Road Accidents Analysis

By:

Mohammad Reza Ghotbizadeh Vahdani

1. Introduction

This comprehensive dataset offers detailed information exclusively on road accidents reported during the year 2022. It encompasses various attributes, including accident status, vehicle and casualty details, demographics, and casualty severity. Key elements such as pedestrian information, types of casualties, involvement of road maintenance workers, and the Index of Multiple Deprivation (IMD) decile for casualties' home areas are included. This dataset provides valuable insights for analyzing road accidents, identifying trends, and implementing targeted safety measures to reduce casualties and enhance road safety specifically for the year 2022. Researchers, policymakers, and analysts can leverage this dataset for evidence-based decision-making to improve overall road transportation systems within the context of the year 2022.

Source 1: https://www.kaggle.com/datasets/juhibhojani/road-accidents-data-2022/data

Source 2: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Source 3: https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-road-safety-open-dataset-data-guide-2023.xlsx

2. Columns Definition

The following provides the definition for each column as outlined in the data guide. After cleaning the data, including checking columns, handling missing values and outliers, histograms are plotted for each specific column to show the distribution of data among categories

2.1. Status: The status of the accident (e.g., reported, under investigation).

2.2. Accident_Index: A unique identifier for each reported accident.

More info: unique value for each accident. The accident_index combines the accident_year and accident_ref_no to form a unique ID. It can be used to join to Vehicle and Casualty.

2.3. accident_year: The year in which the accident occurred.

2.4. accident_reference: A reference number associated with the accident.

More info: In year id used by the police to reference a collision. It is not unique outside of the year, use accident_index for linking to other years.

2.5. vehicle_reference: A reference number for the involved vehicle in the accident (unique value for each vehicle in a singular accident). Can be used to join a Casualty to a vehicle.

Table 1. Vehicle Reference Categories

vehicle_type	
code/format	**label**
1	Pedal cycle
2	Motorcycle 50cc and under
3	Motorcycle 125cc and under
4	Motorcycle over 125cc and up to 500cc
5	Motorcycle over 500cc
8	Taxi/Private hire car
9	Car
10	Minibus (8 – 16 passenger seats)
11	Bus or coach (17 or more pass seats)
16	Ridden horse
17	Agricultural vehicle
18	Tram
19	Van / Goods 3.5 tonnes mgw or under
20	Goods over 3.5t. and under 7.5t
21	Goods 7.5 tonnes mgw and over
22	Mobility scooter
23	Electric motorcycle
90	Other vehicle
97	Motorcycle – unknown cc
98	Goods vehicle – unknown weight
99	Unknown vehicle type (self rep only)
103	Motorcycle – Scooter (1979-1998)
104	Motorcycle (1979-1998)
105	Motorcycle – Combination (1979-1998)
106	Motorcycle over 125cc (1999-2004)
108	Taxi (excluding private hire cars) (1979-2004)
109	Car (including private hire cars) (1979-2004)
110	Minibus/Motor caravan (1979-1998)
113	Goods over 3.5 tonnes (1979-1998)
-1	Data missing or out of range

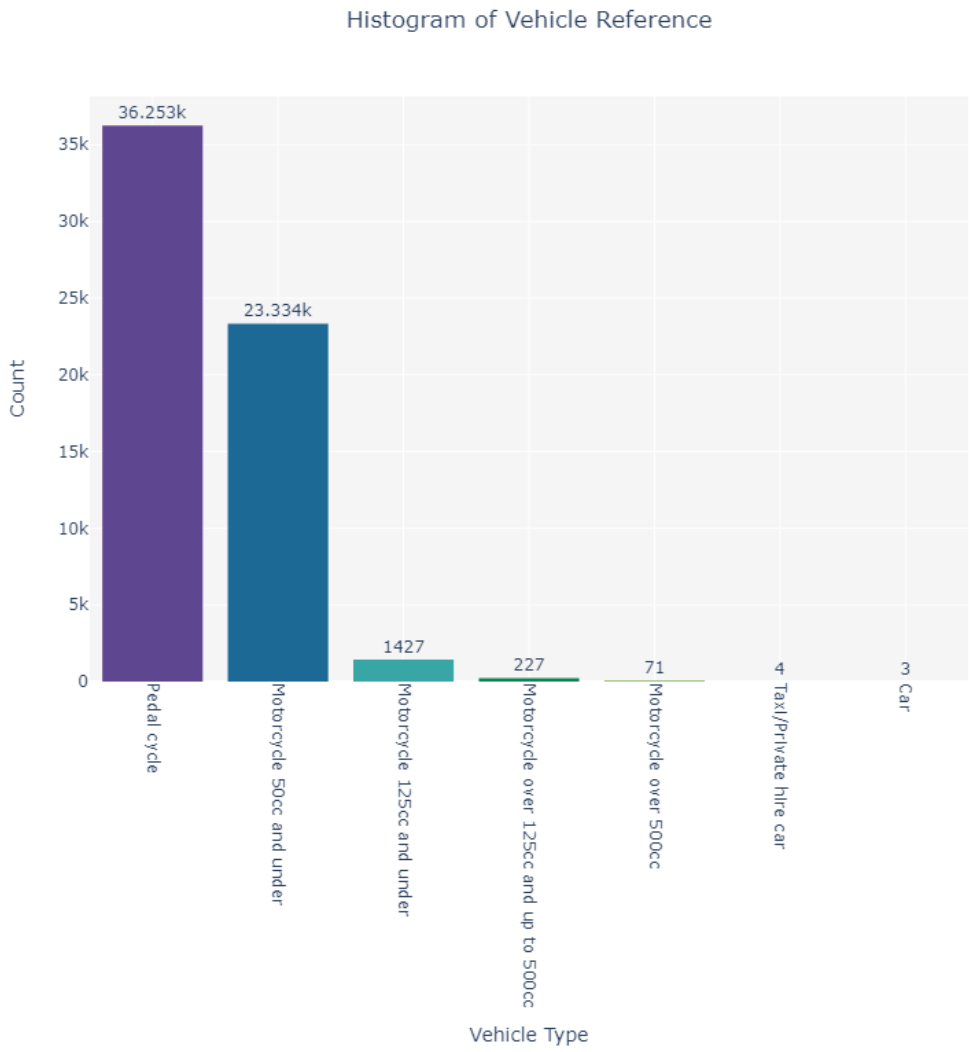


Figure 1. Vehicle Reference Histogram

2.6. casualty_reference: A reference number for the casualty involved in the accident.

More info: unique value for each casualty in a singular accident (historical years may be unique to a singular vehicle in a road accident)

2.7. casualty_class: Indicates the class of the casualty (e.g., driver, passenger, pedestrian).

Table 2. Casualty Class Categories

code/format	label
1	Driver or rider
2	Passenger
3	Pedestrian

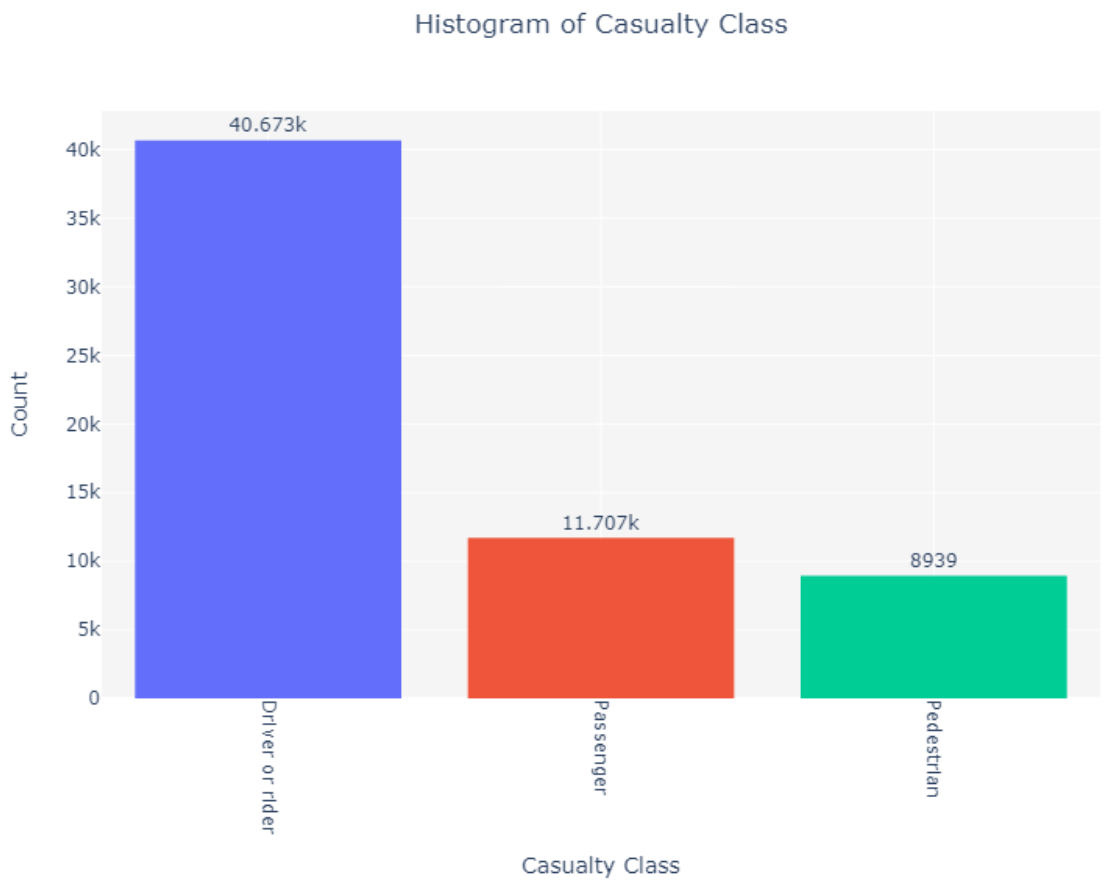


Figure 2. Casualty Class Histogram

2.8. sex_of_casualty:

Table 3. Casualty Gender

code/format	label
1	Male
2	Female
9	unknown (self reported)
-1	Data missing or out of range

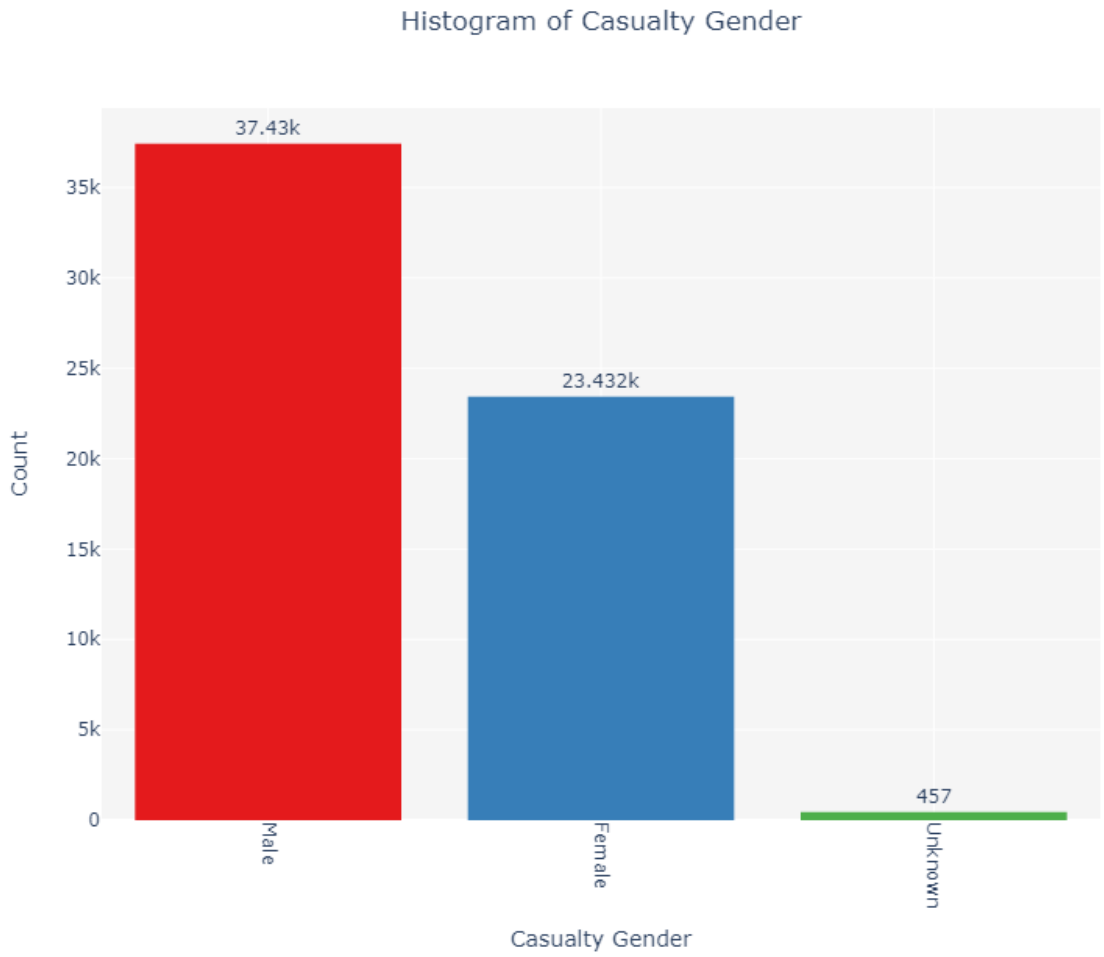


Figure 3. Casualty Gender Histogram

2.9. age_of_casualty: The age of the casualty (-1: Data missing or out of range).

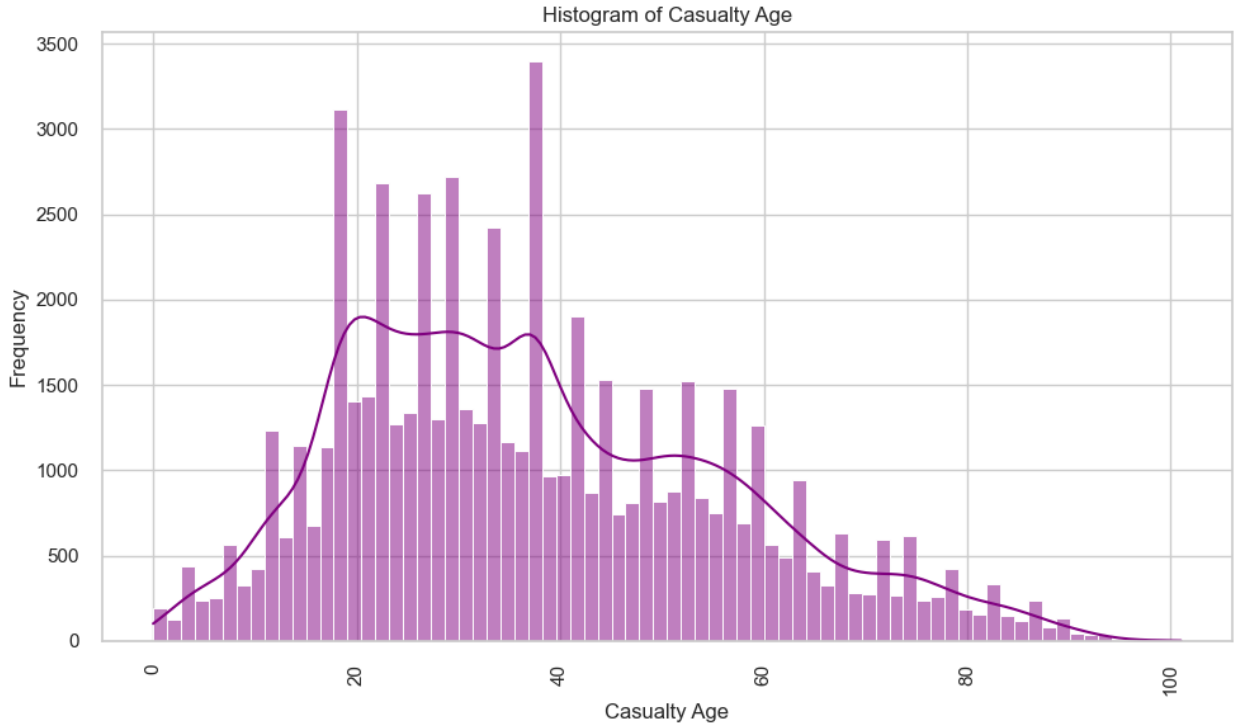


Figure 4. Casualty Age Histogram

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

Table 4. Age Band of Casualty Categories

code/format	label
1	0 - 5
2	6 - 10
3	11 - 15
4	16 - 20
5	21 - 25
6	26 - 35
7	36 - 45
8	46 - 55
9	56 - 65
10	66 - 75
11	Over 75
-1	Data missing or out of range

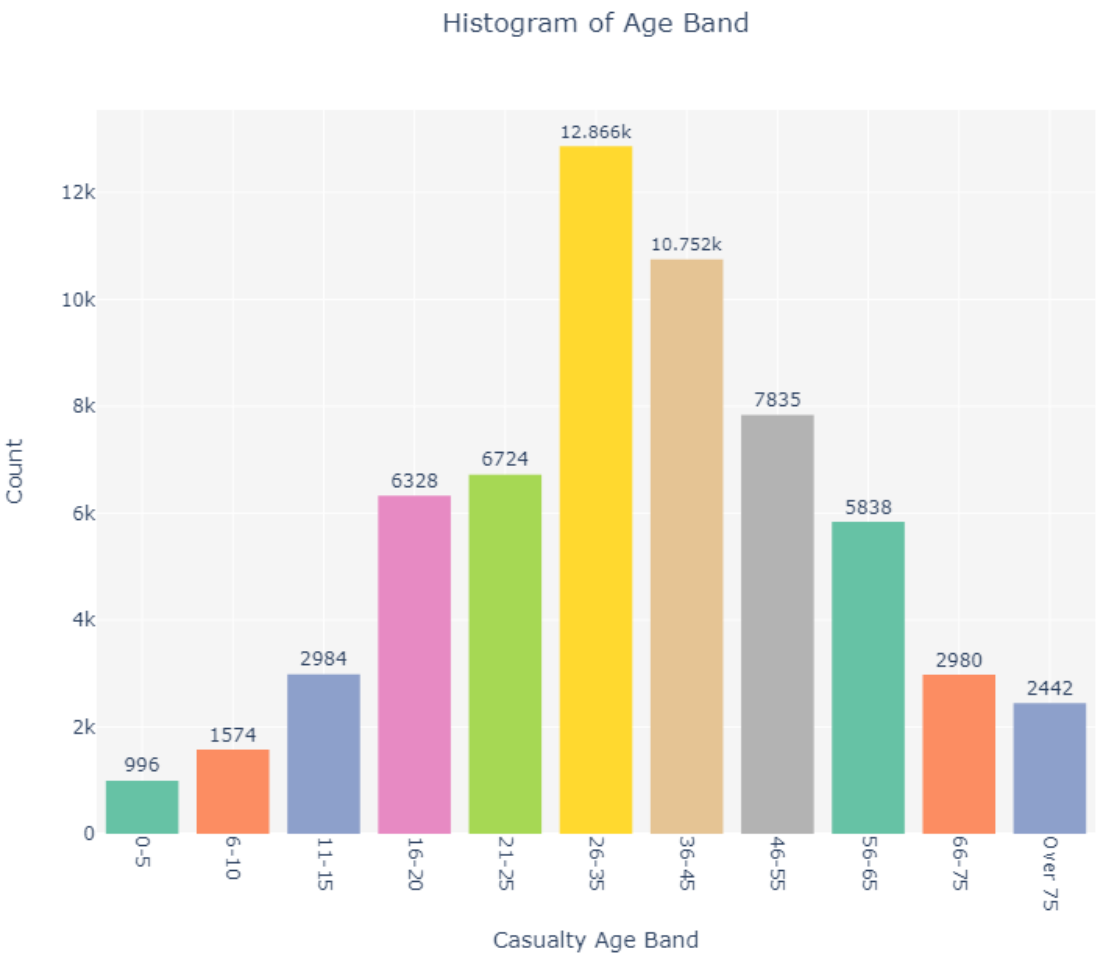


Figure 5. Age Band Histogram

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

Table 5. Casualty Severity

code/format	label
1	Fatal
2	Serious
3	Slight

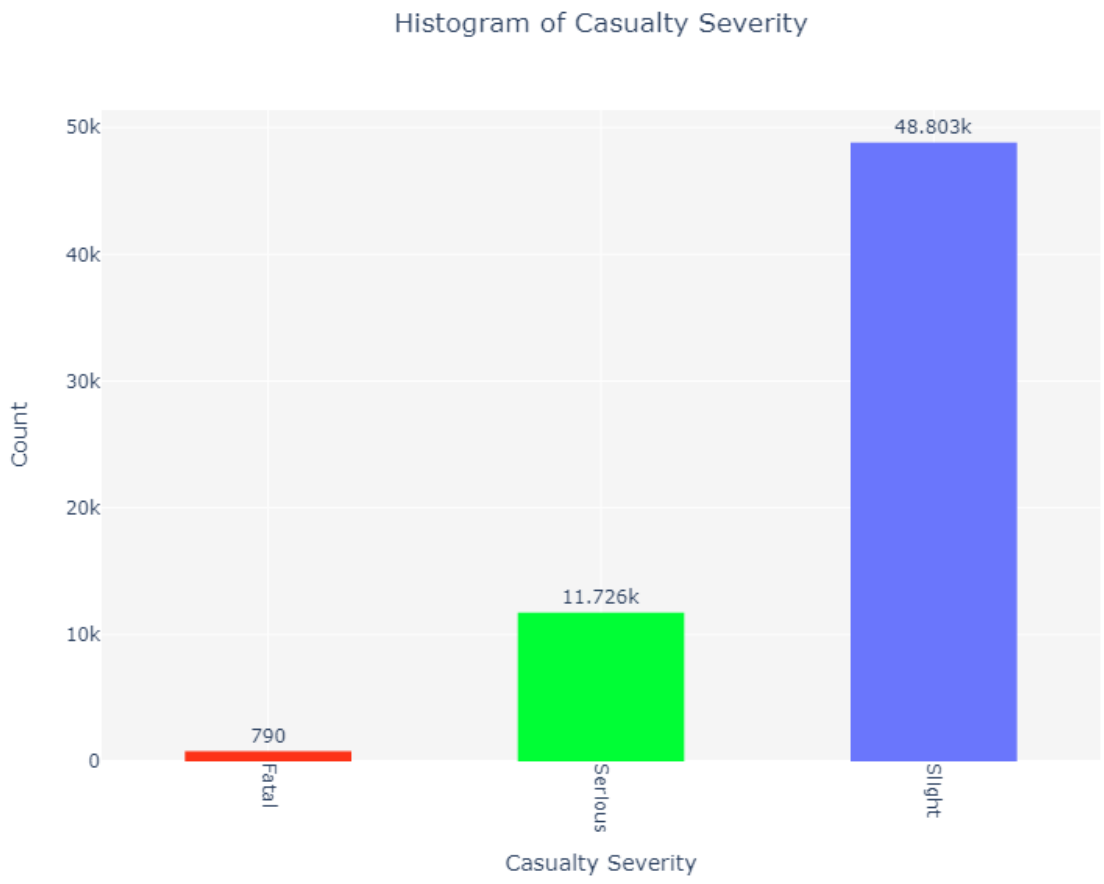


Figure 6. Casualty Severity Histogram

2.12. pedestrian_location: The location of the pedestrian at the time of the accident.

Table 6. Pedestrian Location Categories

code/format	label
0	Not a Pedestrian
1	Crossing on pedestrian crossing facility
2	Crossing in zig-zag approach lines
3	Crossing in zig-zag exit lines
4	Crossing elsewhere within 50m. of pedestrian crossing
5	In carriageway, crossing elsewhere
6	On footway or verge
7	On refuge, central island or central reservation
8	In center of carriageway - not on refuge, island or central reservation
9	In carriageway, not crossing
10	Unknown or other
-1	Data missing or out of range

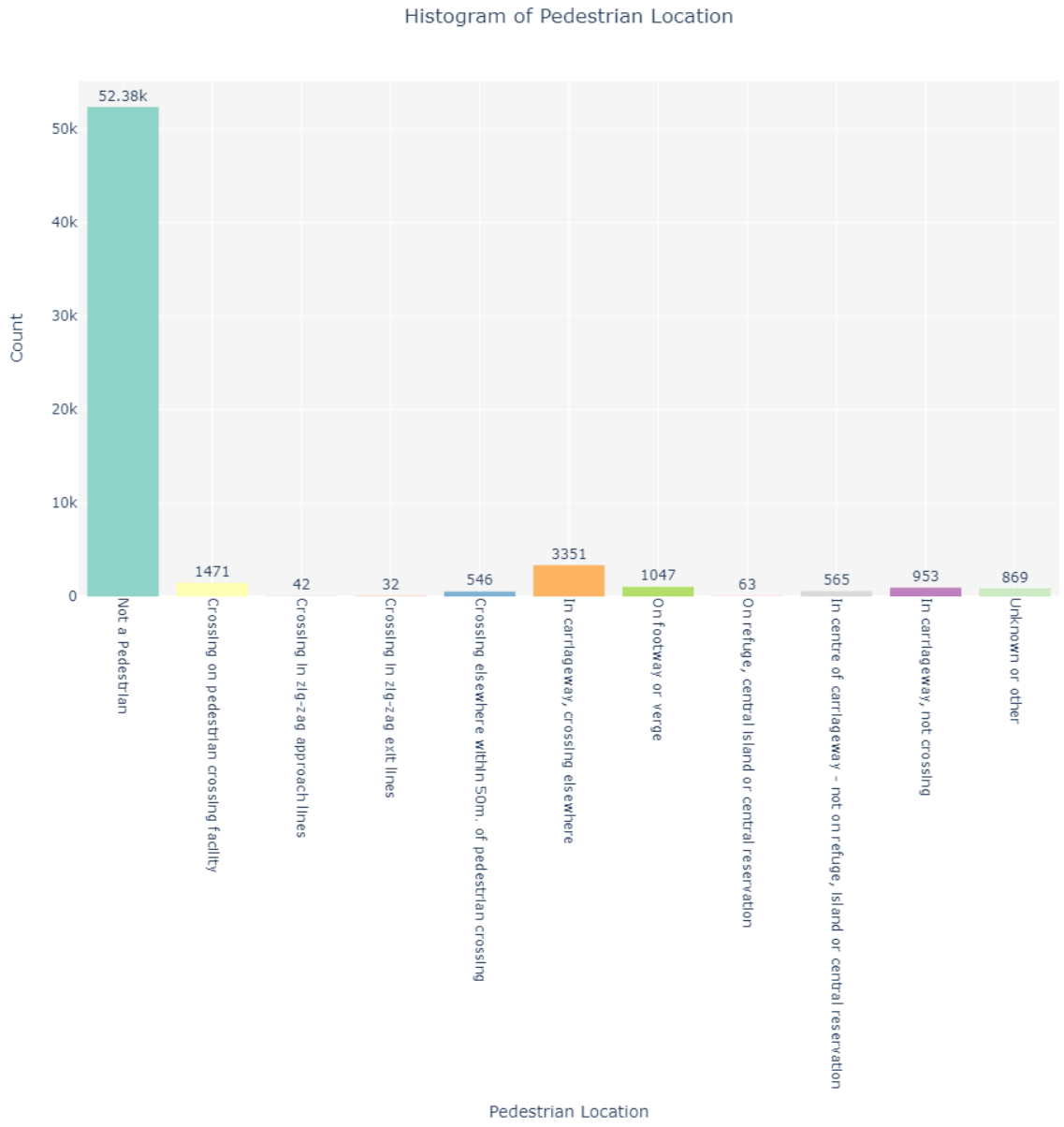


Figure 7. Pedestrian Location Histogram

2.13. pedestrian_movement: The movement of the pedestrian during the accident.

Table 7. Pedestrian Movement Categories

code/format	label
0	Not a Pedestrian
1	Crossing from driver's nearside
2	Crossing from nearside - masked by parked or stationary vehicle
3	Crossing from driver's offside
4	Crossing from offside - masked by parked or stationary vehicle
5	In carriageway, stationary - not crossing (standing or playing)
6	In carriageway, stationary - not crossing (standing or playing) - masked by parked or stationary vehicle
7	Walking along in carriageway, facing traffic
8	Walking along in carriageway, back to traffic
9	Unknown or other
-1	Data missing or out of range

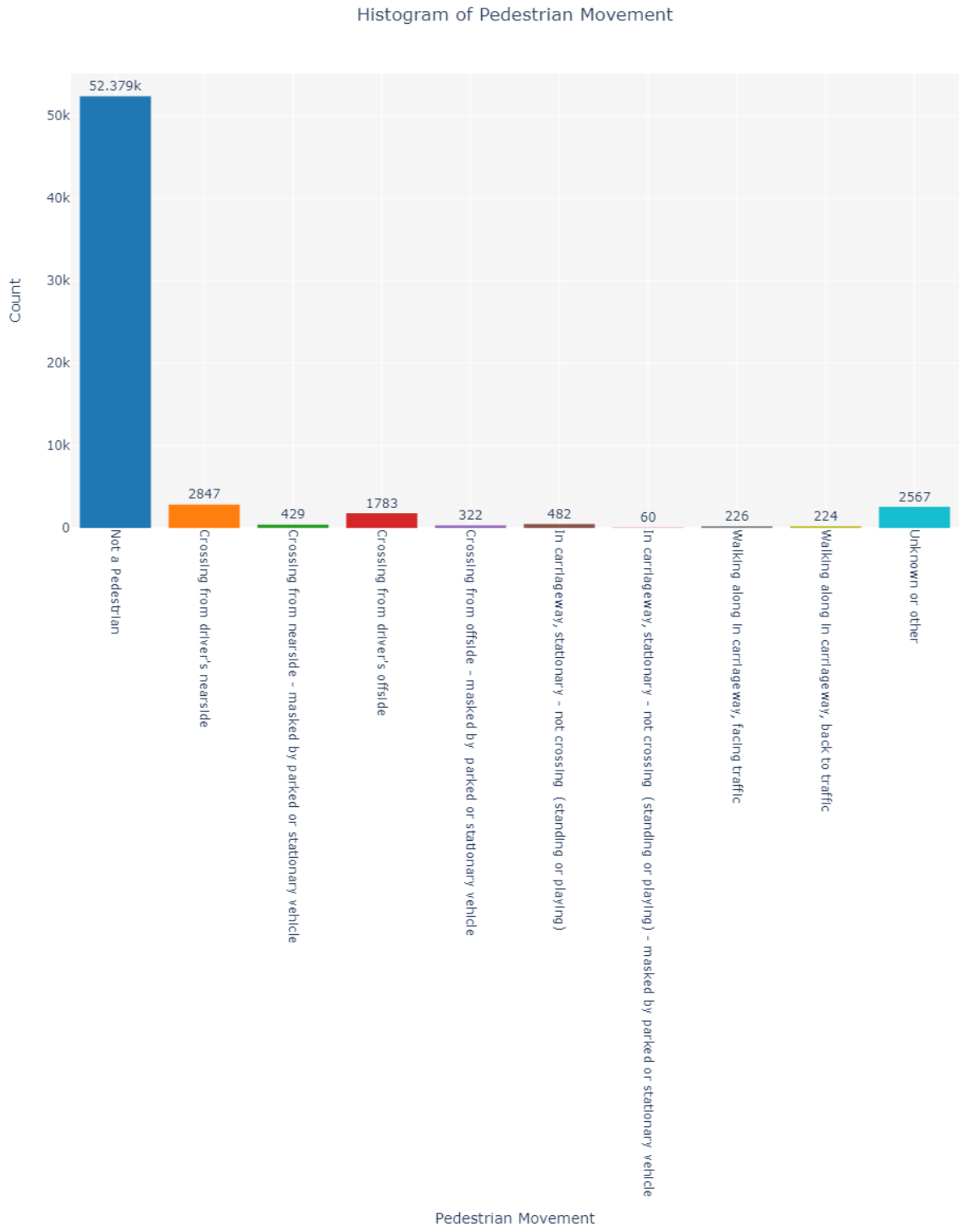


Figure 8. Pedestrian Movement Histogram

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

Table 8. Car Passenger Categories

code/format	label
0	Not car passenger
1	Front seat passenger
2	Rear seat passenger
9	unknown (self-reported)
-1	Data missing or out of range

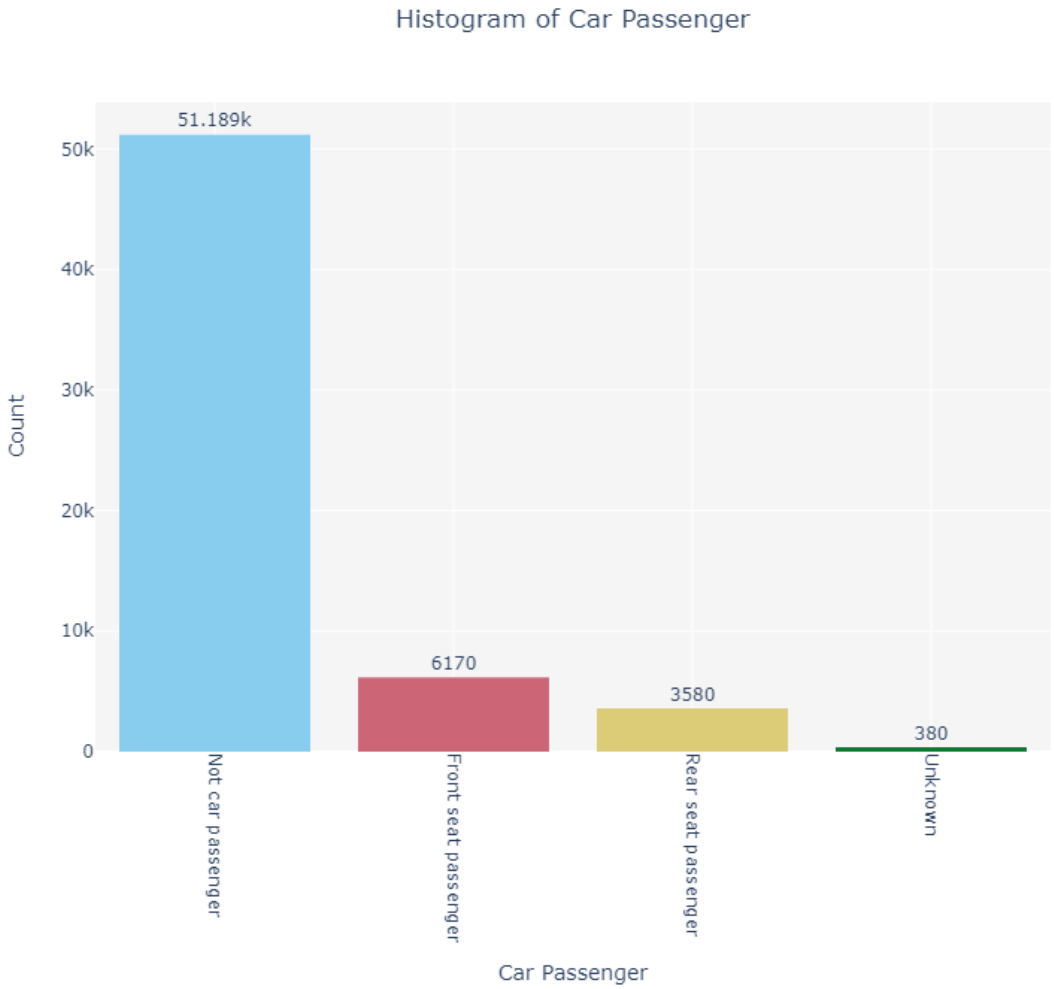


Figure 9. Car Passenger Histogram

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

Table 9. Bus or Coach Passenger Categories

code/format	label
0	Not a bus or coach passenger
1	Boarding
2	Alighting
3	Standing passenger
4	Seated passenger
9	unknown (self-reported)
-1	Data missing or out of range

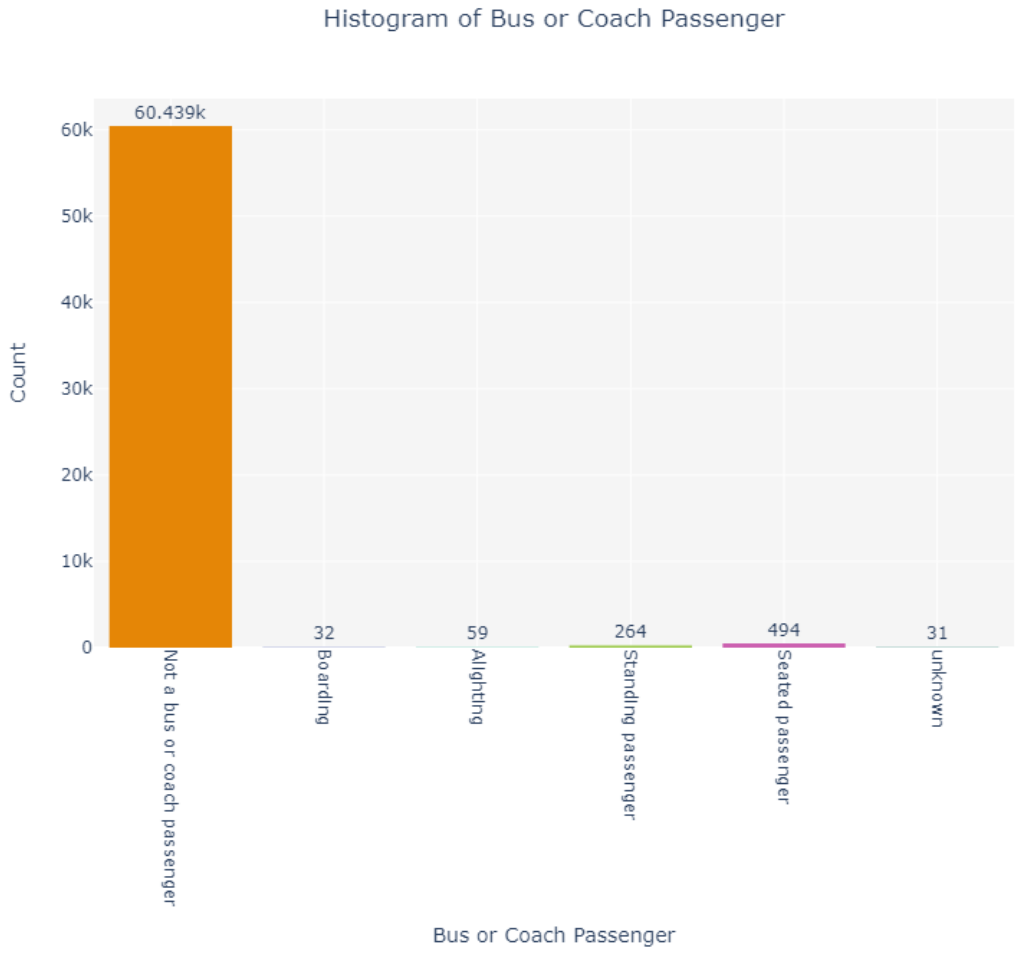


Figure 10. Bus or Coach Passenger Histogram

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

Table 10. Pedestrian Road Maintenance Worker categories

code/format	label	note
0	No / Not applicable	
1	Yes	
2	Not Known	
3	Probable	2005 specification only
-1	Data missing or out of range	

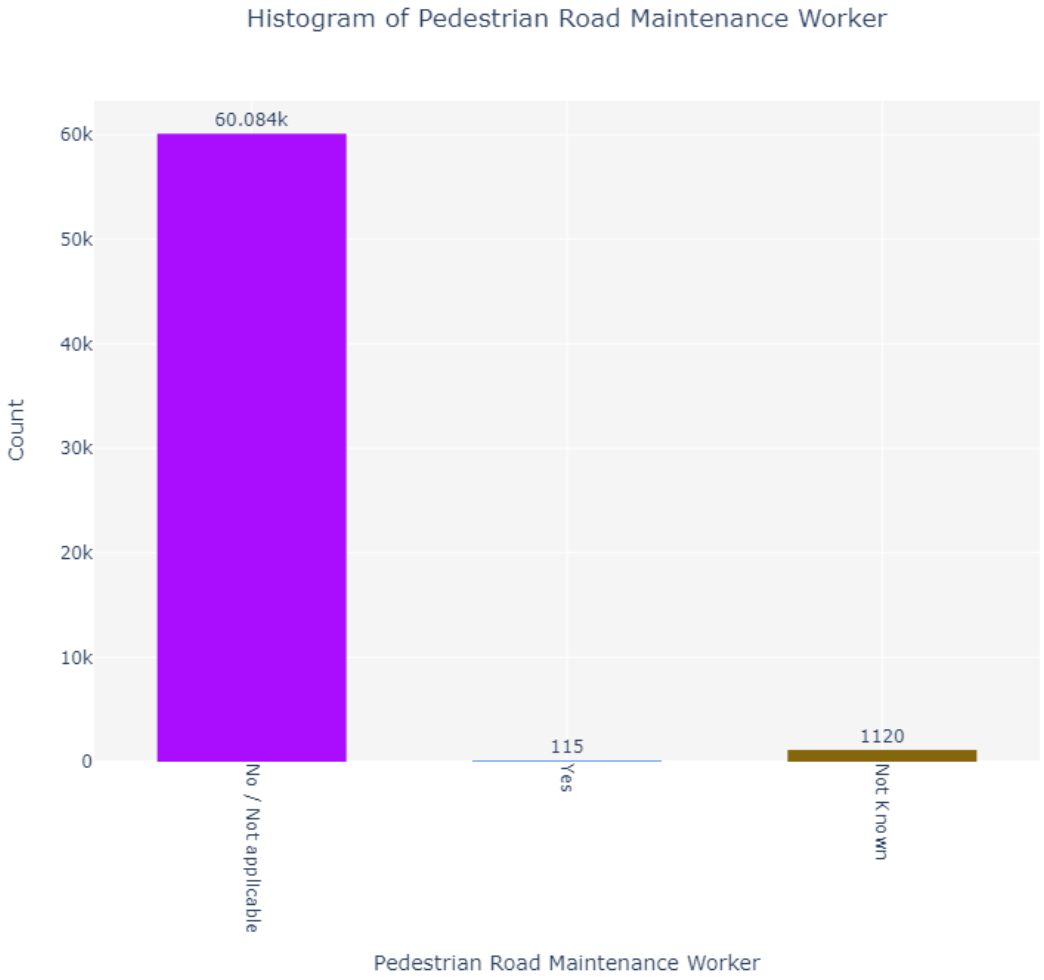


Figure 11. Pedestrian Road Maintenance Worker Histogram

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

Table 11. Casualty Type Categories

code/format	label	note
0	Pedestrian	
1	Cyclist	
2	Motorcycle 50cc and under rider or passenger	
3	Motorcycle 125cc and under rider or passenger	introduced in 1999 specification
4	Motorcycle over 125cc and up to 500cc rider or passenger	introduced in 2005 specification
5	Motorcycle over 500cc rider or passenger	introduced in 2005 specification
8	Taxi/Private hire car occupant	introduced in 2005 specification
9	Car occupant	introduced in 2005 specification
10	Minibus (8 - 16 passenger seats) occupant	introduced in 1999 specification
11	Bus or coach occupant (17 or more pass seats)	
16	Horse rider	introduced in 1999 specification
17	Agricultural vehicle occupant	introduced in 1999 specification
18	Tram occupant	introduced in 1999 specification
19	Van / Goods vehicle (3.5 tonnes mgw or under) occupant	
20	Goods vehicle (over 3.5t. and under 7.5t.) occupant	introduced in 1999 specification
21	Goods vehicle (7.5 tonnes mgw and over) occupant	introduced in 1999 specification
22	Mobility scooter rider	introduced in 2011 specification
23	Electric motorcycle rider or passenger	introduced in 2011 specification
90	Other vehicle occupant	introduced in 2011 specification
97	Motorcycle - unknown cc rider or passenger	introduced in 2011 specification

98	Goods vehicle (unknown weight) occupant	introduced in 2011 specification
99	Unknown vehicle type (self rep only)	introduced in 2011 specification
103	Motorcycle - Scooter (1979-1998)	dropped in 1999 specification
104	Motorcycle (1979-1998)	dropped in 1999 specification
105	Motorcycle - Combination (1979-1998)	dropped in 1999 specification
106	Motorcycle over 125cc (1999-2004)	dropped in 2005 specification
108	Taxi (excluding private hire cars) (1979-2004)	dropped in 2005 specification
109	Car (including private hire cars) (1979-2004)	dropped in 2005 specification
110	Minibus/Motor caravan (1979-1998)	dropped in 1999 specification
113	Goods over 3.5 tonnes (1979-1998)	dropped in 1999 specification
-1	Data missing or out of range	

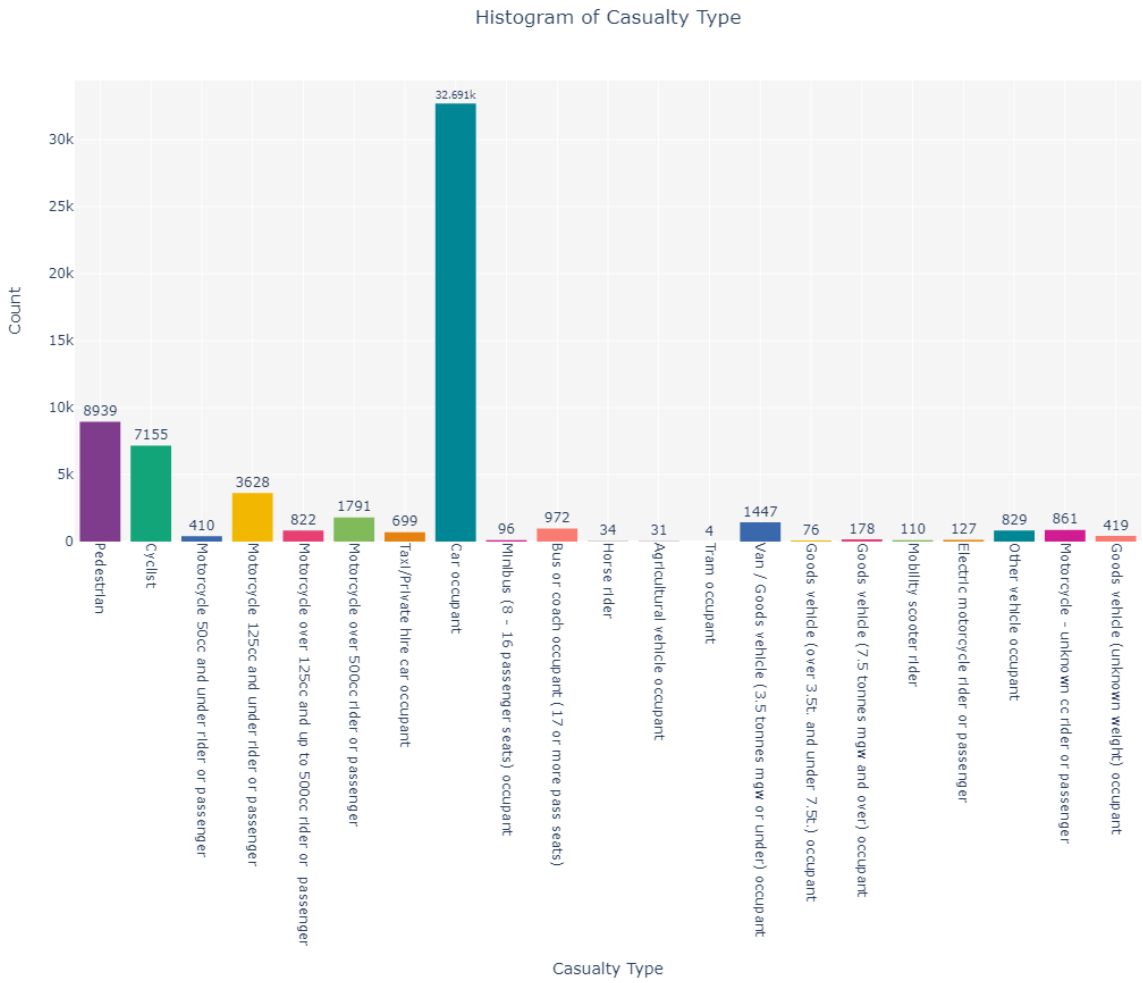


Figure 12. Casualty Type Histogram

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

Table 12. Casualty Home Area Type Categories

code/format	label	note
1	Urban area	field introduced in 1999
2	Small town	field introduced in 1999
3	Rural	field introduced in 1999
-1	Data missing or out of range	field introduced in 1999

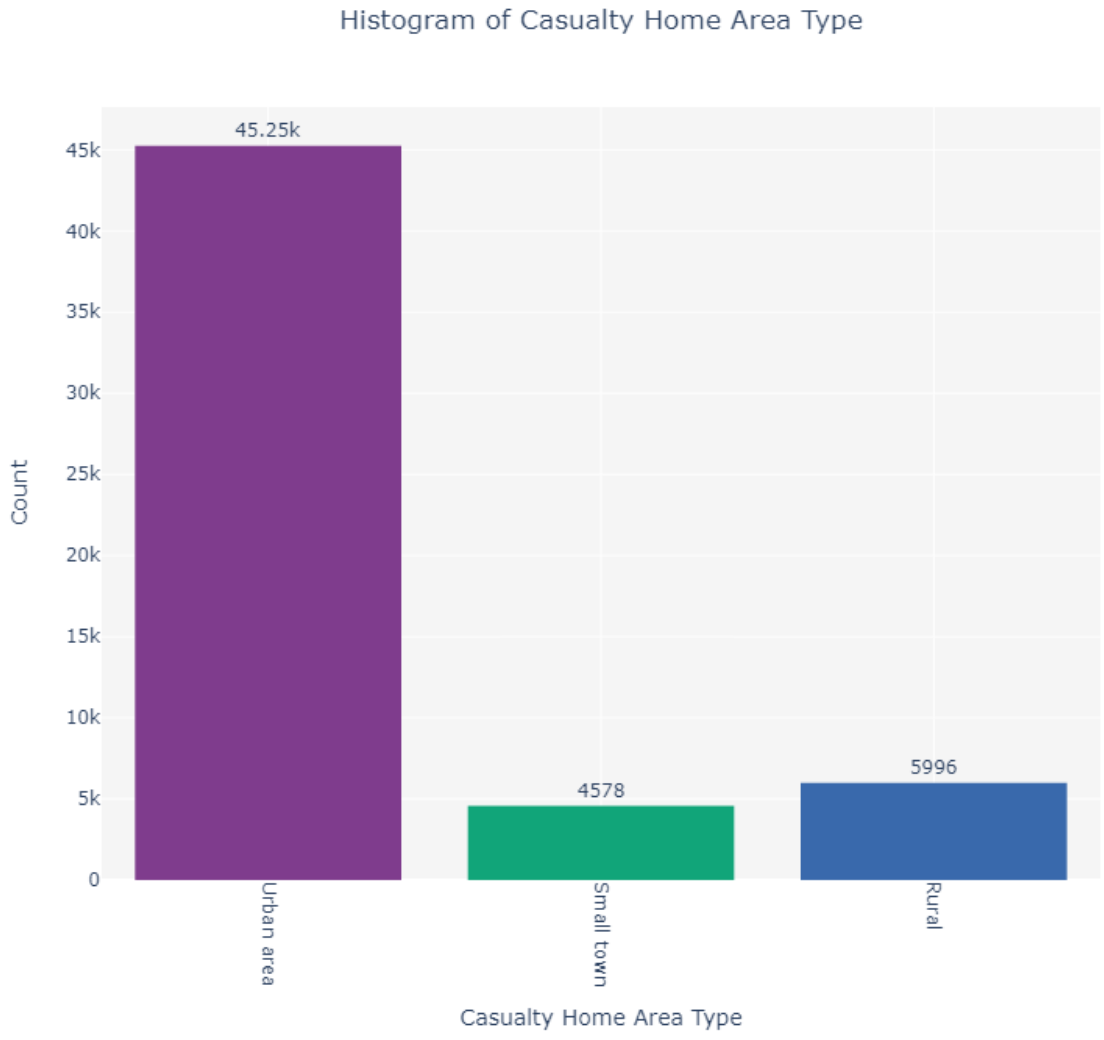


Figure 13. Casualty Home Area Type Histogram

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

Table 13. Casualty IMD decile Categories

code/format	label	note
1	Most deprived 10%	field introduced in 2016
2	More deprived 10-20%	field introduced in 2016
3	More deprived 20-30%	field introduced in 2016
4	More deprived 30-40%	field introduced in 2016
5	More deprived 40-50%	field introduced in 2016
6	Less deprived 40-50%	field introduced in 2016
7	Less deprived 30-40%	field introduced in 2016
8	Less deprived 20-30%	field introduced in 2016
9	Less deprived 10-20%	field introduced in 2016
10	Least deprived 10%	field introduced in 2016
-1	Data missing or out of range	field introduced in 2016

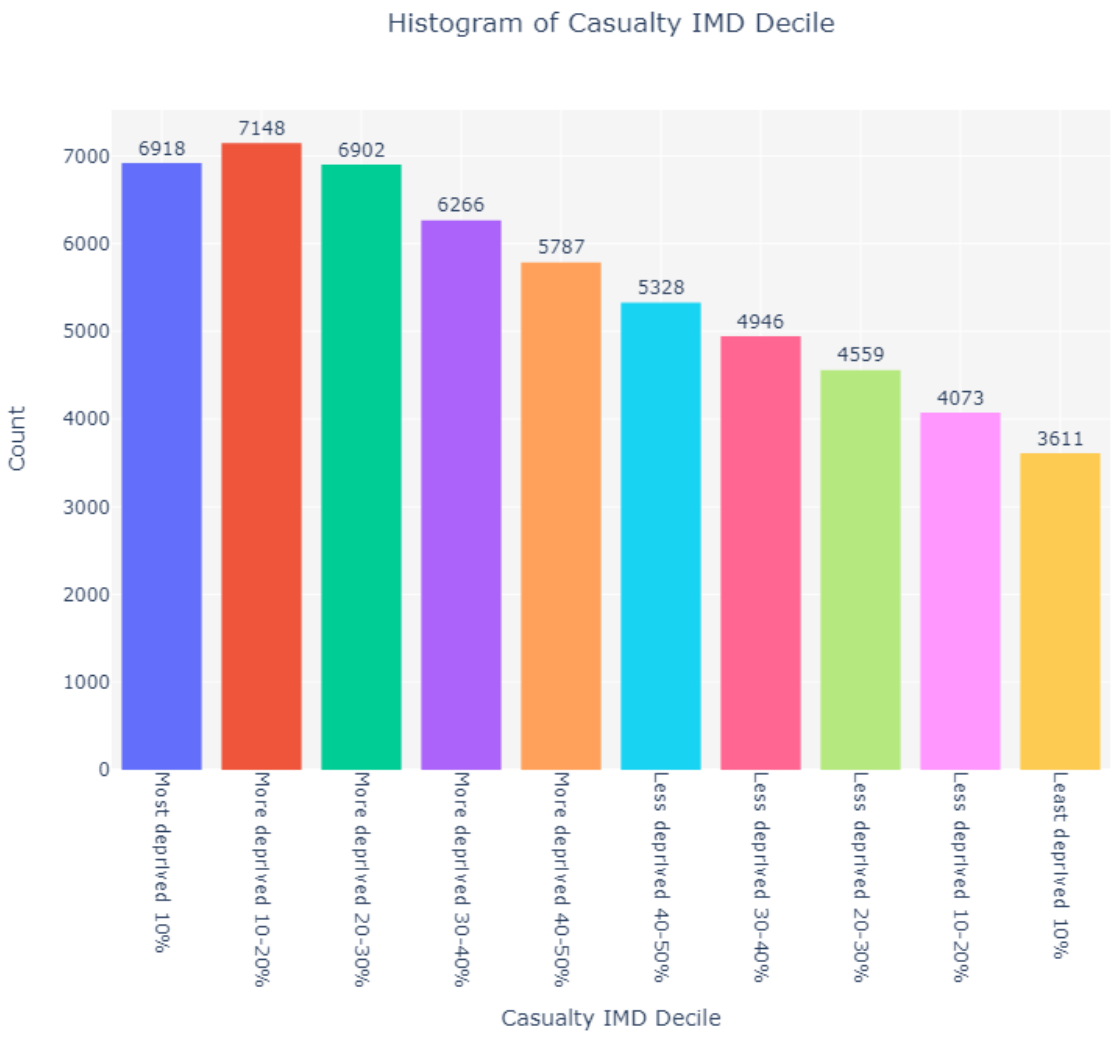


Figure 14. Casualty IMD decile Histogram

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

More info: England and Wales only. See Office for National Statistics (ONS) guidance: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

3. Inferences

1. The histograms suggest that the majority of casualties involved in accidents during 2022 exhibit the following characteristics:

Table 14. Typical Characteristics of casualties in 2022

Feature	Value
vehicle_reference	Pedal cycle
casualty_class	Driver or rider
sex_of_casualty	Male
age_band_of_casualty	26-35
casualty_severity	Slight
pedestrian_location	Not a Pedestrian
pedestrian_movement	Not a Pedestrian
car_passenger	Not car passenger
bus_or_coach_passenger	Not a bus or coach passenger
pedestrian_road_maintenance_worker	No / Not applicable
casualty_type	Car occupant
casualty_home_area_type	Urban area
casualty_imd_decile	More deprived 10-20%

2. This plot illustrates the frequency of individuals involved per accident, predominantly showing a count of one, indicating that the majority of accidents involved only one person.

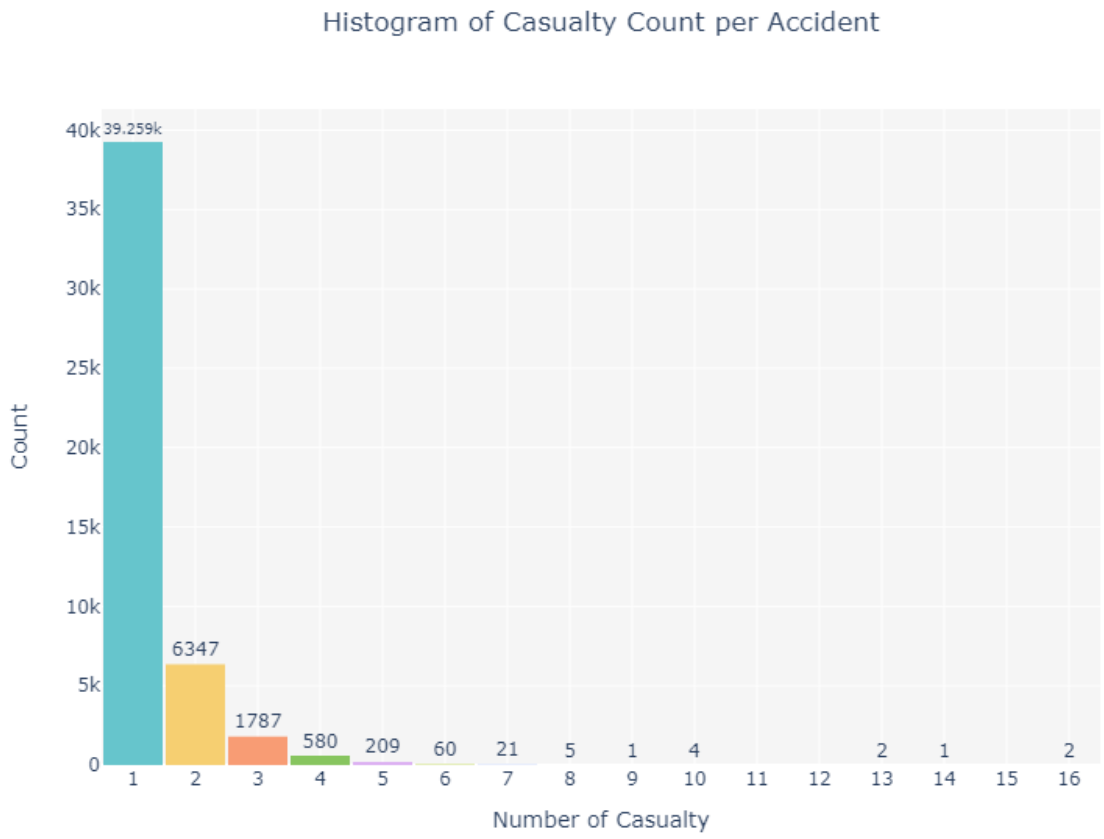


Figure 15. The frequency of individuals involved per accident

3. The figure below displays the correlation values between casualty severity and other features, highlighting which factors are more closely associated with the level of accident severity.

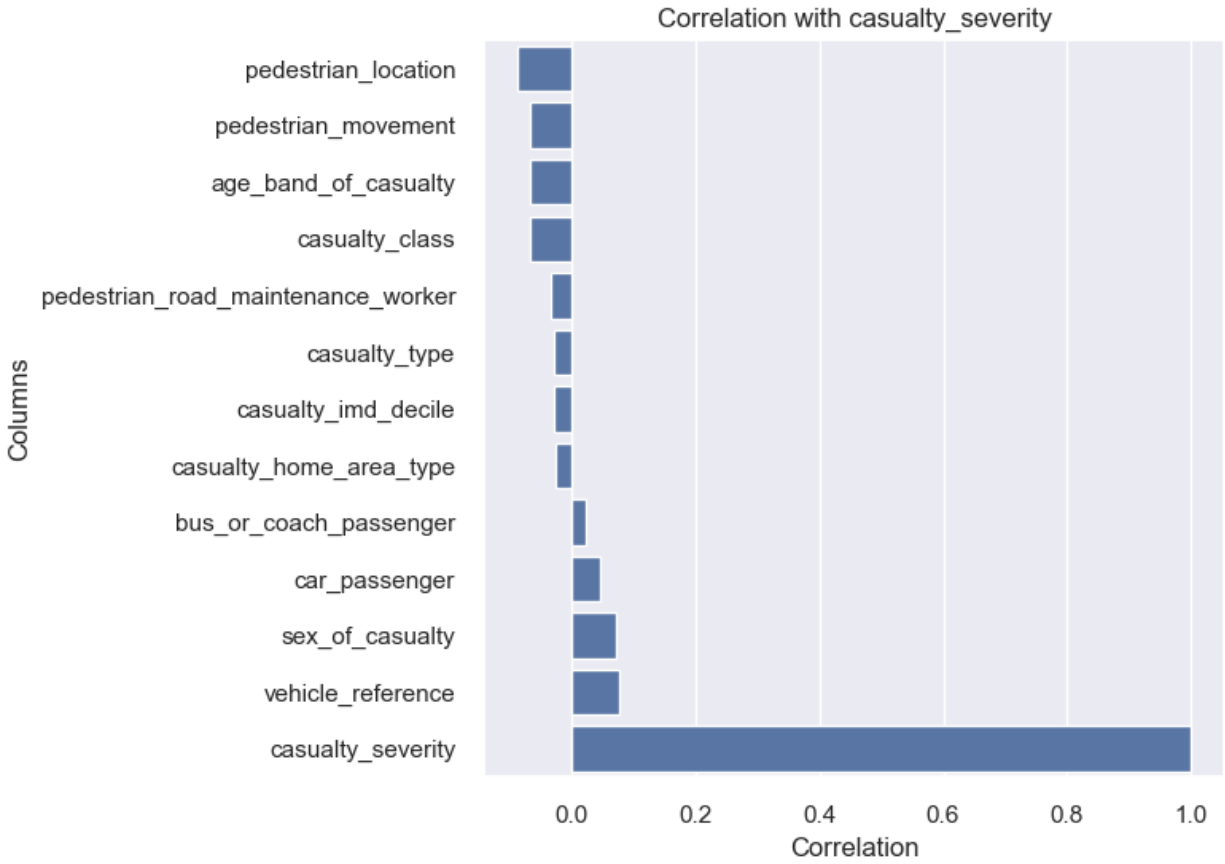
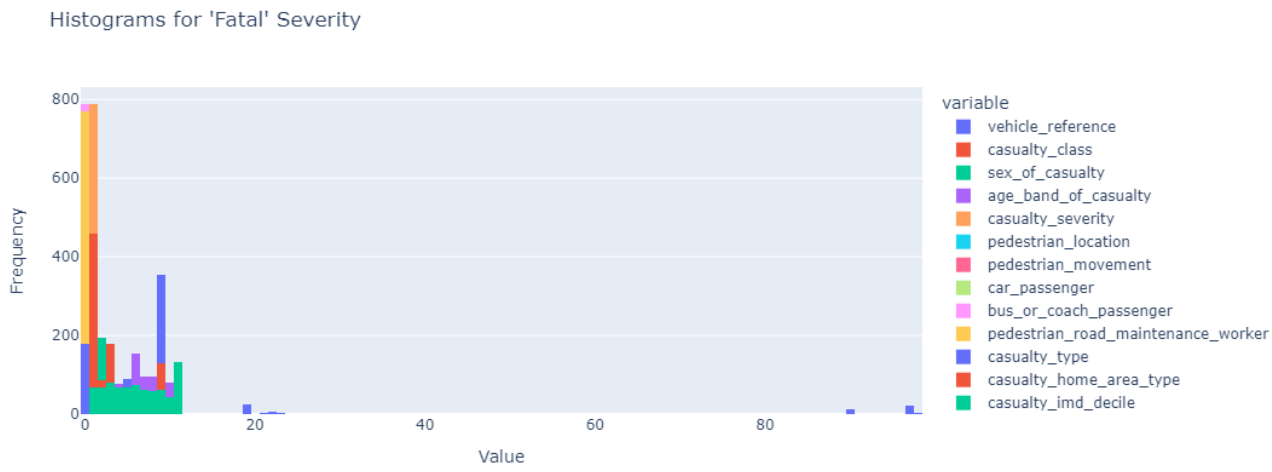
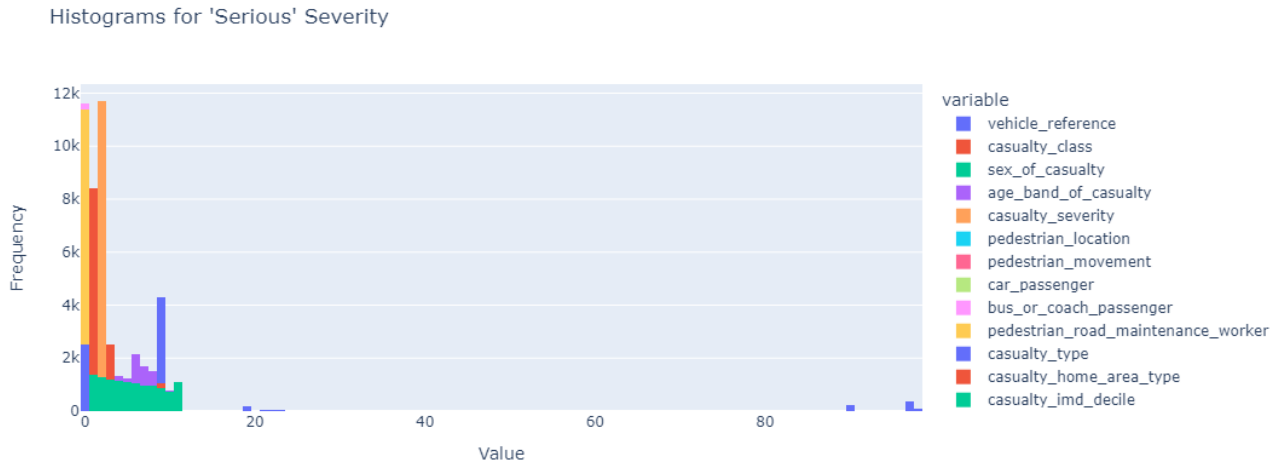


Figure 16. Correlation values between casualty severity and other features

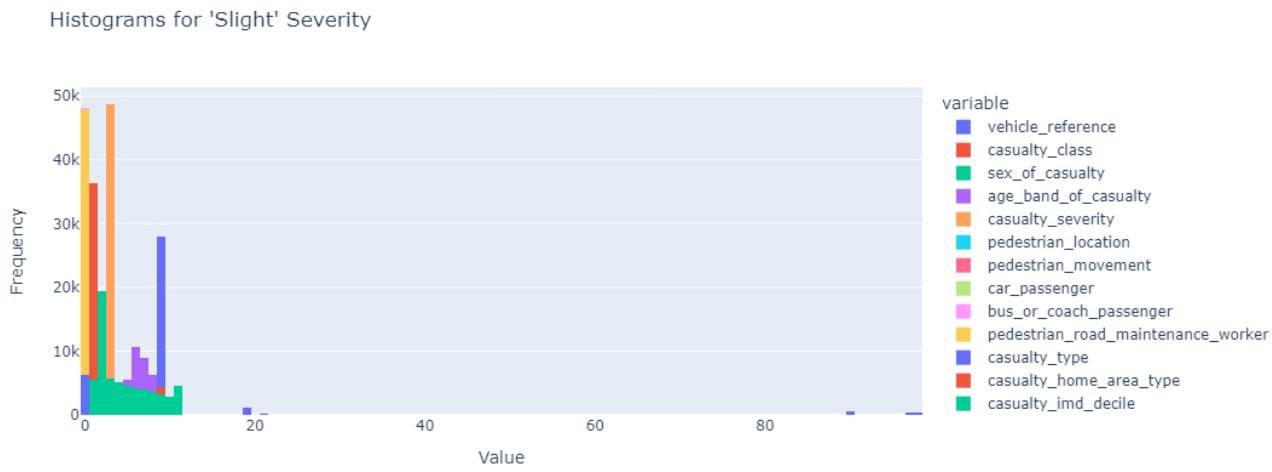
4. Histograms can also be plotted for each specific severity category, including slight, serious, and fatal, to provide further insight into the distribution within these categories.



(A)



(B)



(C)

Figure 17. (A, B, and C) Histograms for Each Severity Category

5. This figure illustrates that the majority of accidents occur when there are no pedestrians present, particularly noting that a higher number of fatal casualties resulted from incidents without pedestrians.

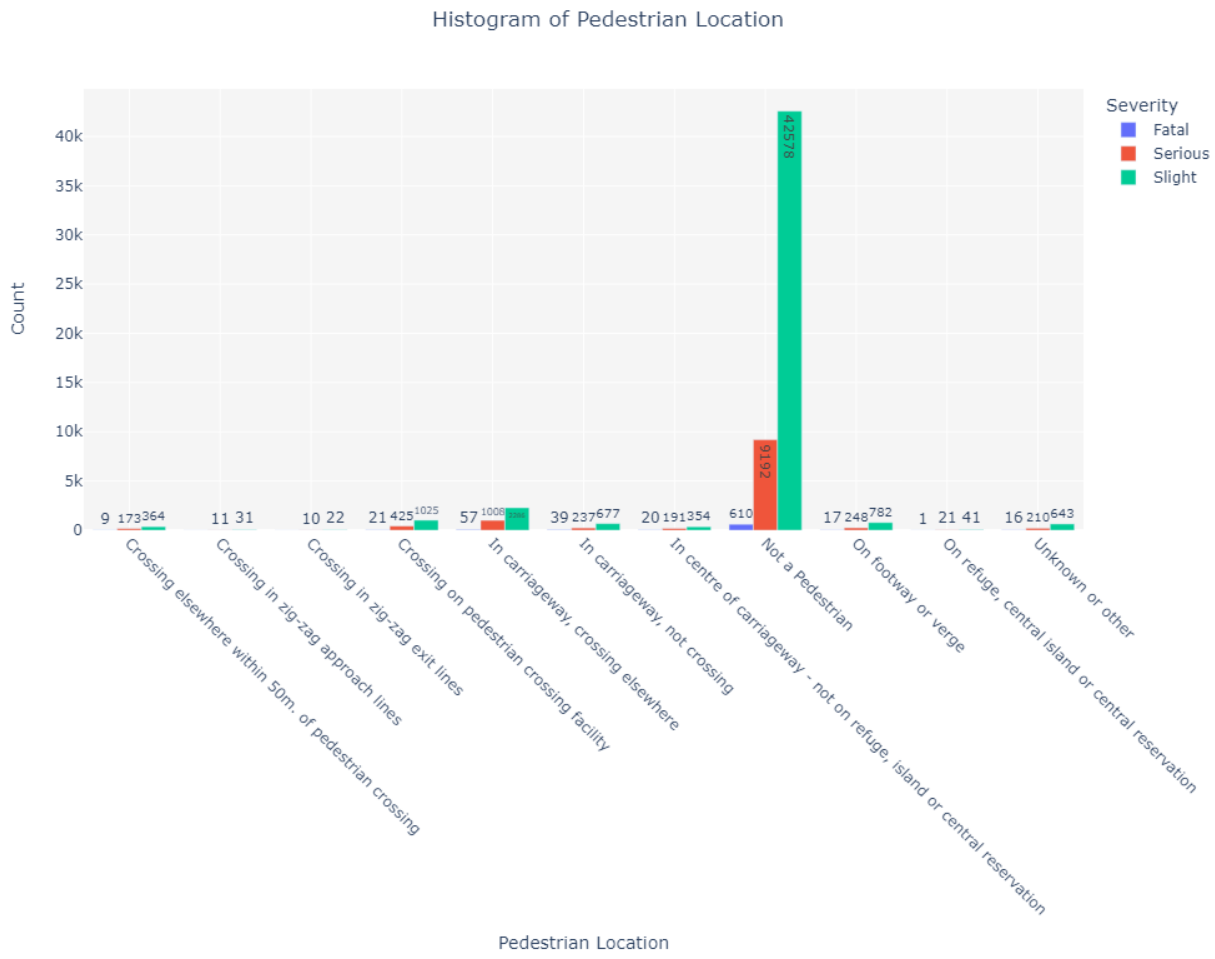


Figure 18. Histogram of Pedestrian Locations by Severity Level

6. Car occupants, pedestrians, and cyclists consistently represent the most frequently involved casualties across all severity levels.

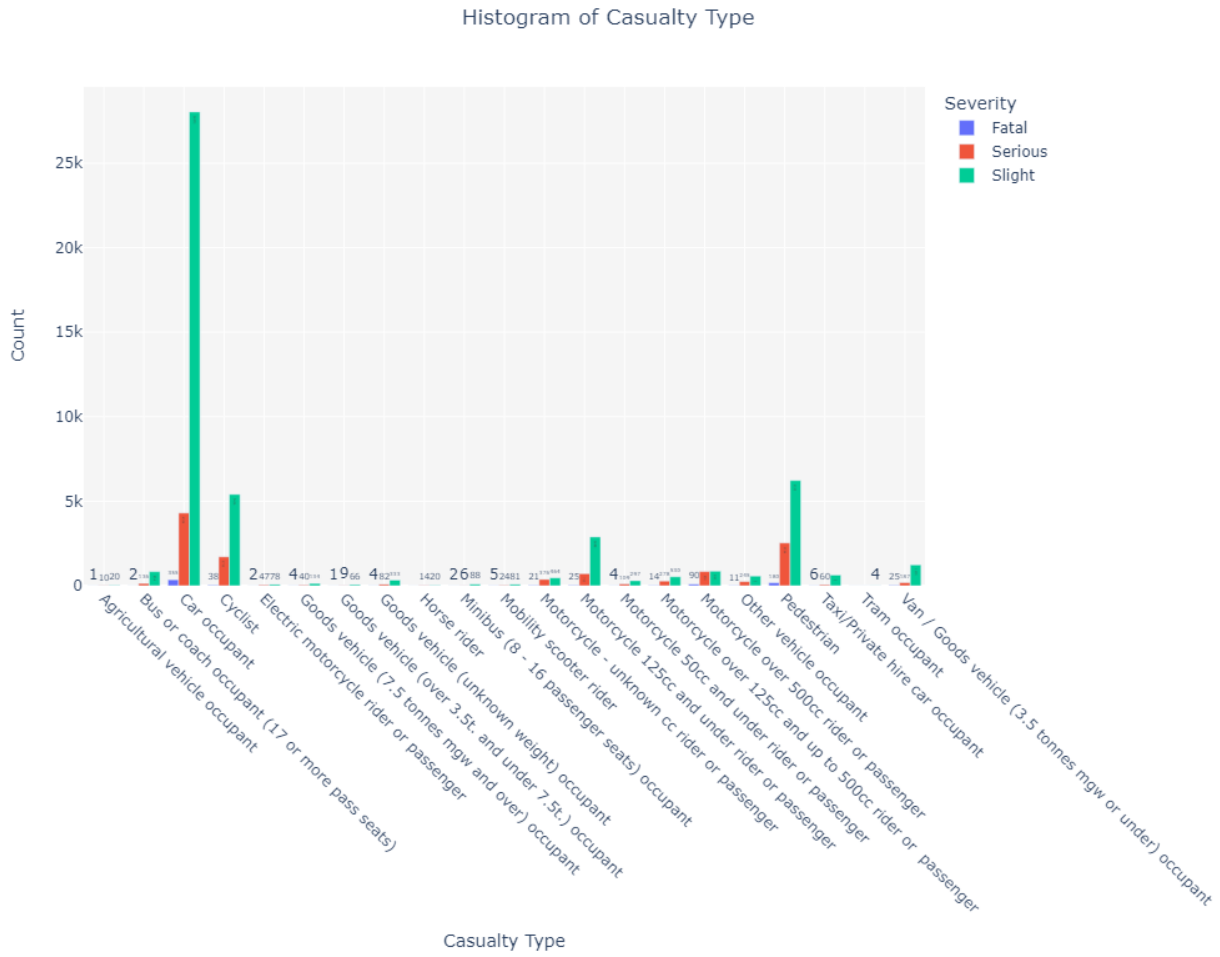


Figure 19. Histogram of Casualty Types by Severity Level

7. It is evident that the majority of both male and female casualties fell within the age range of 26 to 35.

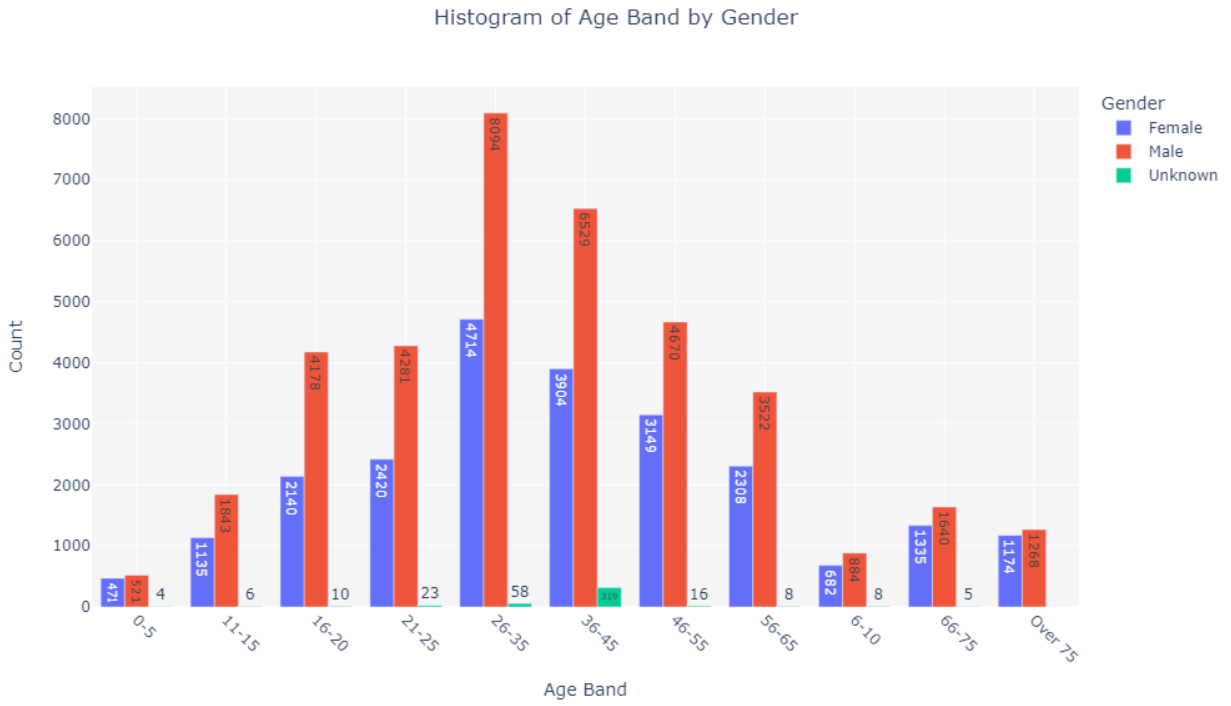


Figure 20. Histogram of Age Bands by Gender Type

Further analysis can reveal additional insights from the dataset, and the ones highlighted here are just a subset of the potential discoveries. We can delve deeper into the data using clustering algorithms, as demonstrated in the Jupyter notebook. Additionally, we can categorize severity levels into two groups: severe (combining fatal and serious) and not severe (slight), assigning each row accordingly. By employing classification algorithms, we can predict the conditions under which accidents are likely to be severe. However, a challenge with this dataset is the imbalance between severe and non-severe cases; (hopefully) there are significantly fewer samples for severe accidents. To address this, techniques such as SMOTE can be utilized to balance the data between classes. Failure to do so may result in lower accuracy when predicting severe accidents. The confusion matrix generated using the random forest algorithm is provided below for reference.

Classification Report:

	precision	recall	f1-score	support
Slight	0.81	0.94	0.87	9756
Fatal/Serious	0.38	0.15	0.21	2508
accuracy			0.78	12264
macro avg	0.59	0.54	0.54	12264
weighted avg	0.72	0.78	0.73	12264

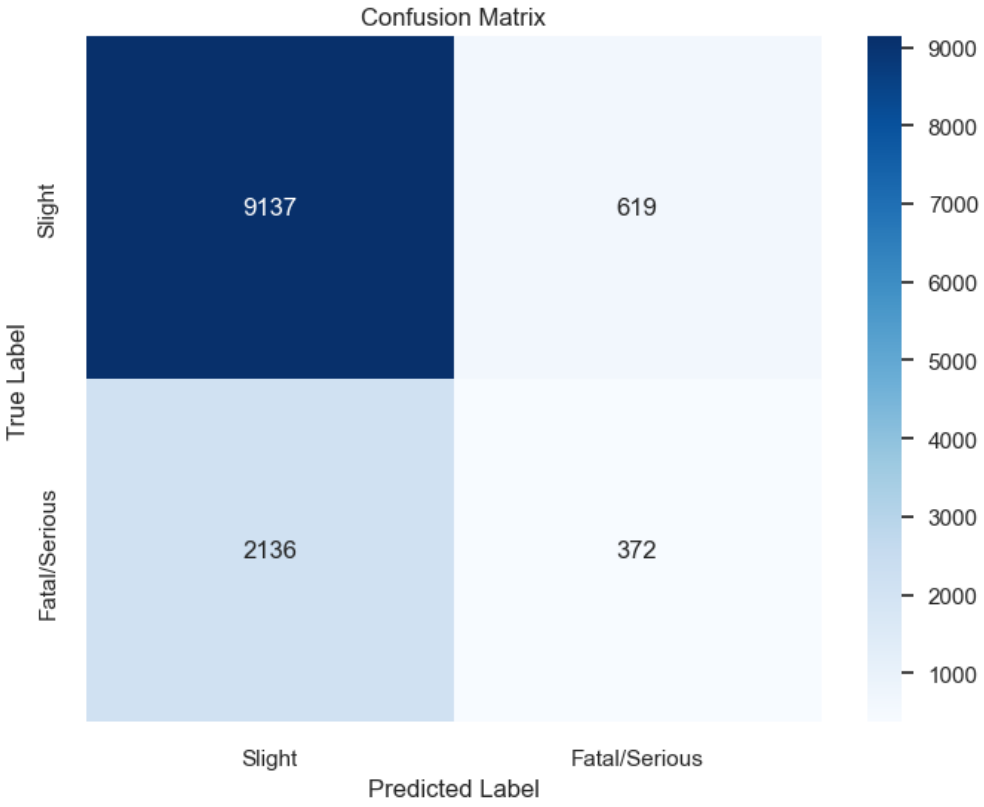


Figure 21. Confusion matrix resulted from Random Forest Classifier

4. Recommendations

1. Since the majority of accidents involve only one person, initiatives should focus on addressing factors contributing to single-vehicle accidents. This could include campaigns targeting driver distraction, fatigue, speeding, and alcohol consumption.

2. Understanding the factors correlated with different levels of accident severity is crucial. Interventions should prioritize addressing these factors.

3. Although most accidents occur without pedestrians present, enhancing pedestrian safety measures at crossings and busy urban areas can further reduce the risk of accidents involving pedestrians.

4. Since car occupants, pedestrians, and cyclists are consistently the most frequently involved casualties, safety measures tailored to these groups should be prioritized. This could include improving cyclist lanes and pedestrian crossings, implementing stricter enforcement of traffic rules concerning interactions between vehicles and vulnerable road users, and raising awareness among all road users about the importance of sharing the road safely.

5. Given that the majority of casualties fall within the age range of 26 to 35, targeted safety programs tailored to this demographic could be effective. These programs could focus on addressing risky driving behaviors, promoting responsible alcohol consumption, and increasing awareness of the importance of adhering to traffic laws.

6. General awareness campaigns focusing on safe driving practices, the dangers of driving under the influence, and the importance of wearing seat belts can have a significant impact on reducing accidents across all demographics.

7. Investing in road infrastructure improvements, such as better lighting, signage, road markings, and junction design, can help mitigate accident risks and improve overall road safety.

8. Given that pedal cyclists are heavily involved in accidents, investing in dedicated cycling lanes separated from motor traffic can reduce the risk of collisions between cyclists and vehicles.

9. Implement measures to enhance safety in urban areas, such as traffic calming measures, improved signage, and better lighting, especially in areas identified as more deprived.

10. Regularly monitor and analyze road accident data to identify emerging trends and adjust road safety strategies accordingly.