**Lorraine Drumm**

**Analysing Visitor Numbers to Ireland Since 2010 and Evaluating the Impact of the Covid-19 Pandemic on Visitor Numbers**

**Contents**

**1 Background and Introduction 3**

1.1 Abstract…………………………………………...........3

1.2 Background………………………………………........3

1.3 Data Selection………………………………………....3

**2 Data Preparation and Visualization 4**

2.1 Data Cleaning and preparation……………………….4

2.2 Preparing the Data for Machine Learning………..….5

2.3 Data Visualization…………………………………..….6

2.4 Tufte’s Principles………………………………….........7

**3 Statistics 7**

3.1 Descriptive Analysis……………………………..….....7

3.2 Discrete Distributions………………………………….11

3.3 Normal Distributions…………………………………..14

**4 Machine Learning 15**

4.1 Project Management Frameworks…………………..15

4.2 My Machine Learning Models………………..………16

4.2.1 Predictive machine modelling…………….………..16

4.2.1 Classification machine modelling……….…………18

**5 Programming 21**

5.1 Programming Paradigms…………………………..….21

**6 Conclusion 22**

**7 References 23**

*List of Figures*

1. Heatmap of Total Visitor Trends
2. Line plot of total visitor trends per year from 2010 to 2024
3. Subplot of the total visitor count per country from 2010 to 2023
4. Lineplot of top 3 arrivals from 2010 to 2024
5. Bar chart of total visitors July 2010 Vs July 2019
6. Boxplot of accumulated visitor count for each country from 2010 to 2024
7. Boxplot of accumulated visitor count from top 3 arrivals excluding outliers
8. Boxplot of total arrivals excluding outliers
9. Binomial Distribution graphs
10. Poisson Distribution graphs
11. Graph showing the transition of the Poisson distribution into Normal Distribution
12. Line plot of Average Expenditure per visitor from Great Britain for each month from April 2023 to January 2024
13. Normal Distribution of Average Expenditure per Visitor
14. Linear Regressor Model graph
15. Decision Tree Regressor Model graph
16. Random Forest Regressor Model graph
17. Confusion matrix for Classification model
18. Scattered boxplot showing classification results

*List of Tables*

1. Data types summary
2. Mean and median for each country
3. % change by country pre pandemic Vs during pandemic
4. Binomial Distribution Results
5. Daily Poisson distribution results
6. Monthly Poisson distribution results
7. Regression model results
8. Classification model results
9. Classification model accuracy and F1 scores

**Analysing Visitor Numbers to Ireland Since 2010 and Evaluating the Impact of the Covid-19 Pandemic on Visitor Numbers**

Lorraine Drumm

**1 Background and Introduction**

*1.1 Abstract*

*This report analyses visitor numbers from 12 countries to Ireland from January 2010 to January 2024.The aim of this study was to understand the makeup of these visitors, investigate the seasonal arrival trends and gain a deeper insight into tourism growth in Ireland. This study found that arrivals to arrival have been steadily increasing since 2010 but dropped significantly in 2020 due to the Covid-19 Pandemic. Numbers are beginning to recover since, and new visitor records are expected in July 2024.*

*1.2 Introduction*

Tourism and overseas visitors to Ireland are a huge source of revenue to the Irish Economy. The number of visitors to Ireland has consistently increased since 2010 however the Covid-19 Pandemic had an adverse effect on this growth. This report analyses the fluctuation in overseas visitors to Ireland from the 12 top countries over the last 15 years with a specific focus on the top 3: Great Britain, Spain and the United States. The CRISP-DM framework is used during this analysis to perform ML on the dataset to understand the impact of Covid-19 on current visitor numbers.

As part of my analysis I wanted to answer 6 questions:

* Which countries contribute most to arrivals?
* How have visitor numbers changed over time and seasonally?
* What is the descriptive statistics of visitor numbers?
* What impact has the Covid-19 Pandemic had on arrivals?
* How many visitors can Ireland expect in July 2024 and which machine learning is most effective to answer this question?
* How are arrival numbers distributed and what is the significance of this distribution?

*1.3 Data Selection:*

The dataset was taken from the CSO’s Open Data Platform and contains data from January 2010 to January 2024 (Data.cso.ie, 2024). This dataset was selected as it is from a reliable source and contains a large sample size. It provided an opportunity to analyse changes in Visitors to Ireland and investigate the impact of the Covid-19 Pandemic leading to meaningful insights.

**2 Data Preparation and Visualisation:**

After reading in the libraries the dataframe was imported as travel\_df. The columns were observed using the Pd.head() function. This displayed the first 5 rows and showed the columns: “STATISTIC Label”, “Month”, “Country”, “Direction”, “UNIT” and “VALUE”. It showed that every country had an arrival and departure visitor number for every month. The df size was observed using the shape function: (6084, 6). The Pandas.tail() function was used to confirm the time range of the dataset.

*2.1 Data cleaning and preparation*

EDA was conducted. For this Pandas.describe() and Pandas.describe(include=object) were incorporated. This showed I had 18 unique country categories. It confirmed that the “STATISTIC Label” column was unnecessary as it only contained 1 unique repeated string. This was dropped using Pandas.drop()

The aim of this analysis was to investigate inbound arrivals therefore all Departure rows were dropped. This removed the need for “Direction” so this column was dropped in the same procedure as “STATISTIC LABEL”. A new column was created called “Visitors” which would replace “UNIT” and “VALUE”. This was accomplished by multiplying “UNIT” by 100 and inserting answers into a column called “Visitors”. The 2 unnecessary columns were dropped.

After inspecting the values in the “Country” column it was observed that some entries referred to groups of counties or variations of a country already entered. For example, “Other Countries(42)”, “Selected EU” and “Other UK”. It was decided to remove country values of this form from the dataset as the focus of this analysis was on individual countries. Two new columns for longitude and latitude was created using Geo Coder. This programme generated coordinates for counties listed and returned NaN values for entries such as “All countries” or “Other Europe”. All rows with NaN values in the “Longitude” column were dropped. It was appropriate to drop these rows as the focus of the analysis was on arrivals from the 12 top contributors only and entries such as “All countries” was just the sum of data already accounted for.

The values in “Month” were converted to datetime. This is essential for effective data filtering and plotting of data (Mckinney, 2022). This column was renamed “Date” and 2 new columns were added to the dataframe: “Year” and “Month” which split the month and year into separate columns allowing for more varied analysis and selection later. Final checks on data such as count, presence of duplicates and number of unique variables in each column were conducted. The rows corresponding to the outliers were found by calculating the IQR for Great Britain using a function and boxplots and they were removed.

*2.2 Preparing Data for Machine Learning:*

Methods to prepare data for ML included imputing missing data, removing rows with missing values, scaling and identifying and removing outliers that could affect the model. Encoding categorical data using various techniques and carrying out feature engineering (Provost and Fawcett, 2013). Feature selection was also incorporated. The data was split into training and testing groups in a 2:8 ratio and then varied to obtain optimum model performance. Appropriate hyperparameters were chosen using GridSearchCV . (Iqbal, 2024).

During EDA it was identified that “Visitors” had 2 missing values. The corresponding rows were identified using the using the print() and .isnull() functions. They corresponded to “Other UK (1)” for January and October 2021. It was decided to replace these missing values with the mean visitor number from Other UK(1) for 2021 (McQuaid, 2024). This method ignores seasonal tourism trends however this was the most appropriate approach as the months were during the Covid-19 pandemic and they were likely not representative of visitor trends from the same month in previous years.

Carrying out descriptive statistics on the data allowed for the identification of extreme outliers from 2020 to 2022. This was carried out by calculating IQRs and filtering outside an appropriate domain. If these extreme outliers hadn’t been removed they would have affected the result of the ML models (Inuwa, 2022).

For the classification model the aggregated sum of all visitors for each month was classified as very low, low, medium or high depending on their total visitors. To carry out the Random Forest Classification these object categories were then encoded as the integers 0,1 and 2 to be interpreted by the model. For this LabelEncoder() is used and fit\_transform. The “fit” component identifies the unique labels and maps it to a unique integer. The transformation refers to replacing the categorical data with this integer.

For all the machine learning models the data had to be split into test and train groups. The ratio of these 2 groups was initially 2:8 however this ratio was altered slightly for individual models to obtain the best test mean squared error result. Optimum parameters were also selected using GridSearchCV and RandomisedSearchCV.

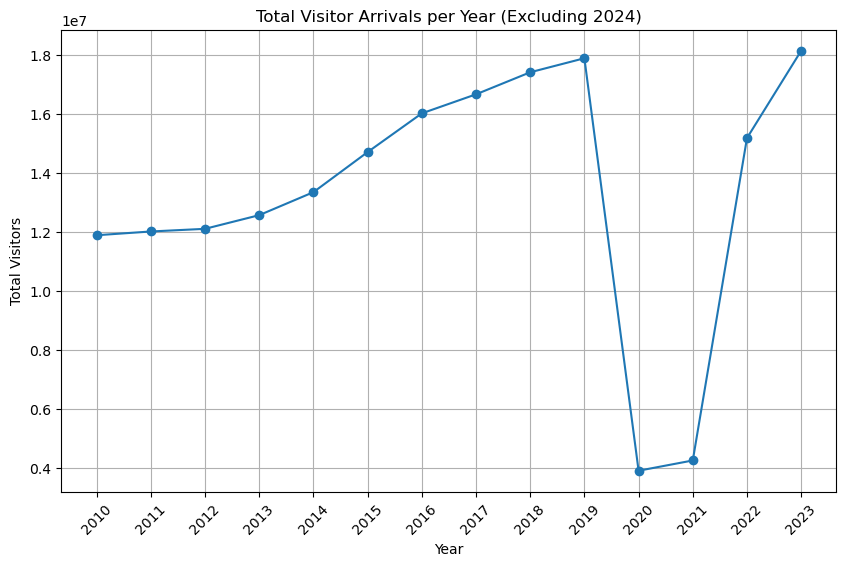
These steps increased the probability of obtaining an accurate result for the models.

I scaled my “Visitor” values but this didn’t improve my ML accuracy so I didn’t incorporate this in the final programme.

*2.3 Visualising the Data:*

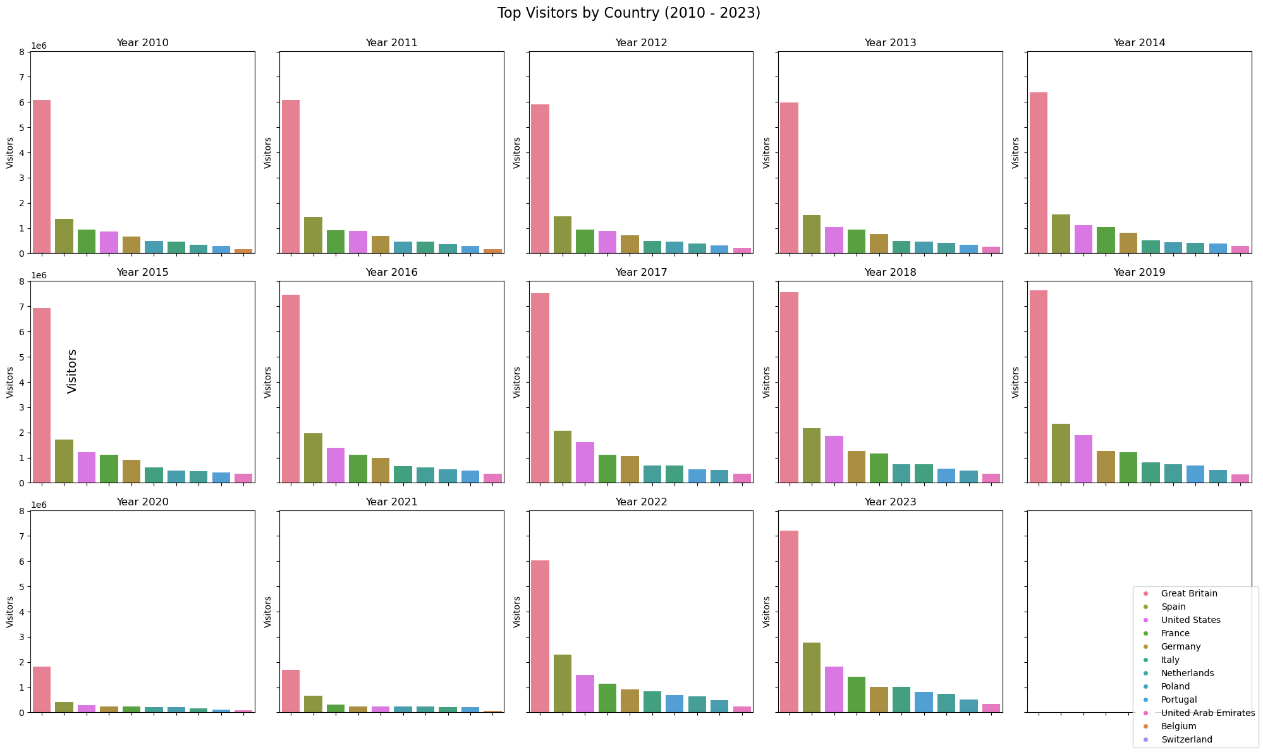
To begin understanding the trend in arrivals to Ireland the total visitors summed. It is vital to understand the bigger picture before focusing on individual countries. Several visual aids such as tables, heat maps, line plots and bar charts were used to represent the change in visitor numbers. The graph selected was based on the type of data it was representing and in what context (McQuaid, 2024b). A table was created which showed the sum of all visitors for every month and the total number of visitors every year. This was effective as it clearly showed the steady growth for specific months and for the total of each year. To compliment this a heatmap was utilised to show the same information but in a more user friendly and visual way. A line plot was used to show the total annual growth only but highlighted the rate of this growth. Th month with the highest arrivals with 1,985,700 visitors was 2024-07 which shows tourism is recovering.

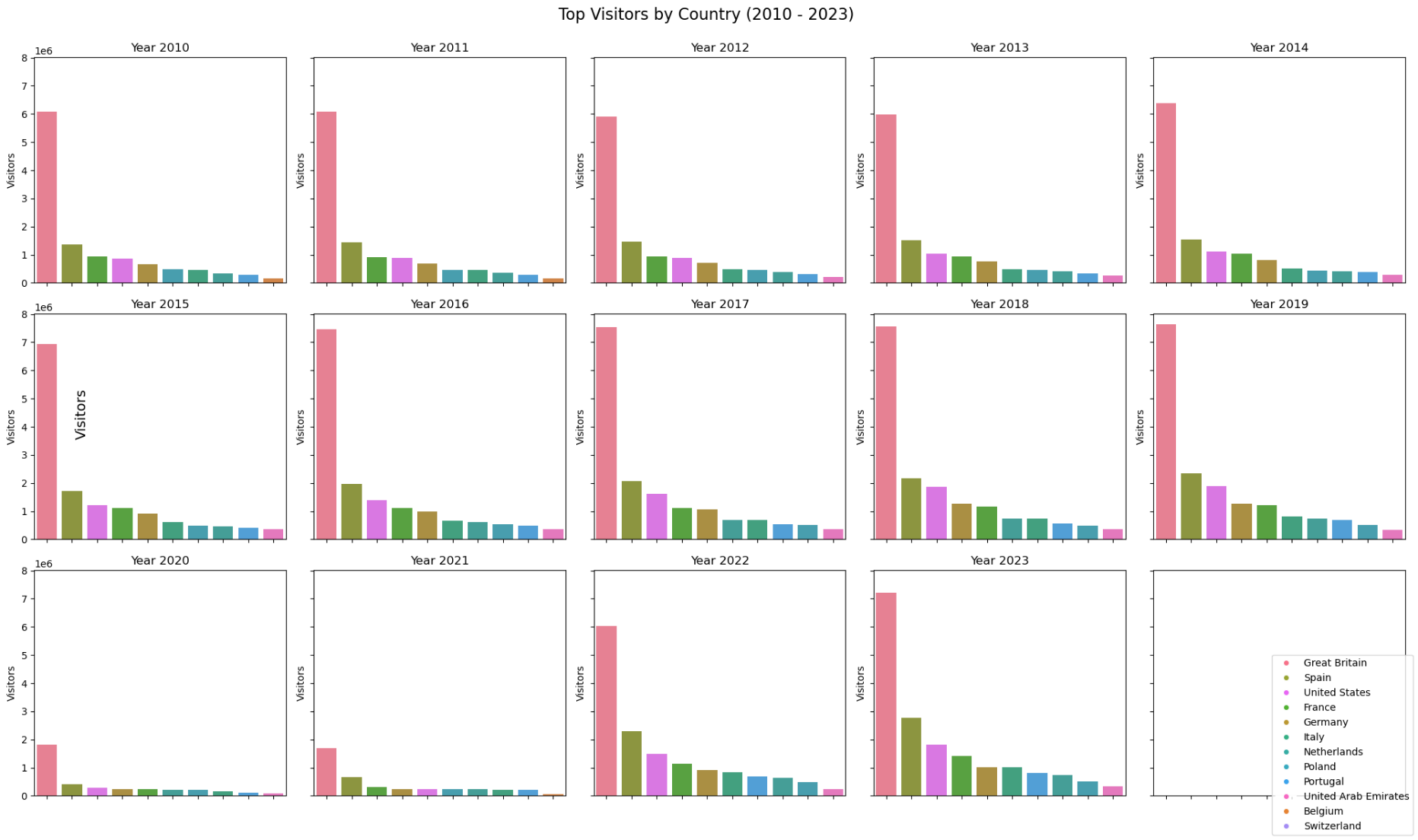
A blue squares with white text

Description automatically generated

*Figure 1: Heatmap of total visitor trends Figure 2: Line plot of total visitor trends annually*

After looking at the total visitor count it was time to investigate the breakdown of where these visitors came from. To visualize this a collection of bar charts which showed the number of visitors from each country annually was created.

****

****

*Figure 3:* *Subplot of the total visitor count per country from 2010 to 2023*

*2.4 Design Features of Visualisations Used (Tufte’s Principles):*

The graphs selected to display the data were selected and refined carefully in alignment with Edward Tufte’s Principles (Tufte, 1983).

Firstly, **the representation of numerical data is accurate and proportionate.** All barcharts used are directly proportional to each other. In Figure 3 it is clear to the observer that Great Britain makes up the largest amount of arrivals. When taking a closer look at the exact visitor numbers in July 2010 in figure 5 we can see how 671,100 people came from Great Britain and 193,300 came from Spain, both identified using labels and a legend. These figures tell us 3 times more people came from Great Britain which matchesthe graph’s scale. Labelling and use of legends used throughout the graphs shows the use of Tufte’s second principle: “**Clear and thorough labelling should be used to defeat graphical distortion and ambiguity**”. The third principle states that graphs should **“show data variation not design variation”.** This is evident in Figure 3. For this subplot the same colour was used for each country for each individual year to avoid misinterpretation. All graphs are in 2d format as this is appropriate for the data. This aligns with Tufte’s 5th principle. The data to ink ratio has been designed to be as low as possible. For example, in figure 1 when the graph was initially generated the values for each month were included however this appeared cluttered and blocked clearly interpreting the tile. All data has been accurately visualised without distorting the data.

**3 Statistics**

*3.1 Descriptive Analysis*

Descriptive Analysis was used to summarize the dataset. Areas of focus included identifying total arrivals and seasonal trends, discovering what countries contributed the most and the mean and median number of visitors from specific countries and to identify the outliers in the dataset.

A summary of the data types in each column can be found in the table below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Country | Visitors | Month | Year | Latitude | Longitude |
| datetime | object | integer | integer | integer | float | float |
| Discrete data | nominal | Discrete data | Discrete data | Discrete data | Continuous | Continuous |

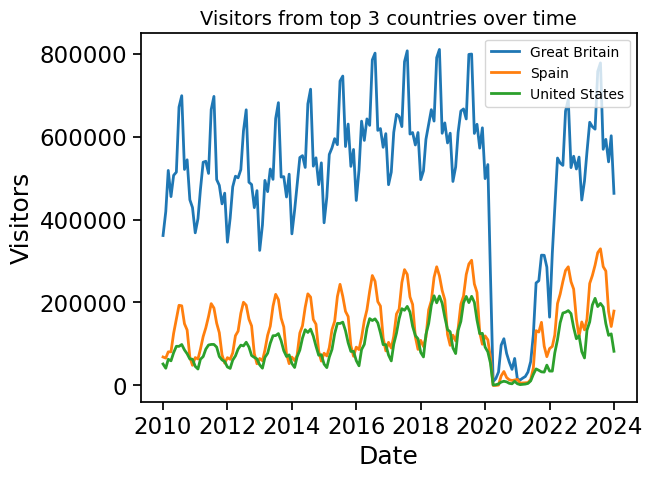
*Table 1: Data Types Summary*

The mean and median number of visitors from each individual country and the aggregated mean of all countries (after removing the outliers) are as follows:

|  |  |  |
| --- | --- | --- |
| Month | Mean (monthly visitors) | Median (monthly visitors) |
| Great Britain | 562,656 | 550,050 |
| Spain | 157,422 | 148,200 |
| United States | 111,856 | 98,950 |
| France | 90,707 | 86,900 |
| Germany | 77,072 | 75,800 |
| Italy | 53,944 | 49,050 |
| Netherlands | 45,436 | 46,150 |
| Poland | 40,920 | 41,450 |
| Portugal | 40,363 | 37,750 |
| United Arab Emirates | 23,193 | 24,450 |
| Belgium | 18,770 | 19,200 |
| Switzerland | 16,505 | 16,000 |
| Aggregated sum of total countries | 1,238,850 | 1,204,700 |

*Table 2: Mean and Median for each country*

It is clear from the Figure 1 that there are seasonal trends in arrivals to Ireland with the summer months of July and August consistently receiving the largest number of arrivals. Total annual visitor numbers increased by 50.5% from the 12 countries mentioned between 2010 and 2019. The seasonal changes can also be seen in figure 4 which shows variation that appears to follow a cycle throughout each year consistently.



*Figure 4: Lineplot of top 3 arrivals from 2010 to 2024*

The growth by country can be seen in Figure 5 and Table 3.

|  |  |  |
| --- | --- | --- |
| Country | % Increase from 2010 to 2019 | % Decrease from 2019 to 2020 |
| Great Britain | 25.3 | -76.29 |
| Spain | 70.69 | -82.33 |
| United States | 117.19 | -85.19 |
| France | 28.14 | -80.3 |
| Germany | 88 | -80.2 |
| Italy | 77.46 | -79.67 |
| Poland | 5.7 | -59.5 |
| Portugal | 132.14 | -82.24 |
| Netherlands | 122.5 | -70.82 |
| United Arab Emirates | 234.24 | -72.62 |
| Belgium | 55.49 | -76.33 |
| Switzerland | 95.26 | -78.51 |
| Total | 50.5 | -78.2 |

*Table 3: % change by country pre pandemic Vs during pandemic*

UAE saw the most growth from 2010 to 2019 while the United States saw the biggest decrease during 2020. The total % decrease in arrivals during the pandemic was 78.2%.

A graph of different colored bars

Description automatically generated with medium confidence

*Figure 5: Bar chart of total visitors July 2010 Vs July 2019*

The very low visitor numbers from March 2020 to January 2022 show extreme outliers which are caused by the Covid-19 pandemic. This highlights there is a need to filter the data of outliers that will skew the statistical findings. This gradual increase and sudden reduction can be seen even clearly in Figure 1.

These outliers were investigated by creating a boxplot which showed extreme outliers in many countries but most obviously from Great Britain. The IQR range of Great Britain was calculated by computing the lower and upper quantiles and printing the visitor values that fell outside this range. The reason for focusing on only Great Britain to filter this data was due to the extreme outliers observed and they are the largest visitor group to Ireland. The original boxplot can be seen below.

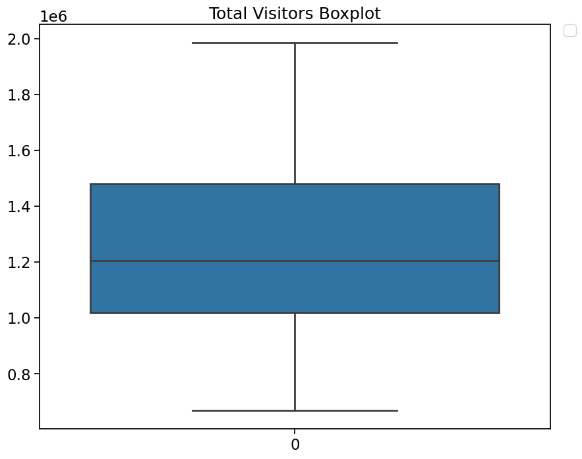
A graph with different colored columns

Description automatically generated

*Figure 6: Boxplot of accumulated visitor count for each country from 2010 to 2024*

When these outliers were located in the dataset, they confirmed the original claim that they occurred due to the Covid-19 Pandemic. March 2020 to January 2022 was removed. An updated boxplot of the top 3 arrivals was created to confirm that the outliers had been filtered out of the dataset.

A graph of a group of people

Description automatically generated with medium confidence

*Figure 7: Boxplot of accumulated visitor Figure 8:* Boxplot of total arrivals excluding outliers *count from top 3 arrivals excluding outliers*

The boxplots show that there are no extreme outliers remaining in the data and are very effective in understanding the spread of the data (Tukey, 2020). Focusing on Figure 5 the median line confirms the value of 1,204,700 obtained earlier and tells us that 50% of the months from January 2010 to January 2024 (excluding outliers) had over this number of visitors. The IQR represented by the box contains visitor numbers from 1,018,025 to 1,479,800 and 50% of months fall within this range. The range is not significantly large in comparison to the median which suggests the monthly visitor numbers are moderately consistent. The whiskers mark the max and min of the data which are equal to 668,800 and 1,985,700 respectively. The minimum value corresponds to January 2012 and the maximum visitors came in August 2023. This once again shows Irish visitor numbers are consistently increasing again post pandemic. The symmetry of the boxplot around the median confirms the even distribution and stability in visitor numbers over time.

*3.2 Discrete Distribution Analysis*

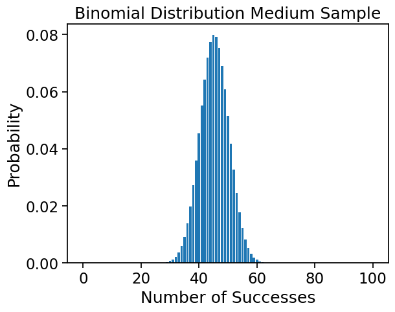
Both Poisson and Binomial distribution were used to identify information about this dataset. Both types of distribution work with discrete data which meant visitor count was an appropriate variable to use in both cases (Researchoptimus.com, 2019). Discrete data refers to numerical data that can only hold specific values and can be counted.

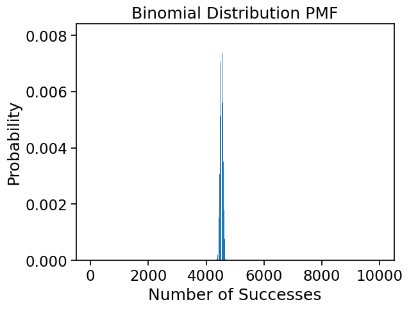
For **Binomial distribution** the question posed was: **“What is the probability of x visitors being from Great Britain?”** There are two possible outcomes: they are from Great Britain, or they are not.

It was found that on average the probability of a visitor being from Great Britain is 0.454176. This relative frequency was calculated based off the dataset df\_no\_covid. Based off this the expected frequency of a visitor being from Great Britain is 454 visitors per 1,000 total. To begin a small sample size of 10 was analysed. It was then increased to 100 and finally to 10000. The results can be seen in Table 3.

|  |  |  |  |
| --- | --- | --- | --- |
| No. of visitors in sample | No. of successes desired | lambda | probability |
| 10 | 4 | 0.4534 | 0.163 |
| 100 | 40 | 0.4534 | 0.0455 |
| 10000 | 400 | 0.4534 | 0.0 |

*Table 3: Binomial Distribution Results*

A graph of a number of successes

Description automatically generated****As sample size increases the the probability decreses.

*Figure 9: Binomial Distribution graphs*

For Poisson distribution the daily arrival rates for Switzerland were investigated. It was chosen to initially work with average daily arrivals as it was observed that the distribution performed better with small sample sizes. The results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Country: | Average daily arrivals: | No. of arrivals to check probability for: | Probability |
| Switzerland | 560.58 | 550 | 0.0154 |
| Switzerland | 560.58 | >550 | 0.6628 |
| Switzerland | 560.58 | =<550 | 0.3372 |

*Table 4: Daily Poisson distribution results*

|  |  |  |  |
| --- | --- | --- | --- |
| Country: | Average monthly arrivals: | No. of arrivals to check probability for: | Probability |
| Switzerland | 17055.48 | 17000 | 0.0028 |
| Switzerland | 17055.48 | >17000 | 0.6627 |
| Switzerland | 17055.48 | =<17000 | 0.3373 |

*Table 5: Monthly Poisson distribution results*

**A graph of a function

Description automatically generated with medium confidence**The results have shown that it is very likely that over 550 visitors per day and 17000 per month visit from Switzerland. However, it is unlikely that these exact numbers will be observed. The probability of predicting the exact number of visitors decreases as the sample size increases.

*Figure 10: Poisson Distribution Graphs*

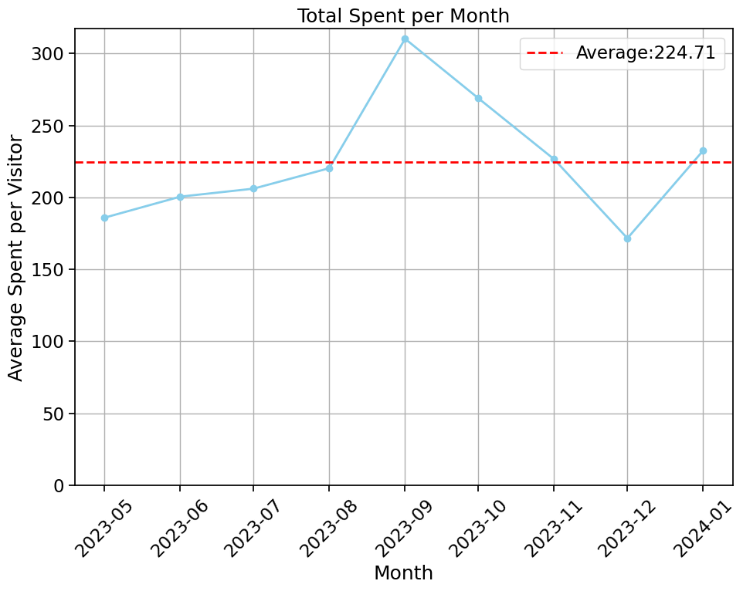
With large samples both Binomial and Poisson distribution start to display characteristics like Normal Distribution (Moore, Mccabe and Craig, 2017). This is known as the Central Limit Theorem and occurs when np>=10 and n(1-p)>=10 where n = no of trials and p= probability of event occurring (Statistics LibreTexts, 2018). When n=10 the distribution was left skewed but when there were 100 trials the graph forms a symmetric curve. As can be seen in the Poisson graphs in figure 10 both graphs form a normal distribution curve. This is because Poisson approaches normal distribution when **λ ≥ 20. A graph with a range of small λ values can be seen in figure X to show the transition from left skewed to the normal distribution curve as λ increases.**

A graph of different events

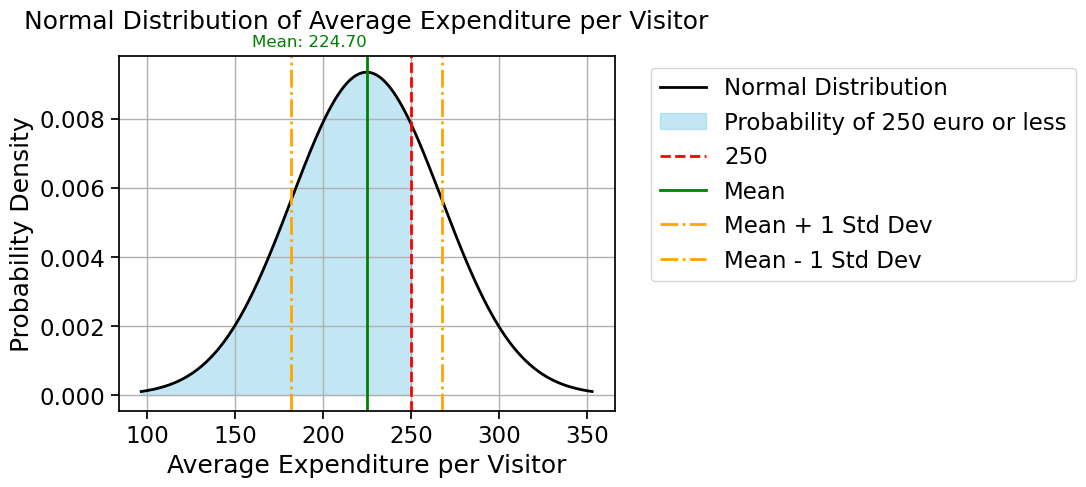
Description automatically generated

*Figure 11: Graph showing the transition of the Poisson distribution into Normal Distribution*

*3.3 Normal Distribution:*

****Normal distribution can only be carried out on continuous data (Iantorno, 2024). A new column was added to the df from the CSO Open Data Portal which contained information on how much money visitors from Great Britain spent from the April 2023 to January 2024(Data.cso.ie, 2024) . Continuous data can assume any numeric value and can be split into smaller parts. Using arrivals from Great Britain for these months the average amount spent per visitor for each month was calculated. The results can be seen in Figure 12.

*Figure 12: Line plot of Average Expenditure per visitor from Great Britain for each month from April 2023 to January 2024*

Normal distribution was used to calculate that the probability of a visitor spending less than **€**250 is 0.723.

*Figure 13: Normal Distribution of Average Expenditure per Visitor*

Applying the Empirical rule to the data set allowed to determine that approximately:

* **68% of visitors spent between €182 - €267.40**
* **95% of visitors spent between €139.30 - €310.10**
* **99.7% of visitors spent between €96.60 - €352.80**

Distributions are important to consider because they give us information such as central tendency, variability, and skewedness (Chip, 2023). They help us make educated guesses for larger samples based on smaller samples. This helps us to understand our data better and give us information on the probability of desired events occurring. As mentioned above the data used for Normal distribution has to be continuous so the data used for the Discrete distributions could not have been used.

**4 Machine Learning:**

*4.1 Project Management Frameworks*

Depending on the specific goals and nature of a data science project CRISP-DM, KDD or SEMMA could be used to as a project management framework . CRISP-DM is widely used and is very applicable in industry. It is the most commonly used methodology used. It consists of 6 stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment (Hotz, 2023).

Business Understanding is the starting point for the project and consists of gaining a deep understanding of the customers needs. This helps to define the project requirements, identify risks and determine what success will look like.

Next it is time to understand the data. The data is collected and EDA is performed. This helps to determine the data quality, visualise it and select interesting subsets.

Data preparation makes up approximately 80% of the project. The data is prepared for modelling through processes such as scaling and encoding. Appropriate variables are identified and the data is filtered of impurities.

It is now appropriate to commence the modelling phase. Here modelling techniques are applied to the data and the model settings are calibrated for optimum results.

The models must now be evaluated to ensure they are effective and to ensure that they have achieved the objectives set out in stage 1. If the evaluation finds the process has been successful the team can move into Deployment.

Deployment phase consists of report creation and a project review. Real-life scenarios where the use of this model is essential for project success include projects in companies such as Google and Microsoft where large teams are required to work together to carry out projects that span multiple teams in multiple countries.

Supervised machine learning was used for the chosen data set in this project as it contained labelled data which I wanted to use in both a classification and prediction application.

*4.2 Machine Learning Models:*

Machine Learning models were used to evaluate: What are the tourist numbers expected for July 2024 and can the months be categorised based off their average visitor count?

*4.2.1 Predictive Machine Modelling*

**Predictions for July 2024 based on 3 regression models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Predicted Visitors | Test RMSE | R squared value | Best score (MSE) |
| Linear Regression | 1586706.277 |  |  |  |
| Decision Tree Regression | 1985700 | 100624.5 | 0.9298 | 10125290000 |
| Random Forest Regression | 1868134 | 107082.75 | 0.9235 | 11466714448.65 |

*Table 6: Regression Model Results*

Comparing the model results the linear model does not perform as well as it’s output is significantly lower. The RMSE scores of the Decision Tree and the Random Forest Regressor are similar however the Decision Tree result is slightly better which indicates better fit (Iqbal, 2024). Both models have a high r squared score of 0.9298 and 0.9235 which shows a good level of fit however the Decision once again slightly outperforms the random forest. Comparing MSE scores, they both seem very high however expected visitor number is measured in millions and this score is proportionally large. Also as MSE squares the errors larger errors are more heavily impacted by this (Stephen Allwright, 2022). The decision tree received a better score.

The Decision Tree Regression model performed the best and suggest that the results are the most accurate for predicting visitor numbers in July 2024. However, it is a more simple model which means it could be overfitted (Hegelich, 2016).

The results of the models can be seen in Figure 14,15 and 16:

A graph with blue dots and red line

Description automatically generated

*Figure 14: Linear Regressor Model graph*

A graph with blue dots

Description automatically generated

Figure 15: Decision Tree Regressor Model graph

A graph of a graph with numbers and lines

Description automatically generated with medium confidence

Figure 16: Random Forest Regressor Model graph

*4.2.2 Classification Machine Modelling*

A random forest classification model was used to classify the months into 4 categories based on their average historical visitor count. This was carried out by creating 4 quantiles: very low, low, medium and high. The results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Very Low | Low | Medium | High |
| January | May | April | June |
| February | October |  | July |
| March |  |  | August |
| November |  |  |  |
| December |  |  |  |

*Table 7: Classification Model Results*

High Visitor Number Category:

The accuracy of the model can be visualized using Figure 17 below. The model predicted 9 months as “High” correctly however it predicted 3 “high” months as “medium. The model is correct 75% and equally balanced between false positive and negatives. This can be seen in table 8.

Medium Visitor Number Category:

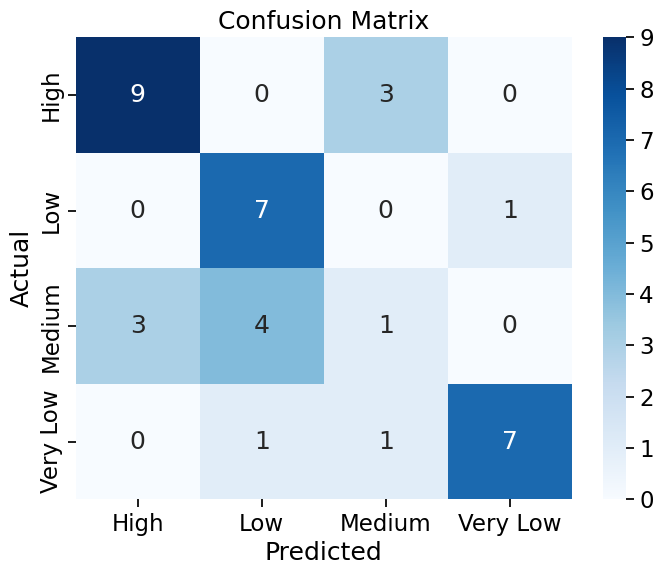
The model did not perform well in this category. It predicted one month correct and predicted 3 months as “high” visitor count. It predicted 4 months incorrectly as “Low”. The category accuracy is 15%.

Low Visitor Number Category:

Seven months in the training set were predicted correctly. One error was made classifying 1 “Low” month as “Very Low”. The model is 70% accurate in this category. It predicts “Low” correctly 88% but has a lower precision score of 58%.

Very Low Visitor Number Category:

Seven months were predicted correctly. There was 2 months incorrectly classified as Low and Medium. The model performs well in this category with an F1 score of 82%.

****

*Figure 17: Confusion matrix for Classification model*

The model is good at categorizing the 2 extreme categories but doesn’t perform as effectively in the “Low” and “Medium” category. It did not mislabel “high” as low or “very low” at any point and the reverse is also true. It can effectively distinguish between extremes. Due to this I believe this model would perform better with 3 quantiles as it often confuses low and medium suggesting many months vary between both categories on a year-by-year basis.

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | F1 score |
| High | 0.75 | 0.75 | 0.75 |
| Medium | 0.2 | 0.12 | 0.15 |
| Low | 0.58 | 0.88 | 0.7 |
| Very Low | 0.88 | 0.78 | 0.82 |
| Accuracy |  |  | 0.65 |
| Average | 0.6 | 0.63 | 0.61 |
| Weighted Average | 0.63 | 0.65 | 0.63 |

*Table 8: Classification model accuracy and F1 scores*

*A graph with colored squares and dots

Description automatically generated with medium confidence*

*Figure 18: Scattered boxplot showing classification results*

**4 Programming**

*Programming Paradigms:*

Programming paradigms included in the development of this project include:

* Imperative Programming:
* Procedural Programming
* Functional Programming
* Object Orientated Programming
* Declarative Programming

Procedural Programming:

I will focus on my use of procedural programming in my analyses and how it impacted my design decisions. I used this paradigm when I wanted to plot the subplot in Figure X. Instead of plotting each graph individually I looped through all the years in my dataset in a logical order. This was carried out in a logical order to achieve the end goal of plotting my graph. I then wanted to focus in on July 2010 and 2019 only. To achieve this my code was focused on step-by-step procedures that performed a task to reach my end goal. This is typical of this paradigm. This made my code more efficient and automated. I incorporated input operations throughout my programme which made it more dynamic and allowed me to change the direction of my project each time I ran the code. When I was completing my machine learning models I broke the steps into functions and distinct sections. This allowed for a more readable programme.

Functional Programming:

I used pure functions throughout my programme. This helped me to easily compare datasets before and after removing outliers without having to rewrite the code. It also allowed me to come back and update my code easily as I filtered and renamed my datasets. I incorporated user input functions throughout my code which helped with comparative analysis for different countries and months. I was then able to directly link the user input into my graphs for efficient plotting.

Declarative Programming:

I found declarative programming especially useful in my programme. It helped me to tackle more complex problems and allowed for more concise lines of code. For example I had to aggregated my data set into categories such as total visitors from all countries or total visitors from a specific country over time. When I would group the data in this name I would create a new updated dataframe which allowed me to answer more diverse questions in a simple way and view the problem from different perspectives.

**Conclusion:**

The results of this analysis show that tourism in Ireland has been growing consistently since 2010 however the Covid-19 pandemic had a negative impact on visitor numbers. However, numbers are once again increasing with the highest visitor count to date expected this July. Most visitors come from Great Britain and more visitors visit Ireland in the summer months.

**References:**

Chip (2023). *Understanding the Role of Distribution in Statistical Analysis -*. [online] quanthub. Available at: https://www.quanthub.com/understanding-the-role-of-distribution-in-statistical-analysis/#:~:text=Distributions%20give%20us%20a%20snapshot%20of%20our%20data.

Data.cso.ie. (2024). *ASMO1 AIR AND SEA TRAVEL*. [online] Available at: [https://data.cso.ie/#](https://data.cso.ie/).

Hegelich, S. (2016). Decision Trees and Random Forests: Machine Learning Techniques to Classify Rare Events. *European Policy Analysis*, 2(1). doi:https://doi.org/10.18278/epa.2.1.7.

Hotz, N. (2023). *What is CRISP DM?* [online] Data Science Project Management. Available at: https://www.datascience-pm.com/crisp-dm-2/.

Iantorno, M. (2024). *Normal Distribution*.

Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media, Inc. October 2016.

Inuwa, M. (2022). *Outliers and Overfitting when Machine Learning Models can’t Reason -*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/outliers-and-overfitting-when-machine-learning-models-cant-reason/.

Iqbal, M. (2024). *Tutorial\_2*. [Lecture].

Mckinney, W. (2022). *Python for data analysis : data wrangling with pandas, NumPy, and IPython*. O’reilly Uuuu-Uuuu.

McQuaid, D. (2024a). *How to Deal with Missing Data in Python*.

McQuaid, D. (2024b). *What Visualisations should I use*.

Moore, D.S., Mccabe, G.P. and Craig, B.A. (2017). *Introduction to the practice of statistics*. New York, Ny: W. H. Freeman And Company.

Provost, F. and Fawcett, T. (2013). *Data science for business : what you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O’Reilly Media.

Researchoptimus.com. (2019). *Normal, Binomial and Poisson Distribution Explained | ROP*. [online] Available at: <https://www.researchoptimus.com/article/normal-binomial-poisson-distribution.php>

Statistics LibreTexts. (2018). *9.1: Central Limit Theorem for Bernoulli Trials*. [online] Available at: <https://stats.libretexts.org/Bookshelves/Probability_Theory/Book%3A_Introductory_Probability_(Grinstead_and_Snell)/09%3A_Central_Limit_Theorem/9.01%3A_Central_Limit_Theorem_for_Bernoulli_Trials>.

Stephen Allwright. (2022). *How to interpret MSE (simply explained)*. [online] Available at: https://stephenallwright.com/interpret-mse/.

Tufte, E.R. (1983). *The Visual display of quantitative information*. Cheshire, Conn.: Graphics Press.

Tukey, J.W. (2020). *Exploratory data analysis*. Hoboken, Nj: Pearson.