BUSINESS INTELLIGENCE CERTIFICATE (MCBI71)

Australian road accident analysis

Assessment 1

RoJESH DHAKAL

October 1, 2023

A toy cars crashed into each other

Description automatically generated

Table of Content

[1. Introduction 2](#_Toc147076436)

[2. Dataset(s) 3](#_Toc147076437)

[3. Data Wrangling 4](#_Toc147076438)

[4. Creating Measures 6](#_Toc147076439)

[5. Data Visualization 7](#_Toc147076440)

[6. Data Dashboard/ Reports 15](#_Toc147076441)

[7. Conclusions 17](#_Toc147076442)

[8. References 18](#_Toc147076443)

[9. Appendix 18](#_Toc147076444)

# Introduction

In Australia, 21.2 million motor vehicles were registered on 31 January 2023 which is an increase of 2.3 percent between January 2022 and January 2023. (Bureau of Infrastructure and Transport Research Economics, 2023) . With this surge in number of vehicles, it is especially important to examine the shifts in crash incidents. This report contains results of exploratory data analysis (EDA), of road transport crash fatalities in Australia which has been reported by the police to the State and Territory road safety authorities monthly. The main objective of this report is to gain valuable information and insight from the dataset and potentially uncover numerous factors, patterns, and trends that contribute to road fatalities. This report aims to understand the underlying factors that cause road crashes, with the help of data visualization and exploration, which might help to create safety policies and measures that can potentially reduce the number of fatalities.

* 1. **Requirements**

1. Data Selection: Identifying the reliable and reputable source for road crash fatality data from credible sources such as road authorities, government agencies, etc.
2. Data Cleaning and Preparation: The data must be cleaned and pre-processed to remove any inconsistencies, and missing values so that it does not affect the analysis.
3. Exploratory Data Analysis (EDA): After the data is cleaned, a robust EDA must be conducted which aids in getting insights and identifying different trends and patterns.
4. Data Visualizations: Creating meaningful and informative data visualization such as charts, graphs, etc. to present the findings from the EDA clearly and understandably.
5. Report Writing: Making a comprehensive report that contains the results and key insights from EDA.
   1. **Business Questions:**
6. **High-Level Questions:**

How can we leverage the insights gained from the EDA of road crash fatalities to develop and implement effective road safety measures, ultimately reducing the number of fatalities and improving overall safety?

1. **Low-Level Questions:**

- What demographic characteristics are associated with a higher risk of being involved in a crash?

- What are the primary contributing factors to road crash fatalities (e.g., road type, overspeeding, etc.)?

- Are crashes more common during specific times of day?

- Can we identify monthly or seasonal patterns in road crashes?

- Does the type of road (e.g., highway, residential street) affect the likelihood of a crash?

# Dataset(s)

The dataset is gathered by the author directly from the Australian Road Death Database and is available at Kaggle. The file is in CSV format, and it can be accessed using this link [https://w](https://www.kaggle.com/datasets/deepcontractor/australian-fatal-car-accident-data-19892021)ww.kaggle.com/datasets/deepcontractor/australian-fatal-car-accident-data-19892021

**Name of the Dataset:** Australian Fatal Road Accident 1989-2021.

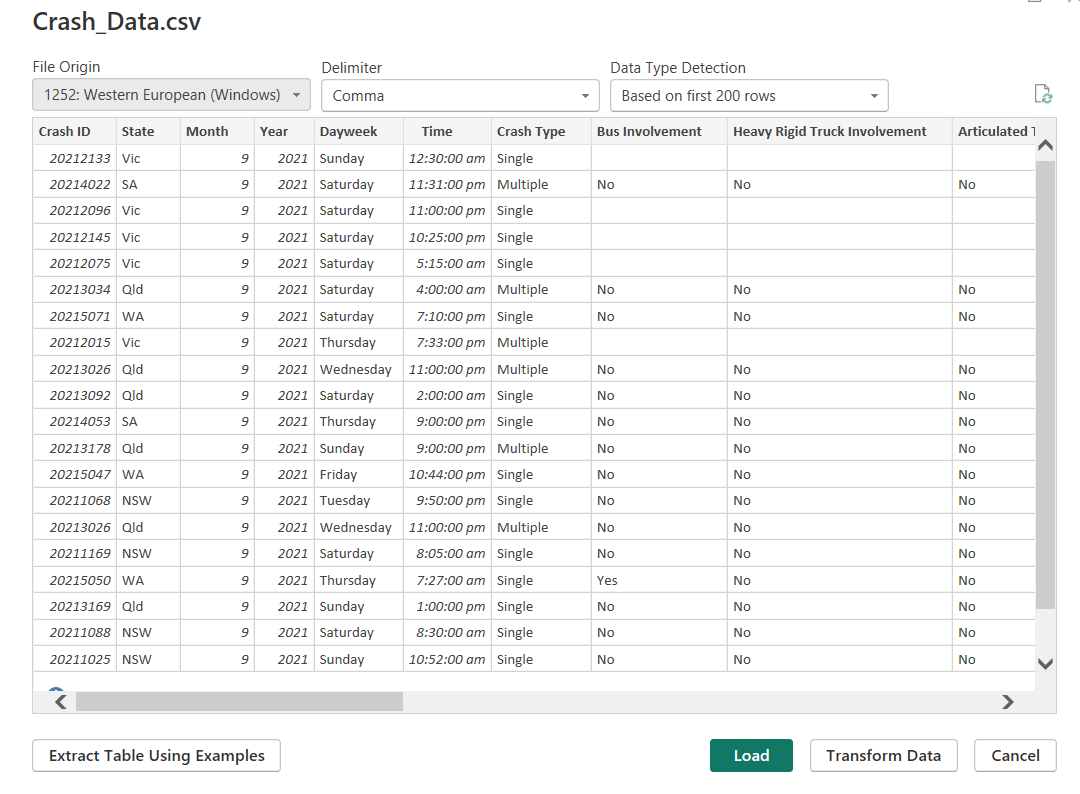
**Features:**

1. Crash Id: Crash Id is unique and contains the details of incidents that occurred in the various regions.
2. State: 6 different states of Australia, VIC (Victoria), QLD (Queensland), NSW (New South Wales), WA (West Australia), TAS (Tasmania), SA (South Australia) and two different territories ACT (Australian Capital Territory) and NT (Northern Territory).
3. Month: Encompasses all twelve months in a year, and the value ranges from 1 to 12.
4. Year: Date ranges from 1989-2021.
5. Day week: All seven days in a week, representing a complete week.
6. Time: The time when the incidents occur which is during day or night.
7. Crash Type: This column shows the value of crash type is classified as single or multiple.
8. Bus Involvement: Whether the bus was involved in a crash.
9. Heavy Rigid Truck Involvement: Whether the truck was involved during the crash.
10. Speed Limit: Speed of vehicle during crash.
11. Road User: Person involved during the crash.
12. Gender: Male or Female population involved in a crash.
13. Age: Age of the person involved during the crash.
14. Christmas Period: Whether the crash occurred during the Christmas period.
15. Easter Period: Whether the crash occurred during the Easter period.
16. Age group: Different range of age groups.
17. Day of week: Did the crash happen during a weekday or weekend?
18. Time of Day: This variable shows the time of crash i.e., Night or Day.

# Data Wrangling

**Data Loading:**

The crash data, which was initially in CSV format, was loaded in Power BI where the data was cleaned, shaped, and transformed using the Power Query Editor. Since the dataset was not fully cleaned, therefore Transform Load option was required. The picture below shows a quick overview of the data.



**Data Cleaning & Transformation**

**State:** The data was not consistent across the board for the state; therefore, basic transformation is done to align them in a consistent form.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Before After

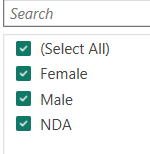
**Age:**  90 rows contain data as -9 for age which does not seem quite right because the age of a person cannot be negative. The data has been converted into positive integers for further analysis.

A screenshot of a computer

Description automatically generated

**Gender:** While analysing this column, some of the data were left blank and it has been filled with No Data Available (NDA) because the overall data is needed for better analysis.

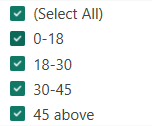
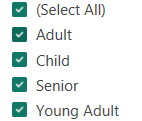
A screenshot of a computer

Description automatically generated 

Before After

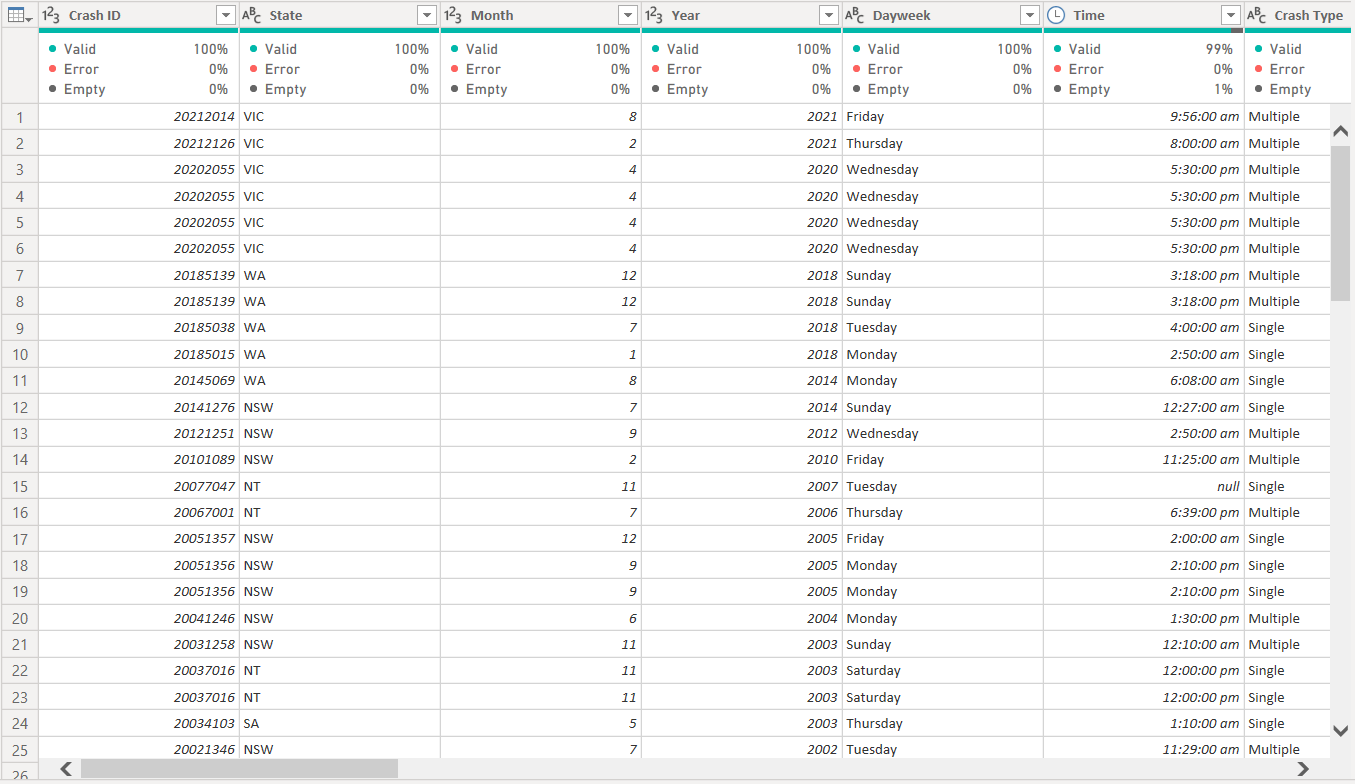
The same procedure, as implemented for the Gender column, has been applied to those columns where the data was left blank. The columns include Bus Involvement, Heavy Rigid Truck Involvement, Articulated Truck Involvement, National Remoteness Area, SA4 Name 2016, and National LGA Name 2017.

The Age Group column was dropped because of the presence of some blank values, and unnecessary subdivision of data which was not required in this business context. Instead, a more streamlined approach was taken by utilizing the Age column to derive the required information. Subsequently, two columns, namely ‘Age Range’ and ‘Age Categories’ were created. This classification allowed for the age data to be organized into four distinct groups, enhancing the relevance and utility of the dataset for analysis.

Age Range Age Categories

Upon, further analysis, there were duplications in the Crash ID column, which have been removed. This was quite essential as these duplications contained identical data.





After conducting a more detailed analysis, it was observed that the gender column contained entries labelled as ‘Unspecified,’ which were replaced by NDA. Additionally, the road user category included a value as Other/-9 which were replaced by Other. These were all the adjustments and changes that were made to enhance the clarity and accuracy of the dataset.

# Creating Measures

Total Crashes = COUNTROWS ('Crash\_Data (2)')

Average Speed = AVERAGE ('Crash\_Data (2)'[Speed Limit])

CY Accidents = TOTALYTD (COUNT ('Crash\_Data (2)'[Crash ID]), 'Calendar'[Date])

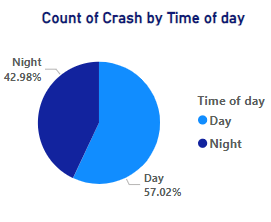
PY Accidents = TOTALYTD (COUNT ('Crash\_Data (2)'[Crash ID]), SAMEPERIODLASTYEAR('Calendar'[Date]))

YoY Accidents = ([CY Accidents]- [PY Accidents])/ [PY Accidents]

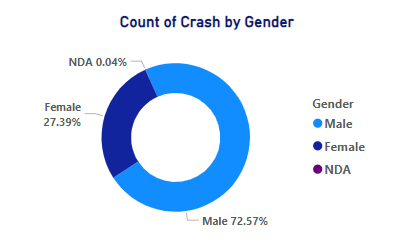
# Data Visualization

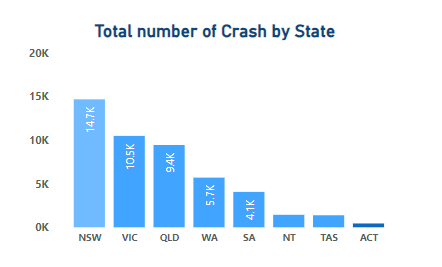
Data visualization is a blend of both art and science. The main objective is to accurately represent the art right without getting the science wrong, and vice versa. A data visualization must not mislead or distort the data and accurately convey the message. When the visualztion is presentable and pleasable, it can easily convey the message. On the other hand, if visuals have contrasting colors, imbalance designs it can be distracting and misleading to the reader (O.Wilke, 2019).

 Over 3 decades, time, there were around 47.57k crash incidents that were reported in Australia which can be seen by the card visuals.

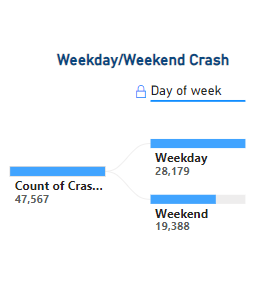


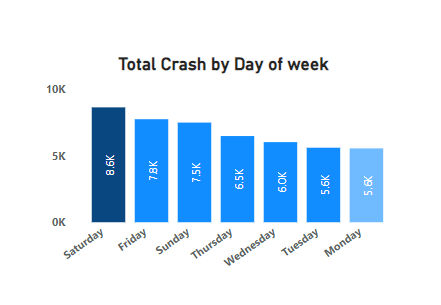
Based on the pie chart, it is evident that the majority of the crashes happened during daytime, accounting for approximately 57.02 percent of total incidents. This suggests that there might be specific factors such as traffic volume, road conditions, etc, which might be the cause, where further analysis can be done. By recognizing these daytime crashes, there is an opportunity to strategically allocate proper resources such as increasing the presence of emergency response teams, police patrols, etc. to improve the efficiency and effectiveness of the crash response.

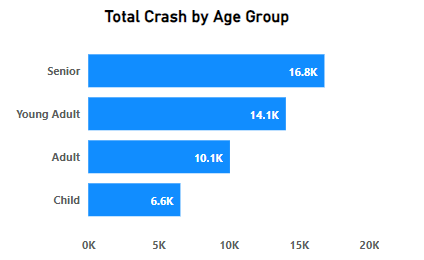


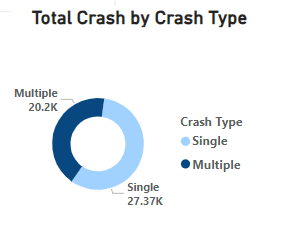
The donut chart illustrates a considerable number of male populations were more involved in the crash which was about 72.57 percent. Also, in the dataset, about 0.04% of the data for gender were missing and were replaced with NDA. This initial observation begs further investigations into specific factors contributing to higher crashes in the male population. Understanding these underlying causes can inform targeted interventions and policies to improve road safety.

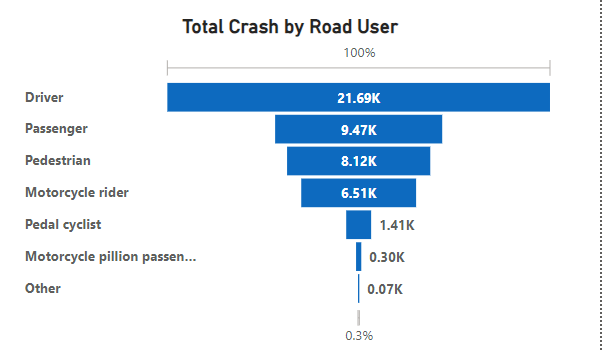
The column chart illustrates New South Wales (NSW) region reported the highest number of crash incidents, totalling 14.7 thousand, while the Australian Capital Territory (ACT) had the least incidents reported which were 0.4k. Victoria ranked as the second highest state in terms of crash involvements. Around 10.5k populations, were involved during a crash. NSW being the most populous state, naturally has a higher number of crash incidents. With ACT reporting the lowest number of incidents, it may be worth examining their road safety policies and practices which can be adopted by other states.

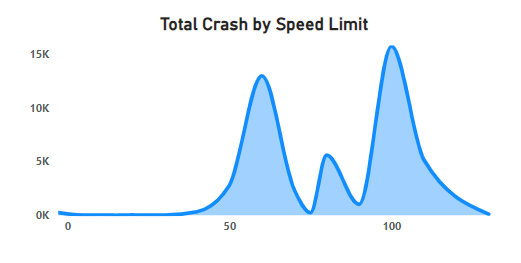
Overall, the total number of crashes during weekdays is higher as compared during weekends. The higher number of crashes during weekdays suggests that there is typically a greater volume of traffic on the roads which could be attributed to factors such as commuters, work-related travel, and school transportation. This could also be, numerous factors such as rush-hour periods, time constraints, etc. Proper resources and strategies could be developed to minimize the crash risks.

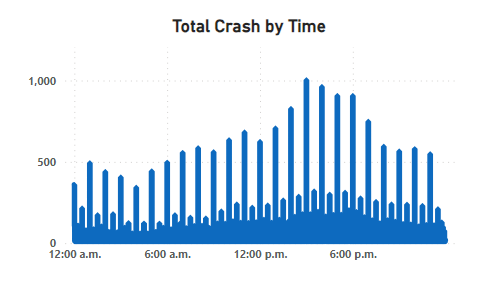
The column chart portrays a visual representation of the total number of crashes that occurred in Australia. Saturday saw a higher number of crash-related incidents which may be due to an increase in recreational activities, higher traffic volumes, etc. Additional data, especially on recreational types and their locations could offer valuable insight. Conversely, Mondays and Tuesdays experienced a notably lower number of reported crash incidents. 

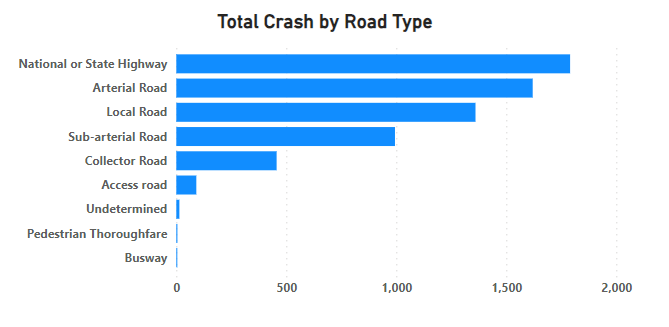
The bar chart represents, the total number of crashes by age group. The X-axis shows the total number of crashes, while the Y-axis represents 4 distinct age groups. It is noteworthy that the senior populations were more involved in the crashes, surpassing other age groups. Following closely, young adults were the second highest demographic in terms of crash involvement. The senior population were more prone to crashes due to slower reaction times, diminished visual sights, etc. Therefore, regular health check-ups and age-related safety measures can be created. 

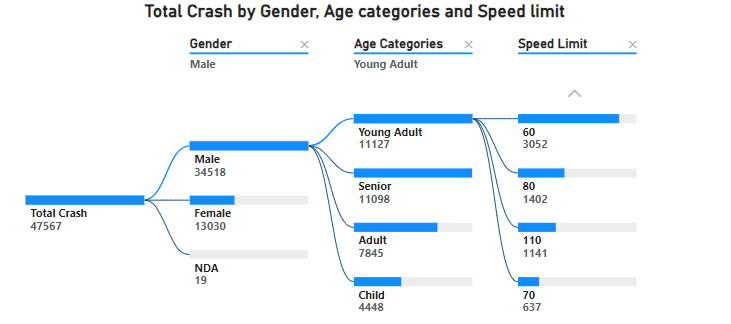
The donut chart provides a visual breakdown of crash involvement, indicating that single individuals accounted for approximately 57.54 percent of the incidents. This begs for the opportunity to emphasize that addresses safe driving practices, hazard recognition, and defensive driving techniques. Further investigations are recommended to understand the specific circumstances and contributing factors surrounding single-driver crashes. 

The waterfall chart depicts drivers were more likely to be involved in crashes. Interestingly, the ratio of passenger and pedestrians involved in crashes were nearly equivalent.

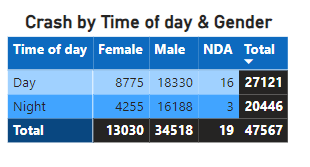
The area chart shows the total number of crashes by speed. The X-axis denotes speed, while the Y-axis represents the total number of crash incidents. Notably, there were 15.7 thousand reported crash-related incidents at a speed of 100, making it the highest recorded speed-related incidents. The second highest occurred at approximately 13 thousand incidents at a speed of 60. This required attention in terms of awareness, enforcement, and potential speed limit adjustments.

The line chart depicts the total crashes over time. More than 1000 crashes were reported during 3 p.m. The majority of crash incidents occurred between 3 pm to 6 pm. This coincides with the typical afternoon rush hour, suggesting increased traffic volumes and potentially rushed behaviours which may contribute to crashes. Campaigns can be launched on reducing distractions, adhering to speed limits, safe driving practises, etc., which might minimize the crash during that period. 

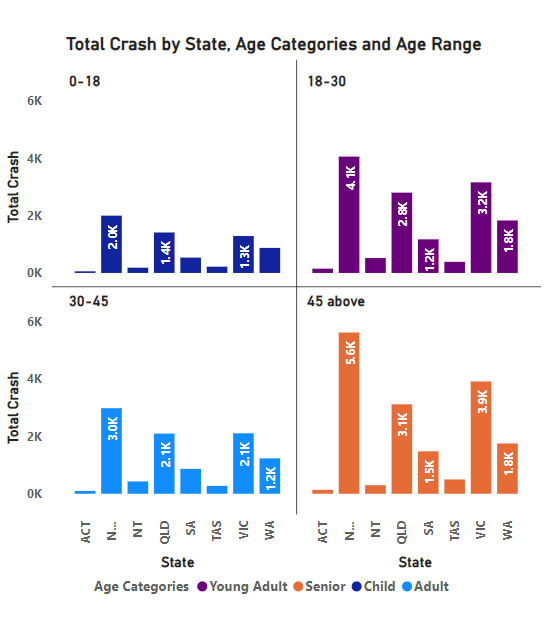
The column chart provides a visual representation of the total number of crashes based on road type. The majority of the crashes occurred on National or State Highway, closely followed by Arterial Road. In contrast, Pedestrian Thoroughfares and Busways experienced the lowest incidence of crashes. Since, the majority of incidents occurred on National Highway and Arterial Road, enhancing signage, speed limit enforcement, and road maintenance may help to reduce crashes.



The decomposition tree depicts an insightful breakdown of the total number of crashes by breaking them into various components. Notably, males were more frequently involved in these crashes. Further analysis within the tree reveals both Young Adults and Senior males were particularly prone to crashes. For Young Adults, a considerable proportion of incidents occurred at a speed of 60, followed by 80 and 110. This information sheds light on specific demographics and circumstances which can help in curating better safety measures and policies.



The matrix offers a visual representation of total crashes categorized by both time of day and gender. Broadly, the crashes were more prevalent during the daytime. When examining gender-specific patterns, it becomes apparent that females experienced notable crashes during the daytime, surpassing their nighttime figures by more than double. On the other hand, males had significantly higher incidences of crashes and the ratio of crashes between day and night for males remained relatively consistent.

The stacked column chart represents, the total crashes categorized by age categories and age range. Notably, the 45 above age group had more involvement in crashes. Within this group, NSW recorded the highest number of crashes, which was about 5.6 thousand, followed by Victoria state with around 3.9 thousand incidents. In contrast, for the 0-18 age group, NSW also reported the highest number of crash incidents, approximately 2 thousand, while QLD stood second with 1.48 thousand.

A pink circle with a number of percentages

Description automatically generated

The donut chart presents the concerning statistic, around 3.02 percent of all total crashes during the Christmas period. The percent may seem modest, but it is quite concerning because of the duration of this holiday season. Therefore, it is important to launch public awareness programs and increasing the police patrols during the holidays which might help to bring the number down.

# Data Dashboard/ Reports

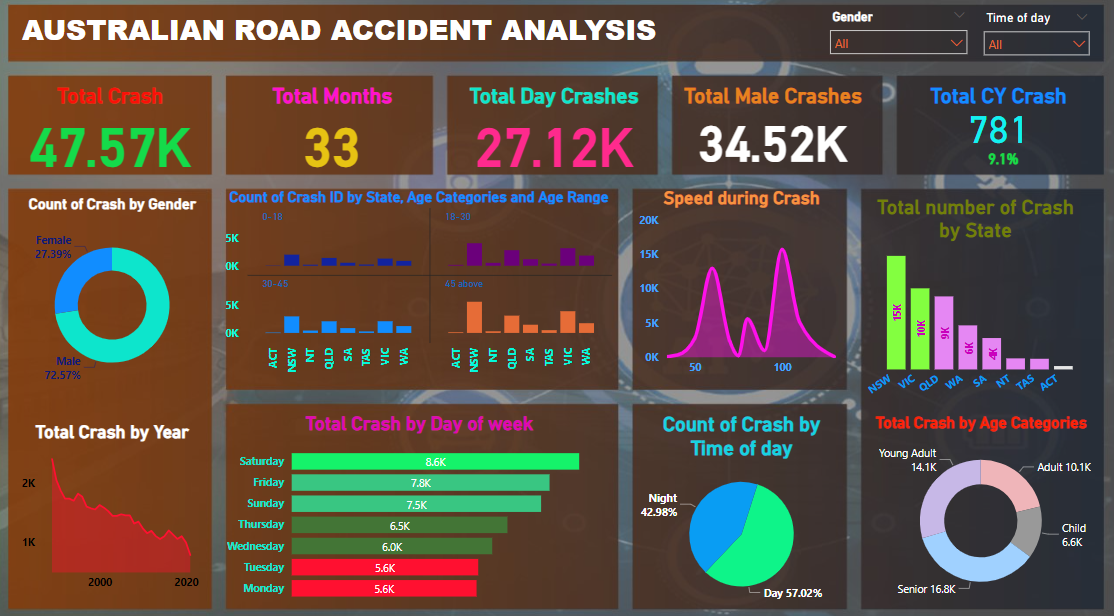


Fig: Main Dashboard

A screenshot of a computer screen

Description automatically generated

Fig: Duration Analysis Dashboard

A screenshot of a computer

Description automatically generated

Fig: Dashboard for Roads

A screenshot of a computer

Description automatically generated

Fig: Dashboard for Demographics

# Conclusions

The comprehensive exploratory data analysis of road crash tragedy presented a valuable opportunity and key insights to enhance road safety measures which can ultimately reduce the number of fatalities. All the key questions related to demographics, road types, timings, and contributing factors have been addressed, through which targeted strategies and new laws and policies can be enforced, which mitigate risks and save lives. The EDA has provided a piece of key information that males were frequently involved in crashes and are identified as higher-risk demographics indicating a potential need for targeted educational and awareness campaigns tailored toward male drivers. Moreover, the senior population were involved in crashes in Australia, which highlights the importance of implementing measures that will address the challenge and specific needs faced by old drivers. From the analysis, speeding also played a key role in crash incidents. Therefore, new policies should be developed regarding speed control measures with proper signage and enforcement which might help to bring crash fatalities down. Also, the insights can be leveraged to allocate the resources available such as the crashed occurred more during the day, which means more resources such as traffic patrollers, emergency response teams, and other resources can be placed strategically. Furthermore, on state highways and arterial roads, it was observed it had higher incidences of crashes. This highlights the need for targeted safety measures on such high-risk road types. Therefore, policy planners can allocate resources for the improvement, expansion, and maintenance of these types of roads. Better signage, road widening, and regular maintenance can help reduce these types of crashes. Therefore, strict speed limit enforcement measures on state highways are essential. In addition, data indicates single drivers were more frequently involved in crashes which underscores the absolute need for targeted involvement and policies that are aimed for driver safety. Various campaigns can be launched to raise awareness, and driver education programs focusing on safe driving practices are required.

To conclude, the insights obtained from this data analysis provides a strong starting point for making roads safer. By focusing on specific actions and policies, there is an enormous potential to decrease the number of crashes and improve road safety.

# References

Bureau of Infrastructure and Transport Research Economics. (2023, January). *Road Statistics*. Retrieved from Australian Government: https://www.bitre.gov.au/publications/2023/road-vehicles-australia-january-2023

O.Wilke. (2019). *Google*. Retrieved from Google Books: https://www.google.co.nz/books/edition/\_/XmmNDwAAQBAJ?hl=en&gbpv=1&pg=PP1&dq=o%27reilly+data+visualization+book

# Appendix

**Glossary:**

EDA- Exploratory Data Analysis

NDA- No Data Available

DAX- Data Analysis Expression

NSW- New South Wales

VIC- Victoria

QLD- Queensland

WA- Western Australia

SA- South Australia

TAS- Tasmania

ACT- Australian Capital Territory

NT- Northern Territory

**Software and Tools Used:** Power BI