COMP20008 Assignment 1: Analytical Discussion

Ed Xiao	Student ID: 1625001
Hao Wang	Student ID: 1548969
Haochen Tang	Student ID: 1335862
Jack Puccini	Student ID: 1733518

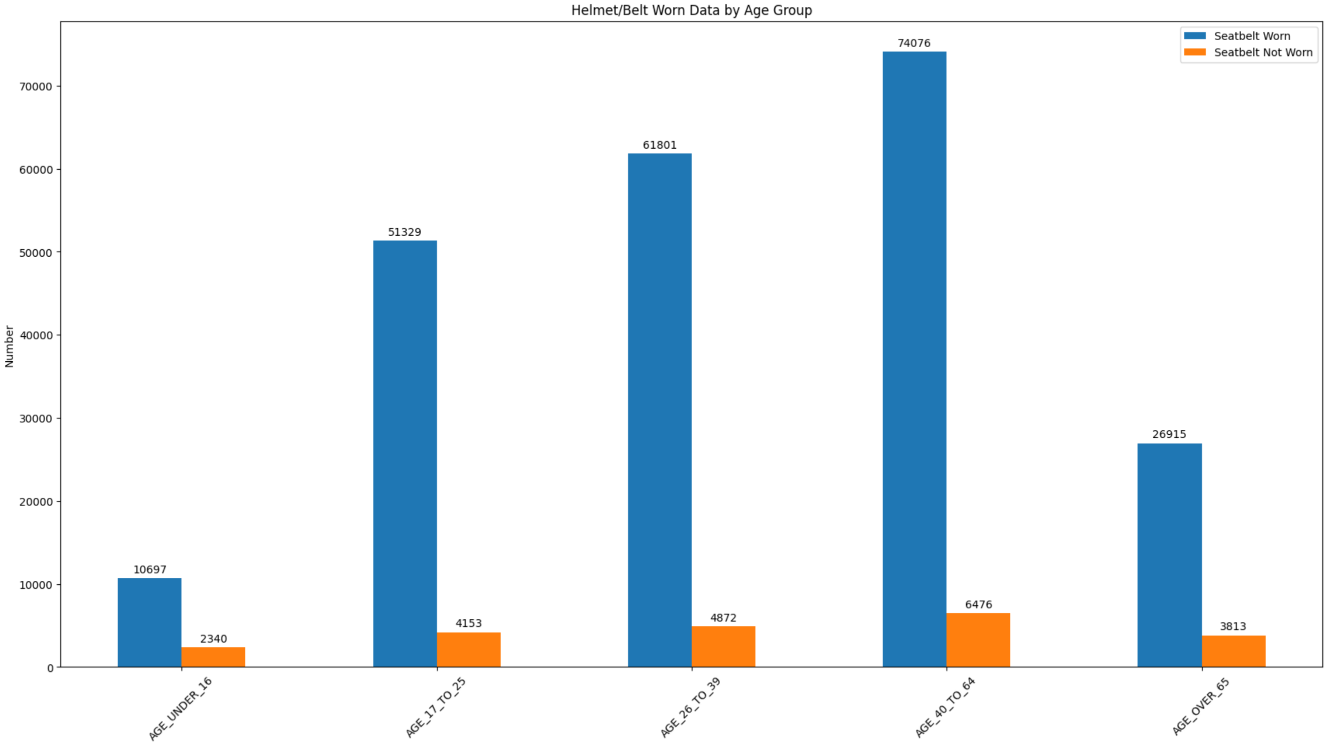
April 16, 2025

Task 1

1. Age Group Analysis

Age Group Analysis - specifically, answer the question "Which age group shows the most risky behaviour regarding seatbelt use?" Justify your answer using the bar chart results.

Answer

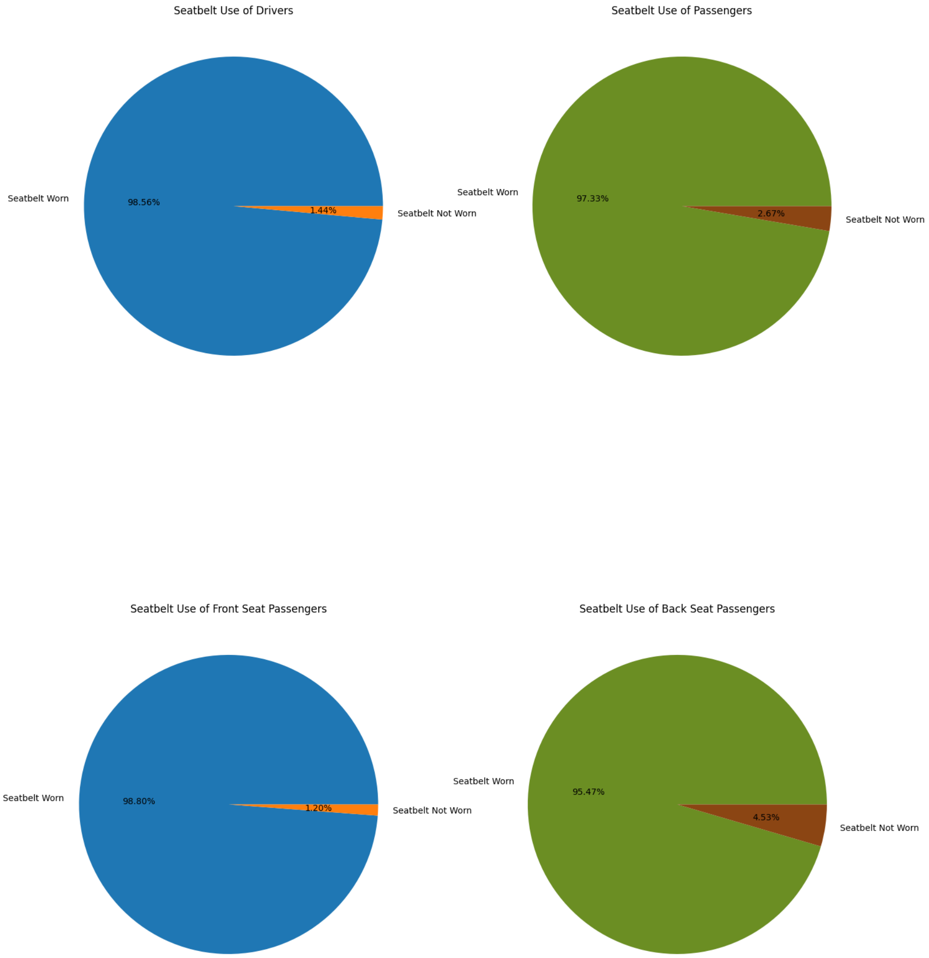


Those aged under 16 show the riskiest behaviour regarding seatbelt use as they have the highest percentage of individuals who do not wear a seatbelt (17.95% calculated from bar chart values). However, the group with the highest overall risk are aged between 40–64, having the highest number of individuals not wearing seatbelts (6476). 40–64-year-olds not wearing seatbelts also make up the greatest proportion of accidents involving unworn seatbelts (29.91%).

2. Vehicle User Group Analysis

Vehicle User Group Analysis - specifically, answer the question ”Which vehicle user group (of drivers, front-seat passengers, or rear-seat passengers) exhibits the most risky behaviour?” Justify your answer using the pie charts.

Answer

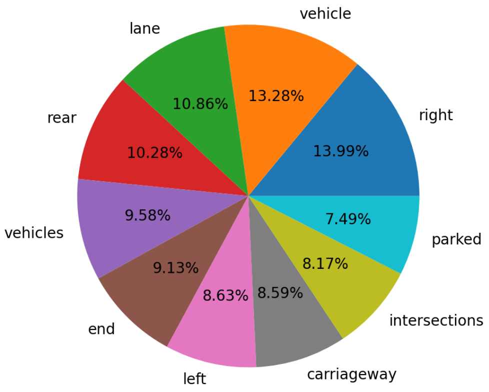


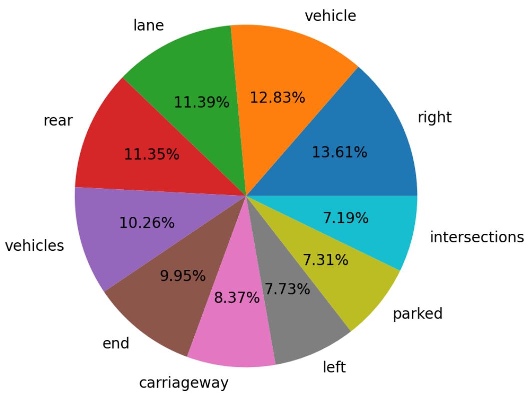
Backseat passengers exhibit the riskiest behaviour regarding seatbelt use. 4.53% of this group were not wearing seatbelts in the accidents, considerably greater than the 1.2% found in front seat passengers and 1.44% in drivers.

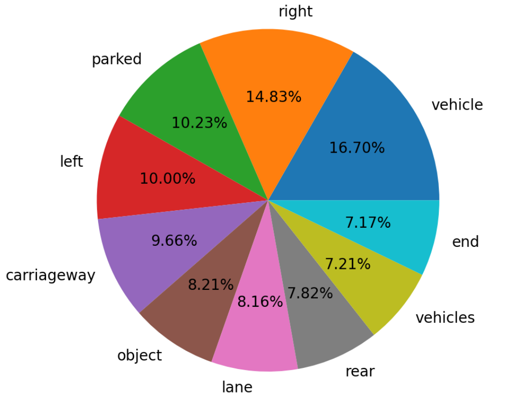
Task 2

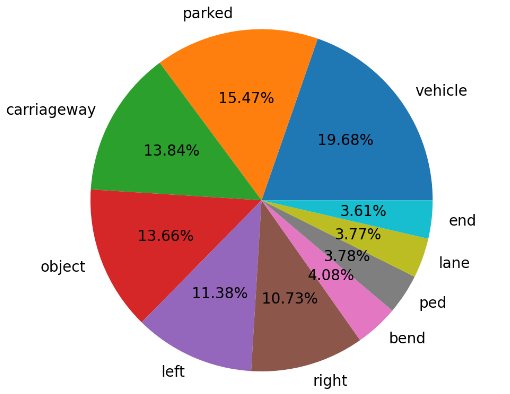
1.

The TIME_OF_DAY pie charts and discuss any patterns or trends you observe and specifically, answer the question "What do the results tell you about the nature of accidents at different times of the day?"


Top 10 words related to Morning accidents


Top 10 words related to Afternoon accidents


Top 10 words related to Evening accidents


Top 10 words related to Late Night accidents

Morning: The frequency of key word 'rear' is significantly higher than Evening and Late night, which is may due to Morning is the peak commute time, citizens can be in a hurry and the average vehicle speed is high. Majority of the accidents are related to lane changing and unproper driving distances.

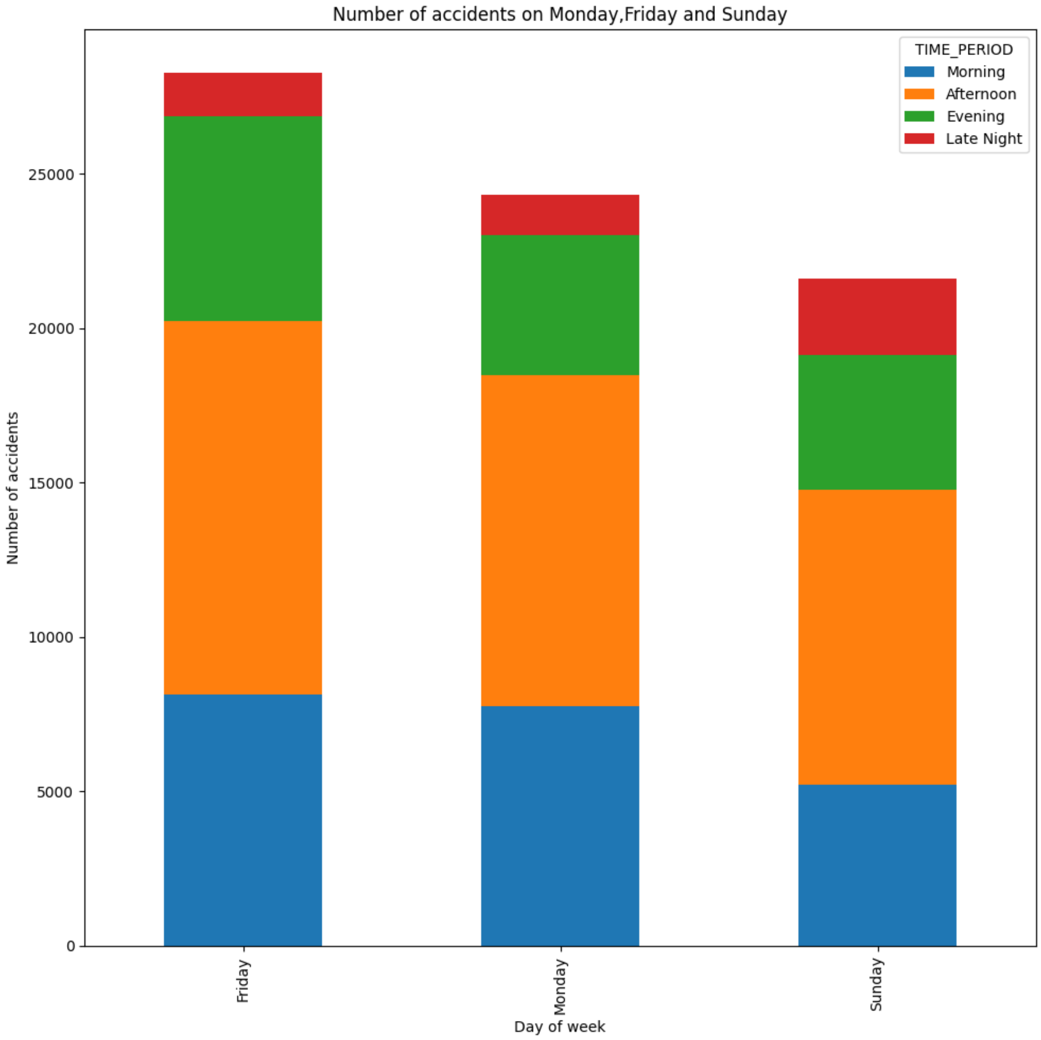
Afternoon: The situation is similar to Morning with more common rear, reflecting speed and space controlling issue. At the same time lane handling accident is still the mainstream.

Evening: The section of key word 'vehicle' is relatively larger than that of Morning and Afternoon, as well as a sharp increase of the percentage of 'right' and 'left' could be witnessed. For pedestrians, accidents are most likely in the Evening. The reason generated the changes may be the fade of sunlight.

Late Night: Proportion of key word 'vehicle' is the highest among the four time periods at roughly a quarter and a new key word 'bend' is contained in this chart, which may be caused by tired driving or drunk driving.

2.

The variation between Mondays, Fridays and Sundays in accidents at different TIME_OF_DAY categories - specifically, answer the question "What do these results tell you about the nature of accidents occurring on these days?"



Friday: Afternoon accidents constitute the majority of the accidents, and the ratio of Evening accidents is also relatively high which is the most prominent of the three days. The reason of this situation may be that Friday is the last day of work week, individuals tend to conduct social activities after work, therefore, the Afternoon and Evening traffic is heavy.

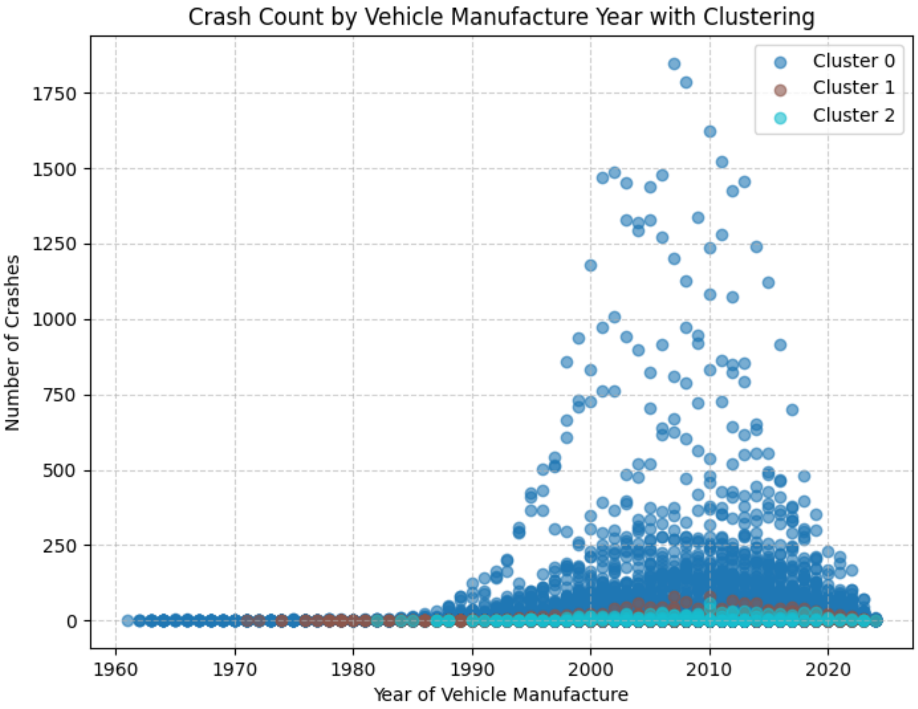
Monday: Total amount of accident is slightly smaller than Friday, but the data distribution is similar to Friday that most of the accidents happens during the commute.

Sunday: Sunday has the smallest accident numbers which is generally contributed by the least Morning accidents. However the volume of Late Night accidents is the most dramatical within the three days, which could be explained by drunk driving or tired driving.

Task 3

1.

Use the produced plots and output from the prior steps to identify what each cluster might mean. Suggest why the data may have clustered this way.



We can identify the meaning of each cluster by looking at the rows from the cluster csv files. For example, in task3_3_cluster0.csv, we see that all vehicles are Sedans from manufacturers like Toyota and Holden, with features such as 4 wheels, 5 seats, and a tare weight around 1300kg. Similarly, task3_3_cluster1.csv consists of P MVR vehicles from the maker KenWth, all with 6 wheels, 2 seats, and tare weights exceeding 9000kg. Lastly, task3_3_cluster2.csv features Buses from Volvo, Scania, and Merc B, with very high seating capacities (around 43–46) and tare weights above 10,000kg. These patterns suggest that each cluster groups vehicles by body style and structural characteristics.

2.

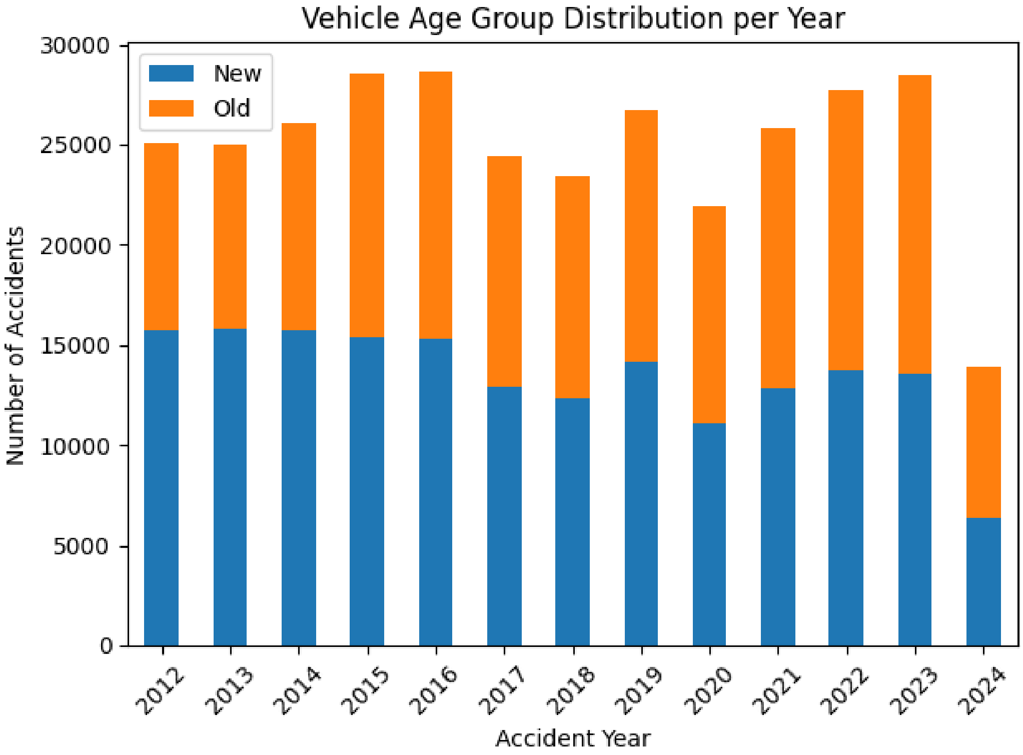
Comment on the usability of the clustering for determining the safer group of vehicles for new vehicle buyers.

The results of this clustering are not directly useful for car purchases, as buyers typically choose vehicles based on their needs - for example, Sedans for households, P MVRs for shipping, and Buses for transporting people - rather than based on crash statistics. Also importantly, the high crash counts seen in cluster0.csv for Sedans likely reflect their popularity rather than a lack of safety. These vehicles are more common on the road, so naturally appear more often in crash data. A more accurate measure of safety would perhaps consider proportional figures, such as crash rates relative to the number of vehicles in use. While clustering reveals useful patterns across vehicle types, it offers limited value for buyers seeking safer cars.

Task 4

1

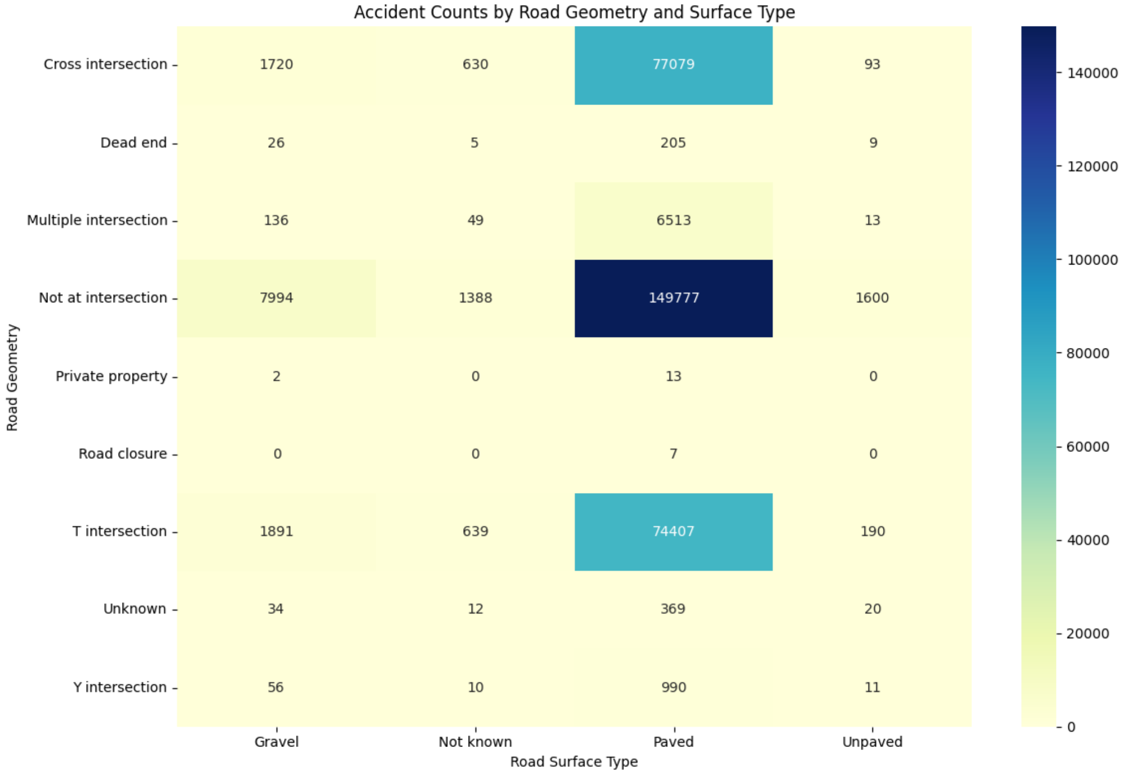
Do older vehicles tend to be involved in more accidents compared to newer vehicles? Analyse the trend over the years using the results from the bar chart.



Referencing the bar chart, from 2012 to around 2018, newer vehicles consistently accounted for a greater number of accidents than older vehicles. However, starting around 2019, the trend shifts, and from 2022 to 2024, older vehicles appear to be involved in as many or more accidents than newer ones. It is worth noting that the drop in counts for 2024 might reflect incomplete data collection due to being partial way through the year.

2

Which intersection type is more accident-prone: Cross intersections or T intersections? Justify your answer using the data from the heatmap.



Referencing the heatmap, we see that the total accident counts across all T intersections is 77,127, whereas for Cross intersections it is 79,522. Therefore, we can potentially conclude that cross intersections may be more accident-prone, due to their higher accident counts. That being said, the difference between the two is relatively small, and further analysis would be required to determine whether this difference is statistically significant.