

Assessment Report: Calderdale Accident Casualty

Author: Dharmarlou Bowen (20123076)

Date: 13/01/2022

Wordcount: 1554

Pagecount: 8

Calderdale Accident Casualty 2019/2020

**Contents**

[Introduction](#_bookmark0) 1

[Data Wrangling](#_bookmark1) 2

[Removing unnecessary columns](#_bookmark2) 6

[Identifying outliers in the ‘Age of Casualty’ column](#_bookmark3) 7

[Data Exploration](#_bookmark4) 9

[Regression](#_bookmark5) 14

[Conclusion](#_bookmark6) 14

# Introduction

This report is comprised of data wrangling, cleaning in addition to the manipulation of Calderdale road data collisions. The data that is used for this analysis is was produced by Calderdale Council available on their website at [‘https://dataworks.calderdale.gov.uk/dataset/](https://dataworks.calderdale.gov.uk/dataset/calderdale-accident-data-) [calderdale-accident-data-](https://dataworks.calderdale.gov.uk/dataset/calderdale-accident-data-)’. A breakdown of all data exploration, regression and cleaning will be documented and reported in the R file named “Assessment2.R” and the CSV files of the cleaned data and the regression predictions included as “regression.csv” and “cleanedSet.csv” respectively.

# Data Wrangling

#### Examining the columns in the data

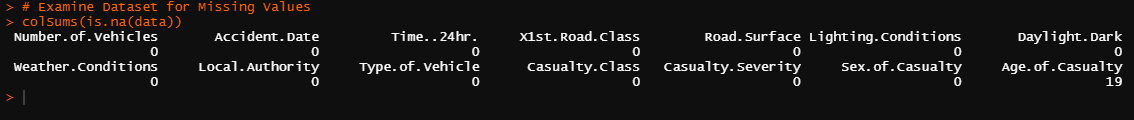
The Accident.csv dataset is comprised of 2069 rows and 14 columns. Columns by the name of “Time 24hr”, “Number of Vehicles”, “Accident Date” in addition to the other relevant columns each play a part in drawing insight from into the occurrence of accidents.

They will provide the ability to perform classifications of the types of vehicles that are commonly encountering accidents in combination with the time of day and the environment of accident. Environmental insight would be drawn from the columns of the names: “Road surface”, “lighting Conditions”, “Daylight/Dark” and “Weather Conditions”.

Insight into the area of occurrence is derived from the column, “Local Authority”. For insight into the victims of the accidents columns by the of “Casualty Class”, ”Age of Casualty”, “Sex of Casualty” would provide greater clarity into their shared characteristics and behaviors.

#### Evaluating Missing data

It was uncovered that 19 missing values in the dataset from the column “Age of Causality ”.



Now these values can all be found in the same column, age. There are 19 cases where the age of the casualty involved was either not discovered or disclosed by the person involved. You can see these 19 cases here:

With these 19 records containing missing values this could be due to the absence of the casualty during the data collection process, but it is also possible that the age wasn’t disclosed. Data can be unavailable because of MAR or MNAR. An example of MNAR would be that a casualty was a young driver and did not consent to their age because of a personal conflict of judgement about their driving proficiency from their peers.

#### Examination & Amending Anomalies

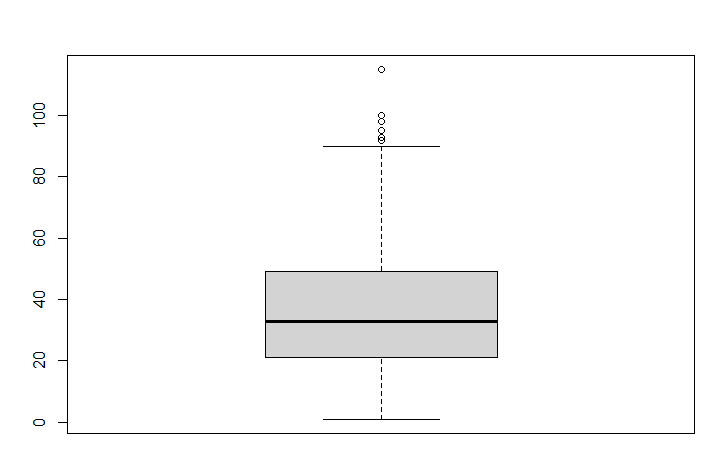
#### An anomaly was encountered in the column known as “Road Surface” and “1st Road Class”. As you can see from the imagine on the right there a distinctive change in data in the fields. The alphanumerical used in “Road Class” becomes strictly numeric and in the “Road surface” column it shares this name inconsistency as each field should contain a environmental category of either (Dry, Wet/Damp, Frost/Ice and Snow).

#### In order to remedy this issue, the follow dplyr functions were used in the image below:

#### 

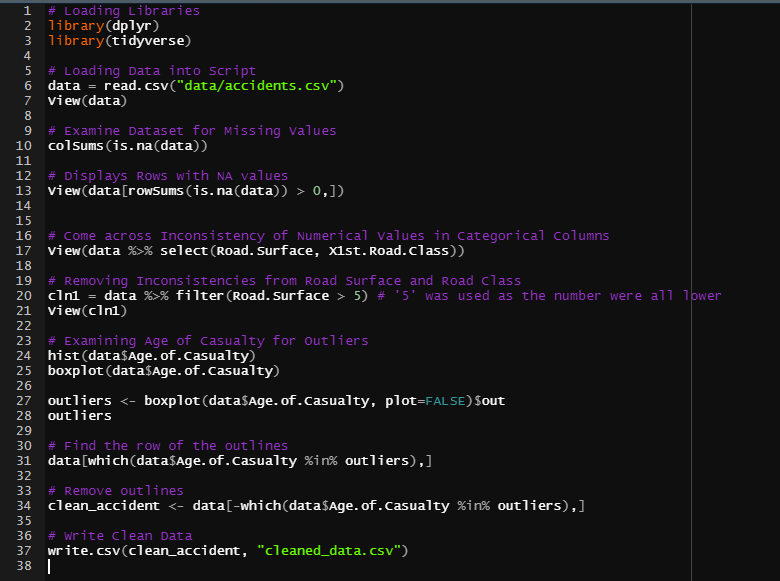
As you gather from the code above the anomalies have been removed which has significantly decreased the number of records to 336. The code removes any record that has a value less than 5 in correlation to road surface. As they are not meant to have numeric values in the first place.

**Examine & Remove Outliers (Age of Casualty)**

After examining the column “Age of Casualty” through the use of outlier discovery method of a boxplot which enabled the encountering of distinctly old ages that seem highly improbable given the average of a driver.

The outliers ages [155, 93, 100, 92, 98] were the outlier values that presented themselves in the diagram.

After also performing analysis with a histogram the data was not clear enough to draw any insight just off a quick glance, but with the box plot diagram the range, IQR and 1st and 3rd quartiles presented greater clarity.

**Save Clean Data**

# Data Exploration

#### Weather conditions and their effects on drivers of different genders

The question here posed is: **“Is there any weather condition where male drivers/riders have more accidents than female drivers?”.**

To visualize the proportion of each gender to each accident here is a bar graph showing how likely either gender is to be involved in a type of accident:

…

And if you want to see how each gender varies in terms of numbers on each category of accident then here are the simple numeric values:

…

Here is a reminder/lookup table to compare what each numeric value translates to in terms of a real-life weather condition scenario:

Table …

#### Casualty Numbers on a year-by-year basis

Since our data doesn’t specify the number of casualties involved in each accident, I am going to assume that there is one casualty per accident. With this assumption in mind, our data set gives us the total number of casualties however it is not grouped by year. To group by year, I made four data frames to collate the data into year groups. Next, I had to convert the column type of ‘Accident.Date’ from factor which is a data type that is tricky to manipulate into a character data type which would allow me to extract the year value from the string.

Here is the code for this below:

…..

From the chart we can see a steadily decreasing number of accidents occurring each year. The year with the highest number of accidents is 2014 with a total number of 623 accidents occurring throughout the year which accounts for ~30% of the accidents in the data.

#### Exploring the relationship between light condition and the severity of the casualty’s condition

This process is known as exploratory data analysis (EDA) and has a distinct set of graphs we can use to find relationships between features in data. I decided the best way to explore the relationship between these two features mentioned was to plot a line graph with each line representing the frequency of crashes of each severity class against each individual lighting condition after which I should be able to both see how the number of crashes resulting in the class of injury occur over each condition but also that I can compare it to the other two classes of injury.

There was quite of lot of manipulation I had to do to make a data frame that I could comfortably plot using GGplot and I am aware that it is very clunky and not succinct so I’ve decided not to include it directly in this report unlike the rest of codebase however the file is packaged with the project and the code can be viewed from there as well as in the raw copy of this report. The rest of the code to create the data frame and plot the graph is included below:

### …..

From this plot we can see how relationship between casualty severity and the lighting conditions at the time of the RTC by plotting how many accidents at each level of severity occurred at each lighting condition. This also helps us compare between the severity levels themselves to see which is more prevalent at each individual light level.

#### Exploring the relationship between weather conditions and the number of vehicles involved in an RTC

To explore this relationship, I’ve decided again to use a line graph however this time instead of using a different line to represent a different number of vehicles that have crashed in a different RTC I will instead take the average number of vehicles involved in crashes at that weather condition and plot that line into the graph.

Much like with the previous plots, this code requires some R code that isn’t succinct enough to fit in the report however it is also included in a separate file and is also bundled in with the raw version of this report, see below for the code to create the plot:

Avg no vehicles involved ~ Weather Conditions

# Regression

#### Training and imputing new values using linear regression

Linear regression follows a uniform path from which we can predict values of (x) based on one or more other values (y, z . . .). In simple graphs often made by hand, a line of best fit is drawn to help visualize the relationship between variables, it is also a great example of linear regression and how it is used to predict values. The first thing that must be done is to split our data into a Training set and a Test set. The training set will have a complete set of values for every feature whereas the test set will be missing values for the feature that we will later attempt to predict. The rest of the explanation will be provided via comments in the following code snippet:

…..

# Conclusion

With this data set we were able to draw a few conclusions. Firstly, that overall, crashes in the Calderdale area were on the decline. Secondly that the majority of accidents occur with only slight injuries with all types of injuries being most frequent in normal conditions or simply windy conditions. Next, that in weather conditions that are snowy and with high winds you’ll see the highest average number of vehicles involved in the crash. Finally, that women are less likely to crash than men when driving in weather conditions that are either snowy with high winds or unknown/other weather conditions