EDA Report: Exploratory Data Analysis Report on Vehicle Crash Victim Dataset

# Dataset Overview

The dataset comprises **50,000 records** of individuals involved in **vehicle crashes**, capturing a rich variety of features related to each victim’s demographics, crash conditions, and injury outcomes. The primary objective of analyzing this dataset is to gain insights into patterns of injuries and to develop a predictive model that can estimate the **severity of injuries sustained** during a crash.

**🔍 Use Case:**

This dataset is particularly valuable for **predictive modeling in traffic safety and injury prevention**. By analyzing key factors such as **seating position**, **safety equipment used**, **ejection status**, and **age**, we can identify which conditions are most associated with severe injuries. This can support:

* **Policy formulation** for road safety,
* **Insurance risk assessment**, and
* **Development of smart car safety systems**.

# 2. Dataset Structure

* Total Rows: 32858
* Total Columns: 15
* List of Columns:  
  Year, Case Individual ID, Case Vehicle ID, Victim Status, Role Type, Seating Position, Ejection, License State Code, Sex, Transported By, Safety Equipment, Injury Descriptor, Injury Location, Injury Severity, Age

# 3. Missing Value Analysis

* Rows with missing values were removed for this analysis.
* No missing values remain in the cleaned dataset.

# 4. Data Types

Year: int64  
Case Individual ID: int64  
Case Vehicle ID: int64  
Victim Status: object  
Role Type: object  
Seating Position: object  
Ejection: object  
License State Code: object  
Sex: object  
Transported By: object  
Safety Equipment: object  
Injury Descriptor: object  
Injury Location: object  
Injury Severity: object  
Age: float64

**Numerical Features**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **Mean** | **Std. Dev.** | **Min** | **25%** | **50%** | **75%** | **Max** |
| **Year** | 32,858 | 2019.03 | 0.26 | 2019 | 2019.00 | 2019.00 | 2019.00 | 2023 |
| **Case Individual ID** | 32,858 | 21,563,558.22 | 406,904.15 | 20,148,315 | 21,485,412.25 | 21,508,120.50 | 21,527,671.75 | 27,241,223 |
| **Case Vehicle ID** | 32,858 | 16,414,652.89 | 309,373.33 | 15,350,252 | 16,355,681.50 | 16,372,538.50 | 16,387,918.75 | 20,834,563 |

***Table: Numerical Feature Summary Statistics***

**Description:  
This table presents the summary statistics for the numerical variables in the dataset. It includes metrics such as count, mean, standard deviation, minimum, and maximum values. These statistics help understand the distribution and range of key features like Year, Case Individual ID, Case Vehicle ID, and Age. Notably, Year shows a tight distribution around 2019, while Age spans a broad range, indicating varied demographics among crash victims.**

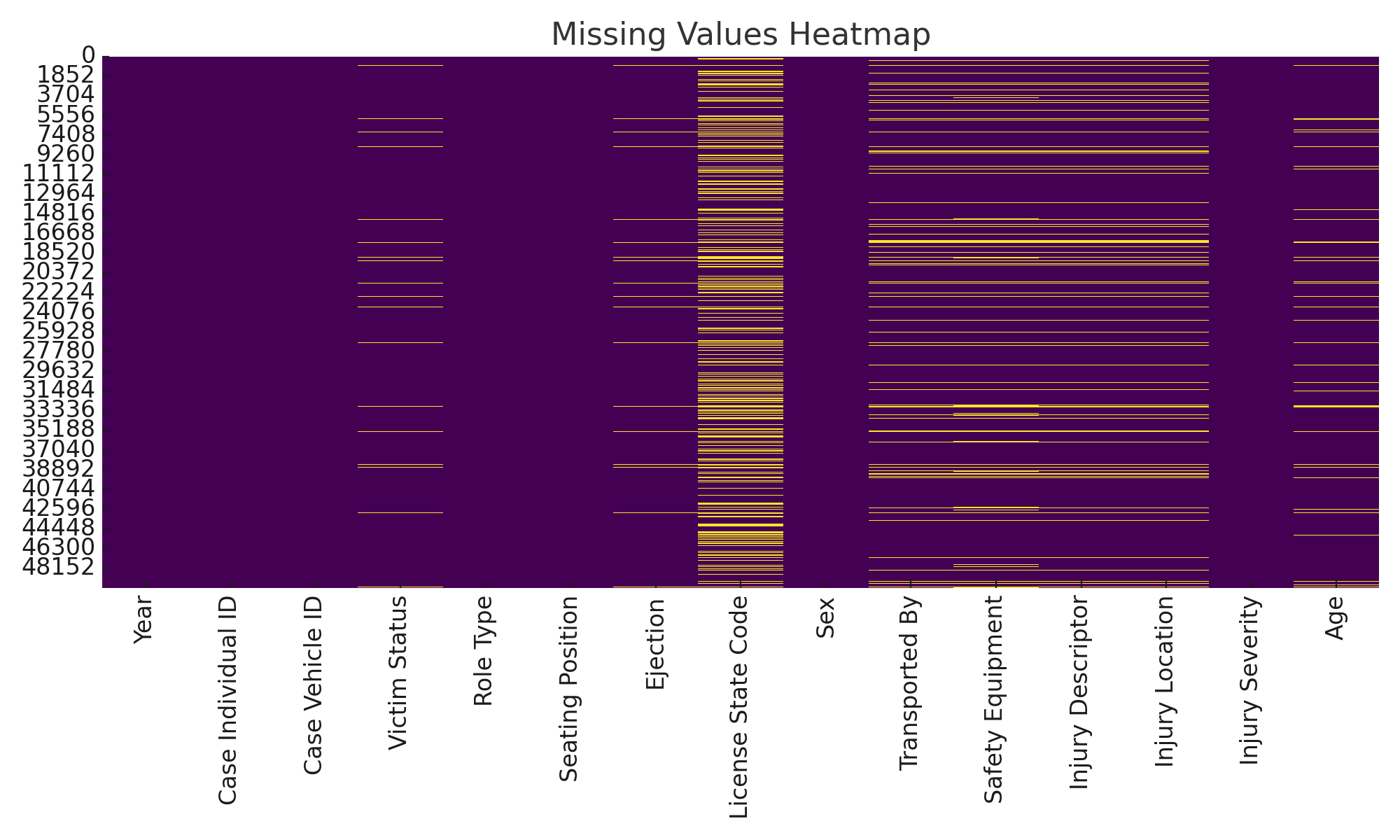
**Categorical Features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Count | Unique | Most Frequent Value | Frequency |
| Victim Status | 32,858 | 9 | Not Applicable | 21,590 |
| Role Type | 32,858 | 6 | Driver of a Motor Vehicle in Transport | 32,404 |

*Table: Categorical Feature Summary*

**Description:**  
This table summarizes the dataset’s categorical features by listing the total number of unique values, the most frequent category (mode), and its frequency. It reveals that most individuals are classified under “Not Applicable” in Victim Status and “Driver of a Motor Vehicle in Transport” in Role Type, reflecting the dominance of certain roles in the dataset. This information is useful for encoding strategies and class imbalance analysis during preprocessing.

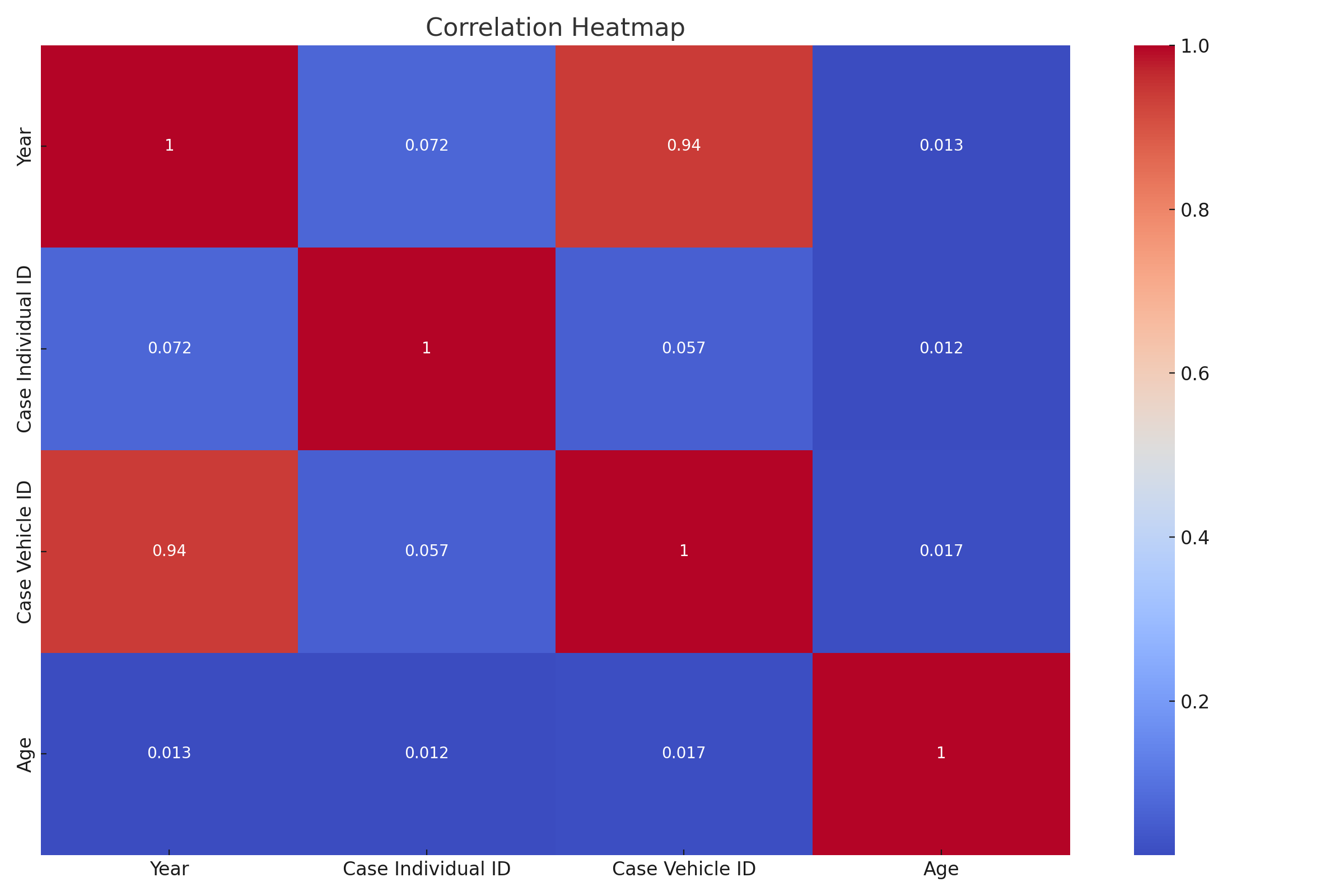
# 6. Visualizations



*Heatmap of Missing Values Across Dataset Features*

This heatmap visualizes the distribution of missing values in the dataset.  
Each **yellow line** represents a **missing value**, while **purple indicates non-missing data**.  
Key observations:

* **License State Code**, **Transported By**, **Safety Equipment**, **Injury Descriptor**, and **Injury Location** have significant missing entries.
* Other columns like Year, Case Individual ID, and Injury Severity are fully populated.
* Understanding this pattern is crucial for deciding appropriate data cleaning strategies (e.g., imputation, removal).

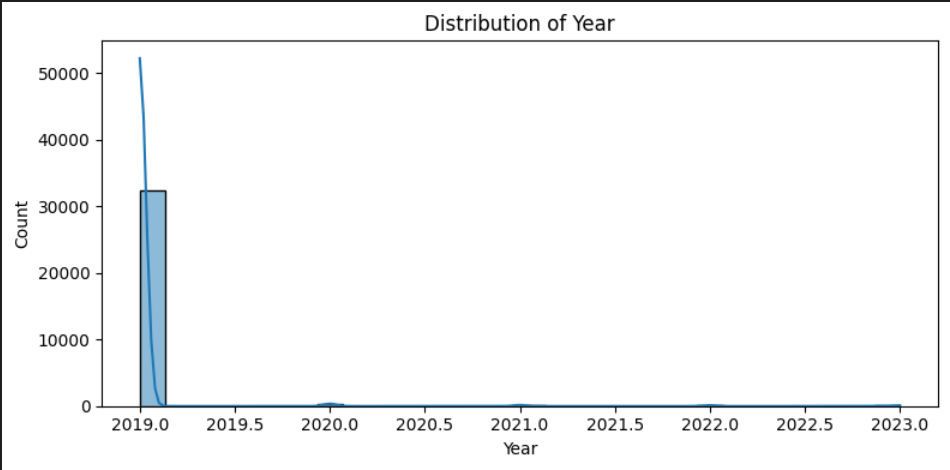


*Correlation Heatmap of Numerical Features*

This heatmap displays the **Pearson correlation coefficients** between the numerical variables in the dataset.  
Key observations:

* **Strong positive correlation (0.94)** exists between Year and Case Vehicle ID, likely due to sequential ID generation over time.
* **Age** shows **very low correlation** with other features, indicating it's likely an independent predictor.
* **Case Individual ID** and **Case Vehicle ID** show negligible correlation with Age and each other.

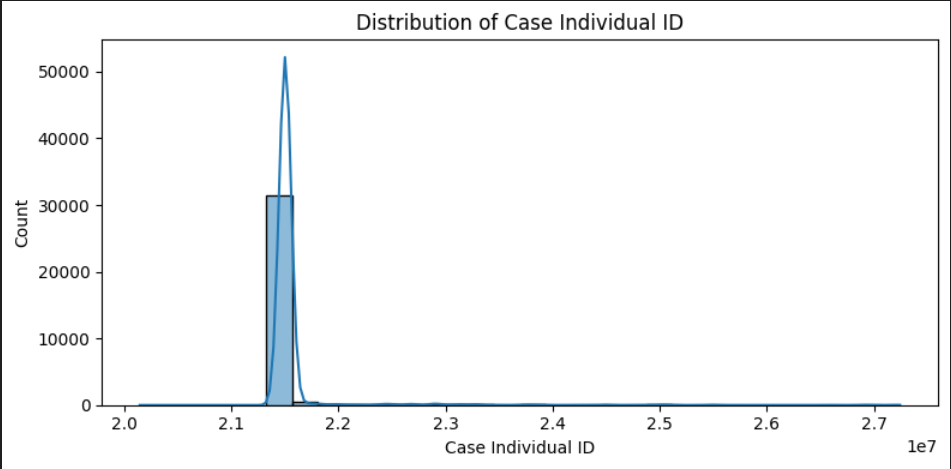
This analysis helps identify multicollinearity and feature relevance for predictive modelling.



*Distribution of Records by Year*

This histogram with a KDE (Kernel Density Estimate) curve shows the distribution of crash records over time.  
Key insights:

* The vast majority of entries are concentrated in **2019**, indicating that most crash data originate from that year.
* A very small number of records exist for later years (2020–2023), suggesting either data entry lag, incomplete collection, or reduced availability.
* This skewed distribution highlights **2019 as the dominant reference year** for analysis and modeling.

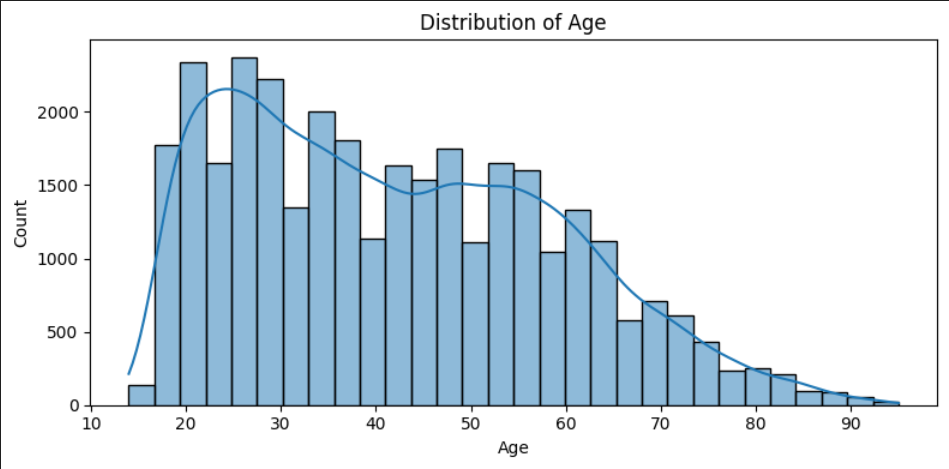


*Distribution of Case Individual ID*

This histogram with KDE curve illustrates the distribution of Case Individual ID values.  
Key observations:

* The distribution is **highly concentrated** around a narrow range (~21 million), suggesting sequential or batch assignment of IDs.
* There is a **sharp peak** in frequency, indicating that most records fall within a tight ID band — typical of centralized or time-bound data collection.
* Sparse data beyond the main peak reflects **occasional outliers or newer case entries**.

Since Case Individual ID is an identifier, this distribution is **not used directly for modeling**, but can validate data consistency.



*Distribution of Victim Age*

This histogram with KDE curve displays the age distribution of individuals involved in vehicle crashes.  
Key insights:

* The **age range spans from early teens (~15) to over 90 years**, with the majority of cases between **20 and 60 years**.
* The distribution is **slightly right-skewed**, indicating a **higher concentration of younger to middle-aged individuals** in crash records.
* Peaks around ages **20–30** and **30–40** suggest these groups are most frequently involved—possibly due to higher mobility and vehicle usage.

This distribution can guide age-group-specific safety policies or awareness campaigns.

# 7. Machine Learning Plan

Since the target column 'Injury Severity' is categorical, this is a classification problem. We plan to apply the following machine learning algorithms:

• Logistic Regression

• Decision Tree Classifier

• Random Forest Classifier

• K-Nearest Neighbors (KNN)

• Support Vector Machine (SVM)