2024 - BUS5DWR – DATA WRANGLING WITH R

Assignment 3: Data Wrangling with R

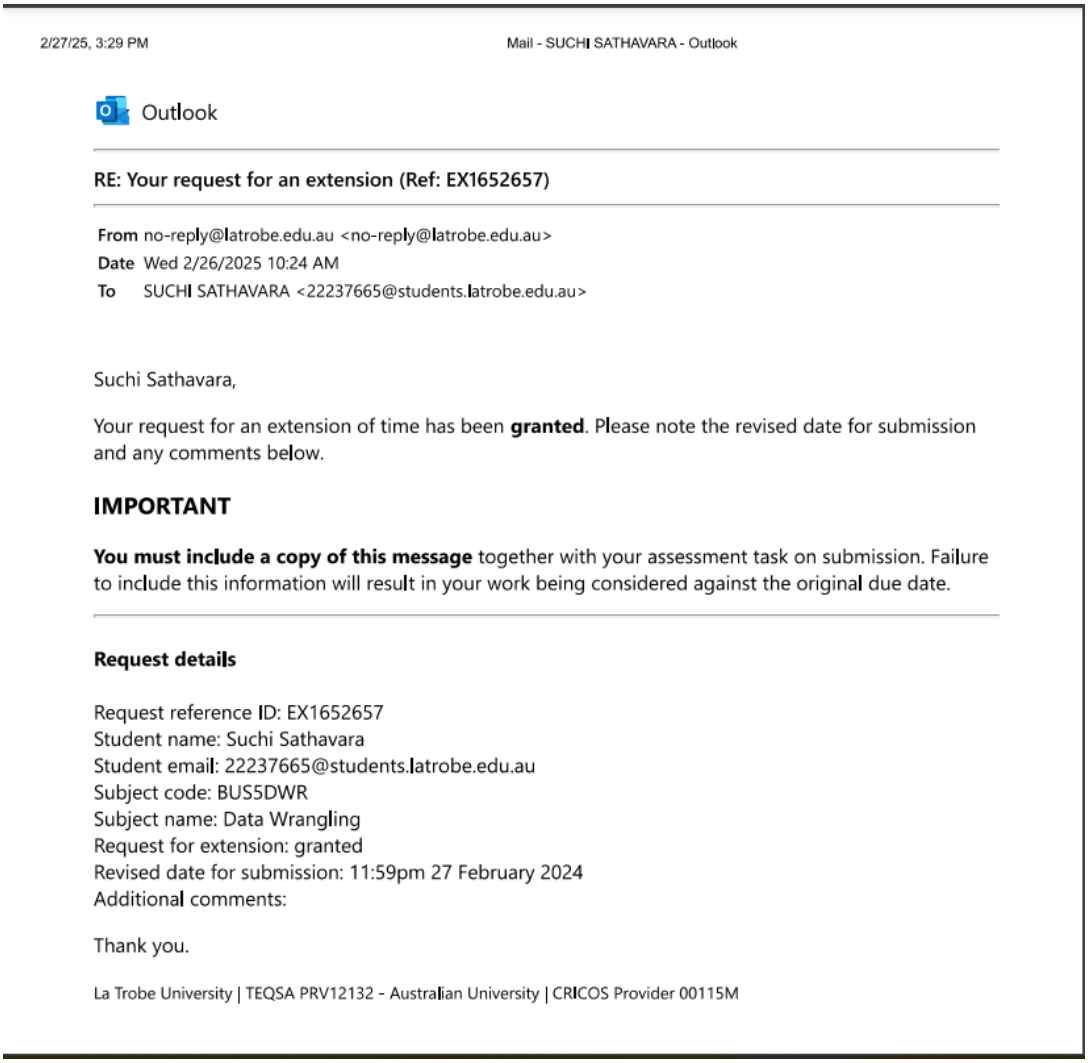
Case Study: Analyzing Road Accidents in Victoria, Australia

Student ID: 2237665

Lecturer: Nadeeka Basnayake La Trobe University

26th February 2025

Deadline Extension



2/27/25, 3:29 PM Mail - SUCHI SATHAVARA - Outlook

Outlook

RE: Your request for an extension (Ref: EX1652657)

From no-reply@latrobe.edu.au <no-reply@latrobe.edu.au>
Date Wed 2/26/2025 10:24 AM
To SUCHI SATHAVARA <22237665@students.latrobe.edu.au>

Suchi Sathavara,

Your request for an extension of time has been **granted**. Please note the revised date for submission and any comments below.

IMPORTANT

You must include a copy of this message together with your assessment task on submission. Failure to include this information will result in your work being considered against the original due date.

Request details

Request reference ID: EX1652657
Student name: Suchi Sathavara
Student email: 22237665@students.latrobe.edu.au
Subject code: BUS5DWR
Subject name: Data Wrangling
Request for extension: granted
Revised date for submission: 11:59pm 27 February 2024
Additional comments:

Thank you.

La Trobe University | TEQSA PRV12132 - Australian University | CRICOS Provider 00115M

Introduction

Road safety is a critical concern for communities worldwide, and understanding the factors contributing to traffic accidents is essential for developing effective prevention strategies. This project focuses on analyzing road accident data in Victoria, Australia, to identify key trends, patterns, and safety concerns that impact accident severity and frequency. Utilizing datasets containing information on accident details, involved persons, and geographical locations, this analysis aims to provide data-driven insights that can inform evidence-based policy decisions and targeted road safety interventions.

Task 1 - Data merging

The given data files were load into individual dataframes. They were explored by head() function. All three datafiles were merged into one named dataset. The dataset was later explored by glimpse() and summary(). To summarize dataset summarytools was imported. To view the summary View(dfSummary(dataset)) was used.

Task 2 – Data cleaning

1.) Handling missing values:

The initial evaluation of dataset started with checking missing values. For that, colSums(is.na(dataset)) was used. To check which row contains missing value: dataset[!complete.cases(dataset)] was used. For column HELMET_BELT_WORN and SEATING_POSITION missing values is handled by “UNKNOWN”. For numerical column filled with mean. And drop the remaining missing values rows.

2.) Convert Data Type:

In the dataset, convert ACCIDENT_DATE from chr data type to Date type. Convert SPEED_ZONE and ACCIDENT_NO as factor from character.

3.) Standardize Gender:

To standardize SEX column use toupper().

4.) Remove duplicate:

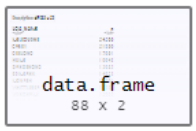
Remove duplicate rows with distinct() function.

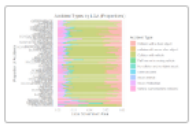
These steps ensured the dataset was clean, consistent, and prepared for further analysis.

Overall, the quality of dataset has both negative and positive reviews. Positive because of the volume of data but it needs quite a bit of attention for data cleaning. In consistency in some column as well as not a proper data format and data types is an bit of an issue.

Task 3 – Suburb and Accident Relationship

Using dataset we analyze accident distribution over suburb. To get the details about accidents in the suburbs use LGA_NAME column. To count accident per suburb use library dplyr function count().



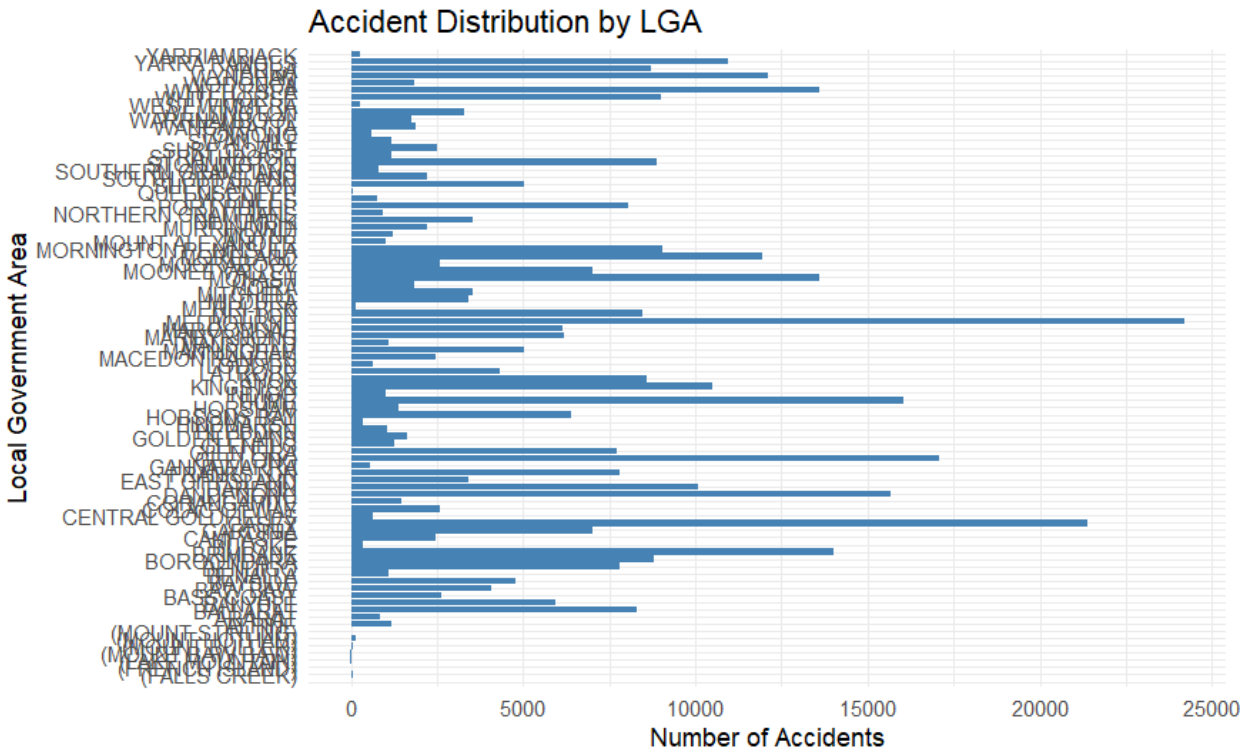
Description: df [88 × 2]

LGA_NAME <chr>	n <int>
MELBOURNE	24209
CASEY	21390
GEELONG	17091
HUME	16043
DANDENONG	15685
BRIMBANK	14028
MONASH	13601
WHITTLESEA	13595
WYNDHAM	12095
MORELAND	11951

1-10 of 88 rows

Previous 1 2 3 4 5 6 ... 9 Next

Here is the suburb with high number of accidents. To visualize this use ggplot library.

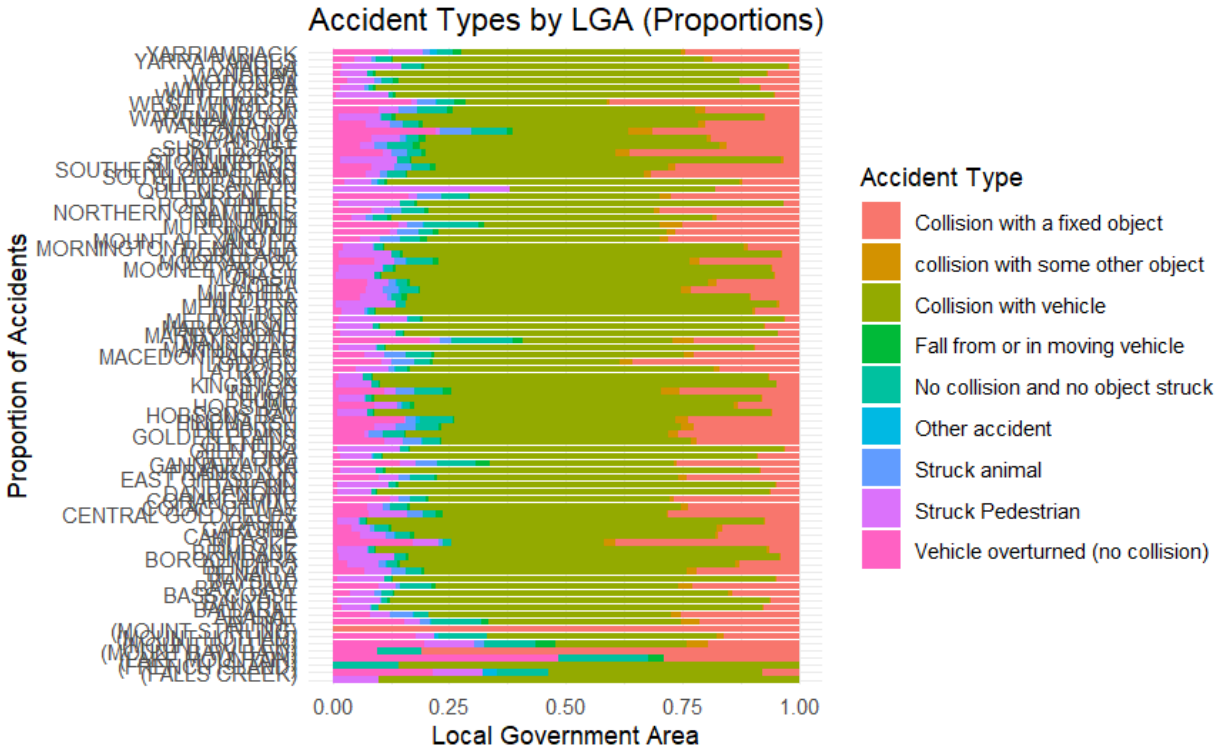


How this accident occurred in specific suburb? To find that out I used cross tabulation between LGA_NAME and ACCIDENT_TYPE_DESC.

LGA_NAME <chr>	Collision with vehicle <int>	Struck Pedestrian <int>
	9	1
(FALLS CREEK)	30	7
(FRENCH ISLAND)	6	0
(LAKE MOUNTAIN)	0	0
(MOUNT BAW BAW)	0	0
(MOUNT BULLER)	13	5
(MOUNT HOTHAM)	61	5
(MOUNT STIRLING)	0	0
ALPINE	461	39
ARARAT	391	38

1-10 of 88 rows | 1-3 of 10 columns Previous 1 2 3 4 5 6 ... 9 Next

For better understanding, I visualize this into,



Data findings:

1.) Melbourne and Geelong city areas have higher number of accidents due to more population.
2.) Rural LGA might have more single accidents than in collision with objects or struck animal.
3.) Industrial area like Geelong has heavy collisions accidents.

Trend:

In a way, large number of accidents might have happened due to one or another type of collisions.

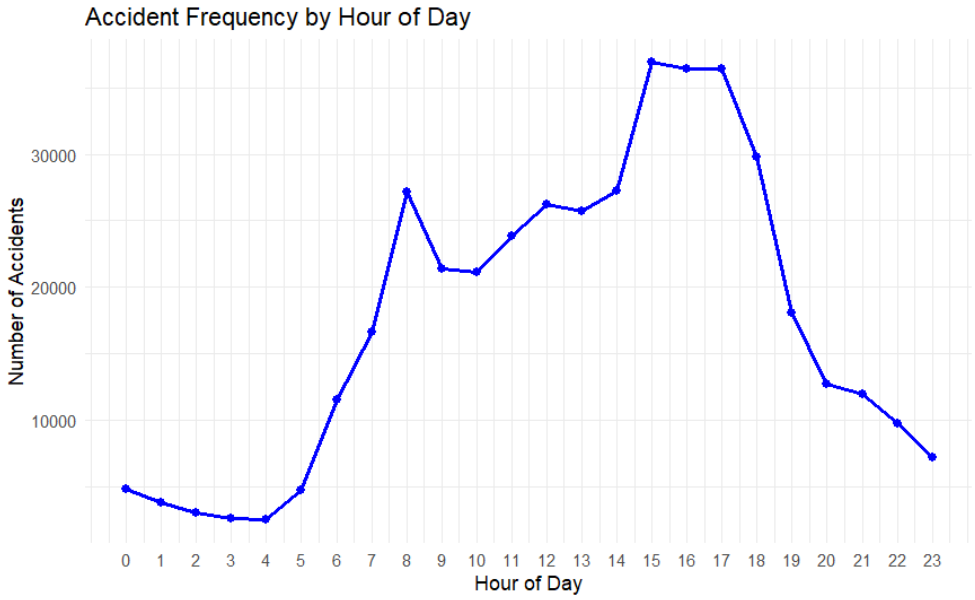
Melbourne has different geometry roads so the accident type can be more varied.

In general, this finding shows the need of road safety measures depending on each LGA and it's characteristic.

Task 4 – Time and Accident Relationship

This is the analysis of accidents over its occurrence. For extracting time hour and day of the week I import lubridate library.

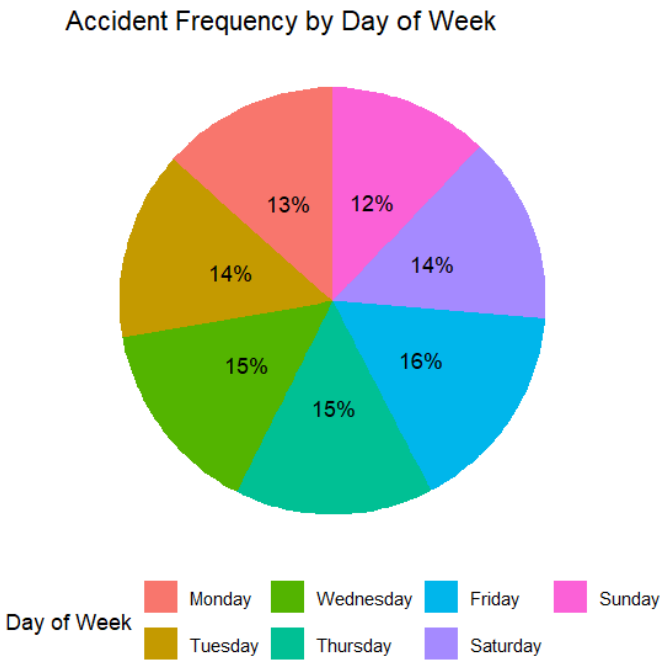
Peak accident hour:



From the visual we can see that peak hour for accident is from rush hour 3pm to 7pm. Where the peak time is 3pm. In the morning, driving to work at 8am is peak time for accidents. The accident rate is lower at nighttime and early morning suggesting it is directly related to traffic hours.

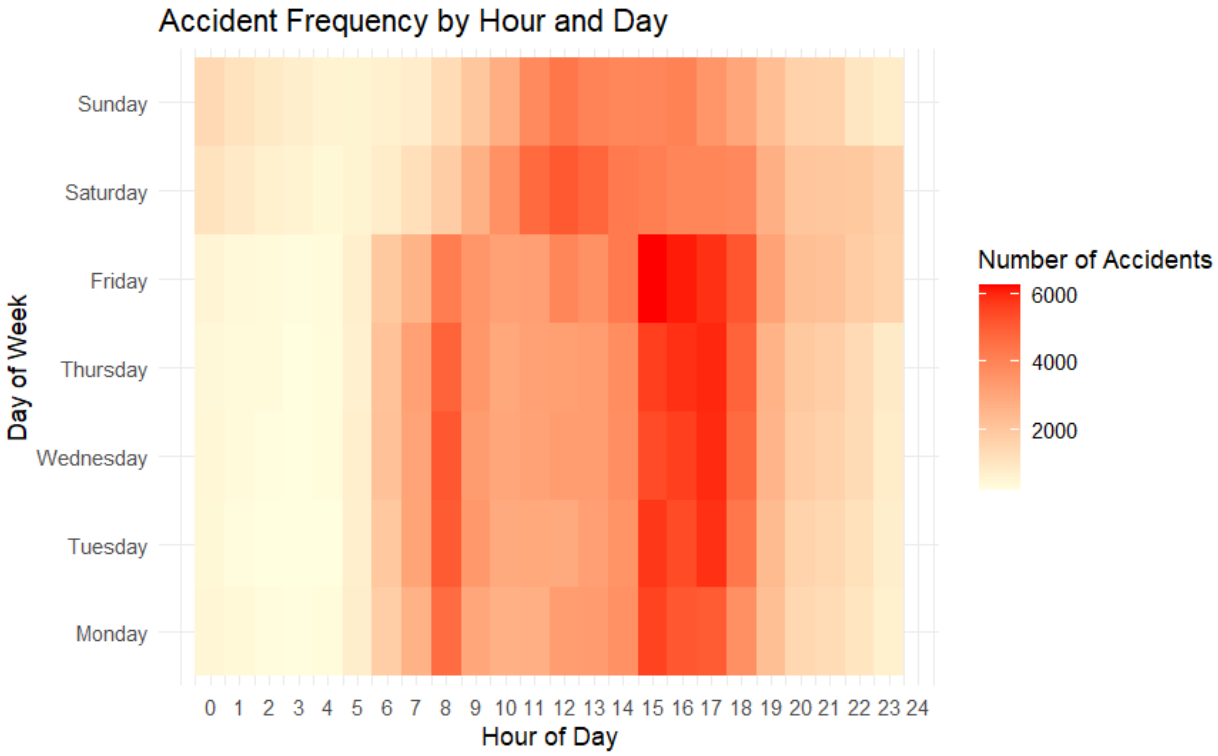
Peak accident day:

ACCIDENT_DAY <fctr>	n <int>
Monday	56737
Tuesday	60126
Wednesday	62643
Thursday	63573
Friday	67765
Saturday	59849
Sunday	51095



Overall, 16% of accidents happen over Friday. Considering it is the start of the weekend means high active time. Whereas Sundays shows the least amount of accident suggesting least activity.

Trend:



Toward the end of work week the frequency of accident is noticeably higher than rest. Weekend activity starting from Friday and Saturday shows dispersion in reasons of accidents including travelling and partying.

Safety Implications:

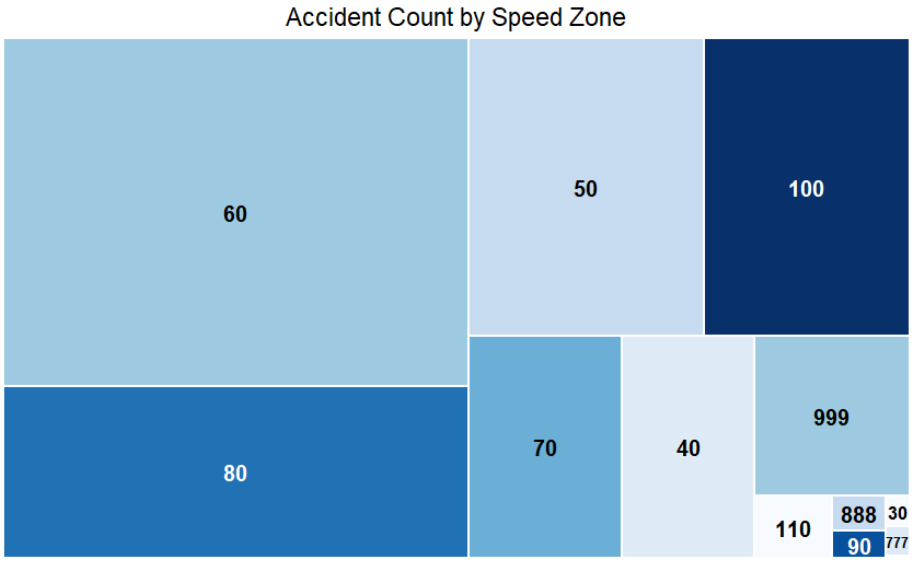
- Infrastructure improvements like broader lanes for less traffic.
- More traffic rules and regulations
- Adjustment in traffic signals
- Hybrid work modules for employees

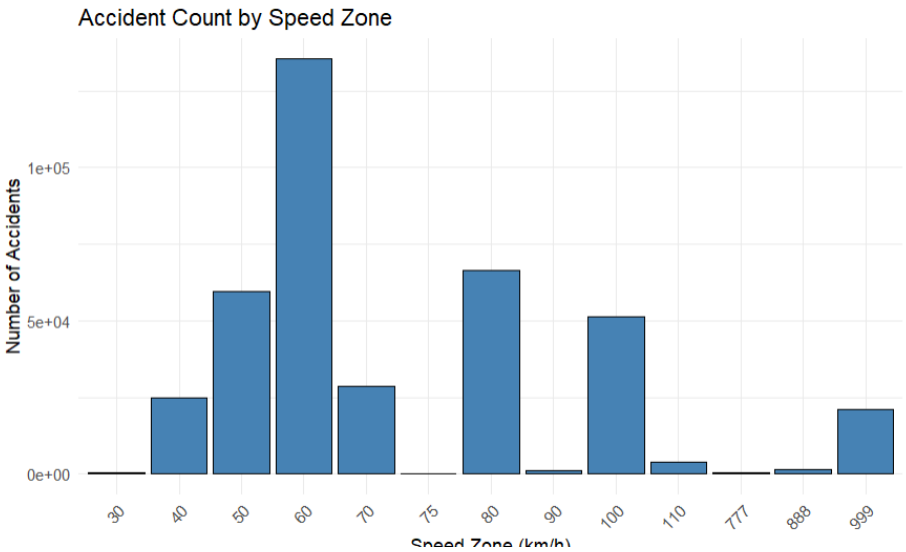
Task 5 – Accident Trends and Safety Concerns

This analyse accident trends in Victoria over four key factors, speed_zone, accident user type, accident severity in age group and time of the accident.

1.) Speed Zone:

30	40	50	60	70	75	80	90	100	110	777	888	999
648	26249	62896	144803	30610	39	71549	1282	54779	4307	617	1695	22314

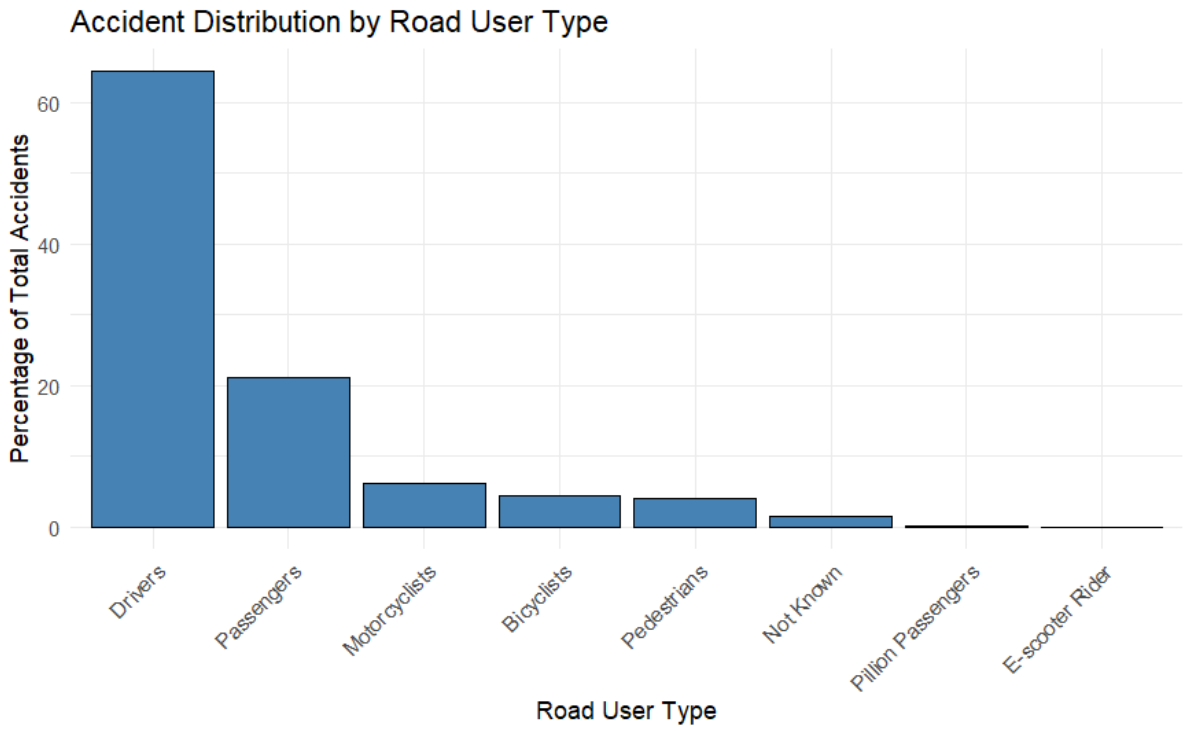




It shows impact on accident severity. In the speed zone 60 to 80 km/h most number of accident occurs. 100km/h suggesting freeway has higher number of accidents as well. Where as, 40km/h shows rural areas has some sort of accidents.

2.) Accident Distriution by Road User type:

ROAD_USER_TYPE_DESC <chr>	n <int>
Drivers	266813
Passengers	87202
Motorcyclists	25649
Bicyclists	18189
Pedestrians	16881
Not Known	6117
Pillion Passengers	935
E-scooter Rider	2

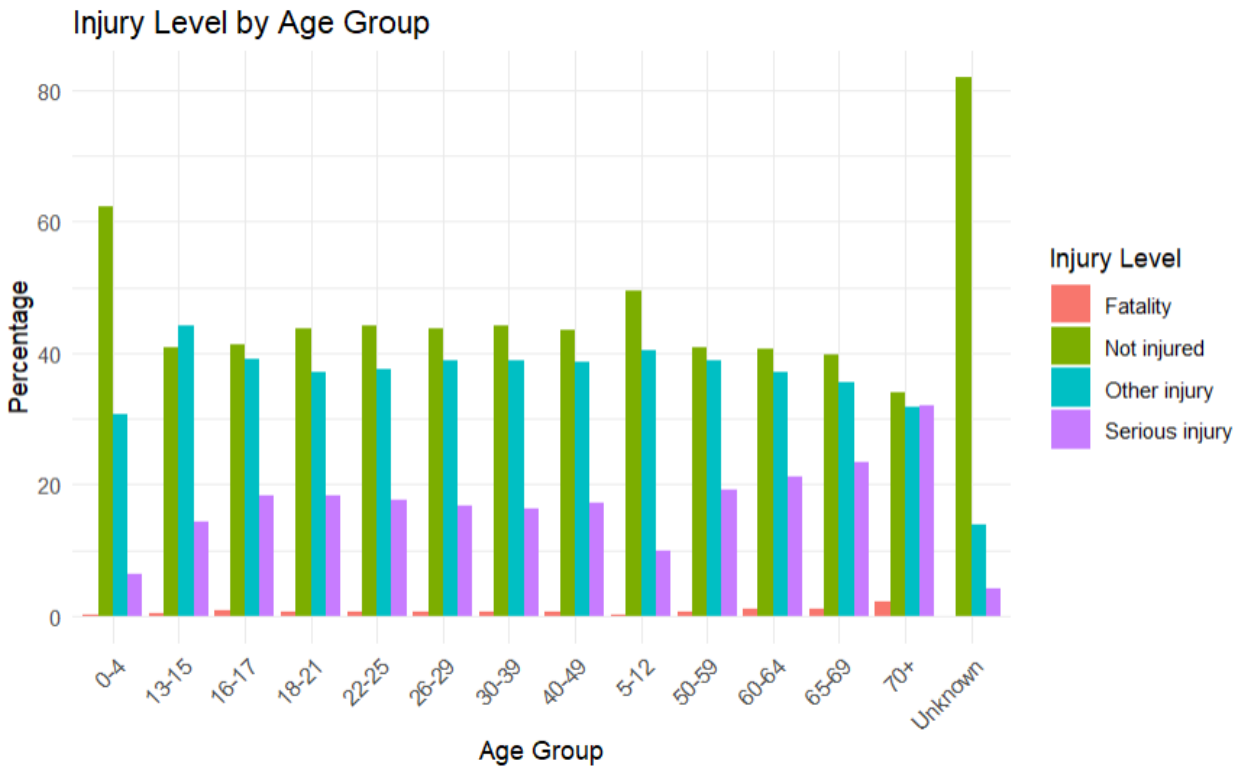


60% of accidents happen involve drivers. Second being passengers getting dragged into accidents. Motorcyclist and Bicyclist has least proportion whereas pillion passengers and e-scooter rider is just few number of accidents.

3.) Injury level in age group:

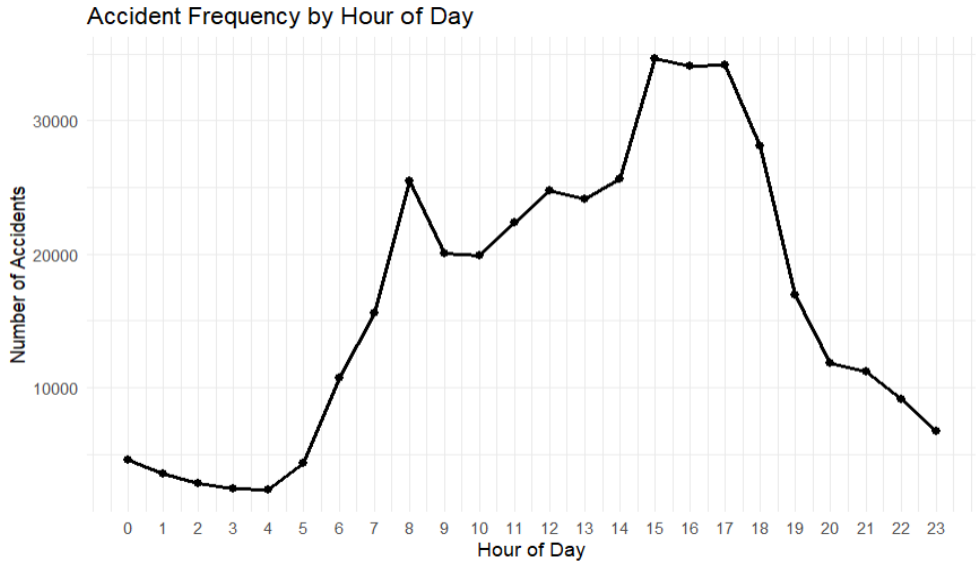
For calculating this, I created dataframe for age group also cross tabs with INJ_LEVEL_DISC. This will help me to analyse injury level due to accident in different age group.

	Fatality	Not injured	Other injury	Serious injury
0-4	0.32460005	62.39276606	30.77904011	6.50359379
13-15	0.50435580	40.91395384	44.29160935	14.29008100
16-17	0.93405038	41.42371922	39.21596377	18.42626663
18-21	0.74137509	43.72889650	37.12013702	18.40959139
22-25	0.58274779	44.10657330	37.54046860	17.77021032
26-29	0.64514430	43.74605014	38.83768696	16.77111860
30-39	0.60819674	44.10022658	38.97361831	16.31795837
40-49	0.63638394	43.55481834	38.62802641	17.18077132
5-12	0.25787188	49.38246471	40.49945711	9.86020630
50-59	0.76926088	40.93720639	38.96414031	19.32939242
60-64	1.03520757	40.69889648	37.02049396	21.24540200
65-69	1.21131494	39.82635485	35.53423890	23.42809130
70+	2.12071205	33.95597065	31.83525859	32.08805871
Unknown	0.07562008	81.90411373	13.84603751	4.17422868



It shows that at the age of 70 people are most likely to suffer fatal injury in accidents, which might be the old age. In all the accident luckily the injury rate is not high. But it balances out the serious injury rate.

4.) Time of Occurrence:



The peak accident occurs during the rush hour of 3 to 6 pm. Early morning and late nights shows less number of accidents, suggesting less amount of traffic.

Recommendation:

- Create awareness in public regarding helmet and seatbelt wearing.
- Tailored campaigns for young and older drivers, emphasizing safe driving practices and the impact of fatigue.
- Increase police presence and enforce traffic laws strictly.
- Make hefty fines for causing an accident.
- To control high speed zones come up with law changes.
- Improve infrastructure, like broader the lanes, protect bicycle lanes and pedestrian path.