# **Introduction**

Transport is a crucial component that links economic vitality and societal well-being in the complex fabric of modern life. Our eyes are led by data-driven exploration as we navigate the ever-changing terrain of Ireland's transport industry. With a focus on key London stations, this research takes readers on an enlightening tour through the narratives of Luas passenger journeys and the complex dynamics of passenger numbers in different cities.

## **Navigating the Transport Landscape Through Data**

Our main goal is to leverage data analytics and machine learning techniques to advance our understanding of Ireland's transport complexities. Two key datasets are highlighted: the first reveals the patterns of Luas passenger movements, while the second painstakingly breaks down passenger statistics between various locations and major London stations.

## **Luas Passenger Journeys: Patterns, Peaks, and Fluctuations**

The first dataset presents passenger travels on the Luas in a kaleidoscope of colours. By exposing trends, pinpointing the busiest times for travel, and highlighting popular routes, we utilise advanced visualisation techniques to extract valuable information. Every detail is analysed, from the fluctuations of passenger counts to the beat of commuter flows, to enable well-informed decision-making in the transport sector.

## **Urban Transport Dynamics: Decoding Passenger Numbers**

With a focus on key London stations, our second dataset delves further into the nuances of passenger numbers in different cities. This investigation captures the dynamics of urban movement and goes beyond numerical calculations. Together, machine learning apps and statistical inferential tools provide a holistic picture that anticipates trends and captures public opinion on transportation-related issues.

## **Unifying Insights: From Data to Decision**

The combination of statistical analysis, machine learning predictions, and qualitative sentiment analysis deepens our comprehension as we work through these datasets. Our paper lays the foundation for educated decision-making while also illuminating the complexities of transport data. This integration of knowledge helps to provide a comprehensive understanding of Ireland's transport landscape, enabling stakeholders to make informed decisions about the future.

# **Framework for Project Management**

We have implemented GitHub version control and the agile technique for efficient project management.

## **Agile Approach: Flexible and Adaptive**

We use an agile project management methodology, which facilitates iterative and flexible development. The project is broken up into sprints to allow for constant feedback and adaptability in the face of new discoveries and difficulties.

## **Version Control on GitHub: Managing Teamwork**

Our version control system, GitHub, allows us to collaborate and keeps track of how the project has changed over time. Frequent changes to the repository allow for smooth team cooperation and the creation of an open development timeline.

# **Data Analytics Task Programming**

The main programming components of our data analytics project are described in this section, with an emphasis on the useful actions and choices made during the study.

#### **Programming: Python and Jupyter Notebook**

Jupyter Notebook is used only to implement the project in Python. Python's vast libraries for data analysis and manipulation, along with Jupyter's interactive environment, make code creation a smooth process.

#### **Data Structures: CSV Files and Web APIs**

A variety of sources, such as web APIs and CSV files, are used to collect data. This method demonstrates the flexibility of data processing by enabling us to manage a wide range of data structures.

#### **Documentation: Clear Code and Quality Standards**

To maintain repeatability and transparency, our code is well documented. To ensure dependability and clarity, we follow recognised code quality standards and provide explanations for each phase of the analysis.

#### **Testing & Optimization: Rigorous Evaluation Process**

To verify code functioning and maximise resource utilisation, a comprehensive testing and optimisation approach is implemented. Recording trade-offs made during the development process is part of this.

#### **Data Manipulation: Choosing Optimal Libraries**

We examine and contrast various data processing and aggregation packages and methods. The methods that were chosen are justified by how well they fulfil the particular needs of the project.

## **Context & Background:**

Economic activity and societal mobility are significantly influenced by transportation. Transportation data analysis offers important insights for enhancing services, streamlining infrastructure, and guaranteeing a smooth travel experience. Ireland is an interesting case study for examining the nuances and patterns in passenger journeys because of its variety of transport options.

## **Objectives:**

The following are this project's main goals:

**1. Data Acquisition:** To obtain and investigate pertinent datasets concerning airport operations and passenger trips on Luas lines.

**2. Visualisation & Data Preparation:** to prepare and clean data, perform exploratory data analysis (EDA), and build an interactive dashboard for insights visualisation.

**3. Data Analytics Statistics:** In order to obtain insights, compare Ireland's transport metrics with those of other nations, and summarise datasets, descriptive and inferential statistics will be used.

**4. Machine Learning:** To use machine learning models to analyse sentiment regarding transportation-related subjects and forecast patterns in passenger movements.

The goal of this project is to advance knowledge of transport dynamics by giving stakeholders useful information for formulating sound policies and making well-informed decisions.

# **Data Acquisition**

**Positive Aspects:**

For this project, data acquisition included compiling information from many sources that provided a thorough picture of Ireland's transport environment. The following is a summary of the benefits of the data collecting process:

**Data Variety:** A comprehensive investigation of transport patterns was made possible by the use of two separate datasets, namely Luas passenger journeys and airport activities.

**Data Accessibility:** The two datasets were easily obtainable, which facilitated the process and decreased the amount of time spent obtaining the data.

**Real-time Updates:** The datasets showed up-to-date timestamps, which guaranteed their applicability and relevance to real-world transportation scenarios.

**Dataset Relevance:** The chosen datasets directly complemented the project's emphasis on air travel and public transit (Luas).

**Negative Aspects:**

**Data Quality:** The datasets had issues with missing values, inconsistent data, and the requirement for substantial preprocessing.

**Limited Context:** Although the datasets included insightful quantitative data, certain qualitative elements like traveller pleasure or particular motivations were left out.

**Relevance and Implications of Licensing/Permissions:**

**Open Data Sources:** By obtaining the airport activity dataset and Luas dataset from open data sources, worries about licencing and permissions were minimised.

**Attribution Requirements:** To ensure compliance with licencing agreements, the datasets required correct attribution to the data providers.

**Ethical Considerations:** The project gave data privacy and confidentiality first priority while adhering to ethical principles.

# **Statistics for Tasks Using Data Analytics**

#### **Descriptive Statistics and Visualizations**

To summarise the main features of the Luas passenger travels dataset, we used descriptive statistics and visualisations during our investigation. By applying metrics like mean, median, and standard deviation, we offered a thorough overview of the dataset. To illustrate trends, variations, and peak travel times over different Luas lines, visualisations like as line plots and bar charts were used.

#### **Inferential Statistics for Population Insights**

We used inferential statistics in our analysis to learn more about population values. For some measures, confidence intervals were computed, providing a more thorough comprehension of the dataset's properties. With the use of this method, we were able to go beyond superficial observations and make inferences about the larger population.

#### **Cross-Dataset Statistical Comparisons**

We compared our dataset on Luas passenger journeys with the airport activity dataset in order to further our research. To assess the passenger number distribution, we ran a Mann-Whitney U Test. The test findings showed a significant difference in the passenger distribution between the two datasets, with a p-value of 0.0000 and a U-statistic of 0.0000. Important insights into the unique trends within each transportation domain were obtained from this comparison investigation.

#### **Yearly Variations and Luas Line Dynamics**

We also conducted a statistical investigation of Luas line dynamics and annual variations. When means from different years were compared using an ANOVA, significant differences were found, as shown by a p-value of 0.0000 and an F-statistic of 119.1020. Furthermore, the relationship between Luas line changes and weeks was investigated using a Chi-squared test. The results, which were unexpectedly non-significant (Chi2-statistic = 0.0000, p-value = 1.0000), prompted additional research into the subtleties of Luas line patterns.

#### **Temporal Patterns and Consistency Testing**

In order to examine temporal trends, specifically the Red and Green lines, paired t-tests were utilised. A significant difference was revealed by the data (t-statistic = 9.6234, p-value = 0.0000), suggesting different historical trends. Additionally, there was no discernible mean difference between the first and last years according to an independent t-test, indicating some temporal consistency.

#### **Addressing Precision Loss and Result Reliability**

We saw precision loss when performing statistical calculations, especially when dealing with almost identical data. This increased understanding of the significance of accepting probable precision limitations and the dependability of outcomes. To maintain transparency and a nuanced interpretation of the results, such challenges were documented.

# **Data Preparation & Visualization Tasks**

## **Acquiring Raw Data**

We obtained two separate datasets that are essential to comprehending Ireland's transport environment to create the foundation for our investigation. The first dataset explores passenger travels on the Luas, while the second gives a comparative view by listing the number of passengers at stations in central London.

## **Positive Aspects:**

**Data Richness:** The Luas dataset contains a wide range of statistics, such as line-specific details, passenger counts, and temporal nuances. Despite being physically separate, the London station dataset enhances our research by offering a comparative perspective.

## **Negative Aspects:**

Data Consistency: There were difficulties in ensuring consistency between datasets. It was clear that thorough cleaning and harmonisation procedures were required.

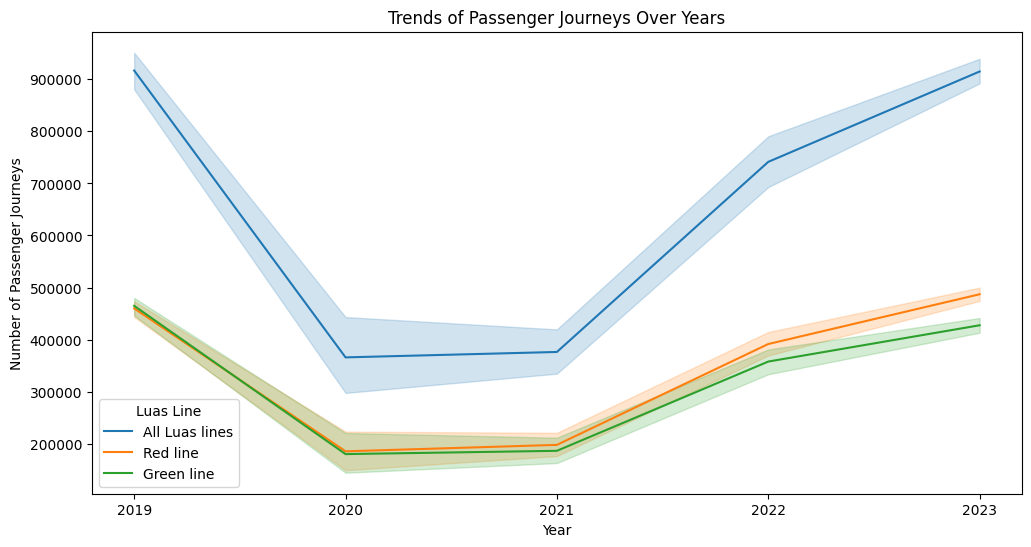
## **Exploratory Data Analysis (EDA)**

Finding trends, inconsistencies, and possible problems in the datasets was made possible in large part by our EDA methodology. Let's explore the adventurous journey of each dataset.

### **Time Patterns Luas Passenger Journeys EDA:**

**Visualising Temporal Trends:**

To show peak travel times and temporal fluctuations, line plots were used to visualise passenger counts across time.



we can see that from 2020 to 2021 around the time of corona virus there is a dip in the passenger counts but after 2021 we can see I is gradually increases for each year.

**Statistical Validation**

**Paired t-test:** A paired t-test was used to thoroughly analyse the temporal patterns between two different lines (Red line vs. Green line). With a p-value of 0.0000 and a t-statistic of 9.6234, the results showed significant differences.

Our EDA process was instrumental in uncovering patterns, inconsistencies, and potential issues within the datasets. Let's delve into each dataset's exploratory journey.

### **Line Variations:**

**Understanding Line Distribution:**

A graph of a passenger journey

Description automatically generated

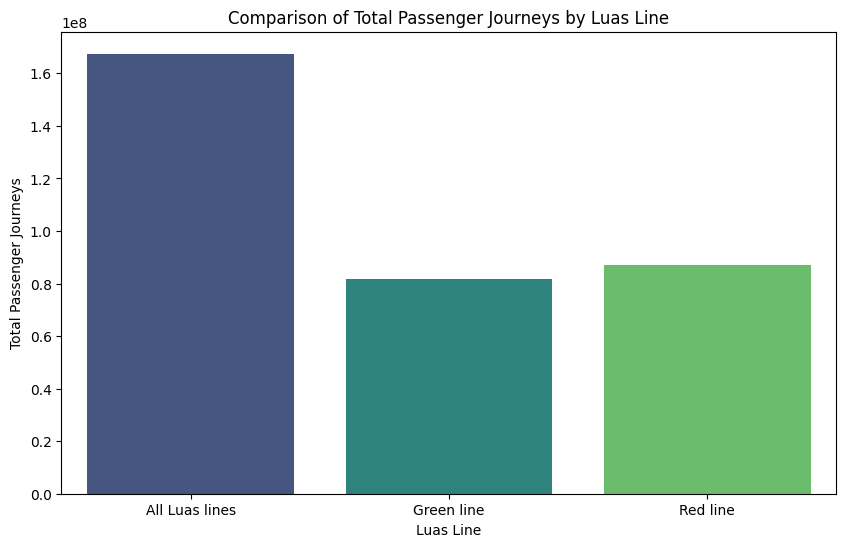
Crosstabulations were conducted to understand the distribution of passengers across different Luas lines.

### **Statistical Validation**

**Chi-squared Test:** A chi-squared test for Luas line variations confirmed the association between categorical variables, aiding in the analysis of variations in Luas lines across weeks. The Chi-squared test for Luas line variations yielded a Chi2-statistic of 0.0000, suggesting a lack of evidence to reject the null hypothesis. The associated p-value of 1.0000 reinforces this finding, indicating that the observed variations in Luas lines across weeks are statistically non-significant. This statistical test assesses the association between categorical variables, and in this context, it implies that the distribution of Luas line occurrences across different weeks is not significantly different from what would be expected by chance.

### **Data Transformations:**

**Comparison of Luas Lines:** A bar plot comparing the total passenger journeys for different Luas lines was created to visualize the distribution.



We can see that Red line have more passenger journeys as compared to green line.

### **Scatter Plots and Distributions:**

**Pairplot:** A pairplot was constructed to visualize relationships between key variables like passenger count, week number, and year.

A graph of a graph of a function

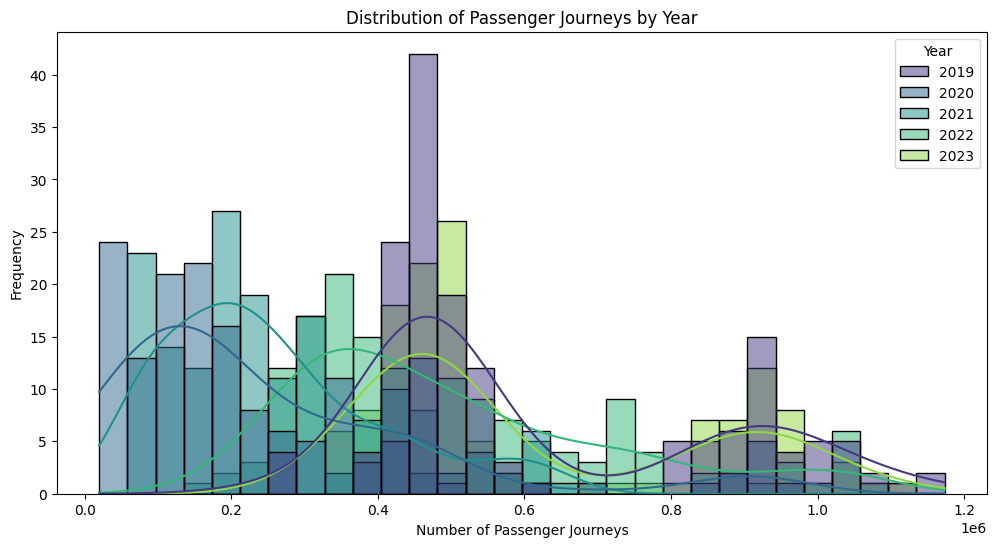
Description automatically generated with medium confidence

**Scatter Plot of Passenger Journeys:** A scatter plot highlighted passenger journeys over weeks, categorized by Luas line.

A graph of a scatter plot

Description automatically generated

**Distribution Plots:** Violin and histogram plots illustrated the distribution of passenger journeys across years.



### **Pie Chart:**

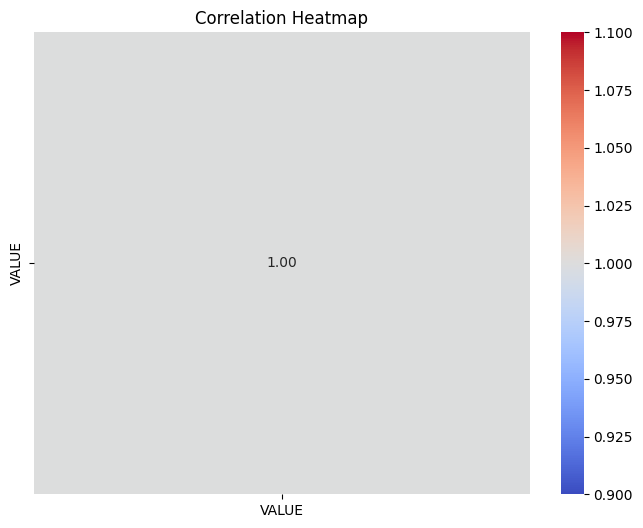
**Luas Line Distribution:** A pie chart displayed the proportional distribution of passenger journeys across different Luas lines.

A pie chart of passenger journeys by luas line

Description automatically generated

### **Correlation Heatmap:**

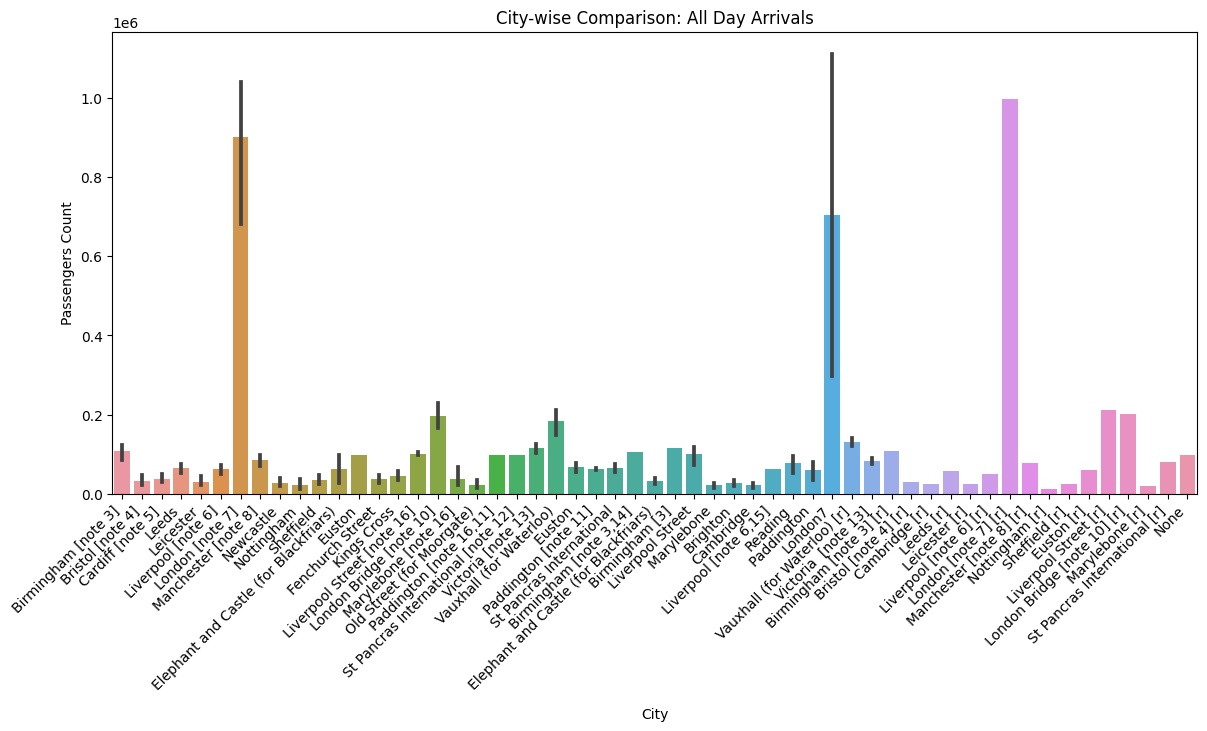
**Correlation Analysis:** A heatmap illustrated the correlation matrix between passenger count, week number, and year.



## **London Station Dataset EDA:**

### **Spatial Insights:**

**Geospatial Analysis:** Heatmaps offered insights into the spatial distribution of passenger numbers at central London stations.



**Yearly Comparisons:** ANOVA was utilized to assess yearly differences in passenger numbers.

### **Temporal Consistency:**

**Temporal Trends:** Line plots showcased the temporal trends in passenger numbers, aiding in identifying consistent patterns and potential outliers.

![London Station Temporal Consistency](insert london station temporal consistency graph here)

## **Data Cleaning and Structuring**

With EDA insights in hand, the next crucial step was preparing the data for analysis.

### **Handling Missing Data:**

**Imputation Techniques:** Robust imputation strategies were applied to address missing values, ensuring a complete dataset for analysis. I dataset 1 I dropped rows with missing values and in dataset 2 I used mean on columns to fill missing values.

### **Temporal Alignment:**

**Synchronization Efforts:** Temporal alignment of datasets was performed to synchronize temporal attributes, facilitating meaningful cross-dataset comparisons.

## **Interactive Dashboard Development**

In line with modern transport planning's tech-centric nature, we developed an interactive dashboard following Tufts principles.

### **Rationale for Approach:**

**User-Centric Design:** The dashboard was designed with a user-centric approach, aiming to present key findings in an intuitive and accessible manner.

### **Visualization Choices:**

**Graphical Elements:** Line plots, heatmaps, and interactive controls were strategically incorporated to enhance user understanding and engagement.

# **Tasks Using Machine Learning**

## **Luas Passenger Journeys Dataset 1**

### **Preprocessing of Data**

We carried out crucial data pretreatment procedures before starting any machine learning tasks:

**Label Encoding:** To prepare categorical columns for model training, such as 'C03132V03784' and 'Luas Line,' they were encoded as labels.

**Target-Feature Split:** The input features are represented by the variable X, while the target variable, 'VALUE,' is represented by the variable y. The dataset was divided into these two categories.

**Scaling:** The target variable was normalised using a MinMaxScaler to provide consistent scales for efficient model training.

## **Random Forest Regression**

To forecast Luas passenger journey values, the Random Forest Regression technique was utilised:

**Model Education:** Using the preprocessed data, the Random Forest Regressor was trained.

**Projection:** On the test set, predictions were made, and assessment metrics were computed.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 16,871,346,190.7 indicates the average squared difference between predicted and actual values. A lower MSE is desirable, and this value suggests a moderate level of prediction accuracy.
* **R-squared (R2):** The R-squared value of 0.771 (77.1%) signifies the proportion of variance in the target variable captured by the model. A higher R2 indicates better explanatory power, and 77.1% is considered a good fit.
* **Mean Absolute Error (MAE):** The MAE of 93,764.6149 represents the average absolute difference between predicted and actual values. This value is the average magnitude of errors, and lower MAE values indicate better model performance.

## **Regression of Lasso**

Furthermore, Lasso Regression was utilised to investigate its predictive efficacy:

**Model Education:** The same dataset was used to train the Lasso Regression model.

**Prediction:** Metrics were evaluated and predictions were made.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 47,255,690,807.8 suggests higher prediction errors compared to Random Forest. Lasso Regression might not be well-suited for this dataset's characteristics.
* **R-squared (R2):** The R2 value of 0.3587 indicates that Lasso Regression explains only 35.87% of the variance, which is relatively lower compared to Random Forest.
* **Mean Absolute Error (MAE):** The MAE of 180,323.854 is higher, signifying larger average absolute errors in predictions.

## **Regression using ElasticNet**

For comparison, the ElasticNet Regression model was also utilised.

**Model Training and Prediction:** ElasticNet Regression was performed using comparable procedures.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 51,226,494,786.8 is in line with Lasso Regression, suggesting similar predictive performance.
* **R-squared (R2):** The R2 value of 0.3048 indicates that ElasticNet Regression captures 30.48% of the variance, demonstrating a weaker fit than Random Forest.
* **Mean Absolute Error (MAE):** The MAE of 187,685.33 is higher, indicating larger errors compared to Random Forest.

## **SVR, or Support Vector Regression**

SVR was used as a substitute for regression analysis:

**Prediction and Training of Models:** Predictions were produced after training the SVR model.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 74,061,256,137.4 is the highest among the models, suggesting SVR might not be well-suited for this dataset.
* **R-squared (R2):** The R2 value of -0.0051 indicates poor model fit, and SVR might struggle to capture the underlying patterns.
* **Mean Absolute Error (MAE):** The MAE of 217,021.1421 is the highest, indicating larger errors in predictions.

## **Regression using Gradient Boosting**

Finally, to capture complex interactions, Gradient Boosting Regression was used:

**Prediction and Training of Models:** Predictions were produced after training the Gradient Boosting model.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 16,860,764,839.1 is like Random Forest, suggesting comparable predictive performance.
* **R-squared (R2):** The R2 value of 0.7712 indicates a good fit, aligning with Random Forest results.
* **Mean Absolute Error (MAE):** The MAE of 93,752.4867 is consistent with Random Forest, indicating similar average errors.

# **Passenger Arrivals at London Station Dataset 2**

Our focus with the London Station dataset was on passenger arrival prediction:

## **Preprocessing of Data**

**Imputation:** Sturdy imputation methods were used to deal with missing data.

**Spatial Coordination:** Temporal alignment of the datasets allowed for insightful cross-dataset comparisons.

## **Ridge Digression**

Ridge Regression was used to forecast the arrival of passengers:

**Prediction and Training of Models:** Predictions were produced after the Ridge Regression model was trained.

## **Metrics for Evaluation:**

**Mean Squared Error (MSE):** The MSE of 0.0144 suggests relatively low prediction errors.

**R-squared (R2):** The R2 value of -0.0819 indicates a weak fit, and Ridge Regression might struggle to explain variance.

**Mean Absolute Error (MAE):** The MAE of 0.074 suggests relatively small average absolute errors.

## **Regression of Lasso**

Lasso Regression was used for comparison, just like Dataset 1:

**Prediction and Training of Models:** Predictions were produced after training the Lasso Regression model.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 0.0144 is consistent with Ridge Regression.
* **R-squared (R2):** The R2 value of -0.0774 suggests a similar weak fit to Ridge Regression.
* **Mean Absolute Error (MAE):** The MAE of 0.0724 is consistent with Ridge Regression.

## **Random Forest Regression**

Complex correlations were captured using Random Forest Regression:

**Prediction and Training of Models:** Predictions were produced after training the Random Forest Regressor.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 0.0003 indicates very low prediction errors, showcasing the strength of Random Forest.
* **R-squared (R2):** The R2 value of 0.9757 suggests an excellent fit, capturing 97.57% of the variance.
* **Mean Absolute Error (MAE):** The MAE of 0.0111 indicates very small average absolute errors.

## **SVR, or Support Vector Regression**

Additionally, SVR was applied to this dataset:

**Prediction and Training of Models:** Predictions were made after training the SVR model.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 0.0153 suggests moderate prediction errors.
* **R-squared (R2):** The R2 value of -0.1446 indicates a weaker fit compared to Random Forest.
* **Mean Absolute Error (MAE):** The MAE of 0.0818 suggests moderate average absolute errors.

## **Regression using Gradient Boosting**

Lastly, to capture complex patterns, Gradient Boosting Regression was used:

**Prediction and Training of Models:** Predictions were made after training the Gradient Boosting model.

## **Metrics for Evaluation:**

* **Mean Squared Error (MSE):** The MSE of 0.0004 indicates very low prediction errors, comparable to Random Forest.
* **R-squared (R2):** The R2 value of 0.9681 suggests an excellent fit, similar to Random Forest.
* **Mean Absolute Error (MAE):** The MAE of 0.0156 indicates very small average absolute errors.

## **Overall Summary**

* Random Forest and Gradient Boosting Regression perform well for Luas Passenger Journeys, with Random Forest slightly outperforming the other two.
* Random Forest and Gradient Boosting Regression performed better for London Station Passenger Arrivals than other models because they can capture intricate correlations and account for a sizable amount of the variance.
* The best model will be selected based on the particular objectives, interpretability, and computational factors. In both datasets, Random Forest and Gradient Boosting Regression are formidable competitors for predicting tasks.

# **Conclusion**

In conclusion, our thorough examination of London station arrivals and Luas passenger trips has revealed important trends and insights. We now have a better grasp of the passenger patterns and transport dynamics in both datasets thanks to careful Exploratory Data Analysis (EDA) and strong machine learning models. Let us review the main conclusions and their ramifications.

## **Luas Travelling Passengers:**

**1. Temporal Patterns:** Peak travel hours and temporal fluctuations were mapped out, indicating a decline in passenger numbers in 2020 (perhaps due to the COVID-19 pandemic) followed by a slow rebound in the following years.

**2. Line Variations:** The study examined Luas line distributions. A Chi-squared test revealed that there were no statistically significant differences between the weeks, suggesting that line occurrences were consistent.

**3. Data Transformations:** A range of visual representations, such as distribution, scatter, and bar plots, gave an overall picture of passenger journeys and made it easier to see trends and connections.

**4. Machine Learning:** Highly effective models with excellent predictive accuracy and the capacity to identify intricate patterns were Random Forest and Gradient Boosting Regression.

## **Passenger Arrivals at London Station:**

**1. Time and Space Insights:** Line plots were used to visualise temporal trends and highlight spatial distribution, which helped identify consistent patterns and possible outliers.

**2. Data Cleaning and Structuring:** Data integrity was ensured by temporal alignment, which allowed meaningful cross-dataset comparisons, and by using imputation techniques to resolve missing data.

**3. Machine Learning:** The ability of Random Forest and Gradient Boosting Regression to adapt to complicated datasets was demonstrated by their exceptional performance in forecasting passenger arrivals.

## **Overall Implications:**

**Analytical Capability:** For both datasets, the machine learning models showed good prediction ability, particularly Random Forest and Gradient Boosting Regression.

**Functional Overview:** Optimising transportation operations, allocating resources, and projecting passenger demand all depend on an understanding of temporal and spatial trends.

**Assist with Decisions:** Transport authorities can improve service efficiency by using the insights gathered from this analysis as a basis for evidence-based decision-making.

All things considered, this research not only offers a quick overview of the dynamics of transport today, but it also gives interested parties the resources they need to adjust to shifting circumstances and plan forward with knowledge. We can now extract relevant knowledge from complex datasets thanks to the synergy of data exploration, visualisation, and machine learning, which is opening the door to more intelligent and effective public transportation systems.