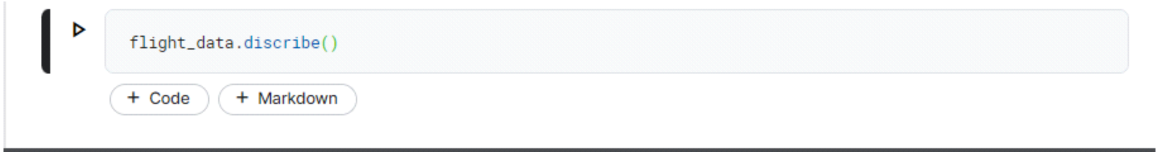
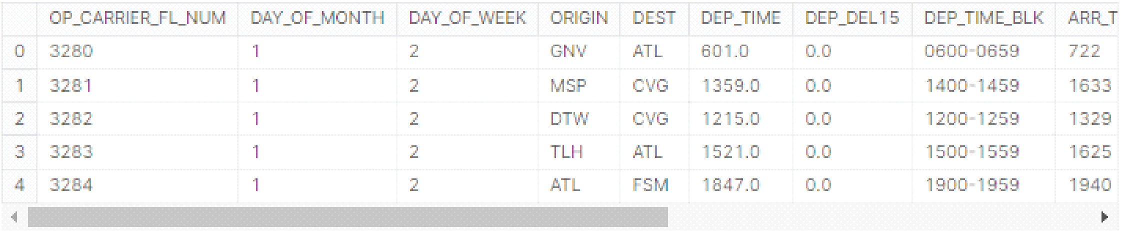
NM2023TMID32012 - Flight delay prediction

Milestone 3 :Exploratory Data Analysis

Activity 1: Descriptive statistical

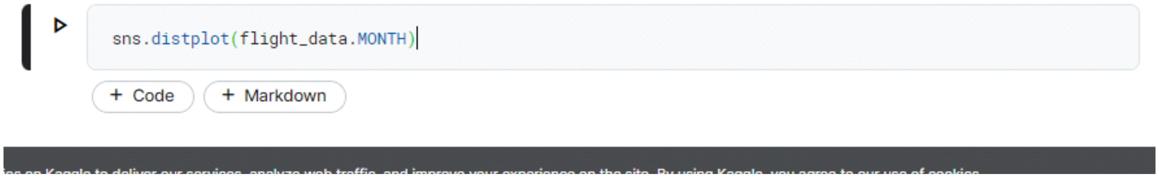


```
flight_data.discribe()
```



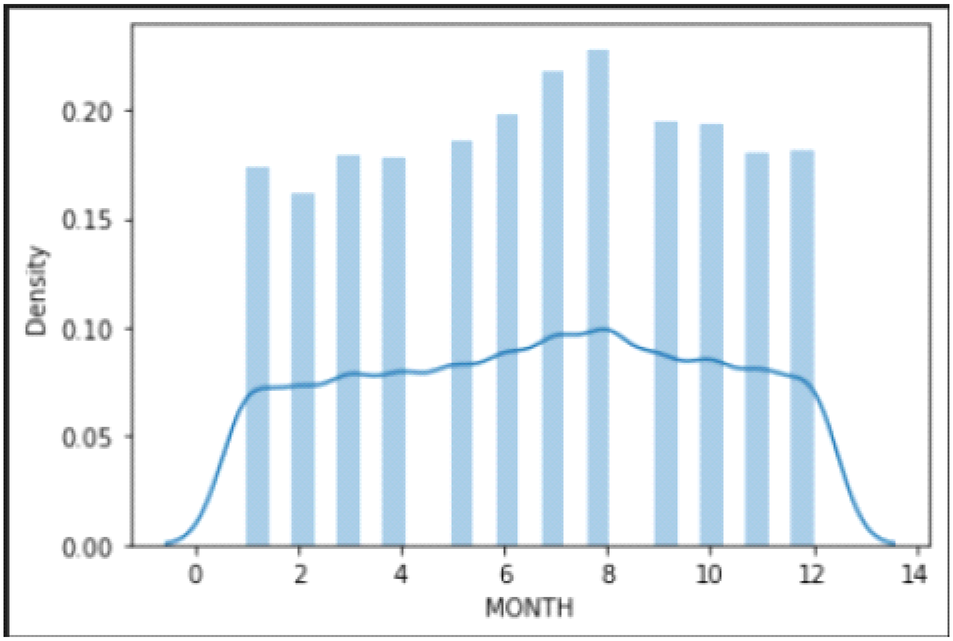
	OP_CARRIER_FL_NUM	DAY_OF_MONTH	DAY_OF_WEEK	ORIGIN	DEST	DEP_TIME	DEP_DEL15	DEP_TIME_BLK	ARR_T
0	3280	1	2	GNV	ATL	601.0	0.0	0600-0659	722
1	3281	1	2	MSP	CVG	1359.0	0.0	1400-1459	1633
2	3282	1	2	DTW	CVG	1215.0	0.0	1200-1259	1329
3	3283	1	2	TLH	ATL	1521.0	0.0	1500-1559	1625
4	3284	1	2	ATL	FSM	1847.0	0.0	1900-1959	1940

Activity 2.1: Univariate analysis

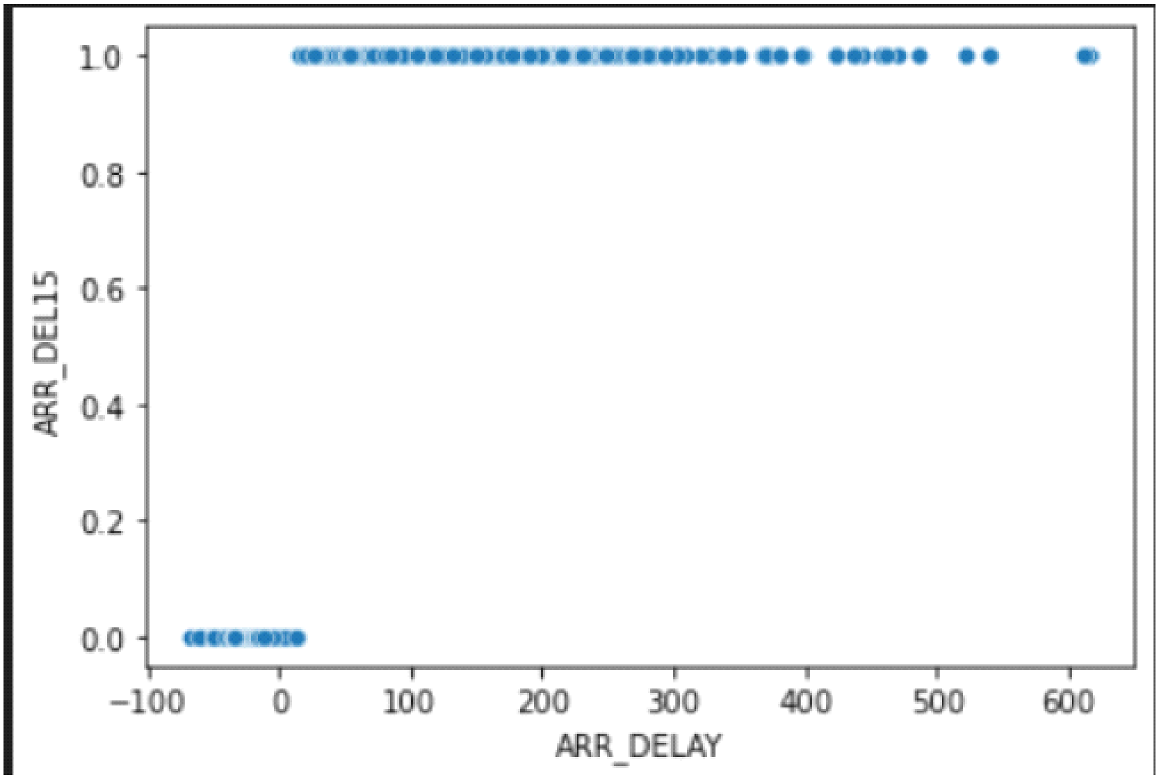


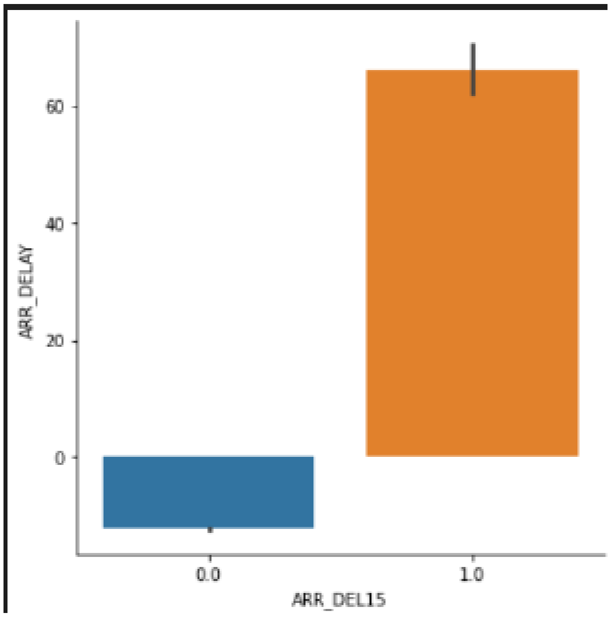
```
sns.distplot(flight_data.MONTH)
```

es on Kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies.

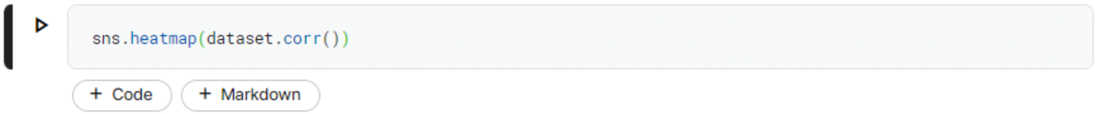


Activity 2.2: Bivariate analysis

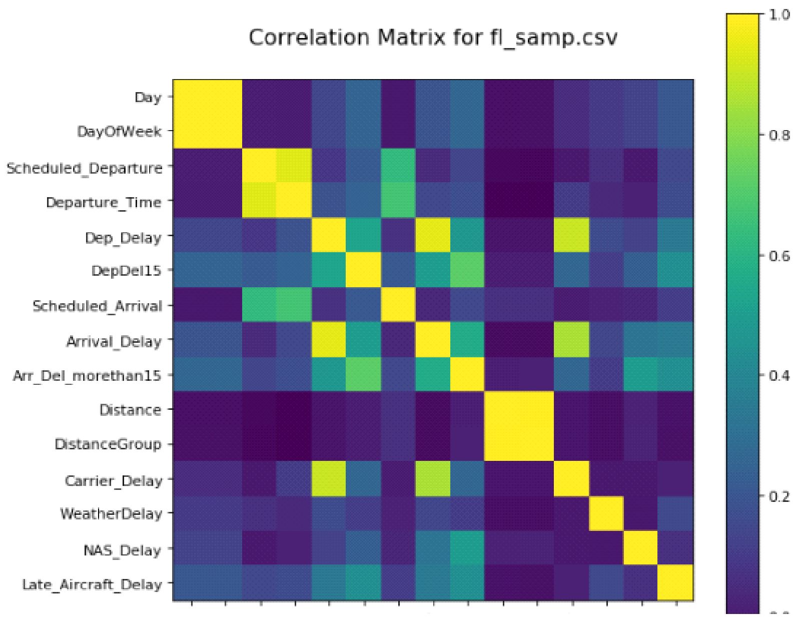




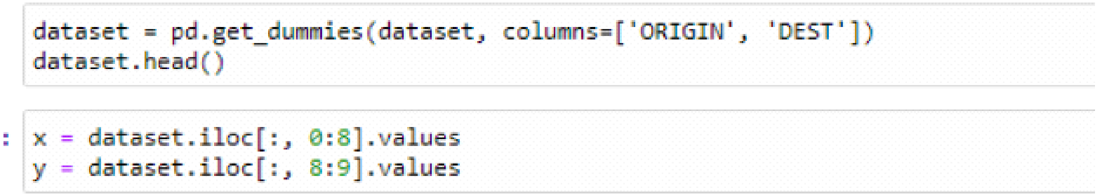
Activity 2.3: Multivariate analysis



```
sns.heatmap(dataset.corr())
```



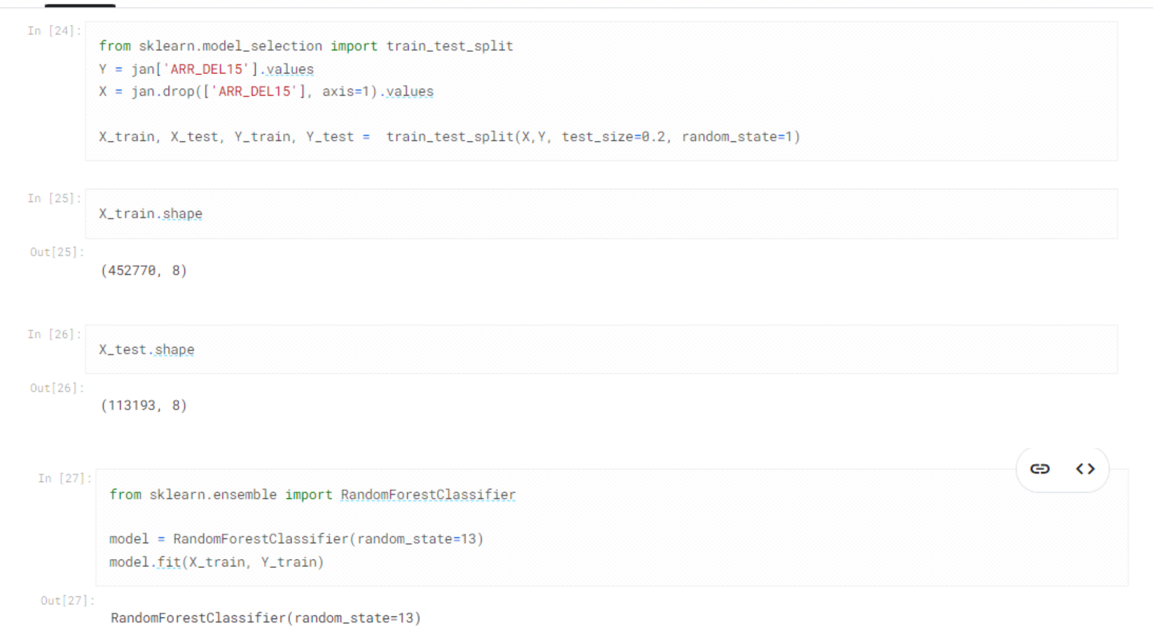
Splitting data into dependent and independent variables



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

```
x = dataset.iloc[:, 0:8].values
y = dataset.iloc[:, 8:9].values
```

Splitting data into train and test



```
In [24]:
from sklearn.model_selection import train_test_split
Y = jan['ARR_DEL15'].values
X = jan.drop(['ARR_DEL15'], axis=1).values

X_train, X_test, Y_train, Y_test =  train_test_split(X,Y, test_size=0.2, random_state=1)
```

```
In [25]:
X_train.shape
```

```
Out[25]:
(452770, 8)
```

```
In [26]:
X_test.shape
```

```
Out[26]:
(113193, 8)
```

```
In [27]:
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(random_state=13)
model.fit(X_train, Y_train)
```

```
Out[27]:
RandomForestClassifier(random_state=13)
```