**AIR TRAFFIC DATA ANALYSIS**

***BY:*** *Disha Hegde*

*BSc CS graduate*

*from Bengaluru City University.*

***Mentor*:** *Bipul Kumar*

**ABSTRACT**

As more commercial airplanes take off to the sky, data generated from air traffic is rapidly increasing in both size and complexity. According to the International Air Transport Association (IATA), Air travel industry is anticipated to annually grow by 3.5% in the next 2 decades (IATA, 2018). This growth raises big questions on how to gather, store, analyze and use the data generated from those flying cities in the sky. In this project, I start by using flight route information to produce a scalable and realistic simulation model for air traffic, and in the second phase we analyze the delay distribution and then evaluate some machine learning algorithms so we can have a better scientific foundation on which machine learning algorithm can produce better and more accurate prediction on flight delays.

**We are developing better algorithm that enable decision support tool development. In order to do so we are using Python programming language, dynamic programming and machine learning. We will take input and trained that input data and plot that on the graph in order to understand the situation.**

**INTRODUCTION**

**Air Traffic Data Analysis**

# Over the last few decades, air transport is increasing in popularity because of its speed and comfort which eventually increases the traffic in the airspace. With the great increase in air traffic comes a large increase in the demand for capacity in airports and airspace. However, the capacity in the aviation industry is often a tight constraint and it cannot keep increasing at the rate necessary to match the rising demand. During peak hours, the demand for resources in both airports and airspace is at its highest.

Airspace congestion and flight delays are two of the most important bottlenecks factors that limit the available resources and causes multiple unhealthy side effects on both the operation of air industry and thus the growth of the economy. It is therefore crucial to have a good trained human resource to manage the airspace and ground operations to avoid any leak of capacity.

In this project we shall investigate the possibilities of using big data technologies to solve such challenges and limitations.

* 1. **Objective**

The objective of this capstone project is to analyze air traffic data to derive insights into flight

patterns, passenger trends, and operational efficiencies within the aviation industry.

* 1. **Background**

# The aviation industry generates vast amounts of data related to flight schedules, passenger

# demographics, aircraft performance, and airport operations. Analyzing this data can provide

# valuable insights for airlines, airports, and aviation authorities to optimize resources, improve

# customer service, and enhance safety measures. Your task is to leverage data science

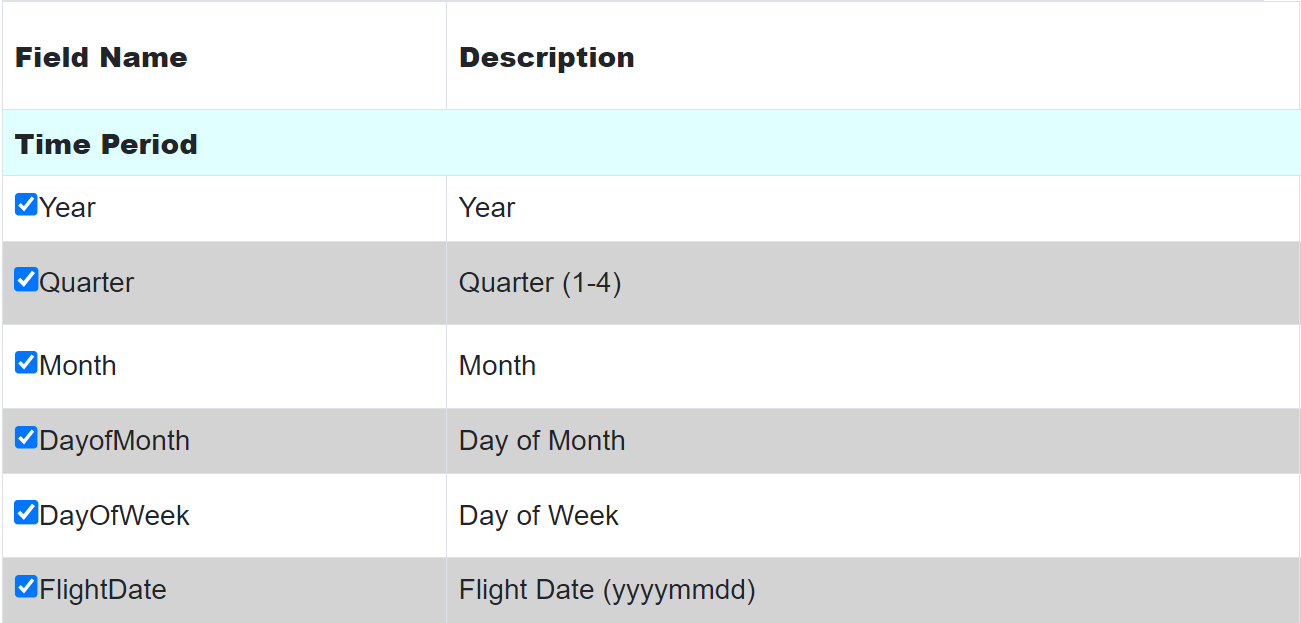
# techniques to explore and analyze air traffic data to uncover meaningful patterns and trends.

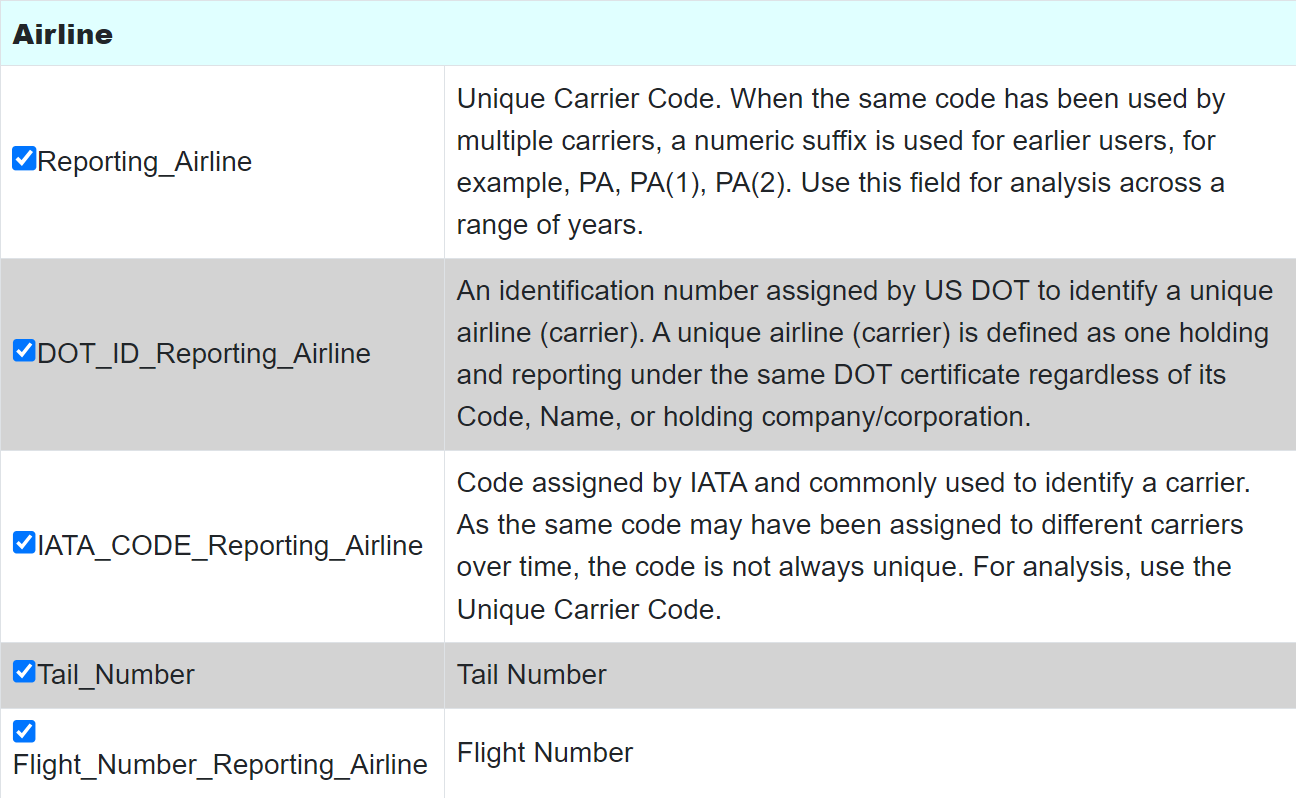
# **PROJECT DESCRIPTION**

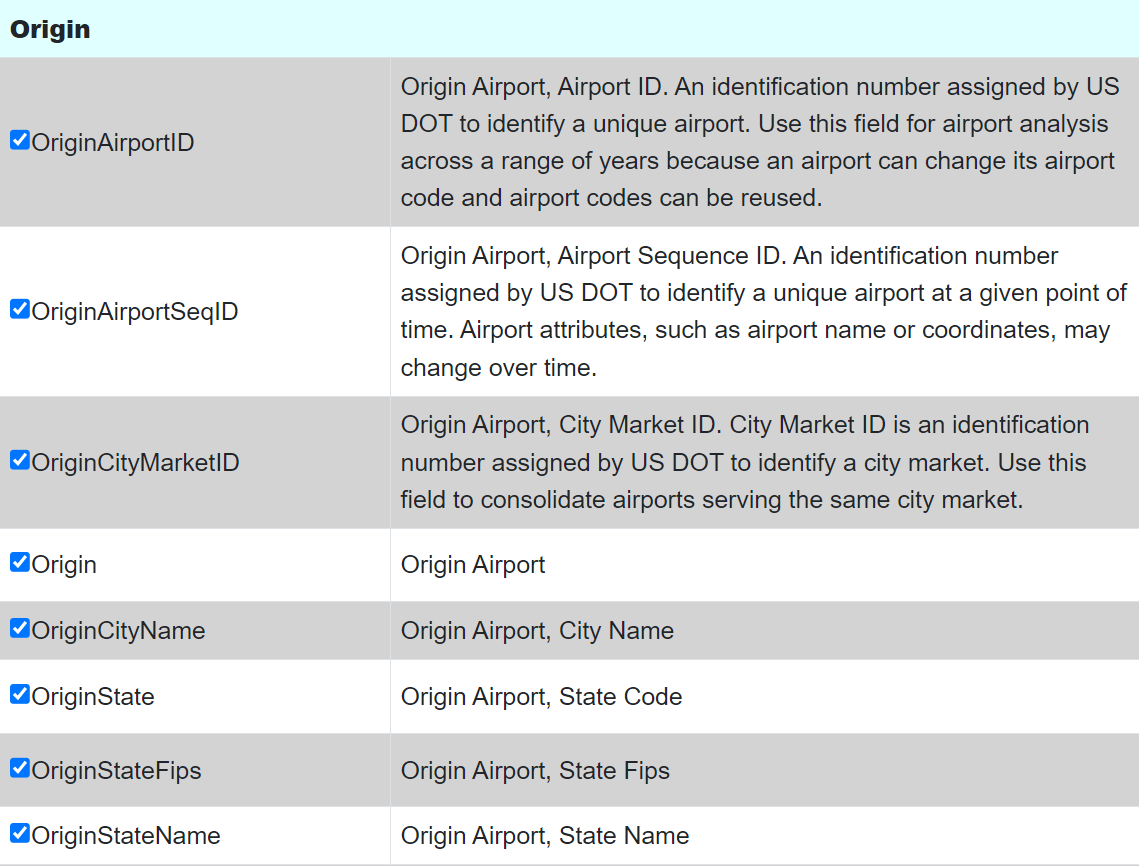
# **2.1. Dataset**

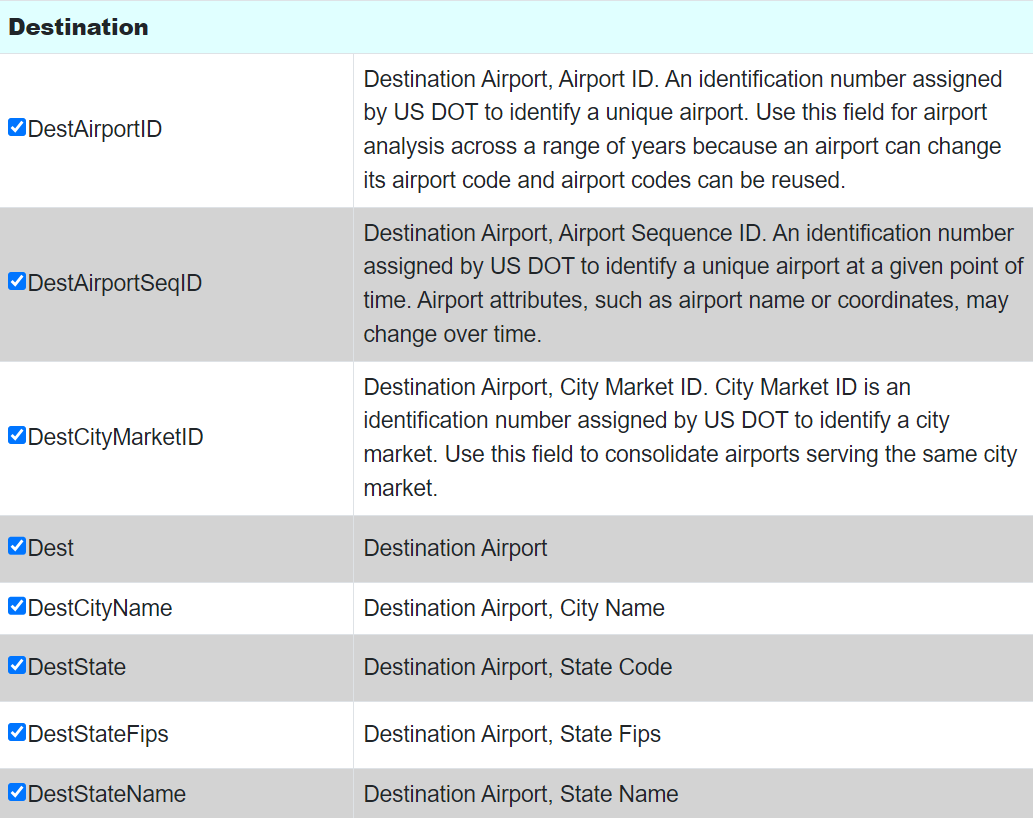
Dataset source: Bureau of Transportation Statistics (BTS) [[dataset source link](https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=)]

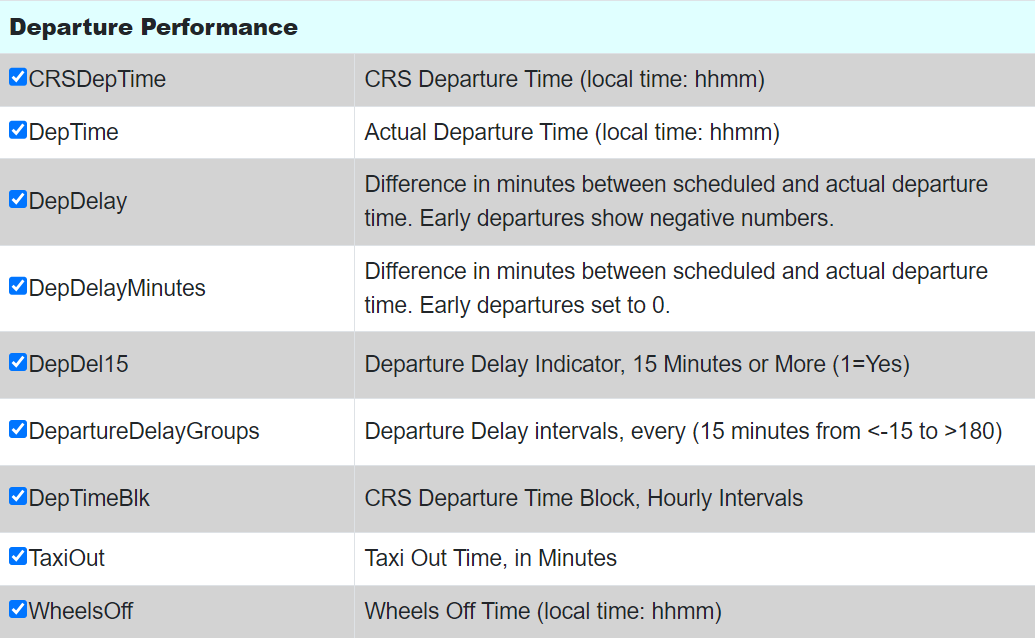
In my dataset, I have chosen the following attributes to build and train my model:

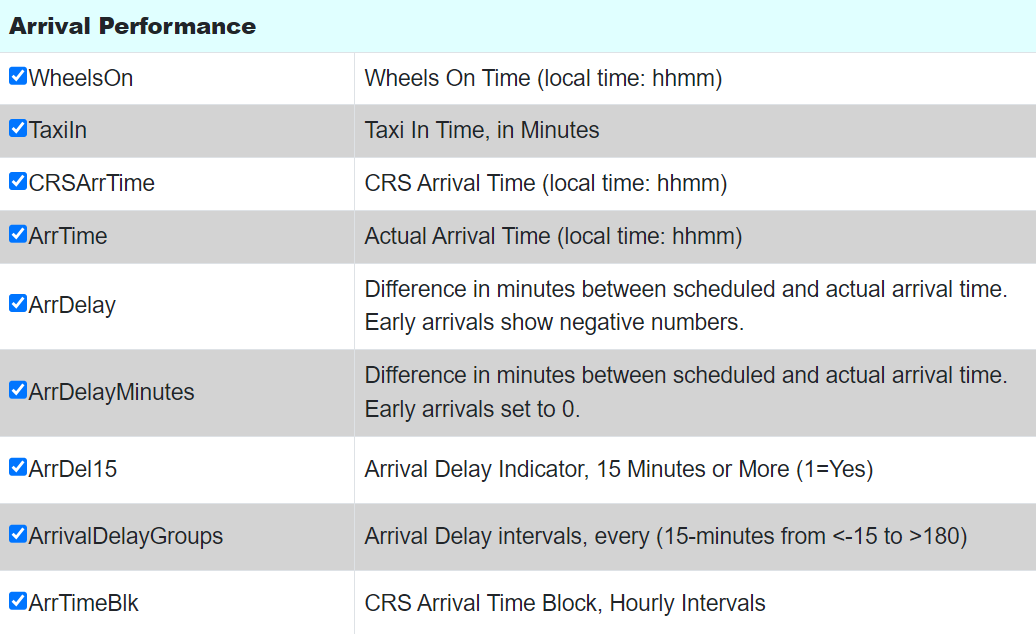


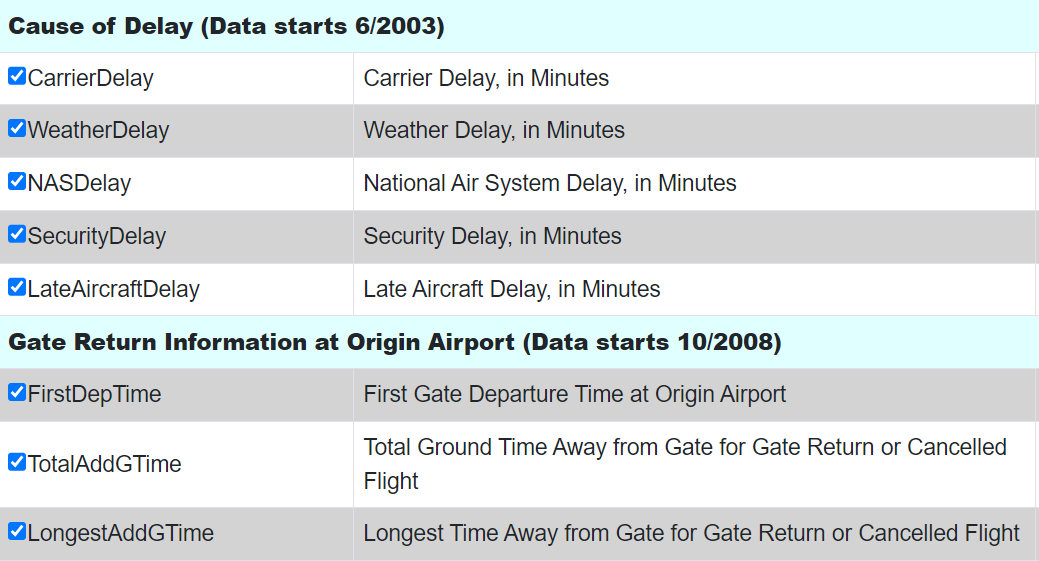












**2.2. Concepts Used**

**Data Preprocessing**:

The dataset undergoes extensive preprocessing to prepare it for modelling. This includes handling missing data, creating new features, and one-hot encoding categorical variables such as carriers and airports. Data normalization is also performed to scale the features, ensuring that they contribute equally to the model’s predictions.

**Regression Modelling**:

Linear Regression is used to model and predict the Total Delay. The model is trained on a portion of the dataset, and its performance is evaluated on a separate test set. Metrics such as Root Mean Square Error (RMSE) and R² scores are used to assess how well the model predicts the delays.

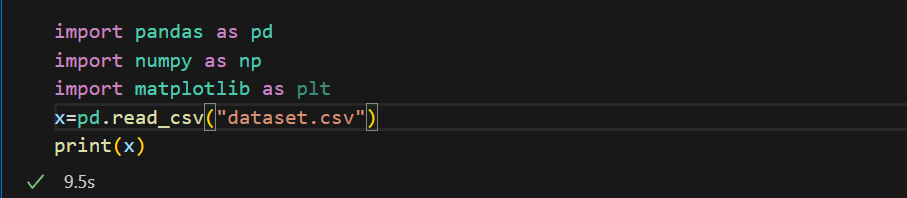
**Model Evaluation and Visualization**:

The performance of the regression models is visualized through scatter plots of actual versus predicted delays. Clustering results are also visualized to show the distribution of delays within different clusters. These visualizations are crucial for interpreting the models' predictions and understanding the underlying data patterns.

**METHODOLOGY**

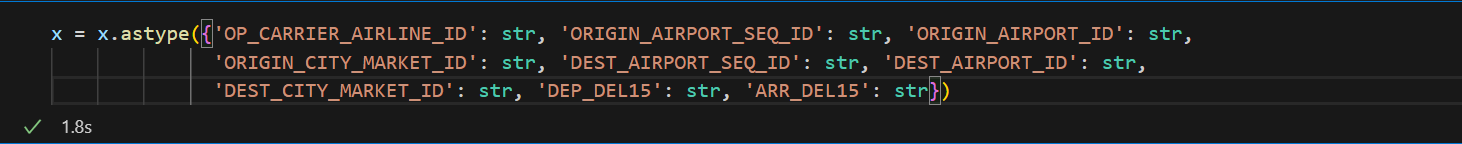
**Data Preprocessing:**

Importing the necessary libraries: Pandas for data manipulation, NumPy for numerical operations, and Matplotlib for plotting. Reading a CSV file named dataset.csv into a Pandas DataFrame. Printing the contents of the DataFrame to the console



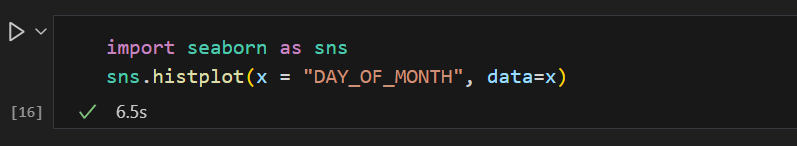
**Data Transformation:**

Changing the data types of specific columns in the DataFrame. This code will ensure the specified columns are treated as strings, which is helpful for data manipulation and analysis tasks where these columns are used as categorical identifiers rather than numerical values.

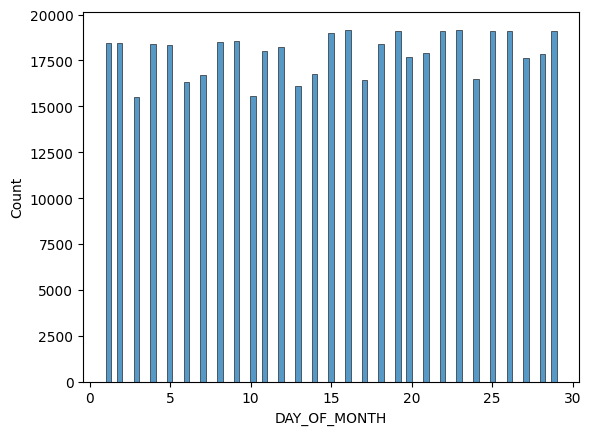


**Exploratory Data Analysis (EDA):**

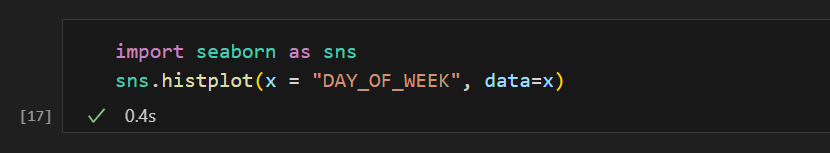
Grouping the flight data by DAY\_OF\_MONTH:



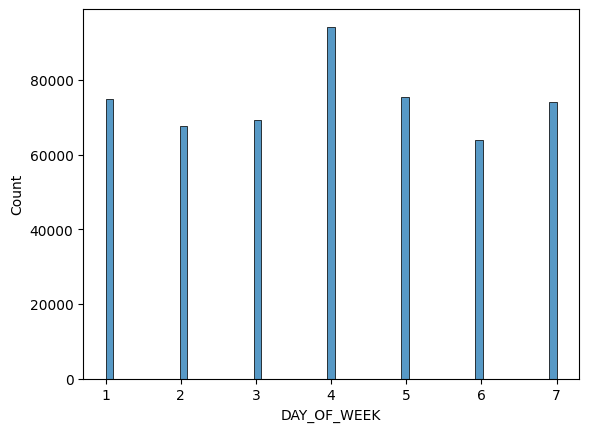
In this section, I have grouped the flight data by day of the month and plotted a graph for the total number of flights that flew through all the airports each day of the month, as seen in the graph below.



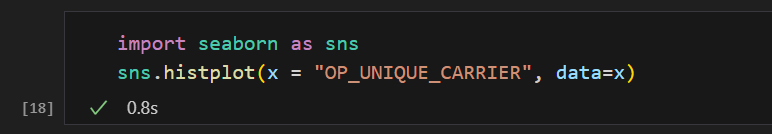
Grouping the flight data by DAY\_OF\_WEEK:



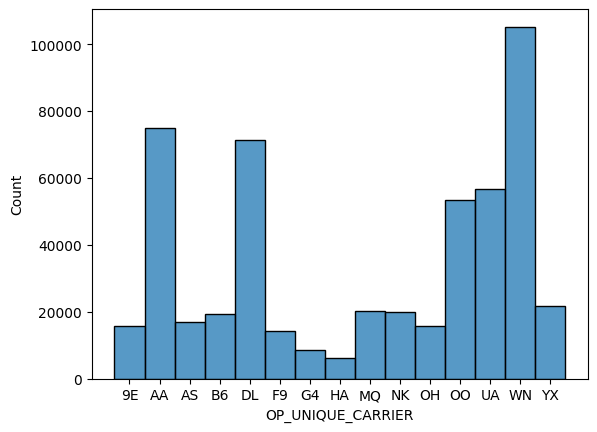
In this section, I have grouped the flight data by day of the week and plotted a graph for the total number of flights that flew through all the airports each day in a week, as seen in the graph below.



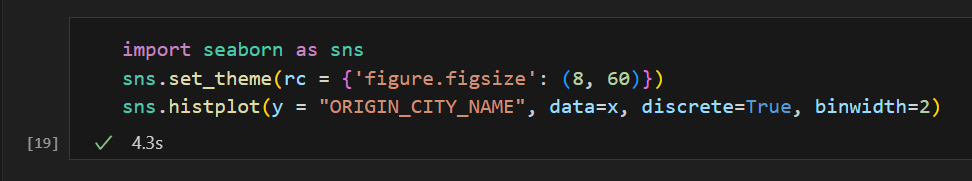
Analyzing the distribution of the type of flights:



This code block plots a histogram of the type of airplanes that flew in the that particular month, as seen below.



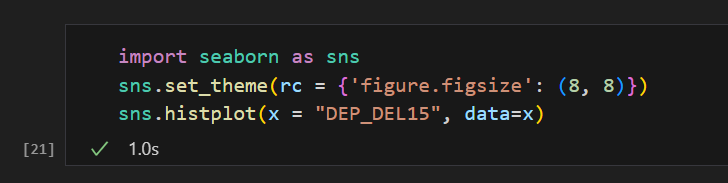
Analyzing the distribution of the flights across all the airports:



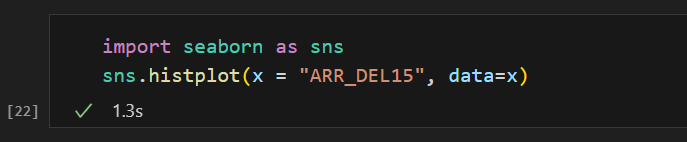
This code block plots a histogram of the number of flights that flew across all the city airports in the that particular month.

**Delay Analysis:**

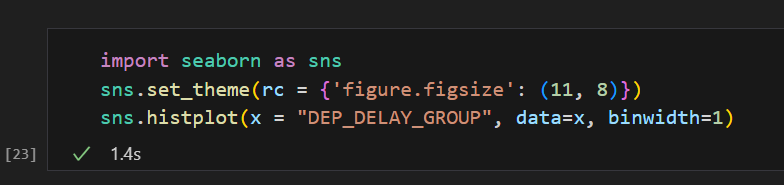
Departure Delay: this tells us about the flights whose departure was delayed by 15 minutes or more (1=yes).



Arrival Delay: this tells us about the flights whose arrival was delayed by 15 minutes or more (1=yes).

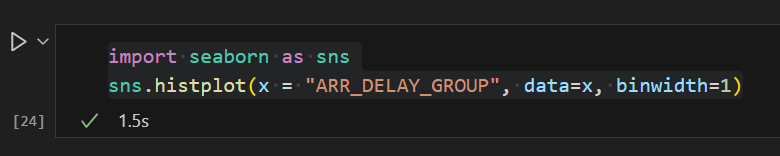


Grouping by Departure delay: visualizing the distribution of flights based on their departure delay groups in 15-minute increments.



* 0: On-time or <15 minutes’ delay.
* Positive values: Delays in 15-minute increments (e.g., 1 for 15-29 minutes).
* Negative values: Early departures (e.g., -1 for 1-14 minutes early).

Grouping by Arrival delay: visualizing the distribution of flights based on their arrival delay groups in 15-minute increments.

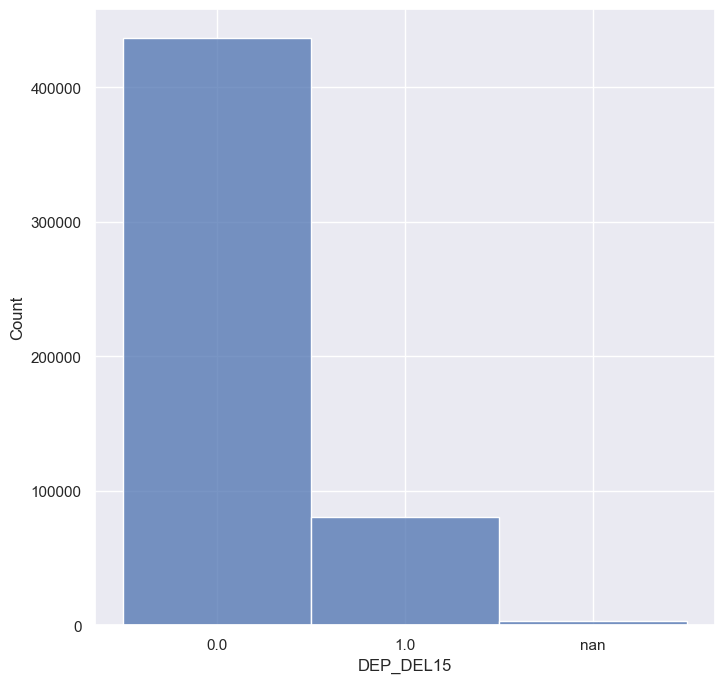


* 0: On-time or <15 minutes’ arrival.
* Positive values: Arrival in 15-minute increments (e.g., 1 for 15-29 minutes).
* Negative values: Early arrival (e.g., -1 for 1-14 minutes early).

Delay Analysis Graphs:

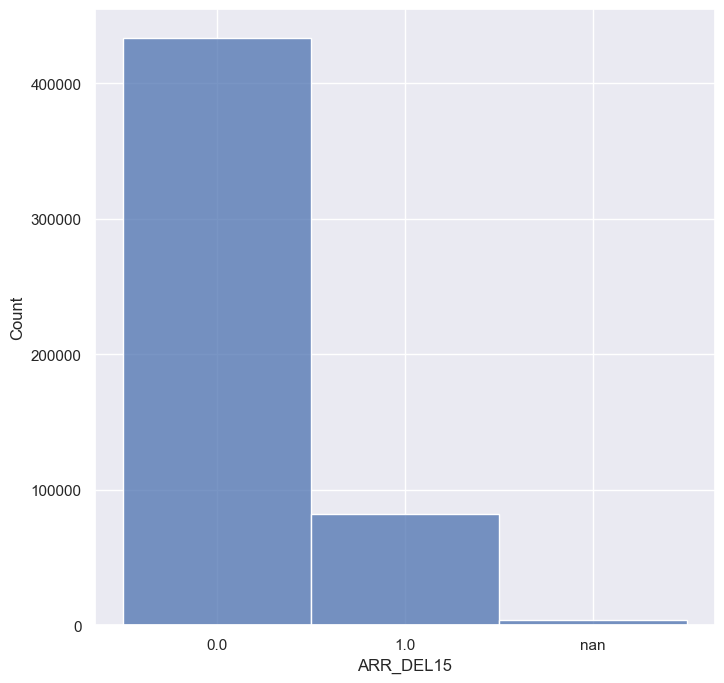
1. Histogram showing the flights delayed by 15 mins.

(1= yes, 0= delayed by < 15 mins, nan= not reported)

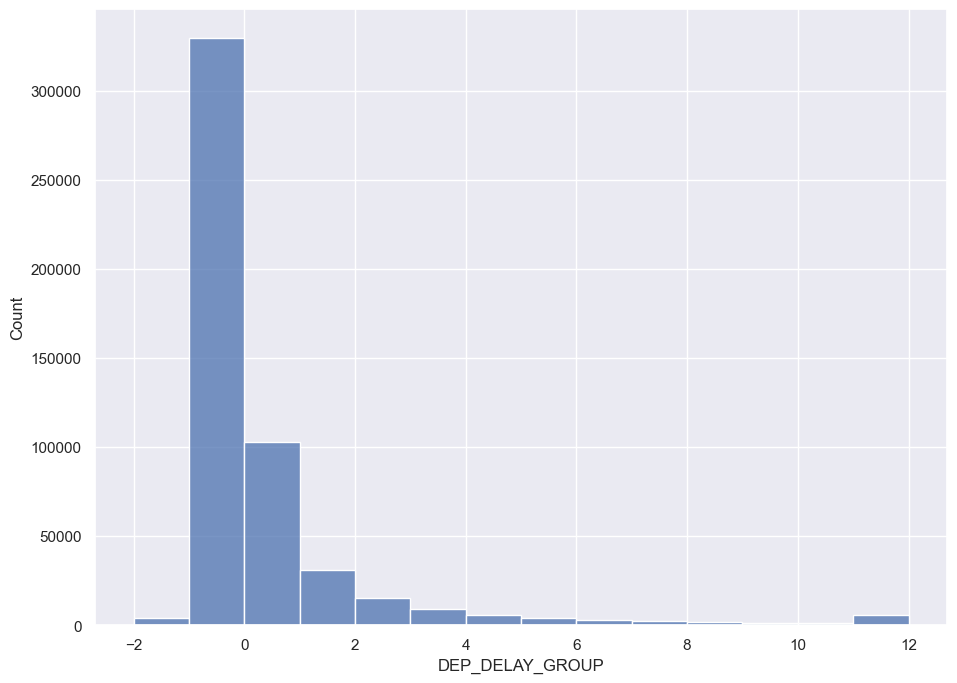


1. Histogram showing the flights delayed by 15 mins.

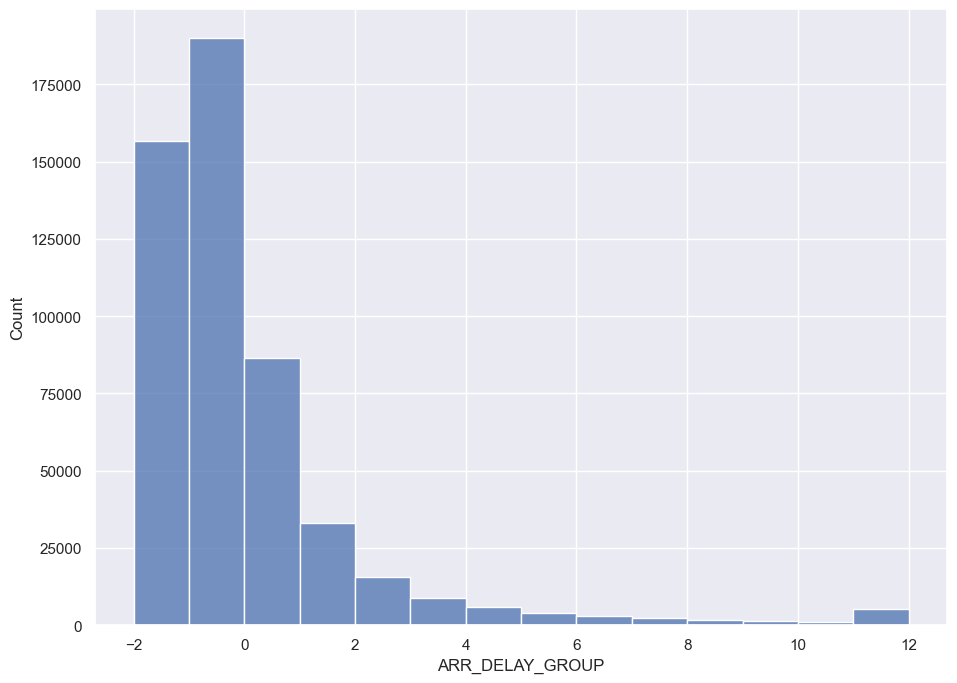
(1= yes, 0= delayed by < 15 mins, nan= not reported)



1. Histogram showing the distribution of flights based on their departure delay groups in 15-minute increments:

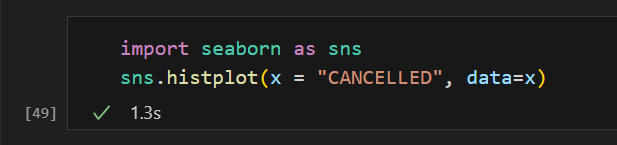


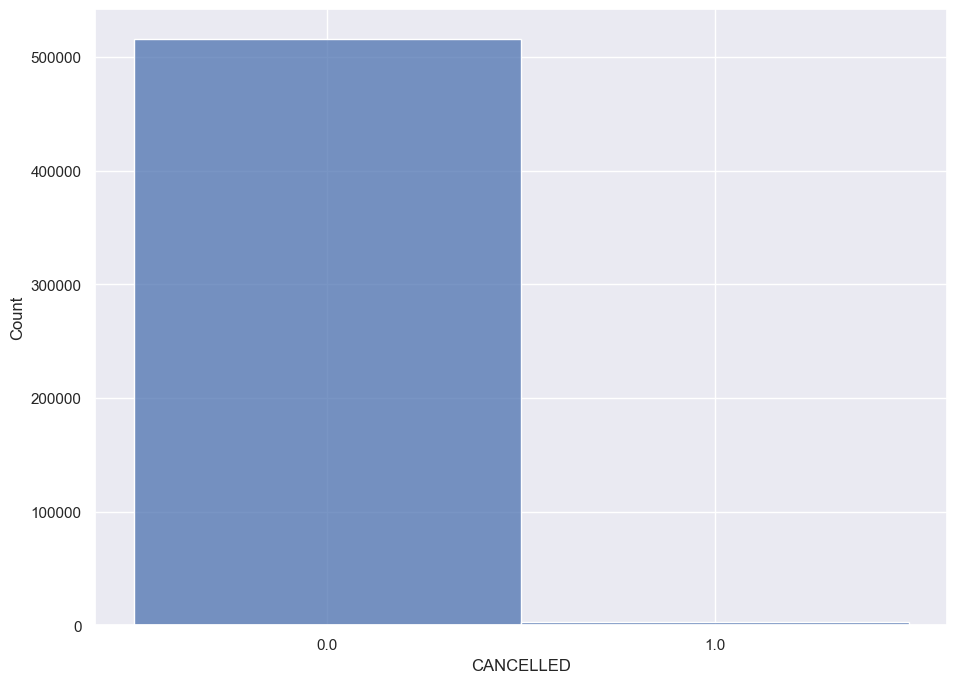
1. Histogram showing the distribution of flights based on their arrival delay groups in 15-minute increments:



**Analysis of Cancellations:**

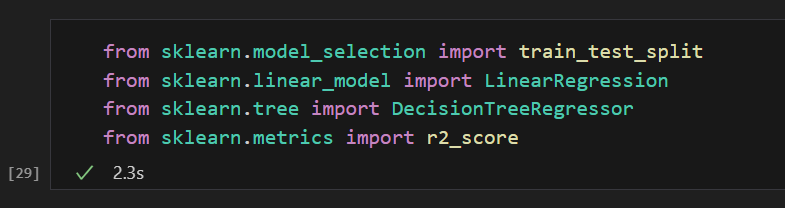
visualizing the distribution of flights based on flight cancellations.

****

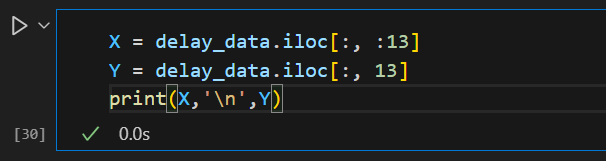


**REGRESSION MODEL TRAINING AND EVALUATION**

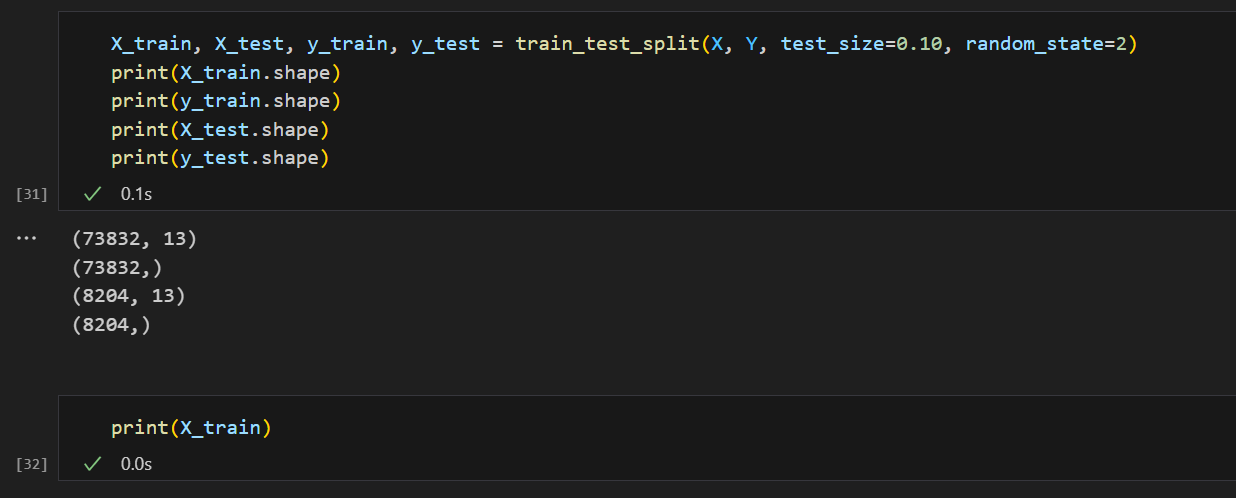
* Setting the environment and importing all the required libraries:

****

* Next, code selects columns from a DataFrame called delay\_data:
* X contains the first 13 columns.
* Y contains the 14th column.
* It then prints X and Y.

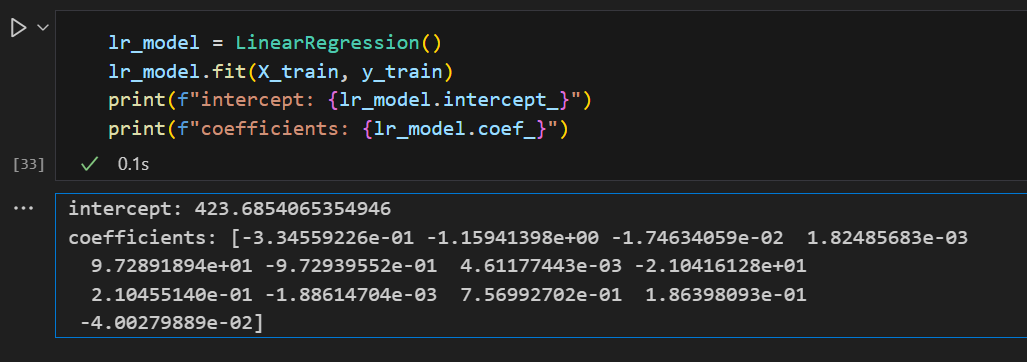


* Training model using train\_split\_test from the sklearn.model:

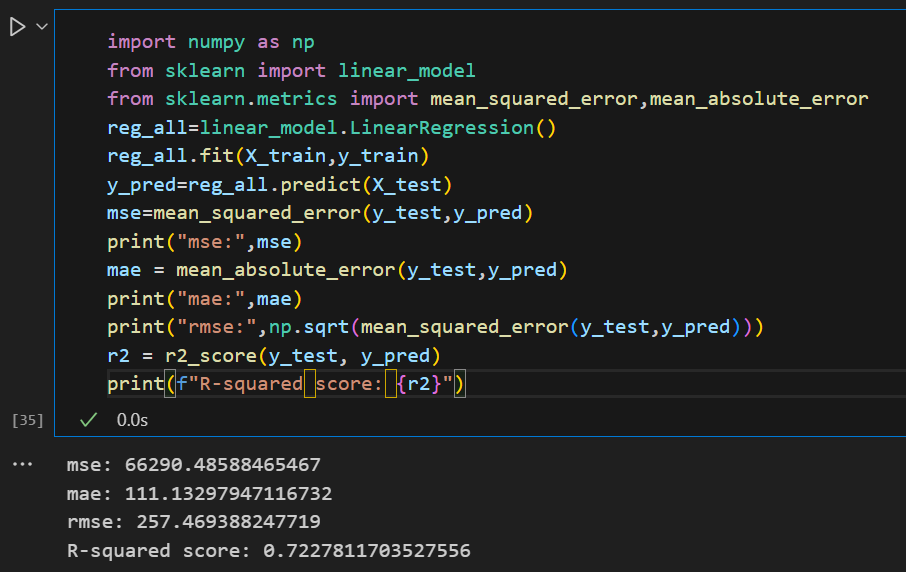


* **Linear Regression:**

1. Model training:

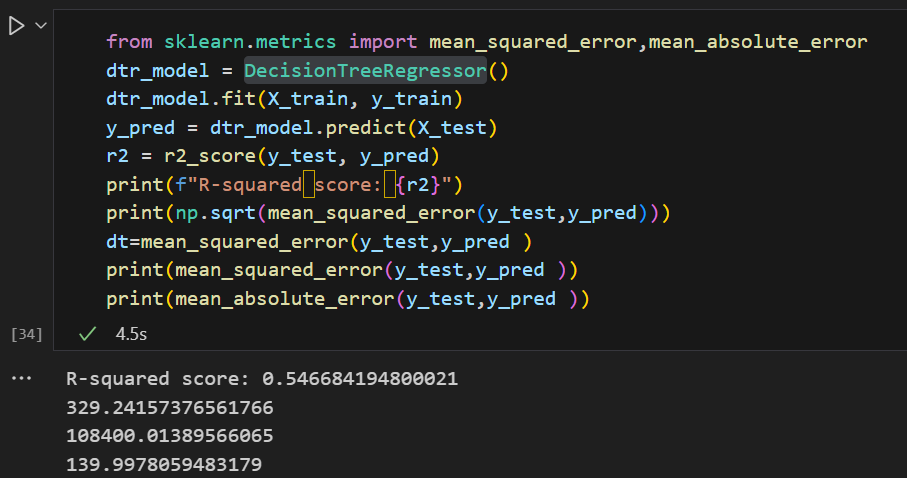


1. Model performance score:



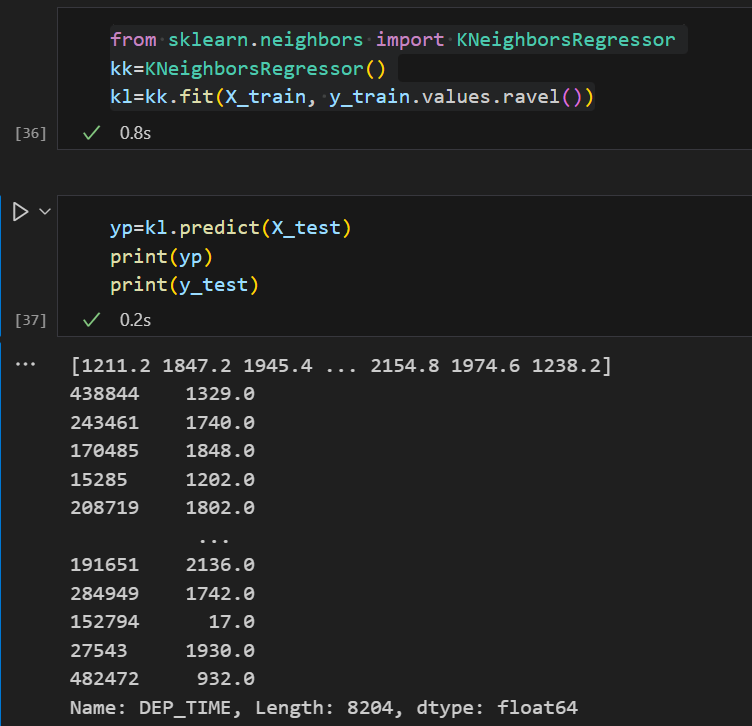
* **Decision Tree Regressor:**

Model training and performance score:

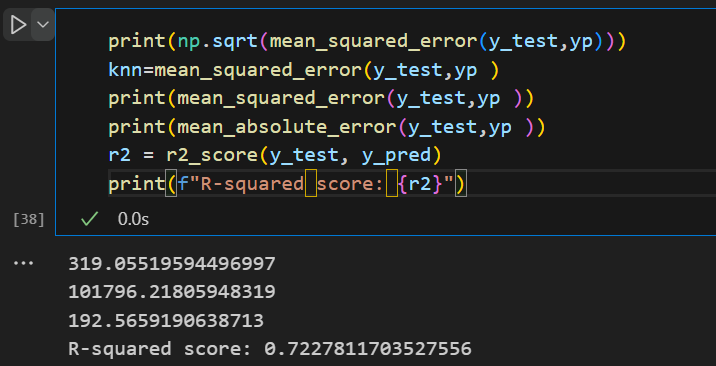


* **KNieghborsRegressor:**

Model training, training and prediction:

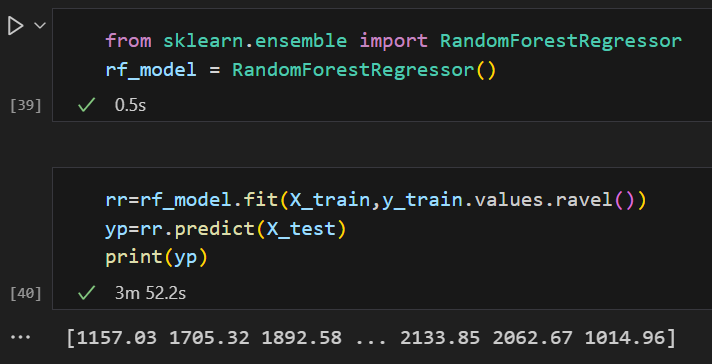


* The code evaluates the performance of a regression model using various metrics. It first calculates the root mean squared error (RMSE) between the actual (y\_test) and predicted (yp) values, printing the result. The mean squared error (MSE) is computed next and stored in the variable knn, then printed. Following that, the mean absolute error (MAE) is calculated and printed to show the average absolute difference between predicted and actual values. Finally, the R-squared score, which measures the proportion of variance in the dependent variable predictable from the independent variables, is computed and printed using r2\_score.



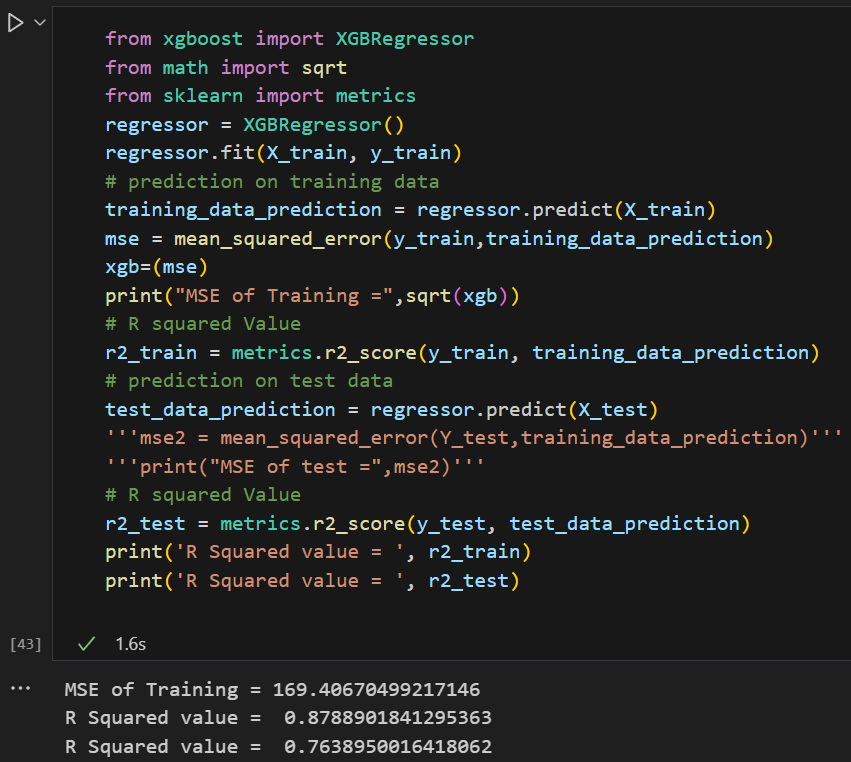
* **Random Forest Regressor:**

Model training and prediction:

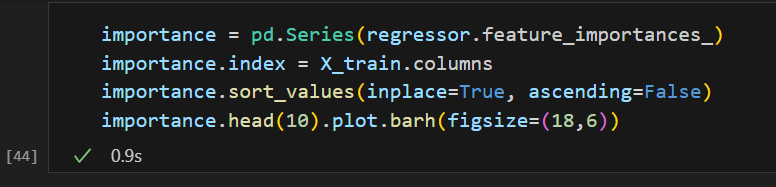


* **XGBRegressor from XGBoost:**

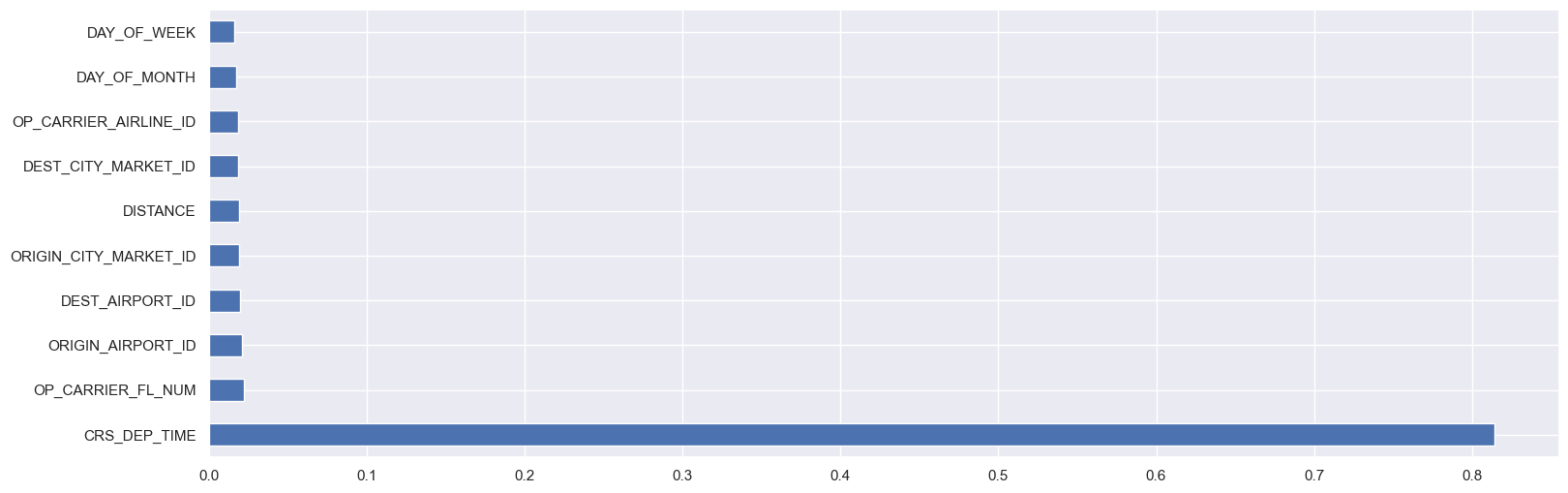
Using XGBRegressor, prediction on training and testing data is made and then the R square value (r2\_score) for both is calculated respectively.



* Plotting a graph that shows the importance of feature columns and highlights the attribute that added the most value in all the prediction models.



Graph depicting the importance of feature column:

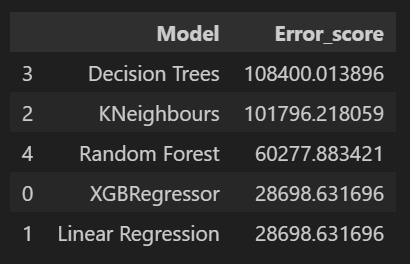


* Plotting a graph that shows the error score of all the machine learning regression models and depicts which model works the best on the dataset and helps best with air traffic data analysis.

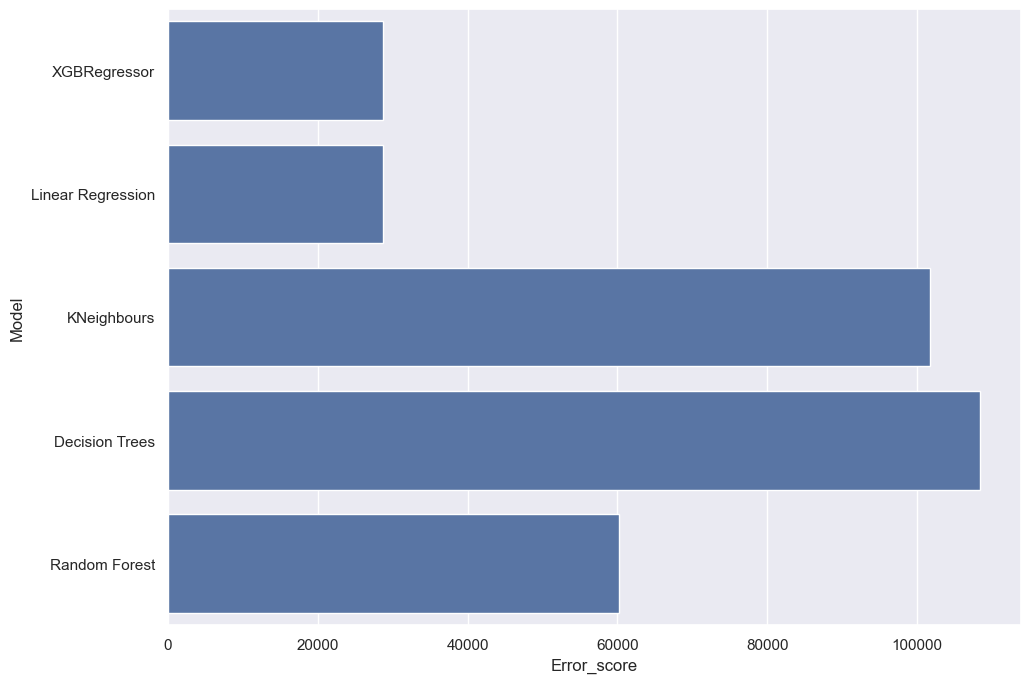
I am also calculating error score for each model separately.



Error score of each regression model:



Graph depicting the error score for each regression model:



**CONCLUSION:**

In this report, a comprehensive analysis was conducted on air traffic delays using a dataset from the Bureau of Transportation Statistics. The dataset undergoes extensive preprocessing to prepare it for modelling. This includes handling missing data, creating new features, and one-hot encoding categorical variables such as carriers and airports. Data normalization is also performed to scale the features, ensuring that they contribute equally to the model’s predictions. Exploratory Data Analysis (EDA) was performed to uncover patterns and relationships within the dataset, including delay and cancellation analysis.

The performance of the regression models is visualized through graphs of actual versus predicted delays. These visualizations are crucial for interpreting the models' predictions and understanding the underlying data patterns.

Based on these findings, several measures can be recommended to mitigate delays. For weather-related delays, improving real-time weather monitoring and implementing adaptive scheduling can help manage and reduce disruptions. For airport management, optimizing operational procedures, enhancing coordination among ground services, and implementing advanced forecasting systems can further decrease delays. These strategies, informed by the detailed analysis and predictive insights, can contribute to more efficient air traffic management and improved passenger experiences.

The Project is also updated on Github, under this link.

**REFERENCES**:

1. Dataset [and](https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=) data information source link
2. Machine learning [models](https://www.geeksforgeeks.org/ml-linear-regression/) guide
3. Reference [project](https://github.com/xiaojunzhao/Airline-Traffic-Data-Analysis) I used to understand the code better