UK Railway Analysis

رواد مصر الرقمية

Presented by:

Omnia Ibrahim Shehata

Gehad Medhat Ali

Nour Walied Ibrahim

Ammar Mohamed Abdelbaset

Muhammed Ateff Muhammed

Mahmoud Khamis Sultan

Agenda

Contents

1. Introduction

This project delves into the intricate world of railway operations by meticulously analyzing a substantial dataset encompassing 31,653 records across 18 key attributes. Far beyond a simple collection of numbers, this data paints a rich and detailed picture of train movements, encompassing a spectrum of operational dynamics and performance indicators. Our mission is to extract meaningful and actionable insights to empower data-driven decision-making within the railway sector, fostering enhanced efficiency and service quality.

Leveraging the powerful capabilities of Microsoft Power BI, we embarked on a comprehensive journey of data refinement, strategic transformation, and insightful visualization. This rigorous process allowed us to dissect complex patterns, pinpoint critical bottlenecks that impede smooth operations, and establish clear visibility into key performance indicators (KPIs) that govern railway effectiveness.

Our analytical framework is thoughtfully structured to provide invaluable support to a diverse group of stakeholders, including operations managers striving for peak efficiency and transportation planners shaping the future of railway networks. By illuminating delay trends with clarity, dissecting performance at the individual station level, and uncovering systemic issues that impact overall service reliability, our analysis aims to provide a holistic understanding of the factors at play.

This documentation serves as a detailed roadmap of our analytical expedition, meticulously outlining each step undertaken – from the foundational data preparation and insightful exploratory data analysis to the creation of dynamic and interactive dashboards that bring the data to life. Furthermore, it synthesizes our key findings into clear and concise takeaways, culminating in practical and impactful recommendations designed to drive tangible improvements in railway operations.

2. Dataset Overview

Spanning the period from December 2023 to April 2024, this rich dataset meticulously chronicles 31,653 individual train journeys, offering a granular perspective on the nuances of railway operations. Beyond mere transactional records, it encapsulates a wealth of interconnected information vital for a deep understanding of passenger behavior and the intricate patterns of service punctuality.

Each journey is characterized by a suite of carefully defined attributes, providing a multidimensional lens through which to analyze performance and customer interactions:

- **Transaction ID:** A unique digital fingerprint assigned to each train ticket purchase, enabling precise tracking and identification of specific transactions within the system.

- **Date of Purchase:** The specific calendar date on which a passenger secured their ticket, providing crucial context for demand forecasting and understanding booking trends over time.

- **Time of Purchase:** The exact moment of ticket acquisition, offering further granularity for analyzing booking patterns and potential correlations with specific times of day.

- **Purchase Type:** Categorizes the point of sale, distinguishing between the convenience of **online** transactions and direct purchases made **at a train station**, illuminating channel preferences and accessibility factors.

- **Payment Method:** Details the mode of payment employed by passengers – **Contactless**, **Credit Card**, or **Debit Card** – offering insights into payment trends and infrastructure utilization.

- **Rail Card:** Indicates the passenger's status regarding the National Railcard program, specifying whether they hold an **Adult**, **Senior**, or

Disabled card, or **None**, allowing for analysis of ridership demographics and the impact of concessionary fares.

- **Ticket Class:** Denotes the level of service selected by the passenger, either **Standard** or **First**, providing a basis for revenue analysis and understanding passenger preferences for comfort and amenities.
- **Ticket Type:** Categorizes the fare structure and usage conditions, distinguishing between discounted **Advance** tickets (requiring purchase at least a day prior), flexible **Off-Peak** tickets (valid outside weekday peak hours of 6-8 am and 4-6 pm), and fully flexible **Anytime** tickets (available for purchase and use at any time), enabling the analysis of pricing strategies and passenger responsiveness to fare incentives.
- **Price:** Represents the revenue generated from each ticket sale, a fundamental metric for financial analysis and performance evaluation.
- **Departure Station:** Identifies the originating station for each train journey, crucial for station-level performance analysis and understanding network flow.
- **Arrival Destination:** Specifies the intended final station for each journey, allowing for the analysis of popular routes and network connectivity.
- **Date of Journey:** The precise calendar date on which the train commenced its travel, forming the basis for temporal analysis of operational performance.
- **Departure Time:** The scheduled time at which the train was intended to leave its originating station, a key reference point for assessing punctuality.

- **Arrival Time:** The scheduled time for the train's arrival at its destination, potentially occurring on the day following departure, is crucial for service planning and passenger expectations.

- **Actual Arrival Time:** The recorded time at which the train effectively reached its destination, which may also fall on the day after departure, serving as the benchmark for delay analysis.

- **Journey Status:** A critical indicator of service delivery, categorizing each journey as **on time**, **delayed**, or **cancelled**, providing a high-level overview of service reliability.

- **Reason for Delay:** When a journey is not on time, this field provides valuable context by specifying the underlying cause of the **delay** or **cancellation**, enabling the identification of recurring operational challenges.

- **Refund Request:** Indicates whether the passenger initiated a request for reimbursement following a service disruption (delay or cancellation), providing a measure of customer impact and service recovery efforts.

3. Data Cleaning and Transformation

To ensure the railway transactions dataset was primed for robust analysis within Power BI, a series of meticulous data cleaning and transformation procedures were executed, focusing on standardization, encoding, and structural refinement:

3.1 Header Promotion and Data Type Refinement

- The initial step involved **promoting the first row to serve as the definitive column headers**, establishing a clear and descriptive naming convention for all attributes.

- Subsequently, each column underwent **rigorous data type conversion** to its most appropriate format. This included:
    1. Representing textual information as **Text**.
    2. Standardizing dates in the **Date** format for accurate temporal analysis (e.g., "Date of Purchase", "Date of Journey").
    3. Formatting time-based data as **Time** (e.g., "Time of Purchase", "Departure Time", "Arrival Time", "Actual Arrival Time").
    4. Ensuring monetary values in the "Price" column were recognized as **Currency** for accurate financial calculations.

3.2 Data Standardization for Consistency

Recognizing the importance of uniformity in categorical data, the "Reason for Delay" column underwent standardization to address inconsistencies in phrasing:

- Variations such as "Signal failure" were uniformly corrected to "**Signal Failure**".
- Related entries like "Weather" and "Weather Conditions" were consolidated into the single, standardized value of "**Weather**".
- The term "Staffing" was consistently replaced with "**Staff Shortage**" for clarity and uniformity.

3.3 Fact and Dimension Tables

In our Power BI model, we structured the dataset using a star schema, where the **fact table** captures transactional data, and **dimension tables** provide descriptive attributes that allow for slicing and filtering the facts. Below is an overview of the dimension tables used in this project :

- **DateTime Dim**

This dimension breaks down the full `DateTime` field into separate components for more flexible time-based analysis:

1. **DateTime**: Full timestamp (e.g., 12/8/2023 12:00:00 AM).
2. **Date**: Date portion only.
3. **Time**: Time portion only.
4. **Shift**: Categorizes the time into “AM” or “PM” shifts, useful for analyzing usage and operations by time of day.

The date range spans from **8th December 2023 to 30th April 2024**, providing a five-month window for temporal analysis.

- **Purchase Type Dim**

Defines the channel through which the ticket was purchased:

1. **Purchase Type**: Includes values like `Online` and `Station`.
2. **PurchaseType ID**: A numerical identifier used to link with the fact table.

This dimension helps track customer behavior across digital and physical sales channels.

- **Payment Method Dim**

Captures the method of payment used during ticket purchase:

1. **Payment Method**: Includes `Contactless`, `Credit Card`, and `Debit Card`.
2. **PaymentMethod ID**: A numerical identifier used to link with the fact table.

Useful for analyzing preferred payment trends and financial data.

- **RailCard Dim**

Stores information about railcard types:

1. **RailCard**: capture types like Adult, Disabled, Senior and None.
2. **RailCard ID**: A numerical identifier used to link with the fact table.

This helps segment passengers based on concessions or eligibility.

- **Ticket Class Dim**

Describes the class of the train ticket booked:

1. **Ticket class**: values include `First Class` or `Standard Class`.
2. **Ticket class ID**: A numerical identifier used to link with the fact table.

Provides insight into customer preferences and revenue breakdowns by class.

- **Ticket Type Dim**

Classifies ticket types:

1. **Ticket Type:** Advance, Off-Peak, Anytime.
2. **Ticket Type ID:** A numerical identifier used to link with the fact table.

Supports operational planning and demand forecasting.

- **Journey Status Dim**

Describes the status of the journey:

1. **Journey Status:** Values like On Time, Delayed, Cancelled.
2. **Journey Status ID:** A numerical identifier used to link with the fact table.

This dimension is key to analyzing punctuality, cancellations.

- **Reason For Delay Dim**

Captures reasons attributed to delayed services:

1. **Reason For Delay**: Signal Failure, Weather Conditions, Technical Issue, Staff Shortage, Traffic or No Delay.
2. **Reason For Delay ID**: A numerical identifier used to link with the fact table.

Facilitates root-cause analysis and mitigation strategies.

- **Refund Request Dim**

Indicates whether a refund request was made:

1. **Refund Request**: Binary (Yes/No).
2. **Refund Request ID**: A numerical identifier used to link with the fact table.

Important for financial and customer service metrics.

- **Station Dim**

Stores metadata about stations:

1. **Departure Station**: e.g., London Paddington, Manchester Piccadilly.
2. **Arrival Destination**: e.g., Liverpool Lime Street, York.
3. **Journey**: Combined route (e.g., London Kings Cross - York).

upports mapping source and destination locations and analyzing station-specific performance.

- **Fact Table**

The Fact Table contains the following critical fields:

1. **Transaction & Journey Details**
   - **Transaction ID**: Unique identifier for each ticket purchase.
   - **Date of Journey** & **Date of Purchase**: Track when journeys were booked and traveled.
   - **Time of Purchase** & **Day of Week**: Analyze booking patterns by time and weekday.
2. **Journey Timing & Delays**
   - **Departure Time**, **Arrival Time**, **Actual Arrival Time**: Measure punctuality and schedule adherence.
   - **Delay** & **Delay (Hr)**: Quantify delays in minutes/hours for performance analysis.
3. **Financial Metrics**
   - **Price**: Ticket price (used for revenue calculations).
4. **Foreign Keys (Connections to Dimension Tables)**
   - **Journey Status ID**: Links to Journey Status Dim (e.g., On Time/Delayed).
   - **Reason For Delay ID**: Links to Reason For Delay Dim (e.g., Signal Failure/Weather).
   - **PaymentMethod ID**, **PurchaseType ID**, **RailCard ID**, **Ticket Class ID**, **Ticket Type ID**, **Station ID**: Connect to respective dimension tables for segmentation.
   - **Refund Request ID**: Flags refunded transactions (Yes/No).

3.4 Removal of Redundant Information

To maintain data integrity and ensure the uniqueness of each recorded transaction, **duplicate rows were identified and subsequently removed** from the dataset.

3.5 Logical Column Reorganization

For enhanced readability and a more structured approach to analysis, the columns within the dataset were strategically reordered. This involved **grouping the original categorical fields alongside their newly created numerical ID columns**, facilitating a more intuitive understanding of the data structure and streamlining the analytical process.

4. Data Modeling

4.1 Why the Star Schema?

The star schema was chosen because it simplifies complex datasets and enhances query performance. It consists of:

- Fact Table: Central table containing measurable, quantitative data (e.g., ticket price, journey time, delays).
- Dimension Tables: Surrounding tables providing descriptive context (e.g., date, ticket type, station, passenger Rail Card).

4.2 Benefits of the Star Schema

- **Improved Performance:** Reduces the complexity of joins during queries.
- **Intuitive Design:** Easier to understand and navigate, especially for non-technical users.

- **Scalability:** Supports growing datasets and evolving business needs.
- **Cleaner DAX Formulas:** Simplifies calculation expressions in Power BI.
  - By implementing a star schema, we ensured the model remained efficient, scalable, and user-friendly.
  - facilitating clearer dashboard interactions and deeper insights.

5. DAX Measures and Calculations

To enhance the analytical capabilities of our Power BI dashboard, we utilized **DAX (Data Analysis Expressions)** to create custom measures and calculated columns. These DAX formulas allowed us to go beyond basic aggregations and unlock deeper insights from the data. From calculating total delays to refund rates, DAX enabled dynamic analysis tailored to our specific business questions.

In this section, we document the key DAX measures developed, explain their logic, and outline their roles within the report visuals and KPIs:

- **Performance and Delay KPIs**:
  1. **Average Delay Duration**: Measures the average length of time trains are delayed, providing insight into the typical impact of disruptions.
  2. **Canceled Percent**: The percentage of train trips that were canceled, indicating the reliability of the service.
  3. **Canceled Trips**: The total number of train trips that were canceled.
  4. **Delayed Percent**: The percentage of train trips that experienced delays.
  5. **Delayed Trips**: The total number of train trips that experienced delays.

6. **Maximum Delay Time**: The longest recorded delay for a single train trip, highlighting the extent of the most severe disruptions.
7. **On Time Percent**: The percentage of train trips that arrived at their destination on schedule, reflecting overall punctuality.
8. **On Time Trips**: The total number of train trips that arrived on time.
9. **PeakPurchaseHour**: The hour of the day during which the highest number of tickets are purchased.

- **Ticket and Revenue KPIs**:

  1. **Average Unit Price**: The average price of a single ticket, providing a general measure of fare levels.
  2. **Frequency of Max Unit Price**: Indicates how often the highest ticket price occurs within the dataset.
  3. **Frequency Of Most Frequent Price**: Indicates how often the most frequent ticket price occurs within the dataset.
  4. **Maximum Unit Price**: The highest price of a single ticket in the dataset.
  5. **Most Frequent Price**: The ticket price that appears most often in the dataset.
  6. **Number of Tickets**: The total number of tickets sold.
  7. **Total Revenue**: The total income generated from ticket sales.
  8. **Top Arrival Destination**: The most frequent destination station among all train journeys.
  9. **Top Departure Destination**: The most frequent departure station among all train journeys.
  10. **Total Journeys**: The total number of train journeys recorded in the dataset.
  11. **Total Refunds**: The total number of refunds issued to passengers.

6. Data Visualization

The UK Train Rides Dashboard is a business intelligence report built using Power BI Desktop, aimed at analyzing and visualizing key metrics associated with train journeys across the United Kingdom. The dashboard consolidates data about journey status, ticket comparisons, delays, and financial coverage to provide railway authorities and stakeholders with actionable insights to enhance operations, passenger experience, and revenue performance.

6.1 Objectives

- Monitor the operational status of train journeys.
- Compare ticket type distribution and sales.
- Analyze service delays and identify bottlenecks.
- Evaluate financial efficiency by examining revenue vs. coverage.

6.2 Dashboard Pages Breakdown

6.2.1 Journey Status Dashboard

- **Purpose:** This dashboard gives a quick overview of train journeys, delays, and passenger patterns to help railway operators improve service and understand traveler behavior.

- **Key Features:**

- Total Journeys: 31,653
- Top Routes: Most journeys start at Manchester Piccadilly and end at Birmingham New Street.
- Peak Purchase Time: 8:25 PM
- Journey Status: 86.8% on time, 7.2% delayed, 5.9% cancelled.
- Rail card Use: Most travelers (20.9K) do not use a railcard.

- Delays: Weather is the main cause of delays across destinations.

- **Insights:**

- Most journeys run on time.
- Birmingham New Street is the top destination.
- Many passengers could benefit from rail cards.
- Delay causes vary by station, but weather leads overall.
- Ticket purchases peak in the evening.

6.2.2 Ticket Dashboard

- **Purpose:** This dashboard explores journey statistics in detail, focusing on delay causes, ticket types, payment methods, and travel routes.
- **Key Features:**

- **Delay Reasons:** Technical issues and staff shortages are the main causes of delays.
- **Payment Methods:** Credit cards are the most used, with high on-time journeys.
- **Purchase Type:** Most tickets are bought online, and they have better on-time performance.
- **Ticket Type:** Advance tickets lead in volume and on-time rates.
- **Journey Breakdown**: Manchester Piccadilly to London Euston is a common route with high punctuality.
- **Yearly Trend**: 2024 shows consistent journey volumes with stable on-time rates.

- **Insights:**

- Advance tickets offer better reliability.
- Online purchases correlate with more on-time journeys.
- Technical and staffing issues need addressing to reduce delays.
- Most of the travel happens on time, despite some delays and cancellations.

6.2.3 Comparison Dashboard

- **Purpose:** Shows overall train punctuality (on-time, delayed, cancelled) and trends**.**

- **Key Visuals:**

- **Big Number Cards:** Display the count and percentage of delayed, cancelled, and on-time trains.
- **Small Line Charts:** Show how the percentage of each status has changed monthly.
- **Filter:** Allows focusing on specific data (currently showing all).

- **Insights:**

- Most trains are on time.
- Delays and cancellations happen but are less frequent.
- The line charts show monthly changes in punctuality.

6.2.4 Delayed Dashboard

- **Purpose:** Analyzes train journey delays, refunds, and station activity.

- **Key Visuals:**

- **Top Left Cards:** Show overall journey stats (total, on-time, cancelled, delayed journeys and their percentages, total refunds, maximum delay time).
- **"Quantifying Delay" Bar Chart:** Shows how different conditions (weather, technical issues, etc.) contribute to delays.
- **"Top Arrival Stations with Delayed Journeys" Bar Chart:** Shows which arrival stations have the most delayed trains.

- **"Delayed Journeys Break Down" Line Chart:** Shows the trend of delayed journeys over recent months.

- **"Stations with the Highest Refund Volumes" Bar Chart:** Shows which stations have the highest amount of refund requests.

- **Insights:**

- Most journeys are on time but delays and cancellations exist.

- Weather conditions are the biggest reported cause of delays.

- Liverpool Lime Street has the most delayed arrivals.

- Delayed journeys peaked in March 2024.

- London Euston has the highest refund volume.

6.2.5 Revenue Dashboard

- **Purpose:** To analyze train ticket revenue based on ticket class, type, rail card usage, payment method, and track revenue trends over time, while also showing overall financial figures.

- **Key Visuals:**

- **Left Side Cards:** Display key financial metrics like the maximum ticket price, average ticket price, peak revenue, total refunds, and total revenue.
- **"Revenue by Ticket Class & Type" Tree map:** Shows how revenue is distributed across different ticket classes (Standard, First Class) and their subtypes (Advance, Off-Peak, Anytime). The size of the boxes represents the revenue amount.
- **"Revenue | Railcard" Bar Chart:** Shows the revenue generated from passengers using different railcard types (No Rail Card, Adult, Disabled, Senior).
- **"Revenue | 2023 - 2024" Line Chart:** Displays the trend of revenue over the past few months (December 2023 to April 2024).
- **"Payment Methods Breakdown | Revenue" Donut Chart:** Shows the proportion of revenue generated by different payment methods (Contactless, Credit Card, Debit Card).
- **Filter (Top Right):** Allows you to filter the data by "Departure Station" (currently showing "All").

- **Insights:**

- Most of the revenue comes from Standard class tickets, specifically Advance purchase.

- Passengers without a railcard contribute the most to revenue.
- Revenue saw a significant increase from December 2023 to January 2024 and has fluctuated since.
- Debit cards are the most common payment method for tickets, generating the largest share of revenue.
- The dashboard allows for filtering by departure station to analyze revenue performance for specific locations.

7. Interactivity and User Controls

- **Interactive Elements Used:**

- Date Slicers
- Category Filters

- **User Navigation:**

Users can click through report tabs to focus on specific areas (e.g., delays, revenue), apply filters to refine data views, and interact with visuals to explore detailed trends.

- **Target Audience**

- Railway Operations Managers
- Finance Teams
- Customer Service Teams
- Marketing and Pricing Teams

8. Conclusion

The UK Train Rides Dashboard offers a robust, interactive platform to monitor operational performance, passenger trends, and financial outcomes for train journeys in the UK. Leveraging the power of Power BI, enables stakeholders to make informed, data-driven decisions to optimize services and enhance the customer experience.