**accidents analysis and severity prediction**

**based on road accidents data 2022 in England**

**Ali Habibnia**

**Introduction**

Road accidents are a significant public health concern, causing immense human suffering and economic loss globally. Understanding the factors contributing to these accidents is crucial for developing effective strategies to prevent them. This report aims to analyze data related to traffic accidents, focusing on identifying patterns and key contributing factors. By shedding light on these factors, we can pave the way for targeted interventions and ultimately reduce the frequency and severity of accidents.

**Scope of the Analysis:**

This report will delve into a dataset containing various features associated with traffic accidents in England 2022, including:

* **Casualty characteristics:** Age band, sex, type of casualty (e.g., pedestrian, motorist), and home area type.
* **Accident circumstances:** Pedestrian location and movement at the time of the accident, car and bus passenger involvement, and whether the pedestrian was a road maintenance worker.

**Methodology:**

The analysis will employ various statistical techniques, potentially including:

* **Descriptive statistics:** To summarize the data and identify initial trends and patterns.
* **Visualizations:** To create charts and graphs that effectively communicate key findings.
* **Chi-squared tests:** To assess the association between categorical variables and casualty severity.
* **Training models:** To extract the most important features some models are trained.

**Objectives and Expected Outcomes:**

This report aims to achieve the following objectives:

* Identify the most common types and severities of traffic accidents represented in the data.
* Uncover patterns and trends associated with casualty characteristics and accident circumstances.
* Determine the key factors that contribute to more severe accident outcomes.
* Provide valuable insights and recommendations for developing targeted interventions and safety measures.

**Tools:**

Python programming language and google colab have been used to perform all analysis.

**Significance:**

The findings of this analysis are expected to contribute significantly to understanding the complex dynamics of traffic accidents. By identifying the most influential factors, this report can assist policymakers, road safety organizations, and other stakeholders in developing effective strategies to reduce accidents and their devastating consequences.

**Literature Review on Traffic Accidents**

Analyzing traffic accidents to identify patterns and effective factors is crucial for reducing their occurrence and severity. Here's a review of relevant literature categorized by factor types:

**Human Factors:**

* **Casualty class:** Driver or rider Passenger or Pedestrian.
* **Gender:** Male or female.
* **Pedestrian position and movement:** at which part of the street was the pedestrian and how and in which direction was moving.
* **Age band:** categories of 5 years bands.

**Vehicle Factors:**

* **Vehicle type:** Bicycle, motorcycle, car, bus, …
* **Car passenger:** Front seat or rear seat.

**Additional Considerations:**

* **Socioeconomic factors:** Income inequality, education levels, and access to public transportation can influence accident risks (e.g., WHO, 2020).
* **Casualty home area type:** Urban or Small town or Rural
* **Casualty imd decile:** The IMD decile in Britain refers to the **Index of Multiple Deprivation (IMD)**, a measure used to identify and rank areas of deprivation across England and Wales. It takes into account several factors related to socio-economic disadvantage.

**Description of analysis methods**

In order to check the data and find the insight and make a hypothesis and check the validity of the hypothesis, various steps are carried out which are described below.

**Exploratory Data analysis (EDA)**

EDA is a crucial step in any Data Science project. It involves analyzing and investigating data sets to understand their main characteristics.

Data scientists use EDA to discover patterns, spot anomalies, test hypotheses, and check assumptions.

EDA helps us determine how to best manipulate data sources to get the answers we need.

Before making assumptions, EDA allows us to look at data.

It helps us identify errors, understand data patterns, and detect outliers.

EDA confirms that we are asking the right questions.

Insights drawn from EDA can be used for more sophisticated data analysis or machine learning.

**First insight**

The first results using Pandas is as below:

The dataset has 61352 rows and 20 columns.

The dataframe columns are:

'status', 'accident\_index', 'accident\_year', 'accident\_reference', 'vehicle\_reference', 'casualty\_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\_type', 'casualty\_home\_area\_type', 'casualty\_imd\_decile', 'lsoa\_of\_casualty'

From these columns the below ones are not used in analysis because of many unique values they have :

'accident\_index', 'accident\_year', 'accident\_reference', 'casualty\_reference', 'status', 'vehicle\_reference', 'lsoa\_of\_casualty'

And 'age\_of\_casualty', is not used because the ‘age\_band\_of\_casualty’ has the same role with less unique values.

The dataset has no null values but 7065 rows should be removed because they contain -1 value that means some features are not registered.

After data cleaning the dataset has 54287 rows and 12 columns.

The count of unique values in each column is shown here:

index 54287

casualty\_class 3

sex\_of\_casualty 3

age\_band\_of\_casualty 11

casualty\_severity 3

pedestrian\_location 11

pedestrian\_movement 10

car\_passenger 4

bus\_or\_coach\_passenger 6

pedestrian\_road\_maintenance\_worker 3

casualty\_type 21

casualty\_home\_area\_type 3

casualty\_imd\_decile 10

the casualty\_severity is our target and has 3 levels:

1: Fatal

2: Serious

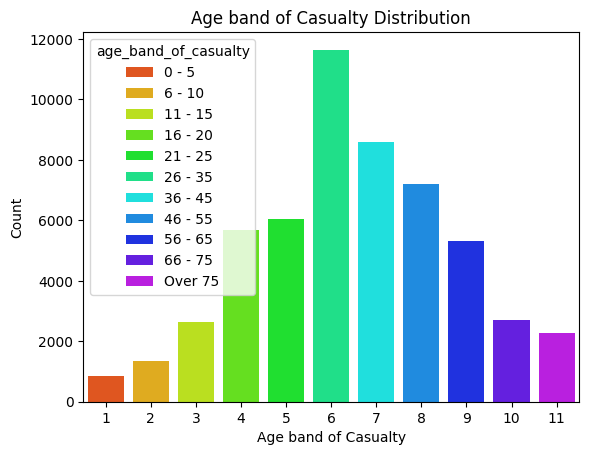
3: Slight

The other 11 columns will be used as features.

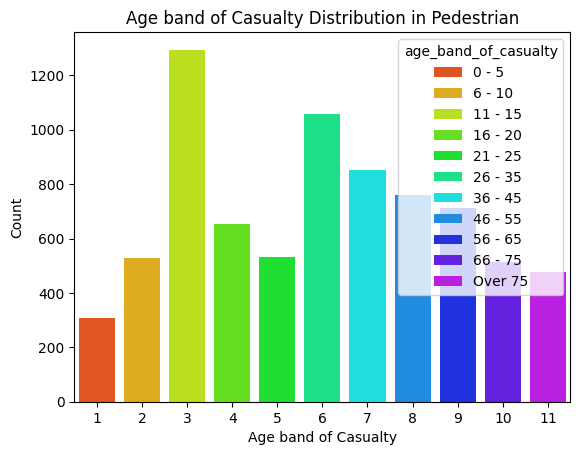
Now it is time to visualize data and find some relations and patterns between variables.

**visualization**

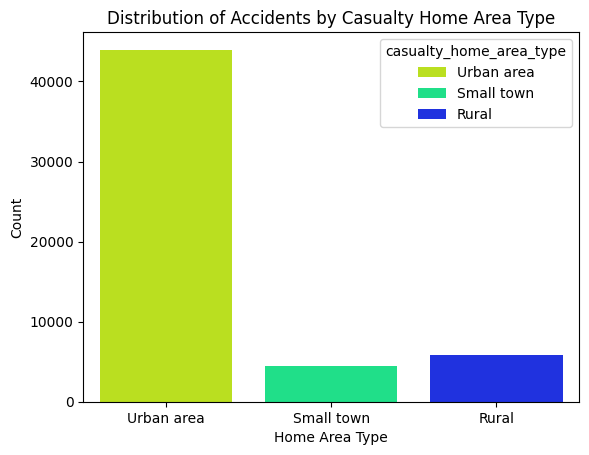
Here, the number of accidents seen in 5-year age ranges is displayed:

****

This chart shows that the most casualties are in age 26 to 35. But chart only for pedestrians is shown here that indicates that in the case of pedestrians, the peak number of casualties is related to 11-15 years old age band. The pedestrian casualties are 7690 persons and other casualties are 46597 pesons.

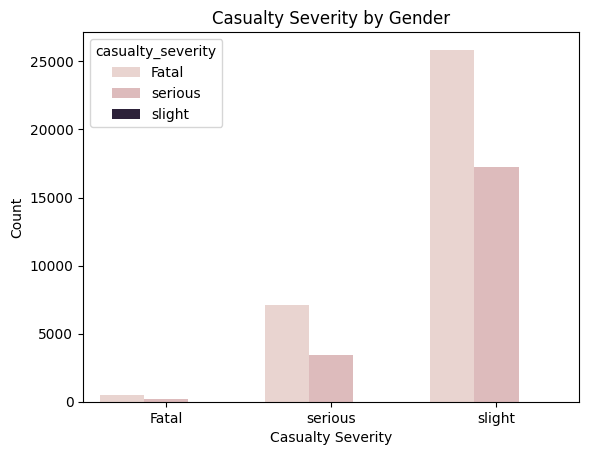


Another feature that shows significant difference between unique values is home area type which are: Urban area, Small town and Rural.

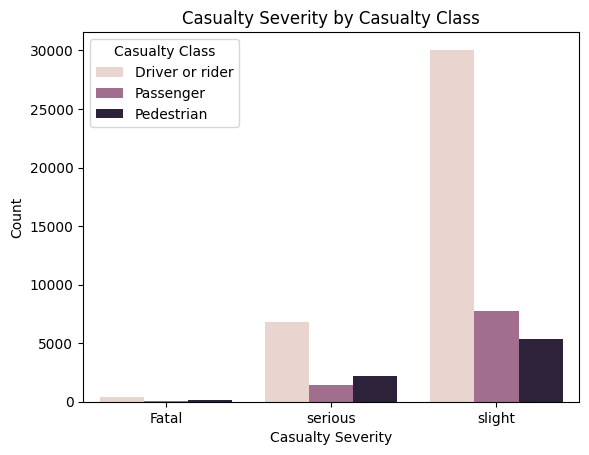


At first glance, it may seem that because the number of accidents is much higher in urban areas, then this feature can be an important factor in the classification of accidents. But it should be noted that until we know what percentage of the entire country's population lives in each of the above three regions in England, it cannot be said that our assumption is correct. In the following sections, we will discuss the importance of this feature.

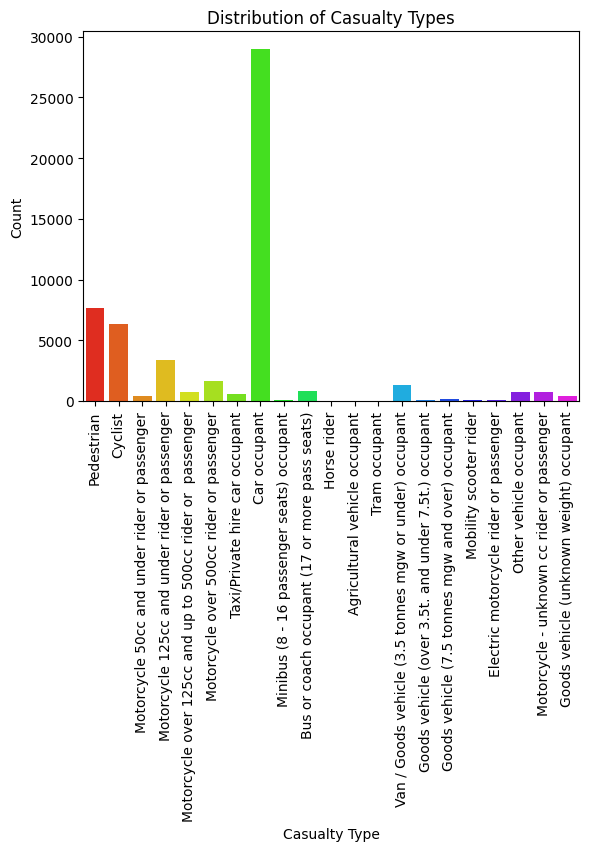
Plotting the relationship between casualty severity and gender reveals that most of the casualties are slight and the number of fatal casualties are very low and this will have a negative impact in our classification process.



Also the relationship between casualty severity and class shows that drivers are significantly more prone to accidents than other passengers and pedestrians.



Another chart that shows the dramatic difference between the types of people affected is shown here:



It is obvious that most casualties are related to personal cars which the exact number of them is 28995 persons and it is more than half of the total casualties.

**Chi Squared test:**

Chi Squared test is a famous test to assess the association between categorical variables and the target. One of the results of this test is P\_Value. In many cases, if this value is less than 0.05, that feature has a significant impact on the target. The result of Chi Squared test of the dataframe is 0 or so close to 0 for all features. Then it is reasonable to use all features in classification.

**Training models**

It is specified in the Chi Squared test that all features should be used in the classification, but this test does not determine the exact amount of their impact on the target. Therefore, the feature and target data are divided into two sets, training and testing. Then, using these sets, we teach several different types of models and determine the most important features of each one along with their effectiveness.

The impact of each feature on the target is shown in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Decision Tree | KNN | SVM | Random Forest |
| casualty\_imd\_decile | 0.269669 | 0.03199 | 0.00017 | 0.30609 |
| age\_band\_of\_casualty | 0.212006 | 0.03394 | 0.00120 | 0.24749 |
| casualty\_type | 0.204407 | 0.04783 | 0.00302 | 0.16775 |
| pedestrian\_location | 0.088156 | 0.01710 | 0.00109 | 0.08224 |
| pedestrian\_movement | 0.062375 | 0.01452 | -0.00013 | 0.06100 |
| casualty\_home\_area\_type | 0.048836 | 0.01202 | 0.00083 | 0.05434 |
| sex\_of\_casualty | 0.043737 | 0.00732 | 0.00037 | 0.01733 |
| casualty\_class | 0.025743 | 0.00789 | 0.00022 | 0.01733 |
| pedestrian\_road maintenance\_worker | 0.025162 | 0.00186 | 0.00006 | 0.01302 |
| car\_passenger | 0.013657 | 0.00264 | 0.00004 | 0.01380 |
| bus\_or\_coach\_passenger | 0.006251 | 0.00033 | -0.00020 | 0.00528 |

According to the above table, it can be evaluated that the three most influential features on the severity of accidents are as follows:

casualty\_type, age\_band\_of\_casualty and casualty\_imd\_decile

And the three features that have the least impact on the severity of accidents are:

bus\_or\_coach\_passenger, car\_passenger, pedestrian\_road maintenance\_worker

**Analysing most effecting features**

Now we can examine the 3 most effective features and see how each of them affects the severity of the accident. The 3 most occurred cases in each of the 3 features are shown in the below tables:

|  |  |
| --- | --- |
| **Casualty\_type** | **counts** |
| Car occupant | 28995 |
| Pedestrian | 7690 |
| Cyclist | 6375 |

|  |  |
| --- | --- |
| **age\_band\_of\_casualty** | **counts** |
| 26 - 35 | 11644 |
| 36 - 45 | 8575 |
| 46 - 55 | 7208 |

|  |  |
| --- | --- |
| **casualty\_imd\_decile** | **counts** |
| More deprived 10-20% | 6946 |
| Most deprived 10% | 6766 |
| More deprived 20-30% | 6709 |

**Actions to be taken:**

potential actions that governments can take:

**Casualty Types:**

**Car Occupants:**

Given the high count of car occupants involved in accidents, governments should focus on improving vehicle safety standards. This includes enforcing seatbelt usage, promoting airbag installation, and encouraging safer driving practices.

**Pedestrians:**

Pedestrian safety is crucial. Governments can invest in pedestrian-friendly infrastructure such as well-marked crosswalks, pedestrian bridges, and traffic calming measures near schools and residential areas.

Most pedestrians are in age band 11 – 15. These are teenagers who are not in the outdoor environment with their parents like children. Therefore, more care should be taken around schools and children's playgrounds.

**Cyclists:**

To reduce cyclist accidents, governments should create dedicated bike lanes, raise awareness about sharing the road, and enforce helmet usage.

**Age Bands of Casualty:**

26-35, 36-45, and 46-55 Age Groups:

These age groups are significantly affected by accidents. Governments should:

**Educate:**

Conduct awareness campaigns targeting safe driving practices, especially for this working-age population.

**Enforce:**

Strengthen traffic laws and penalties for reckless driving.

**Training:**

Provide defensive driving courses and refreshers for older drivers.

**Casualty IMD Decile (Deprivation Level):**

More Deprived 10-20% and Most Deprived 10%:

These groups face higher accident risks. Actions include:

**Infrastructure:**

Improve road conditions in deprived areas to minimize hazards.

**Education:**

Educate residents about road safety, emphasizing pedestrian awareness.

**Community Engagement:**

Involve local communities in safety initiatives.

**Suggestions about data gathering:**

Some parameters are not registered in accidents which can be important. Some of them are mentioned here:

Weather conditions

Daylight

Road maintenance

Also it can be beneficial if 2 dataset have been made for pedestrians and non pedestrians.

Dataset source:

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