**Analysis of Driving Accident Data Report**

April 2024

**Company Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Analysis conducted by: Milad Tahvilzade**

# Introduction

Road traffic accidents constitute a significant global public health concern, with devastating impacts on individuals, families, and societies at large. According to the World Health Organization (WHO), approximately 1.35 million people lose their lives each year due to road traffic crashes, and an additional 20 to 50 million sustain non-fatal injuries, many of which result in permanent disabilities. These accidents not only cause immense human suffering but also impose substantial economic burdens, costing nations around 3% of their gross domestic product.

The disproportionate burden of road traffic injuries falls on low- and middle-income countries, where over 90% of fatalities occur despite these countries accounting for only 60% of the world's vehicles. Within high-income countries, individuals from lower socioeconomic backgrounds are at a heightened risk of being involved in road traffic accidents.

Various factors contribute to the occurrence of road traffic accidents, including speeding, driving under the influence of alcohol or drugs, non-use of safety measures such as seat belts and helmets, distracted driving, unsafe road infrastructure, and inadequate post-crash care. The WHO advocates for a holistic approach to road safety, emphasizing the importance of addressing these factors through interventions spanning multiple sectors, including transportation, law enforcement, health, and education.

The Safe System approach, endorsed by WHO, emphasizes the need to design transport systems that are forgiving of human error and prioritize the safety of all road users. This approach underscores the importance of safe roads and roadsides, safe speeds, safe vehicles, and safe road users in preventing fatal crashes and reducing serious injuries.

In light of the alarming global toll of road traffic accidents, there is a pressing need for comprehensive strategies aimed at preventing these incidents. Effective interventions include designing safer infrastructure, improving vehicle safety standards, enforcing traffic laws, enhancing post-crash care, and raising public awareness. The United Nations has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2030, underscoring the urgency of concerted action.

This report presents an analysis of road accidents using comprehensive datasets sourced from the "Road Accidents Data 2022" by Juhi Bhojani available on Kaggle. The objective is to uncover insights into the underlying causes and circumstances surrounding traffic accidents, employing various analytical methods to facilitate evidence-based decision-making for enhancing road safety.

# Description of Analysis Methods

# Dataset Description:

The data used in this analysis was sourced from the Department for Transport (DfT) for the year 2022. The dataset, titled "Road Accidents Data 2022" by Juhi Bhojani, was obtained from Kaggle. It provides comprehensive information on road accidents reported over multiple years, encompassing various attributes such as accident status, vehicle and casualty details, demographics, and severity of casualties.

The dataset utilized in this analysis serves as a valuable resource for examining road accidents, discerning trends, and comprehending factors that contribute to accidents. With 61,352 rows and 20 columns, it offers a robust collection of information on traffic accidents reported throughout the year 2022.

This dataset provides comprehensive details on various aspects of traffic accidents, including accident status, vehicle and casualty specifics, demographics, and severity of casualties. By leveraging this rich dataset, analysts can delve into nuanced patterns, identify key factors influencing accidents, and inform evidence-based decision-making processes aimed at enhancing road safety measures.

# Dataset columns

# In the provided dataset, each column contains values represented by specific codes. Below, we define the meaning of each code within the dataset:

* **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:** Reference numbers for the involved vehicles in the accident.
* **Casualty\_Reference:** Reference numbers for the casualties involved in the accident.
* The **"Casualty\_Class"** column in the dataset denotes the class of the casualty involved in the accident. Below are the corresponding class codes and their descriptions:
  + **1:** Driver or rider
  + **2:** Passenger
  + **3:** Pedestrian
* The **"Sex\_of\_Casualty**" column in the dataset signifies the gender of the casualty involved in the accident. Below are the corresponding codes and their descriptions:
  + **1:** Male
  + **2:** Female
  + **9:** Unknown (self-reported)
  + **-1:** Data missing or out of range
* The **"Age\_Band\_of\_Casualty"** column in the dataset represents the age group to which the casualty belongs. Below are the corresponding codes and their descriptions:
  + **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
* The **"Casualty\_Severity"** column in the dataset indicates the severity of the casualty's injuries. Below are the corresponding codes and their descriptions:
  + **1:** Fatal
  + **2:** Serious
  + **3:** Slight
* The **"pedestrian\_location"** column in the dataset provides information about the location of the pedestrian at the time of the accident. Below are the corresponding codes and their descriptions:
  + **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
* The **"Pedestrian\_Movement"** column in the dataset indicates the movement of the pedestrian during the accident. Below are the corresponding codes and their descriptions:
  + **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
* The **"Car\_Passeger"** column in the dataset indicates the role of the casualty as a car passenger at the time of the accident. Below are the corresponding codes and their descriptions:
  + **0:** Not a car passenger
  + **1:** Front seat passenger
  + **2:** Rear seat passenger
  + **9:** Unknown (self-reported)
  + **-1:** Data missing or out of range
* The **"Bus\_or\_Coach\_Passenger"** column in the dataset indicates the role of the casualty as a bus or coach passenger at the time of the accident. Below are the corresponding codes and their descriptions:
  + **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
* The **"Pedestrian\_Road\_Maintenance\_Worker"** column in the dataset indicates whether the casualty was a road maintenance worker at the time of the accident. Below are the corresponding codes and their descriptions:
  + **0:** No / Not applicable
  + **1:** Yes
  + **2:** Not Known
  + **3:** Probable
  + **-1:** Data missing or out of range
* The **"Casualty\_Type"** column in the dataset indicates the type of casualty involved in the accident. Below are the corresponding codes and their descriptions:
  + **0:** Pedestrian
  + **1:** Cyclist
  + **2:** Motorcycle 50cc and under rider or passenger
  + **3:** Motorcycle 125cc and under rider or passenger
  + **4:** Motorcycle over 125cc and up to 500cc rider or passenger
  + **5:** Motorcycle over 500cc rider or passenger
  + **8:** Taxi/Private hire car occupant
  + **9:** Car occupant
  + **10:** Minibus (8 - 16 passenger seats) occupant
  + **11:** Bus or coach occupant (17 or more passenger seats)
  + **16:** Horse rider
  + **17:** Agricultural vehicle occupant
  + **18:** Tram occupant
  + **19:** Van / Goods vehicle (3.5 tonnes mgw or under) occupant
  + **20:** Goods vehicle (over 3.5t. and under 7.5t.) occupant
  + **21:** Goods vehicle (7.5 tonnes mgw and over) occupant
  + **22:** Mobility scooter rider
  + **23:** Electric motorcycle rider or passenger
  + **90:** Other vehicle occupant
  + **97:** Motorcycle - unknown cc rider or passenger
  + **98:** Goods vehicle (unknown weight) occupant
  + **-1:** Data missing or out of range
* The **"Casualty\_Home\_Area\_Type"** column in the dataset indicates the type of area where the casualty resides. Below are the corresponding codes and their descriptions:
  + **1:** Urban area
  + **2:** Small town
  + **3:** Rural
  + **-1:** Data missing or out of range
* The **"Casualty\_IMD\_Decile"** column in the dataset represents the Index of Multiple Deprivation (IMD) decile for the casualties' home areas. Below are the corresponding codes and their descriptions:
  + **1:** Most deprived 10%
  + **2:** More deprived 10-20%
  + **3:** More deprived 20-30%
  + **4:** More deprived 30-40%
  + **5:** More deprived 40-50%
  + **6:** Less deprived 40-50%
  + **7:** Less deprived 30-40%
  + **8:** Less deprived 20-30%
  + **9:** Less deprived 10-20%
  + **10:** Least deprived 10%
  + **-1:** Data missing or out of range
* The **"LSOA\_of\_Casualty"** column represents the Lower Layer Super Output Area (LSOA) associated with the casualty's location.

# Data Prepration

* Checking for Duplicates

Before proceeding with data cleaning, we checked for any duplicate rows in the dataset. Fortunately, there were no duplicate entries found.

* Removing Extra Columns

Next, we removed the columns that do not contribute to our analysis or contain redundant information. These columns include:

Status

1. Accident index
2. Accident year
3. Accident reference
4. LSOA of casualty
5. Vehicle reference
6. Casualty reference

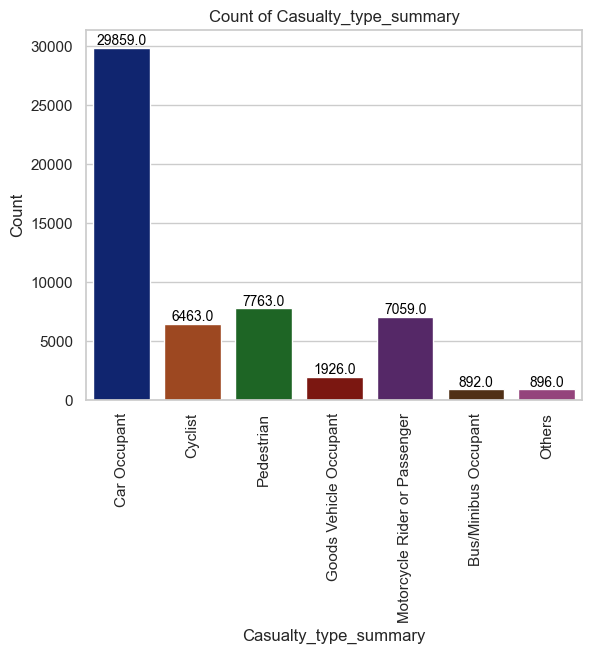
# Summarization and Creation of New Columns

In this section, we'll condense the dataset's columns, focusing on extremely precise categories. By breaking down these columns meticulously, we'll generate new columns to reflect these condensed categories. This process grants us a deeper comprehension of the intricate patterns and trends within the data. Such an approach not only streamlines the analysis but also organizes the data more effectively. Ultimately, it empowers us to make better-informed decisions and implement specific strategies to enhance road safety.

**1. Summarization of Casualty Type**

To simplify the analysis, we summarized the type of casualty involved in the accident into broader categories. The summarized categories, along with their respective counts, are illustrated in Figure 1. The categories are as follows:

* **Car Occupant**: Individuals involved in accidents while driving or riding in a car.
* **Motorcycle Rider or Passenger**: Individuals involved in accidents while riding motorcycles.
* **Goods Vehicle Occupant**: Individuals involved in accidents while riding in goods vehicles.
* **Bus/Minibus Occupant**: Individuals involved in accidents while riding in buses or minibuses.
* **Pedestrian**: Individuals involved in accidents while walking or crossing the road.
* **Cyclist**: Individuals involved in accidents while cycling.
* **Others**: Individuals involved in accidents categorized as other types not covered by the above categories.

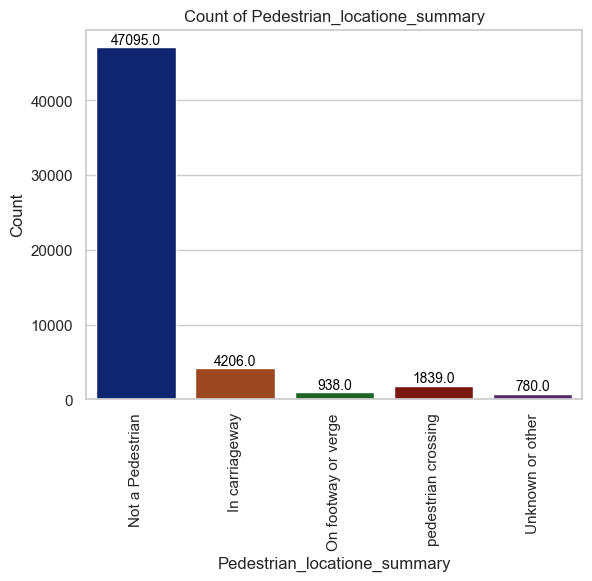


Figure

**2. Summarization of Pedestrian Location**

We categorized the pedestrian's location at the time of the accident into broader categories. The summarized categories, along with their respective counts, are illustrated in Figure 2. The categories are as follows:

* **Unknown or other**: Pedestrian location is unknown or falls into other unspecified categories.
* **Pedestrian Crossing**: Pedestrians involved in accidents while crossing at designated pedestrian crossings or zig-zag approaches.
* **In Carriageway**: Pedestrians involved in accidents while walking or crossing elsewhere on the carriageway.
* **On Footway or Verge**: Pedestrians involved in accidents while on footways or verges adjacent to the road.
* **Not a Pedestrian**: Individuals categorized as non-pedestrians involved in accidents.

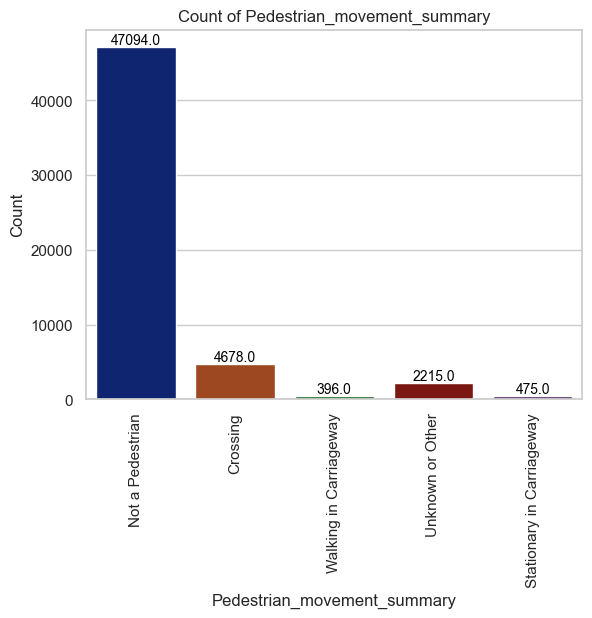


Figure

**3. Categorization of Pedestrian Movement**

We categorized the movement of pedestrians involved in the accident into broader categories. The summarized categories, along with their respective counts, are illustrated in Figure 3. These categories include:

* **Not a Pedestrian**: Individuals categorized as non-pedestrians involved in accidents.
* **Crossing**: Pedestrians involved in accidents while crossing the road.
* **Stationary in Carriageway**: Pedestrians involved in accidents while standing or playing in the carriageway.
* **Walking in Carriageway**: Pedestrians involved in accidents while walking along the carriageway, either facing or back to traffic.
* **Unknown or Other**: Pedestrians involved in accidents with movement patterns not fitting into the above categories.

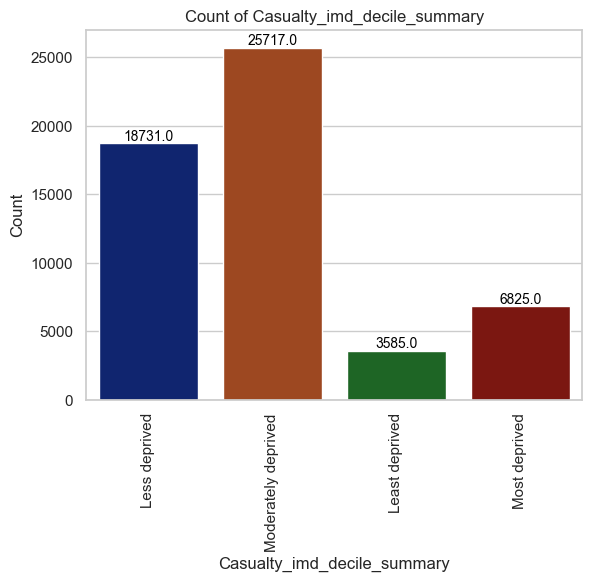


Figure

**4. Summarization of Casualty IMD Decile**

We categorized the Index of Multiple Deprivation (IMD) decile categories into broader summaries using the provided function. These summaries help in simplifying the analysis and understanding the deprivation levels of casualties. The summarized categories, along with their respective counts, are illustrated in Figure 4. The categories are as follows:

* **Most deprived:** Represents the most deprived decile categories.
* **Moderately deprived:** Represents moderately deprived decile categories.
* **Less deprived:** Represents less deprived decile categories.
* **Least deprived:** Represents the least deprived decile categories.



Figure

# New columns

# Masked or not masked

# Combining fetal and serious

# Handling Missing Values

In this dataset, the only numerical column is the "age" column, which contains 1068 missing values. The distribution of this column, as shown in Figure 5, indicates a right-skewed distribution. Therefore, we opt to impute the missing values using the median. Additionally, after imputing missing values in the "age" column, we will update the "age\_band\_of\_casualty" column accordingly to ensure consistency and fill any missing values in this column.

As for other columns with missing values, we prefer to remove them entirely from the dataset.

This approach ensures that we maintain the integrity of the age data while eliminating rows with missing values in other columns, thereby minimizing potential biases in our analysis.

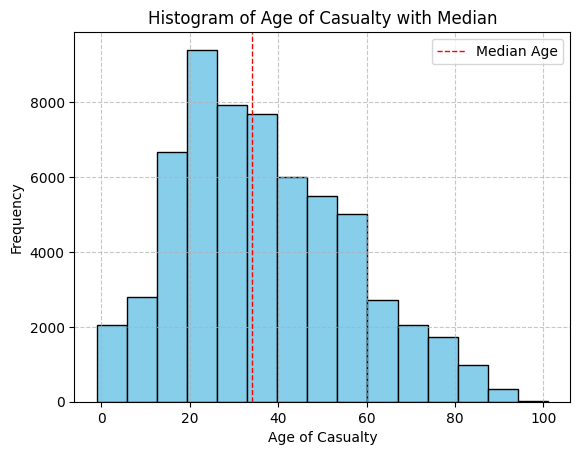
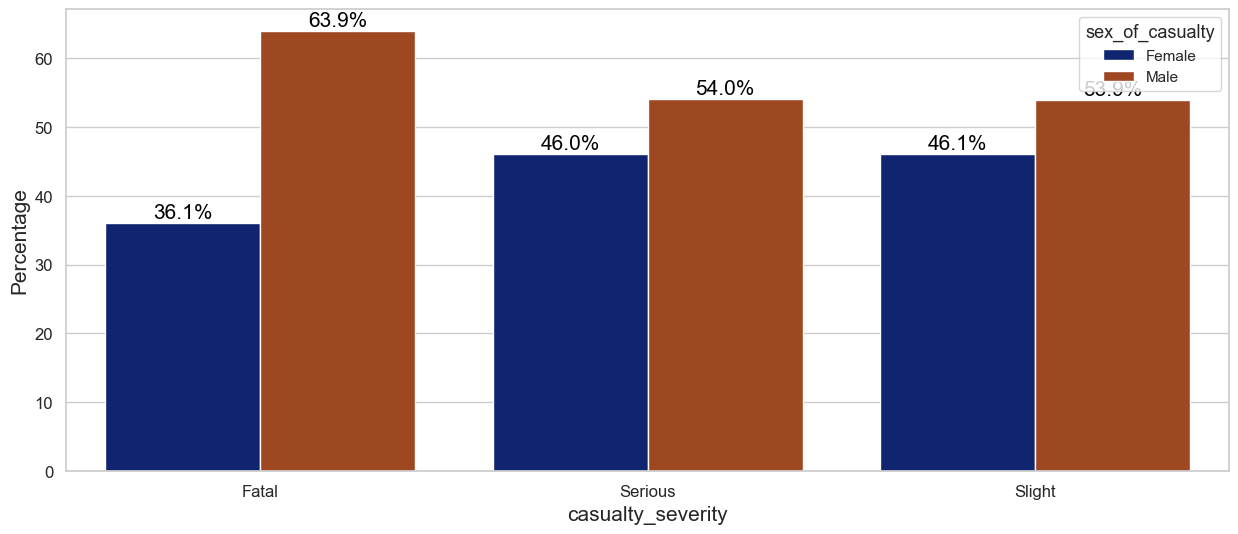


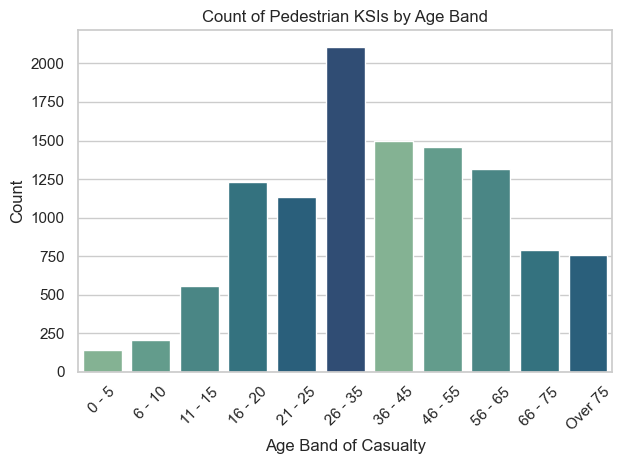
Figure Distribution of age\_of\_casualty Column

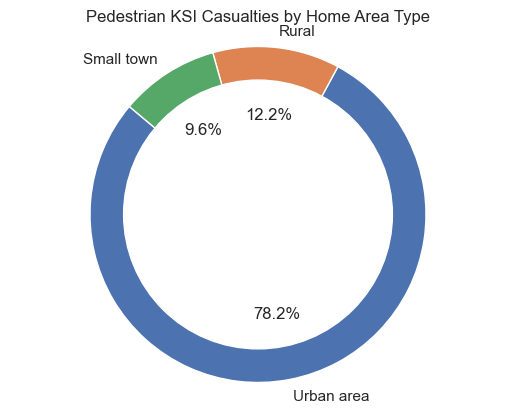
# Exploratory Data Analysis

# Pedestrians, alongside motorcyclists, horse riders, and pedal cyclists, represent some of the most vulnerable user groups on the road. Unlike car users, they lack the protection of a vehicle body, making them more susceptible to injuries and accidents. Additionally, their smaller size and visibility make them harder for other drivers to detect on the road, further increasing their vulnerability. Given the dataset's comprehensive information on pedestrian injuries, separate analyses are conducted for pedestrian injuries and injuries involving other parties.

# 

توزیع سن برای عابرین پیاده که کشته شده یا بصورت جدی اسیب دیده اند



****