1. Data Collection: Data is taken from: dataset

2. Data Cleaning:

Data is cleaning using Python Packages: numpy and pandas in Jupiter notebooks - Detailed code can be found in the below github code

Actions Taken:

1. Corrected the data types of the features.
2. Found no duplicate values.
3. Imputed the missing values.
4. Created new features as required for the analysis.
5. Created a dictionary to store the outputs of the analysis and saved it as a pickle file for reusability.

Source code : **GitHub**

3. Data Analysis:

Objective:

For the Maven Rail Challenge, a company that provides business services to passenger train operators in England, Scotland, and Wales.

The manager requested to create an exploratory dashboard that helps them:

- Identify the most popular routes
- Determine peak travel times
- Analyze revenue from different ticket types & classes
- Diagnose on-time performance and contributing factors

Sharing Insights:

Insights 1 :

- Birmingham New Street is top Arrival Station
- Manchester Piccadilly is the top Departure station
- **Liverpool Lime Street to Manchester Piccadilly is the top Route**

Insights 2:

- The peak travelling times are as in the following order where Morning is in the first and Early morning is in the last position: (Day is divided into 4 parts) Morning 34.70% Afternoon 29.42% Evening/Night 20.88% Early Morning 15.00%
- March & January are the top travelling months
- 6 PM & 6 AM are the busiest travelling hours

Insights 3:

- Revenue for Standard class is high compared to First Class, and as well as the number of passenger who bought the ticket & around 50% of the passengers in both Ticket Classes are Railcard holders. So it wont be a factor for low prices in First Class transactions
- The revenue for Advance & Off Peak booking is high compared to Anytime - This is interesting because Both have discounts/offer for the price but not for Anytime category. The reason for high revenue could be the no of passenger count in each Ticket Type. Also There are 66% of the passengers are having Railcard from Anytime category so this could be one of the factors for low revenue from Anytime category
- Combination: The Standard Class with Advance Type Bookings are high compared to other

Insights 4:

- Most of the trains are On-Time with 86.82% only 7.24% are delayed and rest of the percentage are from cancelled trains
- **Contributing factors for the train delay** in percentages: Weather is the primary factor for the delays

4. Data Visualization:

Plotly Package was used to create Visualizations/Dashboard in Python and hosted the visuals website using streamlit

Website link: https://trainridesproject.streamlit.app/

5. Recommendations:

- **Since the route from Liverpool Lime Street to Manchester Piccadilly and the morning and afternoon times are the busiest, we should ensure that all passenger facilities are available in advance. Additionally, planning for new trains on this route could be beneficial.**
- **As there are a few routes with high delay rates, management needs to take steps to identify the root causes and address them before the impact becomes significant.**