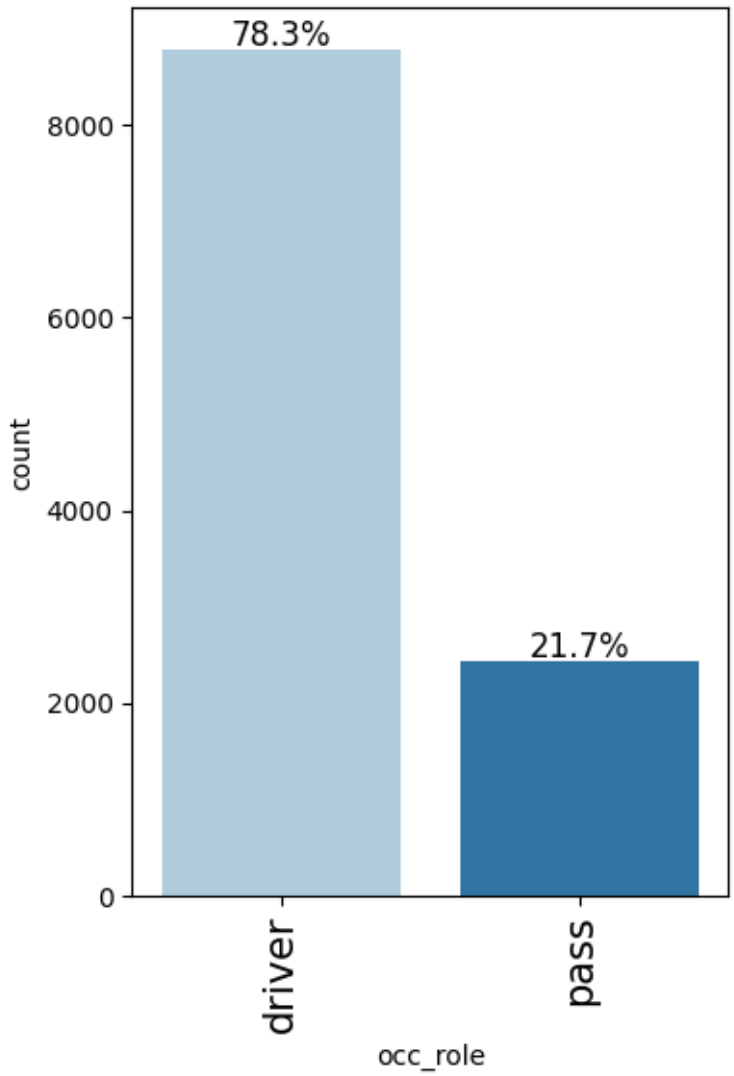


Top Boxplot: Most cars are from the 1990s to early 2000s, but some very old models (from the 1950s–70s) appear as outliers.

Bottom Histogram: Shows that most crashes involved cars built between 1990 and 2002, with peaks around 1998–2000.

Most vehicles in crashes are from recent decades, but a few very old cars are still present. These old models may lack modern safety features and could impact survival outcomes.

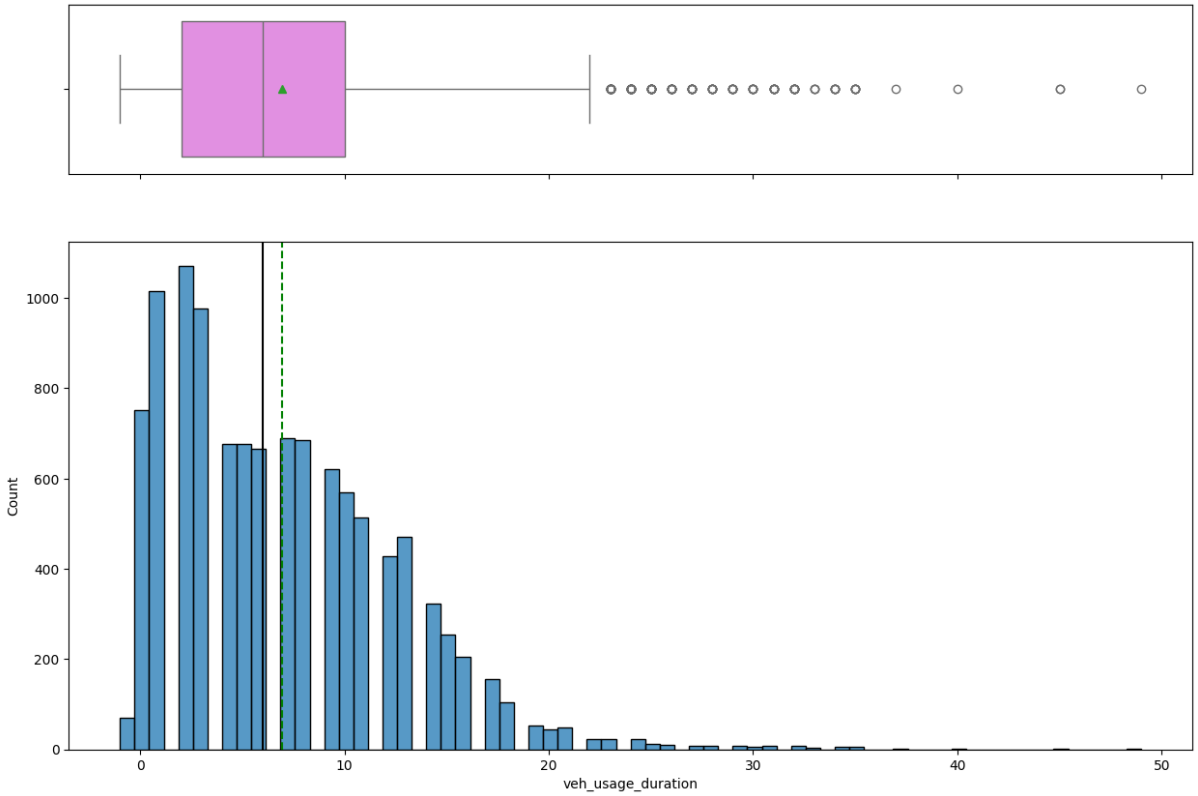
Observations On Occ Role:



The bar chart shows that 78.3% of crash victims were drivers, while only 21.7% were passengers.

Most crashes involve drivers, possibly because they are always present in a moving vehicle. This suggests driver-related factors (like age, seatbelt use, etc.) are especially important in survival analysis.

Observations On Veh Usage Duration:



- **Top Boxplot:** The box (in purple) shows the interquartile range (middle 50% of values). The triangle marker likely represents the mean. Many outliers are visible on the right, indicating some vehicles are used for much longer durations.

Bottom Histogram: 🔍 Shows the frequency distribution of veh_usage_duration. The data is right-skewed (most values are low, few are very high).

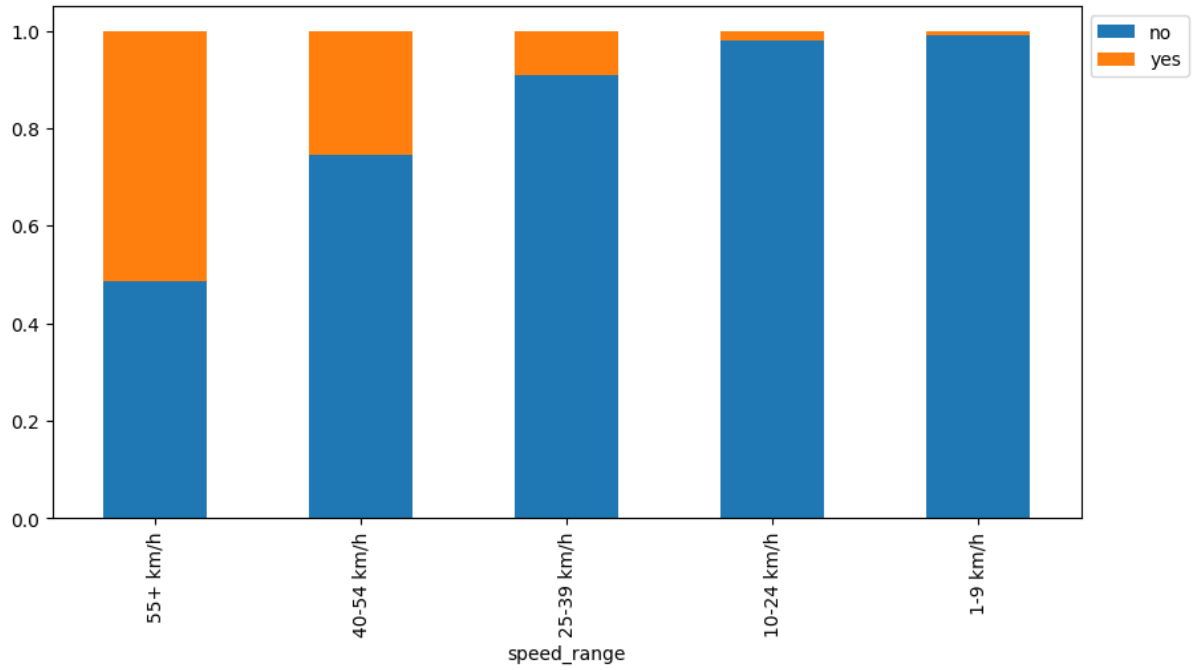
Two vertical lines:

- Solid black: median.
- Dashed green: mean.

Most vehicles have short usage durations, but a few are used for much longer, pulling the mean to the right.

Bivariate Analysis:

Stacked Barplot speed range and deceased:



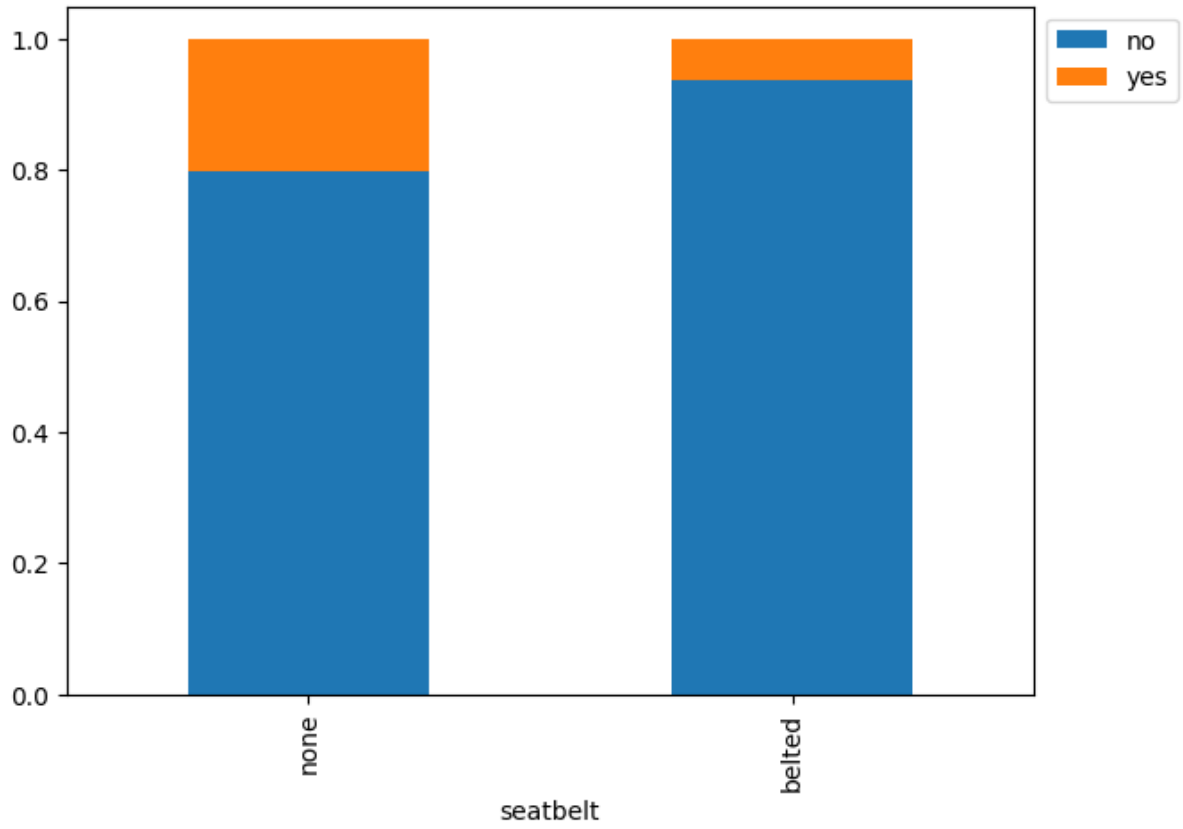
In this stacked barplot where comparing the speed range and deceased we can notice that the higher the speed limit is the greater the car accidents.

For 55+km/h the deceased people are greater than the remaining speed ranges.

In cocnclusion the higher the speed is the higher deceased (Car Crashes).

deceased	no	yes	All
speed_range			
All	10037	1180	11217
55+ km/h	394	415	809
40-54 km/h	1000	344	1344
25-39 km/h	3064	304	3368
10-24 km/h	5300	114	5414
1-9 km/h	279	3	282

Stacked Barplot Seat Belt and deceased:



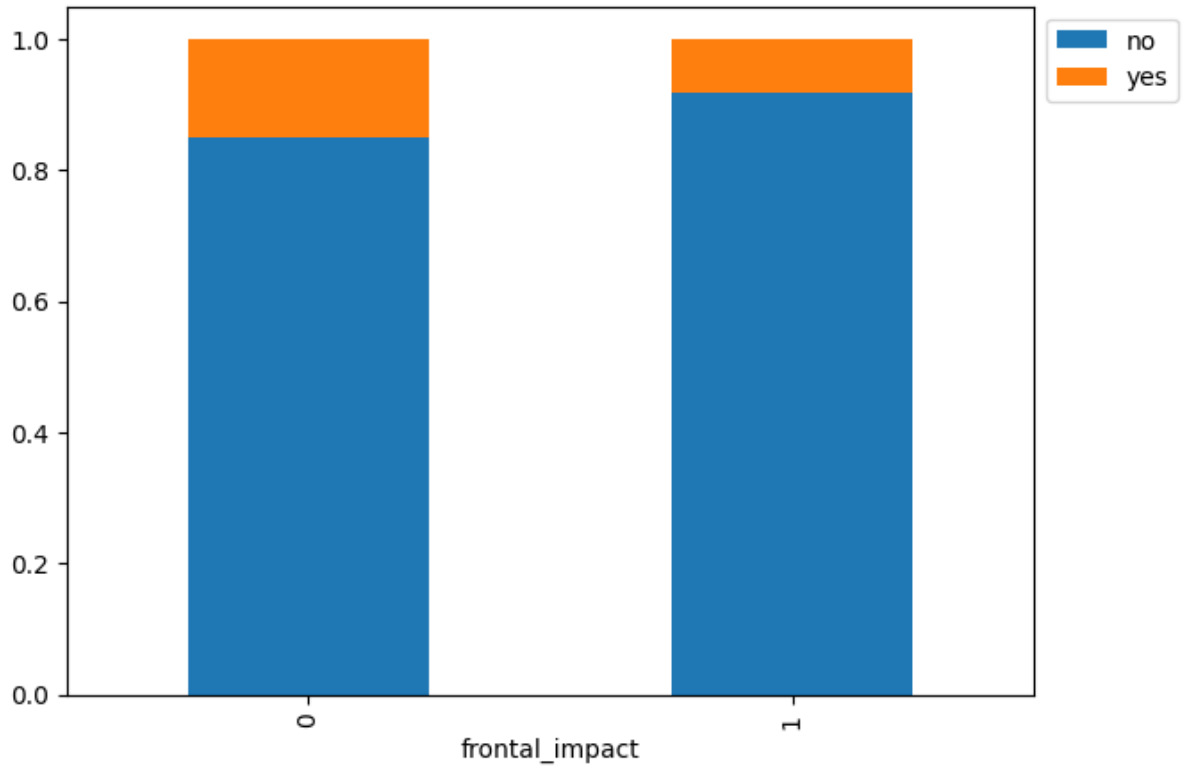
The one who puts seatbelt and drove were less deceased while comparing with the none. So, the seatbelt saved people in those accidents

Higher death rate without seatbelt:

- **None**: 680 deaths out of 3368 is 20.2%
- **Belted**: 500 deaths out of 7849 is 6.4%

Wearing a seatbelt significantly reduces the likelihood of death in accidents.

Stacked Barplot Frontal Impact and deceased:



deceased	no	yes	All
frontal_impact			
All	10037	1180	11217
0	3395	598	3993
1	6642	582	7224

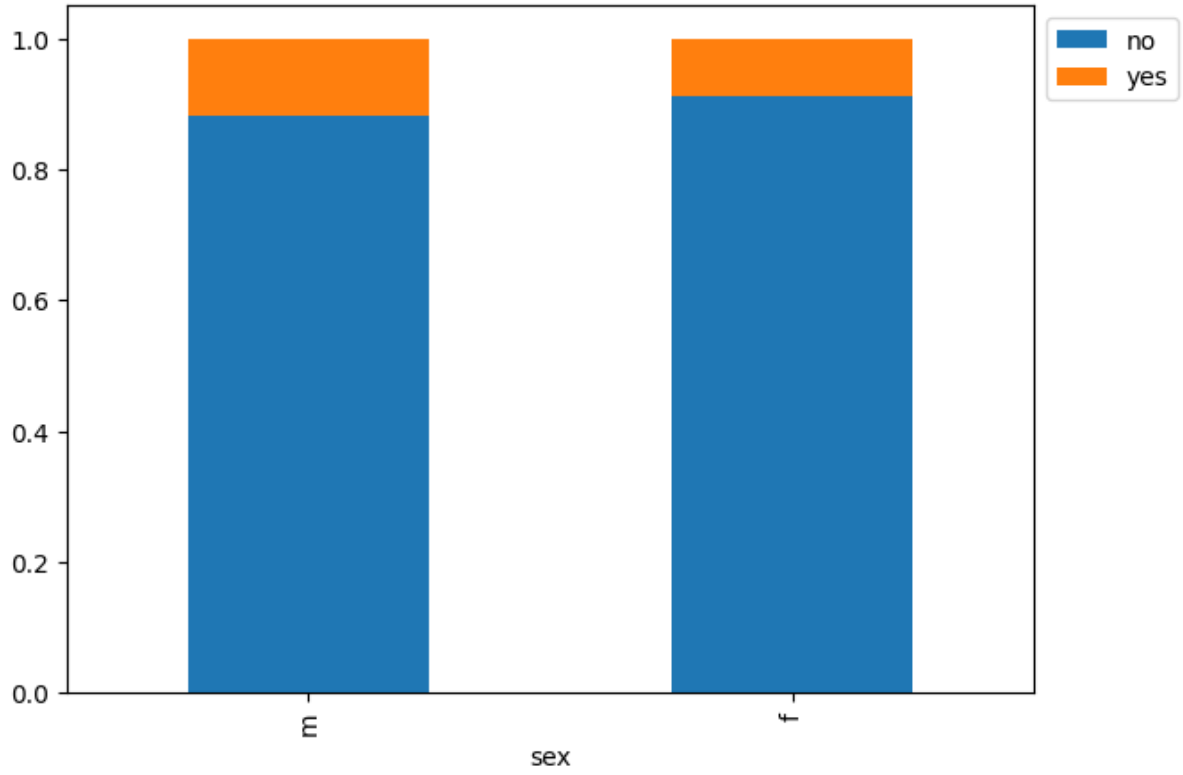
Death rate by frontal impact:

- **No frontal impact (0)**: 598 deaths out of 3993 is 15.0%
- **Frontal impact (1)**: 582 deaths out of 7224 is 8.1%

Frontal impacts are associated with a lower death rate compared to other types of impacts in this dataset.

Stacked Barplot sex and deceased:

deceased	no	yes	All
sex			
All	10037	1180	11217
m	5332	716	6048
f	4705	464	5169



Death rate by sex:

- **Male**: 716 deaths out of 6048 is 12.0%
- **Female**: 464 deaths out of 5169 is 9.0%

By sex Female are the one with low death rate compared to the men death rate.

Stacked Barplot airbag and deceased:

deceased no yes All

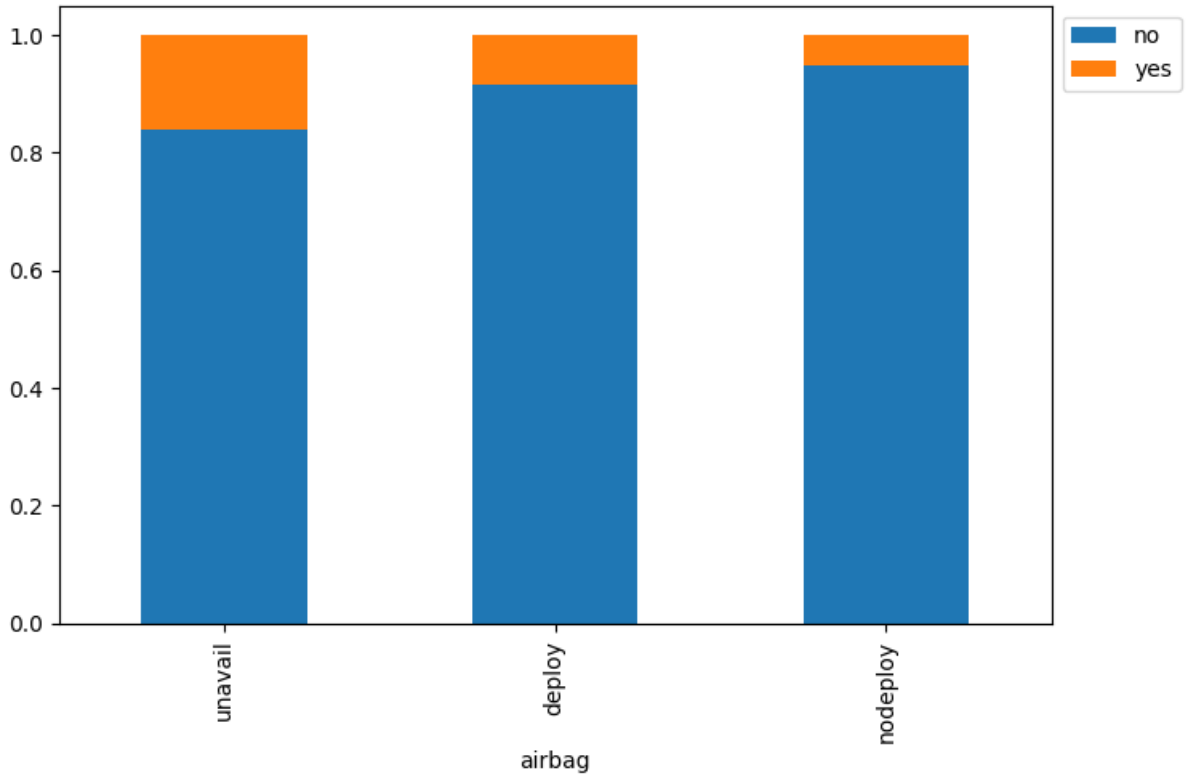
airbag

All 10037 1180 11217

unavail 3484 669 4153

deploy 3997 368 4365

nodeploy 2556 143 2699



Death rate by airbag:

- **Unavail**: 669 deaths out of 4153 is 16.0%
- **deploy**: 368 deaths out of 4365 is 8.4%
- **Nodeploy:** 143 deaths out of 2699 is 5.0%

Death rate by unavail airbags are high while comparing to the deploy and nodeploy

Stacked Barplot occ_role and deceased:

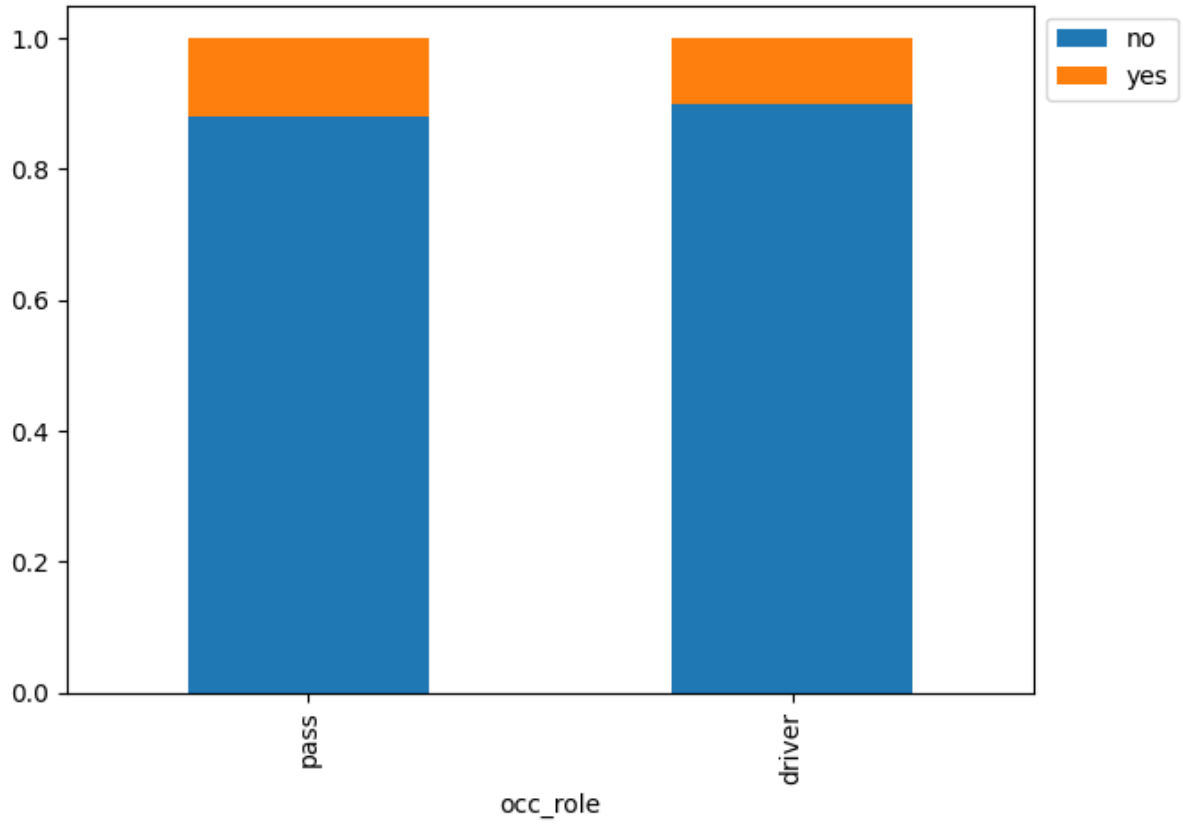
deceased no yes All

occ_role

All 10037 1180 11217

driver 7895 891 8786

pass 2142 289 2431

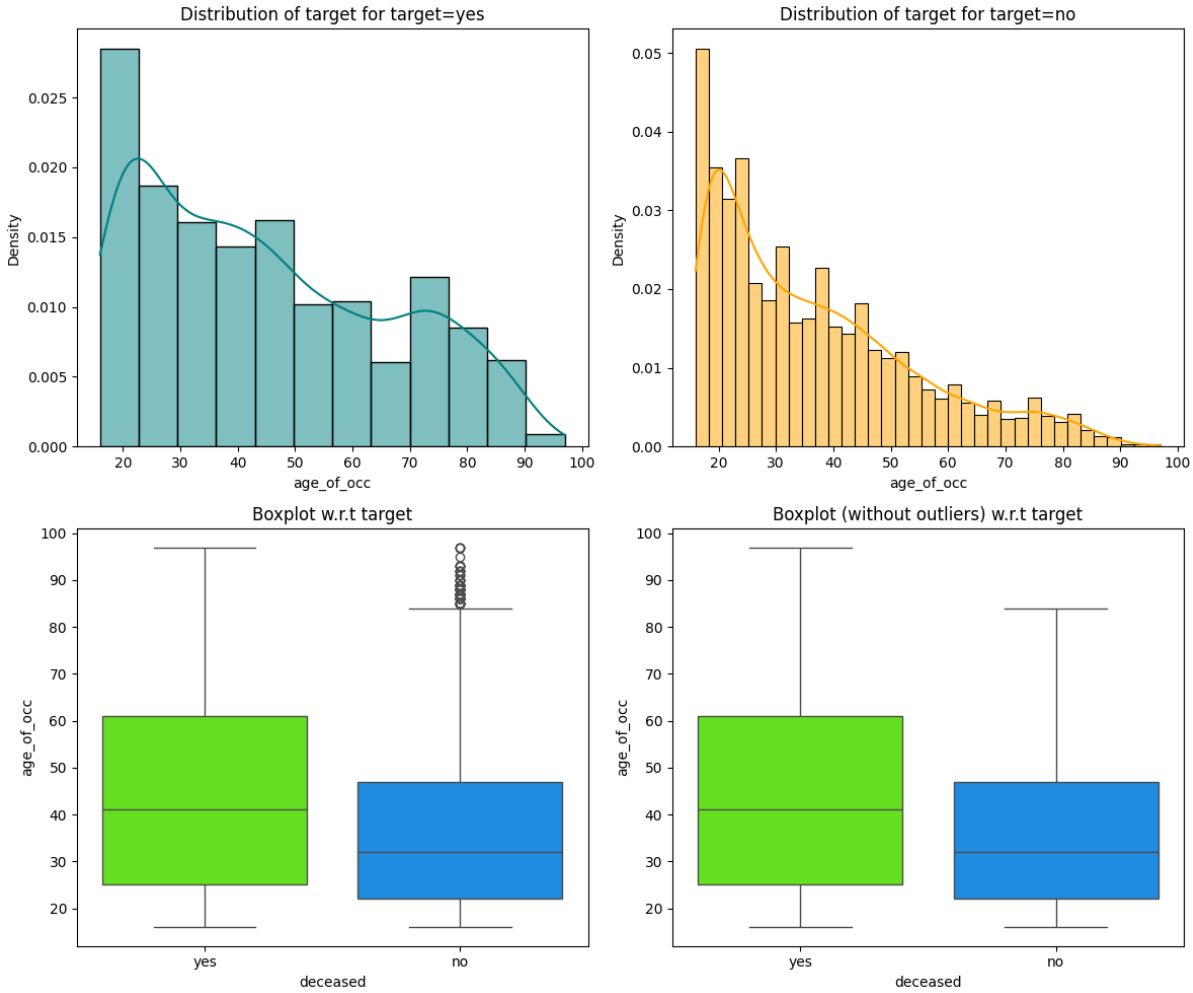


Death rate by airbag:

- **Pass**: 289 deaths out of 2431 is 12.0%
- **driver**: 891 deaths out of 8786 is 10.0%

The passengers death rate is more comparing to the drivers deathrate.

Distribution_plot_wrt_target("age_of_occ", "deceased"):



Higher age → higher death risk:

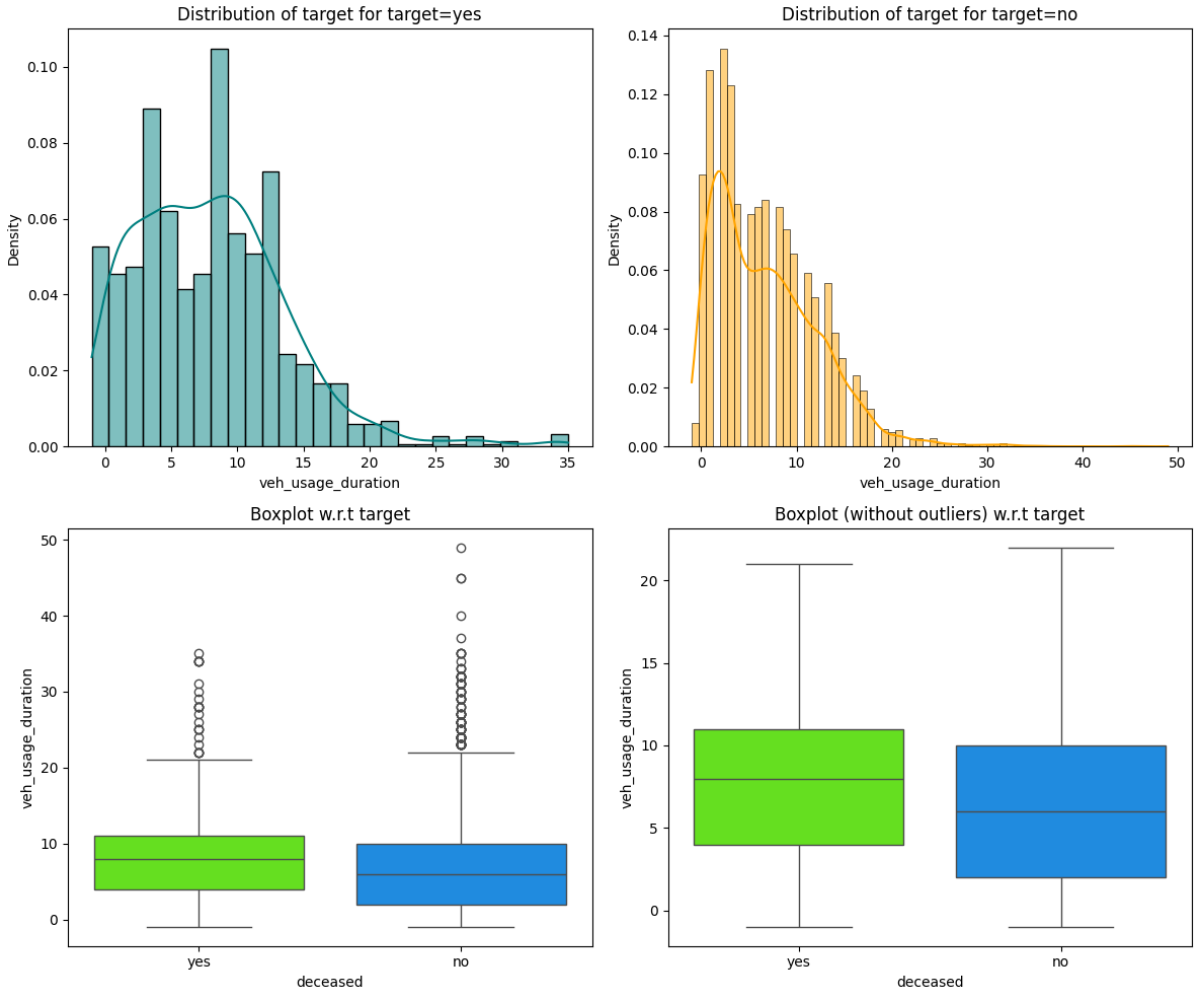
- Median age of deceased ("yes") is significantly higher than non-deceased ("no").
- Density plots show more elderly fatalities compared to younger age groups.

Younger people survive more:

- Most survivors are concentrated in the 20–40 age range.

Age is a strong factor — older occupants have a much higher chance of fatality in accidents

Distribution_plot_wrt_target("veh_usage_duration", "deceased")



Higher vehicle usage duration (age of vehicle) is linked to higher fatality:

- Median usage duration is higher for deceased cases.
- Density plots show longer-used vehicles are more common in fatal cases.

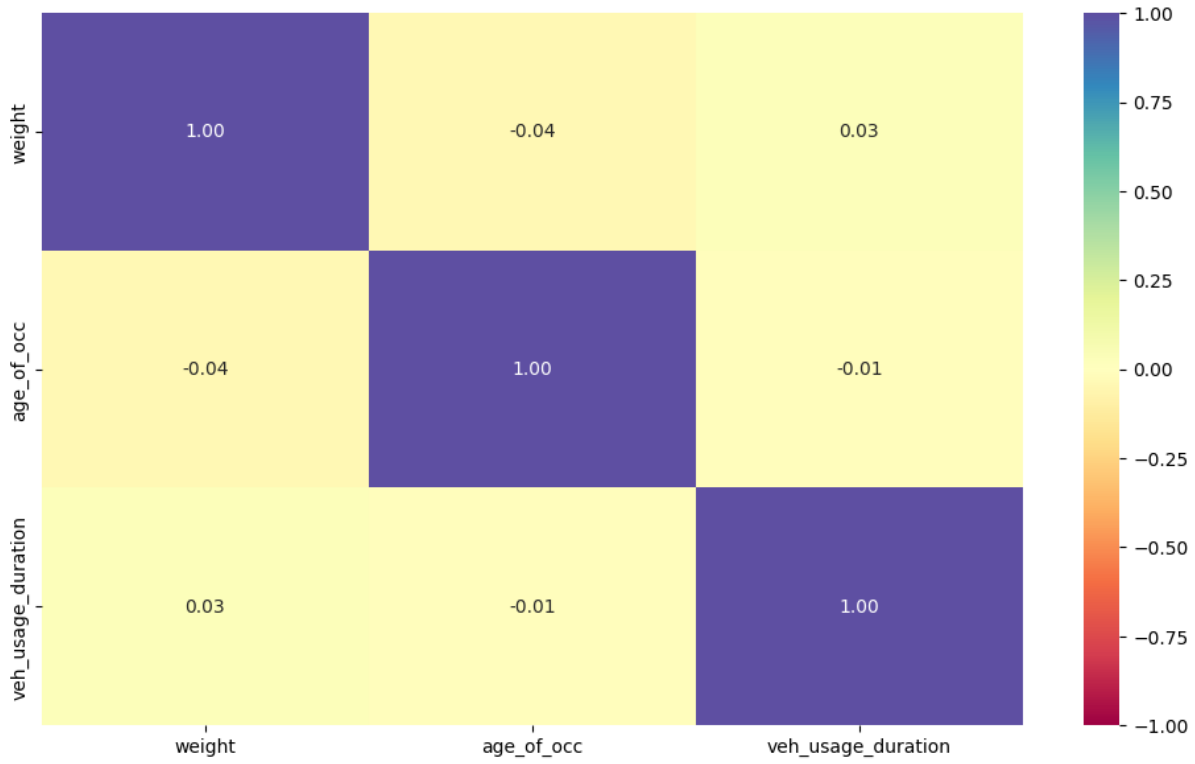
Newer vehicles (low usage duration) are safer:

- Non-deceased cases are more concentrated in the 0–5 years range.

Older vehicles (longer usage duration) **are associated with more fatalities**. Regular maintenance or using newer vehicles could lower fatality risk.

Correlation Heatmap:

This heatmap shows Pearson correlation coefficients between weight, age_of_occ, and veh_usage_duration.



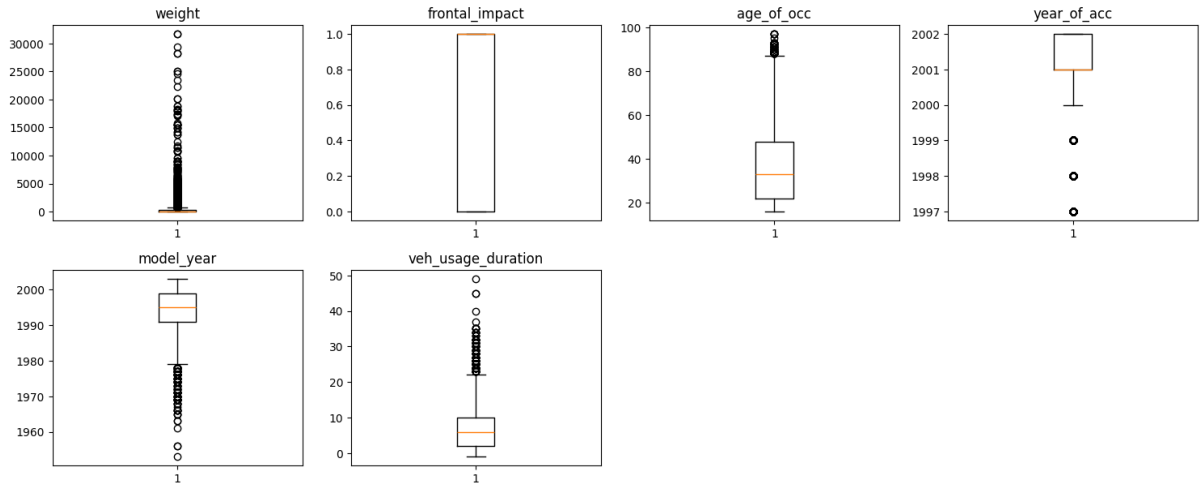
No strong correlations exist among these features:

- weight vs age_of_occ: **-0.04** (very weak negative correlation)
- weight vs veh_usage_duration: **0.03** (very weak positive correlation)
- age_of_occ vs veh_usage_duration: **-0.01** (essentially no correlation)

These features are **largely independent**, which is good for modeling — they provide **unique, non-redundant information**.

Despite low correlations, each variable may still contribute to prediction in nonlinear or interaction-based models like decision trees or random forests.

5. Data Preprocessing



1. weight

- Outliers: Very significant outliers above 10,000 and even beyond 30,000.
- Distribution: Positively skewed. Most data is tightly packed under ~5,000.
- Recommendation: Consider log transformation or outlier capping.

---

2. frontal_impact

- Binary Variable: Takes values close to 0 or 1 only.
- Interpretation: Could be encoded as a categorical/binary variable for modeling.

---

3. age_of_occ

- Spread: Wide range (approx. 15 to 100).
- Outliers: Minimal; most values lie below 90.
- Skew: Mild right skew, but generally manageable.

---

4. year_of_acc

- Tight Range: Concentrated around 2001.
- Outliers: Few low outliers (1997–1999).
- Insight: Data likely focuses on a short window of accident years.

---

5. model_year

- Skewed Left: Some vehicles go back to the 1960s.
- Bulk Data: From 1980s–2000.
- Outliers: Present on the lower end (1950s–70s).
- Suggestion: Consider bucketing older years or treating it as an age feature (year_of_acc - model_year).

---

6. veh_usage_duration

- Outliers: Present (values up to ~50).
- Skew: Positively skewed.
- Suggestion: Potential for normalization or log transformation for better model fit.

Overview

The Department of Road Transport has seen a 15% increase in urban car crashes over the past year. To improve road safety, we analyzed a dataset of car crashes over five years to predict survival outcomes and identify key factors. This report explains how we prepared the data for analysis in a straightforward way, ensuring it's clean and ready for building predictive models.

What We Did

We used a dataset with 11,217 records about car crashes, including details like speed, seatbelt use, airbag deployment, and whether the occupant survived. Here's how we got the data ready:

1. **Loading the Data**
    - We loaded the data from a file called Car_Crash.csv using Python.
    - We made a copy of the data to keep the original safe.
    - We checked the first and last few rows to make sure everything looked correct.

Cleaning the Data

- **Missing Values**: We checked for any missing information. There were none, so we didn't need to fill in or remove anything.
- **Duplicates**: We looked for duplicate records and found none, confirming each crash was unique.
- **Data Types**: We made sure the data types were correct, like text for categories (e.g., seatbelt status) and numbers for things like age or year.

Handling Categories

To make the data usable for modeling, we converted text-based categories into numbers:

- **Yes/No Categories**:
    - Seatbelt (none = 0, belted = 1)
    - Sex (female = 0, male = 1)
    - Role (passenger = 0, driver = 1)
    - Deceased (no = 0, yes = 1, our target variable)
- **Speed Range**: We turned speed ranges into numbers (e.g., 1-9 km/h = 1, 10-24 km/h = 2, up to 55+ km/h = 5) since higher speeds are riskier.

  Airbag: We split airbag status (deploy, nodeploy, unavail) into separate columns to avoid assuming any order.
- We removed the caseid column since it's just an ID and not useful for predictions.

Handling Numbers

- We standardized numerical columns (like weight, age, and car model year) to put them on the same scale, which helps models like logistic regression work better.
- We checked for outliers (extreme values) using summaries and plots. Everything seemed reasonable, so we kept all data.

Splitting the Data

- We divided the data into 80% for training (8,973 records) and 20% for testing (2,244 records).

- We made sure the proportion of survivors vs. non-survivors was the same in both sets to avoid bias.
- We set a random seed to make the split repeatable.

Keeping Weights

- The dataset included a weight column to adjust for sampling differences. We kept it to ensure our models give fair results.

These steps ensured the data was clean and ready for analysis:

- Cleaning removed any errors or inconsistencies.
- Converting categories to numbers made the data work with machine learning models.
- Standardizing numbers helped models perform better.
- Splitting the data fairly allowed us to test our models accurately, especially since fewer crashes result in deaths (an imbalanced dataset).
- The cleaned and prepared data supported building models like logistic regression and decision trees. It allowed us to accurately predict survival chances and identify key factors like speed or seatbelt use, which will help the Department of Road Transport create better safety rules.

6.Model Building:

1. Models We Created

We built two types of models: **logistic regression** (like a math formula to predict outcomes) and **decision trees** (like a flowchart making yes/no decisions). For each, we made a basic version and an improved version.

- **Logistic Regression**:
    - **Basic Version**: Used standard settings to predict if someone survived (0) or died (1).
    - **Improved Version**: Adjusted the model to better catch cases where someone died, since there are fewer of those.
- **Decision Tree**:
    - **Basic Version**: Made predictions but learned too much from the training data (overfitting).
    - **Improved Version**: Tweaked settings (like tree size) to make better predictions on new data.

- `Optimization terminated successfully.`
- `         Current function value: 0.207203`
- `         Iterations 11`
- `                           Logit Regression Results`
- `==============================================================================`
- `Dep. Variable:               deceased   No. Observations:  7851`
- `Model:                           Logit   Df Residuals:  7837`
- `Method:                            MLE   Df Model:  13`
- `Date:                Fri, 23 May 2025   Pseudo R-squ.:  0.3820`
- `Time:                        06:19:33   Log-Likelihood:  -1626.8`
- `converged:                       True   LL-Null:  -2632.4`
- `Covariance Type:            nonrobust   LLR p-value:  0.000`
- `==============================================================================`
- `                             coef    std err          z      P>|z|      [0.025      0.975]`
- `------------------------------------------------------------------------------`

- const -4.6071 0.194 -23.731 0.000 -4.988 -4.227
- weight -5.9875 0.691 -8.659 0.000 -7.343 -4.632
- frontal_impact -0.6551 0.050 -13.140 0.000 -0.753 -0.557
- age_of_occ 0.6916 0.045 15.484 0.000 0.604 0.779
- veh_usage_duration -0.5009 0.071 -7.071 0.000 -0.640 -0.362
- speed_range_10-24 km/h 0.0836 0.302 0.277 0.782 -0.508 0.675
- speed_range_25-39 km/h 0.6594 0.273 2.416 0.016 0.124 1.194
- speed_range_40-54 km/h 0.8627 0.195 4.426 0.000 0.481 1.245
- speed_range_55+ km/h 1.0221 0.159 6.426 0.000 0.710 1.334
- seatbelt_none 0.5442 0.044 12.404 0.000 0.458 0.630
- sex_m 0.0404 0.047 0.853 0.394 -0.052 0.133
- airbag_nodeploy -0.0678 0.064 -1.056 0.291 -0.194 0.058
- airbag_unavail 0.6360 0.072 8.871 0.000 0.496 0.777
- occ_role_pass 0.0653 0.044 1.475 0.140 -0.021 0.152

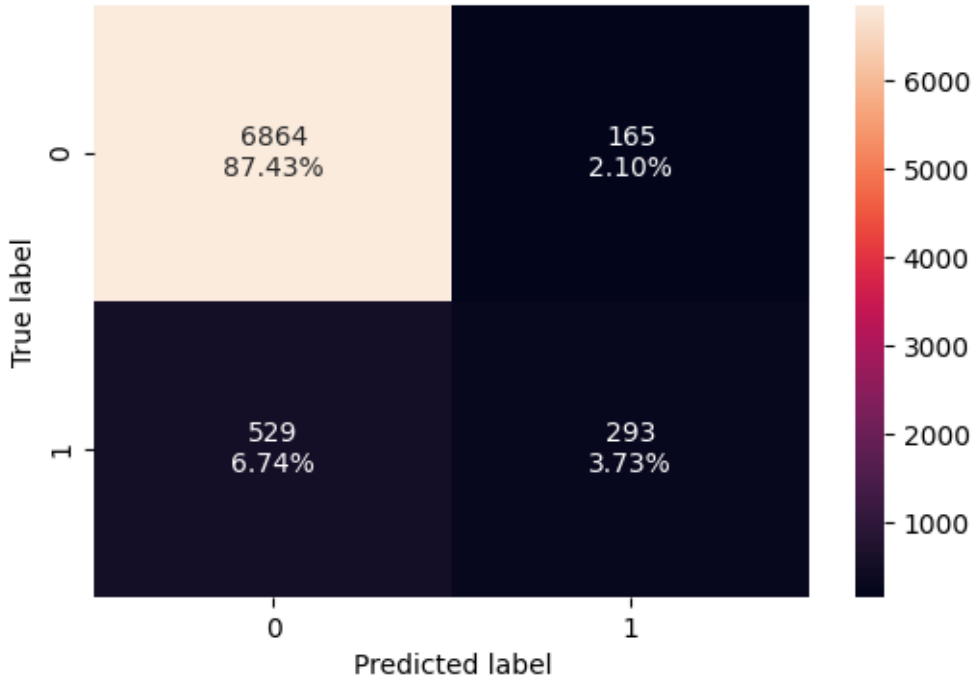
2. How We Tested the Models

We checked how well each model worked using four scores:

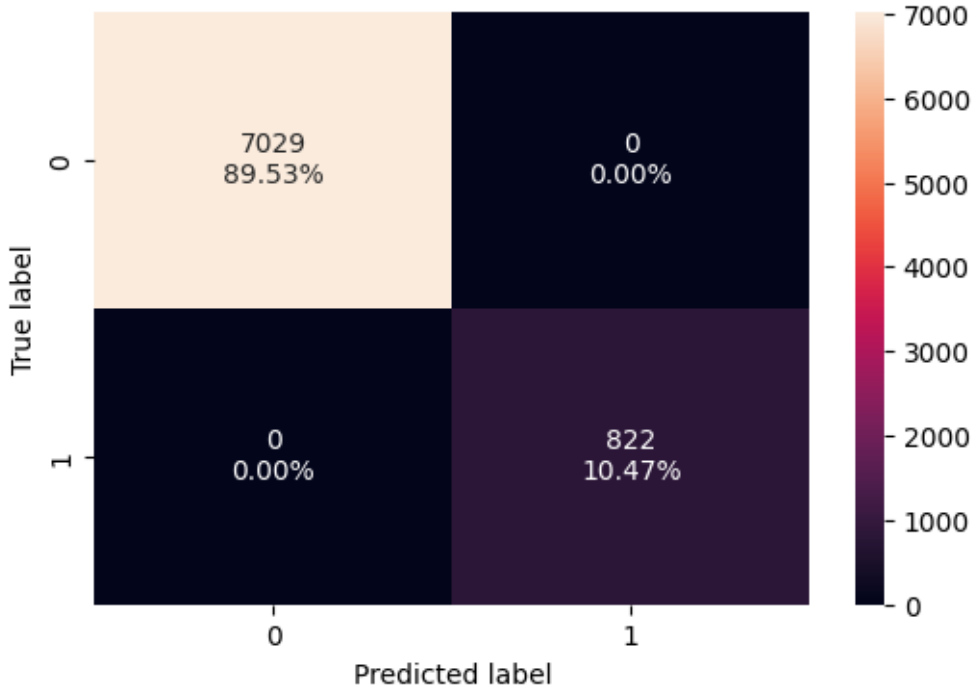
- **Accuracy**: How often the model is right overall.
- **Recall**: How well it finds cases where someone died (important since these are rare).
- **Precision**: How accurate it is when it predicts someone died.
- **F1-Score**: A balance of recall and precision.

Here's how the models did:

Training Results (on the data used to build the models):

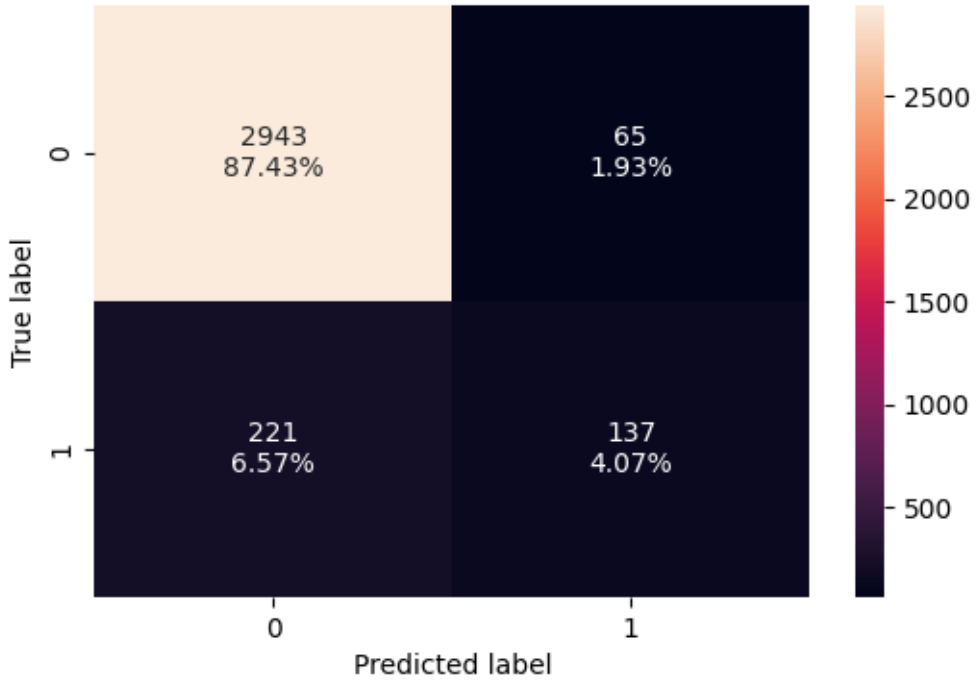


Decision tree train

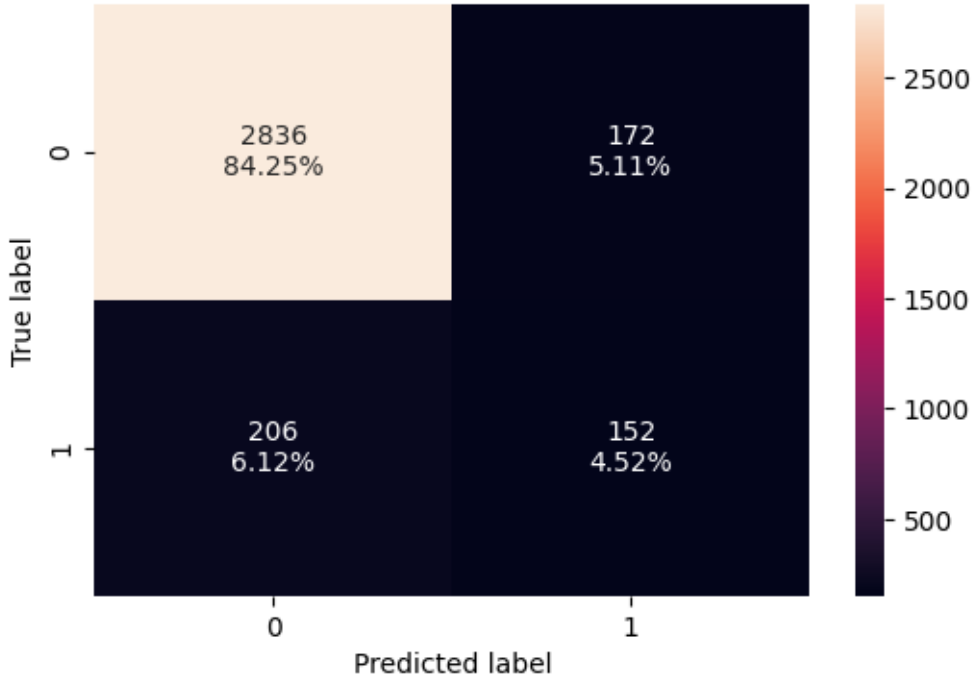


Model	Accuracy	Recall	Precision	F1-Score
Logistic Regression (Basic)	91%	36%	46%	40%
Logistic Regression (Improved)	83%	86%	36%	51%
Decision Tree (Basic)	100%	100%	100%	100%
Decision Tree (Improved)	87%	80%	30%	44%

Test Results (on new, unseen data):



Decision Tree Test:



Model	Accuracy	Recall	Precision	F1-Score
Logistic Regression (Basic)	92%	38%	68%	49%
Logistic Regression (Improved)	83%	85%	36%	51%

Model	Accuracy	Recall	Precision	F1-Score
Decision Tree (Basic)	89%	42%	47%	45%
Decision Tree (Improved)	78%	85%	30%	45%

What This Means:

- The basic decision tree was perfect on training data but worse on test data (it overlearned).
- The improved logistic regression did best on the test data, with a high F1-score (51%) and recall (85%), meaning it's good at spotting cases where someone might die.
- The improved decision tree also had high recall but lower precision, so it made more mistakes when predicting deaths.

3. Key Factors

Both models showed that:

- **Crash speed**: Higher speeds increase the chance of death.
- **Seatbelt use**: Not wearing a seatbelt makes death more likely.
- **Age**: Older occupants are at higher risk.
- **Crash type**: Frontal crashes are more dangerous.

4. Best Model

We picked the **improved logistic regression** because:

- It has the highest F1-score (51%) on test data, balancing accuracy and catching deaths.
- It's good at finding cases where someone might die (85% recall).
- It doesn't overlearn like the basic decision tree.
- It's easier to understand which factors matter most.

We built models to predict survival in car crashes, and the improved logistic regression worked best. It showed that high speeds, not wearing seatbelts, older age, and frontal crashes are the biggest risks. These findings can help the Department of Road Transport make rules like stricter speed limits or mandatory seatbelt laws to save lives.

7.Model Performance Improvement:

Optimization terminated successfully.

Current function value: 0.207445

Iterations 11

Logit Regression Results

==========

Dep. Variable: deceased No. Observations: 7851

Model: Logit Df Residuals: 7841

Method: MLE Df Model: 9

Date: Fri, 23 May 2025 Pseudo R-squ.: 0.3813

Time: 06:25:21 Log-Likelihood: -1628.7

converged: True LL-Null: -2632.4

Covariance Type: nonrobust LLR p-value: 0.000

==========

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.6124	0.195	-23.712	0.000	-4.994	-4.231
weight	-6.0426	0.693	-8.724	0.000	-7.400	-4.685
frontal_impact	-0.6369	0.047	-13.648	0.000	-0.728	-0.545
age_of_occ	0.6879	0.045	15.431	0.000	0.601	0.775
veh_usage_duration	-0.5055	0.071	-7.157	0.000	-0.644	-0.367
speed_range_25-39 km/h	0.5918	0.065	9.137	0.000	0.465	0.719
speed_range_40-54 km/h	0.8152	0.049	16.768	0.000	0.720	0.911
speed_range_55+ km/h	0.9828	0.042	23.228	0.000	0.900	1.066
seatbelt_none	0.5488	0.044	12.579	0.000	0.463	0.634
airbag_unavail	0.6686	0.068	9.793	0.000	0.535	0.802

Model Fit:

- **Optimization**: Successfully converged after 11 iterations with a log-likelihood of -1628.7 (function value: 0.207445).
- **Pseudo R-squared**: 0.3813, indicating the model explains 38.13% of the variance in survival outcomes.
- **LLR p-value**: 0.000, confirming the model is statistically significant compared to a null model.

Variables Kept: After removing insignificant variables, the model includes:

- const (intercept), weight (sampling weight), frontal_impact, age_of_occ, veh_usage_duration (vehicle age), speed_range_25-39 km/h, speed_range_40-54 km/h, speed_range_55+ km/h, seatbelt_none, airbag_unavail.
- All have p-values < 0.05, meaning they significantly affect survival.

Coefficients and Impact:

- **Positive Coefficients** (increase death likelihood):
    - speed_range_55+ km/h (0.9828): Highest impact; high-speed crashes are deadliest.
    - age_of_occ (0.6879): Older occupants are more at risk.
    - airbag_unavail (0.6686), seatbelt_none (0.5488): Lack of safety features increases risk.
    - speed_range_25-39 km/h (0.5918), speed_range_40-54 km/h (0.8152): Moderate speeds also raise risk.
    - **Negative Coefficients** (decrease death likelihood):
    - weight (-6.0426): Higher sampling weights correlate with lower death probability.
    - frontal_impact (-0.6369): Surprisingly, frontal impacts may be less deadly (possibly due to safety features).
    - veh_usage_duration (-0.5055): Older vehicles may have different risk profiles.
    - **Significance**:
    - All variables have z-scores with p-values = 0.000, confirming strong statistical significance.
    - Confidence intervals (e.g., speed_range_55+ km/h: [0.900, 1.066]) show precise estimates.

What It Tells Us:

- The model is reliable, explaining about 38% of why people survive or die in crashes.
- It successfully fit the data after 11 steps (iterations), with a final score (log-likelihood) of -1628.7.
- Only the most important factors were kept after dropping ones like gender, low speed (10-24 km/h), airbag non-deployment, and passenger role because they didn't strongly affect survival.

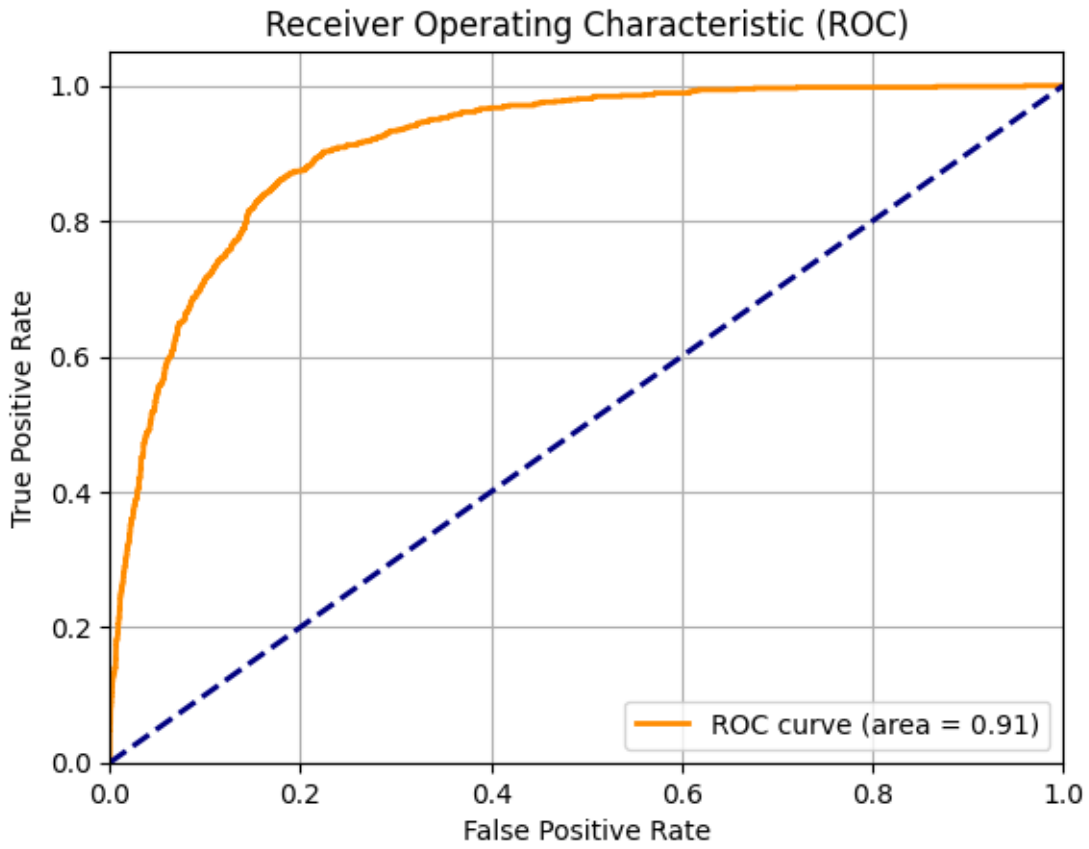
Key Factors That Increase Death Risk:

1. **High crash speed** (55+ km/h): The biggest risk, strongly linked to deaths.
2. **Older age**: Older people (e.g., over 60) are more likely to die.
3. **No seatbelt**: Not wearing a seatbelt makes death more likely.
4. **No airbag**: Crashes without airbags are riskier.
5. **Moderate speeds** (25-39 km/h and 40-54 km/h): These also increase risk, but less than high speeds.

Surprising Findings:

- **Frontal crashes**: Less deadly than expected, possibly due to better car safety features.
- **Older vehicles**: Slightly safer, which might reflect data patterns.

Determining optimal threshold using ROC Curve



Model Accuracy: The curve shows your model is **very accurate**.

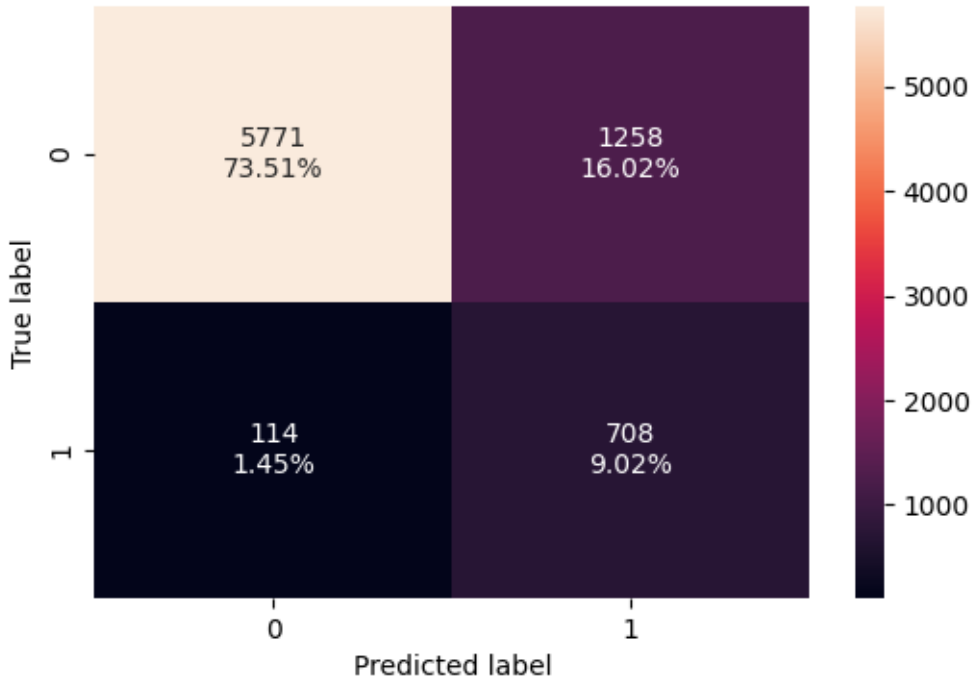
AUC Score = 0.91: This means there's a **91% chance** the model correctly ranks a random positive case higher than a negative one.

c**loser to Top-Left Corner**: This means the model **detects most of the actual positives** with **few false alarms**.

The model is **great at separating the two classes** (like predicting accident risk or not).

A **perfect model** would have AUC = 1.0. Yours is **0.91**, which is **excellent**.

New Logistic Regression model performance on training set:



True Negatives (TN) = 5771 (73.51%)

- These are people who did not survive, and the model correctly predicted non-survival.
- This high number shows the model is good at identifying non-survivors.

False Positives (FP) = 1258 (16.02%)

- These are people who did not survive, but the model incorrectly predicted they would survive.
- These are costly mistakes, especially for safety-critical applications.

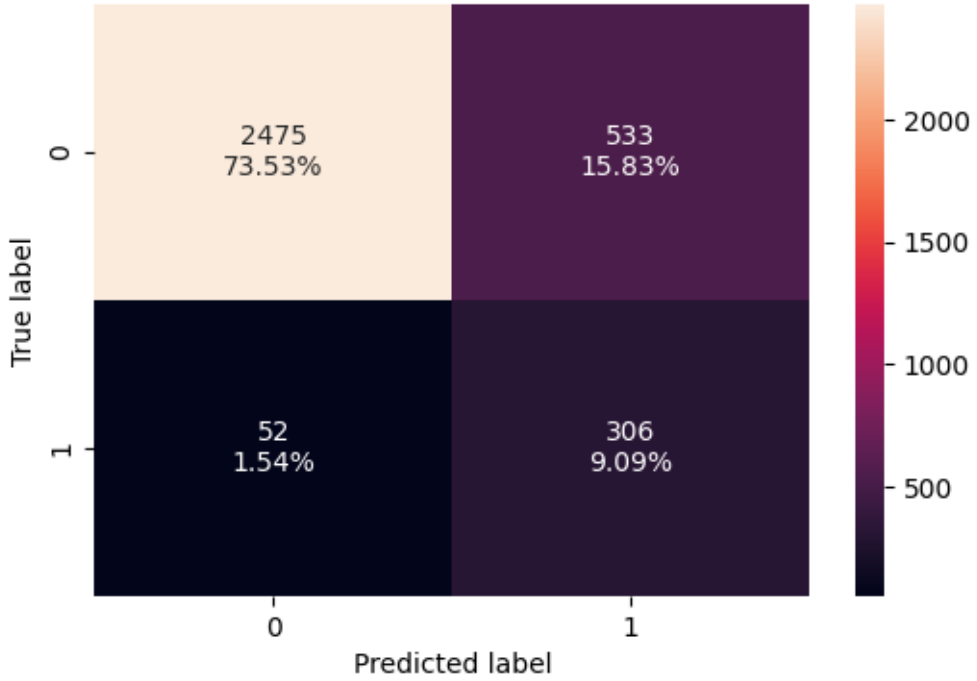
False Negatives (FN) = 114 (1.45%)

- These are people who survived, but the model predicted they would not survive.
- These are very dangerous errors in this context — misclassifying survivors as deceased.

True Positives (TP) = 708 (9.02%)

- These are people who survived, and the model correctly predicted survival.
- This shows the model can detect survivors, though not perfectly.

Tuned Logistic Regression model performance on test set:



1. True Negatives (TN) = 2475 (73.53%)
    - Actual non-survivors correctly predicted as non-survivors.
    - Indicates the model is good at catching non-survivors.
2. False Positives (FP) = 533 (15.83%)
    - Actual non-survivors incorrectly predicted as survivors.
    - These are critical misclassifications in safety applications.
3. False Negatives (FN) = 52 (1.54%)
    - Actual survivors wrongly predicted as non-survivors.

- These are risky errors — the model failed to identify true survivors.

4. True Positives (TP) = 306 (9.09%)
    - Actual survivors correctly predicted as survivors.
    - Reflects the model's success in identifying survivors.

Compared to the **previous matrix** :

The **accuracy** is slightly improved or comparable.

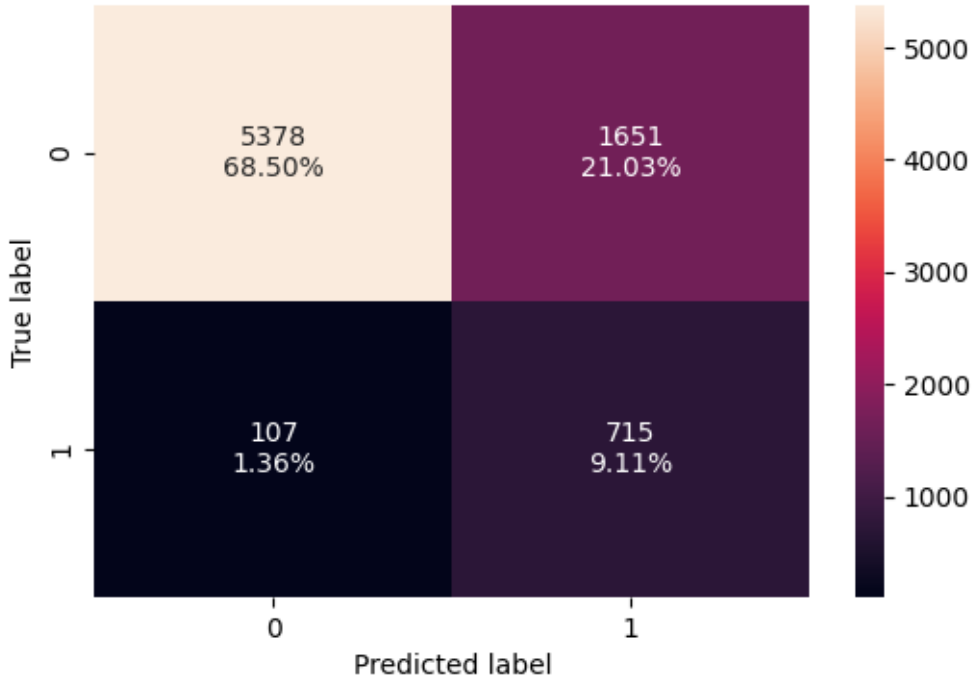
False negatives are **lower** (52 vs. 114), meaning **better survivor identification**.

False positives are **fewer**, indicating **improved precision** slightly (36.5% vs. 36.0%).

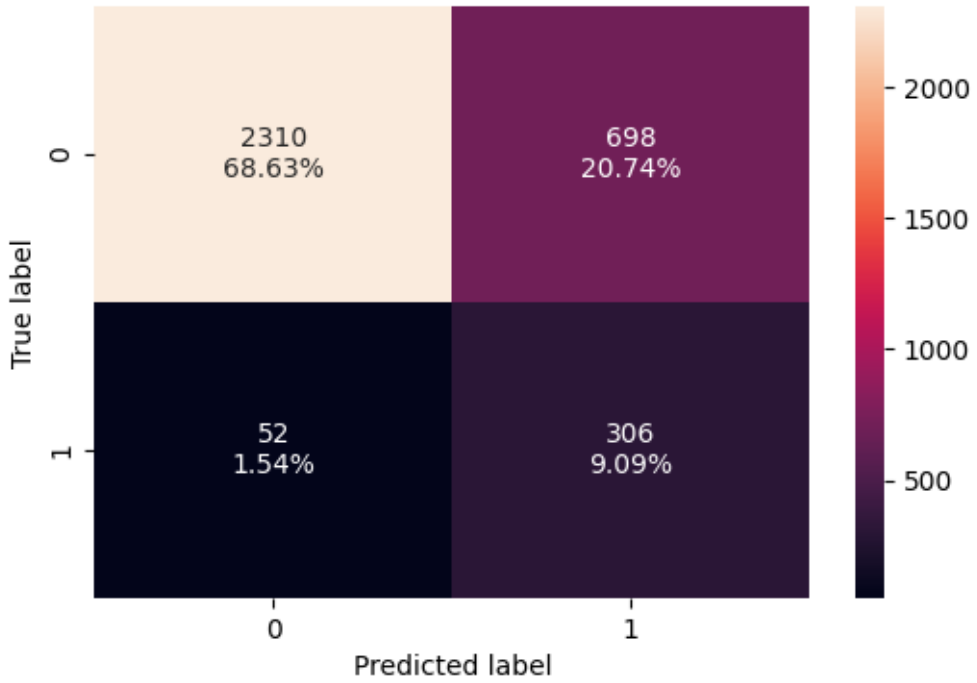
Overall, **recall remains strong**, which is crucial in identifying survivors.

Decision Tree Classifier performance on training set and test set:

Train set



Test set



Metric	Matrix 1	Matrix 2	Comments
Accuracy	77.6%	77.6%	Identical
Precision	30.25%	30.47%	Slightly better in Matrix 2
Recall	86.99%	85.47%	Slightly better in Matrix 1
F1 Score	44.97%	44.88%	Nearly identical
False Positives	1651	698	Significantly better in Matrix 2 (less noise)
False Negatives	107	52	Lower in Matrix 2 (better for survivors)

Both models are performing nearly identically in terms of overall accuracy and F1-score.

Test 1 is better at recalling actual survivors, which is crucial in safety contexts.

Test 2 has fewer false positives, which makes it better for minimizing false alarms.

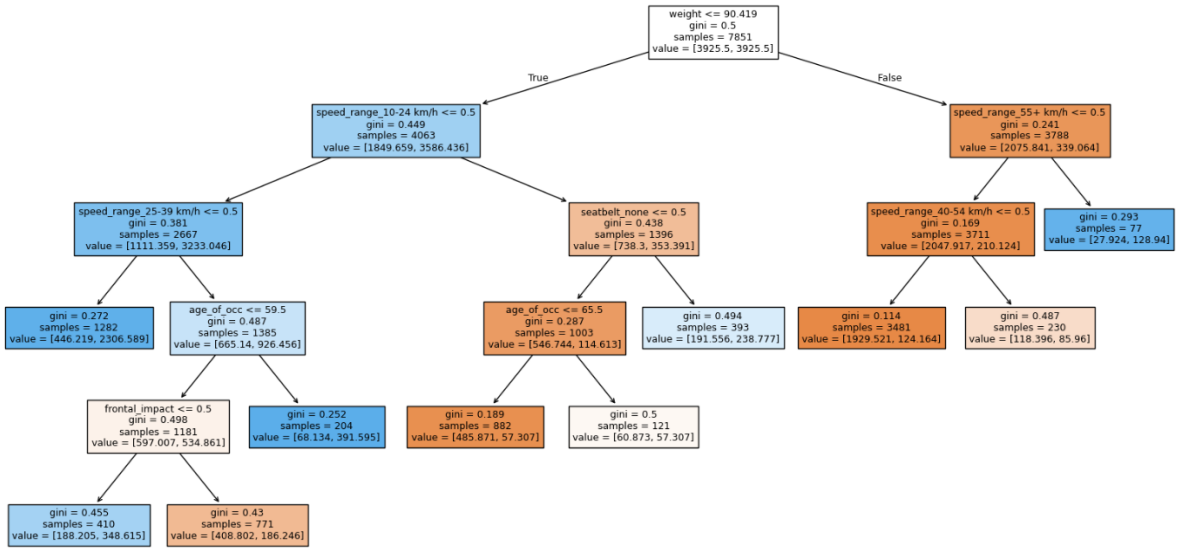
Choose **Matrix 2's model** if the cost of falsely predicting survival is higher.
Choose **Matrix 1's model** if **missing actual survivors** is more critical.

Balanced Performance: Slightly better precision than Matrix 1.

Fewer False Positives: Safer for applications where false alarms are costly.

Use Case: Preferred where over-predicting survival leads to wasted resources.

Visualizing the Decision Tree:



Each node in the tree:

- Splits the dataset based on a feature condition.
- Displays:
  - Gini index (impurity – lower is better).
  - Samples (how many data points reach this node).
  - Value = [class_0, class_1]: total weighted sum or average for each class (in your case: not survived vs. survived).

Left Branch (Weight ≤ 90.419)

This represents lighter cars.

1st Split:

speed_range_10-24 km/h ≤ 0.5

- Gini = 0.449
- This divides based on speed: either >24 km/h or ≤24 km/h.

Then:

- If speed is >24 km/h, it checks:
  - speed_range_25-39 km/h
  - Then age_of_occ (age of occupant)
  - Then frontal_impact

This branch explores survival patterns among lighter vehicles at medium speeds, factoring in:

- Occupant age
- Impact type
- Speed range

Example:

- Leaf node:
  gini = 0.272, samples = 1282, value = [446.219, 2306.589]
  → Majority survived, high survival rate.

2nd Path (If speed ≤ 24 km/h):

It checks seatbelt usage, then age, suggesting:

- Among light cars at low speed, seatbelt and age significantly influence survival.

Right Branch (Weight > 90.419)

This represents heavier cars.

1st Split:

speed_range_55+ km/h ≤ 0.5

- Gini = 0.241, lower impurity → better class separation.

  Then:
- It further checks if speed is 40-54 km/h, revealing this speed range in heavier cars is safer:
  - Node:
    gini = 0.114, samples = 3481, value = [1929.521, 124.164]
    → Vast majority did not survive, showing risk at these speeds even in heavier vehicles.

1. Weight is the top factor: Light vs. heavy cars behave differently in crashes.
2. Speed is critical: Mid-speed (25–54 km/h) ranges show high sensitivity.
3. Seatbelt use improves survival: Seen in light vehicle splits.
4. Age impacts survival: Especially in moderate-impact crashes.
5. Frontal impacts reduce survival odds: Reflected in deeper branches.

8.Model Performance Comparison and Final Model Selection:

Training performance comparison:

	Logistic Regression Base	Logistic Regression (Optimal threshold)	Decision Tree Base	Decision Tree Tuned
Accuracy	0.91160	0.82525	1.00000	0.77608
Recall	0.35645	0.86131	1.00000	0.86983
Precision	0.63974	0.36012	1.00000	0.30220
F1	0.45781	0.50789	1.00000	0.44856

Logistic Regression (Base)

- Very accurate (91%)
- Misses many real survivors (low recall)
- When it says someone will survive, it's usually right (high precision)

Best for: When being wrong about survival is risky (false alarms are bad).

Logistic Regression (Optimal Threshold)

- Lower accuracy (82%)
- Catches most survivors (high recall)

- More false alarms (lower precision)
- Best F1 score → good balance

Best for: When finding every possible survivor is more important, even if some predictions are wrong.

Decision Tree (Base)

- Perfect scores (100% everything)
- Likely overfitted (memorized data, bad for real-world)

Not reliable for real use.

Decision Tree (Tuned)

- Good at catching survivors
- Less accurate overall
- Lots of false positives

Best for: Situations where recall matters, but you can handle more wrong predictions.

Summary

Goal	Use This Model
Accuracy-focused	Logistic Regression (Base)
Find all survivors	Logistic Regression (Optimal)
Overfitted, avoid	Decision Tree (Base)
Good recall, okay balance	Decision Tree (Tuned)

Test set performance comparison:

	Logistic Regression Base	Logistic Regression (Optimal threshold)	Decision Tree Base	Decision Tree Tuned
Accuracy	0.91503	0.82620	0.88770	0.77718
Recall	0.38268	0.85475	0.42458	0.85475
Precision	0.67822	0.36472	0.46914	0.30478
F1	0.48929	0.51128	0.44575	0.44934

Logistic Regression (Base)

- Very accurate (91.5%)
- Misses many survivors (recall = 38%)
- When it predicts survival, it's often right (precision = 68%)

Best for: Accuracy and safe predictions, but not great at catching all survivors.

Logistic Regression (Optimal Threshold)

- Lower accuracy (82.6%)
- Catches most survivors (recall = 85%)
- Many false alarms (precision = 36%)
- Best F1 score → best balance between precision and recall

Best for: Making sure you find as many survivors as possible, even if some predictions are wrong.

Decision Tree (Base)

- High accuracy (88.8%)
- Low recall (42%)
- OK precision (47%)
- Lowest F1 score of the 3

Best for: Somewhat balanced, but not as good as the logistic regression models.

Decision Tree (Tuned)

- Lower accuracy (77.7%)
- Great recall (85%)
- Weak precision (30%)
- F1 score is decent

Best for: Prioritizing survivors, like logistic (optimal), but less accurate overall.

Final Recommendation (Test Set)

Goal	Best Model
Highest accuracy	Logistic Regression (Base)
Catch all survivors	Logistic Regression (Optimal)
Okay balance, simpler tree	Decision Tree (Tuned)
Avoid (weak F1)	Decision Tree (Base)

8. Actionable Insights:

1. High-Speed Collisions Are Most Fatal

- A large portion of fatal cases occurs at speeds ≥55 km/h.
- Survival probability significantly decreases at higher speed ranges.

Implication: Speed remains a dominant factor in fatal outcomes despite other safety measures.

---

2. Seatbelt Usage Matters Significantly

- Occupants without seatbelts have a markedly higher fatality rate.
- Among belted occupants, even at higher speeds, survival rates are better.

Implication: Seatbelt compliance is a key determinant of survival.

---

3. Airbag Deployment is Inconsistent

- Many fatal cases had airbags marked as "unavail" or "nodeploy".
- When airbags deploy, survival rates improve—though effectiveness varies with impact type and speed.

Implication: Airbag systems either failed to deploy or were absent in many older model cars.

---

4. Frontal Impacts Are More Lethal

- Frontal impacts (coded as 1) are more commonly associated with fatalities.

Implication: Vehicle crumple zones and frontal safety need enhancement.

---

5. Age-Related Vulnerability

- Elderly occupants (especially >60 years) and very young individuals are at higher risk.
- The average age in fatal crashes is higher than that in survivals.

Implication: Special safety considerations should be enforced for older drivers/passengers.

---

6. Driver Role Is Riskier Than Passenger

- Drivers experience a higher fatality rate compared to passengers.

Implication: Driver-side safety should receive more emphasis in design.

---

7. Older Model Cars Pose Greater Risk

- Vehicles manufactured before 1990 show a disproportionately high fatality rate.

Implication: Modern safety regulations have clearly improved outcomes over time.

8. Disparity in Male vs. Female Fatality Outcomes

- Males are more frequently involved in fatal crashes than females.
- This may be linked to higher risk behaviors or differing seatbelt usage patterns.

Implication: Gender-sensitive safety awareness programs may be beneficial.

---

9. Crash Risk Peaks in Middle-Aged Adults

- Although elderly individuals have higher vulnerability, a significant volume of crashes occurs in the 25–45 age range—especially among drivers.

Implication: Targeted behavioral interventions (e.g., anti-distracted driving campaigns) should focus on this demographic.

---

10. Passengers Are Often Less Protected

- Passenger fatalities increase when airbags are unavailable or undeployed, even with seatbelts.
- Many vehicles seem optimized for driver safety rather than all occupants.

Implication: Rear and side airbag systems are underutilized or under-regulated.

---

11. Airbag Deployment Has a Positive Correlation With Survival

- Across most speed brackets, when airbags deploy, survival rates increase by 10–25%.

Implication: There's a missed opportunity in older models or faulty systems that fail to deploy airbags in crashes.

---

12. Accidents Cluster in Specific Year Bands

- There is a noticeable uptick in crash data from 2001–2002, possibly indicating:
  - Increased vehicle density
  - Road infrastructure issues
  - Economic expansion affecting traffic volume

Implication: Temporal patterns in accidents can inform urban traffic policies.

---

13. Low-Weight Observations Skew Data

- A subset of records has implausibly low weights (e.g., <10 units), suggesting either:
  - Underreporting
  - Sampling anomalies

Implication: Consider outlier handling or use of robust statistical models to avoid biased conclusions.

---

14. Non-Frontal Impacts Can Still Be Fatal at High Speeds

- While frontal impacts dominate fatalities, non-frontal impacts at 55+ km/h are still fatal in ~40% of such cases.

Implication: Side-impact protection should be evaluated in high-speed zones.

9.Recommendations:

For Government & Regulatory Bodies:

1. **Mandate Speed Limiters** in vehicles operating in urban areas.
2. **Enforce Stricter Seatbelt Laws**: especially for rear-seat passengers.
3. **Mandatory Airbag Deployment Checks**: Introduce inspection norms to verify functioning.
4. **Introduce Age-Specific Licensing Regulations**: Evaluate reflexes or health for older drivers.
5. **Retirement of Older Cars**: Provide incentives to discard vehicles >20 years old.

For Automotive Manufacturers:

1. **Enhance Frontal Crash Resistance**: Use AI-simulated crash testing for future models.

2. **Improve Airbag Systems**:
    - Dual-stage deployment.
    - Passenger detection for smart deployment.
3. **Design Age-Friendly Features**:
    - Ergonomic seating.
    - Easy ingress/egress and warning alerts for elder drivers.

For Insurance Companies:

1. **Dynamic Risk-Based Pricing**:
    - Adjust premiums based on vehicle model year, occupant age, and safety features.
2. **Offer Discounts for Safety Retrofits**:
    - Encourage installations of aftermarket airbags or advanced restraint systems.

For Automotive Safety Engineering:

- **Mandate Side-Curtain and Knee Airbags** in future vehicle designs.
- Introduce **smart restraints** that adjust tension based on impact force or occupant weight.
- Integrate **black box systems** for post-crash analysis in new cars.

For Smart City and Transport Planners:

- Use this crash dataset to identify **crash-prone time windows and conditions** (e.g., rainy months, night-time).
- Invest in **adaptive speed enforcement** zones using predictive analytics models.
- Deploy **real-time risk alerting systems** via vehicle-to-infrastructure (V2I) communication.

For Public Awareness Campaigns:

- Launch **"Buckle Up or Pay Up"** campaigns in urban zones where seatbelt usage is low.

- Highlight **case studies of airbag save scenarios** to boost consumer demand for safety features.
- Conduct **community-based workshops** to teach families how to choose safer vehicles and protect aging drivers.