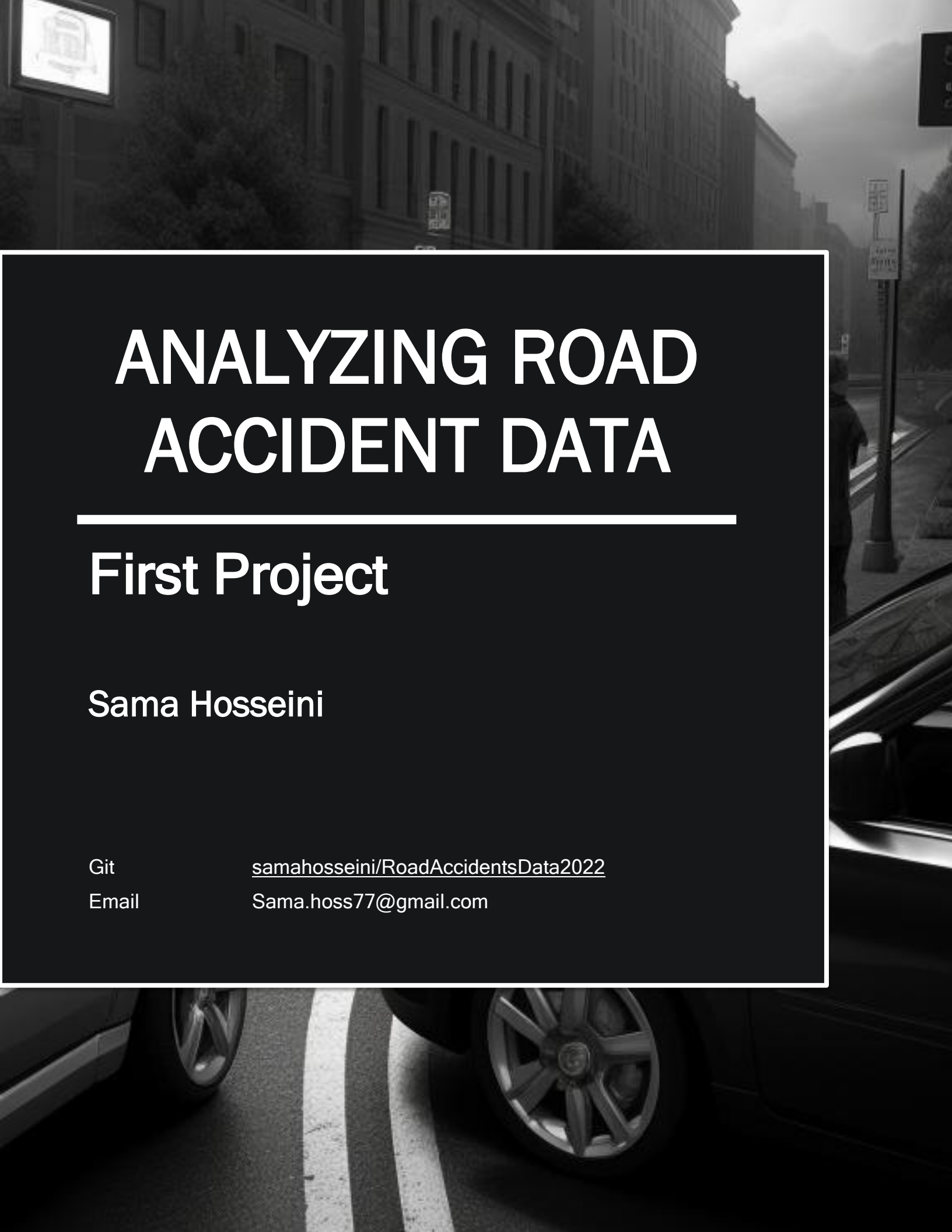


ANALYZING ROAD ACCIDENT DATA

First Project

Sama Hosseini

Git samahosseini/RoadAccidentsData2022

Email Sama.hoss77@gmail.com

Road Accident Data Analyze in Python

Introduction

This report provides an overview and analysis of the "Road Accidents Dataset," a comprehensive dataset containing detailed information on road accidents reported in 2022. The dataset includes various attributes related to accident status, vehicle and casualty references, demographics, and the severity of casualties. It serves as a valuable resource for analyzing road accidents, identifying trends, and implementing safety measures to reduce casualties and enhance road safety.

About Dataset

"Road Accidents Dataset":

This comprehensive dataset provides detailed information on 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.

Columns:

Status: The status of the accident (e.g., reported, under investigation).
Accident_Index: A unique identifier for each reported accident.
Accident_Year: The year in which the accident occurred.
Accident_Reference: A reference number associated with the accident.
Vehicle_Reference: A reference number for the involved vehicle in the accident.
Casualty_Reference: A reference number for the casualty involved in the accident.
Casualty_Class: Indicates the class of the casualty (e.g., driver, passenger, pedestrian).
Sex_of_Casualty: The gender of the casualty (male or female).
Age_of_Casualty: The age of the casualty.

Age_Band_of_Casualty: Age group to which the casualty belongs (e.g., 0-5, 6-10, 11-15).

Casualty_Severity: The severity of the casualty's injuries (e.g., fatal, serious, slight).

Pedestrian_Location: The location of the pedestrian at the time of the accident.

Pedestrian_Movement: The movement of the pedestrian during the accident.

Car_Passenger: Indicates whether the casualty was a car passenger at the time of the accident (yes or no).

Bus_or_Coach_Passenger: Indicates whether the casualty was a bus or coach passenger (yes or no).

Pedestrian_Road_Maintenance_Worker: Indicates whether the casualty was a road maintenance worker (yes or no).

Casualty_Type: The type of casualty (e.g., driver/rider, passenger, pedestrian).

Casualty_Home_Area_Type: The type of area in which the casualty resides (e.g., urban, rural).

Casualty_IMD_Decile: The IMD decile of the area where the casualty resides (a measure of deprivation).

LSOA_of_Casualty: The Lower Layer Super Output Area (LSOA) associated with the casualty's location.

Data Exploration

shape of the Dataset:

- The dataset contains 61,352 rows and 20 columns.

Data Types of Columns:

- The dataset comprises columns of various data types, including object and integer.

Missing Values:

- There are missing values in the dataset, with varying counts across different columns. However, certain columns, such as status, accident_year, accident_reference, and others, have no missing values.

Descriptive Statistics:

Descriptive statistics provide insights into the distribution of numerical columns in the dataset.

- The mean age of casualties is approximately 36.67 years, with a standard deviation of 19.57 years.
- The majority of casualties belong to the age band between 0-5 and 6-10.
- The most common casualty severity is categorized as slight, followed by serious and fatal.
- Various other statistics such as mean, median, minimum, and maximum values provide a comprehensive understanding of the numerical data.
- Unique Value Counts
- Each column has a different count of unique values, indicating the diversity of information captured in the dataset. For example, the casualty_type column has 22 unique values, representing different types of casualties involved in accidents.

DESCRIPTIVE STATISTICS: SUMMARY STATISTICS FOR NUMERICAL VARIABLES

The summary statistics provide insights into the distribution and variability of numerical variables in the dataset. Here's a breakdown of the key statistics:

Accident Year:

Mean: 2022.0

Standard Deviation: 0.0

Minimum: 2022.0

Maximum: 2022.0

Vehicle Reference:

Mean: 1.450368

Standard Deviation: 1.109855

Minimum: 1.0

Maximum: 227.0

Casualty Reference:

Mean: 1.333779

Standard Deviation: 0.981507

Minimum: 1.0

Maximum: 148.0

Other Numerical Variables:

The statistics include mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile), 75th percentile (Q3), and maximum values for each numerical variable.

Missing Values:

Some numerical columns have missing values, as indicated by the difference in the "count" row from the total number of rows (61352).

These statistics serve as a foundation for understanding the central tendency, spread, and range of values present in the numerical variables of the dataset.

FREQUENCY TABLES FOR CATEGORICAL VARIABLES

The frequency tables provide a breakdown of the occurrences of categorical values within each categorical variable in the dataset. Here's a summary of the frequency tables:

Status:

Unvalidated: 61352

Accident Index:

This column contains unique identifiers for each reported accident. The frequency table shows the count of occurrences for each accident index, indicating the number of casualties associated with each accident.

Accident Reference:

Similar to the accident index, this column also contains reference numbers for each reported accident. The frequency table displays the count of occurrences for

each accident reference, providing insight into the distribution of accidents in the dataset.

LSOA of Casualty:

LSOA (Lower Layer Super Output Area) is a geographical area used for statistical purposes in the UK. The frequency table shows the count of occurrences for each LSOA of the casualty, indicating the geographical distribution of casualties across different areas.

These frequency tables offer valuable insights into the distribution and occurrence of categorical values within the dataset, aiding in understanding the composition and diversity of categorical variables.

CORRELATION MATRIX:

The heatmap displays the correlation coefficients between various numerical variables in your dataset. Here's an interpretation of the heatmap:

- **Strong Positive Correlations:**

Age_of_casualty and age_band_of_casualty: This indicates a near-perfect positive correlation, as expected, as the age band is directly derived from the age of the casualty.
car_passenger and vehicle_reference: This suggests a potential link between being a car passenger and the specific vehicle reference value, which might be related to car type or occupancy.
bus_or_coach_passenger and vehicle_reference: Similar to the previous observation, this correlation suggests a possible association between being a bus or coach passenger and the vehicle reference value.

- **Strong Negative Correlations:**

age_of_casualty and casualty_imd_decile: This indicates a negative correlation, suggesting that younger casualties tend to be from more deprived areas (lower IMD decile) compared to older casualties.

age_band_of_casualty and casualty_imd_decile: Similar to the previous observation, this negative correlation suggests a link between younger age groups and lower IMD deciles (more deprived areas).

- **Other Interesting Observations:**

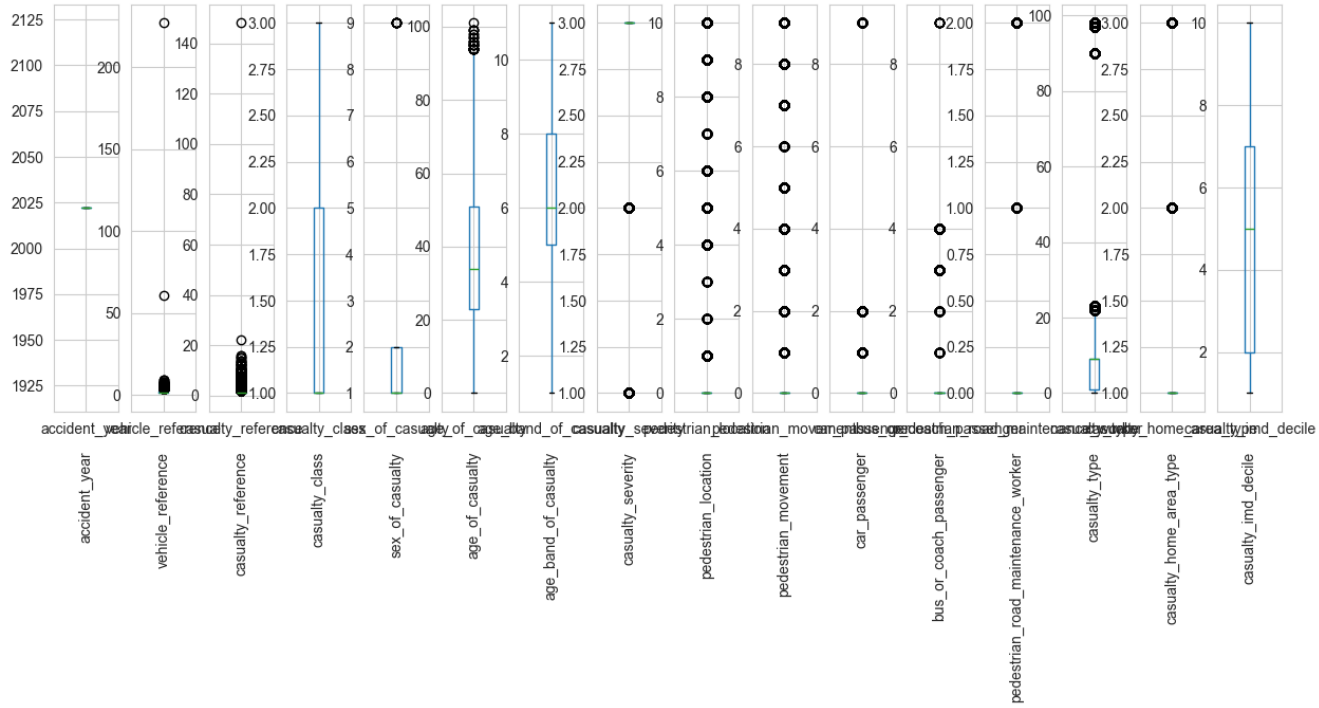
Casualty severity exhibits weak positive correlations with **age_of_casualty** and **age_band_of_casualty**, suggesting that older individuals might be slightly more prone to severe injuries. However, it's crucial to interpret this cautiously due to the weak correlation strength.

Pedestrian_location and **pedestrian_movement** show a weak positive correlation, which might be due to pedestrians in certain locations being more likely to be engaged in specific movements (e.g., walking on sidewalks).

Data Cleaning

OUTLIERS

By examining the data, we can identify outliers or inconsistencies by looking at the distribution of ages using a box plot. This visualization will help us spot any extreme values that deviate significantly from the majority of the data points.



To address outliers, rows with vehicle_reference values of 227 and 61 were removed from the dataset. This decision was made to ensure the integrity of the data analysis, as these values appeared to be outliers compared to the majority of the data.
And for the same reason, the rows with casualty_reference values of 22 and 148 were removed from the dataset.

To refine the dataset for analysis, the following actions were taken:
Data Cleaning Summary:
To refine the dataset for analysis, the following actions were taken:
Column Removal:

Accident Year: Removed as all accidents occurred in 2022.

Accident Reference: Dropped since it doesn't contribute directly to analysis.

Status: Removed as all entries were labeled as "Unvalidated."

Removing Duplicate Records:

Duplicate entries were removed to ensure data integrity and uniqueness.

These steps streamline the dataset, ensuring its accuracy and relevance for analysis.

IMPUTATION OF MISSING VALUES

The missing values in the dataset were imputed using the SimpleImputer class from the scikit-learn library. Two strategies were employed:

- **Imputation with Mode:**

Columns Imputed: 'sex_of_casualty', 'casualty_type', 'car_passenger', 'bus_or_coach_passenger', 'pedestrian_road_maintenance_worker'
Strategy: Most frequent value (mode)
Description: The mode (most frequent value) of each respective column was calculated and used to fill in missing values.

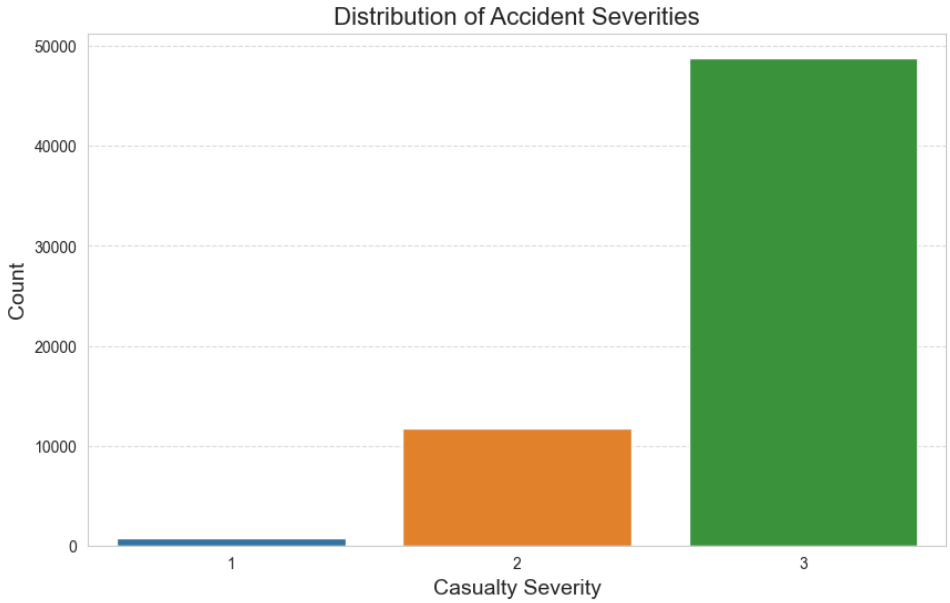
- **Imputation with Median:**

Columns Imputed: 'age_of_casualty', 'age_band_of_casualty'
Strategy: Median value
Description: The median value of each respective column was calculated and used to fill in missing values.
After imputation, any remaining rows with missing values were dropped from the dataset to ensure completeness.

Exploratory Data Analysis (EDA) Report:

CASUALTY SEVERITY DISTRIBUTION:

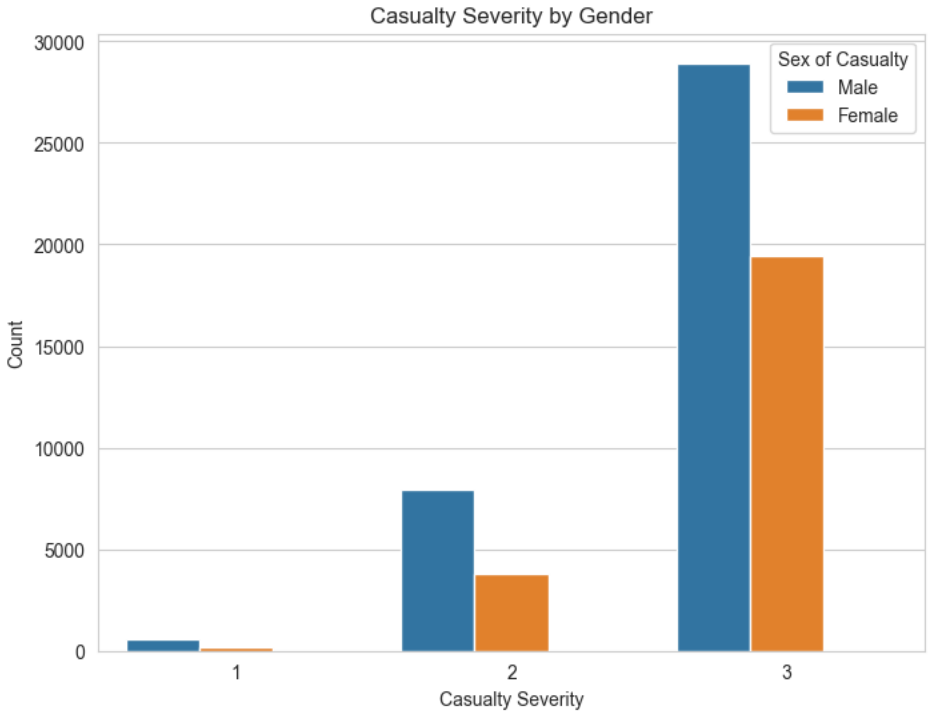
The distribution of accident severities was visualized using a bar plot. The findings indicate that the dataset is dominated by accidents resulting in slight injuries (severity level 3), with over 40,000 occurrences. Serious injuries (severity level 2) are less frequent, with under 10,000 accidents, while fatal accidents (severity level 1) are the least common, with fewer than 500 reported incidents. This distribution highlights the prevalence of less severe accidents in the dataset.



COMPARE THE DISTRIBUTION OF CASUALTY SEVERITY ACROSS DIFFERENT DEMOGRAPHIC GROUPS

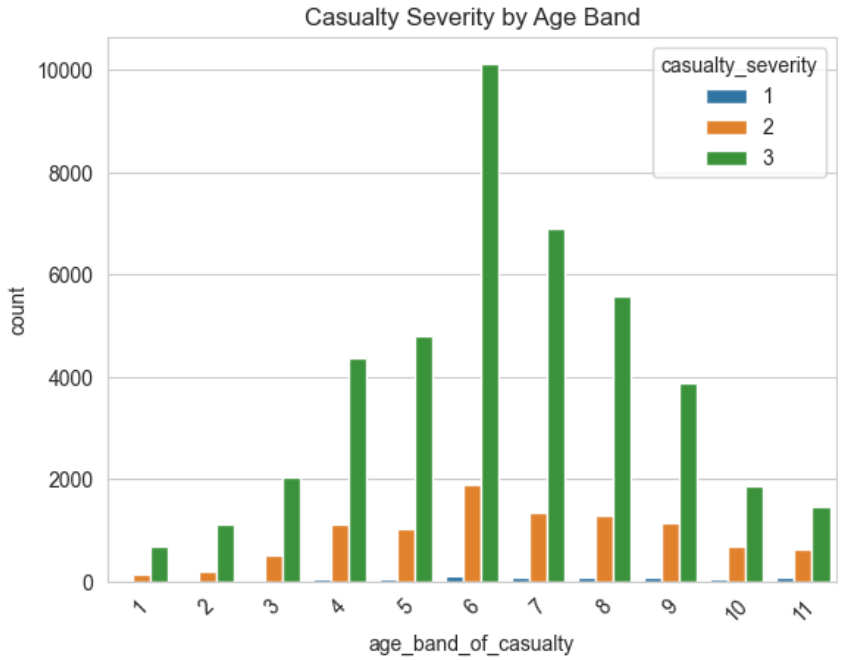
- ### GENDER VS. CASUALTY SEVERITY

The chart reveals a concerning trend: males are disproportionately represented across all casualty severity categories. This suggests they might be more likely to be involved in road accidents, and potentially suffer more serious injuries or fatalities when they do occur.



- AGE BAND VS. CASUALTY SEVERITY

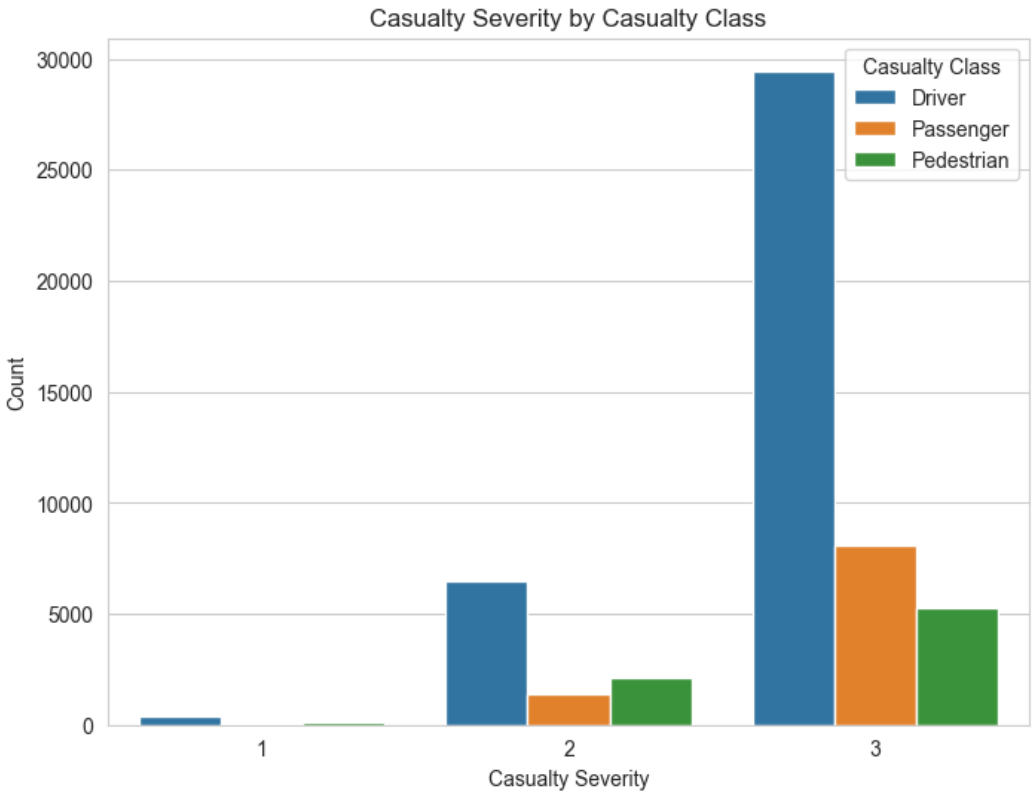
The chart highlights a concerning pattern: younger age groups (0-30) tend to have more "Slight" casualties, while older adults (61+) experience a higher proportion of "Serious" and "Fatal" injuries.



- CASUALTY SEVERITY AND CASUALTY CLASS

Drivers and pedestrians appear to be more vulnerable to severe injuries: The chart highlights a concerning trend where drivers and pedestrians are significantly more likely to be classified as "Serious" or "Fatal" casualties compared to passengers. This could be attributed to several factors, such as the greater impact forces experienced by these individuals during collisions, or their potentially exposed positions within the transportation system.

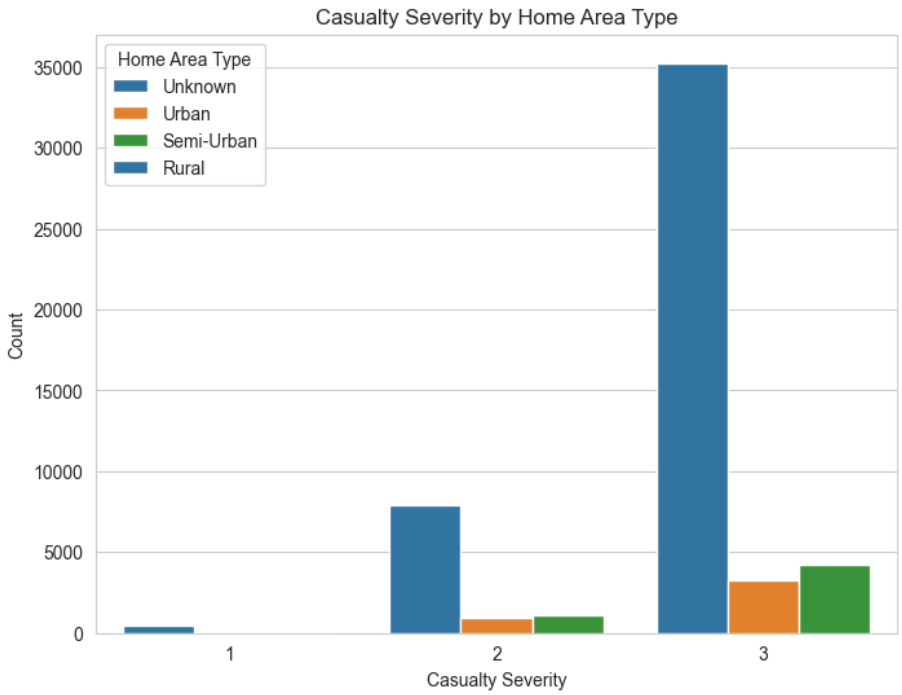
Passengers tend to sustain less severe injuries: The data suggests that passengers are more likely to walk away with "Slight" injuries compared to other casualty classes. This might be due to the protective measures offered by modern vehicles, such as airbags and seatbelts, which are primarily designed for passenger safety. However, it's crucial to remember that this observation should not downplay the potential dangers of being a passenger in a road accident.



- ## CASUALTY SEVERITY AND CASUALTY'S HOME AREA TYPE

Residents of rural areas appear to be disproportionately impacted by severe injuries in road accidents, as indicated by the higher number of "Serious" and "Fatal" casualties compared to their urban counterparts. This concerning trend might be attributed to factors like higher speed limits on rural roads, limited safety infrastructure, and potentially delayed access to medical care in remote locations.

Conversely, individuals residing in urban areas seem to experience a higher incidence of "Slight" injuries. This could be due to lower speed limits, the presence of safer road infrastructure like crosswalks and traffic signals, and potentially quicker access to emergency medical services.



- ## CASUALTY SEVERITY AND CASUALTY TYPES

The most common casualty type in the dataset is "Car", followed by "Motorcycle", "Pedestrian", and "Van". This suggests that car accidents are the most common type of road accident, followed by motorcycle accidents, pedestrian accidents, and van accidents.

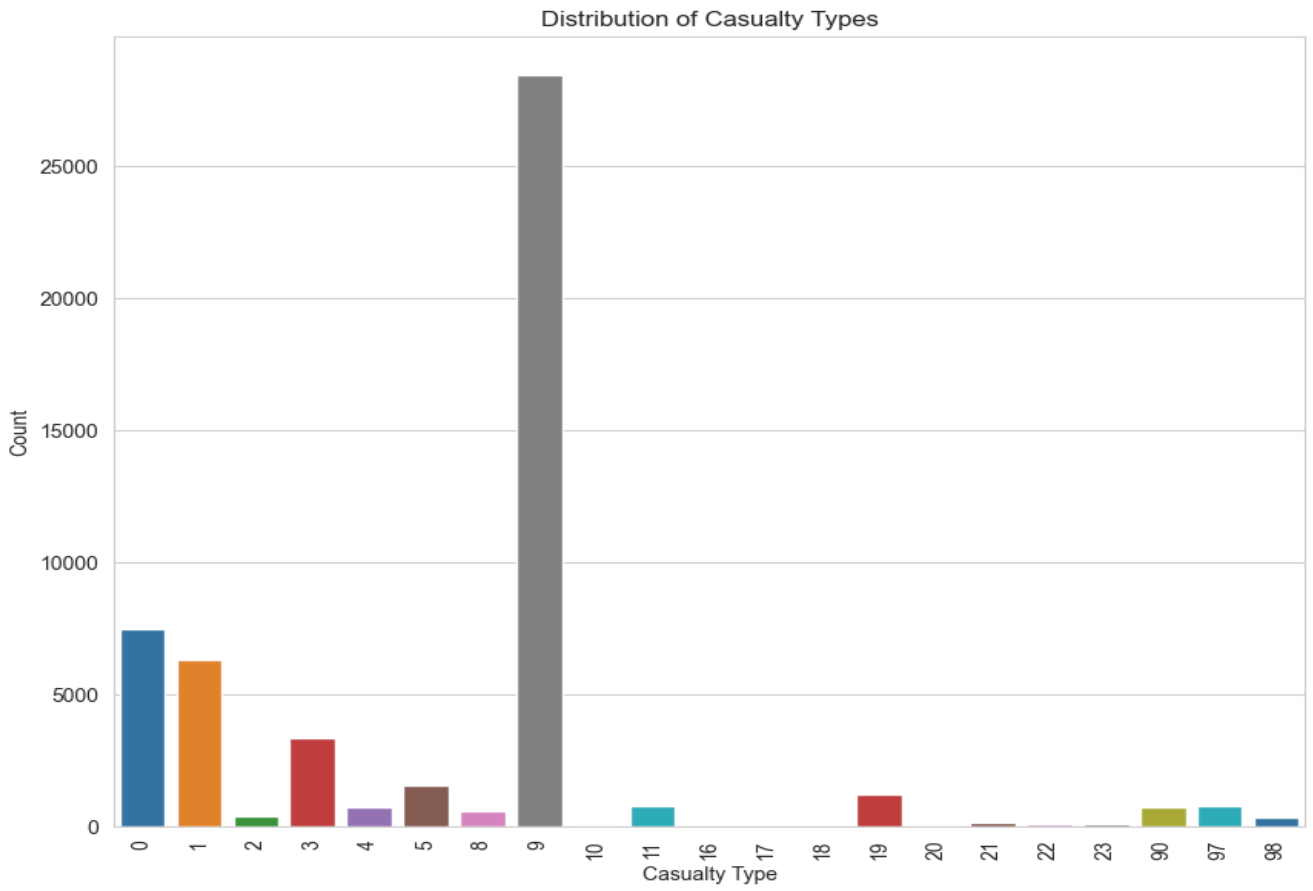
It is important to note that this chart only shows the number of casualties for each type of accident and does not take into account the severity of the injuries sustained. It is also important to remember that this data may not be representative of all road accidents, as it is likely to be based on a specific dataset.

Here are some additional insights that can be drawn from the chart:

The number of car casualties is significantly higher than the number of casualties for any other type of vehicle. This could be due to the fact that cars are the most common type of vehicle on the road.

The number of motorcycle casualties is higher than the number of van casualties. This could be due to the fact that motorcycles are more vulnerable to injury in collisions than vans.

The number of pedestrian casualties is relatively high. This highlights the importance of pedestrian safety measures, such as crosswalks and traffic signals.

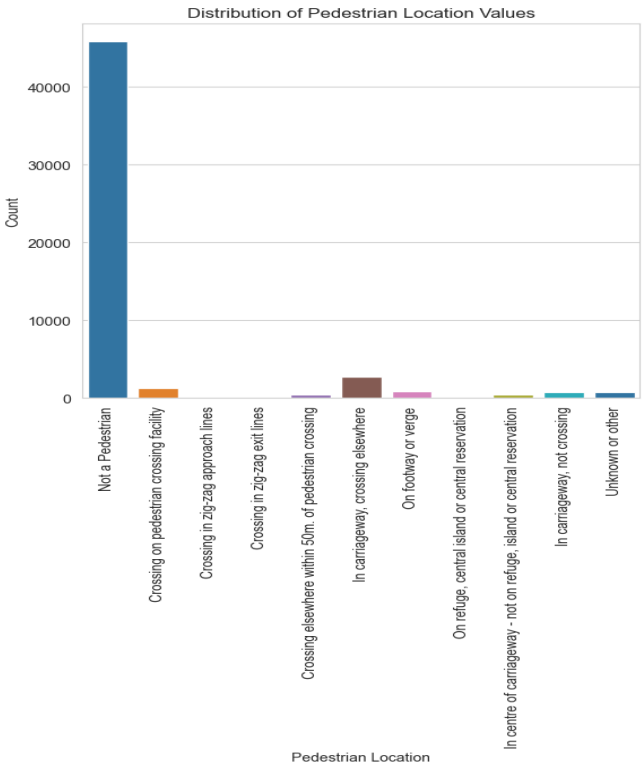


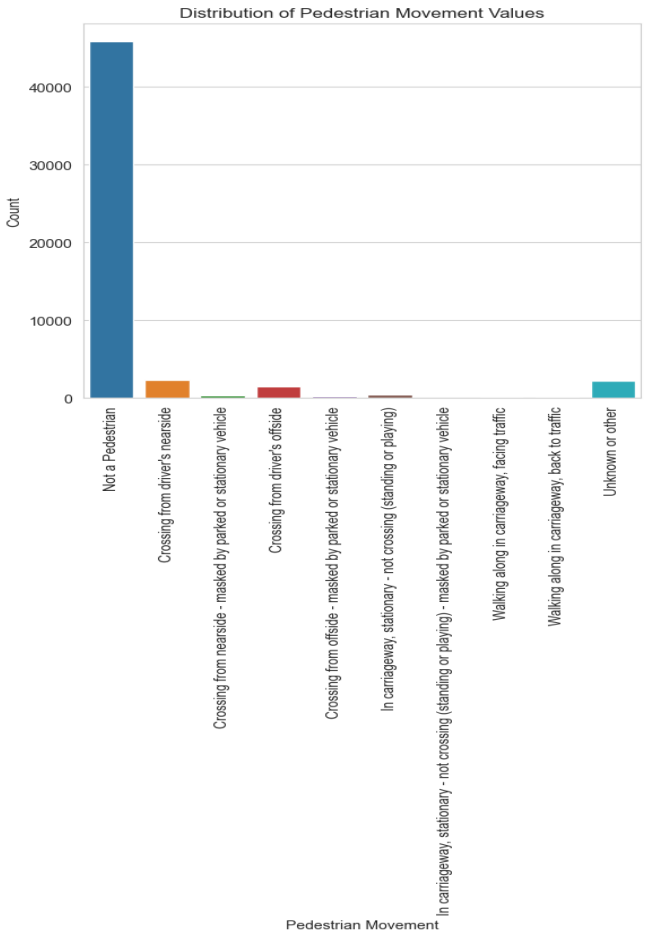
- CASUALTY SEVERITY AND PEDESTRIAN LOCATION AND MOVEMENTS VALUES

The distribution of pedestrian location values and pedestrian movement values is shown in the image. The left side of the image shows the distribution of pedestrian locations, while the right side shows the distribution of pedestrian movements.

Pedestrian Location: The most common pedestrian location is "On Road", followed by "Sidewalk", "Crossing", and "Other". This suggests that a significant proportion of pedestrians are injured while they are on the road, rather than on sidewalks or in designated crossing areas.

Pedestrian Movement: The most common pedestrian movement is "Walking", followed by "Standing", and "Running". This suggests that most pedestrians are injured while they are simply walking, rather than engaged in other activities such as running or standing.





relationship between pedestrian location, pedestrian movement and casualty severity

The boxplots you generated show the distribution of casualty severity across different pedestrian locations and pedestrian movements. The left side of the image shows the boxplots for pedestrian location, while the right side shows the boxplots for pedestrian movement.

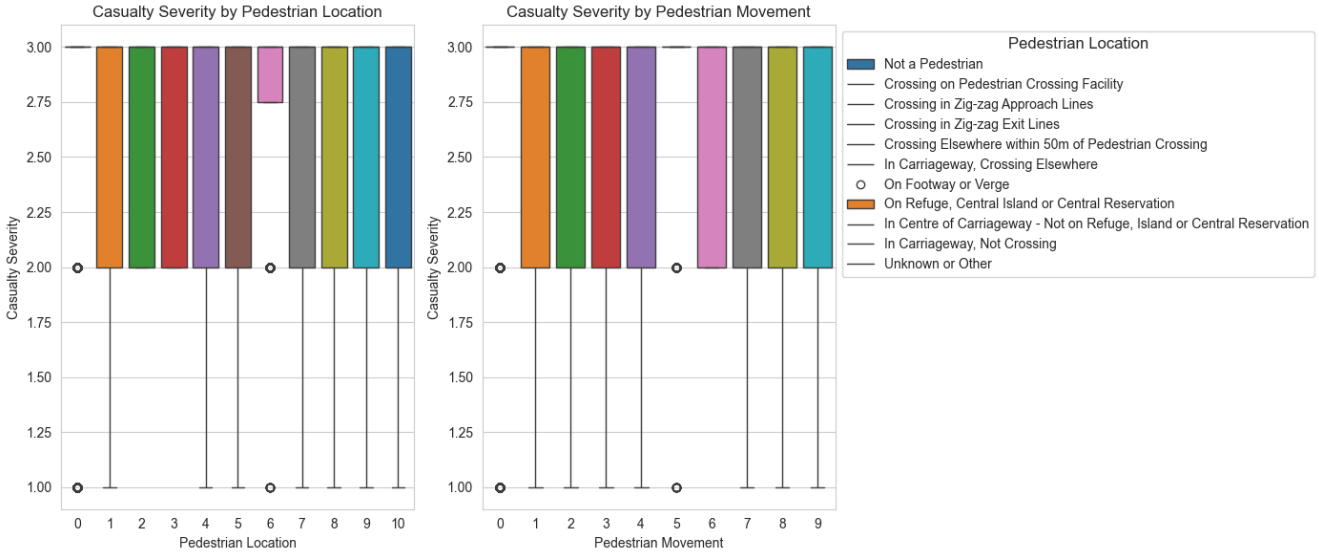
Casualty Severity by Pedestrian Location:

The boxplots reveal that pedestrians who are "In Centre of Carriageway - Not on Refuge, Island or Central Reservation" tend to have the most severe injuries.

Pedestrians who are "Crossing on Pedestrian Crossing Facility" or "On Footway or Verge" tend to have less severe injuries.

Casualty Severity by Pedestrian Movement:

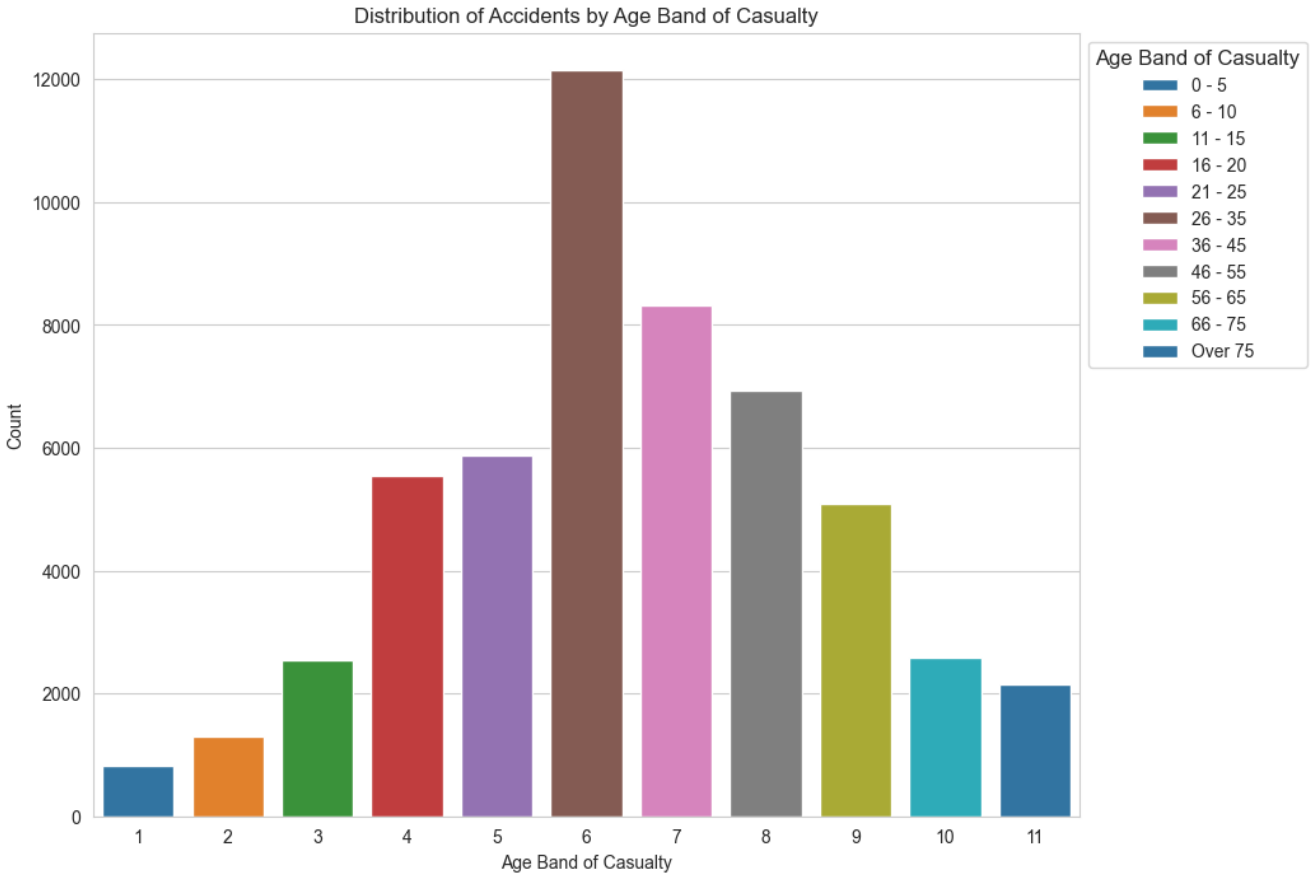
The boxplots suggest that pedestrians who are "Standing" or "Running" tend to have more severe injuries compared to those who are "Walking



- CASUALTY SEVERITY AND AGE BANDS

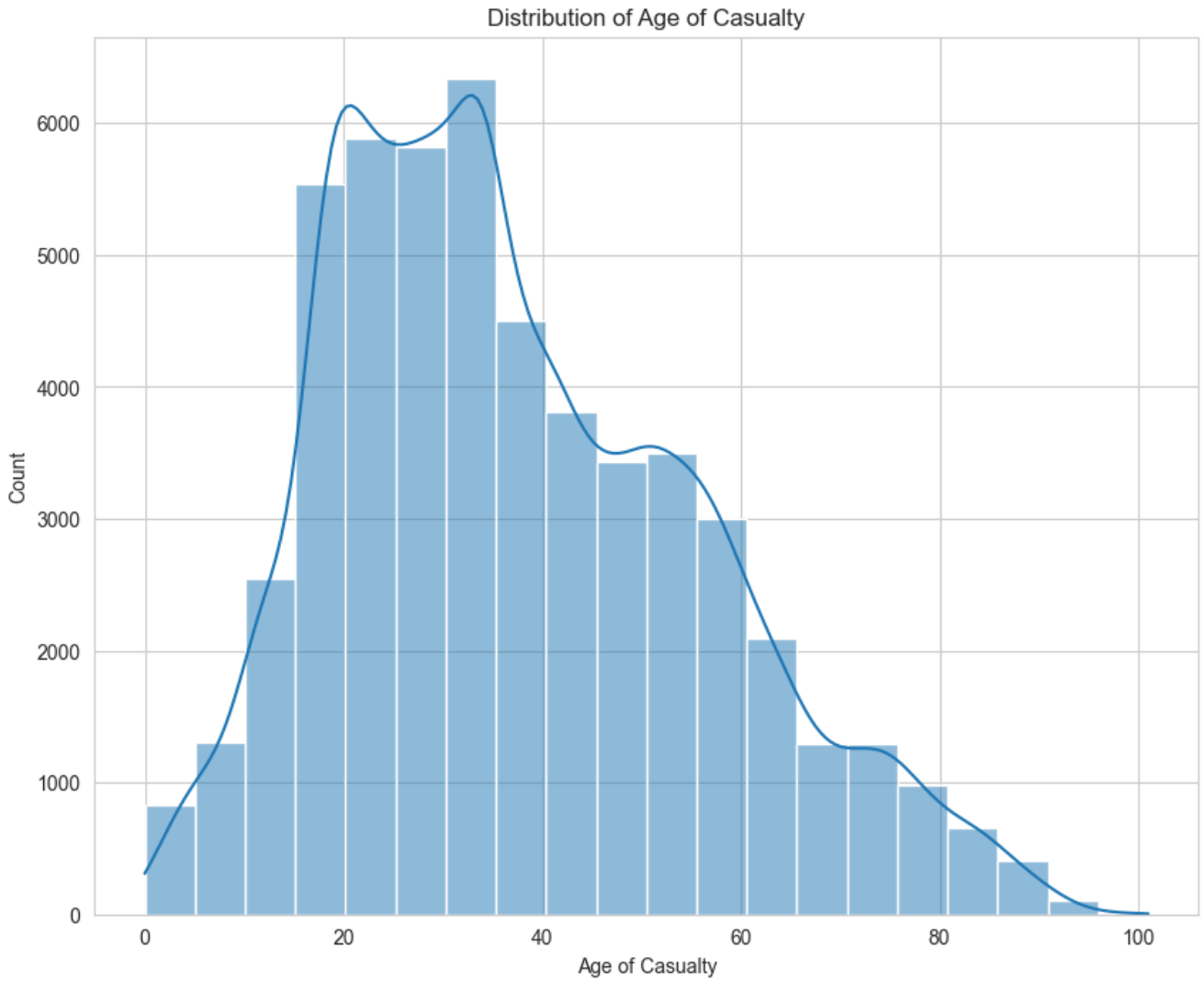
The image you sent me appears to be a bar chart that shows the distribution of accidents across different age bands. The x-axis of the chart shows the different age bands, while the y-axis shows the number of accidents in each age band. The bars are colored differently to distinguish between the different age bands.

It appears that the age band with the most accidents is 21-30 years old, followed by the 31-40 and 41-50 age bands. This suggests that people in these age groups are more likely to be involved in accidents than people in other age groups.



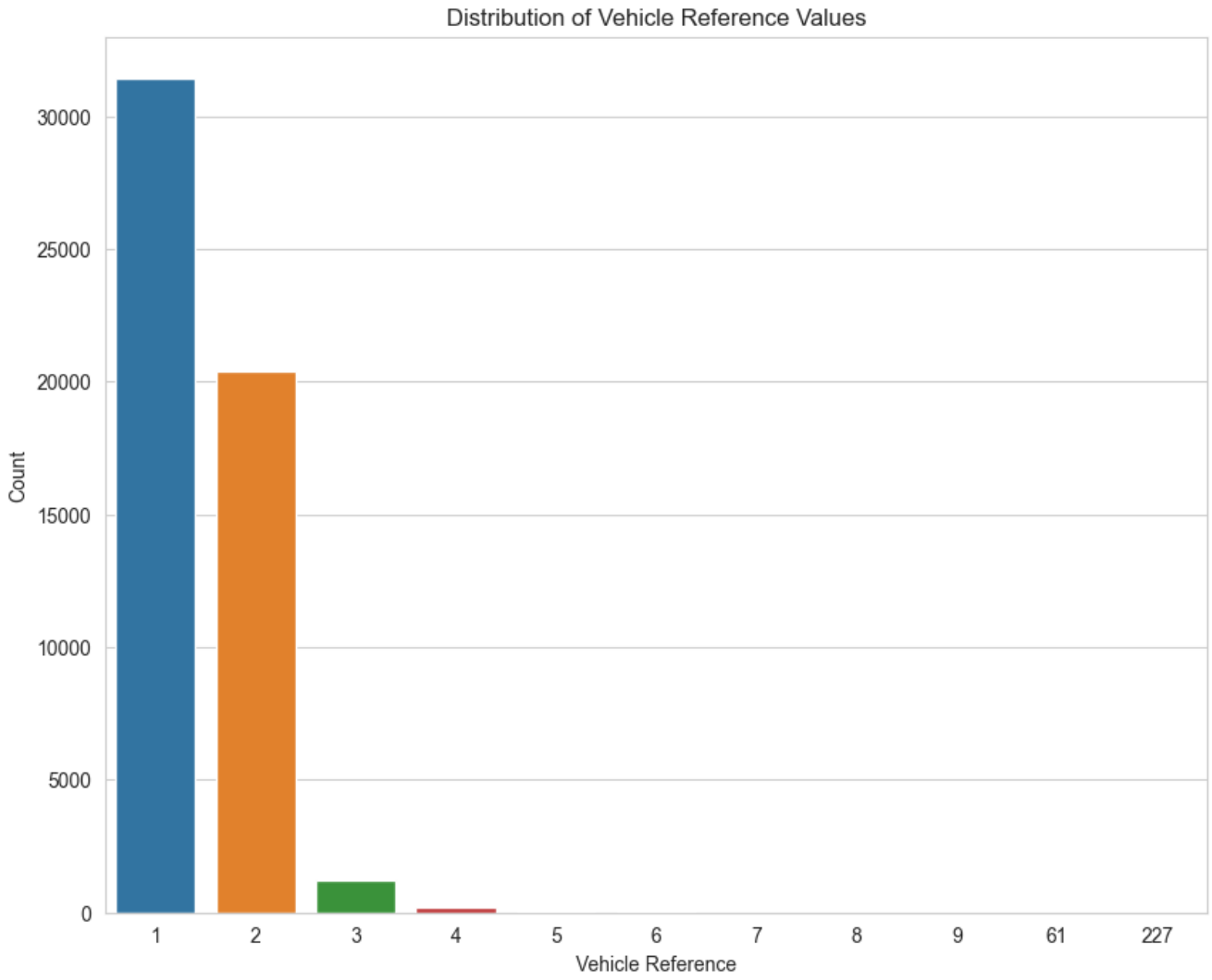
The analysis of casualty IMD deciles reveals a troubling pattern: individuals from more deprived areas (10-40% deciles) appear to be disproportionately impacted by road accidents compared to their less deprived and least deprived counterparts (40-100% deciles). This suggests a potential link between socioeconomic disadvantage and road safety.

Several factors might contribute to this disparity. Residents in deprived areas may face increased exposure to risk factors like poorer road infrastructure, limited access to safe transportation, and financial constraints hindering the use of safety measures. Additionally, they might be more susceptible to engaging in risky behaviors due to factors like stress or lack of access to safe vehicles.



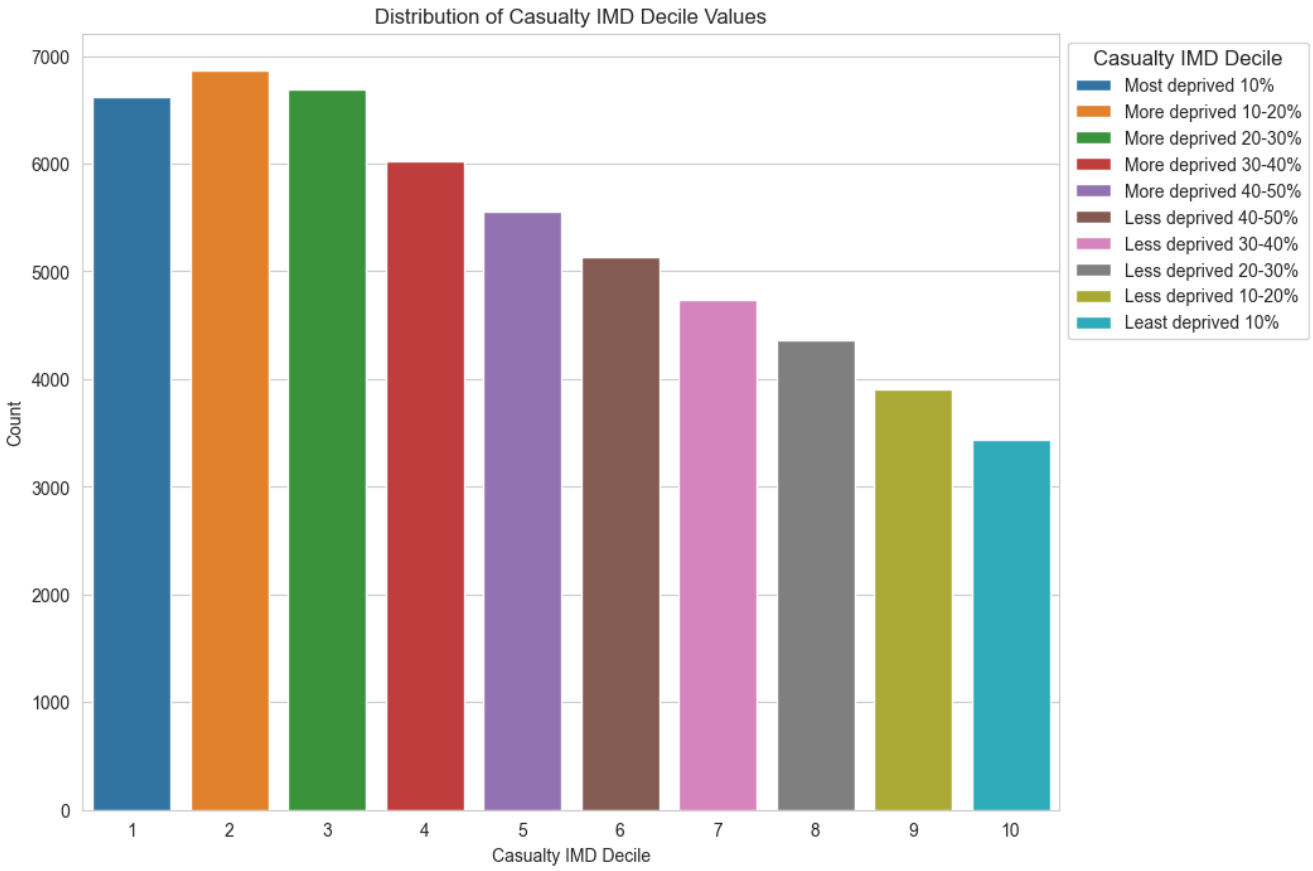
- CASUALTY SEVERITY AND VEHICLE REFERENCE

The image you sent me appears to be a bar chart showing the distribution of vehicle reference values in a road accident dataset. The x-axis of the chart shows the different vehicle reference values, while the y-axis shows the number of accidents for each value. The chart does show that there is a wide range of vehicle reference values represented in the data. The most common value appears to be "1", followed by "2" and "3".



- CASUALTY SEVERITY AND CASUALTY_IMD_DECILE

The x-axis of the chart shows the different casualty IMD deciles, which appear to represent levels of deprivation. The y-axis shows the number of casualties in each decile. The chart suggests that there are more casualties in the more deprived deciles (10-40%) compared to the less deprived and least deprived deciles (40-100%). This could indicate a potential link between socioeconomic disadvantage and road accident risk.



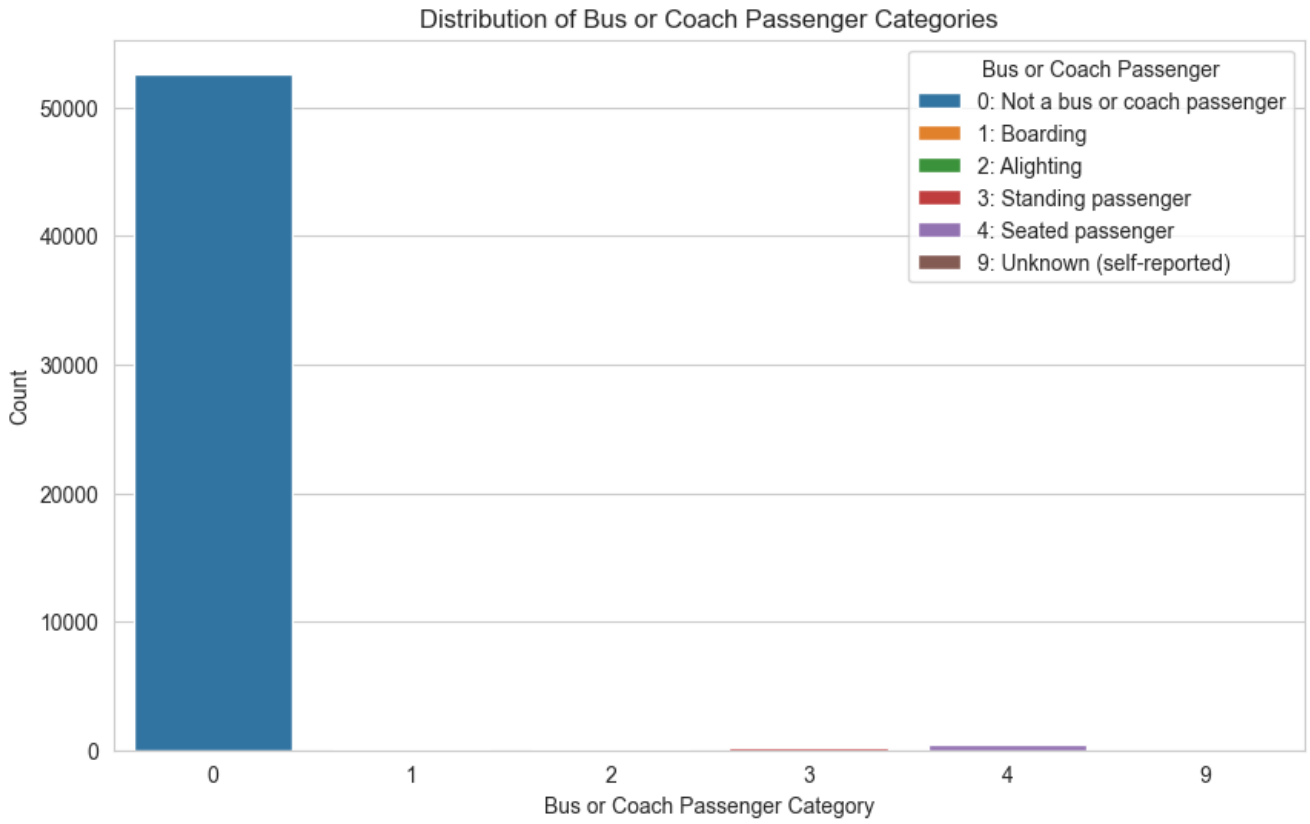
- CASUALTY SEVERITY AND BUS OR COACH PASSENGERS

The x-axis of the chart shows the different passenger categories, while the y-axis shows the number of passengers in each category. The bars are colored differently to distinguish between the different categories.

Here are some key observations from the chart:

Most passengers are not bus or coach passengers: The majority of the data points (52628) fall into the category "Not a bus or coach passenger". This suggests that the data may also include information about other types of road users, such as pedestrians or car drivers.

Seated passengers are the most common type of bus or coach passenger: Among those who are bus or coach passengers, "Seated passenger" is the most common category (396), followed by "Standing passenger" (221), "Boarding" (27), "Alighting" (43), and "Unknown (self-reported)" (6).



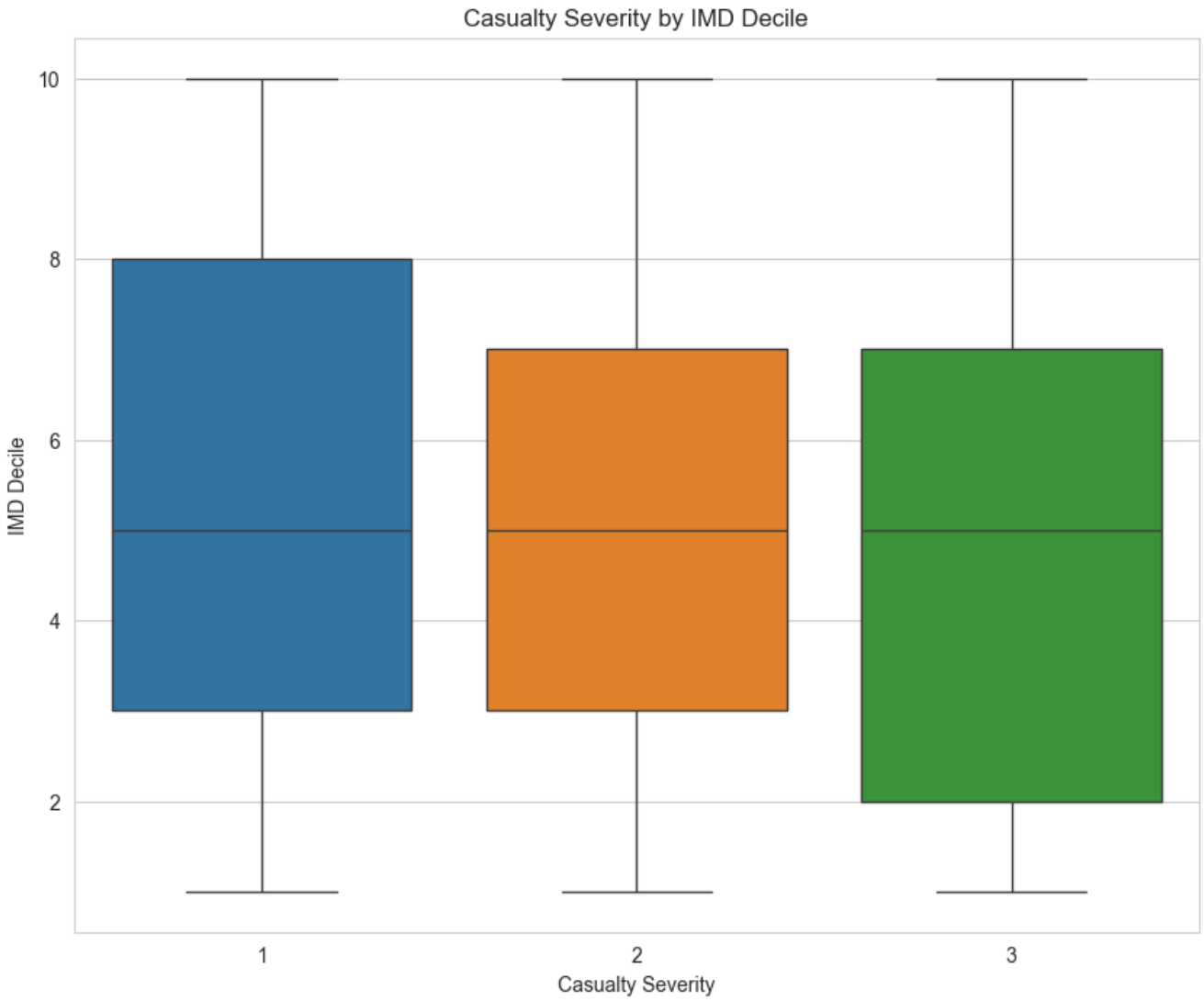
- RELATIONSHIP BETWEEN CASUALTY SEVERITY AND THE IMD DECILE

Here's an interpretation of the relationship between casualty severity and IMD decile:

Casualty Severity: The x-axis of the chart shows the different categories of casualty severity, which include "Fatal", "Serious", and "Slight".

IMD Decile: The y-axis of the chart shows the casualty severity, represented by the different IMD deciles, where 1 represents the most deprived and 10 represents the least deprived areas.

The boxplot suggests that there might be a trend of increasing casualty severity with decreasing IMD decile (increasing deprivation). This means that casualties in more deprived areas (lower deciles) tend to experience more severe injuries compared to those in less deprived areas (higher deciles).



- RELATIONSHIP BETWEEN CASUALTY SEVERITY AND AGE BAND

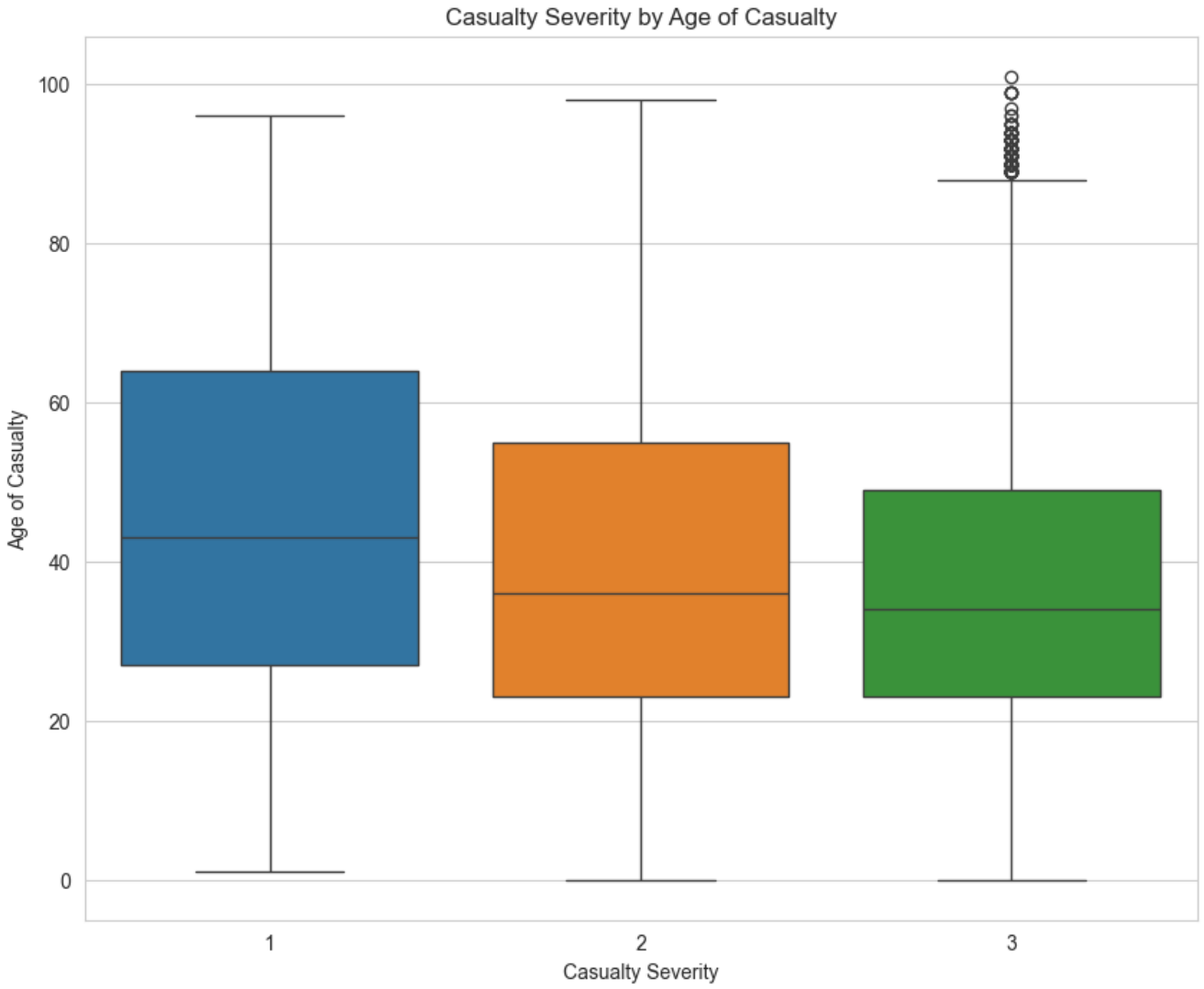
Here's an interpretation of the relationship between casualty severity and age of casualty:

Casualty Severity: The x-axis of the chart shows the different categories of casualty severity, which include "Fatal", "Serious", and "Slight".

Age of Casualty: The y-axis of the chart shows the casualty severity, represented by the different age groups.

The boxplot suggests that there might be a U-shaped relationship between casualty severity and age of casualty. This means that:

Younger age groups (0-20 years old) and **older age** groups (61-80 years old and above) tend to have a higher proportion of casualties with "Serious" and "Fatal" injuries. **Middle-aged adults** (41-60 years old) tend to have a lower proportion of casualties with severe injuries.



Model development

Models Used: Logistic Regression, Decision Tree, Random Forest.
Data Preprocessing: Standard scaling applied to features before training the models.

Training and Testing: Data split into training and testing sets with a 80:20 ratio.

Model Training: Each model trained on the training data.

Model Evaluation: Performance metrics computed for each model - Accuracy, Precision, Recall, and F1-score.

- **Interpretation and Comparison of Results:**

- Logistic Regression: Achieved an accuracy of 80.32%. It performed slightly better than the other models in terms of precision, recall, and F1-score, indicating good overall performance.
- Decision Tree: Lower accuracy compared to logistic regression, at 74.36%. While it has comparable precision, recall, and F1-score, it's slightly less accurate.
- Random Forest: Accuracy lies between logistic regression and decision tree at 77.16%. It shows similar precision, recall, and F1-score to decision tree but outperforms it slightly in terms of accuracy.

Conclusion:

- Logistic Regression stands out as the best-performing model in terms of accuracy and overall performance. It provides a good balance between precision, recall, and F1-score.

- Decision Tree and Random Forest perform relatively well but are slightly less accurate compared to logistic regression. They might be considered for ensemble methods or when interpretability is important.

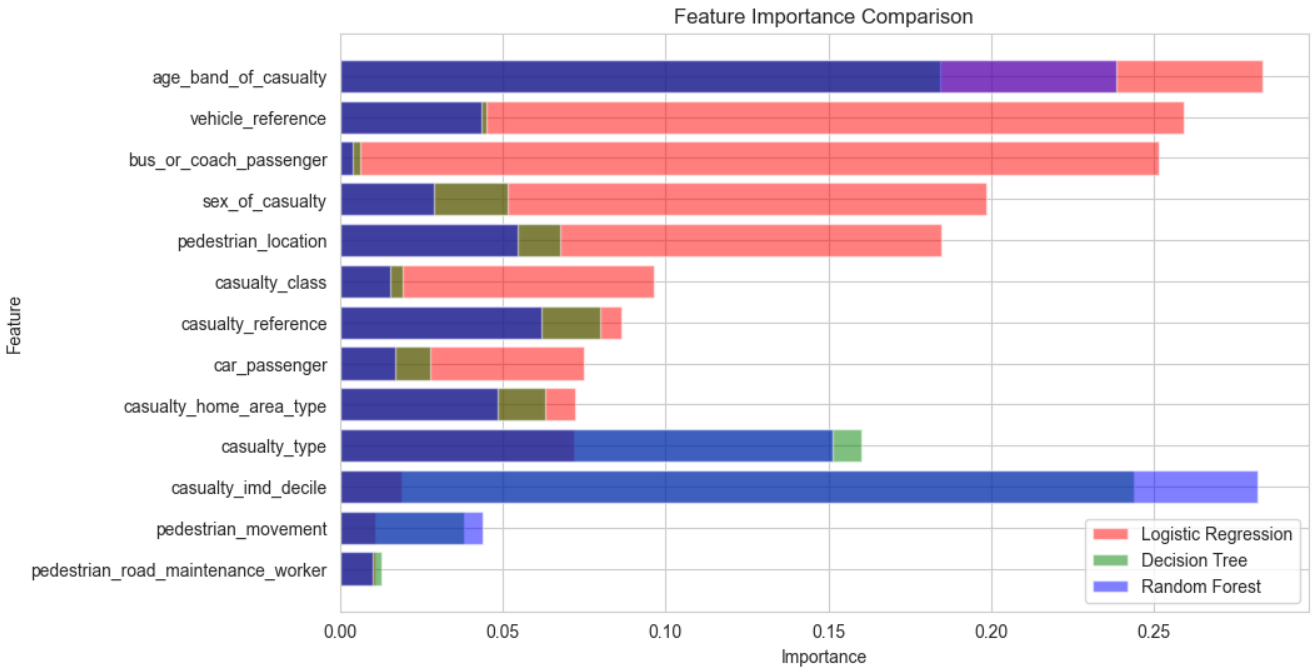
- Overall, while logistic regression performs the best in this scenario, further experimentation and fine-tuning could potentially improve the performance of all models.

Results Table:

	Accuracy	Precision	Recall	F1
Logistic Regression	0.803188	0.738077	0.803188	0.715701
Decision Tree	0.743647	0.713982	0.743647	0.726949
Random Forest	0.771589	0.710718	0.771589	0.731864

- **Feature importance**

The image is a bar chart that compares the feature importance between three machine learning models: Logistic Regression, Decision Tree, and Random Forest.
The x-axis represents the feature importance (how important a feature is to the model's prediction), and the y-axis represents the features (names of the columns in your data). Each colored bar represents a model (red: Logistic Regression, green: Decision Tree, blue: Random Forest), and the length of the bar shows the importance of that feature according to that model.
For example, the feature "pedestrian_location" seems to be the most important feature for the Logistic Regression model, while it seems less important for the Random Forest model based on the length of the bars.
“age_band_casualty ”, “casualty_imd_decile”, “casualty_type ” are some of the most important features.



Suggestions:

To address road accidents effectively, we propose the following solutions:

1. Weather Conditions: Analyzing how weather conditions such as rain, fog, or snow impact accident severity provides insights into weather-related risk factors.

2. Daylight Visibility: Considering the influence of daylight conditions on accidents helps understand visibility-related challenges for drivers and pedestrians.

3. Road Maintenance Quality: Evaluating road maintenance factors like potholes, signage, and road surface conditions contributes to identifying infrastructure-related risks.

4. Separate Datasets: Creating separate datasets for pedestrians and non-pedestrians enables focused analysis, allowing for targeted interventions and safety measures tailored to specific user groups.