UK Trains Railway

The main purpose for this project is to assess the performance of the UK Trains Railway focusing on key aspects of ticket purchases, journey performance, and customer experience. The dataset includes detailed information on transactions such as the time and date of purchase, payment methods, ticket types, as well as journey-specific data like departure and arrival stations, scheduled and actual travel times, delays, and reasons for disruptions. The analysis should help the business owners to know how the system operates, how they can make it more efficient.

Participants & Roles

Role	Person
Data Cleaning	mariam.salem.01234@gmail.com
	Joumana Hossam
Data Modelling	Shaza Allam
Data Visualization	oosha2000@gmail.com
	meramgeorge@gmail.com

Data Structure

The data contains 1 file: ***railway.csv***

Data Description

The columns of the data consist of:

1. **Transaction ID,** Unique identifier for an individual train ticket purchase
2. **Date of Purchase,** Date the ticket was purchased
3. **Time of Purchase,**Time the ticket was purchased
4. **Purchase Type,** Whether the ticket was purchased online or directly at a train station
5. **Payment Method,** "Payment method used to purchase the ticket (Contactless, Credit Card, or Debit Card)"
6. **Railcard,** "Whether the passenger is a National Railcard holder (Adult, Senior, or Disabled) or not (None). Railcard holders get 1/3 off their ticket purchases."
7. **Ticket Class,** Seat class for the ticket (Standard or First)
8. **Ticket Type,** When you bought or can use the ticket. Advance tickets are 1/2 off and must be purchased at least a day prior to departure. Off-Peak tickets are 1/4 off and must

be used outside of peak hours (weekdays between 6-8am and 4-6pm). Anytime tickets are full price and can be bought and used at any time during the day.
9. **Price,** Final cost of the ticket
10. **Departure Station**, Station to board the train
11. **Arrival Destination,** Station to exit the train
12. **Date of Journey,** Date the train departed
13. **Departure Time,** Time the train departed
14. **Arrival Time,** Time the train was scheduled to arrive at its destination (can be on the day after departure)
15. **Actual Arrival Time,** Time the train arrived at its destination (can be on the day after departure)
16. **Journey Status,** "Whether the train was on time, delayed, or cancelled"
17. **Reason for Delay,** Reason for the delay or cancellation
18. **Refund Request,** Whether the passenger requested a refund after a delay or cancellation

Problems of Data

- Multiple cells were empty. (Reason for Delay)
- Duplicates in values. (Same journey exist in different transactions which needed normalization)
- Delay column contained the same name one time with a capital letter and the other time with a small letter. (Reason for Delay)
- Some journeys are marked as Delayed despite the fact that Actual arrival Time equals Arrival Time. (Journey Status)

Data Modelling

Tables in the model are:

- **Fact Table** (for all purchases) consists of:
    - Transaction ID
    - Journey ID (Foreign key for journey)
    - Ticket ID (Foreign key for ticket)
    - Date of Purchase (Foreign key for date)
    - Time of Purchase
    - Hour of Purchase (Calculated Column)
    - Purchase Type
    - Payment Method
    - Ticket Class
    - Ticket Type
    - Price
    - Discount Percentage (added using M Language)
    - Railcard
    - Refund Request

- **Journeys Dimension** consists of:
    - Journey ID (Primary key for journey)
    - Route ID (Foreign key for route)
    - Date of Journey (Foreign key for date)
    - Departure Time
    - Arrival Time
    - Actual Arrival Time
    - Delay Time (Calculated Column)
    - Journey Duration (Calculated Column)
    - Journey Status
    - Reason for Delay

- **Routes Dimension** consists of:
    - Route ID (Primary key for route)
    - Departure Destination (Foreign key for station)
    - Arrival Destination (Foreign key for station)

- **Stations Dimension** consists of:
    - Station Name (Primary key for station)

- **Date Dimension** consists of:
    - Date (Primary key for date)

Key Performance Indicators (KPIs)

KPIs will be categorized into 2 main categories:

- **Sales & Revenue**

    - **Total Revenue**: Sum of all transactions (Not Refunded).
        - **Question:** How much revenue earned during the two years (2023/2024) included in the data without refund?

    - **Total Refunded Amount**: Sum of the refunded amount
        - **Question:** How much refund got on the revenue?

    - **Number of Tickets Sold**: Total count of transactions.
        - **Question:** How many tickets were sold during the two years (2023/2024) included in the data?

    - **Refund Request Rate**: (Number of Refund Requests / Total Number of Transactions) * 100.
        - **Question:** How many of the sold tickets got refund requests?

- **Operational Performance**

    - **Number of Journeys**: Count of all journeys.

        - **Question:** How many journeys were made?
    - **Number of Routes**: Count of routes on the UK railway system.
        - **Question:** How many routes are on the UK railway system?
    - **Number of Stations**: Count of stations on the UK railway system.
        - **Question:** How many stations are on the UK railway system?
    - **Cancellation Rate**: (Number of cancelled journeys / Total number of scheduled journeys) * 100
        - **Question:** What percentage of scheduled journeys are cancelled?
    - **Delay Rate**: (Number of delayed journeys / Total number of scheduled journeys) * 100
        - **Question:** What percentage of scheduled journeys are delayed?
    - **Average Delay Time**: Average of "Actual Arrival Time" - "Arrival Time" for delayed journeys / Number of delayed journeys.
        - **Question:** How many seconds does the journey delay on average?
    - **Average Journey Duration**: Average of "Actual Arrival Time" - "Departure Time" for journeys / Number of journeys.
        - **Question:** How many minutes does the journey take on average?

Charts

charts will be categorized into 3 main categories:

- **Sales & Revenue**
    - **Sales by Payment Method**: No. of Tickets for each "Payment Method" (Contactless, Credit Card).
        - **Question:** How many tickets were purchased in each payment method?
    - **Sales by Ticket Type (with tooltip)**: No. of Tickets for each "Ticket Type" (Advance, etc.).
        - **Question:** How many tickets were purchased in each ticket type?
    - **Sales by Railcard Type (with tooltip)**: No. of Tickets for each "Railcard" type (Adult, Disabled, etc.).
        - **Question:** How many tickets were purchased in each railcard type?
    - **Sales by Purchase Type**: No. of Tickets for each "Purchase Type" (Online, Station).
        - **Question:** How many tickets were purchased in each purchase type?
    - **Sales by Ticket Class**: No. of Tickets for each "Ticket Class" (Standard, etc.).
        - **Question:** How many tickets were purchased in each ticket class?

- **Sales by Purchase Date**: Number of tickets sold on different purchase dates with forecast (1 month).
  - **Question:** How many tickets were purchased on different purchase dates (with prediction for 1 month)?

- **Sales by Purchase Hour**: No. of Tickets Sold by Purchase Hour
  - **Question:** How many tickets were purchased in each purchase hour ?

- **Sales by Journey Date**: Number of tickets sold on different journey dates with forecast (1 month).
  - **Question:** How many tickets were purchased on different journey dates (with prediction for 1 month)?

- **Total Revenue by Journey Date**: Revenue generated on different journey dates with forecast (30 days).
  - **Question:** How much revenue got on different journey dates (with prediction for 30 days)?

- **Demand on each Journey Route (Drill Through)**: Count of purchases on each journey route.
  - **Question:** How many routes are purchased on each route?

Operational Performance

- **Number of Journeys by Journey Date**: Number of journeys done by journey date with forecast (1 month).
  - **Question:** How many journeys are done on different journey dates (with prediction for 1 month)?

- **Peak Time through Day for Departure**: Show time where it has the most count of journeys.
  - **Question:** Which times are the peak for journeys to departure?

- **Number of Delayed Journeys by Departure Time**: Number of delayed journeys on each departure time.
  - **Question:** What is the average delay time on each journey route on each departure time? Which times have the most delay?

- **Delay Reason Analysis**: Frequency of each delay reason (e.g., Signal Failure).
  - **Question:** What is the average delay time in minutes for each delay reason? Which station has the said delay reason ?

- **Journey Status Distribution**: Analyze the distribution of "Journey Status" (On Time, Delayed, Cancelled).
  - **Question:** How many journeys are done on each journey status?

- **Number of Journeys by Station**: Number of journeys done on each station.
  - **Question:** How many journeys are done per station?

- **Delay Reason on each Route (Drill Through)**: Analyze delay reason on each journey route.

        - **Question:** What is the most frequent reason for delay?

    - **Delay Rate on each Route (Drill Through)**: Delay rate on each journey route.
        - **Question:** What is the delay rate on each journey route? Which routes have the most delay rate?

    - **Cancel Rate on each Route (Drill Through)**: Cancel rate on each journey route.
        - **Question:** What is the cancellation rate on each journey route? Which routes have the most cancellation rate?

    - **Average Delay Time on each Route (Drill Through)**: Average delay time "Actual Arrival Time - arrival Time" on each journey route.
        - **Question:** What is the average delay time on each journey route? Which routes have the most delay time?

    - **Average Journey Duration on each Route (Drill Through)**: Average journey duration time "Actual Arrival Time" - "Departure Time" on each journey route.
        - **Question:** What is the average journey duration on each journey route? Which routes have the most average journey duration?

Data Insights

- **Refund Rate on each Ticket Type**
  In spite of the low refund rate **3.53%**, It leads to a great loss on revenue which is almost **38.70K**.
    - **Used KPIs and Charts:**
        - Total Revenue
        - Total Refund
        - Refund Request Rate
        - Sales by Ticket Type (with tooltip)
        - No. of Journeys by Journey Status (with tooltip)

    - **Exploration:**
        - According to the percentage of refund on each Ticket Type in order are:
            - **Off-Peak:** 4.22% most of them are online and using Credit card
            - **Advance:** 3.48% most of them are online and using Credit card
            - **AnyTime:** 2.58% most of them are online and using Credit card
        - The percentage of refund on each Journey Status in order are:
            - **Cancelled**: 30.43% with total refund of 12.54K.
            - **Delayed**: 23.66% with total refund of 25.49K.
            - **On Time**: 0.03% with total refund of 678.

    - **Suggested Explanation:**
        - Off-peak tickets can be used only on off-peak hours (weekdays between 6-8am and 4-6pm), so may be after purchasing them online they can not use it on peak hours.
        - Delay and cancellation of journeys leads to customer dissatisfaction which leads to refund requests.

- **Suggested Strategy:**
  - Discounts can be applied on Anytime tickets that can be used any time even in peak hours (it has the least refund rate).
  - Make a disclaimer about off-peak tickets on the website when purchasing.
  - Resolve the delay and cancellation reasons for journeys.

- **Delay & Cancellation Rate of each Route**
  - **Used KPIs and Charts:**
    - No. of Delayed Journeys by Departure Time
    - No. of Tickets on each Route (Drilled by station)
    - Delay Rate on each Route (Drilled by station)
    - Cancel Rate on each Route (Drilled by station)
    - Avg. Delay Time on each Route (Drilled by station)
    - Delay Reason Analysis (Drilled by station)
    - Avg. Journey Duration on each Route (Drilled by station)
  - **Exploration:**
    - By analysing the **top 10** used stations for journeys:
      - **Manchester Piccadilly**
        - Technical Issue is the most impactful reason for delay.
        - The route with highest demand is to Liverpool Lime Street.
        - The biggest delay is on the route between it and Birmingham New Street.
      - **Birmingham New Street**
        - Weather is the most impactful reason for delay.
        - The route with highest demand is to London Euston.
        - The biggest delay is on the route between it and London Euston.
      - **Liverpool Lime Street**
        - Staff Shortage is the most impactful reason for delay.
        - The route with highest demand is to Manchester Piccadilly.
        - The biggest delay is on the route between it and Manchester Piccadilly.
      - **London Euston**
        - Weather is the most impactful reason for delay.
        - The route with highest demand is to Liverpool Lime Street.
        - The biggest delay is on the route between it and Manchester Piccadilly.
      - **York**
        - Technical Issue is the most impactful reason for delay.
        - The route with highest demand is to London Kings Cross.
        - The biggest delay is on the route between it and London Euston.
      - **London Paddington**
        - Staff Shortage is the most impactful reason for delay.
        - The route with highest demand is to Reading.
        - The biggest delay is on the route between it and Liverpool Lime Street.

- **Reading**
  - Technical Issue is the most impactful reason for delay.
  - The route with highest demand is to London Paddington.
  - The biggest delay is on the route between it and London Paddington.
- **London st. Pancars**
  - No delay.
  - The route with highest demand is to Birmingham New Street.
- **London Kings Cross**
  - Staff Shortage is the most impactful reason for delay.
  - The route with highest demand is to Edinburgh Waverley.
  - The biggest delay is on the route between it and Edinburgh Waverley.
- **Oxford**
  - No delay.
  - The route with highest demand is to London Paddington.

■ The highest number of delayed journeys happens at Departure Time 08:00 AM and 11:00 AM.

- **Suggested Explanation:**
  - Some stations suffer from staff shortage, others from weather conditions and others from technical issues. This leads to delay on those stations which most likely affects the highest demand routes.

- **Suggested Strategy:**
  - Hire more staff on **Liverpool Lime Street**, **London Paddington**, and **London Kings Cross**.
  - Make better schedules to cover all hours along the day (The most delay is at 08:00 AM and 11:00 AM).
  - Track weather forecasting on **Birmingham New Street**, and **London Euston**, to know if there is bad weather to take into account.
  - Solve Technical issues on **Manchester Piccadilly**, **York**, and **Reading**.

- **Purchase Type on each Railcard**
  - **Used KPIs and Charts:**
    - Sales by Railcard (with tooltip)

  - **Exploration:**
    - According to the Purchase Type percentage on each Railcard are:
      - **Adult:** most likely to purchase from station
      - **None, Disabled, Senior:** most likely to purchase online

  - **Suggested Explanation:** Disabled and seniors can not wait in long lines to purchase tickets.

  - **Suggested Strategy:**
    - Encourage all customers to purchase online (which will make lines shorter) by making discounts for those who purchase online.

        - Encourage the disabled to purchase online by making an application with accessibility to people with special needs (like deaf, and blind).

- **Peak Time for Journeys**
    - **Used KPIs and Charts:**
        - Peak Time through Day for Departure
    - **Exploration:**
        - It is noticed that the peak hours for departure are at 06:30 AM, 07:30 AM and 06:45 PM.
    - **Suggested Explanation:** The number of journeys at these hours are big, because of the students and employees commuting.
    - **Suggested Strategy:** As staff shortage and signal failure are the most common reasons for delay, It is needed to have more staff in rush hours and have regular maintenance on the vehicles.

Unavailable Key Performance Indicators (KPIs)

- Financial & Efficiency Metrics
    - **Revenue per Passenger-Kilometer** – Total revenue divided by passenger distance traveled.
    - **Operating Cost per Kilometer** – Total cost per km/mile of train operation.
    - **Energy Efficiency** – kWh consumed per passenger-km.

- Infrastructure & Utilization
    - **Peak vs. Off-Peak Utilization** – Comparison of passenger load during different times.
    - **Seat Occupancy Rate** – % of occupied seats per trip.
    - **Turnaround Time at Stations** – Average time taken for a train to prepare for the next trip.
    - **Platform Utilization** – How busy the platforms are at different times.
    - **Cancellation Reasons** – The reasons for journeys to be cancelled e.g., train maintenance or upgrades, weather issues).
    - **Ticket Refund Processing Time** – Measures the efficiency of refund processing

- Customer Satisfaction & Service Quality
    - **Customer Satisfaction (CSAT)** – Customer satisfaction with various aspects of the journey (booking, journey experience, customer service).
    - **Customer Retention** – How many times a certain customer purchases a ticket? (Indicates how we will retain the customer).