LUAN TRAN - BS22DSY024

FATAL CRASH ROAD PREDICTION

AUSTRALIAN CRASH ROAD ACCIDENT 1989-2021 - **KAGGLE**

The Australian Road Deaths Database provides basic details of road transport crash fatalities in Australia as reported by the police each month to the State and Territory road safety authorities. Road deaths from recent months are preliminary and the series is subject to revision

01

INTRODUCTION

Road accidents are a major cause of deaths and injuries worldwide, with Australia being no exception. This report focuses on the Australian Fatal Road Accident dataset, which includes information on fatal road accidents that occurred in Australia between 1989 and 2021.

DATASET

Australian Crash Road Accident 1989-2021

https://www.kaggle.com/datasets/deepcontractor/australian-fatal-car-accident-data-19892021?resource=download

01

PROBLEM STATEMENT

The problem statement is to analyze and build a model to predict the likelihood of fatal crashes in Australia based on various factors such as road user type, time of day, day of week, speed limit, and age group.

The goal is to identify the most important factors that contribute to fatal crashes and to develop a model that can help to prevent or reduce the number of fatal crashes on Australian roads. The analysis will provide insights for policymakers, law enforcement agencies, and other stakeholders to take necessary actions to improve road safety and save lives.

METHODOLOGY

- The dataset was first cleaned and preprocessed to remove missing values and ensure consistency across attributes.
- The exploratory data analysis was conducted to understand the distribution and correlation of attributes.
- The logistic regression model was then built to predict the likelihood of a fatal crash occurring during the day or night based on various attributes.

02 DATA DESCRIPTION

The data consists of 23 attributes which are :

1. 'Crash ID', 2. 'State', 3. 'Month', 4. 'Year', 5. 'Dayweek', 6. 'Time', 7. 'Crash Type', 8. 'Bus Involvement', 9. 'Heavy Rigid Truck Involvement', 10. 'Articulated Truck Involvement', 11. 'Speed Limit', 12. 'Road User', 13. 'Gender', 14. 'Age', 15. 'National Remoteness Areas', 16. 'SA4 Name 2016', 17. 'National LGA Name 2017', 18. 'National Road Type', 19. 'Christmas Period', 20. 'Easter Period', 21. 'Age Group', 22. 'Day of week', 23. 'Time of day'

Some attributes that needed to be acknowledged :

- Crash ID: A unique identifier for each traffic crash.
- State: The state or territory where the traffic crash occurred.
- Month: The month in which the traffic crash occurred.
- Year: The year in which the traffic crash occurred.
- Dayweek: The day of the week on which the traffic crash occurred.
- Time: The time of day at which the traffic crash occurred.
- Crash Type: The type of traffic crash (e.g., head-on collision, rear-end collision, etc.).
- Speed Limit: The posted speed limit at the location where the traffic crash occurred.
- Road User: The type of road user involved in the traffic crash (e.g., driver, passenger, pedestrian, etc.).
- Gender: The gender of the road user involved in the traffic crash.
- Age: The age of the road user involved in the traffic crash.
- Age Group: A grouping of ages into categories for analysis purposes.
- Day of week: The day of the week on which the traffic crash occurred, as a numerical value.
- Time of day: The time of day at which the traffic crash occurred, as a numerical value.

03

DATA PREPROCESSING

Drop some uninformative and redundant columns

We decided to remove these columns since there are a lot of missing data:
"National Remoteness Areas",
"SA4 Name 2016"
"National LGA Name 2017"
"National Road Type"
"Heavy Rigid Truck Involvement"

Deal with missing value :

For other less missing data columns, we use method called SimpleImputer with the strategy - most_frequent'.

This method is used to replace missing values in the DataFrame with the most frequent value of the corresponding column.

Number of Null value

Crash ID	0
State	0
Month	0
Year	0
Dayweek	0
Time	40
Crash Type	0
Bus Involvement	22
Heavy Rigid Truck Involvement	20515
Articulated Truck Involvement	22
Speed Limit	702
Road User	0
Gender	27
Age	0
National Remoteness Areas	45965
SA4 Name 2016	45951
National LGA Name 2017	45950
National Road Type	45966
Christmas Period	0
Easter Period	0
Age Group	90
Day of week	0
Time of day	0

03

DATA PREPROCESSING

Transform data type

Time data format

Converts the Time column from a string with colon-separated values to a float64 data type, so that it can be used for further analysis.
For example: 12:30 -> 12.30

Deal with numeric attributes

Transform all possible columns to numerical data types includes : Year, Month, Age, Crash ID and Speed limit

Data before transforming

```
Crash ID                          object
State                             object
Month                             object
Year                              object
Dayweek                           object
Time                              object
Crash Type                        object
Bus Involvement                   object
Articulated Truck Involvement     object
Speed Limit                       object
Road User                         object
Gender                            object
Age                               object
Christmas Period                  object
Easter Period                     object
Age Group                         object
Day of week                       object
Time of day                       object
```

Columns with object data type are categorical attributes, or else are numeric attributes

03

DATA PREPROCESSING

Duplicated rows

Check for duplicate data

```
Number of duplicated rows: 157
```

Then drop any duplicated row in the data

Outliers

For any numerical features, check if there is any outlier
It shows that 'Speed Limit' column has 683 outliers so we are going to drop these rows.

```
Outliers in Crash ID: 0
Outliers in Month: 0
Outliers in Year: 0
Outliers in Time: 0
Outliers in Speed Limit: 683
Outliers in Age: 0
```

Categorical and Numeric Attributes

After cleaning the data, we have :

- **Categorical columns:** State, Dayweek, Crash Type, Bus Involvment, Articualted Truck Involvement, Road User, Gender, Christmas Period, Easter Period, Age Group, Day of week, Time of day
- **Numeric columns:** Crash ID, Month, Year, Time, Speed Limit, Age

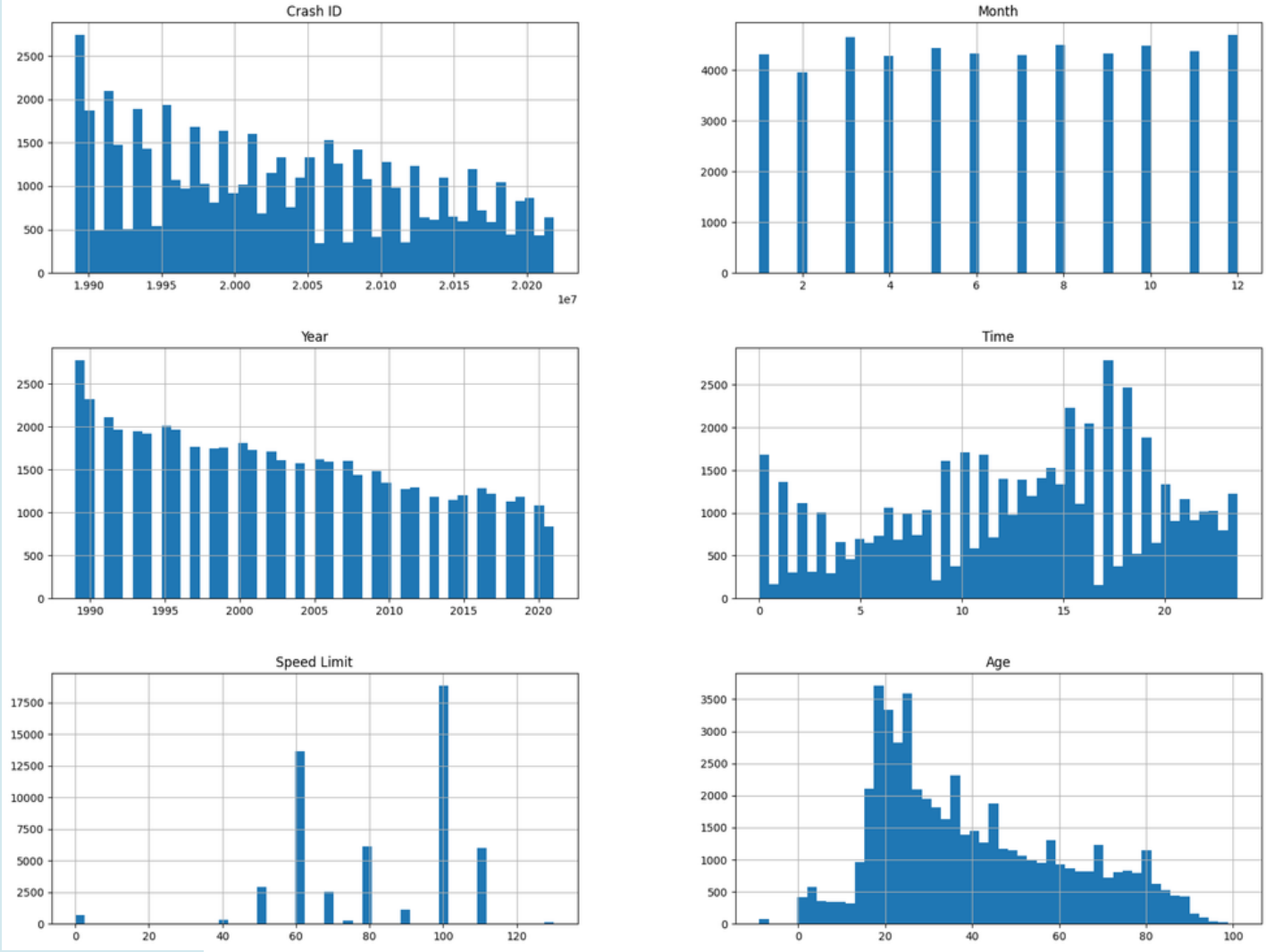
04

EXPLORATION DATA ANALYSIS (EDA)

The next step was to perform EDA to gain a better understanding of the dataset. We used various visualization techniques to explore the distribution of the different features and identify any patterns or correlations.

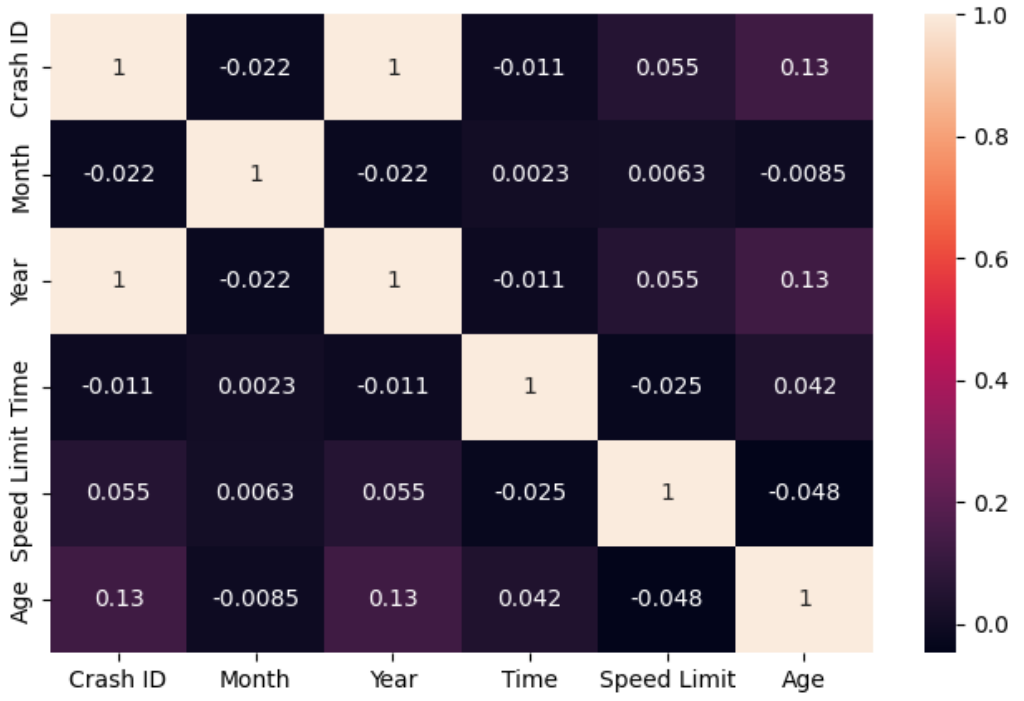
Plot 1 - First, for numeric columns, we visualizing the distribution of each variable

The number of crashes is distributed fairly evenly by month. The majority of accidents occurred between 15 and 20 o'clock during the day. The speed limit of 100 has the highest number of fatal accidents, and the most common age group involved in a crash accident is 17-25.



04

Plot 2 - For numeric variable, use a heatmap to visualize the correlation between different variables in a dataset

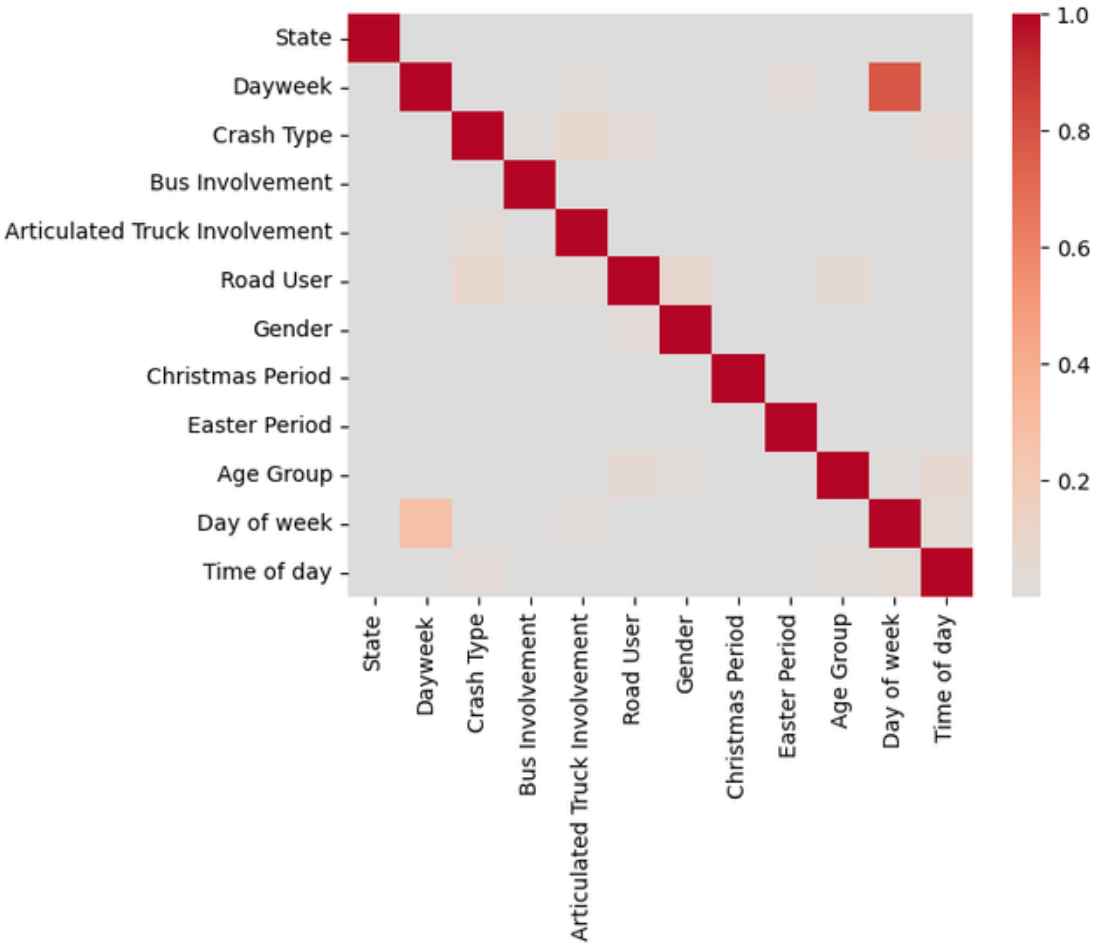


	Crash ID	Month	Year	Time	Speed Limit	Age
Crash ID	1.000000	-0.021941	0.999845	-0.011111	0.055280	0.125267
Month	-0.021941	1.000000	-0.022164	0.002311	0.006282	-0.008523
Year	0.999845	-0.022164	1.000000	-0.011491	0.055419	0.126468
Time	-0.011111	0.002311	-0.011491	1.000000	-0.025467	0.041959
Speed Limit	0.055280	0.006282	0.055419	-0.025467	1.000000	-0.047768
Age	0.125267	-0.008523	0.126468	0.041959	-0.047768	1.000000

- 'Age' and 'Year' are strongly positively correlated (0.126468), indicating that the number of crashes involving older drivers has increased over time.
- 'Year' and 'Month' are weakly negatively correlated (-0.022164), indicating that the number of crashes may be slightly lower in certain months of the year.
- There is no strong correlation between 'Time' and any of the other variables, suggesting that the time of day may not be a strong predictor of crashes.

04

Plot 3 - For categorical variable, use a Theil's U matrix to visualize the correlation between different variables in a dataset

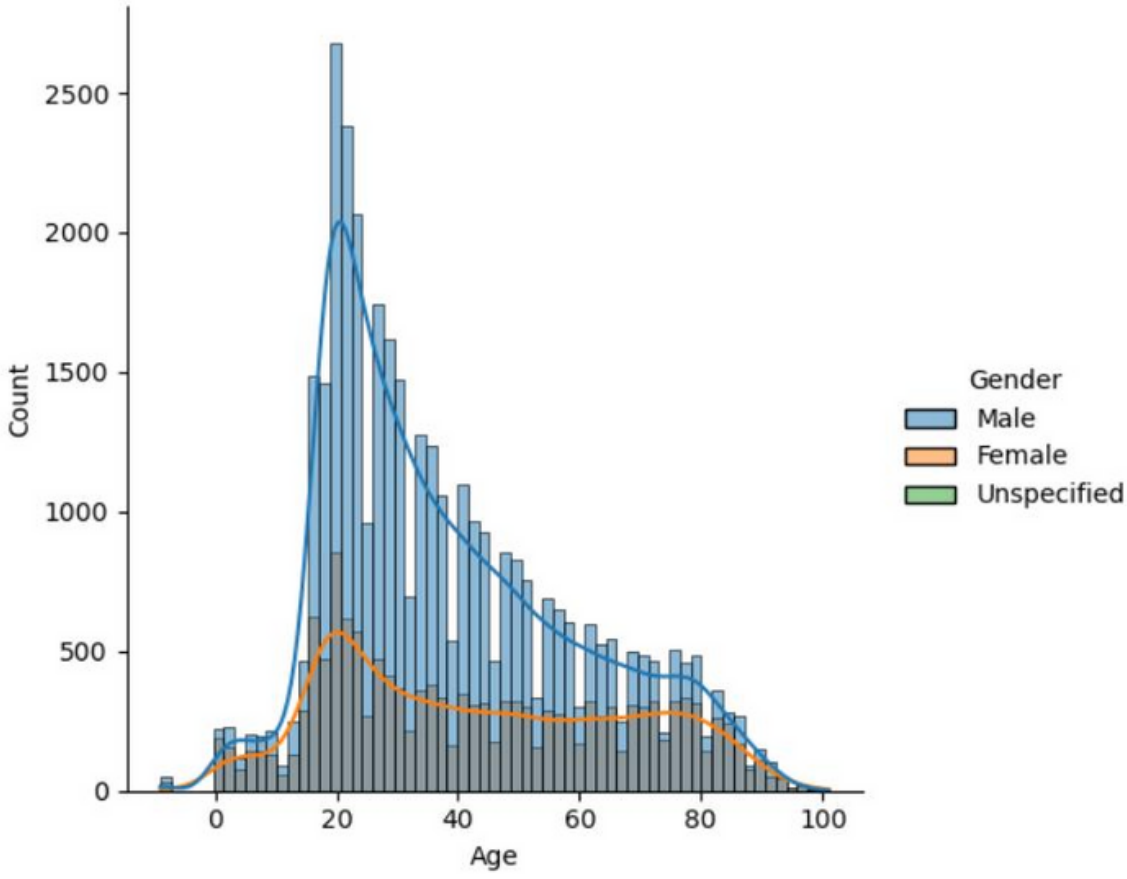


Most of the associations in the Theil's U matrix are nearly 0, it means that the variables are independent of each other and there is little to no predictive power between them

04

Plot 4 - The distribution of Age with respect to Gender

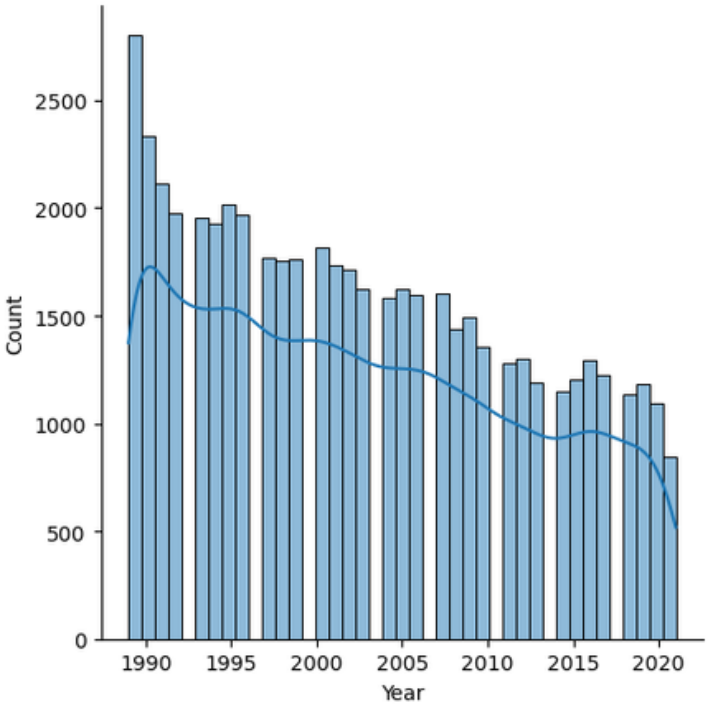
The plot uses a kernel density estimation (KDE) to estimate the probability density function of the variable. It clearly shows that the majority of the crashes involve male, with the majority of the crashes involving individuals in their 20s and 30s.



04

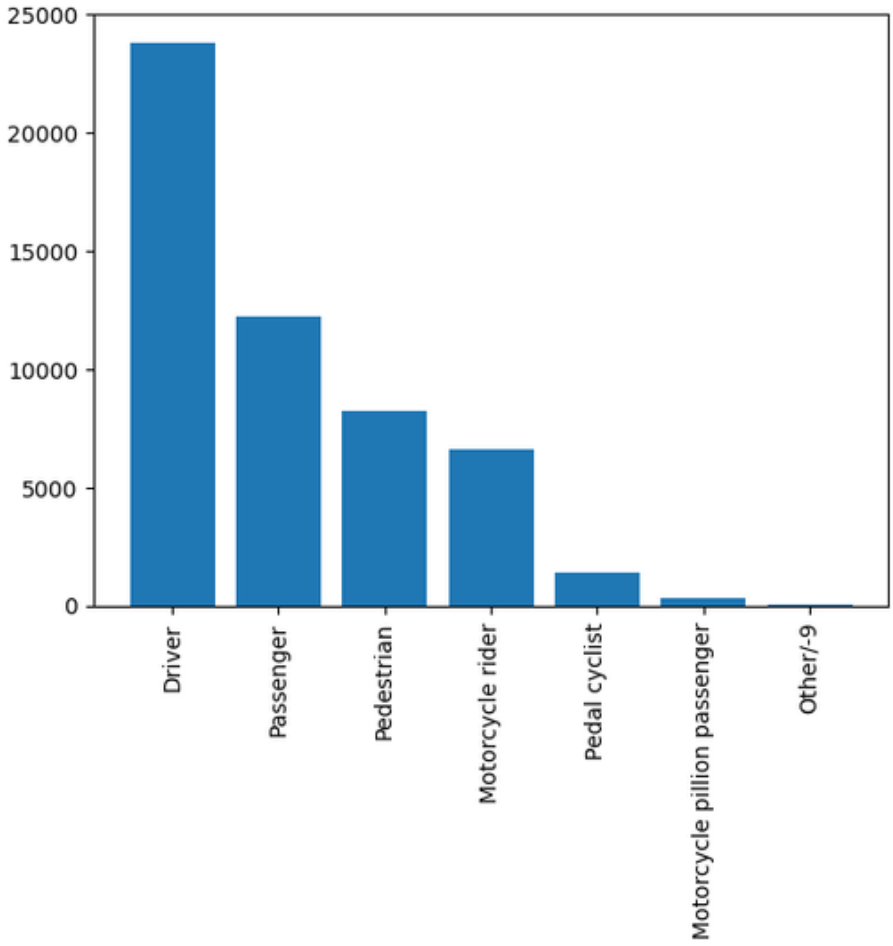
Plot 5 - shows the distribution of Year, which can help in visualizing the frequency of accidents over time.

The number of crashes has been decreasing over the years



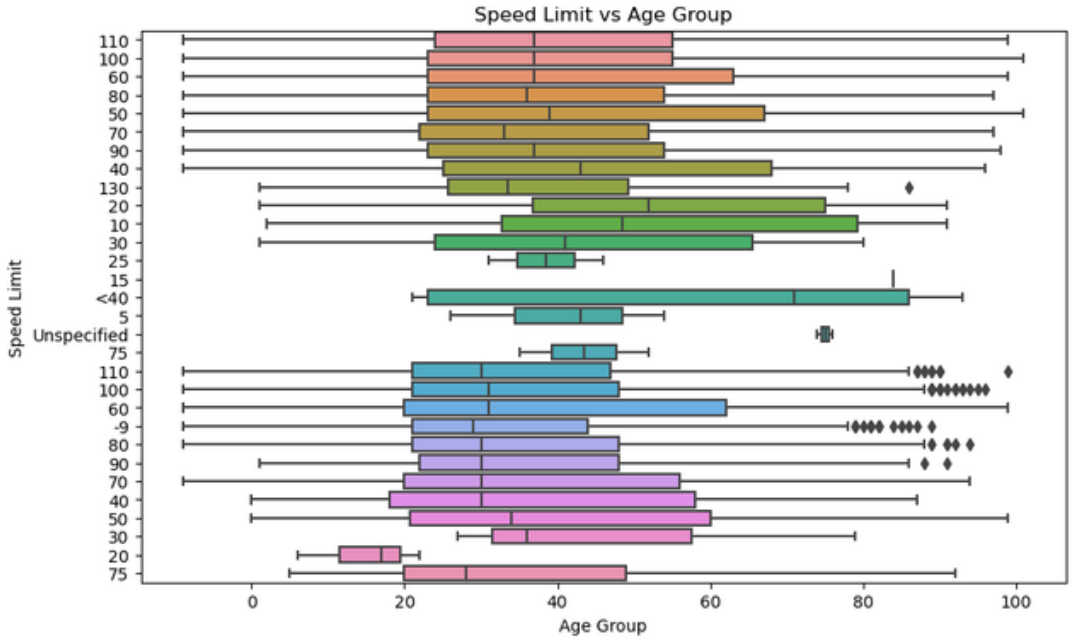
Plot 6 - shows the number of crashes for each type of Road User.

The most frequent road user type is 'Driver'.



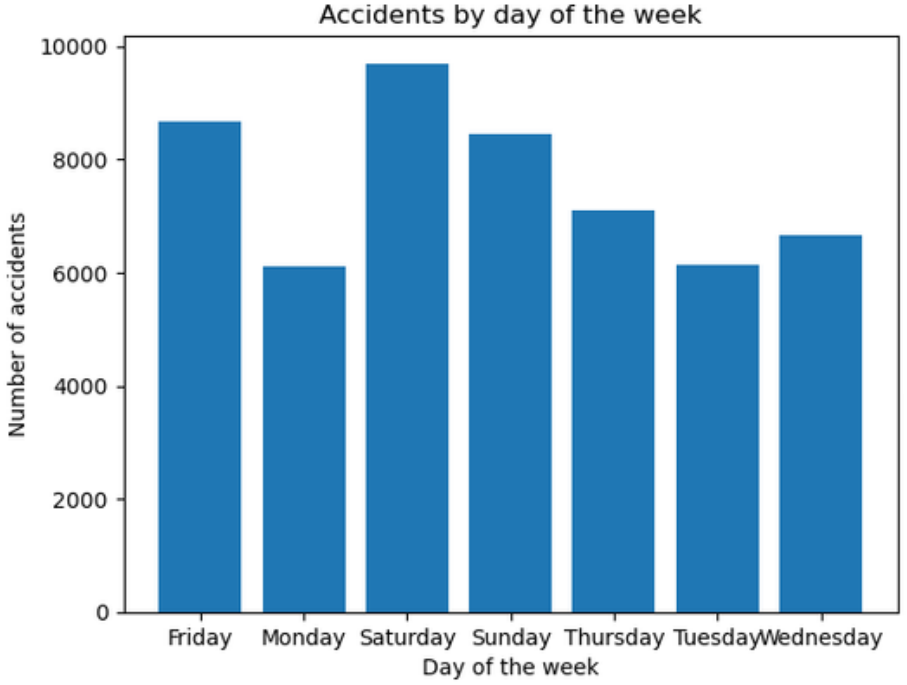
Plot 7 - Box plot - the distribution of Speed Limit for Age Group

Older drivers tend to drive at lower speeds, while younger drivers tend to drive at higher speeds.



Plot 8 - Bar chart shows the distribution of the number of accidents by day of the week.

Friday and Saturday have the highest number of crashes, accounting for more than 30% of all crashes in the dataset. This suggests that weekends may be more dangerous for driving, and more attention may need to be paid to road safety on these days.



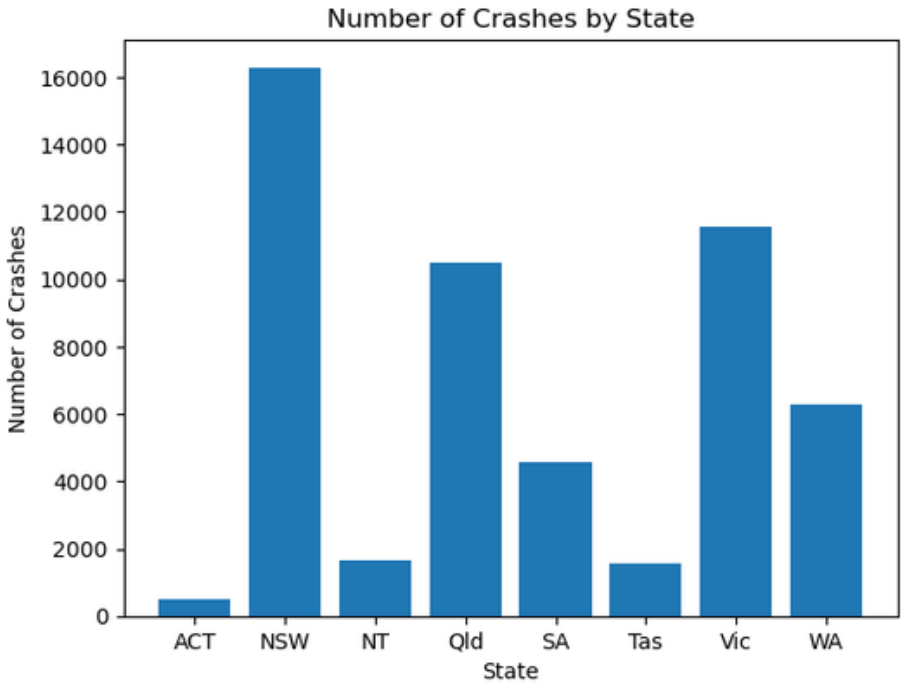
% of Crashes by Day of Week:

Friday	16.397631
Monday	11.558768
Saturday	18.348693
Sunday	16.009689
Thursday	13.447382
Tuesday	11.628787
Wednesday	12.609049

04

Plot 9 - bar chart represents number of crashes by State

The majority of crashes occur in the state of NSW, with Victoria and Qld having the second and third highest number of crashes, respectively

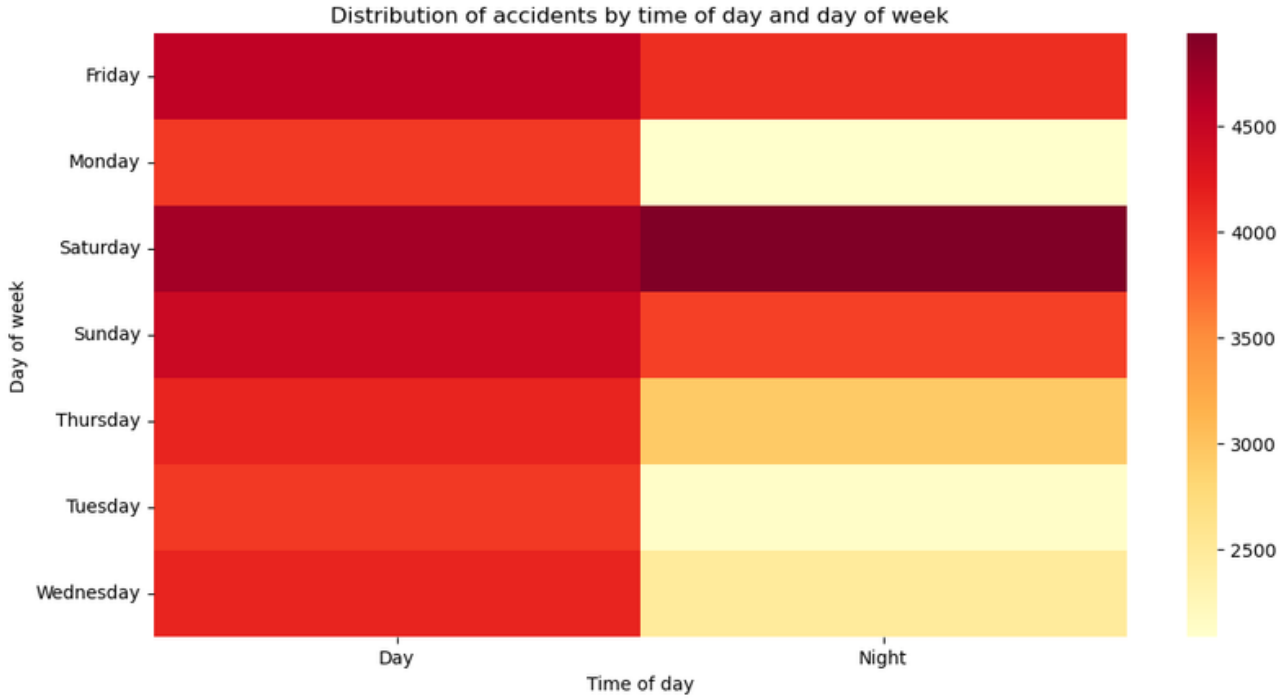


Number of crashes by state:

State	
ACT	478
NSW	16293
NT	1642
Qld	10495
SA	4547
Tas	1550
Vic	11562
WA	6276

Plot 9 - Heatmap visualizes the distribution of accidents by time of day and day of week

There are generally more accidents during the day than at night. Additionally, there appears to be a higher frequency of accidents on Fridays and Saturdays compared to other days of the week.



05

MODELLING

IDENTIFY DEPENDENT AND INDEPENDENT VARIABLES

- **Independent variables :** State, Speed Limit, Road User, Gender, Age, Dayweek
- **Dependent variable (Target value) :** Time of day

PREPROCESSING

- Convert some categorical variables to dummy variables which are : State, Road User, Gender, Dayweek. For example: Dayweek -> Dayweek_Friday, Dayweek_Monday, Dayweek_Saturday, ...

- Convert the target variable - 'Time of day' to binary type : Night - 0, Day - 1

SPLITING THE DATA INTO TRAINNING AND TESTING SET

To avoid overfitting, we split the data into 20% for testing, and the remaining 80% for training. We also set the seed of random number generator of 42 to ensures that the data is split in the same way every time the code is run.

05

LOGISTIC REGRESSION

We perform logistic regression on a binary classification problem, and evaluate its performance using accuracy, precision, recall, and F1 score metrics.

The objective is to predict the probability of a night-time accident on a specific day of week.

The evaluation result is :

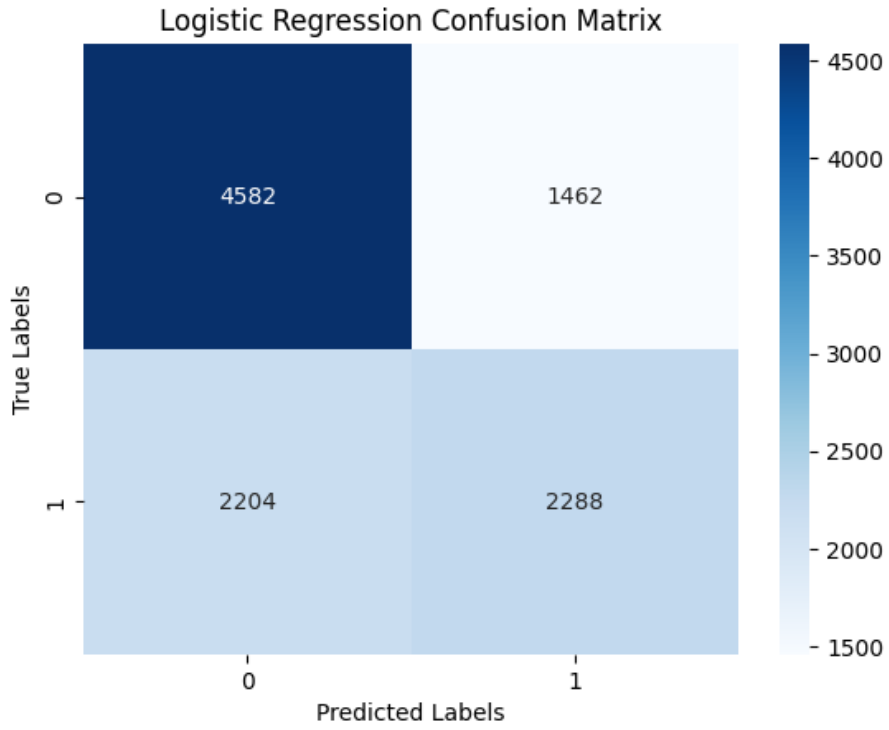
	precision	recall	f1-score	support
0	0.68	0.76	0.71	6044
1	0.61	0.51	0.56	4492
accuracy			0.65	10536
macro avg	0.64	0.63	0.63	10536
weighted avg	0.65	0.65	0.65	10536

The model's overall accuracy is 65%, with precision of 68% and recall of 76% for accidents during the day and precision of 61% and recall of 51% for accidents at night. The f1-score is 71% for day accidents and 56% for night accidents.

Overall, the model shows better performance in predicting accidents during the day than at night.

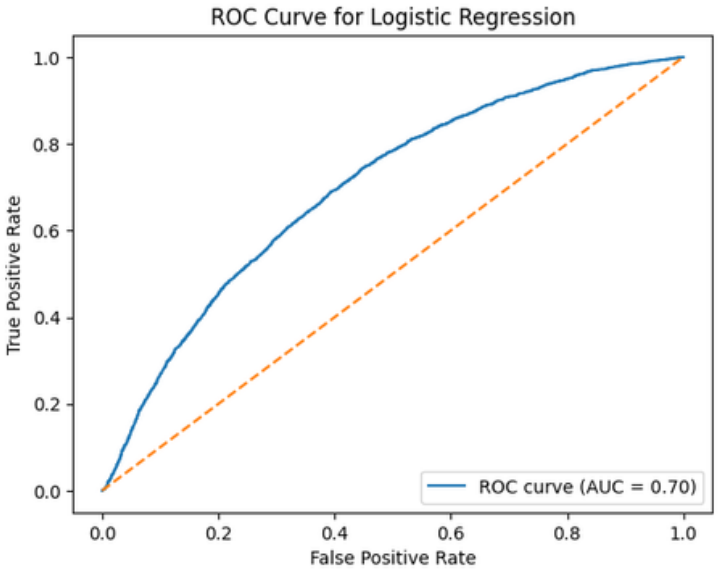
05

Confusion matrix :



- Based on the confusion matrix, the model has more false negatives (2204) than false positives (1462), indicating that it is better at correctly predicting accidents during the day than at night.
- True Positives (4582) and True Negatives (2288) are the highest numbers among the four values in the confusion matrix, it indicates that the model has correctly predicted the majority of both positive and negative instances. This suggests that the model has high accuracy and is performing well.

ROC curve and AUC score for a logistic regression model

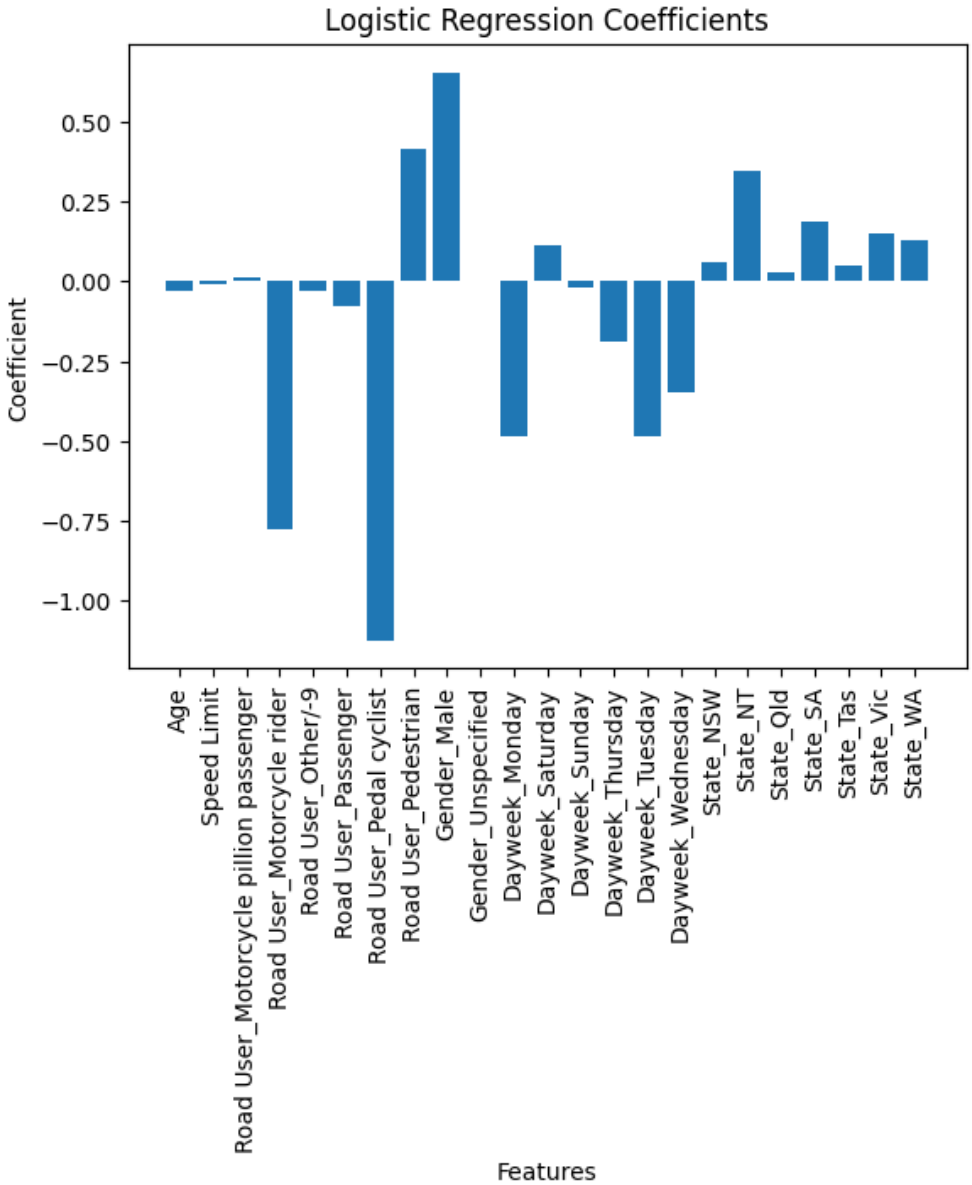


An ROC AUC of 0.70 indicates a moderate level of performance for the classification model. It means that the model has some predictive power, but there is still room for improvement.

06 EVALUATE INDEPENDENT VARIABLES

Since the logistic regression model in this case is a binary classification model with multiple features, we can visualize it by plotting the decision boundary on a two-dimensional plot. However, since the dataset has more than two features, it is not possible to plot the decision boundary in a simple two-dimensional plot.

Another way to visualize the model is to plot the coefficients of the features used in the logistic regression model. We can do this by creating a bar plot of the coefficients, where the height of each bar represents the magnitude of the coefficient. This will give us an idea of which features are most important in predicting the target variable



06 EVALUATE INDEPENDENT VARIABLES

Attribute	Coefficient
Road User_Pedal cyclist	-1.12659
Road User_Motorcycle rider	-0.77897
Gender_Male	0.65725
Dayweek_Monday	-0.487358
Dayweek_Tuesday	-0.484891
Road User_Pedestrian	0.418972
State_NT	0.346911
Dayweek_Wednesday	-0.346511
Dayweek_Thursday	-0.189226
State_SA	0.186618
State_Vic	0.150896
State_WA	0.128812
Dayweek_Saturday	0.115012
Road User_Passenger	-0.0775284
State_NSW	0.0608834
State_Tas	0.0513801
Road User_Other/-9	-0.0313417
State_Qld	0.0276339
Age	-0.0274839
Dayweek_Sunday	-0.0193292
Road User_Motorcycle pillion passenger	0.0140507
Speed Limit	-0.0074258
Gender_Unspecified	0.00351273

Some attributes can be considered important in predicting our target variable which are Road User, Gender, Dayweek and State:

- **Road User_Pedal cyclist:** This has the highest negative coefficient, indicating that being a pedal cyclist is strongly associated with accidents occurring during the day rather than at night.
- **Road User_Motorcycle rider:** This has a negative coefficient as well, indicating that being a motorcycle rider is also associated with accidents occurring during the day rather than at night.
- **Gender_Male:** This has a positive coefficient, indicating that males are associated with accidents occurring at night rather than during the day.
- **Dayweek_Monday, Dayweek_Tuesday** are both negative, indicating that these days have a negative impact on the prediction of the target variable. This suggests that accidents are less likely to occur during the night time on Mondays and Tuesdays compared to other days of the week.
- **State_NT:** This has a positive coefficient, indicating that accidents in Northern Territory are associated with accidents occurring at night rather than during the day.

07

CONCLUSION

Fatal crashes in Australia are a serious concern, with an increasing number of crashes occurring over the years.

> This analysis highlights the importance of factors such as day of week, road user type, and gender in predicting the likelihood of a fatal crash occurring during the day or night. The results of this analysis could be useful in developing interventions and policies aimed at reducing the number of fatal crashes in Australia.

Overall, while the accuracy of the logistic regression model is not extremely high, it still provides valuable insights into the factors that contribute to fatal crashes in Australia. This information can be used by policymakers, road safety experts, and other stakeholders to develop effective strategies for reducing the number of fatal crashes on Australian roads.

- For example, road safety campaigns targeting motorbike riders and pedal cyclists could be developed to raise awareness of the risks associated with these road user types.
- Additionally, policies aimed at reducing the speed limit on certain roads and increasing the number of pedestrian crossings could be developed to reduce the likelihood of fatal crashes involving pedestrians.
- Finally, targeted education and training programs could be developed to improve the driving skills of certain age groups, such as younger and older drivers, who may be at a higher risk of causing a fatal crash.