6.2-EDA_On_Flight_Price_Prediciton

July 17, 2025

0.1 EDA on Flight price prediction

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

[2]:
```
# !pip install openpyxl
```

[3]:
```
df = pd.read_excel('../0-Dataset/flight_price.xlsx')
df.head()
```

[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

[4]:
```
## Basic info about my 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
```

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

[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
```

[6]:
```
## Feature Engineering
df['Date'] = (df['Date_of_Journey'].str.split('/').str[0]).astype('Int64')
df['Month'] = (df['Date_of_Journey'].str.split('/').str[1]).astype('Int64')
df['Year'] = (df['Date_of_Journey'].str.split('/').str[2]).astype('Int64')
```

[7]:
```
df.head()
```

[7]:

	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	24	3	2019
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662	1	5	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	9	6	2019
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218	12	5	2019
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302	1	3	2019

```
3      5  2019
4      3  2019
```

[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  Int64 
 12  Month            10683 non-null  Int64 
 13  Year             10683 non-null  Int64 
dtypes: Int64(3), int64(1), object(10)
memory usage: 1.2+ MB
```

[9]:
```
## Drop the column
df.drop('Date_of_Journey', axis=1, inplace=True)
```

[10]:
```
df['Arrival_Time']= df['Arrival_Time'].apply(lambda x:x.split(' ')[0])
df['Arrival_Time'].head()
```

[10]:
```
0    01:10
1    13:15
2    04:25
3    23:30
4    21:35
Name: Arrival_Time, dtype: object
```

[11]:
```
df['Arrival_hour'] = df['Arrival_Time'].str.split(':').str[0].astype('Int64')
df['Arrival_min'] = df['Arrival_Time'].str.split(':').str[1].astype('Int64')
```

[12]:
```
df.drop('Arrival_Time', axis=1, inplace=True)
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   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
dtypes: Int64(5), int64(1), object(8)
memory usage: 1.2+ MB
```

[13]:
```
df['Departure_hour']= df['Dep_Time'].str.split(':').str[0].astype('Int64')
df['Departure_min'] = df['Dep_Time'].str.split(':').str[1].astype('Int64')
```

[14]:
```
df.drop('Dep_Time', axis=1, inplace=True)
```

[15]:
```
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   Duration         10683 non-null  object
 5   Total_Stops      10682 non-null  object
 6   Additional_Info  10683 non-null  object
 7   Price            10683 non-null  int64
 8   Date             10683 non-null  Int64
 9   Month            10683 non-null  Int64
 10  Year             10683 non-null  Int64
 11  Arrival_hour     10683 non-null  Int64
 12  Arrival_min      10683 non-null  Int64
 13  Departure_hour   10683 non-null  Int64
 14  Departure_min    10683 non-null  Int64
dtypes: Int64(7), int64(1), object(7)
memory usage: 1.3+ MB
```

[16]:
```
df['Total_Stops'].unique()
```

[16]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

[17]:
```
df['Total_Stops'].mode()
```

[17]:
```
0    1 stop
Name: Total_Stops, dtype: object
```

[18]:
```
df['Total_Stops'] = df['Total_Stops'].map({'non-stop':0, '1 stops':1, '2 stop':
↪2, '3 stops':3, '4 stops':4})
```

[19]:
```
df['Total_Stops']=df['Total_Stops'].fillna(0)
```

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

[20]:

	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Departure_hour	Departure_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	0.0	No info	3897	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	7h 25m	0.0	No info	7662	1	5	2019	13	15	5	50

[21]:
```
df.drop('Route', axis=1, inplace=True)
```

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

[22]:

	Airline	Source	Destination	Duration	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	Arrival_min	Departure_hour	Departure_min
0	IndiGo	Banglore	New Delhi	2h 50m	0.0	No info	3897	24	3	2019	1	10	22	20
1	Air India	Kolkata	Banglore	7h 25m	0.0	No info	7662	1	5	2019	13	15	5	50

[23]:
```
df['Duration_hr'] = df['Duration'].str.split(' ').str[0].str.split('h').str[0].
↪fillna(0)
```

```
df['Duration_min'] = df['Duration'].str.split(' ').str[1].str.split('m').str[0].
 ↪fillna(0)
```

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

[24]:
```
       Airline    Source Destination Duration  Total_Stops Additional_Info  \
0       IndiGo  Banglore   New Delhi   2h 50m          0.0         No info
1    Air India   Kolkata    Banglore   7h 25m          0.0         No info
2  Jet Airways     Delhi      Cochin      19h          0.0         No info
3       IndiGo   Kolkata    Banglore   5h 25m          0.0         No info
4       IndiGo  Banglore   New Delhi   4h 45m          0.0         No info

   Price  Date  Month  Year  Arrival_hour  Arrival_min  Departure_hour  \
0   3897    24      3  2019             1           10              22
1   7662     1      5  2019            13           15               5
2  13882     9      6  2019             4           25               9
3   6218    12      5  2019            23           30              18
4  13302     1      3  2019            21           35              16

   Departure_min Duration_hr Duration_min
0             20           2           50
1             50           7           25
2             25          19            0
3              5           5           25
4             50           4           45
```

[25]:
```
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   Duration         10683 non-null  object
 4   Total_Stops      10683 non-null  float64
 5   Additional_Info  10683 non-null  object
 6   Price            10683 non-null  int64
 7   Date             10683 non-null  Int64
 8   Month            10683 non-null  Int64
 9   Year             10683 non-null  Int64
 10  Arrival_hour     10683 non-null  Int64
 11  Arrival_min      10683 non-null  Int64
 12  Departure_hour   10683 non-null  Int64
 13  Departure_min    10683 non-null  Int64
 14  Duration_hr      10683 non-null  object
```

```
 15  Duration_min     10683 non-null  object
dtypes: Int64(7), float64(1), int64(1), object(7)
memory usage: 1.4+ MB
```

[26]:
```
df.drop('Duration', axis=1, inplace=True)
```

[27]:
```
df['Duration_min'] = df['Duration_min'].fillna(0).astype(int)
df['Duration_hr'] = df['Duration_hr'].str.extract('(\d+)').fillna(0).astype(int)
```

[28]:
```
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   Total_Stops      10683 non-null  float64
 4   Additional_Info  10683 non-null  object
 5   Price            10683 non-null  int64
 6   Date             10683 non-null  Int64
 7   Month            10683 non-null  Int64
 8   Year             10683 non-null  Int64
 9   Arrival_hour     10683 non-null  Int64
 10  Arrival_min      10683 non-null  Int64
 11  Departure_hour   10683 non-null  Int64
 12  Departure_min    10683 non-null  Int64
 13  Duration_hr      10683 non-null  int32
 14  Duration_min     10683 non-null  int32
dtypes: Int64(7), float64(1), int32(2), int64(1), object(4)
memory usage: 1.2+ MB
```

[29]:
```
df.head(10)
```

[29]:

	Airline	Source	Destination	Total_Stops	\
0	IndiGo	Banglore	New Delhi	0.0	
1	Air India	Kolkata	Banglore	0.0	
2	Jet Airways	Delhi	Cochin	0.0	
3	IndiGo	Kolkata	Banglore	0.0	
4	IndiGo	Banglore	New Delhi	0.0	
5	SpiceJet	Kolkata	Banglore	0.0	
6	Jet Airways	Banglore	New Delhi	0.0	
7	Jet Airways	Banglore	New Delhi	0.0	
8	Jet Airways	Banglore	New Delhi	0.0	
9	Multiple carriers	Delhi	Cochin	0.0	

	Additional_Info	Price	Date	Month	Year	Arrival_hour	\
0	No info	3897	24	3	2019	1	
1	No info	7662	1	5	2019	13	
2	No info	13882	9	6	2019	4	
3	No info	6218	12	5	2019	23	
4	No info	13302	1	3	2019	21	
5	No info	3873	24	6	2019	11	
6	In-flight meal not included	11087	12	3	2019	10	
7	No info	22270	1	3	2019	5	
8	In-flight meal not included	11087	12	3	2019	10	
9	No info	8625	27	5	2019	19	

	Arrival_min	Departure_hour	Departure_min	Duration_hr	Duration_min
0	10	22	20	2	50
1	15	5	50	7	25
2	25	9	25	19	0
3	30	18	5	5	25
4	35	16	50	4	45
5	25	9	0	2	25
6	25	18	55	15	30
7	5	8	0	21	5
8	25	8	55	25	30
9	15	11	25	7	50

[30]:
```
df['Airline'].unique()
```

[30]:
```
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)
```

[31]:
```
df['Additional_Info'].unique()
```

[31]:
```
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)
```

[32]:
```
df['Destination'].unique()
```

[32]:
```
array(['New Delhi', 'Banglore', 'Cochin', 'Kolkata', 'Delhi', 'Hyderabad'],
      dtype=object)
```

[33]:
```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder()
```

[34]:
```
encoded_value = encoder.fit_transform(df[['Airline', 'Source', 'Destination']]).
 ↪toarray()
encoded_value
```

[34]:
```
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.]])
```

[35]:
```
up_coded = pd.DataFrame(encoded_value, columns=encoder.get_feature_names_out())
up_coded
```

[35]:

	Airline_Air Asia	Airline_Air India	Airline_GoAir	Airline_IndiGo	\
0	0.0	0.0	0.0	1.0	
1	0.0	1.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	1.0	
4	0.0	0.0	0.0	1.0	
…	…	…	…	…	
10678	1.0	0.0	0.0	0.0	
10679	0.0	1.0	0.0	0.0	
10680	0.0	0.0	0.0	0.0	
10681	0.0	0.0	0.0	0.0	
10682	0.0	1.0	0.0	0.0	

	Airline_Jet Airways	Airline_Jet Airways Business	\
0	0.0	0.0	
1	0.0	0.0	
2	1.0	0.0	
3	0.0	0.0	
4	0.0	0.0	
…	…	…	
10678	0.0	0.0	
10679	0.0	0.0	
10680	1.0	0.0	
10681	0.0	0.0	
10682	0.0	0.0	

	Airline_Multiple carriers	Airline_Multiple carriers Premium economy	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	
3	0.0	0.0	
4	0.0	0.0	

```
…                          …                                          …
10678                        0.0                                        0.0
10679                        0.0                                        0.0
10680                        0.0                                        0.0
10681                        0.0                                        0.0
10682                        0.0                                        0.0

       Airline_SpiceJet  Airline_Trujet  …  Source_Chennai  Source_Delhi  \
0                   0.0             0.0  …             0.0           0.0
1                   0.0             0.0  …             0.0           0.0
2                   0.0             0.0  …             0.0           1.0
3                   0.0             0.0  …             0.0           0.0
4                   0.0             0.0  …             0.0           0.0
…                   …               …  …               …           …
10678               0.0             0.0  …             0.0           0.0
10679               0.0             0.0  …             0.0           0.0
10680               0.0             0.0  …             0.0           0.0
10681               0.0             0.0  …             0.0           0.0
10682               0.0             0.0  …             0.0           1.0

       Source_Kolkata  Source_Mumbai  Destination_Banglore  \
0                 0.0            0.0                   0.0
1                 1.0            0.0                   1.0
2                 0.0            0.0                   0.0
3                 1.0            0.0                   1.0
4                 0.0            0.0                   0.0
…                 …              …                     …
10678             1.0            0.0                   1.0
10679             1.0            0.0                   1.0
10680             0.0            0.0                   0.0
10681             0.0            0.0                   0.0
10682             0.0            0.0                   0.0

       Destination_Cochin  Destination_Delhi  Destination_Hyderabad  \
0                     0.0                0.0                    0.0
1                     0.0                0.0                    0.0
2                     1.0                0.0                    0.0
3                     0.0                0.0                    0.0
4                     0.0                0.0                    0.0
…                     …                  …                      …
10678                 0.0                0.0                    0.0
10679                 0.0                0.0                    0.0
10680                 0.0                1.0                    0.0
10681                 0.0                0.0                    0.0
10682                 1.0                0.0                    0.0

       Destination_Kolkata  Destination_New Delhi
```

```
0                     0.0                     1.0
1                     0.0                     0.0
2                     0.0                     0.0
3                     0.0                     0.0
4                     0.0                     1.0
…                     …                       …
10678                 0.0                     0.0
10679                 0.0                     0.0
10680                 0.0                     0.0
10681                 0.0                     1.0
10682                 0.0                     0.0

[10683 rows x 23 columns]
```

[36]:
```
df = pd.concat([df, up_coded], axis=1)
```

[37]:
```
df.drop(['Airline', 'Source', 'Destination'], axis=1, inplace= True)
```

[38]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 35 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Total_Stops                                 10683 non-null  float64
 1   Additional_Info                             10683 non-null  object
 2   Price                                       10683 non-null  int64
 3   Date                                        10683 non-null  Int64
 4   Month                                       10683 non-null  Int64
 5   Year                                        10683 non-null  Int64
 6   Arrival_hour                                10683 non-null  Int64
 7   Arrival_min                                 10683 non-null  Int64
 8   Departure_hour                              10683 non-null  Int64
 9   Departure_min                               10683 non-null  Int64
 10  Duration_hr                                 10683 non-null  int32
 11  Duration_min                                10683 non-null  int32
 12  Airline_Air Asia                            10683 non-null  float64
 13  Airline_Air India                           10683 non-null  float64
 14  Airline_GoAir                               10683 non-null  float64
 15  Airline_IndiGo                              10683 non-null  float64
 16  Airline_Jet Airways                         10683 non-null  float64
 17  Airline_Jet Airways Business                10683 non-null  float64
 18  Airline_Multiple carriers                   10683 non-null  float64
 19  Airline_Multiple carriers Premium economy   10683 non-null  float64
 20  Airline_SpiceJet                            10683 non-null  float64
 21  Airline_Trujet                              10683 non-null  float64
 22  Airline_Vistara                             10683 non-null  float64
```

```
 23  Airline_Vistara Premium economy            10683 non-null  float64
 24  Source_Banglore                            10683 non-null  float64
 25  Source_Chennai                             10683 non-null  float64
 26  Source_Delhi                               10683 non-null  float64
 27  Source_Kolkata                             10683 non-null  float64
 28  Source_Mumbai                              10683 non-null  float64
 29  Destination_Banglore                       10683 non-null  float64
 30  Destination_Cochin                         10683 non-null  float64
 31  Destination_Delhi                          10683 non-null  float64
 32  Destination_Hyderabad                      10683 non-null  float64
 33  Destination_Kolkata                        10683 non-null  float64
 34  Destination_New Delhi                      10683 non-null  float64
dtypes: Int64(7), float64(24), int32(2), int64(1), object(1)
memory usage: 2.8+ MB
```

[39]:
```
df.head()
```

[39]:

	Total_Stops	Additional_Info	Price	Date	Month	Year	Arrival_hour	\
0	0.0	No info	3897	24	3	2019	1	
1	0.0	No info	7662	1	5	2019	13	
2	0.0	No info	13882	9	6	2019	4	
3	0.0	No info	6218	12	5	2019	23	
4	0.0	No info	13302	1	3	2019	21	

	Arrival_min	Departure_hour	Departure_min	…	Source_Chennai	\
0	10	22	20	…	0.0	
1	15	5	50	…	0.0	
2	25	9	25	…	0.0	
3	30	18	5	…	0.0	
4	35	16	50	…	0.0	

	Source_Delhi	Source_Kolkata	Source_Mumbai	Destination_Banglore	\
0	0.0	0.0	0.0	0.0	
1	0.0	1.0	0.0	1.0	
2	1.0	0.0	0.0	0.0	
3	0.0	1.0	0.0	1.0	
4	0.0	0.0	0.0	0.0	

	Destination_Cochin	Destination_Delhi	Destination_Hyderabad	\
0	0.0	0.0	0.0	
1	0.0	0.0	0.0	
2	1.0	0.0	0.0	
3	0.0	0.0	0.0	
4	0.0	0.0	0.0	

	Destination_Kolkata	Destination_New Delhi
0	0.0	1.0

```
1                0.0                0.0
2                0.0                0.0
3                0.0                0.0
4                0.0                1.0

[5 rows x 35 columns]
```