Final Report: UK Train Rides – Digital Egypt Pioneers Initiative (DEPI) Graduation Project

This project was developed by **Team 4: CLS_ONL3_DAT2_G5** as the final graduation requirement for the Digital Egypt Pioneers Initiative (DEPI), Cohort 3, Data Analytics Track. The initiative is under the esteemed patronage of the Ministry of Communications and Information Technology (MCIT).

1.0 Executive Summary

This summary provides a high-level synthesis of the project's entire lifecycle, from the initial business problem to the final strategic recommendations. It encapsulates the core methodologies, key findings, and ultimate value delivered through this data analytics initiative.

The project addressed the critical challenge of the UK rail network's lack of a unified analytics model to efficiently analyze performance drivers across raw, transactional data. The goal was to transform this data into a powerful analytical system using Power BI, unlocking actionable insights into performance, revenue, and demand. This was achieved by engineering a rigorous data pipeline to clean and preprocess a dataset of 31,653 ticket records, constructing a performance-optimized star-schema data model, and developing a multi-page interactive dashboard. The analysis yielded critical insights, pinpointing 'Signal Failure' as the primary operational vulnerability driving the majority of the 1,880 trip cancellations, while simultaneously revealing that a proactive pricing strategy for 'Advance' tickets is the key to maximizing the network's £703.22K in net revenue. The project culminated in the delivery of a comprehensive decision-support system, complete with time-series forecasting models capable of predicting future ride demand and revenue for May 2024, empowering stakeholders to shift from reactive problem-solving to proactive, data-driven management.

This report will now detail the foundational problem statement that guided this work.

2.0 Problem Definition and Project Objectives

This section establishes the strategic business context for the project. It outlines the operational environment of the UK rail network and defines the precise challenges that this analytics system was designed to solve for National Rail stakeholders.

The UK rail network, established in 1825, is the world's oldest and remains one of Europe's busiest, serving 3.5 million passengers daily across a vast network of approximately 2,500 stations. Despite this scale, a core problem persists: railway operators lack a unified analytics model to cohesively understand the critical drivers of performance. The immense volume of raw transactional data makes it difficult to efficiently answer key operational and financial questions regarding the root causes of delays, emerging revenue trends, ticket demand patterns, and station-level performance.

Project Objectives

To address this challenge, the project was structured around four primary objectives:

- **Data Engineering:** To clean, preprocess, and transform raw, unstructured railway journey data into an analysis-ready format.
- **Data Modeling:** To build a robust and validated star-schema data model in Power BI, optimized for performance and scalability.
- **Business Intelligence:** To develop interactive dashboards with key KPIs for delays (target: 87% punctuality), cancellations (target: <6% rate), and financial metrics (e.g., £703.22K Net Revenue).
- **Predictive Analytics:** To create forecasting models for future ride demand, revenue, and ticket class sales.

The following section describes the dataset that served as the foundation for achieving these objectives.

3.0 Dataset Description and Scope

This section details the foundational data that fuels the entire analytical engine. It outlines the source, scope, and key variables of the dataset used to build the models and extract insights.

The analysis is based on a transactional train ticket dataset provided as a single CSV file, representing individual journey records.

The scope of the dataset is defined by the following metrics:

- **Total Tickets Sold:** 31,653
- **Total Trips Analyzed:** 31,653
- **Departure Stations:** 12
- **Arrival Stations:** 32
- **Total Routes Analyzed:** 65

Key Raw Variables

The raw data file contained the following primary fields, which formed the basis for all subsequent transformations and analysis:

- Transaction ID
- Date of Purchase
- Payment Method
- Ticket Class
- Price
- Departure Station
- Arrival Destination
- Journey Status

- Reason for Delay

This raw data required significant transformation to correct inconsistencies and derive new, analytically valuable features before it could be used for modeling.

4.0 Data Transformation and Modeling

This phase represents the "engine room" of the project, where unstructured raw data was systematically forged into a reliable, high-performance analytical model. This process was critical for ensuring the accuracy and integrity of all subsequent findings.

4.1 Data Preprocessing Pipeline

A structured data transformation pipeline was implemented in Power Query to guarantee data quality and consistency. This involved several key stages:

- **Data Quality & Standardization:** The process began with removing duplicate records and rows containing errors. All text-based fields, such as station names, ticket classes, and payment methods, were trimmed, cleaned, and standardized to ensure consistent capitalization and formatting. This step was essential for accurate data grouping, filtering, and model joins.
- **Date/Time Correction:** A critical step was combining separate date and time columns into single DateTime fields. The team applied "midnight-crossing logic"—programmatically adding a day to the arrival date if the arrival time was earlier than the departure time. This was crucial for calculating accurate journey durations and delays for overnight trips.
- **Derived Field Creation:** New features were engineered from the raw data to add analytical depth. Key derived fields include:
  - **Flags:** Boolean fields such as `WasDelayed`, `IsCancelled`, and `ActualArrivalExists` were created to enable simple and efficient filtering and counting for performance analysis.
  - **Metrics:** Core performance metrics like `ActualDelayInMinutes` and `ScheduledDurationInMinutes` were calculated to serve as the basis for all operational KPIs.
  - **Composite Keys:** Composite keys like `Ticket_Key` (combining Railcard-Class-Type) and `RouteKey` (combining Departure-Destination) were created to simplify the data model and facilitate robust, route-level analysis.
  - **Segmentation:** A `FareBucket` field was developed to segment tickets by price range, enabling more granular financial analysis.
  - **Lead Time:** The `LeadTimeDays` metric was calculated to analyze the time between ticket purchase and the journey date, providing crucial insights into passenger booking behavior and the effectiveness of early-booking discounts.

4.2 The Analytical Blueprint: Star Schema

The final data model was built on a validated, optimized star schema architecture. This design ensures fast, efficient querying and provides a clear, logical structure that is easy for analysts and business users to understand.

- **Fact Table (1):**
  - `Fact_Railway`: This is the central table containing all transactional measures, such as `Price` and `ActualDelayInMinutes`, and foreign keys linking to the dimension tables.
- **Dimension Tables (6):**
  - `Dim_Date`: A comprehensive calendar table used for all time-series analysis and trend reporting.
  - `Dim_Tickets`: Contains descriptive attributes of tickets, including `Class`, `Type`, and `Railcard`.
  - `Dim_Stations`: A table detailing all departure and arrival stations.
  - `Dim_Purchase`: Holds information related to the purchase transaction, such as `Purchase Channel` and `Payment Method`.
  - `Dim_Status`: Describes journey outcomes, including the final `Status` and the `Reason for Delay`.
  - `UK_Train_Stations_Coord_Saif`: A supplementary dimension holding the Latitude and Longitude for each station, enabling geospatial analysis and map-based visualizations.
- **Supporting Aggregation Table (1):**
  - `Daily_Summary`: A pre-aggregated table containing daily summaries of key metrics like revenue and ride count. This table was created to improve the performance of high-level dashboard visuals and to serve as an efficient source for the forecasting models.

This robust model serves as the foundation for the key findings presented in the following section.

5.0 Analysis of Findings

This section presents the key findings extracted from the analytical model. It offers a system-wide view of performance and uncovers the critical drivers of revenue, operational disruption, and passenger behavior across the UK rail network.

5.1 System-Wide Operational & Financial Performance

The top-level Key Performance Indicators (KPIs) provide a snapshot of the network's overall health:

Metric	Value
On-Time Rides Percentage	87%
Delayed Trips Percentage	7%
Cancellation Rate	6%
Average Delay Time (Min)	3.25 minutes
Total Tickets Sold	31,653
Net Revenue Generated	£703.22K

While the overall on-time performance is high at 87%, the 6% cancellation rate represents a significant operational disruption. This rate corresponds to **1,880 cancelled trips**, leading to direct revenue loss through refunds and considerable customer dissatisfaction.

5.2 Diagnosing Disruption: Causes and Financial Impact

A deeper analysis reveals the root causes of these disruptions and their specific financial consequences:

- The leading cause of **cancelled** trips is **Signal Failure**, which was responsible for **519** disrupted journeys, followed by Staffing and Weather issues.
- The most frequent cause of **delayed** trips was **Weather**, accounting for **927** separate incidents.
- The analysis uncovered a "hidden cost" of disruptions. While Signal Failure and Weather are the most frequent causes, internal operational issues proved to be the most expensive. **Technical Issues** and **Staffing** problems are the leading causes for refunds, contributing **£15K** and **£12K** respectively to the **£38.70K** total refund amount.
- The route with the highest financial impact from delays was **Liverpool Lime Street -> London Euston**, which accumulated **£13.1K** in lost revenue due to delayed services.

5.3 Uncovering Revenue Streams & Passenger Choices

The model also identified key patterns in revenue generation and passenger purchasing behavior:

- **Revenue by Ticket Type: Advance tickets** are the single most significant revenue driver, generating **£293.6K (41.75%)** of the total net revenue. This indicates a strong passenger preference for pre-booking to secure lower fares.
- **Revenue by Route:** The **London Kings Cross -> York** route is the highest-earning corridor in the network, generating **£179K** in net revenue.
- **Passenger Transaction Flow:** The analysis reveals a dominant passenger profile. The vast majority of journeys (**28,595 of 31,653**) are taken in **Standard Class** by passengers traveling without a Railcard. Within this large segment, 'Advance' tickets are the most popular choice.

These historical analyses provided the necessary foundation for building the forward-looking models discussed next.

6.0 Predictive Analytics: Forecasting for May 2024

Moving beyond historical analysis, the project leveraged the cleaned data to develop time-series forecasting models. These models predict key business metrics for the upcoming month, enabling National Rail management to engage in proactive resource allocation and strategic planning.

The key forecasting outcomes for May 2024 are as follows:

- **Next Month Trip Forecast:** The model predicts a clear increase in the total number of trips for May 2024. This signals a need to prepare for higher passenger demand by ensuring adequate service capacity and staffing levels.

- **Next Month Daily Revenue Forecast:** The model provides a detailed forecast for daily revenue throughout May, complete with confidence intervals (represented by the grey shaded area). This gives management a realistic range for financial planning and performance tracking.
- **Next Month Ticket Demand Forecast:** The system generates forecasts for individual ticket classes, predicting a notable increase in demand for both **Standard** and **First Class** tickets. This granular insight allows for the implementation of targeted pricing, marketing, and resource allocation strategies to maximize revenue from the anticipated surge.

These powerful predictive capabilities are made accessible to end-users through the project's final deliverable: the interactive dashboard system.

7.0 The Interactive Dashboard System

The final deliverable of this project is a comprehensive, multi-page Power BI dashboard. This system is designed to translate complex, granular data into clear, actionable insights for a wide range of stakeholders, from operational managers to financial analysts.

- **Dashboard Pages:** The system is organized into distinct analytical tabs, each designed to answer specific business questions:
  - Overview
  - Financial & Operational Metrics
  - Operational Performance
  - Reasons of Delay & Cancellation
  - Revenue & Price: Segmentation & Flow
  - Revenue & Price: Temporal & Average
  - Statistics & Route: Movement
  - Statistics & Route Analysis: Quality
  - Forecasting & Strategy
- **Key Features:** The user experience is built around intuitive and powerful features, including:
  - Interactive slicers allowing users to filter data by month, day of the week, or route.
  - High-level KPI cards that provide at-a-glance metrics on the network's health.
  - A carefully selected range of visualizations—including bar charts, line charts, and geospatial maps—chosen to best represent the underlying data and reveal hidden patterns.
- **Business Value:** The dashboard's strategic value lies in its ability to provide a unified, dynamic, and near-real-time view of the rail network's health. It empowers stakeholders to monitor performance against targets, diagnose the root causes of operational issues, identify revenue opportunities, and ultimately make faster, more informed, data-driven decisions.

The insights surfaced through this dashboard system directly inform the strategic recommendations that follow.

8.0 Strategic Recommendations

This section translates the analytical findings into a set of actionable, business-ready recommendations. These are designed to improve operational efficiency, enhance the passenger experience, and drive financial performance based on the evidence uncovered in the data.

1. **Prioritize Investment in Signal Infrastructure**
   - **Justification:** "Signal Failure" is the number one cause of all cancellations (519 incidents) and was identified as the primary driver of delays on the network's highest-earning route, 'London Kings Cross -> York'. Investing in signal modernization and maintenance is critical to reducing high-impact disruptions.
2. **Address Core Operational Inefficiencies**
   - **Justification:** While not the most frequent causes of disruption, "Technical Issues" and "Staffing" problems are the most costly. Together, they are responsible for nearly two-thirds of the total refund amount (£27K out of £38.70K). A focused effort on predictive maintenance and workforce management can yield significant financial savings.
3. **Optimize High-Impact Routes**
   - **Justification:** The "Liverpool Lime Street -> London Euston" route should be a primary target for operational review. It suffers from the highest lost revenue due to delays (£13.1K), indicating that performance improvements on this specific corridor will have an outsized positive financial impact.
4. **Leverage Demand Forecasts for Proactive Management**
   - **Justification:** The forecasts for May 2024 predict a clear increase in demand. Management should use these insights to proactively adjust service capacity, plan staffing rosters, and implement targeted marketing campaigns for 'Advance', 'Standard', and 'First Class' tickets to fully capitalize on the anticipated growth in ridership and revenue.

9.0 Project Limitations

To provide a transparent and academically rigorous report, it is important to acknowledge the limitations of this project. These constraints define the boundaries of the current analysis and highlight areas for future enhancement.

- **Data Scope:** The analysis is based on a static, historical dataset covering a specific, finite period. It does not incorporate real-time data feeds, which limits its ability to respond to immediate, ongoing operational events as they unfold.
- **Data Granularity:** While the dataset is rich in transactional detail, it may lack certain external variables that would enable more advanced analysis. For example, the absence of passenger demographics or loyalty program data restricts the potential for sophisticated customer segmentation.
- **External Factors:** The forecasting models are built on historical patterns observed within the dataset. As such, their predictions may not fully account for the impact of unforeseen external events, such as major public holidays not present in the historical data, sudden economic shifts, or large-scale public events that could influence travel patterns.

These limitations should be viewed not as weaknesses, but as clear opportunities for future development and refinement of the analytical system.

10.0 Future Work

This section presents a forward-looking roadmap, outlining key opportunities to build upon the current project's foundation and further enhance the analytical capabilities of the system.

- **Real-Time Monitoring:** The next logical step is to integrate live data feeds from the rail network. This would transform the dashboard from a historical analysis tool into a real-time operational control center, allowing for immediate responses to delays and disruptions.
- **Predictive Maintenance Models:** The historical delay data, especially incidents coded with 'Technical Issue' and 'Signal Failure' reasons, can be used as a training set for machine learning models. These models could predict potential equipment failures before they occur, enabling a shift from reactive to proactive maintenance.
- **Enhanced Customer Segmentation:** Future iterations could incorporate demographic, geographic, or loyalty program data. This would allow for the creation of more granular customer segments, enabling highly targeted marketing campaigns, personalized service offerings, and improved customer retention strategies.
- **Optimization Algorithms:** With a robust historical dataset in place, advanced optimization algorithms could be developed. These models could recommend optimal train timetables and crew rostering schedules based on historical demand and delay patterns, with the goal of minimizing disruptions and maximizing efficiency.

This future work promises to evolve the current system into an even more powerful strategic asset.

11.0 Conclusion

This project successfully engineered a complete end-to-end analytics pipeline, transforming raw, chaotic transactional data into a powerful, interactive decision-support system using Power BI. The analysis successfully identified the primary drivers of operational inefficiency—namely, systemic signal failures and costly staffing shortages—and, for the first time, quantified their direct financial impact on the organization. By moving beyond historical reporting to develop robust predictive models, this work enables a critical shift from reactive problem-solving to proactive, data-informed management. Ultimately, this system transforms complex data into a strategic asset for National Rail stakeholders, providing the insights needed to optimize operations, drive revenue, and improve the travel experience for millions of passengers.