FLIGHT DELAY CLASSIFICATION USING MACHINE LEARNING

EMA1640 – Machine Learning

Final project

GROUP MEMBER

Raseel Alghamdi - [redacted]

Ghada Alsulami - [redacted]

PROJECT AIM

This project aims to classify flight statuses—On Time, Delayed, or Canceled—using historical flight, weather, and congestion data. By applying machine learning techniques, the objective is to develop a predictive model that supports proactive decision-making in airline operations.

PROJECT GOALS

- Predict flight status: On Time, Delayed, or Canceled
- Improve decision-making using machine learning
- Analyze and prepare real-world flight data

- Compare classification models to find the most accurate
- Optimize model performance through hyperparameter tuning

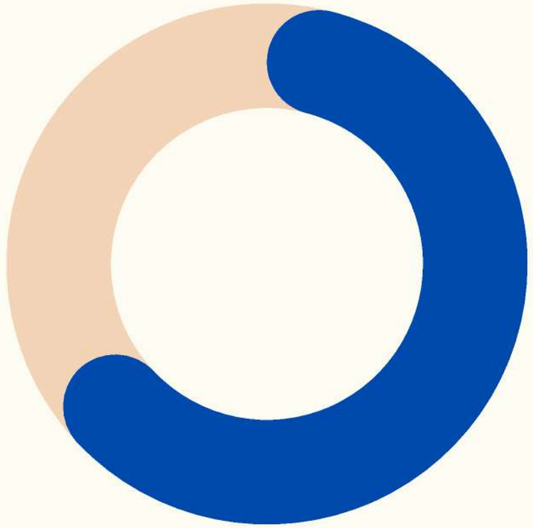
DATASET OVERVIEW

Source: Historical U.S. flight data

- Size: ~30,000 rows after cleaning

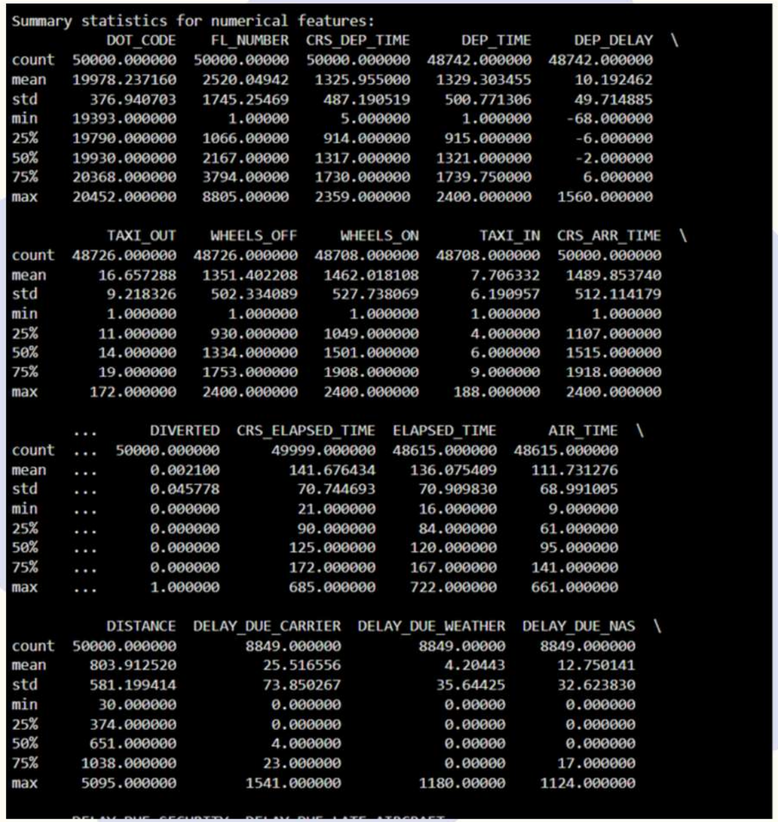
- Features: Flight time, origin, destination, weather, airline, delay reason, etc.

DATA PREPARATION



- Missing Values: Median/mode imputation
- Categorical Encoding: One-hot encoding
- Feature Scaling: Standardization (StandardScaler)
- Class Imbalance: Solved using SMOTE (oversampling minority classes)

DATA SUMMARY



```
Summary statistics for numerical features:
          DOT_CODE     FL_NUMBER  CRS_DEP_TIME      DEP_TIME     DEP_DELAY  \
count  50000.000000  50000.00000  50000.000000  48742.000000  48742.000000
mean   19978.237160   2520.04942   1325.955000   1329.303455     10.192462
std      376.940703   1745.25469    487.190519    500.771306     49.714885
min    19393.000000      1.00000      5.000000      1.000000    -68.000000
25%    19790.000000   1066.00000    914.000000    915.000000     -6.000000
50%    19930.000000   2167.00000   1317.000000   1321.000000     -2.000000
75%    20368.000000   3794.00000   1730.000000   1739.750000      6.000000
max    20452.000000   8805.00000   2359.000000   2400.000000   1560.000000

           TAXI_OUT    WHEELS_OFF     WHEELS_ON       TAXI_IN  CRS_ARR_TIME  \
count  48726.000000  48726.000000  48708.000000  48708.000000  50000.000000
mean      16.657288   1351.402208   1462.018108      7.706332   1489.853740
std        9.218326    502.334089    527.738069      6.190957    512.114179
min        1.000000      1.000000      1.000000      1.000000      1.000000
25%       11.000000    930.000000   1049.000000      4.000000   1107.000000
50%       14.000000   1334.000000   1501.000000      6.000000   1515.000000
75%       19.000000   1753.000000   1908.000000      9.000000   1918.000000
max      172.000000   2400.000000   2400.000000    188.000000   2400.000000

       ...      DIVERTED  CRS_ELAPSED_TIME  ELAPSED_TIME      AIR_TIME  \
count  ...  50000.000000      49999.000000  48615.000000  48615.000000
mean   ...      0.002100        141.676434    136.075409    111.731276
std    ...      0.045778         70.744693     70.909830     68.991005
min    ...      0.000000         21.000000     16.000000      9.000000
25%    ...      0.000000         90.000000     84.000000     61.000000
50%    ...      0.000000        125.000000    120.000000     95.000000
75%    ...      0.000000        172.000000    167.000000    141.000000
max    ...      1.000000        685.000000    722.000000    661.000000

           DISTANCE  DELAY_DUE_CARRIER  DELAY_DUE_WEATHER  DELAY_DUE_NAS  \
count  50000.000000        8849.000000        8849.00000    8849.000000
mean     803.912520          25.516556           4.20443      12.750141
std      581.199414          73.850267          35.64425      32.623830
min       30.000000           0.000000           0.00000       0.000000
25%      374.000000           0.000000           0.00000       0.000000
50%      651.000000           4.000000           0.00000       0.000000
75%     1038.000000          23.000000           0.00000      17.000000
max     5095.000000        1541.000000        1180.00000    1124.000000
```

here is summary for some data

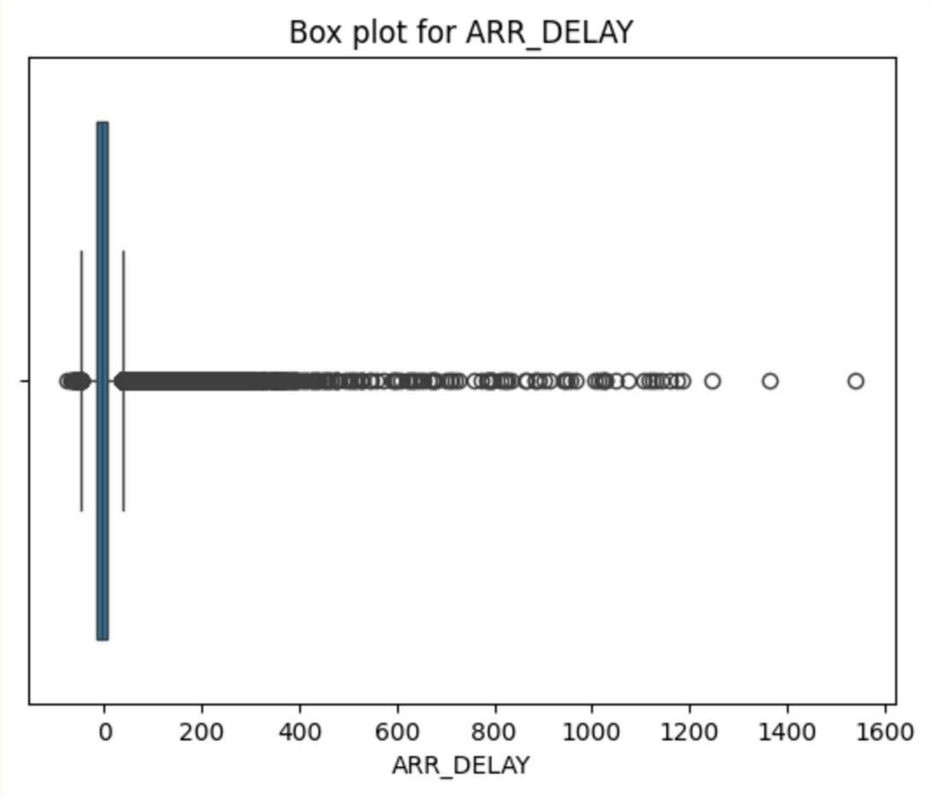
DATA PREPARATION

```
 Missing Values BEFORE Handling (First 50,000 Rows):
DEP_TIME                  2.516
DEP_DELAY                 2.516
TAXI_OUT                  2.548
WHEELS_OFF                2.548
WHEELS_ON                 2.584
TAXI_IN                   2.584
ARR_TIME                  2.584
ARR_DELAY                 2.770
CANCELLATION_CODE        97.440
CRS_ELAPSED_TIME          0.002
ELAPSED_TIME              2.770
AIR_TIME                  2.770
DELAY_DUE_CARRIER        82.302
DELAY_DUE_WEATHER        82.302
DELAY_DUE_NAS            82.302
DELAY_DUE_SECURITY       82.302
DELAY_DUE_LATE_AIRCRAFT  82.302
dtype: float64

 Missing Values AFTER Handling (First 50,000 Rows):
Series([], dtype: float64)
```

We handled data (Missing values)

OUTLIERS DETECTION



here is example for outliers detection in the ARR_DELAY feature

REMOVING OUTLIERS (IQR METHOD)

```
For DOT_CODE, lower bound: 18923.0, upper bound: 21235.0
For FL_NUMBER, lower bound: -3026.0, upper bound: 7886.0
For CRS_DEP_TIME, lower bound: -310.0, upper bound: 2954.0
For DEP_TIME, lower bound: -294.5, upper bound: 2949.5
For DEP_DELAY, lower bound: -24.5, upper bound: 27.5
For TAXI_OUT, lower bound: -1.0, upper bound: 31.0
For WHEELS_OFF, lower bound: -276.5, upper bound: 2959.5
For WHEELS_ON, lower bound: -211.5, upper bound: 3168.5
For TAXI_IN, lower bound: -3.5, upper bound: 16.5
For CRS_ARR_TIME, lower bound: -109.5, upper bound: 3134.5
For ARR_TIME, lower bound: -210.0, upper bound: 3174.0
For ARR_DELAY, lower bound: -46.5, upper bound: 37.5
For CANCELLED, lower bound: 0.0, upper bound: 0.0
For DIVERTED, lower bound: 0.0, upper bound: 0.0
For CRS_ELAPSED_TIME, lower bound: -33.0, upper bound: 295.0
For ELAPSED_TIME, lower bound: -35.0, upper bound: 285.0
For AIR_TIME, lower bound: -55.0, upper bound: 257.0
For DISTANCE, lower bound: -622.0, upper bound: 2034.0
For DELAY_DUE_CARRIER, lower bound: 25.51655554299921, upper bound: 25.51655554299921
For DELAY_DUE_WEATHER, lower bound: 4.204429879082382, upper bound: 4.204429879082382
For DELAY_DUE_NAS, lower bound: 12.75014125889931, upper bound: 12.75014125889931
For DELAY_DUE_SECURITY, lower bound: 0.16928466493389083, upper bound: 0.16928466493389083
For DELAY_DUE_LATE_AIRCRAFT, lower bound: 25.34195954345124, upper bound: 25.34195954345124
```

We Applied the function to each numerical column

DISPLAY DATA AFTER REMOVING OUTLIERS

Data after removing outliers:

	DOT_CODE	FL_NUMBER	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	\
count	33503.000000	33503.000000	33503.000000	33503.000000	33503.000000	
mean	19971.055995	2619.639644	1292.033848	1285.457750	-1.990240	
std	384.465793	1743.149352	481.494367	485.229784	7.070371	
min	19393.000000	1.000000	5.000000	1.000000	-24.000000	
25%	19790.000000	1164.500000	900.000000	855.000000	-6.000000	
50%	19930.000000	2271.000000	1247.000000	1245.000000	-3.000000	
75%	20366.000000	3950.000000	1705.000000	1702.000000	0.000000	
max	20452.000000	7434.000000	2359.000000	2359.000000	27.000000	

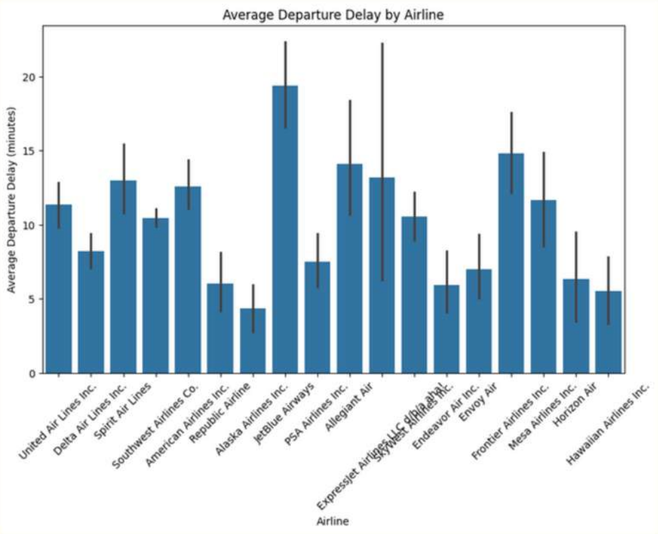
	TAXI_OUT	WHEELS_OFF	WHEELS_ON	TAXI_IN	CRS_ARR_TIME	\
count	33503.000000	33503.000000	33503.000000	33503.000000	33503.000000	
mean	14.725398	1307.803958	1439.930872	6.464914	1459.832881	
std	5.320720	485.207402	495.657476	3.192716	498.722858	
min	3.000000	1.000000	1.000000	1.000000	1.000000	
25%	11.000000	909.000000	1039.000000	4.000000	1055.000000	
50%	14.000000	1258.000000	1428.000000	6.000000	1441.000000	
75%	18.000000	1716.000000	1835.000000	8.000000	1850.000000	
max	31.000000	2400.000000	2400.000000	16.000000	2400.000000	

	...	DIVERTED	CRS_ELAPSED_TIME	ELAPSED_TIME	AIR_TIME	\
count	...	33503.0	33503.000000	33503.000000	33503.000000	
mean	...	0.0	125.324508	116.933021	95.742710	
std	...	0.0	48.912080	48.210545	47.405428	
min	...	0.0	21.000000	16.000000	9.000000	
25%	...	0.0	86.000000	78.000000	58.000000	
50%	...	0.0	116.000000	109.000000	87.000000	
75%	...	0.0	158.000000	149.000000	127.000000	
max	...	0.0	287.000000	264.000000	235.000000	

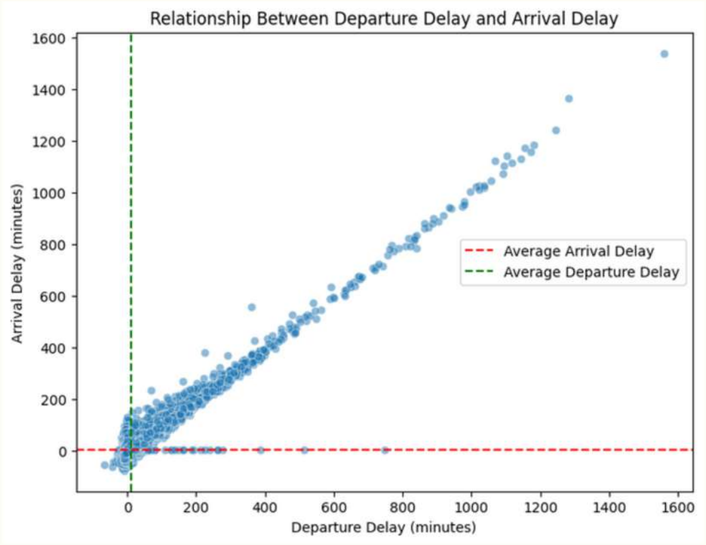
	DISTANCE	DELAY_DUE_CARRIER	DELAY_DUE_WEATHER	DELAY_DUE_NAS	\
count	33503.000000	3.350300e+04	3.350300e+04	3.350300e+04	
mean	674.213444	2.551656e+01	4.204430e+00	1.275014e+01	
std	397.714809	2.230782e-11	1.169748e-12	6.979410e-12	
min	30.000000	2.551656e+01	4.204430e+00	1.275014e+01	
25%	351.000000	2.551656e+01	4.204430e+00	1.275014e+01	
50%	599.000000	2.551656e+01	4.204430e+00	1.275014e+01	
75%	937.000000	2.551656e+01	4.204430e+00	1.275014e+01	
max	1829.000000	2.551656e+01	4.204430e+00	1.275014e+01	

Some data after removing outliers

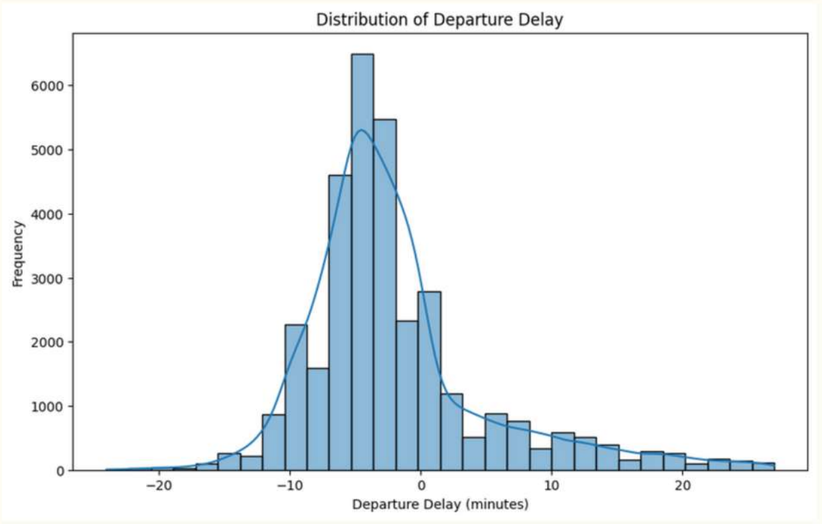
DATA VISUALIZATION



Average Departure Delay by Airline

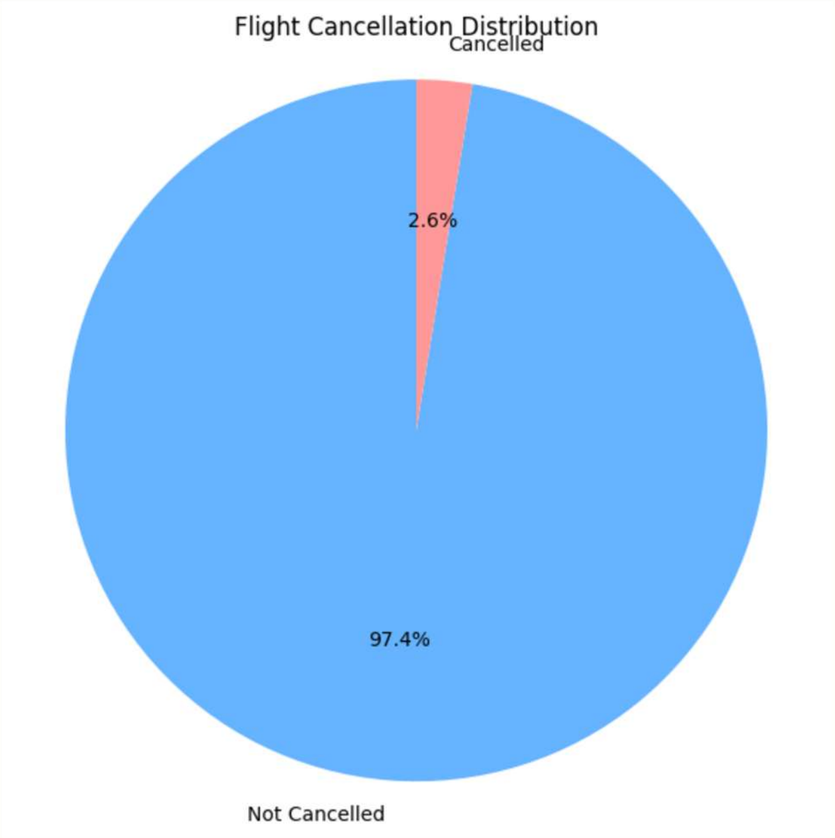


Relationship Between Departure Delay and Arrival Delay



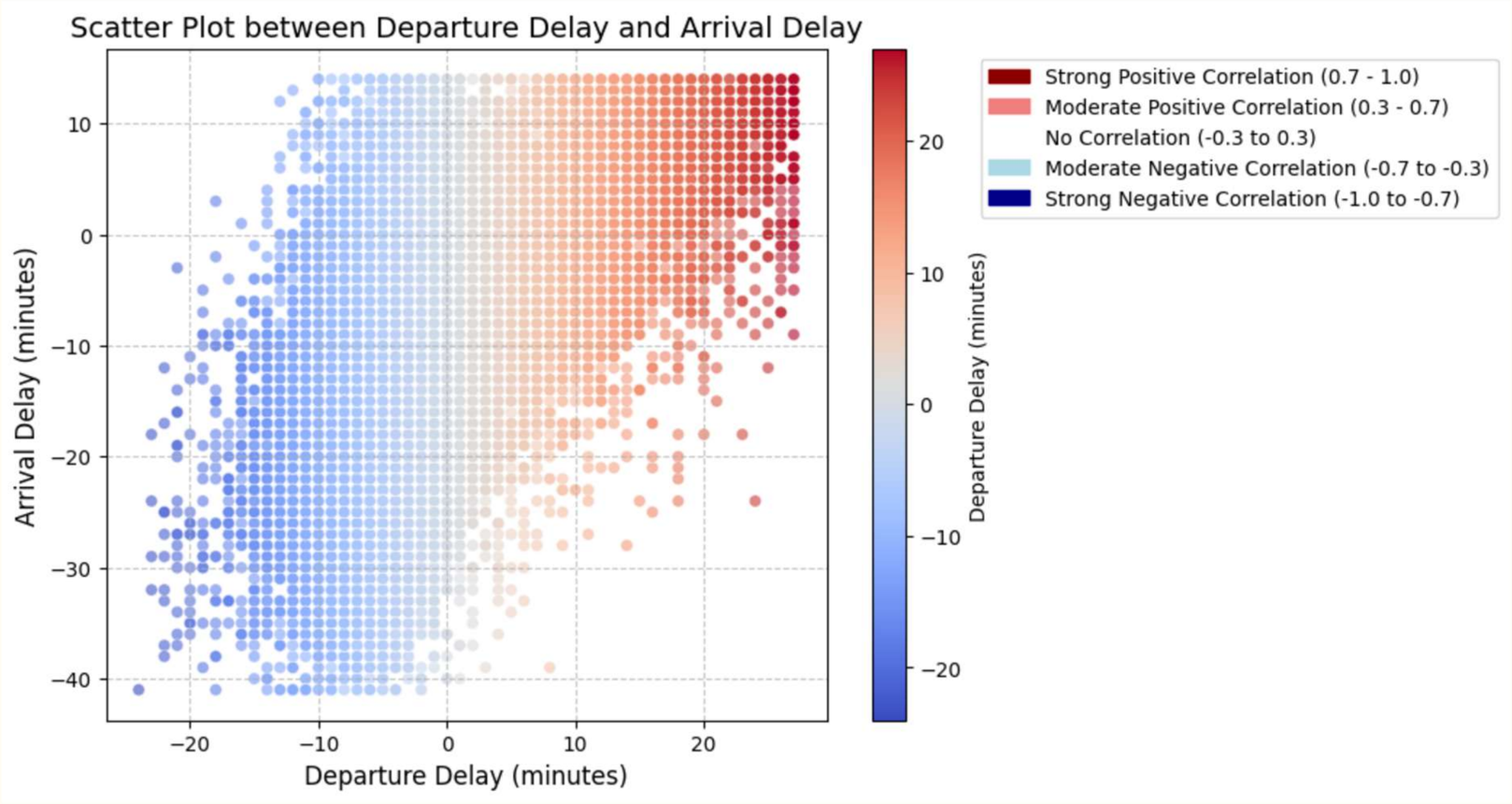
Distribution of Departure Delay for Flights

DATA VISUALIZATION



Flight Cancellation Distribution

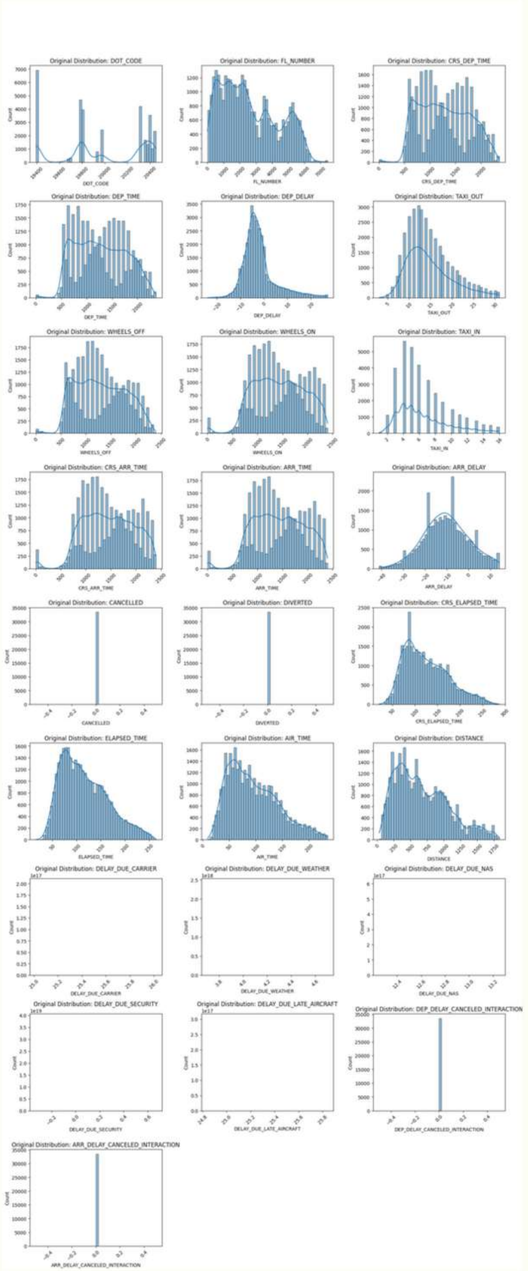
DATA VISUALIZATION



Scatter Plot between DEP_DELAY and ARR_DELAY

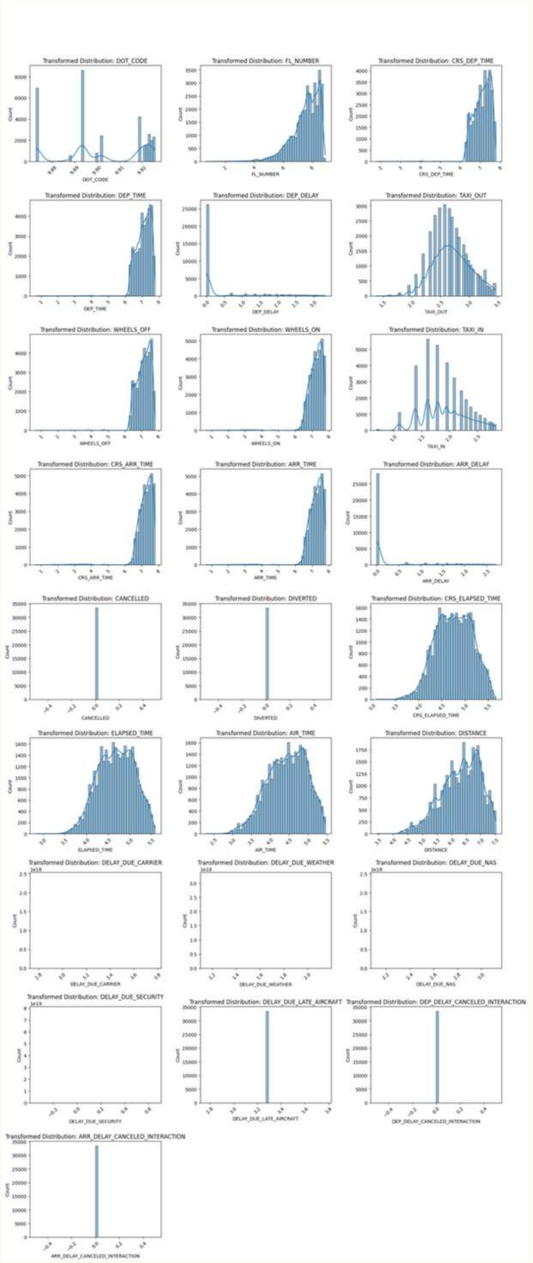
FEATURES DISTRIBUTION

Feature Distributions Before Transformation

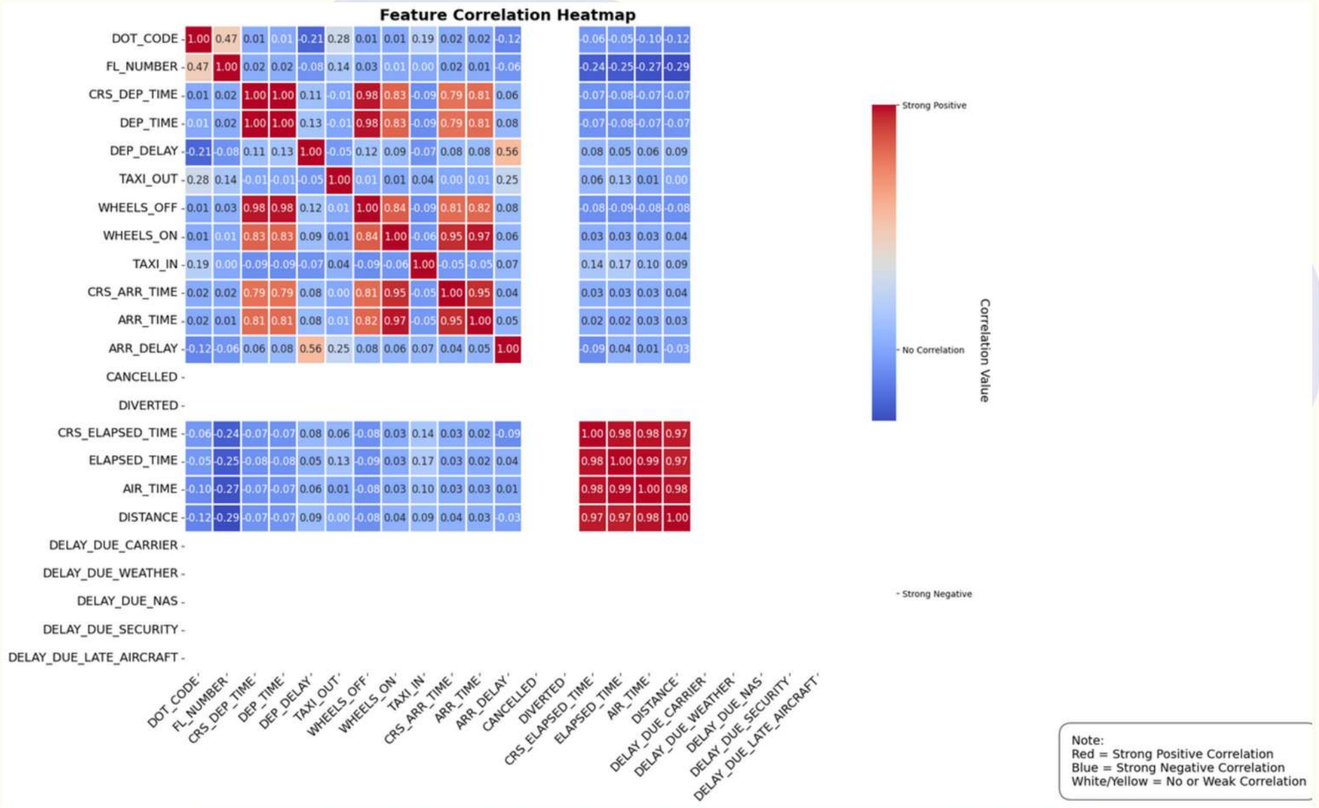


FEATURES DISTRIBUTION

Feature Distributions after Transformation

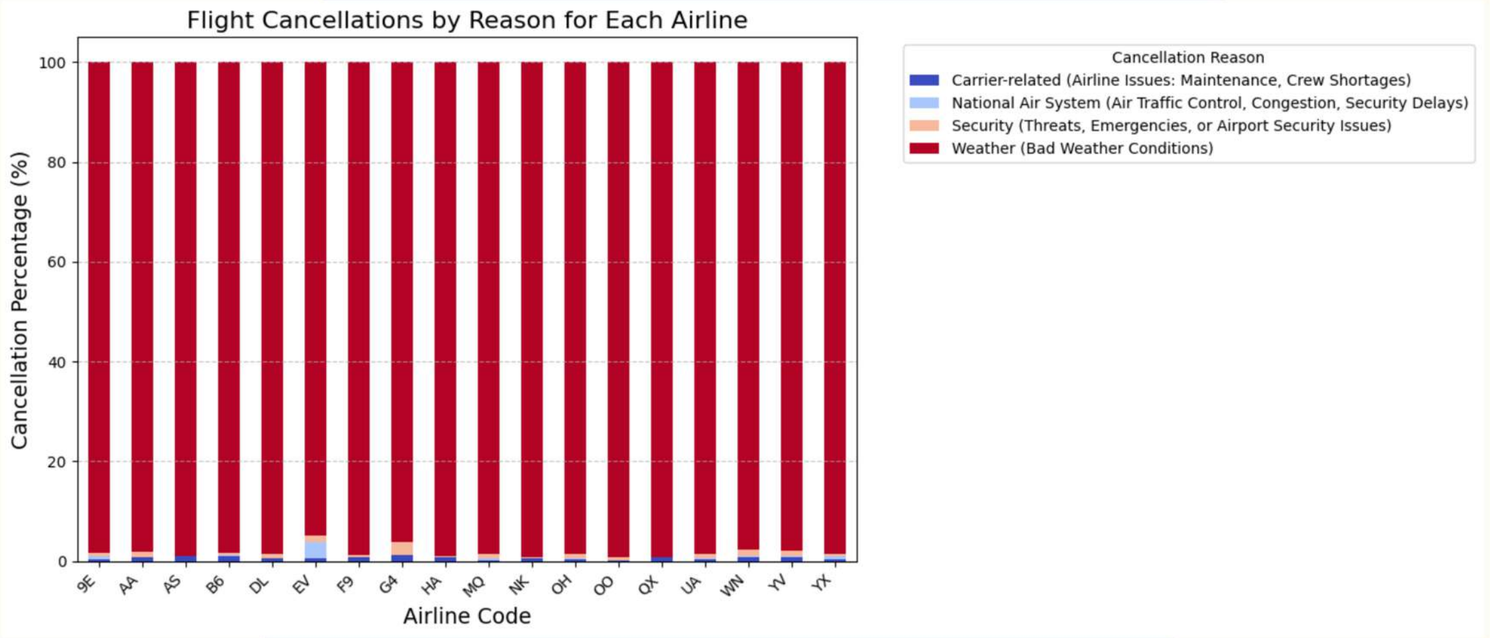


CHARACTERISTICS OF THE DATA



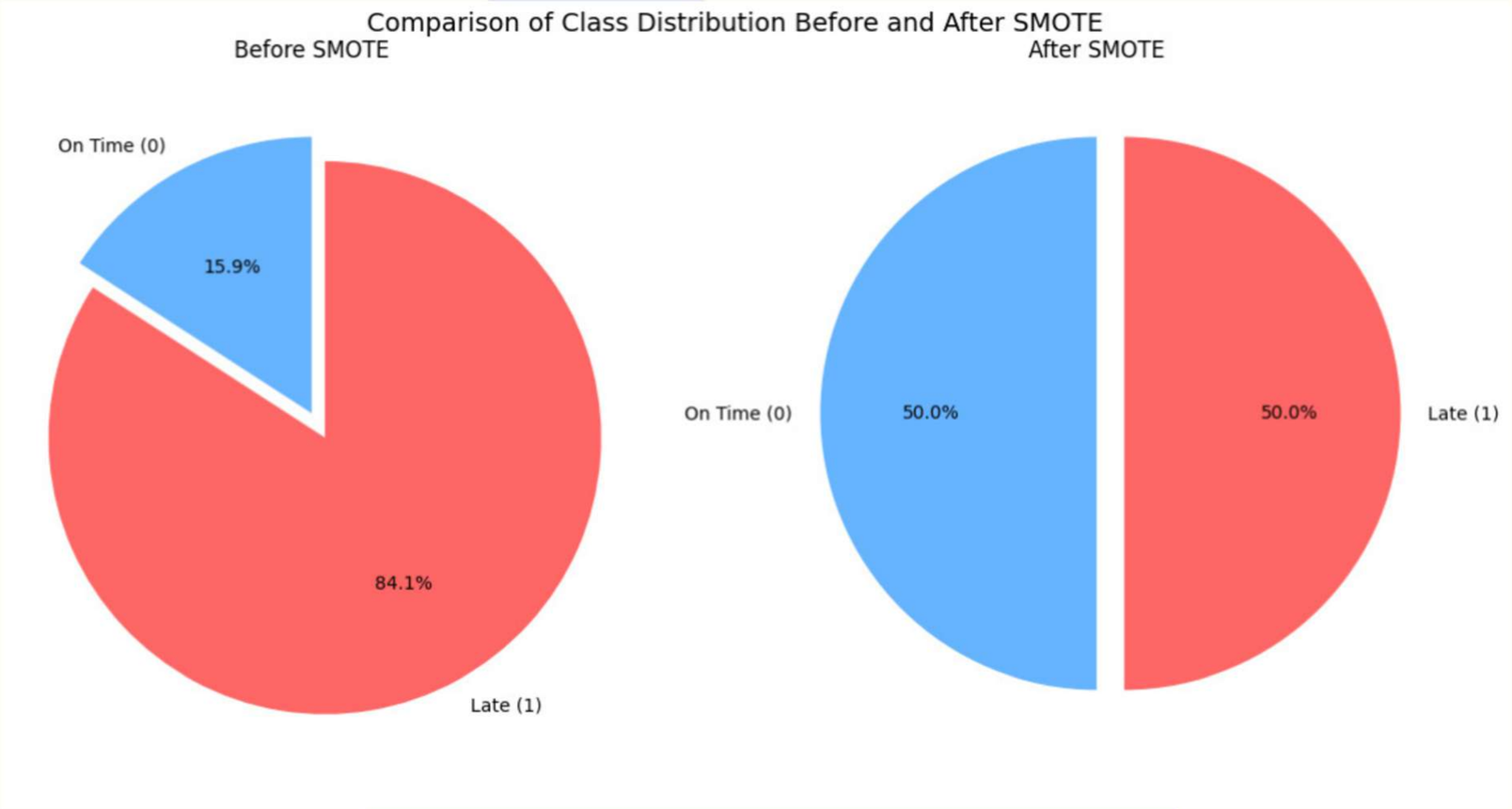
The data was explored and manipulated using Python in a Google Colab notebook, leveraging libraries such as pandas, matplotlib, and seaborn for preliminary analysis and visualization.

FEATURE ENGINEERING



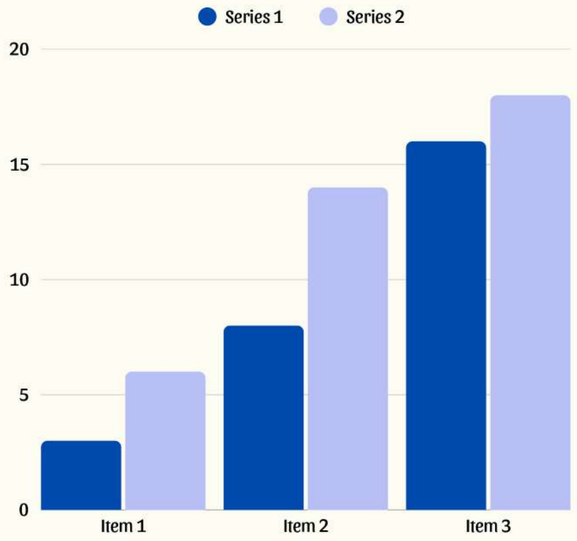
Display cancellation statistics (by bad weather)

HANDLING IMBALANCED DATA



Synthetic Minority Oversampling Technique (SMOTE)

MODELS APPLIED



- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest Classifier

We compared all models based on accuracy, precision, recall, and confusion matrix.

MODEL PERFORMANCE

- Logistic Regression: Accuracy ~98%
- SVM: Accuracy ~93%
- Random Forest: Accuracy ~100%

Random Forest had the best overall performance on all metrics.

HYPERPARAMETER TUNING

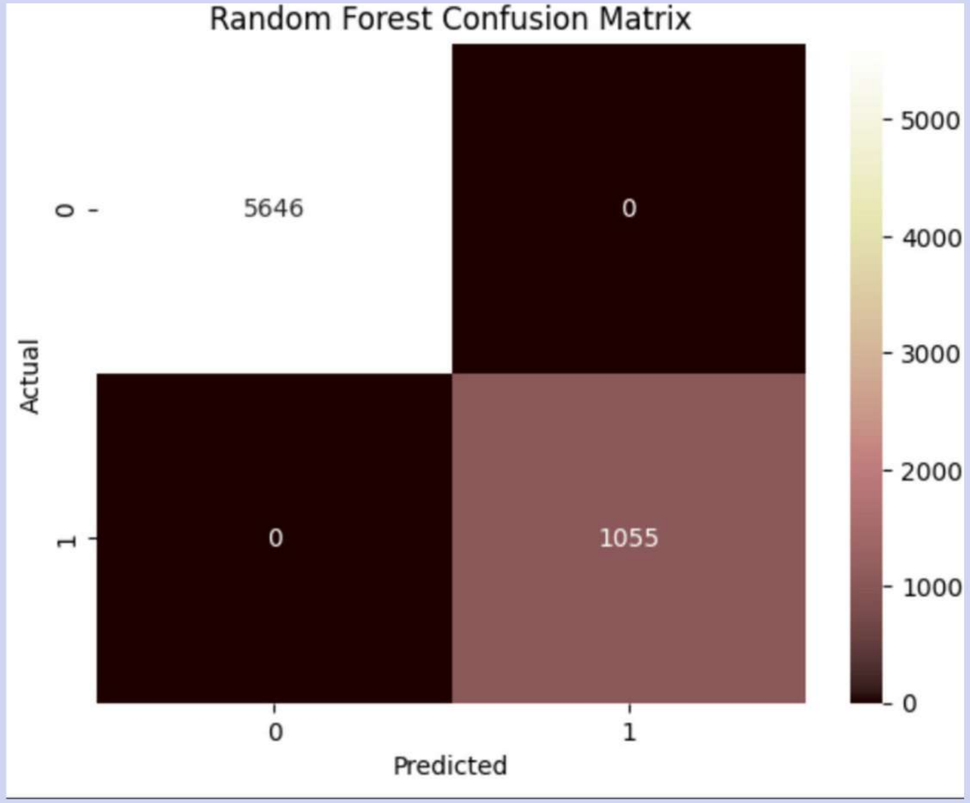
Tool: GridSearchCV

Model Tuned:
Random Forest

Parameters:

- n_estimators
- max_depth
- min_samples_split
- min_samples_leaf

Improved model robustness and accuracy



Random Forest gave the best results

CONCLUSION

- Successfully built a predictive model for flight status
- Random Forest gave the best results
- Data preprocessing and tuning were crucial
- Project demonstrates ML's value in airline analytics

TOOLS & RESOURCES

- Platform: Google Colab
- Libraries: pandas, numpy, seaborn, scikit-learn, imbalanced-learn, matplotlib
- ML Models: Logistic Regression, SVM, Random Forest
- Techniques: SMOTE, GridSearchCV
- https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023
- Cleaned_data.csv

THANK YOU

Any Question?