UK Train Rides-Railway

# Overview

The dataset provides comprehensive details on simulated train ticket sales for National Rail in the United Kingdom, covering the period from January to April 2024. It includes various essential attributes such as the type of ticket purchased, the exact date and time of each journey, the departure and arrival stations, and the ticket price. Additionally, the dataset may encompass other relevant aspects for analysis, including different fare categories, applied discounts, service classes available, and price variations based on demand, travel periods, and specific routes. This information offers valuable insights into ticketing trends, passenger preferences, and potential fluctuations in rail travel patterns during the given timeframe.

# Objectives

The dataset aims to:

* Provide delay time for the journey.
* Explore data and time for purchase ticket and his price.
* The reason for cancelled trips and minimize cancelled trips.
* Understand reason for delayed trips and provide refund request.
* improve stations that departure station and arrival station.
* The peak of purchase ticket per hour and optimize purchasing.

# Data Source and Description

The dataset originates from [UK Train Rides](https://www.kaggle.com/datasets/willianoliveiragibin/uk-train-rides/data) a railway in UK that the passenger in UK use it to travel between stations. It encompasses information on approximately 31653 ticket purchase from January to April 2024. The data include various aspects of the railway and trips, including:

|  |
| --- |
| * **Railway table**: Information on each transaction ID, purchase details, ticket details, railcard and if have railcard or not and payment method. |
| * **Trips table**: trips ID, departure and arrival station, details for delay, details of journey and departure and Arrival Coordinates |
| * **Dim Journey Date table**: date, time and week day typer |
| * **Station:** the station name and coordinates |

* **Ticket details:** ticket price, type and class.
* **Purchase information:** date and time, type, payment method, Days Between Purchase a journey and Purchase Hour.
* **Journey details:** trip ID, date, day of journey, departure and arrival time, actual arrival time, status, delay time, reason for delay.
* **Station information:** coordinates, departure and arrival.

# Key Columns

|  |
| --- |
| **Transaction ID:** Unique identifier for each ticket purchase transaction**.** |
| **Date of Purchase:** The date on which the ticket was purchased**.** |
| **Time of Purchase:** The time on which the ticket was purchased**.** |
| **Purchase Type:** Where the ticket was purchased: Online or at the Station**.** |
| **Payment Method:** Payment method used to purchase the ticket (Contactless, Credit Card, or Debit Card). |
| **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. |
| **Ticket Class:** Seat class for the ticket (Standard or First). |
| **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. |
| **Ticket Price:** Price of the individual ticket. |
| **Departure Station:** The station from which the train journey begins. |
| **Arrival Destination:** The station where the train journey ends. |
| **Journey Date:** The date of the actual train journey. |
| **Departure Time:** Time the train departed. |
| **Arrival Time:** Time the train was scheduled to arrive at its destination (can be on the day after departure). |
| **Actual Arrival Time:** Time the train arrived at its destination (can be on the day after departure). |
| **Journey Status:** Whether the train was on time, delayed, or cancelled. |
| **Reason for Delay:** Reason for the delay or cancellation. |
| **Refund Request:** Whether the passenger requested a refund after a delay or cancellation. |

**Additional Custom Columns Created:**

* **Origin-destination:** merged to column departure station and arrival destination**.**
* **Delay time(minutes):** the difference between arrival time and actual arrival time.
* **Merged column:** merge origin-destination, date of journey and departure time.
* **Purchase Hour:** extracted hour from Purchase Time column.

# Data Analysis questions:

* Performance Analysis

1. What percentage of total trips were cancelled or delayed?
2. Which routes experience the highest number of delays or cancellations?
3. Is there a relationship between the number of trips on a route and the delay/cancellation rate?

* Revenue Analysis

1. Which routes generate the most revenue, and how does that correlate with trip volume or punctuality?
2. What is the revenue loss due to cancelled or delayed trips?
3. Is there a seasonal or monthly trend in revenue, delays, or cancellations?

* Route Optimization

1. Which routes have the lowest performance (e.g., highest cancellations) and lowest revenue — candidates for reevaluation or discontinuation?
2. Which underperforming routes still bring in high revenue, possibly indicating essential services despite inefficiency?

* Operational Efficiency

1. Are certain time periods (days of the week, months) more prone to delays or cancellations?
2. What are the average delay durations per route or trip type?

* Comparative Questions

1. How do high-performing routes differ from low-performing ones in terms of trips, revenue, and delays?
2. what is the trend of trip volume and performance over time — is it improving or declining?

# Data Cleaning Process

Data cleaning was conducted using **Power Query** to ensure consistency and remove any irrelevant information. Below are the detailed steps performed:

# Merging

1. **Merge Departure Station with Arrival Destination**: merge this column in on column called Origin – Destination to know what arrival and departure station.
2. **Merge Departure Station and Arrival Destination Date of Journey and Departure Time**.

# Adding Columns

1. **Add Conditional Column (Have Railcard)**: if passenger have railcard or not.
2. **Add Conditional Column (Journey Duration)**: the difference between arrival time and actual arrival time if actual arrival time greater than arrival time.
3. **Add Custom Column (Trip ID)**: A new column was created to be primary key in trips table and foreign railway table.
4. **Add Custom Column (Duration):** new column to know difference between Arrival Date and Time with Arrival Date and Time

# Removing Irrelevant Data

1. **Trim & Edit** spelling errors in Reason of Delay Column.
2. -**Replace blank Values** with No Delay in Reason of Delay Column.

# Data Modeling:

In this project, a star schema was implemented to efficiently organize the data for reporting and analysis purposes. The star schema consists of a central fact table that captures key transactional data, surrounded by dimension tables that store descriptive information about the entities involved in the transactions.

# Fact Table: Trips

At the core of the star schema is the Trips Table, which serves as the fact table. This table contains transactional data related to trips, including fields like:

**Trips ID**: Links to the trips and railway.

**Departure Time**: Scheduled time of departure.

**Arrival Time**: scheduled time of arrival.

**Actual Arrival Time:** actual arrival time for the journey.

**Journey Status:** if journey already arrived on time or cancelled or delayed for time.

**Reason For Delay:** if journey delayed what is the reason for it.

# Dimension Tables

Each dimension table provides additional context and descriptive information about the data in the fact table:

# Railway:

Contains ticket-specific information such as the ticket type, ticket class, railcard purchase details (Date – Time – Hours - Type) and refund request.

# Journey Status:

Stores status of journey that journey (on time – cancelled – delayed).

# Dim Journey Date:

Provides information dates for all trips (day – month – year – season).

# Station:

All information for the station dimensions on map (station – Latitude - Longitude).

# Relationships

In the star schema, the fact table is connected to each dimension table via a one-to-many relationship. The foreign keys in the fact table (such as trip\_id, station and date) link to the primary keys in the dimension tables, forming the "star" structure.

**The data model**

**A screenshot of a computer

AI-generated content may be incorrect.**

# Analysis Methodology

# DAX

|  |  |
| --- | --- |
| **Total Revenue:** | This measure computes the total revenue. |
| **Average Ticket Price:** | This measure computes the average of total price for tickets. |
| **Cancelled Trips:** | This measure, calculate total cancelled trips. |
| **Delayed Trips:** | This measure, calculate total delayed trips. |
| **On-Time trips:** | This measure, calculate total trips that arrived on-time. |
| **Most Reason for Delay and Cancel** | In this measure, filter reason for delay that the most reason for delay. |
| **Refund Request:** | This measure calculates total refund request. |
| **Tickets** | This measure computes the total ticket of trips. |
| **Total Delay Time:** | This measure calculates total delay time. |
| **Total Trips:** | This measure counts and calculate total trips. |
| **Weekday Trips:** | This is calculated the weekday trips. |
| **Weekend Trips:** | This is calculated the Weekendtrips. |

# Dashboard Creation:

The data was impressively displayed with the five dashboards that were made up:

# Overview Dashboard

Cards: Total Trips, On Time, Cancelled, Delayed, Average Daily Trips, Average Daily Passengers.

# Revenue Dashboard

Cards: Total revenue, Refund, Ticket, Refund Ticket percentage.

# Refund Dashboard

Cards: Total Revenue, Refunded, Number of Ticket, Refunded Tickets.

# Delays & Punctuality Dashboard

Cards: Delayed Trips, Total Delay Time, Delay Percentage, Average Delay Time, Most Delay Reason.

# Cancelled Trips Dashboard

Cards: Cancelled Trips, Cancelled percentage, Week Days Trips, Weekend Trips, Most Cancelled Reason.

# Visualizations:

**Visuals of Overview dashboard:**

* + Pie chart: Total Passengers by Weekday Type.
  + Bar chart: Number of Passengers by Day Name.
  + Column Chart: Trips Peak Hours Per Day.

**Visuals of Revenue dashboard:**

* + Clustered bar chart: Refunded ticket percentage by journey status.
  + Donut chart: Ticket by purchase type, ticket by payment method.
  + Bar plot – Average number of orders per customer.
  + Clustered column chart: Average Ticket Price and Tickets by Ticket Type.
  + table: Revenue by station.

**Visuals for Refund dashboard:**

* + Donat chart: Refunded by Payment Method and Refunded Tickets by Purchase Type.
  + Column chart: Refunded Tickets% by Journey Status, Refunded Tickets by Ticket Type, Refunded Tickets by Reason for Delay / Cancelled and Refunded Tickets% by Departure Station.
  + Bar chart: Refunded Tickets by Ticket Class.

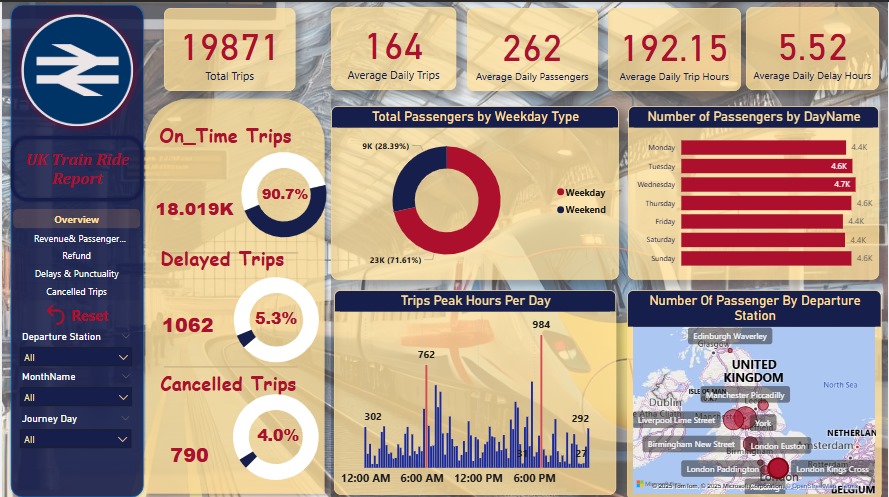
**Visuals for Delays & Punctuality dashboard**:

* + Bar chart: Total Delay Time (h) by Origin - Destination.
  + Line chart: Total Delay Time (h) by Date.
  + Map: Delayed Trips and Average Delay Time(m) by Departure Station.
  + Line and column chart: Average Delay Time(m), Total Trips by Day Name.

**Visuals of Cancelled Trips dashboard:**

* + Column chart: Cancelled trips by Day Name.
  + Line chart: Cancelled trips by Month Name
  + Donut chart: Reasons for Cancelled Trips
  + Bar chart: Cancelled trips by Origin – Destination and Cancelled trips by Departure Station.

**Slicers**: Departure Station - Arrival Destination - Reason for Delay – Month Name.



# Key findings:

* + **Top Day with Most Passengers:**
* **Wednesday** (4.7K passenger).
* **Tuesday, Thursday and Sunday** also show high passenger.
  + **Number of total Trips is 19871 and divided into:**
* **One-Tim** is .
* **Delayed** is 1062.
* **Cancelled** is .

# Analysis Summary:

The journey dashboard is represented total trips and journey status, is it on-time or delayed or cancelled, found have small percentage with delayed and cancelled trips but we need to improve it.



# Key Observations:

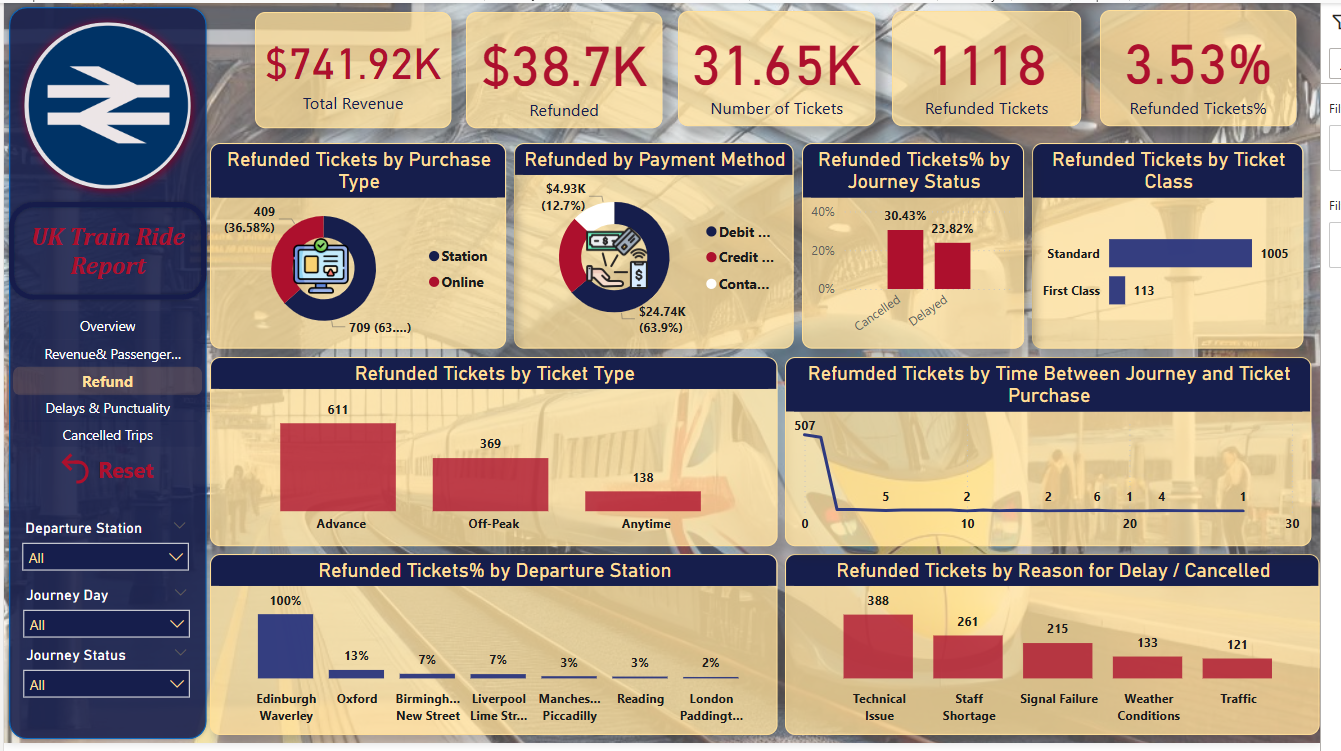
* **Ticket Purchase Method:**
  + Online: 58.51% (18.52K)
  + Station: 41.49% (13.13K)
* **Refund Rate by Journey Status:**
  + Cancelled: 30.43%
  + Delayed: 23.82%
* **Edinburgh Waverley: 100% of revenue was refunded — a significant outlier and may indicate full trip cancellations or major issues.**
* **Top Revenue Stations:**
  + London Kings Cross: $199.65K
  + Liverpool Lime Street: $135.27K
  + London Euston: $112.04K
* **High Refund Rates:**
  + Edinburgh Waverley: 100%
  + Manchester Piccadilly: 12.6%
  + Oxford: 11.6%

**Ticket Type Insights:**

* Advance Tickets are the most sold (17.56K) but have the lowest average price: $17.6
* Anytime Tickets have the highest average price: $39.2 but fewer sales (5.34K)

**Peak Hours Analysis:**

* Peak Online Sales Hour: 15:00 (2,740 tickets)
* Peak Station Sales Hour: 20:00 (2,239 tickets)
* Sales are more active online during midday to afternoon, while station sales peak in the evening.

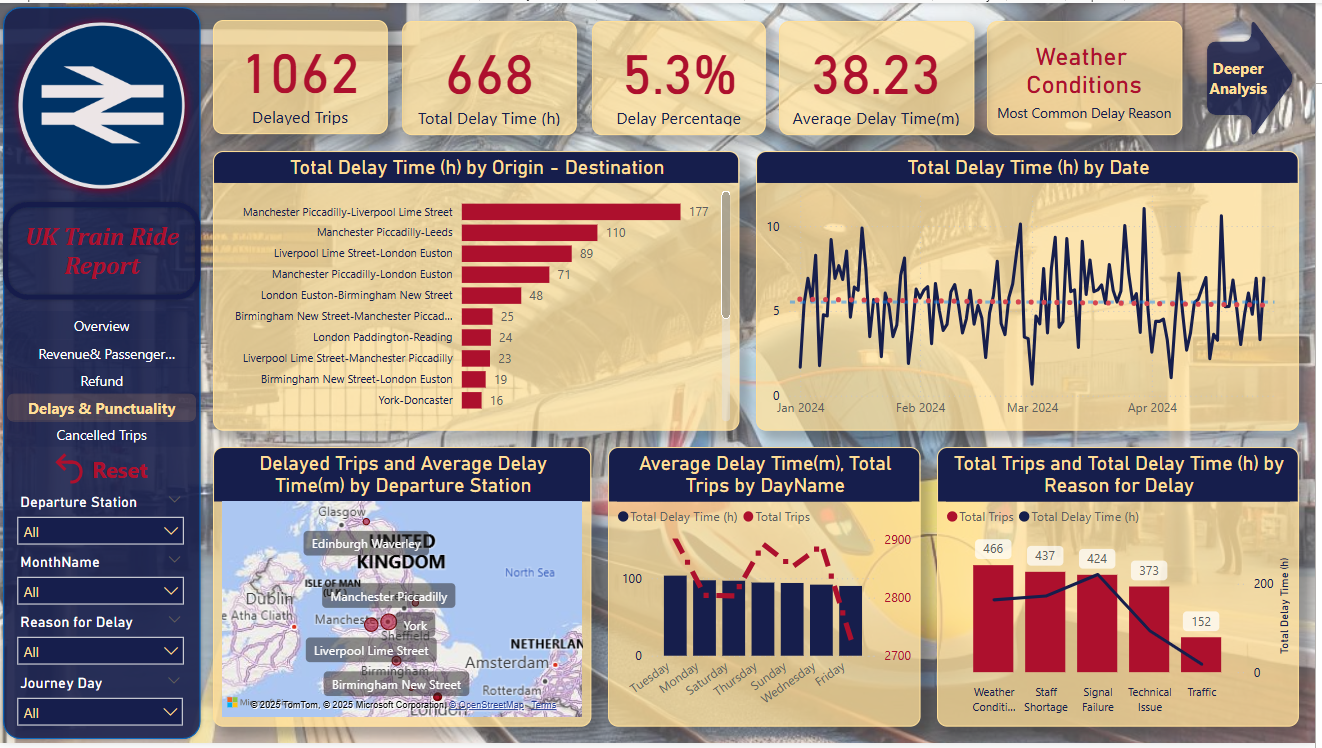


# Key Observations:

* **Edinburgh Waverley: 100% refund rate – major red flag, possibly full cancellations**
* **Most refunds occurred very close to the journey date, particularly within 1 day of purchase — strong indication of last-minute cancellations.**
* **Technical Issues: 388 refunds**
* **Staff Shortage: 261 refunds**
* **Signal Failure: 215 refunds**
* **Weather Conditions: 133 refunds**
* **Traffic: 121 refunds**

**Insights & Recommendations:**

* Advance Tickets may be priced lower but carry higher refund risk.
* Technical issues and staffing problems are key operational challenges — addressable with maintenance and workforce planning.
* Edinburgh Waverley requires urgent investigation — 100% refunds suggest systemic issues.
* Focus on improving service reliability to reduce refund claims, especially for online and credit card customers**.**



# Key Observations:

**Top Delayed Routes (Total Delay Time):**

1. **Manchester Piccadilly – Liverpool Lime Street:** 177 hours
2. **Manchester Piccadilly – Leeds:** 110 hours
3. **Liverpool Lime Street – London Euston:** 89 hours

* **Manchester Piccadilly appears frequently – indicating a key hub with significant punctuality issues.**

**Delay Time Trends Over Time:**

* Delay patterns appear **erratic with frequent spikes**, especially between **February and March 2024**.
* Despite high peaks, the **overall trend line remains steady**, indicating consistent performance issues rather than seasonal ones.

**Geographic Insights (Map):**

* Stations like **Manchester Piccadilly**, **Liverpool Lime Street**, and **Birmingham New Street** experience **both high delay counts and long delay durations.**

**Delays by Day of the Week:**

* **Average Delay Time is highest on Mondays.**
* **Trip volumes peak mid-week (Tuesday to Thursday)**, but **delays remain relatively high throughout.**

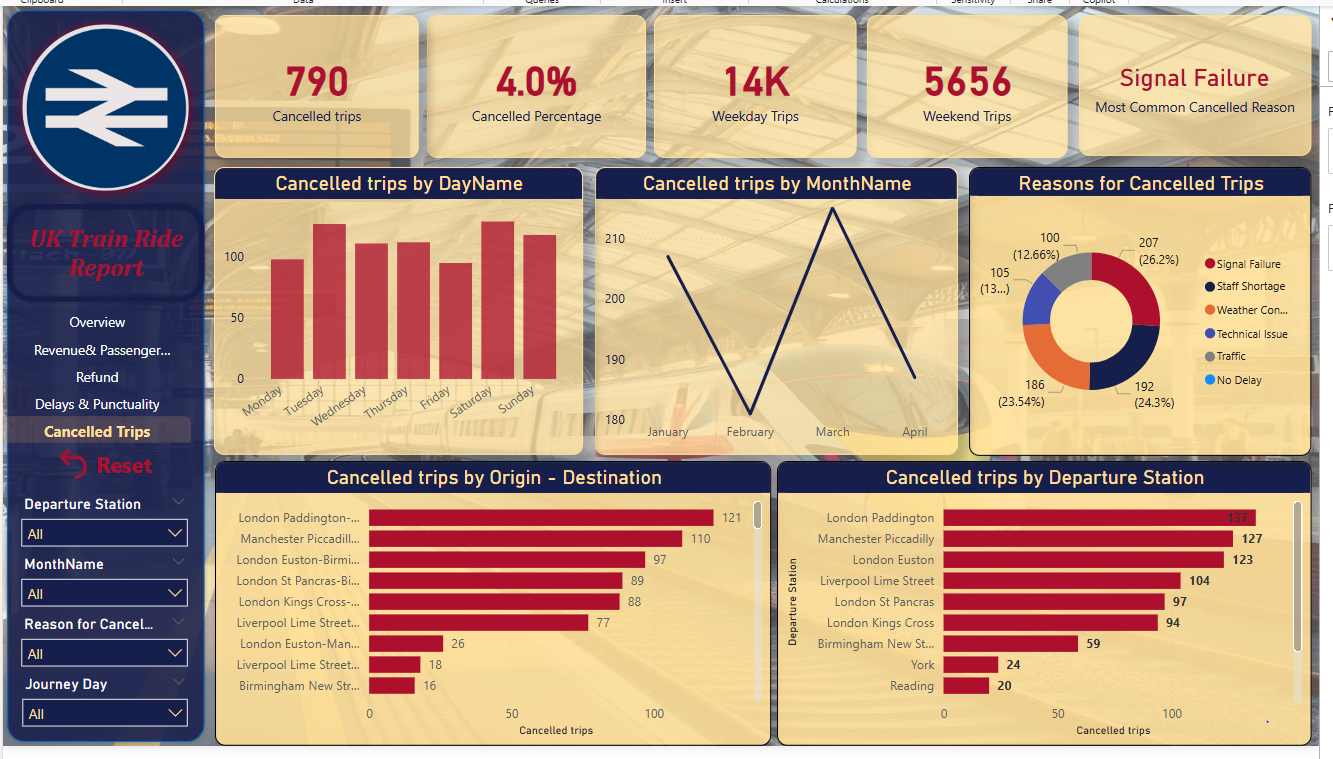
**Top Reasons for Delay (with Trips & Delay Time):**

1. **Weather Conditions:** 466 trips delayed
2. **Staff Shortage:** 437 trips

* **Signal Failures caused fewer trips to be delayed but accounted for the most delay hours, indicating longer impact per incident.**

**Actionable Insights & Recommendations:**

* **Manchester Piccadilly** needs special attention—its connections dominate delay statistics.
* **Signal failure mitigation** could significantly reduce total delay time, even if trip count remains the same.
* **Weather-related delays** are the most frequent—invest in weather-resilient infrastructure and contingency planning.
* **Monday delays** are notably high—optimize staffing and maintenance schedules to reduce early-week service impacts



# Key Observations:

**Cancelled Trips by Day of the Week:**

* **Peak Days:** **Saturday (Highest)** followed closely by **Wednesday and Sunday**
* **Lowest Cancellation Day:** Monday
* **Weekend services, particularly Saturday, face a higher cancellation risk.**

**Cancelled Trips by Month:**

* **March** experienced the **highest cancellations**
* **February** had the **lowest**, suggesting variability across months but no steady trend

**Reasons for Cancelled Trips:**

* **Top 3 Reasons:**
  1. **Signal Failure** – 207 trips (26.2%)
  2. **Staff Shortage** – 192 trips (24.3%)
  3. **Weather Conditions** – 186 trips (23.5%)
* **Other reasons:** Technical Issues (13.3%), Traffic (12.7%)
* **Signal Failures and staffing issues together account for over half of all cancellations.**

**High Cancellation Routes (Origin–Destination):**

* **Top Routes Affected:**
  + **London Paddington – ???:** 121 cancelled trips
  + **Manchester Piccadilly – ???:** 110 trips
  + **London Euston – Birmingham New Street:** 97 trips
  + **London St Pancras – ???:** 89 trips
* **Note: Some route destinations are truncated, but key stations like London Paddington and Manchester Piccadilly are clearly hotspots.**

**Most Affected Departure Stations:**

* **Top 5:**
  1. **London Paddington** – 137 cancellations
  2. **Manchester Piccadilly** – 127
  3. **London Euston** – 123
  4. **Liverpool Lime Street** – 104
  5. **London St Pancras** – 97
* **London-based stations dominate the cancellation list, pointing toward systemic issues in major city hubs.**

**Actionable Insights:**

* **Signal infrastructure upgrades** could significantly reduce cancellations.
* **Staffing reinforcement**, especially on weekends, may help mitigate high Saturday cancellation rates.
* Focus attention on **London Paddington and Manchester Piccadilly** – both top departure and route-based cancellation centers.
* March should be reviewed for **potential seasonal patterns or operational pressures** causing spikes.

# Conclusion:

This project comprehensive analysis of UK train ride data focusing on **revenue, ticket refunds, delays, and cancellations**. The data-driven insights have highlighted several performance areas and improvement opportunities across the rail network.

**Revenue & Refund Insights:**

* **Total Revenue** stood at **$741.92K**, with **$38.7K refunded**, representing a **3.53% refund rate**.
* Majority of refunds were for **Standard class tickets** and **Advance bookings**, primarily purchased **online**.
* **Cancelled journeys (30.4%)** and **delays (23.8%)** were major drivers behind ticket refunds.
* **Edinburgh Waverley** had the **highest refund percentage (100%)**, indicating potential recurring issues in that area.

**Delays & Punctuality Analysis:**

* There were **1,062 delayed trips**, accounting for a **5.3% delay rate**, with an **average delay time of 38.23 minutes**.
* The route **Manchester Piccadilly – Liverpool Lime Street** experienced the **highest total delay time**.
* **Weather conditions** were the most common cause of delays, followed by **staff shortages** and **signal failures**.
* Peak delays occurred mid-week, with **Tuesday and Wednesday** having the most delay hours.

**Cancelled Trips Overview:**

* A total of **790 trips were cancelled**, representing **4.0%** of overall trips.
* **Signal failure** emerged as the leading cancellation cause (26.2%), followed by **staff shortage** and **weather conditions**.
* **Saturday** showed the **highest cancellation rate**, suggesting operational strain over weekends.
* Key affected routes included **London Paddington – [various destinations]** and **Manchester Piccadilly – [various destinations]**.

**Critical Stations Identified:**

* Stations like **London Paddington**, **Manchester Piccadilly**, and **London Euston** were consistently among the **most delayed**, **most cancelled**, and **highest in refund activity**, making them focal points for performance improvements.

**Recommendations**

* 1. Invest in infrastructure, particularly signal systems, to reduce both delays and cancellations.
  2. Reassess staff deployment and training, especially for weekend operations.
  3. Improve real-time communication and contingency planning for weather-related disruptions.
  4. Optimize refund processes and consider predictive analytics to proactively manage high-risk routes or days.
  5. Target operational improvements at high-traffic stations like London Paddington and Manchester Piccadilly.
* This analysis serves as a powerful decision-support tool for rail management, enabling targeted actions to enhance customer satisfaction, improve service reliability, and boost operational efficiency across the UK rail network.

**References**

* **UK Train Rides - Maven Rail Challenge**