**EXPLORATION OF TRAVEL INSIGHTS**

THROUGH SYNTHETIC TRANSIT DATA with

Feature Engineering & EDA

KOSURU VISHNU VARDHAN

AM.EN.U4CSE21032

horizontal line

# Introduction

It is critical to understand passenger preferences and optimize their travel experiences in the constantly evolving travel sector. I'm pleased to present a synthetic dataset I created in order to bring insight into transport structures and provide travel companies with data-driven insights.

This dataset offers an uncommon opportunity to investigate topics like how travel patterns change with the seasons. It includes information on journey durations, modes, expenses, and even customer satisfaction. Which routes have a lot of driving? What elements impact the ratings left by customers? Travel departments may learn a lot from examining these trends, which will help them improve pricing tactics, arrange services more effectively, and ultimately make travel easier and more pleasurable for everyone.

## Introduction to Exploratory Data Analysis (EDA)

Now, before we go into specifics, let's take an important visit called Exploratory Data Analysis (EDA). EDA essentially serves as a link between exploring datasets initially and developing predictive models. It is the process of carefully looking over a dataset to find its underlying features, trends, and possible abnormalities. Assuring we're dealing with a dataset we really understand and disclosing hidden features in a dark environment is similar to doing this.

Using NumPy and Python for Data Exploration, a crucial Python module, to help us navigate this EDA adventure. Python is widely recognised for its adaptability and readable nature, providing an abundance of tools for manipulating and analyzing data. However, NumPy is particularly good at working with numerical data.

## Libraries Used

* **NumPy**

NumPy's efficient array structures provide a solid foundation for storing and manipulating numerical data. We'll leverage its functions to load our dataset from its CSV file and seamlessly perform calculations on various features. Its statistical functions empowers us to compute descriptive statistics, including means, standard deviations, quartiles, and more, revealing central tendencies and dispersion within our variables.

* **Pandas**

Pandas introduces DataFrames, flexible structures that organize data into rows and columns, akin to spreadsheets. This allows us to effortlessly explore, clean, and manipulate our dataset. They empower us to inspect the dataset's dimensions, column names, data types, and identify any missing values, providing a comprehensive overview of its organization.

* **Missingno**

Missingno's visual tools paint a clear picture of missing data patterns within our dataset, aiding in identifying potential biases or areas requiring imputation.

* **Random**

The random library ensures reproducibility of our analysis by generating consistent random numbers, crucial for tasks like sampling or shuffling data.

* **Seaborn**

Seaborn builds upon Matplotlib, offering a high-level interface for crafting informative and aesthetically pleasing visualizations, including histograms, box plots, scatter plots, and more, bringing our data to life.

* **Matplotlib**

While Seaborn provides a foundation for visualizations, Matplotlib offers finer control over plot elements, allowing us to customize and refine visualizations to effectively communicate our findings.

**Description of the dataset:**

The dataset is a synthetic collection of transportation-related information generated for analysis and exploration. It consists of 1200 samples, each representing a travel instance. The data includes details such as the day of the week, time of day, origin, destination, chosen travel mode (bus, train, or aeroplane), travel duration, seasonal information (winter, summer, autumn, spring, or rainy), travel volume, cost of travel, delay duration, and customer rating.

The travel modes have associated characteristics, with bus and train costs subject to seasonal variations during the rainy season. Aeroplane travel is assumed to operate exclusively during the rainy season. Travel durations vary based on the chosen mode, and delays are introduced with multiples of 5 minutes. Customer ratings are influenced by the delay duration, with higher ratings for shorter delays.

This synthetic dataset aims to provide a diverse set of scenarios for exploratory data analysis and serves as a foundation for understanding patterns, relationships, and trends within the transportation domain.

**Feature Labeling:**

* **CATEGORICAL FEATURES:**

**DAY (Data Type: Object):** Represents the day of the week when the travel occurs, providing information about the temporal aspect of travel patterns.

**TIMEOFDAY (Data Type: Object):** Indicates the time of day when the travel takes place, offering insights into travel behaviors based on different parts of the day such as morning, evening, or afternoon.

**ORIGIN (Data Type: Object):** Denotes the starting location of the travel, providing information about the departure points and potential regional patterns.

**DESTINATION (Data Type: Object):** Represents the destination of the travel, offering insights into popular travel destinations and possible regional preferences.

**TRAVELMODE (Data Type: Object):** Indicates the mode of transportation used for travel, offering insights into the distribution of travel modes among the dataset.

**SEASON (Data Type: Object):** Represents the season during which the travel occurs, providing information about the seasonal variations in travel patterns.

* **NUMERICAL FEATURES:**

**TRAVELDURATION (Data Type: Int64):** Represents the duration of the travel in minutes, offering insights into the time taken for different travel modes.

**TRAVELVOLUME (Data Type: Int64):** Indicates the volume of travel, possibly representing the number of passengers or the frequency of travel. This can provide insights into travel demand.

**COSTOFTRAVEL (Data Type: Int64):** Represents the cost associated with the travel, providing information about the financial aspect of different travel modes.

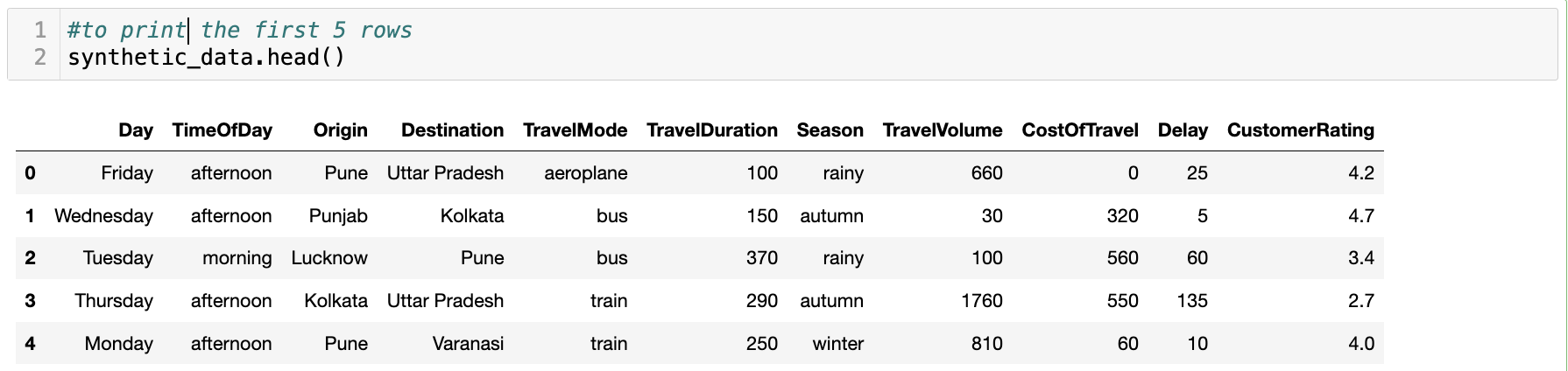
**DELAY (Data Type: Int64):** Denotes the delay in minutes, offering insights into the punctuality and reliability of different travel modes.

**CUSTOMERRATING (TARGET VARIABLE)(Data Type: Float64):** Represents the customer rating given for the travel experience, providing feedback on the satisfaction level of passengers.

**Data Loading**

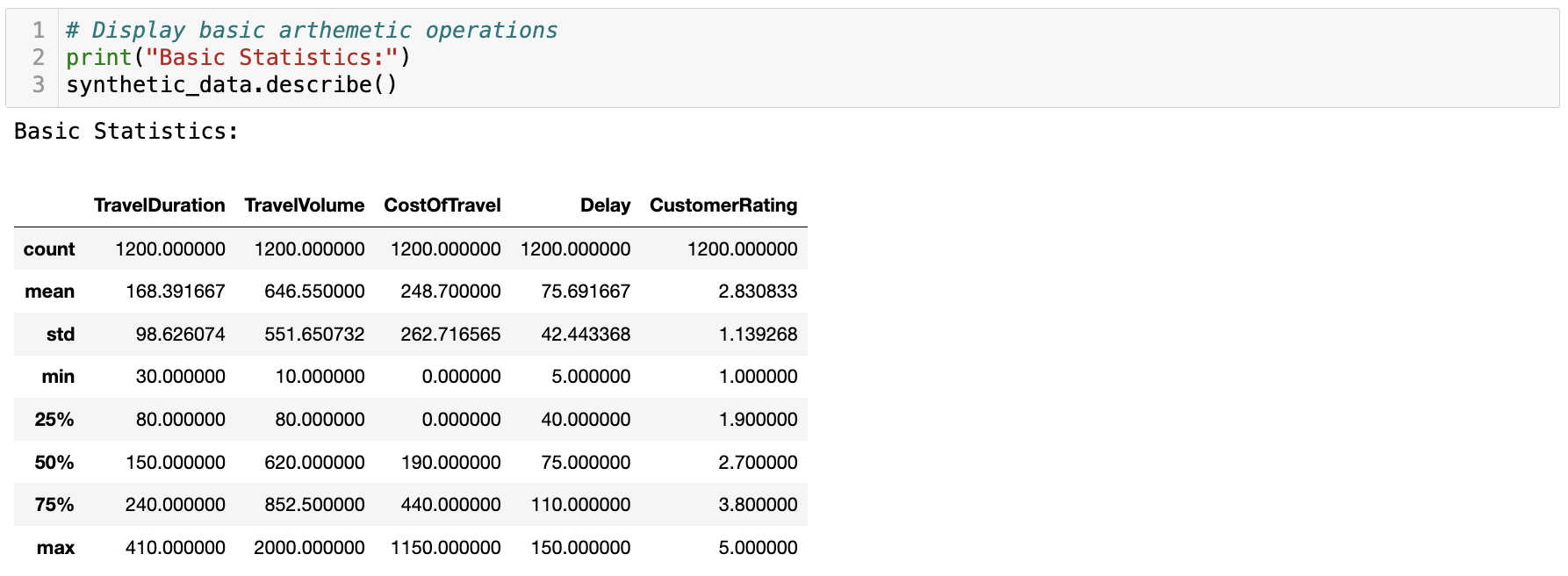
To ensure that my dataset is consistent, I have used random seed in my data creation process using numpy library in Python.

Now we will read the data after loading the dataset into pandas dataframe using **.head()** function.



**Descriptive Statistics of the synthetic dataset**

The .describe() is applied to the dataset inorder to return the summary of descriptive statistics of the dataset including cout, mean, standard deviation, minimum and other quartile information within the synthetic dataset.

****

We can later use these values in further analysis.

**Data Preprocessing**

In order to clean up datasets, deal with problems like missing values and outliers, prepare the data for proper analysis, and glean insightful information, data preparation is essential. Together, let's go through these crucial steps to find the dataset's secrets.

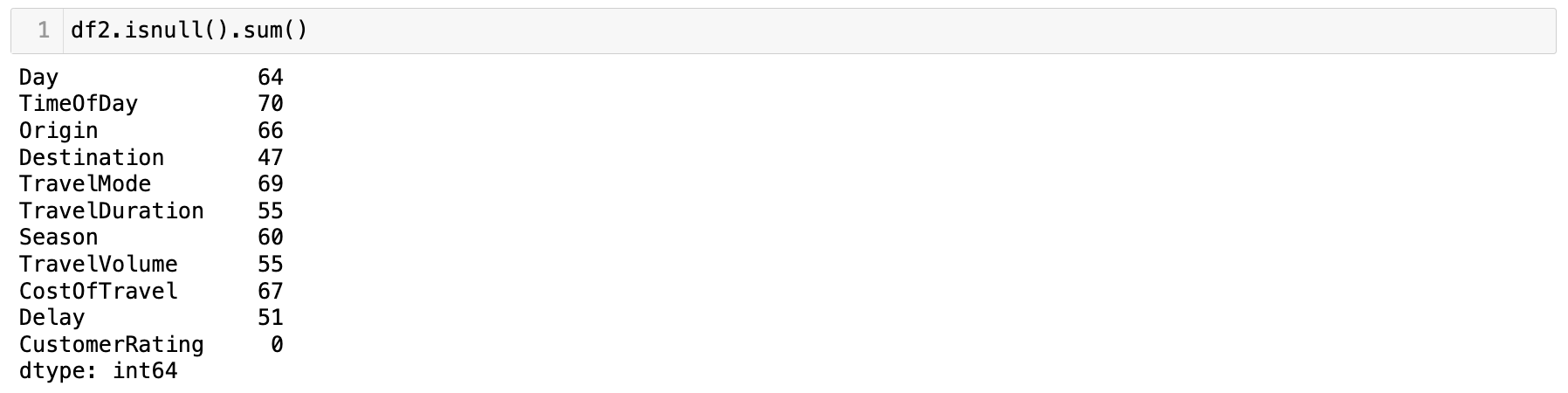
* Creating a copy of a DataFrame using **df2 = synthetic\_data.copy()** is a common practice to avoid unintended modifications to the original DataFrame.



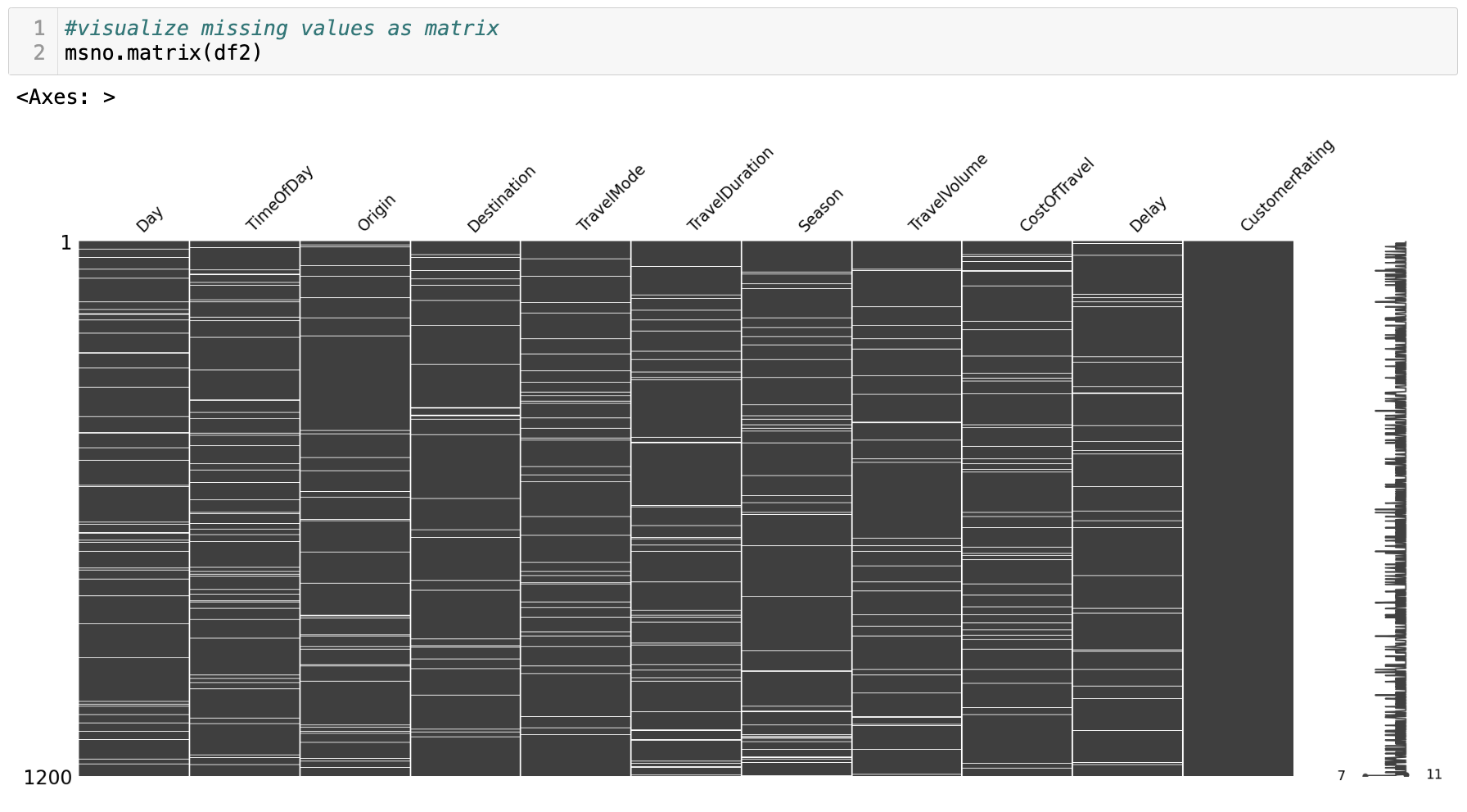
1. **Missing Values**

* **Check for missing values**

To make sure that our dataset has no missing values in them, we need to find the number of available null values in our existing synthetic dataset. **.isnull().sum()** returns the sum of the number of null values found in each feature.

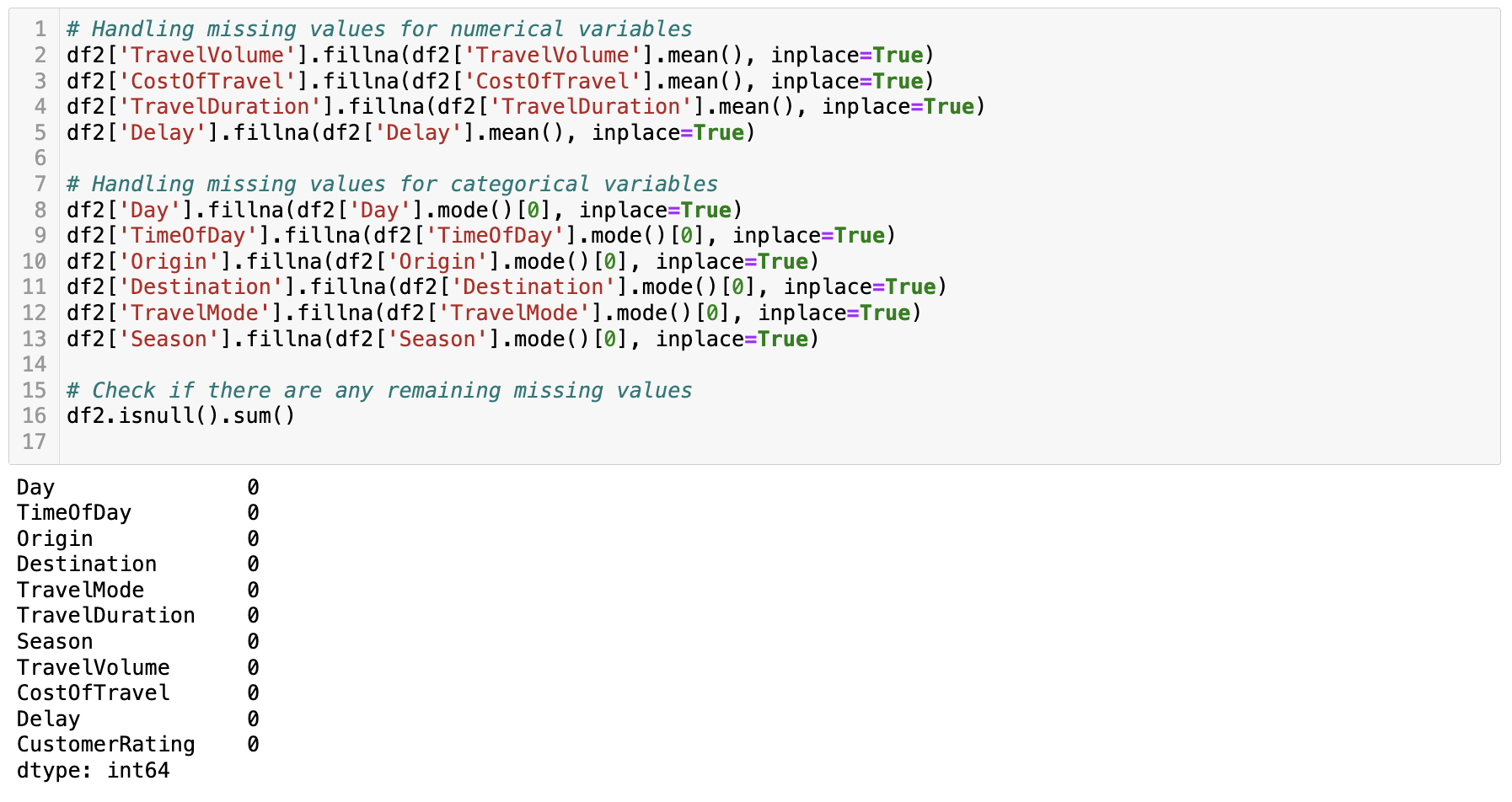


To Properly visualize this we use **missingno** matrix so that we can notice the missing values pattern in the dataset.

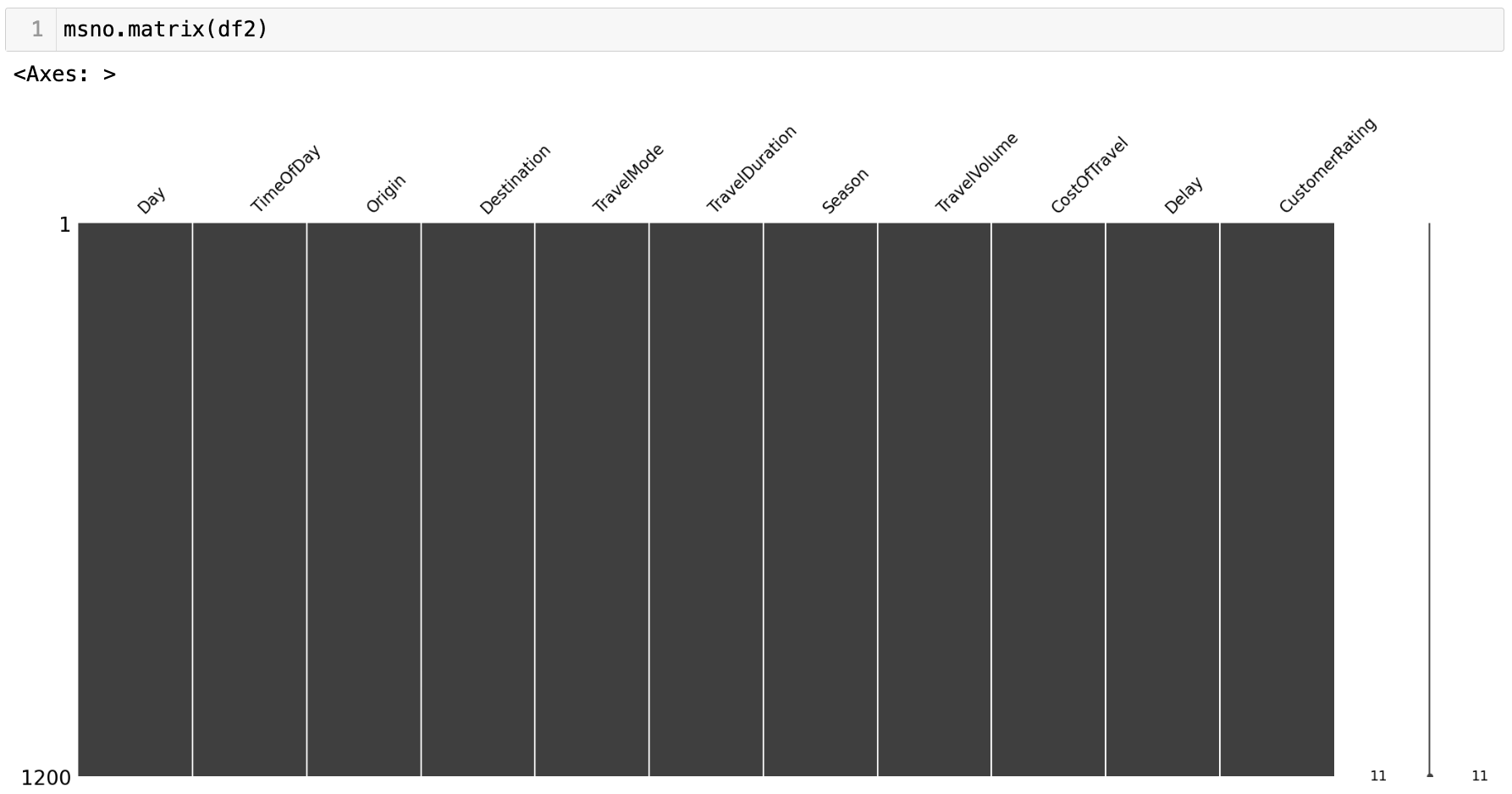


* **Addressing missing values**

To address missing values, we opt for **mean imputation** for numerical data types, ensuring a balanced distribution. For categorical data types, we employ **mode imputation** to uphold consistency in our dataset.



After imputing, the **missingno** matrix shows no empty cell which means that the missing values are filled correctly.

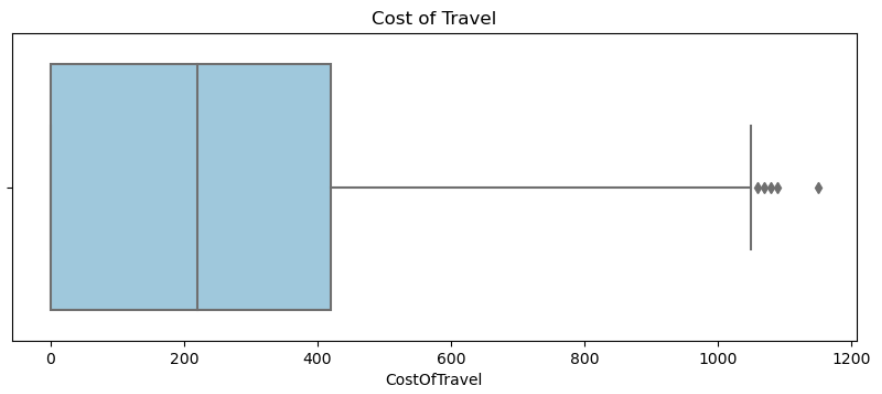


1. **Outliers Detection**

* **Boxplot implementation**

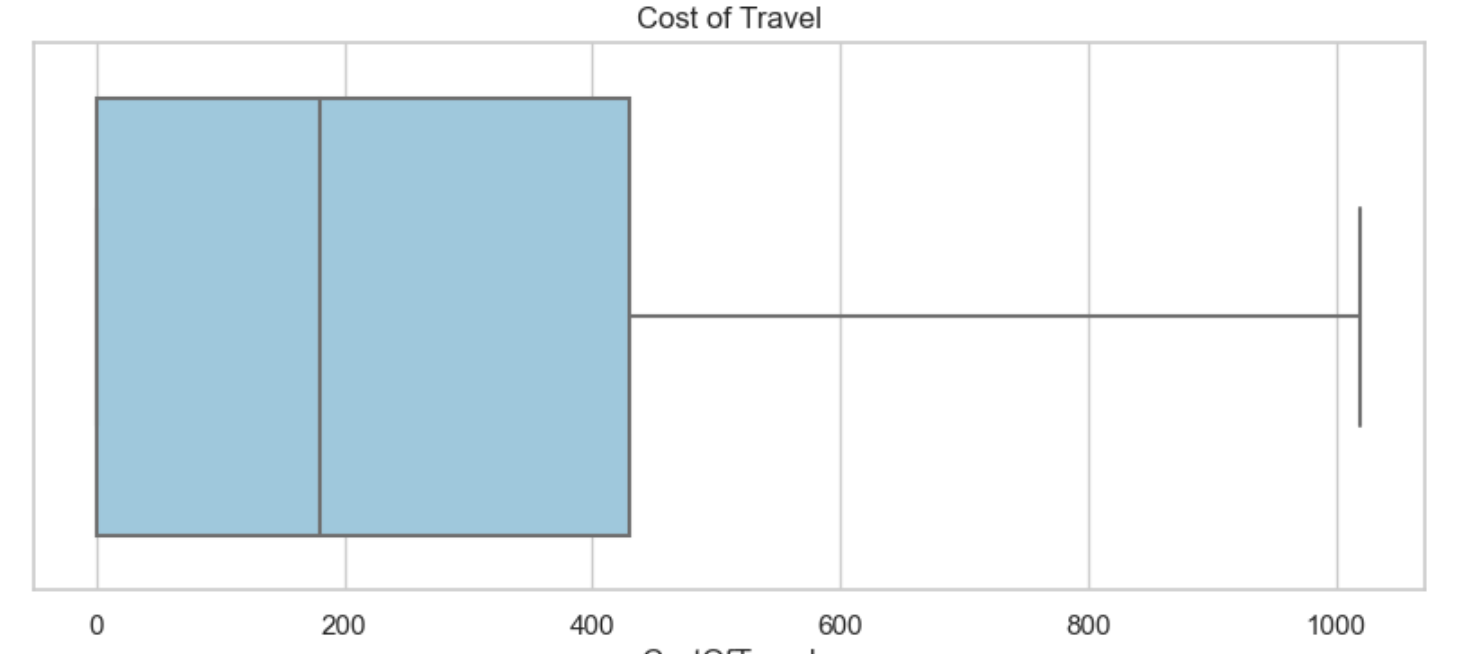
Now lets detect some outliers implementing boxplot for all numerical features.

I have encountered outliers on category **Cost of Travel** after implementing boxplots.



* **Removing outliers**

I explicitly used the **Interquartile Range (IQR)** approach for the **'Cost of Travel'** feature to address outliers in our investigation. This technique focuses on the statistical dispersion within the sample to efficiently identify and remove outliers. We were able to eliminate extreme values from the "Cost of Travel" by computing the IQR and creating a range based on quartiles. This strategy ensures significant insights from the 'CostOfTravel' feature and improves the accuracy of our analysis by striking a balance between the dataset's overall distribution and sensitivity to outliers.



1. **Removing duplicates**

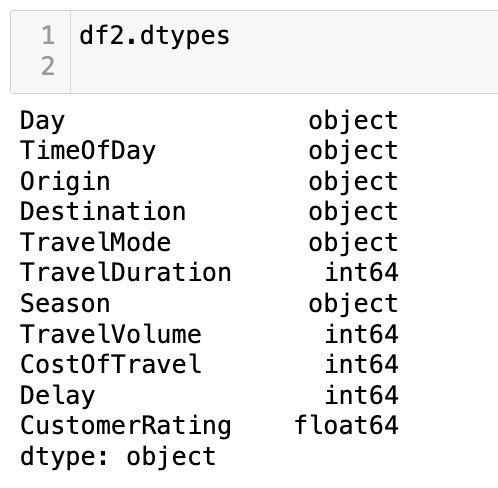
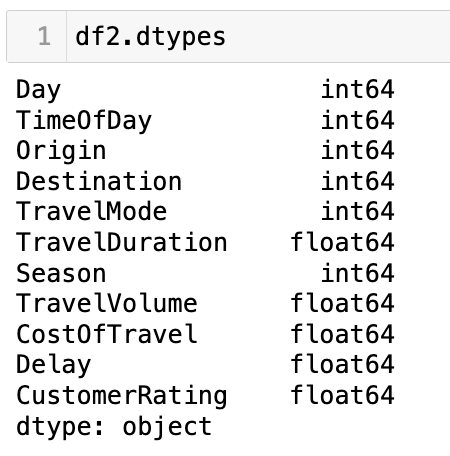
* Handle Duplicate entries in the dataset using **.drop\_duplicates()** which eliminates redundant information.



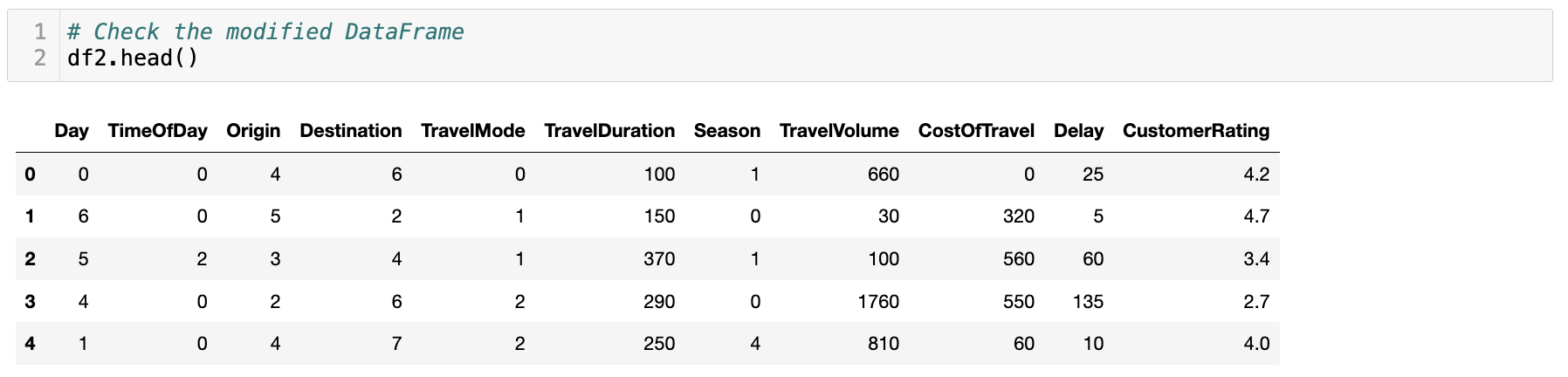
1. **Encoding**

Using encoding techniques to convert categorical variables into a format that is appropriate for machine learning models improves the understanding of data. This optimisation guarantees a more reliable method for data science decision-making that is well-informed.

**Before LabelEncoding** **After LabelEncoding**

Notice the change of datatypes for features like Day, TimeOfDay, Origin & Destination for better effectiveness. **.head()** is used to recheck the dataframe.



**Univariate & Bivariate Analysis**

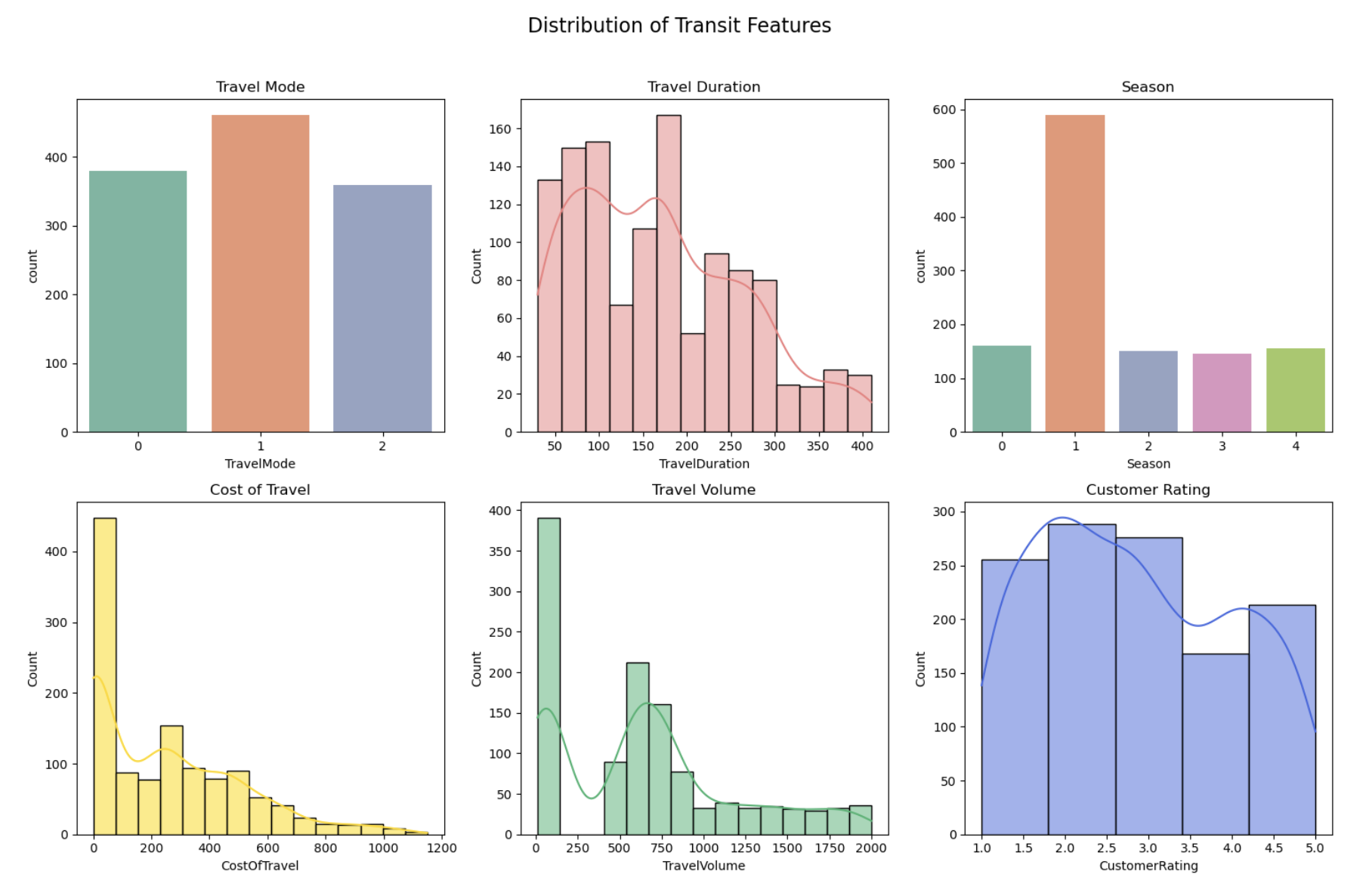
A dataset's individual variables are examined using **univariate analysis**, which provides information on their distributions and properties & In order to find patterns, correlations, or dependencies within the dataset, **bivariate analysis** examines relationships between pairs of variables.

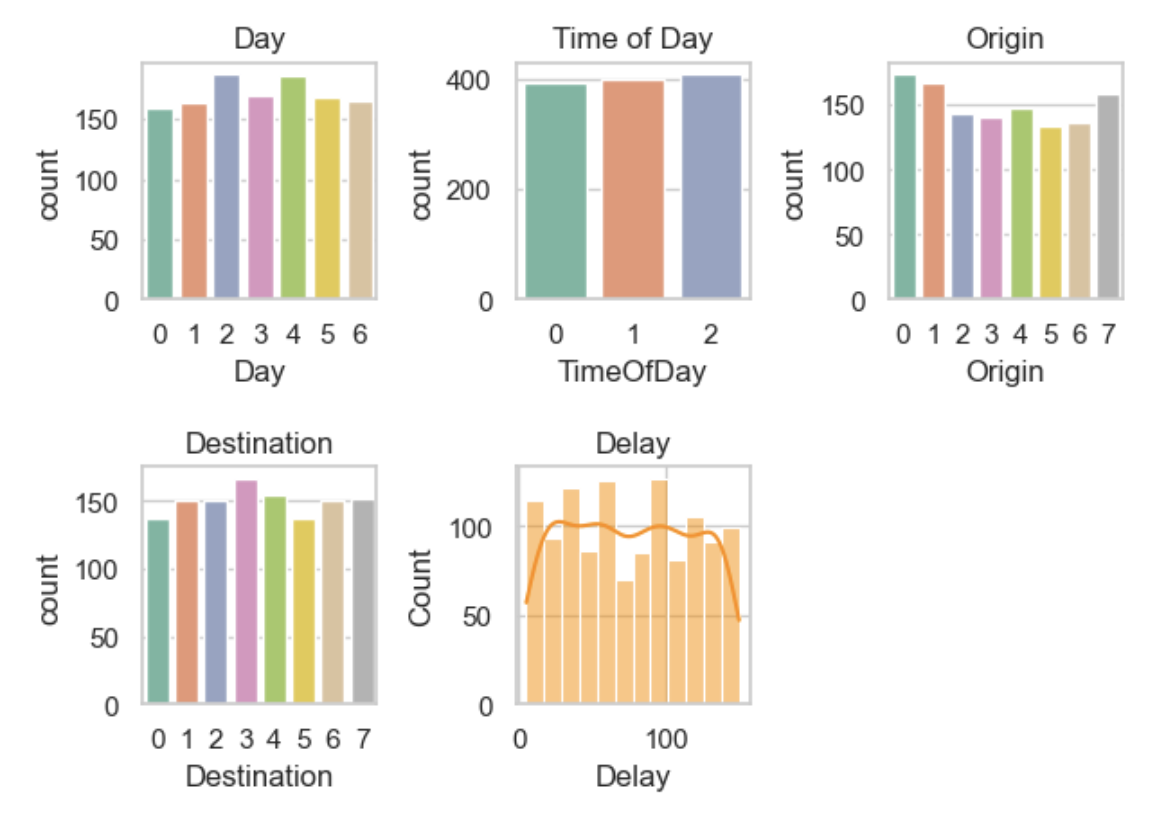
Let's look at a few visualizations that can help us know our data better.

* **Hist Plot & Count Plot**

The **.histplot()** function in Seaborn is used to create histograms, which show the frequency or count of observations within each bin and give a visual depiction of the distribution of a numerical variable.

However, Seaborn's **.countplot()** is specifically made for categorical data. It provides a brief summary of the distribution of categorical values by displaying a bar plot that counts the instances of each distinct category in a categorical variable.

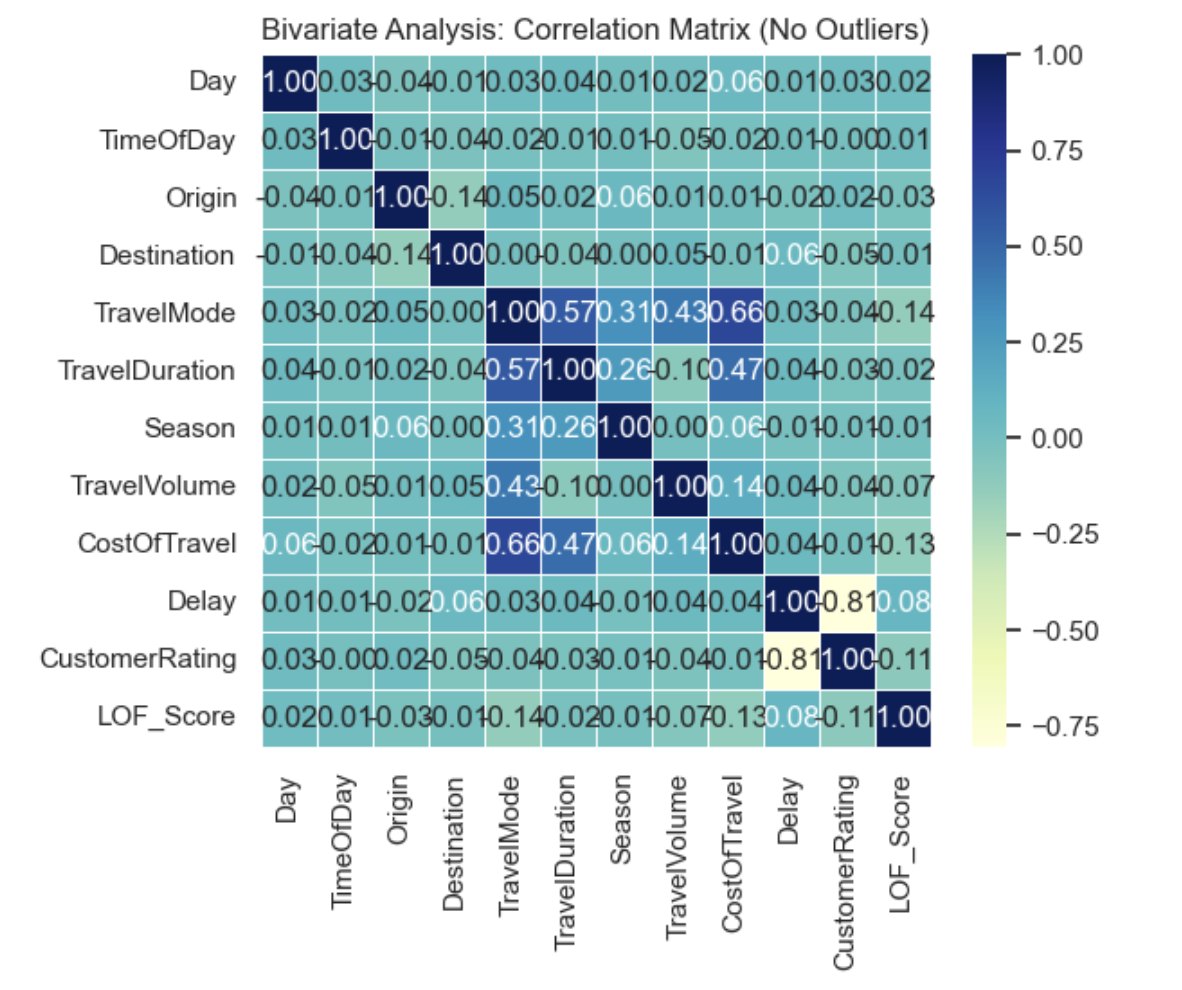




Insights into the general patterns and features of the dataset can be obtained by observing these distributions, which facilitates further analysis and decision-making.

* **Heatmaps**

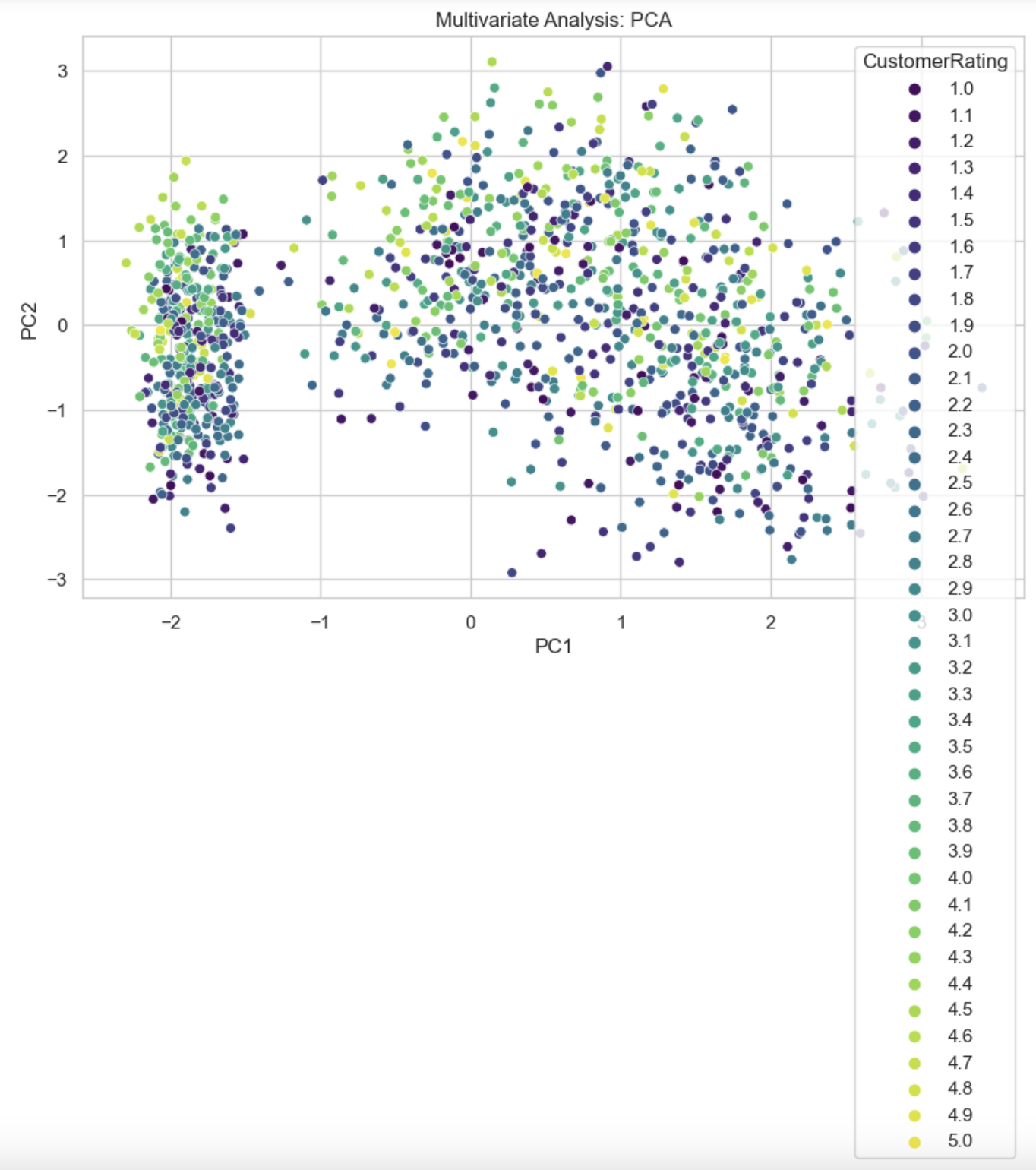
A thorough picture of the correlation map between the numerical features in the transit dataset is shown by the heatmap. The heatmap's color intensity illustrates the direction and strength of relationships between various features. Positive correlations are represented by lighter hues, which suggest that as one attribute rises, the other usually follows suit. Darker hues, on the other hand, indicate negative correlations, which imply that one characteristic tends to diminish as the other grows. Our comprehension of the interactions between transit-related variables is improved by this visual depiction, which also provides insights into possible dependencies and patterns within the dataset.



Finding possible multicollinearity or relationships between variables is made easier by analysing the heatmap. This knowledge directs the selection of features and improves our understanding of the interplay between features, both of which are essential for developing strong predictive models.

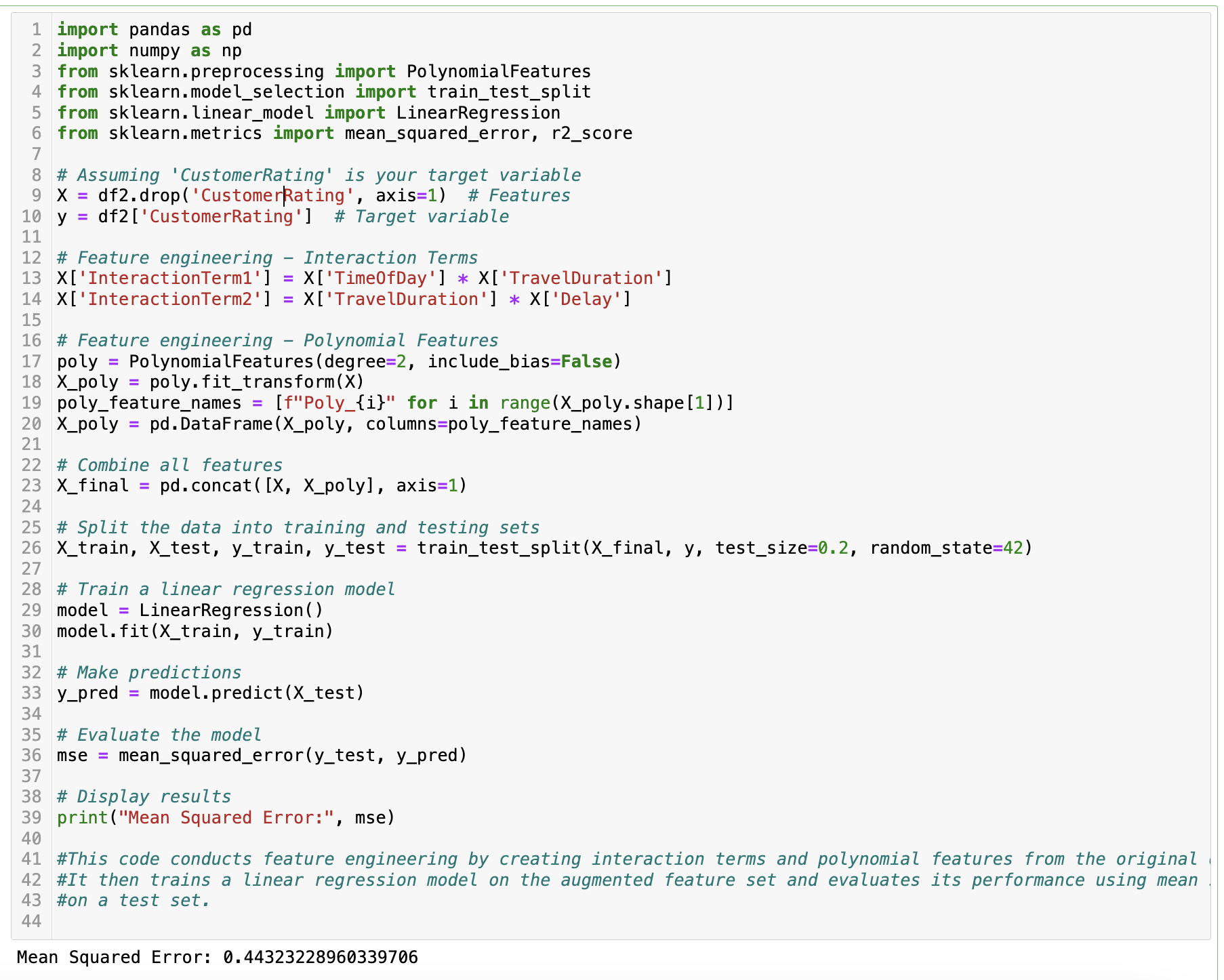
* **PCA**

Principal Component Analysis is used in the multivariate analysis to minimize the transit dataset's dimensionality while maintaining important information.



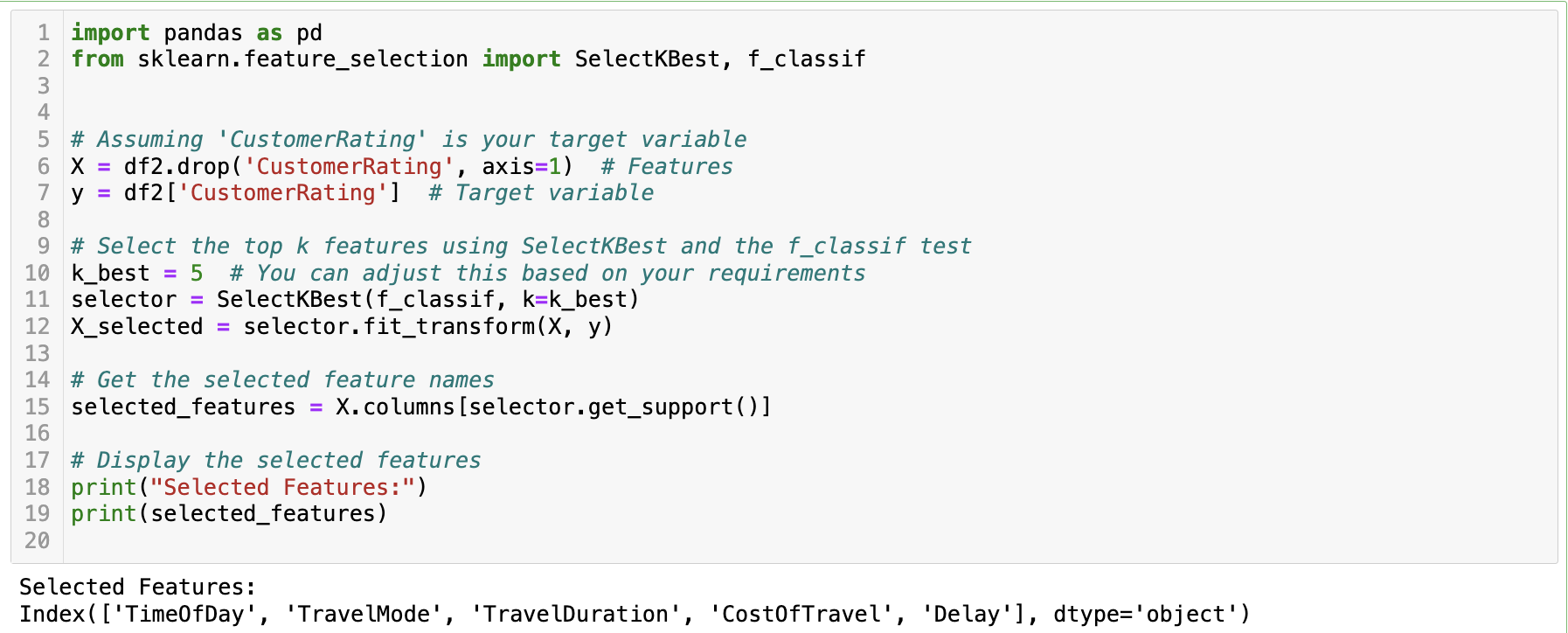
This graphic allows for the identification of travel patterns, which contributes to a deeper comprehension of the underlying relationships in the data. Simplifying complicated datasets for additional investigation and analysis is made easier with the usage of PCA.

* **Initial Feature Engineering**

****

I added polynomial features and interaction terms into the dataset during the first round of feature engineering in an effort to identify complex patterns and correlations between the variables. With the help of this extensive feature set, the model should be able to comprehend the data more deeply and make better predictions for the 'CustomerRating.' Then, in order to improve model performance, a feature selection procedure was implemented, which used the f\_classif test and the SelectKBest method to determine the top 5 statistically significant features. The foundation for a more simplified and effective model was created by the feature set that was further refined following the first engineering phase.

* **Feature Selection**

****

The top 5 statistically significant features were found by applying feature selection, more especially the f\_classif test and the SelectKBest method. By streamlining the feature space and improving interpretability of the model, this technique helped to significantly reduce Mean Squared Error (MSE) by 0.03 and so proved to be crucial in optimizing prediction accuracy for 'CustomerRating.'

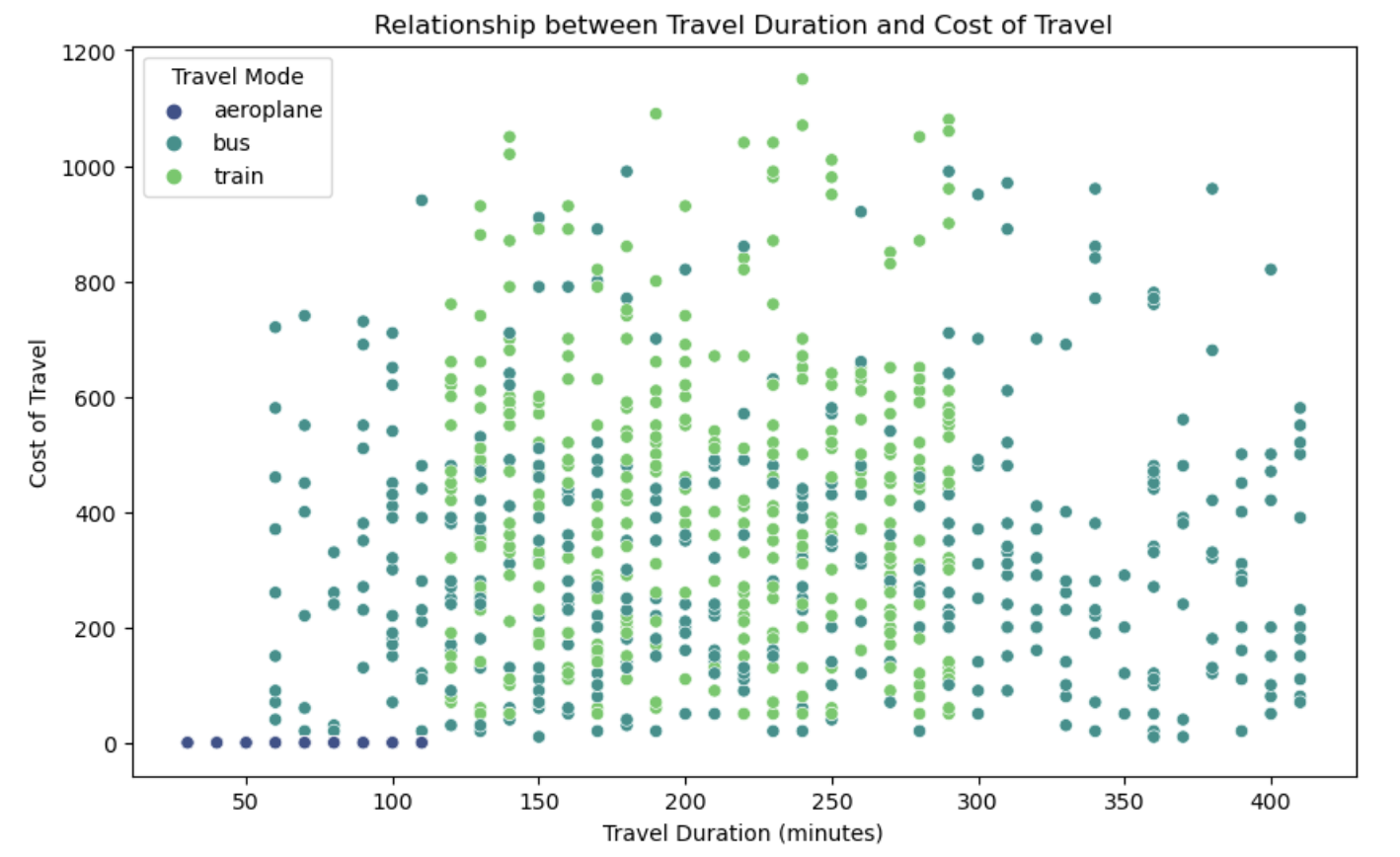
* **phase -2 Feature Engineering**

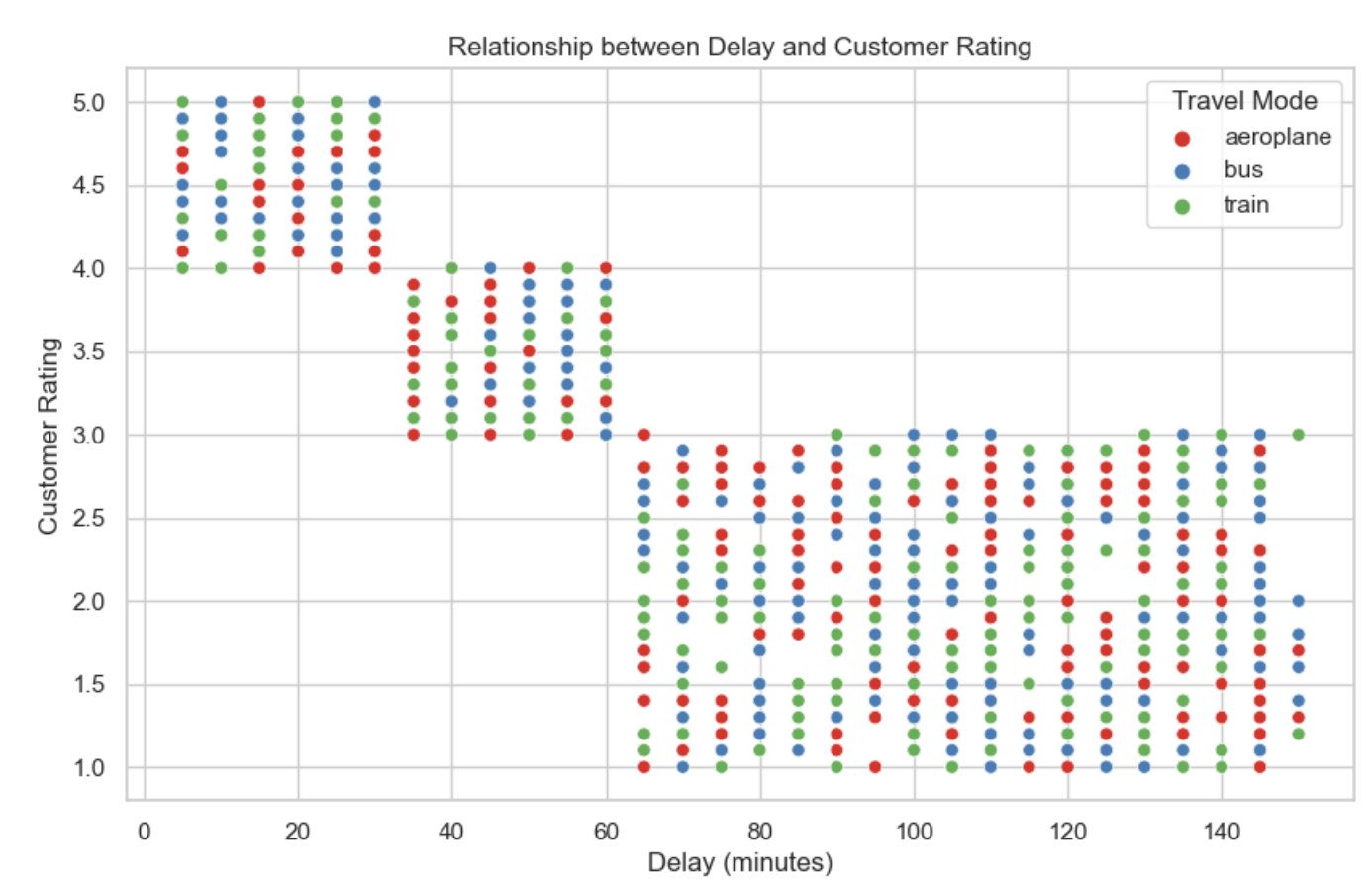


Following feature selection in the second phase, there was a noticeable improvement in the model's performance. In particular, there was a **0.03 reduction** in the Mean Squared Error (MSE) statistic when compared to the initial feature engineering results. This decrease in Mean Squared Error (MSE) highlights the significant influence of feature selection in creating a more targeted and efficient feature set. A focused and deliberate approach to feature engineering is crucial in the data science workflow, as demonstrated by the model's increased accuracy in predicting 'CustomerRating,' which was attained by finding and keeping the most influential characteristics.

* **Scatter Plot**

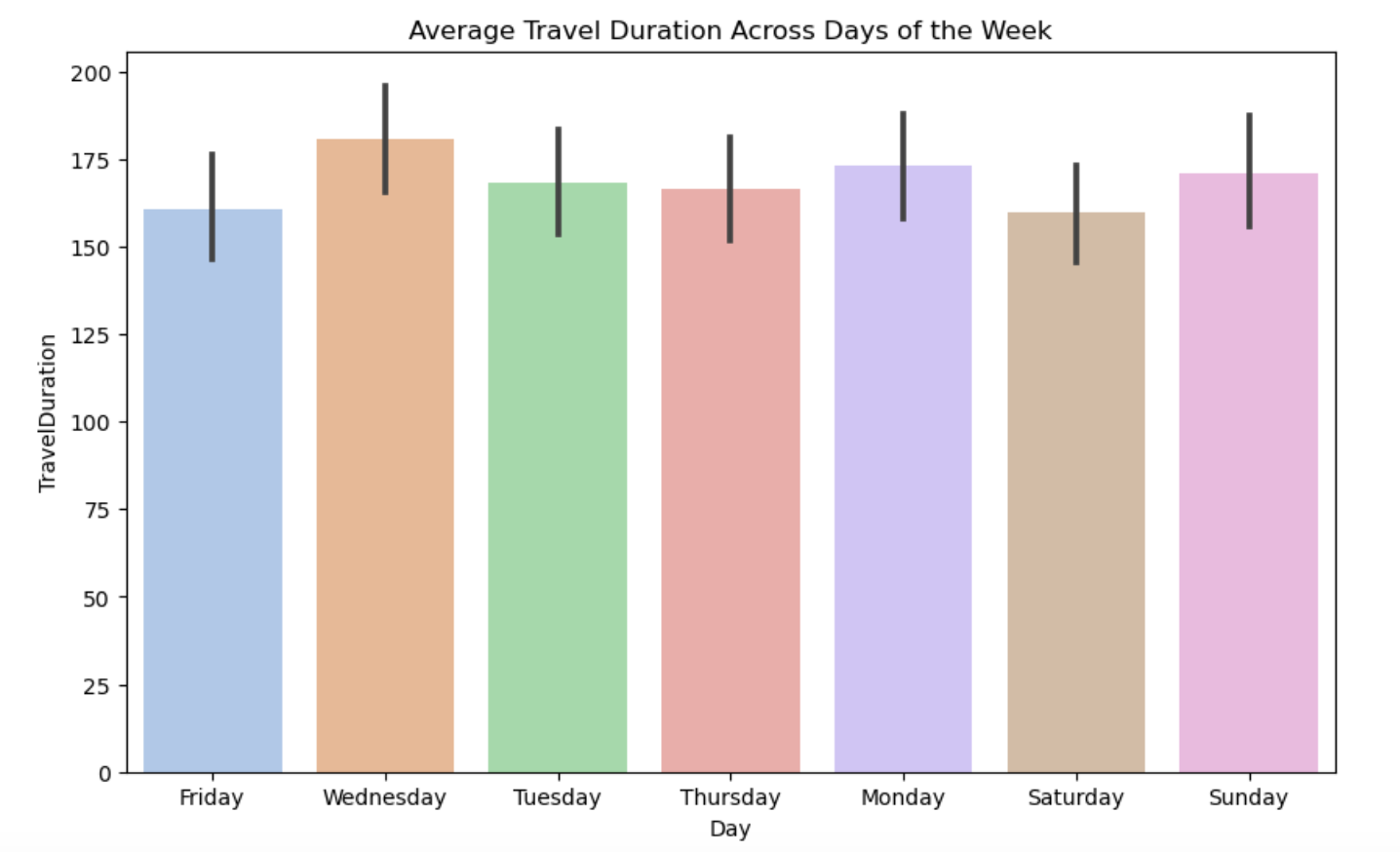
The correlations found in the artificial transit dataset are visually shown by the scatter plots. First, we examine the relationship between "Travel Duration" and "Cost of Travel," with each point denoting a distinct travel occurrence. The portrayal is improved by the use of varied hues to signify different types of transportation.



The second plot explores the relationship between "Delay" and "Customer Rating," illuminating how the length of the delay affects passengers' degree of satisfaction. Each plot's legend clarifies the effects of various types of transportation, advancing our knowledge of these connections throughout the transit domain.  


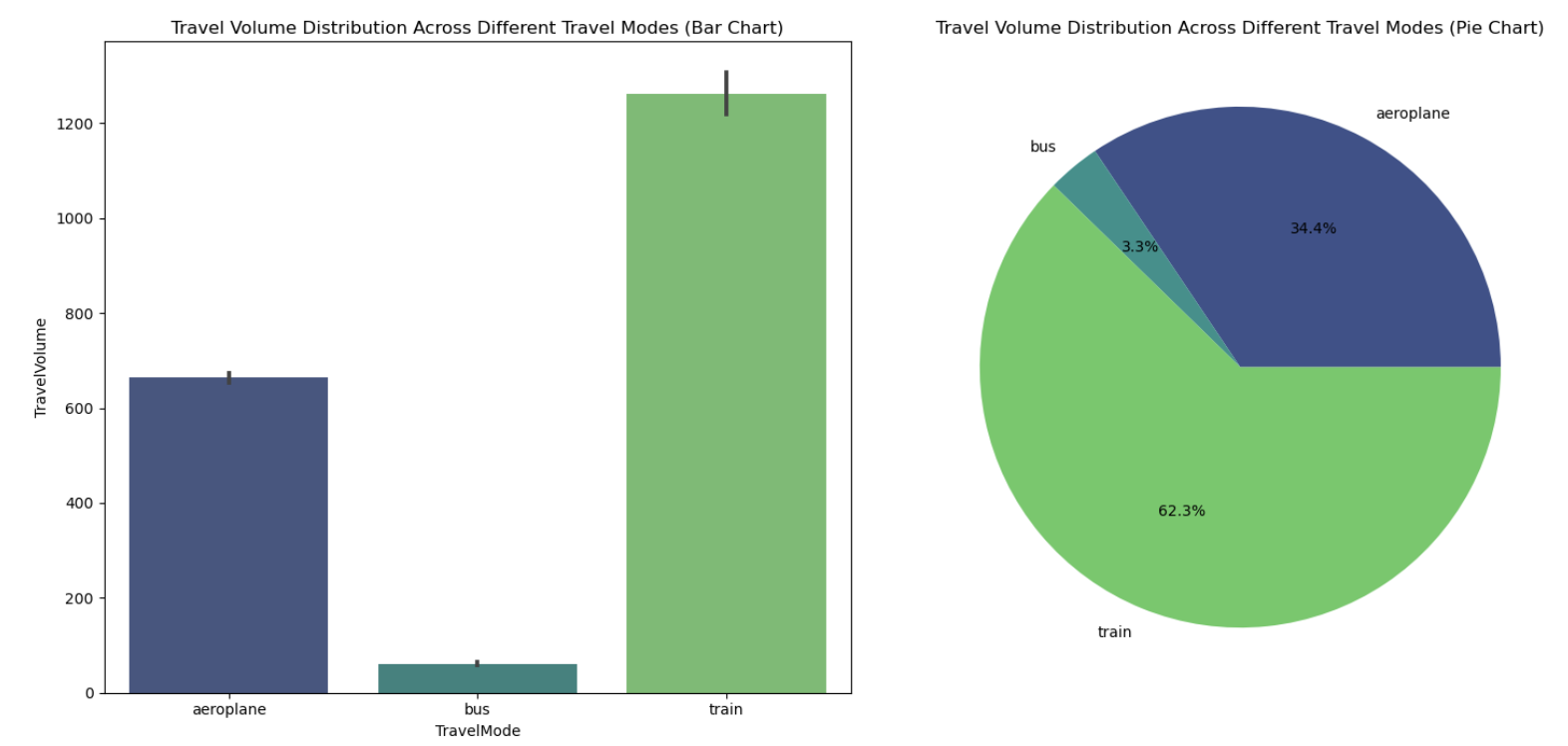
**Questions that are solved using visualizations**

1. **What is the average travel duration for different days of the week?**



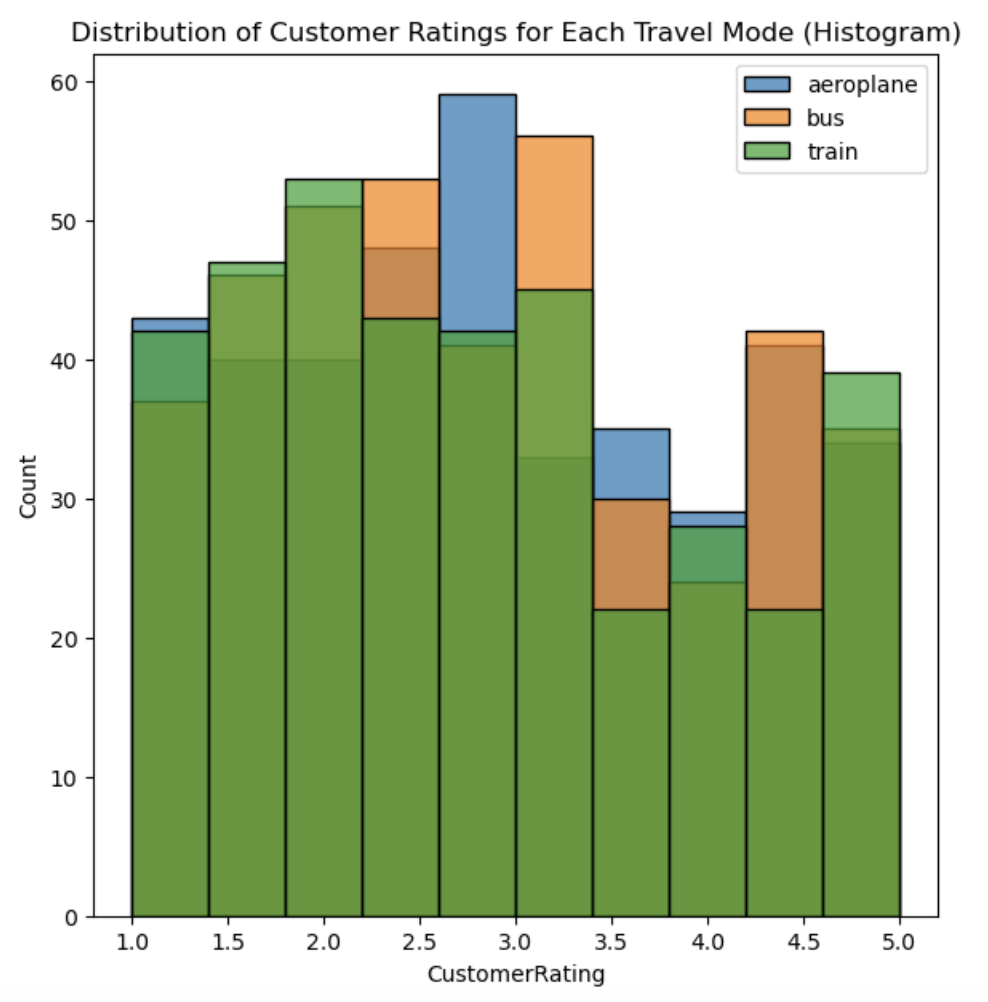
Within the transit dataset, the visualizations provide insight on the temporal dynamics of travel patterns. The bar graph provides a brief summary of temporal fluctuations by displaying the average journey time on several days of the week. In the meantime, the dynamic trends in passenger movement are illustrated by the line chart, which explores the variation in travel volume over the course of the week. These visuals help with strategic decision-making for transport services by providing a more thorough understanding of how trip volume and duration vary across different days.

1. **How is travel volume distributed across different travel modes (bus, train, aeroplane)?**

****

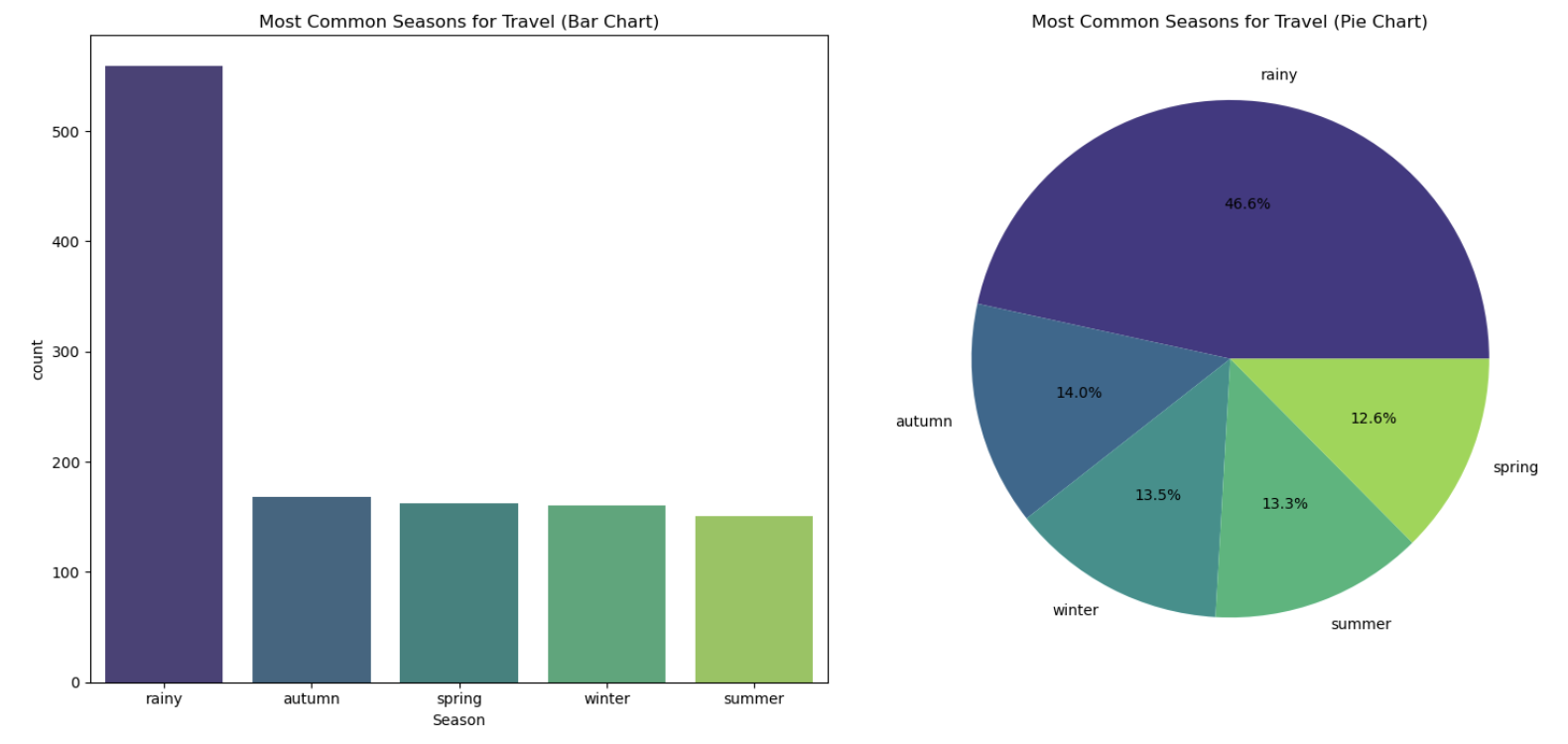
The pairwise displays offer a thorough understanding of the distribution of travel volume among various modes of transport in the transit dataset. The bar chart provides a clear comparison by efficiently displaying the relative travel volumes for the bus, train, and airline modes. In addition, the pie chart shows the distribution percentages for each mode of transportation and offers a proportionate breakdown of the total volume of travel. The popularity and usage patterns of various forms of transportation are better understood by these visualizations.

1. **What is the distribution of customer ratings for each travel mode?**



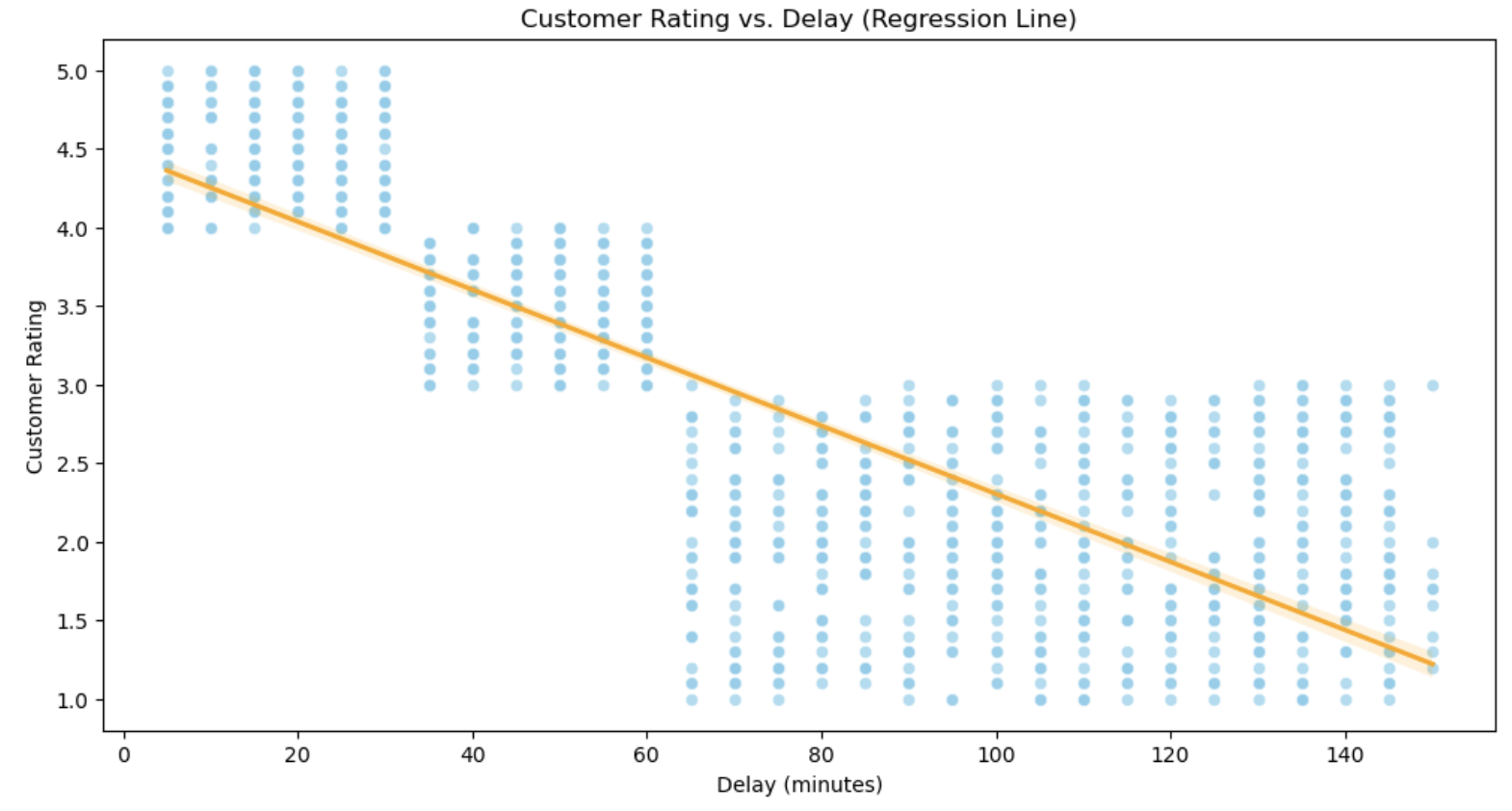
Within the transit dataset, the histogram visualization offers a perceptive representation of the distribution of customer ratings among various modes of transportation. Different color-coded bars correspond to each means of transportation, such as bus, train, and aeroplane. This graphic makes it possible to compare the frequency of different customer evaluations for every mode of transport in an easy-to-understand manner, providing a thorough grasp of the passenger satisfaction levels.

1. **What are the most common seasons for travel?**

****

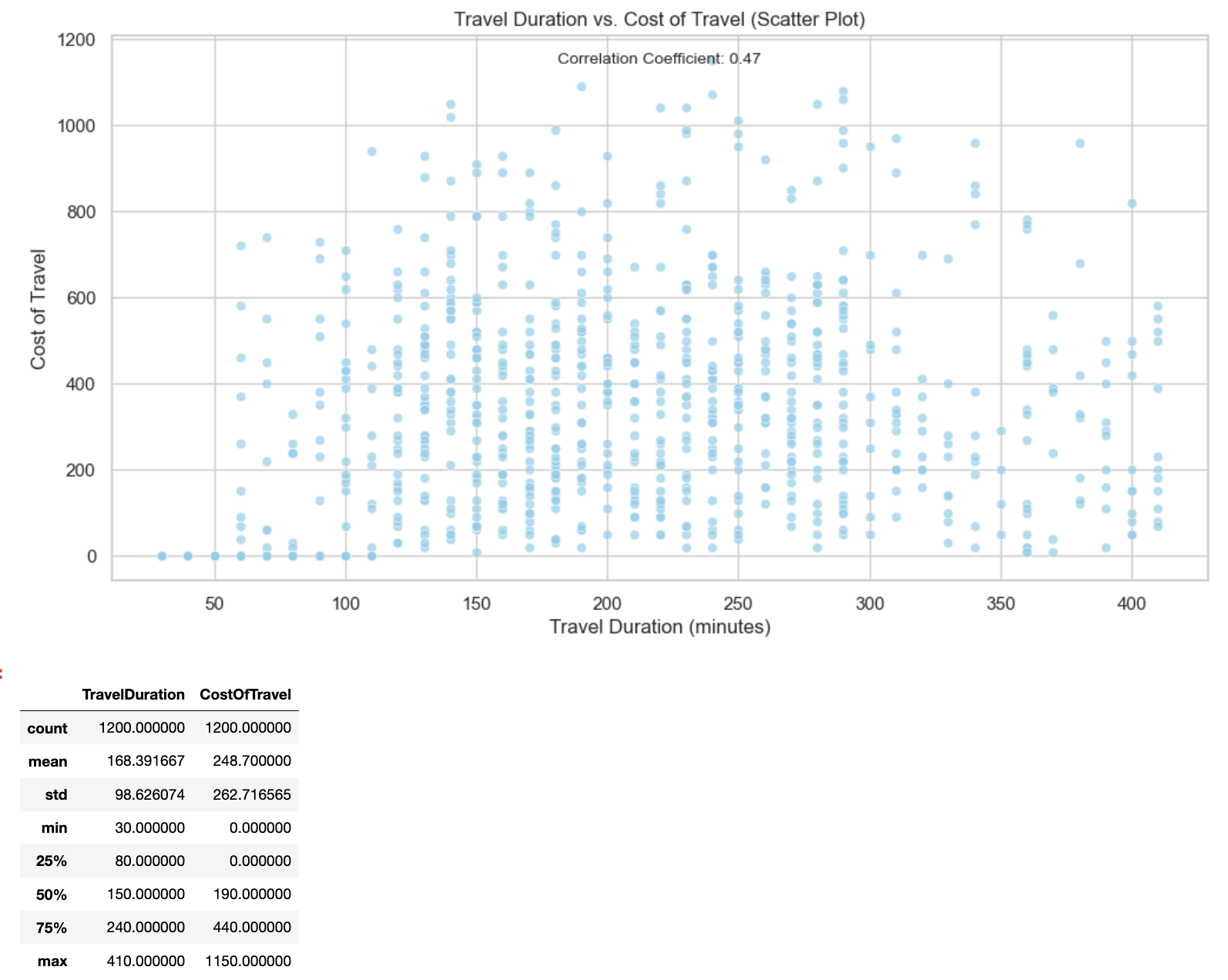
The paired visualizations provide information about the most popular travel seasons in the transport dataset. They consist of a pie chart and a bar chart. With the seasons arranged according to popularity, the bar chart offers a clear picture of how frequently travel occurs throughout the year. In addition, the pie chart provides a percentage-based perspective on the frequency of each season by visually dissecting the overall distribution. The seasonal preferences and trends present in the dataset become more clear .

1. **How does customer rating change with increasing delay?**

****

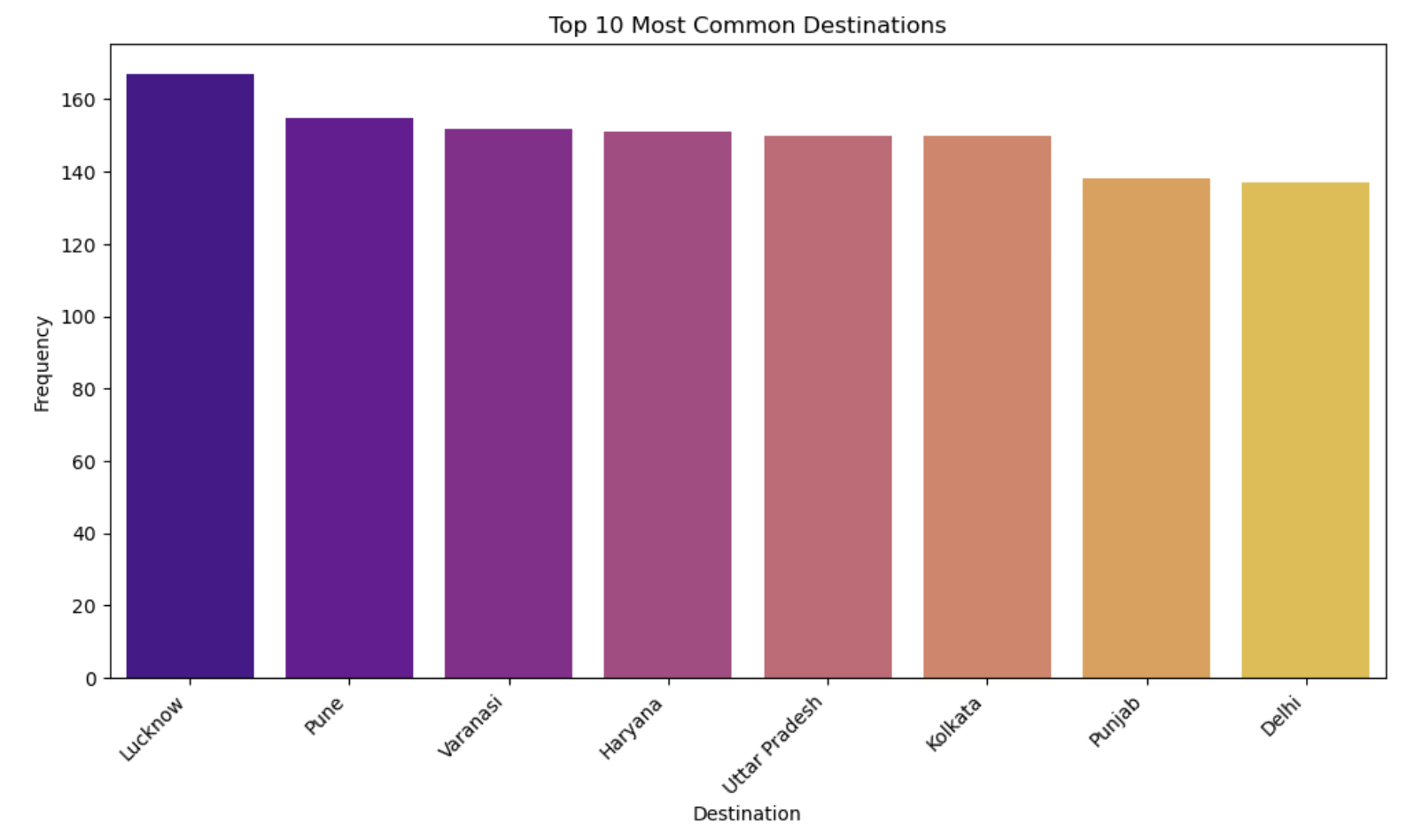
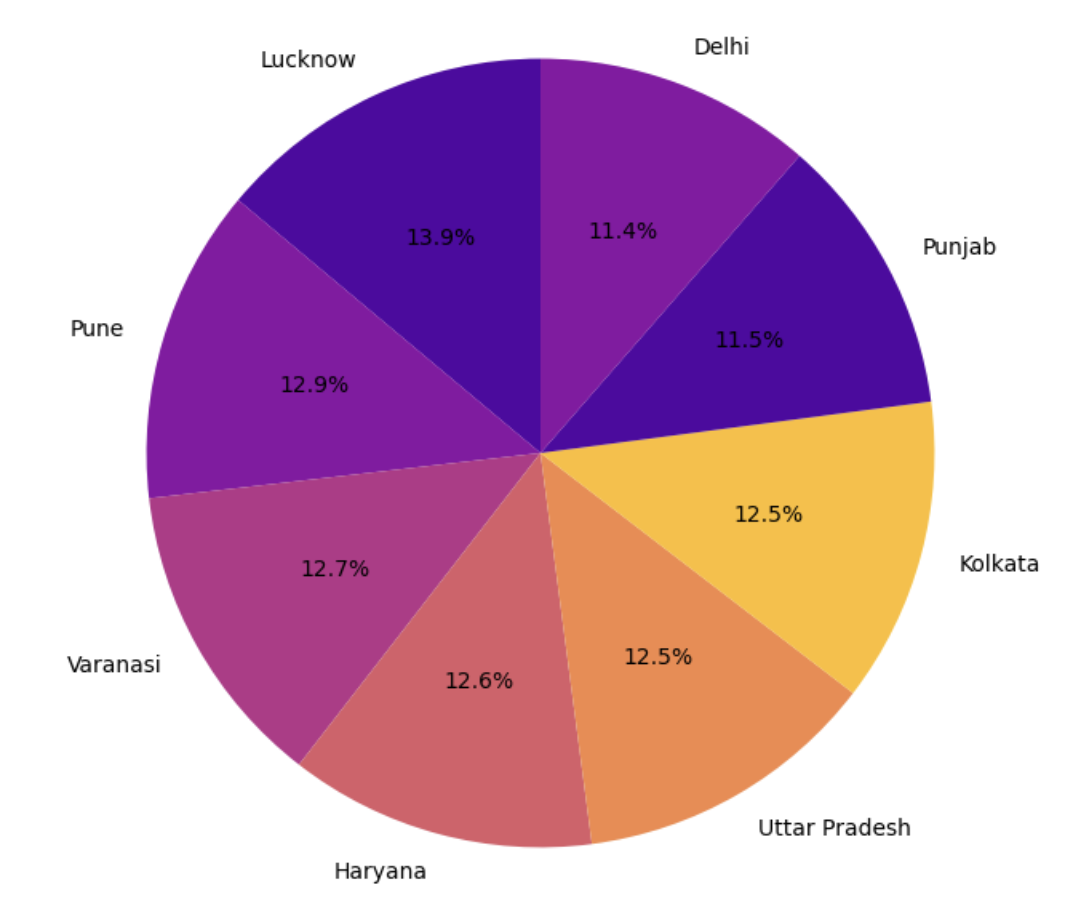
Within the transit dataset, the scatter plot and related regression line provide a visual depiction of the relationship between customer ratings and growing delay. A travel instance is represented by each point on the scatter plot, which shows the correlation between the customer rating and the delay time. The regression line provides more clarity on the general pattern and shows how customer ratings typically fluctuate over time with different delays. This graphic helps explain how delays affect customer satisfaction and can help develop plans for increasing service punctuality.

1. **Scatter plot and correlation coefficient for travel duration vs. cost of travel**

****

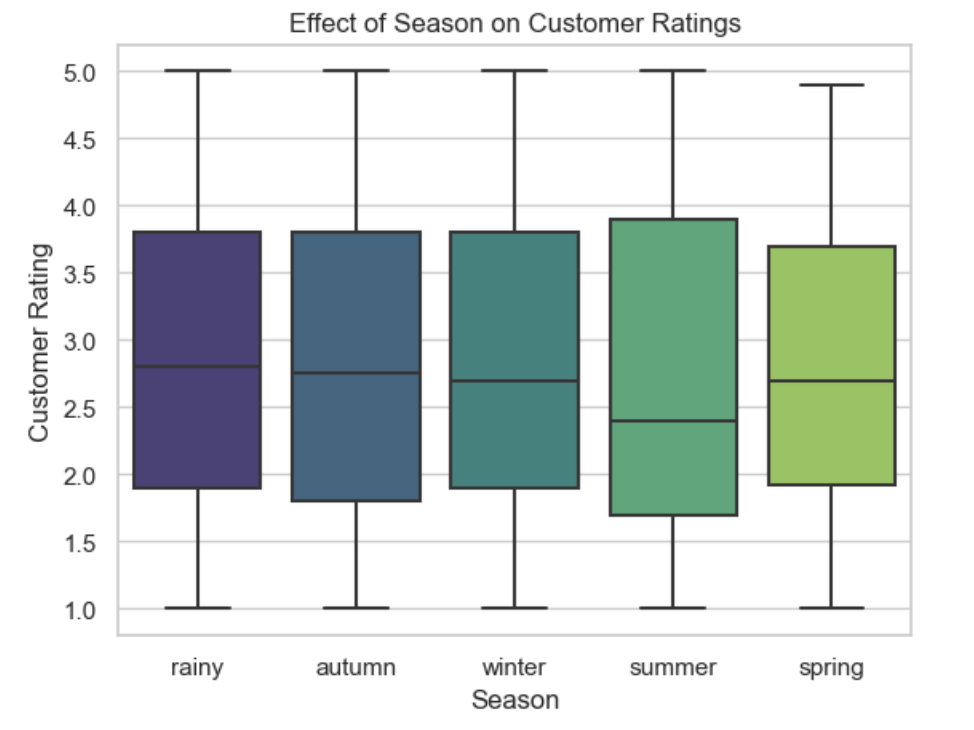
Within the transportation dataset, the scatter plot clearly illustrates the relationship between journey time and travel expense. Plotting each point on the graph as a travel instance gives viewers a visual representation of the relationship between travel time and associated costs. A somewhat positive link between the length of the trip and the cost of the trip is indicated by the correlation coefficient, which was calculated to be 0.47. The scatter plot, which shows a general upward trend indicating that greater travel times are typically linked to higher expenses, supports this information. 'TravelDuration' and 'CostOfTravel' descriptive statistics that go along with the data offer further information on the distribution and central tendencies of these variables, which helps to provide this area of the transportation domain a thorough examination.

1. **What are the most common origins and destinations?**

****

The top ten most popular destinations are displayed in a bar chart, with "Varanasi" coming in at number one. In addition, by displaying the proportionate representation of every destination in the dataset, the pie chart provides a comprehensive understanding of the distribution of destinations. Together, these visualizations improve our knowledge of the transit dataset's most popular travel destinations and points of departure.

1. **Effect of Season on Customer Ratings**

****

Within the synthetic dataset, the boxplot graphically examines how various seasons affect customer evaluations. It offers information on how customer evaluations are distributed throughout different seasons, emphasizing the diversity and core trend for each. The 'viridis' palette helps to distinguish the seasons visually, which makes it easier to comprehend any trends or variances in customer ratings that may be caused by seasonal influences.

**CONCLUSION**

To sum up, this investigation into transit insights using synthetic data has brought a spotlight on a number of aspects related to the transportation industry. I was able to identify patterns in travel behavior that are influenced by dynamics between origin and destination, travel modes, and seasons by carefully analyzing and visualizing my data. With the help of strong Python libraries like NumPy, Pandas, Seaborn, and Matplotlib, I was able to conduct a thorough investigation and obtain insightful information that would help travel agencies improve their service offerings, pricing policies, and overall traveler experiences. By exploring the complexities of the synthetic dataset, I was able to highlight the value of data pretreatment and exploration as well as the potential of data-driven decision-making to optimize the travel industry for both passengers and providers.