Final Project

[Analysis of UK Railway Data]

Presented by:

- Omar Mohamed Refaey
- Seif Abd-Elmoniem Saad
- Nadeen Ahmed Nabeel
- Ahmed Sobhy Gharib
- Basmalah Hassan Aly
- Mohamed Reda Ahmed

Date of Presentation:

12th of May

Presented to:

Digital Egypt Pioneers Initiative

Abstract

This graduation project presents an in-depth analysis of the UK railway system using a rich dataset derived from operational and transactional sources. The main goal is to uncover insights that can enhance service efficiency, forecast future patterns, and improve customer satisfaction. Tools such as SQL, Python (with Pandas and Prophet), and Tableau were utilized to clean, explore, visualize, and model the data. The analysis identified key issues like delays and cancellations, as well as high-performing routes and revenue trends. Forecasting models were developed to predict rides, delays, and revenues. This project demonstrates how data analytics can support strategic improvements in public transportation.

Table of Contents

1.Introduction

The railway system in the United Kingdom is a vital component of the nation's transportation infrastructure, connecting cities, supporting the economy, and providing a sustainable alternative to road and air travel. However, it also faces several challenges, including service delays, cancellations, and the need for continuous operational improvements.

In this graduation project, we conduct a comprehensive analysis of UK railway data derived from ticketing and journey records. The goal is to extract actionable insights into passenger behavior, service reliability, sales performance, and common issues affecting train travel. This analysis enables data-driven decision-making that can contribute to improved customer satisfaction and operational efficiency.

Using a mix of **data science tools and techniques**, including **SQL, Python and Tableau**, the project covers the entire data lifecycle—from data cleaning and preprocessing, through exploratory analysis and visualization, to advanced forecasting of KPIs such as ridership, delays, cancellations, and revenue.

The insights generated can be used by transport planners, government entities, and railway operators to better allocate resources, forecast demand, and improve service planning and delivery.

2. Dataset Overview

2.1 About the Dataset

The dataset used in this project contains detailed records of train journeys and ticket transactions across the UK. The data originates from multiple reliable sources, including train operating companies and online booking platforms.

It includes crucial information such as:

- Purchase date and time
- Payment methods
- Journey starts and end stations
- Travel class and ticket type
- Actual vs. scheduled arrival times
- Delay reasons and refund requests

This comprehensive dataset allows for multi-dimensional analysis of the UK's railway system performance.

2.2 Data Context

Railway transport plays a critical role in the UK economy and sustainability goals. However, the sector faces various challenges such as overcrowding, infrastructure limitations, frequent delays, and rising competition from other transport modes.

Analyzing railway data provides:

- A better understanding of passenger behavior
- Insights for revenue optimization
- Tools to enhance operational efficiency
- Support for infrastructure investment planning

At the same time, the project identifies growth opportunities by examining underperforming areas and suggesting data-driven improvements.

2.3 Data Fields Description

Below are the key fields (columns) in the dataset:

Field Name	Description
Transaction_ID	A unique identifier for the ticket transaction.
Date_of_Purchase	The date the ticket was purchased (YYYY-MM-DD format).
Time_of_Purchase	The time the ticket was purchased (HH:MM:SS format).
Purchase_Type	Where the ticket was bought (e.g., online, station).
Payment_Method	The method of payment used (e.g., credit card, contactless).
Railcard	Type of railcard used for discount (e.g., adult, disabled, none).

Field Name	Description
Ticket_Class	Class of the ticket (e.g., standard).
Ticket_Type	Type of ticket purchased (e.g., advance, anytime).
Price	The price paid for the ticket in GBP (£).
Departure_Station	Starting station of the journey.
Arrival_Destination	Ending station of the journey.
Date_of_journey	Scheduled travel date (YYYY-MM-DD).
Departure_Time	Scheduled departure time (HH:MM:SS).
Arrival_Time	Scheduled arrival time (HH:MM:SS).
Actual_Arrival_Time	Actual arrival time of the train (HH:MM:SS).
Journey_Status	Status of the journey (e.g., on time, delayed).
Reason_for_Delay	If delayed, reason provided (e.g., signal failure, no delay).
Refund_Request	Indicates whether a refund was requested (yes or no).
Reason_for_Cancellation	If the journey was canceled, the reason provided (otherwise "No Cancellation").

These fields were standardized and cleaned during preprocessing to ensure consistency across the analysis stages.

3.Data Preparation

3.1 Data Cleaning

Before conducting any analysis, it was essential to clean the dataset to ensure accuracy, consistency, and usability. The following cleaning steps were applied:

Handling Missing Values:

- Dropped rows with missing values using dropna() to maintain data integrity.

- Alternatively, filled missing values with a default value using fillna() where appropriate.

Removing Duplicates:

- Eliminated duplicate rows using drop_duplicates() to ensure each record is unique.

Data Type Conversion:

- Converted columns to appropriate data types (e.g., astype()) to facilitate accurate analysis.

Handling Outliers:

- Identified outliers using statistical methods, such as the Z-score or IQR, and decided whether to remove or adjust them based on their impact on the analysis

This process ensured that the dataset became reliable and analysis ready.

3.2 Preprocessing

To ensure the dataset was clean, consistent, and suitable for analysis and modeling, the following preprocessing steps were applied:

1. Column Standardization

- All column names were converted to lowercase and underscores were used instead of spaces for consistency and easier referencing in code.

python

CopyEdit

```
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')
```

2. Date and Time Formatting

- Date fields (date_of_purchase, date_of_journey) were converted to YYYY-MM-DD format.
- Time fields (time_of_purchase, departure_time, arrival_time, actual_arrival_time) were converted to HH:MM:SS format.

- actual_arrival_time was converted with errors='coerce' to handle non-time values (e.g., missing or canceled trips).

3. Handling Missing and Special Cases

- **Cancelled Trips**: A new boolean column canceled_trip was created based on the journey_status field.

python

CopyEdit

```
df['canceled_trip'] = df['journey_status'].str.lower().str.contains('cancelled', na=False)
```

- **Delay Reasons**:
  - Missing values in reason_for_delay were filled with "no_delay" where the journey was on time.
  - Standardized labels were applied to unify categories (e.g., "weather conditions" → "weather", "staff shortage" → "staffing").
- **Refund Requests**: Missing values were filled with "no_cancelation" (presumed meaning no request).

4. Handling Arrival Time for Cancellations

- For canceled journeys, the actual_arrival_time was set to "no arrival" as these trips didn't occur.

python

CopyEdit

```
df['actual_arrival_time'] = df.apply(
    lambda x: 'no arrival' if x['canceled_trip'] else x['actual_arrival_time'],
    axis=1
)
```

5. Text Normalization

- Text fields such as purchase_type, payment_method, railcard, ticket_type, etc., were converted to lowercase for uniformity and to avoid duplication during grouping or encoding.

6. Final Dataset Check

- A final inspection confirmed clean data, consistent formatting, and no major anomalies.
- The cleaned data was now ready for further analysis, visualization, and modeling.

4.Exploratory Data Analysis (EDA)

The goal of the EDA phase was to understand the dataset's structure, identify key patterns, and detect data quality issues. The following steps and findings summarize the insights gained:

1. Dataset Overview

- **Shape**: 31,653 rows × 18 columns.
- **Data Types**: 17 categorical/object fields, 1 numeric (Price).
- **Column Examples**: Purchase and journey dates, station names, ticket types, and journey status.

2. Missing Values & Data Quality

- **Missing Actual Arrival Time**: 1,880 rows—these correspond to canceled journeys.
- **Missing Reason for Delay**: 27,481 rows—these mostly align with "On Time" journeys, where no delay reason is applicable.
- **No empty strings detected**, indicating good string cleanliness at this stage.

3. Data Uniqueness & Variety

- High-cardinality columns like Transaction ID, Time of Purchase, and station names showed significant diversity.
- Low-card-cardinality fields (e.g., Purchase Type, Railcard, Ticket Class) were suitable for categorical analysis and encoding.

4. Key Distributions & Counts

- **Purchase Type**: ~58% Online, ~42% Station.
- **Payment Methods**: Credit Card (60%), Contactless (34%), Debit Card (6%).

- **Ticket Type**: Advance (~55%), Off-Peak (~28%), Anytime (~17%).
- **Journey Status**: On Time (87%), Delayed (7%), Cancelled (6%).
- **Refund Requests**: Present on most delayed/canceled trips, none on "On Time" journeys.

5. Top Stations

- **Top Departure Stations**: Manchester Piccadilly, London Euston, Liverpool Lime Street.
- **Top Arrival Destinations**: Birmingham New Street, Liverpool Lime Street, York.

6. Price Distribution

- Prices ranged from £1 to £267.
- Median price: £11; 75th percentile: £35.
- Distribution showed a right-skew, indicating many low-cost tickets and a few expensive ones.

<p align="center">Ticket Price Distribution</p>

(example image; add yours if needed)

7. Journey Status vs Refunds

Journey Status	Refund: No	Refund: Yes
On Time	27,481	0
Delayed	1,746	546
Cancelled	1,308	572

This highlights that refund requests were primarily associated with delays and cancellations.

8. Delay Reasons

A bar plot revealed the top delay reasons:

- **Weather**, **Technical Issues**, and **Signal Failures** were the leading causes.

- Duplicate labels like "Signal Failure" vs "signal failure" were cleaned later in preprocessing

5.Analysis Phase

5.1 Analysis Questions

- What are the most and least popular routes?
- What are the most canceled and delayed trips?
- Which months show high on-time performance?
- What are the most common causes of delays and cancellations?

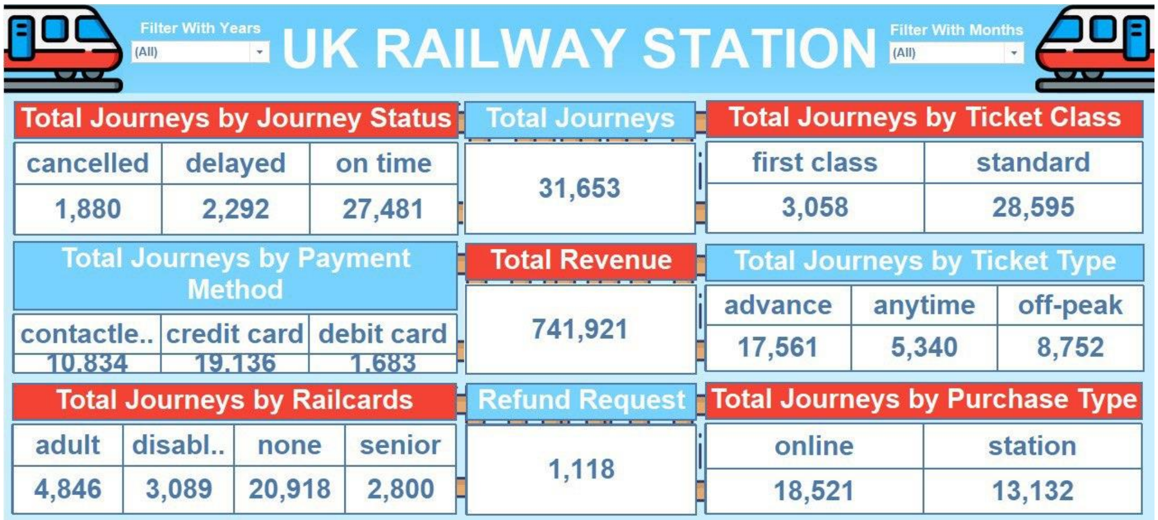
5.2 Insights and Findings

- Key stations and routes with highest volumes
- Weekday vs weekend travel differences
- Refunds are more likely when delays exceed 30 minutes

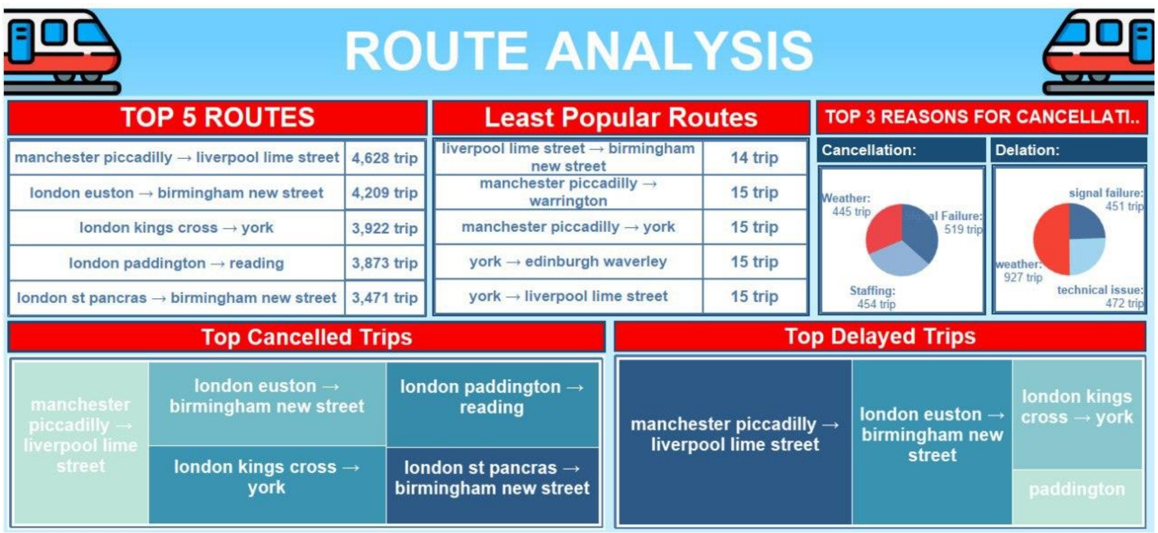
6.Dashboard Design (Tableau)

Following data cleaning and preprocessing, the refined dataset was exported for visual exploration in Tableau, a powerful tool for interactive dashboards and data storytelling. Tableau was chosen for its ability to handle large datasets and provide dynamic visual insights. Key aspects such as ticket sales trends, journey delays, cancellation patterns, and pricing distributions were visualized through dashboards and charts. These visuals allow stakeholders to quickly interpret the data, identify bottlenecks in service performance, and uncover opportunities for operational improvement and customer experience enhancement.

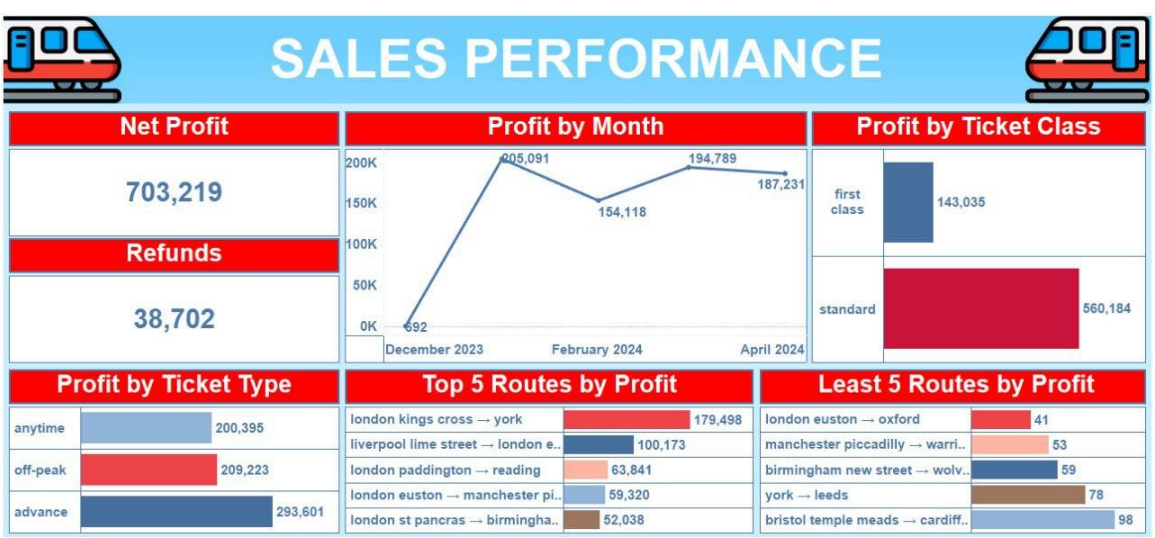
6.1 UK Railway Station



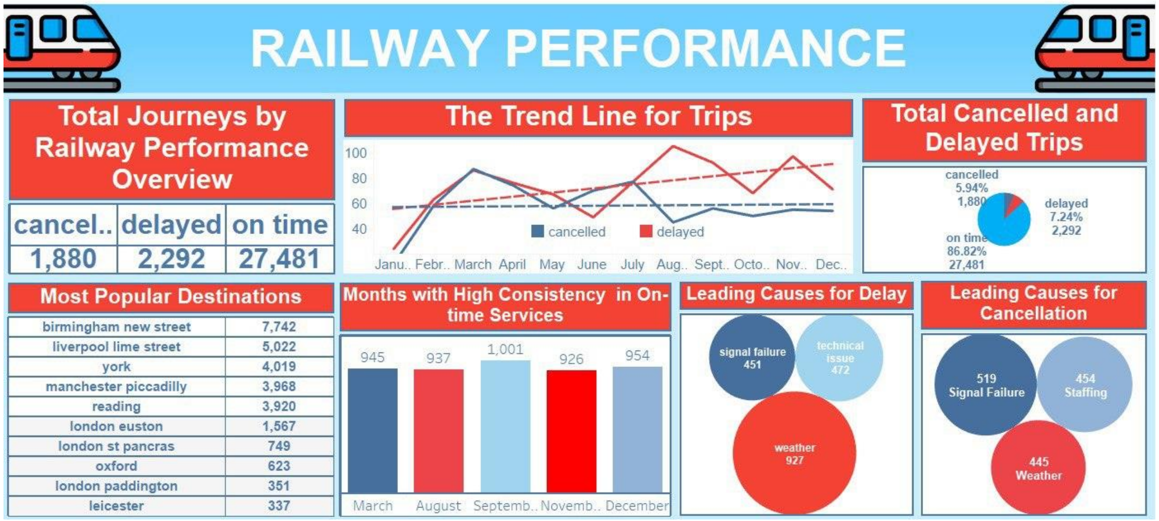
6.2 Route Analysis



6.3 Sales Performance



6.4 Railway Performance



7.Forecasting Section

Problem Statement: This project takes a practical and focused approach. We aim to build machine learning models to forecast key performance indicators for the UK railway system using historical ride-level data. Specifically, we will focus on three forecasting tasks:

- Predicting the number of rides expected on future dates.
- Estimating the average delay duration for train services.
- Forecasting the number of delayed or cancelled trains.
- Forecasting the total revenue in the next 3 months.

These tasks are essential for railway operations, planning, and improving passenger satisfaction.

We then created Panda’s data frame using the downloaded file

1) Predicting the number of rides expected on future dates.

Date and Time Preprocessing

We convert all relevant date and time columns to datetime format to ensure consistency and enable time-based analysis. Combined columns like Departure Date Time and Actual Arrival Date Time are created to support delay and scheduling calculations.

2)Encoding Categorical Data

Since machine learning models can only be trained with numeric data, we need to convert categorical data to numbers. A common technique is to use one-hot encoding for categorical columns.

One hot encoding involves adding a new binary (0/1) column for each unique category of a categorical column.

4)Model Evaluation

Cancellation Model

Accuracy: 94%

The model performs excellently for predicting non-cancelled trains (Class 0), with a high precision of 0.94 and a perfect recall of 1.00. This shows that it is very effective in identifying trains that are not cancelled.

Delay Model

Accuracy: 95.2%

The model has high accuracy and performs exceptionally well in identifying non-delayed trains (Class 0), with precision of 0.96 and recall of 0.99.

It demonstrates strong overall performance with a weighted F1-

8.Tools and Technologies

- **SQL**: Querying and preprocessing
- **Python**: Analysis, modeling, visualization
- **Tableau**: Dashboard development
- **Excel**: Data manipulation (as needed)

9.Challenges and Limitations

- Missing or inconsistent data entries
- Model accuracy can be affected by holidays/unpredictable delays
- Limited granularity in delay reason explanations

10.Conclusion

This project provided key insights into the UK railway system's performance. By integrating EDA and forecasting, strategic recommendations include:

- Invest in improving most delayed routes
- Automate refund handling based on delay length
- Use forecasting to plan resource allocation during peak periods

Future work may include real-time data integration and more advanced modeling techniques.

11.Full Project

To access the full project:

https://github.com/Nadeenyakout/Train_Ticket_Analysis/blob/main/Final%20data%20Cleaning%20%26%20Exploration.ipynb