FINAL PROJECT PRESENTATION

UK RAILWAY DATA ANALYSIS

Agenda

01 Introductions

02 Objectives

03 Data Cleaning

04 Handling Missing Values

05 Analysis Questions and visualization

06 Our Insights

07 Dashboard

Introduction

Since that UK has one of the oldest railway networks in the world where it handle over 5 million train journeys per day across the country.
Our project utilizes a comprehensive dataset and by transforming these raw records that containing ticket purchases, payment methods, journey schedules, delays, and customer refund into actionable insights, The analysis of this data aims to uncover patterns that influence punctuality, demand trends and the operational factors driving delays.

The goal is to support decision making for service optimization, improve customer experience, and enable railway operators to proactively manage performance challenges.

OBJECTIVES

Ticket Pricing &
Revenue Optimization

Operational Performance
Monitoring

Understanding
Passenger Demand
Patterns

Improving Punctuality &
Delay Analysis and
Reduce Financial Loss

data cleaning and handling missing values

```
!pip install pandas
import pandas as pd
```

```
[notice] A new release of pip is available: 25.1.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
Requirement already satisfied: pandas in c:\users\omarm\anaconda3\lib\site-packages (2.2.2)
Requirement already satisfied: numpy>=1.26.0 in c:\users\omarm\anaconda3\lib\site-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\omarm\anaconda3\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\omarm\anaconda3\lib\site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\omarm\anaconda3\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\omarm\anaconda3\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

```
df = pd.read_csv("railway.csv")
```

```
print(df.duplicated()) #checking duplicates
df.duplicated().sum()  #sum of duplicates
```

```
0        False
1        False
2        False
3        False
4        False
         ...
31648    False
31649    False
31650    False
31651    False
31652    False
Length: 31653, dtype: bool
```

```
0
```

[8]:
```
print(df.isnull())   #checking missing missing values
print(df.isnull().sum())  #sum of missing values in each row
df["Reason for Delay"] = df["Reason for Delay"].str.strip().str.title() #editing text in reasons for delay column
df['Reason for Delay'] = df['Reason for Delay'].replace('Weather', 'Weather Conditions')  #editing text in reasons for delay column
```

```
       Transaction ID  Date of Purchase  Time of Purchase  Purchase Type  \
0               False             False             False          False
1               False             False             False          False
2               False             False             False          False
3               False             False             False          False
4               False             False             False          False
...               ...               ...               ...            ...
31648           False             False             False          False
31649           False             False             False          False
31650           False             False             False          False
31651           False             False             False          False
31652           False             False             False          False

       Payment Method  Railcard  Ticket Class  Ticket Type  Price  \
0               False     False         False        False  False
1               False     False         False        False  False
2               False      True         False        False  False
3               False      True         False        False  False
4               False      True         False        False  False
...               ...       ...           ...          ...    ...
31648           False      True         False        False  False
31649           False      True         False        False  False
31650           False      True         False        False  False
31651           False      True         False        False  False
31652           False     False         False        False  False

       Departure Station  Arrival Destination  Date of Journey  \
```

[9]:
```
df.fillna({"Railcard": "No Railcard"}, inplace=True) #replace missing values in railcard column
```

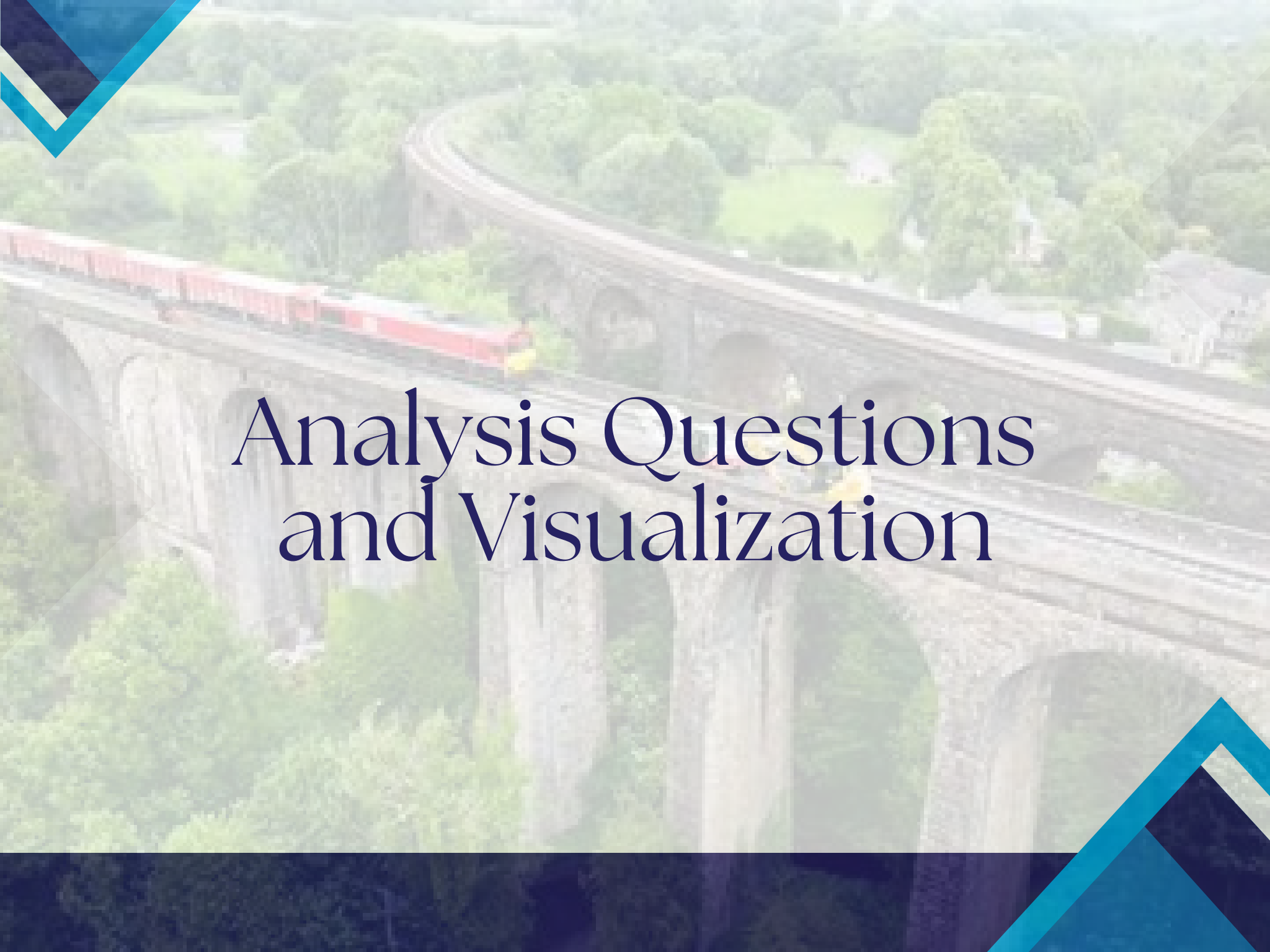
[10]:
```
df.fillna({"Reason for Delay": "On Time"}, inplace=True) #replace missing values in reason for delay
```

[11]:
```
df.fillna({"Actual Arrival Time": "Unknown"}, inplace=True) #replace missing values in actual arrival time
```

[12]:
```
print(df.isnull().sum()) #checking missing values again
```

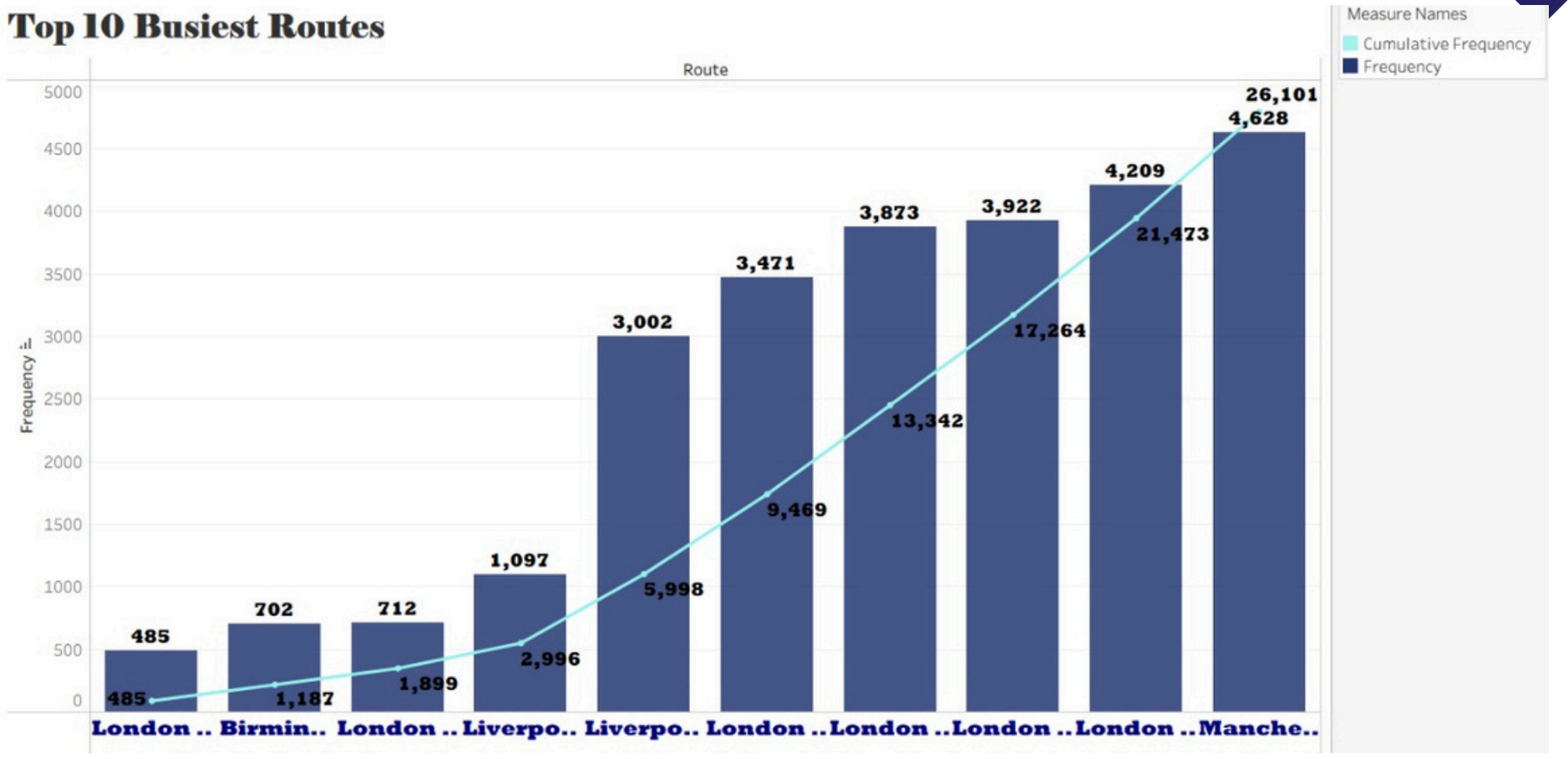
```
Transaction ID         0
Date of Purchase       0
Time of Purchase       0
Purchase Type          0
Payment Method         0
Railcard               0
Ticket Class           0
Ticket Type            0
Price                  0
Departure Station      0
Arrival Destination    0
Date of Journey        0
Departure Time         0
Arrival Time           0
Actual Arrival Time    0
Journey Status         0
Reason for Delay       0
Refund Request         0
dtype: int64
```

[16]:
```
reason_counts = (                       #Q1 ANALYSIS (REASONS OF DELAY)
    df["Reason for Delay"]
    .value_counts()
```



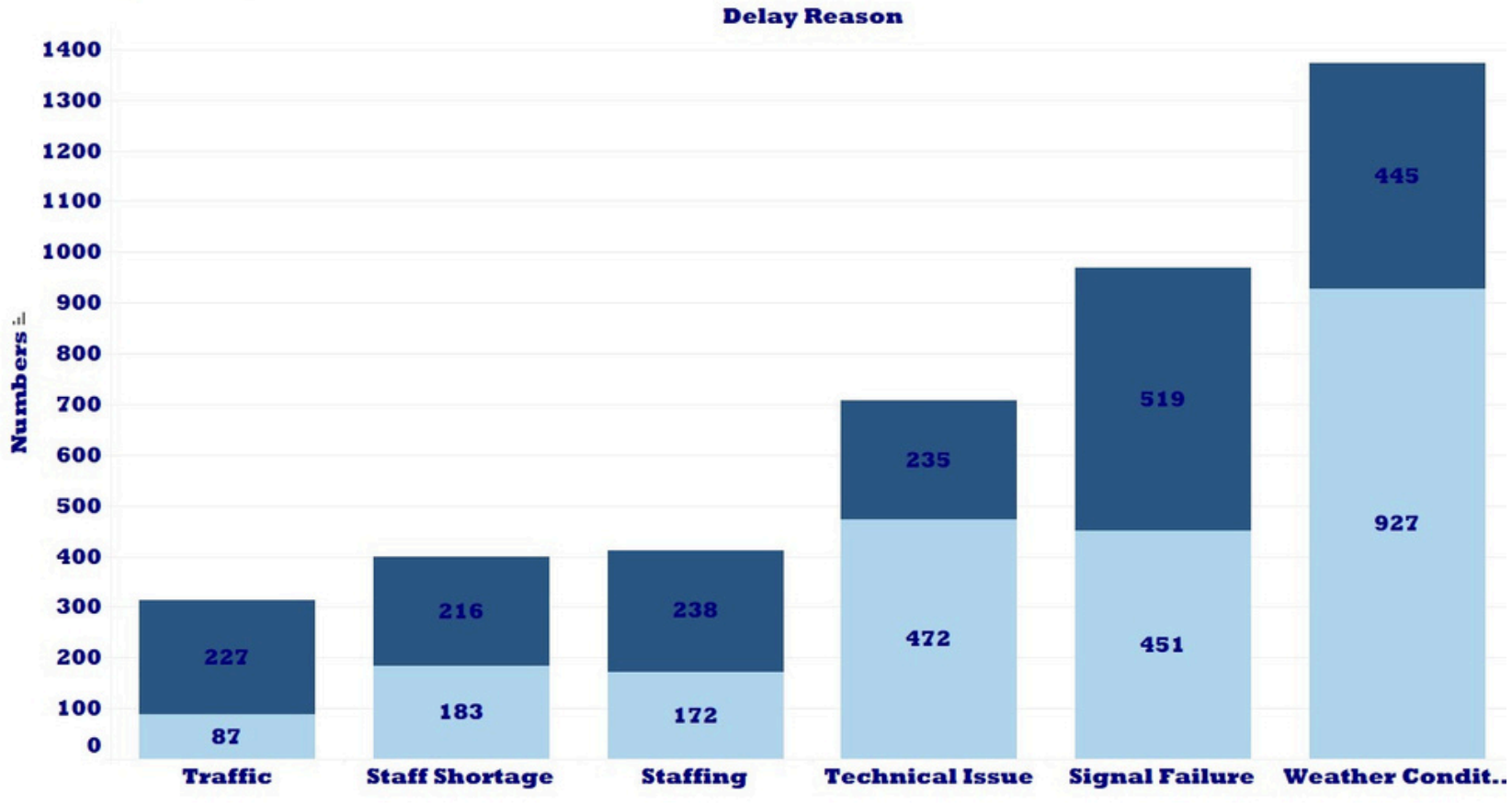
Analysis Questions and Visualization

Top 10 Busiest Routes



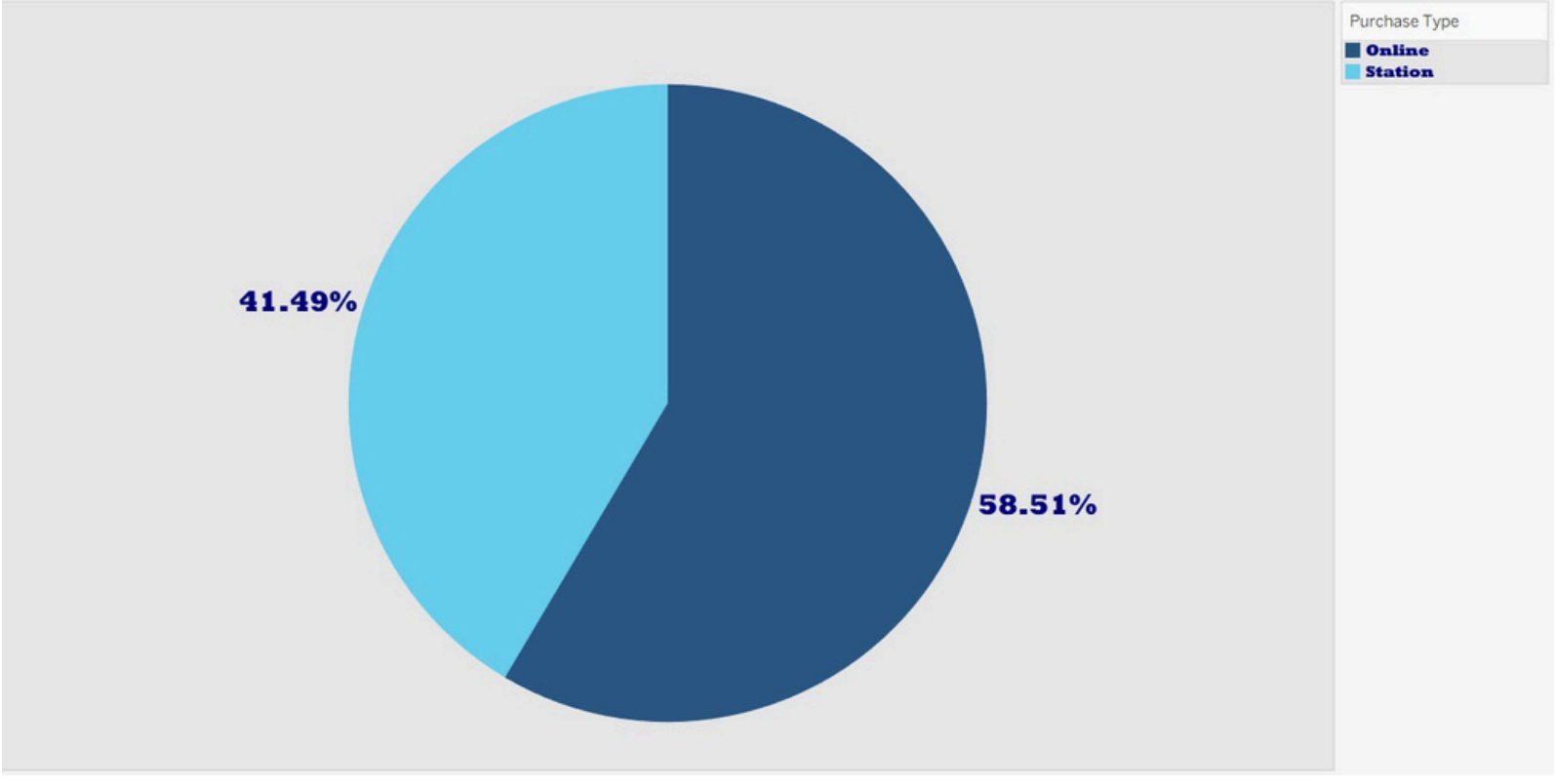
The chart shows the top 10 busiest railway routes with London-based routes dominating the highest traffic levels. The dark blue bars reflect how frequently each route is used, while the light blue cumulative line highlights how these routes collectively contribute to overall demand.

JOURNEY DELAY



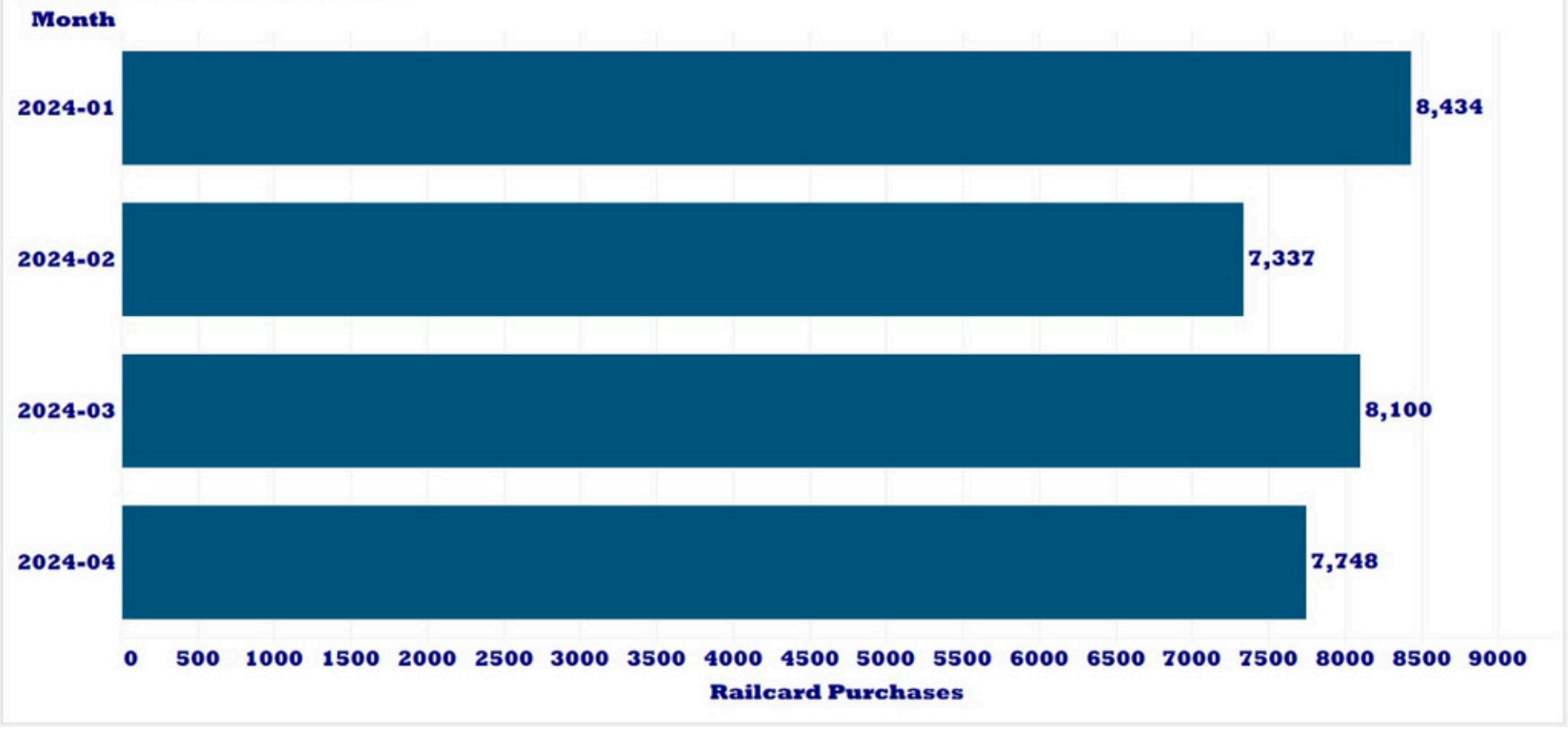
The chart illustrates the distribution of journey delays across six key categories, highlighting which operational issues contribute most to disruptions within the UK railway system. Each bar is split into two segments, representing sub-counts within each delay reason. Overall insight Weather conditions and signal failures stand out as the dominant delay drivers.

Station vs Online sales



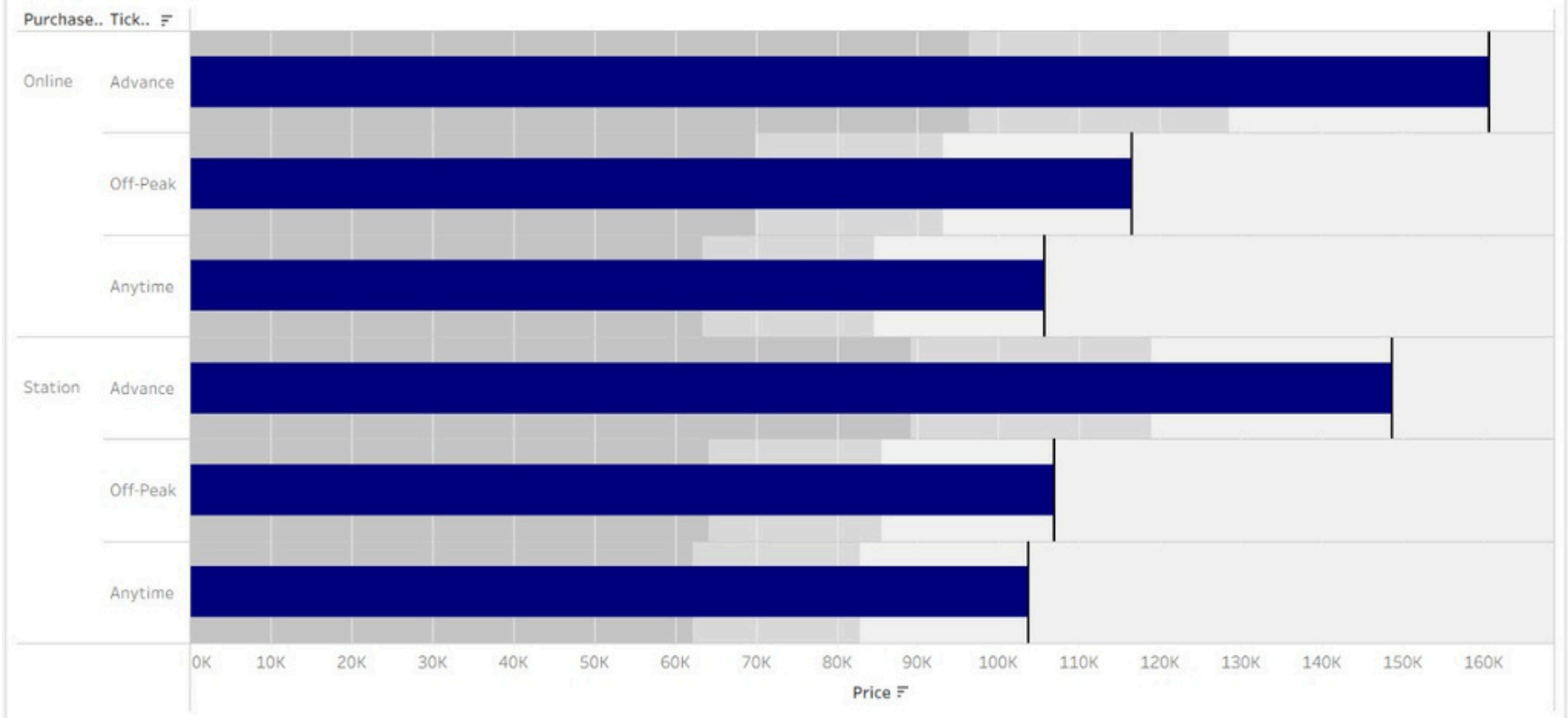
The chart compares the proportion of ticket sales made at physical railway stations versus those completed through online platforms. The distribution shows a clear preference for online channels. Online sales 58.51% and Station Sales 41.49% .

Railcard monthly count



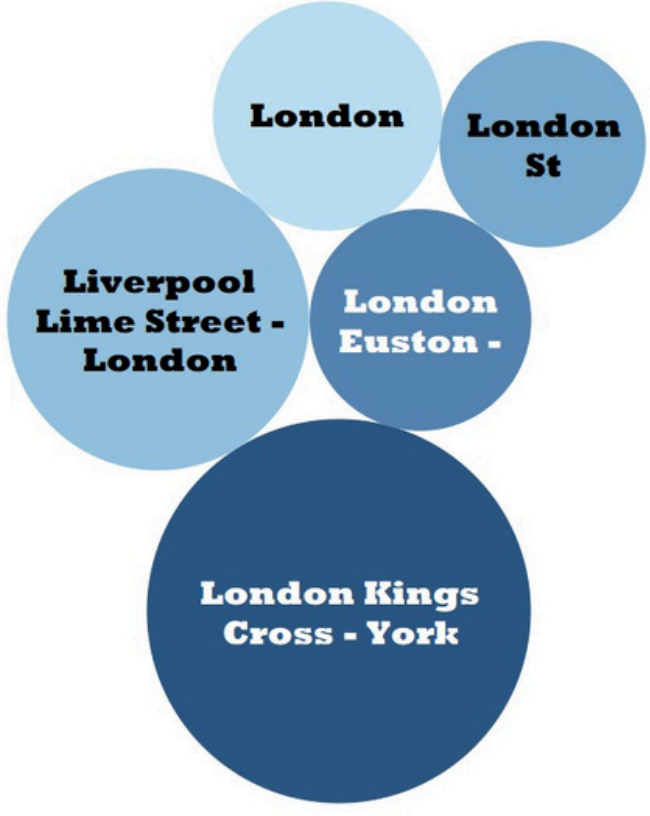
This horizontal bar chart shows the number of railcards purchased across the first four months of 2024, providing insight into seasonal demand and passenger behaviour. Railcard purchases remain consistently high across the four months, with only moderate fluctuations. The pattern highlights strong and stable demand, with January and March standing out as peak months

Ticket Pricing and Sales Performance



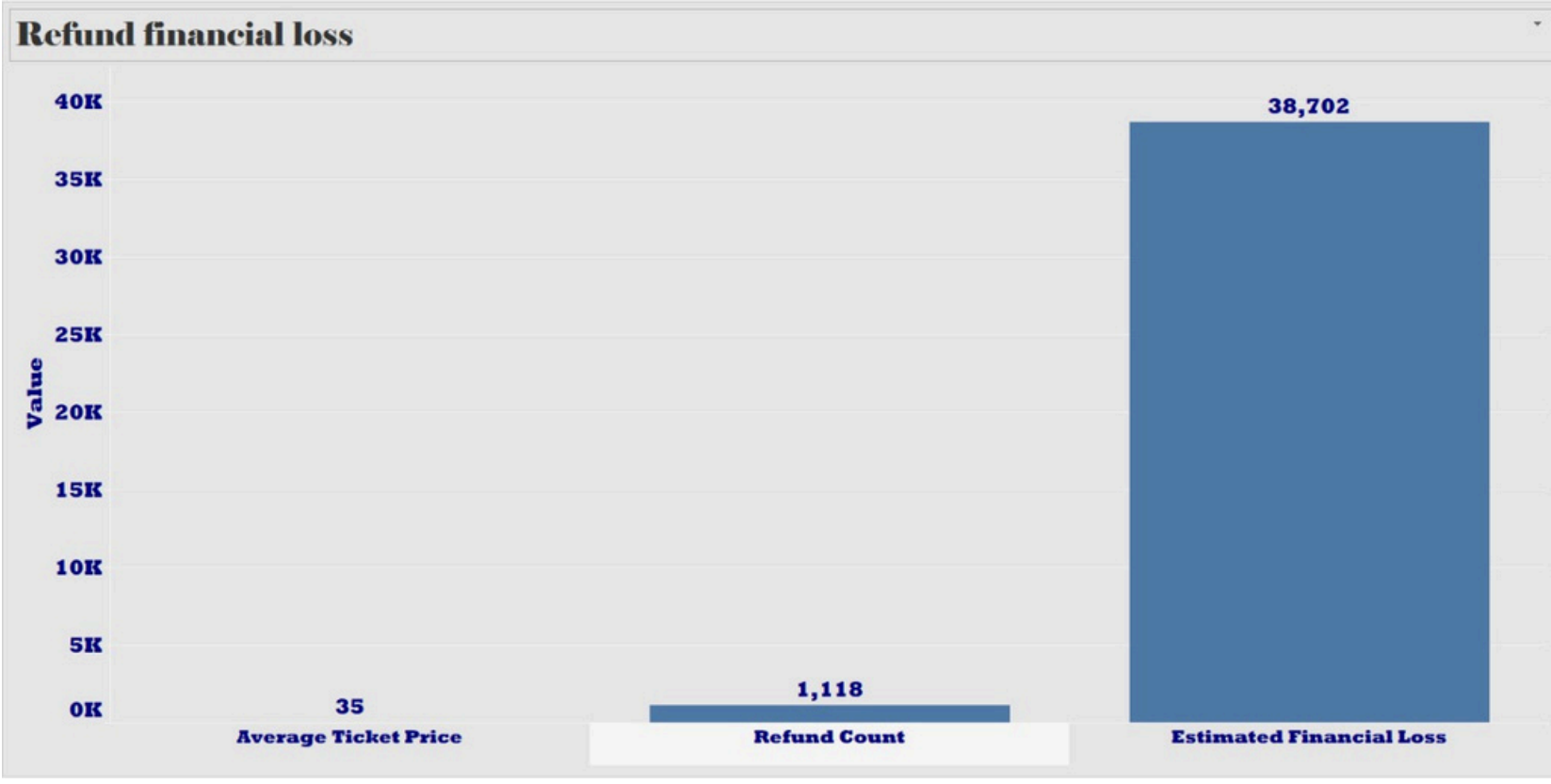
This Bullet graph compares ticket performance across purchase channels (Online vs Station) and ticket types (Advance, Off-Peak, Anytime). Overall, online purchases slightly outperform station purchases, especially for discounted ticket types.

Top 5 Routes by Revenue



This popup graph presents the highest-revenue rail routes, represented by bubbles sized by revenue contribution. Revenue is heavily concentrated on London-linked intercity routes, with London Kings Cross – York generating the greatest revenue.

Refund Financial Loss



The chart compares three key metrics related to ticket refunds: Average Ticket Price – 35, Refund Count – 1,118, Estimated Financial Loss – 38,702.
The high financial loss compared to the small number of refunds indicates that the refund process might be too generous. Improving refund eligibility rules, verifying refund triggers, or reducing avoidable disruptions could significantly cut financial losses without affecting customer satisfaction

OUR INSIGHTS

Railcard purchases peak in January due to annual budget planning and in March as travel increases at the start of spring. This indicates a predictable pattern that can guide marketing timing and promotional campaigns.

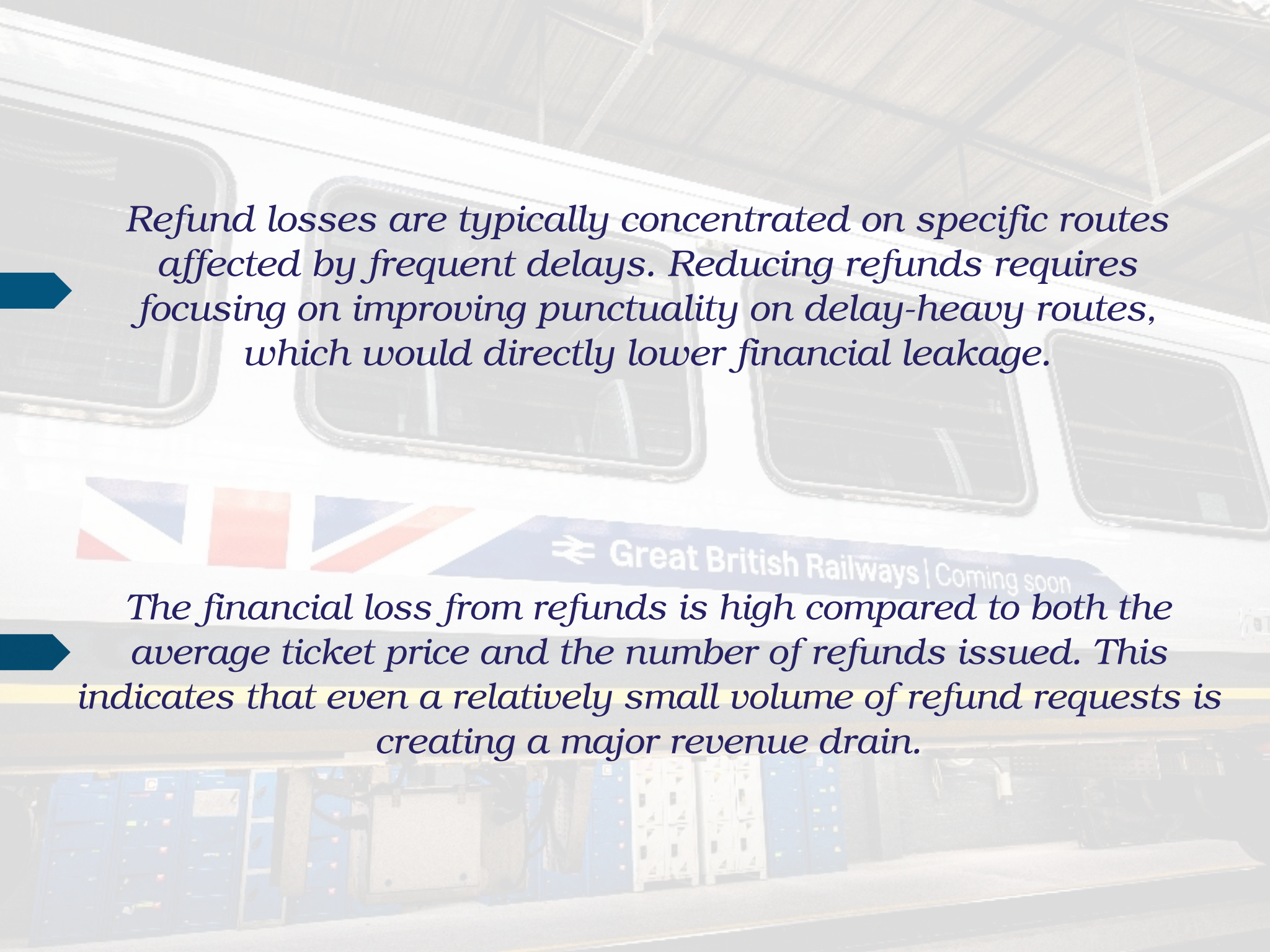
Manchester Piccadilly records the highest traffic volume but does not appear among the top revenue-generating routes. This suggests that high passenger volume does not necessarily translate into financial return. Possible reasons include lower fares, shorter trip distances, or high usage of discounted tickets.

Most of the busiest routes are centered around London, indicating strong demand in and around the capital. This reflects dependency on London's commuter network and suggests that service pressure and operational planning should prioritize these corridors.

Top revenue routes are long-distance intercity lines rather than high-traffic urban routes. Longer journeys with premium fares contribute more to revenue, this shows the importance of maintaining high performance and reliability on intercity connections.

Online sales outperform station sales, especially for discounted ticket types such as Advance and Off-Peak. This indicates that digital users are more price-sensitive and responsive to value-driven offerings.

Weather-related disruptions and signal failures are the leading causes of delays, that means not the staff or scheduling is the main operational challenge. Targeted improvements in signaling systems and weather-proofing would yield the largest punctuality gains.

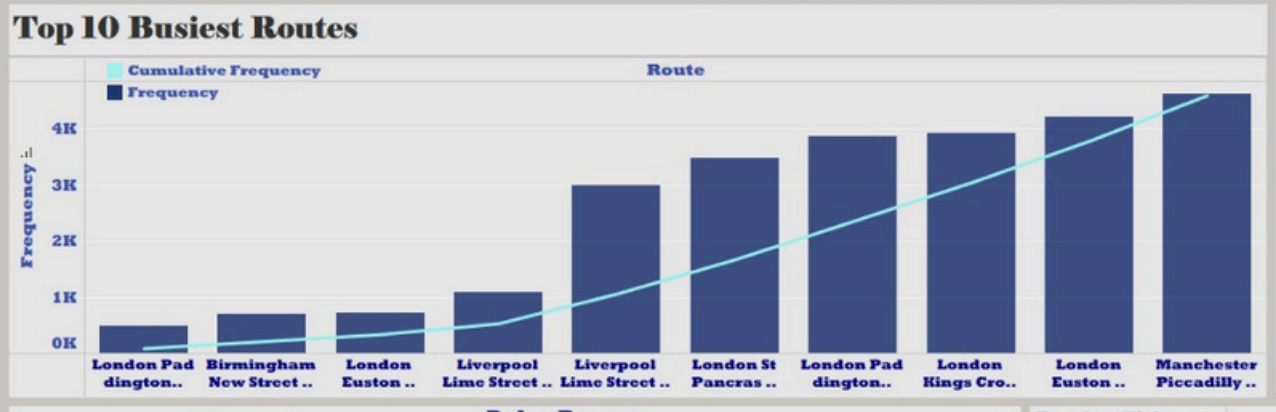


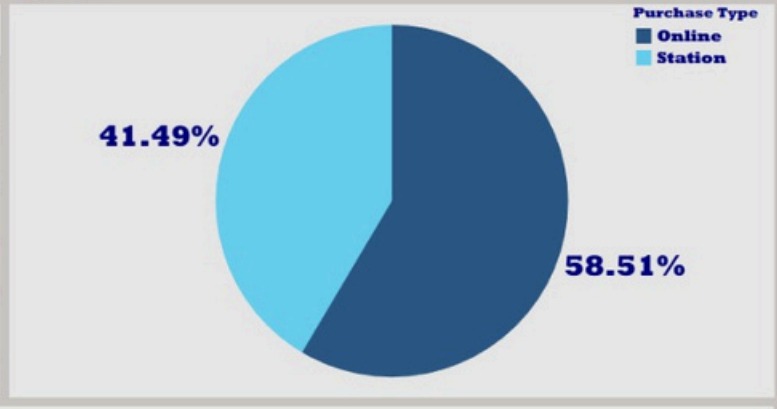
Refund losses are typically concentrated on specific routes affected by frequent delays. Reducing refunds requires focusing on improving punctuality on delay-heavy routes, which would directly lower financial leakage.

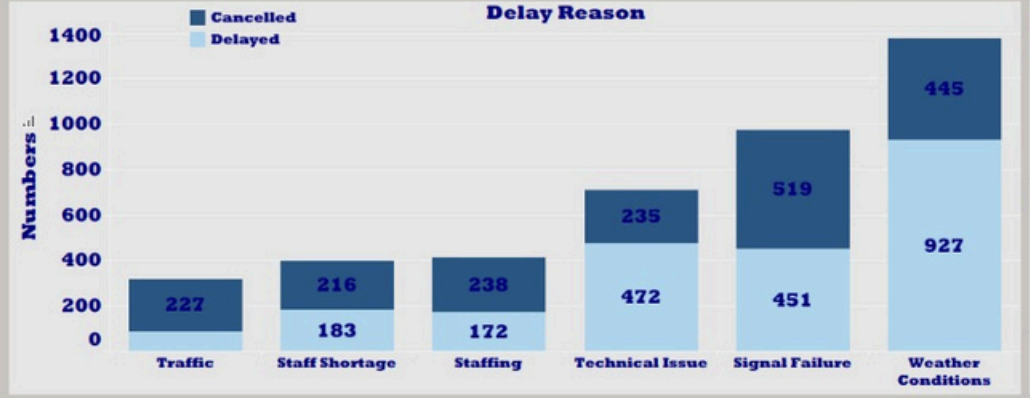
The financial loss from refunds is high compared to both the average ticket price and the number of refunds issued. This indicates that even a relatively small volume of refund requests is creating a major revenue drain.

DASHBOARD

Railway Business Performance Insights

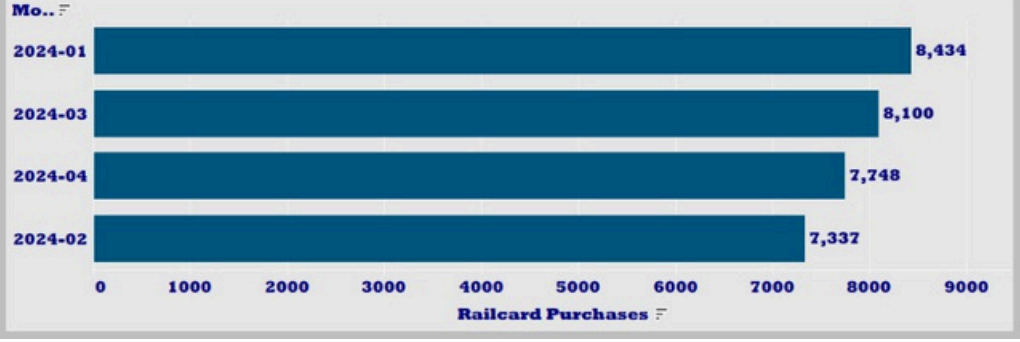






Railway Business Performance Insights

Railcard monthly counts



PRESENTED BY

Omar Mohamed

Ayten Ibrahim

Tasneem Ashraf

Shahd Yasser