**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**

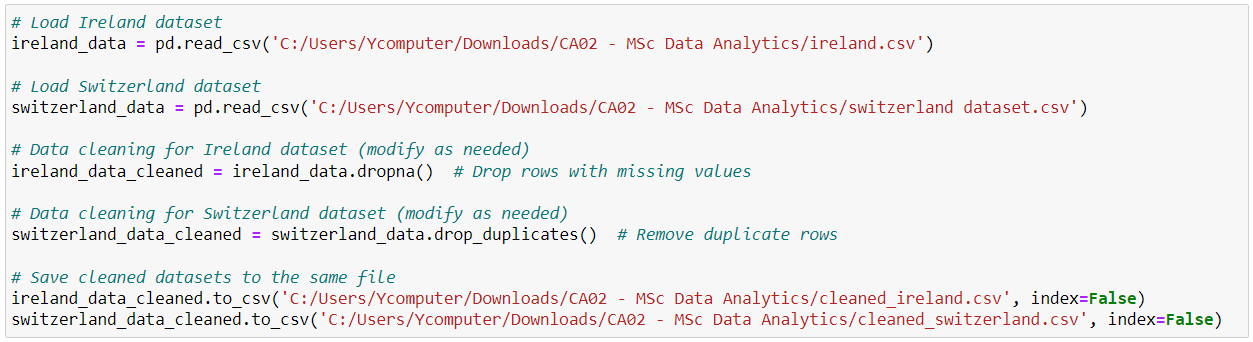


Figure 01. Dataset Load and Cleaning

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.

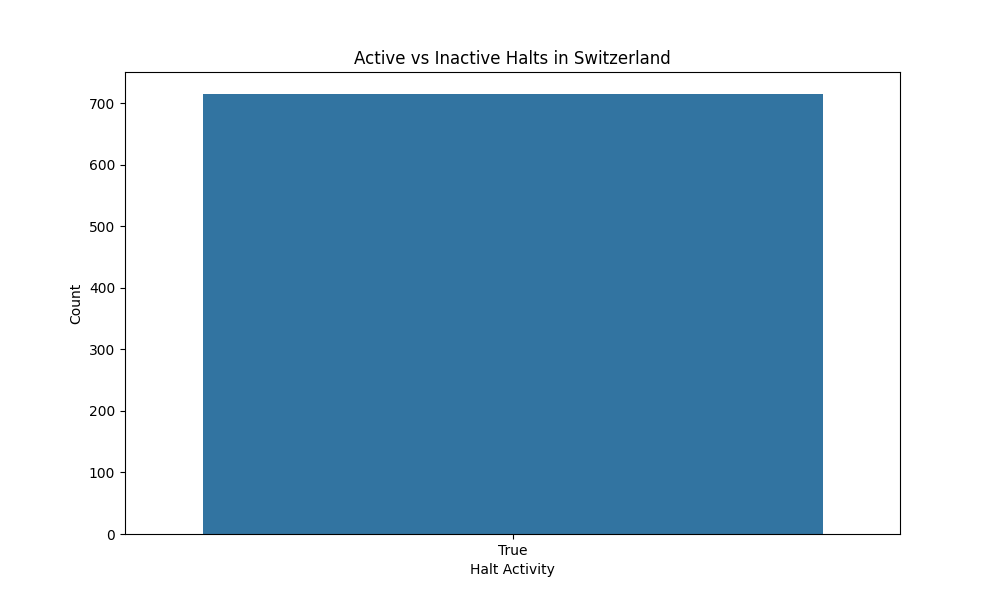
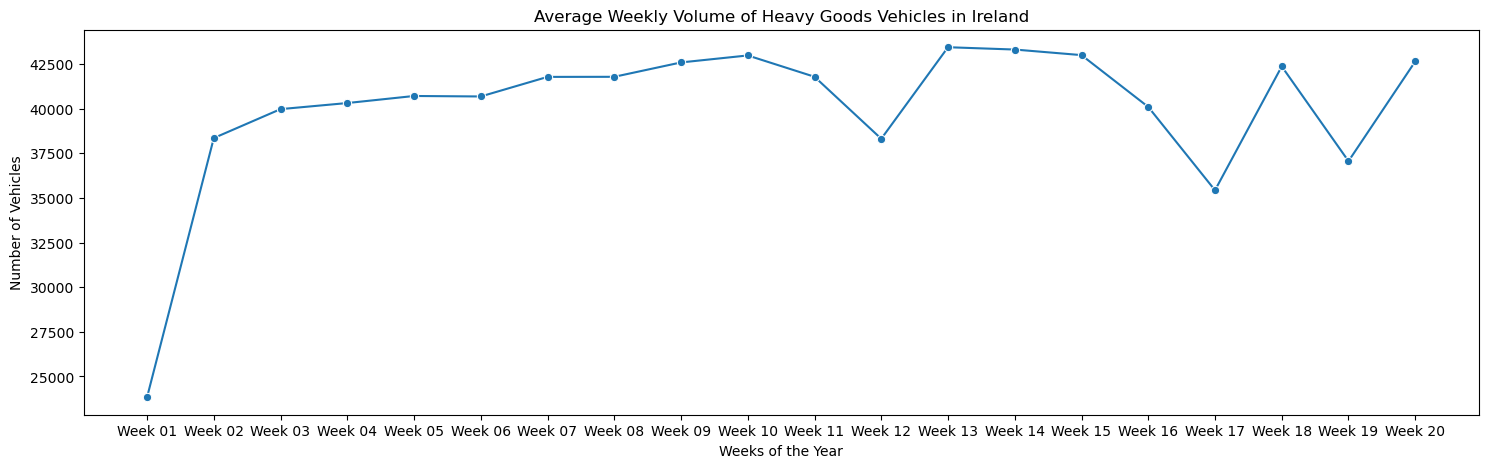
After the datasets are loaded, we handled missing values and eliminate duplicate rows to clean up the data. The pandas drop\_duplicates() and dropna() methods are used for this. The variables cleaned\_ireland and cleaned\_switzerland hold the cleaned data frames.

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**

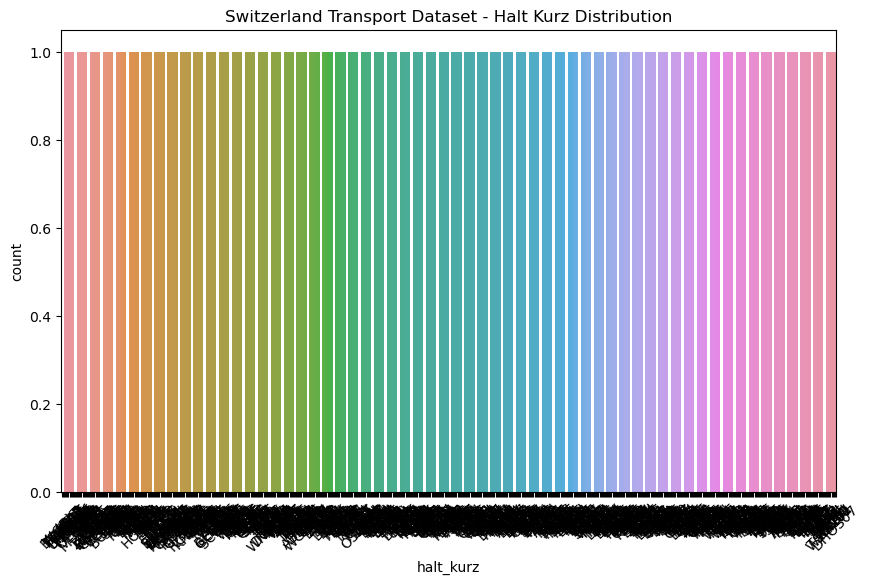
Descriptive Statistics and Visualizations

Figure 04. Visualize Switzerland Dataset Attributes

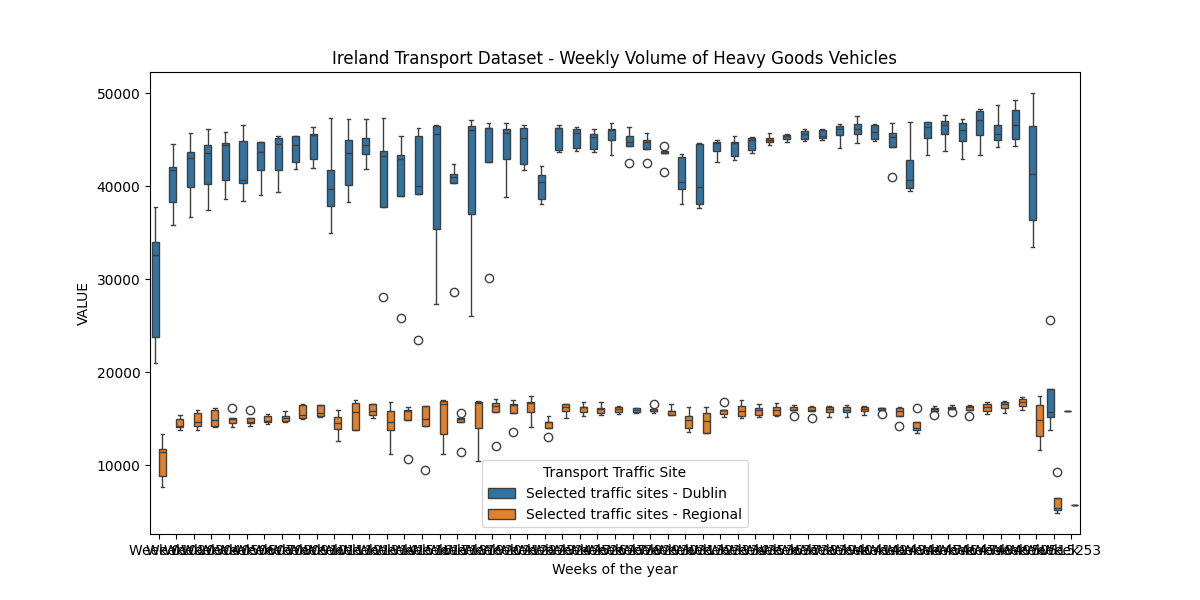
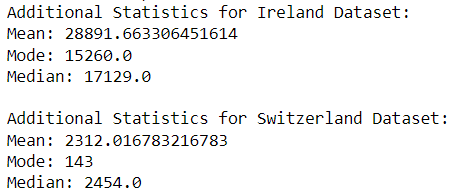
Figure 05. Visualize Ireland Dataset Attributes

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.