Analysis of Euston to Manchester Route Performance

Group Gamma
Project Host:

Avanti West Coast

Ferentinou, Theodora	Petrov, Vladimir	Parth, Shreyas
36458651	36534477	36647998
Analysis, Report	Analysis, Report, Presentation	Report, Presentation
Wei, Chen	Yadav, Dipesh	Zhu, Gary
36441335	36509669	36639227
Analysis, Report	Pre-processing, Analysis, Report	Pre-processing, Report

February 2024

I Introduction

Avanti West Coast is one of the major British train operating companies. Since its creation, in the last quarter of 2019, it has been overseeing the West Coast Main Line, connecting major cities across the United Kingdom, such as London, Birmingham, Manchester, etc. Avanti West Coast distinguish themselves by offering not only high-speed intercity services but unparalleled personal experience with their comfortable coaches, on-train services and customer service. In addition, the company has taken significant steps to stop global warming, some of which include reducing water consumption, improving train efficiency, recycling, etc.

Although Avanti West Coast has managed to maintain superb services throughout the years, one problem has significantly impacted their overall performance - the London Euston - Manchester service. This route currently has the lowest moving annual average performance (36%) among long-distance service groups.

The main objective of this research was to conduct a thorough evaluation of the operational performance of Avanti West Coast from 2017 to 2024, to address that poor performance. In addition, four specific research questions were presented: "What are the top causes of poor performance and who is responsible for these causes? What are the causes of poor performance on a 'good' day (e.g. 25th percentile)? Where is time lost on the Euston to Manchester Route, which is significantly impacting on time performance? Based on the available data, what are the recommendations for improving train service performance?".

In order to answer the questions posed, we were provided with multiple Excel and CSV files containing delays and cancellations information. However, before starting analysing the data, some preprocessing had to be performed. It included understanding the meaning of the features in the data, dealing with missing values and then extracting all the data from the files using Python and R. In addition, since the data files could be divided into three groups, a unique analysis was performed for each separate group.

In the first part of the analysis, it will be investigated if the Avanti West Coast trains experience higher delays than other operators on the route Euston - Manchester. To complete the analysis, a graphical representation of the data in combination with segmentation and computation of averages was performed. The graphics were then used to search for any trends or big differences in delays between Avanti West Coast and other operators.

The second part of the analysis will investigate the differences among the various routes, using a multitude of statistical tests and graphical representations. The ultimate goal of this analysis is to observe the route with the largest number of delays which can lead to a noticeable deterioration in the performance of Avanti West Coast.

The third part of the analysis concentrates on identifying key factors, locations, and service providers responsible for service disruptions within the specified timeframe. It also pays special attention to the Euston - Manchester route, examining the principal reasons behind service interruptions and the managers (managing entities) accountable for these issues. Additionally, this part aims to assess the impact of initial and subsequent delays across different routes. Through the use of periodic summaries and average data visualizations, this analysis seeks to highlight significant findings and propose actionable solutions.

II Preprocessing

During data preprocessing, removing columns with a high amount of missing data and rows lacking crucial information, combined with applying sliding (rolling) imputation to key data, is an effective strategy to prevent bias and enhance the reliability of analysis. As emphasized by Schafer (1997), deleting rows that lack essential information is critical for maintaining the integrity of the dataset. Moreover, leveraging the assumption of temporal correlation in time series data through sliding imputation effectively preserves the continuity and patterns in data, such as train operation times and platform waiting times. This approach follows Schafer's (1997) principle of maintaining dataset completeness while minimizing bias introduction, thus making the dataset more valuable and representative for accurate insights and predictive modelling (Schafer, 1997). The technique is used on the "Quartz" dataset (which consists of punctuality data of each station on the Euston - Manchester route) and the "Acumen" dataset for filtering the missing values.

The dataset "Period Total Punctuality" encompasses various aspects of railway performance, spanning from November 2017 through November 2023, and aligns with the financial year period for railways, defined as 18/01 to 18/13, which translates to 1 April 2017 to 31 March 2018. This dataset is divided into several sheets, each focusing on different metrics of train service performance. More specifically the period from 2019 to 2024 has been denoted by the variable "CP6" (control period 6), which encompasses specific rail projects and work delivered during that time. Part of the dataset concentrates on train punctuality, categorizing it into several time brackets such as "Early", "On Time", "Within 3 mins", etc., with both raw numbers and percentage values for each category. This provides a comprehensive view of punctuality across various time frames. Another part of the dataset tracks the mileage covered by different train service groups, labelled "HF01" through "HF08", over specified periods. This data helps in assessing the operational coverage and intensity of service groups over time.

The "Acumen Cancellation and Delay" dataset included various metrics that are related to train cancellations, such as the unique identifiers, dates, and comprehensive descriptions of each incident. The data includes the "Inc date" and "Inc Cause" with its description, offering specific reasons behind each incident. The location and section of the delay are documented, helping to pinpoint where and under what circumstances the dis-

ruptions occurred. The detailed information about the trains, including their ID, descriptions, journey descriptions, and route details, are provided to give a full picture of the affected services. The dataset also describes the managing entities involved, with "Trust manager desc" identifying the responsible railway management services and their associated routes, respectively.

The data from the "Acumen cancellations and Acumen delays" dataset contains null values for certain variables. During the exploratory data analysis phase, it was determined that multiple of the features (including "Root Cause", "Delay Cause Desc", "Resp Train", etc) were redundant as other more significant correlated variables adequately represented their information. Consequently, these variables were dropped from further analysis. For the remaining variables, instances with missing data accounted for <1% of the dataset and therefore particular rows were dropped from the analysis to ensure data integrity and accuracy.

III Analysis

1 Punctuality on Euston - Manchester Route

To find if Avanti experienced unusually large delays on the Euston – Manchester route, multidimensional data analysis on the "Quartz" dataset was performed over specific periods. In this way, it is possible to identify any existing trends in delays or big differences between operators. The process included splitting the data into two parts - one consisting of Avanti trains only and another consisting of the other operators. Then, for each day, the planned arrival and departure times were compared against the real ones. This comparison gives the increase in delay at each part of the route. However, as some trains might depart a bit early from the stations, leading to a negative increase in delay, all such cases were set to 0 seconds. This ensures that early trains do not have a bigger weight than trains which are on time.

By plotting the average lateness for each day of the week for Avanti West Coast trains compared to other operators (Figure III.1) it could be seen that Avanti do perform worse than the other operators on some of the stations. For example, in Manchester, there could be observed about 2 times higher average increase for each day of the week compared to other operators.

Even though the average increase in delay for a single station is less than half a minute for the majority of days and stations, the accumulation of several such delays over the course of multiple stations can severely lower the overall performance. To find which are the biggest causes for this disturbance, the same dataset was used, as for many of the delayed trains the specific reason was marked down. Singling out the most common issues, it was found that passenger boarding (too many passengers at the station, passengers needing assistance, etc) was the biggest cause for the increase in delay at most stations. From Figure III.2 it can be seen that on average, for small delays (less

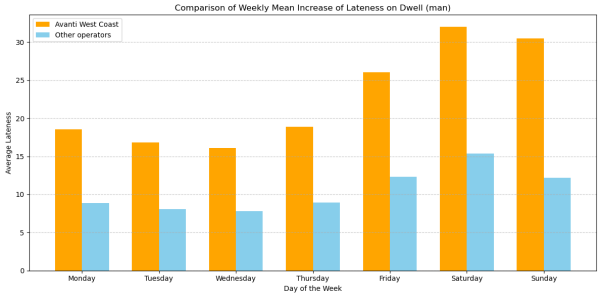


Figure III.1: Bar chart of the average increase in delay in Manchester for each day of the week of Avanti West Coast trains compared to other train operators.

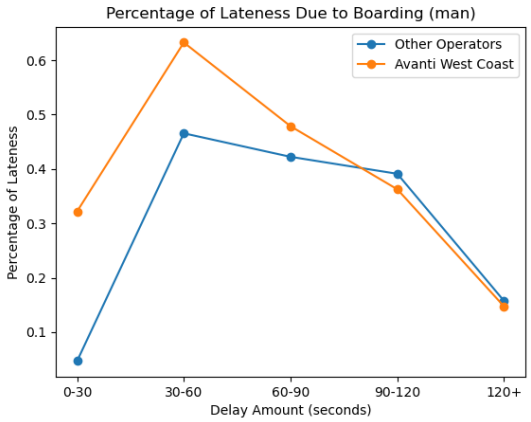


Figure III.2: Lineplot comparing the percentage of delays caused by boarding for Avanti trains vs other train operators.

than 90 seconds), boarding is a bigger issue for Avanti compared to other operators. Hence, Avanti doesn't accurately predict the amount of time needed for boarding, especially on busier days or when there are fewer trains (Saturday and Sunday).

The second biggest delay was following another late train. This is somehow out of the control of any operator and is quite random by nature, because of which not a big weight can be given to it. Furthermore, from Figure III.3 it can be seen that the lost time of Avanti and other operators is almost identical. Hence, it can be deduced that following another late train is not a major reason for the poor performance of the Euston – Manchester route.

In addition to daily trends, a deeper analysis of the hourly data was performed to see if there were any trends. From the initial results, it was found that for some stations the graph peaked around midnight. However, as this is possibly due to these being the last trains of the day, they can be ignored for the rest of the analysis. Taking this into account it could be found that for most stations 12 pm seemed to be the busiest time of the day with graphs usually peaking there. However, no significant trends could be observed among all stations.

Lastly, using this data it could be checked if the performance on good days (the 25 percentile) for the Avanti trains is comparable to the performance of the other operators. Leaving only the good performing days for Avanti and plotting the lateness increase again (Figure III.4) it

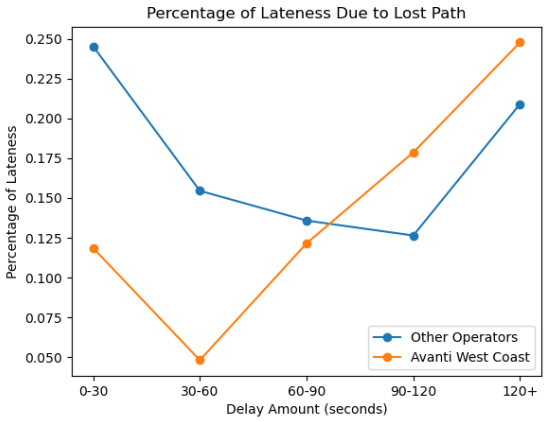


Figure III.3: Lineplot comparing the percentage of delays caused by lost path for Avanti trains vs other train operators.

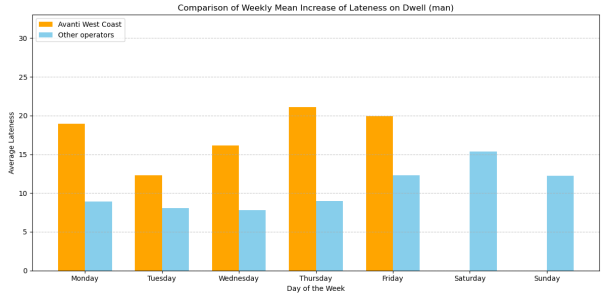


Figure III.4: Bar chart of the average increase in delay in Manchester for each day of the week of Avanti West Coast trains compared to other train operators, taking only the days with good performance (25th percentile).

can be seen that on average the delay is lower. However, a big note is that none of the Saturdays or Sundays are in the 25 percentiles. This again shows that on the days when there are fewer trains, the delays are far bigger. One of the reasons for this might be the fact that if one of the few trains is cancelled, passengers struggle to board and most of the trains become overcrowded. As an alternative, it can be that in general there are too few trains on the weekend. In addition, even on good days, Avanti West Coast gets outperformed by other train operators when it comes to boarding at most stations.

2 Differentiation in Routes

The focus of this part of the analysis is to implement hypothesis tests and visualise the differences among the various routes of Avanti West Coast, based on the time period and date. In particular, it is important to discover the route that records the greatest number of delays and therefore impacts the overall performance of Avanti.

The data that were utilized for the purpose of the analysis are the "CP6" measures obtained from the "Period totals punctuality" dataset. Specifically, the variables concerning the total number of trains that were on time, 20 minutes and 30 minutes late, were used as input for the implementation of the statistical analysis. In addition, the various routes that were tested are the "West Mids", "North Wales", "Manchester", "Liverpool", "North West", "Euston-Scot (via TV)", "Euston-Bhm-Scot" and "Anglo-Scot".

It is worth mentioning that prior to implementing statistical tests to compare the various routes, it is crucial to assess the quality of the data in order to identify interesting patterns, as well as to detect potential anomalies in the various features. The assessment of normality is an important step since it is the primary assumption of parametric testing. The Shapiro-Wilk test is the proposed numerical method, with a null hypothesis that the sample is normally distributed. The results obtained from the normality test indicated that the data have not been generated from a normal distribution, hence why non-parametric tests will be implemented to check whether there are differences among the various routes.

The non-parametric test that will be used in this analysis is the Kruskal-Wallis test which determines whether there are differences among the medians of two or more groups. In cases where the null hypothesis of the test is rejected, it is advisable to identify the groups that differ from each other, using graphical representations such as Boxplots.

With the purpose of examining whether the total number of trains ("On Time" and "Delayed"), depends on the route, the chi-squared test was implemented. In particular, the adjusted residuals of the chi-squared test are used to indicate whether there is an association between the variables, and in the case of this analysis, it indicates the association between the route and the number of delays. The null hypothesis of the chi-squared test is that the two variables "On Time" and "Delayed" are independent, with the assumption that the adjusted residuals have been generated from a standard normal distribution. For that reason, in cases where the adjusted residuals have absolute values of 2.0 or higher in a cell, then the null hypothesis is rejected, indicating that there is an association of the variables in that cell.

According to Table III.1 the Shapiro-Wilk test gives p-values that are smaller than $alpha = 0.05$, signifying that the data are not normally distributed, thus it is advisable to implement the non-parametric Kruskal-Wallis test to check whether there are differences in the "CP6" measures for the various routes. In addition, according to Table III.1 the Kruskal-Wallis test gives p-values that are smaller than $alpha = 0.05$, denoting that the null hypothesis that the CP6 measures are equal among the different routes, is rejected at a significance level of $alpha = 5\%$. In order to observe which routes differ from each other, the Boxplots in Figure III.5 and Figure III.6 are created.

It is apparent from Figure III.5 that the most punctual trains are the ones that follow the "North West" and "Anglo-Scot" routes, in contrast with the trains that follow the "North Wales" route which seem to be delayed the most. In addition, it is evident that there is no statistically significant difference in the performance of the "Euston-Bhm-Scot", "Manchester" and "West Mids" trains since they correspond to approximately 5000 trains that are On Time.

Variables	Shapiro-Wilk	Kruskal-Wallis
CP6: On Time	<.001	<.001
CP6: 20+ minutes	<.001	<.001
CP6: 30+ minutes	<.001	<.001

Table III.1: Shapiro-Wilk and Kruskal-Wallis p-values for the CP6 measures.

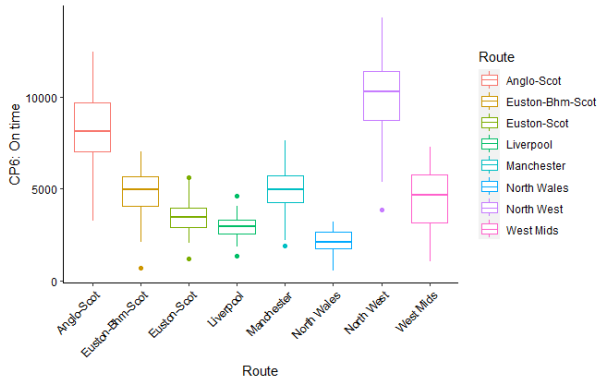


Figure III.5: Boxplot regarding the total number of trains that were On Time for each route.

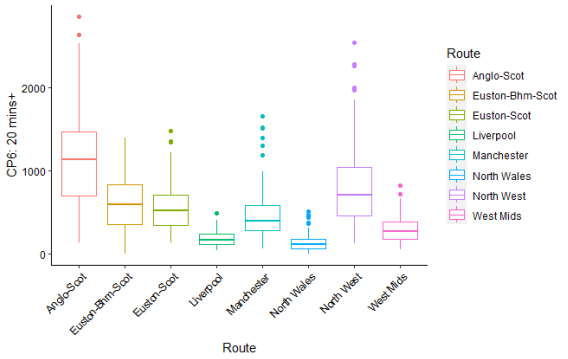
Similarly, the plots in Figure III.6 display several interesting patterns regarding the performance of each route. To begin with, although the trains of the "Anglo-Scot" and "North West" routes were the most punctual according to Figure III.5, they seem to be the ones that get delayed the most as well, since they seem to arrive at the destination 20 and 30 minutes late. Furthermore, there is no large difference in the performance among the "Euston-Bhm-Scot", "Euston-Scot" and "Manchester" trains, as well as among the "Liverpool", "North Wales" and "West Mids" trains.

Due to the fact that the results obtained from the Kruskal-Wallis test and the Boxplots do not give any interpretable and noticeable conclusion, we proceed with implementing the chi-squared test, with the ultimate goal of finding the route that is associated with the largest number of delays.

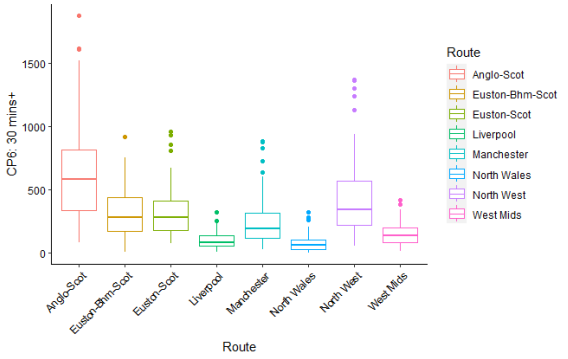
For the purpose of this analysis, the variables that will be tested are the "CP6: On Time" and "CP6: 20+ minutes late". The chi-squared test gives a relatively small p-value, signifying that the null hypothesis that the two variables are independent, is rejected at an $alpha = 5\%$ significance level. In other words, the chi-squared test indicates that there is an association between the trains that are on time and the trains that are delayed. According to Table III.2, the "West Mids", "North Wales" and "Manchester" routes have a strong association with the number of delays, since the adjusted residuals for these specific routes correspond to significantly high values of 105.58247, 52.69645 and 90.40423 respectively.

3 Cancellation and Delay root causes

The analysis of the "Acumen cancellation and delay" dataset involved a detailed periodical examination of cancellations and delays, focusing on the most affected locations, locations where delays occurred, the reasons behind these events, and key intersections. To identify the



(a) Boxplot regarding the number of trains that were 20 minutes late.



(b) Boxplot regarding the number of trains that were 30 minutes late.

Figure III.6: Boxplots regarding the total number of trains that were delayed for each route.

Arrivals		
Route	On Time	Delayed
West Mids	-105.58247	105.58247
North Wales	-52.69645	52.69645
Manchester	-90.40423	90.40423
Liverpool	63.18875	-63.18875
North West	13.08312	-13.08312
Euston-Scot	52.65978	-52.65978
Euston-Bhm-Scot	74.30877	-74.30877
Anglo-Scot	73.20217	-73.20217

Table III.2: Adjusted Residuals of the chi-squared test.

causes, we utilized the "Inc Cause Desc" variable for cancellations and "Delay Cause Desc" for delays, filtering for the top eight occurrences to uncover the most common reasons throughout the entire period. This method allowed us to understand the core issues. Moreover, to pinpoint the service providers that are most accountable for these delays and cancellations, we selected the top eight entries under "Trust Manager Description". This strategy enabled the identification of the frequently involved service providers, facilitating efforts to engage with them for collaboration and to devise solutions to address the recurring failures.

Additionally, our analysis honed in on the Euston-Manchester route, examining the primary reasons for cancellations or delays, along with identifying the responsible managers (see Table III.3). For this specific route, we employed the unique identifier from the

Cause		
Category	Type	Count
Canceled	Driver	922
Canceled	Rail Defect	488
Canceled	Train Manager	451
Delayed	Lost Path - following late train	13086
Delayed	Lost Path - following on time train	5619
Delayed	Lost Path - regulated for late train	4306
Service Provider		
Category	Manager	Count
Canceled	NWE MANCHESTER DU	953
Canceled	WCS AD WEST COAST EXTERNAL	554
Canceled	AWC DRIVERS MANCHESTER	504
Delayed	WCS BLETCHLEY DU	4942
Delayed	WCS STAFFORD DU	4298
Delayed	WCS EUSTON DU	3680

Table III.3: Euston-Manchester route cancellation and delay responsible causes and services.

route manager "Bugle Prof Centre Desc", selecting the top eight values from both "Trust Manager Description" and "Inc Cause Desc". This approach allowed us to catalogue the main factors leading to disruptions on the Euston-Manchester route and pinpoint the service providers accountable. Consequently, this facilitated engagements with service managers to discuss the identified issues and collaborate on developing solutions to mitigate these causes. Further, the locations and junctions were also filtered from the data to identify the hub for most cancellations and delays, which resulted in listing EUSTON, CHU and MANCR PIC as main locations for cancellations and EUSTON, MILTON KC and RUGBY as main locations for delays. The most responsible junctions for delays were HANSLOPEJ-MILTON KC, CAMDENSJN-EUSTON, ATTLBROSJ-NUNEATON and for cancellations MANCR PIC, EUSTON, STOCKPORT-STOCKPORT. It helped in spotting the problem hub locations and the core areas.

To understand the broad impact of various causes over time, a histogram was created (see Figure III.7) that showcases the top five causes, providing a yearly view of the reasons behind cancellations or delays. Additionally, a daily hourly chart was developed to identify peak times and the main causes of the top five that affect daily operations. This chart offers insights into the reasons behind fluctuations and patterns in service disruptions. For cancellations, we plotted the histogram using 'Inc Date' and the top five causes as keys to examine their periodic distribution over time. This analysis was instrumental in identifying the most significant cause for each month within a year, enabling targeted improvements in service areas during specific times to enhance operational efficiency. For delays, a similar approach was adopted, but due to the "Inc Date" column containing only the time "00:00:00", which made it an insignificant variable, we used the "Inc Text" variable instead. By extracting the date using regular expressions, we conducted a similar analysis to achieve comparable insights for delays.

Additional analysis was carried out on the delay dataset, focusing on the primary cause delay minutes and

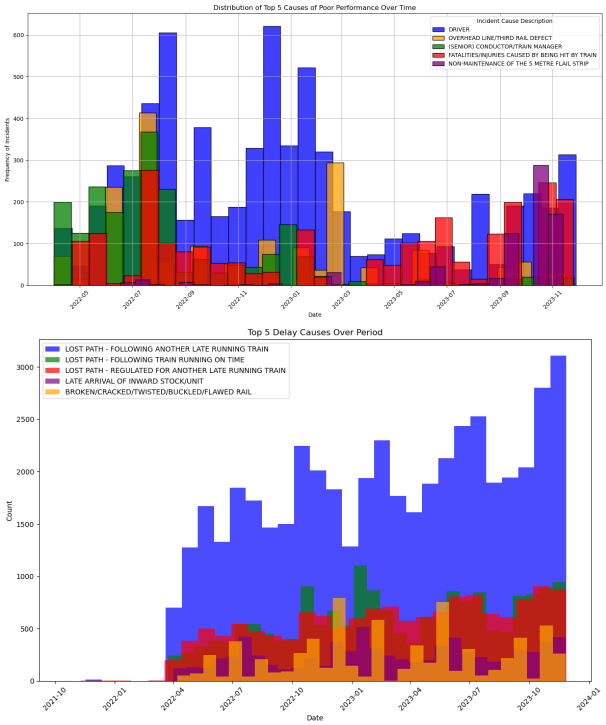


Figure III.7: Periodic Cause distribution for cancellation and delay

Route	Primary Train	Count	Reactionary Train	Count
BIRMINGHAM - SCOTLAND	9M54	2600.0	9M62	5320.78
CL.390 TEST TRAINS	9G49	159.0	9G49	246.00
EUSTON - LIVERPOOL	1F28	1413.0	1F28	2341.00
EUSTON - MANCHESTER	1H74	1804.0	1H73	3548.00
EUSTON - NORTH WALES	1D91	1615.0	1G00	3092.50
EUSTON - SCOTLAND	1S90	3286.5	1M18	5355.00
EUSTON - WEST MIDLANDS	9G48	1839.0	9G47	4009.00

Table III.4: Combined Data Overview

reactionary cause delay minutes to gauge the average delay duration attributable to the initial incident and the subsequent reactionary delays for specific routes per incident (refer to Table III.5). This investigation aids in quantifying the escalation in delay minutes from the primary source to subsequent locations along the route, illustrating how an initial delay not only impacts the immediate operation but also affects the scheduling and timeliness of other trains on the route, thereby contributing to further compounded delay times (see Figure III.8).

Moreover, we catalogued the main trains for the locations concerned, alongside their total accumulated delay time for the entire period for those route locations. We also identified the trains most significantly impacted by reactionary delays to pinpoint those most frequently affected on the routes. By calculating the averages of primary and reactionary delay times, using the route as a key, we derived the average primary and reactionary delay times for the routes. Subsequently, we determined the trains with the highest frequency on the route for primary and reactionary delays as keys to identify the most affected trains (see Figure III.4). This approach enables us to track the trains most impacted over the period, providing a basis for targeted improvements to service to mitigate both primary and reactionary delays.

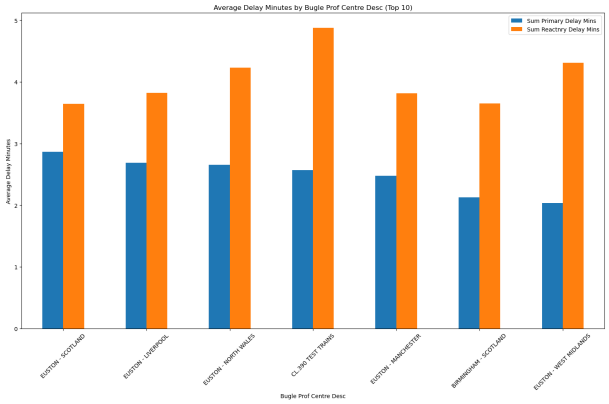


Figure III.8: Primary and reactionary Delay

Route	Primary min	Reactionary min
EUSTON - SCOTLAND	2.867778	3.640743
EUSTON - LIVERPOOL	2.687486	3.823253
EUSTON - NORTH WALES	2.651120	4.228548
CL.390 TEST TRAINS	2.567568	4.873874
EUSTON - MANCHESTER	2.473788	3.815350
BIRMINGHAM - SCOTLAND	2.129749	3.650207
EUSTON - WEST MIDLANDS	2.037480	4.308036

Table III.5: Route primary and reactionary delay time average

IV Conclusion

In this report, we were presented with the task of identifying major issues on the Avanti West Coast's London Euston - Manchester train route. This route has been one of the worst-performing long-distance ones in the UK among all operators. To achieve this, we were presented with multiple CSV and Excel files which contained punctuality data for different stations, operators and cancellations data.

In order to complete any analysis, first the data had to be preprocessed, which included identifying the meaning of all features, dealing with missing data, and converting data types. Then the analysis was split into 3 parts.

The first part consisted of identifying any remarkable problems on the London Euston - Manchester route, by comparing the performance of Avanti to other operators on each station. This was accomplished by visually comparing the average increase in delay on each station of Avanti against other operators. The results showed that one of the biggest reasons for delays is that Avanti often underestimate the number of passengers boarding at many of the stations. In addition, on good days (25th percentile) it was found that weekend trains are the worst-performing ones. In order to fix this, it is recommended that Avanti increase the number of trains issued on this route, especially during the weekend. If this is not possible, a more simple solution would be to allocate more time for boarding.

The second part of the analysis regarding the differentiation in routes determined the train stations that severely affect the overall performance of Avanti on numerous routes. The final outcome derived from the implementation of statistical tests and the graphical representations of the differences in arrival time among the various locations indicated that the "West Mids", "North Wales" and "Manchester" stations had a considerably large number of delays, resulting in the underperformance of Avanti on these specific routes. Henceforth, for improvement purposes, it is advisable to track trains at these stations in real-time, in order to estimate the arrival date of the train and organize the boarding and alighting flow accordingly. Moreover, in cases of overcrowded stations, it is recommended to allocate people who could assist with the smooth boarding process of passengers on trains and improve the platform train interface system.

The third part of the analysis reveals that Euston and Manchester Picadilly are the key locations that are most impacted by cancellations, whereas Euston, Milton KC, and Rugby are most impacted by delays. The primary causes for cancellations were identified as driver shortages, incidents of being struck by a train, and issues related to train management. In terms of delays, the predominant reasons included the loss of path for following trains, the late arrival of incoming rolling stock, and deficiencies in rail infrastructure. For the Euston to Manchester route, the main responsible service providers contributing to cancellations were NWE Manchester DU, WCS AD West Coast External, and AWC Drivers Manchester, while WCS Bletchley DU, WCS Stafford DU, and WCS Euston DU were primarily responsible for delays. The analysis showed that routes with the highest primary delays aren't always those with the highest reactionary delays. Specifically, the Euston to Scotland route faced the most primary delays, while the Euston to West Midlands and Euston to North Wales routes were most affected by reactionary delays. This underscores how delays vary across routes. To improve service reliability, Avanti should focus on the highlighted trains and train managers, prioritize hiring more drivers to address cancellations from driver shortages, enhance rail track security, and collaborate with professional train management services. These steps aim to address the root causes of delays and cancellations, enhancing overall service efficiency.

Finally, in order to improve on our results a few steps could be taken. For the first part of the analysis, to better understand the boarding problems, the overall delay of the train can be compared against the increase in delay at specific stations. This would give a better picture if late trains tend to get further delayed on their route than on-time trains. For the second part of the analysis, it is recommended to implement multiple comparison tests, known as post hoc tests, to identify the groups that differ from each other, with higher accuracy. For the third part of the analysis, the primary delay min average for the junctions can be found and then to narrow down the search to find the actual point of cause of delay on the route. It can help to understand which points are adding up reactionary delay to the route. In addition, outside events (train strikes, weather storms, etc.) can be taken into account when analysing the data as they might negatively impact the long-distance trains offered by Avanti. Lastly, more data for routes which have faced similar issues can be analysed, in order to understand how they have dealt with them.

References

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC.*

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer.

Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S. New York*, Springer.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.*, SAGE Publications.

Hastie, T., Tibshirani, R.,, Friedman, J. (2001). *The Elements of Statistical Learning.* New York, NY, USA: Springer New York Inc.

Tan, Pang-Ning., Steinbach, Michael, and Kumar, Vipin. *Introduction to Data Mining.* 1st ed. Pearson Education UK, 2013.

Berthold, Michael R., Borgelt, Christian, Hoppner, Frank, Klawonn, F, and SpringerLink. *Guide to Intelligent Data Analysis How to Intelligently Make Sense of Real Data. of Texts in Computer Science, 42.* London: Springer London, 2010.

Nussbaumer Knaflic, C. (2015). *Storytelling with data* (C. N. Knaflic, Ed.). John Wiley & Sons.

Schafer, J. (1997).*Analysis of Incomplete Multivariate Data.*United Kingdom: CRC Press.

Ghosh, C. (2022). *Data Pre-processing. In: Data Analysis with Machine Learning for Psychologists.* Springer, Cham.