**CCT College Dublin**

|  |  |
| --- | --- |
| **Module Title:** | MSc. Data Analytics CA 1 REPORT |
| **Assessment Title:** | MSC\_DA\_CA1 |
| **AUTHOR** | OLUFEYI MICHAEL AKINMOLA |
| **Student Number:** | SBS24029 |
| **GITHUB LINK** | https://github.com/Olufeyi24049/MSC\_DA\_CA1 |

Table of Contents

[Section-I (Statistics) 3](#_Toc162456384)

[A. Descriptive Statistics 3](#_Toc162456385)

[1. Dataset Shape: 3](#_Toc162456386)

[2. Column Names: 3](#_Toc162456387)

[3. Data Types: 3](#_Toc162456388)

[4. Summary Statistics: 4](#_Toc162456389)

[5. Missing Values: 5](#_Toc162456390)

[Critical Analysis: 6](#_Toc162456391)

[B. Visualization with Plots 6](#_Toc162456392)

[1. Histogram: Distribution of Tourism Revenue 6](#_Toc162456393)

[2. Heatmap of Correlation Matrix 7](#_Toc162456394)

[3. Bar Graph 8](#_Toc162456395)

[4. Line Graph 9](#_Toc162456396)

[C. Discrete Distributions 10](#_Toc162456397)

[1. Binomial Distribution: 10](#_Toc162456398)

[2. Poisson Distribution: 11](#_Toc162456399)

[D. Normal Distributions 12](#_Toc162456400)

[E. Importance of the distributions used in point 3 and 4 13](#_Toc162456401)

[1. Binomial and Poisson Distributions (Point 3): 13](#_Toc162456402)

[2. Normal Distribution (Point 4): 13](#_Toc162456403)

[SECTION-II 14](#_Toc162456404)

[A. Data preparation and Visualization 14](#_Toc162456405)

[1. Exploratory Data Analysis 14](#_Toc162456406)

[B. The data Methods applied during data preparation for ML 19](#_Toc162456407)

[C. Final Insights 20](#_Toc162456408)

[D. Visualization with Tuft’s Principle 23](#_Toc162456409)

[Section-III (Machine learning for Data Analytics) 24](#_Toc162456410)

[A. a)Project Management Framework 24](#_Toc162456411)

[Justification with real-life Scenario 24](#_Toc162456412)

[b). Selection ML Technique and Models for our Dataset 25](#_Toc162456413)

[B. Approach of Hyperparameter tunning 25](#_Toc162456414)

[Gradient Boosting Regression 25](#_Toc162456415)

[Decision Tree Regression 26](#_Toc162456416)

[Random Forest Regression: 26](#_Toc162456417)

[C. ML modeling comparisons 26](#_Toc162456418)

[D. Demonstration 28](#_Toc162456419)

[References 30](#_Toc162456420)

# Section-I (Statistics)

## A. Descriptive Statistics

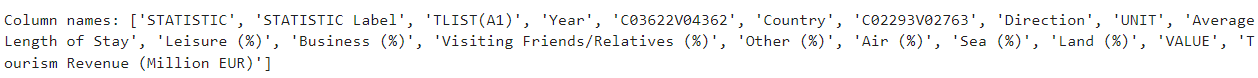
### 1. Dataset Shape:

* The dataset contains 1150 rows and 19 columns.



### 2. Column Names:

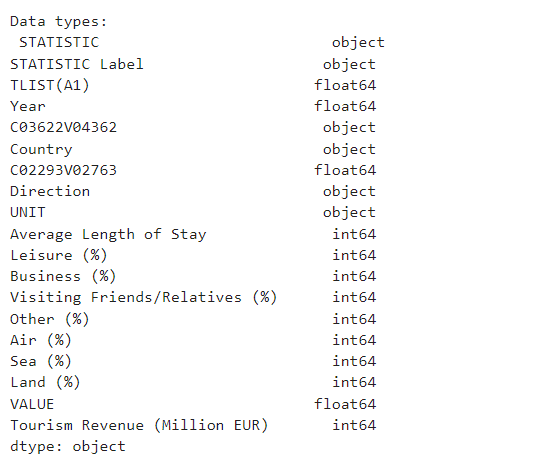
* The dataset has 19 columns with names such as 'STATISTIC', 'STATISTIC Label', 'TLIST(A1)', 'Year', 'Country', 'Direction', 'UNIT', 'Average Length of Stay', 'Leisure (%)', 'Business (%)', 'Visiting Friends/Relatives (%)', 'Other (%)', 'Air (%)', 'Sea (%)', 'Land (%)', 'VALUE', and 'Tourism Revenue (Million EUR)'.



### 3. Data Types:

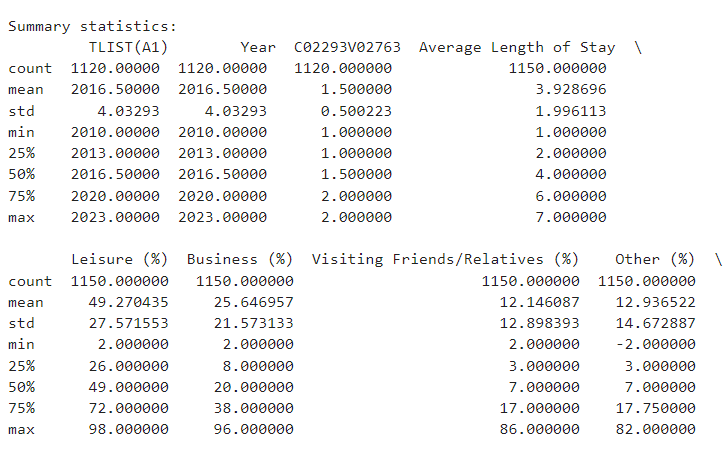
The data types of the columns vary:

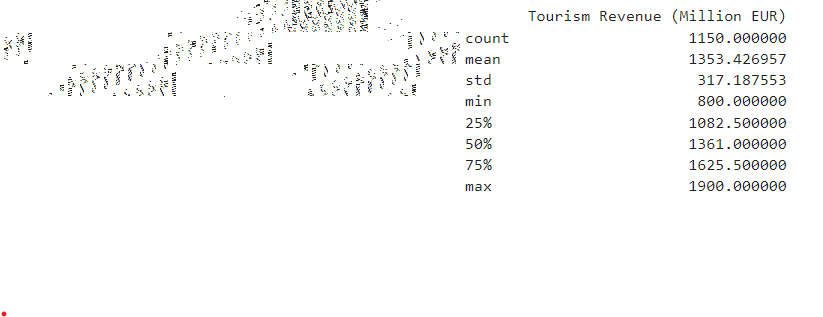
* 'object': STATISTIC, STATISTIC Label, C03622V04362, Country, Direction, UNIT.
* 'float64': TLIST(A1), Year, C02293V02763, VALUE.
* 'int64': Average Length of Stay, Leisure (%), Business (%), Visiting Friends/Relatives (%), Other (%), Air (%), Sea (%), Land (%), Tourism Revenue (Million EUR).



### 4. Summary Statistics:

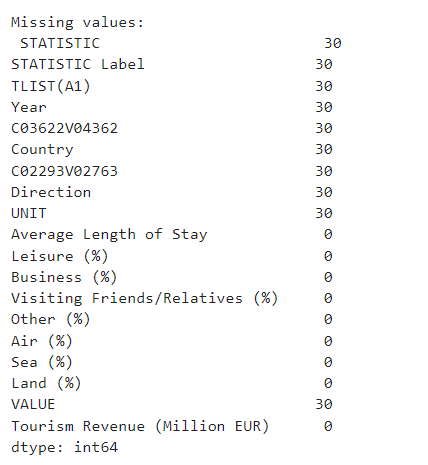
* TLIST(A1): Year column has a mean of approximately 2016.5 with a standard deviation of around 4.03. The data spans from the year 2010 to 2023.
* Average Length of Stay: The average length of stay ranges from 1 to 7 days, with a mean of approximately 3.93 days.
* Leisure (%), Business (%), Visiting Friends/Relatives (%), Other (%): These columns represent the percentage of tourists engaging in different activities. The mean percentages vary, indicating different preferences among tourists.
* Air (%), Sea (%), Land (%): These columns represent the mode of transportation used by tourists. The mean percentages suggest a mix of transportation modes.
* VALUE: The mean value is approximately 736.95, with a large standard deviation of 2583.77, indicating high variability in the data.
* Tourism Revenue (Million EUR): The mean tourism revenue is approximately 1353.43 million EUR, with a standard deviation of 317.19.





### 5. Missing Values:

* There are missing values in some columns (STATISTIC, STATISTIC Label, TLIST(A1), Year, C03622V04362, Country, C02293V02763, Direction, UNIT, VALUE). These missing values need to be handled appropriately before further analysis.



### Critical Analysis:

* The dataset provides insights into tourism-related statistics such as length of stay, activities, transportation modes, and revenue.
* There is variability in the data, as indicated by the standard deviations, suggesting diverse tourism patterns.
* The presence of missing values requires careful handling to ensure the integrity of the analysis.

## B. Visualization with Plots

### 1. Histogram: Distribution of Tourism Revenue

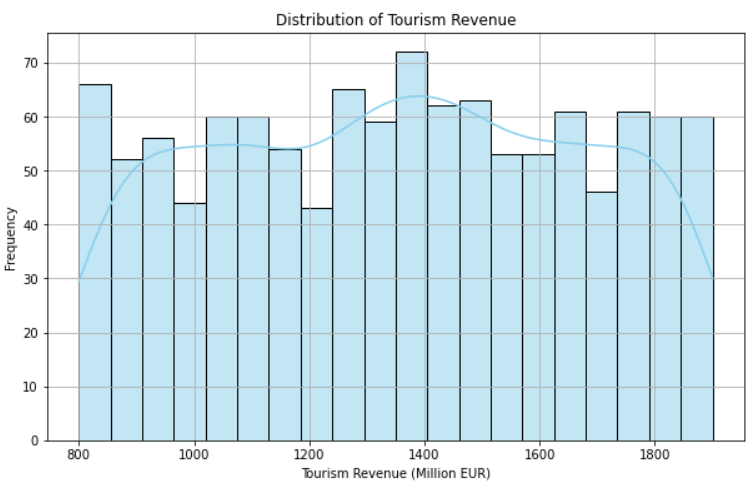
* X-axis: Represents the range of tourism revenue (in million EUR).
* Y-axis: Represents the frequency of occurrence for each revenue range.

**Analysis:**

* Minimum Frequency: The lowest frequency observed in the dataset is approximately 42. This frequency corresponds to a tourism revenue range of around 1200 million EUR.
* Tallest Frequency Bar: The tallest bar in the histogram indicates the highest frequency observed in the dataset, which is approximately 72. This frequency corresponds to a tourism revenue range of around 1400 million EUR.
* Average High Frequencies: The histogram shows that the majority of high frequencies lie within the range of 1300 to 1500 million EUR.

**Interpretation:**

* The histogram indicates the distribution of tourism revenue across different ranges.
* Most of the data points seem to cluster around the 1300 to 1500 million EUR range, as suggested by the higher frequencies observed in this range.



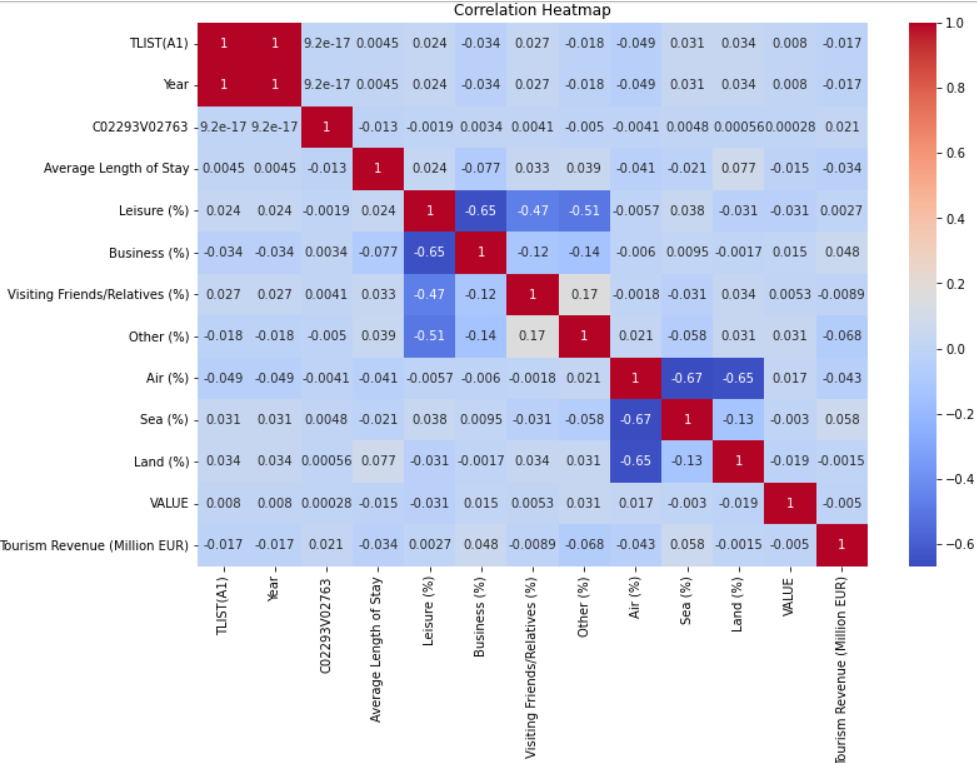
### 2. Heatmap of Correlation Matrix

Correlation matrices visually show the correlation between different variables. It uses a grid of squares where each square shows the correlation between two variables represented by the labels on the axes. The color of the square represents the strength and direction of the correlation.

In this specific correlation matrix, the variables include:

* TLIST(A1)
* Year
* C02293V02763
* Average Length of Stay
* Leisure (%)
* Business (%)
* Visiting Friends/Relatives (%)
* Other (%)
* Air (%)
* Sea (%)
* Land (%)
* Value
* Tourism Revenue (Million EUR)

The correlation coefficient ranges from -1 to 1. A correlation coefficient of 1 represents a perfect positive correlation, which means that as the value of one variable increases, the value of the other variable also increases. A correlation coefficient of -1 represents a perfect negative correlation, which means that as the value of one variable increases, the value of the other variable decreases. A correlation coefficient of 0 means that there is no linear correlation between the two variables.

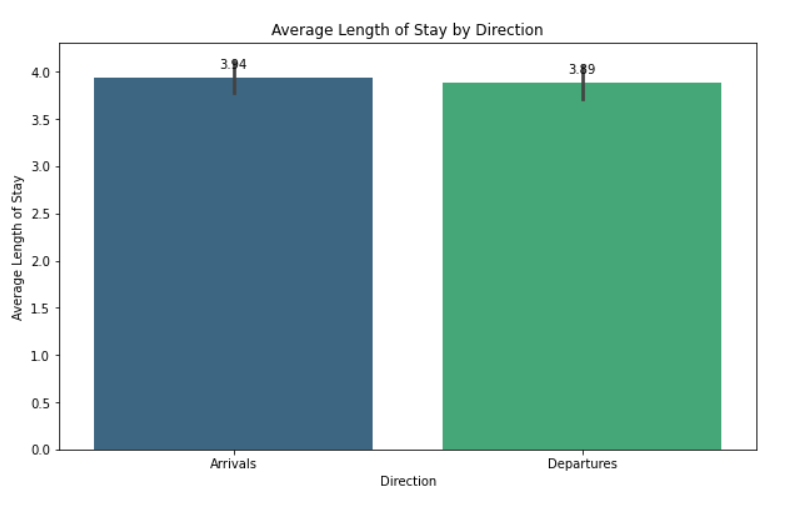


### 3. Bar Graph

The bar graph shows the average length of stay by direction. Here are the different perspectives of the graph:

* The y-axis shows the average length of stay in days.
* The x-axis shows the direction, which is either arrivals or departures.
* The blue bar shows that the average length of stay for arrivals is 3.94 days.
* The green bar shows that the average length of stay for departures is 3.89 days.

While the difference is small, it suggests that people might spend a little more time at their destination upon arrival compared to the place they are departing from.

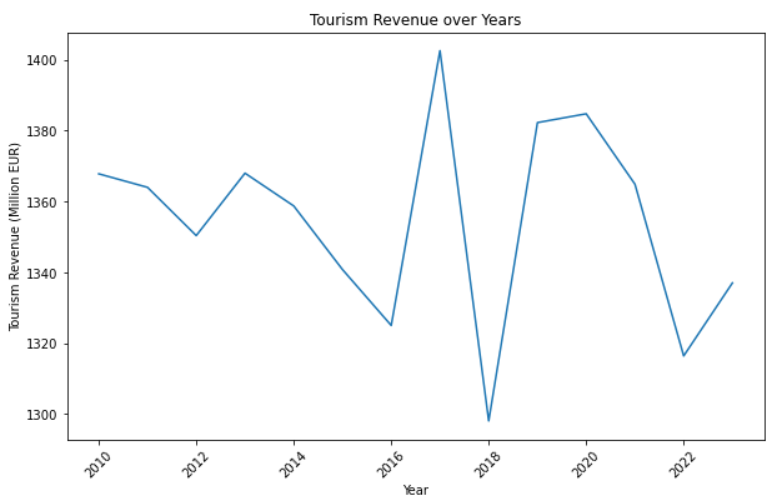


### 4. Line Graph

The line plot you sent me shows tourism revenue over a period of years. The y-axis shows the revenue in millions of Euros and the x-axis shows the years. The line shows that there was a peak in 2017 but fell down afterwards till 2022, and now again increasing the line means that tourism is increasing again after 2022.

Here’s a more detailed explanation of the information in the plot:

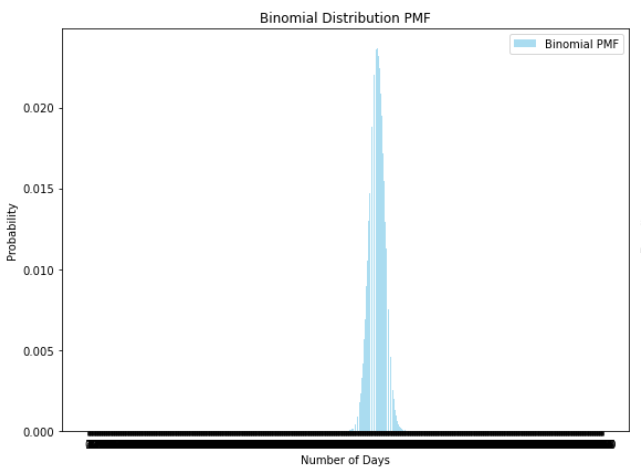
* The y-axis title is “Tourism Revenue (Million EUR)”. This means that the values on the y-axis represent the amount of money that was brought in by tourism in millions of Euros.
* The x-axis title is simply “Year”. This axis shows the years that the data was collected for.
* The line starts at a value around 1370 million euros in 2010 and ends at a value around 1350 million euros in 2022.



## C. Discrete Distributions

### 1. Binomial Distribution:

The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials [[1]](#_References). In the context of the dataset, we can use the binomial distribution to model the probability of tourists spending a certain number of days in a destination, given a probability of staying for a single day.

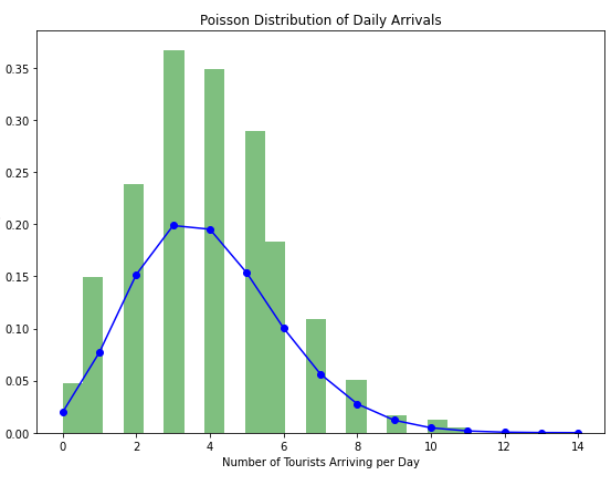


• The x-value of the graph exhibit the number of days between visits, and the y-value display the probability of that days per day between visits occurrence.

• "Binomial PMF" at the graph's top identifies as the binomial probability mass function. The PMF stands for the Probability Mass Function which is a mathematical probability function that helps to find the probability of each outcome in the discrete probability distribution.

### 2. Poisson Distribution:

The Poisson distribution analyze the number of events happening during a single period of time or space given information about the average number of events [[2]](#_References). We can employ the Poisson distribution in modelling the number of creatives arriving on a given destination daily if arrivals are independently distributed and the rate of arrivals is the same.



* The x-axis shows the number of tourists arriving
* The y-axis shows the probability of that number of tourists arriving on a given day.

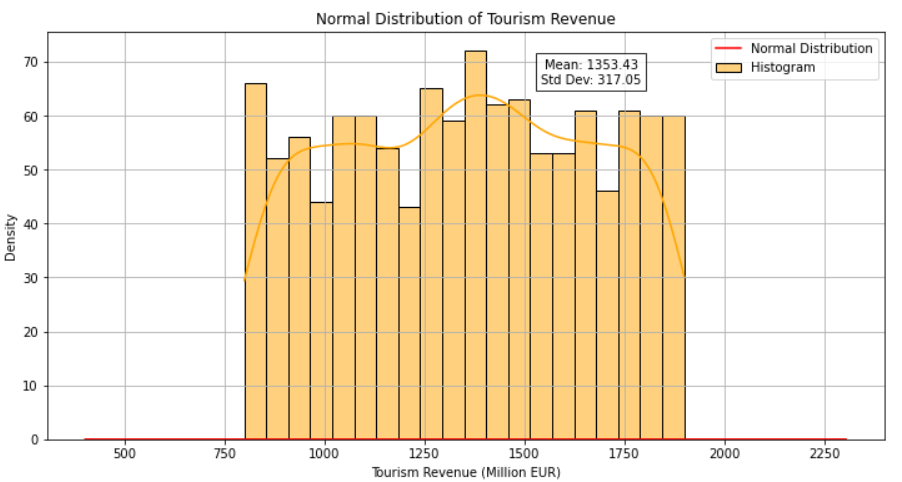
The Poisson distribution would be suitable for the number of daily tourist arrivals fitting as it applies to cases when events happen independently and their impact on the following events does not exist this exponential arrival/travel station assumes a constant rate.

**Explanation:**

* We selected 'Average Length of Stay' as a representative variable for modeling using discrete distributions.
* We fitted a binomial distribution to model the probability of tourists staying for a certain number of days.
* We fitted a Poisson distribution to model the number of tourists arriving in a destination per day.
* The fitted distributions were visualized and compared.
* With large samples, both the binomial and Poisson distributions converge to the corresponding theoretical distributions due to the central limit theorem and law of large numbers.
* In the case of the Poisson distribution, with a large sample size, the distribution of daily arrivals is expected to become more concentrated around the mean arrival rate.

## D. Normal Distributions

The Normal distribution (also known as the Gaussian distribution) is commonly used to model continuous data that are symmetrically distributed around a central mean value. It is characterized by its mean (μ) and standard deviation (σ) [[4]](#_References).



* We selected 'Tourism Revenue (Million EUR)' as a representative variable for modeling using the Normal distribution.
* The mean tourism revenue is approximately 1353.43 million EUR, and the standard deviation is approximately 317.05 million EUR.
* We generated values for the Normal distribution using the calculated mean and standard deviation.
* We plotted the histogram of the 'Tourism Revenue (Million EUR)' data along with the probability density function (PDF) of the Normal distribution.
* The comparison between the histogram and the Normal distribution helps us understand how well the Normal distribution fits the data.
* The statistics displayed on the graph (mean and standard deviation) provide additional information about the central tendency and spread of the data distribution.

## E. Importance of the distributions used in point 3 and 4

Point No 3 and 4 analyzes binary and poison distributions alongside the distribution of normal types. Here's an explanation of their importance and justification of variable choice;

### 1. Binomial and Poisson Distributions (Point 3):

**Representation of Variables:**

• Even though we employed the 'Length of Stay - Avg' and 'Visitor Receipts - Million EUR' to establish normal distribution in point 4, binomial and poison distribution may still offer useful insight into the rest of the variables in the file.

• To this extent, variables related to visitors could be modeled through Binomial or Poisson distributions, for instance by defining "Leisure (%)" or "Business (%)" or "Visiting Friends/Relatives (%)" choices as the probabilities or frequencies of different outcomes.

• Such discrete distributions can be a measure for two types of variables that imply frequencies and proportions, for example, the segment of visitors who reached for leisure purposes and the one for business reasons.

**Justification of Variable Choice:**

• Although we focused on finite and discrete data in Points 3 and 4, the binomial and Poisson distributions take the range of possible datasets distinctly into account.

• With discrete variables like numbers or counts being selected, these distributions could be utilized for the analysis of a different tourism activity variable of either preferences, travel habits, or spending behaviors.

**Could These Variables Be Modeled Using the Normal Distribution?**

• The independent ending values of the discrete distributions (for example, average number of days under stay and daily arrivals) are not suitable for modelling by using the Normal distribution.

• Normal distribution implies the data that is continuous, while data that should be characterized as a mean value must be symmetrically distributed.

• Volumes of tourism(e.g. stay period, arrival numbers) are gaps that usually indicates some kind of skewness or non-normality.

### 2. Normal Distribution (Point 4):

**Representation of Variables:**

• 'Tourism Revenue (Million EUR)' in modeling Normal distribution in point 4 is only a tool to demonstrate how this distribution can serve for finite variables in the dataset.

• Another within category that have normal distribution can be 'VALUE' or some certain year these variables which can be used in analysis of both distribution and the variability.

• We can exploit these Normal to gain better understanding of central tendencies and variabilities or to extract generalized conclusions about the nature of any continuous variable.

**Justification of Variable Choice:**

• While our case study for this task was 'Tourism Revenue (Million EUR)' distribution for the Normal case, other such variables present in the dataset are open for this exercise.

• The Normal distribution is usually employed for continuous variables which are often involved in datasets and have a distinctly different pattern as contrast to discrete variables.

• Through identifying variables which come on a scale of continuity, the underlying Normal distribution becomes a powerful tool to dig deep into other statistical properties that are containing in this dataset.

**Could This Variable Be Modeled Using Discrete Distributions?**

* The variable 'tourism revenue' is not suitable for modeling using discrete distributions like Binomial or Poisson distributions.
* Discrete distributions are designed to model counts or proportions, which are not applicable to continuous variables like revenue.

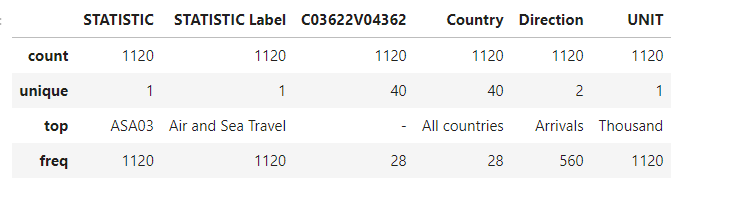
# SECTION-II

## A. Data preparation and Visualization

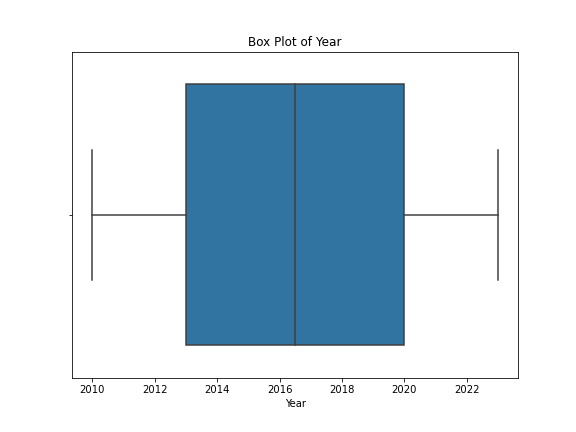
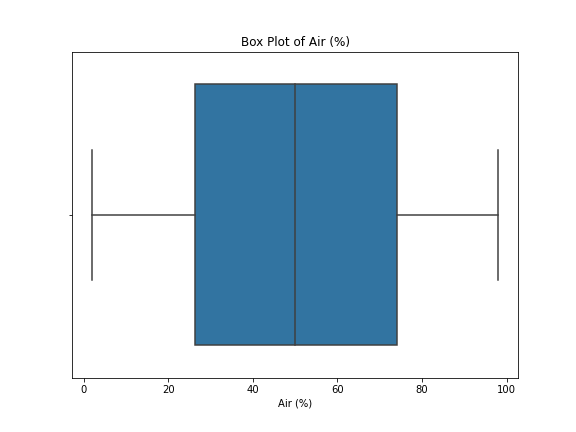
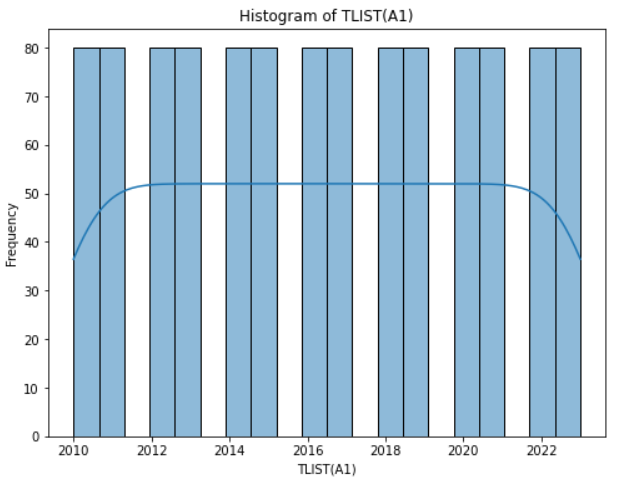
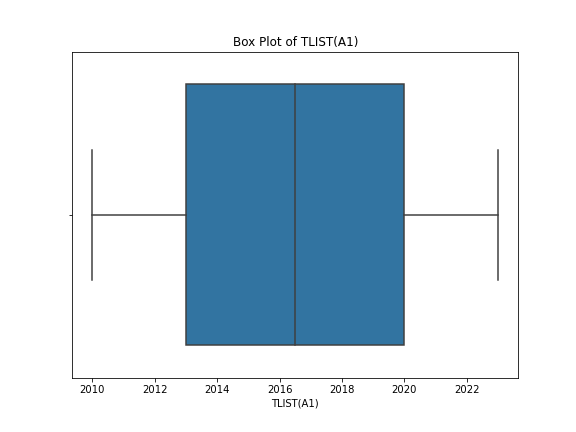
### 1. Exploratory Data Analysis

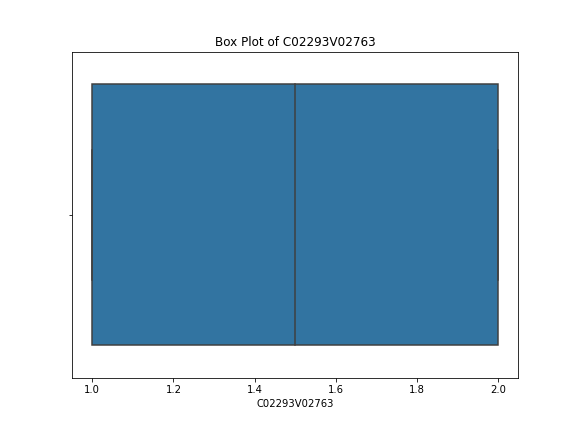
In our EDA, we are going to use summary stats, data visualization, and correlation analysis to capture some of the main insights from the data.

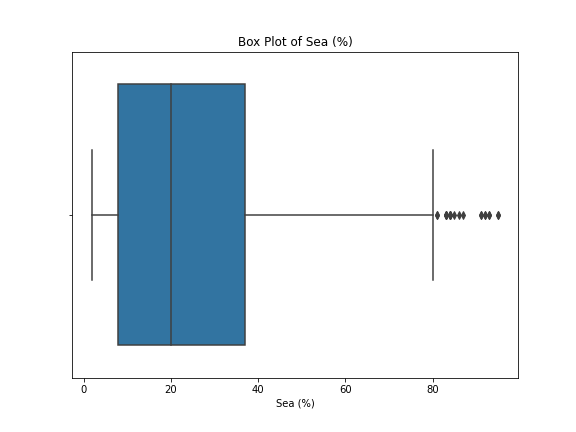
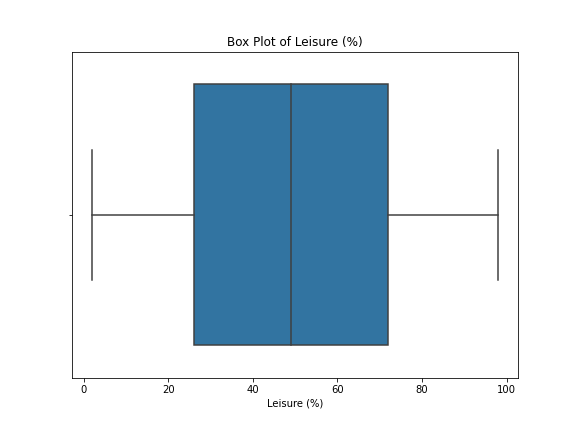
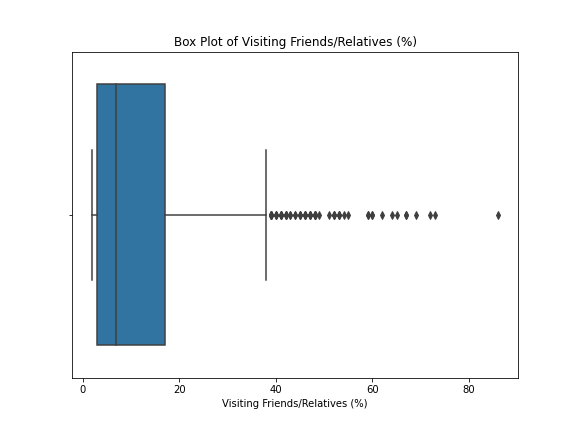
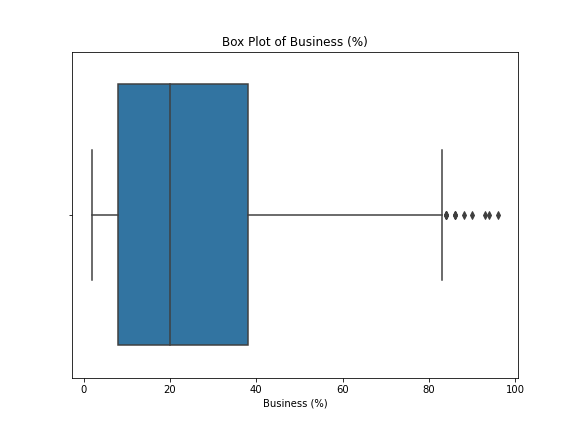
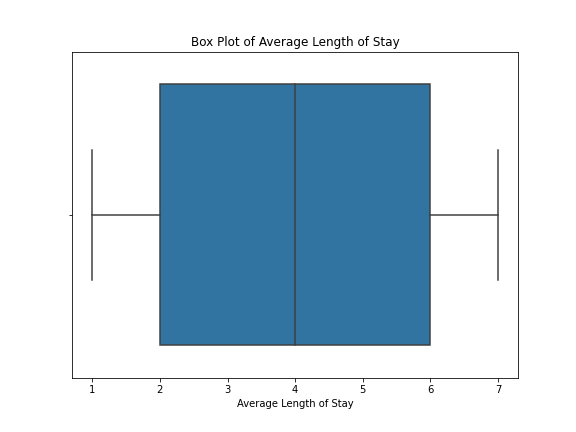
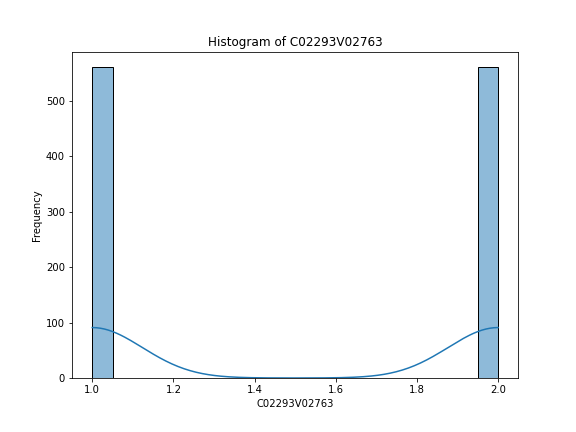
* **Summary Statistics:** Compute descriptive statistics such as mean, median, standard deviation, minimum, and maximum for numerical variables. For categorical variables, calculate frequencies and percentages of different categories.

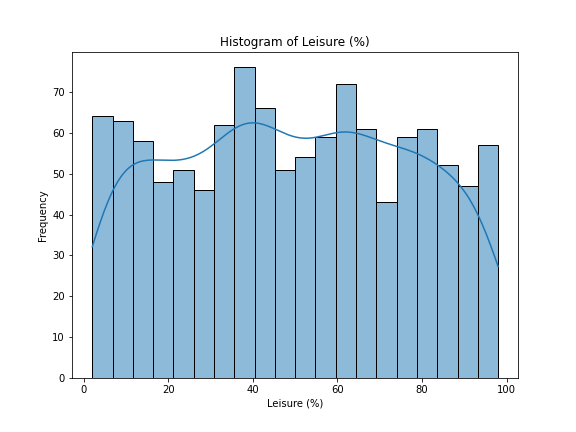
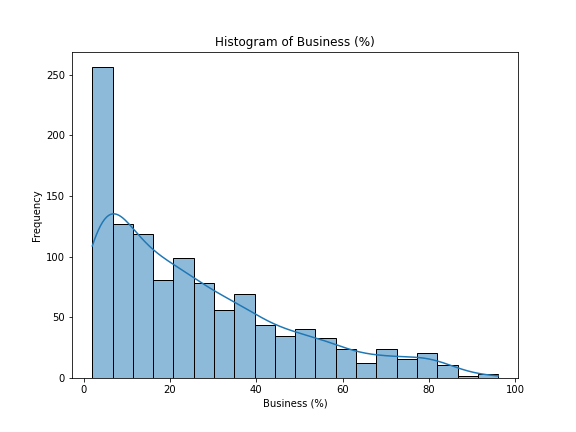


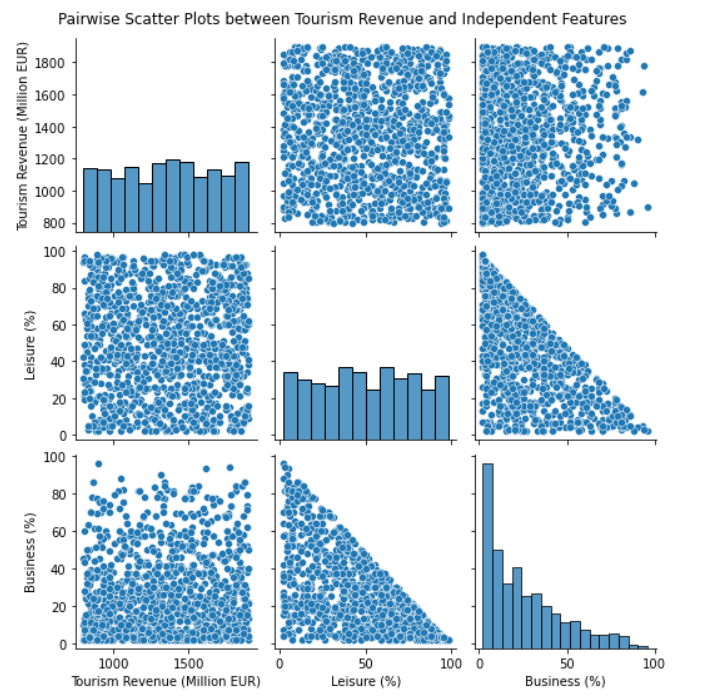
1. Count: This represents the number of non-missing values for each categorical variable. For example, for the 'STATISTIC', 'STATISTIC Label', 'C03622V04362', 'Country', 'Direction', and 'UNIT' columns, there are 1120 non-missing values.
2. Unique: This indicates the number of unique categories within each categorical variable. For instance, the 'STATISTIC' and 'STATISTIC Label' columns have only one unique category, while the 'Country' and 'Direction' columns have 40 unique countries and 2 unique directions, respectively.
3. Top: This represents the most frequent category within each categorical variable . Concretely, if we take an instance of the 'Country' column, 'All countries' is the most common category, appearing 28 times out of 1120 observations.
4. Frequency: It presents the frequency or ratio of most frequent category. Use our AI to write for you about any topic! As shown by the instance above, in the 'Country' column,'All countries' is found 28 times.
   * **Data Visualization**: Use different types of charts like histograms, box plots, scatterplots, barplots, or any other suitable ones to visualize the distribution and relationship among the variables. Such information will help distinguish the normative distribution from outliers and other possible patterns.

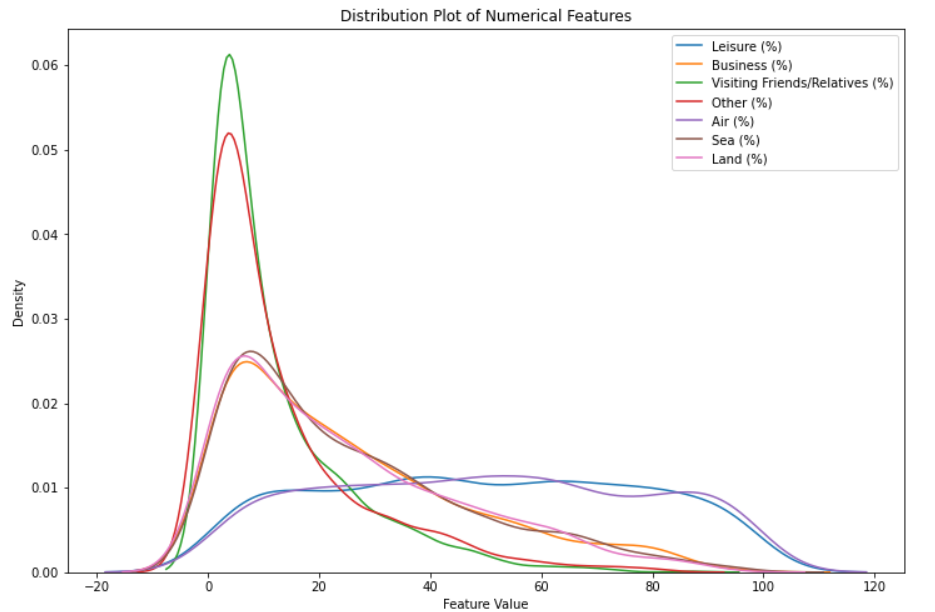


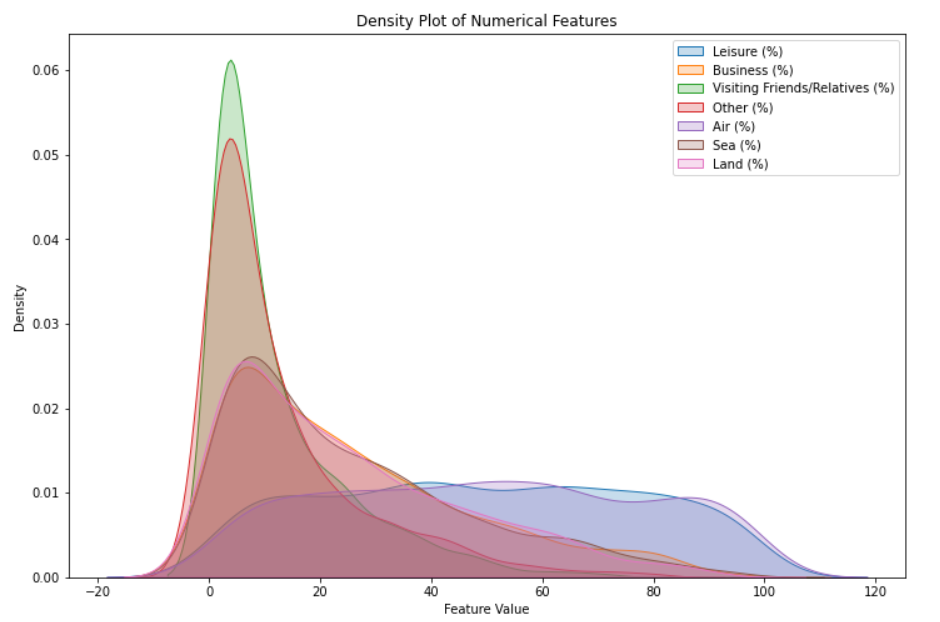


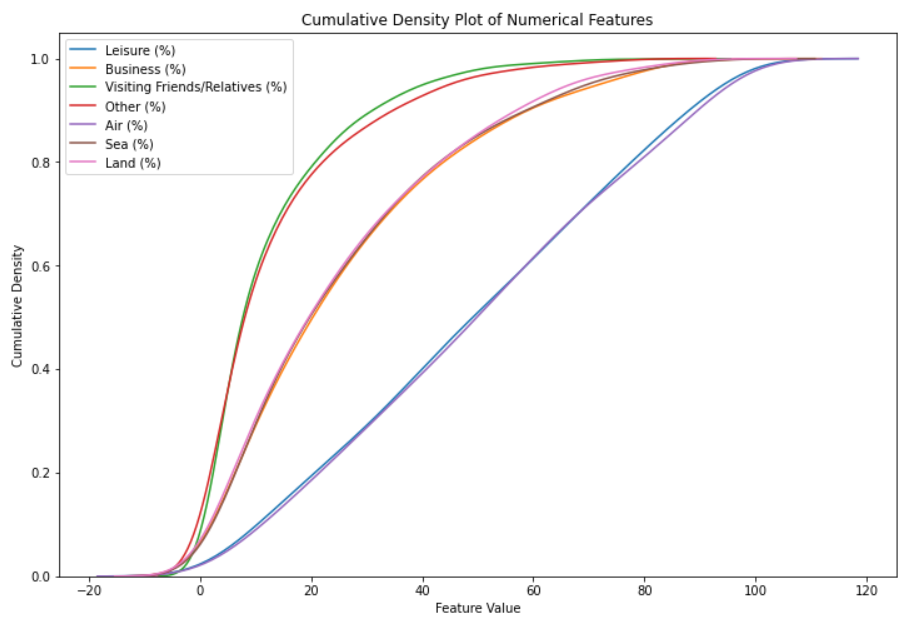


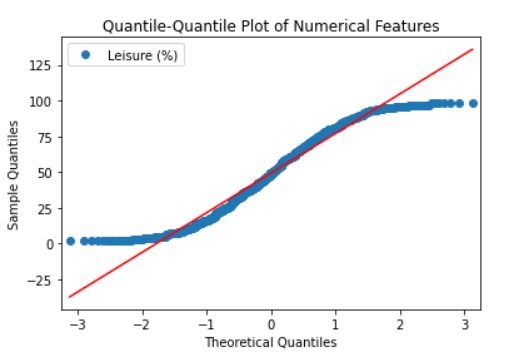
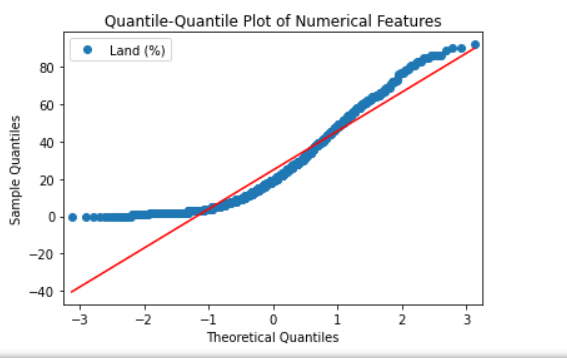




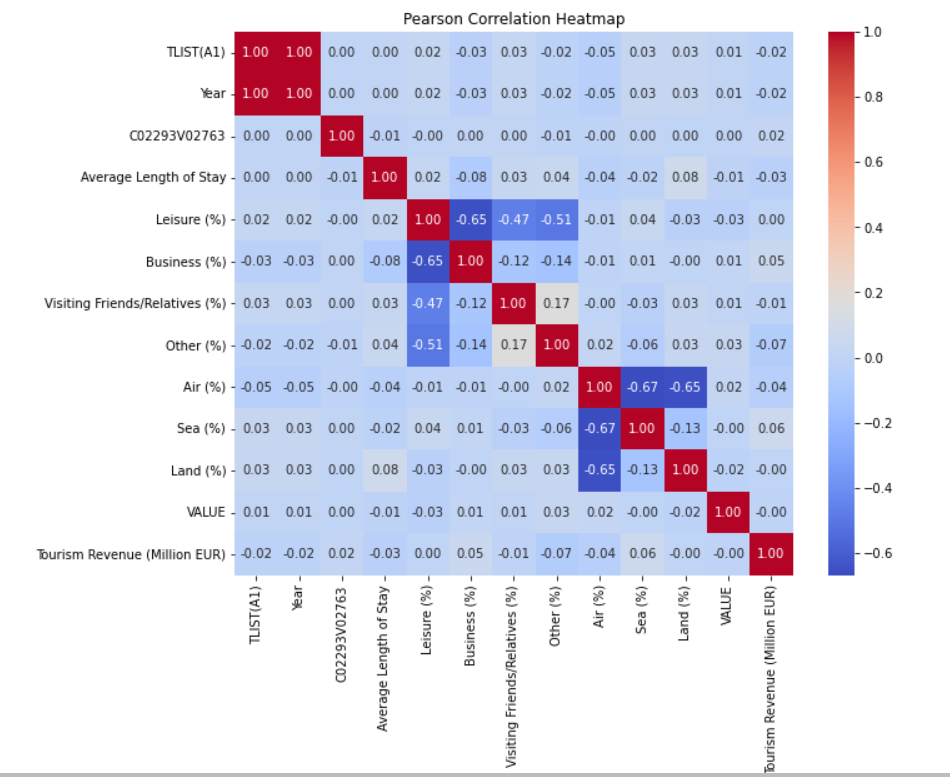




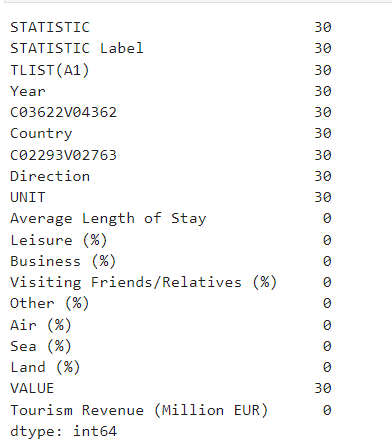




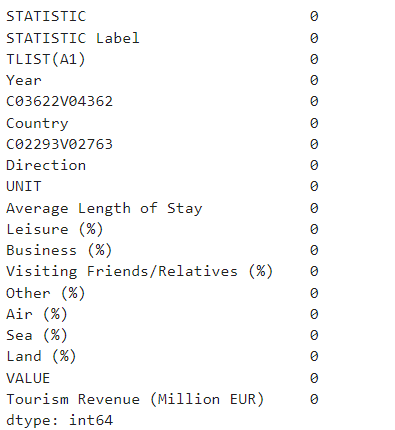
* + **Correlation Analysis:** Uncover correlation coefficients in the case of the numerical variables to discover any linear connection. Visualize relationships using heatmaps to the see the strongest direction and levels of relationships between the variables.



* **Handling Missing Values:** Verify the data set, if it has any missing values and determine strategies for handling it. One example of handling missing values is imputation and removal.



Now look at this after handling the above missing values;



#### Insights and Rationalization:

* **Summary Statistics:** Shows the average of the data and displays how spread out the data points are along with the shape of the distribution of numerical variables. It assists in the remarking of excessive values such as outliers and unsettled issues.
* **Data Visualization**: Can be useful in visualizing how variables are distributed if used for identifying possible outliers as well as seeing if there are any relationships between variables.
* **Correlation Analysis**: It detects linear relationships of numerical features, and they are useful as a criterion for feature selection or dimensionality reduction.
* **Handling Missing Values:** Enables one's choice between filling in values where they are missing or removing them depending on their affect to the analysis.

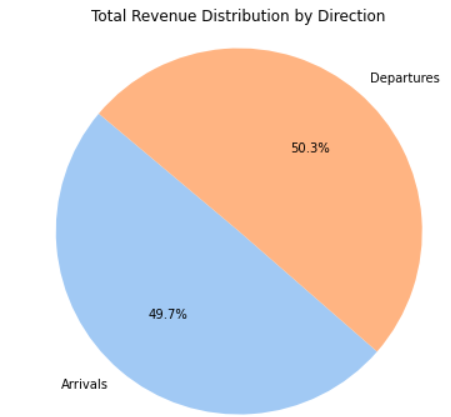
## B. The data Methods applied during data preparation for ML

Usually, some preprocessing steps are performed before the data become suitable for machine learning (ML) models. These steps include data cleaning, data normalization, data scaling and feature encoding. Here's how we can rationalize, justify, and detail the methods used for data preparation:

* **Scaling:** Scaling is very much requisite when the elements in the dataset are by various factors. The main problem that might arise as a result of using gradient descent-based methods or distance-based algorithms is the machine learning algorithms’ sensitivity to feature scale. Statistically transforming to the same scale is pretty much the most common scaling technique, and the one’s name are Min-Max and Standardization (Z-score normalization).
* **Encoding Categorical Variables:** With the majority of ML algorithms depending on numerical input data, it is imperative that categorical variables be converted into numerical formats that they understand. Usually, for a categorical one-hot encoding and a label encoding are in use. One-hot encoding is better off when categories have no ordinal relationship to be it, while label encoding serves as a great method when there is a natural order among categories.
* **Imputation of Missing Values:** Mostly models are made up of missing values so they are likely not to perform at a high level. Depending on the form data is missing, its values can be imputed with several approaches, being the most common are mean, median, or mode imputation and more complex techniques such as k-nearest neighbors (KNN) imputation and iterative imputation methods.
* **Feature Engineering**: Feature engineering embraces the introduction of new features as well as the transformation of the existing features to a representation which accords with the unique patterns existing in the data. For example, these activities may involve polynomials, interaction terms or particular transformations that can help to improve modeling results.
* **Feature Selection:** The use of feature selection approaches allows for the sifting of out the most relevant features for training of the model. Furthermore, it is responsible for the reduction of the dimension of the dataset, thus boosting model accuracy. Finally, it makes the models more readable and generalized. There are many techniques employed including correlational analysis, feature selection via recursive feature elimination, and feature importance ranking.
* **Handling Imbalanced Data:** In the classification problem, uneven class distribution can lead to the lopsided models, that is why we use methods of rebalancing sets of data. Techniques such as oversampling, under-sampling, or the use of ensemble methods like SMOTE (Synthetic Minority Over-sampling Technique) can help address class imbalance issues.

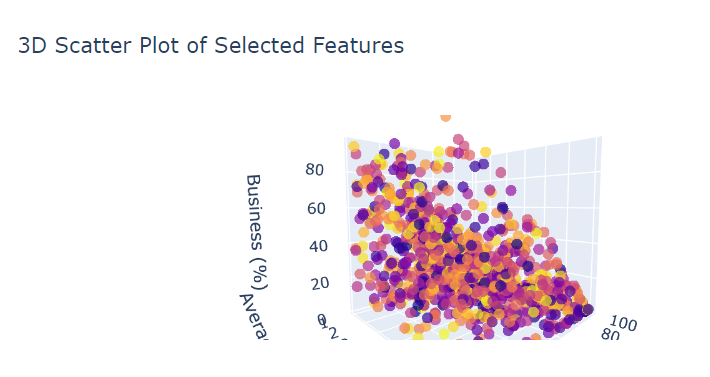


## C. Final Insights

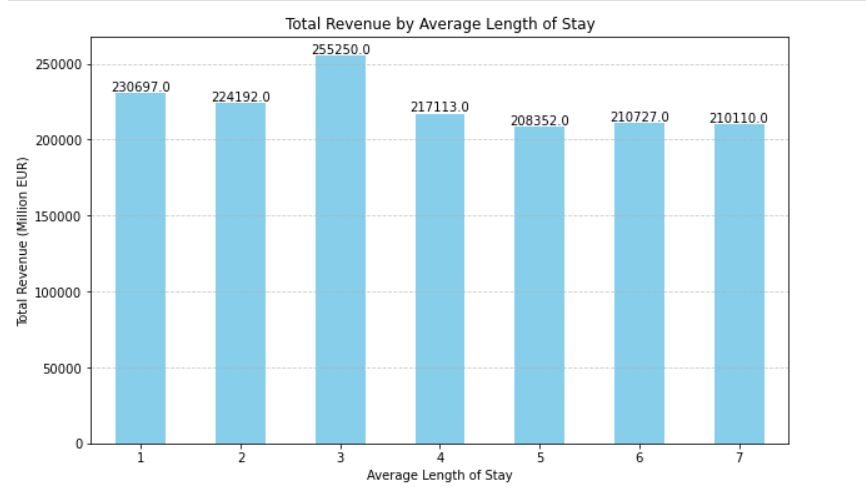


* The pie chart shows the proportion of revenue from arrivals compared to departures.
* The blue slice represents arrivals, and it accounts for 49.7% of the total revenue.
* The orange slice represents departures, and it accounts for 50.3% of the total revenue.

Since the orange slice is slightly larger than the blue slice, we can say that departures account for a slightly higher proportion of the total revenue than arrivals.



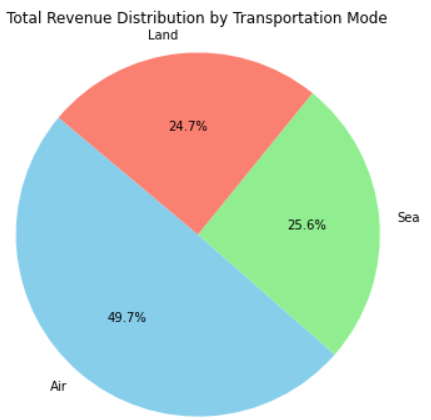
* This 3D Scatter plot shows the effect of the features on the target variable efficiently as we can move the graph. Also, each point on the graph has its own value and by moving cursor on the point shows their value.



* This bar graph shows the total revenue by average length of stay in the Ireland tourism. The y-axis shows the total revenue in millions of Euros and the x-axis shows the average length of stay in days.

Here’s a more detailed breakdown of the information in the graph:

* The title of the graph is “Total Revenue by Average Length of Stay.”
* The y-axis title is “Total Revenue (Million EUR)”. This means that the values on the y-axis represent the amount of money that was brought in by tourism in millions of Euros.
* The x-axis title is “Average Length of Stay.” This axis shows the average number of days that tourist stayed in the Ireland.



The above pie chart shows the total revenue distribution by transportation mode. It breaks down the revenue generated by different modes of transportation used to travel to a destination.

* Air: The largest slice of the pie chart, labeled "Air," accounts for 49.7% of the total revenue.
* Sea: The second-largest slice, labeled "Sea," accounts for 25.6% of the total revenue.
* Land: The smallest slice, labeled "Land," accounts for 24.7% of the total revenue.

## D. Visualization with Tuft’s Principle

**Pie Chart of Arrivals vs. Departures Revenue:**

* Clarity: The use of distinct colors (blue for arrivals and orange for departures) ensures clarity in distinguishing between the two categories. Additionally, the percentage values provide clear quantitative information [[3]](#_References).
* Accuracy: The proportions accurately represent the distribution of revenue between arrivals and departures, providing an accurate depiction of the dataset.
* Relevance: This visualization is relevant as it addresses the specific comparison between revenue generated from arrivals and departures, which is likely of interest in tourism analysis.

**3D Scatter Plot for Feature Analysis:**

• Clarity: The application of 3D scatter plot that can illustrate the three-way profiles, thus, clarify the association between variables and the target variable.

• Accuracy: The interactive design makes it easier to explore the data of interest precisely, leading to a proper interpretation of the values.

• Relevance: The reason for the usefulness of this particular visualization is that it allows drawing conclusions on the presence of patterns and the connections between the features and the dependent variable that can explain the effect of the factors on revenues from tourism.

**Bar Graph of Total Revenue by Average Length of Stay:**

• Clarity: This line graph displays the total income, which is highlighted by different average lengths of stay, in a clear and visually appealing way, with each category accompanied with different bars.

• Accuracy: The graph represents the total revenue connected to each length of stay, capturing the pattern of distribution of a revenue stream as well.

• Relevance: The visualization presented is meaningful and it explains why the average length of stay and total revenue is interconnected which are essential for understanding revenue generation through tourists behavior.

**Pie Chart of Revenue Distribution by Transportation Mode:**

• Clarity: The pie chart is well designed as it segregates the revenues obtained from different modes of transportation with different colors for each of them to aid in the easy understanding of their levels of contribution.

• Accuracy: The figures put in the table proportionally reflect how much the revenues can be contributed to each transport mode correctly.

• Relevance: Such illustration is important due to the fact that it deals with the division of revenue by means of transportation modes, indicating the modes that generate most revenue either individually or jointly.

# Section-III (Machine learning for Data Analytics)

## A. a)Project Management Framework

A data-science project is commonly structured and begun using the CRISP-DM (Cross-industry Standard Process for Data Mining) framework. CRISP-DM is a framework that combines all the phases of a data mining project such as understanding your business objectives, understanding the data, data preparation and modelling, evaluation, and deployment, together into a structured process.

### Justification with real-life Scenario

• Understanding Business Objectives: For example, the first step of CRISP-DM in a real-world situation where one is analyzing customer churn (telecommunication industry) is to define the main business objective. The objective could be reducing churn rates to improve profitability.

• Data Understanding: This step is aimed at data sampling and the first intuitive data analysis. For example, using customer data points like demographics, behavioral pattern, and service subscribes to analyze factors, which cause churn.

• Data Preparation: In this phase, data preprocessing techniques are employed to clean, transform, and to shape up the data to suit the modeling needs. For instance, dealing with missing values, working with categorical variables, and handling the scaling feature of numerical fields.

• Modeling: Based on what the business goals are and what is known about the data, different machine learning techniques can then be selected which can then be trained on the prepared data. Customer churn prediction is made possible via the use of algorithms such as logistic regression, decision trees, or random forests in the field of supervised learning.

• Evaluation: Models’ effectiveness is evaluated using suitable metrics (e.g., accuracy, precision, recall) to determine model performance. Thus while assessing the effectiveness of a churn prediction model, one may study its ability to correctly classify churners and non-churners.

• Deployment: The model is finally deployed in production where is performance is monitored as time goes on. The churn prediction model in telecom case could be interfaced with the customer relationship management system of the company in order to flag customers who are in the risk of churning and take proactive retention measures.

Regarding the choice of supervised, unsupervised, or semi-supervised machine learning technique, it depends on the nature of the dataset and the problem at hand:

**Supervised Learning:** Given the case that the dataset includes labelled data (e.g. historical customer data, with churn labels), and, the goal, is to predict a particular output (e.g. churn or not), supervised learning techniques are applicable. This is because supervised learning agents learn from labeled samples to predict on unlabeled data.

**Unsupervised Learning:** When the data is not labelled or the goal is to detect the hidden patterns or the structures within the data, the unsupervised learning techniques are favored. To illustrate, clustering algorithms like K-means that group customers of common characteristics are one example.

**Semi-Supervised Learning:** In a situation where labeled data is sparse but unlabeled data is rich, these semi-supervised learning techniques can be advantageous. These methods use the labeled as well as the unlabeled data to increase model performance. For example, include a small labeled dataset together with larger unlabeled dataset in training.

## b). Selection ML Technique and Models for our Dataset

**ML Technique:** Supervised Learning

**Problem Type:** Regression (predicting a numerical target variable).

**Dataset Characteristics:**

* **Features:** STATISTIC, Average Length of Stay, Leisure (%), Business (%), and other relevant features.
* **Target Variable:** Tourism Revenue (Million EUR).

**Model Selection Criteria:**

* Interpretability: Prefer a model that provides insights into the relationship between features and the target variable.
* Model Complexity: Balance between model complexity and performance.
* Performance Metrics: Mean Squared Error (MSE), R-squared (R2), and other regression evaluation metrics.

**Model Options:**

* Linear Regression: Provides simple and interpretable results, suitable for understanding linear relationships between features and the target variable.
* Decision Trees: Can capture non-linear relationships and interactions between features, potentially offering higher predictive performance.
* Random Forests: Ensemble of decision trees that often provide better generalization and robustness.

## B. Approach of Hyperparameter tunning

To address this requirement, we'll assess at least two machine learning approaches for our dataset and tune hyperparameters using either GridSearchCV or RandomizedSearchCV. Since we've already discussed using regression for predicting tourism revenue as a common scenario, I'll proceed with two regression approaches: Linear Regression and Random Forest Regression. Here's the plan:

Gradient Boosting Regression**:**

* Approach: This boosting regression model works as combining weak models and from them make a strong model. It improves the performance by minimizing the loss function [[5]](#_References).
* Hyperparameter Tuning: We're going to tune our hyperparameters using GridSearchCV such as the learning rate, the maximum depth of each tree (max\_depth) etc.

### Decision Tree Regression

* Approach: This regression model is actually non-parametric supervised learning that is mainly used for regression and classification both models [[6]](#_References)
* Hyperparameter Tuning: We have used GridSearchCV to find the best combination by using hyperparameter space and the parameters include minimum samples split, and minimum samples leaf etc.

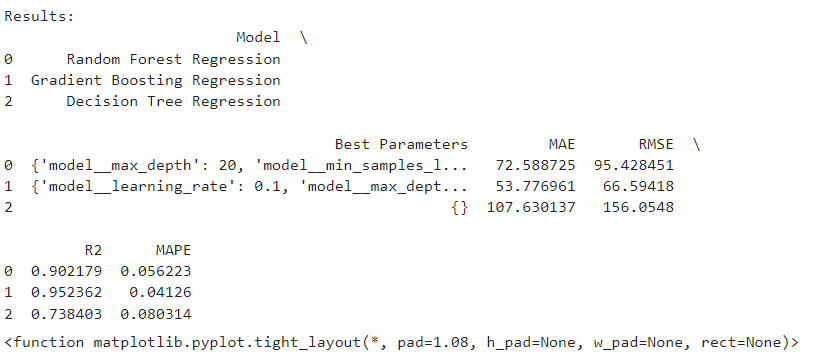
### Random Forest Regression:

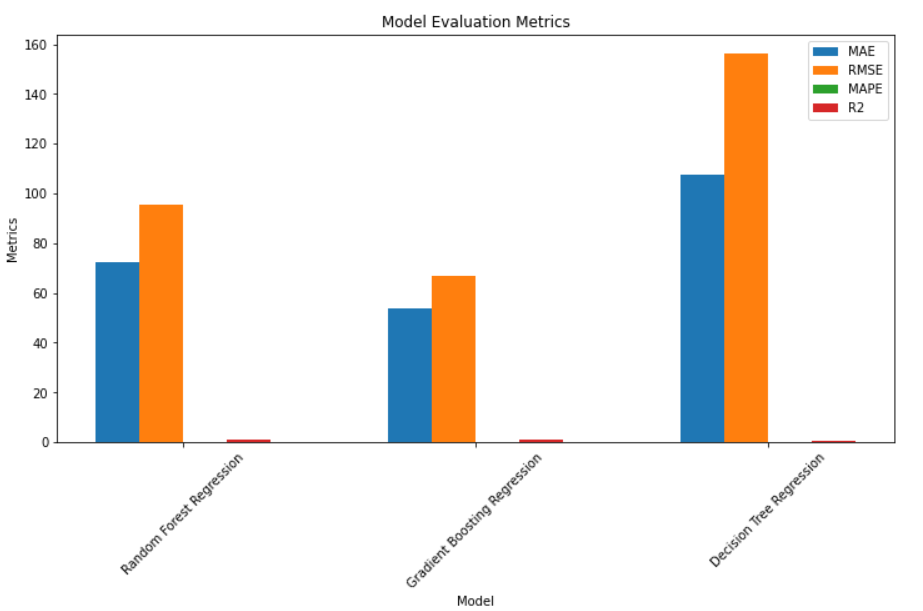
* + Approach: The random forest regression is the ensemble learning method utilizing the decision trees as nodes. It is more widely known for its flexibility and complex relationship capturing ability to do this. [[7]](#_References)
  + Hyperparameter Tuning: We're going to tune our hyperparameters -- the number of trees in the forest (n\_estimators), the maximum depth of each tree (max\_depth), the size of a minimum sample per leaf (min\_samples\_leaf), and the maximum number of available features considered for splitting (max\_features).

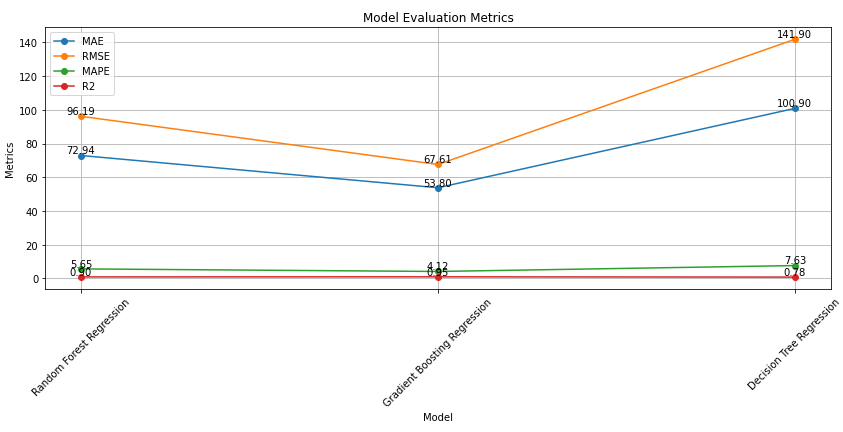
To perform hyperparameter tuning, we'll make use of the GridSearchCV method, which comprehensively searches a designated grid of hyperparameter settings, or the RandomizedSearchCV approach, which randomly samples grid of parameter candidates set by the user. It will be the aim of both procedures to give us the best hyperparameters for each model.

Lastly, we'll adjust hyperparameters of models and train them using indices for their performance that can be Mean Squared Error (MSE), R2 or other similar measures.

## C. ML modeling comparisons







**Random Forest Regression:**

* MAE (Mean Absolute Error): 72.59
* RMSE (Root Mean Squared Error): 95.43
* R2 (R-squared): 0.902
* MAPE (Mean Absolute Percentage Error): 5.62%
* Best Parameters: Maximum depth of 20 and minimum samples leaf of 1

Random Forest Regression came with R-squared value of 0.902, which shows that the 90.2% of the target variable's variance was explained by the chosen model. Thus, the MAE of 72.59 means that, in the model, for about 72.59 million Euros, one should expect the prediction to be off, on average.

**Gradient Boosting Regression:**

* MAE: 53.78
* RMSE: 66.59
* R2: 0.952
* MAPE: 4.13%
* Best Parameters: Learning rate of 0.1 and maximum depth of 5

Gradient Boosting Regression performs even better with an R-squared value of 0.952, indicating that it explains approximately 95.2% of the variance in the target variable. The MAE of 53.78 indicates that, on average, the model's predictions are off by approximately 53.78 million Euros.

**Decision Tree Regression:**

* MAE: 107.63
* RMSE: 156.05
* R2: 0.738
* MAPE: 8.03%
* Best Parameters: No specific parameters

Decision Tree Regression performs the worst among the three models, with an R-squared value of 0.738. The MAE of 107.63 indicates that, on average, the model's predictions are off by approximately 107.63 million Euros.

## D. Demonstration

**Similarities:**

* All three models aim to predict tourism revenue based on various features.
* They use regression techniques to make predictions.
* The models are evaluated using common performance metrics such as MAE, RMSE, R-squared (R2), and MAPE.

**Differences:**

**Performance Metrics**

* MAE: Gradient Boosting Regression has the lowest MAE (53.78), indicating the smallest average prediction error, followed by Random Forest Regression (72.59) and Decision Tree Regression (107.63).
* RMSE: Gradient Boosting Regression also has the lowest RMSE (66.59), indicating better accuracy in predicting revenue compared to the other models.
* R-squared (R2): Gradient Boosting Regression has the highest R2 value (0.952), indicating the highest proportion of variance in the target variable explained by the model, followed by Random Forest Regression (0.902) and Decision Tree Regression (0.738).
* MAPE: Gradient Boosting Regression has the lowest MAPE (4.13%), indicating the smallest percentage of error in predictions relative to the actual revenue values, followed by Random Forest Regression (5.62%) and Decision Tree Regression (8.03%).

**Hyperparameters:**

* Each model has different optimal hyperparameters, which affect their performance. For example, Gradient Boosting Regression has a learning rate of 0.1 and a maximum depth of 5, while Random Forest Regression has a maximum depth of 20 and a minimum samples leaf of 1.

**Model Complexity:**

* Gradient Boosting Regression and Random Forest Regression are ensemble methods that combine multiple weak learners to form a stronger model, whereas Decision Tree Regression relies on a single decision tree.
* Gradient Boosting Regression tends to perform better due to its ability to correct errors of previous models in the ensemble during the training process.

**Relevance and Effectiveness:**

* The findings indicate that Gradient Boosting Regression is the most effective model for predicting tourism revenue, as it achieves the lowest errors and highest predictive accuracy.
* Random Forest Regression also performs well, providing accurate predictions with slightly higher errors compared to Gradient Boosting.
* Decision Tree Regression, while simpler, exhibits higher prediction errors and lower accuracy compared to the other models.

# References

[1] Ghosh, I., Marques, F., & Chakraborty, S. (2023). Bivariate binomial conditional distributions with positive and negative correlations: A statistical study.

[2] Roy, S., Tripathi, R. C., & Balakrishnan, N. (2023). A Conway–Maxwell–Poisson Type Generalization of Hypergeometric Distribution.

[3] Midway, S. R. (2020). Principles of Effective Data Visualization. Pattern Recognition Letters

[4] Wu, J. (2020). Chapter 13: The Normal Distribution. Nanjing University, China.

[5] Dahiya, N., Saini, B., & Chalak, H. D. (In Press). Gradient boosting-based regression modeling for estimating the time period of the irregular precast concrete structural system with cross bracing. Journal of King Saud University - Engineering Sciences. Available online 15 August 2021.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30 (NIPS 2017).

[7] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest, and KNN Models for Text Classification. Augmented Human Research, 5(12).