**The Importance of Tourism in Ireland**

**Author: Andrew Mc Guinn**

**Email: sba24458@student.cct.ie**

**Student ID: sba24458**

4400 Words

*Abstract- Tourism plays a significant role in Irelands economy, making the ability to predict tourism trend and expenditure essential for effective planning and resource allocation. This project aims to analyse the factors influencing Irelands tourism sector from 2012 to 2019 using a combination of descriptive statistics and various regression models, including Linear, Ridge and Lasso regression as well as Decision Tree Regressor. By examining data such as flight arrivals and tourist expenditure, relevant distributions were applied such as Poisson, Binomial and normal distribution to better understand the data and make accurate predictions on visitor behaviour and predict tourism trends. The ML models are evaluated using appropriate metrics such as MAE, MSE and R-squared values. Each of the regression algorithms were trained with 80% of the data, using 5-fold cross validation, and the remaining 20% was used for testing the model.*

*Keywords- Machine Learning, Tourism, Linear Regression, Poisson distribution, Binomial distribution, Normal distribution, Decision Tree.*

1. **Introduction**

## Background on tourism in Ireland

Tourism in Ireland has a rich history that reflects the country’s cultural heritage and natural beauty. The industry began to develop in the 18th century, with increased accessibility due to improvements in infrastructure, such as roads and railways. The establishment of tourist boards and promotional campaigns in the late 19th and early 20th centuries helped popularize iconic attractions such as the Cliffs of Moher and Newgrange (failte, n.d.).

In the 1960’s, the Irish government launched initiatives to promote international tourism, significantly boosting visitor numbers. The hospitality sector expanded rapidly, with new hotels and guesthouses catering to the growing demand. By 2019, tourism had become a major economic driver worth 9.3 billion euros in GDP, contributing significantly to employment for an estimated 265,00 people and attracting millions of visitors each year (itic, 2019).

As Ireland approached the 2020’s, the focus was on enhancing visitor experiences, preserving cultural heritage and promoting sustainable practices. Leveraging data analytics and machine learning techniques became increasingly important for understanding tourism trends and optimising resource allocation in the industry.

## Motivation and potential benefits

Given the vast amount of tourism data available and the challenges faced by policymakers and tourism authorities in gaining real time insights, the application of machine learning algorithms to analyse tourism trends in Ireland presents significant opportunities.

This analysis aims to provide valuable insights that can guide strategic decision making, enhance marketing efforts and optimise resource allocation in the tourism industry. Furthermore, enhanced understanding of tourism dynamics can lead to improved visitor experiences, ultimately contributing to the overall well-being of both tourist and local communities.

## Problem statement

Although vast amounts of data and information on tourism in Ireland are now available, much of it is unstructured , making it difficult to extract meaningful insights efficiently. The substantial volume of tourism data collected by state bodies and tourist agencies, poses challenges for manual analysis, making it difficult to uncover valuable patterns and trends and insights.

While previous studies have analysed various trends and patterns in tourism, there has been limited work focusing on evaluating the accuracy of machine learning models for tourism in Ireland. Past research has generally centred on descriptive statistics or historical trend analysis rather than developing predictive models. This project addresses this gap by using relevant tourism data from Ireland to evaluate and compare the performance of several machine learning algorithms in forecasting tourist expenditures, thereby offering actionable insights to relevant bodies.

## Aim of the research

This project aims to compare the prediction accuracy of several machine learning algorithms such as Linear, lasso, Ridge and Decision Tree Regression, focusing on predicting overseas tourist expenditures in Ireland. The models will be evaluated using relevant metrics such as mean absolute error (MAE), root mean squared error (RMSE) and R2 score.

The research will also aim to gain insights on the trends and patterns from the data in the following:

* Analysing the distribution of the data through relevant graphs and tests
* Analysing the probability of tourists spending a certain number of nights through a Binomial distribution
* Using a Poisson distribution to examine the average number of tourist arrivals
* Using a Normal distribution to analyse tourist expenditure trends
* Analysing expenditure trends over time and any seasonal trends in the data.

# **Related Work**

## Introduction

Tourism is defined as the activity of travelling to and staying in places outside one’s usual environment for leisure business or other purposes (UN, 2008). In recent years, the application of machine learning techniques in tourism analytics has gained momentum, emphasizing data driven approaches for enhancing visitor experiences and economic benefits, this study seeks to contribute to this growing body of literature.

## Literature Review

The capability of Machine learning in tourism management is essential for optimising operations, enhancing customer experiences and improving decision making. Areas such as cultural tourism, nature tourism and MICE (meetings incentives, conferences and exhibitions) significantly benefit from ML applications with supervised learning methods such as decision trees and neural networks particularly effective for predicting tourist spending. Unsupervised learning techniques including k-means are utilised to identify traveller preferences and marketing strategies (Gössling, 2020).

Traditional linear regression models may demonstrate high correlation coefficients but are prone to overfitting, especially when applied to complex datasets with multiple correlated predictors. This can result in models that perform well on training data but fail to generalise to new unseen data. This paper emphasizes the role of regression analysis in data science and the significance of advanced techniques such as Ridge and Lasso regression in addressing challenges related to multicollinearity and overfitting (Mayooran Thevaraja1, 2019).

This paper discusses the mathematical foundation of the Lasso optimisation problem, contrasting it with ordinary least squares (OLS) regression and emphasizing its convex nature, which guarantees a unique solution. Additionally, the authors explore more complex methods such as coordinate descent for implementation and recommend cross-validation techniques for tuning the regularisation parameter lambda λ . Through various applications in fields like genomics and finance, the paper demonstrates Lasso’s effectiveness in identifying significant predictors and compares its performance favourably against Ridge regression and traditional Linear regression (TIBSHIRANIt, 1996).

This study presents a methodology to predict tourist attraction ratings, using models such as linear and Decision Tree Regression models. The motivation is to aid government agencies in identifying high potential tourist sites based on multiple attributes, facilitating informed investment decisions. Decision Tree Regression is used to analyse a dataset that includes tourist site features such as mountains, beaches, climate and other factors.

The DTR model exhibited satisfactory performance in capturing the relationships between various features of tourist sites. The performance was quantified using specific accuracy metrics such as MSE and R-squared values (Anupam Jamatia, 2019).

# RESEARCH METHODOLOGY

## Research Design

For data science projects, choosing the right project management framework is essential for effective analysis. The cross-Industry Approach for Data Mining (CRISP-DM) is ideal due to its structured and flexible approach, comprising of six stages as seen in Figure 1.

This methodology aligns well with real-life scenarios, such as the tourism analytics where understanding customer preferences and seasonal trends is crucial. In such a case, CRISP-DM offers an in-depth exploration of the business objectives, guiding the team to collect relevant data, analyse it effectively and drive actionable insights that meet stakeholder needs (Dubetcky, 2024).

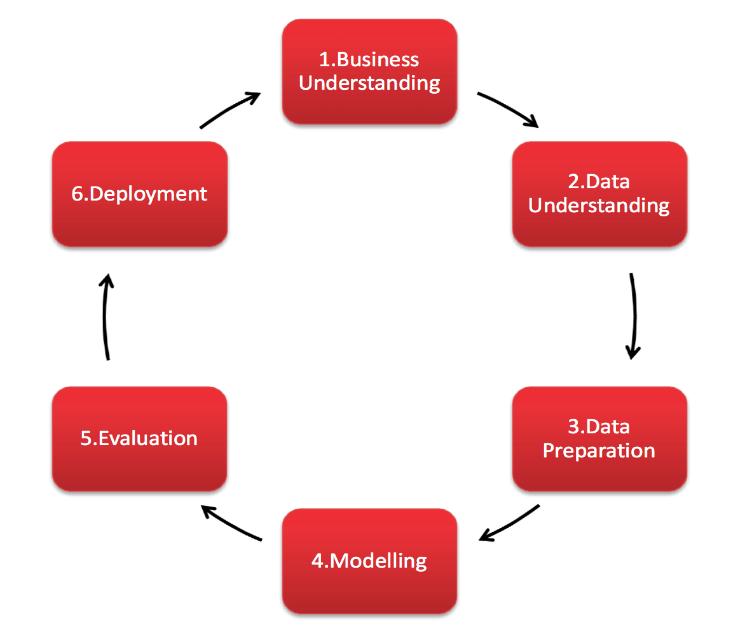


Figure : Diagram of six stages of CRISP-DM

The framework is particularly suitable for analysing tourism data, guiding the process from problem identification to actionable insights.

In comparison, the KDD (Knowledge Discovery in Databases) framework offers a broader perspective on data analysis but can be less adaptable for specific needs. The SEMMA (Sample, Explore, Modify, Model, Assess) framework is focused on technical aspects of modelling without addressing business understanding.

In selecting ML techniques, the decision depends on the nature of the dataset and the project objectives. For tourism, which includes labelled data, a supervised approach would be appropriate, this technique would allow the model to learn from historical data and make accurate predictions for future trends (Lamahewa, 2023).

## Data Understanding

The data was obtained from the central statistics office of Ireland, in total 3 datasets were used, 2 of which were obtained from the CSO central statistics office (cso, 2024) and flight data obtained from (Anon., 2024). Statistics disseminated on these sites are copyright of the government of Ireland and are provided on these sites and are accessible free of charge and licenced under creative commons attribution (version 4.0 cc-by). The configuration of the system used for this research project is ASUS ROG STRIX G513 AMD Ryzen 7 4800H, 2901 mhz, 8 core, 16 logical pr.

## Data Preparation and Cleaning

The jupyter notebook was organised to guide the reader through each step of data preparation, comments were added to justify and manipulation or changes made to the data, manipulation libraries such as Pandas Matplotlib and Numpy were utilised. Initial descriptive statistics on the CSO datasets “df1” and “df2”, revealed that all numeric values were stored within a single column, “Value”, with another column “Unit”, this structure can make analysis more complex as each row represents one measurement. The pivot operation was performed to reorganise the dataset so that the values from the “Unit” column become new columns, this also transformed the columns from datatype object to a numeric datatype.

The datasets were checked for null values and as there were no missing values, no imputation was necessary for the first 2 datasets. To prepare for merging the datasets and setting a universal index, quarters were mapped to starting months and converted to Datetime format, ensuring precise chronological order.

The third dataset was loaded and once again the “Quarter” column was manipulated and transformed from data type object to Datetime. Irrelevant columns that were created during the “Quarter” column transformation were dropped from the dataset and the new “Date” column was set as the index. Irrelevant columns such as” % Average YoY” were dropped as these were just percentage changes from the previous years and contained many missing values with no way of accurate imputation as there is no access to the previous data. Columns were then renamed to more easily interpretable names, as the data type of most of the numeric columns was type “object”, these columns were converted to datatype “float”.

The dataset “df3” was then checked for null values and it contained quite a few. The imputations were made based on temporal relationships in the data. For “%\_yoy\_flight”, backward filling was applied due to its reliance on historical yearly trends. For “%\_QoQ\_flight”, the quarter percentage change was calculated on the percentage change from the column “total\_(m)\_flight” as this was a direct relationship. For “YOY\_flight”, which also depends on previous yearly data which there was no access to, backward filling was once again utilised as the most appropriate method of imputation.

Once all missing values were imputed and the datasets were cleaned, all datasets were merged using outer joins on the “Quarter” and “Date” columns to combine all information. The “Quarter” column was renamed to “Date” for consistency, with a clean and merged dataset, descriptive statistics could now be calculated. The column “Total\_Expenditure” was created for visualisations. All columns related to direct percentage change in percent terms were dropped as the numeric change accounted for this.

## Descriptive Statistics

The final dataset contains 32 entries,11 columns, 10 of numerical datatype and the index being Datetime. The “Euro\_million” column, which represents non-Irish residents spending ,with a mean of 1051.09 and median of 997.5, indicates a reasonably symmetrical distribution, however the standard deviation (SD) is larger relative to the mean, indicating some skewness. For the “Irish\_Million” column which represents expenditure by Irish residents in Ireland, this column appears symmetrical, but the high SD relative to the mean may indicate some outliers. Columns “Euro\_nights” and “Irish\_Nights\_per\_Trip” which represent the average nights spent be non-Irish and Irish residents respectively, both have a close mean and median with small SD. Flight related variables such as “Dublin\_Total\_Flight” and “total\_(m)\_flight” have higher variability, indicating skewness.

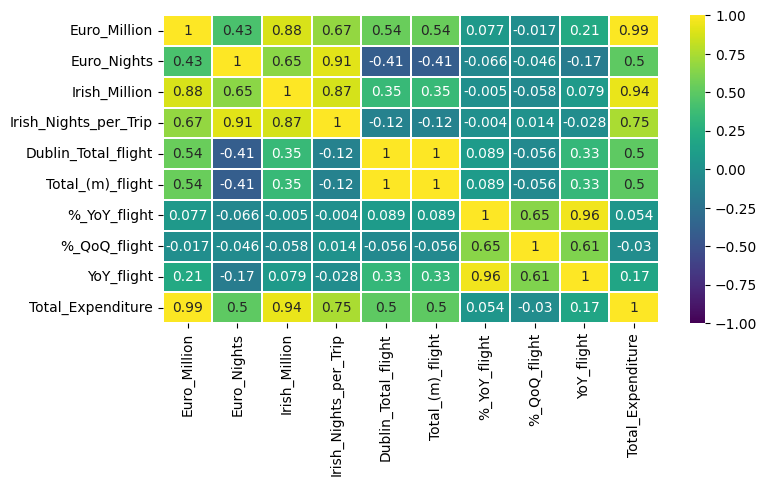


Figure :Correlation matrix of dataset

The correlation matrix as seen in figure 2 shows strong relationships among several variables in the dataset, which could indicate potential issues with multicollinearity and may be difficult to isolate the individual effect of each variable. “Euro\_million” and “Irish\_Million” have a correlation with “Total\_Expenditure” of 0.99 and 0.94 respectively, suggesting these categories may contribute redundantly to predictive models. Similarly, “Irish\_Nights\_per\_Trip” is highly correlated with “Irish\_million” at 0.87, suggesting a close relationship between trip durations and expenditure. Multicollinearity in this data might make it difficult to determine individual variable impacts, to address these issues, techniques such as dimensionality reduction or regularisation may be necessary to improve model stability and interpretability.

A screen shot of a computer

Description automatically generated

Figure :Kurtosis values for dataset

The kurtosis analysis reveals distinct distribution patterns and measures of symmetry in the data. “Dublin\_Total\_Flight” and “total\_(m)\_flight” exhibit high negative kurtosis (-1.534), indicating platykurtic distributions with lighter tails and a broader spread, meaning fewer extreme outliers. In contrast “%\_QoQ\_flight” shows positive kurtosis (1.472), reflecting a leptokurtic distribution with heavier tails and values clustering more around the centre, implying a higher frequency of extreme fluctuations. “Euro\_million” and “Irish\_Million” have moderate negative kurtosis, suggesting broader less concentrated distributions.

A group of blue and black boxes

Description automatically generated

Figure : boxplots of columns in dataset

A graph of different sizes and shapes

Description automatically generated with medium confidence

Figure : Histograms of columns in dataset

Figure 4 and Figure 5 provide insightful visualisations into the distribution of the dataset, ”Euro\_nights” ,“Euro\_million” and “Irish\_Million” , “Irish\_Nights\_per\_Trip” and “Total\_Expenditure” exhibit relatively symmetrical distributions, with values concentrated around the median and no extreme outliers. However, “Dublin\_Total\_Flight” and “total\_(m)\_flight” show right skewness indicating non-normal distributions with extended upper tails. In contrast “%\_QoQ\_flight” and “%\_YoY\_flight” exhibit higher variance and mild skewness. Overall, most distributions approximate normality, but further statistical tests, such as Shapiro-Wilk test, will be useful for a definitive answer (Gupta, 2022).

A graph of different colored lines

Description automatically generated

Figure : Line plot of expenditure from 2012-2019

Figure 6 uses a line plot to identify trends in expenditure over time. By showing the fluctuations in both Euro or Non-Irish and Irish tourists, it helps understand seasonal trends and economic changes. According to Tufts principles, the line plot is effective as it clearly conveys continuous data across time and allows for comparison of multiple data series. The principle of using simple and effective graphics helps avoid cognitive overload and aloes the viewer to interpret the data easily. From the graph, a positive linear relationship can be seen form the data as it slopes up and to the right.

A graph of different colored squares

Description automatically generated

Figure : Average number of nights spent in Ireland by non-Irish

Figure 7 displays the count plot of discretised nights spent in Ireland by non-Irish tourists. This is essential in understanding the distribution of tourist stays in terms of duration. Tuft suggests that when dealing with categorical data such as discretised continuous values, a count plot is effective because it illustrates the frequency of specific ranges.

A graph of different colors

Description automatically generated

Figure : Average number of nights spent in Ireland total

Figure 8 represents a bar plot of average total nights by year, it provides insight into whether tourists are staying longer or shorter over a period. From the data, there is a clear down trend in the data in the average number of nights spent which can be an indicator of the rising cost of tourists coming to Ireland when figure 6 is factored in.

A graph of different colored bars

Description automatically generated

Figure :Expenditure per quarter

Figure 9 displays a bar plot visualising seasonal and annual expenditure patterns. By splitting the data by quarter, it reveals peak seasons for expenditure in Ireland, which is late summer early autumn, the data has a positive upward trend, indicating increased growth yearly.

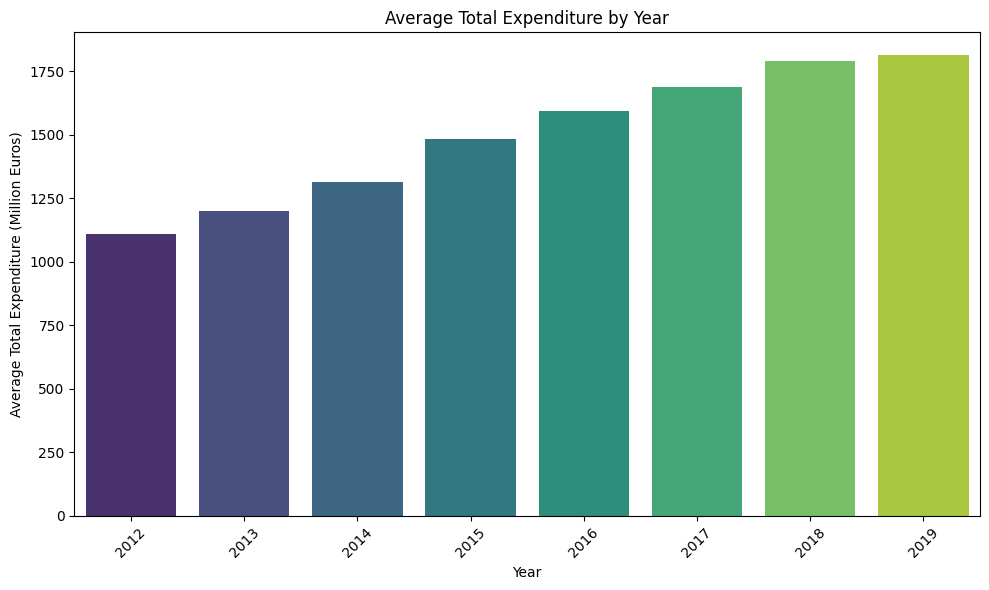
****

Figure : Expenditure per year.

Figure 10 displays yearly total expenditure from tourism in Ireland, allowing easy identification of trends and anomalies. From the data, once again there is a clear positive linear relationship between the data.

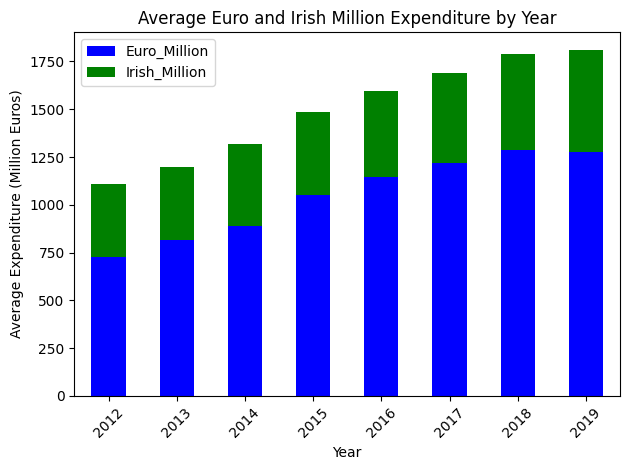
****

Figure :Expenditure for Irish and Non-Irish

Figure 11 displays a stacked bar plot to show total contributions to expenditure. This visualisation helps to understand how total spending evolves over time.

The visualisations adhere to Tufts principles by prioritising clarity, integrity and focus. Bar and line plots simplify data interpretation without unnecessary complexity. Data is accurately represented, avoiding misleading visuals.

# Statistical Analysis through Distributions

## Distributions

A graph and a diagram

Description automatically generated with medium confidence

Figure :KDE Histogram and Q-Q plot

The normal distribution is central to understanding the overall trends in expenditure as it is a continuous variable. Figure 12 shows Histogram with KDE Plot gives an initial look at the distributions shape.

The histogram appears to show a distribution that has slight skewness, which may indicate some deviation from normality. The KDE overlay provides more smoothness to the data visualization and suggests that the data may have slight right skewness. Additionally, a Q-Q plot indicates that most points align with the diagonal line, confirming approximate normality.

To formally test for normality the Shapiro-Wilk test is a formal statistical test used to determine whether a sample comes from a normal distribution. The test provides a test statistic and a p-value.

Null Hypothesis (𝐻0H0): The data follows a normal distribution.

Alternative Hypothesis (𝐻1H1): The data does not follow a normal distribution.

The test outputs are as follows:

Shapiro-Wilk Test Statistic: 0.940

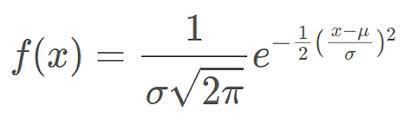
p-value: 0.074

Interpretation

Since the p-value (0.074) is greater than the typical significance level (𝛼=0.05)

we fail to reject the null hypothesis.

This result suggests that there is insufficient evidence to conclude that the distribution of “Total\_Expenditure” significantly deviates from normality.



Equation :PDF normal distribution

The Total Expenditure data exhibits a positive linear relationship, suggesting a steady growth in expenditures over time. By leveraging the mean as a central reference, thresholds based on the distribution can monitor normal fluctuations in expenditure. Deviations above or below these thresholds reveal insights into typical behaviour versus outliers. Tracking expenditures below the mean could indicate areas for improvement, while values above the mean could highlight successful strategies or peak periods.

The probability that total expenditure falls below 1400 is approximately 43%, this suggests a significant portion of time in which total expenditure is lower than the mean. Such instances might correlate with off peak tourism. There is a 36% chance that expenditure will exceed 1600, indicating periods of higher spending. Targeted efforts to identify these periods could see increased overall revenue.

A graph of a number of nights

Description automatically generated

Figure :Probability distribution of number of nights

A graph of a number of numbers

Description automatically generated

Figure :Example of CLT

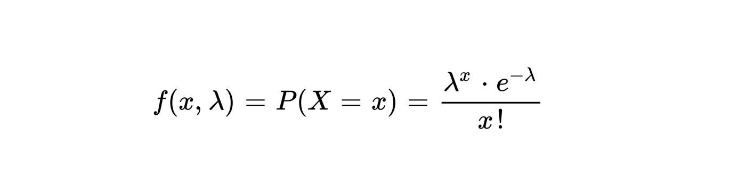
Figures 13 and 14 demonstrate a binomial distribution, to understand tourist behaviour based on nights spent, defining success as tourists who spent above 10 nights, the mean being 9.5. Feature engineering created a binary “success” column to represent long-staying tourists, with an observed probability of success(p=0.25), indicating that 25% of tourists significantly contribute to the Irish economy.

With 32 trials (n=32), eight tourists met the threshold. Binomial probabilities show that the likelihood peaks around 8-10 successes as seen in Figure14. To explore convergence toward a normal distribution, larger sample sizes were examined, visualising probabilities using a PMF function. As expected, the distribution increasingly resembles a normal distribution as n grows.

A graph of a number of flights arriving

Description automatically generated

Figure :Poisson Distribution



Equation :Poisson distribution formula

Figure 15 displays a Poisson Distribution that is applied to model arrival counts based on the assumptions that events are independent, discrete and count based. The average λ of 497 as the parameter was calculated, the original data was quarterly but was then calculated to be weekly arrivals. The probability of more than 500 arrivals per week is calculated at 44%, this would mean 50,000 people arriving weekly as the value was truncated. This arrival data would be at the forefront of policy planning, to get an estimate of the number of people entering Ireland, which then correlates to nights spent and expenditure.The probability of different arrival counts was calculated. Using the CLT, 1000 sample means were generated to demonstrate how the sample means approximate a normal distribution as sample size grows as seen in Figure 16.

A graph of a number of blue bars

Description automatically generated

Figure :CLT applied to Poisson

The analysis combines statistical modelling of nights stayed arrival counts and expenditure using appropriate distributions and methods.

# Machine Learning

## Feature Selection and Preprocessing

From the visualisation of the data, it is evident that there is an approximately linear relationship between the continuous data. Therefore, the first model will be simple linear regression. Linear regression has several underlying assumptions, such as normality of data, linearity of residuals, homoscedasticity and independence or lack of multicollinearity. As the features met most of these assumptions, the feature selection was based on their strong correlation with the target variable.

From the correlation matrix in figure 2, it can be observed that some features such as “Euro\_million”, and “Total\_Expenditure” exhibit high correlation with the target “Irish\_million” making them significant predictors. Other features such as “Euro\_nights”, “Irish\_Nights\_per\_trip” and “Dublin\_Total\_flight” provide valuable information with moderate correlations indicating relevant relationships. “Irish\_million” was chosen as the dependent variable because it represents a key financial metric in the dataset, specifically addressing domestic tourism for Ireland.

No encoding was necessary for the models as there are no categorical variables simplifying the preprocessing steps.

The dataset contains features which have varying scales, to prepare the data for machine learning models, StandardScaler will be implemented as it is the most appropriate given the approximately normal distribution of the data. Given the data’s continuous features and potential outliers, Standardscaler is preferred as other scalers could distort the scaling due to extreme values. The dataset was then split into the standard split of 80% training and 20% testing. (Müller, 2020)

## Modelling

As the dependent variable is continuous, the dataset small and the data visualisations showing an approximately linear trend, the first choice for regression is the Linear regression model from scikit-learn.

A screenshot of a black screen

Description automatically generated

Figure :Training scores table

A screenshot of a graph

Description automatically generated

Figure :Test scores table

The model achieved a training score of 0.99 and a test score of 0.91, training R2 score of 0.99 indicates that the model explains nearly all the variance in the training data, and the test R2 score of 0.91 indicates that the model generalises well to unseen data evidenced in figures 18 and 19.

The coefficients identified “Dublin\_Total\_flight” and “Irish\_Nights\_per\_Trip” as having a strong positive relationship reinforcing the significance of these metrics as predictors.

A graph with blue squares

Description automatically generated

Figure :Linear Regression coefficients

In evaluating the performance of the model, significant differences were observed between the training and test metrics. The training MAE 15.28 and test MAE 44.03 suggest the model performs significantly better on the training data. Similarly, MSE and RMSE are much higher on the test set, indicating large errors on unseen data. The training R2 is 0.986 but drops to 0.907 on the test set , signifying a decrease in the model’s ability to explain variance when applied to unseen data. Cross Validation with 5 folds was applied and the results range from 0.98 to 0.72 with a mean of 0.88, this indicates that the model fluctuates between folds. The variation in fold scores suggests that the model may be sensitive to certain features or subsets of data.

A graph with blue and green dots

Description automatically generated

Figure 20: Ridge Coefficients

Figure 20 displays a visual assessment of the models prediction errors on the training and test sets, from the graph we can identify the one outlier near 700 that lowers the test score.

Lasso was selected as the next model due to its ability to perform both regularisation and feature section and it exhibited similar metrics to the Linear model. Lasso achieved a training score 0.98 slightly more generalised than Linear and scored 0.90 on the test set. Lasso’s MAE and RMSE were similar indicating that both models fit the training data well.

A graph with blue squares

Description automatically generated

Figure 21:Lasso Coefficients

While the models performed similarly on metrics, the coefficients were quite different, Lasso regularises by shrinking the coefficients zeroing out “Total\_(m)\_flight” indicating irrelevance as seen in figure 20.

A graph of a bar graph

Description automatically generated

Figure 22: Ridge Coefficients

Ridge regression was utilized next, unlike Lasso which eliminates irrelevant features, ridge applies a penalty to the magnitude of the coefficients, preventing extreme values. After hyperparameter tuning, the model achieved an alpha value of 0.1. This resulted in similar test and training scores as previous models.

A graph of a line

Description automatically generated

Figure : Residuals for Lasso

Figures 20 and 21 analyse the residuals for the two regularisation models, both models display similar range of values centered on 0. The figures show that both models gave similar results.

A graph with blue lines

Description automatically generated

Figure Residuals for Ridge

The results from the DT modes, both show the base and the hyper tuned versions, show a significant difference in performance when compared to Linear, Lasso and ridge. The DT model even after hypertuning and feature selection, Figure22, performs poorly, particularly on the test set where the R2 score is negative. This indicates that the model struggles to generalise on unseen data, likely due to overfitting on the training set, issues with the results indicate the problem may lie in the data.

A graph with blue squares

Description automatically generated

Figure :Top features for DT

## Model results and Conclusion

The models displayed varying performance, with Linear Regression achieving the best balance between training and test scores. However, it showed significant overfitting, as seen in the higher MAE and RMSE for the test set. Lasso performed similarly but was more effective in feature selection. Ridge regression provided great results too but lacked the same features reduction as Lasso. The DT model, despite hyperparameter tuning, struggled with overfitting, as overfitting was evident in all models, this highlights the need for more data.

# Programming and Code quality

Throughout this project procedural and declarative programming were utilised, the project broke down complex tasks into manageable steps, which greatly improved both the clarity and maintainability of the code. Using procedural programming allowed complex tasks to be break down into manageable steps, enhancing the clarity and maintainability of the code, especially in repetitive tasks like feature engineering and data visualization. However, as the project grew, limitations arose , especially with duplicated code, which became harder to maintain. In a larger-scale project, object-oriented programming (OOP) would have helped to encapsulate these repetitive processes into reusable components, promoting cleaner and more modular code (Shaibu, 2024).

Following best practices, the main naming convention used was snake\_case, camelCase is generally used for oop. Proper spacing and bracketing was essential for writing clear readable and maintainable code. For creating hyperparameter variables or other variables, all were stored in appropriate containers, either a dictionary or a list, depending on use case down the line.

For this project all constants were made with a global scope, as there were few functions defined, allowing the constants to be globally scoped was appropriate. Variable naming is a difficult part of any data science project, names were given precise and meaningful names to limit confusion.

Data preprocessing was kept to a minimum to try and preserve the original structure of the data. Loops are essential constructs, and all loops were kept to the first level, as in no nested for loops.

The concept DRY (Don’t repeat Yourself) was adhered to as best as possible, with though given to many procedures and code, benefits include reduced errors, easier maintenance and enhanced performance.

Functions that were created and designed to return one and only one task, appropriate docstring were attached for clarity. Exception handling was utilised during the project to ensure correct datasets were uploaded, this is crucial part of any data science project (Andrenacci, 2024).

Furthermore, functional programming principles guided the project in designing functions with single responsibilities, which improved readability and maintainability. Overall, the project highlighted the strengths and limitations of procedural programming and recognized that OOP principles like inheritance and encapsulation could further streamline and scale this approach in more complex projects.

# Bibliography

Andrenacci, G., 2024. *Medium.* [Online]   
Available at: https://medium.com/data-bistrot/python-procedural-programming-best-practices-bc20ffeb6e37#:~:text=Procedural%20Programming%20with%20Focus%20on,Data%20analysis%20and%20Machine%20learning&text=Procedural%20programming%20is%20a%20foundational,to%20achieve%20  
[Accessed 08 11 2024].

Anon., 2024. *https://data.smartdublin.ie/dataset/dublin-economic-monitor/resource/f6297d1e-2d25-4a35-a499-164f24511dd7?inner\_span=True.* [Online]   
Available at: https://data.smartdublin.ie/dataset/dublin-economic-monitor/resource/f6297d1e-2d25-4a35-a499-164f24511dd7?inner\_span=True  
[Accessed 05 11 2024].

Anupam Jamatia, U. B. S. P. S. D. &. S. D., 2019. Rating Prediction of Tourist Destinations Based on Supervised Machine Learning Algorithms. *Springer ,* Volume 990, p. 10.

cso, 2024. *cso.* [Online]   
Available at: https://data.cso.ie/  
[Accessed 05 11 2024].

Dubetcky, O., 2024. *Medium.* [Online]   
Available at: https://oleg-dubetcky.medium.com/project-management-for-data-science-kdd-semma-and-crisp-dm-fe9d03d3ab6c  
[Accessed 03 11 2024].

expert.ai, 2022. *expert.ai.* [Online]   
Available at: https://www.expert.ai/blog/machine-learning-definition/  
[Accessed 01 11 2024].

failte, n.d. *https://www.failteireland.ie/.* [Online]   
Available at: https://www.failteireland.ie/  
[Accessed 02 11 2024].

Gössling, S., 2020. Technology, ICT and tourism: from big data to the big picture. *research gate ,* 29(7815), p. 10.

Gupta, S., 2022. *Medium.* [Online]   
Available at: https://towardsdatascience.com/understanding-descriptive-statistics-with-the-help-of-gaussian-left-skewed-and-right-skewed-897b296382c6  
[Accessed 05 11 2024].

itic, 2019. *itic.* [Online]   
Available at: https://www.itic.ie/YE19/index.html  
[Accessed 03 11 2024].

Lamahewa, A., 2023. *Medium.* [Online]   
Available at: https://anujalamahewa.medium.com/supervised-unsupervised-andsemi-supervisedlearning-70095952e2f4  
[Accessed 04 11 2024].

Mayooran Thevaraja1, M. G. A. R., 2019. *Recent Developments in Data Science:,* s.l.: s.n.

Müller, A. C., 2020. *https://amueller.github.io/aml/01-ml-workflow/03-preprocessing.html.* [Online]   
Available at: https://amueller.github.io/aml/01-ml-workflow/03-preprocessing.html  
[Accessed 08 11 2024].

Pandita, I. S., 2022. *Medium.* [Online]   
Available at: https://medium.com/@emperorarthurix/machine-learning-6c26aa789b9f  
[Accessed 02 11 2024].

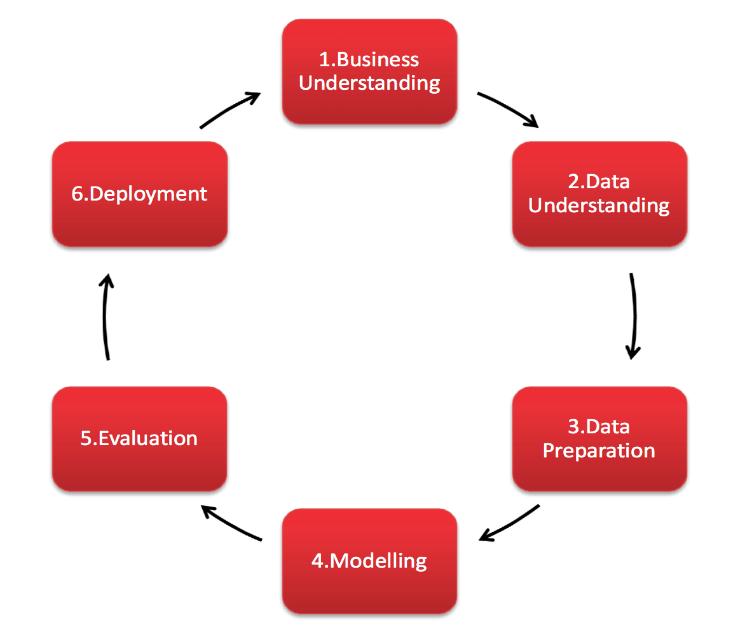
Shaibu, S., 2024. *DataCamp.* [Online]   
Available at: https://www.datacamp.com/blog/introduction-to-programming-paradigms  
[Accessed 08 11 2024].

TIBSHIRANIt, R., 1996. *Regression Shrinkage and Selection via the Lasso,* toronto: watermark.silverchair.

UN, 2008. *https://www.unwto.org.* [Online]   
Available at: https://www.unwto.org/glossary-tourism-terms  
[Accessed 03 11 2024].

Wilder-James, E., 2012. *https://www.oreilly.com.* [Online]   
Available at: https://www.oreilly.com/radar/what-is-big-data/  
[Accessed 02 11 2024].

# Appendix



Appendix 1: Diagram of six stages of CRISP-DM

A chart of numbers and symbols

Description automatically generated with medium confidence

Appendix 2:Correlation matrix of dataset

A screen shot of a computer

Description automatically generated

Appendix 3:Kurtosis values for dataset

A group of blue and black boxes

Description automatically generated

Appendix 4: boxplots of columns in dataset

A graph of different sizes and shapes

Description automatically generated with medium confidence

Appendix 5: Histograms of columns in dataset

A graph of different colored lines

Description automatically generated

Appendix 6: Line plot of expenditure from 2012-2019

A graph of different colored squares

Description automatically generated

Appendix 7: Average number of nights spent in Ireland by non-Irish

A graph of different colors

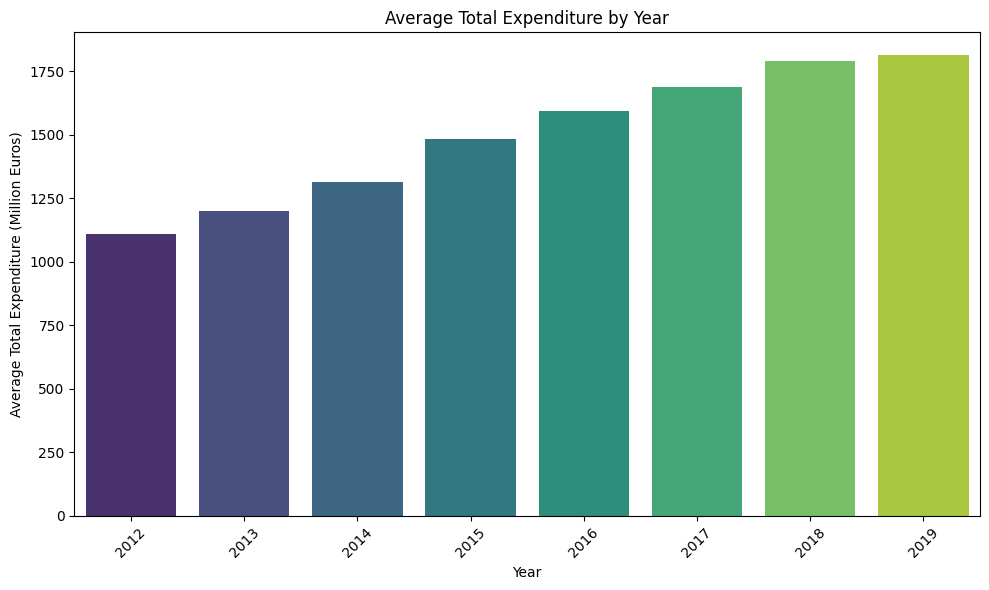
Description automatically generated

Appendix 8: Average number of nights spent in Ireland total

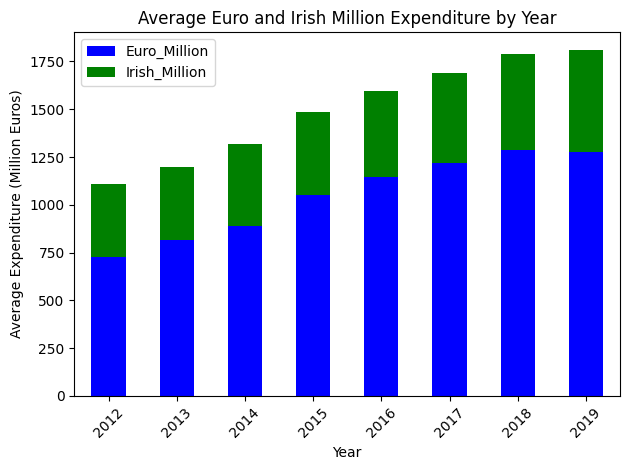
A graph of different colored bars

Description automatically generated

Appendix 9:Expenditure per quarter

****

Appendix 10: Expenditure per year.

****

Appendix 11:Expenditure for Irish and Non-Irish

A graph and a diagram

Description automatically generated with medium confidence

Appendix 12:KDE Histogram and Q-Q plot

A graph of a number of nights

Description automatically generated

Appendix 13:Probability distribution of number of nights

A graph of a number of numbers

Description automatically generated

Appendix 14:Binomial application of CLT

A graph of a number of flights arriving

Description automatically generated

Appendix 15:Poisson Distribution

A screenshot of a black screen

Description automatically generated

Figure 17:Training scores table

A screenshot of a graph

Description automatically generated

Figure 18:Test scores table

A graph with blue squares

Description automatically generated

Appendix 19:Linear Regression coefficients

A graph with blue and green dots

Description automatically generated

Appendix 20: Ridge Coefficients

A graph with blue squares

Description automatically generated

Appendix 21:Lasso Coefficients

A graph of a bar graph

Description automatically generated

Appendix 22: Ridge Coefficients

A graph of a line

Description automatically generated

Appendix 20: Residuals for Lasso

A graph with blue lines

Description automatically generated

Appendix 21Residuals for Ridge

A graph with blue squares

Description automatically generated

Appendix 22:Top features for DT