In [250…

```
import pandas as pd
```

In [251…

```
import pandas as pd

df = pd.read_excel("flight_price (1).xlsx", engine="openpyxl")
print(df)
```

```
           Airline Date_of_Journey    Source Destination  \
0           IndiGo      24/03/2019  Banglore   New Delhi
1        Air India       1/05/2019   Kolkata    Banglore
2      Jet Airways       9/06/2019     Delhi      Cochin
3           IndiGo      12/05/2019   Kolkata    Banglore
4           IndiGo      01/03/2019  Banglore   New Delhi
...            ...             ...       ...         ...
10678     Air Asia       9/04/2019   Kolkata    Banglore
10679    Air India      27/04/2019   Kolkata    Banglore
10680  Jet Airways      27/04/2019  Banglore       Delhi
10681      Vistara      01/03/2019  Banglore   New Delhi
10682    Air India       9/05/2019     Delhi      Cochin

                       Route Dep_Time  Arrival_Time Duration Total_Stops  \
0                  BLR → DEL    22:20  01:10 22 Mar   2h 50m    non-stop
1      CCU → IXR → BBI → BLR    05:50         13:15   7h 25m     2 stops
2      DEL → LKO → BOM → COK    09:25  04:25 10 Jun      19h     2 stops
3            CCU → NAG → BLR    18:05         23:30   5h 25m      1 stop
4            BLR → NAG → DEL    16:50         21:35   4h 45m      1 stop
...                      ...      ...           ...      ...         ...
10678              CCU → BLR    19:55         22:25   2h 30m    non-stop
10679              CCU → BLR    20:45         23:20   2h 35m    non-stop
10680              BLR → DEL    08:20         11:20       3h    non-stop
10681              BLR → DEL    11:30         14:10   2h 40m    non-stop
10682  DEL → GOI → BOM → COK    10:55         19:15   8h 20m     2 stops

      Additional_Info  Price
0             No info   3897
1             No info   7662
2             No info  13882
3             No info   6218
4             No info  13302
...               ...    ...
10678         No info   4107
10679         No info   4145
10680         No info   7229
10681         No info  12648
10682         No info  11753

[10683 rows x 11 columns]
```

In [252…

```
df.head()
```

Out[252…

	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 [253…

```
df.tail()
```

Out[253…

	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 [254…

```
# get the basic info
```

In [255…

```
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 [256…

```
df.describe()
```

Out[256…

	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 [257…

```
df.head()
```

Out[257…

	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 [258…

```
#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 [259…

```
df.head()
```

Out[259…

	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 [260…

```
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 [261…

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

In [262…

```
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 [263…

```
df.drop('Date_of_Journey',axis=1,inplace=True)
```

In [264…

```
df.head()
```

Out[264...

	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 [265...

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

In [266...

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

In [267...

```
df.head()
```

Out[267…

	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 [268…

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

In [269…

```
df.drop('Arrival_Time',axis=1,inplace=True)
```

In [270…

```
df.head()
```

Out[270…

	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
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	7h 25m	2 stops	No
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	19h	2 stops	No
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	18:05	5h 25m	1 stop	No
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	16:50	4h 45m	1 stop	No

In [271…

```
df['Departure_hour']=df['Dep_Time'].str.split(':').str[0]
df['Departure_min']=df['Dep_Time'].str.split(':').str[1]
```

In [272…

```
df.head()
```

Out[272…

	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
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	7h 25m	2 stops	No
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	19h	2 stops	No
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	18:05	5h 25m	1 stop	No
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	16:50	4h 45m	1 stop	No

In [273…

```
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  Departure_hour   10683 non-null  object
 15  Departure_min    10683 non-null  object
dtypes: int64(6), object(10)
memory usage: 1.3+ MB
```

In [274…

```
df['Departure_hour']=df['Departure_hour'].astype(int)
df['Departure_min']=df['Departure_min'].astype(int)
```

In [275…

```
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  Departure_hour   10683 non-null  int64 
 15  Departure_min    10683 non-null  int64 
dtypes: int64(8), object(8)
memory usage: 1.3+ MB
```

In [276…

```
df.drop('Dep_Time',axis=1,inplace=True)
```

In [277…

```
df.head(2)
```

Out[277…

	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 [278…

```
df['Total_Stops'].unique()
```

Out[278…

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

In [279…

```
import numpy as np
```

In [280…

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

In [281…

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

Out[281…

Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Da

In [282…

```
df.drop('Route',axis=1,inplace=True)
```

In [283…

```
df.head(10)
```

Out[283…

	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
5	SpiceJet	Kolkata	Banglore	2h 25m	0	No info	3873	24
6	Jet Airways	Banglore	New Delhi	15h 30m	1	In-flight meal not included	11087	12
7	Jet Airways	Banglore	New Delhi	21h 5m	1	No info	22270	1
8	Jet Airways	Banglore	New Delhi	25h 30m	1	In-flight meal not included	11087	12
9	Multiple carriers	Delhi	Cochin	7h 50m	1	No info	8625	27

In [284…

```
df['Duration_hour']=df['Duration'].str.split(':').str[0]
df['Duration_min']=df['Duration'].str.split(':').str[1]
```

In [286…

```
df['Duration_hour']=df['Duration'].apply(lambda x:x.split(' ')[0])
```

In [290…

```
df.head(2)
```

Out[290…

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	M
0	IndiGo	Banglore	New Delhi	2h 50m	0	No info	3897	24	
1	Air India	Kolkata	Banglore	7h 25m	2	No info	7662	1	

In [291…

```
df.drop('Duration',axis=1,inplace=True)
```

In [292…

```
df.head(2)
```

Out[292…

	Airline	Source	Destination	Total_Stops	Additional_Info	Price	Date	Month	Yea
0	IndiGo	Banglore	New Delhi	0	No info	3897	24	3	201
1	Air India	Kolkata	Banglore	2	No info	7662	1	5	201

In [293…

```
df['Airline'].unique()
```

Out[293…

```
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 [294…

```
df['Source'].unique()
```

Out[294…

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

In [295…

```
df['Additional_Info'].unique()
```

Out[295…

```
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 [296…

```
from sklearn.preprocessing import OneHotEncoder
```

In [297…

```
encoder=OneHotEncoder()
```

In [299…

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

Out[299…

```
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.]])
```

In [300…

```
pd.DataFrame(encoder.fit_transform(df[['Airline' , 'Source', 'Destination']]).tc
```

Out[300…

	Airline_Air Asia	Airline_Air India	Airline_GoAir	Airline_IndiGo	Airline_Jet Airways	Airline_Jet Airways Business	Airl
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 [ ]: