PROJECT DOCUMENTATION

UK TRAIN RIDES DATA ANALYTICS SYSTEM

---

1. Project Overview

1.1 Project Title

UK Train Rides – Data Analytics & Forecasting System

1.2 Project Description

This project aims to build a complete analytics system for UK rail operations. The goal is to transform raw railway journey records into a structured analytical model using Power BI, enabling insights into ridership trends, delays, ticket demand, financial performance, and station-based geospatial analysis. The project was developed as part of the final graduation requirements for the Digital Egypt Pioneers Initiative (DEPI) – Cohort 3, Data Analytics Track using Microsoft Power BI.

1.3 Project Objectives

- Clean and preprocess raw railway journey data.
- Build a validated **star-schema data model** for analytics.
- Create forecasting models for:
  - Ride demand
  - Ticket class demand
  - Revenue
- Provide KPIs for delays, cancellations, punctuality, and financial metrics.
- Develop interactive dashboards, including **map-based visualizations** using station coordinates.
- Deliver full project documentation, source code, and presentation.

---

2. Project Team & Roles

Team Member	Role	Responsibilities
Mohamed Ibrahim (Team Leader)	Lead Data Engineer	Data cleaning, preprocessing, Power Query transformations, data modelling, star schema design, documentation consolidation
Mona Mamdouh	Data Analyst	DAX measures creation, KPI development, preparing data questions, validation
Dina Hassan	BI Developer	DAX measures, dashboard visuals, analysis writing, presentation preparation
Mohannad Abdullah	BI Developer	Contributed to measures and visuals, advanced charts, insight generation
Abdullah Hassan	Dashboard Designer	Power BI dashboard layout, UI/UX design, formatting
Saif Fekry	Data Acquisition Specialist	Collecting station longitude/latitude data, map integration setup, dashboard collaboration with Abdullah

---

3. Project Planning

3.1 Project Proposal Summary

The project focuses on transforming unstructured rail data into a business-intelligent analytical environment using Power BI. The project supports operational and financial decision-making through forecasting, performance measurement, and geospatial insights.

3.2 Timeline (Gantt-Style Overview)

- Week 1: Requirements, Dataset Understanding, Planning
- Week 2: Data Cleaning & Preprocessing, DAX Measures (Mohamed, Mona, Dina, Mohannad)
- Week 3: Data Modelling & Star Schema Design, Dashboard Development (Mohamed, Dina, Mona, Abdullah, Saif)
- Week 4: Documentation, Presentation (Mohamed, Dina, Mona, Mohannad)

3.3 Deliverables

- Clean dataset & model
- Fully documented star schema
- Power BI dashboards
- Presentation file
- GitHub repository
- Documentation file

3.4 Risks & Mitigation

Risk	Impact	Mitigation
Missing or inconsistent data	Delays analysis	Strong preprocessing steps; conditional logic
Incorrect relationships	Wrong KPIs	Proper star schema design
Time constraints	Delivery risk	Parallel task assignment
Common work mistaken as copied	Academic penalty	Clear per-person role separation

---

4. Stakeholder Analysis

Stakeholder	Needs
Passengers (Case-Based)	Better travel experience, fewer delays
Railway Operators	Delay insights, demand forecasting
Management	Financial KPIs, route performance
BI Developers (Team Members)	Clean, well-structured model

---

5. Requirements Gathering

5.1 Functional Requirements

- Import and preprocess raw train ride data.
- Calculate delays, punctuality, revenue, demand, and cancellations.
- Build DAX measures.
- Provide forecasting visualizations.
- Include map-based analysis using station coordinates.

5.2 Non-Functional Requirements

- System must be fast (optimized model).
- Dashboard must be easy to navigate.
- Data transformations must be documented.
- Forecasting must be transparent and explainable.

5.3 User Stories

- *As an operator, I want to view delay patterns so I can improve scheduling.*
- *As a financial analyst, I want revenue metrics and forecasts.*
- *As a planner, I want demand predictions for May.*

---

6. System Analysis & Design

6.1 Problem Statement

The railway operator lacks a unified analytics model for understanding delays, ticket demand, revenue trends, and station-level performance.

6.2 System Goals

- Build a scalable, analytical data model.
- Enable high-quality decision making via Power BI dashboards.

---

7. Data Preprocessing & Cleaning Documentation

A structured preprocessing pipeline was applied in Power Query to prepare the raw ticket-level dataset for modeling. Key steps included:

7.1 Data Quality Fixes

- Removed duplicate rows and records with corrupted values.
- Trimmed, cleaned, and standardized all station names, ticket classes, and categorical text fields.
- Converted Yes/No fields (e.g., Refund Request) into proper Boolean values.
- Standardized data types (Date, DateTime, Numeric).

7.2 Date/Time Corrections

- Combined date and time columns to create:
  - **PurchaseDateTime**
  - **JourneyDateTime_ScheduledDeparture**
  - **JourneyDateTime_Arrival**
  - **ActualArrivalDateTime**
- Applied midnight-crossing logic when arrival time < departure time.
- Calculated:
  - Scheduled duration
  - Actual duration
  - Actual delay in minutes

7.3 Derived Fields

- WasDelayed, IsCancelled, ActualArrivalExists
- FareBucket (price segmentation)
- LeadTimeDays (difference between purchase and journey)
- Ticket_Key and RouteKey composite fields
- Calendar attributes: Year, Month, Week, DayName, IsWeekend

7.4 Integration with Dimensions

- Linked stations with **UK_Train_Stations_Coord_Saif** using Station Name to attach Latitude/Longitude.
- Ensured clean joins with:
  - Dim_Station
  - Dim_Ticket_Class
  - Daily_Summary (via Date Key)

7.5 Fact Table Output (Fact_Railway)

Includes:
TransactionID, datetimes, station fields, ticket attributes, durations, delay metrics, revenue, flags, fare bucket, lead time, and DateKey.

7.6 Validation

- Verified row counts before/after cleaning.
- Checked correctness of midnight adjustments.
- Ensured all DateKey, StationName, TicketClass fields were non-null.
- Compared total price before/after cleaning to ensure no data loss.

8. Database Design & Star Schema

8.1 Data Model Architecture

The final model contains:

Fact Table

- Fact_Railway

Dimension Tables

- Dim_Date
- Dim_Tickets
- Dim_Stations
- Dim_Purchase
- Dim_Status
- UK_Train_Stations_Coord_Saif

Supporting Table

- Daily_Summary

Full details included in final document.

---

9. UI/UX Design

9.1 Dashboard Principles

- Clean layout
- Consistent color scheme
- Icons for clarity
- Interactive slicers
- Optimized for quick insights

9.2 Visuals Used

- Line charts
- Bar charts
- KPI cards
- Maps (using station coordinates from Saif)
- Forecasting visuals

---

10. Implementation Summary

10.1 Tools

- Power BI
- Power Query
- DAX
- GitHub for version control

10.2 Team Implementation Responsibilities

- ✔ Mohamed: Preprocessing, modelling, documentation
- ✔ Mona: Measures + Charts + Handling team meetings
- ✔ Dina: Measures + visuals + Questions
- ✔ Mohannad: Visuals, charts , Analysis
- ✔ Abdullah: Dashboard UI/UX
- ✔ Saif: Coordinates dataset + maps + Dashboard UI/UX

---

11. Testing & Validation

- Data accuracy checks
- Measure validation
- Relationship verification
- Visual accuracy testing
- Performance testing on large visuals

---

12. Final Presentation

Prepared by:

- Dina
- Mona