Insight into the Irish Rail Transport System Compared to 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>

Table of Contents

[1. Abstract 5](#_Toc155902321)

[2. Introduction 6](#_Toc155902322)

[3. Methodology 7](#_Toc155902323)

[4. Results and Discussion 8](#_Toc155902324)

[5. Conclusions 9](#_Toc155902325)

[6. References 10](#_Toc155902326)

No table of figures entries found.

Abbreviations

CSO Central Statistics Office

MS Member State

# Abstract

# 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). 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).

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. The raw data datasets were downloaded in full in the .csv format, except for one dataset which was loaded in as .JSON. Pandas was used to load in the former, and specific libraries were used for the latter (json, pyjstat).

### 3.1.1 Reading in data

All .csv files were read in using Pandas.

Glob used…

.JSON format..

## 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

Simple plot types were preferred to allow ease of interpretation by non-experts. Plots were designed in accordance with Tuft’s Principles. Line plots were used to showcase a variable over time.

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.

To show statistical information, both histograms and boxplots were used.

Visualisations were also used to both validate analyses and illustrate results.

Interactive plot…

Dashboard…

## 3.3 Data pre-processing

The method of data wrangled was catered to the specific requirements in each analysis.   
  
for sentimental analysis, (explain preprocessing steps..)

## 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.

Explain how I applied Tuft’s Principles, esp for the dashboard..

### 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 ….

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.

Section on programming ? or build it into the above?

# Results and Discussion

## 4.1 Insight into the Irish Rail Transport System

Data collected by both the CSO and Iarnród Éireann were used to explore Ireland’s

### 4.1.1 Popular journey types, 2019 and 2021

### 4.1.2 Infrastructure

## 4.2 Sentiment analysis of Iarnród Éireann

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

## 4.4 Ireland vs EU: Impact of other variables

# 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.

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].