In [1]:
```
#importing libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
C:\Users\CODEINE\AppData\Local\Temp\ipykernel_7964\4288724001.py:2: DeprecationWa
rning:
Pyarrow will become a required dependency of pandas in the next major release of
pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better i
nteroperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

  import pandas as pd
```

In [2]:
```
#Loading data
df = pd.read_csv("Flyzy Flight Cancellation.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	Flight ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	Scheduled_Dep
0	7319483	Airline D	475	Airport 3	Airport 2	
1	4791965	Airline E	538	Airport 5	Airport 4	
2	2991718	Airline C	565	Airport 1	Airport 2	
3	4220106	Airline E	658	Airport 5	Airport 3	
4	2263008	Airline E	566	Airport 2	Airport 2	

In [4]:
```
df.tail()
```

Out[4]:

	Flight ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	Scheduled_I
2995	1265781	Airline D	395	Airport 2	Airport 3	
2996	5440150	Airline E	547	Airport 1	Airport 4	
2997	779080	Airline C	461	Airport 1	Airport 3	
2998	4044431	Airline B	464	Airport 3	Airport 3	
2999	2806578	Airline A	369	Airport 1	Airport 2	

In [5]:
```
df.shape
```

Out[5]:
```
(3000, 14)
```

In [6]:
```
df.columns
```

Out[6]:
```
Index(['Flight ID', 'Airline', 'Flight_Distance', 'Origin_Airport',
       'Destination_Airport', 'Scheduled_Departure_Time', 'Day_of_Week',
       'Month', 'Airplane_Type', 'Weather_Score',
       'Previous_Flight_Delay_Minutes', 'Airline_Rating', 'Passenger_Load',
       'Flight_Cancelled'],
      dtype='object')
```

Observation: Most of the column names consist of multiple words seperated by underscores,but 'Flight ID' does not follow this format, therefore we need to change it to keep consistancy.

In [7]:
```
#Changing column name
df.rename(columns={'Flight ID' : 'Flight_ID'}, inplace =True)
```

In [8]:
```
df.head(2)
```

Out[8]:

	Flight_ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	Scheduled_De
0	7319483	Airline D	475	Airport 3	Airport 2	
1	4791965	Airline E	538	Airport 5	Airport 4	

CHECKING DATA TYPES OF EACH COLUMN

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 14 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Flight_ID                     3000 non-null   int64  
 1   Airline                       3000 non-null   object 
 2   Flight_Distance               3000 non-null   int64  
 3   Origin_Airport                3000 non-null   object 
 4   Destination_Airport           3000 non-null   object 
 5   Scheduled_Departure_Time      3000 non-null   int64  
 6   Day_of_Week                   3000 non-null   int64  
 7   Month                         3000 non-null   int64  
 8   Airplane_Type                 3000 non-null   object 
 9   Weather_Score                 3000 non-null   float64
 10  Previous_Flight_Delay_Minutes 3000 non-null   float64
 11  Airline_Rating                3000 non-null   float64
 12  Passenger_Load                3000 non-null   float64
 13  Flight_Cancelled              3000 non-null   int64  
dtypes: float64(4), int64(6), object(4)
memory usage: 328.3+ KB
```

Observation: This results indicate that all columns have the correct data types according to the data they contain

In [10]:
```python
#Checking for duplicates entries
duplicates = df[df.duplicated()]
```

In [11]:
```python
duplicates
```

Out[11]:

Flight_ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	Scheduled_Depa

No duplicates on the dataset

CHECKING FOR MISSING VALUES

In [12]:
```python
df.isnull().sum()
```

Out[12]:
```
Flight_ID                        0
Airline                          0
Flight_Distance                  0
Origin_Airport                   0
Destination_Airport              0
Scheduled_Departure_Time         0
Day_of_Week                      0
Month                            0
Airplane_Type                    0
Weather_Score                    0
Previous_Flight_Delay_Minutes    0
Airline_Rating                   0
Passenger_Load                   0
Flight_Cancelled                 0
dtype: int64
```

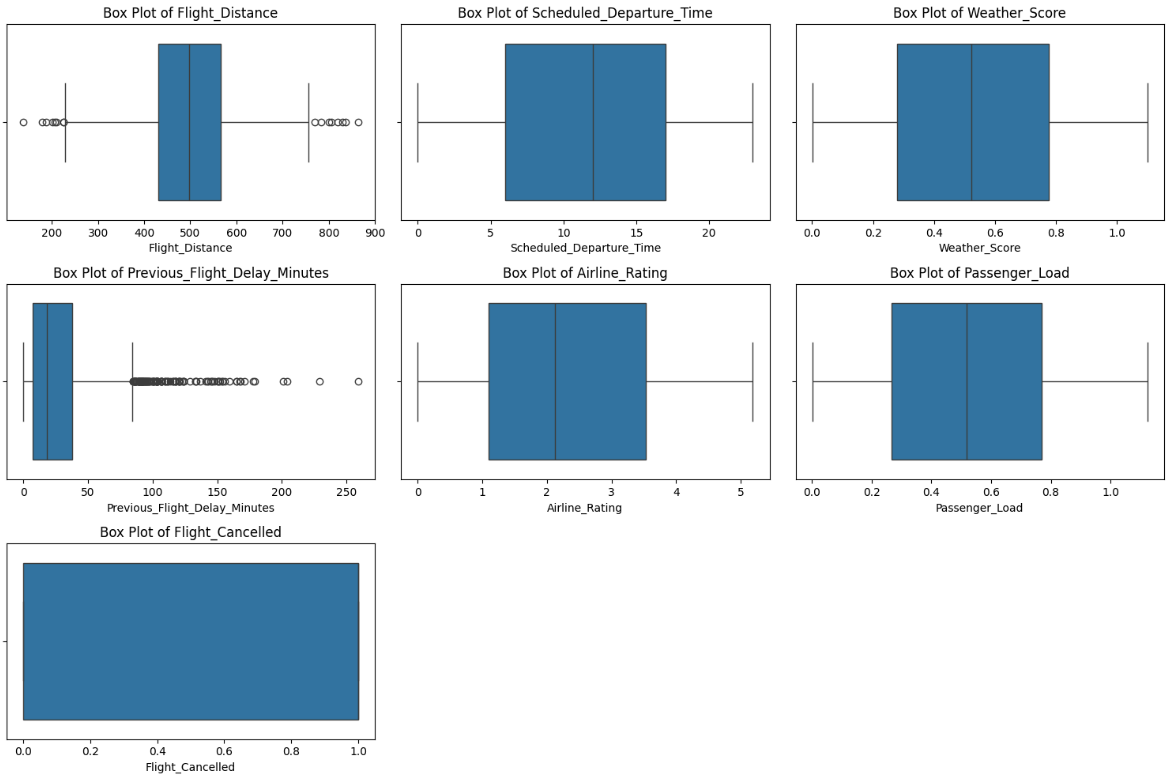
There are no missing values

CHECKING FOR OUTLIERS

In [13]:
```
#Used boxplot to visually check outliers
```

In [14]:
```
columns_to_check =['Flight_Distance',
          'Scheduled_Departure_Time',
          'Weather_Score',
          'Previous_Flight_Delay_Minutes',
          'Airline_Rating', 'Passenger_Load',
          'Flight_Cancelled'
         ]
```

In [15]:
```
plt.figure(figsize =(15,10))
for i, col in enumerate(columns_to_check, 1):
    plt.subplot(3,3, i)
    sns.boxplot(x=df[col])
    plt.title(f'Box Plot of {col}')
plt.tight_layout()
plt.show()
```



This plots shows that the following columns have outliers and have to be handled

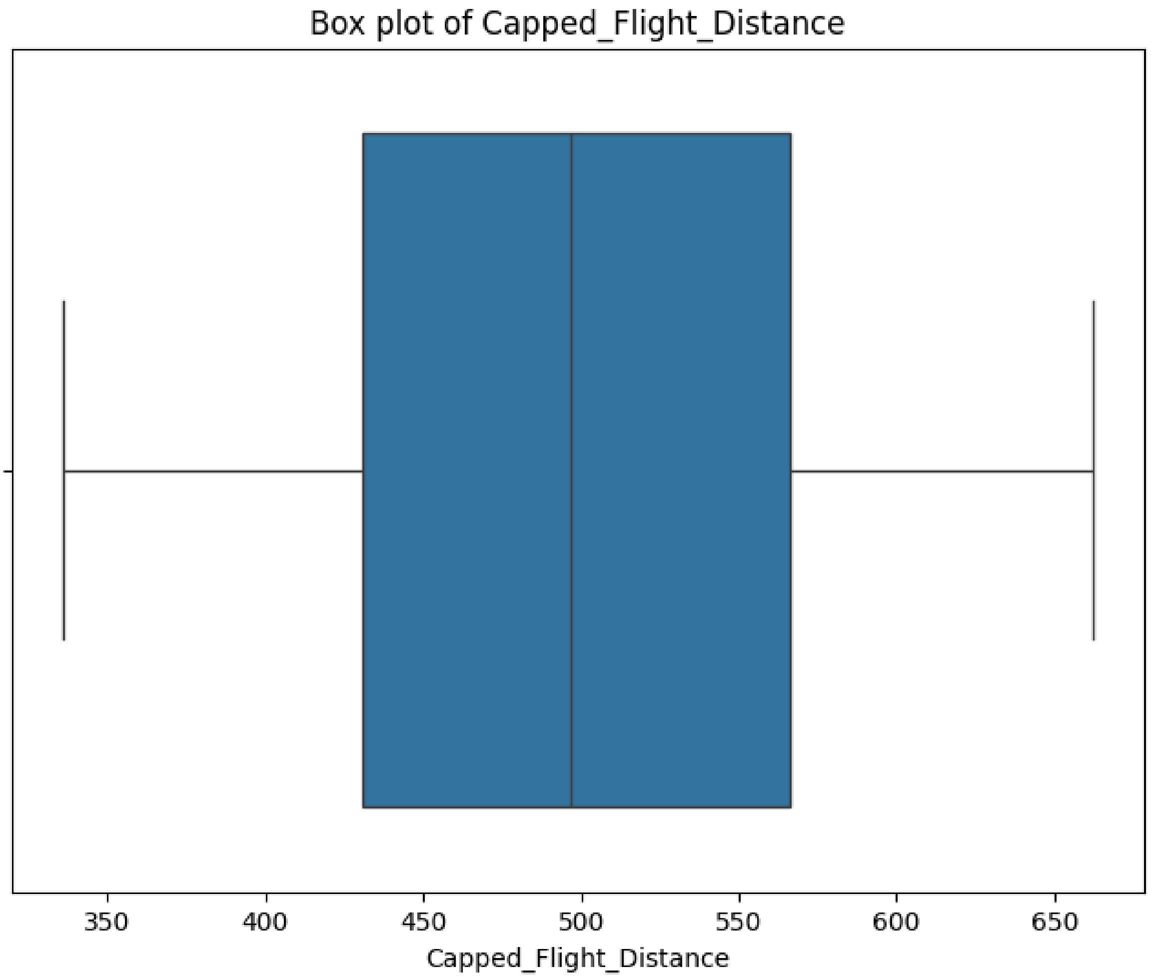
1. Flight_Distance
2. Previous_Flight_Delay_Minutes

1. Handling outliers for Flight_Distance column using Capping method Because it reduces the impact of extreme outliers, which can distort the analysis.

In [16]:
```
#Handling outliers on Flight Distance using Capping approach

#applying threshold
cap_max = df['Flight_Distance'].quantile(0.95)
```

```
cap_min =df['Flight_Distance'].quantile(0.05)
#Apply capping
df['Capped_Flight_Distance'] = np.clip(df['Flight_Distance'], cap_min, cap_max)
```

In [17]:
```
#Plotting transformed column
plt.figure(figsize =(8,6))
sns.boxplot(x=df['Capped_Flight_Distance'])
plt.title('Box plot of Capped_Flight_Distance')
plt.show()
```

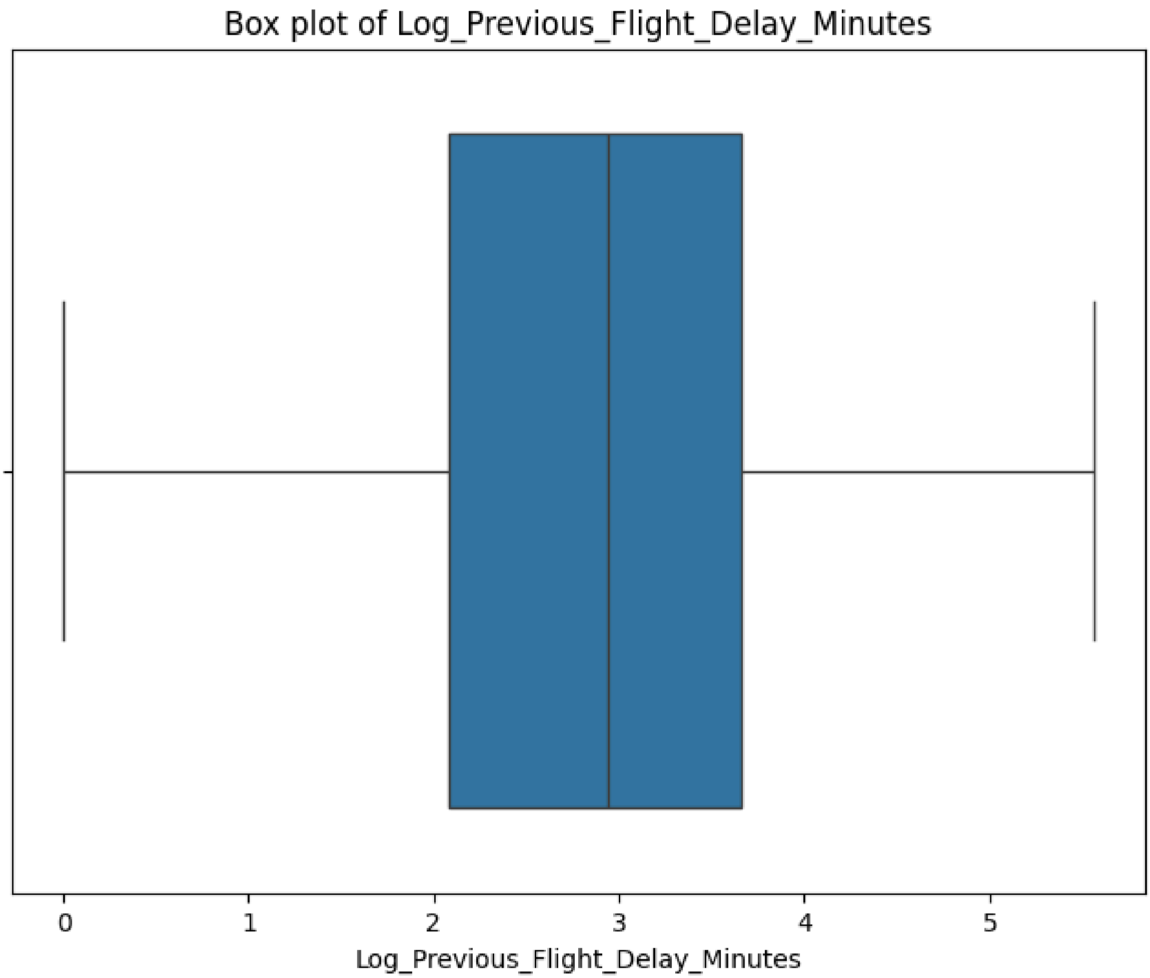


The results shows no more outliers for Flight_Distance

2. Handling Outliers for Previous_Flight_Delay_Minutes Using Log Transformation because data is skewed, compressing the range of delay times, reducing the impact of extreme values.

In [18]:
```
#creating a new column and applying the log
df['Log_Previous_Flight_Delay_Minutes'] = np.log1p(df['Previous_Flight_Delay_Min
```

In [19]:
```
#Plotting transformed column
plt.figure(figsize =(8,6))
sns.boxplot(x=df['Log_Previous_Flight_Delay_Minutes'])
plt.title('Box plot of Log_Previous_Flight_Delay_Minutes')
plt.show()
```



Now the outliers were handled and not showing on the plot

In [20]:
```
df.head(2)
```

Out[20]:

	Flight_ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	Scheduled_Dep
0	7319483	Airline D	475	Airport 3	Airport 2	
1	4791965	Airline E	538	Airport 5	Airport 4	

In [21]:
```
df.shape
```

Out[21]: (3000, 16)

EXPLORATORY DATA ANALYSIS

DESCRIPTIVE STATISTICS

In [22]:
```
df.describe()
```

Out[22]:

	Flight_ID	Flight_Distance	Scheduled_Departure_Time	Day_of_Week	Mo
count	3.000000e+03	3000.000000	3000.000000	3000.000000	3000.00
mean	4.997429e+06	498.909333	11.435000	3.963000	6.38
std	2.868139e+06	98.892266	6.899298	2.016346	3.47
min	3.681000e+03	138.000000	0.000000	1.000000	1.00
25%	2.520313e+06	431.000000	6.000000	2.000000	3.00
50%	5.073096e+06	497.000000	12.000000	4.000000	6.00
75%	7.462026e+06	566.000000	17.000000	6.000000	9.00
max	9.999011e+06	864.000000	23.000000	7.000000	12.00

The above shows statistical analysis for various features.

Examples,We can see the longest flight delay of 250 minutes from the previous flight delay minutes and also highest and lowest Airline ratings amongst other observations

DATA DISTRIBUTION

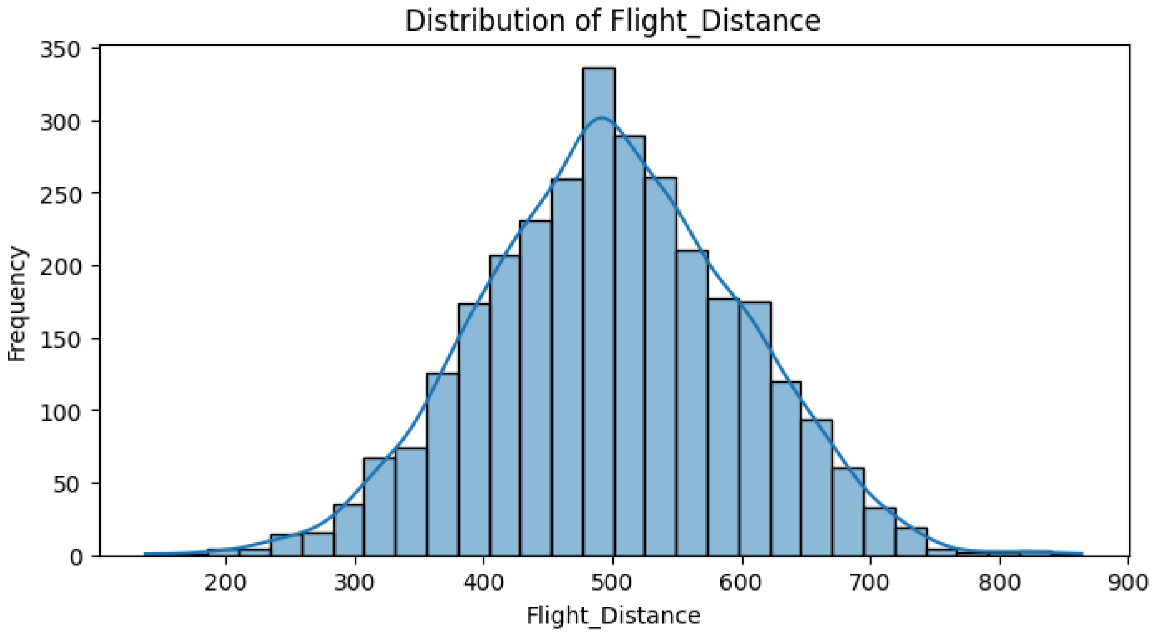
Numerical Columns

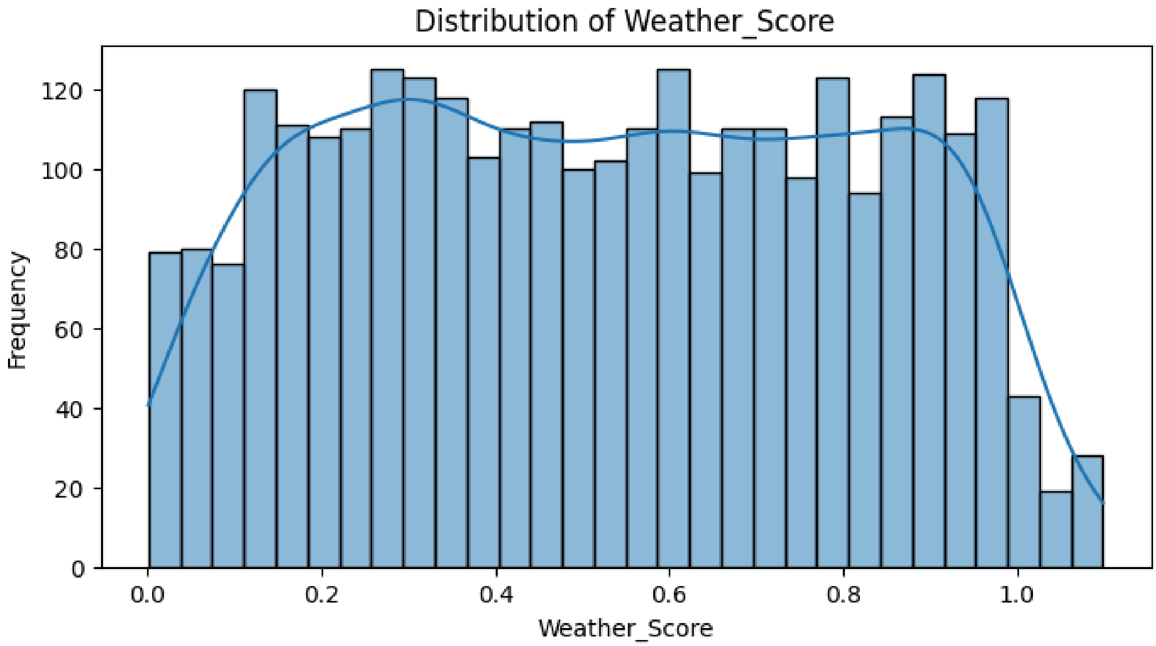
In [23]:
```
df.columns
```

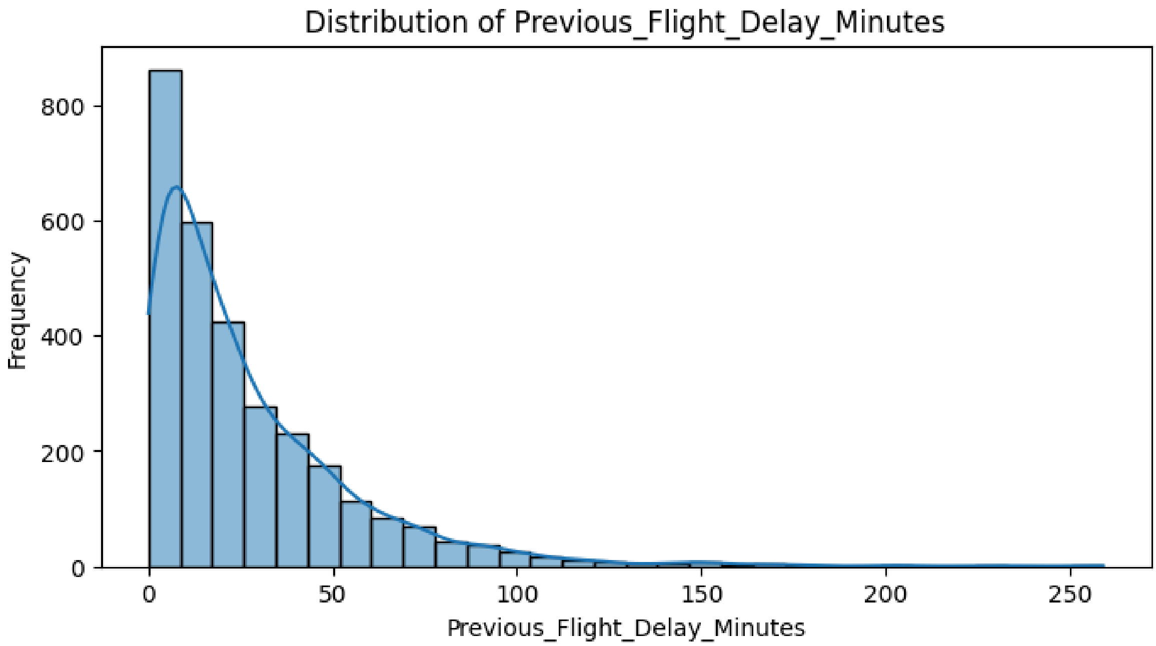
Out[23]:
```
Index(['Flight_ID', 'Airline', 'Flight_Distance', 'Origin_Airport',
       'Destination_Airport', 'Scheduled_Departure_Time', 'Day_of_Week',
       'Month', 'Airplane_Type', 'Weather_Score',
       'Previous_Flight_Delay_Minutes', 'Airline_Rating', 'Passenger_Load',
       'Flight_Cancelled', 'Capped_Flight_Distance',
       'Log_Previous_Flight_Delay_Minutes'],
      dtype='object')
```

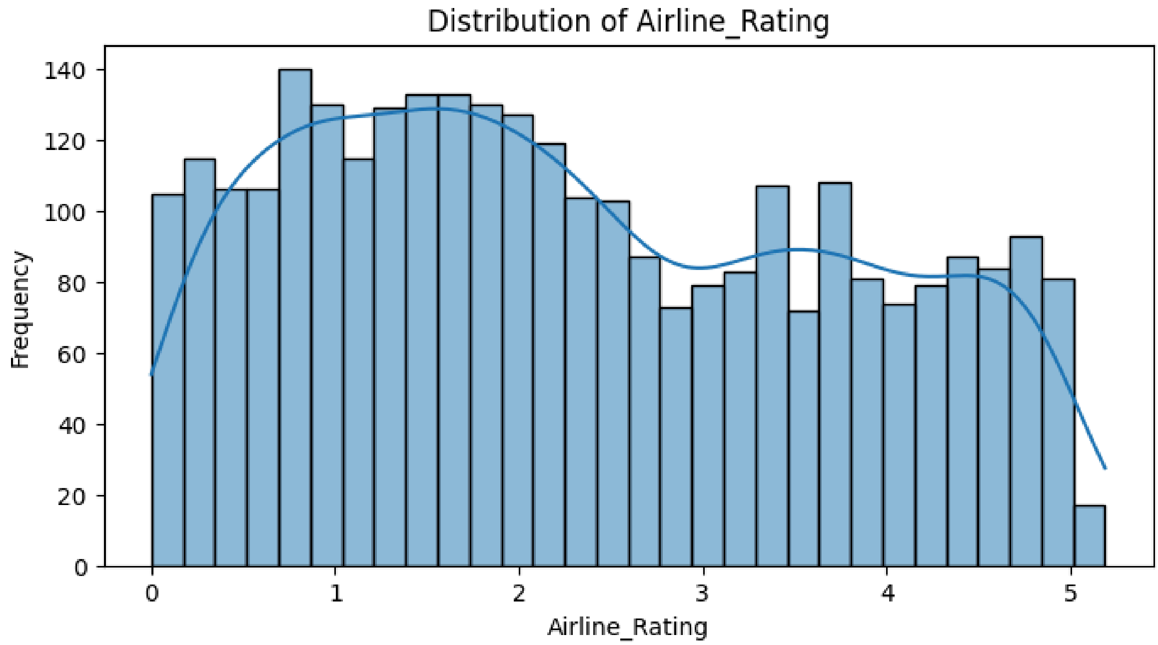
In [24]:
```
#Selecting relevant numerical columns
numerical_columns = [
    'Flight_Distance', 'Weather_Score', 'Previous_Flight_Delay_Minutes',
    'Airline_Rating', 'Passenger_Load', 'Capped_Flight_Distance',
    'Log_Previous_Flight_Delay_Minutes'
]
```

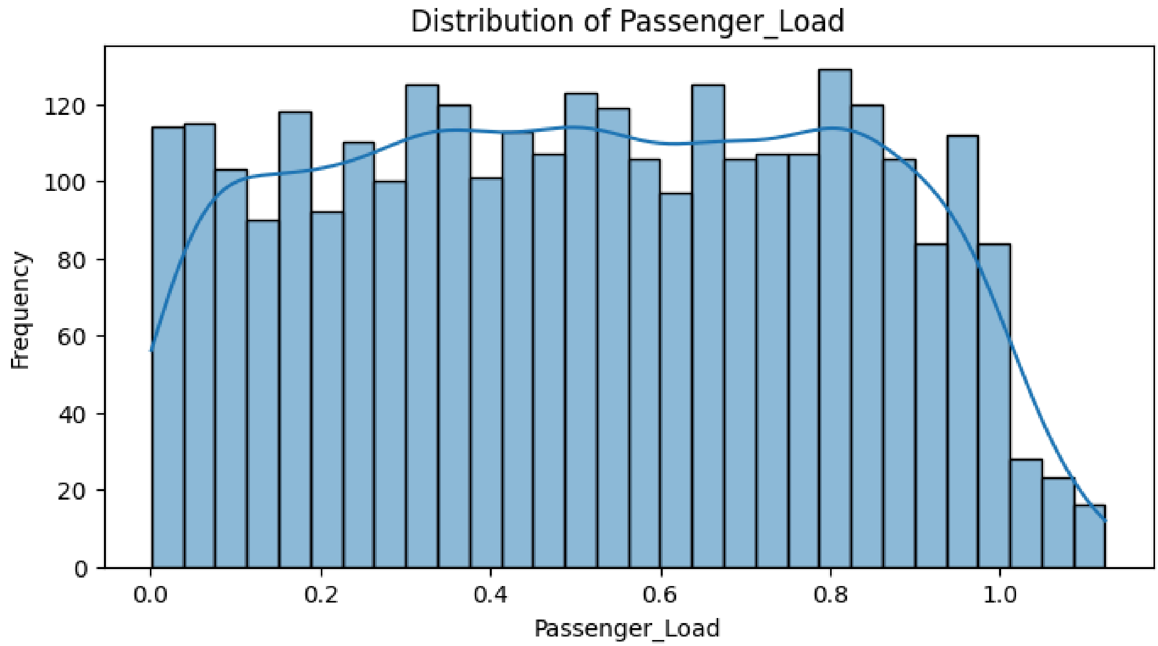
In [25]:
```
# Plotting the distribution for each numerical column
for col in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.histplot(df[col], kde=True, bins=30)  #
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.show()
```

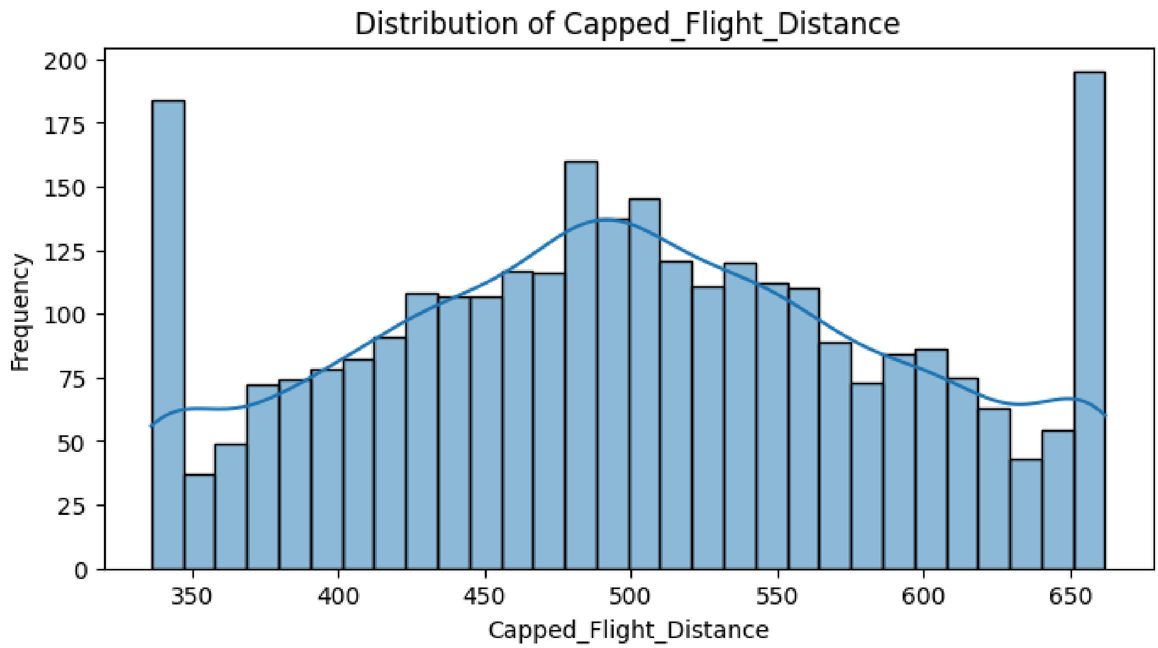

Distribution of Flight_Distance
Frequency
0
50
100
150
200
250
300
350
200
300
400
500
600
700
800
900
Flight_Distance

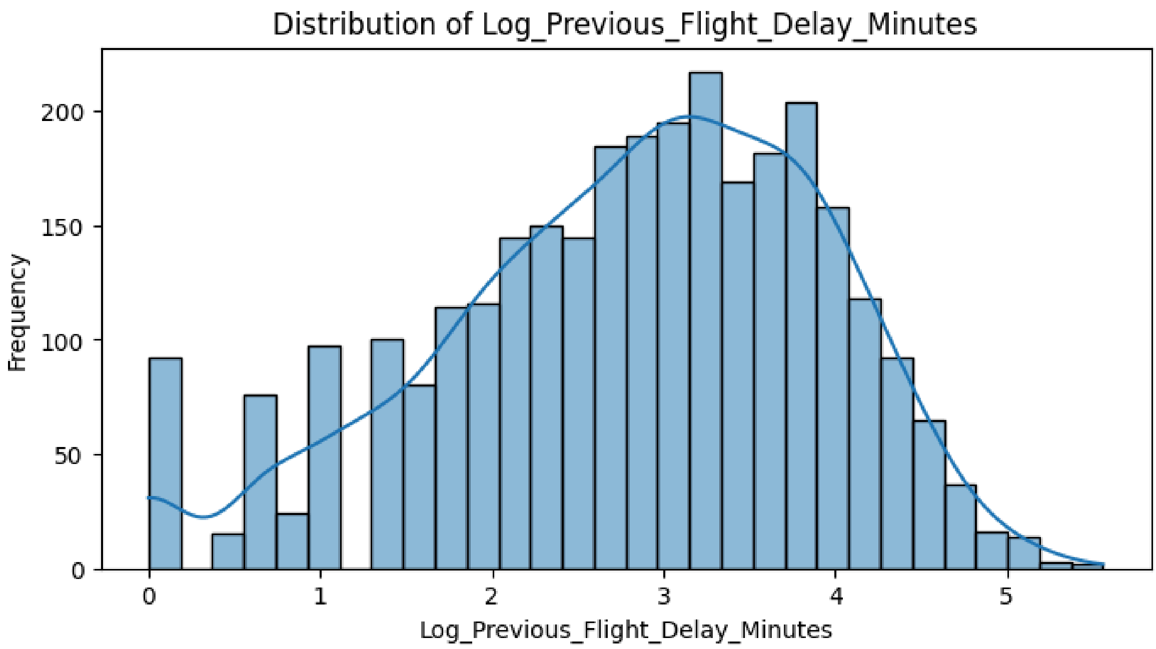

Distribution of Weather_Score
Frequency
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
Weather_Score


Distribution of Previous_Flight_Delay_Minutes
Frequency
0
200
400
600
800
0
50
100
150
200
250
Previous_Flight_Delay_Minutes


Distribution of Airline_Rating
Frequency
0
20
40
60
80
100
120
140
0
1
2
3
4
5
Airline_Rating


Distribution of Passenger_Load
Frequency
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
Passenger_Load


Distribution of Capped_Flight_Distance
Frequency
0
25
50
75
100
125
150
175
200
350
400
450
500
550
600
650
Capped_Flight_Distance

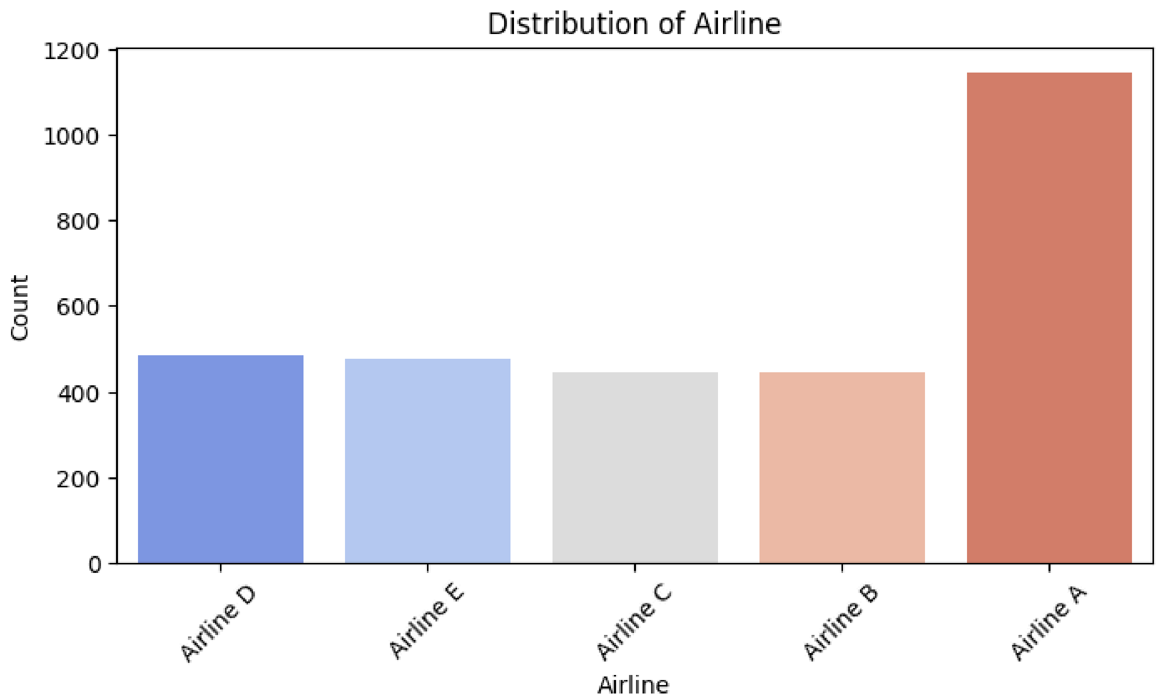


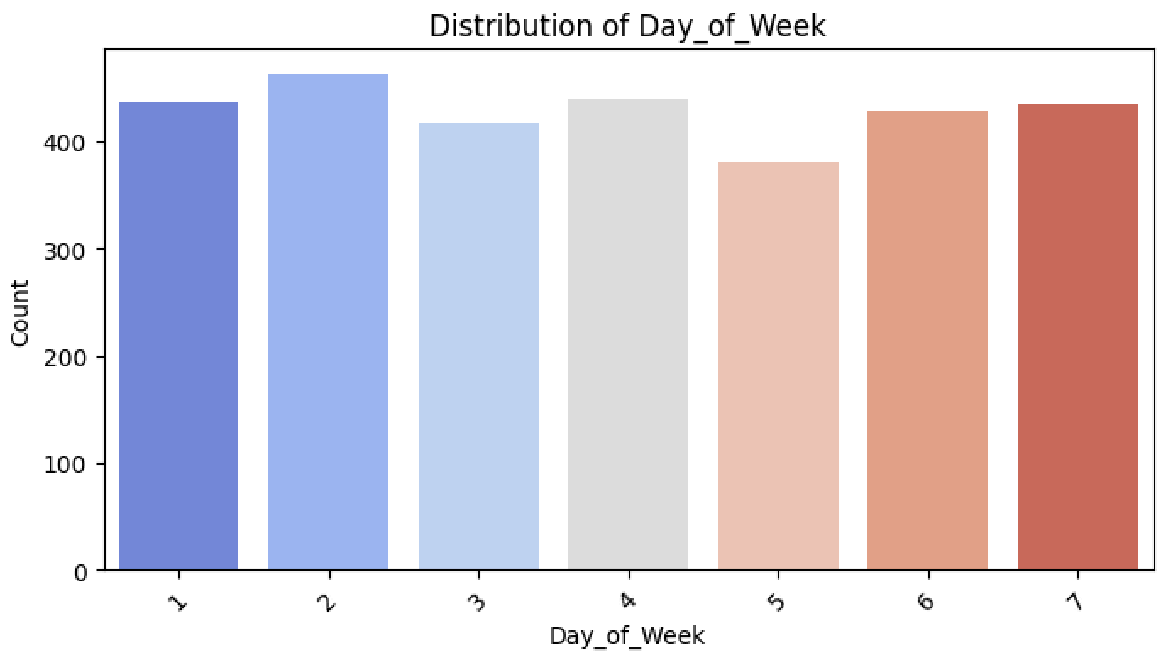
From this , we can see various distribution of the numerical columns

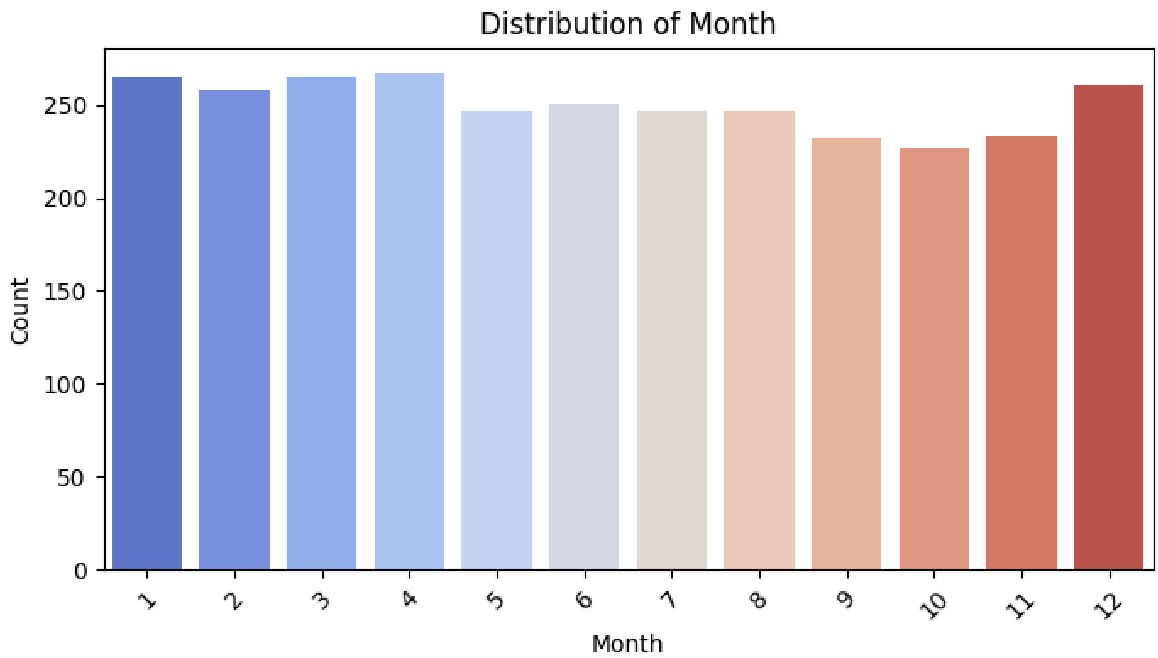
Now checking categorial columns

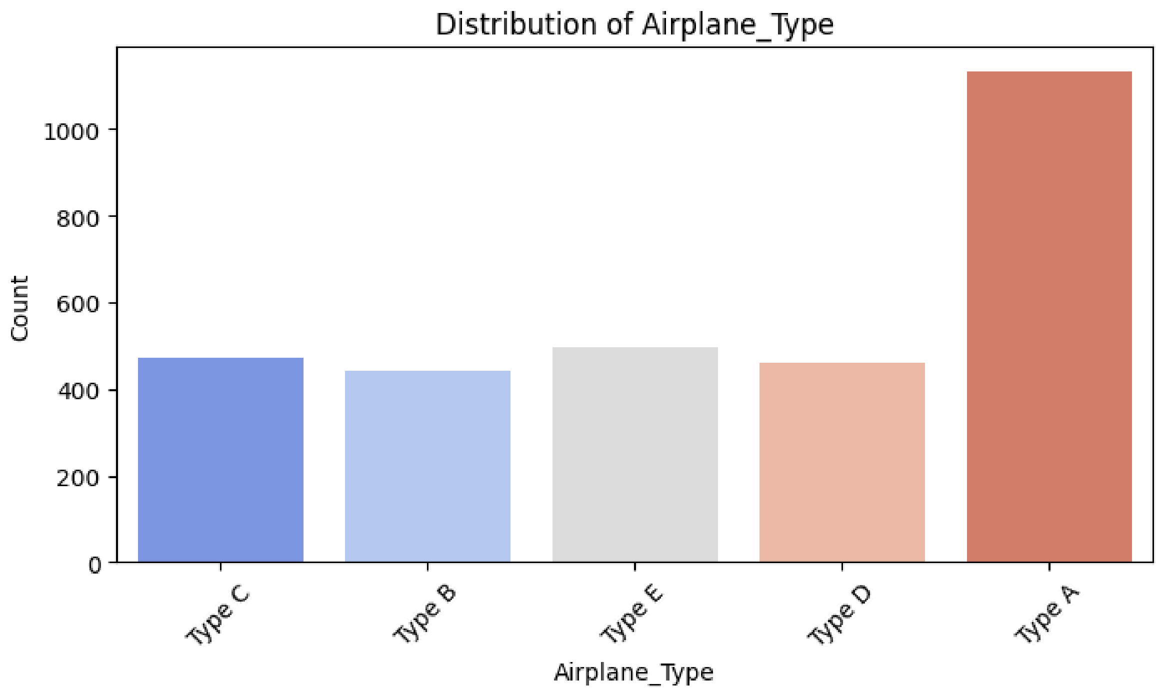
In [26]:

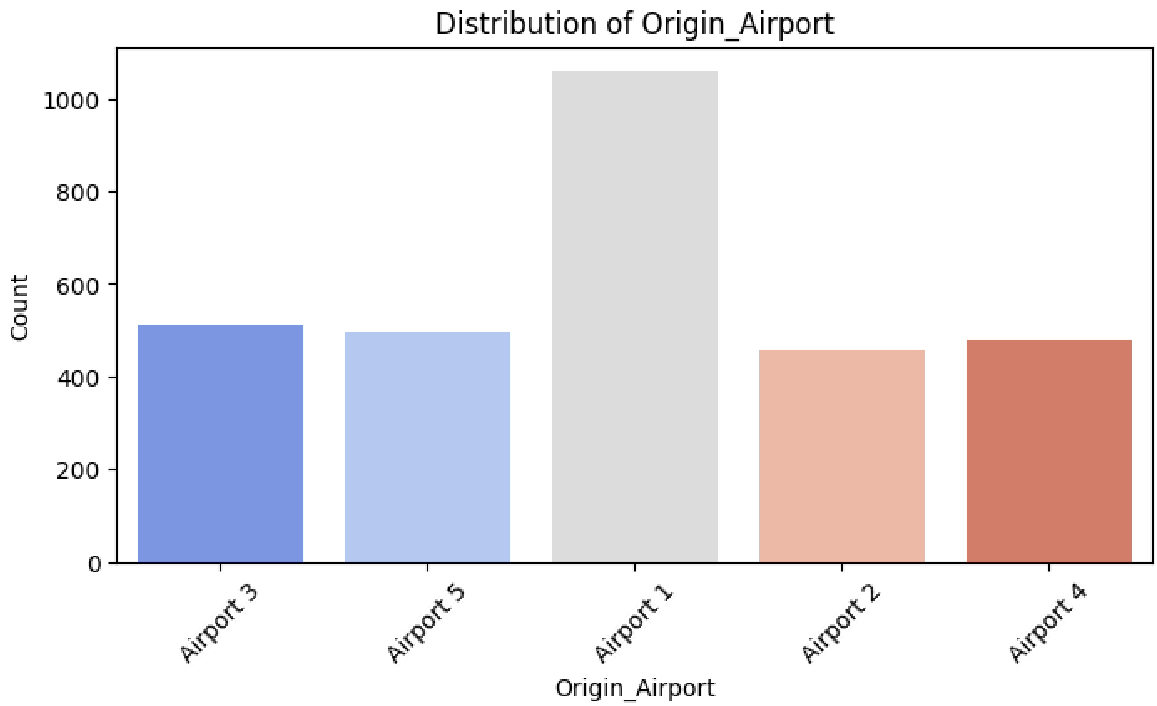
```
import warnings
warnings.filterwarnings("ignore")
#selecting columns to check
categorical_columns = ['Airline','Day_of_Week', 'Month', 'Airplane_Type','Origin

# Plotting the distribution for each categorical column
for col in categorical_columns:
    plt.figure(figsize=(8, 4))
    sns.countplot(x=df[col], palette='coolwarm')  # Countplot to show the freque
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Count')
    plt.xticks(rotation=45)  # Rotate x labels if needed for readability
    plt.show()
```

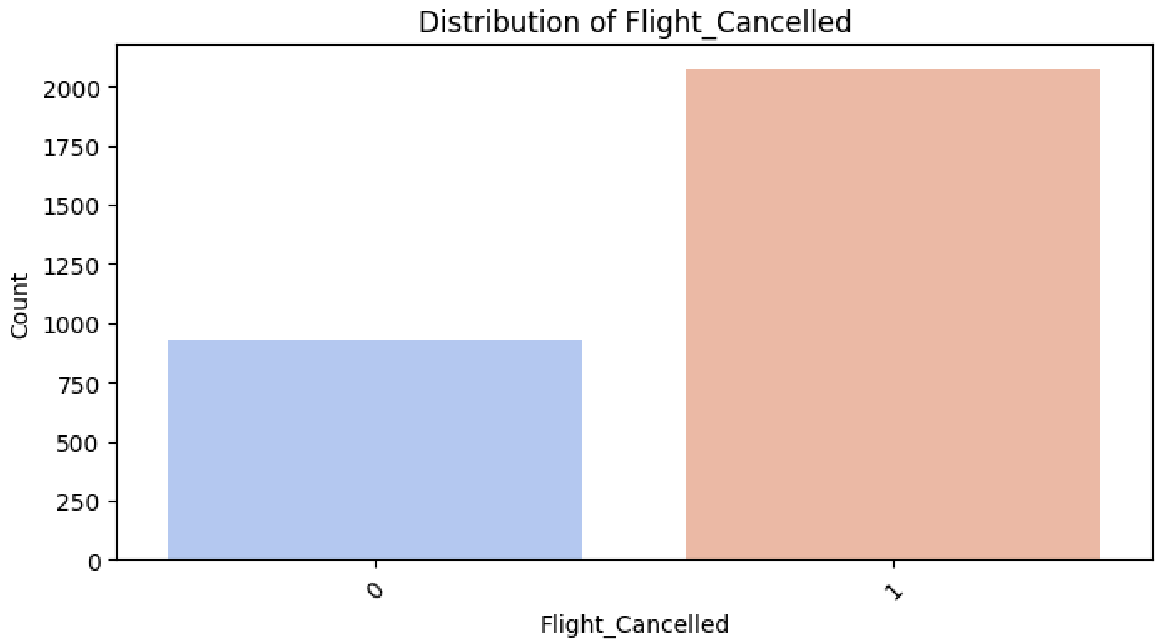

Distribution of Airline
Count
1200
1000
800
600
400
200
0
Airline D
Airline E
Airline C
Airline B
Airline A
Airline


Distribution of Day_of_Week
Count
400
300
200
100
0
1
2
3
4
5
6
7
Day_of_Week


Distribution of Month
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
Month


Distribution of Airplane_Type
Count
1000
800
600
400
200
0
Type C
Type B
Type E
Type D
Type A
Airplane_Type





Data from Flight_Cancelled target column is highly imbalanced, we have more occurances of cancelled flights than non cancelled.

There is also a noticeable high occurance of :

TypeA flight than other flights

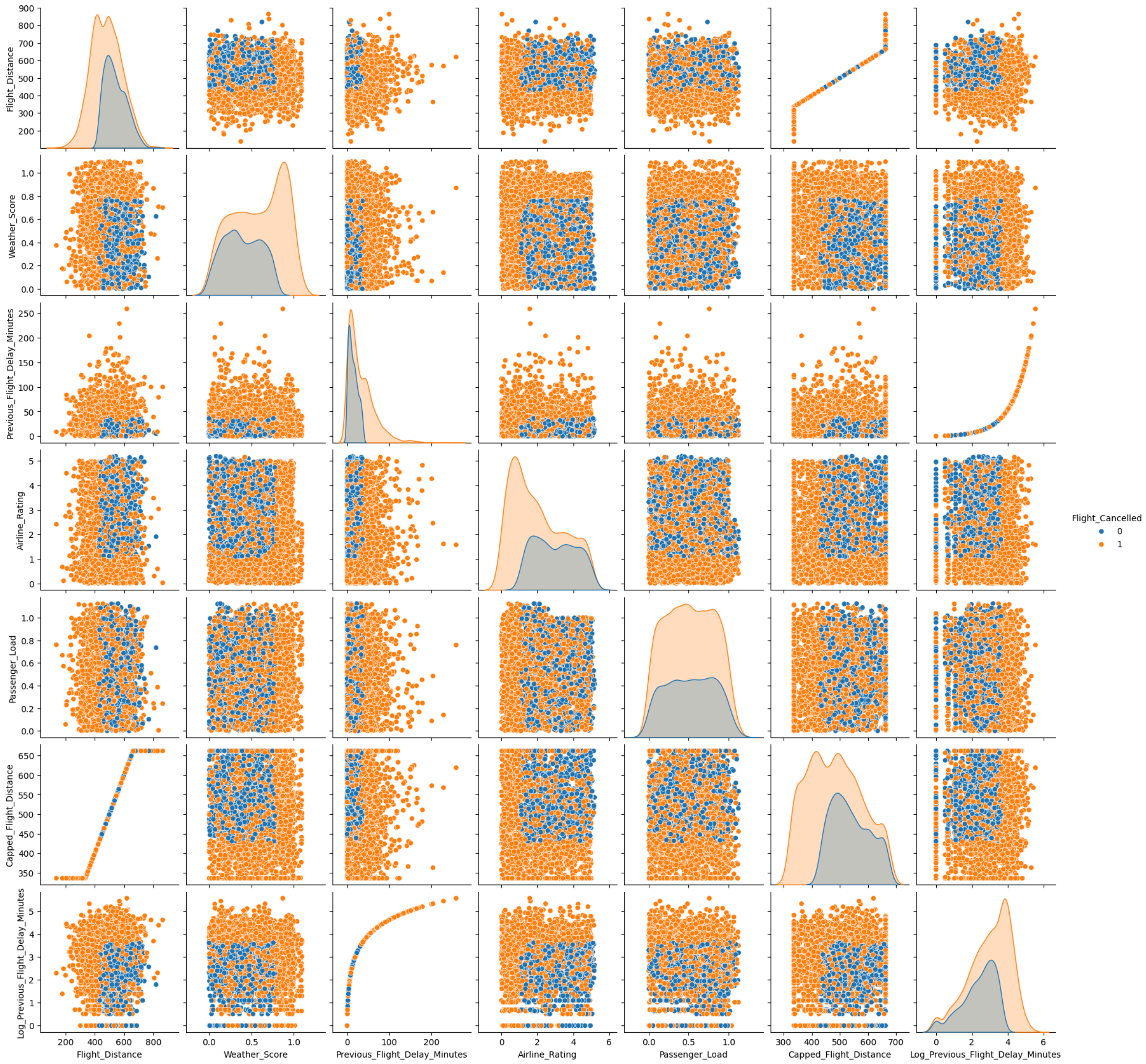
AirlineA than other airlines

Airport1 than other airports

Month and day of the week: there is slight difference between various months and various days of the week

RELATIONSHIP BETWEEN FEATURES:

In [27]:
```
#pairplots
sns.pairplot(df, hue ='Flight_Cancelled', vars=numerical_columns)
plt.show()
```



In [28]:
```
#Correlation matrix to understand the relationships better
df.select_dtypes(include ="number").corr()
```

Out[28]:

	Flight_ID	Flight_Distance	Scheduled_Departure_Time
Flight_ID	1.000000	-0.007541	0.006207
Flight_Distance	-0.007541	1.000000	0.039727
Scheduled_Departure_Time	0.006207	0.039727	1.000000
Day_of_Week	-0.012384	0.024455	-0.011834
Month	-0.025743	0.019573	0.018319
Weather_Score	-0.002007	0.010139	-0.023682
Previous_Flight_Delay_Minutes	0.006172	0.018413	-0.036318
Airline_Rating	0.043170	0.042128	0.040739
Passenger_Load	0.009312	-0.018627	0.046556
Flight_Cancelled	-0.009101	-0.277471	-0.043733
Capped_Flight_Distance	-0.007407	0.988170	0.033209
Log_Previous_Flight_Delay_Minutes	0.011250	0.008974	-0.032594

In [29]:

```
#Visualising the relationships using Heatmap
plt.figure(figsize =(12,10))
sns.heatmap(df.select_dtypes(include ="number").corr(), annot = True)
```

Out[29]: <Axes: >

Relationship Insights:

Flight_Distance & Capped_Flight_Distance are strongly correlated since"Capped_Flight_Distance" is derived from "Flight_Distance" while managing outliers.

Similarly for Previous_Flight_Delay_Minutes & Log_Previous_Flight_Delay_Minutes (0.827)

Flight_Cancelled & Weather_Score (0.306): There's a moderate positive correlation indicating that poor weather may be associated with more cancellations.

Flight_Cancelled & Previous_Flight_Delay_Minutes (0.303): A moderate positive correlation suggesting that flights with previous delays might have a higher chance of being cancelled.

Scheduled_Departure_Time & Passenger_Load (0.047): Slight positive correlation but not strong.

Flight_Distance & Passenger_Load (-0.0186): A weak negative correlation

Airline_Rating & Previous_Flight_Delay_Minutes (-0.0360): A weak negative correlation, indicating that higher delays are not strongly related to airline ratings.

Airline_Rating & Flight_Cancelled (-0.314): A moderate negative correlation, indicating that lower airline ratings might be associated with higher chances of cancellation.

Day_of_Week, Month: weak correlations with other variables, indicating that the day of the week and month might not have strong impacts on flight cancellations.

In [ ]:

RELATIONSHIP BETWEEN FEATURES AND TARGET COLUMN

Based on the investigations from the Correlation matrix and Heatmap above, below are the observations:

Flight_Distance - there is a moderate negative correlation between Flight_Distance and Flight_Cancelled, suggesting flight distance might not be highly influencial

Scheduled_Departure_Time , correlation is very close to zero indicating weak correlation with Flight_Cancelled, therefore might not be a significant predictor for flight cancellations

Day_of_Week and Month, also shows weak correlation with the target column , suggesting not much impact on flight cancellations

Weather_Score, shows moderate postive correlation with target column indicating worse weather conditions can influence flight cancellations, this is very important for predicting cancellations

Previous_Flight_Minutes_Minutes, there is moderate postive correlation suggesting previous delays could influence cancellations

Airline_Rating, shows moderate negative correlation with the target, it indicate flights with lower airline rating are likely to be cancelled

Passenger_Load, weak correlation indicating number of passengers not impacting flight cancellations

Capped_Flight_Distance, similar to original Flight_Distance, moderate negative correlation

Log_Transformed_Delay_Minutes, also similar to the original Previous_Flight_Delay_Minutes although slitly lower than orignal but it still indicates delays are positively associated with cancellations.

TASK 3 : DATA PRE-PROCESSING AND MODEL BUILDING

##Dealing cyclical data

In [30]:

```
#Cyclical data Month and Day of Week
df['Day_of_Week_sin'] =np.sin(2*np.pi *df['Day_of_Week']/7)
df['Day_of_Week_cos'] =np.cos(2*np.pi *df['Day_of_Week']/7)
df['Months_sin'] = np.sin(2*np.pi*df['Month'] /12)
df['Months_cos'] = np.cos(2*np.pi*df['Month'] /12)
```

```
df.drop(['Day_of_Week','Month'], axis =1).head()
```

Out[30]:

	Flight_ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	Scheduled_De
0	7319483	Airline D	475	Airport 3	Airport 2	
1	4791965	Airline E	538	Airport 5	Airport 4	
2	2991718	Airline C	565	Airport 1	Airport 2	
3	4220106	Airline E	658	Airport 5	Airport 3	
4	2263008	Airline E	566	Airport 2	Airport 2	

##Encoding categorical variables:

In [31]:
```
#One hot encoding for categorial variables
categorical_columns =['Airline', 'Origin_Airport', 'Destination_Airport', 'Airpl
df_encoded =pd.get_dummies(df, columns =categorical_columns, drop_first =True)
```

In [50]:
```
df_encoded.head()
```

Out[50]:

	Flight_ID	Flight_Distance	Scheduled_Departure_Time	Day_of_Week	Month	Weat
0	7319483	-0.241812	-1.077826	1.010411	-1.549203	
1	4791965	0.395351	0.081906	-1.469735	-0.109691	
2	2991718	0.668421	0.806738	-0.477676	0.754016	
3	4220106	1.608995	-1.512725	-1.469735	0.466114	
4	2263008	0.678535	1.096671	1.506441	1.617723	

5 rows × 31 columns

##Feature Scaling:

In [37]:
```
#Numerical columns to scale
numerical_features = [ 'Weather_Score','Scheduled_Departure_Time', 'Airline_Rati
       'Passenger_Load', 'Capped_Flight_Distance',
       'Log_Previous_Flight_Delay_Minutes','Flight_Distance','Previous_Flight_De

#apply standard scaler
from sklearn.preprocessing import StandardScaler
#initialise standard scaler
scaler =StandardScaler()

#apply scaler
df_encoded[numerical_features]= scaler.fit_transform(df_encoded[numerical_featur
```

In [38]:
```
df_encoded.head()
```

Out[38]:

	Flight_ID	Flight_Distance	Scheduled_Departure_Time	Day_of_Week	Month	Weat
0	7319483	-0.241812	-1.077826	1.010411	-1.549203	
1	4791965	0.395351	0.081906	-1.469735	-0.109691	
2	2991718	0.668421	0.806738	-0.477676	0.754016	
3	4220106	1.608995	-1.512725	-1.469735	0.466114	
4	2263008	0.678535	1.096671	1.506441	1.617723	

5 rows × 31 columns

##Model Building:

In [40]:
```
#splitting the data
from sklearn.model_selection import train_test_split

#Dropping original columns (to use only the transformed ones) and also target co
#
X =df_encoded.drop(['Flight_Cancelled', 'Flight_ID','Flight_Distance',"Month","D

#Target variable
y = df_encoded['Flight_Cancelled']

#Splitting the data
X_train, X_test, y_train,y_test = train_test_split(X,y,test_size =0.2, random_st
```

In [41]:
```
X.head(3)
```

Out[41]:

	Scheduled_Departure_Time	Weather_Score	Airline_Rating	Passenger_Load	Capped_F
0	-1.077826	-1.028402	-0.115698	-0.130868	
1	0.081906	-1.595333	-0.501109	-1.204954	
2	0.806738	-1.479818	1.460975	-0.876504	

3 rows × 25 columns

In [51]:
```python
#Training the Logistic Regression Model
from sklearn.linear_model import LogisticRegression
#initialize the model
model = LogisticRegression()
#Train the model
model.fit(X_train,y_train)
```

Out[51]:
```
▾ LogisticRegression
LogisticRegression()
```

In [43]:
```python
#making predictions on test set
y_pred = model.predict(X_test)
y_proba = model.predict_proba(X_test)[:, 1]
```

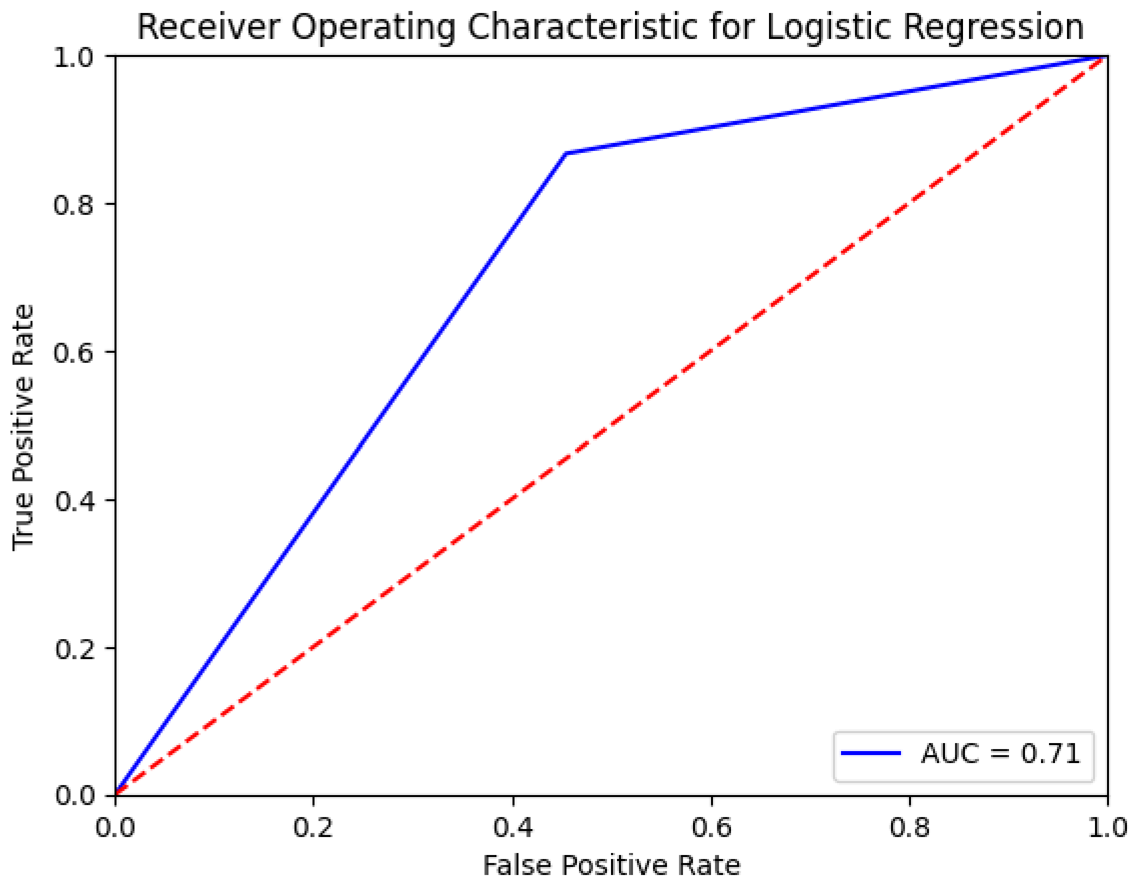
##Model Evaluation

In [52]:
```python
#Evaluate the model
from sklearn.metrics import accuracy_score,precision_score,recall_score, confusi
#Accuracy
accuracy = accuracy_score(y_test,y_pred)
print(f'Accuracy: {accuracy: 2f}')

#precision,Recall, F1-score
print(classification_report(y_test,y_pred))

#confusion matrix
conf_matrix = confusion_matrix(y_test,y_pred)
print('Confusion Matrix: ')
print(conf_matrix)
```

```
Accuracy:  0.766667
              precision    recall  f1-score   support

           0       0.65      0.55      0.59       187
           1       0.81      0.87      0.84       413

    accuracy                           0.77       600
   macro avg       0.73      0.71      0.71       600
weighted avg       0.76      0.77      0.76       600

Confusion Matrix: 
[[102  85]
 [ 55 358]]
```

In [47]:
```
import sklearn.metrics as metrics
fpr, tpr, threshold = metrics.roc_curve(y_test, y_pred)
print(fpr)
print(tpr)
print(threshold)
roc_auc = metrics.auc(fpr, tpr)
print(roc_auc)

# method I: plt
plt.title('Receiver Operating Characteristic for Logistic Regression')
plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

```
[0.         0.45454545 1.        ]
[0.         0.86682809 1.        ]
[inf  1.  0.]
0.7061413163108079
```



Our model achived an accuracy of 77%, we can try other classification methods to see how they perform

In [ ]:

In [ ]: