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

In [2]:
```
df_training=pd.read_excel("Data_Train.xlsx")
df_train.head()
```

Out[2]:

	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 [3]:
```
df_testing=pd.read_excel("Test_set.xlsx")
df_test.head()
```

Out[3]:

	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 [4]:
```
df_train.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_train.shape
```

Out[5]:
```
(10683, 11)
```

In [6]:
```
df_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2671 entries, 0 to 2670
Data columns (total 10 columns):
#   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
0   Airline          2671 non-null   object
1   Date_of_Journey  2671 non-null   object
2   Source           2671 non-null   object
3   Destination      2671 non-null   object
4   Route            2671 non-null   object
5   Dep_Time         2671 non-null   object
6   Arrival_Time     2671 non-null   object
7   Duration         2671 non-null   object
8   Total_Stops      2671 non-null   object
9   Additional_Info  2671 non-null   object
dtypes: object(10)
memory usage: 208.8+ KB
```

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

Out[7]:
```
(2671, 10)
```

In [8]:
```
#Concating the two dataset.
df=df_train.append(df_test)
```

```
C:\Users\VIVEK PANDEY\AppData\Local\Temp\ipykernel_12532\3594449827.py:1: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
  df=df_train.append(df_test)
```

In [9]:
```
df
```

Out[9]:

	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
...	...	...	...	...	...	...	...	...	...	...	...
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	27/05/2019	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

13354 rows × 11 columns

In [10]:
```
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 [11]:
```
#Featuring Engineering  Process
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 [12]:
```
df.head()
```

Out[12]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year
0	IndiGo	24/03/2019	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	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662.0	1	05	2019
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	9	06	2019
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218.0	12	05	2019
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302.0	01	03	2019

In [13]:
```
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   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
11  Date             13354 non-null  object
12  Month            13354 non-null  object
13  Year             13354 non-null  object
dtypes: float64(1), object(13)
memory usage: 1.5+ MB
```

In [14]:
```
df['Date']=df['Date'].astype(int)
df['Month']=df['Month'].astype(int)
df['Year']=df['Year'].astype(int)
```

In [15]:
```
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   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
11  Date             13354 non-null  int32 
12  Month            13354 non-null  int32 
13  Year             13354 non-null  int32 
dtypes: float64(1), int32(3), object(10)
memory usage: 1.4+ MB
```

In [16]:
```
df.drop('Date_of_Journey',axis=1,inplace=True)
```

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

Out[17]:

	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

In [18]:
```
df['Arrival_Time']=df['Arrival_Time'].str.split(' ').str[0]
```

Out[18]:
```
0        01:10
1        13:15
2        04:25
3        23:30
4        21:35
         ...  
2666     20:25
2667     16:55
2668     04:25
2669     19:15
2670     19:15
Name: Arrival_Time, Length: 13354, dtype: object
```

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

In [20]:
```
df['Arrival_hour']=df['Arrival_Time'].str.split(':').str[0]
df['Arrival_min']=df['Arrival_Time'].str.split(':').str[1]
```

In [21]:
```
df.head()
```

Out[21]:

	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	01:10	2h 50m	non-stop	No info	3897.0	24	3	2019	01	10

In [22]:
```
df['Arrival_hour']=df['Arrival_hour'].astype(int)
df['Arrival_min']=df['Arrival_min'].astype(int)
```

In [23]:
```
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   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  Date             13354 non-null  int32 
11  Month            13354 non-null  int32 
12  Year             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: 1.6+ MB
```

In [24]:
```
df.drop('Arrival_Time',axis=1,inplace=True)
```

In [25]:
```
df.head(1)
```

Out[25]:

	Airline	Source	Destination	Route	Dep_Time	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	2h 50m	non-stop	No info	3897.0	24	3	2019	1	10

In [26]:
```
df['Dep_Time']=df['Dep_Time'].str.split(':').str[0]
df['Dep_hour']=df['Dep_Time'].str.split(':').str[1]
df['Dep_min']=df['Dep_Time'].str.split(':').str[2]
df.drop('Dep_Time',axis=1,inplace=True)
```

In [27]:
```
df.head(1)
```

Out[27]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	non-stop	No info	3897.0	24	3	2019	1	10	22	20

In [28]:
```
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      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_hour     13354 non-null  int32 
12  Arrival_min      13354 non-null  int32 
13  Dep_hour         13354 non-null  object
14  Dep_min          13354 non-null  object
dtypes: float64(1), int32(5), object(9)
memory usage: 1.7+ MB
```

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

Out[29]:
```
1
```

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

Out[30]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

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['Total_Stops'].isnull().sum()
```

Out[32]:
```
0
```

In [33]:
```
df.head(1)
```

Out[33]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min
9039	Air India	Delhi	Cochin	Nan	23h 40m	NaN	No info	7480.0	6	5	2019	9	25	9	45

In [34]:
```
df.head(2)
```

Out[34]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	0.0	No info	3897.0	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	2.0	No info	7662.0	1	5	2019	13	15	5	50

In [35]:
```
df.drop('Route',axis=1,inplace=True)
```

In [36]:
```
df.head(2)
```

Out[36]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min
0	IndiGo	Banglore	New Delhi	2h 50m	0.0	No info	3897.0	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	7h 25m	2.0	No info	7662.0	1	5	2019	13	15	5	50

In [37]:
```
df['Additional_Info'].unique()
```

Out[37]:
```
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 layovers'], dtype=object)
```

In [38]:
```
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   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   Date             13354 non-null  int32 
8   Month            13354 non-null  int32 
9   Year             13354 non-null  int32 
10  Arrival_hour     13354 non-null  int32 
11  Arrival_min      13354 non-null  int32 
12  Dep_hour         13354 non-null  object
13  Dep_min          13354 non-null  object
dtypes: float64(2), int32(5), object(7)
memory usage: 1.5+ MB
```

In [39]:
```
df['Duration_hour']=df['Duration'].str.split(' ').str[0].str.split('h').str[0]
```

In [40]:
```
df[df['Duration_hour']=='5m']
```

Out[40]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min	Duration_hour
6474	Air India	Mumbai	Hyderabad	5m	2.0	No info	17327.0	6	3	2019	16	15	16	10	5m
2660	Air India	Mumbai	Hyderabad	5m	2.0	No info	NaN	12	3	2019	16	15	16	10	5m

In [41]:
```
df.drop(6474,axis=0,inplace=True)
df.drop(2660,axis=0,inplace=True)
```

In [42]:
```
df.head(2)
```

Out[42]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min	Duration_hour
0	IndiGo	Banglore	New Delhi	2h 50m	0.0	No info	3897.0	24	3	2019	1	10	22	20	2
1	Air India	Kolkata	Banglore	7h 25m	2.0	No info	7662.0	1	5	2019	13	15	5	50	7

In [43]:
```
df['Duration_hour']=df['Duration_hour'].astype(int)
```

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

In [45]:
```
df.head(1)
```

Out[45]:

	Airline	Source	Destination	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min	Duration_hour
0	IndiGo	Banglore	New Delhi	0.0	No info	3897.0	24	3	2019	1	10	22	20	2

In [46]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13352 entries, 0 to 2670
Data columns (total 14 columns):
#   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
0   Airline          13352 non-null  object
1   Source           13352 non-null  object
2   Destination      13352 non-null  object
3   Total_Stops      13351 non-null  float64
4   Additional_Info  13352 non-null  object
5   Price            10681 non-null  float64
6   Date             13352 non-null  int32 
7   Month            13352 non-null  int32 
8   Year             13352 non-null  int32 
9   Arrival_hour     13352 non-null  int32 
10  Arrival_min      13352 non-null  int32 
11  Dep_hour         13352 non-null  object
12  Dep_min          13352 non-null  object
13  Duration_hour    13352 non-null  int32 
dtypes: float64(2), int32(6), object(6)
memory usage: 1.4+ MB
```

In [47]:
```
df.shape
```

Out[47]:
```
(13352, 14)
```

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

Out[48]:

	Airline	Source	Destination	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Dep_hour	Dep_min	Duration_hour
2666	Air India	Kolkata	Banglore	1.0	NaN	NaN	6	6	2019	20	25	20	30	23
2667	IndiGo	Kolkata	Banglore	0.0	NaN	NaN	27	3	2019	16	55	14	20	2
2668	Jet Airways	Delhi	Cochin	1.0	NaN	NaN	6	3	2019	4	25	21	50	6
2669	Air India	Delhi	Cochin	1.0	NaN	NaN	6	3	2019	19	15	4	0	15
2670	Multiple carriers	Delhi	Cochin	1.0	NaN	NaN	15	6	2019	19	15	4	55	14

In [49]:
```

```