PROJECT 1 - DATA WAREHOUSE DESIGN

1. Introduction

Background

Since the beginning of record keeping in 1925 there have been over 190,000 lives lost on Australian roads [1] Today road safety and rising death tolls remains a critical public health, economic and social issue in Australia. The current fatality rate is 4.8 deaths per 100,000 people. In comparison to Iceland, a global key leader in road safety, which has a rate of 2.1 deaths per 100,000 people, Australia significantly falls behind globally in Road Safety. As of July 2024, Australia's road death toll has reached 761 lives lost, with the 12-month total being 1,327 road deaths: a 10 percent increase from last year. Today, the Australian Government aims to reduce road trauma deaths and serious injuries by 2050 [1]

Purpose of the Data Warehouse

This project aims to develop a data warehouse that is comprehensive in its analysis of historical data. Our database consists of fatal crashes and resultant fatalities that have occurred in Australia. By integrating and structuring crash and fatality data, our aim is to provide a robust analysis that enables:

1. Identification of patterns and risk factors in fatal road crashes
2. Analysis of temporal, geographic, and demographic dimensions of road fatalities
3. Evidence-based recommendations for targeted road safety interventions
4. Support and insight into government decision-making on resource allocation, regarding the safety of our roads

Datasets

This analysis utilizes two key datasets from the Australian Road Deaths Database (ARDD):

1. **Fatal Crashes - December 2024**: Contains information about crash events, including location, timing, and circumstances

2. **Fatalities - December 2024**: Contains detailed information about each individual fatality, including demographics and road user type

These datasets provide a source of information on road fatalities and crashes. This allows us to perform a multi-dimensional analysis across factors such as location, time, vehicle types, and road user demographics.

2. Dimensional Modeling Process

Utilizing Kimball's four-step dimensional modeling approach [2], we created a data warehouse to analyze fatal crashes in Australia. This structured approach enabled us to transform raw crash and fatality data into an analytical framework that supports complex multi-dimensional analysis, with the goal of improving road safety outcomes across Australia [2].

Step 1: Identify the Process Being Modeled

The primary business process being modeled is **road fatality incidents** in Australia. This process represents a critical event where one or more individuals lose their lives in a traffic crash. Each fatality event contains attributes about:

- The crash event (location, time, conditions)
- The individuals involved (demographics, role in the crash)
- The vehicles involved (types, involvement factors)

The business goal is to analyze these events to identify patterns and risk factors that can inform road safety interventions. As stated by the Department, "Delivering safer roads is not solely a government problem, nor solely a transport problem – everyone's actions matter" [1]. By modeling this process, we can answer key questions like:

- Which factors contribute most significantly to fatal crashes?
- Where and when are fatalities most likely to occur?
- Which demographic groups are at highest risk?
- What combinations of factors create the most dangerous scenarios?

Step 2: Determine the Grain at Which Facts Can Be Stored

We determined that the most appropriate grain for our fact table is **one row per person killed in a crash**. This approach aligns with Kimball's recommendation to define grain at ‘the most atomic level possible’ [2]. This person-level grain was selected as:

1. It preserves detailed information about each fatality (age, gender, road user type)
2. It allows for demographic analysis that would be lost at a higher level of aggregation
3. It maintains the relationship between individual fatalities and the crash event

4. It supports both person-level queries and crash-level aggregations

Each fact record represents a single fatality linked to the crash event, allowing analysis at both individual and crash levels. This granularity provides maximum analytical flexibility while maintaining data integrity, which is essential for understanding the complex factors contributing to Australia's road death toll [2].

Step 3: Choose the Dimensions

Based on the business questions and available data, we identified eight key dimensions that provide the analytical depth needed to understand fatal crash patterns. As Kimball notes, dimensions fall out of the question: 'how do businesspeople describe the data that results from the business process?'[2].

Step 4: Identify the Numeric Measures for the Facts

For our fact table, we identified the following numeric measures:

1. **Crash Count** (always 1)
    - Counts crashes in a distinct manner when aggregating
    - Enables normalization for per-crash metrics
2. **Number Fatalities** (from the crash record)
    - Total fatalities associated with the crash
    - Supports severity analysis and multi-fatality crash identification

These measures, combined with the dimensional attributes, support analytical operations such as:

- Count and percentage calculations
- Comparative analysis between categories
- Severity analysis (fatalities per crash)
- Temporal trend analysis
- Geographic distribution analysis

By focusing on these core metrics, we maintain a straightforward fact table structure while enabling complex analytical capabilities through dimensional attributes.

3. Concept Hierarchies for Fatalities Data Warehouse

We created concept hierarchies that support multi-dimensional analysis across all facets of road fatalities such as:

1. Date Dimension (DateDim)

Primary Key: Date Key
Hierarchy: Year → Quarter → Month → Day Name → Day Type → Time of Day

- **Year**: (1989-2024) - Top level for temporal analysis
- **Quarter**: (1-4) - Groups months into quarters
- **Month**: (1-12) - Monthly breakdown of data
- **Day Name**: (Monday-Sunday) - Specific day of the week
- **Day Type**: (Weekday, Weekend) - Categorizes days into weekday/weekend
- **Time of Day**: (Day, Night) - Distinguishes between daytime and nighttime crashes
- **Additional Flags**:
    - Christmas Period (0/1)
    - Easter Period (0/1)
    - Is Holiday (0/1) - Combined indicator for any holiday period

2. Location Dimension (LocationDim)

Primary Key: Location Key
Hierarchy: State → Remoteness → Road Type → SA4 Name → LGA Name

- **State**: (NSW, VIC, QLD, SA, WA, TAS, NT, ACT) - Top geographic level
- **Remoteness**: (Major Cities, Inner Regional, Outer Regional, Remote, Very Remote) - Population density classification
- **Road Type**: (National or State Highway, Arterial Road, Local Road, Collector Road) - Road classification
- **SA4 Name**: Statistical Area Level 4 regions - Mid-level geographic grouping
- **LGA Name**: Local Government Area - Most granular level of geographic detail

3. Speed Dimension (SpeedDim)

Primary Key: Speed Key
Hierarchy: Speed Category → Speed Limit

- **Speed Category**: (Low, Medium, High, Missing) - Grouped speed ranges

- Low: ≤ 50 km/h
- Medium: 51-90 km/h
- High: > 90 km/h

- **Speed Limit**: Actual numeric speed limit in km/h

4. Age Dimension (AgeDim)

Primary Key: Age Key
Hierarchy: Age Group

- **Age Group**: Categorization of individual ages

  - 0_to_16: Children and young teenagers
  - 17_to_25: Young adults
  - 26_to_39: Adults
  - 40_to_64: Middle-aged adults
  - 65_to_74: Younger elderly
  - 75_or_older: Elderly
  - Missing: Unknown age

5. Road User Dimension (RoadUserDim)

Primary Key: Road User Key
Hierarchy: User Type → Road User

- **User Type**: High-level categorization

  - Vehicle Occupant: Includes drivers, passengers, motorcycle riders
  - Non-Occupant: Includes pedestrians and cyclists
  - Other: Other road users not in the above categories
  - Unknown: Missing road user type

- **Road User**: Specific road user role (Driver, Passenger, Pedestrian, etc.)

6. Crash Type Dimension (CrashTypeDim)

Primary Key: Crash Type Key
Hierarchy: Crash Type

- **Crash Type**:

  - Single: Single vehicle crashes
  - Multiple: Multiple vehicle crashes

7. Gender Dimension (GenderDim)

Primary Key: Gender Key
Hierarchy: Gender Known → Gender

- **Gender Known**: (0/1) - Indicates whether gender information is available
- **Gender**: (Male, Female, Missing)

8. Vehicle Dimension (VehicleDim)

Primary Key: Vehicle Key
Hierarchy: Bus Involvement, Heavy Rigid Truck Involvement, Articulated Truck Involvement

- **Bus Involvement**: (0/1/-1) - Indicates bus involvement in crash (Yes/No/Unknown)
- **Heavy Rigid Truck Involvement**: (0/1/-1) - Indicates heavy rigid truck involvement (Yes/No/Unknown)
- **Articulated Truck Involvement**: (0/1/-1) - Indicates articulated truck involvement (Yes/No/Unknown)

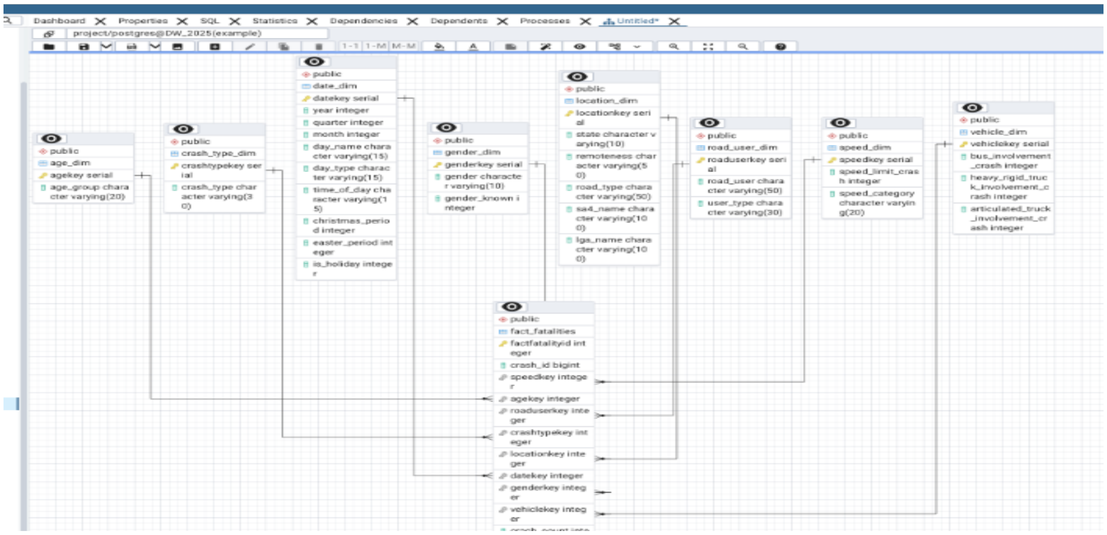
Fact Table (Fact Fatalities)

Primary Key: Fact Fatality ID
Foreign Keys: All dimension keys (SpeedKey, AgeKey, RoadUserKey, CrashTypeKey, LocationKey, DateKey, GenderKey, VehicleKey)
Measures:

- **Crash ID**: Original identifier for the crash event
- **Crash Count**: Always 1, allows for counting distinct crashes
- **Number Fatalities**: Total number of fatalities in the crash

These dimensions were selected to provide comprehensive analytical capabilities across all facets of road fatalities, while maintaining a manageable and efficient schema structure. Essentially it follows following Kimball's principle that dimensions should represent 'all possible descriptions' [2].

4. Data Warehouse Schema And Fact Table Design



Our data warehouse implements a **star schema** design, which consists of a central fact table (Fact Fatalities) surrounded by eight-dimension tables. This schema was chosen for several key reasons:

1. **Analysis Speed**: Road safety data analysis requires quick response times for decision-makers, which star schema delivers through simpler joins. As Kimball notes, ‘Star schemas are specifically designed to address these requirements with a simple, symmetrical, extensible, and predictable structure ‘[2].
2. **Business User Accessibility**: Many stakeholders in road safety (including non-technical government officials) need to understand and work with this data. According to Kimball and Ross, star schemas allow data to be ‘easily sliced and diced any which way your business users want to’[2].
3. **Flexibility for Ad-hoc Queries**: Star schema better supports the varied and unpredictable nature of road safety analysis questions. Kimball's approach prioritizes creating flexible data models that can evolve in response to business changes [2].
4. **Direct Mapping to OLAP Cubes**: Our star schema design directly maps to the multidimensional OLAP cube structure used for analysis, making implementation more straightforward. This alignment supports the performance benefits described by Kimball as a core advantage of dimensional modeling [2].

While alternative schemas like snowflake offer better storage efficiency through normalization, this advantage is outweighed by the performance benefits of the star schema for our analytical needs. As Kimball states in his dimensional modelling approach, query performance and user comprehension should be prioritized over storage efficiency [2].

Query Footprints and OLAP Operations

Our data warehouse supports six key business questions through specific query footprints on our StarNet diagram. Each query uses different dimensions and OLAP operations to answer important road safety questions.

Query 1: Weekend vs Weekday Fatal Crash Patterns

Business Question: How do fatal crash patterns differ between weekends and weekdays across states, and what role does time of day play?

Techniques:

- Used the Location dimension (State level) and Date dimension (Day Type, Time of Day)
- Counted fatalities for each combination
- Filtered all records with missing values
- Compared patterns between weekdays/weekends across different states and times

Analysis Approach:

- Grouped data by state, day type, and time of day
- Could drill down from day type to specific states and times
- Could roll up from specific times to get overall day type patterns
- Sliced away missing data to focus on complete records

Query 2: Risk Analysis of Speed, Road Type, and Time

Business Question: **What specific combinations of speed limits, road infrastructure, and timing factors pose the greatest risk for fatal crashes?**

Techniques:

- Combined Speed dimension (speed categories), Location dimension (road types), and Date dimension (time of day)
- Counted crashes and fatalities for each combination
- Calculated the severity (fatalities per crash) for each combination
- Removed all missing values to ensure data quality

Analysis Approach:

- Created a three-way analysis across speed, road, and time dimensions
- Examined patterns starting from any dimension
- Ranked combinations by total fatalities to identify highest-risk scenarios
- Used dice operations to analyse specific combinations of interest

Query 3: High-Risk Local Government Areas

Business Question: **What are the most dangerous local government areas (LGAs) in terms of road fatalities, and what specific crash characteristics distinguish these high-risk zones?**

Techniques:

- Focused on the most detailed level of Location dimension (LGA name)
- Added Crash Type and Speed Category dimensions
- Counted crashes and fatalities
- Calculated percentage of LGA's total fatalities
- Measured average fatalities per crash to assess severity

Analysis Approach:

- Drilled down to the most detailed geographic level
- Compared patterns within each LGA across crash types and speed categories

- Created percentage metrics to understand relative importance of different factors
- Focused on patterns with sufficient data (5+ crashes)

Query 4: Demographic Analysis of Fatal Crashes

Business Question: What demographic patterns exist in road fatalities when examining the intersection of age, gender, and road user type?

Techniques:

- Combined three demographic dimensions: Age, Gender, and Road User
- Counted crashes and fatalities for each demographic combination
- Calculated each combination's percentage of total fatalities
- Removed all missing values to ensure complete demographic profiles

Analysis Approach:

- Created three-way demographic analysis to identify high-risk groups
- Examined how different demographic factors interact
- Ranked combinations by fatality count to identify priorities
- Used percentage calculations to understand relative risk

Query 5: Holiday Period Risk Analysis

Business Question: How do fatal crash patterns on different road types vary between holiday periods and regular days across urban and remote regions?

Techniques:

- Created special period categories (Christmas, Easter, Regular) from holiday flags
- Combined with Location dimension (Road Type, Remoteness)
- Added Time of Day from Date dimension
- Counted crashes and fatalities for each combination
- Removed missing values from all dimensions

Analysis Approach:

- Created categories from holiday flags to enable special period analysis
- Analyzed patterns across different location types and remoteness levels
- Compared holiday patterns to regular day patterns
- Examined day vs. night differences within each period type

Query 6: Heavy Vehicle Involvement Analysis

Business Question: How does the involvement of different heavy vehicle types influence the number and severity of fatal crashes?

Techniques:

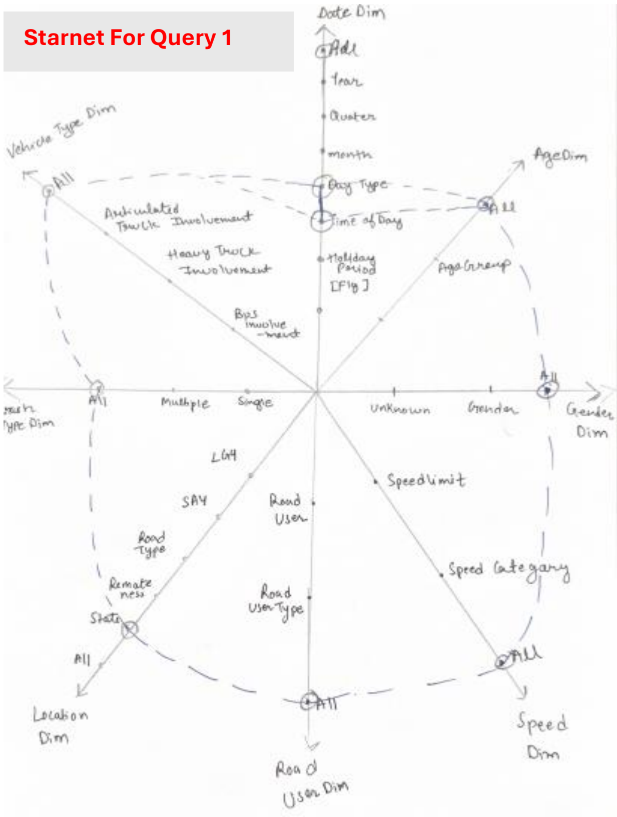
- Used only the Vehicle dimension with three involvement flags (Bus, Heavy Rigid Truck, Articulated Truck)
- Examined all possible combinations of vehicle types
- Counted crashes and fatalities for each combination
- Calculated severity (fatalities per crash) for each combination

Analysis Approach:

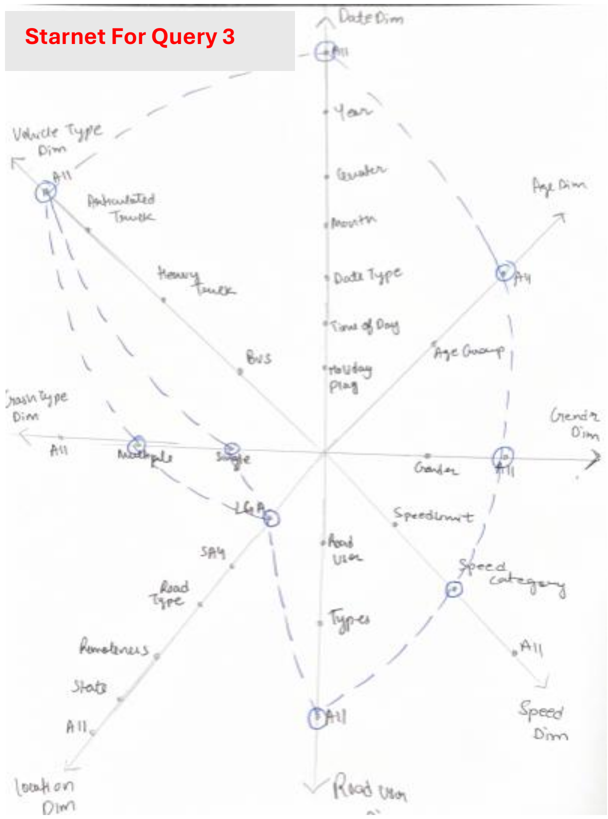
- Created all possible combinations of the three vehicle involvement flags
- Determined how different vehicle combinations affect crash severity
- Identified particularly dangerous vehicle type combinations
- Ranked by fatality count to prioritize safety interventions

These query footprints demonstrate how our star schema effectively supports diverse analytical questions using different dimension combinations and OLAP operations.

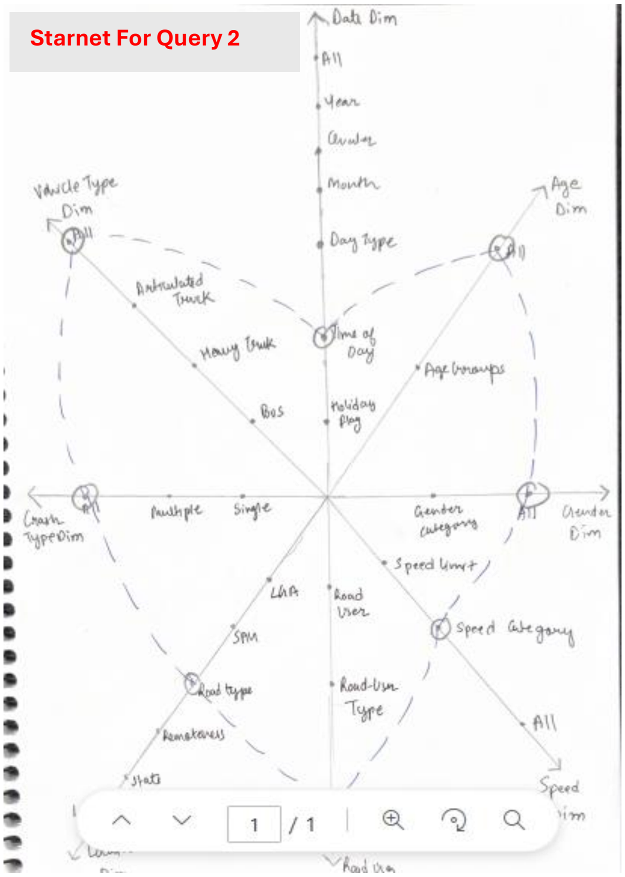
Starnet For Query 1



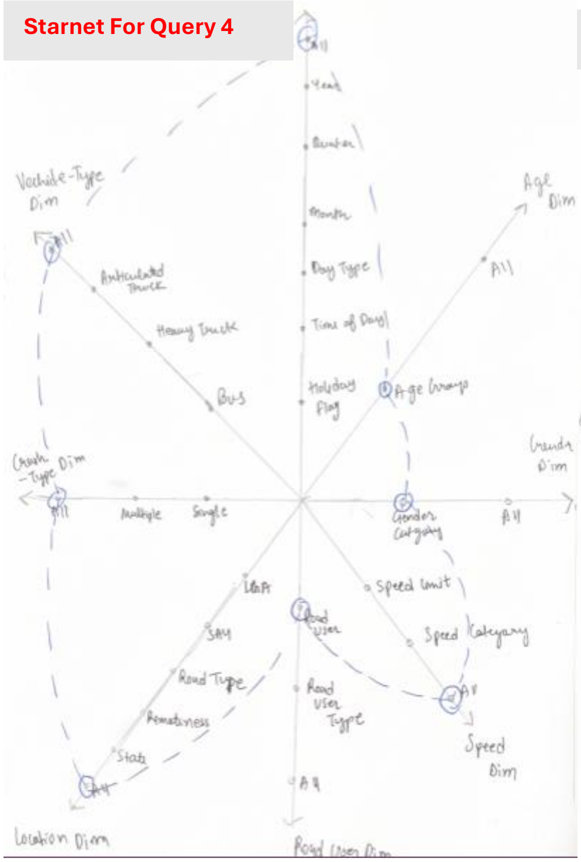
Starnet For Query 3



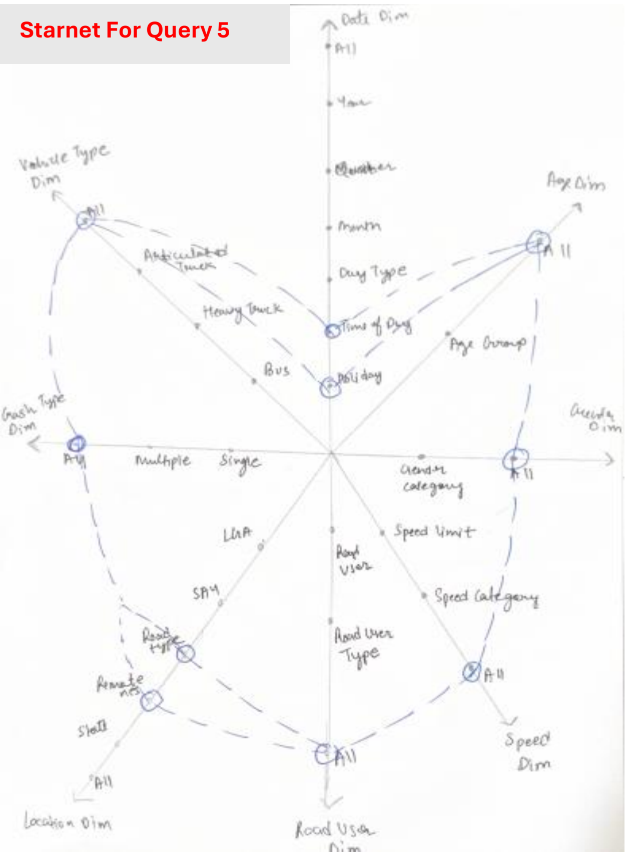
Starnet For Query 2



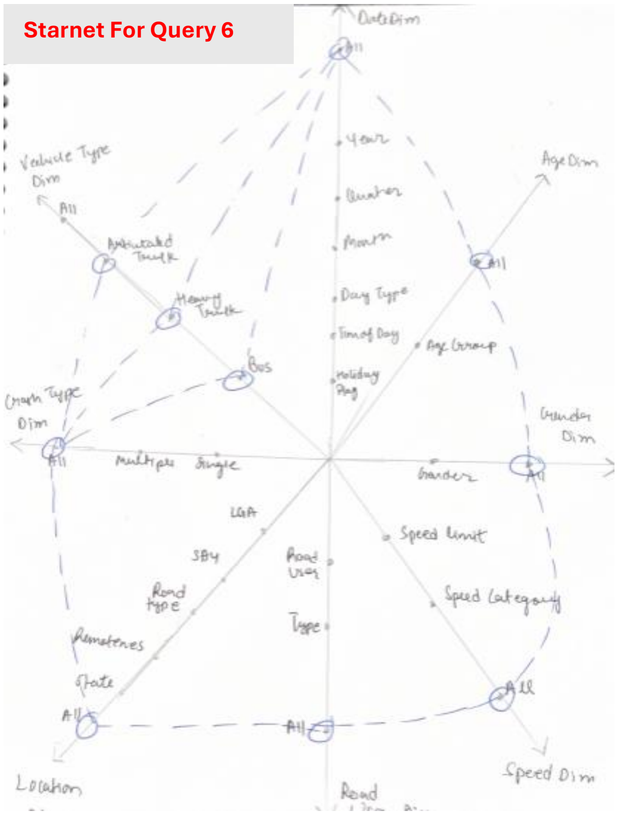
Starnet For Query 4



Starnet For Query 5



Starnet For Query 6

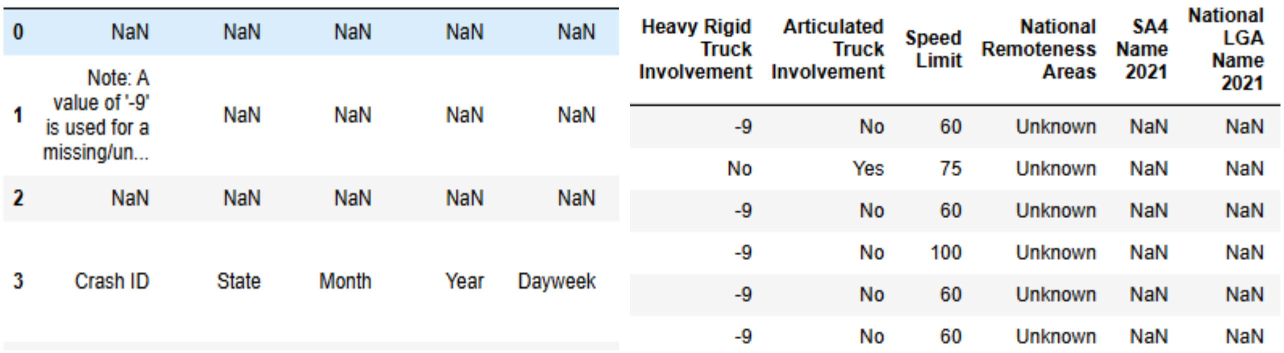


5. Data cleaning, Preprocessing, and ETL Process:

We began our ETL process by cleaning each dataset individually, which involved standardizing headers, stripping whitespace, and handling missing or placeholder values like blank spaces and '-9' respectively. We then analysed the schema of both datasets: the **Fatalities** dataset contained demographic and crash factor information, while **Fatal Crashes** included broader crash metadata. **Crash ID** was identified as the common key; hence, we performed an inner join to merge them into a combined dataset combining individual and crash information. From this combined dataset, we constructed 8-dimension tables and one fact table, forming a star schema ready to be loaded into PDGB admin.

Extraction:

- Datasets were manually exported to CSV files and loaded in with Pandas 'pd. read_csv'
- Once loaded in there was a multiple NaNs and the true header was in the fourth row

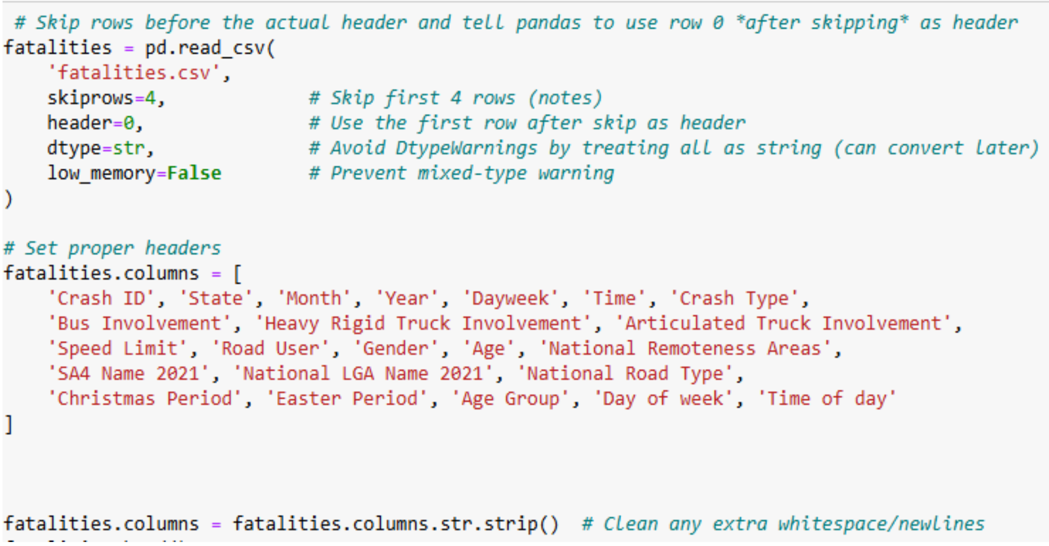


0	NaN	NaN	NaN	NaN	NaN
1	Note: A value of '-9' is used for a missing/un...	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN
3	Crash ID	State	Month	Year	Dayweek

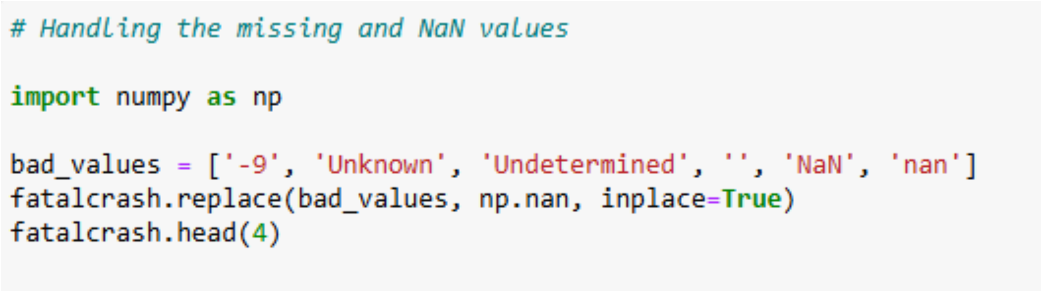
Heavy Rigid Truck Involvement	Articulated Truck Involvement	Speed Limit	National Remoteness Areas	SA4 Name 2021	National LGA Name 2021
-9	No	60	Unknown	NaN	NaN
No	Yes	75	Unknown	NaN	NaN
-9	No	60	Unknown	NaN	NaN
-9	No	100	Unknown	NaN	NaN
-9	No	60	Unknown	NaN	NaN
-9	No	60	Unknown	NaN	NaN

Figure 1: How both Fatal Crashes & Fatalities displayed prior to data cleansing. Displayed NaNs, Misplaced Header, and Placeholder values like '-9' for missing data

Code Details:



```
 # Skip rows before the actual header and tell pandas to use row 0 *after skipping* as header
fatalities = pd.read_csv(
    'fatalities.csv',
    skiprows=4,              # Skip first 4 rows (notes)
    header=0,                # Use the first row after skip as header
    dtype=str,               # Avoid DtypeWarnings by treating all as string (can convert later)
    low_memory=False         # Prevent mixed-type warning
)

# Set proper headers
fatalities.columns = [
    'Crash ID', 'State', 'Month', 'Year', 'Dayweek', 'Time', 'Crash Type',
    'Bus Involvement', 'Heavy Rigid Truck Involvement', 'Articulated Truck Involvement',
    'Speed Limit', 'Road User', 'Gender', 'Age', 'National Remoteness Areas',
    'SA4 Name 2021', 'National LGA Name 2021', 'National Road Type',
    'Christmas Period', 'Easter Period', 'Age Group', 'Day of week', 'Time of day'
]



fatalities.columns = fatalities.columns.str.strip()  # Clean any extra whitespace/newlines
```

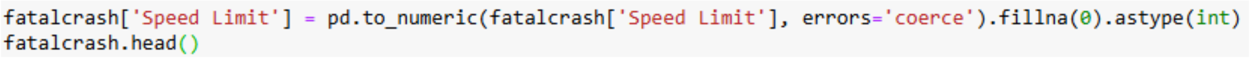


```
# Handling the missing and NaN values

import numpy as np

bad_values = ['-9', 'Unknown', 'Undetermined', '', 'NaN', 'nan']
fatalcrash.replace(bad_values, np.nan, inplace=True)
fatalcrash.head(4)
```

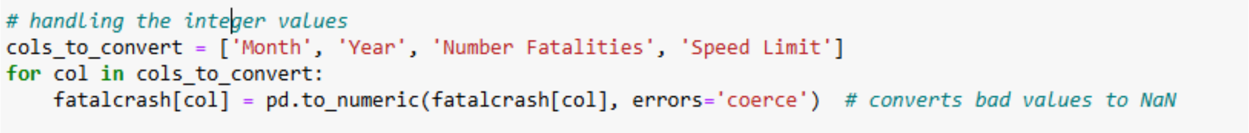
- Skipped metadata rows using skiprows=4 to ignore note rows and reach the true header.
- Set column names manually to match the actual header row (row 5 in the dataset).
- Used *dtype=str* to read all data as strings, avoiding DtypeWarnings and allowing custom conversion.
- Removed extra whitespace from column names using .str.strip() to prevent issues with merging and filtering.
- Replaced invalid or missing values (-9, Unknown, Undetermined, blank, NaN) with np.nan to standardize nulls. Then replaced all NaN values with the label 'Missing'
- We opted for this approach as a high proportion of rows for 'SA4 Name 2021' and 'National LGA Name 2021' were NaNs, and removing these rows would exceed more than 5% of the total dataset. Hence, we decided to use a common label called 'Missing' for all unknown values to preserve data integrity and use this category to differentiate against our valid entries in our analysis and visualization.

Transformation:

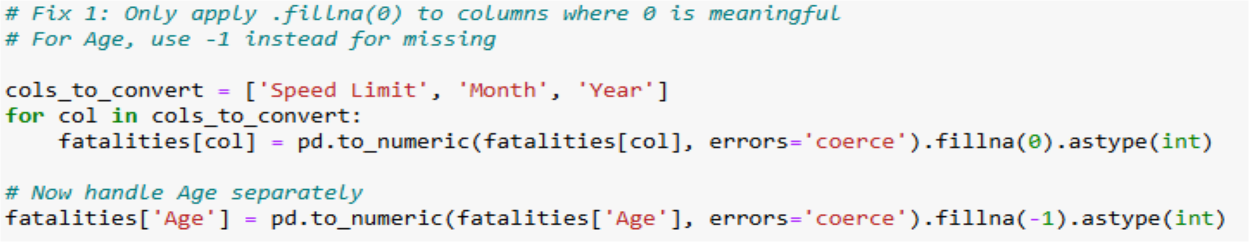
Code Details (Type Conversion & Missing Values):



```
fatalcrash['Speed Limit'] = pd.to_numeric(fatalcrash['Speed Limit'], errors='coerce').fillna(0).astype(int)
fatalcrash.head()
```



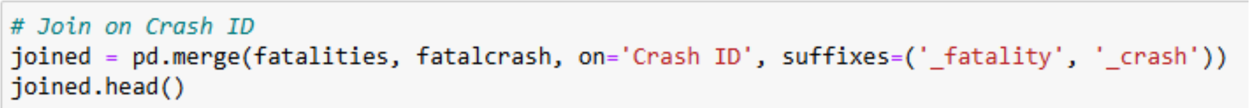
```
# handling the integer values
cols_to_convert = ['Month', 'Year', 'Number Fatalities', 'Speed Limit']
for col in cols_to_convert:
    fatalcrash[col] = pd.to_numeric(fatalcrash[col], errors='coerce')  # converts bad values to NaN
```



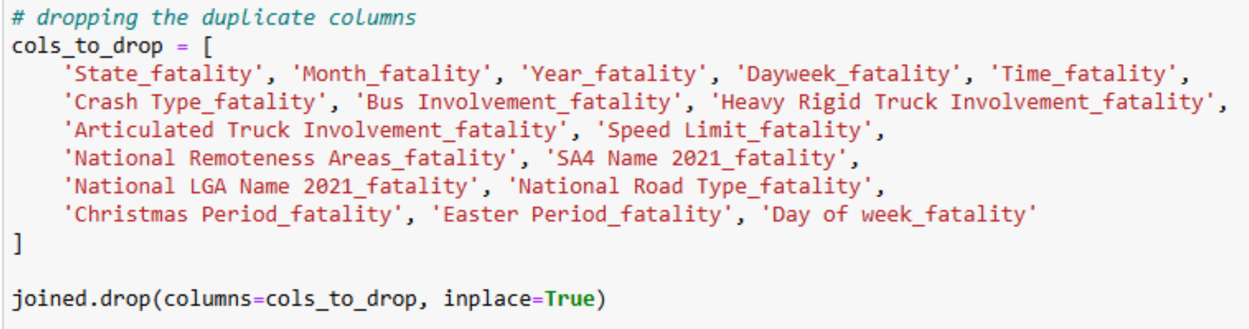
```
# Fix 1: Only apply .fillna(0) to columns where 0 is meaningful
# For Age, use -1 instead for missing

cols_to_convert = ['Speed Limit', 'Month', 'Year']
for col in cols_to_convert:
    fatalities[col] = pd.to_numeric(fatalities[col], errors='coerce').fillna(0).astype(int)

# Now handle Age separately
fatalities['Age'] = pd.to_numeric(fatalities['Age'], errors='coerce').fillna(-1).astype(int)
```

- We selected critical columns like: Age, Speed Limit, Month, Year to be converted into numeric values (integer type) from String
- To do this we used pd.to_numeric (..., errors='coerce') to convert and fill bad/missing values with 0 and a special case for Age where missing values will be ‘-1’ (0 is an age value in the dataset). The results were casted to integer using .astype(int)
- We filled all remaining missing values using fillna('Missing') to simplify downstream processing
- The final datasets were clean, numeric where needed, and consistent for merging with crash data via Crash ID

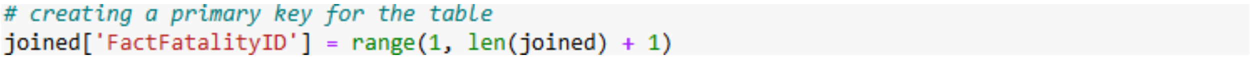
Code Details (Merging Datasets & Generating Primary Key):



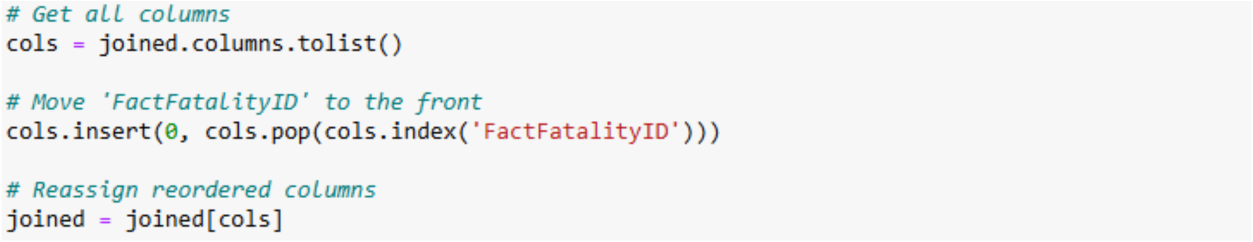
```
# Join on Crash ID
joined = pd.merge(fatalities, fatalcrash, on='Crash ID', suffixes=('_fatality', '_crash'))
joined.head()
```



```
# dropping the duplicate columns
cols_to_drop = [
    'State_fatality', 'Month_fatality', 'Year_fatality', 'Dayweek_fatality', 'Time_fatality',
    'Crash Type_fatality', 'Bus Involvement_fatality', 'Heavy Rigid Truck Involvement_fatality',
    'Articulated Truck Involvement_fatality', 'Speed Limit_fatality',
    'National Remoteness Areas_fatality', 'SA4 Name 2021_fatality',
    'National LGA Name 2021_fatality', 'National Road Type_fatality',
    'Christmas Period_fatality', 'Easter Period_fatality', 'Day of week_fatality'
]

joined.drop(columns=cols_to_drop, inplace=True)
```



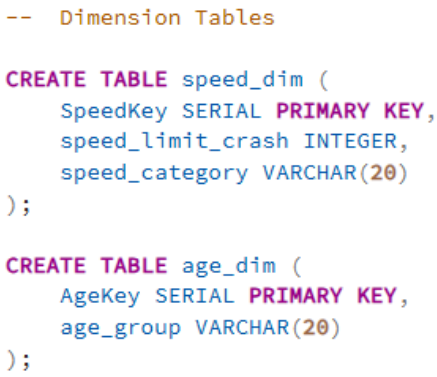
```
# creating a primary key for the table
joined['FactFatalityID'] = range(1, len(joined) + 1)
```



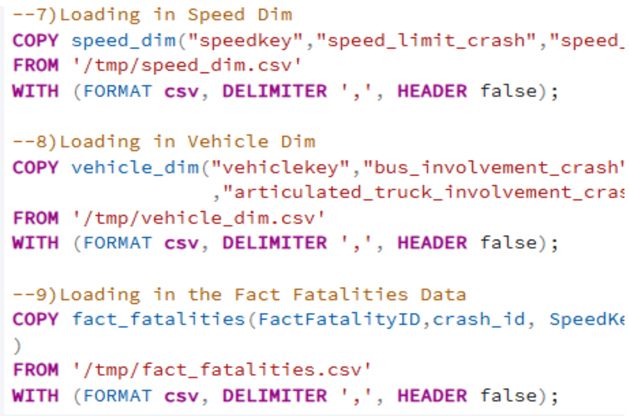
```python
# Get all columns
cols = joined.columns.tolist()

# Move 'FactFatalityID' to the front
cols.insert(0, cols.pop(cols.index('FactFatalityID')))

# Reassign reordered columns
joined = joined[cols]
```

- Performed an inner join on the shared Crash ID column using pd.merge()
- Ensured each person-level fatality record had their own crash information
- Used suffixes (_fatality, _crash) to differentiate columns that appeared in both datasets; hence removing duplicate or redundant columns from the fatalities dataset
- Dropped columns: State_fatality, Speed Limit_fatality, etc.
- Support for this decision was that these attributes were duplicate columns already in the fatal crash dataset and including them would cause redundancy
- Created a new surrogate primary key called **FactFatalityID**
- Assigned a unique ID to every fatality (row) in the combined dataset to differentiate each fatality from each in an individual crash; hence **FactFatalityID** is used as the primary identifier in the fact table.
- Reordered columns to move FactFatalityID to the front.This makes the table easier to read and matched the ordering convention used in dimension and fact table structures.
- Final result: a clean, combined dataset with one row per person killed in a crash, with consistent crash attributes and ready for dimension joins and fact table construction.

Loading:



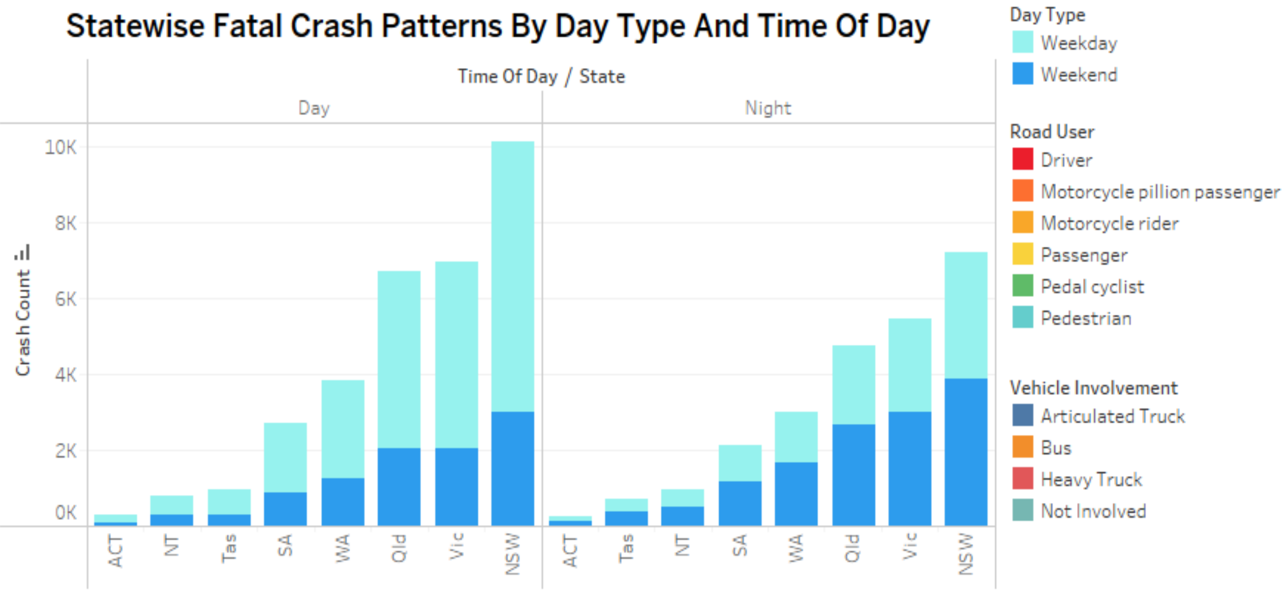
```sql
--  Dimension Tables

CREATE TABLE speed_dim (
    SpeedKey SERIAL PRIMARY KEY,
    speed_limit_crash INTEGER,
    speed_category VARCHAR(20)
);

CREATE TABLE age_dim (
    AgeKey SERIAL PRIMARY KEY,
    age_group VARCHAR(20)
);
```

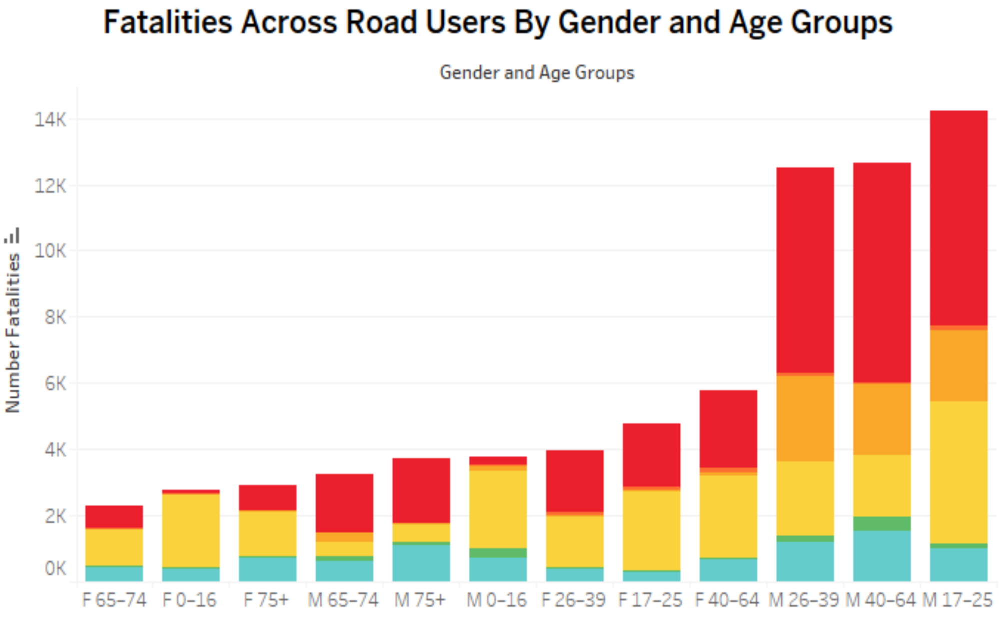


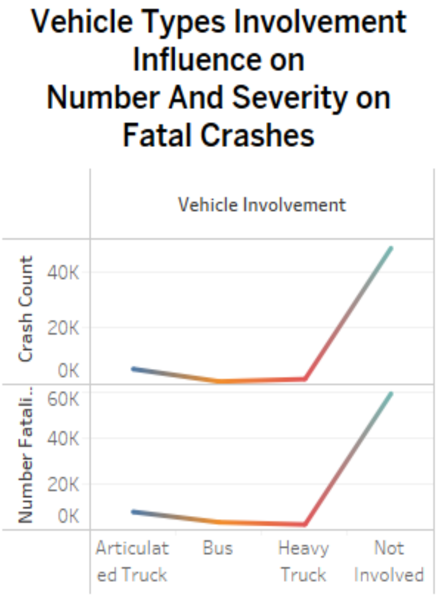
```sql
--7)Loading in Speed Dim
COPY speed_dim("speedkey","speed_limit_crash","speed_
FROM '/tmp/speed_dim.csv'
WITH (FORMAT csv, DELIMITER ',', HEADER false);

--8)Loading in Vehicle Dim
COPY vehicle_dim("vehiclekey","bus_involvement_crash'
                 ,"articulated_truck_involvement_cras
FROM '/tmp/vehicle_dim.csv'
WITH (FORMAT csv, DELIMITER ',', HEADER false);

--9)Loading in the Fact Fatalities Data
COPY fact_fatalities(FactFatalityID,crash_id, SpeedKe
)
FROM '/tmp/fact_fatalities.csv'
WITH (FORMAT csv, DELIMITER ',', HEADER false);
```

We first loaded all dimension and fact tables using their respective schemas. Then, we performed bulk insert operations to efficiently load the data into the destination system of the data warehouse, which is a PostgreSQL database. To ensure data integrity, we defined foreign key relationships between the fact table and the corresponding dimension tables. This process ensured that the data was consistently and reliably loaded, making it ready for efficient querying and visualization in Tableau during the later steps of our analysis.

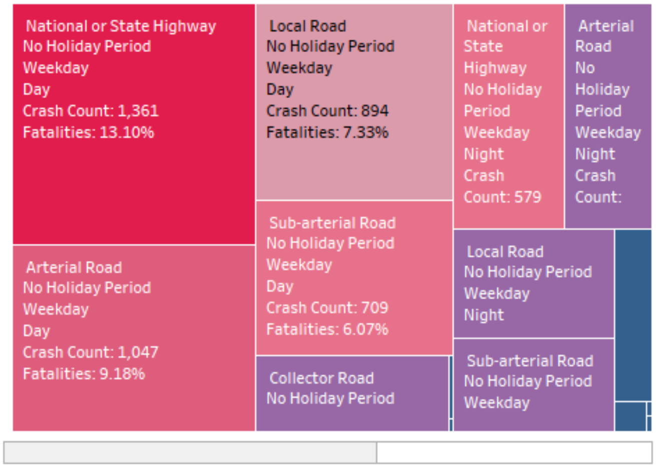
6. Query Results and Visualisation Using Tableau


Statewise Fatal Crash Patterns By Day Type And Time Of Day

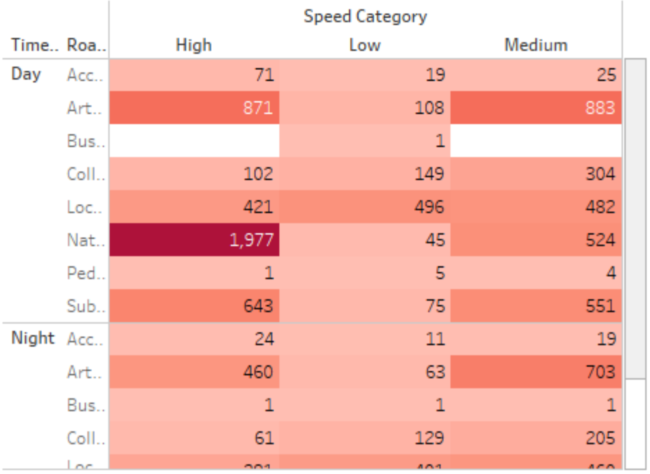

Fatalities Across Road Users By Gender and Age Groups


Vehicle Types Involvement Influence on Number And Severity on Fatal Crashes

Fatality Rates And Crash Patterns Across Road Types, Times of Day, and Holiday Periods

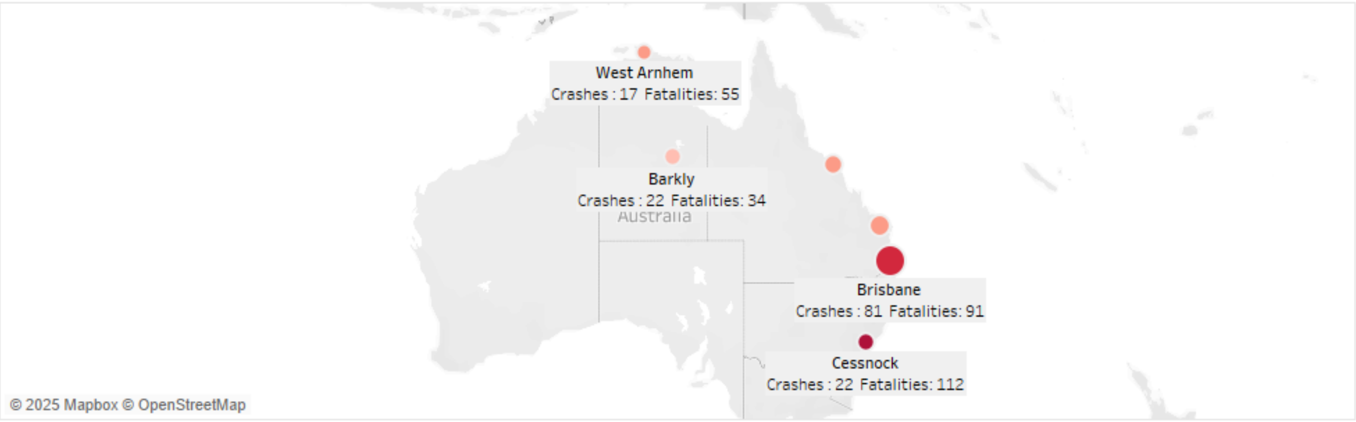


Number Of Fatalities by Speed Limit, Road Type, and Time of Day



		Speed Category		
Time..	Roa..	High	Low	Medium
Day	Acc..	71	19	25
	Art..	871	108	883
	Bus..		1	
	Coll..	102	149	304
	Loc..	421	496	482
	Nat..	1,977	45	524
	Ped..	1	5	4
	Sub..	643	75	551
Night	Acc..	24	11	19
	Art..	460	63	703
	Bus..	1	1	1
	Coll..	61	129	205
	Loc..	[illegible]	[illegible]	[illegible]

Dangerous LGA Zones By Combinations Of Crash Type,Speed And Remoteness


West Arnhem
Crashes : 17 Fatalities: 55
Barkly
Crashes : 22 Fatalities: 34
Australia
Brisbane
Crashes : 81 Fatalities: 91
Cessnock
Crashes : 22 Fatalities: 112
© 2025 Mapbox © OpenStreetMap

Query 1: How do fatal crash patterns differ between weekends and weekdays across states, and what role does time of day play?

Our key insights found that NSW had the highest number of fatal crashes resulting in 7122 fatalities during the weekday. Referring to the bar plot titled '**Statewise Fatal Crash Patterns By Day And Time Of Day**' We found weekday daytime driving is generally most dangerous across all states while weekend nights showed a disproportionately higher fatal crash rate. The plot displays the total number of crashes by each state where 'Time of Day' only includes 'Day' and 'Night'; excluding 'Missing' as we believed it would not significantly impact our visual analysis. The plot offers different reveals of the different risk profiles for each State. For example, ACT would be considered a low crash risk state as the proportion of fatal crashes is the lowest for both 'Time of Day' categories.

Query 2: What specific combinations of speed limits, road infrastructure, and timing factors pose the greatest risk for fatal crashes?

In reference to our heatmap titled **'Number Of Fatalities by Speed Limit, Road Type, and Time of Day '** where the intensity of color in each column block represented the severity of fatalities in fatal crashes for each crash risk. From the visual interpretation of our query, we found that: 1. High-speed, on a National or State Highway during the day was the greatest risk for fatal crashes based on the severity of the number of fatalities. This is supported by reference to the result of our query where National or Highway for this combination had a crash count of 1,252 and 1,977 fatalities.

Query 3: What are the most dangerous local government areas (LGAs) in terms of road fatalities, and what specific crash characteristics distinguish these high-risk zones?

Based on the geographical plot titled '**Dangerous LGA Zones By Combinations Of Crash Type, Speed And Remoteness**' The most dangerous LGA zone is shown by the color intensity which represents the severity of fatalities. Cessnock (NSW) stands out as particularly dangerous with 112 fatalities from 22 Crashes. Specific crash characteristics for Cessnock included: single-vehicle crashes in medium-speed zones. Brisbane (QLD) was the second most dangerous LGA with 81 crashes and 91 fatalities involving multiple-vehicles in medium-speed zones. West Arnhem (NT) shows a concerning pattern with 100% of its fatalities (55) occurring in high-speed single-vehicle crashes. Our map included remote areas like Whitsunday and Bundaberg which had a particularly high number of fatalities in high-speed zones.

Query 4: What demographic patterns exist in road fatalities when examining the intersection of age, gender, and road user type?

Our stacked bar plot titled **'Fatalities Across Road Users by Gender and Age Groups'** highlights the top demographic categories involved in road fatalities: Drivers who are young males (17–25), adult males (26–39), and middle-aged males (40–64). Based on our query results, these groups combined account for 26.64% of all road fatalities, and have the highest fatalities across all Road User Type, indicating a clear gender disparity in the fatalities of fatal crashes.

Query 5: How do fatal crash patterns on different road types vary between holiday periods and regular days across urban and remote regions?

The titled treemap, **'Fatality Rates And Crash Patterns Across Road Types, Times of Day, and Holiday Periods**' show that National or State Highways in regional areas during regular days have the highest fatal crash count (1,361). The second highest fatal crashes occurred during the Christmas Period on National/State Highways in Inner Regional Australia. Some key differences between the Christmas and Easter holiday periods indicated by our query results are: 1.

Christmas fatalities are spread across different road types unlike Easter fatalities 2. Outer Regional highways show higher fatality counts during Christmas (43 fatalities) than Easter (20 fatalities) 3. Daytime driving during holidays poses a higher crash risk than nighttime, especially on major highways.

Query 5: How does the involvement of different heavy vehicle types influence the number and severity of fatal crashes?

Based on the line graph titled **'Vehicle Types Involvement Influence on Number And Severity on Fatal Crashes'** No heavy vehicle involvement accounts for 50.5% of our total dataset (36,728 fatalities). This can be shown by the highest peak in the line graph colored blue for 'Not Involved'. We have intentionally included this value as we wanted to preserve data integrity and offer accurate insight. To draw comparisons across vehicle types, Articulated Trucks were the most severe in fatalities and crash counts. Buses also demonstrated severity in fatalities, but this is likely due to the volume of passengers involved. Heavy truck contributed to the least number of fatalities as evidenced by its position as the closest point to the axis in the graph.

7. Association Rules Mining

Apriori Algorithm:

We opted to use Apriori Algorithm as an unsupervised machine learning algorithm for our Association Rules Mining; data mining technique that identifies frequent patterns, connections and dependencies among different groups of items (item sets) [3] The algorithm operates on the principle that all subsets of a frequent itemset must also be frequent. This property, known as the anti-monotonicity of support, allows for efficient pruning of the search space, making it efficient to analyse high-dimensional datasets. We use this mining technique to identify the underlying reoccurring pattern between various factors and road user types involved in fatal crashes. Hence, able to provide insightful recommendations to the Australian Government to improve road safety of our roads.

Our basis for choosing Apriori Algorithm:

1. Categorical Data: Road User type, Age Group, State is categorical data in our dataset that align well with itemset-based approach of Apriori

2. Interpretability: The algorithm produces clear association rules that connect antecedents (conditions) with consequents (outcomes)

3. Quantifiable Strength of Associations: Support-confidence-lift provides clear guide to assess how significant and reliable patterns are in data

Minimum support (1%) and confidence (10%) were chosen to find meaningful patterns while filtering out weak associations

Data Preparation:

The dataset contained 56,874 road fatality records with 26 attributes. Several preprocessing steps were undertaken:

1. Missing Value Analysis: Multiple columns had a high proportion of ‘Missing’ values; therefore, columns with over 80% of missing data were excluded. These columns were: SA4 Name 2021_crash, National LGA Name 2021_crash,National Road Type_crash, and National Remoteness Areas_crash

2. Feature Selection: Five key columns were selected for association rule mining:

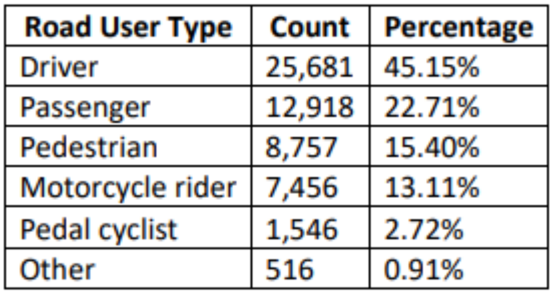
- Road User (target variable as consequent)
- Age Group
- State_crash
- Speed Category (derived)
- Time of Day

3. Feature Engineering: A new "Speed Category" variable was created to transform the numeric "Speed Limit_crash" into a meaningful categorical variable with three levels such as: Low (≤50 km/h),Medium (51-80 km/h), High (>80 km/h)

4. Transaction Encoding: The data was transformed into a transaction format suitable for the Apriori algorithm, with each attribute-value pair encoded as an item (e.g., "Age Group=75_or_older").

Results:

Before examining the association rules, it’s imperative to understand the baseline distribution of fatalities across road user types:

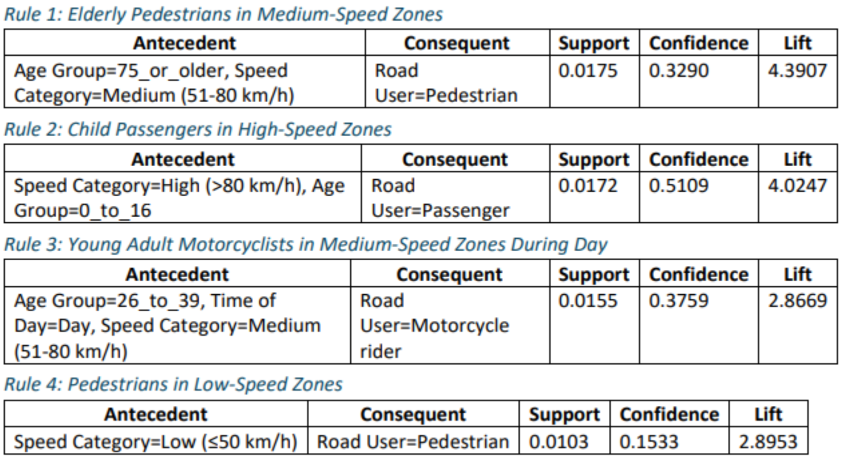


Road User Type	Count	Percentage
Driver	25,681	45.15%
Passenger	12,918	22.71%
Pedestrian	8,757	15.40%
Motorcycle rider	7,456	13.11%
Pedal cyclist	1,546	2.72%
Other	516	0.91%

Despite drivers constituting the largest category of fatalities, the association rule analysis identified stronger patterns (measured by lift) for other road user types under specific conditions. This highlights the value of the lift metric in identifying disproportionate relationships beyond raw frequency counts.

Top Association Rules with "Road User" as Consequent

The following tables present the four most significant non-duplicate association rules identified by lift, with "Road User" as the consequent:



Rule 1: Elderly Pedestrians in Medium-Speed Zones

Antecedent	Consequent	Support	Confidence	Lift
Age Group=75_or_older, Speed Category=Medium (51-80 km/h)	Road User=Pedestrian	0.0175	0.3290	4.3907

Rule 2: Child Passengers in High-Speed Zones

Antecedent	Consequent	Support	Confidence	Lift
Speed Category=High (>80 km/h), Age Group=0_to_16	Road User=Passenger	0.0172	0.5109	4.0247

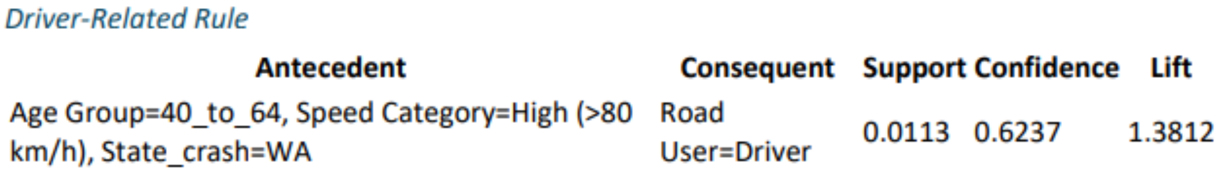
Rule 3: Young Adult Motorcyclists in Medium-Speed Zones During Day

Antecedent	Consequent	Support	Confidence	Lift
Age Group=26_to_39, Time of Day=Day, Speed Category=Medium (51-80 km/h)	Road User=Motorcycle rider	0.0155	0.3759	2.8669

Rule 4: Pedestrians in Low-Speed Zones

Antecedent	Consequent	Support	Confidence	Lift
Speed Category=Low (≤50 km/h)	Road User=Pedestrian	0.0103	0.1533	2.8953

For comparison, the highest-lift rule with "Driver" as the consequent:



Driver-Related Rule

Antecedent	Consequent	Support	Confidence	Lift
Age Group=40_to_64, Speed Category=High (>80 km/h), State_crash=WA	Road User=Driver	0.0113	0.6237	1.3812

Plain English Interpretation of Top Rules

Rule 1: Elderly Pedestrians in Medium-Speed Zones: When a road fatality involves an elderly person (75 or older) in a medium-speed zone (51-80 km/h), there is a significantly elevated likelihood that the victim is a pedestrian. This pattern occurs in about 1.75% of all fatalities, and under these specific conditions, the victim is a pedestrian 32.9% of the time—4.4 times higher than would be expected by random chance. This indicates a pronounced vulnerability of elderly pedestrians in these speed environments, likely due to reduced mobility resulting in slower crossing times.

Rule 2: Child Passengers in High-Speed Zones: When a road fatality involves a child (0-16 years) on a high-speed road (>80 km/h), there is a strong association with the victim being a passenger. This pattern appears in 1.72% of all fatalities, and in these specific circumstances, the victim is a passenger 51.1% of the time—4 times higher than would be expected by random chance. This highlights the vulnerability of children as passengers in high-speed crashes, due to both the severity of the crashes and their physical sensitivity, even when properly restrained.

Rule 3: Young Adult Motorcyclists in Medium-Speed Zones During Day: When a road fatality involves a young adult (26-39 years) during daylight hours on a medium-speed road (51-80 km/h), there is a strong association with the victim being a motorcycle rider. This pattern occurs in about 1.55% of all fatalities, and under these conditions, the victim is a motorcycle rider 37.6% of the time—2.9 times higher than expected by chance. This suggests a particular risk for young adults who are motorcyclists who fit these criteria.

Rule 4: Pedestrians in Low-Speed Zones: Fatalities in low-speed zones (≤50 km/h) show a strong association with pedestrian victims. This pattern appears in 1.03% of all fatalities, and in these environments, the victim is a pedestrian 15.3% of the time—2.9 times higher than expected by chance. This finding, which may seem counterintuitive, suggests that despite lower vehicle speeds, urban and residential areas with low-speed limits pose significant risks to pedestrians, likely due to higher pedestrian density as well as higher rates of interaction between pedestrians and vehicles.

Key Insights from Association Rules

Age-Specific Vulnerability Patterns - different age groups show distinct vulnerability patterns as road users:

- Children (0-16): Highly vulnerable as passengers, particularly in high-speed environments
- Young Adults (17-39): Show stronger associations with motorcycle fatalities
- Middle-Aged Adults (40-64): More commonly involved as drivers, particularly in highspeed zones in Western Australia
- Elderly (75+): Extremely vulnerable as pedestrians across multiple speed environments

Speed Environment Impact - Different speed zones exhibit distinct fatality patterns:

- High-Speed Zones (>80 km/h): Strong association with passenger fatalities, especially for children, and driver fatalities for middle-aged adults
- Medium-Speed Zones (51-80 km/h): Associated with pedestrian fatalities for elderly and motorcycle fatalities for young adults
- Low-Speed Zones (≤50 km/h): Despite lower speeds, show strong association with pedestrian fatalities

Influence of Time of Day

- Daytime: Stronger association with young adult motorcycle fatalities in medium-speed zones and child passenger fatalities in high-speed environments
- Nighttime: Higher association with pedestrian fatalities, especially for middle-aged adults in medium-speed environments

State-Specific Patterns

Geographic differences exist across Australian states:

- New South Wales: Strong association between children (0-16) and passenger fatalities (Lift: 2.49)
- Victoria: Medium-speed zones show association with pedestrian fatalities (Lift: 1.89)
- Queensland: Stronger association with motorcycle rider fatalities, especially in daytime medium speed zones (Lift: 1.83)
- Western Australia (WA): Middle-aged adults (40-64) in high-speed zones show association with driver fatalities (Lift: 1.38)

Recommendations for Improving Road Safety

1) Age-Targeted Safety Interventions

Implement safety programs for vulnerable age groups

Elderly Pedestrians - Extend pedestrian crossing times at intersections in areas with high elderly populations

Child Passengers - Strengthen the enforcement of child restraint laws, especially on high-speed roads

Young Adult Motorcyclists - Implement targeted enforcement of protective gear requirements

Middle-Aged Drivers - Enforcement of speeding and distraction laws on high-speed roads

The association rules clearly demonstrate that different age groups face distinct risks. By tailoring interventions, resources are used where it will have the greatest impact.

2) Speed Zone-Specific Engineering Solutions

Redesign road environments based on the specific risk patterns identified in different speed zones.

Low-Speed Zones - Redesign urban intersections to improve pedestrian visibility

Medium-Speed Zones - Install median barriers to prevent pedestrians from crossing at undesignated locations

High-Speed Zones - Install cable barriers or concrete barriers to reduce the severity of run-off-road crashes

The association rules show that different speed environments are associated with different types of fatalities. Engineering solutions tailored to the specific risks in each environment would address the underlying mechanisms of fatal crashes.

3) Safety Programs For Each State

Develop targeted safety initiatives that address the unique patterns identified in each state.

New South Wales - Implement pedestrian-priority zones in areas with high pedestrian activity

Victoria - Enhance street lighting at pedestrian crossings on medium-speed roads

Queensland - Implement motorcycle-specific infrastructure improvements on popular routes

Western Australia - install smart systems on roads that automatically monitor and adjust driving speed based on the road conditions; Which can be to used dangerous rural roads

By addressing each unique risk identified in each state, resources can be allocated more efficiently and effectively

Conclusion:

While drivers are the largest category of fatalities, the association rule analysis demonstrates the value of looking beyond raw numbers to identify specific high-risk scenarios where interventions might be most effective. A comprehensive road safety strategy should incorporate both broad approaches to address the large number of driver fatalities and targeted interventions for the distinct risks discussed in the analysis. By implementing the recommended age-targeted, speed zone-specific, and state safety programs, significant progress could be made in reducing road fatalities across all road user groups in Australia.

While our created data warehouse is comprehensive in its analysis, a limitation would be the significant portion of missing data (specifically geographic and vehicle information) and simplistic categorizations which reduces the precision of our analysis.

References (IEEE)

[1] “Road safety,” Department of Infrastructure, Transport, Regional Development, Communications and the Arts, Nov. 27, 2024. https://www.infrastructure.gov.au/infrastructure-transport-vehicles/roads/road-safety

[2] “Kimball’s Dimensional Data Modeling | The Analytics Setup Guidebook,” *www.holistics.io*. https://www.holistics.io/books/setup-analytics/kimball-s-dimensional-data-modeling/

[3] IBM, “Apriori Algorithm,” Ibm.com, Jun. 09, 2024. https://www.ibm.com/think/topics/apriori-algorithm