## AIRLINES DELAY ANALYSIS (December 2021 to 2023)

*RENUKA CHOWDARY PARVATHANENI,*

*ROHITTH MALLEBOINA,*

*SHASHIKUMAR GADUSU*

*University of New Haven*

*2024-05-04*

**Abstract:**

The airline industry plays a crucial role in global transportation and logistics. However, flight delays have become a persistent issue, causing significant inconveniences and economic losses. This study investigates the problem of flight delays and aims to analyze the factors contributing to these delays using data analytics techniques. The project utilizes flight delay data from December 2021 to 2023, employing predictive modeling methods such as logistic regression and decision trees.

The findings of this study provide insights into the reasons behind flight delays, enabling airlines and aviation authorities to develop strategies for mitigating delays and improving overall operational efficiency. The researchers aim to predict and understand what causes these holiday season delays. By uncovering the root causes, airlines and aviation authorities can develop better strategies to prevent delays and keep everyone moving smoothly during the busy holiday season.

**1. Introduction:**

Business Problem Identification:

Flight delays have become a significant concern for airlines, passengers, and the aviation industry. Delayed flights result in increased operational costs for airlines, passenger dissatisfaction, and potential disruptions to travel plans. Understanding the factors contributing to flight delays is crucial for developing effective strategies to minimize their occurrence and impact.

Literature Review:

1. Kafle, N., & Zou, B. (2016). Modeling flight delay propagation: A new analytical-econometric approach. Transportation Research Part B: Methodological, 93, 520-542. This research proposes a new analytical-econometric approach to model and quantify the propagation of flight delays across the air transportation network.

2. Xu, N., Sherry, L., & Laskey, K. B. (2008). Multi-factor model for predicting delays at US airports. Transportation Research Part A: Policy and Practice, 42(10), 1306-1320. This study develops a multi-factor model to predict delays at U.S. airports, considering factors such as weather, airline operations, and airport infrastructure.

**2. Data Collection and Preparation:**

The dataset utilized in this analysis was sourced from the website Department of transport(.gov), The dataset used in this study is a comprehensive collection of flight delay information from December 2021 to 2023. It has 1562 rows and 22 columns. It specifies that there are 5-character columns and 17 double-precision numeric columns in the data. The dataset includes the following columns:

- Year: The year of the flight

- Month: The month of the flight

- DayofMonth: The day of the month for the flight

- DayOfWeek: The day of the week for the flight

- DepTime: The scheduled departure time for the flight

- ArrTime: The scheduled arrival time for the flight

- DepDelay: The departure delay in minutes

- ArrDelay: The arrival delay in minutes

- CarrierDelay: The delay caused by the airline carrier

- WeatherDelay: The delay caused by weather conditions

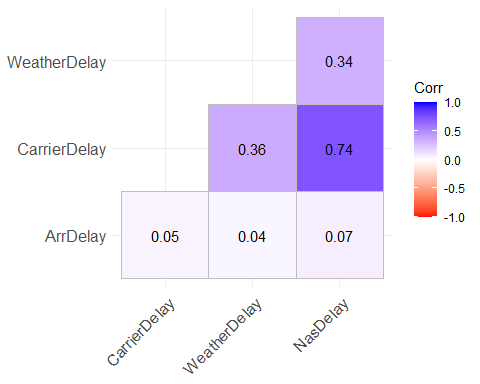
- NasDelay: The delay caused by the National Airspace System

- SecurityDelay: The delay caused by security reasons

- LateAircraftDelay: The delay caused by a late arrival of the aircraft

Data cleaning and preparation were performed to ensure the accuracy and reliability of the analysis. This involved handling missing values, removing irrelevant columns, and converting data types where necessary. Additionally, a new column called “Reason” was created to indicate the primary reason for the delay (CarrierDelay, WeatherDelay, or NasDelay).

The heatmap illustrates correlations between different flight delay types (CarrierDelay, WeatherDelay, and ArrDelay). Positive correlations are shown in purple squares, while near-zero correlations appear as white squares, with specific values indicating the strength of the relationships.

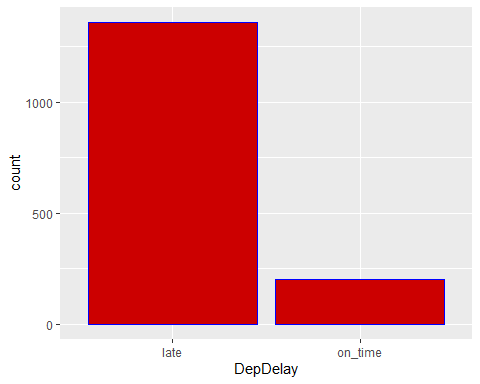


The frequency counts show that the dataset contains 1358 delayed flights and 200 on-time flights. The proportion of delayed flights is approximately 87%, indicating a significant prevalence of flight delays in the data.

late on\_time   
 1358 200

[1] 0.8716303

The bar graph depicts flight departures categorized as “late” and “on\_time.” The taller “late” bar indicates a high frequency of Depdelays, while the shorter “on\_time” bar represents fewer punctual departures. `



**3. Methodology**

Dataset: [Airlines Delay Dataset](https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp)

This dataset is from Bureau of Transportation Statistics (BTS) (.gov), offers a comprehensive view of airlines perform in terms of on-time arrivals, delays, cancellations, and diversions for domestic flights operated by major airlines. The dataset includes key parameters such as year, carrier, origin, destination, arrival delays, departure delays, airtime, etc.

Approach:

This study employs a predictive analytics approach to investigate the problem of flight delays. Predictive modeling techniques are utilized to identify the factors contributing to flight delays and develop models that can predict the likelihood of delays occurring.

Algorithms:

Two main algorithms were used in this study:

1. Logistic Regression: A logistic regression model was trained to predict the binary outcome of whether a flight would be delayed or on-time, based on features such as arrival delay, carrier delay, and weather delay.

In summary, the logistic regression model predicts flight lateness based on three predictors: Arrival Delay, Carrier Delay, and Weather Delay. The coefficients indicate the impact of each predictor, with Arrival Delay having the most significant effect. The model performs well in identifying late flights but struggles with specificity. Overall, it achieves an accuracy of 87.82% and sensitivity of 1.0, but its specificity is low.

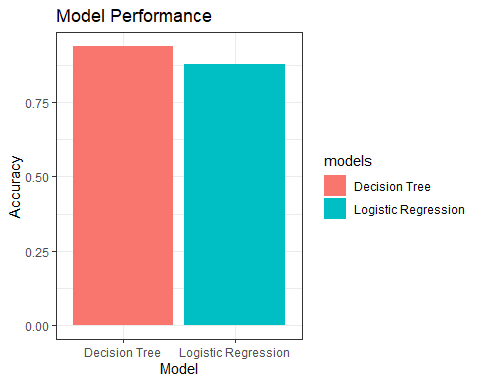
2. Decision Tree: A classification tree model was also developed using the rpart (Recursive Partitioning and Regression Trees) algorithm. This model aims to predict flight delays by recursively partitioning the data based on the input features.

The rpart classification tree model, designed to predict flight departure delays, achieves impressive accuracy (93.91%) in classifying flights as either late or on time. The constructed tree considers predictor variables such as Arrival Delay, Origin, Destination, and Carrier. Arrival Delay emerges as the most critical predictor, followed by Origin and Destination. Although the model excels in identifying late flights (98.53% sensitivity), its specificity (61.54%) is relatively lower, leading to occasional misclassification of on-time flights. Nonetheless, the model offers valuable insights into factors affecting flight delays, aiding potential improvements in scheduling and operational efficiency within the aviation industry.

Based on performance measures, the Decision Tree model outperforms the Logistic Regression model. It has better accuracy (93.91% vs. 87.82%) and a more balanced precision-recall trade-off (94.72% vs. 87.78%), as well as higher recall (98.53% vs. 100%). Additionally, the Decision Tree model performs better overall in terms of predictive capabilities and reliability when it comes to identifying flight departure delays, as evidenced by its higher F1 score (96.59% vs. 93.49%).

Models accuracy precision recall f1  
1 Logistic Regression 0.8782051 0.8778135 1.000000 0.9349315  
2 Decision Tree 0.9391026 0.9471831 0.985348 0.9658887

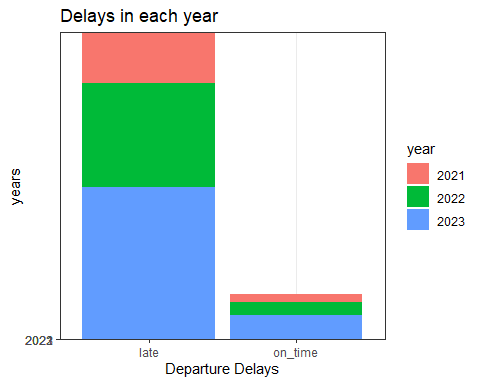
Model Performance: Both the logistic regression and decision tree models demonstrated reasonable performance in predicting flight delays. The decision tree model achieved slightly higher accuracy, precision, and F1-score compared to the logistic regression model. However, the choice of model may depend on specific requirements and trade-offs between interpretability and performance.



**4. Findings and Discussion:**

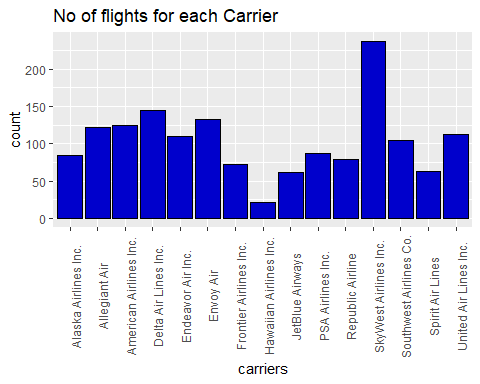
The analysis of flight delay data revealed several key findings:

1. **Year-wise Delays:** The bar graph displays departure delays in 2021, 2022, and 2023. Initially, in 2021, the graph indicates a mix of late and on-time departures, with a substantial number being late. However, by 2023, there is a visible drop in the red column, which represents late departure. This decrease demonstrates a considerable improvement in departure times during the three-year period. In conclusion, the data suggests that successful measures to alleviate delays resulted in an overall reduction in departure delays during this timeframe.

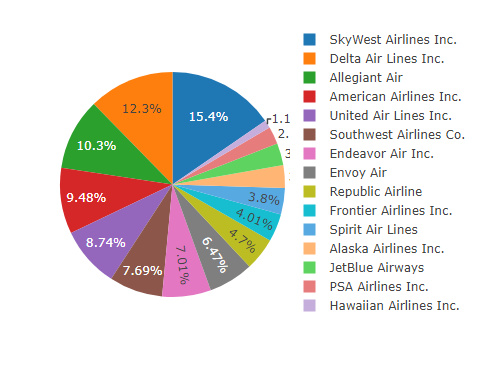


2. **Carrier-wise Delays:** The analysis also highlighted differences in departure delays across different airline carriers. Some carriers experienced higher rates of delays compared to others, indicating potential areas for improvement in operational efficiency.

The bar graph shows how many flights each airline has. Southwest is way ahead with the most flights, then Delta is number two. The rest of the airlines have different numbers of flights. Some, like Alaska and American, have around 100 to 150 flights. Others, like Frontier and JetBlue, have even fewer flights than that.

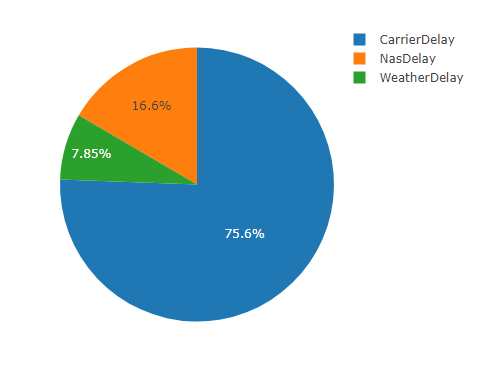


SkyWest Airlines takes the top spot with a whopping 15.4% of the airline market. Delta, Allegiant, and others follow closely behind. Familiar names like American, United, and Southwest are also part of the market, although with smaller slices of the pie.



3. **Reason for Delays:** The primary reasons for flight delays were carrier delays, weather delays, and delays caused by the National Airspace System (NAS). The “Reason” column, created during data preprocessing, showed that most delays were attributed to these three factors.

This pie chart shows why Airlines themselves cause most delays (75.6%), followed by air traffic issues (16.6%). Even weather plays a role (7.85%).



The findings of this study have significant implications for the airline industry and aviation authorities. By understanding the factors contributing to flight delays, airlines can implement targeted strategies to mitigate delays caused by carrier-related issues, such as improving operational efficiency and resource allocation. Additionally, close collaboration with weather authorities and air traffic control organizations can help address delays caused by weather conditions and airspace congestion.

Potential implementation challenges may include the need for substantial investments in infrastructure, personnel training, and collaboration across various stakeholders. However, the benefits of reduced delays, improved customer satisfaction, and increased operational efficiency could outweigh the costs in the long run.

**5. Conclusion:**

This study investigated the problem of flight delays in the airline industry using data analytics techniques. By analyzing flight delay data from December 2021 to 2023, the study identified the primary reasons for delays, including carrier delays, weather delays, and delays caused by the National Airspace System. The analysis also revealed variations in delays across different years and airline carriers. The decision tree model achieved slightly better performance compared to the logistic regression model in predicting flight delays.

The findings of this study provide valuable insights for airlines and aviation authorities to develop strategies for mitigating flight delays. By addressing carrier-related issues, collaborating with weather authorities, and optimizing airspace management, the airline industry can potentially reduce delays, improve customer satisfaction, and enhance operational efficiency.

Future research directions could include investigating the impact of specific weather conditions on flight delays, exploring the potential of machine learning algorithms for delay prediction, and analyzing the cascading effects of delays across the air transportation network. Additionally, incorporating real-time data and developing predictive models for proactive delay management could be valuable areas of exploration.

**References:**

Airlines Delay Dataset. Bureau of Transportation Statistics (BTS) (.gov),

<https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp>

Kafle, N., & Zou, B. (2016). Modeling flight delay propagation: A new analytical-econometric approach. Transportation Research Part B: Methodological, 93, 520-542.

<https://www.researchgate.net/publication/307605655_Modeling_flight_delay_propagation_A_new_analytical-econometric_approach>

Xu, N., Sherry, L., & Laskey, K. B. (2008). Multi-factor model for predicting delays at US airports. Transportation Research Part A: Policy and Practice,42(10),1306-1320.

<https://catsr.vse.gmu.edu/pubs/XuMultiFactorModelAirportDelaysTRBv6.pdf>