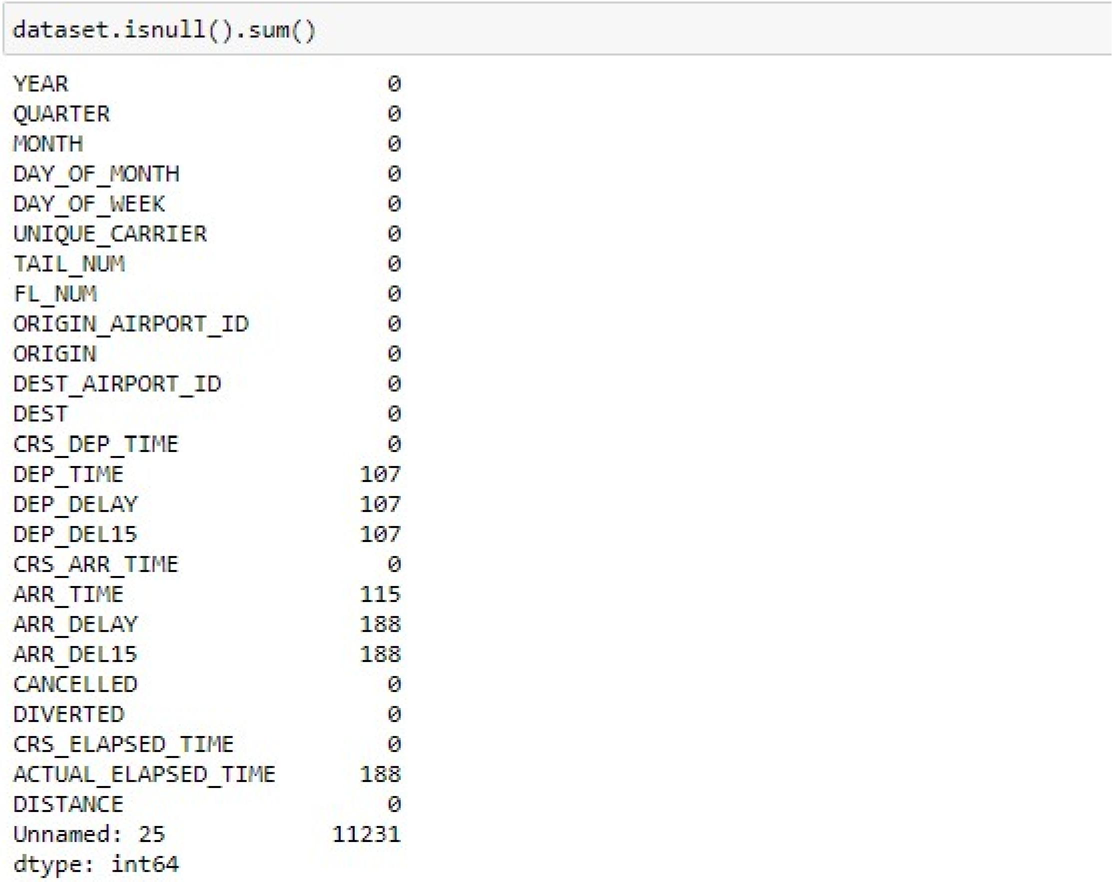
DEVELOPING A FLIGHT DELAY MODEL USING MACHINE LEARNING

TEAM ID: PNT2022TMID27775

Handling Missing Values

- Sometimes you may find some data missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common ideas to handle the problem is to take a mean of all the values for continuous and for categorical we make use of mode values and replace the missing data.



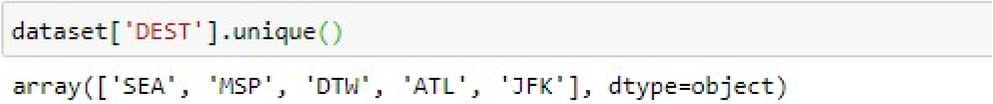
```
dataset.isnull().sum()
```

```
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
```

- Word “True” that the particular column has missing values, we can also see the count of missing values in each column by using isnull().sum function.

Check unique values in dataset

- Often, a DataFrame will contain columns that are having some unique values from which we can find out the unique records which are present in the dataset.



```
dataset['DEST'].unique()
```

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