رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

UK Train

Data Analysis Project

10/14/24

eyouth®

Problem Statement:

- The UK train system faces continuous challenges such as unpredictable delays, fluctuating passenger demand, unclear ticket choices, and revenue uncertainty.
- Passengers struggle to select the best ticket option, while decision-makers lack a unified tool to analyze performance, understand trends, and predict future behaviors.

Proposed Solution:

- We developed an intelligent, data-driven dashboard that analyzes train journeys, predicts operational risks, and provides smart ticket recommendations.
- The system combines traditional analytics with Machine Learning to offer:
- Passenger demand forecasting: predict the volume of passengers per month
- Journey performance prediction: predict cancelled journeys per month
- Operational risk analysis: predict average delay time per month
- Delay prediction and risk scoring
- All integrated into an interactive, user-friendly dashboard that helps both passengers and decision-makers make informed choices.

Unique Value Proposition:

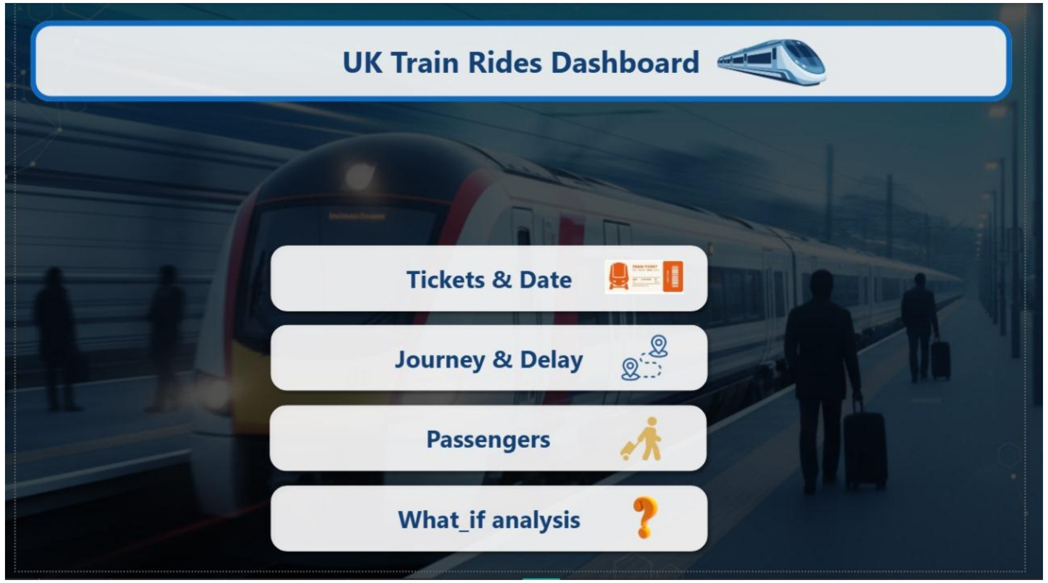
Our project stands out by transforming raw railway data into a clear, interactive, and insight-driven dashboard.

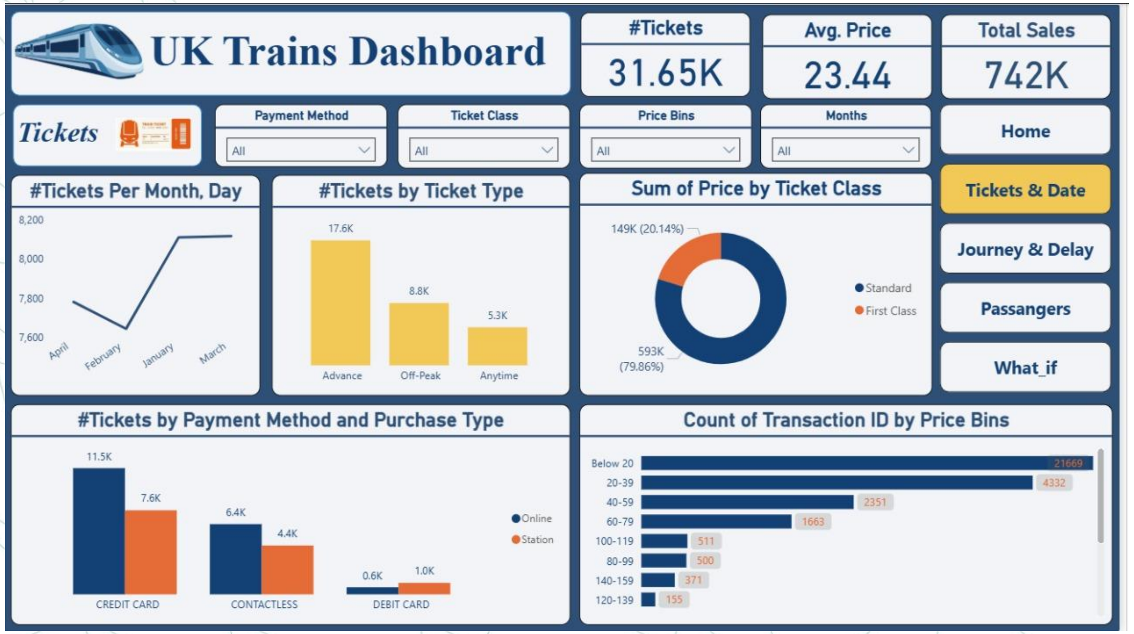
Unlike regular reports, our solution offers:

- Comprehensive trend analysis for passenger demand and journey activity, including predictions for the volume of passengers per month
- Detailed delay insights using fields such as Journey Status, Actual Arrival Time, and Journey Delay (min), alongside predicted average delay time per month
- Forecasting of cancelled journeys per month to help anticipate operational disruptions
- Route performance comparison across different departure and arrival stations
- Ticket analysis by Ticket Class, Ticket Type, Railcard, and Purchase Method
- Revenue and pricing breakdowns supported by Price, Price Bins, and booking times
- A structured star-schema data model that improves accuracy, flexibility, and reporting performance

This makes the system not just a visual dashboard, but a practical tool that helps understand travel patterns, optimize operations, and support data-driven decision-making with predictive insights.

Dashboard:

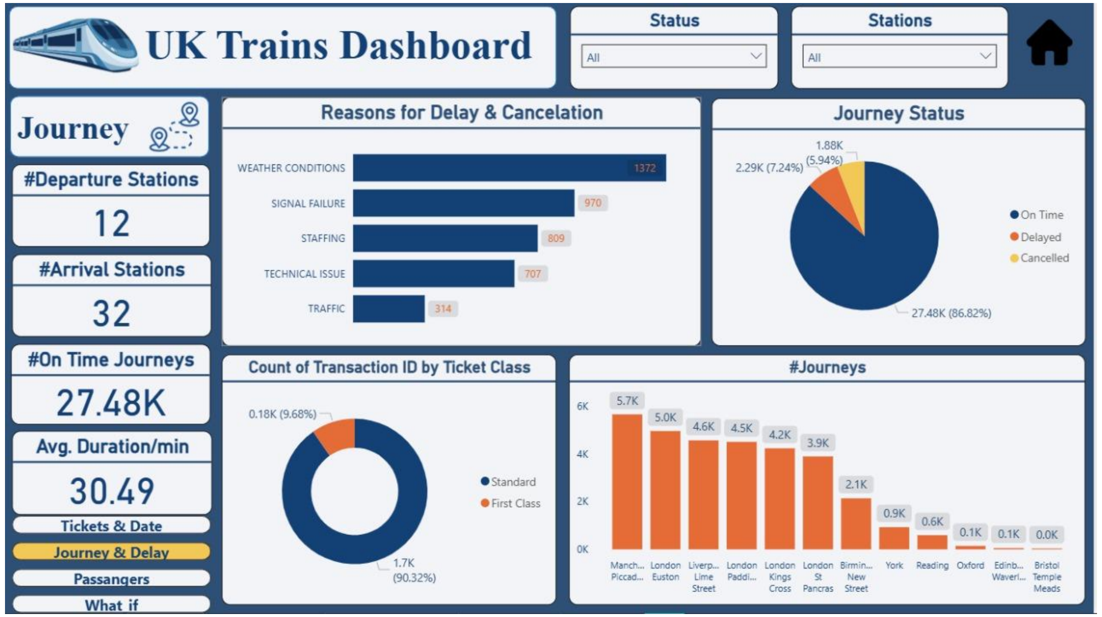



UK Trains Dashboard

#Tickets	Avg. Price	Total Sales
31.65K	23.44	742K

Tickets

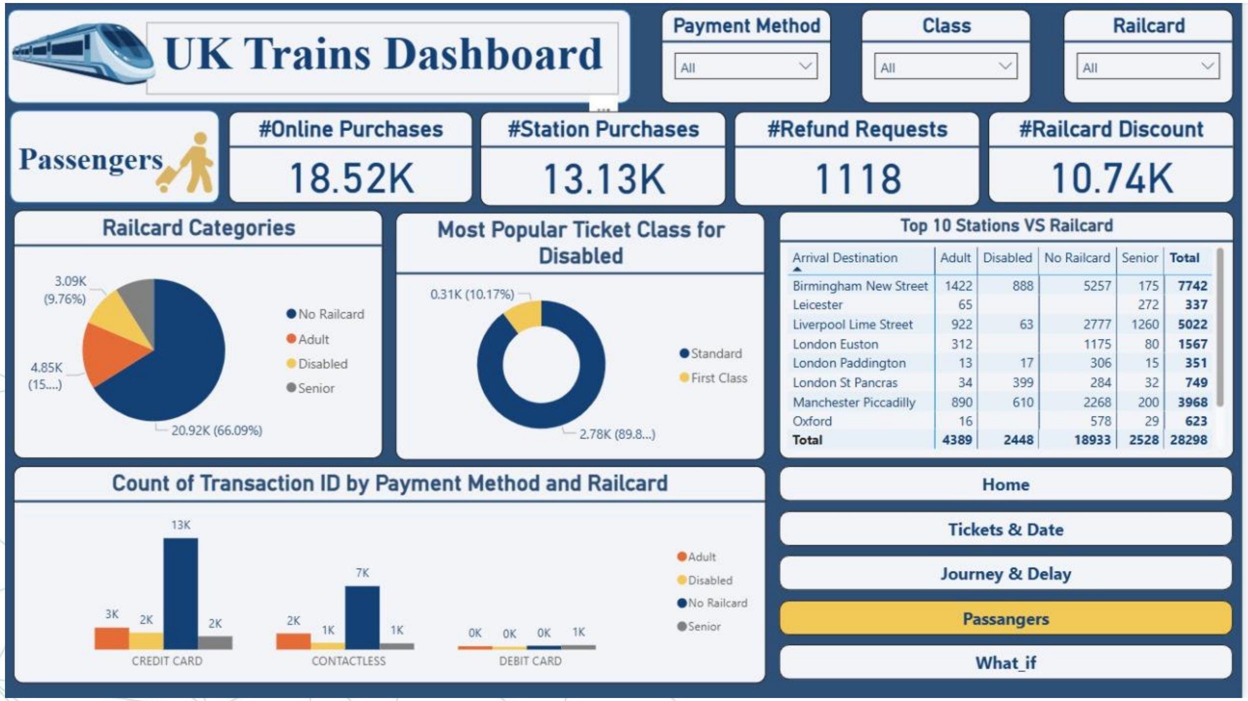
Payment Method	Ticket Class	Price Bins	Months
All	All	All	All


UK Trains Dashboard

Status	Stations
All	All

Journey

#Departure Stations	#Arrival Stations	#On Time Journeys	Avg. Duration/min
12	32	27.48K	30.49



UK Trains Dashboard

Payment Method: All | Class: All | Railcard: All

Passengers

#Online Purchases	#Station Purchases	#Refund Requests	#Railcard Discount
18.52K	13.13K	1118	10.74K

Railcard Categories

Most Popular Ticket Class for Disabled

Top 10 Stations VS Railcard

Arrival Destination	Adult	Disabled	No Railcard	Senior	Total
Birmingham New Street	1422	888	5257	175	**7742**
Leicester	65			272	**337**
Liverpool Lime Street	922	63	2777	1260	**5022**
London Euston	312		1175	80	**1567**
London Paddington	13	17	306	15	**351**
London St Pancras	34	399	284	32	**749**
Manchester Piccadilly	890	610	2268	200	**3968**
Oxford	16		578	29	**623**
Total	**4389**	**2448**	**18933**	**2528**	**28298**

Count of Transaction ID by Payment Method and Railcard

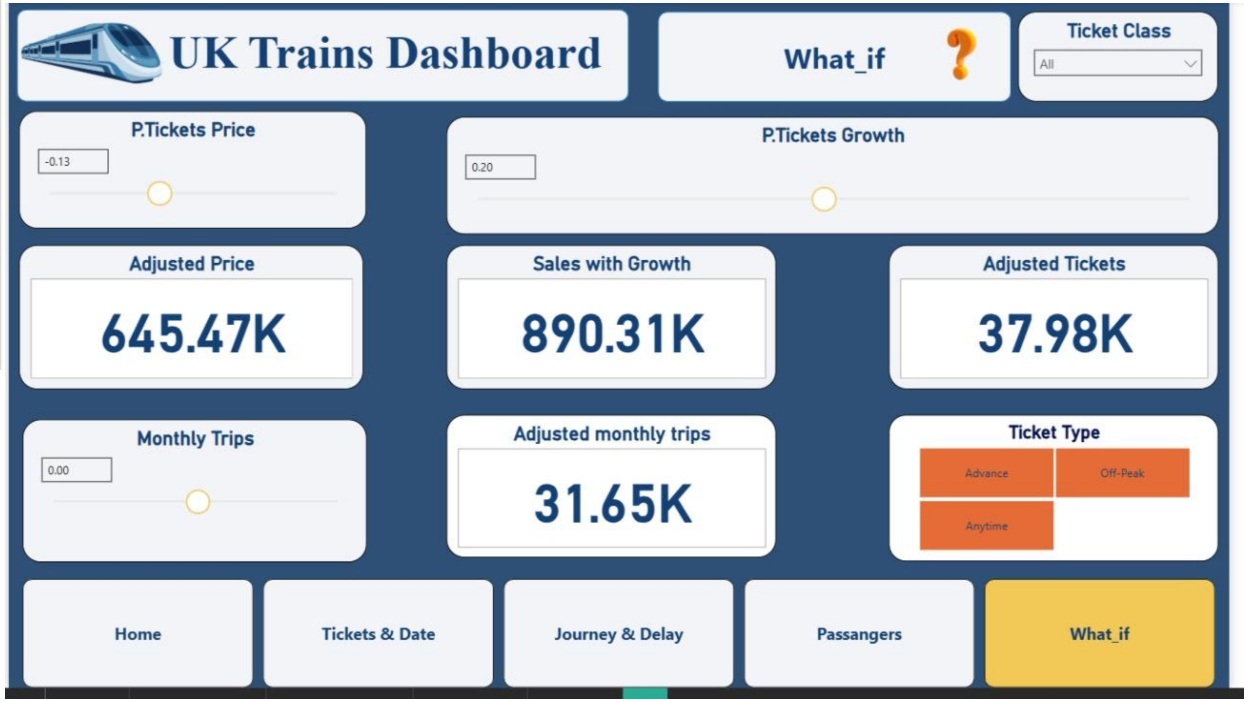
Home

Tickets & Date

Journey & Delay

Passangers

What_if



UK Trains Dashboard

What_if ?

Ticket Class: All

P.Tickets Price: -0.13

P.Tickets Growth: 0.20

Adjusted Price	Sales with Growth	Adjusted Tickets
645.47K	890.31K	37.98K

Monthly Trips: 0.00

Adjusted monthly trips: 31.65K

Ticket Type

Home | Tickets & Date | Journey & Delay | Passangers | What_if

End Users + Features:

- Passengers / Travelers: Need to select the best ticket, avoid delays, and plan trips efficiently.
- Railway Operators / Businesses: Monitor journey performance, passenger volumes, revenues, and optimize operations.
- Decision-makers / Planners: Require accurate forecasts for passenger demand, cancellations, delays, and financial insights to make strategic decisions.

Key Features & Insights:

- Passenger Demand Forecasting → Predict monthly passenger volume
- Delay & Risk Prediction → Estimate journey delays and average delay times
- Cancelled Journey Forecasting → Forecast number of cancelled journeys
- Revenue & Pricing Analysis → Analyze revenue and prices by Ticket Class and Payment Method

How Features Solve User Problems:

User Persona	Problem	Feature/ Insight	Benefit
Passengers	Difficulty selecting the best ticket	Ticket Recommendation, Ticket Analysis	Choose the best ticket, save money and time
Passengers	Uncertainty about delays or cancellations	Delay & Cancel Forecasting	Plan trips better, avoid risky journeys
Operators	Hard to monitor operations and resources	Passenger Demand Forecasting, Revenue Analysis	Optimize staffing, train allocation, and sales strategies
Decision-Makers	Need accurate operational predictions	Predictive Insights (Delay, Cancel, Passenger Volume)	Make strategic, data-driven decisions

Data Structure:

Database Architecture:

- Data stored in CSV files, representing structured tabular format.
- Flat file structure, no relational database used, easy to load into Python/Power BI.

Key Entities & Features:

- Tickets: Transaction ID, Price, Ticket Class/Type, Railcard, Payment Method, Purchase Date
- Passengers: Passenger categories (Adult, Senior, Disabled, No Railcard), Count per ticket type
- Journey: Departure & Arrival Stations, Journey Status, Actual Arrival Time, Delay (min), Reasons for Delay

Data Sape & Quality:

- Total rows: +30k Row
- Columns: 32 features including categorical, numerical, and date/time fields
- Data cleaning applied: removed duplicates, fixed date/time formats, handled missing values, converted times to numeric, encoded categorical values
- Data validation performed to ensure consistency and correctness
- Some imbalance present (e.g., delayed vs. on-time journeys), considered in predictive models

Data Flow:

- Collection: Train booking systems export CSV files
- Storage: Stored locally or on server as CSV
- Processing: Loaded into Python/Power BI for analysis and predictive modeling
- Access: Filtered and visualized in the dashboard; used for forecasting and AI predictions

Programming Languages & Frameworks

Main Languages:

Python (data preprocessing, ML models, forecasting)

Frameworks & Tools:

- Pandas, NumPy (data manipulation and analysis)
- Scikit-learn (machine learning models: Random Forest, Logistic Regression)
- Matplotlib, Seaborn (data visualization)

Supporting Technologies:

- CSV files as primary data source
- Power BI (visualization, dashboard deployment)

Programming Languages & Frameworks

Main Languages:

Python (data processing, ML models, forecasting)

Frameworks & Tools:

- **Pandas, NumPy** (data manipulation and analysis)
- **Sklearn.linear_model, sklearn.preprocessing , numpy, matplotlib, pandas** (machine learning models: Random Forest, Logistic Regression)
- **Matplotlib, Seaborn** (data visualization)

Supporting Technologies:

- CSV files as primary data source
- Power BI (visualization, dashboard deployment)

Final Project Deliverables

- Fully interactive UK Trains Analytics Dashboard (Power BI)
- Cleaned & validated CSV dataset (after removing duplicates, fixing dates, and applying validation rules)
- Machine Learning notebooks for:
  1. Passenger demand forecasting
  2. Cancelled journeys prediction
  3. Average delay time prediction
- Python script for data preprocessing

Documentation Provided

- Data Cleaning & Preparation Document
- Exploratory Data Analysis (EDA) Report
- Forecasting
- Power BI Dashboard Guide
- Testing & Validation Report
- Project Summary + Conclusions

Source Code & Files

- Power BI project file (.pbix)
- GitHub repository including:
    1. Notebooks
    2. Python script
    3. Code of forecasting
    4. Original Data
    5. Cleaned Data
    6. Dashboard file

Timeline & Milestones

- Week 1–2: Data cleaning + validation
- Week 3: Data modeling + dashboard draft
- Week 4: ML modeling & forecasting
- Week 5: Integration + testing
- Week 6: Final dashboard + documentation delivery

Team Members & Their Key Responsibilities:

- **Marina Youssef:** Data Cleaning(Python), Forecasting(Python), Visualization(Power BI), Deployment(Streamlit), Analysis(Python)
- **Helana Hany:** Data Cleaning(Python), Forecasting(Python), Analysis(Python)
- **Salma Hazem:** Data Cleaning(Python), Forecasting(Python), Visualization(Power BI), Deployment(Streamlit), Analysis(Python)
- **Mariam Ahmed:** Data Cleaning(Python), Forecasting(Python), Visualization(Power BI), Analysis(Python)
- **Menna-Allah Mahmoud:** Data Cleaning(Python), Analysis(Python)

Team Members & Roles:

- **Marina Youssef:** Data Analysis, Visualization, Forecasting
- **Helana Hany:** Data Analysis, Forecasting
- **Salma Hazem:** Data Analysis, Visualization, Forecasting
- **Mariam Ahmed:** Data Analysis, Visualization, Forecasting
- **Menna-Allah Mahmoud:** Data Analysis

Collaborations Methods:

- Communication: Zoom / WhatsApp / Google meeting
- Code Sharing: GitHub Repository
- Project Management: Agile approach (sprints, task assignment)
- Version Control: Git + GitHub branches

Thank You!

GitHub Repository

Deployment Link