UK National Rail Analytical Project Report

Instructor:

- Ahmed Abd Ellatif (G1)

Data Analysis Team:

- Asmaa Ibrahim
- Heba Essam
- Maha Muhammed
- Mahmoud Abd El-Ghani
- Mahmoud El-Sherif
- Ahmed Sakr
- Mohammad Hussain

Project Overview

This project focuses on analyzing UK National Rail journeys to uncover patterns in passenger behavior, ticket purchasing, travel performance, and operational efficiency. Using a cleaned dataset of train ride transactions, the analysis aims to identify key trends in ticket types, pricing, travel times, delays, and refund requests. The insights generated help demonstrate practical data-analysis skills, including data preparation, dashboard design, and the extraction of meaningful conclusions from real-world transport data.

Dataset Description

The dataset used in this project contains **31,653 train ride transactions** from the UK National Rail network during 2024. It includes detailed information about each journey, covering ticket purchases, passenger attributes, travel timing, and service performance. The dataset originates from the Maven Rail Challenge and is publicly available on Kaggle.

- **Size**: The dataset contains **31,653 train ride records** with a wide range of journey, ticket, passenger, and performance attributes.
- **Grain**: Each row represents **a complete train-ride transaction** for a single passenger**, including ticket purchase details and the journey's actual outcome**.
- **Purpose**: The dataset is designed to support **operational analysis**, including ticketing behavior, travel performance, delay patterns, and customer refund activity. It provides the foundation for building dashboards that reveal key insights about passenger habits and railway efficiency.

Key Components of the Dataset:

- **Transaction & Purchase Details:**
  Columns: Transaction ID, Date of Purchase, Time of Purchase, Purchase Type (Online or Station), Payment Method.
- **Passenger & Ticket Information:**
  Columns: Railcard type, Ticket Class, Ticket Type, Ticket Price.
- **Journey Information:**
  Columns: Departure Station, Arrival Destination, Date of Journey, Scheduled Departure Time, Scheduled Arrival Time.
- **Performance & Delay Metrics:**
  Columns: Actual Arrival Time, Journey Status (On Time, Delayed, Cancelled), Reason for Delay.
- **Customer Response:**
  Columns: Whether a Refund Request was submitted.

Dataset Highlights:

- Covers a broad range of ticket types and purchase behaviors.
- Includes both scheduled and actual arrival times, enabling delay analysis.
- Provides rich categorical data suitable for segmentation and performance insights.
- Ideal for dashboarding and operational analytics in Power BI.

Data Limitations

- The dataset covers only **four months** of the year, limiting the ability to analyze full-year seasonality, such as summer travel peaks, holiday periods like Thanksgiving, and long-term monthly trends.
- The dataset lacks a user demographics **or customer dimension table**, making it challenging to analyze **individual passenger behavior**, loyalty patterns, or segmentation by user characteristics.
- Journey records do not include **distance or route-length information**, preventing analysis linking **delay duration to journey distance**, route complexity, or operational constraints.

Key Diagnostic Questions

The objective of this analysis is to provide actionable insights into the operational performance, financial health, and customer behavior across the UK National Rail journeys dataset. The following questions frame the scope of this investigation:

I. Operational Performance & Reliability

1. **Punctuality vs. Benchmark:** How does the "On Time" performance directly compare to national industry punctuality standards (e.g., PPM)? Is the remaining non-punctual margin indicative of acceptable volatility or systemic failure?

2. **Root Cause Identification:** What are the most dominant causes of service delay or cancellation, and what is the percentage impact of the single most significant factor (e.g., Signal Failure) on overall service availability?
3. **Service Disruption Severity:** To what extent does the delay and cancellation rate exceed acceptable industry benchmarks, and which routes or time periods are most affected by these severe disruptions?

II. Financial Health & Cost of Disruption

4. **Revenue Stream Segmentation:** How is the total revenue distributed across key passenger segments, particularly focusing on the split between different **Ticket Types** (Advance, Anytime, Off-Peak)?
5. **Financial Leakage from Refunds:** What is the total monetary cost and rate of refunds, and how does this 'cost of disruption' correlate with the volume of delayed or cancelled services?
6. **Refund Policy Effectiveness:** Given the Incident Refund Rate, does the current refund mechanism effectively manage customer expectation, or does it represent an area of significant preventable financial risk?

III. Customer Behavior & Growth Momentum

7. **Channel Preference:** What is the preferred channel for ticket purchasing (Online vs. Station), and does this preference correlate with a specific journey type, Railcard usage, or ticket value?
8. **Growth Sustainability:** With a strong Month-over-Month Growth, how can the findings regarding operational failures be leveraged to ensure service reliability scales effectively to sustain this positive financial trajectory?

Data Cleaning & Preparation

During the preparation phase, several transformations and enhancements were applied to ensure data quality and analytical value using Power Query:

Data Cleaning

- Promoting Headers
- Changing Column Types
- Performed **text trimming and capitalization** across key categorical fields to standardize formatting.
- **Merged the Departure and Arrival Station** columns to create a unified **Route** field for easier journey-level analysis.
- Standardized inconsistent values in the **Delay Reason** column. For example:
    - Unified variations such as *“Weather”* and *“Weather Conditions”* into a single category.
    - Consolidated staffing-related issues, such as *“Staffing”* and *“Staff Issue,”* under one consistent label.

Data Enrichment & Dimension Tables

Several new dimension tables were created to improve modeling and enable a more structured, star-schema design:

- **Stations Dimension Table**
  Contains Station Code, Latitude, Longitude, Region, County, and Network Rail Region. These attributes were sourced from official UK rail station information.
- **Routes Dimension Table**
  Generated from unique Departure and Arrival combinations in the primary dataset to support cleaner journey-level filtering.
- **Railcards Dimension Table**
  Grouped railcards into standardized categories such as **Adult**, **Senior**, and **Disabled**, enabling more precise segmentation.

- **Time Dimension Table**
  Built using **M code in Power Query**, containing hour-based and hierarchy-friendly fields used for time-of-day and scheduling analysis.

Exploratory Data Analysis (EDA)

An initial exploratory analysis was conducted to understand the data's structure and distribution. This included examining ticket type frequencies, pricing patterns, travel demand across weekdays, and variations in trip duration. Delay trends were also explored to identify when disruptions are most likely to occur. The EDA provided a clear understanding of passenger behavior and highlighted patterns that shaped the design of the final dashboards.

Data Modeling

- The data model follows a **star schema design**, with *UK Train Rides* as the central flat table and several supporting dimension tables.
- The **Date dimension** connects to the fact table through **active and inactive relationships on Date of Journey and Date of Purchase**, enabling calendar-based filtering and time-series analysis.
- The **DimHour table** is linked to the flat table via **active and inactive relationships** on the Departure and Arrival time fields.
  This relationship is activated in specific time-related measures.
- The **Stations dimension** is connected to the fact table via an **active and inactive one-to-many relationship** on **Departure and Arrival Station Code**, enabling station- and geography-based analysis.
- The **Route dimension** has an active relationship with the flat table using **TripID**, allowing route-level filtering and breakdowns.
- The model includes a **dedicated Measures table** that centralizes DAX calculations, keeping the model structured and maintainable.

- Overall, the schema is designed to optimize performance, enable intuitive navigation, and support flexible slicing across time, routes, stations, ticket types, and railcards.

DAX Measures Overview

A comprehensive set of DAX measures was created to support the report's analytical requirements. These measures enable the calculation of performance KPIs, operational metrics, financial indicators, and month-over-month comparisons. The measures are organized to ensure clarity, reusability, and efficient reporting throughout the Power BI model.

Core Aggregation Measures

- Total Revenue
- Total Rides
- Avg Revenue Per Day
- Avg Rides Per Day
- Avg Ticket Price

Operational Performance Measures

- Average Ride Time
- Average Delay Time
- On-Time Rides
- Delayed Rides (%)
- Cancelled Rides (%)
- Service Failure Rides (%)
- Incidents Delay Time
- PPM% (Public Performance Measure)

Passenger Segmentation Measures

- **Railcard Passengers**: Counts journeys associated with railcard holders.
- **Non-Railcard Passengers**: Counts journeys without railcards.
- **Revenue by Railcard**: Calculates total revenue generated per railcard category.

Busyness Metrics

- Busiest Arrival Time
- Busiest Departure Time
- Busiest Arrival Count / Departure Count

Refund & Customer Behavior Measures

- Refunded Tickets
- Refund Amount
- Overall Refund Rate (%)
- Incidents Refund Rate (%)

Month-over-Month (MoM) Measures

- **MoM Revenue Growth %**
- **MoM Revenue Indicator**
- **MoM Rides Change %**
- **MoM Rides Indicator**
- **Revenue LM / Rides LM**: Revenue and ride counts for Last Month.
- **Revenue MTD / Revenue Share %** – Month-to-date revenue and its share of total.

Dashboard Walkthrough: Visual Structure

1. Executive Summary Dashboard

The purpose of this page is to provide a high-level overview of the project's most critical financial and performance metrics.

- **Key Performance Indicators (KPIs) & Cards:**
  - Ticket Sales (Total Rides)
  - Total Revenue
  - Refund Amount (£)
  - Refund Rate (%)
  - On Time (%)

- **Performance & Trend Charts (General):**
    - Revenue Over Time (Multi-Line Chart with KPI Indicators)
    - Rides Status (Pie Chart)
- **Segmentation & Specialized Charts:**
    - Top 5 Route By Revenue (Bar Chart)
    - Ticket Types Revenue (Pie Chart)
    - RailCard Holder Rides (Bar Chart)
    - Card Status (Treemap)

2. Flow Analysis Dashboard

This page focuses on passenger movement, time-based trends, and the spatial distribution of rides.

- **Trend & General Charts:** Rides Over Time (Multi-Line Chart with KPI Indicators)
- **Behavioral & Specialized Charts:**
    - Rides by Day and Hour of Day (Combined Matrix and Column Charts)
    - Departure & Arrival Stations Rides and Behavioral Trends (Bar Chart)

3. Rail Performance Dashboard

Dedicated to service reliability, tracking disruptions, and identifying the root causes of incidents.

- **Key Performance Indicators (KPIs) & Cards:**
    - Month-to-Date (MTD) Rides KPI Card
    - Ont Time Rides %
    - Average Ride Time
    - Average Delay Time (min)
- **Trend & General Charts:**
    - On Time and Incidents Over Time (Multi-Line Chart)
    - Incidents (Pie Chart)
- **Root Cause & Specialized Charts:**

    - Incidents Reason (Stacked Bar Chart) - *Crucial for root cause analysis.*
    - Incidents per Route (Bar Chart) - *Critical for identifying problematic routes.*

4. Sales Dashboard

A deep dive into revenue generation, ticket types, and the financial contribution of various passenger segments.

- **Key Performance Indicators (KPIs) & Cards:**
    - MTD Rides KPI Card
    - MTD Revenue Card
    - Refund Cards: Tickets, Amount, and Incidents refunds %
- **Trend & General Charts:**
    - Revenue over Time with Forecast (Line Chart)
- **Segmentation & Specialized Charts:**
    - Ticket Class Revenue (Pie Chart)
    - Revenue Share by Ticket Type (100% Stacked Bar Chart)
    - Revenue by Route (Bar Chart)
    - Railcard Types Revenue (Column Chart with KPI Indicator)

Key Insights & Findings

This section details the key operational and financial findings extracted from the dataset, benchmarking service performance against established UK rail industry standards.

Key Performance Figures (from dataset):

- Total Revenue: **£742,000**
- Total Rides: **32,000**
- Delayed Rides: **2,292**
- Delayed Rides (%): **7.24%**
- Cancelled Rides: **1,880**
- Cancelled Rides (%): **5.94%**
- On-Time Rides: **≈ 27,000**
- On-Time Rides (%): **86.82%**
- Refunded Tickets: **1,118**
- Refund Amount: **£39,000**
- Overall Refund Rate (%): **5.53%**
- Incident-related Refund Rate (%): **26.80%**

On-Time Performance in Context

According to the latest release by the Office of Rail and Road (ORR) for Great Britain, the “Time to 3” punctuality measure — i.e., percentage of station stops arriving within 3 minutes of scheduled time — is **86.3%** in the most recent quarter (ORR Data Portal).

The dataset’s on-time rides rate of **86.82%** is therefore **very close** to this national benchmark, suggesting that the sample reflects national-level punctuality performance — at least in the period covered. However, this match should be interpreted cautiously because:

    - The national benchmark is based on *station-stop punctuality*, not necessarily full-journey punctuality.
    - The dataset covers only a limited time span (4 months), which may exclude seasonal variations, major holidays, strikes, or winter disruptions that affect punctuality.

Rout Analysis

- **KGX - YRK:** Highest revenue (~£183K) with 3,922 rides and 329 delay/cancel incidents (8.39%).
- **PAD - RDG:** 3,873 rides generating ~£65K revenue with 352 delay/cancel incidents (9.1%).
- **STP - BHM:** 3,471 rides, ~£53K revenue, and 273 delay/cancel incidents (7.87%).
- **LIV - EUS:** 1,097 rides producing ~£113K revenue but extremely high disruptions: 879 incidents (80%).
- **MAN - LIV:** Highest rides count (4,628) but low revenue (~£17K) and 644 delay/cancel incidents (13.9%).
- **EUS - BHM:** 4,209 rides, ~£50K revenue, and 543 delay/cancel incidents (12.9%).

Total Rides VS. Delay and Cancellation Timelines

- January recorded **8,111 journeys**, reflecting increased travel following the Christmas and New Year holidays.
- February saw a drop to **7,644 journeys**, partly attributable to having only **29 days**, which naturally reduced total ride volume compared to January and March.
- March recorded the highest journey count (8,117**)** across the four months.
- Journey volume declined again in April to **7,781 journeys**.
- Journey distribution across weekdays aligns logically with **workdays vs. weekends**, reflecting expected demand fluctuations.
- Journeys by day-of-month show a **noticeable dip in the first two days**, likely due to early-January public holidays that depress totals across all months.

- The last few days of each month show lower counts because **February has only 29 days**, which reduces cumulative values for days 29–31.
- January recorded the **lowest delay & cancellation rate (12.8%)**, while March showed the **highest rate (14%)**, coinciding with its peak journey volume.

Conclusions

- A substantial majority of rides (≈ 87%) arrive on time, indicating generally reliable service during the sampled period.
- Total delayed rides (~7.2%) and cancelled rides (~5.9%) represent non-trivial proportions and may point to operational inefficiencies or external disruptions.
- The refund rate (5.53% overall and 26.8% for incidents) suggests that a significant share of delays/cancellations result in compensation claims or refunds — implying customer dissatisfaction or disrupted journeys.
- Combining revenue data with delay/cancellation statistics could provide a basis for financial risk analysis, e.g., estimating the costs of service failures, lost rides, and refund payouts.
- The **LIV-EUS** line stands out as a significant outlier, with an exceptionally high disruption rate of **80%**, indicating severe operational challenges likely tied to infrastructure constraints, congestion, or staffing issues.
- Despite high ridership, lines such as **MAN-LIV** and **EUS-BHM** experience elevated disruption rates (**13–14%**), suggesting systemic reliability issues that disproportionately affect frequent commuter routes.
- Conversely, **KGX-YRK**, **PAD-RDG**, and **STP-BHM** lines demonstrate **more stable operations**, maintaining moderate disruption levels (~8-9%) while contributing meaningfully to revenue.
- The mismatch between **revenue and disruption severity**—particularly on **LIV-EUS**—indicates that high-value routes are at risk of customer dissatisfaction, compensation payouts, and long-term revenue erosion.

- Variations in total journeys across months are strongly shaped by **seasonal travel behaviors** and **calendar-related factors**, particularly the number of days in each month.
- The reduced journey count in February is structural, mainly rather than demand-driven, reinforcing the importance of **contextualizing monthly comparisons**.
- The peak in March journeys places **additional operational pressure** on the network, which correlates with the higher delay & cancellation rate recorded during the same month.
- The consistently lower journey counts on the first two days and the last three days of each month suggest that **holiday effects and month-length differences** significantly influence travel patterns.
- Comparing punctuality ratios across months highlights that **higher traffic volumes tend to coincide with more disruptions**, particularly in months with heavier operational loads.

Recommendations / Actions

- **Prioritize operational recovery on the LIV-EUS corridor,** focusing on root-cause diagnostics such as track capacity, timetable pressure, and fleet availability to bring disruption levels closer to national benchmarks.
- **Implement targeted timetable adjustments** on high-frequency lines (MAN-LIV, EUS-BHM) to reduce congestion-induced delays and improve punctuality during peak hours.
- **Enhance resource allocation**—especially rolling stock maintenance and crew scheduling—on routes with consistently high disruption rates to stabilize day-to-day performance.
- **Introduce revenue-protection strategies** on high-value lines like KGX-YRK by maintaining punctuality, as even modest reliability improvements can drive substantial financial impact.

- **Deploy predictive maintenance and incident early-warning analytics** using historical disruption patterns, with priority given to lines exceeding 12% disruption rates.
- **Engage with infrastructure partners** to address structural bottlenecks, particularly on the northbound long-distance routes where disproportionate delays are most evident.
- **Adjust resource allocation dynamically** during high-volume months, such as March, to ensure sufficient rolling stock and crew availability, thereby reducing operational strain.
- **Reinforce operational readiness during post-holiday surges**, particularly early January and late March, to maintain punctuality despite elevated demand.
- **Introduce predictive scheduling models** that account for expected seasonal peaks and calendar-driven variations, helping proactively manage congestion and reduce disruption risks.
- **Implement targeted monitoring** for days with consistently lower journey volumes (e.g., the first two days of each month) to optimize under-utilized capacity and identify opportunities for service balancing.
- **Strengthen incident-prevention measures** during months with historically high journey volumes, as increased traffic has been strongly correlated with disruption rates.

Forecasts

Projected Revenue for Q2–Q3 2024

- The revenue forecast for **May 2024** is **£193,761**, with an upper bound of **£220,000** and a lower bound of **£167,597**.
- Revenue levels are projected to **remain stable through August**, staying within the same forecast range.
- The projected Q2–Q3 average is **higher than April's actual revenue (£187,782)**, indicating a moderate upward trend.
- However, these forecasts remain **below the strong performance observed in January (£199,618)** and **March (£195,147)**.
- Overall, revenue is expected to stabilize at a **mid-range level**, suggesting steady but not peak-period performance during these months.

Appindex

Data Model Measures (DAX Formulas)

The following DAX measures serve as the foundation for the analysis, categorized by the key business areas they address:

1. Core Volume & Financial Metrics (Base Metrics)

These measures calculate the fundamental counts and monetary values for the entire dataset.

Measure Name	DAX Formula	Description
Total Rides	COALESCE(COUNTROWS('UK Trains Rides'),0)	Total number of individual train journeys/transactions in the dataset.
Total Revenue	SUM('UK Trains Rides'[Price])	Total revenue generated from all ticket sales.
Refund Amount	CALCULATE([Total Revenue],'UK Trains Rides'[Refund Request] = "Yes")	Total monetary value of tickets for which a refund was requested.
Refunded Tickets	CALCULATE([Total Rides],'UK Trains Rides'[Refund Request] = "Yes")	Total number of tickets for which a refund was requested.

PPM %	0.86	A hardcoded benchmark for the Public Performance Measure (PPM).

2. Operational Performance & Punctuality

These measures quantify service reliability, delays, and failures, which are crucial for analyzing operational efficiency.

Measure Name	DAX Formula	Description
On-Time Rides	[Total Rides]-[Cancelled Rides]-[Delayed Rides]	Total number of journeys that were not delayed or cancelled.
On-Time Rides (%)	DIVIDE([On-Time Rides],[Total Rides],0)	Percentage of total journeys that were completed on time.
Delayed Rides	CALCULATE([Total Rides],'UK Trains Rides'[Journey Status] = "Delayed")	Total count of rides categorized as 'Delayed'.
Delayed Rides (%)	DIVIDE([Delayed Rides],[Total Rides],0)	Percentage of total rides categorized as 'Delayed'.
Cancelled Rides	CALCULATE([Total Rides],'UK Trains Rides'[Journey Status] = "Cancelled")	Total count of rides categorized as 'Cancelled'.

Cancelled Rides (%)	DIVIDE([Cancelled Rides],[Total Rides],0)	**Percentage of total rides categorized as 'Cancelled'.**
Service Failure Rides	[Delayed Rides]+[Cancelled Rides]	**Combined count of disrupted journeys (Delayed or Cancelled).**
Service Failure Rides %	[Delayed Rides (%)]+[Cancelled Rides (%)]	**Combined percentage of disrupted journeys.**
Average Ride Time	AVERAGE('UK Trains Rides'[Ride Time])	**Average scheduled duration of all train journeys.**
Average Delay Time	AVERAGE('UK Trains Rides'[Delay Time])	**Overall average duration of delay across all services.**
Avg. Incidents Delay Time	DIVIDE(SUM('UK Trains Rides'[Delay Time]),[Delayed Rides],0)	**Average delay time specifically for journeys that were delayed.**

3. Financial & Revenue Analysis

Measures focused on pricing, revenue distribution, and the financial impact of customer refunds.

Measure Name	DAX Formula	Description
Avg Ticket Price	DIVIDE([Total Revenue], [Total Rides], 0)	Average revenue generated per ride.
Overall Refund Rate %	DIVIDE([Refunded Tickets],[Total Rides],0)	Percentage of total rides that resulted in a refund request (overall rate).
Incidents Refund Rate %	Calculate(DIVIDE([Refunded Tickets],[Total Rides],0),'UK Trains Rides'[Journey Status] IN {"Cancelled", "Delayed"})	Refund rate specifically calculated on disrupted services (Cancelled or Delayed).
Revenue Share %	DIVIDE([Total Revenue],CALCULATE([Total Revenue], ALL('UK Trains Rides')))	Proportion of current revenue compared to the total revenue for context/filters.

Revenue by Railcard	CALCULATE([Total Revenue], ALLEXCEPT('UK Trains Rides', 'UK Trains Rides'[Railcard]))	Total revenue segmented by the passenger's Railcard status.

4. Time Intelligence & Growth Metrics

Measures used for tracking performance trends and comparing data over different time periods.

Measure Name	DAX Formula	Description
Revenue LM	CALCULATE(SUM('UK Trains Rides'[Price]),DATEADD('Date'[Date], -1, MONTH))	Total Revenue for the Last Month (used for MoM calculations).
Rides LM	CALCULATE([Total Rides],DATEADD('Date'[Date], -1, MONTH))	Total Rides for the Last Month (used for MoM calculations).
Revenue MTD	TOTALMTD(SUM('UK Trains Rides'[Price]),'Date'[Date])	Total Revenue accumulated Month-To-Date.
MoM Revenue Growth %	DIVIDE([Total Revenue] - [Revenue LM],[Revenue LM])	Percentage change in Total Revenue Month-over-Month.

MoM Rides Change %	DIVIDE([Total Rides] - [Rides LM],[Rides LM])	Percentage change in Total Rides Month-over-Month.

5. Segmentation & Passenger Demographics

Measures used to segment the passenger base by specific attributes, such as Railcard usage.

Measure Name	DAX Formula	Description
Rail Card Passengers	CALCULATE(COUNT('UK Trains Rides'[Transaction ID]),'UK Trains Rides'[Railcard] <> "None")	Count of passengers who purchased a ticket using any type of Railcard.
Non-Rail Card Passengers	CALCULATE(COUNT('UK Trains Rides'[Transaction ID]),'UK Trains Rides'[Railcard] = "None")	Count of passengers who did not use a Railcard for the transaction.