ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Module Code: CT051-3-M-DM

Module Name: Data Management

Assignment – Part 1

Initial Data Exploration and Evaluation of Data Warehouse Concepts for UK National Rail

Student Name: Muhammad Yousouf Ali Budullah

TP Number: TP086704

Intake Code: APDMF2501DSBA(BI)(PR)

Programme: MSc Data Science & Business Analytics

Module Lecturer: Dr. Murugananthan Velayutham

Date of Submission: May 5, 2025

Table of Contents

1 INTRODUCTION 3

2 INITIAL DATA EXPLORATION 3

2.1 METADATA AND ATTRIBUTE IDENTIFICATION 3
2.1.1 Nominal Attributes *3*
2.1.2 Ordinal Attributes *4*
2.1.3 Interval Attributes *4*
2.1.4 Ratio Attributes *5*
2.2 DESCRIPTIVE STATISTICS AND VISUAL EXPLORATIONS 6
2.2.1 Nominal and Ordinal Attributes *6*
2.2.1.1 Purchase Type 6
2.2.1.2 Payment Method 7
2.2.1.3 Railcard 8
2.2.1.4 Ticket Class 9
2.2.1.5 Ticket Type 10
2.2.1.6 Journey Status 11
2.2.1.7 Reason for Delay 12
2.2.1.8 Refund Reequest 13
2.2.1.9 Departure Station 14
2.2.1.10 Arrival Destination 14
2.2.2 Interval Attributes *15*
2.2.2.1 Date of Purchase 15
2.2.2.2 Time of Purchase 16
2.2.2.3 Date of Journey 17
2.2.2.4 Departure Time 18
2.2.2.5 Arrival Time 19
2.2.2.6 Actual Arrival Time 20
2.2.3 Ratio Attributes *21*
2.2.3.1 Price 21
2.3 DATA QUALITY ASSESSMENT 22
2.3.1 Missing Values *22*
2.3.2 Inconsistencies *23*
2.3.2.1 Ticket Type 23
2.3.2.2 Journey Status 23
2.3.2.3 Reason for Delay 23
2.3.2.4 Time Format and Date Format Variants 23
2.3.2.5 Station Name Variants 23
2.3.3 Outliers *24*

3 EVALUATION OF DATA WAREHOUSE CONCEPTS 25

3.1 DATA INTEGRATION AND CLEANING CAPABILITY 25
3.2 TIME VARIANT DATA TRACKING 26
3.3 MULTIDIMENSIONAL ANALYSIS SUPPORT 27
3.4 SCALABILITY AND QUERY PERFORMANCE 28
3.5 ANALYTICAL USEFULNESS AND DECISION SUPPORT 29

5 CONCLUSION 30

BIBLIOGRAPHY 31

APPENDIX A – SAS CODE 32

APPENDIX A.1 – SAS CODE FOR PURCHASE_TYPE 32
APPENDIX A.2 – SAS CODE FOR PAYMENT_METHOD 32
APPENDIX A.3 – SAS CODE FOR RAILCARD 32

Appendix A.4 – SAS Code for Ticket_Class 32
Appendix A.5 – SAS Code for Ticket_Type 33
Appendix A.6 – SAS Code for Journey_Status 33
Appendix A.7 – SAS Code for Reason_for_Delay 33
Appendix A.8 – SAS Code for Refund_Request 33
Appendix A.9 – SAS Code for Departure_Station 33
Appendix A.10 – SAS Code for Arrival_Destination 34
Appendix A.11 – SAS Code for Date_of_Purchase 34
Appendix A.12 – SAS Code for Time_of_Purchase 34
Appendix A.13 – SAS Code for Date_of_Journey 35
Appendix A.14 – SAS Code for Departure_Time 35
Appendix A.15 – SAS Code for Arrival_Time 35
Appendix A.16 – SAS Code for Actual_Arrival_Time 36
Appendix A.17 – SAS Code for Price 36
Appendix A.18 – SAS Code for Missing Values 37
Appendix A.19 – SAS Code for Price Outliers 37

List of Tables and Figures

Table 1- Frequency of Ticket Purchase Types 6
Table 2 - Frequency of Payment Methods 7
Table 3 – Frequency of Railcard Usage 8
Table 4 – Distribution of Ticket Classes 9
Table 5 - Frequency of Ticket Types 10
Table 6 - Frequency of Journey Statuses 11
Table 7 - Frequency of Delay Reasons 12
Table 8 - Frequency of Refund Requests 13
Table 9 - Summary Statistics of Ticket Prices 21
Table 10 - Missing Value Summary for Interval and Time-Based Attributes 22
Table 11- Boxplot of Ticket Prices 24

Figure 1- Bar Chart of Purchase Types 6
Figure 2 – Bar Chart of Payment Methods 7
Figure 3 - Bar Chart of Railecard Usage 8
Figure 4 - Bar Chart of Ticket Types 9
Figure 5 – Bar Chart of Ticket Classes 10
Figure 6 – Bar Chart of Journey Statuses 11
Figure 7 - Bar Chart of Delay Reasons 12
Figure 8 - Bar Chart of Refund Requests 13
Figure 9 - Bar Chart of Departure Stations 14
Figure 10 - Bar Chart of Arrival Destinations 14
Figure 11 - Line Chart of Purchases Over Time 15
Figure 12 - Histogram of Hourly Purchases 16
Figure 13 - Line chart of Journeys Over Time 17
Figure 14 - Histogram of Departure Hours 18
Figure 15 - Histogram of Arrival Hours 19
Figure 16 - Histogram of Actual Arrival Hours 20
Figure 17 - Histogram of Ticket Prices 21

1 Introduction

Understanding and preparing the dataset is the most important step that lays the foundation for future analytics. Part one of this report is centred around the initial exploration of a UK National Rail dataset containing ticket transaction and journey details between January and April 2024. It describes the metadata and attributes present in the data, descriptive statistics and data quality issues such as missing values and inconsistencies. Furthermore, it proposes and evaluates key data warehouse concepts, aligning with the UK National Rail dataset.

2 Initial Data Exploration

Initial data exploration is the first step in data preparation, it allows us to gain a comprehensive understanding of the dataset's features and structure. This section identifies the metadata and attribute types, followed by an analysis of each attribute using descriptive statistics and visualisations.

2.1 Metadata and Attribute Identification

2.1.1 Nominal Attributes

Attribute Name	Description	Attribute Type	SAS Data Type	Format	Example
Transaction ID	Unique identifier for an individual train ticket purchase	Nominal	Character	$23	e098ea8c-682f-4b97-ad28
Purchase Type	Whether the ticket was purchased online or directly at a train station	Nominal	Character	$7	Online
Payment Method	Payment method used to purchase the ticket.	Nominal	Character	$11	Contactless
Railcard	Whether the passenger is a National Railcard holder or not.	Nominal	Character	$8	Adult

Ticket Type	When you bought or can use the ticket.	Nominal	Character	$8	Off-Peak
Departure Station	Station to board the train	Nominal	Character	$21	London Paddington
Arrival Destination	Station to exit the train	Nominal	Character	$21	Reading
Journey Status	Whether the train was on time, delayed, or cancelled	Nominal	Character	$9	Delayed
Reason for Delay	Reason for the delay or cancellation	Nominal	Character	$15	Technical Issue
Refund Request	Whether the passenger requested a refund after a delay or cancellation	Nominal	Character	$2	No

2.1.2 Ordinal Attributes

Attribute Name	Description	Attribute Type	SAS Data Type	Format	Example
Ticket Class	Seat class for the ticket.	Ordinal	Character	$11	Standard

2.1.3 Interval Attributes

Attribute Name	Description	Attribute Type	SAS Data Type	Format	Example
Date of Purchase	Date the ticket was purchased	Interval	Numeric	YYMMDD10.	4/3/24
Time of Purchase	Time the ticket was purchased	Interval	Numeric	TIME20.3	7:30:38
Date of Journey	Date the train departed	Interval	Numeric	YYMMDD10.	4/3/24
Departure Time	Time the train departed	Interval	Numeric	TIME20.3	17:30:00
Arrival Time	Time the train was scheduled to arrive at its destination	Interval	Numeric	TIME20.3	19:45:00

Actual Arrival Time	Time the train arrived at its destination	Interval	Numeric	TIME20.3	20:05:00

2.1.4 Ratio Attributes

Attribute Name	Description	Attribute Type	SAS Data Type	Format	Example
Price	Final cost of the ticket	Ratio	Numeric	BEST12.	18

2.2 Descriptive Statistics and Visual Explorations

Note: Any missing values, inconsistencies, or outliers related to the attributes seen in this section will be discussed in detail in 2.3 Data Quality Assessment

2.2.1 Nominal and Ordinal Attributes

2.2.1.1 Purchase Type

Purchase_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Online	1805	60.17	1805	60.17
Station	1195	39.83	3000	100.00

Table 1- Frequency of Ticket Purchase Types

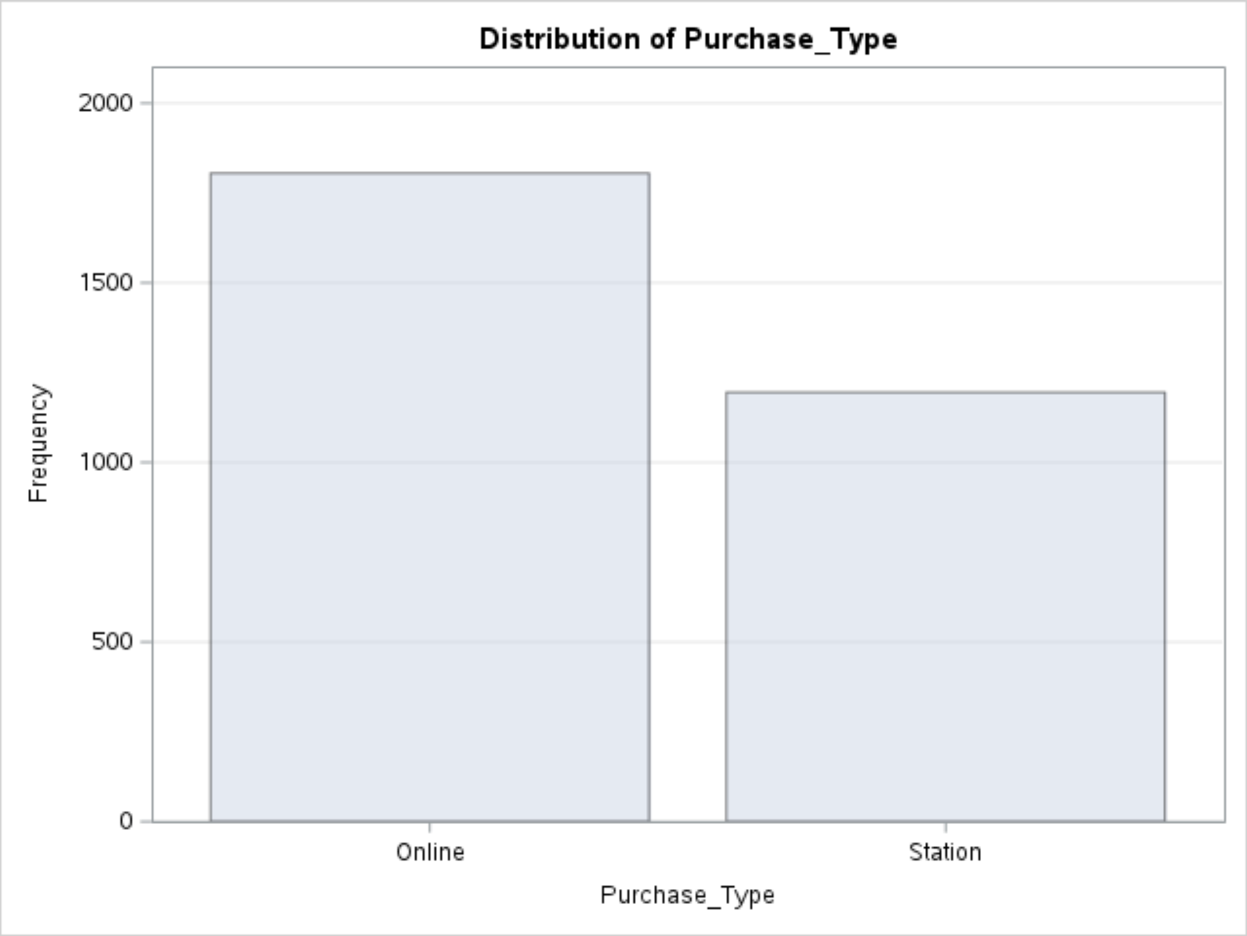
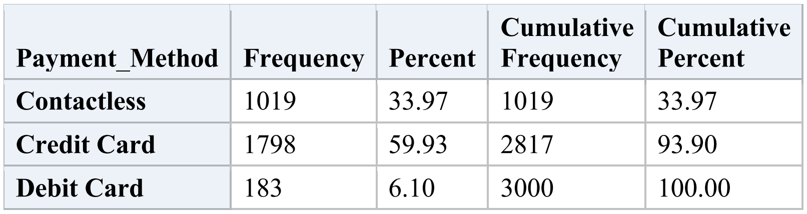


Figure 1- Bar Chart of Purchase Types

Table 1 displays the frequency and percentage of ticket purchase types. More tickets (60.17%) were purchased online, while 39.83% were bought at the station. Figure 1 further emphasizes this observations, clearly showing online purchase outnumber station purchases.

Refer to Appendix A.1 – SAS Code for Purchase_Type for SAS code.

2.2.1.2 Payment Method



Payment_Method	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Contactless	1019	33.97	1019	33.97
Credit Card	1798	59.93	2817	93.90
Debit Card	183	6.10	3000	100.00

Table 2 - Frequency of Payment Methods

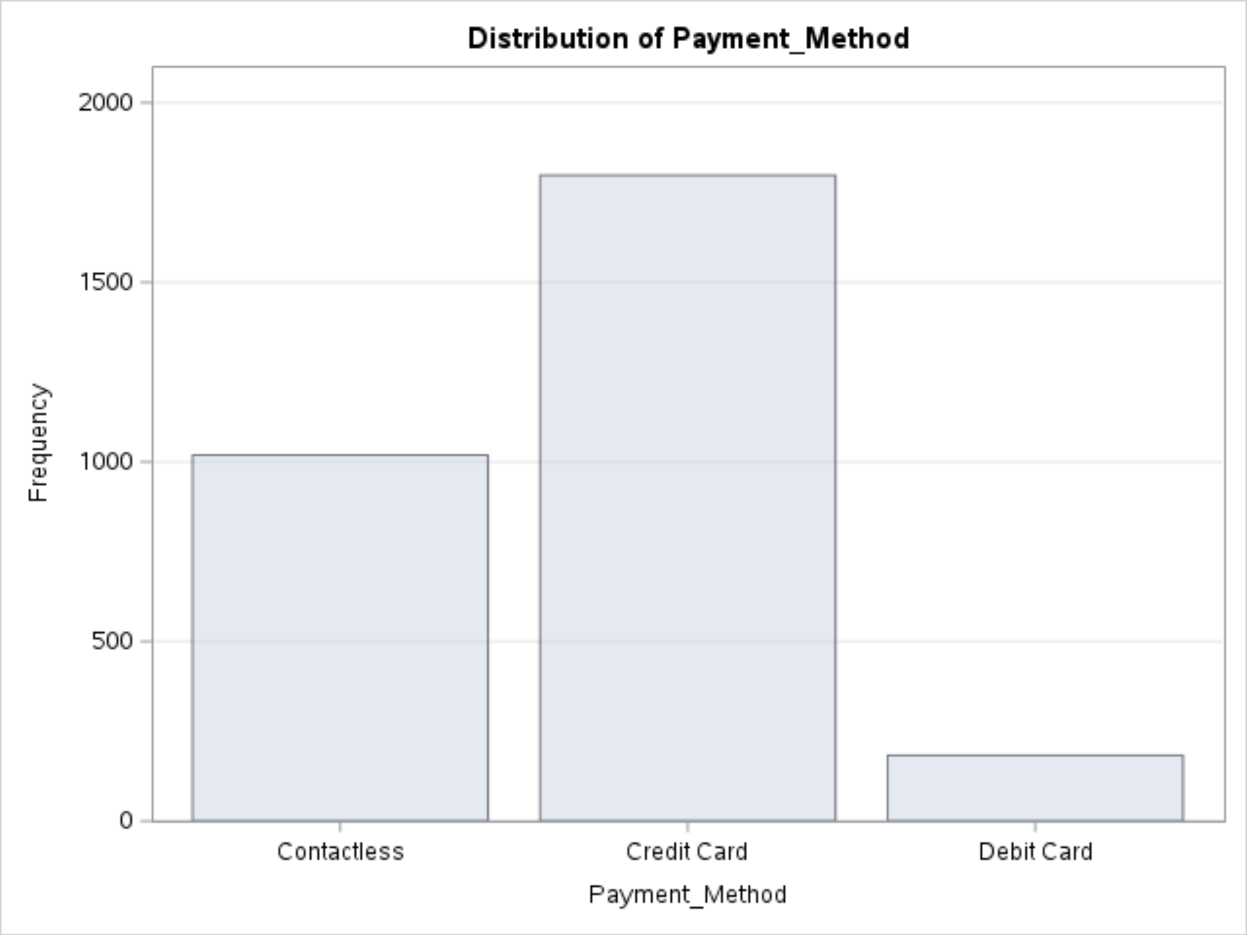
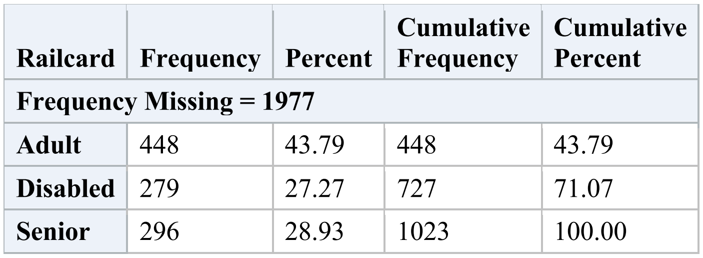


Figure 2 – Bar Chart of Payment Methods

Table 2 displays the frequency and percentage distribution of payment methods used for ticket purchases. The most common method of payment was credit card, accounting for 59.93% of all transactions. Followed by contactless payments at 33.97% and debit card at 6.10%. Figure 2 visualises this distribution showing a clear preference for credit card payments.

Refer to Appendix A.2 – SAS Code for Payment_Method for SAS code.

2.2.1.3 Railcard



Railcard	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Frequency Missing = 1977				
Adult	448	43.79	448	43.79
Disabled	279	27.27	727	71.07
Senior	296	28.93	1023	100.00

Table 3 – Frequency of Railcard Usage

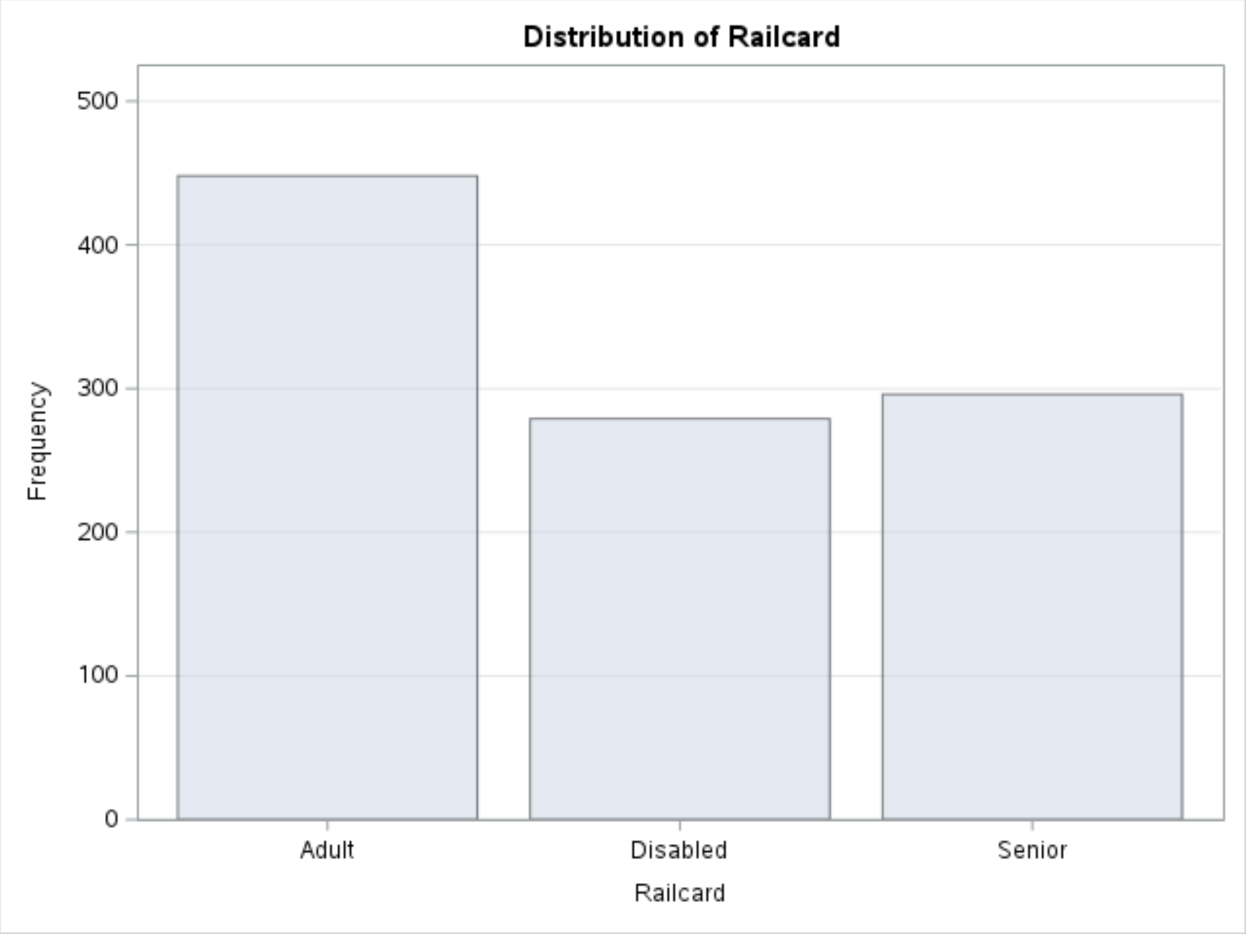


Figure 3 - Bar Chart of Railecard Usage

Table 3 displays the distribution of rail card types among ticket holders. The most frequent used railcard was the Adult Railcard. Followed by the Senior Railcard and then Didabled with 1977 passengers opting not to use any railcard. Figure 3 illustrates these differences, clearly showing Adult Railcards are the most used.

Refer to Appendix A.3 – SAS Code for Railcard for SAS code.

2.2.1.4 Ticket Class

Ticket_Class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
First Class	284	9.47	284	9.47
Standard	2716	90.53	3000	100.00

Table 4 – Distribution of Ticket Classes

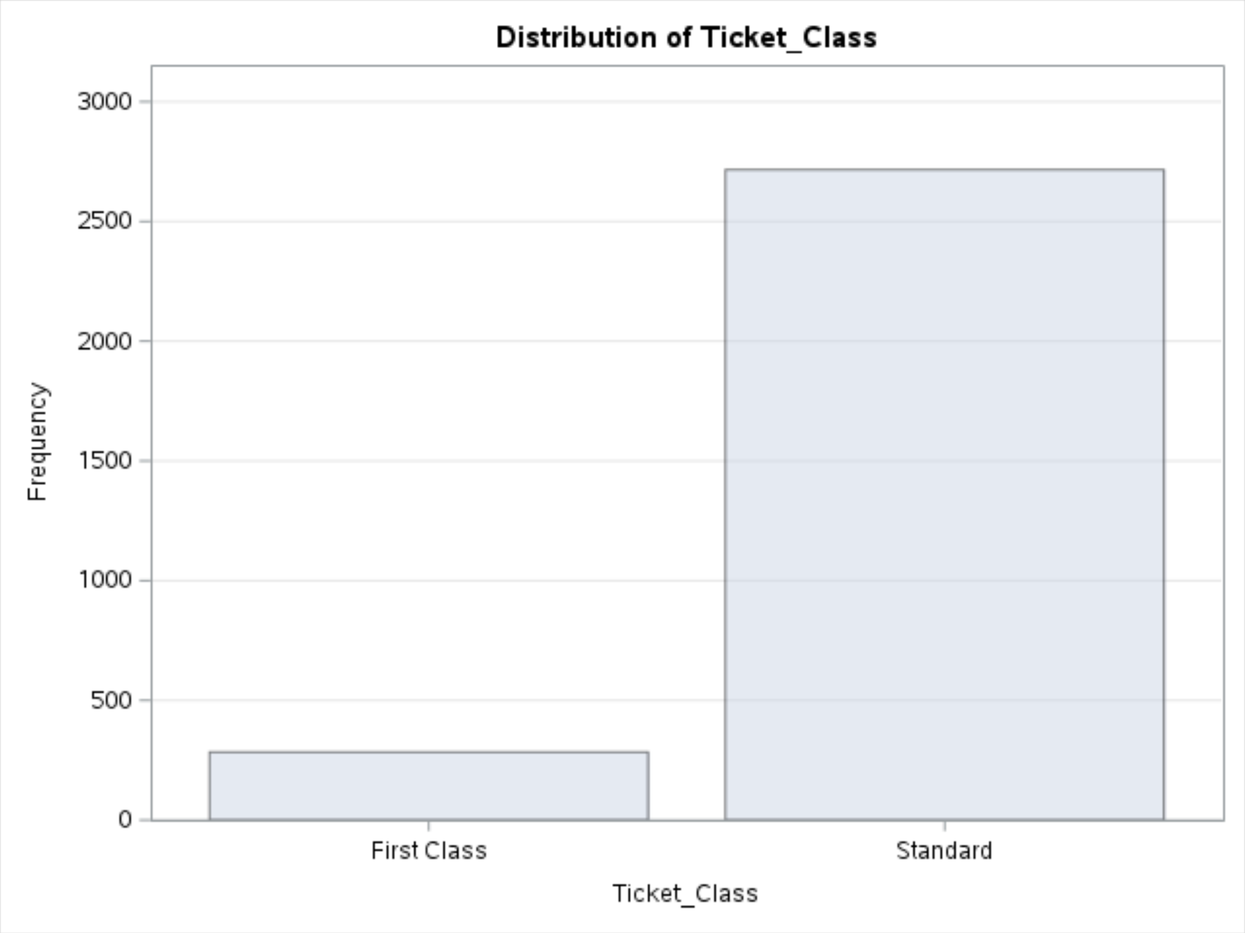
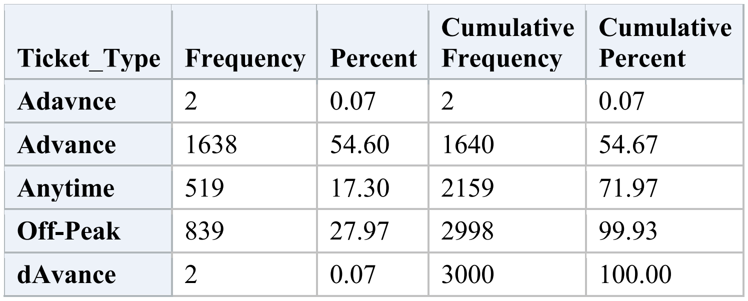


Figure 4 - Bar Chart of Ticket Types

Table 4 displays the distribution of ticket classes purchased by passengers. The vast majority, 90.53%, chose to travel in Standard Class while only 9.47% chose to travel in First Class. Figure 4 visualises this disparity highlighting the preference for most passengers to be standard Class.

Refer to Appendix A.4 – SAS Code for Ticket_Class for SAS code.

2.2.1.5 Ticket Type



Ticket_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Adavnce	2	0.07	2	0.07
Advance	1638	54.60	1640	54.67
Anytime	519	17.30	2159	71.97
Off-Peak	839	27.97	2998	99.93
dAvance	2	0.07	3000	100.00

Table 5 - Frequency of Ticket Types

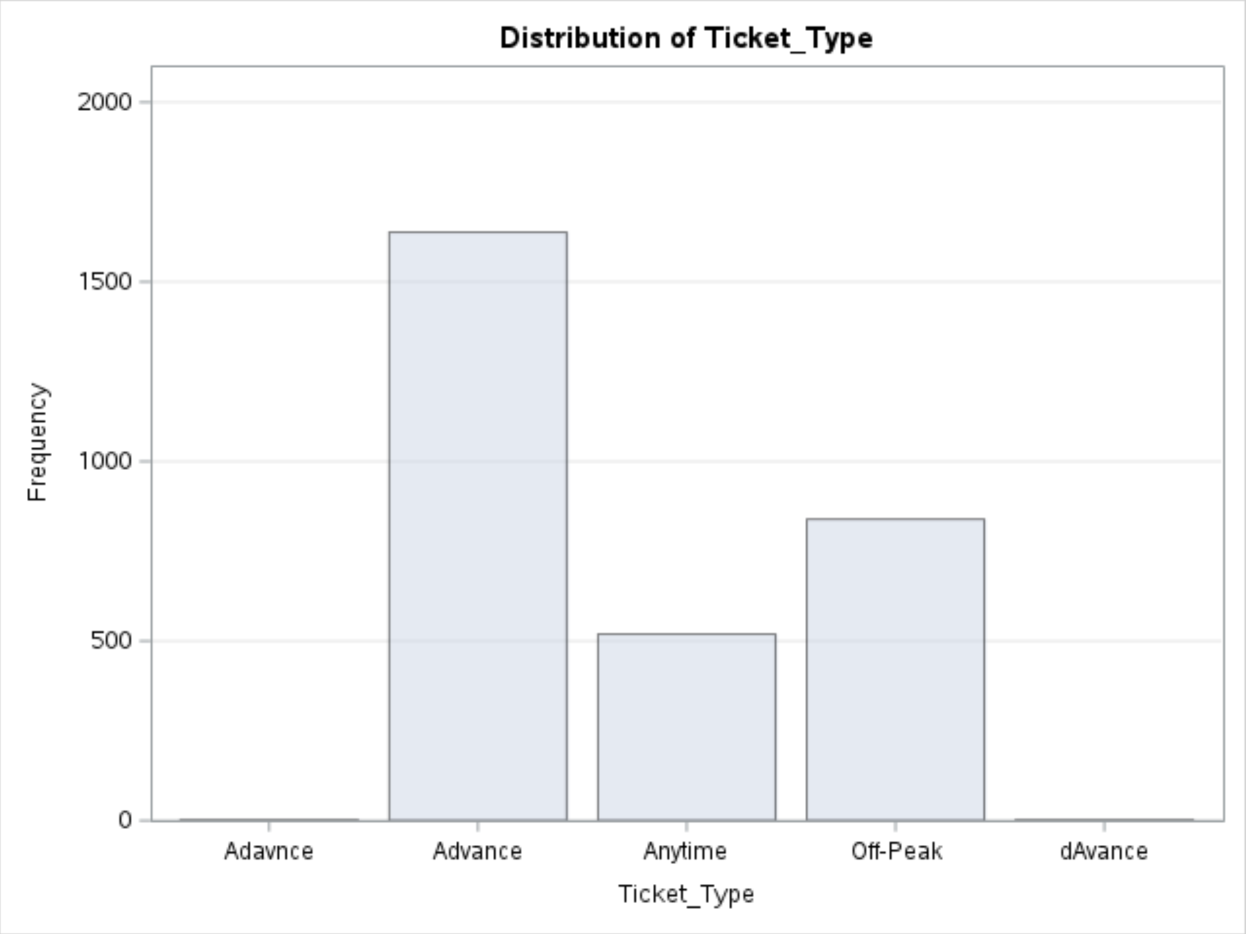
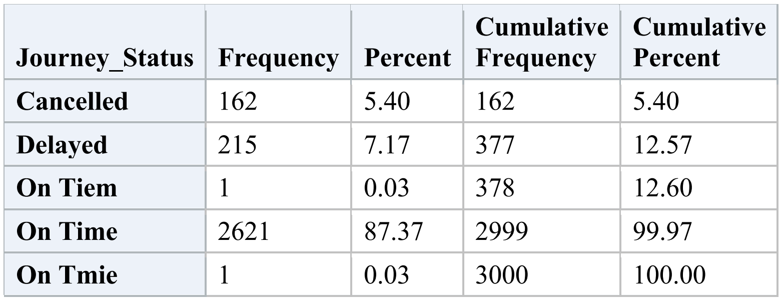


Figure 5 – Bar Chart of Ticket Classes

Table 5 shows the frequency distribution of different ticket types. The most purchased type was Advance, followed by Off-Peak and Anytime. Figure 5 reinforces this observation.

Refer to Appendix A.5 – SAS Code for Ticket_Type for SAS code.

2.2.1.6 Journey Status



Journey_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cancelled	162	5.40	162	5.40
Delayed	215	7.17	377	12.57
On Tiem	1	0.03	378	12.60
On Time	2621	87.37	2999	99.97
On Tmie	1	0.03	3000	100.00

Table 6 - Frequency of Journey Statuses

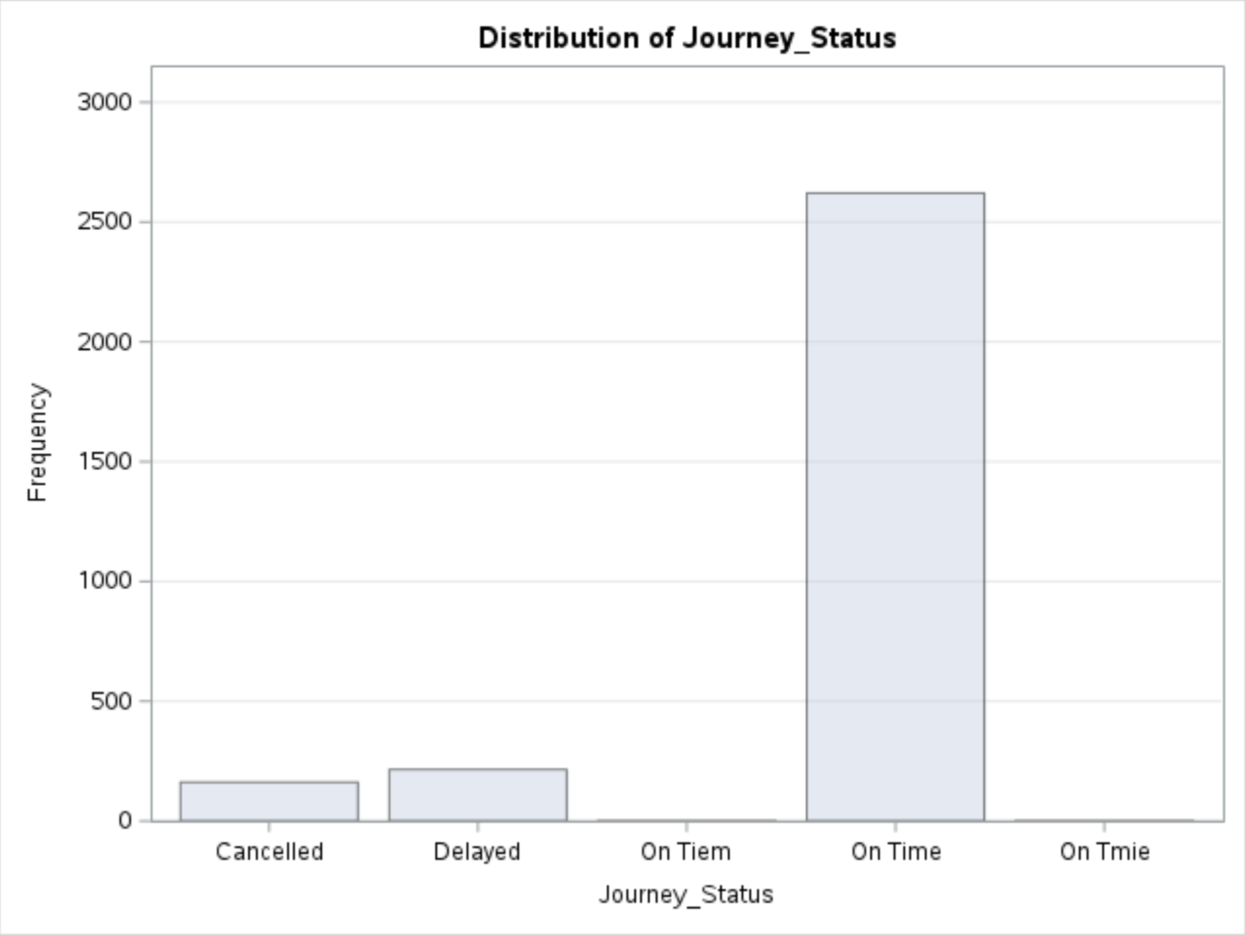


Figure 6 – Bar Chart of Journey Statuses

Table 6 shows the frequency distribution of journey statuses. A vast majority of trips arrived on time with only a few being cancelled of delayed. Figure 6 reinforces this observation.

Refer to Appendix A.6 – SAS Code for Journey_Status for SAS code.

2.2.1.7 Reason for Delay

Reason_for_Delay	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Frequency Missing = 2623				
Signal Failure	45	11.94	45	11.94
Signal failure	43	11.41	88	23.34
Staff Shortage	29	7.69	117	31.03
Staffing	41	10.88	158	41.91
Technical Issue	71	18.83	229	60.74
Traffic	33	8.75	262	69.50
Weather	83	22.02	345	91.51
Weather Conditi	32	8.49	377	100.00

Table 7 - Frequency of Delay Reasons

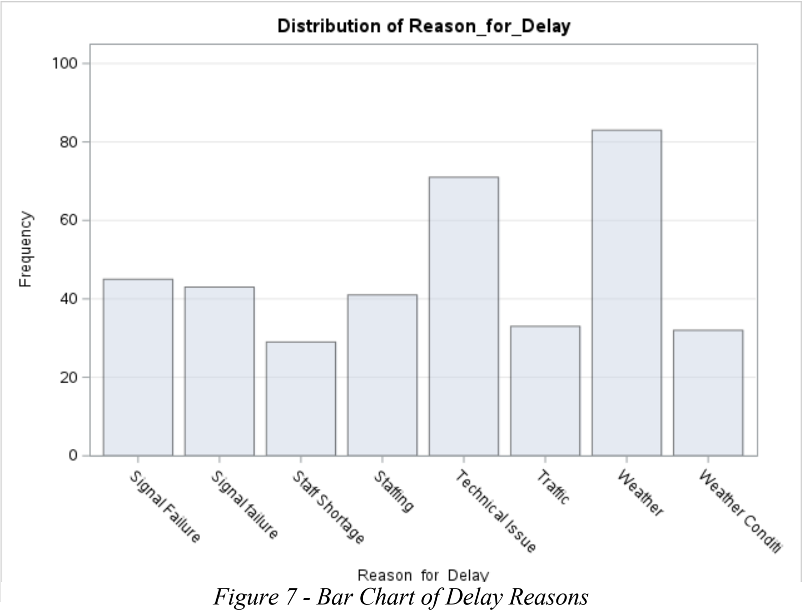
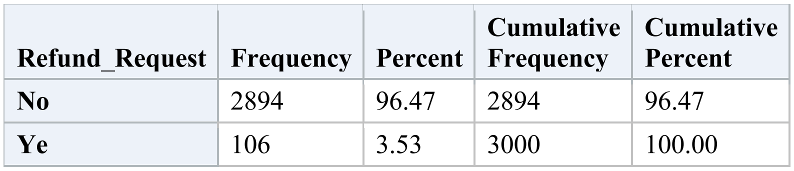


Figure 7 - Bar Chart of Delay Reasons

Table 7 and Figure 7 together display the range of reasons recorded for train delays and cancellations. The most common reasons that can be seen are weather and technical issues and signal failures.

Refer to Appendix A.7 – SAS Code for Reason_for_Delay for SAS code.

2.2.1.8 Refund Reequest



Refund_Request	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	2894	96.47	2894	96.47
Ye	106	3.53	3000	100.00

Table 8 - Frequency of Refund Requests

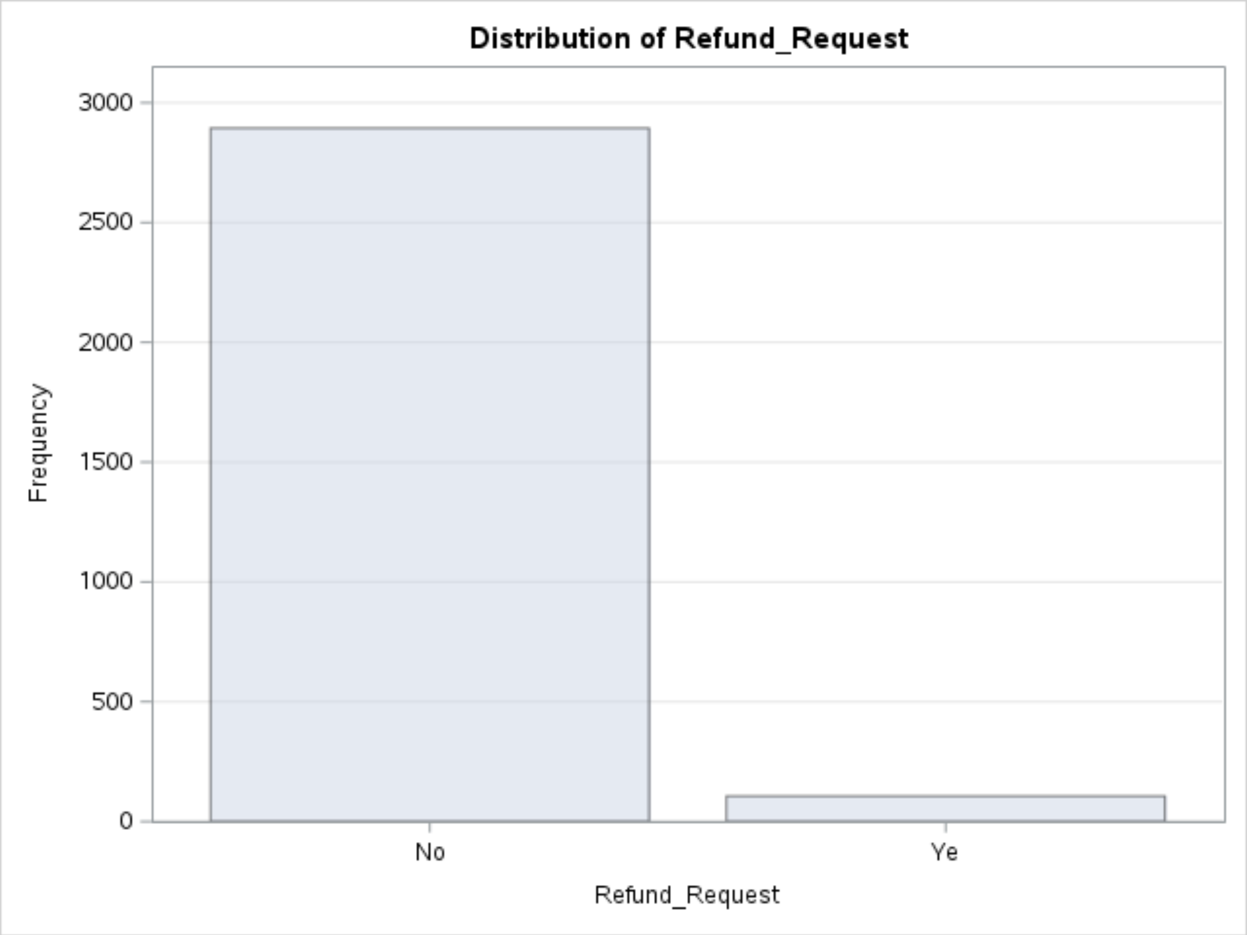


Figure 8 - Bar Chart of Refund Requests

Table 8 and Figure 8 show that an overwhelming amount of passengers, 96.47% , did not request a refund even though their trip was either cancelled or delayed.

Refer to Appendix A.8 – SAS Code for Refund_Request for SAS code.

2.2.1.9 Departure Station

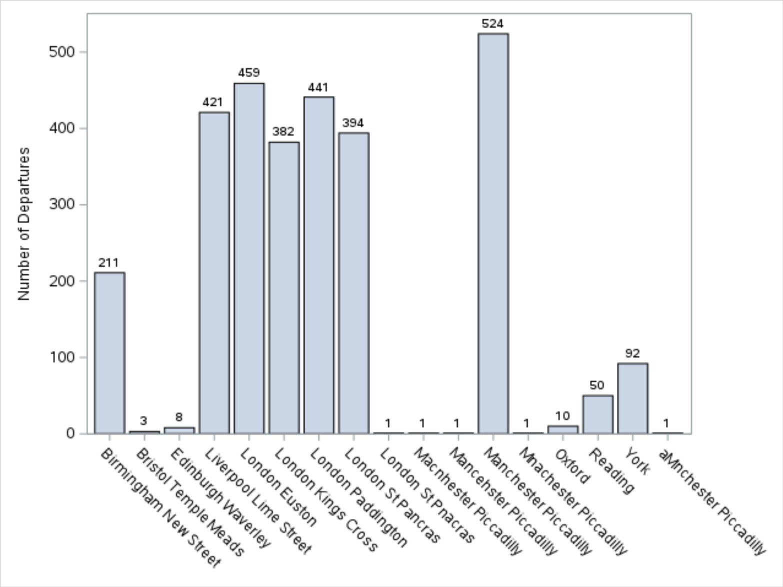


Figure 9 - Bar Chart of Departure Stations

Figure 9 displays the number of train departures from each station. Manchester Picadilly recorded the highest number of departures, followed by London Euston, London Paddington and London's King Cross. *Refer to Appendix A.9 – SAS Code for Departure_Station for SAS code.*

2.2.1.10 Arrival Destination

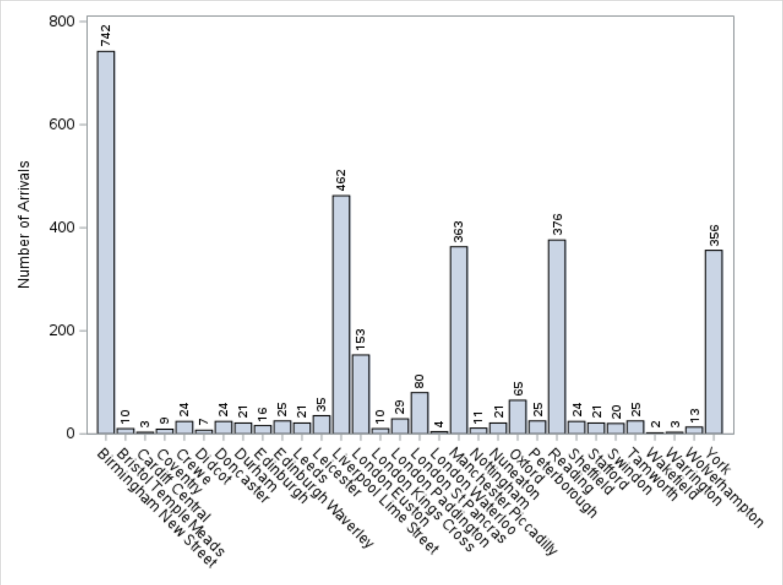


Figure 10 - Bar Chart of Arrival Destinations

Figure 10 shows the frequency of arrivals at various destinations. Birmingham New Street had the highest arrival count, with notable volumes from Liverpool Lime Street, Manchester Picadilly, Reading and York. *Refer to Appendix A.10 – SAS Code for Arrival_Destination for SAS code.*

2.2.2 Interval Attributes

2.2.2.1 Date of Purchase

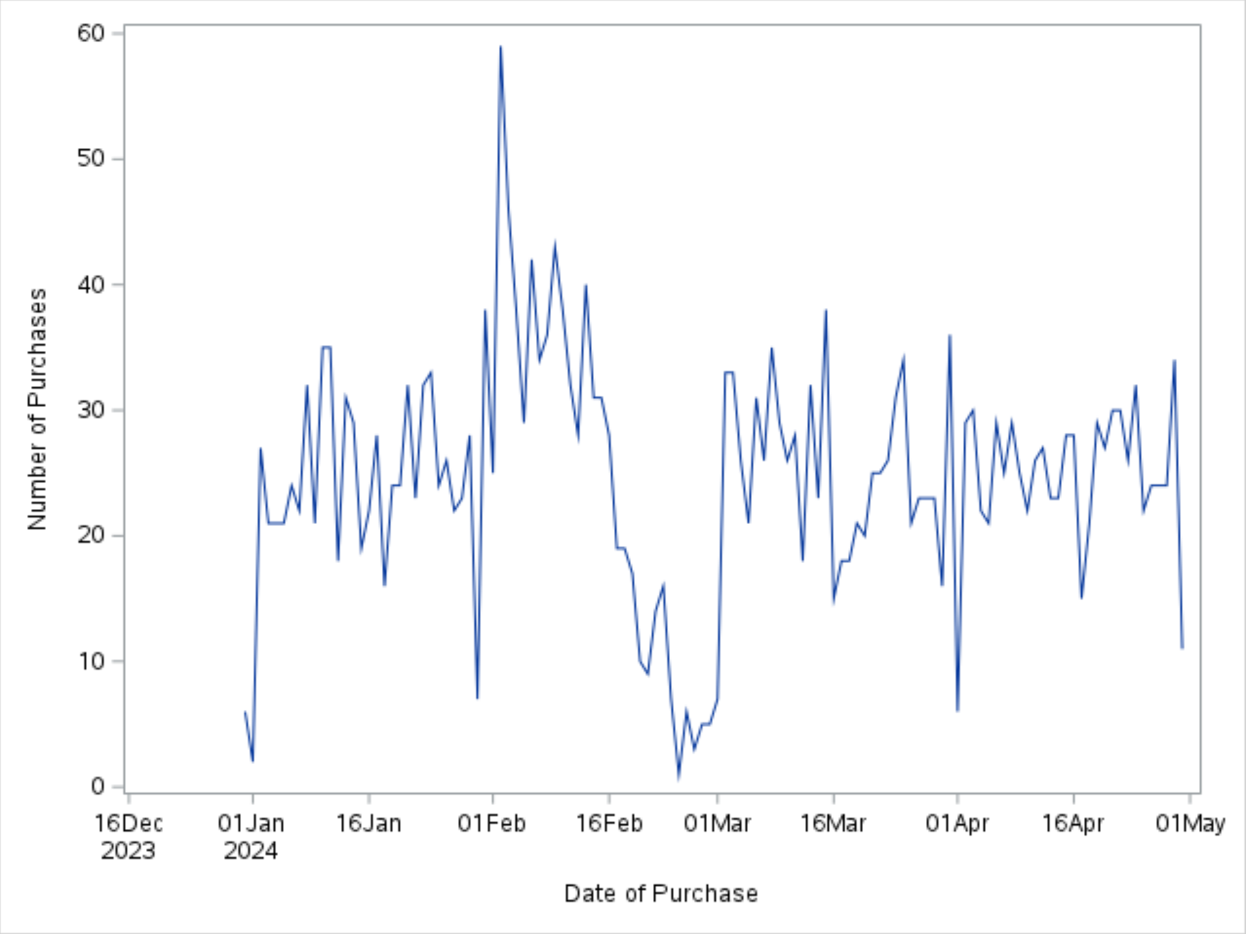


Figure 11 - Line Chart of Purchases Over Time

Figure 11 displays the number of ticket purchases made per day over the period from late December 2023 to April 2024. The trend shows noticeable variations with several spikes in early February and March. It also reveal intermittent drops suggesting periods of lower activity.

Refer to Appendix A.11 – SAS Code for Date_of_Purchase for SAS code.

2.2.2.2 Time of Purchase

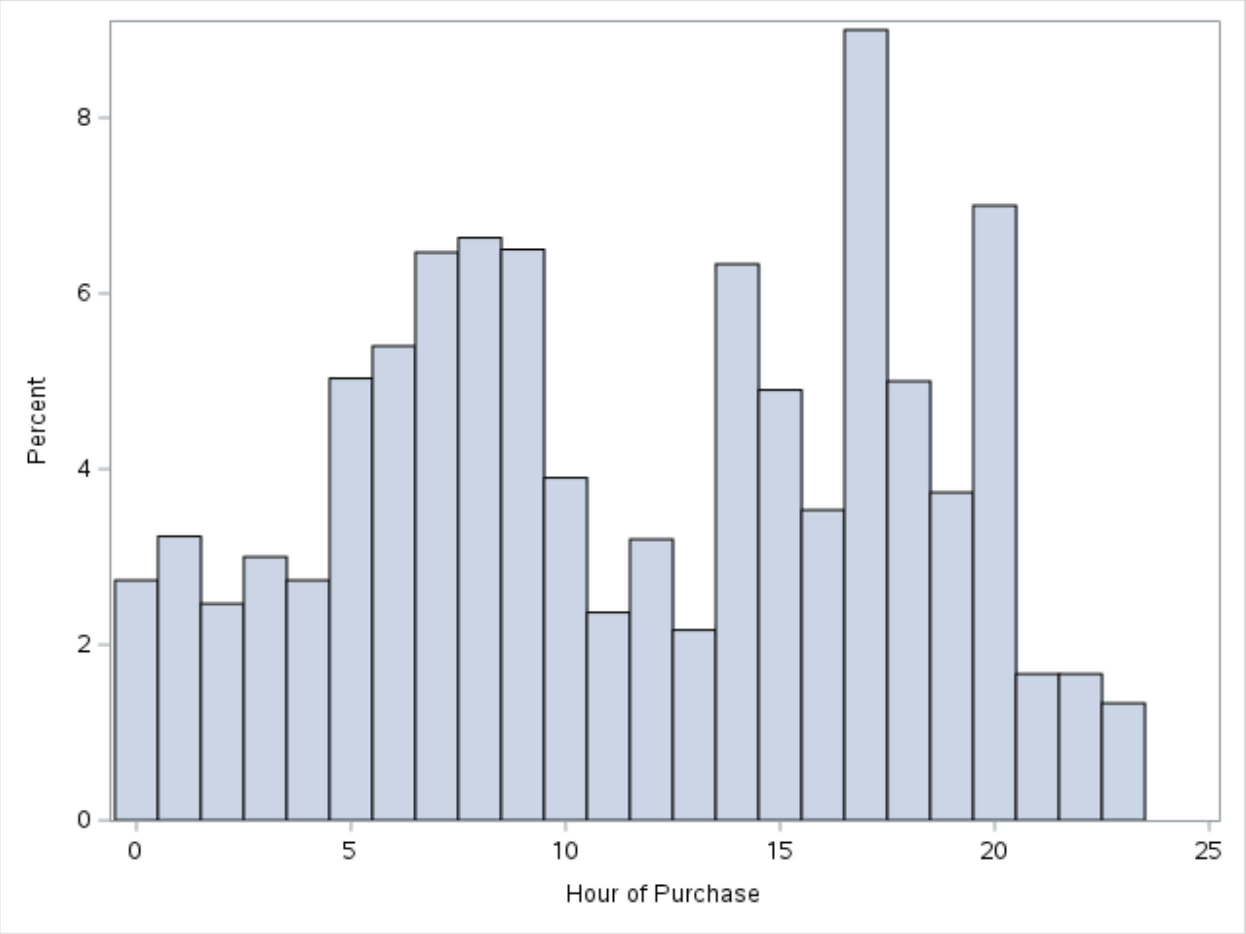


Figure 12 - Histogram of Hourly Purchases

Figure 12 presents the distribution of purchase by hour of the day. The histogram shows us that gradually increase from early morning and peak at 17h00 (5pm) with secondary rises in the morning and early evening.

Refer to Appendix A.12 – SAS Code for Time_of_Purchase for SAS code.

2.2.2.3 Date of Journey

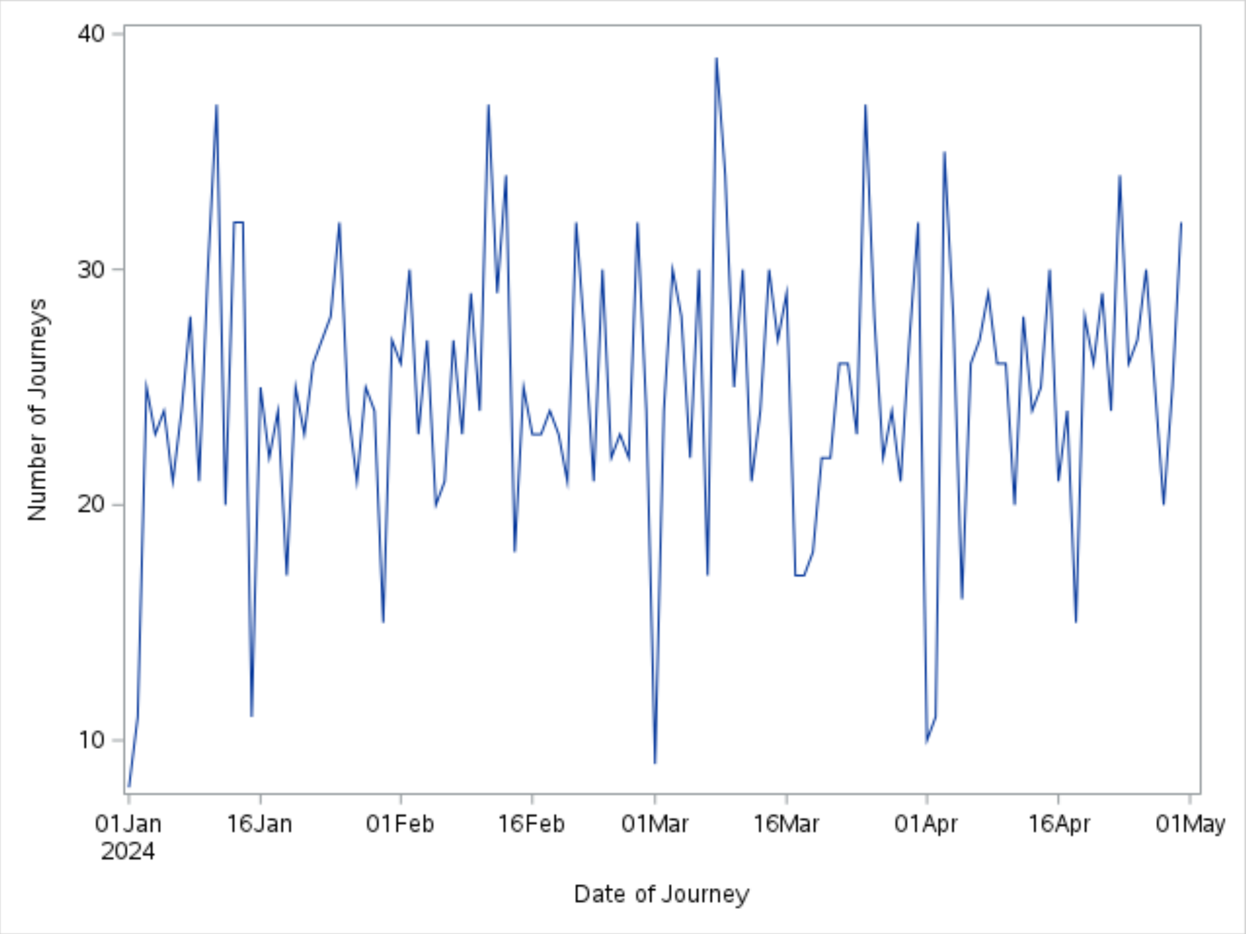


Figure 13 - Line chart of Journeys Over Time

Figure 13 displays the number of journeys taken per day from January 2024 to April 2024. It reveals a consistent level of travel activity, with frequent fluctuations in journey volume. Peaks and dips occur regularly.

Refer to Appendix A.13 – SAS Code for Date_of_Journey for SAS code.

2.2.2.4 Departure Time

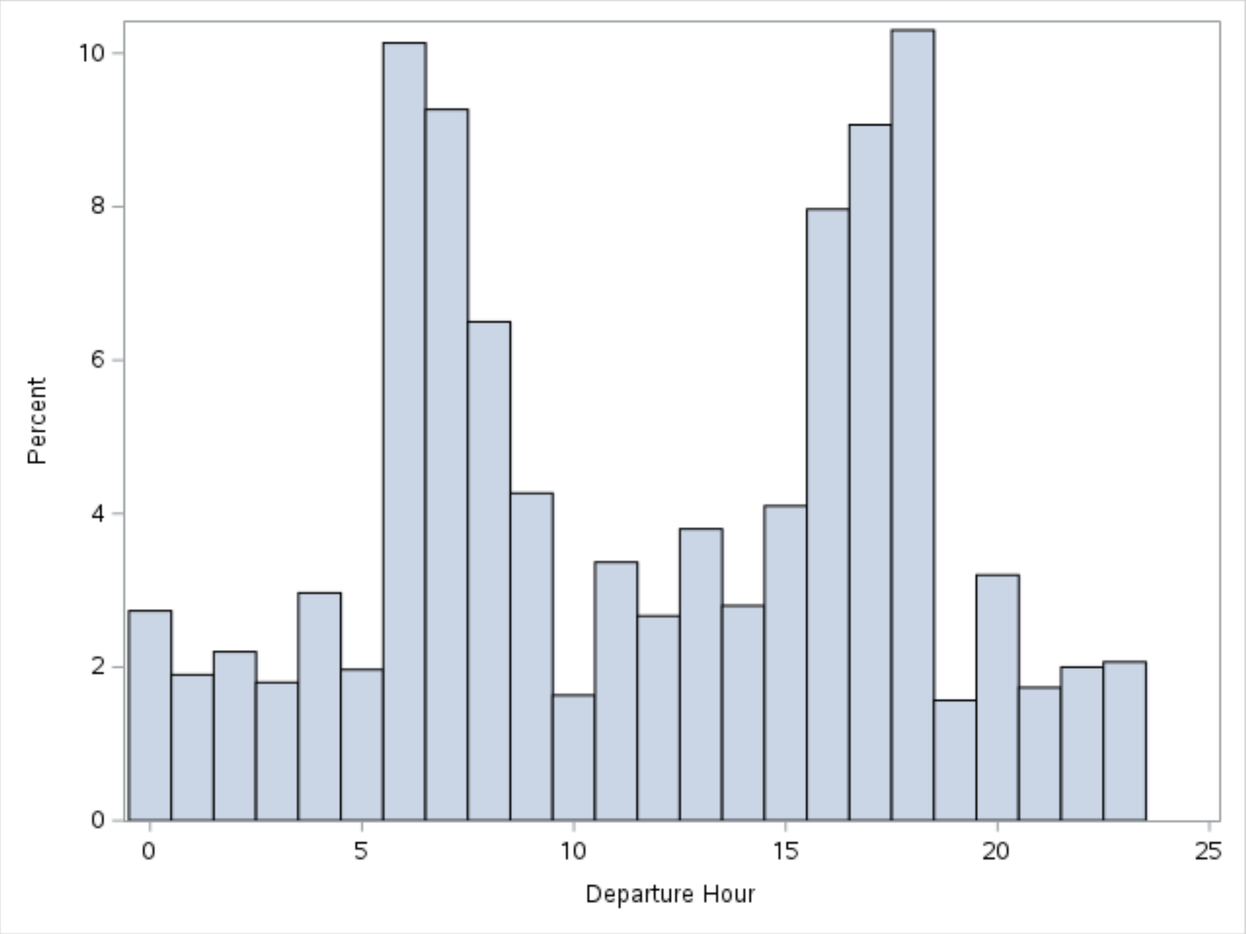


Figure 14 - Histogram of Departure Hours

Figure 14 shows the distribution of train departures by hour of the day. Two notable peaks occur at 6h00 (6 am) to 8h00 (8am) and at 16h00 (4pm) to 18h00 (6pm).

Refer to Appendix A.14 – SAS Code for Departure_Time for SAS code.

2.2.2.5 Arrival Time

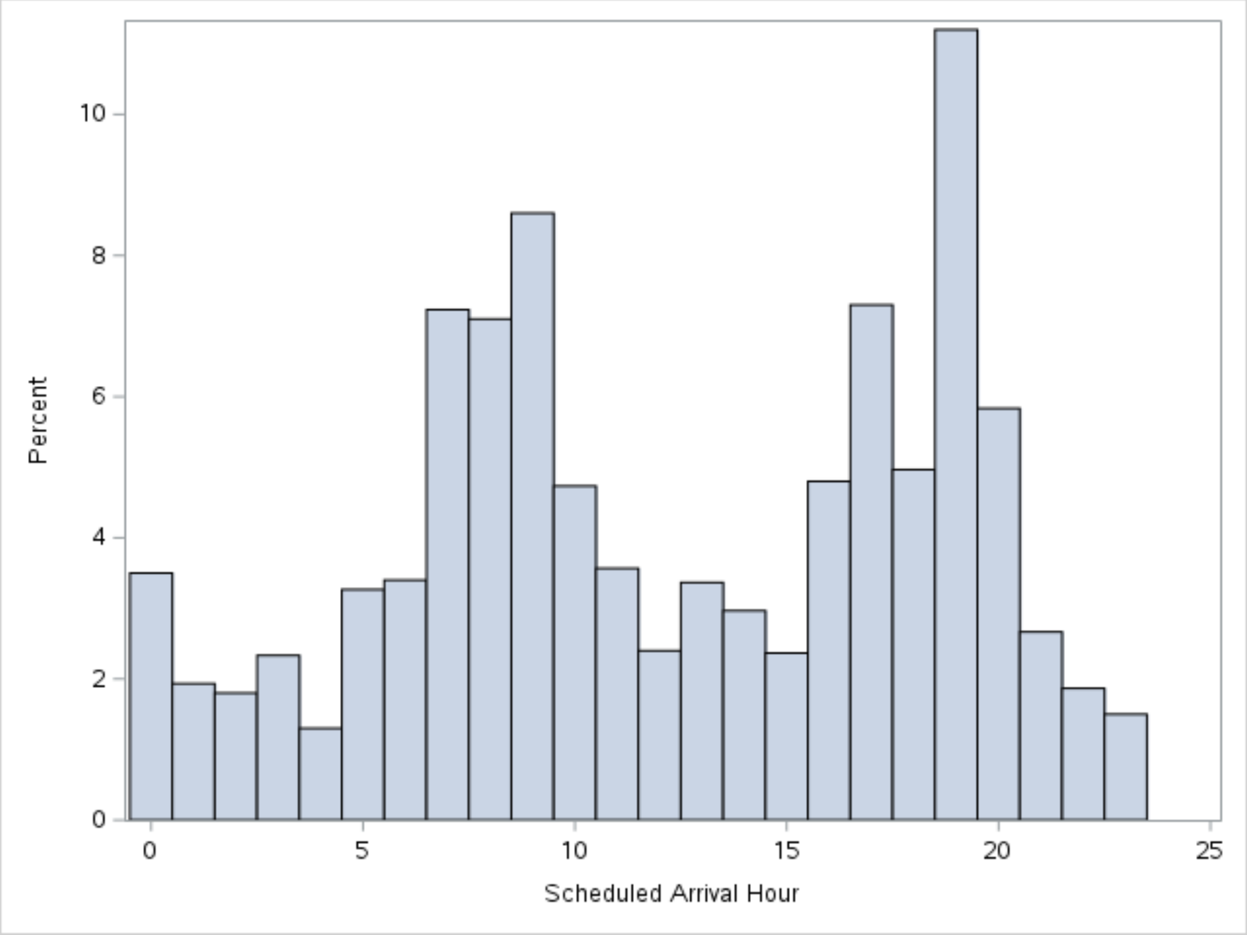


Figure 15 - Histogram of Arrival Hours

Figure 15 illustrate the distribution of scheduled train arrivals by the hour. It shows a wide spread of arrivals throughout the day with two noticeable peaks at 9h00 (9am) and 19h00 (7pm).

Refer to Appendix A.15 – SAS Code for Arrival_Time for SAS code.

2.2.2.6 Actual Arrival Time

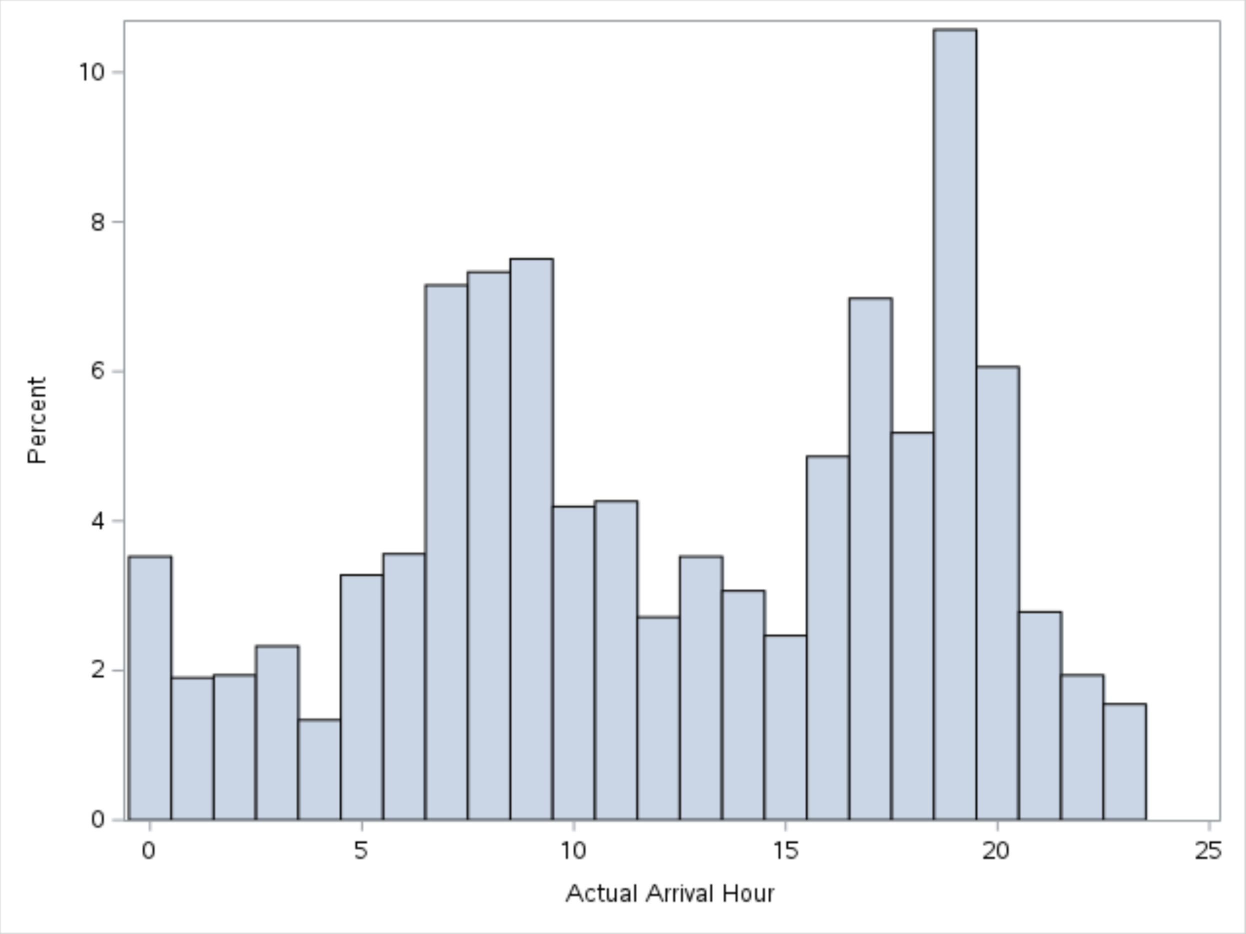


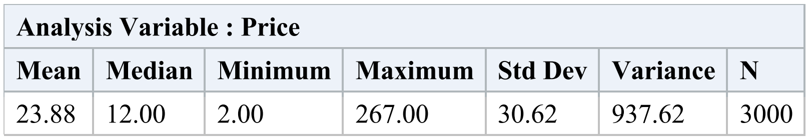
Figure 16 - Histogram of Actual Arrival Hours

Figure 16 shows the distribution of actual arrival times by hour of day. Similar to scheduled arrival times there are 2 clear peaks at 9h00 (9am) and 19h00 (7pm). The trend closely mirrors that of scheduled arrivals but slight shifts in frequency could reflect delays or early arrival across stations.

Refer to Appendix A.16 – SAS Code for Actual_Arrival_Time for SAS code.

2.2.3 Ratio Attributes

2.2.3.1 Price



Analysis Variable : Price						
Mean	Median	Minimum	Maximum	Std Dev	Variance	N
23.88	12.00	2.00	267.00	30.62	937.62	3000

Table 9 - Summary Statistics of Ticket Prices

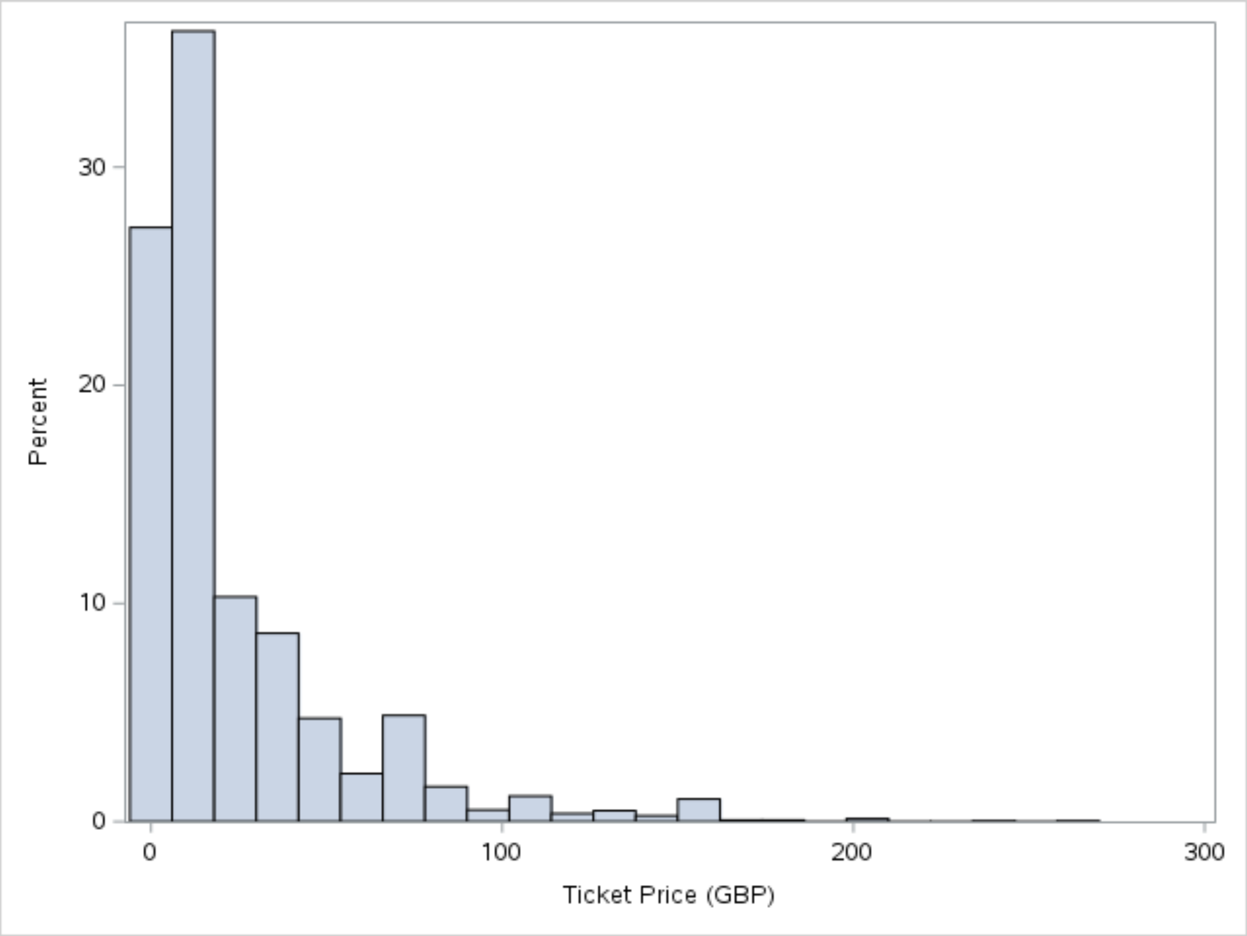


Figure 17 - Histogram of Ticket Prices

Table 9 gives the summary statistics for price. The mean price is 23.88 while the median is 12.00 indicating a positive skewness in the data. The minimum and maximum values range from 2.00 to 267.00 with a standard deviation of 30.62 sugeting significant price variability in tickets.

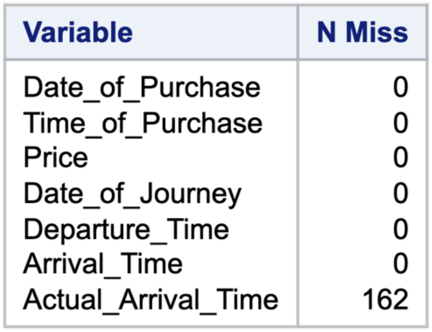
Figure 17 visually confirms the right skewed distribution, where most ticket prices are concentrated at the lower end, and a long tail extends towards the higher prices.

Refer to Appendix A.17 – SAS Code for Price for SAS code.

2.3 Data Quality Assessment

Preceding any analysis to be performed, the data’s quality must be evaluated to ensure accuracy and reliability. This section assesses common data quality issues such as missing values, inconsistencies and outliers. In addition, each issue is examined, and suggestions are made to ensure data integrity for future use.

2.3.1 Missing Values



Variable	N Miss
Date_of_Purchase	0
Time_of_Purchase	0
Price	0
Date_of_Journey	0
Departure_Time	0
Arrival_Time	0
Actual_Arrival_Time	162

Table 10 - Missing Value Summary for Interval and Time-Based Attributes

Refer to Appendix A.18 – SAS Code for Missing Values for SAS code.

Missing values were observed in several key attributes, particularly Railcar, Reasons_For_Delay and Actual_Arrival_Time. In Table 3 – Frequency of Railcard Usage we can see there are 1977 missing records. However, these omission do not represent data collection errors but most instead passenger who did not use a railcard. Thus, the best action would be to replace values with an appropriate label such as “No Railcard” .

In Table 7 - Frequency of Delay Reasons we can also see 2623 missing values which could initially be seen as incomplete data. However, it is most like missing because they correspond with on-time journey hence no need to state a reason for delay. Similar to Railcard it would be best to replace these missing values with a place holder such as “No Delay” to provide better clarity.

Lastly, in Table 10 - Missing Value Summary for Interval and Time-Based Attributes we can note missing values for Actual_Arrival_Time. These occur when a journey was cancelled and didn’t arrive thus not recorded or considered unnecessary.

2.3.2 Inconsistencies

2.3.2.1 Ticket Type

In Table 5 - Frequency of Ticket Types we can see multiple variations of what appear to be the same ticket type. This inconsistency fragment frequency counts as seen in the table and impact future trend analysis

2.3.2.2 Journey Status

In Table 6 - Frequency of Journey Statuses we can see multiple variations of the on-time category. Like the inconsistency in Ticket Type this cause fragmentation in frequency and can also impact future trend analysis.

2.3.2.3 Reason for Delay

In Table 7 - Frequency of Delay Reasons the categories showed several duplicates such as Signal Failure vs Signal failure, Staffing vs Staff Shortage, and Weather vs Weather Conditi. These are most likely due to manual entry and could lead to inaccurate frequency analysis.

2.3.2.4 Time Format and Date Format Variants

The original dataset contained mixed time and date formats (e.g., 2:00 PM vs. 14:00:00, or March 1, 2024 vs. 01/03/24). However, upon import into SAS, these fields were automatically converted into consistent internal formats.

2.3.2.5 Station Name Variants

As seen in Figure 9 - Bar Chart of Departure Stations and Figure 10 - Bar Chart of Arrival Destinations, Departure_Station and Arrival_Destination occasionally displayed inconsistent capitalisation or minor misspellings. This can prevent correct grouping and route analysis.

2.3.3 Outliers

As mentioned in 2.2.3.1 Price, the Price variable shows a right-skewed distribution, with most tickets priced under £50, but with a few records reaching as high as £267. These upper-end values may reflect long-distance First Class fares or unusual pricing scenarios.

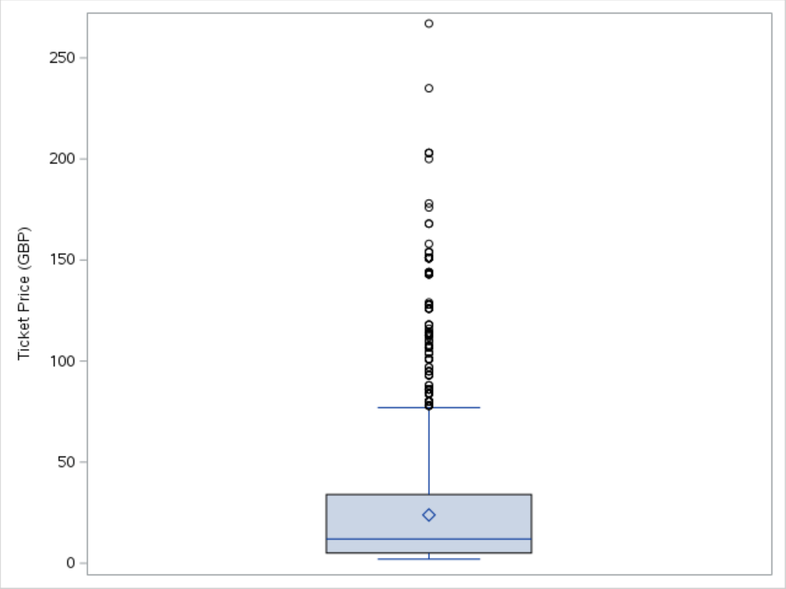


Table 11- Boxplot of Ticket Prices

Refer to Appendix A.19 – SAS Code for Price Outliers for SAS code

The boxplot reinforces this observation, with several data points lying well beyond the upper whisker a strong visual indicators of outliers.

3 Evaluation of Data Warehouse Concepts

A data warehouse plays a pivotal role in storing, consolidating and analysing large volumes of data. This section evaluates five key data warehouse concepts and is based around the needs of the UK National Rail context, supported by academic literature and examples from the dataset.

3.1 Data Integration and Cleaning Capability

"*Can the data warehouse effectively integrate and clean data from multiple sources?"*

Data integration and cleaning are essential for the reliability of a data warehouse. Integration combines data from various systems, for instance ticketing platforms, delay logs and other external sources into a single consolidated structure. Cleaning ensures that the data is accurate, consistent and without errors. Without integration and cleaning processes, any analysis can become fragments, flawed or incorrect.

In the context of a national rail system, poor integration and cleaning could lead to misleading insights for delays, refund patterns or route performance. The UK National Rail dataset includes fields such as ticket types, departure stations and reason for delays where values may be recorded differently or incorrectly. This can be seen in the initial data exploration in 2.3.2 Inconsistencies where ticket types and stations are misspelt. Without proper adjustments and standardisation, queries for any relevant insight would provide erroneous conclusions.

However, there are challenges to integration and cleaning. Namely, it can be resource-intensive, especially if dealing with older legacy systems, inconsistent formats and real-time data. Hence, metadata management and transformation logic must be approached carefully.

As noted by Tzika-Kostopoulou et al. (2024), integrating multiple datasets with traditional systems remains a titular challenge in unlocking the value of big data for traffic operations. This highlights the need to consolidate heterogeneous inputs such that inconsistent labels are resolved, and do not significantly impact the ability to generate accurate reports. Similarly, Golfarelli and Rizzi (2020) remark "a data warehouse should provide a unified view of all the data" (p. 16). This further reinforces the idea that integration and consistency in any data warehouse system is crucial for clean, accurate insights.

3.2 Time Variant Data Tracking

"*Does the data warehouse preserve historical data for long-term analysis"*

Time-variant data is a distinguishing feature of data warehousing, it allows users to analyse how data changes over time and not just its current state. Contrary to transactional systems, data warehouses retain historical records, which are crucial for gathering trends, patterns and forecasting from data.

For transport systems like the UK National Rail, accumulating timestamped data enables the study of delay trends across months, impact of scheduling changes, or monitor recurring refund patters. Without this historical context, it would be difficult for planners to detect seasonality, forecast peak hours, or track long-term performance metrics.

Nevertheless, time-variant data tracking does introduce complexity. Warehouse must handle Slowly Changing Dimensions (SCD), increased storage capacity for growing historical data, and optimise queries for time-based analytics. These must all be taken into consideration and require thoughtful schema planning and maintenance,

As Golfarelli and Rizzi (2020) clarify, modern business intelligence systems use time-flows to assemble interrelated events, enabling decision makers to understand both the cause and effect of relationships to create predictive insights. This is particularly important for UK National Rail data, where understanding journey delays, refund frequency and cheduling outcomes rely heavily on long-term tracking. Tzika-Kostopoulou et al. (2024) support this view in the transport domain, stressing that most urban mobility studies rely on historical data to detect spatial-temporal trends and predict travel demand. In the UK National Rail context, fields such as Date of Journey, Departure Time and Actual Arrival Time require accurate logging. Without a time aware structure, delay forecasting and service improvement analytics would not be possible.

3.3 Multidimensional Analysis Support

"Does the data warehouse support multidimensional analysis for flexible and efficient querying?"

Multidimensional analysis enables the exploration of datasets from a variety of perspectives through processes such as slicing, dicing, drill-down and roll-up. These processes are implemented through Online Analytical Processing (OLAP) techniques, which structures data into measures and hierarchies, which assists in gathering comprehensive insights.

In the context of UK National Rail dataset, this analysis would allow stakeholders to investigate delay patternrs by route, compare refunds by ticket type, or even specific days with peak issues. These techniques would support not only reporting but also strategic planning and service optimisation.

However, designing and maintaining a multidimensional schema can be quite challenging, especially when dealing with evolving data structures, high-dimensional queries and a large volume of data. It would require handling hierarchies such as day, month and quarter for example, while also assuring performance tuning across complex joins.

As Martinez-Mosquera et al. (2024) note, OLAP cubes enable the examination of data through multiple dimensions, allowing for more complete and efficient aggregation and insights across large volumes of data. This is especially important for the UK National Rail dataset, which includes multiple dimensions like reason for delays, ticket types, journey dates and stations. The capability to slice this data by route, roll up by month or even drill into refund patterns by ticket class is important for understanding operational trends and customer behaviour. By having the ability to navigate complex dataset though multidimensional structure it ensures that the data warehouse can support analysis across many fields that influence national rail performance.

3.4 Scalability and Query Performance

"Can the data warehouse scale to acocomodate growing data volumes while maintaining efficient query performance"

Scalability and performance are essential to ensuring the long-term efficiency of a data warehouse. Scalability pertains to the system's ability to handle the increasing volume of data while performance stresses the responsiveness and speed at which queries are executed. Without scalability and performance, analytical insights could be delayed or lost, reducing the data warehouse's effectiveness.

In regard to the UK National Rail dataset, the warehouse would store large volumes of data which increase daily. In time, the aggregation of the data would demand more storage capacity, faster indexing and optimised queries. The UK National Rail dataset which include both transaction and time-series data, would require a warehouse to handle consistent performance as an abundance of records are stored and queried over long periods.

Nonetheless, to achieve scalability and performance to meet those demands introduce its own challenges. This includes the need for optimised data models, portioning strategies, parallel processing and adaptive resource tuning. Without an appropriate design approach for optimisation, queries could become unnecessarily slow or resource intensive.

As Ahmadi (2023) explains, machine learning plays a crucial part in improving the responsiveness of cloud data warehouses by identifying performance bottlenecks, optimising query operations and adjusting resource allocation. This is especially important for transport systems like the UK National Rail, where sudden spikes in activity such as during major delays can overwhelm ill-equipped systems. Comparably, Golfarelli and Rizzi (2020) note that many warehouse architectures fail due to poor scalability and inadequate capacity for future data. The significance on creating systems that can easily accommodate growing data and user needs support the argument for why scalability and performance must be a priority consideration.

3.5 Analytical Usefulness and Decision Support

“Does the data warehouse enable meaningful analytical insights and support operational or strategic decision-making?”

A core goal of data warehousing is not just to store large amounts of data, but to also transform data into actionable insights. Analytical usefulness describes how well a data warehouse enables analysis on applicable business dimensions, whereas decision support places emphasis on how the analysis aids stakeholders in making informed decisions.

Within the context of UK National Rail, decision support would include determining which routes experience the most delays or gauging the effectiveness of refund policies. These insights rely on the data warehouse being structured around key subjects such as reason for delays, ticket types and stations which aligns with the subject orientated nature of data warehousing.

Despite that, providing such value does not rely solely on data quality and structure but also on user accessibility, metadata clarity, and meaningful aggregation of data. A data warehouse with excellent technical capabilities that cannot support the decision-making needs of stakeholder ultimately falls short of its purpose.

Golfarelli and Rizzi (2020) describe data warehousing as a set of techniques intended to help decision makers such as managers and analysts turn data into actionable insights. They further emphasize Inmon’s principle that a warehouse must be subject-oriented, consistent and time-variant, designed specifically to support organisational decision making rather than just operational tasks. When applied to the UK National Rail, this implies enabling relevant stakeholders to interpret complex operational data in such a way that it directly informs planning and strategy. As an example, this could be the ability to track delays across routes. By organising the data around business relevant subjects like routes, delays and refunds, the warehouse would facilitate decision makers to derive targeted, evidence-based conclusions.

5 Conclusion

The exploration od the UK National Rail dataset and context revealed meaningful insights, and spotlights the dataset's possible usage for predictive modelling and operational insights. Furthermore, the evaluation of the data warehouse concepts magnifies the importance of integration, scalability and time-variant tracking in bolstering effective decision-making. Together, these insights set a foundation for further analysis in part two of the report.

Bibliography

Ahmadi, S. (2023). Optimizing Data Warehousing Performance through Machine Learning Algorithms in the Cloud. *International Journal of Science and Research (IJSR)*, *12*(12), 1859–1867. https://doi.org/10.21275/sr231224074241

Golfarelli, Matteo., & Rizzi, Stefano. (2009). *Data warehouse design : modern principles and methodologies*. McGraw-Hill.

Martinez-Mosquera, D., Navarrete, R., Luján-Mora, S., Recalde, L., & Andrade-Cabrera, A. (2024). Integrating OLAP with NoSQL Databases in Big Data Environments: Systematic Mapping. In *Big Data and Cognitive Computing* (Vol. 8, Issue 6). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/bdcc8060064

Tzika-Kostopoulou, D., Nathanail, E., & Kokkinos, K. (2024). Big data in transportation: a systematic literature analysis and topic classification. *Knowledge and Information Systems*, *66*(8), 5021–5046. https://doi.org/10.1007/s10115-024-02112-8

Appendix A – SAS Code

Appendix A.1 – SAS Code for Purchase_Type

```
/* Purchase Type - Frequency count + bar chart */
proc freq data=clean_railway;
  tables Purchase_Type / plots=freqplot;
run;
```

Appendix A.2 – SAS Code for Payment_Method

```
/* Payment Method - Frequency table + bar chart */
proc freq data=clean_railway;
  tables Payment_Method / plots=freqplot;
run;
```

Appendix A.3 – SAS Code for Railcard

```
/* Railcard - Frequency count + bar chart */
proc freq data=clean_railway;
  tables Railcard / plots=freqplot;
run;
```

Appendix A.4 – SAS Code for Ticket_Class

```
/* Ticket Class - Ordered categories */
proc freq data=clean_railway;
  tables Ticket_Class / plots=freqplot;
run;
```

Appendix A.5 – SAS Code for Ticket_Type

```
/* Ticket Type - e.g., Advance, Anytime, Season */
proc freq data=clean_railway;
  tables Ticket_Type / plots=freqplot;
run;
```

Appendix A.6 – SAS Code for Journey_Status

```
/* Journey Status - Delayed vs On Time */
proc freq data=clean_railway;
  tables Journey_Status / plots=freqplot;
run;
```

Appendix A.7 – SAS Code for Reason_for_Delay

```
/* Reason for Delay - Frequency + bar chart */
proc freq data=clean_railway;
  tables Reason_for_Delay / plots=freqplot;
run;
```

Appendix A.8 – SAS Code for Refund_Request

```
/* Refund Request - Frequency of Yes/No */
proc freq data=clean_railway;
  tables Refund_Request / plots=freqplot;
run;
```

Appendix A.9 – SAS Code for Departure_Station

```
proc sgplot data=Freq_Departure;
  vbar Departure_Station / response=Count datalabel;
  xaxis display=(nolabel) fitpolicy=rotate;
  yaxis label="Number of Departures";
run;
```

Appendix A.10 – SAS Code for Arrival_Destination

```
/* Arrival Destination – Frequency + full bar chart */
proc freq data=clean_railway noprint;
  tables Arrival_Destination / out=Freq_Arrival;
run;

proc sgplot data=Freq_Arrival;
  vbar Arrival_Destination / response=Count datalabel;
  xaxis display=(nolabel) fitpolicy=rotate;
  yaxis label="Number of Arrivals";
run;
```

Appendix A.11 – SAS Code for Date_of_Purchase

```
/* Date of Purchase – Frequency distribution and line chart
*/
proc freq data=clean_railway noprint;
  tables Date_of_Purchase / out=Freq_Date_of_Purchase;
run;

proc sgplot data=Freq_Date_of_Purchase;
  series x=Date_of_Purchase y=Count;
  xaxis label="Date of Purchase";
  yaxis label="Number of Purchases";
run;
```

Appendix A.12 – SAS Code for Time_of_Purchase

```
/* Time of Purchase – Convert to hour and visualize
distribution */
data clean_railway;
  set clean_railway;
  Hour_of_Purchase = hour(Time_of_Purchase);
run;

proc freq data=clean_railway;
  tables Hour_of_Purchase;
run;

proc sgplot data=clean_railway;
  histogram Hour_of_Purchase / binwidth=1;
  xaxis label="Hour of Purchase";
run;
```

Appendix A.13 – SAS Code for Date_of_Journey

```
/* Date of Journey - Frequency table and travel trend */
proc freq data=clean_railway noprint;
  tables Date_of_Journey / out=Freq_Date_of_Journey;
run;

proc sgplot data=Freq_Date_of_Journey;
  series x=Date_of_Journey y=Count;
  xaxis label="Date of Journey";
  yaxis label="Number of Journeys";
run;
```

Appendix A.14 – SAS Code for Departure_Time

```
/* Departure Time - Analyse by hour */
data clean_railway;
  set clean_railway;
  Hour_of_Departure = hour(Departure_Time);
run;
```

Appendix A.15 – SAS Code for Arrival_Time

```
/* Arrival Time - Analyse by hour */
data clean_railway;
  set clean_railway;
  Hour_of_Arrival = hour(Arrival_Time);
run;

proc freq data=clean_railway;
  tables Hour_of_Arrival;
run;

proc sgplot data=clean_railway;
  histogram Hour_of_Arrival / binwidth=1;
  xaxis label="Scheduled Arrival Hour";
run;
```

Appendix A.16 – SAS Code for Actual_Arrival_Time

```
/* Actual Arrival Time - Analyse by hour */
data clean_railway;
  set clean_railway;
  Hour_of_Actual_Arrival = hour(Actual_Arrival_Time);
run;

proc freq data=clean_railway;
  tables Hour_of_Actual_Arrival;
run;

proc sgplot data=clean_railway;
  histogram Hour_of_Actual_Arrival / binwidth=1;
  xaxis label="Actual Arrival Hour";
run;
```

Appendix A.17 – SAS Code for Price

```
/* Price - Descriptive statistics */
proc means data=clean_railway mean median min max std var n
maxdec=2;
  var Price;
run;

/* Price - Histogram to show distribution */
proc sgplot data=clean_railway;
  histogram Price;
  xaxis label="Ticket Price (GBP)";
run;
```

Appendix A.18 – SAS Code for Missing Values

```
/* Summary of missing values for key fields */
proc freq data=clean_railway;
  tables
    Purchase_Type
    Payment_Method
    Railcard
    Ticket_Class
    Ticket_Type
    Departure_Station
    Arrival_Destination
    Journey_Status
    Reason_for_Delay
    Refund_Request
  / missing;
run;

proc means data=clean_railway nmiss;
run;

proc means data=clean_railway n nmiss;
  var Actual_Arrival_Time;
run;
```

Appendix A.19 – SAS Code for Price Outliers

```
/* Price - Boxplot to identify outliers visually */
proc sgplot data=clean_railway;
  title "Boxplot of Ticket Prices";
  vbox Price;
  yaxis label="Ticket Price (GBP)";
run;
```