Road Accident Data Analysis

Fatemeh Chajaei

1. Introduction

Road accidents pose significant challenges to public safety and transportation systems worldwide. Understanding the factors contributing to accidents and their severity is crucial for implementing effective safety measures. This report aims to analyze road accident data and identify key insights to inform strategies for accident prevention and mitigation.

2. Related Works

Numerous studies have investigated the causes and consequences of road accidents. Previous research has highlighted factors such as driver behavior, road conditions, vehicle characteristics, and environmental conditions as contributing elements to accidents. Understanding these factors is essential for developing targeted interventions to reduce accident rates.

For example, Hernandez Garcia and Lozano (2024) conducted a comprehensive analysis of road accidents, focusing on the relationships between accident occurrence and various factors such as types of vehicles, external conditions like precipitation, and the day of the week. They utilized Poisson Regression and Negative Binomial models to study the data, which included information on road accidents, traffic counts per vehicle type, and precipitation levels [1].

Similarly, Gothane and Sarode (2016) investigated factors influencing road accidents using the WEKA tool and CCTV camera data. Their study highlighted several key factors contributing to accidents, including drunken driving, overloaded vehicles, and poor vehicle fitness. They emphasized the importance of strict driving license verification and proposed measures to address overcrowding and vehicle fitness issues [2].

These examples illustrate the diverse methodologies employed in accident research and underscore the importance of considering multiple factors in accident analysis and prevention strategies.

3. Methodology

3.1. Data Collection

The data utilized in this project was sourced from the UK government's official statistics on road traffic accidents. This dataset offers comprehensive insights into road accidents reported over multiple years. The dataset encompasses various attributes related to accident status, vehicle and casualty references, demographics, and severity of casualties. It includes essential factors such as pedestrian details, casualty types, road maintenance worker involvement, and the Index of Multiple Deprivation (IMD) decile for casualties' home areas. The dataset serves as a valuable resource for

analyzing road accidents and understanding the factors contributing to accident occurrence and severity.

Table 1. Data Description for Road Accident Analysis

Attribute	Description	Data Type	Missing Values
Status	The status of the accident (e.g., reported, under investigation)	object	0
Accident_Index	A unique identifier for each reported accident	object	0
Accident_Year	The year in which the accident occurred	int64	0
Accident_Reference	A reference number associated with the accident	object	0
Vehicle_Reference	A reference number for the involved vehicle in the accident	int64	0
Casualty_Reference	A reference number for the casualty involved in the accident	int64	0
Casualty_Class	Indicates the class of the casualty	int64	0
Sex_of_Casualty	The gender of the casualty	float64	448 (7.3%)
Age_of_Casualty	The age of the casualty	float64	1350 (2.2%)
Age_Band_of_Casualty	Age group to which the casualty belongs	float64	1350 (2.2%)
Casualty_Severity	The severity of the casualty's injuries (e.g., fatal, serious, slight)	int64	0
Pedestrian_Location	The location of the pedestrian at the time of the accident	int64	0
Pedestrian_Movement	The movement of the pedestrian during the accident	int64	0
Car_Passenger	Indicates whether the casualty was a car passenger at the time of the accident	float64	314 (5.1%)
Bus_or_Coach_Passenger	Indicates whether the casualty was a bus or coach passenger	float64	23 (0.4%)
Pedestrian_Road_Maintenance_Worker	Indicates whether the casualty was a road maintenance worker	float64	113 (1.8%)
Casualty_Type	The type of casualty	float64	45 (0.7%)
Casualty_Home_Area_Type	The type of area in which the casualty resides	float64	5500 (9%)
Casualty_IMD_Decile	The IMD decile of the area where the casualty resides (a measure of deprivation)	float64	5784 (9.3%)
LSOA_of_Casualty	The Lower Layer Super Output Area (LSOA) associated with the casualty's location	object	8027 (15.1%)

3.2. Data Preprocessing

The collected data underwent preprocessing steps, including handling missing values, removing irrelevant and highly correlated columns, and standardizing numerical variables using standard scaling. Also, relevant features were selected, so the model can achieve better generalization

performance and avoid overfitting. These steps aimed to improve data quality and prepare it for analysis and modeling.

3.3. Exploratory Data Analysis (EDA)

Descriptive statistics, visualizations, and correlation analysis were performed to gain insights into the dataset.

3.4. Model Development

For modeling the relationship between various factors and casualty severity, logistic regression was chosen as the primary predictive model. Logistic regression is a widely-used algorithm for multi-class classification tasks, making it an appropriate choice for predicting casualty severity levels (slight, serious, or fatal). Additionally, logistic regression provides interpretable coefficients, allowing us to assess the impact of each predictor on the likelihood of severe casualties.

The logistic regression model was evaluated using multiple metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and accuracy. MAE and MSE provide insights into the model's predictive accuracy and error magnitude, while accuracy measures the overall correctness of the model's predictions.

Overall, logistic regression offers a balance between simplicity, interpretability, and predictive performance, making it a suitable choice for modeling casualty severity in road traffic accidents.

3.5. Feature Importance Analysis

The feature importance of the logistic regression model was examined to identify significant predictors of casualty severity. This analysis helps prioritize factors that contribute significantly to the likelihood of severe accidents, enabling stakeholders to focus resources and interventions effectively.

4. Results and Discussion

4.1. Exploring Factors Influencing Casualty Severity: Insights from Relationship Analysis

Correlation Matrix

The correlation matrix of numerical variables reveals insights into the relationships between different features in the dataset. Fig.1 heatmap displays the correlations between various attributes in your data. The color intensity and the sign (positive or negative) indicate the strength and direction of the relationship between each pair of attributes.

Of particular note is the exceptionally high correlation between "age_band_of_casualty" and "age_of_casualty" (0.98), indicating a strong linear relationship between these two variables. The high correlation (0.98) between "age_band_of_casualty" and "age_of_casualty" can be attributed to the inherent relationship between the two variables. "Age_band_of_casualty" represents age groups or bands that individuals belong to, while "age_of_casualty" represents the actual age of

individuals involved in accidents. Since "age_band_of_casualty" is derived from "age_of_casualty" by grouping ages into specific bands, it's expected that there would be a strong correlation between them.

The high correlation (0.83) between "Pedestrian_Location" and "Pedestrian_Movement" can be explained by the inherent relationship between the two variables and the common scenarios they represent in pedestrian accidents. The Pedestrian_Location variable describes the specific location of the pedestrian at the time of the accident, such as crossing on a pedestrian crossing facility, walking along the carriageway, or standing on a footway. The Pedestrian_Movement variable characterizes the movement of the pedestrian during the accident, including actions like crossing from the driver's nearside, walking along the carriageway facing or back to traffic, or standing stationary. The high correlation indicates that certain locations correspond strongly with specific movements during pedestrian accidents. For example: Pedestrians crossing on pedestrian crossing facilities (Pedestrian_Location category 1) are likely to be associated with movements such as crossing from the driver's nearside (Pedestrian_Movement category 1) or crossing from the driver's offside (Pedestrian_Movement category 3).

In summary, the high correlation between pedestrian location and movement with casualty class primarily stems from the dominance of category 0 (not a pedestrian) in the data, which inherently aligns with casualty classes 1 and 2 (drivers and passengers). This emphasizes the importance of considering data biases and contextual factors when interpreting correlation results.

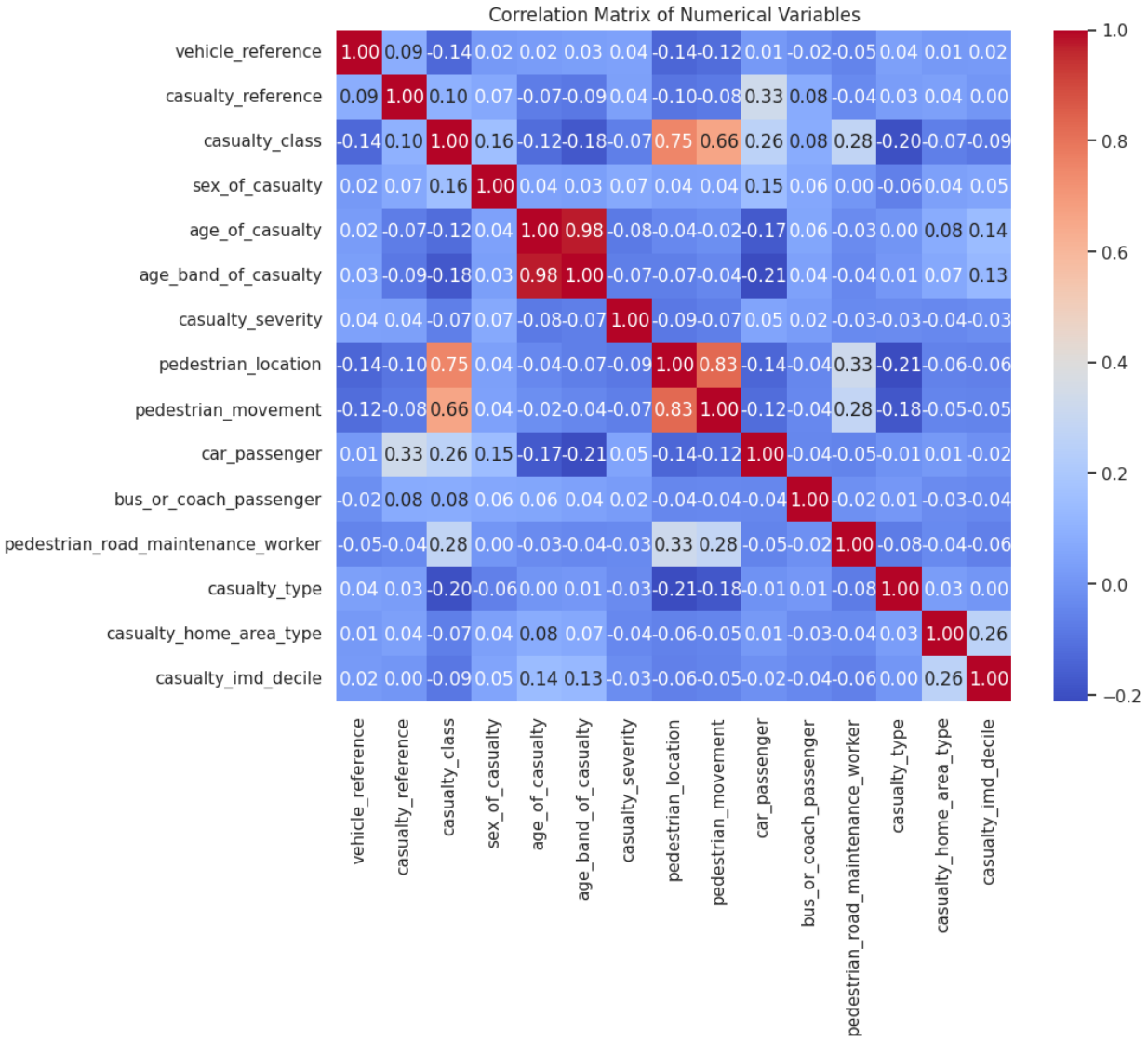

Fig 1. Correlations Between Various Attributes

These findings underscore the importance of considering various factors when analyzing casualty severity in road accidents, including demographic information, vehicle types, and accident circumstances. Further exploration and analysis of these correlations can provide valuable insights for road safety interventions and policy decisions.

Casualty Severity Distribution

The dataset comprises casualties classified into three severity levels: fatal (3), serious (2), and slight (1). The analysis of casualty severity revealed that fatal accidents accounted for approximately 80.2% of all reported incidents, making it the most prevalent severity level, indicating the severity and potential loss of life associated with road accidents. Serious injuries were the next most common, comprising approximately 18.6% of the total casualties. Slight

injuries constituted a smaller proportion at 1.1%. These findings underscore the significant impact of road accidents, with a notable portion resulting in severe or fatal outcomes.

Further analysis, including identifying contributing factors associated with different severity levels, can inform targeted interventions and policy development efforts aimed at reducing the overall impact of road accidents on individuals and communities.

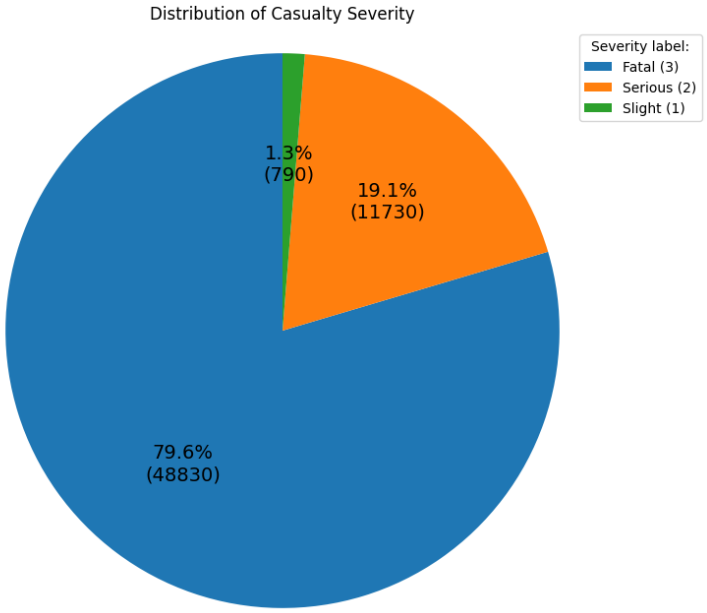

Fig2. Distribution of Casualty Severity

Casualty Class Distribution

The dataset includes casualties classified into three main categories: driver (1), passenger (2), and pedestrian (3). The majority of casualties in the dataset are drivers (40,702 cases), indicating that individuals operating vehicles are most commonly involved in road accidents. Passenger casualties represent a significant proportion (11,710 cases), highlighting the risk faced by individuals traveling as passengers in vehicles. Pedestrian casualties, while fewer in number (8,940 cases) compared to drivers and passengers, underscore the vulnerability of individuals walking or crossing roads. Understanding the distribution of casualty classes is essential for identifying at-risk populations and implementing targeted interventions to improve road safety.

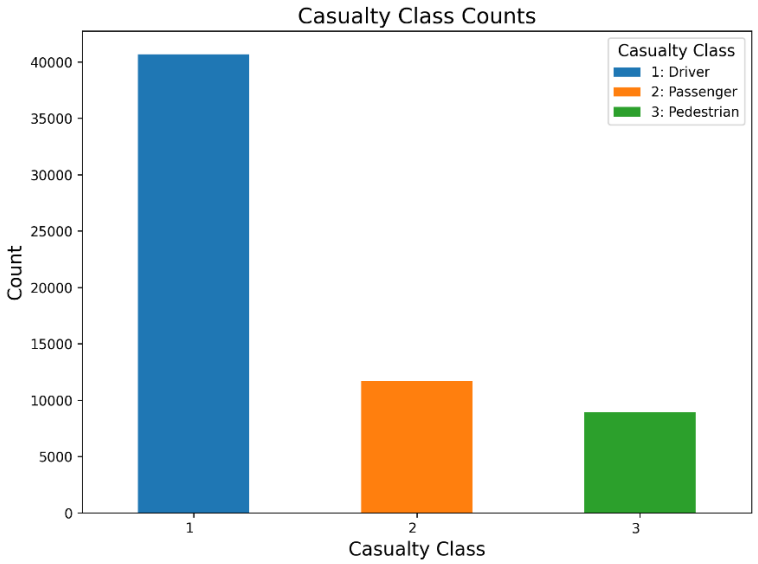

Fig 3. Distribution of Casualty Counts by Casualty Class

Strategies aimed at reducing road accidents should address the specific needs and vulnerabilities of different casualty classes. Interventions targeting drivers should focus on promoting safe driving practices, enforcing traffic laws, and raising awareness about the risks of reckless behavior. Measures to enhance passenger safety may include ensuring the use of seat belts, improving vehicle safety features, and educating passengers about their rights and responsibilities. Efforts to protect pedestrians should prioritize infrastructure improvements, such as designated crosswalks, pedestrian signals, and traffic calming measures, to create safer environments for walking and commuting.

Gender Distribution

Gender distribution reveals that a significant proportion of road accidents involve male individuals, comprising 61.5% of all accidents, while females account for 38.5%. Male casualties outnumber female casualties across all three categories of casualty severity.

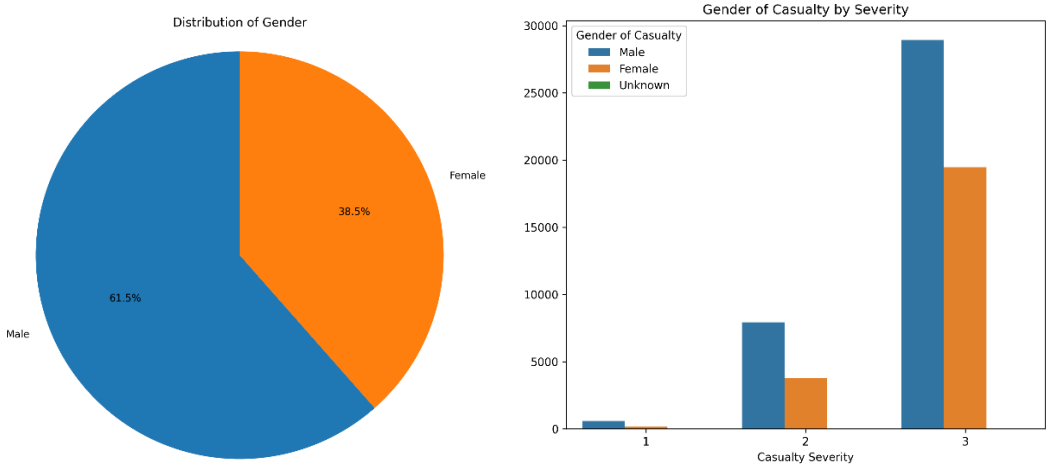

Fig 4. Analysis of Casualty Gender and Severity in Road Accident

The higher involvement of males in road accidents may be attributed to various factors, including differences in driving behavior, risk-taking tendencies, and exposure to road traffic. Understanding gender-based differences in accident involvement is crucial for designing gender-sensitive road safety interventions and policies. Strategies aimed at reducing road accidents should consider gender-specific risk factors and tailor interventions to address the unique needs and challenges faced by male and female road users.

Efforts to improve road safety should prioritize interventions that target the specific risk factors contributing to accidents among male drivers and passengers. Gender-responsive road safety programs may include initiatives to promote safe driving behaviors, increase awareness about the risks of reckless driving, and provide support services for male road users.

Age of Casualty Distribution

The age distribution of casualties reveals insights into the demographic profile of individuals involved in road accidents. Age band 6, corresponding to individuals aged between 26 and 35 years, emerges as the most represented age group among casualties, with the highest count. Following, age band 7 (36 to 45 years) also exhibits a notable frequency of casualties.

Notably, the distribution of casualty severity across age bands indicates that younger age groups, particularly those in age band 6, experience a higher frequency of severe accidents.

The concentration of casualties in age bands 6 and 7 suggests that individuals in their late twenties to mid-forties are more susceptible to road accidents. The higher prevalence of severe accidents among younger age groups underscores the vulnerability of this demographic to road traffic injuries, highlighting the need for targeted interventions to address risk factors associated with young drivers and pedestrians. The age distribution histogram (Fig. 5) provides a visual representation of the frequency of casualties across different age groups, offering further insights into the age dynamics of road accidents.

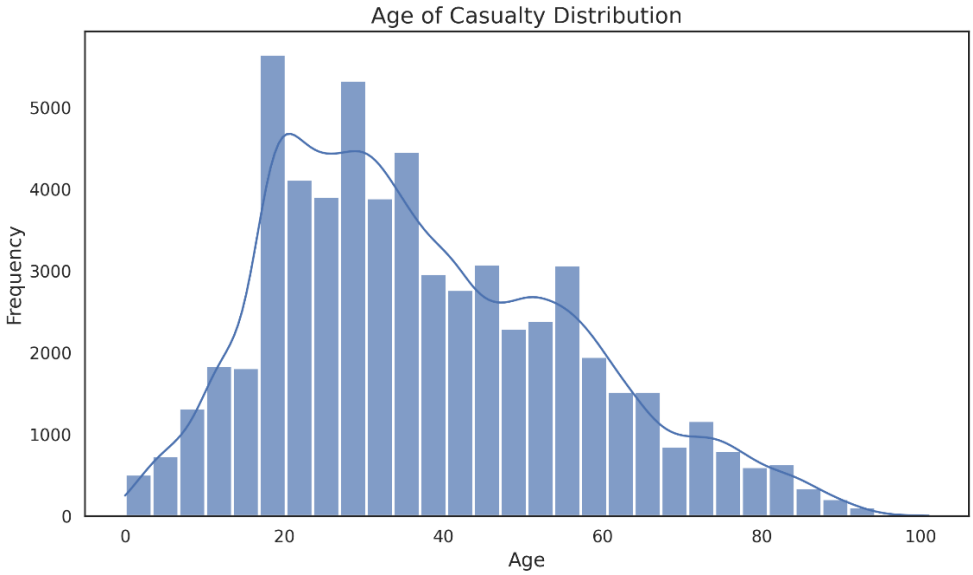

Fig 5. Age of Casualty Distribution Histogram

According to Fig 6, The box plot for fatal casualties exhibits the narrowest spread of ages, with the median age potentially being the lowest among the severity levels. This indicates that fatal accidents may be more concentrated within a specific age group compared to slight and serious casualties, potentially involving younger individuals. The interquartile range (IQR), represented by the size of the box in the box plot, is wider for slight casualties compared to serious and fatal casualties. This suggests a greater variability in ages for slight casualties, indicating a more diverse age distribution within this severity level. Understanding the age dynamics of casualty distribution across severity levels is crucial for developing targeted interventions and preventive measures tailored to the specific age demographics most affected by road accidents.

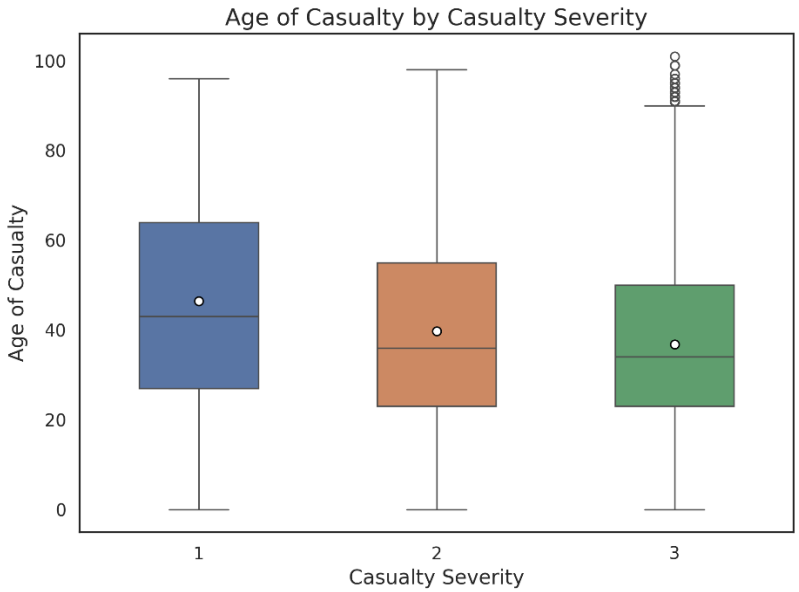

Fig 6. Box plot of Age of Casualty by Casualty Severity

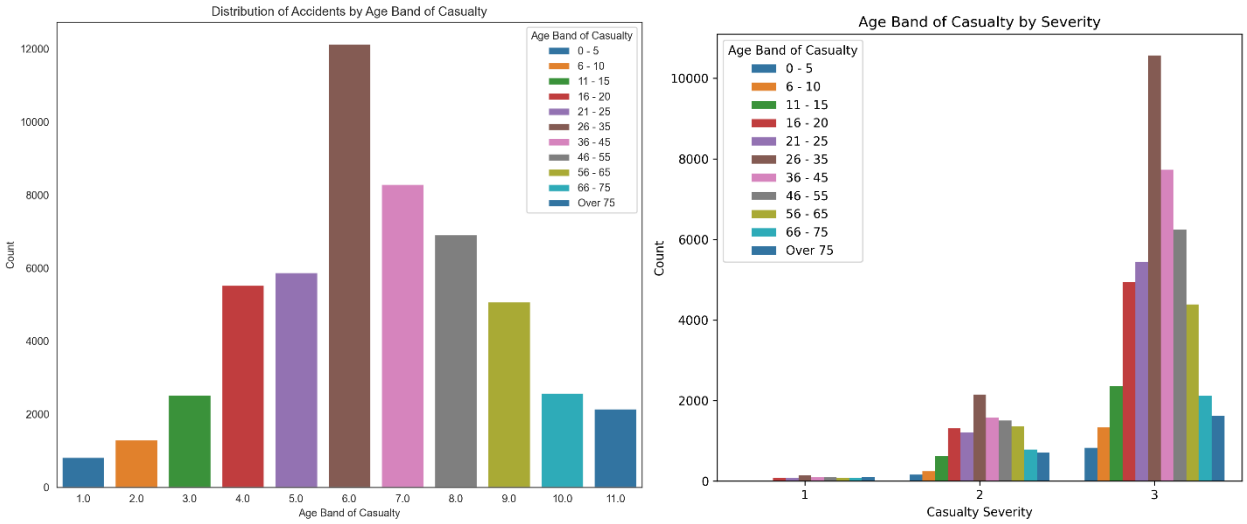

Fig 7. Analysis of Age Band and Severity in Road Accident

Road safety initiatives should prioritize interventions aimed at young adults aged between 25 and 30, as well as individuals in the 36 to 45 age brackets, to mitigate the risk of accidents and reduce the severity of injuries. Education and awareness campaigns targeting specific age groups can promote safe driving behaviors, increase awareness of road hazards, and encourage responsible road use among vulnerable demographics. Collaborative efforts involving government agencies, educational institutions, and community organizations are essential for implementing effective road safety measures tailored to the needs of different age cohorts.

Casualty Type Distribution

The distribution of casualty types provides insights into the prevalence of different types of casualties involved in road accidents. Understanding the distribution of casualty types is essential for developing targeted interventions and accident prevention measures tailored to specific road user groups.

Car occupants (Class 9) represent the most prevalent casualty type, with a count of 32,715. This suggests that individuals traveling in cars are most commonly involved in road accidents, highlighting the importance of implementing safety measures within vehicles to reduce the risk of injuries. Pedestrians (Class 0) constitute the second-highest casualty type, with a count of 8,940. Pedestrian safety measures such as improved infrastructure, crosswalks, and pedestrian education programs are crucial for reducing pedestrian-related accidents and injuries. Cyclists (Class 1) represent another significant casualty type, with a count of 7,155. Enhancing cycling infrastructure, such as dedicated bike lanes and cyclist awareness campaigns, can help mitigate the risk of accidents involving cyclists on the road. Other casualty types, including motorcyclists, taxi occupants, van/goods vehicle occupants, and agricultural vehicle occupants, also contribute to the overall casualty count but with smaller proportions compared to car occupants, pedestrians, and cyclists.

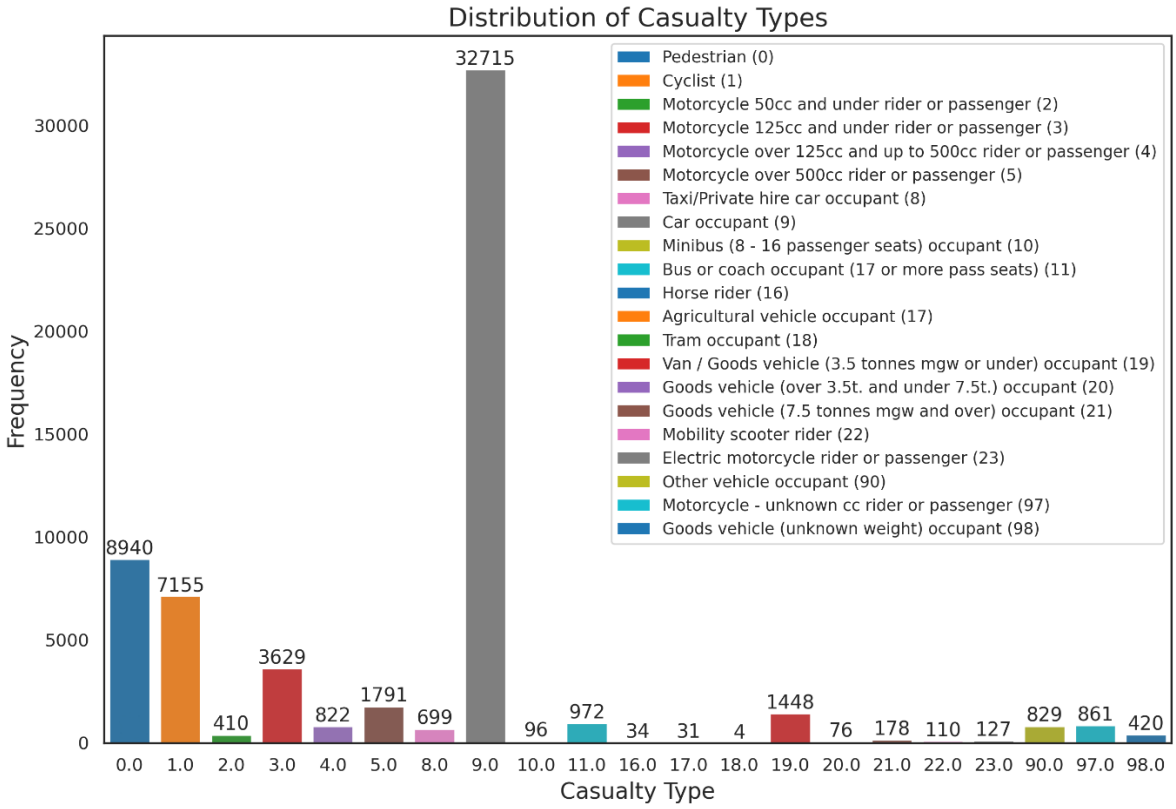

Fig 8. Distribution of Casualty Type

The high prevalence of car occupants, pedestrians, and cyclists underscores the need for targeted safety measures tailored to these road user groups. Implementing measures such as seat belt enforcement, pedestrian-friendly infrastructure, and cyclist safety initiatives can help reduce the risk of accidents and injuries. Investing in infrastructure improvements, such as dedicated pedestrian walkways, cycling lanes, and traffic calming measures, can enhance road safety for vulnerable road users such as pedestrians and cyclists. Public awareness campaigns focusing on safe driving practices, pedestrian safety, and cyclist awareness can help raise awareness among road users and promote safer behaviors on the road.

Pedestrian Location and Movement

The distribution of pedestrian locations reveals that the majority of accidents involve individuals who are not pedestrians, accounting for the highest count. This observation aligns with the distribution of casualty classes, where pedestrian casualties have lower counts compared to other classes such as drivers and passengers. Among pedestrian accidents, the most common location is "In carriageway, crossing elsewhere," suggesting that accidents frequently occur when pedestrians attempt to cross roads outside designated crossing facilities.

In terms of pedestrian movement during accidents, the distribution highlights that accidents often occur when pedestrians are crossing from the driver's nearside or offside. This suggests that accidents involving pedestrians frequently occur near vehicles, emphasizing the importance of pedestrian awareness and driver vigilance, particularly at intersections and crossings. Additionally,

the presence of pedestrians' stationary in the carriageway or walking along facing or away from traffic indicates potential hazards for both pedestrians and motorists.

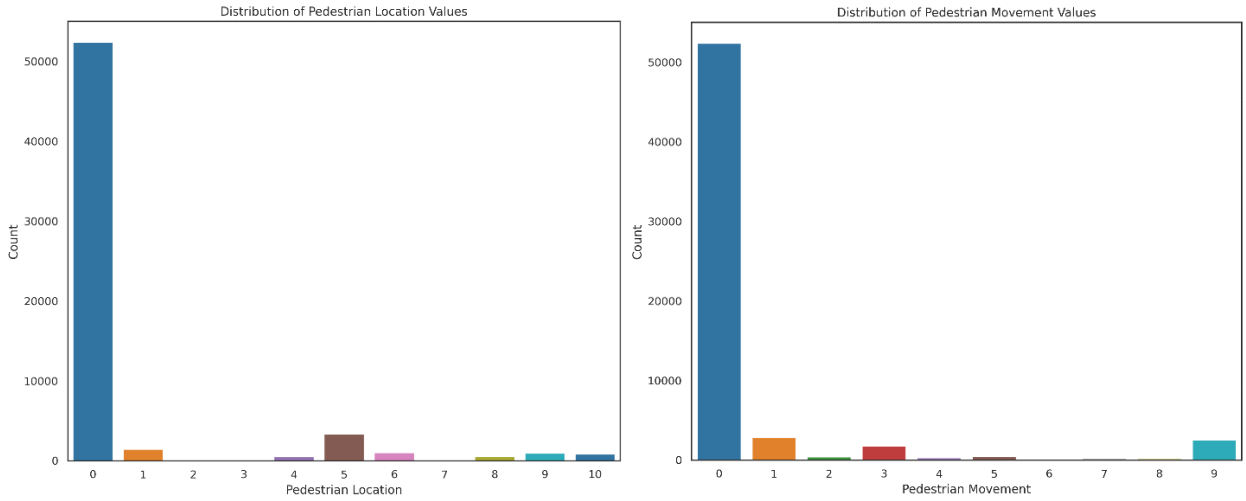

Fig 9. Distribution of Pedestrian Movement and Location

Overall, understanding the distribution of pedestrian locations and movements can inform targeted interventions and infrastructure improvements aimed at reducing pedestrian accidents and enhancing road safety for all road users.

Home Area Type

The distribution of casualty home area types provides insights into the geographic locations where casualties reside and their association with road accident severity. Understanding the distribution of home area types is crucial for implementing targeted interventions and policies aimed at mitigating the risk of accidents in different geographic settings.

Urban areas represent the most prevalent home area type among casualties, accounting for 81.1% of cases. This suggests that a significant proportion of casualties reside in densely populated urban settings where traffic volumes and potential accident risks are higher. Rural areas are the second most prevalent home area type, with 10.7% of cases. While rural areas typically have lower population densities and traffic volumes compared to urban areas, they still experience a notable number of road accidents, highlighting the importance of addressing road safety concerns in rural communities. Small towns have the lowest prevalence of casualties among the three home area types, with 8.2% cases. While small towns may have intermediate population densities and traffic levels compared to urban and rural areas, they still contribute to the overall casualty count and warrant attention in road safety initiatives.

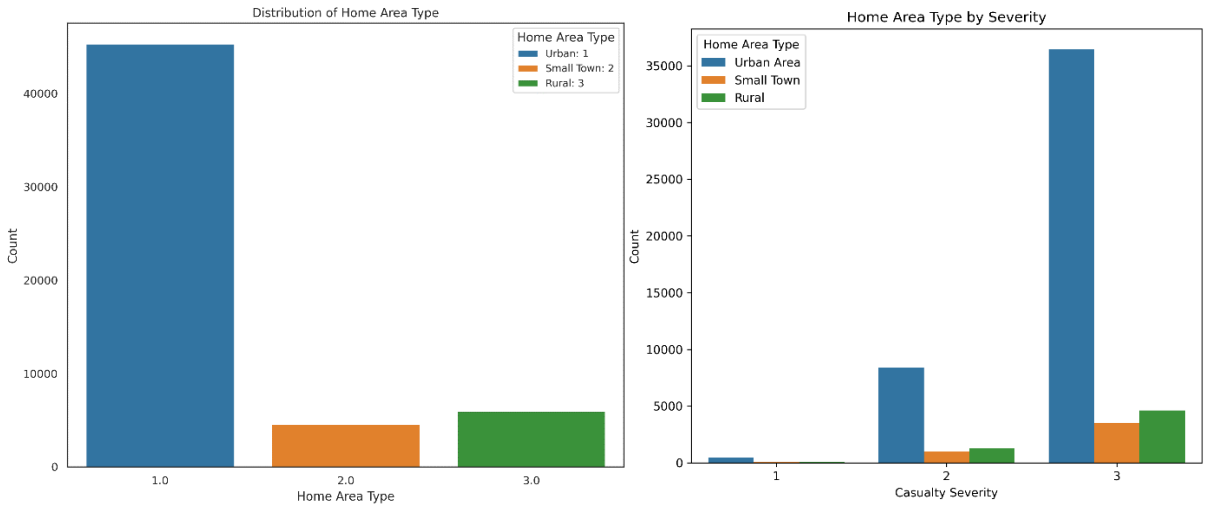

Fig 10. Analysis of Type of Area Which Casualty Occur and Casualty Severity

Given the higher prevalence of casualties in urban areas, targeted road safety measures and infrastructure improvements should be prioritized in urban settings to address the specific challenges associated with dense traffic, pedestrian activity, and vehicle congestion. While rural areas have fewer casualties compared to urban areas, the severity of accidents in rural settings can be higher due to factors such as higher speeds and limited access to emergency services. Therefore, road safety initiatives in rural areas should focus on speed management, enhanced signage, and improved emergency response capabilities. Implementing a comprehensive approach to road safety that addresses the unique characteristics and challenges of different home area types is essential for reducing the incidence of road accidents and minimizing their impact on casualties. This approach may include infrastructure improvements, public awareness campaigns, enforcement of traffic laws, and collaboration with local communities and stakeholders.

Casualty IMD Decile Distribution

The IMD decile of the area where a casualty resides serves as a measure of deprivation, providing insights into the socioeconomic conditions of the communities affected by road accidents. Understanding the distribution of IMD deciles among casualties can shed light on the association between socioeconomic factors and accident severity, thereby informing targeted interventions and policy initiatives aimed at addressing disparities in road safety outcomes.

The analysis reveals that casualties residing in the most deprived areas, represented by IMD decile 1 (Most deprived 10%), exhibit a higher prevalence of severe casualties, particularly those classified as serious or fatal. This suggests a possible correlation between higher levels of deprivation and increased likelihood of severe outcomes in road accidents. Conversely, casualties residing in less deprived areas, characterized by IMD deciles 7 to 10, show a lower prevalence of severe casualties compared to their counterparts in more deprived areas. This trend indicates a potential protective effect associated with lower levels of deprivation, contributing to reduced severity of road accidents in these areas. Regional disparities in casualty IMD decile distribution highlight variations in socioeconomic conditions and road safety outcomes across different geographic areas. Policymakers and stakeholders can use this information to target resources and

interventions towards communities with higher levels of deprivation, aiming to address underlying socioeconomic factors contributing to road accidents and their severity.

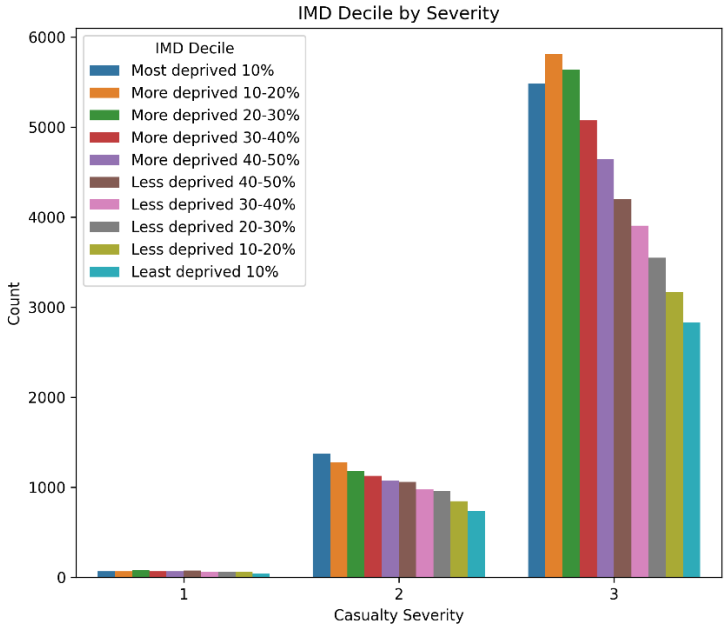

Fig 11. Analysis of IMD Decile and Severity of Road Accident

Tailoring road safety interventions to address the specific needs and challenges faced by communities in different IMD decile categories is essential for maximizing their effectiveness. Strategies may include improving infrastructure, enhancing access to education and employment opportunities, and implementing measures to mitigate the impact of socioeconomic disparities on road safety outcomes. Engaging with local communities and stakeholders to identify and address the root causes of deprivation and associated road safety risks is crucial for developing sustainable solutions. Collaborative efforts that involve community members in decision-making processes can lead to more effective interventions and greater acceptance and adoption of road safety initiatives. Continuously monitoring and analyzing casualty IMD decile distribution alongside other road safety indicators can provide valuable insights for policymakers and practitioners. By leveraging data-driven approaches, decision-makers can prioritize resources, evaluate the impact of interventions, and adapt strategies to address evolving road safety challenges.

4.2. Exploring Factors Influencing Casualty Severity: Insights from Logistic Regression Model

Based on the feature importance results from the logistic regression model, the factors contributing most significantly to casualty severity are as follows:

Vehicle Reference (20.43%): The specific vehicle involved in the accident plays a substantial role in determining the severity of casualties.

Age Band of Casualty (19.69%): The age group to which the casualty belongs significantly influences the severity of the outcome.

Casualty Class (11.73%): Whether the casualty is a driver, passenger, or pedestrian impacts the severity of the accident.

Sex of Casualty (13.41%): Gender differences contribute to variations in casualty severity.

Bus or Coach Passenger (11.31%): The involvement of bus or coach passengers affects the severity of casualties.

Car Passenger (7.76%): Similarly, being a car passenger is associated with different levels of severity.

Pedestrian Location (7.12%): The location of pedestrians at the time of the accident influences casualty severity.

Casualty Home Area Type (6.76%): The type of area where the casualty resides contributes to varying levels of severity.

Casualty Type (4.75%): The type of casualty, such as pedestrian, cyclist, or vehicle occupant, affects severity.

Pedestrian Movement (3.62%): The movement of pedestrians during accidents impacts casualty severity.

Casualty Reference (6.66%): The specific reference number of the casualty also plays a role in determining severity.

Casualty IMD Decile (2.63%): The Index of Multiple Deprivation (IMD) decile for casualties' home areas contributes to severity.

Pedestrian Road Maintenance Worker (0.92%): Involvement of pedestrian road maintenance workers slightly affects casualty severity.

These results highlight the multifaceted nature of factors influencing casualty severity in road accidents, encompassing both individual characteristics and contextual factors.

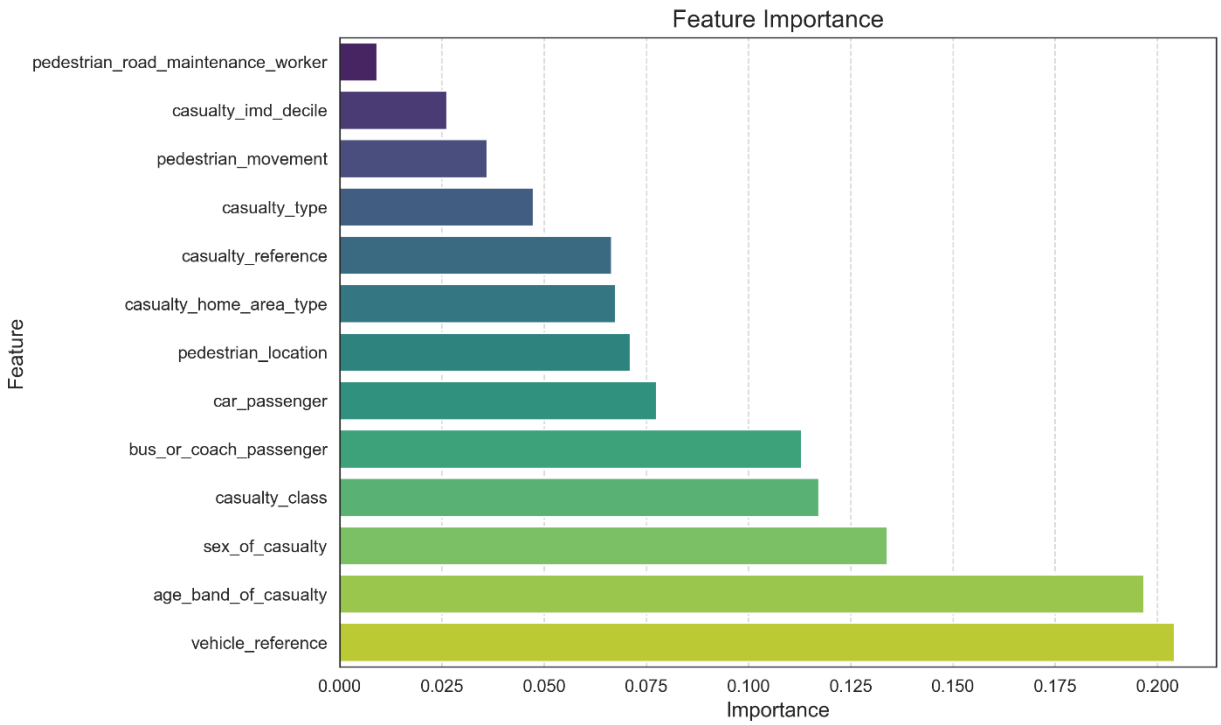

Fig 12. Feature Importance from Logistic Regression Model

These findings provide valuable insights into the factors driving casualty severity, helping to prioritize interventions and mitigate risks on the road.

5. Conclusion

In conclusion, this report provides valuable insights into road accident analysis and offers recommendations for reducing accidents and enhancing road safety. By leveraging data-driven approaches and implementing targeted interventions, stakeholders can work towards the goal of creating safer roadways for all users.

References

[1] H. G. M. Eliane and L. Angelica, "Analysis of road accidents in two mixed industrial urban zones, using nested Poisson and Negative Binomial models," *Transportation Research Procedia,* vol. 78, pp. 377-383, 2024/01/01/ 2024, doi: https://doi.org/10.1016/j.trpro.2024.02.048.

[2] S. Gothane and M. Sarode, *Analyzing Factors, Construction of Dataset, Estimating Importance of Factor, and Generation of Association Rules for Indian Road Accident*. 2016, pp. 15-18.