# 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.

## Data Optimization

## Code Validation and Assurance

In our analysis, it is crucial to ensure the precision and dependability of the code that has been applied. The following techniques were used to confirm and validate the code's integrity:

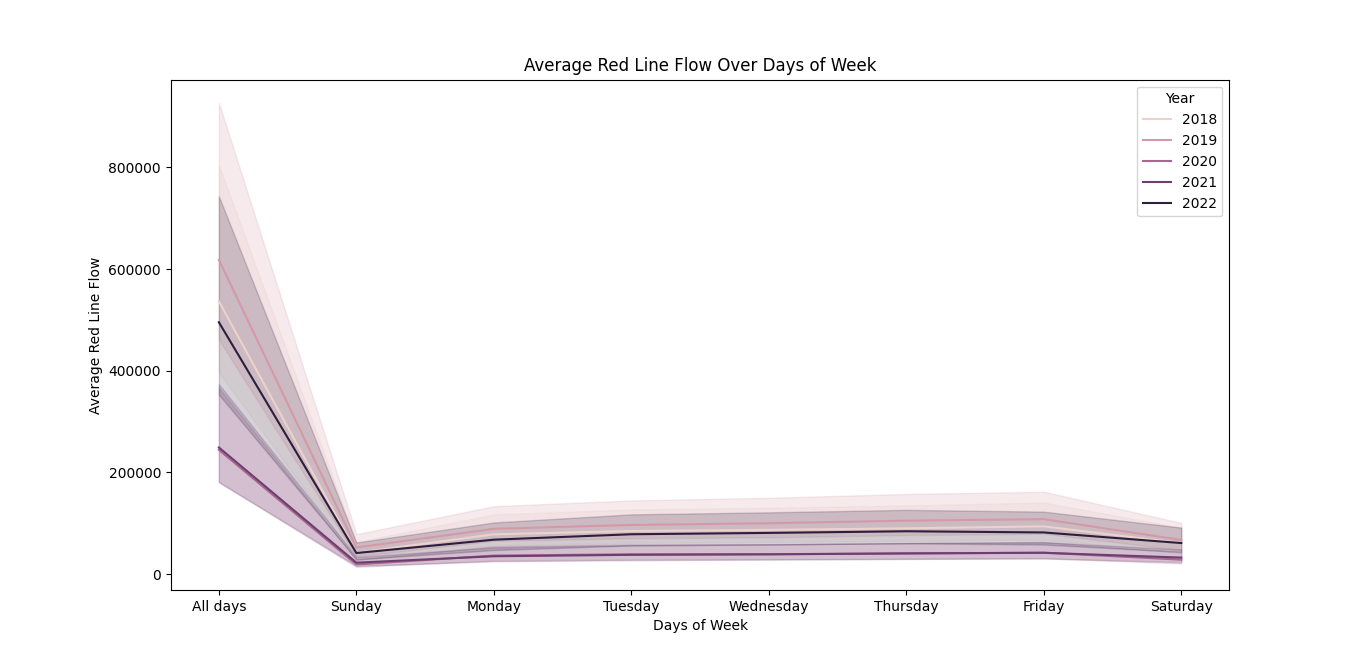
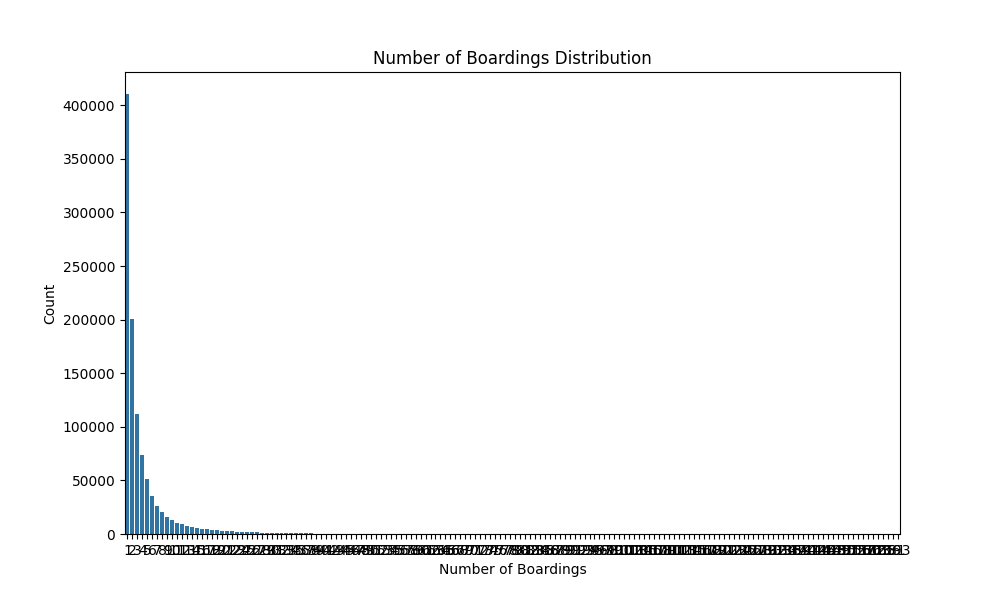
* **Unit Testing**: To ensure that every function and component in the codebase is correct, each one was thoroughly tested separately.
* **Integration Testing**: To guarantee smooth cooperation and adherence to the main goals of the analysis, the integration of numerous modules and components was carefully verified.
* **Data Consistency Checks**: Throughout the analysis pipeline, routine checks were carried out to guarantee the consistency and integrity of the dataset(s).

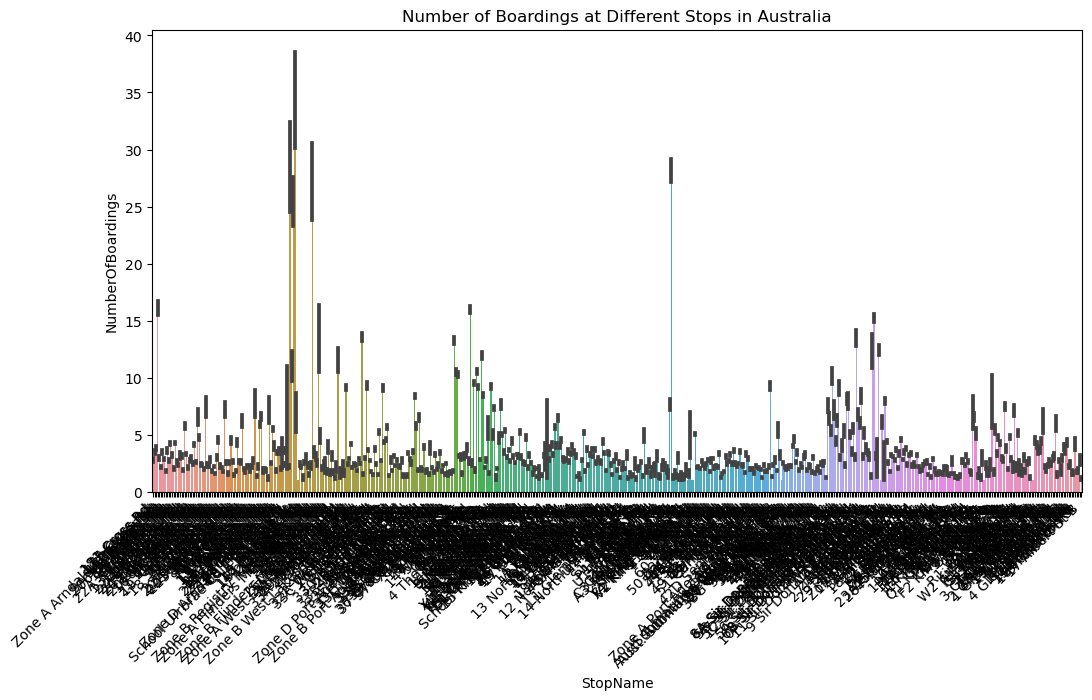
## Optimization and Resource Management

* Memory Management: To minimize memory usage, especially when managing big datasets, data was arranged and processed in segments.
* Parallel Processing: In order to maximize computer resources and speed up computations, parallel processing techniques were studied.
* Algorithmic Efficiency: Care was taken when choosing and using the mathematical techniques to guarantee computation efficiency and minimize temporal complexity.

## Data Visualization

Information perception is essential to extracting meaningful insights from datasets because it provides a graphical aid for understanding structures, communicating ends, and identifying discrepancies. It supports comparative analysis, provides general well-informed guidance, and facilitates the investigation of the interrelationships among its components. Perspectives enhance narration and aid in exploration investigation of information by revealing patterns and ephemeral examples when they create an argument surrounding the data. Information representation is a fundamental device for both specialized and non-specialized crowds, since it can make an interpretation of complicated data into effortlessly grasped experiences.





A graph showing different days of the week

Description automatically generated