Documentation: Cleaning Rail Equipment Accident Incident Data (Form 54)

My Goal:

The purpose of this cleaning step was to simplify the dataset so that it focuses on the variables that matter for analysis of railroad incomes, logistics, and business impact. I wanted to remove redundant identifiers, technical codes, and metadata that added no analytical value, while retaining financial, operational, and organizational fields. The result is a lean, transparent dataset that is ready for risk, resilience, and business analysis.

Steps Taken

1. Reading the raw dataset

```
df = pd.read_csv('Rail_Equipment_Accident_Incident_Data_(Form_54)_20260121.csv')
```

I loaded the original accident/incident dataset into a pandas DataFrame. The file contained many fields, including multiple accident identifiers, technical codes, and metadata that were not useful for business or logistics analysis.

2. Converting the date column

```
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
```

I converted the Date column into proper datetime format. Using errors='coerce' ensured that invalid entries were safely converted to NaT, which avoids calculation errors in time-based filtering.

3. Filtering by time window

```
df = df[(df['Date'] >= '1/1/2024') & (df['Date'] <= '1/1/2026')]
```

```
df = df.reset_index(drop=True)
```

I restricted the dataset to records between **January 1, 2024 and January 1, 2026**. This narrowed the scope to the most relevant period for current analysis and reduced the dataset size significantly.

4. Dropping redundant or misleading columns

```
cols_to_drop = ['Accident Number','Other Accident Number','Maintenance Accident Number',
'Division','Division Code','Other Railroad Code','Accident Type Code','Visibility Code',
'Weather Condition Code','Track Type Code','Train Direction Code','Equipment Type Code',
'Signalization Code','Method of Operation Code','Adjunct Code 1','Adjunct Code 2','Adjunct Code 3',
'Remote Control Locomotive Code','First Car Initials','First Car Number','First Car Position',
'Causing Car Initials','Causing Car Number','Causing Car Position','Hours Engineers On Duty',
'Minutes Engineers On Duty','Hours Conductors On Duty','Minutes Conductors On Duty','PDF Link',
'Incident Key','Report Key','Special Study 1','Special Study 2','County Code','State Code',
'Temperature','Visibility','Weather Condition','Other Accident Month','Grade Crossing ID']
df = df.drop(columns=cols_to_drop)
```

I removed columns that either:

- Added no analytical value (multiple accident numbers, adjunct codes, internal report keys).
- Were redundant (county/state codes duplicated full names).
- Introduced unnecessary complexity (crew duty minutes, weather codes).
- Were metadata only (PDF links, special study flags).

This step sharpened the dataset to focus on financial, logistical, and organizational variables.

5. Handling missing values

df = df.fillna('NaN')

df.replace('NaN', np.nan, inplace=True)

I first replaced missing values with the placeholder "NaN" to ensure clarity and consistency across the dataset. Then, I converted those placeholders back into proper np.nan values. This two-step process ensures that missing values are standardized and can be correctly interpreted by analytical tools, while avoiding blank cells that might cause confusion.

6. Saving the cleaned dataset

df.to_csv('Clean_Rail_Equipment_Accident_Incident_Data_(Form_54)_20260121.csv', index=False)

Finally, I exported the cleaned dataset. The result is a leaner file that highlights the essentials: damage costs, train logistics, hazmat and evacuation counts, and railroad business identifiers.

Key Decisions and Rationale

- Accident identifiers dropped: Multiple accident numbers and codes were redundant and added no value to analysis.
- Weather and visibility removed: These fields were not central to income/logistics/business analysis and risked distracting from core metrics.
- Crew duty minutiae excluded: Hours/minutes on duty were too granular and not relevant to the financial/logistics focus.
- Business focus sharpened: By keeping railroad identifiers, damage costs, tonnage, and car counts, the dataset now directly supports analysis of operational resilience and business impact.
- Transparency ensured: Every cleaning step was documented, making the dataset reproducible and audit-ready.

Data Source: https://data.transportation.gov/Railroads/Rail-Equipment-Accident-Incident-Data-Form-54-/85tf-25kj