Applying Machine Learning to Aviation Data for Prediction of Flight Delay Propagation

Dorcas Engasia Misango[a]

[a]*Admission No:, 100235,*

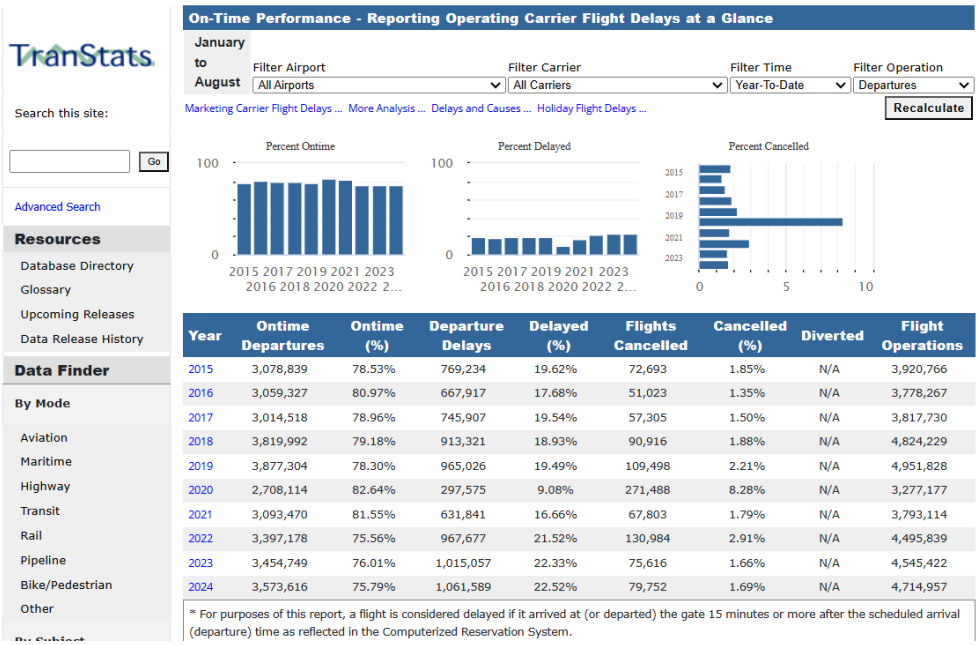
Abstract

In the aviation sector, flight delay propagation is a serious problem since a single flight delay can cause considerable operational disruptions and monetary losses throughout a network. Through the analysis of historical flight data from a local airline, spanning from September 2023 to August 2024, this study aims to comprehend and mitigate delay propagation. This study investigates the factors influencing flight delays and their propagation within a "hub and spoke" aviation network, with a focus on Nairobi (NBO) as a central hub. Flights are divided into categories according to their tendency for delays using sophisticated clustering techniques, with the most and least delay-prone groups being created. To find trends and drivers of delay propagation, important variables such as scheduled and actual arrival/departure timings, turnaround time (TAT), aircraft registration ID, and delay reasons are examined. To forecast delay durations and evaluate the effect of delay propagation across aircraft clusters, machine learning models such as KMeans Clustering, Random Forest Regressor and HistGradientBoostingRegressor are utilized. Additional features and hyperparameter tuning were also employed to help the models make more accurate predictions.

This study gives airlines practical insights to prioritize interventions, improve resource allocation, and implement targeted measures to minimize delay propagation by grouping flights into delay-prone categories. The practical deployment and usability for stakeholders are ensured by incorporating these insights into the current airline management systems. Reduced operational disruptions, financial savings, and an enhanced passenger experience are among the expected results. This study opens the door to a more robust and effective aviation network by providing a data-driven method for understanding and controlling the spread of aircraft delays.

1. Introduction

It is undeniable that the use of air transport has grown over the years. Air travel seems to have picked up, and so have the number of delays being recorded (Lesgourgues and Malavolti, 2023).In the US, the number of delays experienced this year is about 20% of all departures that took place in various airports. This number is a slight increase from last year but as compared to before COVID-19, there is a significant rise. A delay in the airline industry causes a ripple effect on various stakeholders, including passengers, airport handlers (GSA), and the crew of the airline of interest. With more sectors of the economy depending on air transport to deliver their goods and also efficiently get them across vast distances so they can conduct their business or even enjoy a vacation, the impact of these delays goes beyond just financial consequences to the airlines (**JOUR**).In Kenya, air travel plays a crucial role in the economy, where air travel has created approximately 410,000 jobs and In total, 4.6 percent of the country's GDP is supported by inputs to the air transport sector and foreign tourists arriving by air ("IATA Kenya Report", n.d.)(Tchouamou Njoya et al., 2020). These delays can have substantial effects that go beyond even the airport; to businesses that rely on



Year	Ontime Departures	Ontime (%)	Departure Delays	Delayed (%)	Flights Cancelled	Cancelled (%)	Diverted	Flight Operations
2015	3,078,839	78.53%	769,234	19.62%	72,693	1.85%	N/A	3,920,766
2016	3,059,327	80.97%	667,917	17.68%	51,023	1.35%	N/A	3,778,267
2017	3,014,518	78.96%	745,907	19.54%	57,305	1.50%	N/A	3,817,730
2018	3,819,992	79.18%	913,321	18.93%	90,916	1.88%	N/A	4,824,229
2019	3,877,304	78.30%	965,026	19.49%	109,498	2.21%	N/A	4,951,828
2020	2,708,114	82.64%	297,575	9.08%	271,488	8.28%	N/A	3,277,177
2021	3,093,470	81.55%	631,841	16.66%	67,803	1.79%	N/A	3,793,114
2022	3,397,178	75.56%	967,677	21.52%	130,984	2.91%	N/A	4,495,839
2023	3,454,749	76.01%	1,015,057	22.33%	75,616	1.66%	N/A	4,545,422
2024	3,573,616	75.79%	1,061,589	22.52%	79,752	1.69%	N/A	4,714,957

Figure 1: US IATA Numbers
(“Bureau of Transportation Statistics”, n.d.)

timely delivery of either raw product or their finished goods. Flight delays in the United States result in significant costs to airlines, passengers and society (Rebollo and Balakrishnan, 2014)

For airlines, delays and cancellations lead to rebooking costs, compensation to passengers, lost revenue from unsold seats, and increased operational expenses due to extended crew duty hours and standby staff. In addition, airlines may incur expenses for accommodating stranded passengers in hotels that meet their standards and these includes meals and accommodation. These expenses may also come in the form of the following scenario: flight delays severely decrease customer satisfaction and may lead to customers choosing a different airport or airline, in the long term impacting the load factor of affected airlines and leading to low profits (**JOUR**). The rise in air traffic demand and the expansion of highly connected air transportation networks have led to increased congestion of busy hubs, which ultimately contribute to flight delays.

In recent decades, flight delays have become a major concern in airport management and flight scheduling, affecting the efficiency of air transportation system operations and influencing passenger choices (Li et al., 2024).The article (Zámková et al., 2022) looks into various causes of delays which include operational reasons, delays caused by passengers, delays that are caused by airport restrictions and reactionary delays which are the result of a previous delayed flight. A recent empirical study by Mayer and Sinai (Mayer and Sinai, 2003) finds that air traffic congestion due to airline hubbing and over-scheduling of flights at airport facilities are the primary causes of flight delays (Mayer and Sinai, 2003). Brueckner et al. (2019) suggest the late-arriving aircraft could be the major source of flight delays in US, outweighing other reasons like bad weather and air traffic control (Tan et al., 2021).

In the context of airline operations, a single flight delay can have a butterfly effects, disrupting schedules across an entire network. This phenomenon is known as delay propagation (Kafle and Zou, 2016).Delay propagation occurs when an initial delay in one segment of an airline’s schedule affects subsequent flights due to the interconnected nature of air transport operations as well as shared resources; mostly aircrafts.

Other ways the interconnectedness arises is in the form of reuse of aircrafts, crew, and airport resources, as well as tightly planned schedules aimed at maximizing efficiency and revenue.The desire to maximize aircraft utilization reduces the time buffer between arrivals and departures, increasing the likelihood of delay propagation (Rebollo and Balakrishnan, 2014) and therefore if these flights are delayed it impacts subsequent flights by arriving late for its next scheduled departure to other segments of its schedule.

In a network structure, where delays can have an even greater impact—without adequate slack to absorb an initial root delay, subsequent flights may also be delayed as they await aircraft and crews from the initially delayed flight (AhmadBeygi et al., 2008). However, we can see that airlines deliberately insert buffers into flight schedules and ground turnaround operations to try and mitigate the effects of these propagated delays (Kafle and Zou, 2016). The examination of flight delay and the delving into delay propagation mechanisms have long captivated scholars within the aviation domain. In comparison to delay prediction research, delay propagation inquiries may prove more captivating yet challenging, as they enable the identification of delay origins, the computation of reactionary delays and the comprehension of delay evolution.

Such understanding paves the way for implementing potent measures to counteract the cascading effect of delays (Li et al., 2024).The ultimate goal of the paper is to explore how the contribution of propagated departure delays to arrival delays varies with route, airport, and airline characteristics. This research article aims to contribute to the existing body of knowledge on flight delay propagation by exploring the following research questions:

1. **What are the primary factors influencing flight delays and its propagation within and across air transportation in a hub and spoke network?**
2. **How do different operational strategies impact the extent and severity of delay propagation?**
3. **Is there a way to classify delays into their severity of affecting the next flight leg delay and effectively find the origin of major flight delays?**

By addressing these questions, this study seeks to provide valuable insights into the dynamics of flight delay propagation and to come up with the development of evidence-based strategies to enhance the reliability and efficiency of global air transportation networks. In this research, The aim is to explore the mechanisms of flight delay propagation by analyzing data from airline operations. Specifically, I want to investigate how resource inter-dependencies, network configurations, and external disruptions contribute to the cascading effects of delays. By integrating statistical analyses and predictive modeling approaches, the aim is to provide potential actionable insights for airlines. The goal is to also identify strategies that can identify delay origins, mitigate delay propagation and improve the robustness of airline schedules, thereby enhancing the overall efficiency and reliability of air travel.

2. Literature Review on Flight Delays and Disruption Management

Reactionary delays have a large impact in the air transportation system; both at an operational and economical point of view (Ciruelos et al., 2015).The essence of this literature review is to compile the findings from multiple studies where the topic revolves around flight delays and come up up with a summary of the causes the effects and possible mitigation efforts that have been put in place to manage the delays and disruptions.

2.1. Causes of Flight Delays

Flight delay can be divided into two categories which are the root and propagated delays (Kim and Park, 2021).Research identifies multiple drivers of flight delays, ranging from operational inefficiencies to external factors. Departure delays can arise due to mechanical problems, weather delays, ground holds, and

other sources. Flights that depart on time can still be delayed in arrival, due to such things as air traffic control issues or re-routings done to avoid bad weather (AhmadBeygi et al., 2008).Delays are not only influenced by the operation of one particular flight but on previous flights. If the inbound flight arrives late, its subsequent outbound flight is also likely to depart late. This is because the incoming aircraft or crews could be used in a next departure flight (Tan et al., 2021).

Some of the causes are broken down further:

1. **Air Traffic Congestion:** Researchers like Mayer and Sinai highlight how airport congestion, driven by over-scheduling and airline hubbing, remains a leading cause of delays (Rupp, 2007)
2. **Weather Conditions:** Weather plays a substantial role, accounting for a significant percentage of delays in regions like China, where severe weather contributes to nearly half of all delays (Gui et al., 2020). From the analysis of METARS, delay and diversion incidents, and delay and diversion durations, it was found that the highest contributors are fog/mist, thunderstorm and low level clouds, followed by rain not originating from convective clouds and wind causing runway change.15(https://erepository.uonbi.ac.ke/handle/11295/164209)
3. **Airline and Airport Operations:** Factors such as crew scheduling, aircraft maintenance, and airport restrictions also significantly contribute to delays.The large number of shared resources(The connected resources can be the crew, passengers and airport resources) (Kafle and Zou, 2016) together with aircraft result in the propagation of delays through the network (Rebollo and Balakrishnan, 2014). analysis also shows that larger-capacity aircraft are particularly prone to longer delays due to operational complexities (Zámková et al., 2022)(Kim and Park, 2021).Other papers show that among all the causes, airlines' management related causes are the ones with highest contribution to the total delay (Ciruelos et al., 2015).
4. **Reactionary Delays:** Delays from one flight often propagate to subsequent flights, compounding their impact across networks. Airports like Jeju International Airport are particularly notable for delay propagation (Kim and Park, 2021). Propagated delay occurs because of connected resources involved in an initially delayed flight and flights downstream. For example, the same aircraft flies multiple flight legs in a day. Delay of an earlier flight can sustain in the subsequent flights of the same aircraft. (Kafle and Zou, 2016)
5. **Strikes:** Three types of scenarios are considered: Air traffic controllers' strikes, implemented reducing the capacity in the affected areas. Airport staff's strikes, modelled increasing the minimum turnaround time in the affected airports and Pilots' strike, implemented modifying the crew connectivity parameter. (Ciruelos et al., 2015)

2.2. Impacts of Flight Delays

The cost comes from various sources, including additional use of crew, fuel, and aircraft maintenance; increase in passenger travel time; greater environmental externalities; and the macroeconomic impact of flight delay on other economic sectors (Kafle and Zou, 2016). Delays have far-reaching consequences

1. **Economic Costs:** Delays impose substantial financial burdens on airlines due to increased operational costs and compensation claims under regulations like EU261 (Zámková et al., 2022).Flight delays in the United States result in significant costs to airlines, passengers and society. The annual cost of domestic flight delays to the US economy was estimated to be $31–40 billion in 2007 (Rebollo and Balakrishnan, 2014).The cost comes from various sources, including additional use of crew, fuel, and aircraft maintenance, increase in passenger travel time, greater environmental impacts and the macroeconomic impact of flight delay on other economic sectors. As a result of delays airlines also

incur a lot of expenses in the form of penalties, accommodation compensation and loss of potential market (Kafle and Zou, 2016).

2. **Environmental Impact:** Delayed flights increase fuel consumption and emissions, exacerbating the environmental footprint of aviation ("Delay Impacts Assessment", n.d.)(Gui et al., 2020). An example is if there is a delay caused by an ATC(Air Traffic Control) restriction might impact the flight while it is still airborne it leads to a longer than necessary flight meaning more fuel is being consumed.
3. **Passenger Experience:** Long delays lead to dissatisfaction, with cascading effects on airline reputations and passenger loyalty (Rupp, 2007)(Zámková et al., 2022). An ill experience by a customer on their flight leaves a lasting impression; especially for first-time travellers. An airline that eventually gets a lot of delays starts to get a reputation that is hard to shake off.

2.3. Strategies for Delay Mitigation

A range of solutions has been proposed to mitigate flight delays:

1. **Technological Integration:** Improved data integration and machine learning tools have shown promise in predicting delays and managing disruptions. Researchers have used various methods to analyze and model delay propagation, including statistical analysis, probabilistic models, network representation, operational research, and machine learning (Kim and Park, 2021).For instance, random forest models and LSTM networks have achieved over 90% accuracy in delay prediction, leveraging aviation big data (Gui et al., 2020)(Kim and Park, 2021).Researchers like Jetzki studied the propagation of delays in Europe, with the goal of identifying the main delay sources. Others developed a model for estimating flight departure delay distributions, and used the estimated delay information in a strategic departure delay prediction model (Rebollo and Balakrishnan, 2014). Kim (Kim and Park, 2021) highlights the use of statistical analysis, probabilistic models, network representation, operational research, and machine learning techniques to analyze and predict flight delay propagation.
   One approach to modeling delay propagation is to use the susceptible-infected-recovered (SIR) model from epidemiology. This model is based on the idea that delays can spread through a network like a disease.In the airport SIR (ASIR) model, airports are classified as susceptible, infected, or recovered (Zhang et al., 2020).
   - Susceptible airports are those that have not yet experienced a delay.
   - Infected airports are those that are currently experiencing a delay.
   - Recovered airports are those that have experienced a delay but have since recovered.

   The ASIR model can be used to simulate the spread of delays through a network and to identify factors that influence the probability of delay propagation . proposes the use of the Airport Susceptible-Infected-Recovered (ASIR) model, inspired by the epidemic spreading mechanism, to simulate delay propagation in networks, taking into account factors like airport connectivity, traffic, and turnaround efficiency. (Zhang et al., 2020) Zhang explains that the ASIR model considers the probability of delay propagation (infection rate) as a function of network configuration, airport traffic, and turnaround service level.
   Another approach is using a delay propagation multiplier is a value which when multiplied with the initial delay yields the sum of all potential downstream delays plus the initial delay. Beatty et al. constructed delay trees by considering three causes for delay propagation: aircraft equipment, cockpit crew, and flight attendants. A delay tree can contain up to 50–75 flights for a flight early in the day that is connected to the rest of the system (Kafle and Zou, 2016).

2. **Operational Optimization:** Measures like congestion-based pricing, strategic buffer times in scheduling,optimizing flight schedules, and implementing time-based air traffic management are suggested to alleviate delays at congested airports (Rupp, 2007)(Zámková et al., 2022)(Kafle and Zou, 2016).The delay that occurred in the previous flight can be mitigated at the connecting flight when B/T(Block time) and G/T(Ground Time) are configured less tightly compared to the actual period of flight and turnaround. On the contrary, if B/T and G/T for a specific route or airport are insufficient, there is an inevitable adverse delay on other flights (Kim and Park, 2021).
3. **Prioritize flights at high-traffic airports:** Airports with larger traffic volumes are more likely to experience delays, which can then propagate to other airports. Prioritizing flights at these airports can help to minimize the impact of delays. This could involve giving priority to flights that are already delayed, or to flights that are connecting to other flights (Ogunsina et al., 2021)(Zhang et al., 2020).
4. **Expand airport facilities:** Expanding airport facilities can help to increase capacity and reduce delays. This could involve building new runways or terminals, or expanding existing facilities. (Kim and Park, 2021).However, this is a long-term solution that requires significant investment and planning.

3. Methodology

This study adopts a mixed-methods approach, combining empirical analysis of real-world data with simulation modeling to investigate flight delay propagation. The methodology follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, encompassing the following stages:

3.1. Business Understanding

The objective that this paper aims to achieve is to utilize real world data and simulate and analyse flight delay propagation. Flight delays seem to be an inevitable issue that most if not all airlines run into. The delays are caused by a number of factors such as air traffic control restrictions, weather and also other delays from previous flights(Rupp, 2007)(Gui et al., 2020)(Kim and Park, 2021).These delays cause airlines billions in terms of fuel consumption, penalties and passenger compensations. The financial impact is just one of the few impacts of flight delays; others include: passenger dissatisfaction and environmental impact due to fuel emissions. Airlines need to understand how these factors and their relation can be minimized to offset the flight delays and reduce their occurrences to as low as possible.

As long as passenger numbers continue to rise, the need to develop tools that help airlines predict delays and therefore come up with disruption management systems before they happen is even more necessary. This will not only cut costs that would have been incurred by the airlines but also give them an advantage over their competitors. This paper aims to develop a data driven approach to understand and quantify the propagation effects of delays across flight networks and how they are impacting subsequent flights operated by the same aircraft or from the same hub. However in the pursuit of modeling flight delay propagation some of the challenges that have risen is data integration. Trying to consolidate data sources of weather conditions, airport operations and flight schedules seems to be a difficult task. However with the rise of technology, systems that incorporate all these reasons have become popular. These systems are able to correctly display the scheduled flight plan, the actual takeoff and landing time, turnaround time, the delay time and the reason for the delays.

3.2. Data Understanding

3.2.1. Source:

The data was collected from AIMS(Airplane Information Management System). This system is a product of Boeing and is used by airlines for various reasons such as crew planning, operations control and

commercial planning. The system is able to plan flight schedules and generate multiple scenarios in the mission of monitoring if flight schedules are well integrated. The system is also able to generate reports on OTP(On Time Performance) and this is the report that was used in this paper. The system is well integrated with other systems used in the airline industry giving it the ability to consolidate a lot of data together. The dataset collected has 50,00 0records of flight history. This is their scheduled and actual arrival and departure time, the scheduled and achieved turn around time, the delay time if any, the plate number of the aircraft and the route description which is basically the origin and destination. Some information however will need to be encoded to avoid violating the data protection act.

3.2.2. Data Variables and Types:

Var Name	**Data Type**	**Description**
Date	Datetime	This is the date the flight took place.
Origin	String	The origin of the flight in the form of IATA 3 letter city codes.
VIA	String	Where the flight did its turnaround. This is where the plane was heading to and since the flights are considered returns, they are also the origin point of the return flights.
Destination	String	This is the end of the flight plan in the form of IATA 3 letter city code.
Flight_No	String	This is the flight number of that specific route. Inbound flights always end in odd numbers while outbound flights end in even numbers.
AC	String	The aircraft type in the standard IATA aircraft type designator that are used for aircraft models.
Reg_ID	String	This is the plate number for a specific aircraft.
Sched_ARR	Time	The scheduled time the flight is supposed to arrive in the city where the turn around will happen('VIA').
Sched_DEP	Time	The scheduled time the flight is supposed to leave the VIA station.
Actual_ARR	Time	he actual time the aircraft successfully landed in the outstation('VIA').
Actual_DEP	Time	The actual time the aircraft operating that flight successfully took off from the outstation.
TAT_Min	Time	The minimum allowed time by industry standards between arrival and departure for an aircraft operating a return flight
TAT_Sched	Time	The scheduled time between arrival and departure for the aircraft.
TAT_Achieved	Time	The actual time between arrival and departure that the aircraft took to successfully turn around.
Total_NR_Delay	Time	The time in minutes and hours of the departure delays caused by various factors except by reactionary delays. These are root delays specific to only departures.
Delay_Time1	Time	The time in minutes and hours caused by the first reason for the delay.

Var Name	Data Type	Description
Delay_Reason1	String	The first reason as to why the delay occurred. These are in the form of IATA codes as acronyms of the actual delay.
Delay_Time2	Time	The time in minutes and hours caused by the second reason for the delay if any exists.
Delay_Reason2	String	The second reason as to why the delay occurred.
Delay_Time3	Time	The time in minutes and hours caused by the third reason for the delay.
Delay_Reason3	String	The third reason as to why the delay occurred.

3.3. Data Cleaning and Processing

3.3.1. Handling missing data

Due to the nature of the data obtained from the source, original elimination of rows with missing data was not considered a good idea as it meant alot of the information would be removed and therefore the results of the EDA and machine learning models would change. However, while conducting some feature engineering, missing values were checked in various ways.

1. **pd.to_numeric(...,errors='coerce')**: Converts strings to numbers, converting invalid values to 'NaN'.
2. **fillna(-1)**: Fills NaN values with a placeholder (-1) before string conversion.
3. **lambda x: x.hour * 60 + x.minute if x is not None else None)**: Handles potential None values in time calculations.

3.3.2. Data Type Conversion

Some of the data types that were changed were dates. *pd.to_datetime(...):* Converts strings to datetime objects for proper date/time handling and calculations. the reason for carrying out this cleaning process is because it allows for a wide range of powerful date and time operations. For example, we can easily calculate time differences between events, extract specific components like day of the week or month, and perform time-based aggregations. This is particularly useful for time series analysis, where we can identify trends, seasonality, and other temporal patterns in the data.Some of the time columns were also changed from strings to time/datetime to be able to carry out analysis on them.

3.3.3. Column splitting and renaming

Some of the columns were renamed for clarity purposes and as the original one came with came with some columns and rows merged, further action was needed to split them to generate more descriptive visualizations and conduct better performing models. For instance, the "Origin/VIA/Destination" column was split into separate columns for "Origin," "VIA," and "Destination" to facilitate easier analysis. Additionally, the delay breakdown column was split into multiple columns to capture different delay reasons and their corresponding times. Splitting strings helps to extract meaningful features from the data, such as flight numbers, origin, and destination airports.This step is crucial as it allows for a more detailed analysis of delay causes, which is essential for predictive modeling.

3.4. Exploratory Data Analytics (EDA)

Exploratory Data Analytics (EDA) is essential to understand the underlying patterns and relationships within the dataset. Initial EDA revealed that the dataset contains flight information over several months,

with multiple flights per day.Alot of these flight had their destinations(in this case VIA as Destination is the same as the Origin in round trips) as well as origins in Nairobi (NBO) which is easily explained as the hub for the airline is in Nairobi. Another observation seen from univariate analysis is that the main cause of delays as seen in delay reason 1 column is RF(Crew Alignment) The delay times and reasons were analyzed to identify common patterns. For instance, certain airports or routes may have higher delays due to weather conditions or operational issues. In bivariate analysis, we can see that alot of these trends occur between the months of March and May. Further investigation is needed to asses why the spike in delays occured around these months. We can use boxplots to also see that although the distribution of the TAT_Achieved(turn around time achieved by the ground handling crew) is consistent there are still some outliers. Outliers can also be seen in the total delays recorded. Visualizations such as histograms, box plots, and scatter plots were used to identify trends and outliers. This step is critical for identifying potential features that could influence the predictive model.

3.5. Feature Engineering/Selection

Feature engineering is the process of developing additional features from an existing dataset to improve model performance. This stage is critical since it minimizes the dimensionality of the dataset and concentrates on the most important features, boosting model efficiency and accuracy. In this working, features like ”month” and ”is weekend” have been calculated by extracting the date values and checking the against a calender. More columns such as ' Origin', 'VIA', and the delay reasons were extracted by splitting their original columns that had merged these values.

Random Forest was used to determine each feature's relevance in predicting the target variable. The Random Forest method gives significance scores to each feature based on how much it contributes to lowering the impurity of the tree nodes. Features with higher significance ratings are deemed more significant for predicting the target variable. Using the feature importance ratings, the top 12 most significant variables were selected for the model. They assisted with:

1. Reduced Noise: By reducing irrelevant or redundant features from the data, we may make the models more accurate and reliable.
2. Improved Model Interpretability: With fewer characteristics, the model becomes easier to grasp and interpret
3. Reduced Overfitting: Overfitting leads to poor performance on new, untested data. Feature selection helps to reduce overfitting by limiting the amount of parameters the model must learn.
4. Faster Training and Inference: Using fewer features results in faster training and inference times, particularly for large datasets and complicated models.
5. Improved Generalization: By focusing on the most important properties, the model is more likely to generalize effectively to new, previously unknown data, resulting in greater predicted performance.

3.6. Machine Learning Modelling

Data Normalization/Scaling: Standard Scaler was used to normalize numerical features by eliminating the mean and scaling them to unit variance. This guarantees that all features provide an equal contribution to the model, preventing larger scale characteristics from dominating the learning process. This results in more accurate and fair model projections. It can also help to speed up the convergence of some optimization procedures. This step is essential for algorithms that are sensitive to the scale of the data.

Data Splitting: The data was divided into training and testing sets using train_test_split at an 80/20 ratio (80% for training and 20% for testing). This ensures that the model is trained on a large percentage of the

data while maintaining a separate set for performance evaluation. Cross-validation was also used to ensure that the model can generalize effectively to new data. Splitting the data into training and testing sets enables an unbiased evaluation of the model's performance on previously unknown data. This helps to prevent overfitting and guarantees that the model can generalize effectively to fresh data.

Algorithms:

- **Random Forest Regressor:** A potent ensemble learning technique that uses several decision trees to generate predictions is the Random Forest Regressor. A varied ensemble of models is produced by training each decision tree in the forest on a random subset of the data and characteristics. In order to produce a more reliable and precise prediction, the Random Forest averages the predictions of each individual tree. This ensemble method enhances generalization performance and lessens overfitting. Because Random Forest can efficiently handle high-dimensional data, it is especially well-suited for huge datasets with plenty of features. It is a strong and adaptable algorithm because it is also comparatively insensitive to outliers and missing variables. It was selected due to its resilience and capacity to manage huge, highly dimensional datasets. parameters like n_estimators (number of trees) and max_depth (maximum depth of each tree) can be tuned for better performance.

- **HistGradientBoostingRegressor:** This is a gradient boosting algorithm that builds an additive model in a forward stagewise fashion. It iteratively adds trees to the model, focusing on correcting the errors made by previous trees. This approach allows the model to capture complex patterns in the data and achieve high accuracy. HistGradientBoostingRegressor is known for its robustness to outliers and missing values, and it can handle both numerical and categorical features effectively.parameters like max_iter (maximum number of iterations) and learning_rate can be adjusted.

- **Kmeans Clustering:**This was used in the source to perform classification of delay-prone aircraft. Specifically, the analysis grouped aircraft by aircraft registration ID and total delay minutes, then used K-means clustering to identify distinct clusters of aircraft based on their delay characteristics.By distinguishing these clusters, airlines can develop proactive strategies, such as optimizing maintenance schedules, improving turnaround processes, or adjusting flight assignments to minimize overall delays.

3.7. Performance Evaluation

Mean Squared Error (MSE): Measures the average squared difference between the predicted and actual values. MSE is a common metric for regression problems, but it can be sensitive to outliers. It is often used in conjunction with other metrics, such as RMSE and R-squared, to provide a more comprehensive evaluation of the model's performance.

R-Squared: Also known as the coefficient of determination, is a statistical measure that evaluates how well a regression model explains the variance of the dependent variable. It represents the proportion of the total variability in the dependent variable that can be attributed to the independent variables. R-Squared value ranges from 0 to 1, where 0 means the model explains none of the variability, and 1 means it explains all of it.

3.8. Model Optimization

Hyperparameter Tuning: parameters such as the number of trees, maximum depth, and minimum samples per leaf were tuned to improve performance. This step is crucial for maximizing the model's predictive accuracy and ensuring it generalizes well to new data.

4. Results

The Results section presents the results objectively based on the methods used to analyze the dataset. The findings are organized logically and supplemented by illustrative materials such as tables and figures. The findings are summarized below, with a focus on major patterns, trends, and statistical analysis.

4.1. Data Preparation and Cleaning

The baseline dataset was 50,057 rows and 18 columns, with some missing values and unstructured data. After cleaning, the dataset was reduced to 50,049 rows, with extraneous columns deleted, including "Unnamed: 2," "Unnamed: 3," and "Unnamed 4" The "Origin/VIA/Destination" column was divided into three columns: "Origin," "VIA," and "Destination," allowing for a more detailed analysis of flight routes. Furthermore, the "NR_breakdown" column, which comprised delay reasons and times, was divided into several columns to capture individual delay reasons and their related periods. This preprocessing phase ensured that the dataset was properly formatted and prepared for analysis.

4.2. Explaratory Data Analysis(EDA

Exploratory Data Analysis revealed several key trends and patterns in the dataset. The most frequent delay reasons included "RF", "RE", and "RG" . The latter two reasons seem to have the same frequency of occurrence. Further analysis onto their actual average delays would indicate which of the two causes a significantly higher delay than the other. Additionally, certain airports, such as NBO (Nairobi) and ABJ(Abidjan), experienced more traffic compared to others. The NBO airport traffic could be explained by the fact that NBO is the hub of the airport. ABJ, MBA (Mombasa) and CPT(Cape Town) seem to have also been popular destinations during this time.

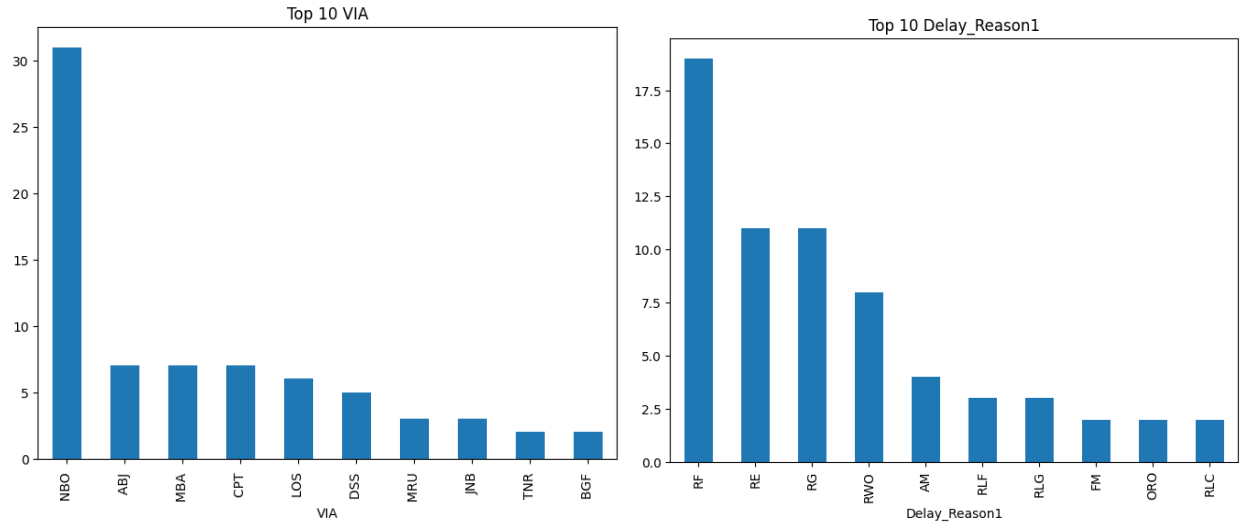


Figure 2: Top 10 Airports and Delay Reasons

Trends of key variables which are total delays and achived TATs show spikes around the month of August indicating that the delays are usually high around the summer peak and so are the minutes taken to turn the aircraft around. Most of the delays and even the Turn Around Time for the aircrafts are concentrated in a specific range with occasional outliers. This could be explained by the regulations set in place by their authorizing body such as IATA and KCAA that ensures certain aircraft have a minimum time they are

allowed to take between one flight and the subsequent one ensuring the aircraft is safe for takeoff and flying to its intended destination.

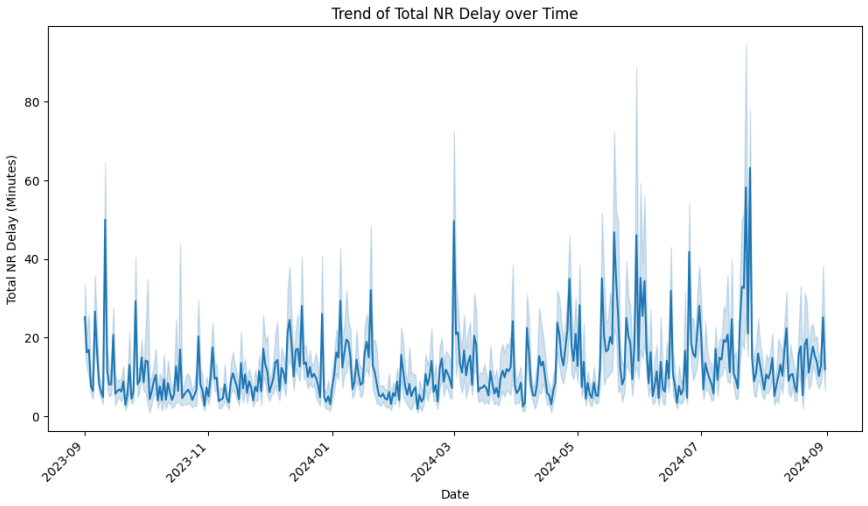


Figure 3: Total Delay Trend

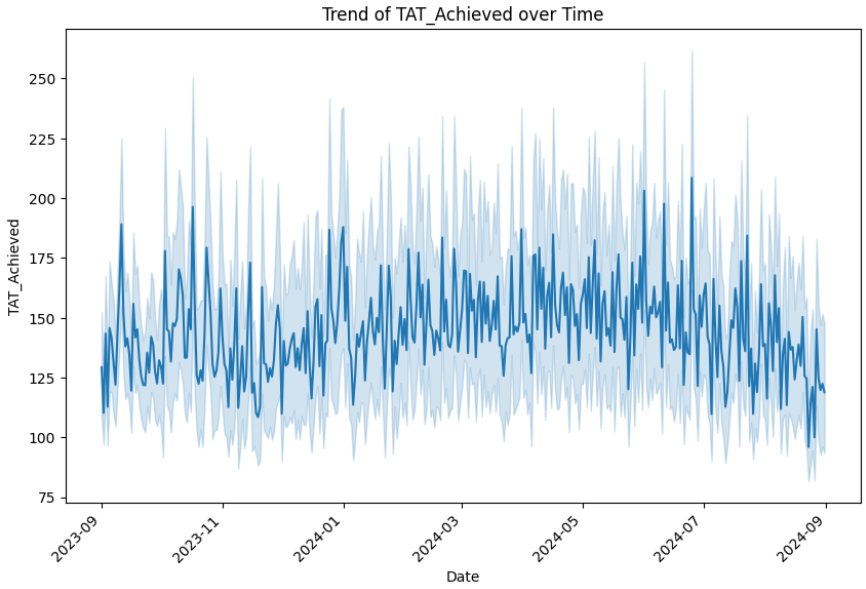


Figure 4: Total TAT Achieved over time

Pairplot of key variables revealed a strong positive correlation between scheduled Turnaround and Achieved Turnaround, indicating that most flights achieved their assigned time to turn the aircraft around. There is a strong linear relationship between TAT_Achieved_Minutes and TAT_Sched_Minutes. TAT_Min_Minutes seems to have distinct values, indicating predefined operational constraints. The relationship between Total_NR_Delay_Minutes and other variables appears non-linear, suggesting other factors influence delays. An interesting observation is also that the flights with little turnaround time scheduled or achieved recorded the most total delays.

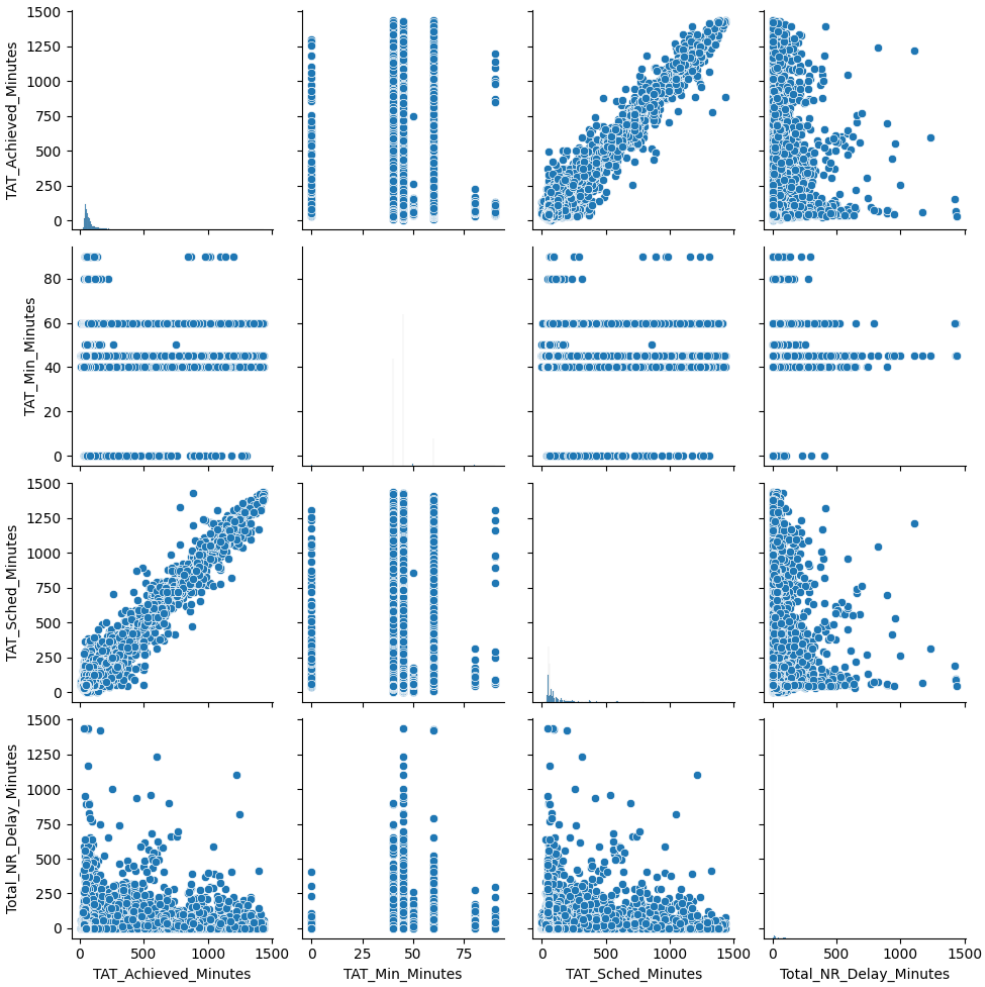


Figure 5: Pairplot of some key variables

4.3. Feature Correlation

The correlation matrix revealed significant relationships between Total_NR_Delay and factors such as the date and the flight number. This suggests that some delays were definitely caused by the weather which went on to affect more than a few flights. Another correlation to note is the aircraft type and the delay reason 1. This shows that some aircraft types might have the same reason as to why they would be delayed which is amplified by the scheduled urn around time which is constant for all aircrafts of the same type. The date, flight number, origin and destination will definitely have some significant correlation as they operate on a set schedule with predefined route descriptions. Although the term significant is relative as the highest correlation that can be seen is of 0.43 for positive correlation and -0.4 for negative correlation.

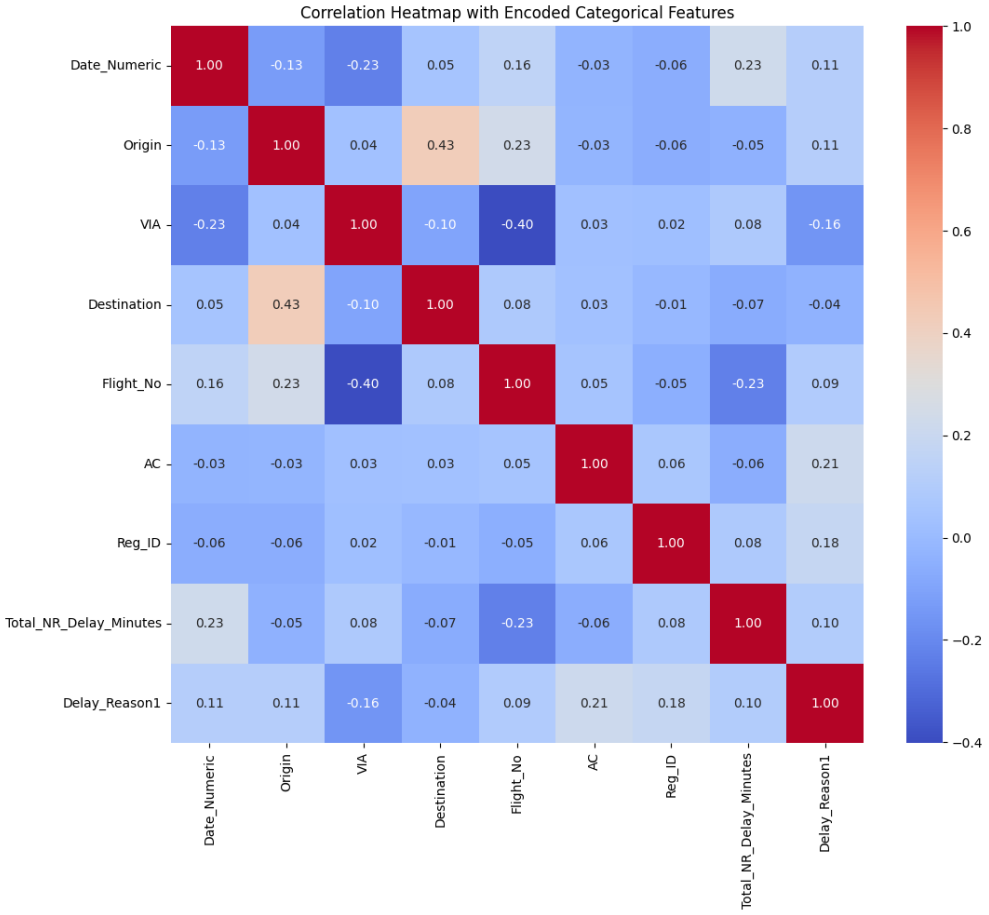


Figure 6: Correlation Heatmap

4.4. Machine Learning and Model Optimization

Three machine learning models—Random Forest, Gradient Boosting, and KMeans—were trained and evaluated on the dataset. The models were configured with default parameters initially, and their performance was evaluated using metrics such as accuracy. Curating the perfect features for the models (especially Random Forest) performed differently leading to an assumption that ”Red_ID” which is the plate number of the aircraft should be included to generate more accurate results.

Based on the likelihood of delays, flights were divided into discrete groups using K-Means clustering. According to the data, which are displayed below, flights in Cluster 3 are most likely to experience large average delays. These flights' effective recovery time most likely played a role in their placement into Cluster 3. The turnaround time (TAT) for a later flight is frequently shortened to make up for a higher-than-normal delay. This modification preserves overall scheduling efficiency and lessens cascading delays.

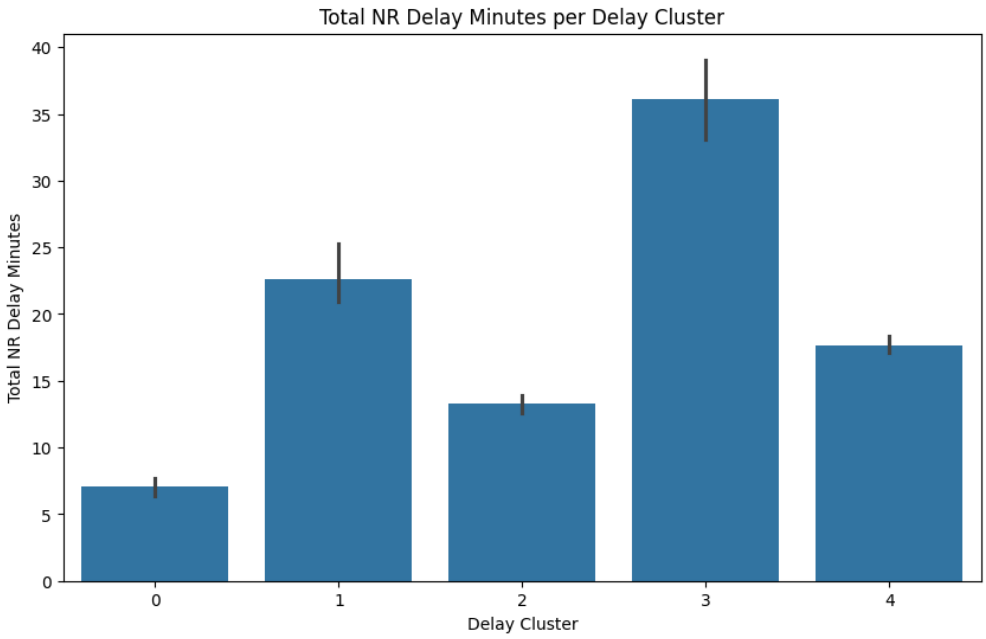


Figure 7: Histgradient Boosting Predictions

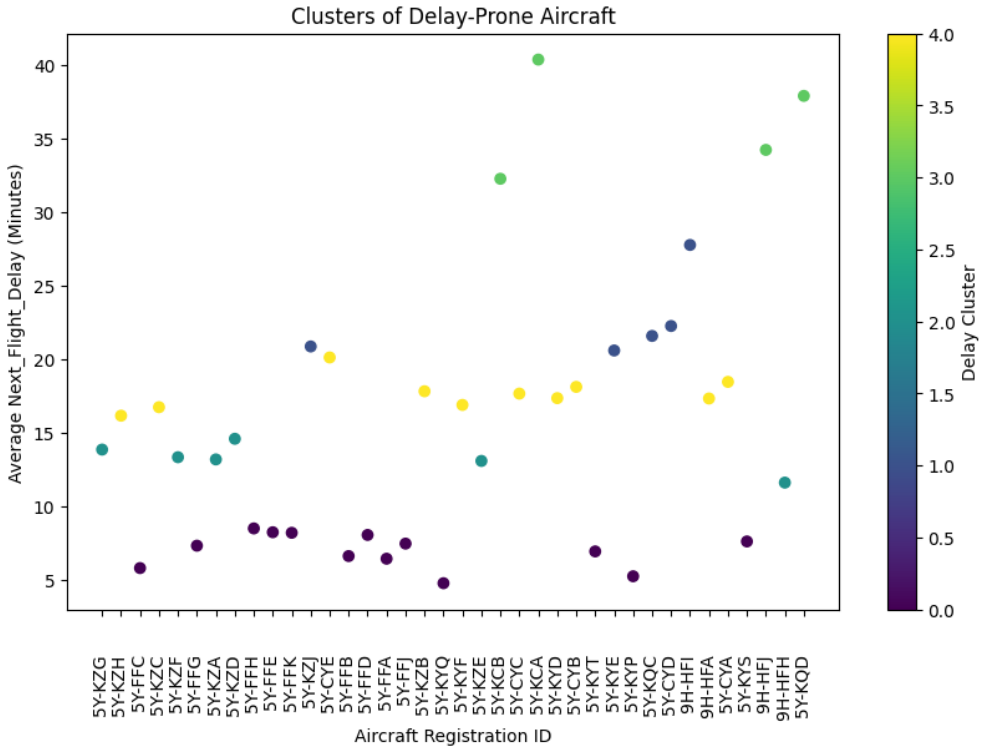


Figure 8: Delay Prone Aircraft Clustering

Performance metrics of different regression models (Random Forest, HistGradientBoostingRegressor) on the test set show that the Random Forest where the n_estimators were slightly adjusted achieved the best performance with. Parameters such as the number of estimators, maximum depth, and minimum samples per split were tuned for both the models to achieve optimal results. The optimized model achieved a 3% improvement in R^2 compared to the default configuration.

Model	R-Squared
Random Forest Regressor	0.993
Random Forest Regressor+Clusters	0.9804
Hist Gradient Boosting Regressor	0.9399

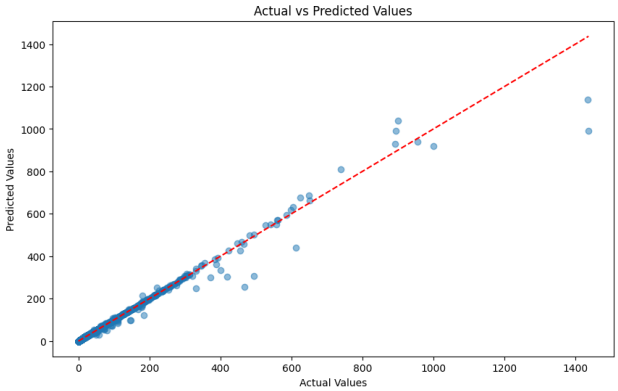


Figure 9: Random Forest Predictions

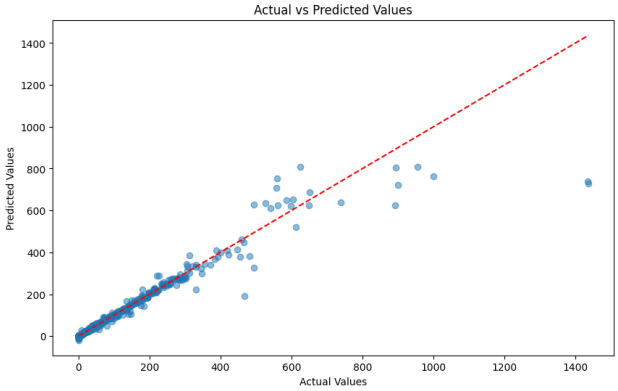


Figure 10: Histgradient Boosting Predictions

5. Discussion

The results from the data modeling and visualization provide critical insights into the dynamics of flight delays, their underlying causes, and their potential propagation effects. A key finding is Nairobi's Jomo Kenyatta International Airport (NBO) as the central hub in a "hub and spoke" network structure.

This model, where flights originate from and return to a central hub while connecting to various spoke destinations, can concentrate delays at the hub (Baumgarten et al., 2014). For instance, delays at NBO can ripple through the network, affecting multiple flights and amplifying disruptions across the system. This highlights the vulnerability of such structures to systemic delays and underscores the need for targeted operational improvements at hub airports to mitigate cascading effects (Grove and O'Kelly, 2005).

The analysis also reveals the significant role of aircraft models and operational patterns in delay dynamics. The Embraer E90, the most frequently used aircraft in the dataset, is predominantly deployed for short-haul flights, which are more susceptible to delays due to factors like tight turnaround times and airport congestion. Additionally, specific plate numbers such as 5Y-KYP and 5Y-FF, often associated with E90s, suggest that maintenance schedules, operational routes, or other aircraft-specific factors may contribute to delays. Furthermore, reactionary delays—caused by issues such as baggage handling, check-in processes, and crew availability—are a major bottleneck. These delays often trigger a chain reaction, where a disruption in one area leads to delays in subsequent flights, emphasizing the interconnected nature of airport operations.

Seasonal trends also play a significant role in flight delays, with peaks observed during August (summer) and December (winter). These periods coincide with higher passenger volumes and adverse weather conditions, which strain airport resources and increase the likelihood of delays (Wu et al., 2019). We can also see that most delays are relatively short (under 50 minutes). Additionally, the strong linear relationship between achieved and scheduled turnaround times (TAT) shows the cres dedication to adhering to TAT schedules for maintaining on-time performance. However, the non-linear relationship between total delay minutes and other variables suggests that delays are influenced by a complex interplay of factors, including but not limited to weather, equipment issues, and air traffic control, rather than timing alone.

From a modeling perspective, the findings highlight the cascading nature of delays, with a strong linear relationship between the propagation ratio and subsequent flight delays. This indicates that initial delays often lead to a domino effect, disrupting connected flights. The high R2 value (0.993) for flight clustering and delay prediction demonstrates the model's accuracy in capturing data variance(though it is essential to interpret this alongside other metrics to avoid overfitting). Key features such as Delay Time and Propagation Ratio were identified as critical drivers of delays. Moreover, classifying aircraft based on delay proneness enables targeted interventions to address issues specific to delay-prone aircraft. The Random Forest Regressor's strong performance, with an $R\hat{2}$ score of 0.980, further validates the model's predictive capability, making it a valuable tool for anticipating and mitigating flight delays. Together, these findings offer a comprehensive understanding of delay patterns and provide a foundation for developing more efficient and resilient aviation operations.

5.1. Future Works

- **Expanding the Dataset:** The current study focused on data from a single airline. Future work could include data from multiple airlines and different regions to understand how delay propagation varies across different operational contexts especislly in other types of network configurations used by other airlines.The data will also extend the dataset to cover a longer time period to capture seasonal variations and long-term trends in flight delays.

- **Enhancing Predictive Models:** Explore other machine learning models e.g., LSTM networks to improve the accuracy of delay predictions and to handle more robust data. Another aim is to develop models that can predict delays in real-time, incorporating live data feeds from air traffic control, weather services, and airline operations.

- **Incorporating External Factors:** Integrate more detailed weather data to better understand its impact on delays and improve predictive accuracy. Include data on air traffic control restrictions and congestion to better model delays caused by these factors.
- **Technological Innovations:** Explore the use of automation and AI in various aspects of airline operations, from scheduling to real-time decision-making, to reduce delays.

6. Summary and conclusions

This study underscores the multifaceted nature of flight delays, which are influenced by a combination of operational inefficiencies, seasonal travel patterns, and aircraft-specific factors. Using the latest methods in machine learning, this study examined the factors causing aircraft delays and found that turnaround time, the causes of delays, and flight-specific characteristics are important factors in determining how long delays last and their cascading effect to the subsequent flights. Outperforming previous models and proving the usefulness of machine learning in resolving operational inefficiencies, the RandomForest Regressor model became the most successful predictive analytics tool in aviation operations.

We have seen that the "hub and spoke" network structure, while efficient for connectivity, concentrates delays at central hubs thus amplifying their impact across the network. Reactionary delays and seasonal peaks further strain airport resources, highlighting the need for robust operational planning and resource allocation. The propagation of delays, as demonstrated by the strong linear relationship between initial and subsequent delays, emphasizes the importance of proactive measures to prevent cascading disruptions.These results give airlines useful information that they may use to improve scheduling, better allocate resources, and lessen the ripple effects of delays throughout their networks.

This discovery has effects that go beyond short-term operational enhancements. Airlines can employ focused tactics to reduce delays, increase punctuality, and boost passenger satisfaction by using clustering and predictive modeling to identify and rank high-risk flights. By guaranteeing scalability and user-friendliness for all parties involved, the incorporation of these models into current airline management systems provides a viable route for practical implementation.

In order to facilitate dynamic scheduling adjustments, future research should concentrate on integrating real-time data streams into prediction models. This strategy would further improve operational resilience and reliability by enabling airlines to react proactively to changing circumstances. In addition to tackling the urgent problem of flight delays, this study opens the door for a more effective, data-driven future in aviation by expanding the application of predictive analytics in the industry.

References

AhmadBeygi, S., Cohn, A., Guan, Y., & Belobaba, P. (2008). Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, *14*(5), 221–236. https://doi.org/https://doi.org/10.1016/j.jairtraman.2008.04.010

Baumgarten, P., Malina, R., & Lange, A. (2014). The impact of hubbing concentration on flight delays within airline networks: An empirical analysis of the us domestic market. *Transportation Research Part E: Logistics and Transportation Review*, *66*, 103–114. https://doi.org/10.1016/j.tre.2014.03.007

Bureau of transportation statistics. (n.d.). https://www.transtats.bts.gov/homedrillchart.asp

Ciruelos, C., Arranz, A., Etxebarria, I., Peces, S., Campanelli, B., Fleurquin, P., Eguíluz, V., & Ramasco, J. J. (2015). Modelling delay propagation trees for scheduled flights.

Delay impacts assessment. (n.d.). https://assets.publishing.service.gov.uk/media/5a7cfa0040f0b60aaa29371a/AC08_tagged.pdf

Grove, P., & O'Kelly, M. (2005). Hub networks and simulated schedule delay. *Papers in Regional Science*, *59*, 103–119. https://doi.org/10.1111/j.1435-5597.1986.tb00985.x

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, *69*(1), 140–150. https://doi.org/10.1109/TVT.2019.2954094

Iata kenya report. (n.d.). https://www.iata.org/contentassets/0fc44e59164d44579f17356da0cc98fd/iata_kenya_report.pdf

Kafle, N., & Zou, B. (2016). Modeling flight delay propagation: A new analytical-econometric approach. *Transportation Research Part B: Methodological*, *93*, 520–542. https://doi.org/https://doi.org/10.1016/j.trb.2016.08.012

Kim, M., & Park, S. (2021). Airport and route classification by modelling flight delay propagation. *Journal of Air Transport Management*, *93*, 102045. https://doi.org/https://doi.org/10.1016/j.jairtraman.2021.102045

Lesgourgues, A., & Malavolti, E. (2023). Social cost of airline delays: Assessment by the use of revenue management data. *Transportation Research Part A: Policy and Practice*, *170*, 103613. https://doi.org/https://doi.org/10.1016/j.tra.2023.103613

Li, C., Mao, J., Li, L., Wu, J., Zhang, L., Zhu, J., & Pan, Z. (2024). Flight delay propagation modeling: Data, methods, and future opportunities. *Transportation Research Part E: Logistics and Transportation Review*, *185*, 103525. https://doi.org/https://doi.org/10.1016/j.tre.2024.103525

Mayer, C., & Sinai, T. (2003). Network effects, congestion externalities, and air traffic delays: Or why not all delays are evil. *American Economic Review*, *93*(4), 1194–1215. https://doi.org/10.1257/000282803769206269

Ogunsina, K., Bilionis, I., & DeLaurentis, D. (2021). Exploratory data analysis for airline disruption management. *Machine Learning With Applications*. https://doi.org/10.48550/arXiv.2102.03711

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, *44*, 231–241. https://doi.org/https://doi.org/10.1016/j.trc.2014.04.007

Rupp, N. (2007). Further investigations into the causes of flight delays.

Tan, X., Jia, R., Yan, J., Wang, K., & Bian, L. (2021). An exploratory analysis of flight delay propagation in china. *Journal of Air Transport Management*, *92*, 102025. https://doi.org/https://doi.org/10.1016/j.jairtraman.2021.102025

Tchouamou Njoya, E., Semeyutin, A., & Hubbard, N. (2020). Effects of enhanced air connectivity on the kenyan tourism industry and their likely welfare implications. *Tourism Management*, *78*. https://doi.org/10.1016/j.tourman.2019.104033

Wu, W., Zhang, H., Feng, T., & Witlox, F. (2019). A network modelling approach to flight delay propagation: Some empirical evidence from china. *Sustainability*, *11*(16). https://doi.org/10.3390/su11164408

Zámková, M., Rojík, S., Prokop, M., & Stolín, R. (2022). Factors affecting the international flight delays and their impact on airline operation and management and passenger compensations fees in air transport industry: Case study of a selected airlines in europe. *Sustainability*, *14*, 14763. https://doi.org/10.3390/su142214763

Zhang, H., Wu, W., Zhang, S., & Witlox, F. (2020). Simulation analysis on flight delay propagation under different network configurations. *IEEE Access*, *8*, 103236–103244. https://doi.org/10.1109/ACCESS.2020.2999098