

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.

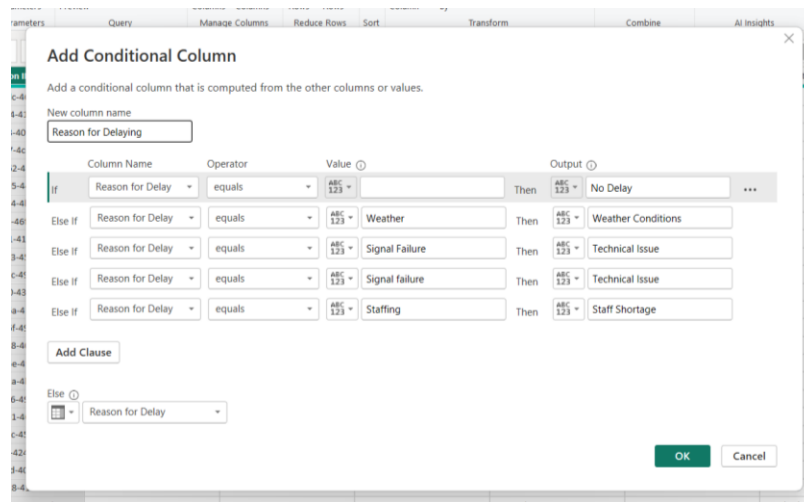
Checking the Data (Assumptions & Data Quality)

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 Failure, Signal failure), (Weather, Weather Conditions),(Staffing, Staff Shortage) and it contains blanks.
Note : Old column "Reason for Delay" , New column "Reason for Delaying"

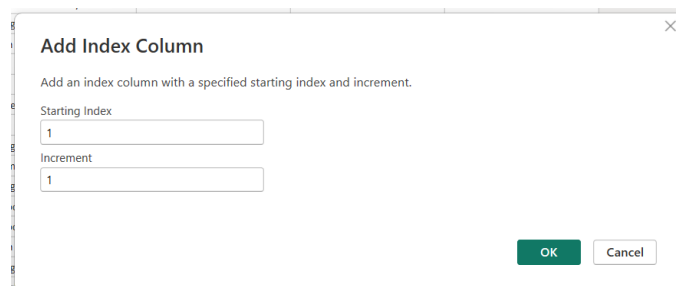


Step 2: Splitted data into 2 tables

- **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.

Merge

Select a table and matching columns to create a merged table.

fRailway

Departure Station	Arrival Destination	Date of Journey	Departure Time	Arrival Time	Actual Arrival Time
London Paddington	Liverpool Lime Street	1/1/2024	11:00:00 AM	1:30:00 PM	1:30:00 PM
London Kings Cross	York	1/1/2024	9:45:00 AM	11:35:00 AM	11:40:00 AM
Liverpool Lime Street	Manchester Piccadilly	1/2/2024	6:15:00 PM	6:45:00 PM	6:45:00 PM
London Paddington	Reading	1/1/2024	9:30:00 PM	10:30:00 PM	10:30:00 PM

dJourney

Departure Station	Arrival Destination	Departure Time	Arrival Time	Journey ID
London Paddington	Liverpool Lime Street	11:00:00 AM	1:30:00 PM	1
London Kings Cross	York	9:45:00 AM	11:35:00 AM	2
Liverpool Lime Street	Manchester Piccadilly	6:15:00 PM	6:45:00 PM	3
London Paddington	Reading	9:30:00 PM	10:30:00 PM	4
Liverpool Lime Street	London Euston	4:45:00 PM	7:00:00 PM	5

Join Kind

Left Outer (all from first, matching from second)

☐ Use fuzzy matching to perform the merge

↳ Fuzzy matching options

✓ The selection matches 31653 of 31653 rows from the first table.

OK Cancel

Final Clean Dataset (Before & After)

Before Cleaning

1 Table
fRailway had 18 columns
31653 rows

27481 missing values in
"Reason for Delaying"

After Cleaning

2 Tables.
15 columns.
31653 rows.
0 missing values.

Queries [2]

Table.SelectColumns("#Expanded dJourney",{"Transaction ID", "Date of Purchase", "Time of Purchase", "Purchase Type", "Payment Method", "Railcard", "Ticket Class", "Source", "Promoted Headers", "Added Conditional Column", "Changed Type", "Merged Queries", "Expanded dJourney", "Removed Other Columns", "Removed Duplicates", "Added Index", "Renamed Columns", "Changed Type" })

Transaction ID	Date of Purchase	Time of Purchase	Purchase Type	Payment Method	Railcard	Ticket Class
da8a0ab-b5d-4677-9176	12/6/2023	12:41:11 PM	Online	Contactless	Adult	Standard
b0cd1310-7214-4197-9e53	12/16/2023	11:23:01 AM	Station	Credit Card	Adult	Standard
b2b9f180-176a-466c-8861	1/2/2024	8:19:27 AM	Station	Contactless	Adult	Standard
bcd179fe-d358-466a-6316	1/2/2024	8:28:15 AM	Station	Contactless	Adult	Standard
f0ba7496-f713-4009-9629	12/19/2023	7:51:17 PM	Online	Credit Card	None	Standard
b2471711-4fe7-4c57-9ab4	12/20/2023	11:00:36 PM	Station	Credit Card	None	Standard
f45cd210-beac-4326-a700	1/2/2024	11:01:56 PM	Station	Credit Card	None	Standard
3be0b4d3-0762-425e-a7a3	12/27/2023	6:22:56 PM	Online	Contactless	None	Standard
4e1dc885-3495-44af-99fa	12/30/2023	7:56:06 PM	Online	Credit Card	None	Standard
15f1839c-33c7-4528-9e7b	1/2/2024	4:54:54 AM	Online	Credit Card	None	Standard
e5e8c888-b22e-40f5-aaf6	1/2/2024	4:47:49 AM	Online	Contactless	Adult	Standard
8c236178-ccb3-4a41-9c8a	1/2/2024	4:58:57 AM	Online	Debit Card	None	Standard
861746f0-7913-4248-b086	1/3/2024	4:56:47 AM	Station	Contactless	None	Standard

Queries [2]

Table.TransformColumnTypes("#Renamed Columns",({ "Departure Station", type text }, {"Arrival Destination", type text }, {"Departure Time", type text }, {"Arrival Time", type text }, {"Journey ID", type text }))

Departure Station	Arrival Destination	Departure Time	Arrival Time	Journey ID
London Paddington	Liverpool Lime Street	11:00:00 AM	1:30:00 PM	1
London Kings Cross	York	9:45:00 AM	11:35:00 AM	2
Liverpool Lime Street	Manchester Piccadilly	6:15:00 PM	6:45:00 PM	3
London Paddington	Reading	9:30:00 PM	10:30:00 PM	4
Liverpool Lime Street	London Euston	4:45:00 PM	7:00:00 PM	5
London Kings Cross	York	6:15:00 AM	6:05:00 AM	6
London Euston	Coventry	10:30:00 AM	11:40:00 AM	7
Liverpool Lime Street	Manchester Piccadilly	12:00:00 AM	12:30:00 AM	8
London Euston	York	12:00:00 AM	1:50:00 AM	9
London Paddington	Reading	1:30:00 AM	2:30:00 AM	10
York	Durham	1:45:00 AM	2:35:00 AM	11
London Paddington	Reading	2:15:00 AM	3:15:00 AM	12
Manchester Piccadilly	Liverpool Lime Street	4:15:00 AM	4:45:00 AM	13

Query Settings

Properties

Name

fRailway

Applied Steps

Source

Promoted Headers

Added Conditional Column

Changed Type

Merged Queries

Expanded dJourney

Removed Other Columns

Removed Duplicates

Added Index

Renamed Columns

Changed Type

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".