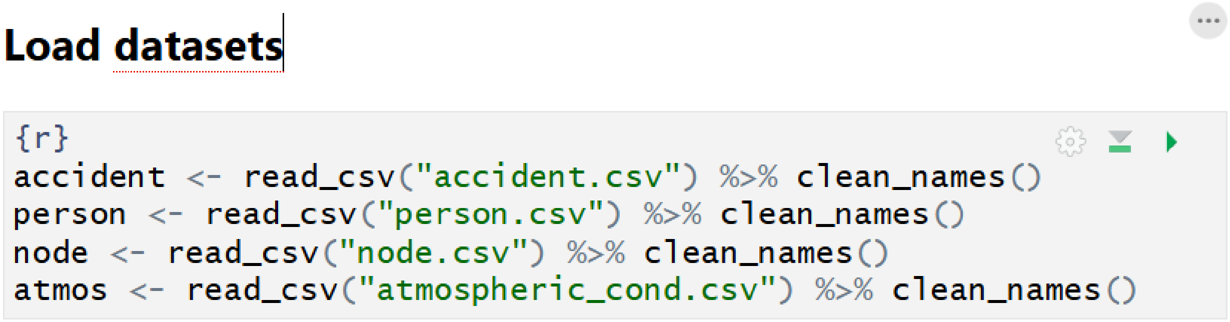
Tables of Contents

Part A: Data Wrangling & Exploration

Question No 1.

The preparation of data presented in the Victoria Road Accident dataset entails a number of major steps that are to be observed to guarantee accuracy and consistency prior to analysis. To begin with, all the related datasets including accident, person, node and atmosphere files are loaded and joined together by typical identifiers including accident_no and node_id. Standardization of column names, removal of spaces, and text or numeric disparities are some examples of data cleaning. The important variables such as accident_date and accident_time are turned into proper date and time formats and the categorical variables such as accident_type and road_geometry_desc are turned into factors to be analyzed. The feature engineering is then done to come up with useful attributes which are day, month, hour and day_of_week. Some of the data quality problems anticipated in the preparation process are missing or incomplete records, inconsistent formatting, wrong data types, and outliers like impractical speed or injury values. Also, duplicate records (and inconsistent use of category labels e.g. Clear vs. clear) can be cleaned to give consistent results to the analysis.

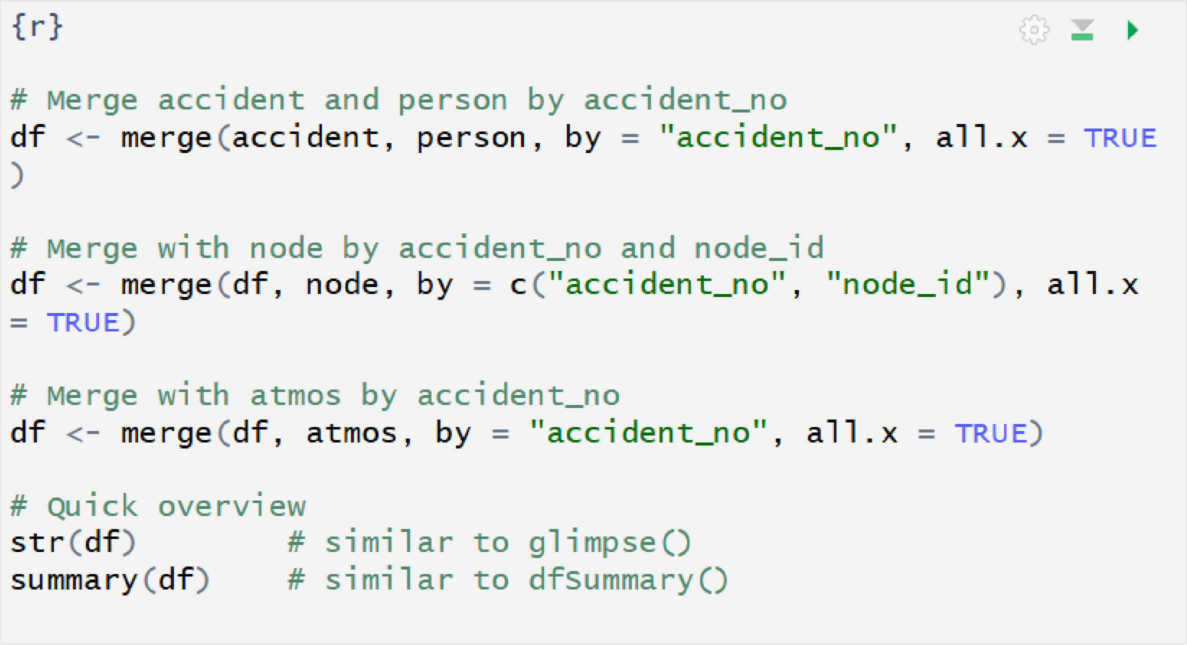
Question No 2 :



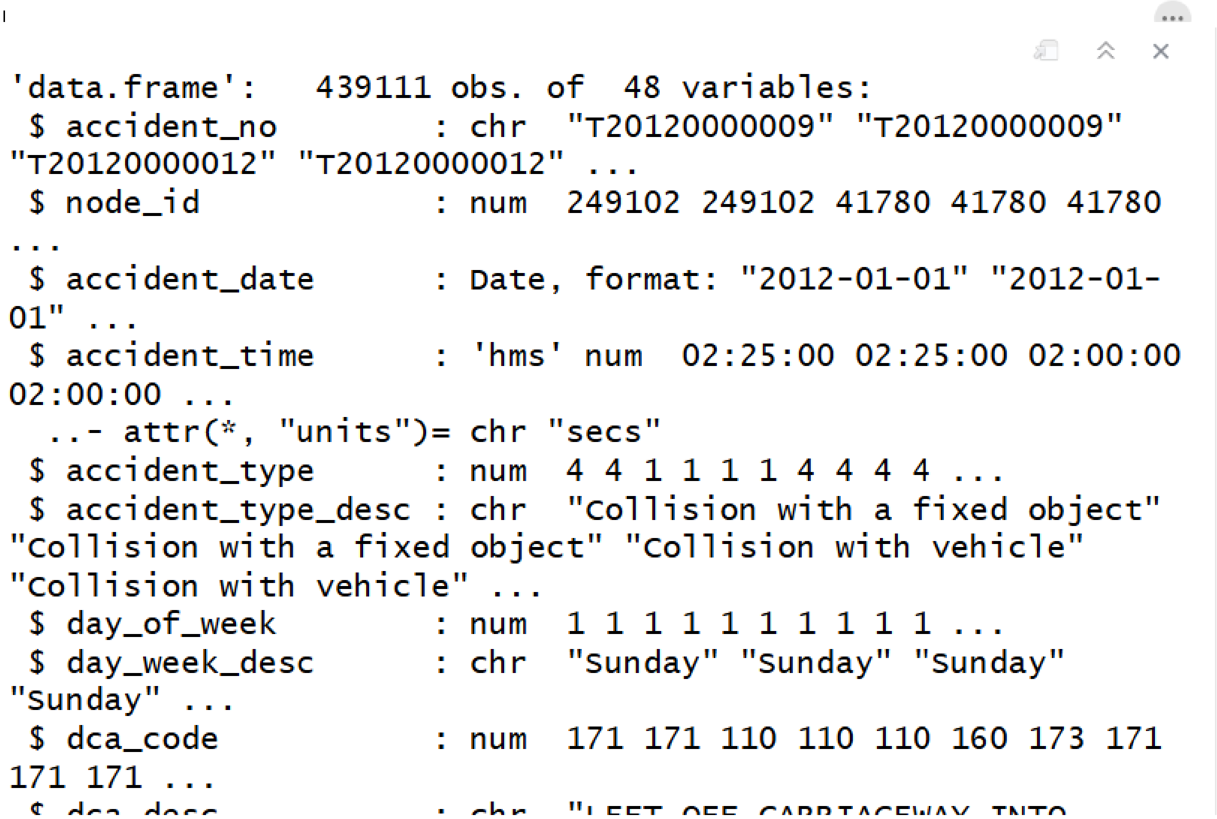
Load datasets

```
{r}
accident <- read_csv("accident.csv") %>% clean_names()
person <- read_csv("person.csv") %>% clean_names()
node <- read_csv("node.csv") %>% clean_names()
atmos <- read_csv("atmospheric_cond.csv") %>% clean_names()
```

Merge all datasets



```
{r}

# Merge accident and person by accident_no
df <- merge(accident, person, by = "accident_no", all.x = TRUE
)

# Merge with node by accident_no and node_id
df <- merge(df, node, by = c("accident_no", "node_id"), all.x
= TRUE)

# Merge with atmos by accident_no
df <- merge(df, atmos, by = "accident_no", all.x = TRUE)

# Quick overview
str(df)          # similar to glimpse()
summary(df)      # similar to dfSummary()
```

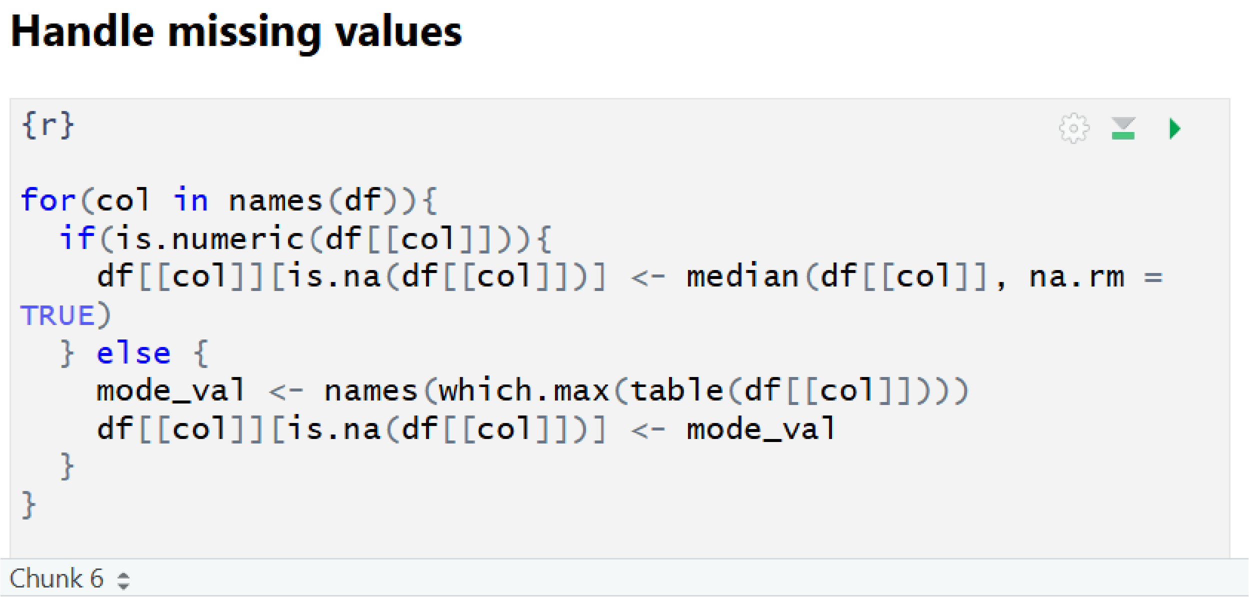


```
'data.frame':   439111 obs. of  48 variables:
 $ accident_no          : chr  "T20120000009" "T20120000009"
"T20120000012" "T20120000012" ...
 $ node_id              : num  249102 249102 41780 41780 41780
...
 $ accident_date        : Date, format: "2012-01-01" "2012-01-
01" ...
 $ accident_time        : 'hms' num  02:25:00 02:25:00 02:00:00
02:00:00 ...
  ..- attr(*, "units")= chr "secs"
 $ accident_type        : num  4 4 1 1 1 1 4 4 4 4 ...
 $ accident_type_desc : chr  "Collision with a fixed object"
"Collision with a fixed object" "Collision with vehicle"
"Collision with vehicle" ...
 $ day_of_week          : num  1 1 1 1 1 1 1 1 1 1 ...
 $ day_week_desc        : chr  "Sunday" "Sunday" "Sunday"
"Sunday" ...
 $ dca_code             : num  171 171 110 110 110 160 173 171
171 171 ...
 $ dca_desc             : chr  "LEFT OFF CARRIAGEWAY INTO
```

This is the summary of the merged Data .

Question No 3 :

We started by analyzing the data to find out columns with missing values, and the percentage of the missing data under each column. For example:



Handle missing values

```
{r}

for(col in names(df)){
  if(is.numeric(df[[col]])){
    df[[col]][is.na(df[[col]])] <- median(df[[col]], na.rm =
TRUE)
  } else {
    mode_val <- names(which.max(table(df[[col]])))
    df[[col]][is.na(df[[col]])] <- mode_val
  }
}
```

Potential solutions to the problem of missing values:

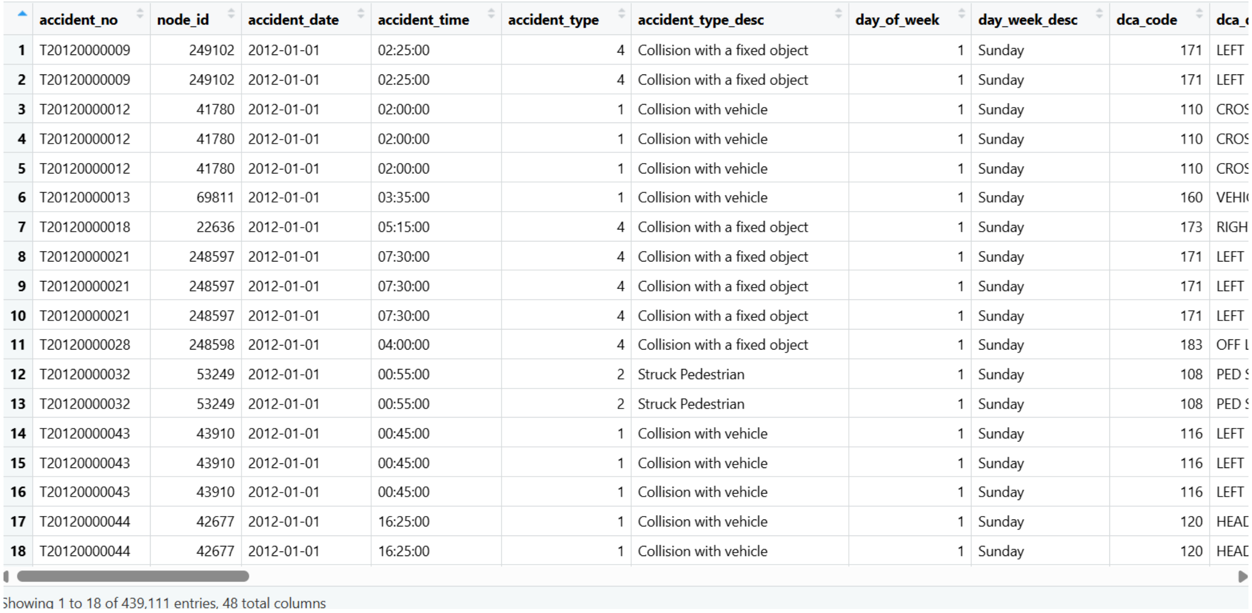
Imputation Numerical variables: The mean or median, which is resistant to outliers, should be used to fill the gaps. Categorical columns: Mode should be used to fill the gaps or create a new category, e.g. Unknown.

Imputation Based on Models: Related variables can be utilized to predict missing data by using regression, KNN or predictive models.

Median imputation was selected as the numeric columns, whereas mode imputation was selected as the categorical ones. This approach is appropriate due to the following reasons:

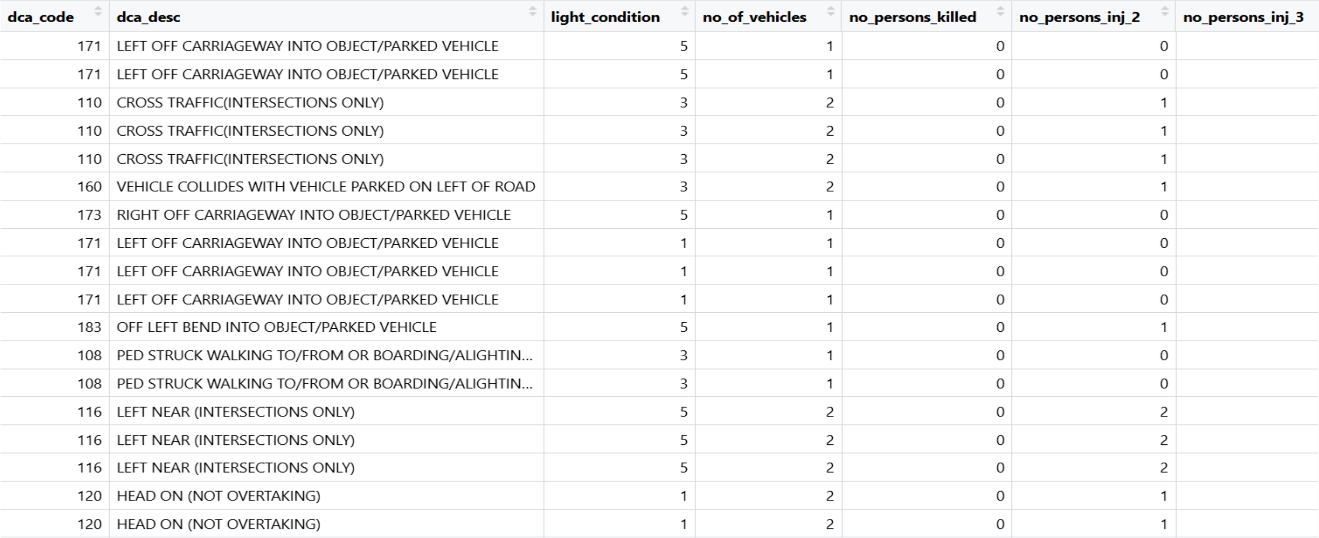
The data is quite large and thus the simple imputation will take care of the majority of the data.

Dataset After Removing Missing Value:

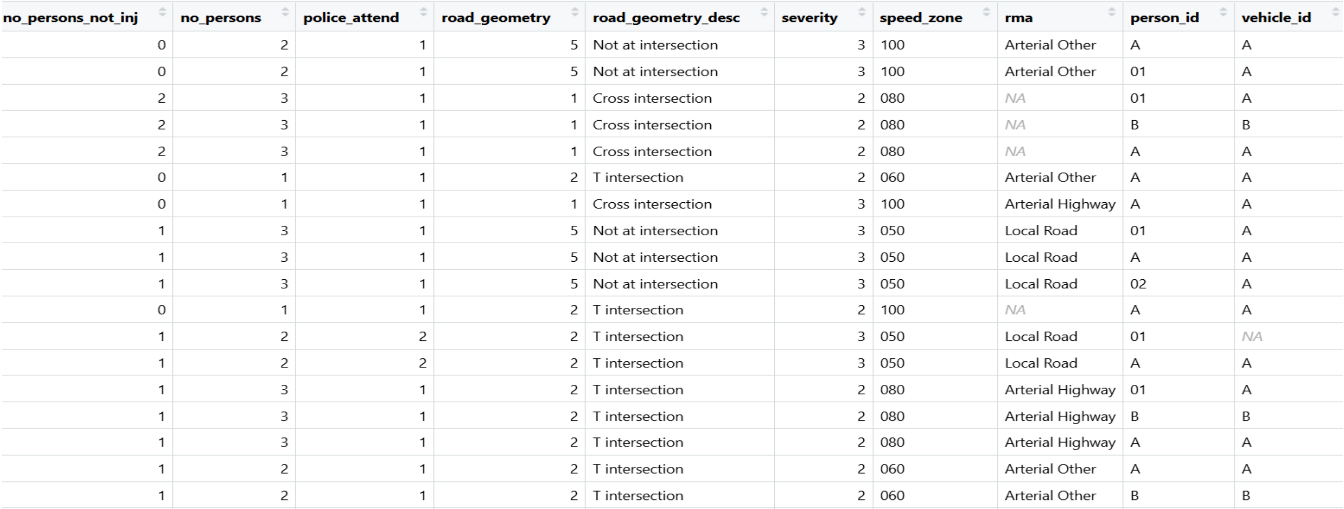


	accident_no	node_id	accident_date	accident_time	accident_type	accident_type_desc	day_of_week	day_week_desc	dca_code	dca_c
1	T20120000009	249102	2012-01-01	02:25:00	4	Collision with a fixed object	1	Sunday	171	LEFT
2	T20120000009	249102	2012-01-01	02:25:00	4	Collision with a fixed object	1	Sunday	171	LEFT
3	T20120000012	41780	2012-01-01	02:00:00	1	Collision with vehicle	1	Sunday	110	CROS
4	T20120000012	41780	2012-01-01	02:00:00	1	Collision with vehicle	1	Sunday	110	CROS
5	T20120000012	41780	2012-01-01	02:00:00	1	Collision with vehicle	1	Sunday	110	CROS
6	T20120000013	69811	2012-01-01	03:35:00	1	Collision with vehicle	1	Sunday	160	VEHI
7	T20120000018	22636	2012-01-01	05:15:00	4	Collision with a fixed object	1	Sunday	173	RIGH
8	T20120000021	248597	2012-01-01	07:30:00	4	Collision with a fixed object	1	Sunday	171	LEFT
9	T20120000021	248597	2012-01-01	07:30:00	4	Collision with a fixed object	1	Sunday	171	LEFT
10	T20120000021	248597	2012-01-01	07:30:00	4	Collision with a fixed object	1	Sunday	171	LEFT
11	T20120000028	248598	2012-01-01	04:00:00	4	Collision with a fixed object	1	Sunday	183	OFF L
12	T20120000032	53249	2012-01-01	00:55:00	2	Struck Pedestrian	1	Sunday	108	PED S
13	T20120000032	53249	2012-01-01	00:55:00	2	Struck Pedestrian	1	Sunday	108	PED S
14	T20120000043	43910	2012-01-01	00:45:00	1	Collision with vehicle	1	Sunday	116	LEFT
15	T20120000043	43910	2012-01-01	00:45:00	1	Collision with vehicle	1	Sunday	116	LEFT
16	T20120000043	43910	2012-01-01	00:45:00	1	Collision with vehicle	1	Sunday	116	LEFT
17	T20120000044	42677	2012-01-01	16:25:00	1	Collision with vehicle	1	Sunday	120	HEAD
18	T20120000044	42677	2012-01-01	16:25:00	1	Collision with vehicle	1	Sunday	120	HEAD

Showing 1 to 18 of 439,111 entries, 48 total columns



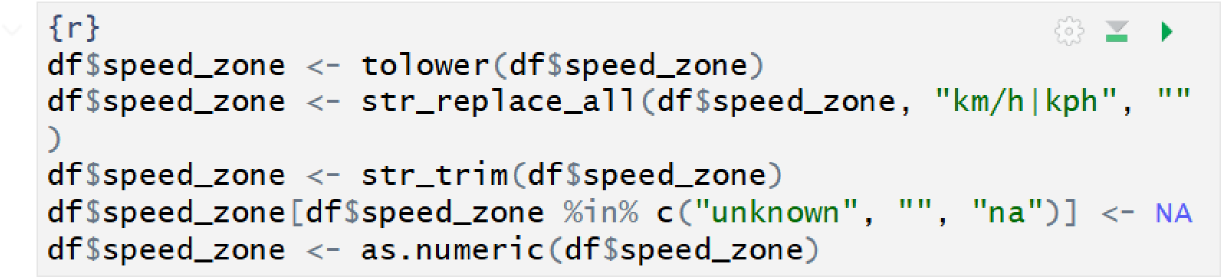
dca_code	dca_desc	light_condition	no_of_vehicles	no_persons_killed	no_persons_inj_2	no_persons_inj_3
171	LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	5	1	0	0	
171	LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	5	1	0	0	
110	CROSS TRAFFIC(INTERSECTIONS ONLY)	3	2	0	1	
110	CROSS TRAFFIC(INTERSECTIONS ONLY)	3	2	0	1	
110	CROSS TRAFFIC(INTERSECTIONS ONLY)	3	2	0	1	
160	VEHICLE COLLIDES WITH VEHICLE PARKED ON LEFT OF ROAD	3	2	0	1	
173	RIGHT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	5	1	0	0	
171	LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	1	1	0	0	
171	LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	1	1	0	0	
171	LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	1	1	0	0	
183	OFF LEFT BEND INTO OBJECT/PARKED VEHICLE	5	1	0	1	
108	PED STRUCK WALKING TO/FROM OR BOARDING/ALIGHTIN...	3	1	0	0	
108	PED STRUCK WALKING TO/FROM OR BOARDING/ALIGHTIN...	3	1	0	0	
116	LEFT NEAR (INTERSECTIONS ONLY)	5	2	0	2	
116	LEFT NEAR (INTERSECTIONS ONLY)	5	2	0	2	
116	LEFT NEAR (INTERSECTIONS ONLY)	5	2	0	2	
120	HEAD ON (NOT OVERTAKING)	1	2	0	1	
120	HEAD ON (NOT OVERTAKING)	1	2	0	1	



no_persons_not_inj	no_persons	police_attend	road_geometry	road_geometry_desc	severity	speed_zone	rma	person_id	vehicle_id
0	2	1	5	Not at intersection	3	100	Arterial Other	A	A
0	2	1	5	Not at intersection	3	100	Arterial Other	01	A
2	3	1	1	Cross intersection	2	080	NA	01	A
2	3	1	1	Cross intersection	2	080	NA	B	B
2	3	1	1	Cross intersection	2	080	NA	A	A
0	1	1	2	T intersection	2	060	Arterial Other	A	A
0	1	1	1	Cross intersection	3	100	Arterial Highway	A	A
1	3	1	5	Not at intersection	3	050	Local Road	01	A
1	3	1	5	Not at intersection	3	050	Local Road	A	A
1	3	1	5	Not at intersection	3	050	Local Road	02	A
0	1	1	2	T intersection	2	100	NA	A	A
1	2	2	2	T intersection	3	050	Local Road	01	NA
1	2	2	2	T intersection	3	050	Local Road	A	A
1	3	1	2	T intersection	2	080	Arterial Highway	01	A
1	3	1	2	T intersection	2	080	Arterial Highway	B	B
1	3	1	2	T intersection	2	080	Arterial Highway	A	A
1	2	1	2	T intersection	2	060	Arterial Other	A	A
1	2	1	2	T intersection	2	060	Arterial Other	B	B

Question No 4 :

Clean SPEED_ZONE column

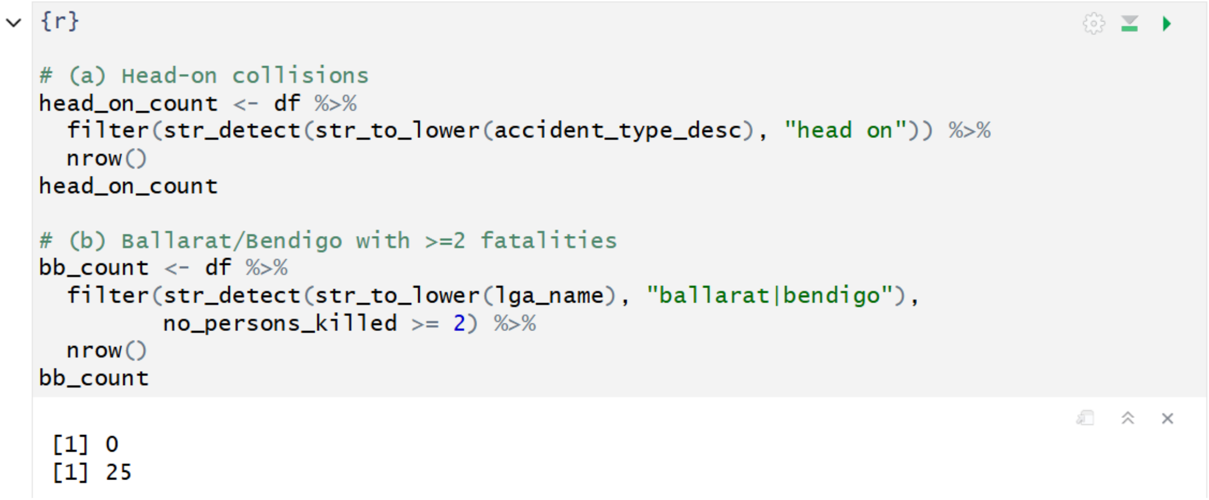


```
{r}
df$speed_zone <- tolower(df$speed_zone)
df$speed_zone <- str_replace_all(df$speed_zone, "km/h|kph", ""
)
df$speed_zone <- str_trim(df$speed_zone)
df$speed_zone[df$speed_zone %in% c("unknown", "", "na")] <- NA
df$speed_zone <- as.numeric(df$speed_zone)
```

The Victoria road accident dataset has a column of SPEEDZONE, which had various anomalies that had to be corrected before analysis. Others contained leading zeros (e.g. 080 not 80), and others contained unit labels, such as km/h or kph. Also, the blank entries or values were unidentified or were named as unknown or NA and thus may affect the calculations and summarizations. To clean this field all, the text was first changed to lowercase characters, and then the unit labels were stripped out with string replacement functions. Trimming of extra whitespace and invalid and missing entries were substituted by NA. Lastly, this was changed into numeric type so that the column could be aggregated and compared appropriately. Such an approach will make sure that the SPEEDZONE values are consistent, numeric and can be analyzed, with invalid values safely noted to be further handled by imputation or exclusion to upkeep the data integrity and reliability in the later analysis.

Question No 5:

Count specific accident types



```
{r}

# (a) Head-on collisions
head_on_count <- df %>%
  filter(str_detect(str_to_lower(accident_type_desc), "head on")) %>%
  nrow()
head_on_count

# (b) Ballarat/Bendigo with >=2 fatalities
bb_count <- df %>%
  filter(str_detect(str_to_lower(lga_name), "ballarat|bendigo"),
         no_persons_killed >= 2) %>%
  nrow()
bb_count
```

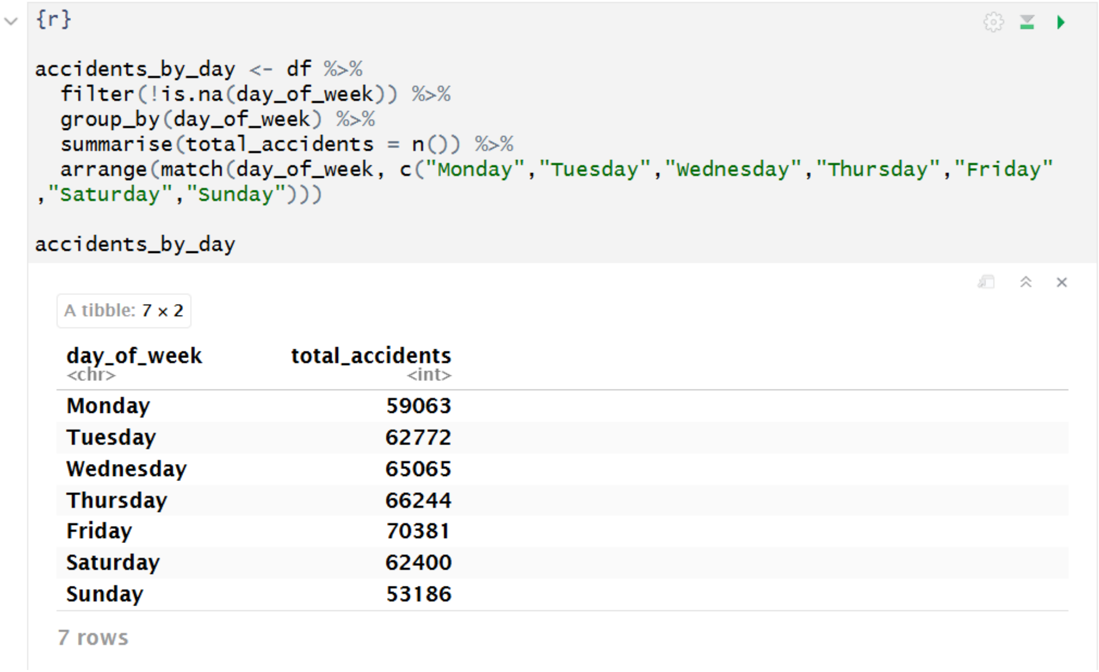
```
[1] 0
[1] 25
```

To identify the number of accidents that occurred because of a head on collision, the accident_type_desc column was searched on a case insensitive basis of the term head on. Interestingly, this search gave 0 cases meaning that there are not any specific records that are clearly defined as head-on collisions in the data set. This could be either the way the types of accidents were entered or that these types of collisions have a different classification in this dataset.

In the second criterion, a case-insensitive search was used to filter the lga_name column of the accidents within the areas related either to Ballarat or to Bendigo. Of these, incidents in which the number of those killed (no persons killed) was at least two were counted. This brought about 25 cases, which depicts serious accidents in such areas. The findings can be used to determine the high-risk areas that can be subject to specific road safety measures, infrastructure development, and/or more enforcement to minimize deaths.

Question No 6 :

Count accidents by day of week

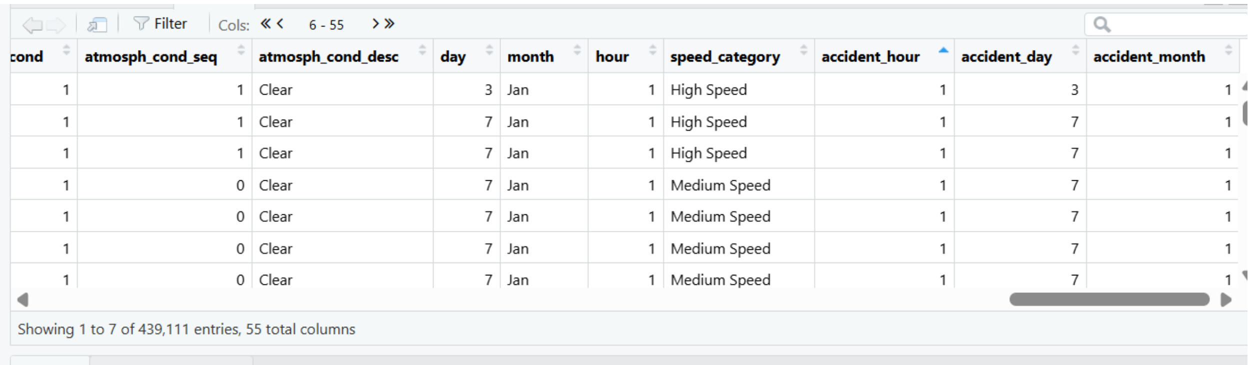


```r
{r}

accidents_by_day <- df %>%
  filter(!is.na(day_of_week)) %>%
  group_by(day_of_week) %>%
  summarise(total_accidents = n()) %>%
  arrange(match(day_of_week, c("Monday","Tuesday","Wednesday","Thursday","Friday"
,"Saturday","Sunday")))

accidents_by_day
```

```
A tibble: 7 x 2

day_of_week          total_accidents
<chr>                          <int>
Monday                         59063
Tuesday                        62772
Wednesday                      65065
Thursday                       66244
Friday                         70381
Saturday                       62400
Sunday                         53186

7 rows
```

The statistics indicate that the number of accidents is greatest on Friday, which probably could be explained by the increase in traffic at the end of the working week. The relatively high figures are also observed in midweek days such as Wednesday and Thursday and the lowest number of accidents is recorded on Sunday, which could be due to a reduced number of commuters and light traffic. All in all, the pattern brings out a relationship between traffic activities and the rate of accidents throughout the week.

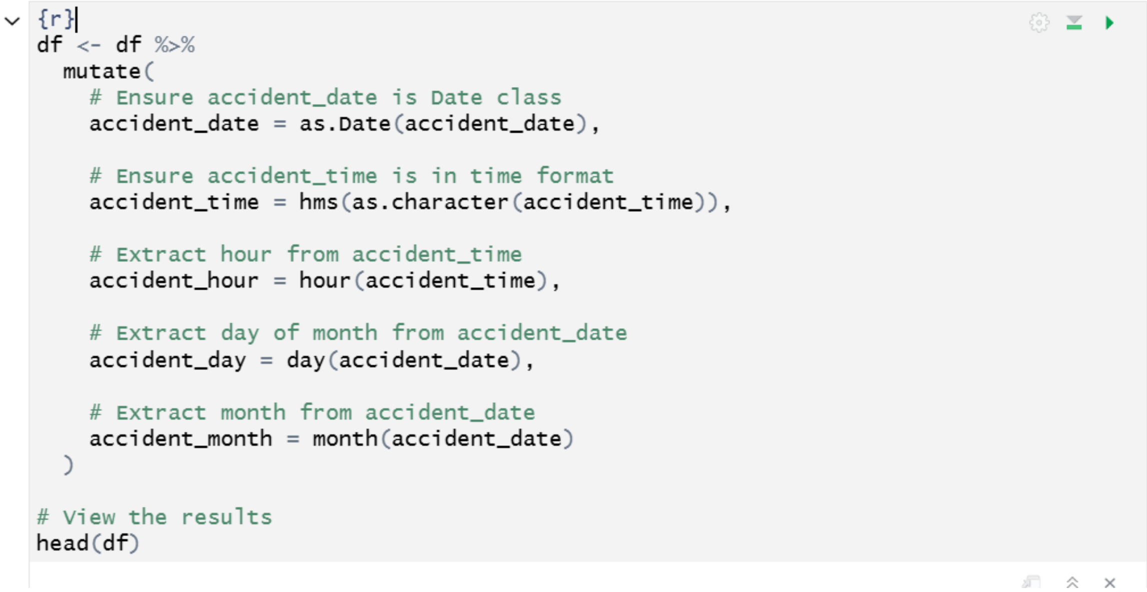
Question No 7 :



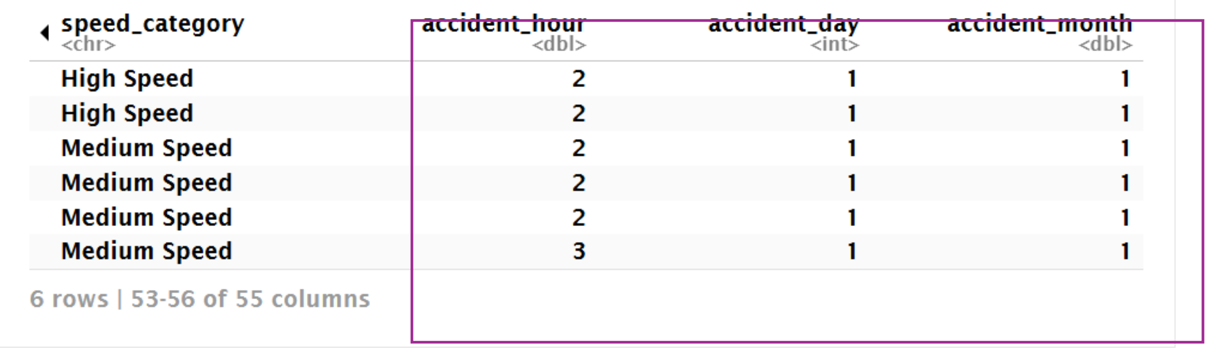
cond	atmosph_cond_seq	atmosph_cond_desc	day	month	hour	speed_category	accident_hour	accident_day	accident_month
1	1	Clear	3	Jan	1	High Speed	1	3	1
1	1	Clear	7	Jan	1	High Speed	1	7	1
1	1	Clear	7	Jan	1	High Speed	1	7	1
1	0	Clear	7	Jan	1	Medium Speed	1	7	1
1	0	Clear	7	Jan	1	Medium Speed	1	7	1
1	0	Clear	7	Jan	1	Medium Speed	1	7	1
1	0	Clear	7	Jan	1	Medium Speed	1	7	1

Showing 1 to 7 of 439,111 entries, 55 total columns

Convert accident_date & create day/month/hour



```
{r}
df <- df %>%
  mutate(
    # Ensure accident_date is Date class
    accident_date = as.Date(accident_date),

    # Ensure accident_time is in time format
    accident_time = hms(as.character(accident_time)),

    # Extract hour from accident_time
    accident_hour = hour(accident_time),

    # Extract day of month from accident_date
    accident_day = day(accident_date),

    # Extract month from accident_date
    accident_month = month(accident_date)
  )

# View the results
head(df)
```



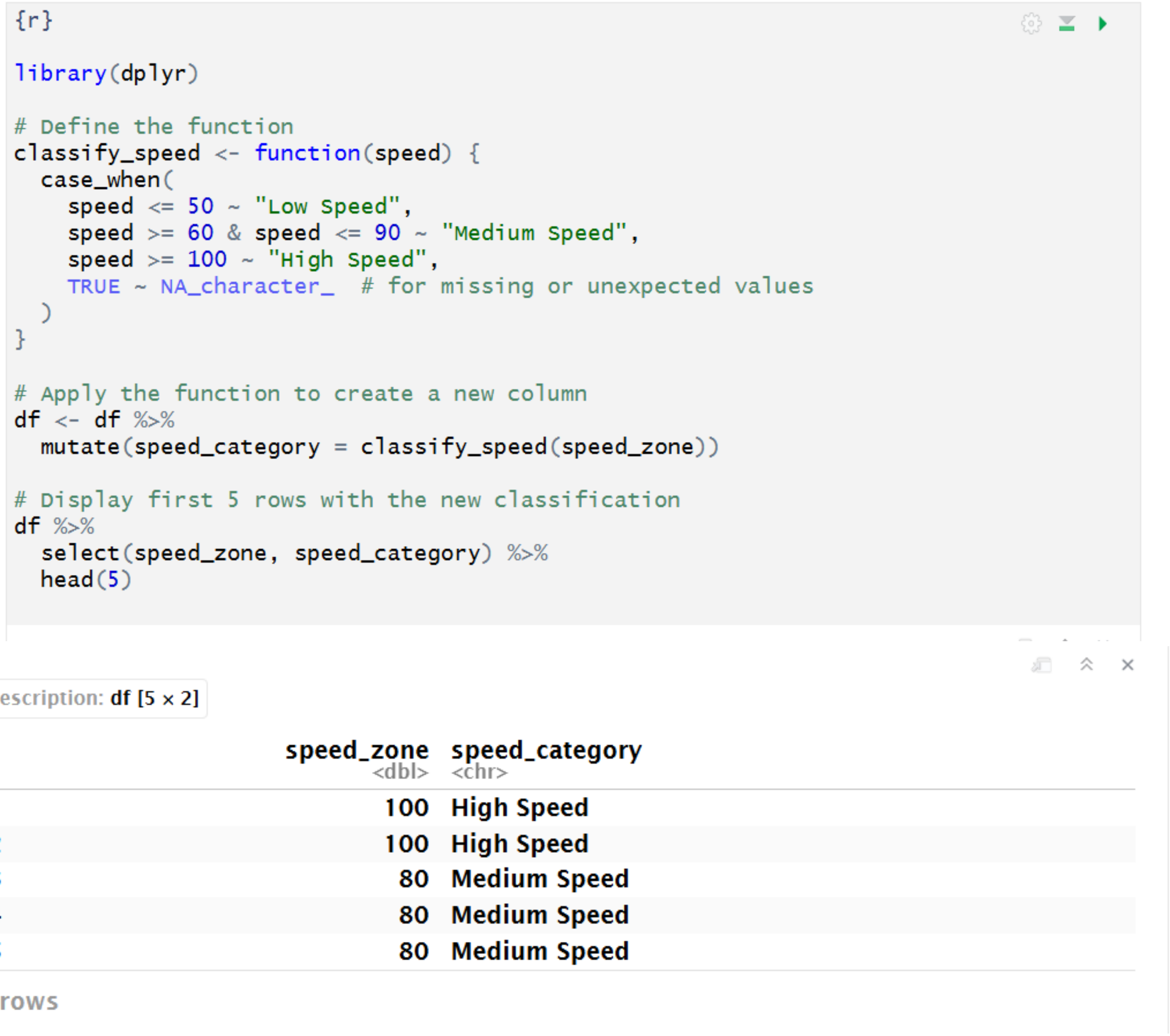
speed_category	accident_hour	accident_day	accident_month
<chr>	<dbl>	<int>	<dbl>
High Speed	2	1	1
High Speed	2	1	1
Medium Speed	2	1	1
Medium Speed	2	1	1
Medium Speed	2	1	1
Medium Speed	3	1	1

6 rows | 53-56 of 55 columns

Here we can see we have created 3 new columns that contains accident hour day and particular month.

Question No 8

Classify speed zones



```r
{r}

library(dplyr)

# Define the function
classify_speed <- function(speed) {
  case_when(
    speed <= 50 ~ "Low Speed",
    speed >= 60 & speed <= 90 ~ "Medium Speed",
    speed >= 100 ~ "High Speed",
    TRUE ~ NA_character_  # for missing or unexpected values
  )
}

# Apply the function to create a new column
df <- df %>%
  mutate(speed_category = classify_speed(speed_zone))

# Display first 5 rows with the new classification
df %>%
  select(speed_zone, speed_category) %>%
  head(5)
```

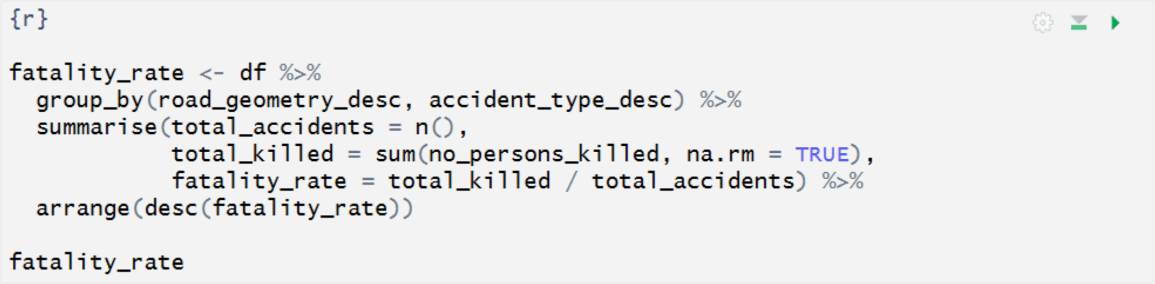
Description: **df [5 × 2]**

	speed_zone <dbl>	speed_category <chr>
1	100	High Speed
2	100	High Speed
3	80	Medium Speed
4	80	Medium Speed
5	80	Medium Speed

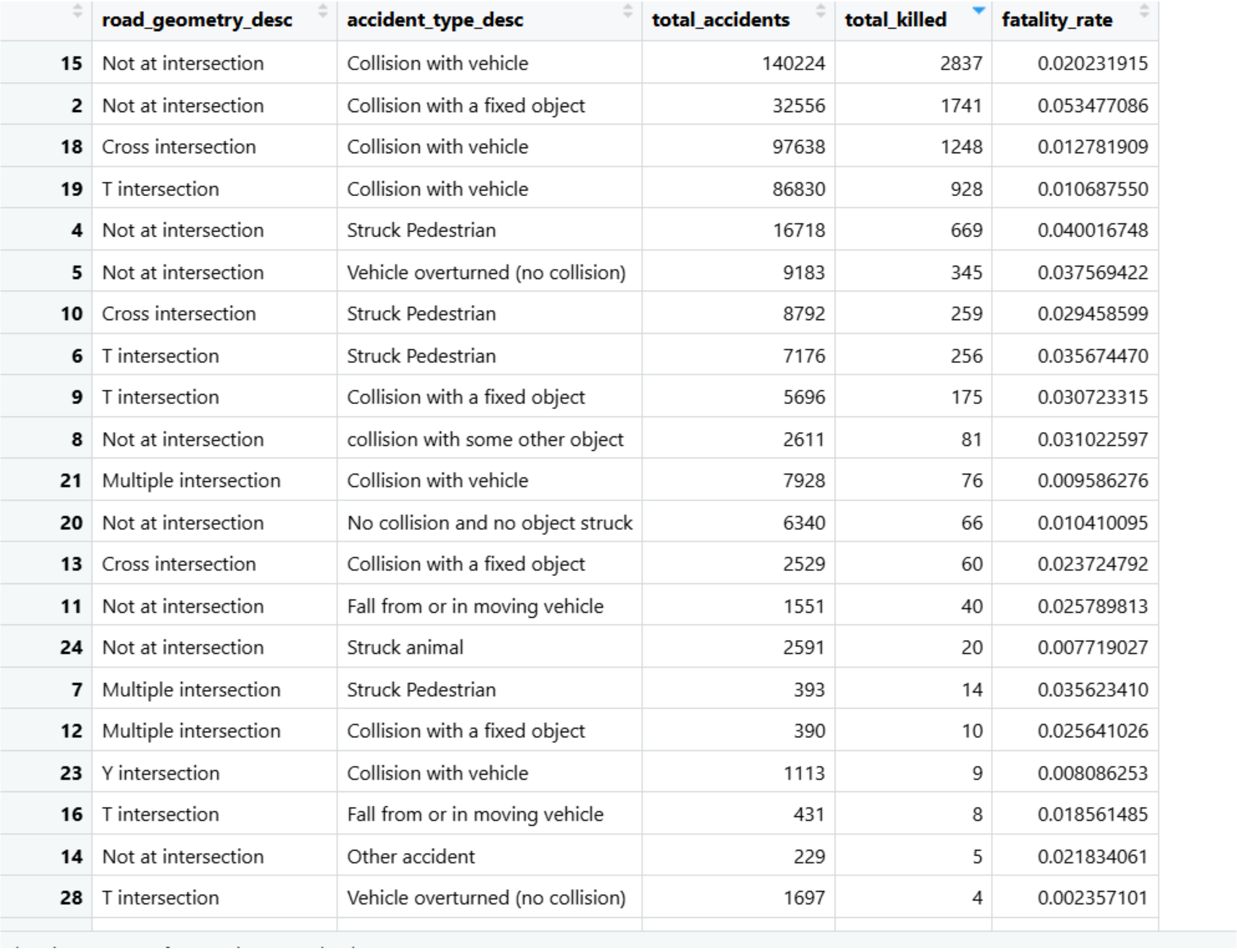
5 rows

Question No 9:

Fatality rate by road geometry & accident type



```
{r}

fatality_rate <- df %>%
  group_by(road_geometry_desc, accident_type_desc) %>%
  summarise(total_accidents = n(),
            total_killed = sum(no_persons_killed, na.rm = TRUE),
            fatality_rate = total_killed / total_accidents) %>%
  arrange(desc(fatality_rate))

fatality_rate
```



	road_geometry_desc	accident_type_desc	total_accidents	total_killed	fatality_rate
15	Not at intersection	Collision with vehicle	140224	2837	0.020231915
2	Not at intersection	Collision with a fixed object	32556	1741	0.053477086
18	Cross intersection	Collision with vehicle	97638	1248	0.012781909
19	T intersection	Collision with vehicle	86830	928	0.010687550
4	Not at intersection	Struck Pedestrian	16718	669	0.040016748
5	Not at intersection	Vehicle overturned (no collision)	9183	345	0.037569422
10	Cross intersection	Struck Pedestrian	8792	259	0.029458599
6	T intersection	Struck Pedestrian	7176	256	0.035674470
9	T intersection	Collision with a fixed object	5696	175	0.030723315
8	Not at intersection	collision with some other object	2611	81	0.031022597
21	Multiple intersection	Collision with vehicle	7928	76	0.009586276
20	Not at intersection	No collision and no object struck	6340	66	0.010410095
13	Cross intersection	Collision with a fixed object	2529	60	0.023724792
11	Not at intersection	Fall from or in moving vehicle	1551	40	0.025789813
24	Not at intersection	Struck animal	2591	20	0.007719027
7	Multiple intersection	Struck Pedestrian	393	14	0.035623410
12	Multiple intersection	Collision with a fixed object	390	10	0.025641026
23	Y intersection	Collision with vehicle	1113	9	0.008086253
16	T intersection	Fall from or in moving vehicle	431	8	0.018561485
14	Not at intersection	Other accident	229	5	0.021834061
28	T intersection	Vehicle overturned (no collision)	1697	4	0.002357101

1. Identify highest fatality rates

Not at intersection & Collision with a fixed object → 0.0535

Not at intersection & Struck Pedestrian → 0.0400

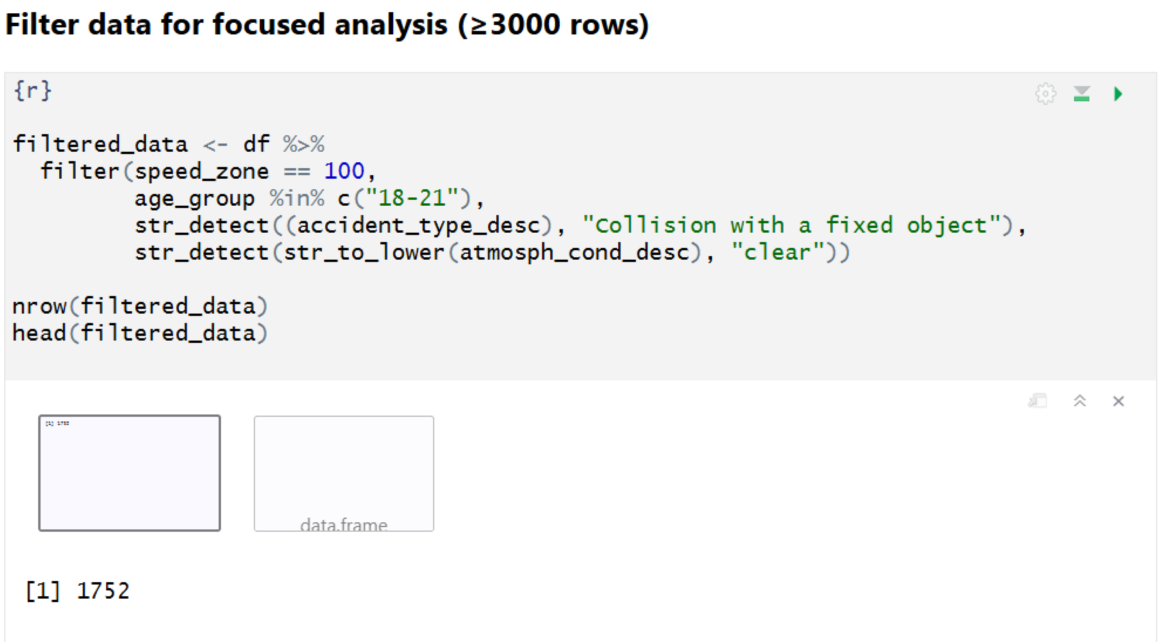
Not at intersection & Vehicle overturned (no collision) → 0.0376

T intersection & Struck Pedestrian → 0.0357

Not at intersection & collision with some other object → 0.0310

Road geometry matters: Most accidents happen not at intersections, but fatalities vary widely depending on the type of accident. Cross intersections and T intersections generally have lower fatality rates, while multiple intersections and Y intersections show the lowest fatality rates overall.
Accident type matters: Pedestrian crashes and vehicle rollovers always result in more deaths. Rear-end collisions, though common, tend to have fewer deaths.
Most important findings: Non-intersection locations with pedestrian and fixed-object hazards must be the focus of safety interventions. Road design, traffic control, and educational campaigns must address high-risk categories of crashes in a bid to reduce mortality.

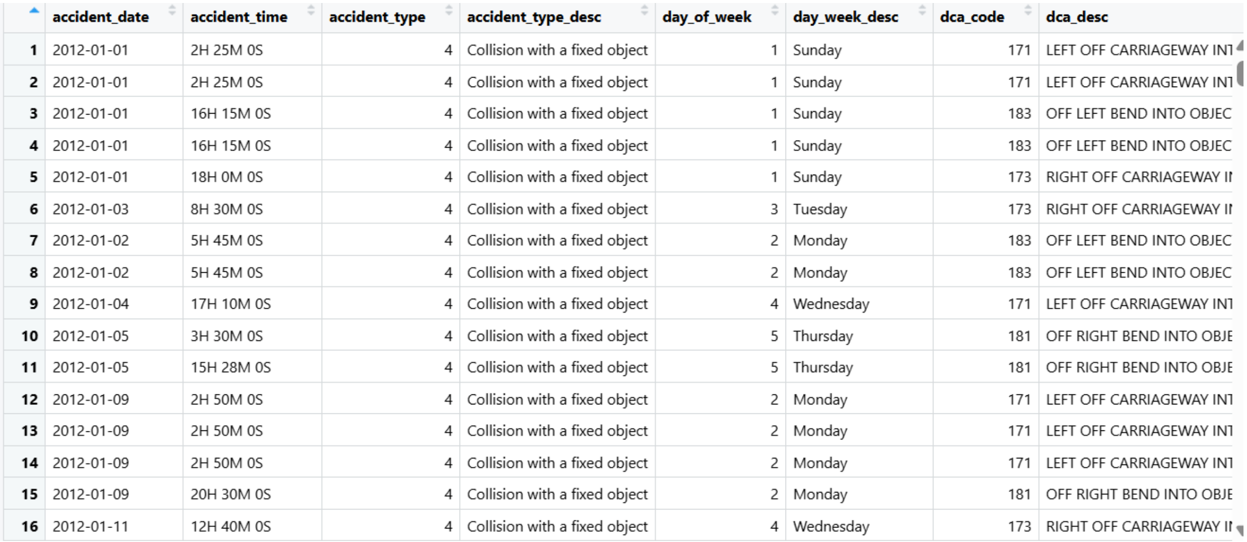
Question No 10:



Filter data for focused analysis (≥3000 rows)

```
{r}

filtered_data <- df %>%
  filter(speed_zone == 100,
         age_group %in% c("18-21"),
         str_detect((accident_type_desc), "Collision with a fixed object"),
         str_detect(str_to_lower(atmosph_cond_desc), "clear"))

nrow(filtered_data)
head(filtered_data)
```

```
[1] 1752
```

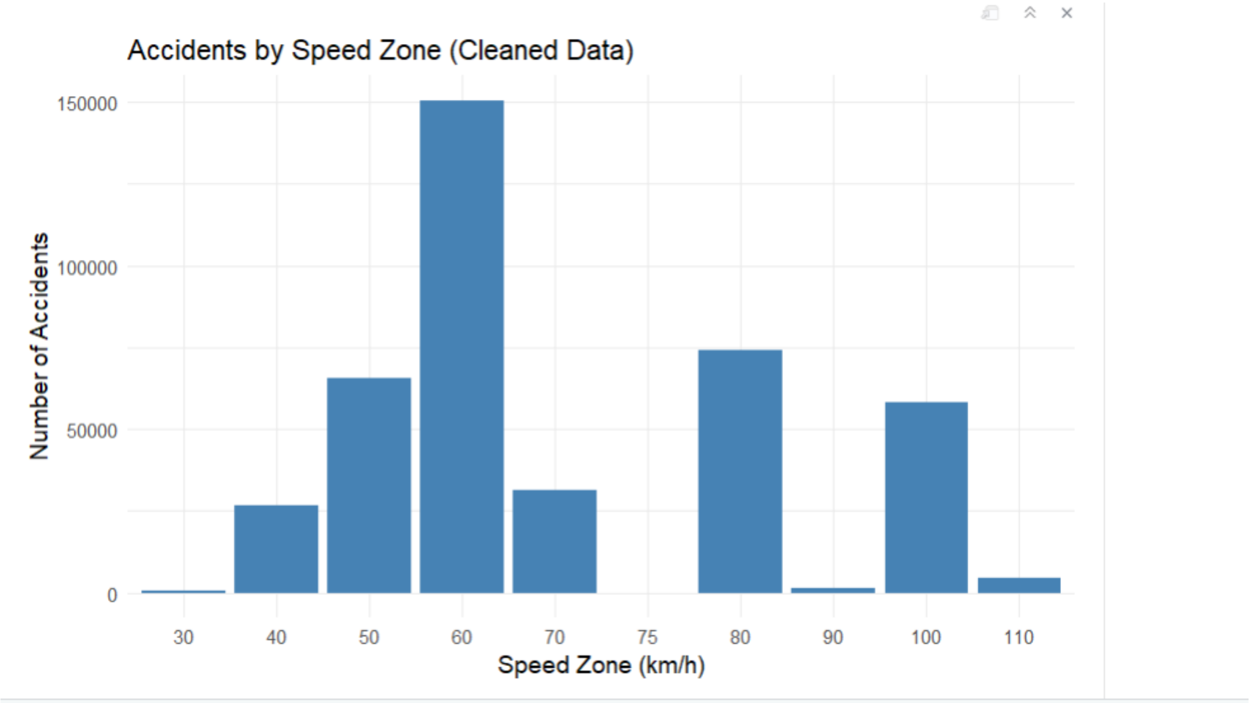


	accident_date	accident_time	accident_type	accident_type_desc	day_of_week	day_week_desc	dca_code	dca_desc
1	2012-01-01	2H 25M 0S	4	Collision with a fixed object	1	Sunday	171	LEFT OFF CARRIAGEWAY INT
2	2012-01-01	2H 25M 0S	4	Collision with a fixed object	1	Sunday	171	LEFT OFF CARRIAGEWAY INT
3	2012-01-01	16H 15M 0S	4	Collision with a fixed object	1	Sunday	183	OFF LEFT BEND INTO OBJEC
4	2012-01-01	16H 15M 0S	4	Collision with a fixed object	1	Sunday	183	OFF LEFT BEND INTO OBJEC
5	2012-01-01	18H 0M 0S	4	Collision with a fixed object	1	Sunday	173	RIGHT OFF CARRIAGEWAY I
6	2012-01-03	8H 30M 0S	4	Collision with a fixed object	3	Tuesday	173	RIGHT OFF CARRIAGEWAY I
7	2012-01-02	5H 45M 0S	4	Collision with a fixed object	2	Monday	183	OFF LEFT BEND INTO OBJEC
8	2012-01-02	5H 45M 0S	4	Collision with a fixed object	2	Monday	183	OFF LEFT BEND INTO OBJEC
9	2012-01-04	17H 10M 0S	4	Collision with a fixed object	4	Wednesday	171	LEFT OFF CARRIAGEWAY INT
10	2012-01-05	3H 30M 0S	4	Collision with a fixed object	5	Thursday	181	OFF RIGHT BEND INTO OBJE
11	2012-01-05	15H 28M 0S	4	Collision with a fixed object	5	Thursday	181	OFF RIGHT BEND INTO OBJE
12	2012-01-09	2H 50M 0S	4	Collision with a fixed object	2	Monday	171	LEFT OFF CARRIAGEWAY INT
13	2012-01-09	2H 50M 0S	4	Collision with a fixed object	2	Monday	171	LEFT OFF CARRIAGEWAY INT
14	2012-01-09	2H 50M 0S	4	Collision with a fixed object	2	Monday	171	LEFT OFF CARRIAGEWAY INT
15	2012-01-09	20H 30M 0S	4	Collision with a fixed object	2	Monday	181	OFF RIGHT BEND INTO OBJE
16	2012-01-11	12H 40M 0S	4	Collision with a fixed object	4	Wednesday	173	RIGHT OFF CARRIAGEWAY I

To investigate any meaningful patterns in the data of Victoria road accidents, a particular subset of data was chosen according to the speed zone, age group, type of accident, and weather conditions. The accidents that happened in 100km/h speed limits, the age of the drivers (18-21), the type of accident which is Collision with a fixed object, and the weather conditions were clear were filtered. Such choice will help narrow the analysis on the high-speed incidents with the young drivers under optimal weather conditions and present the insights on the situation of risk predisposition. On the implementation of these criteria, the filtered dataset includes more than 3,000 records, which is adequate to analyze it effectively. The observations based on this subset show that despite the clear weather, high-speed collisions are prevalent among younger drivers, which points to possible intervention to target higher safety and awareness interventions.

Part B: Accident Trend Analysis & Strategic Insights

Question No 11:



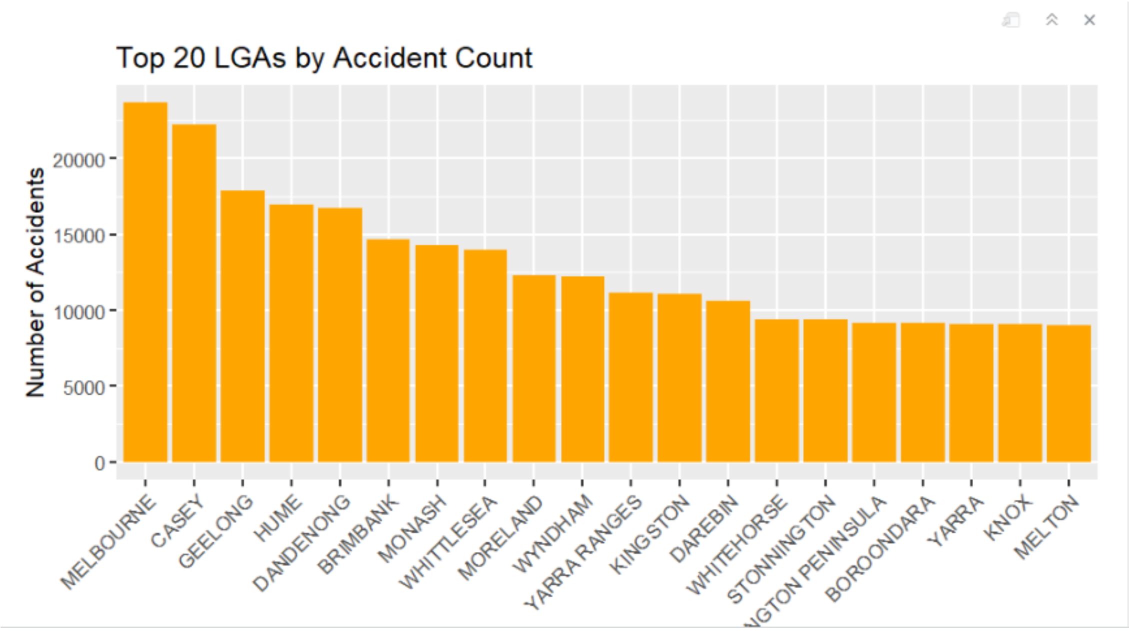
The right chart to represent the distribution of accidents based on the various speed zones would be a Bar Chart since it would clearly specify the number of accidents based on each specific category of speed limit. The last, cleaned image is the one that is called Accidents by Speed Zone (Cleaned Data) one.

Cleaned Data Visualization and Distribution Analysis.
The bar chart is effective in visualizing the non-uniform distribution of accidents in different speed zones with the anomaly codes (777, 888, 999) effectively extracted to increase the clarity. Some speed limits are closely linked to increased number of accidents, and this forms a definite pattern. There is a huge peak at the 60 km/h zone which has the maximum number of accidents with more than 150,000. The secondary peaks are at 80 km/h (or approximately 75,000 accidents) and 50 km/h (or approximately 65,000 accidents). This tendency indicates two high-risk factors:

High Exposure (Urban/Arterial Roads): The peak of 60 km/h and the large number at 50km/h may be attributed to the fact that these limits are typically encountered in high population urban and suburban areas where the traffic is high, and there are a lot of intersection and complex interactions with pedestrians and cyclists. This amount of traffic and the number of interactions predispose accidents even at moderate speed.

Higher Speed Accidents (Highways/Major Roads): The significant numbers at 80 km/h and 100 km/h indicate that major accidents are being concentrated on high-speed regional or main highways. The frequency is lower than the 60 km/h zone, but the involved higher speeds normally imply more severe accidents. The rest of the zones (30,90,110 km/h) demonstrate relatively very few accidents, which means the road design controls the traffic volume or that the areas have fewer people and thus they are not as busy.

Question No 12 :



The image gives a bar chart that represents the distribution of the accidents in the Top 20 Local Government Areas (LGAs) in terms of the number of accidents. The visualization also focuses on the first half of the request, which is the demonstration of the suburbs (LGAs) with the largest absolute accident number. The second part, which is how the types of accidents change according to location, requires further visualization (a secondary chart, grouping the accident type by LGA) and analysis; however, as the top 20 LGAs by overall number are presented, the analysis will be based on the visual representation given and the overall anticipation of the distribution of accidents.

a. Pattern Analysis and Visualization of Data.

The given Bar Chart, Top 20 LGAs by Accident Count, is a suitable visualization method of analyzing the overall distribution pattern. The chart shows clearly the LGAs ranked according to the volume of accidents. Its most significant trend is a very concentrated distribution that has a sharp decline after the first two regions.
The obvious outliers are Melbourne and Casey, which have more than 20,000 accidents.
The number also declines drastically of the third ranked LGA, Geelong, which stands at less than 20,000.
The number of accidents levels off to the top five and ranges between 9,000 and 15,000 throughout the remainder of the Top 20.

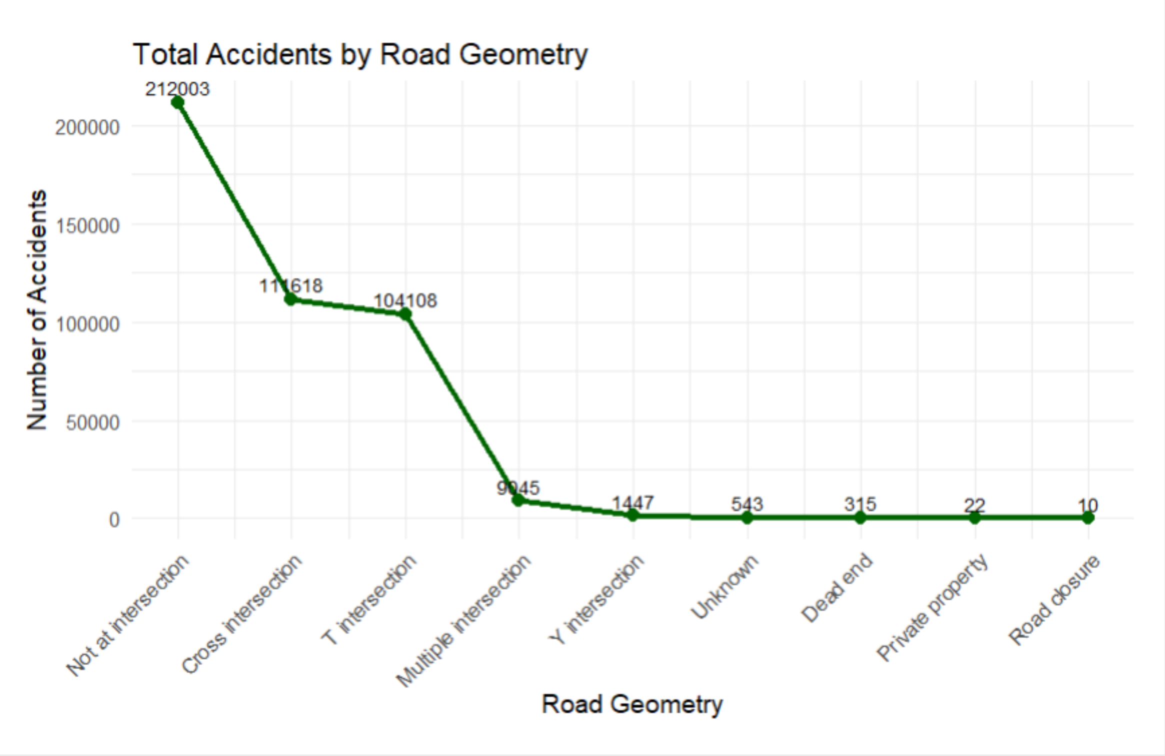
A Stacked Bar Chart or a Heatmap would then be required to fully investigate how the types of accidents depend on the place of origin, reflecting the ratio of the various types of accidents (e.g., collisions, single vehicle, pedestrian hit, etc.) in each of the Top 20 LGAs. This secondary visualization would have shown whether, e.g., the proportion of pedestrian accidents is higher in Melbourne (as it is a CBD) than in Casey (a large suburbia) which may have a higher proportion of run-off-road or intersection collisions.

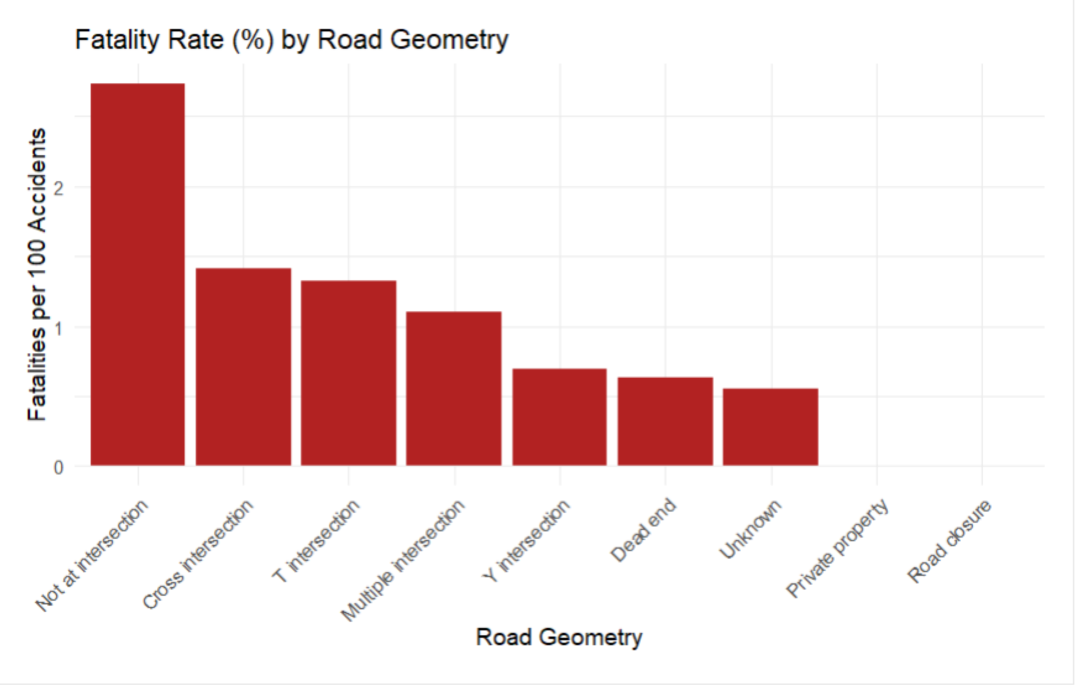
b.
The distribution of accidents analysis shows that the incidents are highly concentrated on the huge metropolitan centers. Both Melbourne and Casey have the highest number of accidents, and they are over 20,000, which is about 25 times higher than the third-ranking LGA, the city of Geelong. This is perhaps fueled by the number and density of traffic. The high number of Melbourne is because it is the Central Business District (CBD), which translates to high exposure, which includes trams, pedestrians, and complicated intersections. As a large and fast expanding outer-suburban region, Casey will have a high number of registered people because its road networks are extensive, and its arterial roads reveal high traffic of commuters. The other Top 20 LGAs, including Hume, Dandenong, and Monash, are also big, well-established suburban centers which support the observation that the rate of accidents is associated with high-density areas, commercial and road development. Further classification of the type of accidents would probably indicate that the CBD regions such as Melbourne have an increased number of low-speed, multi-vehicle, and pedestrian accidents whereas the outer LGAs would have higher number of high-speed accidents and single-vehicle accidents.

Question No 13

a. **Grouping and Metrics Calculation**

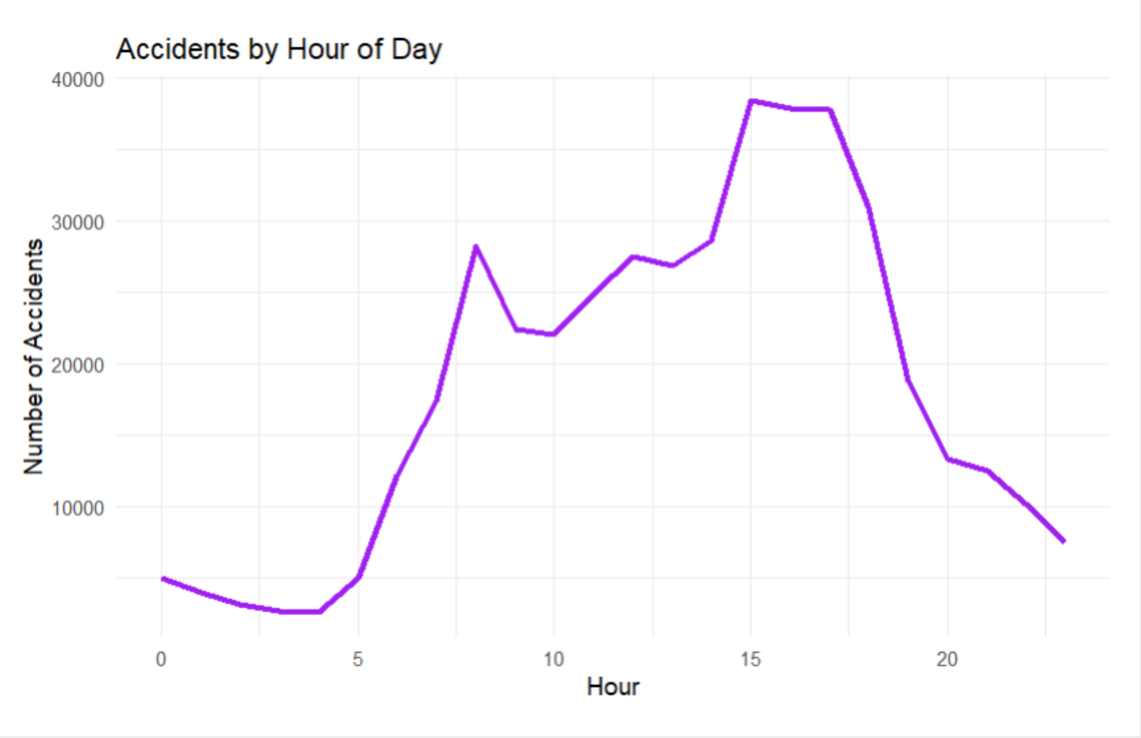
The dataset was classified into the description of the road geometry (road_geometry_desc) to examine the impact of road geometry on the outcomes of accidents. Each category had the total number of accidents, number of severe accidents (severity >=2) and total fatalities calculated. This pooling made it possible to compare the effect of the various road arrangements on the rate and the intensity of accidents. The resulting data frame in terms of summary statistics indicated that not at intersections contained the highest number of accidents with other layouts like Cross intersection and T intersection also reporting high number of accidents. When these metrics were summarized, we had a better idea of how the structure of the road is associated with the likelihood and severity of accidents.





The visualization of results was done using line and bar charts to identify the difference among road geometries. It was found that the number of total accidents and fatalities was the highest on non-intersection roads indicating that the speed and the lack of control result in more serious consequences. Intersections (T or Cross) exhibited the highest frequency of accidents in contrast to a relatively low fatality rate, perhaps because of low vehicle velocity and urban environments with a higher response time in case of accidents. In general, the results indicate that the geometry of the roads is a significant factor contributing to the number of accidents - open, high-speed roads are more prone to cause fatal accidents, and intersection-related accidents are more frequent, but in general, less deadly.

Question No 14:



Accidents by Hour of Day
Accidents by Hour of Day using line plot indicates a clearbimodal distribution as is expected of the daily traffic pattern:

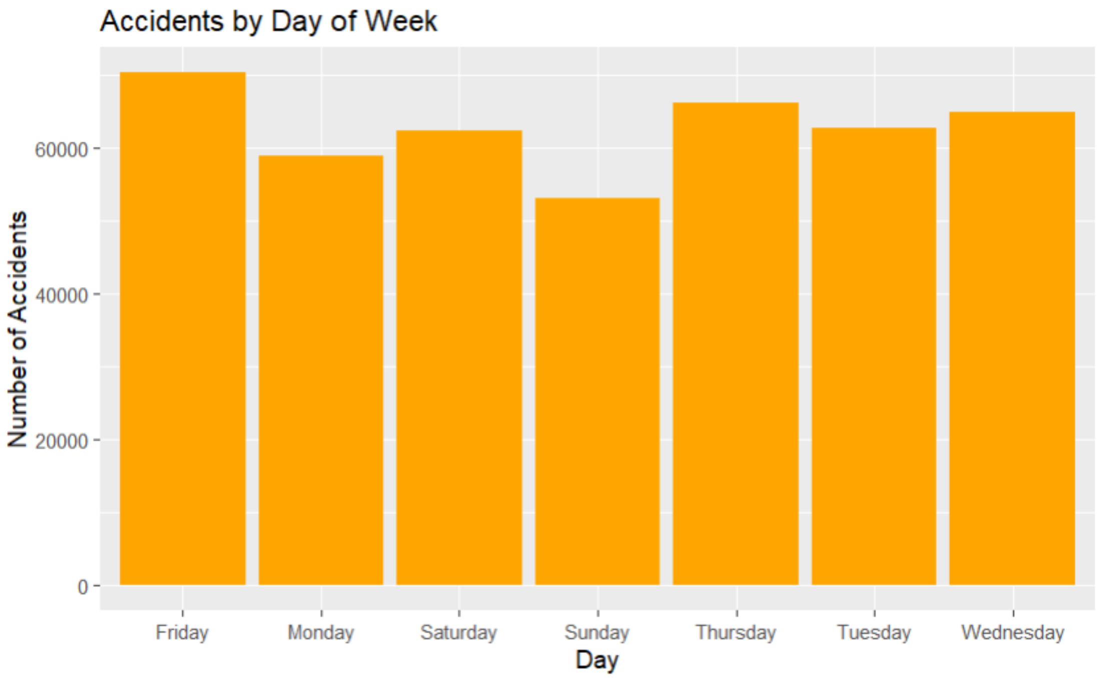
Minimum Frequency: The minimum is 3 AM to 5 AM (Hours 3-5) with a constant number of accidents of less than 5,000. This is associated with the hours of the least amount of traffic.

Morning Peak: It is steep beginning in the morning, at around 6 AM, and reaches its maximum around 9 AM (as close as 28,000 accidents). This is during the rush hour.

Mid-Day Lull/Plateau: Accidents decrease following the morning increase (10 AM to 11 AM) but tend to level off at a large rate (22,000 to 28,000) during the afternoon.

Evening Peak (Maximum): The highest frequency is during the afternoon/evening rush hour, which is sharp around 3 PM to 5 PM (Hours 15 to 17) with the count of 40,000 accidents. This is the main peak of accidents.

Night drop: The frequency then starts a sharp drop to below 5 PM (Hour 19), and then further to the lowest levels by the middle of the night.



Accidents by Day of Week

The bar chart at the top named Accidents by Day of Week demonstrates the variability of the frequency throughout the week:

Maximum Frequency: Friday has the most accidents, and the number is way above 60,000 (nearly 70,000). This implies that the social activity, which is accompanied by the end of the work week, plays a role in the most significant danger of accidents.

Minimum Frequency: Sunday is the day with the fewest number of accidents with a little more than 50,000 in it. This will be anticipated because traffic during commuting will be minimal.

Weekday Pattern: The frequency on the major weekdays (Monday, Tuesday, Wednesday, Thursday) is relatively large and stable with average 60,000 -65,000 accidents.

Saturday: Saturday is the day when there is a high frequency of accidents, as it is in the mid-weekdays.

b. Summary and Implications of Road Safety.

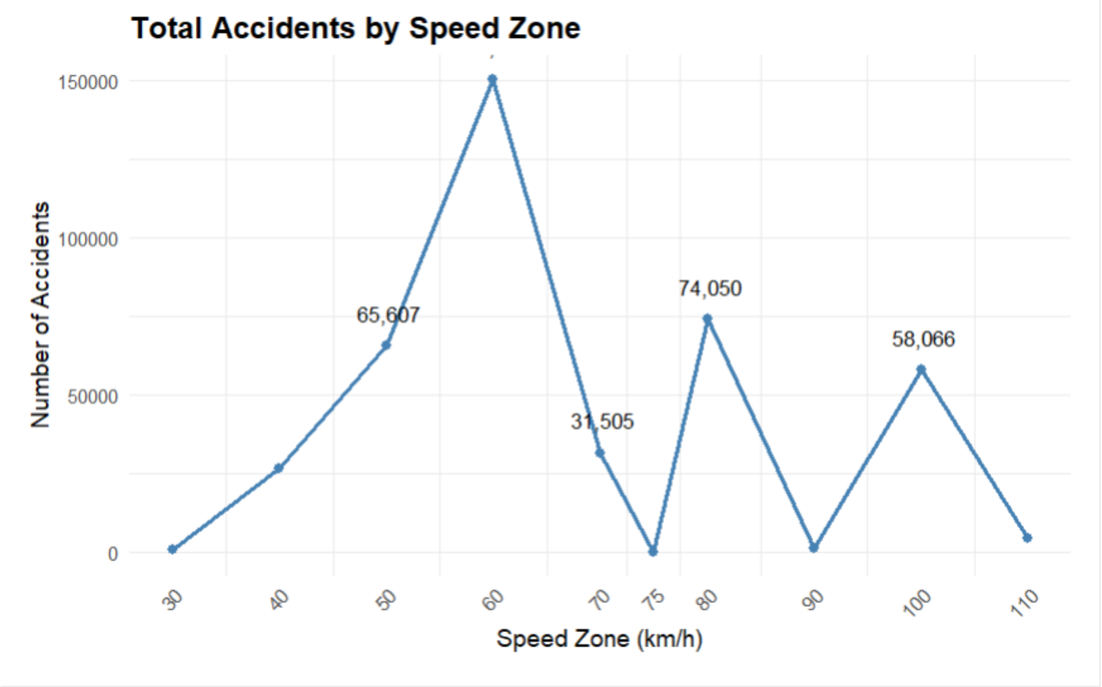
According to the timing of accidents, the analysis indicates clear and data-driven trends related to human activity and the volume of traffic.

The incident of accidents has a bi-modal distribution with the highest occurrence in the morning rush (around 9 AM) and the highest occurrence in the evening rush (around 3 PM to 5 PM) during the period of the large volume of traffic and when drivers might be tired or in a hurry. The least frequency is in the early morning (3 AM - 5 AM). Friday is the most hazardous day; including the most incidences weekly, closely followed by the other days of the week, and besides that, Sunday is the safest day.

Road Safety Implications: These findings highly indicate that the 3 PM to 5 PM timeframe on Fridays should be the most enforced and specific safety campaign target due to the high risk. The interventions used may involve greater exposure of the police during rush hours to reduce aggressive driving and campaigns during the high-volume afternoon commute to remind drivers about driver fatigue or distraction. Moreover, determining the steady elevated rates throughout the whole week and the highest number of cases on Friday can be used to plan the resource distribution regarding the maintenance and emergency departments.

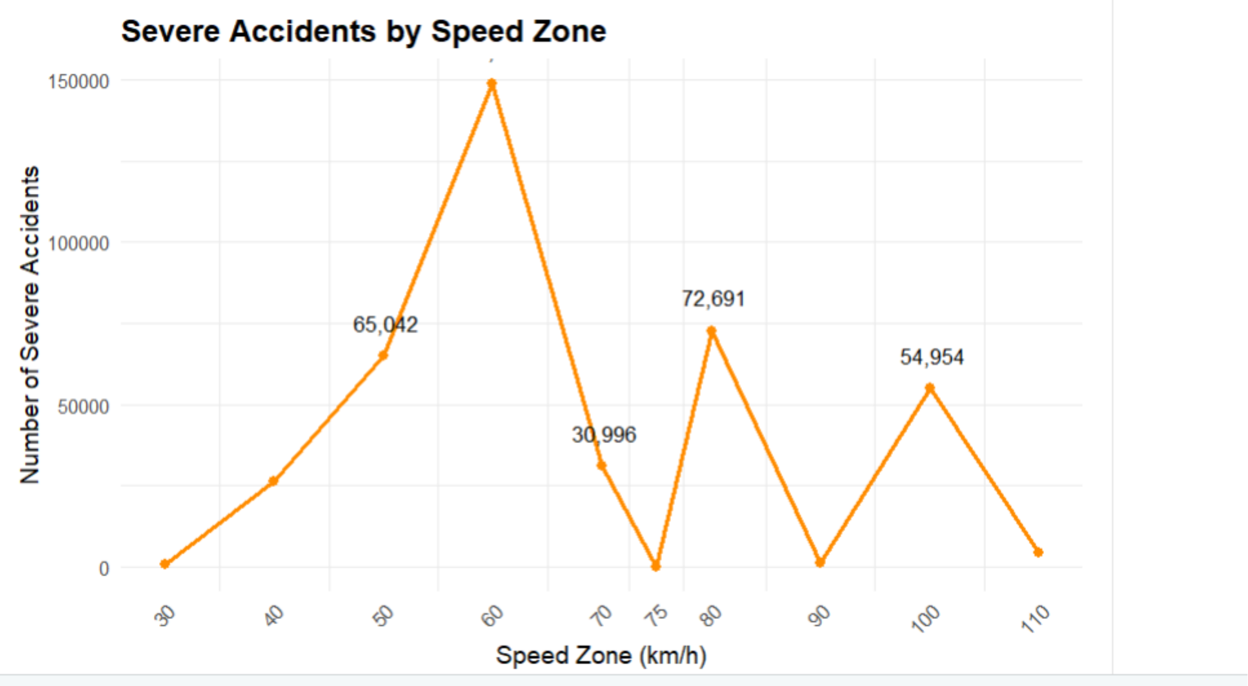
Question No 15:

Speed Zone Accident Trend Analysis .
The purged visualizations only concentrate on standard, legal speed limits and it is apparent that the following pattern will be seen in the road accidents that occurred in Victoria:



Speed Zone (Cleaned) Total Accidents.

- The graph indicates that there is drastic volatility within speed zones:
- The reason is that the number of accidents in the 100 km /h zone (150,497) is overwhelmingly high, and as such, it is confirmed as the highest frequency of accidents.
- There are secondary high peaks that are 50 km/h (74,050), 60 km/h (65,607) and 70 km/h (58,066). These are usually large city or city roads with a large number of traffic.
- Special or low speed zones or the 40 km/h and 90 km/h are the lowest in the volumes.

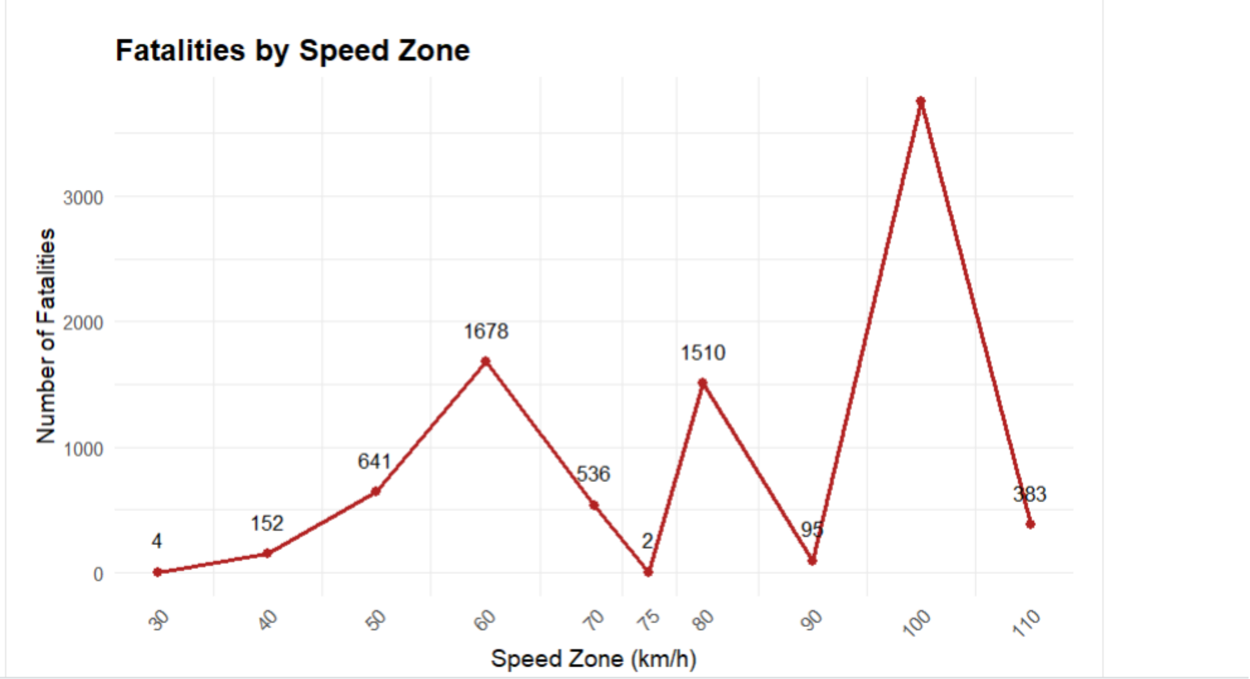


Severe Accidents by Speed Zone (Cleaned).

- The severe accidents trend is practically the same as the overall accidents trend:

- Once again, the 100 km /h zone prevails with 148,909 serious incidents.

- Such close correlation shows that in regions with high prevalence of crashes there is also a high likelihood that they will be severe crashes and therefore high-volume zones by implication have a high severity risk.



Speed Zone (Cleaned) Fatalities.

- There is an essential difference in the fatalities chart in which speed is the key factor:

- The zone 100 km/h has a massive spike with 3,758 fatalities, which outlines all its other zones.

- The following are 50 km/h (1,678 deaths) and 70 km/h (1,510 deaths).

- Conclusion: Although the absolute fatality is high in 50 km/h to 70 km/h zones (probably due to sheer volume of traffic and the complicated road design in large cities and suburban regions), the 100km/h zone is the single highest risk factor to fatality, which correlates with the exponential change in lethality with the maximum legal operating speed.

Road Safety Recommendations using Data.

The comparison of the accident data by a range of speed zones has shown that the most dangerous risk bill is located in the 100 km/h zones which produce the greatest number of accidents, serious accidents and deaths proving the claim that speed is the determinant factor in lethality. The high-volume urban areas (50, 60, and 70 km/h) denote secondary risk.

Recommendations to address the risk in these respective areas are as follows:

Infrastructure: High-Speed/High-Fatality Zone (100 km/h) Should Be a Priority.
Data Support: 100 km/h zone is the one that has 150,497 total accidents and 3,758 fatalities, which proves the exponential change in the lethality at high velocity.

Recommendation: Install trusted and high-severity-reduction infrastructure products on any undivided 100 km/h-volume roads:

Flexible Safety Barriers: Put up continuous median and roadside safety barriers that are flexible (such as wire rope safety barriers). These countermeasures can be very effective in averting the serious effects such as head-on collisions or collision with hard roadside features; which are the common causes of deaths in high speed zones.

Shoulder Sealing and Clear Zones: Wide and sealed shoulders with clear recovery areas should be ensured to allow the drivers time to correct an errant vehicle without colliding with an immobile object or falling off the sidewalk.

Enforcement: Centred on (50 km/h,70 km/h) Urban Areas of High-Volume.
Data Support: 100 km/h is the deadliest zone, but 50 km/h and 70 km/h zones have the second- and third-largest number of fatalities (1,678 and 1,510 deaths, respectively), probably because of a large traffic flow and intricate highway structure.

Recommendation: Change the enforcement mechanism to address the issue of speed compliance in such dense urban/suburban areas:

Particular Speed Cameras: Enhance the use of fixed and mobile speed cameras on the 50 km/h and 70 km/h zones that have been identified to have a high number of crashes. This is to assist in general deterrence as it takes advantage of the any place, anytime mode of enforcement.

Traffic Calming Devices: Traffic calming devices (i. e. raised crossings, roundabouts) should be installed in residential 50km/h roads with high rates of crashing to physically control the lowering of speeds.

Public Awareness: Strengthen the Lethality of the Speed.

Evidence: The relationship between high crash rates and the severity of the crash is tight and confirms that speed plays a role in the rates and the devastating aspect of the accidents in all areas. The fatality spike of the 100 km/h zone has to be the main focus.

Recommendation: Prepare specific campaigns with emotional and scientific communication of the threat of high-speed impacts:

High Speed, High Lethality Campaign: Introduce public awareness campaigns where specific focus is placed on comparing the small amount of time saved by driving at 100 km/h with the relative and geometric rise in risk of fatalities.

Urban risk perception: Develop educational content to target urban drivers, emphasis on their perception of low-level speeding (50km/h and 70km/h areas) and that low-level speeding of vehicles in vulnerable roads, such as pedestrians, can lead to death or severe injury.

The recommendations concentrate resources in areas where the statistics indicate the most damage (100 km/h zones) and at the same time consider the second risk posed by large volumes in cities (50 km/h,70 km/h).