**CA1 - MSc in Data Analytics**

**Tourism Arrival numbers to Ireland Analytics Study**

Author: Sallam Salman

e-mail: [sbs24003@student.cct.ie](mailto:sbs24003@student.cct.ie)  
Student ID: sbs24003

Github Link: https://github.com/CCT-Dublin/ca1-Sallam-Noor

Table of Contents

[Abstract 3](#_Toc163412855)

[1 Introduction 3](#_Toc163412856)

[1.1Datasets 3](#_Toc163412857)

[1.2 Methodology 3](#_Toc163412858)

[1.3 Report Structure 4](#_Toc163412859)

[2 Data Preparation and Visualization 5](#_Toc163412860)

[2.1 ASM03: International Travel (Arrivals, Departures) To Ireland Counts 5](#_Toc163412861)

[2.2 ASM01: Modes of Travel (Arrivals, Departures) To Ireland Counts 7](#_Toc163412862)

[2.3 ASM02: Route of Travel (Arrivals, Departures) To Ireland Counts 8](#_Toc163412863)

[2.4 mly532: Weather Data Dublin Airport Station 9](#_Toc163412864)

[2.5 The Combined Dataset 9](#_Toc163412865)

[2.5 Additional Data Preparation 10](#_Toc163412866)

[3 Statistic Distributions (Binomial / Poisson Case Studies) and Normal Distribution 16](#_Toc163412867)

[3.1 Binomial Distribution Case Study 16](#_Toc163412868)

[3.2 Poisson Distribution Case Study 17](#_Toc163412869)

[3.3 Normal Distribution 18](#_Toc163412870)

[4 Machine Learning 20](#_Toc163412871)

[4.1 Random Forest Regression (RFR) 20](#_Toc163412872)

[4.2 Support Vector Regressor (SVR) 21](#_Toc163412873)

[4.3 Ridge Regressor 22](#_Toc163412874)

[5 Programming 23](#_Toc163412875)

[6 Results and Discussion 23](#_Toc163412876)

[6.1 Results 23](#_Toc163412877)

[6.2 Discussion 24](#_Toc163412878)

[7 Conclusion 25](#_Toc163412879)

Tables

[Table 1:Data Type 5](#_Toc163412880)

[Table 2: Numerical Column Descriptive Statistics 5](#_Toc163412881)

[Table 3: Categorical Columns Descriptive Statistics 5](#_Toc163412882)

[Table 4: Null Values 6](#_Toc163412883)

[Table 5: Data Type 7](#_Toc163412884)

[Table 6: Numerical Column Descriptive Statistics 7](#_Toc163412885)

[Table 7: Categorical Columns Descriptive Statistics 7](#_Toc163412886)

[Table 8: The Sum of the Null values 7](#_Toc163412887)

[Table 9: Data Type 8](#_Toc163412888)

[Table 10: Numerical Column Descriptive Statistics 8](#_Toc163412889)

[Table 11: Categorical Columns Descriptive Statistics 8](#_Toc163412890)

[Table 12: Weather Data Detailed Description 9](#_Toc163412891)

[Table 13 The Combined Dataset Some Features 10](#_Toc163412892)

[Table 14: comparison between ML models performance 24](#_Toc163412893)

Figures

[Figure 1:Other UK (Arrival & Departures) Counts During Covid-19 period. With the Missing Value 6](#_Toc163412894)

[Figure 2: Travel Data Pattern 11](#_Toc163412895)

[Figure 3: Distributions plots and the Box plots for the combined dataset 14](#_Toc163412896)

[Figure 4: The heatmap 15](#_Toc163412897)

[Figure 5: Binomial Distribution for a Sample of 100 Air Mode Trips 16](#_Toc163412898)

[Figure 6: Probability of exactly 90 Travellers by Air 17](#_Toc163412899)

[Figure 7: Probability of more than 10 Travellers by Sea 17](#_Toc163412900)

[Figure 8: Poisson Distribution 17](#_Toc163412901)

[Figure 9:Probability of Cross Chanel Trips between 380 and 420 18](#_Toc163412902)

[Figure 10: Distribution of Sample Means of Total Arrivals 19](#_Toc163412903)

[Figure 11: RFR Feature Importance Score 20](#_Toc163412904)

[Figure 12: SVR Feature Importance Score 21](#_Toc163412905)

[Figure 13: Ridge Coefficients 22](#_Toc163412906)

## Abstract

*Recently, Machine Learning (ML) has been used for widespread application in problem-solving across various domains. In this study, three different machine learning algorithms were analysed to develop a model that can accurately predict the arrivals count into Ireland, with the conclusion that Ridge Regression performed the best particularly when hyperparameter tuning was applied. Performance metrics such as the test R-squared, Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) were computed for the test set, revealing a remarkably high level of performance.*

## 1 Introduction

"Data will talk to you if you're willing to listen." - Jim Bergeson

The main purpose of this report is to specify valuable datasets related to tourism in Ireland to conduct a statistical analysis, data preparation, and visualization, while also implementing machine learning models that serve a certain goal all using Python programming.

In this study the counts of arrivals into Ireland, the travel means counts - by sea and by air- and the travel routes counts, additionally, weather features such as temperature, rainfall, wind speed, and sunshine duration play a crucial role in this analysis. All these features were combined based on monthly data from January 2010 to January 2024.

Supervised learning proved to be a valuable method for examining traveller counts into Ireland through machine learning models. Utilizing historical data, three distinct machine learning algorithms were employed, focuses on the factors influencing travel behaviour and enabling the forecasting of future trends. Moreover, supervised learning assisted in identifying trends and patterns in traveller counts over time, thus assisting in decision-making and strategic planning.

### 1.1Datasets

The arrival dataset formed from 15 features and include 169 observations. this dataset was built up from 4 different datasets, described below.

* ASM03: International Travel (Arrivals, Departures) To Ireland Counts
* ASM01: Modes of Travel (Arrivals, Departures) To Ireland Counts
* ASM02: Route of Travel (Arrivals, Departures) To Ireland Counts
* mly532: Weather Data Dublin Airport Station

### 1.2 Methodology

In this study, the Knowledge Discovery in Databases (KDD) framework was employed to guide the methodology. Based on Osei-Bryson and Barclay (2015, p12) that fayyad et al.'s(1996), the nine-step process modelling approach outlines a systematic method for enhancing organizational performance through knowledge discovery. This process involves a series of steps aimed at uncovering valuable insights from data to consolidating discovered knowledge. The nine steps typically include:

1. Develop and understand the application domain.

2. Creating Target dataset.

3. Data cleaning and preprocessing

4. Data reduction and projection

5. Choosing the task

6. Choosing the Algorithm

7. Data Modelling

8. Interpretation and Evaluation

9. Consolidating discovered knowledge

This nine-step process modelling approach provides a structured framework to leverage data-driven insights effectively, ultimately enhancing their performance and competitiveness.

The first step was to explore different aspects of tourism in Ireland, then data observation and collection took a place, the data scientist focus was on analysing the arrivals trips to Ireland. Through the analysis of three datasets covering traveller counts from various countries, mode of travel counts, and counts of travellers by different routes. These datasets served as the foundation for understanding the patterns and dynamics of tourism activity. Additional dataset (weather data) was added to provide better insight into the patterns. Descriptive statistics were then applied to comprehensively grasp the datasets. Data preparation involved a series of steps, including cleaning, constructing, joining, reduction, and preparing the datasets for analysis. Exploratory Data Analysis (EDA) techniques were utilized to find deeper insights before selecting suitable machine learning algorithms for modelling based on the data type and distribution. Evaluation and cross-validation of these models were conducted to ensure their effectiveness in capturing the complexities of tourism arrival counts to Ireland. The study aimed to uncover actionable insights to inform decision-making and strategic planning in the tourism sector.

### 1.3 Report Structure

In section 2 a detailed explanatory data analysis was covered, it also focused on data preparation and visualization techniques applied to the dataset. While in section 3 a statistic description of some case study was done. In section 4, various machine learning models utilized in the study were discussed. Section 5 covers the programming aspects of the study. Section 6 presents the discussion and results obtained from the analysis. Finally, the conclusion summarized the findings and insights found from the study.

## 2 Data Preparation and Visualization

In this section, each of the datasets were described, manipulated, and visualized. The primary objective of this step was to prepare the datasets for joining with others containing the key features and observations essential for this analysis.

### 2.1 ASM03: International Travel (Arrivals, Departures) To Ireland Counts

#### 2.1.1 Exploratory Data Analysis (EDA)

This dataset consists of 6084 entries rows with a total of 6 columns, Table 1 show the specific types of this dataset:

|  |  |
| --- | --- |
| **Column** | **Type** |
| Statistics Label | Categorical |
| Month | Categorical |
| Country | Categorical |
| Direction | Categorical |
| UNIT | Categorical |
| VALUE | Numerical (float64) |

Table 1:Data Type

Descriptive Statistic is for the numerical column (VALUE column), and the categorical are show below in table 2 and table 3 respectively:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **VALUE** | | | | | | | |
| **Count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| 6080.00 | 136.604638 | 316.126348 | 0.00 | 12.80 | 36.10 | 85.40 | 2270.70 |

Table 2: Numerical Column Descriptive Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **STATISTIC Label** | **Month** | **Country** | **Direction** | **UNIT** |
| **count** | 6084 | 6084 | 6084 | 6084 | 6084 |
| **unique** | 1 | 169 | 18 | 2 | 1 |
| **top** | Air and Sea Travel | 2010 January | Great Britain | Arrivals | Thousand |
| **freq** | 6084 | 36 | 338 | 3042 | 6084 |

Table 3: Categorical Columns Descriptive Statistics

A discrepancy was noted in the observation count between the numerical "VALUE" column and the categorical columns, indicating missing values. Looking at the standard deviation value, it is noted that the data shows a high level of variability within the dataset. Additionally, the columns labelled "STATISTIC" and "UNIT" contain only one observation each, limiting their informative value.

#### 2.1.2 Dataset Manipulation and Cleaning

Various operations have been undertaken to prepare this dataset. Firstly, a combined column of country name and direction (arrival or departure) was added to facilitate table pivoting. This step provides unique information, which would serve as the new column names (features) in the pivoted table. This approach enabled the isolation of the desired feature for analysis. Therefore, table pivoting was performed using the 'Month' date as the index. The 'Month' date column was then reshaped from yyyy MonthName (%Y %B) to yyyy-mm (%Y-%m) format and sorted in ascending order to start from 2010-01 to 2024-01, establishing the foundation column for dataset collection.

Certain values were found to be missing from the dataset as shown in table 4. In order to gain clarity on these missing values, particularly those occurring during the COVID-19 lockdown period, this period was isolated and plotted to obtain a general understanding of the pattern, as shown in figure 1 below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Date** | **…** | **Other UK (1)\_Arrivals** | **Other UK (1)\_Departures** | **…** |
| **132** | 2021-01 | … | NaN | NaN | … |
| **141** | 2021-10 | … | NaN | NaN | … |

Table 4: Null Values

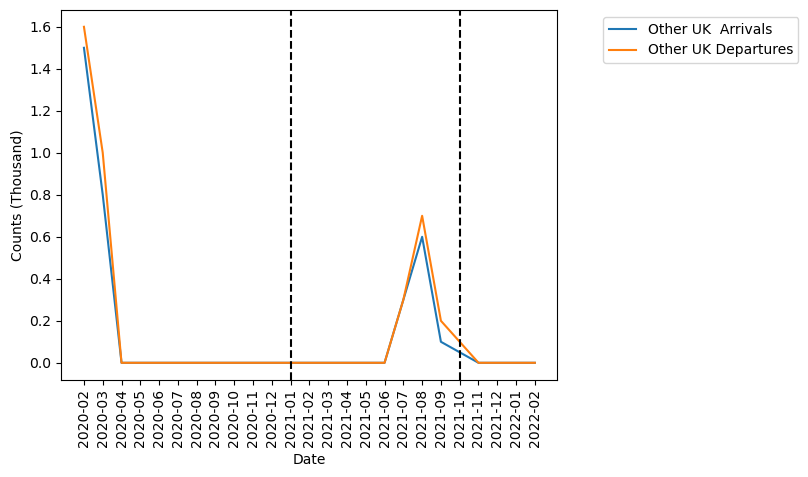


Figure 1:Other UK (Arrival & Departures) Counts During Covid-19 period. With the Missing Value

The interpolate method was used to deal with these missing values to ensure continuity and preserve the integrity of the data. This approach allows to estimation the missing values based on the observed data points around the specific missing value , thereby maintaining the temporal and spatial relationships within the dataset. Finally, the unnecessary columns were dropped, and the datasets were tidied up.

### 2.2 ASM01: Modes of Travel (Arrivals, Departures) To Ireland Counts

#### 2.2.1 Exploratory Data Analysis (EDA)

This dataset consists of 1014 entries rows with a total of 6 columns, table 5 show the specific types of this dataset:

|  |  |
| --- | --- |
| **Column** | **Type** |
| Statistics Label | Categorical |
| Month | Categorical |
| Direction | Categorical |
| Mode | Categorical |
| UNIT | Categorical |
| VALUE | Numerical (float64) |

Table 5: Data Type

Descriptive Statistic is for the numerical column (VALUE column), and the categorical are show below in table 6 and table 7 respectively:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **VALUE** | | | | | | | |
| **Count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| 1014 | 819.083531 | 653.410945 | 4 | 109.6 | 870.3 | 1323.825 | 2270.7 |

Table 6: Numerical Column Descriptive Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **STATISTIC Label** | **Month** | **Direction** | **Mode** | **UNIT** |
| **count** | 1014 | 1014 | 1014 | 1014 | 1014 |
| **unique** | 1 | 169 | 2 | 3 | 1 |
| **top** | Air and Sea Travel | 2010 January | Arrivals | All modes of transport | Thousand |
| **freq** | 1014 | 6 | 507 | 338 | 1014 |

Table 7: Categorical Columns Descriptive Statistics

From the tables above, the columns labelled "STATISTIC" and "UNIT" contain only one observation each, these can be omitted later in the dataset joining stage.

#### 2.2.2 Dataset Manipulation and Cleaning

Various operations have been undertaken to prepare this dataset. Firstly, the data set was checked for null values as shown in table 8 below no null value was found.

A white background with black text

Description automatically generated

Table 8: The Sum of the Null values

Similar steps were incorporated as previous pivoted dataset above, all was detailed in jupyter notebook cell numbers (32-37).

### 2.3 ASM02: Route of Travel (Arrivals, Departures) To Ireland Counts

#### 2.3.1 Exploratory Data Analysis (EDA)

This dataset consists of 1014 entries rows with a total of 6 columns, table 9 show the specific types of this dataset:

|  |  |
| --- | --- |
| **Column** | **Type** |
| Statistics Label | Categorical |
| Month | Categorical |
| Direction | Categorical |
| Route | Categorical |
| UNIT | Categorical |
| VALUE | Numerical (float64) |

Table 9: Data Type

Descriptive Statistic is for the numerical column (VALUE column), and the categorical are show below in table 10 and table 11 respectively:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **VALUE** | | | | | | | |
| **Count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| 1690 | 491.449941 | 505.845834 | 0.5 | 59.625 | 387.1 | 719.675 | 2270.7 |

Table 10: Numerical Column Descriptive Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **STATISTIC Label** | **Month** | **Direction** | **Route** | **UNIT** |
| **count** | 1690 | 1690 | 1690 | 1690 | 1690 |
| **unique** | 1 | 169 | 2 | 5 | 1 |
| **top** | Air and Sea Travel | 2010 January | Arrivals | All routes of travel | Thousand |
| **freq** | 1690 | 10 | 845 | 338 | 1690 |

Table 11: Categorical Columns Descriptive Statistics

The columns labelled "STATISTIC" and "UNIT" contain only one observation each, these can be omitted later in the dataset joining stage.

#### 2.3.2 Dataset Manipulation and Cleaning

The same prosses has been followed as the previous datasets. Firstly, the data set was check for null values as shown jupyter notebook cell (42). Then a combined column of route of transport and direction was added, table pivoting was performed using the 'Month' date as the index. Finally, dataset was tidied up. All was detailed in cells number (42-47).

### 2.4 mly532: Weather Data Dublin Airport Station

#### 2.4.1 Description and Manipulation

This dataset consists of 988 entries rows with a total of 12 columns, the focus was to involve additional data into the analysis to help in identify the pattern and get better prediction results. Some weather data such as mean temp, mean wind speed, rain amount perception and sunshine duration, these features were detailed in the table 12.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mean Air**  **Temperature(C)** | **Precipitation**  **Amount(mm)** | **Mean Wind**  **Speed(knot)** | **Sunshine**  **duration(hours)** |
| **count** | 169 | 169 | 169 | 169 |
| **mean** | 9.598225 | 64.481065 | 10.095858 | 125.449112 |
| **std** | 3.897503 | 35.847649 | 1.921878 | 53.310163 |
| **min** | -0.1 | 4.8 | 6.5 | 30.5 |
| **25%** | 6.2 | 39.7 | 8.8 | 79.7 |
| **50%** | 9.1 | 57.7 | 9.6 | 121 |
| **75%** | 13.1 | 83.5 | 11 | 159.8 |
| **max** | 16.7 | 193.5 | 16.3 | 295 |

Table 12: Weather Data Detailed Description

From the tables above, the mean of (mean air Temperature) is slightly more than the median, the distribution of this data to be positively skewed, the standard deviation value suggesting less variability and greater consistency.

This dataset was checked for Null values, non was detected. The weather data was sorted monthly so minimum amount of cleaning and manipulation was required.

### 2.5 The Combined Dataset

All the above datasets were merged using inner join, this function keeps the rows of a specific column indices (Date) exist in all the datasets. Unmatched rows from any dataset were discarded. This process was illustrated in jupyter notebook cells (57-66)

#### 2.5.1 Exploratory Data Analysis (EDA)

This dataset consists of 169 observations with a total of 15 columns, table 13 some of these features statistics description. All of the features are numerical.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Statistics** | **Total Arrivals** | **Total Departure** | **Air Arrivals** | **Sea**  **Arrivals** | **Continental Eur Rute** | **Cross channel Rute** | **Other Rute** |
| **count** | 169 | 169 | 169 | 169 | 169 | 169 | 169 |
| **mean** | 1227.49704 | 1229.7574 | 1129.8355 | 97.657988 | 577.139645 | 503.81716 | 36.406509 |
| **std** | 518.133699 | 516.4088 | 474.693749 | 63.111201 | 267.351749 | 188.35303 | 18.228572 |
| **min** | 16.1 | 17.8 | 12.1 | 4 | 3.5 | 9.3 | 2 |
| **25%** | 911.4 | 895.1 | 846 | 48.9 | 389.2 | 455.1 | 23.7 |
| **50%** | 1236.6 | 1256.4 | 1171.7 | 87.6 | 571.1 | 531.3 | 37.7 |
| **75%** | 1576 | 1577.6 | 1450.3 | 122.5 | 759 | 621.9 | 47.8 |
| **max** | 2270.7 | 2194.3 | 2070.2 | 279.5 | 1168.1 | 814.8 | 85.4 |

Table 13 The Combined Dataset Some Features

The mean of total arrival is approximately close to the median, indicating a relatively balanced distribution. However, the substantial standard deviation suggests significant variability within the dataset.

### 2.5 Additional Data Preparation

#### 2.5.1 Visualization

The Tufts Principles were considered in the design of data visualization within the report. By prioritizing clarity, accuracy, and relevance, these principles ensured effective visualizations, communicated insights and facilitate understanding for the audience.

As shown in the figure 2 below, the travel data follows a trend of seasonality with growth. It shows a continuous consistent pattern until restrictions were enforced, resulting in a unique pattern that subsequently recovered and resumed following the initial trend.

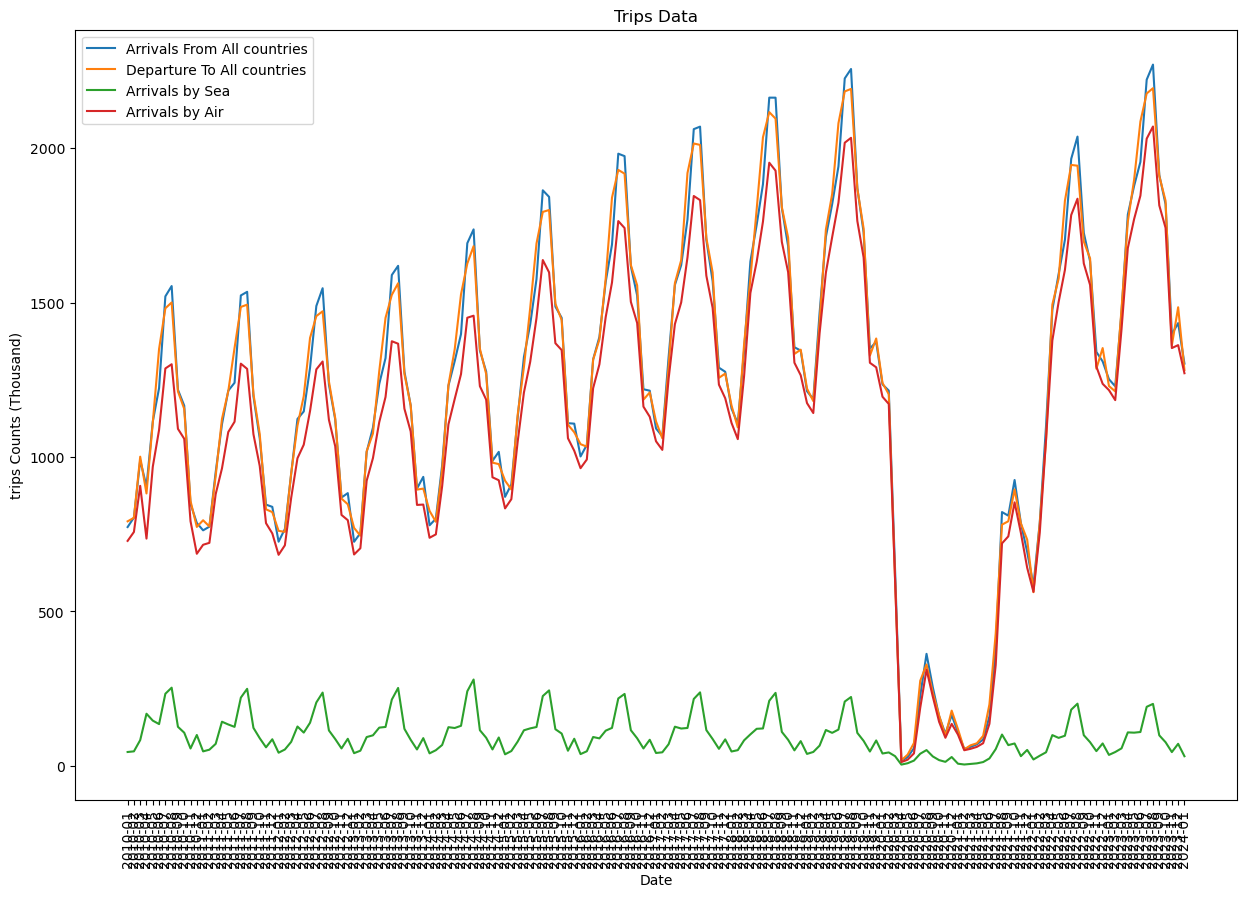


Figure 2: Travel Data Pattern

The unique pattern can be justified by the scientific article by Ilin et al. (2021) suggests that lockdown policies by the decision-makers and governmental restrictions led to substantial decreases in human mobility, which helped to reduce the spread of COVID-19 infections globally. Also, according to the article conducted by Ito et al. (2022), human behaviour and decision-making were affected to minimize the risk of contracting illnesses.

The distributions plots were used to visually understand the values frequency for each feature in the dataset. Additionally, box plots were employed to visualize the statistical information for each feature. The box plot is important to identify whether there are outliers in the feature values. Distribution plots and box plots for the combined dataset features are shown in figure 3.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Figure 3: Distributions plots and the Box plots for the combined dataset

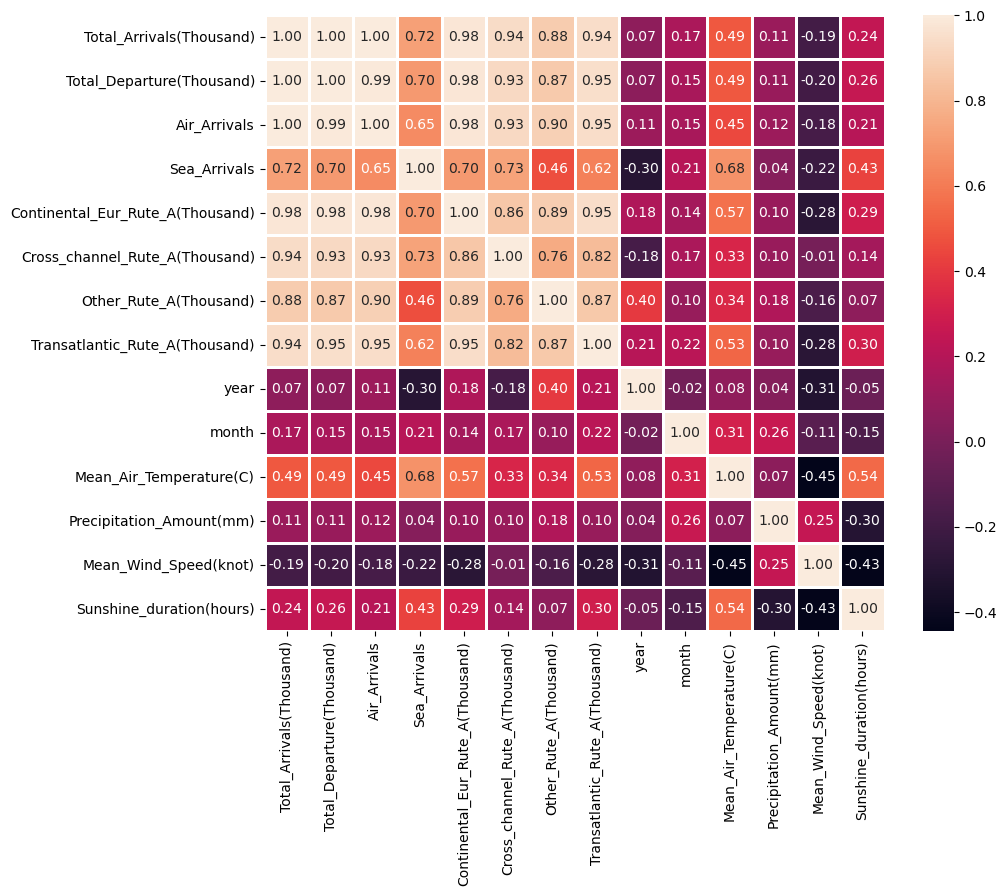


Figure 4: The heatmap

The heatmap (figure 4) shows the correlation of the arrival trips to Ireland (Total arrivals (Thousands)) with independent variables:

The total arrivals are highly (positively) correlated with the total departures, temperature, travel routes and travel means (by sea or Air). It also negatively correlated with wind speed.

Correlation among independent variables: Many independent variables are highly correlated (At the top-left part of matrix).

#### 2.5.2 Scaling

The dataset was scaled in later stages to be analysed using machine learning; due to the variety of numerical features, such as arrival trip counts in thousands and mean temperatures less than 17 degrees Celsius. The scaler was chosen based on analyzing the distribution plots and box plots, which revealed the presence of outliers and non-normally distributed data in the dataset due to the impact of the pandemic on traveler arrivals into Ireland and weather data spikes. It is recommended to use the robust scaler instead of the standard scaler. Robust scaler is less affected by outliers and is designed to handle such data more effectively. By employing a robust scaler, the data can be preprocessed to overcome the effects of outliers and non-normal distributions, making it more suitable for machine learning algorithms such as Support Vector Machines (SVM) and Ridge Regression.

## 3 Statistic Distributions (Binomial / Poisson Case Studies) and Normal Distribution

Two discrete distributions (Binomial and/or Poisson) were applied for two case studies related to the dataset (arrivals\_df) in order to identify some information about the dataset. In the first case study, the Binomial distribution was explored, while in the second the Poisson distribution was examined. Another subsection was dedicated for the Normal distribution.

### 3.1 Binomial Distribution Case Study

In this case, a sample of one hundred individuals from the population was utilized to analyse the count of people arriving via air travel mode. This type of distribution required three pieces of information: the probability, the sample size, and the number of trials.

The probability was calculated by

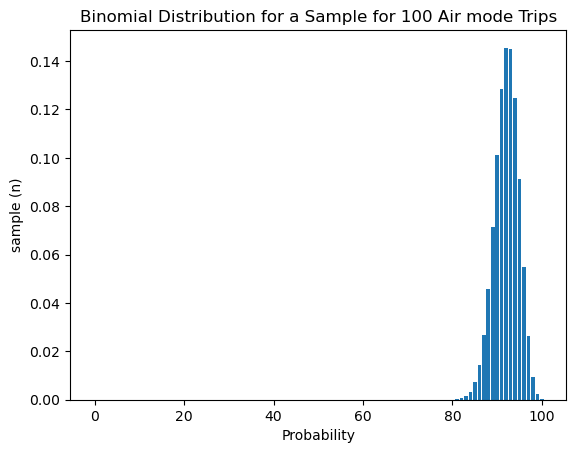


Figure 5: Binomial Distribution for a Sample of 100 Air Mode Trips

While the sample size was already defined. The Binomial distribution was applied to calculate the probability of exactly 90 travellers arriving by air.

The probability of exactly 90 travellers arriving by air = 0.10112642.

The probability of having more than 10 arrivals via sea to Ireland within the 100 samples was also calculated by first determining the probability of travel via sea.

Applying the Binomial distribution equation.

Then the probability of more than 10 travellers out of 100 came by sea= 0.17130389.

Below, visualization figures illustrate the results of the Binomial distribution calculation above.

|  |  |
| --- | --- |
|  |  |
| Figure 6: Probability of exactly 90 Travellers by Air | Figure 7: Probability of more than 10 Travellers by Sea |

### 3.2 Poisson Distribution Case Study

In this case, the focus was on the route type. The average number of arrivals by the cross-channel route from January 2010 to January 2020 was calculated, lambda determined to be 417. The Poisson disruption was plotted in figure 8.

A blue graph with white text

Description automatically generated

Figure 8: Poisson Distribution

The probability of having between 380 and 420 arrivals into Ireland via the cross-channel route was calculated using the Poisson distribution equation.

A graph of a distribution of a number

Description automatically generated with medium confidence

Figure 9:Probability of Cross Chanel Trips between 380 and 420

### 3.3 Normal Distribution

The central limit theorem is critical concept in statistics, states that the distribution of the statistical mean within a sample from a population tends towards a normal distribution with increasing sample size. Essentially, even if data obtained from independent random samples initially appears skewed, maximizing the sample size eventually leads to a reflection of the probable statistical average. This theorem forms a fundamental basis of statistics, enabling researchers to draw conclusions about entire populations by analysing data from smaller sample sizes. (Sheposh, 2023)

It works for populations that follow a binomial distribution, as long as a certain condition is met regarding the sample size and probability of success. This theorem helps to use the normal probability model to make predictions about population averages based on sample averages. (LibreTexts, 2014)

In Section 2.5.1, it was observed that although most of the features showed non-normal distribution graphs, a similarity to a normal distribution was noticed in the distribution of Total\_Arrivals(Thousand) feature. Therefore, multiball samples were applied to this feature to demonstrate the central limit theorem.

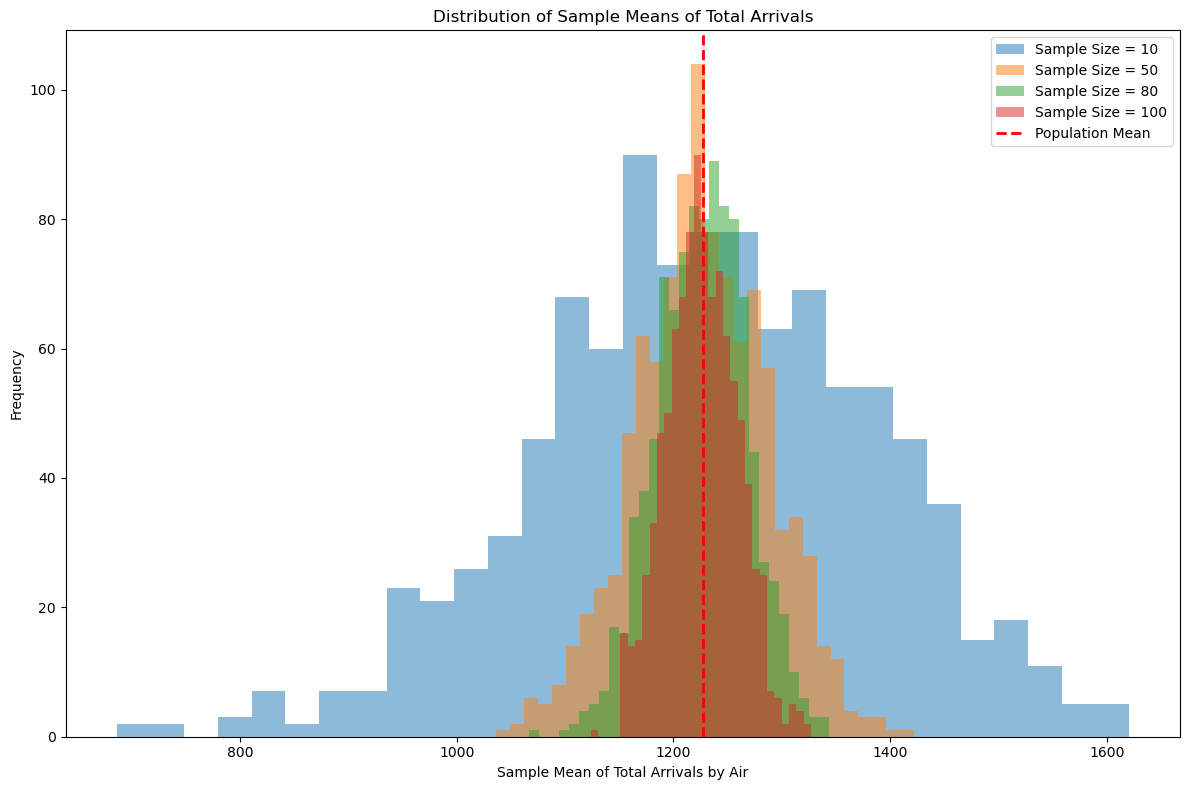


Figure 10: Distribution of Sample Means of Total Arrivals

It is important to mention that the study of the statistical characteristics and distribution plays a crucial role in data analysis. Through statistical observation, it will ease the process of choosing suitable machine learning algorithms that would work effectively with the data type and distribution pattern.

## 4 Machine Learning

The supervised machine learning was used for predicting arrival numbers of people traveling into Ireland, based on the data distribution and outliers, a robust regression algorithm was employed. The data scientist intention was to use the following algorithms:

* Random Forest Regression
* Support Vector Machine Regressor
* Ridge Regressor and
* Time Series\*\*

Total\_Arrivals (Thousand) was specified as an independent feature; the dataset was split into the 80% training data 20% testing data with random state of 42.

### 4.1 Random Forest Regression (RFR)

This model is robust to outliers due to its ensemble nature. C. Müller and Guido (2017, p. 83) mention that Random Forest overcomes the drawback of decision trees by building multiple random subsets of the data and combining their predictions, which reduces the impact of outliers on the overall prediction. This method does not assume normality in the data distribution and can handle non-linear relationships between features and the target variable without requiring data transformation.

The data was trained using the basic RFR with the parameters of 200 trees with random state of 42.

The feature importance was analysed for RFR model, this helps to choose which features to include. Rremoving less important features can make the model simpler, easier to understand, and less resource-intensive to train and use.

A graph with text and numbers

Description automatically generated with medium confidence

Figure 11: RFR Feature Importance Score

After completing the initial training process for the RFR (basic RFR), the second iteration of the model was adjusted to include only the 7 most important features, and the same model parameters were applied again.

Random Forest regression was fine-tuned using RandomizedSearchCV, a technique that randomly explores the hyperparameter space to find the optimal configuration for the model, the optimal parameters was obtained after fitting 5 folds for each of 100 candidates, totalling 500 fits, the best parameters were:

* ‘n\_estimators': 400 - 'min\_samples\_split': 2 - 'min\_samples\_leaf': 1
* 'max\_features': 'sqrt' - 'max\_depth': None - 'bootstrap': False

### 4.2 Support Vector Regressor (SVR)

Support Vector Regressor (SVR) can be useful for predicting the number of people traveling, especially when dealing with non-normally distributed data and outliers. SVR is known for its flexibility in handling non-linear relationships and its ability to generalize well to unseen data.

Initially the basic SVR model was applied to 80 -20 train test data split, this model was re-assed after fine-tunning the parameters using RandomizedSearchCV, the optimal parameters were obtained after fitting 5 folds for each of 100 candidates, totalling 500 fits, the best parameters were:

* 'kernel': 'linear' - 'gamma': 'scale' - 'epsilon': 0.001 - 'C': 1

The best parameters for the SVR model were determined, and it was decided by the data scientist not to create another iteration using the important features, considering the high accuracy of the test results and the minimal expected effect.

A graph with colorful bars

Description automatically generated

Figure 12: SVR Feature Importance Score

### 4.3 Ridge Regressor

It is a leaner model for regression, that can be a suitable choice for predicting the number of people traveling, especially when dealing with non-normally distributed data and moderate levels of outliers. It offers a good balance between simplicity, interpretability, and performance, making it worth to be considered for a specific task.

The best Ridge regression model was applied using the fine-tuned parameters by applying RandomizedSearchCV, the optimal parameters was obtained after fitting 5 folds for each of 24 candidates, totalling 120 fits, the best parameters were:

* 'solver': 'auto' - 'fit\_intercept': True - 'alpha': 0.01

The Ridge Coefficients were determined, and it was decided by the data scientist not to create another iteration, considering the high accuracy of the test results and the minimal expected effect.

A graph with blue and white text

Description automatically generated

Figure 13: Ridge Coefficients

## 5 Programming

Python has emerged as the preferred language for many data science tasks due to its versatility and ease of use. It combines the power of general-purpose programming languages with the simplicity of domain-specific scripting languages, in this code python language was used.

During the project design, the data scientist considered the influence of different programming paradigms to ensure effective decision-making. The imperative programming paradigm, with its step-by-step instructions, was adopted for most of the program specially with exploring the datasets from cell number 8 to 16. This approach guided the structure of the code’s logic, making it easy to read and allowing control over the code flow to change, modify, and optimize the program. (C. Müller and Guido, 2017. P.5)

For complex analysis, as demonstrated in cell number (94), the procedural programming paradigm was used. The code is compacted into one cell and executed sequentially, step by step, with each step tasked to perform a specific function. This programming approach made it easy to organize related elements, ensuring that the results of this code section were clear and directly relevant to its intended task.

The declarative programming paradigm was also utilized, leveraging imported libraries incorporating predefined functions and objects that involve Object-Oriented Programming (OOP) principles. By doing this, the data scientist expressed the intent to utilize these libraries effectively, importing and employing various functions and objects within them as needed.

In the function’s cells of the program, the data scientist constructed several functions to facilitate the reuse of code by simply calling these functions when necessary. It is recommended for future programs to create a separate function file to enhance program readability and compactness.

## 6 Results and Discussion

### 6.1 Results

The results of this study which aims to predict the number of arrivals into Ireland using different machine learning models. RFR, SVR and Ridge Regression performances were scored for R-squared, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as indicated in the table below; R2, RMSE and MAPE are the most used scoring parameter for regression (C. Müller and Guido, 2017, p. 301), where R2 Score measures how well the predicted data approximates the actual data. A score close to 0 means that the model did not fit well with the actual data. An R2 score of negative means that the model performed worse than a simple baseline model.

RMSE metric measures the average magnitude of the errors between the predicted and actual values, the RSME is used to assess the accuracy of the model’s prediction. A lower RSME indicates better accuracy.

MAPE measures the percentage absolute difference between the predicted value and the actual value. A MAPE score of zero means that the model has a perfect prediction accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Performance** | **Training R- 2** | **Test R-2** | **Test RMSE** | **Test MAPE** |
| **Basic RFR** | 0.9992949150888747 | 0.9921561306323602 | 0.055388879517696415 | 21.620232082279355 |
| **RFR \_Top 7 features** | 0.9993460882898371 | 0.9931141269407507 | 0.05189636144815242 | 25.032731123683828 |
| **Best RFR** | 1.0 | 0.9925980708665324 | 0.05380589564533354 | 25.45267825989262 |
| **Basic SVR** | 0.992996545315271 | 0.9572347216817699 | 0.1293311248909469 | 58.22209884060964 |
| **Best RFR** | 0.9999992298356158 | 0.9999986419276727 | 0.0007288176030316592 | 0.6630043417644301 |
| **Basic Ridge** | 0.9997558648774527 | 0.9995979463171548 | 0.012540054555886164 | 10.447693288706738 |
| **Best Ridge** | 0.999999749375581 | 0.999999670061898 | 0.00035923086117486486 | 0.27048257264600434 |

Table 14: comparison between ML models performance

### 6.2 Discussion

In justifying the decision to keep outliers in the dataset despite identifying them due to the COVID period and weather instability events, a few important points are highlighted. Firstly, outliers are seen as essential parts of the dataset, providing a true picture of real-world situations. By including these extreme values, the dataset stays genuine and accurate in reflecting the actual patterns in the data. Additionally, having unique data, including outliers, gives insights into the worst-case scenarios and unique challenges faced during this time. By considering these outliers, the analysis becomes more thorough, allowing for a stronger assessment of future expectations and aiding in planning for resilience.

Hyperparameter Tunning

Hyperparameter tuning is the process of optimizing the settings of a machine learning model to achieve the best performance. In this crucial step, the parameters will be fine-tuned that govern the behaviour of the models, enhancing their accuracy and robustness. (So et al., 2020, p.393)

RandomizedSearchCV has a couple of advantages over GridSearchCV. Firstly, it can be more efficient when dealing with large search spaces because it only looks at a subset of possible combinations instead of evaluating all of them. This means it can save time and computational resources. (So et al., 2020, p.419)

Secondly, RandomizedSearchCV can be more robust against overfitting. With GridSearchCV, there's a risk of finding hyperparameters that work well only on the training data but not on unseen data, especially when the search space is very large, and the model is simple. RandomizedSearchCV helps reduce this risk by randomly sampling from the search space, which makes it less likely to overfit. (So et al., 2020, p.425)

## 7 Conclusion

Supervised learning models can accurately generate a prediction based on historical data. By training the model on past data and associated traveller counts, the model can learn patterns and relationships that enable it to make accurate predictions for future travellers counts. it will also provide insights into the factors that influence traveller counts. By analysing the feature importances, that have significant impact on traveller counts.

In conclusion, the report provided insights on the performance of RFR, SVR and Ridge models on the combined dataset. the finding indicated that ridge model performs better in predicting the total number of travellers arrived to Ireland. However, the number of data available limited this study and may have an impact with the accuracy of the result. The data scientist recommends further testing using Time series model with large datasets to enhance the robustness of the findings of this study.

## Reference list

Brownlee, J. (2020). *Random Forest for Time Series Forecasting*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/random-forest-for-time-series-forecasting/> [Accessed 2 Apr. 2024].

C. Müller, A. and Guido, S. (2017). *Introduction to Machine Learning with Python A Guide for Data Scientists*. O’Reilly.

Central Statistic Office (2024a). *ASM01*. [online] cso.ie. Available at: <https://data.cso.ie/> [Accessed 25 Mar. 2024].

Central Statistic Office (2024b). *ASM02*. [online] cso.ie. Available at: <https://data.cso.ie/> [Accessed 25 Mar. 2024].

Central Statistic Office (2024c). *ASM03*. [online] cso.ie. Available at: <https://data.cso.ie/> [Accessed 25 Mar. 2024].

Department of the Taoiseach (2020). *gov.ie - Resilience and Recovery 2020-2021: Plan for Living with COVID-19*. [online] web.archive.org. Available at: [https://web.archive.org/web/20200919042930/https://www.gov.ie/en/campaigns/resilience-recovery-2020-2021-plan-for-living-with-covid-19/](https://web.archive.org/web/20200919042930/https:/www.gov.ie/en/campaigns/resilience-recovery-2020-2021-plan-for-living-with-covid-19/) [Accessed 29 Mar. 2024].

GeeksforGeeks. (2022). *Comparing Randomized Search and Grid Search for Hyperparameter Estimation in Scikit Learn*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/comparing-randomized-search-and-grid-search-for-hyperparameter-estimation-in-scikit-learn/> [Accessed 2 Apr. 2024].

Ilin, C., Annan-Phan, S., Tai, X.H., Mehra, S., Hsiang, S. and Blumenstock, J.E. (2021). Public mobility data enables COVID-19 forecasting and management at local and global scales. *Scientific Reports*, 11(1). doi:<https://doi.org/10.1038/s41598-021-92892-8>.

Ito, K., Kanemitsu, S., Kimura, R. and Omori, R. (2022). Future behaviours decision-making regarding travel avoidance during COVID-19 outbreaks. *Scientific Reports*, 12(1). doi:<https://doi.org/10.1038/s41598-022-24323-1>.

LibreTexts (2014). *6.2: the Sampling Distribution of the Sample Mean*. [online] Statistics LibreTexts. Available at: <https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Introductory_Statistics_(Shafer_and_Zhang)/06%3A_Sampling_Distributions/6.02%3A_The_Sampling_Distribution_of_the_Sample_Mean#:~:text=If%20the%20population%20is%20normal> [Accessed 7 Apr. 2024].

Met Eireann (2019). *Historical Data - Met Éireann - The Irish Meteorological Service*. [online] Www.met.ie. Available at: <https://www.met.ie/climate/available-data/historical-data> [Accessed 31 Mar. 2024].

Osei-Bryson, K.-M. and Barclay, C. (2015). *Knowledge Discovery Process and Methods to Enhance Organizational Performance*. Auerbach Publications.

RCSI, UNIVERSITY OF MEDICINE AND HEALTH SCIENCES (2022). *2020 – 2021: Years like no other Emergency Medicine in Ireland during the COVID-19 pandemic*. [online] *https://www.hse.ie/*, p.15. Available at: <https://www.hse.ie/eng/about/who/cspd/ncps/emp/achievements/emergency%20medicine%20in%20ireland%20during%20the%20covid-19%20pandemic-years%20like%20no%20other.pdf> [Accessed 29 Mar. 2024].

Scikit-learn Developers (n.d.). *3.3. Metrics and scoring: Quantifying the Quality of Predictions — scikit-learn 0.22 Documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics> [Accessed 2 Apr. 2024].

Sheposh, R. (2023). Central limit theorem.  *Salem Press Encyclopedia of Science*. [online] Available at: <https://research.ebsco.com/linkprocessor/plink?id=da09da5a-1f69-399a-a15b-89bfdff707ce.> [Accessed 7 Apr. 2024].

So, A., Joseph, T.V., John, R.T., Worsley, A. and Asare, S. (2020). *The the Data Science Workshop*. 2020.

Zhao, Y., Zhang, W. and Liu, X. (2024). Hyper-parameter optimization for financial time series forecasting. *ScienceDirect*, 154(1568-4946).

* AI tools were used to check for spelling and grammar errors in this report.