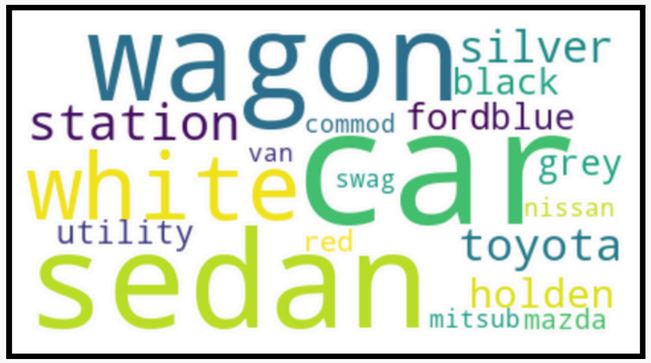
May 29, 2025

How Do Vehicle Characteristic Impact Severity of Accident &What are the possible High-Risk Vehicle Profile

EODP
COMP20008
W10G07

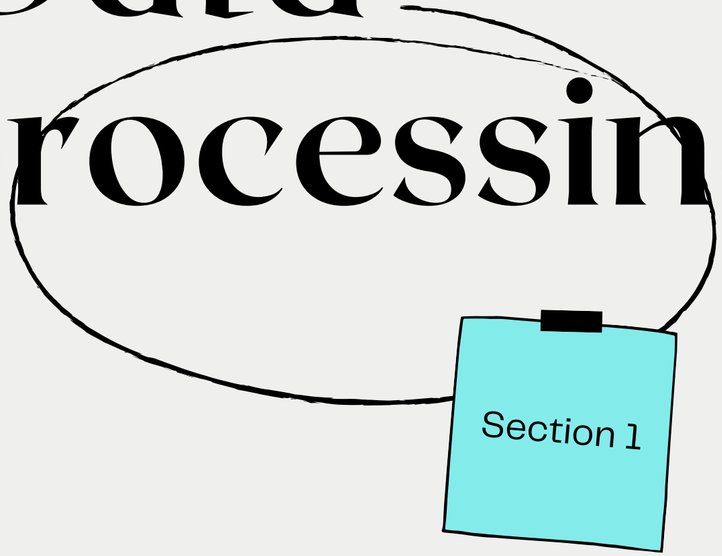
Word Cloud



Research Tasks

1. Data Pre-Processing
2. Correlation Analysis
3. Supervised Learning
   Prediction Model
4. Unsupervised learning
   Clustering
5. High Risk Vehicle Profile

Data Pre-processing



Vehicle Characteristic

Choosing our **feature variables** that are related to vehicle characteristic

We chose it based on **5 categories**

1. physical structure
2. weight and size
3. mechanical and engine
4. identification and manufacturing
5. appearance and visibilities

Dataset before processing

1. VEHICLE_TYPE,
   VEHICLE_BODY_STYLE,
   SEATING_CAPACITY,
   CONSTRUCTION_TYPE.
2. TARE_WEIGHT,
   VEHICLE_WEIGHT,
   CARRY_CAPACITY.
3. NO_OF_CYLINDERS,
   CUBIC_CAPACITY,
   VEHICLE_POWER,
   FUEL_TYPE.
4. VEHICLE_MAKE,
   VEHICLE_MODEL,
   VEHICLE_YEAR_MANUF.
5. VEHICLE_COLOUR_1,
   VEHICLE_COLOUR_2.

Pre-Processing

1. **Missing data removal** – column wise delete those where missing value > 80%
2. **Delete constant variables** – column with single value are removed
3. **Missing data fill in** – numerical use mean, categorical use mode
4. **Outlier detection** – five number summary and box plot
5. **Text cleaning** – remove stop words and punctuation.
6. **Create 'AGE' column** = Year of accident – Year of vehicle manufactured
7. **Datasets Merging** – merge accident.csv and filtered_vehicle.csv

Processed Data

- VEHICLE_MAKE
- VEHICLE_TYPE
- FUEL_TYPE
- NO_OF_CYLINDER
- SEATING_CAPACITY
- TARE_WEIGHT
- VEHICLE_BODY_STYLE
- VEHICLE_COLOUR_1
- AGE
- SEVERITY

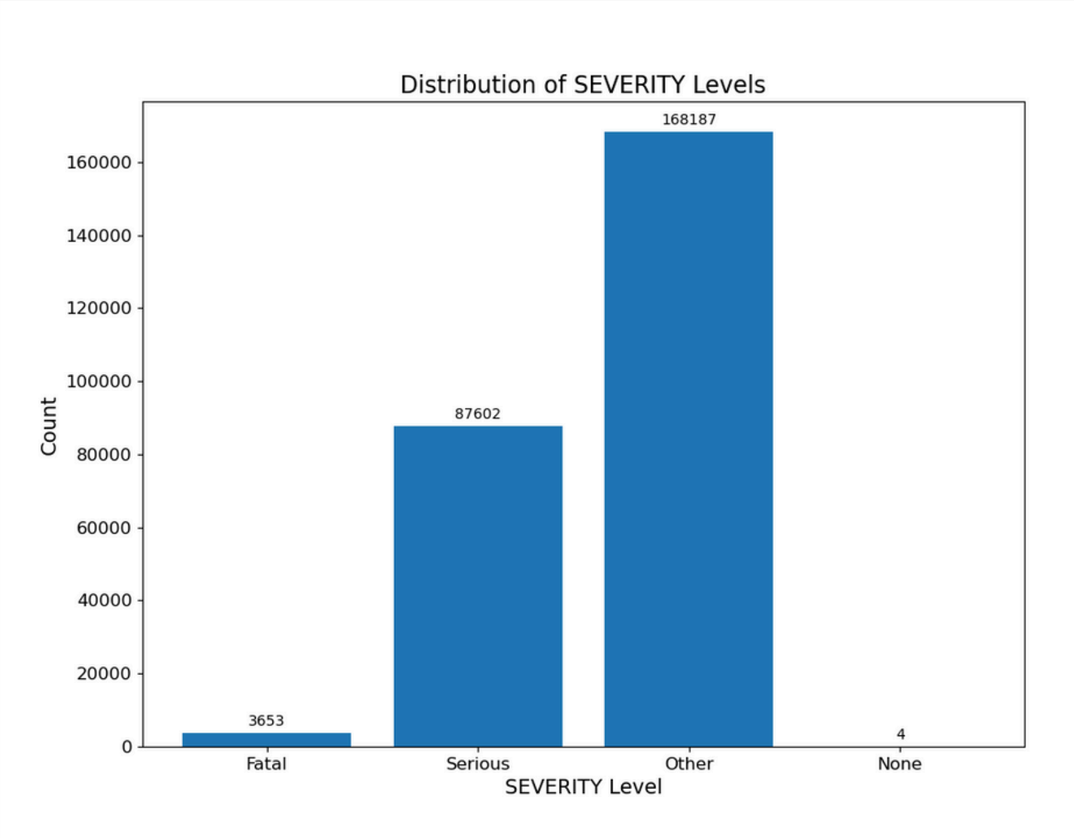
Accident Severity

Choose **target variable** “Severity” and create categories for severity levels

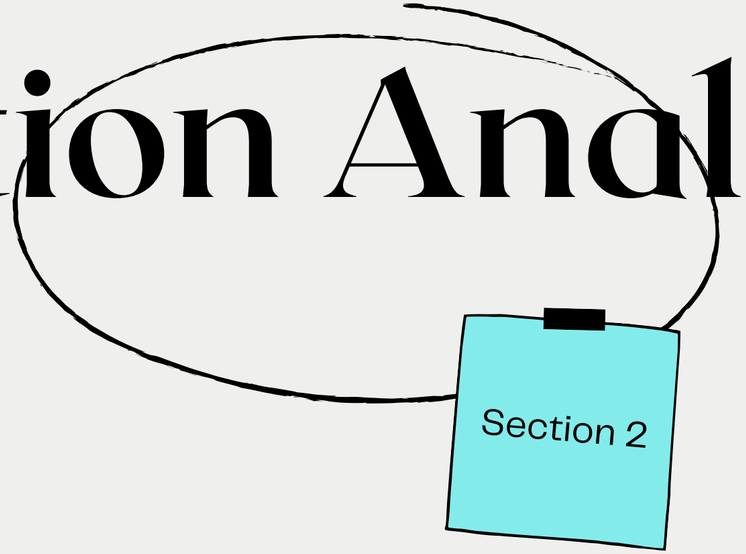
- 1 = **Fatal** accident
- 2 = **Serious** accident
- 3 = **Other injury** accident
- 4 = **None injury** accident

We decide to drop severity level = 4 (none injury) with only 4 samples

Severity Distribution



Correlation Analysis

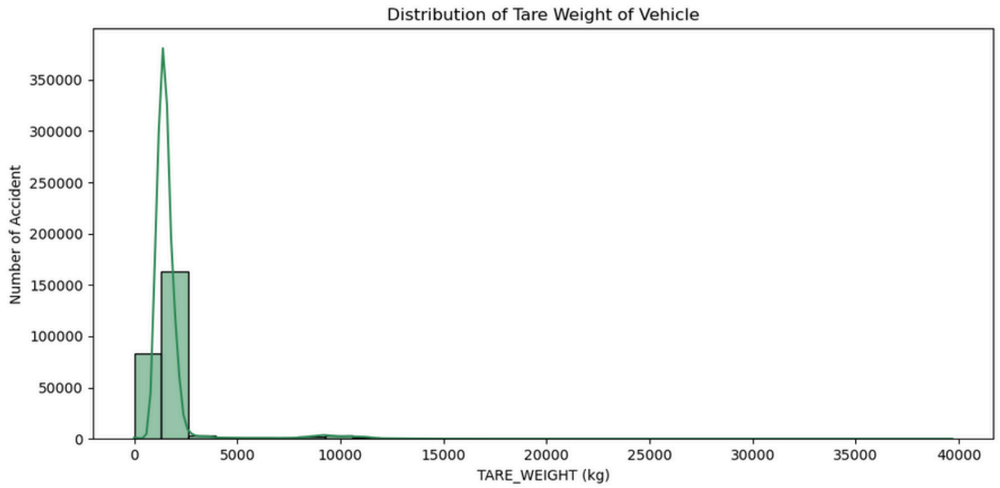


Pearson Correlation

Linear relationship & continuous variables

- Exploring **linear relationship** between continuous variable (**tare weight** and **age**) and **severity levels** (**1, 2, 3**)
- **P-value of 0** and **Pearson correlation (r<0.05)** suggest a very **weak negative linear relationship** between weight & age of vehicle and accident severity levels
- However, they could have **non-linear relationship**
- **Imbalanced data** of vehicle tare weight

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



Mutual Information

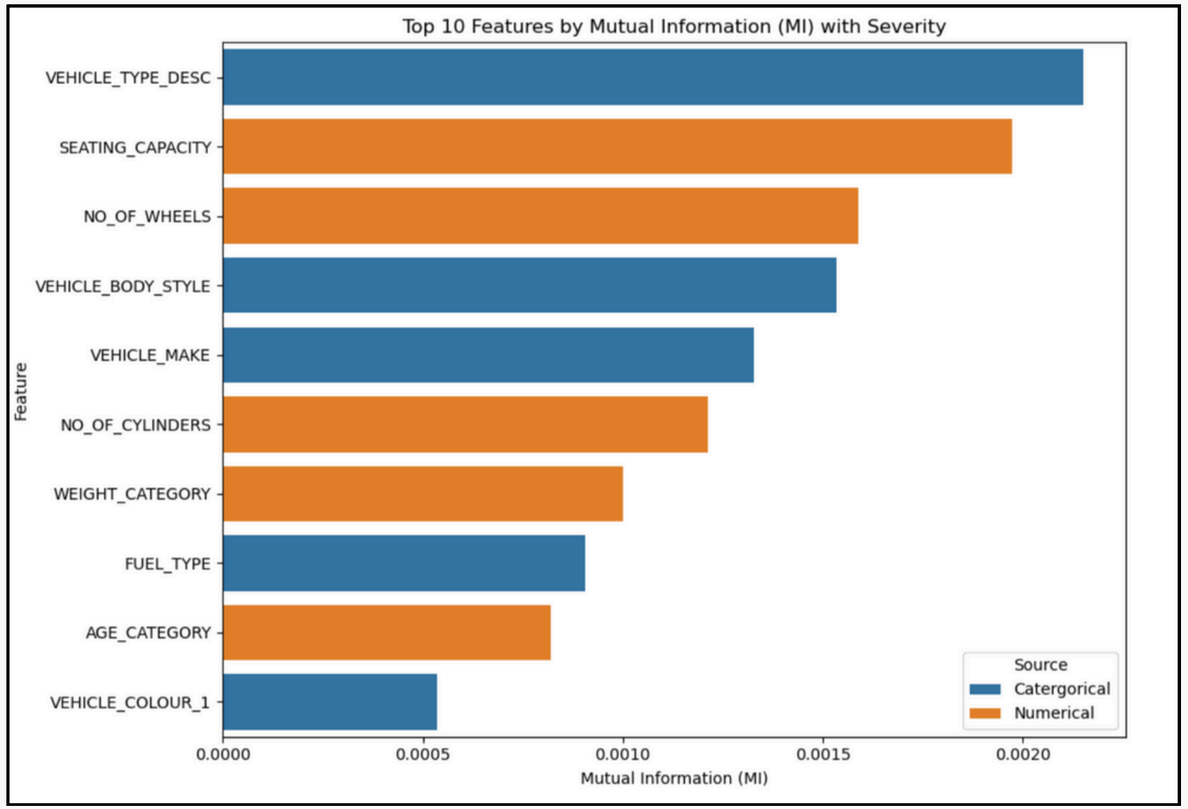
Both linear & non-linear relationships

1 VEHICLE_TYPE_DESC

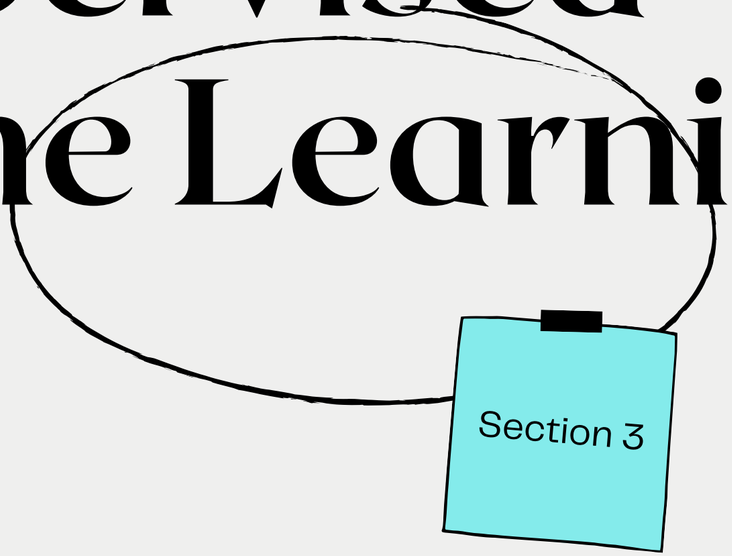
2 SEATING_CAPACITY

3 NO_OF_WHEELS

- For **continuous variable**, we **discretise** them before calculate the entropy
- Applied **ordinal encoding** to **categorical variable** to avoid artificial order
- **Vehicle type**, **seating capacity** and **number of wheels** could be more associated with accident severity

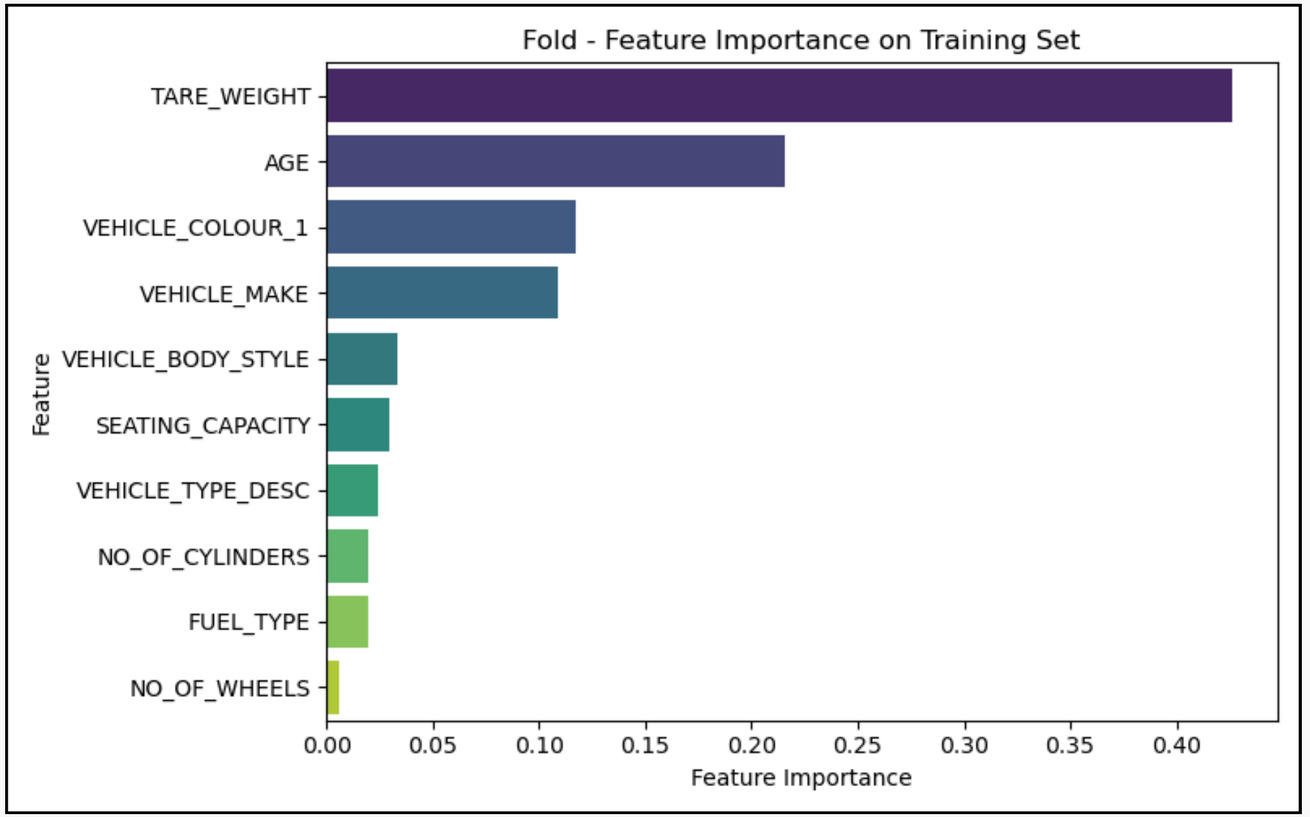


Supervised Machine Learning

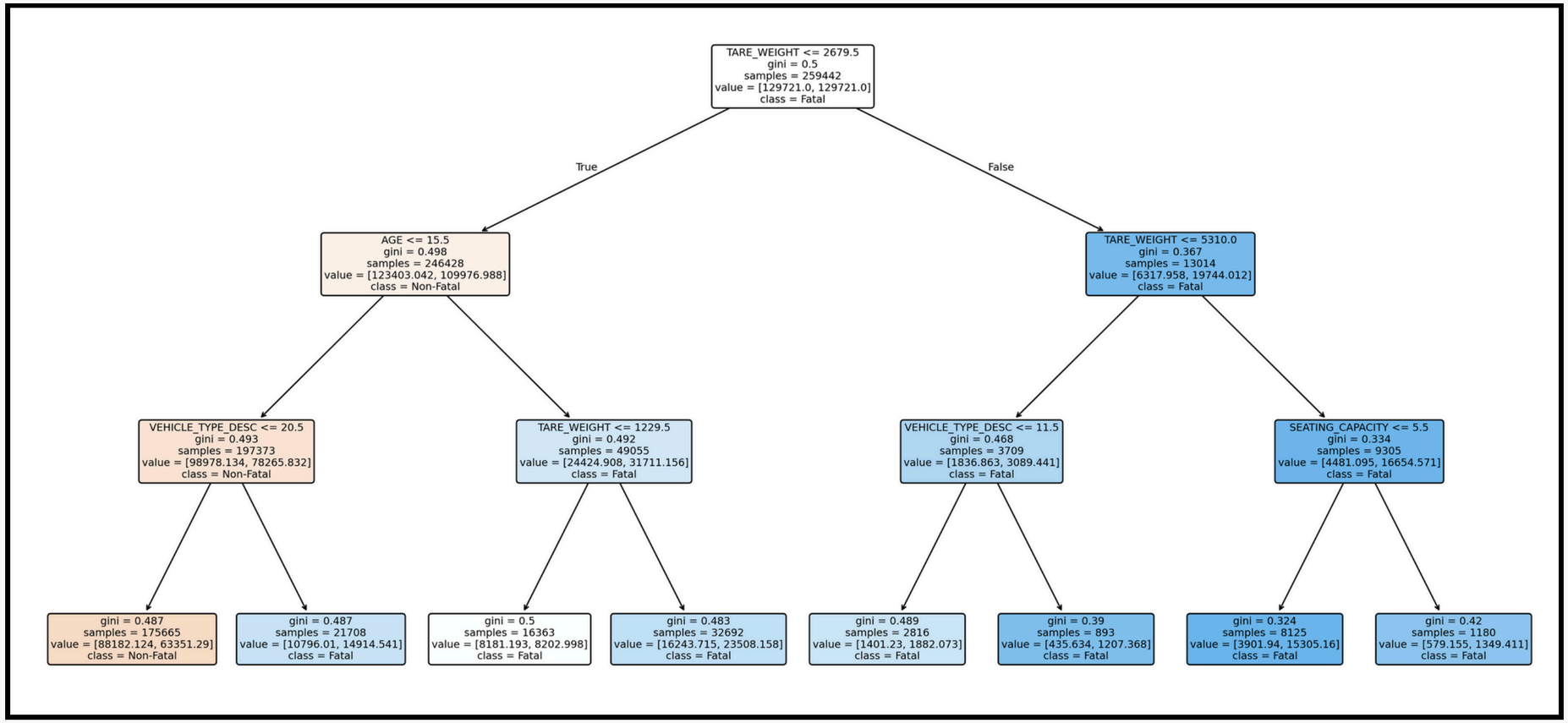


Feature Importance

- **Top 10 important features** selected by Random Forest Classifier
- **Tare weight** is the most important feature



Decision Tree (depth = 3)



1 **Primary Factors**
Top node of the tree

TARE WEIGHT

AGE

VEHICLE TYPE

SEATING CAPACITY

2 **Subsequent Factors**
Nodes on the branches

Section 3 - Supervised Learning

Classification Report

- **Zero-R Baseline** shows the highest accuracy, but 0 recall on fatal and serious accidents
- **Random Forest** model has the most balanced recall for each severity levels of accidents
- **Logistic Regression** model has the best recall on fatal accident but relatively low precision
- **Low generalisation ability** on fatal accident due to **imbalanced data**

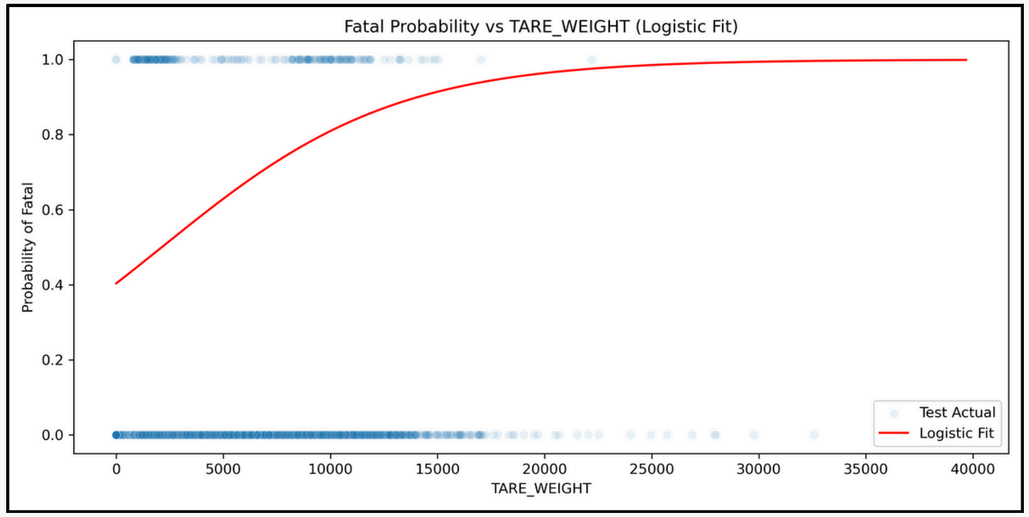
Zero-R Baseline						
	Fatal	Serious	Other	Macro Avg	Weighted Avg	Accuracy
Precision	0	0	0.65	0.22	0.42	
Recall	0	0	1	0.33	0.65	
F1-score	0	0	0.79	0.26	0.51	0.64
Support	1096	26281	50456	77833	77833	77833

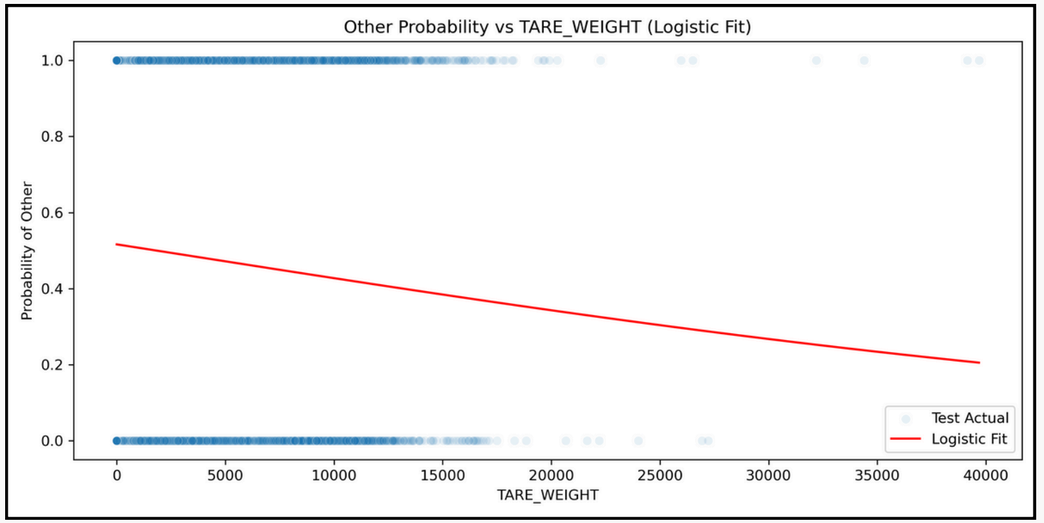
Random Forest Model						
	Fatal	Serious	Other	Macro Avg	Weighted Avg	Accuracy
Precision	0.01	0.34	0.65	0.33	0.54	
Recall	0.02	0.4	0.58	0.33	0.51	
F1-score	0.02	0.37	0.61	0.33	0.52	0.51
Support	3653	87602	168187	259442	259442	259442

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

Section 3 - Supervised Learning

Logistic Regression

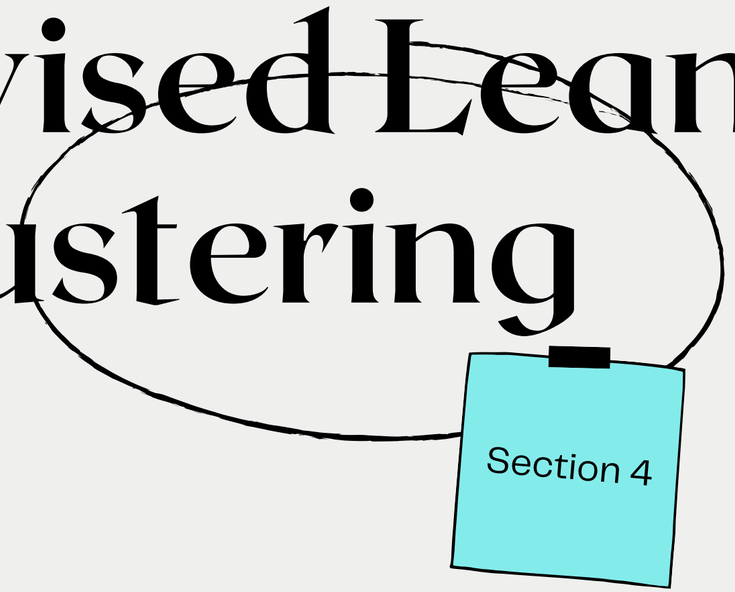




Fitting Curves

- **Possibility of fatal accident** could be **increased** as vehicle weight increased
- **Other injury possibility** of accident could be **decreased** as vehicle weight increased

Unsupervised Learning
Clustering



Feature selection

For distance-based clustering KMeans

Groupby (categorical)

- VEHICLE_MAKE
- VEHICLE_BODY_STYLE
- VEHICLE_TYPE_DESC
- FUEL_TYPE
- VEHICLE_COLOUR_1

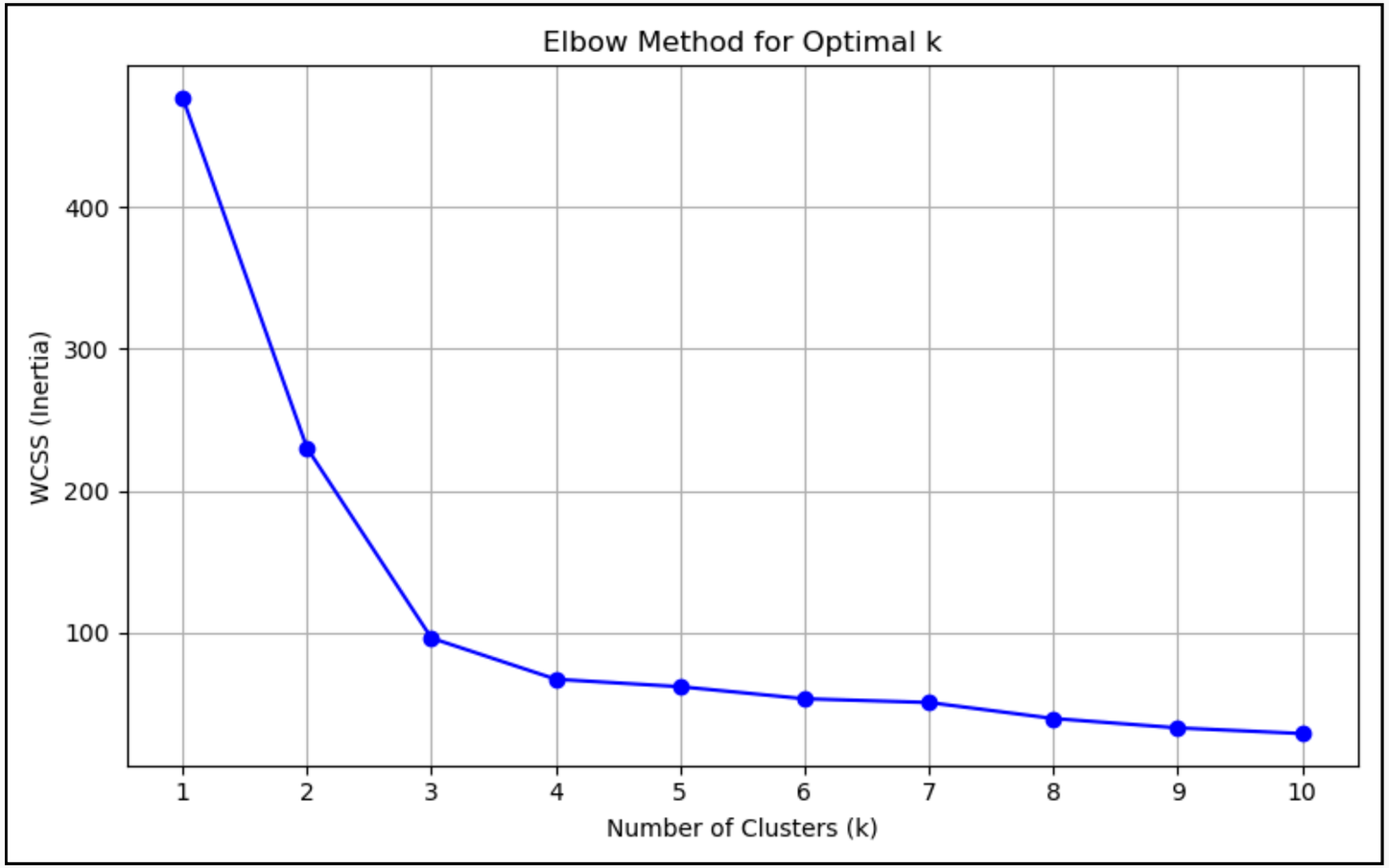
Clustering (numerical)

- TARE_WEIGHT
- SEATING_CAPACITY
- NO_OF_WHEELS
- NO_OF_CYLINDERS

Elbow method

- The point at which the rate of improvement in **WCSS** significantly slows down

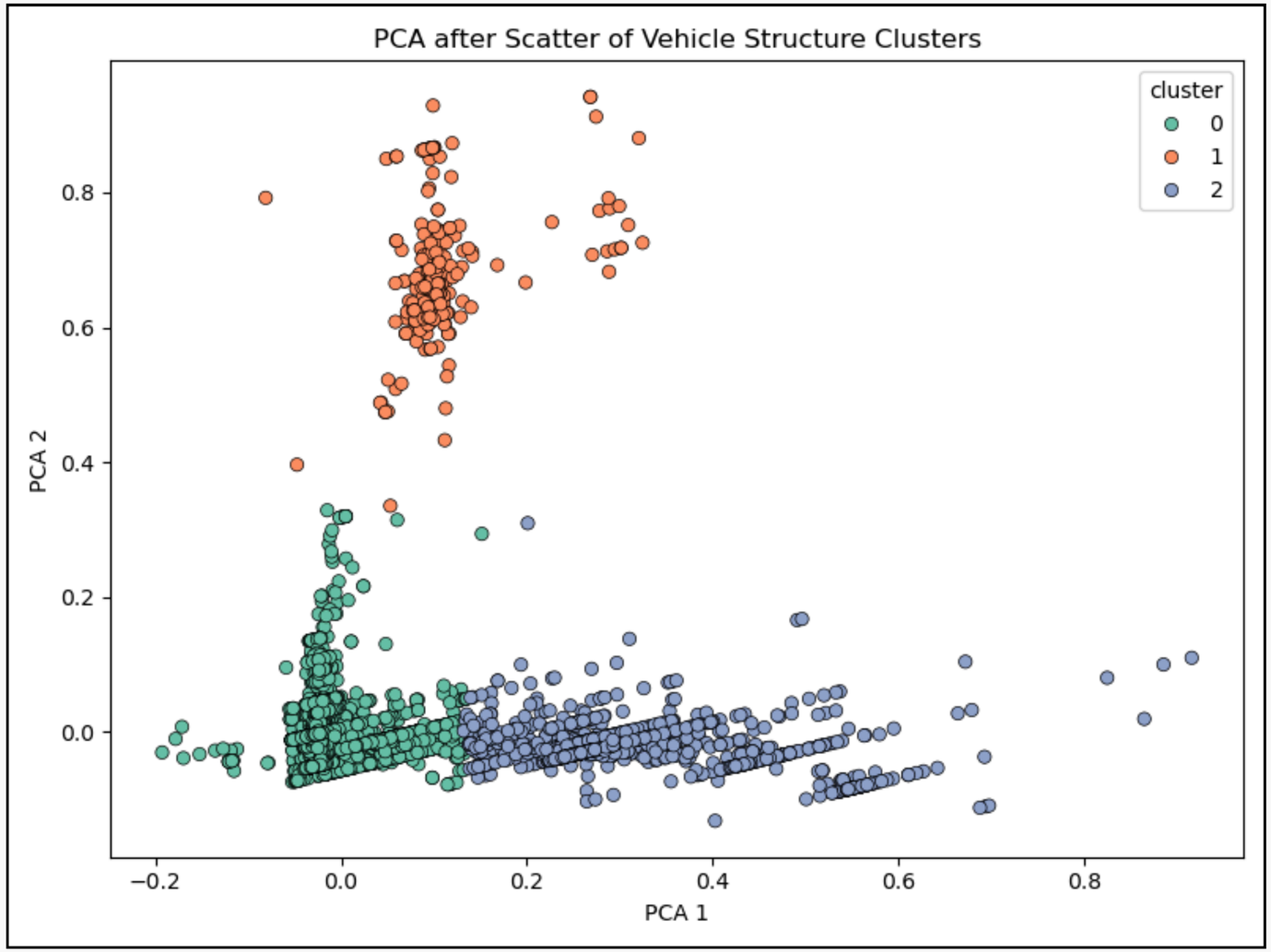
- We choose **k = 3** as our number of clusters



KMeans & PCA

- **Clustered using KMeans**, then reduced to **2D via PCA** to enable visualisation while preserving **data interpretability**
- First **2 components** explain **91.32% of variance** in vehicle structure

- Clusters show compact intra-group distance, indicating **strong internal cohesion**
- Clusters are well-separated with large inter-cluster distances, suggesting **clear groups distinctions**



Cluster Profile

Numerical Data - median

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

Categorical Data - mode

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

- **Vehicle** with average **age 9–10**, color in **white**, and made by **Toyota** is most likely to be invloved in accident recorded by the VIC government

0 **Cars**, low tare weight, 5 seating capacity, fewer cylinders and wheels

1 **Buses & coaches**, large seating capacity, highest tare weight

2 **Prime Mover**, high tare weight, small seating capacity, more wheels

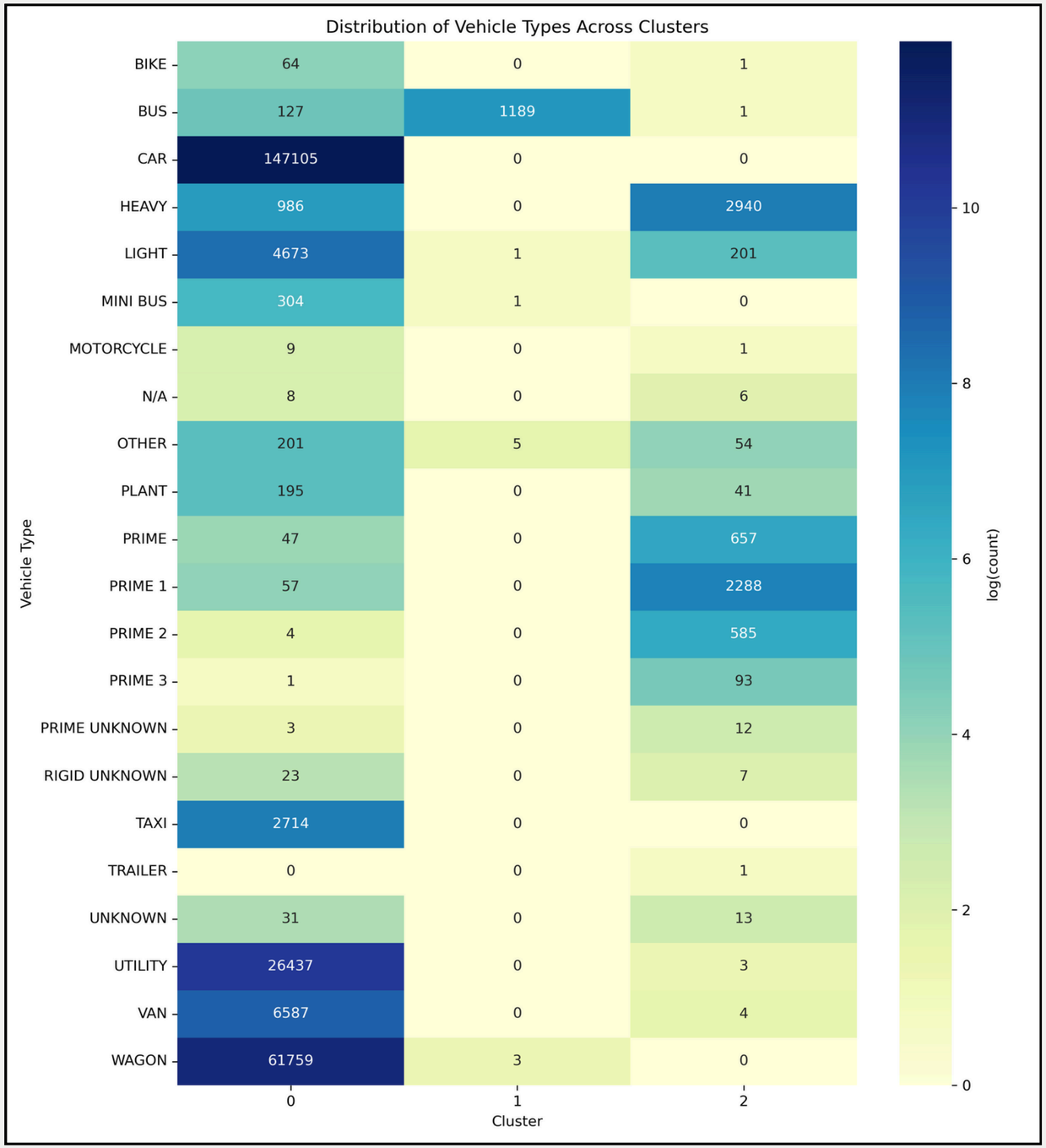
Risk Profile

W10G07
May 29, 2025

Section 5 - Risk Profile

Distribution of Vehicle Type Across Clusters

- Reveals the fact of **data imbalance**
- **Light vehicles** contribute to a **large number of accidents**
- While **heavy vehicles** are involved **far less often**
- **Car and wagon** are **widely used** and exposed, especially in urban areas



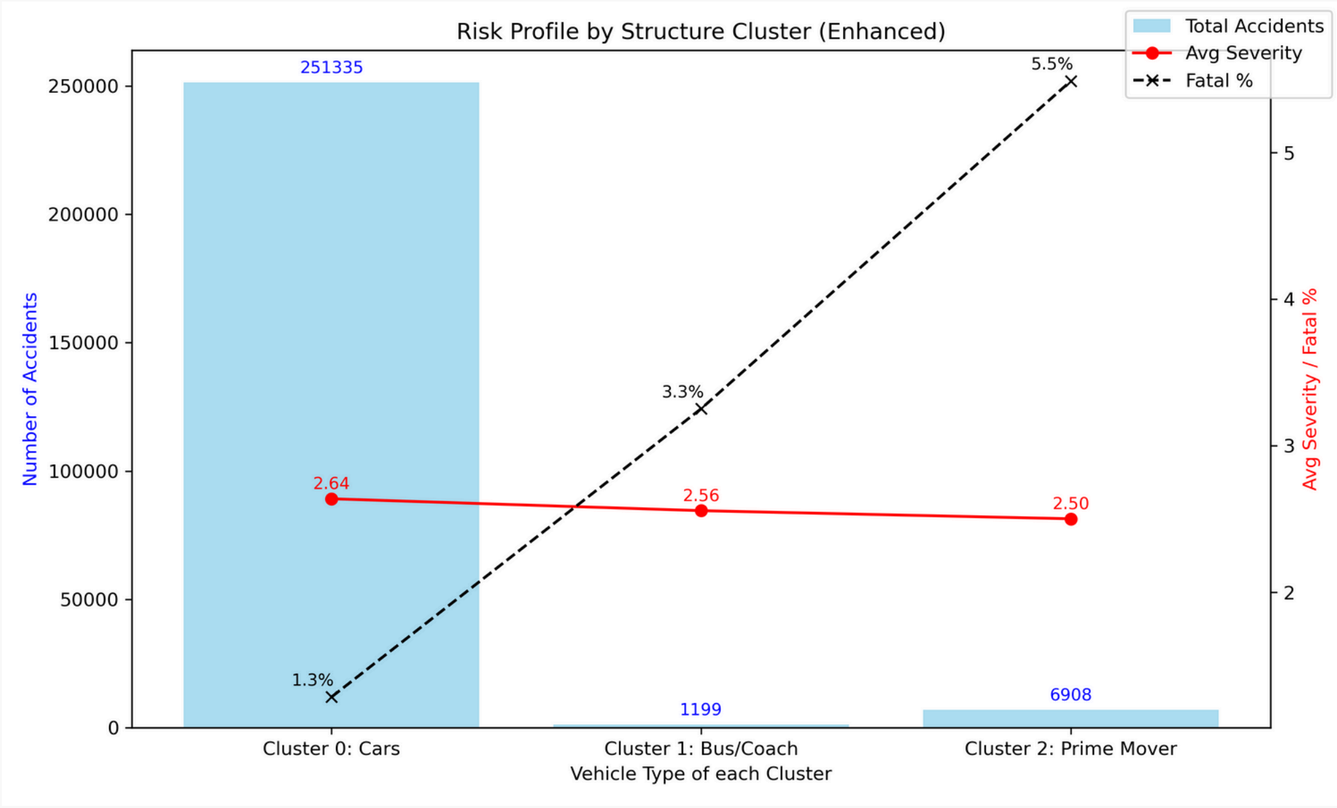
Vehicle Risk Profile

For imbalanced Data

Heavy Vehicle

Light Vehicle

- **Fatal rate** (dash line) reveals **structural risks** – **heavy vehicles** could be more likely to be involved in **fatal outcomes**
- Prime Movers and Buses may have **higher momentum** and **multiple passengers**, which may increase the severity in collisions.
- Small cars have **greater agility**, which may **reduce collision severity**
- The **average severity** (red line) across clusters appears similar



Vehicles with different weight and size exhibit different risk patterns across accident severity levels

- It is difficult to determine which vehicle types are at higher risk due to data imbalances and small sample sizes for some categories.

Limitations & Improvements

1. Our investigation is only focus on vehicle characteristic
2. Ignore other factors may influence accident severity of different vehicles
3. Limitations of vehicle age

Driver

Driver for heavy vehicles could be more professional and more experienced

Environment

Visibility of vehicles with different colors could depend on different light condition

Vehicle Age

We considered vehicle age but ignored that newer models may perform better due to technology improvements

Vehicle Owner

Older cars may degrade, but this could depends on maintenance and driving habits


Thank you!
Have a good weekend!