Maryna Botas

ACHIEVEMENT 6
Advanced Analytics & Dashboard Design

Exercise 6.01

RAILWAY INCIDENT DATASET (1975-2022)

Data Source: The Railway Incident Dataset, published by the Federal Railroad Administration (FRA) Office of Railroad Safety, provides comprehensive data on railway incidents from 1975 to 2022. This dataset is available on Kaggle and is licensed under the public domain, making it an open-source resource.

The dataset can be accessed on Kaggle via the following link: Railroad Accident & Incident Data.

This dataset is licensed under the public domain license available at http://www.usa.gov/publicdomain/label/1.0/.

Reasons for Database Choice

This dataset meets all project requirements:

It is provided by a state agency, ensuring accessibility and reliability.

All columns selected for analysis have clear names and descriptions.

The dataset covers the period from 1975 to 2022.

Four continuous variables selected for analysis are 'Train Speed,' 'Total Damage Cost,' 'Total Persons Killed,' and 'Total Persons Injured.'

Five categorical variables selected for analysis are 'Accident Type,' 'Passengers Transported,' 'Temperature,' 'Visibility,' and 'Weather Condition.'

The initial dataset contains 215,849 rows.

The geographical component is represented by the State and FRA District.

Data Understanding

Given the extensive nature of the Railway Incident dataset, which contains 160 columns, a focused subset of key variables was selected for the initial analysis.

This selection is driven by the constraints of time and volume constraints of the project, while still providing a robust basis for understanding the nature of railway incidents.

The selected variables are:

Variable	Description
‘Reporting Railroad Code’ and ‘Reporting Railroad Name’	Identifying the specific railway companies involved in incidents helps identify areas for targeted safety improvements and interventions.
‘Report Year’	Analyzing the year of each incident allows for the examination of temporal trends and the effectiveness of safety measures over time.
‘Accident Number’	This unique identifier ensures that each incident can be distinctly tracked and referenced throughout the analysis.
‘Accident Month’	Examining the month of occurrence helps identify seasonal patterns and specific times of the year when incidents are more likely to occur, aiding in seasonal safety planning.
‘Accident Type’	Categorizing incidents by type (e.g., collisions, derailments) allows for a more granular analysis and tailored safety recommendations.
‘State Name’ and ‘FDA District’	Geographic variables are crucial for spatial analysis, helping to identify regions with higher incident rates and understand geographic influences on railway safety.
‘Train Speed’	Analyzing train speed at the time of incidents helps understand its role in accident severity and informs speed-related safety protocols.
‘Temperature’, ‘Visibility’, and ‘Weather Condition’	These variables can be used to assess how weather conditions affect the frequency and severity of incidents.
‘Passengers Transported’	This variable indicates whether there were passengers on the train at the time of the incident. This variable

	provides a quicker way to identify incidents involving passenger trains.
‘Total Damage Cost’	The financial impact of incidents is important for cost-benefit analyses of safety improvements and for understanding the economic burden of rail accidents.
‘Total Persons Killed’ and ‘Total Persons Injured’	The analysis of fatalities and injuries is essential for assessing the incident severity and prioritizing actions to protect human life.

Data Cleaning and Consistency Checks

Renaming columns. All selected columns were renamed according to name standards and logic of the subset.

Original Name	New Name
‘Reporting Railroad Code’	‘railroad_code’
‘Reporting Railroad Name’	‘railroad_name’
‘Report Year’	‘year’
‘Accident Number’	‘accident_id’
‘Accident Month’	‘month’
‘Accident Type’	‘accident_type’
‘State Name’	‘state’
‘District’	‘fda_district’
‘Train Speed’	‘train_speed’
‘Temperature’	‘temperature’
‘Visibility’	‘visibility’
‘Weather Condition’	‘weather_condition’
‘Passengers Transported’	‘has_passengers’
‘Total Damage Cost’	‘damage_cost’
‘Total Persons Killed’	‘persons_killed’
‘Total Persons Injured’	‘persons_injured’

Duplicates. There were indicated and removed 5177 duplicates.

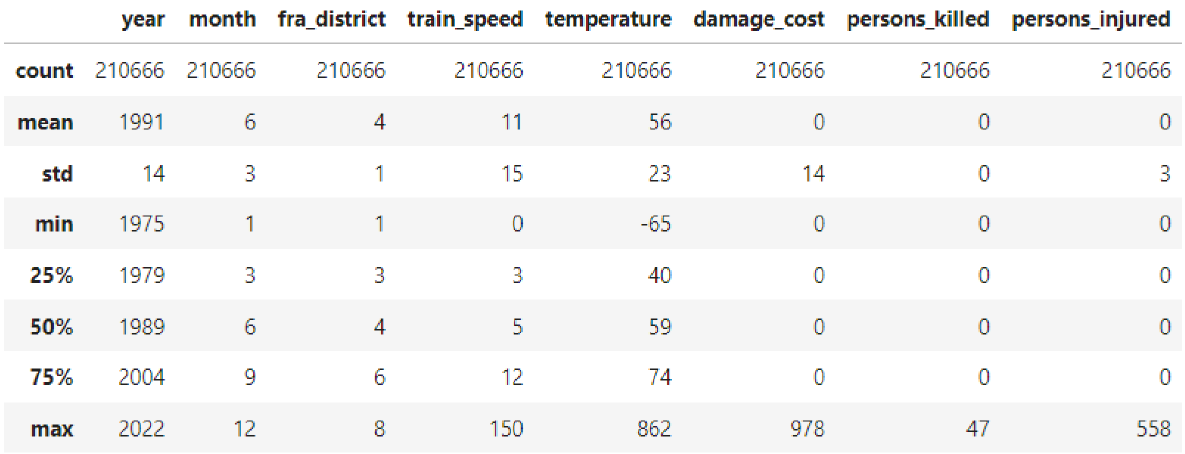
Missing values. The columns 'visibility', 'weather_condition' and 'has_passengers' have the most missing values and object datatypes. Therefore, we replace missing values with common category 'Unknown' for these variables.¶

Other columns have 1 to 3 missing values, and dropping them in a data set with 210672 rows should not skew the data any more than any other way of addressing.

Mixed data types. No mixed data type was detected. Changed the type of the 'year', 'month' and 'fda_district' variables from float to integer. Changed the type of the 'damage_cost' variable from object to float.

Data Profile

Descriptive statistics for numerical variables



	year	month	fra_district	train_speed	temperature	damage_cost	persons_killed	persons_injured
count	210666	210666	210666	210666	210666	210666	210666	210666
mean	1991	6	4	11	56	0	0	0
std	14	3	1	15	23	14	0	3
min	1975	1	1	0	-65	0	0	0
25%	1979	3	3	3	40	0	0	0
50%	1989	6	4	5	59	0	0	0
75%	2004	9	6	12	74	0	0	0
max	2022	12	8	150	862	978	47	558

Descriptive statistics for numerical variables give us the following observations:

- The 'year' variable covers the period from 1975 to 2022, but 50% of the data covers the first 14 years of the period shown.
- The variables 'month' and ‘fra_district’ show the normal distribution.
- Train speeds vary widely from 0 to 150. The median train speed is 5, indicating that at least half of the train speeds are relatively low.
- The maximum value for temperature 862 degrees does not look normal for any thermometer scale. Considering that the USA uses the Fahrenheit scale, values higher than 134 degrees Fahrenheit are impossible. To solve this problem, we use the mean imputation.
- The ‘damage cost’ values range from 0 to $978K. The median damage cost is $59K, indicating that most incidents have relatively low costs, with some high-cost outliers.
- The majority of records of the variable ‘persons killed’ and ‘persons injured’ have 0 value, as indicated by the mean, median, and percentiles all being 0. The maximum values of 47 persons killed and 558 people injured

indicates that there are incidents with significant fatalities or high number of injuries, but these are rare.

Mode and Cardinality:
reporting_railroad_code: Mode = UP, Cardinality = 990
railroad_name: Mode = Union Pacific Railroad Company, Cardinality = 987
accident_id: Mode = 001, Cardinality = 166889
accident_type: Mode = Derailment, Cardinality = 13
state: Mode = ILLINOIS, Cardinality = 50
visibility: Mode = Day, Cardinality = 5
weather_condition: Mode = Clear, Cardinality = 7
has_passengers: Mode = No, Cardinality = 3

Mode and Cardinality for categorical variables give us the following observations:

- The dataset includes a variety of railroad codes, but UP is the most common and indicating that the Union Pacific Railroad Company has the highest number of reported incidents or operations.
- There are 166,889 unique accident IDs in the dataset, which is less than the total number in the clean dataset. Since we have already removed duplicates, this could be a result of matching report numbers in different companies.
- Derailments are the most common type of accident reported, with 13 unique accident types in the dataset.
- Illinois has the highest number of reported incidents, which could be due to a higher volume of rail traffic or more thorough reporting. The cardinality of 50 indicates that the dataset covers all U.S. states.
- There are 4 unique visibility conditions in the dataset and the most common visibility condition is during the day.
- There are 6 unique weather conditions in the dataset and the most common weather condition is clear weather.
- The most common value indicating the presence of passengers is No, which means that the majority of reported incidents do not involve passenger trains.

Limitations

- Data Completeness: Although the dataset spans several decades, there might be gaps or inconsistencies in the data due to changes in reporting standards, methodologies, or incomplete records from earlier years.

- Data Accuracy: The accuracy of the dataset depends on the quality of the initial incident reports submitted by various railway companies. There may be variations in reporting practices and thoroughness.
- Temporal Relevance: Given the dataset covers up to 2022, it may not reflect the most recent changes in railway safety practices or emerging trends after 2022.

Ethical Considerations

- Privacy and Confidentiality: The dataset contains sensitive information such as accident details, fatalities, and injuries. While the data is anonymized, care must be taken to ensure that personally identifiable information (PII) is not inadvertently disclosed.
- Data Use: The use of this data should aim to improve railway safety and inform policy making. Ethical considerations include ensuring that the data is not used to unfairly penalize railway companies or misrepresent the state of railway safety.
- Reporting Bias: Potential biases in the initial reporting process should be acknowledged. Different railway companies may have different levels of diligence in reporting incidents, which could affect the consistency and fairness of the data set.

Exploratory Questions for the Railway Incident Dataset

Trend analysis: How have the number and severity of railway incidents changed over the years from 1975 to 2022? Are there any noticeable trends in incident frequency or severity over specific time periods?

Company-specific analysis: Which railway companies have the highest and lowest incident rates? Are there particular companies with consistently high or low numbers of incidents? (incident rates)

Incident type and frequency: What are the most common types of railway incidents reported? How does the frequency of different types of incidents change over time?

Geographical analysis: Which states, or FDA districts report the highest number of railway incidents? Are there geographic patterns or hotspots for certain types of incidents?

Environmental factors: How do weather conditions (such as temperature, visibility, weather) affect the frequency and severity of railway incidents? Are incidents more frequent during certain weather conditions or seasons?

Operational factors: How does train speed correlate with the occurrence and severity of incidents? Do certain types of the railway incidents occur more frequently at the higher speeds?

Safety and damage analysis: What is the total damage cost associated with railway incidents over the years?
How does the number of people killed or injured in incidents vary by incident type, company, or geographical location?

Temporal patterns: Are there certain months or times of the year when incidents are more likely to occur?