# Scenario

The use of smartcard ticketing systems in Ireland's public transportation is covered in the scenario. These devices record copious amounts of data that mirror commuter behavior, enabling a thorough evaluation of transportation requirements. The gathered datasets allow for precise study of travel patterns on an individual and group level, classifying travelers according to fare kinds such as senior citizens or students. Acquiring exact insights into public transportation user behavior is intended to enable service optimization, group tailoring, and general advances in transportation efficiency.

# Dataset Selection

## For Transport in Ireland

**Dataset:** TOA02 - Average weekly flow of Luas passengers

**Published by**: Transport Infrastructure Ireland

**Licensed under**: Creative Commons Attribution 4.0

**Category**: Government

## For another Country (Australia)

**Dataset:** Public Transport Services

**Published by**: Department for Transport

**Licensed under**: Creative Commons Attribution 3.0 Australia

**Category**: Government

# Data Preparation

To prepare raw data for analysis, it must be cleaned, transformed, and arranged. These covers encoding categorical variables, converting data types, handling outliers, and dealing with missing values. Data from various sources may be combined and duplicates eliminated. The dataset is improved by feature engineering, normalization, and aggregation; unbalanced data and skewed distributions are taken care of. Activities like lag generation and resampling can be done with time series data. For model evaluation, the dataset is frequently divided into training and testing sets, and the entire procedure is documented for transparency's sake. Accuracy, completeness, and relevance in ensuing analytical and modeling activities are guaranteed by efficient data preparation.

**Code Reasoning**

We preprocessed the Ireland dataset using a Python script as part of the data cleaning procedure. Starting with the tab ('\t') as the delimiter, we read the raw data from the given file location. We examined the dataset's metadata to determine its structure after putting the data into a panda Data Frame. We addressed missing values and eliminated rows that were duplicates to improve the quality of the data, guaranteeing a clean dataset for further research.

We gave the columns new names and more illustrative labels in an effort to increase uniformity and legibility. We also changed the 'Year' and 'VALUE' columns to numeric formats to fix any possible flaws or discrepancies in the original data. By substituting NaN for all non-numeric values, the 'to numeric' function with the 'errors' option set to 'coerce' made this conversion easier.

At last, we produced a summary of the Ireland dataset that had been cleaned, displaying the initial few rows. This data cleaning script provides a well-processed dataset for the project's next phases, laying the groundwork for additional investigation.