Australian Road Accident Analysis

Assessment No 2



Rojesh Dhakal

November 12, 2023

Table of Content

[1. Introduction 3](#_Toc150521273)

[2. Descriptive Data Analysis 4](#_Toc150521274)

[3. Diagnostic Analysis 14](#_Toc150521275)

[4. Business Questions 20](#_Toc150521276)

[5. Exploratory Data Analysis 22](#_Toc150521277)

[6. Feature Selection 23](#_Toc150521278)

[7. Conclusions 24](#_Toc150521279)

[8. References 25](#_Toc150521280)

[9. Appendix 25](#_Toc150521281)

**List of Figures**

[Figure i: Main Dashboard 5](#_Toc150521305)

[Figure ii: Boxplot of Speed Limits 6](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521306)

[Figure iii: Descriptive Statistics of Speed Limit 6](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521307)

[Figure iv: Histogram of Speed Limits 7](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521308)

[Figure v: Descriptive Statistics of Age 7](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521309)

[Figure vi: Boxplot of Age 8](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521310)

[Figure vii: Histogram of Age distribution 9](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521311)

[Figure viii: Descriptive Statistics of Year 9](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521312)

[Figure ix: Distribution of Gender 10](#_Toc150521313)

[Figure x: Monthly Crashes 11](#_Toc150521314)

[Figure xi: Crash Timings 11](#_Toc150521315)

[Figure xii: Distribution of age groups in crashes 12](#_Toc150521316)

[Figure xiii: Crash distribution by different days 13](#_Toc150521317)

[Figure xiv: Weekdays/Weekends Crash 13](#_Toc150521318)

[Figure xv: Road User involvements during crash 14](#_Toc150521319)

[Figure xvi: Distribution of crash by time of the day 14](#_Toc150521320)

[Figure xvii: Total crash by various crash types 15](#_Toc150521321)

[Figure xviii: Analysis of crash by age group and gender 15](#_Toc150521322)

[Figure xix: Analysis of crash by gender and speed 16](#_Toc150521323)

[Figure xx: Analysis of proportion of gender involved in different crash types. 17](#_Toc150521324)

[Figure xxi: Analysis of proportion of gender involved in crash during night/day. 17](#_Toc150521325)

[Figure xxii: Analysis of different age groups involved in different crash types in various state. 18](#_Toc150521326)

[Figure xxiii: Analysis of various age groups with their gender involved in different crash types. 19](#_Toc150521327)

[Figure xxiv: Analysis of crash by different age group, gender during day/night. 20](#_Toc150521328)

[Figure xxv: Hypothesis testing between age and speed. 21](#_Toc150521329)

[Figure xxvi: Hypothesis testing between age and crash. 22](#_Toc150521330)

[Figure xxvii: Hypothesis testing between crash and speed. 22](#_Toc150521331)

[Figure xxviii: Hypothesis testing between crash with the years. 23](#_Toc150521332)

[Figure xxix: Young Adult Crashes 23](#_Toc150521333)

[Figure xxx: Involvement in crashes during 100 speeds. 24](file:///C:\Users\User\Desktop\Future%20Skills%20Assements%202.docx#_Toc150521334)

[Figure xxxii: Crash in NSW by Gender and Age group 26](#_Toc150521335)

[Figure xxxiii: Weekends Crash by Timings 26](#_Toc150521336)

# Introduction

This report presents extensive analysis of the Australian Road Accident data spanning from 1989 to 2021, with primary focus on understanding the complex dynamics of road safety, particularly in relation to gender and age groups. A targeted analysis was conducted to explore the dynamics of age, speed, and their implications on road safety. This report zooms in on the characteristics of two specific age groups- seniors and young adults- aiming to recognize patterns and trends that could significantly contribute to understanding of crash fatalities.

The main objective of this report is to gain valuable insights and information from the Australian crash dataset, which sheds lights upon various factors, patterns and trends that play a vital role in crash fatalities. The report seeks to comprehensively grasp the main causes of crash incidents through the power of inferential statistics, data visualization and exploration, with the primary aim of informing for the creation of safety measures and policies which could help to reduce the tragic accidents that occurs on the road and ultimately save lives.

To achieve this objective, the dataset has been analysed comprehensively utilizing inferential statistics. With the help of thorough data preparation and exploration, the key factors that has contributed to crash have been identified. Leveraging inferential statistics, evidence-based recommendations has been formulated for safety policies and measures. These strategies aim to reduce the tragic toll of lives lost during crash and foster safer and secure environment for road users.

While preparing the report, various considerations has been considered. First and foremost, the quality of the data has been enhanced during data preparation and cleaning process. Any issues regarding the missing data, outliers, or errors have been identified and dealt accordingly. Proper statistical techniques have been used to ensure that findings are robust, and suggestions are strong and reliable.

# Descriptive Data Analysis

Descriptive analysis encompasses the processes of gathering, cleaning, and summarizing data from various sources to facilitate well-informed decisions because it provides clear and concise summary of the data. The information is conveyed in a way that is easily comprehensible, with the help of charts, tables, and graphs. This method aids in recognizing strengths and weaknesses, tracking progress over the period, and make informed decisions based on historical data (Hurwitz et al., 2015). According to Greasley (2019), by employing suitable tools and techniques, descriptive analytics can empower business to maintain competitive edge and achieve long term success.

**2.1. Dataset Overview**

The dataset is about Australian Road Accident that ranges from 1989 to 2021.The dataset in this report was collected by the author directly from the Australian Road Death Database and is available at Kaggle. The dataset is provided in CSV format, and it can be accessed using this link <https://www.kaggle.com/datasets/deepcontractor/australian-fatal-car-accident-data-19892021>

**2.2. Variables**

|  |  |  |
| --- | --- | --- |
| Variable Category | Datasheet Name: | |
| Variable | Data Type |
| Unique Identifier | Crash Id | Categorical, Nominal |
| City | State | Categorical, Nominal |
| Time information | Month | Categorical, Ordinal |
| Year | Discrete, Continuous |
| Day week | Categorical, Nominal |
| Time | Continuous |
| Day of week | Ordinal, Categorical |
| Time of day | Nominal, Categorical |
| Type | Crash Type | Categorical, Nominal |
| Involvement | Bus Involvement | Categorical, Nominal |
| Heavy Rigid Truck Involvement | Categorical, Nominal |
| Articulated Truck Involvement | Categorical, Nominal |
| Limits | Speed limit | Ordinal |
| User | Road User | Categorical |
| Demographics | Gender | Categorical, Nominal |
| Age | Continuous |
| Age Group | Categorical |
| Age Categories | Categories, Ordinal |
| Holidays | Christmas Period | Categorical, Nominal |
| Easter Period | Categorical, Nominal |
| Roads | National Road Type | Nominal, Categorical |
| National Remoteness Area | Nominal, Categorical |
| National LGA Name 2017 | Nominal, Categorical |

**2.3. Dashboard**

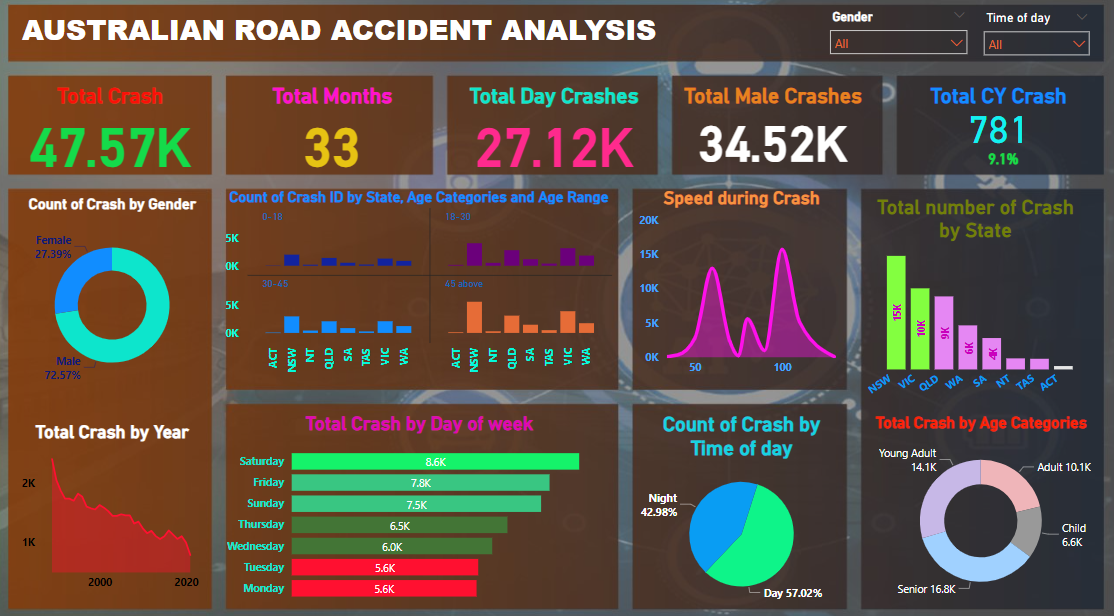


Figure : Main Dashboard

The dashboard above shows the brief overview of the data. Over the course of 33 years, approximately 47.57 thousand crashes were recorded. Notably, total number of daytime crashes were about 27.12 thousand. Surprisingly, males were involved in 34.5 thousand crashes. Saturday saw the higher number in crashes. Senior population was the top contributor to overall crash count. NSW stood out among other regions with a higher incidence of crashes. Notably, at speed of 100, there was an increased number of accidents.

**2.4. Descriptive Statistics for Quantitative Variables**

**a. Speed Limit**

A graph with numbers and a box

Description automatically generated A screenshot of a table

Description automatically generatedThe figure depicts the descriptive statistics of speed limit during crashes provided valuable insights into these crashes. The average speed limit when the crash occurred was 81.19, with a median of 80, indicating that the majority of crashes occurred during the central value. Additionally, the mode is 100, which shows that most of crashes happened at the speed of 100. Further analysis and investigation can be done to understand why and where these crashes occurred and implement necessary plans and actions which can help to mitigate the risks. The maximum speed when the crash occurred was at 130 and surprisingly crash did happen at the speed of 5. The standard deviation is 22.12 which suggests that the speed limit values are relatively spread out from the mean. The negative kurtosis (0.35) suggests that the distribution of speed limit values is slightly platykurtic, meaning it has lighter tails than a normal distribution. The negative skewness (0.47) indicates that the speed limit distribution is slightly skewed to the left, which implies that most accident occurred at lower speed.

Figure ii: Boxplot of Speed Limits

Figure iii: Descriptive Statistics of Speed Limit

The boxplot of speed limit offers key statistics related to the dataset. The minimum point at which the speed occurred was at 5 and the maximum was at 130. The first quartile (Q1) is 60, which indicates 25% of the data falls below this value. The third quartile (Q3) is 100, which indicates 75% of the data lies beneath this threshold. The height of the box is equal to IQR, which is calculated as Q3-Q1 which is 40, measures the spread of the central data values. Larger IQR indicates greater variability in the central values.

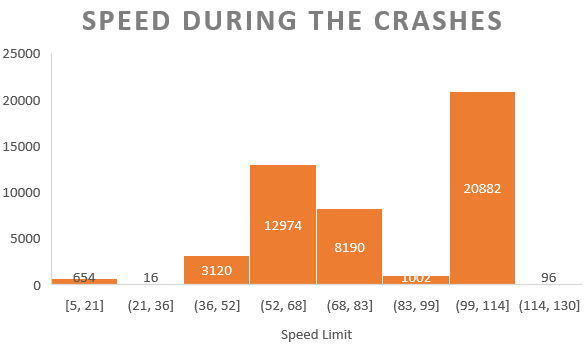
 The figure histogram of speed limits provides the distribution of speed related crashes. A significant number of the crashes occurred at the speed between 99 to 114, accounting for 20.8 thousand crashes. For high-frequency speed range, initiatives such as speed limit enforcement, road patrols, awareness campaigns can be effective to reduce accidents. Likewise, the second highest crashes occurred at the speed between 52 to 68 which accounted for 12.9 thousand.

Figure iv: Histogram of Speed Limits

**b. Age**

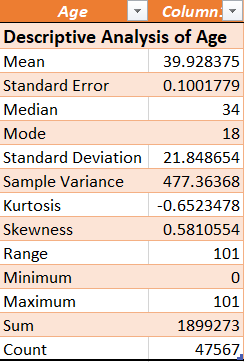
The figure depicting the descriptive statistics of age of the person involved during crashes provides valuable insights. The average age of a individuals involved in a crash was 39.93, with a median age of 34, reflecting the central tendency of the age distribution. It is intriguing that the mode is 18, which signifies that most of crashes occurred at the age of 18, which warrants for in-depth analysis and investigation, as to understand why and where these crashes occurred and implement necessary plans and actions which can help to mitigate the risks among the young generations. The maximum age at which person got in the crash was at 101 and newborn were also involved during the crash. The standard deviation is 21.85 which suggests that the speed limit values are relatively spread out from the mean. The negative kurtosis (0.65) suggests that the distribution of age values is slightly platykurtic, meaning it has lighter tails than a normal distribution. The positive skewness (0.58) indicates that the age distribution is slightly skewed to the right, which implies there are more individuals with ages greater than mean.

Figure v: Descriptive Statistics of Age

A green and black graph

Description automatically generated The boxplot of age offers key statistics related to the dataset. The minimum point at which the person involved at crash was at 0 and the maximum was at 101. The first quartile (Q1) is 22, which indicates 25% of the data falls below this value. The third quartile (Q3) is 56, which indicates 75% of the data lies beneath this threshold. The height of the box is equal to IQR, which is calculated as Q3-Q1 which is 34, measures the spread of the central data values. Larger IQR indicates greater variability in the central values.

Figure vi: Boxplot of Age

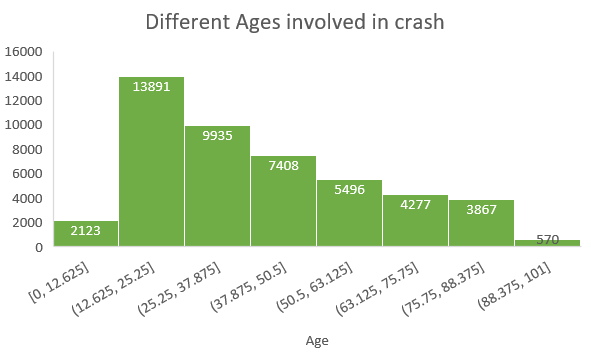


Figure vii: Histogram of Age distribution

The picture depicts histogram of the ages of individuals involved in a crash. The histogram shows significant of the crashes involves younger individuals, particularly from the age range of 12 to 25, accounting for approximately 13.8 thousand cases. The age bracket of 25 to 37 also experienced a notable number of accidents. Conversely, individuals aged 88 to 101 had a less involvement in the crashes, which may be attributed to their reduced mobility or perhaps more cautious approach to driving.

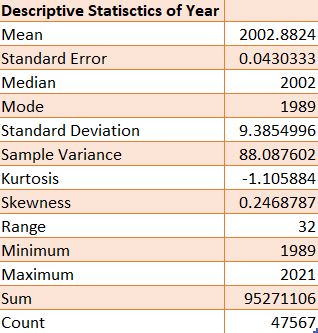
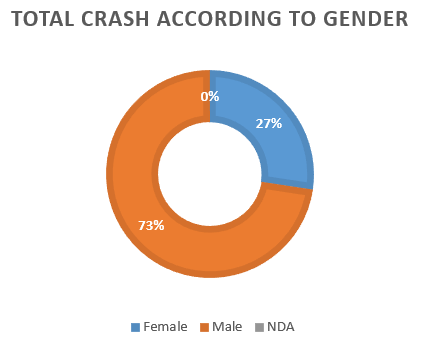
**c. Year**

Figure viii: Descriptive Statistics of Year

The figure depicts the descriptive statistics of year when the crash occurred. The average year is approximately 2002.88, with a median year of 2002, reflecting the central tendency of the age distribution. It is intriguing that the mode is 1989, which signifies that most of crashes occurred at the year. The negative kurtosis (1.11) suggests that the distribution of year values is slightly platykurtic, meaning it has lighter tails than a normal distribution. The positive skewness (0.25) indicates that the age distribution is slightly skewed to the right, which implies there are more years greater than mean.

**2.5. Descriptive Statistics for Qualitative Variables**

**a. Gender**

 A screenshot of a computer

Description automatically generated

Figure : Distribution of Gender

The figure above reveals a notable gender disparity in crash involvement, with a significant proportion of male involved in the crash accounting for 73%. Additionally, about 19 entries had missing data which was replaced by NDA. This initial observation emphasizes the need for further investigations into specific factors contributing to crashes in male population. Identifying the underlying causes, risk factors and behaviour patterns associated with male drivers can help to develop targeted interventions and policies which aims to reduce the incidents and enhance road safety.

**b. Month**

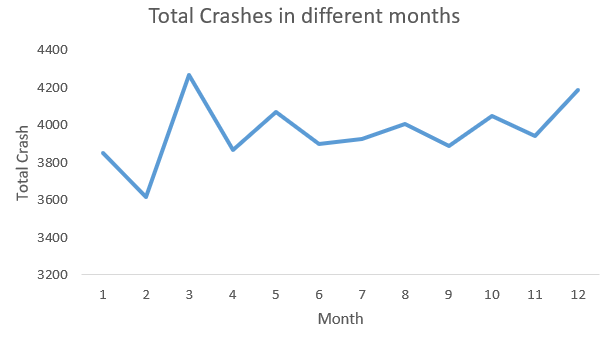


Figure : Monthly Crashes

The line chart above shows insights into crash occurrences across various month highlighting some noteworthy patterns. There was significant spike in crashes that occurred in month of April, followed closely by December. Further analysis can be done to gain deeper understanding of these trends, taking account of other factors such as weather conditions, holiday seasons etc. In contrast, February had the least number of accidents as compared to other months.

**c. Time**

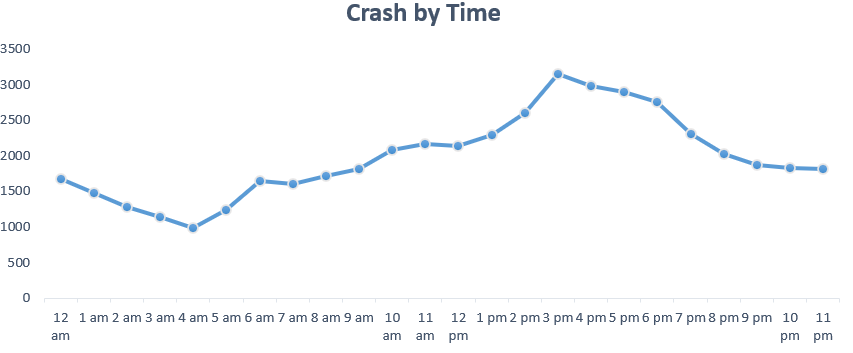


Figure : Crash Timings

The line chart above shows the timings of the crashes highlighting some noteworthy patterns. There was significant number of crashes that occurred between 3pm to 6 pm, which coincides with the traditional afternoon rush hour. This temporal correlation suggests some kind of connection between increased traffic volumes and potentially rushed behaviours which might contribute to crash. However, from 2am to 4 am, there is reduction in number of crashes, due to low traffic flow and reduced activities during early mornings.

**d. Age group**

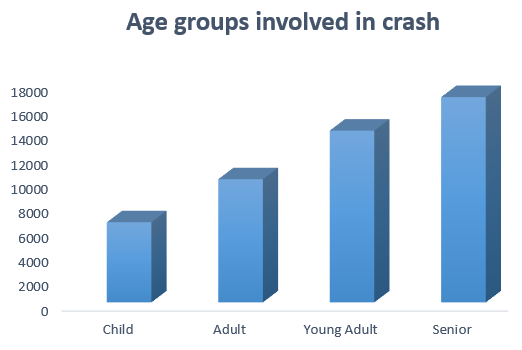
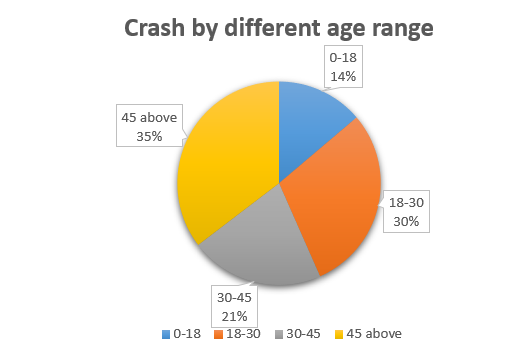
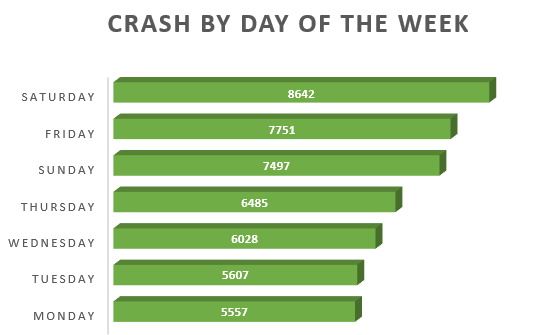
 

Figure : Distribution of age groups in crashes

The chart above represents the total number of crashes by various age groups. Senior population stands out as the group most frequently involved in crashes, surpassing all other categories. This might be due to various factors such as diminished eye sights, slower reaction times etc. Therefore, to address these issues, encourage seniors to do regular health checkups and implement age- related safety measures such as vision checks and adapted driving techniques. Following closely, young adult were the second highest groups in terms of crash involvement. This might be due to various factors such as inexperience, risky behaviour, distracted behaviour, etc. Therefore, proper safety measures can be launched, and proper education and training can be given to these age groups.

**e. Weekdays**

 A screenshot of a graph

Description automatically generated

Figure : Crash distribution by different days

The bar chart depicts total number of crashes in Australia on different days. Notably, Saturday surpasses all other day in reported crashes, which might be due to increase in recreational activities, traffic volume, carefree driving activities during weekend etc. Conversely, Monday and Tuesday, experienced a lower number of reported crash incidents.

**f. Day of week**

A diagram of a graph

Description automatically generated A screenshot of a computer

Description automatically generated

Figure : Weekdays/Weekends Crash

The donut chart above illustrates, a higher number of crashes during the weekdays as compared to weekdays. This observation suggests that there might be higher amount of traffic on roads primarily driven by work-related travel, commuters, school transportation etc. Therefore, extra resources and strategies needs to be developed during this period to minimize the crash rates.

**g. Road User**

A graph of a crash by road user

Description automatically generated A screenshot of a computer

Description automatically generated

Figure : Road User involvements during crash

The chart above shows the breakdown of crash incidents by various road users. Drivers, accounted for higher number of crashes at 21,688, followed by passengers with 9740 crashes. The higher number of crashes involving drivers suggest that proper driver education and training should be given, promote health and wellness, advance driver assistance system can be launched.

**h. Time of the day**

A graph of a pie chart

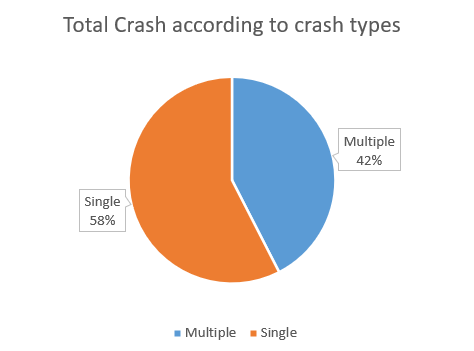
Description automatically generated A screenshot of a computer

Description automatically generated

Figure : Distribution of crash by time of the day

The donut chart illustrates that, majority of the crashes happened during daytime. These might be due to increased numbers of traffic volume, road conditions etc which might have attributed to this crash. When the patterns are recognized, proper resources can be allocated such as police patrols, emergency response teams which helps to prevent the crash numbers.

**I. Crash Types**

 A screenshot of a computer

Description automatically generated

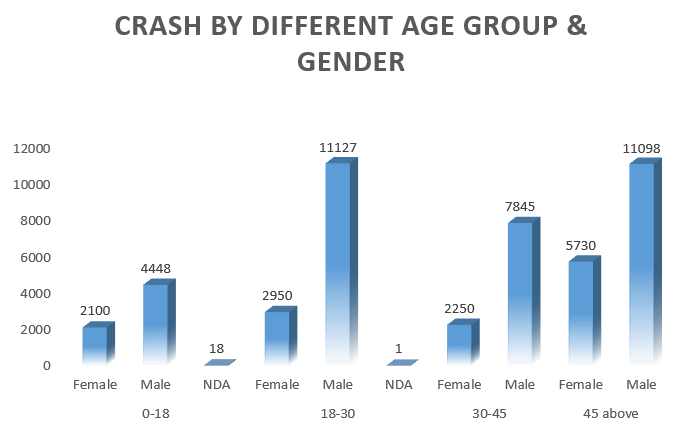
Figure : Total crash by various crash types

The pie chart provides a visual breakdown of crash types, indicating that single individuals accounted for approximately 57.54 percent of the incidents. Furthe analysis can be done, what has contributed to single driver crashes and emphasize can be given to safe driving practices, hazard recognition and defensive driving techniques.

# Diagnostic Analysis

**3.1. Bivariate Analysis**

**a. Age group & Gender**

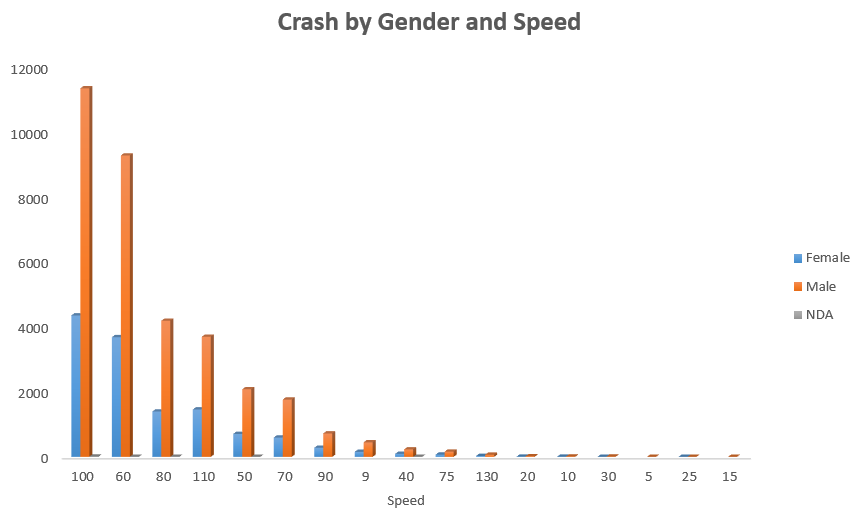
 A screenshot of a computer

Description automatically generated

Figure : Analysis of crash by age group and gender

The above chart clearly illustrates that, male population is more likely to be involved in the crashes across all age groups. Amongst the various age groups, younger male between the age 18-30 stand out as the highest-risk demographic, with significant figure of 11.12 thousand crashes. Following closely, 45 above male individuals recorded approximately 11.1 thousand crashes. Among the female population, 45 and above age group, stood the highest in all group, accounting for roughly 5.7 thousand crashes.

**b. Gender & Speed**

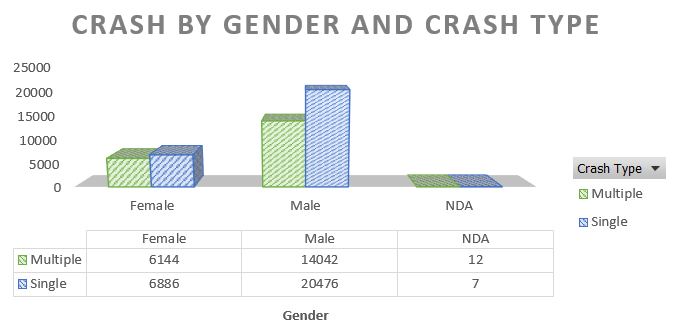
 A screenshot of a computer

Description automatically generated

Figure : Analysis of crash by gender and speed

The above figure depicts a compelling story, that male population experienced a higher number of crashes across all speed limits. When analysing the data, it is evident that crashes were more frequent at the speed of 100, where female accounted for 4.3 thousand crashes. However, it’s worth noting that the male population significantly surpassed this number, with 11.3 thousand incidents- more than double the female count. The second highest crashes occurred at the speed of 60 where male accounted for 9.2 thousand incidents whereas female had roughly one-third of male population, approximately 3.6 thousand crashes. This data highlights some trends in gender disparity in crash rates across various speed limits.

**c. Gender & Crash Type**

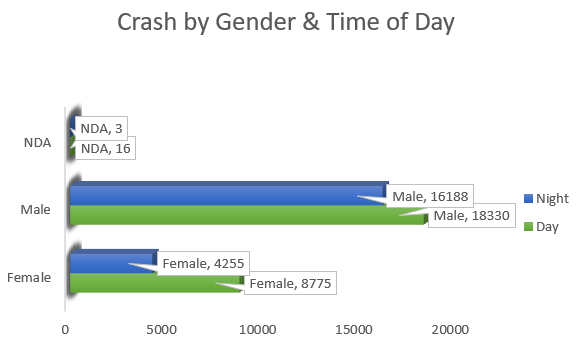
 A screenshot of a computer

Description automatically generated

Figure : Analysis of proportion of gender involved in different crash types.

The above picture highlights a significant disparity in involvement in single crashes where male had staggering number of 20 thousand incidents. This striking difference begs further questions and in-depth analysis to uncover underlying factors behind this trend. On the other hand, female population accounts for roughly one third of male total incidents, which accounts for 6.1 thousand incidents. It is noteworthy that the proportion of female involved in single and multiple crashes is nearly identical.

**d. Gender & Time of day**

 A screenshot of a computer

Description automatically generated

Figure : Analysis of proportion of gender involved in crash during night/day.

The figure above presents comprehensive overview of gender-related crash incidents during both daytime and nighttime. It is evident that males are consistently prominent group involved in crashes, with roughly equal number of accidents recorded during both day and night. Surprisingly, females exhibit a notable variation in crash involvement, likely to be more involved in daytime crashes which accounted for 8.7k crashes however, this number dropped roughly to half during nighttime crashes approx. 4.2k incidents. This intriguing difference between daytime and nighttime crashes in female helps to further investigate and analysis to better understand the underlying causes and potential preventive measures.

**3.2. Multivariate Analysis**

**a. Age group, Crash Types & State**

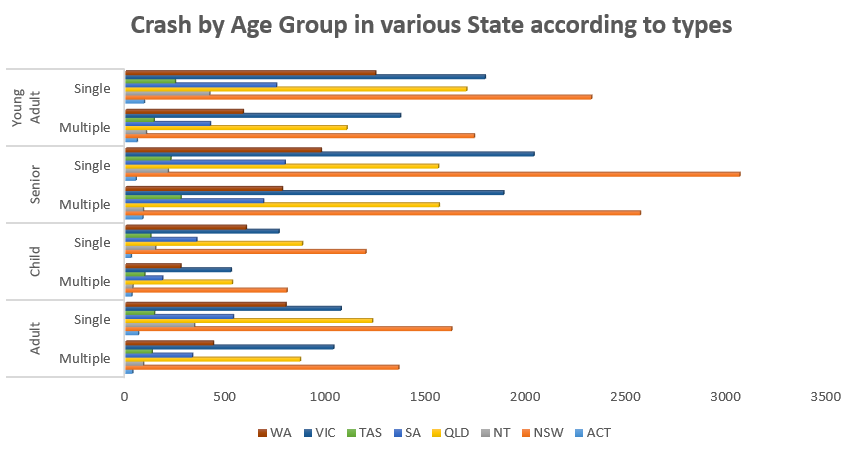
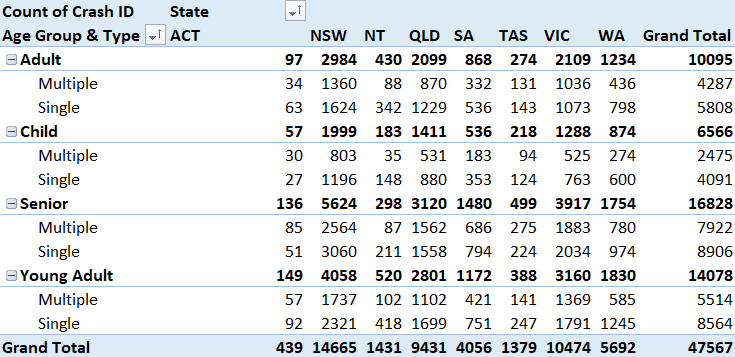
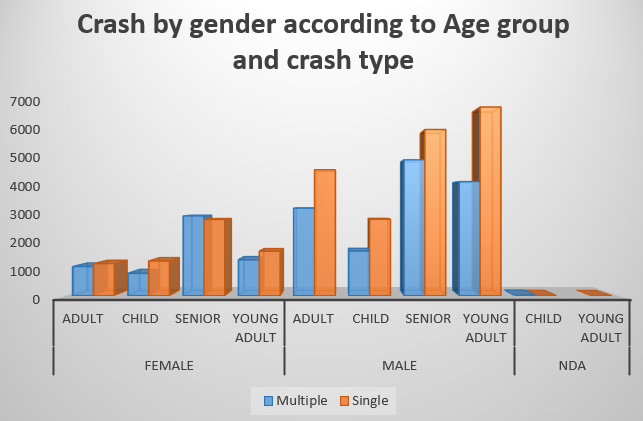


Figure : Analysis of different age groups involved in different crash types in various state.



The figure illustrates the distribution of crash incidents in Australia across different age groups, crash types and states. The state of NSW had a greater number of crashes incidents beating other regions. The senior population stands out as a significant contributor to these crashes, both in terms of their total involvement and prevalent in various crash types. Young adults represent the second highest demographics to involve in a crash where majority of the crashes were in NSW, totalling around 4058, followed by VIC at 3160. Notably, senior, and young adults, who were involved in single crash accidents accounted for more than 8 thousand incidents. In contrast, ACT had surprisingly few crashes, with a total of 439 incidents.

**b. Age Group, Gender, and Crash Type**

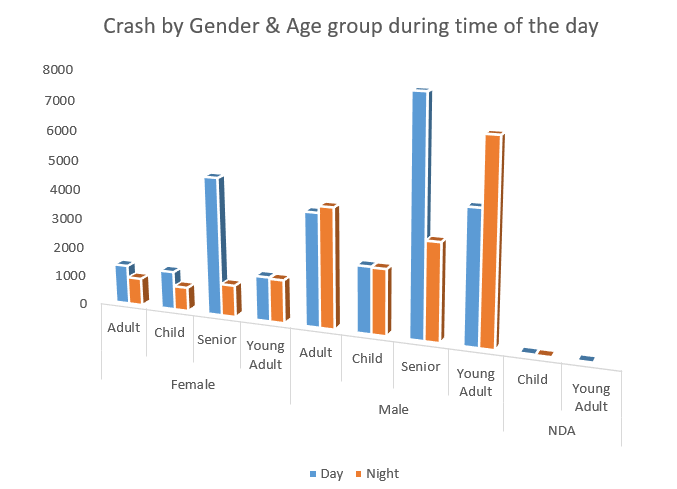
 A screenshot of a computer

Description automatically generated

Figure : Analysis of various age groups with their gender involved in different crash types.

The provided data highlights the significant details of the crash categorized by various age groups, gender, and crash types. The senior age group, in both genders, stands out as significant contributors to crashes, emphasizing the need for tailored strategies for older generations. Additionally, the data highlights young male as vulnerable groups with significant number of crashes. Therefore, to address this issues, target measures, and proper trainings and knowledge should be given to this age group which helps in reduction in crashes.

**c. Gender, Age group and Time of the day**

 A screenshot of a data

Description automatically generated

Figure : Analysis of crash by different age group, gender during day/night.

The provided data highlights the significant details of the crash categorized by various gender, age groups, and time of day. The distinction between day and night is particularly significant, with 20.4 thousand crashes occurring during night and 27.1 thousand crashes occurring during day. Notably, senior individuals in both age groups, were the higher contributor to the crashes which begs for importance of tailoring the safety measures for this age group. It is also noteworthy that male young adults were involved in crash more during the night as compared to other age group. This could be linked to specific lifestyle or behaviour patterns which begs for more analysis and develop some strategy which helps to mitigate risks.

# Business Questions

**4.1. Business Questions**

How can road safety laws and policies be improved, to better address the needs of vulnerable groups like young adults and seniors, and how can innovative strategies and campaigns be developed to effectively raise awareness and encourage safer road behaviours within these demographics?

**4.2. Business Questions using Descriptive Statistics**

* How does distribution of driving speeds vary across various age groups, and what trends in speed behaviours can be identified for the road safety?
* How can new speed limits and safety guidelines be recommended, by looking at the trends?
* What is the relationship between age and crash rates, and how can this data be leveraged to create precise and targeted safety measures tailored to specific age groups?

**4.3. Hypothesis Testing**

1. Analysis between speed and age.

Null hypothesis (H0): There is no significant correlation between age and crash speed.

Alternate hypothesis (H1): Age is corelated with the speed of vehicles involved in crashes.

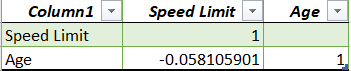


Figure : Hypothesis testing between age and speed.

The significance level (alpha) is 0.05 and p-value from the analysis is -0.05 from the testing, and p value is lower than the alpha. It implies that, there is a negative and very weak correlation between them. As age increases, the other variable, speed tends to decrease and vice-versa, but the relationship is not that strong. It implies that, null hypothesis can be rejected. Also, it suggests that relationship is so weak its unlikely to have any substantial predictive or explanatory power.

2. Analysis of crash with age.

Null hypothesis (H0): There is no significant corelation between age and road accidents.

Alternate hypothesis (H1): There is a significant corelation between age and road accidents.

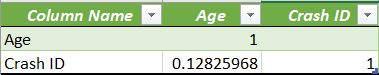


Figure : Hypothesis testing between age and crash.

The significance level (alpha) is 0.05 and p-value from the analysis is 0.128 from the testing, and p value is slightly greater than the alpha. It implies that, there is a positive and weak correlation between them. As age increases, the accidents tend to increase, but the relationship is not that strong. It implies that, null hypothesis cannot be rejected. Also, it suggests that there is no strong evidence to conclude that there is statistically significant relationship between age and frequency of accidents.

3. Analysis of crash with speed limit

Null hypothesis (H0): There is no significant association between speed limits and the frequency of road accidents.

Alternate hypothesis (H1): There is a significant association between speed limits and the frequency of road accidents.

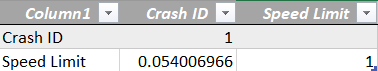


Figure : Hypothesis testing between crash and speed.

The significance level (alpha) is 0.05 and p-value from the analysis is 0.054 from the testing, and p value is slightly greater than the alpha. It implies that, null hypothesis cannot be rejected. Also, it suggests that there is no strong evidence to conclude that there is statistically significant relationship between speed limits and frequency of accidents.

4.Analysis of crash with the year.

Null hypothesis (H0): The number of accident crashes remains constant over the years.

Alternate hypothesis (H1): The number of accident crashes changes over the years.

A close up of a price tag

Description automatically generated

Figure : Hypothesis testing between crash with the years.

The significance level (alpha) is 0.05 and p-value from the analysis is 0.99 from the testing, and p value is greater than the alpha and indicates a linear relationship. It implies that, alternate hypothesis should be selected. Also, it suggests that there is strong positive relationship, to conclude that there is statistically significant relationship between them.

# Exploratory Data Analysis

1. **Young Adult crash timings investigations**

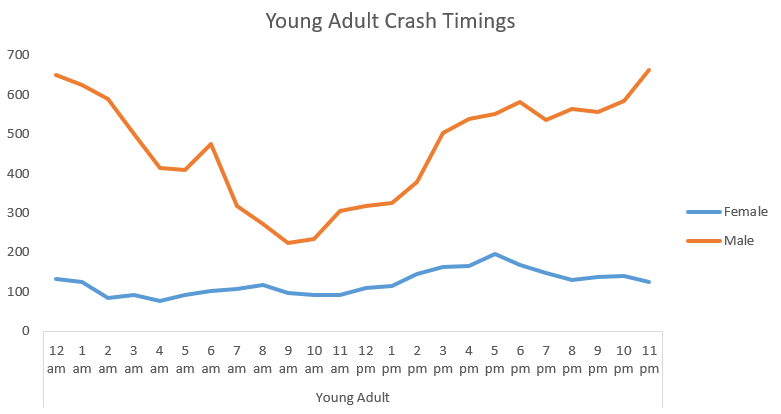


Figure : Young Adult Crashes

The figure illustrates a clear visual representation of the timings of crashes of young adults. Notably, young adults are more likely to be involved in crash during nighttime, particularly between 10 pm-12am. This observation shows interesting trend and intriguing questions about their social life, lifestyle choices or impaired visions during late hours. Further investigation and research can be done to explore the contributing factors to these crashes and targeted interventions can be made. Surprisingly, pattern for young female crash remains somewhat consistent, majority of them being from 4pm to 6pm. The highest crashes occurred from 4pm to 6 pm, which may be due to rush hour congestion, and work activities.

1. **Age related distribution of crashes at the speed of 100.**

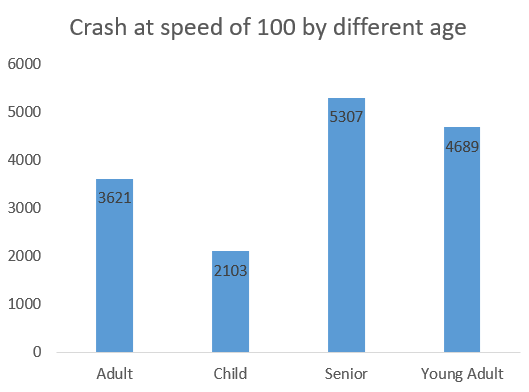
 Adding up to the previous analysis, it’s clear that crashes at the speed of 100 are a significant concern, because it is nearly one third of total crashes. This emphasizes the need for comprehensive safety measures. Among different age groups, senior and young adults were prominent in these types of crashes. This begs for the importance of effective education and awareness program, young driver trainings, creating data-driven interventions and improvement in the road. Apart from that, further analysis can be done to see where and why the crashes are prevalent at 100, and adjusting the speed limits, enforcing traffic laws, and enhancing safety measures in high-risk zones.

Figure xxx: Involvement in crashes during 100 speeds.

# Feature Selection

In the comprehensive analysis of road accident, impact of various variables has been explored on road accidents and five important variables has been identified: Speed limit, Age, Year, Gender, and Age Categories. Speed is selected because it has p-value marginally above the chosen significance level of 0.05. It also has practical importance in the domain of road safety. During thorough analysis, it was discovered speed had a meaningful impact on the crashes. While there is weaker corelation between age and crashes, it is essential to acknowledge that significance level alone does not always capture the complete picture. When p-value is slightly above the significance level it still might be meaningful in real-world applications. Also, year has been added because it has strong relationships, and suggests that it has substantial predictive power in explaining crashes. It also helps to identify the trend, and significant changes over the years. Gender is a categorical variable, even in the absence of corelation analysis, categorical variable can hold substantial insights into analysis. It has captured difference in driving behaviours at various time of day, crash involvement etc. Comprehensive understanding of crashes is possible when age is selected. Age categories was chosen for granular level understanding of crash patterns. These selections aids to provide comprehensive outlook on road accidents, ensuring that the analysis incorporates both statistically significant and practically relevant variables.

# Conclusions

Improving road safety laws and policies to better address the needs of vulnerable groups such as young adults and seniors, and creating innovative strategies and campaigns to raise awareness and encourage safer road behaviours within these demographics is complex but crucial task. The analysis revealed how different age groups approach speed limits and propensity for risky driving behaviours. Seniors were more likely to engaged in accidents, therefore further analysis can be done on what was contributing to these crashes. As seniors were more prone to crashes, further investigations can be done to identify contributing factors, and new policies can be introduced to these population such as regular health checkup and specialized driver training programs, to address age-specific challenges and emphasize safe driving practises. The senior driver might need to do fitness to drive test yearly.

Additionally, from the analysis, speeding was major issues during the crashes, most crash happening at the speed of 100 where young adults and seniors contributed most to the numbers. This clearly provides opportunity to look further where these crashes were occurring and why the number is higher and design a new speed limit so that the road will be safer for all. This analysis helps to consider reducing speed limits in areas with higher number of crashes. This analysis can be used by law enforcement agency, to increase enforcement of speed limits, targeting young adult and senior’s drivers. Young male adults were more prone to crashes between 10 pm to 12am which indicates that they have different social life, lifestyle choices etc as other groups. Police can be vigilant during this period and increase the patrolling numbers mainly targeting the young adult males which will help to bring the crash numbers down.

To conclude, the analysis done from this analysis serve as a robust foundation to create targeted interventions and create targeted market campaigns. By looking at various specific actions and implementing targeted policies, the crashes number might reduce significantly and make road safe.

# References

Hurwitz, J., Kaufman, M., & Bowles, A. (2015). *Cognitive Computing and Big Data Analytics*. New York: Wiley

Greasley, A. (2019). *Simulating Business Processes for Descriptive, Predictive, and Prescriptive Analytics*. Boston: deGruyter

# Appendix

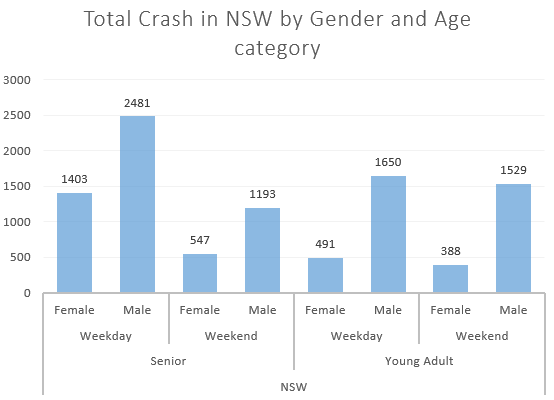


Figure : Crash in NSW by Gender and Age group

A graph of a crash

Description automatically generated

Figure : Weekends Crash by Timings