ASSESSMENT 2: DATA ANALYSIS

Part B: Report

36103 Statistical Thinking for Data Science

University Of Technology Sydney

Group Name

Retail Group

Group Members

Leah Nguyen

Tony Tan

Anmol Mittal

Ben McKinnon

Kasun Caldera

Manasa Burli Nagendra

Date

October, 2021

Contents

1 Executive Summary

2 Introduction

2.1 Global Road Fatalities

Road traffic accidents are the leading cause of injuries and deaths worldwide. According to the WHO, road traffic accidents kill roughly 1.3 million people and injure up to 50 million people each year (World Health Organization, 2021).

If current trends continue, road traffic accidents are expected to be the third leading cause of disease and injury worldwide by 2020. (Murray el al., 1996). The burden of traffic-related fatalities, disabilities, and injuries has a significant impact on the health and social and economic development of many countries, particularly low and middle-income countries (Nantulya et al., 2002).

2.2 Victoria Road Fatalities

For many decades, Victoria has been a leader in road safety in Australia. Victoria's fatality rate per 100,000 population in December 2019 was 4.06, compared to the national average of 4.68. (Australian Automobile Association, 2020). However, this does not change the fact that progress in decreasing the road toll in Victoria has halted, and motor vehicle crashes remain a major cause of death and injury, with significant health and cost consequences.

Having said that, the entire annual cost of road casualties in Victoria is projected to be well over $1 billion (including property damage), and the total lifetime medical and associated costs for motor vehicle traffic-related injury in 1993-94 were $570 million (VicRoads and TAC, 1997; Watson and Ozanne- Smith; 1997). Furthermore, Minister for Road Safety – Luke Donnellan (Towards Zero, 2020) indicated that over the next ten years, roughly 2,500 people in Victoria will die in automobile accidents, and 50,000 people will be hospitalized with serious and life-changing injuries.

Despite the fact that the Victoria management authority has been addressing the situation by enacting a number of rules and safety measures, serious road accidents are nearly always the result of driver carelessness and irresponsibility. This requires an analysis of the relative influence of factors on the number of deaths in Victoria in order to develop effective methods to minimize the rate of road fatalities.

Therefore, this study aims to:

- Analyze crash fatalities data in Australia according to several factors, including age, causes of crash deaths, and type of road users, locations, etc. to finding factors related to fatalities conditioned on there being an accident in the first place.

- Develop a prediction model for crash fatalities in Australia based on crash fatalities data for the period 2006-2020 by using the logistics modelling method.

3 Data

3.1 Accident Data

The national road crash database obtained from the Road Crash Information System was the major source of data used in this study (RCIS). The database contains 9 tables, 500,000+ crash records, and 100 features spanning a 4-year period from 2006 to 2020. (Figure 1, 2).

With such a vast number of features, we can investigate several methods for predicting the fatal probability associated with different factors. The dataset is made up of features in multiple formats that include both numerical and categorical data. Because the aim of this research project was to discover the factors that potentially contribute to fatal road injuries, data was refined and parsed in the Data Pre-processing section for further analysis.

3.2 Weather Data

In order to assist our analysis, our group has also gathered weather data from Victoria from 2006 to 2020. In our study, we adopted the Visual Crossing Weather API to make an API request for the location based on a collection of all the distinct postcodes from the node dataset. Due to the time required to scrape data from the API, the program has divided down the collection to collect weather data for all suburbs yearly. The collection contains approximately 1,000,000+ weather records in various forms, each with 25 features (Figure 3).

As a result, the weather data for each year was combined, the address columns were split to match by postcode, and the postcode was eventually matched. This will also be included in the merge data set.

3.3 Overview of final dataset

Since the API response contains a list of records for the date range specified in the query, we then use the location's postcode and the date to match each of our accident data to the weather. The final data is a merged version of the accident and weather data with the features of interest that were chosen. It is made up of 52 features (Figure 4) and 550,578 records, with each record representing a victim involved in an accident.

4 Exploratory Data Analysis

5 Method

5.1 Regression Model

In our study, we chose multivariate Logistic Regression Model to tackle the research questions. According to Menard (2002), logistic regression models are used to model the probability of a certain event based on independent predictor variables.

While more sophisticated machine learning (ML) techniques have arisen and been applied to accident investigations over the last decade, logistic regression offers several advantages over ML techniques that support its use in our group study.

To begin, regression model findings are easy to interpret. Only the independent variables and their coefficients are required to represent the model in a single formula. The model's coefficients can be used to determine critical variables, as well as the amount and direction of association between each independent variable (i.e., risk factor) and the dependent variable (fatalities rate).

Second, once risk factors are discovered, creating a logistic regression model is straightforward and does not require tweaking multiple hyperparameters, as machine learning methods do. Due to this property, logistic regression is frequently used as the first classifier in predictive research and serves as a valid baseline for more advanced classifiers.

Third, while association rules might be effective for identifying latent patterns in huge data sets, they are fundamentally distinct from classification methods such as logistic regression modeling.

5.2 Data Preparation

5.2.1 Data Filtering

Traditionally, building statistical models starts with selecting variables that can result in a parsimonious model (i.e., having as few variables as possible). In order to do that, the first step is to edit it so that each point is genuinely useful, as larger is not necessarily better (Christensen, 2020). One simple solution is to clearly understand what problems we expect to resolve from our dataset.

Since the aim of this project was to predict the factors that are likely to contribute to deaths from road accidents, we filtered the data points containing the variables 'Drivers' and 'Motorcyclists' of the Road User Types to avoid any bias during analysis. After filtering out, the number of rows is reduced to 311,199 records.

5.2.2 Data Pre-processing

Next, an essential component of statistical modelling is analyzing the data set to ensure the data is tidy and in a compliant format for desired modelling technique. Our merged dataset is unstructured and contains significant superfluous data (defined as not contributing significantly to the prediction process). Since large datasets require longer training times, data preprocessing is, therefore, required to overcome this limitation. Preprocessing involves various tasks including dealing with following problems:

- Handling Outliers
- Handling Missing Values
- Handling Skewness
- Encoding
- Data Imbalance

Therefore, within the dataset, the following data issues were identified and handled accordingly.

5.2.2.1 Dealing with Missing Values Handling missing values within the data can be tedious. While some methods can be convenient, they can also be at the expense of the data's integrity. For instance, while handling numerical variables was straightforward, dealing with missing categorical data had challenges.

In our dataset, there are total of 135,303 missing values, which equivalent to 0.83% of our dataset. As illustrated in Figure 5, the majority of the important variables have little or no missing values while most of the missing values are associated with various vehicle information. In order to deal with missing data, our group has come up with 6 common techniques, including:

- Drop / Remove all missing values
- Imputation Using Mode Values
- Imputation Using (Mean/Median) Values
- Amelia predictive model (Multiple Imputation)
- K-nearest neighbor
- Random Imputation

in which each of them is fully explained in Figure 6.

However, these methods have their pros and cons. Based on our research problem and technical requirement, we have chosen to reject some of the approaches as illustrated in Table 1.

Techniques	Problems
Drop / Remove all missing values	“Upon first inspection, multiple variables contained large amounts of missing fields within numerical and categorical data fields. Dropping would result in thousands of lost rows; hence this method was rejected.”
Imputation Using Mode Values	“Due to the skewness of several categorical variables, replacing missing values with the mode made this skew even larger; hence this method was rejected.”
Imputation Using (Mean/Median) Values	Significantly reduce the model’s accuracy and bias the results since it can has an impact on attributes variability.
Amelia predictive model (Multiple Imputation)	“A multiple imputation method that replaces missing values with a bootstrap approach. This approach required 50 gigabytes of system memory to perform, which was resource-intensive for a home computer.”
K-nearest neighbor	“This imputation method was trialed and, similarly to Amelia, is computationally expensive since KNN only works by storing the whole training dataset in memory.”

-> **Table 1**. Missing data dealing techniques rejection <-

As final, we chose **Random Imputation** as the technique for dealing with missing values before constructing regression model (Figure []). This method eliminates the imputation variance of the estimator of a mean or total, and at the same time preserves the distribution of item values (Chen, et al., 2000).

5.2.2.2 Dealing with skewed data As illustrated in Figure [], there were degrees of skew on some of our numerical data, specifically in precipitation, cloud cover and relative humidity columns. Skewed data is troublesome as logistic regression assumes a normal distribution (Bill, 2014).

We need to handle this skewness, as our modelling algorithms assumes a normal distribution (Bill, 2014). As our skewed data follows closely to beta distributions, it was appropriate to use log transformation (Hammouri, Sabo and Alsaadawi, 2020).

5.2.2.3 Standardizing numerical data Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled (Bhandari, 2021). This is important to give each data point greater meaning by transforming the data to comparable scales.

Since our numerical data is spread over on a relatively large scale and needs to be dealt for more effective modelling, standardization need to be employed. In our case, Normal Standardization is the approach that is considered most effective as our numerical data is transformed or already follows the Gaussian Distribution (Lakshmanan, 2019).

As we are rescaling our data for standardization, we use the scale method in R with the default settings, so that the data has a mean of 0 and standard deviation of 1 (Geller, 2019; see also Saporta, 2013; Scale Function - RDocumentation, n.d.).

5.2.2.4 Handling Outliers Outliers in data can have significant impacts on analysis of the data. If outliers are present in the data during analysis, they may increase the error variance, reduce the strength of statistical tests, and can bias the results used in estimates of model parameters. Consequently, it is vital to scan and address outliers in the pre-processing phase (Dolgun, 2020).

The decision to remove outliers was based on the belief that they were present due to reporting errors or the presumption that they would impact results (3). Within our meaningful variables, we found `AGE` column to have several outliers (Figure []).

Taking a closer group in terms of different age groups, according to Figure [], all outliers are found in age group 70+ using the univariate box plot approach to detect outliers in AGE for a given `AGE GROUP`. Another strategy is using z-score approach to find out outliers of AGE variable.

First, our group calculated z-scores. Then I will find all observations that have the absolute value of AGE's z-score is greater than 3 (Figure []). From our observation, the minimum z score is -2.4227 and the maximum is 3.9373; there are only 373 out of 311,199 observations, which is only 0.1% of observations, are outliers. In this case, our group decided not to do anything to change the outliers because the age of people should not be changed.

5.2.2.5 Categorical Data Encoding In this study, the features collected are the combinations of categorical and numerical data. Categorical variables possess a vast detail of information that links to the target variables; However, as we will use machine learning approach is logistic regression, it cannot operate on categorical values directly and require the input variables and the output variables to be numeric. Therefore, in order to represent categorical information, 2 common techniques as One-Hot Encoding and Label Encoding are proposed (Figure []).

Since our categories values contain different variables with little relationships to each other, we adopted this technique to deal with categorical variables, specifically, variables that were deemed to be primarily present in fatal accidents (Figure []).

5.2.2.6 Data imbalance In our dataset, the class imbalance problem presents an important challenge. In particular, it was observed that value counts of non-fatalities and fatalities were 540,266 and 10,312, respectively. In this case, the imbalance data happens to be the "Rare Class Problem", in which the number of examples of one class is more than the others (Maheshwari et al., 2018).

Working with such imbalanced datasets presents the difficulty that most machine learning algorithms ignore, and so perform poorly on, the minority class, despite the fact that performance on the minority class is often the most significant. Therefore, the chosen methods to solve this problem were:

- **Under sampling:** The class of focus total occurrences were found, and then a random sample of the overpopulated class was taken to create a subsetted data set. This essentially achieves the same as randomly eliminating cases from the majority class (Pykes, 2020).
- **SMOTE:** Oversampling is another way to solve this problem, which has an advantage of no data loss compared to undersampling, however it risks overfitting due to replicated observations

(Vidhya, 2020). The article follows up with a better approach called the synthetic minority oversampling technique (SMOTE), which apparently generates artificial data by using bootstrapping and k-nearest neighbors. We can see the introduction and a more in-depth explanation of this technique from Chawla et al. (2002) where page 329 has pseudocode of the logic. Luckily, the article from Vidhya (2020) introduces the DmwR package, which has the SMOTE function to help with applying the SMOTE technique quickly.

5.2.3 Features Selection

5.2.4 Modelling

6 Evaluation

7 Limitations

8 Conclusion

9 References

10 Appendix

10.1 Figure 1. Accident Data Tables Description

Table	Description
accident	“basic accident details, time, severity, location”
person	“person based details, age, sex etc”
vehicle	“vehicle based data, vehicle type, make etc”
accident_event	“sequence of events e.g. left road, rollover, caught fire”
road_surface_cond	“whether road was wet, dry, icy etc”
atmospheric_cond	“rain, winds etc”
sub_dca	detailed codes describing accident
accident_node	master location table - NB subset of accident table
Node	Lat/Long references

10.2 Figure 2. Accident Data Attributes Description

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
ACCIDENT_DATE	Text	255	Accident Date	dd/mm/yyyy. (e.g.: 10 July 1995 = 10/07/1995)
ACCIDENT_TIME	Text	255	Accident Time	hh.mm.ss
STAT_DIV_NAME	Text	40	STAT_DIV_NAME is a character field indicating the Metro Melbourne or Country region where the crash occurred.	“Metro, Country.”
ACCIDENT_NO	Text	12	“From November 2005 the accident number field was changed to be 12 character field, starting with T (for example, T20060123456) Where characters 2 to 5 are the year in which accident was registered; Where characters 6 to 12 are a numeric sequencing numbers”	“Example: 12001012345, T20060006259”
ACCIDENTDATE	Date		Date of accident. Australian format DD/MM/YYYY	(e.g.: 10 July 1995 = 10/07/1995)
ACCIDENTTIME	Text	225	“hh.mm.ss. Original date stored in 24 hour format (ie 1pm = 1300 hours) Note the common practice used by the Police, when originally coding up the accident details, of ‘rounding off the time’ to the nearest 5 minutes or even nearest hour. This naturally occurs because in the vast majority of accidents police arrive at the scene well after the accident occurred and so the ‘REAL’ time of the accident is never precisely known.”	Examples of various PC time formats: 24 Hour format 2:35:00 PM = 14:35 or 12 Hour format 2:35:00 PM = 02:35PM 9999 Unknown time midnight = 00:00

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
ACCIDENT_TYPE	Number		“the type of accident. It is a basic description of what occurred, based on nine categories.”	1 Collision with vehicle 2 Struck pedestrian 3 Struck animal 4 Collision with a fixed object 5 Collision with some other object 6 Vehicle overturned (no collision) 7 Fall from or in moving vehicle 8 No collision and no object struck 9 Other accident
DAY_OF_WEEK	Number		the day of the week upon which the accident occurred	1 Sunday 2 Monday 3 Tuesday 4 Wednesday 5 Thursday 6 Friday 7 Saturday
DCA_CODE Part 1	Text	3	the Definitions for Classifying Accidents	
LIGHT_CONDITION	Number		the light condition or level of brightness at the time of the accident.	1 Day 2 Dusk/dawn 3 Dark street lights on

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				4 Dark street lights off
				5 Dark no street lights
				6 Dark street lights unknown
				9 Unknown
NO_PERSONS	Number	4	the number of people involved in the accident	
NO_PERSONS_KILLED	Number	4	Number of people with a given injury level	
NO_PERSONS_INJ_2	Number	4	Number of people with a given injury level	
DCA_CODE Part 2	Text	3		
NO_PERSONS_INJ_3	Number	4	Number of people with a given injury level	
NO_PERSONS_NOT_INJ	Number	4	the number of people that were not injured in the accident	
NO_OF_VEHICLES	Number	4	“the number of vehicles involved in the accident. Includes bicycles but not objects, property, toys (skate boards), etc.”	
POLICE_ATTEND	Number		Whether or not the police attended the scene of the accident.	1 Yes
				2 No
				9 Not known
ROAD_GEOMETRY	Number		The layout of the road where the accident occurred	1 Cross intersection
				2 ‘T’ Intersection
				3 ‘Y’ Intersection
				4 Multiple intersections
				5 Not at intersection
				6 Dead end

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				7 Road closure 8 Private property 9 Unknown
SEVERITY	Text		VicRoads estimation of the severity or seriousness of the accident	1 Fatal accident 2 Serious injury accident 3 Other injury accident 4 Non injury accident
DIRECTORY	Text		indicates the name of the street directory used to provide a map reference for the accident.	MEL Melway directory VCD Vic Roads directory
EDITION	Text	70	the edition or version of the street directory used to provide a map reference for the accident	MEL Melway directory VCD Vic Roads directory e.g. ED30
PAGE	Text	70	the page number of the street directory used to provide a map reference for the accident	MEL Melway directory VCD Vic Roads directory e.g. 91A
GRID_REFERENCE_X	Text	70	the grid reference in the x direction of the cell in the street directory used to provide a map reference for the accident.	
GRID_REFERENCE_X	Text	70	the grid reference in the y direction of the cell in the street directory used to provide a map reference for the accident.	

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
SPEED_ZONE	Text	3	the speed zone at the location of the accident. The speed zone is generally assigned to the main vehicle involved.	040 40 km/hr 050 50 km/hr 060 60 km/hr 075 75 km/hr 080 80 km/hr 090 90 km/hr 100 100 km/hr 110 110 km/hr 777 Other speed limit “888 Camping grounds, off road” 999 Not known
NODE_ID	Text	70	The node id of the accident. It starts with 1 and incremented by one when a new accident location is indentified.	e.g. 43078
EVENT_SEQ_NO	Number	4	It starts with 1 and incremented for more than one event in the same accident.	
EVENT_TYPE	Text	1	type of incident event	0 Not applicable 1 Rollover on/off carriageway 2 Fell from vehicle 3 Ran off carriageway 4 Mechanical failure

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				5 Struck by stone/projectile/load 6 Fell in vehicle 8 Other 9 Not known C Collision
VEHICLE_1_ID	Text	1	first vehicle involved in the event	
VEHICLE_1_COLL_PT	Text	1	collision point on the vehicle.	0 Towed unit 1 Right front corner 2 Right side (forwards) 3 Right side (rearwards) 4 Right rear corner 5 Left front corner 6 Left side (forwards) 7 Left side (rearwards) 8 Left rear corner 9 Not known or Not Applicable F Front N None R Rear S Sidecar T Top/Roof U Undercarriage
VEHICLE_2_ID	Text	1	second vehicle involved in the event.	

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
VEHICLE_2_COLL_PT	Text	1	collision point on the vehicle.	0 Towed unit 1 Right front corner 2 Right side (forwards) 3 Right side (rearwards) 4 Right rear corner 5 Left front corner 6 Left side (forwards) 7 Left side (rearwards) 8 Left rear corner 9 Not known or Not Applicable F Front N None R Rear S Sidecar T Top/Roof U Undercarriage
PERSON_ID	Text	2	person involved in the specific accident event	
OBJECT_TYPE	Text	2	object involved in the specific accident event	1 Pole (telephone/electricity) 2 Tree (shrub/scrub) 3 Fence/Wall (including gates) 17 Traffic island
COMPLEX_INT_NO	Number	4	“the segment is part of a complex intersection. If accident is located in complex intersection, the field has non zero value.”	0 Not part of a complex intersection 1-n Valid complex intersection number

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
ROAD_ROUTE_1	Number	4	This is the primary road/route number for road_name_1.	Group Classifications are: 2000-2999 Freeways or Highways 3000-3999 Forest Rds 4000-4999 Tourist Rds 5000-5999 Main Rds 7000-7999 Ramps (mainly Freeway ramps) 9999 Unclassified Roads e.g. Council / Local roads
ROAD_NAME	Text	45	highest priority road at intersection OR road on which accident took place.	
ROAD_TYPE	Text	15	type of Road_Name	
ROAD_NAME_INT	Text	45	the primary name of the intersecting road	
ROAD_TYPE_INT	Text	15	the type or suffix of the intersecting road	
DISTANCE_LOCATION	Number	4	the distance (in metres) of the accident from the nearest intersecting road (if the crash is a non-intersection or mid-block accident).	Eg: 153
DIRECTION_LOCATION	Text	2	the direction of the accident from the nearest intersecting road (if the crash is a non-intersection or mid-block accident)	N North
				NE North East
				E East
				SE South East
				S South

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				SW South West W West NW North West UK Not known
NEAREST_KM_POST	Number	4	the distance (in metres) of the accident to the nearest or closest kilometre post	
NEAREST_KM_POST	Number	4	the distance (in metres) of the accident to the nearest or closest kilometre post	
OFF_ROAD_LOCATION	Text	40	the name of the closest landmark or marker to the accident	
ATMOSPH_COND	Text	1	atmospheric condition	1 Clear 2 Raining 3 Snowing 4 Fog 5 Smoke 6 Dust 7 Strong winds 9 Not known
ATMOSPH_COND_SEQ	Number	4	1 and incremented by 1 if more than one atmospheric condition is entered for the same incident	
LONGITUDE	Double	8	Geographical coordinates	
LATITUDE	Double	8	Geographical coordinates	
NODE_TYPE	Number	1	location type identified by the RCIS spatial system	I Intersection

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				N Non-Intersection O Off Road U Unknown
AMG_X	Double	8	“AMG coordinate X value. With the emergence of digital mapping (mid 1980s), the (then) Lands Department of Victoria defined a projection which would allow Victoria to be viewed as a single, continuous map coverage, rather than as multiple zones. This projection, known in VicRoads as Pseudo AMG, is based on AGD 66, but uses a UTM modified to have scale distortion of 1.0 at its centre, a centre based on 145 degrees longitude (Melbourne) and a single zone covering the whole state.”	e.g. 2519154.655
AMG_Y	Double	8	“AMG coordinate Y value. With the emergence of digital mapping (mid 1980s), the (then) Lands Department of Victoria defined a projection which would allow Victoria to be viewed as a single, continuous map coverage, rather than as multiple zones. This projection, known in VicRoads as Pseudo AMG, is based on AGD 66, but uses a UTM modified to have scale distortion of 1.0 at its centre, a centre based on 145 degrees longitude (Melbourne) and a single zone covering the whole state.”	e.g. 2390265.155
LGA_NAME	Text	25	the LGA name	e.g. DANDENONG

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
SEX	Text	1	the sex or gender of the person	M Male F Female U Not known
AGE	Number	4	how old the person was at the time of the accident. It is calculated by subtracting the person’s birth date from the accident date to give the person’s age in years	
INJ_LEVEL	Text	1	the level or degree of injury that the person has experienced as a result of the accident. It is calculated field using inj_police_level and taken_hospital	1 Fatality 2 Serious injury 3 Other injury 4 Not injured
SEATING_POSITION	Text	2	where the person was located on the vehicle	CF Centre-front CR Centre-rear D Driver or rider LF Left-front LR Left-rear NA Not applicable NK Not known OR Other-rear PL Pillion passenger PS Motorcycle sidecar passenger RR Right-rear

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
HELMET_BELT_WORN	Text	1	whether or not the person was wearing a helmet or seatbelt at the time of the accident	1 Seatbelt worn
				2 Seatbelt not worn
				3 Child restraint worn
				4 Child restraint not worn
				5 Seatbelt/restraint not fitted
				6 Crash helmet worn
				7 Crash helmet not worn
				8 Not appropriate
				9 Not known
ROAD_USER_TYPE	Text	2	the role of the person was at the time of the accident. It is calculated field using person_status and vehicle_type from vehicle table	1 Pedestrian
				2 Driver (of V-type 1-9 17 60-63 70-71)
				3 Passenger (of V-type 1-9 17 60-63 70-71)
				4 Motorcyclist
				5 Pillion Passenger
				6 Bicyclist (incl. passengers)
				7 Other driver (V-type 14-16 99)

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				8 Other passenger (V-type 14-16 99) 9 Not known
LICENCE_STATE	Text	1	the state of issue of the person’s driver license	A Australian Capital Territory B Commonwealth D Northern Territory N New South Wales O Overseas Q Queensland S South Australia T Tasmania V Victoria W Western Australia Z Not known _ Not available (Blank value entered)
PEDEST_MOVEMENT	Text	1	“indicates the movement or travel of the person, if classified as a pedestrian”	0 Not applicable 1 Crossing carriageway 2 Working/playing/lying or standing on carriageway 3 Walking on carriageway with traffic

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				4 Walking on carriageway against traffic 5 Pushing or working on vehicle 6 Walking to/from or boarding tram 7 Walking to/from or boarding other vehicle 8 Not on carriageway (e.g. footpath) 9 Not known
POSTCODE	Number	4	the postcode where the owner of the vehicle resides	
TAKEN_HOSPITAL	Text	1	whether or not the person was taken to hospital	Y Yes N No _ Not Known
EJECTED_CODE	Text	1	whether or not the person was ejected or thrown out of the vehicle	0 Not applicable 1 Total ejected 2 Partially ejected 3 Partial ejection involving extraction _ Not known
SURFACE_COND	Text	1	road surface condition	1 Dry

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				2 Wet 3 Muddy 4 Snowy 5 Icy 9 Unknown
SURFACE_COND_SEQ	Number	4	starts with 1 and incremented by 1 if more than one road surface condition is entered for the same incident.	
SUB_DCA_CODE	Text	3	SUB_DCA code of the accident. Link to DCA Chart and Sub DCA Codes https://vicroads-public.sharepoint.com/InformationAccess/Shared%20Documents/Road%20Safety/Crash/Accident/DCA_Chart_and_Sub_DCA_Codes.PDF	
SUB_DCA_CODE	Number	4	starts with 1 and incremented by 1 if more than one sub_dca is entered for the same incident Link to DCA Chart and Sub DCA Codes https://vicroads-public.sharepoint.com/InformationAccess/Shared%20Documents/Road%20Safety/Crash/Accident/DCA_Chart_and_Sub_DCA_Codes.PDF	
VEHICLE_YEAR_MANUF	Number	4	indicates the year in which the vehicle was built or manufactured. The data is stored in yyyy format.	

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
VEHICLE_DCA_CODE	Text	1	“links the vehicle with the movement depicted in the DCA table. For example, if the DCA code for the accident is 111 and the vehicle DCA code is 2, then an inspection of the DCA chart will show that the second vehicle involved in the accident was turning right.”	1 Vehicle 1
				2 Vehicle 2
				3 Not known which vehicle was number 1
				8 Not involved in initial event
INITIAL_DIRECTION	Text	2	“the initial or first direction of travel of the vehicle. For a vehicle that is turning, the initial direction will be different to the final direction. For a non-turning vehicle, the initial direction will be the same as the final direction.”	E East N North NE North east NW North west S South SE South east SW South west W West NK Not known
ROAD_SURFACE_TYPE	Text	1	Prior to 1990 only one road surface was stored. This value is stored with the first vehicle. Road surface for 1990 is available for each vehicle in the collision.	1 Paved
				2 Unpaved
				3 Gravel
				9 Not known
REG_STATE	Text	1	the state which is the vehicle is registered in	A Australian Capital Territory
				B Commonwealth
				D Northern Territory

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				N New South Wales O Overseas Q Queensland S South Australia T Tasmania V Victoria W Western Australia Z Not known _ (Blank value entered)/Not available
VEHICLE_BODY_STYLE	Text	6	the body type of the vehicle	
VEHICLE_MAKE	Text	6	the vehicle make or manufacturer	
VEHICLE_MODEL	Text	6	the model of the vehicle	E.g. FALCON 0 Unknown 66 Sleeper 75 Tow
VEHICLE_POWER	Number	4	“the power of the vehicle, in CCs or horsepower. For motor cycles, motor scooters and mopeds, the units will be CCs and for all other vehicles the units are rated horsepower.”	0 Unknown 1-1000 Horsepower 1-9999 CCs
VEHICLE_TYPE	Text	2	the type or category of vehicle	
VEHICLE_WEIGHT	Number	4	the weight or mass of the vehicle. The unit of measurement is kilograms.	
CONSTRUCTION_TYPE	Text	1	the construction or formation of the vehicle	A Articulated

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				P Interpretation is not known
				R Rigid _ (Blank value entered)
				Unknown
FUEL_TYPE	Text	1	the type of fuel used by the vehicle	D Diesel
				E Electric
				G Gas
				M Multi
				P Petrol
				R Rotary
				Z Unknown
NO_OF_WHEELS	Number	4	the number of wheels that the vehicle has	
NO_OF_CYLINDERS	Number	4	the number of engine cylinders that the vehicle has	
SEATING_CAPACITY	Number	4	the number of seats in the vehicle	
TARE_WEIGHT	Number	4	the tare or unladen weight of the vehicle. The unit of measurement is kilograms	
TOTAL_NO_OCCUPANTS	Number	4	indicates the number of occupants or people in the vehicle at the time of the accident	
CARRY_CAPACITY	Number	4	the carry or load capacity of the vehicle. The unit of measurement is kilograms	
CUBIC_CAPACITY	Number	4	indicates the cubic capacity of the engine of the vehicle. The unit of measurement is cubic centimetres	

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
FINAL_DIRECTION	Text	2	“the final or last direction of travel of the vehicle. For a vehicle that is turning, the initial direction will be different to the final direction. For a non-turning vehicle, the initial direction will be the same as the final direction”	E East
				N North
				NE North east
				NW North west
				S South
				SE South east
				SW South west
				W West
				NK Not known
FINAL_DIRECTION	Text	2	what the driver of the vehicle was attempting to undertake at the time of the accident. This information is meant to obtain via an interview of the vehicle’s driver.	
VEHICLE_MOVEMENT	Text	2	the actual movement of the vehicle prior to the accident.	
TRAILER_TYPE	Text	1	“the type of trailer towed by the vehicle involved in the accident, as reported by the police.”	
VEHICLE_COLOUR_1	Text	3	the primary or main colour of the vehicle.	
VEHICLE_COLOUR_1	Text	3	the secondary colour of the vehicle	
CAUGHT_FIRE	Text	1	whether or not the vehicle caught fire as a result of the accident.	0 Not applicable

FIELD NAME	TYPE	WIDTH	DEFINITION	DOMAIN
				1 Yes 2 No 9 Not known
INITIAL_IMPACT	Text	1	the position on the vehicle where the initial impact occurred.	
LAMPS	Text	1	whether the lamps or headlights for the vehicle (under the ambient lighting conditions) were alight (on).	0 Not applicable 1 Yes 2 No 9 Not known
LEVEL_OF_DAMAGE	Text	1	the damage level of the vehicle.	1 Minor 2 Moderate (driveable vehicle) 3 Moderate (unit towed away) 4 Major (unit towed away) 5 Extensive (unrepairable) 6 Nil damage 9 Not known
OWNER_POSTCODE	Number	4	the postcode where the owner of the vehicle resides.	

10.3 Figure 3. Weather API Data Attributes Description

Domain	Description
Address	“is the address, partial address or latitude,longitude location for which to retrieve weather data. You can also use US ZIP Codes.”
Date time	“ISO formatted date, time or datetime value indicating the date and time of the weather data in the locale time zone of the requested location”
Minimum Temperature	minimum temperature at the location.
Maximum Temperature	maximum temperature at the location.
Temperature	temperature at the location
Dew Point	dew point temperature
Relative Humidity	relative humidity in %
Heat Index	" a value between 0 and 10 indicating the level of ultra violet (UV) exposure for that hour or day. 10 represents high level of exposure, and 0 represents no exposure. The UV index is calculated based on amount of short wave solar radiation which in turn is a level the cloudiness, type of cloud, time of day, time of year and location altitude. Daily values represent the maximum value of the hourly values."
Wind Speed	average wind speed over a minute
Wind Gust	instantaneous wind speed at a location – May be empty if it is not significantly higher than the wind speed.
Wind Direction	direction from which the wind is blowing
Wind Chill	
Precipitation	the amount of precipitation that fell or is predicted to fall in the period
Precipitation Cover	the proportion of hours where there was non-zero precipitation
Snow Depth	the depth of snow on the ground
Visibility	distance at which distant objects are visible
Cloud Cover	how much of the sky is covered in cloud ranging from 0-100%
Sea Level Pressure	the sea level atmospheric or barometric pressure in millibars (or hectopascals)
Weather Type	
Latitude	“is the address, partial address or latitude,longitude location for which to retrieve weather data. You can also use US ZIP Codes.”

Domain	Description
Longitude	"is the address, partial address or latitude,longitude location for which to retrieve weather data. You can also use US ZIP Codes."
Resolved Address	"is the address, partial address or latitude,longitude location for which to retrieve weather data. You can also use US ZIP Codes."
Name	"is the address, partial address or latitude,longitude location for which to retrieve weather data. You can also use US ZIP Codes."
Info	NA
Conditions	textual representation of the weather conditions.

10.4 Figure 4. Dealing with missing values techniques

Accident attributes:

- **X:** The ID for the record
- **ACCIDENT_NO:** The accident id that the person was associated with
- **FATAL_ACCIDENT:** Categorical variable on whether the person was involved in an accident with fatalities. This will be our target variable.
- **ACCIDENTDATE:** The date of the accident
- **ACCIDENTTIME:** The time of the accident
- **DAY_OF_WEEK:** The day of week in numerical form
- **Day_Week_Description:** The day of week in categorical form
- **NO_OF_VEHICLES:** The number of vehicles involved in the accident
- **NO_PERSONS:** The number of people involved in the accident
- **NO_PERSONS_INJ_2:** The number of people with an injury level of 2
- **NO_PERSONS_INJ_3:** The number of people with an injury level of 3
- **NO_PERSONS_KILLED:** The number of people died
- **NO_PERSONS_NOT_INJ:** The number of people not injured
- **Accident_Type_Desc:** Description of the accident type
- **DCA_Description:** 'Definition for Coding Accidents' code description. Basically, the category of the accident.
- **SEVERITY:** The severity of the accident in numerical form

Accident location and environmental attributes:

- **SPEED_ZONE:** The speed zone of where the accident occurred
- **Road_Geometry_Desc:** The road geometry description

- **ROAD_NAME:** Name of the road the accident occurred on
- **ROAD_TYPE:** Type of the road the accident occurred on
- **ROAD_NAME_INT:** Name of the road intersection the accident is closest to
- **ROAD_TYPE_INT:** Type of the road intersection the accident is closest to
- **LGA_NAME:** The local government area name
- **Road_Surface_Type_Desc:** The road surface type description
- **Surface_Cond_Desc:** The road surface condition
- **LIGHT_CONDITION:** The light condition in numerical form
- **Light_Condition_Desc:** The light condition in categorical form

Weather attributes:

- **Atmosph_Cond_Desc:** The atmosphere condition description
- **Temperature:** The average temperature throughout the day
- **DewPoint:** The average dewpoint throughout the day
- **RelativeHumidity:** The average relative humidity throughout the day
- **WindSpeed:** The average windspeed throughout the day
- **Precipitation:** The total precipitation throughout the day
- **Visibility:** The average visibility throughout the day. The distance that can seen in daylight.
- **CloudCover:** The average cloud cover throughout the day
- **WeatherType:** The weather types throughout the day
- **Conditions:** The weather conditions throughout the day

Vehicle attributes:

- **VEHICLE_YEAR_MANUF:** When the vehicle that the person was riding was manufactured

- **VEHICLE_BODY_STYLE:** The shape and body of the vehicle the person was on when the accident happened
- **VEHICLE_MAKE:** The brand of the vehicle the person was on
- **VEHICLE_MODEL:** The model of the brand (specific product) of the vehicle
- **NO_OF_CYLINDERS:** The number of cylinders which power the car

Person attributes:

- **PERSON_ID:** Unique ID for the person involved in the accident
- **SEX:** The gender of the person
- **AGE:** The age of the person
- **Age_Group:** Grouped-up version of age
- **INJ_LEVEL:** The injury level in numerical form
- **Inj_Level_Desc:** The injury level in categorical form
- **SEATING_POSITION:** Indicates the seating position of where the person was sitting in a vehicle
- **Road_User_Type_Desc:** Type of road user the person is
- **POSTCODE:** Postcode of where the person lives
- **TOTAL_NO_OCCUPANTS:** The number of people in the vehicle the person was on

10.5 Figure 5. Dealing with missing values techniques

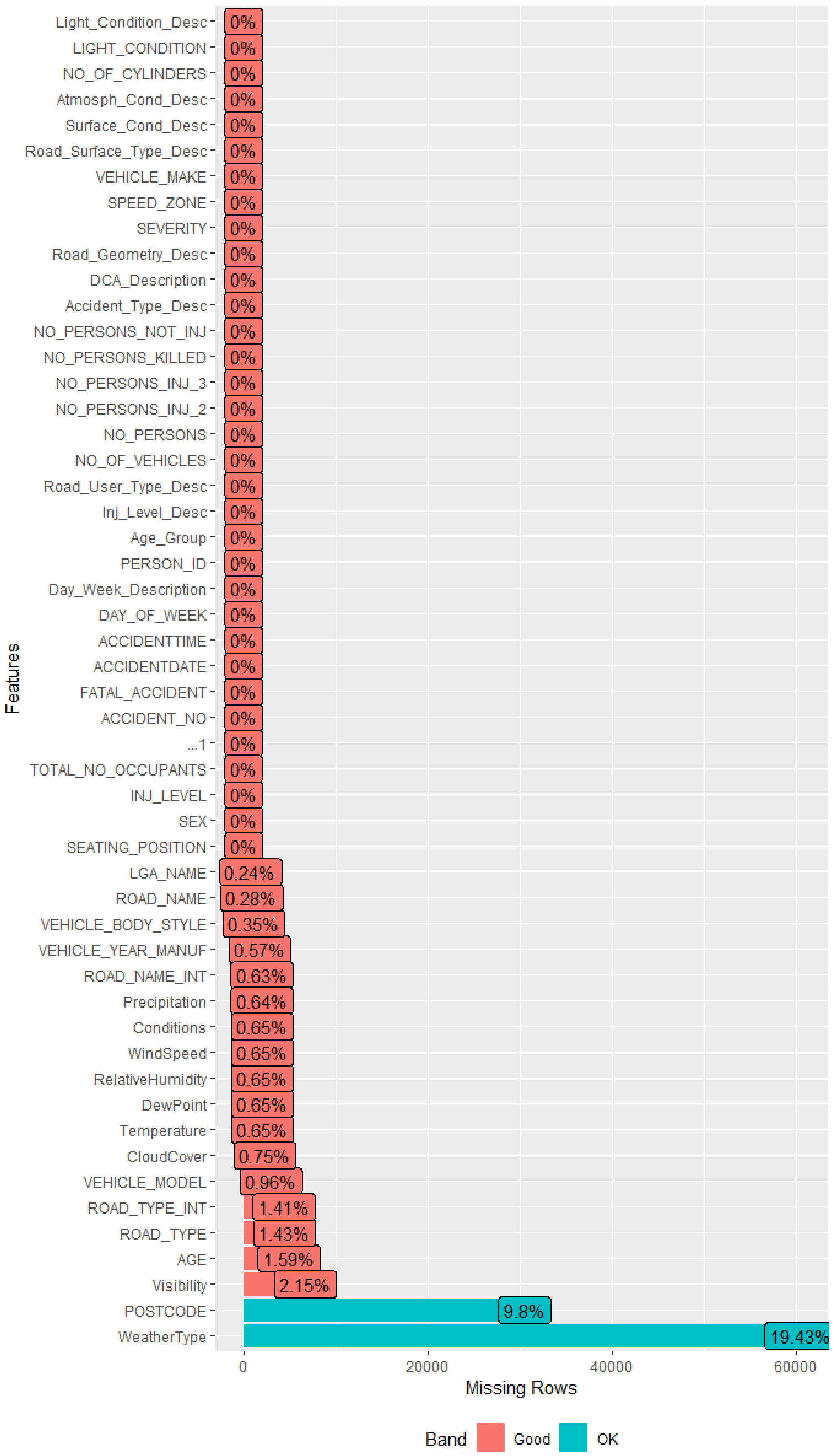

Figure 1: title

10.6 Figure 6. Dealing with missing values techniques

Techniques	Definition
Drop / Remove all missing values	The easiest method of dealing with missing values. This generally will exclude all those missing values out of the dataset by removing/deleting column rows.
Imputation Using Mode Values	“Mode/Most frequent is another statistical strategy to impute missing values. It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column (Badr, 2019).”
Imputation Using (Mean/Median) Values	“A widely common and accepted approach to missing data is to replace the NA with the mean. The nature of this method does not distort or change the distribution of values within each feature. Furthermore, the simplicity to implement this was the driving factor in this method being chosen (Badr, 2019).”
Amelia predictive model (Multiple Imputation)	“Multiple imputation involves imputing values for each missing cell in your data matrix and creating “completed” data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data. (Amelia II: A Program for Missing Data, n.d.)."
K-nearest neighbor	The k nearest neighbors is an algorithm that is used for simple classification. The algorithm uses ‘feature similarity’ to predict the values of any new data points. “This is useful in making predictions about the missing values by finding the k’s closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighborhood (Badr, 2019).”
Random Imputation	“On the other hand, replace missing numerical data with mean and replace missing categorical data at existing distribution are the two techniques we chose to deal with missing values in our dataset.”