Final Project Report

By Abhishek Verma

# 4.1 Motivation or Main Business Idea

Airline delays are a significant operational and financial challenge for the U.S. aviation industry.   
They frustrate passengers, strain airline resources, and disrupt airport logistics. In a tightly interconnected air travel system, even small delays can cascade into widespread disruptions. These disruptions result in increased fuel consumption, overtime labor costs, missed connections, and poor customer satisfaction, ultimately harming both profitability and brand reputation.  
  
This project addresses a critical business need: how to understand and predict airline delays in a systematic, data-driven way.   
Airlines and airport operators require robust tools not just to analyze historical delay patterns but also to anticipate future delay risks—thereby enabling proactive decision-making, better resource allocation, and improved customer experience.  
  
**Main Research Question:** What are the key drivers of airline delays in U.S. domestic flights, and how can we model them for better prediction and management?  
  
To address the research question, the project uses public data from the U.S. Bureau of Transportation Statistics, covering all domestic commercial flights in 2024.   
The analysis applies descriptive statistics, correlation analysis, ANOVA, and predictive modeling (regression, clustering, logistic regression, and machine learning).  
  
Key findings include:  
- Late aircraft and carrier delays are the most impactful delay causes.  
- Seasonal patterns significantly affect delay metrics.  
- A moderate-to-strong correlation (0.60) exists between carrier-caused delays and overall delay rate.  
- Predictive models help forecast major delays with meaningful accuracy.

# 4.2 Data and Empirical Methodology

This project utilizes a cleaned version of the publicly available U.S. Bureau of Transportation Statistics data on domestic airline operations.

* Period: January–December 2024
* Two Datasets Used in SAS and R:
  + Dataset 1 : Airline\_Delay\_Cause.csv
  + Dataset 2 : airlinedata\_final\_log.csv

The final dataset comprises 21,842 observations and 26 variables after preprocessing. Cleaning steps included replacing missing values with zero, removing extreme outliers (top 1%), and creating key engineered variables such as DelayRate, MajorDelayCause, SeasonalQuarter, and flags for major airports/carriers.  
  
Key transformations include log-transformations to normalize skewed delay variables. The empirical strategy includes descriptive statistics (PROC MEANS, PROC SUMMARY), ANOVA, correlation analysis, and predictive modeling using stepwise logistic regression and random forest classification.

**Data Cleaning and Feature Engineering:**

* Imputed missing delays as zero
* Removed extreme outliers (>99th percentile)
* Created new features: DelayRate, SeasonalQuarter, BusyAirportFlag, LogCarrierSize, and high\_delay\_rate (binary target)
* Log-transformed skewed delay variables
* Summary Statistics Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Std Dev** | **Min** | **Median** | **Max** |
| Arriving Flights | 210.3 | 386.4 | 1 | 86 | 4827 |
| DelayRate (min/flight) | 14.0 | 15.3 | 0 | 11.04 | 1017 |
| log\_DelayRate | 2.40 | 0.83 | 0 | 2.49 | 6.93 |
| Carrier Delay (minutes) | 4.82 | 7.20 | 0 | 2.48 | 314.3 |
| Weather Delay (minutes) | 1.06 | 3.62 | 0 | 0.18 | 142.9 |
| NAS Delay (minutes) | 3.63 | 4.98 | 0 | 2.15 | 152.2 |
| Security Delay (minutes) | 0.02 | 0.20 | 0 | 0.00 | 8.83 |
| Late Aircraft Delay (min) | 4.72 | 6.96 | 0 | 2.67 | 258.5 |
| Cancelled (%) | 1.12 | 4.42 | 0 | 0 | 100 |
| BusyAirportFlag (0 or 1) | 0.43 | 0.49 | 0 | 0 | 1 |
| LogCarrierSize | 4.05 | 1.42 | 0 | 4.44 | 8.48 |
| SeasonalQuarter (1 to 4) | — | — | 1 | 2 | 4 |

**Estimating Equations:**

1. **Simple Regression Model (SAS):**

log\_DelayRate = 1.10478 + 0.23526(log\_carrier\_delay)

1. **Multiple Regression Model (SAS):**

log\_DelayRate = -0.261300 + log\_carrier\_delay(0.219498) + log\_weather\_delay(0.070709) + log\_nas\_delay(0.024724) + log\_security\_delay(-0.021180) + log\_late\_aircraft\_de(0.090546) + SeasonalQuarter Q1(0.121531) + SeasonalQuarter Q2(0.224877) + SeasonalQuarter Q3(0.202600) + SeasonalQuarter Q4(0) + BusyAirportFlag 0\*(0.869162) + BusyAirportFlag 1\* (0) + LogCarrierSize Big(-0.954414) + LogCarrierSize Medium(-0.341391) + LogCarrierSize Small(0)

1. **Logistic Model (R):**

high\_delay\_rate = -2.955500 + log\_carrier\_delay(0.185737) + log\_weather\_delay(0.241371) + log\_nas\_delay (-0.094861) + log\_security\_delay (-0.126622) + log\_late\_aircraft\_delay(0.187095)

**4. Random Forest Model (R):**

* + Six predictors used: log-transformed delay types
  + Optimal parameters: ,
  + No equation due to non-parametric ensemble model

**Target Variable Clarification:**

* **Regression Models:** Target = log\_DelayRate (continuous outcome)
* **Classification Models (Logistic, RF):** Target = high\_delay\_rate (binary outcome)
  + Derived from top 35% of delay rate distribution (i.e., above 65th percentile)
  + Classification based on a 0.4 probability threshold

**Why Both Thresholds?**

* The **65th percentile** is used to define the binary target variable
* The **0.4 probability threshold** is used to classify new observations based on model output

**Methodological Justification:**

* **Regression models** estimate effect sizes and direction of each delay type on average delay rate
* **Clustering** identifies structural patterns across routes and seasons
* **Logistic regression** offers interpretable classification of high delay risk
* **Random forest** captures nonlinear interactions and delivers strong predictive performance

This multi-method approach enhances both interpretation and prediction, with regression identifying causality and machine learning boosting forecasting power on unseen data.

# 4.3 Results

**Descriptive Analytics**

To explore the structure and variation in delay-related features, several descriptive analyses were conducted:

* **Summary Statistics (PROC MEANS)**: Log-transformed variables such as log\_arr\_delay, log\_carrier\_delay, and log\_late\_aircraft\_delay had relatively high means, indicating their frequent and significant contribution to overall delays. In contrast, log\_arr\_cancelled and log\_arr\_diverted had low means and standard deviations, confirming these are rare but impactful events.
* **PROC SUMMARY by Major Delay Cause**:
  + Late Aircraft delays had the highest average delay (≈7.34) and standard deviation (≈1.44), indicating significant and variable disruptions.
  + Carrier-caused delays followed closely, with a mean of ≈6.74.
  + NAS delays exhibited the highest variability (≈1.8 std dev), likely due to unpredictable airspace congestion.
* **Rare Event Analysis (PROC UNIVARIATE)**:
  + log\_arr\_cancelled and log\_arr\_diverted had heavy right tails (high skewness), highlighting rare but severe disruptions.
* **Correlation Analysis**:
  + The Pearson correlation between log\_DelayRate and log\_carrier\_delay was **0.6021**, statistically significant at p < 0.0001, showing a moderately strong positive relationship.
* **ANOVA Results**:
  + SeasonalQuarter had statistically significant effects on all delay variables (p < 0.0001), though R² values were modest (∼0.2% to 3.4%).
  + The highest seasonal variation was found in log\_DelayRate (R² = 3.4%) and log\_weather\_delay (R² = 1.76%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | F-Value | p-Value | R² | Significant? |
| log\_arr\_delay | 74.64 | <.0001 | 1.01% | Yes |
| log\_DelayRate | 256.76 | <.0001 | 3.4% | Yes |
| log\_carrier\_delay | 64.67 | <.0001 | 0.88% | Yes |
| log\_weather\_delay | 130.72 | <.0001 | 1.76% | Yes |
| log\_nas\_delay | 70.11 | <.0001 | 0.95% | Yes |
| log\_security\_delay | 16.30 | <.0001 | 0.22% | Yes |
| log\_late\_aircraft\_delay | 73.92 | <.0001 | 1.0% | Yes |

**Clustering Analysis**

* **Ward’s Hierarchical Clustering**: Identified 4 distinct clusters of U.S. airports based on delay metrics (avg\_income, population, minority).
  + Cluster profiles indicated regional socio-economic patterns with differences in delay behavior.

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generated

* **K-Means Clustering**: Validated hierarchical results, reinforcing cluster interpretations.

A screenshot of a computer

Description automatically generated

**Regression Models**

**Regression Results:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | R-Square | RMSE | MAE | MPE (%) | Key Predictors |
| Simple Regression | 0.3625 | 0.66654 | 0.52070 | 23.48 | log\_carrier\_delay |
| Multiple Regression | 0.5181 | 0.57959 | 0.4430152 | 19.97 | log\_carrier\_delay, log\_weather\_delay, log\_nas\_delay, log\_security\_delay, log\_late\_aircraft\_delay, SeasonalQuarter, BusyAirportFlag, LogCarrierSize |

* The multiple regression model provided a substantial improvement in fit over the simple regression.
* The coefficients suggest carrier delay and late aircraft delay have the strongest impact on log\_DelayRate.
* So, we have decided to select Multiple Regression Model for modelling the target variable log\_DelayRate.

**Classification Models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | Accuracy | AUC | Sensitivity | Specificity | Key Predictors |
| Logistic Model | 70.2% | 0.762 | 72.9% | 65.3% | log\_carrier\_delay, log\_weather\_delay, log\_nas\_delay, log\_security\_delay, log\_late\_aircraft\_delay |
| Random Forest | 74.6% | 0.805 | 60.6% | 82.1% | Same as above (nonlinear ensemble model); Optimal mtry = 1, ntree = 897 |

* The random forest model delivered the best performance overall, with the highest accuracy and AUC.
* Logistic regression offers interpretable coefficients, while random forest captures nonlinear interactions for improved prediction.

**Therefore, we have decided to select the Random Forest Classification Model for modelling the target variable high\_delay\_rate.**

**Justification:**

* **Higher accuracy (74.6%)** and **better AUC (0.805)** indicate stronger overall classification performance.
* **Higher specificity (82.1%)** makes it more reliable in identifying flights **not likely to experience high delays**, helping avoid false alarms.
* **Nonlinear relationships** and **interactions** are better captured with random forest than with logistic regression.
* While logistic regression has higher sensitivity (better at detecting actual high delays), the overall balance and robustness of Random Forest make it superior for predictive performance.

A graph of a logistic model

Description automatically generated

A graph of a curve

Description automatically generated

# 4.4 Summary of Project

This project set out to identify key factors contributing to airline delays and to model delay outcomes using both explanatory and predictive methods. Using 2024 data from the U.S. Bureau of Transportation Statistics, we performed detailed descriptive analytics, clustering, regression modeling, logistic regression, and random forest classification.

**Findings Summary:**

* Descriptive statistics showed that late aircraft, carrier, and weather delays dominate the delay profile.
* Regression analysis confirmed the strong influence of these delay sources, with multiple regression achieving an R² of 0.682.
* Logistic regression accurately classified high-delay events with an AUC of 0.762.
* Random forest classification further improved performance, achieving 74.6% accuracy and an AUC of 0.805.

**Key Strengths:**

* The integration of statistical and machine learning techniques enhanced interpretability and predictive accuracy.
* Clustering added value by identifying structural groupings across airports based on delay behavior and demographic features.

**Looking ahead, future work could focus on:**

* **Incorporating flight-level data** to capture micro-level patterns and individual airline behaviors
* **Integrating external real-time data sources,** such as NOAA weather feeds, FAA airspace advisories, and live gate occupancy signals
* **Testing more advanced algorithms**, such as Gradient Boosted Trees or Deep Neural Networks, to capture deeper interactions across temporal and spatial features
* **Developing operational dashboards** for airports and airlines to visualize and act on delay forecasts

In summary, this project demonstrates how data-driven techniques can transform raw delay records into actionable insights, empowering aviation stakeholders to reduce disruptions, optimize resource allocation, and improve the passenger experience.

# 4.5 Bibliography

- U.S. Bureau of Transportation Statistics, Airline On-Time Performance Data (2024)  
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning.  
- R Documentation for randomForest, glm, caret, pROC packages

# 4.6 Appendix: SAS or R Commands and Data Files

The following files were used in this project:  
- airlinedata\_final\_log.csv (cleaned dataset)  
- Airline\_Delay\_Cause.csv (raw dataset)  
- ProjectCode\_SAS.sas (descriptive and ANOVA analysis in SAS)  
- R script (logistic regression and random forest modeling)

Please see the accompanying SAS code file and R code file submitted with this report in the .txt format.