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

In [2]:
```
df = pd.read_excel('flight_price.xlsx')
df.head()
```

Out[2]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Du
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4

In [3]:
```
df.tail()
```

Out[3]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time
10678	Air Asia	9/04/2019	Kolkata	Banglore	CCU → BLR	19:55	22:25
10679	Air India	27/04/2019	Kolkata	Banglore	CCU → BLR	20:45	23:20
10680	Jet Airways	27/04/2019	Banglore	Delhi	BLR → DEL	08:20	11:20
10681	Vistara	01/03/2019	Banglore	New Delhi	BLR → DEL	11:30	14:10
10682	Air India	9/05/2019	Delhi	Cochin	DEL → GOI → BOM → COK	10:55	19:15

In [4]:
```
# Get the basic information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64 
dtypes: int64(1), object(10)
memory usage: 918.2+ KB
```

In [5]:
```
df.describe()
```

Out[5]:

	Price
count	10683.000000
mean	9087.064121
std	4611.359167
min	1759.000000
25%	5277.000000
50%	8372.000000
75%	12373.000000
max	79512.000000

In [6]:
```
df.head()
```

Out[6]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Du
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4

In [7]:
```
# Feature Engineering
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 [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64 
 11  Date             10683 non-null  object
 12  Month            10683 non-null  object
 13  Year             10683 non-null  object
dtypes: int64(1), object(13)
memory usage: 1.1+ MB
```

In [9]:
```
df['Date'] = df['Date'].astype(int)
df['Month'] = df['Month'].astype(int)
df['Year'] = df['Year'].astype(int)
```

In [10]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64 
 11  Date             10683 non-null  int64 
 12  Month            10683 non-null  int64 
 13  Year             10683 non-null  int64 
dtypes: int64(4), object(10)
memory usage: 1.1+ MB
```

In [11]:
```
# Drop Date of Journey column
df.drop('Date_of_Journey', axis = 1, inplace = True)
```

In [12]:
```
df.head()
```

Out[12]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop

In [13]:
```
df['Arrival_Time'] = df['Arrival_Time'].apply(lambda x:x.split(' ')[0])
```

In [14]:
```
df['Arrival_hour'] = df['Arrival_Time'].str.split(':').str[0]
df['Arrival_min'] = df['Arrival_Time'].str.split(':').str[1]
```

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

Out[15]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	01:10	2h 50m	non-stop
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25	19h	2 stops
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop

In [16]:
```
df['Arrival_hour'] = df['Arrival_hour'].astype(int)
df['Arrival_min'] = df['Arrival_min'].astype(int)
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10683 non-null  object
 1   Source           10683 non-null  object
 2   Destination      10683 non-null  object
 3   Route            10682 non-null  object
 4   Dep_Time         10683 non-null  object
 5   Arrival_Time     10683 non-null  object
 6   Duration         10683 non-null  object
 7   Total_Stops      10682 non-null  object
 8   Additional_Info  10683 non-null  object
 9   Price            10683 non-null  int64
 10  Date             10683 non-null  int64
 11  Month            10683 non-null  int64
 12  Year             10683 non-null  int64
 13  Arrival_hour     10683 non-null  int64
 14  Arrival_min      10683 non-null  int64
dtypes: int64(6), object(9)
memory usage: 1.2+ MB
```

In [18]:
```
df.drop('Arrival_Time', axis = 1, inplace = True)
```

In [19]:
```
df.head(2)
```

Out[19]:

	Airline	Source	Destination	Route	Dep_Time	Duration	Total_Stops	Additional_I
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	2h 50m	non-stop	No i
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	7h 25m	2 stops	No i

In [20]:
```
df['Depature_hour'] = df['Dep_Time'].str.split(':').str[0]
df['Depature_min'] = df['Dep_Time'].str.split(':').str[1]
```

In [21]:
```
df['Depature_hour'] = df['Depature_hour'].astype(int)
df['Depature_min'] = df['Depature_min'].astype(int)
```

In [22]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10683 non-null  object
 1   Source           10683 non-null  object
 2   Destination      10683 non-null  object
 3   Route            10682 non-null  object
 4   Dep_Time         10683 non-null  object
 5   Duration         10683 non-null  object
 6   Total_Stops      10682 non-null  object
 7   Additional_Info  10683 non-null  object
 8   Price            10683 non-null  int64
 9   Date             10683 non-null  int64
 10  Month            10683 non-null  int64
 11  Year             10683 non-null  int64
 12  Arrival_hour     10683 non-null  int64
 13  Arrival_min      10683 non-null  int64
 14  Depature_hour    10683 non-null  int64
 15  Depature_min     10683 non-null  int64
dtypes: int64(8), object(8)
memory usage: 1.3+ MB
```

In [23]:
```
df.drop('Dep_Time', axis = 1, inplace = True)
```

In [24]:
```
df.head(2)
```

Out[24]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	non-stop	No info	3897
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	2 stops	No info	7662

In [25]:
```
df['Total_Stops'].unique()
```

Out[25]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

In [26]:
```
df[df['Total_Stops'].isnull()]
```

Out[26]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price
9039	Air India	Delhi	Cochin	NaN	23h 40m	NaN	No info	7480

In [27]:
```
df['Total_Stops'].mode()
```

Out[27]:
```
0    1 stop
Name: Total_Stops, dtype: object
```

In [28]:
```
df['Total_Stops'] = df['Total_Stops'].map({'non-stop':0, '1 stop':1, '2 stops':2
```

In [29]:
```
df[df['Total_Stops'].isnull()]
```

Out[29]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Da

In [30]:
```
df.head(2)
```

Out[30]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	0	No info	3897
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	2	No info	7662

In [31]:
```
df.drop('Route', axis = 1, inplace = True)
```

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

Out[32]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date
0	IndiGo	Banglore	New Delhi	2h 50m	0	No info	3897	24
1	Air India	Kolkata	Banglore	7h 25m	2	No info	7662	1
2	Jet Airways	Delhi	Cochin	19h	2	No info	13882	9
3	IndiGo	Kolkata	Banglore	5h 25m	1	No info	6218	12
4	IndiGo	Banglore	New Delhi	4h 45m	1	No info	13302	1

In [33]:
```
# df['Duration'] = df['Duration'].map({np.nan:'0h 0m'})
```

In [34]:
```
df['Duration_Hour'] = df['Duration'].str.split(' ').str[0].str.split('h').str[0]
```

In [35]:
```
df['Duration_Minute'] = df['Duration'].str.split(' ').str[1].str.split('m').str[
```

In [36]:
```
# df[df['Duration'].isnull()]
```

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

Out[37]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date
0	IndiGo	Banglore	New Delhi	2h 50m	0	No info	3897	24
1	Air India	Kolkata	Banglore	7h 25m	2	No info	7662	1
2	Jet Airways	Delhi	Cochin	19h	2	No info	13882	9
3	IndiGo	Kolkata	Banglore	5h 25m	1	No info	6218	12
4	IndiGo	Banglore	New Delhi	4h 45m	1	No info	13302	1

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

Out[38]:
```
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 [39]:
```
df['Source'].unique()
```

Out[39]:
```
array(['Banglore', 'Kolkata', 'Delhi', 'Chennai', 'Mumbai'], dtype=object)
```

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]:
```
from sklearn.preprocessing import OneHotEncoder
```

In [42]:
```
encoder = OneHotEncoder()
```

In [43]:
```
encoder.fit_transform(df[['Airline', 'Source', 'Destination']]).toarray()
```

Out[43]:
```
array([[0., 0., 0., ..., 0., 0., 1.],
       [0., 1., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 1.],
       [0., 1., 0., ..., 0., 0., 0.]], shape=(10683, 23))
```

In [44]:
```
pd.DataFrame(encoder.fit_transform(df[['Airline', 'Source', 'Destination']]).toa
```

Out[44]:

	Airline_Air Asia	Airline_Air India	Airline_GoAir	Airline_IndiGo	Airline_Jet Airways	Airline_Jet Airways Business	Air
0	0.0	0.0	0.0	1.0	0.0	0.0	
1	0.0	1.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	1.0	0.0	
3	0.0	0.0	0.0	1.0	0.0	0.0	
4	0.0	0.0	0.0	1.0	0.0	0.0	
...	...	...	...	...	...	...	
10678	1.0	0.0	0.0	0.0	0.0	0.0	
10679	0.0	1.0	0.0	0.0	0.0	0.0	
10680	0.0	0.0	0.0	0.0	1.0	0.0	
10681	0.0	0.0	0.0	0.0	0.0	0.0	
10682	0.0	1.0	0.0	0.0	0.0	0.0	

10683 rows × 23 columns

In [ ]:

In [ ]: