PROJECT DEVELOPMENT PHASE

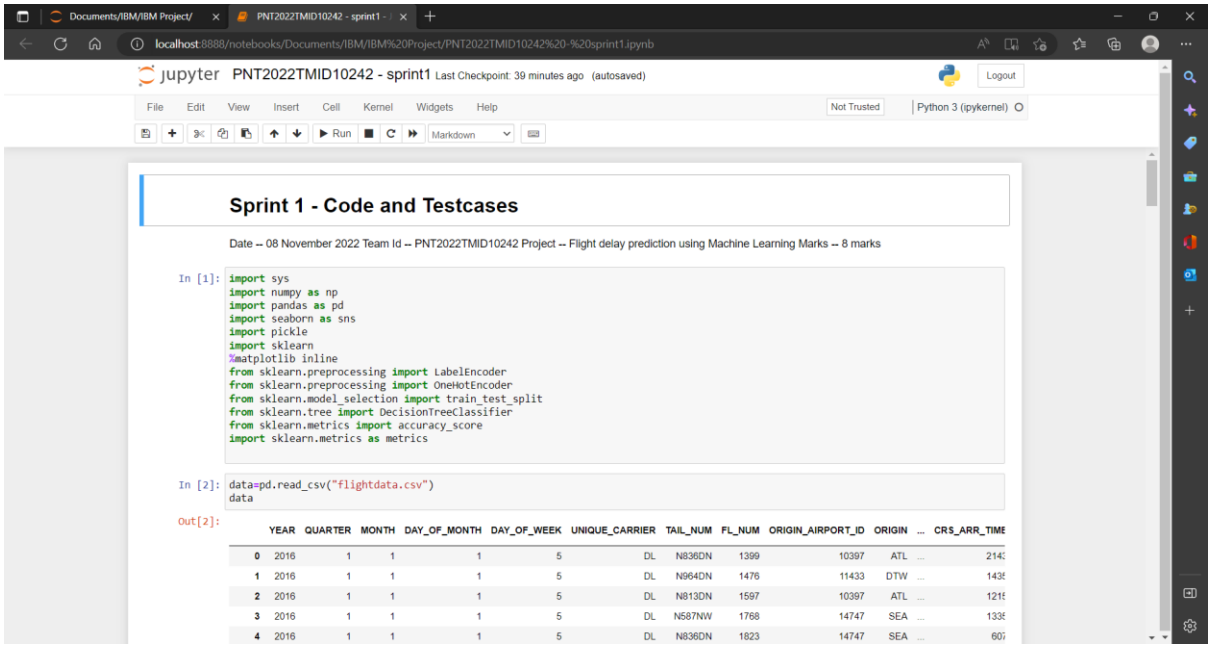
SPRINT 2 – CODE AND TESTCASE

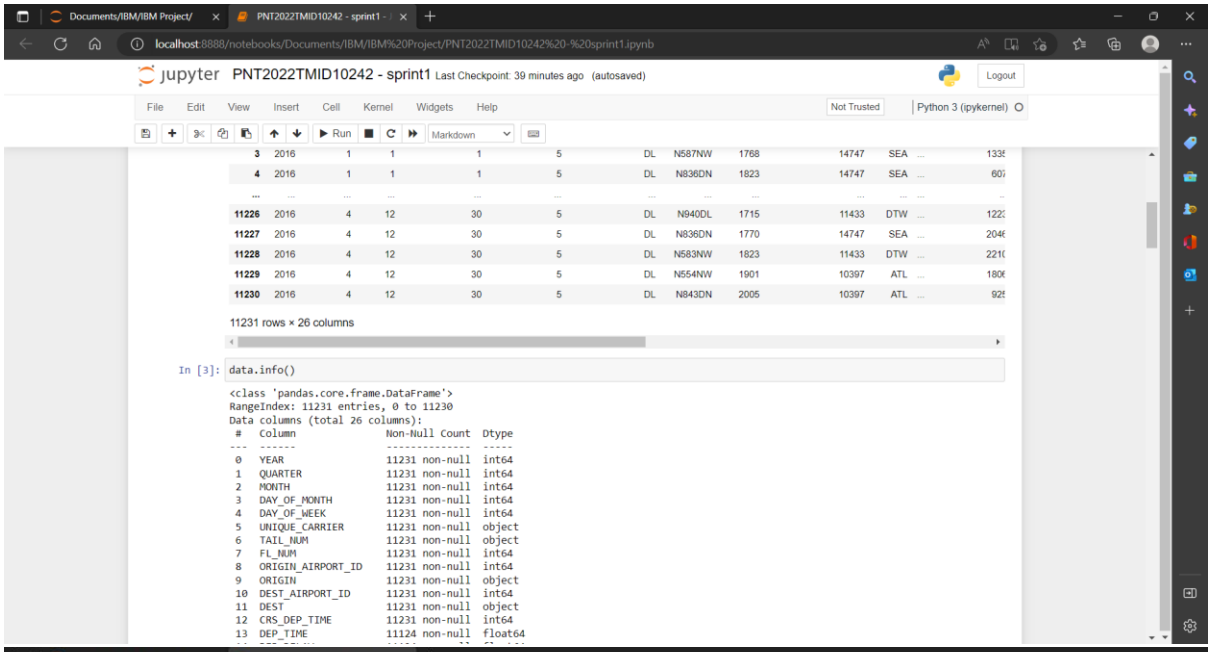
Date	10 November 2022
Team ID	PNT2022TMID10242
Project	Flight delay prediction using Machine learning
Marks	8 Marks

We have performed the uploading the Dataset and performed the Data Pre-processing and also we have split the dataset into train data and Test dataset in this Sprint development phase.

Jupyter notebook :

Screenshots :



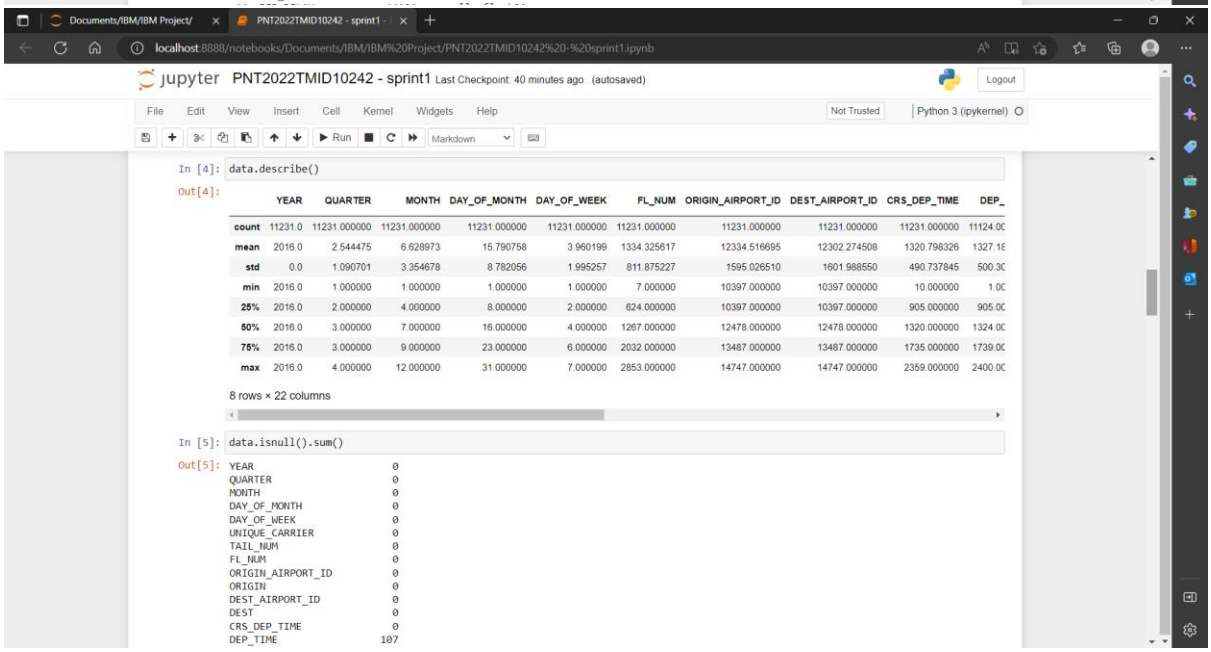


3	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	...	133
4	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	...	607
...	...	...	...	...	...	...	...	...	...	...	...	...
11226	2016	4	12	30	5	DL	N940DL	1715	11433	DTW	...	122
11227	2016	4	12	30	5	DL	N836DN	1770	14747	SEA	...	204
11228	2016	4	12	30	5	DL	N583NW	1823	11433	DTW	...	221
11229	2016	4	12	30	5	DL	N554NW	1901	10397	ATL	...	180
11230	2016	4	12	30	5	DL	N843DN	2005	10397	ATL	...	92

11231 rows × 26 columns

In [3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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
```



In [4]:
```
data.describe()
```

Out[4]:

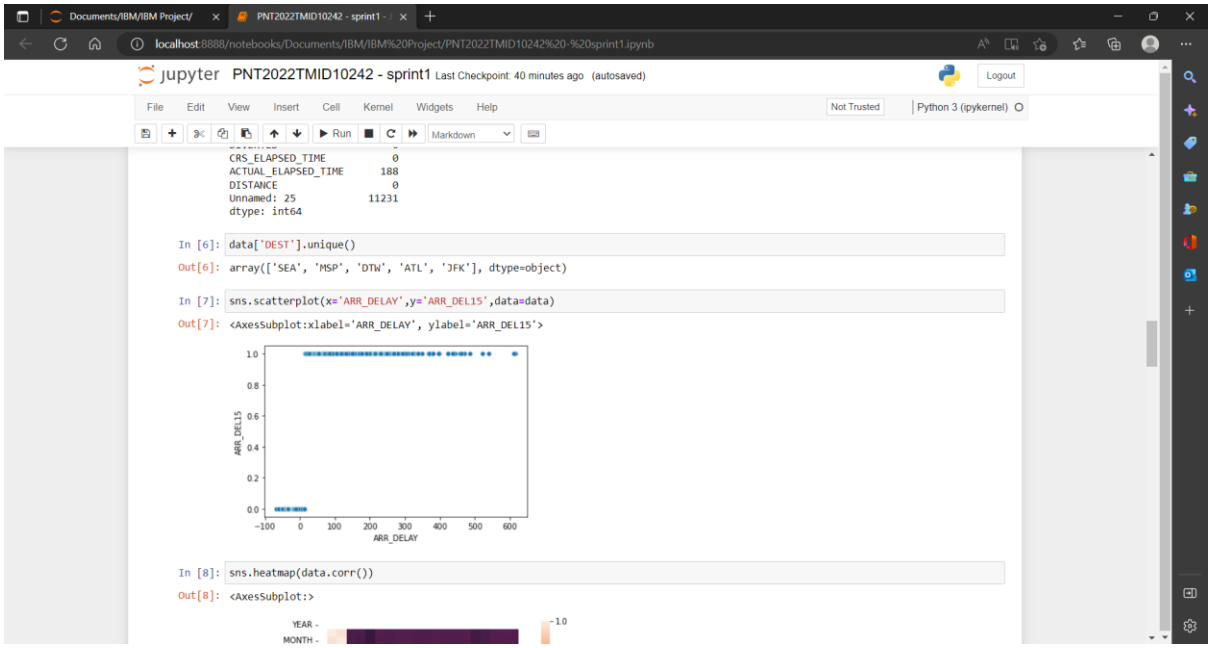
	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_NUM	ORIGIN_AIRPORT_ID	DEST_AIRPORT_ID	CRS_DEP_TIME	DEP_
count	11231.0	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11124.0
mean	2016.0	2.544475	6.628973	15.790758	3.960199	1334.325617	12334.516695	12302.274508	1320.798326	1327.1
std	0.0	1.090701	3.354678	8.782056	1.995257	811.875227	1595.026510	1601.988550	490.737845	500.3
min	2016.0	1.000000	1.000000	1.000000	1.000000	7.000000	10397.000000	10397.000000	10.000000	1.0
25%	2016.0	2.000000	4.000000	8.000000	2.000000	624.000000	10397.000000	10397.000000	905.000000	905.0
50%	2016.0	3.000000	7.000000	16.000000	4.000000	1267.000000	12478.000000	12478.000000	1320.000000	1324.0
75%	2016.0	3.000000	9.000000	23.000000	6.000000	2032.000000	13487.000000	13487.000000	1735.000000	1739.0
max	2016.0	4.000000	12.000000	31.000000	7.000000	2853.000000	14747.000000	14747.000000	2359.000000	2400.0

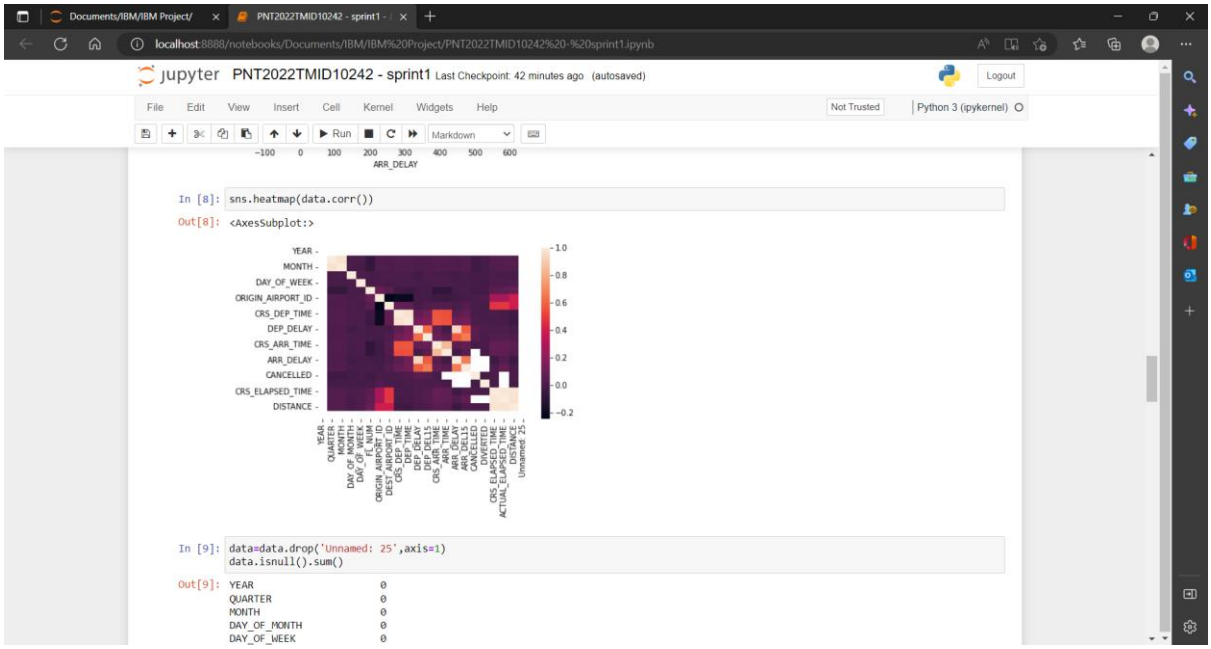
8 rows × 22 columns

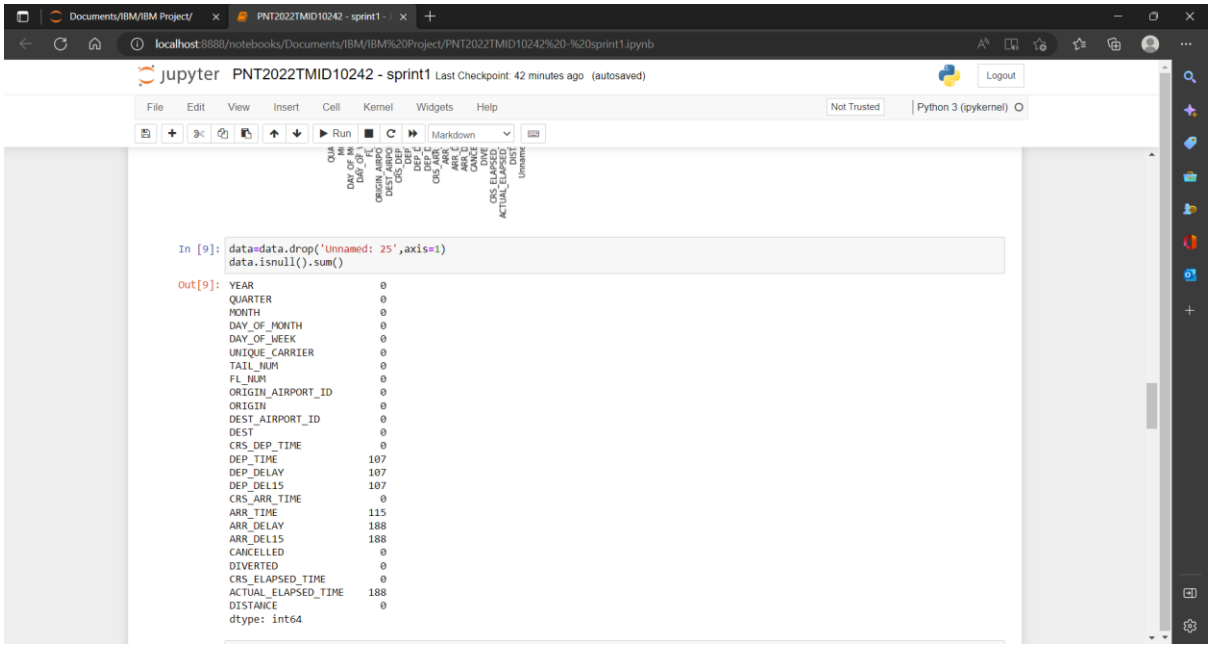
In [5]:
```
data.isnull().sum()
```

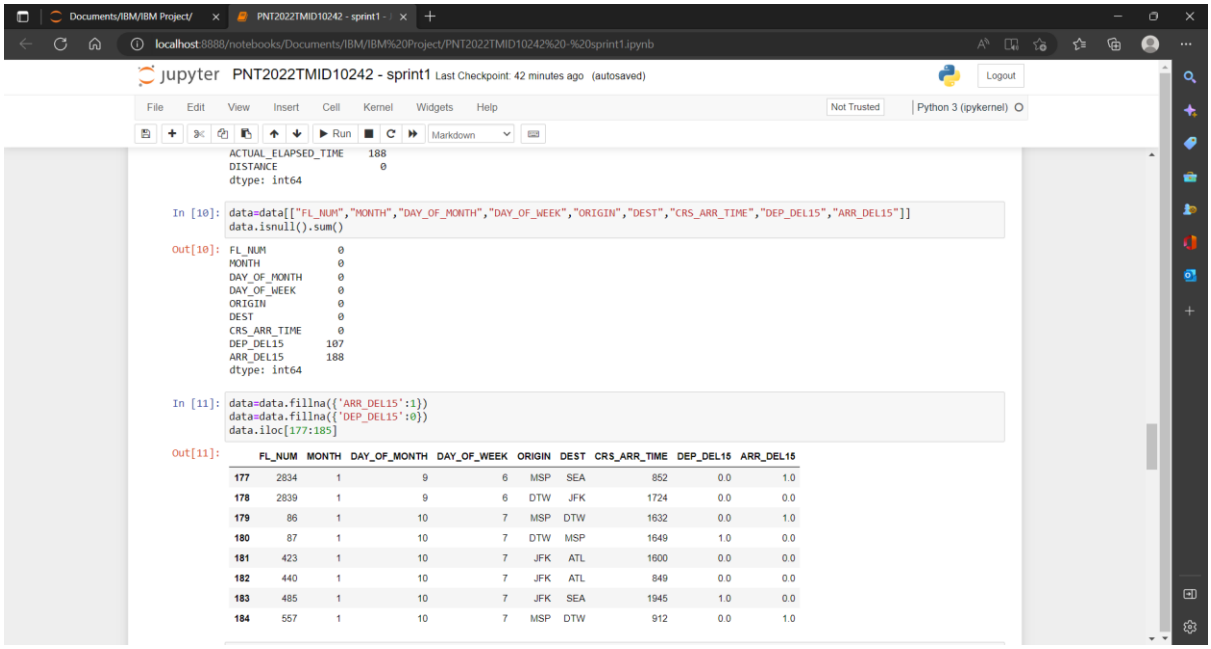
Out[5]:

```
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
```

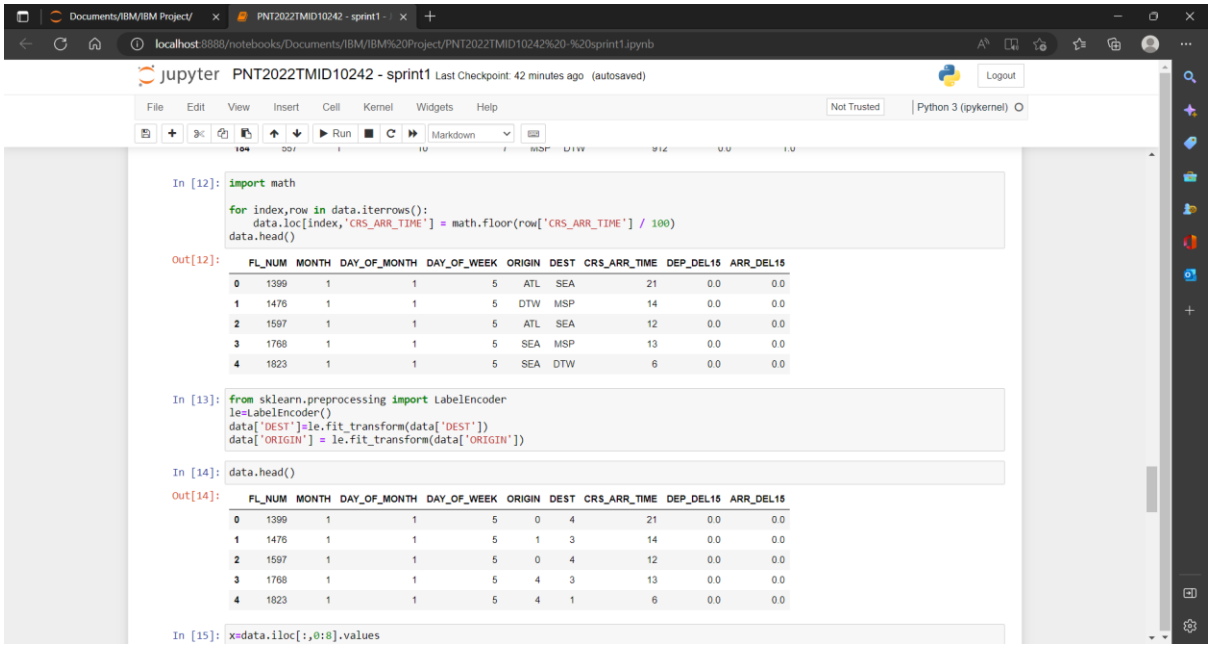

Documents/IBM/IBM Project/
PNT2022TMID10242 - sprint1
localhost:8888/notebooks/Documents/IBM/IBM%20Project/PNT2022TMID10242%20-%20sprint1.ipynb
jupyter PNT2022TMID10242 - sprint1 Last Checkpoint: 40 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
CRS_ELAPSED_TIME 0
ACTUAL_ELAPSED_TIME 188
DISTANCE 0
Unnamed: 25 11231
dtype: int64
In [6]:
data['DEST'].unique()
Out[6]:
array(['SEA', 'MSP', 'DTW', 'ATL', 'JFK'], dtype=object)
In [7]:
sns.scatterplot(x='ARR_DELAY',y='ARR_DEL15',data=data)
Out[7]:
<AxesSubplot:xlabel='ARR_DELAY', ylabel='ARR_DEL15'>
ARR_DEL15
ARR_DELAY
In [8]:
sns.heatmap(data.corr())
Out[8]:
<AxesSubplot:>
YEAR
MONTH

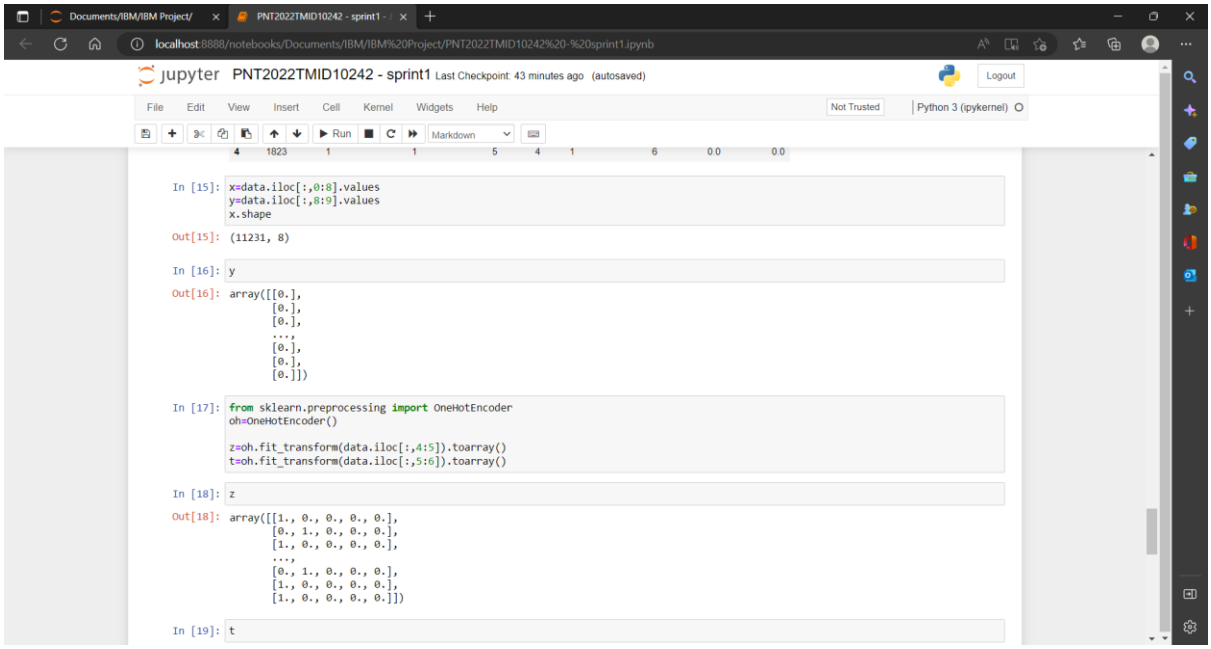

Documents/IBM/IBM Project/
PNT2022TMID10242 - sprint1
localhost:8888/notebooks/Documents/IBM/IBM%20Project/PNT2022TMID10242%20-%20sprint1.ipynb
jupyter PNT2022TMID10242 - sprint1 Last Checkpoint: 42 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
ARR_DELAY
In [8]:
sns.heatmap(data.corr())
Out[8]:
<AxesSubplot:>
YEAR
MONTH
DAY_OF_WEEK
ORIGIN_AIRPORT_ID
CRS_DEP_TIME
DEP_DELAY
CRS_ARR_TIME
ARR_DELAY
CANCELLED
CRS_ELAPSED_TIME
DISTANCE
In [9]:
data=data.drop('Unnamed: 25',axis=1)
data.isnull().sum()
Out[9]:
YEAR 0
QUARTER 0
MONTH 0
DAY_OF_MONTH 0
DAY_OF_WEEK 0

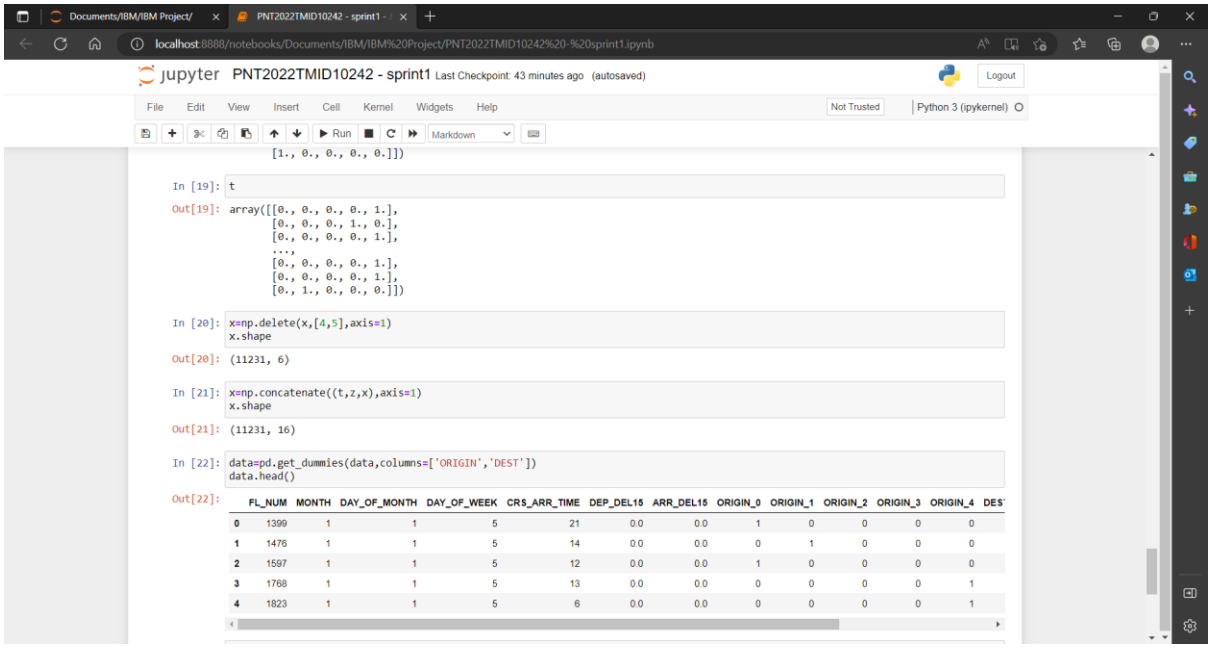


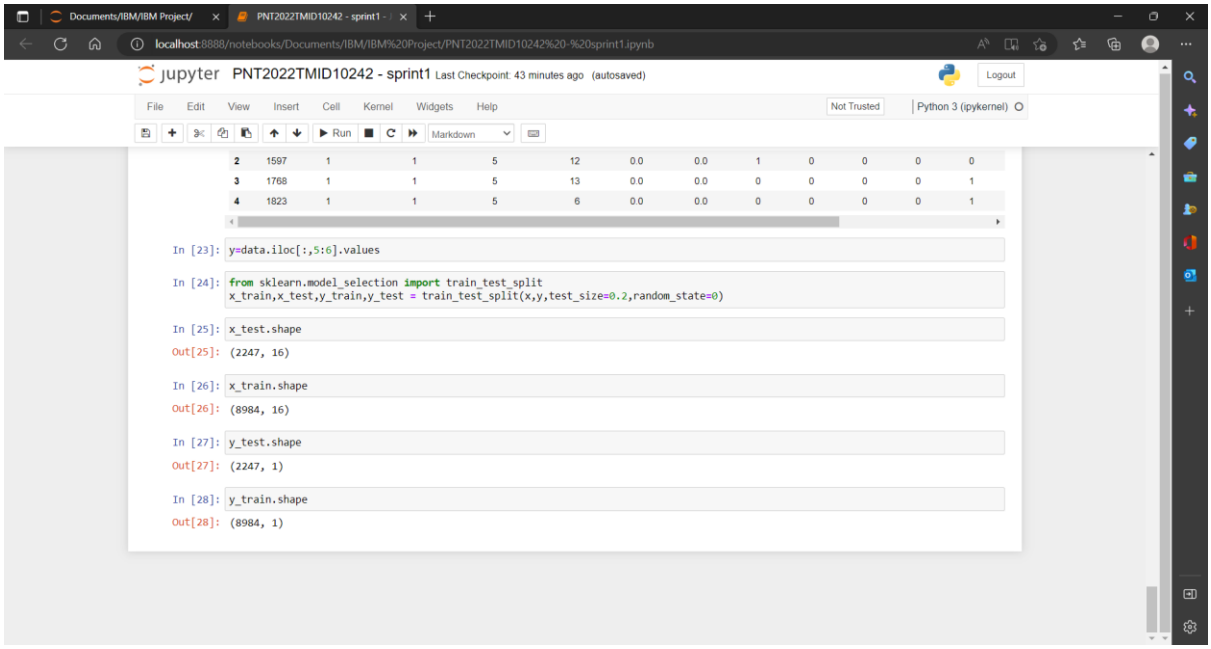


	FL_NUM	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	ORIGIN	DEST	CRS_ARR_TIME	DEP_DEL15	ARR_DEL15
177	2834	1	9	6	MSP	SEA	852	0.0	1.0
178	2839	1	9	6	DTW	JFK	1724	0.0	0.0
179	86	1	10	7	MSP	DTW	1632	0.0	1.0
180	87	1	10	7	DTW	MSP	1649	1.0	0.0
181	423	1	10	7	JFK	ATL	1600	0.0	0.0
182	440	1	10	7	JFK	ATL	849	0.0	0.0
183	485	1	10	7	JFK	SEA	1945	1.0	0.0
184	557	1	10	7	MSP	DTW	912	0.0	1.0


Documents/IBM/IBM Project/
PNT2022TMID10242 - sprint1
localhost:8888/notebooks/Documents/IBM/IBM%20Project/PNT2022TMID10242%20-%20sprint1.ipynb
jupyter PNT2022TMID10242 - sprint1 Last Checkpoint: 42 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
In [12]:
import math
for index,row in data.iterrows():
data.loc[index,'CRS_ARR_TIME'] = math.floor(row['CRS_ARR_TIME'] / 100)
data.head()
Out[12]:
FL_NUM MONTH DAY_OF_MONTH DAY_OF_WEEK ORIGIN DEST CRS_ARR_TIME DEP_DEL15 ARR_DEL15
0 1399 1 1 5 ATL SEA 21 0.0 0.0
1 1476 1 1 5 DTW MSP 14 0.0 0.0
2 1597 1 1 5 ATL SEA 12 0.0 0.0
3 1768 1 1 5 SEA MSP 13 0.0 0.0
4 1823 1 1 5 SEA DTW 6 0.0 0.0
In [13]:
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
data['DEST']=le.fit_transform(data['DEST'])
data['ORIGIN'] = le.fit_transform(data['ORIGIN'])
In [14]:
data.head()
Out[14]:
FL_NUM MONTH DAY_OF_MONTH DAY_OF_WEEK ORIGIN DEST CRS_ARR_TIME DEP_DEL15 ARR_DEL15
0 1399 1 1 5 0 4 21 0.0 0.0
1 1476 1 1 5 1 3 14 0.0 0.0
2 1597 1 1 5 0 4 12 0.0 0.0
3 1768 1 1 5 4 3 13 0.0 0.0
4 1823 1 1 5 4 1 6 0.0 0.0
In [15]:
x=data.iloc[:,0:8].values


Documents/IBM/IBM Project/
PNT2022TMID10242 - sprint1
localhost:8888/notebooks/Documents/IBM/IBM%20Project/PNT2022TMID10242%20-%20sprint1.ipynb
jupyter PNT2022TMID10242 - sprint1 Last Checkpoint: 43 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
4 1823 1 1 5 4 1 6 0.0 0.0
In [15]:
x=data.iloc[:,0:8].values
y=data.iloc[:,8:9].values
x.shape
Out[15]:
(11231, 8)
In [16]:
y
Out[16]:
array([[0.],
[0.],
[0.],
...,
[0.],
[0.],
[0.]])
In [17]:
from sklearn.preprocessing import OneHotEncoder
oh=OneHotEncoder()
z=oh.fit_transform(data.iloc[:,4:5]).toarray()
t=oh.fit_transform(data.iloc[:,5:6]).toarray()
In [18]:
z
Out[18]:
array([[1., 0., 0., 0., 0.],
[0., 1., 0., 0., 0.],
[1., 0., 0., 0., 0.],
...,
[0., 1., 0., 0., 0.],
[1., 0., 0., 0., 0.],
[1., 0., 0., 0., 0.]])
In [19]:
t


Documents/IBM/IBM Project/
PNT2022TMID10242 - sprint1
localhost:8888/notebooks/Documents/IBM/IBM%20Project/PNT2022TMID10242%20-%20sprint1.ipynb
jupyter PNT2022TMID10242 - sprint1 Last Checkpoint: 43 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
[1., 0., 0., 0., 0.]])
In [19]: t
Out[19]: array([[0., 0., 0., 0., 1.],
[0., 0., 0., 1., 0.],
[0., 0., 0., 0., 1.],
...,
[0., 0., 0., 0., 1.],
[0., 0., 0., 0., 1.],
[0., 1., 0., 0., 0.]])
In [20]: x=np.delete(x,[4,5],axis=1)
x.shape
Out[20]: (11231, 6)
In [21]: x=np.concatenate((t,z,x),axis=1)
x.shape
Out[21]: (11231, 16)
In [22]: data=pd.get_dummies(data,columns=['ORIGIN','DEST'])
data.head()
Out[22]:
FL_NUM MONTH DAY_OF_MONTH DAY_OF_WEEK CRS_ARR_TIME DEP_DEL15 ARR_DEL15 ORIGIN_0 ORIGIN_1 ORIGIN_2 ORIGIN_3 ORIGIN_4 DES
0 1399 1 1 5 21 0.0 0.0 1 0 0 0 0
1 1476 1 1 5 14 0.0 0.0 0 1 0 0 0
2 1597 1 1 5 12 0.0 0.0 1 0 0 0 0
3 1768 1 1 5 13 0.0 0.0 0 0 0 0 1
4 1823 1 1 5 6 0.0 0.0 0 0 0 0 1


Documents/IBM/IBM Project/
PNT2022TMID10242 - sprint1
localhost:8888/notebooks/Documents/IBM/IBM%20Project/PNT2022TMID10242%20-%20sprint1.ipynb
jupyter PNT2022TMID10242 - sprint1 Last Checkpoint: 43 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
2 1597 1 1 5 12 0.0 0.0 1 0 0 0 0
3 1768 1 1 5 13 0.0 0.0 0 0 0 0 1
4 1823 1 1 5 6 0.0 0.0 0 0 0 0 1
In [23]: y=data.iloc[:,5:6].values
In [24]: 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 [25]: x_test.shape
Out[25]: (2247, 16)
In [26]: x_train.shape
Out[26]: (8984, 16)
In [27]: y_test.shape
Out[27]: (2247, 1)
In [28]: y_train.shape
Out[28]: (8984, 1)