# Case Study: Predicting Bus Travel Times

**Overview**
In this case study, you are tasked with building a model that predicts the travel time of bus journeys for a specific bus line. The goal is to forecast the time it takes for a bus to travel from first to final stop at any given point in the future. Your model (and analyses) will support bus providers (like EBS, RET & HTM) in scheduling their future bus trips.

You will be provided with three CSV files containing real-world data:

- bus_trips.csv
- stops.csv
- check_ins.csv

**Data Description**
*bus_trips.csv*
This file contains detailed data on realized and planned travel times for bus lines 1 & 2 from first to final stop. The variables included are:

- lineplanningnumber: The identifier used for the bus line by the provider (in this case “1” and “2”).
- journeynumber: A unique identifier for each journey along the bus line.
- vehiclenumber: The number of the vehicle operating the trip.
- userstopcode_start: The code representing the bus stop where the journey begins.
- userstopcode_end: The code representing the bus stop where the journey ends.
- messagetype_begin: The type of message or signal indicating the start of the journey.
- messagetype_end: The type of message or signal indicating the end of the journey.
- operatingday: The date on which the journey took place.
- departure_time: The actual departure time of the bus.
- realized_time: The actual travel time recorded for the journey.
- planned_time: The historical scheduled journey time planned by the bus provider (which, as you might know, is not necessarily accurate).

*stops.csv*
This file contains information about the bus stops referenced in the bus_trips data. The variables include:

- userstopcode: A unique code identifying each bus stop.
- userstopareacode: A code representing the geographical area of the stop.
- stopname: The name of the bus stop.

*check_ins.csv*
This file provides data on hourly public transport check-ins across The Netherlands. Note that the check-in numbers are scaled by a factor of 1,000 (e.g., a value of 200 corresponds to 200,000 nation-wide check-ins). The variables include:

- id: A composite identifier following the structure %year_%month_%day_%hour (for example, 2024_05_05_18 indicates May 5th, 2024, at 18:00).
- number_of_check_ins: The number of check-ins recorded at the specified hour on the given date.

**Task Objectives**

Using the provided datasets, your task is to develop a model that can predict bus travel times. You can then use this model to find answers to the following cases:

1. *Effect of COVID on Driving Time:*
   Analyze how the COVID pandemic has influenced bus driving times. What could be the reason for this effect?
2. *Future Prediction:*
   Predict the expected drive time for bus lines 1 & 2 for journeys departing at 18:00 on May 5th, 2024, when training a model on data until and including 2023. You are free in choosing which model(s) to use.
3. ***Bonus: Model Effectiveness Validation:***
   Validate the effectiveness of your predictive model by comparing its output against the provider's planned_time for all journeys throughout the year 2024 (when training up until and including 2023). Discuss any discrepancies and possible improvements.

**Deliverables**

- *Code Repository:*
  Submit your code base with all the necessary scripts to preprocess data, build models, and generate predictions.
- Final Report:
  Include a short report (<1 A4) that explains your methodology, analysis, results, and conclusions. Address each of the tackled task objectives.

**Tips**

- *Think of the physical problem:*
  When analyzing the data and developing your model, make sure to remember the physical problem. For example, bus lines usually go both ways.
- *Modelling Techniques:*
  You are free to use any appropriate techniques ranging from general statistical analysis to advanced machine learning models such as Regression, Gradient Boosting, or Time Series Forecasting. You are not required to use all the data and variables included in the datasets. Feel free to disregard certain identifiers from your model and analyses.
- *Feature Engineering:*
  Feel free to create your own variables to improve model performance. This might include binary variables to capture seasonality, trends, or other relevant factors.
- *Programming languages, packages and IDE:*
  You are allowed to use any programming language to solve this problem, but Python is recommended (and SQL might be useful if this is available to you). Feel free to use Jupyter Notebook for programming the different tasks (or a different IDE if you like). You are free in choosing which packages to use, but pandas, numpy, sklearn and matplotlib might be useful.
- *Documentation and Reproducibility:*
  Document your methodology, assumptions, and findings clearly. Your code should be well-organized and reproducible, with clear comments and a logical structure.

**We look forward to seeing your innovative solutions and insights into predicting bus travel times. Good luck!**