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.

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:

[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 researched, I soon came across a number of obstacles which forced me to rethink the subject matter of my project. I will discuss details of this over the course of my report. You may also find details of this in Appendix 2.

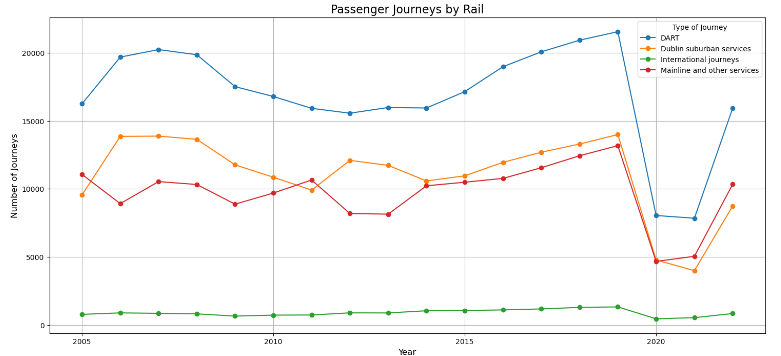
The Irish Rail network has been in a state of decline over the last 100 years. Mainly due to the increased popularity of personal cars. However, the automotive industry is currently in a state of transition, with increasing greenhouse gas regulation and a population who are now more concerned about the 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, sources of which are outlined within my notebook.

The first dataset I will be taking a look at contain information on the number of passengers taking certain rail journeys, split into 4 categories (DART, Mainline, Dublin Suburban and International Journeys).

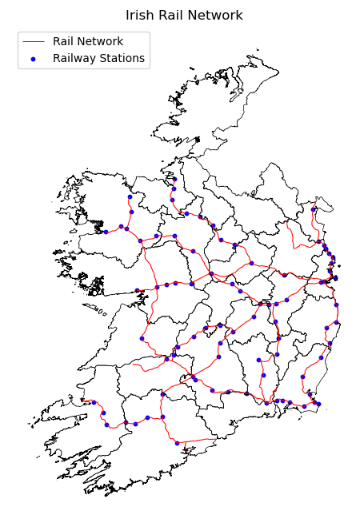
Whilst this data is not very granular it is the only dataset, I have access to the moment. I requested a more detailed dataset from Irish Rail but unfortunately, I did not receive a response.

After reading in the data, I trimmed the data to only include 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 excluded any years prior to 2005.

I then visualised the data using a line chart:

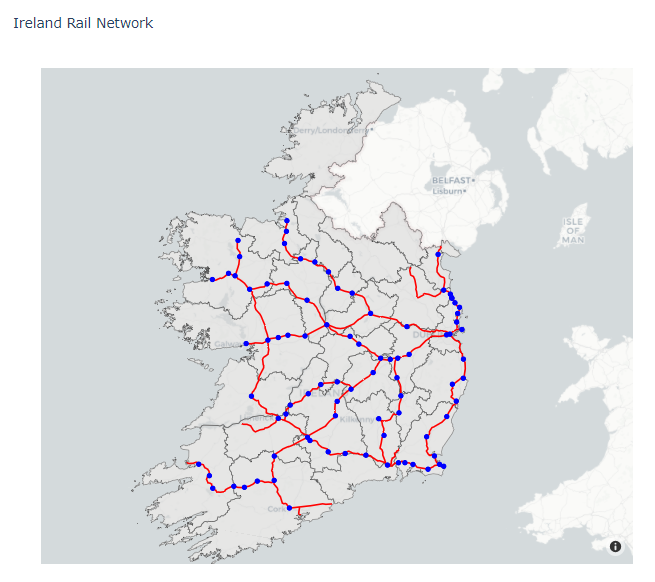
We can really see the dramatic impact that COVID-19 had on the number of passenger journeys made by train, this will be a recurring trend we will notice going forward. We are seeing passenger numbers increase recently but it is still apparent that that have not yet reached pre-covid levels.

It is only appropriate to geographically visualise the current rail network in the Ireland. I began by sourcing a number of GeoJSON. Again, the 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.

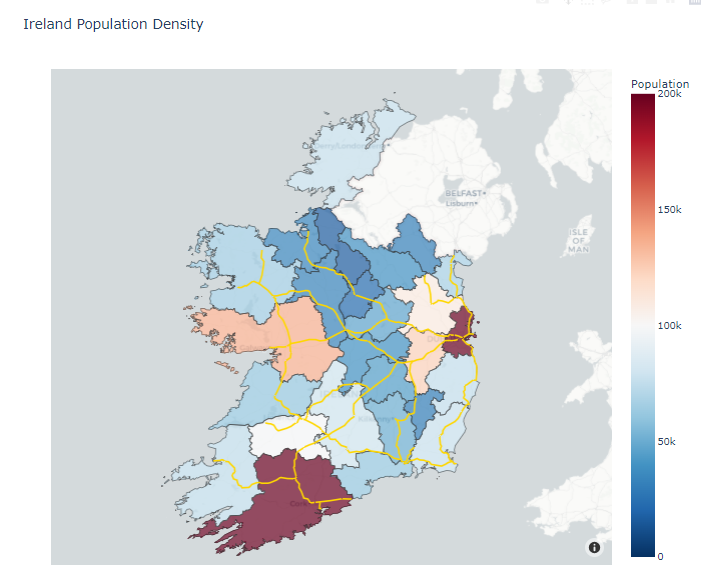
Whilst this is a good visualisation of Ireland’s railway system, I wanted to take it a step further and create an interactive map, which would allow the user to take a more personal look at the current rail network.

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

I was keen to take this interactive map a step further in order to try and visualise population density also. 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 contains the working population by county. I joined this to my existing dataframe to create a choropleth map based on the 26 county regions.

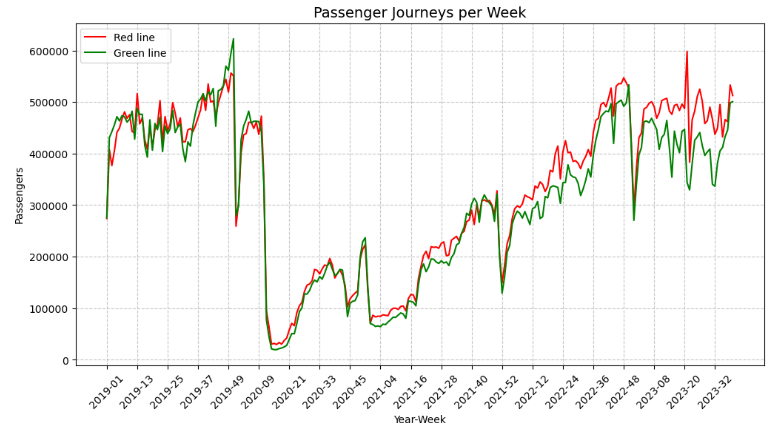
After some successful data manipulation, I merged the two dataframes in preparation for my interactive choropleth map.



I think the most notable thing to note 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 also be justified given the huge population discrepancy that is present.

After struggling to find sufficiently granular data to perform an adequate investigation of the 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 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 Dublin city area.

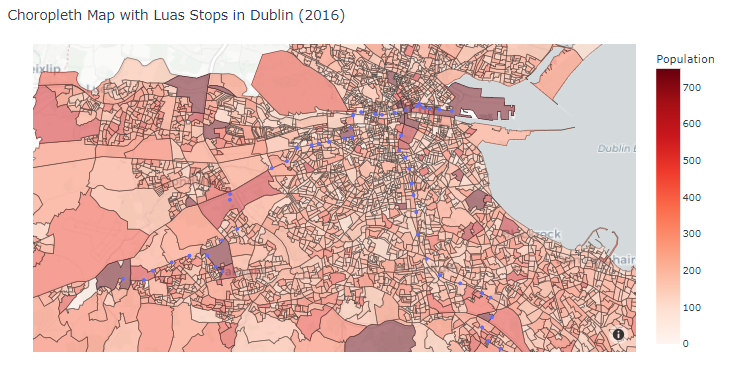
In order to understand the current situation regarding the Luas, I wanted to take a closer look at the recent trends regarding passenger numbers and I was soon able to find a dataset outlining this exact topic. After analysing and performing some basic manipulation on the dataset, I was able to plot an 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 one another. We can again see the impact that COVID-19 had on the passenger numbers.

The next aspect of the Luas I was keen to investigate was how accessible the lines are 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. These areas are quite subjective, but the code can be easily manipulated to include a wider range of areas.

I then merged the population data from the small area census to this dataframe. As evident in Notebook 1, I had a lot of practice in geographical plotting, so I was able to rely heavily on the research I had performed earlier to produce the below visualisation.

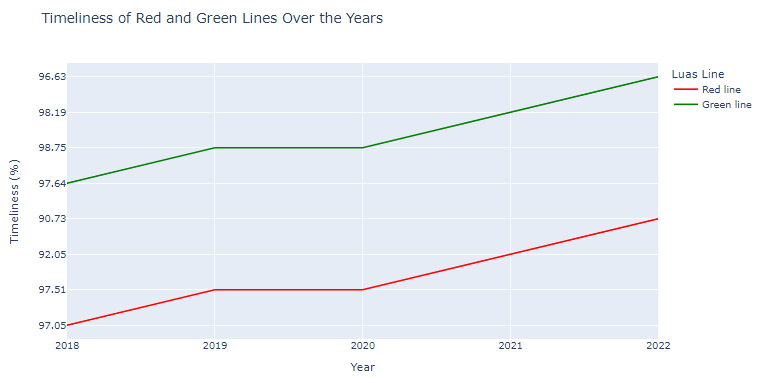


This gives us insight into the current accessibility of the Luas and potential for future development through extensions or indeed new lines. After analysing this map, I arrived on a number of potential developments regarding the current Luas system. More details on this can be found in my Jupyter notebook.

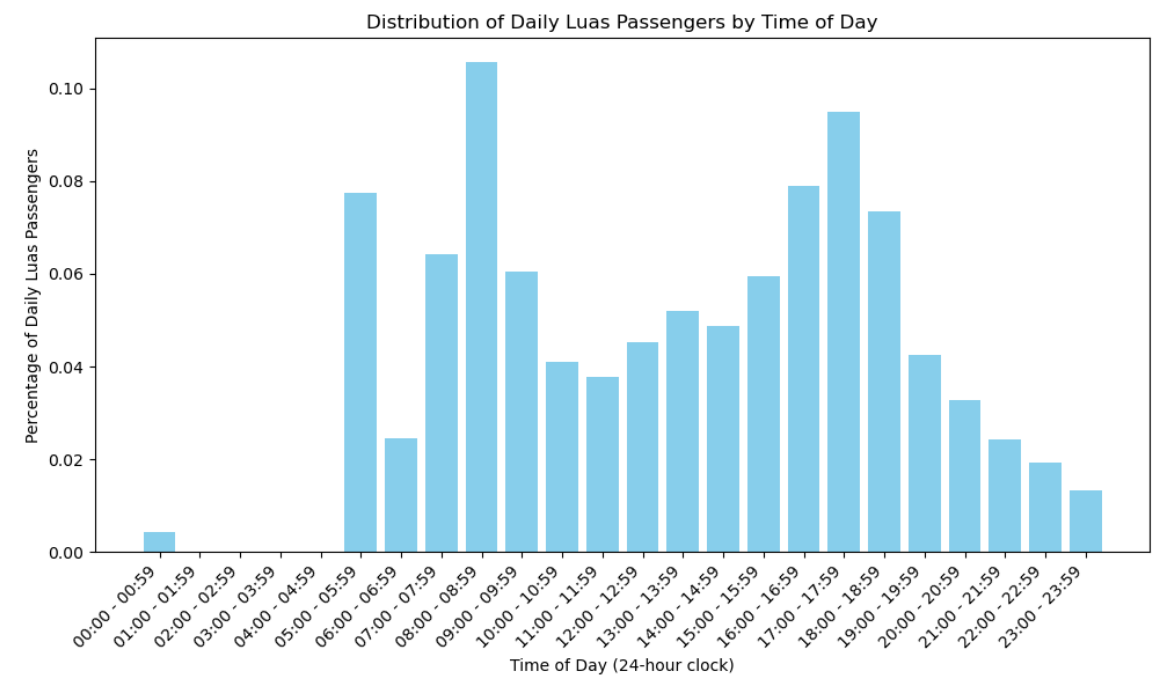
1. Extensions to Blanchardstown, Leixlip – It is clear that a reasonably small offshoot of the red line could satisfy a large population centre to the east, primarily Blanchardstown.
2. Dublin Airport – This would be a huge project but after researching the Edinburgh Light Rail system (Section 3.1 – 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 these areas.
4. North Dublin – The complete lack of Luas lines in north Dublin is quite strange, this is just a general observation but could be grounds for an entirely new line.
5. Stadiums, Universities – The Luas system could take a leave out of this Edinburgh’s 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 in place. 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)

I then plotted a quick graph to assess the Luas’s reliability.

With this new dataset I was also able to plot the popularity of each Luas line per hour of the day. This lead to the plotting of the below graph:



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

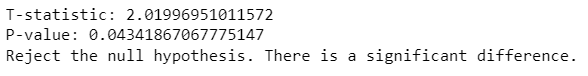
I was keen on creating an appropriate dataset that could be used to perform predictive modelling. I noticed that I had one dataset which outlines the number of passenger journeys per day and another dataset which outlines the percentage of daily passengers per hour of the day. By assuming that the percentage 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 more detailed data from Luas themselves, but I was once again met with no response.

Before proceeding with my predictive modelling, I wanted to perform a few statistical tests on my data. Firstly, I wanted to gain insight into my data by calculating confidence intervals surrounding passengers per hour-of-the-day.

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:

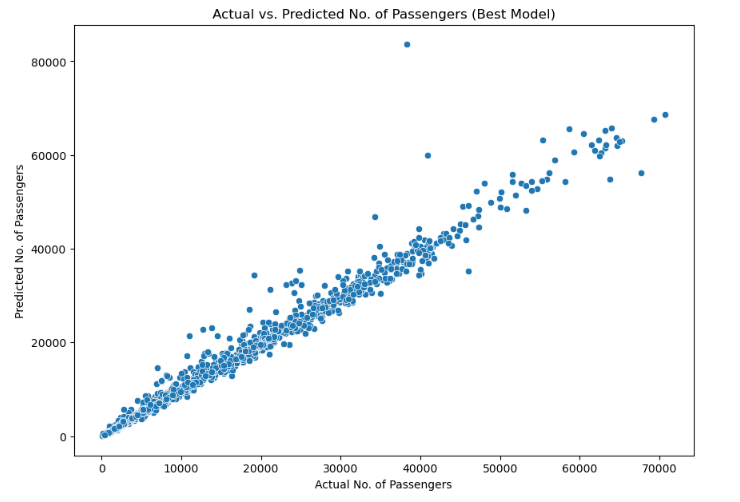


I also wanted to test whether the average number of passengers is significantly different between the red and green lines. I began by separating out my data and performing the t-test using the ‘ttest\_ind’ function. The results of this test can be found below:

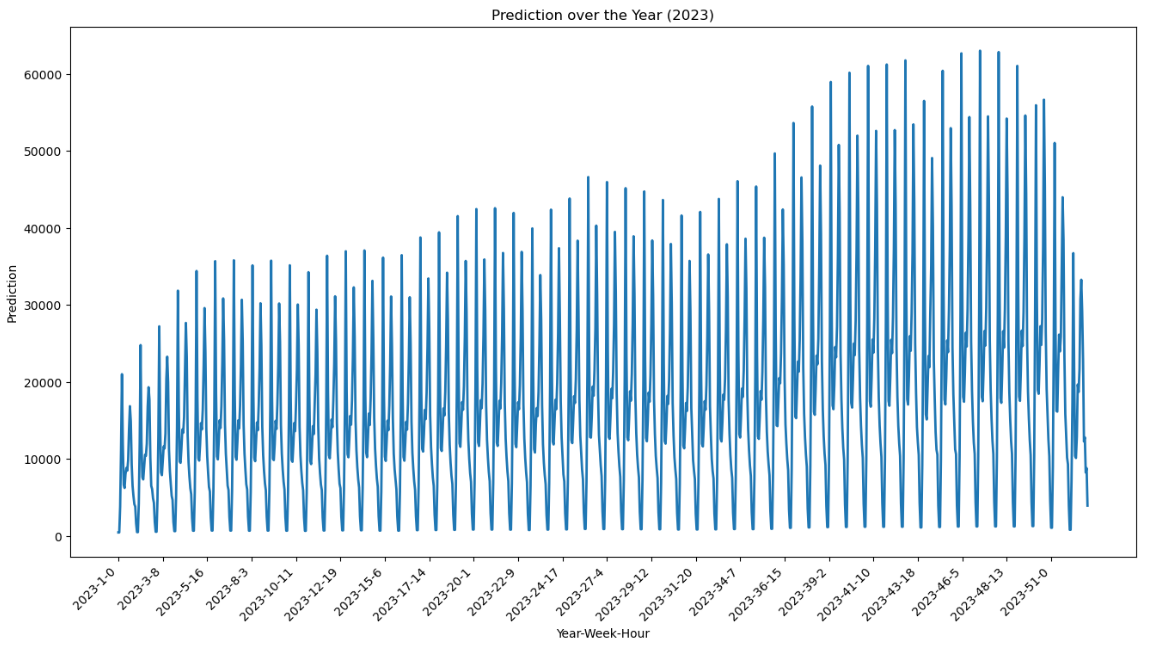


As intended, I wanted to create a predictive model to predict the number of passengers on any given hour of the day. I firstly encoded my data and then trained my model appropriately using a random forest regression model. The initial model I created performed well with a good mean-squared score.

I also plotted a scatter plot based on the predicted and actual values which looked 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:



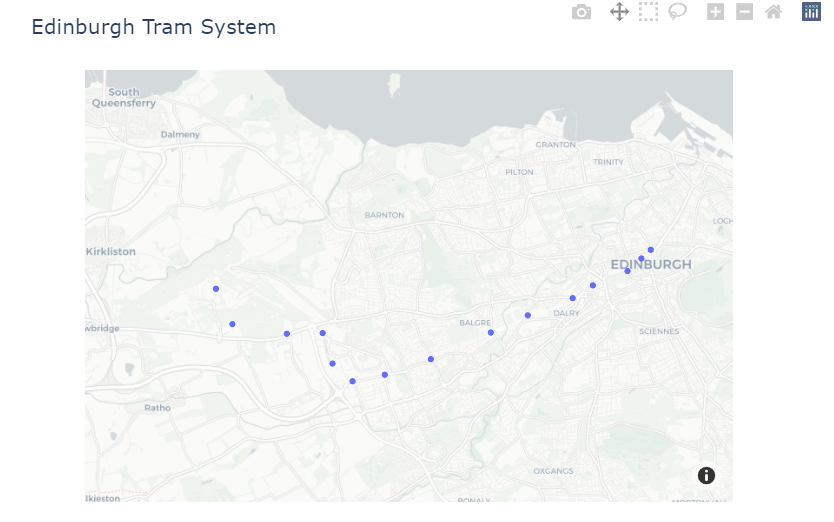
I experimented with numerous alternate models such as a linear regression model, lasso regression and SVM model with predictably poor results.

I then created a 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**

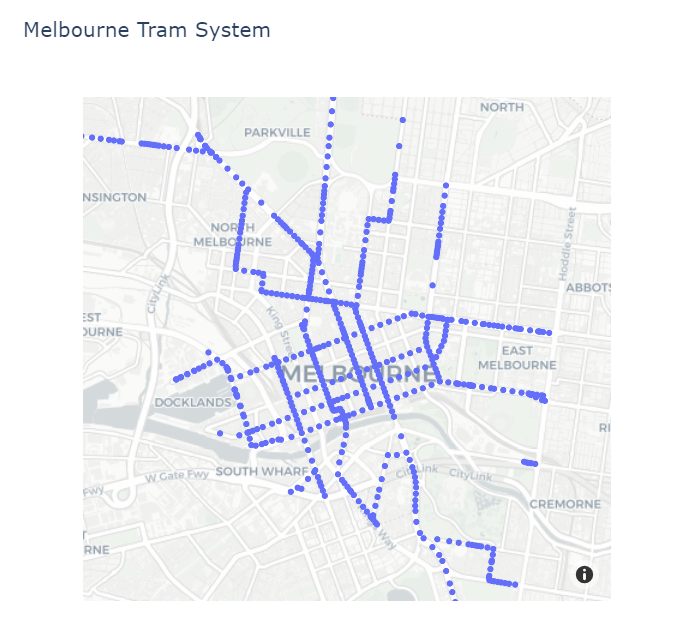
In order to choose an appropriate tram system to compare with the Luas, I had to consider a number of factors, including population, density, current rail infrastructure. I spend a significant amount of time researching in order to make an appropriate selection and I concluded that the most suitable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

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

I then sourced a number of key statistics related to the tramline, with the aim of performing a comparison to the Luas. Unfortunately, I was 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.

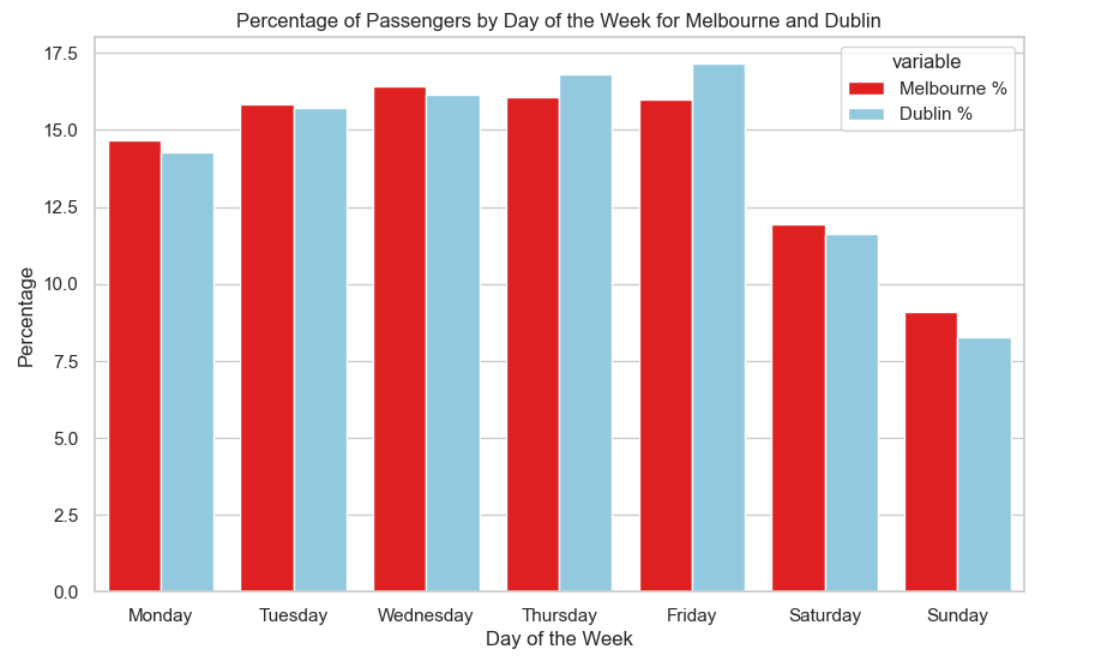
As I was forced to turn my attention elsewhere, I began to research once again. This time I prioritised data availability before I committing to comparison work. After my research I found that Melbourne tramlines would be a great system for comparison and the local government place a huge emphasis on open data. This would make it a lot easier to collect data.

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

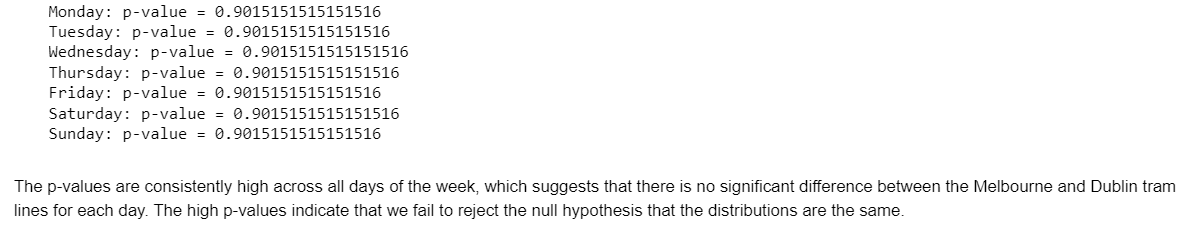


I made the decision to plot this map using centroid positions of each tramline as it significantly reduces the processing power required to generate such a map, whilst still giving us a good gauge of the size of the network.

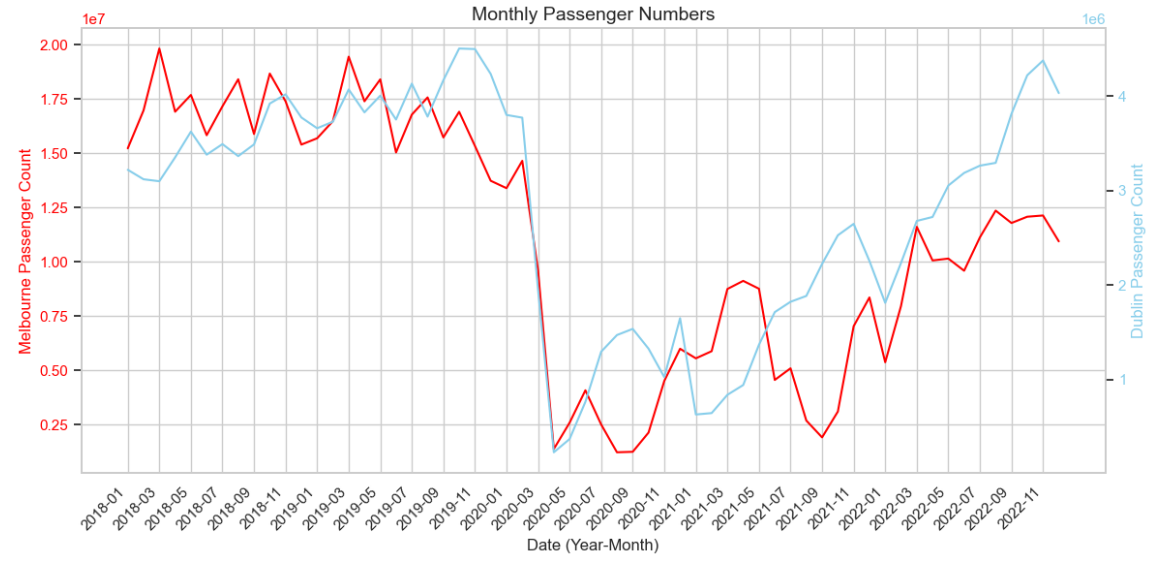
I then gathered data on the number of passengers per day of the week in order to compare this to the Luas. I collected, manipulated, and aggregated my data until it was in an ideal format for visualisation. I decided that a stacked bar chart would be most appropriate for this scenario.

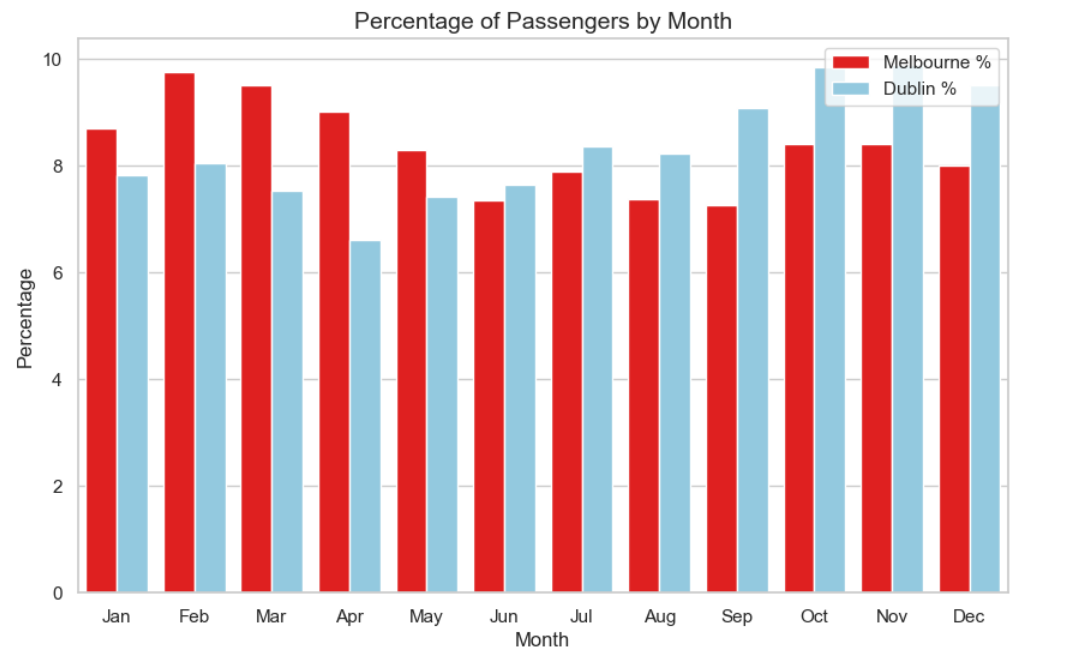


From the chart above we can see that weekly patterns are very similar between the two tram systems. This led me to complete a Mann-Whitney test on our data. The results of which can be found below.



I also wanted to take a closer look at the trends in passenger numbers over recent years



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.

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

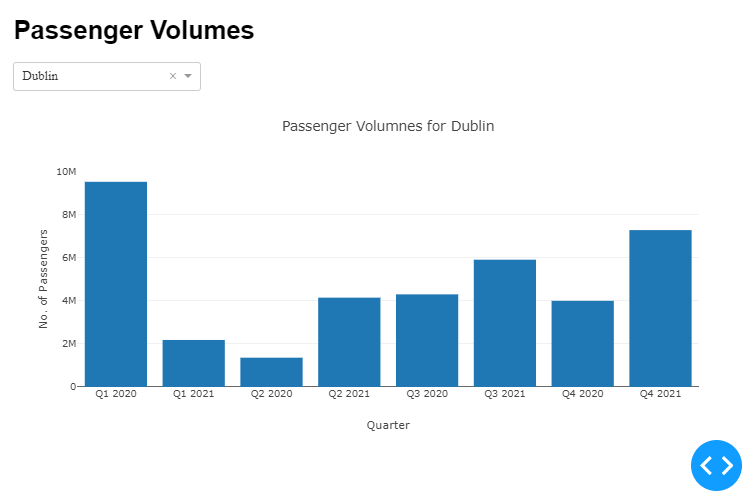


This section is very straightforward, and its purpose is simply to organise, filter and tidy the data that has been downloaded from the UN website.

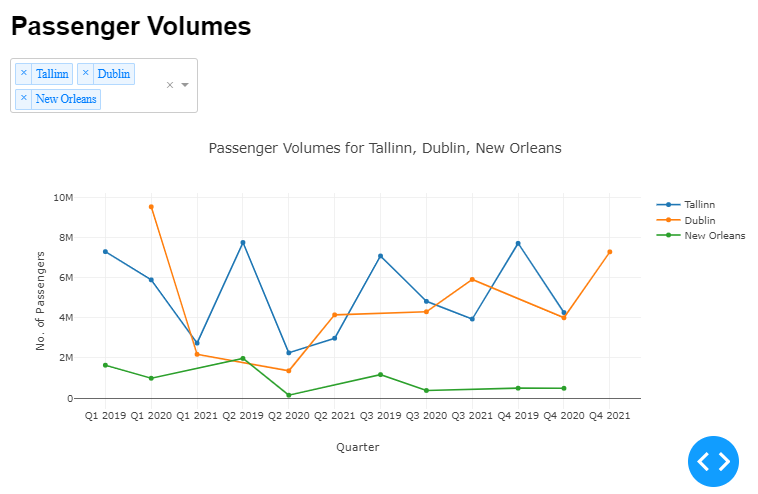
I had to make a key decision here, which was whether to only include the data that was labelled “Tram” or whether to group the data together by City. There are pros and cons for each, which is evident in my Jupyter notebook, but ultimately decided to only include the tram data. I have however, left in 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 make an interaction dashboard to grasp this large dataset. It is only correct to visualise this in a way which allows the user optionality when viewing the statistics. I created a dashboard showing the passenger volumes for one particular city that was selected. The results were displayed using a bar chart.



However, I felt that this did not serve any great purpose. 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.



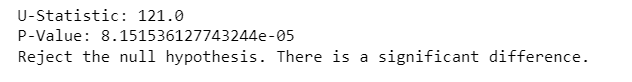
I wanted to perform a few statistical tests on this dataset. I appreciate that the dataset is quite small when we compare individual cities but as more data becomes available over the coming years this will increase the reliability of the results. However, for the time being this is the granularity and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

Please see my notebook for a greater discussion on the Welsh’s T-test. Please see below for an example output:



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

I then wrote code to allow the user to complete a Mann-Whitney test comparing passenger kilometres between cities, I have included an example below which comparing Tallinn and Sofia:



## 4: Sentiment Analysis

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. LinkedIn 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 was not possible to use these comments in my analysis. So therefore I turned my attention to Reddit who have more lenient privacy regulations. I was able to find a very similar post on Reddit surrounding the topic of the ‘All-Ireland Strategic Rail Review’.

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 credentials. I then included a separate piece of code to ensure the Reddit API was initialised correctly. If successful it will print my username.

I then 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.

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)

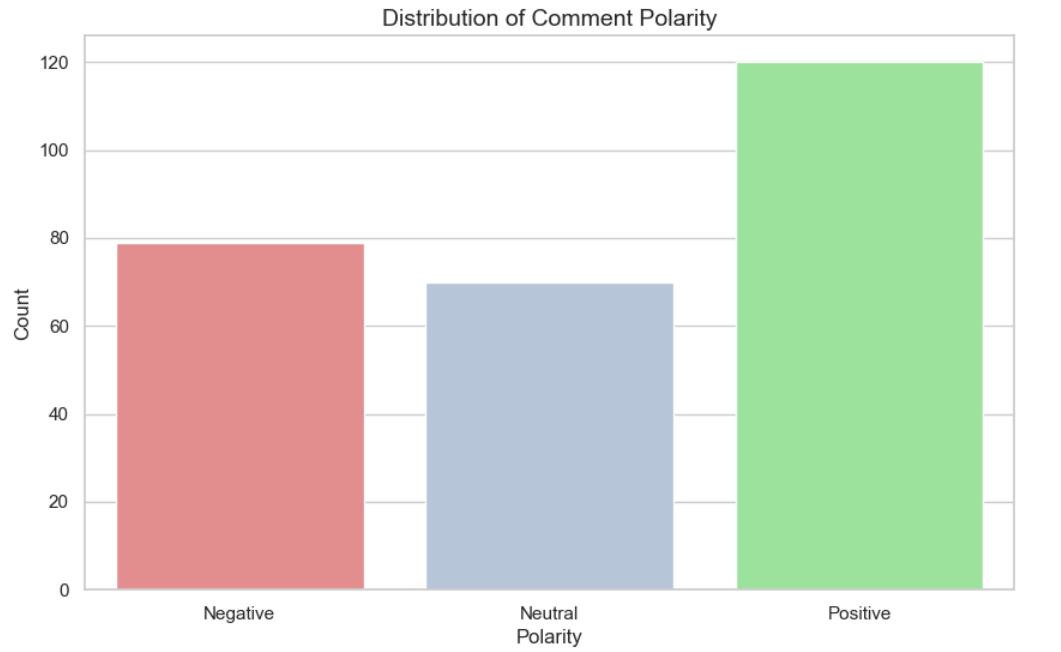
More details can be found in my Jupyter notebook.

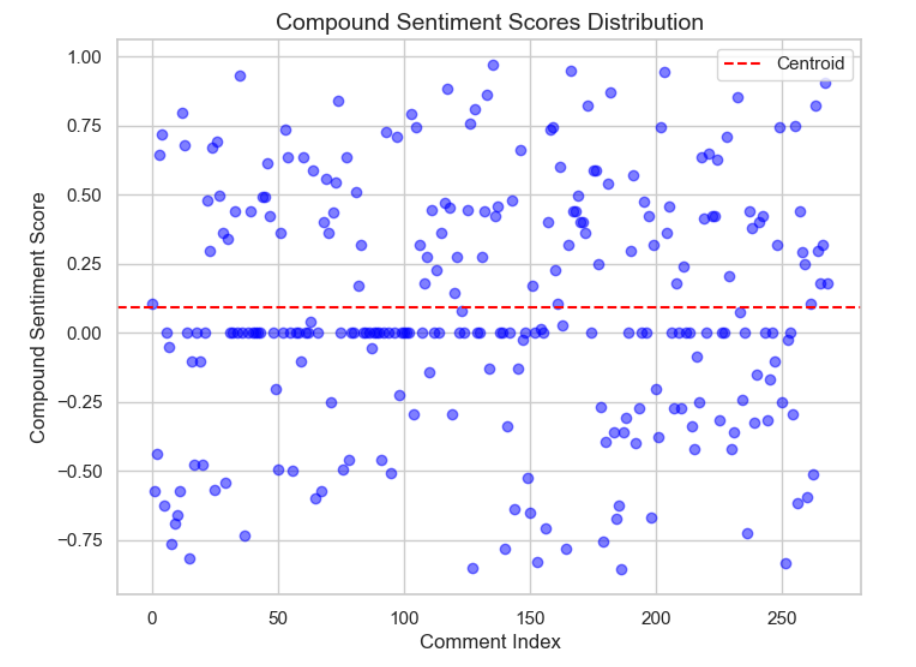
Firstly, I initialised SentimentIntensityAnalyzer class, which is a part of the VADER library and is used to analyse sentiment in text. I then iterate through each row (or comment) in the dataframe and calculating and storing a compound sentiment score for each in an additional column.

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.



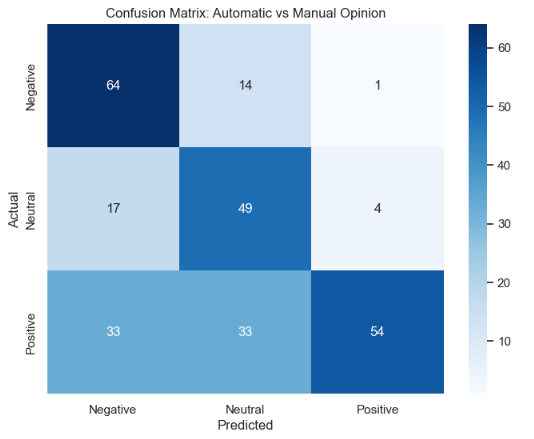


As a disclaimer, this section is very subjective and the outcomes of this are not intended to further develop any of my research. I simply completed this section to satisfy my own curiosity and practice my analytical techniques.

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.

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:



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.