DECEMBER 12, 2024

PROJECT REPORT

MACHINE LEARNING

ZAIN UL WAHAB (22I-0491)
FAST NUCES
Islamabad

Table of Contents

Problem Statement

Flight departure delays are a critical challenge in the aviation industry. Such delays affect passenger satisfaction, airline operations, and overall efficiency. You are provided with raw Excel files (test, train, and weather data) and are tasked with calculating departure delays. Using these datasets, you will analyze delay patterns and build predictive models to identify key factors contributing to delays.

Phase 1: Data Preprocessing and Feature Engineering

The below steps were taken in order:

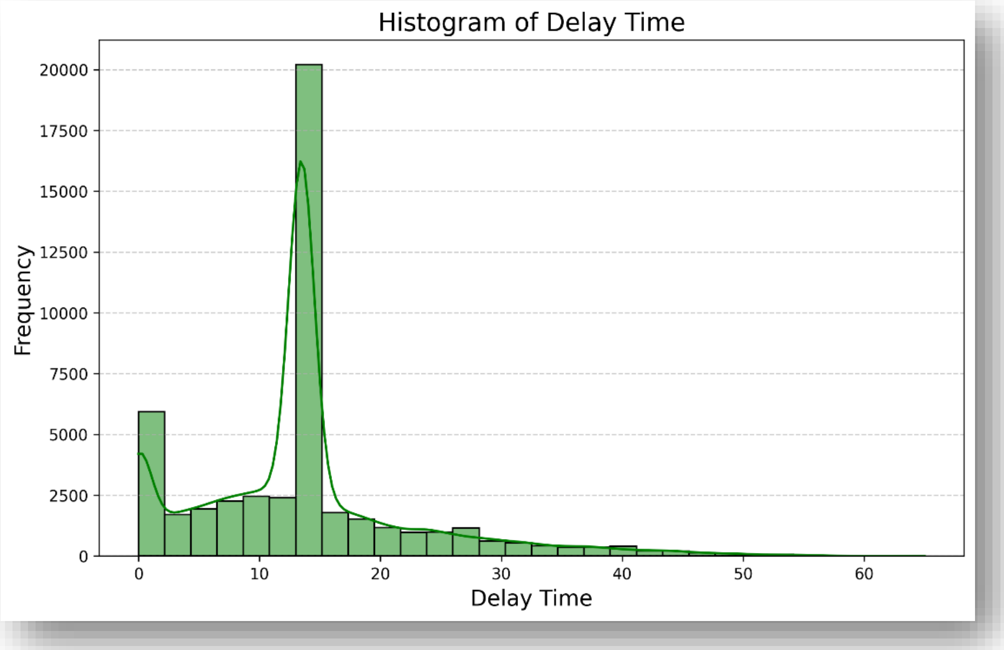
1. In Phase 1 first of all I extracted data from the docx files stored in the "Train" Folder to make my test.csv.
2. After extracting all the data from the docx files I stored it in the "output.csv". This was done in the "reading_data_from_docx.ipynb" jupyter notebook file.
3. Then I extracted the weather data, properly structured it and stored it in the "Weather_formatted" folder.
4. Then I combined all the weather data into a single csv called "combined_weather.csv". The above two steps were done in the "Formatting_weather.ipynb" jupyter notebook file.
5. Then I joined the weather data and the flight data on the common columns which was "Month" and "Day" and stored it in the "test.csv". This was done in the "joining_weather_and_flight.ipynb".
6. Then I made the test.csv in the same method as before and stored it in the "test.csv". This was done in the "combined_weather.csv".
7. Then I revisited the "joining_weather_and_flight.ipynb" and did the feature enginerring.
8. The feature engineering included the below new features:
    a. Calculated delay time
    b. Found day of the week
    c. Found hour of the day
    d. Filling any nan values in the "delay_time" column.
9. That's it for Phase 1

Phase 2: Exploratory Data Analysis (EDA)

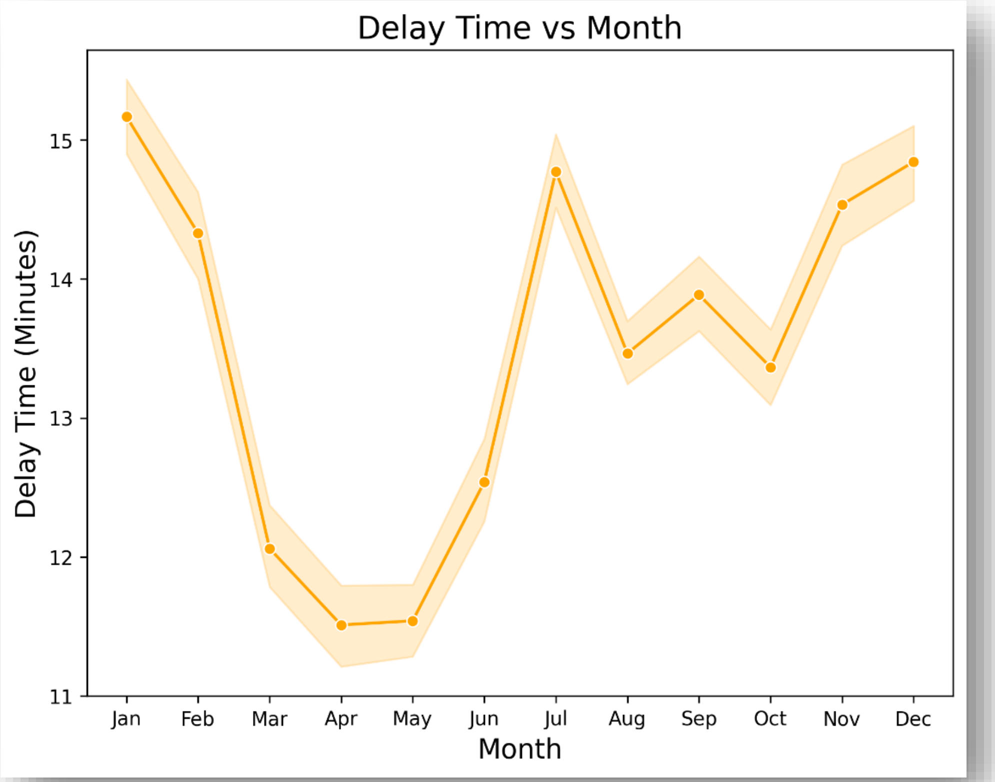
In this phase, the primary objective is to thoroughly explore and understand the dataset prepared in Phase 1. Exploratory Data Analysis (EDA) serves as a crucial step in the data pipeline, as it helps uncover patterns, relationships, and potential anomalies within the data. By leveraging various statistical and visualization techniques, we aim to gain valuable insights that will guide the subsequent stages of model building and optimization.

This phase involves analyzing the distribution of key features, identifying correlations, detecting outliers, and ensuring data quality. Additionally, EDA will help verify the effectiveness of the feature engineering performed in Phase 1 and highlight areas for further refinement, if necessary. Ultimately, this process lays the foundation for developing a robust and accurate predictive model.

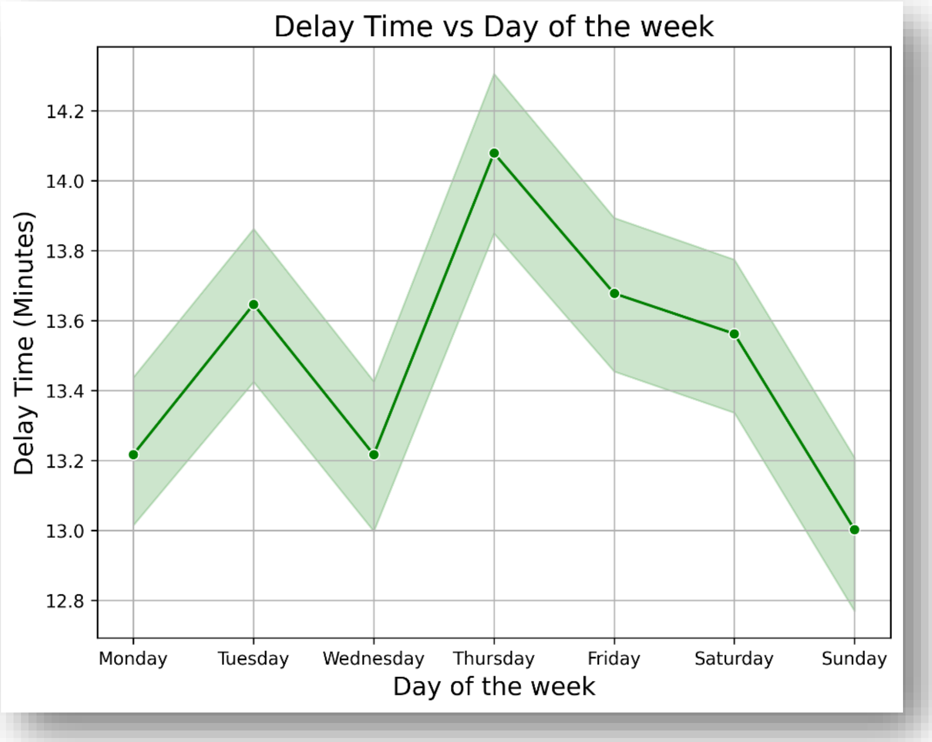
Histogram of delay time



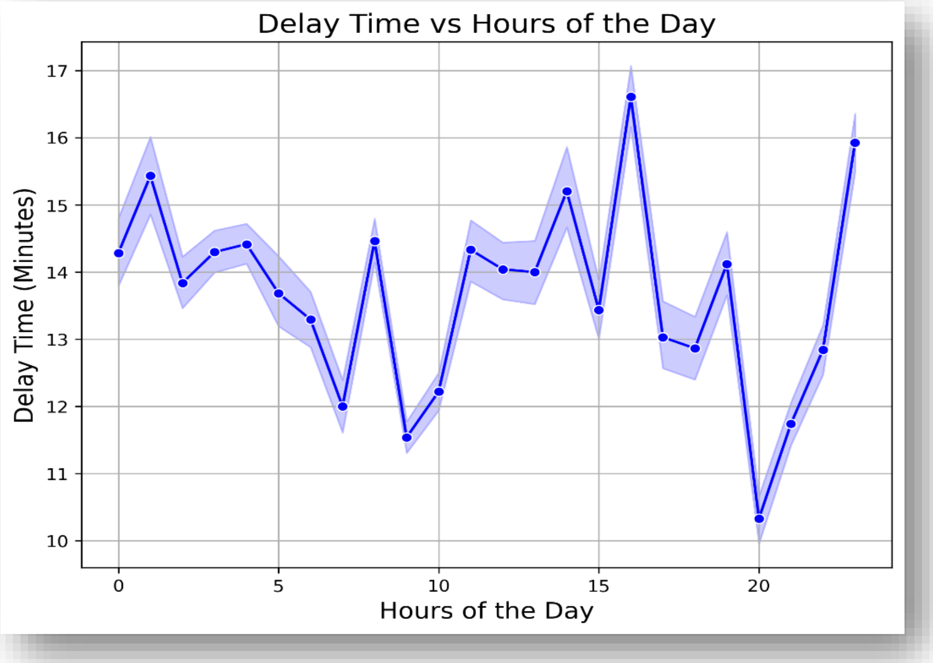
Delay Time vs Month



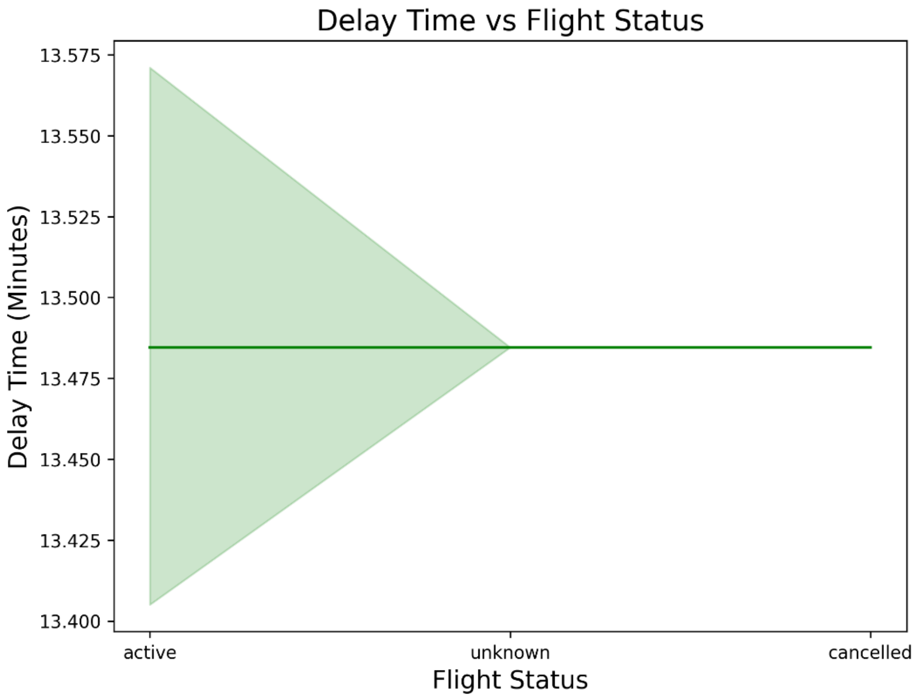
Delay Time vs Day of the week



Delay time vs Hours of the day

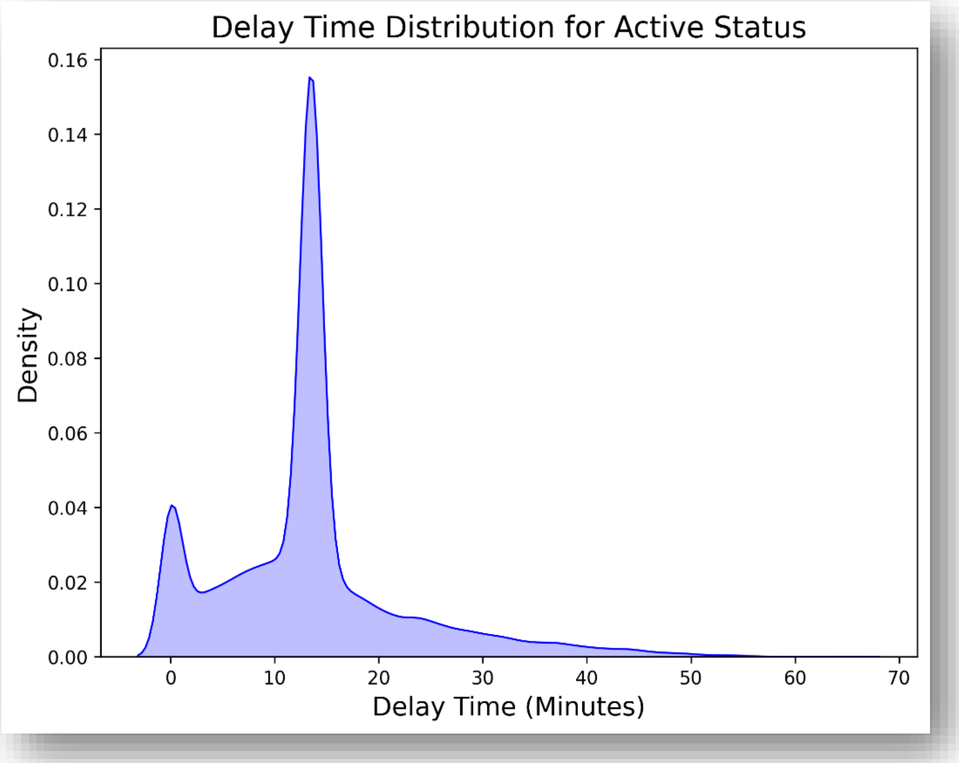


Delay Time vs Flight Status

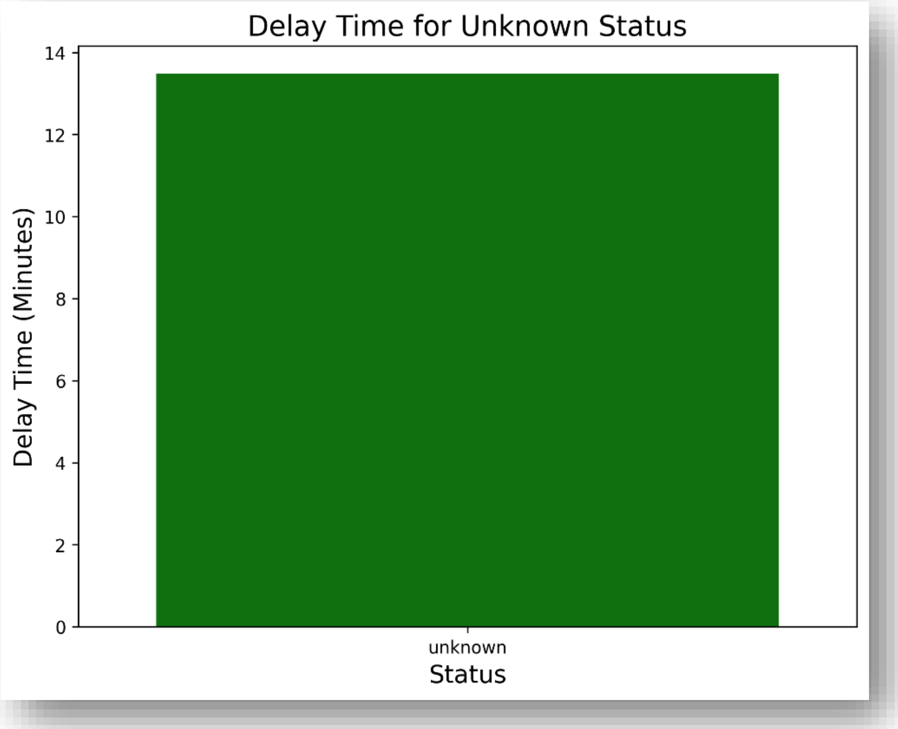


Individual Plots

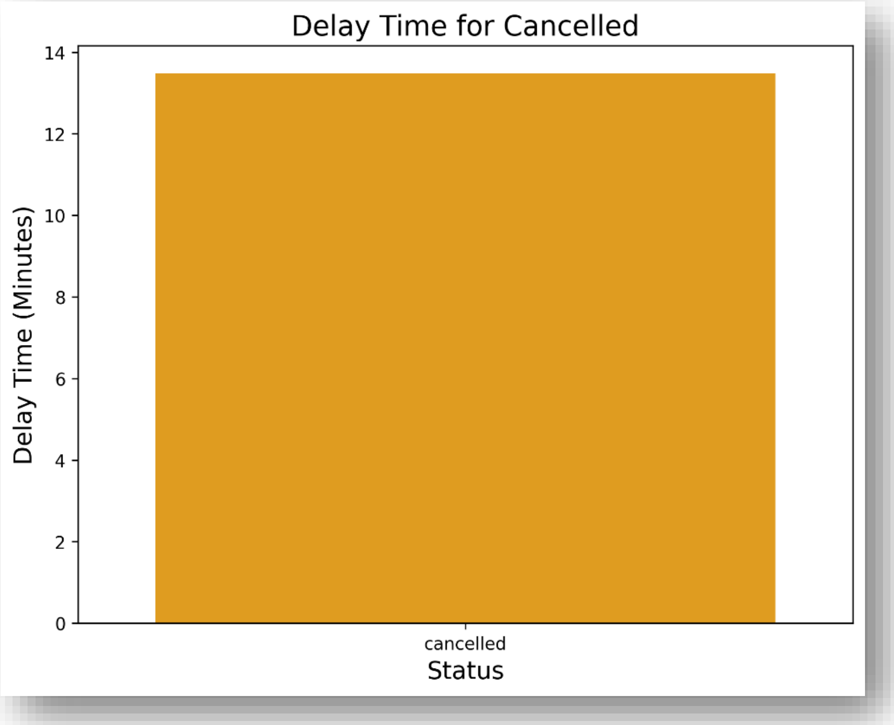
Active



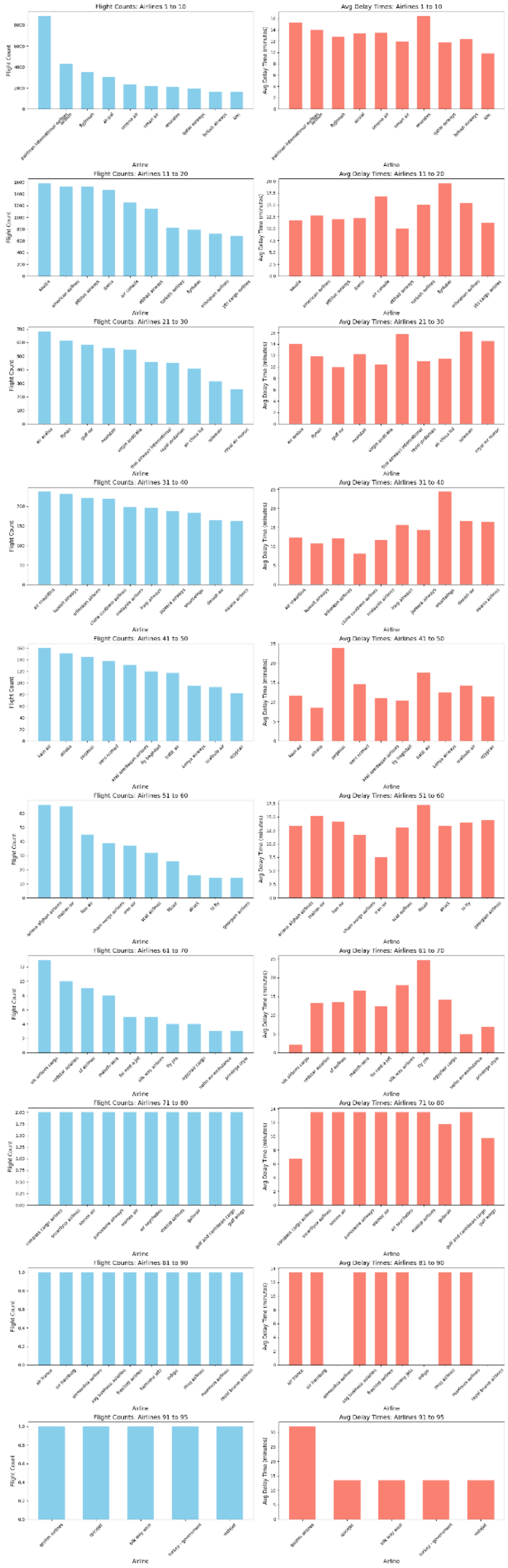
Unknown



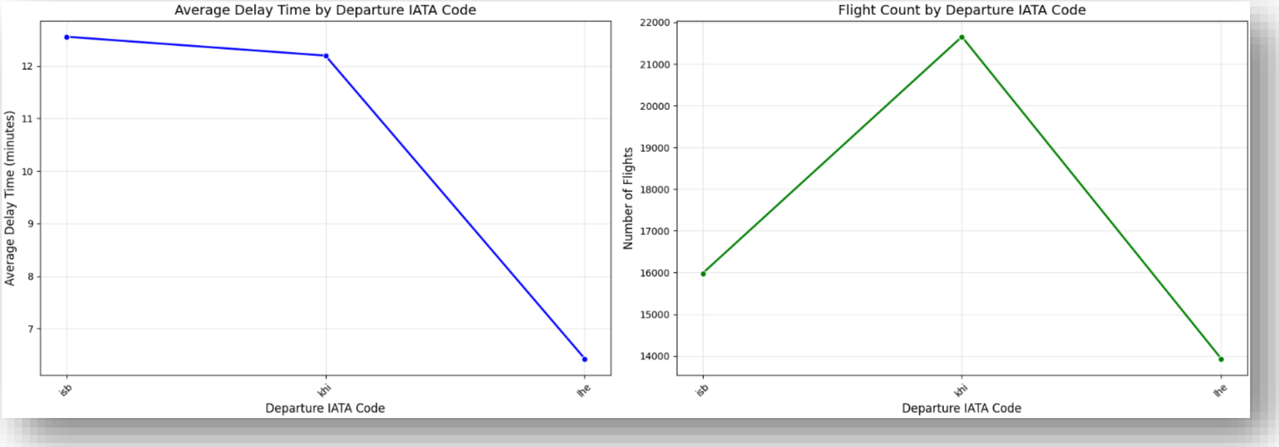
Cancelled



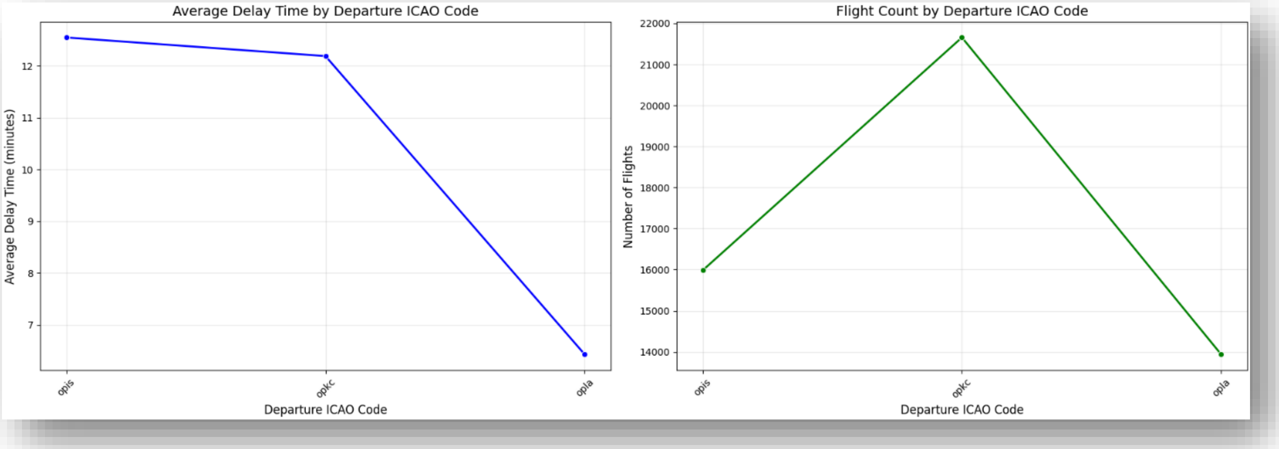
Delay Time vs Airline



Delay Time & Flight Count vs Departure IATA Code



Delay Time & Flight Count vs Departure ICAO Code

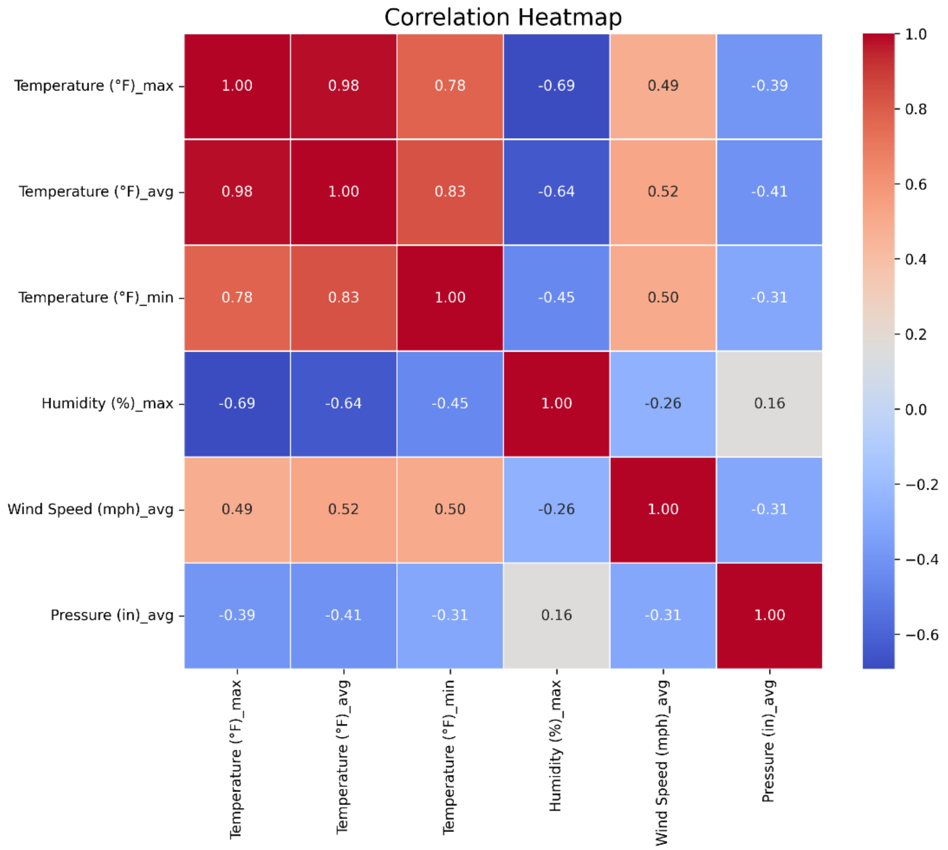


Correlation Analysis

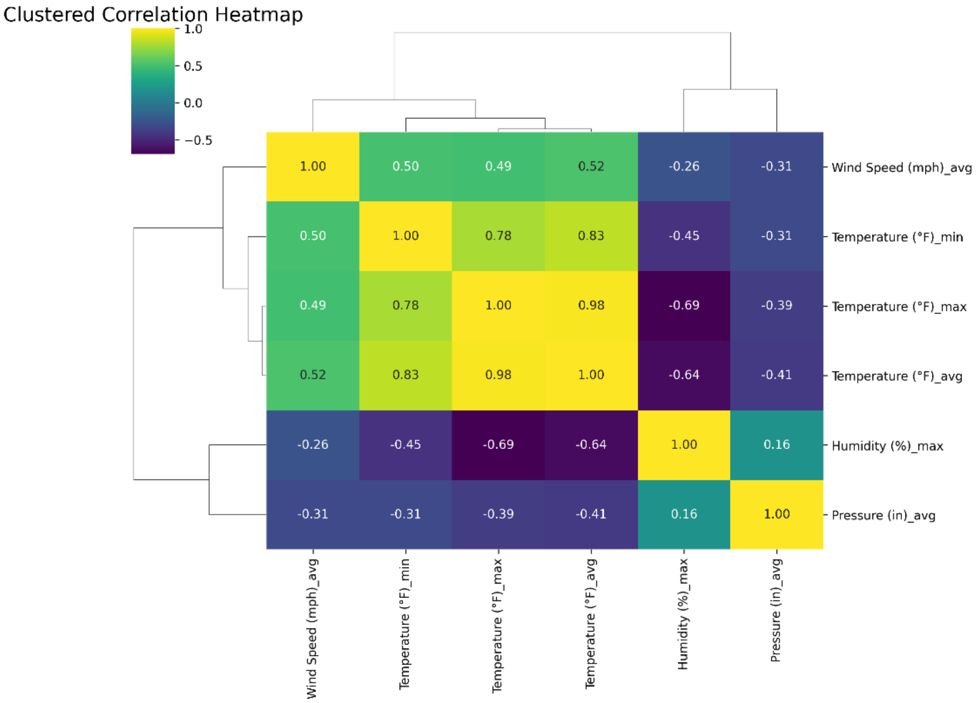
Features Selected:

- Temperature Max
- Temperature Avg
- Temperature Min
- Humidity Max
- Wind Speed Avg
- Pressure Avg

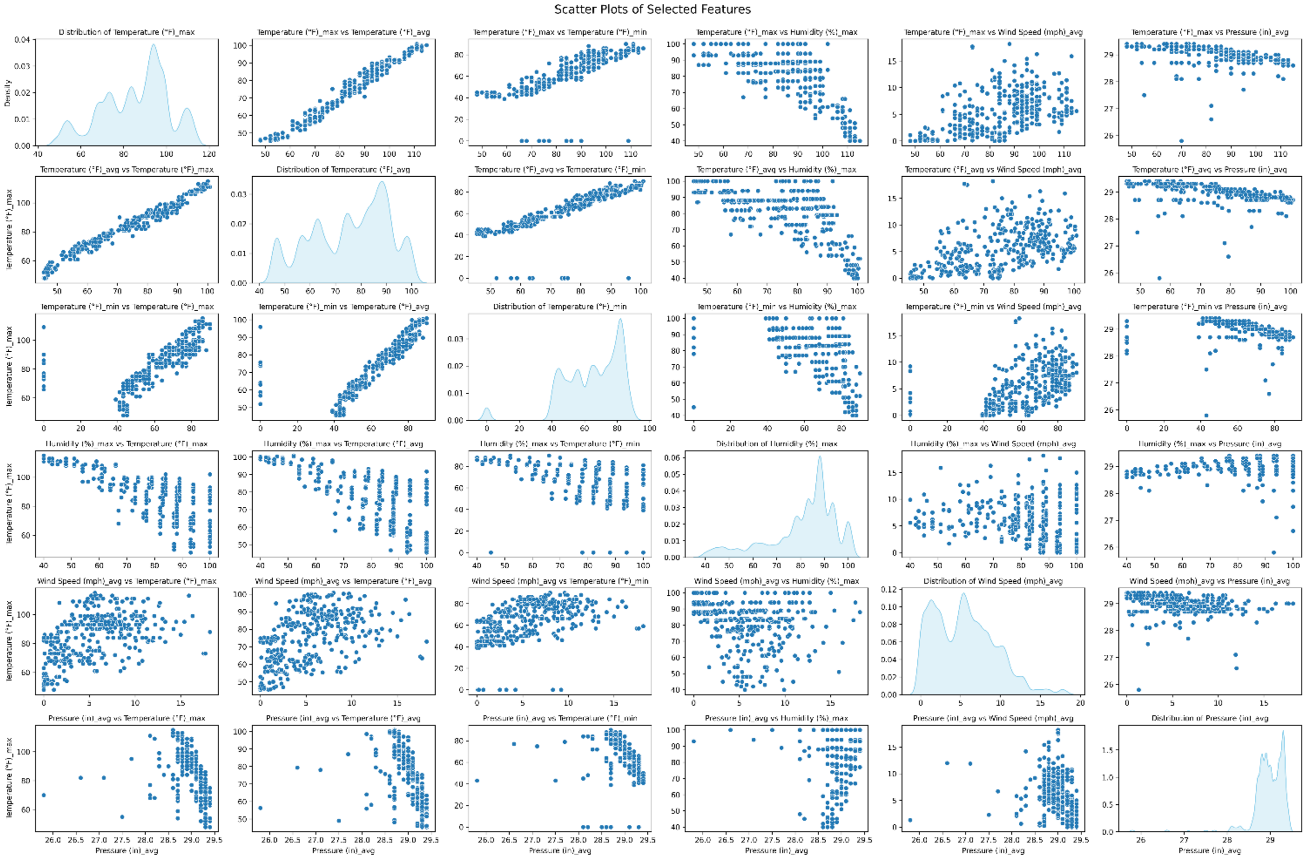
Heatmap



Clustered Heatmap



Scatter Plots



Phase 3: Analytical and Predictive Tasks

In this phase, the focus shifts to applying machine learning techniques to achieve key objectives. Three primary tasks are addressed:

1. **Binary Classification:** Predicting outcomes with two possible categories.
2. **Multi-Class Classification:** Distinguishing between multiple categories.
3. **Regression:** Predicting continuous values, such as flight delay durations.

These tasks leverage the insights and features prepared in earlier phases, ensuring accurate and meaningful predictions tailored to the problem at hand.

Binary Classification

Steps Taken:

Preprocessing

1. Hot encoding the airline names
    a. I encoded the top 10 airlines individually and then the other airlines I categorized as "other".
2. Hot encoding the day of the week.

3. Hot encoding the Month.
4. Hot encoding the departure ICAO code.
5. Hot encoding the departure IATA code.
6. Hot encoding the status colun.
7. Saving the encoded data to a csv.

Preparing Data

1. Using multiple ways to fill nan values
2. Selecting needed features
3. Converting delay time to delay time binary.
4. Oversampling Data
5. Applying PCA to reduce number of dimensions.

Prediction Making and Analysis

1. Scaled the data
2. Trained and Predicted using Linear Regression
3. Finding all the evaluation Metrics
    a. Accuracy
    b. Precision-Recall
    c. F1-Score
    d. Class-wise Precision-Recall
    e. Confusion matrix.
4. Trained and predicted once without PCA and once with PCA.
5. Lead to a reduction of accuracy of **0.02**.

Multi Class Classification

Steps Taken:

Preprocessing

1. Hot encoding the airline names
    a. I encoded the top 10 airlines individually and then the other airlines I categorized as "other".
2. Hot encoding the day of the week.
3. Hot encoding the Month.
4. Hot encoding the departure ICAO code.
5. Hot encoding the departure IATA code.
6. Hot encoding the status column.
7. Saving the encoded data to a csv.

Preparing Data

1. Using multiple ways to fill nan values
2. Selecting needed features
3. Converting delay time to delay time categories.
4. Oversampling Data
5. Applying PCA to reduce number of dimensions.

Prediction Making and Analysis

1. Scaled the data
2. Trained and Predicted using Random Forest Classifier and KNN.
3. Finding all the evaluation Metrics
    a. Accuracy
    b. Precision-Recall
    c. F1-Score
    d. Class-wise Precision-Recall
    e. Confusion matrix.
4. Trained and predicted once without Oversampling and once with Oversampling.
5. Lead to an increase of accuracy from **0.89** to **0.94**.

Multi Class Classification

Steps Taken:

Preprocessing

1. Hot encoding the airline names
    a. I encoded the top 10 airlines individually and then the other airlines I categorized as "other".
2. Hot encoding the day of the week.
3. Hot encoding the Month.
4. Hot encoding the departure ICAO code.
5. Hot encoding the departure IATA code.
6. Hot encoding the status column.
7. Saving the encoded data to a csv.

Preparing Data

1. Using multiple ways to fill nan values
2. Selecting needed features
3. Converting delay time to a proportion of the day using the following formula:
    a. $X = x / 24 * 60$
4. Applying PCA to reduce number of dimensions.

Prediction Making and Analysis

1. Scaled the data
2. Trained and Predicted using Random Forest Regressor and Linear Regression.
3. Finding all the evaluation Metrics
    a. Mean Absolute Error (MAE)
    b. Root Mean Square Error (RMSE)
4. Got an MAE of 0.02 and RMSE of 0.03.

Phase 4: Model Optimization and Evaluation

This phase is dedicated to refining predictive models to maximize their performance and reliability. The focus is on fine-tuning, validation, and comparison to identify the best-performing models for the given tasks.

1. **Hyperparameter Tuning:**
   Techniques like grid search and random search are employed to optimize model parameters, ensuring the models are well-suited to the data.
2. **Validation:**
   K-fold cross-validation is applied to rigorously assess model performance, minimizing the risk of overfitting and ensuring robust generalization to unseen data.
3. **Model Comparison:**
   Finally, the performance of different models is systematically compared to select the most accurate and efficient model for deployment.

This phase ensures that the selected models are not only accurate but also reliable and optimized for the intended predictive tasks.

Phase 5: Model Testing:

In the final phase, the trained models are evaluated on the test dataset to generate predictions and assess their real-world performance. This step ensures that the models generalize well to new, unseen data.

1. **Prediction Generation:**
   The trained models are used to make predictions on the test dataset. Depending on the task, predictions include:
   - **Regression:** Predicting the exact delay in minutes.
   - **Classification:** Predicting delay categories or binary outcomes such as "on-time" or "delayed".
2. **Formatting Predictions for Submission:**
   The predictions are saved in the Kaggle submission format. For classification tasks, the delay column is represented in a string format (e.g., "on-time" or "delayed") instead of numerical values like 0 or 1, adhering to the required submission guidelines.

This phase ensures that the results are both interpretable and aligned with the competition requirements, ready for final submission.