Title: Collision and Casualty Analysis (2022 Dataset)

Author: Mohsen Amiri Amjad

Github: https://github.com/MohsenAmiri79/Traffic-Accidents-and-Casualty-Analysis-2022-only

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

Introduction

This report presents a comprehensive analysis of the “Car Accidents 2022” dataset obtained from Kaggle, a renowned platform for data science and machine learning. The dataset provides an extensive compilation of data on road accidents reported over the year 2022, encompassing a wide range of attributes related to accidents, vehicles, and casualties.

The objective of this analysis is to identify patterns and correlations within the data that could provide valuable insights into the factors contributing to road accidents and their severity. By understanding these factors, we aim to propose data-driven recommendations for improving road safety and minimizing casualties.

The report is structured as follows: First, we present an overview of the dataset and its features. Next, we detail the data cleaning and preprocessing steps undertaken to ensure the reliability of our analysis. We then delve into the data exploration and analysis, utilizing various statistical techniques and visualizations to uncover patterns and correlations. Finally, we conclude with a summary of our findings and propose future recommendations based on these insights.

Through this analysis, we hope to contribute to the ongoing efforts to enhance road safety and reduce the impact of road accidents on individuals and communities. We believe that data-driven decision-making can play a crucial role in these efforts, and we hope that our analysis will serve as a valuable resource for researchers, policymakers, and analysts in this field.

Overview

This dataset offers an extensive compilation of data on road accidents reported over 2022. It covers a wide range of attributes pertaining to the status of accidents, references for vehicles and casualties, demographic details, and the severity of injuries. Key elements such as pedestrian specifics, types of casualties, involvement of road maintenance workers, and the Index of Multiple Deprivation (IMD) decile for the home areas of casualties are included.

FEATURE	DESCRIPTION
STATUS	Current state of the accident (e.g., reported, under investigation).
ACCIDENT INDEX	Unique identifier assigned to each reported accident.
ACCIDENT YEAR	The year the accident took place.
ACCIDENT REFERENCE	Reference number linked to the accident.
VEHICLE REFERENCE	Reference number assigned to the vehicle involved in the accident.
CASUALTY REFERENCE	Reference number assigned to the casualty in the accident.
CASUALTY CLASS	Class of the casualty (e.g., driver, passenger, pedestrian).
SEX OF CASUALTY	Gender of the casualty (male or female).
AGE OF CASUALTY	Age of the casualty.
AGE BAND OF CASUALTY	Age group of the casualty (e.g., 0-5, 6-10, 11-15).
CASUALTY SEVERITY	Severity of the casualty's injuries (e.g., fatal, serious, slight).
PEDESTRIAN LOCATION	Location of the pedestrian when the accident occurred.
PEDESTRIAN MOVEMENT	Movement of the pedestrian at the time of the accident.
CAR PASSENGER	Indicates if the casualty was a car passenger during the accident (yes or no).
BUS OR COACH PASSENGER	Indicates if the casualty was a bus or coach passenger (yes or no).
PEDESTRIAN ROAD MAINTENANCE WORKER	Indicates if the casualty was a road maintenance worker (yes or no).
CASUALTY TYPE	Type of casualty (e.g., driver/rider, passenger, pedestrian).
CASUALTY HOME AREA TYPE	Type of area where the casualty resides (e.g., urban, rural).
CASUALTY IMD DECILE	IMD decile of the casualty's residential area (a measure of deprivation).
LSOA OF CASUALTY	The Lower Layer Super Output Area (LSOA) linked to the casualty's location.

This dataset is a valuable resource for analyzing road accidents, identifying patterns, and formulating safety measures to minimize casualties and improve road safety. It can be utilized by researchers, policymakers, and analysts for data-driven decision-making and enhancement of overall road transportation systems.

Literature Review

The study of road accidents and their contributing factors has been the subject of extensive research over the years. Various studies have explored the role of different factors such as driver behavior, vehicle characteristics, road conditions, and environmental factors in road accidents.

Driver Behavior: Numerous studies have highlighted the role of driver behavior in road accidents. Factors such as speeding, distracted driving, and driving under the influence have been identified as major contributors to road accidents. For instance, a study by the World Health Organization (WHO) found that speeding contributes to about 30% of deaths on the road.

Vehicle Characteristics: The characteristics of the vehicle involved in the accident also play a significant role. Studies have shown that vehicle age, type, and safety features can significantly impact the severity of an accident.

Road and Environmental Conditions: Road conditions, including road design, signage, and lighting, have been found to influence accident rates. Similarly, environmental conditions such as weather and time of day also play a role.

Demographic Factors: Demographic factors such as the age and gender of the driver have also been associated with accident risk. For example, young drivers are often found to be at a higher risk of being involved in accidents.

The “Car Accidents 2022” dataset provides a comprehensive compilation of data on road accidents, allowing for an in-depth analysis of these and other factors. By leveraging this dataset, we aim to contribute to the existing body of knowledge on road safety and provide data-driven insights to inform policy-making and intervention design.

Data Preprocessing

The project involves cleaning and preprocessing a dataset related to road casualties. The dataset comprises 13 columns, most of which are categorical, with ‘age of casualty’ being the only numerical column. The rest of the columns were not useful for this analysis and were removed in the beginning. The columns include various attributes such as ‘vehicle reference’, ‘casualty class’, ‘sex of casualty’, ‘age of casualty’, ‘age band of casualty’, ‘casualty severity’, ‘pedestrian location’, ‘pedestrian movement’, ‘car passenger’, ‘bus or coach passenger’, ‘pedestrian road maintenance worker’, ‘casualty home area type’, and ‘is pedestrian’.

The initial dataset overview revealed missing values in several columns. The ‘casualty home area type’ column had the most missing values (5498), followed by ‘age of casualty’ and ‘age band of casualty’ (both with 1350 missing values). Other columns with missing values included ‘sex of casualty’ (448), ‘car passenger’ (314), ‘bus or coach passenger’ (23), and ‘pedestrian road maintenance worker’ (113).

The data cleaning process involved handling these missing values using a Decision Tree Classifier model. The model was trained on the non-missing values and used to predict the missing values. This approach was applied to the following columns: ‘sex of casualty’, ‘age of casualty’, ‘age band of casualty’, ‘car passenger’, ‘bus or coach passenger’, and ‘pedestrian road maintenance worker’.

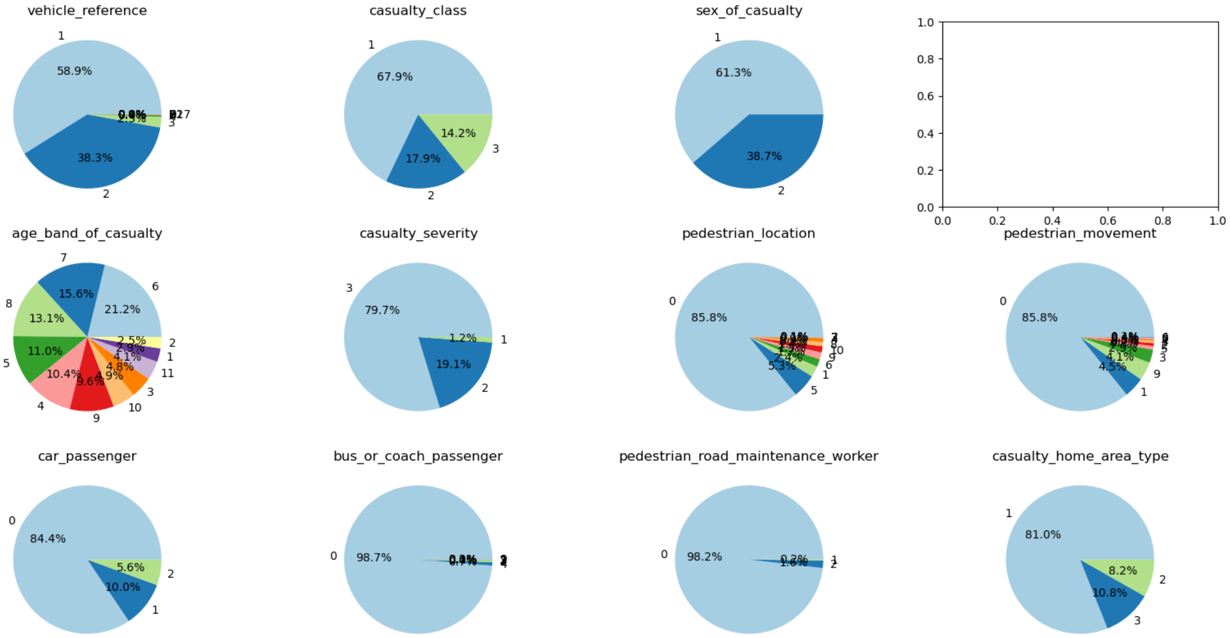
After cleaning, the dataset was further refined to only include rows where ‘age of casualty’ is greater than or equal to 0, ‘sex of casualty’ is either 1 or 2, and ‘car passenger’ is between 0 and 2 (inclusive).

This comprehensive data cleaning and preprocessing approach has ensured that the dataset is now ready for further analysis or modeling, with missing values appropriately handled and the data in a suitable format for downstream tasks. The cleaned dataset maintains the integrity of the original data while providing a more robust foundation for deriving insights related to road casualties.

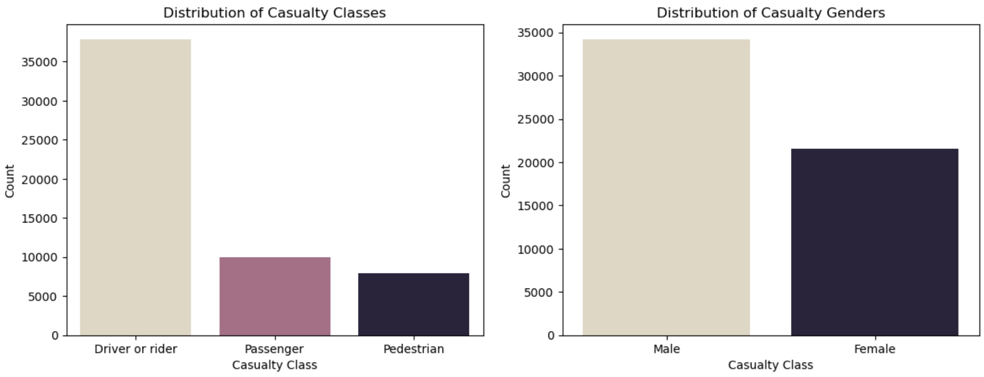
Exploring the Data

The initial dataset comprises approximately 61.3K accident records, each accompanied by casualty information. Following a thorough cleaning and preprocessing phase, the dataset is reduced to 55.8K records, all of which are free from unknown or missing data entries.

The distribution of each category is visually represented in the pie charts below.

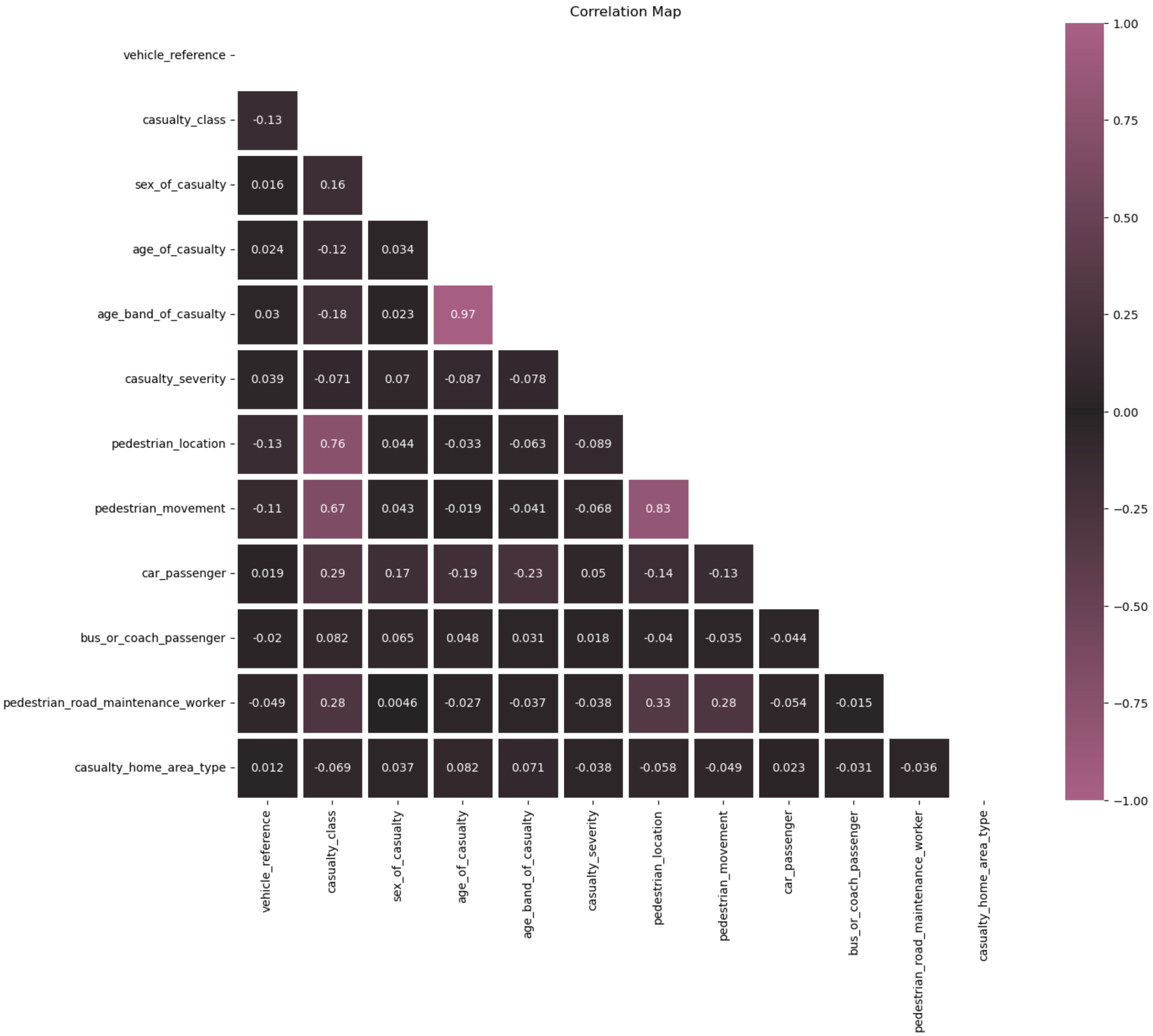


Out of these 55.8K accidents, men account for 34.2K casualties, while women represent a smaller portion with 21.6K casualties. The driver or rider of the vehicle was the primary casualty in more than two-thirds (67.89%) of the incidents. Passengers (17.95%) and pedestrians (14.16%) constitute the remaining casualties as can be seen in the figures below.

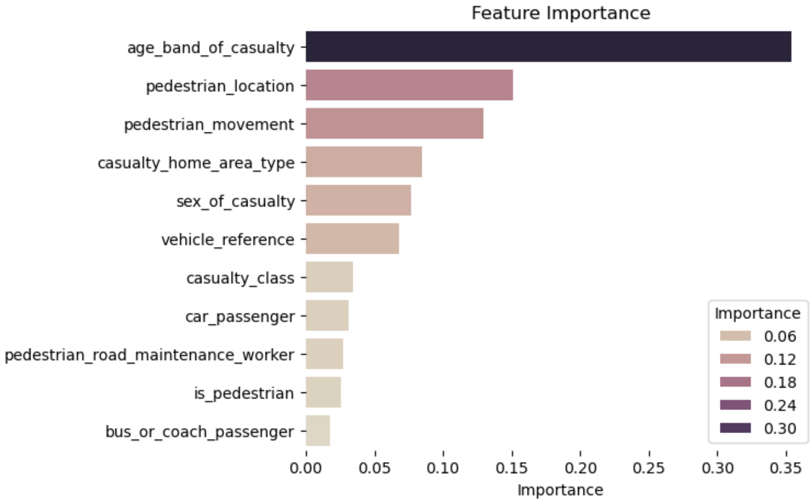


The correlation matrix of accident features reveals a low to moderate correlation among the features. The most significant correlation, excluding the expected high correlation between age and age band, is observed between the pedestrian's movement involved in the accident and their location.

However, our primary interest lies in identifying the features that correlate most significantly with the severity of casualties caused by these accidents. A preliminary correlation analysis indicates that none of the columns exhibit a substantial correlation with the severity of the casualties. The age of the individual injured in the accident and their location on the street (if the injured was a pedestrian) have the highest correlations with casualty severity, albeit with a correlation score of less than 0.1.



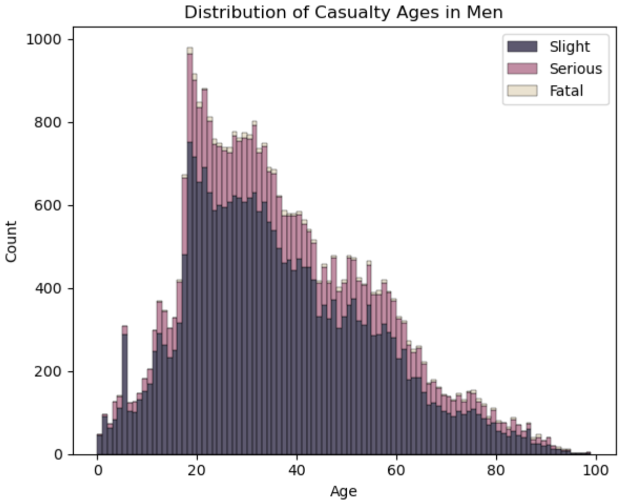
For further analysis, we can utilize a model (in our case, a random forest) to find the most important feature for predicting casualty severity. Adding on to our previous findings, these calculations also put age group and pedestrian location and movement as the most important telling features for prediction of accident severity. The results are as follow:

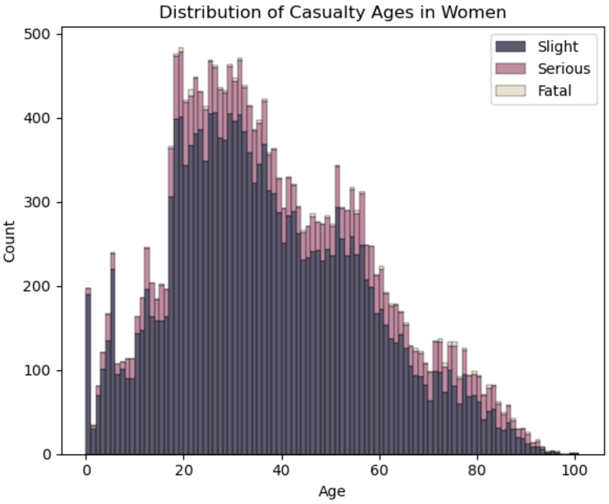


Analyses

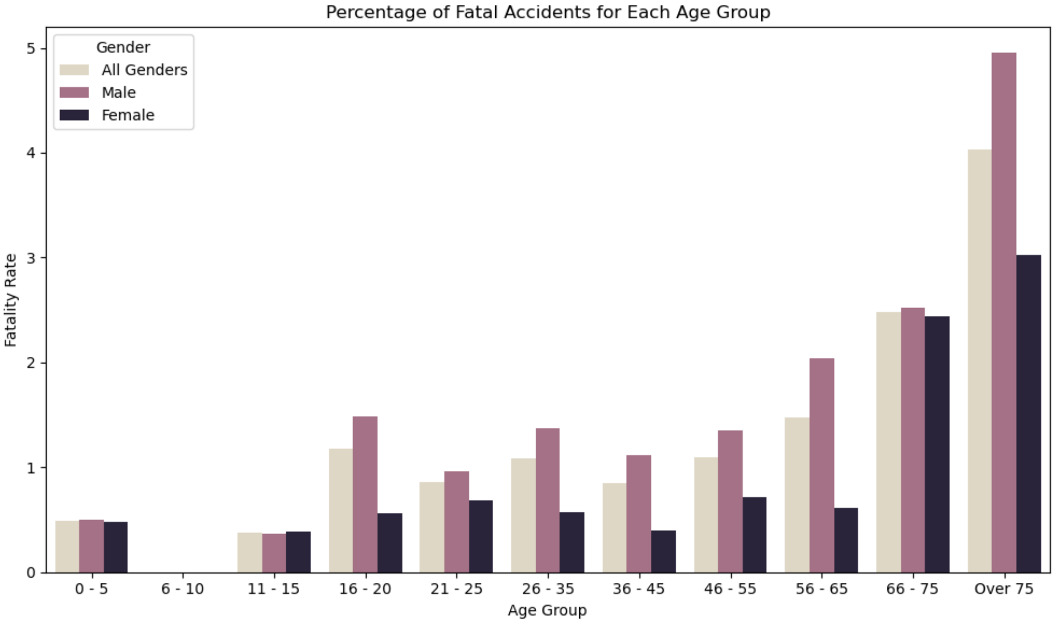
Age Groups

The charts below, show the distribution of casualties in men and women grouped by their age. As it can be seen from the charts, men are twice as much prone to being injured in a car accident compared to women. Furthermore, the data suggests that most casualties involve people in the 20-60 range of age, with men having the highest casualty rate in their early twenties, and women having a distributed peak between ages 20 to 40. It can also be understood from the following charts that slight casualties are three to four times more probable than any serious or fatal casualties.



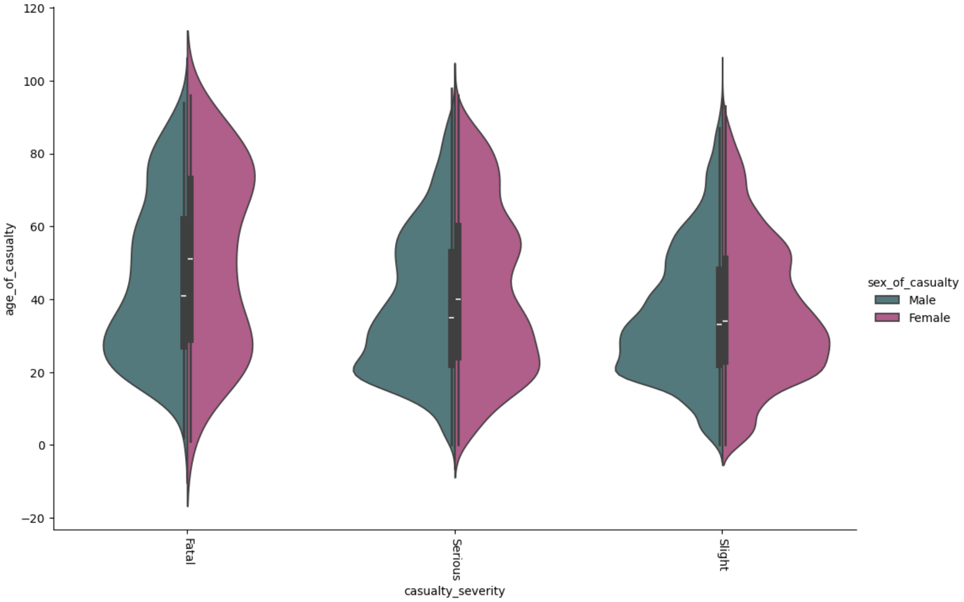


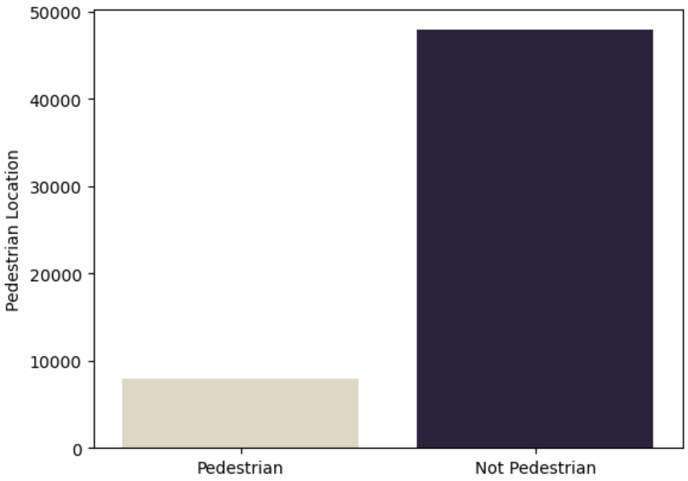
We can delve a bit deeper into the correlation between age groups and the fatality of car accidents. As has been shown below, older age groups are more probable to face a fatal injury when involved in a car accident. Moreover, in almost all age groups, male casualties have higher fatality rate compared to the average, while females of all age groups, except age group of 11 to 15, show a lower fatality rate in accident casualties. All of the data that we have explored up to this point is showing a higher involvement of men in casua`lties of car accidents which can be examined further in another analysis.



Now that we have explored the relationship between the age, gender, and the severity of casualties, we can move on to the second most correlated feature to severity, which is the location and movement of the pedestrian casualty. For the following analyses, we have excluded non-pedestrian casualties from the dataset to improve visual discrimination of differences in plots.

In addition, we can visualize the previous plots better by using a violin plot. As you can see, male casualties are more concentrated age-wise in the 20 to 40 years of age, while female casualties are a bit more evenly distributed along different age groups.

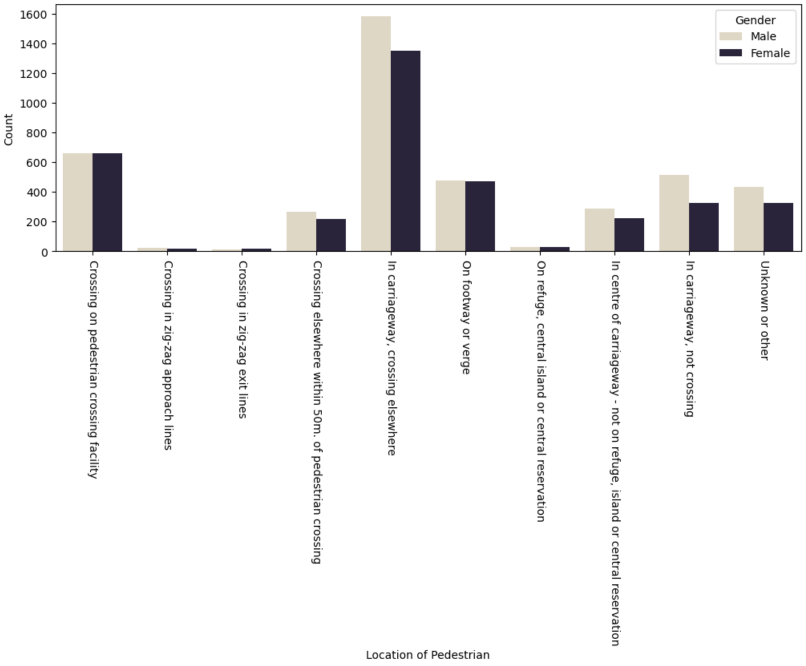




Pedestrian Location

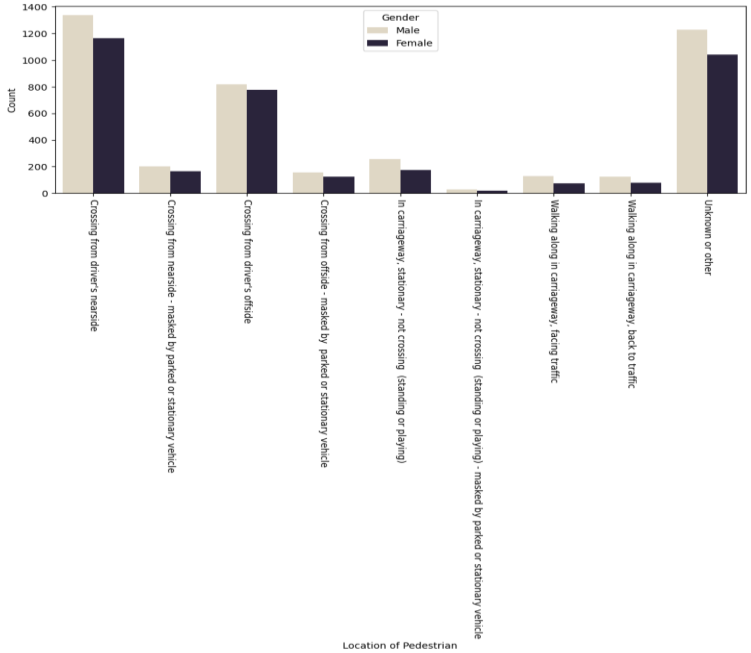
First, we can see that only 7.9K of the casualties are pedestrians while non-pedestrians (drivers, riders, and passengers) make up the other 47.9K casualties.

In addition, as we can see from the following plot, pedestrians crossing the street on non-crossing locations have the highest casualty rate with nearly 3.4k casualties happening to pedestrians. Using crossing facilities lowers the risk down by a considerable factor, as casualties for those using these facilities are about 2.5 times lower at around 1.3k.



Furthermore, the data shows a higher accident chance when pedestrians are crossing from the near side of the driver, meaning there is a higher probability of accidents and casualty when the pedestrian is crossing the street from the middle towards the pavement.

Moreover, we can see that although there are more male casualties in most situations, the difference is not as much as in non-pedestrian cases, which can mean male drivers have a higher casualty count compared to women. This case will be investigated in later analyses.



Conclusion and Future Recommendations

The analysis of the 2022 road accidents dataset reveals significant insights into the factors contributing to road casualties. The data suggests that men, particularly those in their early twenties, are more likely to be involved in accidents. Pedestrians crossing the street at non-crossing locations also represent a significant proportion of casualties. The severity of injuries is most correlated with the age of the casualty and the location of the pedestrian at the time of the accident.

Several measures can be proposed to reduce road casualties based on this analysis:

- **Public Awareness Campaigns:** Targeted campaigns can be developed to educate specific demographics, such as men in their early twenties, about road safety.
- **Infrastructure Improvements:** Enhancing the visibility and accessibility of pedestrian crossings could encourage their use and reduce accidents involving pedestrians.
- **Policy Changes:** Policies could be implemented to improve road safety, such as stricter enforcement of speed limits and harsher penalties for jaywalking pedestrians.
- **Further Research:** More research could be conducted to understand why certain demographics are more prone to accidents. This could involve analyzing additional factors not included in the current dataset, such as traffic flow, city planning, weather conditions or time of day.

These are just suggestions based on the current analysis. Real-world implementation of these measures would require further study and collaboration with local authorities and communities.