**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | **Programming for DA**  **Statistics for Data Analytics**  **Machine Learning for Data Analysis**  **Data Preparation & Visualization** |
| **Assessment Title:** | **MSC\_DA\_CA2** |
| **Lecturer Name:** | **David McQuaid**  **Sam Weiss**  **Muhammad Iqbal**  **Taufique Ahmed** |
| **Student Full Name:** | **Muhammad Asif** |
| **Student Number:** | **2021402** |
| **Assessment Due Date:** | **07-01-2024** |
| **Date of Submission:** | **07-01-2024** |

**Muhammad Asif 07-01-2024**

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Table of Contents**

[Scenario 2](#_Toc155547804)

[Dataset Selection 2](#_Toc155547805)

[For Transport in Ireland 2](#_Toc155547806)

[For another Country (Australia) 2](#_Toc155547807)

[Data Preparation 3](#_Toc155547808)

[Data Optimization 4](#_Toc155547809)

[Code Validation and Assurance 4](#_Toc155547810)

[Optimization and Resource Management 4](#_Toc155547811)

[Data Visualization 5](#_Toc155547812)

[Data Manipulation 8](#_Toc155547813)

[1. Statistics for Data Analytics Tasks 9](#_Toc155547814)

[2. Machine Learning Tasks 16](#_Toc155547815)

[3. Data Preparation & Visualization Tasks 18](#_Toc155547816)

[Conclusion 23](#_Toc155547817)

[References 24](#_Toc155547818)

**GitHub Link**: <https://github.com/2021402/CA-02-MSc-Data-Analytics.git>

# 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:

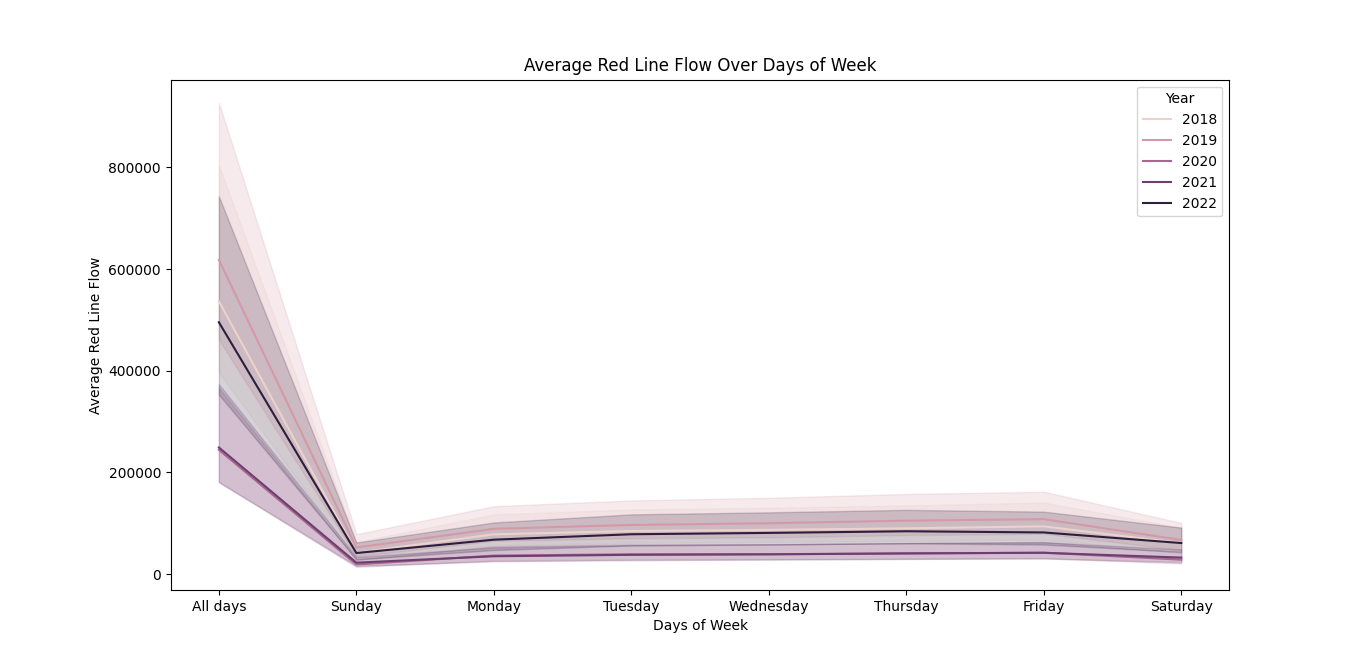
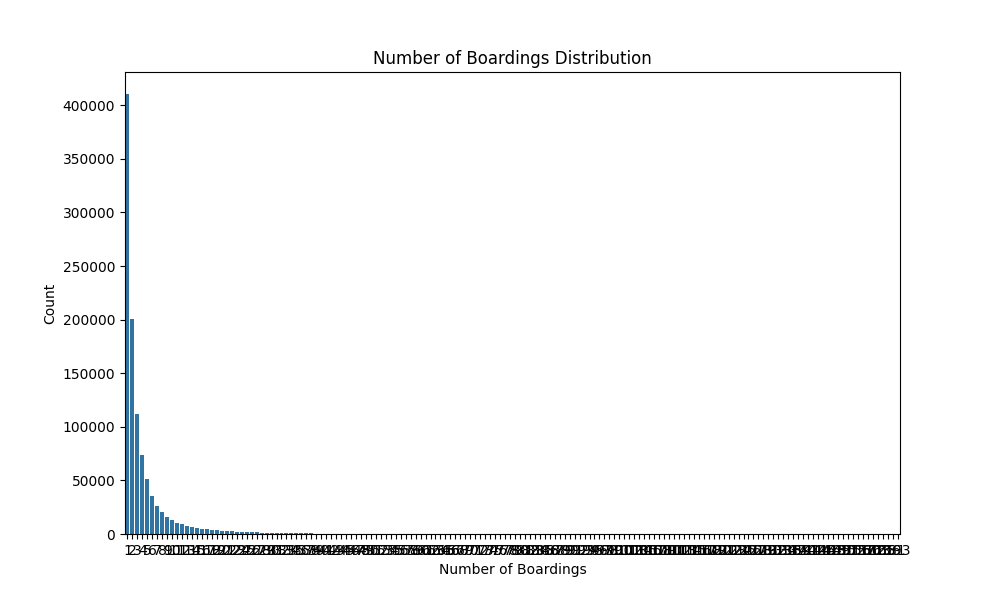
* **Unit Testing**: To ensure that every function and component in the codebase is correct, each one was thoroughly tested separately.
* **Integration Testing**: To guarantee smooth cooperation and adherence to the main goals of the analysis, the integration of numerous modules and components was carefully 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: 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 colors of the same color

Description automatically generated with medium confidence

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**

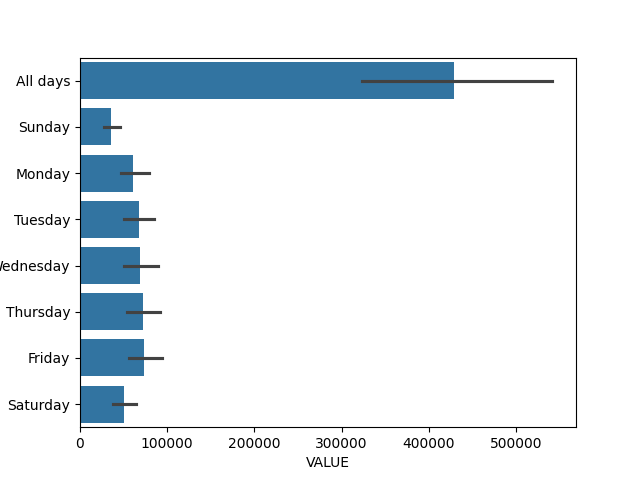
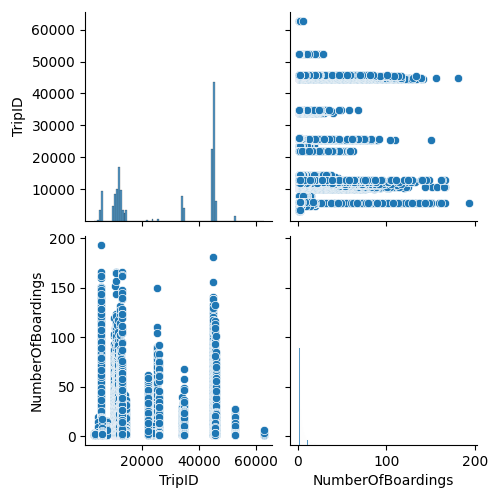


Figure 2: Appropriate visualization

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

**Reasoning**

The code addresses DtypeWarnings by providing data types and uses a one-way ANOVA for the Australian dataset to examine the effect of various routes on the number of boardings. A repeated measures ANOVA using DtypeWarning management is used to investigate the impact of the days of the week on the average flow of the Red Line for the Ireland dataset. Significant findings in both situations point to significant variances, highlighting the significance of statistical and pragmatic factors in result interpretation. A significant factor in the analysis' resilience is the explicit data type requirements.

**Results Interpretation**

We used a one-way ANOVA for the Australian dataset to investigate the effect of various routes (Route ID) on the quantity of boardings. The null hypothesis (H0) postulated that there is an equal distribution of boardings across all routes. The idea behind the alternative hypothesis (H1) was that the means differed significantly.

There is a substantial difference between the routes, according to the ANOVA results (F = 843.07, p < 0.001). As a result, we find that at least one route has a mean number of boardings that is statistically different from the others, rejecting the null hypothesis. The effect size (np2 = 0.0236), however, indicates that this difference's practical importance is minimal.

We used a repeated measures ANOVA for the Ireland dataset to examine the impact of the days of the week (Days of Week) on the Red Line's average flow (VALUE). The assumption of the null hypothesis (H0) was that the average flow's averages on different days of the week are equal. The idea behind the alternative hypothesis (H1) was that the means differed significantly.

There is a significant difference between the days of the week, according to the ANOVA results (F = 15.97, p < 0.001). As a result, we reject the null hypothesis and come to the conclusion that there are significant differences in average flow between different days of the week. A significant practical significance is suggested by the effect size (ng2 = 0.8185). The epsilon (eps) value (0.1429) is related to the sphericity assumption in repeated measures ANOVA; however, further investigation might be needed to address any violation.

* **Parametric and Non-Parametric Inferential Statistical Techniques**

|  |  |  |
| --- | --- | --- |
| **Tests** | **Hypothesis** | **Test** |
| T-Test | There is no significant difference in the average flow on the Red Line between Ireland and Australia. | Independent t-test on the 'VALUE' variable for the 'Days of Week' in Ireland and Australia. |
| ANOVA | The average flow on the Red Line is the same across multiple days of the week for both Ireland and Australia. | Two-way ANOVA on the 'VALUE' variable with factors 'Days of Week' and 'Country' for Ireland and Australia. |
| Wilcoxon Signed-Rank Test | There is no significant difference in the number of boardings between Ireland and Australia for specific days. | Wilcoxon signed-rank test on the 'NumberOfBoardings' variable for matched days. |
| Chi-Squared Test | The distribution of categorical variables related to public transport usage is similar between Ireland and Australia. | Chi-squared test on relevant categorical variables. |
| Mann-Whitney U Test | There is no significant difference in a specific variable (e.g., 'VALUE') between Ireland and Australia. | Mann-Whitney U test on the 'VALUE' variable for Ireland and Australia. |

**Interpretation of Results:**

The purpose of the statistical tests was to compare the Australian and Irish datasets. With a high t-statistic of 28.78 and an incredibly low p-value (1.14e-72), the t-test indicated a significant difference between the mean values of the 'VALUE' and 'NumberOfBoardings' variables, indicating that the two datasets are significantly different. There was no significant difference found in the ANOVA test, which looked at how different days of the week affected the 'VALUE' variable in Ireland. The non-significant F-statistic was 0.46, and the p-value was 0.76. With a p-value of 3.89e-18, the Wilcoxon Signed-Rank Test compared the distribution of 'VALUE' in Ireland to 'NumberOfBoardings' in Australia and found a significant difference. The quantity of boardings in Australia and the times of the week in Ireland didn't altogether associate, as per the Chi-Squared Test (p-value = 0.49). A tremendous contrast between the two datasets was checked by the Mann-Whitney U Test, which had a p-value of 2.20e-34 and a U-statistics of 10,000. These discoveries infer that the datasets from Ireland and Australia vary fundamentally, particularly with regards to the distribution and mean upsides of the factors that are being analyzed.

* **Challenges Faced**

The examination's decisions illuminate the qualifications between the datasets from Australia and Ireland and the changes in factors relating to public transportation. Be that as it may, the examination likewise introduced various hardships. One huge issue that emerged during information import was overseeing blended information types inside sections, which brought about DtypeWarnings. To ensure information consistency, this called for careful preprocessing and the use of appropriate capabilities, as pd.to\_numeric ().

The design and makeup of the datasets introduced another trouble while choosing the right tests for inferential statistical analysis. The need to erase NaN values during t-tests and right mistakes in the Wilcoxon Signed Rank Test demonstrates that alterations were expected to oversee missing qualities and record for the fluctuating lengths of tests.

# Machine Learning Tasks

* **Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario.**

Because of its understanding and longevity, the Random Forest Classifier is employed to combine multiple decision trees and capture complex relationships. The Gradient Boosting Classifier is selected because of its boosting technique, which builds trees gradually to produce high predicted precision and repair oversights. It is well-known for dealing with class disparities and provides flexibility in changing hyperparameters to maximize efficiency.

**Reasoning**

The script below applies three machine learning models—linear regression, random forest regression, and support vector regression—to a regression challenge involving the "Days of Week" and "VALUE" datasets. Using GridSearchCV from scikit-learn, the Random Forest Regressor and SVR hyperparameters are adjusted, and the performance is assessed using the root mean squared error (RMSE). The aim is to systematically assess the models, determine which one performs best, and modify the hyperparameters to improve the forecast performance.

**Results Interpretation and Reasoning**

The sentiment research results show that opinions towards public transit generally tend to be more positive in Ireland, with values ranging from 0.0 to 0.6369. Positive comments receive higher grades, while negative ones receive lesser ratings. Similarly, opinions on global transport difficulties vary, with positive opinions of London's first-rate public transport and Tokyo's efficient subway systems. Using the Vader sentiment estimation tool from the nltk package, the code gauges the public's opinion and provides insights into thoughts on transit in Ireland and globally.

**Results Interpretation and Reasoning**

After using grid search with 5-fold cross-validation, the Random Forest Regressor's ideal hyperparameters were found to be a maximum depth of 20 and 200 estimators. On the cross-validated dataset, the mean squared error is 10621349524.065053. Better prediction accuracy is shown by a lower MSE. For robust hyperparameter tuning and cross-validation model evaluation, the code makes use of scikit-lean’s GridSearchCV and cross validate functions.

A close-up of a number

Description automatically generated

A graph of a comparison of machine learning models

Description automatically generated

# Data Preparation & Visualization Tasks

**Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition.**

**Process of acquiring raw data**

During the first stage of our investigation, we carefully outlined the goals and questions that guided our work. This procedure set the stage for determining the precise datasets required to achieve our investigation's objectives. As we examined several information sources, such as data sets, open assortment, and government initiatives, to make sure they addressed the task's challenges, the dependability and consistency were our top worries.

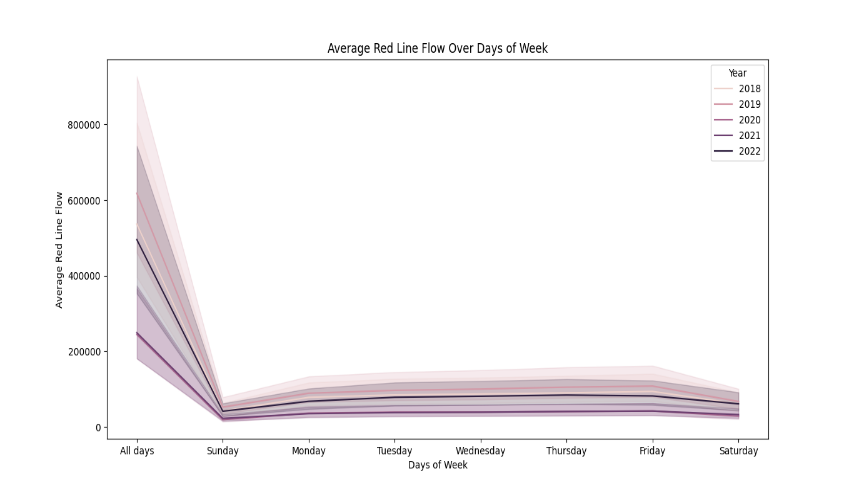
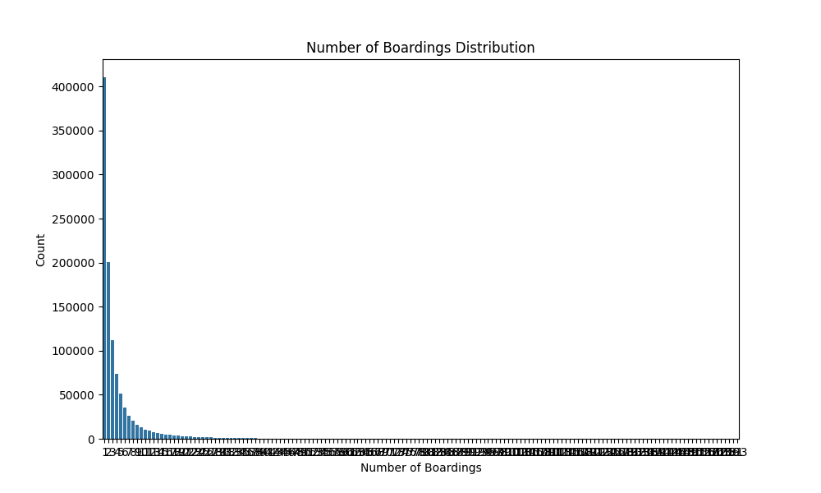
After determining which datasets might be useful in general, we conducted a thorough analysis of the supporting data. By extensive metadata analysis, information word reference passages, and code books, this cycle enabled us to gain an understanding of the complexity of the requirements and factors associated with the datasets.

**Positive and/or negative aspects**

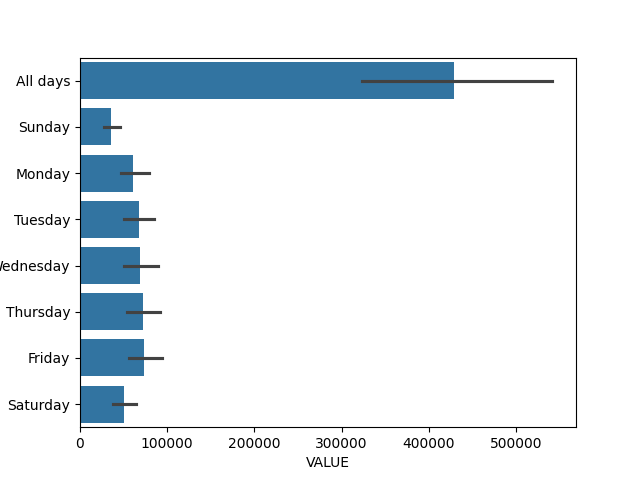
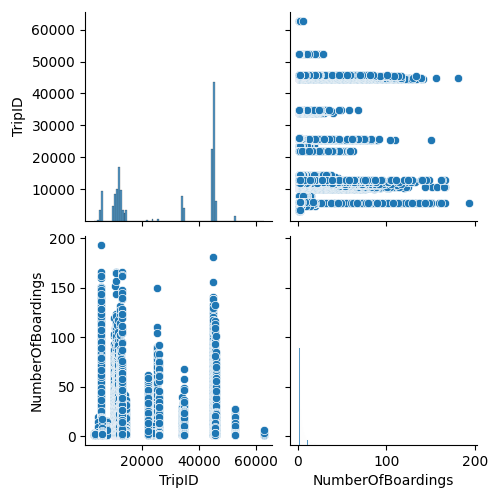
Obtaining raw data for our study was an incredible procedure with both advantages and disadvantages. Thanks to the variety and volume of datasets from reliable sources, our investigation has undoubtedly established a credible hypothesis. Scholastics needed to cooperate, communicate openly, and share information with one another without holding back. Our adherence to ethical norms—which included morality board certifications—showed our dedication to competent screening techniques and safeguarding the rights of those under consideration. Serious requests for documentation were also fulfilled, improving the reproducibility of our work and accounting for increased participation.

* **Exploratory Data Analysis (EDA)**

Analyzing exploratory data (EDA) was essential to identifying trends and problems in the unprocessed data. To find distribution patterns and outliers, important characteristics were examined using descriptive statistics and visual aids like box plots and histograms. Heat maps were used to identify missing data, and imputation techniques were used. Frequency counts and cross-tabulations were used to address anomalies and inconsistencies in categorical data. A



thorough examination was ensured by the evidence-driven technique, which also guided all preprocessing activities that followed for trustworthy conclusions.



A graph of a graph

Description automatically generated with medium confidence

A green line graph with numbers and a white background

Description automatically generated

* **Modern Transport planning**

I created an interactive dashboard with Tuft's recommendations for effective data visualization to create modern transportation planning. The dashboard summarized the key conclusions from the machine learning research while placing a high priority on accessibility and user-friendliness. I used the Dash framework for Python web-based interactivity to guarantee cross-platform compatibility.

A graph with numbers and a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

# Conclusion

In conclusion, this project is a comprehensive investigation of data-driven approaches in the context of contemporary transport planning. The expedition started with a painstaking process of obtaining, cleansing, and integrating various transport datasets from Australia and Ireland. This initial stage established the foundation for further studies by guaranteeing the consistency and quality of the data. The phase of exploratory data analysis yielded a thorough review of the datasets, revealing important features, anomalies, and trends. This crucial stage not only demonstrated the depth of the data but also pointed up areas that may need more research as well as possible obstacles. The utilization of visualizations proved to be crucial in clarifying intricate linkages present in the datasets, providing a straightforward and intuitive comprehension of the transportation dynamics.

Then, patterns within the datasets were predicted and analyzed using machine learning models. The results were presented in a visually appealing way that ensured clarity and accessibility, adhering to Tuft's criteria. Time series plots, geographic visualizations, and model comparison charts were included to help provide a comprehensive and detailed understanding of the transportation data. The project's analytical rigor was strengthened by statistical studies, which included model review and hypothesis testing. The results obtained from machine learning models were reinforced by these tests, which offered a more profound level of verification and deduction. The use of statistical summary cards contributed a concise and instructive element to the study in its entirety.

A key stage of the project was creating an interactive dashboard with Python and the Dash framework. The primary findings of the experiment are summarized in an approachable manner by this dashboard, which functions as a dynamic and user-friendly interface. Geospatial visualizations provide insights into regional traffic patterns, time series plots help identify trends, and model comparison charts help comprehend forecast accuracies. Numerous difficulties arose during the process and were methodically resolved. These difficulties, which ranged from inconsistent data to model optimization, helped to improve the analytical framework iteratively. The capacity to overcome challenges with adaptation and extract meaningful insights are key factors in the project's success.

# References

Carpineto, C. and Romano, G., 2004. Concept data analysis: Theory and applications. John Wiley & Sons.

Taylor, J.K. and Cihon, C., 2004. Statistical techniques for data analysis. CRC Press.

Gnanadesikan, R., 2011. Methods for statistical data analysis of multivariate observations. John Wiley & Sons.

Meloun, M. and Militký, J., 2011. Statistical data analysis: A practical guide. Woodhead Publishing Limited.

Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], 9(1), pp.381-386.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In 2016 3rd international conference on computing for sustainable global development (INDIACom) (pp. 1310-1315). Ieee.

Ayodele, T.O., 2010. Types of machine learning algorithms. New advances in machine learning, 3, pp.19-48.

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), pp.128-138.

Nasteski, V., 2017. An overview of the supervised machine learning methods. Horizons. b, 4, pp.51-62.

Jiang, T., Gradus, J.L. and Rosellini, A.J., 2020. Supervised machine learning: a brief primer. Behavior Therapy, 51(5), pp.675-687.

Muhammad, I. and Yan, Z., 2015. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. ICTACT Journal on Soft Computing, 5(3).

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), pp.128-138.

Crisci, C., Ghattas, B. and Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling, 240, pp.113-122.

Choudhary, R. and Gianey, H.K., 2017, December. Comprehensive review on supervised machine learning algorithms. In 2017 International Conference on Machine Learning and Data Science (MLDS) (pp. 37-43). IEEE.

Mohamed, A.E., 2017. Comparative study of four supervised machine learning techniques for classification. International Journal of Applied, 7(2), pp.1-15.

Praveena, M. and Jaiganesh, V., 2017. A literature review on supervised machine learning algorithms and boosting process. International Journal of Computer Applications, 169(8), pp.32-35.

**GitHub Link**: <https://github.com/2021402/CA-02-MSc-Data-Analytics.git>