Discussion for Traffic Accident Analytics Project

Analysis of Vehicle Dataset Processing

Significance of Clusters in Crash Analysis of Vehicle Groups

The k-means clustering analysis grouped vehicles into different clusters based on 5 features being: number of wheels, number of cylinders, seating capacity, tare weight, and number of occupants. I am thinking that each cluster most likely represents a category of vehicles with similar characters, which would also relate to their crash frequency.

One cluster we see in the plot is lightweight, low seating capacity vehicles, such as sedans or hatchbacks. This category of vehicles has fewer seats and lower tare weight. If the crash frequency is high in this group, it may be since these vehicles are more present in urban traffic, have less crash protection due to the small size, or younger driver demographic.

Another cluster I thought of looking at the plot could be mid-sized vehicles with higher seating capacity and weight, such as SUVs, Tilters, Tippers, Vans and more. These vehicles are most likely driven by families or in a more specialised field and are mostly used in less dense areas. If this cluster shows lower crash rates, it may be due to safer driving conditions, and less vehicles being present.

A third cluster we see in the plot is high seating capacity vehicles, more specifically buses. Their usage patterns and requirements might influence crash rates differently. For example, for buses to be driven professional drivers that have a bus license can only drive them, and buses have more strict safety inspections.

The data is likely clustered this way because attributes like weight and capacity correlate strongly with how and where vehicles are driven. Moreover, crash frequency patterns aren't just strictly based on the vehicle's attributes but also human behaviour and the environmental context which can be somewhat captured through the vehicle's attributes. This explains the different clustering of vehicles based on their attributes.

Limitations of Clustering for Vehicle Safety Insights

Clustering acts as a tool for identifying patterns in the vehicle crash data, but its usability for determining the safest vehicles for new buyers must be considered carefully with other factors taken into account.

The clusters in the plot highlight groups of vehicles with shared characteristics and similar crash frequencies, which offers insights into which types of vehicles tend to be involved in more or fewer crashes. For example, if a cluster containing heavier vehicles with more seating capacity shows lower crash counts, a new vehicle buyer might think that these vehicles are safer, and this could guide new vehicle buyers towards vehicles with similar specifications.

However, it's important to realise that clustering reflects correlation and not causation. A low crash frequency in a group may result from many factors unrelated to the vehicles safety features such as: driver information, driving behaviour, where the vehicle is used. On the other

hand, some vehicles might appear in a high-crash cluster simply because they are more common or driven more frequently in high-risk areas.

Moreover, the features used in clustering: tare weight, number of cylinders, etc, do not directly measure the safety of those vehicles. For clustering to be more usable for new vehicle buyers it should be combined with crash severity data, driver behaviour information, and vehicle safety ratings. Even then, it would still just serve as an aid rather than a proper guide.

Analysis of Accident Dataset Processing

Text Mining Insights: Accident Types Across Different Times of Day

From the TIME_OF_DAY pie charts, we can observe trends and other key terms predominant at certain times of day to make hypotheses as to what nature of accidents occur at various times of the day.

Firstly, we observe that intersections is one of the ten most frequent terms in the morning and afternoon, whilst there is a decreasing trend throughout the day in the frequency of the terms ‘lane’ and ‘rear’ from the top ten terms for a given time of day. A potential explanation for this could be a greater volume of cars on the roads and speeding during rush hours in mornings and afternoons when parents rush to drop and pick children to and from school and other extracurricular activities, and adults need to commute to and from workplaces promptly. Consequently, traffic and speeding during mornings and afternoons when compared to nighttime may lead to more frequent crashes at intersections, more sideswiping because of quick lane changing and the sheer number of cars on the roads. Furthermore, speeding vehicles in a rush may also be more prone to rear-end collisions.

Moreover, the terms ‘bend’ and ‘ped’ only appear in the ten most frequent terms in key words from accident descriptions at late night (from midnight to 5:59am) and ‘object’ appears only in evening and late night. This may be explained by sunlight fading during the evening and for most of this time, drivers often face visibility issues because of insufficient light. Further, drivers may also face excessive fatigue after a long and difficult day. As a result, there may be greater collisions with pedestrians and various other objects around the road. There is also an increasing trend in the term ‘carriageway’ throughout the day, suggesting cars may swerve off roads potentially because of a delayed reaction to lane markings and barriers because of low-light conditions and fatigue. Visibility and drowsiness may also explain why ‘parked’ also increases in frequency out of top ten words in accident-related descriptions. Hence, during the late night, drivers are more prone to different types of accidents, specifically collisions with objects and pedestrians, possibly at bends, veering off the carriageway.

Another key observation is that the terms ‘right’ and ‘left’ are in top ten most frequent terms for each time of day, with ‘right’ usually being more frequent than ‘left’, except for late night. Furthermore, ‘right’ was the most frequent accident-related term during mornings and afternoons. A possible explanation for this is that in Victoria, right lanes are overtaking lanes, where drivers are often speeding beyond the legal speed limit, potentially resulting in greater cases of sideswiping. Earlier discussion of greater volume of cars during the day may also explain why ‘right’ is the most frequent accident-related term during the day (mornings and afternoons).

Accident Frequency and Patterns by Day and Time

From the stacked bar chart 'Frequency of Accidents by Day', we can observe differences in accident counts by day and time of day and can try to provide possible explanations and conjecture about the nature of accidents occurring. Friday has the highest aggregate collisions compared to Mondays and Sundays, with Sundays having the least accident count.

Firstly, we observe a large frequency of accidents occurring on Friday afternoons into evenings and on Sunday's 'late night' (midnight to 5:59am), where Sunday late night is effectively the very early hours of Sunday. After a long and tiring week, people are quite fatigued and attend parties on Friday afternoons into the next morning, where they may consume large volumes of alcohol. Consequently, there may be greater incidents of drink driving or drowsiness. Further, the use of recreational drugs on weekend evenings in social settings can impair drivers' reaction times and coordination, compared to drivers on evenings from Sunday to Thursday, because of work-related commitments. Such usage can cause several types of accidents such as rear-end collisions or swerving off the road and colliding with barriers. Moreover, youths may choose to go out socially to recreational places from Friday afternoons onwards. Younger drivers are known to be riskier drivers, maybe avoiding certain precautions like speed limits, that makes them more prone to high-speed crashes especially on freeways. Hence, our analysis of the greater accident counts during certain times of the day compared to other weekdays has uncovered potential nature of accidents such as crashes because of reduced reaction time by speeding or consumption of drugs.

It is evident from the stacked bar chart that cumulatively, Monday and Friday mornings and afternoons have a greater number of accidents, when compared to Sunday. A possible explanation for this could be that weekday mornings and afternoons are often peak hour traffic, commuting to and from workplaces and schools. This may result in crashes with pedestrians in school zones, or collisions in high traffic areas or even sideswiping by frantic lane changing. Tiredness because of not having enough sleep the night before could also be a major factor, possibly leading to accidents due to reduced concentration, such as lane departures. The multi-faceted effect of fatigue and peak-hour traffic during the day, along with the suboptimal decision-making during recreational activities in the evenings, also explains our earlier observation of greater collisions on Fridays, compared to Mondays and Sundays.

Analysis of Person Dataset Processing

Risky Seatbelt Behaviour by Age Group

The bar chart (Seatbelt Use across Age Groups) displays a clustered bar chart of the 5 age groups (and the 'unknown' category) against seatbelt distribution. The distribution is taken from the 'person' dataset in the 'Dataset Overview Report' of Victorian road crashes/accidents. The two clusters are either 'Seat Belt Worn' (in orange) or 'Seatbelt Not Worn' (in blue).

From the results of the bar chart, the age groups 17-25 and 40-64 years have the highest bars, hence, the highest values of not wearing a seatbelt. Hence, they demonstrate the riskiest behaviour regarding seatbelt use from the dataset.

For the 17–25 age group, this trend may reflect risk-taking tendencies, inexperience, or a sense of invulnerability that younger drivers and passengers often have. They may underestimate the dangers of not wearing a seatbelt, or overestimate their ability to avoid serious accidents

This suggests that despite possibly having more driving experience, individuals aged 40–64 may be more complacent or less compliant with seatbelt regulations (overconfidence), thereby demonstrating the riskiest behaviour in terms of seatbelt non-use compared to other age groups. This non-use is significant to assessing 'risky' behaviour, since seatbelt use is a key factor in reducing injury severity and fatalities in road accidents.

Seatbelt Compliance by Vehicle Occupant Position

The first plot of pie charts compares 'Seatbelt Use of Drivers,' which shows 1.46% of Drivers do not wear their seatbelts, and 'Seatbelt Use of Passengers,' which records 2.77% of passengers not wearing their seatbelts.

The second plot depicts pie charts of 'Seatbelt Use of front-seat passengers' and 'Seatbelt Use of rear-seat passengers.' So, of the 2.77% of overall passengers not wearing their seatbelt, 1.20% are front-seat passengers and 4.53% are rear-seat passengers.

Since the plots of pie charts depict structured data (percentages), which are normalized by vehicle user group size, they can be compared as they are. Hence, 1.46% of drivers do not wear their seatbelt, 1.20% of front-seat passengers and 4.53% of rear-seat passengers. So, the group exhibiting the riskiest behaviour (not wearing seatbelts) is rear-seat passengers, with a result of 4.53%.