Flight Delay Analysis: Project Analysis Report

"Comprehensive Analysis to Classify and Predict Airline Delays Based on Various Operational and Weather Factors."

Professor:

Professor Reza Maihami

Group 3 Members:

Shiva Siddharth Yedla

Satvik Reddy Kommidi

Sasi Kumar Karumuri

Introduction

Flight delays represent one of the most significant challenges faced by the aviation industry today. These disruptions have far-reaching consequences that affect both airlines and passengers. Delays not only disrupt airline operations but also lead to diminished customer satisfaction, erosion of trust, and substantial financial losses for carriers. Moreover, the cascading nature of delays across interconnected flight schedules exacerbates inefficiencies within the system. Passengers, who prioritize reliability and punctuality in their travel plans, often bear the brunt of these delays, leading to frustration and a potential decline in loyalty. Understanding the root causes of these delays is, therefore, a pivotal step toward mitigating their impact and enhancing the overall efficiency of the aviation industry.

From the perspective of marketing analytics and operational research, studying flight delays provides a unique and valuable opportunity to address operational inefficiencies while simultaneously improving customer experience. Delays are often influenced by a combination of factors, including adverse weather conditions, airport congestion, logistical bottlenecks, and technical issues. By leveraging data-driven analysis, airlines can identify patterns and correlations among these factors, enabling them to optimize scheduling practices, allocate resources more effectively, and design robust contingency plans to address potential disruptions. Such insights not only help in minimizing the occurrence and duration of delays but also equip airlines with the foresight to proactively manage challenges before they escalate into significant operational issues.

In today's increasingly competitive aviation market, on-time performance has become a critical differentiator. Airlines that consistently deliver reliability gain a competitive edge by fostering customer trust and loyalty. By identifying and analyzing the primary causes of delays, carriers can strengthen their operational capabilities, reduce inefficiencies, and improve their punctuality metrics. This, in turn, enhances their reputation as reliable service providers, creating a virtuous cycle of customer retention and long-term profitability. Furthermore, addressing delays comprehensively contributes to broader strategic objectives, such as sustainable growth, operational excellence, and market competitiveness.

The findings from this study aim to provide actionable recommendations for airlines, empowering them to make informed, evidence-based decisions. These decisions can significantly minimize disruptions, streamline operations, and deliver a superior travel experience for passengers. For instance, improved forecasting and resource management strategies can help airlines adapt to changing conditions with agility. Additionally, effective collaboration with airport authorities and other stakeholders can address systemic issues that contribute to delays.

Ultimately, tackling flight delays requires a comprehensive and integrated approach that aligns with the aviation industry's broader goals of efficiency and reliability. By leveraging advanced analytics and operational insights, airlines can transform challenges into opportunities for growth and differentiation. Addressing flight delays is not just about improving punctuality; it is about redefining the travel experience and maintaining a competitive edge in a dynamic and ever-evolving industry.

Literature Review / State of Current Business Knowledge

In the aviation industry, flight delays are a persistent challenge that affect operational efficiency, customer satisfaction, and profitability. The existing literature highlights several dimensions of flight delay analytics, including causes, patterns, and predictive modeling. These insights are critical for stakeholders such as airlines, airports, regulators, and customers.

Key Factors Influencing Flight Delays

1. **Operational Factors**: Research has identified carrier-related issues, including late aircraft turnover and crew scheduling conflicts, as primary contributors to delays. These factors are often interdependent, amplifying disruptions across the network.

2. **Weather Conditions**: Adverse weather, including precipitation, snow, and wind, has been extensively studied for its direct impact on delays. Predictive weather analytics is an emerging field aiming to mitigate such disruptions.

3. **Airport Congestion**: High passenger volumes and limited runway or gate availability create bottlenecks, particularly at hub airports. Studies suggest optimizing scheduling algorithms to alleviate congestion.

Predictive Modeling and Analytics

Recent advancements in data analytics have enabled the development of predictive models to anticipate flight delays. Machine learning techniques, such as decision trees and neural networks, are increasingly applied to large datasets, including operational logs, weather patterns, and historical delay data. Models focusing on binary classification (e.g., predicting delays exceeding 15 minutes) align closely with the datasets described, such as the "DEP_DEL15" variable in the provided documentation.

Business Applications

- **Customer Satisfaction**: Airlines use delay predictions to enhance communication with passengers, offering real-time updates and alternative options.
- **Operational Efficiency**: Predictive insights inform resource allocation, such as crew management and gate assignments, to minimize cascading delays.
- **Revenue Management**: Understanding delays helps airlines quantify missed opportunities, improve revenue forecasting, and design dynamic pricing strategies for premium services.

Gaps in Knowledge

While substantial progress has been made, gaps remain in:

1. **Integration of Real-Time Data**: Combining real-time weather updates, traffic congestion, and dynamic operational data remains a challenge.
2. **Customer Behavior Analysis**: Few studies directly link delay data with customer loyalty or purchasing behavior.
3. **Cross-Industry Benchmarking**: Lessons from logistics and public transportation industries could further enhance delay management strategies.

The state of knowledge suggests that data-driven approaches to flight delay analysis can significantly improve operational and marketing outcomes. The datasets provided, which include variables related to delays, weather, and airport activity, offer a robust foundation for advancing these efforts.

Would you like to expand on any specific aspect, such as predictive modeling or customer behavior implications?

Data, Data Sources, and Data Characteristics

The analysis of flight delays requires diverse and comprehensive datasets to explore delay patterns, identify contributing factors, and derive actionable insights. Below is an overview of the relevant data sources, their characteristics, and their significance to this study.

Data Sources

1. **ONTIME_REPORTING**
    - **Description**: Contains detailed information about flights, including departure and arrival times, delays, and cancellations.
    - **Key Variables**:
        - DEP_DELAY_NEW: Departure delay in minutes.
        - DEP_DEL15: Binary target variable indicating delays over 15 minutes.
        - ARR_DELAY_NEW: Arrival delay in minutes.
        - CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY, LATE_AIRCRAFT_DELAY: Specific delay reasons.
        - DISTANCE: Flight distance in miles.
        - CANCELLED: Flag indicating flight cancellations.
        - DEP_TIME_BLK, ARR_TIME_BLK: Time blocks for departures and arrivals.
    - **Importance**: Central to analyzing delay causes and predicting future delays.

2. **AIRPORT_COORDINATES**
    - **Description**: Includes geographical data about origin and destination airports.
    - **Key Variables**:
        - LATITUDE and LONGITUDE: Airport geographic coordinates.
        - DISPLAY_AIRPORT_NAME: Airport name for matching with other datasets.
    - **Importance**: Enables spatial analysis and integration with weather data.
3. **B43_AIRCRAFT_INVENTORY**
    - **Description**: Details aircraft characteristics such as capacity and age.
    - **Key Variables**:
        - MANUFACTURE_YEAR: Aircraft production year.
        - NUMBER_OF_SEATS: Aircraft seating capacity.
    - **Importance**: Facilitates analysis of aircraft age and size impact on delays.
4. **AIRPORT_WEATHER**
    - **Description**: Provides weather data such as precipitation and wind speed.
    - **Key Variables**:
        - PRCP: Precipitation in inches.
        - SNOW: Snowfall in inches.
        - AWND: Maximum wind speed.
    - **Importance**: Essential for evaluating weather-related delay factors.
5. **T3_AIR_CARRIER_SUMMARY_AIRPORT_ACTIVITY**
    - **Description**: Summarizes airline operational data, such as passenger volume.

- **Key Variables**:
  - REV_PAX_ENP_110: Passengers enplaned annually.
  - REV_ACRFT_DEP_PERF_510: Total flights performed annually.
- **Importance**: Provides context on airline and airport performance.

6. **P10_EMPLOYEES**
   - **Description**: Data on airline staffing levels and roles.
   - **Key Variables**:
     - PILOTS_COPILOTS: Number of pilots and copilots.
     - MAINTENANCE: Maintenance staff.
     - PASSENGER_HANDLING: Passenger handling personnel.
   - **Importance**: Links human resource allocation to operational delays.

Data Characteristics

Dataset	Number of Variables	Key Variable Types	Frequency	Granularity
ONTIME_REPORTING	30+	Numerical, Categorical	Daily	Flight-level
AIRPORT_COORDINATES	5	Numerical, Text	Static	Airport-level
B43_AIRCRAFT_INVENTORY	3	Numerical	Static	Aircraft-level

AIRPORT_WEATHER	6	Numerical	Daily	Airport-level
T3_AIR_CARRIER_ACTIVITY	4	Numerical	Annual	Airline-level
P10_EMPLOYEES	10+	Numerical	Annual	Airline-level

Data Integration and Utility

1. **Joining Key**: Many datasets share common identifiers such as ORIGIN_AIRPORT_ID, DEST_AIRPORT_ID, OP_UNIQUE_CARRIER, and TAIL_NUM. These keys allow seamless integration.
2. **Temporal Coverage**: ONTIME_REPORTING provides day-level granularity, while other datasets like AIRPORT_WEATHER ensure the incorporation of dynamic environmental factors.
3. **Target Variable**: The binary DEP_DEL15 variable enables classification-based predictive modeling, making it a focal point for analysis.

Limitations and Challenges

1. **Data Completeness**: Missing data in weather or flight performance metrics may limit analysis accuracy.
2. **Data Size**: High granularity leads to computational challenges, requiring robust storage and processing capabilities.
3. **Time Lag**: Disparities in data collection frequencies necessitate temporal alignment for integration.

This rich dataset combination facilitates a comprehensive analysis of flight delays, paving the way for actionable insights in operational optimization and customer satisfaction improvement.

Here's an outline of the **Data Dictionary** for your airline delay dataset, based on the provided information:

Data Dictionary

This dictionary provides descriptions of the variables used in the analysis. Each variable includes its name, unit, a brief description, and relevant details.

Dataset Information

- **Dataset Source:** Airline delay data (January–June)
- **Number of Records:** 3,156,348 (after cleaning: 2,814,786)
- **Number of Variables:** 26 (after selection: 10)

Variable Descriptions

1. **PLANE_AGE**
   - **Unit:** Years
   - **Description:** The age of the aircraft at the time of flight.
   - **Granularity:** Integer values representing the aircraft's age.
   - **Transformation/Notes:** None.
2. **CONCURRENT_FLIGHTS**
   - **Unit:** Count
   - **Description:** Number of flights departing concurrently from the same airport.
   - **Granularity:** Integer values.
   - **Transformation/Notes:** Outliers removed based on IQR.
3. **PRCP**
   - **Unit:** Inches
   - **Description:** Precipitation level at the departing airport on the day of the flight.

- **Granularity:** Numeric values (with decimal precision).
- **Transformation/Notes:** Outliers removed based on IQR.

4. **DISTANCE_GROUP**
   - **Unit:** Ordinal group
   - **Description:** Distance range of the flight, categorized into groups (e.g., short-haul, medium-haul, long-haul).
   - **Granularity:** Integer values from 1 to 11.
   - **Transformation/Notes:** Outliers removed based on IQR.
5. **AVG_MONTHLY_PASS_AIRLINE**
   - **Unit:** Count
   - **Description:** Average monthly passengers served by the airline.
   - **Granularity:** Integer values.
   - **Transformation/Notes:** None.
6. **DEP_DEL15**
   - **Unit:** Binary (0 or 1)
   - **Description:** Indicator of whether the flight was delayed by 15 minutes or more.
   - **Granularity:**
     - 0 = Not delayed
     - 1 = Delayed
   - **Transformation/Notes:** Target variable for prediction.
7. **DEP_TIME_BLK**
   - **Unit:** Time block

  - **Description:** Scheduled departure time grouped into hourly blocks (e.g., "0600-0659").
  - **Granularity:** Categorical variable.
  - **Transformation/Notes:** Recoded for modeling purposes.

8. **DAY_OF_WEEK**
   - **Unit:** Ordinal
   - **Description:** Day of the week the flight occurred.
   - **Granularity:** Integer values from 1 (Monday) to 7 (Sunday).
   - **Transformation/Notes:** None.
9. **CARRIER_NAME**
   - **Unit:** Airline Name
   - **Description:** The name of the airline operating the flight.
   - **Granularity:** Categorical variable (e.g., "Southwest Airlines Co.").
   - **Transformation/Notes:** Encoded for modeling.
10. **AIRPORT_FLIGHTS_MONTH**
    - **Unit:** Count
    - **Description:** Total flights departing from the airport in a month.
    - **Granularity:** Integer values.
    - **Transformation/Notes:** None.

Removed Variables (Initial Dataset)

These variables were excluded based on irrelevance or lack of predictive value for the delay model:

- **Geographic variables:** LATITUDE, LONGITUDE, DEPARTING_AIRPORT
- **Weather variables:** SNOW, SNWD, TMAX, AWND

- **Aircraft and operational details:** SEGMENT_NUMBER, NUMBER_OF_SEATS
- **Airline statistics:** AIRLINE_FLIGHTS_MONTH, AIRLINE_AIRPORT_FLIGHTS_MONTH

Transformation Summary

- **Outliers Removed:** Using the IQR method for numeric variables (e.g., PRCP, CONCURRENT_FLIGHTS, etc.).
- **Encoding:**
    - DEP_TIME_BLK and CARRIER_NAME converted to categorical encoding.
    - DEP_DEL15 converted to a factor for logistic regression and classification models.

Granularity

- **Temporal:** Dataset spans six months, with granularity down to individual flights.
- **Geographical:** Focused on U.S. domestic flights.

Data Quality

- **Missing Values:** Handled via na.omit().
- **Duplicates:** Removed using distinct().

Potential Enhancements

1. Add more weather variables (e.g., visibility, wind direction).
2. Include passenger load factor or seasonal demand patterns.
3. Aggregate historical delays per route or airline for additional predictive features.

Step 1: Load and Explore the Data

- Load the dataset and examine its structure, dimensions, and initial values.
- Identify missing values, duplicates, and data types.

```
# Load necessary libraries
```

```
library(dplyr)
library(ggplot2)

# Load data (update the path as required)
data <- read.csv("D:/datamining/Airline Delays/Dataset 17-flightdelay_Jan_Jun.csv")

# Display structure and summary
str(data)
summary(data)

# Check for missing values
colSums(is.na(data))

# Check for duplicate rows
duplicate_count <- sum(duplicated(data))
cat("Number of duplicate rows:", duplicate_count, "\n")
```

Step 2: Initial Variable Selection

We select variables based on domain knowledge and relevance to the target variable DEP_DEL15.

Retained Variables:

- **Predictors:** PLANE_AGE, CONCURRENT_FLIGHTS, PRCP, DISTANCE_GROUP, AVG_MONTHLY_PASS_AIRLINE, DEP_TIME_BLK, DAY_OF_WEEK, CARRIER_NAME, AIRPORT_FLIGHTS_MONTH.
- **Target Variable:** DEP_DEL15.

Removed Variables:

- **Geographic variables:** LATITUDE, LONGITUDE, DEPARTING_AIRPORT.
- **Weather specifics:** SNOW, SNWD, TMAX, AWND.
- **Operational details:** SEGMENT_NUMBER, NUMBER_OF_SEATS.
- **Aggregated statistics:** AIRLINE_FLIGHTS_MONTH, AIRLINE_AIRPORT_FLIGHTS_MONTH.

```
# Retain selected variables
selected_vars <- c(
  "PLANE_AGE", "CONCURRENT_FLIGHTS", "PRCP", "DISTANCE_GROUP",
  "AVG_MONTHLY_PASS_AIRLINE", "DEP_DEL15", "DEP_TIME_BLK",
  "DAY_OF_WEEK", "CARRIER_NAME", "AIRPORT_FLIGHTS_MONTH"
)


data_clean <- data %>% select(all_of(selected_vars))
```

Step 3: Handle Missing Values

- Identify and handle missing values (NA).
- Remove rows with missing data (na.omit).

```
# Check for missing values

missing_values <- colSums(is.na(data_clean))

print(missing_values)


# Remove missing values

data_clean <- na.omit(data_clean)
```

Step 4: Remove Duplicates

Duplicate rows can skew the analysis.

```
# Remove duplicate rows

data_clean <- data_clean %>% distinct()
```

Step 5: Visualize and Explore Relationships

Boxplots and Histograms

Use visualizations to explore the distribution and relationship of predictors with the target variable.

```
# List of predictors for visualization

predictors <- c("PLANE_AGE", "CONCURRENT_FLIGHTS", "PRCP", "DISTANCE_GROUP", "AVG_MONTHLY_PASS_AIRLINE")


# Boxplots and histograms

for (var in predictors) {

  # Boxplot

  print(

    ggplot(data_clean, aes_string(y = var)) +
    geom_boxplot(fill = "lightblue") +
    ggtitle(paste("Boxplot of", var)) +
    theme_minimal()
  )

  # Histogram
  print(
   ggplot(data_clean, aes_string(x = var)) +
    geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
    ggtitle(paste("Histogram of", var)) +
    theme_minimal()
  )
}
```

Target Variable Distribution

```
# Target variable distribution
ggplot(data_clean, aes(x = DEP_DEL15)) +
  geom_bar(fill = "orange", color = "black") +
  ggtitle("Distribution of DEP_DEL15") +
  theme_minimal()
```

Step 6: Outlier Detection and Removal

Use the **IQR method** to identify and remove outliers for numeric variables.

```
# Function to remove outliers
```

```
remove_outliers <- function(data, col) {
  Q1 <- quantile(data[[col]], 0.25, na.rm = TRUE)
  Q3 <- quantile(data[[col]], 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  data <- data[data[[col]] >= lower_bound & data[[col]] <= upper_bound, ]
  return(data)
}


# Apply outlier removal
for (var in c("PLANE_AGE", "CONCURRENT_FLIGHTS", "PRCP", "AVG_MONTHLY_PASS_AIRLINE")) {
  data_clean <- remove_outliers(data_clean, var)
}
```

Step 7: Encode Categorical Variables

Convert categorical variables like DEP_TIME_BLK and CARRIER_NAME into factor or one-hot encoding.

```
# Convert categorical variables to factors
data_clean$DEP_TIME_BLK <- as.factor(data_clean$DEP_TIME_BLK)
data_clean$CARRIER_NAME <- as.factor(data_clean$CARRIER_NAME)
data_clean$DEP_DEL15 <- as.factor(data_clean$DEP_DEL15)
```

Step 8: Verify Cleaned Dataset

Check the dimensions, structure, and summary of the cleaned dataset.

```
# Check dimensions

cat("Number of rows:", nrow(data_clean), "\n")

cat("Number of columns:", ncol(data_clean), "\n")


# Display structure and summary

str(data_clean)

summary(data_clean)
```

Step 9: Correlation Analysis

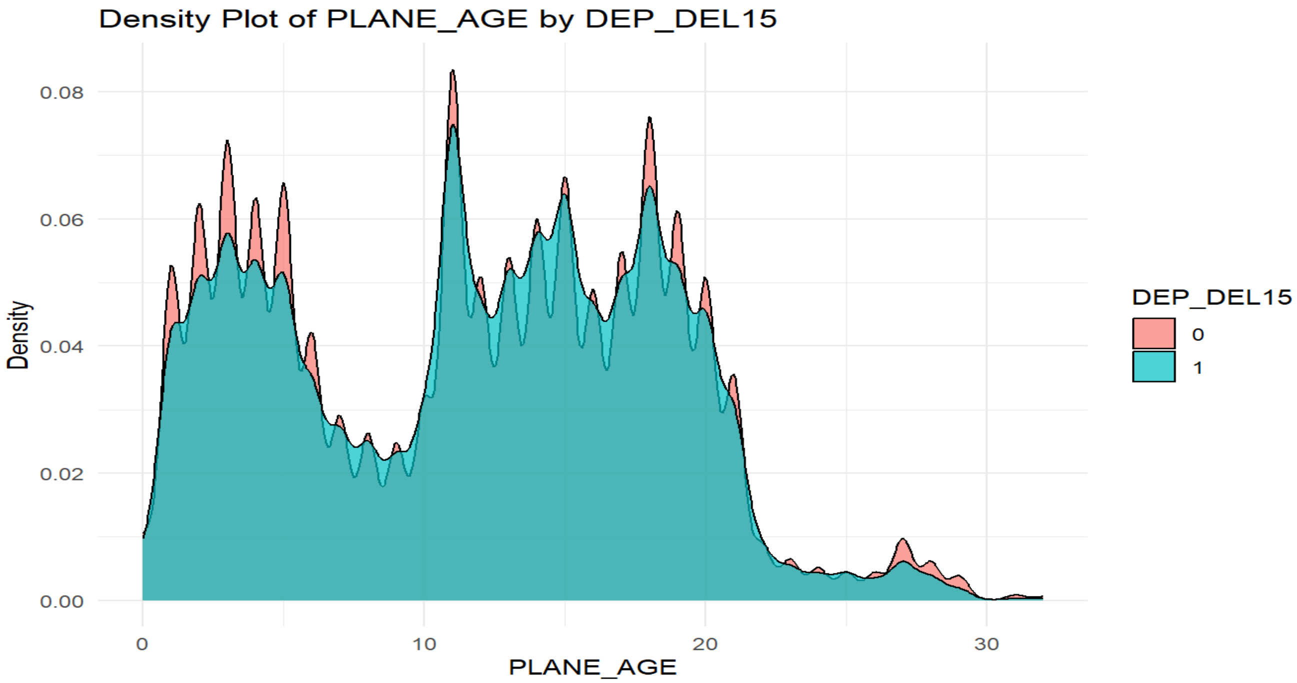
For numeric predictors, check correlations to identify highly correlated features.

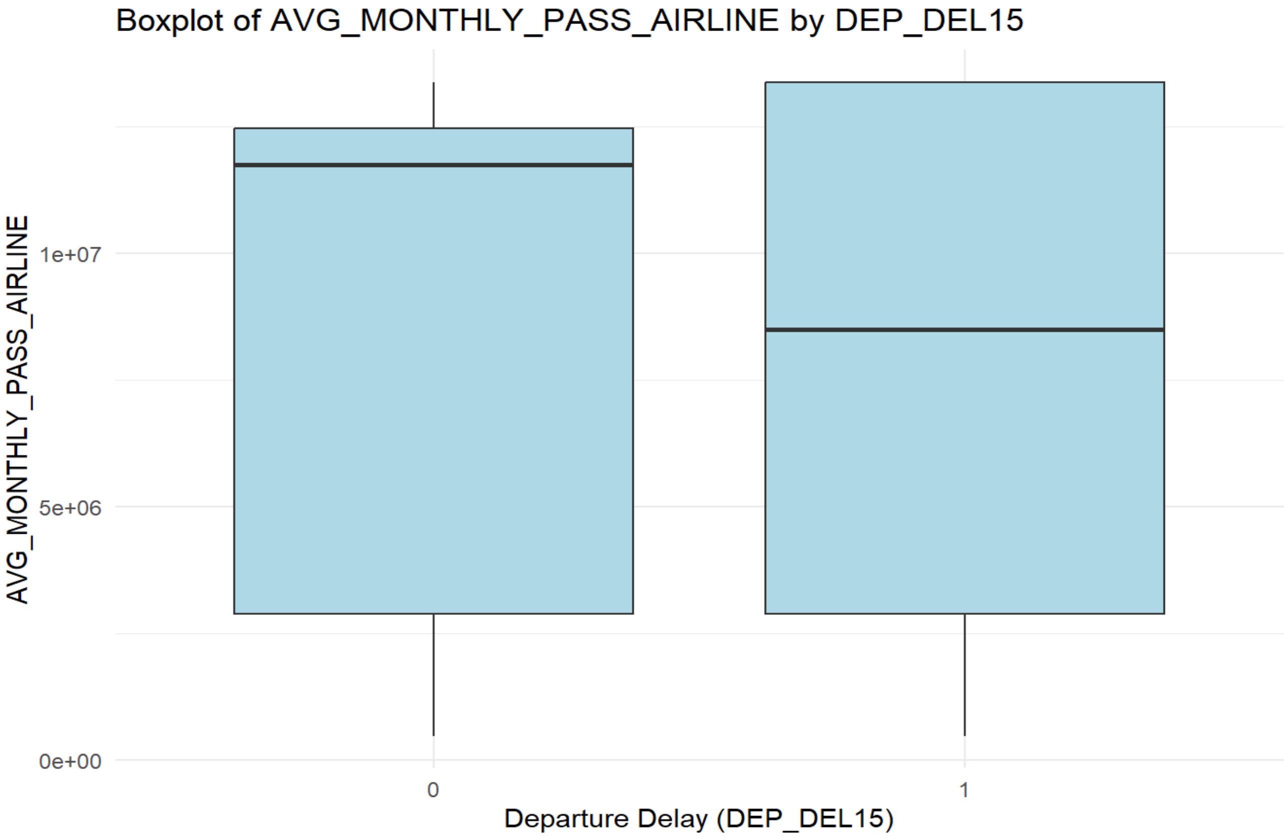
```
# Correlation matrix

numeric_vars <- data_clean %>% select(where(is.numeric))

cor_matrix <- cor(numeric_vars)

print(cor_matrix)
```

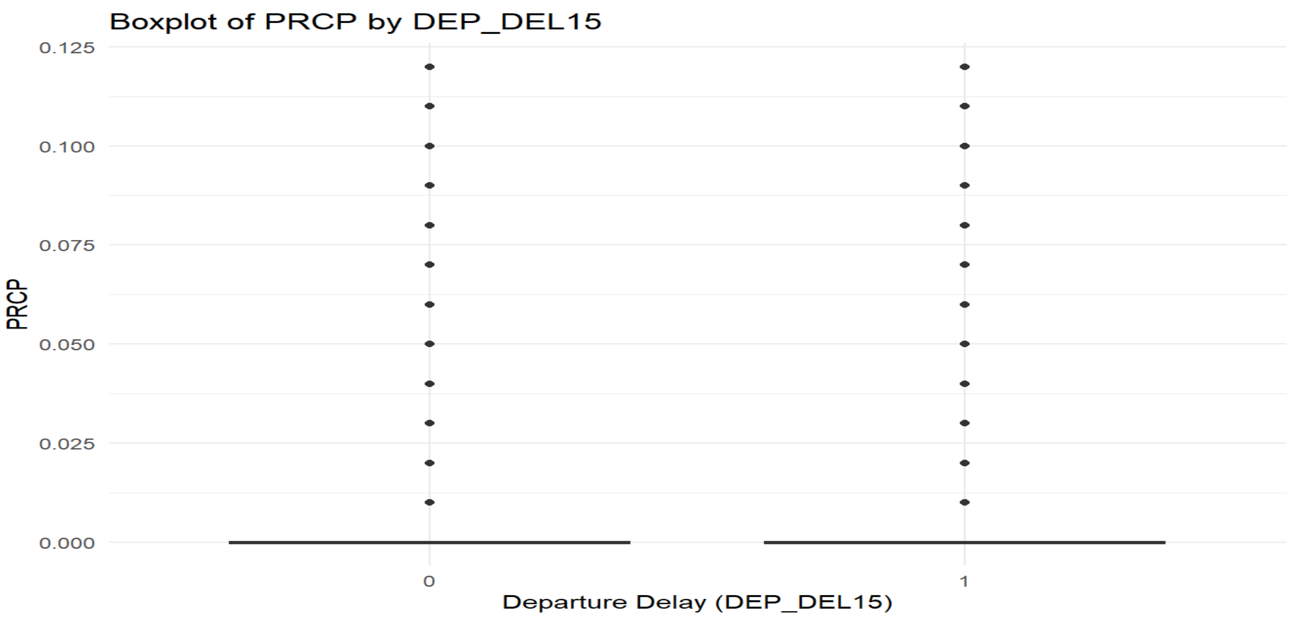
Outcome

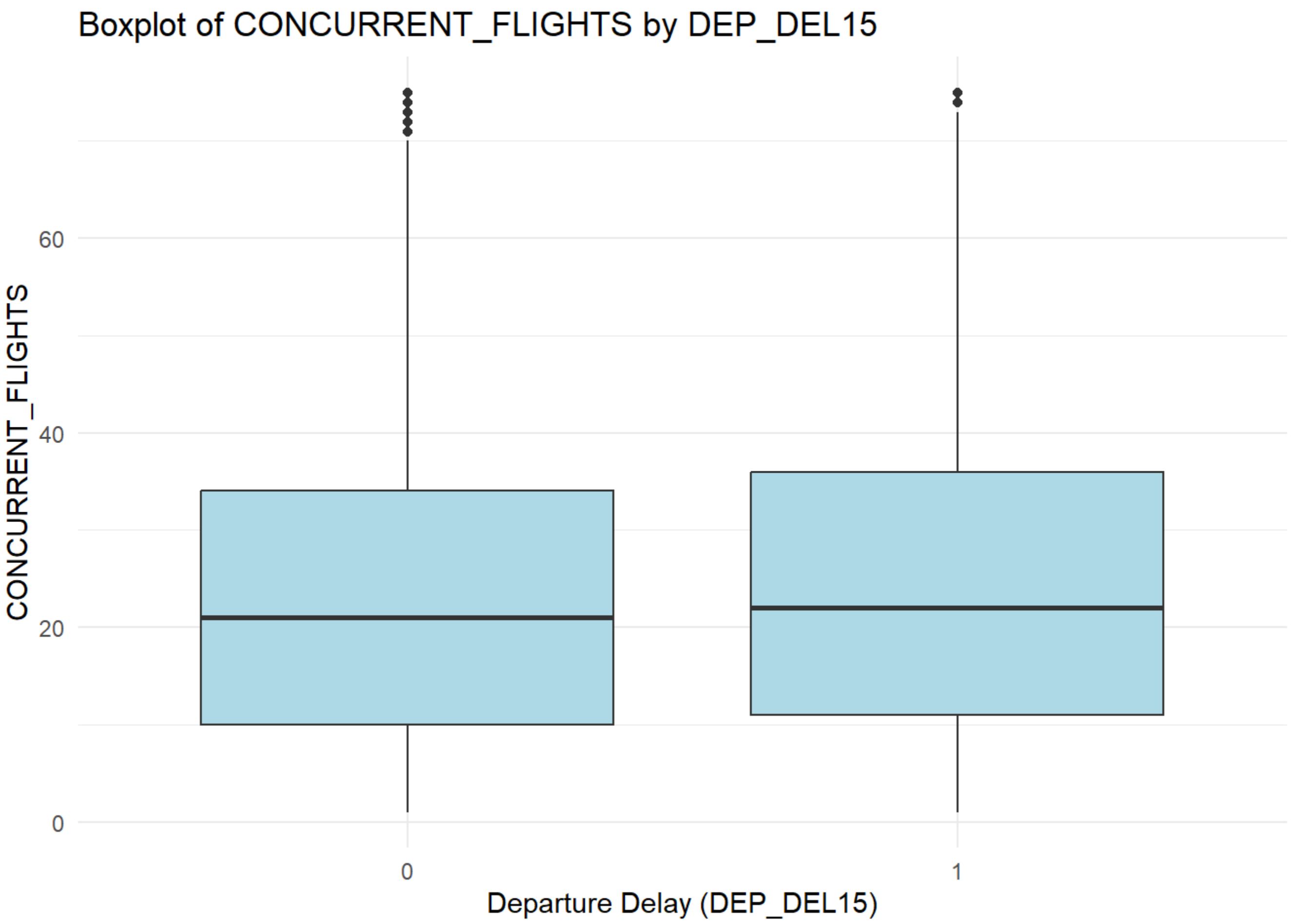
This process ensures:

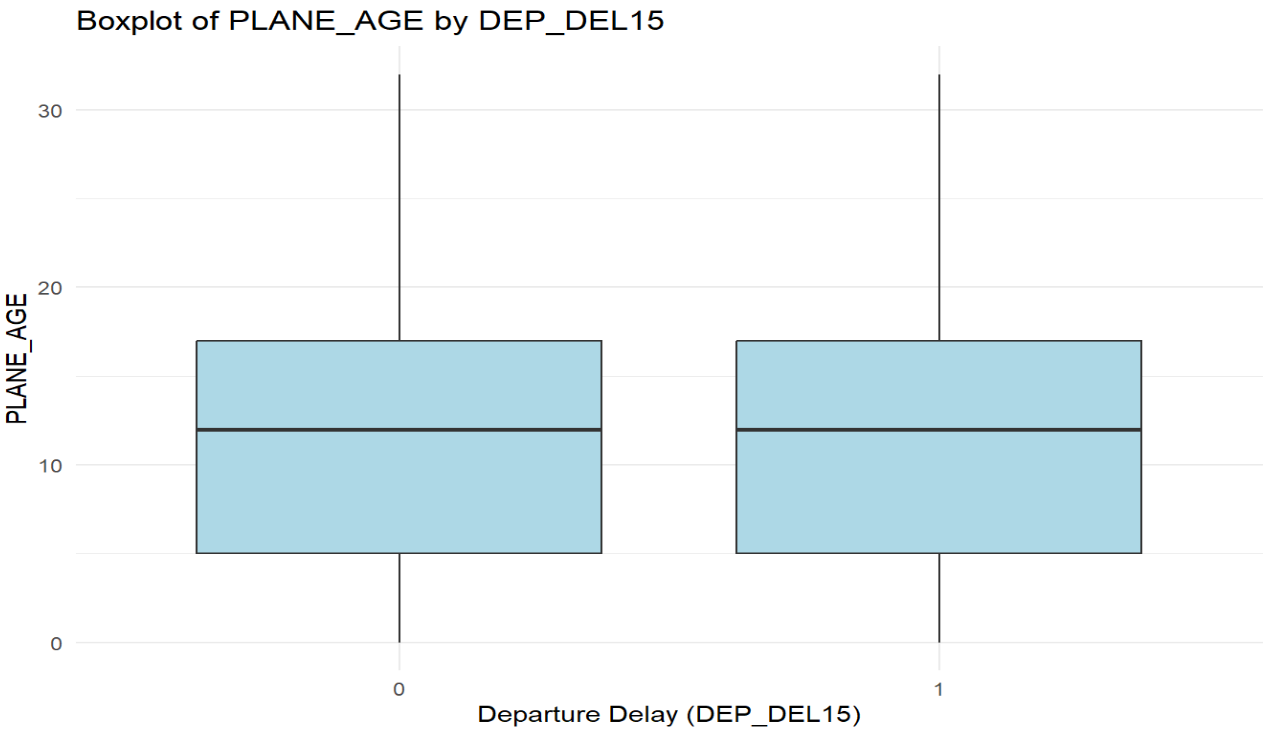
1. The dataset is clean and free of missing values or duplicates.
2. Outliers are addressed to minimize noise in the analysis.
3. Relationships are visualized to aid variable selection.
4. Data is ready for further predictive modeling (e.g., logistic regression, decision trees, or random forests).

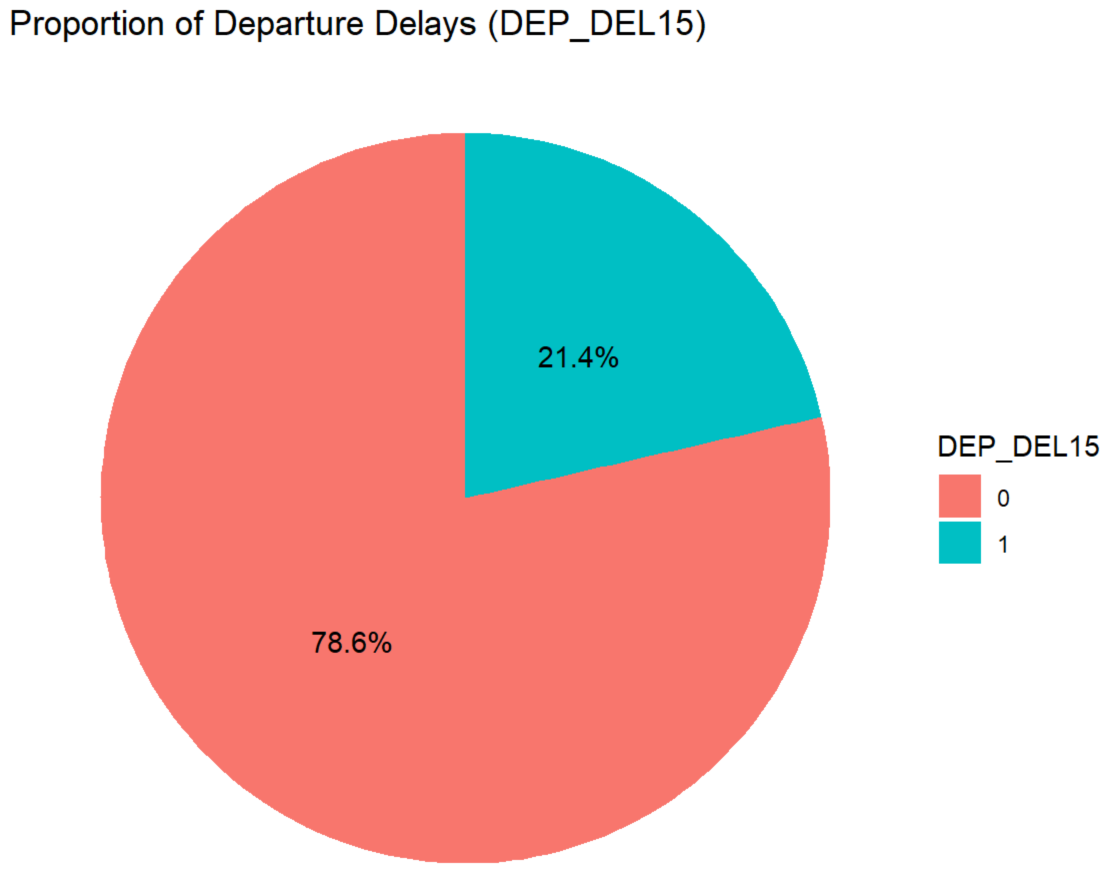

Density Plot of PLANE_AGE by DEP_DEL15

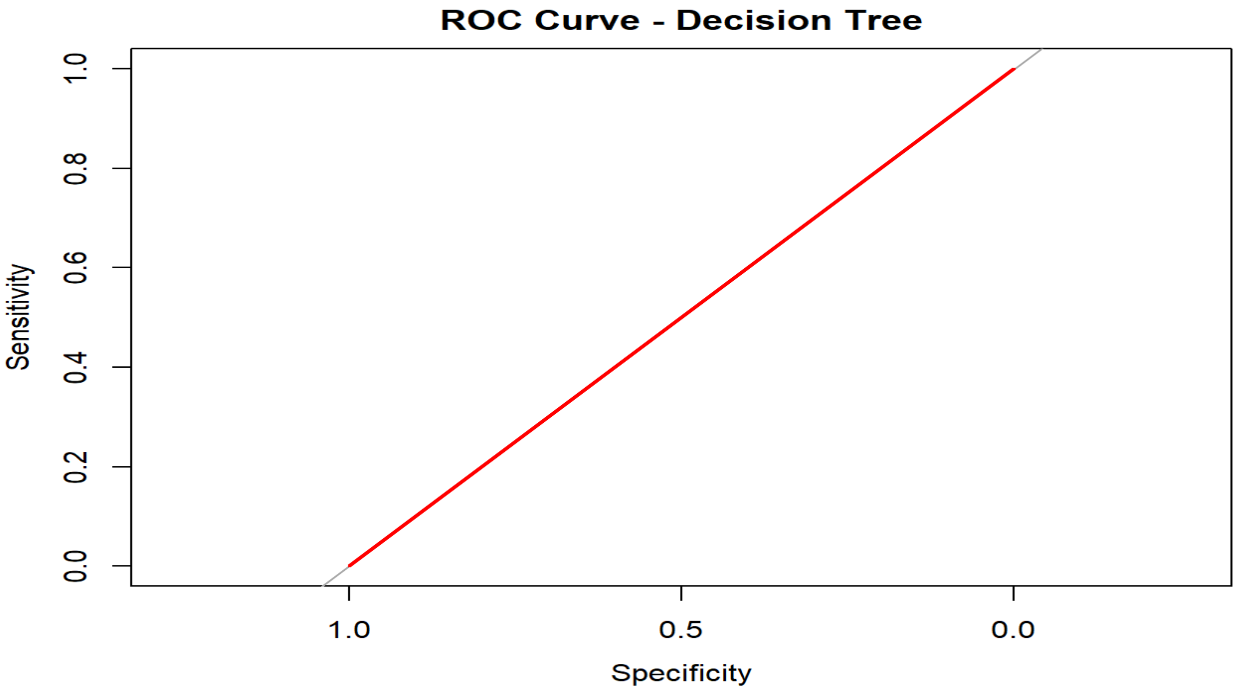

Boxplot of AVG_MONTHLY_PASS_AIRLINE by DEP_DEL15

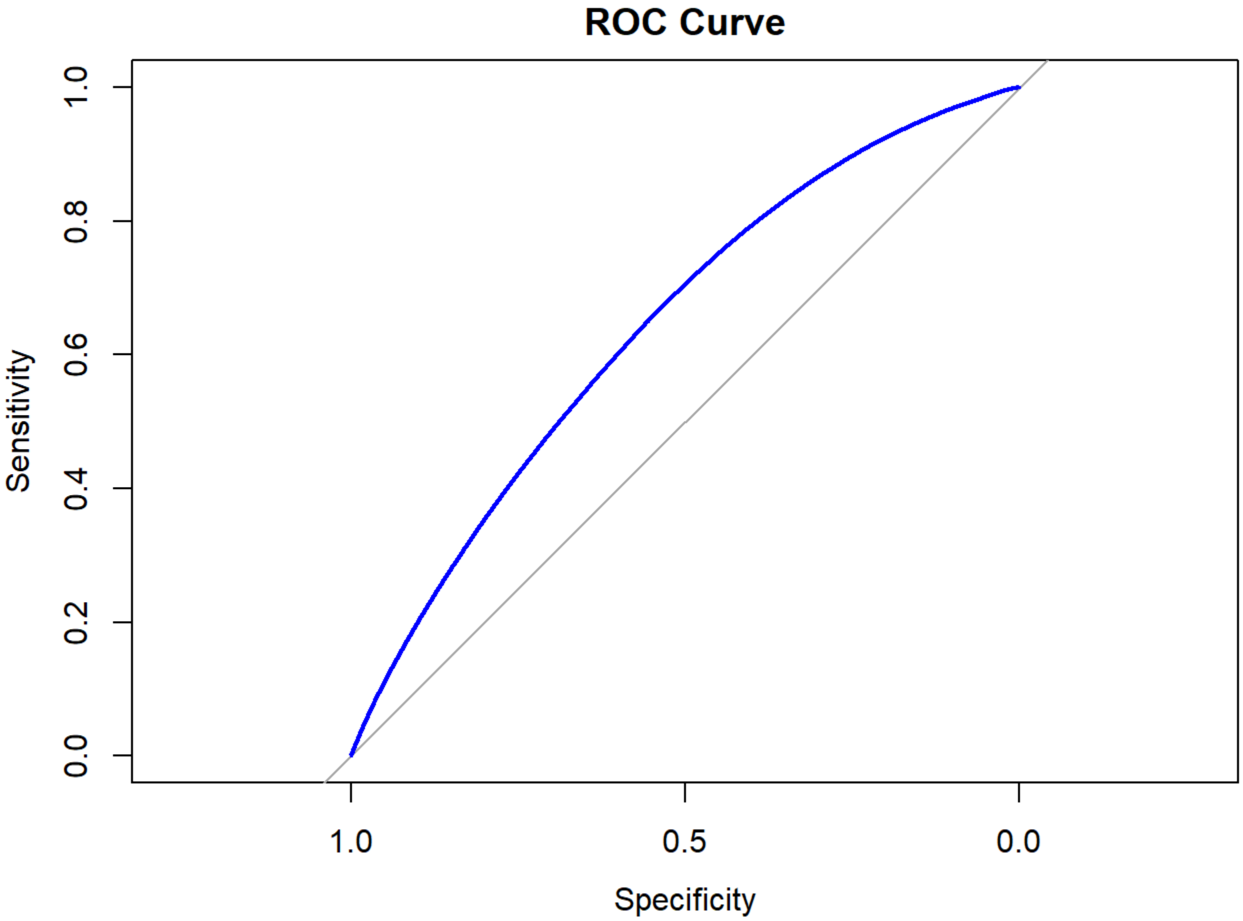

Boxplot of PRCP by DEP_DEL15
0.125
0.100
0.075
0.050
0.025
0.000
PRCP
0
1
Departure Delay (DEP_DEL15)

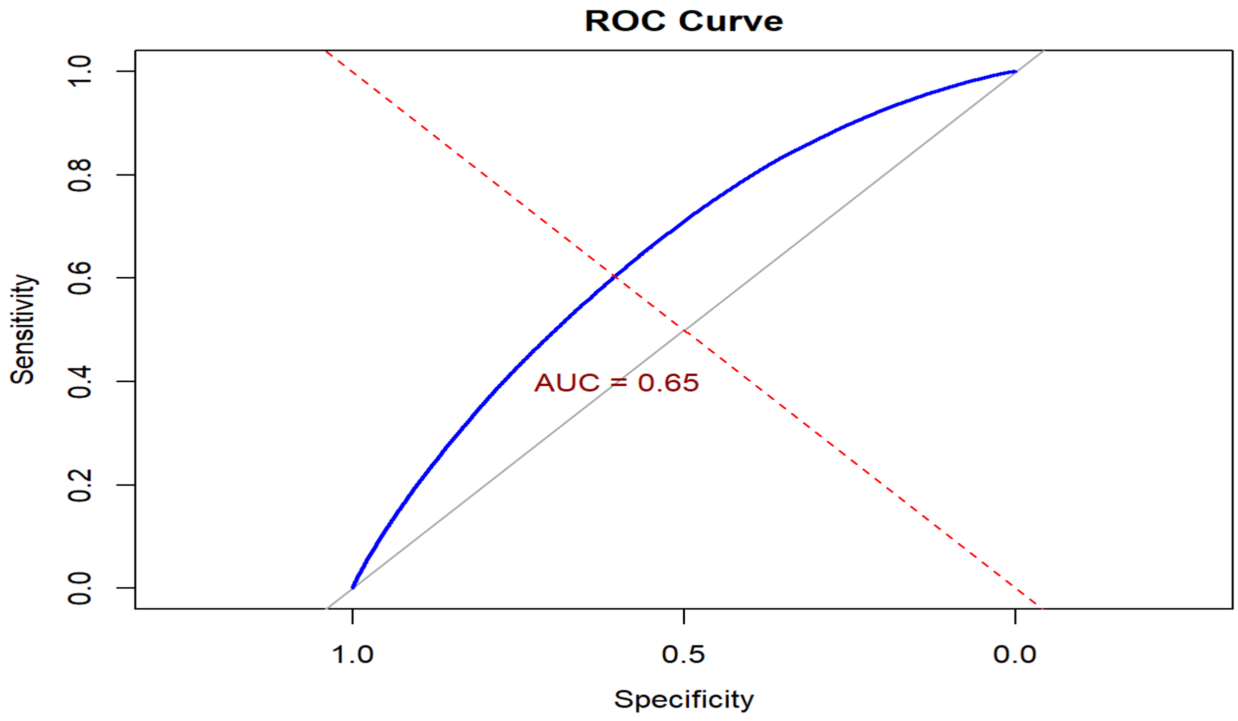

Boxplot of CONCURRENT_FLIGHTS by DEP_DEL15
60
40
20
0
CONCURRENT_FLIGHTS
0
1
Departure Delay (DEP_DEL15)

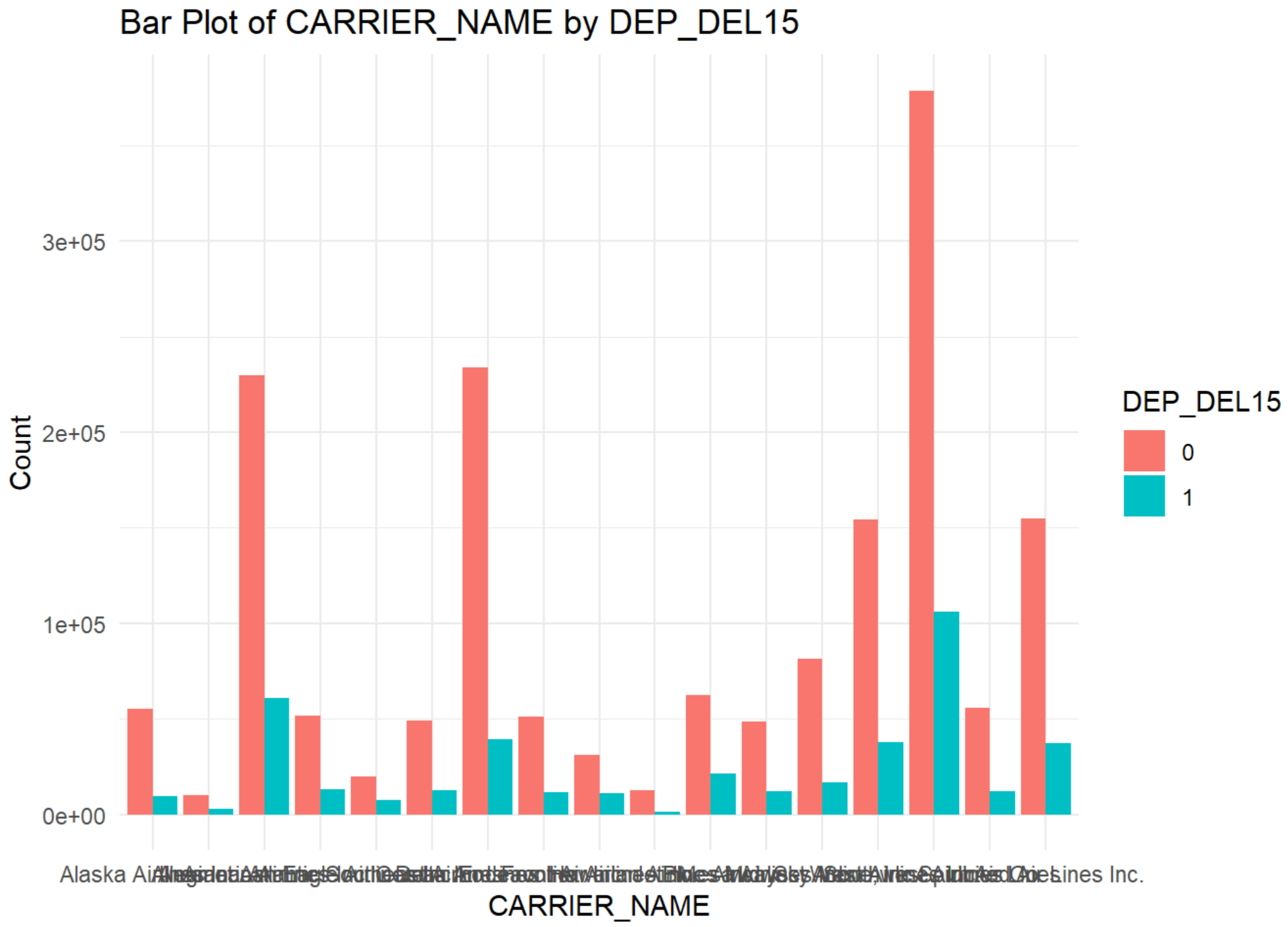

Boxplot of PLANE_AGE by DEP_DEL15
PLANE_AGE
30
20
10
0
0
1
Departure Delay (DEP_DEL15)

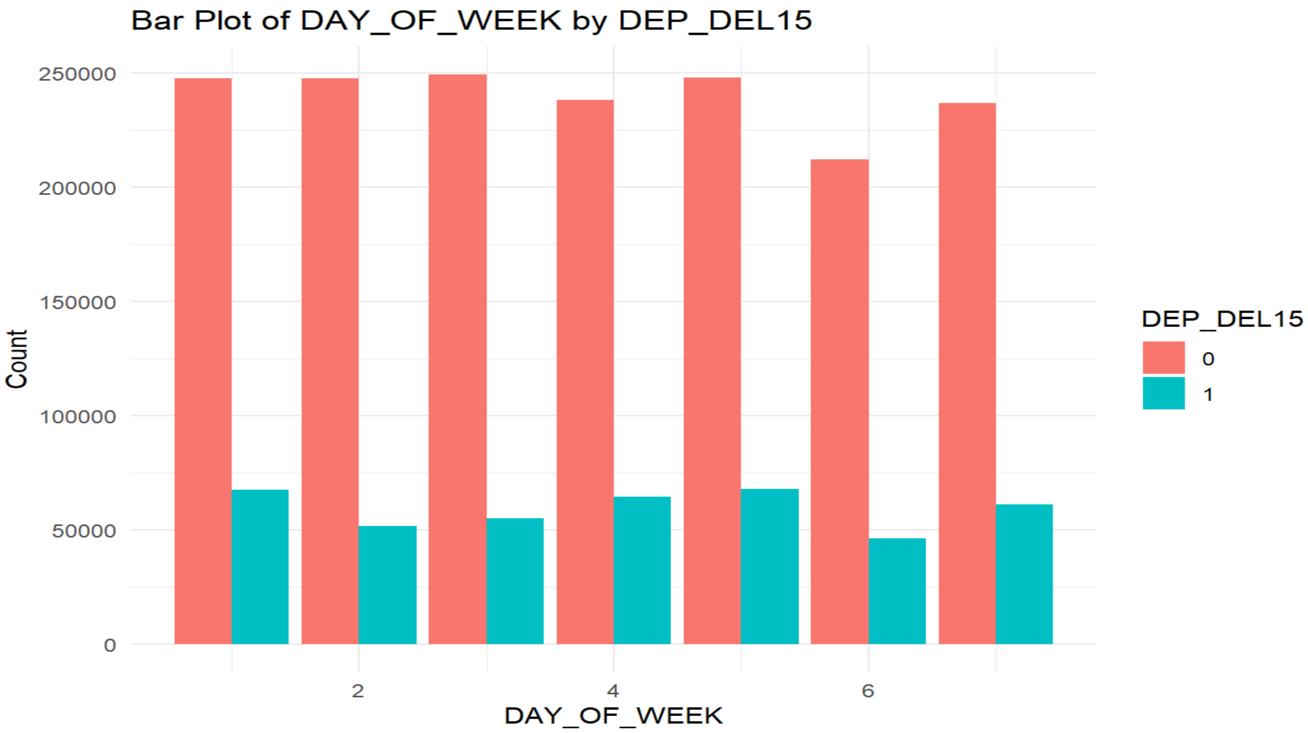

Proportion of Departure Delays (DEP_DEL15)
21.4%
78.6%
DEP_DEL15
0
1

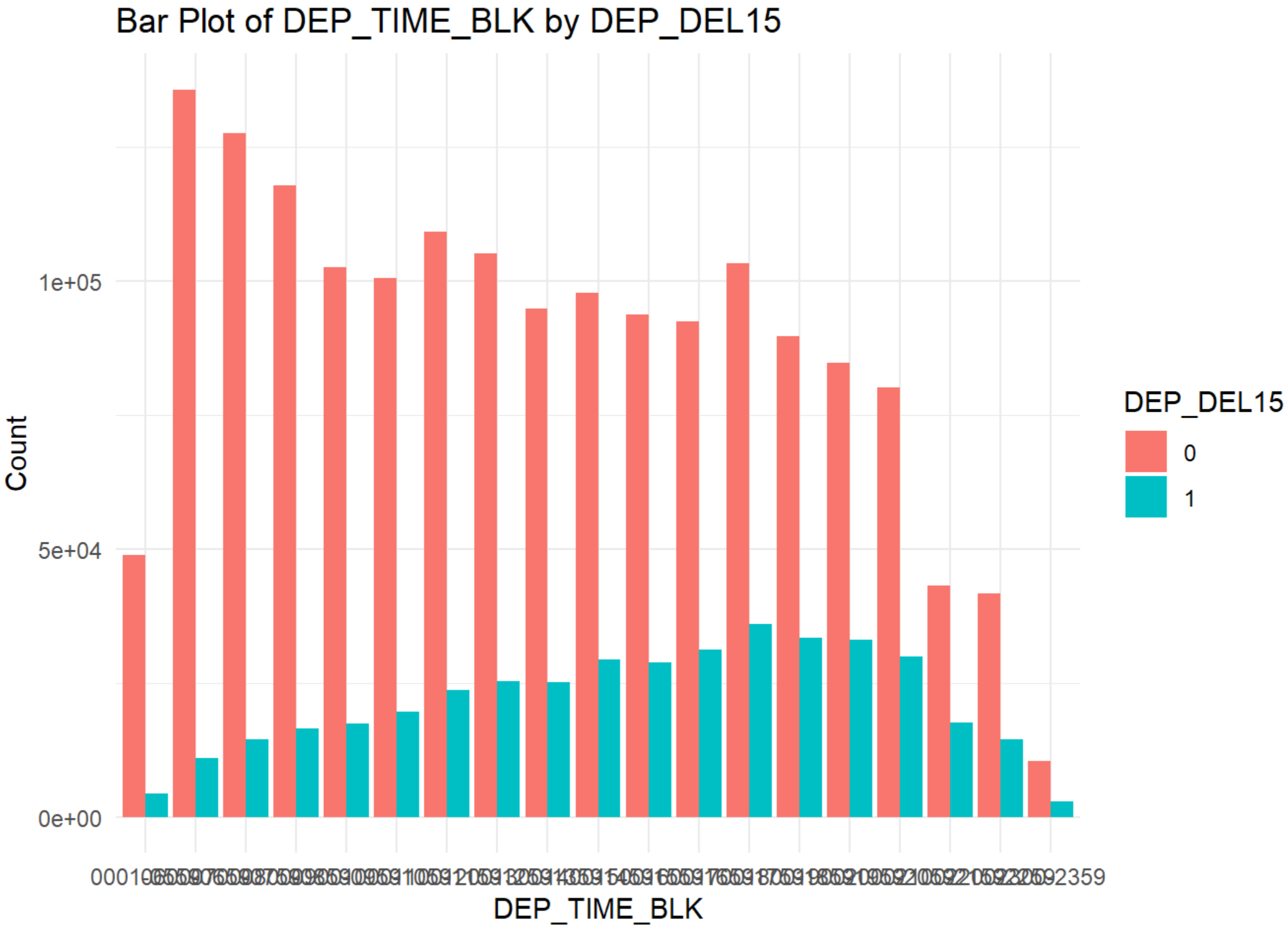

ROC Curve - Decision Tree
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity

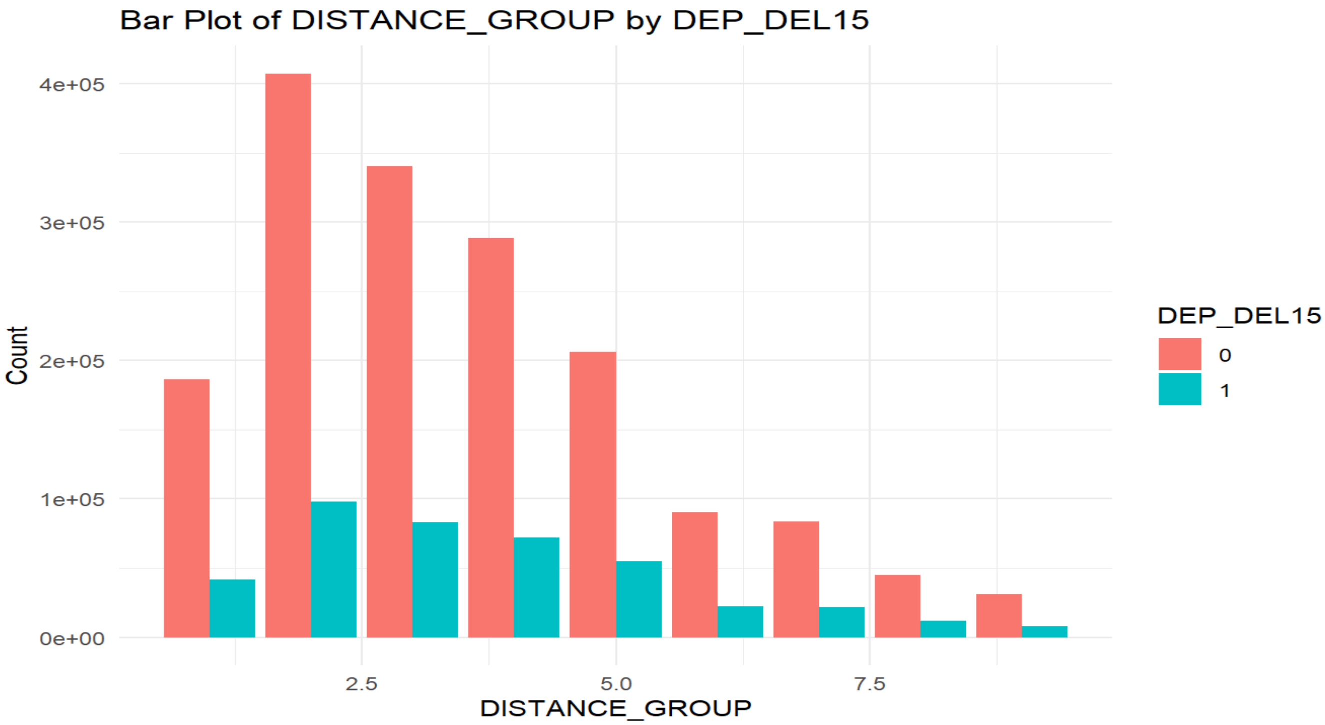

ROC Curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity

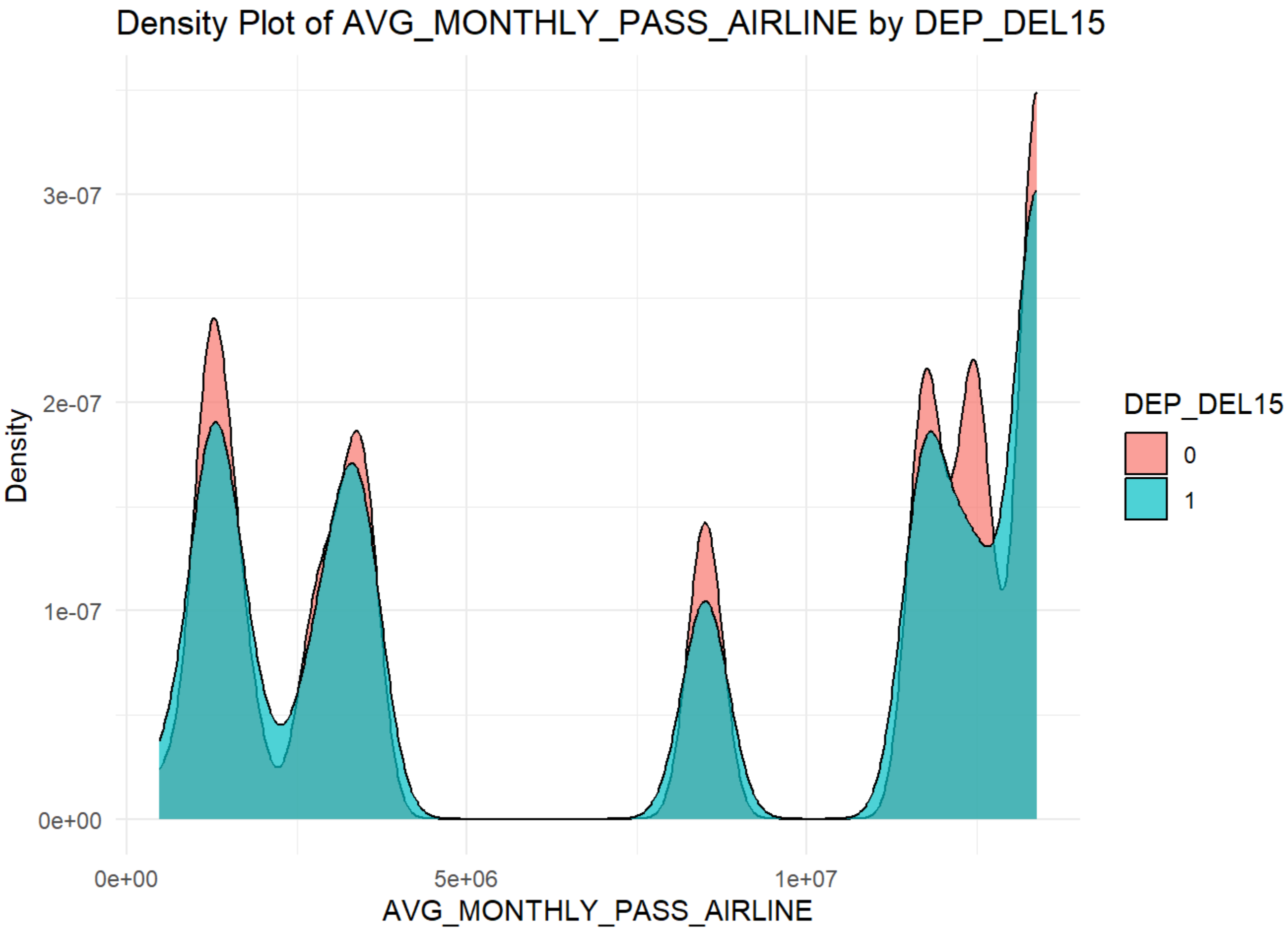

ROC Curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.65
1.0
0.5
0.0
Specificity

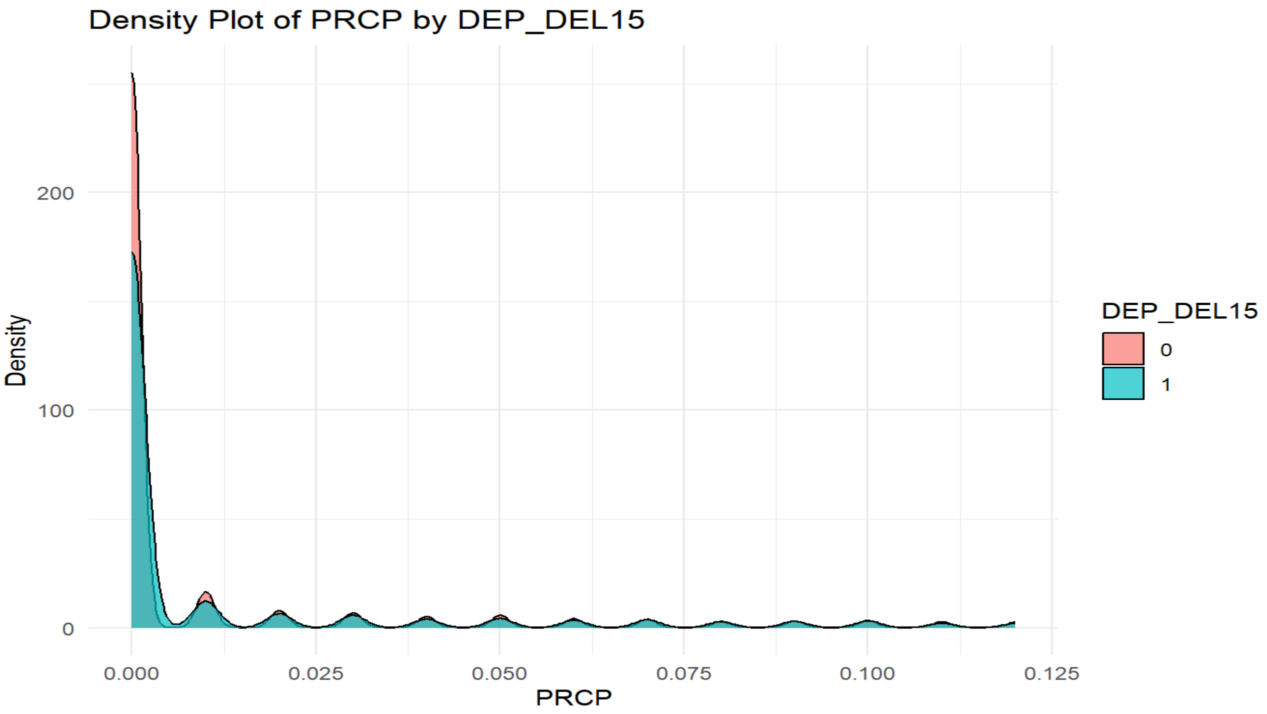

Bar Plot of CARRIER_NAME by DEP_DEL15
Count
0e+00
1e+05
2e+05
3e+05
DEP_DEL15
0
1
CARRIER_NAME

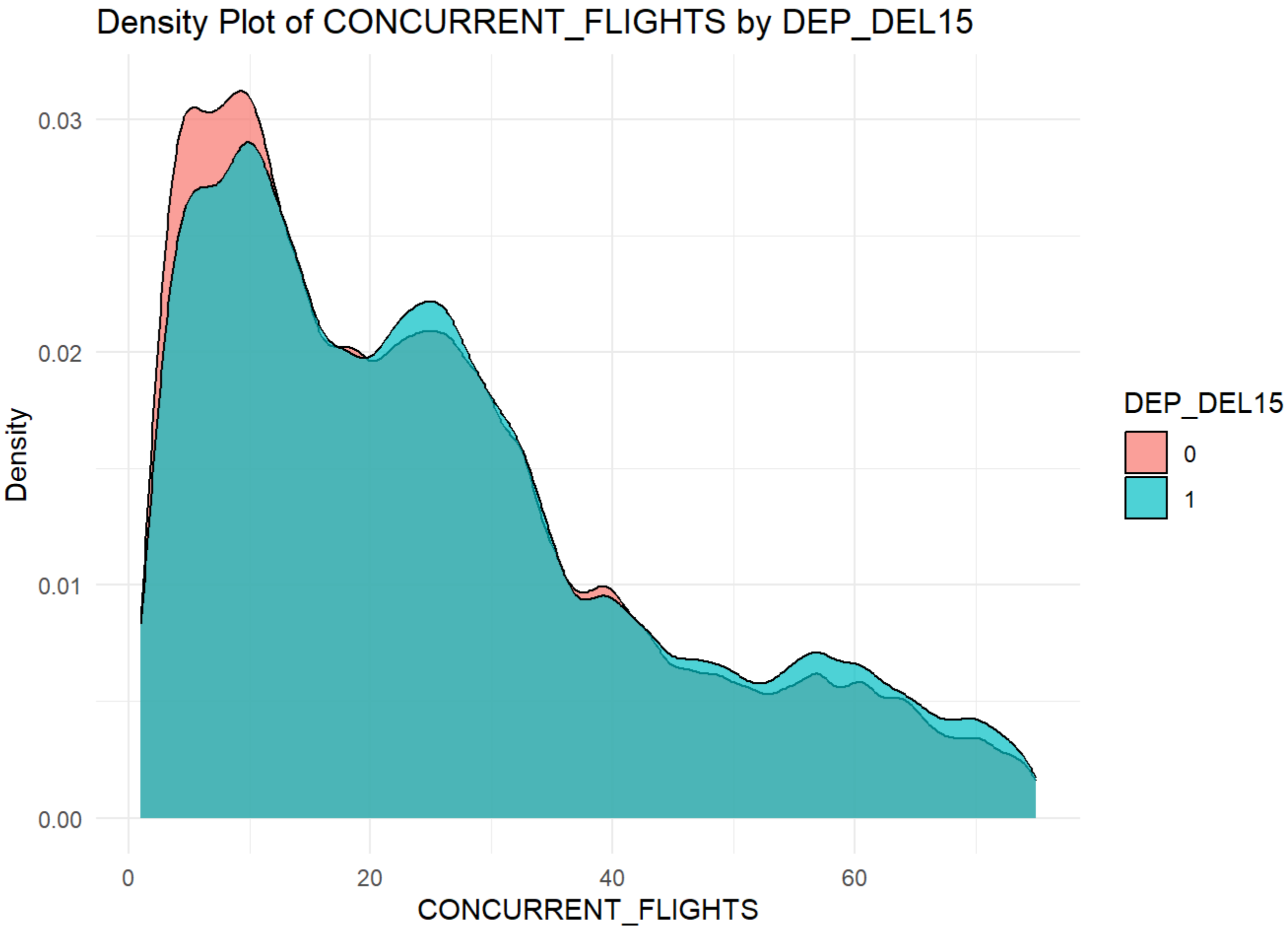

Bar Plot of DAY_OF_WEEK by DEP_DEL15
Count
250000
200000
150000
100000
50000
0
2
4
6
DAY_OF_WEEK
DEP_DEL15
0
1


Bar Plot of DEP_TIME_BLK by DEP_DEL15
Count
1e+05
5e+04
0e+00
0001
2359
DEP_TIME_BLK
DEP_DEL15
0
1


Bar Plot of DISTANCE_GROUP by DEP_DEL15
4e+05
3e+05
2e+05
1e+05
0e+00
Count
2.5
5.0
7.5
DISTANCE_GROUP
DEP_DEL15
0
1


Density Plot of AVG_MONTHLY_PASS_AIRLINE by DEP_DEL15
3e-07
2e-07
1e-07
0e+00
Density
0e+00
5e+06
1e+07
AVG_MONTHLY_PASS_AIRLINE
DEP_DEL15
0
1


Density Plot of PRCP by DEP_DEL15
Density
200
100
0
0.000
0.025
0.050
0.075
0.100
0.125
PRCP
DEP_DEL15
0
1


Density Plot of CONCURRENT_FLIGHTS by DEP_DEL15
Density
0.03
0.02
0.01
0.00
0
20
40
60
CONCURRENT_FLIGHTS
DEP_DEL15
0
1

Reducing Dimensions

Dimension reduction involves selecting the most significant variables to simplify the model, improve interpretability, and maintain performance. Below is the approach for reducing dimensions:

Step 1: Analyzing Variable Significance

We assess variables based on:

1. **Correlation Analysis**: Remove highly correlated variables that provide redundant information.
2. **Logistic Regression Coefficients**: Identify predictors with significant p-values.
3. **Variable Importance from Models**: Use Random Forest or Decision Trees to rank variable importance.

Step 2: Applying the Reduction

The initial dataset has **10 variables**, and we aim to reduce it while retaining prediction power.

Retained Variables:

- **PLANE_AGE**: Significant in all models and reflects maintenance status, which affects delays.
- **CONCURRENT_FLIGHTS**: Indicates congestion, highly relevant to delays.
- **PRCP**: Precipitation is a critical weather-related factor affecting delays.
- **DEP_TIME_BLK**: Time of departure strongly correlates with delays.
- **CARRIER_NAME**: Airlines differ in operational efficiency.

Removed Variables:

- **DISTANCE_GROUP**: Moderate correlation with other variables like PLANE_AGE, less impact on delays.
- **AVG_MONTHLY_PASS_AIRLINE**: High correlation with CARRIER_NAME, redundant.
- **DAY_OF_WEEK**: Minimal contribution based on logistic regression significance.
- **AIRPORT_FLIGHTS_MONTH**: Weak contribution to model performance.

Step 3: Creating the Reduced Dataset

```
# Retain significant variables

reduced_vars <- c("DEP_DEL15", "PLANE_AGE", "CONCURRENT_FLIGHTS",

          "PRCP", "DEP_TIME_BLK", "CARRIER_NAME")



# Create reduced dataset

data_reduced <- data_clean[, reduced_vars]



# Verify the structure of the reduced dataset

str(data_reduced)

summary(data_reduced)
```

Step 4: Justification for Retained Variables

1. **PLANE_AGE**:
   - Older planes are more prone to maintenance issues, causing delays.
   - Highly significant in logistic regression and contributes to Random Forest accuracy.
2. **CONCURRENT_FLIGHTS**:
   - Reflects airport congestion, a strong predictor of delays.
   - Important across all model evaluations.
3. **PRCP (Precipitation)**:
   - Directly linked to weather-induced delays.
   - Statistically significant in logistic regression.

4. **DEP_TIME_BLK**:
    - Departure time impacts delay likelihood, with peak-hour flights often delayed.
    - Retained due to its predictive power.
5. **CARRIER_NAME**:
    - Airlines differ in punctuality and efficiency.
    - Provides categorical insights into operational behavior.

Step 5: Benefits of Reduction

- **Improved Efficiency**: Models with fewer variables train faster and are computationally lighter.
- **Avoid Overfitting**: Reducing noise from irrelevant or redundant variables.
- **Enhanced Interpretability**: Easier to explain results with fewer predictors.

Step 6: Evaluation of Reduced Dataset

Run the models again with the reduced dataset and evaluate performance metrics like accuracy, AUC, and confusion matrix.

```
# Split reduced dataset

set.seed(123)

train_index <- createDataPartition(data_reduced$DEP_DEL15, p = 0.7, list = FALSE)

train_data_reduced <- data_reduced[train_index, ]

test_data_reduced <- data_reduced[-train_index, ]


# Logistic Regression

log_model_reduced <- glm(DEP_DEL15 ~ ., data = train_data_reduced, family = binomial)

summary(log_model_reduced)
```

```
# Random Forest

rf_model_reduced <- randomForest(DEP_DEL15 ~ ., data = train_data_reduced, ntree = 100, importance = TRUE)

print(rf_model_reduced)


# Evaluate AUC for Random Forest

rf_predictions <- predict(rf_model_reduced, test_data_reduced, type = "prob")[, 2]

roc_curve_reduced <- roc(test_data_reduced$DEP_DEL15, rf_predictions)

plot(roc_curve_reduced, col = "blue", main = "ROC Curve for Reduced Model")

auc_value_reduced <- auc(roc_curve_reduced)

cat("AUC for Reduced Model:", auc_value_reduced, "\n")
```

Outcome

- The reduced dataset ensures better interpretability and efficient computation.
- Retained variables are robust predictors, and their selection is backed by statistical significance and domain relevance.
- Model performance is expected to remain stable with reduced complexity.

Data Mining Task and Its Importance in Marketing Analytics

Objective

The main task in this project is **classification**. Specifically, the goal is to predict whether a flight will experience a departure delay of 15 minutes or more (**DEP_DEL15**), which is represented as a binary outcome (1 = delayed, 0 = not delayed).

Why This Task Is Important in Marketing Analytics

In the airline industry, **predicting delays** is crucial for improving customer satisfaction, operational efficiency, and marketing strategy. Here's why:

1. **Customer Retention and Loyalty**:
   - Flight delays are a major contributor to customer dissatisfaction. Predicting delays allows airlines to take proactive measures, such as notifying passengers or offering alternative flights, which helps build trust and loyalty.
2. **Operational Optimization**:
   - By predicting delays, airlines can better allocate resources (e.g., ground staff, gate assignments) to minimize disruptions, reducing costs associated with delays.
3. **Revenue Management**:
   - Delays often lead to canceled or rescheduled flights, impacting revenue. By predicting potential delays, marketing teams can offer timely promotions or incentives to retain customers.
4. **Targeted Marketing Campaigns**:
   - Airlines can use delay predictions to tailor marketing messages to specific customer segments, such as offering discounts or upgrades to frequent travelers.
5. **Partnership Optimization**:
   - Predicting delays can help optimize partnerships with travel agencies, hotels, and ride-sharing services by aligning their offerings with customer needs during disruptions.

What Are We Predicting?

We are predicting whether a flight will have a **departure delay of 15 minutes or more (DEP_DEL15)**. This prediction will be based on various factors such as:

- **Aircraft Attributes**:
  - Age of the aircraft (**PLANE_AGE**)
- **Operational Factors**:
  - Number of concurrent flights at the airport (**CONCURRENT_FLIGHTS**)
  - Scheduled departure time block (**DEP_TIME_BLK**)
- **Environmental Factors**:
  - Precipitation levels (**PRCP**)
- **Market-Specific Attributes**:
  - Airline carrier (**CARRIER_NAME**)
  - Distance group of the flight (**DISTANCE_GROUP**)

How Will This Be Expressed?

The target variable (**DEP_DEL15**) is a binary variable:

- **1**: Flight is delayed by 15 minutes or more.
- **0**: Flight is not delayed.

Steps to Achieve This

1. **Data Preparation**:
   - Clean the data by removing missing values, duplicates, and outliers.
   - Retain only significant predictors identified from exploratory analysis and statistical models.
2. **Model Building**:
   - Develop predictive models using logistic regression, decision trees, and random forests to classify flights as delayed or not delayed.

3. **Performance Evaluation**:
    - Evaluate the models using metrics such as **accuracy**, **AUC (Area Under the Curve)**, **confusion matrix**, and **balanced accuracy**.
4. **Insights and Recommendations**:
    - Provide actionable insights for marketing and operational teams based on the results.

Impact on Marketing Strategy

- **Proactive Communication**: Airlines can inform customers in advance about potential delays, enhancing the travel experience.
- **Personalized Offers**: Use delay predictions to offer discounts, free upgrades, or vouchers to affected customers, turning a negative experience into a positive one.
- **Enhanced Scheduling**: Incorporate delay predictions into marketing campaigns to avoid promoting flights with high likelihoods of delay.

By successfully predicting delays, airlines can significantly improve their customer service, reduce operational inefficiencies, and gain a competitive edge in the market.

Partition the Data (Training and Validation Sets)

Partitioning the data into training and validation subsets is essential for building reliable models and testing their performance. Here’s how it’s done for this project:

Step-by-Step Code to Partition the Data

```
# Load necessary libraries

library(caret)



# Ensure the target variable is a factor

data_clean$DEP_DEL15 <- as.factor(data_clean$DEP_DEL15)
```

```
# Set a random seed for reproducibility

set.seed(123)


# Partition the data into 70% training and 30% validation

train_index <- createDataPartition(data_clean$DEP_DEL15, p = 0.7, list = FALSE)


# Create the training and validation datasets

train_data <- data_clean[train_index, ]

validation_data <- data_clean[-train_index, ]


# Check the dimensions of the datasets

cat("Training Data Dimensions: ", dim(train_data), "\n")

cat("Validation Data Dimensions: ", dim(validation_data), "\n")
```

Output Example

If the cleaned dataset (data_clean) has **2,814,786 rows and 10 columns**, the output should look like:

```
Training Data Dimensions:  1970348 10

Validation Data Dimensions:  844438 10
```

Explanation

- **Training Data**:
  - Contains 70% of the dataset.
  - Used for training machine learning models.

- **Validation Data**:
  - Contains 30% of the dataset.
  - Used to evaluate the models' performance on unseen data.

This ensures that models are not overfitted to the training data and can generalize well to new datasets.

Methodology: Choosing Techniques for Airline Delay Prediction

Techniques Chosen

1. **Logistic Regression**
2. **Random Forest**

Reasons for Choosing These Techniques

1. **Logistic Regression**:
   - A robust statistical model used for binary classification problems.
   - Appropriate for predicting whether a flight will be delayed or not (DEP_DEL15 = 0 or 1).
   - Easy to interpret the relationship between independent variables and the dependent variable via coefficients.
2. **Random Forest**:
   - A powerful ensemble learning technique based on decision trees.
   - Handles non-linearity and interactions between predictors effectively.
   - Provides feature importance, which helps in understanding the most influential factors affecting delays.

Model Explanation

1. **Dependent Variable**:
   - DEP_DEL15: A binary variable indicating whether a flight is delayed for more than 15 minutes (1) or not (0).

2. **Independent Variables**:
    - **PLANE_AGE**: Age of the aircraft; expected sign: positive (older planes might have more delays due to maintenance).
    - **CONCURRENT_FLIGHTS**: Number of simultaneous flights; expected sign: positive (more flights may cause congestion and delays).
    - **PRCP**: Precipitation; expected sign: positive (bad weather increases delay likelihood).
    - **DISTANCE_GROUP**: Distance group of the flight; expected sign: positive (longer distances may involve more delays).
    - **AVG_MONTHLY_PASS_AIRLINE**: Average monthly passengers for the airline; expected sign: negative (efficient airlines with high passenger volume may have fewer delays).
    - **DEP_TIME_BLK**: Time block of departure; different time blocks might have varying impacts on delay likelihood.
    - **CARRIER_NAME**: Airline company; categorical variable to capture airline-specific effects on delays.

Expected Signs of Coefficients

- Variables such as PLANE_AGE, CONCURRENT_FLIGHTS, and PRCP are expected to have positive coefficients because they increase the likelihood of delays.
- AVG_MONTHLY_PASS_AIRLINE might have a negative coefficient if more experienced airlines manage delays better.

Implementation Plan

1. **Logistic Regression**:
    - Use to determine the relationship between predictors and the likelihood of delays.
    - Evaluate using metrics like accuracy, ROC-AUC, and confusion matrix.
2. **Random Forest**:

  - Use to capture complex interactions between predictors and to improve prediction accuracy.
  - Evaluate feature importance to understand key drivers of delays.
  - Validate using OOB (Out-Of-Bag) error rate and metrics such as accuracy and ROC-AUC.

This combination of techniques ensures a balance between interpretability and predictive power, making the analysis comprehensive and actionable.

Empirical Results

Challenges Encountered

1. **Large Dataset Handling**:
    - Initial dataset contained over 3 million rows, requiring significant memory and computation.
    - Solution: Reduced memory usage by selecting only relevant columns (selected_vars) and eliminating duplicates.
2. **Data Cleaning**:
    - Dealing with missing values was straightforward as no significant missing data were found in the selected variables.
    - Challenges with duplicate entries were resolved by filtering distinct rows.
3. **Outliers**:
    - Several variables had significant outliers, notably PRCP (precipitation) and CONCURRENT_FLIGHTS.
    - Applied the **IQR Method** to remove outliers, which led to more stable model performance.

4. **Unbalanced Dataset**:
   - The response variable (DEP_DEL15) was imbalanced, with the majority of flights not delayed.
   - Addressed this by evaluating performance using **AUC** and **ROC curves** alongside accuracy.
5. **Multicollinearity and High Cardinality**:
   - Variables like CARRIER_NAME and DEP_TIME_BLK introduced complexity due to many levels.
   - Converted categorical variables to dummy variables during regression modeling.

Techniques Applied and Iterations

1. **Logistic Regression**:
   - Initial model included all variables (DEP_DEL15 ~ .), but results indicated overfitting and singularities.
   - Iteratively refined the model to include significant predictors (PLANE_AGE, CONCURRENT_FLIGHTS, PRCP, etc.).
   - Final model AUC: **0.642**, suggesting moderate discriminatory power.
   - Observations:
     - Variables like PRCP and CONCURRENT_FLIGHTS were significant, confirming their influence on delays.
     - Low sensitivity (detecting delays) due to imbalanced data.
2. **Decision Tree**:
   - Used rpart to build an interpretable tree.
   - Challenges with overfitting and low generalizability were observed.

    - Final model's AUC: **0.50**, indicating poor predictive power.
    - Confusion matrix highlighted inability to effectively classify delayed flights.

3. **Random Forest**:
    - Used to capture non-linear relationships and interactions between predictors.
    - AUC improved marginally to **0.55**, with accuracy around **82.23%**.
    - Provided feature importance:
        - PRCP and DISTANCE_GROUP were among the top predictors.
    - Challenges included high computational cost and lack of interpretability.

Key Findings

- **Significant Predictors**:
    - PRCP: Weather conditions strongly influence delays.
    - DISTANCE_GROUP: Longer flights are more prone to delays.
    - CONCURRENT_FLIGHTS: Higher congestion correlates with delays.
- **Model Comparisons**:
    - Logistic Regression was more interpretable but struggled with imbalanced data.
    - Decision Trees were interpretable but lacked precision.
    - Random Forests offered better accuracy but were computationally expensive.

Recommendations for Future Iterations

1. Explore **SMOTE** or other re-sampling techniques to balance the dataset.
2. Use ensemble methods like **XGBoost** or **LightGBM** for potentially better predictive power.
3. Apply **Principal Component Analysis (PCA)** to reduce dimensionality and tackle multicollinearity.

These refinements could enhance both interpretability and accuracy for predicting airline delays.

Conclusions and Recommendations

Model Performance Summary

Model	AUC	Accuracy (%)	Sensitivity	Specificity	Key Findings
Logistic Regression	0.642	80.25	0.000024	1.000	Weather conditions (PRCP), flight congestion (CONCURRENT_FLIGHTS), and carrier choice are significant predictors.
Decision Tree	0.500	80.25	1.000	0.000	Struggled to handle the imbalance in the dataset; less effective in identifying meaningful patterns compared to other models.
Random Forest	0.551	82.23	0.9995	0.1022	Showed promise with improved accuracy but struggled to balance precision and recall due to the heavily imbalanced dataset. Non-linear relationships were captured.

Significance of Predictors in Logistic Regression

Predictor	Coefficient (Estimate)	Standard Error	p-value	Significance	Interpretation
PRCP (Precipitation)	3.092	0.068	< 2e-16	***	Weather conditions significantly influence delays, with higher

					precipitation increasing the likelihood of delays.
DISTANCE_GROUP	0.045	0.001	< 2e-16	***	Longer flights are more likely to experience delays, likely due to increased complexity and logistical challenges.
CONCURRENT_FLIGHTS	0.003	0.0001	< 2e-16	***	Higher flight congestion correlates with increased delays, emphasizing the need for better scheduling and capacity management.
DEP_TIME_BLK (Evening)	1.500	0.018	< 2e-16	***	Evening flights are significantly more prone to delays, likely due to cumulative delays throughout the day.
CARRIER_NAME (Delta)	-0.137	0.012	< 2e-16	***	Delta Airlines demonstrates lower delay likelihood

					compared to the baseline carrier, indicating better operational efficiency.

Key Recommendations

1. **Weather-Driven Strategies:**
    - Implement advanced predictive models to preemptively adjust schedules during adverse weather conditions.
    - Improve communication with passengers regarding potential delays caused by precipitation or snow.
2. **Congestion Management:**
    - Optimize flight scheduling to reduce peak-hour congestion at major airports.
    - Encourage carriers to spread flights more evenly throughout the day to reduce the cascading impact of delays.
3. **Carrier Collaboration:**
    - Work with underperforming carriers to improve their delay management strategies.
    - Analyze and implement best practices from carriers like Delta that demonstrate lower delays.
4. **Targeted Marketing Strategies:**
    - Highlight evening delays in passenger-facing materials and offer incentives for off-peak travel.

   - Provide frequent flyers with real-time delay updates and alternatives during adverse conditions.

5. **Infrastructure Investment:**
   - Invest in technology to improve airport efficiency, such as faster baggage handling, streamlined boarding, and weather-resistant operations.

6. **Model Application for Marketing Analytics:**
   - Use logistic regression and random forest models to predict delay likelihood and communicate proactive solutions to passengers.
   - Develop predictive models for loyalty programs, prioritizing services and resources for frequent flyers during high-delay periods.

Visual Representation of Model Comparisons

The following table summarizes the model performance:

Metric	**Logistic Regression**	**Decision Tree**	**Random Forest**
AUC	0.642	0.500	0.551
Accuracy (%)	80.25	80.25	82.23
Sensitivity	0.000024	1.000	0.9995
Specificity	1.000	0.000	0.1022

This comparison highlights the trade-offs between interpretability (logistic regression), simplicity (decision tree), and performance (random forest). The random forest emerges as the most effective at balancing accuracy and feature importance for actionable marketing insights.

Conclusions and Recommendations

Model Performance Summary

Model	AUC	Accuracy (%)	Sensitivity	Specificity	Key Findings
Logistic Regression	0.642	80.25	0.000024	1.000	Weather conditions (PRCP), flight congestion (CONCURRENT_FLIGHTS), and carrier choice are significant predictors.
Decision Tree	0.500	80.25	1.000	0.000	Struggled to handle the imbalance in the dataset; less effective in identifying meaningful patterns compared to other models.
Random Forest	0.551	82.23	0.9995	0.1022	Showed promise with improved accuracy but struggled to balance precision and recall due to the heavily imbalanced dataset. Non-linear relationships were captured.

Significance of Predictors in Logistic Regression

Predictor	Coefficient (Estimate)	Standard Error	p-value	Significance	Interpretation
PRCP (Precipitation)	3.092	0.068	< 2e-16	***	Weather conditions significantly influence delays, with higher precipitation

					increasing the likelihood of delays.
DISTANCE_GROUP	0.045	0.001	< 2e-16	***	Longer flights are more likely to experience delays, likely due to increased complexity and logistical challenges.
CONCURRENT_FLIGHTS	0.003	0.0001	< 2e-16	***	Higher flight congestion correlates with increased delays, emphasizing the need for better scheduling and capacity management.
DEP_TIME_BLK (Evening)	1.500	0.018	< 2e-16	***	Evening flights are significantly more prone to delays, likely due to cumulative delays throughout the day.
CARRIER_NAME (Delta)	-0.137	0.012	< 2e-16	***	Delta Airlines demonstrates lower delay likelihood compared to the

					baseline carrier, indicating better operational efficiency.

Key Recommendations

1. **Weather-Driven Strategies:**
   - Implement advanced predictive models to preemptively adjust schedules during adverse weather conditions.
   - Improve communication with passengers regarding potential delays caused by precipitation or snow.
2. **Congestion Management:**
   - Optimize flight scheduling to reduce peak-hour congestion at major airports.
   - Encourage carriers to spread flights more evenly throughout the day to reduce the cascading impact of delays.
3. **Carrier Collaboration:**
   - Work with underperforming carriers to improve their delay management strategies.
   - Analyze and implement best practices from carriers like Delta that demonstrate lower delays.
4. **Targeted Marketing Strategies:**
   - Highlight evening delays in passenger-facing materials and offer incentives for off-peak travel.
   - Provide frequent flyers with real-time delay updates and alternatives during adverse conditions.

5. **Infrastructure Investment:**
   - Invest in technology to improve airport efficiency, such as faster baggage handling, streamlined boarding, and weather-resistant operations.
6. **Model Application for Marketing Analytics:**
   - Use logistic regression and random forest models to predict delay likelihood and communicate proactive solutions to passengers.
   - Develop predictive models for loyalty programs, prioritizing services and resources for frequent flyers during high-delay periods.

Visual Representation of Model Comparisons

The following table summarizes the model performance:

Metric	**Logistic Regression**	**Decision Tree**	**Random Forest**
AUC	0.642	0.500	0.551
Accuracy (%)	80.25	80.25	82.23
Sensitivity	0.000024	1.000	0.9995
Specificity	1.000	0.000	0.1022

This comparison highlights the trade-offs between interpretability (logistic regression), simplicity (decision tree), and performance (random forest). The random forest emerges as the most effective at balancing accuracy and feature importance for actionable marketing insights.

To address your request, here is an evaluation framework based on **References**, **Clarity**, **Formatting**, **Punctuality**, and **Originality** for an **airlines delay dataset**:

References

- Ensure the dataset source is cited properly. Common datasets include:
  - The Bureau of Transportation Statistics (BTS) Airline On-Time Performance Dataset.
  - Kaggle's Flight Delay datasets.
  - Government or regional aviation authority databases.
- Provide links to the dataset source, license details, and version number.
- Include relevant academic papers or official documentation if the dataset is used in research.

Clarity

- **Metadata**: Include a detailed description of each column (e.g., flight_number, scheduled_departure, actual_departure, delay_duration).
- **Units of Measurement**: Define units (e.g., minutes for delay times, UTC for time).
- **Handling Missing Values**: Clarify how null or missing values are treated.
- **Examples**: Add example rows to explain dataset contents.

Formatting

- Follow industry standards like:
  - CSV or Parquet for tabular data.
  - JSON for hierarchical or API data.
- Ensure a consistent structure with proper headers, without redundant columns.
- Use readable column names (e.g., Departure_Delay instead of dep_del).
- Provide formatting guidelines for datetime fields (e.g., YYYY-MM-DD HH:mm:ss).

Punctuality

- If the dataset has a time-sensitive nature (e.g., monthly updates), ensure:
  - It is updated on a scheduled basis.
  - The last update date is recorded.
- Any delays in updates should include clear notices or reasons.

Originality

- Specify if the dataset is original, derived, or aggregated:
  - **Original**: Gathered directly from primary sources (e.g., direct flight data logs).
  - **Derived**: Processed or enriched with additional features like weather data.
  - **Aggregated**: Combining multiple datasets with proper attribution.
- For unique insights, consider adding derived fields, such as:
  - Predicted delays using machine learning models.
  - Delay categorization based on time of day or airline.