In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df_train = pd.read_excel(r'C:\Users\MALU\Downloads\Data_Train.xlsx')
```

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

Out[3]:

	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
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

In [4]:
```
df_test = pd.read_excel(r'C:\Users\MALU\Downloads\Test_set.xlsx')
```

In [5]:
```
df_test.head()
```

Out[5]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
0	Jet Airways	6/06/2019	Delhi	Cochin	DEL → BOM → COK	17:30	04:25 07 Jun	10h 55m	1 stop	No info
1	IndiGo	12/05/2019	Kolkata	Banglore	CCU → MAA → BLR	06:20	10:20	4h	1 stop	No info
2	Jet Airways	21/05/2019	Delhi	Cochin	DEL → BOM → COK	19:15	19:00 22 May	23h 45m	1 stop	In-flight meal not included
3	Multiple carriers	21/05/2019	Delhi	Cochin	DEL → BOM → COK	08:00	21:00	13h	1 stop	No info
4	Air Asia	24/06/2019	Banglore	Delhi	BLR → DEL	23:55	02:45 25 Jun	2h 50m	non-stop	No info

In [6]:
```
df_train.shape
```

Out[6]: (10683, 11)

In [7]:
```
df_test.shape
```

Out[7]: (2671, 10)

In [8]:
```
df = pd.concat([df_train,df_test], axis=0)
```

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

Out[9]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
2666	Air India	6/06/2019	Kolkata	Banglore	CCU → DEL → BLR	20:30	20:25 07 Jun	23h 55m	1 stop	No info	NaN
2667	IndiGo	27/03/2019	Kolkata	Banglore	CCU → BLR	14:20	16:55	2h 35m	non-stop	No info	NaN
2668	Jet Airways	6/03/2019	Delhi	Cochin	DEL → BOM → COK	21:50	04:25 07 Mar	6h 35m	1 stop	No info	NaN
2669	Air India	6/03/2019	Delhi	Cochin	DEL → BOM → COK	04:00	19:15	15h 15m	1 stop	No info	NaN
2670	Multiple carriers	15/06/2019	Delhi	Cochin	DEL → BOM → COK	04:55	19:15	14h 20m	1 stop	No info	NaN

In [10]:
```
df.shape
```

Out[10]: (13354, 11)

In [11]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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 [12]:
```
df.isnull().sum()
```

Out[12]:
```
Airline               0
Date_of_Journey       0
Source                0
Destination           0
Route                 1
Dep_Time              0
Arrival_Time          0
Duration              0
Total_Stops           1
Additional_Info       0
Price              2671
dtype: int64
```

In [13]:
```
# Feature engineering
```

In [14]:
```
df['Date_of_Journey'] = pd.to_datetime(df['Date_of_Journey'])
df['year'] = df['Date_of_Journey'].dt.year
df['month'] = df['Date_of_Journey'].dt.month
df['day'] = df['Date_of_Journey'].dt.day
```

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

Out[15]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	year	month	day
0	IndiGo	2019-03-24	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897.0	2019	3	24
1	Air India	2019-05-01	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0	2019	5	1
2	Jet Airways	2019-06-09	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882.0	2019	6	9
3	IndiGo	2019-05-12	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218.0	2019	5	12
4	IndiGo	2019-03-01	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302.0	2019	3	1

In [16]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13354 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          13354 non-null  object
 1   Date_of_Journey  13354 non-null  datetime64[ns]
 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
 11  year             13354 non-null  int32
 12  month            13354 non-null  int32
 13  day              13354 non-null  int32
dtypes: datetime64[ns](1), float64(1), int32(3), object(9)
memory usage: 1.4+ MB
```

In [17]:
```
df.drop('Date_of_Journey',axis=1,inplace=True)
```

In [18]:
```
df.head()
```

Out[18]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	year	month	day
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897.0	2019	3	24
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0	2019	5	1
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882.0	2019	6	9
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218.0	2019	5	12
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302.0	2019	3	1

In [19]:
```
df['Arrival_Time']=df['Arrival_Time'].str.split(' ').str[0]
```

In [20]:
```
df.sample(5)
```

Out[20]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	year	month	day
3826	Jet Airways	Delhi	Cochin	DEL → AMD → BOM → COK	23:05	04:25	29h 20m	2 stops	In-flight meal not included	11150.0	2019	6	27
496	Jet Airways	Delhi	Cochin	DEL → BOM → COK	20:55	04:25	7h 30m	1 stop	No info	14714.0	2019	6	6
3748	Multiple carriers	Delhi	Cochin	DEL → BOM → COK	10:20	18:50	8h 30m	1 stop	No info	11999.0	2019	3	21
3330	IndiGo	Mumbai	Hyderabad	BOM → HYD	19:05	20:35	1h 30m	non-stop	No info	2754.0	2019	6	12
1499	Jet Airways	Kolkata	Banglore	CCU → BOM → BLR	14:05	04:40	14h 35m	1 stop	In-flight meal not included	NaN	2019	6	9

In [21]:
```
df['Arrival_hour']=df['Arrival_Time'].str.split(':').str[0]
df['Arrival_min']=df['Arrival_Time'].str.split(':').str[1]
```

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

Out[22]:

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

In [23]:
```
df['Arrival_hour']=df['Arrival_hour'].astype(int)
df['Arrival_min']=df['Arrival_min'].astype(int)
```

In [24]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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   Dep_Time         13354 non-null  object
 5   Arrival_Time     13354 non-null  object
 6   Duration         13354 non-null  object
 7   Total_Stops      13353 non-null  object
 8   Additional_Info  13354 non-null  object
 9   Price            10683 non-null  float64
 10  year             13354 non-null  int32
 11  month            13354 non-null  int32
 12  day              13354 non-null  int32
 13  Arrival_hour     13354 non-null  int32
 14  Arrival_min      13354 non-null  int32
dtypes: float64(1), int32(5), object(9)
memory usage: 1.4+ MB
```

In [25]:
```
df.drop('Arrival_Time',axis=1,inplace=True)
```

In [26]:
```
df['Dept_hour'] = df['Dep_Time'].str.split(':').str[0]
df['Dept_min'] = df['Dep_Time'].str.split(':').str[1]
df['Dept_hour'] = df['Dept_hour'].astype(int)
df['Dept_min'] = df['Dept_min'].astype(int)
df.drop('Dep_Time',axis=1,inplace=True)
```

In [27]:
```
df.head(2)
```

Out[27]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	year	month	day	Arrival_hour	Arrival_min	Dept_hour	Dept_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	non-stop	No info	3897.0	2019	3	24	1	10	22	20
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	2 stops	No info	7662.0	2019	5	1	13	15	5	50

In [28]:
```
df['Total_Stops'].unique()
```

Out[28]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

In [30]:
```
df[df['Total_Stops'].isnull()]
```

Out[30]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	year	month	day	Arrival_hour	Arrival_min	Dept_hour	Dept_min
9039	Air India	Delhi	Cochin	NaN	23h 40m	NaN	No info	7480.0	2019	5	6	9	25	9	45

In [31]:
```
df['Total_Stops']=df['Total_Stops'].map({'non-stop':0, '1 stop':1, '2 stops':2, '3 stops':3, '4 stops':4, 'nan':1})
```

In [32]:
```
df.drop('Route',axis=1,inplace=True)
```

In [33]:
```
df.sample(3)
```

Out[33]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	year	month	day	Arrival_hour	Arrival_min	Dept_hour	Dept_min
5997	Multiple carriers	Delhi	Cochin	10h	1.0	No info	8099.0	2019	6	24	19	15	9	15
6393	Jet Airways	Kolkata	Banglore	28h 5m	1.0	No info	14151.0	2019	5	15	21	5	17	0
9913	Air India	Kolkata	Banglore	15h 20m	2.0	No info	10408.0	2019	4	1	1	20	10	0

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

Out[34]:
```
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 [36]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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   Duration         13354 non-null  object
 4   Total_Stops      13353 non-null  float64
 5   Additional_Info  13354 non-null  object
 6   Price            10683 non-null  float64
 7   year             13354 non-null  int32
 8   month            13354 non-null  int32
 9   day              13354 non-null  int32
 10  Arrival_hour     13354 non-null  int32
 11  Arrival_min      13354 non-null  int32
 12  Dept_hour        13354 non-null  int32
 13  Dept_min         13354 non-null  int32
dtypes: float64(2), int32(7), object(5)
memory usage: 1.2+ MB
```

In [37]:
```
df['Duration_hour'] =df['Duration'].str.split(' ').str[0].str.split('h').str[0]
```

In [38]:
```
df[df['Duration_hour']=='5m']
```

Out[38]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	year	month	day	Arrival_hour	Arrival_min	Dept_hour	Dept_min	Duration_hour
6474	Air India	Mumbai	Hyderabad	5m	2.0	No info	17327.0	2019	3	6	16	55	16	50	5m
2660	Air India	Mumbai	Hyderabad	5m	2.0	No info	NaN	2019	3	12	16	55	16	50	5m

In [39]:
```
df.drop(6474,axis=0,inplace=True)
df.drop(2660,axis=0,inplace=True)
```

In [40]:
```
df['Duration_hour']= df['Duration_hour'].astype('int')
```

In [41]:
```
df['Duration-min']=df['Duration'].str.split(' ').str[1].str.split('m').str[0]
```

In [42]:
```
df[df['Duration-min'].isnull()]
```

Out[42]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	year	month	day	Arrival_hour	Arrival_min	Dept_hour	Dept_min	Duration_hour	Duration-min
2	Jet Airways	Delhi	Cochin	19h	2.0	No info	13882.0	2019	6	9	4	25	9	25	19	NaN
18	Air India	Delhi	Cochin	23h	2.0	No info	13381.0	2019	6	12	19	15	20	15	23	NaN
33	Jet Airways	Delhi	Cochin	22h	2.0	In-flight meal not included	10919.0	2019	6	15	12	35	14	35	22	NaN
44	Multiple carriers	Delhi	Cochin	12h	1.0	No info	13062.0	2019	3	21	21	0	9	0	12	NaN
53	IndiGo	Banglore	Delhi	3h	0.0	No info	3943.0	2019	6	18	0	15	21	15	3	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2588	Air India	Delhi	Cochin	3h	0.0	No info	NaN	2019	3	27	17	10	14	10	3	NaN
2598	Jet Airways	Delhi	Cochin	11h	1.0	No info	NaN	2019	6	9	19	0	8	0	11	NaN
2604	Multiple carriers	Delhi	Cochin	10h	1.0	No info	NaN	2019	6	12	19	15	9	15	10	NaN
2607	Multiple carriers	Delhi	Cochin	13h	1.0	No info	NaN	2019	3	3	21	0	8	0	13	NaN
2622	Jet Airways	Banglore	Delhi	3h	0.0	In-flight meal not included	NaN	2019	5	3	20	45	17	45	3	NaN

1283 rows × 16 columns

In [43]:
```
df.drop('Duration',axis=1,inplace=True)
```

In [44]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13351 entries, 0 to 2670
Data columns (total 15 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      13350 non-null  float64
 4   Additional_Info  13351 non-null  object
 5   Price            10681 non-null  float64
 6   year             13351 non-null  int32
 7   month            13351 non-null  int32
 8   day              13351 non-null  int32
 9   Arrival_hour     13351 non-null  int32
 10  Arrival_min      13351 non-null  int32
 11  Dept_hour        13351 non-null  int32
 12  Dept_min         13351 non-null  int32
 13  Duration_hour    13351 non-null  int32
 14  Duration-min     12068 non-null  object
dtypes: float64(2), int32(8), object(5)
memory usage: 1.2+ MB
```

In [45]:
```
df['Airline'].unique()
```

Out[45]:
```
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 [46]:
```
df['Source'].unique()
```

Out[46]:
```
array(['Banglore', 'Kolkata', 'Delhi', 'Chennai', 'Mumbai'], dtype=object)
```

In [47]:
```
df['Destination'].unique()
```

Out[47]:
```
array(['New Delhi', 'Banglore', 'Cochin', 'Kolkata', 'Delhi', 'Hyderabad'],
      dtype=object)
```

In [48]:
```
df['Additional_Info'].unique()
```

Out[48]:
```
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 [49]:
```
from sklearn.preprocessing import LabelEncoder
labelencoder=LabelEncoder()
```

In [50]:
```
df['Airline'] = labelencoder.fit_transform(df['Airline'])
df['Source'] = labelencoder.fit_transform(df['Source'])
df['Destination'] = labelencoder.fit_transform(df['Destination'])
df['Additional_Info'] = labelencoder.fit_transform(df['Additional_Info'])
```

In [51]:
```
df.head(3)
```

Out[51]:

	Airline	Source	Destination	Total_Stops	Additional_Info	Price	year	month	day	Arrival_hour	Arrival_min	Dept_hour	Dept_min	Duration_hour	Duration-min
0	3	0	5	0.0	8	3897.0	2019	3	24	1	10	22	20	2	50
1	1	3	0	2.0	8	7662.0	2019	5	1	13	15	5	50	7	25
2	4	2	1	2.0	8	13882.0	2019	6	9	4	25	9	25	19	NaN

In [52]:
```
df.shape
```

Out[52]: (13351, 15)

In [53]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13351 entries, 0 to 2670
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          13351 non-null  int32
 1   Source           13351 non-null  int32
 2   Destination      13351 non-null  int32
 3   Total_Stops      13350 non-null  float64
 4   Additional_Info  13351 non-null  int32
 5   Price            10681 non-null  float64
 6   year             13351 non-null  int32
 7   month            13351 non-null  int32
 8   day              13351 non-null  int32
 9   Arrival_hour     13351 non-null  int32
 10  Arrival_min      13351 non-null  int32
 11  Dept_hour        13351 non-null  int32
 12  Dept_min         13351 non-null  int32
 13  Duration_hour    13351 non-null  int32
 14  Duration-min     12068 non-null  object
dtypes: float64(2), int32(12), object(1)
memory usage: 1.0+ MB
```

In [54]:
```
df.columns
```

Out[54]:
```
Index(['Airline', 'Source', 'Destination', 'Total_Stops', 'Additional_Info',
       'Price', 'year', 'month', 'day', 'Arrival_hour', 'Arrival_min',
       'Dept_hour', 'Dept_min', 'Duration_hour', 'Duration-min'],
      dtype='object')
```

In [55]:
```
df['Duration_hour']=df['Duration_hour']* 60
```

In [56]:
```
df['Duration_hour']
```

Out[56]:
```
0        120
1        420
2       1140
3        300
4        240
        ... 
2666    1380
2667     120
2668     360
2669     900
2670     840
Name: Duration_hour, Length: 13351, dtype: int32
```

In [57]:
```
df['Duration-min']=df['Duration-min'].fillna(0)
```

In [58]:
```
df['Duration-min']=df['Duration-min'].astype('int')
```

In [59]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13351 entries, 0 to 2670
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          13351 non-null  int32
 1   Source           13351 non-null  int32
 2   Destination      13351 non-null  int32
 3   Total_Stops      13350 non-null  float64
 4   Additional_Info  13351 non-null  int32
 5   Price            10681 non-null  float64
 6   year             13351 non-null  int32
 7   month            13351 non-null  int32
 8   day              13351 non-null  int32
 9   Arrival_hour     13351 non-null  int32
 10  Arrival_min      13351 non-null  int32
 11  Dept_hour        13351 non-null  int32
 12  Dept_min         13351 non-null  int32
 13  Duration_hour    13351 non-null  int32
 14  Duration-min     13351 non-null  int32
dtypes: float64(2), int32(13)
memory usage: 990.9 KB
```

In [60]:
```
df['Total_Duration(min)'] = df['Duration_hour'] + df['Duration-min']
```

In [61]:
```
df.drop(['Duration_hour','Duration-min'],axis=1,inplace=True)
```

In [62]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13351 entries, 0 to 2670
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Airline              13351 non-null  int32
 1   Source               13351 non-null  int32
 2   Destination          13351 non-null  int32
 3   Total_Stops          13350 non-null  float64
 4   Additional_Info      13351 non-null  int32
 5   Price                10681 non-null  float64
 6   year                 13351 non-null  int32
 7   month                13351 non-null  int32
 8   day                  13351 non-null  int32
 9   Arrival_hour         13351 non-null  int32
 10  Arrival_min          13351 non-null  int32
 11  Dept_hour            13351 non-null  int32
 12  Dept_min             13351 non-null  int32
 13  Total_Duration(min)  13351 non-null  int32
dtypes: float64(2), int32(12)
memory usage: 938.7 KB
```