Data Checking

May 24, 2024

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

[2]:
```
#loading the dataset
df= pd.read_excel('Flyzy Flight Cancellation.xlsx')
```

[3]:
```
#EXPLORING THE DATA
#finding the shape ofthe dataset
df.shape
```

[3]:
```
(3000, 14)
```

[4]:
```
#head of the data
df.head()
```

[4]:

	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
```

[5]:
```
#the tail of the dataset
df.tail()
```

[5]:
```
      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
```

[6]:
```
#getting the information of the dataset
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: 328.2+ KB
```

[9]:
```
#identifying and dropping null values and finding missing values
df.isnull().sum()
```

[9]:
```
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
```

[8]:
```
#describing the dataset
df.describe()
```

[8]:

	Flight ID	Flight_Distance	Scheduled_Departure_Time	Day_of_Week	\
count	3.000000e+03	3000.000000	3000.000000	3000.000000	
mean	4.997429e+06	498.909333	11.435000	3.963000	
std	2.868139e+06	98.892266	6.899298	2.016346	
min	3.681000e+03	138.000000	0.000000	1.000000	
25%	2.520313e+06	431.000000	6.000000	2.000000	
50%	5.073096e+06	497.000000	12.000000	4.000000	

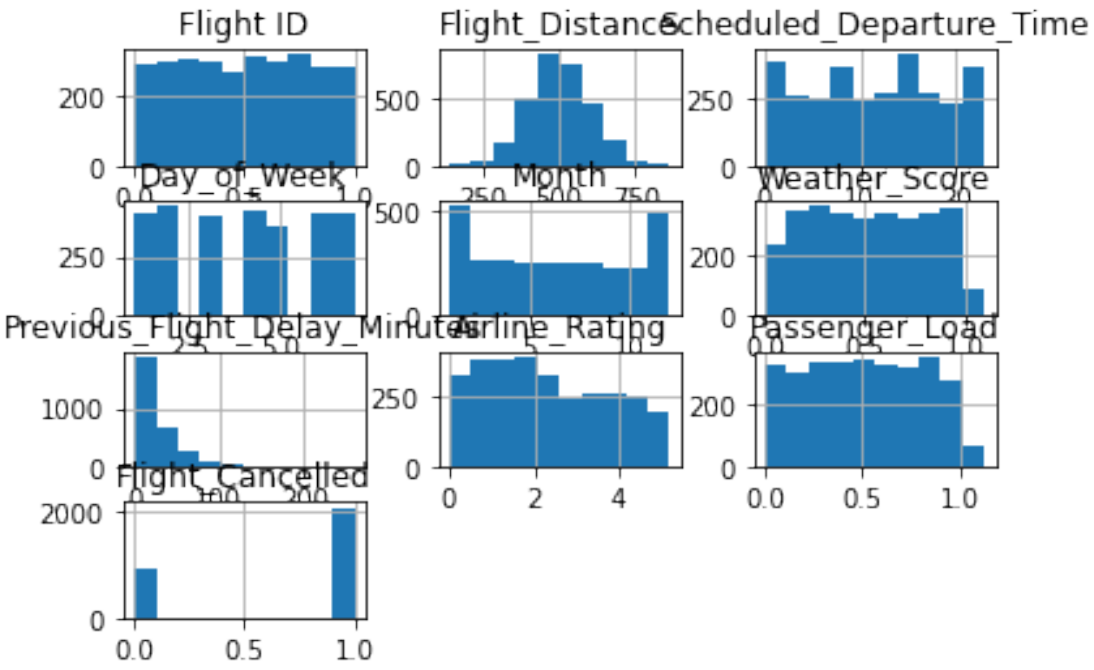
```
75%    7.462026e+06       566.000000                 17.000000     6.000000
max    9.999011e+06       864.000000                 23.000000     7.000000

             Month  Weather_Score  Previous_Flight_Delay_Minutes  \
count  3000.000000    3000.000000                    3000.000000
mean      6.381000       0.524023                      26.793383
std       3.473979       0.290694                      27.874733
min       1.000000       0.000965                       0.000000
25%       3.000000       0.278011                       7.000000
50%       6.000000       0.522180                      18.000000
75%       9.000000       0.776323                      38.000000
max      12.000000       1.099246                     259.000000

       Airline_Rating  Passenger_Load  Flight_Cancelled
count     3000.000000     3000.000000       3000.000000
mean         2.317439        0.515885          0.690667
std          1.430386        0.295634          0.462296
min          0.000103        0.001039          0.000000
25%          1.092902        0.265793          0.000000
50%          2.126614        0.517175          1.000000
75%          3.525746        0.770370          1.000000
max          5.189038        1.123559          1.000000
```

[10]:
```
#identifying the trend of the data using histogeams
df.hist()
```

[10]:
```
array([[<AxesSubplot: title={'center': 'Flight ID'}>,
        <AxesSubplot: title={'center': 'Flight_Distance'}>,
        <AxesSubplot: title={'center': 'Scheduled_Departure_Time'}>],
       [<AxesSubplot: title={'center': 'Day_of_Week'}>,
        <AxesSubplot: title={'center': 'Month'}>,
        <AxesSubplot: title={'center': 'Weather_Score'}>],
       [<AxesSubplot: title={'center': 'Previous_Flight_Delay_Minutes'}>,
        <AxesSubplot: title={'center': 'Airline_Rating'}>,
        <AxesSubplot: title={'center': 'Passenger_Load'}>],
       [<AxesSubplot: title={'center': 'Flight_Cancelled'}>,
        <AxesSubplot: >, <AxesSubplot: >]], dtype=object)
```


Flight ID
Flight_Distance
Scheduled_Departure_Time
Day_of_Week
Month
Weather_Score
Previous_Flight_Delay_Minutes
Airline_Rating
Passenger_Load
Flight_Cancelled

[ ]: