A picture containing icon

Description automatically generated

**An Analysis of Ireland’s Current Passenger Rail Transport**

A Report Submitted in Partial Fulfilment of the requirements for the Degree of MSc Data Analytics

November 2023

*Supervisors:*

Dr. Muhammad Iqbal

David Gonzalez

David McQuaid

Marina Iantorno

*Cathal Nugent*

*CCT College, Dublin*

Word Count:

**Contents**

[Background 3](#_Toc155370936)

[Planning 3](#_Toc155370937)

[1: The Current State of the Irish Rail Network 4](#_Toc155370938)

[1.1: Introduction 4](#_Toc155370939)

[1.2: Data Analysis (EDA) 4](#_Toc155370940)

[1.3: Geographical Visualisation 5](#_Toc155370941)

[1.4: Interactive Map 5](#_Toc155370942)

[1.5: Choropleth Map 6](#_Toc155370943)

[2: Analysis of Dublin’s ‘Luas’ system 7](#_Toc155370944)

[2.1: Introduction 7](#_Toc155370945)

[2.2: Passenger Journeys 7](#_Toc155370946)

[2.3: Accessibility 8](#_Toc155370947)

[2.4: Reliability 9](#_Toc155370948)

[2.5: Hour of the Day 9](#_Toc155370949)

[2.6: Data Preparation 10](#_Toc155370950)

[2.7: Inferential Statistics 10](#_Toc155370951)

[2.8: Random Forrest Regression Model 10](#_Toc155370952)

[2.9: SVM Model 11](#_Toc155370953)

[2.10: Predictions 11](#_Toc155370954)

[3: Comparable Tram Systems 12](#_Toc155370955)

[**Section 1: Edinburgh** 12](#_Toc155370956)

[**Section 2: Melbourne** 13](#_Toc155370957)

[**Section 3: UN Tram and Light Rail Data** 15](#_Toc155370958)

[4: Sentiment Analysis 18](#_Toc155370959)

[4.1: Introduction 18](#_Toc155370960)

[4.2: Reddit API 18](#_Toc155370961)

[4.3: Pre-processing the Data 18](#_Toc155370962)

[4.4: VADER Sentiment Analysis 19](#_Toc155370963)

[4.5: Self-Assessment of Model 20](#_Toc155370964)

[4.6: Word Cloud 21](#_Toc155370965)

[Conclusion: 22](#_Toc155370966)

[Appendix 1: Examination Criteria 23](#_Toc155370967)

[Appendix 2: Planning 26](#_Toc155370968)

[Bibliography 27](#_Toc155370969)

## Background

My analysis will focus on current passenger rail transport in Ireland. I will compare and contrast aspects of Ireland’s current rail network to a country with a similar rail network.

I will also analysis Ireland’s new proposed rail network and compare this to a country with a similar rail network. I will use the CRISP-DM Project management framework whilst completing this analysis. This is a widely used framework and is appropriate for the task at hand.

## Planning

When first reading the subject matter for this project I immediately though about an article I had read on LinkedIn just a few weeks prior. It detailed ‘The All-Ireland Strategic Rail Review’ which was published by Iarnród Éireann. A link to the article can be found here:

Article Link: [All-Island Strategic Rail Review](https://www.linkedin.com/feed/update/urn:li:activity:7089642579767517186?updateEntityUrn=urn%3Ali%3Afs_feedUpdate%3A%28V2%2Curn%3Ali%3Aactivity%3A7089642579767517186%29)

This article seemed to generate a lot of attention and discussion and I thought it would be the perfect subject matter to base my project on. However, as I delved deeper into my research, I soon came across a number of obstacles and setbacks which forced me to rethink the subject matter of my project. I will discuss details of this over the course of my investigation. You may also find details of this in Appendix 2, which outlines the planning process and timelines I had set for this project.

## 1: The Current State of the Irish Rail Network

### 1.1: Introduction

The Irish Rail network has unfortunately been in a state of decline over the last 100 years. Mainly due to the increased popularity of personal cars. However, the automotive industry is currently in a state of transition, with increasing greenhouse gas regulation and a population who are now more aware and more concerned about the impact of their carbon footprint than ever before, we could potentially see the popularity of rail transport skyrocket once again.

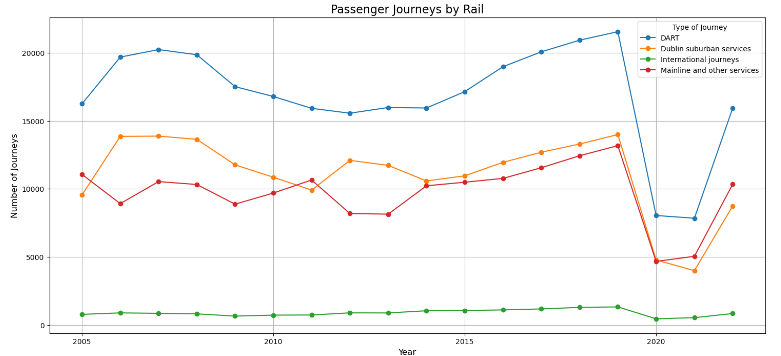
I would firstly like to first assess details of Ireland’s current rail network by analysing some public datasets available on the Central Statistics office (‘CSO’) website.

### 1.2: Data Analysis (EDA)

The first dataset I will be taking a look at will be looking at Passenger Data by Rail. This dataset has been downloaded from the CSO website and informs us on the amounts of passengers taking certain rail journeys split into 4 main groups which are DART, Mainline, Dublin Suburban and International Journeys.

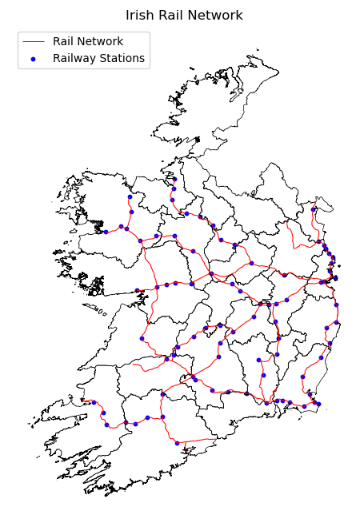
Whilst this data is not as granular as I would have liked it is the most granular dataset, I have access to the moment. I did request a more granular dataset from Irish Rail but unfortunately, I did not receive a response.

After reading in the data, I trimmed the data to only include appropriate and necessary columns. I performed a number of EDA (Early Data Analysis) tasks in order to get comfortable with this dataset. After this, it is clear that the complete dataset in only available in it’s entirely from 2005 onwards so I took this decision to exclude any years prior to 2005.

I then visualised the data using an appropriate line chart:

From this chart we can clearly see the dramatic impact that COVID-19 had on the number of passenger journeys made by train, this will be a recurring trend we will see going forward. We are seeing journeys increase once again but it is still apparent that that have not yet reached pre-covid level.

### 1.3: Geographical Visualisation

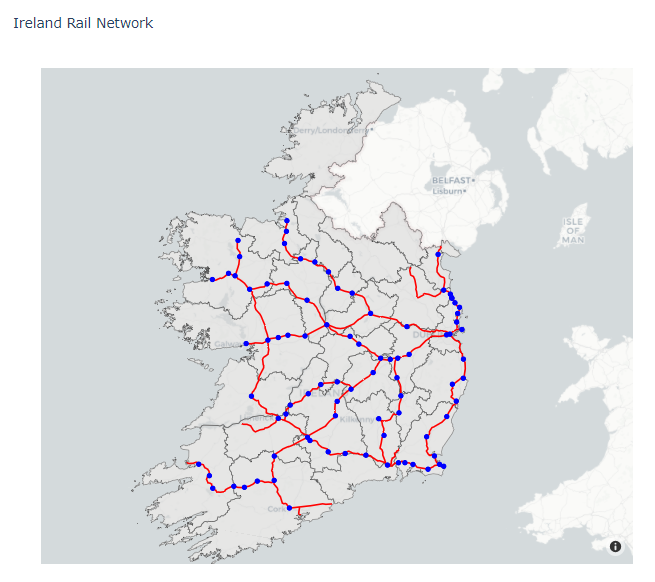
I felt that it was necessary to visualise the current rail network in the Republic of Ireland. I began by sourcing a number of GeoJSON files online outlining Ireland’s Rail Network, Railway Stations and County Borders. Sources for these files can be found in the Jupyter notebook.

I began by reading in each GeoJSon file and plotting the current rail network as it currently exists, detailing each public station with a blue dot and each trainline with a red line.

Whilst this is a good visualisation of Ireland’s railway system, I wanted to take it a step further and create an interactive map. Allowing the user to freely move take a more detailed look at the current rail network.

### 1.4: Interactive Map

I spent a lot of time researching interactive mapping techniques in Python and after a lot of effort I was able to manipulate the GeoJSON file in an appropriate way in order to achieve the desired outcome. A detail commentary of the code can be found within the Jupyter notebook.

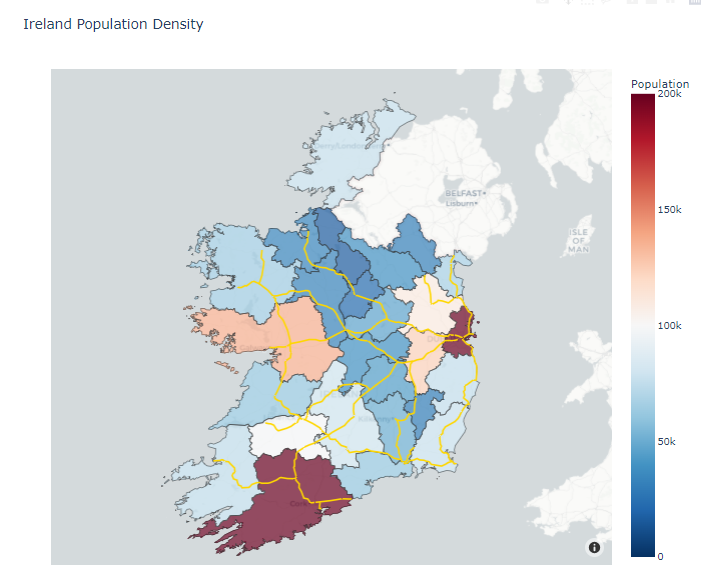


### 1.5: Choropleth Map

I was keen to take this interactive map a step further in order to try and visualise the population density of Ireland. In order to do this I firstly had to source an appropriate dataset which matches the scale I am aiming for.

I was able to find a population dataset which breaks down population of each county from the CSO website. I could join this to my existing Counties GeoJSON file and create a choropleth map based on the 26 county regions. After reading in the data I performed some basic data manipulation to aggregate the data into each of the 26 counties.

After this data manipulation, I merged the population data onto the Counties Geographical dataframe in preparation for creating my interactive choropleth map.



I think the most notable think here would be the lack of railway infrastructure present in Donegal. The border counties are also noticeably barren in this regard but given the lack of population here it could be somewhat justifiable.

It is also noticeable how other population hubs are quite inefficiently linked to each other, such as Galway and Cork. It is clear that the current rail system is extremely centre around Dublin which could be understandable given the huge population discrepancy that is present.

## 2: Analysis of Dublin’s ‘Luas’ system

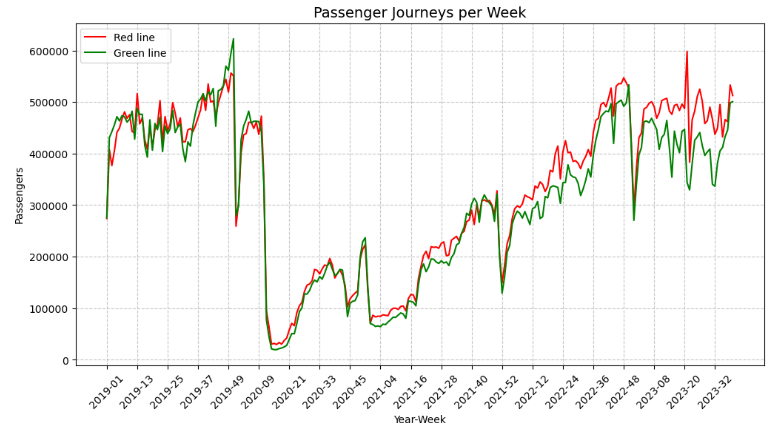
### 2.1: Introduction

After struggling to find sufficiently granular data needed to perform an adequate investigation of Ireland’s current rail network, I was forced to shift my focus elsewhere. After some research I decided that there was enough available data to investigate Dublin’s current tram system (the Luas).

The Luas is currently comprised of two lines, a red line and a green line. These lines have been operational since 2004 and are generally used for commuting purposes in the greater Dublin area.

### 2.2: Passenger Journeys

In order to understand the current situation regarding the Luas, I wanted to take a closer look at the recent trends regarding the number of passenger journeys made.

I was able to find a dataset on the CSO website outlining this exact topic. After analysing and performing some basic manipulation on the dataset, I was able to plot an appropriate graph showing trends in the number of passenger journeys on the Luas over the last number of years.

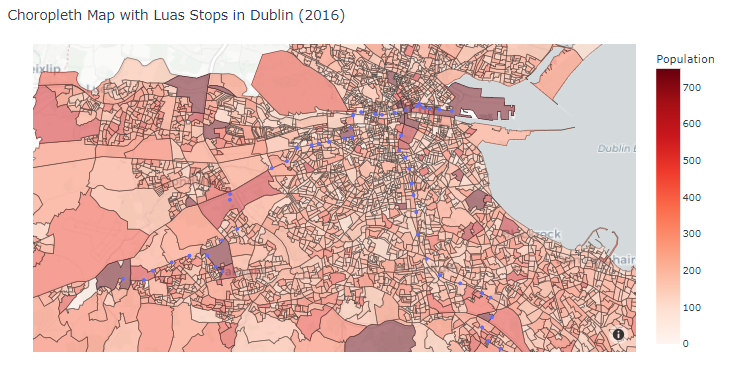
We can see that the number of journeys completed on both the red and green lines are quite similar to each other and often mimic each other in terms of ebbs and flows. Similar to our earlier analysis of Ireland railways, we can clearly see the significant impact that COVID-19 had on the number of passenger journeys.

### 2.3: Accessibility

The next aspect of the Luas I was keen to investigate was how accessible the lines where to the general public of Dublin. After some research, I realised that it was possible to create a choropleth map using data from Ireland’s small area census, which was last performed in 2016.

I was able to find a GeoJSON file online outlining the “small areas” as they were in 2016. I then trimmed this data to only include necessary areas of interest to us, such as Dublin City, South Dublin, Dún Laoghaire-Rathdown and Wicklow. These areas are quite subjective, and the code can be easily manipulated to include a wider range surrounding the Dublin area.

I then merged the population data from the small area census to this dataframe and we had all the necessary data to plot our visualisation. As evident in Notebook 1, I had a lot of practice, so I was able to rely heavily on the research I had performed earlier.



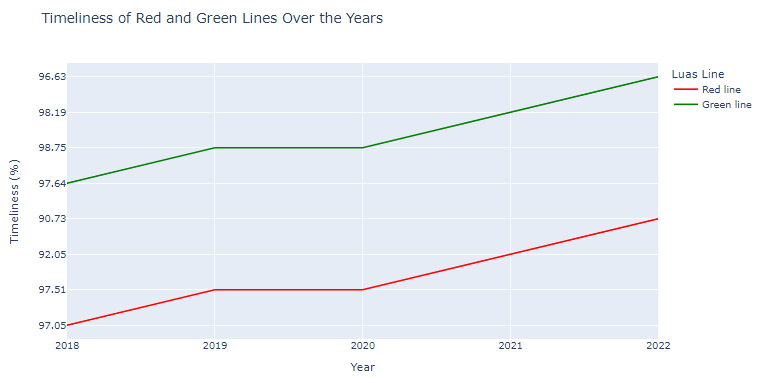
From this visualisation, we can gain a great insight into the current accessibility of the Luas and potential for future development through extensions or indeed new lines. After analysing this map, I arrived on a number of potential developments regarding the current Luas system. More details on this can be found in my Jupyter notebook.

1. Extensions to Blanchardstown, Leixlip – It is clear that a reasonably small offshoot of the red line could satisfy a large population centre to the east, primarily Blanchardstown.
2. Dublin Airport – This would be a huge project but after researching the Edinburgh Light Rail system (Section 3.1 – 3.5), I can clearly see how positive this could potentially be.
3. Greystones, Bray – After analysing the southern tip of the Luas line it is quite clear that a reasonably small extension to the green line could service areas like Greystones and Bray.
4. North Dublin – The complete lack of Luas lines in north Dublin is quite strange, this is just a general observation but could be
5. Aviva – As referenced earlier, the Edinburgh light rail system has seen success less as a commuter transport and more tourism/events focused. The Luas system could take a leave out of this playbook and link to big event centres like the Aviva, Croke Park and even universities like UCD or DCU.

I purposely came up with these recommendations independently from any planned future developments, but it is clear that some of my thoughts on this are shared by others with plans already. More details of this can be found in the article below:

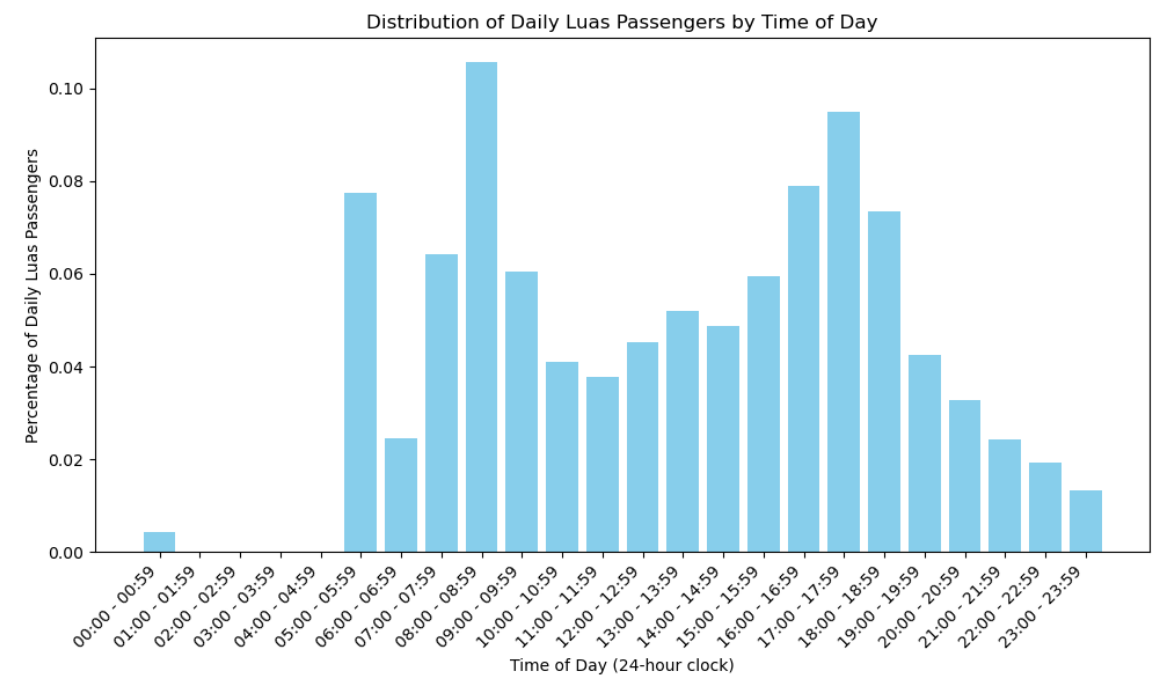
[Plans for new Luas lines and Metrolink published by National Transport Authority](Plans%20for%20new%20Luas%20lines%20and%20Metrolink%20published%20by%20National%20Transport%20Authority)

### 2.4: Reliability

I then performed a quick exercise to analyse the Luas’s reliability, resulting in the plot below.

### 2.5: Hour of the Day

With this new dataset I wanted to plot the popularity of each Luas line per hour of the day. So I manipulated the dataset and plotted the below line chart.



We can clearly see strong evidence of the Luas’s primary use as commuter transport, with peaks around 9a.m and 5p.m.

### 2.6: Data Preparation

For this section, I was keen on creating an appropriate dataset that could be used to perform predictive modelling. I noticed that I had one dataset which outlines the number of passenger journeys per day and another dataset which outlines the percentage of daily passengers per hour of the day. By assuming that the percentage data was appropriate for each and every week within that year I could create a large dataset outlining the number of passenger journeys made for a certain hour of any given week. Of course this is a reasonably crude approximation, but I did attempt to request the more precise data from Luas themselves, but I was once again met with no response.

### 2.7: Inferential Statistics

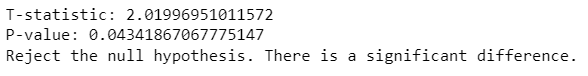
Before proceeding with my predictive modelling as intended, I wanted to perform a few statistical tests on my dataset using a hypothetical scenario. I wanted to gain insight into my data by calculating confidence intervals surrounding my average passengers per time-of-day data.

**Test 1: Confidence Intervals**

The code itself is easily adjustable to whatever the user might desire. Below is an example of the number of passengers using the Luas between 17:00 - 17:59 on any given day:

**Test 2: T-test**

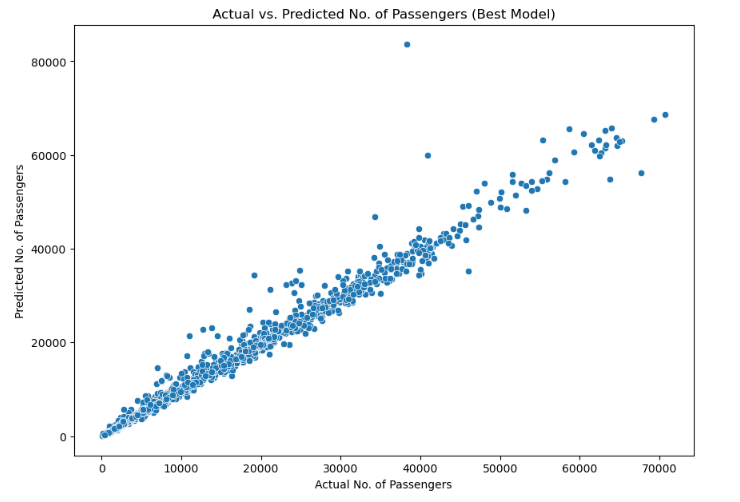
For my next test I thought of a hypothetical scenario, I decided to test whether the average number of passengers is significantly different between the red and green lines. I began by separating out my data and performing the t-test using the ‘ttest\_ind’ function. The results of this test can be found below:



### 2.8: Random Forrest Regression Model

Next, I wanted to create a predictive model to predict the number of passengers on any given hour of the day. I firstly encoded my data and then trained my model appropriately using a random forest regression model. The initial model I create appeared to perform well with a mean-squared score of 854.58.

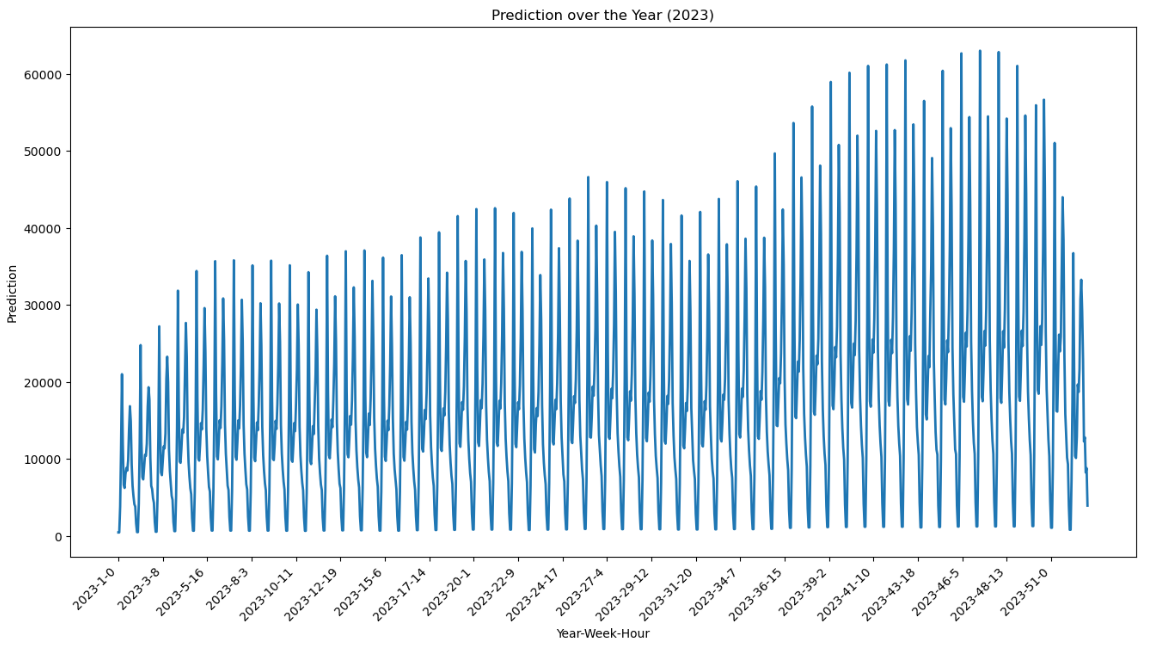
I also plotted a scatter plot based on the predicted and actual values which looks very promising, showing a clear diagonal line which tells us that the model is accurate. Despite this, I used GridSearchCV to tune my model in an attempt to achieve an even higher degree of accuracy, which was successful. The best model parameters and scatter plot are shown below:



### 2.9: SVM Model

I experimented with numerous alternate models such as a linear regression model, lasso regression and SVM model with predictably poor results.

### 2.10: Predictions

I then created a dummy dataframe to make predictions for 2023 based on my best model. The results of this can be found below:

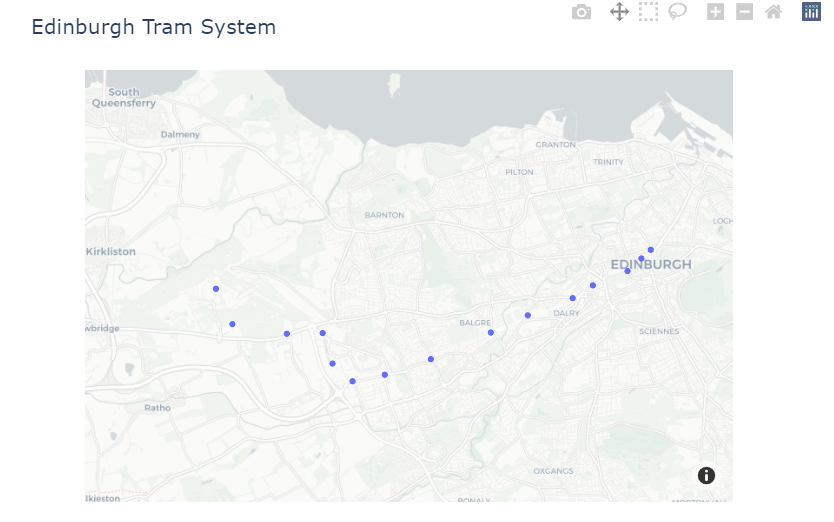
## 3: Comparable Tram Systems

### **Section 1: Edinburgh**

#### 3.1: Introduction

In order to choose an appropriate tram system to compare with the Luas, I had to consider a number of factors, including population, population density, current rail infrastructure amongst other things. I spend a significant amount of time researching in order to make an appropriate selection and I concluded that the most suitable system for comparison was Edinburgh’s Light Rail system.

#### 3.2: Geographical Visualisation

I firstly visualised the tramline in order to get a better understanding of its purpose and scale.

#### 3.3-3.5: Key Statistics

I then sourced a number of key statistics related to the tramline, with the aim of performing a comparison to the Luas. I sourced a number of statistics but was unfortunately left disappointed by the level of granularity in the data available. Despite researching heavily for more data I was forced to accept that there was very limited available data, so I went back to the drawing board.

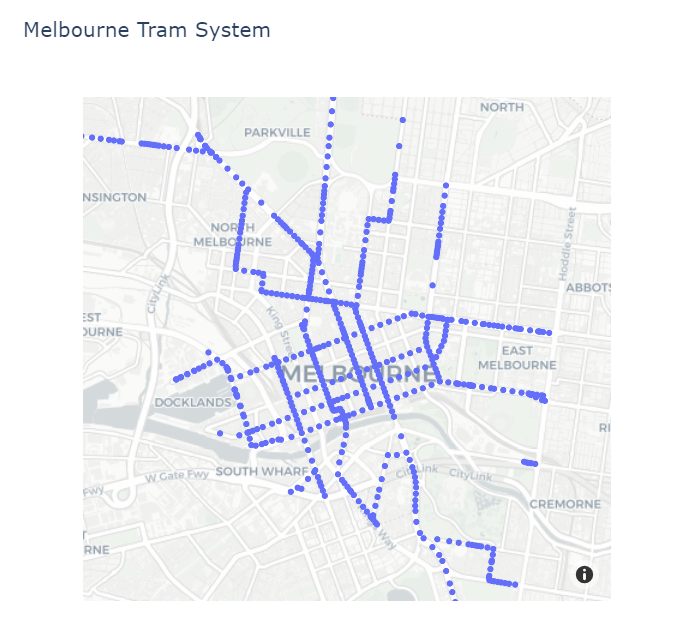
### **Section 2: Melbourne**

#### 3.6: Introduction

After a disappointing lack of data was available for Edinburgh’s tram, I was forced to turn my attention elsewhere. This time I prioritised data availability before I committed to my comparison work. After some research I found that Melbourne tramlines would be a great system for comparison and the local government in Melbourne place a huge emphasis on open data. So this time it would be a lot easier to collect and analyse data.

#### 3.7: Geographical Visualisation

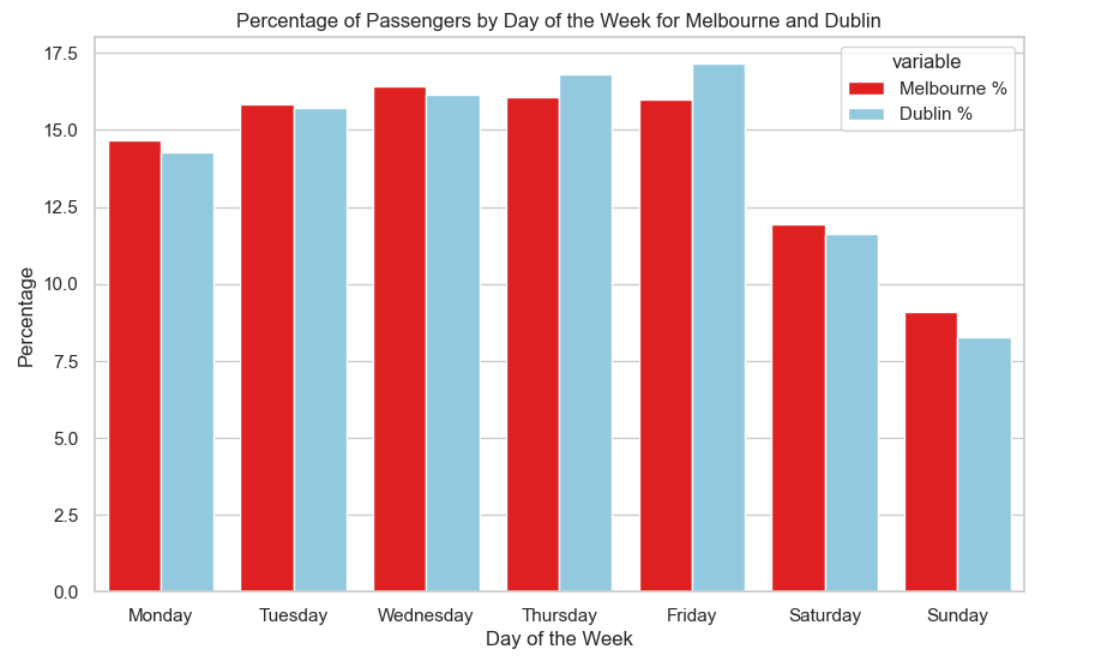
As is typical, I plotted a geographical visualisation to get a better understanding of the scale and comparability of Melbourne’s tram system.



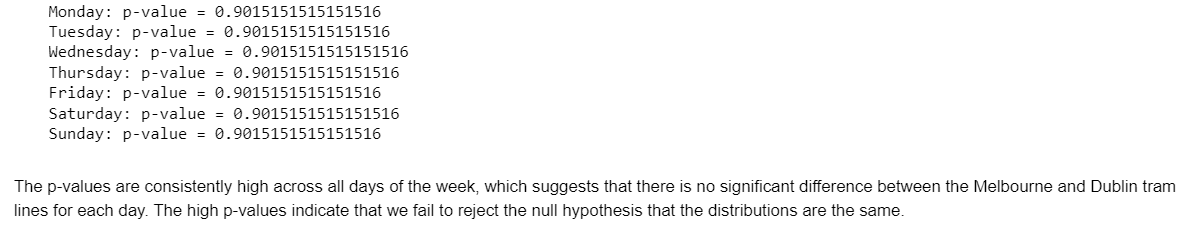
I made the decision to plot this map using centroid positions of each tramline as it significantly reduces the processing power required to generate such a map, whilst still giving us a good gauge of the size and scale of the network.

#### 3.8: Passenger Numbers by Days of the Week

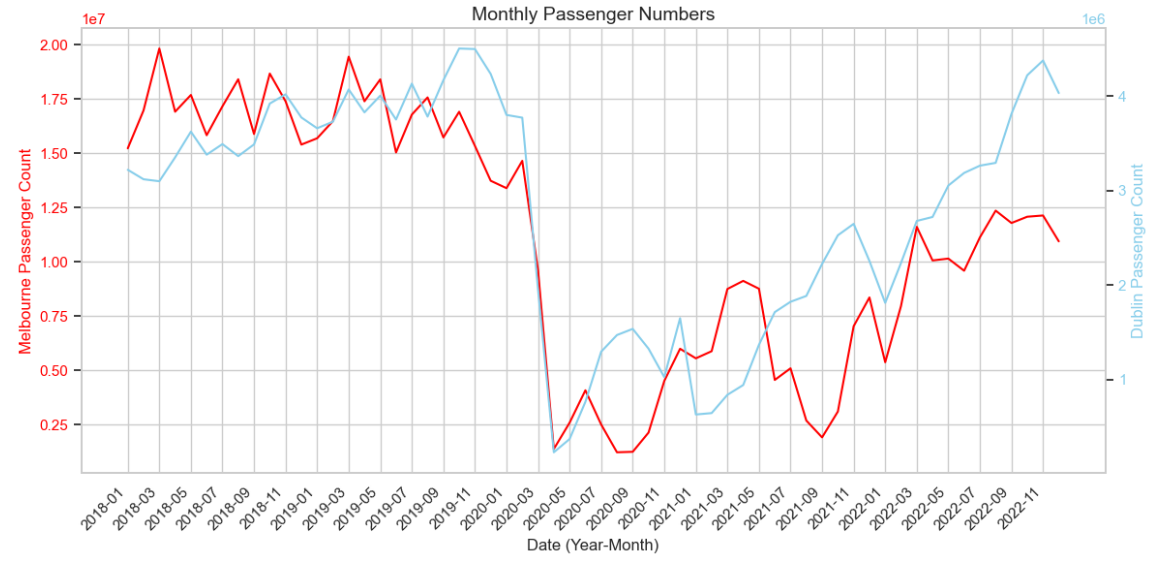
Naturally, I then gathered data on the number of passengers per day of the week in order to compare this to the Luas system. I collected, manipulated, and aggregated my data until it was in an ideal format for visualisation. I decided that a bar chart would be most appropriate for this scenario.



From the chart above we can see that weekly patterns are very similar between the two tram systems. This led me to complete a Mann-Whitney test on our data. The results of which can be found below.

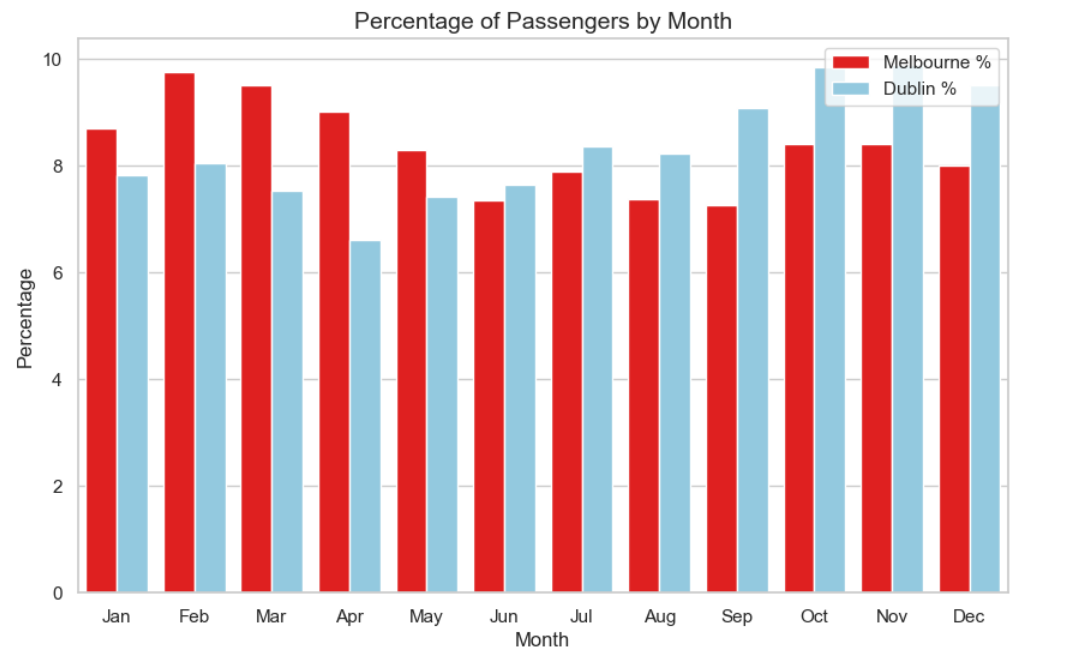


#### 3.9: Recent Trends in Passenger Numbers

I also wanted to take a closer look at the trends in passenger numbers over recent years. In particular I wanted to see the impact COVID-19 had on passenger journeys.

Interesting to see the trends in passenger number are actually quite similar despite the scale differences. Following this, I performed a both a t-test and a Kolmogorov-Smirnov test on the 2 datasets, more details of these tests can be found in the Jupyter notebook.

#### 3.10: Average Passengers by Month

I then plotted a stacked bar chart showing percentage of passengers by month, very similar to the one I had plotted in Section 3.8.

#### 3.11: Reliability

I then took a quick look at some reliability (or timeliness) statistics for the two tramlines. Results of this are shown below:



### **Section 3: UN Tram and Light Rail Data**

I was able to source a dataset from the UN website detailing passenger volumes and passenger kilometres for major cities and region around the world. I thought this dataset would provide myself with a great opportunity for visualisations and statistical testing.

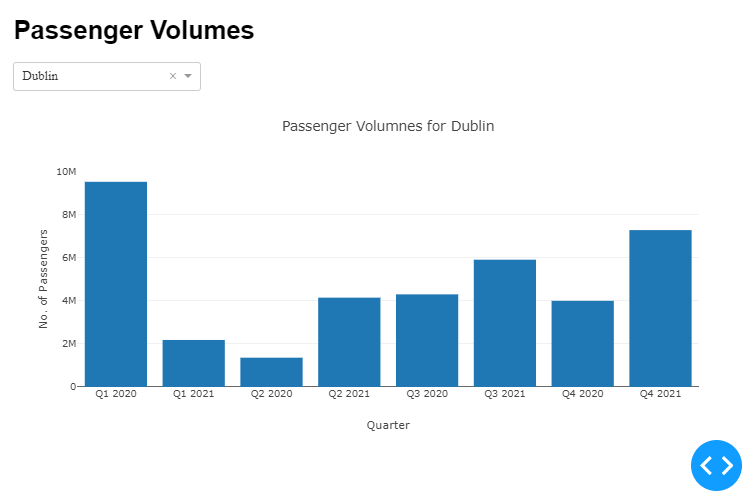
#### 3.12: Data Manipulation

This section is very straightforward, and its purpose is simply to organise, filter and tidy the data that has been downloaded as a CSV file from the UN website. I will not discuss this section in great detail simply as to not repeat myself.

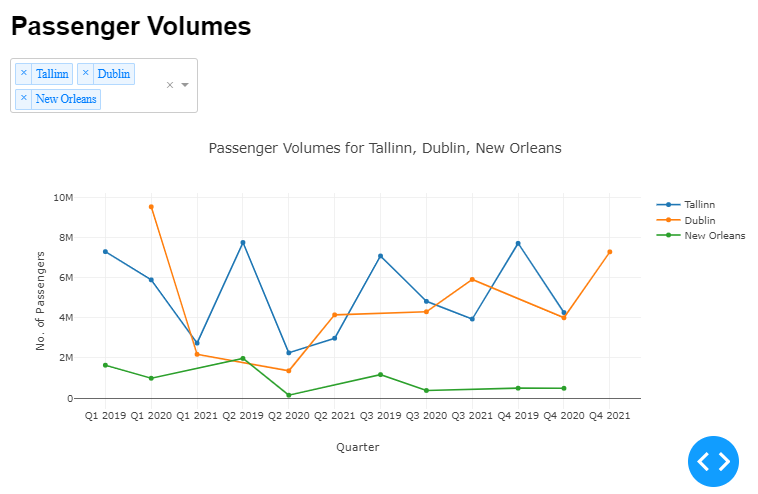
I will say that I had to make a key decision here, which was whether to only include the data that was labelled “Tram” or whether to group the data together by City. There are pros and cons for each, which is evident in my Jupyter notebook, but ultimately decided to only include the tram data. I have however, left in some code which would allow the user to easier change this decision and group the data together by city.

#### 3.13: Passenger Volumes Dashboard

I was very keen to use Dash in order to provide interaction visualisation to grasp this large dataset. As the data is collected from a large sample of differing sizes and scales, I felt it is only correct to visualise this using a dashboard which allows the user optionality when viewing the statistics. I firstly created a dashboard showing the passenger volumes for one particular city that was selected. The results were displayed using a bar chart.



After creating this dashboard, I felt that it did not serve the purpose I intended. Initially I wanted to create a dashboard to allow comparison between cities or regions, so I decided to recreate this dashboard displaying the results in a line chart and allowing the user to select multiples cities/regions. I have included an example below:



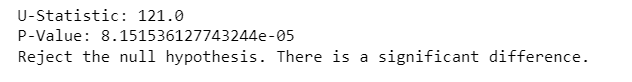
#### 3.14: Passenger Volumes Welch's T-test

As mentioned earlier, I also wanted to perform a few statistical tests on this dataset. I appreciate that the dataset is quite small when we compare individual cities against one another but as more data becomes available over the coming years this will increase the reliability of the results. However, for the time being this is the granularity and amount of data we currently have at our disposal.

The reasoning behind performing a Welsh’s T-test and not a T-test can be found in my Jupyter notebook.

#### 3.15: Passenger Kilometres Mann-Whitney test

I then created some code allowing the user to compare passenger kilometres between 2 desired cities, unfortunately these statistics are not available for Dublin, but I have included an example below, comparing Tallinn and Sofia:



## 4: Sentiment Analysis

### 4.1: Introduction

As I mentioned earlier, there has been a recent proposal to overhaul Irelands current rail network, this proposal is known as the ‘All-Ireland Strategic Rail Review’. This proposal has gained quite a lot of momentum over recent months and years.

I identified a post on LinkedIn which generated a significant amount of interest with over 400 comments under the post. LinkedIn understandable outline certain terms and conditions which users must adhere in order to protect other users, I have read these terms and have concluded that it is not possible to use these comments in my analysis. So therefore I turned my attention to Reddit which has a more lenient view of this. I was able to find a very similar post on Reddit surrounding the topic of the ‘All-Ireland Strategic Rail Review’.

### 4.2: Reddit API

Firstly, I had to gather comments. After researching I found that it was recommended to store using user-specific credentials in a JSON file. So I began by creating a JSON file with all the necessary components included.

I then utilised the PRAW library to initialise the Reddit API using the extracted user values and finally I return the Reddit object. I then included a separate snippet of code below this to ensure the Reddit API was initialised correctly. If successful it will print my username.

After outlining the post URL, I created another function which is designed to scrape (or gather) comments from a single post. The function itself is quite straightforward, printing the post’s title and all comments below. I also converted the comments list into and dataframe as I find this much easier to work with. I made the personal choice not to include any usernames as to remain prudent about user privacy.

### 4.3: Pre-processing the Data

Naturally, before performing sentiment analysis, I had to process my data. This was a very straightforward sequence of steps, which I have outlined below.

* Converting all characters to lower case
* Removal of special characters
* Tokenisation (breaking down text into individual units, known as tokens)
* Removal of ‘Stop Words’
* Lemmatisation (reducing words to their base form)

More details can be found in my Jupyter notebook.

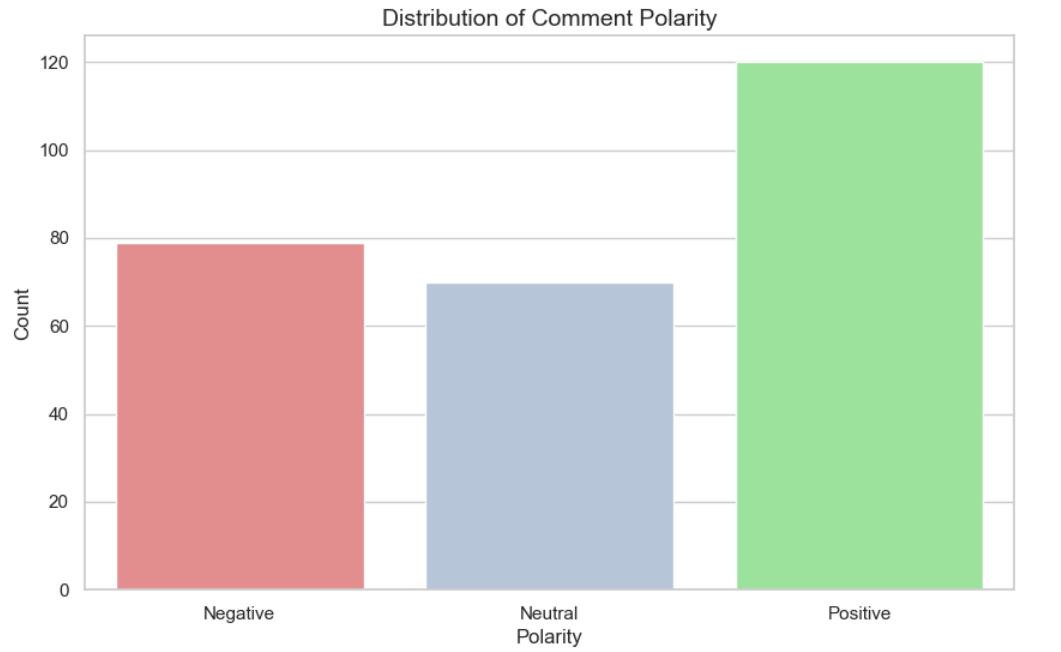
### 4.4: VADER Sentiment Analysis

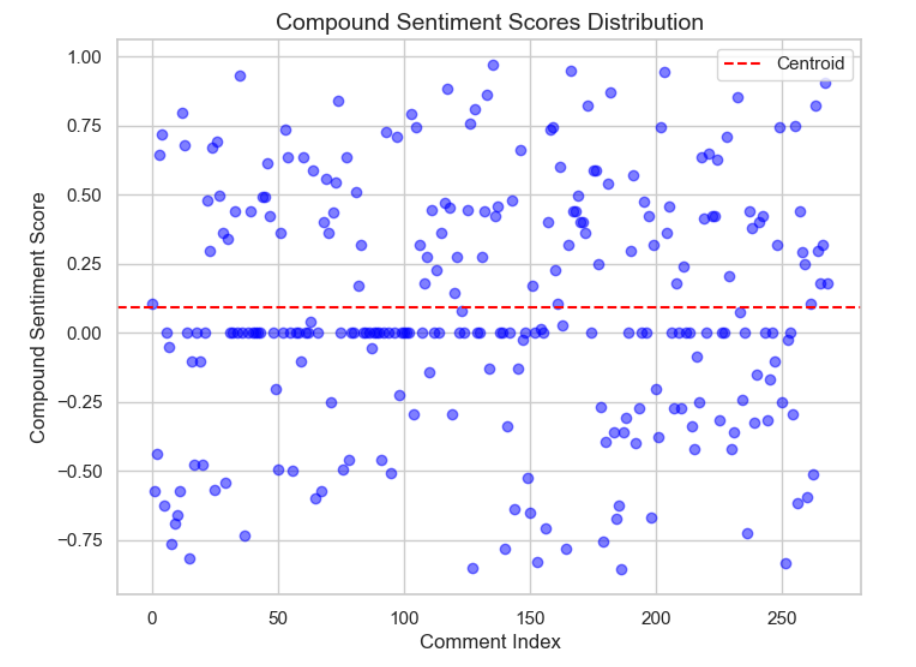
Firstly, I initialised SentimentIntensityAnalyzer class, which is a part of the VADER library and is used to analyse sentiment in text. I then iterate through each row (or comment) in the dataframe and calculating and storing a compound sentiment score for each in an additional column.

Our data now looks like this:

As this number can be quite difficult to interpret, I decided to add a polarity column which groups comment into specific categories which are Positive, Negative and Neutral.

I then visualised the results using a couple of simple visualisations, which follow Tuft’s principles.





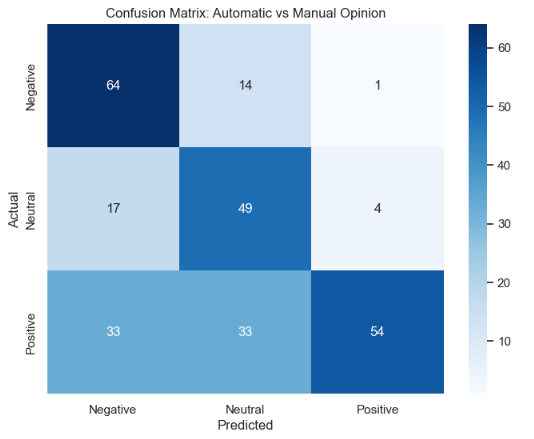
### 4.5: Self-Assessment of Model

As a disclaimer, this section is very subjective and the outcomes of this are not intended to further develop any of my research. I simply completed this section to satisfy my own curiosity and practice my analytical techniques.

After outputting the results of my sentiment analysis I was curious to see how my own judgement would match up to the model outcomes. I independently categorised each of the comments, disregarding the outcomes of the sentiment analysis.

This exercise was actually incredibly beneficial for me personally, it really showed how subjective sentiment analysis can be. There were numerous comments which were clearly sarcastic which wouldn’t have been picked up by our model. There were also a large number of comments which weren’t even discussing the article directly and instead going off on an obscure tangent.

The results of this exercise can be found below:



I hope this exercise isn’t taken as part of my project as a whole and can be seen as more of an exploratory standalone exercise which was very beneficial to me personally.

### 4.6: Word Cloud

My final section I wanted to visualise words that were commonly used surrounding this post. I found a library online called ‘wordcloud’ which was perfect for this visualisation. The results of this are shown below:



## Conclusion:

Overall, I was very happy with how this project went, despite all the setbacks regarding data available and privacy restrictions. There was a huge emphasis on independent research for this project which was time-consuming but also incredible rewarding.

I hope that my report has given the reader some interesting insights into Rail transport in Ireland whilst attempting to satisfy each exam requirement.

## Appendix 1: Examination Criteria

Programming for DA Tasks

1) Programming: The project must be explored programmatically: this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook.

The entirety of my Jupyter notebooks can show evidence of this.

2) Data structures: You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format.

I have gathered a processed data from both CSV, GeoJSON and JSON format.

3) Documentation: The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied.

Documentation is provided in my report and also within my notebooks.

Testing & Optimisation: You are required to document and evaluate a testing and optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). Note any trade-offs that you've made in these areas.

I have often discussed processing and computing time within my Jupyter notebook, see section 3.7 and modelling section 2.8 and 2.9.

Data manipulation: For each of the different data sources, compare and contrast at least two relevant libraries and techniques for a) processing and b) aggregating the respective data, in order to justify your chosen libraries/techniques.

I’ve often compared techniques such as interactive and non-interactive plotting,

Statistics for Data Analytics Tasks:

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

Number visualisations were used to help justify my choice of Regression model, ultimately leading me to favour the Random Forrest regression model.

2) Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with public transport, you could find a confidence interval for the population proportion of users commuting to Dublin by train).

Please see section 2.7.

3) Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests.

I have used over 5 statistical tests throughout my project.

4) Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process.

I was able to use the outcomes of the tests I performed in Section 3: UN data to research specific cities which were comparable to Dublin. Unfortunately data availability was a big issue here.

Machine Learning Tasks

1) Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

Please see sections 2.8-2.10 and Sections 4.

2) Collect and develop a dataset based on the transport topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate transport topic (e.g., public transport, freight movement etc…) for producers and consumers point of view in Ireland.

See section 4.

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

See sections 2.8-2.10.

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

See sections 2.8-2.9. Please see commentary in Notebooks also.

Data Preparation & Visualisation Tasks

1) Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data.

Discussed throughout the report, discussed issues when trying to get LinkedIn data.

2) Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.

Discussed and performed throughout report and notebook.

Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.

See section 4 for data preparation for sentiment analysis. Also present in other sections.

Modern Transport planning has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern Transport planning, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development. Note you may not use Powerbi, rapidminer, tableau or other such tools to accomplish this (at this stage).

Made use of Dash library and made number interactive maps. See section 1 and section 3.13.

## Appendix 2: Planning

Please see below for my planning/schedule regarding my project. It was very difficult to plan in advance. I wrote out a schedule and had my report/code sections all planned out but continuously hit barriers and roadblock which forced me to refocus my attention on another topic.

Please note this section is quite informal.

* 30/11/2023: Start project and begin researching
* 01/12/2023: Kick off writing code and complete initial visualisation of current rail networks in Ireland. Having issues with large GeoJSON files which make my notebook quite slow, but I really like the visualisations, so I think I am happy to make this trade-off.
* 03/12/2023: Begin report and plan out sections, main focus is on idea generation. What can I model? How will I generate a large enough dataset to perform modelling?
* 04/12/2023: Really wanted to perform sentiment analysis on a particular LinkedIn article but discovered that this was simply impossible given their strict privacy regulations. After discussing with my classmate, Stephen, he directed me towards considering Reddit as an option, so I have to credit him for this. I was soon able to find a similar article on Reddit.
* 07/12/2023: Worked on Sentiment analysis and emailed numerous bodies in an attempt to get a larger dataset, I contacted Irish Rail and also Luas.
* 09/12/2023: Still no contact from either. Decided to shift my focus to the Luas as there is a real lack of granular data regarding the Irish Rail Network. Began conglomerating numerous datasets surrounding the Luas.
* 12/12/2023: I accepted that I couldn’t rely on being sent a large dataset from either Luas or Irish Rail, so I decided to create my own data using the data available to me.
* 13/12/2023: Began researching similar tramlines throughout the world, after some research I selected Edinburgh, given its similar size and scale. Both considering the tramline itself and also the cities.
* 15/12/2023: Continue to work on Sections 2 and 4. Section 1 is now complete. Report needs updated.
* 17/12/2023: Developed Section 3, began diving into Melbourne datasets.
* 18/12/2023: Worked on Report, updating sections 2 and 3.
* 19/12/2023: Plan and section finalised, idea generation and research in mostly complete. Now I will just write my code and complete my report.
* 19/12/2023 – 26/12/2023: Continue to write and tidy code. Documenting my code in the report.
* 27/12/2023: Finish off any loose ends, ensuring code is correctly annotated and removing of any tedious sections of ‘workings’ code.
* 02/01/2023: Complete project and make sure all points in the exam criteria are hit. Also need to ensure GitHub is up-to-date and accurate.
* 03/01/2023: Tidy up my report, ensure sections and datasets are correctly labelled. Work on appendices.

## Bibliography

A, M.R.L. (2021). A Beginners Guide to Create a Cloropleth Map in Python using GeoPandas and Matplotlib. [online] Medium. Available at: https://towardsdatascience.com/a-beginners-guide-to-create-a-cloropleth-map-in-python-using-geopandas-and-matplotlib-9cc4175ab630 [Accessed 5 Jan. 2024].

Arora, S. (2022). Sentiment Analysis Using Python. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/.

Brownlee, J. (2018). How to Make Predictions with scikit-learn. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/make-predictions-scikit-learn/.

Campbell, K. (2023). ‘Litany of avoidable failures’ in Edinburgh tram project. BBC News. [online] 19 Sep. Available at: https://www.bbc.co.uk/news/uk-scotland-edinburgh-east-fife-66854342.

Carrell, S. and editor, S.C.S. (2023). True cost of Edinburgh tram line has exceeded £1bn, says report. The Guardian. [online] 19 Sep. Available at: https://www.theguardian.com/uk-news/2023/sep/19/edinburgh-tram-line-cost-public-inquiry-report.

GeeksforGeeks. (2020a). Data Manipulation in Python using Pandas. [online] Available at: https://www.geeksforgeeks.org/data-manipulattion-in-python-using-pandas/.

GeeksforGeeks. (2020b). Introduction to Dash in Python. [online] Available at: https://www.geeksforgeeks.org/introduction-to-dash-in-python/.

GeeksforGeeks. (2021). Choropleth Maps using Plotly in Python. [online] Available at: https://www.geeksforgeeks.org/choropleth-maps-using-plotly-in-python/.

Juanola, M.P. (2022). French company to build 100 new Melbourne trams under $1.85 billion deal. [online] The Age. Available at: https://www.theage.com.au/national/victoria/french-company-to-build-100-new-melbourne-trams-under-1-85-billion-deal-20220421-p5af0t.html.

Mogyorosi, M. (n.d.). Sentiment Analysis: First Steps With Python’s NLTK Library – Real Python. [online] realpython.com. Available at: https://realpython.com/python-nltk-sentiment-analysis/.

O’Broin, C. (2023). NTA publishes plans for Luas, DART and Metrolink extensions. [online] Irish Mirror. Available at: https://www.irishmirror.ie/news/irish-news/nta-publishes-plans-dublins-transport-29060541.

Pascual, F. (2022). Getting Started with Sentiment Analysis using Python. [online] huggingface.co. Available at: https://huggingface.co/blog/sentiment-analysis-python.

Pieters, M. (2020). Martin Peters. [online] martinbpeters.github.io. Available at: https://www.martinpeters.ie/2020/01/25/ireland-geospatial/ [Accessed 5 Jan. 2024].

RG (2019). One-stop Guide to Data Manipulation in Python. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/python-data-manipulation-fb86d0cdd028.

Selvaraj, N. (2020). A Beginner’s Guide to Sentiment Analysis with Python. [online] Medium. Available at: https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6.

unece.org. (n.d.). UNECE releases new dataset for tram and metro statistics, ‎supporting informed policy responses for sustainable post-COVID ‎mobility | UNECE. [online] Available at: https://unece.org/transport/press/unece-releases-new-dataset-tram-and-metro-statistics-supporting-informed-policy [Accessed 5 Jan. 2024].

Zhang, Y. (2020). How to Plan and Organize a Data Science/Analytics Project. [online] Medium. Available at: https://towardsdatascience.com/how-to-plan-and-organize-a-data-science-analytics-project-a9418c12c808.