Vehicle Recall Patterns in the United States: Clustering Recalls by Severity

Anastasia Haidar, Viviana Alba, Hyunseo Kim, Charley Wang

1. Introduction/Background

Vehicle recalls are vital for ensuring consumer safety, as they identify when vehicles fail to meet federal safety standards.[1] The National Highway Traffic Safety Administration (NHTSA) tracks all recall campaigns across manufacturers in the U.S., providing valuable data for examining long-term trends in vehicle safety and manufacturing performance. This project utilizes nearly 60 years of NHTSA recall data[3] (1966–2025) to explore patterns in recall frequency and severity.

While recall counts are widely reported, policymakers and manufacturers often lack the tools to evaluate recall severity, such as evaluating how many vehicles are potentially affected or how significant the underlying defect may be. Prior research on recall data has largely focused on descriptive statistics or overall recall frequency,[2] leaving limited insight into the structural factors that drive high-severity recalls. Although machine learning methods have started being used in the automotive industry, its application to recall data, specifically using unsupervised learning methods such as clustering, remains limited.[2]

This project addresses this gap by clustering U.S. vehicle recalls based on severity, defined by the number of potentially affected vehicles, and by examining how severity patterns differ across manufacturers, defect categories, and time. By identifying and characterizing these severity-based clusters, this study aims to reveal systemic safety vulnerabilities, support more effective regulatory prioritization, enhance manufacturers' risk management and quality-control strategies, and provide consumers with clearer information when evaluating vehicle safety and purchase decisions.

2. Description of Dataset and Pre-Processing

The dataset used contains 29,367 rows and 15 columns of U.S. vehicle recall information from 1966 to October 1, 2025. Each record includes recall dates, affected components, associated safety risks, the number of potentially affected units, and how each issue was addressed. Prior to completing clustering analysis, we cleaned and preprocessed the data by standardizing dates and values, managing missing data, and assessing multivariate and bivariate relationships.

Missingness was assessed to identify potential biases. Columns with the highest rates of missingness were Consequence Summary (16.1%), Recall Description (8.5%), Component

(24.2%), and Completion Rate (62.2%). Due to its high missingness rate, the Completion Rate column was excluded from the analysis. Rows missing Manufacturer or Report Received Date were dropped, given that these fields are critical for temporal and manufacturer-based analyses. Missing values in the Consequence Summary and Component were imputed with "No description" and "Unknown", respectively. A chi-squared test of independence revealed that missing values of Consequence Summary were not independent of manufacturer (χ^2 = 9,863, degrees of freedom = 1,839, p-value ≈ 0.0), suggesting that some manufacturers are more likely to omit consequence descriptions than others.

During exploratory data analysis, temporal trends and manufacturer-specific patterns in vehicle recalls were examined. Looking at manufacturer-level patterns, the top three manufacturers by total number of recalls are General Motors, LLC (1,675 recalls), Ford Motor Company (1,611 recalls), and Chrysler (1,407 recalls). However, the manufacturer associated with the highest average number of potentially affected vehicles per recall is Pennzoil Quaker State Company, with an average of about 16.2 million vehicles potentially affected.

An increasing trend in recalls over time was also discovered, which may reflect regulatory changes, shifts in manufacturing practices, or improved reporting efforts. Monthly recall activity remained relatively consistent, indicating no strong seasonal trends, and day-of-week trends showed only a greater concentration of midweek recalls (Tuesday through Thursday), aligning with standard business operations. Further, the number of potentially affected vehicles was heavily right-skewed, so a logarithmic transformation was applied to normalize the distribution.

3. Description of Methods Used

Analyses were conducted collaboratively using Python (pandas, seaborn, matplotlib, numpy, scipy, and scikit-learn). Project tasks were divided as follows: Charley and Anastasia led exploratory data analysis; Viviana and Hyunseo led preprocessing and clustering. All members contributed to writing, coordination/planning, and review.

Given our goal of clustering recalls by severity, three attributes were identified as most important: (1) the number of potentially affected vehicles (used as the primary severity indicator), (2) recall year, and (3) manufacturer. These features directly support our goal of forming severity-based clusters and evaluating patterns across time and manufacturers. For model stability, the dataset was restricted to entries labeled as 'Vehicle' within the 'Recall Type' column (25,564 rows). Only the top 100 manufacturers were retained; all remaining manufacturers were grouped into an "Other" category. The dataset was further prepared for clustering analysis by standardizing numeric features ('Potentially_Affected_LogTransform', 'Year') using z-score normalization and one-hot encoding categorical variables ('Manufacturer').

K-Means clustering was selected for its ability to scale efficiently to large datasets, strong performance on large datasets with mixed feature types (after encoding), and its ease of interpretation. The optimal number of clusters was determined using both the Elbow and Silhouette methods, applied to fixed random samples (10,000 for Elbow and 8,000 for Silhouette) to reduce computational cost and limit noise sensitivity. The Elbow method identified the point of diminishing returns in within-cluster variance, while the Silhouette score evaluated the balance between cluster cohesion and separation. The Elbow optimum was used as the primary criterion, ensuring that the Silhouette score remained locally high near the same K. In cases where the two measures diverged, the more parsimonious Elbow solution was chosen following visual inspection. After determining the optimal K using a representative subset of the data, the final K-Means model was retrained on the full dataset to assign cluster labels.

Recall clusters were categorized into low, medium, and high severity levels based on the median number of vehicles affected within each cluster. Clusters were ranked by their median recall size, and these rankings were mapped to ordered severity levels, ensuring that clusters with larger medians corresponded to higher severity classifications. Following clustering, we analyzed severity patterns across years, manufacturers, and component categories. This included evaluating both frequency-based and proportional (severity rate) distributions at the manufacturer level, as well as examining component-level differences across clusters to identify defect types most strongly associated with each severity group.

4. Experiment Setup, Model Evaluation, and Analysis

K-Means clustering was applied using three features: the log-transformed number of potentially affected vehicles, recall year, and manufacturer. Both the Elbow and Silhouette criteria identified k = 3 as the optimal cluster solution. The resulting clusters were labeled "low," "medium," and "high" severity based on their median recall size. These three clusters revealed clear gradations in severity: low-severity recalls, which accounted for approximately 45% of all cases, typically involved tens to a few hundred vehicles; medium-severity recalls (25%) generally affected several thousand vehicles; and high-severity recalls (30%) often impacted tens

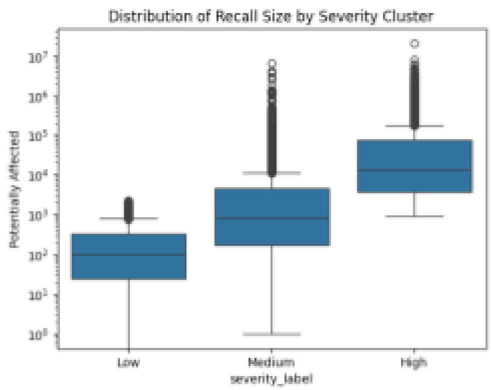

Figure 1.0: Distribution of Recall Size by Severity Cluster

of thousands and occasionally millions of vehicles. The distinct separation in recall magnitudes indicates that K-Means successfully identified meaningful severity categories. Figure 1.0 visually reinforces this separation by illustrating the distribution of recall size, which increases consistently from low to high severity clusters.

Temporal analysis revealed that medium severity recalls dominated the industry up until the 1990s, after which their frequency declined. Between 1990 and 2000, there was a sharp increase in both low and high severity recalls, likely due to stricter regulations and improved defect detection methods. From the 2010s onward, low and high severity recalls continued to rise, with low severity events remaining the most common overall.

Examining recall patterns across top manufacturers highlights both frequency and proportional differences in the data. Figure 1.1 reveals that the top 3 manufacturers (General Motors (GM), Ford, Chrysler) dominate both recall count and severity. General Motors, which owns Chevrolet, Buick, GMC, and Cadillac, reports the highest absolute number of high-severity recalls (739), followed closely by Ford (806) and Chrysler (687). A similar pattern is observed in medium-severity and low-severity recalls, where these large manufacturers consistently appear at the top of each category.

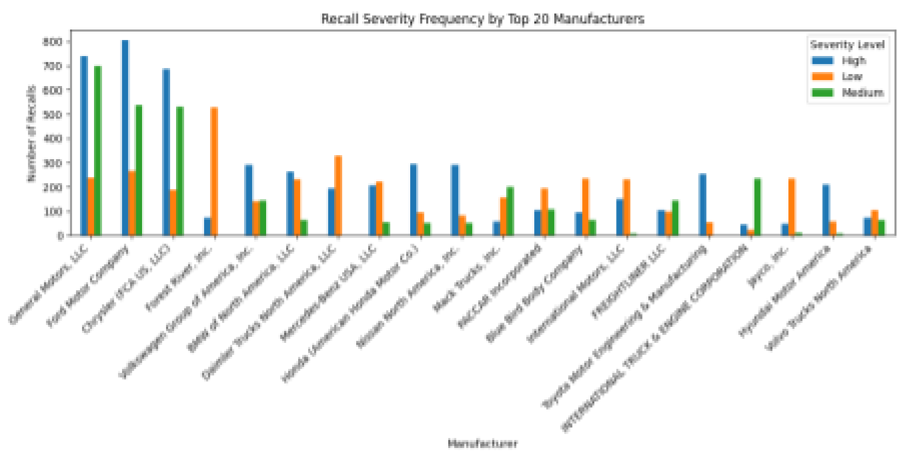

Figure 1.1: Recall Severity Frequency by Top 20 Manufacturers

However, raw frequencies naturally reflect differences in production scale, meaning that manufacturers with larger output tend to report more recalls across all severity levels. To obtain a more comparable perspective, we evaluated the proportional distribution of severity levels within each manufacturer (as seen in Figure 1.2). While GM, Ford, and Chrysler have the highest

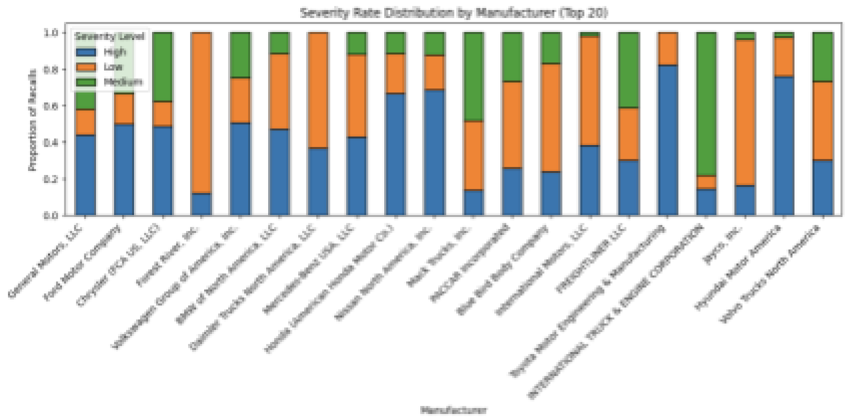

Figure 1.2: Severity Rate Distribution by Manufacturer

counts, their high-severity rates range only from 44% to 50%. In contrast, smaller manufacturers exhibited disproportionately high rates of severe recalls. Toyota shows an 82.3% high-severity rate, Hyundai 76.1%, Honda 66.7%, and Nissan 68.6%. Conversely, manufacturers such as Forest River and Jayco report predominantly low-severity recalls, with low-severity proportions of 87.7% and 79.7%, respectively.

This comparison shows a key distinction. Frequency-based analysis identifies the manufacturers most affected in absolute terms, reflecting production scale and market presence, whereas severity rate analysis uncovers underlying risk tendencies that are obscured when relying solely on raw counts. In other words, GM, Ford, and Chrysler dominate in total volume, but manufacturers such as Toyota and Hyundai exhibit a higher relative likelihood of issuing severe recalls when recalls do occur. Therefore, the contrast between absolute and proportional measures provides a more complete understanding of manufacturer-level safety performance.

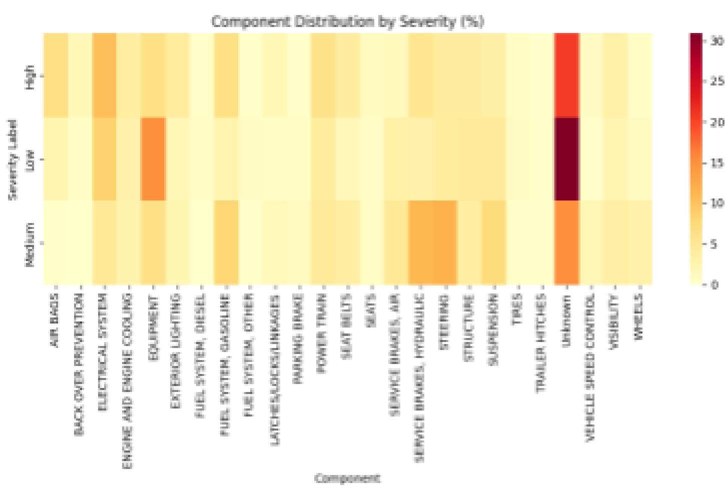

Figure 1.3. Component Distribution by Severity.

We also sought to understand whether the nature of the defect itself contributes to severity differences. To better understand why severity levels differ, we extended our analysis to examine whether specific defect types are disproportionately associated with certain severity clusters. Using component frequency distributions, we observed distinct patterns across the three severity groups with a heatmap, as seen in Figure 1.3 and a bar graph. In the high-severity cluster, the most frequently reported components were Electrical System (9.7%), Equipment (7.7%), Air Bags (6.0%), and Fuel System; Gasoline

(6.0%). The medium severity cluster showed higher proportions of Steering (10.6%), Service Brakes; Hydraulic (10.3%), and Fuel System; Gasoline (7.1%). In the low-severity cluster, the most common components were Equipment (15.2%), Electrical System (8.4%), and Suspension (4.7%). The Unknown category appeared at consistently high rates across all clusters. Across all three clusters, Equipment and Electrical System appeared consistently, indicating that these components are commonly reported regardless of severity. These results indicate that each severity level is associated with a distinct distribution of frequently reported components.

To assess clustering quality, we calculated several metrics that jointly assess intra-cluster cohesion and inter-cluster separation. As seen in Table 1, these included the Davies-Bouldin index, which quantifies cluster compactness and separation through pairwise comparisons, the Silhouette coefficient, which evaluates the relative distance between data points within and across clusters, and the Calinski-Harabasz index, which examines the ratio of inter- to intra-cluster variance. The Davies-Bouldin index value of 1.297 and Silhouette score of 0.266 for k=3 both indicate moderate inter–cluster separation and moderate intra-cluster cohesion, expected for noisy, high-variance, real-world data. The Calinski-Harabasz index value of 10,583.5 was the highest among candidate k values, further supporting k=3 as the optimal k-value. Taken together with the separation in recall magnitudes and meaningful differences observed across time, manufacturers, and defect categories, the results demonstrate that the clustering framework successfully identified patterns in vehicle recall severity.

K	Calinski-Harabasz	Davies-Bouldin
2	9219.25	1.51434
3	10583.5	1.29764
4	8611.12	1.6522
5	7528.96	1.71692
6	6782.87	1.66665

Silhouette score for (k=3): 0.266

Table 1. Cluster Validation Metrics Across K-Values

5. Observations

Clustering analysis reveals several important insights into how recall severity emerges in the U.S. automotive industry. First, severity is not randomly distributed; rather, it is systematically shaped by both the type of defect and manufacturer-specific patterns. The identification of three distinct severity clusters reflects meaningful differences in subsystem criticality. High-severity recalls are disproportionately concentrated in safety-critical components such as electrical systems, fuel systems, and airbags, where failures can lead to fires, stalling, loss of control, or fatal injury. In contrast, low-severity recalls more often involve equipment or non-critical structural components, indicating that the technical nature of the defect strongly influences the magnitude of its impact.

At the manufacturer level, clustering uncovers patterns that raw recall counts alone hide. While high-volume manufacturers such as General Motors, Ford, and Chrysler record the greatest number of severe recalls in absolute terms, which is expected given their production scale, proportional severity rates reveal a different risk landscape. Manufacturers including Toyota, Hyundai, Honda, and Nissan exhibit substantially higher relative rates of high-severity recalls, suggesting underlying vulnerabilities that may be associated with specific design choices, technological platforms, or supply chain dependencies. These manufacturer-specific severity tendencies point to structural safety challenges that extend beyond production volume. Temporal trends reinforce these findings. The rise in high-severity recalls beginning in the 1990s aligns with increasing vehicle complexity, widespread adoption of electronic control systems, and evolving regulatory legislation. This suggests that severity trends reflect broader shifts in engineering practices and regulatory oversight, such as the Intermodal Surface Transportation Efficiency Act of 1991, which required all cars and light trucks sold in the US to have airbags on both sides of the front seat, reducing the risk of dying in a head-on collision by 30 percent.[4]

Evaluation metrics further support the validity of the three-cluster solution. The Silhouette coefficient and Davies–Bouldin index indicate moderate but meaningful separation between clusters, which are appropriate for a heterogeneous real-world dataset, while the Calinski–Harabasz index clearly favors the k=3 solution. The consistent alignment between cluster labels and recall size distributions suggests that residual overlap is attributable to noise inherent in large, multi-decadal administrative datasets rather than methodological error. Taken together, these metrics confirm that the clustering successfully identifies structurally meaningful variations in recall severity.

6. Future Work / Conclusion

This project demonstrates that recall severity is shaped by an interplay of subsystem risk, organizational design, manufacturing tendencies, and long-term industry evolution. By moving beyond simple frequency counts, the clustering approach applied in this study provides a more nuanced understanding of how and why severe recalls emerge. The identification of three severity clusters (low, medium, and high) revealed meaningful distinctions in defect criticality, manufacturer vulnerability patterns, and temporal shifts associated with increasing vehicle and regulatory complexities. These findings underscore the value of unsupervised learning as a tool for uncovering structural patterns not apparent through descriptive statistics alone.

Several opportunities for future work arise from this analysis. These include predictive modeling of high-severity recall risk, component-level safety forecasting, and targeted regulatory or quality-control interventions for manufacturers with elevated severity profiles. Methodologically, the analysis could be extended by exploring alternative clustering frameworks, such as density-based or hierarchical approaches, or by incorporating additional features, including natural language processing of the “Consequence Summary” and training with additional variables in the dataset, while carefully monitoring for overfitting. These enhancements may improve robustness in the presence of noise and heterogeneous component types. Linking recall severity to real-world crash or injury outcomes using federal Department of Motor Vehicle (DMV) data and developing standardized consumer-facing safety indicators can also serve as promising avenues for translating analytical findings into actionable insights that contribute to safer roads across the United States. These directions highlight the broader relevance of this work: understanding recall severity at scale can inform regulatory oversight, support manufacturer quality-control strategies, and ultimately enhance transparency for consumers, allowing them to make more informed purchasing decisions in an increasingly complex automotive market.

7. References

[1] Automotive Research Center, "What You Need to Know about Vehicle Safety Recalls," AAA – ACE, American Automobile Association, Mar. 4, 2019. [Online]. Available: https://www.ace.aaa.com/automotive/advocacy/automotive-safety-recalls-explained.html.

[2] K. Ahsan, “Trend Analysis of Car Recalls: Evidence from the US Market,” International Journal of Managing Value and Supply Chains, vol. 4, pp. 1–16, December 2013, doi: 10.5121/ijmvsc.2013.4401.

[3] National Highway Traffic Safety Administration, "Recalls Data," U.S. Data.gov, updated Oct. 5, 2025. [Online]. Available: https://catalog.data.gov/dataset/recalls-data.

[4] U.S. Department of Transportation, "A Moment in Time: Highway Safety Breakthrough," FHWA Highway History Website. [Online]. Available: https://highways.dot.gov/highway-history/general-highway-history/moment-time-highway-safety-breakthrough.