Final Year Project

Data Analysis of Train Delay Factors to Optimise the Railway Network in Great Britain

BSc (Hons) Computer Science

Frances Hatton

12019397

21/05/25

University of Staffordshire

Overview

Research Background

- The efficiency of Great Britain's railway network is challenged by ageing infrastructure, rising demand, weather disruptions, and operational inefficiencies (ORR, 2024). Regions like Wales and Western England are especially affected, prompting strategic improvement plans (Network Rail, 2023).
- This project analyses Office of Rail and Road (ORR) historical train delay data to uncover root causes and recurring patterns.
- Using data analysis techniques, the aim is to generate insights that support better decision-making, optimise scheduling, and highlight areas for infrastructure improvement to enhance reliability of railway network performance.

(ORR, 2024)

(Network Rail, 2023)

Project Aim & Objectives

Project Aim	Project Objectives
Analyse historical train delay data	Retrieve historical train delay data from the ORR.
Transform raw data into valuable insights	Apply descriptive analysis to identify trends and root causes of delays (create content and awareness).
Communicate insights effectively to the audience	Use visualisations tools (e.g., Seaborn, Matplotlib python) to clearly present insights.
Support decision making	Apply diagnostic analysis to support decision making by uncovering causes.
Predict and support maintenance planning and improve scheduling	Apply predictive (ML) models (e.g., decision trees, regression, neural networks) to predict delays.

Hypothesis:
Train delays are influenced by factors like weather, infrastructure, and operations.
Data analysis can reveal hidden patterns and root causes, uncovering trends across lines, regions, or seasons.

Project Deliverables

Deliverable	Description
Literature Review	Summary of research and best practices in data analysis and ML for railways.
Decision-Making Framework	Selection of suitable methods for data preprocessing and predictive modelling.
Data Analysis & Visualisation	Insightful visualisations using Python to communicate delay patterns.
Machine Learning Insights	Predictive models highlighting key delay factors and trends.
Processed Dataset	Cleaned, structured dataset ready for analysis.
Final Report	Comprehensive documentation of methodology, findings, and recommendations.
Presentation	Summary of key insights and outcomes for stakeholders and VIVA presentation.

Literature Review Key Findings

Data Analytics Overview	✓ Enables trend identification, insight generation, and evidence-based decision-making. ✓ Categorised into descriptive, diagnostic, predictive, and prescriptive analytics.
Data Analytics Process (Descriptive)	✓ Involves collection, organisation, analysis, visualisation, and interpretation. ✓ Emphasises data preparation, preprocessing, and feature engineering for model accuracy.
Visualisation Techniques	✓ Tools i.e. Python Matplotlib, Seaborn, and Plotly enhance insight communication and pattern detection. ✓ Interactive dashboards improve real-time decision-making for stakeholders.
Machine Learning & Predictive Modelling (Predictive)	✓ Techniques include decision trees, random forests, regression, neural networks. ✓ Applied for forecasting delays and improving planning and operations.
CRISP-DM Framework	✓ Structured 6-phase approach for managing data mining projects: Business Understanding → Data Understanding → Preparation → Modelling → Evaluation → Deployment.
Prescriptive Insights & Reporting	✓ Translates analysis into actionable strategies aligned with organisational goals. ✓ Supports sustainable improvements through documentation, monitoring, and feedback loops.

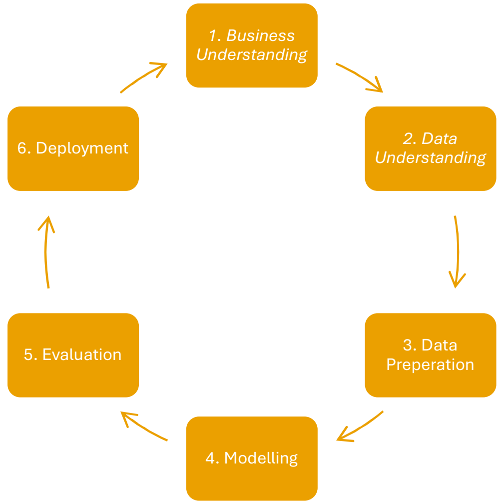
Decision-Making Framework: Research Methodology

This project adopts a hybrid approach which integrates elements of both top-down and bottom-up strategies whilst also following CRISP-DM Framework explained below.

- **Top-Down Approach**: A top-down approach begins with defining high-level goals or problems (often from stakeholders) and then breaks them down into data tasks to align analysis with strategic objectives.

- **Bottom-Up Approach**: A bottom-up approach starts with analysing raw data to uncover patterns and insights, which are then used to inform and shape higher-level strategies and decisions.

Phase	Description
Business Understanding	Identify key delay factors i.e., Network Management, Track, Non-Track, Weather & Structures and External.
Data Understanding	Collected ORR dataset (*Table 3184*), assessed quality, structure and relevance, verified source credibility and data compliance.
Data Preparation	Cleaned & transformed data e.g., handling missing values, feature engineering. Used Python libraries for preprocessing.
Modelling	Built descriptive trend analysis and predictive machine learning models, applied algorithms e.g., regression decision trees, conducted EDA, dimensionality reduction and feature selection.
Evaluation	Assessed model accuracy, validity, alignment with business goals. Incorporated feedback i.e., survey insights.
Deployment	Decided upon the most useful visualisations for stakeholders with the most relevant information which are included in the final report.



CRISP-DM Framework (Jackson, 2002)

Data Sources

Table 3184 - Delay Minutes by Operator and Cause (Periodic)

- Sourced from the ORR Data Portal (ORR Data Portal, 2025).
- Covers delay metrics submitted by Network Rail every 4 weeks (13 times/year).
- Used for monitoring punctuality, performance trends, and delay causes.

Data Quality Assurance:

- Verified via ORR's *Passenger Rail Performance Quality & Methodology Report* (Lunn, 2025).
- Stored in centralised warehouse with strict QA checks and ISO/IEC 27040 compliance (Lunn, 2025).
- Licensed under Open Government License (OGL) for ethical use (Lunn, 2025).

Exploration & Use

- Explored using summary statistics, visualisations, and queries.
- Identified initial patterns, data structure, and potential issues.
- Ensured reliability by handling NaN values and validating integrity for analysis.

Data Preparation

1. Data Cleaning – Removed NaN values, corrected errors, and eliminated duplicates using Pandas functions:

- .dropna() – Remove missing values.
- .fillna() – Impute missing entries.
- .astype() – Convert data types.
- For example, a crucial step in data cleaning is handling missing values. Using the .dropna() function in Pandas, incomplete records, such as those missing values in key columns like 'JPIP_Category_Group_Description' (Delay Causes), 'Financial_Period_Year2' (Year), and 'Adjusted_Pfpi_Minutes5' (Delay Minutes), are removed to ensure data accuracy and prevent skewed analysis results.

2. Feature Engineering – Created/selected variables to enhance model performance.

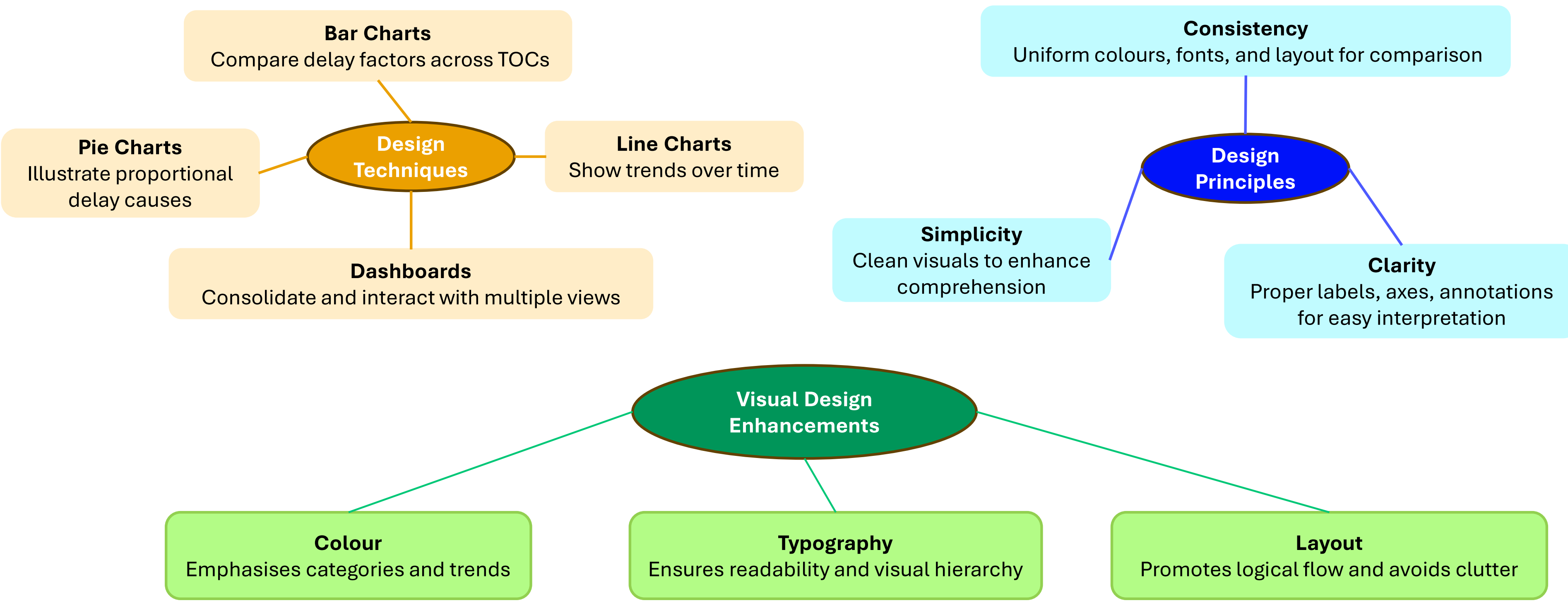
3. Data Formatting – Transformed data types e.g. *Textbox24* from object to float64 for accurate plotting and machine learning use.

📊 Prepared data for visualisation & predictive modelling.

Data Visualisation Approach: Design Methods Used

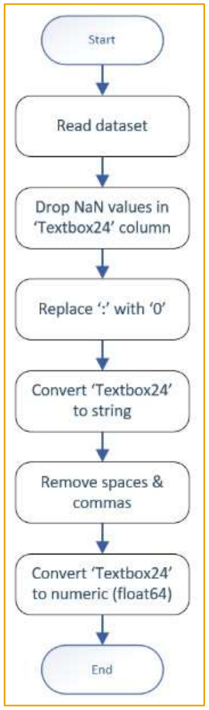
Objective:

Transform complex train delay data into **clear, engaging, and insightful visuals** that support decision-making and performance improvement.



Artefact Development and Implementation

Step	Action
1	Read dataset into a DataFrame (pd.read_csv())
2	Apply .dropna() to remove rows with NaN values in 'Textbox24'
3	Replace all occurrences of ':' with 0
4	Convert all values in 'Textbox24' to strings
5	Strip leading/trailing whitespaces and remove commas
6	Use pd.to_numeric() to convert cleaned values to float64 (with error handling)
7	Confirm successful type conversion using .info()



Outcome:

Prepared dataset enabled accurate visualisation and insight generation for time-series, comparative, and proportional analyses.

Case Study 1: Comparative Time-Series Analysis

Purpose:

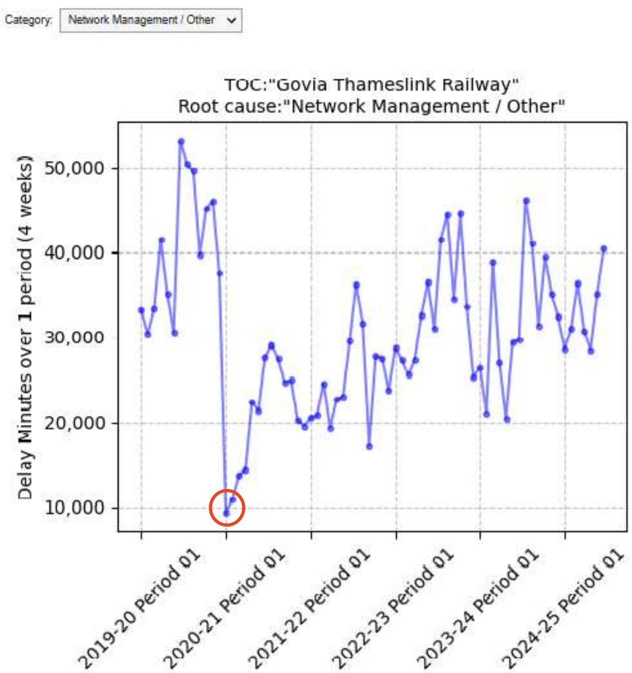
- To identify **trends, fluctuations, and seasonal patterns** in train delay minutes across Train Operating Companies over five years 2019-2025, 7 periods (6 months) amongst 24 TOCs.

Key Insights:

Major delay causes:

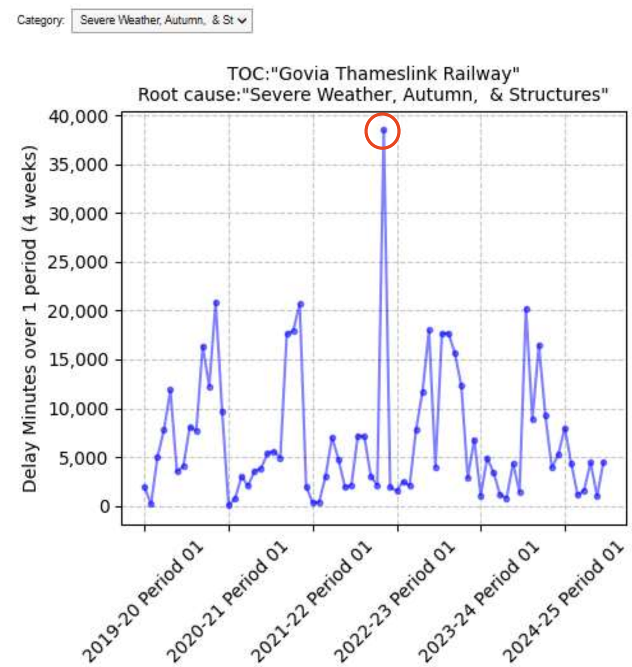
- *Network Management*
- *Severe Weather, Autumn & Structures*
- *Track-related*
- *Non-Track Assets*
- *External Factors*

- **GTR** and Northern Trains show the highest fluctuations across all factors.
- Caledonian Sleeper, Hull Trains, and Heathrow Express maintain low, stable delay levels.
- COVID-19 lockdown led to temporary reductions in delay minutes.
- Seasonal weather causes cyclical delay spikes (winter & summer extremes).Infrastructure resilience and geography greatly impact performance.



Graph 1: Network Management Delays GTR

- Shows fluctuation in delay minutes from 2019–2025.
- Significant drop during 2020-21 likely due to COVID-19 lockdown.
- Gradual rise post-pandemic reflects service resumption challenges.
- Delay peaks indicate recurring operational inefficiencies.



Graph 2: Severe Weather, Autumn & Structures GTR

- Clear seasonal peaks in colder/hotter periods.
- Major spike during 2021-22 correlates with extreme weather events (heatwave).
- Lower average values compared to network management but higher volatility.

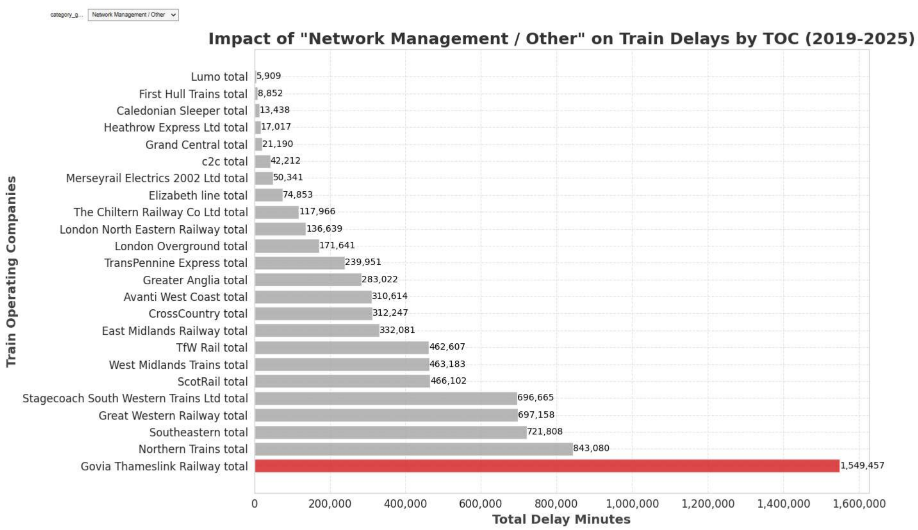
Case Study 2: Comprehensive Analysis of Delay Factors

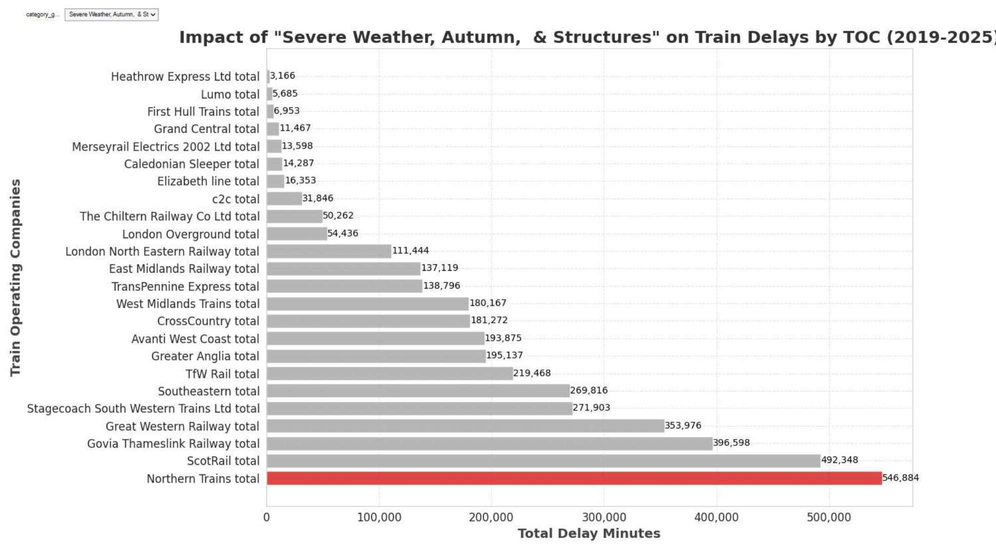
Purpose:

- Compare how various **Train Operating Companies** are impacted by **different delay causes** over a 5-year period 2019-2025**.**

Key Insights:

- GTR consistently records the highest delay minutes across all categories, indicating system-wide vulnerabilities in management, infrastructure, and environment resilience.
- Northern Trains and Great Western Railway also rank highly, especially in weather and asset-related delays, likely tied to network size and geographical exposure.
- Smaller TOCs like Lumo, Hull Trains, and Caledonian Sleeper experience significantly fewer delays, suggesting benefits from simpler routes, new company lacks data history, or lower operational complexity.
- Non-track asset failures and track-related issues (e.g., signalling, power supply, track faults) point to the need for infrastructure upgrades and predictive maintenance.
- The variation in delay causes by TOC shows the need for customised, not one-size-fits-all, strategies for delay reduction.





Graph 3: Network Management Delays (GTR)

- GTR recorded the highest total delays: 1.55 million minutes.
- Northern Trains, Southeastern, and Great Western Railway follow with significant values.
- Smaller TOCs like Lumo and Hull Trains had minimal network management issues.
- Insight: Larger networks face greater coordination and scheduling challenges, highlighting the need for better timetabling and traffic management.

Graph 4: Severe Weather, Autumn & Structures (Northern Trains)

- Northern Trains leads with 547k delay minutes, indicating high exposure to weather-related disruptions.
- ScotRail and GTR also significantly affected, especially in coastal and rural regions.
- Lower delays for operators like Heathrow Express and Lumo, likely due to protected or urban routes.
- Insight: Investment in weather-resilient infrastructure and seasonal response strategies is critical.

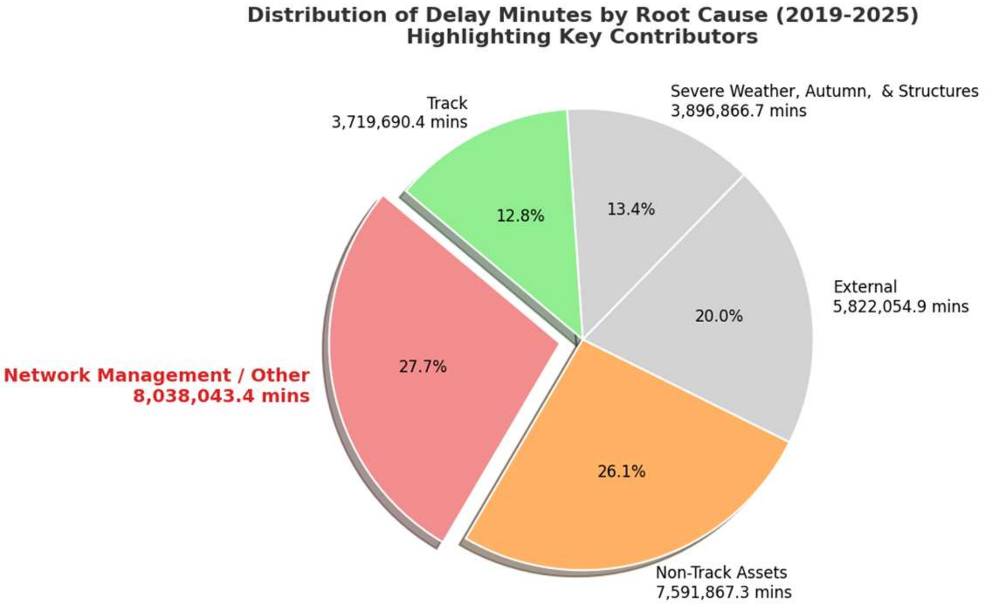
Case Study 3: Proportional Analysis of Delay Factors

Purpose:

- To visualise the relative contribution of each root cause to total train delays which helps prioritise the most impactful issues affecting railway performance.

Key Insights:

- Network Management / Other is the top contributor (27.7%, 8M minutes)
  → Includes scheduling issues, signalling failures, and planning inefficiencies.
- Non-Track Assets account for 26.1% (7.6M minutes)
  → Delays due to faults in stations, telecoms, and electrical systems.
- External Factors contribute 20.0% (5.8M minutes)
  → Includes vandalism, trespassing, road incidents, and power supply faults.
- Severe Weather, Autumn & Structures: 13.4% (3.9M minutes)
  → Seasonal delays from flooding, storms, and low adhesion.
- Track Faults: 12.8% (3.7M minutes)
  → Lower impact, but still significant for performance.
- Focus should be placed on operational efficiency, asset resilience, and external risk mitigation to reduce delays.



Stakeholder Feedback: Primary Research

Method: Structured survey targeting **41 Network Rail professionals** across roles (engineering, management, planning, etc.).

Key Findings:

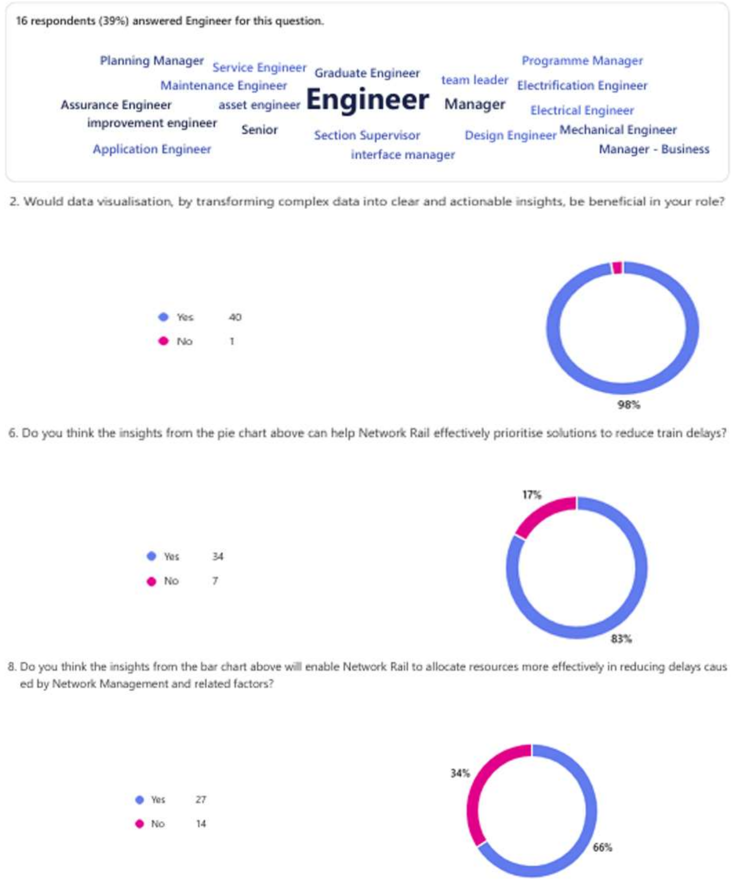
- **98%** agreed data visualisation enhances decision-making in railway operations.
- **83%** regularly experience train delays (Always, Often, Sometimes).

Top Delay Factors Identified:

- Non-Track Assets
- Network Management (e.g. timetable, signalling)
- External factors (e.g. vandalism, fatalities)
- Severe Weather
- Track-related Issues

Insight Usefulness (Visualisation):

- **83%** found ORR pie charts effective for solution prioritisation.
- **66%** agreed bar charts helped with resource allocation.



Applying Machine Learning

Goal: Use ML to predict train delays by modelling the relationship between delay times and key factors. Linear, Polynomial, and Random Forest regressors were tested, Random Forest delivered the highest accuracy, capturing complex, non-linear patterns.

Machine Learning Categories:

Supervised Learning

- Classification (e.g. KNN, Decision Trees, Random Forest)
- Regression (Linear, Polynomial, Random Forest Regressor)

Unsupervised Learning

- Clustering (e.g. K-Means)
- Association & Dimensionality Reduction (e.g. Random Forest, Correlation Filtering)

(Banerjee, 2023)

Machine Learning Workflow

	Stages	Description
1	Define Problem & Collect Data	Identify inputs/outputs, set prediction goal, gather relevant data.
2	Preprocess Data	Clean and format data for analysis.
3	Feature Engineering	Select or create key variables to improve model learning.
4	Model Training	Choose and train ML models using historical data.
5	Evaluate & Validate	Test model accuracy and generalisation using performance metrics.
6	Tune Hyperparameters	Adjust model settings to enhance results.
7	Deploy & Monitor	Implement the model in real systems and track performance over time.

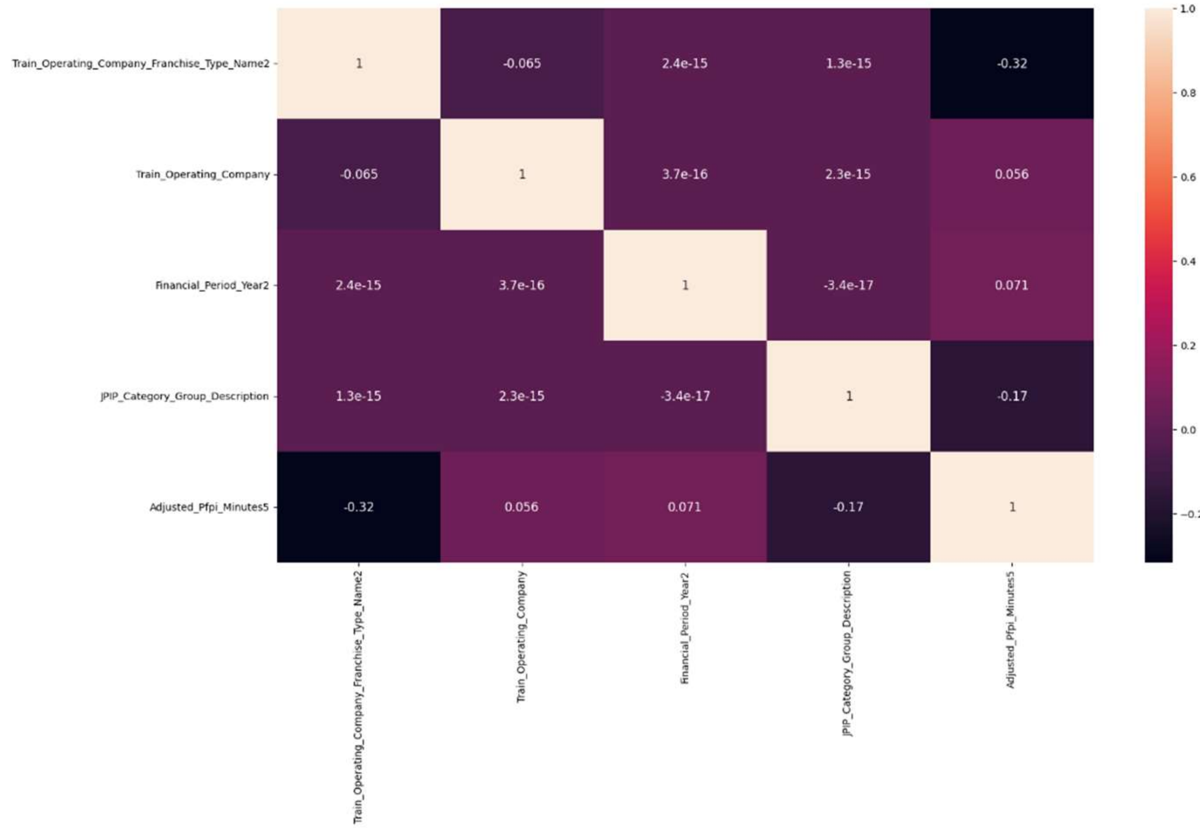
(Chollet, 2018)

Regression Models – Results & Evaluation

Model Performance Comparison

Model	R^2 Score	MAE (min)	RMSE (min)
Linear Regression	0.133	4033	6070
Polynomial Regression	0.208	3843	5803
Random Forest Regressor	**0.763**	**1676**	**3173**

Pearson Correlation Matrix



Critical Evaluation

- **Data Quality Issues** Incomplete, inconsistent or biased data impacts model accuracy (McLoughlin, 2025).
- **Model Interpretability**: Advanced models such as Random Forest can act as ‘black boxes’ (Bashar & Torres Machi, 2024).
- **Implementation Barriers**: High computational costs and lack of expertise (McLoughlin, 2025).
- **Ethical and Legal concerns**: Risks of data misuse, privacy breaches, and algorithmic bias (McLoughlin, 2025).
- **Operational Limitations**: Legacy systems, weather variability, and infrastructure unpredictability can limit machine learning effectiveness.

Project Conclusion

Objective: To analyse train delay factors in Great Britain using data analytics, machine learning and data visualisation.

Approach & Methods:

- CRISP-DM framework used for structured data analysis.
- Descriptive & diagnostic analysis revealed key delay causes:
  - Network Management, Non-Track Asset Failures, Severe Weather, External Incidents.
- Predictive modelling applied using Random Forest Regressor.
  - Achieved R^2 score of 76.31%, showing strong predictive capability.

Key Outcomes:

- Identified patterns & root causes of delays.
- Delivered actionable insights for improving railway performance.
- Validated with industry feedback and primary research.

Impact: Supports proactive maintenance, improved scheduling, and data-driven decision-making for Network Rail and other stakeholders.

References

- Banerjee, J. (2023). Artificial Intelligence Overview. University of Staffordshire.
- Bashar, M. Z., & Torres-Machi, C. (2024). Machine learning to enhance the management of highway pavements and bridges. Infrastructure Asset Management, 11(3), 119–127. https://doi.org/10.1680/JINAM.22.00031
- Chollet, F. (2018). Deep Learning with Python. Manning.
- Jackson, J. (2002). Data Mining; A Conceptual Overview. Communications of the Association for Information Systems, 8, 267–296. https://doi.org/10.17705/1CAIS.00819
- Lunn, M. (2025). Passenger rail performance, Quality and Methodology Report. https://dataportal.orr.gov.uk/media/lvqlsyh3/passenger-performance-quality-report.pdf
- McLoughlin, A. (2025). Bias in Business Data. In Decision Analytics COMP60022. University of Staffordshire.
- McLoughlin, A. (2025). Knowledge Discovery and Management Concepts. University of Staffordshire.
- Network Rail. (2023). England & Wales Strategic Business Plan Control Period 7. https://www.networkrail.co.uk/wp-content/uploads/2023/05/England-and-Wales-CP7-Strategic-Business-Plan.pdf
- Office of Rail and Road. (2024). Punctuality at Recorded Station Stops Data. https://app.powerbi.com/view?r=eyJrIjoiODYyM2FmMGUtMmY1Ni00MTRhLTkxZTAtMWZjOGFkMjVkMDI3IiwidCI6IjIzMjM3OTk2LTdmM2EtNDM5NC04MGY1LTQ2MGNiYzA3NjEzYiJ9
- ORR Data Portal. (2025). Table 3184 - Delay minutes by operator and cause (periodic) . https://dataportal.orr.gov.uk/statistics/performance/passenger-rail-performance/table-3184-delay-minutes-by-operator-and-cause-periodic/