Engineering Safer Roads: A Clustering Analysis of Road Accident Data in Canada

Abstract

This study employs k-means clustering to analyze vehicular accident data in Canada from 1999 to 2014, aiming to uncover trends and inform safety enhancements. The clustering identified rear-end collisions as the most frequent accident type, emphasizing the need for engineering interventions to prevent rear-end collisions. This research advocates for developing driver support systems like automatic braking, intelligent speed adaptation, tire technology, and road materials improvements to address adverse weather conditions. The study concludes that integrating these engineering solutions is crucial for reducing accidents and enhancing road safety, highlighting the significant impact of technology on public health and safety.

1. Introduction

Car accidents remain a significant concern around the world, posing a substantial threat to public safety. Statistics reveal an alarming trend of vehicular accidents, resulting in a considerable number of fatalities and injuries. Each year, 1.35 million people are killed on roadways in the world. Every day, almost 3,700 people are killed globally in crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians (WHO, 2018). This situation underscores an urgent need to address the issue proactively. Enhancing road conditions and integrating advanced driver-support systems are pivotal strategies in reducing these incidents. By prioritizing these aspects, there's a potential to significantly diminish the frequency and severity of accidents, thereby safeguarding lives and improving overall road safety. This study aims to delve into the intricacies of car accidents in Canada from 1999 to 2014, utilizing unsupervised clustering methods to uncover underlying trends and inform effective engineering perspective interventions.

Engineering plays a crucial role in mitigating the risks associated with car accidents. The application of engineering principles can lead to the development of safer road infrastructures, such as better-designed road alignments, improved signage, and optimized traffic control mechanisms. Additionally, advancements in vehicle technology, including the integration of driver-assistance systems, promise to reduce accident rates significantly. These systems, engineered to enhance vehicle safety, can aid drivers in navigating complex traffic situations, detect potential hazards, and even automate critical responses in emergency scenarios. Even though the potential of engineering interventions is significant, there are limited studies to investigate their potential from real accident statistic data. Most studies with accident data have revealed the cause of accidents, but they mainly focus on policy interventions (Nikam 2020; Aljaban 2021). By understanding the patterns and factors contributing to accidents, engineering solutions can be more effectively tailored to address specific risks and vulnerabilities.

The landscape of driver support systems has seen remarkable advancements in recent years, introducing many technologies to enhance road safety. Examples of these innovations include adaptive cruise control, lane departure warnings, and automatic emergency braking systems. These systems represent significant strides in vehicular technology, designed to assist drivers in various aspects of vehicle operation and situational awareness. However, despite the availability of these diverse systems, a critical question remains unanswered: which of these technologies is most effective in reducing the incidence of car accidents? Vehicle companies

provide their own studies, showing how many accidents can be avoided by these technologies. However, there is no research to study the impact of these technologies in real traffic conditions based on the accident statistics data. It is crucial first to understand the root causes of these accidents and the specific contexts in which they occur. This knowledge is essential for determining which driver support systems are most relevant and effective in different scenarios, ensuring that the development and implementation of these systems are strategically aligned with the actual needs and situations encountered on Canadian roads.

The utility of unsupervised clustering methods in understanding car accidents is profound, especially when compared to more traditional data analysis techniques that might rely solely on mode values (Shi et al. 2022; Suarez-Del Fueyo et al. 2021). For instance, while a surface-level examination of data might suggest that most accidents occur on sunny days, this could be misleading without considering the overall frequency of sunny weather in Canada. Unsupervised clustering goes beyond such simplistic interpretations by identifying patterns and correlations within the data that are not immediately apparent. This method enables the segmentation of accident data into meaningful clusters based on various attributes like weather conditions, road features, and driver demographics. Doing so provides a nuanced understanding of the conditions and factors that commonly precede accidents. This deeper insight is invaluable for developing targeted strategies and interventions to enhance road safety, as it moves beyond generalizations to uncover the specific circumstances and combinations of factors that most often lead to accidents.

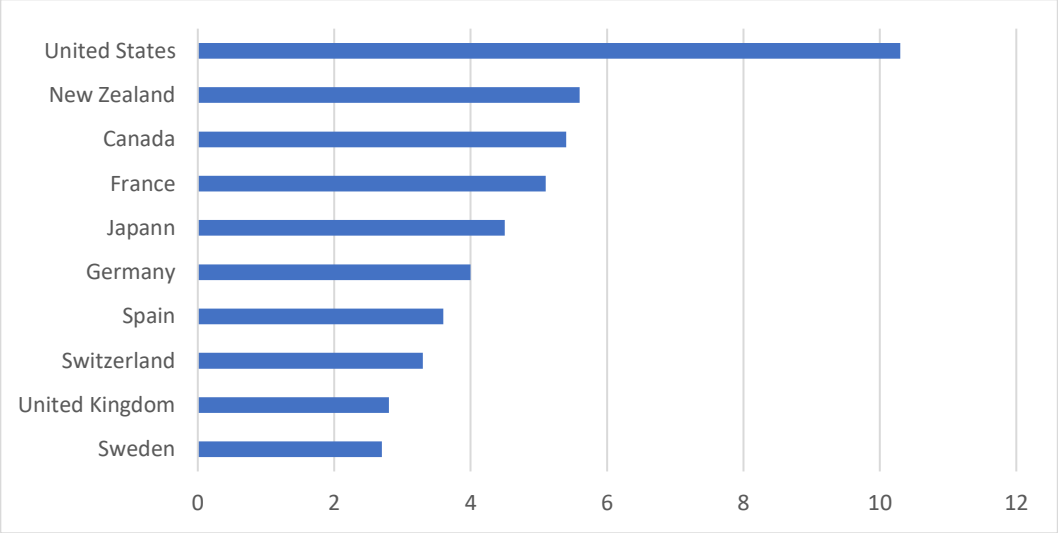

Figure 1: Motor Vehicle Crash Death in 10 High-Income Countries (WHO, 2015)
The value indicates deaths per 100,000 people in 2013.

2. Methodology

2.1 Data

I collect the collision data for car accidents in Canada from 1999 to 2014 from the Kaggle dataset (https://www.kaggle.com/datasets/tbsteal/canadian-car-accidents-19942014/data?select=drivingLegend.pdf). This dataset is created based on the data provided from Transport Canada and Statistics Canada. This dataset contains various features of accidents

that happened in Canada, such as time of day, driver gender, traffic control, weather, and so on. Among 22 features, I selected the following variables to study the possibility of engineering interventions to reduce accidents: collision hour, collision configuration, road configuration, weather, road surface, road alignment, traffic control, vehicle type, vehicle year, driver sex, and driver age. The dataset contains 3,853,678 collisions. To reduce the computational burden, I randomly selected 10,000 collisions and analyzed them in our study. Table 1 presents the description of each variable.

Table 1: Variables and their Description and Mode Values

Variable	Description	Mode Value
Hour	Collision hour	16:00 to 16:59
Collision Configuration	Collision configuration	Rear-end collision
Road Configuration	Roadway configuration	At an intersection of at least two public roadways
Weather	Weather condition	Clear and sunny
Road Surface	Road surface	Dry, normal
Road Alignment	Road alignment	Straight and level
Traffic Control	Traffic control	No control present
Vehicle Type	Vehicle type	Light Duty Vehicle
Vehicle Year	Vehicle model year	2000
Driver Sex	Driver sex	Female
Driver Age	Driver age	18

2.2 Unsupervised Learning Method

In this research, we employed the k-means clustering method to unravel the complex features of collision accidents. This choice of methodology was driven by the need for a robust and interpretable analysis of the multifaceted accident data collected across Canada. K-means clustering was selected due to its efficiency and effectiveness in handling large datasets. This algorithm is renowned for its simplicity and its capability to produce clear, distinct clusters based on variance minimization. In the context of accident analysis, where data points represent unique incidents with diverse attributes, k-means facilitate the grouping of these incidents into clusters with similar characteristics. This enables us to discern patterns and commonalities among accidents, which might not be apparent in an un-clustered dataset.

Prior to clustering, I standardized the variables to ensure uniformity in scale. This step is crucial in k-means clustering, as the algorithm relies on distance metrics to assign data points to clusters. Variables measured on different scales can disproportionately influence the clustering outcome, leading to skewed results. Standardization mitigates this issue by converting all variables to a common scale without distorting differences in the ranges of values or losing information. This ensures that each variable contributes equally to the clustering process, making the resulting clusters more reliable and interpretable.

The use of principal component analysis (PCA) in our methodology was strategic for reducing the dimensionality of our dataset to two dimensions. PCA is a powerful technique for simplifying complex datasets with numerous variables, extracting the most important features

without significant loss of information. Reducing the dataset to two principal components facilitated a more manageable and visually interpretable clustering process. It allowed us to capture the essence of the data's variability and reveal patterns that are not immediately obvious in a high-dimensional space.

I calculated silhouette scores for cluster numbers ranging from two to seven to ascertain the most appropriate number of clusters. The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates a model with better-defined clusters. We discovered that a four-cluster solution had the highest silhouette score of 0.51, signifying an optimal balance between the number of clusters and the distinctness of each cluster. Figure 2 illustrates the variation of silhouette scores in relation to the number of clusters, guiding our selection of the four-cluster model as the most appropriate for our analysis. This approach ensures that the chosen clustering model not only captures significant patterns in the data but also maintains clarity and distinction among the different accident groups.

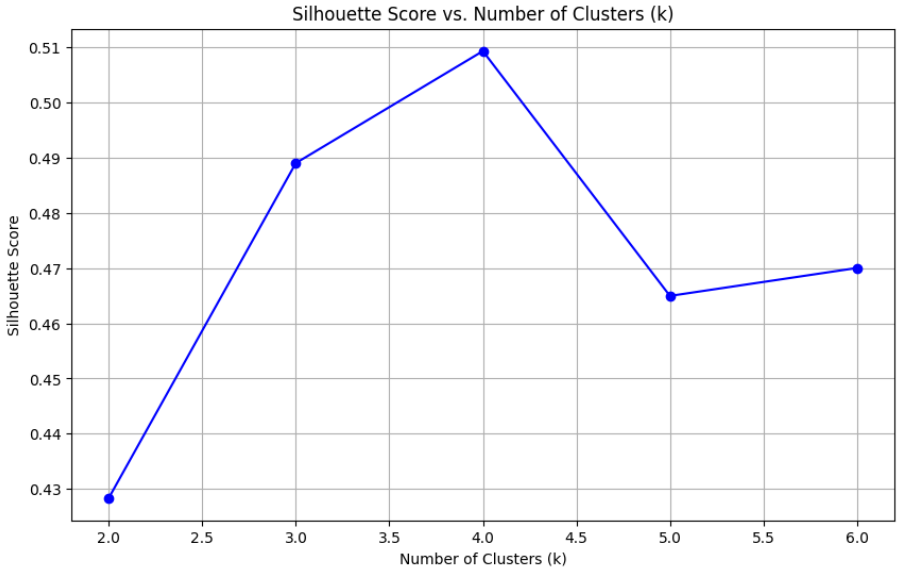

Figure 2: The Trend of Silhouette Score Based on the Number of Clusters

3. Results & Discussion

Figure 3 and Table 2 show the results of the clustering of the data. The results present that clusters 0 and 1 are characterized by rear-end collisions during clear and sunny weather conditions at 16:00, suggesting a correlation with the late afternoon traffic rush. The main difference between these two clusters is road configuration. Cluster 0 distinguishes itself by accidents occurring at non-intersections without traffic control, hinting at potential driver distraction or misjudgment in stopping distances. In contrast, Cluster 1 is marked by accidents at traffic signal-controlled intersections, which may indicate that issues such as delayed driver reactions to traffic signals due to increased congestion during peak hours could be contributing factors.

Cluster 2 shares the 16:00 timing with the first two clusters but stands apart due to its snowy conditions and consequent wet road surface, emphasizing the impact of adverse weather on driving safety. Similar to Cluster 0, these rear-end collisions also occur at non-intersections with no traffic control, suggesting that drivers might be unprepared for the sudden change in road conditions, leading to accidents. Moving to Cluster 3, the time shifts slightly to 17:00, possibly indicating the tail end of peak traffic conditions. The collision scenario remains consistent with rear-end impacts; however, the accidents in this cluster are compounded by rainy weather and wet road surfaces at intersections with traffic signals, pointing towards potential visibility issues or challenges in judging stopping distances in slippery conditions.

When considering vehicle and driver demographics across the clusters, a consistent theme emerges with all incidents involving light-duty vehicles from the early 2000s, driven by young individuals aged 17 to 19. Although driver gender varies—female drivers being involved in the first three clusters and a male in the last—the narrow age range suggests that driver inexperience could be a significant factor in these accidents.

It becomes evident that engineering interventions could profoundly impact reducing the incidence of rear-end collisions, which our results have identified as the most prevalent type of accident. The development of driver support systems that mitigate the risks of rear-end collisions must be a priority. Such systems, including automatic braking when a vehicle approaches another too closely, are already in various stages of implementation. However, their efficacy must extend across all weather conditions. There are some research articles that claim that the driver-assist system fails in rainy conditions (Jhung and Kim 2021). The system should be able to work properly, presenting unique visibility and stopping distance challenges.

The data also suggests that improvements in tire technology, road surface materials, camera, and sensor technology could significantly reduce accidents, particularly in adverse weather conditions that contribute to slipping and providing better visuals for drivers. Engineering advancements in tire tread designs that enhance grip and road pavements that facilitate better water runoff could be crucial in preventing collisions.

Table 2: Mode Value for Each Cluster & Variables

Cluster	Hour	Collision Configuration	Road Configuration	Weather	Road Surface	Road Alignment	Traffic Control	Vehicle Type	Vehicle Year	Driver Sex	Driver Age
0	16	Rear-end Collision	Non-intersection	Clear & Sunny	Dry Normal	Straight & Level	No Controlled Traffic	Light Duty Vehicle	2002	Female	18
1	16	Rear-end Collision	Intersection	Clear & Sunny	Dry Normal	Straight & Level	Traffic Signal	Light Duty Vehicle	2002	Female	17
2	16	Rear-end Collision	Non-intersection	Snowing	Wet	Straight & Level	No Controlled Traffic	Light Duty Vehicle	2000	Female	19
3	17	Rear-end Collision	Intersection	Raining	Wet	Straight & Level	Traffic Signal	Light Duty Vehicle	2000	Male	18

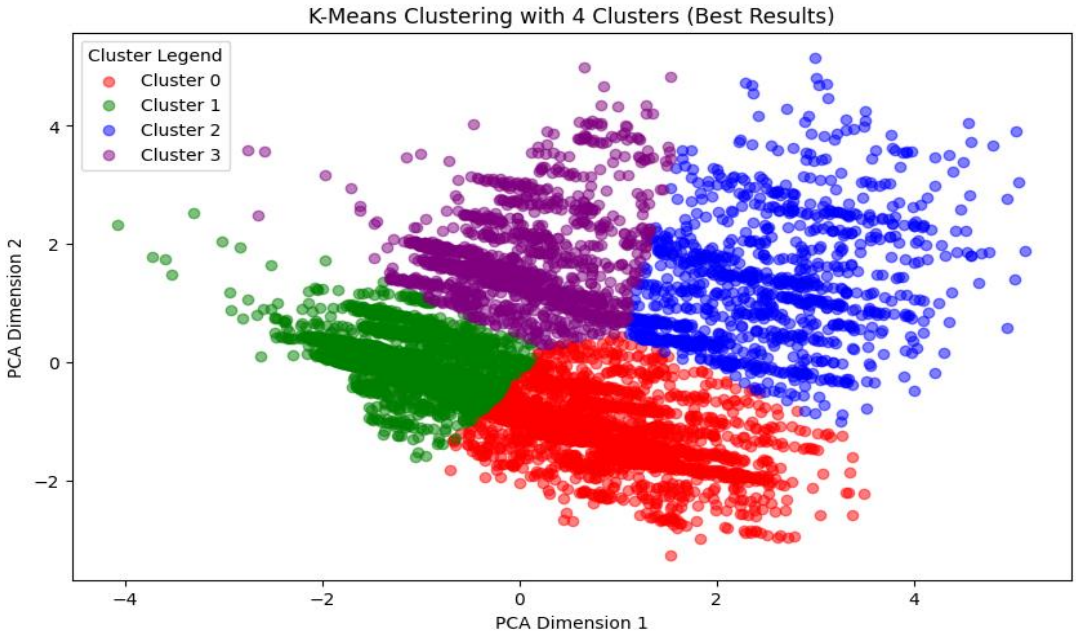

Figure 3: Clustering Result

4. Conclusion

This project analyzes the road accident data in Canada to find out how technologies can contribute to reducing the number of accidents. Using an unsupervised machine learning method, k-means clustering, the results provide a deep understanding of the cause and features of accidents. I find that the most prevalent accidents were rear-end collisions in any weather conditions. Therefore, there is a need to develop a system to prevent rear-end collisions, which already has been employed in many vehicles. However, there is a need to upgrade this system so that it is capable of preventing collisions in any weather condition. Another possible approach is tire and pavement improvements to ensure vehicles stop properly.

There are many other new technologies that could reduce accidents in the future. For example, connective vehicle technology. This system helps vehicles communicate with each other so that they can warn drivers about potential hazards and their surrounding information. Even though it is required to conduct a more detailed analysis to investigate how engineering approaches and technologies can decrease car accidents, this study provides potential fields engineers should focus on.

Reference

Aljaban, Mohamed. 2021. “Analysis of Car Accidents Causes in the USA.” Rochester Institute of Technology.

Jhung, Junekyo, and Shiho Kim. 2021. “Behind-The-Scenes (BTS): Wiper-Occlusion Canceling for Advanced Driver Assistance Systems in Adverse Rain Environments.” *Sensors* 21 (23): 8081. https://doi.org/10.3390/s21238081.

Nikam, Swapnil Kisan. 2020. “ANALYSIS OF US ACCIDENTS AND SOLUTIONS.” California State University, San Bernardino.

Shi, Xiupeng, Yiik Diew Wong, Chen Chai, Michael Zhi-Feng Li, Tianyi Chen, and Zeng Zeng. 2022. "Automatic Clustering for Unsupervised Risk Diagnosis of Vehicle Driving for Smart Road." *IEEE Transactions on Intelligent Transportation Systems* 23 (10): 17451–65. https://doi.org/10.1109/TITS.2022.3166838.

Suarez-Del Fueyo, Rocio, Mirko Junge, Francisco Lopez-Valdes, H. Clay Gabler, Lucas Woerner, and Stefan Hiermaier. 2021. "Cluster Analysis of Seriously Injured Occupants in Motor Vehicle Crashes." *Accident; Analysis and Prevention* 151 (March): 105787. https://doi.org/10.1016/j.aap.2020.105787.

World Health Organization. 2018. "Global status report on road safety 2018. Geneva". Licence: CC BYNC-SA 3.0 IGO.

World Health Organization. 2015. "Global status report on road safety 2015"