Importing Libraries

In [3]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib
import seaborn as sns
import warnings
%matplotlib inline
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

Data Preparation

In [6]:
```
df = pd.read_excel("C:/Users/Titan Rafi/Dropbox/PC/Desktop/ML Projects - Self/Data_Train.xlsx")
df.head()
```

Out[6]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

In [7]:
```
df.columns = df.columns.str.lower()
```

In [10]:
```
df.describe().T
```

Out[10]:

	count	mean	std	min	25%	50%	75%	max
price	10683.0	9087.064121	4611.359167	1759.0	5277.0	8372.0	12373.0	79512.0

In [11]:
```
df.isna().sum()
```

Out[11]:
```
airline            0
date_of_journey    0
source             0
destination        0
route              1
dep_time           0
arrival_time       0
duration           0
total_stops        1
additional_info    0
price              0
dtype: int64
```

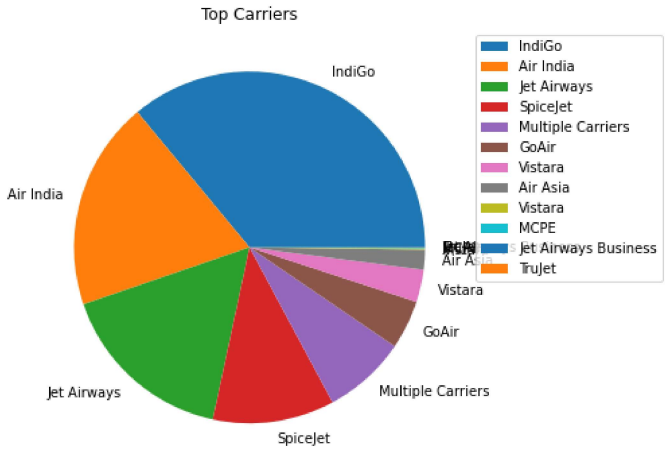
In [12]:
```
df.dropna(inplace=True)
```

In [13]:
```
df.dtypes
```

Out[13]:
```
airline            object
date_of_journey    object
source             object
destination        object
route              object
dep_time           object
arrival_time       object
duration           object
total_stops        object
additional_info    object
price               int64
dtype: object
```

1.Top Carriers by No of Flight :

In [15]:
```
top_carriers = np.array(df['airline'].value_counts(sort=True))
labels = ['IndiGo','Air India','Jet Airways','SpiceJet','Multiple Carriers','GoAir','Vistara','Air Asia','Vistara','MCPE',
          'Jet Airways Business','TruJet']
plt.figure(figsize=(8,6))
plt.pie(top_carriers, labels=labels,shadow=False)
plt.legend(labels=labels,bbox_to_anchor=(1,1))
plt.title('Top Carriers')
plt.show()
```



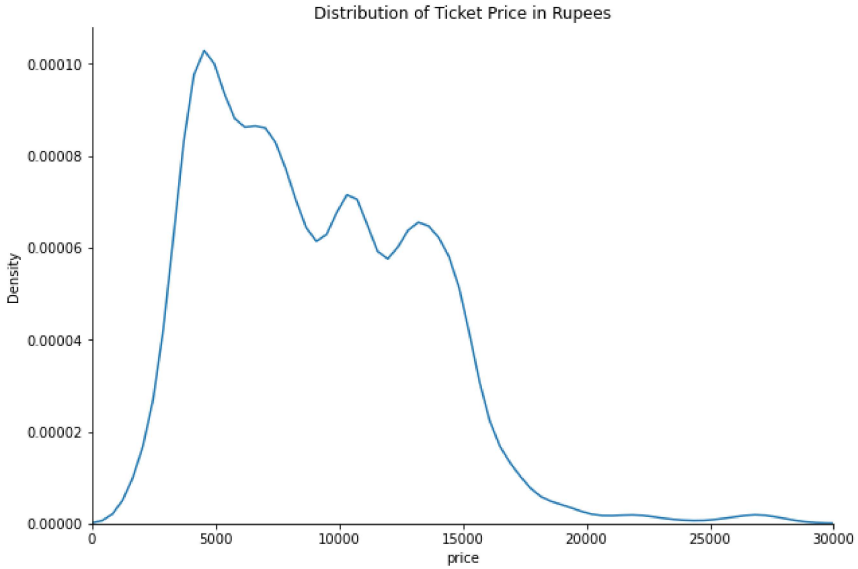
Conclusion:

* IndiGo Operates the Most No.of Flights.

2.Distribution of Ticket Prices :

In [18]:
```
ticket_price = sns.displot(data=df, x='price', kind='kde', legend='True')
plt.title('Distribution of Ticket Price in Rupees')
ticket_price.fig.set_figwidth(10)
ticket_price.fig.set_figheight(6)
ticket_price.set(xlim=(0, 30000))
```

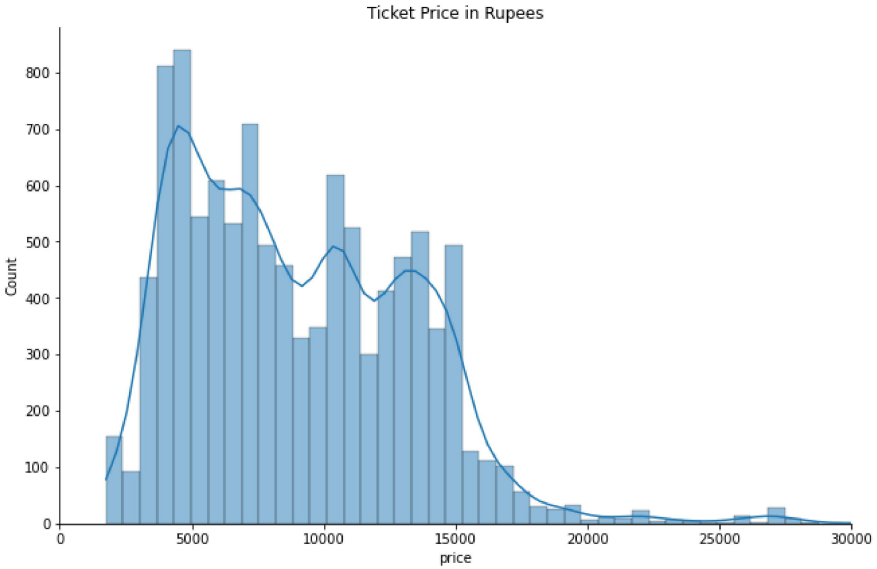
Out[18]: <seaborn.axisgrid.FacetGrid at 0x1cba81ac340>



3.Histogram of Ticket Prices :

In [19]:
```
ticket_price = sns.displot(x=df['price'], data=df, kde=True)
plt.title('Ticket Price in Rupees')
ticket_price.fig.set_figwidth(10)
ticket_price.fig.set_figheight(6)
ticket_price.set(xlim=(0, 30000))
```

Out[19]: <seaborn.axisgrid.FacetGrid at 0x1cba6426e20>

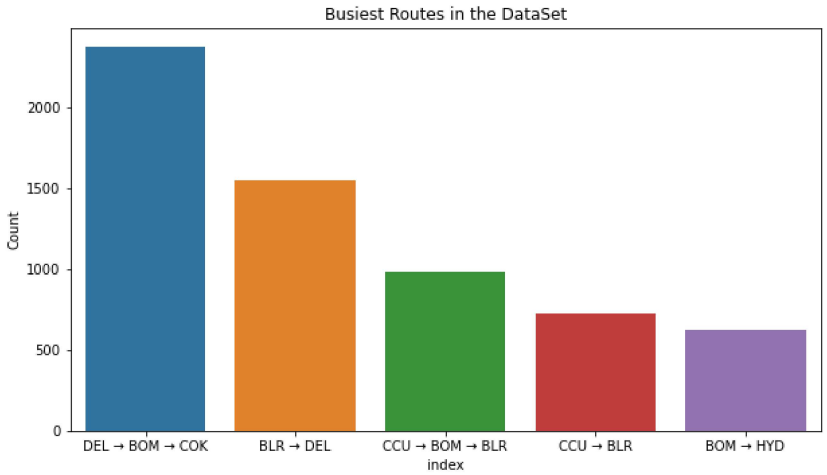


4.Busiest Air Routes in the Dataset :

In [20]:
```
busy_routes = df['route'].value_counts().reset_index().set_index('index')
busy_routes = busy_routes.head()
print(busy_routes)
```

```
                 route
index
DEL → BOM → COK   2376
BLR → DEL         1552
CCU → BOM → BLR    979
CCU → BLR          724
BOM → HYD          621
```

In [21]:
```
plt.figure(figsize=(10,5.5))
sns.barplot(x=busy_routes.index, y=busy_routes.route)
plt.title('Busiest Routes in the DataSet')
plt.ylabel('Count')
plt.show()
```

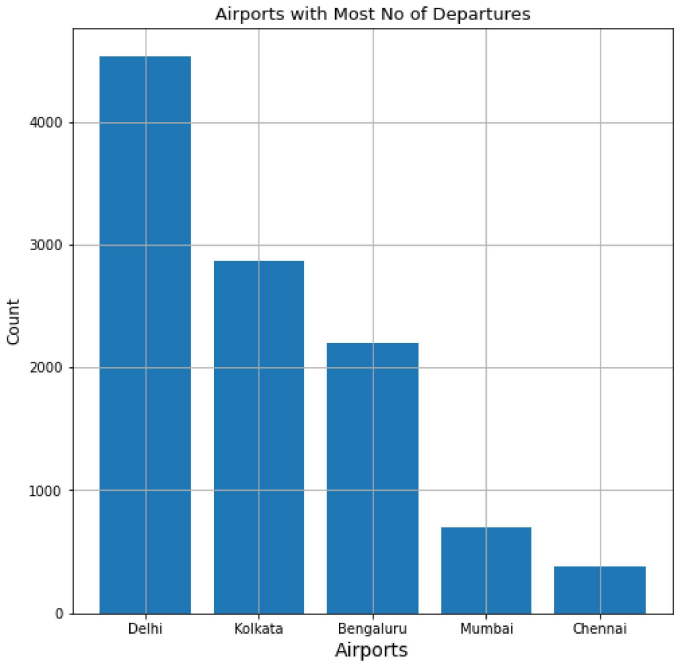


Conclusions :

* Delhi to Cochin Via Mumbai is the Busiest Route With 2376 Flights.
* Followed by Bengaluru to Delhi

5.Airports with Most No of Departure :

In [22]:
```
count = df.source.value_counts()
departure = np.array(['Delhi','Kolkata','Bengaluru','Mumbai','Chennai'])
fig, ax = plt.subplots()
ax.bar(departure,count)
ax.set_title('Airports with Most No of Departures',fontdict={'size':13})
ax.set_xlabel('Airports',fontdict={'size':14})
ax.set_ylabel('Count',fontdict={'size':12})
ax.grid()
fig.set_size_inches(8,8)
plt.show()
```

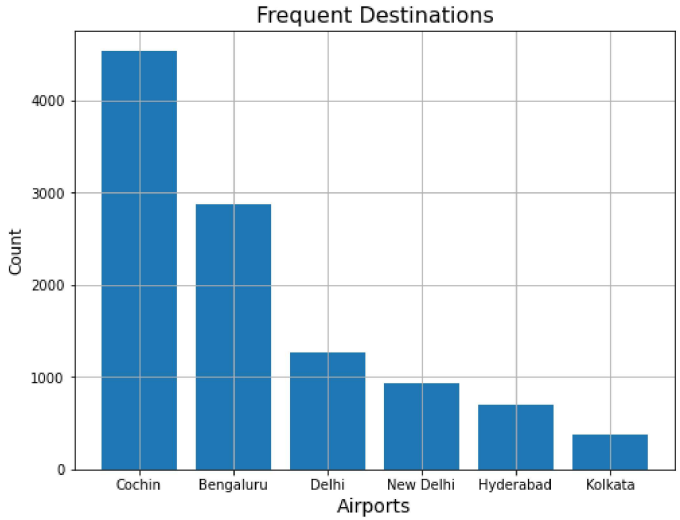


Conclusions :

* Delhi is the Busiest Airport with 4536 Flights followed by Kolkata & Bengaluru Airports.

6.Frequent Destination :

In [24]:
```
count1 = df.destination.value_counts()
departure = np.array(['Cochin','Bengaluru','Delhi','New Delhi','Hyderabad','Kolkata'])
fig, ax = plt.subplots()
ax.bar(departure,count1)
ax.set_title('Frequent Destinations',fontdict={'size':16})
ax.set_xlabel('Airports',fontdict={'size':14})
ax.set_ylabel('Count',fontdict={'size':12})
ax.grid()
fig.set_size_inches(8,6)
plt.show()
```



Conclusions :

* Cochin is the Frequent Destination followed by Bengaluru & Delhi

7.Peak Day of the Month

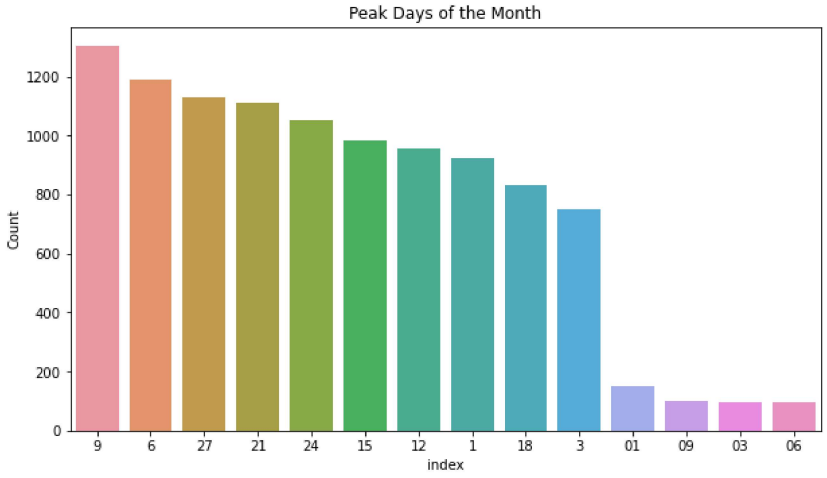
In [26]:
```
df_1 = df.copy()
df_1['date']=df_1['date_of_journey'].str.split('/').str[0]
df_1['month']=df_1['date_of_journey'].str.split('/').str[1]
df_1['year']=df_1['date_of_journey'].str.split('/').str[2]
```

In [28]:
```
df_1.head().T
```

Out[28]:

	0	1	2	3	4
airline	IndiGo	Air India	Jet Airways	IndiGo	IndiGo
date_of_journey	24/03/2019	1/05/2019	9/06/2019	12/05/2019	01/03/2019
source	Banglore	Kolkata	Delhi	Kolkata	Banglore
destination	New Delhi	Banglore	Cochin	Banglore	New Delhi
route	BLR → DEL	CCU → IXR → BBI → BLR	DEL → LKO → BOM → COK	CCU → NAG → BLR	BLR → NAG → DEL
dep_time	22:20	05:50	09:25	18:05	16:50
arrival_time	01:10 22 Mar	13:15	04:25 10 Jun	23:30	21:35
duration	2h 50m	7h 25m	19h	5h 25m	4h 45m
total_stops	non-stop	2 stops	2 stops	1 stop	1 stop
additional_info	No info	No info	No info	No info	No info
price	3897	7662	13882	6218	13302
date	24	1	9	12	01
month	03	05	06	05	03
year	2019	2019	2019	2019	2019

In [29]:
```
peak_month = df_1['date'].value_counts().reset_index().set_index('index')
plt.figure(figsize=(10,5.5))
sns.barplot(x=peak_month.index, y=peak_month.date)
plt.title('Peak Days of the Month')
plt.ylabel('Count')
plt.show()
```

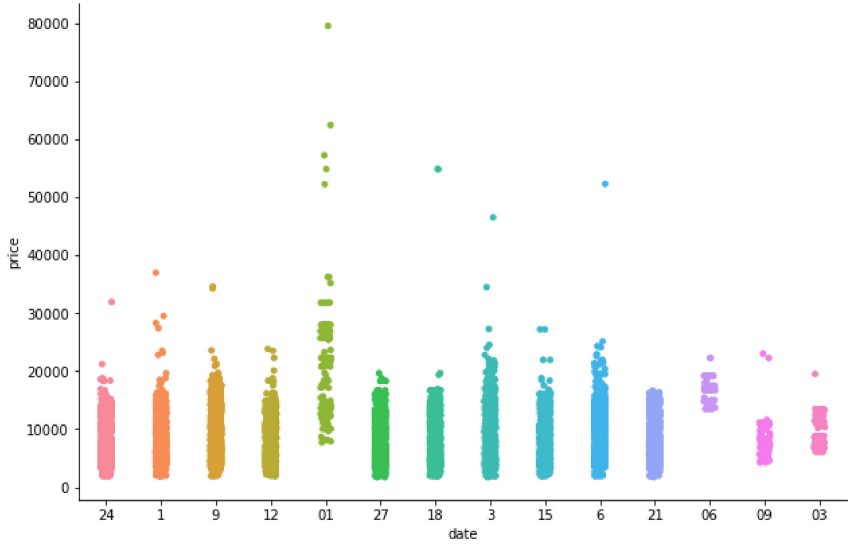


Conclusions :

* Day 9 is the Peak Day of the Month

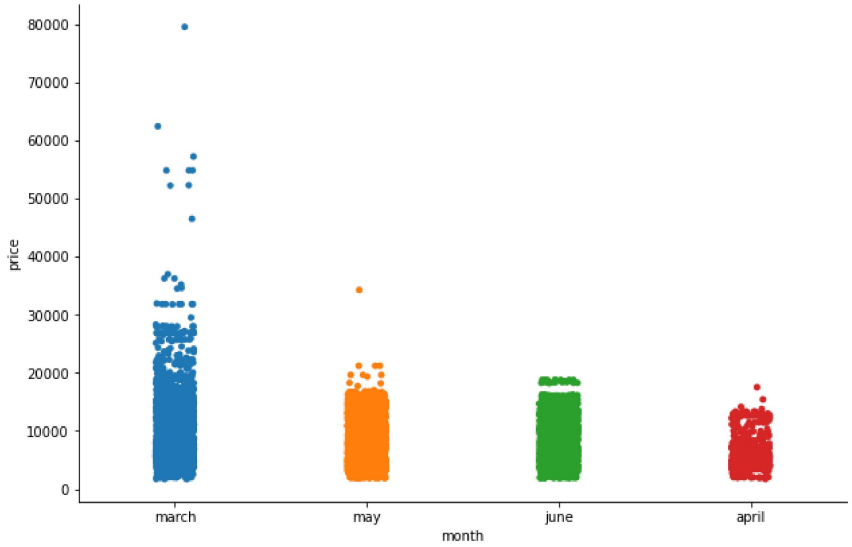
8.Tariff on Individual Days :

In [30]:
```
tariff = sns.catplot(x='date', y='price', data=df_1)
tariff.fig.set_figwidth(10)
tariff.fig.set_figheight(6)
```



9.Tariff on Individual Month :

In [32]:
```
df_1['month']=df_1['month'].map({'03':'march','05':'may','06':'june','04':'april'})
tariff = sns.catplot(x='month', y='price', data=df_1)
tariff.fig.set_figwidth(10)
tariff.fig.set_figheight(6)
```



Conclusion :

* March has the Highest Tariffs, While April has the Lowest.

In [34]:
```
#Splitting Departure Time into Hour and Min Column:

df_1['dept_hour']=df_1['dep_time'].str.split(':').str[0]
df_1['dept_min']=df_1['dep_time'].str.split(':').str[1]

df_1['dept_hour']=df_1['dept_hour'].astype(int)
df_1['dept_min']=df_1['dept_min'].astype(int)

#Splitting Arrival Time into Hour and Min Column:

df_1['arrival_hour']=df_1['arrival_time'].str.split(':').str[0]
df_1['arrival_min']=df_1['arrival_time'].str.split(':').str[1]

df_1['arrival_hour']=df_1['arrival_hour'].astype(int)
df_1['arrival_min']=df_1['arrival_hour'].astype(int)
```

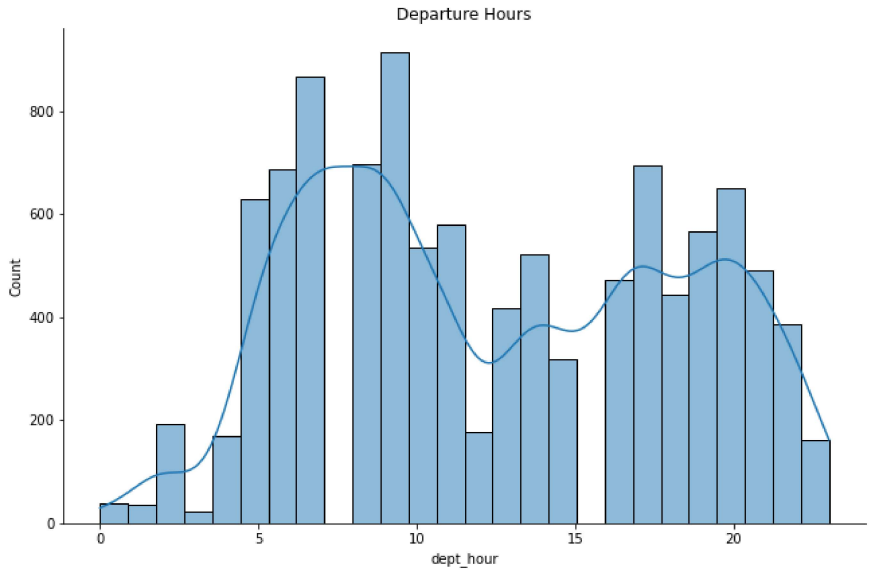
In [35]:
```
df_1.head().T
```

Out[35]:

	0	1	2	3	4
airline	IndiGo	Air India	Jet Airways	IndiGo	IndiGo
date_of_journey	24/03/2019	1/05/2019	9/06/2019	12/05/2019	01/03/2019
source	Banglore	Kolkata	Delhi	Kolkata	Banglore
destination	New Delhi	Banglore	Cochin	Banglore	New Delhi
route	BLR → DEL	CCU → IXR → BBI → BLR	DEL → LKO → BOM → COK	CCU → NAG → BLR	BLR → NAG → DEL
dep_time	22:20	05:50	09:25	18:05	16:50
arrival_time	01:10 22 Mar	13:15	04:25 10 Jun	23:30	21:35
duration	2h 50m	7h 25m	19h	5h 25m	4h 45m
total_stops	non-stop	2 stops	2 stops	1 stop	1 stop
additional_info	No info	No info	No info	No info	No info
price	3897	7662	13882	6218	13302
date	24	1	9	12	01
month	march	may	june	may	march
year	2019	2019	2019	2019	2019
dept_hour	22	5	9	18	16
dept_min	20	50	25	5	50
arrival_hour	1	13	4	23	21
arrival_min	1	13	4	23	21

10.Peak Hour for Departure :

In [36]:
```
dept_hour = sns.displot(x=df_1['dept_hour'], data=df_1, kde=True)
plt.title('Departure Hours')
dept_hour.fig.set_figwidth(10)
dept_hour.fig.set_figheight(6)
```

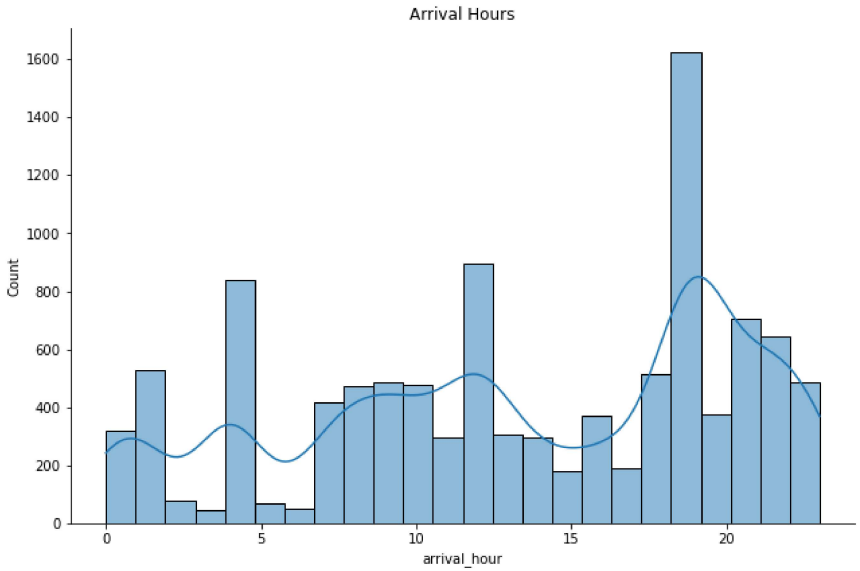


Conclusions :

* 9 AM is the Peak Hour for Departure.

11.Peak Hour for Arrival :

In [38]:
```
arr_hour = sns.displot(x=df_1['arrival_hour'], data=df_1, kde=True)
plt.title('Arrival Hours')
arr_hour.fig.set_figwidth(10)
arr_hour.fig.set_figheight(6)
```

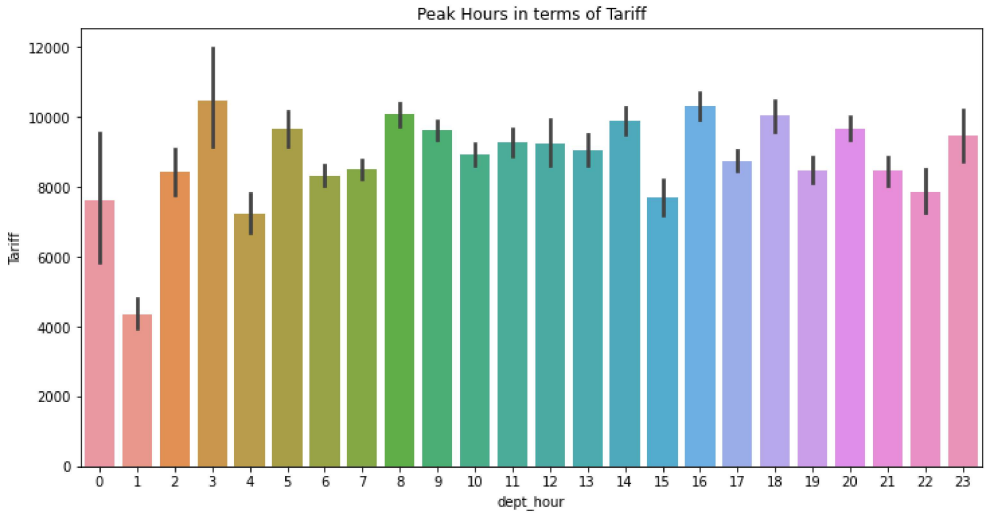


Conclusion :

* 6 PM is the Peak Hour for Arrival.

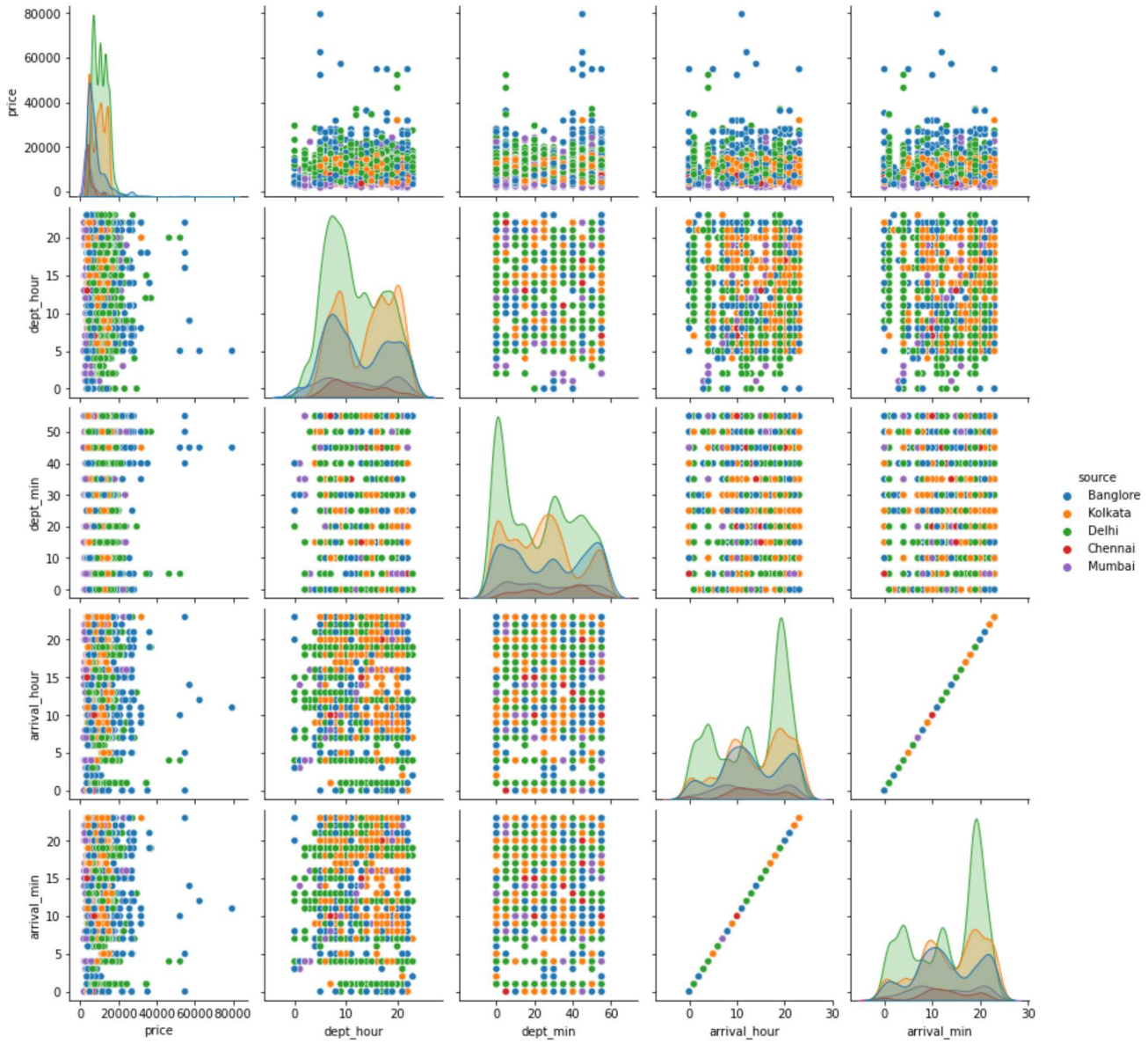
12.Peak Hours in terms of Tariff :

In [39]:
```
plt.figure(figsize=(12,6))
sns.barplot(x=df_1.dept_hour, y=df_1.price)
plt.title('Peak Hours in terms of Tariff')
plt.ylabel('Tariff')
plt.show()
```



13.Plotting Cities with Other Variables :

In [40]:
```
sns.pairplot(df_1,hue = 'source')
plt.show()
```



In [ ]: