

# EDA\_flight\_price\_practice\_2

February 24, 2024

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

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

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

```
[3]:
```

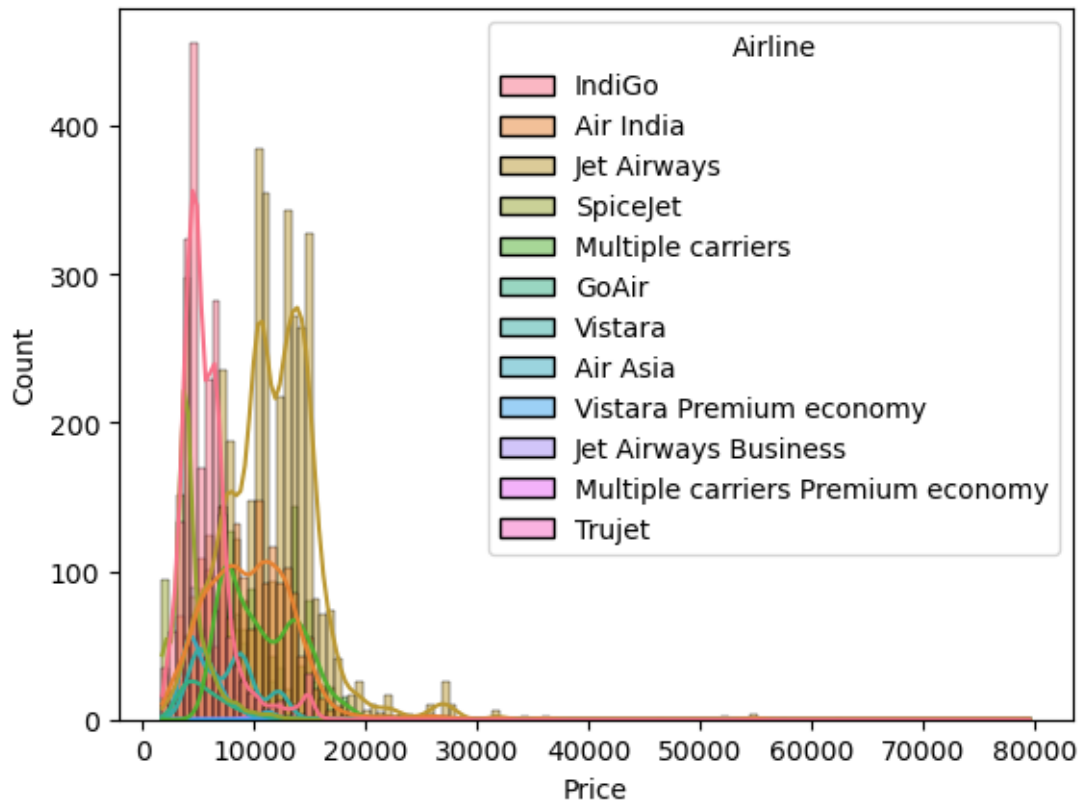
	Airline	Date_of_Journey	Source	Destination	Route	\
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	

	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	05:50	13:15	7h 25m	2 stops	No info	7662

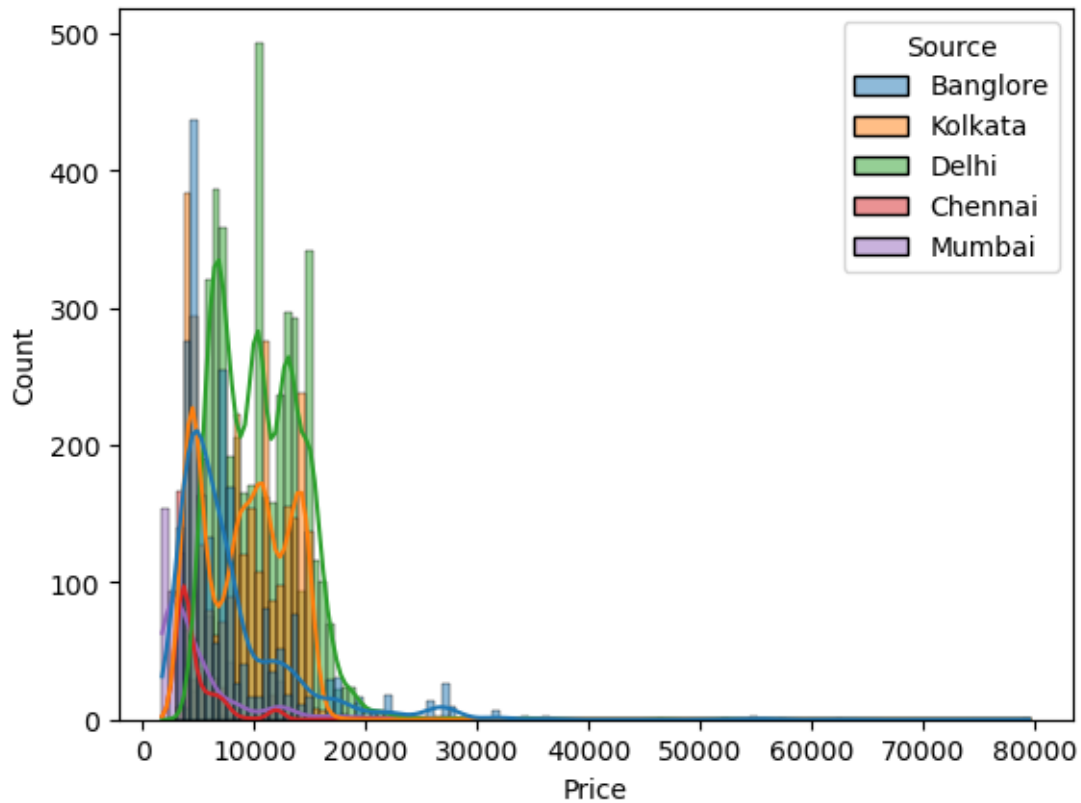
```
[4]: sns.histplot(data = df, x = df['Price'], hue = df['Airline'], kde = True)
```

```
[4]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



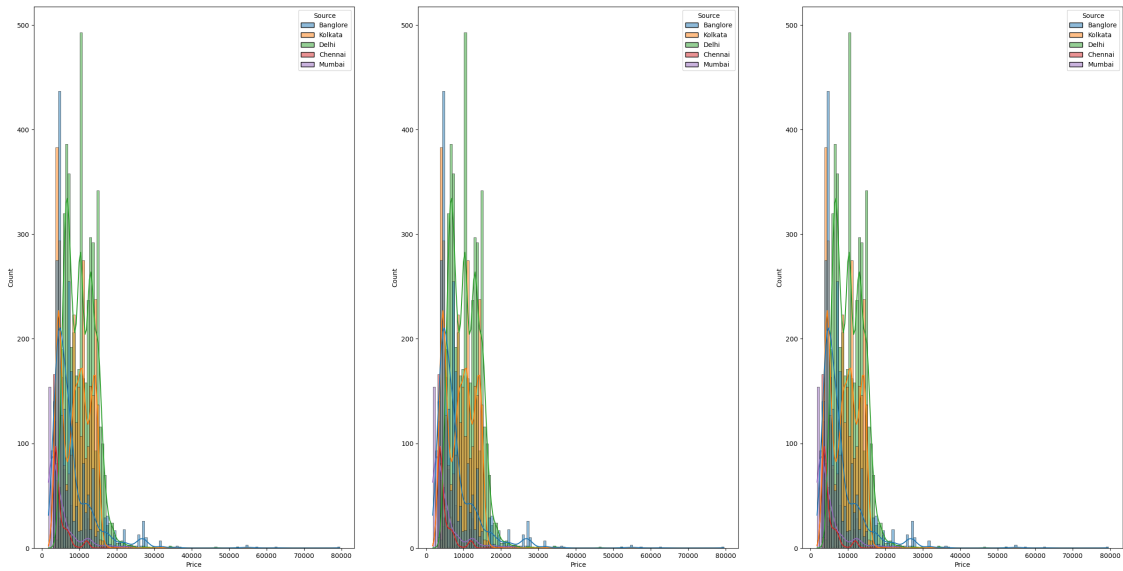
```
[5]: sns.histplot(data = df, x = df['Price'], hue = df['Source'], kde = True)
```

```
[5]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



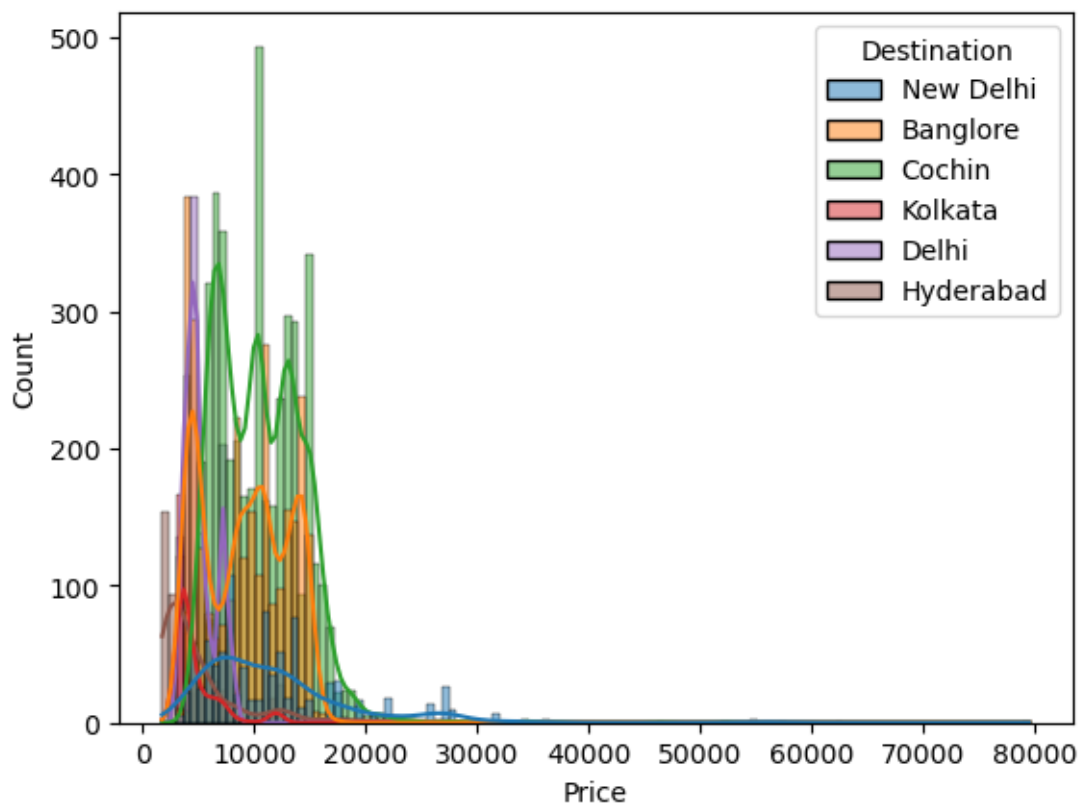
```
[6]: plt.subplots(1,3,figsize=(30,15))
plt.subplot(131)
sns.histplot(data = df[df.Airline == 'Indigo'], x = df['Price'], hue = df['Source'], kde = True)
plt.subplot(132)
sns.histplot(data = df[df.Airline == 'Jet Airways'], x = df['Price'], hue = df['Source'], kde = True)
plt.subplot(133)
sns.histplot(data = df[df.Airline == 'Trujet'], x = df['Price'], hue = df['Source'], kde = True)
```

```
[6]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



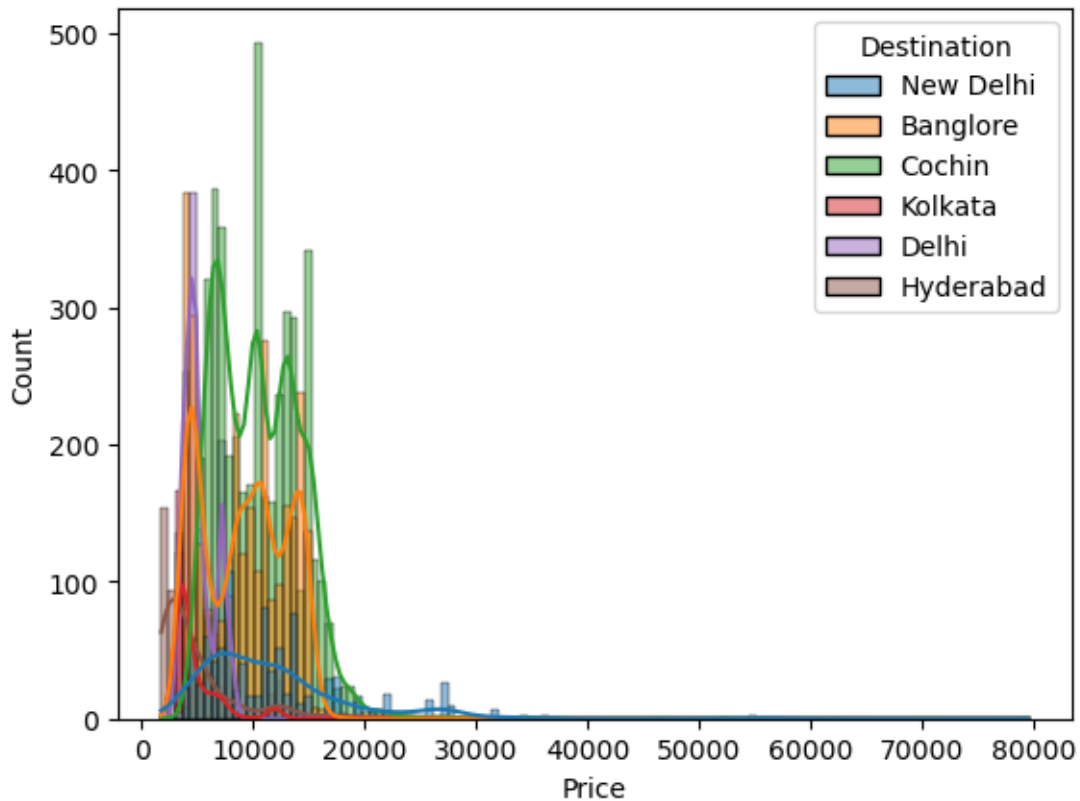
```
[7]: sns.histplot(data = df, x = df['Price'], hue = df['Destination'], kde = True)
```

```
[7]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



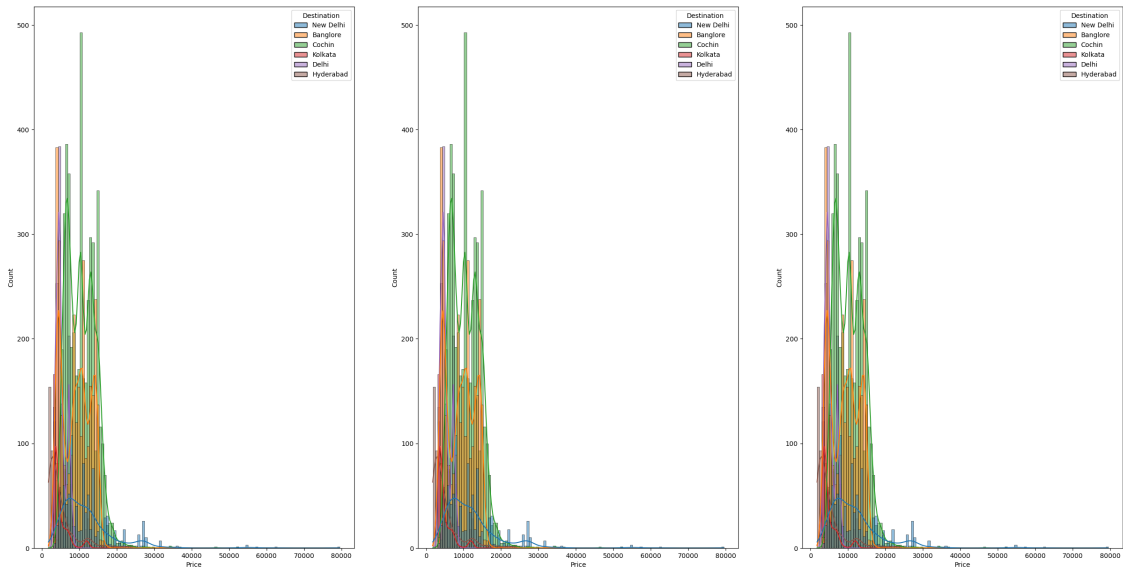
```
[8]: sns.histplot(data = df[df.Airline == 'Indigo'], x = df['Price'], hue = df['Destination'], kde = True)
```

```
[8]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



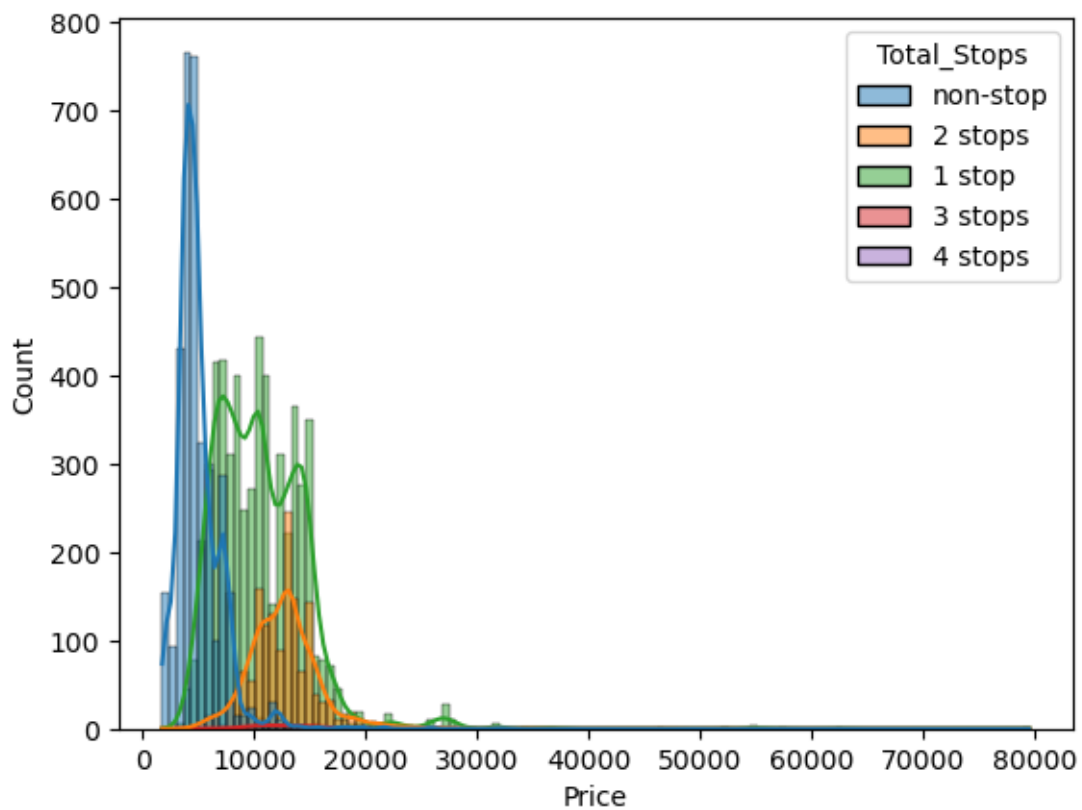
```
[9]: plt.subplots(1,3,figsize=(30,15))
plt.subplot(131)
sns.histplot(data = df[df.Airline == 'Indigo'], x = df['Price'], hue = df['Destination'], kde = True)
plt.subplot(132)
sns.histplot(data = df[df.Airline == 'Jet Airways'], x = df['Price'], hue = df['Destination'], kde = True)
plt.subplot(133)
sns.histplot(data = df[df.Airline == 'Trujet'], x = df['Price'], hue = df['Destination'], kde = True)
```

```
[9]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



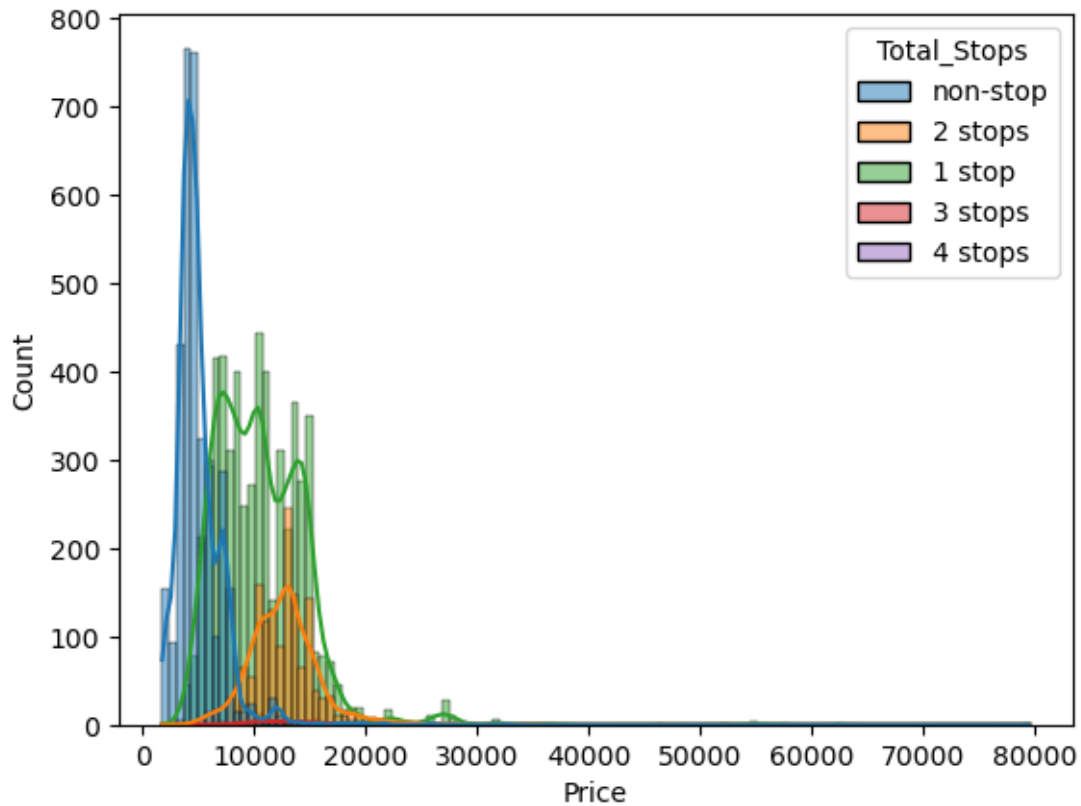
```
[10]: sns.histplot(data = df, x = df['Price'], hue = df['Total_Stops'], kde = True)
```

```
[10]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



```
[11]: sns.histplot(data = df[df.Airline == 'Indigo'], x = df['Price'], hue = df['Total_Stops'], kde = True)
```

```
[11]: <AxesSubplot: xlabel='Price', ylabel='Count'>
```



```
[ ]: plt.subplots(1,3,figsize=(30,15))
plt.subplot(131)
sns.histplot(data = df[df.Airline == 'Indigo'], x = df['Price'], hue = df['Total_Stops'], kde = True)
plt.subplot(132)
sns.histplot(data = df[df.Airline == 'Jet Airways'], x = df['Price'], hue = df['Total_Stops'], kde = True)
plt.subplot(133)
sns.histplot(data = df[df.Airline == 'Trujet'], x = df['Price'], hue = df['Total_Stops'], kde = True)
```

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

## 0.1 Insights

1. Indigo has cheapest flight
2. Most cheapest flights have departed from Chennai than else where irrespective of flight
3. Most flights departed from delhi are costliest.
4. Most cheapest flights have arrived at Kolkata irrespective of airlines.
5. Costliest flights have arrived at Cochin, irrespective of airlines.
6. Most costliest flights have 2 stop, irrespective of airline.
7. Most cheapest flights have no stop, irrespective of airline.
8. Jet Airways is costliest flight among all

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

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

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

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

```
[ ]: encoded_df = pd.DataFrame(encoded, columns = encoder.get_feature_names_out())
```

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

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

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

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

```
[ ]: df.drop(['Total_Stops'],axis=1)
```

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

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

```
[ ]: df['Additional_Info'] = df['Additional_Info'].map({'No info':0, 'In-flight meal_  
    ↪not included':1,  
    'No check-in baggage included':2, '1 Short layover':3, 'No Info':0,  
    '1 Long layover':4, 'Change airports':5, 'Business class':6,  
    'Red-eye flight':7, '2 Long layover':8})
```

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

```
[ ]: df['Day'] = 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]
```



```

[ ]: df['Departure_Hour'] = df['Dep_Time'].str.split(':').str[0]
    df['Departure_Minute'] = df['Dep_Time'].str.split(':').str[1]

[ ]: df = df.drop(['Dep_Time'],axis=1)

[ ]: df = df.drop(['Date_of_Journey'],axis=1)

[ ]: df.head(2)

[ ]: df['Arrival_hour'] = df['Arrival_Time'].str.split(' ').str[0].str.split(':').
    ↪str[0]

[ ]: df['Arrival_minute'] = df['Arrival_Time'].str.split(' ').str[0].str.split(':').
    ↪str[0]

[ ]: df = df.drop(['Arrival_Time'],axis=1)

[ ]: df.head()

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

[ ]: df['Duration_hour'].unique()

[ ]: df['Duration_hour'] = df['Duration_hour'].replace('5m','0.083')

[ ]: df['Duration_hour'].unique()

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

[ ]: df['Duration_min'] = df['Duration_min'].str.split('m').str[0]

[ ]: df['Duration_min']

[ ]: df = df.drop(['Duration'], axis = 1)

[ ]: df.head(2)

[ ]: df = df.sort_values(by = 'Duration_hour')

[ ]: df.head()

[ ]: sns.histplot(data = df, x = df['Price'], hue = df['Duration_hour'], bins = 40,
    ↪kde = True)

[ ]: df = df.sort_values(by = 'Departure_Hour')
    sns.histplot(data = df, x = df['Price'], hue = df['Departure_Hour'], kde = True)

```

## 0.2 Insighst

1. Cheapest flight have very less duration hour
2. Costliest and cheapest flights have departed early hour
3. Peak Arrival seasn is around may and june.
4. Bangalore and Cochin are the places where most people have arrived.

[ ]:

[ ]:

[ ]:

[ ]: