Factors Affecting Road Safety and SPER Debt Collection in Queensland, 2021

PREPARED BY GROUP 10

Nihlahtuzzahra | Jenlarp Wattanakul

Jin Chai | Yifan Zhang

Yusi Cheng | Mansi Mohan Darekar

We give consent for this report and the video of our final presentation to be used as a teaching resource.

Executive Summary

Background:

Queensland, like any other state in Australia, is facing road safety challenges that affect the well-being and economic growth of the state. The high number of road accidents and the inability to recover SPER debts affect not only the state's revenue, but also the safety of its citizens. In 2021, several datasets were analyzed to investigate the factors affecting road safety and SPER debt collection in Queensland. The availability of these datasets provides a unique opportunity to identify the factors that affect road safety and SPER debt recovery and to develop strategies to mitigate them.

Aim and Motivation:

This project aims to investigate the factors affecting road safety and SPER debt collection in Queensland in 2021. By analyzing four datasets, the project aims to identify the factors that contribute to road accidents and SPER debt accumulation, specifically investigating the relationship between road accidents and factors such as speed, drunk driving, fatigue, and defective vehicles, as well as the relationship between SPER debt and traffic offenses such as tolling, speeding, driving, and parking. The project aims to improve road safety and SPER debt collection in Queensland by providing evidence-based insights to policymakers, law enforcement agencies, road safety advocates, insurance companies, and the public. The findings of this project will help develop effective interventions, targeted enforcement strategies, and evidence-based campaigns to promote road safety and reduce SPER debt accumulation.

Results:

Data analysis indicated that Brisbane City, Gold Coast City, Logan City, Moreton Bay Region, and Sunshine Coast Region had the highest occurrence of car crashes in 2021. Key factors, such as driver speed, road conditions, drinking while driving, driver speed, fatigue, and defective vehicles, played a significant role in predicting car crashes and associated casualties. Moreover, speeding and tolling offenses showed a notable positive correlation with outstanding balances and number of debts. Furthermore, the presence of speed cameras does not consistently lead to a reduction in crash accidents in certain regions. Despite the higher camera installation in the Brisbane and Gold Coast, accidents still occurred without camera coverage. Conversely, areas with fewer cameras, such as Moreton Bay and the Sunshine Coast, also experienced accidents despite the presence of cameras. This suggests the possibility that other factors contribute to these accidents. Thus, installing speed cameras alone may not provide a comprehensive solution for reducing accidents, indicating the need to consider additional factors.

1. Problem Solving with Data

1. 1 Introduction

Road safety is a crucial aspect of modern society, with millions of lives lost or affected by road accidents globally every year. In Australia, road accidents remain a major cause of fatalities and injuries, with Queensland experiencing a significant number of incidents annually. Additionally, unpaid traffic fines and SPER debts have become an increasing concern for the state, and the inability to recover these debts affects the state's revenue and road safety. To address these issues, it is essential to investigate the factors that contribute to road accidents and SPER debt accumulation in Queenslands. Understanding these factors can provide insights into effective interventions to improve road safety and SPER debt collection.

By following the data science process and applying design thinking, we aimed to develop well-targeted solutions to the problem of road safety and SPER debt collection in Queensland. Design thinking is used to formulate an authentic data science problem and identify necessary datasets while considering data privacy and sampling issues. We then explore, transform, and enrich the datasets to ensure their fitness for the intended problem. Next, we design and implement analytical methods to extract insights and/or foresight related to data science problems. Finally, we create visual representations and narratives that facilitate effective decision-making by stakeholders.

The insights gained from this project will be helpful to policymakers, traffic police, road safety advocates, insurance companies, and the public in developing evidence-based strategies to improve road safety and reduce SPER debt accumulation.

1. 2 Research Question

1. Which factors are significantly associated with car accidents in Queensland in 2021?
2. Can the number and locations of cameras be used to reduce the average speed of drivers and the number of crashes in Queensland in 2021?
3. What are the most significant factors contributing to the State Penalties Enforcement Registry (SPER) debt by the SA4 regions in Queensland for 2021?

1. 3 Key Stakeholders

In any data science project, it is essential to consider stakeholders who will benefit from the project's findings. In our projects, the key stakeholders are as follows:

1. Government: Speed camera programs are essential for road safety. The findings of this project can help governments make informed decisions on the effective implementation and management of these programs.

2. Traffic Police: Police officers are crucial in enforcing speed limits and ensuring road safety. The insights gained from this project can help develop targeted enforcement strategies.
3. Drivers: Understanding the impact of speed cameras on road safety can help drivers comply with speed limits and avoid traffic fines, thereby ultimately reducing the number of road accidents.
4. Insurance Companies: Insurance companies provide policies that protect drivers and vehicles. The findings of this project can help develop suitable insurance policies that consider the impact of speed cameras on road safety.
5. General Public: The general public needs to understand the importance of speed cameras in promoting road safety. The insights gained from this project can help educate the public about the benefits of speed camera programs and promote safe driving practices.

2. Getting the Data

2. 1 Data Sources

In this data science project, we utilized four datasets to investigate the factors affecting road safety and SPER debt collection in Queensland for 2021. The datasets were obtained from the Open Data Portal Queensland Government and contained information on road accidents, speed camera locations, factors contributing to road accidents, and outstanding SPER debts.

2.2 About The Data

The first dataset is the crash data from Queensland roads, which contains information on the location and characteristics of crashes in Queensland for all reported Road Traffic Crashes that occurred from January 1 to December 31, 2021. This dataset provides valuable information on the factors contributing to road accidents in Queensland.

The second dataset was the speed camera location dataset, which consisted of fixed speed cameras, mobile speed cameras, road safety camera trailers, point-to-point speed cameras, combined red light and speed cameras, and parked mobile speed camera sites. This dataset provides information on the location of speed cameras in Queensland, which will help us understand the impact of speed cameras on road safety.

The third dataset comprises the factors in the road crash dataset, which contains information on alcohol, speed, fatigue, and defective vehicle involvement in crashes within Queensland for all reported Road Traffic Crashes from January 1 to December 31, 2021. This dataset provides detailed information on the factors that contribute to road accidents in Queensland.

The fourth and final dataset is the State Penalties Enforcement Registry (SPER) debt by SA4 region and the top 10 offence groups in Queensland. This dataset will provide information on outstanding SPER debts, which will help us understand the relationship between SPER debt accumulation and traffic offenses such as tolling, speeding, driving, and parking.

3. Is The Data Fit for Use

3.1 Check for completeness

We ensured that all the required data were available and checked for missing values. By using Python, we confirmed that the data were complete. (See appendix 1.1)

3.2 Check for outliers

3.2.1 Outliers of Factor of Crash dataset

We Look for any extreme values in the data that may be erroneous or skew the analysis. We used statistical methods and boxplots to visualize the distribution of the data. Using Python code, we identified outliers in our dataset, specifically in the Factor of Crash dataset for the columns Count_Crashes, Count_Fatality, Count_Medically_Treated, Count_Minor, and Count_All_Casualties. These outliers were caused by the fact that the values in these columns depended on the other columns in the dataset. Despite the presence of outliers, we determined that they were acceptable and justifiable in the context of our study. Therefore, we decided to retain them in the dataset. (See appendix 1.2)

Loc_Police_Region	Crash_Severity	Involving_Drink_Driving	Involving_Driver_Speed	Involving_Fatigued_Driver	Involving_Defective_Vehicle	Count_Crashes	Count_Fatality	Count_Hospitalised	Count_Medically_Treated	Count_Minor_Injury	Count_All_Casualties
Brisbane	Fatal	No	No	No	No	18	18	8	4	0	30
Brisbane	Fatal	No	No	Yes	No	2	2	0	1	0	3
Brisbane	Fatal	No	Yes	No	No	7	7	0	0	0	7
Brisbane	Fatal	Yes	No	No	No	4	6	0	3	0	9
Brisbane	Fatal	Yes	Yes	No	No	3	3	1	0	0	4
Brisbane	Hospitalisation	No	No	No	No	1217	0	1385	86	97	1568
Brisbane	**Hospitalisation**	**No**	**No**	**No**	**Yes**	**14**	0	16	1	1	18
Brisbane	**Hospitalisation**	**No**	**No**	**Yes**	**No**	**27**	0	37	3	1	41
Brisbane	Hospitalisation	No	Yes	No	No	32	0	47	2	6	55
Brisbane	Hospitalisation	No	Yes	Yes	No	1	0	1	0	0	1
Brisbane	Hospitalisation	Yes	No	No	No	109	0	127	12	7	146
Brisbane	Hospitalisation	Yes	No	No	Yes	2	0	2	0	0	2
Brisbane	Hospitalisation	Yes	No	Yes	No	2	0	3	0	0	3
Brisbane	Hospitalisation	Yes	Yes	No	No	10	0	15	0	0	15
Brisbane	Medical treatment	No	No	No	No	1525	0	0	1737	176	1913
Brisbane	Medical treatment	No	No	No	Yes	8	0	0	8	1	9
Brisbane	Medical treatment	No	No	Yes	No	13	0	0	15	0	15
Brisbane	Medical treatment	No	Yes	No	No	15	0	0	16	1	17
Brisbane	Medical treatment	Yes	No	No	No	43	0	0	53	5	58
Brisbane	Medical treatment	Yes	Yes	No	No	2	0	0	3	0	3
Brisbane	Minor injury	No	No	No	No	642	0	0	0	738	738
Brisbane	Minor injury	No	No	No	Yes	8	0	0	0	11	11
Brisbane	Minor injury	No	No	Yes	No	6	0	0	0	6	6
Brisbane	Minor injury	No	Yes	No	No	8	0	0	0	9	9
Brisbane	Minor injury	Yes	No	No	No	14	0	0	0	15	15
Brisbane	Minor injury	Yes	Yes	No	No	1	0	0	0	1	1

To read the count column data, we read it horizontally from left to right, rather than vertically. For example, we can see that there were 14 crashes in Brisbane with a Hospitalized crash severity that were caused by a Defective Vehicle, while there were 27 crashes that were caused by an Involving Fatigued Driver

3.2.2 Outliers of SPER dataset

We identified outliers in the SPER dataset for the column numbers of debt and standing balances. These outliers were caused by the fact that the values in these columns depended on the other columns in the dataset. Despite the presence of outliers, we determined that they were acceptable and justifiable in the context of our study. Therefore, we decided to retain them in the dataset. (See appendix 1.2)

SA4 Mon	Loc_ABS_Statistical_Area	Top 10 Offence Group by Outstanding Balanc	Number of Deb	Outstanding Balanc
January	Brisbane - East	Speeding	31.087	$9.632.000
January	Brisbane - East	Tolling	52.552	$10.447.000
January	Brisbane - East	Driving	15.017	$4.706.000
January	Brisbane - East	Vehicle	11.782	$4.075.000
January	Brisbane - East	Other	565	$3.127.000
January	Brisbane - East	Offender Debt Recovery	184	$2.011.000

To read the count column data, we read it horizontally from left to right, rather than vertically. For example, we can see that in January, in the Brisbane East area, there were 11,782 debtors who were categorized as having committed a vehicle offence, while there were only 565 debtors categorized as having committed other types of offences in that time and area.

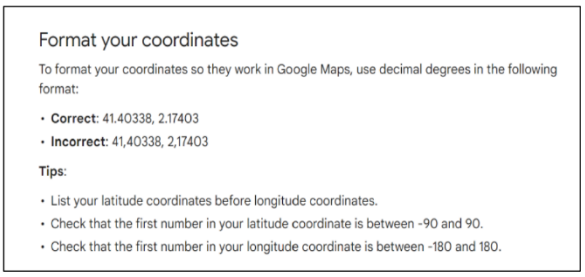
3.3 Check for Accuracy

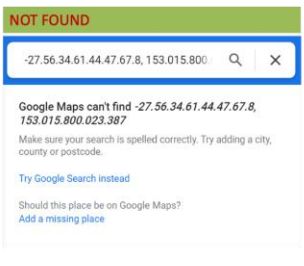
3.3.1 Accuracy of the crash location

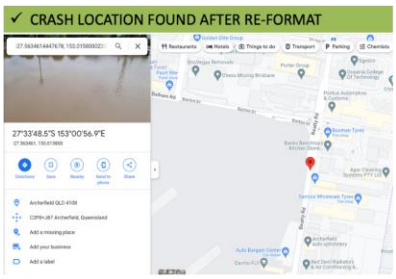
The crash locations of crash datasets do not show a standard format. If we check it on Google Maps, the location is not found. So, we re-format it using Decimal Degrees Format and store it into one column

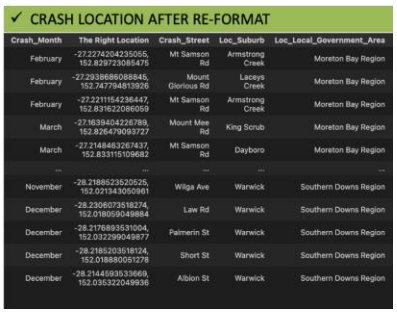
BEFORE RE-FORMAT

Crash_Longitude	Crash_Latitude
153.015.800.023.387	-27.56.34.61.44.47.67.8
153.051.538.014.882	-27.60.02.34.24.86.46.8
153.029.773.018.299	-27.59.17.90.44.63.94
153.054.874.013.435	-27.60.88.17.44.90.49.8
153.081.437.011.786	-27.59.92.02.45.16.44.9







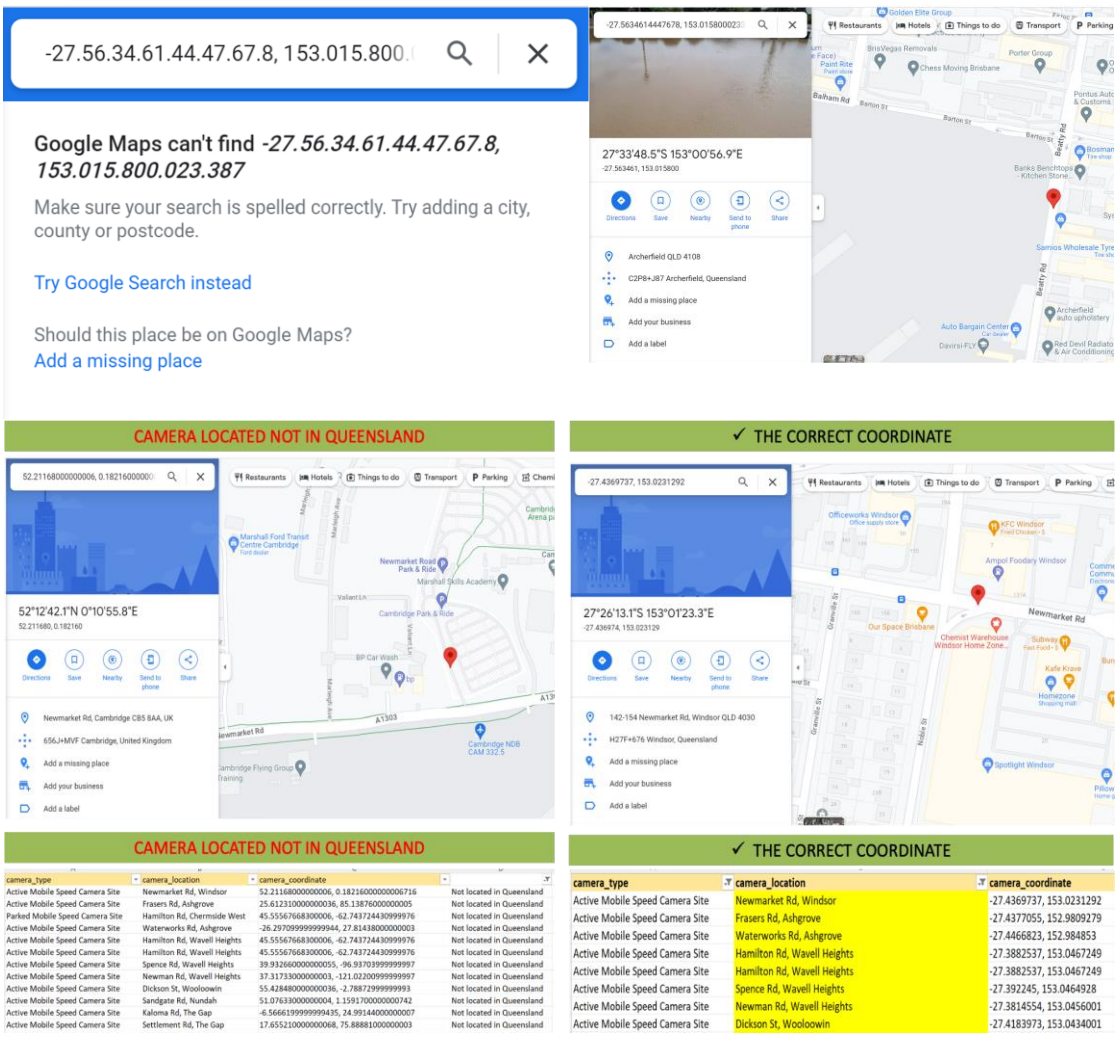


3.3.2 Consistency of the camera location

We discovered that no coordinate locations were available for the Active Mobile Speed Camera and Road Safety Camera Trailers. However, using a Python query, we were able to determine the coordinates.(see Appendices 1.3 and 1.4)

3.3.3 Accuracy of the camera location

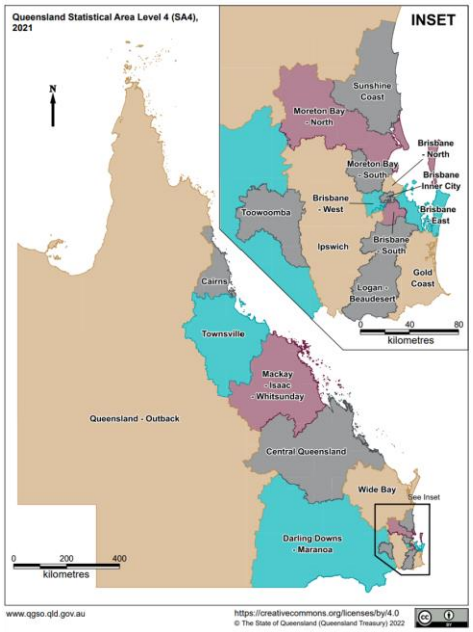
We randomly checked the locations of some of the cameras and found that they were not located in the Queensland. Some of these are located in other states or countries. After discovering this, we used a Python query to ensure that all coordinates were located within Queensland. (see Appendices 1.5, 1.6, and 1.7)

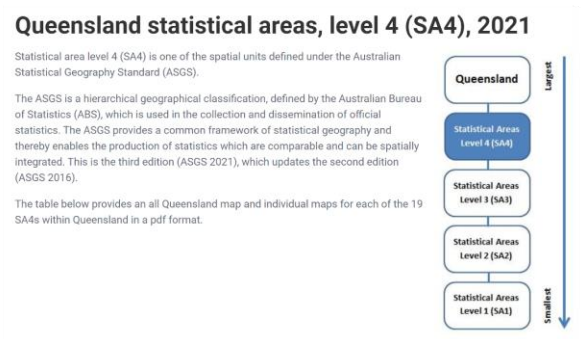


After identifying the incorrect coordinates, we used Python code to find the correct coordinates for each location, ensuring that they were all located within Queensland (see Appendix 1.7)

3.3.4 Accuracy and Consistency of The Number of Local ABS (Australian Bureau of Statistics) Statistical Area 4

ABS SA4 consists of 19 areas based on the Australian Statistical Geography Standard (ASGS), ABS SA4 consists of 19 areas. In the crash dataset, we have a column for Local ABS Statistical Area 4; therefore, in the SPER dataset, we verified that the ABS SA4 column only includes these 19 areas with regard to the Outstanding Balance. We used Python code to perform this verification. (see Appendix 1.8)





✓ SA4 NAME CONSISTS OF 19 AREAS

	SA4 Name	Top 10 Offence Group by Outstanding Balance	Number of Debts	Outstanding Balance
0	Brisbane - East	Tolling	49,528	$9,817,000
1	Brisbane - East	Speeding	28,461	$8,454,000
2	Brisbane - East	Driving	14,003	$4,376,000
3	Brisbane - East	Vehicle	10,970	$3,786,000
4	Brisbane - East	Other	5,518	$2,826,000
...	...	...	...	...
2175	Wide Bay	Public Order Offences	12,613	$4,441,000
2176	Wide Bay	Tolling	16,991	$3,318,000
2177	Wide Bay	Drug	8,504	$3,020,000
2178	Wide Bay	Stealing	6,318	$2,503,000
2179	Wide Bay	Fraud, Deception and Extortion Related Offences	1,423	$1,608,000

2160 rows × 4 columns

3.3.5 Year Consistency

Because the SPER Dataset only contains data from 2021, we wanted to ensure that the Crash Data Set and the Factor of Crash Data Set also only included data from 2021 for balance and consistency. We used Python code to filter the data sets accordingly. (see Appendix 1.9)

✓ FACTOR OF CRASH (2021 ONLY)

	Crash_Year	Crash_Police_Region	Crash_Severity	Involving_Drink_Driving	Involving_Driver_Speed	Involving_Fatigued_Driver
4613	2021	Brisbane	Fatal	No	No	No
4614	2021	Brisbane	Fatal	No	No	Yes
4615	2021	Brisbane	Fatal	No	Yes	No
4616	2021	Brisbane	Fatal	Yes	No	No
4617	2021	Brisbane	Fatal	Yes	Yes	No
...	...	...	...	...	...	...
4794	2021	Southern	Minor injury	No	No	No
4795	2021	Southern	Minor injury	No	No	Yes
4796	2021	Southern	Minor injury	No	Yes	No
4797	2021	Southern	Minor injury	No	Yes	Yes
4798	2021	Southern	Minor injury	Yes	No	No

186 rows × 13 columns

3.4 Data Transformation

3.4.1 Factor of Crash

We converted four columns of Factor of Crash Data from string type to numerical type, replacing the values of 'yes' and 'no' with 1 and 0, respectively. This conversion will make it more convenient for us to perform data analysis on these columns. (See Appendix 1.11)

Data in string

Involving_Drink_Driving	Involving_Driver_Speed	Involving_Fatigued_Driver	Involving_Defective_Vehicle
No	No	No	No
No	No	Yes	No
No	Yes	No	No
Yes	No	No	No
Yes	Yes	No	No
No	No	No	No
No	No	No	Yes
No	No	Yes	No
No	Yes	No	No
No	Yes	Yes	No
Yes	No	No	No
Yes	No	No	Yes
Yes	No	Yes	No
Yes	Yes	No	No
No	No	No	No
No	No	No	Yes
No	No	Yes	No
No	Yes	No	No
Yes	No	No	No
Yes	Yes	No	No
No	No	No	No
No	No	No	Yes
No	No	Yes	No
No	Yes	No	No
Yes	No	No	No
Yes	Yes	No	No
No	No	No	No
No	No	Yes	No
No	Yes	No	No
No	Yes	No	Yes
Yes	No	No	No
Yes	No	No	Yes
Yes	Yes	No	No

✓ Data in numeric

Involving_Drink_Driving	Involving_Driver_Speed	Involving_Fatigued_Driver	Involving_Defective_Vehicle
0	0	0	0
0	0	1	0
0	1	0	0
1	0	0	0
1	1	0	0
0	0	0	0
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	0	1
1	0	1	0
1	1	0	0
0	0	0	0
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0
1	1	0	0
0	0	0	0
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0
1	1	0	0
0	0	0	0
0	0	1	0
0	1	0	0
1	0	0	0
1	1	0	0
0	0	0	0
0	0	1	0
0	1	0	1
1	0	0	0

3.4.2 Outsanding Balance of SPER

We found that the outstanding balance in the SPER dataset was in the string format, which prevented us from performing calculations on it. To address this issue, we converted the column into a numerical format.

BEFORE CONVERTED

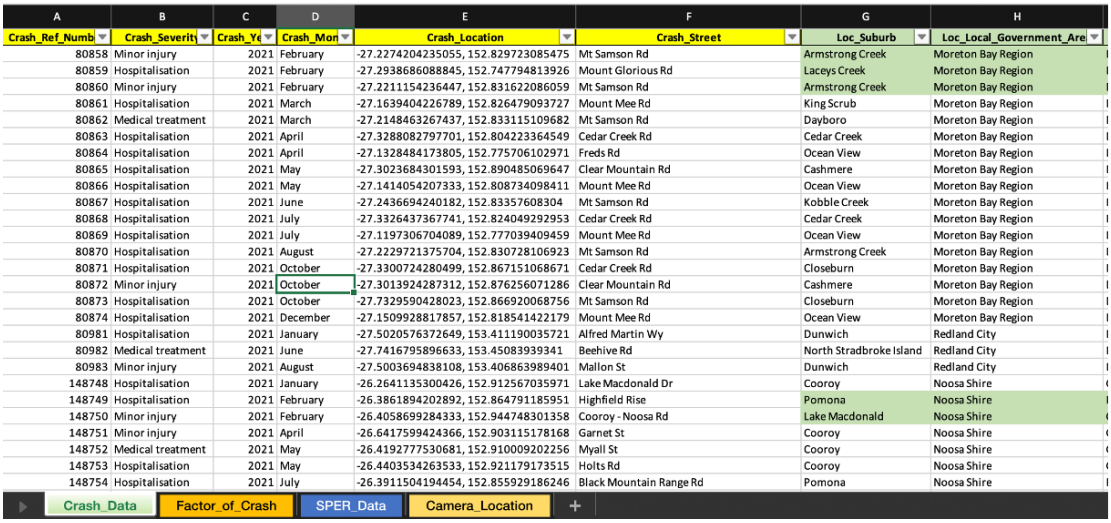
Top 10 Offence Group by Outstanding Balance	Number of Debts	Outstanding Balance
Tolling	49,528	$9,817,000
Speeding	28,461	$8,454,000
Driving	14,003	$4,376,000
Vehicle	10,970	$3,786,000
Other	5,518	$2,826,000
Fraud, Deception and Extortion Related Offences	778	$1,989,000
Fare Evasion	7,064	$1,788,000
Offender Debt Recovery	178	$1,767,000
Public Order Offences	4,273	$1,535,000
Parking	11,035	$1,496,000
Speeding	30,482	$8,925,000

✓ AFTER CONVERTED

Top 10 Offence Group by Outstanding Balance	Number of Debts	Outstanding Balance
Tolling	49,528	$9.817.000
Speeding	28,461	$8.454.000
Driving	14,003	$4.376.000
Vehicle	10,97	$3.786.000
Other	5,518	$2.826.000
Fraud, Deception and Extortion Related Offences	778	$1.989.000
Fare Evasion	7,064	$1.788.000
Offender Debt Recovery	178	$1.767.000
Public Order Offences	4,273	$1.535.000
Parking	11,035	$1.496.000
Speeding	30,482	$8.925.000

3.5 Data Integration

To make our data analysis tasks more efficient, effective, and insightful, we used Python code to integrate multiple datasets into one data file. (see Appendix 1.10)



A	B	C	D	E	F	G	H
Crash_Ref_Numb	Crash_Severity	Crash_Ye	Crash_Mon	Crash_Location	Crash_Street	Loc_Suburb	Loc_Local_Government_Are
80858	Minor injury	2021	February	-27.2274204235055, 152.829723085475	Mt Samson Rd	Armstrong Creek	Moreton Bay Region
80859	Hospitalisation	2021	February	-27.2938686088845, 152.747794813926	Mount Glorious Rd	Laceys Creek	Moreton Bay Region
80860	Minor injury	2021	February	-27.2211154236447, 152.831622086059	Mt Samson Rd	Armstrong Creek	Moreton Bay Region
80861	Hospitalisation	2021	March	-27.1639404226789, 152.826479093727	Mount Mee Rd	King Scrub	Moreton Bay Region
80862	Medical treatment	2021	March	-27.2148463267437, 152.833115109682	Mt Samson Rd	Dayboro	Moreton Bay Region
80863	Hospitalisation	2021	April	-27.3288082797701, 152.804223364549	Cedar Creek Rd	Cedar Creek	Moreton Bay Region
80864	Hospitalisation	2021	April	-27.1328484173805, 152.775706102971	Freds Rd	Ocean View	Moreton Bay Region
80865	Hospitalisation	2021	May	-27.3023684301593, 152.890485069647	Clear Mountain Rd	Cashmere	Moreton Bay Region
80866	Hospitalisation	2021	May	-27.1414054207333, 152.808734098411	Mount Mee Rd	Ocean View	Moreton Bay Region
80867	Hospitalisation	2021	June	-27.2436694240182, 152.83357608304	Mt Samson Rd	Kobble Creek	Moreton Bay Region
80868	Hospitalisation	2021	July	-27.3326437367741, 152.824049292953	Cedar Creek Rd	Cedar Creek	Moreton Bay Region
80869	Hospitalisation	2021	July	-27.1197306704089, 152.777039409459	Mount Mee Rd	Ocean View	Moreton Bay Region
80870	Hospitalisation	2021	August	-27.2229721375704, 152.830728106923	Mt Samson Rd	Armstrong Creek	Moreton Bay Region
80871	Hospitalisation	2021	October	-27.3300724280499, 152.867151068671	Cedar Creek Rd	Closeburn	Moreton Bay Region
80872	Minor injury	2021	October	-27.3013924287312, 152.876256071286	Clear Mountain Rd	Cashmere	Moreton Bay Region
80873	Hospitalisation	2021	October	-27.7329590428023, 152.866920068756	Mt Samson Rd	Closeburn	Moreton Bay Region
80874	Hospitalisation	2021	December	-27.1509928817857, 152.818541422179	Mount Mee Rd	Ocean View	Moreton Bay Region
80981	Hospitalisation	2021	January	-27.5020576372649, 153.411190035721	Alfred Martin Wy	Dunwich	Redland City
80982	Medical treatment	2021	June	-27.7416795896633, 153.45083939341	Beehive Rd	North Stradbroke Island	Redland City
80983	Minor injury	2021	August	-27.5003694838108, 153.406863989401	Mallon St	Dunwich	Redland City
148748	Hospitalisation	2021	January	-26.2641135300426, 152.912567035971	Lake Macdonald Dr	Cooroy	Noosa Shire
148749	Hospitalisation	2021	February	-26.3861894202892, 152.864791185951	Highfield Rise	Pomona	Noosa Shire
148750	Minor injury	2021	February	-26.4058699284333, 152.944748301358	Cooroy - Noosa Rd	Lake Macdonald	Noosa Shire
148751	Minor injury	2021	April	-26.6417599424366, 152.903115178168	Garnet St	Cooroy	Noosa Shire
148752	Medical treatment	2021	May	-26.4192777530681, 152.910009202256	Myall St	Cooroy	Noosa Shire
148753	Hospitalisation	2021	May	-26.4403534263533, 152.921179173515	Holts Rd	Cooroy	Noosa Shire
148754	Hospitalisation	2021	July	-26.3911504194454, 152.855929186246	Black Mountain Range Rd	Pomona	Noosa Shire

Crash_Data | Factor_of_Crash | SPER_Data | Camera_Location

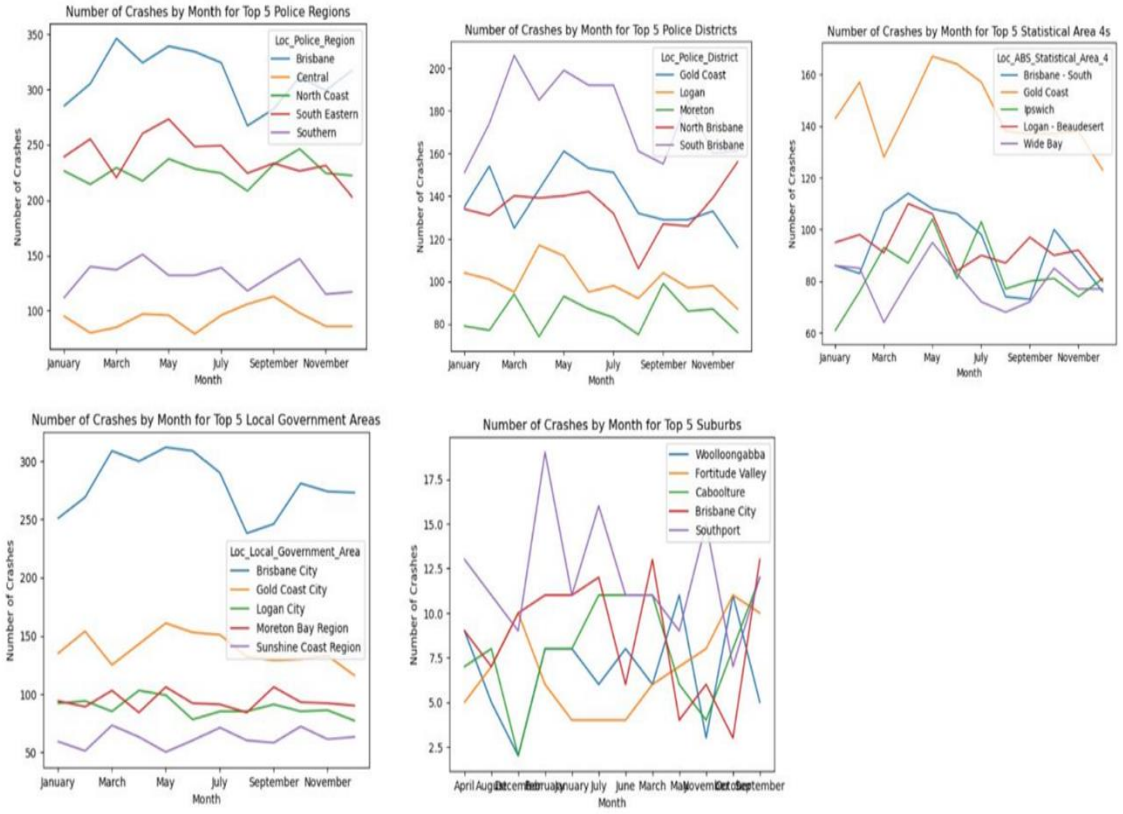
4. Making the data confess

The fourth step of the data science process involves making the data confess. During this step, we analyzed and visualized the data to uncover hidden insights and patterns. By doing so, we aim to extract meaningful information that can help us answer our research questions or make predictions.

4.1 Exploratory Data Analysis

4.1.1 Exploration Top 5 Crash Location in The Queensland

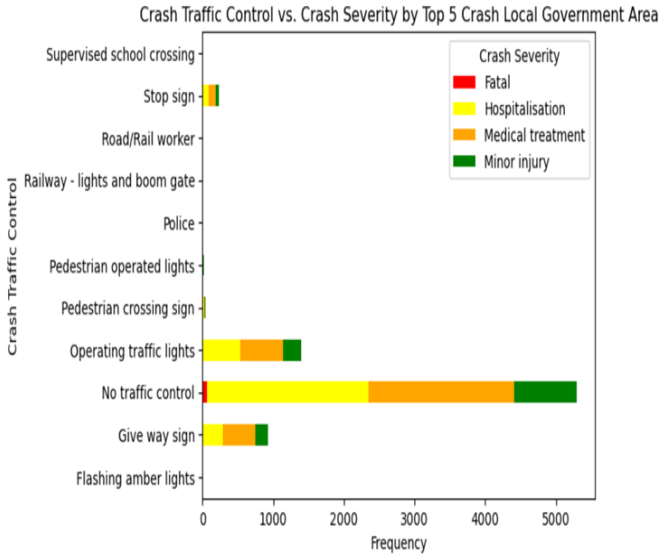
We are attempting to identify patterns in the occurrence of crashes throughout the year in various locations, such as the police, police, SA4, local governments, and suburbs. (Code in Appendix 2.1)

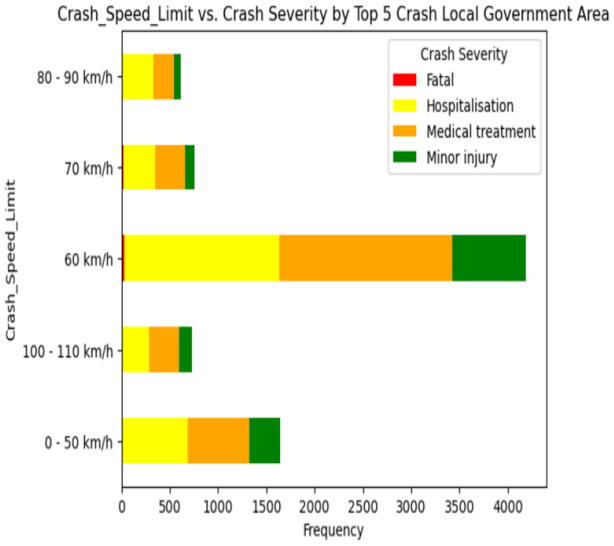


The number of crashes across all areas fluctuated significantly throughout 2021. When considering Local Government Areas, Brisbane City, Gold Coast City, Logan City, Moreton Bay Region, and Sunshine Coast Region were the top five areas with the highest number of crashes throughout the year. Additionally, the Brisbane Police Region had the highest crash rate in 2021, followed by the North Brisbane Police District. In terms of the Local ABS SA4 areas, the Gold Coast had the highest number of crashes throughout the year 2021. In conclusion, further investigation is required to identify the factors contributing to the high number of crashes in the top five cities in 2021.

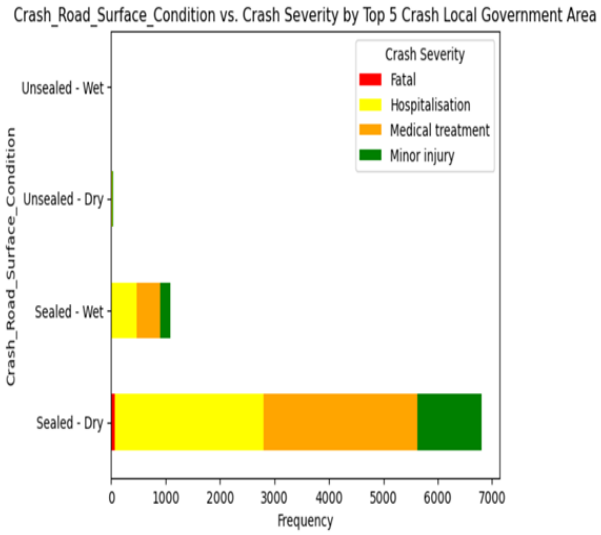
4.1.2 factors that contributed to the severity of crashes in the top five Local Government Areas.

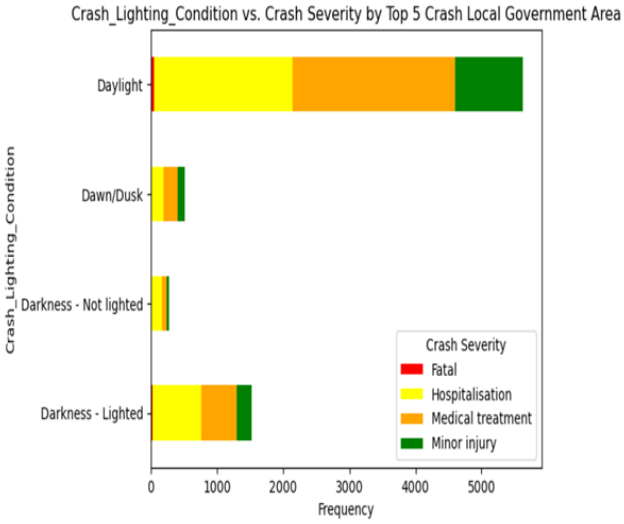
Based on previous findings, we know that the top five areas with the highest number of crashes throughout the year are the Local Government Areas of Brisbane City, Gold Coast City, Logan City, Moreton Bay Region, and Sunshine Coast Region. Therefore, further investigations should be conducted to identify the factors that contribute to crashes in these areas. (See the Code in Appendix 2.2).

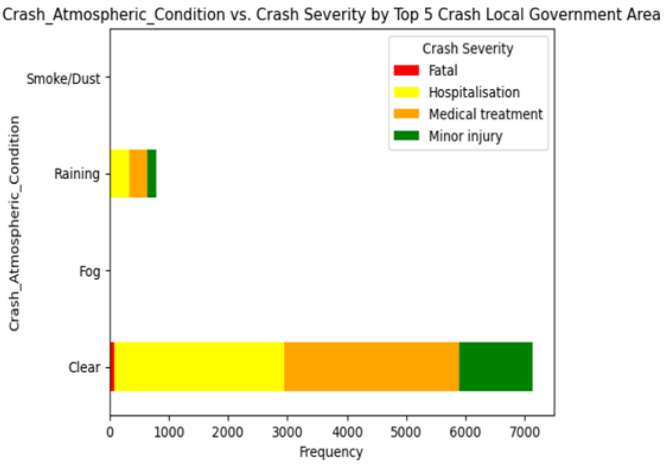




Based on the diagram, the majority of crash severity incidents were classified as hospitalizations and medical treatments. The three primary causes of crashes were the absence of traffic control, which accounted for 5000 cases, followed by operating traffic lights, which were responsible for approximately 1500 cases, and giving way signs, which caused 1000 cases. Interestingly, when examining the driver speed, the highest number of crashes occurred at 60 km/h, rather than at speeds of 100-110 km per hour.



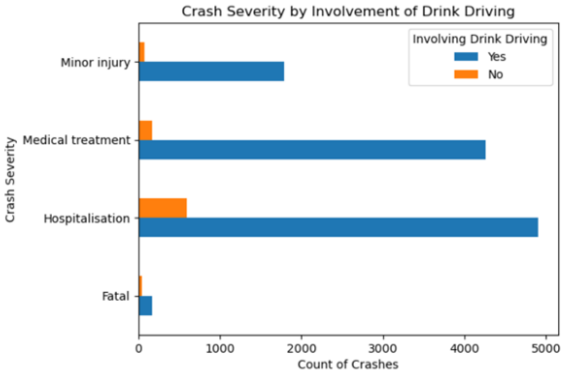


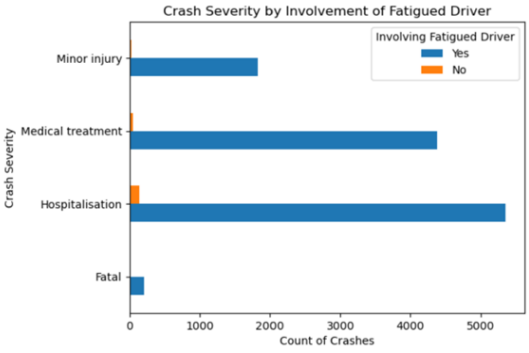


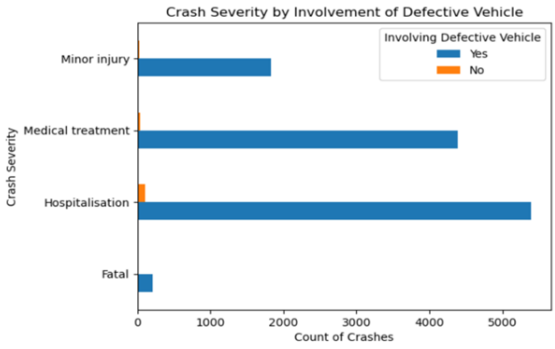
Based on the stacked bar graph, the most frequent crash severity levels were hospitalization and medical treatment. The highest number of crashes occurred under the road surface conditions of sealed-dry and sealed-wet, with 6900 and 1000 cases, respectively. Interestingly, these crashes occurred under clear atmospheric and daylight lighting conditions.

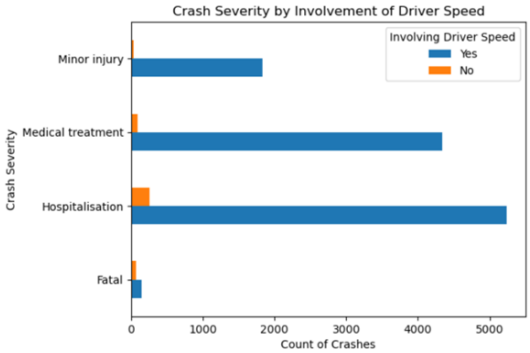
4.1.3 Involvement factors that contributed to the severity of the crashes in the top 5 Local Police Regions

Based on our previous findings, the top five Local Police Regions with the highest number of crashes throughout the year were Brisbane, Central, North Coast, South Eastern, and Southern. To identify the factors that contributed to the severity of crashes in these regions, we used the involvement factors of the crash, specifically, drink driving, driver speed, fatigued drivers, and defective vehicles. Please refer to Appendix 2.2 for the code used in this investigation.









All the graphics demonstrate that hospitalization is the most prevalent crash severity resulting from the involvement of drinking while driving, driver's speed, fatigued driving, and defective vehicle.

4.1.4 Association strength between factors of crash to the crash severity

To determine the influence of each factor on crash severity, we used Cramer's V, a measure of the association strength between nominal variables. Cramer's V is a measure of the strength of the association between two nominal variables, and its value ranges from zero to one. (Code in Appendix 2.3)

(IN ALL AREAS)

```
Cramer's V between Crash_Traffic_Control and Crash_Severity: 0.040
Cramer's V between Crash_Speed_Limit and Crash_Severity: 0.061
Cramer's V between Crash_Road_Surface_Condition and Crash_Severity: 0.025
Cramer's V between Crash_Atmospheric_Condition and Crash_Severity: 0.014
Cramer's V between Crash_Lighting_Condition and Crash_Severity: 0.050
```

(CRASHES IN THE TOP 5 LOCAL GOVERNMENT AREAS)

```
Cramer's V between Crash_Traffic_Control and Crash_Severity: 0.044
Cramer's V between Crash_Speed_Limit and Crash_Severity: 0.036
Cramer's V between Crash_Road_Surface_Condition and Crash_Severity: 0.019
Cramer's V between Crash_Atmospheric_Condition and Crash_Severity: 0.013
Cramer's V between Crash_Lighting_Condition and Crash_Severity: 0.045
```

Cramer's V values between the predictor variables (Crash_Traffic_Control,Crash_Speed_Limit,Crash_Road_Surface_Condition,Crash_Atmospheric_Condition, and Crash_Lighting_Condition) and the response variable (Crash_Severity) indicate the strength of the association between them, with values ranging from 0 to 1. A higher value indicates a stronger association between the variables.

In all areas of crashes, Cramer's V values ranged from 0.014 to 0.061, indicating a weak to moderate association between the predictor variables and Crash_Severity. However, when we focused on the top five areas of crashes (Brisbane City, Gold Coast City, Logan City, Moreton Bay Region, and Sunshine Coast Region), the Cramer's V values ranged from 0.013 to 0.045, indicating a slightly weaker association between the predictor variables and Crash_Severity compared to all areas of crashes.

Overall, these results suggest that while there is some association between the predictor variables and Crash_Severity, the strength of these associations is not particularly strong in all areas or the top five areas of crashes.

4.1.4 Association strength between involvement factors that contributed to the severity of the crashes

To determine the influence of each factor on crash severity, we used Cramer's V, a measure of the association strength between nominal variables. Cramer's V is a measure of the strength of the association between two nominal variables, and its value ranges from zero to one. (Code in Appendix 2.4)

(IN ALL AREAS)

```
Cramer's V between Involving_Drink_Driving and Crash_Severity: 0.089
Cramer's V between Involving_Driver_Speed and Crash_Severity: 0.072
Cramer's V between Involving_Fatigued_Driver and Crash_Severity: 0.051
Cramer's V between Involving_Defective_Vehicle and Crash_Severity: 0.120
```

(CRASHES IN THE TOP 5 LOCAL GOVERNMENT AREAS)

```
Cramer's V between Involving_Drink_Driving and Crash_Severity: 0.069
Cramer's V between Involving_Driver_Speed and Crash_Severity: 0.087
Cramer's V between Involving_Fatigued_Driver and Crash_Severity: 0.087
Cramer's V between Involving_Defective_Vehicle and Crash_Severity: 0.116
```

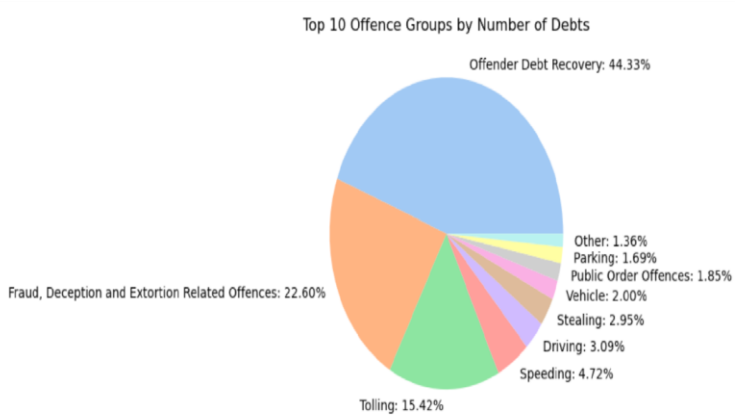
The output shows the strength of association between each predictor variable (Involving_Drink_Driving, Involving_Driver_Speed, Involving_Fatigued_Driver, Involving_Defective_Vehicle) and the response variable (Crash_Severity) using Cramer's V statistic. The strongest association was observed between Involving_Defective_Vehicle and Crash_Severity (Cramer's V = 0.12), followed by Involving_Drink_Driving and Crash_Severity (Cramer's V = 0.089). In contrast, the lowest association was observed between Involving_Fatigued_Driver and Crash_Severity (Cramer's V = 0.051).

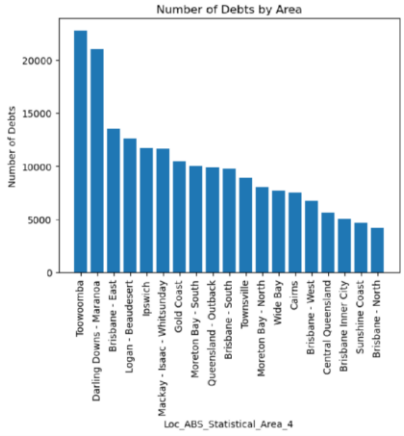
When focusing on the top five areas in the Loc_Police_Region, the results were similar, with Involving_Defective_Vehicle and Crash_Severity showing the strongest association (Cramer's V = 0.116) and Involving_Fatigued_Driver showing the weakest association (Cramer's V = 0.087). However, there is a difference in the association strength between Crash_Severity and Involving_Driver_Speed and Involving_Fatigued_Driver, where their association strength in the top five areas is stronger than in all areas of the crash.

Overall, the results suggest that Involving_Defective_Vehicle and Involving_Drink_Driving are the two most important predictor variables associated with Crash_Severity in all areas of the crash, and this finding is consistent with the top five areas in Loc_Police_Region.

4.1.5 Top 10 Offence Contributed to Number of Debtors

Nearly half of the total number of debts is caused by offender debt recovery (44.33%), followed by fraud Offence (22.60 %) and tolling (15.42 %). The regions with the highest number of debts are Toowoomba, Darling Downs, and Brisbane East.





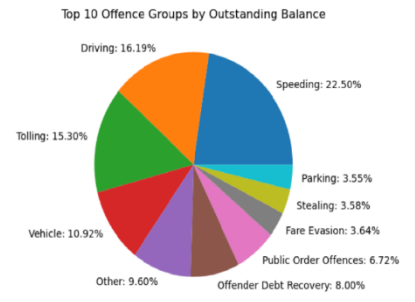
4.1.6 Proportion of Offence Contributed to Number of Debtors

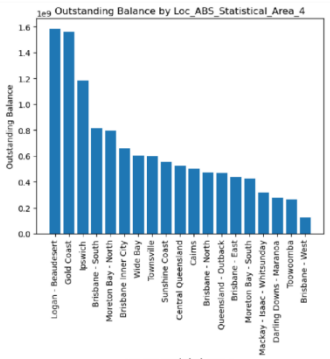
Based on the previous graph, it can be observed that the five areas with the highest number of debtors are located in Toowomba, Darlig Downs, Brisbane-East, Logan, and Ipswich. To obtain more detailed insight into the factors that contribute to the high number of debtors, a pie chart can be created to determine the proportion of each offence group that has contributed to the number of debtors.

Across all five regions, speeding consistently contributed approximately 20% to the number of debtors. Tolling is the main contributor to the number of debtors in Logan, accounting for 41%, followed by Brisbane East (37%), and Ipswich (30%).

4.1.7 Top 10 Offence Contributed to Outstanding Balance

According to the chart, Speeding, Driving, and Tolling were the top three offenses that contributed the most to the outstanding balance, accounting for 22.5%, 16.19%, and 15.3%, respectively. The trend showed that Logan having the highest outstanding balance of 1.6 billion, followed by Gold Coast with 1.5 billion, and Ipswich in third place with an outstanding balance of 1.2 billion. (See Appendix 2.6)





SUMMARY STATISTICS OF OUTSTANDING BALANCE

	Group.1	x.mean	x.median
1	Animal	837500	837500
2	Bail, Probation and Fail to appear Offences	1459667	1461500
3	Driving	8259597	7626500
4	Drug	2125652	2165000
5	Fail to Vote	298200	295000
6	Fare Evasion	3495150	3130000
7	Fraud, Deception and Extortion Related Offences	2174043	1663500
8	Offender Debt Recovery	3774671	3400500
9	Other	4835423	5188000
10	Other ²	8671000	8671000
11	Parking	3825168	3338000
12	Public Order Offences	3227097	2739000
13	Speeding	11878991	9192500
14	Stealing	2354402	2064000
15	Tolling	9399141	7500000
16	Unknown	3139600	3134000
17	Vehicle	5642213	4724500

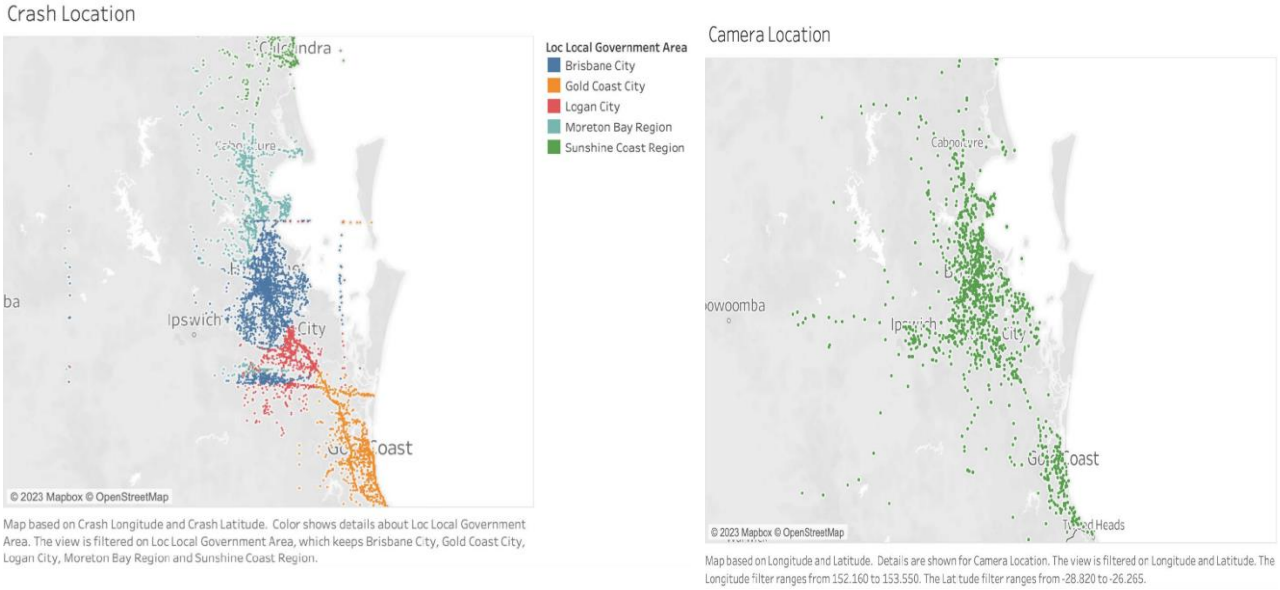
4.1.8 Proportion of Offence Contributed to Outstanding Balance

Based on the previous graph, it can be observed that the five areas with the highest number of outstanding balances are located in Logan, Gold Coast, Ipswich, Brisbane-South, and Moreton Bay North. To obtain a more detailed insight into the factors that contribute to the high number of outstanding balances, a pie chart can be created to determine the proportion of each offence group that has contributed to the number of outstanding balances. (Code in Appendix 2.7)

Almost all regions consistently showed that speeding was the leading factor contributing to outstanding balances, with an average proportion of 25%. However, Logan differs from the others in that tolling is the largest contributor. Furthermore, Tolling consistently ranks as the second-largest contributor to the accumulation of outstanding balances, accounting for almost 20% of each region.

4.1.9 Cameras and Crash Locations

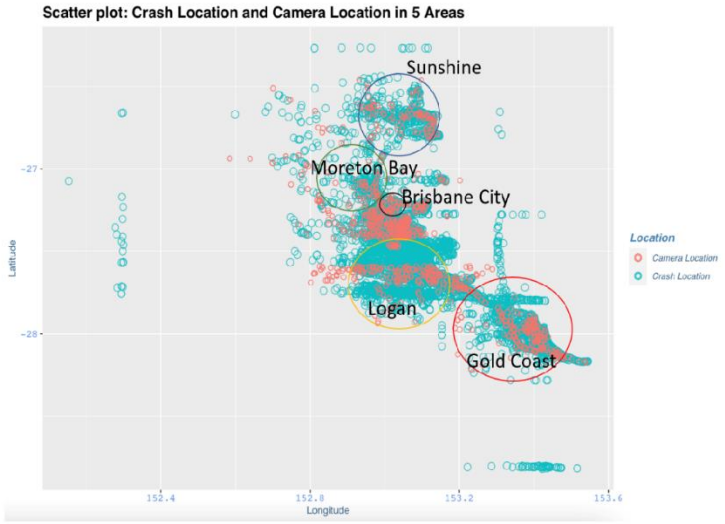
Plotting maps indicates that the number of installed speed cameras in Brisbane City is the largest and also has the highest rate of crash accidents, followed by the gold coast. Compared to other areas, although there are not many installed cameras, crash accidents rarely occur, such as in the Sunshine coast region and Moreton Bay region. Therefore, there might be other factors relevant to these accidents, as both maps indicate that the installation of speed cameras may not ensure a decrease in crash accidents.

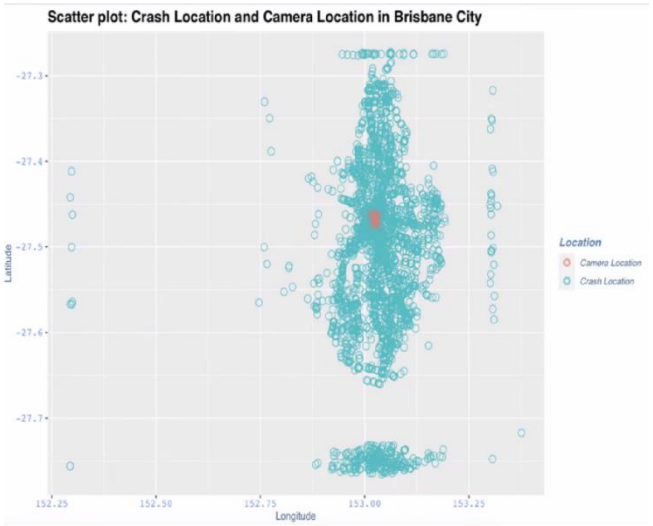


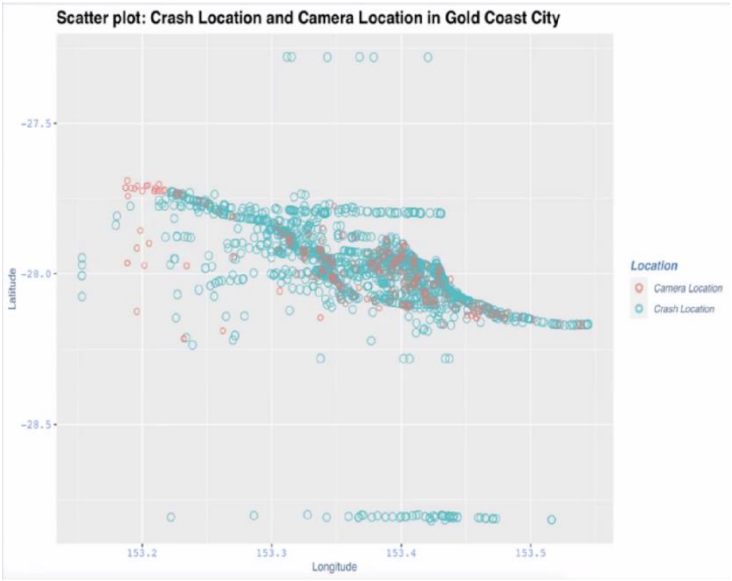
Map based on Crash Longitude and Crash Latitude. Color shows details about Loc Local Government Area. The view is filtered on Loc Local Government Area, which keeps Brisbane City, Gold Coast City, Logan City, Moreton Bay Region and Sunshine Coast Region.

Map based on Longitude and Latitude. Details are shown for Camera Location. The view is filtered on Longitude and Latitude. The Longitude filter ranges from 152.160 to 153.550. The Latitude filter ranges from -28.820 to -26.265.

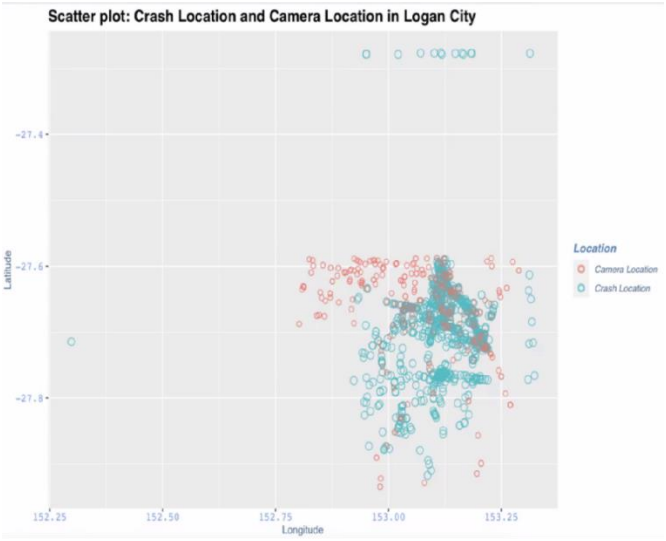
4.1.10 Cameras and Top 5 Crash Locations in Local Government Area

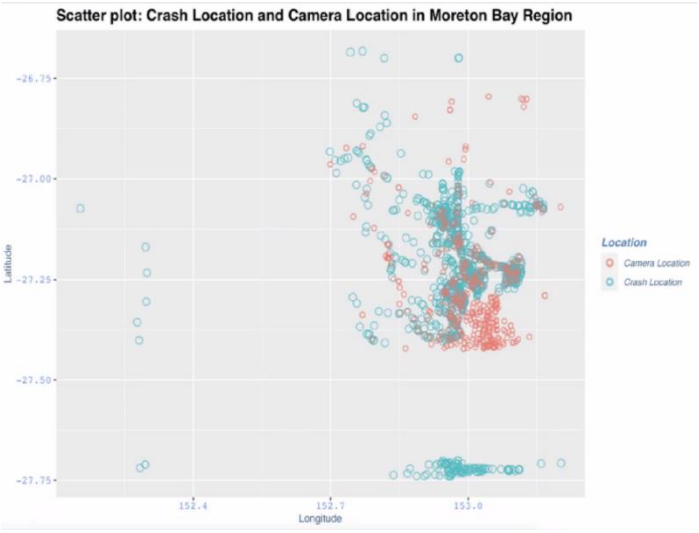


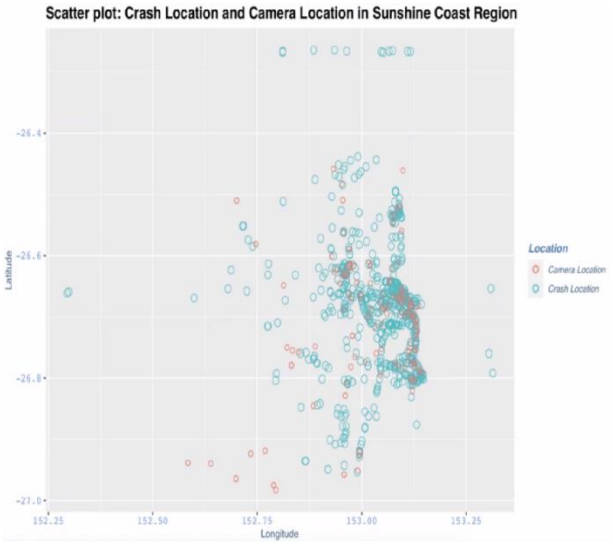




In Brisbane City, the number of cameras is insufficient compared with the number of crashes that occur. A significant number of accidents occur in areas where cameras are absent. On the Gold Coast, although some cameras are present in locations where crashes occur, incidents of crashes still occur without any camera coverage.







Logan City also shows that some cameras are present in locations where crashes occur, and that there are still incidents of crashes that occur without any camera coverage. In the Moreton Bay Region, it is apparent that cameras are present in some locations where crashes occur, yet many incidents of crashes still occur without any camera coverage. In contrast, there are areas where cameras are present, but no accidents have been recorded. In the Sunshine Coast Region, it is apparent that cameras are present in some locations where crashes occur, yet many incidents of crashes still occur without any camera coverage. In contrast, there are areas where cameras are present, but no accidents have been recorded.

4.2 Statistical Data Analysis

4.2.1 Regression Analysis to Outstanding Balance and Number of Debts

The regression model output indicates that the model has a multiple R-squared value of 0.3186, suggesting that the model explains approximately 31.86% of the variance in the outstanding balance variable. The adjusted R-squared value of 0.3135 was slightly lower than the multiple R-squared values, indicating that adding more predictors did not improve the explanatory power of the model. The F-statistic has a value of 62.63 and a p-value of less than 2.2e-16, which indicates that at least one of the predictors is significantly related to the outcome.

REGRESSION MODEL TO Y (OUTSTANDING BALANCE)

```
Call:
lm(formula = Outstanding_Balance ~ method, data = spea)

Residuals:
     Min       1Q   Median       3Q      Max
-8924991 -2051477  -493543  1180093 31549859

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                        837500    3228662
methodBail, Probation and Fail to appear Offences  622167    3487354
methodDriving                                     7422097    3243575
methodDrug                                        1288152    3257362
methodFail to Vote                                -539300    3820205
methodFare Evasion                                2657650    3255457
methodFraud, Deception and Extortion Related Offences 1336543 3251642
methodOffender Debt Recovery                      2937171    3243575
methodOther                                       3997923    3243645
methodOther ²                                     7833500    5592207
methodParking                                     2987668    3258697
methodPublic Order Offences                       2389597    3243575
methodSpeeding                                   11041491    3243575
methodStealing                                    1516902    3247711
methodTolling                                     8561641    3245435
methodUnknown                                     2302100    3820205
methodVehicle                                     4804713    3243575
                                                 t value Pr(>|t|)
(Intercept)                                        0.259 0.795355
methodBail, Probation and Fail to appear Offences  0.178 0.858421
methodDriving                                      2.288 0.022220 *
methodDrug                                         0.395 0.692544
methodFail to Vote                                -0.141 0.887749
methodFare Evasion                                 0.816 0.414381
methodFraud, Deception and Extortion Related Offences 0.411 0.681087
methodOffender Debt Recovery                       0.906 0.365284
methodOther                                        1.233 0.217882
methodOther ²                                      1.401 0.161422
methodParking                                      0.917 0.359335
methodPublic Order Offences                        0.737 0.461375
methodSpeeding                                     3.404 0.000676 ***
methodStealing                                     0.467 0.640499
methodTolling                                      2.638 0.008399 **
methodUnknown                                      0.603 0.546831
methodVehicle                                      1.481 0.138673
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4566000 on 2143 degrees of freedom
Multiple R-squared:  0.3186,    Adjusted R-squared:  0.3135
F-statistic: 62.63 on 16 and 2143 DF,  p-value: < 2.2e-16

Call:
lm(formula = Number_of_Debts ~ method, data = spea)
```

The coefficient values suggest that the “Speeding” method has a significantly positive relationship with outstanding balance, as it has the highest coefficient value and a low p-value (0.000676). This finding implies that individuals with outstanding balances resulting from speeding offenses tend to have higher outstanding balances than those with other offenses.

REGRESSION MODEL TO Y (NUMBER OF DEBTS)

```
Residuals:
   Min     1Q Median     3Q    Max
-42091  -5361   -279   2389 159530

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                        2418.5    13467.5
methodBail, Probation and Fail to appear Offences  2520.9    14546.5
methodDriving                                     22980.1    13529.7
methodDrug                                         4183.5    13587.2
methodFail to Vote                                 -747.9    15934.9
methodFare Evasion                                10921.6    13579.2
methodFraud, Deception and Extortion Related Offences -1144.1 13563.3
methodOffender Debt Recovery                       -2078.6   13529.7
methodOther                                        8130.9    13530.0
methodOther ²                                     13161.5    23326.4
methodParking                                     26850.1    13592.8
methodPublic Order Offences                        7522.9    13529.7
methodSpeeding                                    37603.9    13529.7
methodStealing                                     3416.2    13546.9
methodTolling                                     45023.4    13537.4
methodUnknown                                     -2063.3    15934.9
methodVehicle                                     14145.3    13529.7
                                                 t value Pr(>|t|)
(Intercept)                                        0.180 0.857499
methodBail, Probation and Fail to appear Offences  0.173 0.862432
methodDriving                                      1.698 0.089560 .
methodDrug                                         0.308 0.758190
methodFail to Vote                                -0.047 0.962570
methodFare Evasion                                 0.804 0.421320
methodFraud, Deception and Extortion Related Offences -0.084 0.932785
methodOffender Debt Recovery                       -0.154 0.877915
methodOther                                        0.601 0.547935
methodOther ²                                      0.564 0.572655
methodParking                                      1.975 0.048359 *
methodPublic Order Offences                        0.556 0.578250
methodSpeeding                                     2.779 0.005494 **
methodStealing                                     0.252 0.800929
methodTolling                                      3.326 0.000896 ***
methodUnknown                                     -0.129 0.896988
methodVehicle                                      1.045 0.295910
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19050 on 2143 degrees of freedom
Multiple R-squared:  0.3802,    Adjusted R-squared:  0.3755
F-statistic: 82.15 on 16 and 2143 DF,  p-value: < 2.2e-16
```

Regarding the number of debts, the model has a multiple R-squared value of 0.3802, indicating that it explains approximately 38.02% of the variance in the number of debt variables. The adjusted R-squared value of 0.3755 was slightly lower than the multiple R-squared values, indicating that adding more predictors did not improve the explanatory power of the model. The F-statistic of 82.15 with 16 and 2143 degrees of freedom and a p-value of less than 2.2e-16, indicates that at least one of the predictors is highly significant in explaining the variance in the outcome variable.

The coefficient values suggest that the " tolling method has a significantly positive relationship with the number of debts, as it has the highest coefficient value and a low p-value. This implies that individuals with outstanding balances resulting from tolling offenses tend to have more debts than those with other offenses

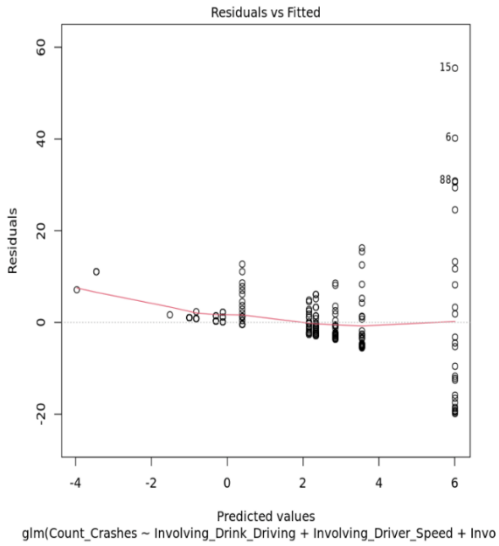
4.2.2 Poisson Regression for Analyzing Involvement Factors to Count of Crashes

In our factor of the crash dataset, both the "Crash Count" and "Count all casualties" variables involve counting and accumulation of data, making them more suitable for modeling using count data analysis methods such as Poisson regression. We create 5 models:

No	Y	X	$\hat{y}$	Significance at $\alpha = 0.05$	Residual Deviance	AIC	Overall Interpretation
1	Count_Crashes	X_1: Involving_Drink Driving	4.67-2.0$_{X1}$	Yes	42047 on 184 degrees of freedom	42804	the model suggests that Involving Drinking Driving is a significant predictor of the count of crashes, and that the model explains some but not all of the variability in the count of crashes
2	Count_Crashes	X_1: Involving_Driver_Speed	4.65-2.5 $_{X1}$	Yes	42047 on 184 degrees of freedom	42804	the model suggests that Involving Driver Speed is a significant predictor of the count of crashes, and that the model explains some but not all of the variability in the count of crashes
3	Count_Crashes	X_1: Involving_Fatigued_Driver	4.49-2.5 $_{X1}$	Yes	44153 on 184 degrees of freedom	44911	the model suggests that Involving Fatigued Driver is a significant predictor of the count of crashes, and that the model explains some but not all of the variability in the count of crashes
4	Count_Crashes	X_1: Involving_Defective_Vehicle	4.50-2.8 $_{X1}$	Yes	43770 on 184 degrees of freedom	44527	the model suggests that Involving Defective Vehicle is a significant predictor of the count of crashes, and that the model explains some but not all of the variability in the count of crashes
5	Count_Crashes	X_1: Involving_Drink Driving X_2: Involving_Driver_Speed X_3: Involving_Fatigued_Driver X_4: Involving_Defective_Vehicle	6.00-2.4 $_{X1}$-3.1 $_{X2}$-3.6 $_{X3}$-3.8 $_{X4}$	Yes	14214 on 184 degrees of freedom	14977	the model suggests that the linear relationship between the combination of four variables and the occurrence of car accidents. When the occurrences of the four situations are low, the probability of a car accident is also low. However, as these situations increase, the probability of a car accident also increases. There is a linear relationship between these situations and the number of car accidents, as shown by the red line in the graph

The results of the Poisson regression analysis suggest that drinking driving, evolving driver speed, evolving fatigue driver, and evolving defective vehicle are all significant predictors of the crash count. These variables can explain some, but not all, of the variability in the count of crashes. Additionally, a linear relationship appears to exist between the combination of these four variables and the occurrence of car accidents. As the occurrence of such situations increases, the probability of car accidents also increases. The red line in the graph illustrates the linear relationship between these situations and number of car accidents. Overall, these findings provide important insights into the factors that contribute to car accidents, and can inform strategies for reducing the occurrence of car accidents on the road.

```
Call:
glm(formula = Count_Crashes ~ Involving_Drink_Driving + Involving_Driver_Speed +
    Involving_Fatigued_Driver + Involving_Defective_Vehicle,
    family = poisson(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-27.398   -3.268   -0.223    1.692   42.366

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                6.007916   0.009352  642.42   <2e-16 ***
Involving_Drink_Driving   -2.455051   0.032500  -75.54   <2e-16 ***
Involving_Driver_Speed    -3.155913   0.045214  -69.80   <2e-16 ***
Involving_Fatigued_Driver -3.670080   0.060664  -60.50   <2e-16 ***
Involving_Defective_Vehicle -3.850533   0.070092  -54.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 48637  on 185  degrees of freedom
Residual deviance: 14214  on 181  degrees of freedom
AIC: 14977

Number of Fisher Scoring iterations: 5
```



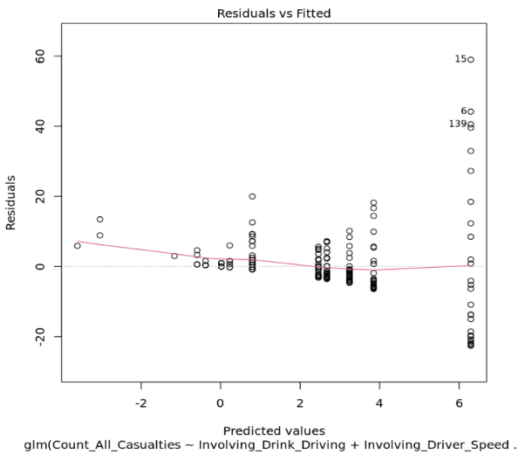
4.2.3 Poisson Regression for Analyzing Involvement Factors to Count of Casualties

Poisson regression is primarily used for modeling count data. In our factor of the crash dataset, both the "Crash Count" and "Count all casualties" variables involve counting and accumulation of data, making them more suitable for modeling using count data analysis methods such as Poisson regression. We create 5 models:

No	Y	X	$\hat{y}$	Significance at $\alpha = 0.05$	Residual Deviance	AIC	Overall Interpretation
1	Count all casualties	X_1: Involving_Drink Driving	4.9-1.9 x_1	Yes	55241 on 184 degrees of freedom	56053	the model suggests that Involving Drinking Driving is a significant predictor of the count all casualties, and that the model explains some but not all of the variability in the count of crashes
2	Count all casualties	X_1: Involving_Driver_Speed	4.9-2.4 x_1	Yes	54195 on 184 degrees of freedom	55008	the model suggests that Involving Driver Speed is a significant predictor of the count all casualties, and that the model explains some but not all of the variability in the count of crashes
3	Count all casualties	X_1: Involving_Fatigued_Driver	4.7-2.5 x_1	Yes	58089 on 184 degrees of freedom	58902	the model suggests that Involving Fatigued Driver is a significant predictor of the count all casualties, and that the model explains some but not all of the variability in the count of crashes
4	Count all casualties	X_1: Involving_Defective_Vehicle	4.7-2.8 x_1	Yes	57513 on 184 degrees of freedom	58326	the model suggests that Involving Defective Vehicle is a significant predictor of the count all casualties, and that the model explains some but not all of the variability in the count of crashes
5	Count all casualties	X_1: Involving_Drink Driving X_2: Involving_Driver_Speed X_3: Involving_Fatigued_Driver X_4: Involving_Defective_Vehicle	6.2-2.4 x_1-3.0 x_2-3.6 x_3-3.8 x_4	Yes	18581 on 184 degrees of freedom	19399	the model suggests that the linear relationship between the combination of four variables and the occurrence of casualties. When the occurrences of the four situations are low, the probability of casualties is also low. However, as these situations increase, the probability of casualties also increases. There is a linear relationship between these situations and the casualties, as shown by the red line in the graph

Based on the fifth models, the last models give the lowest AIC.The results of the model indicate that involving driving while under the influence of alcohol, speeding, driving while fatigued, and driving a defective vehicle are all significant predictors of the count of all casualties in car accidents. However, the model does not explain all the variability in the number of casualties. Additionally, the model suggests a linear relationship between the combination of these four variables and the occurrence of casualties. As the occurrence of these situations increases, the probability of casualties also increases; this relationship is indicated by the red line in the graph. Overall, these findings can help inform efforts to reduce the occurrence of car accidents and the associated casualties.

```
Call:
glm(formula = Count_All_Casualties ~ Involving_Drink_Driving +
    Involving_Driver_Speed + Involving_Fatigued_Driver + Involving_Defective_Vehicle,
    family = poisson(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-30.729   -4.112   -0.226    2.082   45.703

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 6.293238   0.008107  776.28   <2e-16 ***
Involving_Drink_Driving    -2.442808   0.027959  -87.37   <2e-16 ***
Involving_Driver_Speed     -3.052139   0.037277  -81.88   <2e-16 ***
Involving_Fatigued_Driver  -3.620605   0.051312  -70.56   <2e-16 ***
Involving_Defective_Vehicle -3.831642   0.060168  -63.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 64006  on 185  degrees of freedom
Residual deviance: 18581  on 181  degrees of freedom
AIC: 19399

Number of Fisher Scoring iterations: 5
```



5. Data Storytelling

5.1. Conclusions

In summary, the findings of this analysis highlight important insights for stakeholders who may not be familiar with the technical aspects of the study, but are interested in understanding the research outcomes. Analysis of the data revealed that Brisbane City, Gold Coast City, Logan City, Moreton Bay Region, and Sunshine Coast Region had the highest number of car crashes throughout the year 2021. Factors such as driver speed, road conditions, involvement in drinking while driving, driver speed, fatigued driving, and defective vehicles were found to be significant predictors of car crashes and associated casualties. Additionally, speeding and tolling offenses were found to have a significant positive relationship with outstanding balances and number of debts.

Furthermore, installation of speed cameras may not necessarily lead to a decrease in the number of crash accidents in certain regions. Although Brisbane and Gold Coast have a relatively high number of cameras installed, there are still incidents of crashes that occur without any camera coverage. On the other hand, regions such as Moreton Bay and the Sunshine Coast have a lower number of installed cameras, yet there are areas where accidents still occur despite the presence of cameras.

Thus, other factors may be relevant to these accidents.

5.2. Suggestions to Stakeholders

1. **The government** should invest in improving road conditions and implementing stricter laws and regulations to ensure that drivers adhere to the traffic rules. Additionally, it may be beneficial to conduct further research to better understand the effectiveness of speed cameras and other measures in reducing accidents, as well as to identify other relevant factors.
2. **The traffic police** should increase the number of random breath tests and roadside checks to curb the occurrence of drink driving. They should also enforce speed limits and ensure that drivers adhere to traffic rules to reduce the probability of car accidents.
3. **Drivers** should be encouraged to practice safe driving by adhering to traffic rules, avoiding driving while under the influence of alcohol or drugs, and ensuring that their vehicles are in a good condition. They should also consider taking regular breaks during long drives to prevent fatigue and to reduce the probability of accidents.
4. **Insurance Companies** should consider offering discounts to drivers who adhere to traffic rules, have good driving records, and have their vehicles regularly serviced to incentivize safe driving behavior.
5. **The general public** should be encouraged to report any incidents of reckless driving, defective vehicles, or road hazards to the appropriate authorities. They should also be educated on the importance of safe driving practices to reduce the probability of car crashes and the associated costs.

6. Response to feedback

One important aspect of the feedback received in Trial Presentations pertains to the text-heavy nature of the slides, which hinders the audience's ability to read and absorb information while simultaneously listening to the speaker. To address this concern, we revised the slides to reduce the amount of text and utilize concise bullet points to highlight key information. By doing so, the audience can focus on essential points without feeling overwhelmed by excessive details.

Another observation made by the feedback provider suggests that the presentation may have been sped up to fit within the allotted 10-minute timeframe. This resulted in difficulties for the audience to follow along. To rectify this issue, we will prioritize minimizing irrelevant information, streamlining the presentation, and ensuring a smoother flow of content. By doing so, we aim to deliver a more coherent and easily understandable final presentation.

The feedback also highlights the appreciation of the insightful sections of the presentation. However, this suggests the need to summarize the information and present it using dot points to facilitate viewer comprehension. Considering this valuable suggestion, we revised the insight sections accordingly, presenting them in a concise format with clearly articulated key points. This approach enables viewers to read and grasp their insights more effectively.

We appreciate the feedback received as it provides crucial insights for improving the overall presentation. By implementing the suggested changes, we aim to create a more engaging and accessible experience for the audience, ensuring that the research results are effectively communicated to stakeholders who may not possess technical expertise in the analyzed datasets.

7. Data Sources

1. Active Mobile Speed Camera Sites
   https://www.data.qld.gov.au/datastore/dump/f6b5c37e-de9d-4041-8c18-f4d4b6c593a8?bom=True
2. Road Safety Camera Trailer Sites
   https://www.data.qld.gov.au/dataset/active-mobile-speed-camera-sites/resource/d059503f-3685-4669-8c43-df5b74da8ba8
3. Fixed speed camera, red light camera, combined red light/speed, point-to-point speed camera
   https://maps.google.com/maps/ms?msid=207652404510330023662.0004a7c174068b1f77c4e&msa=0
4. Road Crash Locations
   https://www.data.qld.gov.au/dataset/f3e0ca94-2d7b-44ee-abef-d6b06e9b0729/resource/e88943c0-5968-4972-a15f-38e120d72ec0/download/crash_data_queensland_1_crash_locations.csv
5. Factors in road crashes
   https://www.data.qld.gov.au/dataset/f3e0ca94-2d7b-44ee-abef-d6b06e9b0729/resource/18ee2911-992f-40ed-b6ae-e756859786e6/download/crash_data_queensland_e_alcohol_speed_fatigue_defect.csv
6. SPER Debts
   https://www.data.qld.gov.au/dataset/sper-debt-by-sa4-regions-and-top-10-offence-groups-in-queensland

Appendix

The analysis for this project was performed using the programming language (R and Python).

Is The Data Fit for Use?

1.1 Check for Completeness

```
import pandas as pd

# Read the Excel file
file_path = r'D:\Semester 1\Intro to Data Science\ASSESMENT-2-GROUP PROJECT\Clean_datasets.xlsx'
df_dict = pd.read_excel(file_path, sheet_name=None)

# Check for missing values
for sheet_name, df in df_dict.items():
    print(f'Missing values in {sheet_name}:')
    print(df.isnull().sum())
    print()

# Check for outliers
for sheet_name, df in df_dict.items():
    print(f'Outliers in {sheet_name}:')
    print(df.describe())
    print()
```

```
Missing values in Crash_Data:
Crash_Ref_Number                0
Crash_Severity                  0
Crash_Year                      0
Crash_Month                     0
Crash_Location                  0
Crash_Street                    0
Loc_Suburb                      0
Loc_Local_Government_Area       0
Loc_Police_Division             0
Loc_Police_District             0
Loc_Police_Region               0
Loc_ABS_Statistical_Area_4      0
Crash_Traffic_Control           0
Crash_Speed_Limit               0
Crash_Road_Surface_Condition    0
Crash_Atmospheric_Condition     0
Crash_Lighting_Condition        0
```

```
Missing values in Factor_of_Crash:
Crash_Year                      0
Loc_Police_Region               0
Crash_Severity                  0
Involving_Drink_Driving         0
Involving_Driver_Speed          0
Involving_Fatigued_Driver       0
Involving_Defective_Vehicle     0
Count_Crashes                   0
Count_Fatality                  0
Count_Hospitalised              0
Count_Medically_Treated         0
Count_Minor_Injury              0
Count_All_Casualties            0
```

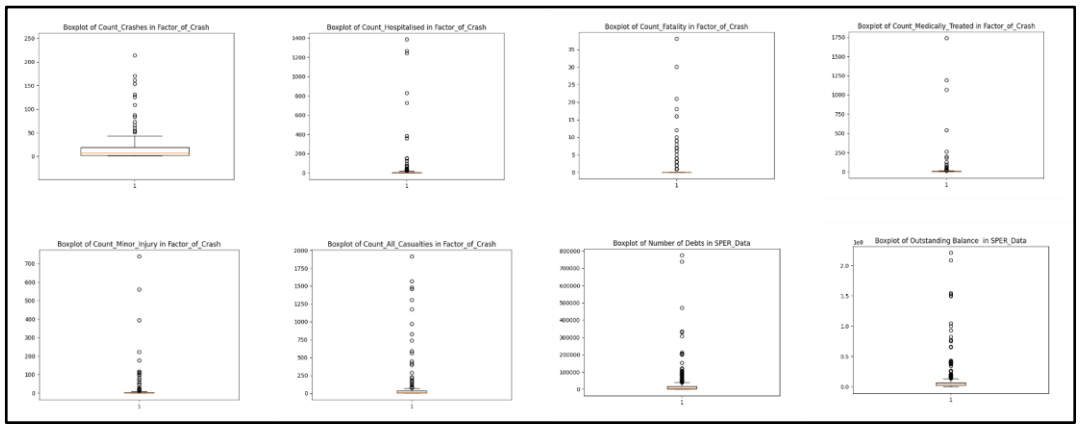
```
Missing values in SPER_Data:
SA4 Month                                   0
Loc_ABS_Statistical_Area_4                  0
Top 10 Offence Group by Outstanding Balance 0
Number of Debts                             0
Outstanding Balance                         0
```

```
Missing values in Camera_Location:
camera_type          0
camera_location      0
camera_coordinate    0
```

1.2 Check for Outliers

```python
# Check for outliers by descriptive statistics
for sheet_name, df in df_dict.items():
    print(f'Outliers in {sheet_name}:')
    print(df.describe())
    print()

# Check for outliers using boxplot
import matplotlib.pyplot as plt
for sheet_name, df in df_dict.items():
    print(f'Outliers in {sheet_name}:')
    for col in df.select_dtypes(include=['int64', 'float64']):
        plt.boxplot(df[col])
        plt.title(f'Boxplot of {col} in {sheet_name}')
        plt.show()
```

```
Outliers in Camera_Location:
                           camera_type  ...                    camera_coordinate
count                             3813  ...                                 3813
unique                               7  ...                                 2337
top     Active Mobile Speed Camera Site  ...  -26.187649999999962, 152.6591400000001
freq                              3038  ...                                   29
```




```
Outliers in Factor_of_Crash:
       Crash_Year  Count_Crashes  ...  Count_Minor_Injury  Count_All_Casualties
count       186.0     186.000000  ...          186.000000            186.000000
mean       2021.0      71.790323  ...           18.870968             96.112903
std           0.0     219.543591  ...           77.955528            298.390107
min        2021.0       1.000000  ...            0.000000              1.000000
25%        2021.0       2.000000  ...            0.000000              3.000000
50%        2021.0       7.000000  ...            1.000000              9.000000
75%        2021.0      19.000000  ...            4.000000             29.750000
max        2021.0    1525.000000  ...          738.000000           1913.000000

[8 rows x 7 columns]

Outliers in SPER_Data:
       Number of Debts  Outstanding Balance
count      1506.000000         1.506000e+03
mean      16431.815767         6.754566e+06
std       40917.107320         1.378936e+07
min           1.020000         2.750000e+02
25%         675.250000         2.000750e+06
50%        6610.500000         4.275000e+06
75%       16374.000000         6.790250e+06
max      774786.000000         2.207940e+08
```

1.3 Find the coordinate of mobile speed camera

```python
import csv
from geopy.geocoders import ArcGIS

# create an instance of the ArcGIS geocoder
nom = ArcGIS(timeout=10)

# open the CSV file containing the addresses
with open("C:/Users/Nihlahtuzzahra/Downloads/addresses.csv", 'r') as file:
    # create a CSV reader object
    reader = csv.reader(file)
    # create a new list to store the rows with the geocode data
    rows = []
    # iterate over each row in the CSV file
    for i, row in enumerate(reader):
        # get the address from the row
        address = row[0]
        # use the geocode() method to get the geocode for the address
        location = nom.geocode(address)
        # check if the geocode was found
        if location is not None:
            # add the latitude and longitude to the row
            row.append(f"{location.latitude}, {location.longitude}")
        else:
            # if no geocode was found, add an empty string to the row
            row.append('')
        # add the row to the list of rows
        rows.append(row)

# write the modified CSV file with the geocode data
with open("addresses_with_coordinates.csv", 'w', newline='') as file:
    # create a CSV writer object
    writer = csv.writer(file)
    # write the rows with the geocode data
    writer.writerows(rows)

print("File saved as addresses_with_coordinates.csv")
```

1.4 Find the coordinate of road safety camera

```python
from geopy.geocoders import ArcGIS

# create an instance of the ArcGIS geocoder
nom = ArcGIS(timeout=5)

# open the CSV file containing the addresses
with open("C:/Users/Nihlahtuzzahra/Downloads/rsct_coordinate.csv", 'r') as file:
    # create a CSV reader object
    reader = csv.reader(file)
    # create a new list to store the rows with the geocode data
    rows = []
    # iterate over each row in the CSV file
    for i, row in enumerate(reader):
        # get the address from the row
        address = row[0]
        # use the geocode() method to get the geocode for the address
        location = nom.geocode(address)
        # check if the geocode was found
        if location is not None:
            # add the latitude and longitude to the row
            row.append(f"{location.latitude}, {location.longitude}")
        else:
            # if no geocode was found, add an empty string to the row
            row.append('')
        # add the row to the list of rows
        rows.append(row)

# write the modified CSV file with the geocode data
with open("rsct_with_coordinates.csv", 'w', newline='') as file:
    # create a CSV writer object
    writer = csv.writer(file)
    # write the rows with the geocode data
    writer.writerows(rows)

print("File saved as rsct_with_coordinates.csv")
```

1.5 Checking the coordinate of camera

```python
import pandas as pd

# Load the excel file
df = pd.read_excel(r"C:\Users\Nihlahtuzzahra\Downloads\crash_coordinate.xlsx", sheet_name="Sheet2")

# Define the Queensland latitude and longitude boundaries
north_lat = -9.1428
south_lat = -29.1777
west_long = 138.0027
east_long = 153.5493

# Define a function to check if a coordinate is within Queensland
def is_in_queensland(latitude, longitude):
    # Convert the latitude and longitude to floats
    latitude = float(latitude)
    longitude = float(longitude)

    # Check if the coordinate is in Queensland
    if south_lat <= latitude <= north_lat and west_long <= longitude <= east_long:
        return "Located in Queensland"
    else:
        return "Not located in Queensland"

# Apply the is_in_queensland function to each row of the DataFrame
df["location_check"] = df.apply(lambda row: is_in_queensland(row["camera_coordinate"].split(",")[0],
                                                             row["camera_coordinate"].split(",")[1])
                                if len(row["camera_coordinate"].split(",")) >= 2
                                else "Invalid coordinate format",
                                axis=1)

# Save the updated DataFrame to a new excel file
df.to_excel(r"C:\Users\Nihlahtuzzahra\Downloads\crash_coordinate_updated.xlsx", index=False)
```

1.6 Checking the coordinate of crash

```python
import pandas as pd
# Load the excel file
df = pd.read_excel(r"C:\Users\Nihlahtuzzahra\Downloads\crash_coordinate.xlsx", sheet_name="Sheet1")

# Define the Queensland latitude and longitude boundaries
north_lat = -9.1428
south_lat = -29.1777
west_long = 138.0027
east_long = 153.5493

# Define a function to check if a coordinate is within Queensland
def is_in_queensland(latitude, longitude):
    # Convert the latitude and longitude to floats
    latitude = float(latitude)
    longitude = float(longitude)

    # Check if the coordinate is in Queensland
    if south_lat <= latitude <= north_lat and west_long <= longitude <= east_long:
        return "Located in Queensland"
    else:
        return "Not located in Queensland"


# Apply the is_in_queensland function to each row of the DataFrame

df["location_check"] = df.apply(lambda row: is_in_queensland(row["crash_coordinate"].split(",")[0],
                                                             row["crash_coordinate"].split(",")[1]),
                                axis=1)

# Save the updated DataFrame to a new excel file
df.to_excel(r"C:\Users\Nihlahtuzzahra\Downloads\crash_coordinate_updated.xlsx", index=False)
```

1.7 Finding the right camera location (located in Queensland)

```python
import pandas as pd
from geopy.geocoders import Nominatim
from geopy.exc import GeocoderTimedOut

# Load the Excel file with the street names
df = pd.read_excel(r'C:\Users\Nihlahtuzzahra\Downloads\camera_location.xlsx')

# Create a geolocator object from Nominatim
geolocator = Nominatim(user_agent='my_application')

# Define a function to get the coordinates from a street name
def get_coordinates(address):
    try:
        location = geolocator.geocode(address + ', Queensland, Australia', timeout=10)
        if location is not None:
            return location.latitude, location.longitude
        else:
            return None
    except (GeocoderTimedOut, Exception) as e:
        print(f"Error geocoding {address}: {e}")
        return None

# Apply the function to the camera_location column and create a new column
df['camera_coordinate'] = df['camera_location'].apply(get_coordinates)

# Save the updated Excel file
df.to_excel(r'C:\Users\Nihlahtuzzahra\Downloads\camera_location_with_coordinates.xlsx', index=False)
```

1.8 Create consistency of ABS SA4 Areas

```
import pandas as pd
import glob
files = glob.glob('2021/*.csv')
dfs = []
for file in files:
    df = pd.read_csv(file)
    print(df.shape[0])
    dfs.append(df)
combined_df = pd.concat(dfs, ignore_index=True)
combined_df.to_excel('3.outstanding balances.xlsx', index=False)
df_excel = pd.read_excel('3.outstanding balances.xlsx')
combined_df_filter = df_excel[df_excel['SA4 Name'] != 'Other QLD']
combined_df_filter = combined_df_filter[combined_df_filter['SA4 Name'] != 'Total QLD']
combined_df_filter = combined_df_filter[combined_df_filter['SA4 Name'] != 'QLD - All']
combined_df_filter = combined_df_filter[combined_df_filter['SA4 Name'] != 'Queensland - All']
combined_df_filter = combined_df_filter[combined_df_filter['SA4 Name'] != 'Queensland - Outback']
combined_df_filter = combined_df_filter[['SA4 Name', 'Top 10 Offence Group by Outstanding Balance','Number of Debts', 'Outstanding Balance']]
combined_df_filter.to_excel('3.outstanding balances2.xlsx', index=False)
```

1.9. Year Consistency (Only use 2021)

Crash Factor Dataset

```
import pandas as pd

read_factor = pd.read_csv("factor_of_crash.csv", delimiter= ';')
factor_2021 = read_factor[read_factor['Crash_Year'] > 2020]
factor_2021.to_excel('2.factor_to_crash.xlsx', index=False)
```

The Crash Dataset

```
import pandas as pd
df = pd.read_excel('crashdata.xlsx')
df_filter_data = df[['Crash_Ref_Number', 'Crash_Severity', 'Crash_Year',    'Crash_Month', 'The Right Location'
                    ,'Crash_Street', 'Loc_Suburb', 'Loc_Local_Government_Area', 'Loc_Police_Division', 'Loc_Police_District'
                    , 'Loc_Police_Region', 'Loc_ABS_Statistical_Area_4', 'Crash_Traffic_Control', 'Crash_Speed_Limit'
                    , 'Crash_Road_Surface_Condition','Crash_Atmospheric_Condition', 'Crash_Lighting_Condition']]

df_filter_data_2021 = df_filter_data[df_filter_data['Crash_Year'] > 2020]

df_filter_data_2021.to_excel('crashdatafilter.xlsx')
```

1.10 Merging 4 datasets

```
import pandas as pd

writer = pd.ExcelWriter('Merge Datasets.xlsx', engine='xlsxwriter')
df1 = pd.read_excel('1.crashdatafilter.xlsx')
df2 = pd.read_excel('2.factor_to_crash.xlsx')
df3 = pd.read_excel('3.oustanding balances.xlsx')
df4 = pd.read_excel('4.camera_location_datasets.xlsx')

df1.to_excel(writer, sheet_name='Crash_Data')
df2.to_excel(writer, sheet_name='Factor_of_Crash')
df3.to_excel(writer, sheet_name='SPER_Data')
df4.to_excel(writer, sheet_name='Camera_Location')

writer.close()
```

1.11 Data Transformations Factor of Crash

```
import pandas as pd

# Read the "Factor_of_Crash" sheet from the Excel file
file_path = r'D:\Semester 1\Intro to Data Science\ASSESMENT-2-GROUP PROJECT\Clean_datasets.xlsx'
df = pd.read_excel(file_path, sheet_name='Factor_of_Crash')

# Define a dictionary to map "Yes" and "No" to 1 and 0, respectively
mapping = {"Yes": 1, "No": 0}

# Apply the mapping to the specified columns
cols_to_convert = ['Involving_Drink_Driving', 'Involving_Driver_Speed', 'Involving_Fatigued_Driver',
                   'Involving_Defective_Vehicle']
df[cols_to_convert] = df[cols_to_convert].replace(mapping)

# Save the updated DataFrame to a new Excel file
new_file_path = r'D:\Semester 1\Intro to Data Science\ASSESMENT-2-GROUP PROJECT\Updated_datasets.xlsx'
df.to_excel(new_file_path, index=False)

# Print a message to confirm that the file has been saved
print(f"Updated dataset saved to {new_file_path}")
```

Cameras and Crash Locations

```r
library(readxl)
library(ggplot2)
Crash_data <- read_excel("Clean_datasets_2.xlsx")
Crash_data$Latitude = as.numeric(Crash_data$Latitude)
Crash_data$Longitude = as.numeric(Crash_data$Longitude)
Camera_data <- read_excel("Clean_datasets_2.xlsx", sheet= 4)
Camera_data$Latitude <- as.numeric(Camera_data$Latitude)
Camera_data$Longitude <- as.numeric(Camera_data$Longitude)
Crash_longitude <- Crash_data$Longitude
Crash_lagitude <- Crash_data$Latitude
df_Crash <- data.frame(Crash_longitude, Crash_lagitude)
Camera_longitude <- Camera_data$Longitude
Camera_latitude <- Camera_data$Latitude
df_Camera <- data.frame(Camera_longitude, Camera_latitude)
scatter_plot <- ggplot() +
                geom_point(data = df_Crash,
                           aes(x = df_Crash$Crash_longitude,
                               y = df_Crash$Crash_lagitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = df_Camera,
                           aes(x = df_Camera$Camera_longitude,
                               y = df_Camera$Camera_latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location',
                     color = "Location")
theme_plot <- theme(
plot.title = element_text(family = "Helvetica", face = "bold", size = (15)),
legend.title = element_text(colour = "steelblue", face = "bold.italic", family = "Helvetica"),
legend.text = element_text(face = "italic", colour = "steelblue4", family = "Helvetica"),
axis.title = element_text(family = "Helvetica", size = (10), colour = "steelblue4"),
axis.text = element_text(family = "Courier", colour = "cornflowerblue", size = (10))
)
```

```r
Crash_brisbane <- Crash_data[Crash_data$Loc_Local_Government_Area == 'Brisbane City', ]
Crash_goldcoast <- Crash_data[Crash_data$Loc_Local_Government_Area == 'Gold Coast City', ]
Crash_logan <- Crash_data[Crash_data$Loc_Local_Government_Area == 'Logan City', ]
Crash_moreton <- Crash_data[Crash_data$Loc_Local_Government_Area == 'Moreton Bay Region', ]
Crash_sunshine <- Crash_data[Crash_data$Loc_Local_Government_Area == 'Sunshine Coast Region', ]
df_crash_brisbane <- data.frame(Crash_brisbane$Longitude, Crash_brisbane$Latitude)
df_crash_goldcoast <- data.frame(Crash_goldcoast$Longitude, Crash_goldcoast$Latitude)
df_crash_logan <- data.frame(Crash_logan$Longitude, Crash_logan$Latitude)
df_crash_moreton <- data.frame(Crash_moreton$Longitude, Crash_moreton$Latitude)
df_crash_sunshine <- data.frame(Crash_sunshine$Longitude, Crash_sunshine$Latitude)

brisbane_camera <- Camera_data[Camera_data$Longitude > 153.010503 & Camera_data$Longitude < 153.035599
                  & Camera_data$Latitude > -27.480858 & Camera_data$Latitude < -27.460001,]
goldcoast_camera <- Camera_data[Camera_data$Longitude > 153.184329 & Camera_data$Longitude < 153.551850
                  & Camera_data$Latitude > -28.265145 & Camera_data$Latitude < -27.690399,]
logan_camera <- Camera_data[Camera_data$Longitude > 152.799330 & Camera_data$Longitude < 153.290340
                  & Camera_data$Latitude > -27.938566 & Camera_data$Latitude < -27.587315,]
moreton_camera <- Camera_data[Camera_data$Longitude > 152.651273 & Camera_data$Longitude < 153.207221
                  & Camera_data$Latitude > -27.422384 & Camera_data$Latitude < -26.792279,]
sunshine_camera <- Camera_data[Camera_data$Longitude > 152.551100 & Camera_data$Longitude < 153.150790
                  & Camera_data$Latitude > -26.984769 & Camera_data$Latitude < -26.432195,]
five_areas <- rbind(brisbane_camera, goldcoast_camera,logan_camera, moreton_camera, sunshine_camera)
```

```r
scatter_plot_brisbane <- ggplot() +
                geom_point(data = df_crash_brisbane,
                           aes(x = df_crash_brisbane$Crash_brisbane.Longitude,
                               y = df_crash_brisbane$Crash_brisbane.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = brisbane_camera,
                           aes(x = brisbane_camera$Longitude,
                               y = brisbane_camera$Latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location in Brisbane City',
                     color = "Location")

scatter_plot_goldcoast <- ggplot() +
                geom_point(data = df_crash_goldcoast,
                           aes(x = df_crash_goldcoast$Crash_goldcoast.Longitude,
                               y = df_crash_goldcoast$Crash_goldcoast.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = goldcoast_camera,
                           aes(x = goldcoast_camera$Longitude,
                               y = goldcoast_camera$Latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location in Gold Coast City',
                     color = "Location")

scatter_plot_logan <- ggplot() +
                geom_point(data = df_crash_logan,
                           aes(x = df_crash_logan$Crash_logan.Longitude,
                               y = df_crash_logan$Crash_logan.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = logan_camera,
                           aes(x = logan_camera$Longitude,
                               y = logan_camera$Latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location in Logan City',
                     color = "Location")

scatter_plot_moreton <- ggplot() +
                geom_point(data = df_crash_moreton,
                           aes(x = df_crash_moreton$Crash_moreton.Longitude,
                               y = df_crash_moreton$Crash_moreton.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = moreton_camera,
                           aes(x = moreton_camera$Longitude,
                               y = moreton_camera$Latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location in Moreton Bay Region',
                     color = "Location")
```

```r
scatter_plot_sunshine <- ggplot() +
                geom_point(data = df_crash_sunshine,
                           aes(x = df_crash_sunshine$Crash_sunshine.Longitude,
                               y = df_crash_sunshine$Crash_sunshine.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = sunshine_camera,
                           aes(x = sunshine_camera$Longitude,
                               y = sunshine_camera$Latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location in Sunshine Coast Region',
                     color = "Location")

scatter_plot_5_areas <- ggplot() +
                geom_point(data = df_crash_brisbane,
                           aes(x = df_crash_brisbane$Crash_brisbane.Longitude,
                               y = df_crash_brisbane$Crash_brisbane.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = df_crash_goldcoast,
                           aes(x = df_crash_goldcoast$Crash_goldcoast.Longitude,
                               y = df_crash_goldcoast$Crash_goldcoast.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = df_crash_logan,
                           aes(x = df_crash_logan$Crash_logan.Longitude,
                               y = df_crash_logan$Crash_logan.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = df_crash_moreton,
                           aes(x = df_crash_moreton$Crash_moreton.Longitude,
                               y = df_crash_moreton$Crash_moreton.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = df_crash_sunshine,
                           aes(x = df_crash_sunshine$Crash_sunshine.Longitude,
                               y = df_crash_sunshine$Crash_sunshine.Latitude,
                               color = "Crash Location"),
                               shape = 21, size = 3) +
                geom_point(data = five_areas,
                           aes(x = five_areas$Longitude,
                               y = five_areas$Latitude,
                               color = "Camera Location"),
                               shape = 21, size = 2) +
                labs(x = "Longitude", y = "Latitude",
                     title = 'Scatter plot: Crash Location and Camera Location in 5 Areas',
                     color = "Location")
```

Regression Analysis to Outstanding Balance and Number of Debts

```r
spea <- read.csv("newdata.csv")
cor(spea$Number_of_Debts, spea$Outstanding_Balance)
model1 <- lm(Outstanding_Balance ~ method, data = spea)
summary(model1)
model2 <- lm(Number_of_Debts ~ method, data = spea)
summary(model2)
```

Poisson Regression for Analyzing Involvement Factors to Count of Crashes

```
data <- read.csv("clean_datas.csv")
# Splitting the dataset into training and testing sets.
library(caTools)
set.seed(123)
split <- sample.split(data$Count_Crashes, SplitRatio = 0.7)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)
```

```
model <- glm(Count_Crashes ~ Involving_Driver_Speed+
             Involving_Drink_Driving+Involving_Defective_Vehicle+
             Involving_Fatigued_Driver, data = data, family = poisson())
summary(model)
plot(model, which = 1)
```

Poisson Regression for Analyzing Involvement Factors to Count of Casualties

```
data <- read.csv("clean_datas.csv")
# Splitting the dataset into training and testing sets.
library(caTools)
set.seed(123)
split <- sample.split(data$Count_All_Casualties, SplitRatio = 0.7)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)
```

```
model <- glm(Count_All_Casualties ~ Involving_Driver_Speed+
             Involving_Drink_Driving+Involving_Defective_Vehicle+
             Involving_Fatigued_Driver, data = data, family = poisson())
summary(model)
plot(model, which = 1)
```