Project Report: Road Crash Fatality Analysis and Visualization Using Tableau & PostgreSQL

Course: CITS5504

Student: Kundan Jha

Date: April 2025

1. Introduction

The overall objectives of this project are to build a data warehouse using real-world datasets and to carry out a basic data mining activity, specifically association rule mining. This project focuses on analyzing Australian road crash fatality data to support decision-making, improve road safety, and uncover key spatial and temporal patterns using data warehousing and visualization.

2. Datasets and Problem Domain

Road Safety Context

Every nation is working toward lowering road tolls for safer transportation. In 2023, Australia recorded a fatality rate of 4.8 deaths per 100,000 people, significantly higher than Iceland's 2.1, which is the world's safest. As of July 2024, 761 road deaths were recorded in Australia, with a 12-month total of 1327 — a 10% increase from the previous year.

Project Datasets

- **ARDD: Fatal Crashes—December 2024 (XLSX)**
- **ARDD: Fatalities—December 2024 (XLSX)**
- **GeoJSON boundary files** (LGA, SA4, STE)
- **Dwelling Count Data (2021)** from ABS

GeoJSON files were used for spatial mapping in Tableau. The crash and fatality data were the core inputs to the data warehouse, while dwelling data supported deeper demographic analysis.

3. Data Warehouse Design, Implementation, and Usage

Following Kimball's 4-step dimensional modelling approach:

A. Process Being Modeled

Fatal crashes and associated details in Australia during December 2024.

B. Grain

One row per crash in the fact table.

C. Dimensions Chosen

- Date (with hierarchy: Day → Month → Year)
- Time (Hour → Time of Day)
- Location (Suburb → LGA → State)
- Road (Segment → Type)
- Vehicle (Model → Type)
- Event (Weather → Cause)
- Dwelling (Suburb → LGA)
- Population (LGA → Region)

These dimensions support multidimensional OLAP-style querying.

D. Measures Identified

- Number of crashes
- Number of fatalities
- Number of dwellings

Schema and StarNet Design

A **star schema** was selected due to its simplicity and performance for slicing and dicing. It allows for fast aggregations on large volumes of data. Also, it is easier to maintain and understand.

Entity-Relationship Diagram:

Each business question (e.g., crashes by area, time, weather) corresponds to a footprint over the StarNet.

Business Queries Answered:

- Which LGAs have the highest number of fatal crashes?
- Do rural areas experience more fatalities than urban?
- Is there a pattern between vehicle type and fatality?
- Are fatal crashes more frequent at specific times?

4. ETL Process: Data Cleaning and Preprocessing

ETL Tools and Environment

- Python (pandas, psycopg2)
- PostgreSQL for staging and final tables
- Tableau for final visualization

ETL Steps

1. **Extract:** Loaded Excel files into staging tables.
2. **Transform:**
    a. Removed duplicates
    b. Normalized fields (e.g., consistent suburb names)
    c. Derived fields like day of week, time of day
    d. Removed corrupt data
    e. Validated LGA names for GeoJSON join
3. **Load:** Inserted clean data into PostgreSQL star schema.

Data Loss Control

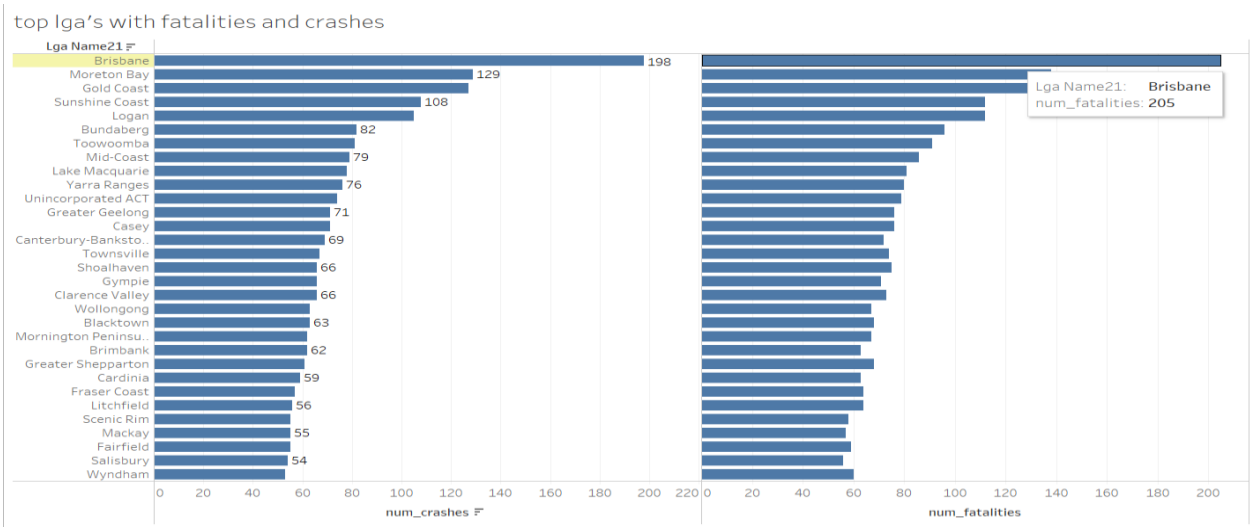
- Less than 2.5% of rows were removed (mainly due to null critical fields).

5. Visualizations and Key Insights

The following visualizations were created using Tableau and are presented below for visual analysis and interpretation.

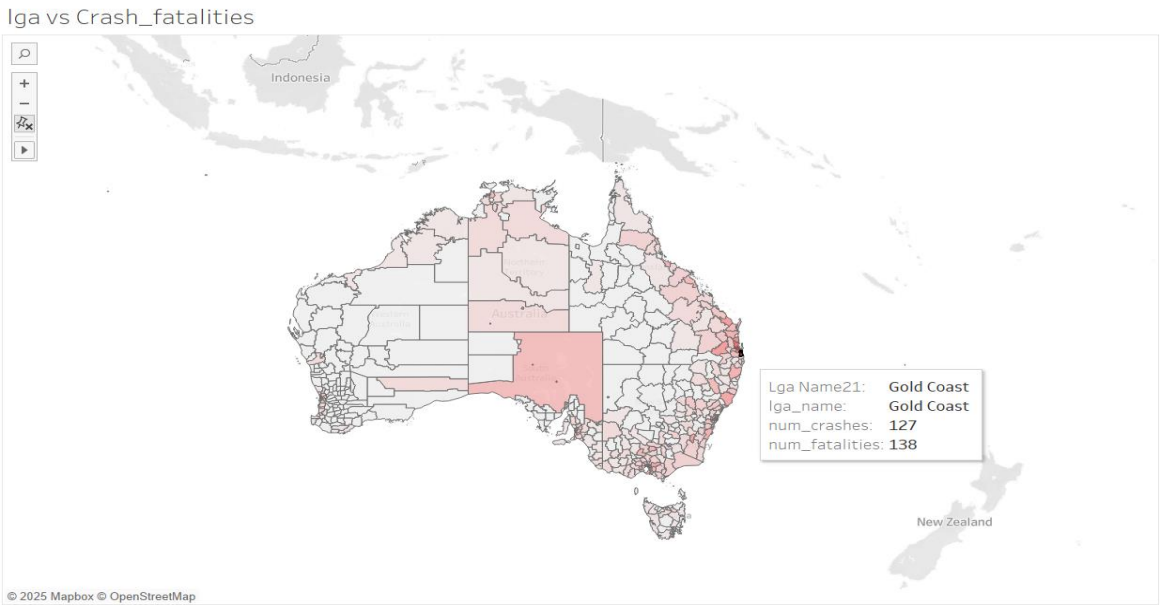
Visualization 1: Bar Chart – Top 10 LGAs by Fatal Crashes

- Highlights spatial disparities; some LGAs report far more fatalities.



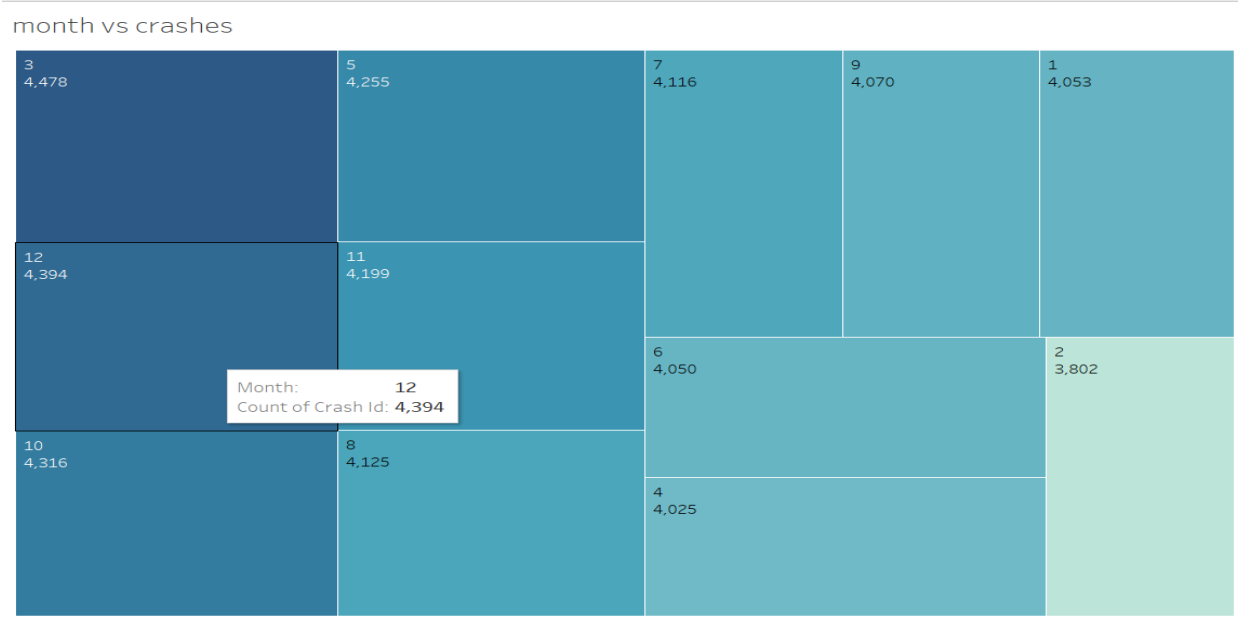
- Highlights spatial disparities; some LGAs report far more fatalities.
- Darker regions on the map reveal high-risk areas.
- This visualization supports prioritizing safety improvements in densely populated LGAs.
- Brisbane stands out as the LGA with the **highest number of crashes and fatalities**, indicating it as a critical focus area for road safety intervention.

Visualization 2: Choropleth Map – Fatal Crashes by LGA



- **Brisbane** remains the most severely affected LGA, both in terms of absolute crash volume and fatalities.
- Other prominent LGAs include **Gold Coast**, **Greater Geelong**, and **Moreton Bay**.
- Many high-fatality LGAs are found in **urban coastal areas**, possibly due to dense traffic and high-speed corridors.
- It is evident from the visual that Eastern Australia has more accidents than Western Australia.

Visualization 3: Time Series – Monthly Fatalities



- December shows a spike, correlating with holiday travel.
- March has many crashes as well due to rainy Season in different parts of country
- Whole Summer as well more accidents are observed due to increase in traffic activity in this time

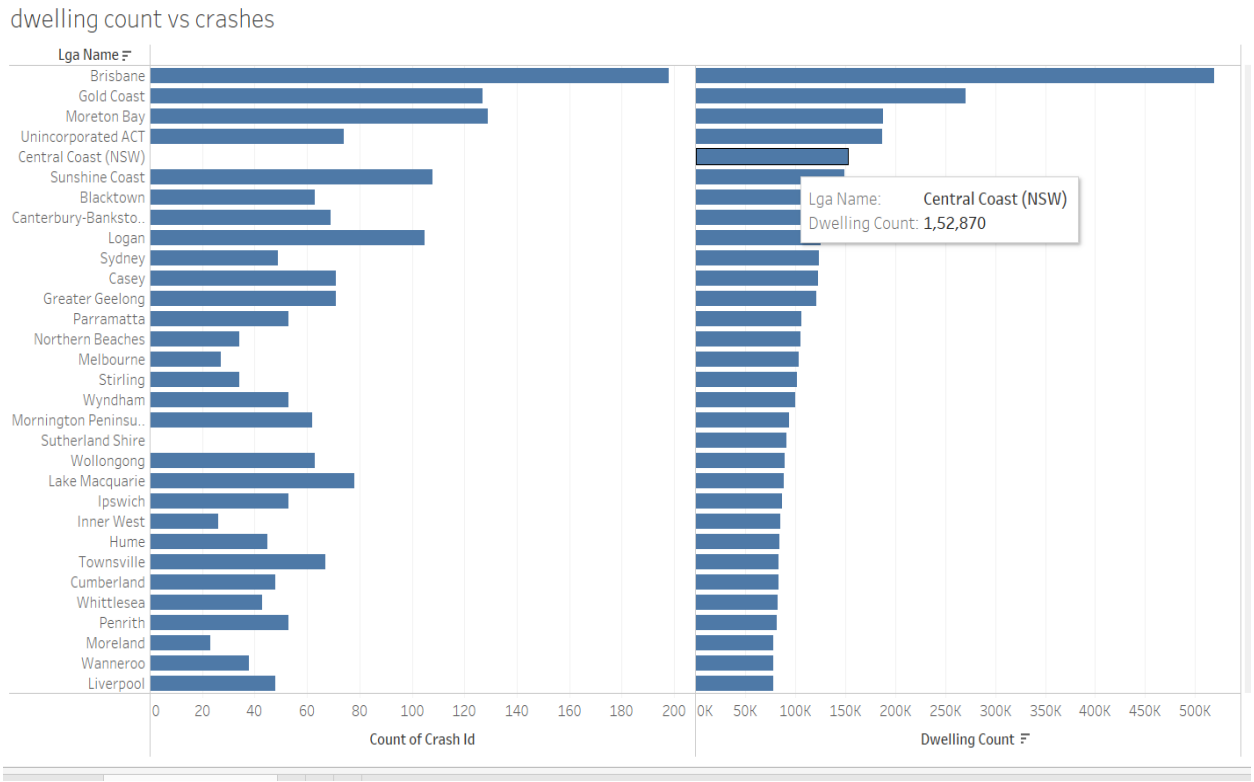
Visualization 4: Urban vs Rural Area Impact

- Regional Areas are prone to accidents as well maybe due to non-maintainence of roads mainly and people not following traffic rules.

Visualization 5: Vehicle Type Contribution

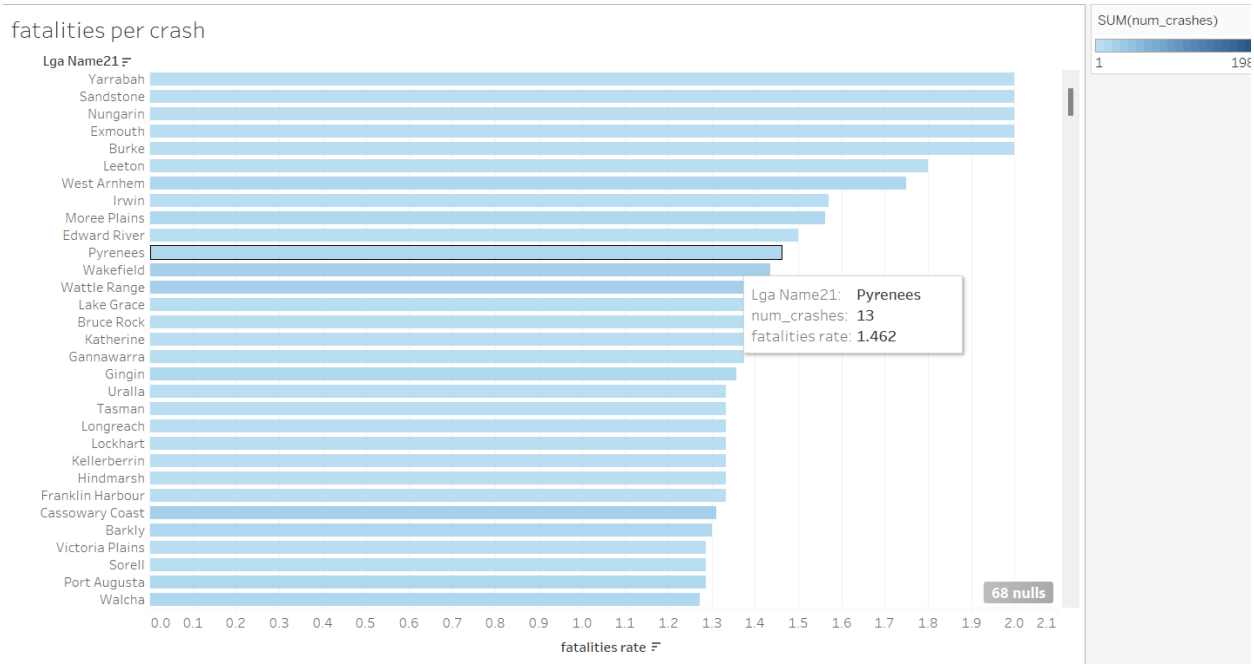
- Shows distribution of fatal crashes by vehicle type.
- It can be seen above that articulated trucks involvement is highest among bus, rigid trucks and Articulated trucks.
- More experienced drivers should be allowed to drive these trucks
- Turns should be designed keeping in mind these kind of trucks.

Visualization 6: Bar chart of dwelling count vs crashes



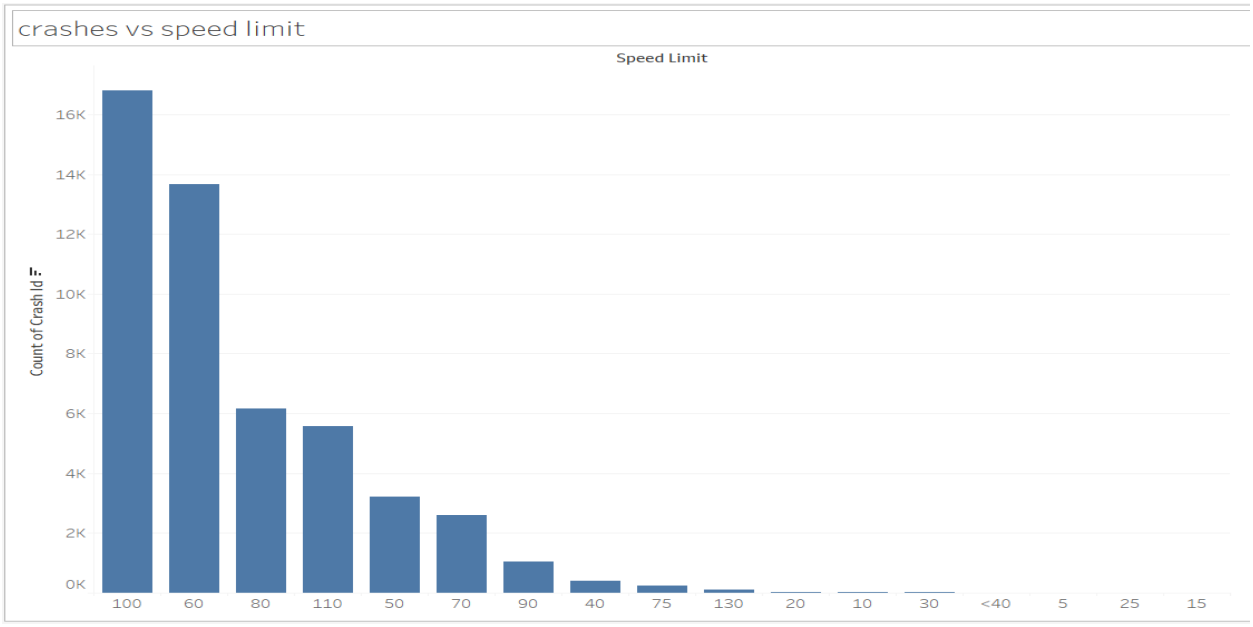
- It shows the dwelling count and crash count for different LGAs.
- It is evident that LGAs with more dwelling count has more crashes.
- Surprisingly **Central coast** has no accidents although it comes in top 5 in terms of dwelling count

Visualization 7: Bar chart of fatalities per crash



- Shows the LGAs with maximum fatalities per crash with the one having more crashes colored dark.
- It can be noticed that though Pyrenes has lesser fatalities rate but it has more crashes. Hence these are crucial LGAs.
- Yarrabah leading this chart with maximum fatalities rate.

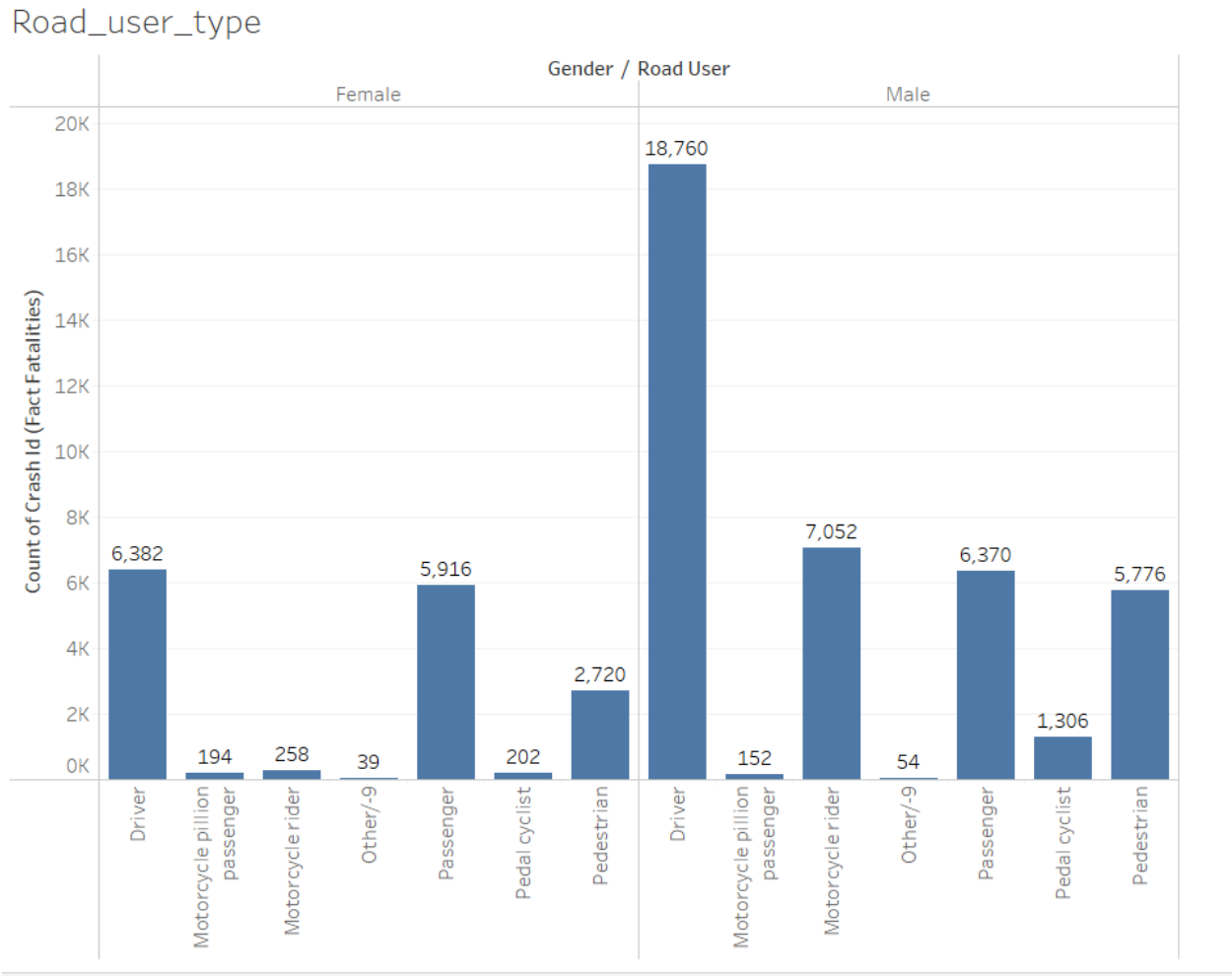
Visualization 8: Histogram of crashes in different speed limit



- Shows the number of crashes at different speed limits.

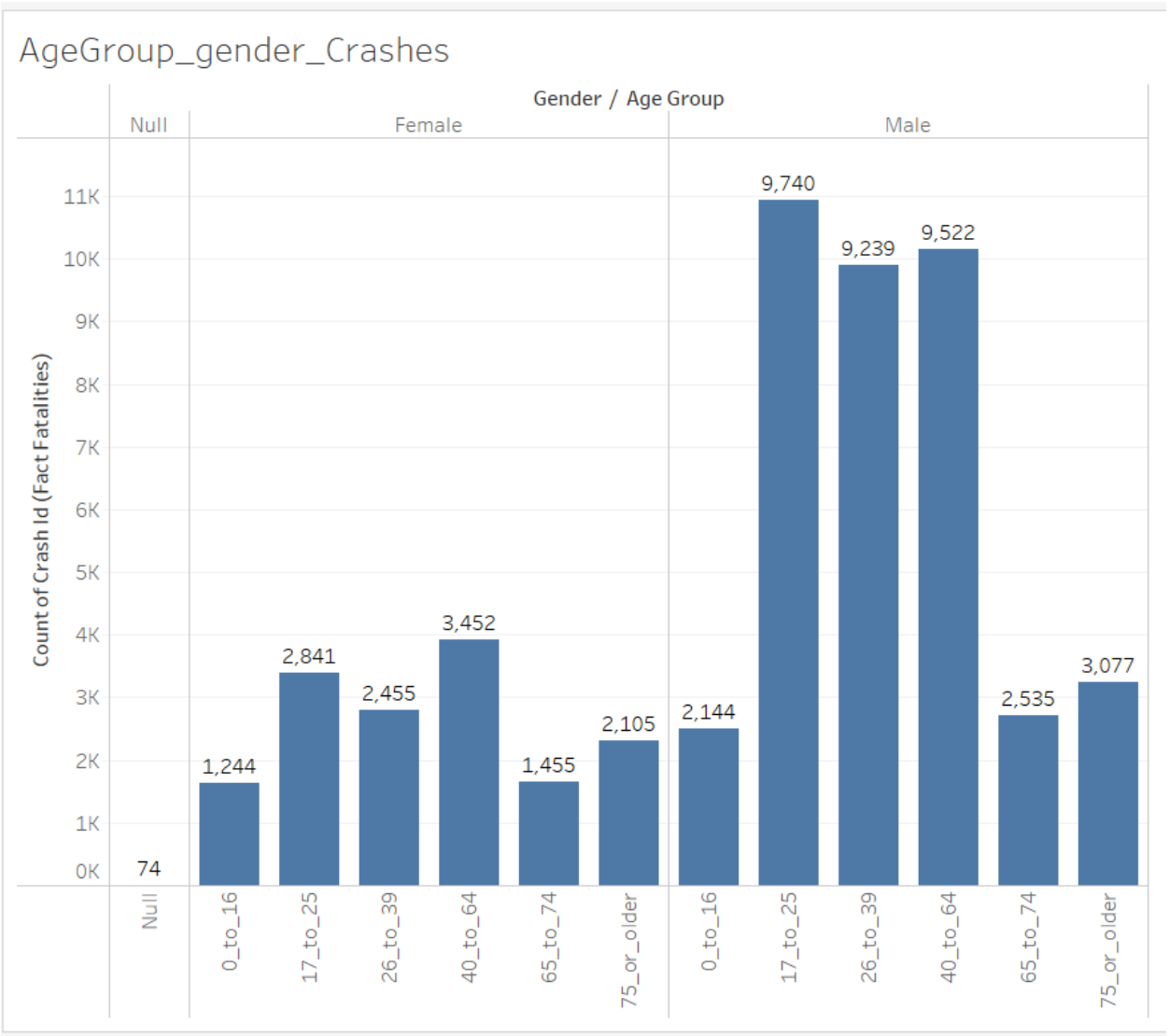
- It can be observed that majority of crashes are happening at 100 and 60.
- It is difficult to control vehicle at turns when you are at 100. Also, majority of the accidents might be happening at entry and exit of freeways.

Visualization 8: Histogram of crashes for different road user



- Shows the number of crashes for different road user.
- Male motorcyclists are more prone to accidents due to not having good control in high speeds
- This can be reduced by encouraging motorcyclists to maintain safe speed and organising safe driving campaigns for them

Visualization 8: Histogram of crashes for different age group



- Shows the number of crashes for different age group/gender.
- **Males aged 17–64** are overwhelmingly represented in fatal crashes
- **Females** have significantly lower crash counts across all age groups.
- Male motorcyclists are more prone to accidents due to not having good control in high speeds
- Targeted **awareness campaigns** and **behavioral interventions** for males in the 17–64 range could significantly reduce fatalities.

Business Query Answers (Q01–Q06)

++Queries are attached in sql file.
++Starnet diagram with footprints is also attached

Q01. Which LGAs have the highest number of fatal crashes?

Answer: *Brisbane has the highest number of crashes and fatalities*.

Recommendations: Prioritize infrastructure audits and enforcement strategies in high-risk zones.

Q02. Urban vs Rural Fatalities Comparison

Answer: *Rural areas show a higher fatality-per-crash ratio compared to urban zones*.

Recommendations: Invest in better lighting, emergency response, and public awareness in rural areas.

Q03. Fatal crashes by month of year

Answer: *Crashes are spread across the year, but peaks align with march and December. This can be due to increased traffic activity in summer and rainy season accidents due to poor weather conditions*.

Recommendations: Schedule special road patrols and campaigns in Summer. Roads should be maintained better in rainy days.

Q04. Crash count by speed limit

Answer: *Most crashes happen in areas with 100 km/h speed limits*.

Recommendation: Freeway entry and exits should have proper lane marking and should be properly designed.

Recommendations: Reinforce signage and speed calming in suburban/urban corridors.

Q05. Public holiday crash impact

Answer: *Both Christmas and Easter periods see higher crash counts*.

Recommendations: Schedule special road patrols and campaigns in these periods.

Q06. Crash count by gender and age group

Answer: *Males aged 17–64 are most involved in fatal crashes, peaking between 17–39*.

Recommendations: Target this demographic with high-impact education and driver behavior programs.

6. Association Rule Mining

++python script for this is attached in a seperate .py file

Algorithm Used

- **Apriori Algorithm** implemented using Python's `mlxtend` library.
- Evaluation metrics: **Support**, **Confidence**, and **Lift**.

Implementation Process

The crash dataset was cleaned to retain key categorical attributes such as:

- Crash Type
- Vehicle Involvement (Bus, Articulated Truck, Heavy Rigid Truck)
- Road Type
- Remoteness Area
- Time of Day
- Day of Week
- Number of Fatalities

Rows were transformed into transactional format and encoded using the Transaction Encoder. Rules were generated using the Apriori algorithm.

Python Code – It is attached in submission documents

Top Rules with 'Fatal' as Consequent

Below are the top 10 rules extracted from the crash dataset where the **consequent includes 'Fatal'**. These rules are sorted based on **Lift** and **Confidence**, key metrics in association rule mining.

Support indicates how frequently the rule appears in the dataset.

Confidence measures the likelihood of the consequent occurring when the antecedent is present.

Lift evaluates the strength of the rule compared to random chance.

The analysis focused on rules with high Lift (>2.0) and Confidence (>0.65).

Antecedents	Consequents	Support	Confidence	Lift
('Yes', 'Day')	('Fatal', 'Weekday', 'Multiple')	0.0608	0.6588	2.38
('Yes', 'No', 'Day')	('Fatal', 'Weekday', 'Multiple')	0.0605	0.6582	2.37
('Yes', 'Day')	('Fatal', 'Weekday', 'No', 'Multiple')	0.0605	0.6554	2.37
('Unknown', 'Yes', 'Day')	('Fatal', 'Undetermined', 'Multiple')	0.0538	0.7766	2.30
('Unknown', 'Yes', 'No', 'Day')	('Fatal', 'Undetermined', 'Multiple')	0.0536	0.7758	2.30
('Unknown', 'Yes', 'Day')	('Fatal', 'No', 'Undetermined', 'Multiple')	0.0536	0.7732	2.30
('Yes', 'Day', 'Undetermined')	('Fatal', 'Unknown', 'Multiple')	0.0538	0.7705	2.29
('Yes', 'No', 'Day', 'Undetermined')	('Fatal', 'Unknown', 'Multiple')	0.0536	0.7697	2.29
('Yes', 'Day', 'Undetermined')	('Fatal', 'Unknown', 'No', 'Multiple')	0.0536	0.7672	2.28
('Weekday', 'Unknown', 'Yes')	('Fatal', 'Undetermined', 'Multiple')	0.0605	0.7416	2.20

Interpretation

- The top rules suggest that crashes occurring during the **day**, involving **uncategorized vehicle conditions (Yes, No, Unknown)** and **multiple road user types**, show higher fatality likelihood.
- Fatalities also appear more frequently under combinations involving **weekday crashes**, suggesting commuter risk.

Insights

- Daytime crashes involving multiple or ambiguous conditions (like 'Yes', 'No', 'Unknown') are strongly associated with fatal outcomes.
- Improving the classification and recording of vehicle and crash types can significantly enhance safety analytics.

Recommendations Based on Rules

1. **Improve rural lighting and signage** – Especially at intersections and high-crash corridors.

2. **Enhance daytime awareness campaigns** – Since a significant number of fatal crashes also occur during the day, more public messaging can help.
3. **Investigate unknown or unclassified crash attributes** – A higher share of fatal crashes appears in records with vague classifications; efforts should improve crash investigation and documentation.

- **High-Risk LGA Audits:** Focus infrastructure reviews on top 10 LGAs.
- **Night-Time Enforcement:** Increase patrols and alcohol checks.
- **Targeted Education:** Run area- and vehicle-specific campaigns.
- **Data-Driven Resource Allocation:** Use heatmaps to prioritize emergency service distribution.

8. Conclusion

This project successfully met the goals of CITS5504: creating a functional data warehouse, building insights through Tableau visualizations, and applying association rule mining in Python. The integrated approach reveals actionable insights for reducing Australia's road toll and aligns with global road safety objectives.

9. References:

GeeksForgeeks: for referring python etl script
chatgpt: for taking template of association rule minning script.

End of Report