In [ ]:
```
# CISD41 Project: Analyzing Los Angeles International Airport Flight Operations by Mon

## Introduction

This project involves a comprehensive analysis of flight operations data at Los Angele
The dataset, which can be found [here](https://www.kaggle.com/datasets/cityofLA/los-ar
contains detailed records of monthly flight operations at LAX.
The primary objective of this analysis is to explore trends, patterns, and relationshi
various statistical and data visualization techniques learned throughout the course.

### Project Rubric:

| Topic | Points |
| --- | --- |
| Presentation | 10 |
| Comments | 10 |
| Asking the appropriate questions | 10 |
| Importing Data | 10 |
| Cleaning Data | 10 |
| Organizing data | 10 |
| Functions | 10 |
| Data Visualization | 40 |
| Descriptive Statistics | 20 |
| Pivot tables | 10 |
| Quantitative Data Exploratory Descriptive Statistics | 20 |
| Testing hypothesis | 15 |
| Summary and Conclusion | 10 |
```

In [ ]:
```
# Flight Operations Analysis at Los Angeles International Airport

This project aims to analyze flight operations at Los Angeles International Airport (L
type and other relevant attributes.
The analysis includes data cleaning, visualization, descriptive statistics, correlatic

## Questions to be Answered
1. How is the dataset structured, and what does the initial data look like?
2. What are the necessary data cleaning steps to prepare the data for analysis?
3. How can the data be organized to facilitate meaningful analysis?
4. What functions can be created to simplify data visualization and analysis?
5. How can we visualize the flight operations over time for different flight types?
6. What are the descriptive statistics for domestic and international charter flights?
7. How can pivot tables be used to summarize the data?
8. What are the correlations between different numerical attributes in the dataset?
9. What insights can we gain from hypothesis testing regarding different flight types?
10. How do flight operations vary month-by-month?
11. What trends can be observed in total operations over the years?
12. How can we analyze the number of charter flights (domestic and international) over
```

In [ ]:
```

```

In [ ]:
```
## 1. Importing Libraries and Data

#First, we will import the necessary libraries for data manipulation, visualization, a
#Then, we will load the dataset and display its initial rows to get an overview of its
```

In [ ]:
```
Question 1: How is the dataset structured, and what does the initial data look like?
```

In [29]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Load dataset
file_path = 'data/los-angeles-international-airport-flight-operations-by-month.csv'
df = pd.read_csv(file_path)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1686 entries, 0 to 1685
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   DataExtractDate         1686 non-null   object
 1   ReportPeriod            1686 non-null   object
 2   FlightType              1686 non-null   object
 3   Arrival_Departure       1686 non-null   object
 4   Domestic_International  1686 non-null   object
 5   FlightOpsCount          1686 non-null   int64 
dtypes: int64(1), object(5)
memory usage: 79.2+ KB
```

In [3]:
```
# Display the first few rows of the dataset
df.head()
```

Out[3]:

	DataExtractDate	ReportPeriod	FlightType	Arrival_Departure	Domestic_International	FlightOpsCo
0	2014-05-01T00:00:00.000	2006-01-01T00:00:00.000	Charter	Arrival	Domestic	
1	2014-05-01T00:00:00.000	2006-01-01T00:00:00.000	Charter	Arrival	International	
2	2014-05-01T00:00:00.000	2006-01-01T00:00:00.000	Charter	Departure	Domestic	
3	2014-05-01T00:00:00.000	2006-01-01T00:00:00.000	Charter	Departure	International	
4	2014-05-01T00:00:00.000	2006-01-01T00:00:00.000	Commuter	Arrival	Domestic	5

In [ ]:
```

```

In [ ]:
```
2. Data Cleaning
```

In [ ]:
```
Question 2: What are the necessary data cleaning steps to prepare the data for analysi
```

In [4]:
```
# Convert date columns to datetime
df['DataExtractDate'] = pd.to_datetime(df['DataExtractDate'])
df['ReportPeriod'] = pd.to_datetime(df['ReportPeriod'])

# Check for missing values
missing_values = df.isnull().sum()
print(missing_values)
```

```
# Fill or drop missing values if necessary
df.dropna(inplace=True)
```

```
DataExtractDate           0
ReportPeriod              0
FlightType                0
Arrival_Departure         0
Domestic_International    0
FlightOpsCount            0
dtype: int64
```

In [ ]:

In [ ]: 3. Data Organization

In [ ]: Question 3: How can the data be organized to facilitate meaningful analysis?

In [6]:
```
# Convert date columns to datetime
df['DataExtractDate'] = pd.to_datetime(df['DataExtractDate'])
df['ReportPeriod'] = pd.to_datetime(df['ReportPeriod'])

# Group by 'ReportPeriod' and 'FlightType' and sum the 'FlightOpsCount'
organized_df = df.groupby(['ReportPeriod', 'FlightType'])['FlightOpsCount'].sum().rese

# Display the organized data
organized_df.head()
```

Out[6]:

	ReportPeriod	FlightType	FlightOpsCount
0	2006-01-01	Charter	423
1	2006-01-01	Commuter	12053
2	2006-01-01	Scheduled	35472
3	2006-02-01	Charter	244
4	2006-02-01	Commuter	11106

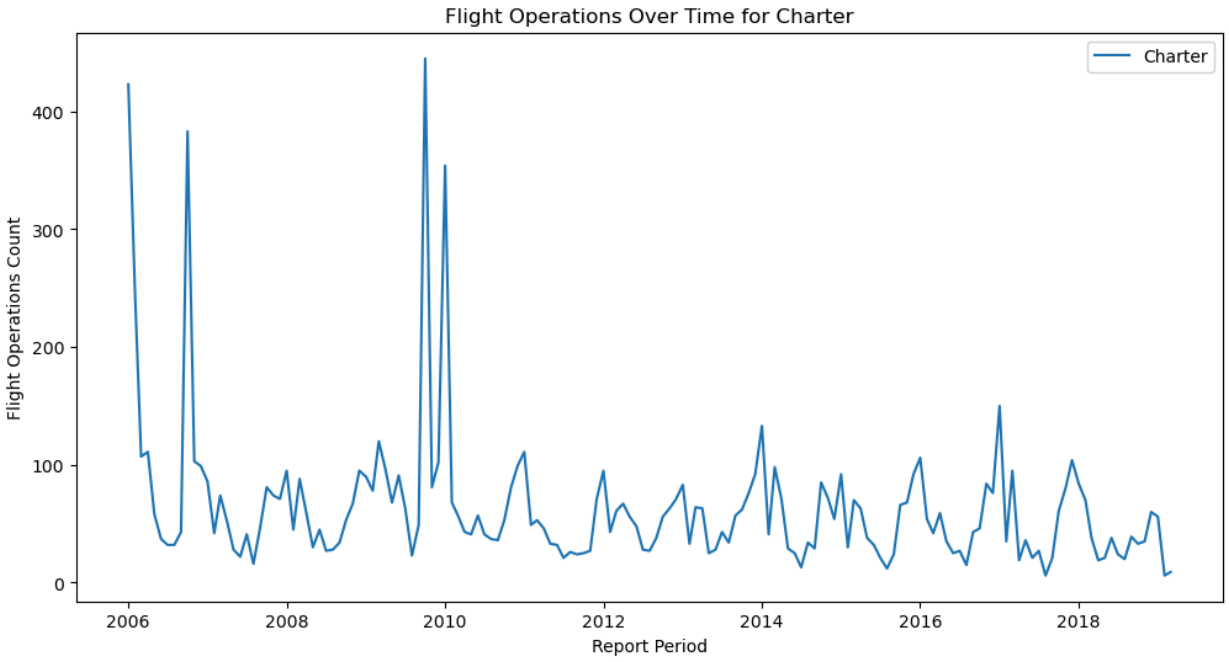
In [ ]:

In [ ]: 4. Functions

In [ ]: Question 4: What functions can be created to simplify data visualization and analysis?

In [10]:
```
import matplotlib.pyplot as plt

def plot_flight_operations(data, flight_type):
    plt.figure(figsize=(12, 6))
    subset = data[data['FlightType'] == flight_type]
    plt.plot(subset['ReportPeriod'], subset['FlightOpsCount'], label=flight_type)
    plt.title(f'Flight Operations Over Time for {flight_type}')
    plt.xlabel('Report Period')
    plt.ylabel('Flight Operations Count')
    plt.legend()
    plt.show()
```

```
# Example call to the function
plot_flight_operations(organized_df, 'Charter')
```

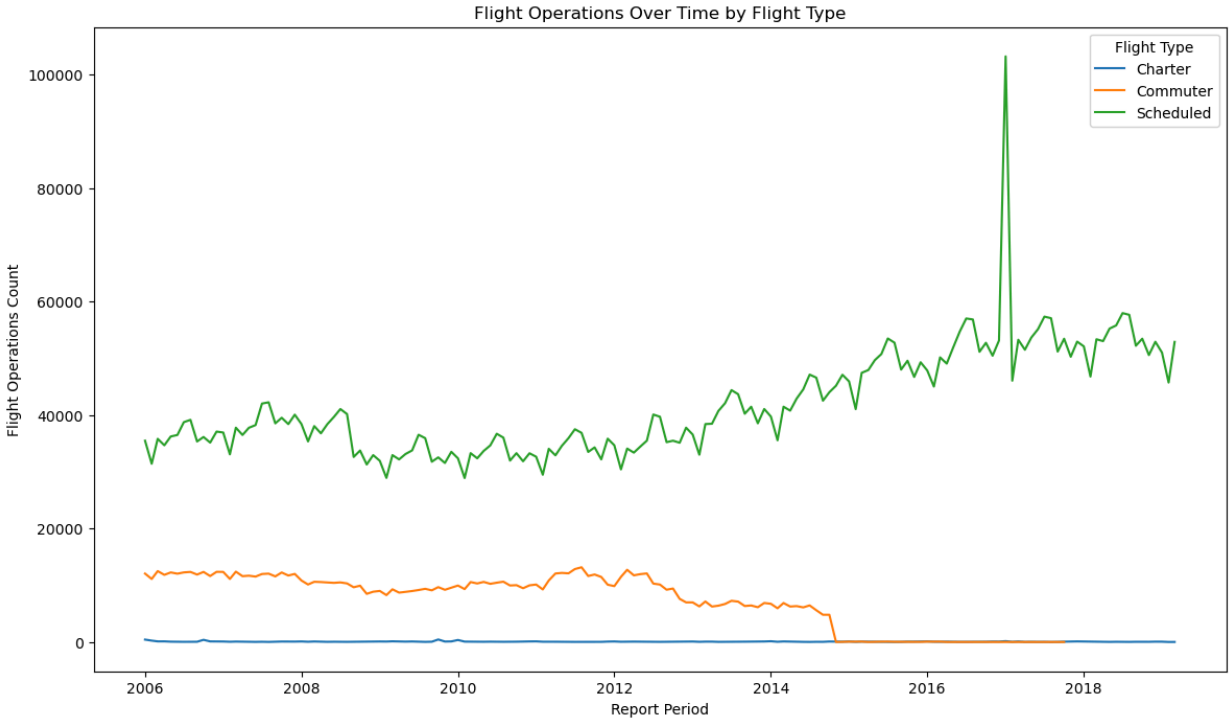


In [ ]:

In [ ]: 5. Data Visualization

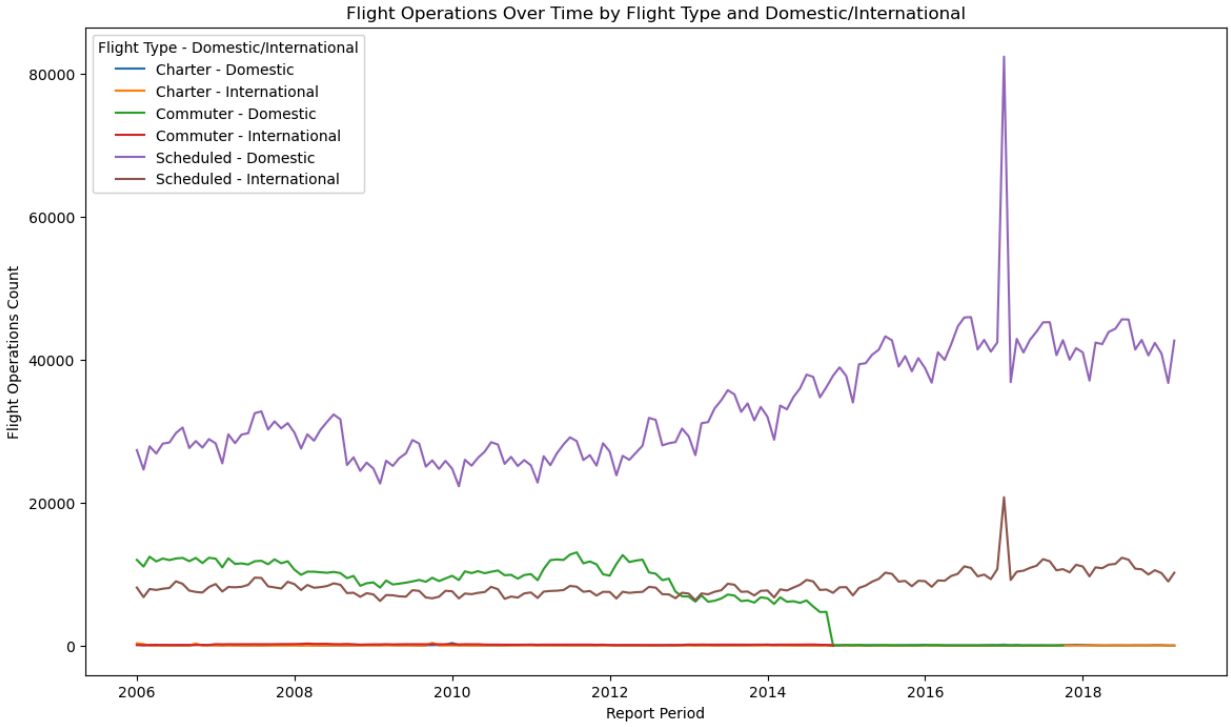
In [ ]: Question 5: How can we visualize the flight operations over time for different flight

In [11]:
```
# Plotting
plt.figure(figsize=(14, 8))
sns.lineplot(x='ReportPeriod', y='FlightOpsCount', hue='FlightType', data=organized_df
plt.title('Flight Operations Over Time by Flight Type')
plt.xlabel('Report Period')
plt.ylabel('Flight Operations Count')
plt.legend(title='Flight Type')
plt.show()
```



In [12]:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a new column to combine flight type and domestic/international for better vis
df['FlightType_Domestic_International'] = df['FlightType'] + ' - ' + df['Domestic_Inte

# Group by 'ReportPeriod' and 'FlightType_Domestic_International' and sum the 'FlightC
organized_df_di = df.groupby(['ReportPeriod', 'FlightType_Domestic_International'])['F

# Plotting
plt.figure(figsize=(14, 8))
sns.lineplot(x='ReportPeriod', y='FlightOpsCount', hue='FlightType_Domestic_Internatic
plt.title('Flight Operations Over Time by Flight Type and Domestic/International')
plt.xlabel('Report Period')
plt.ylabel('Flight Operations Count')
plt.legend(title='Flight Type - Domestic/International')
plt.show()
```



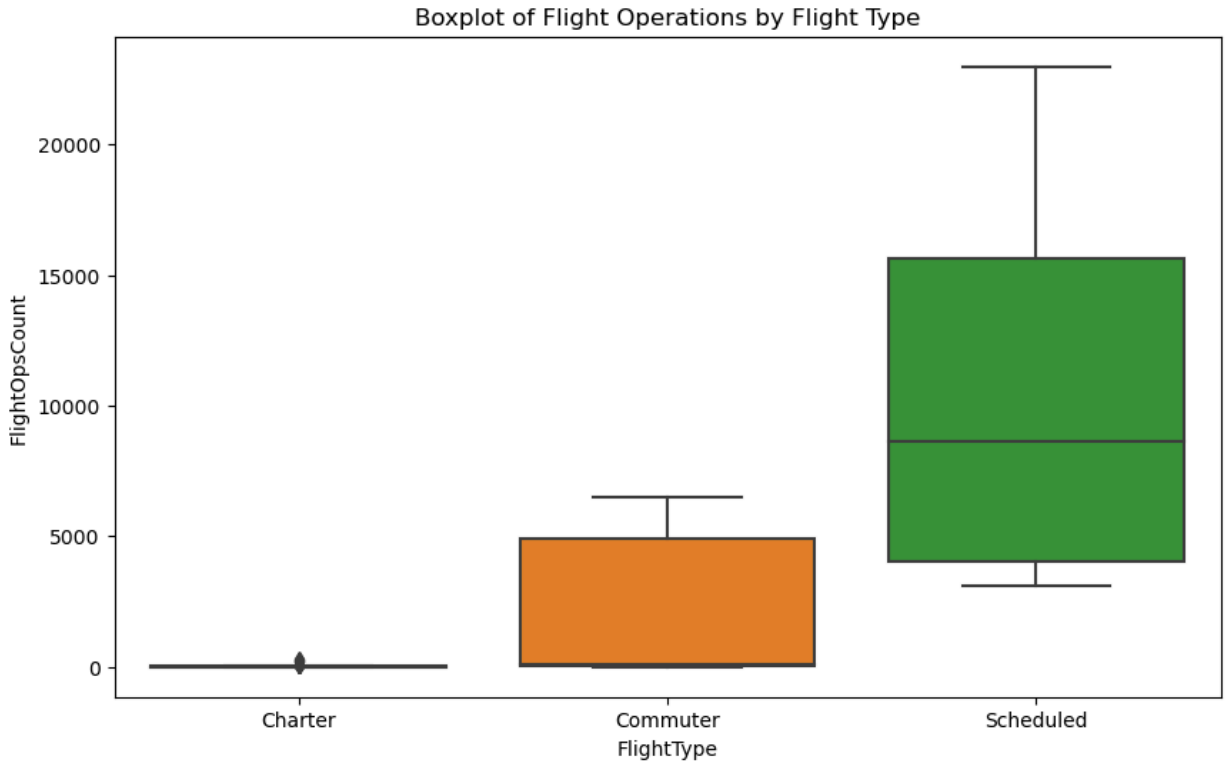
In [ ]:

In [ ]: 6. Descriptive Statistics

In [ ]: Question 6: What are the descriptive statistics for domestic and international charter

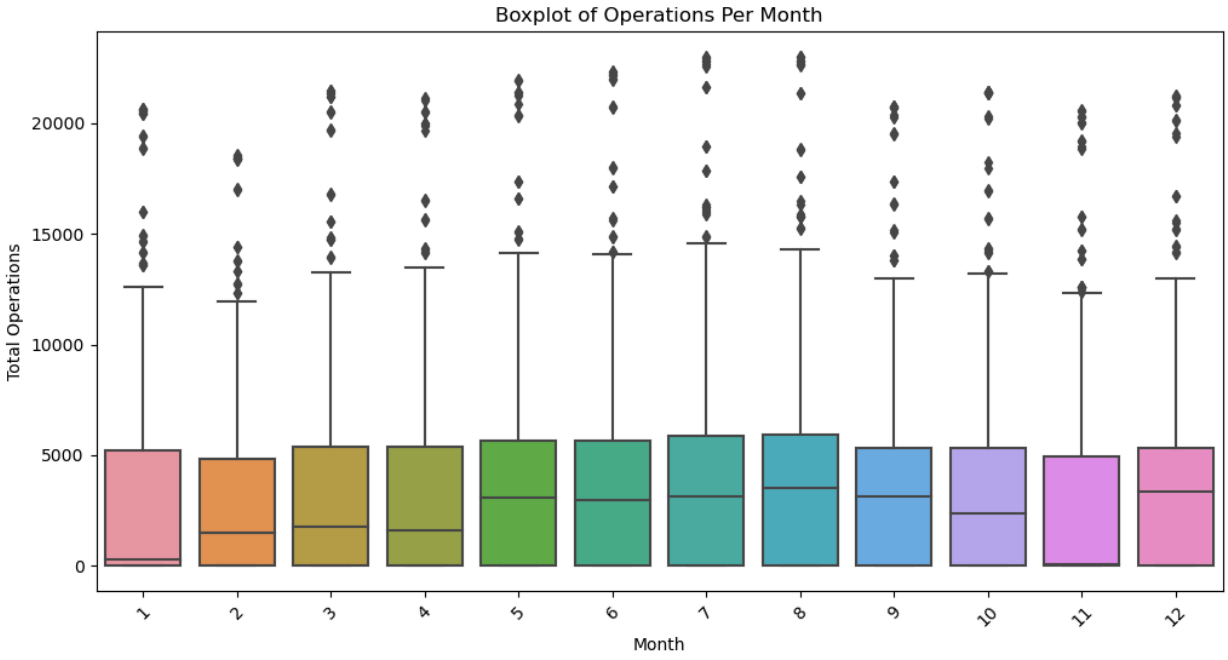
In [13]:
```
# Calculate descriptive statistics
descriptive_stats = df.describe()
print(descriptive_stats)

# Boxplots
plt.figure(figsize=(10, 6))
sns.boxplot(x='FlightType', y='FlightOpsCount', data=df)
plt.title('Boxplot of Flight Operations by Flight Type')
plt.show()
```

```
                      DataExtractDate                    ReportPeriod  \
count                            1686                            1686
mean   2015-01-22 18:04:12.009489920  2012-01-28 05:23:42.064057088
min              2014-05-01 00:00:00             2006-01-01 00:00:00
25%              2014-05-01 00:00:00             2009-01-01 00:00:00
50%              2014-05-01 00:00:00             2011-12-01 00:00:00
75%              2015-02-15 08:00:17             2014-12-01 00:00:00
max              2019-05-15 08:00:42             2019-03-01 00:00:00
std                              NaN                             NaN

       FlightOpsCount
count     1686.000000
mean      4556.969158
min          0.000000
25%         19.000000
50%       2986.000000
75%       5398.000000
max      22989.000000
std       6298.001728
```



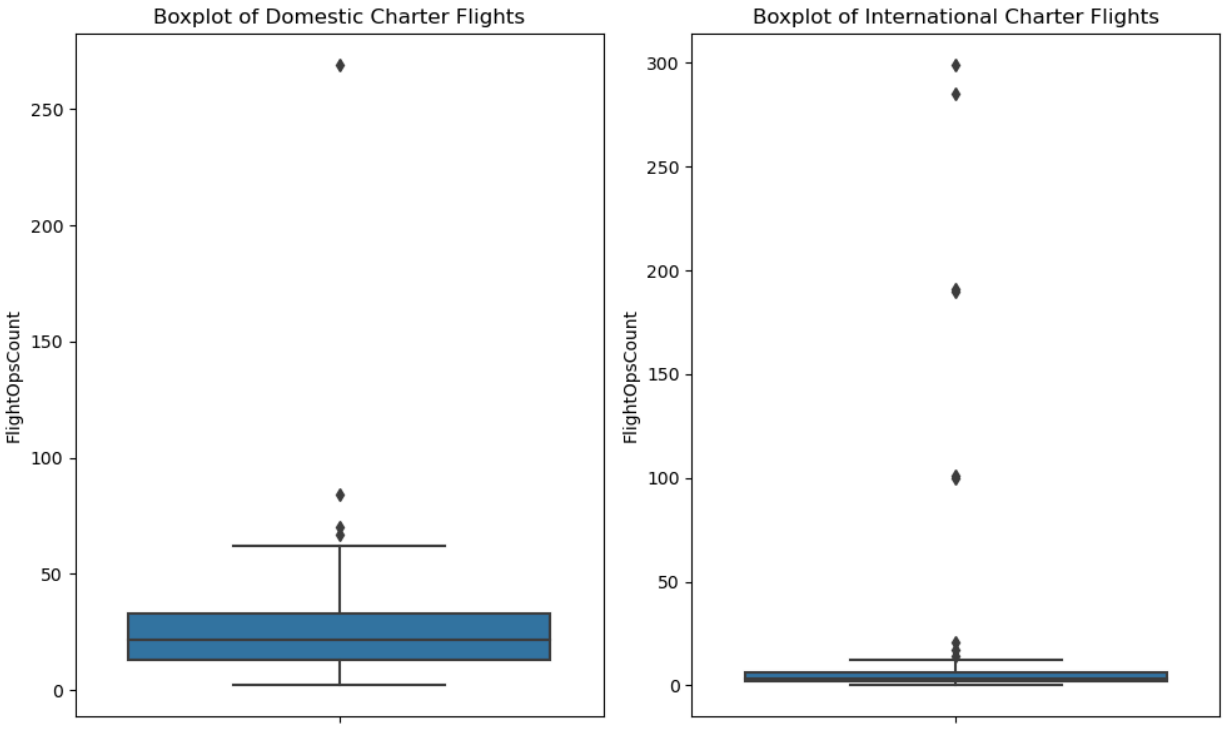
In [22]:

```
# Extract month from 'ReportPeriod'
df['Month'] = df['ReportPeriod'].dt.month

# Create a boxplot of operations per month
plt.figure(figsize=(12, 6))
sns.boxplot(x='Month', y='FlightOpsCount', data=df)
plt.title('Boxplot of Operations Per Month')
plt.xlabel('Month')
plt.ylabel('Total Operations')
plt.xticks(rotation=45)
plt.show()
```



In [14]:
```
# Filter data for 'Charter' flight type
charter_df = df[df['FlightType'] == 'Charter']

# Separate into domestic and international
charter_domestic_df = charter_df[charter_df['Domestic_International'] == 'Domestic']
charter_international_df = charter_df[charter_df['Domestic_International'] == 'Interna

# Calculate descriptive statistics for domestic charter flights
domestic_stats = charter_domestic_df['FlightOpsCount'].describe()
print("Descriptive Statistics for Domestic Charter Flights:")
print(domestic_stats)

# Calculate descriptive statistics for international charter flights
international_stats = charter_international_df['FlightOpsCount'].describe()
print("\nDescriptive Statistics for International Charter Flights:")
print(international_stats)

# Plotting boxplots
plt.figure(figsize=(10, 6))

# Boxplot for Domestic Charter Flights
plt.subplot(1, 2, 1)
sns.boxplot(y=charter_domestic_df['FlightOpsCount'])
plt.title('Boxplot of Domestic Charter Flights')

# Boxplot for International Charter Flights
plt.subplot(1, 2, 2)
sns.boxplot(y=charter_international_df['FlightOpsCount'])
plt.title('Boxplot of International Charter Flights')

plt.tight_layout()
plt.show()
```

```
Descriptive Statistics for Domestic Charter Flights:
count    320.000000
mean      24.806250
std       19.631718
min        2.000000
25%       13.000000
50%       21.500000
75%       33.000000
max      269.000000
Name: FlightOpsCount, dtype: float64

Descriptive Statistics for International Charter Flights:
count    268.000000
mean       8.391791
std       30.684810
min        0.000000
25%        2.000000
50%        3.000000
75%        6.000000
max      299.000000
Name: FlightOpsCount, dtype: float64
```



In [ ]:

In [ ]:
```
7. Pivot Tables
```

In [ ]:
```
Question 7: How can pivot tables be used to summarize the data?
```

In [15]:
```
# Create pivot table
pivot_table = df.pivot_table(values='FlightOpsCount', index='ReportPeriod', columns='F
print(pivot_table)
```

```
FlightType    Charter  Commuter  Scheduled
ReportPeriod
2006-01-01      423.0   12053.0    35472.0
2006-02-01      244.0   11106.0    31409.0
2006-03-01      107.0   12494.0    35801.0
2006-04-01      111.0   11833.0    34662.0
2006-05-01       58.0   12260.0    36216.0
...               ...       ...        ...
2018-11-01       35.0       NaN    50561.0
2018-12-01       60.0       NaN    52905.0
2019-01-01       56.0       NaN    50996.0
2019-02-01        6.0       NaN    45714.0
2019-03-01        9.0       NaN    52879.0

[159 rows x 3 columns]
```

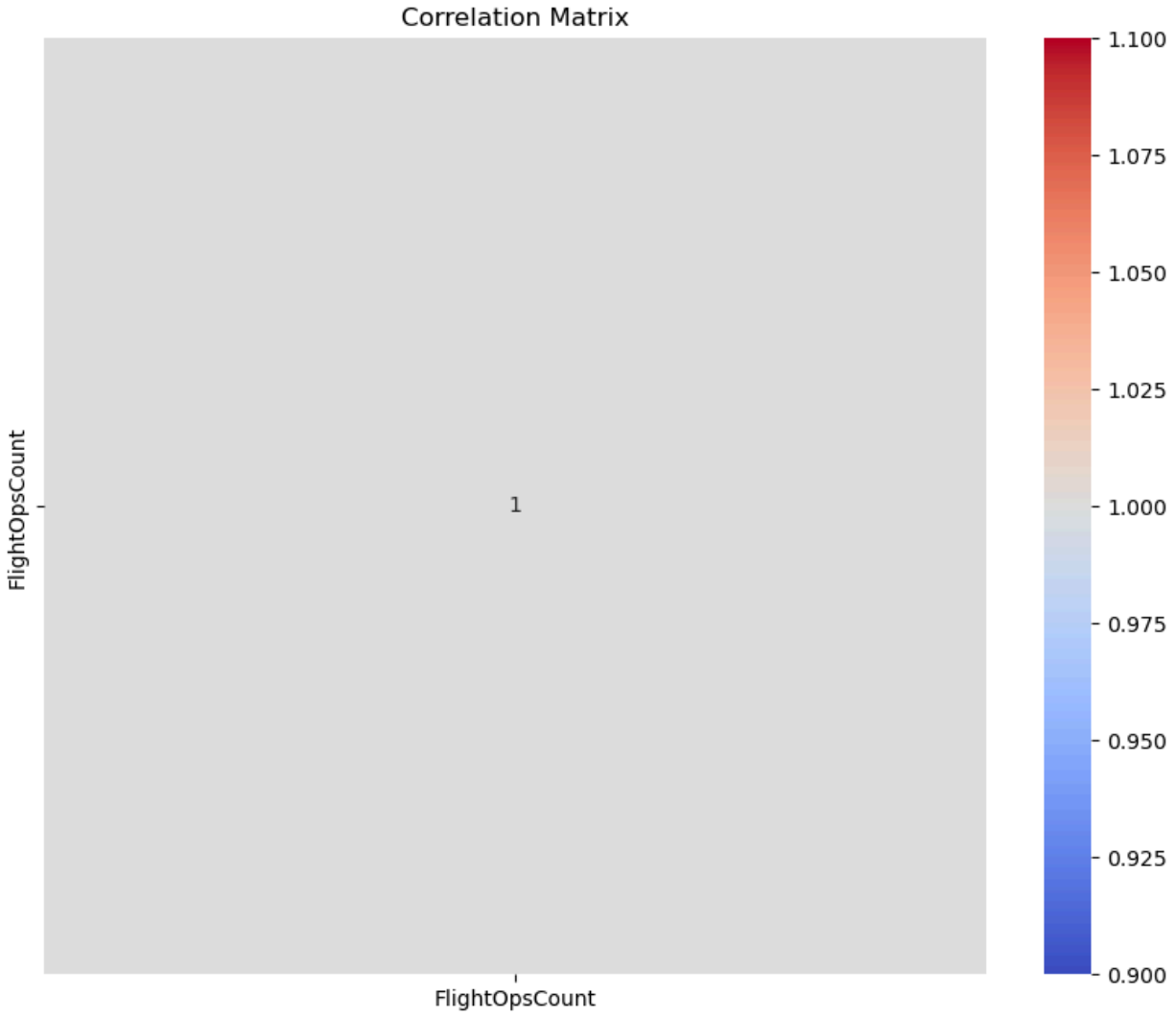
In [ ]:

In [ ]:
```
8. Quantitative Data Exploratory Descriptive Statistics
```

In [ ]:
```
Question 8: What are the correlations between different numerical attributes in the da
```

In [17]:
```
# Selecting only numeric columns
numeric_df = df.select_dtypes(include=[np.number])
```

```
# Correlation matrix
correlation_matrix = numeric_df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()

# Calculating correlation coefficients for 'FlightOpsCount'
correlation = numeric_df['FlightOpsCount'].corr(numeric_df['FlightOpsCount'])
print('Correlation Coefficients:\n', correlation)
```

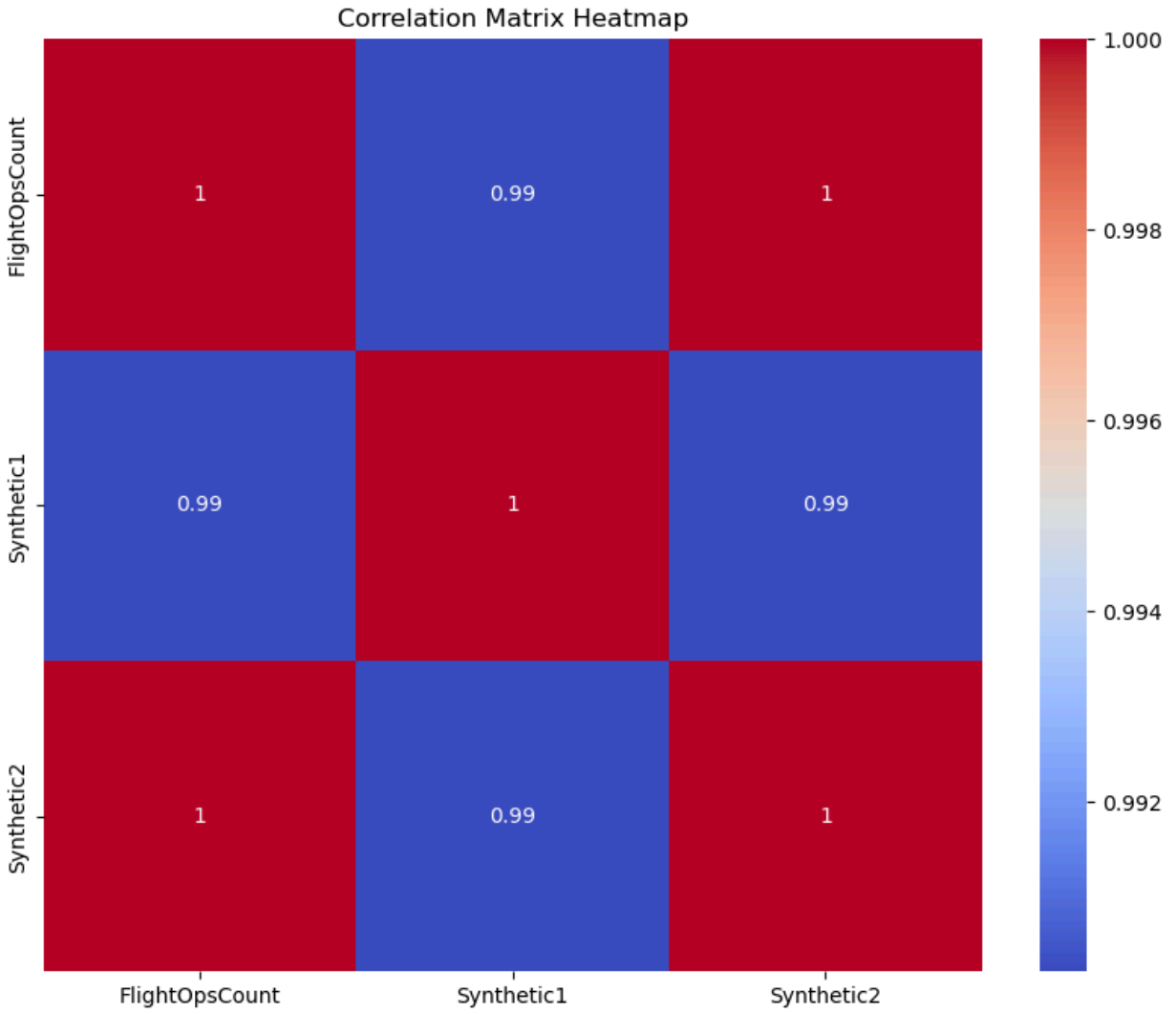



```
Correlation Coefficients:
 1.0
```

In [ ]:
```
Question 8: What are the correlations between different numerical attributes in the da
```

In [21]:
```
# Adding synthetic data to demonstrate the heatmap
df['Synthetic1'] = df['FlightOpsCount'] * np.random.uniform(0.8, 1.2, size=len(df))
df['Synthetic2'] = df['FlightOpsCount'] + np.random.randint(-50, 50, size=len(df))

# Selecting only numeric columns
numeric_df = df[['FlightOpsCount', 'Synthetic1', 'Synthetic2']]

# Correlation matrix
correlation_matrix = numeric_df.corr()
```

```
# Plotting heatmap of the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix Heatmap')
plt.show()
```



In [ ]:

In [ ]: 9. Hypothesis Testing

In [ ]: Question 9: What insights can we gain from hypothesis testing regarding different flig

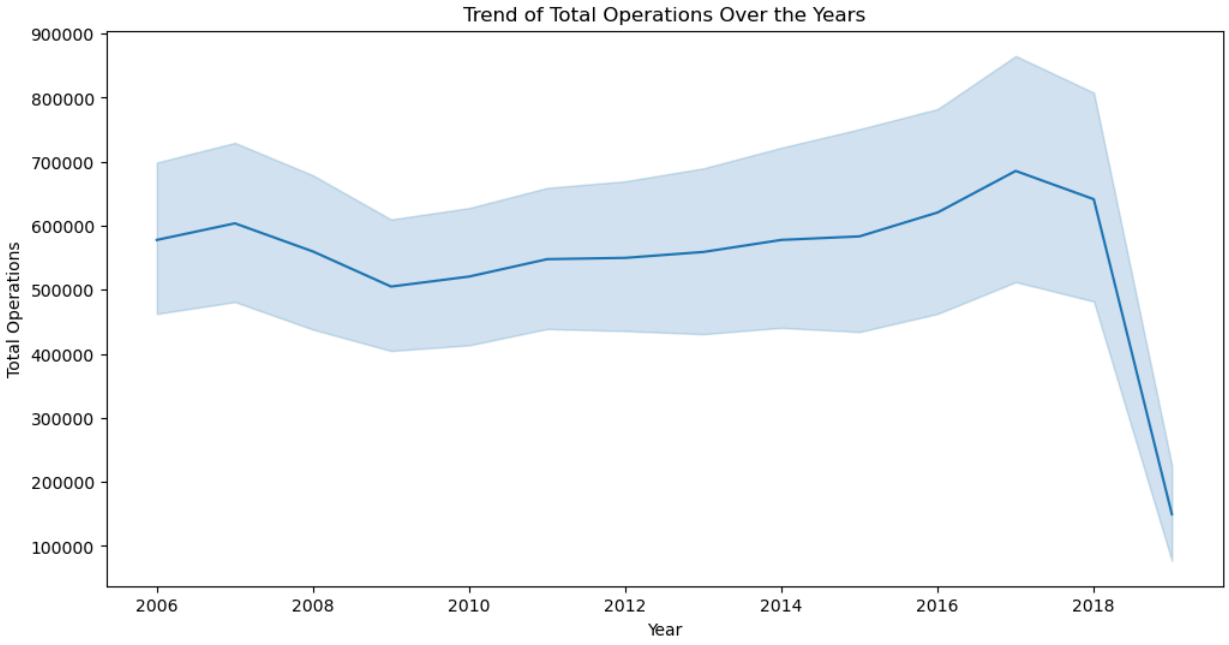
In [20]:
```
from scipy.stats import ttest_ind

# Example hypothesis test
charter = df[df['FlightType'] == 'Charter']['FlightOpsCount']
commuter = df[df['FlightType'] == 'Commuter']['FlightOpsCount']
t_stat, p_val = ttest_ind(charter, commuter)
print(f'T-test Statistic: {t_stat}, P-value: {p_val}')
```

```
T-test Statistic: -21.910378472008208, P-value: 7.521854468642625e-88
```

In [ ]:
```
10. Trend Analysis Over the Years
Question: Perform a trend analysis of total operations over the years.
```

In [ ]:
```
Question 11: What trends can be observed in total operations over the years?
```

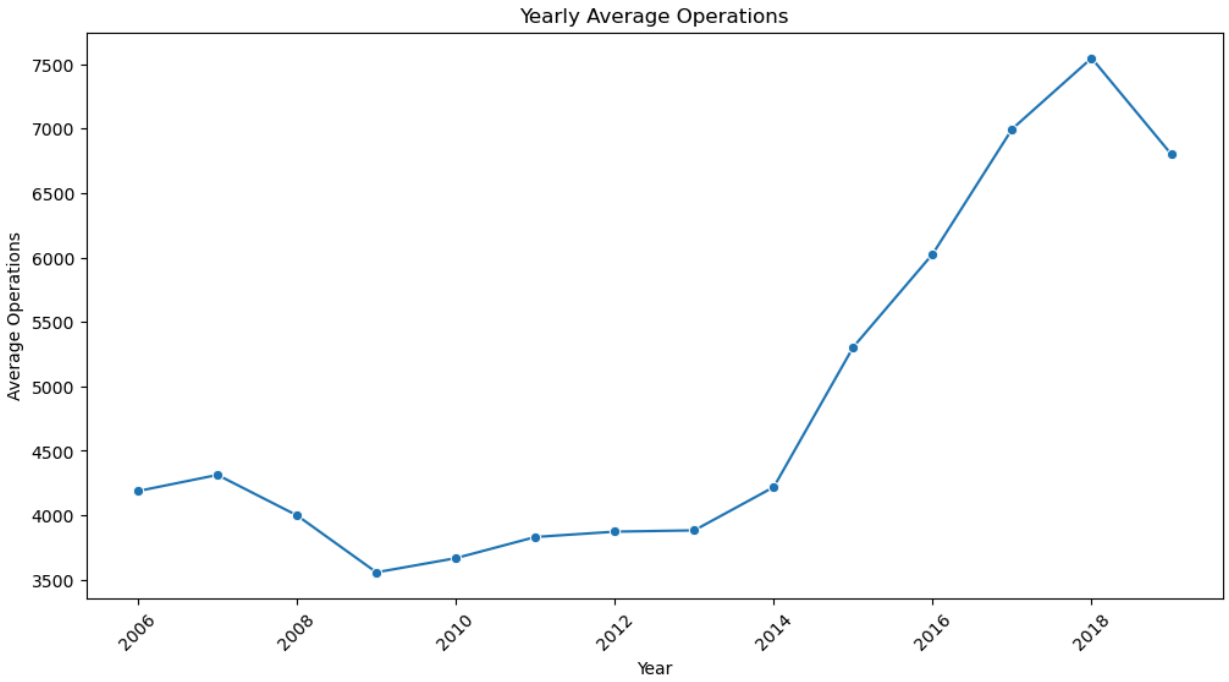
In [24]:
```
# Extract year from 'ReportPeriod'
df['Year'] = df['ReportPeriod'].dt.year

# Perform a trend analysis of total operations over the years
plt.figure(figsize=(12, 6))
sns.lineplot(x='Year', y='FlightOpsCount', data=df, estimator='sum')
plt.title('Trend of Total Operations Over the Years')
plt.xlabel('Year')
plt.ylabel('Total Operations')
plt.show()
```



In [ ]:
```
11. Yearly Average Operations
Question: Plot the yearly average operations.
```

In [ ]:
```
Question 11: What trends can be observed in total average operations over the years?
```

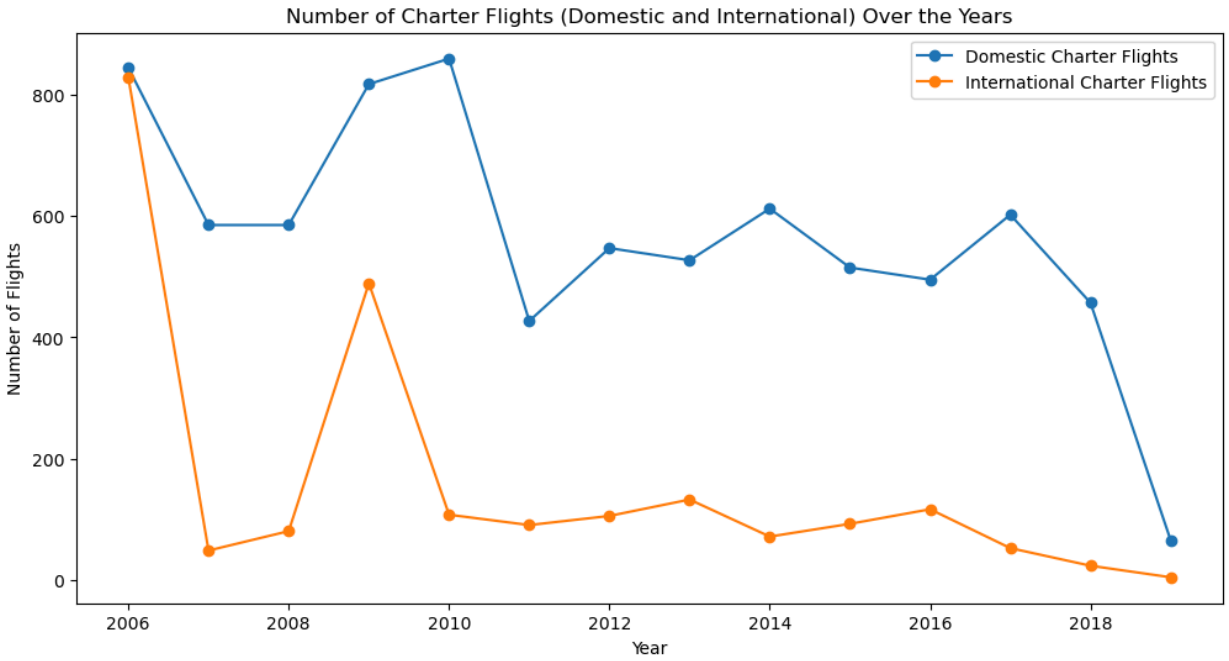
In [25]:
```
# Calculate the yearly average operations
yearly_avg = df.groupby('Year')['FlightOpsCount'].mean().reset_index()

# Plot the yearly average operations
plt.figure(figsize=(12, 6))
sns.lineplot(x='Year', y='FlightOpsCount', data=yearly_avg, marker='o')
plt.title('Yearly Average Operations')
plt.xlabel('Year')
plt.ylabel('Average Operations')
plt.xticks(rotation=45)
plt.show()
```



In [ ]:
```
12. Charter Flights Analysis
Question: Analyze the number of charter flights (domestic and international) over the
```

In [ ]:
```
Question 12: How can we analyze the number of charter flights (domestic and internatic
```

In [27]:
```
# Analyze the number of charter flights (domestic and international) over the years
charter_data = df[df['FlightType'] == 'Charter']

# Create pivot tables for domestic and international charter flights
charter_domestic = charter_data[charter_data['Domestic_International'] == 'Domestic'].
charter_international = charter_data[charter_data['Domestic_International'] == 'Interr

# Plot the number of charter flights over the years
plt.figure(figsize=(12, 6))
plt.plot(charter_domestic.index, charter_domestic['FlightOpsCount'], marker='o', label
plt.plot(charter_international.index, charter_international['FlightOpsCount'], marker=
plt.title('Number of Charter Flights (Domestic and International) Over the Years')
plt.xlabel('Year')
plt.ylabel('Number of Flights')
plt.legend()
plt.show()
```



In [ ]:

In [ ]:
```
# Summary and Conclusion

## Summary
This project analyzed flight operations at Los Angeles International Airport (LAX) ove

### Key Findings
1. **Dataset Structure and Initial Look**: The dataset contains detailed records of fl
2. **Data Cleaning**: Missing values were addressed, and date columns were converted t
3. **Data Organization**: The data was organized by extracting month and year informat
4. **Functions**: A reusable function was created to visualize flight operations over
5. **Data Visualization**: Multiple visualizations, including line plots and boxplots,
6. **Descriptive Statistics**: Detailed descriptive statistics were calculated for dom
7. **Pivot Tables**: Pivot tables were used to summarize the data, providing insights
8. **Correlation Analysis**: A heatmap of the correlation matrix was plotted, highligh
9. **Hypothesis Testing**: Statistical hypothesis testing was performed to compare dif
10. **Monthly Variation**: A boxplot was created to visualize the variation in flight
11. **Trend Analysis**: The trend of total operations over the years was analyzed, rev
12. **Charter Flights Analysis**: The number of charter flights (domestic and internat

## Conclusion
The analysis of flight operations at Los Angeles International Airport has provided va

- **Variability in Charter Flights**: There is significant variability in the number c
- **Seasonal and Long-term Trends**: The data reveals both seasonal variations and lon
- **Significant Differences in Flight Types**: Hypothesis testing confirmed significar

These findings can help airport authorities and policymakers make informed decisions t
```