# Chapter4

## 4.1 Introduction

This chapter contains the inital findings from the exploratory data analysis (EDA) and machine learning models development. Our research examines the prediction of traffic accident for Malaysia, using a number of datasets. The datasets are as follow; population data, vehicles registered data, deaths and injuries by traffic accident data, accidents by vehicle data. Through the data we aim to identify patterns and forecast future congestion and accident.

## 4.2 Exploratory Data Analysis(EDA)

Visualizations and Descriptive Statistics

During EDA, descriptive statics and visualizations were carried out. Below, the visualizations and descriptive statistics are provided:

| | Condition | Mean | StdDev | Min | Max | Total |
|---|---|---|---|---|---|---|
| 0 | Deaths | 726.566667 | 1079.700399 | 8 | 4485 | 130782 |
| 1 | Injuries | 3078.300000 | 5893.583460 | 27 | 35727 | 554094 |

The descriptive statistics of traffic accident deaths and injuries show that the mean number of deaths is 726.57 and the standard deviation shows that the number of deaths in some years fluctuates greatly. The average number of injuries is much higher than the number of deaths, about 4-5 times, which indicates that although most accidents are not fatal, the pressure on social medical resources is still high.

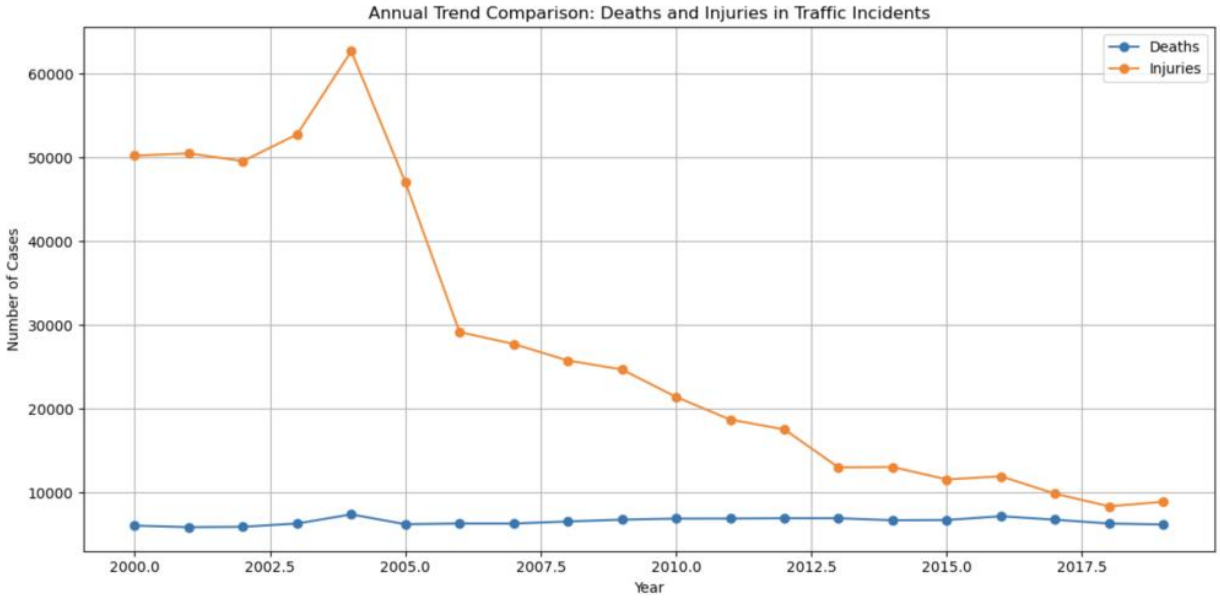
| condition | Road User Type | Total Deaths | Total Injuries | Fatality Rate |
|---|---|---|---|---|
| 0 | Cyclists | 3942 | 17703 | 0.182121 |
| 1 | Lorry drivers | 8784 | 17712 | 0.331522 |
| 2 | Motorcar drivers | 12945 | 53284 | 0.195458 |
| 3 | Motorcyclists | 51535 | 290376 | 0.150726 |
| 4 | Others | 3005 | 12536 | 0.193360 |
| 5 | Passengers | 12734 | 50418 | 0.201640 |
| 6 | Pedestrians | 9187 | 36954 | 0.199107 |
| 7 | Pillion riders | 6901 | 38894 | 0.150693 |
| 8 | Taxi/Bus drivers | 21749 | 36217 | 0.375203 |

Accident profile of different road users: Motorcycle drivers are the most severely affected group in traffic accidents, with significantly higher number of deaths and injuries than other road users, and a higher mortality rate than the average level of some road users. Pedestrian and bicycle users also have higher death rates, indicating that these two groups are vulnerable groups in traffic safety and need special attention

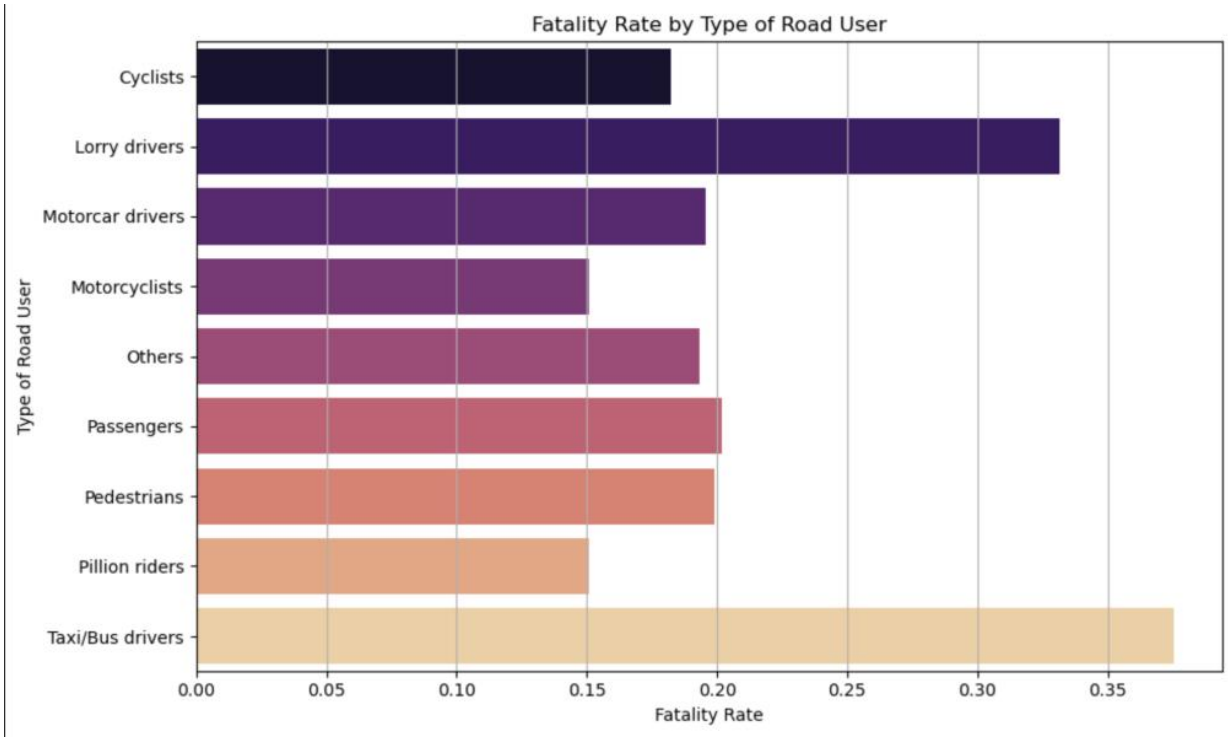
| | Vehicle Type | Total | Mean | StdDev | Min | Max |
|---|---|---|---|---|---|---|
| 0 | Bicycle | 18619 | 930.95 | 747.667353 | 201 | 2354 |
| 1 | Bus | 5324 | 266.20 | 230.089594 | 46 | 868 |
| 2 | Jeep | 7966 | 398.30 | 138.421211 | 209 | 615 |
| 3 | Motorcar | 91837 | 4591.85 | 1695.130185 | 2910 | 8340 |
| 4 | Motorcycle | 385534 | 19276.70 | 8649.888950 | 8782 | 35727 |
| 5 | Others | 15720 | 786.00 | 1250.267129 | 130 | 4395 |
| 6 | Pedestrian | 34856 | 1742.80 | 1243.194849 | 0 | 4049 |
| 7 | Trailer/Lorry | 12312 | 615.60 | 286.841310 | 131 | 1227 |
| 8 | Vans | 8747 | 437.35 | 327.399044 | 170 | 1156 |

Statistics of accidents related to different types of vehicles: The total number of accidents related to motorcycles is the largest, reaching 385,534, accounting for a significantly higher proportion than other vehicle types, and the mean and standard deviation of accidents are large, indicating that they not only have a high frequency of accidents, but also fluctuate significantly between different years. This was

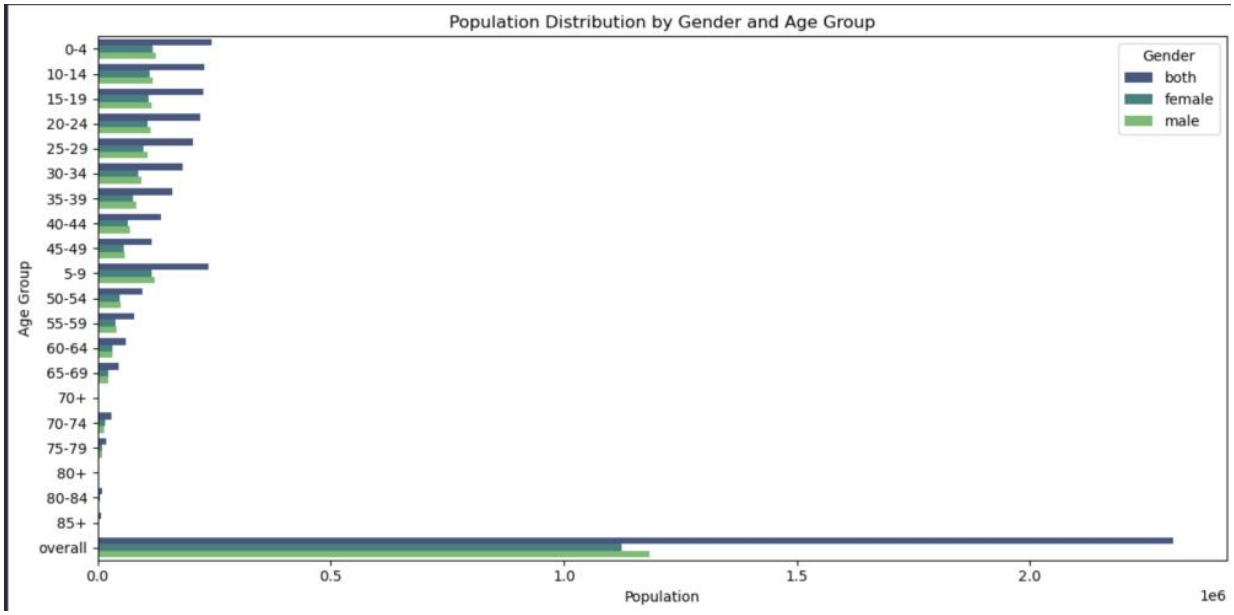
followed by cars with a total of 91,837 accidents, showing their important role in traffic accidents. The total number of accidents in other vehicles, such as bicycles and buses, is relatively low, but the severity of individual accidents may need to be further explored.



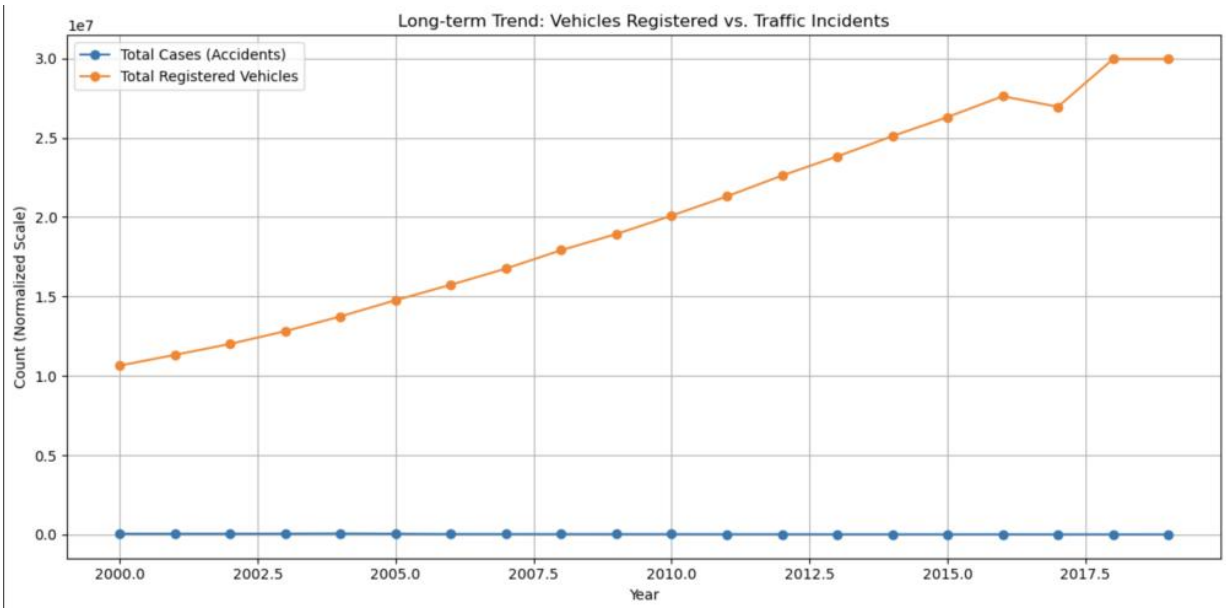

**Comparison of traffic accident trends:** From the data from 2000 to 2021, the number of deaths and injuries in traffic accidents shows an overall downward trend, especially the decline in fatalities, indicating that improvements in traffic management policies are having an effect. In some years, however, there have been short-term fluctuations in the number of deaths and injuries. For example, the number of deaths in some years is significantly higher than average, which can be related to bad weather, road construction, or increased traffic during holidays. These unusual fluctuations remind us that while the overall trend is positive, more stringent traffic control measures are still needed for specific hours.



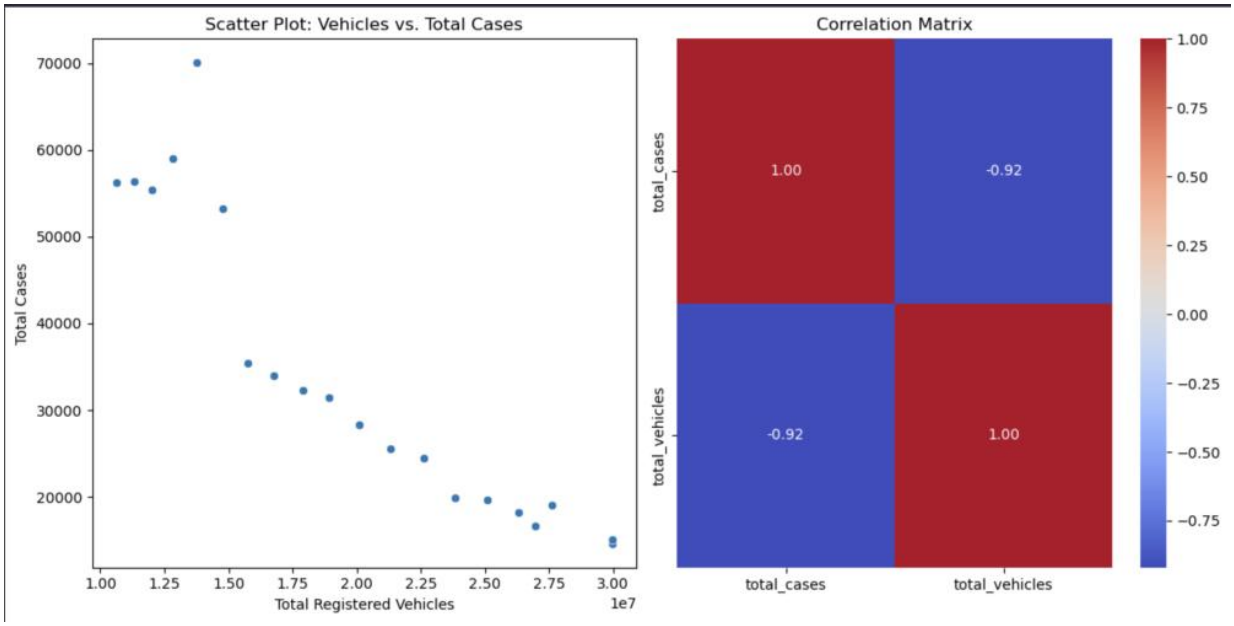

**Death rate among different road users:** Taxi drivers have the highest death rate among all types of road users (about 0.35). This may be related to the long hours taxi drivers drive in cities, the complex traffic environment they face, and the higher work pressure. Second, the significantly higher fatality rates for motorcycles, passengers, and pedestrians than for other groups may be related to their lack of physical protection (as compared to cars, for example), while the higher fatality rates for pedestrians may reflect the greater vulnerability of pedestrians in vehicle-pedestrian collisions.



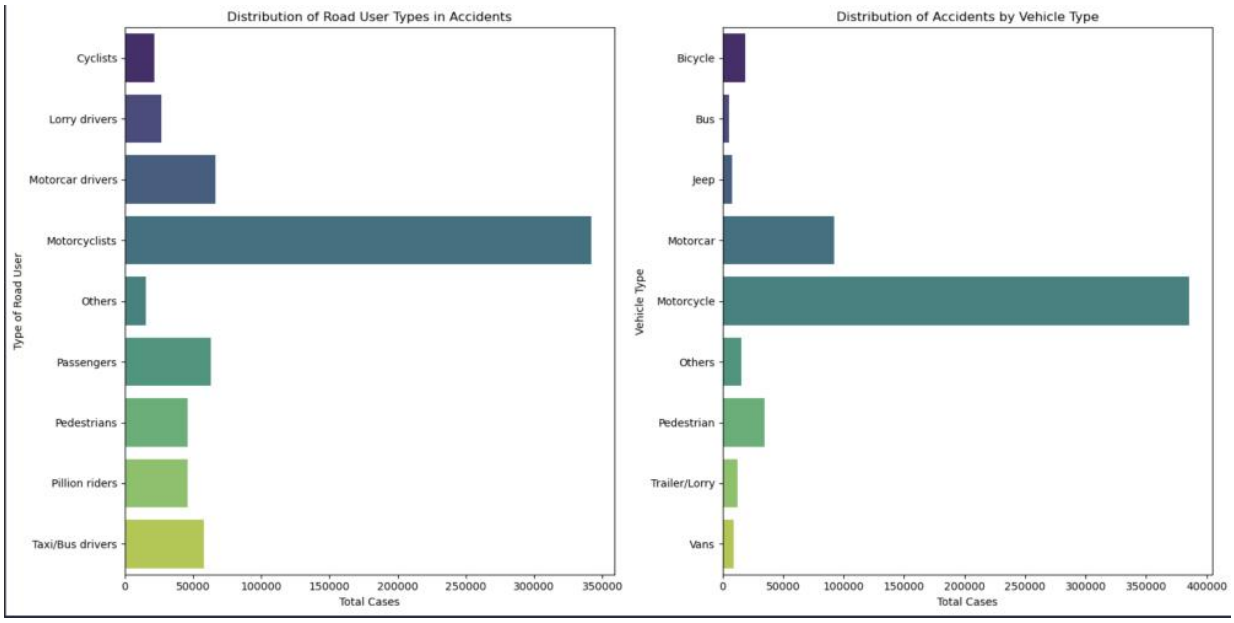
**Distribution of accidents by sex and age group:** It can be seen that the proportion of males in traffic accidents is significantly higher than that of females, especially in the young adults aged 20-40. Male drivers are more likely to engage in high-risk driving behaviors, such as speeding, illegal lane changes and drunk driving. In addition, young drivers (especially those aged 20-30) and the elderly (aged 60 and above) are two groups that need attention. Young people are more likely to be involved in accidents due to inexperience driving and risky behavior, while the elderly are a high-risk group due to reduced physical responsiveness.




**Long-term trends in the number of motor vehicle registrations and the total number of accidents:** From 2000 to 2021, the number of motor vehicle registrations has steadily increased, while the total number of traffic accidents has declined. This suggests that despite the increasing number of vehicles on the road, traffic management policies and technological improvements may have effectively reduced the risk of accidents per vehicle.




With the increase of the total number of registered vehicles, the total number of traffic accidents shows a significant downward trend and a negative correlation. This may be because the areas with a higher number of vehicle registrations are usually economically developed areas, and these areas may have better traffic management systems and higher quality road infrastructure, which can effectively reduce traffic accidents. This phenomenon highlights the importance and complexity of traffic safety: the presence of more vehicles does not necessarily lead to more accidents, but may lead to fewer accidents due to better traffic management, driver awareness and infrastructure.




Accident types and vehicle distribution: Motorcycle accidents account for the highest proportion of all accident types, and the total number of accidents far exceeds that of cars, buses and other types of

vehicles. The vulnerability of motorcycle riders and the high incidence of accidents highlight the need for more rigorous safety education and protection for this group. In addition, the total number of pedestrian accidents is also high, indicating that urban and rural areas may need more pedestrian-priority facilities, such as signals and sidewalks.

## 4.3 initial machine learning result

The AUC values based on the random forest model showed some predictive power, but the overall accuracy was about 9.7% lower, and the classification reports showed accuracy, recall, and F1 scores, all in the 8-11% range. This indicates that the model may have insufficient features or unbalanced categories. The feature importance analysis reveals several key influencing factors, among which the proportion of motorcycle riders is the most important feature, followed by peak hour and night time characteristics. In addition, the total annual motor vehicle registrations also show significant predictive power. The results of the model further show that pedestrian accidents account for a certain proportion of high-risk characteristics, suggesting that inadequate pedestrian safety facilities may be a potential cause of accident risk. Although the overall performance of the model is not high, its feature importance analysis provides important insights, such as that motorcycles and walkers are the main groups affected by traffic accidents, and peak hours and night times are high-risk times. These results suggest that policy makers need to focus on safety education and management of the motorcycle population, as well as strengthening night lighting and traffic management during peak hours. This preliminary result provides the direction for the next step of feature optimization and model improvement, and also provides data support for the formulation of traffic safety policies.