Untitled8

May 31, 2024

```
[71]: #importing the needed libraries
      import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sns
```

```
[2]: #LOADING THE DATASET
     cancelled_3= pd.read_excel('Flyzy Flight Cancellation (3).xlsx')
```

```
[3]: #EXPLORING THE DATA
     cancelled_3.head()
```

```
[3]:    Flight ID    Airline  Flight_Distance Origin_Airport Destination_Airport  \
     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

        Scheduled_Departure_Time  Day_of_Week  Month Airplane_Type  Weather_Score  \
     0                         4            6      1        Type C       0.225122
     1                        12            1      6        Type B       0.060346
     2                        17            3      9        Type C       0.093920
     3                         1            1      8        Type B       0.656750
     4                        19            7     12        Type E       0.505211

        Previous_Flight_Delay_Minutes  Airline_Rating  Passenger_Load  \
     0                            5.0        2.151974        0.477202
     1                           68.0        1.600779        0.159718
     2                           18.0        4.406848        0.256803
     3                           13.0        0.998757        0.504077
     4                            4.0        3.806206        0.019638

        Flight_Cancelled
     0                 0
     1                 1
     2                 0
```

```
3                 1
4                 0
```

[4]:

```
cancelled_3.tail()
```

[4]:

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

     Destination_Airport  Scheduled_Departure_Time  Day_of_Week  Month  \
2995           Airport 3                         0            6      1
2996           Airport 4                        22            4      7
2997           Airport 3                         8            3      1
2998           Airport 3                         5            5      3
2999           Airport 2                         1            1     10

     Airplane_Type  Weather_Score  Previous_Flight_Delay_Minutes  \
2995        Type B       0.190018                        1.00000
2996        Type E       0.719271                       91.00000
2997        Type B       0.458724                        3.00000
2998        Type E       0.443373                       46.00000
2999        Type A       0.704563                       18.66667

      Airline_Rating  Passenger_Load  Flight_Cancelled
2995        2.451216        0.283440                 1
2996        0.027039        0.665294                 1
2997        1.131315        0.991307                 0
2998        0.968651        0.254808                 1
2999        1.879411        0.532486                 1
```

[5]:

```
cancelled_3.columns
```

[5]:

```
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')
```

[6]:

```
cancelled_3.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: 328.2+ KB
```

[8]:

```
#converting all null values to numeric
from sklearn.preprocessing import LabelEncoder
label_encoder=LabelEncoder()
```

[9]:

```
cancelled_3['Airline']=label_encoder.fit_transform(cancelled_3['Airline'])
cancelled_3['Origin_Airport']=label_encoder.
  ↪fit_transform(cancelled_3['Origin_Airport'])
cancelled_3['Destination_Airport']=label_encoder.
  ↪fit_transform(cancelled_3['Destination_Airport'])
cancelled_3['Airplane_Type']=label_encoder.
  ↪fit_transform(cancelled_3['Airplane_Type'])
```

[12]:

```
#all null values are now numeric
cancelled_3.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   int64
 2   Flight_Distance              3000 non-null   int64
 3   Origin_Airport               3000 non-null   int64
 4   Destination_Airport          3000 non-null   int64
 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   int64
 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(10)
memory usage: 328.2 KB
```

[13]:
```
cancelled_3.isnull().sum()
```

[13]:
```
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
```

[14]:
```
cancelled_3.describe()
```

[14]:
```
          Flight ID      Airline  Flight_Distance  Origin_Airport  \
count  3.000000e+03  3000.000000      3000.000000     3000.000000
mean   4.997429e+06     1.567333       498.909333        1.631667
std    2.868139e+06     1.513350        98.892266        1.499805
min    3.681000e+03     0.000000       138.000000        0.000000
25%    2.520313e+06     0.000000       431.000000        0.000000
50%    5.073096e+06     1.000000       497.000000        1.000000
75%    7.462026e+06     3.000000       566.000000        3.000000
max    9.999011e+06     4.000000       864.000000        4.000000

       Destination_Airport  Scheduled_Departure_Time  Day_of_Week  \
count          3000.000000               3000.000000  3000.000000
mean              0.911667                 11.435000     3.963000
std               1.147012                  6.899298     2.016346
min               0.000000                  0.000000     1.000000
25%               0.000000                  6.000000     2.000000
50%               0.000000                 12.000000     4.000000
75%               2.000000                 17.000000     6.000000
max               3.000000                 23.000000     7.000000
```

```
             Month  Airplane_Type  Weather_Score  \
count  3000.000000    3000.000000    3000.000000
mean      6.381000       1.582000       0.524023
std       3.473979       1.515049       0.290694
min       1.000000       0.000000       0.000965
25%       3.000000       0.000000       0.278011
50%       6.000000       1.000000       0.522180
75%       9.000000       3.000000       0.776323
max      12.000000       4.000000       1.099246

       Previous_Flight_Delay_Minutes  Airline_Rating  Passenger_Load  \
count                    3000.000000     3000.000000     3000.000000
mean                       26.793383        2.317439        0.515885
std                        27.874733        1.430386        0.295634
min                         0.000000        0.000103        0.001039
25%                         7.000000        1.092902        0.265793
50%                        18.000000        2.126614        0.517175
75%                        38.000000        3.525746        0.770370
max                       259.000000        5.189038        1.123559

       Flight_Cancelled
count       3000.000000
mean           0.690667
std            0.462296
min            0.000000
25%            0.000000
50%            1.000000
75%            1.000000
max            1.000000
```

[15]:

```
cancelled_3.shape
```

[15]:

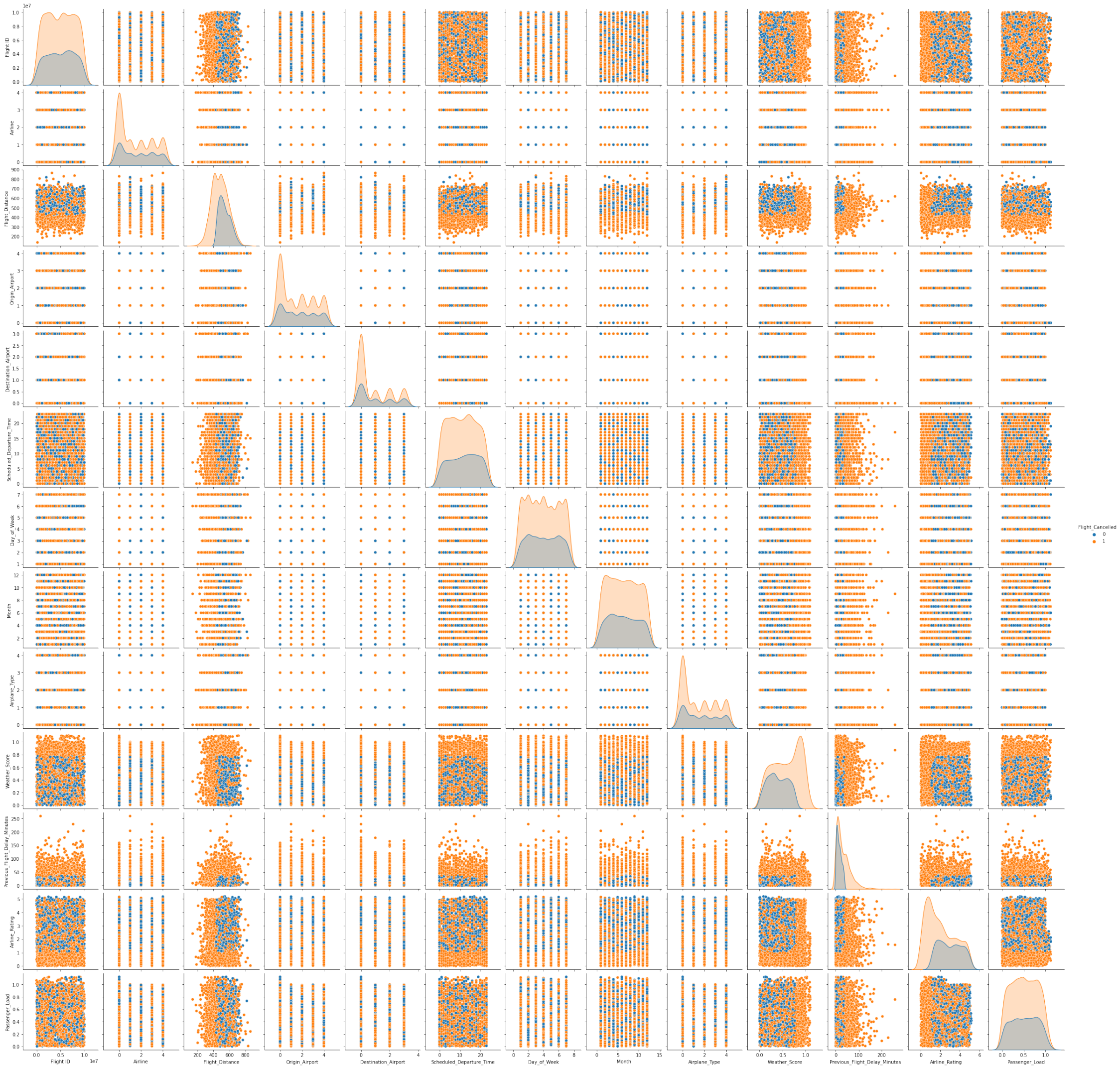
```
(3000, 14)
```

[16]:

```
sns.pairplot(cancelled_3,hue='Flight_Cancelled')
```

[16]:

```
<seaborn.axisgrid.PairGrid at 0x7f00e9db6c50>
```



[28]:

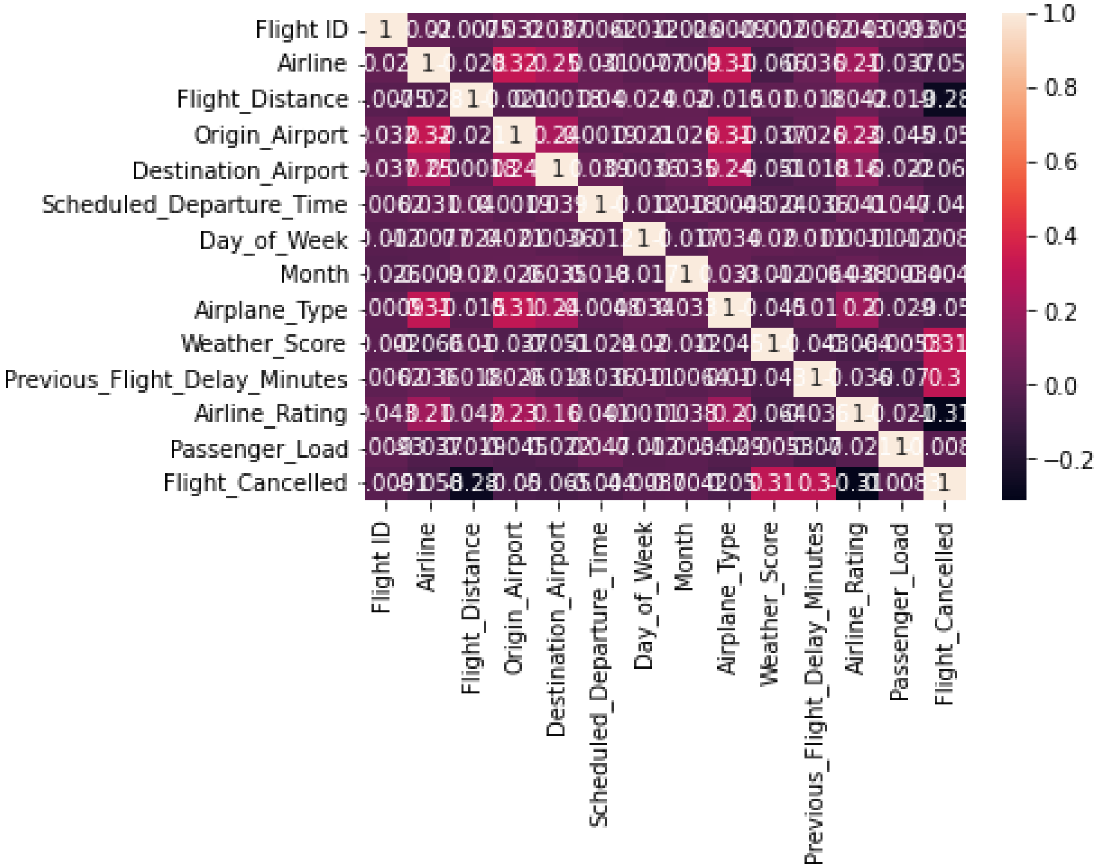
```
#Relationship analysis
corelation=cancelled_3.corr()
```

[31]:

```
sns.heatmap(corelation, xticklabels=corelation.columns, yticklabels=corelation.
columns
,annot=True)
```

[31]:

```
<AxesSubplot: >
```



[55]:
```
x=cancelled_3[['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']]
y=cancelled_3['Flight_Cancelled']
```

[34]:
```
correlation=np.corrcoef(x['Flight ID'], y)[0, 1]
print("Correlation coefficient:", correlation)
```

```
Correlation coefficient: -0.0091005441027218
```

[35]:
```
correlation=np.corrcoef(x['Airline'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.0579152207425434
```

[36]:
```
correlation=np.corrcoef(x['Origin_Airport'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.049925451318859365
```

[37]:

```python
correlation=np.corrcoef(x['Destination_Airport'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.06475308532828986
```

[38]:

```python
correlation=np.corrcoef(x['Destination_Airport'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.06475308532828986
```

[39]:

```python
correlation=np.corrcoef(x['Scheduled_Departure_Time'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.043732799217209566
```

[40]:

```python
correlation=np.corrcoef(x['Day_of_Week'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.008705376908751991
```

[41]:

```python
correlation=np.corrcoef(x['Month'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.004242162010093053
```

[42]:

```python
correlation=np.corrcoef(x['Airplane_Type'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.04994233077062403
```

[43]:

```python
correlation=np.corrcoef(x['Weather_Score'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: 0.30576162505311555
```

[44]:

```python
correlation=np.corrcoef(x['Previous_Flight_Delay_Minutes'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: 0.302804640547877
```

[45]:

```python
correlation=np.corrcoef(x['Airline_Rating'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.31409863751544576
```

[46]:

```python
correlation=np.corrcoef(x['Passenger_Load'],y)[0,1]
print("Correlation coefficient:",correlation)
```

```
Correlation coefficient: -0.008319756091970014
```

[47]:
```
from sklearn.linear_model import LinearRegression
```

[48]:
```
model=LinearRegression()
model.fit(x, y)
print("Intercept:", model.intercept_)
print("Coefficients:", model.coef_)
```

```
Intercept: 1.1322946462844234
Coefficients: [-7.78946167e-11 -1.00440931e-03 -1.28197407e-03  3.22053796e-03
 -1.33395687e-03 -2.33343775e-04 -2.62244908e-03  2.19967335e-03
  3.11880877e-03  4.85104135e-01  5.16096743e-03 -8.89749801e-02
  7.14363056e-03]
```

[49]:
```
#EVALUATION  R-SQUARED.
from sklearn.metrics import r2_score
```

[50]:
```
y_pred=model.predict(x)
r2=r2_score(y, y_pred)
print("R-squared:", r2)
```

```
R-squared: 0.3495804329516544
```

[51]:
```
dataset_2=pd.read_excel("Flyzy Flight Cancellation.xlsx")
```

[62]:
```
x=cancelled_3.drop(columns=['Flight ID', 'Airline','Flight_Distance',␣
 ↪'Origin_Airport','Origin_Airport',␣
 ↪'Destination_Airport','Scheduled_Departure_Time','Day_of_Week',
'Month', 'Airplane_Type', 'Weather_Score',␣
 ↪'Previous_Flight_Delay_Minutes','Airline_Rating','Passenger_Load'])
y=cancelled_3['Flight_Cancelled']
```

[73]:
```
#model feeding.
model=LinearRegression()
model.fit(x_train, y_train)
```

[73]:
```
LinearRegression()
```

[74]:
```
y_pred=model.predict(x_test)
```

[75]:
```
r_squared=r2_score(y_test, y_pred)
print("R-squared:", r_squared)
```

```
R-squared: 1.0
```