**Problem statement**

Forecast bus trip travel time (or delay delta per trip) using historical operations data and engineered features, so each **trip** gets an individual prediction and we can report an overall performance score suitable for operations.

**Methodology**

**Exploration → iterative feature engineering → validation.**

1. **Open exploration:** sanity checks, time-zone normalization (Europe/Amsterdam, DST-safe), date-windowing with a **history buffer**, early de-duplication, and merge validation. Checked basic seasonality (hour/day/week), lag structure, and demand proxies.
2. **Iterative features:** built time features (hour-of-week, Fourier sin/cos), **lags/rolling** of target, reliability indicators for missing lags, and a **time-weighted check-ins feature** that integrates hourly network demand across each trip’s actual interval (start + duration). Evaluated incremental utility each iteration.
3. **Models & pipelines (sklearn):**
   * **HistGradientBoostingRegressor** (point ETA).
4. **Validation:** time-aware splits (TimeSeriesSplit), headline metric **MAE (min)** with **MASE vs baseline** (e.g., timetable or naïve), plus P90 |e| and % within ±1/±2 min. Tracked micro (overall) and macro (by line×direction) views.

**Key findings & what I did with them**

* **Previous driving times by line×direction help.** → Kept grouped lags/rolling; used them as core predictors. Positive impact on model performance.
* **Network check-ins showed weak raw correlation (~0.18).** → Engineered **time-weighted** check-ins over the trip interval; kept as a feature with a coverage flag, expecting value once interactions/nonlinearity are learned.

**Proposed improvements**

* **Outlier handling:** replace crude cutoffs with **segment-wise quantile/MAD rules**; prefer **clipping (winsorizing)** learned on train to dropping, and consider **relative-error** thresholds with a minimum absolute error on short trips.
* **Demand dynamics:** add **lags/rolling of check-ins** (e.g., last 15/30/60 min) and peak/off-peak interactions.
* **Contextual data:** incorporate **location/geometry** (stop/corridor), **distance**, roadworks/incidents, and weather; move from network-wide to **line/stop-level demand** if available.
* **Quality & Cardinality:** stop dataset seems to have some issues (non-uniqueness, missing values). Furthermore, the stop id columns in bus\_trip dataset have a cardinality of N to N with stop id in stop data set. More robust pipeline needs to be built to cleanup the stop set. Potentially provide client with cleaned & enriched dataset.

**Questions answered**

* **Effect of covid on bus times**: we see fewer people checkin, while the mean difference between planned & realized increases. Ideally, this is something we want to account for in our ML pipeline. Either by correcting this dataset, adding a feature to indicate it was covid period, or remove it altogether.
* **Model validation**: we see marginal improvements with the gradient boosting model over the planned values. The model is now tailored to just 2 bus lines, which are extremely similar. I expect model performance to drop when we add in new bus lines with different travel times, locations etc.