In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
train=pd.read_excel("C:\Users\Admin\Downloads\Flight_Data_Train.xlsx")
```

In [3]:
```
test=pd.read_excel("C:\Users\Admin\Downloads\Flight_Test_set.xlsx")
```

In [4]:
```
df=pd.concat([train,test])
df.head()
```

Out[4]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897.0
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882.0
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218.0
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302.0

In [13]:
```
df.shape
```

Out[13]:
```
(13354, 11)
```

In [14]:
```
df=train.append(test)
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_16732\2043221198.py:1: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a fu
ture version. Use pandas.concat instead.
  df=train.append(test)
```

In [15]:
```
df.head()
```

Out[15]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897.0
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882.0
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218.0
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302.0

In [16]:
```
df.shape
```

Out[16]:
```
(13354, 11)
```

In [17]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13354 entries, 0 to 2670
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13354 non-null  object
 1   Date_of_Journey  13354 non-null  object
 2   Source           13354 non-null  object
 3   Destination      13354 non-null  object
 4   Route            13353 non-null  object
 5   Dep_Time         13354 non-null  object
 6   Arrival_Time     13354 non-null  object
 7   Duration         13354 non-null  object
 8   Total_Stops      13353 non-null  object
 9   Additional_Info  13354 non-null  object
 10  Price            10683 non-null  float64
dtypes: float64(1), object(10)
memory usage: 1.2+ MB
```

In [18]:
```
df['Date']=df['Date_of_Journey'].str.split('/').str[0]
df['Month']=df['Date_of_Journey'].str.split('/').str[1]
df['Year']=df['Date_of_Journey'].str.split('/').str[2]
```

In [19]:
```
df.drop(columns='Date_of_Journey',axis=1,inplace=True)
```

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

Out[20]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897.0	24	03	2019
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0	1	05	2019

In [21]:
```
df['Month']=df['Month'].astype(int)
df['Date']=df['Date'].astype(int)
df['Year']=df['Year'].astype(int)
```

In [22]:
```
df.head(2)
```

Out[22]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897.0	24	3	2019
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0	1	5	2019

In [23]:
```
df['Route'].str.replace('→','to')
```

Out[23]:
```
0              BLR to DEL
1          CCU to IXR to BBI to BLR
2          DEL to LKO to BOM to COK
3              CCU to NAG to BLR
4              BLR to NAG to DEL
                  ...
2666           CCU to DEL to BLR
2667           CCU to BLR
2668       DEL to BOM to COK
2669       DEL to BOM to COK
2670       DEL to BOM to COK
Name: Route, Length: 13354, dtype: object
```

In [24]:
```
df['Arrival_Time']=df['Arrival_Time'].str.split(' ').str[0]
```

In [25]:
```
df['Arrival_Time']
```

Out[25]:
```
0       01:10
1       13:15
2       04:25
3       23:30
4       21:35
        ...  
2666    20:25
2667    16:55
2668    09:25
2669    19:15
2670    19:15
Name: Arrival_Time, Length: 13354, dtype: object
```

In [27]:
```
df['Arrival_hours']=df['Arrival_Time'].str.split(':').str[0]
df['Arrival_min']=df['Arrival_Time'].str.split(':').str[1]
```

In [28]:
```
df.drop(columns='Arrival_Time',axis=1,inplace=True)
```

In [29]:
```
df['Arrival_hours']=df['Arrival_hours'].astype(int)
df['Arrival_min']=df['Arrival_min'].astype(int)
```

In [30]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13354 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13354 non-null  object
 1   Source           13354 non-null  object
 2   Destination      13354 non-null  object
 3   Route            13353 non-null  object
 4   Dep_Time         13354 non-null  object
 5   Duration         13354 non-null  object
 6   Total_Stops      13353 non-null  object
 7   Additional_Info  13354 non-null  object
 8   Price            10683 non-null  float64
 9   Date             13354 non-null  int32 
 10  Month            13354 non-null  int32 
 11  Year             13354 non-null  int32 
 12  Arrival_hours    13354 non-null  int32 
 13  Arrival_min      13354 non-null  int32 
dtypes: float64(1), int32(5), object(8)
memory usage: 1.4+ MB
```

In [31]:
```
df['Dep_hours']=df['Dep_Time'].str.split(':').str[0]
df['Dep_min']=df['Dep_Time'].str.split(':').str[1]
df['Dep_hours']=df['Dep_hours'].astype(int)
df['Dep_min']=df['Dep_min'].astype(int)
df.drop(columns='Dep_Time',axis=1,inplace=True)
```

In [32]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13354 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13354 non-null  object
 1   Source           13354 non-null  object
 2   Destination      13354 non-null  object
 3   Route            13353 non-null  object
 4   Duration         13354 non-null  object
 5   Total_Stops      13353 non-null  object
 6   Additional_Info  13354 non-null  object
 7   Price            10683 non-null  float64
 8   Date             13354 non-null  int32 
 9   Month            13354 non-null  int32 
 10  Year             13354 non-null  int32 
 11  Arrival_hours    13354 non-null  int32 
 12  Arrival_min      13354 non-null  int32 
 13  Dep_hours        13354 non-null  int32 
 14  Dep_min          13354 non-null  int32 
dtypes: float64(1), int32(7), object(7)
memory usage: 1.7+ MB
```

In [33]:
```
df.head(2)
```

Out[33]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	non-stop	No info	3897.0	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	2 stops	No info	7662.0	1	5	2019	13	15	5	50

In [34]:
```
df['Total_Stops'].unique()
```

Out[34]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

In [35]:
```
df['Total_Stops']=df['Total_Stops'].map({'non-stop':0,'1 stop':1,'2 stops':2,'3 stops':3,'4 stops':4,'nan':1})
df.head()
```

Out[35]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	0	No info	3897.0	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	2	No info	7662.0	1	5	2019	13	15	5	50
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	19h	2	No info	13882.0	9	6	2019	4	25	9	25
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	5h 25m	1	No info	6218.0	12	5	2019	23	30	18	5
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	4h 45m	1	No info	13302.0	1	3	2019	21	35	16	50

In [36]:
```
from sklearn import preprocessing
label_encoder=preprocessing.LabelEncoder()
df['Total_Stops']=label_encoder.fit_transform(df['Total_Stops'])
```

In [37]:
```
df.head()
```

Out[37]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	4	No info	3897.0	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	1	No info	7662.0	1	5	2019	13	15	5	50
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	19h	1	No info	13882.0	9	6	2019	4	25	9	25
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	5h 25m	0	No info	6218.0	12	5	2019	23	30	18	5
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	4h 45m	0	No info	13302.0	1	3	2019	21	35	16	50

In [38]:
```
df['Total_Stops'].unique()
```

Out[38]:
```
array([4, 1, 0, 2, 5, 3])
```

In [39]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13354 entries, 0 to 2670
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13354 non-null  object
 1   Source           13354 non-null  object
 2   Destination      13354 non-null  object
 3   Route            13353 non-null  object
 4   Duration         13354 non-null  object
 5   Total_Stops      13354 non-null  int32 
 6   Additional_Info  13354 non-null  object
 7   Price            10683 non-null  float64
 8   Date             13354 non-null  int32 
 9   Month            13354 non-null  int32 
 10  Year             13354 non-null  int32 
 11  Arrival_hours    13354 non-null  int32 
 12  Arrival_min      13354 non-null  int32 
 13  Dep_hours        13354 non-null  int32 
 14  Dep_min          13354 non-null  int32 
dtypes: float64(1), int32(8), object(6)
memory usage: 1.2+ MB
```

In [40]:
```
df['Additional_Info'].unique()
```

Out[40]:
```
array(['No info', 'In-flight meal not included',
       'No check-in baggage included', '1 Short layover', 'No Info',
       '1 Long layover', 'Change airports', 'Business class',
       'Red-eye flight', '2 Long layover'], dtype=object)
```

In [41]:
```
df['Duration_hour']=df['Duration'].str.split(' ').str[0].str.split('h').str[0]
```

In [42]:
```
df['Duration_hour'].unique()
```

Out[42]:
```
array(['2', '7', '19', '5', '4', '15', '21', '25', '13', '12', '26', '22',
       '23', '20', '18', '9', '11', '6', '16', '3', '17', '1', '14', '8',
       '10', '24', '28', '29', '27', '34', '30', '35', '36',
       '47', '33', '32', '31', '42', '39', '37', '38', '41', '40'], dtype=object)
```

In [43]:
```
df['Duration_hour']=='5m'
```

Out[43]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min	Duration_hour
6474	Air India	Mumbai	Hyderabad	5m	1	No info	17327.0	6	3	2019	16	55	16	50	5m
2660	Air India	Mumbai	Hyderabad	5m	1	No info	NaN	12	3	2019	16	55	16	50	5m

In [44]:
```
df.drop(columns='Route',axis=1,inplace=True)
```

In [45]:
```
df.drop(6474,axis=0,inplace=True)
df.drop(2660,axis=0,inplace=True)
```

In [46]:
```
df['Duration_hour']=='5m'
```

Out[46]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min	Duration_hour

In [47]:
```
df['Duration_hour']=df['Duration_hour'].astype('int')
```

In [48]:
```
df.head()
```

Out[48]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min	Duration_hour
0	IndiGo	Banglore	New Delhi	2h 50m	4	No info	3897.0	24	3	2019	1	10	22	20	2
1	Air India	Kolkata	Banglore	7h 25m	1	No info	7662.0	1	5	2019	13	15	5	50	7
2	Jet Airways	Delhi	Cochin	19h	1	No info	13882.0	9	6	2019	4	25	9	25	19
3	IndiGo	Kolkata	Banglore	5h 25m	0	No info	6218.0	12	5	2019	23	30	18	5	5
4	IndiGo	Banglore	New Delhi	4h 45m	0	No info	13302.0	1	3	2019	21	35	16	50	4

In [49]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13351 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13351 non-null  object
 1   Source           13351 non-null  object
 2   Destination      13351 non-null  object
 3   Total_Stops      13351 non-null  int32 
 4   Additional_Info  13351 non-null  object
 5   Price            10681 non-null  float64
 6   Date             13351 non-null  int32 
 7   Month            13351 non-null  int32 
 8   Year             13351 non-null  int32 
 9   Arrival_hours    13351 non-null  int32 
 10  Arrival_min      13351 non-null  int32 
 11  Dep_hours        13351 non-null  int32 
 12  Dep_min          13351 non-null  int32 
 13  Duration_hour    13351 non-null  int32 
dtypes: float64(1), int32(9), object(3)
memory usage: 1.6+ MB
```

In [50]:
```
df.drop('Duration',axis=1,inplace=True)
```

In [51]:
```
df['Airline'].unique()
```

Out[51]:
```
array(['IndiGo', 'Air India', 'Jet Airways', 'SpiceJet',
       'Multiple carriers', 'GoAir', 'Vistara', 'Air Asia',
       'Vistara Premium economy', 'Jet Airways Business',
       'Multiple carriers Premium economy', 'Trujet'], dtype=object)
```

In [52]:
```
df['Airline']=label_encoder.fit_transform(df['Airline'])
```

In [53]:
```
df['Source']=label_encoder.fit_transform(df['Source'])
```

In [54]:
```
df['Destination']=label_encoder.fit_transform(df['Destination'])
```

In [55]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13351 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13351 non-null  int64 
 1   Source           13351 non-null  int32 
 2   Destination      13351 non-null  int32 
 3   Total_Stops      13351 non-null  int32 
 4   Additional_Info  13351 non-null  object
 5   Price            10681 non-null  float64
 6   Date             13351 non-null  int32 
 7   Month            13351 non-null  int32 
 8   Year             13351 non-null  int32 
 9   Arrival_hours    13351 non-null  int32 
 10  Arrival_min      13351 non-null  int32 
 11  Dep_hours        13351 non-null  int32 
 12  Dep_min          13351 non-null  int32 
 13  Duration_hour    13351 non-null  int32 
dtypes: float64(1), int32(11), int64(1), object(1)
memory usage: 998.0+ KB
```

In [56]:
```
df['Additional_Info'].unique()
```

Out[56]:
```
array(['No info', 'In-flight meal not included',
       'No check-in baggage included', '1 Short layover', 'No Info',
       '1 Long layover', 'Change airports', 'Business class',
       'Red-eye flight', '2 Long layover'], dtype=object)
```

In [57]:
```
df['Additional_Info']=label_encoder.fit_transform(df['Additional_Info'])
```

In [58]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13351 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          13351 non-null  int64 
 1   Source           13351 non-null  int32 
 2   Destination      13351 non-null  int32 
 3   Total_Stops      13351 non-null  int32 
 4   Additional_Info  13351 non-null  int64 
 5   Price            10681 non-null  float64
 6   Date             13351 non-null  int32 
 7   Month            13351 non-null  int32 
 8   Year             13351 non-null  int32 
 9   Arrival_hours    13351 non-null  int32 
 10  Arrival_min      13351 non-null  int32 
 11  Dep_hours        13351 non-null  int32 
 12  Dep_min          13351 non-null  int32 
 13  Duration_hour    13351 non-null  int32 
dtypes: float64(1), int32(11), int64(2)
memory usage: 938.7 KB
```

In [59]:
```
df.shape
```

Out[59]:
```
(13351, 14)
```

In [60]:
```
df.head()
```

Out[60]:

	Airline	Source	Destination	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hours	Arrival_min	Dep_hours	Dep_min	Duration_hour
0	3	0	5	4	8	3897.0	24	3	2019	1	10	22	20	2
1	1	3	0	1	8	7662.0	1	5	2019	13	15	5	50	7
2	4	2	1	1	8	13882.0	9	6	2019	4	25	9	25	19
3	3	3	0	0	8	6218.0	12	5	2019	23	30	18	5	5
4	3	0	5	0	8	13302.0	1	3	2019	21	35	16	50	4

In [ ]:
```

```