RAILWAY DATA ANALYSIS REPORT

PRESENTED TO
ENG. MENNA TAREK

RAILWAY RAW DATA DESCRIPTION

Transaction ID: Unique identifier for an individual train ticket purchase.

Date of Purchase: Date the ticket was purchased.

Time of Purchase: Time the ticket was purchased.

Purchase Type: Whether the ticket was purchased online or directly at a train station.

Payment Method: Payment method used to purchase the ticket (Contactles, Credit Card, or Debit Card).

Railcard: Whether the passenger is a National Railcard holder (Adult, Senior, or Disabled) or not.

Ticket Class: Seat class for the ticket (Standard or First).

Ticket Type: When you bought or can use the ticket. Advance tickets are 1/2 off and must be purchased at least a day prior to departure. Off-Peak tickets are 1/4 off and must be used outside of peak hours (weekdays between 6-8am and 4-6pm). Anytime tickets are full price and can be bought and used at any time during the day.

Price: Final cost of the ticket.

Departure Station: Station to board the train.

Arrival Destination: Station to exit the train.

Date of Journey: Date the train departed.

Departure Time: Time the train departed.

Arrival Time: Time the train was scheduled to arrive at its destination (can be on the day after departure).

Actual Arrival Time: Time the train arrived at its destination (can be on the day after departure).

Journey Status: Whether the train was on time, delayed, or cancelled.

Reason for Delay: Reason for the delay or cancellation.

Refund Request: Whether the passenger requested a refund after a delay or cancellation.

CLEANING STEPS

- **CLEANING ON EXCEL POWER QUERY:**

1. removing duplicates.
2. changing data type of the time of purchase , interarrival time and actual arrival time to time.
3. adding column named purchase time interval to determine Morning and Evening
4. splitting the Arrival stations column to Arrival station-city and Arrival Station-place to ease anlysis on city and display a map .
5. adding calculated field named Arrival delay [Actual Arrival time - Arrival time]
6. adding calculated field named journey Time [Actual Arrival time - Departure time]

- **CLEANING USING PYTHON:**

1. filling nulls in Reason for Delay field with "Non Cancelled nor Delayed".
2. detected for outliers on the price field but didnot remove them as they are not a data false they were descrbtive.

POWER BI

- **MEASURES ADDED(DAX):**

1. **Company Debt :** is the sum of price where refund request = "NO" and journey status = "cancelled".
2. **No.of Transactions :** count of transactions.
3. **ON time Percentage :** is the percentage of journey status = "On time" to see the KPI of the train company.
4. **Revenue** : sum of price.

Displaying them on cards.

- **OPTAINED INSIGHTS:**

- **the data contains of 32k row .**

- **Revenue equals 741.9k$.**

- **percent of on time arrival 86.82%.**

- **percentage of ticket type:**
  1. Anytime: 47.61%.
  2. off-peak: 31%.
  3. Advanced: 21.39%.

- **Average price of ticket type:**
  1. Anytime: 39.2$.
  2. off-peak: 25.5$
  3. Advanced: 17.6$.

- **count of Delay contribution factors:**
  1. Weather: 995.
  2. signal failure: 970.
  3. Staffing: 410.
  4. Traffic: 314.

- **Top 5 Destinations:**
  1. Durham: 258.
  2. Didcot: 48.
  3. cardiff Central: 16
  4. wakefield : 15.
  5. warrington: 15.

- **Avg Revenue per Month**
  1. the highest revenue avg month is February
  2. the least revenue avg month is october.

- **Ticket Class Revenue:**
  1. standard , Online : 303,198$
  2. standard,Station : 289,324$
  3. FirstClass, Online : 79,556$
  4. FirstClass , Station:69,843$

REPORTED BY:

- SALMA SAMEH.
- AYA BADAWY.
- HABIBA AMR.
- MAYADA OUN.
- SHEROUK YASSER.

THANKS!