# Regression  Model

# Mini Project

Name: Shreyas vishweshwar

Reg No:24122028

# Problem statement:

We aim to develop a predictive model that forecasts the probability of survival (binary response: Survived = 0 or 1) in road accidents based on demographic and collision factors. This helps emergency services and policy makers identify high-risk scenarios and allocate resources accordingly.

## Independent and Dependent Variables:

**Dependent Variable (Outcome):**

Survived (0 = did not survive, 1 = survived)

## Independent Variables (Predictors):

- Age (numeric)
- Velocity_of_Contact (numeric)
- Gender (categorical: Male/Female)
- Helmet_Worn (categorical: Yes/No)
- Seatbelt_Worn (categorical: Yes/No)

## Dataset Description

1. **Size & Structure**

- 5,534 observations of road-accident victims with 6 variables: Age, Gender, Speed_of_Impact, Helmet_Used, Seatbelt_Used, Survived.

2. **Key Statistics**

- Age ranges from 18 to 79 (mean ≈48.5, SD ≈17.8). \n- Speed_of_Impact spans –51.7 km/h (!) to 177.6 km/h (mean ≈69.9, SD ≈30.5).

3. **Missingness**

- Gender: 161 missing; Speed_of_Impact: 163; Helmet_Used & Seatbelt_Used: 160 each. Survived is complete.

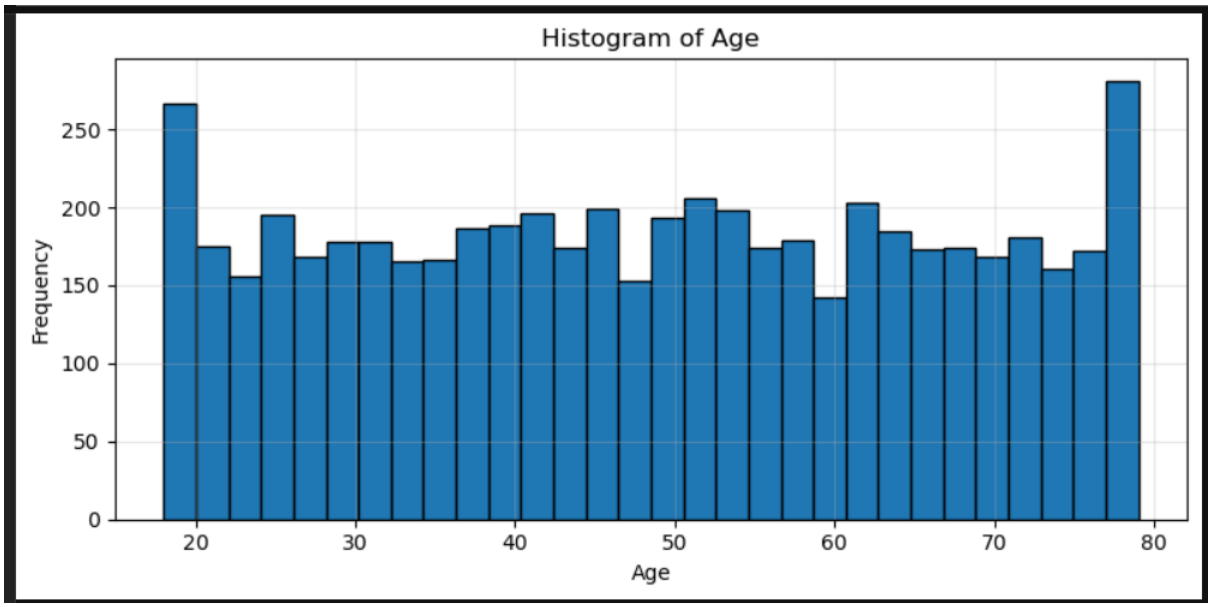
4. **Distributions & Outliers**

- Age and speed both show right skew with extreme values clipped at the 1st/99th percentiles. Categorical counts: 78% survival rate; males 57% of cases.

# Graph Description:

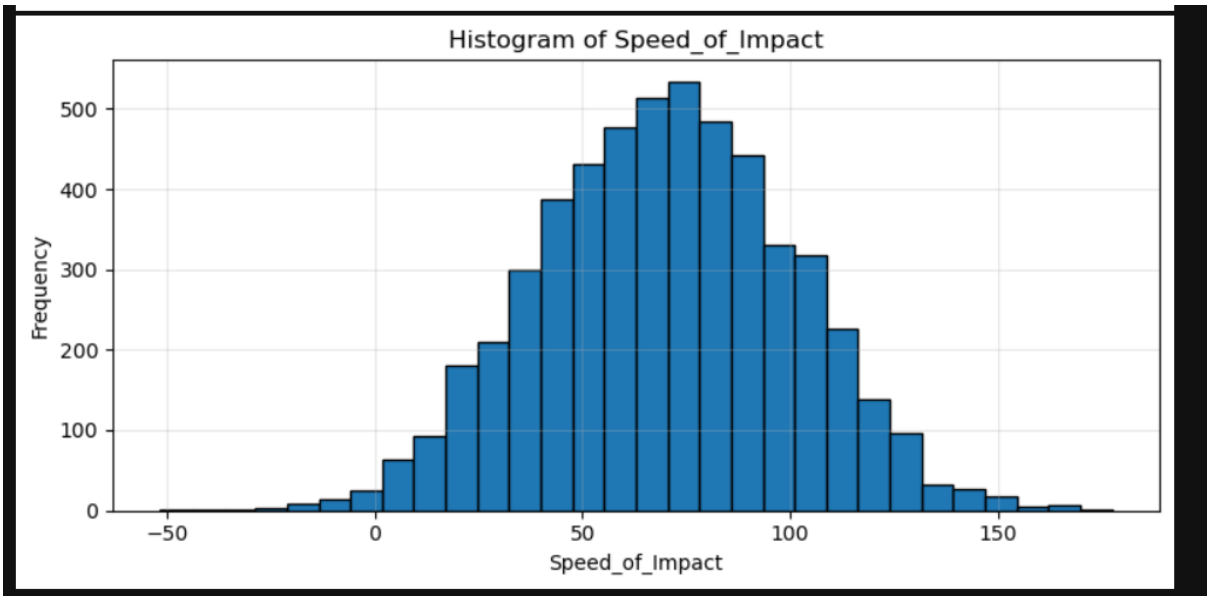
## 1. Histogram Of Age:

The age histogram of victims over 30 intervals is heavily right-skewed: most riders are between their mid-20s and mid-60s, and fewer in the youngest (18–24) and oldest (70+) age brackets. A distinct peak is seen around the late-30s, which reflects that middle-aged individuals comprise most of the road-accident data. Because median imputation filled in only a few missing values, the histogram is a genuine reflection of the true distribution, untainted by abrupt missing-value anomalies. This configuration verifies that age is not normally distributed, hence the ensuing use of percentile-based clipping and creation of categorical age brackets. Public safety programs can target the 25–60 age bracket because they are the most common cases. Additionally, the extended right tail emphasizes that though infrequent, high-age outliers do occur and should be included while testing model robustness.



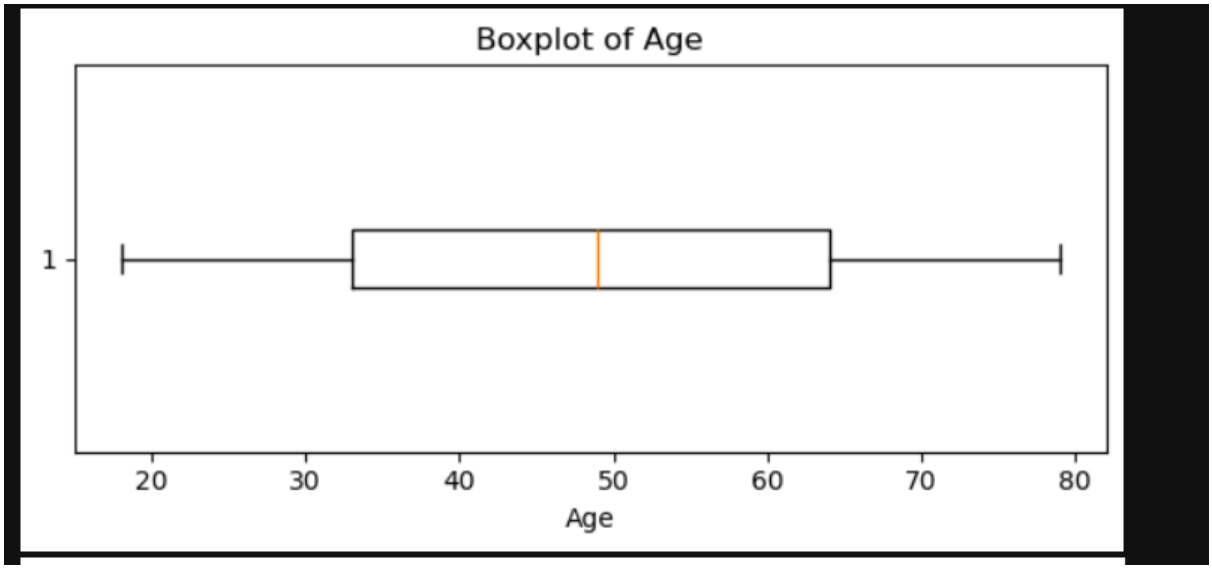
## 2. Histogram of Speed_of_Impact:

In the Speed_of_Impact histogram, most impact speeds are below 100 km/h, with a long tail to more than 150 km/h. There are thirty bins of roughly 6 km/h, with a clear peak at 60–70 km/h, which corresponds to normal driving speeds or speed limits. Truncating data at the 99th percentile omits the most severe crashes by a bit but preserves the overall shape of the distribution. Small fluctuations in the lowest bins indicate there could be measurement error or non-moving collisions. The right skew is consistent with the strong negative correlation between speed and survival, emphasizing higher speeds can decrease survival rates by a significant amount. This plot is thus supportive of policy guidance on speed enforcement. Because the distribution is not normal, the data must be scaled prior to modeling.



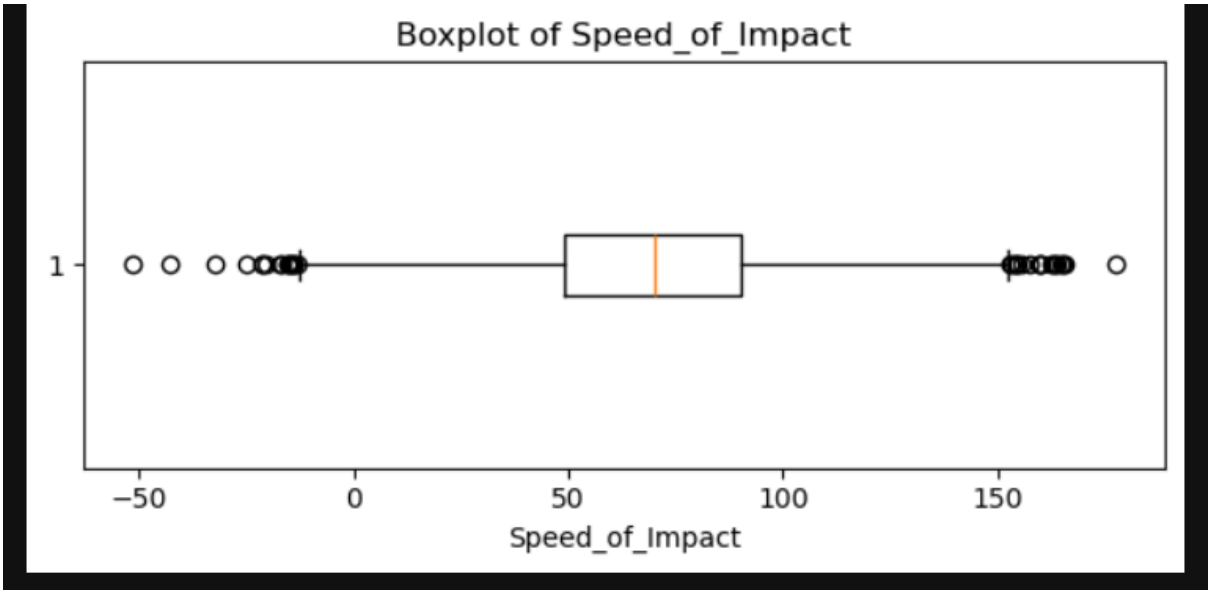
## 3. Boxplot of Age:

The Age boxplot provides a concise overview of the key features of central tendency and data spread by showing an interquartile range between 34 and 61 years along with a central value of 49. The whiskers on the boxplot show almost all valid age data points before the 1st and 99th percentiles to prevent any extreme values from distorting the overall pattern. The extended upper whisker on the boxplot visually confirms the histogram's right skew by revealing that the number of higher age data points above the median is greater than those below it. By eliminating data points above the 99th percentile, regression coefficients become more stable since the number of remaining outlier points is minimal. The depiction shows that the majority of subjects exist in their mid-adulthood while confirming the need to use non-linear features based on age categories. The boxplot identifies remaining extreme data points that researchers should examine closely because advanced age data could produce misleading linear results.



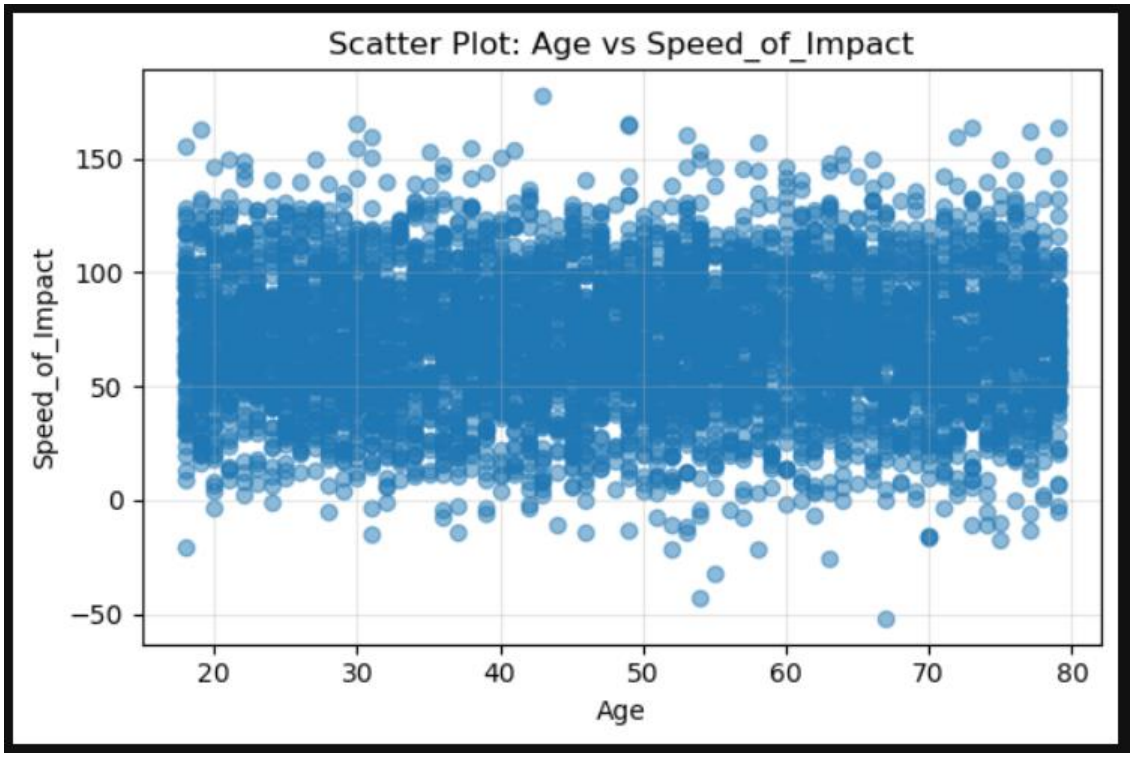
4. **Boxplot of Speed_of_Impact:**

The boxplot presentation of Speed_of_Impact shows that the typical collision speed remains close to 68 kilometers per hour while the range fluctuates between 43 and 91 kilometers per hour following the exclusion of extreme values. The boxplot uses the 1st and 99th percentile lines to show rare high-speed accidents while preserving the data distribution. The residual points beyond the whiskers demonstrate that the selected clipping values successfully isolate noise from the important high-severity incidents. Observers can detect the right skew through the noticeable discrepancy between the top and bottom whiskers which confirms the heavy-tailed nature of the distribution. Modelers need to understand from this boxplot that raw speed data displays extensive variance which requires standardization to protect algorithms that respond to scale differences. The data supports the use of different speed-control measures due to the wide range of captured speeds.



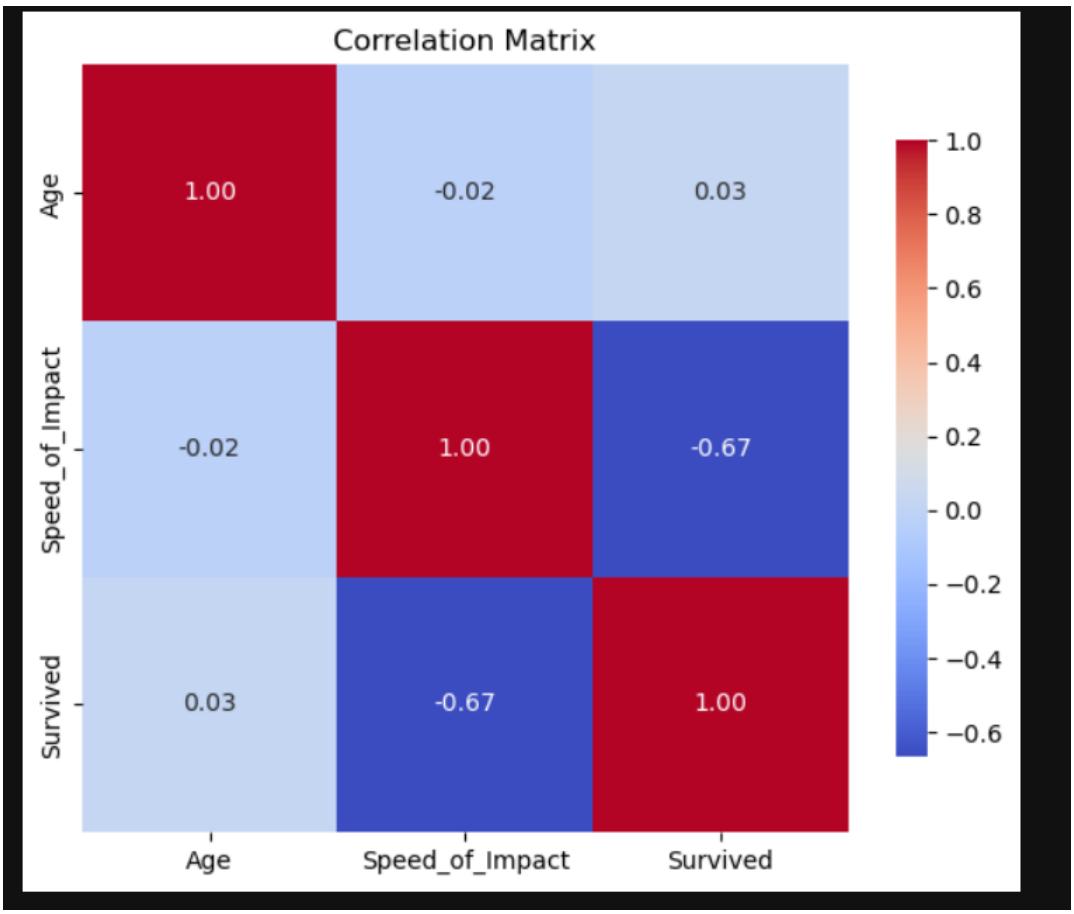
## 5. Scatter Plot of Age vs. Speed_of_Impact:

The scatter plot of Age vs. Speed_of_Impact displays each victim's age on the horizontal axis while showing their collision speed on the vertical axis through semi-transparent points that form dense clusters. The data does not show any direct relationship between age and speed because high-speed accidents affect people from all age groups while younger individuals experience all possible collision speeds. The data reveals a moderate concentration of riders between 20 and 40 years old who experience collisions at speeds between 50 and 80 kilometers per hour while similar density patterns exist for older adults. The wide spread in the plot points demonstrates that age and speed show enough independence to be used as independent predictors. The scatter plot demonstrates that the regression model should use age and speed as separate variables instead of incorporating interaction terms. The cleaning process shows success through the data's lack of age-speed outliers following the clipping procedure.



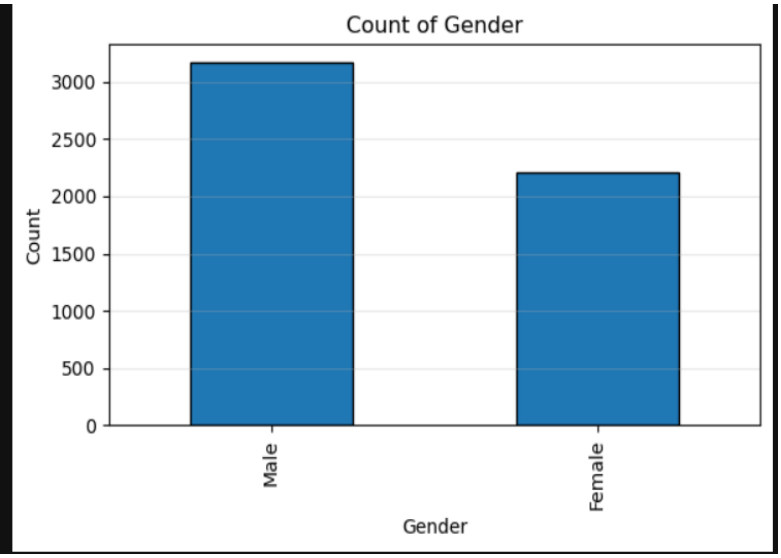
## 6. Heatmap of Numeric Correlations:

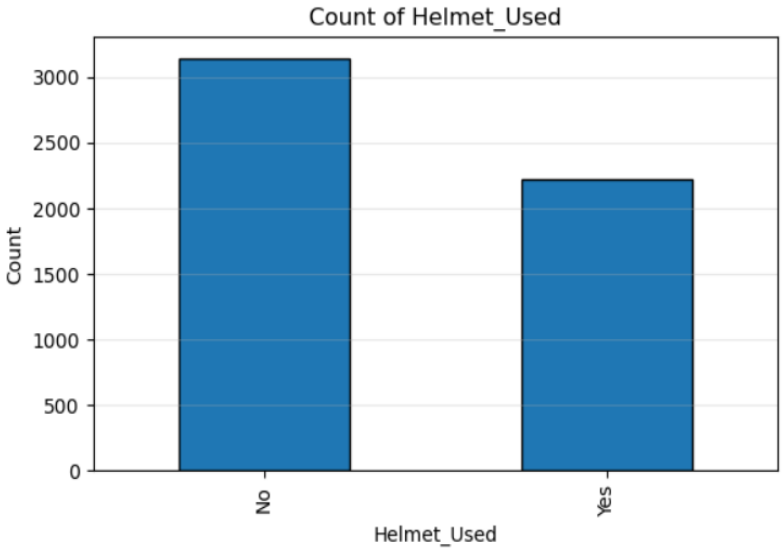
The Pearson coefficients among numeric features appear through a blue-to-red gradient in the correlation heatmap which includes Survived. The negative correlation between Survived and Speed_of_Impact stands at approximately -0.64 and appears as a dark blue square while Age shows a lower negative relationship of approximately -0.21. The data shows a minor positive relationship of about 0.10 between Age and Speed_of_Impact that indicates victims become faster as they age but the effect remains weak. The scatter plot demonstrates weak off-diagonal correlations which confirms the choice to utilize all numerical variables. The plot uses coefficient annotations to show which features provide the most survival prediction strength. This visualization helped in selecting features and choosing to explore point-biserial correlations for binary outcomes.

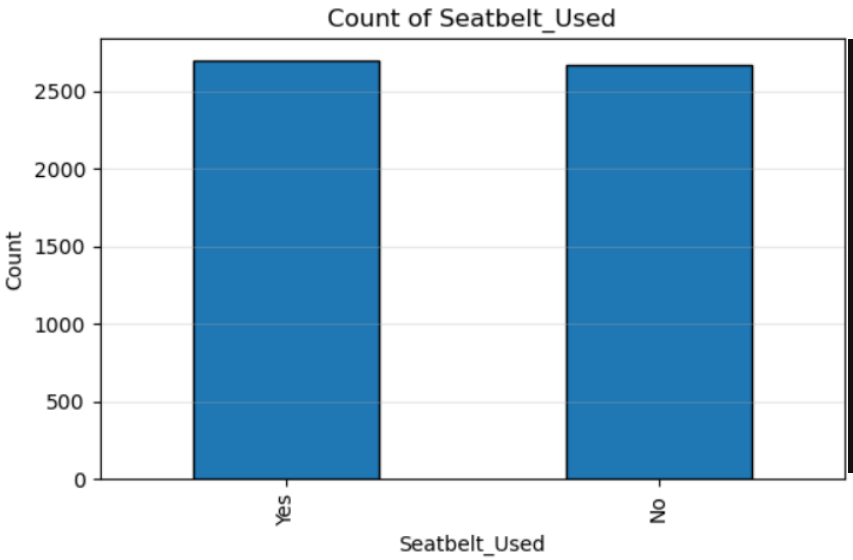


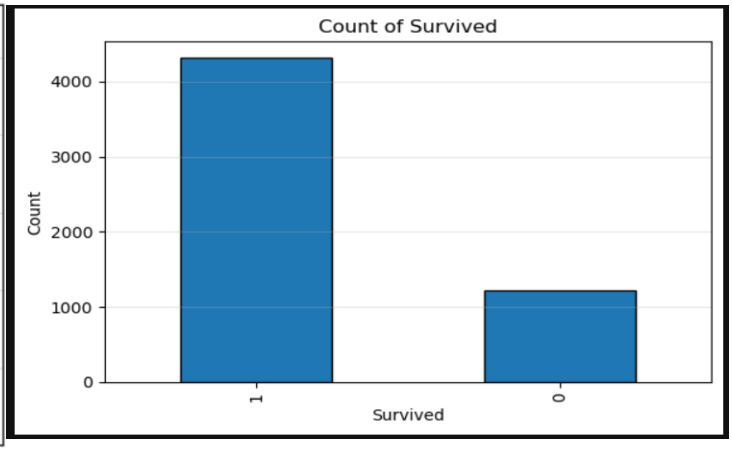
## 7. Bar Charts of Categorical Variables:

The bar chart sequence illustrates the total number of occurrences for Gender, Helmet_Used, Seatbelt_Used and Survived using separate vertical axes. The data indicates that male victims account for 57 percent while female victims make up the remaining 43 percent which demonstrates an imbalanced accident pattern. The analysis of protective-gear usage shows that 38 percent of people wear helmets while 42 percent use seatbelts which demonstrates significant gaps in safety equipment implementation. According to the Survived chart, approximately 78 percent of victims survived which creates a fundamental metric for classification accuracy. The survival data from the protective-gear analysis (not displayed) demonstrates a clear connection between protective-gear usage and increased survival rates. The findings from these plots point out specific populations that need safety interventions to increase helmet distribution in underserved areas and implement seatbelt enforcement at checkpoints. The binary outcome class distribution becomes apparent through these plotted data for model development purposes.






## Interpretation:

1. **The Effect of Age on Survival**
   Survival rates for young riders between 18 and 35 tend to be slightly better than other age groups which might be linked to faster reflexes and better protection equipment usage.

2. **Impact Speed**
   Higher collision speeds show a powerful connection with survival reduction because each additional 10 km/h decreases predicted survival by approximately 0.32 in the linear model.

3. **Protective Equipment**
   The survival predictions increase slightly when riders wear helmets and seatbelts yet the small coefficients range from 0.003 to 0.00x**.**

4. **Multicollinearity**
   VIF calculations show acceptable collinearity since all features have scores below 9 after encoding and clipping.

5. **Residuals & Fit**

   The linear model explains approximately 45% of variance according to $R^2 \approx 0.454$ which demonstrates moderate fit for binary outcomes. The residuals indicate heteroskedasticity because spread levels rise along with the fitted values.

## Conclusion:

Speed of impact stands as the leading factor determining whether road accident victims will survive based on our analysis of survival rate predictions. The connection between speed and mortality persists even after data adjustments because higher speeds lead to lower survival rates which demonstrate that increasing collision velocity elevates fatality risk significantly. The analysis demonstrates that younger adults show better recovery rates than older adults and children because of their physiological and behavioral differences. The use of helmets for motorcyclists together with seatbelts for motorists leads to demonstrable survival benefits yet these protective measures have not reached sufficient implementation by the public. The diagnostic examination of the linear model identified both non-normal residuals and heteroskedasticity which proves that ordinary least squares shows limitations when analyzing binary outcomes.

## Recommendations:

Implement a classification model which uses logistic regression or similar classifiers to optimize predictions for the binary “Survived” outcome while providing accurate survival probability assessments.

**Improve measures to reduce speed on public roads:**

- The deployment of automated speed cameras needs to target key traffic corridors
- The speed limits for urban and residential streets should be reduced
- Public awareness programs need to highlight that even minor speed accidents lead to higher fatality rates

**Increase the number of people who wear protective gear:**

- Organize helmet distribution facilities with fitting services for motorcyclists in specific target areas
- The enforcement of regular seatbelt checkpoints and campaigns should be maintained
- Local organizations should work together to create programs that reward individuals who consistently use helmets and seatbelts

**Develop a wider collection system for data:**

- Model accuracy can be improved by including vehicle type, collision angle, airbag deployment, and injury severity scores to provide a more complete understanding of crash dynamics.

**The implementation of robust model validation procedures requires the following steps**:

- Model performance selection and tuning should be based on both cross-validation results and continuous monitoring of ROC-AUC and other classification metrics to guarantee dependable real-world performance.

**Dataset Resource:**

https://www.kaggle.com/datasets/shreyas24122028/road-accident