**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for Data Analytics*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *Analytical Study: Volume of Passengers of Trams/Bus Systems in Dublin and Basel* |
| **Lecturer Name:** | *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid*  Taufique Ahmed |
| **Student Full Name:** | Bruno Conti Souza Paes da Silva |
| **Student Number:** | 2023387 |
| **Assessment Due Date:** | 05/12/2024 |
| **Date of Submission:** | 05/12/2024 |

**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. |

**Data Verification for volume of Passengers for Trams/Luas for Dublin and Basel**

# Abstract

*This research aims on verifying passenger volume data for buses and trams in Dublin and Basel, employing datasets from the Central Statistics Office Ireland (CSO) and Basler Verkehrs-Betriebe (BVB). It was utilized techniques from Statistics, Machine Learning, Programming, and Data Preparation and Visualization. Additionally, sentiment analysis is applied to the Reddit page "LUAS" using Python code with the PRAW, TextBlob, and tabulate libraries.*

*In the statistical analysis using Jupyter notebook, we tackle challenges with limited data, showcasing adaptability and collaboration. Specific functions, like "calculate\_confidence\_interval\_binary" and "t\_test\_total\_passengers," provide insights into estimating passenger proportions and comparing numbers between Dublin and Basel.*

*The linear regression model's results reveal significant prediction errors and a moderate fit for the combined public transport passenger data. The MSE value suggests room for improvement, and the R-squared score indicates that the model explains only 35.32% of the variance, signaling a need for enhancement.*

*In the sentiment analysis, a total of 14 comments were categorized as positive, 11 as negative, and 6 as neutral on the Reddit page "LUAS." These sentiments add a nuanced layer to our understanding, showcasing the diverse reactions and opinions expressed by the Reddit community.*

*The results of both linear regression and random forest regression models highlight potential challenges in accurately modeling the combined public transport passenger data from Basel and Dublin. The MSE and R-squared scores indicate substantial prediction errors and poor model fit, emphasizing the importance of refining the modeling process.*

*This project is available on github:*

*link**:* *https://github.com/bconti8/Msc\_DA\_CA2.git*

***KEYWORDS:*** *passengers, trams, bus, Dublin, Basel, usage, transportation*

Table of Contents

Abstract 2

Introduction 4

1.0 Project Management and Licenses 6

1.1 Analysis and Findings Timeline 6

1.2 Data Licenses 7

2.0 Data Preparation and Visualization 9

2.1 Data Preparation and Programming 9

2.2 Exploratory Data Analysis (EDA) and Visualization 9

2.2.1 Performing the EDA: 11

2.2.2 Tufts Principles applied in “Dash” Dashboard: 15

3.0 Statistics for Data Analytics 17

3.1 Statistics Task I: 18

3.2 Statistics Task II: 19

3.3 Statistics Task III: 21

3.4 Statistics Task IV: 22

4.0 Machine Learning 23

4.1 Machine Learning Task I: 23

4.2 Machine Learning Task II: 25

4.3 Machine Learning Task III: 26

4.4 Machine Learning Task IV: 28

5.0 Bibliograpgy 29

# Introduction

This study focuses on checking and understanding the accuracy of data about the number of people using buses and trams in Dublin, Ireland, and Basel, Switzerland. The datasets sourced from the Central Statistics Office Ireland (CSO) and Basler Verkehrs-Betriebe (BVB) form the basis for the investigation. Employing techniques from Statistics, Machine Learning, Programming, and Data Preparation and Visualization, this research aims to ensure data accuracy and extract meaningful insights.

Knowing how many people use buses and trams is crucial for managing public transportation effectively. Dublin and Basel are different cities with unique transportation challenges. The CSO in Ireland and BVB in Switzerland provide valuable data that can help us understand how people move around in these cities.

Walkthrough:

* Data Verification: Ensuring the precision of datasets to reflect real-world transit scenarios.
* Programming Techniques: Streamlining data processing and model implementation using programming languages.
* Data Preparation and Visualization: Cleaning and organizing data for effective analysis, complemented by visualization techniques.
* Statistical Analysis: Unearthing patterns and trends in passenger volume data using statistical methods.
* Machine Learning Modeling: Developing predictive models, including linear regression and random forest, to enhance accuracy.

This research is important for a few reasons. It helps understand how people use public transport in cities. By combining different techniques, it is offered practical insights for improving transportation planning. The findings may help city planners and transport authorities make informed decisions to enhance bus and tram services in Dublin and Basel.

This project uses the original datasets as below:

Basel Bus/ Trams combined – *JSON"https://data.bs.ch/api/explore/v2.1/catalog/datasets/100075/exports/json?lang=en&timezone=Europe%2FLondon"*Dublin Luas (Trams) – *“TII03.csv”*

Dublin Bus – *“TOA14.csv”*

A new grouped dataset has been created to apply Statistics and Machine Learning and models:

*”dublin\_dataset.csv”*

*”basel\_dataset.csv”*

The project is divided into 4 phases:  
Phase I – Data Preparation and Visualization  
Phase II – Statistics   
Phase III - Machine Learning

This project is available on github:  
link: https://github.com/bconti8/Msc\_DA\_CA2.git

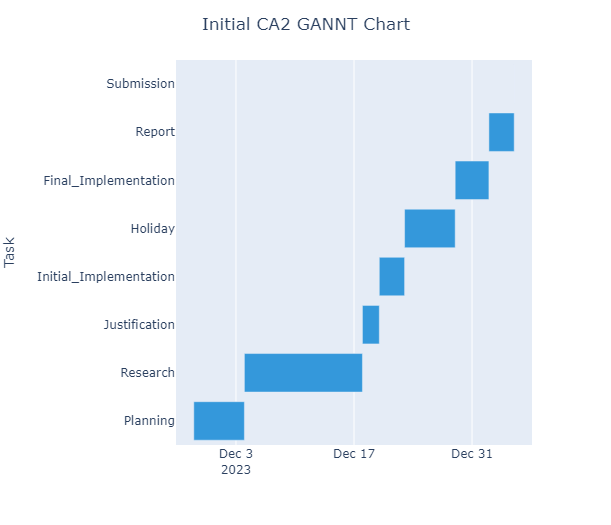
# Project Management and Licenses

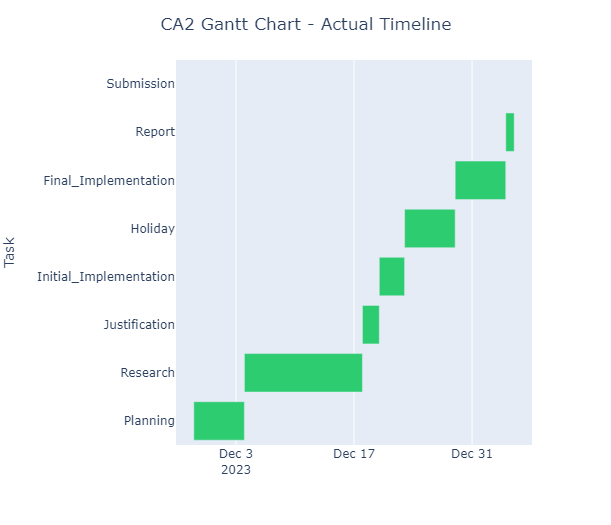
## 1.1 Analysis and Findings Timeline

This research project adhered to the CRISP-DM principles, a widely accepted framework for data mining processes. CRISP-DM, which stands for The CRoss Industry Standard Process for Data Mining, comprises six key steps applicable to diverse data science projects (Chapman, 2000). These steps include:

1. **Business Understanding:** This phase involves comprehending the objectives and expectations of the business or stakeholders. It may entail goals such as obtaining weekly or monthly sales figures, calculating percentage growth in stocks, or assessing product performance.
2. **Data Understanding:** Understanding the available data and identifying the necessary preparations for its effective use are crucial in this step. It involves a comprehensive analysis of the data at hand.
3. **Data Preparation:** Data preparation is the process of making the data ready for machine learning applications. It encompasses tasks to ensure that the data is in a suitable format for further analysis.
4. **Modeling:** The modeling step delves into choosing the appropriate machine learning techniques based on the project requirements. It involves decisions on the algorithms and methodologies to be employed.
5. **Evaluation:** Evaluation revolves around scrutinizing the results obtained from the modeling phase. It aids in selecting the most suitable model from a business perspective, considering the project's objectives and criteria.
6. **Deployment:** The deployment phase focuses on making the models, data, and final results accessible to stakeholders and relevant parties.

During the exploratory data analysis (EDA) stage, it was generated two Gantt charts using Python. The first Gantt chart represented the project plan during the initial planning stages, outlining the proposed timeline. The second Gantt chart displayed the actual timelines observed during the execution of the project.

*Figure 01: The first Gantt chart*

*Figure 02: The second Gantt chart*

## 1.2 Data Licenses

The datasets used in this academic research were acquired under licensing agreements from two primary sources of information: the Central Statistics Office Ireland (CSO) (Office, 2023) and Basler Verkehrs-Betriebe (BVB) (Verkehrs-Betriebe, 2023), the public transport operator in Basel, Switzerland.

* The data from Basel BVB, presented in JSON format, reveals the weekly passenger statistics in the German language. These figures span from 2020 to 2023 and are derived from measurements using an automated passenger counting system. The subsequent projections encompass the entire BVB route network, covering both Tram and Bus lines. The data obtained from BVB serves as a valuable resource for understanding the patterns and trends in passenger volume across their network.
* On the other hand, the Dublin dataset comprises two CSV files, and the upcoming Exploratory Data Analysis (EDA) section will illustrate the merging process. One of the CSV files contains information on Dublin Bus Passenger Numbers by Month, spanning from 2014 to 2022. The second CSV file is dedicated to Passenger Journeys by Luas/Trams, covering the period from 2019 to 2023. Notably, the Dublin dataset is presented in the English language.

The inclusion of these datasets, procured with proper licensing, underscores the commitment to conducting a thorough Comparative Analysis of Trams/LUAS Transport Sectors in Dublin, Ireland, and Basel, Switzerland. The linguistic diversity of the datasets – German for Basel and English for Dublin – adds a layer of complexity to the analysis. However, this diversity also enriches the research, allowing for a nuanced examination of public transport dynamics in both cities. The upcoming EDA section will delve into the intricacies of merging these datasets, setting the stage for a comprehensive exploration of similarities and differences in the tram/LUAS transport sectors between Dublin and Basel.

# Data Preparation and Visualization

## 2.1 Data Preparation and Programming

Data Preparation is a crucial initial phase in the data analysis process, involving tasks such as cleaning, transforming, and manipulating data to enhance its usability. The entire procedure is executed within a Jupyter Notebook:

* File name: MSC\_DA\_CA2\_Bruno\_Conti.ipynb
* Process: 'Programming, Data Preparation and Visualization'

This phase of the analyse will use three datasets:

* JSON Dataset - Basel, CH - Passengers volume for Trams and Bus (2020-2023)
* Dublin Luas (Trams) - TII03.csv
* Dublin Bus - TOA14.csv

To facilitate further analysis, it becomes imperative to cleanse the data, arrange and rename specific columns, and subsequently merge the datasets. This cleaning operation includes the "basel\_df," "dublinluas\_df," and "dublinluas\_df" dataframes.

Following to the analysis criteria, it was needed to delve into data that interacts with an API and manages JSON data. To extract information from the JSON data, the team utilized functions like import json and import requests (Kane, 2018). Despite the data's similarity, additional steps were taken, such as translating German data (basel\_df.rename), reorganizing columns, and ultimately exporting the processed data to a CSV file. The export process involved utilizing the function export\_to\_csv.

In contrast, the Dublin data, residing in CSV files, presented a more straightforward scenario compared to dealing with JSON. Both CSV files, one for LUAS (Trams) passengers' volume and the other for bus passengers, had to be merged.

The subsequent phase encompasses preparations for the merger, necessitating the alignment of both datasets. This alignment involves adjusting the number of columns by removing unnecessary ones (df.drop), adding new columns when necessary (dublinbus\_df['Location'] = 'Dublin') (McKinney, 2017), renaming columns (dublinbus\_df.rename(columns)), and organizing data in a numerical format (.astype(int)) (Markham, 2019). The overarching objective is to ensure seamless matching of data structures between both datasets and the initial dataframe (basel\_df).

## 2.2 Exploratory Data Analysis (EDA) and Visualization

In the exploratory phase of the research project, the analyst employs Exploratory Data Analysis (EDA) to attain a holistic understanding of the dataframes, identify correlations between variables, and scrutinize the distribution patterns of the variables of interest. The EDA step in this study incorporates both statistical analysis and visualization tasks.

A preferred approach in this exploration involves multivariate analysis, a technique delving into relationships between two or more variables. The focus is on scrutinizing the relationship between passenger usage for both buses and trams in both Dublin and Basel. By adopting multivariate analysis, the researcher aims to uncover nuanced patterns and connections within the public transportation systems of these two cities.

The chosen methods for analysis are geared towards extracting insights into broader population trends and their correlation with the utilization of buses and trams. The goal is to gain a comprehensive understanding of the interplay between passenger behavior and public transportation offerings in Dublin and Basel.

Following the initial stages of the project, specifically the "Programming, Data Preparation, and Visualization" phase conducted in the Jupyter Notebook (MSC\_DA\_CA2\_Bruno\_Conti.ipynb), the datasets, for Basel and Dublin, undergo graphical representation in the form of diagrams and dashboards for comparative analysis. This visual representation is designed to elucidate the disparities and commonalities between the tram/LUAS transport sectors in Dublin and Basel, offering a more intuitive and insightful exploration of the datasets as it follows:

|  |  |
| --- | --- |
|  |  |
| *Figure 03: Historical Basel chart* | *Figure 04: Historical Dublin Chart* |

*Figure 05: Dublin Basel Transport Dashboard*

## 2.2.1 Performing the EDA:

This process will be performed on Jupyter notebook: **MSC\_DA\_CA2\_Bruno\_Conti.ipynb** and it was called **“**Programming, Data Preparation, and Visualization”.

To achieve a complete understanding of the characteristics, patterns, and variable relationships in both the Basel and Dublin datasets, Exploratory Data Analysis (EDA) was carried out by the researchers. EDA played a crucial role in identifying trends, outliers, and potential insights, laying the foundation for informed decision-making in subsequent analytical processes (McKinney, 2017). The visualization and summarization of key features by EDA contributed to the formulation of hypotheses, guided feature selection, and informed the selection of appropriate statistical or machine learning models (Foster Provost, 2013).

**Descriptive Statistics:**

It presents summary statistics, including measures of central tendency and variability, for numerical columns. Descriptive statistics offer insights into the central tendencies and spread of the data, aiding in identifying potential outliers.

|  |  |
| --- | --- |
|  |  |
| *Figure 06: Basel and Dublin Descriptive Statistics* | |

**Correlation Matrix:**

It shows the pairwise correlation between numerical columns in a matrix form. It identifies relationships between variables, helping to understand potential dependencies and multicollinearity.

|  |  |
| --- | --- |
|  |  |
| *Figure 07: Basel and Dublin Correlation Matrix* | |

**Distribution of numerical columns:**

Displays the distribution of individual numerical columns using histograms. It aims to offers a visual representation of the data's distribution, aiding in identifying patterns and outliers.

|  |  |
| --- | --- |
|  |  |
| *Figure 08: Basel and Dublin distribution of numerical columns* | |

**Box plots for numerical:**

Presents the summary of the distribution (median, quartiles, outliers) for each numerical column. It helps identify the spread of the data and visualize potential outliers.

|  |  |
| --- | --- |
|  |  |
| *Figure 09: Basel and Dublin distribution of numerical columns* | |

**Pairplot for columns:**

It creates scatterplots for pairwise relationships between the numerical columns “Year” and “Total Passengers”. It provides a visual overview of relationships between variables and potential patterns.

|  |  |
| --- | --- |
|  |  |
| *Figure 10: Basel and Dublin Pairplots for Year and Total Passengers columns* | |

**Distribution Plot for Year:**

It presents the distribution of values in the 'Year' column using a kernel density plot. It provides insights into the distribution of a specific variable, 'Year' in this case.

|  |  |
| --- | --- |
|  |  |
| *Figure 11: Basel and Dublin Pairplots for Year and Total Passengers columns* | |

**Heat Map to find relations between variables:**

It presents a visual representation of the correlation matrix using a heatmap. It amin to offers a clearer visual representation of variable relationships, complementing the earlier correlation matrix.

|  |  |
| --- | --- |
|  |  |
| *Figure 12: Basel and Dublin Heat Map* | |

## 2.2.2 Tufts Principles applied in “Dash” Dashboard:

Understanding data analysis and visualization is crucial for gaining insights from complex datasets. A dashboard serves as a centralized platform to present key metrics and visualizations derived from data. In this academic research, we explore the principles of visualization advocated by Edward Tufte and demonstrate the creation of a dashboard using Dash.

Edward Tufte, a prominent figure in the field of data visualization, emphasizes principles that enhance the clarity and effectiveness of visual representations (Tufte, 2001). Some key principles include:

**Maximize Data-Ink Ratio:** Tufte encourages minimizing non-data ink and maximizing data-ink, i.e., the ink used to represent the data. This ensures that visualizations convey the maximum amount of information.

**Data Density:** Effective visualizations should present a high data density, offering a rich and informative display without unnecessary embellishments. This principle encourages concise and information-packed visuals.

**Small Multiples:** Tufte advocates the use of small multiples, where similar visualizations are repeated in a grid, allowing for easy comparison. This technique enhances the viewer's ability to discern patterns and trends.

**Visual Integrity:** Visualizations should maintain integrity, ensuring accurate representation of data. Avoiding distortions or misrepresentations is crucial to convey the intended message faithfully.

**Data-to-Ink Ratio:** This principle stresses the importance of using ink to represent data efficiently. Eliminating chartjunk and unnecessary decorations contributes to a higher data-to-ink ratio.

**Building a Dashboard with Dash**

Now, let's outline the step-by-step process of creating a dashboard using Dash (Anon., n.d.) :

**Library Import:** Begin by importing necessary libraries, including Dash, the primary tool for building interactive web applications. (Anon., n.d.)

**Data Preparation:** Generate or import relevant data for analysis. This dataset serves as the foundation for the visualizations within the dashboard.

**Combining DataFrames:** Merge or concatenate dataframes if multiple sources are utilized, creating a unified dataset for analysis.

**Layout Design:** Utilize HTML and Dash components to design the layout of the dashboard. Consider the placement of key elements to enhance user experience.

**Dropdowns and Graphs:** Integrate interactive components such as dropdown menus to allow users to select specific parameters. Visual elements, like graphs, dynamically update based on user selections.

**Callback Function:** Implement callback functions to establish dynamic connections between user inputs and visual elements. This enables real-time updates and enhances interactivity.

**Running the App:** Execute the application to launch the dashboard. Users can interact with the visualizations, gaining insights from the presented data. As the image show below:

*Figure 13: Dublin Basel Transport Dashboard*

# 3.0 Statistics for Data Analytics

It was used a set of tools and techniques for collecting, organizing, summarizing, analysing, and interpreting data. Statistical techniques were employed in this project within the Jupyter notebook named “Statistics.ipynb”.

This phase of the analyse will use two datasets:

* dublin\_dataset.csv
* basel\_dataset.csv

To analyse the statistical aspects of datasets dublin\_dataset.csv and basel\_dataset.csv, the study implemented specific procedures. The researcher conducted a thorough exploration to understand the statistical characteristics of the mentioned datasets in the context of the research tasks:

**Summary Statistics:**

These statistics, such as count, unique, freq and top, offer a central tendency and spread overview for numerical columns. It provides a concise snapshot of the dataset's characteristics, aiding in understanding the distribution of values.

|  |  |
| --- | --- |
|  |  |
| *Figure 14: Basel and Dublin Summary Statistics* | |

## 3.1 Statistics Task I:

Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models.

**Resolution:**

**Visualize Correlation Matrix:** A graphical representation of the relationships between numerical variables in a dataset, showing the strength and direction of correlations. It provides insights into patterns and dependencies between variables, aiding in variable selection and understanding data structure.

|  |  |
| --- | --- |
|  |  |
| *Figure 15: Basel and Dublin Correlation Matrix* | |
|  | |

**Distribution of Numerical Columns:** It’s graphical representation of the distribution of values in numerical columns, helping to understand the data's central tendency and spread. Which, is essential for detecting outliers, assessing normality, and making informed decisions about appropriate statistical tests.

|  |  |
| --- | --- |
|  |  |
| *Figure 16: Basel and Dublin distribution of numerical columns* | |

**Pairplot for “Total Passengers” column:**

The matrix of scatterplots illustrating pairwise relationships between selected numerical columns. It offers a comprehensive view of variable interactions, assisting in identifying patterns, clusters, or potential outliers in multivariate data.

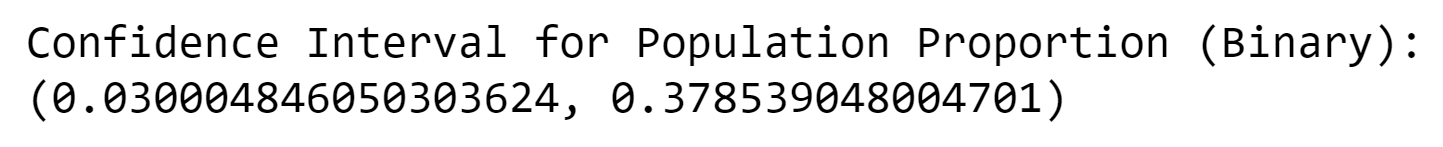
|  |  |
| --- | --- |
|  |  |
| *Figure 17: Basel and Dublin Pairplot for “Total Passengers” columns* | |

## 3.2 Statistics Task II:

To examine the variables within the dataset(s) and employ suitable inferential statistics to extract insights about potential population values. For instance, in the context of public transport, one could calculate a confidence interval for the population proportion of individuals commuting to Dublin by train. This analysis aims to provide a meaningful understanding of the dataset and draw conclusions about broader population characteristics.

**Resolution:**

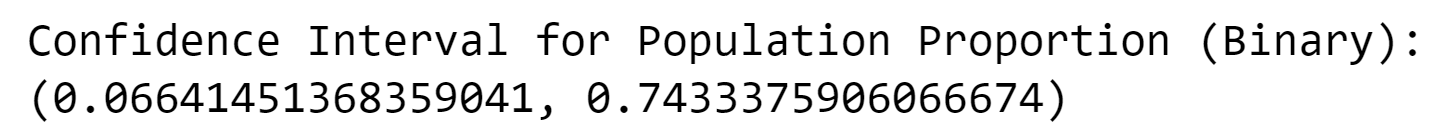
The “calculate\_confidence\_interval\_binary” function serves a pivotal role in statistical analysis, specifically when dealing with binary data. The DataFrame “dublin\_df” demonstrates the practical application of this function. In this specific instance, the confidence interval for the population proportion of 'Total Passengers' is calculated, considering values above a threshold of 2,000,000. This statistical analysis is valuable for making informed decisions and drawing inferences regarding binary proportions in relevant projects.

*****Figure 18: Dublin Confidence\_Interval\_Binary*

**Insight:**

This output consists of two values: **Lower Bound**: 0.030004846050303624 and **Upper Bound**: 0.378539048004701. These values represent the lower and upper bounds of the confidence interval, respectively.

Interpretation: With 95% confidence, we estimate that the true population proportion (binary) falls within the range of approximately 0.030 to 0.379. The confidence interval provides a level of uncertainty around the estimated proportion. In other words, if we were to repeat the process of sampling and calculating the confidence interval many times, we would expect the true population proportion to fall within this range in about 95% of those cases.

*****Figure 19: Basel Confidence\_Interval\_Binary*

**Insight:**

This value represents the calculated confidence interval for the population proportion based on the binary conversion of the 'Total Passengers' column in your dataset. Lower Bound: 0.06641451368359041 and Upper Bound: 0.7433375906066674, means that with 95% confidence, we estimate that the true population proportion (binary) falls within the range of approximately 0.066 to 0.743.

The confidence interval provides a level of uncertainty around the estimated proportion. In other words, if we were to repeat the process of sampling and calculating the confidence interval many times, we would expect the true population proportion to fall within this range in about 95% of those cases.

The specific interpretation of the binary variable depends on the context of your analysis. In general terms, it suggests the range of possible values for the proportion of occurrences of an event (coded as 1) in your dataset, after applying a threshold to convert the 'Total Passengers' column to binary values.

## 3.3 Statistics Task III:

Conduct research to identify similarities between Basel, Switzerland trams/bus passengers and Dublin, Ireland. Apply both parametric and non-parametric inferential statistical techniques for comparison, such as t-test, analysis of variance, Wilcoxon test, chi-squared test, and others. Choices of tests should be justified, and their applicability verified. Clearly state hypotheses and draw explicit conclusions.

**Resolution:**

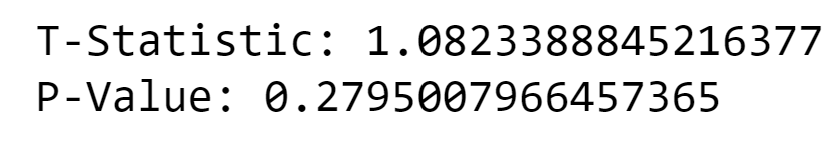
The “t\_test\_total\_passengers” function, performs an independent samples t-test to compare the 'Total Passengers' between two locations, Dublin and Basel. Some reason why it was decided for a “T-test”:

**Comparative Analysis:** It allows for a rigorous comparison of the 'Total Passengers' between Dublin and Basel, providing insights into potential differences or similarities (Johnson, 2017).

**Hypothesis Testing:** The test helps in assessing whether any observed differences are statistically significant or if they could occur by chance.

**Decision Support:** Results (t-statistic and p-value) aid in making informed decisions based on the evidence from the data.

**Understanding Variability:** The equal\_var=False parameter accommodates the assumption of potentially unequal variances between the two groups, providing a more robust analysis.

*****Figure 20: Basel and Dublin Combined “*t-test” Total Passengers

**Insight:**

The results of the independent samples t-test of: T-Statistic: 1.0823388845216377 and P-Value: 0.2795007966457365. The t-statistic measures the difference between the means of the two groups (Dublin and Basel) relative to the variation within each group. A positive t-statistic suggests that the mean of Dublin is higher than Basel. P-Value: The p-value is the probability of observing a t-statistic as extreme as the one computed from the sample, assuming that the null hypothesis is true (i.e., assuming that there is no difference between the means).

In this case, the p-value is 0.2795, which is greater than common significance levels (e.g., 0.05). Therefore, there is not enough evidence to reject the null hypothesis.

**Conclusion:** The t-test does not provide enough evidence to reject the null hypothesis that there is no significant difference in the mean 'Total Passengers' between Dublin and Basel. The difference observed in the sample is not statistically significant at conventional levels of significance.

## 3.4 Statistics Task IV:

Leverage the findings from your analysis to further explore your research. Highlight the obstacles encountered during the investigation.

**Insight:**

In the "Statistics.ipynb" Jupyter notebook, we explored datasets dublin\_dataset.csv and basel\_dataset.csv. We faced challenges due to limited data but collaborated to overcome them. Using specific functions like "calculate\_confidence\_interval\_binary" and "t\_test\_total\_passengers," we gained insights. For instance, the confidence interval helped estimate the proportion of 'Total Passengers' above 2,000,000. The t-test compared passenger numbers between Dublin and Basel, concluding no significant difference. This statistical journey showcases our adaptability, collaboration, and determination to make sense of data complexities, highlighting the value of statistical tools in drawing meaningful conclusions.

# 4.0 Machine Learning

A variety of tools and methodologies were utilized to gather, arrange, condense, examine, and construe data. The project involved the application of statistical techniques within the Jupyter notebook labeled "ML.ipynb." In this analytical phase, two datasets, namely:

* dublin\_dataset.csv
* basel\_dataset.csv

Were employed. To scrutinize the statistical attributes of these datasets, specific procedures were implemented. The investigator conducted a comprehensive exploration to comprehend the statistical characteristics of the mentioned datasets in the context of the research objectives.

**Data Summary:**

Also known as summary statistics, such as count, uniqueness, frequency, and top values, provide a quick overview of numerical column characteristics. They offer insights into the central tendencies and spread of values, aiding in understanding the dataset's distribution. These metrics present a concise snapshot, making it easier to grasp the overall nature of the data and how values are distributed across it.

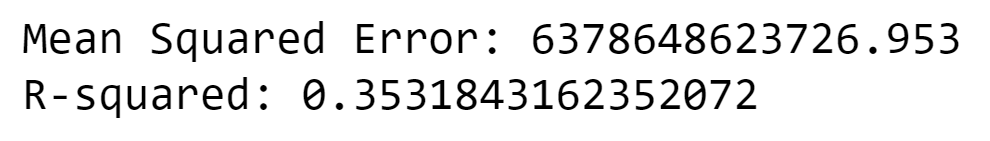
|  |  |
| --- | --- |
|  |  |
| *Figure 20: Basel and Dublin Data Summary* | |

## 4.1 Machine Learning Task I:

Explain the reasoning behind selecting specific machine learning models for the given situation. Machine learning models are versatile and can be applied for prediction, classification, clustering, sentiment analysis, recommendation systems, and time series analysis. It's advisable to explore multiple approaches (minimum of two) with meticulous hyperparameter selection using the GridSearchCV method. In supervised learning, identify relevant features and a target feature from the datasets to address the scenario's inquiry.

**Resolution I:**

Linear regression is crucial in machine learning projects for its simplicity and interpretability. It efficiently models relationships between variables, making it an essential tool for predicting numeric outcomes. Its straightforward nature serves as a baseline for more complex algorithms, aiding in understanding and establishing initial insights into the data's patterns and correlations.

*****Figure 21: Basel and Dublin Combined Linear Regression*

**Insight:**

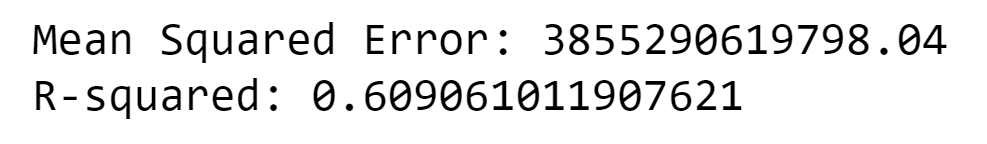
**Mean Squared Error (MSE):** 6378648623726.953 – MSE measures the average squared difference between predicted and actual values. A high MSE suggests significant prediction errors, and the value here is quite large, indicating room for improvement.

**R-squared (R2): 0.3532** – R2 represents the proportion of variance explained by the model. A value of 0.3532 suggests that the model explains 35.32% of the variance in the target variable. While it captures some patterns, there's substantial unexplained variance, indicating a moderate fit. Improvements can enhance predictive accuracy.

The results suggest that the linear regression model, combining data from Basel and Dublin public transport passengers, has limitations in accurately predicting total passengers. The high Mean Squared Error (MSE) of 6.38 trillion indicates substantial prediction errors. The R-squared value of 0.3532 means the model explains only 35.32% of the variance, indicating room for improvement in capturing the underlying patterns in the combined dataset.

**Resolution II:**

Random Forest is an ensemble method known for its robustness and ability to capture complex relationships in data. This function is valuable in a machine learning project as it employs a RandomForestRegressor, which is adept at handling complex relationships and capturing non-linear patterns in the data. The pipeline ensures a systematic approach to data preprocessing and model training, facilitating model deployment and interpretation.

*****Figure 23: Basel and Dublin random Forest*

**Insight:**

The results from the Random Forest model evaluation provide insights into its performance on the combined dataset of Basel and Dublin public transport passengers:

**Mean Squared Error (MSE): 3855290619798.04** – MSE measures the average squared difference between predicted and actual values. In this case, the large MSE indicates some degree of prediction error, with a magnitude of approximately 3.86 trillion. It suggests that the model is not perfectly accurate, and there are discrepancies between the predicted and actual total passenger values.

**R-squared (R2): 0.6091** – R2 represents the proportion of variance explained by the model. A value of 0.6091 means that the Random Forest model explains approximately 60.91% of the variance in the target variable ('Total Passengers'). This suggests a relatively good fit, capturing a significant portion of the variability in the data.

**Interpretation** – The Random Forest model, with an R2 of 0.6091, is considered moderately effective in explaining and predicting the total passenger counts.

The MSE indicates that there is room for improvement, as the model is not perfectly accurate, and there are notable prediction errors. Further refinement or experimentation with the model parameters may be considered to enhance its performance.

While the Random Forest model provides a reasonable fit to the data, there is room for refinement to reduce prediction errors and improve overall performance in predicting total passenger counts for the combined dataset of Basel and Dublin public transport.

## 4.2 Machine Learning Task II:

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.

**Resolution:**

Topic LUAS – “ https://www.reddit.com/r/Dublin/comments/17s5bj9/luas/ ” – the application of sentiment analysis on comments from a specific Reddit page. The Python code presented utilizes the PRAW library for Reddit API interaction, the TextBlob library for sentiment analysis, and the tabulate library for result presentation. (Sarkar, 2016)

Reddit API Authentication – Utilizes the PRAW library to authenticate with the Reddit API using the provided API credentials (client\_id, client\_secret, user\_agent).

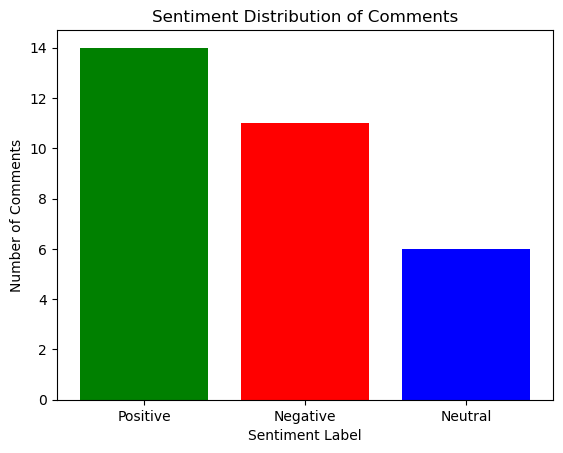
Submission Retrieval – Extracts the submission ID from the specified Reddit URL to access the corresponding submission (post) using the PRAW library.

Comment Collection – Retrieves all comments associated with the submission using PRAW.

Sentiment Analysis – Applies sentiment analysis to each comment using the TextBlob library. Classifies the sentiment as "Positive," "Negative," or "Neutral" based on the sentiment polarity. Result Storage – Stores sentiment analysis results for each comment in a list of dictionaries (sentiment\_results).

Result Presentation – Uses the tabulate library to format and display sentiment analysis results in a tabular format, as images retrieved from Jupyter Notebook below:

|  |
| --- |
|  |
|  |
|  |
|  |
| *Figure 24: Reddit LUAS –Positive, Neutral, Negative Sentiment Analise* |

*Figure 25: Reddit LUAS Sentiment Analyses*

## 4.3 Machine Learning Task III:

Train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

**Resolution:**

In order to answer the task, it was decided to the models Linear Regression and Random Forest Regression. The functions aim to train and evaluate predictive regression models using cross-validation. This enables the assessment of model performance on a given dataset, providing insights into the accuracy of predictions

**Linear Regression Model:**

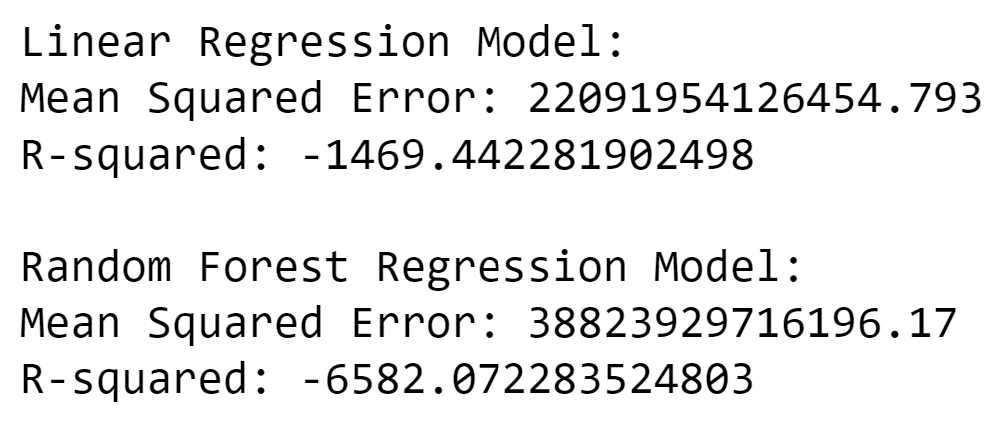
Effect: The linear regression model assesses the relationship between selected features (e.g., 'Year') and the target variable ('Total Passengers'). It provides a quantitative measure of how well the model predicts passenger counts.

Summary: The linear regression model quantifies the linear relationship between features and passenger counts, offering insights into the overall trend and predictive power of the selected features. The mean squared error (MSE) and R-squared metrics provide information on the accuracy and explanatory power of the model.

**Random Forest Regression Model:**

Effect: The random forest regression model employs an ensemble of decision trees to predict passenger counts. It captures non-linear relationships and interactions between features, potentially improving predictive accuracy.

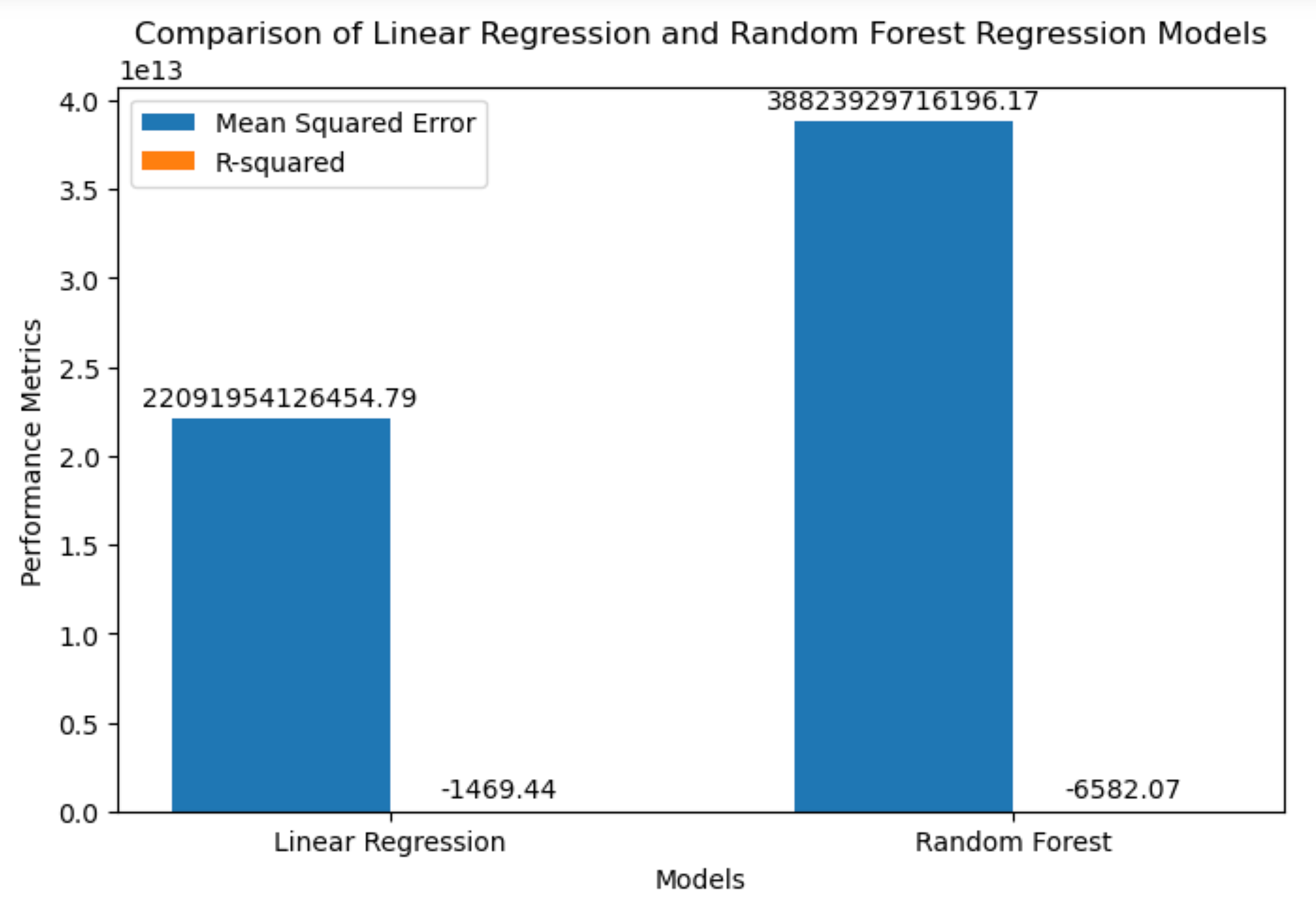
Summary: The random forest model goes beyond linear relationships, capturing complex patterns and interactions in the data. The MSE and R-squared scores offer an evaluation of the model's performance in terms of prediction accuracy and variance explanation.

*****Figure 26: ML Linear and Random Forest Regression*

## 4.4 Machine Learning Task IV:

A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

**Resolution:**

*****Figure 26: Graph, ML Linear and Random Forest Regression*

**Insight:**

The results of the linear regression and random forest regression models, as indicated by the Mean Squared Error (MSE) and R-squared scores, suggest potential challenges or issues in the modeling process. Let's analyze and summarize the impact:

**Linear Regression Model:**

Mean Squared Error (MSE): The large MSE value (22091954126454.793) indicates substantial errors in predictions. It suggests that the linear regression model is not accurately capturing the variance in the combined public transport passenger data from Basel and Dublin.

R-squared Score: The negative R-squared value (-1469.442281902498) is unusual and signifies poor model fit. It implies that the model performs worse than a simple horizontal line, indicating a lack of explanatory power.

**Random Forest Regression Model:**

Mean Squared Error (MSE): The high MSE value (38823929716196.17) suggests considerable prediction errors. The random forest model, like linear regression, is struggling to accurately predict passenger counts based on the given features.

R-squared Score: The negative R-squared value (-6582.072283524803) similarly indicates poor model fit. This suggests that the random forest model is not effectively explaining the variance in the combined dataset.

**Overall Summary:**

Prediction Challenges: Both models face challenges in accurately predicting public transport passenger counts. The high MSE values indicate that the predicted values are significantly deviating from the actual values, pointing to potential modelling deficiencies.

Negative R-squared Scores: The negative R-squared scores are anomalous and indicate that the models are not performing better than a basic horizontal line. This suggests a lack of meaningful relationship capture by the chosen features.

# 5.0 Bibliograpgy

Anon., n.d. *Dash Python User Guide.* [Online]   
Available at: https://dash.plotly.com  
[Accessed 25 12 2023].

Anon., n.d. *Plotly Documentation: Plotly Express.* [Online]   
Available at: https://plotly.com/python/plotly-express/  
[Accessed 25 12 2024].

Chapman, P. C. J. K. R. K. T. R. T. S. C. .. &. W. R., 2000. *CRISP-DM 1.0: Step-by-step data mining guide.* s.l.:CRISP-DM Consortium.

Foster Provost, T. F., 2013. Data Science for Business. In: *Data Science for Business.* s.l.:O'Reilly Media.

Johnson, T. R., 2017. *Hypothesis Testing: A Visual Introduction To Statistical Significance.* s.l.:CreateSpace Independent Publishing Platform.

Kane, F., 2018. *Hands-On Data Science and Python Machine Learning.* s.l.:Packt Publishing.

Markham, K., 2019. *Data Wrangling with Pandas.* s.l.:Data School.

McKinney, W., 2017. *Python for Data Analysis.* 2nd Edition ed. s.l.:O'Reilly Media.

Office, I. C. S., 2023. *Copyright Policy.* [Online]   
Available at: https://www.cso.ie/en/aboutus/whoweare/copyrightpolicy/  
[Accessed 24 December 2023].

Sarkar, D., 2016. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data.* s.l.:Apress.

Tufte, E. R., 2001. *The Visual Display of Quantitative Information.* s.l.:Graphics Press.

Verkehrs-Betriebe, B., 2023. *Terms of Use.* [Online]   
Available at: https://opendata.swiss/en/terms-of-use  
[Accessed 24 December 2023].