Untitled

August 26, 2024

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
data = pd.read_csv('Flyzy Flight Cancellation - Sheet1.csv')
data.iloc[1:4, 2:3]= np.NaN
data.iloc[1:4, 3:4]= "NA"
data.iloc[1:4, 4:5]= ""
data["None_col"]= None
data.head()
```

[2]:

	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	None_col
0	7319483	Airline D	475.0	Airport 3	Airport 2	4	6	1	Type C	0.225122	5.0	2.151974	0.477202	0	None
1	4791965	Airline E	NaN	NA		12	1	6	Type B	0.060346	68.0	1.600779	0.159718	1	None
2	2991718	Airline C	NaN	NA		17	3	9	Type C	0.093920	18.0	4.406848	0.256803	0	None
3	4220106	Airline E	NaN	NA		1	1	8	Type B	0.656750	13.0	0.998757	0.504077	1	None
4	2263008	Airline E	566.0	Airport 2	Airport 2	19	7	12	Type E	0.505211	4.0	3.806206	0.019638	0	None

[3]:
```
null= pd.isnull(data)
null.head()
```

[3]:

	Flight ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	\
0	False	False	False	False	False	
1	False	False	True	False	False	
2	False	False	True	False	False	
3	False	False	True	False	False	
4	False	False	False	False	False	

	Scheduled_Departure_Time	Day_of_Week	Month	Airplane_Type	Weather_Score	\
0	False	False	False	False	False	
1	False	False	False	False	False	
2	False	False	False	False	False	
3	False	False	False	False	False	
4	False	False	False	False	False	

	Previous_Flight_Delay_Minutes	Airline_Rating	Passenger_Load	\
0	False	False	False	
1	False	False	False	
2	False	False	False	
3	False	False	False	
4	False	False	False	

	Flight_Cancelled	None_col
0	False	True
1	False	True
2	False	True
3	False	True
4	False	True

[4]:
```
pd.isnull(data).sum().sum()
```

[4]: 3003

[5]:
```
missing_vals= ['NA',"", None, np.NaN]
missing= data.isin(missing_vals)
missing.head()
```

[5]:

	Flight ID	Airline	Flight_Distance	Origin_Airport	Destination_Airport	\
0	False	False	False	False	False	
1	False	False	True	True	True	
2	False	False	True	True	True	
3	False	False	True	True	True	
4	False	False	False	False	False	

	Scheduled_Departure_Time	Day_of_Week	Month	Airplane_Type	Weather_Score	\

```
0                    False        False  False          False          False
1                    False        False  False          False          False
2                    False        False  False          False          False
3                    False        False  False          False          False
4                    False        False  False          False          False

   Previous_Flight_Delay_Minutes  Airline_Rating  Passenger_Load  \
0                          False           False           False
1                          False           False           False
2                          False           False           False
3                          False           False           False
4                          False           False           False

   Flight_Cancelled  None_col
0             False      True
1             False      True
2             False      True
3             False      True
4             False      True
```

[6]:
```
data.fillna(0).head()
```

[6]:
```
   Flight ID    Airline  Flight_Distance Origin_Airport Destination_Airport  \
0    7319483  Airline D            475.0      Airport 3           Airport 2
1    4791965  Airline E              0.0             NA
2    2991718  Airline C              0.0             NA
3    4220106  Airline E              0.0             NA
4    2263008  Airline E            566.0      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  None_col
0                 0         0
1                 1         0
2                 0         0
```

```
3                  1         0
4                  0         0
```

[7]:
```
missing_vals= ['NA' ,"" ,None, np.NaN]
missing= data.isin(missing_vals)
data.mask(missing, "missing").head()
```

[7]:
```
   Flight ID    Airline Flight_Distance Origin_Airport Destination_Airport  \
0    7319483  Airline D           475.0      Airport 3           Airport 2
1    4791965  Airline E         missing        missing             missing
2    2991718  Airline C         missing        missing             missing
3    4220106  Airline E         missing        missing             missing
4    2263008  Airline E           566.0      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 None_col
0                 0  missing
1                 1  missing
2                 0  missing
3                 1  missing
4                 0  missing
```

[8]:
```
missing_values = data.isnull().sum()
print("Missing values per column (before handling):")
print(missing_values)
```

```
Missing values per column (before handling):
Flight ID                         0
Airline                           0
Flight_Distance                   3
Origin_Airport                    0
Destination_Airport               0
Scheduled_Departure_Time          0
Day_of_Week                       0
```

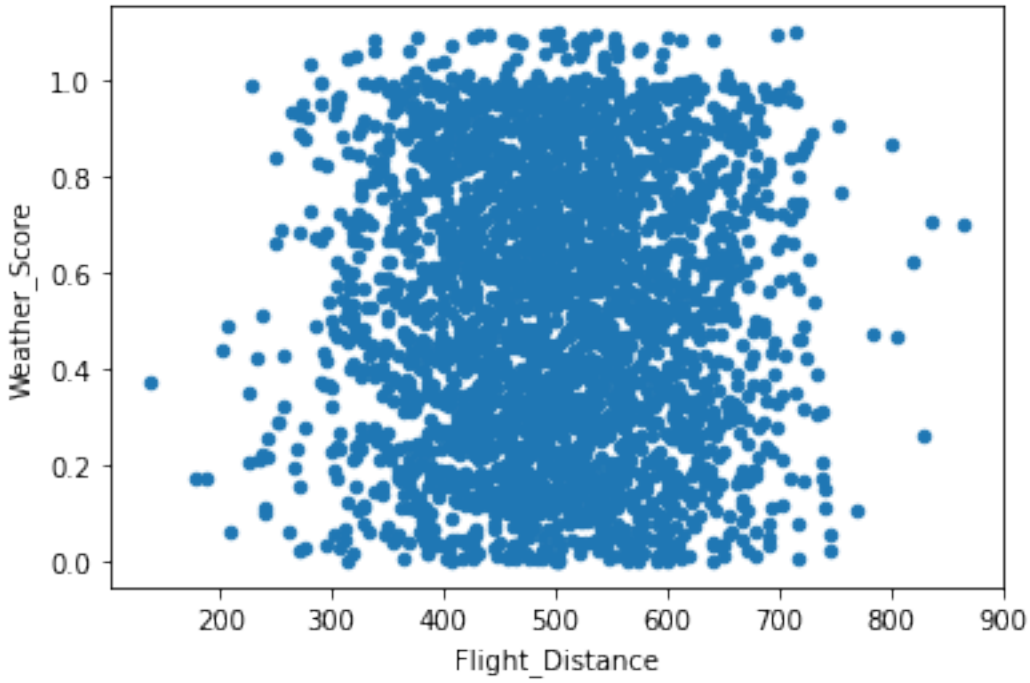
```
Month                              0
Airplane_Type                      0
Weather_Score                      0
Previous_Flight_Delay_Minutes      0
Airline_Rating                     0
Passenger_Load                     0
Flight_Cancelled                   0
None_col                        3000
dtype: int64
```

[9]:
```
data = data.drop('None_col' , axis=1)
```

[10]:
```
missing_values = data.isnull().sum()
print("Missing values per column (after Handling):")
print(missing_values)
```

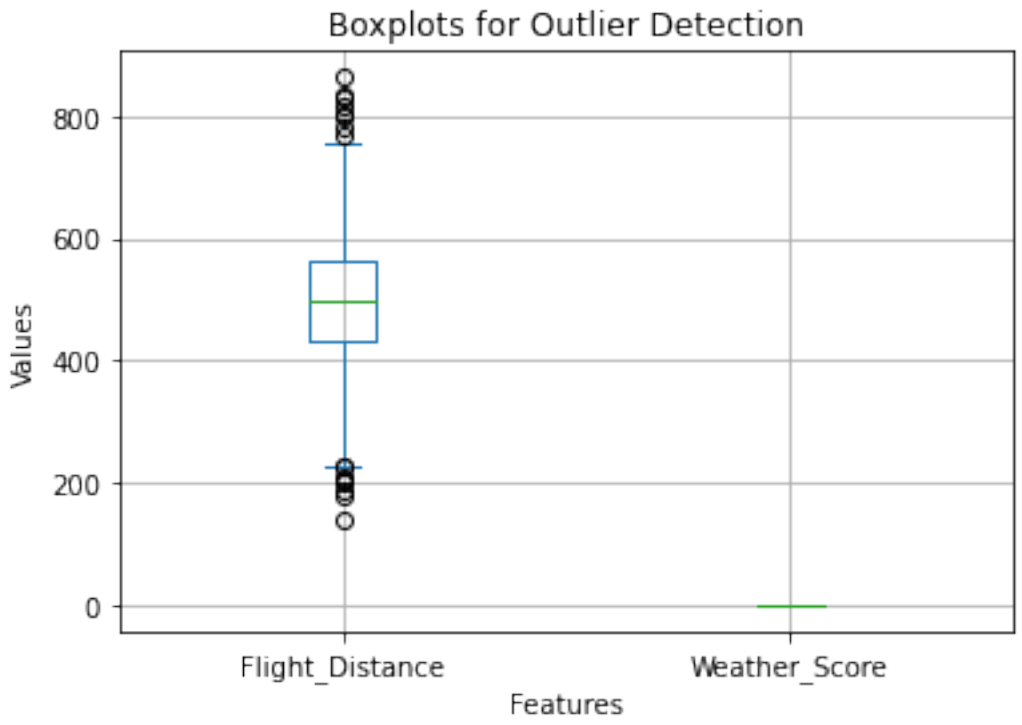
```
Missing values per column (after Handling):
Flight ID                          0
Airline                            0
Flight_Distance                    3
Origin_Airport                     0
Destination_Airport                0
Scheduled_Departure_Time           0
Day_of_Week                        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
```

[11]:
```
import pandas as pd
import matplotlib.pyplot as plt
data.plot(kind='scatter' , x= 'Flight_Distance', y= 'Weather_Score')
plt.show()
```



[12]:
```
import matplotlib.pyplot as plt
import pandas as pd
columns_to_check = ['Flight_Distance', 'Weather_Score']
plt.figure(figsize=(10, 6))
data[columns_to_check].plot(kind= 'box')
plt.title('Boxplots for Outlier Detection')
plt.xlabel('Features')
plt.ylabel('Values')
plt.grid(True)
plt.show()
```

```
<Figure size 720x432 with 0 Axes>
```



[13]:
```
data_types = data.dtypes
print("Data types of each column:")
print(data_types)
```

```
Data types of each column:
Flight ID                          int64
Airline                           object
Flight_Distance                  float64
Origin_Airport                    object
Destination_Airport               object
Scheduled_Departure_Time            int64
Day_of_Week                         int64
Month                               int64
Airplane_Type                     object
Weather_Score                    float64
Previous_Flight_Delay_Minutes    float64
Airline_Rating                   float64
Passenger_Load                   float64
Flight_Cancelled                    int64
dtype: object
```

[ ]: