Data Cleaning Report

Graduation Project – Data Analysis Track (Power BI Engineer)

Project Info

- Project Title: UK Train Rides Analysis Project
- Team Members: Abdelrahman Maher, Amira Atef, Aya Anwar, Seif Marzouk, Ziad Mohamed.
- Team Leader: Mohamed Alaa.
- Date Submitted: 4/16/2025.
- Dataset Used: UK Railway Train Rides.

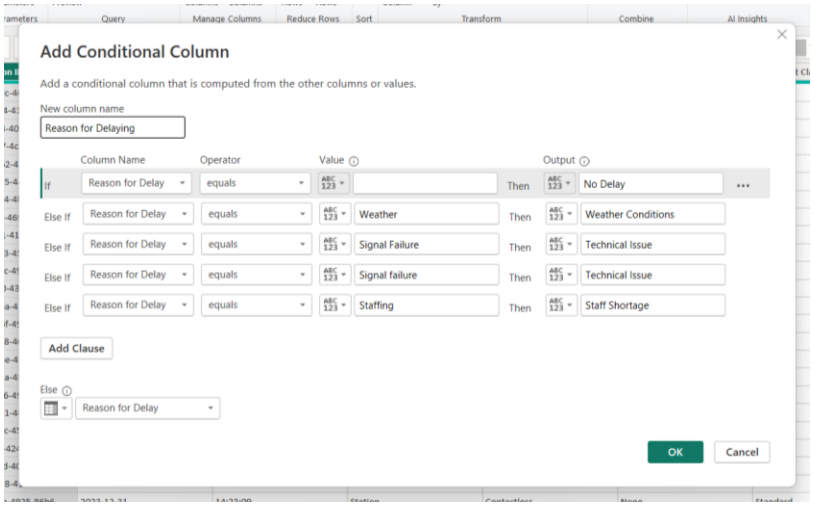
Checking the Data (Assumptions & Data Quality)

What You Checked	What You Found
Missing Values	1880 missing values in "Actual Arrival Time". 27481 missing values in "Reason for Delay".
Duplicates	No Duplicates.
Redundancy	There's repetition in journey info.
Inconsistent Values	"Reason for Delay" consists of many different values with the same meaning.

Steps You Did in Power Query

Step 1: Added Conditional Column "Reason for Delaying"

- **How:** Added some conditions to fix "Reason for Delay" column problems.
- **Reason:** A lot of values in "Reason for Delay" were having the same meaning such as : (Technical Issue, Signal **F**ailure, Signal **f**ailure), (Weather, Weather Conditions),(Staffing, Staff Shortage) and it contains blanks.
  Note : Old column "Reason for Delay" , New column "Reason for Delay**ing**"

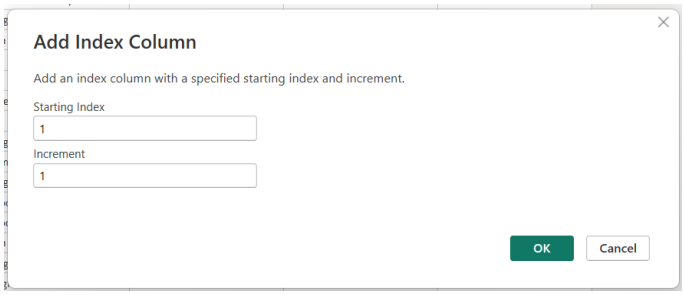


Step 2: Splitted data into 2 tables

- **How:** Created dimension table “dJourney” and a fact one “fRailway”
- **Reason:** Normalizing the data to build efficient and scalable model.

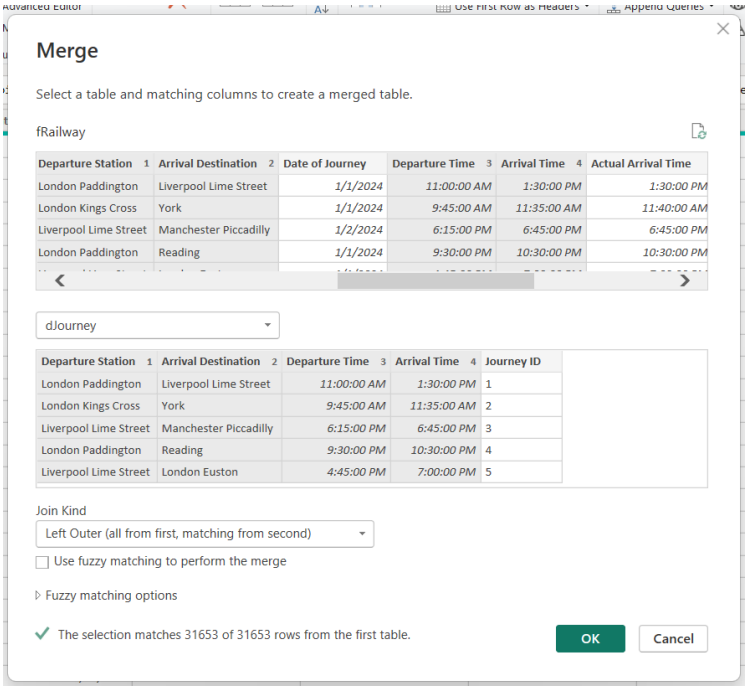
Step 3: Adding primary key for the dimension table

- **How:** Added index column in “dJourney” and renamed it “Journey ID” after removing duplicated rows.
- **Reason:** To merge it with “fRailway” table.



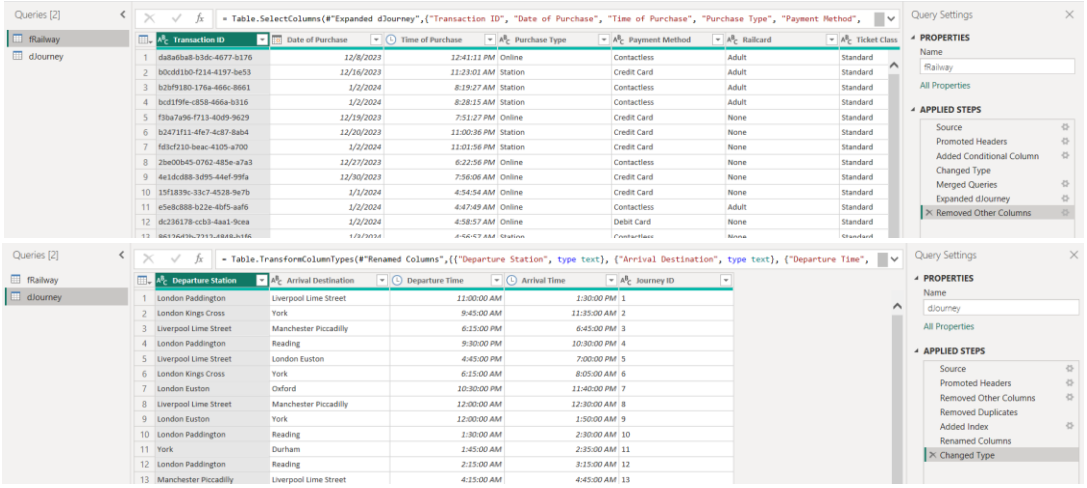
Step 4: Merged queries

- **How:** Merged “fRailway” with “dJourney” by [Departure Station, Arrival Destination, Departure Time, Arrival Time].
- **Reason:** To organize data for better analysis and decision-making.



Final Clean Dataset (Before & After)

Before Cleaning	After Cleaning
1 Table **fRailway had 18 columns** **31653 rows**	2 Tables. 15 columns. 31653 rows.
27481 missing values in "Reason for Delaying"	0 missing values.



Problems You Faced & How You Solved Them

Problem: Rides which were cancelled, left missing values in "Actual Arrival Time".
Solution: Kept them as they are because any time value will affect data integrity and any other value will change column type which is necessary in measures.

Problem: "Reason for Delay" column had many problems.
Solution: Added conditional column "Reason for Delaying" and replaced all the unwanted values then deleted the old column.

Problem: No primary key for “dJourney”.
Solution: removed duplicated rows then added index column starts from 1 and named “Journey ID”.