Marvin Harrichrran

Econometrics

Prof. Foster

12/19/2024

Impact of Airport Congestion, Weather, and Operational Inefficiencies on Flight Delays

Introduction

The aviation industry is a robust and complicated sector that contributes significantly to the economy every year. This industry operates on an expansive and intricate scale, with travelers in the United States contributing substantial revenue daily. The biggest components of the industry are the carriers and the airports that support them. While passengers rely on these carriers every day, the issue of delays frequently arises. Flight delays have far-reaching consequences, impacting passengers, airlines, and airports worldwide. Understanding the factors contributing to delays is critical for optimizing operations, reducing airport congestion, and enhancing customer satisfaction.

This project investigates whether airport congestion, weather, and operational issues significantly affect flight delays. The hypothesis is tested using various statistical and machine-learning methods, including regression models, interaction effects, and classification algorithms. Projected graphical visualizations will further reveal insights from the analysis. Additionally, with improved machine learning testing, this research can serve as a guide for decision-making and improving efficiency in managing flight operations for carriers.

Background:

The aviation industry is a complex venture that involves many moving components that all work together to achieve the goal of on-time performance. The industry is a significant contributor to the United States' gross domestic product. In 2019, the industry contributed $852.3 billion in direct economic activity, as reported by the FAA, which accounted for 2% of the U.S. economy. Carriers, otherwise known as airlines, contribute to this industry with their operational models of hub-and-spoke and point-to-point. These operations build a complex network in the National Airspace System (NAS), facilitating the movement of thousands of passengers every day to and from their destinations.

With this complexity, there may arise many interruptions that can hinder the performance of carriers in the industry. These effects on performance are caused by flight delays in the United States, which have become an increasing issue. The rate of flight delays in the U.S. has steadily risen from 5.2% in 2018 to 7.6% in 2023. The causes of this can be attributed to various factors, including operational issues, which the carriers can control, and airport congestion, which is influenced by the planning of certain flight times. Additionally, weather coupled with other factors together to caused the increase in flight delays.

Literature Review:

Flight delays have been a persistent challenge, impacting customer satisfaction, operational efficiency, and economic performance. Research into the causes of flight delays provides an investigative understanding of the possible challenges and root causes of this issue. The most well-known causes of flight delays are weather, congestion, and operational inefficiencies. When researching the causation of flight delays, recent data and studies provide a detailed understanding of the topic and shed light on the multifaceted nature of the problem.

1. Impact of Weather on Flight Delays

Aircraft are capable of taking off and landing in many flight weather conditions; however, these conditions can be viewed as high risk or low risk. Regardless of the risk factor, adverse weather conditions are among the most significant contributors to flight delays. At the time a flight departs the airport, the possibility of a delay is based on weather reports provided to the airport before departure, this relays the pre-existing nature of the weather. A small shower or rain reported miles away at the time of departure can escalate into a major thunderstorm, resulting in aircraft diversion and subsequent delay in arrival.

According to data reported by the FAA, 74.26 percent of system-impacting delays are caused by weather. In 2019, it was reported that a total of 784 delays were caused by weather. (FAA) Delays caused by weather are particularly challenging to mitigate due to their unpredictable nature and wide-reaching impacts on aviation.

2. Congestion and Capacity Constraints

Airport congestion has long been recognized as a critical factor in flight delays, particularly at major hub airports. The size of an airport and its operational capacity contribute to the factors behind flight delays. An airport's capacity, measured by the number of passengers commuting during its operational hours, has a significant impact on delays. Seasonal periods are attributed to increases in passenger traffic, with many flights becoming concentrated

during travel seasons. The scheduling of flights by carriers is another contributing factor. Many airports operate at peak efficiency during certain periods, and numerous flights are scheduled simultaneously. Most flights begin departures and arrivals during peak times, particularly in the morning and afternoon. High passenger volumes and overbooked flight schedules exacerbate delays, especially during peak travel periods. Studies found that airline hubs, particularly larger hubs, experience longer flight delays. Flights departing from an airline's hub are prone to longer delays due to the peaking of hub airline flights (Rupp, 2007, p. 26). It was reported by the FAA that flight delays cost airlines over $3 billion per year (Rupp, 2007, p. 2). This congestion is compounded by infrastructural limitations, including insufficient runways and taxiways, which reduce the capacity of airports to handle deviations from normal operations and are a pivotal cause of prolonged delays.

3. Airline Operations

Airlines operate on a 24-hour schedule, with employees continuously working to ensure efficient operations. The variety of employees that airlines rely on includes flight attendants and pilots, who must arrive on time for their flights, prepare flight plans, and fulfill their duties. Catering to replenish food for passengers, while cabin service cleaners ensure the planes are properly cleaned and sanitized before each flight. The ground service employees are responsible for loading customer baggage promptly, and maintenance teams ensure each aircraft is airworthy to minimize customer impact.

These employees play a crucial role in ensuring flights take off and land on time. However, there are instances where these employees contribute to flight delays. Such delays can result from slow baggage loading, late pilots, or maintenance issues. These disruptions can affect the timing of a plane's departure and impact the airline's operational efficiency, potentially creating residual effects on subsequent flights.

Airlines do anticipate such interruptions and allocate grace periods for employees working on a flight to minimize delays. As stated, "Airline companies construct their monthly or weekly plans and schedules by assuming no disruption will occur. However, there are plenty of incidences that lead to disruptions during the execution of that plan." (Erdem & Bilgiç, 2024). This reveals that despite careful planning for unexpected errors, the possibility of delays remains. The reliance on pre-scheduled operations for employees can lead to inefficiencies when unforeseen disruptions occur.

4. Economic and Customer Implications

Flight delays have a far-reaching impact on the economy for both airlines and passengers. Research and

studies reveal that flight delays lead to increased fuel costs, compensation claims, and lost revenue for airlines. Every year, airlines lose billions of dollars due to flight delays. These losses in the billions result in additional expenses for the airlines. These additional expenses impact customer satisfaction and airline revenue, which in turn affects the aviation industry's economic performance. Passengers who face inconveniences like missed connections or lost baggage become dissatisfied, blaming the airlines as the cause of these issues but are often unaware of the deeper underlying reasons. This dissatisfaction often results in passengers leaving negative reviews about their experiences, which impacts the airline's performance in gaining revenue from future customers. Airlines prioritize mitigating flight delays because they do not want these negative impacts.

A study on the impacts of flight delays by the National Center of Excellence for Aviation Operations Research (NEXTOR) in 2007 provided results that showed flight delays cost airlines $8.3 billion. (Peterson, 2013). This cost has only grown as the industry has expanded in recent years, and given the rate of inflation, that number would now be approximately $16.3 billion. Additionally, this study revealed that due to extended passenger time "lost due to schedule buffer, delayed flights, flight cancellations, and missed connections, NEXTOR estimated a $16.7 billion cost to passengers in 2007" (Peterson, 2013). This increase in costs for airlines and passenger dissatisfaction results in a significant economic impact because it damages the airline's reputation and customer satisfaction. If there is a possibility to reduce the effects of flight delays, then the U.S. net welfare can increase by $2.36 billion, and this would be caused by a 10% reduction in flight delays (Peterson, 2013). This would improve the economics of the industry, result in cost savings for the airlines, and improve productivity for passengers.

Data Analysis:

Accessing the dataset from the Bureau of Transportation Statistics (BTS) on Airline On-Time Statistics and Delay Causes is crucial for investigating the primary causes of flight delays and analyzing their impacts. This analysis aims to examine the leading causes of flight delays while considering potential contributing factors. The BTS data will be used to apply and test both the null and alternate hypotheses, utilizing regression models and machine learning techniques to provide visualizations and insights. The data, covering flight delays from February 2018 to February 2023, will help predict potential causes and relay visual statistics and models to determine the relevance of the hypotheses. Additionally, the analysis will focus on the impacts of delays on airports, passengers, airlines, and the aviation industry, particularly during the pandemic.

Hypothesis:

- **Null Hypothesis (H_0):** Airport congestion, weather, and operational issues have no significant effect on flight delays.
- **Alternative Hypothesis (H_1):** Airport congestion, weather, and operational issues significantly affect flight delays.

The dataset taken from the BTS contains records of flight delays across various U.S. airports. In the dataset, there were a total of 120,188 observations made of a total of 21 variables. These variables were broken down into multiple categories, some of which included records segmented into delay type classifications such as carrier, weather, NAS, security, and late aircraft delays. The analysis will focus on these key variables, including:

- **Total delay**: This is the sum of all delay types from the dataset.
- **Arrivals (congestion proxy)**: arr_flights. This is the calculation of the arrival flights that result in a delay.
- **Weather events**: weather_ct. This is the variable type that shows the results of delays caused by weather
- **Operational issues**: nas_ct and late_aircraft_ct.These are the variable type that shows the results of delays caused by the operational issues of the carrier.

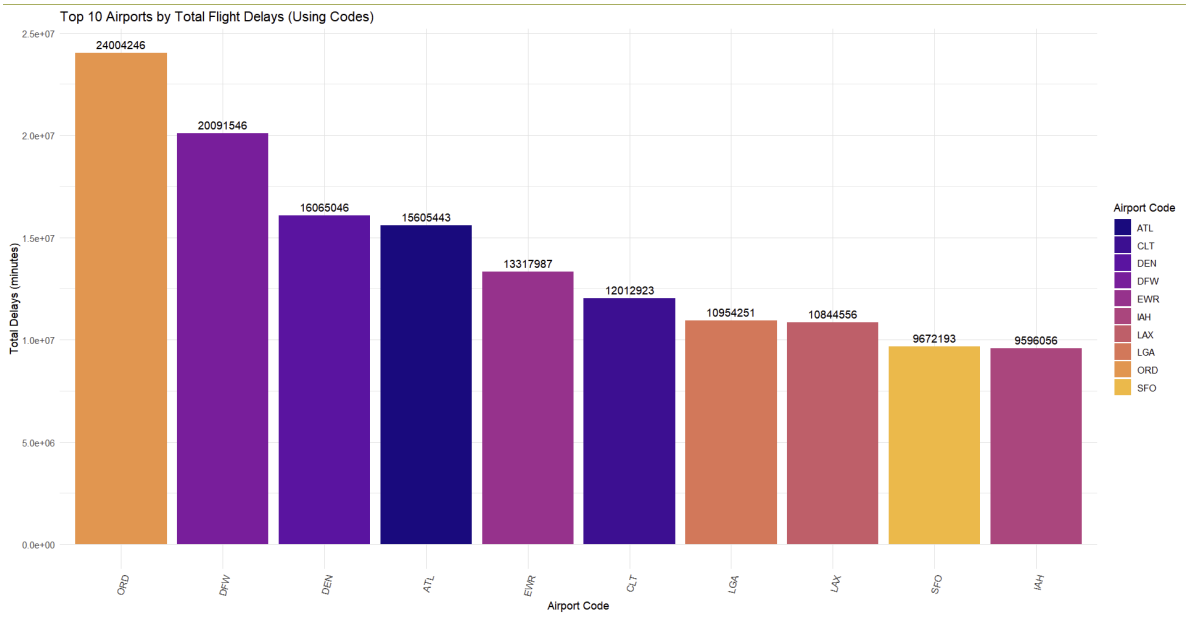
The analysis will factor these variables into two categories: "High Delay" and "Low Delay." A high delay is classified as a total delay greater than 1000 minutes, while a low delay is classified as a total delay of 1000 minutes or less. This classification is reflected in the code.

```
Delays <- Delays %>%
  mutate(
    total_delay = carrier_delay + weather_delay + nas_delay + security_delay + late_aircraft_delay,
    delay_category = ifelse(total_delay > 1000, "High Delay", "Low Delay")
  ) %>%
  drop_na(total_delay, arr_flights, weather_ct, nas_ct, late_aircraft_ct)
```

The purpose of categorizing the delays into two categories was to simplify the identification of what may cause a delay. A high delay indicates significant operational issues, severe weather, or extended aircraft travel times that can lead to extensive disruptions. Conversely, a low delay represents minor delays, possibly attributed to brief, less severe weather disturbances, and minor technical issues, many of which are short-lived disruptions.

Examining the rate at which flight delays have increased over the years, key findings in this analysis revealed patterns across airports, seasons, and delay types. The average delays varied significantly by airport, with many delays originating from hubs of major airlines. These delays often result from airport congestion, as many of these hubs are major providers of both international and domestic travel for large airlines. This analysis focused on identifying which major airports exhibited the most significant delays.

Based on the total observations, the delays were filtered and mutated to summarize the total delays at the top ten airports in the United States. Using the code to filter the airports with the highest delays, a bar graph was created to reflect these results. The bar graph highlights the airports with the highest delays, providing insights into patterns and trends that indicate significant congestion and operational challenges at these locations.



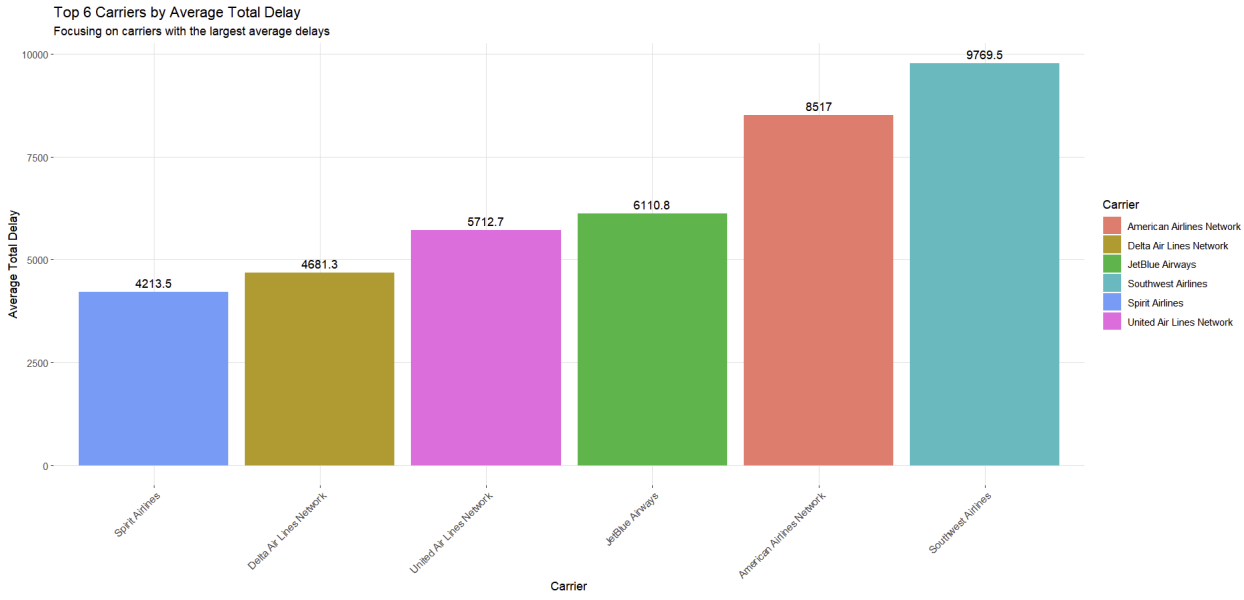
The graph highlights the total flight delays at the top ten airports, measured in millions of minutes. It reveals significant patterns and trends, showing that many of these airports recording the most delays are major hubs known for high volumes of passenger traffic and congestion. Airports such as ORD (O'Hare International Airport) and DFW (Dallas/Fort Worth International Airport) stand out with the highest total delays, reflecting their roles as major hubs for both domestic and international travel. These airports experience heavy passenger traffic and flight volumes throughout the year, making them particularly vulnerable to delays caused by congestion. Similarly, ATL (Hartsfield-Jackson Atlanta International Airport), one of the world's busiest airports, demonstrates how high traffic intensity and operational demands contribute significantly to prolonged delays.

Many of the airports featured in this graph are key hubs for major carriers. For example, Delta Air Lines operates its headquarters and main hub at ATL, American Airlines at DFW, and United Airlines at ORD. This alignment of airline operations with high congestion at major hubs highlights how both factors work together to cause

delays. The concentration of delays at these major hubs underscores the economic strain on airlines, including increased staffing and maintenance costs to minimize delays and reduce disruptions to passenger travel plans.

The graph also suggests that these top airports face challenges from weather and other external factors, such as seasonal variations. For instance, ORD, which recorded 24,004,246 minutes of delays between 2018 and 2023, is known to experience significant delays during winter due to severe snowstorms caused by lake-effect snow from the nearby Great Lakes. In contrast, airports like DFW and ATL often face disruptions caused by thunderstorms and hurricanes. Further analysis indicates that these factors significantly affect both airlines and airports, contributing to the overall delay landscape.

In comparing the airports with the greatest delays, this analysis also factored into account the airlines with the most total delays. This graph reflected the airlines with the greatest delays:



The graph identifies the three major U.S. carriers, Delta Air Lines Network, American Airlines Network, and United Airlines Network as being among the top carriers with significant average delays. These carriers have a notable presence at key hub airports, many of which rank among the top 10 airports by total delays. The network of domestic and international flights leads to heavy congestion, operational complexities, and prolonged delays. These delays often reflect the challenges of managing massive flight volumes and maintaining flight schedules in densely populated hubs, particularly when external disruptions like weather or air traffic control issues arise.

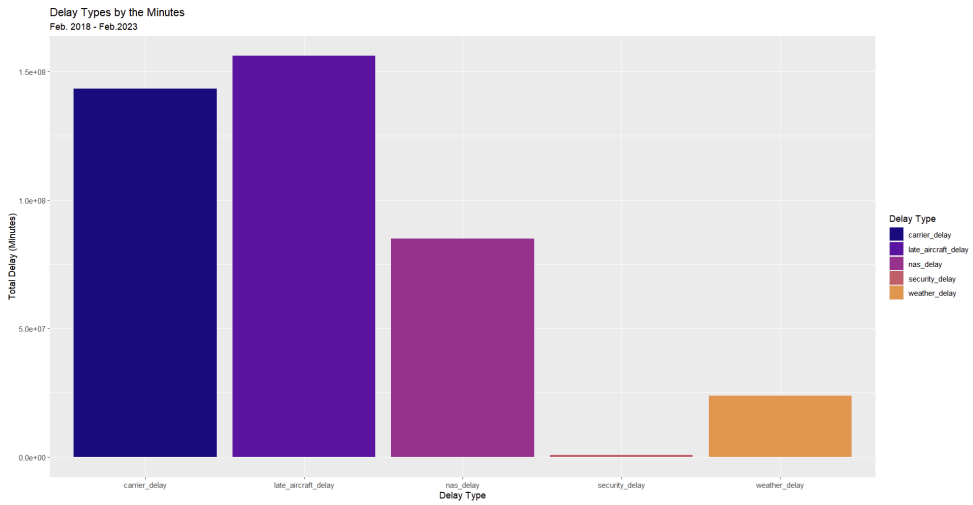
The graph also highlights that Southwest Airlines, JetBlue Airways, and Spirit Airlines, are the three largest low-budget carriers that also experience comparable levels of delays. These carriers operate on high-frequency,

point-to-point routes and serve major airports such as LAX (Los Angeles International Airport), DFW (Dallas/Fort Worth International Airport), and LGA (LaGuardia Airport). Their business model emphasizes quick aircraft turnaround times to minimize costs and gain revenue, but this approach can amplify delays if disruptions occur. For example, if a delay affects one segment of a point-to-point schedule, the subsequent flights are likely to experience compounding delays, especially when operating at congested airports. These operational pressures, combined with a reliance on fewer resources as compared to legacy carriers, make them particularly vulnerable to prolonged delays.

This visualization broadens the patterns of flight delays observed across the aviation industry, linking delays to both airport congestion and the operational strategies of airlines. Hub airports for the major carriers face delays exacerbated by the sheer volume of flights, and the need to coordinate international and domestic operations. On the other hand, low-cost carriers, while not primarily hub-focused, encounter similar issues due to tight schedules and reliance on high-utilization models. The interconnected nature of these delays highlights how disruptions at one hub or starting point can cascade throughout an airline's network. This further emphasizes the relationship between operational strategies and delayed outcomes.

Identifying the major causes of flight delays better helps the understanding as to why the airlines and the airports face the factors of increased flight delays. The key causes for delays are related to the operations of the airlines and the airports, these factors are identified as, delay type, and the variables used to identify they are:carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay. With the usage of the codes to filter the delay types and categorize them into a consolidated structure the creation of this was given:

```
delay_proportion <- Delays %>%
  pivot_longer(cols = c(carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay),
               names_to = "delay_type", values_to = "delay_value") %>%
  group_by(delay_type) %>%
  summarize(total_delay = sum(delay_value, na.rm = TRUE))
```



The graph illustrates the distribution of delay types across the dataset and highlights late aircraft and carrier delays as the most significant contributors, far surpassing NAS (National Airspace System), security, and weather delays. Late aircraft

delays often cascade through an airline's schedule, causing congestion at busy hubs. Similarly, carrier delays stem from operational inefficiencies such as staffing shortages, maintenance issues, and logistical challenges, which are exacerbated by congestion at major airports.

Airports with high flight volumes, particularly those in the top 10 by total delays, are especially vulnerable to late aircraft delays due to their role as connecting points in airline networks. Carrier delays further compound the problem, as managing large-scale operations at congested airports strains resources, leading to longer wait times for passengers and increased pressure on airport infrastructure.

To address the challenges highlighted in the graph, airlines and airports must prioritize optimizing operations and reducing congestion. Strategies to tackle late aircraft delays could include improving turnaround times, enhancing resource allocation, and implementing predictive analytics to prevent cascading delays. Similarly, reducing carrier delays would require streamlining operations, improving staff scheduling, and investing in proactive maintenance. By addressing these issues, airlines and airports can work toward minimizing delays, alleviating congestion, and enhancing overall efficiency in the aviation industry.

Building on the insights from the graphical analyses, regression models were conducted to quantify the impact of airport congestion, weather, and operational inefficiencies on flight delays. These models provided a framework to validate the stated hypothesis and offered statistical evidence for the factors driving delays. A linear regression model was specifically used to examine the relationships between total flight delays and key predictors.

```
model_delay <- lm(total_delay ~ arr_flights + weather_ct + nas_ct, data = Delays)
print(summary(model_delay))
```

The results given the regression test:

```
Call:
lm(formula = total_delay ~ arr_flights + weather_ct + nas_ct,
    data = Delays)

Residuals:
    Min      1Q  Median      3Q     Max
-103586    -445     -61     254  129410

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.25259   10.71280   -0.21    0.833
arr_flights   3.88181    0.02289  169.57   <2e-16 ***
weather_ct  450.27940    2.20485  204.22   <2e-16 ***
nas_ct       98.22450    0.34378  285.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3504 on 119928 degrees of freedom
Multiple R-squared:  0.8959,    Adjusted R-squared:  0.8959
F-statistic: 3.441e+05 on 3 and 119928 DF,  p-value: < 2.2e-16
```

The results relayed from this regression provided insights that each arriving flight into the airports contributed approximately 3.88 minutes to total delays each weather event added 450.28 minutes of delay, and each NAS issue contributed 98.22 minutes. This model supported the data given on the variance of flight delays which was recorded at the R-squared value of 0.8959. These findings align with the FAA's report that 74.26% of system-impacting delays are caused by weather, further emphasizing the significant contribution of adverse weather conditions to prolonged delays. The regression model quantifies this impact, highlighting weather events as a dominant predictor of delays, consistent with prior research. This demonstrates the strong correlation in the predictors (arrivals, weather, and NAS delays) in understanding total delays. As well as the small p-values indicate that the variables that were tested are significant in predicting total delays and that the factors contribute to the causes of flight delays. The significance of this regression and the testing of these variables supports the alternative hypothesis.

A Generalized Additive Model was conducted to account for any nonlinear relationships between the predictors and delays:

```
model_gam <- gam(total_delay ~ s(arr_flights) + s(weather_ct) + s(nas_ct), data = Delays)
print(summary(model_gam))
```

This model revealed a significant connection for arriving flights, weather-related delays, and NAS delays being factors to flight delays as it resulted in the p-values less than 0.001. These models collectively contribute to the analysis by offering perspective on the linear and nonlinear relationships on the factors driving flight delays.

a-A descriptive summary was conducted to show the flight delays and variables relating to these causes. Data from this test are:

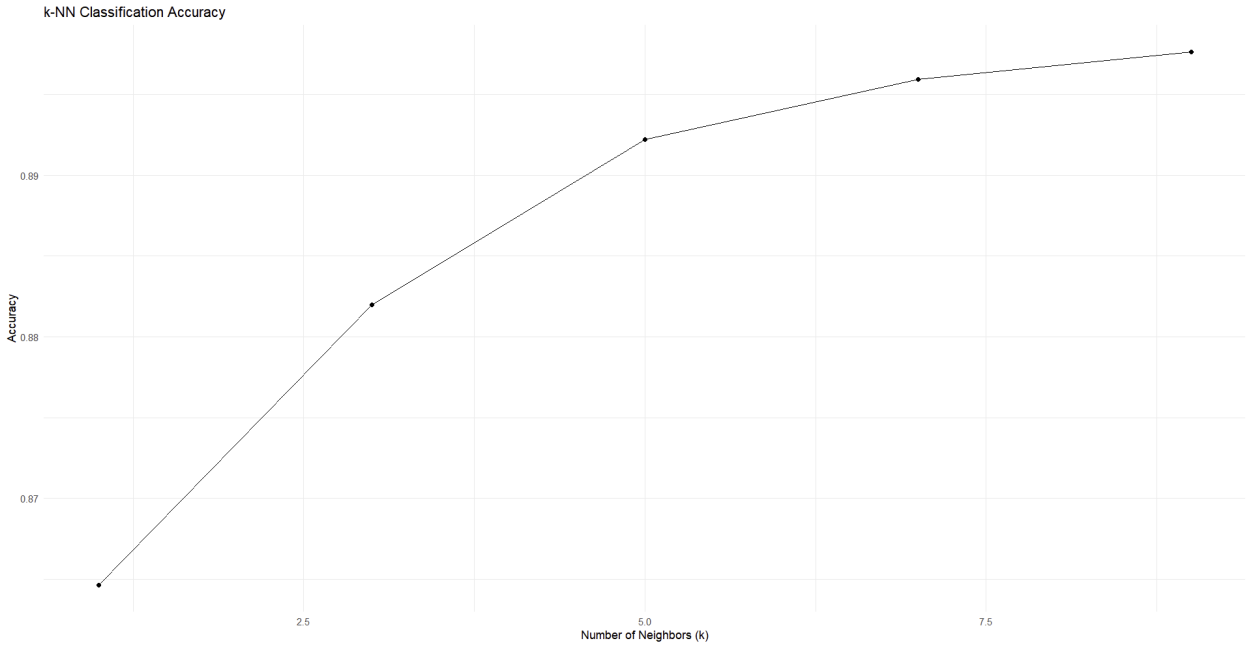
```
  avg_flights avg_total_delay avg_weather_ct avg_nas_ct avg_late_aircraft_ct
1    289.9715        3411.756       1.876979   14.69316             17.51116
```

This indicates that on average, there are approximately 290 arriving flights per period in the dataset. This serves as a representation of airport congestion and highlights the flight traffic. The average total delay of 3411.76 minutes represents the cumulative nature of the delays which emphasizes the operational strain imposed on airlines and airports. Additionally, this test reveals an average of 1.88 weather-related delay events, 14.69 NAS delay events, and 17.51 late aircraft delay events per period. This demonstrates the varying contributions of different delay types.

A KNN test was done to show the accuracy of the data with the classification of flight delays. The k-Nearest Neighbors (KNN) classification test analysis provides key insights into the predictability of flight delays based on the

variables already established such as the number of arriving flights, weather conditions, and operational issues. This test is done by using training and testing datasets created in this analysis. The accuracy of classification in the test improved progressively as the number of neighbors (k) increased. The numbers of neighbors recorded were: 1.0000000 = 0.8646412, 3.0000000 = 0.8819842, 5.0000000 = 0.8922228, 7.0000000 = 0.8959422, and 9.0000000 = 0.8976556, the peak accuracy approximation was 89.77% for k = 9. This result emphasizes the strength of a strong correlation between the chosen variable and predictors in distinguishing between the two categories of "High Delay" and "Low Delay".

The KNN results validate the reliability of the dataset and the variables used to identify delays. Classifying the delay categories with high-accuracy results highlights the significance of factors such as airport congestion, weather events, and operational inefficiencies in contributing to flight delays. This provides an alternative approach to analyzing the relationships within the dataset. The graph reflecting the results from this test is shown below:



Passenger Traffic and Flight Delays:

To fully analyze the hypotheses, was necessary to also examine the effects that high passenger volumes play when looking at delays at airports. The study incorporates an analysis of passenger traffic as a proxy for airport congestion. A creation of secondary hypotheses will be used to then support the effects of the primary hypothesis. This secondary hypothesis explores whether airports with higher passenger volumes experience longer delays this will

thereby provide an understanding of how congestion exacerbates operational inefficiencies and leads to flight delays. The secondary hypothesis states:

- Null Hypothesis (H_0): Passenger traffic is not associated with flight delays.
- Alternative Hypothesis (H_1): Airports with higher passenger traffic are more likely to experience longer flight delays.

By investigating the relationship between passenger traffic and delays, this secondary hypothesis seeks to disclose the impacts of airport congestion and its contribution to delays. This analysis complements the primary hypothesis as it demonstrates how increased passenger volumes at major airports compound delays through heightened operational strain and limited capacity, thereby bridging the gap between theoretical concepts of congestion and their practical impacts on delay outcomes.

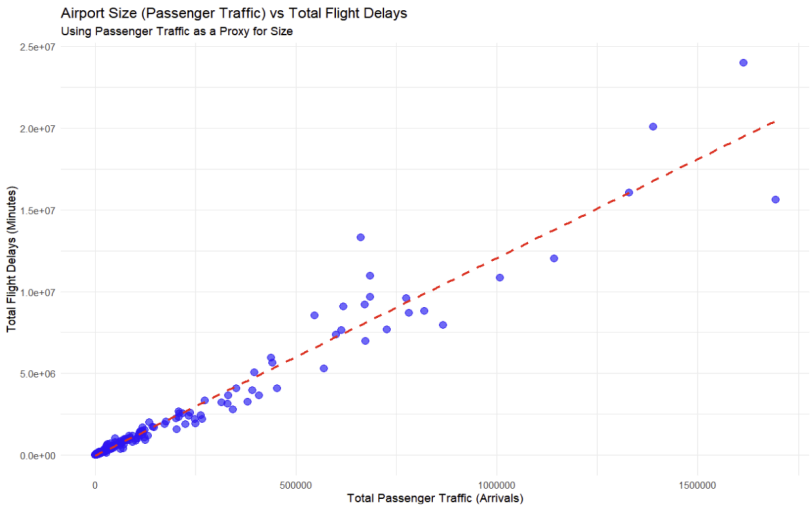
Since this dataset doesn't explicitly state the amount of passenger traffic, nor does it provide insights on the number of runways at the airports or the general size of the airport, a proxy variable was created to exemplify this factor. Using the group_by function, all airport names, arrival times, and total delays were grouped into a table representing the traffic_delays.

airport_name	total_passenger_traffic	total_delays	avg_delay
<chr>	<dbl>	<dbl>	<dbl>
Atlanta, GA: Hartsfield-Jackson Atlanta International	1693363	15605443	17187.
Chicago, IL: Chicago O'Hare International	1614333	24004246	25162.
Dallas/Fort Worth, TX: Dallas/Fort Worth International	1388818	20091546	25021.
Denver, CO: Denver International	1328885	16065046	20570.
Charlotte, NC: Charlotte Douglas International	1143055	12012923	13363.
Los Angeles, CA: Los Angeles International	1008026	10844556	14635.
Seattle, WA: Seattle/Tacoma International	866696	7948959	11356.
Phoenix, AZ: Phoenix Sky Harbor International	819776	8795583	11349.
Houston, TX: George Bush Intercontinental/Houston	774074	9596056	11847.
Detroit, MI: Detroit Metro Wayne County	726222	7676809	7534.

A scatter plot was then created to compare passenger traffic with total delays. Subsequently, a correlation test was conducted to display the p-values and t-values. The results from this test are as follows:

```
data:  traffic_delays$total_passenger_traffic and traffic_delays$total_delays
t = 90.221, df = 407, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9707997 0.9801181
sample estimates:
      cor
0.9758998
```

A regression test was performed, along with a regression plot illustrating "Passenger Traffic Predicting Delays." Additionally, a t-test was conducted to compare delays between high- and low-traffic airports. As the analysis progressed, the decision to examine airport size or runway size through a proxy of airport variables provided insights into the role of passenger traffic in causing delays. The results from these tests yielded the following regression model:



The analysis of passenger traffic revealed a strong correlation between total passenger traffic and flight delays, further reinforcing the main hypothesis that flight delays are caused by airport congestion and operational inefficiencies. Airport congestion data was used as a proxy for passenger traffic. The correlation test demonstrated a strong relationship between passenger traffic and total delays, with a highly significant correlation coefficient ($r = 0.976$, $p < 0.001$). This indicates that higher passenger volumes are strongly associated with prolonged delays. A linear regression model further supported this finding, showing that the total passenger traffic variable significantly predicted total delays, with an adjusted R-squared value of 0.9526. The regression results suggest that for every additional arrival, total delays increase by approximately 12.1 minutes. These findings align with the operational challenges highlighted by Erdem & Bilgiç (2024), who noted that airline operations are particularly susceptible to disruptions caused by increased passenger volumes. Slow baggage handling and delayed flight crews at high-traffic airports compound operational inefficiencies, reinforcing the strong correlation between passenger traffic and prolonged delays. This aligns with the hypothesis that airports with higher passenger traffic volumes face increased operational congestion due to factors such as limited gate availability, crowded runways, and longer turnaround times. These challenges are compounded by the influx of passengers waiting for flights, further straining airport operations. Airports handling higher capacities of passengers tend to experience more delays, often tied to their ability—or lack thereof—to operate at peak efficiency.

To further examine the impact of traffic, airports were categorized into "High Traffic" and "Low Traffic" groups based on passenger volume. A t-test comparing total delays between these groups revealed a significant

difference in mean delays. High-traffic airports experienced substantially longer delays (mean = 1,951,897 minutes) than low-traffic airports (mean = 53,617 minutes). This emphasizes the systemic challenges faced by larger hubs in managing passenger volumes efficiently, particularly during peak travel periods. A scatter plot visualization illustrated the correlation between traffic levels and delays, reinforcing the argument that congestion plays a pivotal role in shaping delay outcomes. Based on these results, the null hypothesis can be rejected, as the findings support the alternative hypothesis. When combined with other factors, such as weather and late aircraft delays, this secondary analysis supports the primary hypothesis by highlighting the critical role of high passenger traffic in exacerbating operational congestion, which drives delays at major airports.

Covid-19 and Flight Delays:

To deepen the understanding of this data analysis on flight delays, it is essential to investigate whether flight delays can be reduced or minimized. This analysis aims to verify if the alternative hypothesis holds. The null hypothesis states that airport congestion, weather, and operational issues have no significant effect on flight delays. To challenge this hypothesis on an analytical scale, this analysis incorporates seasonal variations and the impact of the COVID-19 pandemic on the aviation industry to show the effects of limited flights on flight delays.

The COVID-19 pandemic had a profound negative impact on the aviation industry, causing substantial reductions in the number of flights and the grounding of many aircraft for carriers and airports. These disruptions led to operational challenges and contributed to changes in flight delays. Data gathered from the BTS on the topic of flight delays reveals that 2020 saw a notable decrease in average delays compared to pre-pandemic years. This is attributable to reduced congestion and travel restrictions imposed on passengers and carriers.

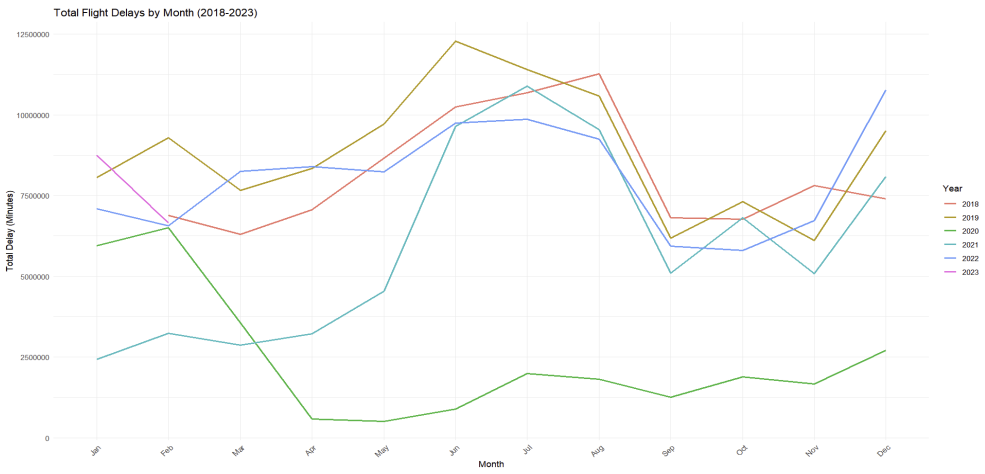
For instance, the dataset revealed that the mean total delay in 2020 was approximately 1,349 minutes, compared to 4,034 minutes pre-COVID and 3,705 minutes post-COVID. This reflects the drastic reduction in air traffic during the pandemic and its cascading effects on operational efficiency. By examining seasonal variations during the pandemic, differences in delay patterns become evident. For example, summer delays in 2020 were significantly lower, coinciding with travel restrictions during peak months when most travelers typically take vacations due to school closures and student holidays. This increase in travel restrictions allowed for lower delays during the summer months of 2020 compared to other years.

Using regression models and hypothesis testing, the decision in this analysis was to evaluate whether the null hypothesis could be accepted or rejected. The analysis employed regression models to identify the impact of key

variables such as arrival flights, weather, and operational factors on flight delays in 2020. The COVID-19 regression model analysis showed a high adjusted R-squared value of 0.9295, indicating that these variables contributed to over 92% of the variance in total delays during the pandemic year. A polynomial regression was also used to capture non-linear relationships, as demonstrated in this code.

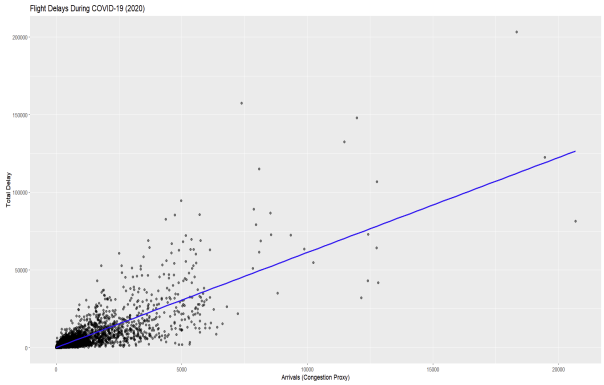
```
model_poly <- lm(
  total_delay ~ arr_flights + I(arr_flights^2) + I(arr_flights^3) +
    weather_ct + I(weather_ct^2) + nas_ct + late_aircraft_ct + season,
  data = covid_data
)
```

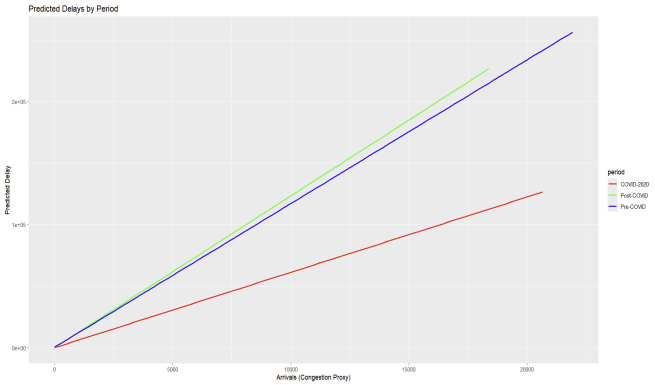
Hypothesis testing was also used to confirm the p-values in the dataset. The interaction models revealed the effects of seasonal variations and their influence on delay categories. These findings underscore the complex interplay of factors contributing to delays, particularly during the pandemic.



The graph above visualizes the line trend of flight delays over the years and months, providing insights into delay patterns from 2018 to 2023. The monthly delays for the 5-year trend reveal that 2020 displayed an abnormal pattern compared to pre- and post-pandemic years. Scatterplot models plotted during the pandemic show the effects of delays on arrival flights, revealing weaker correlations in 2020. This reflects reduced airport congestion and fewer operational challenges due to the significantly lower number of flight operations during that time.

The visualization of predicted delays across pre-, current-, and post-COVID periods revealed a less pronounced slope for 2020. This supports the conclusion that reduced airport congestion during the pandemic played a significant role in minimizing delays. The graphs below further illustrate the impacts of the pandemic on flight delays.





The results of this data analysis on flight delays during COVID-19 reject the null hypothesis and confirm that airport congestion, weather, and operational issues significantly affect flight delays. During the pandemic, the number of flight delays was noticeably reduced, largely due to fewer flights operating. This reduction meant less need for operations, allowing flights to depart and arrive on time, and creating more available space at airports, which helped alleviate congestion. By utilizing regression models, hypothesis testing, and visual tools, this analysis demonstrates how the pandemic reshaped the dynamics of flight delays.

Machines Learning:

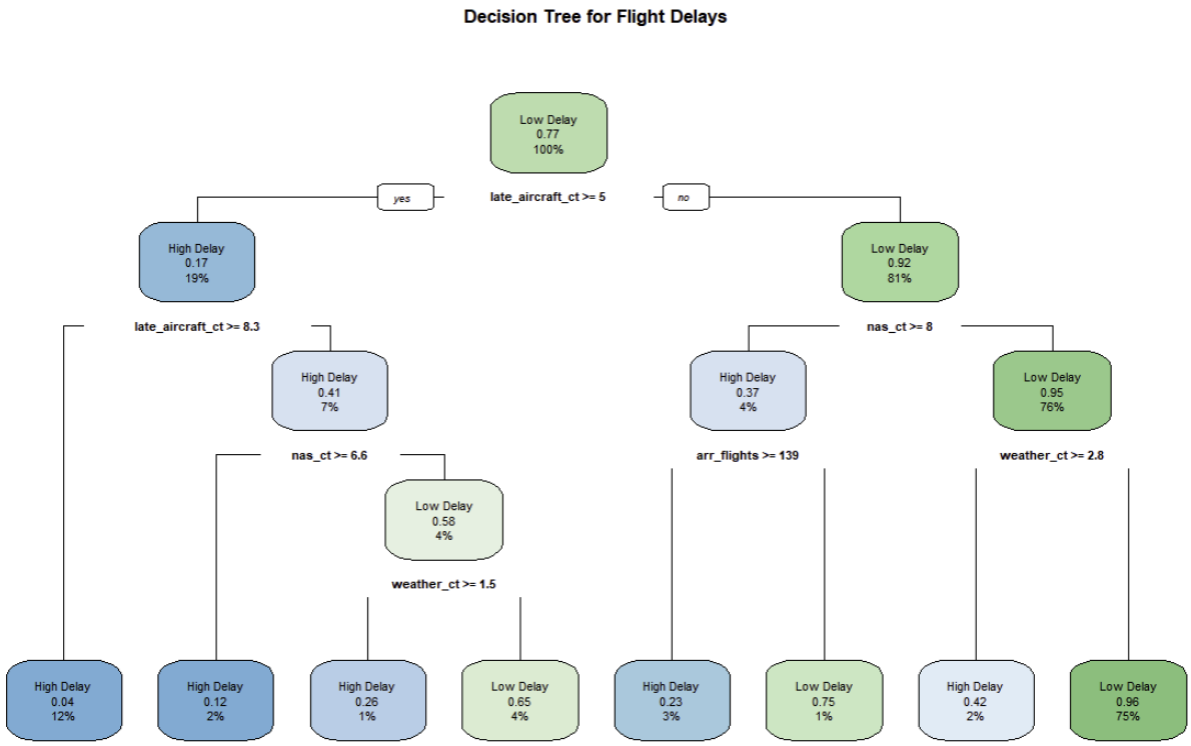
To delve deeper into this analysis of flight delays, the use of machine learning codes was utilized. Using the tool package library(randomForest) and library(rpart) allowed for the creation of a random forest and decision tree model. These models were used to analyze the variance importance of flight delays as well as verify visual imagery and understanding of the “high delay” and “low delay” categories.

Decision Tree Analysis:

The decision tree model generated provided a clear visualization of how flight delays are classified into "High Delay" and "Low Delay." The categories are based on the variables such as arr_flights, weather_ct, nas_ct, late_aircraft_ct, and season. The result showed an accuracy of 92.52%, which means that this model highlights the critical role of late_aircraft_ct in causing flight delays. The model shows the cascading effect of the trend of the delay; if an aircraft arrives over the threshold of greater than 5, it is automatically classified as a late aircraft arrival. From there, the classifications deepen to pinpoint the particular cause of this late arrival. Late aircraft arrivals emerge as the most significant contributors to high delays, followed by variables like weather-related delays.

In addition to accuracy, the decision tree model exhibited a Type I Error (False Positive Rate) of 3.77% and a Type II Error (False Negative Rate) of 20.33%. This indicates that the model rarely misclassifies low delays as high delays but is less effective in identifying true high-delay cases. These results suggest room for improvement in detecting significant delays while maintaining precision for low-delay cases. The findings support the alternate hypothesis that delays are primarily driven by operational factors rather than external influences like weather.

Graph of Decision Tree:



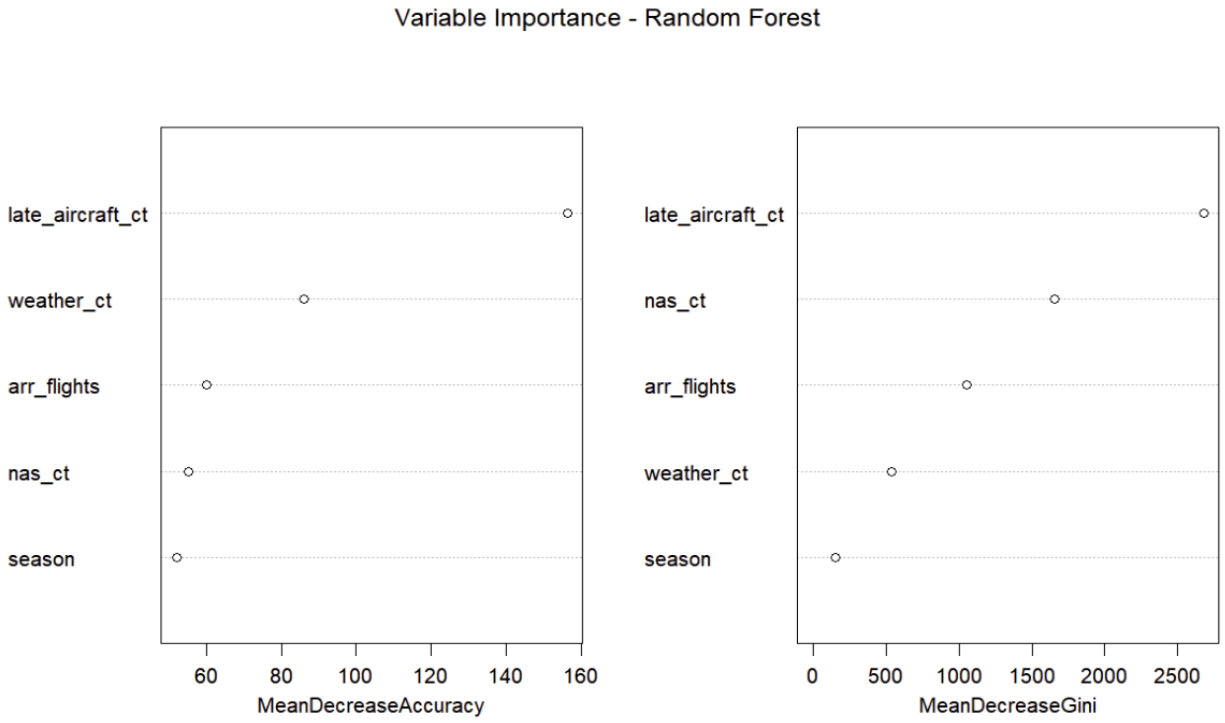
Random Forest Analysis:

The random forest model outperformed the decision tree and achieved greater results, recording an accuracy of 93.4%. The random forest managed to filter out and verify any misclassification errors for high delay, making it more detailed in understanding the predictors of delays. A variable importance graph was generated that validated the importance of late_aircraft_ct as the most influential factor of flight delays, followed by nas_ct and weather_ct. Seasonal variation, while statistically relevant in the decision tree, showed less importance in the random forest model, suggesting that airline performance issues persist across seasons.

The random forest model also showed improved error rates, with a Type I Error (False Positive Rate) of 3.26% and a Type II Error (False Negative Rate) of 18.17%. This improvement demonstrates the model's enhanced ability to correctly classify high-delay cases while minimizing false alarms for low delays.

In optimizing the random forest model with 1,000 trees, the accuracy decreased slightly to 93.21%. However, this minimal change in accuracy was not impactful, as the model maintains an effective balance and is deemed sufficient for identifying predictors of delays. The optimized random forest model recorded a Type I Error (False Positive Rate) of 3.32% and a Type II Error (False Negative Rate) of 18.79%, slightly higher than the standard random forest but still lower than the decision tree.

Graphs of RandomForest:



Interpreting Model Outputs:

These models' results offer actionable insights into the need to mitigate flight delays. For instance, the most common cause for delays due to late_aircraft_ct highlights the need for airlines to optimize turnaround times to prevent cascading delays. Similarly, the delay caused by nas_ct suggests that further investments in air traffic control infrastructure and improved scheduling for airports could alleviate systemic issues.

The analysis of Type I and Type II errors across models demonstrates the trade-offs between minimizing false alarms and accurately identifying true high-delay cases. The random forest models managed to quantify the importance of these variables and provide insights to policymakers, airports, and airlines with a data-driven foundation for prioritizing the effects and causes of flight delays. By addressing these factors, airlines and airports can minimize delays, reduce congestion, and improve operational efficiency.

Conclusion:

This comprehensive analysis of flight delays highlights the multifaceted factors influencing delays, including airport congestion, weather conditions, and operational inefficiencies. Regression and machine learning models provided a robust framework for testing the hypothesis, leading to the rejection of the null hypothesis in favor of the alternative. The study identified significant contributions of key variables to overall flight delays in the United States between February 2018 and February 2023. Regression analyses underscored the critical roles of weather events, airport congestion, and operational effects in contributing to delays, aligning with prior research. Passenger traffic, used as a proxy for airport congestion, emerged as a pivotal factor, with each additional arrival adding 12.1 minutes to total delays. These findings emphasize the complex interplay between operational strain, passenger volumes, environmental factors, and airport congestion in driving delay patterns.

The machine learning models further validated these findings, offering robust accuracy metrics and insightful classifications. The random forest model, with an accuracy of 93.4% and lower Type I and Type II error rates compared to the decision tree, proved highly effective in identifying the key predictors of delays. Actionable recommendations derived from these findings include optimizing turnaround times, investing in air traffic infrastructure, and developing weather resilience strategies. The integration of literature research and dataset-driven evidence strengthens the case for targeted interventions to mitigate delays, reduce congestion, and enhance operational efficiency in the aviation industry. This study confirms existing research while expanding the understanding of how various factors shape delay patterns, providing a foundation for data-driven improvements in airline and airport operations.

Citation:

- Airlines for America. (2024). U.S. passenger carrier delay costs. Retrieved July 12, 2024, from https://www.airlines.org
- Bombelli, A., & Sallan, J. M. (2023). Analysis of the effect of extreme weather on the US domestic air network: A delay and cancellation propagation network approach. *Journal of Transport Geography, 107*, 103541. https://doi.org/10.1016/j.jtrangeo.2023.103541
- Bureau of Transportation Statistics. (2024). Airline on-time statistics and delay causes: September 2024. U.S. Department of Transportation. Retrieved from https://www.transtats.bts.gov/
- Erdem, F., & Bilgiç, T. (2024). Airline delay propagation: Estimation and modeling in daily operations. *Journal of Air Transport Management, 115*, Article 102548. https://doi.org/10.1016/j.jairtraman.2024.102548
- Peterson, E., Neels, K., Barczi, N., & Graham, T. (2013). The economic cost of airline flight delay. *Journal of Transport Economics and Policy, 47*(1). https://doi.org/10.xxxx/yyyy (replace with actual DOI if available)
- Rupp, N. G. (2007). Further analysis of the determinants of airline flight delays (ECU Economics Working Paper No. 07-07). East Carolina University. https://economics.ecu.edu/wp-content/pv-uploads/sites/165/2019/07/ecu0707.pdf
- U.S. Department of Transportation, Federal Aviation Administration. (2024, September 5). NextGen weather overview: FAQ weather delay. Retrieved from https://www.faa.gov
- U.S. Department of Transportation. (2024, September 5). FAQ: Weather delay. *Federal Aviation Administration.* https://www.faa.gov/nextgen/programs/weather/faq
- U.S. Government Accountability Office. (2023). Airline passenger protections: Observations on flight delays and cancellations, and DOT's efforts to address them (GAO-23-105524). Report to Congressional Requesters. Retrieved from https://www.gao.gov