Insight into Ireland’s Rail Transport and How It Compares to Other European Member States.

Author: Emma McCready

Student ID: sba23001

MSc in Data Analytics

Postal address: Meenagowan, Lettermacaward, Co. Donegal, F94 F961.

E-mail: [sba23001@student.cct.ie](mailto:sba23001@student.cct.ie)

Github link: <https://github.com/EmmaMcC1802/CA2>

Word count: 3016

Table of Contents

[Table of Figures 3](#_Toc155995316)

[1. Abstract 4](#_Toc155995317)

[2. Introduction 5](#_Toc155995318)

[3. Methodology 6](#_Toc155995319)

[3.1 Datasets 6](#_Toc155995320)

[3.1.1 Reading in data 6](#_Toc155995321)

[3.1.2 Aggregating datasets 6](#_Toc155995322)

[3.2 Exploratory Data Analysis 6](#_Toc155995323)

[3.2.1 Data cleaning 6](#_Toc155995324)

[3.2.2 Data visualisation 7](#_Toc155995325)

[3.3 Data pre-processing 7](#_Toc155995326)

[3.4 Data analysis 7](#_Toc155995327)

[3.4.1 Statistical analysis 7](#_Toc155995328)

[3.4.2 Machine Learning 7](#_Toc155995329)

[3.4.3 Sentiment Analysis 8](#_Toc155995330)

[3.5 Visualising Machine Learning Model Results using an Interactive Dashboard 8](#_Toc155995331)

[4. Results and Discussion 9](#_Toc155995332)

[4.1 Insight into the Irish Rail Transport System 9](#_Toc155995333)

[4.2 Sentiment analysis of Iarnród Éireann 9](#_Toc155995334)

[4.3 Ireland vs EU: Passenger Number, and Investment into Infrastructure 10](#_Toc155995335)

[4.3.1 Statistical Analysis 12](#_Toc155995336)

[4.4 Ireland vs EU: Impact of other variables 12](#_Toc155995337)

[5. Conclusions 13](#_Toc155995338)

[6. References 14](#_Toc155995339)

# Table of Figures

[Figure 1. The annual passenger count on Irish Rail for the years 1981 to 2022. 9](#_Toc155995340)

[Figure 2. Sentiment Analysis of TripAdvisor Reviews. 10](#_Toc155995341)

[Figure 3. Sentiment Analysis of TrustPilot reviews, showing the amount of each sentiment category that was detected in each rating 1 - 5 out of 5. 10](#_Toc155995342)

[Figure 4. The amount of passengers travelling per kilometer in each EU Member State. The bottom plot excludes the 4 countries with the highest count to showcase the counts for the remaining MS. 11](#_Toc155995343)

[Figure 5. The monetary investment into rail infrastructure in each EU Member State. The bottom plot excludes the same 4 countries as Figure 4. to allow for consistent comparison. 11](#_Toc155995344)

# Abstract

*Open-access data collected by the public sector was used to explore the Irish rail system. Trends and public opinion were analysed to garner sentiment and use of the rail system, both in Dublin and outside of Dublin. Ireland was then compared to other European Member States, and relations were explored using different features. A supervised random forest regressor was most successful in predicting the total amount of passengers travelling on a train in a given year. The results of this model was built into an interactive dashboard.*

# Introduction

In a world in dire need of a shift towards sustainability, fit-for-purpose public transport is a necessity. This requires sufficient investment from the infrastructure government, but also requires appetite from the average person to use public transport. Within Ireland, the Central Statistics Office (CSO) collects information collects information on public transport use and investment, and publishes annual datasets with this information. For this reason, the use of the rail transport system in Ireland can be explored and analysed, and crucially to assess if there’s a relationship between factors such as infrastructure investment and passenger counts. Similarly, Eurostat is the statistical office of the entire European Union, and collects information on a wide range of areas, including transport, for each Member State (MS). Comparison between different MSs is thus possible, and showcases the Irish rail system’s performance within the EU.

# Methodology

## 3.1 Datasets

The data sets relating solely to Ireland were obtained from Ireland’s open-access governmental data repository (Ireland, The Open Data Unit, 2014). The scope was limited to passenger trains, and did not include trams. Two European datasets, one on passenger transport and the other on infrastructure investment, were obtained from the OECD Data Explorer (Organisation for Economic Co-operation and Development, OECD Data Explorer). Both OECD and Eurostat had to be used as OECD had data for more years, but Eurostat’s database contained a greater number of datasets. The links to each individual dataset can be obtained in Codebook 1 for all datasets except those which were sourced from Eurostat, which are linked in Codebook 2. All data were obtained from open-access repositories, which is free to use. This is additionally beneficial as it allows for reproducibility of the analyses within this project. The advantage of the chosen repositories is that they are reputable sources and the data is already in an accessible format in its raw form.

### 3.1.1 Reading in data

Most datasets were downloaded as raw .csv files, and were read in as dataframes using Pandas. One dataset was read into Codebook 1 in Section 4 in the .JSON format. Specific libraries had to be used, and overall took 6 lines of code whereas reading in .csv files only takes one. However, .JSON format has the advantage that it’s linked to the source database via a URL, so it’s automatically updated in real-time, whereas a .csv file would have to be downloaded manually and read in again when the data is updated. In Codebook 2, 17 individual datasets were downloaded as .csv files from EUROSTAT. For efficiency, the glob module was used to find all the relevant files and then they were loaded in and aggregated in the same for loop. This was much quicker and easier than loading in .csv files one-by-one.

### 3.1.2 Aggregating datasets

The benefit of loading in .csv files using the glob module was that they could be aggregated in the same ‘for loop’. This was much more efficient in terms of code, and also made the   
Codebook much tidier. The ‘for loop’ could then be improved to improve the way it was aggregated during EDA, which was much easier than finding and editing multiple lines of code. The other method I used to aggregate was through the .merge() function, which involved planning ahead and taking extreme caution that the wrong merge wasn’t used (i.e. using left merge instead of inner).

## 3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was employed throughout the project to explore each dataset, primarily throughout Codebook 1. The process was cyclical and revisited throughout each analysis as more information was discovered about a dataset. A typical approach was taken, wherein subsections of the dataset were viewed to gain preliminary insight, and then the data was cleaned before being visualised. For the large dataset used to build the machine learning models, a library designed for automating EDA and generating reports, YData Profiling, was used (See Section 2, Codebook 2). The resulting reports are available in the project’s Github repository.

### 3.2.1 Data cleaning

The data was cleaned to remove unnecessary information and make the datasets more accessible for analysis. Missing data was explored to identify why the data was missing, and then removed if necessary. In Section 3 of Codebook 2, the efficiency of data cleaning was assessed using a YData profile.

### 3.2.2 Data visualisation

Visualisations were used throughout each Codebook to explore data, to illustrate results and to validate analyses. Simple plot types were preferred to allow ease of interpretation by non-experts. Line plots were useful to show trends over time. To show statistical information, both histograms and boxplots were favoured. For the Eurostat dataset, a correlation matrix heatmap was used to showcase the correlations between variables, as there were too many variables for a pair plot. Section 6.3, Codebook 1 contains an interactive plot which is accessible and easy to interpret, with a colour gradient included.

Plots were designed in accordance with Tuft’s Principles, and care was taken not to add any additional and unnecessary features to plots. The plots were all labelled with axes titles and plot titles, and legends were used where required. The default colour scheme in Seaborn was used as this is in line with what viewers are most accustomed to, and it’s optimised for representing categorical data. A conscious effort was made to preserve assigned colours across variables during the same subsection of analysis. For example, in Section 6, Codebook 1, specific colours were mapped to categorical data, but given the volume of different labels a palette had to be used from the colorcet library. Moreover, it was ensured that there was adequate spacing if subplots were too close together.

## 3.3 Data pre-processing

The method of data wrangled was catered to the specific requirements in each analysis. When choosing pre-processing methods, e.g. for feature normalisation, consideration was given to the analysis itself and the requirements of the model. Rationale is explained in the relevant sections in the codebook.

## 3.4 Data analysis

All analysis was conducted using Python in JupyterLab Notebooks, which are found in the supplementary codebooks available on Github. For statistical analysis, the SciPy library was used. For Machine Learning, Scikit Learn was used. An ongoing issue with the datasets concerned the distribution, as no feature was normally distributed as a result of outliers and yearly variability. Thus, models were either chosen taking variability and outliers in the dataset into account.

### 3.4.1 Statistical analysis

*See Section 1 of Codebook 2.* The focus of this analysis was to assess the relationship between the amount of passengers travelling per kilometre and infrastructure investment across the EU, with a focus on Ireland. A Student’s T-test was used to calculate the average annual number of passengers, but the data had to be normalised by bootstrapping it first. A number of inferential statistic tests were used to gain insights on possible population values, which are discussed in Section 4.3.1 of this report. In general, parametric inferential statistical techniques were avoided owing to the lack of normally distributed data. Attempts were made to normalise data using multiple different techniques, for example by log transforming it and Z-score normalisation. After each attempt, a Q-Q plot was used to find the type of distribution, and the Shapiro-Wilk test was used to test the normality. To test the hypothesis that there is a linear relationship between the number of passengers transported per kilometre and infrastructure investment, and used a linear regression model. Two different methods of standardising data were used to compare their impact on the model.

### 3.4.2 Machine Learning

*See Section 2 of Codebook 2.* For machine learning, a large dataset containing features both directly related and unrelated to rail transport in the EU was collated with a goal of checking for correlation between features. A correlation matrix heatmap was used to give an initial insight. A supervised machine learning model was used to predict the amount of passengers travelling on a train in a MS in a given year. Random Forest Regressor and Support Vector Regression models were selected for this purpose, as they tend to be more capable of handling missing values and are less sensitive to the magnitude of variables. The model was normalised using standardisation to make the model more efficient and converge faster, and standardisation was chosen as it’s less sensitive to outliers. The countries were one-hot encoded to include the data in the model but also to avoid an artificial relationship that isn’t actually there. In each model, hyperparameters were altered to optimise the model and cross validation was used to the machine learning models to evaluate its performance. An unsupervised machine learning model was used to identify hidden relationships between features. For this, the data was first clustered using KMeans, and then dimensionality reduction via Principle Component Analysis (PCA) was applied. The clusters were then visualised.

### 3.4.3 Sentiment Analysis

The full sentiment analysis is located in Codebook 3. The method was used to determine and categorise the sentiment expressed in the text of reviews of Iarnród Éireann, from two popular websites: Trustpilot (Trustpilot, 2024) and TripAdvisor (TripAdvisor, 2024). These websites were scraped using a web scraping software, Octoparse (Octoparse, 2024), for the purpose of efficiency. The extracted reviews were pre-processed by removing unnecessary characters and punctuation to reduce noise, and making text lowercase to ensure uniformity. It was then tokenised (broken into smaller parts) and stop words were removed. A stemmer was then applied to remove word suffixes to normalise text. A module from the Natural Language Toolkit (NLTK) library, VADER (Valence Aware Dictionary and sEntiment Reasoner), was used to analyse the sentiment of each review and label them as either “positive”, “negative” or “neutral”. The output was then visualised using bar charts to observe trends in the sentiment.

## 3.5 Visualising Machine Learning Model Results using an Interactive Dashboard

*See Codebook 4.* The best performing supervised machine model (Random Forest Regressor) in predicting the annual total passenger count was identified and included in a dashboard. Difficulties were encountered at this phase in the analysis, as the Windows computer used throughout the project was no longer accessible and the project had to be finished on a Mac. Owing to the differences in system architecture and operating system, different results are observed for machine learning models despite the a constant random state value. As the results had already been explored on the Windows computer, the decision was made to create the dashboard on a separate Codebook. Interestingly, the model performed much better on Mac. Different approaches were attempted when creating this dashboard but ultimately were not successful.  
A simple user friendly dashboard was designed with Tuft’s Principles in mind, with different tabs included so that the dashboard doesn’t contain too much information at once. This allows for easier interpretation. The plots had a minimalist approach as to erase non-data ink. A blue colour scheme was followed as it’s the default colour used throughout the project. The first tab is just text and serves as a key for the labels in the other two tabs. The second tab is a feature importance plot, showcasing which features were most important as a predictor for the desired feature. The final tab contains a plot with the actual versus the predicted values.

# Results and Discussion

## 4.1 Insight into the Irish Rail Transport System

Data collected by both the CSO and Iarnród Éireann revealed interesting trends in the use of trains in Ireland. Figure 1 illustrates the total amount of passengers annually from 1981 to 2022. This total seemed to be on the updated trend until around 2009, where there was a slight dip, but still peaked at 2019 before there was a dramatic drop in the annual passenger count. It can be assumed that the former was a result of the economic crisis, and that the latter decrease was a result of COVID-19. The data from 2022 shows that it hasn’t returned to pre-pandemic levels. The boxplot in this figure proves the variance in the data, but notably there are no outliers. This information is explored more in Section 3 of Codebook 1, including a breakdown of train journey types that accounted for the numbers. Further, Section 4 of the same Codebook explores the breakdown of journey types through visualisations. This information is included in Codebook 1 but not the report, as it’s not strictly relevant to the analysis, but still is interesting.

A graph of passenger passengers travelling by rail

Description automatically generated

Figure 1. The annual passenger count on Irish Rail for the years 1981 to 2022.

An attempt was made to explore the impact of infrastructure on passenger counts in Section 5 of Codebook 1, but infrastructure seems to have remained fairly constant throughout the years, so it was not beneficial to explore any relationship.

## 4.2 Sentiment analysis of Iarnród Éireann

*See Codebook 3.* The text of 411 reviews of Iarnród Éireann were scraped from TripAdvisor. Of this, approximately 74% of the reviews contained mostly positive sentiment (Figure 2.). This corresponds relatively well with Iarnród Éireann’s rating on TripAdvisor, which was 3 out of 5 stars.

A graph with different colored squares

Description automatically generated

Figure 2. Sentiment Analysis of TripAdvisor Reviews.

In the absence of an individual rating corresponding to each review, it was difficult to evaluate the performance of the sentiment analysis model. A second sentiment analysis was conducted on a different website for this reason, this time on TrustPilot. There were fewer reviews (n = 55), but they had a rating attached, allowing for the analysis model to be evaluated. The results are shown in Figure 3, and displays an interesting and expected trend. Higher ratings had higher positive sentiment, and the opposite was true for negative ratings. Neutral sentiment was fairly constant across all ratings.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 3. Sentiment Analysis of TrustPilot reviews, showing the amount of each sentiment category that was detected in each rating 1 - 5 out of 5.

## 4.3 Ireland vs EU: Passenger Number, and Investment into Infrastructure

When comparing Ireland to the rest of the EU, a focus was put on passenger counts and the amount of euros invested into infrastructure annually.

A graph of different colored lines

Description automatically generated

Figure 4. The amount of passengers travelling per kilometer in each EU Member State. The bottom plot excludes the 4 countries with the highest count to showcase the counts for the remaining MS.

A graph of different colored lines

Description automatically generated

Figure 5. The monetary investment into rail infrastructure in each EU Member State. The bottom plot excludes the same 4 countries as Figure 4. to allow for consistent comparison.

### 4.3.1 Statistical Analysis

*See Section 1, Codebook 2.* In order to investigate the relationship between passenger counts and infrastructure investment, statistical tests were used. The full list of tests and rationale can be located in the relevant sections of Codebook 2. Following a linear regression analysis wherein different methods of normalisation were employed, it was discovered that there is a decent correlation at lower investment values, however as investment increases the data points are a lot more dispersed and there are a lot of outliers. Other points of interest are that there is a statistical difference in the distribution of infrastructure investment among Ireland and all other EU Member States, which isn’t surprising when Figure 5 is referred to. There is a clear variability across the EU which proved to be a challenge, so a focus was put on Denmark to compare it to Ireland as it is relatively similar in terms of land size and population. However, it was concluded that there is a statistically much higher amount of passengers travelling per kilometer, implying they have a larger, more popular rail system than Ireland. Another area of interest is the decrease in train use pre- and post- the pandemic, and it was concluded that there is not enough evidence to conclude a significant difference in the number of passengers transported per kilometer in Ireland before and after the pandemic.

## 4.4 Ireland vs EU: Impact of other variables

*See Section 2, Codebook 2 for full explanation of the data involved and the rationale used.* One goal of the project was to train a machine learning model to predict the annual passenger count using a range of features, both directly related to rail and also not. Two models were used, Random Forest Regressor and Support Vector Regression, as these two models are less susceptible to the missing values in the dataset. Random Forest Regressor was the best fit and most accurate following hyperparameter tuning with the GridSearchCV method. The results of the tuning can be seen in Table 1. The full explanation of these results can be found in the Section 2.1.2 of Codebook 2. However, there were challenges applying hyperparameter tuning to the Support Vector Regression model, as this model was a lot more demanding on computer resources and it took a long time for the cell to be executed if at all.

*Table 1. Evaluation of the fit of machine learning method following hyperparameter tuning.*

|  |  |  |  |
| --- | --- | --- | --- |
| Hyperparameters | R2 (test set) | Root Mean Squared Error | Mean Absolute Percentage Error |
| 'max\_depth': 100, 'n\_estimators': 100 | 0.98990 | 54063.98574 | 10.826072 |
| 'max\_depth': 100, 'n\_estimators': 10 | 0.98652 | 62444.61232 | 13.30183 |
| 'max\_depth': 75, 'n\_estimators': 20 | 0.984165 | 67695.06055 | 13.02569 |

# Conclusions

One goal of this analysis was to showcase the importance of the investment of public funds into infrastructure. It was difficult to show the relationship between this factor and annual passenger counts, due to variability and outliers in the data. A second goal of this project was to make the data accessible to non-data analysts, which was achieved through the use of an interactive plot and dashboard. Using sentiment data extracted from two popular websites for reviews, it is shown that the public opinion of Irish Rail, the national rail provider, is mixed but leaning towards positive.

# References

Organisation for Economic Co-operation and Development, OECD Data Explorer, *OECD Data Explorer*, available: <https://data.oecd.org/> [accessed 16 Dec 2023].

Ireland, The Open Data Unit (2014) *IRELAND'S OPEN DATA PORTAL*, available: <https://data.gov.ie/> [accessed 16 Dec 2023].

Octoparse (2024) *Easy Web Scraping for Anyone*, available: <https://www.octoparse.com/> [accessed 5 Jan 2024].

TripAdvisor (2024) *Irish Rail*, available: <https://www.tripadvisor.ie/Attraction_Review-g186605-d7940415-Reviews-Irish_Rail-Dublin_County_Dublin.html> [accessed 5 Jan 2024].

Trustpilot (2024) *Iarnród Éireann Irish Rail*, available: <https://ie.trustpilot.com/review/irishrail.ie> [accessed 5 Jan 2024].