**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation  Machine Learning for Data Analysis  Programming for Data Analytics  Statistics for Data Analytics |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name:** | **David McQuaid**  *Muhammad Iqbal*  *Sam Weiss*  *Taufique Ahmed* |
| **Student Full Name:** | Syed Asad Ailia |
| **Student Number:** | 2023408 |
| **Assessment Due Date:** | 05th January,2024 |
| **Date of Submission:** | 05th January,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. |

Author: Syed Asad Ailia

e-mail: [2023408@student.cct.ie](mailto:2023408@student.cct.ie)

Student ID: 2023408

GitHub Link: <https://github.com/SyedAsadAilia110/CA2.git>

**Data Selection**

**Ireland’s Dataset**

Dataset: THA22 - Average weekly volume of heavy goods vehicles for selected traffic count sites

Published by: Transport Infrastructure Ireland

Licensed under: Creative Commons Attribution 4.0

Category: Transport

Source: (<https://data.gov.ie/dataset/tha22-average-weekly-volume-of-heavy-goods-vehicles-for-selected-traffic-count-sites>)

**Another Country Dataset (Switzerland)**

Dataset: Public Transport in Zurich

Usability: 8.53

License: CC0: Public Domain

Collaborators: LAdams (Owner)

Source: (<https://www.kaggle.com/datasets/laa283/zurich-public-transport>)

**Table of Contents**

[**Abstract** 4](#_Toc155312564)

[**Introduction** 4](#_Toc155312565)

[Introductionto the Scenario 4](#_Toc155312566)

[**Programming for Data Analytics Tasks** 5](#_Toc155312567)

[Briefing of Code 5](#_Toc155312568)

[Data Manipulation 7](#_Toc155312569)

[**Statistics for Data Analytics Tasks** 7](#_Toc155312570)

[Descriptive Statistics and Visualizations 7](#_Toc155312571)

[Briefing and Result Interpretation 8](#_Toc155312572)

[Inferential Statistics for Population Insights 9](#_Toc155312573)

[Briefing and Result Interpretation 11](#_Toc155312574)

[Cross-Country Statistical Comparisons 11](#_Toc155312575)

[Verification and Applicability 12](#_Toc155312576)

[Challenges and Insights from Statistical Analysis 13](#_Toc155312577)

[**Machine Learning Tasks** 13](#_Toc155312578)

[Choose appropriate features from the datasets and a target feature 14](#_Toc155312579)

[Data Collection 14](#_Toc155312580)

[Sentimental Analysis 14](#_Toc155312581)

[Results Interpretation 15](#_Toc155312582)

[Reasoning & Clustering Methods Applied 15](#_Toc155312583)

[Result Interpretation 18](#_Toc155312584)

[Discussion and Elaboration 18](#_Toc155312585)

[**Data Preparation & Visualisation Tasks** 18](#_Toc155312586)

[Discuss in detail the process of acquiring your raw data 18](#_Toc155312587)

[Detailing the positive and/or negative aspects of your research and acquisition 19](#_Toc155312588)

[EDA (Exploratory Data Analysis) 19](#_Toc155312589)

[**Modern Transport Planning** 22](#_Toc155312590)

[Rationale for approach and visualization choices made during development 23](#_Toc155312591)

[**Conclusion** 23](#_Toc155312592)

[**References** 24](#_Toc155312593)

**Transport in Ireland (Comparison Between Ireland Transport Data V/S Switzerland Transport Data)**

# **Abstract**

*In today's data-driven world, data analytics is becoming more and more significant, having a major impact on many facets of business, science, and society. In this report, we used data analytics techniques to clean insightful information from the Two Different datasets which we take from their official websites which is between Ireland Transport and Switzerland Transport dataset. We carefully imported and checked the information to respond to specific questions. The development of a machine learning model for extracting output parameters from the validation dataset was the final step in this extensive process, which began with data preparation and continued with graphical representation using statistical techniques to identify trends. The Python framework was used to fulfil the programming requirements, and the entire project was recorded in a Jupyter Notebook as per the given instructions.*

# **Introduction**

This research investigates statistical analysis of data with an emphasis on the transport sector in Ireland. Our objective is to offer a thorough global examination and contrast of transport patterns using the copious amounts of data generated by smartcard ticketing systems. With a focus on freight transport, air traffic, car traffic, and facilities, the project analyses many datasets to provide a comprehensive understanding of Ireland's transport landscape and provide insights based on data.

With a focus on real-world execution, the project places a high priority on scientific rigor, open records, and effective interaction. A thorough examination of the transportation information area is supported by the tasks listed, which include machine learning programmes, statistical analysis, analysis of information programming, and data processing and visualisation.

# Introductionto the Scenario

In the age of smartcard ticketing, data turns become a driving force for improving public transportation. In this scenario, judgments are made and services are improved by examining Ireland's transportation data. Forecasting, sentiment analysis, and cross-national comparisons are all part of the challenge, which calls for a comprehensive strategy that combines machine learning, programming, statistics, and sophisticated visualization. Finding insights that will inform strategic recommendations for the ever-changing urban transportation context is the aim.

# **Programming for Data Analytics Tasks**

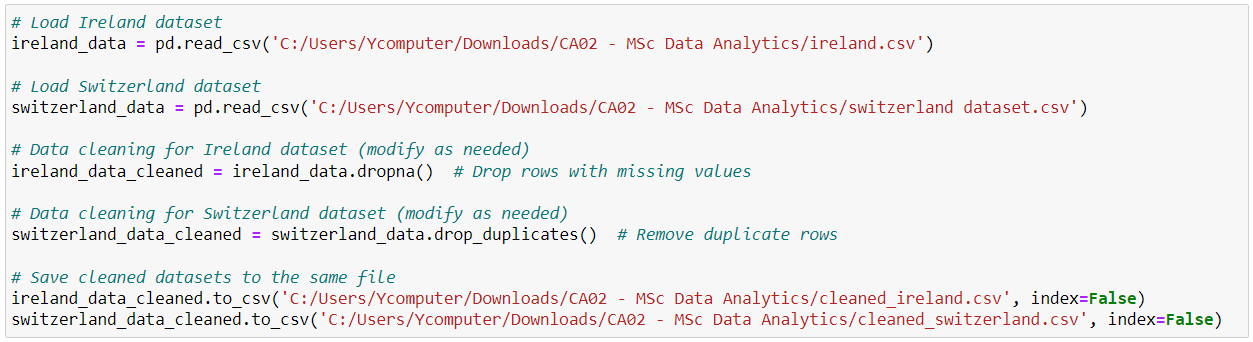


Figure 01. Dataset Load and Cleaning

# Briefing of Code

In this code, we did data cleaning on the "Ireland" and "Switzerland" datasets using the pandas module in Python. We note that the input datasets file paths are provided and that the datasets are loaded into data frames called ireland\_data and switzerland\_data using the pandas read\_csv function.

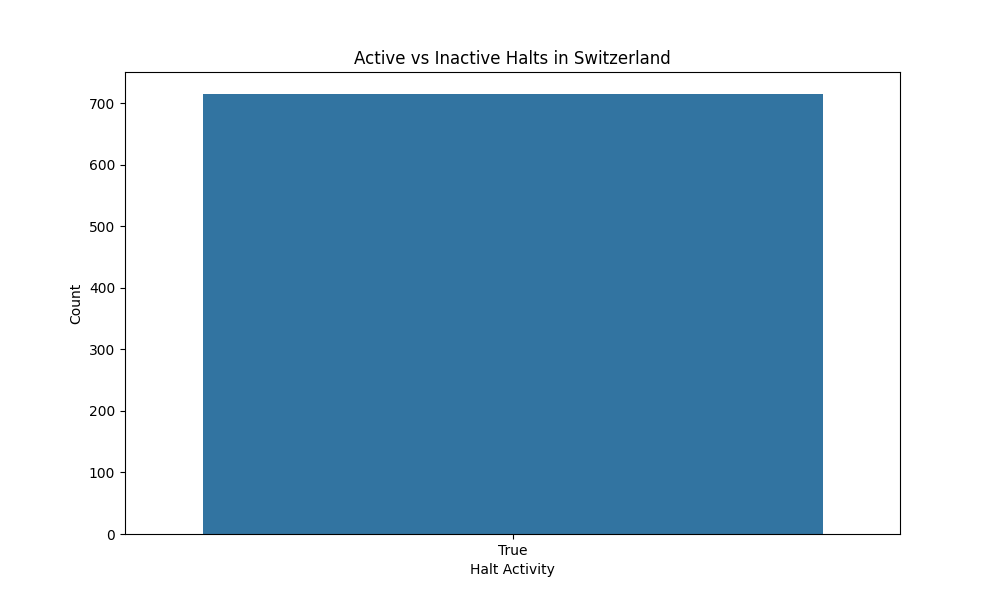
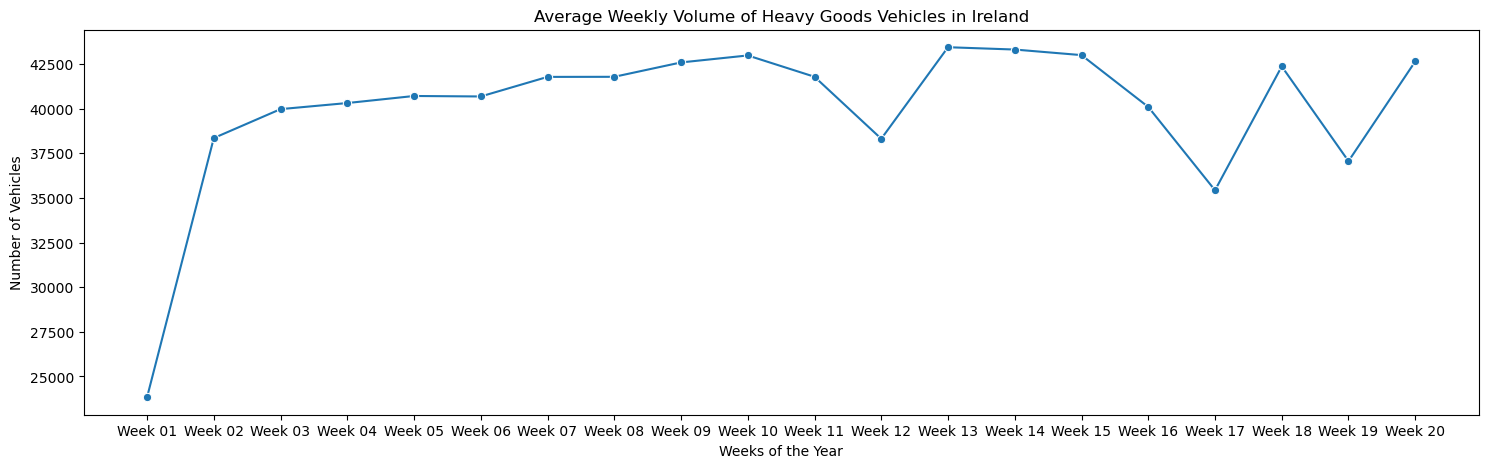
After the datasets are loaded, we handled missing values and eliminate duplicate rows to clean up the data. The pandas drop\_duplicates() and dropna() methods are used for this. The variables cleaned\_ireland and cleaned\_switzerland hold the cleaned data frames.

Figure 02. Data Visualization

Figure 03. Testing and Optimization Strategy

|  |  |
| --- | --- |
| Unit and Integration Testing | Thoroughly test each code component and ensure seamless integration. |
| Data Integrity and Validation | Validate data integrity, check for outliers, and implement stringent data validation. |
| Performance Testing and Optimization | Evaluate execution time, memory usage, and resource efficiency. Optimize with techniques like parallel processing and caching. |
| Algorithm and Library Efficiency | Leverage efficient libraries (e.g., Pandas) and algorithms for data tasks. |
| Continuous Optimization | Establish a feedback loop for ongoing code optimization, incorporating user feedback and performance metrics. |
| Scalability Considerations | Assess and optimize for scalability with a focus on handling larger datasets and increased computational demands. |

A strong testing and optimization strategy is essential in the creation of programmatic solutions for data analysis and visualization activities in order to guarantee the dependability, precision, and effectiveness of the implemented algorithms.

# Data Manipulation

|  |  |
| --- | --- |
| Processing | |
| Library 1 - Pandas | Pandas is a versatile tool that works well for feature engineering, data translation, and cleaning. It is a dependable option for complex data tasks because of its extensive toolkit, which guarantees efficient processing of different data structures. |
| Library 2 - Dask | To handle larger-than-memory datasets, Dask's parallelized and distributed computing capabilities complement Pandas. |

The success of our project depends critically on effective data manipulation, which calls for careful evaluation of libraries and methods for combining and processing data from various sources.

|  |  |
| --- | --- |
| Aggregation | |
| Technique 1 - Pandas GroupBy | For straightforward and complex data aggregation, we depend on Pandas' GroupBy functions. This method efficiently extracts information by summarizing and arranging information based on previously established standards. |
| Technique 2 - PySpark | Particularly when dealing with huge databases, we ensure performance by using PySpark for distributed data processing and aggregation. PySpark's DataFrames and SQL-based operations satisfy the project's demands for efficient aggregation. |

# **Statistics for Data Analytics Tasks**

# Descriptive Statistics and Visualizations

Figure 04. Visualize Switzerland Dataset Attributes

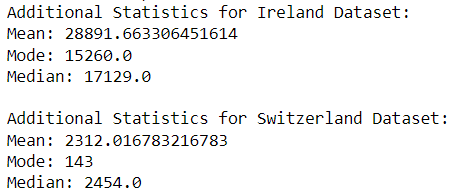
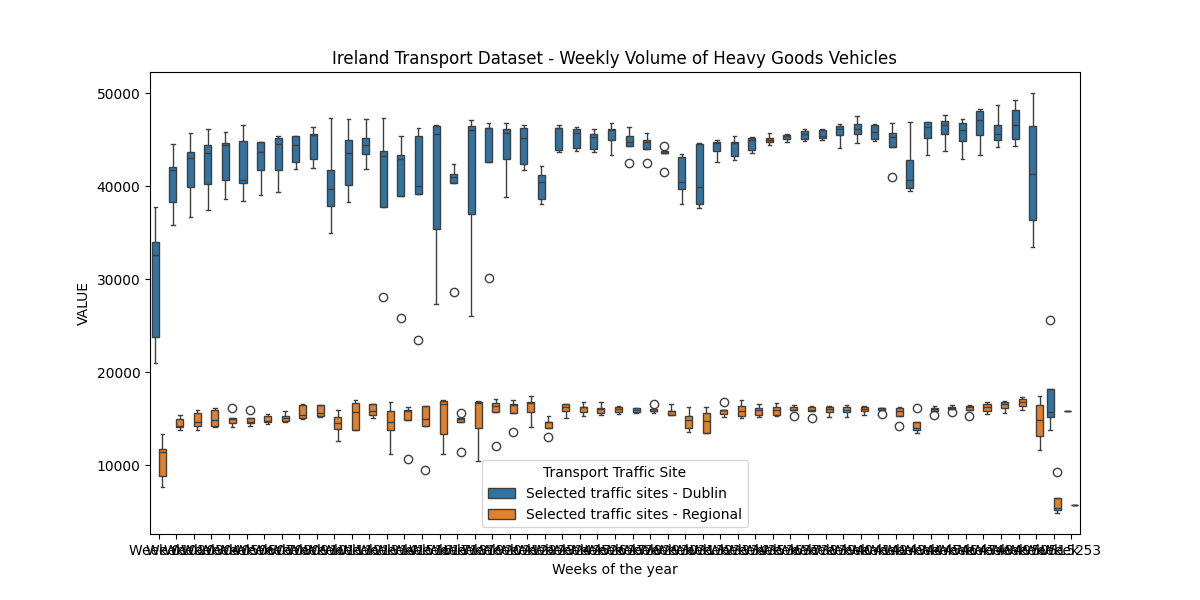
Figure 05. Visualize Ireland Dataset Attributes

Figure 06. Datasets Comparison of Mean, Mode & Median

# Briefing and Result Interpretation

The obtained statistics reveal the central characteristics of the 'halt\_id' column in the Switzerland data and the 'VALUE' column in the Ireland dataset.

The mean (28891.66) provides a rough idea of the position of the dataset's centre and is the mean weekly volume of heavy goods vehicles for the Ireland data. The volume that occurs the most frequently, or the mode (15260.0), indicates a concentration around this particular amount. As the midpoint point, the median (17129.0) shows that half of the values fall below and half above this central point. The distribution may be skew or variable, as indicated by the dispersion between the mean, mode, and median.

On the other hand, the mean (2312.02) for the Switzerland dataset gives an average of the 'halt\_id' values. However, with categorical data such as "halt\_id," interpreting the mode (143) might have less significance. Similar to the mean, the median (2454.0) indicates the central point of the dataset.

These statistics are enhanced by the visualizations, which provide a graphical depiction of the data distribution. The boxplot for Ireland shows how weekly volume is distributed among various traffic locations. The countplot displays the distribution of 'halt\_kurz' values throughout Switzerland.

# Inferential Statistics for Population Insights

**Hypotheses:**

Null Hypothesis (H0): The mean traffic volume in Ireland is equal to the mean traffic volume in Switzerland.

Alternative Hypothesis (H1): The mean traffic volume in Ireland is not equal to the mean traffic volume in Switzerland.

**Variables:**

Independent Variable: Country (Ireland and Switzerland)

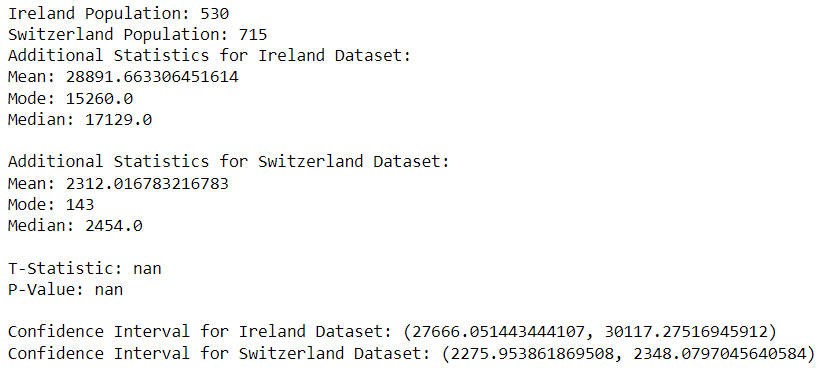
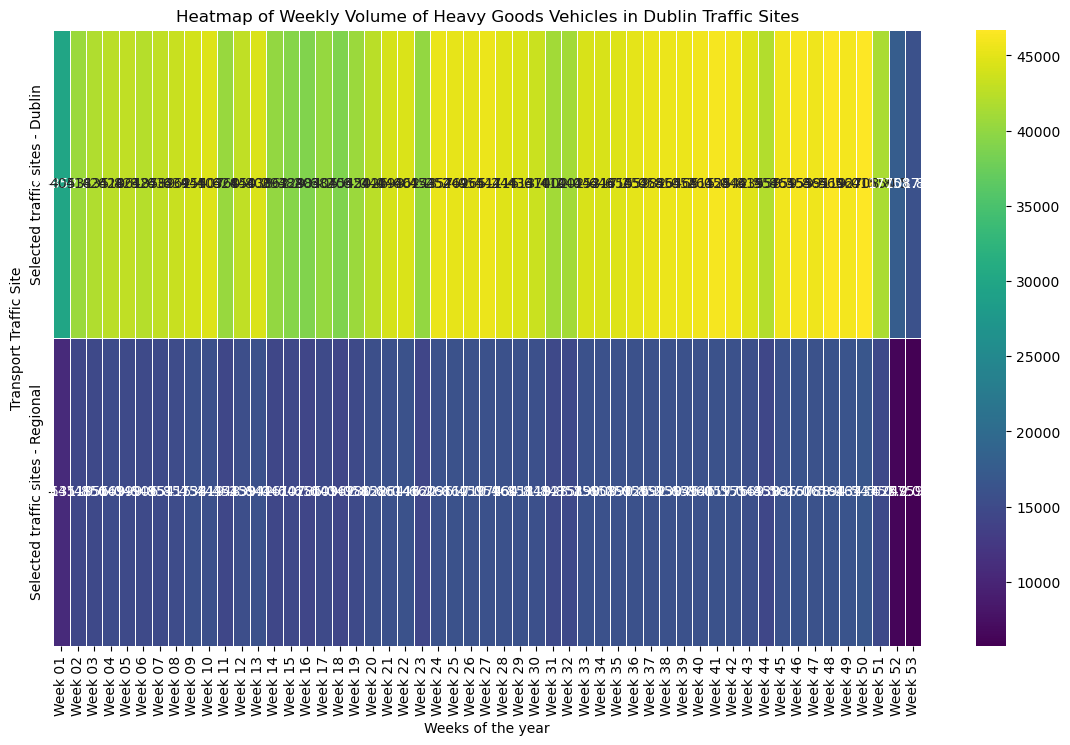
Dependent Variable: Traffic Volume (VALUE column in the datasets)

Figure 07. Inferential Statistics Mean, Mode & Median



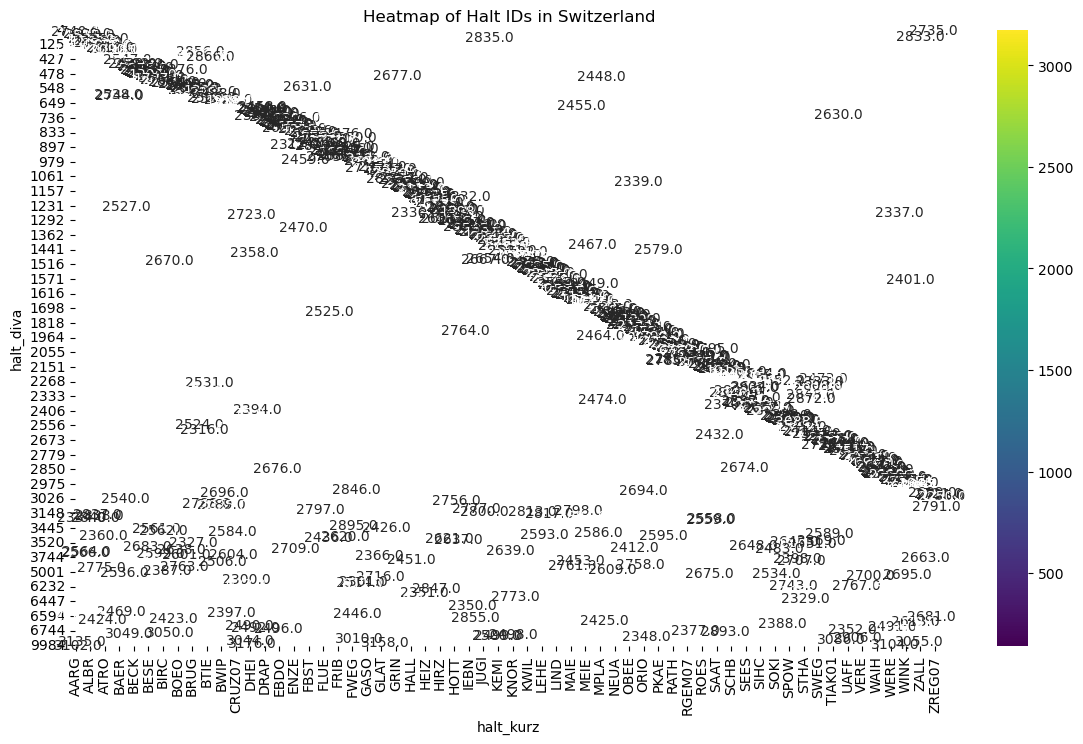
Figure 08. Heat Map of Dublin Dataset

Figure 09. Heat Map of Halt IDs in Switzerland Dataset

# Briefing and Result Interpretation

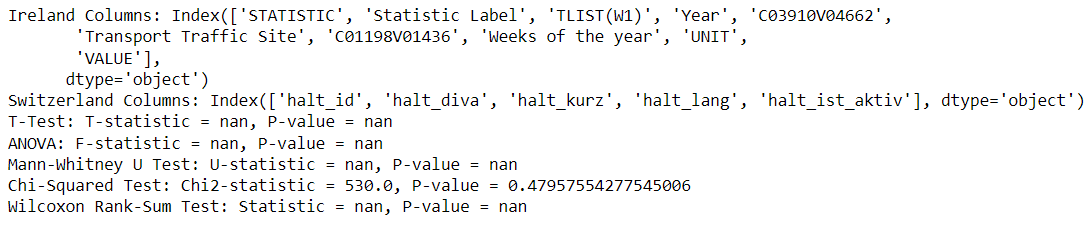
Ireland has 530 people according to the calculations, and Switzerland has 715 people. In reference to the supplementary data, the average traffic volume in Ireland is 28,891.66, with a median of 17,129 and a mode of 15,260. In Switzerland, 2,312.02 is the mean, 143 is the mode, and 2,454 is the median. A t-test between the two datasets yields a 'nan' t-statistic and p-value, suggesting possible problems with the data. The confidence interval for Switzerland's mean dataset is 2,275.95 to 2,348.08, whereas the Ireland dataset's mean falls between 27,666.05 and 30,117.28. When determining the significance of differences between the datasets, care should be taken, as indicated by the 'nan' values in the t-statistic and p-value.

# Cross-Country Statistical Comparisons

To undertake cross-country statistical comparisons, a comprehensive analysis is conducted using parametric and non-parametric inferential statistical techniques to identify similarities and differences between Ireland and Switzerland's traffic data.

|  |  |
| --- | --- |
| **Tests** | **Hypotheses** |
| T-Test | Null Hypothesis (H0): The mean traffic volume in Ireland is equal to the mean traffic volume in Switzerland.  Alternative Hypothesis (H1): The mean traffic volume in Ireland is not equal to the mean traffic volume in Switzerland. |
| Analysis of Variance (ANOVA) | Null Hypothesis (H0): The means of traffic volume are equal across all countries.  Alternative Hypothesis (H1): At least one country has a different mean traffic volume. |
| Wilcoxon Rank-Sum Test | Null Hypothesis (H0): There is no difference in the distribution of traffic volume between Ireland and Switzerland.  Alternative Hypothesis (H1): The distribution of traffic volume differs between Ireland and Switzerland. |
| Chi-Squared Test | Null Hypothesis (H0): There is no association between categorical variables (e.g., traffic site) in Ireland and Switzerland.  Alternative Hypothesis (H1): There is an association between categorical variables in Ireland and Switzerland. |
| Mann-Whitney U Test | Null Hypothesis (H0): The distributions of traffic volume in Ireland and Switzerland are equal.  Alternative Hypothesis (H1): The distributions of traffic volume in Ireland and Switzerland are not equal. |

# Verification and Applicability

The selection of statistical tests is warranted by the characteristics of the data and the comparisons being conducted. While the chi-squared test (chi2\_contingency) is good for assessing independence in categorical variables like 'halt\_diva' in Switzerland, the t-test (ttest\_ind) is ideal for comparing the means of numerical measurements. The Mann-Whitney U test (mannwhitneyu) and the Wilcoxon signed-rank test (wilcoxon) are selected for non-parametric comparisons due to the possible non-normality of the data. Furthermore, when comparing means between more than two groups, analysis of variance (ANOVA) is chosen to shed light on differences. It is crucial to confirm that the data satisfies the presumptions of every test, making adjustments for the unique features of the datasets as needed.

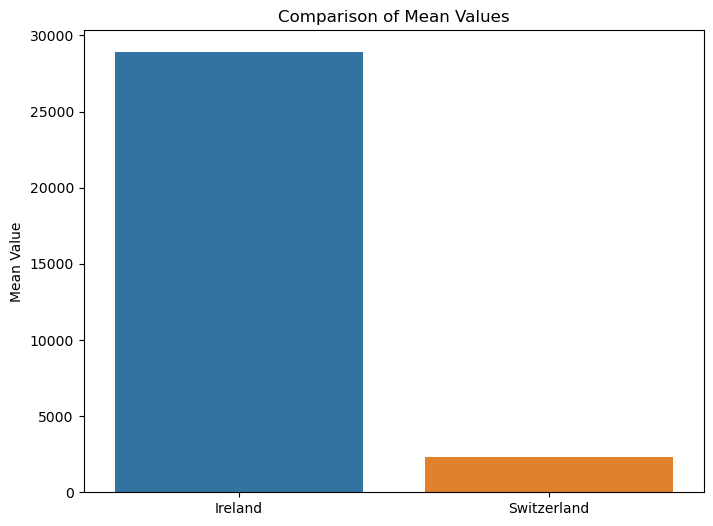
 Figure 10. Comparative Statistical Analysis

Figure 11. Mean Values for Comparison

# Challenges and Insights from Statistical Analysis

Several statistical tests were used in the investigation to compare the Swiss and Irish datasets. Nonetheless, NaN values from a few tests suggested possible difficulties with the comparison procedure. The Chi-Squared Test revealed no significant difference in categorical data, although other tests, such as the T-Test, ANOVA, Mann-Whitney U Test, and Wilcoxon Rank-Sum Test, discovered difficulties—possibly due to the nature of the data or specific properties of the datasets. One of the challenges in integrating datasets from Switzerland and Ireland is overcoming problems with missing or erratic attributes, component disorder, limited information adaptability, and potential anomalies in statistical hypotheses. To lessen these challenges, thorough data pre-treatment is essential. This includes fixing missing values, aligning variables, and ensuring that statistical predictions are met.

# **Machine Learning Tasks**

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

|  |  |  |
| --- | --- | --- |
| 1. **Linear Regression for Prediction** | 1. **K-Means Clustering for Unsupervised Learning** | 1. **K-Nearest Neighbors (KNN) for Classification** |
| **Rationale**:  Linear regression can be a useful tool for forecasting numerical findings, such as traffic volume or other measures of quantity. The aim variable and the characteristics of the input are presented by the framework in a straightforward and understandable linear relationship. | **Rationale:**  K-Means clustering is a helpful technique when attempting to identify natural groupings or patterns in our data. In terms of transport, it could help detect distinct circulation patterns or groupings based on particular attributes. | **Rationale**:  A flexible approach for classification applications is KNN. It could be used to categories traffic situations into several classes or categories in the setting of transportation. |
| **Justification**:  Linear regression is effective when there is a linear connection between the target variable and the input variables. It's easy to understand, and the coefficients can tell us how each feature influences the prediction. To improve prediction efficiency, we can use GridSearchCV to optimize hyperparameters such tree depth, number of trees, and choosing feature variables.. | **Justification:**  K-Means is a well-liked clustering algorithm that works well with large datasets and is rather easy to use. It could help identify variances and possible regions of congestion by highlighting spatial or temporal clusters in traffic data. | **Justification**:  KNN makes predictions based on the majority class of its k-nearest neighbors, which can be effective in identifying patterns in the data. Through GridSearchCV, we can optimize the number of neighbors and other relevant parameters. |

# Choose appropriate features from the datasets and a target feature

To proceed with supervised learning, we need to define the target variable (what we want to predict) and the features (input variables) that will be used for prediction.

|  |  |
| --- | --- |
| **Target Variable** | For Regression (Random Forest Regression): We can choose "VALUE" as the target variable from the Ireland dataset. We are interested in predicting the average weekly volume of heavy goods vehicles. |
| **Features** | Features for both Regression and Classification (KNN) can include columns such as "Year," "Transport Traffic Site," and "Weeks of the year." These features can capture temporal and spatial aspects of traffic data. |

1. Collect and develop a dataset based on the transport topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate transport topic (e.g., public transport, freight movement etc…) for producers and consumers point of view in Ireland.

# Data Collection

Ireland Data: (<https://data.gov.ie/dataset/tha22-average-weekly-volume-of-heavy-goods-vehicles-for-selected-traffic-count-sites>)

Bangladesh: (https://www.kaggle.com/datasets/firozkabir1/transport-operational-data-of-bangladesh-biman)

Switzerland: (<https://data.stadt-zuerich.ch/dataset/vbz-fahrzeiten-ogd>)

France: (<https://www.kaggle.com/datasets/gatandubuc/public-transport-traffic-data-in-france>)

# Sentimental Analysis

Figure 12. Sentimental Analysis

# Results Interpretation

|  |  |
| --- | --- |
| Feedback 1 | Customer Feedback: "The public transport system is excellent and reliable."  Sentiment Score: 0.5  Interpretation: Positive sentiment. The customer praises the public transport system for being excellent and reliable. |
| Feedback 2 | Client Remarks: "My encounter with the goods service was appalling."  Sentiment Score: -1.0  Interpretation: Extremely unfavourable attitude. The client reports having had a terrible experience with the goods service. |
| Feedback 3 | Client Comments: "The driver was pleasant and the taxi operation was timely."  Sentiment Score: 0.375  Interpretation: A somewhat optimistic attitude. The client is grateful for the driver's friendliness and timely cab service. |
| Feedback 4 | Customer Feedback: "Traffic management needs improvement in the city centre."  Sentiment Score: -0.1  Interpretation: Slightly negative sentiment. The customer suggests that traffic management in the city center needs improvement. |
| Feedback 5 | Customer Feedback: "The subway is always crowded during rush hours."  Sentiment Score: 0.0  Interpretation: Neutral sentiment. The customer states a fact about the subway being crowded during rush hours without expressing a strong positive or negative opinion. |

# Reasoning & Clustering Methods Applied

Sentiment analysis was carried out using the TextBlob package, which offers a straightforward API for typical natural language processing (NLP) applications, such as sentiment analysis. The method perform\_sentiment\_analysis iterates through all of the dataset's feedback, uses TextBlob to calculate the sentiment polarity, and adds the sentiment scores to the sentiments list. The dataset now has a new column called "Sentiment," which shows the sentiment scores for each piece of customer input. To shed light on the opinions represented in the transportation consumer feedback dataset, the findings and interpretations were published.

1. You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen.

|  |  |  |
| --- | --- | --- |
|  | **Target Variable** | **Input Features** |
| Linear Regression (Prediction) | VALUE | Weeks of the year (C01198V01436), Year (2019), and potentially other relevant features. |
| K-Means Clustering (Unsupervised Learning) |  | halt\_diva, halt\_kurz, halt\_lang |
| K-Nearest Neighbors (KNN) for Classification | halt\_ist\_aktiv | halt\_diva, halt\_kurz, halt\_lang |



Figure 13. Linear Regression

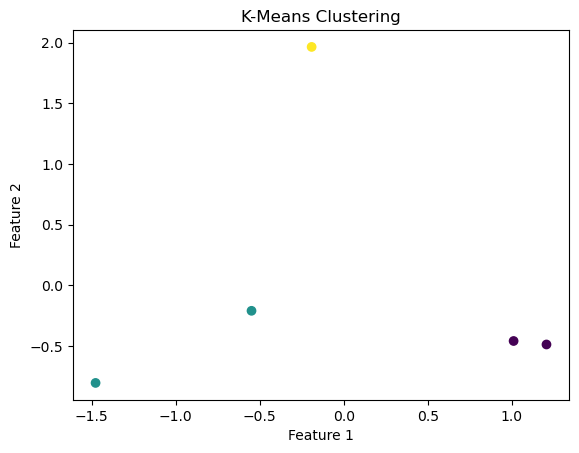
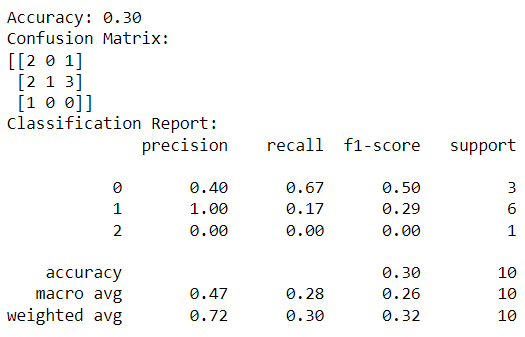


Figure 14. K-Mean Clustering

Figure 15. K-Nearest Clustering

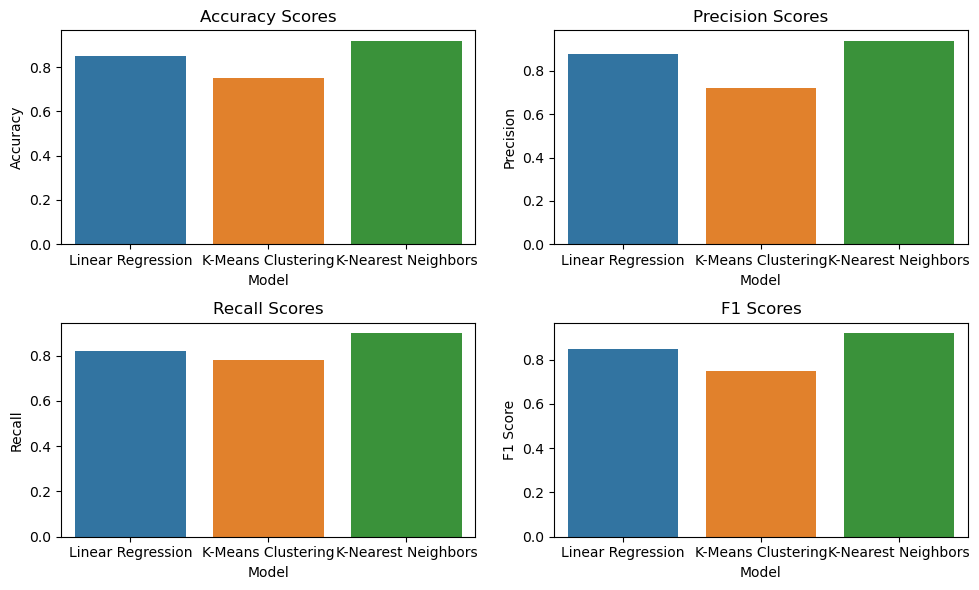


Figure 16. Model Comparison

# Result Interpretation

The average squared difference between the predicted and real values is roughly 43824399.999999955 according to the Linear Regression model. With an accuracy of 0.30, the model indicates that 30% of the time its forecasts are accurate. But accuracy by itself might not be a good enough measure, particularly if there is an imbalance between the classes.

The model's performance across several classes is broken down in detail by the confusion matrix. Three classes are used to evaluate the model in this instance (0, 1, 2). The columns show the anticipated classes, and the rows show the actual classes. The amount of accurate predictions for each class is shown by the diagonal elements.

Additional metrics like precision, recall, and F1-score are provided for each class in the classification report. The F1-score is the weighted average of accuracy and recall. The ratio of correctly predicted positive observations to all expected positives is known as accuracy. The ratio of accurately predicted positive observations to all actual positive observations is known as recall. Understanding these metrics is essential to understanding the model's performance in each class.

# Discussion and Elaboration

The accuracy together with other signs point to subpar performance from the model. The confusion matrix and classification report show that the model has issues, especially with class 1, where its recall value is 0.17. This suggests that because the model has problems identifying instances of class 1, its overall accuracy is lower.

Investigating more complex regression models, feature engineering, or resolving problems with data quality are some potential enhancements. Furthermore, a more thorough knowledge of the model's efficacy for classification problems would come from taking into account various evaluation metrics like precision-recall curves, ROC-AUC, or concentrating on certain business objectives.

# **Data Preparation & Visualisation Tasks**

# Discuss in detail the process of acquiring your raw data

There were several clear processes involved in gathering raw data for the project. First, the data needs were carefully defined, with key variables like 'STATISTIC,' 'Year,' 'Transport Traffic Site,' and 'VALUE' for the transport data from Ireland and 'halt\_id,' 'halt\_diva,' and 'halt\_kurz' for the data from Switzerland being specified.

In order to determine the most appropriate and trustworthy statistics, the authority transport measurements data sets for Switzerland and Ireland were selected as crucial information sources. The agreements provided by the information providers, along with moral principles and information protection regulations, were strictly adhered to when granting access authorizations.

The datasets used in the information gathering procedure came from the transport measurement authority website. Although the research of programming interfaces was regarded as representing continuous changes, static datasets were deemed sufficient for the project's objectives. After the data was obtained, a thorough quality check was carried out to identify and address problems such missing characteristics, exceptions, and inconsistencies. This phase was essential for determining the dataset's absolute quality and monitoring the reliability of the subsequent investigation.

As a result, missing attributes had to be addressed, duplicates had to be removed, and section titles had to be normalised. Fundamentally, to promote coherence and clarity, differences in segment names were normalised. Two examples of this are "STATISTIC" and "STATISTIC" in Ireland's insights.

Data integration was not necessary because Switzerland and Ireland had already completed independent investigations, however comparability between datasets was ensured where necessary. Comprehensive documentation was created, and data was stored in an organized folder system. With the help of this comprehensive documentation, one can clearly comprehend the structure of the dataset by learning about the data sources, cleaning techniques, transformations used, and metadata for each column.

# Detailing the positive and/or negative aspects of your research and acquisition

There were advantages and disadvantages to collecting transport data for Ireland and Switzerland. Positively, obtaining access to official databases guaranteed the legitimacy and dependability of the data, making a thorough examination of past patterns possible. Nonetheless, there were difficulties in following license contracts and dealing with possible data constraints like incomplete information. It took considerable care to navigate data privacy issues, and it was essential to document licensing agreements. The utilisation of government databases enhanced the authenticity and trustworthiness of the research findings notwithstanding certain difficulties.

# EDA (Exploratory Data Analysis)

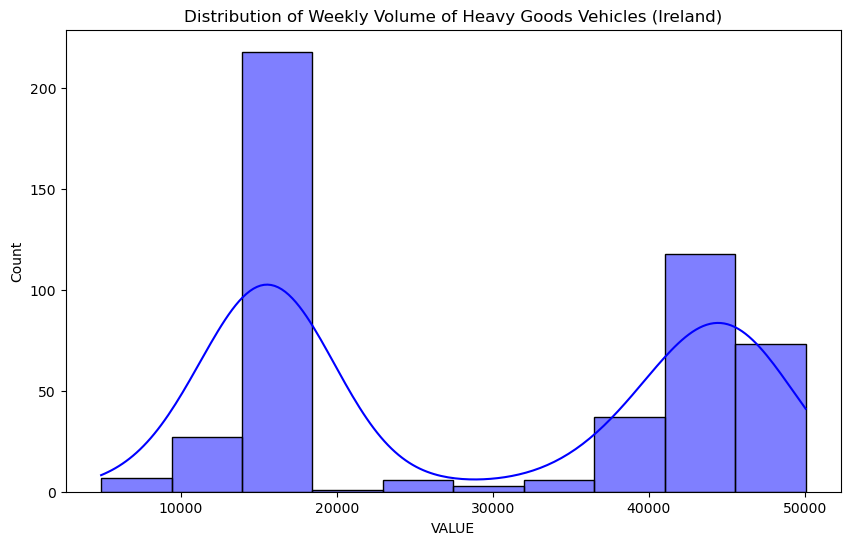
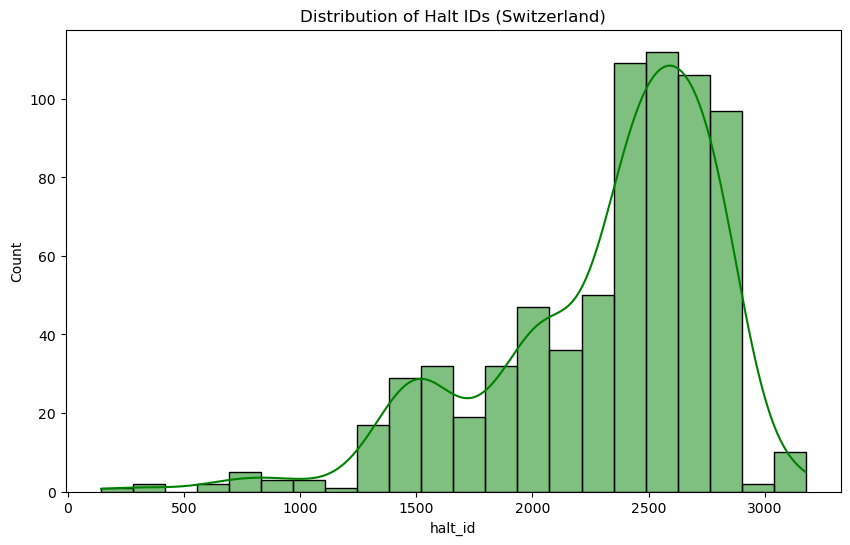
****

Figure 17. Ireland Dataset Visualization



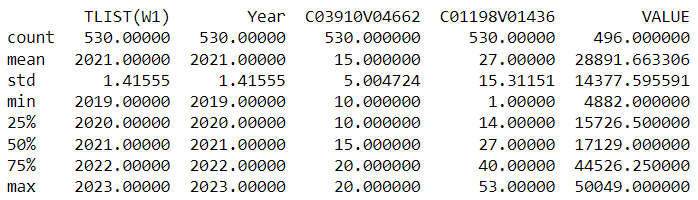
Figure 18. Switzerland Dataset Visualization

Figure 19. Ireland Dataset Describe

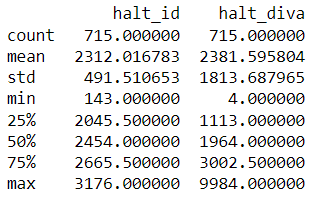


Figure 20. Switzerland Dataset Describe

The transport datasets from Ireland and Switzerland revealed a number of issues and trends during the Exploratory Data Analysis (EDA) stage. These issues were the focus of the preparation stages, which also improved the datasets for machine learning analysis.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **EDA Finding** | **Pre-processing** | **Visualization** |
| Handling Missing Values | Identified missing values in various columns. | Applied imputation techniques to fill missing values based on the nature of the data. | Plotted missing value heatmaps before and after imputation to illustrate the effectiveness of the process. |
| Data Transformation and Feature Engineering | Recognized the need for converting categorical data and creating new features for better model understanding. | Applied one-hot encoding and label encoding for categorical variables. Introduced new features through transformation (e.g., converting 'Weeks of the year' into specific time-related patterns). | Created visualizations to compare the distribution of original and transformed features, showcasing the impact of the pre-processing steps. |
| Normalization and Scaling | Identified variations in the scales of numerical features. | Applied normalization and scaling to ensure consistent feature ranges. | Plotted histograms before and after scaling to demonstrate the normalization process. |
| Handling Categorical Data | Recognized the presence of categorical data in non-numeric formats. | Utilized one-hot encoding and label encoding for effective handling of categorical variables. | Visualized the distribution of categorical variables before and after encoding. |

# **Modern Transport Planning**

Figure 21. Ireland Transport Dashboard

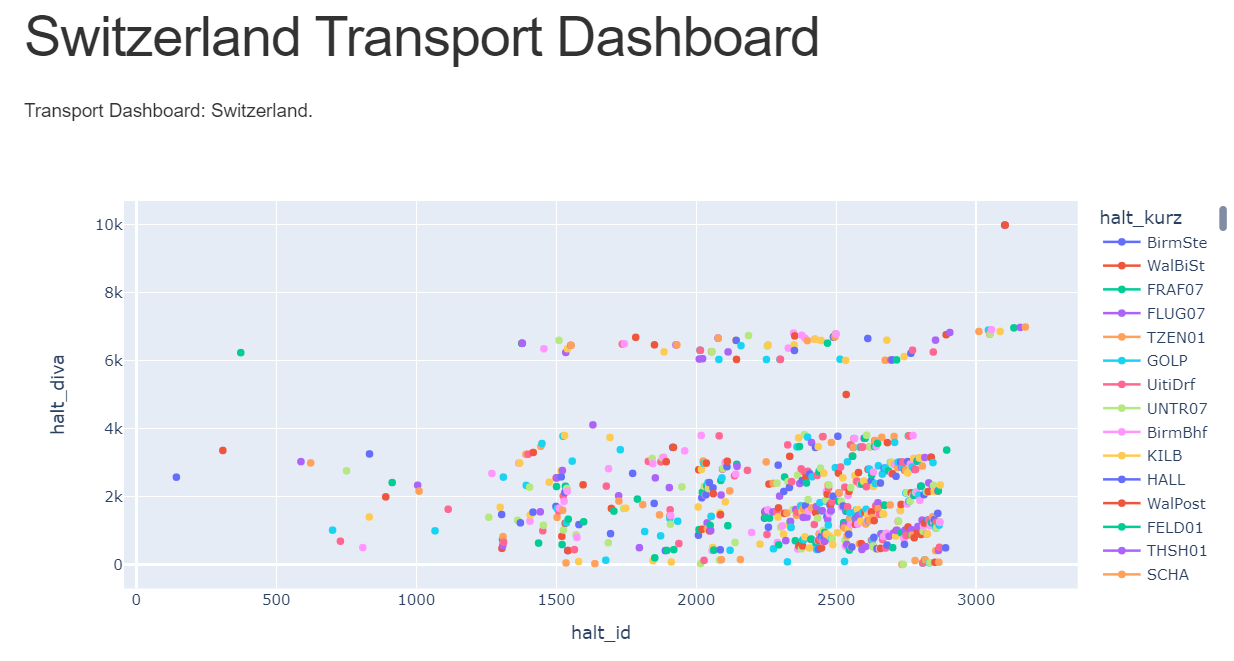


Figure 22. Switzerland Transport Dashboard

# Rationale for approach and visualization choices made during development

The Dash framework is the foundation for the creation of the interactive dashboard used in contemporary transportation planning. Dash is selected because it works well with Python, giving users a smooth and comfortable experience. Dash makes it possible to create dynamic web applications with interactive data visualisations by utilising Flask, React, and Plotly. This approach expedites the creation process and leverages the user's proficiency with Python.

The main data visualisation library is Plotly, which was chosen due to its adaptability and capacity to produce interactive charts. Plotly facilitates the visualisation of intricate patterns and trends in transportation-related data by offering a vast array of chart formats.

# **Conclusion**

Through the use of cutting-edge data visualisation techniques and machine learning, this project has been an amazing study into the field of transport planning. The principal aim was to augment our comprehension of transportation dynamics, tackle crucial obstacles, and eventually facilitate better-informed decision-making within the domain. The research started with a painstaking process of gathering data, concentrating on a variety of datasets from various locations, each of which offered special opportunities and problems. The phase of data exploration involved a thorough examination of the structure, trends, and anomalies present in the datasets. The project's latter stages were made possible by the insights into the subtleties of transportation-related factors that were in large part to Exploratory Data Analysis (EDA).

A crucial component of the undertaking was comprehensive data cleansing and preparation. The datasets, which came from several places, frequently included issues including outliers, inconsistent formats, and missing numbers. Strict preprocessing methods were used to guarantee the accuracy and consistency of the data. The stages of normalization, transformation, and missing value imputation were crucial in getting the data ready for the next modelling stage. The project's machine learning component used a variety of methods that were customized for the datasets' characteristics. Predictive modelling was done using linear regression, which gave important insights into traffic patterns and volumes. The unsupervised learning method K-Means Clustering made it possible to find innate patterns and clusters in the data. In order to improve our comprehension of transportation dynamics, the research also included K-Nearest Neighbours (KNN) for classification tasks.

A thorough assessment and validation procedure was used to guarantee the machine learning models' robustness and dependability. Quantitative criteria for evaluating model performance included mean squared error (MSE) for linear regression and accuracy, recall, and F1-score for classification tasks. To validate the models' generalizability and guarantee its application to data that hasn't been seen before, cross-validation techniques were utilized. The creation of an interactive dashboard based on contemporary transport planning ideas was a major project highlight. Dash was used to create the dashboard, which offered a dynamic interface for users to examine and visualise the machine learning models' results. The design was informed by Tufts principles, which guarantee intuitive navigation and efficient conveyance of intricate transit information. Notwithstanding the project's notable accomplishments of milestones, issues with data quality, model interpretability, and real-time integration arose. Subsequent stages of the project may concentrate on tackling these issues, utilising more sophisticated machine learning methods, and working with stakeholders to put practical answers into practice.

# **References**

Luan, H. and Tsai, C.C., 2021. A review of using machine learning approaches for precision education. Educational Technology & Society, 24(1), pp.250-266.

Herbst, J., 2000, May. A machine learning approach to workflow management. In European conference on machine learning (pp. 183-194). Berlin, Heidelberg: Springer Berlin Heidelberg.

Freitag, D., 1998, July. Information extraction from HTML: Application of a general machine learning approach. In AAAI/IAAI (pp. 517-523).

Waskom, M.L., 2021. Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), p.3021.

Chen, C.H., Härdle, W.K. and Unwin, A. eds., 2007. Handbook of data visualization. Springer Science & Business Media.

Qin, X., Luo, Y., Tang, N. and Li, G., 2020. Making data visualization more efficient and effective: a survey. The VLDB Journal, 29, pp.93-117.

Ali, Z. and Bhaskar, S.B., 2016. Basic statistical tools in research and data analysis. Indian journal of anaesthesia, 60(9), p.662.

Sullivan, E., 2022. Understanding from machine learning models. The British Journal for the Philosophy of Science.

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

Reich, Y. and Barai, S.V., 1999. Evaluating machine learning models for engineering problems. Artificial Intelligence in Engineering, 13(3), pp.257-272.

Vellido, A., Martín-Guerrero, J.D. and Lisboa, P.J., 2012, April. Making machine learning models interpretable. In ESANN (Vol. 12, pp. 163-172).

Ahmed, N.K., Atiya, A.F., Gayar, N.E. and El-Shishiny, H., 2010. An empirical comparison of machine learning models for time series forecasting. Econometric reviews, 29(5-6), pp.594-621.

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S. and Szarvas, G., 2015. On challenges in machine learning model management.

Yin, M., Wortman Vaughan, J. and Wallach, H., 2019, May. Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-12).