Investigating the Impact of Vehicle Characteristics on Accident Severity and Identifying High-Risk Vehicle Profiles
W10G07

Group members:

- Kakanang Kullatee, 1649978, kkullatee@student.unimelb.edu.au
- Qingyue Zhao, 1449162, qingzhao@student.unimelb.edu.au
- Yujie Ren, 1565480, yujier1@student.unimelb.edu.au
- KEMAN LI, 1421046, kemanl@student.unimelb.edu.au

Research Question

" How do vehicle characteristics impact the severity of accidents and what are the possible high risk vehicle profiles? "

Executive Summary

This report investigates how specific vehicle characteristics impact the severity of road accidents across Victoria. By integrating data from *filtered_vehicle.csv* anand *accient.csv*, we used a combination of data cleaning, correlation analysis, and both supervised and unsupervised machine learning techniques to uncover patterns associated with high-risk vehicle profiles The analysis found that tare weight and type of vehicle are strongly linked to accident severity. It was found that heavy-duty vehicles such as prime movers are often associated with fatal outcomes, however there is less data which suggest they may be less frequent. On the other hand, light passenger vehicles like sedans are involved in lower severity accidents, however given the large amount of data, they may be more frequent. This suggests different types of vehicles present different forms of risk. However, the imbalance of data limits the reliability of these interpretations.

Introduction

Understanding the factors contributing to injury severity in road accidents is essential for improving vehicle safety standards and policies. This report investigates how specific vehicle characteristics influence the severity of injuries sustained in road accidents by integrating detailed vehicle characteristic data from filtered_vehicle.csv with accident severity data from the victoria road crash data and timing information from accident.csv. Furthermore the report will analyse and seek to uncover structural patterns associated with high-risk vehicle profiles.

Methodology

- Data cleaning & integration: merge vehicle characteristics and severity, outlier detection using five number summary. Column-wise deletion and row-wise fill in for missing data.
- Feature engineering: create accident severity index, upsampling for unbalanced labels for supervised learning models training.
- Feature selection & correlation analysis: mutual information calculations through entropy and feature importance via random forest model.
- Encoding: convert categorical data into numerical using ordinal encoding for random forest model and one hot encoding for mutual information calculation.
- Supervised learning & prediction model: tree based models include decision tree and random forest. Other regression models include the logistic regression model as a comparison group.
- Unsupervised learning & Data clustering: clustering method k-means , elbow method to find k. Use PCA to reduce the dimension.
- Visualization: bar chart for top data, box plot for outliers, histogram for discrete continuous variables, heatmap for confusion matrix and cluster, shallow decision tree visualisation.

Data Exploration and Analysis

1. Defining Vehicle Characteristic

To understand the relationship between vehicle characteristics and accident severity, we began by identifying relevant features from the filtered_vehicle.csv file. To ensure objectivity in selecting which data qualifies as vehicle characteristics, we conducted preliminary research to better understand the dataset's context. According to a publication by the Australian government[1], vehicle characteristics include fuel type, average age, cylinders and tare weight. However, to broaden our analysis and avoid excluding relevant variables, we chose our vehicle characteristic based on five categories. First, physical structure and design features include VEHICLE_TYPE, VEHICLE_BODY_STYLE, SEATING_CAPACITY, and CONSTRUCTION_TYPE related to the form and configuration of the vehicle. Second, mass and size attributes, including TARE_WEIGHT, VEHICLE_WEIGHT, and CARRY_CAPACCITY. Third, mechanical and engine specifications: such as NO_OF_CYLINDERS, CUBIC_CAPACITY, VEHICLE_POWER, and FUEL_TYPE. Fourth, identification and manufacturing features such as VEHICLE_MAKE, VEHICLE_MODEL, and VEHICLE_YEAR_MANUF. Lastly, vehicle appearance and visibility features: VEHICLE_COLOUR_1 and VEHICLE_COLOUR_2. Together, these five categories encompass our definition of vehicle characteristics in this report.

2. Data Pre-Processing

To initiate the analysis, we extracted the vehicle characteristic from *filtered_vehicle.csv* and merged it with SEVERITY and ACCIDENT_DATE from *accident.csv*. We then began cleaning the dataset. The first step is assessing the completeness of each feature by calculating both the number and percentage of missing values. A visual summary of top 10 variables with the highest percentage of missing data across features is presented in Figure 1 below.

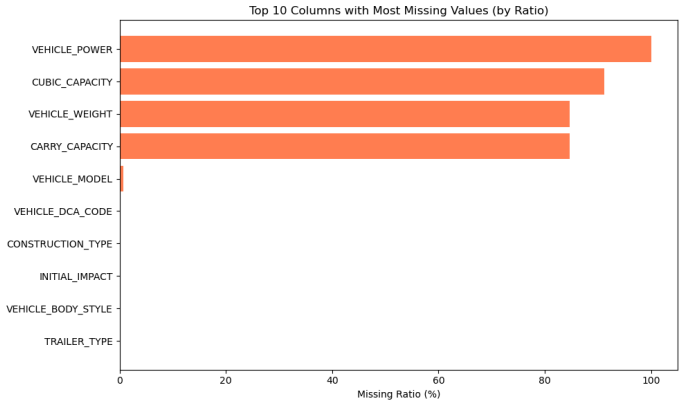


Figure 1. Top 10 variables with the highest proportion of missing data

Features with more than 80% missing data were removed from the dataset. Such a high proportion of missing data could due to limited recording, modelled directly with these data may be non-reliable, which may lead to biased predictions. Furthermore, features that exhibited a constant value across all records were removed, as they provided no meaningful information for the analysis. The remaining features were retained and subjected to imputation. For categorical variables, missing values were filled using the mode after removing punctuation and stopwords to ensure the validity and consistency of the entries. For

[1] Bureau of Infrastructure and Transport Research Economics. (2021, October 9). *Australia's light vehicle fleet - some insights*.

numerical features, missing values were imputed using the median, and a five-number summary was employed to assess and minimize the impact of outliers and skewed distributions. In addition, for features with less than 20% of missing data, we create a category for unknown data.

Since the *filtered_vehicle.csv* includes the manufacturing year of each vehicle, but the accident date varies across records, we could not rely solely on the manufacturing year to determine the vehicle's age. Therefore, we extracted the year of the accident and subtracted the year manufactured to obtain the age of vehicle in a new AGE, then removed the VEHICLE_YEAR_MANUF and ACCIDENT_DATE columns.

As a result, our final set of data, *data_cleaned.csv*, include: VEHICLE_MAKE, VEHICLE_TYPE, FUEL_TYPE, NO_OF_CYLINDER, SEATING_CAPACITY, TARE_WEIGHT, VEHICLE_BODY_STYLE, VEHICLE_COLOUR_1, AGE and SEVERITY.

3. Correlation Analysis

Now, we proceeded to explore potential relationships between these characteristics and accident severity. Firstly, we begin by plotting the distribution of the target variable SEVERITY using our cleaned dataset *data_cleaned.csv*. The result reveals a significant class imbalance — specifically, the "Fatal" category is heavily underrepresented, while severity level 4 ("None") contains only four samples.

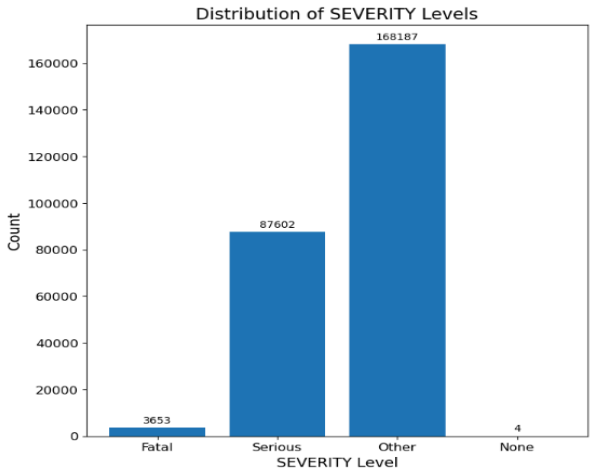


Figure 2. Distribution plot of severity level

As part of our analysis, we created a custom classification standard for severity, focusing on three levels: *Fatal (1)*, *Serious (2)*, and *Other (3)*. Since *non-injuries (4)* fall outside the scope of our research on injury-related outcomes and are also too rare and extreme (only 4 cases) to provide meaningful insight as shown on *Figure 2*, we excluded it from all subsequent analysis and focused only on severity levels 1–3.

Next, we extract all continuous variables, AGE and TARE_WEIGHT to explore whether linear relationships are present using Pearson correlations.

Feature	Pearson r	p-value
TARE WEIGHT	-0.0429	0
AGE	-0.0298	0

Table 1. Pearson correlation result for continuous variable vs severity level

The results show that although both variables yield statistically significant results p-value = 0 and the Pearson correlation coefficients with accident severity are extremely weak, which is -0.0429 for TARE_WEIGHT and -0.0298 for AGE. This indicates that while there are slight differences in mean values of these variables across severity categories, there is no meaningful linear relationship between either feature and severity level. However, there could still be non-linear relationships between these features and accident severity.

Following the analysis of severity distribution and linear relationship analysis, we computed Mutual Information (MI) to explore any relationship between each vehicle-related feature and the severity level. To capture the relationship of the continuous variables we discretize AGE into AGE_CATEGORY and TARE_WEIGHT into WEIGHT_CATEGORY before calculating MI. We also applied ordinal encoding to categorical variables to convert into sparse matrices to avoid any artificial order.

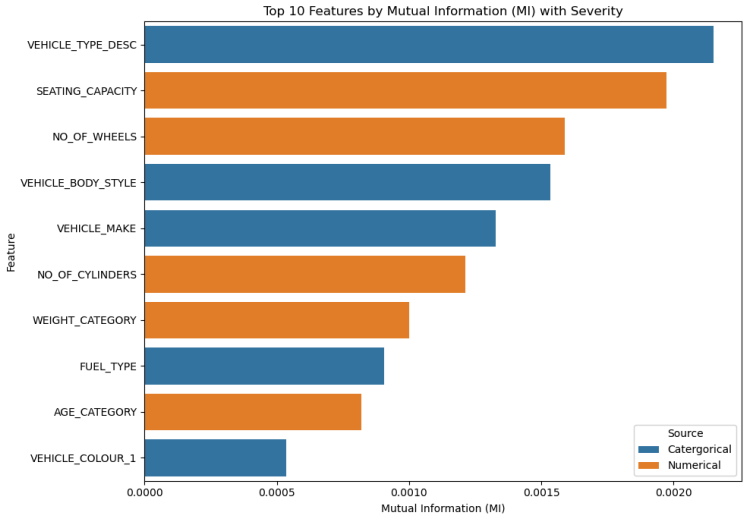


Figure 3. Top 10 Feature with highest Mutual Information

As shown on *figure 3*, the VEHICLE_TYPE, SEATING_CAPACITY and NO_OF_CYLINDER showed the highest MI value suggesting it may be an important predictor and have a strong correlation to severity. For categorical data, although they showed some correlations, they are weaker than numerical data overall.

4. Supervised Learning

4.1 Feature Importance by Random Forest Classifier

However, the mutual information assumes features are independent. In order to investigate which combination of features could be strongly associated with the accident severity level, we trained the random forest classification models that could automatically select the most important features.
We visualised the top 10 most important features extracted by the Random Forest Classifier model in a bar chart as shown on *figure 4*, suggesting that TARE_WEIGHT was the most influential feature accounting for an importance score of over 0.4, which is significantly higher than the rest. This highlights a strong relationship between vehicle weight as a primary factor, with other features contributing in combination. As a result, we are further investigating the impact of TARE_WEIGHT these important features on accident severity.

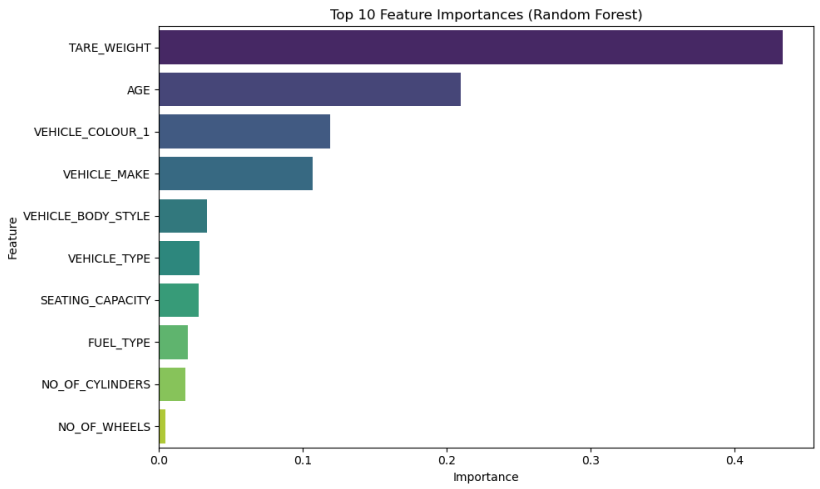


Figure 4. Top 10 most important features selected by Random Forest Classifier

4.2 Evaluation of Prediction Models

To evaluate the performance of our trained random forest model, we split our dataset into training and test sets and used the K-Fold cross validation test while ensuring that the model does not have the information of the importance ranking of each feature before testing to avoid any overfitting. In addition, the training set was upsampled to avoid underfitting and keep the test set unchanged to ensure that performance evaluation reflects real-world conditions and avoids artificial inflation of results due to duplicated data.

Zero-R Baseline						
	Fatal	**Serious**	**Other**	**Macro Avg**	**Weighted Avg**	**Accuracy**
Precision	0	0	0.65	0.22	0.42	
Recall	0	0	1	0.33	0.65	
F1-score	0	0	0.79	0.26	0.51	0.64
Support	1096	26281	50456	77833	77833	77833

Table 2. Zero-R Baseline

Random Forest Model						
	Fatal	**Serious**	**Other**	**Macro Avg**	**Weighted Avg**	**Accuracy**
Precision	0.01	0.34	0.65	0.33	0.54	
Recall	0.02	0.4	0.58	0.33	0.51	
F1-score	0.02	0.37	0.61	0.33	0.52	0.51
Support	3653	87602	168187	259442	259442	259442

Table 3. Random Forest Model classification report

From Table 2 above, we observe that the Zero-R baseline achieved a higher overall accuracy of 64.83% by always predicting the majority class ("Other"). However, it completely failed to identify any Fatal or Serious cases, with precision, recall, and F1-score all at 0 for those classes. This highlights its inability to provide meaningful insights into high-risk accidents, despite its superficially high accuracy.

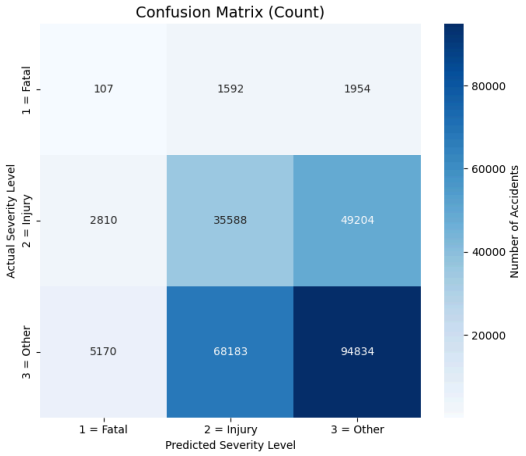
In contrast, the report of the Random Forest model shown on *Table 3*, evaluated using Stratified K-Fold Cross Validation to assess the model's generalization ability. The model produced an overall accuracy of 51.32%. However, it demonstrated a more balanced performance across severity levels, detecting a portion of Serious cases with f1-score at 37% and even some Fatal cases with f1-score at 2%. Although the recall for Fatal accidents remains low, this is expected due to the highly imbalanced datasets, as shown in *Figure 2 Distribution plot of severity level*.

Logistic Regression Model						
Fatal	**Fatal**	**Serous**	**Other**	**Macro Avg**	**Weight Avg**	**Accuracy**
Precision	0.02	0.34	0.67	0.34	0.55	
Recall	0.52	0.16	0.53	0.4	0.4	
F1-score	0.04	0.22	0.59	0.28	0.46	0.41
Support	1079	26051	50092	77222	77222	77222

Table 4. Logistic Regression Model classification report

We also evaluated a Logistic Regression model using the same setup. This model achieved a lower overall accuracy of 40.52%, and although it had a much higher recall for Fatal cases at 52% but lower precision at 2%. Furthermore, its recall for Serious accidents at 16% was significantly worse than the Random Forest at 40% indicating reduced reliability in detecting fatal accidents. The macro and weighted average F1-scores were also lower than those of the Random Forest model.

In conclusion, while Logistic Regression was able to capture more Fatal accidents in recall, this came at the low precision and overall reliability. In contrast, the Random Forest model provided more balanced, consistent, and interpretable predictions across all severity levels. Therefore, despite its modest accuracy, Random Forest is the preferred model in this context, particularly for real-world traffic accident risk profiling where identifying high-risk (Fatal or Serious) outcomes is of utmost importance.



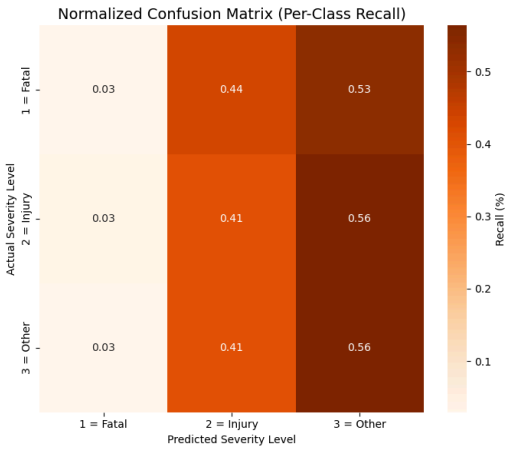


Figure 5. Confusion Matrix (Number of accidents) & *Figure 6. Normalised Confusion Matrix (Per-Class Recall)*

To further evaluate the predictive performance of the Random Forest model, we plot the confusion matrix in both absolute count and normalized to present the comparison of actual versus predicted classifications, revealing that a substantial number of Fatal cases (label 1) were misclassified as either Serious or Other. Specifically, as shown on *figure 5*, only 107 out of 3,653 Fatal accidents were correctly identified, with the majority being incorrectly predicted as Serious and Other accidents. A similar trend was observed for Serious cases, many of which were classified as Other, while predictions for the majority class "Other" were more accurate, as expected given its dominance in the dataset. The normalized confusion matrix on figure 6 provides a clearer picture of per-class recall by expressing each row as proportions. The recall for Fatal accidents was extremely low, at only 3%, reflecting the model's significant difficulty in identifying these high-risk events. In contrast, the recall for Serious accidents reached 41%, and the model correctly identified 56% of Other cases.

These results align with the earlier classification report and the confusion matrix emphasize the impact of class imbalance on model performance even though we use upsampling on the training set, but the generalization ability of the model is weak due to duplication and repeated samples in training set, which leads to an result that the model performs well for Other cases, it remains largely ineffective in detecting Fatal accidents.

4.3 Modeling Vehicle Weight Patterns Across Severity Levels Using Logistic Regression

By summarizing all of the above analysis, we observed that TARE_WEIGHT was ranked as a highly influential feature in predicting accident severity, as indicated by its strong importance score in the Random Forest model. However, its Pearson correlation with severity was extremely low, suggesting little to no linear relationship. This contrast prompted us to apply a logistic regression model, which has the best performance in predicting the fatal class to examine how change in TARE_WEIGHT influences the probability of each accident severity level.

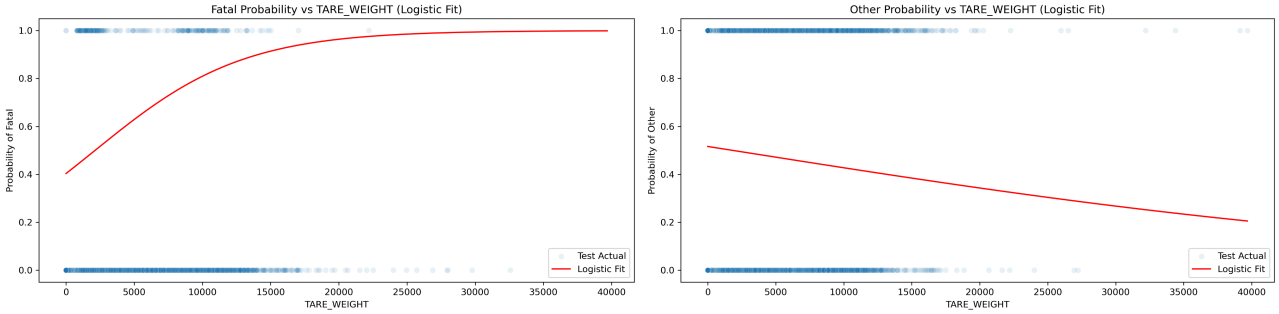


Figure 7. Fatal Probability vs Tare Weight & Figure 8. Other Probability vs Tare Weight

The logistic regression curves for both *Fatal* and *Other* accident outcomes consistently reflect a relationship between vehicle tare weight and accident severity. As shown on *figure 7 & 8*, the predicted probability of fatal accidents increases with tare weight, while the likelihood of Other severity outcomes decreases. While these trends are derived from simplified single-variable models with absence of interaction effects of other variables. As a result, these trends should not be interpreted as causal, they provide converging evidence that vehicle mass may significantly influence severity outcomes. Notably, the model's performance on predicting Other severity outcomes is relatively strong by achieving high precision and recall, this suggesting that the observed downward trend could be helpful for us to identify the risk profile of light and heavy vehicles.

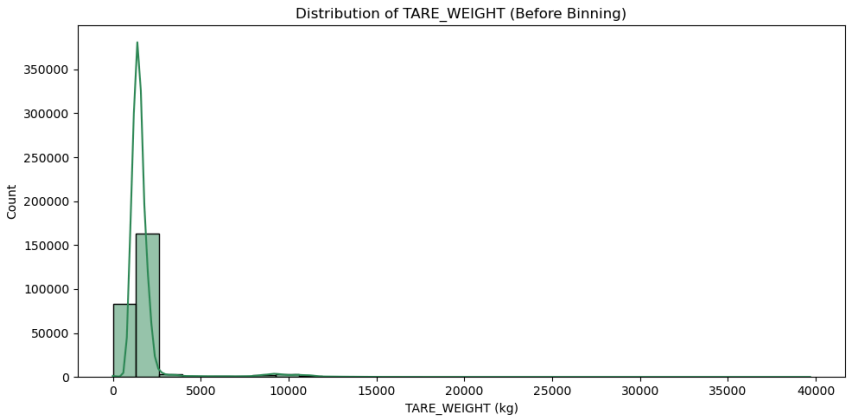


Figure 9. Histogram of vehicle weight distribution

Previously, as shown in *Figure 2 Distribution plot of severity level*, the number of Fatal accidents is extremely low compared to other severity levels. However, as illustrated in *Figure 9*, such heavy vehicles constitute only a small fraction of the entire dataset. This issue reveals a structural imbalance of our dataset, which significantly limits the model's ability to learn meaningful patterns specific to fatal cases and explain the result of the classification report tables above.

4.4 Vehicle Risk profile analysis using decision tree model

In addition to the random forest classification model analysis, a shallow decision tree model was trained and visualized to generate interpretable risk patterns across severity levels.

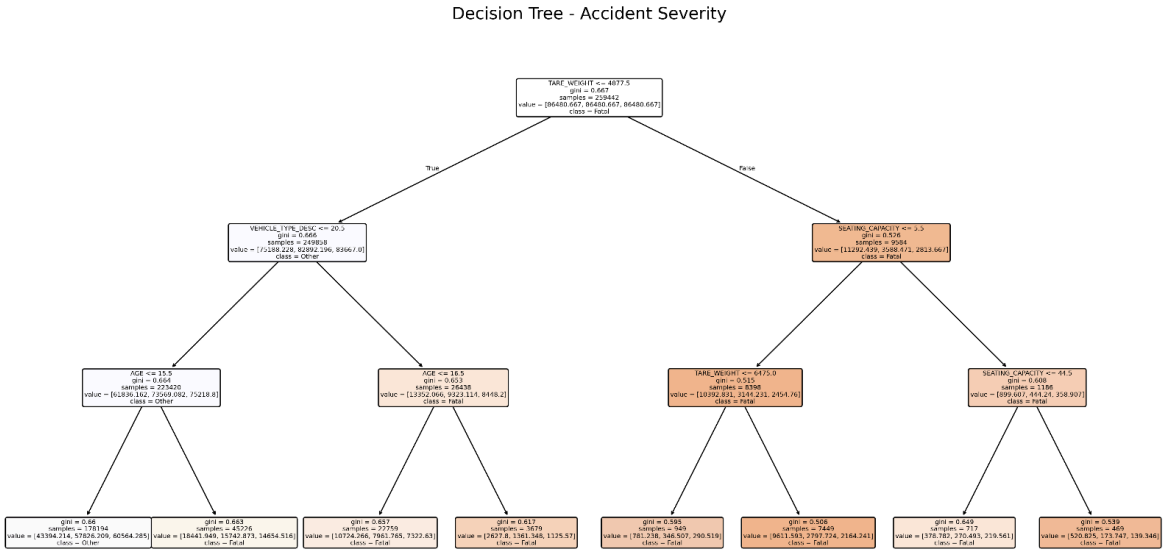


Figure 10. Visualization of Decision Tree model (Depth = 3)

As shown in Figure 10, TARE_WEIGHT again emerges as the root node, highlighting its dominant role in determining accident severity. The initial split at approximately 4677.5 kg effectively separates lighter vehicles from heavier ones. Vehicles with a tare weight above this threshold tend to follow decision paths that culminate in Fatal classifications, as shown by the orange colored nodes on the right-hand side of the tree. On the right subtree, the decision further refines high-risk profiles using SEATING_CAPACITY and

VEHICLE_TYPE_DESC. Specifically, vehicles with SEATING_CAPACITY ≤ 5.5 in combination with high tare weights, are often classified as fatal cases. In contrast, on the left side of the tree, it shows that AGE plays a more important role among lighter vehicles. Vehicles with AGE > 16.5 years are more frequently associated with fatal outcomes, even among those with lower tare weight.

However, these interpretable paths consistent with the earlier prediction results from logistic regression: heavier vehicles though sparse in the dataset may play a disproportionate role in fatal accidents. By summarizing the observation from supervised learning models, we could conclude a possible prediction that vehicles with high and low tare weight may have different risk patterns with different characteristics. Therefore, in the subsequent clustering analysis, we further examine the high-risk vehicle profiles using unsupervised learning models and compare the pattern against the predictions and correlation analysis.

5. Unsupervised Learning

5.1 Feature selection for KMeans Clustering

In order to explore the risk patterns of vehicles, we are going to use unsupervised model clustering. For high-dimensional data categorical variable such as VEHICLE_BODY_STYLE that have multiple unique values, hierarchical clustering requires recursive matrices and recursive merging, which is computationally intensive; DBSCAN may eliminates some noise data but is extremely sensitive to high dimensionality categorical data with high time complexity. Therefore, K-Means offers relatively low computational complexity and generates easily interpretable cluster centroids. Its scalability and adaptability to large datasets make it well-suited for the objectives of our investigation.

In order to explore the impact of vehicle feature combinations on severity level, we choose numerical variables include NO_OF_WHEELS, TARE_WEIGHT, SEATING_CAPACITY, NO_OF_CYLINDERS, AGE that are suitable for K-Means Euclidean distance calculations. In order to capture distinct structural and functional differences between vehicles, we aggregated vehicles by key categorical features according to the result of feature selection above, include VEHICLE_MAKE, VEHICLE_BODY_STYLE, VEHICLE_TYPE_DESC, FUEL_TYPE, VEHICLE_COLOUR_1. For each group, numerical attributes were averaged to form representative vehicle profiles, which improves clustering quality by focusing on structural characteristics and reduces redundancy and noise.

5.2 Elbow method find K

Firstly, we applied the Elbow method to determine the appropriate number of clusters (*k*), by plotting the within-cluster sum of squares (WCSS) against increasing values of *k*. As shown in the *Figure 11 & 12* below, the choice of all numerical features resulted in a non- smooth curve and non-compact points for each cluster group.

Although AGE is considered to be a vehicle characteristic and may be associated with severity according to our previous analyses, it can also be influenced by external factors such as maintenance condition or mileage and is not an inherent structural property of the vehicle. Since our clustering objective focuses on structural profiles, AGE was removed from the final clustering feature set due to its limited structural interpretability, however, it was still incorporated during the group-wise aggregation process alongside categorical features. This allows AGE to contribute to the representation of each vehicle group without distorting the clustering outcome.

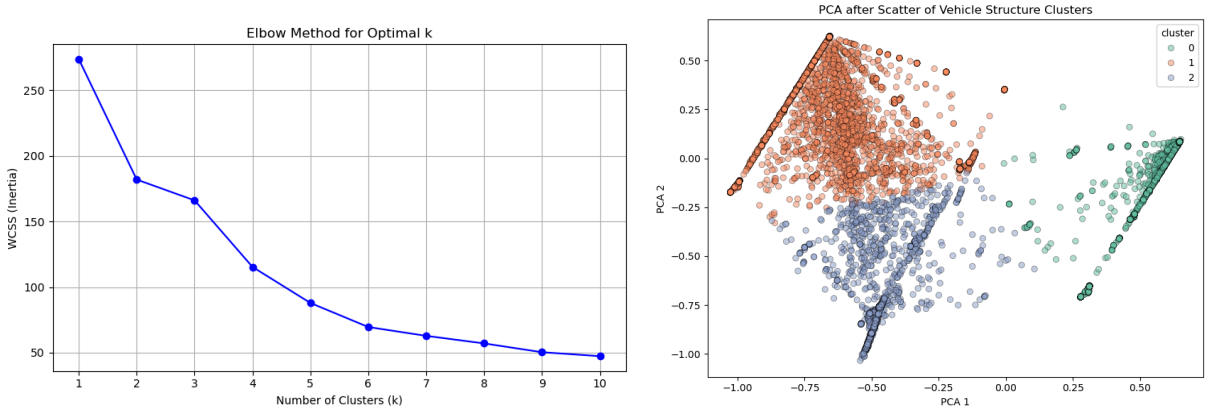


Figure 11. Elbow Method to Find k & Figure 12. PCA after cluster include vehicle age

Based on the Elbow plot shown in *Figure 13* below, the optimal number of clusters was determined to be $k = 3$, where the marginal reduction in WCSS becomes minimal. This number of clusters provides a good trade-off between capturing structural diversity and maintaining interpretability.

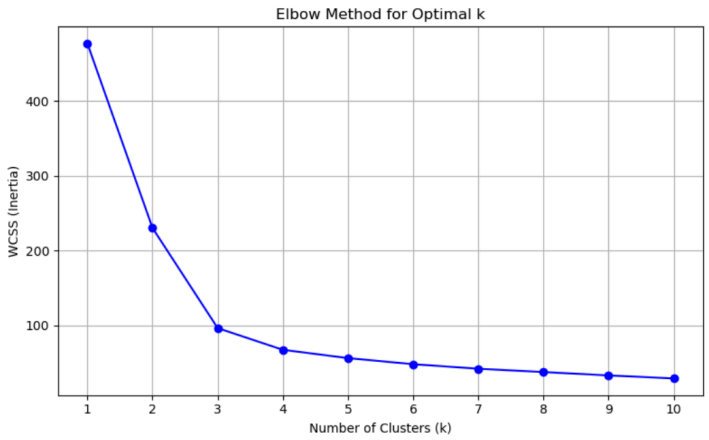


Figure 13. Elbow Method to Find k (exclude vehicle age)

5.3 Combination of Clustering and Dimensionality Reduction

We applied Principal Component Analysis (PCA) in combination with clustering to compress the high-dimensional data into two dimensions for visualization purposes. However, PCA was applied clustering instead of before. This decision was made due to several limitations of PCA. Specifically, PCA transforms the original features into composite components, which leads to a loss of interpretability. Particularly in our study, the meaning of structural vehicle attributes is important for risk profile analysis. Therefore, PCA is a linear method and may fail to capture nonlinear patterns in the data. To preserve the original structural semantics of vehicle features during clustering, we first performed clustering using the original features and then applied PCA solely for visualization.

The first two principal components (PCA 1 and PCA 2) explain 91.32% of the total variance in the dataset, with PCA 1 alone capturing 59.28% and PCA 2 capturing an additional 32.04%. This high cumulative variance indicates that the majority of structural variation among vehicle types can be effectively represented in just two dimensions.

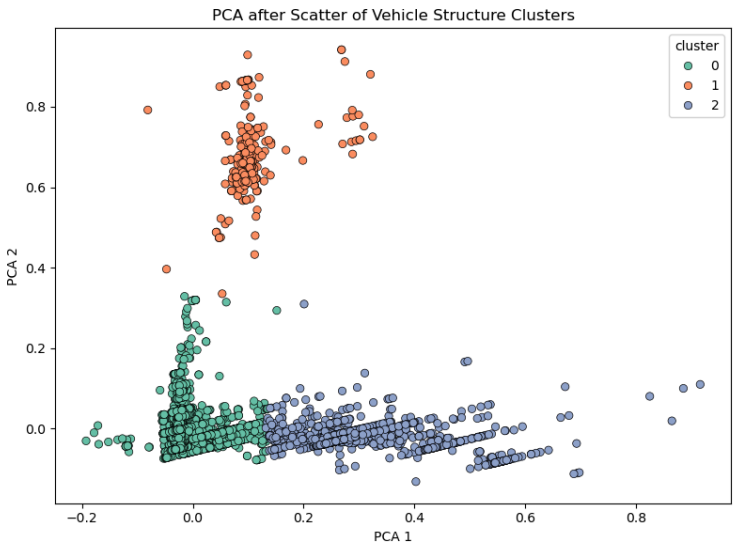


Figure 14. PCA after Cluster with Selected Features

The resulting scatter plot in *Figure 14* above shows that the clusters exhibit clear separation and minimal overlap, indicating that input features include tare weight, seating capacity, no of wheels and cylinders capture structural difference among vehicle types. Cluster 1(orange) appears tightly packed, suggesting high intra-cluster similarity, while Cluster 2 (blue) is more dispersed, possibly indicating internal structural variation among heavy vehicles. Overall, the pattern shows a relatively low intra-cluster distance and high inter-cluster distance compared to *Figure 12 PCA after cluster include vehicle age.*

5.4 Interpretation of Cluster Groups

To better understand the structure of each cluster, we examined both numerical and categorical feature centroids as shown on *Tables 5 & 6* below. For continuous numeric features such as TARE_WEIGHT ,we used mean, for discrete numeric features such as SEATING-CAPACITY,we used median, and for categorical features like VEHICLE_TYPE_DESC and MAKE, we used mode.

Cluster 0 is characterized by light vehicles with low tare weight (1,547 kg), moderate seating capacity (5 seats), and fewer cylinders and wheels. It is primarily composed of TOYOTA sedans, representing common passenger vehicles. Cluster 1 is composed of large-capacity vehicles, such as buses and coaches, with very high seating capacity (43), heavier tare weight (10,695 kg). Cluster 2 contains heavy industrial vehicles, such as prime movers (PMVR), with high tare weight (9,546 kg), low seating capacity (2), and 6 wheels, they are likely commercial or freight vehicles.

cluster	TARE_ WEIGHT	AGE	SEATING_ CAPACITY	NO_OF_ CYLINDERS	NO_OF_ WHEELS
0	1547.83	9.94	5.0	4.0	4.0
1	10695.07	8.89	43.0	6.0	4.0
2	9546.22	9.39	2.0	6.0	6.0

Table 5. Numerical features centroids

cluster	VEHICLE_ TYPE_DESC	VEHICLE_ MAKE	VEHICLE_ BODY_STYLE	VEHICLE_ COLOUR_1	FUEL_ TYPE
0	CAR	TOYOTA	SEDAN	WHI	P
1	BUS/COACH	UNKNOWN	BUS	WHI	D
2	HEAVY VEHICLE	UNKNOWN	PMVR	WHI	D

Table 6. Categorical features centroids for each clustering groups

Moreover, we can observe from the tables above that accidents tend to concentrate among vehicles with an average age of around 9 to 10 years, suggesting that mid-life vehicles may present higher operational risk due to potential wear or outdated safety features and this risk may be further amplified by the fact that vehicles in this age range are often still in frequent use.Additionally, white-colored vehicles appear as the dominant colour across all clusters, particularly among high-risk categories. For compact cars, the most frequently occurring maker is TOYOTA, indicating its high representation in the dataset.

5.6 High-Risk Vehicle Profile Analysis

Based on the three structural clusters identified earlier, we mapped each group back to the dataset (*data_cleaned.csv*) to evaluate its real-world accident risk, using average severity score and fatality rate as outcome indicators.

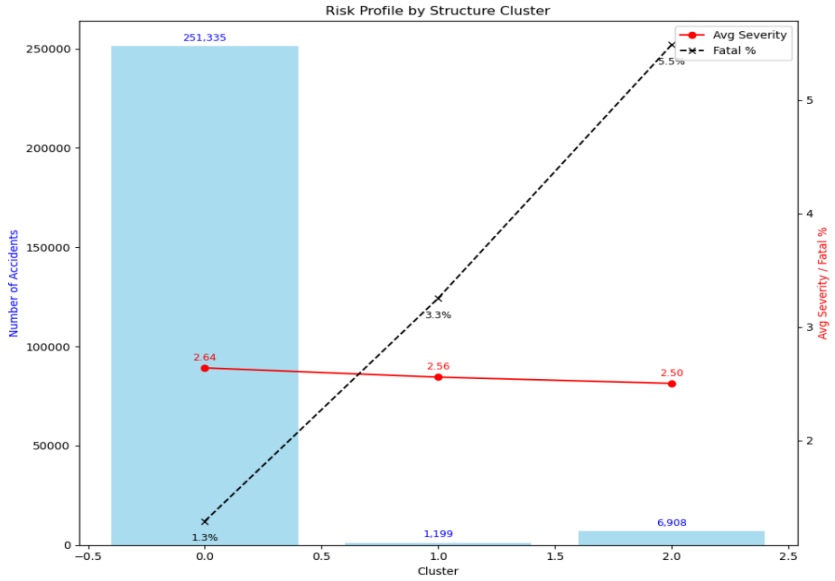


Figure 15. Bar chart of average severity and fatal rate for each cluster groups

As shown on *Figure 15* above, although Cluster 0 accounts for the overwhelming majority of accidents (~250,000), its fatality rate is the lowest (1.3%), indicating that compact passenger cars could have relatively lower structural risk. In contrast, Cluster 2 corresponds to heavy industrial vehicles that have the highest fatality rate (5.5%) despite representing the smallest average severity.

These risk profiles support the conclusion that both vehicle mass and size are critical features associated with accident severity. Furthermore, the probability of fatal outcomes may differ substantially between

compact and large vehicles, with heavier or larger vehicles more likely to be involved in high-severity or fatal accidents.

These structural groupings align well with prior findings from our machine learning models and correlation analysis, which suggested that heavy vehicles with low seat capacity and large-capacity passenger carriers, as well as compact sedans could be associated with fatal severity accidents. Although heavy vehicles offer greater protection to their own occupants, their larger inertia could make them more dangerous to others. In the event of a collision, they may tend to inflict more severe damage on surrounding vehicles and greater harm to vulnerable road users. Moreover, some heavy vehicles like prime movers are used to transport chemical fuels or other hazardous substances[2], could lead to a heightened risk of secondary disasters such as explosions or fires following a crash. In contrast, light vehicles such as compact passenger cars may provide less structural protection, but their greater maneuverability may help avoid certain accidents. Furthermore, their lower mass and inertia reduce the force of impact, potentially resulting in less damage to others in a crash. However, compact cars also represent a disproportionately large share of total accidents. While their fatality rate is relatively low, the absolute number of serious and minor injuries associated with these vehicles is significantly higher than that of heavy-duty vehicles.

Given this imbalance in accident volume across vehicle types, it is difficult to definitively conclude which type is inherently more dangerous. Instead, these insights indicate that severity-based safety strategies should account for both the structural characteristics and real-world exposure of different vehicle types, rather than targeting a single high-risk group in isolation.

Limitations and improvement opportunities

While the analysis produced valuable insights, several limitations were encountered. One major limitation lies in supervised learning is the modeling of fatal accidents (SEVERITY = 1). Due to the extreme class imbalance and the small number of fatal samples, the classification models struggled to reliably distinguish fatal cases from non-fatal ones based solely on vehicle characteristics. One possible remedy would be to merge fatal and serious injuries (SEVERITY = 1 and 2) into a single “severe” class to create a binary classification model. Although this approach improves predictive performance by increasing the effective sample size of high-severity cases. However, such a simplification deviates from the primary objective of this study, as well as reducing the granularity needed for detailed risk profiling. Additionally, excluding variables with over 30% missing values may have removed potentially informative features.

For the feature selection of high risk profile analysis, while AGE was identified as an important feature by several machine learning models, further investigation revealed its limited explanatory power in practice. Across the logistic regression analyses, we did not observe a clear or consistent distinction in accident severity between newer and older vehicles. Similarly, in the clustering analysis, vehicle age did not emerge as a decisive factor in defining any high-risk profile. This suggests that the relationship between vehicle age and accident outcomes is weak. To address this ambiguity, future work could involve age-specific modeling, such as stratifying the data into age groups or introducing interaction terms with other features.

Moreover, it is important to acknowledge that some observed trends, such as the dominance of certain makes (e.g., Toyota) or colors (e.g., white), may be influenced by underlying data imbalances or external factors, such as disproportionate market share or sampling distribution within the database and weather or light condition could affect the visibility of vehicles with different color. Therefore, while these patterns offer useful insights, they should be interpreted with caution and not generalized without further

[2] *Combined cooling, heating and Power Systems: A survey*.

validation. These limitations collectively highlight the need for deeper segmentation, richer data, and more targeted modeling strategies in future research on accident severity and vehicle risk profiling.

Conclusion

In conclusion, we examined how vehicle characteristics may influence accident severity through data cleaning, correlation analysis, and a combination of supervised and unsupervised machine learning models to identify high-risk vehicle profiles. The findings suggest, with correlation analysis, vehicle type showed the strongest relationship to severity that vehicle tare weight could play an important role in accident outcomes, on the other hand, supervised learning suggest that tare weight is the most important feature. For the Unsupervised learning model we observed different clusters indicating that heavy-duty vehicles such as prime movers may be associated with higher fatality risk despite their relatively low involvement in total accident counts, commonly used light vehicles sedans contribute to a large portion of accidents but tend to be involved in fewer fatal outcomes.

Overall, these results revealed a high-risk profile which identifies different forms of risk for light and heavy vehicles. Heavy and large vehicles are associated with high-impact but low-frequency industrial vehicles, and light and small vehicles are associated with high-frequency but lower-severity passenger vehicles. However, due to imbalanced sample sizes and potential dataset biases, we cannot definitively conclude which vehicle type is more dangerous. Instead, these results support the interpretation that different vehicle categories may carry different forms of risk, which could inform future severity-focused regulatory strategies.

Reference

Department of Transport and Planning. (2024, December 11). *Victoria road crash data* [Data set]. Victoria Government Data Directory. Retrieved May 15, 2025, from https://discover.data.vic.gov.au/dataset/victoria-road-crash-data

Xu, C., Ding, Z., Wang, C. and Li, Z. (2019). Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. 2019.09.001. *Journal of Safety Research*, from :https://doi.org/10.1016/j.jsr

Bureau of Infrastructure and Transport Research Economics. (2021, October 9). *Australia's light vehicle fleet - some insights*. https://www.bitre.gov.au/publications/2021/australias-light-vehicle-fleet-some-insights#:~:text=Vehicle%20characteristics%20considered%20include%20fuel,and%20utilities%20are%20also%20examined

Author links open overlay panelMingxi Liu a, a, b, cooling, A. combined, Oliveira, A. C., Lior, N., Huangfu, Y., Bilgen, E., Kong, X. Q., Havelsky, V., Wu, D. W., Xu, J., Dong, L., Martens, A., Pilavachi, P. A., Weisser, D., Dincer, I., Agostini, P., Floros, N., … Bourgeoisa, T. G. (2014, April 12). *Combined cooling, heating and Power Systems: A survey*. Renewable and Sustainable Energy Reviews. https://www.sciencedirect.com/science/article/abs/pii/S1364032114002263