# 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:

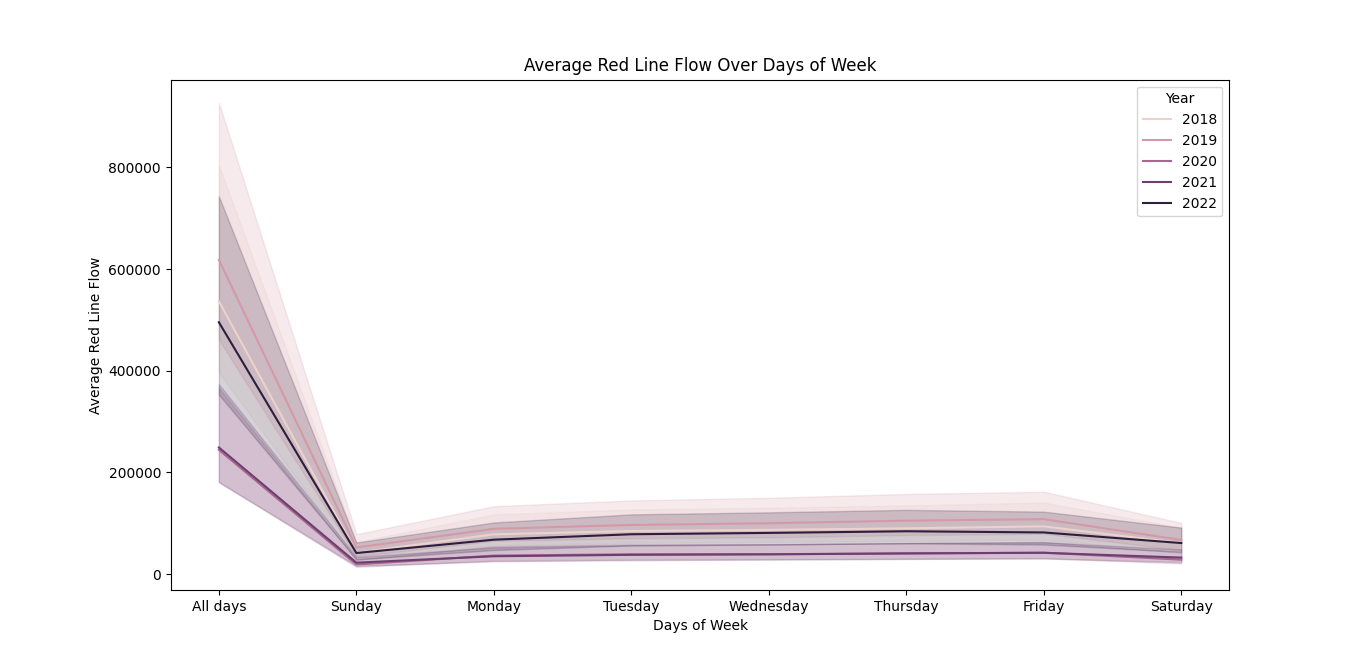
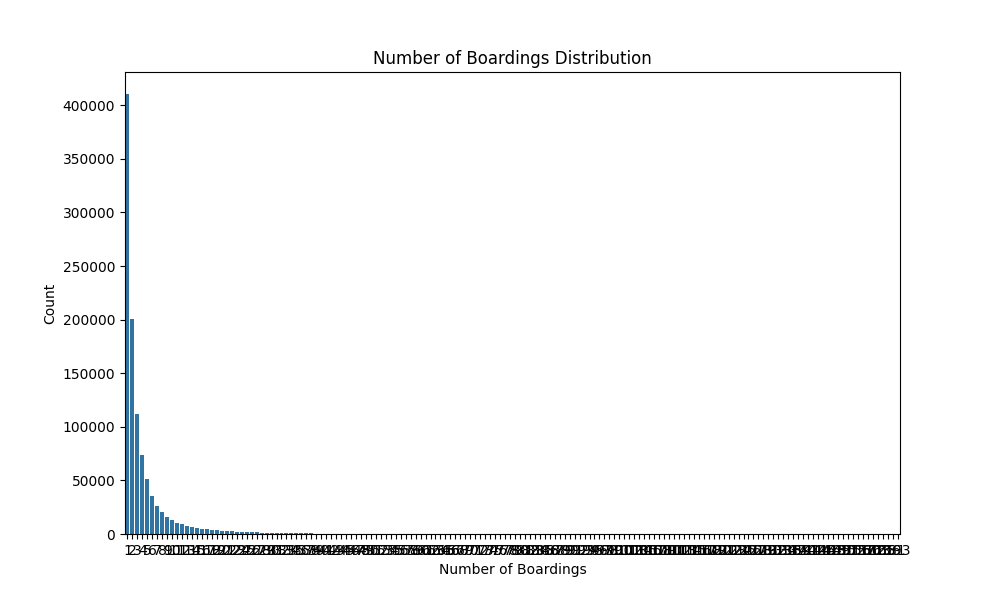
* **Unit Testing**: To ensure that every function and component in the codebase is correct, each one was thoroughly tested separately.
* **Integration Testing**: To guarantee smooth cooperation and adherence to the main goals of the analysis, the integration of numerous modules and components was carefully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

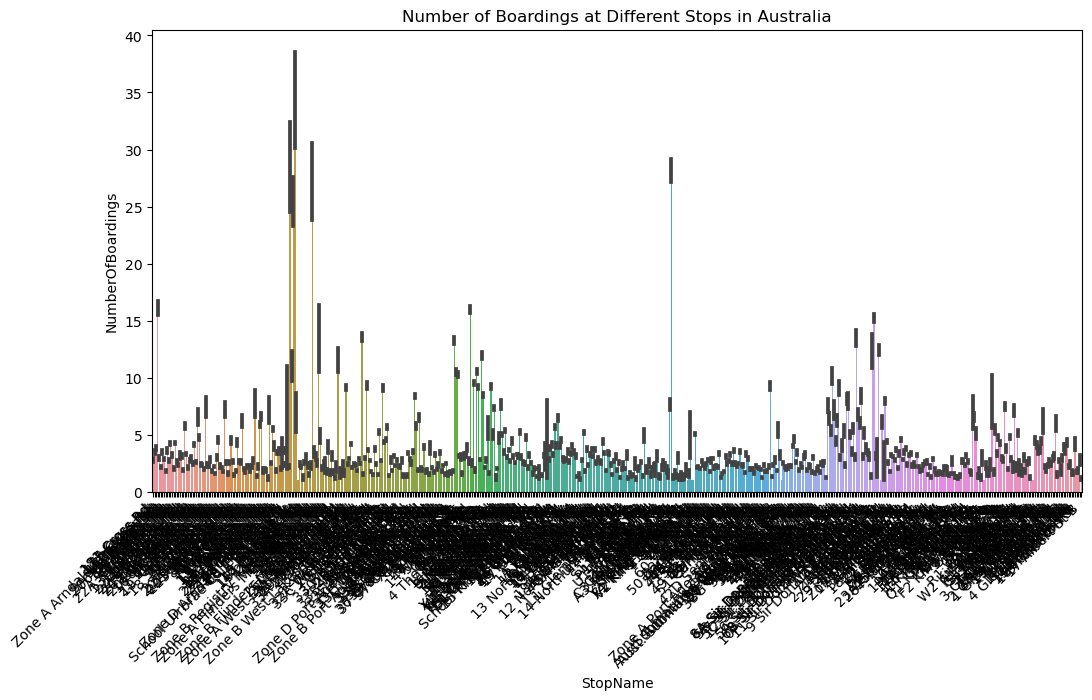
## Optimization and Resource Management

* 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

## Data Manipulation

|  |  |
| --- | --- |
| **Data Processing Libraries** | |
| **Pandas** | **NumPy** |
| **Tabular Data Handling**  With a significant spotlight on tabular data structures, Pandas is a vigorous Python tool stash for information control and examination.  Filtering, grouping, and merging are only a couple of the confounded information tasks made simpler by the major Data Frame object. Provides improved readability and usability for jobs requiring organized data. | **Numerical Operations**  Supporting massive, multi-dimensional arrays and matrices, NumPy is a foundational library for numerical computing.  Provides a large number of mathematical operations and functions, especially useful for computations involving numerical arrays.  Effective management of numerical data, which is necessary for a number of mathematical and statistical investigations. |
| **Structured Aggregation**  The group by function in Pandas is essential for organizing data into groups according to predetermined standards, allowing for further actions on these groups.  Makes it easier for groups to filter, transform, and aggregate data.  Offers a high degree of group-wise operating flexibility, improving the capacity to extract significant insights. | **Aggregation Functions**  The fundamental aggregating functions like sum, mean, min, max, and so on are included in NumPy.  Makes it simple to compute summary statistics for numerical data arrays.  Effective and succinct for simple aggregate requirements; especially well-suited for numerical summarization. |

# Statistics for Data Analytics Tasks

The dataset(s) under review comprise an extensive set of data relevant to transport datasets. These databases, which contain comprehensive records and insights into a variety of transportation patterns are extremely valuable resources.

There are several reasons to perform statistical analysis on these datasets. A useful toolkit for extracting significant patterns, trends, and insights from unprocessed data is the statistical approach. Using descriptive statistics and representations, we want to introduce a succinct outline of the dataset(s), featuring significant factors and their traits.

Moreover, the objective of acquiring a more profound comprehension of the populace esteems that underlie the noticed information drives the utilization of inferential measurements. To help information driven independent direction and empower a nuanced perspective on the bigger climate, this involves exploring certainty stretches for relevant factors.

Moreover, the correlation review with other country means to distinguish shared traits and contrasts, using different measurable tests to recognize designs that upgrade perception of the dataset(s). A devotion to factual meticulousness persuades the insightful choice of these tests, ensuring the legitimacy and significance of our decisions.

* **Descriptive Statistics and Appropriate Visualizations**

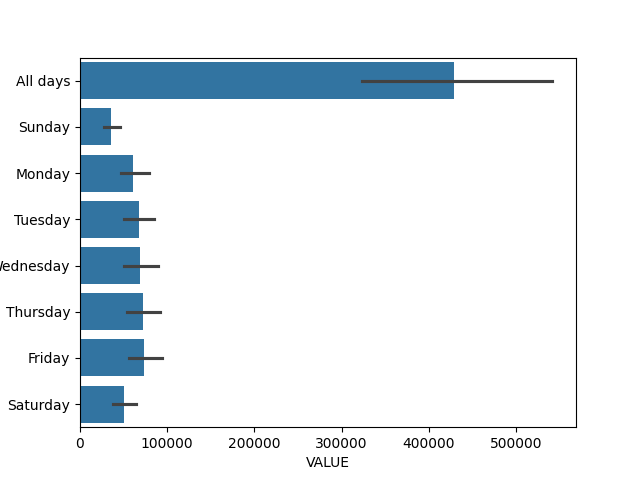
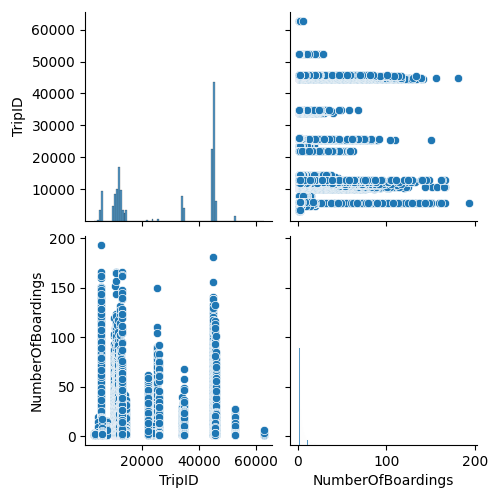


Figure 2: Appropriate visualization

The descriptive statistics and representations shed light on the elements of the datasets from Australia and Ireland.

We see the summary statistics for factors like "Trip ID," "Route ID," and "Number Of Boardings" for the Australian dataset. A careful comprehension of the relationships between these factors are given by the pair plot show. While demonstrating the utilization of public transportation, the dissemination of "Number Of Boardings" may be vital, and the connection among it and "Trip ID" and "Route ID" can assist with picking the best model. The average flow (also called "VALUE") for every day is shown in the bar plot depiction. Understanding the traffic patterns on various days with the use of this information can help choose the right models for traffic forecasting.

The particular analysis aims would determine which models were selected for each of the two datasets. Regression models could be investigated for the Australian dataset in order to forecast the number of boardings according to trip and route variables. Time series models could be useful for capturing the daily trends in traffic flow for the Ireland dataset. By offering a clear knowledge of the data distribution and correlations between variables, the statistics and visualizations aid in the justification of these decisions.

* **Analyze the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values.**

To perform inferential statistics on the datasets, we need a specific hypothesis or question to address.

|  |  |
| --- | --- |
| **Australian Dataset** | **Ireland Dataset** |
| **Hypothesis**: Is there a significant difference in the average number of boardings on different routes? | **Hypothesis**: Is there a significant difference in the average traffic flow ('VALUE') between different days of the week? |
| **Approach**: We can use a one-way ANOVA test to compare the means of 'Number Of Boardings' for different routes. If the p-value is significant, it indicates that there is a significant difference in the average number of boardings between at least two routes. Post-hoc tests can identify which routes differ. | **Approach**: We can use a repeated measures ANOVA or a Friedman test (non-parametric alternative) to assess whether there are significant differences in the average flow across different days of the week. Post-hoc tests can identify specific days that differ. |

|  |  |
| --- | --- |
| **Variables** | |
| **Australia Dataset** | **Ireland Dataset** |
| Route ID | Days of Week |
| Number Of Boardings | VALUE |

A screenshot of a graph

Description automatically generated