Table of Contents

[Section-I (Statistics) 2](#_Toc162384625)

[A. Descriptive Statistics 2](#_Toc162384626)

[1. Dataset Shape: 2](#_Toc162384627)

[2. Column Names: 2](#_Toc162384628)

[3. Data Types: 2](#_Toc162384629)

[4. Summary Statistics: 3](#_Toc162384630)

[5. Missing Values: 4](#_Toc162384631)

[Critical Analysis: 5](#_Toc162384632)

[B. Visualization with Plots 5](#_Toc162384633)

[1. Histogram: Distribution of Tourism Revenue 5](#_Toc162384634)

[2. Heatmap of Correlation Matrix 6](#_Toc162384635)

[3. Bar Graph 7](#_Toc162384636)

[4. Line Graph 8](#_Toc162384637)

[C. Discrete Distributions 9](#_Toc162384638)

[1. Binomial Distribution: 9](#_Toc162384639)

[2. Poisson Distribution: 10](#_Toc162384640)

[D. Normal Distributions 11](#_Toc162384641)

[E. Importance of the distributions used in point 3 and 4 12](#_Toc162384642)

[1. Binomial and Poisson Distributions (Point 3): 12](#_Toc162384643)

[2. Normal Distribution (Point 4): 12](#_Toc162384644)

[SECTION-II 13](#_Toc162384645)

[A. Data preparation and Visualization 13](#_Toc162384646)

[1. Exploratory Data Analysis 13](#_Toc162384647)

[B. The data Methods applied during data preparation for ML 18](#_Toc162384648)

[C. Final Insights 19](#_Toc162384649)

[D. Visualization with Tuft’s Principle 22](#_Toc162384650)

# Section-I (Statistics)

## A. Descriptive Statistics

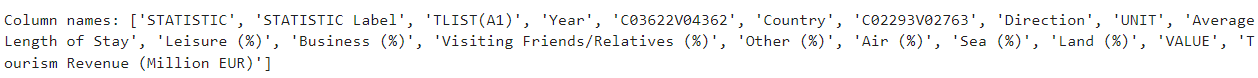
### 1. Dataset Shape:

* The dataset contains 1150 rows and 19 columns.



### 2. Column Names:

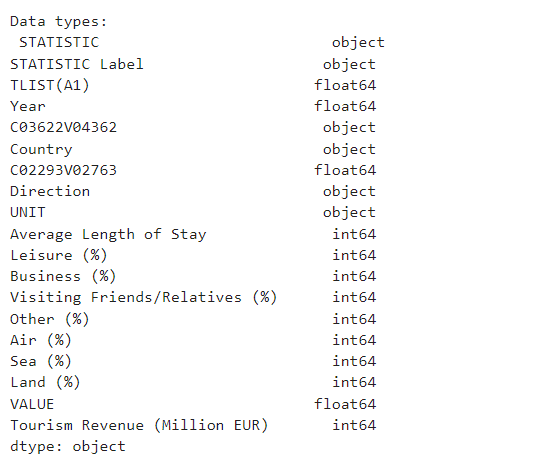
* The dataset has 19 columns with names such as 'STATISTIC', 'STATISTIC Label', 'TLIST(A1)', 'Year', 'Country', 'Direction', 'UNIT', 'Average Length of Stay', 'Leisure (%)', 'Business (%)', 'Visiting Friends/Relatives (%)', 'Other (%)', 'Air (%)', 'Sea (%)', 'Land (%)', 'VALUE', and 'Tourism Revenue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



### 3. Data Types:

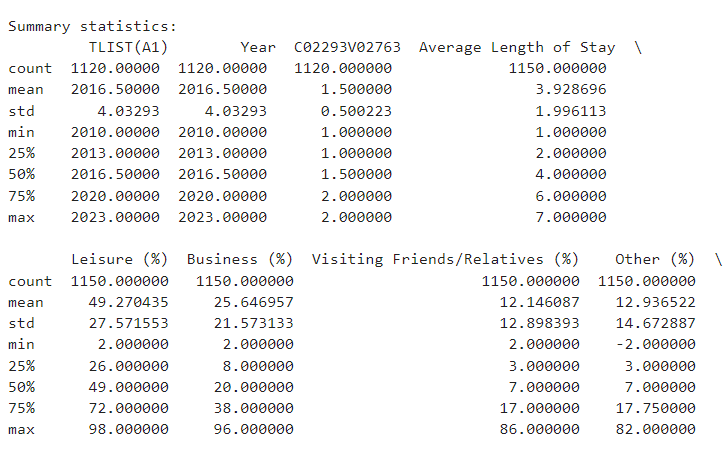
The data types of the columns vary:

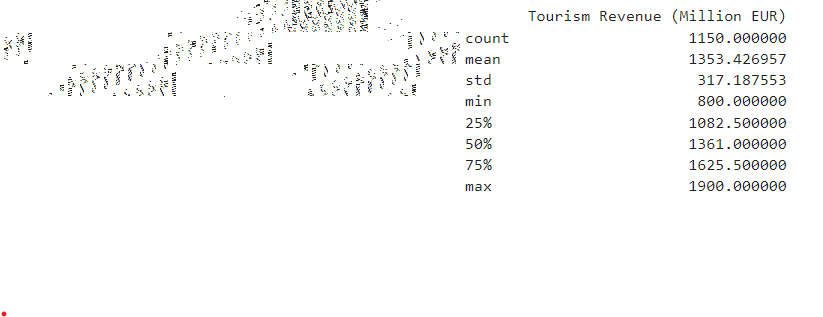
* 'object': STATISTIC, STATISTIC Label, C03622V04362, Country, Direction, UNIT.
* 'float64': TLIST(A1), Year, C02293V02763, VALUE.
* 'int64': Average Length of Stay, Leisure (%), Business (%), Visiting Friends/Relatives (%), Other (%), Air (%), Sea (%), Land (%), Tourism Revenue (Million EUR).



### 4. Summary Statistics:

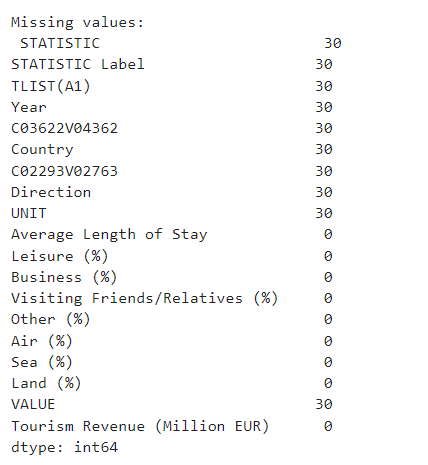
* TLIST(A1): Year column has a mean of approximately 2016.5 with a standard deviation of around 4.03. The data spans from the year 2010 to 2023.
* Average Length of Stay: The average length of stay ranges from 1 to 7 days, with a mean of approximately 3.93 days.
* Leisure (%), Business (%), Visiting 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

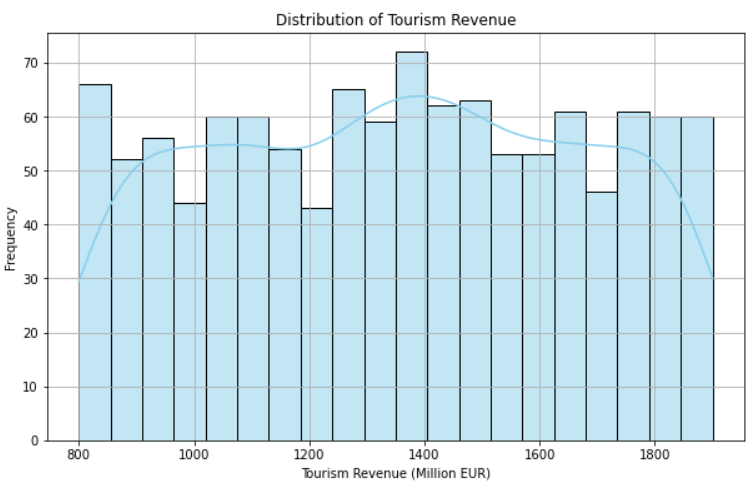
* X-axis: Represents the range of tourism revenue (in million EUR).
* Y-axis: Represents the frequency of occurrence for each revenue range.

**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.



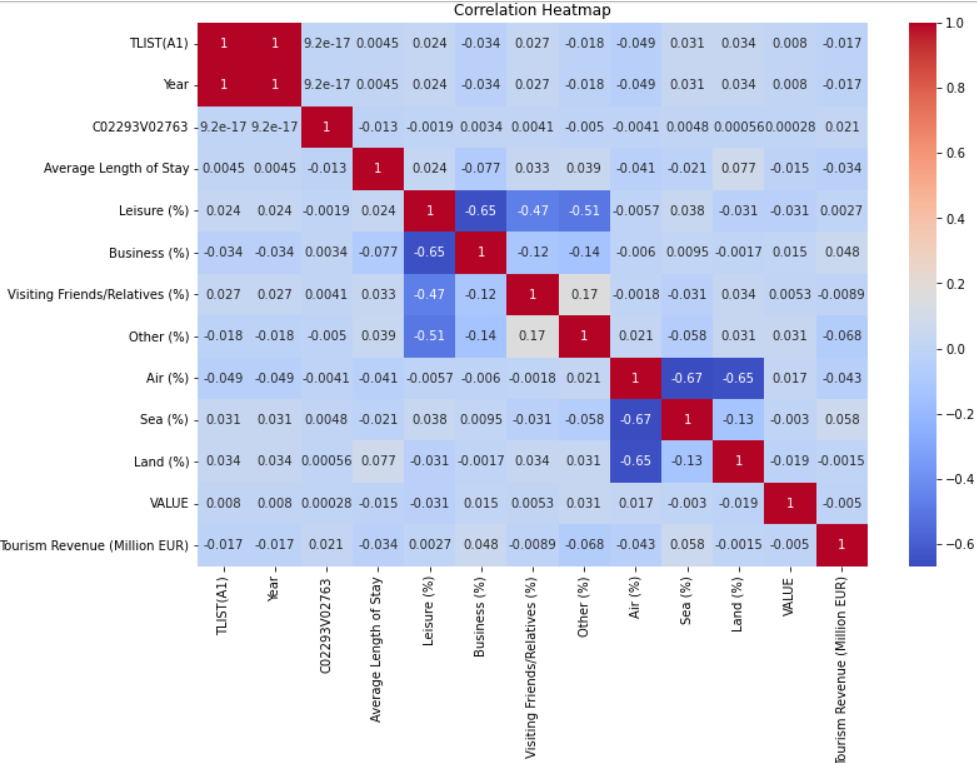
### 2. Heatmap of Correlation Matrix

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.

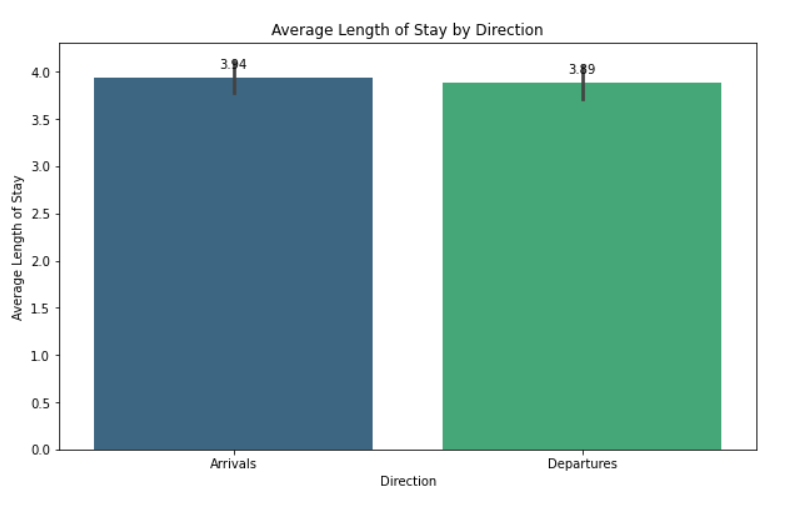


### 3. Bar Graph

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.

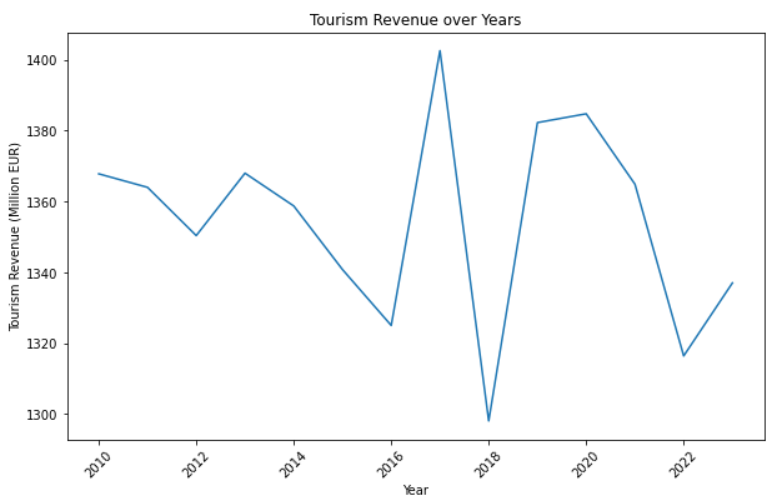


### 4. Line Graph

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:

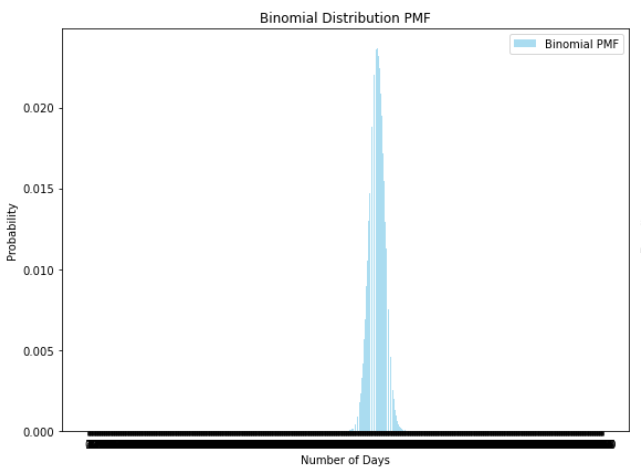
* The y-axis title is “Tourism Revenue (Million EUR)”. This means that the values on the y-axis represent the amount of money that was brought in by tourism in millions of Euros.
* The x-axis title is simply “Year”. This axis shows the years that the data was collected for.
* The line starts at a value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

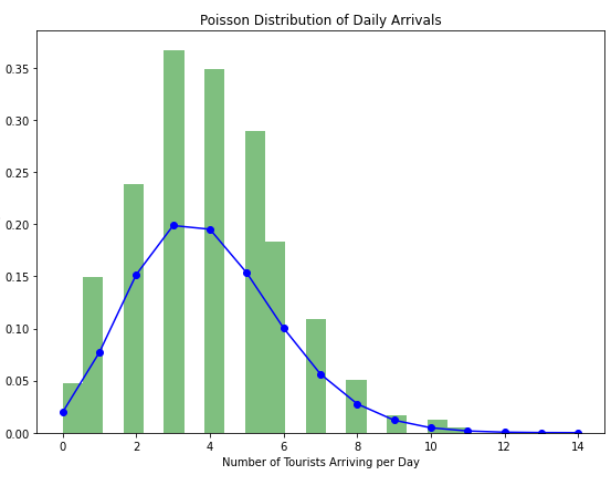


• The x-value of the graph exhibit the number of days between visits, and the y-value display the probability of that days per day between visits occurrence.

• "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 analyse the number of events happening during a single period of time or space given information about the average number of events. 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.

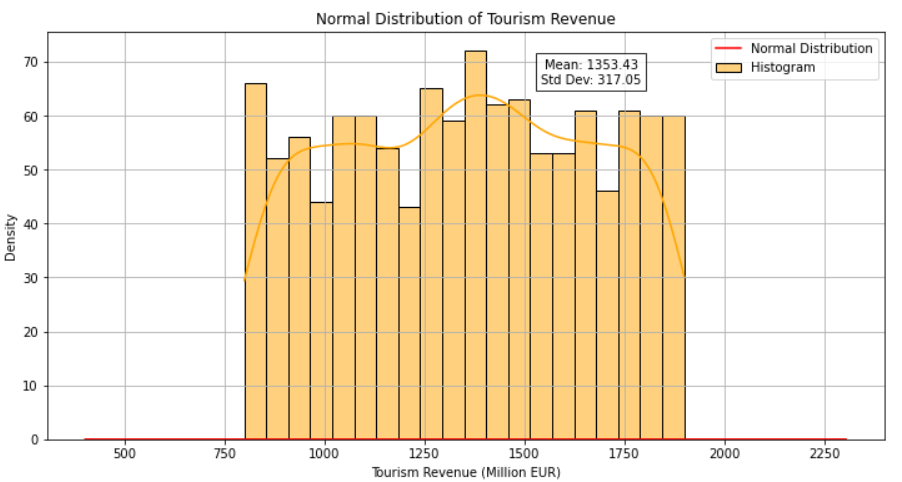
The Poisson distribution would be suitable for the number of daily tourist arrivals fitting as it applies to cases when events happen independently and their impact on the following events does not exist this exponential arrival/travel station assumes a constant rate.

**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).



* 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.

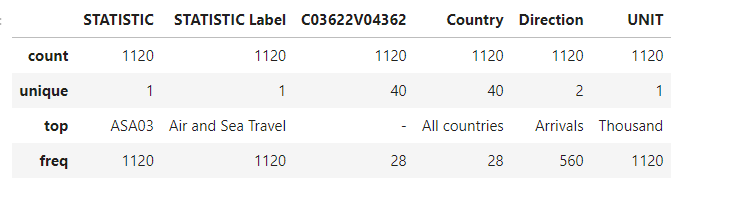
# SECTION-II

## A. Data preparation and Visualization

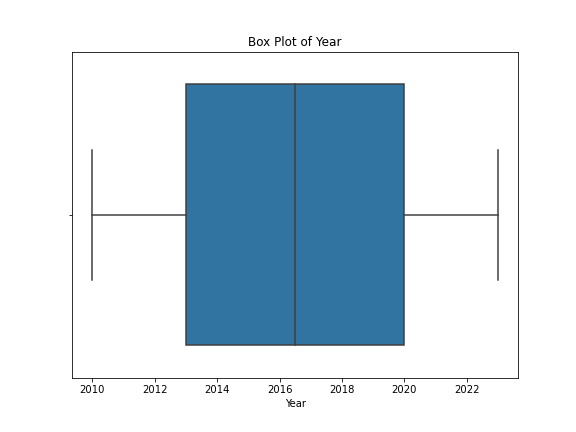
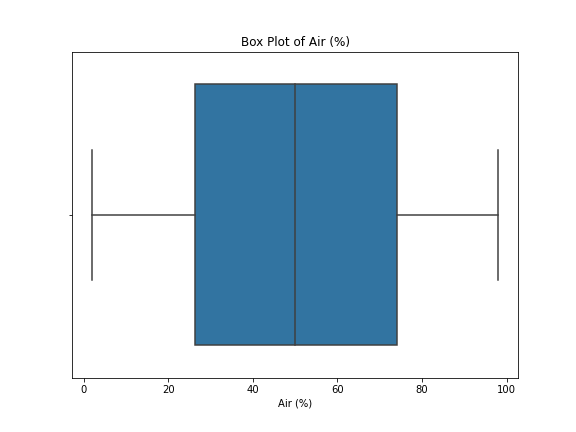
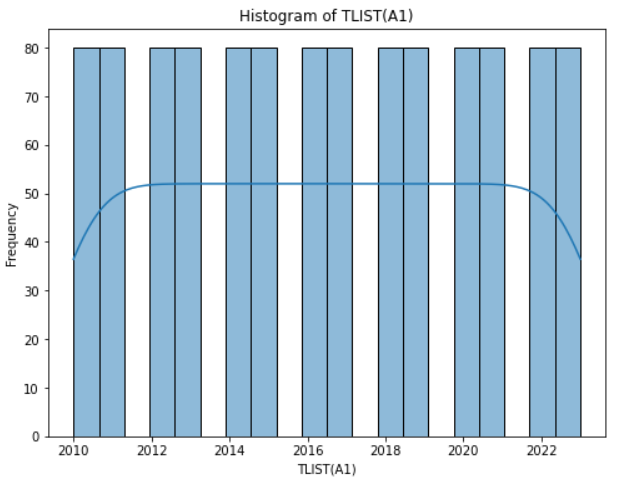
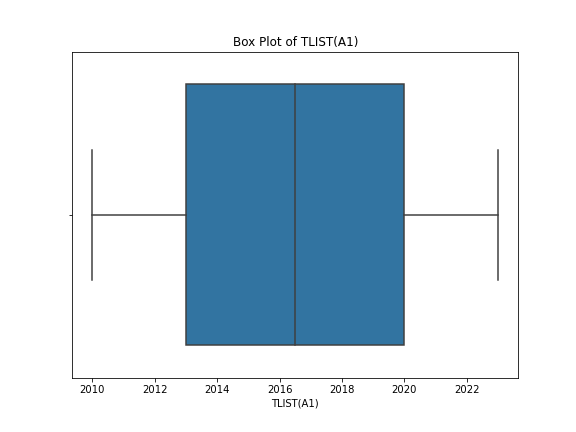
### 1. Exploratory Data Analysis

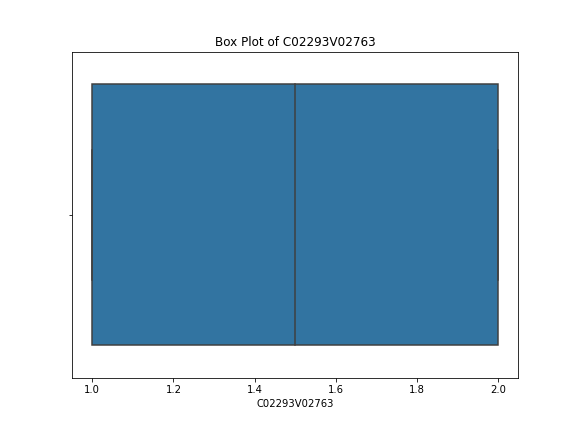
In our EDA, we are going to use summary stats, data visualization, and correlation analysis to capture some of the main insights from the data.

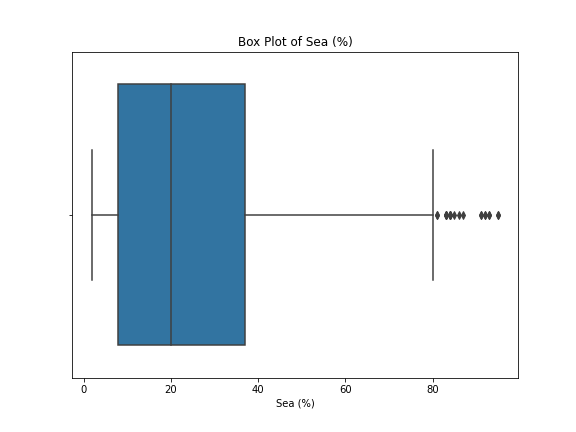
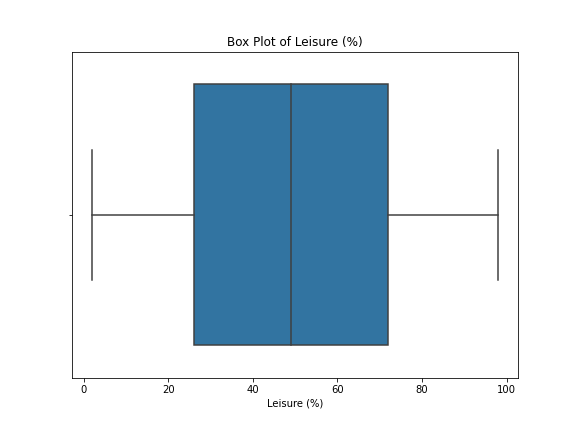
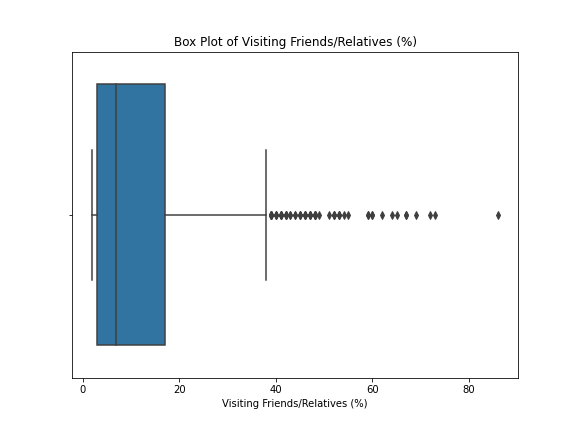
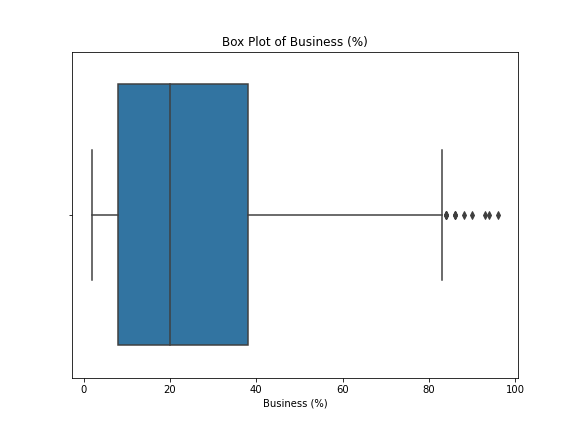
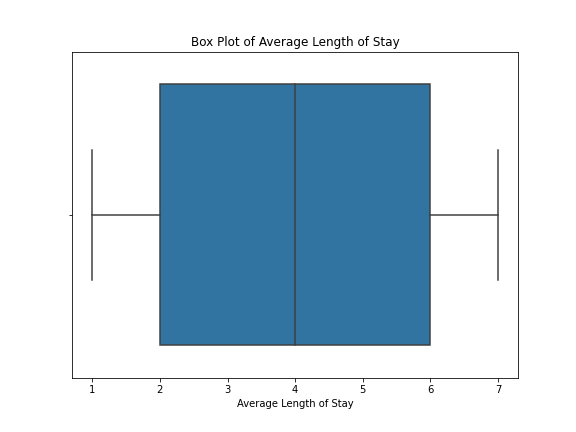
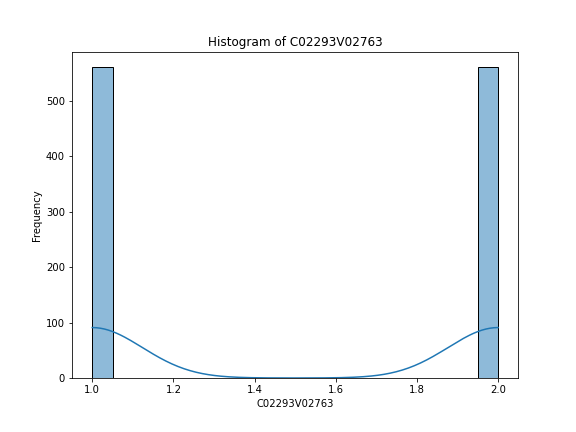
* **Summary Statistics:** Compute descriptive statistics such as mean, median, standard deviation, minimum, and maximum for numerical variables. For categorical variables, calculate frequencies and percentages of different categories.

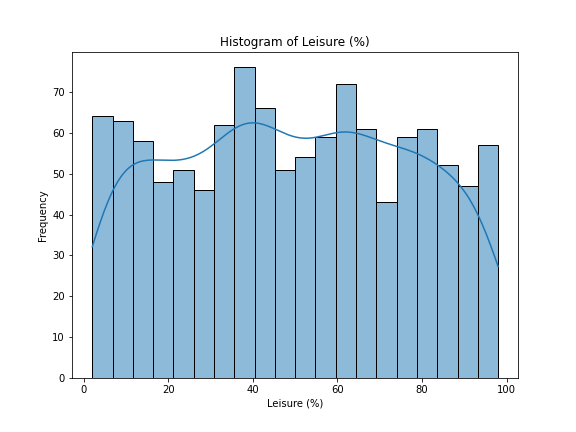
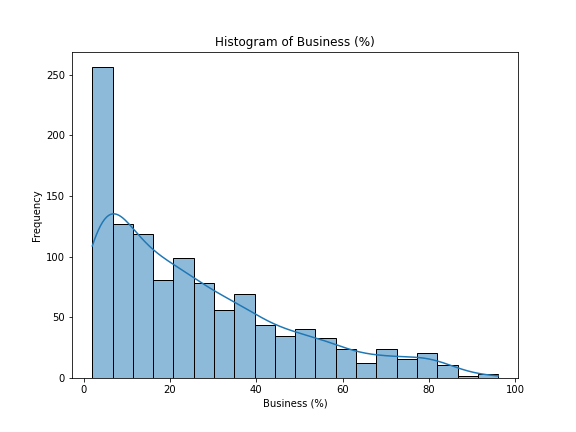


1. Count: This represents the number of non-missing values for each categorical variable. For example, for the 'STATISTIC', 'STATISTIC Label', 'C03622V04362', 'Country', 'Direction', and 'UNIT' columns, there are 1120 non-missing values.
2. Unique: This indicates the number of unique categories within 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.

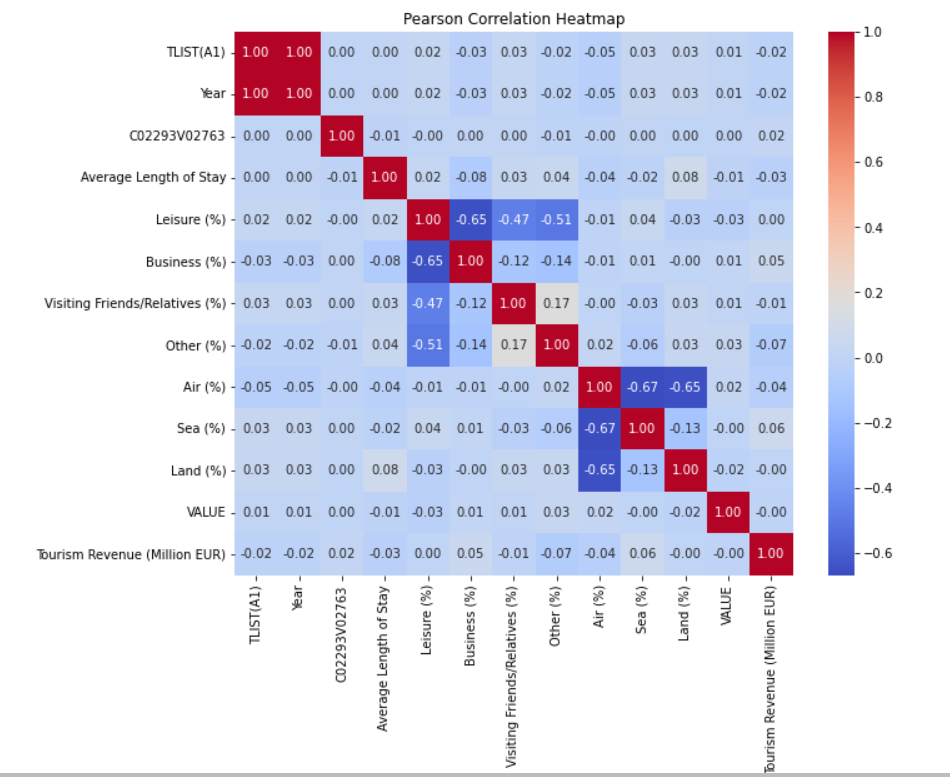




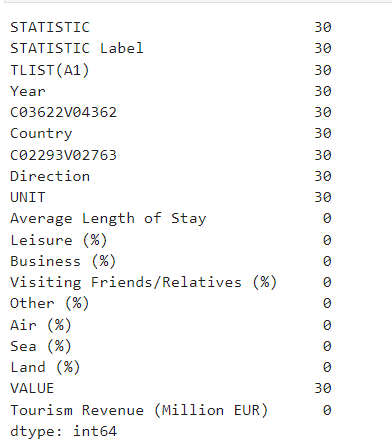




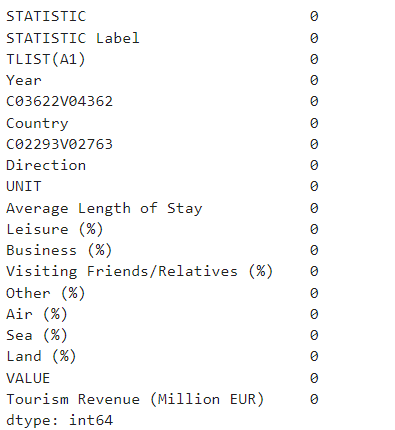
* + **Correlation Analysis:** Uncover correlation coefficients in the case of the numerical variables to discover any linear connection. Visualize relationships using heatmaps to the see the strongest direction and levels of relationships between the variables.



* **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.

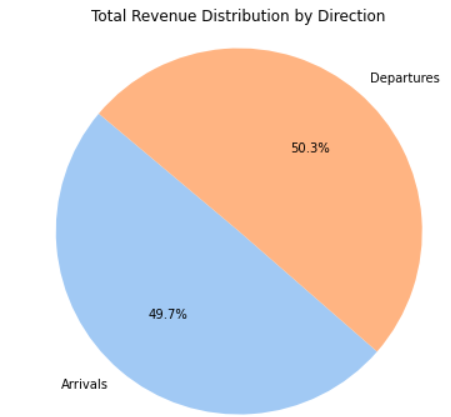
## B. The data Methods applied during data preparation for ML

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.

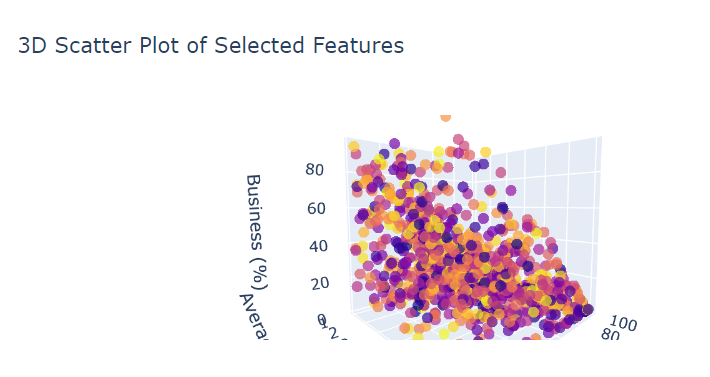


## C. Final Insights

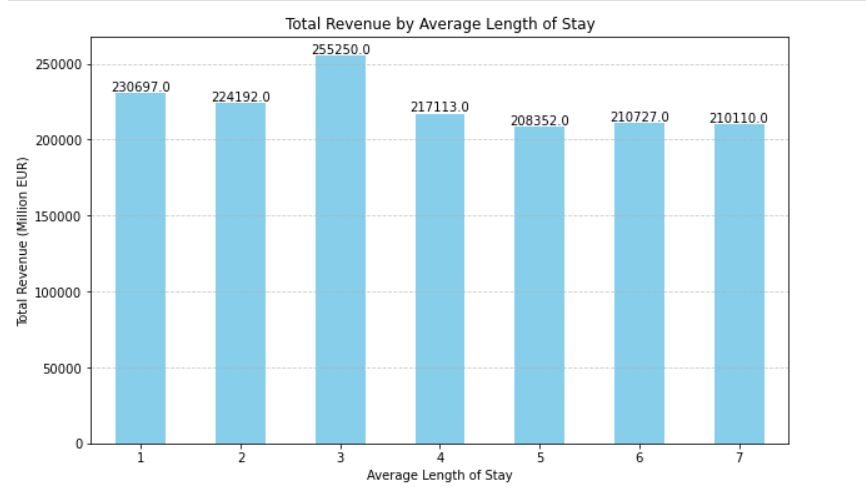


* The pie chart shows the proportion of revenue from arrivals compared to departures.
* The blue slice represents arrivals, and it accounts for 49.7% of the total revenue.
* The orange slice represents departures, and it accounts for 50.3% of the total revenue.

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.



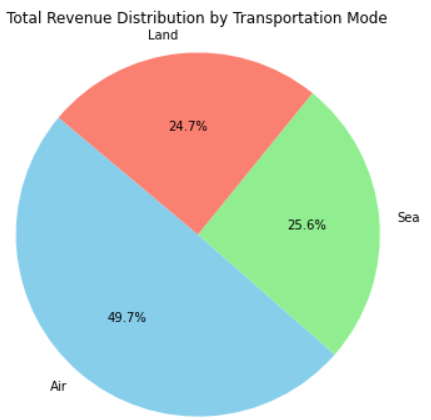
* This 3D Scatter plot shows the effect of the features on the target variable efficiently as we can move the graph. Also, each point on the graph has its own value and by moving cursor on the point shows their value.



* 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.
* 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.