Team 3: Aakash Hariharan, Abhilasha Singh, Trisha Singh, Vishal Fulsundar

DATS 6103: Summary Report

Professor: Divya Narula

December 17, 2024

**US Highway Railroad Crossing Accident**

Railroad crossings are vital intersections where highways and railroads meet, creating unique challenges for traffic safety. Accidents at these crossings are a significant concern, often resulting in fatalities, severe injuries, and property damage. Our team selected this topic due to the critical need to enhance safety measures and reduce the frequency of such incidents. Despite advancements in transportation infrastructure, railroad crossing accidents remain a persistent risk, particularly in the United States.

The study of **US Highway Railroad Crossing Accidents** aims to understand the causes, impacts, and prevention strategies for these incidents. This involves analyzing variables such as crossing infrastructure, highway user behavior, train characteristics, environmental conditions, and the effectiveness of warning systems. These factors play a crucial role in identifying risk patterns and implementing targeted safety solutions.

The **US Highway Railroad Crossing Accident Dataset** serves as an essential resource for this analysis. It offers detailed information, including railroad specifications, incident records, geographic locations, crossing types, user behaviors, train attributes, environmental conditions, and warning system performance. By leveraging this dataset, researchers can uncover the root causes of accidents, evaluate existing safety measures, and design strategies to improve railroad crossing safety. Data visualization further enhances the process, turning complex data into actionable insights that inform decision-making and promote safer outcomes at railroad crossings across the United States.

**SMART Questions**

Smart questions are the foundation of effective inquiry and problem-solving. They help clarify goals, focus efforts, and improve the quality of insights, making them essential for any research or data-driven project. For our dataset, we have developed the following smart questions, primarily focused on prediction:

1. **How can we predict the severity of a driver’s injury in a railroad crossing accident using external factors?**
2. **How can we identify accident-prone locations in the USA based on accident frequency over the past 46 years?**
3. **How can we predict the presence of crossing warning signs during railroad accidents using historical data?**

**Literature Review**

Highway-railroad crossing accidents remain a critical safety concern in the U.S., causing significant fatalities, injuries, and economic losses annually. Research highlights various contributing factors to these accidents, such as driver inattention, risk-taking behaviors, and poor visibility during adverse weather conditions (Liu et al., 2012). The disparity in size and weight between trains and vehicles exacerbates the severity of collisions, leading to serious injuries and fatalities.

Improved technology and safety systems have proven effective in reducing accidents. Automated warning systems, enhanced signal systems, and risk reduction measures, as highlighted by Arthur D. Little, Inc. (1996), play a crucial role in mitigating incidents. Additionally, studies on derailment factors emphasize the importance of infrastructure improvements and risk assessments in reducing hazardous material transportation risks (Barkan et al., 2003).

Freight train operations also influence accident rates. Empirical analyses by Zhang et al. (2022) indicate that derailment rates vary between unit trains and manifest trains, emphasizing the need for tailored risk management strategies. Despite advancements, safety remains a challenge, necessitating ongoing improvements to warning systems, visibility enhancements, and public awareness initiatives.

**Dataset Overview**

The dataset, titled "Highway-Rail Grade Crossing Accident Data," was sourced from Kaggle and contains 46 years of historical data on railroad crossing incidents across the United States. Initially, the dataset included 239,487 observations and 141 variables. However, many variables were highly correlated, either presenting the same features in different ways or conveying the same information with slight variations. As a result, we removed the redundant columns. The final dataset now contains 120,365 observations and 46 variables, covering information on railroad incidents, location details, highway and crossing specifics, vehicle data, train information, and environmental conditions.

**Preparing The Data**

As part of preprocessing the data we implemented the following steps:

**Handling Missing Values:** Missing data can compromise the integrity of the dataset, making it less reliable for analysis. In our dataset, we found no missing values.

**Managing Outliers:** Outliers can affect model accuracy and statistical analysis. Given that each incident in our dataset is crucial for identifying causes, we retained all data points, including outliers, to preserve valuable information.

**Removing Duplicate Rows:** Duplicate entries can skew analysis results. We identified and removed four duplicate rows from our dataset.

**Data Type conversion:** We converted the relevant categorical variables into factorial values to facilitate better analysis.

**Data Analysis**

**EDA and Graphs:**

Data visualization serves as a critical tool to transform this dataset into actionable insights. Let’s begin by visualizing the US Highway Railroad Crossing Accident Dataset to uncover meaningful patterns and trends.

**1.** **State Analysis:** Texas has the significantly highest number of incidents, followed by Illinois and Indiana. In contrast, Hawaii and District of Columbia have almost no incidents.

A graph of a number of income

Description automatically generated

**Possible reason:** The size of the rail network, population density, the amount of freight vs. passenger traffic, and the geography and climate of each area. Larger states like Texas, Illinois, and Indiana have bigger rail systems, which can lead to more incidents due to the challenges of managing such large networks. On the other hand, Hawaii and DC have smaller, more manageable rail systems, which results in fewer incidents.

**2. Yearly and Monthly Analysis:** The number of incidents was significantly higher in 1980, but there has been a notable decline in the following years. January and December record the highest number of incidents, while April, May, and June see a slight decrease of about 3,000 incidents.

A graph of a distribution of data

Description automatically generated A graph of a number of blue lines

Description automatically generated with medium confidence

**Possible reason:** The sharp decline in incident numbers since 1980 may be attributed to advancements in technology, infrastructure, safety measures, and regulations. The higher incident counts in January and December could be influenced by winter weather conditions and increased travel during the holiday seasons. Conversely, the drop in incidents during April, May, and June is likely due to better weather, improved track conditions, and a decrease in travel demand, all contributing to fewer accidents during these months.

**3.Weather and Visibility Analysis:** Clear weather is associated with a considerable number of incidents. Furthermore, the visibility data reveals that 50% of incidents occur during daylight hours, while 38% take place at night.

A graph with different colored squares

Description automatically generated A blue circle with different colored numbers

Description automatically generated

**Possible reason:** Clear weather may result in increased train activity, and while there are more incidents during daylight hours due to higher traffic, nighttime incidents are likely caused by reduced visibility, fatigue, and other environmental factors.

**4. Type of Track Analysis:** The majority of accidents took place on the main track, with only a few incidents occurring on industry, yard, and siding tracks.

A graph with different colored bars

Description automatically generated

**Possible reason:** The higher number of accidents on the main track is likely due to its heavier use, increased train traffic, and higher speeds, all of which elevate the risk of incidents. In contrast, industry, yard, and siding tracks see less traffic and slower operations, reducing the likelihood of accidents. These tracks are designed for localized purposes, resulting in fewer incidents overall.

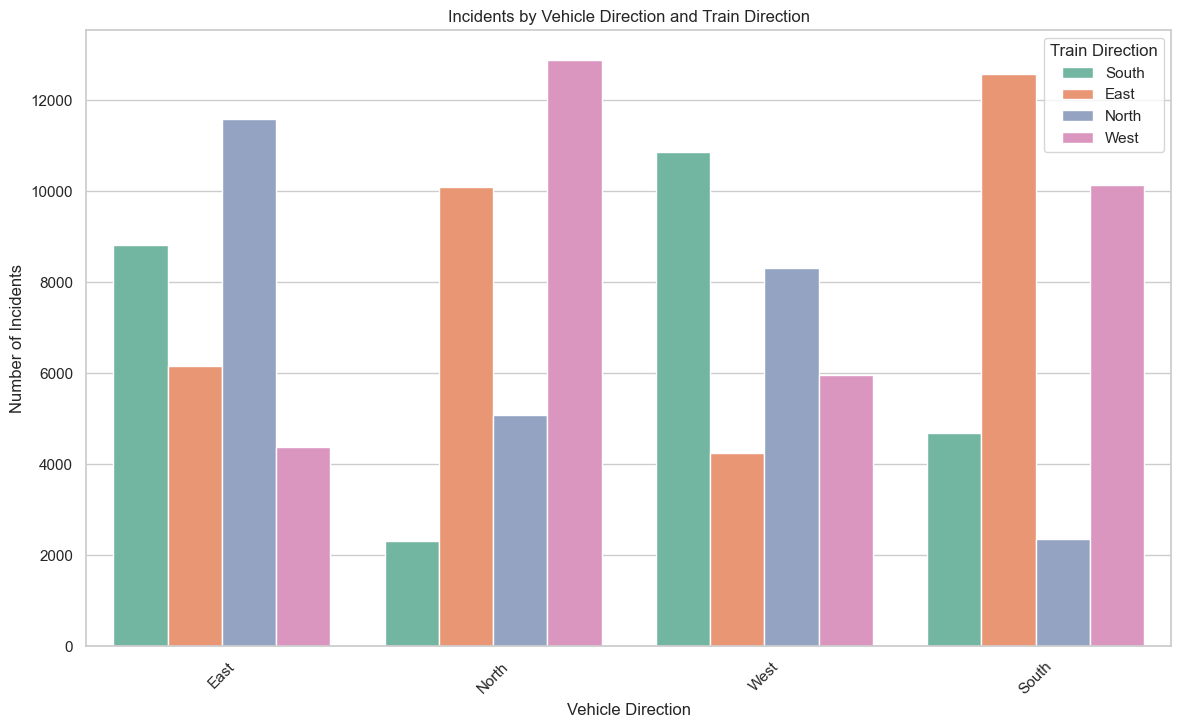
**5.User action and type of highway user analysis:** The graph shows that most accidents occurred due to vehicles failing to stop, with autos being the most involved type, having the highest accident count.

A graph with different colored bars

Description automatically generated

**Possible reason:** The failure to stop is often linked to driver inattention, distractions, or misjudgment, especially at railroad crossings. Autos, being the most common type of vehicle on the road, are naturally more involved in accidents. Additionally, drivers of autos may be less aware of crossing signals or may underestimate the danger of approaching trains, leading to a higher incidence of accidents compared to other vehicle types.

**6. Vehicle Direction and Train Direction Analysis:** The graph shows that most accidents occurred when the vehicle and train were traveling in opposite directions, rather than running parallel to each other.



**Possible reason:** Accidents are more likely when vehicles and trains are traveling in opposite directions because drivers may misjudge the timing or distance of the oncoming train, leading to a failure to stop in time. In contrast, when vehicles and trains are running parallel, drivers have more time to assess the situation and make decisions, reducing the risk of collisions.

**Addressing Key Questions through Model Development**

**Conclusion**

**References**

Arthur D. Little, Inc. (ADL). Risk Assessment for the Transportation of Hazardous Materials by Rail, Supplementary Report: Railroad Accident Rate and Risk Reduction Option Effectiveness Analysis and Data, 2nd rev.ADL, Cambridge, Mass., 1996.

Barkan, C. P. L., C. T. Dick, and R. Anderson. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *In Transportation Research Record: Journal of the Transportation Research Board*,No. 1825, Transportation Research Board of the National Academies,Washington, D.C., 2003, pp. 64–74.

Liu, Xiang & Saat, Rapik & Barkan, Christopher. (2012). Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates. *Transportation Research Record: Journal of the Transportation Research Board*. 2289. 154-163. 10.3141/2289-20.

Zhang Z, Lin C-Y, Liu X, et al. An Empirical analysis of freight train derailment rates for unit trains and manifest trains. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit.* 2022;236(10):1168-1178. doi:10.1177/09544097221080615