# Group ID - MSc in Data Analytics

Author: Dominic Nnamdi Nwachukwu

e-mail: s[ba24103@student.cct.ie](mailto:ba24103@student.cct.ie)

Student ID: sba24103

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | MSC\_DA\_CA1 |
| **Assessment Title:** | PREDICTION OF TOURISM VALUE IN IRELAND USING MACHINE LEARNING |
| **Lecturer Names:** | *Taufique Ahmed/Cristiani Eccher*  *Kayoum Khbuli*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Dominic Nnamdi Nwachukwu |
| **Student Number:** | Sba24103 |
| **Assessment Due Date:** | 8th November 2024 |
| **Date of Submission:** | 7th November 2024 |

**Declaration**

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

**EXECUTIVE SUMMARY**

The primary purpose of this study is to evaluate the factors affecting tourism demand in Ireland and predict tourism value in Ireland over time. In this context, secondary quantitative data has been collected from the Central Statistics Office (CSO) related to the tourism sector in Ireland. Based on the collected data, statistical analysis (including discrete and continuous distribution, EDA, summary statistics, and supervised machine learning models (Regression)) was performed using Python programming in Jupyter Notebook.

The obtained accuracy of the Random Forest regression model (without and with hyperparameter optimisation using GridSearchCV method) is slightly higher than the Bagging regression and Gradient Boosting regression model. This reflects high applicability of the Random Forest regression model in the prediction of tourism value in Ireland.

**TABLE OF CONTENTS**

[1. Introduction 4](#_Toc181093910)

[1.1 Background context 4](#_Toc181093911)

[1.2 Aim and Objectives 4](#_Toc181093912)

[1.3 Questions 5](#_Toc181093913)

[1.4 Relevance of the project with Data Analytics 5](#_Toc181093914)

[2. Criteria of Analysis 6](#_Toc181093915)

[2.1 Statistics 6](#_Toc181093916)

[2.1.1 Summarisation of dataset 6](#_Toc181093917)

[2.1.2 Evaluation of the Dataset using Discrete Distributions 10](#_Toc181093918)

[2.1.3 Evaluation of information of the dataset using the concept of Normal distribution 13](#_Toc181093919)

[2.1.4 Explanation of Discrete Distributions 13](#_Toc181093920)

[2.2 Data Preparation and Visualisation 14](#_Toc181093921)

[2.2.1 EDA 14](#_Toc181093922)

[2.2.2 Data preparation for ML modelling 15](#_Toc181093923)

[2.2.3 Visualisations (Tufte Principle) 20](#_Toc181093924)

[2.3 Machine Learning for Data Analytics 24](#_Toc181093925)

[2.3.1 Project Management Framework: SEMMA 24](#_Toc181093926)

[2.3.2 Programming paradigms 24](#_Toc181093927)

[2.3.3 Supervised ML Modelling: Regression 25](#_Toc181093928)

[2.3.4 Model comparisons 32](#_Toc181093929)

[2.3.5 Logical interpretation 32](#_Toc181093930)

[3. Conclusion 33](#_Toc181093931)

[References 34](#_Toc181093932)

[Appendices 37](#_Toc181093933)

[Appendix 1: Box plot of VALUE 37](#_Toc181093934)

# 1. Introduction

## 1.1 Background context

The United Nations World Tourism Organisation (UN-WTO) refers to ‘tourism’ as the fundamental activities of the population travelling to and staying in places outside their usual environment. Tourism is an essential measure of the economic activity of a state that drives employment in hospitality and food (Central Statistics Office, 2023). Thus, the revenue generated through domestic and foreign visitors during the travel trip contributes adequately to both payment and national accounts. Likewise, tourism in Ireland contributes significantly (approximately 8%) to the total GDP of Ireland, with more than 6.3 million foreign visitors and a total spending of €7.3 billion in 2023 (Statista, 2024). Ireland's rich cultural heritage and natural landscape have contributed significantly to the growth of the travel and tourism industry. Within this study, the main motive is to evaluate the market value of Irish tourism through the implication of statistical analysis and Machine Learning modelling.

## 1.2 Aim and Objectives

Aim:

This study aims to evaluate the demand (value) for tourism services in Ireland using Statistical analysis and Machine Learning models.

Objectives:

* To explore the changes in demand for different tourism services (inbound and outbound) in Ireland
* To perform probabilistic distribution (Binomial and Gaussian distribution) to explore the distribution of the value of different statistics levels in Irish tourism (such as number of Trips by Irish Residents, number of nights stayed by travellers and length of stay)
* To perform data preprocessing (treatment of missing and duplicate records) and transformation (such as encoding for categorical features) for preparing the data for statistical analysis and ML modelling
* To develop supervised machine learning models (Regression) (Bagging regressor, Random Forest and XGBoost Regressor) to estimate the value of Irish tourism
* Perform hyperparameter optimization using the GridSearchCV method to fine-tune Regression models' parameters and increase the machine learning models' predictive accuracy.

## 1.3 Questions

* How has the value of inbound and outbound tourism in Ireland changed?
* What is the implication of Data Analytics in evaluating data distribution related to tourism value?
* What is the significance of ML modelling in predicting tourism value in Ireland?
* What is the significance of Hyperparameter Optimisation using GridSearchCV technique for optimising predictive performance of the ML models (Regression models)?

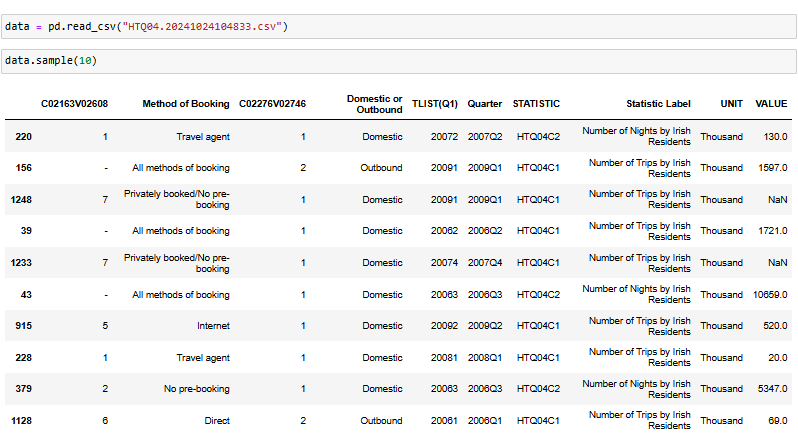
## 1.4 Relevance of the project with Data Analytics

The execution of this study is closely linked with the concept of data analytics as this study leverages statistical analysis (probability distribution, descriptive statistics and visualisations) and Supervised ML techniques to understand and predict demand dynamics in the Irish tourism sector. Through probabilistic distribution analysis (using Binomial and Gaussian distributions), the project intends to reveal the variability and patterns in key tourism metrics, like the number of trips by Irish residents and travellers' length of stay. This can be helpful in understanding trends in inbound and outbound tourism in Ireland, which can allow businesses in the hotel and hospitality sector to optimise their resources to serve a maximum number of visitors. Moreover, supervised ML modelling (including Bagging Regressor, Random Forest Regressor and Gradient Boosting Regressor) to predict the value of Irish tourism.

# 2. Criteria of Analysis

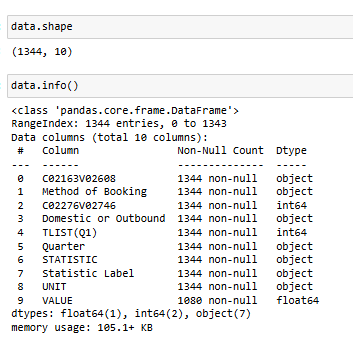
## 2.1 Statistics

### 2.1.1 Summarisation of dataset

******

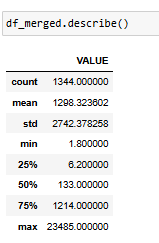
***Figure 1: Loading the dataset***

The ***‘HTQ04.20241024104833.csv’*** dataset has been loaded into the Python environment using the ‘read\_csv ()’ method from the pandas library, as the collected dataset from the ***‘Central Statistics Office’*** database was in CSV format (***Refer to Figure 1***).

******

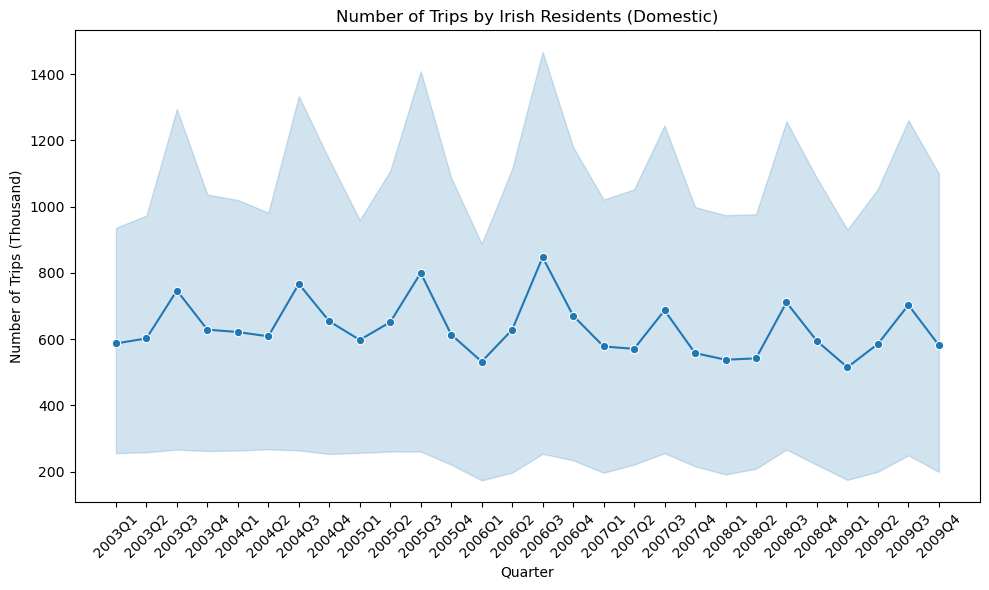
***Figure 2: Shape and information of the dataset***

Shape of the dataset has been checked using the ‘shape’ method from pandas to explore the number of rows and columns in the dataset. As per the viewpoint of Stańczyk, Zielosko and Baron (2024), exploration of shape and data info is important within the data analytics process as it helps in understanding data types and size of the dataset, which can help in the selection of appropriate statistical methods. ***Figure*** ***2*** shows that the dataset contains 1344 observations and 10 variables, out of which 3 are numerical variables (indicated by int64 and float64) and 7 are categorical variables (object).

******

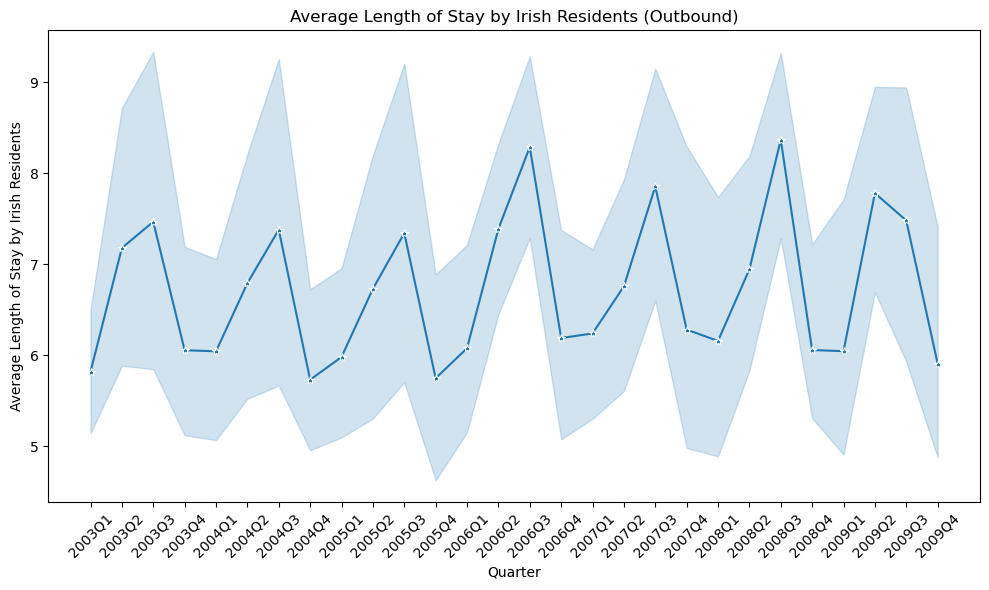
***Figure 3: Summary statistics***

The summary statistics for the variable ‘***demand of tourism (VALUE)’*** provide detailed insights into the distribution and central tendency (***Refer to Figure 3***) (***Refer to Appendix 1***). The mean value of tourism demand is 1298.32, which indicates a significantly high demand for tourism in Ireland. The report published by the Irish Examiner revealed that the scenic beauty and natural landscape of Ireland have led to higher visitor footfalls, where 14% of travellers visit Ireland to discover newness (Murray, 2024). High standard deviation of 2742.37 reflects high variations in ‘VALUE’ due to variability in demand of inbound and outbound tourism mainly due to seasonal demand of tourism in Ireland.

******

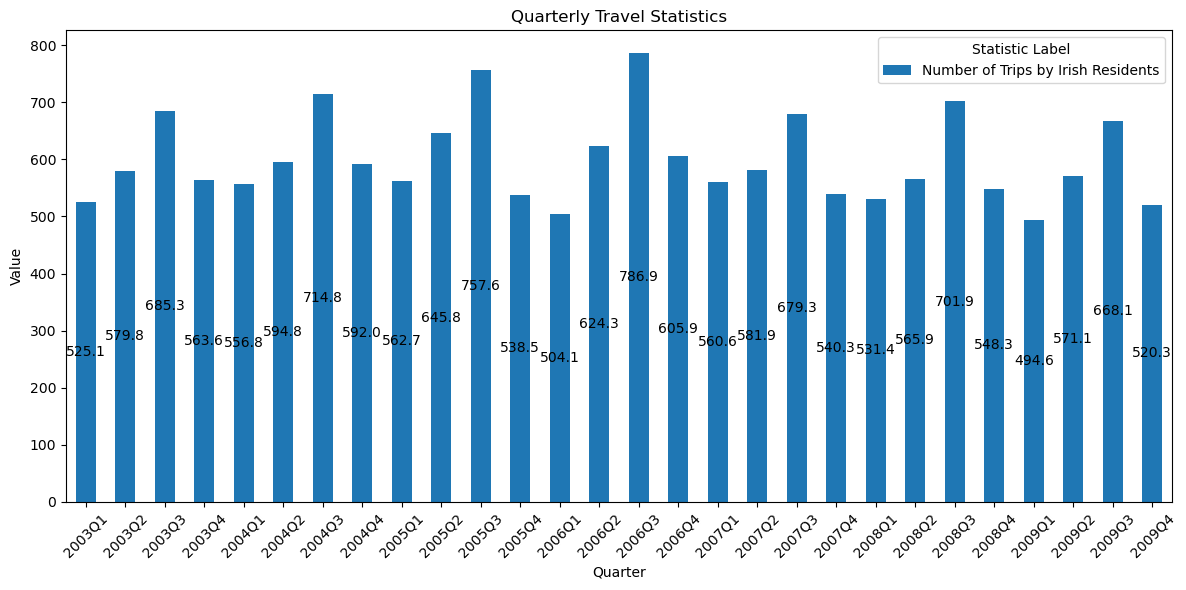
***Figure 4: Number of trips by Irish residents (domestic)***

The line plot shows the variations in number of trips by Irish residents over time, from which it can be observed that there was a low fluctuation in the number of trips (domestic tourism) in Ireland. It has been found that over the past 15 years, inbound tourism in Ireland has slowly increased by more than 27%, justifying the continuous growth or lower fluctuations in the number of trips (Irish Tourism Industry Confederation, 2023). ***Figure 4*** shows a constantly high number of trips made by Irish residents over time, which is indicated by trends in trips (***Refer to Figure 4***).

******

***Figure 5: Average length of stay by Irish residents***

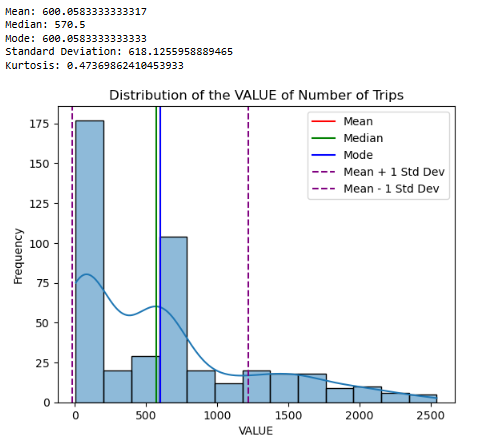
The length of stay by Irish residents has shown a high fluctuation in the average length of stay by Irish residents with outbound tourism (***Refer to Figure 5***). Irish residents often take longer trips during holiday seasons, such as summer or around public holidays (Wang et al., 2024). This leads to substantially high variations in the average length of stay in Ireland over time.

******

***Figure 6: Quarterly travel statistics***

Quarterly travel statistics (for the number of trips in Ireland for inbound and outbound tourism) have been shown using a bar plot. From this, it can be observed that the number of trips in Ireland follows constant trends, which can be due to the low level of international tourist visits.

### 2.1.2 Evaluation of the Dataset using Discrete Distributions

******

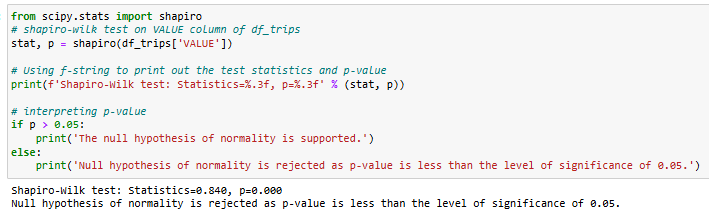
***Figure 7: Distribution of the Value of Number of Trips***

The distribution of Value of the number of trips in Ireland has followed a right-skewed distribution with a mean of 600.05 and a standard deviation of 618.12, reflecting high variability in the distribution (***Refer to Figure 7***). The distribution highlights variability and asymmetry in the number of trips taken by Irish residents, with most data points concentrated in the lower range and a few instances of much higher numbers. This shows that the value of the number of trips in certain regions (such as Dublin, the South West, and the West) is considerably higher than in other regions, which has potentially led to a right-skewed distribution. The high variability (as seen from the widespread standard deviation) suggests that while many take fewer trips, a subset of residents travels significantly more, causing a right-skewed distribution.

The concept of the Central Limit Theorem states that for a large sample size (n = 1000), both Binomial and Poisson distribution are approximated to Normal distribution. Thus, the normality of the distribution (Number of Trips) has been checked using the Normality test (Shapiro-Wilk test).

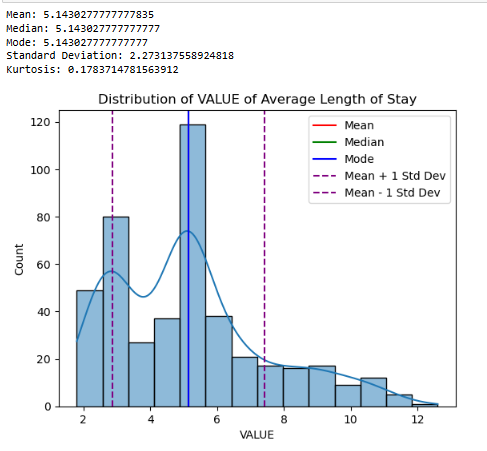
H0: The variable ‘Number of Trips’ has come from a normal distribution.

Ha: The variable ‘Number of Trips’ has not come from a normal distribution.

******

***Figure 8: Normality test for Number of Trips***

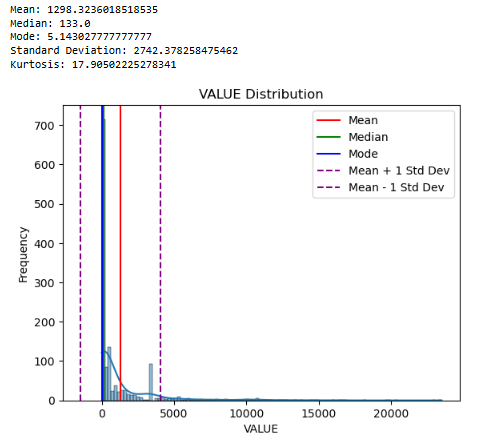
The results obtained from the Shapiro-Wilk Normality test (0.840 with a p-value = 0.000) reject the null hypothesis indicating that the data has not followed a normal distribution, which is true as this is a real dataset. This suggests that Gaussian techniques may not be suitable for this data due to its high variability.

******

***Figure 9: Distribution for Average length of stay of visitors in Ireland***

***Figure 9*** shows the distribution of average length of stay of visitors in Ireland, from which it can be observed that it has followed an approximated normal distribution with a nearly equal mean, median and mode (5.14). The distribution of the average length of stay shows that most trips taken by Irish residents are concentrated around the 5-day mark. It has been observed that the majority of travellers in Ireland prefer shorter vacations, often between 3 to 7 days (RTÉ News, 2024). This is preferably due to a 5-day trip offering a perfect balance for those looking to take a break from routine without requiring extensive time off work or significant planning.

### 2.1.3 Evaluation of information of the dataset using the concept of Normal distribution

******

***Figure 10: Distribution of ‘VALUE’***

***Figure 10*** shows the normal distribution for the variable ‘VALUE’, which shows a slightly right-skewed distribution. According to Brereton (2014), the concept of normal distribution of a variable can be characterised as the three parameters (mean, median and mode) centre around the mean. In this study, the mean (1298.32), median (133.0) and mode (5.14) have shown high variations, reflecting that the variable has deviated significantly from the assumption of normality.

### 2.1.4 Explanation of Discrete Distributions

#### 2.1.4.1 Insights captured from distribution

The distribution of ***‘Number of Trips’*** can be considered as essential for evaluating the extent to which individuals or groups travel within Ireland, which reflects the variability in travel. The right-skewed distribution of ***‘Number of Trips’***, where majority of values are lower with a few very high ones, provides a realistic view of travel patterns in Ireland. For instance, high-frequency travel in regions like Dublin and the Southwest reflects that these are popular or economically dynamic areas where people either visit more often or there are more residents with higher travel habits. Additionally, the continuous distribution of the variable ***‘Average length of stay’*** provides insights regarding the common travel behaviour of travellers in Ireland. The observation of normal distribution for this variable indicates the travel duration of tourists in Ireland is approximately 4-5 days.

#### 2.1.4.2 Justification of choice of variables

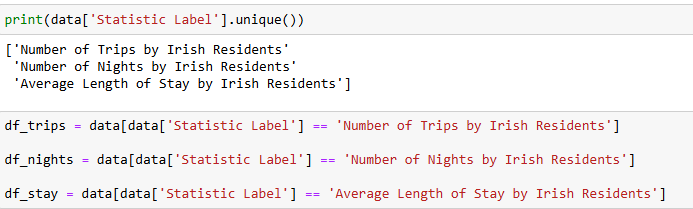
The consideration of the variable ‘Number of Trips’ indicates the overall frequency of travel, which fundamentally varies considerably based on variation in region. As per the viewpoint of Malichová, Cornet and Hudák (2022), exploration of frequency of trips as an indicator of travelling behaviour allows evaluation of influence of travelling behaviour on the growth of the tourism sector. Therefore, the evaluation of ‘number of trips’ is crucial in the exploration of travelling frequency of travellers, which can offer a discrete perspective on travel behaviour. Additionally, the variable ‘length of stay’ captures the duration trend of travellers, which can be beneficial for exploration of preferences of visitors in Ireland in staying at tourist places.

## 2.2 Data Preparation and Visualisation

### 2.2.1 EDA

The section involves data preparation for ML modelling and EDA to visualise the patterns and relationships of the features. Exploratory data analysis is a significant step of data-driven research, which explains the possible structure of the data, trends, and hidden patterns, guiding the selection and tuning of subsequent models (Thanos et al., 2023). The study involves EDA to lay foundational insights into Irish tourism data, prepared so that the best method to prepare the data is represented towards its possible application in ML and statistical analysis. All the crucial transformation steps include converting the VALUE column into categorised statistics by types of trips, nights stayed, and length of stay for targeted insights of tourists. This structured approach to EDA has helped clarify trends in tourism demand and has provided beneficial inputs into forecasting of tourists and the dynamics of the tourism sector of Ireland.

### 2.2.2 Data preparation for ML modelling



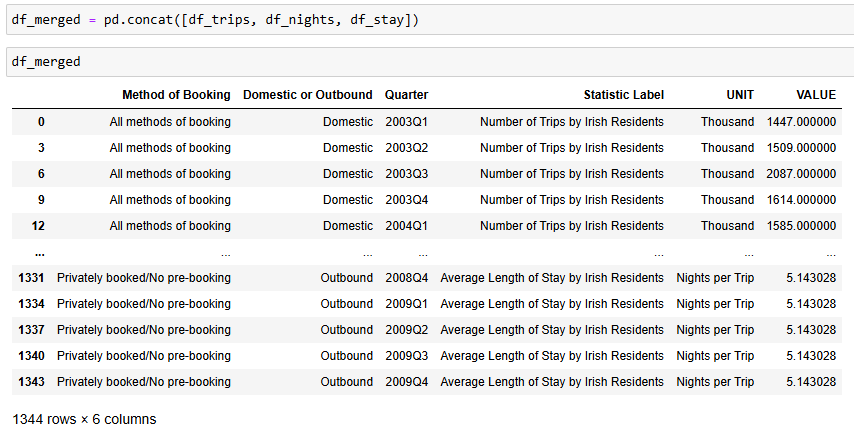
***Figure 9: Segmenting the 'VALUE column into three different columns based on Statistic Label***

***Figure 9*** shows the VALUE column divided into three columns depending on the Statistical Label, such as the Number of trips, Nights, and Average Length of Stay by Irish Residents, to simplify the data. Data under different labels can be matched and categorised with unique tourism metrics in that precise analysis, enabling the development of insights on travel behaviour and tourism demand in Ireland.

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

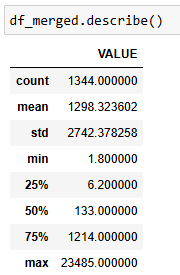
***Figure 10: Handling Null Values by Mean Imputation***

***Figure 10*** helps to ensure that the integrity of the results is preserved by addressing missing values in the VALUE column of the three segmented data frames using mean imputation. Addressing missing values by replacing mean values (mean imputation) confirms data reliability for further analysis, which helps reduce biases in the data (Jadhav, Pramod, and Ramanathan, 2019). This process filled all 88 missing entries for each data, making complete datasets available to enhance the reliability of the insights into demand for Irish Tourism.

******

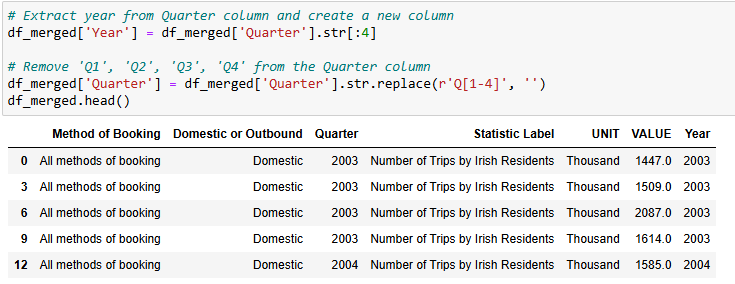
***Figure 11: Concatenating the three data frames***

***Figure 11*** shows the segmented data frames concatenated into one main data to make the entire structure more suitable for further analysis. This aggregation easily allows for an all-encompassing analysis, providing an overview of tourism statistics, such as the number of trips, the number of nights spent in the states, and, on average, the length of stay, and, therefore, more powerful insights into tourism demand in Ireland.

******

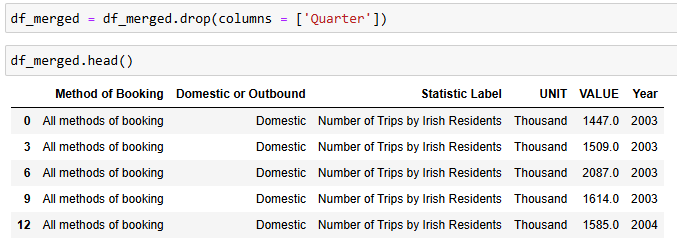
***Figure 12: Descriptive Statistics***

The descriptive statistics of merged data in ***Figure*** ***12*** give a detailed overview of the Irish tourism metrics. The mean was around 1298, and the standard deviation went up to 2742, emphasising high volatility. This helps to understand that Ireland has, over the years, seen fluctuation in the number of tourists coming, which might happen due to seasonal or geopolitical changes. The entire spread between 1.8 and 23485 signifies diversified trends in tourism demand that required further exploration.



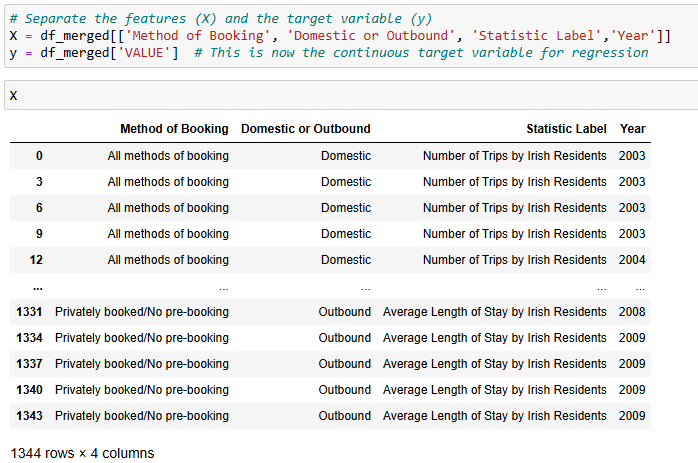
***Figure 13: Extracting Years from the date column***

***Figure 13*** shows the data transformation, starting with extracting the year from the “Quarter” column to create a new column called Year. After extracting, the Quarter entries are reduced by removing Q1, Q2, Q3, and Q4 from the Quarter column. This step simplifies the quarter column by clarifying the analytics for trends in Irish tourism for each of the years, and stakeholders want to look a little further by using the data.



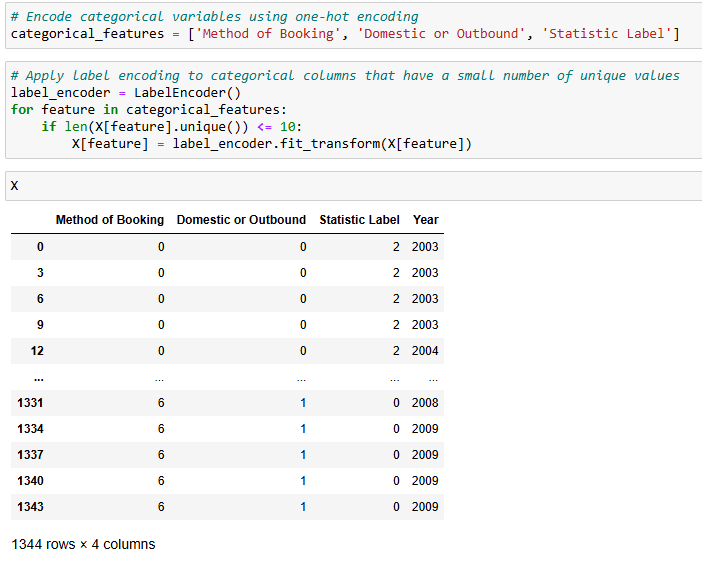
***Figure 14: Dropping the 'Quarter Column'***

After extracting ‘Year’ from Quarter (containing Year and Quarter), the ‘Quarter’ column has been dropped from the data frame to avoid data duplication (***Refer to Figure 14***).



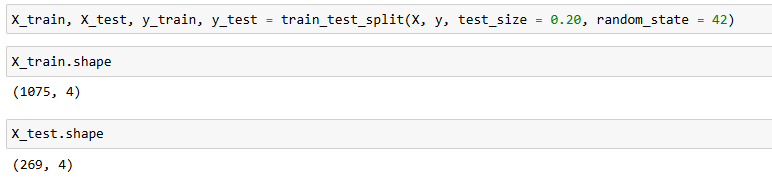
***Figure 15: Separating features and target variable***

As a part of the ML modelling, the features (‘Methods of Booking’, ‘Domestic or Outbound’, ‘Statistics Label’ and ‘Year’) and target variable ‘VALUE’ have been defined by creating two data frames ‘X’ and ‘y’ (***Refer to Figure 15***).



***Figure 16: Categorical encoding for categorical features***

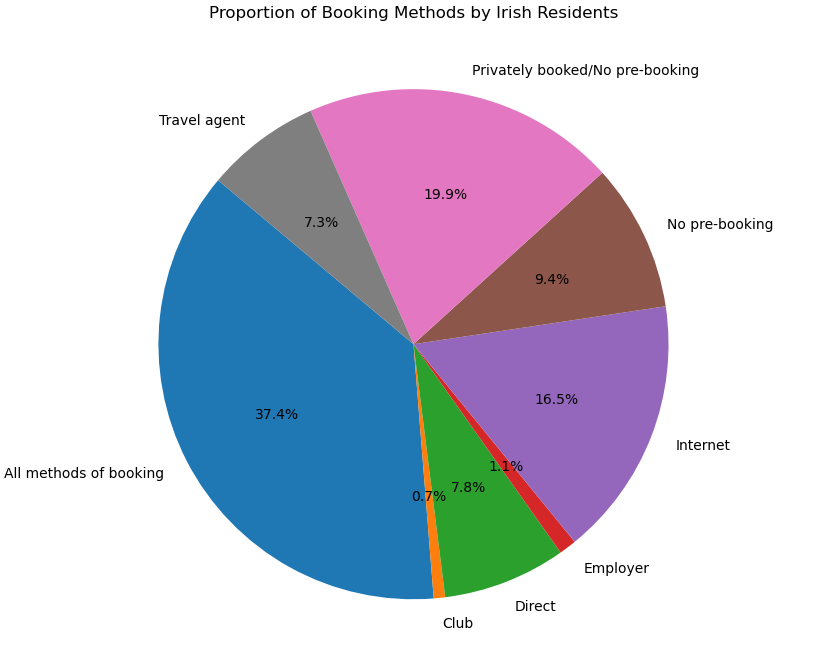
Categorical encoding has been performed using ‘LabelEncoder’ to convert the categorical features (Method of Booking, Domestic or outbound and Statistics Label) into numerical labels. According to Hancock and Khoshgoftaar (2020); Pargent et al. (2022), the use of ‘LabelEncoder’ class from ‘Scikit-learn’ framework enables encoding of categorical class of categorical features into numerical labels, which ensures effective training of ML models. Therefore, the application of categorical encoding has ensured numerical conversions of categorical variables, leading to effective training of supervised ML models as ML models are trained only on numerical labels.

******

***Figure 17: Data Splitting***

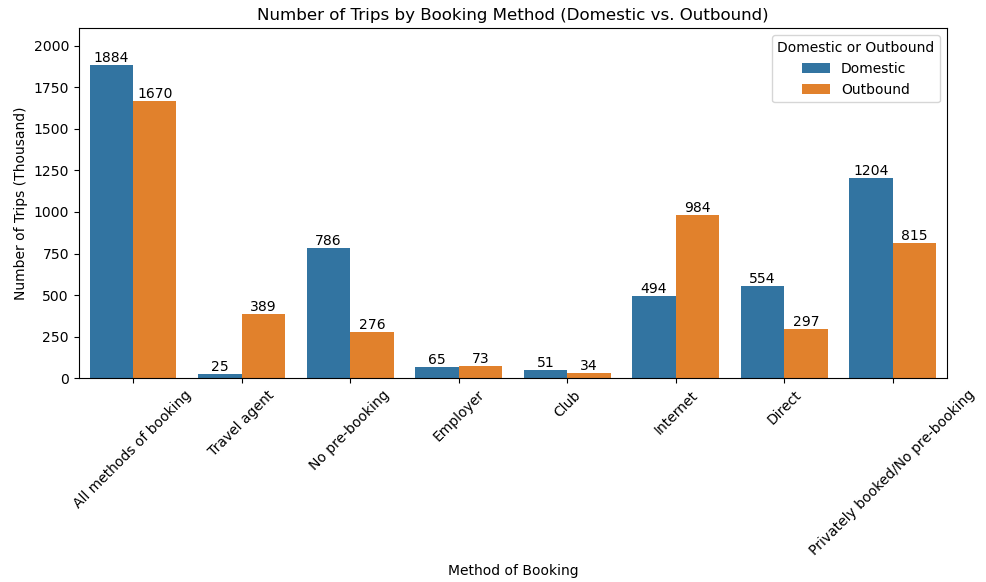
An 80-20% splitting has been performed to segregate the data into train (80%) and test (20%) sets, where training of supervised ML models (Regression) has been performed on 80% of data and remaining 20% of data has been utilised for evaluating model performance (***Refer to Figure 17***). The selection of 80-20% splitting ensures the maintenance of a trade-off between adequate training of supervised models and evaluation of predictive performance of the models.

### 2.2.3 Visualisations (Tufte Principle)



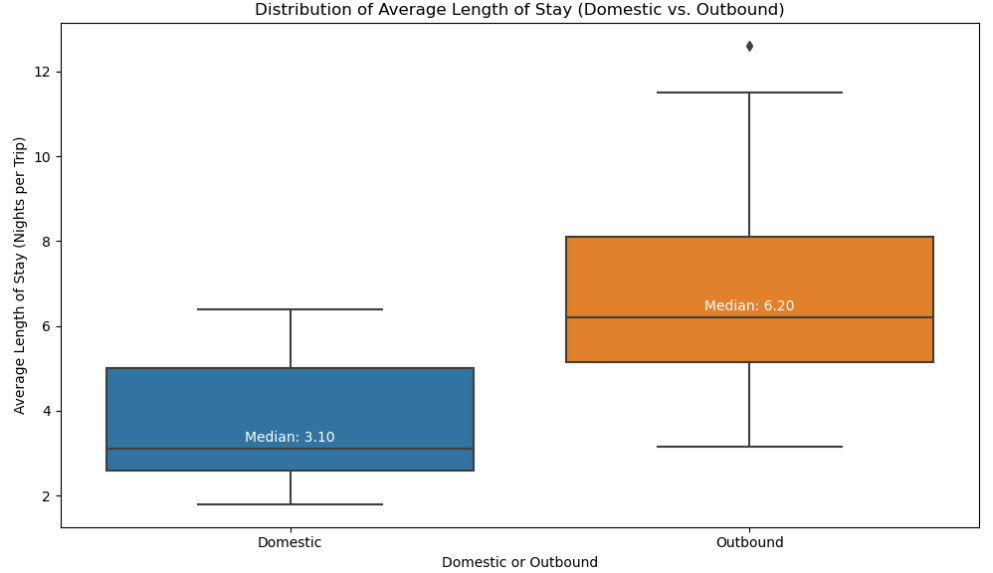
***Figure 18: Distribution of Booking methods by Irish Residents***

From ***Figure 18***, it can be observed that over the years, tourists in Ireland have preferred either ‘all methods of booking’ (37.4%) or ‘privately booked/No pre-booking’ (19.9%), which includes aspects of Tufte principles. The concept of the ‘***Tufte Principle***’ includes ***documentation*** (detailed titles, attribution, and measurements to increase the credibility of the visualisations) and ***context*** (context of the chart) (Tufte, 2001; Midway, 2020). The implication of Tufte’s principle in this report has allowed in showing credibility of the visualisation by integrating elements like detailed titles, data labels and category labels, allowing users to evaluate the distribution of booking methods obtained by Irish tourists. From the pie chart, it can be inferred that many travellers favouring flexible or spontaneous travel arrangements, which highlights diverse travel preferences, appealing to both organised as well as last-minute tourists.



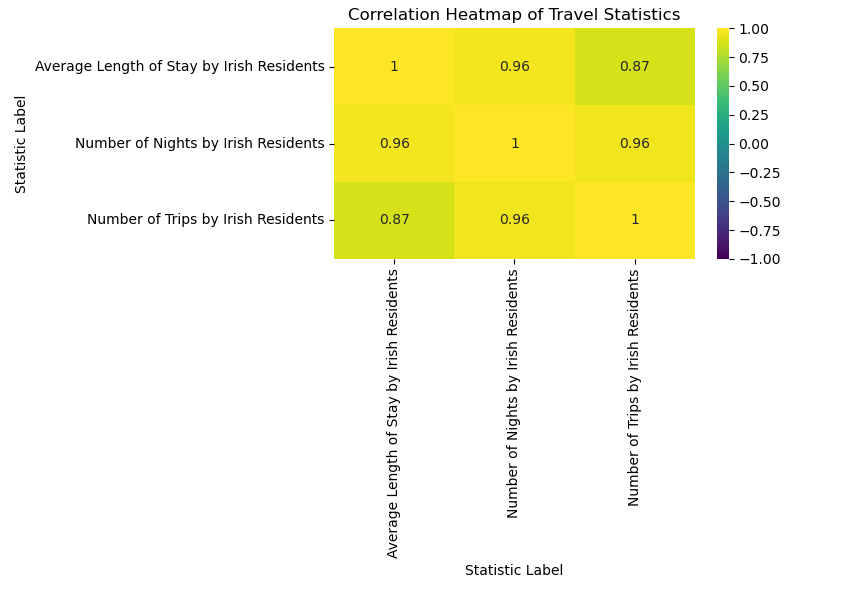
***Figure 19: Number of Trips by Booking Methods***

***Figure 19*** (stacked bar chart) shows the distribution of number of trips in terms of booking methods, which fulfils the ‘Comparisons’ and ‘Multivariate’ aspects of Tufte Principle. From ***Figure 19***, it can be observed that domestic tourists in Ireland have preferred ‘all methods of booking’, on the other hand, outbound tourists have preferred Internet booking methods more as compared to domestic tourists over the years (indicated by the internet column bar in Figure 19). This shows that the offering of internet bookings for international tourists has influenced international tourists to select Ireland as a preferable tourist destination.



***Figure 20: Distribution of average length of stay***

The spread of average length of stay of outbound tourists is higher compared to domestic tourists, which is indicated by a box plot (***Refer to Figure 20***). The report published by Central Statistics Office (CSO) revealed that in 2022, approximately 16.2 million domestic same-day visits and 1.3 million outbound same-day visits were observed over the years (Central Statistics Office, 2023a). The tendency of domestic tourists to make same-day visits leads to a lower spread compared to outbound tourism in Ireland.

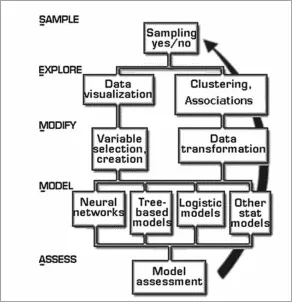


***Figure 21: Correlation Heatmap of Travel Statistics***

The correlation coefficient between variables like ‘Number of nights by Irish Residents’ and ‘Average length of stay by Irish Residents’ is 0.87, reflecting a strong correlation (***Refer to Figure 21***). Due to this, instead of having multicollinearity (Correlation coefficient outside the range of ± 0.70), these variables have been included in this study as they add up in the representation of tourism demand (Value) in Ireland.

## 2.3 Machine Learning for Data Analytics

### 2.3.1 Project Management Framework: SEMMA

******

***Figure 22: Process flow in SEMMA framework***

(Source: Sarnovsky, Bednar and Smatana, 2019)

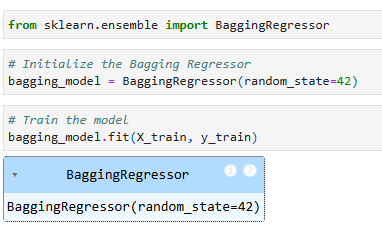
The concept of the ***SEMMA framework*** has been implemented in this study for data acquisition, preparation and modelling for predicting tourism value in Ireland over the years. As per the viewpoint of Plotnikova, Dumas and Milani (2020); Sarnovsky, Bednar and Smatana (2019), SEMMA frameworks include aspects like ***‘Sample’, ‘Explore’, ‘Modify’, ‘Model’*** and ***‘Asesss’*** that enables the development and assessment of predictive modelling (***Refer to Figure 22***). In this study, the implication of the ‘***SEMMA framework’*** has ensured collection of data (***Sample***) (related to tourism sector in Ireland), ***Exploration*** of data (using distribution and EDA), ***Modification*** of data (transformation of data using techniques like categorical encoding), ***Modelling*** of supervised ML models (Regression) and ***Assessment*** of model performance (using metrics like R2 score, Mean-squared error and mean absolute error).

### 2.3.2 Programming paradigms

The target variable (VALUE) is a continuous variable (indicating value of tourism in Ireland) and is labelled in the dataset. According to Espinheira et al. (2019), Regression is a supervised ML technique (works on labelled data) that is suitable for prediction of continuous measure. In this study, the labelled as well as continuous nature of target variable has justified selection of Regression for prediction of tourism value in Ireland over time. The development and training of ML models (Bagging Regressor, Random Forest Regressor and XGBoost Regressor) have been performed using Python programming, as Python provides in-built libraries and frameworks like Pandas, Numpy, Scikit-learn, which makes data handling and model training effective. Additionally, Jupyter Notebook has been selected as the Integrated Development Environment (IDE), which is suitable for this study due to the notebook format of the IDE.

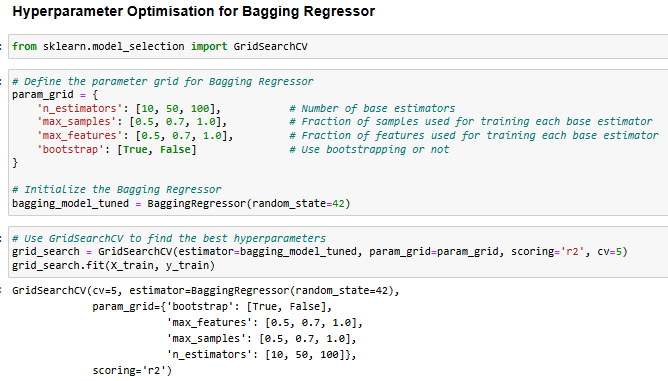
### 2.3.3 Supervised ML Modelling: Regression

#### 2.3.3.1 Bagging regressor (with and without hyperparameter tuning)

******

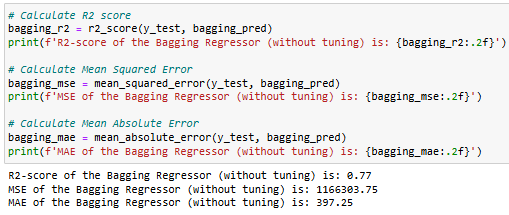
***Figure 23: Model architecture of Bagging Regressor***

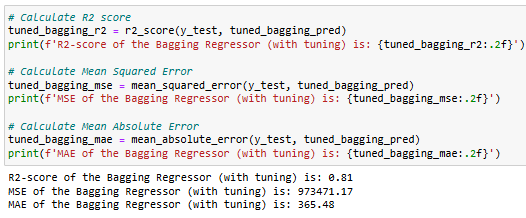
The Bagging Regressor model has been initialised on top of a default Decision Tree Regressor model to improve the overall predictive accuracy. The capability of Bagging Regressor model in minimising model variance and enhancing generalisation of the model makes Bagging method a suitable technique for Regression problem (Scikit-learn, 2024). A random\_state = 42 has been set as a parameter in this Bagging Regressor model to maintain reproducibility of the model (***Refer to Figure 23***).

******

***Figure 24: Hyperparameter optimisation of Bagging Regressor model***

A GridSearchCV method has been applied to the Bagging Regressor model to identify and train the models based on optimal values of a pool of hyperparameters (max\_features, max\_samples, n\_estimators) (***Refer to Figure 24***). According to Verma and Yadav (2024), GridSearchCV method enables fine-tuning of ML models by finding the best combination of hyperparameters from the pre-defined grid, making the identification of hyperparameters more systematic. Therefore, GridSearchCV method has opted in this study over RandomSearchCV method to optimally utilise computational resources for finding optimal parameters of the model. The identification of optimal values of these parameters has helped in more robust training of the Bagging Regressor model, which has led to higher predictive performance for prediction of tourism value in Ireland.

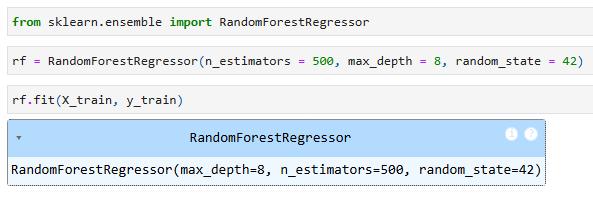


******

***Figure 25: Model performance of Bagging Regressor (without and with hyperparameter optimisation)***

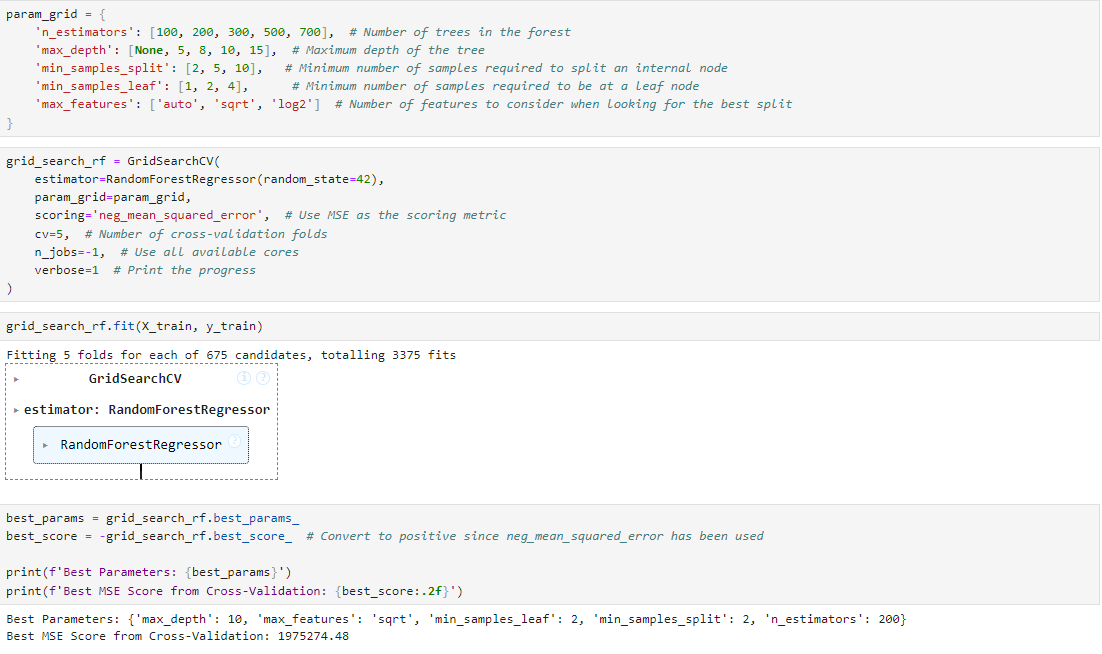
R2 score obtained from the Bagging Regressor is 0.77 (indicating 77% of the variability in tourism value over time can be explained by the features), which has increased to 0.81 (81%) after implication of hyperparameter optimisation technique. Additionally, the Mean-squared error (MSE) and mean absolute error (MAE) of the model have reduced from 1166303.75 to 973471.17 and from 397.25 to 365.48 after application of hyperparameter optimisation technique (***Refer to Figure 25***). This reflects that training of Bagging Regressor model with optimal values of parameters (max\_features, max\_samples, n\_estimators) has led to a higher predictive performance of the model.

#### 2.3.3.2 Random Forest Regressor (Without and with hyperparameter optimisation)

******

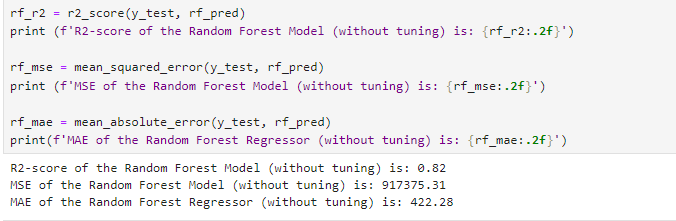
***Figure 26: Model architecture for Random Forest regressor***

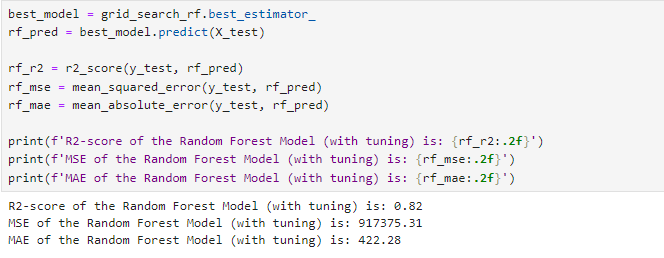
The Random Forest regression model was initialised with a max\_depth = 8 and n\_estimators = 500 to maintain the ***model complexity*** and ***predictive efficacy*** trade-off. According to Hwang et al. (2023), Random Forest Regression model is robust to outliers in target variable and has the capability to accurately capture variations in target variable based on multi-dimensional features. Therefore, to minimise the effect of outliers in target variable and adequately capture the multi-dimensional influence of features (like booking method and type of travel) on target variable (tourism value), Random Forest regression model been considered.

******

***Figure 27: Hyperparameter optimisation for Random Forest Regressor***

The implication of GridSearchCV technique has been applied to identify the optimal value of the hyperparameters. The identified optimal values of hyperparameters are respectively max\_depth = 10, max\_features = ‘sqrt’, min\_sample\_leaf = 2, min\_sample\_split = 2 and n\_estimators = 200 (***Refer to Figure 27***). The neg\_mean\_squared\_error scoring metric in scikit-learn is used to allow error minimization as the optimisation goal during cross-validation, however, it represents scores as negative values (the negative of the mean squared error). This approach helps the GridSearchCV function identify the minimum MSE by maximising the negative value.

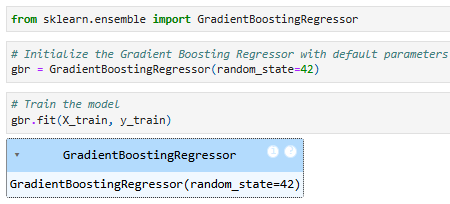
******

******

***Figure 28: Predictive performance of the Random Forest Regressor (without and with hyperparameters optimisation)***

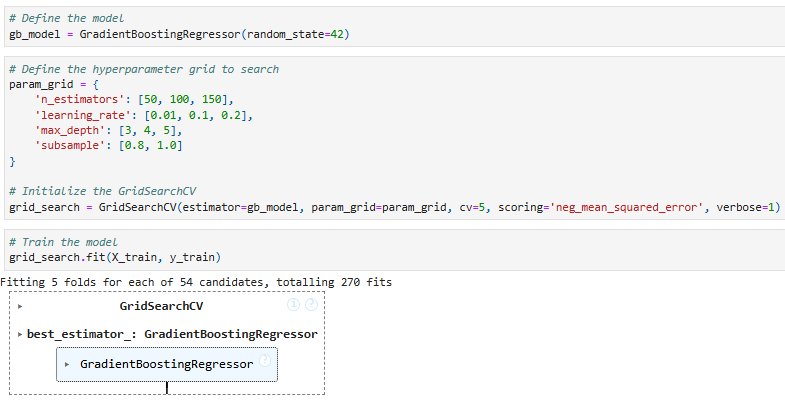
The R2 score of the Random Forest regressor model is 0.82, reflecting the model has the capability to explain 82% variability in target variable (tourism value), which is slightly higher than Bagging Regressor (***Refer to Figure 28***). However, the implication of hyperparameter optimisation has failed to improve the predictive performance of the model, which may be due to optimal training in the base Random Forest model itself.

#### 2.3.3.3 GradientBoosting Regressor (without and with hyperparameter optimisation)

******

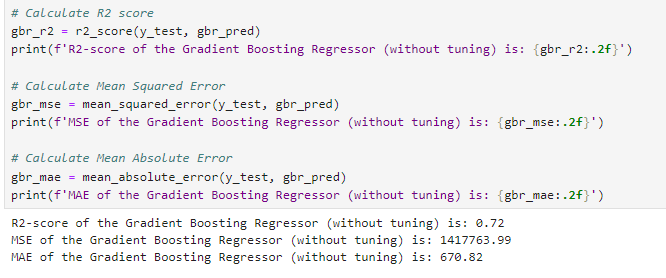
***Figure 29: Architecture GradientBoosting Regressor***

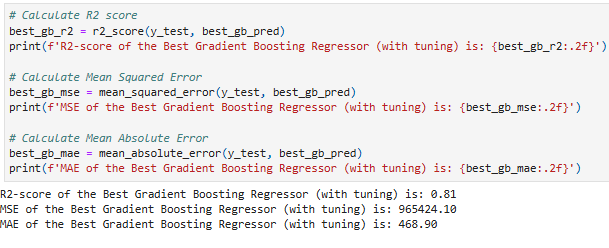
The Gradient Boosting Regressor model has been initialised from the ‘sklearn. ensemble’ module with a random\_state = 42, which has ensured reproducibility of the model (***Refer to Figure 29***).

******

***Figure 30: Hyperparameter optimization of Gradient Boosting Regressor***

The base Gradient Boosting Regressor model has been further tuned using the GridSearchCV method through identifying the optimal value of the hyperparameters (n\_estimators, learning\_rate, max\_depth, subsample) from a grid (***Refer to Figure 30***).

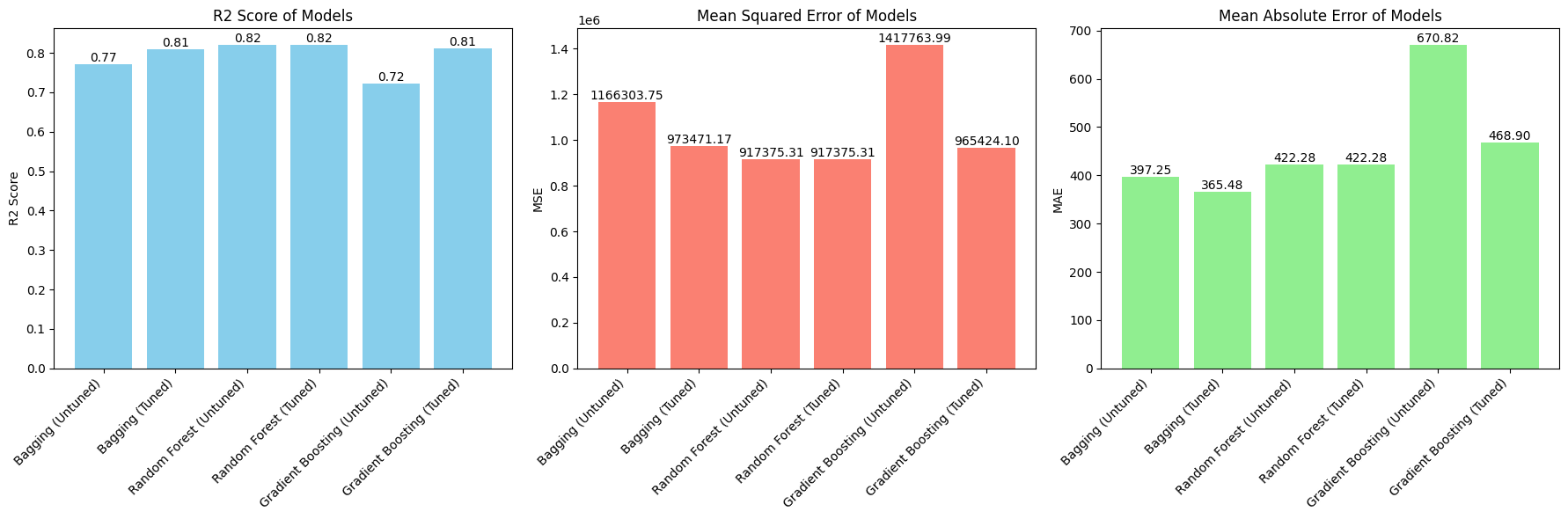


******

***Figure 31: Predictive performance of Gradient Regression model (without and with hyperparameter optimisation)***

The obtained R2 score of the Gradient Boosting regressor model is 0.72 (explaining 72% variability in target variable), which has increased to 0.81 after implication of hyperparameter optimization (***Refer to Figure 31***). Additionally, the MSE and MAE have dropped significantly from 1417763 to 965424 and from 670.82 to 468.90 respectively, reflecting tuning of models based on optimal hyperparameter value has improved predictive performance of the model.

### 2.3.4 Model comparisons

******

***Figure 32: Model comparisons***

The predictive performance of the Random Forest regressor model (with and without hyperparameter optimisation) is slightly better (with R2 score = 0.82, MSE = 917375.31) than the Bagging and Gradient Boosting regressor model (***Refer to Figure 32***). This shows the high applicability of the Random Forest regressor model in the prediction of tourism value (overall number of tourists) in Ireland.

### 2.3.5 Logical interpretation

#### 2.3.5.1 Similarities and differences between the results obtained from the machine learning models

The predictive performance of the Random Forest regression model is slightly higher in the prediction of tourism value within Irish tourism sector over time. Other models like the Bagging Regressor and Gradient Boosting regressor, have shown a comparatively lower predictive accuracy, however, the implication of hyperparameter optimisation has improved the predictive accuracy of the model. On the other hand, past studies have emphasised on the implication of forecasting methods such as ARIMA for the prediction of demand for Irish tourism, however, the obtained predictive performance is comparatively lower (below 70%) (Wu et al., 2024). Thus, it can be inferred that state-of-the-art ML models (such as Random Forest Regressor) have high reliability in prediction of tourism value in Ireland.

#### 2.3.5.2 Relevance, suitability and practical implications of findings

The findings of this study primarily emphasised the positive influence of the availability of multiple booking methods and payment options on the overall growth of market value of Irish tourism over time. This can provide detailed insights into the behaviour of travellers while booking hotels and hospitality services near the tourist destination. Additionally, the incorporation of factors like type of travel (domestic or outbound) and type of travel within the predictive modelling has provided insights into the multi-dimensional factors that can potentially affect the demand for tourism in Ireland.

# 3. Conclusion

Based on the above discussion, it can be summarised that factors like type of travel, payment method and booking method have a substantial positive influence on demand for tourism in Ireland. Additionally, within the context of SEMMA methodological framework, implication of data preparation and transformation steps like treatment of missing values and categorical encoding of features has ensured effective training of three different regression models (Bagging, Random Forest and Gradient Boosting model). The application of hyperparameter optimisation using GridSearchCV method has improved predictive performance except for Random Forest due to high capability of handling outliers and overfitting problems, which has enhanced overall applicability of the model in real-world scenarios for prediction of tourism value in Ireland.

# References

Brereton, R.G. (2014). The normal distribution. *Journal of Chemometrics*, [online] 28(11), pp.789–792. doi: https://doi.org/10.1002/cem.2655.

Central Statistics Office (2023a). *Household Travel Survey Quarter 4 and Year 2022 - CSO - Central Statistics Office*. [online] www.cso.ie. Available at: https://www.cso.ie/en/releasesandpublications/ep/p-hts/householdtravelsurveyquarter4andyear2022/ [Accessed 28 Oct. 2024].

Central Statistics Office (2023b). *Tourism Information - CSO - Central Statistics Office*. [online] www.cso.ie. Available at: https://www.cso.ie/en/interactivezone/statisticsexplained/tourismandtravel/tourisminformation/ [Accessed 25 Oct. 2024].

Espinheira, P., da Silva, L., Silva, A. and Ospina, R. (2019). Model Selection Criteria on Beta Regression for Machine Learning. *Machine Learning and Knowledge Extraction*, [online] 1(1), pp.427–449. doi: https://doi.org/10.3390/make1010026.

Hancock, J.T. and Khoshgoftaar, T.M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, [online] 7(1), pp.1–41. doi: https://doi.org/10.1186/s40537-020-00305-w.

Hwang, S.-W., Chung, H., Lee, T.-K., Kim, J., Kim, Y., Kim, J.-C., Kwak, H.S., Choi, I.-G. and Yeo, H. (2023). Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar. *Journal of Wood Science*, [online] 69(1), pp.1–12. doi: https://doi.org/10.1186/s10086-022-02073-y.

Irish Tourism Industry Confederation (2023). *The Competitiveness of Irish Tourism – Review & Outlook - June 2023*. [online] ITIC - Industry Recovery Roadmap. Available at: https://www.itic.ie/RECOVERY/competitiveness-2023/ [Accessed 25 Oct. 2024].

Jadhav, A., Pramod, D. and Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, [online] 33(10), pp.913–933. doi: https://doi.org/10.1080/08839514.2019.1637138.

Malichová, E., Cornet, Y. and Hudák, M. (2022). Travellers’ use and perception of travel time in long-distance trips in Europe. *Travel Behaviour and Society*, [online] 27, pp.95–106. doi: https://doi.org/10.1016/j.tbs.2021.12.003.

Midway, S.R. (2020). Principles of Effective Data Visualization. *Patterns*, [online] 1(9), p.100141. doi: https://doi.org/10.1016/j.patter.2020.100141.

Murray, S. (2024). *Landscapes and scenery the main reasons tourists come to Ireland; survey finds*. [online] Irish Examiner. Available at: https://www.irishexaminer.com/news/arid-41495384.html [Accessed 25 Oct. 2024].

Pargent, F., Pfisterer, F., Thomas, J. and Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, [online] 37, pp.2671–2692. doi: https://doi.org/10.1007/s00180-022-01207-6.

Plotnikova, V., Dumas, M. and Milani, F. (2020). Adaptations of data mining methodologies: a systematic literature review. *PeerJ Computer Science*, [online] 6, p.e267. doi: https://doi.org/10.7717/peerj-cs.267.

RTÉ News (2024). Domestic holidays taken last year up 8% to over 1 million. *RTE.ie*. [online] doi: urn: epic:1460587.

Sarnovsky, M., Bednar, P. and Smatana, M. (2019). Cross-Sectorial Semantic Model for Support of Data Analytics in Process Industries. *Processes*, [online] 7(5), p.281. doi: https://doi.org/10.3390/pr7050281.

Scikit-learn (2024). *BaggingRegressor*. [online] scikit-learn. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.BaggingRegressor.html [Accessed 28 Oct. 2024].

Stańczyk, U., Zielosko, B. and Baron, G. (2024). Importance of Characteristic Features and Their Form for Data Exploration. *Entropy*, [online] 26(5), pp.404–404. doi: https://doi.org/10.3390/e26050404.

Statista (2024). *Travel & Tourism - Ireland | Statista Market Forecast*. [online] Statista. Available at: https://www.statista.com/outlook/mmo/travel-tourism/ireland [Accessed 25 Oct. 2024].

Thanos, C., Meghini, C., Bartalesi, V. and Coro, G. (2023). An exploratory approach to data driven knowledge creation. *Journal of Big Data*, 10(1). doi: https://doi.org/10.1186/s40537-023-00702-x.

Tufte, E.R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.

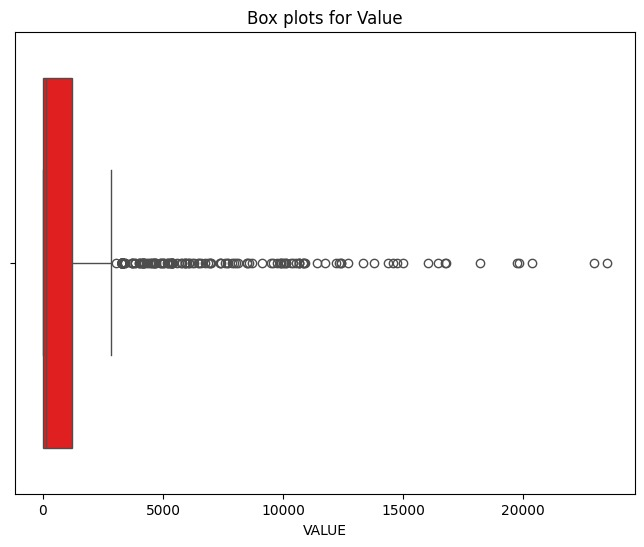
Verma, B.K. and Yadav, A.K. (2024). Advancing Software Vulnerability Scoring: A Statistical Approach with Machine Learning Techniques and GridSearchCV Parameter Tuning. *SN Computer Science*, [online] 5(5), p.595. doi: https://doi.org/10.1007/s42979-024-02942-x.

Wang, X., Oh, Y., Park, S., Hong, S. and Lai, P. (2024). Factors influencing travelers’ intention by environmental changes of destination: Cross‐country evidence from Far‐East Asia. *International Journal of Tourism Research*, [online] 26(4), p.e2705. doi: https://doi.org/10.1002/jtr.2705.

Wu, D.C., Zhong, S., Wu, J. and Song, H. (2024). Tourism and Hospitality Forecasting with Big Data: A Systematic Review of the Literature. *Journal of Hospitality & Tourism Research*. [online] doi: https://doi.org/10.1177/10963480231223151.

# Appendices

## Appendix 1: Box plot of VALUE

******

***Figure 33: Box plot for VALUE***