**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Public Transport Analysis |
| **Assessment Title:** | Public Transport Analysis |
| **Lecturer Name:** | *Marina Iantorno/Taufique Ahmed*  *Sam Weiss/David Gonzalez*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Zhe Luo |
| **Student Number:** | 2023234 |
| **Assessment Due Date:** | 07/01/2024 |
| **Date of Submission:** | 07/01/2024 |
| **Github** | https://github.com/ZheLuo2023/CA2.git |
| **Email:** | 2023234@student .cct.ie |

**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. |

**Public Transport Analysis**

**Abstraction**

The role of public transportation in a nation's economic development is critical. In this research, public transportation datasets from Ireland and other countries were compared. Using cutting edge data analysis techniques, including machine learning, supervised and unsupervised learning, time series analysis, sentiment analysis, and visualization techniques, the research sought to answer two main questions:

(1) Are there any countries that are similar to Ireland in particular ways?

(2) How satisfied are the Irish people with their public transportation system?

Following these questions as a focal point, the analysis employed a variety of analytical methods and techniques to delve deeply into public transportation data from Ireland and other countries. The overall goal of this research is to offer valuable insights regarding the critical perspectives of the Irish population regarding their transportation needs, as well as insightful perspectives on Ireland's railway transportation sector in comparison to other countries.

**Key words: Public Transport Data Analysis Machine Learning Ireland Satisfaction**

**Introduction**

The final selection of the dataset for this study was made after a thorough evaluation of up to thirty different types of datasets; these include population and per capita GDP datasets, railway traffic data for Ireland and other countries, and a dataset of user comments obtained via the Reddit website API, among other datasets.

The study is divided into four main modules, each of which examines how different techniques from different disciplines are applied in the context of this study. For example, in the statistical analysis section, different parametric and non-parametric methods will be used to compare data between different countries; in the machine learning section, different machine learning models will demonstrate their performance in regression tasks; and the study will delve into the application of machine learning models to machine learning tasks.

The programming section will provide a thorough overview of Python's powerful applications in the field of data science, including handling various types of data, enhancing readability, improving code quality, and using programming techniques to improve model performance. The data visualization section will feature a wide range of visualization techniques, demonstrating how they aid in gaining key insights from the dataset and presenting research findings.

**Data Sourse**

|  |
| --- |
| Data Set Name: Rail freight transport.csv  Description: Ireland, Finland, Luxembourg, Japan, Rail freight transport data from 1981 to 2022, in millions.  Data source: [https://stats.oecd.org/#](https://stats.oecd.org/)  Key word Search : Transport  Copyright Notice: <https://www.oecd.org/termsandconditions/>  Please Read : I. Use of Material -> (c) Data |
| Data Set Name: Rail passenger transport.csv  Description: Ireland, Finland, Luxembourg, Japan, Rail passenger transport data from 1981 to 2022, in millions.  Data source: [https://stats.oecd.org/#](https://stats.oecd.org/)  Key word Search : Transport  Data source: <https://www.oecd.org/termsandconditions/>  Please Read : I. Use of Material -> (c) Data |
| Data Set Name: Rail infrastructure investment.csv  Description: Ireland, Finland, Luxembourg, Japan, Rail infrastructure investment data from 1995 to 2022, in Euros  Data source: [https://stats.oecd.org/#](https://stats.oecd.org/)  Key word Search : Transport  Data source: <https://www.oecd.org/termsandconditions/>  Please Read : I. Use of Material -> (c) Data |
| Data Set Name: GDP per head of population\_USD, constant prices, 2015 PPPs.csv  Description: Ireland, Finland, Luxembourg, Japan, GDP per head of population\_ data from 1995 to 2022, in USD, constant prices, 2015 PPPs  Data source: [https://stats.oecd.org/#](https://stats.oecd.org/)  Key word Search : GDP  Data source: <https://www.oecd.org/termsandconditions/>  Please Read : I. Use of Material -> (c) Data |
| Data Set Name : World\_population.CSV  Description: World population data for each country in the world, from 1950 to 2100, with estimates starting from 2022  Data source: <https://population.un.org/wpp/Download/Standard/MostUsed/>  Copyright Notice:  Copyright © 2022 by United Nations, made available under a Creative Commons license CC BY 3.0 IGO: http://creativecommons.org/licenses/by/3.0/igo/ Suggested citation: United Nations, Department of Economic and Social Affairs, Population Division (2022). World Population Prospects 2022, Online Edition. |
| Data Set Name: Reddit customer comments  Description: Publiced on the Reddit website, the dataset of comments related to the public transportation experiences in Ireland and Argentina.  Data source: <https://www.reddit.com/r/irishtourism/comments/17y3yl5/experiences_using_public_transport_in_ireland/>  <https://v.redd.it/n7uzzgdfecdb1>  Copyright Notice: <https://www.redditinc.com/policies/user-agreement/> |

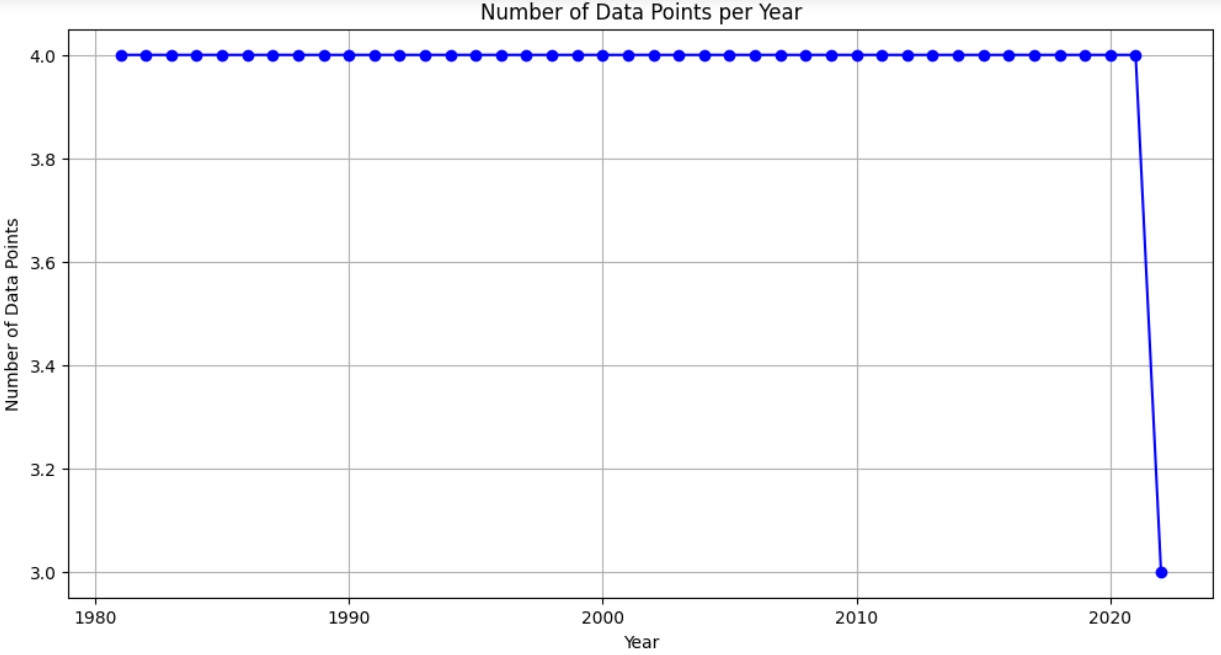
***Data Preparation & Visualisation***

*Chapter 1 . Code* *1- 96 line.*

A dataset containing data on public transportation from various countries, including Ireland, was eventually found on The Organisation for Economic Co-operation and Development (OECD) website; however, it was noted that many of these resources had significant issues such as extensive missing values, short data periods, limited data volume, and the ability to download only single-variable datasets at a time. It became quite difficult to find a dataset that simultaneously encompassed multiple countries, including Ireland, and had data on multiple transportation variables. This led to the idea of collecting multiple datasets and merging them.

Following thorough comparisons, the research focused on four variables for railway transportation in Ireland, Luxembourg, Finland, and Japan. The national population variable was taken into consideration because of its substantial influence on transportation; data for this variable was obtained from the United Nations website, albeit with the disadvantage that it was difficult to process and had a complex structure.

*Chapter* 2 .Code *1- 96 lines*



After conducting Exploratory Data Analysis (EDA) on each original dataset, including methods like **head()**, **info()**, **describe()**, and creating Missing Values Heatmaps and so on , it was noticed that data dimensions didn't match expectations.

A more detailed investigation of the OECD data showed a pattern: rows that had missing values for the primary variables were completely eliminated; requiring a unique methord to locate these gaps. For instance, in the Rail freight transport dataset, the results of this search are depicted in the figure.

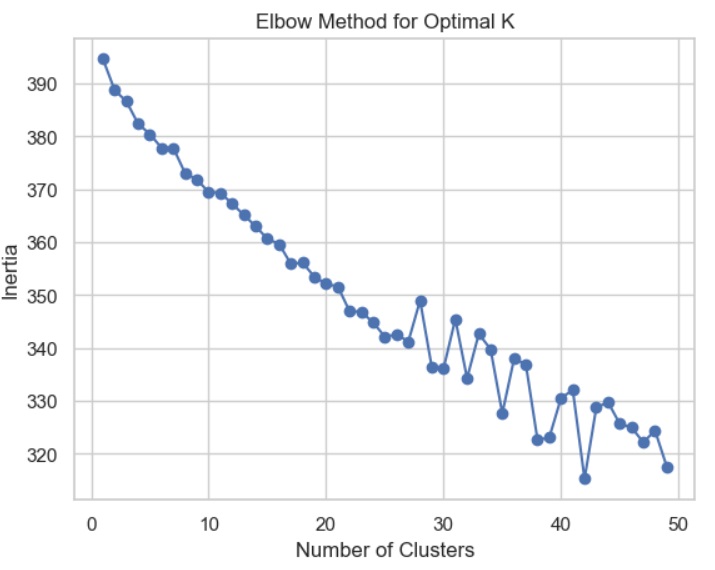
For population data, a custom search class was constructed following **EDA(**Exploratory Data Analysis**)** to locate relevant variables, and the necessary results were obtained after data segmentation. Finally, a search for missing years was undertaken, but no missing values were detected.

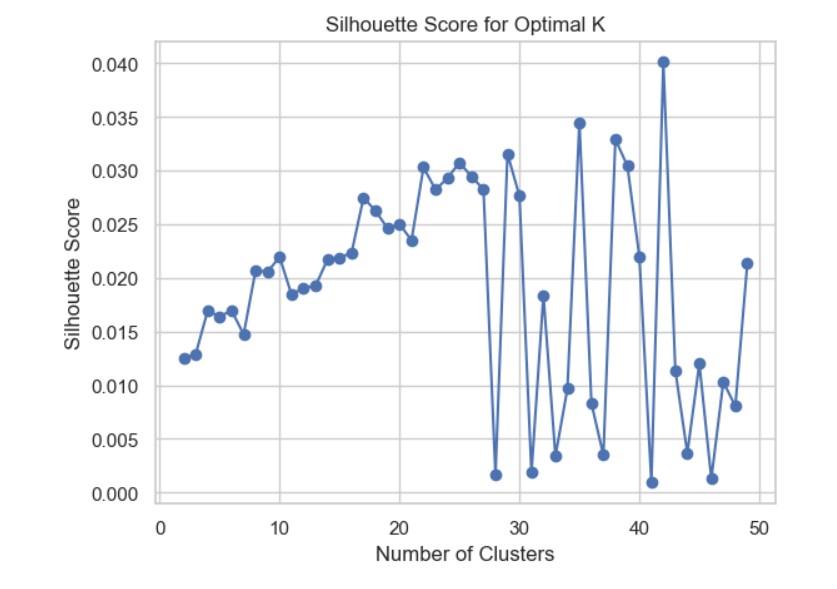
*Chapter* 3 Code 98-126，179-185

Because of the handling of missing values during previous data cleaning steps, the focus of EDA for the integrated dataset has shifted to obtaining distribution information for each country and monitoring outliers, with the goal of providing guidance for subsequent statistical and machine learning analyses. Initially, extensive descriptive statistical methods were used to compute various statistics for each variable in different countries.

The generated country distribution and outlier information serve as a critical foundation for later statistical parameter analysis, such as the selection of parameter analysis targets and the removal of outliers discovered using scatter plots to make the data conform to a normal distribution.

Because most countries and variables do not follow a normal distribution, MinMaxScaler was used to scale the data for the machine learning section. Following the supervised learning studies, a bar chart was used to clearly compare the performance of different models. Bar charts were chosen because they effectively display subtle differences between numerical values.





During the sentiment analysis phase, using the K-means model for unsupervised learning, the optimal hyperparameter, the number of clusters, was determined using the **Elbow** Method and **Silhouette Score** Method, visualizing the best number of clusters as 30. Following sentiment classification, a pie chart was used to represent the sentiment classification results for Ireland and Argentina.

*Chapter* 4 code 101

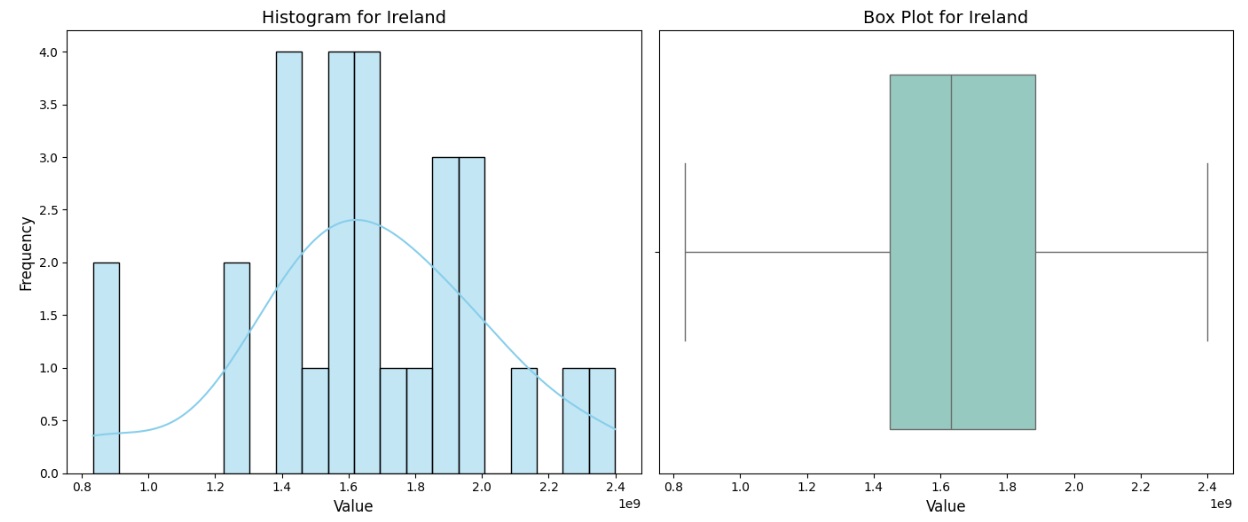
Based on the merged overall dataset, this dashboard perfectly adheres to Tuft's four crucial principles (Tufte, 2001): clear visualization, easy comprehensibility, time dimension scales, and overall strategy. The importance of the lie factor and data ink ratio was noted in the design of the charts, and as a result, they were meticulously adjusted to ensure the optimal ratio for presenting data information as clearly as possible while maximizing the data space.

The brilliance of this dashboard lies in its ability to allow users to compare multiple countries. Using callback functions, the titles and axis labels dynamically adjust to transform based on the selected countries, making it exceptionally easy to compare the same parameters across countries.

By comparing charts of various parameters from different countries, it was discovered that Ireland and Luxembourg have remarkably similar data distributions in the "Rail infrastructure investment (Euro)" variable, directing future research towards uncovering commonalities between these two countries in the field of railway transportation.

**Statistic for Data Analytics**

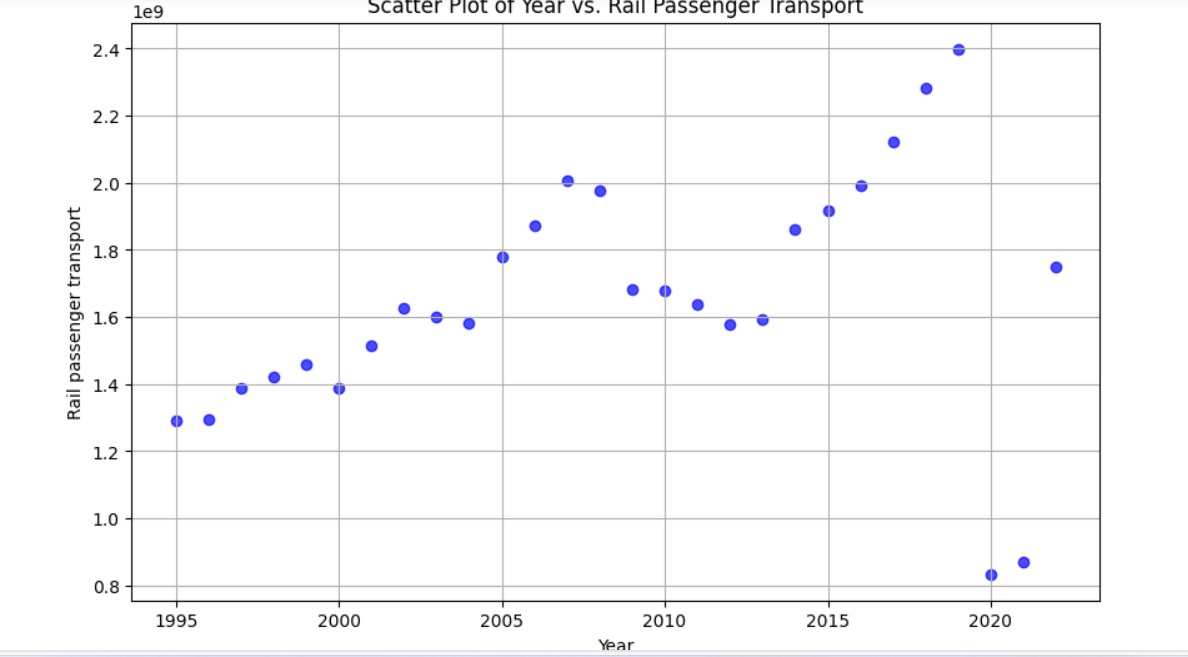
*Chapter* 1 code:98-117



A total of five different datasets were used for this research project, and a detailed descriptive statistical approach was not directly used to integrate them at the initial stage. Instead, a preliminary exploration of the five datasets was conducted, examining each dataset's size and identifying missing values. Because all of these datasets include a temporal dimension, with data points arranged in chronological order, utilizing time series methods was deemed the most appropriate.

Following the completion of the entire dataset, a comprehensive exploration will begin, encompassing detailed descriptive statistics and methodologies, such as the computation of measures of central tendency, such as **Mean, Median, and Mode**, as well as measures of dispersion, such as **Variance** and **Standard Deviation** ,For particular calculation results, please see the code section.Once the descriptive statistical analysis is completed, the transition to inferential statistics (David S. Moore, 2014) takes precedence, with a particular emphasis on the normal distribution.

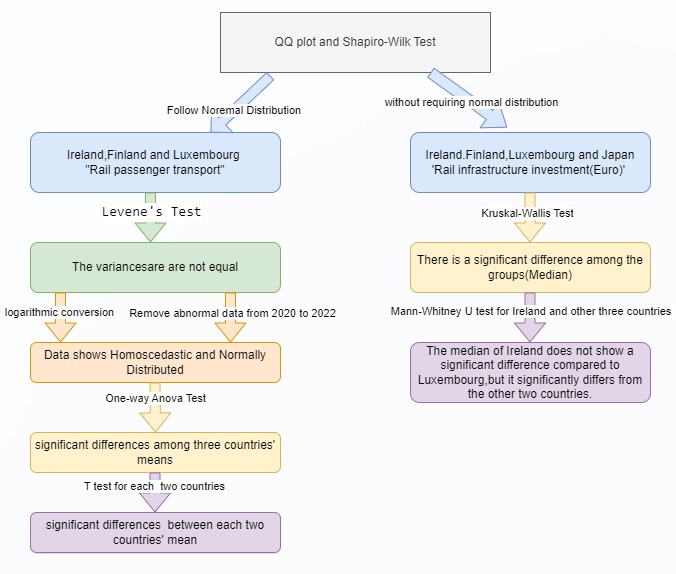
*Chapter* 2 code :118-120



For the next step, inferential statistics (David S. Moore, 2014)were applied to three major variables in Ireland: total rail freight transport, rail passenger transport, and rail infrastructure investment (Euro). Scatterplots were also used for visualization. This non-parametric statistical method was chosen because it is independent of certain distribution assumptions and is appropriate for ordinal data. The results showed that rail passenger transport showed a weak positive correlation (0.4636) over time, indicating an increasing acceptance of train travel, while rail freight transport showed a strong negative correlation (-0.9271), likely as a result of the development of alternative, more effective freight transport options. Rail infrastructure investment (Euro) did not correlate (0.1738), most likely owing to policy difference.

These statistical tools allow for a better understanding of the advancements in Ireland's railway industry by illuminating the underlying population values among variables. This knowledge can be essential for policymakers in terms of investment, transportation, and economic growth.

*Chapter* 3Code 120-139



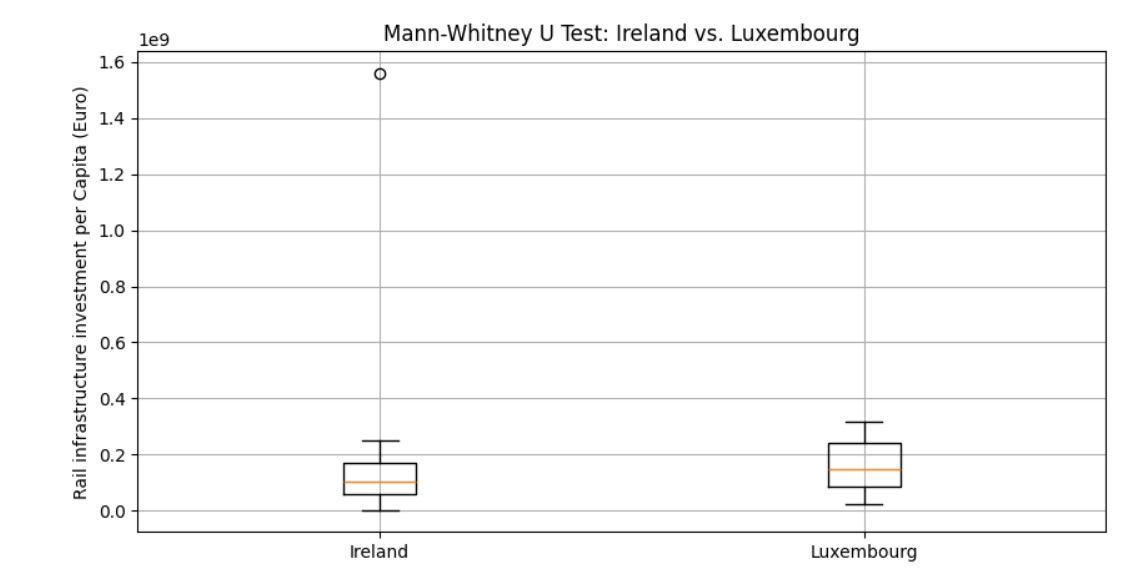
This part does a formal parameter analysis of the railway industry, comparing Ireland to Luxembourg, Finland, and Japan, using the provided flowchart and categorizing hypothesis testing as parametric or non-parametric based on data distribution.

First, the Shapiro-Wilk Test validates that the "Rail passenger transport" columns for Ireland, Finland, and Luxembourg are chosen. Following the verification of data homoscedasticity and distribution, a one-way ANOVA Test indicates substantial mean differences across these three countries, which are then confirmed by paired T-tests.

Following that, the Kruskal-Wallis hypothesis test is applied to the "Rail infrastructure investment (Euro) per Capita" variables for Ireland, Finland, Luxembourg, and Japan, and Ireland is compared separately with each of the other three countries using Mann-Whitney U tests. The results show that the median for Ireland differs significantly from Luxembourg but not from the other two countries.

Finally, while Ireland's railway sector development is very similar to Luxembourg's, it differs greatly from the paths taken by Finland and Japan, due to a variety of factors such as population, land area, and economic conditions in each of these four nations.

*Chapter* 4Code



At first, it is important to note that the dataset contains other significant variables besides "Year" and "Country," but these variables actually come from different data sources. The dataset was integrated after thorough data selection and preprocessing, and because all of the missing data represents time-series data, we used a time series **ARIMA** model for prediction for the small amount of missing data.

Second, following data preprocessing and descriptive statistical analysis, it was discovered that most data did not have a normal distribution, which could impede future research. As a result, QQ-plots and Shapiro-Wilk tests were performed in multiple countries to identify normally distributed variables suitable for parameter analysis.

The data from the three selected countries—Ireland, Finland, and Luxembourg—did not meet the assumption of homoscedasticity, so a logarithmic transformation was performed to achieve homoscedasticity, but this transformation resulted in some country-specific data deviating from normality. Interestingly, scatterplots revealed significant deviations in data trends beginning in 2019, which coincided with the COVID-19 pandemic outbreak. To counteract these outside influences, data from 2020 to 2022 were omitted, resulting in data from these three countries reverting to normal distribution properties.

**Machine Learning for Data Analytics**

|  |  |  |  |
| --- | --- | --- | --- |
| machine learning models used for regression | | | |
| Model Name | Pros | Cons | Hyperparameters |
| ****Linear Regression**** | Simple, efficient for large datasets. | Limited to linear relationships, sensitive to outliers. |  |
| ****Ridge Regression:**** | Handles multicollinearity, prevents overfitting. | Requires tuning, feature standardization. | Alpha (Regularization strength) |
| ****Lasso Regression**** | Selects important features, prevents overfitting. | May pick few features, tuning needed. | Alpha (Regularization strength) |
| ****Decision Tree Regression**** | Captures non-linearity, easy to visualize | Prone to overfitting, sensitive to noise | Maximum depth, Minimum samples per leaf, |
| ****Random Forest Regression**** | Combats overfitting, scalable. | Complex, requires hyperparameter tuning | Number of trees, Maximum depth,  ,min\_samples\_split, min\_samples\_leaf |
| ****SVM Regression**** | Works in complex problems. | Sensitive to settings, computationally intensive. | Kernel type, C (Regularization parameter), Epsilon (epsilon-SVR model) |
| ****KNN Regression**** | suitable for small datasets. | Slow for large datasets, sensitive to outliers | Number of neighbors, Distance metric,  Weighting method |

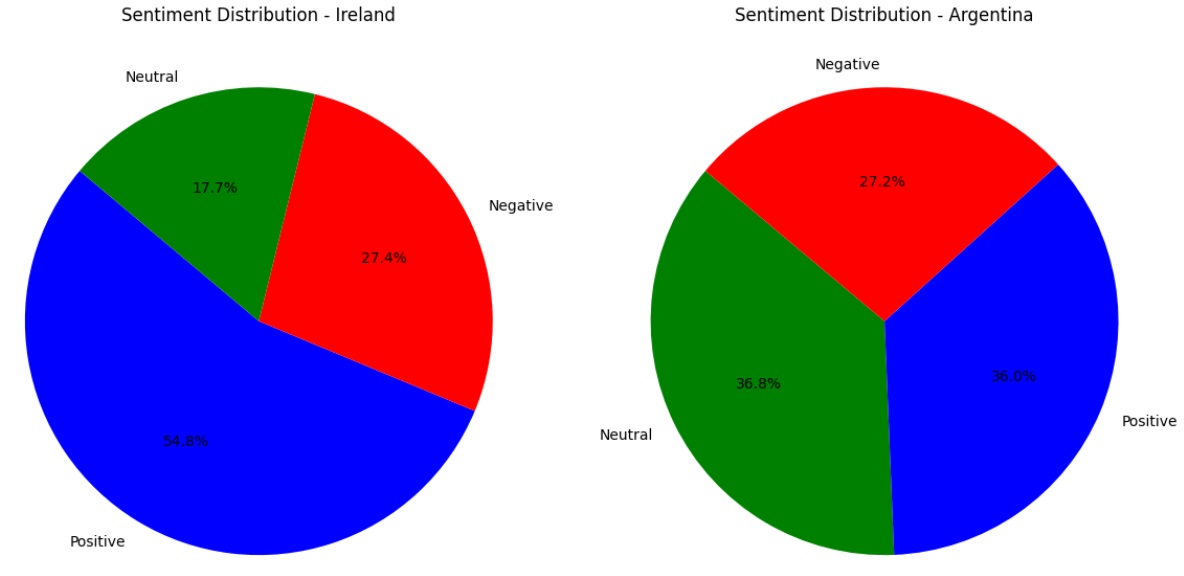
*Chapter* 1 Code 11-15 ,142-145

Prior to formally beginning our journey into machine learning analysis, a few datasets had missing data discovered during the preprocessing and data cleaning stages. Because all datasets are based on sequential order, **ARIMA** models are frequently used for time series data. To address this, the pmdarima package was used to predict missing values using **ARIMA** models.

Following the selection of Luxembourg as the comparison country and Ireland as the base country, regression analysis was performed on the target feature "Rail infrastructure investment (Euro)." The data was divided into training and testing sets, and because the data in both countries did not follow a normal distribution, Standardization was accomplished using **MinMaxScaler**.

All of the machine learning models indicated in the previous figure were considered, and the **GridSearchCV** approach was used to choose the best hyperparameters for regression analysis, resulting in predictions.

*Chapter* 2Code :162-189

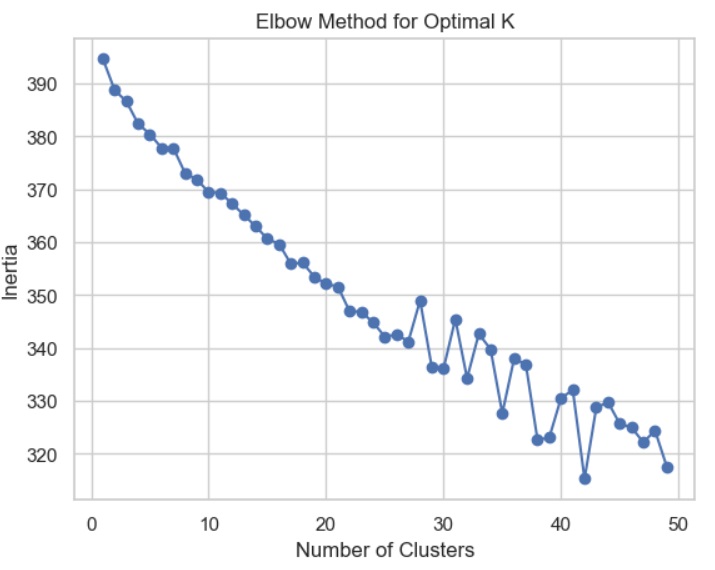
****

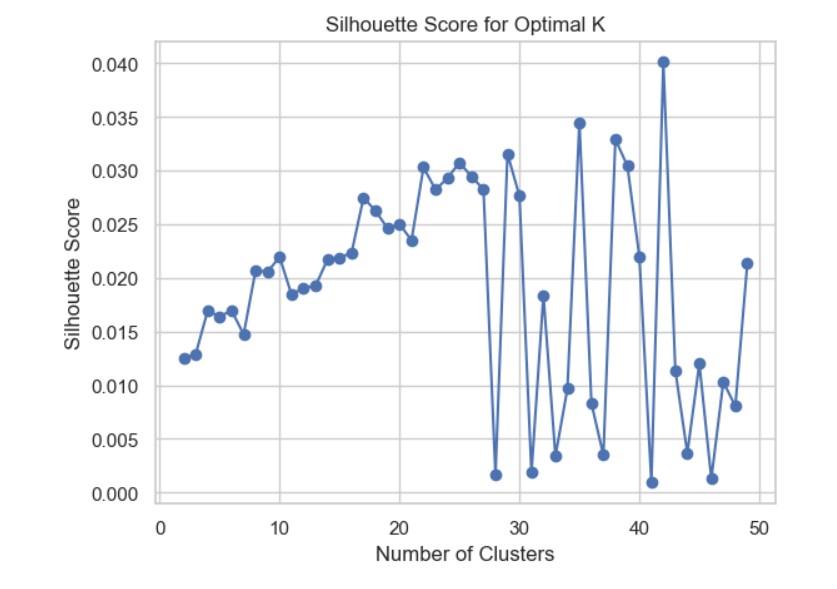
Following the use of the Reddit API to collect user comments from two threads that explored experiences with public transportation in Argentina and Ireland, data preparation was carried out. This included stopword removal, NLTK word tokenization, and **PorterStemmer()** word stemming. Sentiment analysis was performed using both the textblob and vaderSentiment packages; the latter was chosen due to its larger usage, and a threshold of ±0.05 was manually determined. According to the sentiment analysis results, both countries had a higher percentage of positive sentiments than negative sentiments, with Ireland having somewhat higher satisfaction levels (54.8%).

*Chapter* 3 Code 142-160,162-189

The supervised learning procedure included the selection of a model and hyperparameters, the partitioning of all models into training and testing sets, and the deployment of optimal learning algorithms.

Interestingly, the Random Forest model was improved by using the **Feature\_importance()** technique to minimize the number of parameters to four and three; these configurations were also compared.



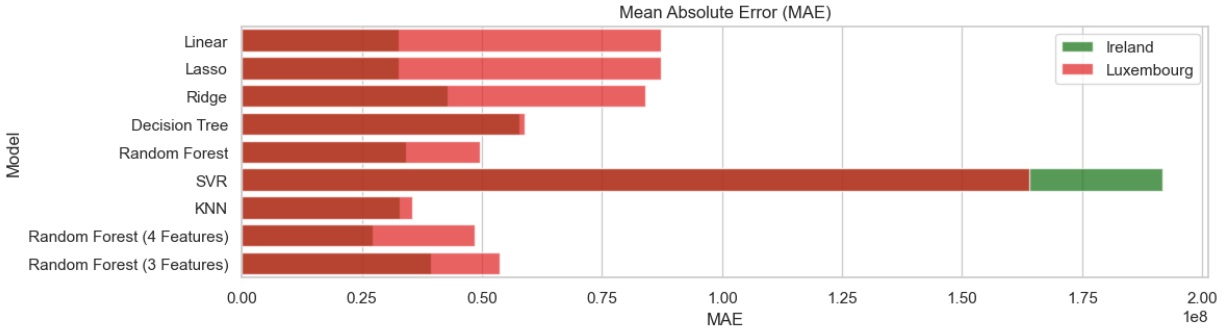


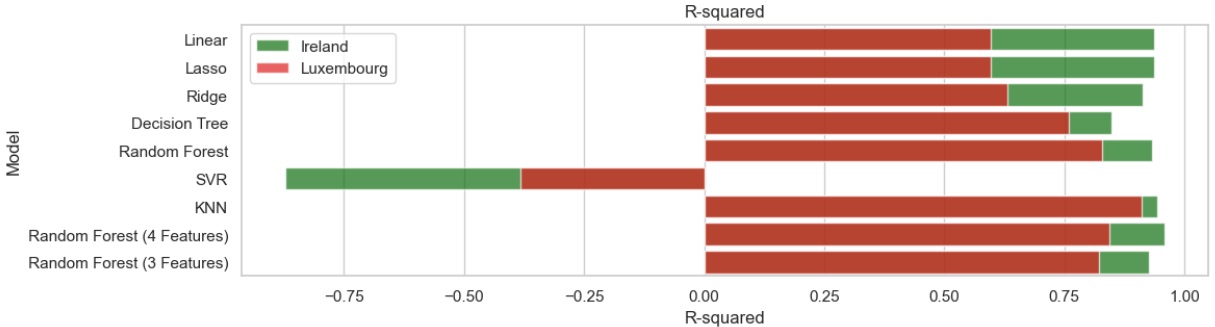
In unsupervised learning, high-volume comments from Argentina's public transportation system were converted into numerical features using the **Bag of Words** and **TF-IDF** methods, allowing grouping with the K-Means algorithm. The **Elbow Method** and **Silhouette\_score Method** were used to identify the optimal number of clusters; results revealed large oscillations over 30 clusters, indicating overclustering. As a result, three clusters were chosen for the K-Means analysis, and the clustering result was not optimal, most likely due to noise or varying perspectives in the comments. Future study paths could include employing larger, higher-quality datasets or reexamining feature selection, both of which could provide significant insights for future investigations.

In addition, the **Latent Dirichlet Allocation (LDA)** model was used to highlight the top keywords for each topic and arrange comments into a variable number of themes.

*Chapter* 4 code 161







As a result, SVM was shown to be less adaptable in this study's data characteristics, however k-Nearests Neighbors (KNN) and Random Forest models performed remarkably well. KNN regularly outperformed other models on the Luxembourg dataset, and it outperformed Random Forest without feature reduction on the Ireland dataset.However, when feature importance-based feature selection was included, Random Forest outperformed KNN in Ireland's dataset and became the best-performing model. Reducing the features to three dimensions affected model performance dramatically, making it worse than not performing feature selection—possibly due to the loss of essential features and information.

Finally, KNN demonstrated its ease of use and suitability for smaller datasets, whilst Random Forest and Decision Tree models—with their ability to select parameters—proved to be effective tools that, when utilized correctly in actual scenarios, can considerably increase model performance.

**Programing**

*Chapter* 1 Code: 1-189

The database for this study was chosen with care and deliberateness to provide optimal analysis and presentation results. Python's built-in techniques such as **head(), info(), shape()**, and the **Pandas** package were heavily used in data preprocessing and cleaning. Pandas provided extensive features for data structure, reading, cleaning, and transformation, as well as strong indexing and labeling capabilities. The **Arima** technique from the **Pmdarima** library was used for its user-friendly automatic parameter search in missing value prediction.

The Scipy and Statistics libraries were mostly used for statistical analysis, while several algorithms from the **Sklearn** library were used for machine learning, and data visualization was primarily achieved using **Seaborn** and **Matplotlib**, both sophisticated data visualization tools.

*Chapter* 2 Code:1- 98 162-165



The key data sources for this study's statistical analysis and supervised machine learning were CSV and Xlsx files, as shown in the data source table. To suit the research criteria, these files were cleaned and processed. JSON format comment data was gathered mostly from the Reddit Web API for the unsupervised learning and sentiment analysis parts of machine learning. This data was additionally cleaned and preprocessed, resulting in key information being transformed into DataFrames for research purposes. In addition, a storage path in **MongoDB** was created to meet the storage requirements of huge datasets and to ease future research activities.

*Chapter* 3 code: 13,15,69,101,113,121,

In this study, all of the data was modularized, with each module having its own set of duties, and we used Object-Oriented programming approaches extensively, as described in the **Programming Paradigm** (Peter Van Roy, 2004) section. Custom classes were constructed for a variety of uses, including leveraging Arima from the pmdarima package for missing value prediction during initial data preprocessing. This approach was also visible in the handling of population data, such as the search for key parameter indices and the exploration of various variables for different countries within the integrated dataset, with custom classes built throughout to facilitate ease of use and reduce code redundancy, thereby minimizing code volume.

Loops and functions were used frequently, particularly in the sentiment analysis and visualization sections, which considerably improved the overall code's intuitiveness, reusability, and maintainability, making the entire development process more efficient.

*Chapter* 4 code :68,101,150-153

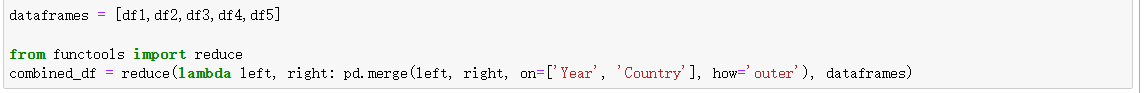
**Unit testing** was performed numerous times on each module in this research project to confirm that the code works as expected. This resulted in the identification of various defects and possible difficulties, one notable example being the Dashboard code blocks, which did not always perform as intended due to server occupancy, and a specific port number was assigned to assure its functionality. Furthermore, during the search for key indices in population data, a **try...except** structure was built to collect and handle exceptions, assuring the program's ongoing execution.

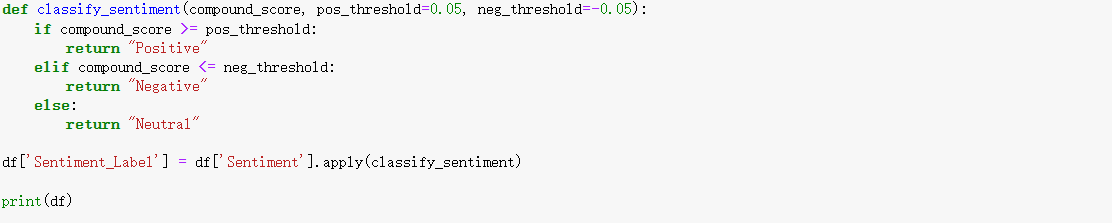
Given the project's multitasking nature, **Various integration tests** were also run in the later stages to ensure correct interaction between different code units.

Furthermore, in this study, an effort was made to reduce the computational resources and runtime utilized by the code. One notable example is optimizing the execution of machine learning supervised learning models such as Random Forest Regression and SVM Regression, which need intensive grid searching of hyperparameters, resulting in significant time and computing resource consumption. To address this, the hyperparameter search process was abstracted and reused on both datasets using functions, resulting in increased efficiency for both models.

Overall, these testing and optimization procedures were critical in assuring the code's dependability, performance, and efficiency throughout the study project.

*Chapter* 5 code:76, 97,111,112,168-174





When processing individual datasets in the early stages of the project, population data in the form of Excel (xlsx) files was saved in two different sheets. Because the structures of these two portions were identical after data preprocessing, the **concat** technique from the Pandas library was used to merge them into a single dataset. However, given the availability of five sub-datasets, the **reduce, lambda**, and **merge** methods from the Pandas package were used to merge all of the different datasets into one complete dataset. This method enabled for the seamless merging of the datasets in a single line of code by utilizing the merge function's **on=[]** option to indicate the joining keys.

As the project advanced into the middle phase, descriptive statistical analysis became necessary, and variance and standard deviation calculations were performed using methods from the **Statistics** library rather than **Pandas'** Var and Std methods. Because the latter methods utilize different formulas, one based on samples and the other on the population, depending on population-based calculations may add bias.

In the project's later stages, Both the **TextBlob** and **VaderSentiment** libraries were used as sentiment analyzers for conducting sentiment analysis on JSON data. While TextBlob is simple and easy to use, vaderSentiment has the advantage of adjustable thresholds, therefore that method was finally picked for further research and analysis.

**Conclusion**

This research project has taken a significant journey, and the research findings are indeed exciting: in a cross-sectional comparison across five dimensions, Ireland and Finland, Luxembourg, and Japan exhibited roughly similar levels of overall railway investment over the past nearly 30 years. However, based on their respective circumstances, these four countries demonstrated significant differences in other areas. When machine learning was applied to regression analysis, it was discovered that the K-Nearest Neighbors (KNN) model performed remarkably well, owing to its adaptability to small datasets and user-friendliness. The Random Forest model, with its distinct feature selection capabilities, also produced excellent results.

In a study of public transportation satisfaction in Ireland and Argentina, it was discovered that the Irish population was 54.8% satisfied with their country's transportation system, indicating a high level of positivity and optimism about their country's public transportation.

The study also identified some challenges, such as the K-Means model's suboptimal performance in comment clustering analysis, which could be attributed to factors such as dataset selection, model selection, data preprocessing, and feature extraction, providing valuable directions for future research.

Finally, data science has proved its important significance in practical research, becoming an indispensable tool for tackling real-world challenges, and we have every reason to think that data science technologies will continue to surprise us in future research undertakings.

**Reference**

David S. Moore, G. P. (2014). *Introduction to the Practice of Statistics.* W. H. Freeman, 2014.

Peter Van Roy, S. H. (2004). *Concepts, Techniques, and Models of Computer Programming.* Prentice-Hall.

Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.).

**Word count:** 3210