UK Train Rides

Project by:

"Insight Innovators"

YAT422B_GIZ2_DAT1_G3

1. Sarah Farag
2. Sahar Hamdi
3. Nada Saad
4. Mona Hassan
5. Sara Mohamed
6. Dina Yasser

UK Train Rides Dataset

Introduction:

This project aims to analyze railway journey data from January 1st, 2024, to April 30th, 2024. The dataset includes crucial information regarding railway card usage, journey status, delay reasons, and other relevant details. By comprehensively examining this data, we aim to uncover significant insights and patterns to improve railway operations and enhance passenger experience.

Business Outcomes:

- Improve operational efficiency by identifying delay patterns.
- Enhance passenger satisfaction through better resource allocation and scheduling.
- Increase revenue by understanding booking trends and passenger behavior.

Problem Statement:

The dataset reveals significant issues, such as high numbers of delays and missing railway card data. These issues impact operational efficiency and passenger experience.

Pain Points:

- Frequent delays causing passenger dissatisfaction.
- Incomplete data affecting the accuracy of analysis.
- Lack of insights into the reasons for delays and their impact on revenue.

Objectives:

1. ***Data Cleaning and Preparation:***
    - The dataset contains over 20,000 null values in the "Railway Card" column. These null values will be replaced with "Unknown" to ensure completeness.
    - All null values in “Reason for Delay” will be replaced with “No Delay”
    - All null values in “Actual Arrival Time” will be replaced with “Cancelled” or "N/A" or “unique time”
    - All necessary columns with object data types will be converted to appropriate date/time formats for accurate analysis.

2. ***Standardization:***

   - The values in the "Reason for Delay" column will be unified into specific categories:

     1. "Weather" will be "Weather Conditions,"
     2. "Signal failure" will be "Signal Failure,"
     3. and "Staffing" will be "Staff Shortage."

3. ***Research and Analysis:***

   - Conduct research on various weather conditions that affected railway journeys during the specified period.
   - Investigate the impact of public holidays on travel patterns, including increased bookings and revenue trends for specific destinations.
   - Analyze ticket types to study the behavior of passengers, particularly focusing on the demand for refunds and compensations in cases of cancellations or delays.

By achieving these objectives, we aim to gain a comprehensive understanding of the factors influencing railway journeys and delays. The insights derived from this analysis will be instrumental in identifying areas for improvement and implementing strategies to enhance the overall efficiency and passenger satisfaction in railway services.

Roles and Responsibilities with Timelines

Phases	Milestones	Tools	Timelines	Deliverables	Assigned to
Data Exploration and Preprocessing	Examining the data to understand its structure, main characteristics, and discover patterns or relationships.	MS Excel Python (pandas, Matplotlib)	2 weeks	Project Documentation	The Whole Group
Data Cleaning and Preparation	Perform data cleaning, transformation, and analysis.	MS Excel Python (pandas, Matplotlib)	2 weeks	1. Cleaned dataset ready for analysis 2. Data preprocessing notebook.	Sarah Farag Sahar Hamdi Nada Saad
Data Analysis and Insights Generation	Provide insights into railway operations and help interpret analysis results.	MS Excel Python (pandas, Matplotlib)	2 weeks	Set of analysis questions that can be answered via the dataset	Mona Hassan Sara Mohamed Dina Yasser
Forecasting Questions Phase	Forecasting questions must include the prediction of number of rides for the next month. Then, accordingly, highlighting the forecasted revenue during each day of the next month. Also, you need to specify the demand on different ticket classes.	Python (pandas, Matplotlib)	2 weeks	Visualization plots answering forecasting questions	
Visualization Dashboard and Final Presentation	Build a Tableau visualization dashboard that visualize the answers to all answered questions.	Tableau	2 weeks	1. Visualization dashboard 2. Final report and presentation.	The Whole Group

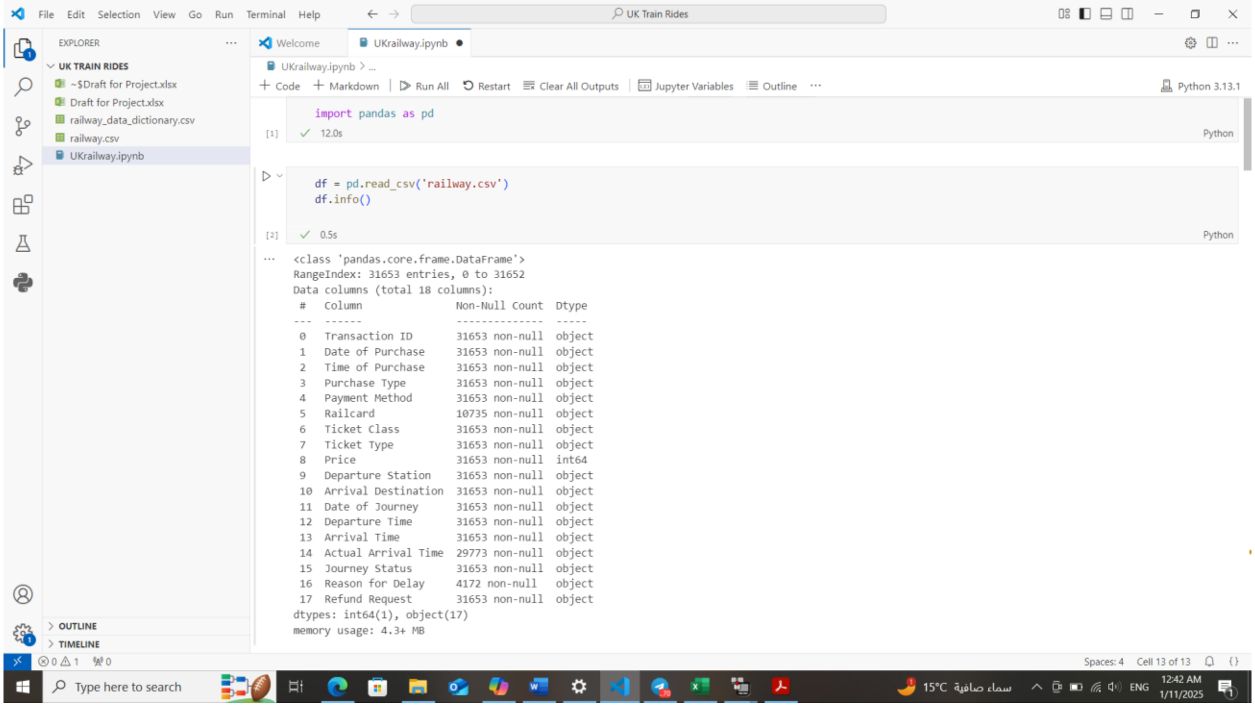
1st Phase: Data Exploration and Preprocessing

Data Exploration with Python: Unveiling Insights from UK Train Rides

The efficient operation and punctuality of train services play a crucial role in ensuring the seamless transportation of passengers across the UK. By using the powerful capabilities of Python, we can delve into the vast datasets related to UK train rides, uncovering valuable insights that inform decisions, improve services, and enhance customer satisfaction. Python, with its libraries such as Pandas, NumPy, and Matplotlib, serves as a versatile and efficient tool for data exploration and analysis. These libraries enable us to clean, manipulate, and visualize data effectively, transforming raw datasets into meaningful information.

In this exploration, we aim to analyze various aspects of UK train rides, including purchase patterns, ticket pricing, punctuality, and delays. By examining the data, we can identify trends, uncover anomalies, and derive actionable insights that can help optimize train services and improve the overall passenger experience.

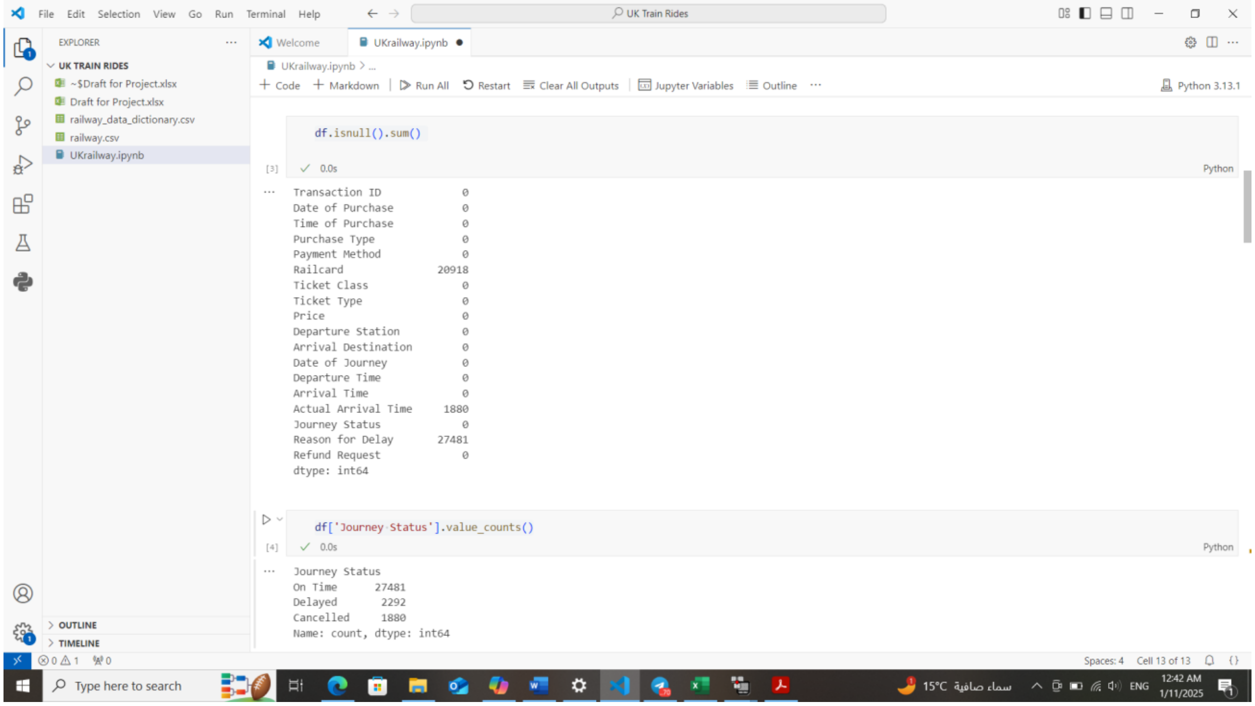
Our data exploration journey begins with data cleaning and preprocessing, where we handle missing values, standardize formats, and ensure data consistency. Next, we dive into descriptive statistics and visualizations, providing a comprehensive overview of the dataset. Through this process, we aim to answer key questions about purchase methods, ticket types, station activities, and delay reasons. By harnessing the power of Python, we embark on a systematic and insightful exploration of UK train rides, paving the way for data-driven decision-making and enhanced operational efficiency.



```
[1]
import pandas as pd
✓ 12.0s
```

```
[2]
df = pd.read_csv('railway.csv')
df.info()
✓ 0.5s
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31653 entries, 0 to 31652
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   Transaction ID       31653 non-null  object
 1   Date of Purchase     31653 non-null  object
 2   Time of Purchase     31653 non-null  object
 3   Purchase Type        31653 non-null  object
 4   Payment Method       31653 non-null  object
 5   Railcard             10735 non-null  object
 6   Ticket Class         31653 non-null  object
 7   Ticket Type          31653 non-null  object
 8   Price                31653 non-null  int64 
 9   Departure Station    31653 non-null  object
 10  Arrival Destination  31653 non-null  object
 11  Date of Journey      31653 non-null  object
 12  Departure Time       31653 non-null  object
 13  Arrival Time         31653 non-null  object
 14  Actual Arrival Time  29773 non-null  object
 15  Journey Status       31653 non-null  object
 16  Reason for Delay     4172 non-null   object
 17  Refund Request       31653 non-null  object
dtypes: int64(1), object(17)
memory usage: 4.3+ MB
```

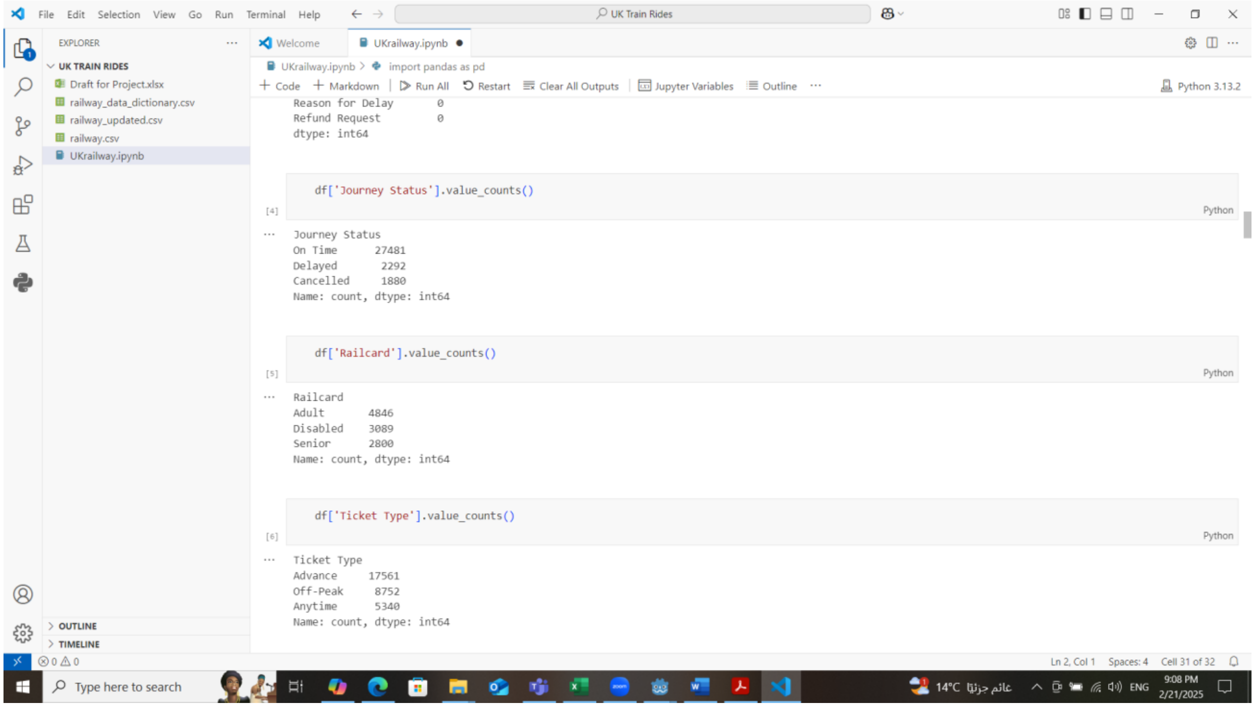


```
df.isnull().sum()
```

```
Transaction ID             0
Date of Purchase           0
Time of Purchase           0
Purchase Type              0
Payment Method             0
Railcard               20918
Ticket Class               0
Ticket Type                0
Price                      0
Departure Station          0
Arrival Destination        0
Date of Journey            0
Departure Time             0
Arrival Time               0
Actual Arrival Time     1880
Journey Status             0
Reason for Delay       27481
Refund Request             0
dtype: int64
```

```
df['Journey Status'].value_counts()
```

```
Journey Status
On Time      27481
Delayed       2292
Cancelled     1880
Name: count, dtype: int64
```



```
Reason for Delay       0
Refund Request         0
dtype: int64
```

```
df['Journey Status'].value_counts()
```

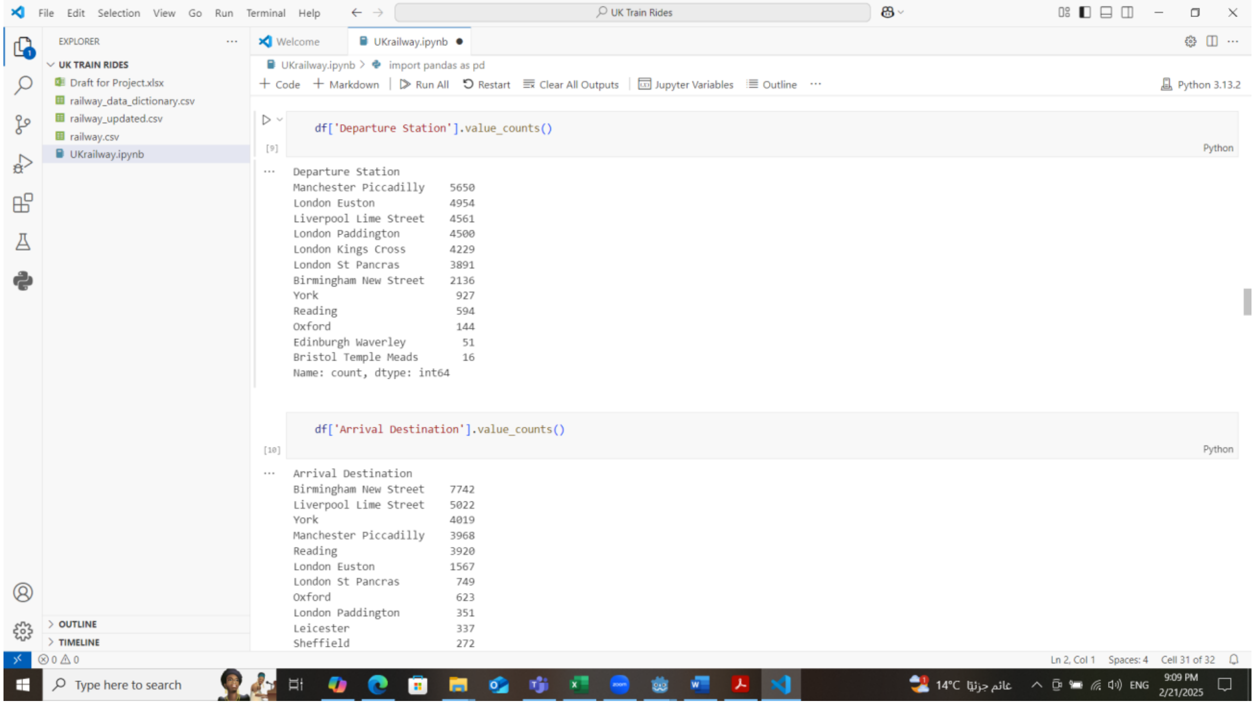
```
Journey Status
On Time      27481
Delayed       2292
Cancelled     1880
Name: count, dtype: int64
```

```
df['Railcard'].value_counts()
```

```
Railcard
Adult       4846
Disabled    3089
Senior      2800
Name: count, dtype: int64
```

```
df['Ticket Type'].value_counts()
```

```
Ticket Type
Advance     17561
Off-Peak     8752
Anytime      5340
Name: count, dtype: int64
```

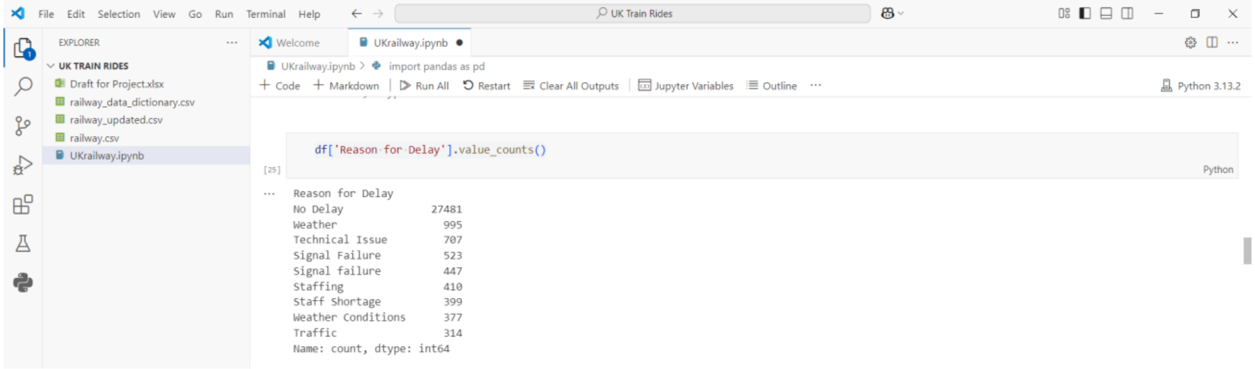

File Edit Selection View Go Run Terminal Help
UK Train Rides
EXPLORER
UK TRAIN RIDES
Draft for Project.xlsx
railway_data_dictionary.csv
railway_updated.csv
railway.csv
UKrailway.ipynb
Welcome
UKrailway.ipynb > import pandas as pd
+ Code + Markdown | Run All Restart Clear All Outputs Jupyter Variables Outline
Python 3.13.2
OUTLINE
TIMELINE
Ln 2, Col 1 Spaces: 4 Cell 31 of 32
Type here to search
14°C
ENG
9:09 PM
2/21/2025

```
[9]
df['Departure Station'].value_counts()
```

```
Departure Station
Manchester Piccadilly    5650
London Euston            4954
Liverpool Lime Street    4561
London Paddington        4500
London Kings Cross       4229
London St Pancras        3891
Birmingham New Street    2136
York                      927
Reading                   594
Oxford                    144
Edinburgh Waverley         51
Bristol Temple Meads       16
Name: count, dtype: int64
```

```
[10]
df['Arrival Destination'].value_counts()
```

```
Arrival Destination
Birmingham New Street    7742
Liverpool Lime Street    5022
York                     4019
Manchester Piccadilly    3968
Reading                  3920
London Euston            1567
London St Pancras         749
Oxford                    623
London Paddington         351
Leicester                 337
Sheffield                 272
```


File Edit Selection View Go Run Terminal Help
UK Train Rides
EXPLORER
UK TRAIN RIDES
Draft for Project.xlsx
railway_data_dictionary.csv
railway_updated.csv
railway.csv
UKrailway.ipynb
Welcome
UKrailway.ipynb > import pandas as pd
+ Code + Markdown | Run All Restart Clear All Outputs Jupyter Variables Outline
Python 3.13.2

```
[25]
df['Reason for Delay'].value_counts()
```

```
Reason for Delay
No Delay              27481
Weather                 995
Technical Issue         707
Signal Failure          523
Signal failure          447
Staffing                410
Staff Shortage          399
Weather Conditions      377
Traffic                 314
Name: count, dtype: int64
```

2nd Phase: Data Cleaning and Preparation

Assigned to (Sarah Farag – Sahar Hamdi – Nada Saad)

After conducting Data exploration using python to see the data in a collective perspective, we've found out that certain fields are containing the most missing values. The decision was to use Python as the first tool to clean the data because Python's efficient data structures and optimized libraries enable us to process this large dataset quickly and effectively. This makes data cleaning more manageable, even when dealing with massive amounts of data.

Also, the date/time format needed to be adjusted for the dataset's usability, enabling more accurate and meaningful data analysis and visualization. We decided on using Power Query as it provides an intuitive, user-friendly interface that allows us to manage and format date/time data without requiring advanced coding skills. Power Query seamlessly integrates with Excel, allowing us to leverage Excel's powerful data analysis and visualization tools. Using Power Query ensures that date/time formatting is consistent across the dataset. This is crucial for accurate data analysis, as inconsistent formats can lead to errors and misinterpretations.

Data Cleaning and Handle Missing Data: Identify and handle missing values, duplicates, and erroneous or outlier data points.

Replacing "Null" values in the "Railcard" column with "Unknown" provides a couple of key advantages:

1. **Clarity**: When data analysts and other stakeholders review the dataset, they can clearly see that the value is unknown rather than being unsure if it's a mistake or a genuine missing value.
2. **Consistency**: It ensures that all entries in the "Railcard" column have a value, making data handling and processing more straightforward. This also helps maintain data integrity and reduces the risk of errors during analysis.
3. **Querying**: When querying the dataset, it becomes easier to filter and analyze data. Instead of checking for "Null" values, which can be interpreted differently by various systems or software, users can simply filter for "Unknown".
4. **Documentation**: It provides a clear indication that the data point was either not captured or not applicable, which can be helpful for documenting the dataset and understanding its limitations.

The decision to set the "Actual Arrival Time" column entries to a unique time of "23:59:59" was made to address the issue of missing values in a clear and consistent manner. This approach offers several key benefits:

1. **Distinct Placeholder**: The time "23:59:59" was chosen as it is highly unlikely to be a valid arrival time, thereby clearly indicating that the actual arrival time is not available or was not recorded.

2. **Data Integrity**: By using a consistent placeholder, we ensure that the dataset remains structurally sound and free of null values, which can lead to complications in data processing and analysis.

3. **Ease of Identification**: The unique time "23:59:59" allows for easy identification and filtering of missing or placeholder entries during data analysis. This makes it straightforward to isolate these entries for further investigation or handling.

4. **Standardization**: This method ensures that all missing actual arrival times are uniformly represented, facilitating more accurate and reliable data analysis and reporting.

The decision to replace the "Reason for Delay" column entries with "No Delay" was made to address the issue of missing values and ensure data consistency. This approach offers several key benefits:

1. **Clarity**: By replacing missing or null entries with "No Delay," we provide a clear indication that there was no recorded reason for a delay, making it easier for analysts to understand the dataset.

2. **Data Integrity**: Ensuring that every entry in the "Reason for Delay" column has a value maintains the structural integrity of the dataset. This helps prevent issues that can arise from null values during data processing and analysis.

3. **Consistency**: Using a consistent placeholder like "No Delay" helps standardize the data, making it easier to analyze and interpret. This uniformity aids in generating accurate and reliable reports.

4. **Ease of Analysis**: The "No Delay" entry allows for straightforward filtering and querying of the dataset, enabling analysts to quickly identify and separate entries with no recorded delays from those with specific reasons for delays.

By adopting this approach, we maintain the integrity and usability of the dataset, enabling more efficient and effective data analysis.

Transform Data: Ensure that our data is in the right format (e.g., date formats, numerical values, categorical encoding).

This decision was driven by several key factors:

1. **Accuracy**: Converting these columns to the "Date/Time" format ensures that all date and time values are accurately recognized and processed. This helps to eliminate errors that can occur when dates and times are stored as text or other formats.

2. **Consistency**: Standardizing the data types across these columns provides consistency in how dates and times are represented. This uniformity is essential for accurate comparisons, calculations, and aggregations.

3. **Enhanced Analysis**: With the data in the correct "Date/Time" format, it becomes easier to perform time-based analyses, such as calculating durations, identifying trends, and generating time series visualizations.

4. **Improved Visualization**: Many data visualization tools and software rely on proper "Date/Time" formats to create accurate and meaningful visual representations. By ensuring that the data is in the correct format, we enable more effective and insightful visualizations.

5. **Data Integrity**: Adjusting the data types helps maintain the integrity of the dataset by ensuring that dates and times are stored in a standardized and reliable manner. This reduces the risk of data corruption and enhances the overall quality of the dataset.

Through these adjustments, we enhance the dataset's usability, enabling more accurate and meaningful data analysis and visualization.

In order to effectively manage and analyze the dataset, we employed a combination of Python and Power Query with Excel to address several key data quality issues and standardization needs.
Using Python:

- To handle missing values, we substituted "Null" entries in specific columns with predefined values:
    - The "Railcard" column entries were replaced with "Unknown".
    - The "Actual Arrival Time" column entries were set to a unique time "23:59:59".
    - The "Reason for Delay" column entries were replaced with "No Delay".
- To standardize the values in the "Reason for Delay" column, we unified various entries to ensure consistency:
    - Entries like "Weather" and "Weather Conditions" were merged under "Weather Conditions".
    - Entries like "Signal Failure" and "Signal failure" were consolidated under "Signal Failure".
    - Entries like "Staffing" and "Staff Shortage" were unified under "Staff Shortage".

Using Power Query and Excel:

- To ensure proper data analysis and visualization, we adjusted the data types of the following columns to "Date/Time" format:
    - "Date of Purchase"
    - "Time of Purchase"
    - "Date of Journey"
    - "Departure Time"
    - "Arrival Time"
    - "Actual Arrival Time"

Through these steps, we transformed the dataset into a more structured and standardized form, facilitating more accurate and meaningful insights.

3rd Phase: Data Analysis and Insights Generation

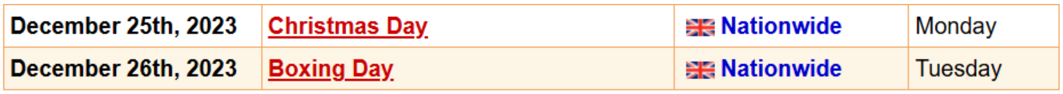
Assigned to (Mona Hassan – Sara Mohamed – Dina Yasser)

At this phase we aim to fully understand the dataset to better extract the valuable information that will aid in generating the best insights.

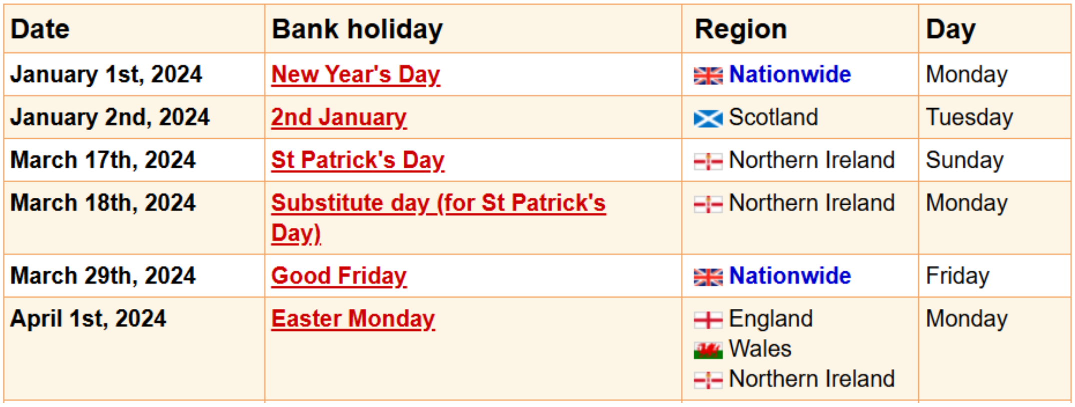
Dates of the bookings are between **December 8th 2023 – April 30th 2024** which highlights the holidays seasons in winter and spring.

Dates of the journeys are between **January 1st 2024 – April 30th 2024.**

- Christmas Day – Monday, December 25 Nationwide
- Boxing Day – Tuesday, December 26 Nationwide
- New Year's Day – Monday, January 1. Nationwide
- Good Friday – Friday, March 29 Nationwide
- Easter Monday – Monday, April 1.



December 25th, 2023	Christmas Day	Nationwide	Monday
December 26th, 2023	Boxing Day	Nationwide	Tuesday



Date	Bank holiday	Region	Day
January 1st, 2024	New Year's Day	Nationwide	Monday
January 2nd, 2024	2nd January	Scotland	Tuesday
March 17th, 2024	St Patrick's Day	Northern Ireland	Sunday
March 18th, 2024	Substitute day (for St Patrick's Day)	Northern Ireland	Monday
March 29th, 2024	Good Friday	Nationwide	Friday
April 1st, 2024	Easter Monday	England Wales Northern Ireland	Monday

Source: Bank Holidays 2024 in the UK, with printable templates

Here are the railway stations affected by Storm Henk:

1. **Winchester and Micheldever**: Landslide blocked all railway lines between these stations on January 4th.
2. **Yeovil Junction and Exeter**: Landslide near Crewkerne in south Somerset blocked all railway lines between these stations on January 5th.
3. **Maidstone East**: Landslide blocked several lines by this station in Kent.
4. **Robertsbridge**: Landslide hit the line at this station in East Sussex.
5. **Arlesey**: Landslide on the railway line between Stevenage and Peterborough.
6. **London Paddington and south Wales**: Flooding caused train services to be diverted between these stations.
7. **Coventry and Birmingham International**: Flooding hit the railway between these stations.
8. **London Paddington and Maidenhead**: Electricity failure caused route blockages between these stations.
9. **Watford Junction and London Euston**: Damage to overhead wires caused earlier closures between these stations.
10. **Richmond and Willesden Junction**: Urgent repairs on the track at Gunnersbury affected services between these stations.
11. **Swindon and Bristol Parkway**: Flooding blocked several parts of the network between these stations.

These disruptions were caused by landslides, flooding, and power failures due to the severe weather conditions brought by Storm Henk.

Source: Storm Henk triggers landslide chaos across UK rail network | Ground Engineering

Source: Storm Henk batters UK leading to power outages, travel disruption and flooding - BBC News

The key points from Payment Options:

1. **Pay As You Go with Contactless or Oyster**:
   - No need to buy tickets in advance.
   - Touch in and out with a contactless card/device or Oyster card to be charged the correct fare.
   - Contactless cards/devices are accepted on National Rail, London Underground, DLR, London Buses, and London Trams within London and at some stations across the South East.
   - Daily and weekly (Monday to Sunday) capping ensures you won't be charged more than a certain amount.
2. **Travelcards**:
   - Purchased ahead of your journey for unlimited travel within specific zones.
   - Available for 1 day, 7 days, 1 month, or any period between 1 month and 1 year.
   - Travelcards can be used on National Rail services, London Underground, DLR, London Buses, and London Trams in specified zones.

Contactless Payment Specifics:

- Charged the cheapest adult fare for your journey on the day and time of travel.
- Railcard and other discounts are not available with contactless payment.
- Capping ensures cost-efficiency for frequent travel.

Oyster Card Specifics:

- Allows storage of up to £90 of pay as you go credit.
- Usable on a wide range of transport services within London Zones 1 to 9 and some services beyond.
- Capping limits daily and weekly travel costs.
- Railcard discounts can be applied for up to 1/3 off travel.
- Child rate fares available; under 5s travel free with a fare-paying adult.
- Can store up to 3 Travelcard or London Bus & Tram Pass Season tickets.

Travelcards:

- Provide unlimited travel within specified zones.
- Available for different durations: 1 day, 7 days, 1 month, etc.
- Paper One Day Travelcards; longer duration Travelcards added to Oyster or Smartcard.
- Boundary Zone tickets available for travel beyond permitted zones.
- Return tickets from home counties, South, South East, or further afield may include a One Day Travelcard.

The ticket details, discounts, and refund policies for different ticket types:

Advance Tickets

- **Details**:
  - Great value for long journeys, must be bought in advance.
  - Valid only on the date and train specified.
  - Single journeys only, can combine tickets for return trips.
  - Sold in limited numbers, subject to availability.
  - Available up to 12 weeks ahead of travel, sometimes until the day of travel.
- **Discounts**:
  - 50% off for children aged 5 to 15.
  - 50% off for 16-17 Saver Railcard holders on adult Standard class Advance fares.
  - 1/3 off Standard class Advance fares with any National Railcard.
  - 1/3 off First Class Advance fares with 16-25, 26-30, Senior, Two Together, HM Forces, Veterans, or Disabled Persons Railcards.
- **Refund Policy**:
  - Non-refundable unless the train is delayed or cancelled and you choose not to travel.
  - Can amend the ticket up to the time of travel, with a £10 fee and any difference in fare.

Off-Peak and Super Off-Peak Tickets

- **Details**:
  - Available for travel at less busy times on weekdays, all day on weekends.
  - Cheaper but may require travel at specified times, days, or routes.
  - Valid for 1 month from the date shown on the ticket.
  - Can buy at any time before travel.
- **Discounts**:
  - 50% off for children aged 5 to 15.
  - 50% off for 16-17 Saver Railcard holders on adult Off-Peak and Super Off-Peak fares.
  - 1/3 off Standard Class Off-Peak and Super Off-Peak fares with any National Railcard.
- **Refund Policy**:
  - Refundable with no admin fee if the service is delayed or cancelled and you choose not to travel.
  - Can get a refund and rebook with no admin fee if traveling on another day not covered by the original ticket.
  - Refundable with a maximum £5 admin fee if canceling for other reasons.

Anytime Tickets

- **Details**:
  - Fully flexible with no time restrictions.
  - Can travel on any train on the route shown.
  - Anytime Singles valid for 2 days, Anytime Returns for 5 days (outward) and 1 month (return).
  - Can buy at any time before travel.
- **Discounts**:
  - 50% off for children aged 5 to 15.
  - 50% off for 16-17 Saver Railcard holders on adult Anytime fares.
  - 1/3 off Standard Class Anytime fares with any National Railcard.
- **Refund Policy**:
  - Refundable with no admin fee if the service is delayed or cancelled and you choose not to travel.
  - Can get a refund and rebook with no admin fee if traveling on a different day.

Find out more about our Railcards.

See the Advance tickets terms and conditions for full information.

Find out more about pay as you go with contactless(external link, opens in a new tab) and check out the pay as you go with contactless map.

Purchase & Ticketing Insights

These questions have been designed to extract valuable insights from the dataset and enhance our understanding of the purchase and ticketing patterns. Here's why each question is useful:

1. **What are the most common purchase methods (Online vs. Station)?**
   - Understanding the preferred purchase method helps identify customer preferences and can guide decisions on where to invest more resources or marketing efforts.
2. **What percentage of tickets are purchased with railcards, and which type is most used?**
   - This insight reveals the popularity of railcards, which can help in tailoring promotional offers and understanding the needs of different customer segments.
3. **How does the ticket price vary by ticket type (Advance, Standard, etc.)?**
   - Analyzing price variations by ticket type can assist in pricing strategies and identifying which ticket types are most attractive to customers.
4. **Which stations have the highest ticket sales?**
   - Identifying the busiest stations can help allocate resources, improve services, and plan for infrastructure developments.
5. **What are the peak times for ticket purchases?**
   - Knowing peak purchase times allows for better staffing and resource allocation, as well as targeted marketing campaigns during these periods.
6. **What are the busiest routes based on ticket purchases?**
   - Understanding the most popular routes can inform service improvements, route planning, and capacity management.
7. **Which ticket class (Standard vs. others) is the most popular?**
   - Analyzing the popularity of ticket classes can guide service enhancements and pricing adjustments to meet customer demand.
8. **How does ticket price vary across different departure stations?**
   - Examining price variations by departure station can reveal pricing trends and potential areas for adjustment to optimize revenue.
9. **Are refunds more common for specific routes or ticket types?**

- Identifying patterns in refund requests can help address service issues, improve customer satisfaction, and refine refund policies.

10. **How do ticket prices impact refund requests?**

    - Understanding the relationship between ticket prices and refund requests can inform pricing strategies and identify areas where customers may feel overcharged.

11. **Do delays increase the likelihood of a refund request?**

    - Analyzing the impact of delays on refund requests can help improve service reliability and address customer concerns proactively.

By exploring these questions, we can uncover critical insights that will drive better decision-making, enhance customer satisfaction, and optimize overall operations.