**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *MSc in Data Analytics* |
| **Assessment Title:** | *MSC\_DA\_CA2* |
| **Lecturer Name:** | *Sam Weiss*  *Taufique Ahmed*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Leda Mercedes Gale Torre |
| **Student Number:** | 2023205 |
| **Assessment Due Date:** | 07/01/2024 |
| **Date of Submission:** | 07/01/2024 |

**Declaration**

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

**Index**

[Abstract 1](#_Toc155544853)

[Introduction 1](#_Toc155544854)

[Approaches 2](#_Toc155544855)

[Programming 2](#_Toc155544856)

[Data Preparation and Visualisation 3](#_Toc155544857)

[Statistics 5](#_Toc155544858)

[Machine Learning 8](#_Toc155544859)

[References 11](#_Toc155544860)

[Appendixes 12](#_Toc155544861)

## Abstract

This report outlines a detailed analysis of Ireland's transport sector, comparing it with other European countries. The study involves Programming, Statistics, Machine Learning, and Data Preparation & Visualization.

Pyton tools in a Jupyter Notebook have been used focusing on coding efficiency. Data structures such as csv, tsv, and API were needed to get enough data to achieve the project’s goals. The corresponding Exploratory Data Analysis (EDA) was performed followed by the adequately cleaned and engineered data for Machine Learning. An interactive dashboard was developed to present the insights. Descriptive statistics was implemented to analyze the datasets, as well as inferential statistics for hypothesis testing to compare the different data sources. Supervised Machine learning models (e.g., Linear Regression, Polynomial Regression, Decision Tree, and Random Forest) applying hyperparameter tuning and cross-validation were generated pretending to get an efficient model to make predictions. Evaluation metrics, such as R2 scores, are used to assess model performance.

Data collected from Reddit was gathered to analyze opinions on the public services in the countries included in the study. Sentiment analysis techniques categorized the opinions.

The word count of the project is 3,140 words with figures in the appendix as support of the text and including the needed referencing.

## Introduction

The present CA report aims to detail the step-by-step work done in the chosen dataset, describing each section included in the analysis.

According to the guides of the project, the main task corresponds to gathering datasets from Ireland’s transport to compare the transport sector with other countries of choice.

The study has been based on two different datasets, the first one was taken from Ireland’s open data portal (<https://data.gov.ie/organization/transport-infrastructure-ireland>), and the second one was taken from (<https://ec.europa.eu/>) which includes records from several different European countries.

The programming language used is Python, its code is clear to read and write and facilitates getting results in fewer lines than other languages. Libraries such as Pandas, NumPy, seaborn, and others helped achieve the project goals.

In this project 4 approaches were taken: Programming, Statistics, Machine Learning, and Data Preparation &Visualisation.

The stated approaches are amplified in the following section.

## Approaches

### Programming

* **Programming**

The project's exploration was elaborated by implementing Python tools to develop and execute the entire analysis workflow within a Jupyter Notebook environment to fulfill the analysis requirements. Libraries like the previously mentioned and more specific ones such as scipy.stats, sklearn, nltk, and dash were necessary to extract some knowledge from the raw data.

Functions like dataframe\_uniques, replace\_values, label\_graph, update\_dataframe, general\_table, save\_results, show\_results, sentiment\_analysis, getPolarity, result\_analysis, and is\_english were created to avoid repetition on the code and make it easier to extract information from the data. All aspects of the analysis, from data manipulation to model training and evaluation, should be meticulously executed within the confines of a Jupyter Notebook. The project aims to achieve a transparent, and reproducible analysis, aligning with best practices in the field of data science and programming.

* **Data structures**

Data stored in formats such as CSV, TSV, and API was used in the development of this project

* **Documentation**

Prioritizing code reproducibility and transparency, a clear and consistent naming convention to declare the variables, and functions was used, and corresponding inline comments and markdown cells were included to aid in maintaining the logic and purpose behind each code segment. Decisions such as splitting the dataset, re-indexing columns, renaming them, and creating functions are some actions taken to keep the code readable.

* **Testing & Optimisation**

Functions like save\_results and add\_years were tested to confirm that the results thrown from these functions were desired. In save\_results data for test, train score, and test score were stored, and after this values were checked to see if they were stored for the code given. In add\_years the data frame result was checked in shape to have the number of rows for the given number of years per month.

* **Data manipulation**

In the manipulation and processing of CSV format data Pandas library is versatile has rich functionality and is designed for in-data processing for moderate-size files, other libraries like Dask are designed for parallel and distributed computing and is suitable when managing more extensive datasets. The data used in the project fits comfortably into memory which makes Pandas an appropriate choice, it is well-documented and user-friendly, and it has a large community which makes it easier to find resources and support.

In the manipulation and processing of API Data the requests with Pandas, allow you to interact with APIs by sending HTTP requests (GET, POST, etc) and receiving responses, this technique was chosen over some other options like PyTorch and JSON for its simplicity and effectiveness, for example, once retrieved the API, Pandas excels in structuring and manipulating data, making easier to perform various data analysis. PyTorch and JSON are more suitable for more complex structures that need custom processing.

### Data Preparation and Visualisation

* **Process of acquiring the data**

The acquisition of the datasets needed was challenging and very time-consuming. There are some delicate areas regarding privacy, and ethical considerations that make privacy, and licensing terms are essential components of responsible and impactful data-driven research and acquisitions.

On the positive side, extensive research can lead to great discoveries. Having access to diverse datasets allows for the development of more accurate models.

Ireland’s dataset was collected from the data.gov.ie which has an Open Data Licence held in Feb – Mar 2015. It states that under the CC-BY Licence, users must acknowledge the source of the information in their product or application (<https://data.gov.ie/pages/opendatalicence>).

Other countries from Europe dataset was collected from Eurostat, which has a policy of encouraging free re-use of its data, both non-commercial and commercial purposes (https://ec.europa.eu/eurostat/about-us/policies/copyright).

* **Exploratory Data Analysis**

Exploratory Data Analysis (EDA), involves exploring and summarizing data to identify patterns and relationships.

There are several goals for an EDA, which are: To determine if there are any problems with your dataset, to determine whether the questions you are asking can be answered by the data you have, and to develop a sketch of the answer to your questions. (Peng & Matsui, 2016)

Libraries such as pandas, seaborn, matplotlib.pyplot, and NumPy were imported to initiate the exploration. Ireland’s dataset “TII03” was imported, using the method .head() and .shape it could be identified that the dataset originally had 8 features with 768 observations.

The columns “STATISTIC”, “Statistic Label” and `UNIT` were considered unnecessary, hence they were dropped.“TLIS(W1)” and `Week` contain the same information making one redundant. The feature “Week” was split into 2 “Year” and “Week\_Year”.

Null and duplicate revisions showed the presence of 9 values missing in the attribute “VALUE” corresponding to Week 53 in the non-leap years therefore these missing values were dropped.

The columns were renamed, before proceeding to develop the unique() and nununique() revision. After, the following conclusions were made: There are records for 5 years, the attribute “Luas lines” needed to be separated into df\_general (“All Luas lines”), and df\_expanded (“Green line” & “Red line”).

Data visualization provides tools to facilitate the understanding of trends and/ or patterns. Edward Tufte wrote what is probably the most relevant book about data visualization. The 6 principles are: 1. The representation of numbers should be directly proportional to the numerical quantities represented, 2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity, 3. Show data variation, not design variation, 4. In time series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units, 5. The number of information-carrying dimensions depicted should not exceed the number of dimensions in the data, 6. Graphics must not quote data out of context. (Tufte, 2007)

The graphic “Passengers using the Luas Services by Year” shows the number of passengers over the years. The plot which is clear, detailed and thoroughly labeled, shows the data variation over the years included in the dataset incorporating mainly the 2nd and 3rd of Tufte’s principles. This is an interactive plot that updates itself every time you change in the dropdown list which year you are looking to visualize (See Appendix Figure 1).

Another interactive plot was elaborated and it shows not only the passengers using the Luas services in general, but also the section of them that uses the Red Line or Green Line or the three cases (See Appendix Figure 2).

The same techniques mentioned above were used in the second dataset (dropping of columns, revision of missing and duplicated values, renaming of columns, etc.) some of the relevant insights found in the process were that the dataset contains records in two different units “MIO\_PKM” and “THS\_PAS” after noticing this the values under the unit of “MIO\_PKM” were dropped. Also, a dataset with codes was used to join it to the Europe dataset to identify what the “Country codes” stand for. Initially, the countries selected for the analysis were: Spain, France, Italy, Denmark, Austria, Portugal, and Finland.

* **Data cleaning, engineering for Machine Learning**

Some of the cleaning steps mentioned before such as the dropping of missing values, the separation of the attribute “Week” into “Year” and “Week\_Year”, the incorporation in the following sections, of the attribute month based on the “Year” and “Week\_Year” to have more data that helps the model to have things to learn form. Also, it was decided to exclude the year 2019 due to the big drop shown from this point to 2020 because of Covid-19, since this is an atypical event that could make the model take into consideration a pattern that is not usual.

* **Interactive Dashboard**

Following Tuft’s principles of visualization, such as clarity, and efficiency a dashboard to summarize the project’s insights has been elaborated, prioritizing simplicity, and avoiding unnecessary information displayed. The mentioned dashboard shows the passengers using Luas Services allowing you to select the year of interest with the option to choose the Luas line information needed. Also, it shows the machine learning model chosen as the one with the best performance (Random Forest Regression Model) giving the user the option to select the number of years for which the user is seeking to predict data (See Appendix Figure 3.1 and Appendix Figure 3.2).

### Statistics

Statistics is the science of Learning from Data. (Lyman Ott & Longnecker, 2015). As simple as that, with this science studies or experiments can be designed, and they can be modeled and analyzed to make data-driven decisions.

* **Descriptive statistics**

There are many ways to organize and describe a data set. Important characteristics to look for when organizing and describing a dataset are its center, its variability (or spread), and its shape.(Larson & Farber, 2012). The characteristics mentioned before have been analyzed in the datasets of this project.

* Shape of the data

Appendix Figure 4 and Appendix Figure 7 shows the distribution of the dataset, the values are grouped into bins on the x-axis and the height indicates how many values of the dataset fall into that bin.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Ireland´s Dataset | | Europe’s Dataset | |
| Value | Observation | Value | Observation |
| Modes | bimodal | Asymmetrical two modes left-skewed | unimodal | Right-skewed |
| Coefficient of skewness (S) | -0.42 | Left-skewed | 1.85 | Right-skewed |
| Kurtosis (K) | -1.16 | Platykurtic more flatter than normal spread out | 2.94 | Leptokurtic sharper and peak |

Table 1. Shape of the data parameters.

* Central tendency

Table 2 shows the values of central tendency of both datasets.

|  |  |  |
| --- | --- | --- |
| Parameter | Ireland´s Dataset | Europe’s Dataset |
| Value | Value |
| Median | 739,584.00 | 324,144,500.00 |
| Max Value | 1,173,473.00 | 2,938,023,000.00 |
| Min Value | 49,393.00 | 55,009,000.00 |
| Range | 1,124,080.00 | 2,883,014,000.00 |

Table 2. Central tendency.

* Variablity

Quantiles are values that split sorted data or a probability distribution into equal parts. They are three values that split sorted data into four parts each with an equal number of observations.

|  |  |  |
| --- | --- | --- |
| Parameter | Ireland´s Dataset | Europe’s Dataset |
| Value | Value |
| Min Value | 49,393.00 | 55,009,000.00 |
| q1 | 386,208.00 | 165,635,000.00 |
| q2 | 739,584.00 | 324,144,500.00 |
| q3 | 919,255.00 | 885,281,500.00 |
| Max Value | 1,173,473.00 | 2,938,023,000.00 |
| iqr | 533,047.00 | 719,646,500.00 |
| std | 307,021.91 | 755,615,144.93 |

Table 3. Measures of variability.

Appendix Figure 5 and Appendix Figure 8 show a Box plot, which is a standardized way of displaying the distribution of the data based on the following numbers; minimum, first quartile (q1), median, third quartile (q3), and maximum.

* **Inferential statistics**
* Testing normality

The Q-Q plot is a technique that compares two distributions (A normal and the actual distribution) by plotting their quantiles against each other. Appendix Figure 6 and Appendix Figure 9 clearly show how the datasets do not follow a normal distribution. Shapiro–Wilk test allows you to determine normality in the dataset, the hypotheses were: H0 (Accepted): The data are normally distributed (Pvalue > 0.05) and Ha (Rejected): The data are not normally distributed. According to the results of the test, Ireland’s dataset is supposed to be normally distributed which has been proved wrong, it is important to mention that this test might not detect non-normality in two-tail distributions.

* Hypothesis testing

A hypothesis test is a process that uses sample statistics to test a claim about the value of a population parameter. (Larson & Farber, 2012)

Having a sample of Ireland’s dataset (See Appendix Figure 10) a Z test was applied with a Null hypothesis as if the mean passengers of the sample were the same as the entire population. A z-score of 0.36 and a p\_value of 0.64 leads to the rejection of the null hypothesis. A confidence level was obtained getting values of (606,769.06 – 731,529.34) of weekly passengers.

A one-sample proportion test was performed to investigate if the passengers using the Green Line is the 50% of the total passengers, a p\_value of 0.0015 leads to the rejection of the null hypothesis. Having a sample of 80 records where 34 are successful as Green Line it was found that this line is not the most used.

* **Inferential statistics to find similarities between the datasets**

Parametric statistics tests, used for continuous data, were applied to compare Ireland’s passengers and the second dataset of rail passengers from Spain, France, Italy, Germany, Denmark, Austria, Portugal, and Finland. Stratified sampling was used in this case, as it was pretended to get insights about particular countries against Ireland. The following tests were applied:

Independent T-test two-tailed with a Null Hypothesis as if the mean passengers in Ireland is the same as the mean on Europe’s dataset. Getting a T-statistic of -4.21 and a p-value of 0.000057. The Null Hypothesis was rejected.

Independent T-test left-tailed with a Null Hypothesis as if the mean passengers in Ireland is greater or equal to the mean of passengers on Europe’s dataset. Getting a T-statistic of -4.21 and a p-value of 0.000029. The Null Hypothesis was rejected. The mean of the passengers in Ireland is inferior to data from the other countries.

ANOVA test to confirm the results of the first test with a Null Hypothesis as if the mean passengers in Ireland is the same as the mean on Europe’s dataset. Getting an F-Statistic of 52.70 and P-value 1.40e-17. The Null Hypothesis was rejected, confirming the result of the first test.

Another set of different countries of Europe (Estonia, Croatia, Turkey, Bulgaria) was selected to compare Ireland once again with other European transports. The following tests were applied:

ANOVA test with a Null Hypothesis as if the mean passengers in Ireland is the same as the mean on Europe’s dataset. Getting an F-Statistic of 50.01 and P-value 7.91e-10. The Null Hypothesis was rejected.

Independent T-test right-tailed with a Null Hypothesis as if the mean passengers in Ireland is less or equal to the mean of passengers on Europe’s dataset. Getting a T-statistic of -3.36 and p-value of 1.00. The Null Hypothesis was rejected. The mean of the passengers in Ireland is greater than the data from the other countries.

To run the tests the formula for each type of test was coded and in the case of the ANOVA the Scipy library was used.

### Machine Learning

Machine learning is about extracting knowledge from data. It is a research field at the intersection of statistics, artificial intelligence, and computer science and is also known as predictive analytics or statistical learning. (C. Muller & Guido, 2017)

* **Description of choice of machine learning model**

After inspecting the data it was concluded the data frame was suitable for Supervised Learning models which are used whenever we want to predict a certain outcome from a given input, and we have examples of input/output pairs. (C. Muller & Guido, 2017).

Specifically, Regression models, to predict numerical features such as the number of passengers using public transport. Models such as Linear, Polynomial, Ridge, Decision Tree, and Random Forest regression will be performed to compare which model gives more accurate results.

* **Hyperparameter and Cross-Validation**

Cross-validation is a way to evaluate a given algorithm on a specific dataset. However, it is often used in conjunction with parameter search methods like grid search. For this reason, many people use the term cross-validation colloquially to refer to grid search with cross-validation. (C. Muller & Guido, 2017)

Adequate hyperparameters (number of degrees, alpha value, max\_depth,) and cross-validation (K-fold) were chosen for each model to verify how the results improved with these techniques. The internal process the model follows is the one shown in Appendix Figure 11.

* **Table of Machine Learning models results**

|  |  |  |
| --- | --- | --- |
| Model | Train Score | Test Score |
| Linear Regression Model | 0.63 | 0.44 |
| Lagged Linear Regression | 0.96 | 0.94 |
| Polynomial Regression Model | 0.70 | 0.60 |
| Polynomial Regression Model – GridSearchCV | 0.85 | 0.78 |
| Ridge Regression Model | 0.63 | 0.44 |
| Ridge Regression Model - GridSearchCV | 0.84 | 0.75 |
| Decision Tree Regression Model | 0.77 | 0.68 |
| Decision Tree Regression Model - GridSearchCV | 0.93 | 0.89 |
| Random Forest Regression Model | 0.93 | 0.91 |
| Random Forest Regression Model - GridSearchCV | 0.93 | 0.89 |

Table 4. Results Machine Learning Models perform.

As seen in the table above, the models show a significant improvement when the GridSearch and the Cross-Validation are applied. Linear, Polynomial, and Ridge Regression models that address linear relationships in the data, throw a significantly lower performance, showing a possible non-linearity or complexity in the data. Therefore, the suitable models based on the research are the Decision Tree and Random Forest Regression Models. It is important to mention that a Lagged Linear Regression was performed getting a better result in training and test scores, it seems like the creation of more data gives the model sufficient tools to build a successful model.

* **Sentiment Analysis**

Sentiment analysis, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, appraisals, attitudes, and emotions towards entities and their attributes expressed in written text. (Liu, 2020)

This analysis will be used to investigate people’s opinions on the public services of Ireland and another European country of choice to make a comparison about it. The data was gathered from Reddit connecting through an API, and all the titles and comments under the subreddit “Luas” for Ireland and subreddit “tram” in the case of Germany, were collected successfully. Cleaning techniques such as converting to lowercase, taking away any special character, removal of stop words, and correcting using text blob were used. Tools such as Stemmer and Lemmatizer were used for text normalization, helping to reduce words to their base or root forms, making it easier to analyze and compare text.

A sentiment polarity model was used to categorize into three main classes, positive, negative, and neutral. It was noticed the Lemmatizer is less rigorous in the classification, consequently, the analysis was based on the results of the Stemmer model in both countries. In general, the perception of the people about the public transport mentioned is more positive in Germany with 42% compared to 25% for Ireland, however, for the last country 57% was classified as neutral this could be an opportunity to make a more profound analysis to get more insights about it. Appendix 12.1, 12.2, 13.1, and 13.2 show the results of the sentimental analysis.

## References

C. Muller, A. & Guido, S., 2017. *Introduction to Machine Learning with Python.* First Edition ed. s.l.: O'Reilly.

Larson, R. & Farber, B., 2012. *Elementary Statistics.* 5Th ed. s.l.:Pearson Education, Inc..

Liu, B., 2020. *Sentiment Analysis.* Second Edition ed. s.l.: Cambridge University Press.

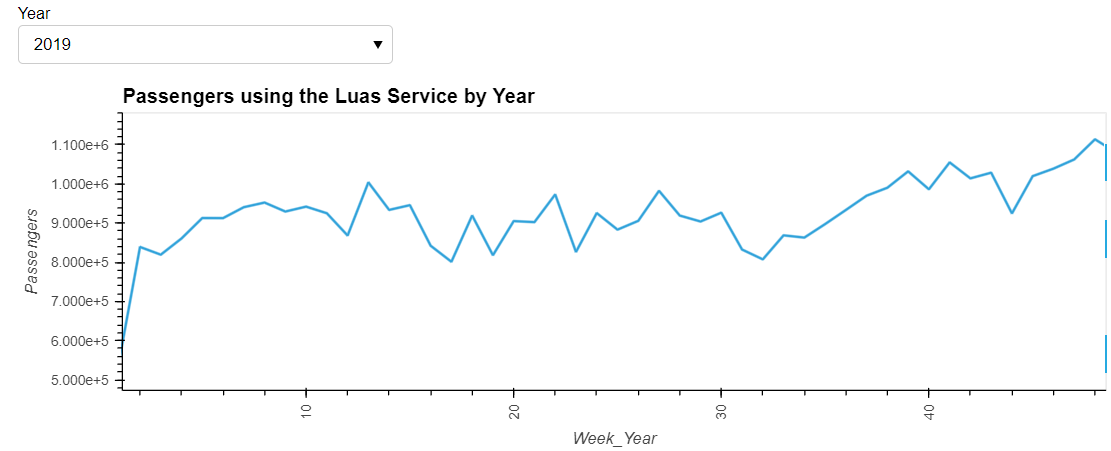
Lyman Ott, R. & Longnecker, M., 2015. *An Introduction to Statistical Methods & Data Analysis.* Seventh Edition ed. s.l.:Cengage Learning.

Peng, R. D. & Matsui, E., 2016. *The Art of Data Science.* s.l.:s.n.

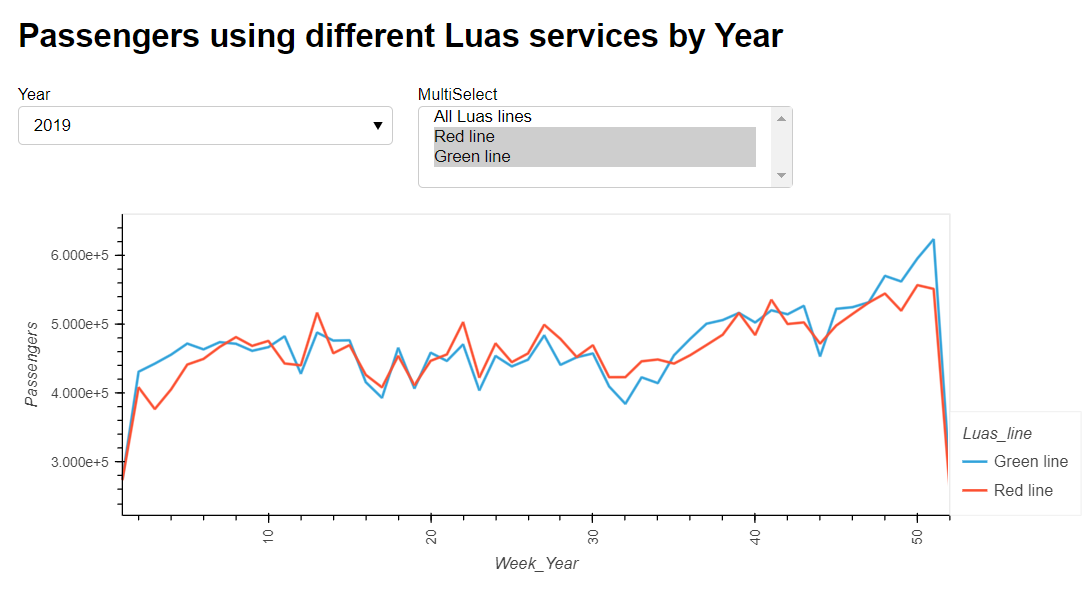
Tufte, E. R., 2007. *The Visual Display of Quantitative Information.* Second Edition ed. United States of America: Graphics Press LLC.

## Appendixes

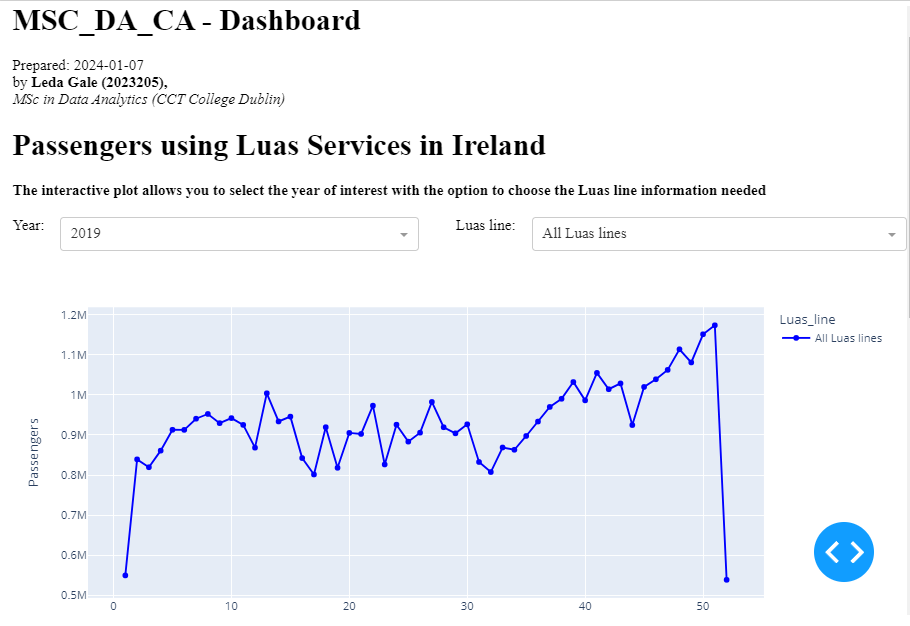
Appendix Figure 1 Passengers using the Luas Service by Year.



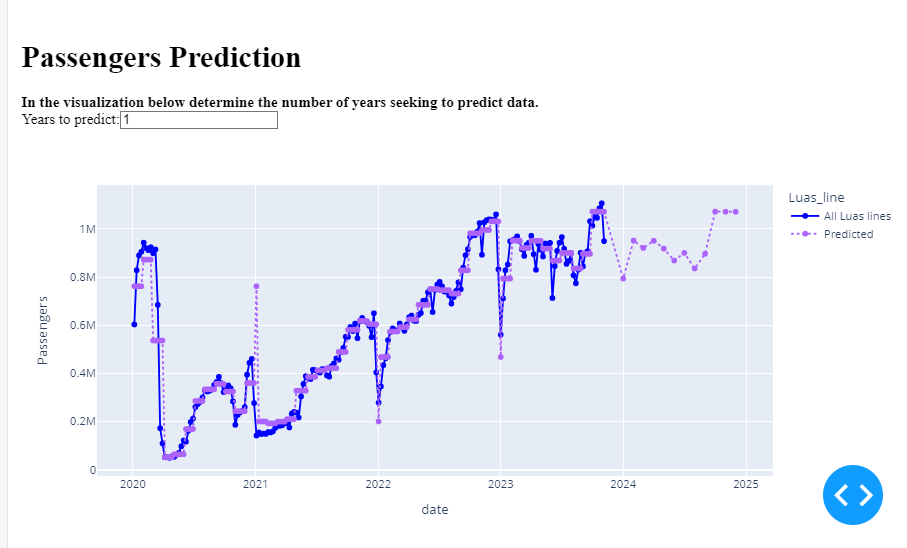
Appendix Figure 2. Passengers using different Luas services by Year.



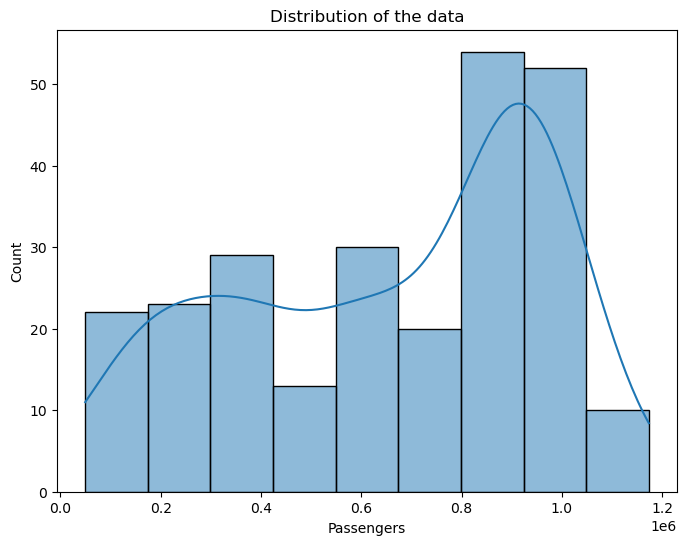
Appendix Figure 3.1. Dashboard Part 1.



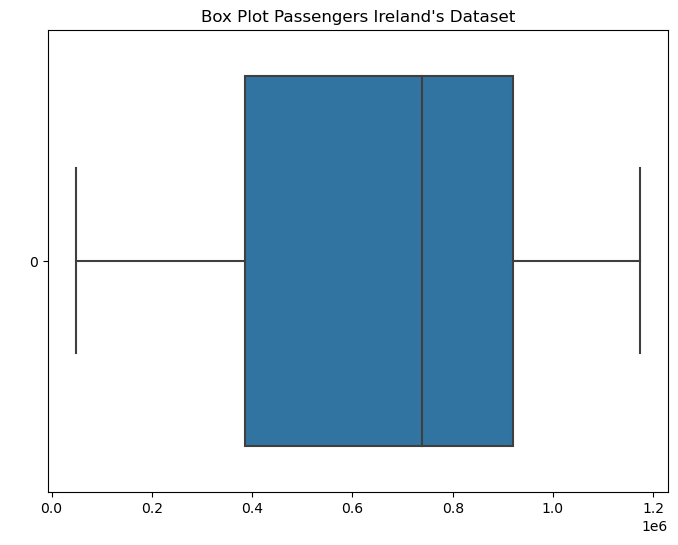
Appendix Figure 3.2. Dashboard Part 2.



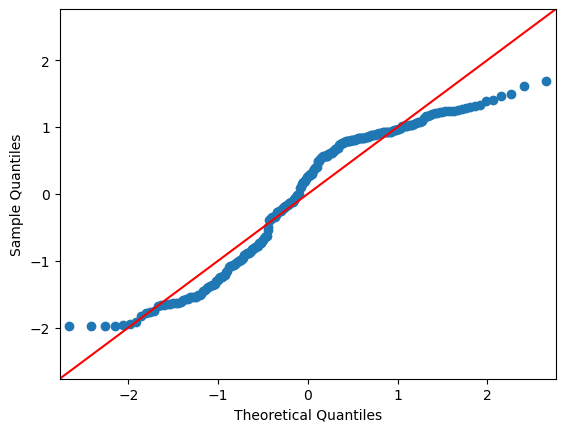
Appendix Figure 4. Distribution of the data (Ireland’s Dataset)



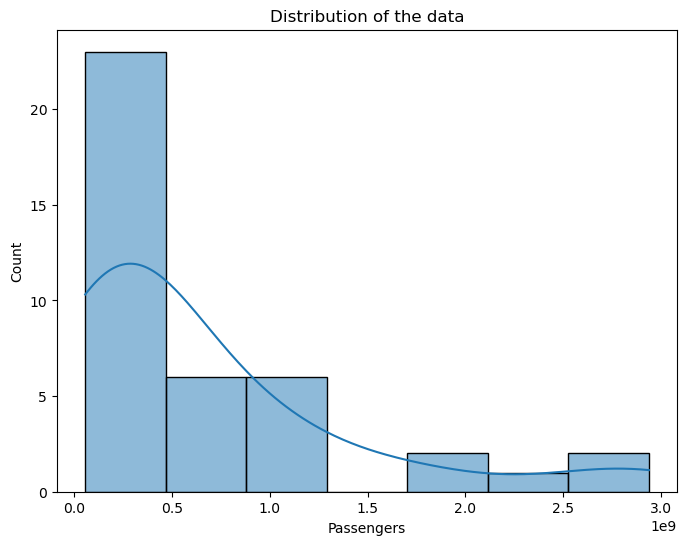
Appendix Figure 5. Box Plot (Ireland’s Dataset)



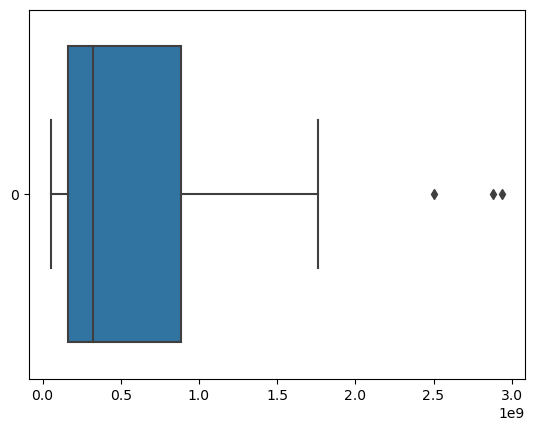
Appendix Figure 6. Q-Q plot (Ireland’s Dataset)



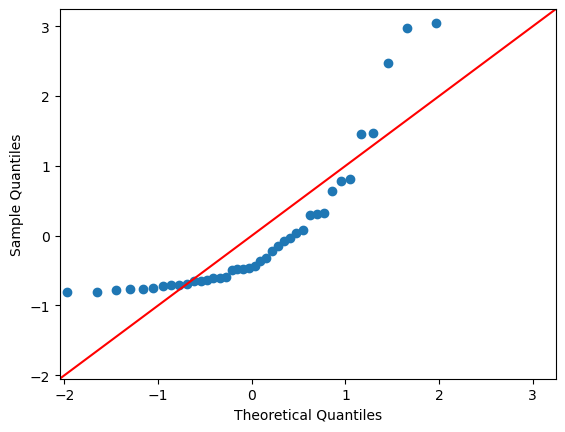
Appendix Figure 7. Distribution of the data (Europe’s Dataset)



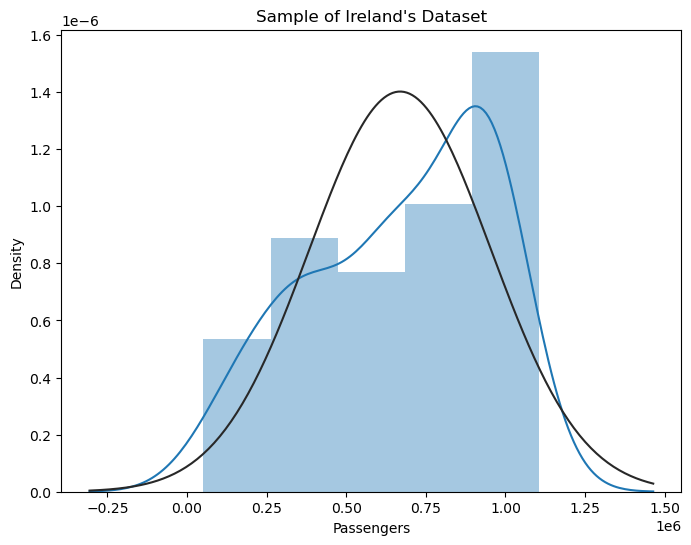
Appendix Figure 8. Box plot (Europe’s Dataset)



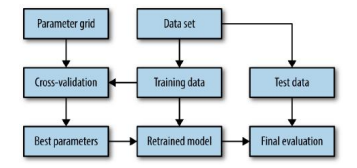
Appendix Figure 9. Q-Q plot (Europe’s Dataset)



Appendix Figure 10. Sample Ireland’s Dataset.

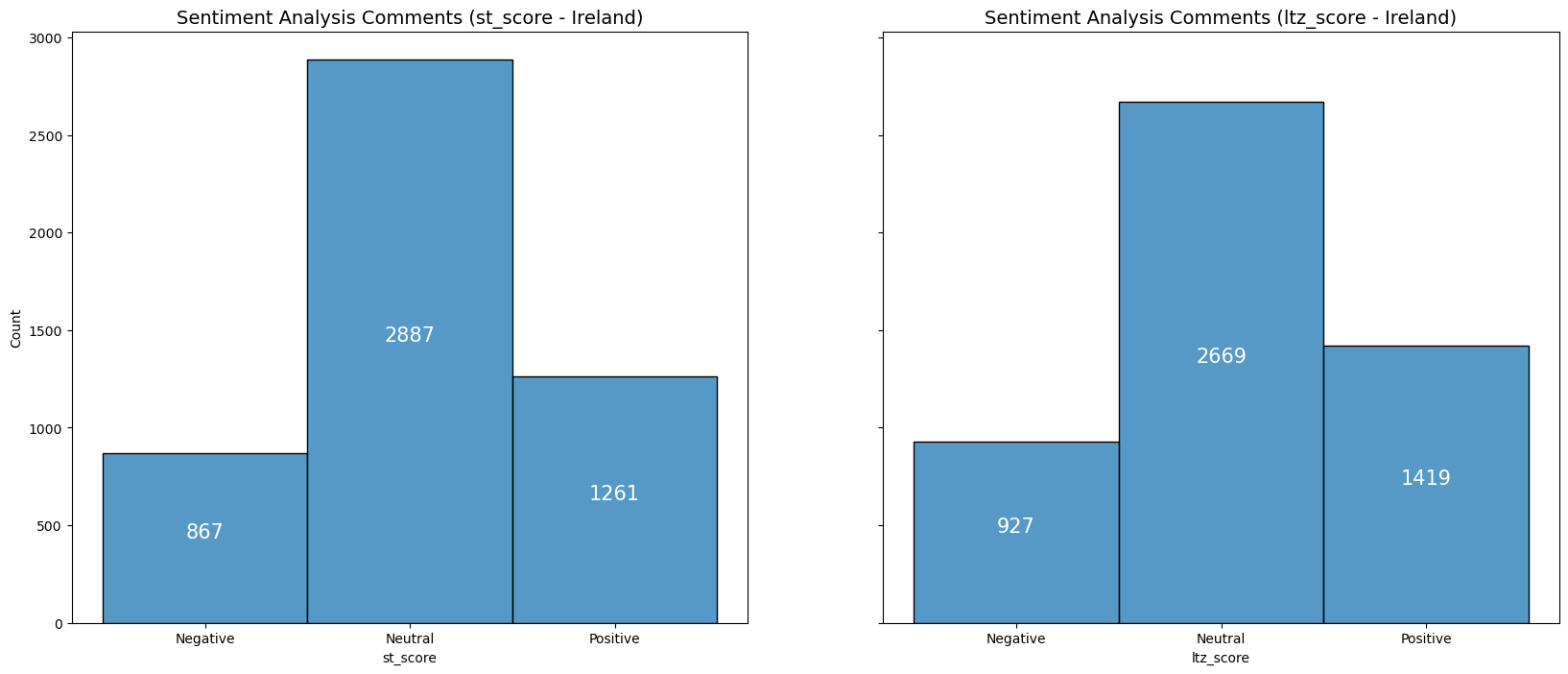


Appendix Figure 11. Process of hyperparameter selection and Cross-Validation.

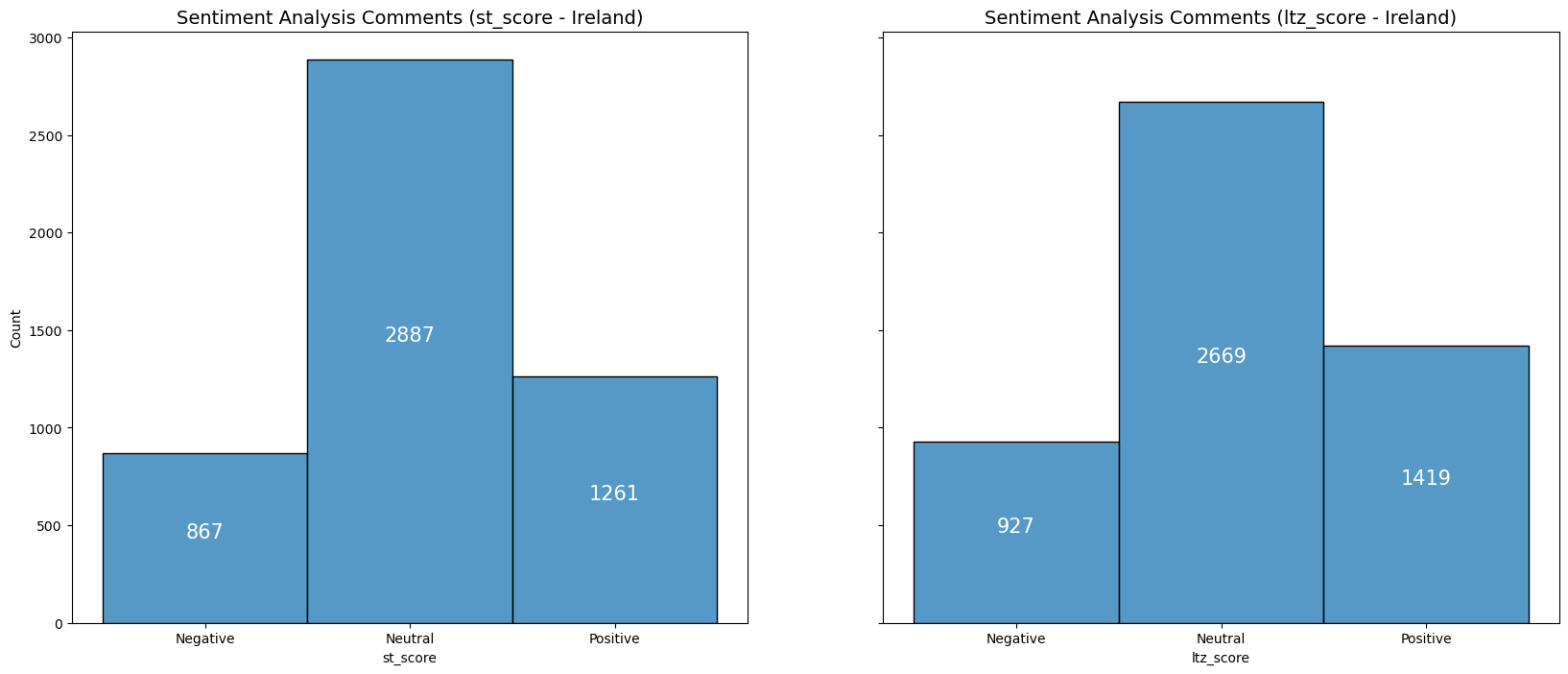


Source: (C. Muller & Guido, 2017)

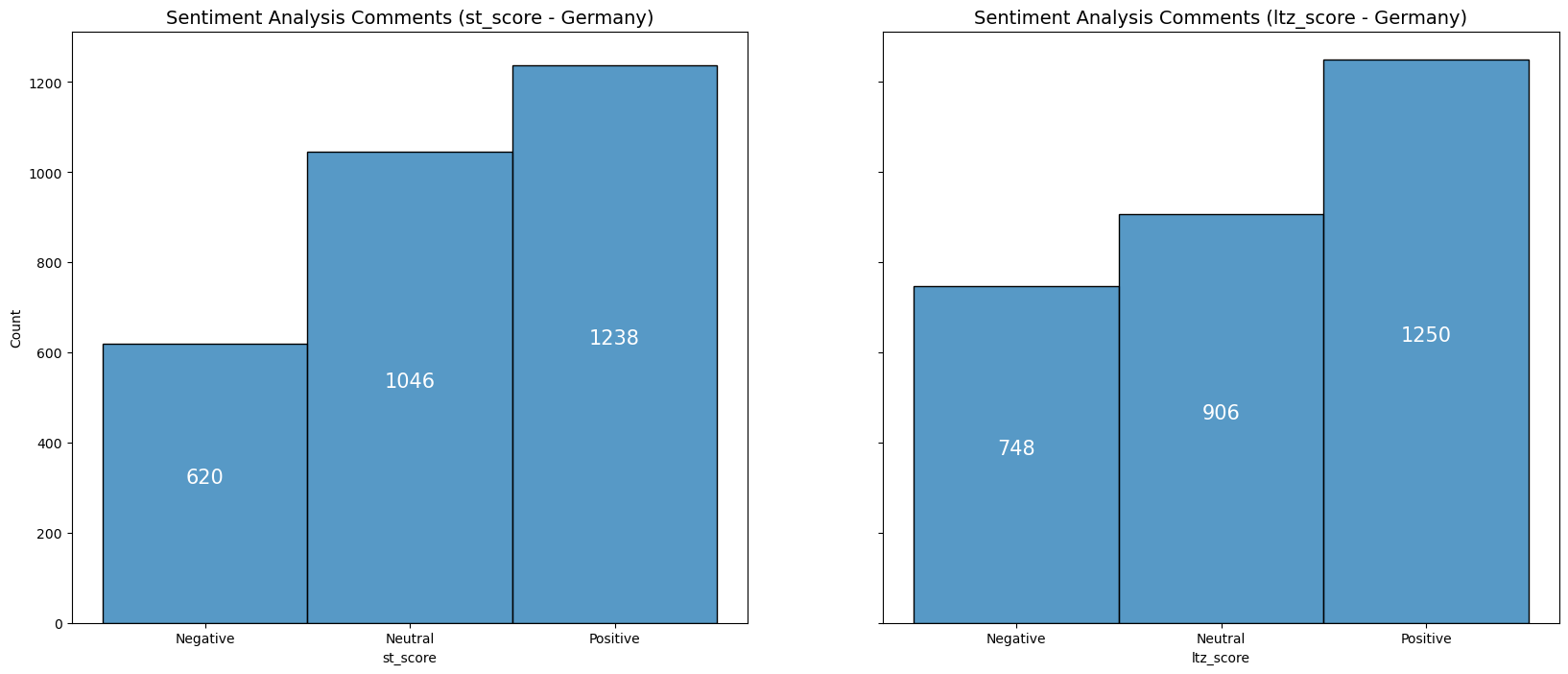
Appendix Figure 12.1 Sentiment Analysis Comments (st\_score – Ireland)



Appendix Figure 12.2 Sentiment Analysis Comments (ltz\_score – Ireland)



Appendix Figure 13.1 Sentiment Analysis Comments (st\_score – Germany)



Appendix Figure 13.2 Sentiment Analysis Comments (ltz\_score – Germany)

