Flight Price EDA

March 24, 2023

1 Flight Price Feature Engineering & EDA

1.1 Import requirments

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

1.2 Import the DataSet

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

[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

1.2.1 summarize the data

[3]:
```
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
```

1.2.2 Check columns

[4]:
```
df.columns
```

[4]:
```
Index(['Airline', 'Date_of_Journey', 'Source', 'Destination', 'Route',
       'Dep_Time', 'Arrival_Time', 'Duration', 'Total_Stops',
       'Additional_Info', 'Price'],
      dtype='object')
```

1.2.3 Split the Date_of_Journey column

[5]:
```
df["Date_of_Journey"]
```

[5]:
```
0        24/03/2019
1         1/05/2019
2         9/06/2019
3        12/05/2019
4        01/03/2019
            ...
10678     9/04/2019
10679    27/04/2019
10680    27/04/2019
10681    01/03/2019
10682     9/05/2019
Name: Date_of_Journey, Length: 10683, dtype: object
```

[6]:
```
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]
```

[7]:
```
df
```

[7]:
```
            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 Date Month  Year
0             No info   3897   24    03  2019
1             No info   7662    1    05  2019
2             No info  13882    9    06  2019
3             No info   6218   12    05  2019
4             No info  13302   01    03  2019
…                …    …  …   …    …
10678         No info   4107    9    04  2019
10679         No info   4145   27    04  2019
10680         No info   7229   27    04  2019
10681         No info  12648   01    03  2019
10682         No info  11753    9    05  2019

[10683 rows x 14 columns]
```

Now drop the "Date_of_Journey" column

[8]:
```
df.drop("Date_of_Journey",axis=1,inplace=True)
```

[9]:
```
df
```

[9]:
```
            Airline    Source Destination                  Route Dep_Time  \
0            IndiGo  Banglore   New Delhi              BLR → DEL    22:20
1         Air India   Kolkata    Banglore  CCU → IXR → BBI → BLR    05:50
2       Jet Airways     Delhi      Cochin  DEL → LKO → BOM → COK    09:25
3            IndiGo   Kolkata    Banglore        CCU → NAG → BLR    18:05
4            IndiGo  Banglore   New Delhi        BLR → NAG → DEL    16:50
…               …       …          …                     …       …
10678     Air Asia   Kolkata    Banglore              CCU → BLR    19:55
10679    Air India   Kolkata    Banglore              CCU → BLR    20:45
10680  Jet Airways  Banglore       Delhi              BLR → DEL    08:20
10681      Vistara  Banglore   New Delhi              BLR → DEL    11:30
10682    Air India     Delhi      Cochin  DEL → GOI → BOM → COK    10:55

       Arrival_Time Duration Total_Stops Additional_Info  Price Date Month  \
0      01:10 22 Mar   2h 50m    non-stop         No info   3897   24    03
1             13:15   7h 25m     2 stops         No info   7662    1    05
2      04:25 10 Jun      19h     2 stops         No info  13882    9    06
3             23:30   5h 25m      1 stop         No info   6218   12    05
4             21:35   4h 45m      1 stop         No info  13302   01    03
…               …       …          …              …     …   …   …
10678         22:25   2h 30m    non-stop         No info   4107    9    04
10679         23:20   2h 35m    non-stop         No info   4145   27    04
10680         11:20       3h    non-stop         No info   7229   27    04
10681         14:10   2h 40m    non-stop         No info  12648   01    03
10682         19:15   8h 20m     2 stops         No info  11753    9    05

       Year
0      2019
1      2019
2      2019
3      2019
4      2019
…       …
10678  2019
10679  2019
10680  2019
10681  2019
10682  2019

[10683 rows x 13 columns]
```

1.2.4 As we already have Source and Destinations, Drop the Route column

[10]:
```
df.drop("Route",axis=1,inplace=True)
```

[11]:
```
df
```

[11]:
```
            Airline    Source Destination Dep_Time  Arrival_Time Duration  \
0            IndiGo  Banglore   New Delhi    22:20  01:10 22 Mar   2h 50m
1         Air India   Kolkata    Banglore    05:50         13:15   7h 25m
2       Jet Airways     Delhi      Cochin    09:25  04:25 10 Jun      19h
3            IndiGo   Kolkata    Banglore    18:05         23:30   5h 25m
4            IndiGo  Banglore   New Delhi    16:50         21:35   4h 45m
…              …       …          …      …           …       …
10678     Air Asia   Kolkata    Banglore    19:55         22:25   2h 30m
10679    Air India   Kolkata    Banglore    20:45         23:20   2h 35m
10680  Jet Airways  Banglore       Delhi    08:20         11:20      3h
10681      Vistara  Banglore   New Delhi    11:30         14:10   2h 40m
10682    Air India     Delhi      Cochin    10:55         19:15   8h 20m

      Total_Stops Additional_Info  Price Date Month  Year
0        non-stop         No info   3897   24    03  2019
1         2 stops         No info   7662    1    05  2019
2         2 stops         No info  13882    9    06  2019
3          1 stop         No info   6218   12    05  2019
4          1 stop         No info  13302   01    03  2019
…             …              …    …  …   …    …
10678    non-stop         No info   4107    9    04  2019
10679    non-stop         No info   4145   27    04  2019
10680    non-stop         No info   7229   27    04  2019
10681    non-stop         No info  12648   01    03  2019
10682     2 stops         No info  11753    9    05  2019

[10683 rows x 12 columns]
```

1.2.5 Split Dep_Time column

[12]:
```
df["Dep_Hour"]=df["Dep_Time"].str.split(":").str[0]
df["Dep_Minute"]=df["Dep_Time"].str.split(":").str[1]
```

[13]:
```
df
```

[13]:
```
            Airline    Source Destination Dep_Time  Arrival_Time Duration  \
0            IndiGo  Banglore   New Delhi    22:20  01:10 22 Mar   2h 50m
1         Air India   Kolkata    Banglore    05:50         13:15   7h 25m
2       Jet Airways     Delhi      Cochin    09:25  04:25 10 Jun      19h
3            IndiGo   Kolkata    Banglore    18:05         23:30   5h 25m
4            IndiGo  Banglore   New Delhi    16:50         21:35   4h 45m
…              …       …          …      …           …       …
10678     Air Asia   Kolkata    Banglore    19:55         22:25   2h 30m
10679    Air India   Kolkata    Banglore    20:45         23:20   2h 35m
10680  Jet Airways  Banglore       Delhi    08:20         11:20      3h
10681      Vistara  Banglore   New Delhi    11:30         14:10   2h 40m
10682    Air India     Delhi      Cochin    10:55         19:15   8h 20m
```

```
      Total_Stops Additional_Info  Price Date Month  Year Dep_Hour Dep_Minute
0        non-stop         No info   3897   24    03  2019       22          20
1         2 stops         No info   7662    1    05  2019       05          50
2         2 stops         No info  13882    9    06  2019       09          25
3          1 stop         No info   6218   12    05  2019       18          05
4          1 stop         No info  13302   01    03  2019       16          50
…             …               …    …  …   …   …      …        …
10678    non-stop         No info   4107    9    04  2019       19          55
10679    non-stop         No info   4145   27    04  2019       20          45
10680    non-stop         No info   7229   27    04  2019       08          20
10681    non-stop         No info  12648   01    03  2019       11          30
10682     2 stops         No info  11753    9    05  2019       10          55

[10683 rows x 14 columns]
```

Now, as we have seperate columns for Dep Hour and Dep Minute, Drop Dep_Time

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

[15]:
```
df
```

[15]:
```
            Airline    Source Destination  Arrival_Time Duration Total_Stops  \
0            IndiGo  Banglore   New Delhi  01:10 22 Mar   2h 50m    non-stop
1        Air India   Kolkata    Banglore         13:15   7h 25m     2 stops
2      Jet Airways     Delhi      Cochin  04:25 10 Jun      19h     2 stops
3            IndiGo   Kolkata    Banglore         23:30   5h 25m      1 stop
4            IndiGo  Banglore   New Delhi         21:35   4h 45m      1 stop
…               …       …          …            …       …          …
10678     Air Asia   Kolkata    Banglore         22:25   2h 30m    non-stop
10679    Air India   Kolkata    Banglore         23:20   2h 35m    non-stop
10680  Jet Airways  Banglore       Delhi         11:20       3h    non-stop
10681      Vistara  Banglore   New Delhi         14:10   2h 40m    non-stop
10682    Air India     Delhi      Cochin         19:15   8h 20m     2 stops

      Additional_Info  Price Date Month  Year Dep_Hour Dep_Minute
0             No info   3897   24    03  2019       22          20
1             No info   7662    1    05  2019       05          50
2             No info  13882    9    06  2019       09          25
3             No info   6218   12    05  2019       18          05
4             No info  13302   01    03  2019       16          50
…                 …    …  …   …   …      …        …
10678         No info   4107    9    04  2019       19          55
10679         No info   4145   27    04  2019       20          45
10680         No info   7229   27    04  2019       08          20
10681         No info  12648   01    03  2019       11          30
10682         No info  11753    9    05  2019       10          55
```

[10683 rows x 13 columns]

1.2.6 Now split the Arrival_Time column

[16]:
```
df["Arrival_Time"]=df["Arrival_Time"].str.split(" ").str[0]
```

[17]:
```
df
```

[17]:

	Airline	Source	Destination	Arrival_Time	Duration	Total_Stops	\
0	IndiGo	Banglore	New Delhi	01:10	2h 50m	non-stop	
1	Air India	Kolkata	Banglore	13:15	7h 25m	2 stops	
2	Jet Airways	Delhi	Cochin	04:25	19h	2 stops	
3	IndiGo	Kolkata	Banglore	23:30	5h 25m	1 stop	
4	IndiGo	Banglore	New Delhi	21:35	4h 45m	1 stop	
…	…	…	…	…	…	…	
10678	Air Asia	Kolkata	Banglore	22:25	2h 30m	non-stop	
10679	Air India	Kolkata	Banglore	23:20	2h 35m	non-stop	
10680	Jet Airways	Banglore	Delhi	11:20	3h	non-stop	
10681	Vistara	Banglore	New Delhi	14:10	2h 40m	non-stop	
10682	Air India	Delhi	Cochin	19:15	8h 20m	2 stops	

	Additional_Info	Price	Date	Month	Year	Dep_Hour	Dep_Minute
0	No info	3897	24	03	2019	22	20
1	No info	7662	1	05	2019	05	50
2	No info	13882	9	06	2019	09	25
3	No info	6218	12	05	2019	18	05
4	No info	13302	01	03	2019	16	50
…	…	…	…	…	…	…	
10678	No info	4107	9	04	2019	19	55
10679	No info	4145	27	04	2019	20	45
10680	No info	7229	27	04	2019	08	20
10681	No info	12648	01	03	2019	11	30
10682	No info	11753	9	05	2019	10	55

[10683 rows x 13 columns]

[18]:
```
df["Arrival_Hour"]=df["Arrival_Time"].str.split(":").str[0]
df["Arrival_Minute"]=df["Arrival_Time"].str.split(":").str[1]
```

[19]:
```
df
```

[19]:

	Airline	Source	Destination	Arrival_Time	Duration	Total_Stops	\
0	IndiGo	Banglore	New Delhi	01:10	2h 50m	non-stop	
1	Air India	Kolkata	Banglore	13:15	7h 25m	2 stops	
2	Jet Airways	Delhi	Cochin	04:25	19h	2 stops	
3	IndiGo	Kolkata	Banglore	23:30	5h 25m	1 stop	

```
4           IndiGo  Banglore   New Delhi        21:35   4h 45m      1 stop
…              …       …          …          …      …         …
10678     Air Asia   Kolkata    Banglore        22:25   2h 30m    non-stop
10679    Air India   Kolkata    Banglore        23:20   2h 35m    non-stop
10680  Jet Airways  Banglore       Delhi        11:20      3h     non-stop
10681      Vistara  Banglore   New Delhi        14:10   2h 40m    non-stop
10682    Air India     Delhi      Cochin        19:15   8h 20m     2 stops

      Additional_Info  Price Date Month  Year Dep_Hour Dep_Minute  \
0             No info   3897   24    03  2019       22         20
1             No info   7662    1    05  2019       05         50
2             No info  13882    9    06  2019       09         25
3             No info   6218   12    05  2019       18         05
4             No info  13302   01    03  2019       16         50
…                …    …  …   …   …      …        …
10678         No info   4107    9    04  2019       19         55
10679         No info   4145   27    04  2019       20         45
10680         No info   7229   27    04  2019       08         20
10681         No info  12648   01    03  2019       11         30
10682         No info  11753    9    05  2019       10         55

      Arrival_Hour Arrival_Minute
0               01             10
1               13             15
2               04             25
3               23             30
4               21             35
…              …              …
10678           22             25
10679           23             20
10680           11             20
10681           14             10
10682           19             15

[10683 rows x 15 columns]
```

Now drop the Arrival_Time column

[20]:
```
df.drop("Arrival_Time",axis=1,inplace=True)
```

[21]:
```
df
```

[21]:
```
           Airline    Source Destination Duration Total_Stops Additional_Info  \
0           IndiGo  Banglore   New Delhi   2h 50m    non-stop         No info
1        Air India   Kolkata    Banglore   7h 25m     2 stops         No info
2      Jet Airways     Delhi      Cochin      19h     2 stops         No info
3           IndiGo   Kolkata    Banglore   5h 25m      1 stop         No info
```

```
4           IndiGo  Banglore   New Delhi   4h 45m      1 stop         No info
…              …       …          …      …          …              …
10678     Air Asia   Kolkata    Banglore   2h 30m    non-stop         No info
10679    Air India   Kolkata    Banglore   2h 35m    non-stop         No info
10680  Jet Airways  Banglore       Delhi       3h    non-stop         No info
10681      Vistara  Banglore   New Delhi   2h 40m    non-stop         No info
10682    Air India     Delhi      Cochin   8h 20m     2 stops         No info

       Price Date Month  Year Dep_Hour Dep_Minute Arrival_Hour Arrival_Minute
0       3897   24    03  2019       22         20           01             10
1       7662    1    05  2019       05         50           13             15
2      13882    9    06  2019       09         25           04             25
3       6218   12    05  2019       18         05           23             30
4      13302   01    03  2019       16         50           21             35
…      …  …    …   …       …          …           …              …
10678   4107    9    04  2019       19         55           22             25
10679   4145   27    04  2019       20         45           23             20
10680   7229   27    04  2019       08         20           11             20
10681  12648   01    03  2019       11         30           14             10
10682  11753    9    05  2019       10         55           19             15

[10683 rows x 14 columns]
```

1.2.7 Now let's convert Duration into minutes

[22]:
```
df["Duration_Hours"]=df["Duration"].str.split("h").str[0]
df["Duration_Minutes"]=df["Duration"].str.split("h").str[1].str.split("m").
  ↪str[0]
```

[23]:
```
df
```

[23]:
```
           Airline    Source Destination Duration Total_Stops Additional_Info  \
0           IndiGo  Banglore   New Delhi   2h 50m    non-stop         No info
1        Air India   Kolkata    Banglore   7h 25m     2 stops         No info
2      Jet Airways     Delhi      Cochin      19h     2 stops         No info
3           IndiGo   Kolkata    Banglore   5h 25m      1 stop         No info
4           IndiGo  Banglore   New Delhi   4h 45m      1 stop         No info
…              …       …          …      …          …              …
10678     Air Asia   Kolkata    Banglore   2h 30m    non-stop         No info
10679    Air India   Kolkata    Banglore   2h 35m    non-stop         No info
10680  Jet Airways  Banglore       Delhi       3h    non-stop         No info
10681      Vistara  Banglore   New Delhi   2h 40m    non-stop         No info
10682    Air India     Delhi      Cochin   8h 20m     2 stops         No info

       Price Date Month  Year Dep_Hour Dep_Minute Arrival_Hour Arrival_Minute  \
0       3897   24    03  2019       22         20           01             10
1       7662    1    05  2019       05         50           13             15
```

```
2       13882    9    06  2019        09          25            04              25
3        6218   12    05  2019        18          05            23              30
4       13302   01    03  2019        16          50            21              35
…         …  …    …   …     …         …           …             …
10678    4107    9    04  2019        19          55            22              25
10679    4145   27    04  2019        20          45            23              20
10680    7229   27    04  2019        08          20            11              20
10681   12648   01    03  2019        11          30            14              10
10682   11753    9    05  2019        10          55            19              15

      Duration_Hours Duration_Minutes
0                  2               50
1                  7               25
2                 19
3                  5               25
4                  4               45
…                …                 …
10678              2               30
10679              2               35
10680              3
10681              2               40
10682              8               20

[10683 rows x 16 columns]
```

[24]:
```
def func(x):
    if x=="":
        return 0
    else:
        return x
```

[25]:
```
df["Duration_Minutes"]=df["Duration_Minutes"].apply(func)
```

[26]:
```
df=df[df["Duration_Hours"]!="5m"]
```

[27]:
```
df["Duration_Hours"]=df["Duration_Hours"].astype(int)
df["Duration_Minutes"]=df["Duration_Minutes"].astype(int)
```

[28]:
```
df
```

[28]:
```
            Airline    Source Destination Duration Total_Stops Additional_Info  \
0            IndiGo  Banglore   New Delhi   2h 50m    non-stop         No info
1         Air India   Kolkata    Banglore   7h 25m     2 stops         No info
2       Jet Airways     Delhi      Cochin      19h     2 stops         No info
3            IndiGo   Kolkata    Banglore   5h 25m      1 stop         No info
4            IndiGo  Banglore   New Delhi   4h 45m      1 stop         No info
…              …        …          …         …          …              …
```

```
10678     Air Asia   Kolkata    Banglore   2h 30m    non-stop         No info
10679    Air India   Kolkata    Banglore   2h 35m    non-stop         No info
10680  Jet Airways  Banglore       Delhi       3h    non-stop         No info
10681      Vistara  Banglore   New Delhi   2h 40m    non-stop         No info
10682    Air India     Delhi      Cochin   8h 20m     2 stops         No info

       Price Date Month  Year Dep_Hour Dep_Minute Arrival_Hour Arrival_Minute  \
0       3897   24    03  2019       22         20           01             10
1       7662    1    05  2019       05         50           13             15
2      13882    9    06  2019       09         25           04             25
3       6218   12    05  2019       18         05           23             30
4      13302   01    03  2019       16         50           21             35
…        …   …    …   …          …          …            …              …
10678   4107    9    04  2019       19         55           22             25
10679   4145   27    04  2019       20         45           23             20
10680   7229   27    04  2019       08         20           11             20
10681  12648   01    03  2019       11         30           14             10
10682  11753    9    05  2019       10         55           19             15

       Duration_Hours  Duration_Minutes
0                   2                50
1                   7                25
2                  19                 0
3                   5                25
4                   4                45
…                 …                …
10678               2                30
10679               2                35
10680               3                 0
10681               2                40
10682               8                20

[10682 rows x 16 columns]
```

[29]:
```
df["Total_Duration_In_Minutes"]=(60*df["Duration_Hours"])+df["Duration_Minutes"]
```

[30]:
```
df
```

[30]:
```
           Airline    Source Destination Duration Total_Stops Additional_Info  \
0           IndiGo  Banglore   New Delhi   2h 50m    non-stop         No info
1        Air India   Kolkata    Banglore   7h 25m     2 stops         No info
2      Jet Airways     Delhi      Cochin      19h     2 stops         No info
3           IndiGo   Kolkata    Banglore   5h 25m      1 stop         No info
4           IndiGo  Banglore   New Delhi   4h 45m      1 stop         No info
…             …       …          …      …          …              …
10678     Air Asia   Kolkata    Banglore   2h 30m    non-stop         No info
10679    Air India   Kolkata    Banglore   2h 35m    non-stop         No info
```

```
10680  Jet Airways  Banglore       Delhi       3h    non-stop         No info
10681      Vistara  Banglore   New Delhi   2h 40m    non-stop         No info
10682    Air India     Delhi      Cochin   8h 20m     2 stops         No info

       Price Date Month  Year Dep_Hour Dep_Minute Arrival_Hour Arrival_Minute  \
0       3897   24    03  2019       22         20           01             10
1       7662    1    05  2019       05         50           13             15
2      13882    9    06  2019       09         25           04             25
3       6218   12    05  2019       18         05           23             30
4      13302   01    03  2019       16         50           21             35
…        …  …    …   …       …         …           …              …
10678   4107    9    04  2019       19         55           22             25
10679   4145   27    04  2019       20         45           23             20
10680   7229   27    04  2019       08         20           11             20
10681  12648   01    03  2019       11         30           14             10
10682  11753    9    05  2019       10         55           19             15

       Duration_Hours  Duration_Minutes  Total_Duration_In_Minutes
0                   2                50                        170
1                   7                25                        445
2                  19                 0                       1140
3                   5                25                        325
4                   4                45                        285
…                 …                  …                         …
10678               2                30                        150
10679               2                35                        155
10680               3                 0                        180
10681               2                40                        160
10682               8                20                        500

[10682 rows x 17 columns]
```

Now we can drop Duration , Duration_hours and Duration_minutes columns

[31]:
```
df.drop(["Duration","Duration_Hours","Duration_Minutes"],axis=1,inplace=True)
```

[32]:
```
df
```

[32]:
```
            Airline    Source Destination Total_Stops Additional_Info  Price  \
0            IndiGo  Banglore   New Delhi    non-stop         No info   3897
1         Air India   Kolkata    Banglore     2 stops         No info   7662
2       Jet Airways     Delhi      Cochin     2 stops         No info  13882
3            IndiGo   Kolkata    Banglore      1 stop         No info   6218
4            IndiGo  Banglore   New Delhi      1 stop         No info  13302
…              …       …          …         …              …     …
10678     Air Asia   Kolkata    Banglore    non-stop         No info   4107
10679    Air India   Kolkata    Banglore    non-stop         No info   4145
```

```
10680  Jet Airways  Banglore       Delhi    non-stop         No info   7229
10681      Vistara  Banglore   New Delhi    non-stop         No info  12648
10682    Air India     Delhi      Cochin     2 stops         No info  11753

      Date Month  Year Dep_Hour Dep_Minute Arrival_Hour Arrival_Minute  \
0       24    03  2019       22         20           01             10
1        1    05  2019       05         50           13             15
2        9    06  2019       09         25           04             25
3       12    05  2019       18         05           23             30
4       01    03  2019       16         50           21             35
…     …   …    …       …         …           …              …
10678    9    04  2019       19         55           22             25
10679   27    04  2019       20         45           23             20
10680   27    04  2019       08         20           11             20
10681   01    03  2019       11         30           14             10
10682    9    05  2019       10         55           19             15

       Total_Duration_In_Minutes
0                            170
1                            445
2                           1140
3                            325
4                            285
…                            …
10678                        150
10679                        155
10680                        180
10681                        160
10682                        500

[10682 rows x 14 columns]
```

1.2.8 Now change Total_Stops to Numeric Values

[33]:
```
df["Total_Stops"].unique()
```

[33]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

[34]:
```
df["Total_Stops"].isnull().sum()
```

[34]:
```
1
```

[35]:
```
modee=df["Total_Stops"].mode()
modee
```

[35]:
```
0    1 stop
Name: Total_Stops, dtype: object
```

[36]:
```
df["Stops"]=df["Total_Stops"].map({'non-stop':0, '2 stops':2, '1 stop':1, '3
↪stops':3, np.nan:1, '4 stops':4})
```

[37]:
```
df
```

[37]:

	Airline	Source	Destination	Total_Stops	Additional_Info	Price	\
0	IndiGo	Banglore	New Delhi	non-stop	No info	3897	
1	Air India	Kolkata	Banglore	2 stops	No info	7662	
2	Jet Airways	Delhi	Cochin	2 stops	No info	13882	
3	IndiGo	Kolkata	Banglore	1 stop	No info	6218	
4	IndiGo	Banglore	New Delhi	1 stop	No info	13302	
…	…	…	…	…	…	…	
10678	Air Asia	Kolkata	Banglore	non-stop	No info	4107	
10679	Air India	Kolkata	Banglore	non-stop	No info	4145	
10680	Jet Airways	Banglore	Delhi	non-stop	No info	7229	
10681	Vistara	Banglore	New Delhi	non-stop	No info	12648	
10682	Air India	Delhi	Cochin	2 stops	No info	11753	

	Date	Month	Year	Dep_Hour	Dep_Minute	Arrival_Hour	Arrival_Minute	\
0	24	03	2019	22	20	01	10	
1	1	05	2019	05	50	13	15	
2	9	06	2019	09	25	04	25	
3	12	05	2019	18	05	23	30	
4	01	03	2019	16	50	21	35	
…	…	…	…	…	…	…	…	
10678	9	04	2019	19	55	22	25	
10679	27	04	2019	20	45	23	20	
10680	27	04	2019	08	20	11	20	
10681	01	03	2019	11	30	14	10	
10682	9	05	2019	10	55	19	15	

	Total_Duration_In_Minutes	Stops
0	170	0
1	445	2
2	1140	2
3	325	1
4	285	1
…	…	…
10678	150	0
10679	155	0
10680	180	0
10681	160	0
10682	500	2

[10682 rows x 15 columns]

1.2.9 Now Drop Total_Stops column

[38]:
```
df.drop("Total_Stops",axis=1,inplace=True)
```

[39]:
```
df
```

[39]:

	Airline	Source	Destination	Additional_Info	Price	Date	Month	\
0	IndiGo	Banglore	New Delhi	No info	3897	24	03	
1	Air India	Kolkata	Banglore	No info	7662	1	05	
2	Jet Airways	Delhi	Cochin	No info	13882	9	06	
3	IndiGo	Kolkata	Banglore	No info	6218	12	05	
4	IndiGo	Banglore	New Delhi	No info	13302	01	03	
…	…	…	…	…	…	…	…	
10678	Air Asia	Kolkata	Banglore	No info	4107	9	04	
10679	Air India	Kolkata	Banglore	No info	4145	27	04	
10680	Jet Airways	Banglore	Delhi	No info	7229	27	04	
10681	Vistara	Banglore	New Delhi	No info	12648	01	03	
10682	Air India	Delhi	Cochin	No info	11753	9	05	

	Year	Dep_Hour	Dep_Minute	Arrival_Hour	Arrival_Minute	\
0	2019	22	20	01	10	
1	2019	05	50	13	15	
2	2019	09	25	04	25	
3	2019	18	05	23	30	
4	2019	16	50	21	35	
…	…	…	…	…	…	
10678	2019	19	55	22	25	
10679	2019	20	45	23	20	
10680	2019	08	20	11	20	
10681	2019	11	30	14	10	
10682	2019	10	55	19	15	

	Total_Duration_In_Minutes	Stops
0	170	0
1	445	2
2	1140	2
3	325	1
4	285	1
…	…	…
10678	150	0
10679	155	0
10680	180	0
10681	160	0
10682	500	2

```
[10682 rows x 14 columns]
```

1.2.10 Now let's handle Additional_Info

[40]:
```
df["Additional_Info"].unique()
```

[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)
```

[41]:
```
df["Additional_Info"].value_counts()
```

[41]:
```
No info                         8344
In-flight meal not included     1982
No check-in baggage included     320
1 Long layover                    19
Change airports                    7
Business class                     4
No Info                            3
1 Short layover                    1
Red-eye flight                     1
2 Long layover                     1
Name: Additional_Info, dtype: int64
```

1.2.11 As most of the data is of NO INFO, let's delete this column

[42]:
```
df.drop("Additional_Info",axis=1,inplace=True)
```

[43]:
```
df
```

[43]:

	Airline	Source	Destination	Price	Date	Month	Year	Dep_Hour	\
0	IndiGo	Banglore	New Delhi	3897	24	03	2019	22	
1	Air India	Kolkata	Banglore	7662	1	05	2019	05	
2	Jet Airways	Delhi	Cochin	13882	9	06	2019	09	
3	IndiGo	Kolkata	Banglore	6218	12	05	2019	18	
4	IndiGo	Banglore	New Delhi	13302	01	03	2019	16	
…	…	…	…	…	…	…	…	…	
10678	Air Asia	Kolkata	Banglore	4107	9	04	2019	19	
10679	Air India	Kolkata	Banglore	4145	27	04	2019	20	
10680	Jet Airways	Banglore	Delhi	7229	27	04	2019	08	
10681	Vistara	Banglore	New Delhi	12648	01	03	2019	11	
10682	Air India	Delhi	Cochin	11753	9	05	2019	10	

	Dep_Minute	Arrival_Hour	Arrival_Minute	Total_Duration_In_Minutes	Stops
0	20	01	10	170	0

```
1              50          13              15                        445       2
2              25          04              25                       1140       2
3              05          23              30                        325       1
4              50          21              35                        285       1
…             …           …               …                    …     …
10678          55          22              25                        150       0
10679          45          23              20                        155       0
10680          20          11              20                        180       0
10681          30          14              10                        160       0
10682          55          19              15                        500       2

[10682 rows x 13 columns]
```

1.2.12 As price is Target Column, Lets create another column for price at last and drop the current one

[44]:
```
df["Amount"]=df["Price"]
df.drop("Price",axis=1,inplace=True)
```

[45]:
```
df
```

[45]:
```
            Airline    Source Destination Date Month  Year Dep_Hour Dep_Minute  \
0            IndiGo  Banglore   New Delhi   24    03  2019       22         20
1        Air India   Kolkata    Banglore    1    05  2019       05         50
2      Jet Airways     Delhi      Cochin    9    06  2019       09         25
3            IndiGo   Kolkata    Banglore   12    05  2019       18         05
4            IndiGo  Banglore   New Delhi   01    03  2019       16         50
…              …       …          …  …   …   …      …        …
10678     Air Asia   Kolkata    Banglore    9    04  2019       19         55
10679    Air India   Kolkata    Banglore   27    04  2019       20         45
10680  Jet Airways  Banglore       Delhi   27    04  2019       08         20
10681      Vistara  Banglore   New Delhi   01    03  2019       11         30
10682    Air India     Delhi      Cochin    9    05  2019       10         55

      Arrival_Hour Arrival_Minute  Total_Duration_In_Minutes  Stops  Amount
0               01             10                        170      0    3897
1               13             15                        445      2    7662
2               04             25                       1140      2   13882
3               23             30                        325      1    6218
4               21             35                        285      1   13302
…              …              …                    …    …     …
10678           22             25                        150      0    4107
10679           23             20                        155      0    4145
10680           11             20                        180      0    7229
10681           14             10                        160      0   12648
10682           19             15                        500      2   11753
```

```
[10682 rows x 13 columns]
```

1.3 Now convert all the numeric values as INT

[46]:
```
lst=['Date','Month','Year','Dep_Hour','Dep_Minute','Arrival_Hour','Arrival_Minute','Total_Dura
```

[47]:
```
for i in lst:
    df[i]=df[i].astype(int)
```

[48]:
```
df
```

[48]:

	Airline	Source	Destination	Date	Month	Year	Dep_Hour	\
0	IndiGo	Banglore	New Delhi	24	3	2019	22	
1	Air India	Kolkata	Banglore	1	5	2019	5	
2	Jet Airways	Delhi	Cochin	9	6	2019	9	
3	IndiGo	Kolkata	Banglore	12	5	2019	18	
4	IndiGo	Banglore	New Delhi	1	3	2019	16	
…	…	…	…	…	…	…	…	
10678	Air Asia	Kolkata	Banglore	9	4	2019	19	
10679	Air India	Kolkata	Banglore	27	4	2019	20	
10680	Jet Airways	Banglore	Delhi	27	4	2019	8	
10681	Vistara	Banglore	New Delhi	1	3	2019	11	
10682	Air India	Delhi	Cochin	9	5	2019	10	

	Dep_Minute	Arrival_Hour	Arrival_Minute	Total_Duration_In_Minutes	\
0	20	1	10	170	
1	50	13	15	445	
2	25	4	25	1140	
3	5	23	30	325	
4	50	21	35	285	
…	…	…	…	…	
10678	55	22	25	150	
10679	45	23	20	155	
10680	20	11	20	180	
10681	30	14	10	160	
10682	55	19	15	500	

	Stops	Amount
0	0	3897
1	2	7662
2	2	13882
3	1	6218
4	1	13302
…	…	…
10678	0	4107
10679	0	4145
10680	0	7229

```
10681      0   12648
10682      2   11753

[10682 rows x 13 columns]
```

1.4 Let's see whether they were converted to Int or not

[49]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10682 entries, 0 to 10682
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Airline                    10682 non-null  object
 1   Source                     10682 non-null  object
 2   Destination                10682 non-null  object
 3   Date                       10682 non-null  int64
 4   Month                      10682 non-null  int64
 5   Year                       10682 non-null  int64
 6   Dep_Hour                   10682 non-null  int64
 7   Dep_Minute                 10682 non-null  int64
 8   Arrival_Hour               10682 non-null  int64
 9   Arrival_Minute             10682 non-null  int64
 10  Total_Duration_In_Minutes  10682 non-null  int64
 11  Stops                      10682 non-null  int64
 12  Amount                     10682 non-null  int64
dtypes: int64(10), object(3)
memory usage: 1.1+ MB
```

1.5 Check for missing Data

[50]:
```
df.isnull().sum()
```

[50]:
```
Airline                    0
Source                     0
Destination                0
Date                       0
Month                      0
Year                       0
Dep_Hour                   0
Dep_Minute                 0
Arrival_Hour               0
Arrival_Minute             0
Total_Duration_In_Minutes  0
Stops                      0
Amount                     0
```

```
dtype: int64
```

1.6 Check for Duplicate Data

[51]:
```
df.duplicated().sum()
```

[51]:
```
222
```

1.7 Drop the Duplicate Values

[52]:
```
df.drop_duplicates(inplace=True)
```

[53]:
```
df.duplicated().sum()
```

[53]:
```
0
```

[54]:
```
df
```

[54]:

	Airline	Source	Destination	Date	Month	Year	Dep_Hour	\
0	IndiGo	Banglore	New Delhi	24	3	2019	22	
1	Air India	Kolkata	Banglore	1	5	2019	5	
2	Jet Airways	Delhi	Cochin	9	6	2019	9	
3	IndiGo	Kolkata	Banglore	12	5	2019	18	
4	IndiGo	Banglore	New Delhi	1	3	2019	16	
…	…	…	…	…	…	…	…	
10678	Air Asia	Kolkata	Banglore	9	4	2019	19	
10679	Air India	Kolkata	Banglore	27	4	2019	20	
10680	Jet Airways	Banglore	Delhi	27	4	2019	8	
10681	Vistara	Banglore	New Delhi	1	3	2019	11	
10682	Air India	Delhi	Cochin	9	5	2019	10	

	Dep_Minute	Arrival_Hour	Arrival_Minute	Total_Duration_In_Minutes	\
0	20	1	10	170	
1	50	13	15	445	
2	25	4	25	1140	
3	5	23	30	325	
4	50	21	35	285	
…	…	…	…	…	
10678	55	22	25	150	
10679	45	23	20	155	
10680	20	11	20	180	
10681	30	14	10	160	
10682	55	19	15	500	

	Stops	Amount
0	0	3897

```
1           2    7662
2           2   13882
3           1    6218
4           1   13302
…       …      …
10678       0    4107
10679       0    4145
10680       0    7229
10681       0   12648
10682       2   11753

[10460 rows x 13 columns]
```

1.8 Now, let's convert categorical values to Numerical

[55]:
```
df["Airline"].nunique(),df["Source"].nunique(),df["Destination"].nunique()
```

[55]:
```
(12, 5, 6)
```

[56]:
```
from sklearn.preprocessing import LabelEncoder
```

[57]:
```
label=LabelEncoder()
```

[58]:
```
airlines=label.fit_transform(df[["Airline"]])
sources=label.fit_transform(df[["Source"]])
destinations=label.fit_transform(df[["Destination"]])
```

[59]:
```
df["Airline_converted"]=airlines
df["Source_converted"]=sources
df["Destination_converted"]=destinations
```

[60]:
```
df
```

[60]:
```
            Airline    Source Destination  Date  Month  Year  Dep_Hour  \
0            IndiGo  Banglore   New Delhi    24      3  2019        22
1        Air India   Kolkata    Banglore     1      5  2019         5
2      Jet Airways     Delhi      Cochin     9      6  2019         9
3            IndiGo   Kolkata    Banglore    12      5  2019        18
4            IndiGo  Banglore   New Delhi     1      3  2019        16
…              …       …          …    …    …     …   …
10678     Air Asia   Kolkata    Banglore     9      4  2019        19
10679    Air India   Kolkata    Banglore    27      4  2019        20
10680  Jet Airways  Banglore       Delhi    27      4  2019         8
10681      Vistara  Banglore   New Delhi     1      3  2019        11
10682    Air India     Delhi      Cochin     9      5  2019        10

       Dep_Minute  Arrival_Hour  Arrival_Minute  Total_Duration_In_Minutes  \
```

```
0              20              1              10                      170
1              50             13              15                      445
2              25              4              25                     1140
3               5             23              30                      325
4              50             21              35                      285
…             …               …              …                       …
10678          55             22              25                      150
10679          45             23              20                      155
10680          20             11              20                      180
10681          30             14              10                      160
10682          55             19              15                      500

       Stops  Amount  Airline_converted  Source_converted  \
0          0    3897                  3                 0
1          2    7662                  1                 3
2          2   13882                  4                 2
3          1    6218                  3                 3
4          1   13302                  3                 0
…      …      …                      …                 …
10678      0    4107                  0                 3
10679      0    4145                  1                 3
10680      0    7229                  4                 0
10681      0   12648                 10                 0
10682      2   11753                  1                 2

       Destination_converted
0                          5
1                          0
2                          1
3                          0
4                          5
…                        …
10678                      0
10679                      0
10680                      2
10681                      5
10682                      1

[10460 rows x 16 columns]
```

[61]:
```
df["Airline"].nunique(),df["Source"].nunique(),df["Destination"].nunique()
```

[61]:
```
(12, 5, 6)
```

1.9 Let's make 2 columns for Duration Minutes and Amount to use it as scaled Data

[62]:
```
from sklearn.preprocessing import MinMaxScaler
```

[63]:
```
scaler=MinMaxScaler()
```

[64]:
```
df["Duration_Minutes_Scaled"]=scaler.
 ↪fit_transform(df[["Total_Duration_In_Minutes"]])
df["Amount_Scaled"]=scaler.fit_transform(df[["Amount"]])
```

[65]:
```
df
```

[65]:

	Airline	Source	Destination	Date	Month	Year	Dep_Hour	Dep_Minute	Arrival_Hour	Arrival_Minute	Total_Duration_In_Minutes	Stops	Amount	Airline_converted	Source_converted
0	IndiGo	Banglore	New Delhi	24	3	2019	22	20	1	10	170	0	3897	3	0
1	Air India	Kolkata	Banglore	1	5	2019	5	50	13	15	445	2	7662	1	3
2	Jet Airways	Delhi	Cochin	9	6	2019	9	25	4	25	1140	2	13882	4	2
3	IndiGo	Kolkata	Banglore	12	5	2019	18	5	23	30	325	1	6218	3	3
4	IndiGo	Banglore	New Delhi	1	3	2019	16	50	21	35	285	1	13302	3	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
10678	Air Asia	Kolkata	Banglore	9	4	2019	19	55	22	25	150	0	4107	0	3
10679	Air India	Kolkata	Banglore	27	4	2019	20	45	23	20	155	0			
10680	Jet Airways	Banglore	Delhi	27	4	2019	8	20	11	20	180	0			
10681	Vistara	Banglore	New Delhi	1	3	2019	11	30	14	10	160	0			
10682	Air India	Delhi	Cochin	9	5	2019	10	55	19	15	500	0			

```
10679      0    4145                  1                  3
10680      0    7229                  4                  0
10681      0   12648                 10                  0
10682      2   11753                  1                  2

       Destination_converted  Duration_Minutes_Scaled  Amount_Scaled
0                          5                 0.034111       0.027497
1                          0                 0.132855       0.075920
2                          1                 0.382406       0.155917
3                          0                 0.089767       0.057348
4                          5                 0.075404       0.148457
…                        …                    …              …
10678                      0                 0.026930       0.030198
10679                      0                 0.028725       0.030687
10680                      2                 0.037702       0.070351
10681                      5                 0.030521       0.140046
10682                      1                 0.152603       0.128535

[10460 rows x 18 columns]
```

2 Visualization

[66]:
```
df["Year"].value_counts()
```

[66]:
```
2019    10460
Name: Year, dtype: int64
```

2.1 Only one year's Data is Given. i.e, For this data Year may not have much significance

[67]:
```
df["Month"].value_counts()
```

[67]:
```
5    3396
6    3311
3    2675
4    1078
Name: Month, dtype: int64
```

2.2 Data is present for months March, April, May and June of 2019

[68]:
```
df.groupby("Month")["Amount"].mean()
```

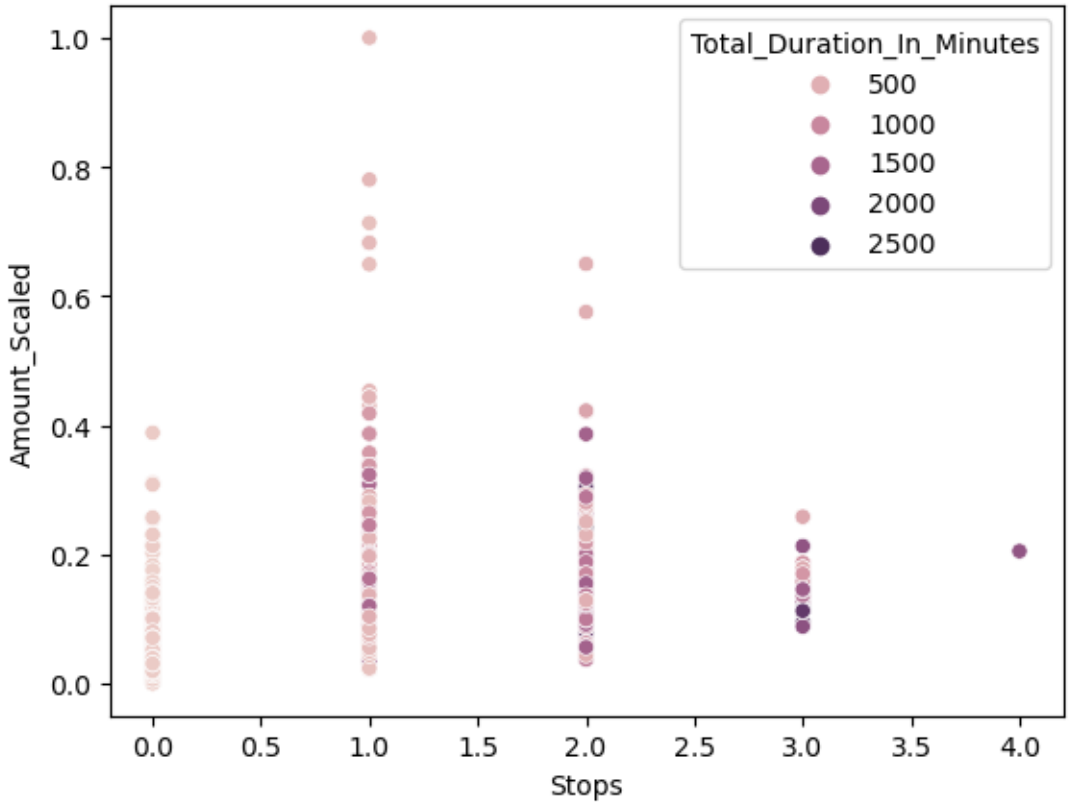
[68]:
```
Month
3    10696.395140
4     5766.545455
```

```
5      9028.783569
6      8736.152522
Name: Amount, dtype: float64
```

2.3 March has much flight ticket rates followed by May , June and April has the least

[69]:
```
sns.scatterplot(y=df.Amount_Scaled,x=df.
↪Stops,hue=df["Total_Duration_In_Minutes"])
```

[69]:
```
<AxesSubplot: xlabel='Stops', ylabel='Amount_Scaled'>
```

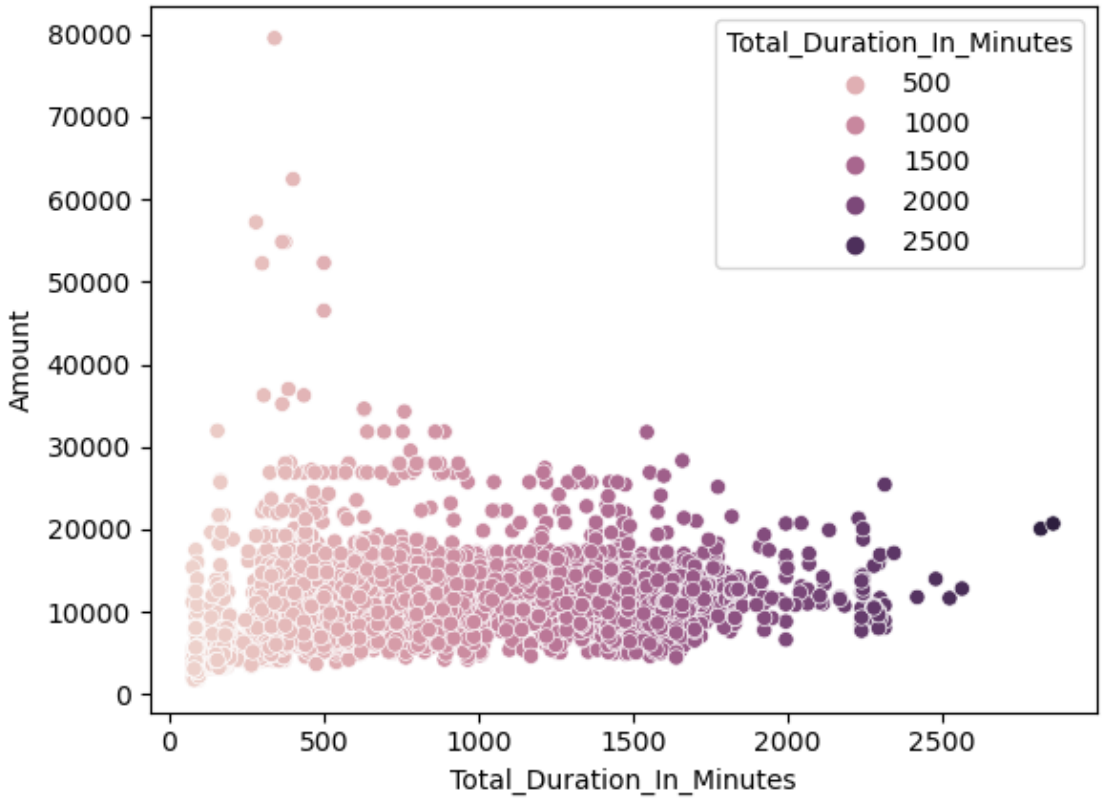


2.4

- Amount is less when there are “Zero” stops, and all the duration in minutes ranges beyond 500 only.
- Amount is slightly high if there is 1 stop compared to others, and all duration minutes are less than 1500.
- When there are 3 stops, amount is less but , Total Distance is more.

[70]:
```
sns.scatterplot(y=df.Amount,x=df.
↪Total_Duration_In_Minutes,hue=df["Total_Duration_In_Minutes"])
```

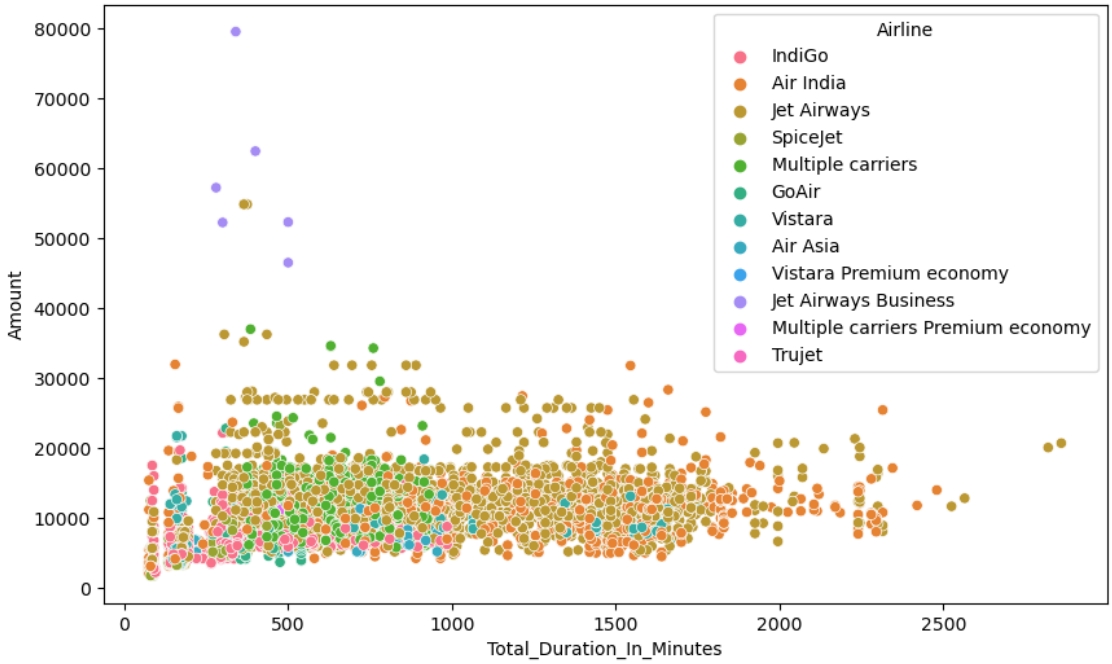
[70]: <AxesSubplot: xlabel='Total_Duration_In_Minutes', ylabel='Amount'>



2.5 Amount is high when Duration time is Less

[71]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(y=df.Amount,x=df.Total_Duration_In_Minutes,hue=df["Airline"])
```

[71]: <AxesSubplot: xlabel='Total_Duration_In_Minutes', ylabel='Amount'>

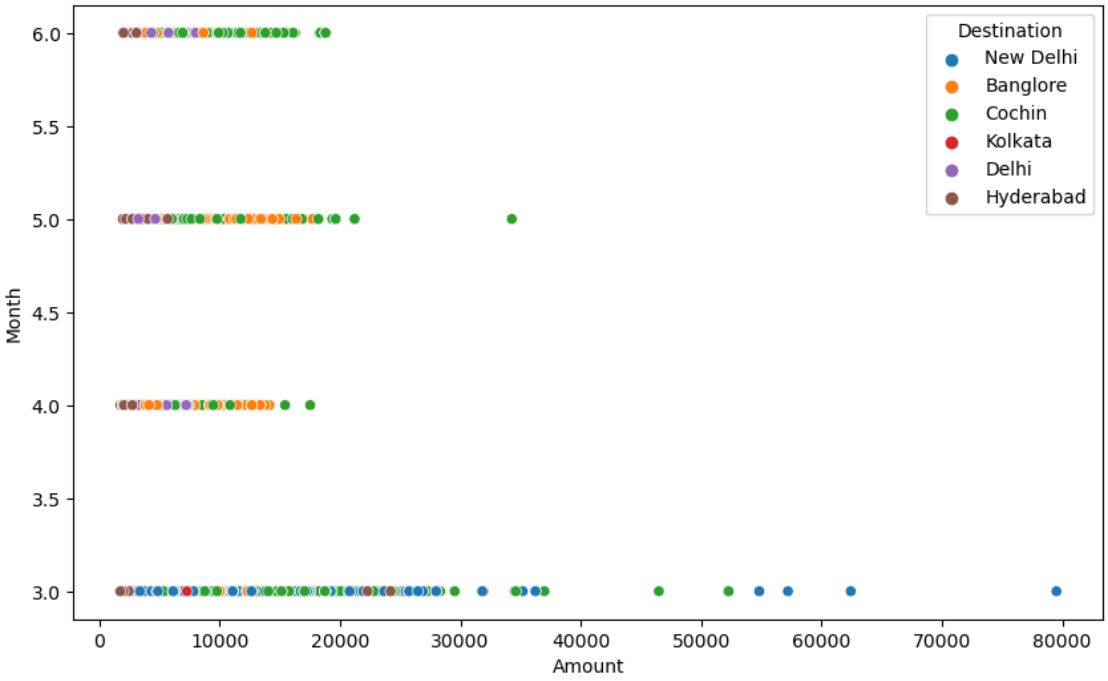


2.6

- Amount of Jet Airway Business is High compared to others but the travel duration is very less
- Trujet covers only below 1000 minutes of duration
- Jet airways Dominates Airline Industry with more customers, followed by Air India and Trujet

[72]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(y=df.Month,x=df.Amount,hue=df["Destination"])
```

[72]:
```
<AxesSubplot: xlabel='Amount', ylabel='Month'>
```

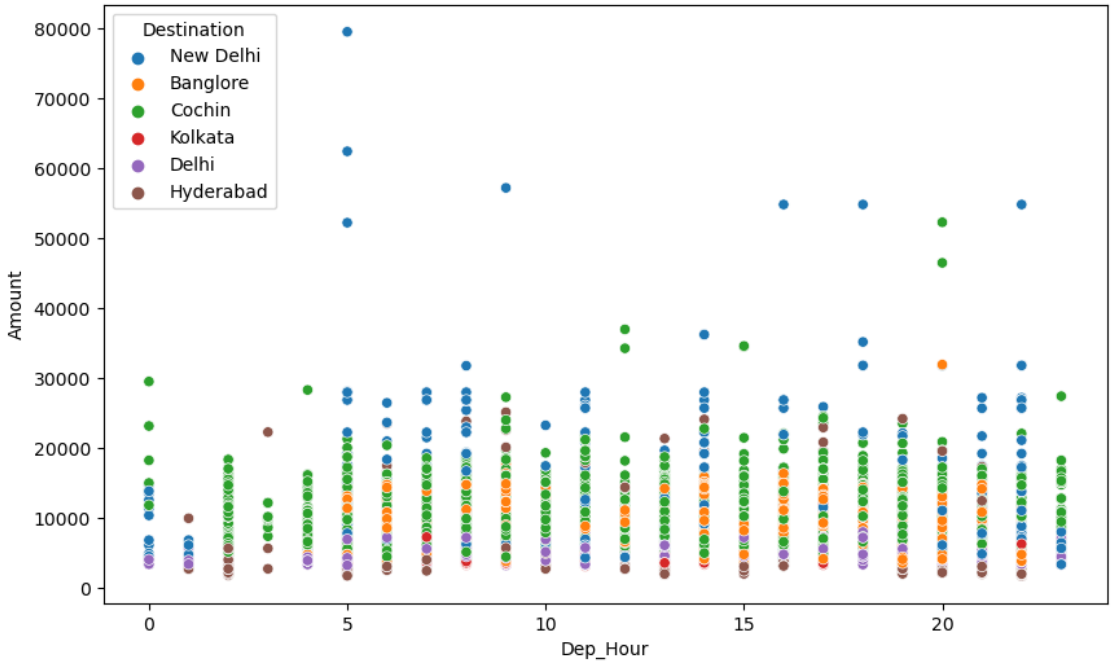


2.7

- on April most people were travelling to Delhi, and prices were also somewhat high
- Cochin is the most travelled place among all the Destinations.
- For Hydrebad most of the times Tickets were cheaper.

[73]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(x=df.Dep_Hour,y=df.Amount,hue=df["Destination"])
```

[73]:
```
<AxesSubplot: xlabel='Dep_Hour', ylabel='Amount'>
```

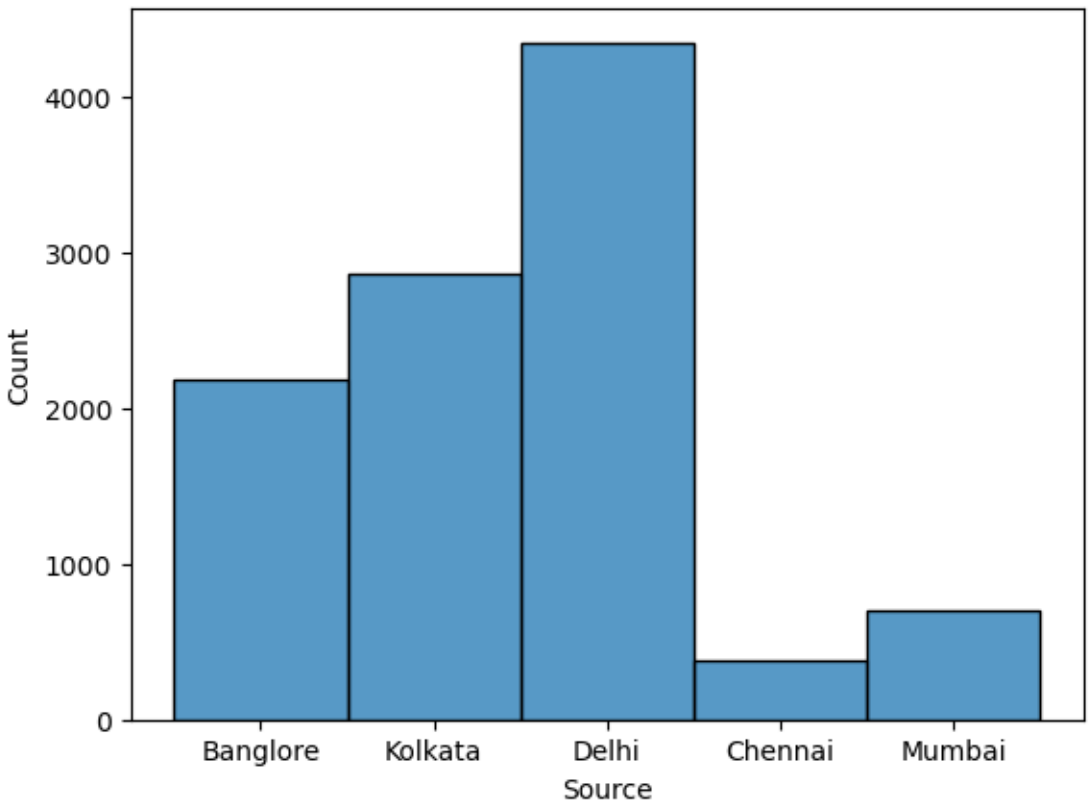


2.8

- Cochin is the most travelled place at most times
- Delhi has high prices in most of the times

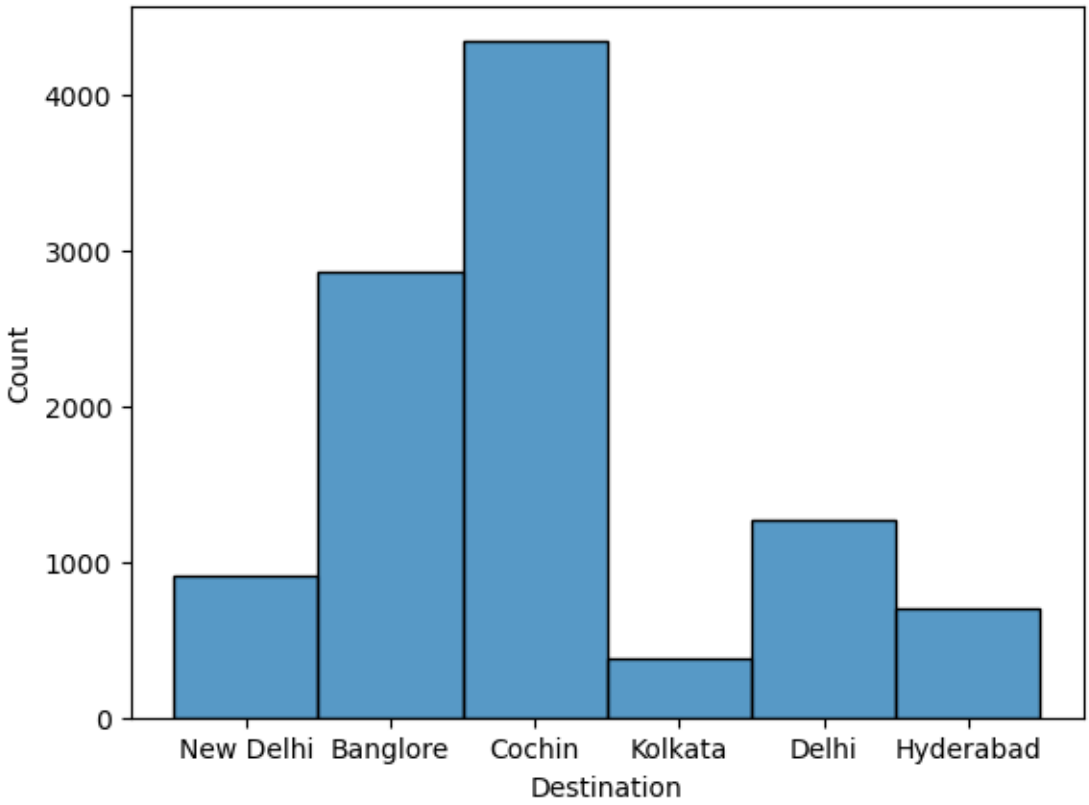
[74]:
```
sns.histplot(df.Source)
```

[74]:
```
<AxesSubplot: xlabel='Source', ylabel='Count'>
```



[75]:
```
sns.histplot(df.Destination)
```

[75]:
```
<AxesSubplot: xlabel='Destination', ylabel='Count'>
```

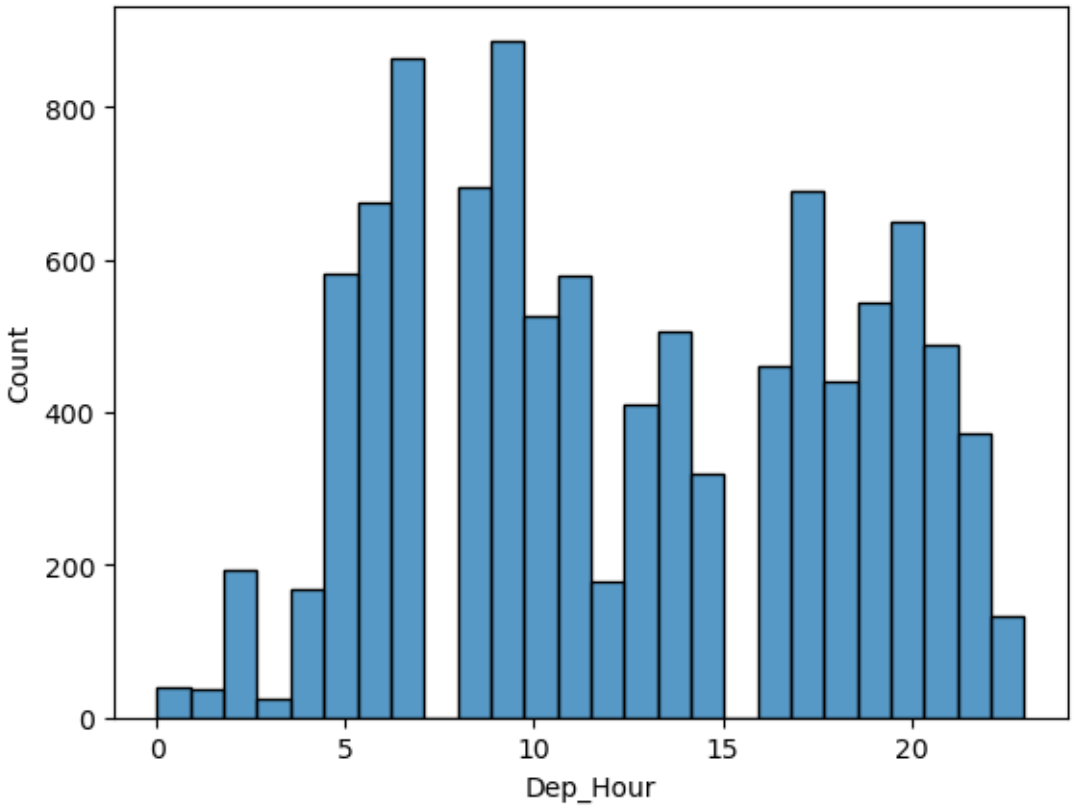


2.9

- No one has started their flight from Hydrebad and Cochin
- No one travelled to Mumbai and Chennai

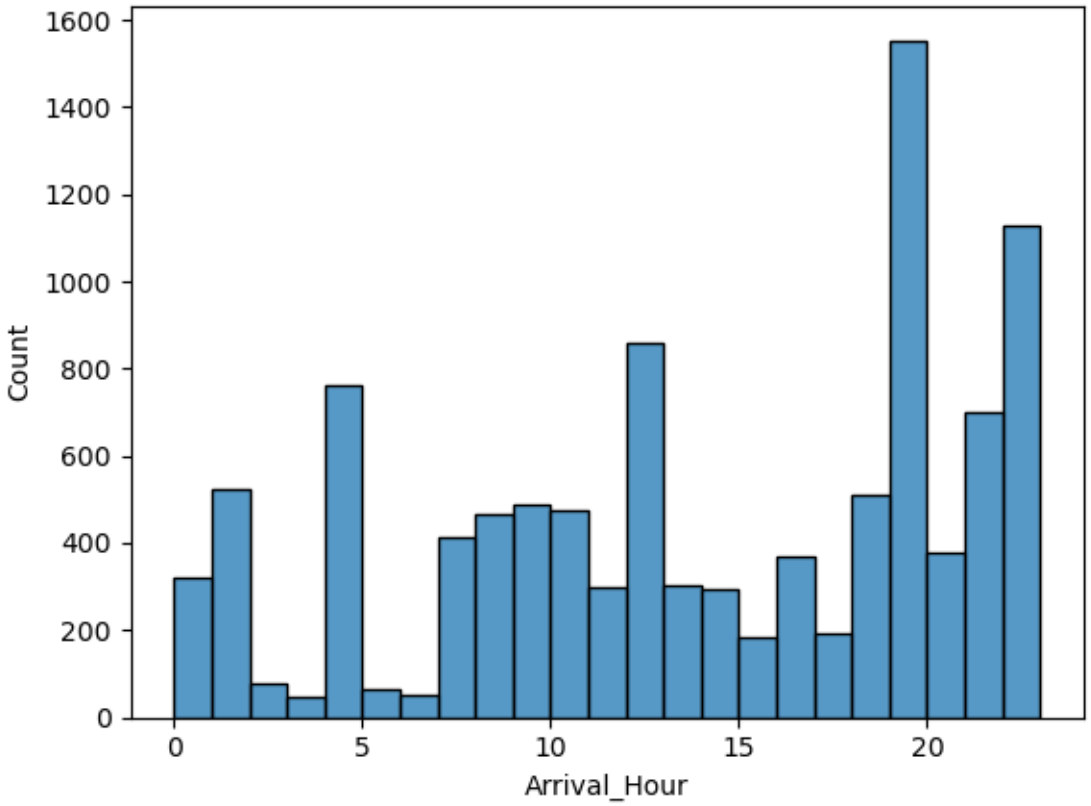
[76]:
```
sns.histplot(df.Dep_Hour)
```

[76]:
```
<AxesSubplot: xlabel='Dep_Hour', ylabel='Count'>
```



[77]:
```
sns.histplot(df.Arrival_Hour)
```

[77]:
```
<AxesSubplot: xlabel='Arrival_Hour', ylabel='Count'>
```



2.10 From mid night 12 to morning 6 people were less likely to arrive or departure