

Data Collection and Preprocessing Phase

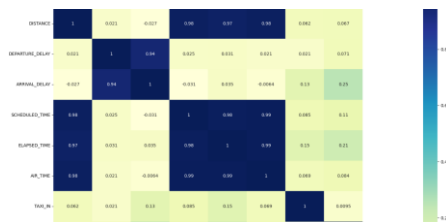
Date	03-10-2024
Team ID	LTVIP2024TMID24897
Project Title	Flight delays prediction using ML
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																																																					
Data Overview	dimension: 8rows X 26 columns Descriptive statistics:																																																																																																																					
	<table><thead><tr><th></th><th>YEAR</th><th>MONTH</th><th>DAY</th><th>DAY_OF_WEEK</th><th>FLIGHT_NUMBER</th><th>SCHEDULED_DEPARTURE</th><th>DEPARTURE_TIME</th><th>DEPARTURE_DELAY</th><th>TAXI_OUT</th><th>WHEELS_OFF</th><th>...</th><th>SCHEDULED_J</th></tr></thead><tbody><tr><td>count</td><td>2332420</td><td>2.332420e+06</td><td>2.332420e+06</td><td>2.332420e+06</td><td>2.332420e+06</td><td>2.332419e+06</td><td>2.281203e+06</td><td>2.281203e+06</td><td>2.279954e+06</td><td>2.279954e+06</td><td>_</td><td>2.332</td></tr><tr><td>mean</td><td>2015.0</td><td>3.001552e+00</td><td>1.530177e+01</td><td>3.920830e+00</td><td>2.212780e+03</td><td>1.327548e+03</td><td>1.335321e+03</td><td>9.564441e+00</td><td>1.610641e+01</td><td>1.358011e+03</td><td>_</td><td>1.489</td></tr><tr><td>std</td><td>0.0</td><td>1.399672e+00</td><td>8.569885e+00</td><td>1.981082e+00</td><td>1.780452e+03</td><td>4.787265e+02</td><td>4.906983e+02</td><td>3.704553e+01</td><td>9.182326e+00</td><td>4.916444e+02</td><td>_</td><td>4.993</td></tr><tr><td>min</td><td>2015.0</td><td>1.000000e+00</td><td>1.000000e+00</td><td>1.000000e+00</td><td>1.000000e+00</td><td>1.000000e+00</td><td>1.000000e+00</td><td>-6.800000e+01</td><td>1.000000e+00</td><td>1.000000e+00</td><td>_</td><td>1.000</td></tr><tr><td>25%</td><td>2015.0</td><td>2.000000e+00</td><td>8.000000e+00</td><td>2.000000e+00</td><td>7.440000e+02</td><td>9.200000e+02</td><td>9.240000e+02</td><td>-5.000000e+00</td><td>1.100000e+01</td><td>9.390000e+02</td><td>_</td><td>1.115</td></tr><tr><td>50%</td><td>2015.0</td><td>3.000000e+00</td><td>1.500000e+01</td><td>4.000000e+00</td><td>1.691000e+03</td><td>1.322000e+03</td><td>1.330000e+03</td><td>-1.000000e+00</td><td>1.400000e+01</td><td>1.343000e+03</td><td>_</td><td>1.522</td></tr><tr><td>75%</td><td>2015.0</td><td>4.000000e+00</td><td>2.300000e+01</td><td>6.000000e+00</td><td>3.395000e+03</td><td>1.727000e+03</td><td>1.737000e+03</td><td>8.000000e+00</td><td>1.900000e+01</td><td>1.750000e+03</td><td>_</td><td>1.918</td></tr><tr><td>max</td><td>2015.0</td><td>5.000000e+00</td><td>3.100000e+01</td><td>7.000000e+00</td><td>9.794000e+03</td><td>2.359000e+03</td><td>2.400000e+03</td><td>1.988000e+03</td><td>2.250000e+02</td><td>2.400000e+03</td><td>_</td><td>2.400</td></tr></tbody></table>		YEAR	MONTH	DAY	DAY_OF_WEEK	FLIGHT_NUMBER	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS_OFF	...	SCHEDULED_J	count	2332420	2.332420e+06	2.332420e+06	2.332420e+06	2.332420e+06	2.332419e+06	2.281203e+06	2.281203e+06	2.279954e+06	2.279954e+06	_	2.332	mean	2015.0	3.001552e+00	1.530177e+01	3.920830e+00	2.212780e+03	1.327548e+03	1.335321e+03	9.564441e+00	1.610641e+01	1.358011e+03	_	1.489	std	0.0	1.399672e+00	8.569885e+00	1.981082e+00	1.780452e+03	4.787265e+02	4.906983e+02	3.704553e+01	9.182326e+00	4.916444e+02	_	4.993	min	2015.0	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	-6.800000e+01	1.000000e+00	1.000000e+00	_	1.000	25%	2015.0	2.000000e+00	8.000000e+00	2.000000e+00	7.440000e+02	9.200000e+02	9.240000e+02	-5.000000e+00	1.100000e+01	9.390000e+02	_	1.115	50%	2015.0	3.000000e+00	1.500000e+01	4.000000e+00	1.691000e+03	1.322000e+03	1.330000e+03	-1.000000e+00	1.400000e+01	1.343000e+03	_	1.522	75%	2015.0	4.000000e+00	2.300000e+01	6.000000e+00	3.395000e+03	1.727000e+03	1.737000e+03	8.000000e+00	1.900000e+01	1.750000e+03	_	1.918	max	2015.0	5.000000e+00	3.100000e+01	7.000000e+00	9.794000e+03	2.359000e+03	2.400000e+03	1.988000e+03	2.250000e+02	2.400000e+03	_	2.400
		YEAR	MONTH	DAY	DAY_OF_WEEK	FLIGHT_NUMBER	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS_OFF	...	SCHEDULED_J																																																																																																									
count	2332420	2.332420e+06	2.332420e+06	2.332420e+06	2.332420e+06	2.332419e+06	2.281203e+06	2.281203e+06	2.279954e+06	2.279954e+06	_	2.332																																																																																																										
mean	2015.0	3.001552e+00	1.530177e+01	3.920830e+00	2.212780e+03	1.327548e+03	1.335321e+03	9.564441e+00	1.610641e+01	1.358011e+03	_	1.489																																																																																																										
std	0.0	1.399672e+00	8.569885e+00	1.981082e+00	1.780452e+03	4.787265e+02	4.906983e+02	3.704553e+01	9.182326e+00	4.916444e+02	_	4.993																																																																																																										
min	2015.0	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	-6.800000e+01	1.000000e+00	1.000000e+00	_	1.000																																																																																																										
25%	2015.0	2.000000e+00	8.000000e+00	2.000000e+00	7.440000e+02	9.200000e+02	9.240000e+02	-5.000000e+00	1.100000e+01	9.390000e+02	_	1.115																																																																																																										
50%	2015.0	3.000000e+00	1.500000e+01	4.000000e+00	1.691000e+03	1.322000e+03	1.330000e+03	-1.000000e+00	1.400000e+01	1.343000e+03	_	1.522																																																																																																										
75%	2015.0	4.000000e+00	2.300000e+01	6.000000e+00	3.395000e+03	1.727000e+03	1.737000e+03	8.000000e+00	1.900000e+01	1.750000e+03	_	1.918																																																																																																										
max	2015.0	5.000000e+00	3.100000e+01	7.000000e+00	9.794000e+03	2.359000e+03	2.400000e+03	1.988000e+03	2.250000e+02	2.400000e+03	_	2.400																																																																																																										
Univariate Analysis																																																																																																																						
Bivariate Analysis																																																																																																																						

Multivariate Analysis



Outliers and Anomalies

There is no need of outliers ,as we using regressions and classifiers

Data Preprocessing Code Screenshots

Loading Data

```
flightsinfo = pd.read_csv("flights.csv")

list(flightsinfo.columns)

['YEAR',
 'MONTH',
 'DAY',
 'DAY_OF_WEEK',
 'AIRLINE',
 'FLIGHT_NUMBER',
 'TAIL_NUMBER',
 'ORIGIN_AIRPORT',
 'DESTINATION_AIRPORT',
 'SCHEDULED_DEPARTURE',
 'DEPARTURE_TIME',
 'DEPARTURE_DELAY',
 'TAXI_OUT',
 'WHEELS_OFF',
 'SCHEDULED_TIME',
 'ELAPSED_TIME',
 'AIR_TIME',
 'DISTANCE',
 'WHEELS_ON',
 'TAXI_IN',
 'SCHEDULED_ARRIVAL',
 'ARRIVAL_TIME',
 'ARRIVAL_DELAY',
 'DIVERTED',
 'CANCELLED']
```

Handling Missing Data

```
airport.isnull().sum()

IATA_CODE    0
AIRPORT      0
CITY          0
STATE        0
COUNTRY      0
LATITUDE     3
LONGITUDE    3
dtype: int64
```

```
Click to add a breakpoint: flightsinfo.isnull().sum()*100/flightsinfo.shape[0]
flightsinfo_NULL
```

Variable	Count
YEAR	0.000000
MONTH	0.000000
DAY	0.000000
DAY_OF_WEEK	0.000000
AIRLINE	0.000000
FLIGHT_NUMBER	0.000000
TAIL_NUMBER	0.395340
ORIGIN_AIRPORT	0.000043
DESTINATION_AIRPORT	0.000043
SCHEDULED_DEPARTURE	0.000043
DEPARTURE_TIME	2.195874

Data Transformation

There is no need of standardization and normalization of our dataset , as we using histogram techniques.

Feature Engineering	<pre> Applying SMOTE to deal with class imbalance (Inport) SMOTE: Any from imblearn.over_sampling import SMOTE import "imblearn.ova /usr/local/lib/python3.6/dist-packages/sklearn/externals/six.py:31: "(https://pypl.org/project/six/).", DeprecationWarning) : sm = SMOTE() X_train, y_train = sm.fit_sample(X_train, y_train) pd.Series(y_train).value_counts() 1 2294725 0 2294725 dtype: int64 </pre>
Save Processed Data	-