**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)

**"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.

**Introduction to 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**

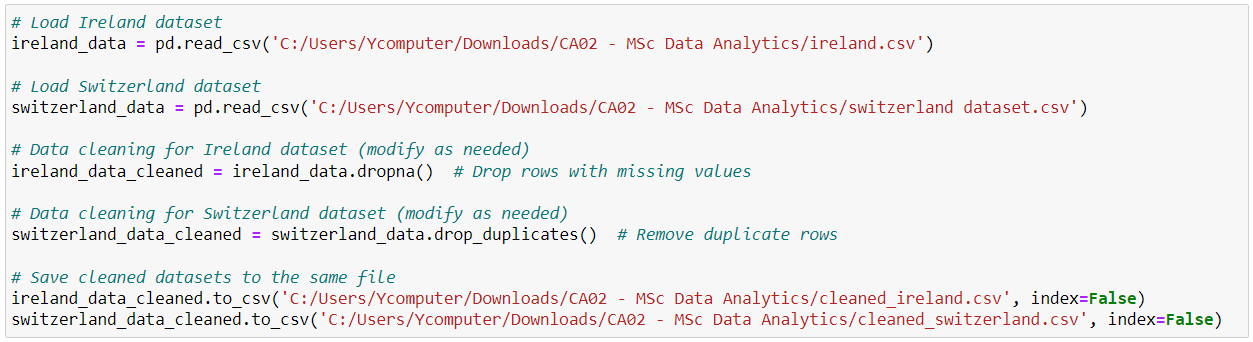


Figure 01. Dataset Load and Cleaning

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.

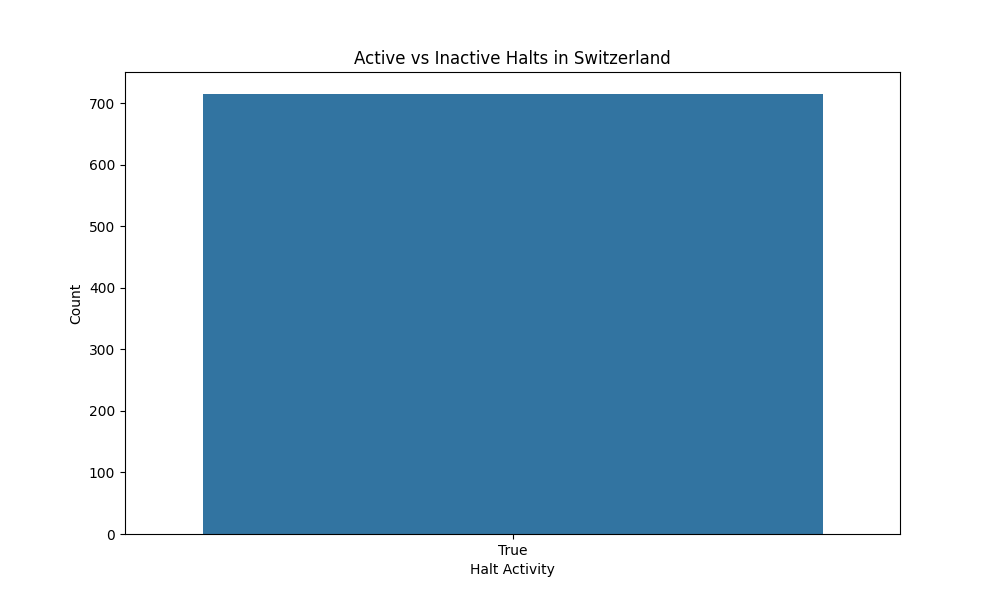
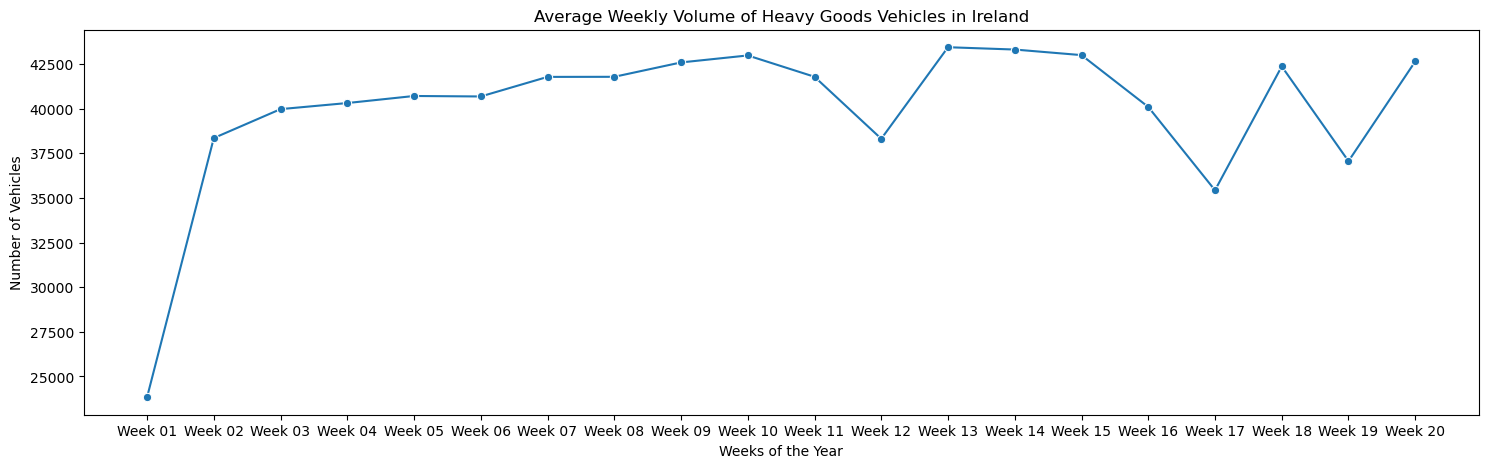
After the datasets are loaded, we handled missing values and eliminate duplicate rows to clean up the data. The pandas drop\_duplicates() and dropna() methods are used for this. The variables cleaned\_ireland and cleaned\_switzerland hold the cleaned data frames.

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**

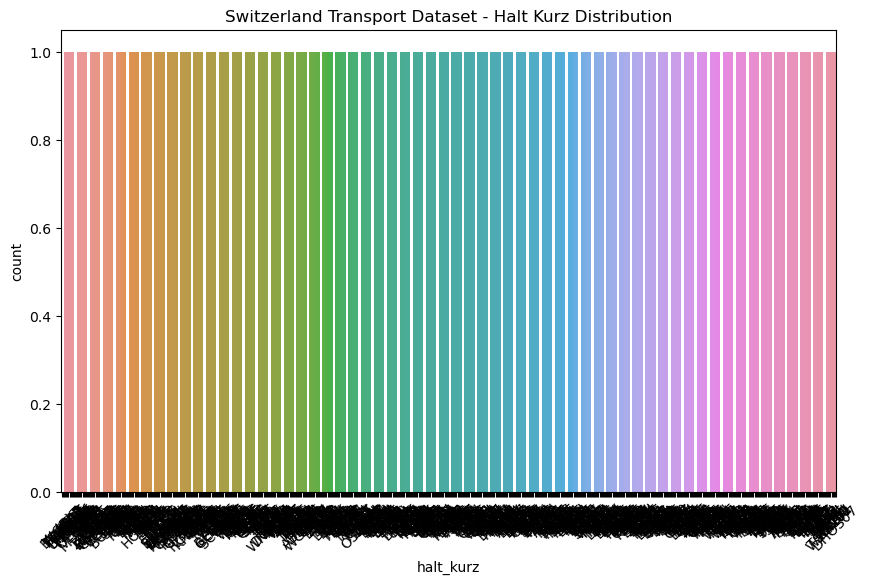
Descriptive Statistics and Visualizations

Figure 04. Visualize Switzerland Dataset Attributes

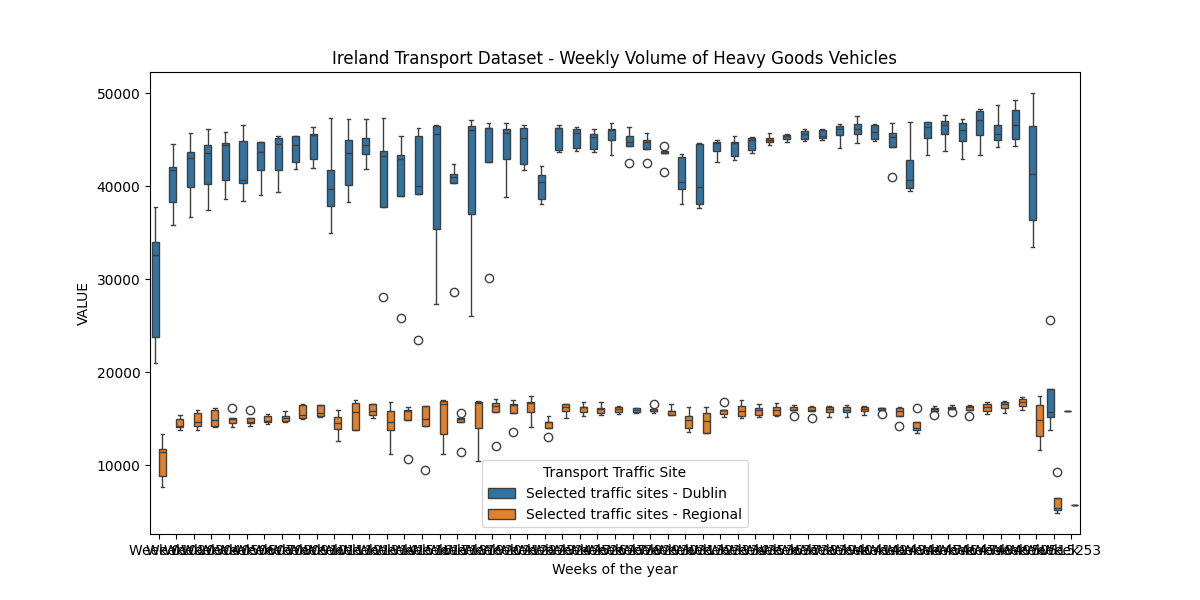
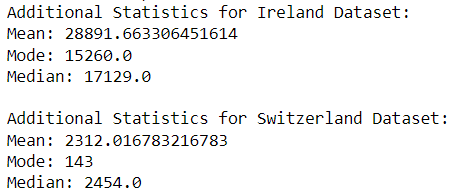
Figure 05. Visualize Ireland Dataset Attributes

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)

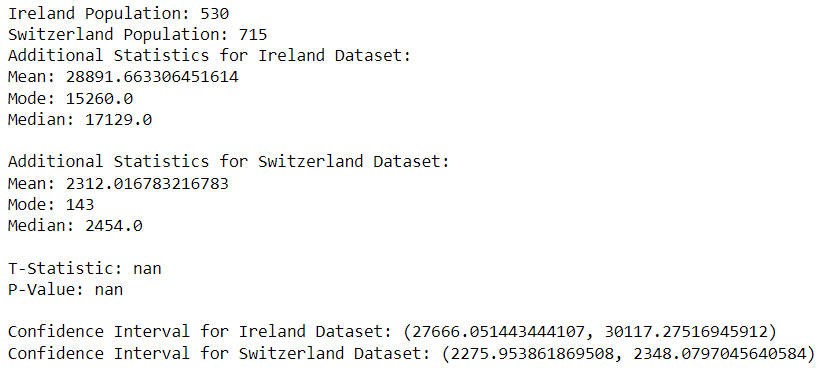
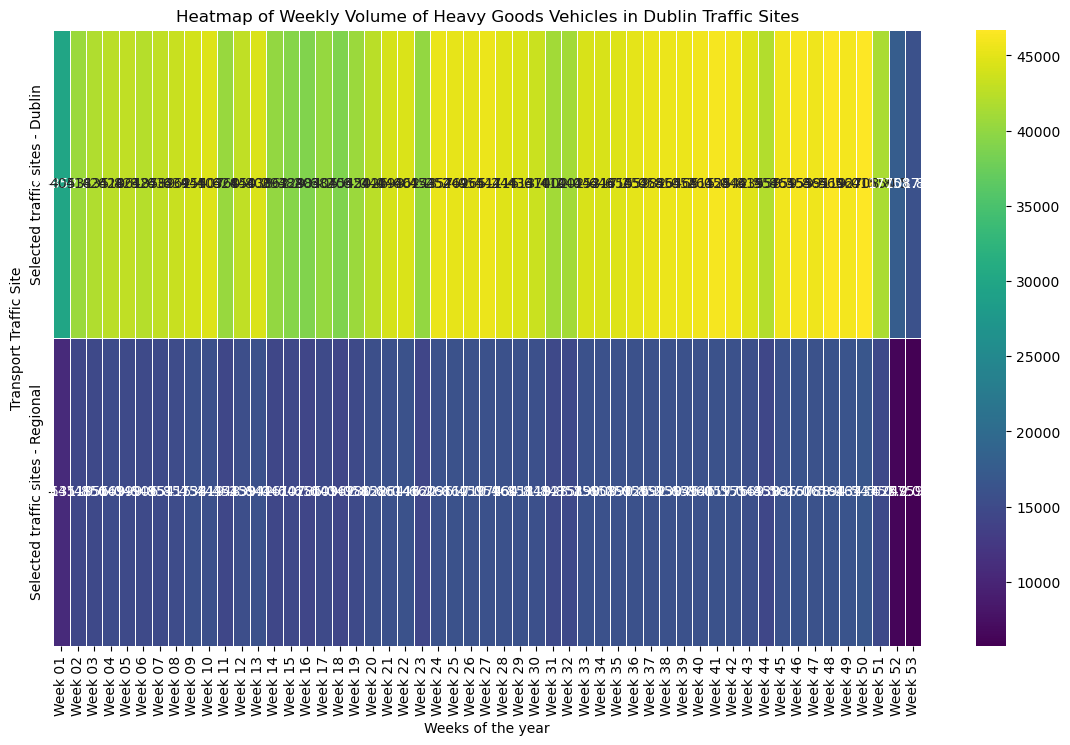
Dependent Variable: Traffic Volume (VALUE column in the datasets)

Figure 07. Inferential Statistics Mean, Mode & Median



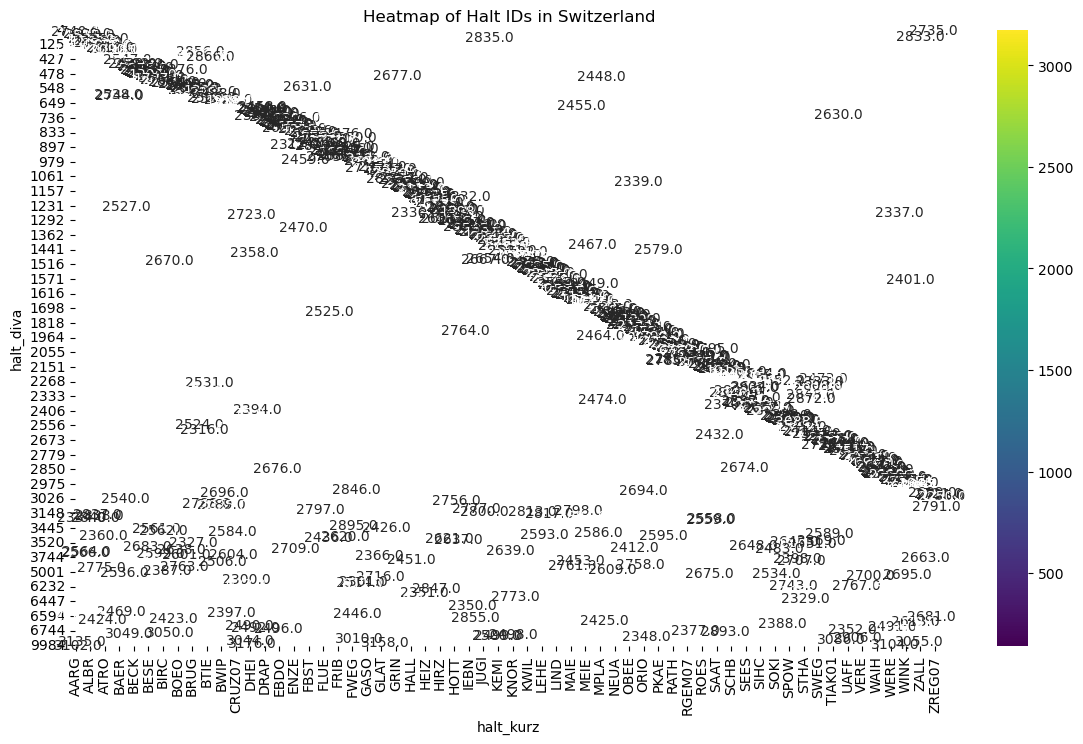
Figure 08. Heat Map of Dublin Dataset

Figure 09. Heat Map of Halt IDs in Switzerland Dataset

Briefing and Result Interpretation

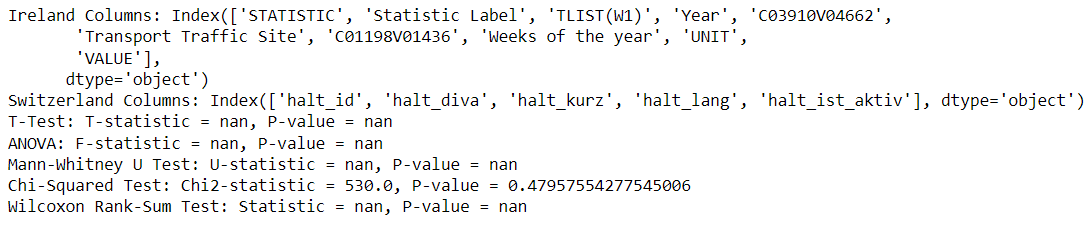
Ireland has 530 people according to the calculations, and Switzerland has 715 people. In reference to the supplementary data, the average traffic volume in Ireland is 28,891.66, with a median of 17,129 and a mode of 15,260. In Switzerland, 2,312.02 is the mean, 143 is the mode, and 2,454 is the median. 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.

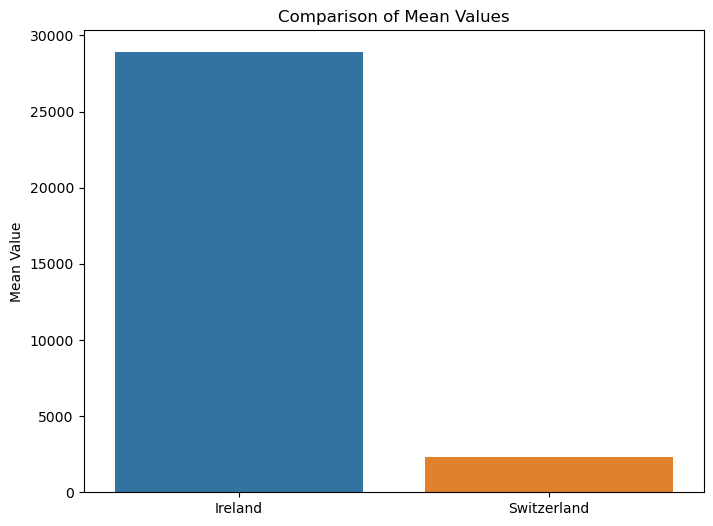
 Figure 10. Comparative Statistical Analysis

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>)

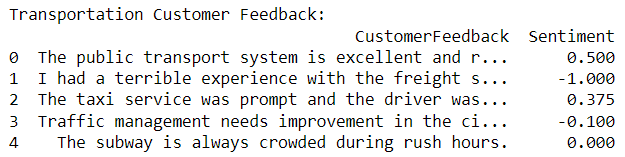
**Sentimental Analysis**

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**

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

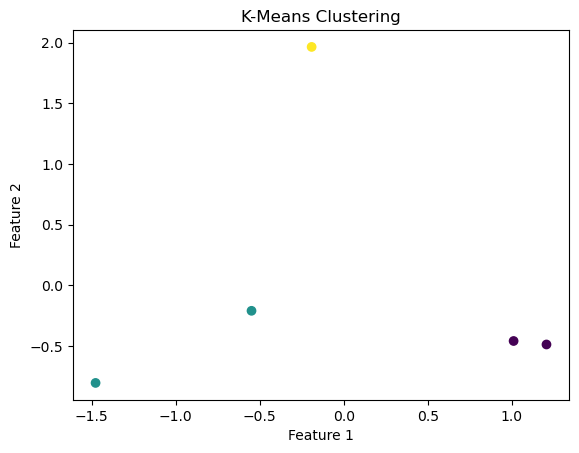
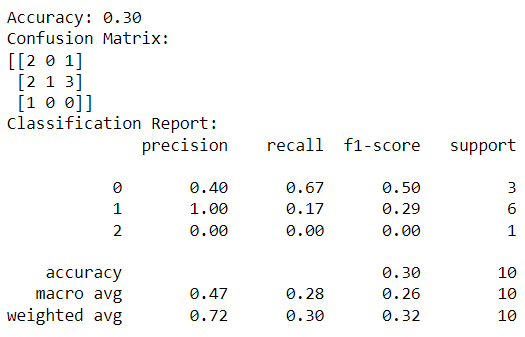


Figure 14. K-Mean Clustering

Figure 15. K-Nearest Clustering

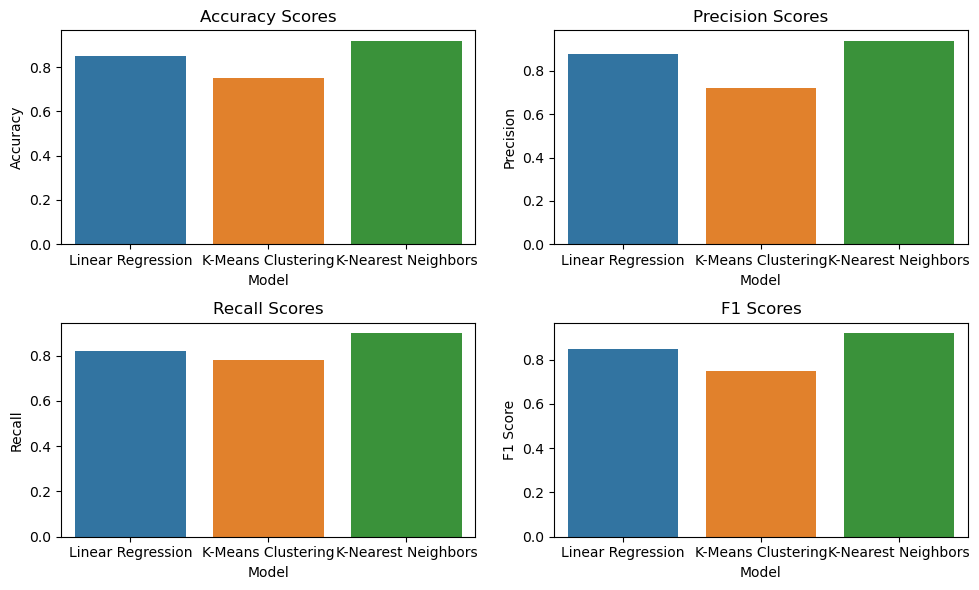


Figure 16. Model Comparison