🚆 Train Delay Analysis (Deutsche Bahn)

📌 Project Overview

This project explores **train delay patterns in Germany** using real Deutsche Bahn data from **January–July 2024**.
The workflow demonstrates a complete data analytics pipeline : from raw data cleaning in Python, SQL exploration, statistical analysis, and finally, building an **interactive dashboard in Looker Studio**.

🎯 Motivation

Train delays affect thousands of commuters daily. By analyzing delay data, we can:

- Enable **data-driven decisions** for transport planning
- Show the **value of analytical insights** in real-world scheduling
- Support **smarter operations** and improve passenger experience

🛠️ Tools & Technologies

- **Python (Google Colab)** → Data cleaning, transformation, feature engineering
- **SQL (DuckDB)** → Querying and validating results
- **Looker Studio** → Interactive dashboard visualization
- **Matplotlib / Pandas** → Exploratory plots and statistical checks

📂 Dataset

- Source: Deutsche Bahn train delay records (Jan–Jul 2024)
- Key fields:
- scheduled_time (date/time of service)
- delay_minutes (numeric delay in minutes)
- delay_type (No Delay / Short Delay / Moderate Delay)
- platform_number (extracted from route)
- hbf (station name)

- train_model (train type, e.g., S1, S2, Bus S7)

🔧 Methodology

1. **Data Cleaning**
    - Converted date columns into proper datetime format
    - Combined date + time fields
    - Handled missing or invalid timestamps
    - Removed duplicates and irrelevant bus entries
2. **Feature Engineering**
    - Created delay_minutes
    - Categorized delays into delay_type
    - Extracted platform_number
3. **SQL Analysis**
    - Explored delay distribution by minutes
    - Analyzed delays by station (hbf)
    - Validated Python results with SQL queries
4. **Statistical Analysis**
    - Average delay rate by station
    - Delay trends by time of day
    - Top stations by number of trips
    - Frequency of train models
5. **Visualization in Looker Studio**

- Imported cleaned dataset from Colab
- Built interactive components:
- Table: total delay minutes by platform
- Pie chart: delay category distribution
- Bar chart: record count across delay categories
- Table: average delay rates by station
- Filters: station selector, date range

📊 Key Insights

- **Platforms 1 & 2** recorded the highest total delays (8,594 and 8,055 minutes)
- Delay categories were evenly distributed (~33% each for No Delay, Short Delay, Moderate Delay)
- **München Hbf** had the highest average delay (1.96 minutes)
- **Hamburg Hbf** and **Berlin Hbf** showed lower averages (1.43 and 1.36 minutes)

✅ Conclusion

This project demonstrates:

- End-to-end data analytics workflow
- Cleaning and preparation in Python
- SQL validation and statistical exploration
- Interactive dashboard design in Looker Studio
- Actionable insights for transport planning