LONDON TUBE PERFORMANCE

YACINE ND.

TABLE OF CONTENTS

Project Overview

This project analyses London Underground line-level performance using Excess Journey Time (EJT) data. The analysis conducted on the dataset will help identify which Underground lines are the most performant, how performance varies over time, and how consistently each line delivers service. The project is implemented entirely in SQLite, with results visualised in Tableau and documented for portfolio presentation on GitHub.

Dataset information

The London underground performance dataset was sourced from "Zenodo". It was uploaded on October 24th, 2019, by a user named William Gilks (Gilks, 2019). The original data was provided by "Mango Solutions", a consultancy with transport data expertise (Open Aire, 2025). The dataset was published under the creative commons attribution 4.0 international license (CC-BY 4.0), meaning it's open for reuse and to build upon (Zenodo,2025).

The dataset contains 1,050 rows and 9 columns (Line, Month, Scheduled, Excess, Total, Opened, Length, Type and Stations). The Elizabeth line did not exist as a London underground line at the time this dataset was created and as such is not included (London Assembly, 2025). The lines concerned are the Bakerloo, Central, Circle & Ham, District, Jubilee, Metropolitan, Northern, Piccadilly, Victoria and Waterloo & City.

Business Question & KPI

The analysis is conducted to help answer the question as to "Which London Underground line is the most performant based on excess journey time, and how does performance vary over time?" This interrogation focuses on relative performance across lines, temporal variation in performance, and consistency versus volatility of service quality. The analysis is built around industry-standard TFL metrics:

- Average Excess Journey Time (EJT) per Line: measures overall line performance, lower values indicate better performance.
- Monthly Excess Journey Time Trend: tracks how performance changes over time.
- Performance Rank by Line: ranks lines from most to least performant.
- Performance Consistency: is measured using variance or standard deviation of excess journey time per line.

Phase 1: Dataset Understanding & Data Quality Audit

The purpose of this phase is to understand the structure, scope, and quality of the London Underground performance dataset before performing any cleaning or analysis. This ensures that subsequent transformations and KPIs are grounded in verified data characteristics.

1.1 Dataset Structure and Grain

The dataset is stored in a single table named “LTD”. An initial inspection of the table schema confirms that the data is structured at a line-by-month grain**,** where each row represents aggregated performance metrics for a London Underground line in a given month. Key fields include:

- Line: Name of the Underground line
- Month: Numeric representation of the reporting month
- Scheduled: Planned journey time
- Excess: Excess Journey Time (performance metric)
- Total: Actual journey time
- Line metadata: Length, number of stations, opening year, and line type

This structure confirms that the dataset does not contain event-level or timestamped train records and is therefore suitable for line-level performance analysis over time**.**

1.2 Temporal Coverage

To assess the time span of the dataset and ensure adequate historical coverage, the minimum and maximum values of the “Month” field are reviewed, along with the total number of records. This confirms the range of months covered and whether each line is consistently represented over time. The key questions addressed include how many months are included, whether all lines have similar temporal coverage, and the presence of gaps in reporting?

The output of the query confirms the number of rows and shows that the data covers 1 through 105 months.

1.3 Line Coverage and Distribution

The dataset is reviewed to identify elements such as the total number of London Underground lines included, the distribution of records across lines, and observations repartition among lines. This step will ensure that performance comparisons across lines are statistically meaningful and not driven by uneven data availability.

The query returns 10 rows for each underground line and confirms that each of them have been covered for 105 months.

1.4 Missing Value Assessment

The critical fields required for analysis and KPI computations are checked for missing values. The fields concerned are "Line", "Month", and "Excess". Any missing values identified at this stage will inform cleaning decisions in the next phase.

The presence of missing values in secondary metadata fields is also noted but treated separately if they do not directly affect KPI calculations.

The query checks each column of critical fields. If the value is null, then it is counted as 1, otherwise it shows 0 to confirm the lack of missing value. The overall output shows 0 across all critical fields.

1.5 Numeric Range Validation

First, key numeric fields (Excess, Scheduled, and Total) are examined using minimum (0), maximum (22.25), and average (5.77) values as baseline expectations. This will help detect implausible values further down the line.

Then, scheduled journey time are compared to the actual ones and the output from the query shows the minimum scheduled time across all rows 7.8 with a total of 9.36. The scheduled time across rows equals 47.27 with the total maximum being 59.62.

Additionally, extreme Excess Journey Time values are identified by flagging observations that fall well outside the average range. The output shows similar metrics to that of the above-mentioned comparison between scheduled and actual time metrics.

Finally, the detection of anomalous excess values is queried but given that the output is perfectly aligned with that of the journey time comparison, it can be safely deduced that there are no anomalies.

1.6 Metadata Consistency

Line-level metadata such as “Length”, “Stations”, “Opened”, and “Type” are checked for consistency within each line. This ensures that descriptive attributes remain stable across months and can be reliably used during analysis and visualisation. The output shows 1 across all columns for each of the 10 lines concerned, further confirming the reliability of the metadata over time.

1.7 Audit Outcomes

The outcomes expected from the first phase of this project include confirming dataset grain and suitability for analysis, verifying overall temporal and line-level coverage, identifying the presence of extreme but valid operational values, and establishing confidence in metadata stability. These findings directly inform the next phase which focuses on data cleaning and feature engineering where derived analytical fields are introduced without altering the integrity of the source data.

Phase 2: Data Cleaning & Feature Engineering

The purpose of this phase is to prepare the London Underground performance dataset for KPI computation and comparative analysis. Based on the findings from Phase 1, transformation decisions are made to ensure analytical consistency, semantic clarity, and suitability for downstream aggregation and visualisation. This phase does not alter the fundamental grain of the dataset. However, it will enhance it with derived fields required to answer the research question.

2.1 Data Cleaning Scope and Principles

The previous phase's audit confirms that the dataset is largely complete and structurally consistent, with no missing values in critical analytical fields such as Line, Month, Scheduled, Excess, or Total. As a result, no row-level deletions or imputations are required at this stage. Cleaning efforts in this phase are therefore limited to ensuring logical consistency between numeric fields, handling derived fields with missing values, and standardising analytical definitions instead of modifying raw source values. This approach preserves the integrity of the original dataset while ensuring analytical reliability.

2.2 Validation of Performance Metrics

Excess Journey Time is retained as the primary performance indicator for the project. Phase 1 confirms that "Excess" represents the difference between scheduled and actual journey times and that "Total" aligns conceptually with "Scheduled" plus "Excess". The output of the validation query shows a minimum difference of -0.01 and a maximum difference of 0.01 between "Total" and "Scheduled" + "Excess". These near-zero differences indicate minor rounding effects rather than structural inconsistencies, confirming that Excess is consistently defined and can be safely reused as a delay metric without transformation.

Upon validating the performance metrics, a business-readable KPI field called "delay minutes" is created. The latter contains values directly copied from "Excess".

2.3 Peak and Off-Peak Period Classification

A peak-period indicator is retained in the dataset schema to support structural extensibility. However, the dataset is aggregated at a monthly, line-level grain and does not contain time-of-day or event-level information. As a result, reliable peak and off-peak classification cannot be derived without introducing unsupported assumptions.

To avoid false precision, all records are assigned a non-peak classification. This is done by assigning “non-peak” (0) to all rows where “is-peak” is null; existing values are not overwritten. The peak indicator is therefore preserved as a placeholder rather than an active analytical dimension and is not used in KPI computation. Overall line-level performance remains the sole focus of the analysis.

Following this assignment, a validation step is performed to confirm analytical completeness. A null check across derived fields verifies that both *delay minutes* and the peak indicator are fully populated, confirming that no gaps remain after feature engineering. A final analytical snapshot of the dataset is then reviewed to ensure that key analytical columns are correctly populated at the line-by-month level. Together, these checks confirm that the dataset is structurally sound and ready for KPI computation without requiring further cleaning or adjustment.

2.4 Phase Outcomes

By the end of Phase 2, the dataset remains unchanged at the source level while including the derived fields required for analysis. All analytical columns are fully populated, and unsupported segmentation has been avoided. As a result, the dataset is stable, consistent, and ready for KPI computation and line-level performance comparison in the next phase.

Phase 3: KPI Computation & Line Performance Ranking

The purpose of this phase is to compute performance Key Performance Indicators (KPIs) from the cleaned and feature-engineered dataset and to rank London Underground lines based on operational performance. This phase directly addresses the project's core research question by translating validated performance data into comparable, line-level metrics suitable for interpretation and visualisation.

3.1 KPI 1: Average Delay Per Line

The primary KPI for this analysis is the average delay per line, calculated as the mean of the "delay minutes" column, which is derived from Excess Journey Time. This metric captures the additional journey time experienced beyond scheduled expectations, with lower values indicating better operational performance. A secondary KPI "delay variability" is calculated as the standard deviation of "delay minutes" across months for each line and provides context on how consistent each line's performance is over time.

The output is a table that ranks all 10 lines based on average delay in minutes. The top-performing lines are Waterloo & City (2.06 minutes) and Bakerloo (5.05 minutes). They are well below the dataset's overall average delay of 5.77 minutes (from numeric range validation). The lowest-performing lines are Circle & Ham (7.17 minutes) and Metropolitan (8.55 minutes). They are considerably above the dataset mean, highlighting performance disparities across the network. This contextual comparison demonstrates not only relative performance between lines but also how each line performs against typical network-wide expectations.

3.2 KPI 2: Delay Variability

The second KPI focuses on delay variability, which measures how consistent each line's performance is over time by quantifying the month-to-month fluctuations in delays. Lines with low variability tend to have stable and predictable service, while lines with high variability experience more inconsistent performance. This is calculated by examining how much each month's delay differs from the line's average delay and summarizing these differences across all months. This produces a single variability value per line.

The output ranks all 10 lines by their delay variability. The lines with the most stable performance are Jubilee (0.80 minutes) and Bakerloo (0.66 minutes). They are well below the dataset's overall average delay variability of 1.17 minutes which is calculated across all lines.

Conversely, the most variable lines are Piccadilly (1.78 minutes) and Northern (2.12 minutes), indicating that their monthly delays fluctuate more dramatically compared to the network norm. This comparison allows for the contextualization of operational reliability alongside average performance. A line with a low average delay but high variability might still experience unpredictable service, whereas a slightly slower line with low variability may provide a more consistent passenger experience.

3.3 Tableau-Ready KPI Output

To support downstream visualisation and ensure a single source of truth, a Tableau-ready KPI view is created. This view consolidates all key line-level performance metrics into one virtual table. For each London Underground line, it reports the average delay in minutes across all months, the variability of delays to indicate how consistent performance is over time, and the total number of monthly observations used to compute these metrics. Grouping the dataset by line in this view allows for stable, ready-to-use performance summaries that can be queried directly in Tableau or other reporting tools. This approach ensures that averages and variability do not need to be recalculated for each visualisation, providing consistent and reproducible results across analyses.

3.4 Phase Outcomes

The queries executed in this phase produce clearly defined performance KPIs, including average delay and delay variability, along with a defensible ranking of all London Underground lines. These results are consolidated into a stable, Tableau-ready KPI view, providing a single source of truth for visualisation and further analysis. Together, these outputs offer a complete, data-driven answer to the project’s core research question, enabling network-wide performance comparisons, identification of consistently high- or low-performing lines, and the exploration of operational reliability patterns through dashboards and narrative interpretation.

REFERENCES

Gilks, 2019. “London Underground Performance Data”. Available at https://zenodo.org/records/3518376

Open Aire, 2025. “London Underground Performance Data”. *Summary.* Available at: https://explore.openaire.eu/search/result?pid=10.5281%2Fzenodo.3518376

Zenodo, 2025. “General Policies”. Available at: https://about.zenodo.org/policies/

London Assembly, 2025. “London’s Elizabeth line at three: data analysis”. Available at: https://www.london.gov.uk/who-we-are/what-london-assembly-does/london-assembly-research-unit-publications/londons-elizabeth-line-three-data-analysis