Data Collection and Preprocessing Phase

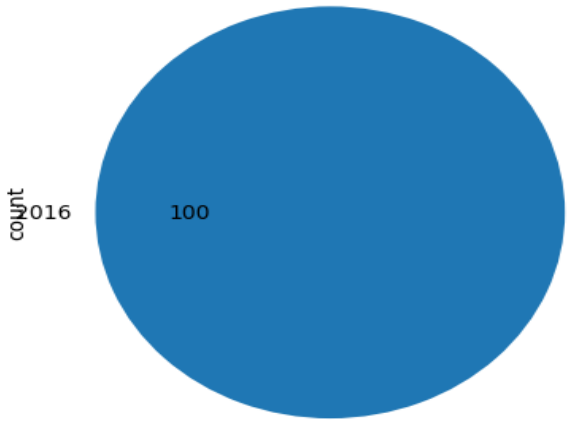
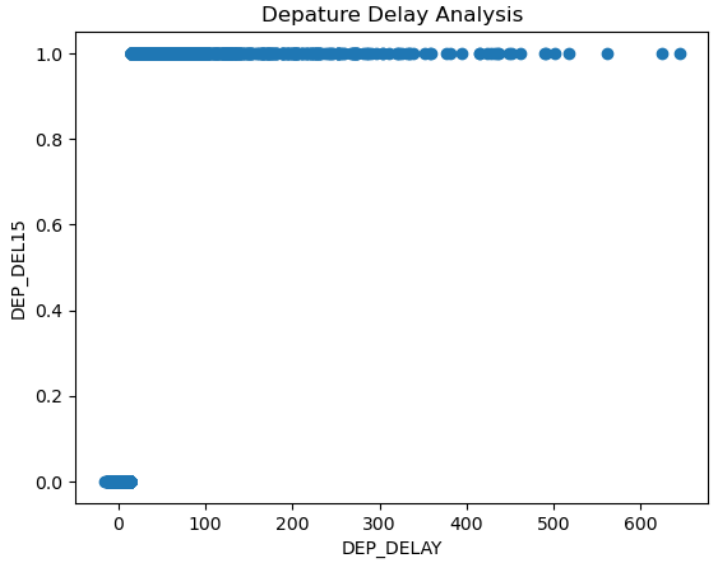
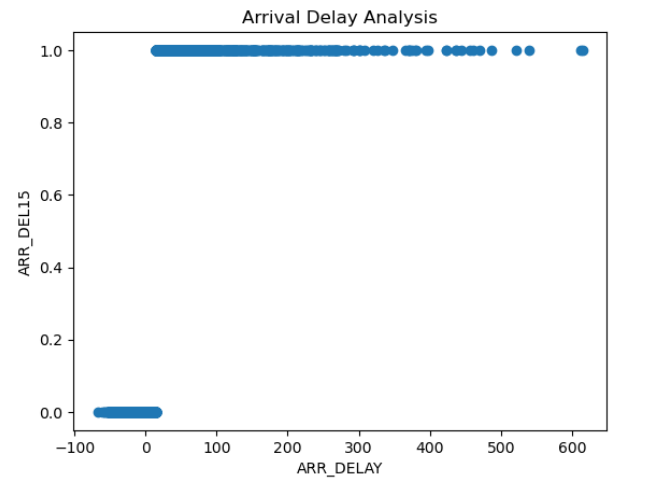
Date	15 March 2024
Team ID	739674
Project Title	Smart Lender- Flight delay Prediction
Maximum Marks	6 Marks

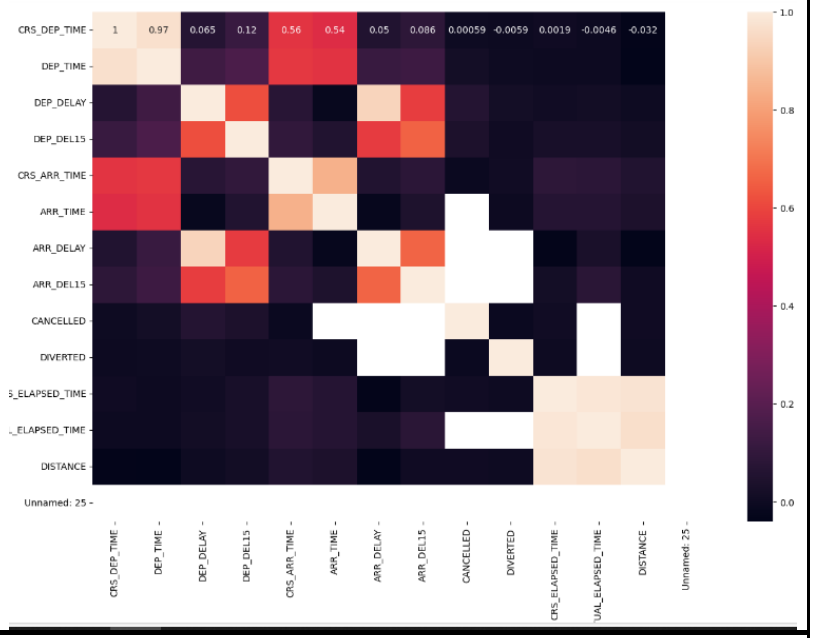
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employedforpreprocessingtaskslikenormalizationandfeatureengineering.Datacleaningwill address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 11231rows×26columns Descriptive statistics:
Univariate Analysis	

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIME
0	2016	1	1	1	5	DL	N836DN	1399	10397	ATL	...	214
1	2016	1	1	1	5	DL	N964DN	1476	11433	DTW	...	143
2	2016	1	1	1	5	DL	N813DN	1597	10397	ATL	...	121
3	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	...	133
4	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	...	60
...	...	...	...	...	...	...	...	...	...	...	...	..
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

	
Bivariate Analysis	 

Heat map Analysis	
Loading Data	


```
1 dataset=pd.read_csv("flightdata.csv")
2 dataset
```

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	...	CRS_ARR_TIM
0	2016	1	1	1	5	DL	N836DN	1399	10397	ATL	...	214
1	2016	1	1	1	5	DL	N964DN	1476	11433	DTW	...	143
2	2016	1	1	1	5	DL	N813DN	1597	10397	ATL	...	121
3	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	...	133
4	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	...	60
...	...	...	...	...	...	...	...	...	...	...	...	
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

Handling Missing Data & Replacing null Values

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

	FL_NUM	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	DEP_DEL15	CRS_ARR_TIME	ARR_DEL15
count	11231.000000	11231.000000	11231.000000	11231.000000	11124.000000	11231.000000	11043.000000
mean	1334.325617	6.628973	15.790758	3.960199	0.142844	15.067314	0.124513
std	811.875227	3.354678	8.782056	1.995257	0.349930	5.023534	0.330181
min	7.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000
25%	624.000000	4.000000	8.000000	2.000000	0.000000	11.000000	0.000000
50%	1267.000000	7.000000	16.000000	4.000000	0.000000	15.000000	0.000000
75%	2032.000000	9.000000	23.000000	6.000000	0.000000	19.000000	0.000000
max	2853.000000	12.000000	31.000000	7.000000	1.000000	23.000000	1.000000

```
dataset=dataset.fillna({'DEP_DEL15':dataset['DEP_DEL15'].mode()[0],
                        'ARR_DEL15':dataset["ARR_DEL15"].mode()[0]})
```

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

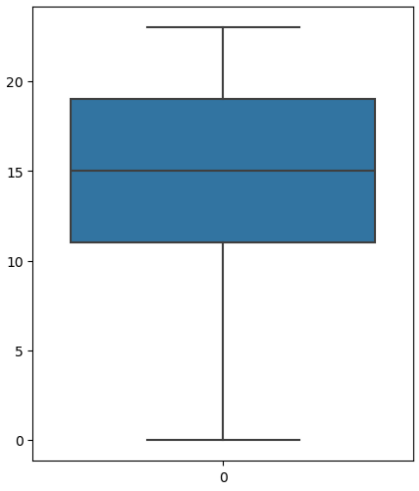
```
FL_NUM          0
MONTH           0
DAY_OF_MONTH    0
DAY_OF_WEEK     0
ORIGIN          0
DEST            0
DEP_DEL15       0
CRS_ARR_TIME    0
ARR_DEL15       0
dtype: int64
```

```
dataset["ARR_DEL15"].value_counts()
```

```
ARR_DEL15
0.0    9856
1.0    1375
Name: count, dtype: int64
```

Handling outliers

```
fig, ax=plt.subplots(figsize=(5,6))
sb.boxplot(data=dataset["CRS_ARR_TIME"])
plt.show()
```



Handling Categorical values

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

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

```
le=LabelEncoder()
dataset["DEST"]=le.fit_transform(dataset["DEST"])
dataset["ORIGIN"]=le.fit_transform(dataset["ORIGIN"])
dataset.head()
```

	FL_NUM	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	ORIGIN	DEST	DEP_DEL15	CRS_ARR_TIME	ARR_DEL15
0	1399	1	1	5	0	4	0.0	21	0.0
1	1476	1	1	5	1	3	0.0	14	0.0
2	1597	1	1	5	0	4	0.0	12	0.0
3	1768	1	1	5	4	3	0.0	13	0.0
4	1823	1	1	5	4	1	0.0	6	0.0

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11231 entries, 0 to 11230
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   FL_NUM         11231 non-null  int64
 1   MONTH          11231 non-null  int64
 2   DAY_OF_MONTH   11231 non-null  int64
 3   DAY_OF_WEEK    11231 non-null  int64
 4   ORIGIN         11231 non-null  int32
 5   DEST           11231 non-null  int32
 6   DEP_DEL15      11231 non-null  float64
 7   CRS_ARR_TIME   11231 non-null  int64
 8   ARR_DEL15      11231 non-null  float64
dtypes: float64(2), int32(2), int64(5)
memory usage: 702.1 KB
```

Splitting data into independent and dependent Variables

Splitting Dataset into Independent and Dependent Variables ¶

```
X=dataset.drop(columns=["ARR_DEL15"]) #independent variables
Y=dataset[["ARR_DEL15"]]#dependent variables
#converting to 1-D array to train model
X=X.values
Y=Y.values
```