In [56]:
```
#import required libraries
import sys
import numpy as np #Linear Algebra
import pandas as pd #Data Processing
import seaborn as sns #Data Visualizaton.
import pickle
%matplotlib inline
from matplotlib import pyplot as plt
from sklearn.preprocessing import LabelEncoder #LabelEncoding From Sklearm
from sklearn.preprocessing import OneHotEncoder #One-Hot Encoding From Sklearn
from sklearn.model_selection import train_test_split #split Data in Train & Test Array
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier#ml Algorithm
from sklearn.metrics import accuracy_score #Calculate Accuracy Score
import sklearn.metrics as metrics #Confusion Matrix
```

In [57]:
```
#import dataset
df=pd.read_csv('C:\\Users\\ELCOT\\Downloads\\flightdata.csv')
```

In [8]:
```
df
```

In [8]:
```
df
```

Out[8]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIME	ARR_TIME	ARR_DEL
0	2016	1	1	1	5	DL	N836DN	1399	10397	ATL	...	2143	2102.0	-4
1	2016	1	1	1	5	DL	N964DN	1476	11433	DTW	...	1435	1439.0	
2	2016	1	1	1	5	DL	N813DN	1597	10397	ATL	...	1215	1142.0	-3
3	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	...	1335	1345.0	1
4	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	...	607	615.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
11226	2016	4	12	30	5	DL	N940DL	1715	11433	DTW	...	1223	1148.0	-3
11227	2016	4	12	30	5	DL	N836DN	1770	14747	SEA	...	2046	2100.0	1
11228	2016	4	12	30	5	DL	N583NW	1823	11433	DTW	...	2210	2154.0	-1
11229	2016	4	12	30	5	DL	N554NW	1901	10397	ATL	...	1806	1801.0	-
11230	2016	4	12	30	5	DL	N843DN	2005	10397	ATL	...	925	913.0	-1

11231 rows × 26 columns

In [58]:
```
#Analyze the data
```

In [9]:
```
df.info()
```

```
RangeIndex: 11231 entries, 0 to 11230
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   YEAR                  11231 non-null  int64
 1   QUARTER               11231 non-null  int64
 2   MONTH                 11231 non-null  int64
 3   DAY_OF_MONTH          11231 non-null  int64
 4   DAY_OF_WEEK           11231 non-null  int64
 5   UNIQUE_CARRIER        11231 non-null  object
 6   TAIL_NUM              11231 non-null  object
 7   FL_NUM                11231 non-null  int64
 8   ORIGIN_AIRPORT_ID     11231 non-null  int64
 9   ORIGIN                11231 non-null  object
 10  DEST_AIRPORT_ID       11231 non-null  int64
 11  DEST                  11231 non-null  object
 12  CRS_DEP_TIME          11231 non-null  int64
 13  DEP_TIME              11124 non-null  float64
 14  DEP_DELAY             11124 non-null  float64
 15  DEP_DEL15             11124 non-null  float64
 16  CRS_ARR_TIME          11231 non-null  int64
 17  ARR_TIME              11116 non-null  float64
 18  ARR_DELAY             11043 non-null  float64
 19  ARR_DEL15             11043 non-null  float64
 20  CANCELLED             11231 non-null  float64
 21  DIVERTED              11231 non-null  float64
 22  CRS_ELAPSED_TIME      11231 non-null  float64
 23  ACTUAL_ELAPSED_TIME   11043 non-null  float64
 24  DISTANCE              11231 non-null  float64
 25  Unnamed: 25           0 non-null      float64
dtypes: float64(12), int64(10), object(4)
memory usage: 2.2+ MB
```

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

Out[10]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
0	2016	1	1	1	5	DL	N836DN	1399	10397	ATL	...	2143	2102.0	-41.0
1	2016	1	1	1	5	DL	N964DN	1476	11433	DTW	...	1435	1439.0	4.0
2	2016	1	1	1	5	DL	N813DN	1597	10397	ATL	...	1215	1142.0	-33.0
3	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	...	1335	1345.0	10.0
4	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	...	607	615.0	8.0

5 rows × 26 columns

In [11]:
```
df.describe()
```

Out[11]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_NUM	ORIGIN_AIRPORT_ID	DEST_AIRPORT_ID	CRS_DEP_TIME	DEP_TIME	...	CRS_ARR_TIM
count	11231.0	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11124.000000	...	11231.00000
mean	2016.0	2.544475	6.628973	15.790758	3.960199	1334.325617	12334.516695	12302.274508	1320.798326	1327.189410	...	1537.31279
std	0.0	1.090701	3.354678	8.782056	1.995257	811.875227	1595.026510	1601.988550	490.737845	500.306462	...	502.51249
min	2016.0	1.000000	1.000000	1.000000	1.000000	7.000000	10397.000000	10397.000000	10.000000	1.000000	...	2.00000
25%	2016.0	2.000000	4.000000	8.000000	2.000000	624.000000	10397.000000	10397.000000	905.000000	905.000000	...	1130.00000
50%	2016.0	3.000000	7.000000	16.000000	4.000000	1267.000000	12478.000000	12478.000000	1320.000000	1324.000000	...	1559.00000
75%	2016.0	3.000000	9.000000	23.000000	6.000000	2032.000000	13487.000000	13487.000000	1735.000000	1739.000000	...	1952.00000
max	2016.0	4.000000	12.000000	31.000000	7.000000	2853.000000	14747.000000	14747.000000	2359.000000	2400.000000	...	2359.00000

8 rows × 22 columns

In [59]:
```
#Handling missing values
```

In [12]:
```
df.isnull().any()
```

Out[12]:
```
YEAR                 False
QUARTER              False
MONTH                False
DAY_OF_MONTH         False
DAY_OF_WEEK          False
UNIQUE_CARRIER       False
TAIL_NUM             False
FL_NUM               False
ORIGIN_AIRPORT_ID    False
ORIGIN               False
DEST_AIRPORT_ID      False
DEST                 False
CRS_DEP_TIME         False
DEP_TIME              True
DEP_DELAY             True
DEP_DEL15             True
CRS_ARR_TIME         False
ARR_TIME              True
ARR_DELAY             True
ARR_DEL15             True
CANCELLED            False
DIVERTED             False
CRS_ELAPSED_TIME     False
ACTUAL_ELAPSED_TIME   True
DISTANCE             False
Unnamed: 25           True
dtype: bool
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
YEAR                    0
QUARTER                 0
MONTH                   0
DAY_OF_MONTH            0
DAY_OF_WEEK             0
UNIQUE_CARRIER          0
TAIL_NUM                0
FL_NUM                  0
ORIGIN_AIRPORT_ID       0
ORIGIN                  0
DEST_AIRPORT_ID         0
DEST                    0
CRS_DEP_TIME            0
DEP_TIME              107
DEP_DELAY             107
DEP_DEL15             107
CRS_ARR_TIME            0
ARR_TIME              115
ARR_DELAY             188
ARR_DEL15             188
CANCELLED               0
DIVERTED                0
CRS_ELAPSED_TIME        0
ACTUAL_ELAPSED_TIME   188
DISTANCE                0
Unnamed: 25         11231
dtype: int64
```

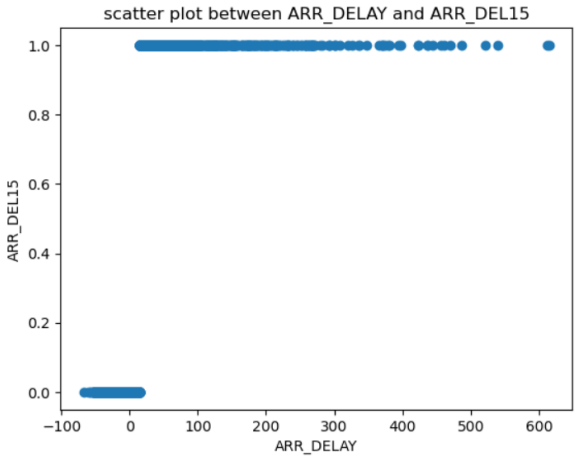
In [14]:
```
df['DEST'].unique()
```

Out[14]:
```
array(['SEA', 'MSP', 'DTW', 'ATL', 'JFK'], dtype=object)
```

In [60]:
```
#Data visualization
```

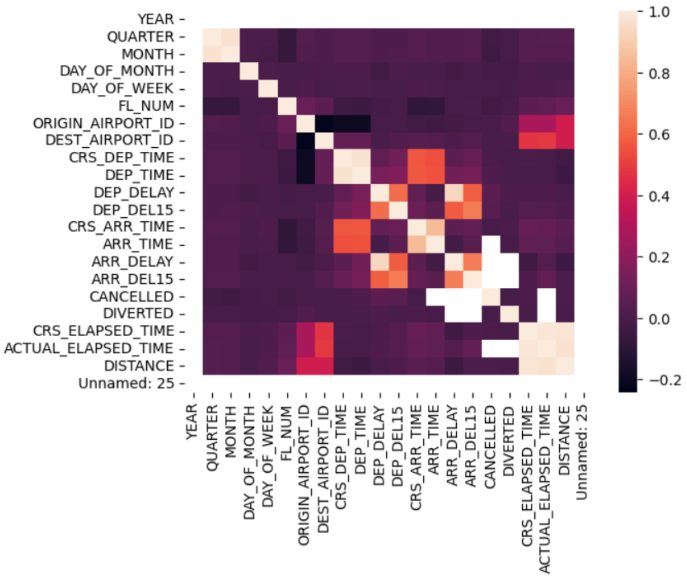
In [18]:
```
plt.scatter(df['ARR_DELAY'],df['ARR_DEL15'])
plt.xlabel('ARR_DELAY')
plt.ylabel('ARR_DEL15')
plt.title('scatter plot between ARR_DELAY and ARR_DEL15')
```

Out[18]:
```
Text(0.5, 1.0, 'scatter plot between ARR_DELAY and ARR_DEL15')
```



In [15]:
```
sns.heatmap(df.corr())
```

Out[15]:



In [61]:
```
#Droping unnecessary columns
```

In [19]:
```
dataset = df.drop('Unnamed: 25',axis=1)
dataset.isnull().sum()
```

Out[19]:
```
YEAR                   0
QUARTER                0
MONTH                  0
DAY_OF_MONTH           0
DAY_OF_WEEK            0
UNIQUE_CARRIER         0
TAIL_NUM               0
FL_NUM                 0
ORIGIN_AIRPORT_ID      0
ORIGIN                 0
DEST_AIRPORT_ID        0
DEST                   0
CRS_DEP_TIME           0
DEP_TIME             107
DEP_DELAY            107
DEP_DEL15            107
CRS_ARR_TIME           0
ARR_TIME             115
ARR_DELAY            188
ARR_DEL15            188
CANCELLED              0
DIVERTED               0
CRS_ELAPSED_TIME       0
ACTUAL_ELAPSED_TIME  188
DISTANCE               0
dtype: int64
```

In [20]:
```
dataset = df[["FL_NUM", "MONTH", "DAY_OF_MONTH", "DAY_OF_WEEK", "ORIGIN", "DEST", "CRS_ARR_TIME","DEP_DEL15", "ARR_DEL15"]]
dataset.isnull().sum()
```

Out[20]:
```
FL_NUM            0
MONTH             0
DAY_OF_MONTH      0
DAY_OF_WEEK       0
ORIGIN            0
DEST              0
CRS_ARR_TIME      0
DEP_DEL15       107
ARR_DEL15       188
dtype: int64
```

In [21]:
```
dataset = df.fillna ({'ARR_DEL15': 1})
dataset = df.fillna ({'DEP_DEL15': 0})
dataset.iloc[177:185]
```

Out[21]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
177	2016	1	1	9	6	DL	N3743H	2834	13487	MSP	...	852	1151.0	NaN
178	2016	1	1	9	6	DL	N975AT	2839	11433	DTW	...	1724	1709.0	-15.0
179	2016	1	1	10	7	DL	N924DN	86	13487	MSP	...	1632	NaN	NaN
180	2016	1	1	10	7	DL	N671DN	87	11433	DTW	...	1649	1703.0	14.0
181	2016	1	1	10	7	DL	N319NB	423	12478	JFK	...	1600	1607.0	7.0
182	2016	1	1	10	7	DL	N587NW	440	12478	JFK	...	849	835.0	-14.0
183	2016	1	1	10	7	DL	N813DN	485	12478	JFK	...	1945	1955.0	10.0
184	2016	1	1	10	7	DL	N922DX	557	13487	MSP	...	912	1500.0	NaN

8 rows × 26 columns

In [23]:
```
import math
for index, row in df.iterrows():
    df.loc[index, 'CRS_ARR_TIME'] = math.floor(row['CRS_ARR_TIME'] / 100)
df.head()
```

Out[23]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
0	2016	1	1	1	5	DL	N836DN	1399	10397	ATL	...	0	2102.0	-41.0
1	2016	1	1	1	5	DL	N964DN	1476	11433	DTW	...	0	1439.0	4.0
2	2016	1	1	1	5	DL	N813DN	1597	10397	ATL	...	0	1142.0	-33.0
3	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	...	0	1345.0	10.0
4	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	...	0	615.0	8.0

5 rows × 26 columns

In [62]:
```
#One hot encoder
```

In [48]:
```
from sklearn.preprocessing import OneHotEncoder
oh=OneHotEncoder()
```

In [49]:
```
z=oh.fit_transform(x[:,4:5]).toarray()
t=oh.fit_transform(x[:,5:6]).toarray()
```

In [50]:
```
z
```

Out[50]:
```
array([[0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 1., 0., 0.],
       ...,
       [0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 1., 0., 0.]])
```

In [51]:
```
t
```

Out[51]:
```
array([[1.],
       [1.],
       [1.],
       ...,
       [1.],
       [1.],
       [1.]])
```

In [63]:
```
#Label encoder
```

In [24]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['DEST'] = le.fit_transform(df [ 'DEST'])
df['ORIGIN'] =le.fit_transform(df['ORIGIN'])
```

In [25]:
```
df.head (5)
```

Out[25]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
0	2016	1	1	1	5	DL	N836DN	1399	10397	0	...	0	2102.0	-41.0
1	2016	1	1	1	5	DL	N964DN	1476	11433	1	...	0	1439.0	4.0
2	2016	1	1	1	5	DL	N813DN	1597	10397	0	...	0	1142.0	-33.0
3	2016	1	1	1	5	DL	N587NW	1768	14747	4	...	0	1345.0	10.0
4	2016	1	1	1	5	DL	N836DN	1823	14747	4	...	0	615.0	8.0

5 rows × 26 columns

In [26]:
```
x=df.iloc[:, 4:5].values
y=df.iloc[:, 5:6].values
x.shape
```

Out[26]: (11231, 1)

In [27]:
```
y.shape
```

Out[27]: (11231, 1)

In [28]:
```
df =pd.get_dummies (dataset, columns=['ORIGIN', 'DEST'])
df.head()
```

Out[28]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	DEST_AIRPORT_ID	...	ORIGIN_ATL	ORIGIN_DTW	OR
0	2016	1	1	1	5	DL	N836DN	1399	10397	14747	...	1	0	
1	2016	1	1	1	5	DL	N964DN	1476	11433	13487	...	0	1	
2	2016	1	1	1	5	DL	N813DN	1597	10397	14747	...	1	0	
3	2016	1	1	1	5	DL	N587NW	1768	14747	13487	...	0	0	
4	2016	1	1	1	5	DL	N836DN	1823	14747	11433	...	0	0	

5 rows × 34 columns

In [64]:
```
#split data into dependant and independent variables
```

In [29]:
```
x=df.iloc[:, 0:8].values
y=df.iloc[:, 8:9].values
```

In [ ]:
```
#splitting the dataset into trainset and test set
```

In [30]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.2, random_state=0)
```

In [31]:
```
x_test.shape
```

Out[31]: (2247, 8)

In [32]:
```
x_train.shape
```

Out[32]: (8984, 8)

In [33]:
```
y_train.shape
```

Out[33]: (8984, 1)

In [34]:
```
y_train.shape
```

Out[34]: (8984, 1)