Good morning/afternoon everyone. Today, we'll be presenting our analysis of railway transaction data. Our goal is to uncover insights about passenger behaviour, journey patterns, and ticketing trends using a data-driven approach. I'm Savan Golakiya, and I'm here with Maitrik Rajai.

On this slide, we outline the key purpose of our analysis, which is to explore trends in railway transactions. This analysis is significant for railway companies looking to improve their services and optimize operations. We focused on questions like identifying popular stations, understanding delays, and analyzing ticket prices. For this, we used R programming to process and visualize the data effectively.

Our dataset contains over 31,653 records of train ticket sales and journeys, offering a comprehensive look into railway transactions. The key fields in the dataset can be categorized into three areas. First, ticket information, which includes ticket class, type, price, and purchase method. Second, journey details, such as departure and arrival stations, journey times, delays, and reasons for delays. And third, customer behaviour, including purchase timing, railcard usage, and refund requests.

From this dataset, we can derive several insights. For instance, we can identify the most popular routes by analyzing departure and arrival stations. Revenue trends can be explored by looking at how ticket type and class influence earnings. Delay analysis allows us to uncover factors contributing to journey delays, and customer experience can be better understood by examining refund requests and their links to service performance. These key attributes form the basis of our analysis moving forward.

The raw dataset presented several challenges. One significant issue was missing data, particularly for the 'Reason for Delay' field, which was empty for most records. Additionally, date and time fields were inconsistent and needed to be standardized. Finally, we identified redundant or irrelevant fields that were removed to streamline our analysis.

To address these issues, we handled missing data by using placeholders or analyzing such records separately. We also converted date and time fields to a uniform format and filtered out unnecessary fields. This preprocessing ensured that we were working with a clean dataset, allowing us to focus on deriving meaningful insights.

Exploratory Data Analysis is an essential step in understanding our dataset. It helps us identify patterns, trends, and potential issues that require deeper investigation. For example, we found that ticket prices are primarily clustered between $20 and $100, with some outliers.

The top departure stations include Manchester Piccadilly, London Euston, and Liverpool Lime Street, reflecting high passenger volumes in these areas. In terms of journey status, the majority of journeys are 'On Time,' but delays and cancellations still account for a significant number, approximately 15% combined.

Here are some visualizations highlighting these findings, such as the distribution of ticket prices and the breakdown of journey statuses.

Next, we analyzed customer behavior to understand how passengers interact with the railway system. First, we looked at purchase timing and found that a majority of tickets are purchased online, with many being booked on the same day as the journey.

Railcard usage was seen in around 30% of transactions, predominantly for standard class tickets, which indicates cost-saving preferences among customers.

Refund requests were relatively low, at about 3.5% of transactions, and were mostly linked to delays or cancellations. These findings help provide insights into customer preferences and how they respond to service disruptions.

Slide 1: Railway Transaction Analysis

"Good [morning/afternoon], everyone. Today, we’ll be presenting our analysis on railway transactions in the UK. This is a data-driven approach to understanding passenger behavior and improving railway operations.

We’ll explore patterns in ticket purchases, identify delay trends, and demonstrate how machine learning can improve railway management. Let’s get started."

Slide 2: Motivation and Background

"Railway transportation is vital for millions of daily passengers in the UK. However, challenges like delays, inconsistent pricing, and refund policies often impact service quality.

Using a mock dataset of train ticket sales and journey data, our goal was to address questions like: Which are the most popular routes? What causes delays? And how do ticket prices vary under different conditions?"

Slide 3: Dataset Overview

"Our dataset includes over 31,653 records of ticket sales and journeys. Key attributes include ticket information, journey details, and customer behavior, such as refund requests.

With this dataset, we explored trends in popular routes, revenue generation, and delay analysis to uncover actionable insights."

Slide 4: Data Cleaning and Preprocessing

"Before diving into analysis, we faced several challenges. Some data, like 'Reason for Delay,' had missing entries, and time formats were inconsistent.

We resolved these by using placeholders, standardizing time fields, and filtering unnecessary fields. The cleaned dataset provided a solid foundation for our analysis."

Slide 5: Exploratory Data Analysis (EDA)

"EDA helped us identify key trends in the dataset:

Ticket prices mostly range between $20 and $100.

Popular departure stations include Manchester Piccadilly, London Euston, and Liverpool Lime Street.

Most journeys are on time, but delays and cancellations account for 15% of records."

Slide 6: Customer Behavior Analysis

"We analyzed customer behavior and found:

Most tickets are purchased online, often on the same day as the journey.

Railcards are used in 30% of transactions, primarily for standard-class tickets.

Refund requests occur in 3.5% of transactions, with only 27% of delayed or canceled tickets being refunded.

These insights highlight areas to improve, like encouraging advance purchases and simplifying refund processes."

Slide 7: Revenue Insights

"When analyzing revenue, we found that advance tickets generate the highest revenue due to their volume, despite being cost-effective.

Our analysis also showed that the top revenue-generating route is London Kings Cross to York, contributing $183,193. This is likely due to high demand, frequent services, and its economic importance."

Slide 8: Predicted vs. Actual Price (Linear Regression)

"We used a linear regression model to predict ticket prices. The model performed well for tickets priced between $0–$100, but deviations occurred at higher prices.

These deviations indicate that the model could benefit from additional features or more advanced techniques like non-linear regression."

Slide 9: Predicting Actual Arrival Times (Random Forest)

"Using Random Forest, we predicted actual arrival times. The model achieved high accuracy, with most predictions aligning closely with actual times.

Key factors influencing predictions were scheduled arrival time, departure time, and journey status. These insights can be used to improve punctuality and notify passengers of delays in advance."

Slide 10: Conclusion and Insights

"To conclude:

Most tickets are purchased online and on the day of travel.

Railcards are commonly used but mostly for standard-class tickets.

Refund policies need improvement, as only 27% of delayed or canceled tickets have refunds requested.

Signal failure is the leading cause of delays, and advance tickets generate the most revenue due to their high sales volume.

Our Random Forest model demonstrated the potential of predictive analytics in improving operations and customer satisfaction."

Slide 11: Future Work

"Looking forward, there are several opportunities for further work:

Develop dedicated delay prediction models incorporating external factors like weather and passenger volume.

Use clustering to segment customers by preferences for personalized recommendations.

Explore revenue optimization through dynamic pricing and ticket elasticity analysis.

Collaborate with railway companies to test real-time tracking and enhance operational planning.

These steps can significantly advance the use of data analytics in railway management."

Final Notes

Close your presentation with a thank you and open the floor for questions:

"Thank you for your attention. We’re happy to take any questions or discuss our findings further!"