TECHNICAL REPORT
PROJECT 1

UNIT: CITS5504

Date: 08 April 2025

Student 1: Cedrus Dang - ID: 24190901
Student 2: Bhavesh Agarwal - ID: 23933845

Abstract

This technical report presents the design, implementation, and usage of a data warehouse for analysing Australian road fatality data in 2024. Following the Kimball dimensional modeling methodology [1], the system integrates historical fatality crash data and a population with multiple dimensions into a Star schema. The resulting analytical cube supports OLAP operations with multidimensional queries and is connected with PowerBI software for visualization analysis. Additionally, association rule mining was conducted to derive insights related to road user type.

I. Introduction

The data sources are three Australian data sets:

1. Fatal Crashes - December 2024.
2. Fatalities - December 2024.
3. Population estimates by LGA, Significant Urban Area, Remoteness Area, Commonwealth Electoral Division and State Electoral Division, 2001 to 2023.

Road safety remains a significant concern for Australia, with an average fatality rate of 4.8 deaths per 100,000 people in 2023 [2]. As of July 2024, Australia's national road toll is sitting at 761 lives lost so far this year, while the 12-month total of 1327 road deaths is up 10% on the year prior [3].

This project seeks to assist decision-makers by developing an infrastructure for Online Analytical Processing (OLAP) to support the government and public understanding about the situation and insights of Australia road safety and create suggestions to lower traffic risks using a Starnet schema, a Kimball’s data warehouse model with one fact table and no neted dimensions in dimension [1]. The project focuses on the following key objectives:

1. Build a data warehouse to store historical data on fatal crashes.
2. Use the created data warehouse to present key insights with a dashboard.
3. Support decision-making with data mining techniques.
4. List and describe a few suggestions for improving road safety for the government.

The tech stack that is implemented into this project are:

- Jupyter Python notebook is used for EDA pre-development, ETL, and data mining.
- PostgresSQL for the database server with SQL as query language.
- Tableau/PowerBI data analysing.
- The library for ETL is Pandas, and for data mining, it is the FP-Growth Algorithm.

II. Methodology

A. Dimensional Design

The data warehouse was designed based on Kimball's Four Steps [1] as follows:

1. **Select the business process:** Recording of crash fatalities.
2. **Declare the grain:** Each row in the fact table represents a single fatality case.
3. **Identify the dimensions:** Time, Location, Gender, Age, Speed Limit, Crash Type, Road User, Road Type.
4. **Identify the facts:** The number of fatalities, derived by aggregating individual fatality records.

The dimension selection was based on Kimball's definition: "Dimensions provide the 'who, what, where, when, why, and how' context surrounding a business process event. Dimension tables contain the descriptive attributes used by BI applications for filtering and grouping the facts. With the grain of a fact table firmly in mind, all the possible dimensions can be identified. Whenever possible, a dimension should be single valued when associated with a given fact row." [1]

Some sample questions that can be answered by the chosen dimensions are:

1. Time: What are the trends in road fatalities across different months and days of the week?
2. Location: Which states have the highest fatality counts and rates when adjusted for population?
3. Gender: How does the distribution of fatalities differ between male and female road users?
4. Age: Which age groups are most frequently involved in fatal crash incidents?
5. Speed Limit: What is the distribution of road fatalities across various posted speed limits?
6. Crash Type: Which crash types result in the highest number of fatalities?

7. Road User: How do fatality rates vary by road user categories?
8. Road Type: What types of roads are most commonly associated with fatal accidents?

While many dimensions can be created by this definition, some dimensions and attributes were not inserted into the data warehouse as follows:

1. **Lower-than-state location code dimension:** Only the State code was used as the other location codes were missing values higher than 30%.
2. **Junk dimension for bus, heavy rigid truck, articulated truck involvement**: In pre-deploy data exploration, those flag values are highly dominated with negative values at >90%. Additionally, heavy rigid truck involvement also accounts for 36% of the missing data. Based on that, the future values of those flag attributes are expected to be significantly low and will heavily impact the data mining process, creating uninformative rules. Therefore, they were decided not to be implemented into the data warehouse and only be treated as potential data for future deployment if needed.
3. **Christmas period, Easter period:** The holiday flags are very useful for analysis; however, in this case, the date is not provided—only the day of the week is available. This prevents the ability to perform continuous time series analysis. In addition, more than 97% of the flag values are negative, introducing a strong imbalance that increases the risk of misleading outcomes. Therefore, these attributes were excluded from the current data warehouse schema and are only considered potential inputs for future deployments if complete timestamp data is available.
4. **Day of week (Weekend flag):** The ABS weekend flag is defined as 'Weekday' from 6:00 AM Monday to 5:59 PM Friday. However, this business-centric definition does not reflect real-world commuting patterns, as weekday activities and heavy traffic often extend until 11:59 PM Friday. It is unrealistic to assume that individuals abruptly stop commuting or change behavior immediately after 6 PM. Therefore, treating Friday evenings as part of the weekend introduces potential misclassification of crash timing. Instead, a more detailed analysis should be conducted by looking at the time and day of the week for any related analysis.

The Starnet Query Model of the design is shown below:

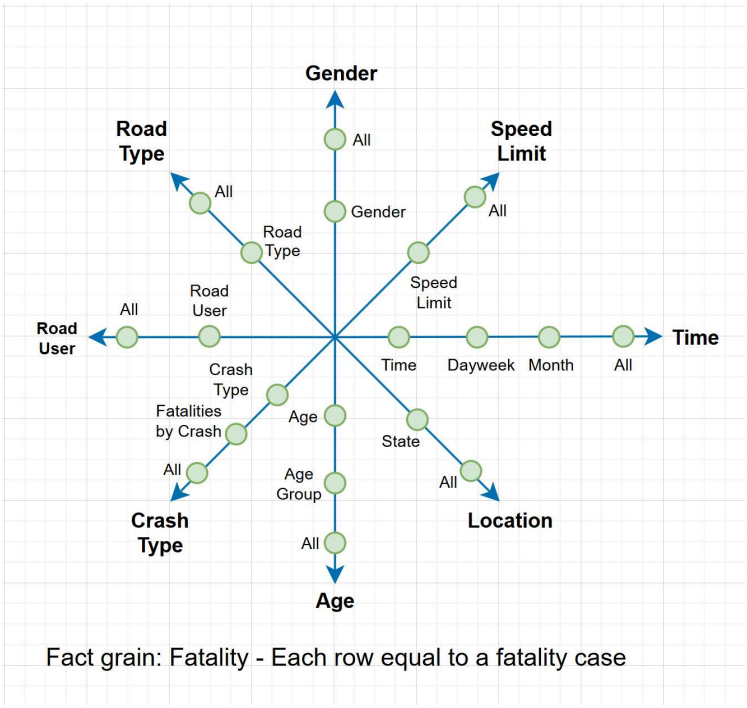


Diagram 1: Starnet Query Model

B. Star Schema Logical Design

Below is the structure of the star model in this data warehouse

Fact Table: fact_fatality (Primary Key: crash_sk)

Dimensions:

1. dim_time (Primary Key: time_sk, Attributes: time, dayweek, month, year)
2. dim_location (Primary Key: state_id, Attributes: state_name, population)
3. dim_gender (Primary Key: gender_sk, Attributes: gender)
4. dim_age (Primary Key: age_sk, Attributes: age, age_group)
5. dim_speed_limit (Primary Key: speed_limit_sk, Attributes: speed_limit)
6. dim_crash_type (Primary Key: crash_type_sk, Attributes: crash_type, fatalities_by_crash)
7. dim_road_user (Primary Key: road_user_sk, Attributes: road_user)

Each dimension uses a surrogate key for referential integrity, join performance, and decoupling from natural keys to ensure stability and save computational resources (Kimball's suggestion [1]), except for state_id. The state_id remains a natural key to prevent potential errors and mismatches in joint operations with current and future population data updates. All foreign keys in the fact_fatality table refer to these surrogate keys and ID.

Additionally, while the data scope in fatality in 2024 and population in 2023 were without a year attribute, the year attribute in the time dimension was not removed to ensure the data warehouse's scalability for future development, where new data with years were added to the data warehouse. However, in this project query and analysis, this preserved attribute will not be used.

The star schema implemented in PostgreSQL is visualised below:

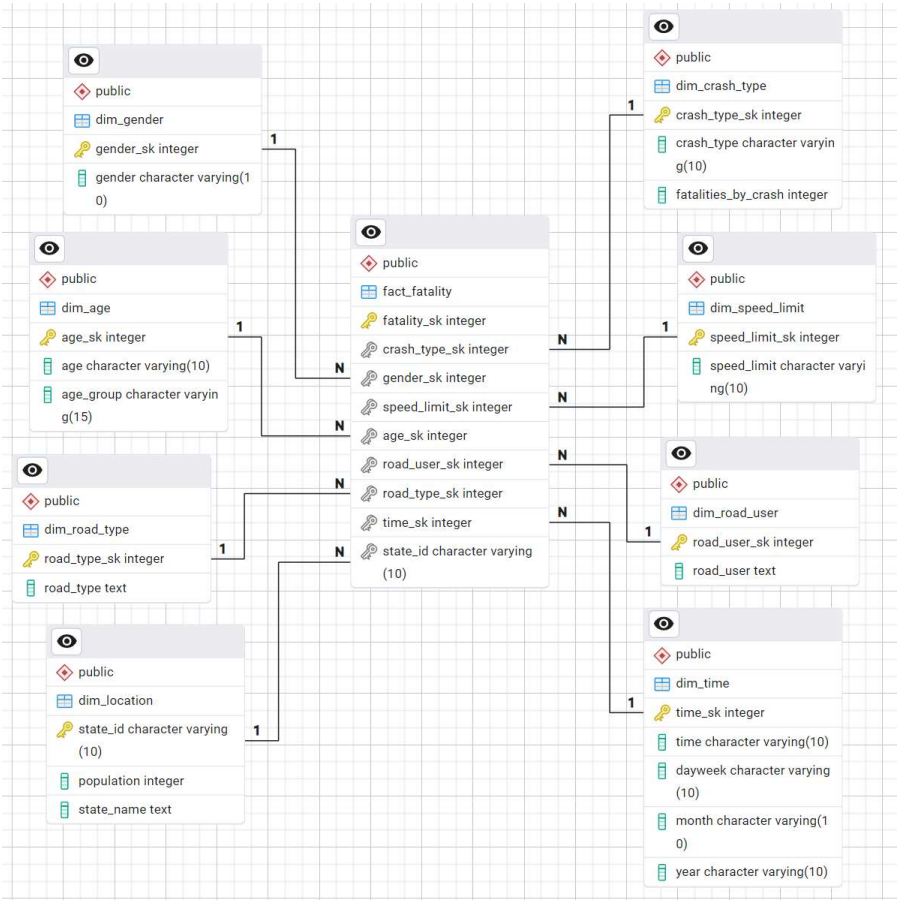


Diagram 2: Starnet Query Model

III. Implementation Details

A. Data cleaning, preprocessing, and ETL process:

ABS's data often contain multiple meta rows and meta tabs, which contain unuseful information and can create corruption when used. Therefore, a manual pre-processing explanatory data analysis (EDA) has been conducted using Excel. Then, many further, simple EDA processes were also performed during the ETL using Jupyter Notebook

functions. However, most of this code has been removed from the source code to ensure clarity and is only left with the most important functions to ensure the industrial-grade quality of the source code.
In the ETL pipeline, cleansings of meta rows, columns, and tabs were embedded to ensure automatic flow was enabled, with only a limited decrease in run-time efficiency.
The technical methodology of the ETL is doing all steps in temporary memory, using Jupyter Python variables to store data to increase efficiency without creating extended storage files.

1. Extract

Data source:

- bitre_fatal_crashes_dec2024.xlsx
- bitre_fatalities_dec2024.xlsx
- ABS Population Estimates.xlsx

Tools: openpyxl, pandas
The Extract only selects 3 tabs that contain the used data and then removes the index, dictionary, and other irrelevant data tabs. As the data set is small and has multiple meta rows, Python data framework Pandas was chosen instead of PySpark to focus on EDA and fast transformation ability instead of Big data and PySpark's efficiency.

2. Transformation

There are n steps in the Transformation section:

1. Clearing the meta rows and not data rows. Additionally, there are no meta columns.
2. Handling missing values with -9 mask value for missing values, NaN, Unknown, other and similar values. Specifically, those values will be changed to Unknown to ensure clarity while not impacting the analysis's result as it is similar in attributes.
3. Standardizing the time column by changing it from HH:MM:SS to HH, from full time with hours, minutes and seconds to hours for short. This grouping will ensure the analysis will not be too micro details later.
4. Unpivolt population data to long data, filter the year to the latest one (2023), and drop unused columns, leaving only population and state code for future join processes.
5. Enrich location data with the state's name instead of just the state's code to create stability for mapping. This prevents software and libraries from mistaking Australia's state code for other nations.
6. Create dimension by extracting columns from the main data set.
7. Resetting indexes creates surrogate keys, and copying indexes to new columns requires using Pandas's reset function twice.

8. Remove duplicated rows in dimension tables to group them to categorical, finish creating dimension tables.
9. Join dimension tables back to the main data to add surrogate keys to the main data.
10. Create a surrogate key for fatality cases in the fact table.
11. Create a fact table by extracting surrogate keys and state_id.

The ETL step did not include removing any rows containing empty values, as this is predicted to not affect further analysis, and empty value-containing rows account for more> than 30% of the data.

Used tools: pandas

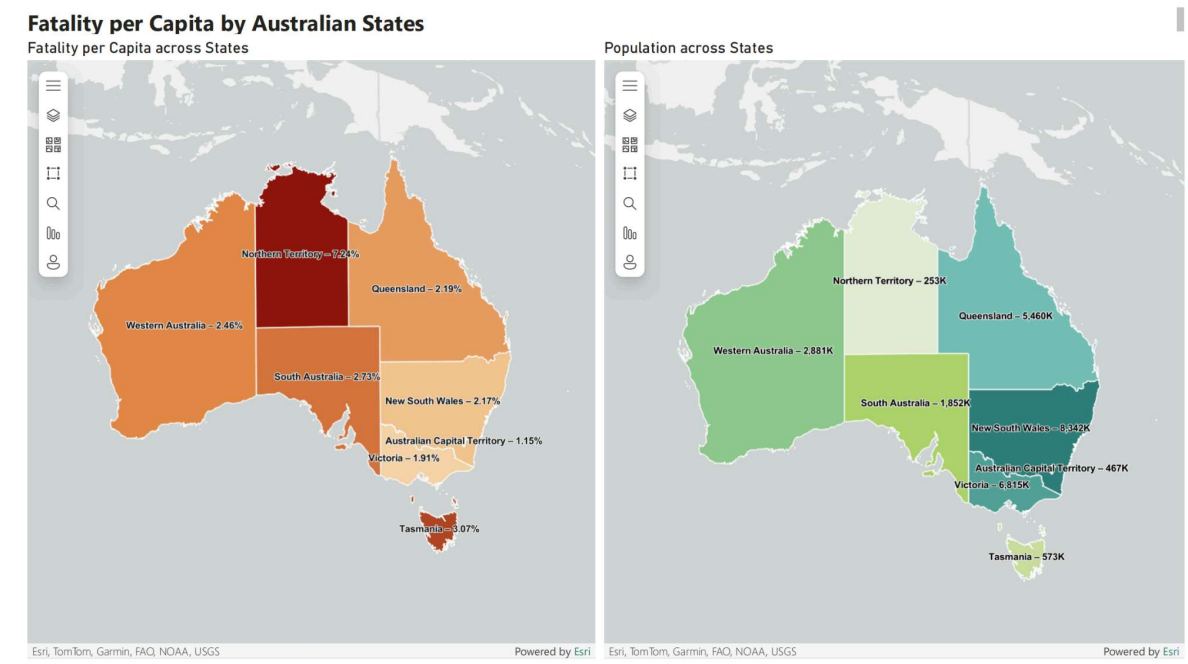
3. Load

The Sqlalchemy library creates a connection to a pre-prepared PostgresSQL server. To ensure the best cybersecurity practice, sign in using secret keys in the .env file. A SQL query code contained in an SQL script will be called to clear and create tables, and data in dimension variables storing temporary memory will then be loaded directly to the database.
An optional function has also been embedded inside the source code that provides extended .csv files for dimension and fact tables for any alternative usage, disabled as the default.

IV. Query Design and OLAP Capabilities

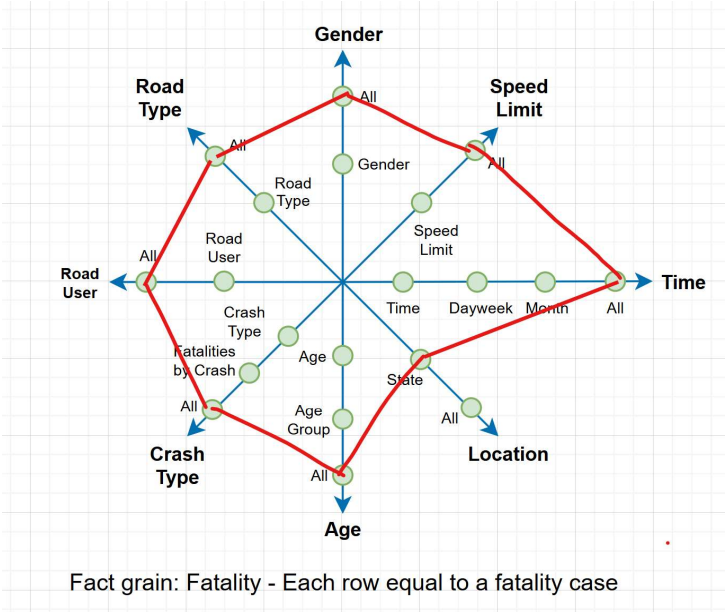
A. Visualisation of query footprint and results

Below are five PowerBI dashboards with the application connected directly to the PostgresSQL data warehouse and their query footprints using automatic query functions from the application:

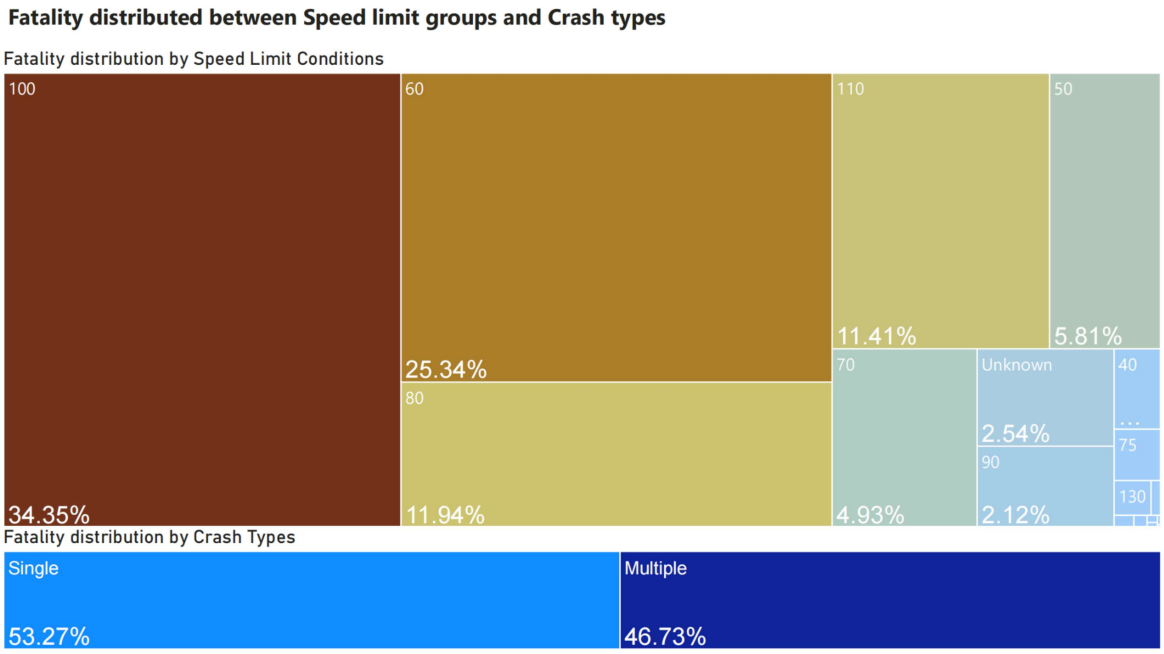


Dash Broad 1: Fatality per Capita be Australian States

The map shows that places with fewer people have higher fatality rates than others. For example, the Northern Territory has the smallest population but the highest fatality rate. In contrast, New South Wales and Victoria have large populations and lower fatality rates. This suggests remote areas might have higher road risks than densely populated places.

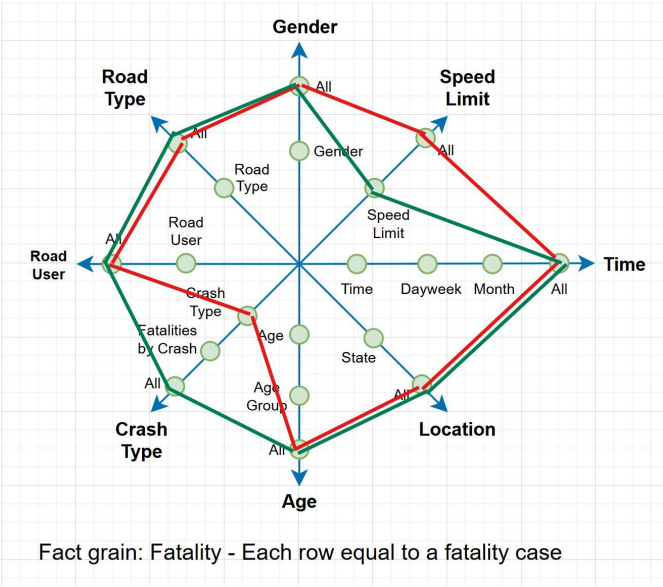


Query Footprint 1: Fatality per Capita be Australian States

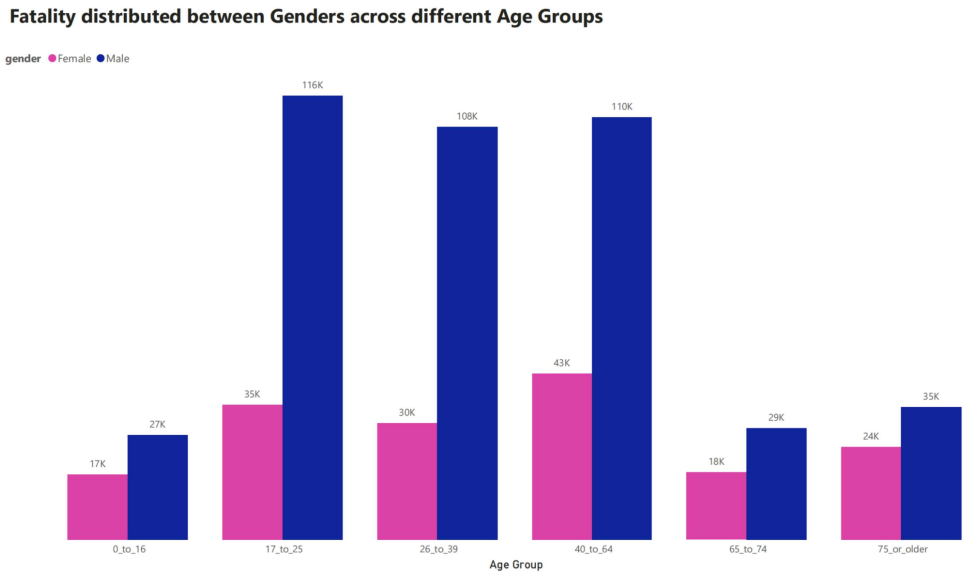


Dash Broad 2: Fatality distributed between Speed limit groups and Crash types

Fatality distributed between Speed limit groups and Crash types shows that the top four speed zones with the highest fatality shares are 100 km/h (34.35%), 60 km/h (25.34%), 80 km/h (11.94%), and 110 km/h (11.41%). Additionally, the near-equal distribution between single-vehicle (53.27%) and multi-vehicle (46.73%) crashes suggests that self-crash accidents have the same rate of fatality as multi-vehicle crashes.

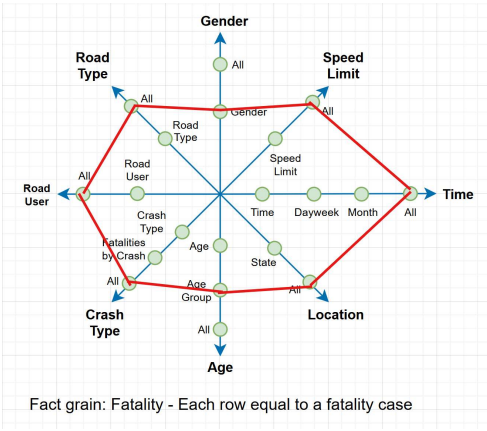


Query Footprint 2: Fatality distributed between Speed limit groups and Crash types

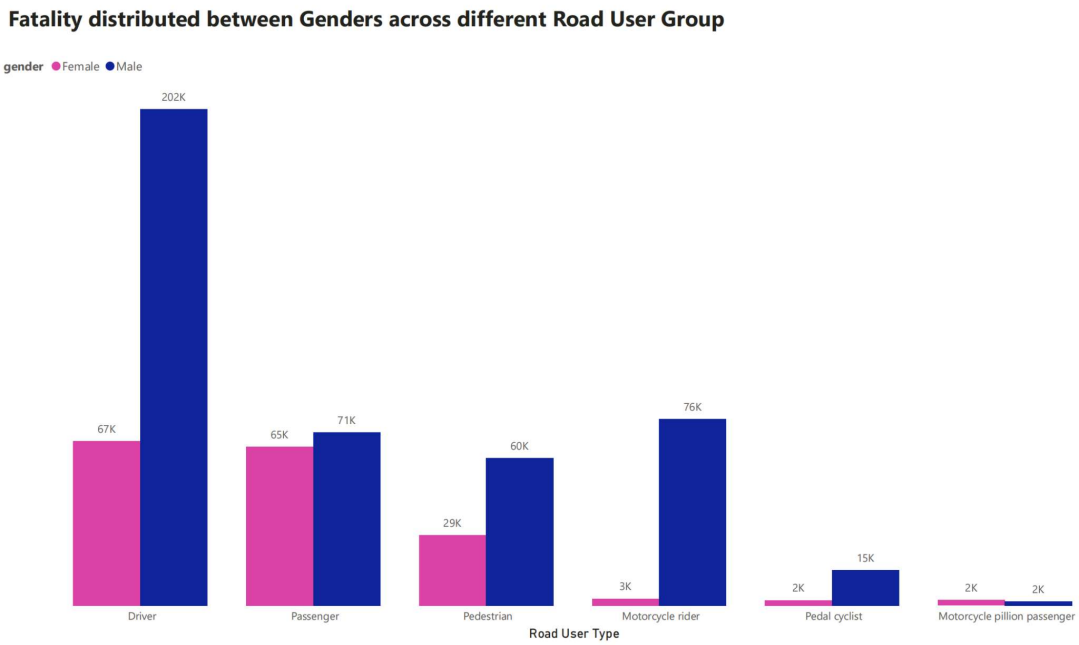


Dash Broad 3: Fatality distributed between Genders across different Age Groups

Fatality distributed between Genders across different Age Groups shows that male fatalities are consistently higher than female across all age groups, with the highest counts in the 17–25 (116K), 26–39 (108K), and 40–64 (110K) male groups. The gap is most pronounced in the young to middle-aged brackets, indicating that males in these age ranges are at significantly higher risk of fatal crashes compared to females. However, an additional analysis of the gender demographics of the Australian population is necessary.

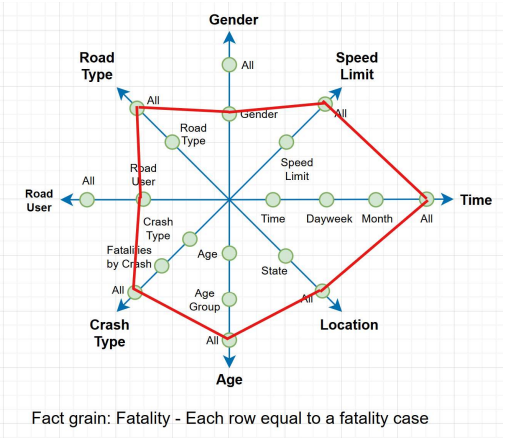


Query Footprint 3: Fatality distributed between Genders across different Age Groups

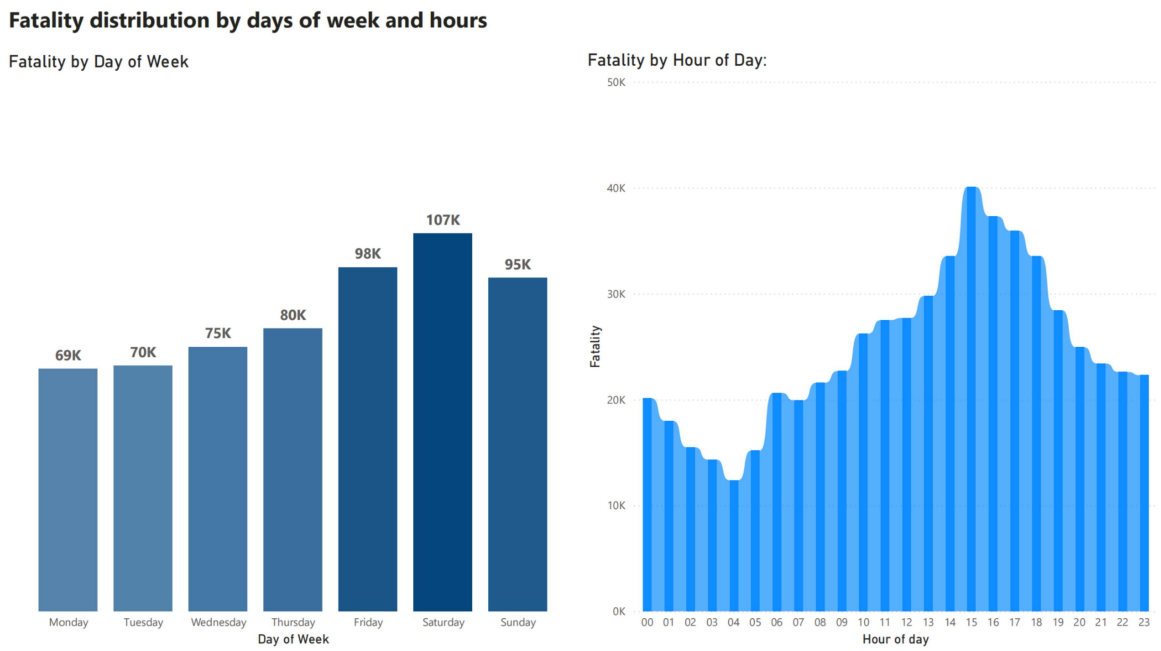


Dash Broad 4: Fatality distributed between Genders across different Road User Groups

Fatality distributed between Genders across different Road User Groups shows that male fatalities are significantly higher among drivers (202K vs. 67K), indicating a strong gender imbalance in fatal crash involvement for those in control of vehicles. Passenger fatalities are nearly equal between males (71K) and females (65K), suggesting that risk exposure differs by role rather than presence. Among motorcycle riders, males account for an overwhelming majority (76K vs. 3K), and even pedestrian deaths are notably higher in males (60K vs. 29K). These patterns suggest a potential relationship between gender and fatality likelihood, possibly linked to behavioral, exposure, or role-based differences on the road.

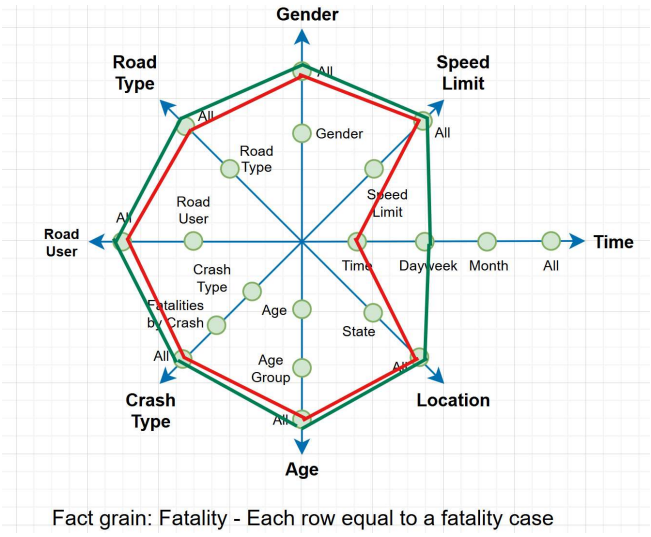


Query Footprint 4: Fatality distributed between Genders across different Road User Groups



Dash Broad 5: Fatality distribution by days of week and hours

Fatality distributed by days of the week and hours shows a clear rise towards the end of the week, peaking on Saturday (107K) and remaining high on Friday (98K) and Sunday (95K), suggesting increased exposure to risk during weekend activities. The hourly trend reveals a gradual buildup in fatalities from early morning, peaking sharply between 15:00 and 17:00, then slowly declining. This pattern indicates the influence of accumulated fatigue, end-of-day travel, and potentially higher risk behavior during off-work hours, aligning with the weekend spike.



Query Footprint 5: Fatality distribution by days of week and hours

B. Association Rule Mining

Mining methodology: FP-Tree Construction: The algorithm begins with two dataset scans. The first determines the frequency of items, eliminating those below the minimum support threshold and ordering the rest by frequency. In the second scan, transactions are processed to build the FP-tree, merging common prefixes and tracking item counts (Han, Kamber, & Pei, 2012).

Frequent Pattern Mining: The algorithm recursively generates frequent itemsets by constructing conditional FP-trees for each item. This divide-and-conquer strategy continues until all patterns are discovered. [4][5]

Key Evaluation Metrics

- **Support:** Proportion of transactions containing both antecedent and consequent.
- **Confidence:** Likelihood of the consequent appearing when the antecedent is present.
- **Lift:** Strength of the association; a value >1 indicates a positive correlation.

Justification for Use:

- **Efficiency:** For a dataset with ~ 57,000 records, FP-Growth significantly outperforms Apriori, particularly under lower support thresholds, by eliminating the costly candidate generation process.
- **Scalability:** While the dataset is not “big data,” FP-Growth scales well, making it ideal for future expansions or high-frequency patterns.
- **Dimensional Suitability:** The algorithm efficiently handles high-dimensional, sparse categorical data, such as the current dataset with 14 crash-related attributes.
- **Memory Management:** Although FP-tree construction can be memory-intensive, it uses less than Apriori.

The hyperparameters and filter rules for the data mining process:

1. Minimum support is 0.15, meaning minimum support must be >= 0.15, where the pattern has to be appearch in 15% of the data
2. Minimum confidence is 0.5, meaning the confidence level of this pattern must be >= 50%
3. Filter: Only take rows with the consequent field containing solely "**Road User**" K - rules is >= 1 with k as the number of top rules to extract, ranked by lift and then confidence.

The top k-rules of the mining association:

1. **Rule 1:**

- **Association:** Antecedent is Speed limit = 100 and Consequent is Road User = Driver
- **Confidence** is 0.5697
- **Lift** is 1.2581
- **Interpretation:** This rule indicates that when the accident happens in a speed limit area of 100 km/h, there is approximately a 57% chance that the fatality is a driver. The lift value greater than 1 confirms a positive association, meaning the occurrence of the antecedent (Speed Limit = 100) is positively correlated with the consequent (Road User = Driver).

2. **Rule 2:**
    - Association: Antecedent is Speed limit = 100 and Road Type is “Unknown” and consequents is Road User = Driver.
    - Confidence is 0.5609
    - Lift is 1.2386
    - **Interpretation:** This rule combines a 100 km/h speed limit with an unspecified road type, as the “Unknown” value is for missing and unknown values, into the antecedents. Under these conditions, the probability of the road user being a driver in the fatality case is around 56% if the accident happens in a 100km/h speed limit zone. The confidence and Lift are slightly lower than Rule 1. This suggests that known road types correlate more strongly with Road users as drivers in fatality cases, although not very significantly, compared with incidents in unknown ones.

Key Insights

- **Fatality as a Driver in High-Speed Zones:** Both rules point to a consistent pattern: An incident in a 100 km/h speed limit zone is more likely to involve a driver fatality with confidence as 56%. This insight could be seen as a sign of a need for investigation in such areas. Although the confidence level is not significantly higher than 50%, it still shows a possible relationship between the consistency and consequency with high lift.
- **Influence of Road Type Data:** The second rule introduces the condition of an “Unknown” road type. While driver involvement remains likely, the reduced lift and confidence suggest that better road type specification could refine the understanding of incident patterns.

Recommendation to government: Based on the patterns identified through the FP-Growth algorithm and the subsequent analysis of frequent item sets associated

with crash incidents, the following recommendations are proposed to improve road safety outcomes:

1. **Enhanced Driver-Focused Safety Campaigns for 100 km/h Zones:** Given that both rules indicate a higher likelihood of drivers being involved in incidents where the speed limit is 100 km/h, the government should implement targeted safety campaigns specifically aimed at drivers in these zones. These campaigns could emphasize safe driving practices at higher speeds.
2. **Improved Data Collection for Road Type Information**: Rule 2 highlights incidents occurring in 100 km/h zones where the road type is "Unknown." To gain a more precise understanding of the factors contributing to these incidents and to potentially develop more targeted interventions, the government should invest in improving data collection protocols and standardization for road type information to reduce the unclear road type values and gain better insights.
3. **Further Investigation into Factors Contributing to Driver Involvement in 100 km/h Zones:** While the rules indicate a higher association of drivers with incidents in 100 km/h zones, the confidence levels are only around 56-57%. This implies that many incidents at this speed limit involve other road users. The government should conduct further in-depth studies to identify the specific factors contributing to incidents involving drivers in these zones.

References

[1] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.

[2] European Commission. "2023 figures show stalling progress in reducing road fatalities in too many countries". Mobility and Transport. https://transport.ec.europa.eu/news-events/news/2023-figures-show-stalling-progress-reducing-road-fatalities-too-many-countries-2024-03-08_en (accessed Feb. 24, 2025).

[3]S. Guthrie. "The country with the safest roads in the world". Drive. https://www-drive-com-au.webpkgcache.com/doc/-/s/www.drive.com.au/caradvice/the-country-with-the-safest-roads-in-the-world/ (accessed Feb. 24, 2025).

[4] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in *Proc. 2000 ACM SIGMOD Int. Conf. Management of Data*, Dallas, TX, USA, 2000, pp. 1–12.
[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Burlington, MA, USA: Morgan Kaufmann, 2012.
[6] OpenAI, *ChatGPT (Mar 14 version)* [Online]. Available: https://chat.openai.com/chat. Accessed: Apr. 11, 2025.