Table of Contents

[Section-I (Statistics) 2](#_Toc162384500)

[A. Descriptive Statistics 2](#_Toc162384501)

[1. Dataset Shape: 2](#_Toc162384502)

[2. Column Names: 2](#_Toc162384503)

[3. Data Types: 2](#_Toc162384504)

[4. Summary Statistics: 3](#_Toc162384505)

[5. Missing Values: 4](#_Toc162384506)

[Critical Analysis: 5](#_Toc162384507)

[B. Visualization with Plots 5](#_Toc162384508)

[1. Histogram: Distribution of Tourism Revenue 5](#_Toc162384509)

[2. Heatmap of Correlation Matrix 6](#_Toc162384510)

[3. Bar Graph 7](#_Toc162384511)

[4. Line Graph 8](#_Toc162384512)

[C. Discrete Distributions 9](#_Toc162384513)

[1. Binomial Distribution: 9](#_Toc162384514)

[2. Poisson Distribution: 10](#_Toc162384515)

[D. Normal Distributions 11](#_Toc162384516)

[E. Importance of the distributions used in point 3 and 4 12](#_Toc162384517)

[1. Binomial and Poisson Distributions (Point 3): 12](#_Toc162384518)

[2. Normal Distribution (Point 4): 12](#_Toc162384519)

# Section-I (Statistics)

## A. Descriptive Statistics

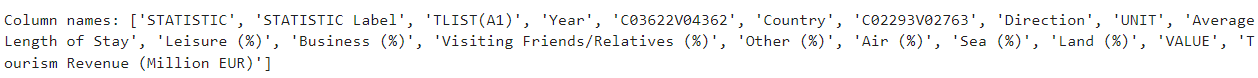
### 1. Dataset Shape:

* The dataset contains 1150 rows and 19 columns.



### 2. Column Names:

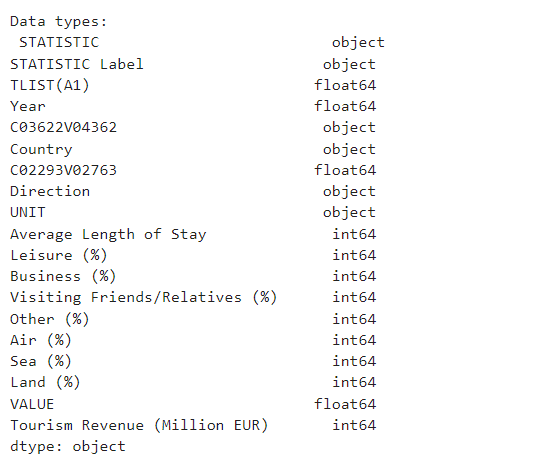
* The dataset has 19 columns with names such as 'STATISTIC', 'STATISTIC Label', 'TLIST(A1)', 'Year', 'Country', 'Direction', 'UNIT', 'Average Length of Stay', 'Leisure (%)', 'Business (%)', 'Visiting Friends/Relatives (%)', 'Other (%)', 'Air (%)', 'Sea (%)', 'Land (%)', 'VALUE', and 'Tourism Revenue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



### 3. Data Types:

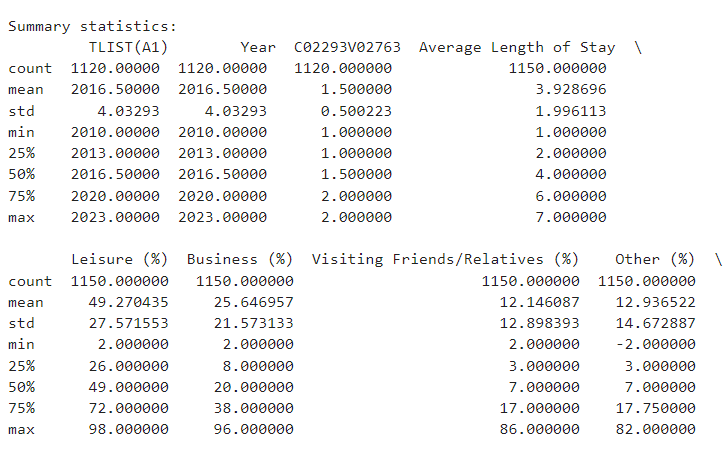
The data types of the columns vary:

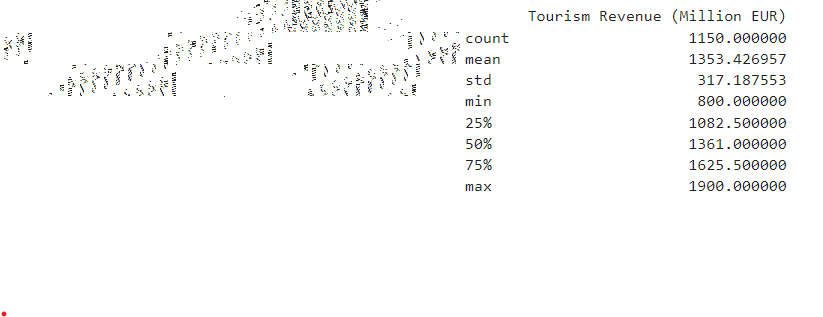
* 'object': STATISTIC, STATISTIC Label, C03622V04362, Country, Direction, UNIT.
* 'float64': TLIST(A1), Year, C02293V02763, VALUE.
* 'int64': Average Length of Stay, Leisure (%), Business (%), Visiting Friends/Relatives (%), Other (%), Air (%), Sea (%), Land (%), Tourism Revenue (Million EUR).



### 4. Summary Statistics:

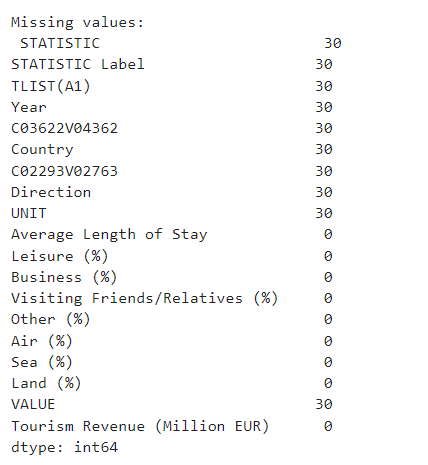
* TLIST(A1): Year column has a mean of approximately 2016.5 with a standard deviation of around 4.03. The data spans from the year 2010 to 2023.
* Average Length of Stay: The average length of stay ranges from 1 to 7 days, with a mean of approximately 3.93 days.
* Leisure (%), Business (%), Visiting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

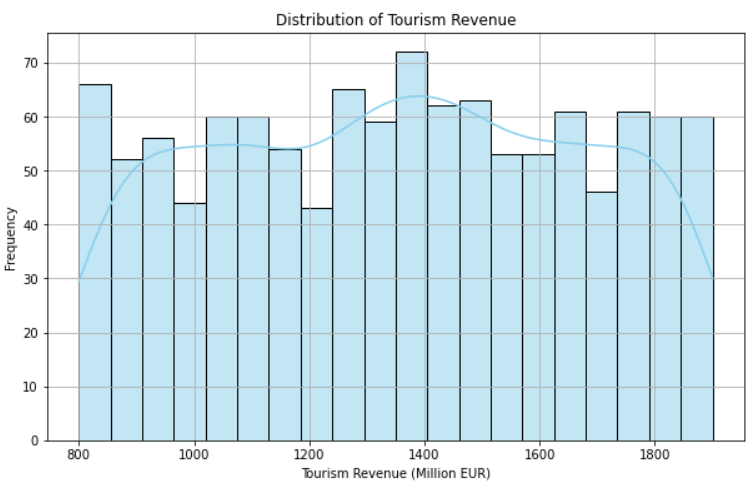
* X-axis: Represents the range of tourism revenue (in million EUR).
* Y-axis: Represents the frequency of occurrence for each revenue range.

**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.



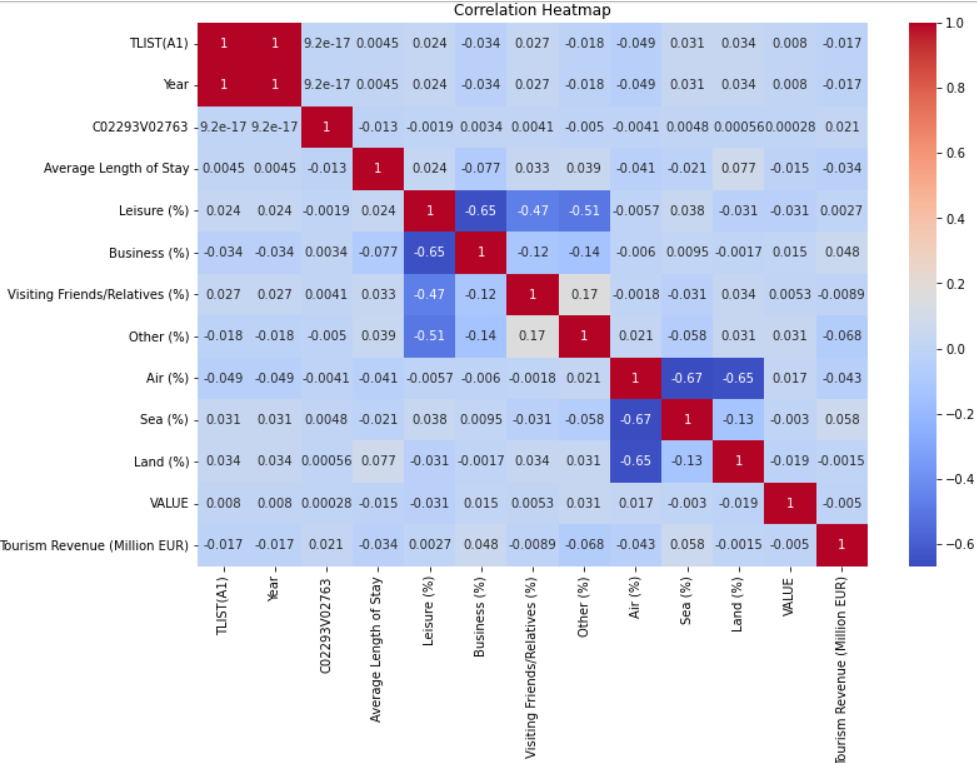
### 2. Heatmap of Correlation Matrix

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.

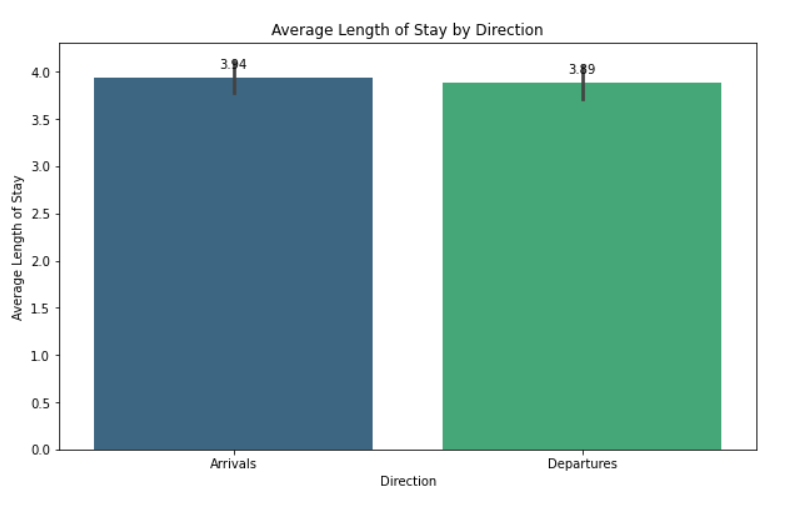


### 3. Bar Graph

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.

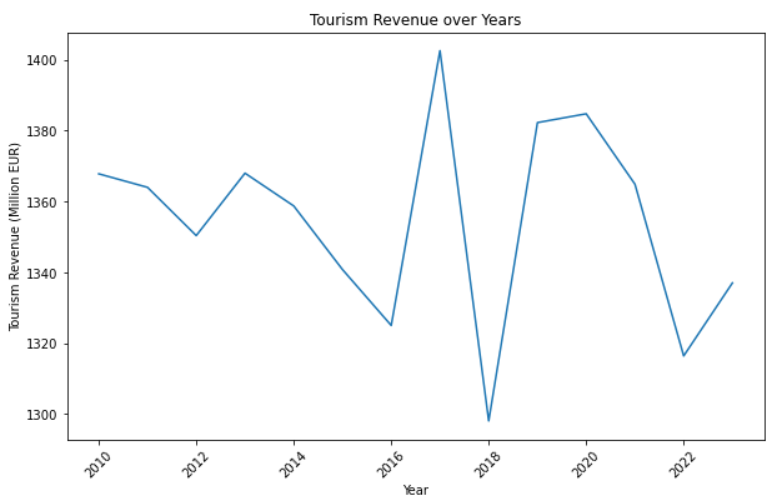


### 4. Line Graph

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:

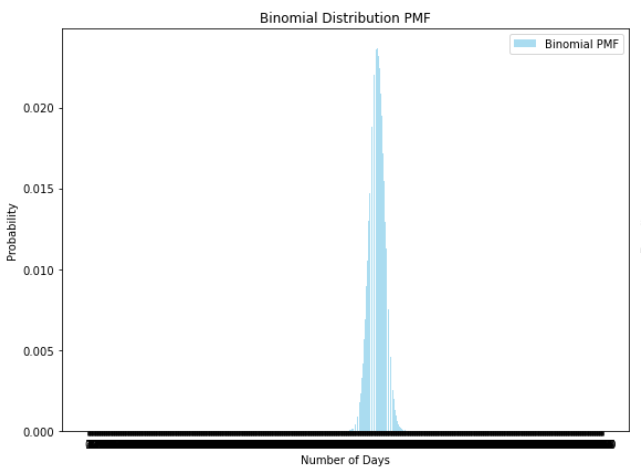
* The y-axis title is “Tourism Revenue (Million EUR)”. This means that the values on the y-axis represent the amount of money that was brought in by tourism in millions of Euros.
* The x-axis title is simply “Year”. This axis shows the years that the data was collected for.
* The line starts at a value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

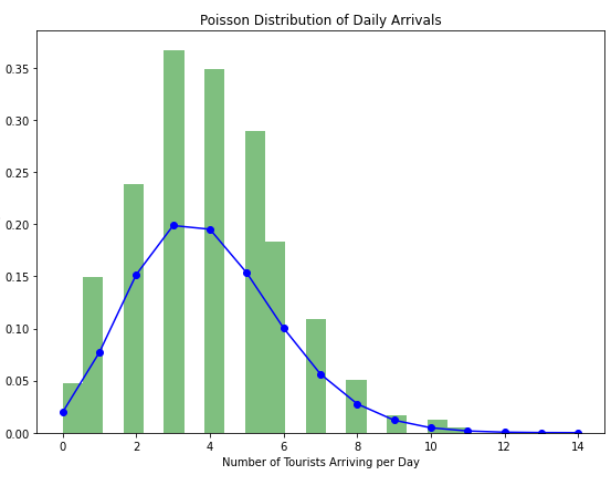


• The x-value of the graph exhibit the number of days between visits, and the y-value display the probability of that days per day between visits occurrence.

• "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.

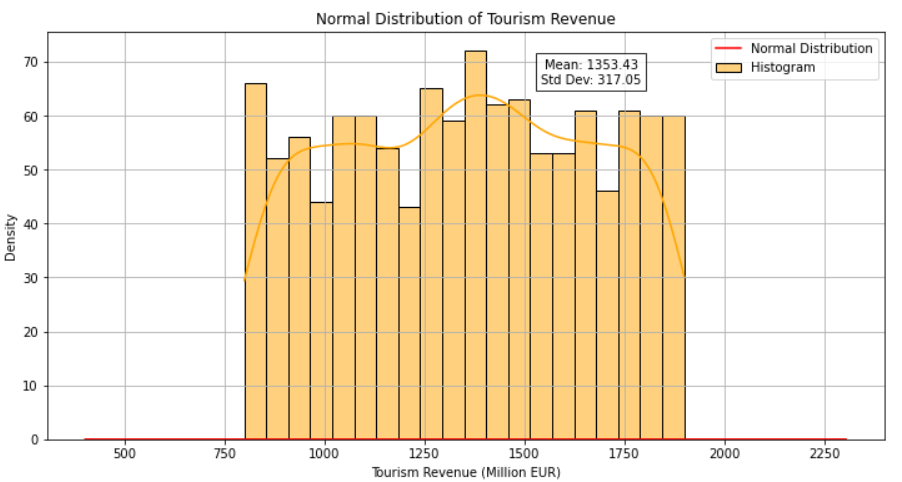
The Poisson distribution would be suitable for the number of daily tourist arrivals fitting as it applies to cases when events happen independently and their impact on the following events does not exist this exponential arrival/travel station assumes a constant rate.

**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.