Data Analytics – Data Visualisation Coursework

Ross Hunter[1]

[1] School of Computing, Edinburgh Napier University
40478443@live.napier.ac.uk

Abstract. This report will cover the analysis of data provided, from UK road accidents. It identifies the presence of outliers in two of the attributes, shown in different forms with explanation of why they are considered outliers, why they have been retained and associated visualisations. Then, the believed five relationships between the various attributes, including what they are, how they were identified and visualisations showing this for each. The report concludes with an evaluation of a provided visualisation of fast-food chain sales, including positive aspects and mostly criticism, and followed by sketch to improve the visualisation and associated explanations of why it is better.

1 Outliers

1.1 Speed

One set of outliers identified is Speed (shown in Fig 1.). These would be considered outliers as group of several items existing in the 130-140 mph range while the majority speeds accidents took place in ends at around 90 mph.

This was discovered during an initial survey of generally what the data was like, such as finding any ranges data had and specifically if any immediate outliers could be identified.

This data has been retained as it displays the small number of accidents to take place beyond the legal speed limit, and does therefore not seem erroneous.

1.2 Damage

The other set of outliers are in Damage, appearing when plotted with another attribute such Road Type, shown by Fig 2.

This was initially shown with the damage/severity scatterplot, created while comparing various attributes in a 2D charts and later refined to confirm that the outliers were in damage and not severity.

These have been retained in the data set for the visualisations as they highlight some variation in the trend of damage per accident and also, don’t affect the data-ink ratio of any of the relationship visualisations.

2 Relationships

2.1 Damage and Severity by Road Type

The resulting Damage and the Severity of an accident have a positive correlation, as shown in Fig 3. This is also linked to the road type the accident occurred on, with unclassified roads having the least severe and damaging accidents while motorways have the most.

This was identified first by two separate relationships, Damage and Severity scatterplot then Severity and Road Type boxplot. The latter, by arranging the Road Types in the order U,C,B,A,M would create an increasing Severity, same as its relation to Damage.

2.2 Casualty Class and Brightness

As shown by Fig 4., a relationship exists between the type of casualty in an accident and the brightness of the road environment where the accident took place. The average number of ‘other’ casualties is generally higher than drivers or passengers in environments with greater brightness, with an average around 75,000 lux.

This relationship was found during first, simple 2-D visualising of the various attributes within the data set. It was later expanded using juxtaposition and small multiples to verify that no other attribute contributed to the relationship.

2.3 Age and Road Type by Gender

Another relationship identified from the data was the Age and Gender of the person involved and on which Road Type the accident occurred on. As Fig. 5 shows, there appears to be no males over the age of 30 that are involved in any accidents on motorways.

Identified after first plotting Road Type and Age on a boxplot, during initial 2D comparisons, showing a reduced average age on motorways. Subsequently, juxtaposed other metrics like Junction Type and Casualty Class without success before then identifying Gender as the other metric.

2.4 Year and Speed by Junction Type and Casualty Class

The fourth relationship found is where and when the accident took place. From Fig. 6, both years 1998 and 2002 have accidents at ‘Other’ type of junctions and only involving a driver that are different from the other years. For 1998 the speeds are only at 80-90mph while in 2002 they were down between 30 and 45mph.

The relationship was found when firstly plotting Year and Speed with a boxplot and noticing a small increase and dip for the years above. Then juxtaposed other discrete types, with Junction Type continuing this difference, Finally, used small multiples to check for any other verity, revealing Driver Casualty Class.

2.5 Age and Severity by Gender

The final relationship is between Age, Severity and Gender, shown by Fig 7. This highlights that males in younger age groups, those being 17-27, more likely to be in a more severe accidents than other ages and more than female counterparts. It also shows older females, generally above 40, are more slightly more likely to be in more severe accidents than male counterparts.

Identified by plotting the two variables against each other. This was first on a scatterplot, which yielded a mess, then chose to groping the ages into ranges of 10 and switch to a boxplot displayed it and expanded to juxtapose other attributes, settling on gender.

3 Visualisation Evaluation

3.1 Evaluation

The visualisation, on fast-food chains sales provided, is not well executed, and doesn't visualise the data well.

While the idea of using logos for the brands is interesting, it gives the visualisation a poor data-ink ratio with all the extra colour and these elements can be considered 'chart junk'. Additionally, the sizing results in a few of the logos being unreadable.

There is no labelling, for both the title and x-axis, with the label for y-axis being incorrectly placed. The presence of extra labelling for the exact sales/GDP of each brand/country highlights that it must no be clear to read the top against the axis.

While the inclusion of Afghanistan for a comparison is helpful, the positioning of the country's behind the brands makes it less clear and harder to spot and also has no distinction with the it not being a fast food brand.

3.2 Improvement

Fig 8., a sketch, shows how the visualisation could be improved. It has been stripped back to just simple bars to reduce the data-ink ratio and 'chart junk'. The colours have been selected from the respective logos.

The comparison for the country of Afghanistan has been placed further from the others to highlight the distinction between them.

For the data to be clearly understood, a title and good axes with corresponding labels have been provided.

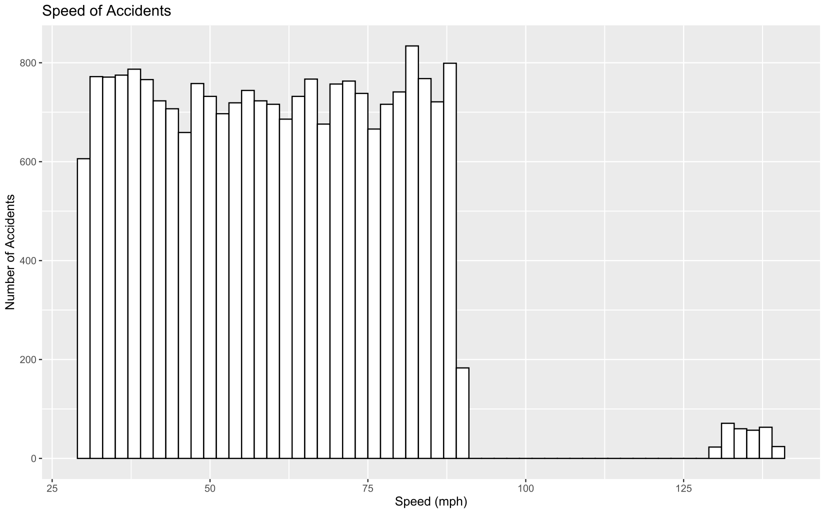

Fig. 1. Histogram of the speeds each vehicle was doing at each accident case, showing outliers above 125mph.

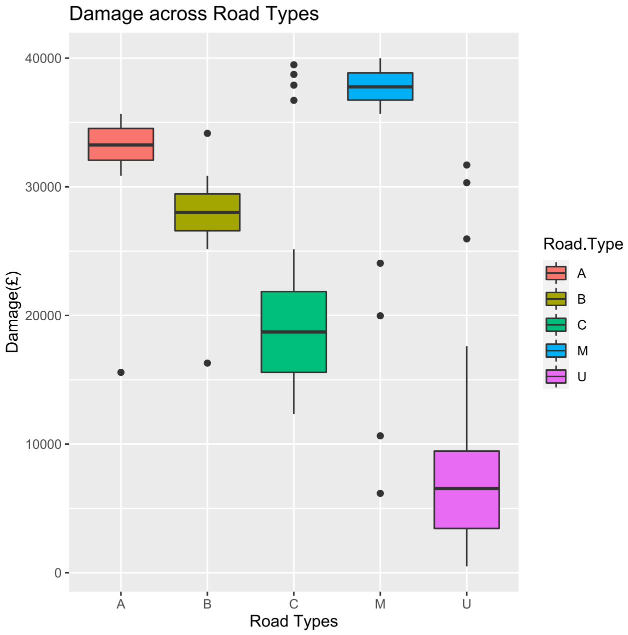

Fig. 2. Boxplot showing Damage of each accident separated by Road Type, highlighting the outliers in Damage.

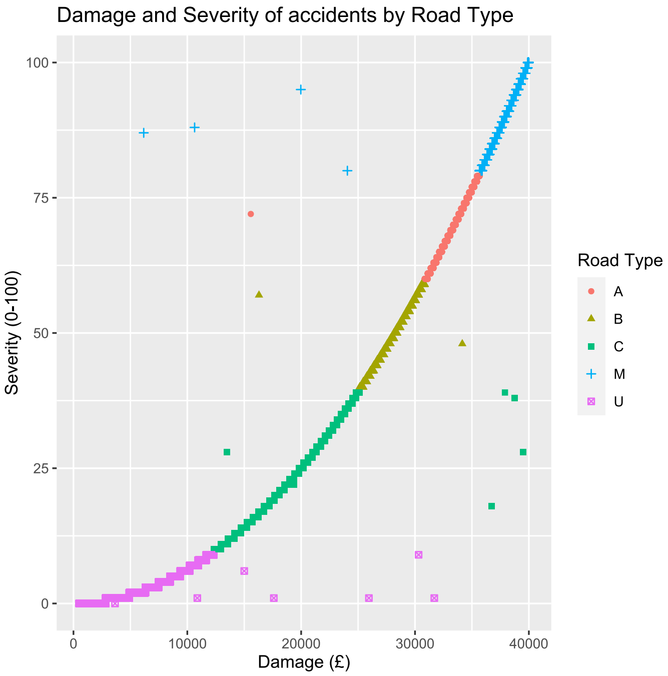

Fig. 3. Scatterplot showing positive correlation between Damage and Severity of accidents, which is also reflected in Road Types.

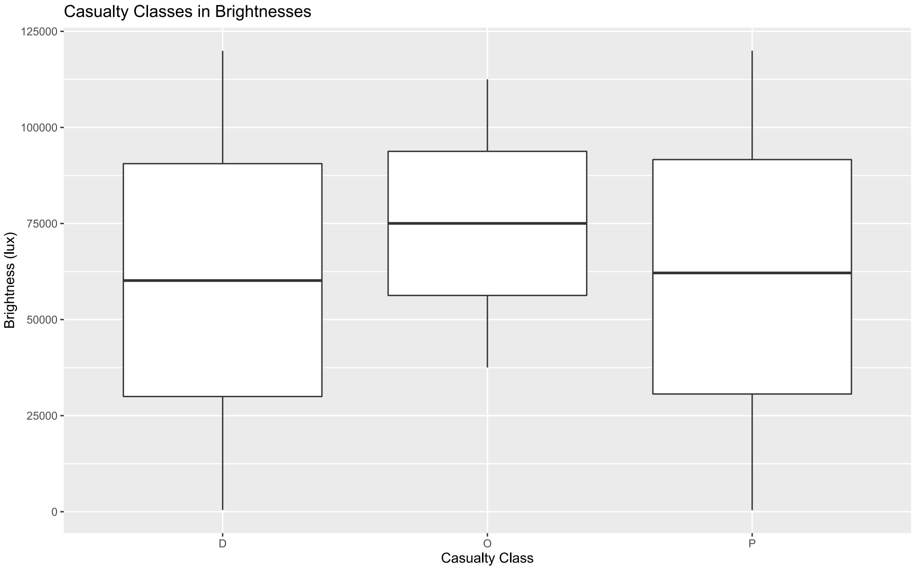

Fig. 4. Boxplot of casualty classes in accidents with brightness, showing ‘other’ class more in brighter environments.

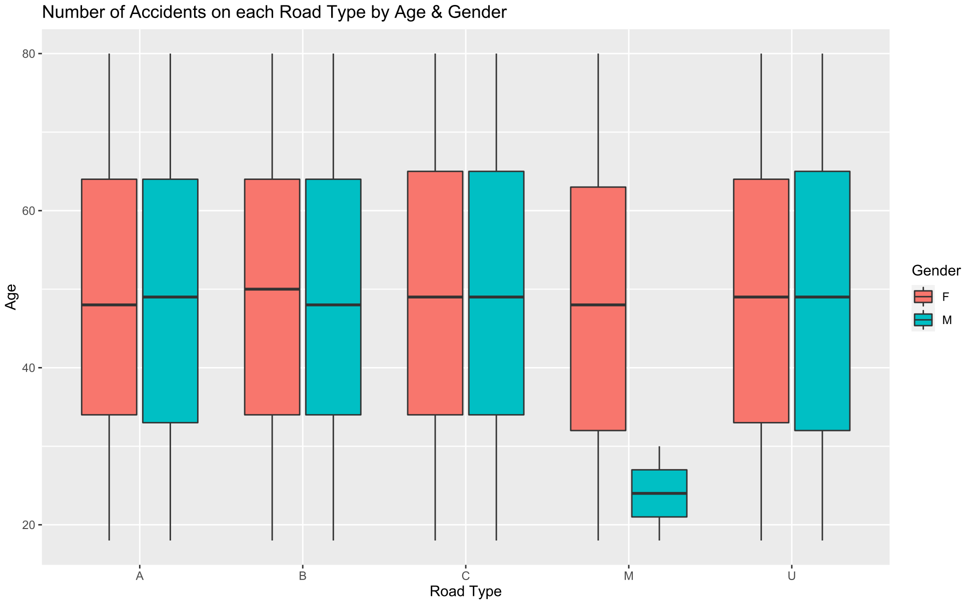

Fig. 5. Boxplot showing the accidents on each Road Type by Age and Gender of casualty, older males on motorways have no accidents

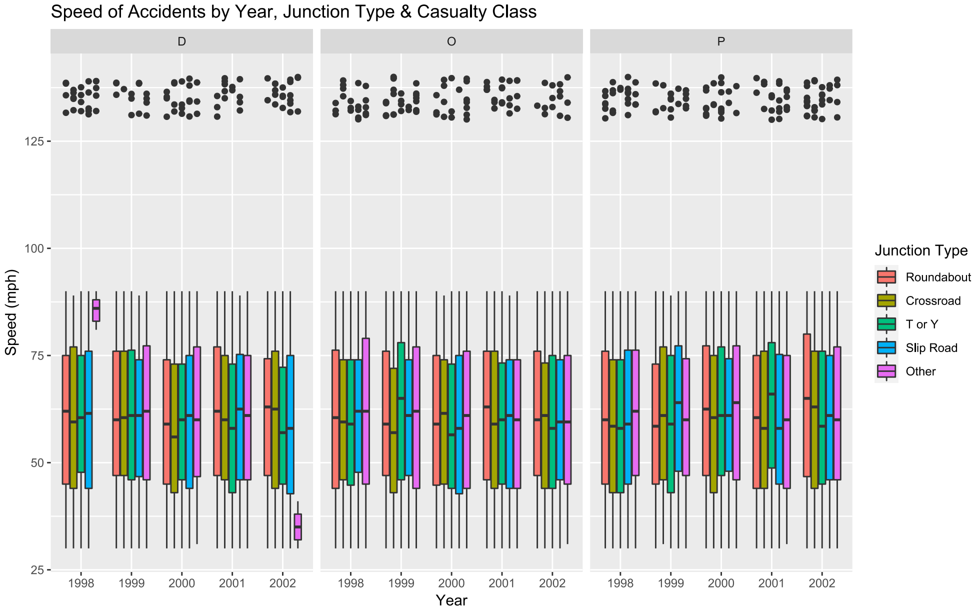

Fig. 6. Boxplot of speed of accidents by Year, Junction Type and Casualty Class, showing differences for Drivers at ‘Other’ junction in 1998 & 2002

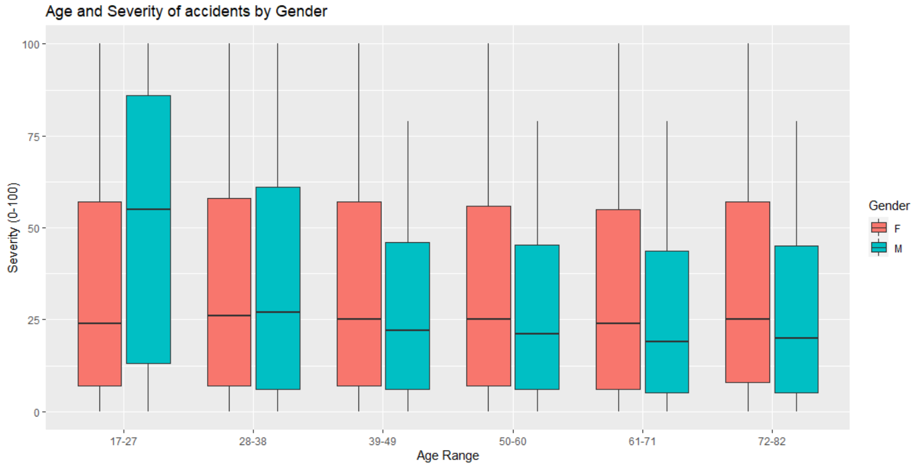

Fig. 7. Boxplot with Age Ranges against the Severity of accidents, juxtaposed with Gender, showing younger males generally are more severely affected.

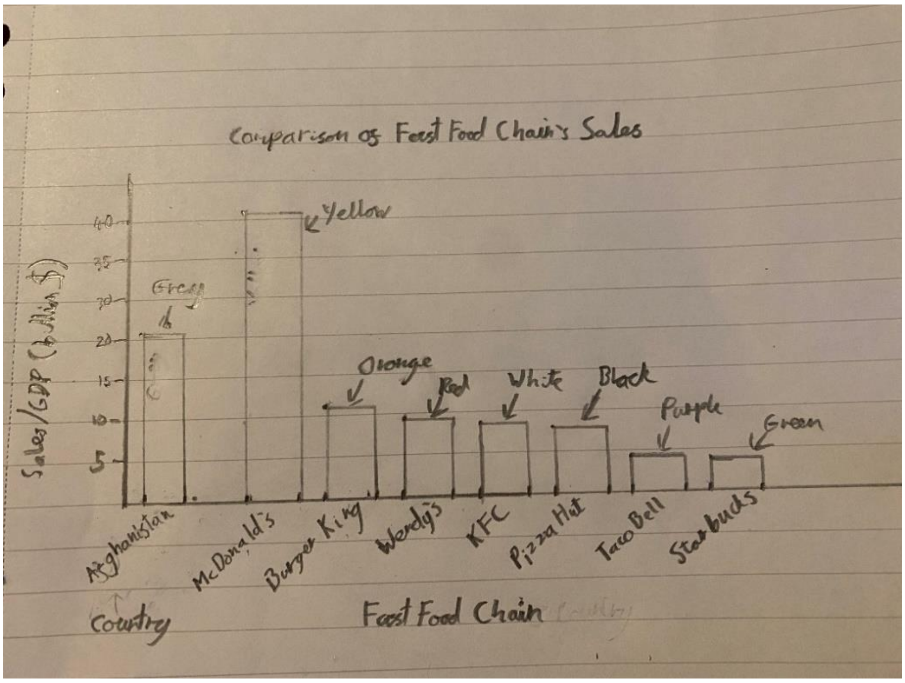

Fig. 8. Sketch of bar chart showing a comparison of fast-food chain sales, improvement to provided visualisation