A comparative study of public transport use in Dublin and Canberra

Nayane Oliveira Araújo

*GitHub:* <https://github.com/NOA-Data1/2021264-MSC-DA-CA-2>



MSc in Data Analytics

Programming for Data Analytics, Machine Learning, Statistics for Data Analytics, and Data Preparation and visualization

January 2024

Dublin - Ireland

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for Data Analytics*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualization* |
| **Assessment Title:** | *A comparative study of public transport use in Dublin and Canberra* |
| **Lecturer Name:** | *Sam Weiss*  *David Gonzalez*  *Taufique Ahmed*  *David McQuaid* |
| **Student Full Name:** | Nayane Oliveira Araújo |
| **Student Number:** | 2021264 |
| **Assessment Due Date:** | 07/01/2024 |
| **Date of Submission:** | 07/01/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**CONTENTS**

[Abstract 4](#_Toc155553919)

[Introduction 5](#_Toc155553920)

[ **Project Description** 5](#_Toc155553921)

[1 - Data Preparation and Visualization 6](#_Toc155553922)

[**1.1 Description of the raw data acquisition process:** 6](#_Toc155553923)

[**1.2 Exploratory Data Analysis (EDA)** 7](#_Toc155553924)

[ **Visualization Dublin (Ireland)** 8](#_Toc155553925)

[ **Visualization Canberra (Australia)** 9](#_Toc155553926)

[2 - Statistics for Data Analytics 11](#_Toc155553927)

[**2.1 Distribution and correlation of all modes of transport in Dublin** 11](#_Toc155553928)

[ **Descriptive statistics** 11](#_Toc155553929)

[**Empirical Cumulative Distribution function** 17](#_Toc155553930)

[ **Parametric Test (t-test)** 18](#_Toc155553931)

[ **Nonparametric Test (Wilcoxon)** 18](#_Toc155553932)

[ **Chi-Square Test** 19](#_Toc155553933)

[ **Hypothesis testing (comparing of means using t-test)** 19](#_Toc155553934)

[3 - Machine Learning 19](#_Toc155553935)

[**SGBRegressor** 20](#_Toc155553936)

[**Sentiment Analysis** 21](#_Toc155553937)

[**Sentiment analysis with VADER** 23](#_Toc155553938)

[4 - Programming 25](#_Toc155553939)

[**4.1 – Programming** 25](#_Toc155553940)

[**4.2 - Data Structures** 25](#_Toc155553941)

[**4.3 Documentation** 25](#_Toc155553942)

[**4.5 Data manipulation** 26](#_Toc155553943)

[Conclusion 26](#_Toc155553944)

[References 27](#_Toc155553945)

# Abstract

This report provides a comprehensive analysis of public transport in Dublin, Ireland's capital, and Canberra, Australia's capital. Several statistical tests, including parametric and non-parametric inferential tests like the t-test, analysis of variance, Wilcoxon test, and chi-squared test, were performed as part of the analysis, producing different insights for data visualization. To apply the sets to the time series of Machine Learning, all data preparation steps were followed. Using a third dataset in JSON format from an investigation collected in Dublin on the reasons and opinions for not using public transport, as well as the Reddit from web APIs, sentiment analysis and clustering are performed. The structure used in the project was the CRISP-DM method. The data was researched and obtained from government sources in each country, with their licenses duly verified, and this project is available in the GitHub repository through the link previously provided.

***Word count****: 3808 words (excluding code, code comments, titles, references, and citations)*

# Introduction

The public transport system in Ireland transport supplies citizens as well as tourists with a wide range of options. It is essential to the connection of neighborhoods, cities, and suburban areas. In particular, Dublin's infrastructure is diversified and substantial, including the rent of bicycle systems, buses, trams (Luas), suburban trains (Dart), and other transport options constituting an integrated urban mobility system.

The Dublin Bus company's bus service covers an extensive area, traveling throughout Dublin and its surrounding districts. Buses are common for most urban transport due to their variable and typical routes. The Luas tram system connects the heart of the city to the suburbs and several significant places via well-planned routes. The Dart train service travels along the coast, connecting the city to suburban areas. Furthermore, suburban trains facilitate travel outside of Dublin's closest suburbs by linking it to other urban and rural regions, in addition, the Dublinbikes bicycle rental program promotes a healthy and environmentally friendly choice for short city trips.

Using the Leap Card travel card, the integration of several transportation modes makes it easier to take advantage of services and costs less. In summary, Ireland’s public transport system offers various options to meet travel needs, standing out for its comprehensiveness and integration, facilitating mobility in the capital and beyond.

This project will analyze the flow of people who use specific kinds of transport in Dublin, the capital of Ireland, and compare it with Canberra, the capital of Australia, utilizing data that could be collected through the automation of the electronic ticket system.

The capitals chosen for the study developed on the use of public transport are Dublin in Ireland, which in the last census carried out by the government statistics centre counted a total of 1,458,154 million inhabitants, as well as Canberra in Australia, which in the last census released by the government carried out in 2021, had approximately 453,890 inhabitants.

## **Project Description**

**1 - Business Understanding**

The objective of the business is to analyze and understand trends in the use of public transport in the capital of Ireland, Dublin, and the capital of Australia, Canberra.

**2 - Understanding the data**

The analysis will focus on three sets of data: two sets of data with a collection of the flow of public transport use in each capital, making it possible to carry out statistical calculations, visualization of manipulated data, and application of machine learning as predictions of travel volume. A dataset in JSON format, and a collection of Reddit reviews, for sentiment analysis.

**3 - Data Preparation**

This stage includes all data preparation and exploration to produce data visualization, Machine Learning, and Statistics results.

**4 – Modelling**

To predict the number of passengers per week, a regression time series analysis was performed using three models: Lars, Decision Tree, and a new model called XGBRegressor, also using GridSearchCV to select the best hyperparameters, and sentiment analysis.

**5 - Assessment**

The model's effectiveness was evaluated by comparing the R² metrics obtained from each applied regression model and statistical tests.

**6 - Implementation**

After comparing the performance of the regression models, the best **forecasting model** can be determined.

# 1 - Data Preparation and Visualization

During the data collection stage, some points made it difficult to choose certain data sets, such as a good time series, to carry out Machine learning tasks, and the difference between the types of data collected for each country, which would make comparison difficult. The datasets, and even the collection periods with a large difference in years between countries, in addition to the licenses assigned in each set of data found, and what was allowed to be done in each of them. The Australian capital Canberra, firstly was due to compatibility in the years of collection, with the chosen Irish dataset, and in research on the place, despite having a relatively low population compared to Dublin, there is a much larger territorial extension, which perhaps it could influence in a more frequent use of public transport. Data preparation consisted of data collection, data cleaning, and data manipulation, this step was critical for adjusting the metrics of two different datasets to be compared.

## **1.1 Description of the raw data acquisition process:**

* Dublin (Ireland) dataset: “Passenger Journeys by Public Transport.csv”.

This dataset consists of 1040 entries and 5 columns, containing a weekly collection of passenger flow by public transport category in the years 2019 until 2023.

License: The dataset was collected from data.gov.ie and downloaded as a CSV file, under the “Creative Commons Attribution 4.0 (CC BY 4.0)” license.

* Canberra (Australia) dataset: "Daily Public Transport Passenger Journeys by Service Type 20231226.csv".

The dataset is available on the Open Data portal ACT government, has 1639 entries and 7 columns, and was collected daily from 2019 until December 2023.

License: The dataset was downloaded as a CSV file, under the “Creative Commons Attribution 4.0 (CC BY 4.0)” license.

The third dataset is “NTA56.20240103164753.csv” - **Reasons for not using bus services more frequently.** The dataset was acquired from the website data.gov.ie and has 336 entries and 12 columns, containing the population's opinions on the use of public transport.

License: The dataset was downloaded as a CSV file, under the “Creative Commons Attribution 4.0 (CC BY 4.0)” license.

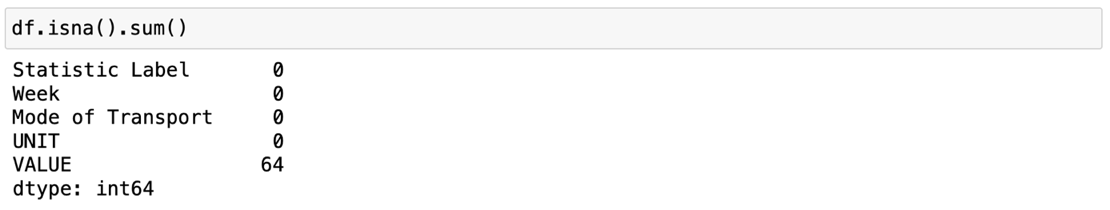
## **1.2 Exploratory Data Analysis (EDA)**

In data exploration, data cleaning was carried out, duplicate lines were considered, and the decision on null values ​​found was to be filled with zero, this will be detailed later, the transformation and manipulation value units were standardized for the dataset and creation of new columns, the sampling Statistics technique with 5% and 20% was applied for statistical inferential tests and normalization of values ​​to perform time series in Machine Learning.

* **Analysis**

In the first steps using the isna function plus the sum, it was discovered that the dataset presents 64 missing values ​​in total. Using the fillna function, all empty cells in this Dataframe were filled with the value zero. The decision between filling the missing values ​​with zeros or eliminating the lines was decided in the context and the impact these missing values ​​could have on the model being built.

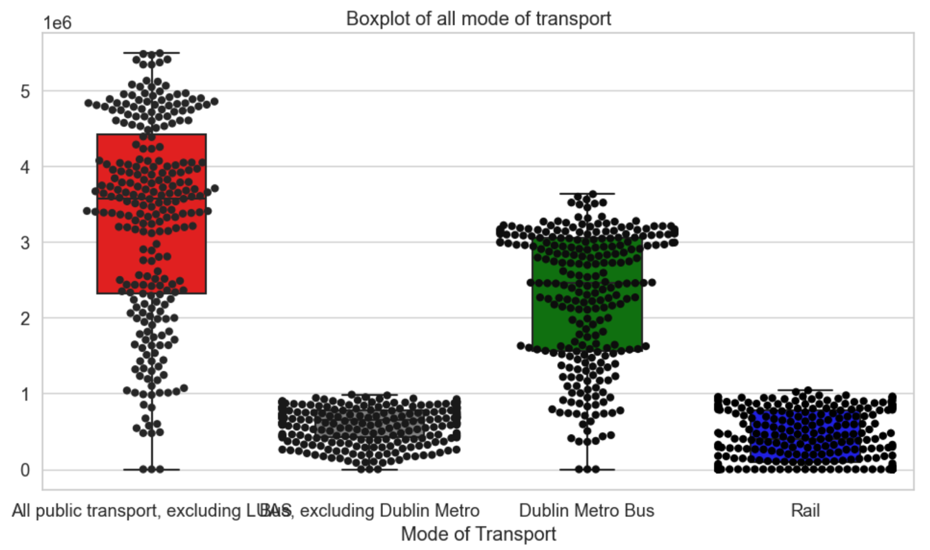
Filling in with zero was a decision taken, considering the advantage of data preservation, keeping all observations so as not to lose valuable information. The disadvantages are that descriptive statistics distortions and measures of central tendency can occur, and in some cases, filling in with zero can introduce a bias in modelling, in the general context, by preserving data, instead of removing missing values, the decision was to fill in zero value.



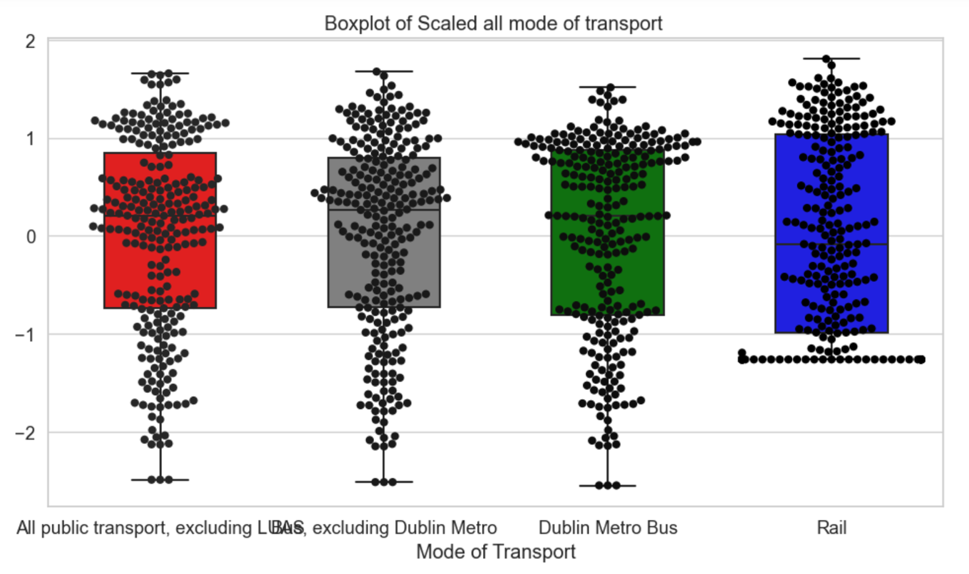
### **Visualization Dublin (Ireland)**

The first principle is stated as “Show comparisons, contrasts, differences” (Tufte, 2001:56).

Tufte presents six visualization principles; as part of the generation of data in graphical form, models, colors, and an interface were selected to make the data presented understandable.

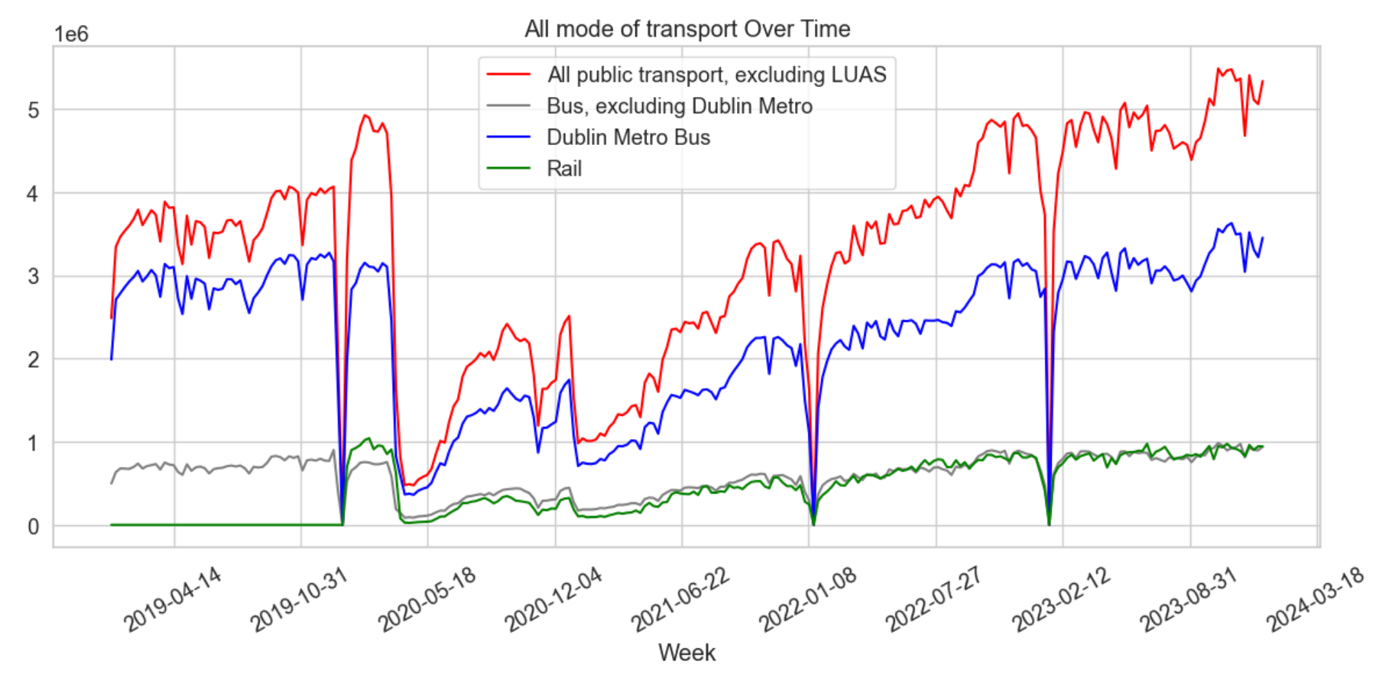


The boxplot and its data points appear greatly reduced, so the scaling method will be applied to display a clear view of all boxplots.



* Looking at the boxplot before and after interpolation, it is clear that it ended up smoothing the data by reducing the number of points outside the whiskers.
* After applying the standard scaler, not only it is possible to observe how the data is distributed but also to identify outliers in the data set.
* The outliers in the dataset are valid data points found over time, and despite being outliers, they are facts, so based on this, the outliers remained in the analysis.
* It is important to remember that the Rail transport category has its data with a zero value for all weeks of 2019.

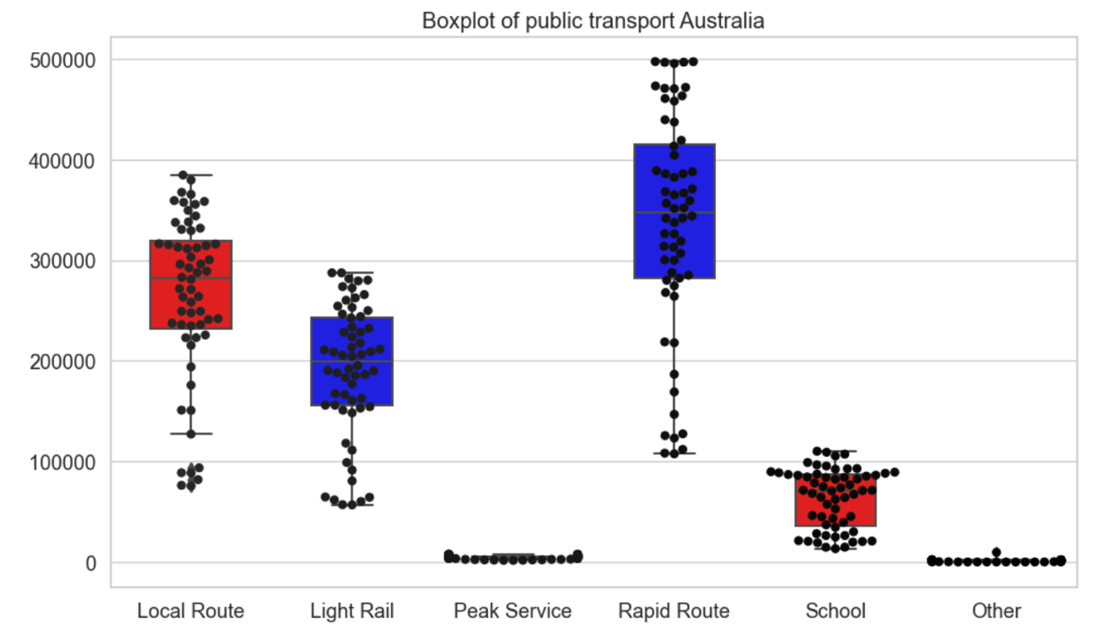
The next plot to be created in the lineplot, in this case, has the goal of observing the trend and pattern of the data in Dublin.



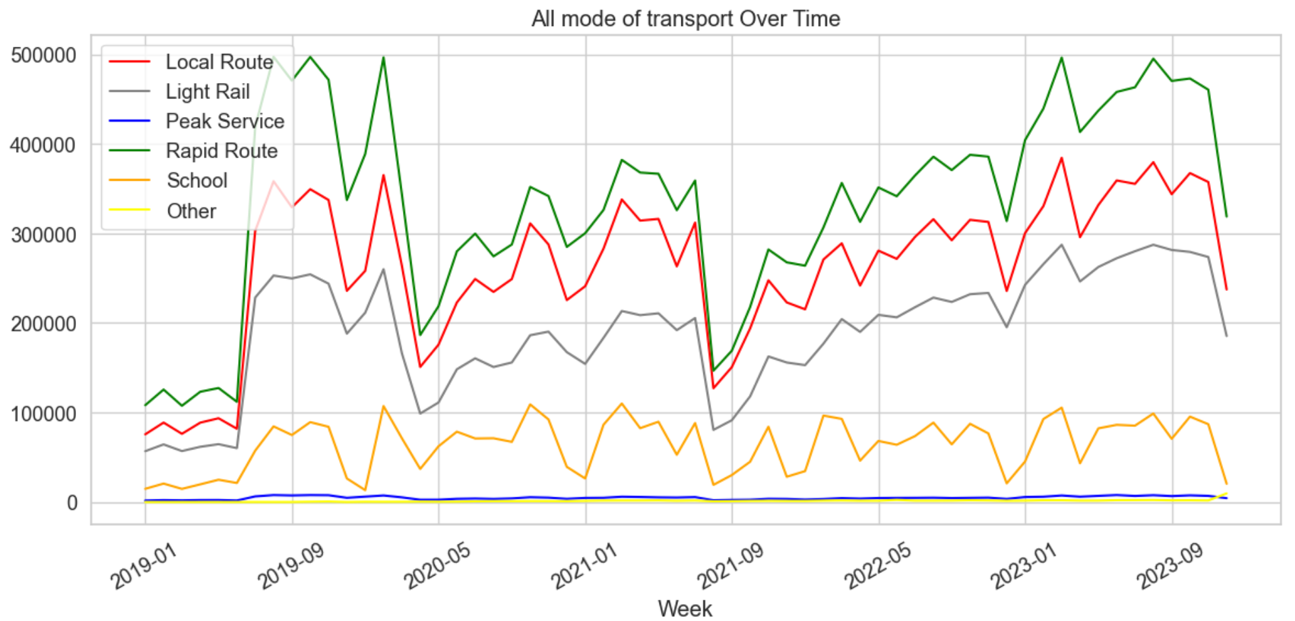
In the graph above, we can visualize how the trend in public transport use behaved from 2019 to the end of 2023, and analyze various categories of public transport and trend lines over time.

### **Visualization Canberra (Australia)**

The Australian dataset had only three missing values, which were also filled with the value zero.



* By observing the boxplot, it is possible to observe how the data is distributed, and also identify outliers in the data set.
* The outliers in the dataset are valid data points found over time and although they are outliers, they are facts, so on this basis, the outliers remained in the analysis.



Using the graph above, you can see how the weekly trend in the use of public transport behaved from 2019 until the end of 2023. Notable downward spikes, such as holidays and variations in lockdown periods, are included in this graph.

# 2 - Statistics for Data Analytics

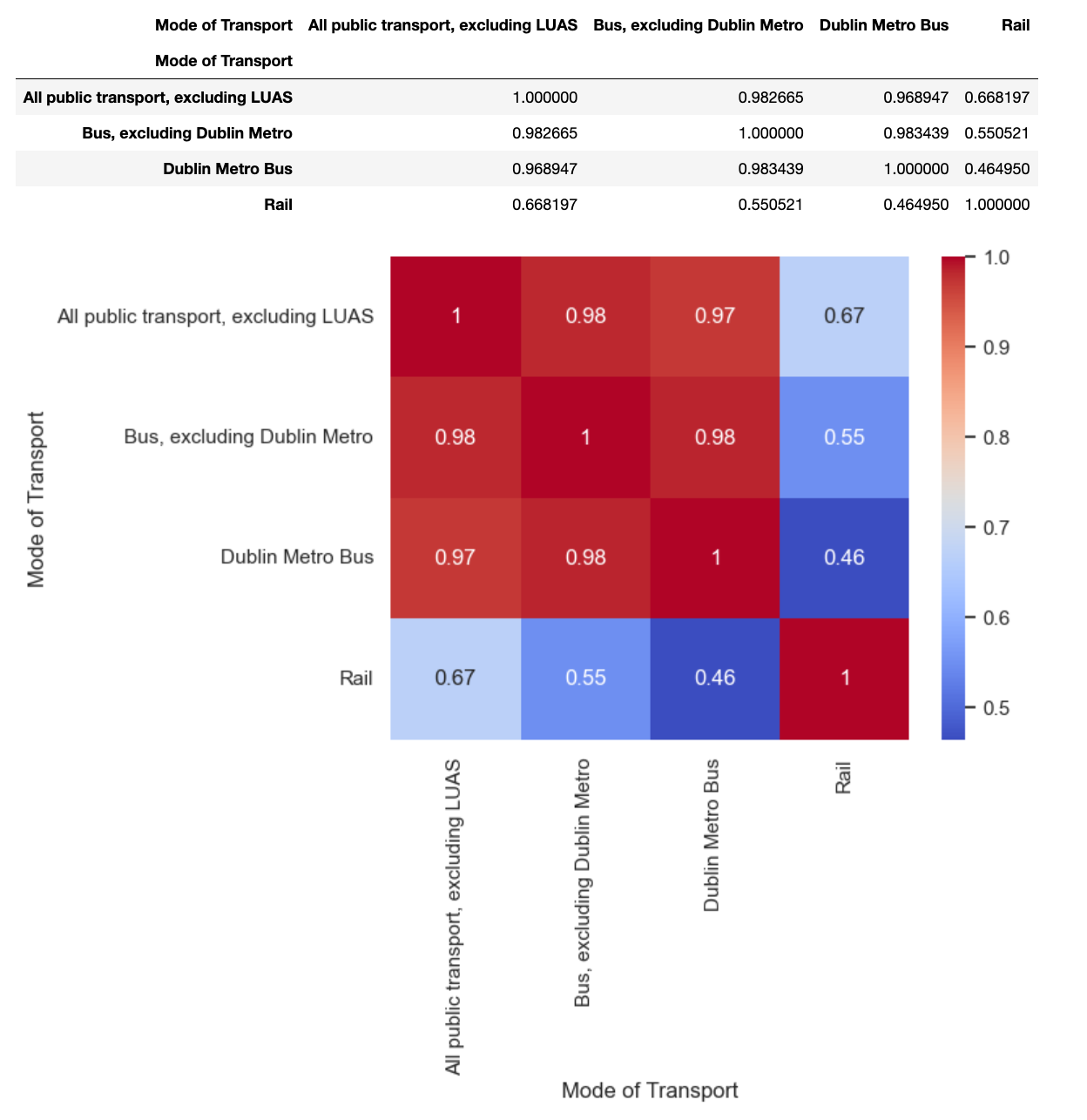
## **2.1 Distribution and correlation of all modes of transport in Dublin**

## **Descriptive statistics**

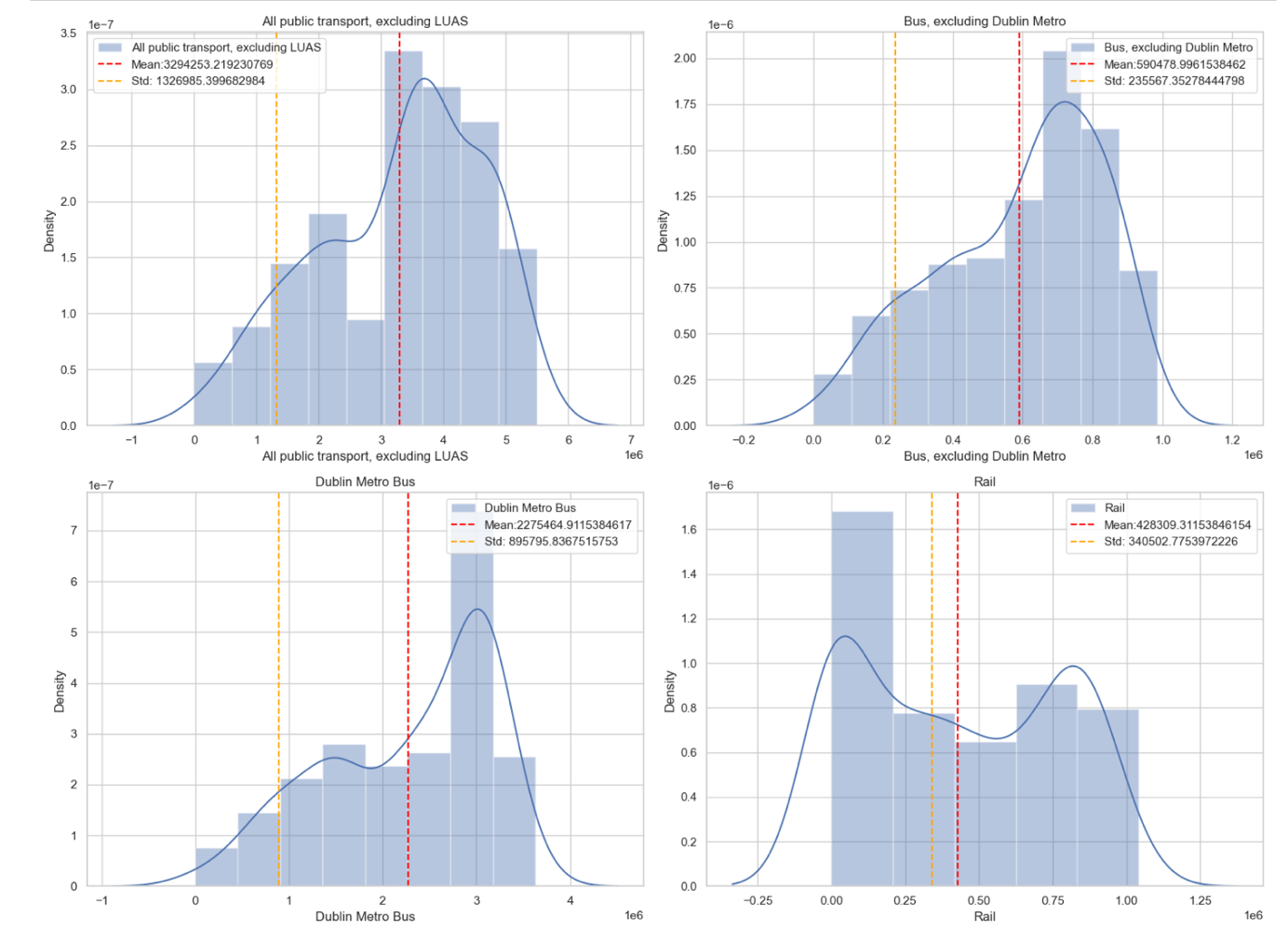
Through descriptive statistics, it is possible to identify patterns, trends, and relationships between data. It is divided into measures of central tendency and measures of variability.

Measures of central tendency describe the centre of the data set, while measures of dispersion describe how spread out the data is.

**Dublin (Ireland)**

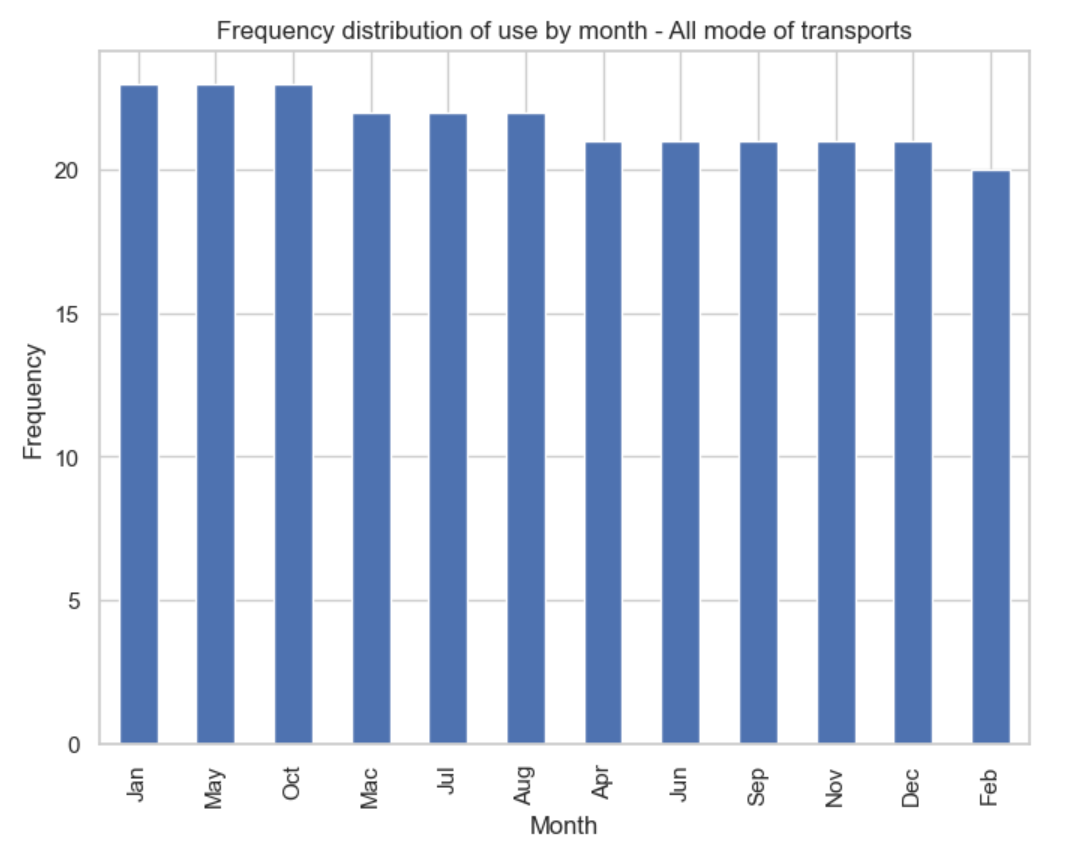


* Using the heat map, it is possible to analyze that there is a strong correlation between the bus category and Dublin Metro, and a low correlation with the Rail category, considering an independent variable

****

The graphs show the density distributions for different modes of transport from the Ireland dataset. Density distribution is a way of visualizing the probability of a value occurring in a set of data. The greater the height of the curve, the greater the probability that a value will be in this range

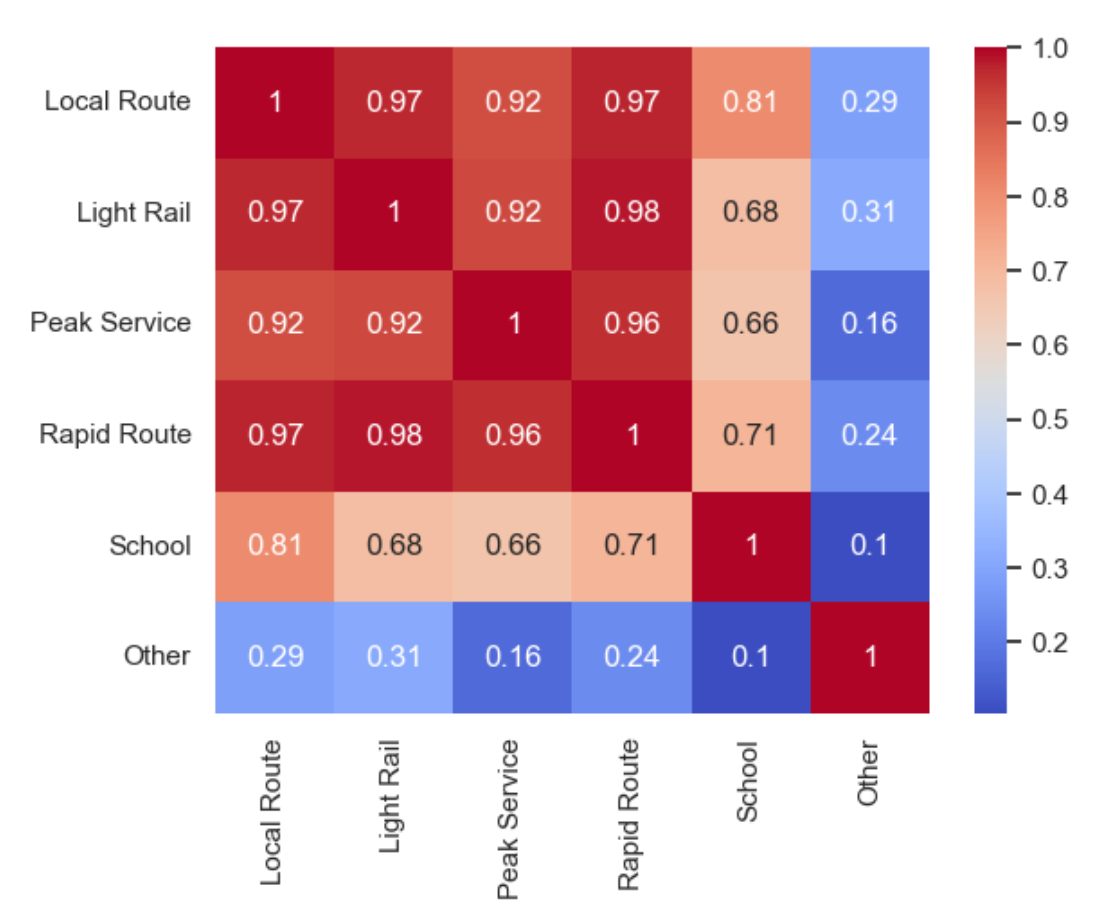
The red vertical lines represent the means and the yellow lines represent the standard deviation limits, which measure the variation of the data around the mean. The narrower the curve, the smaller the standard deviation and the more concentrated the data around the mean



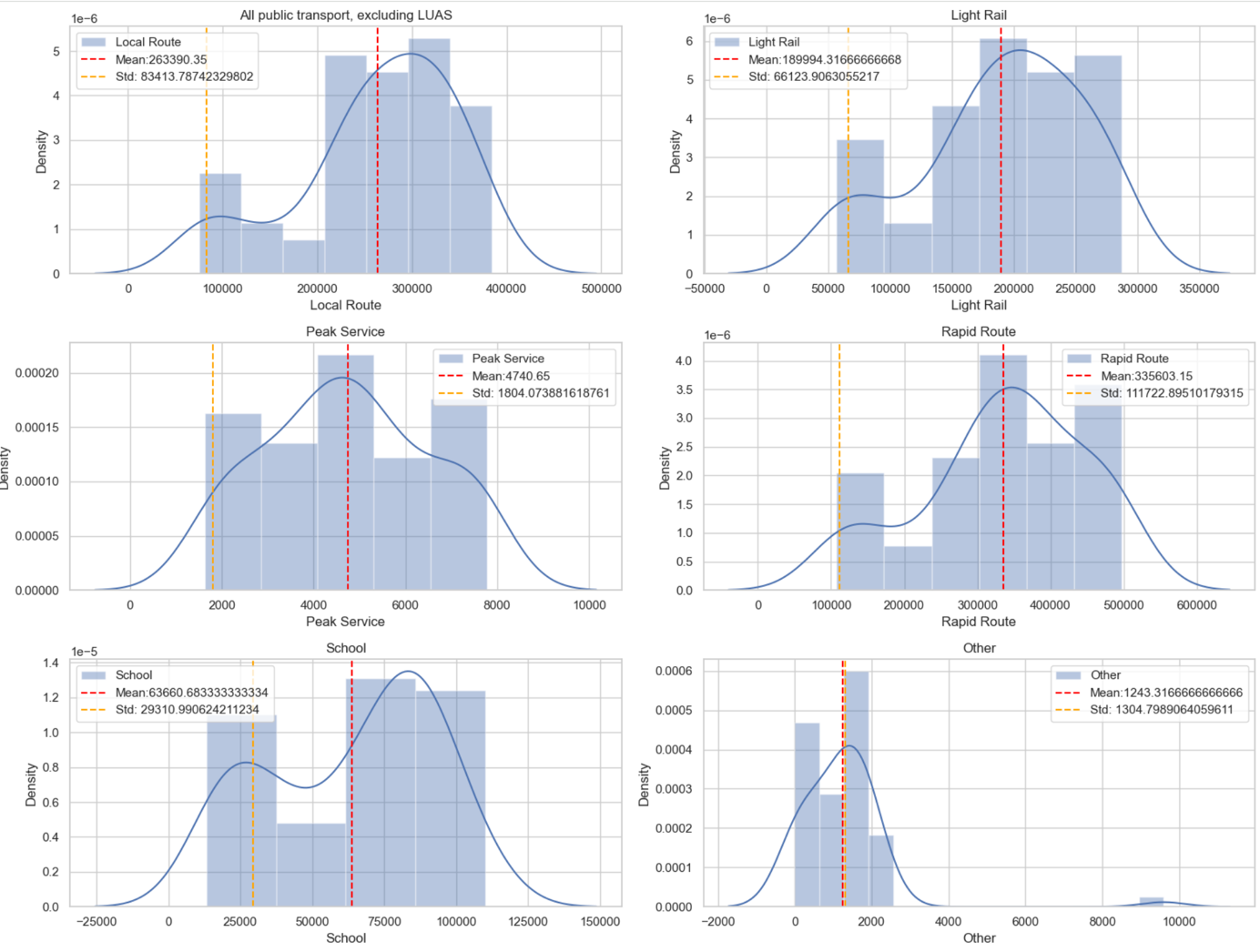
The bar chart above displays the distribution of frequency of use per month for all modes of transport present in the dataset, showing the amount of transport activity for each month represented by the month names on the x-axis and the frequency on the y-axis.

Identifying the months with the highest usage rates of January, May, October, and February with a lower usage frequency.

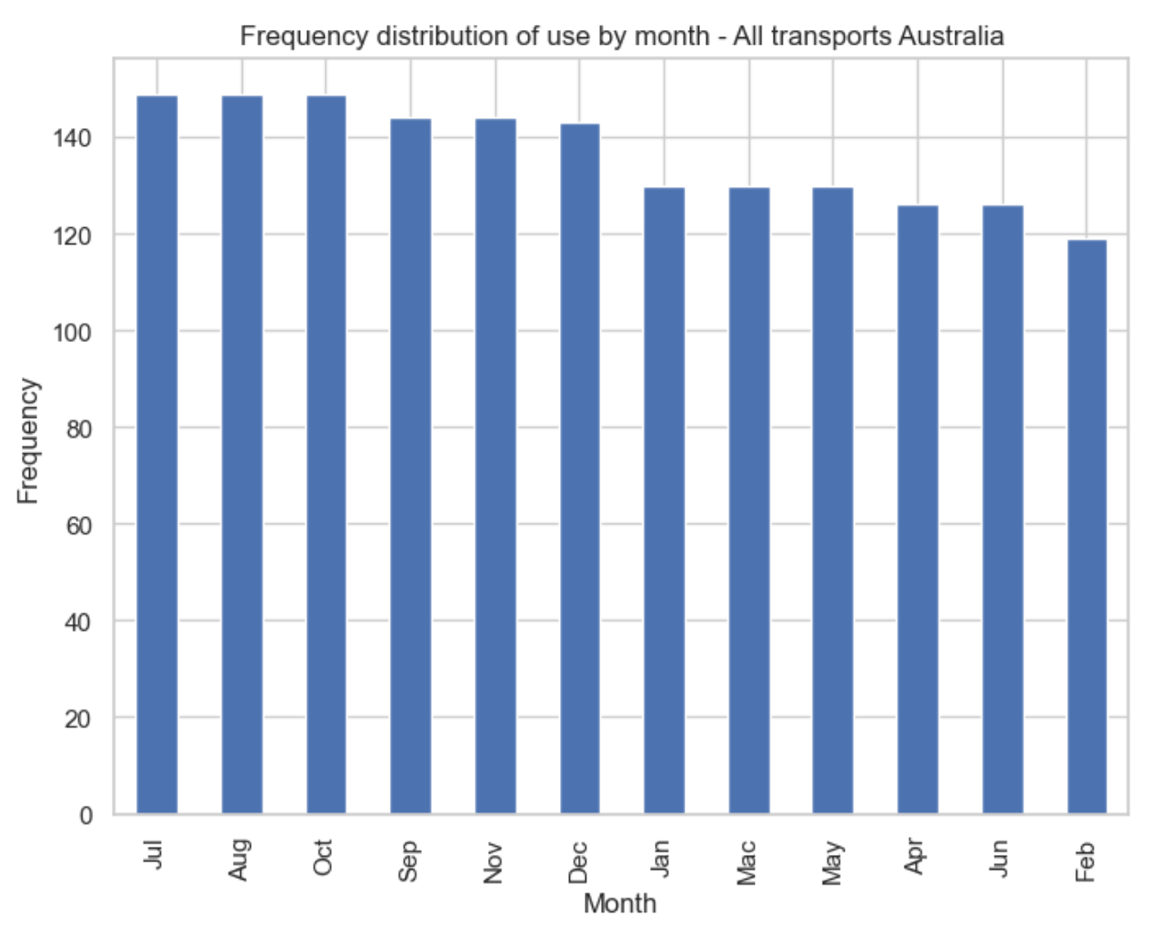
**Canberra (Australia)**



* Using the heat map, it is possible to analyze that there is a strong correlation between the bus local route, Light Rail, Peak Service and Rapid Route category and School and Other category shows a low correlation, considering an independents variable.



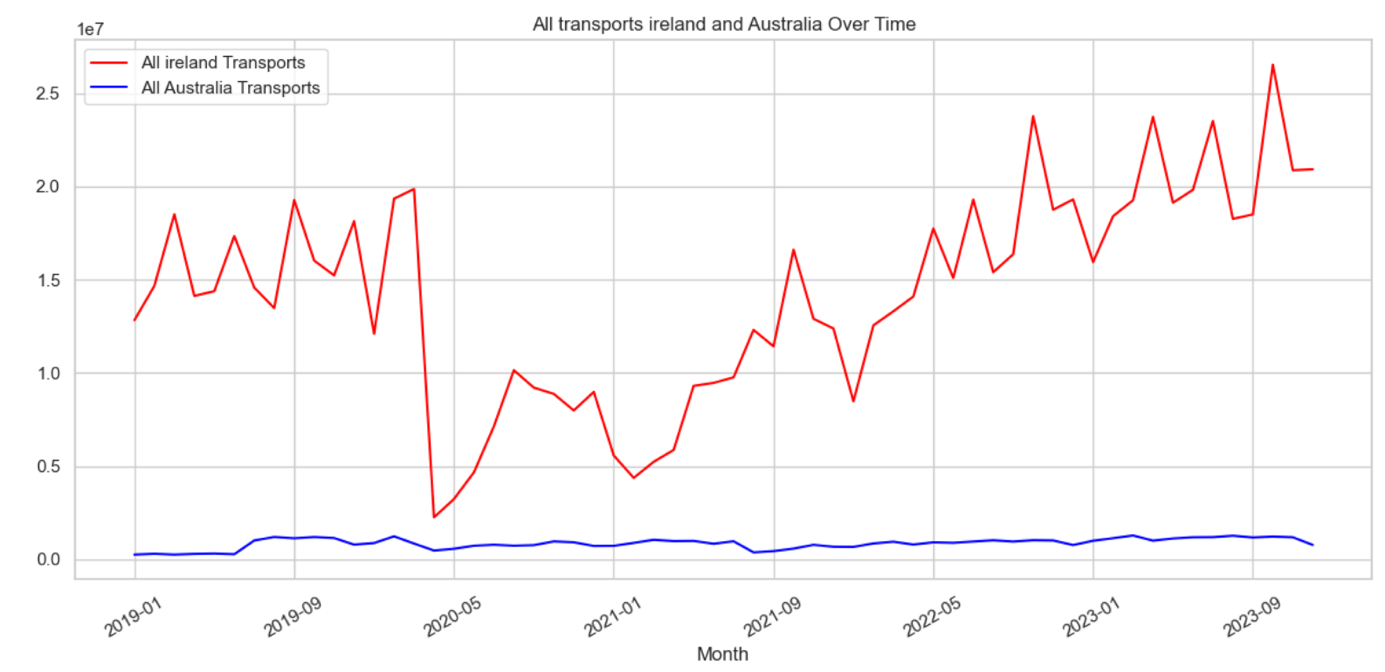
* The graphs show the density distributions for different modes of transport from the Australian dataset. Density distribution is a way of visualizing the probability of a value occurring in a set of data. The greater the height of the curve, the greater the probability that a value will be in this range
* The red vertical lines represent the means and the yellow lines represent the standard deviation limits, which measure the variation of the data around the mean. The narrower the curve, the smaller the standard deviation, and the more concentrated the data around the mean



The bar chart above displays the distribution of frequency of use per month for all modes of transport present in the dataset, showing the amount of transport activity for each month represented by the month names on the x-axis and the frequency on the y-axis.

Identifying the months with the highest usage rates of July, August, and October and April, June, and February with a lower usage frequency.

**Compare All transports in Ireland and Australia**

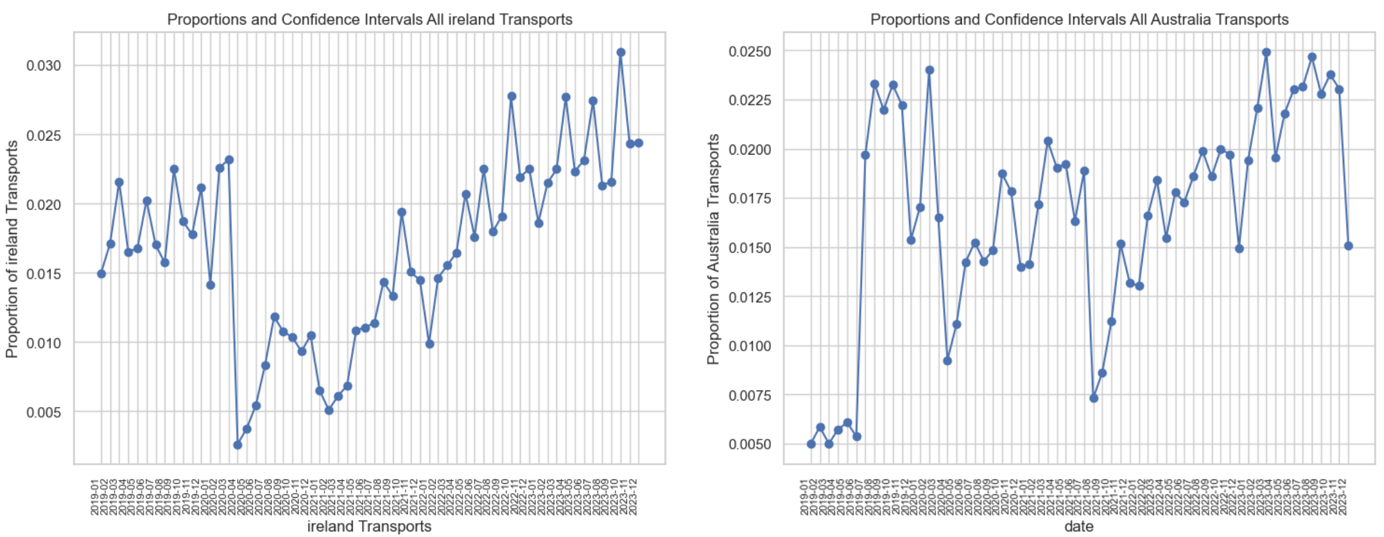
****

* This graph shows the monthly flow of public transport usage in Ireland x Australia, and we can already see the discrepancy between the two cities, Dublin shows to have extremely high numbers compared to Canberra.

**Sampling**

A suitable sampling technique can be used to strike a balance between accuracy and efficiency.

For this project, a 20% sample was used, which failed to produce a good result in the evaluation (using ECDFs - Empirical Cumulative Distribution Function) which resulted in a good result.

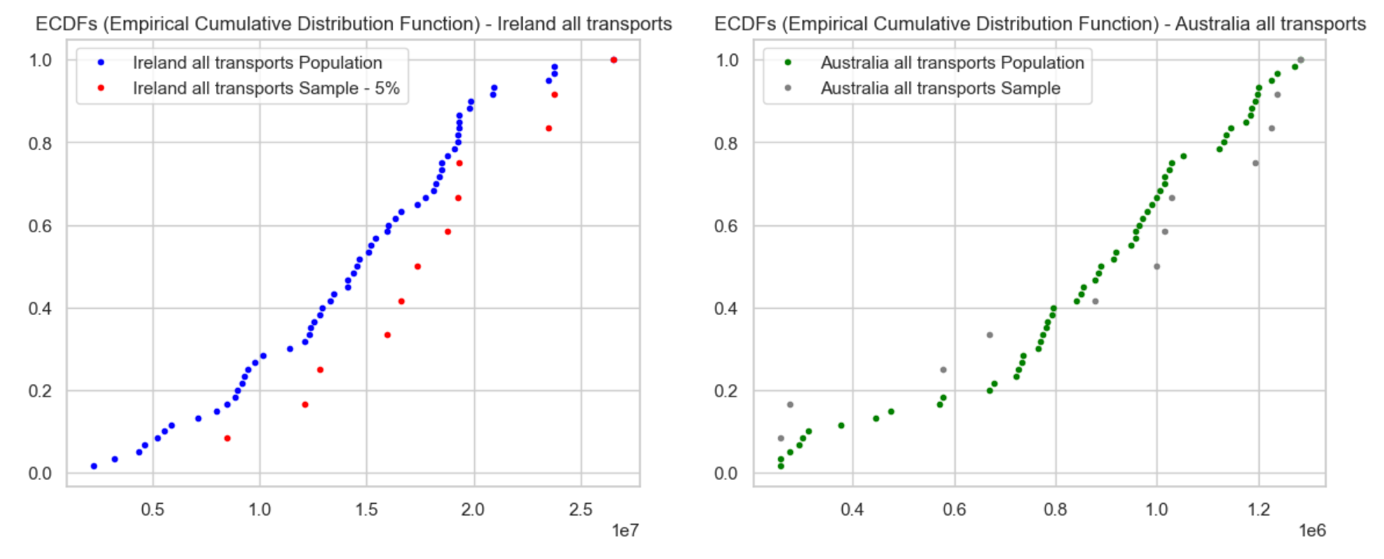


* This is a line graph with confidence intervals for the proportions of the two data sets relating to transport in Ireland and Australia. Overall it is a comparison of the proportions of transport in Ireland and Australia over time, with confidence intervals highlighted.

## **Empirical Cumulative Distribution function**

The Empirical Cumulative Distribution function calculates the empirical cumulative distribution function for a set of data. The function returns an array of ordered values ​​and a probability array associated with those ordered values, representing the proportion of data points that are less than or equal to each value.

The “ecdf” function is used to calculate the empirical cumulative distribution function for transport-related datasets in Ireland and Australia. Then, ECDF plots are plotted for these datasets, representing the cumulative distributions of the data.

****

In the graphs generated above, the blue and green dots show the confidence interval, and the red and gray dots show the data with the 20% sampling. The behavior of the Australian dataset appears to be better than that of Dublin, but it should be noted that the Irish dataset had 64 missing values ​​filled with the value zero, which may have resulted in greater dispersion, in general, they show acceptable behavior.

### **Parametric Test (t-test)**

This test is used to determine whether there is a significant difference between the means of two samples.

t-test: t-statistic = 18.468848958110282, p-value = 1.3090227661951214e-36

Conclusions: There is a significant difference between the two data sets.

### **Nonparametric Test (Wilcoxon)**

This test is a non-parametric technique used to evaluate whether two independent samples were selected from populations with the same distribution. This is an alternative to independent samples, especially when the data does not follow normal distributions.

Wilcoxon test: U-statistic = 3600.0, p-value = 3.5565709749847226e-21

Conclusions: There is a significant difference between the two data sets.

### **Chi-Square Test**

A statistical technique used to evaluate the association or independence between two categorical variables. It is used when you want to compare the frequency distribution observed in a crossover table with an expected distribution under the null hypothesis of independence of variables.

Suppose we want to compare the distribution of days of the week between the two data sets

Chi-square test: Chi2 statistic = 3540.0, p-value = 0.23850194637923627

Conclusions:

There is no significant difference between the two data sets.

### **Hypothesis testing (comparing of means using t-test)**

Performs a t-test for two independent samples with unequal variances and interprets the result based on the p-value. The t\_stat is the value of the t-statistic, which measures the difference between sample means in terms of standard deviations. The p\_value is the p-value associated with the statistical test.

In this case, the t-test is used to determine whether the means of the two samples, Ireland Transports and Australia Transports, are different in a statistically significant way, based on the p-value calculated from the observed samples.

T-statistic: 18.468848958110286

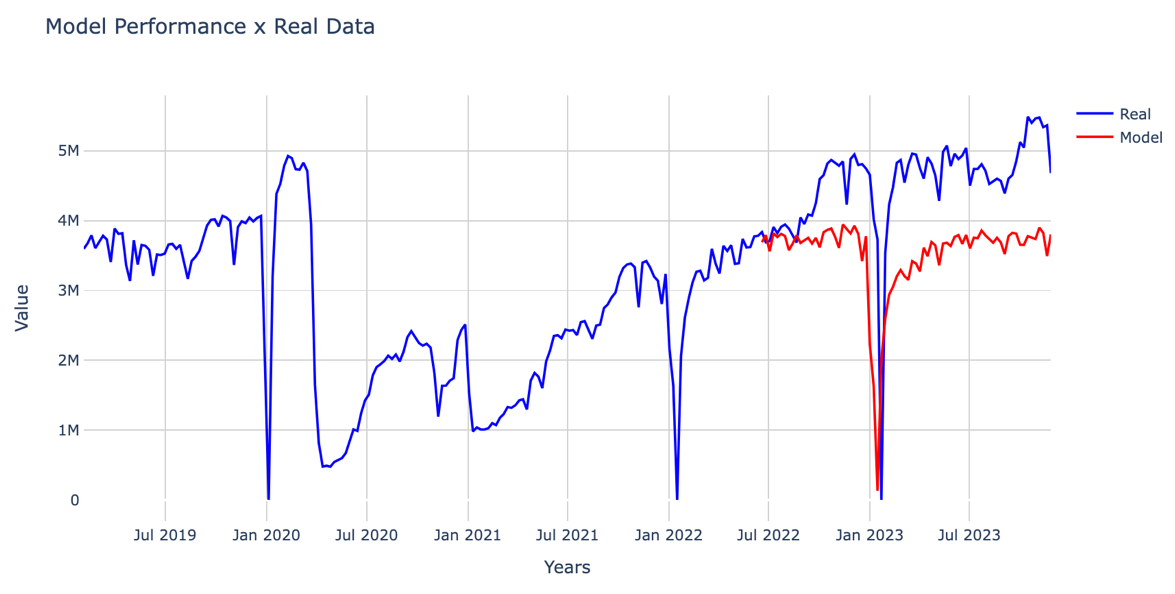
P-value: 2.845959797458664e-26

Reject the null hypothesis: Ireland uses more public transport than Australia.

# 3 - Machine Learning

Three methods were used to find the best regression model. The first is a relatively newer method, SGBRegressor, which runs through the XGBoost library and is considered an efficient implementation of gradient boosting; the technique is frequently used for predictive modeling. For this model comparison, the other two models are Lars and the Decision tree using the sklearn library.

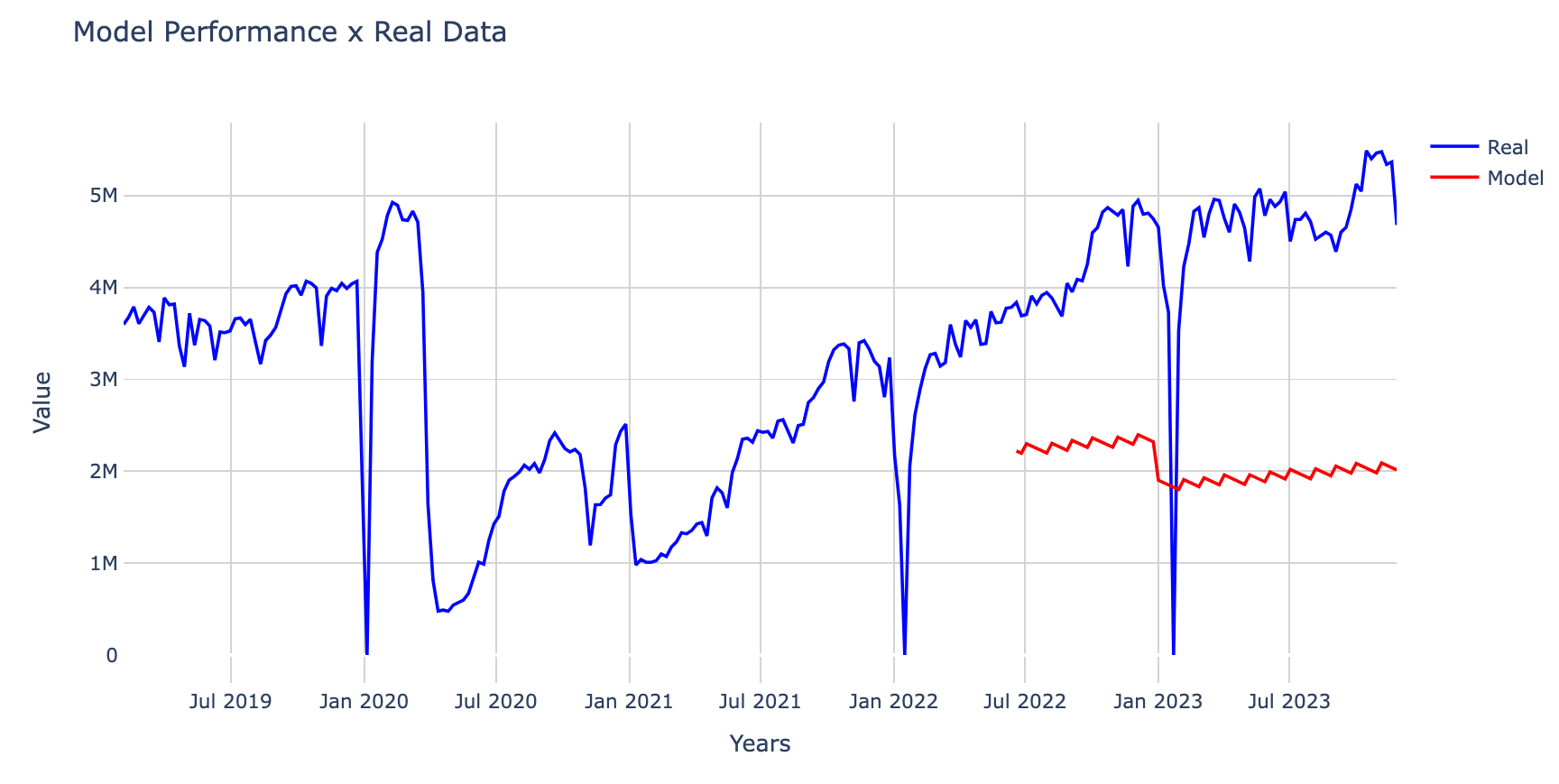
## **SGBRegressor**



'R2 Score: 1486982414494.3496'

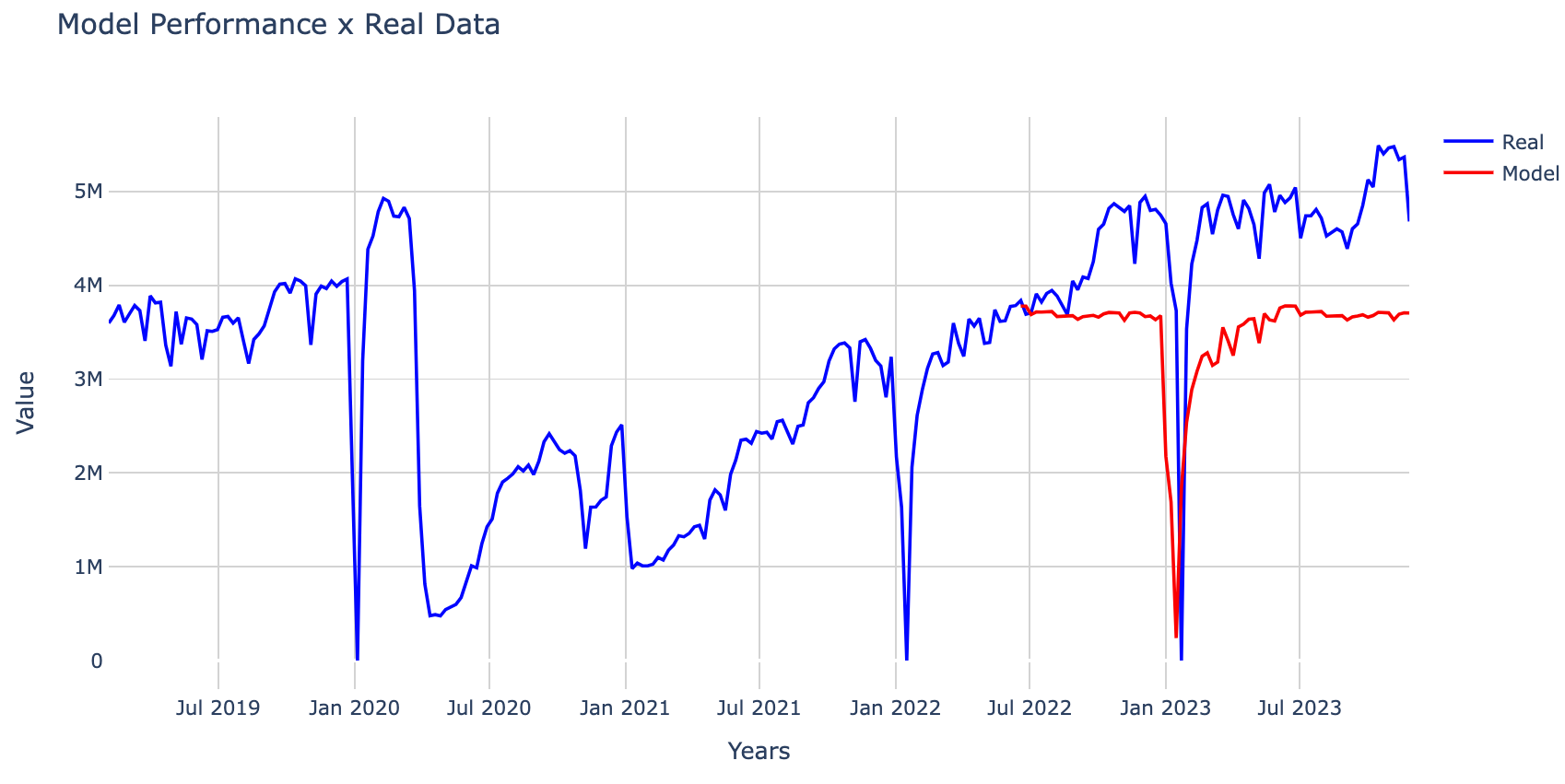
The commented hyperparameters are due to the high computational cost of executing Gridsearch.

**Lars Model**



'R2 Score: 3577259995699.883'

**Decision Tree model**



'R2 Score: 1479219964111.491'

SGBRegressor was the algorithm that, among the three tested, presented the most acceptable result, and this can be proven graphically.

## **Sentiment Analysis**

**Loaded data through the API, and transformed the JSON into a pandas Dataframe**

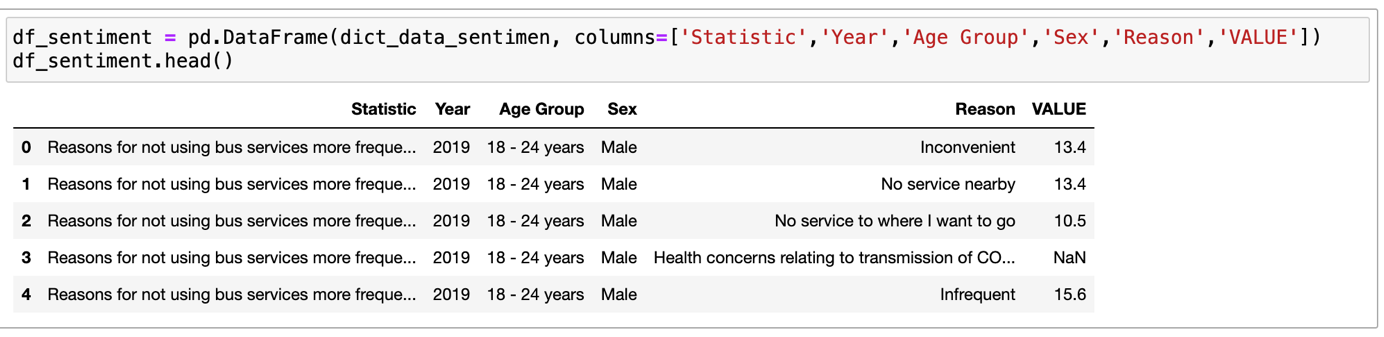
"STATISTIC: dict\_values(['Reasons for not using bus services more frequently'])"

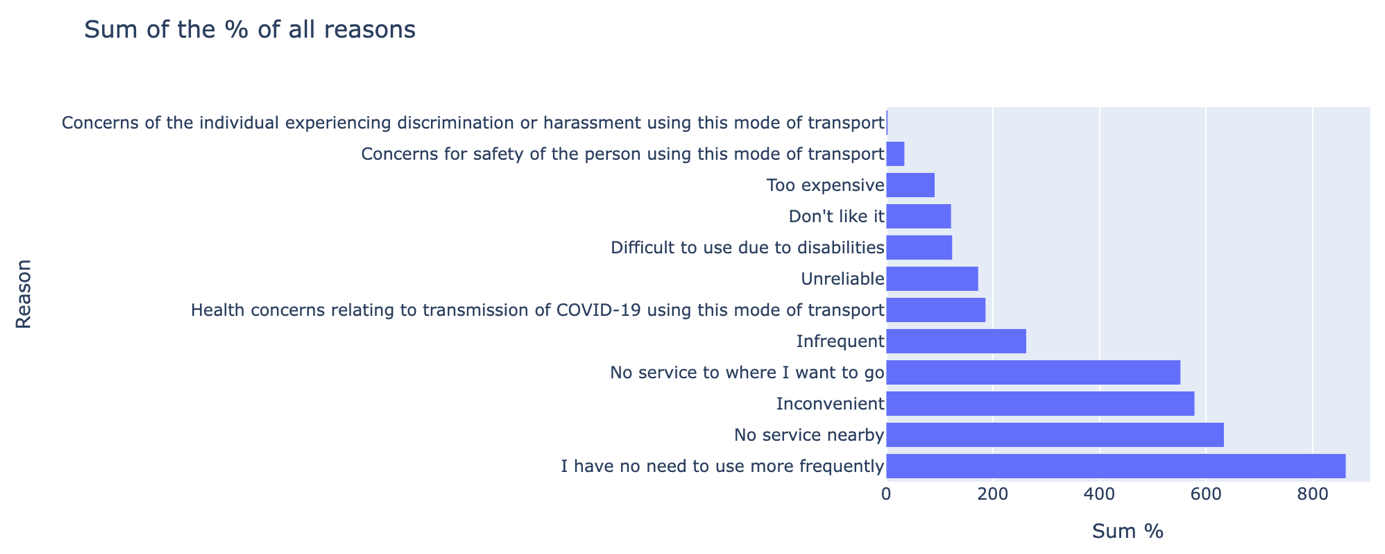
"TLIST(A1): dict\_values(['2019', '2021'])"

"C02076V02508: dict\_values(['18 - 24 years', '25 - 34 years', '35 - 44 years', '45 - 54 years', '55 - 64 years', '65 - 74 years', '75 years and over'])"

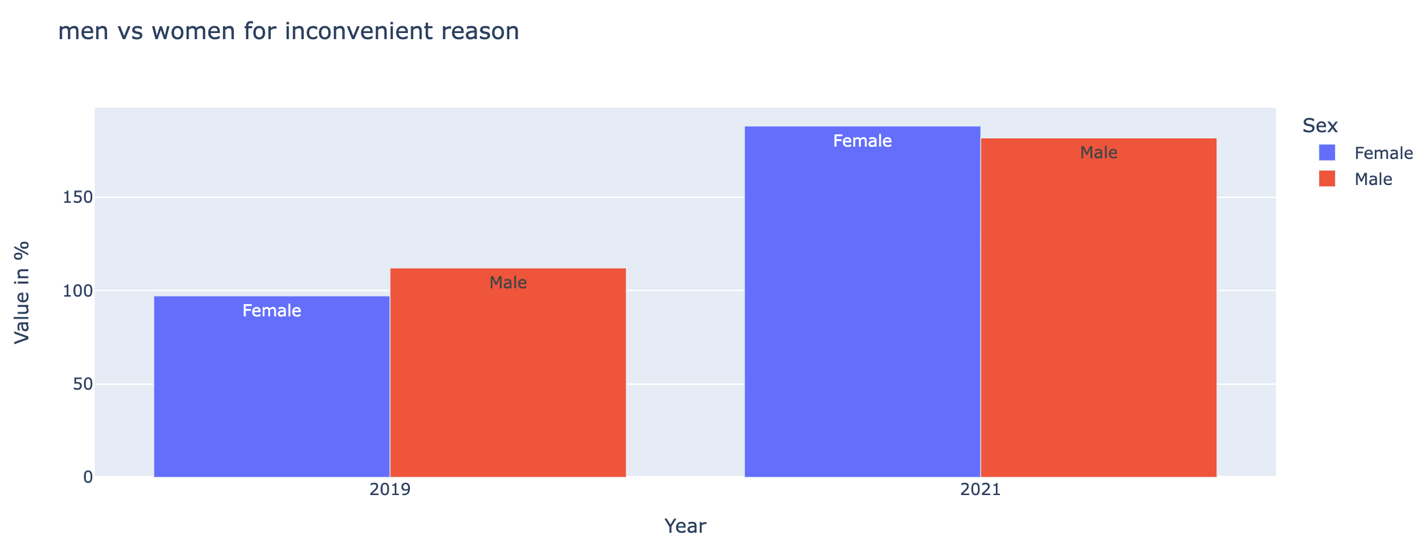
"C02199V02655: dict\_values(['Male', 'Female'])"

'C03660V04402: dict\_values([\'Inconvenient\', \'No service nearby\', \'No service to where I want to go\', \'Health concerns relating to transmission of COVID-19 using this mode of transport\', \'Infrequent\', \'Unreliable\', "Don\'t like it", \'Difficult to use due to disabilities\', \'Too expensive\', \'Concerns for safety of the person using this mode of transport\', \'Concerns of the individual experiencing discrimination or harassment using this mode of transport\', \'I do not need to use more frequently\'])'





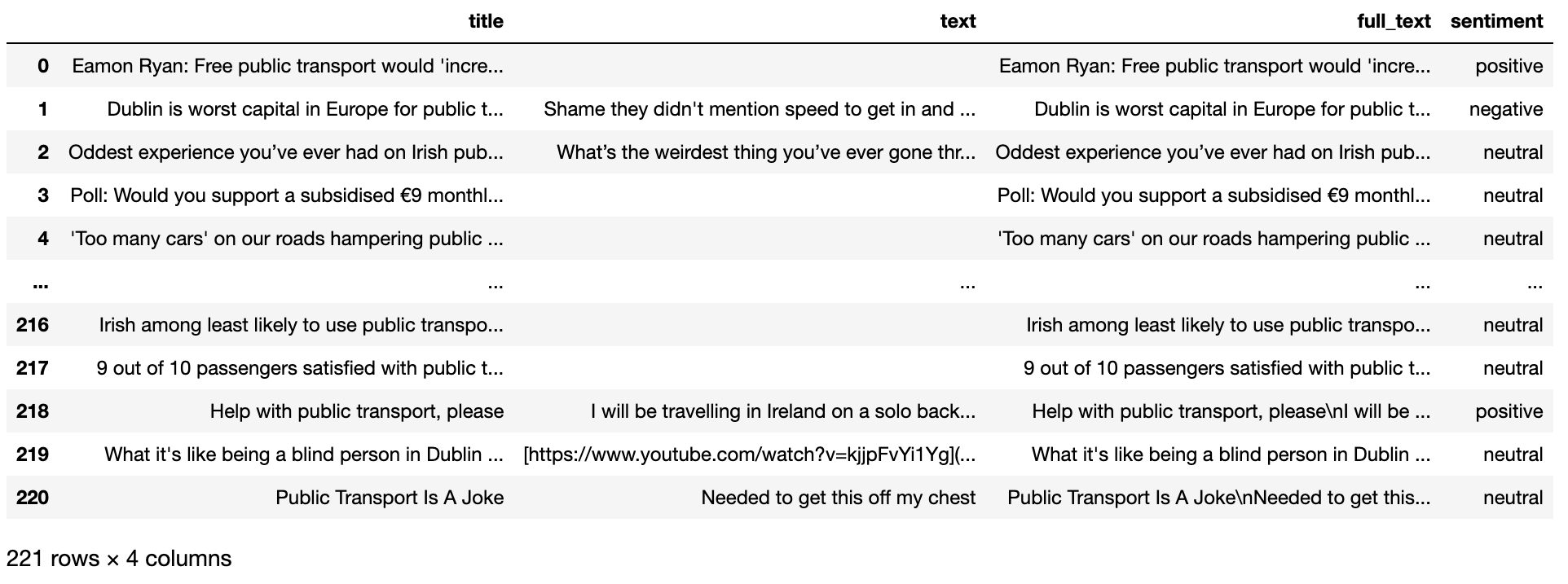
The biggest causes are not needing to use public transport frequently, there is no bus operation nearby and the service is inconvenient.



Complaints about inconvenience in the service in 2021 were more commented on by men, while in 2023 women provided more of this type of opinion.

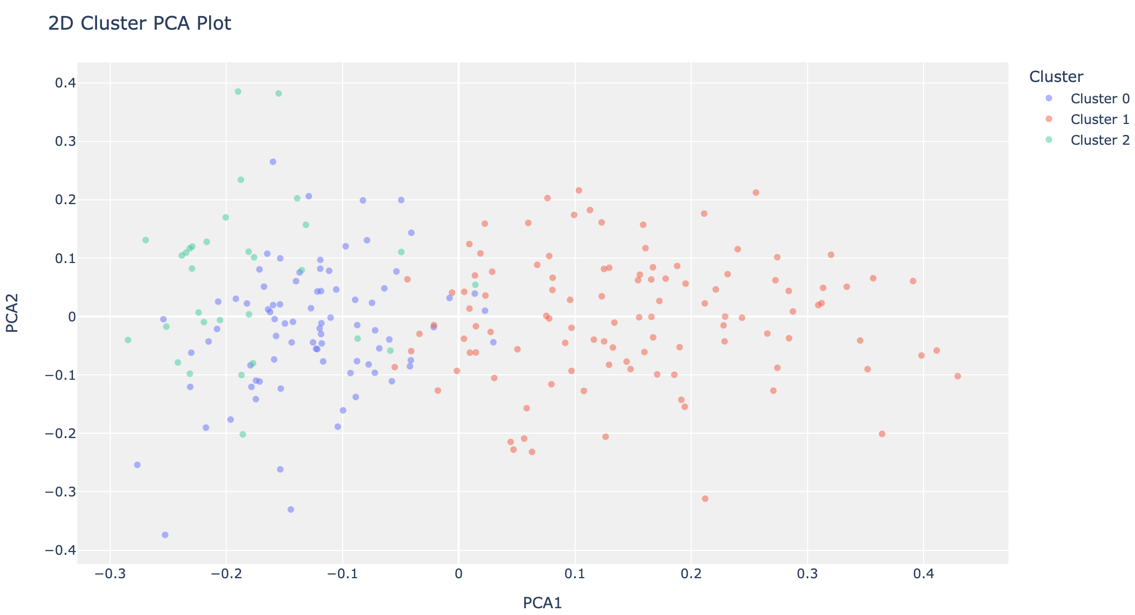
## **Sentiment analysis with VADER**

For Sentiment Analysis, another technique applied was VADER (Valence Aware Dictionary and Sentiment Reasoner), chosen because it was specifically designed for sentiment analysis in social media texts and was trained with this data. Additionally, the clustering of learning models using the more automated library Pycaret was used to improve the accuracy of sentiment classification.

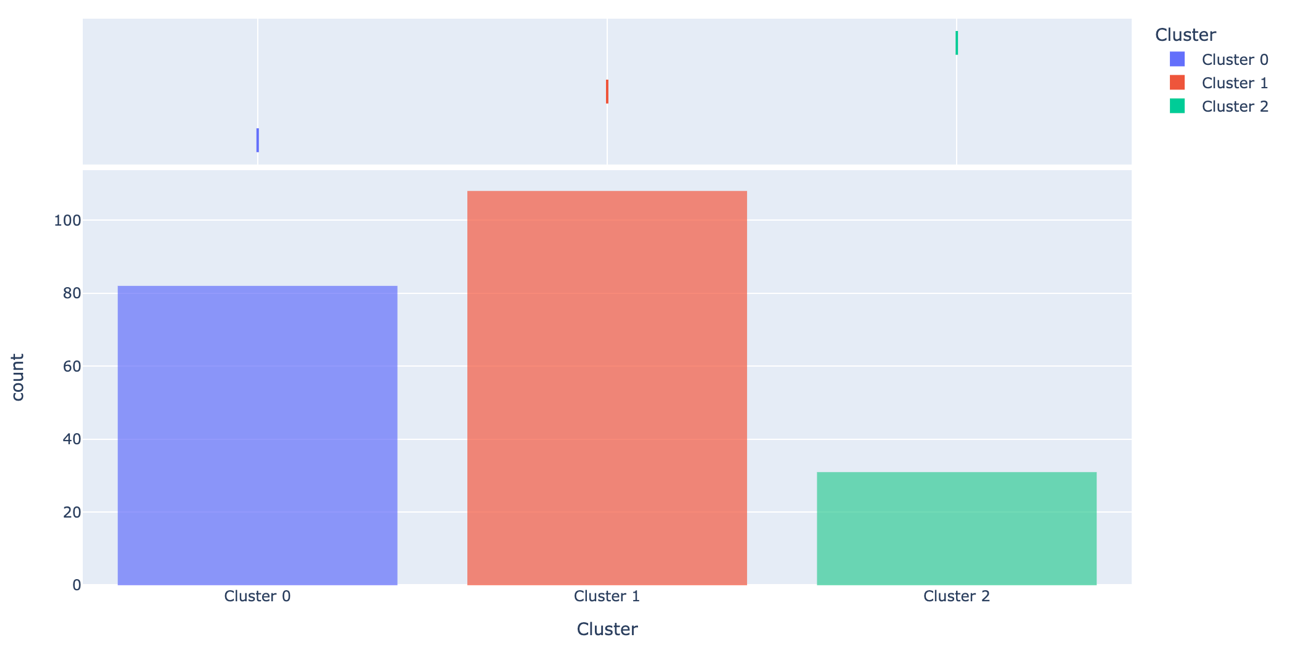


**Performing simple clustering with PyCaret**

**2D clustering PCA Plot**

****

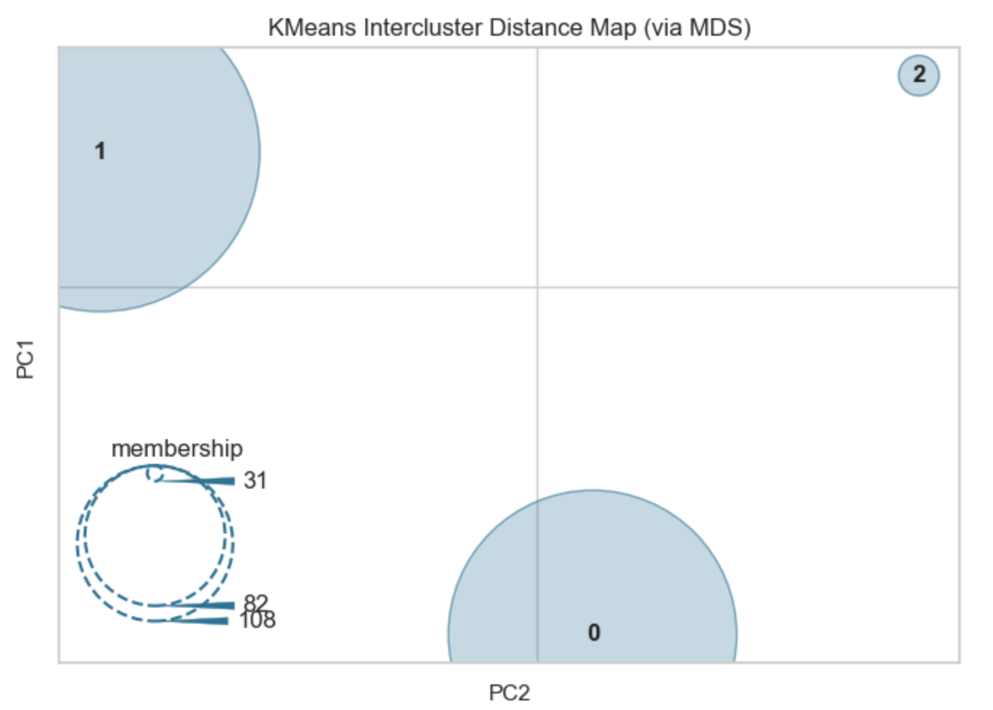
**Clustering Distribution**

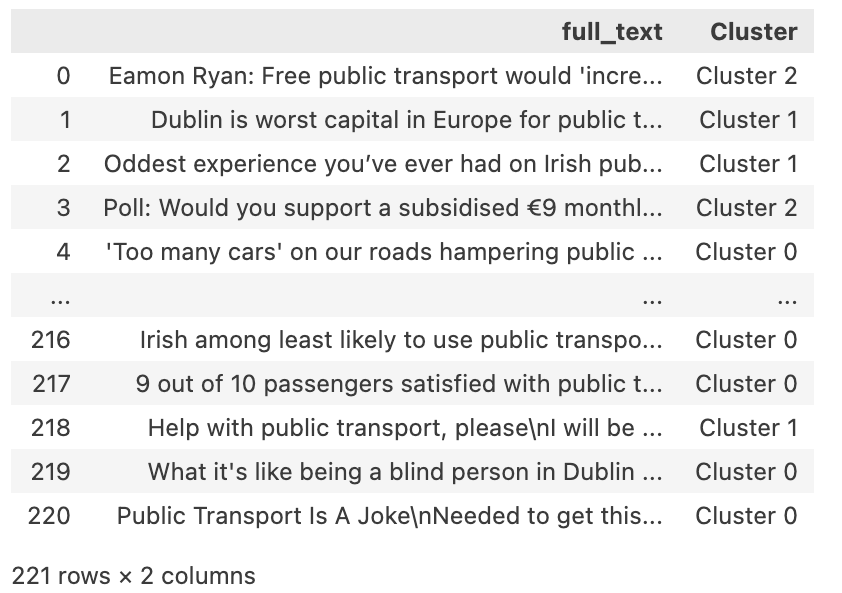
****

**KMens Intercluster Distance Map**

Through descriptive statistics, it is possible to identify patterns, trends, and relationships between data. It is divided into measures of central tendency and measures of variability.

Measures of central tendency describe the centre of the data set, while measures of dispersion describe how spread out the data is.





From this clustering, it is possible to analyze some lines of each cluster to check if it is possible to define the clusters as being ``positive``, ``negative`` or ``neutral`` comments.

A more precise analysis could be carried out based on the comments on each Reddit publication, and better data processing could be carried out, as when it comes to texts, good cleaning, and preparation directly influence the final result of the clustering. This would require more working time. But in the general context, tests were carried out, comments were demonstrated and clustering was presented.

# 4 - Programming

## **4.1 – Programming**

Evidence in Jupyter notebook

## **4.2 - Data Structures**

The data collection process was done in two main formats: CSV files, sentiment analysis using a JSON dataset, and Reddit through APIs, in JSON format.

## **4.3 Documentation**

Evidence in the Jupyter Notebook and Report.

## **4.5 Data manipulation**

a- Processing: From the pandas library, pivot was used to create new columns transforming them into date and time.

b - Aggregation of the respective Praw was the library used for API wrapper for Reddit, extracting comments about Dublin public transport in sentiment analysis

# Conclusion

This study shows that, even though the Australian dataset has more transport categories, Dublin has a higher frequency of public transport use than Canberra because of the significant population difference between the two cities. Even when there were discrepancies between the datasets and zero values were used to fill in missing values, the applied models displayed acceptable behavior. Rerunning the analyses with the lines containing missing data dropping and comparing the models would be a valuable insight for the ongoing work.

# References

TUFTE, E. R. The visual display of quantitative information. Connecticut: Graphics Press LLC, 2001.

CHAPMAN, Pete et al . CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, v. 16, 2000.

[Sarah Guido](https://www.amazon.co.uk/Sarah-Guido/e/B0788K4D8N/ref=dp_byline_cont_book_1) and Andreas C. Mueller (2016) **Introduction to Machine Learning with Python: A Guide for Data Scientists.** O'Reilly Media; 1st edition.

<https://machinelearningmastery.com/xgboost-for-regression/> (*(Accessed 18 December 2023)*

Brownlee, J. (2017) **Machine Learning Mastery.** ‘How to handle missing data with python’ Available at: https://machinelearningmastery.com/handle-missing-data-python/ *(Accessed 08 December 2023)*;

Dan, A., (2020) Medium. ‘Exploratory Data Analysis (EDA) in Python’ Available at: <https://medium.com/@atanudan/exploratory-data-analysis-eda-in-python-893f963cc0c0> *(Accessed 15 December 2023)*;

Bonthu, H., (2021) **Analytics Vidhya.** ‘Detecting and Treating Outliers | Treating the odd one out!’ Available at: h[ttps://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/](https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/) *(Accessed 17 December 2023)*;

Kumar, V. (2021) **Machine Learning Knowledge.** ‘Categorical Data Encoding with Sklearn LabelEncoder and OneHotEncoder. Available at:

https://machinelearningknowledge.ai/categorical-data-encoding-with-sklearn-labelencoder-and-onehotencoder/ *(Accessed 17 December 2023)*;

[Yaddanapudi](https://www.orskl.com/author/administrator/), P. (2019) **Our Skills Our Future.** ‘Will highly correlated variables impact Linear Regression?’ Available at:https://www.orskl.com/will-highly-correlated-variables-impact-linear-regression/*(Accessed 29 December 2023)*

Van Dorpe, S. (2018) **Towards Data Science.** ‘Preprocessing with sklearn: a complete and comprehensive guide’ Available at:<https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9>/*(Accessed 28 December 2023)*