A logo of a globe with yellow rings around it

AI-generated content may be incorrect.

**GROUP ASSIGNMENT**

**TECHNOLOGY PARK MALAYSIA**

**CT127-3-2-PFDA**

**PROGRAMMING FOR DATA ANALYSIS**

**HAND OUT DATE: 4 AUGUST 2025**

**HAND IN DATE: 26 SEPTEMBER 2025**

**WEIGHTAGE: 50%**

**INSTRUCTIONS TO CANDIDATES:**

1. **Students are advised to underpin their answers with the use of references (cited using the American Psychological Association (APA) Referencing).**
2. **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.**
3. **Cases of plagiarism will be penalized.**
4. **Submit the assignment to APU Learning Management System.**
5. **You must obtain 0% overall to pass this module.**

**Group No: 28**

**Member 1: CHAN MIN HUEY TP083261**

**Member 2: GONG YEE CHENG TP081910**

**Member 3: KANG HONG QIAN TP081205**

**Member 4: YAP LI SHAN TP080968**

Table of Contents

[1.0 Introduction 5](#_Toc211025227)

[1.1 Data Description 6](#_Toc211025228)

[1.2 Hypothesis and Objectives 7](#_Toc211025229)

[2.0 Data Preparation 8](#_Toc211025230)

[2.1 Data Import 8](#_Toc211025231)

[2.2 Data Cleaning 10](#_Toc211025232)

[2.2.1 Convert hhmm variables to date-time variable 10](#_Toc211025233)

[2.2.2 Handle abnormal flights (overnight / same time) 12](#_Toc211025234)

[2.2.3 Remove unneeded variables 14](#_Toc211025235)

[2.2.4 Missing Value analysis 15](#_Toc211025236)

[2.2.5 Cancelled flights logic 16](#_Toc211025237)

[2.2.6 Check for duplicate records 19](#_Toc211025238)

[2.2.7 Check for Outlier records 20](#_Toc211025239)

[2.2.8 Identify abnormal airport codes 21](#_Toc211025240)

[2.2.9 Factorize airports and group rare ones 23](#_Toc211025241)

[2.3 Data Verification 25](#_Toc211025242)

[3.0 Analysis 29](#_Toc211025243)

[3.1 Objective 1: To investigate the relationship between departure delay and arrival delay, including the impact of additional flight factors (e.g., distance, air time and flight date), and explore their overall correlation and trend. – Chan Min Huey (TP083261) 29](#_Toc211025244)

[3.1.1 Analysis 1-1: To analyse the distribution of departure delay and arrival delay in order to understand their basic characteristics. 29](#_Toc211025245)

[3.1.2 Analysis 1-2: To examine the relationship between departure delay and arrival delay. 36](#_Toc211025246)

[3.1.3 Analysis 1-3: To evaluate whether schedule depature with flight factors like flight distance, and air time significantly predict arrival delay. 40](#_Toc211025247)

[3.1.4 Analysis 1-4: To provide recommendations on managing different departure delay ranges and evaluate whether weekday/weekend patterns influence arrival delays. 44](#_Toc211025248)

[3.1.5 Extra Features 48](#_Toc211025249)

[3.1.6 Conclusion 48](#_Toc211025250)

[3.1.7 Recommendation 48](#_Toc211025251)

[3.2 Objective 2: To evaluate the performance of different airlines in managing flight delays – Kang Hong Qian (TP081205) 49](#_Toc211025252)

[3.2.1 Analysis 2-1: Flight Delay Distribution by Airline 49](#_Toc211025253)

[3.2.2 Analysis 2-2: Impact of Departure Time and Airline on Delay 53](#_Toc211025254)

[3.2.3 Analysis 2-3: Assessing Chaining Delay Risk Using Logistic Regression 58](#_Toc211025255)

[3.2.4 Analysis 2-4: Identifying Monthly and Seasonal Delay Patterns 61](#_Toc211025256)

[3.2.5 Extra Features 63](#_Toc211025257)

[3.2.6 Conclusion 63](#_Toc211025258)

[3.2.7 Recommendation 64](#_Toc211025259)

[3.3 Objective 3: To analyse how various factors , including time of day, day of week, flight distance, and departure delay, influence flight arrival delays, and to identify the relationships that contribute to overall delay performance. – Gong Yee Cheng (TP081910) 65](#_Toc211025260)

[3.3.1 Analysis 3-1: Assessing Time-of-Day and Flight Distance Interaction Effects on Flight Arrival Delays using Two-Way ANOVA 65](#_Toc211025261)

[3.3.2 Analysis 3-2: Exploring Weekly Delay Trends through Descriptive Statistical Profiling 69](#_Toc211025262)

[3.3.3 Analysis 3-3: Examining the Relationship between Flight Distance and Arrival Delay using Spearman Rank Correlation 71](#_Toc211025263)

[3.3.4 Analysis 3-4: Evaluating the Predictive Influence of Departure Delay on Arrival Delay through Linear Regression Modelling 73](#_Toc211025264)

[3.3.5 Extra Features 75](#_Toc211025265)

[3.3.6 Conclusion 75](#_Toc211025266)

[3.3.7 Recommendation: 75](#_Toc211025267)

[3.4 Objective 4: To analyse the propagation mechanism of flight delays, with a focus on the relationship between Late Aircraft Delay and Arrival Delay and find out the key drivers that amplify delay propagation as well as potential intervention points for mitigation. – Yap Li Shan (TP080968) 76](#_Toc211025268)

[3.4.1 Analysis 4-1: Identify the primary attributable factors to flight delays and analysing the localized delay risk profile across the Top 15 busiest origin airports. 76](#_Toc211025269)

[3.4.2 Analysis 4-2: To diagnose how ground operational efficiency (measured by TAXI\_OUT) moderates the relationship between LATE\_AIRCRAFT\_DELAY and ARRIVAL\_DELAY. 82](#_Toc211025270)

[3.4.3 Analysis 4-3: To develop a pre-departure predictive model by focused on pre-flight observable factors (ORIGIN\_AIRPORT, AIRLINE, MONTH) to identify inherent flight delay risks before take-off for preventive intervention. 88](#_Toc211025271)

[3.4.4 Analysis 4-4: Quantify each airport's average total contribution in minutes across the two core risk factors and identify the top 15 origin airports with the highest minute. 93](#_Toc211025272)

[3.4.5 Extra Features 95](#_Toc211025273)

[3.4.6 Conclusion 95](#_Toc211025274)

[3.4.7 Recommendation 95](#_Toc211025275)

[4.0 Conclusion 97](#_Toc211025276)

[4.1 Overall Discussion 97](#_Toc211025277)

[4.2 Recommendation 97](#_Toc211025278)

[4.3 Limitation and Future Direction 98](#_Toc211025279)

[5.0 References 99](#_Toc211025280)

[6.0 Workload Matrix 100](#_Toc211025281)

# 1.0 Introduction

Flights delay is a significant challenge to the airline industry as the delays will directly impacts passenger satisfaction and airtime reliability. In this project, “Flight Delay Prediction and Airline Reliability Assessment”, it aims to analyse flight operation data using R programming to identify the key factors that causes delays. The dataset “flights\_100k.csv” include flight schedules, operational details and delay causes type. By evaluating airline reliability and flight delay patterns, the analysis technique like data exploration, manipulation and visualisation will provide a valuable insights and recommendations to enhance airline operational efficiency and performance.

## 1.1 Data Description

The dataset “flights\_100k.csv” provides comprehensive information about flights operation and delays, including the variable such as date and time (year, month, day), airline, flight number, tail number, origin airports and destination airports. Each flights record is recorded with flight details like scheduled departure/arrival and actual departure/arrival time, with others delay reasons categorized by weather, security, airline-related factors and more. Other than that, the dataset also includes flights duration fields like taxi-out time and air time to allow more detailed analysis. Two additional files “iata\_airline\_codes” (IATA airline codes) and “iata\_airport\_codes” (IATA airport codes) are provided to identify the airline and airport names. Before analysis, data cleaning should be done like handling missing values and removing duplicates is required to ensure data accuracy and consistency.

## 1.2 Hypothesis and Objectives

**Hypothesis**

Departure delay is the primary factor causing arrival delay, and airlines differ systematically in their efficiency at managing departure delays and controlling the propagation of subsequent delays, which directly impacts their overall reliability.

**Objectives**

1. To investigate the relationship between departure delay and arrival delay, including the impact of additional flight factors (e.g., distance, air time and flight date), and explore their overall correlation and trend.

2.To evaluate the performance of different airlines in managing flight delays.

3. To analyse how various factors, including time of day, day of week, flight distance, and departure delay, influence flight arrival delays, and to identify the relationships that contribute to overall delay performance.

4. To analyse the propagation mechanism of flight delays, with a focus on the relationship between Late Aircraft Delay and Arrival Delay, including the impact of additional flights factors (e.g. Airline Delay, Origin Airports, Taxi Out), and find out the key drivers that amplify delay propagation as well as potential intervention points for mitigation.

# 2.0 Data Preparation

## 2.1 Data Import

文本

AI 生成的内容可能不正确。

Figure 2.1.1: Code of read dataset’s csv file

**Output:**

1. Glimpse Flights Dataset

图片包含 日历

AI 生成的内容可能不正确。

Figure 2.1.2: Glimpse of the Raw Flight Dataset

1. Summary Flights Dataset

图形用户界面

AI 生成的内容可能不正确。

Figure 2.1.3: Summary of the Raw Flight Dataset

Read the dataset using the read\_csv() function and inspected with glimpse() and summary() to understand their structure, variable types, and overall data completeness.

## 2.2 Data Cleaning

### 2.2.1 Convert hhmm variables to date-time variable

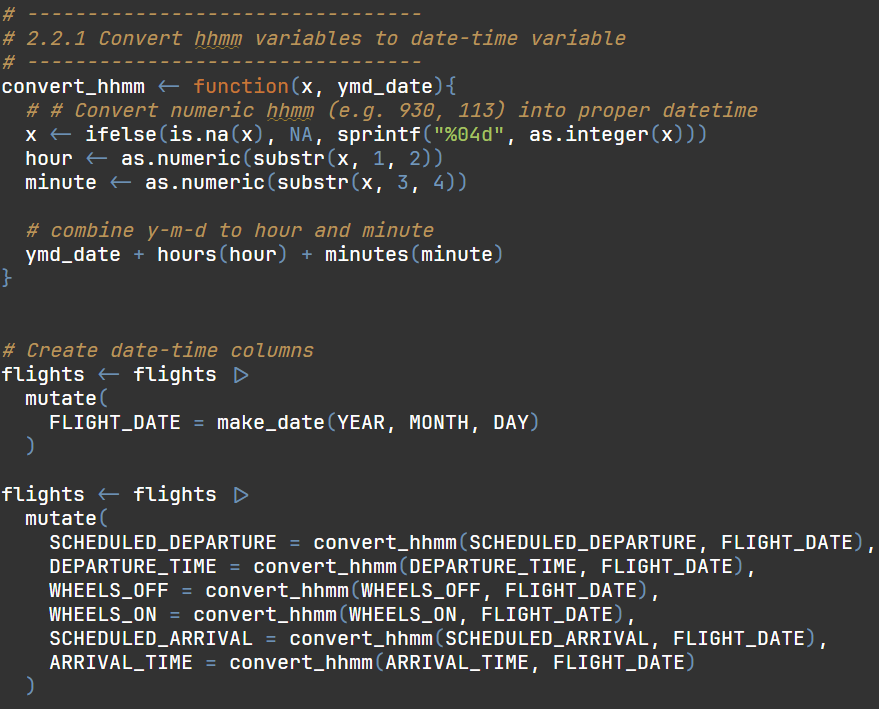


Figure 2.2.1.1**:** Code of Converted Datetime Fields

**Output:**

表格

AI 生成的内容可能不正确。

Figure 2.2.1.2**:** Example of Converted Datetime Fields

Time variables in the dataset were originally stored as four-digit numbers in hhmm format. convert\_hhmm() was defined to convert these into proper datetime objects by combining year, month, and day with hour and minute components. This ensured all time variables were in a consistent and readable format.

### 2.2.2 Handle abnormal flights (overnight / same time)

1. Flights with ARRIVAL < DEPATURE

文本

AI 生成的内容可能不正确。

Figure 2.2.2.1: Code of count Flights with Overnight Flights

**Output:**



Figure 2.2.2.2: Number of Flights with Overnight

Some flights had an ARRIVAL\_TIME earlier than their DEPARTURE\_TIME, indicating overnight flights. These were corrected by adding one day to arrival-related timestamps.

1. Flights with ARRIVAL = DEPARTURE

文本

AI 生成的内容可能不正确。

Figure 2.2.2.3: Code of count Flights with Identical Departure and Arrival Times

**Output:**



Figure 2.2.2.4: Flights with Identical Departure and Arrival Times

Flights where ARRIVAL\_TIME equalled DEPARTURE\_TIME were identified as logically invalid and removed.

1. Fix overnight and not logical flights

电脑屏幕截图

AI 生成的内容可能不正确。

Figure 2.2.2.5: Code of fix Overnight and not Logical Flights

This process ensured temporal consistency and prevented negative or zero travel durations.

### 2.2.3 Remove unneeded variables

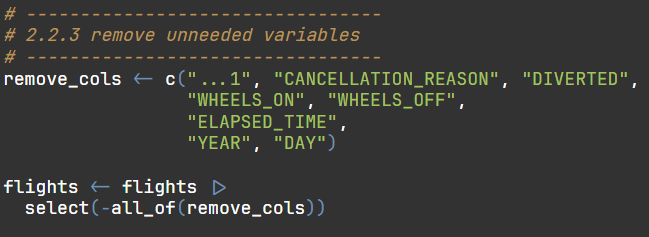


Figure 2.2.3.1: List of Removed Columns from Dataset

Irrelevant columns such as "CANCELLATION\_REASON", "DIVERTED", "WHEELS\_ON", "WHEELS\_OFF", "ELAPSED\_TIME", and time-identifying fields like "YEAR" and "DAY" were removed using select(-all\_of()). This step simplified the dataset, keeping only relevant variables for later analysis.

### 2.2.4 Missing Value analysis

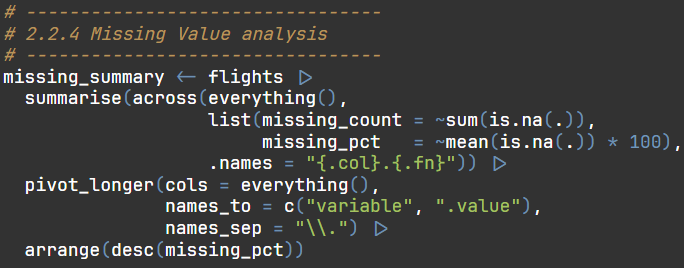


Figure 2.2.4.1: Code of Summary of Missing Values per Variable

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.4.2: Summary of Missing Values per Variable

A missing value summary was generated using summarise(across()) to calculate both missing counts and percentages.

### 2.2.5 Cancelled flights logic

1. Replace NA in delay columns with 0 and check illogical flights

文本

AI 生成的内容可能不正确。

Figure 2.2.5.1: Code of Replace NA in delay columns with 0 and check illogical Flights

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.5.2: Summary of illogical Flights

Replace NA values in delay columns (AIR\_SYSTEM\_DELAY, SECURITY\_DELAY, AIRLINE\_DELAY, LATE\_AIRCRAFT\_DELAY, WEATHER\_DELAY) with 0 using the mutate(across()) function, as missing values indicate “no delay” in this dataset.

Logical checks were performed to verify the completeness of key time variables (DEPARTURE\_TIME, ARRIVAL\_TIME, DEPARTURE\_DELAY, ARRIVAL\_DELAY, AIR\_TIME) for non-cancelled flights.

1. Check for missing key time fields for non-cancelled flights

文本

AI 生成的内容可能不正确。

Figure 2.2.5.3: Code of Cross-tabulation of Cancelled vs Non-Cancelled Flights

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.5.4: Cross-tabulation of Cancelled vs Non-Cancelled Flights

Cross-tabulation confirmed that cancelled flights (CANCELLED = 1) contained missing time values as expected, while valid flights (CANCELLED = 0) need retained complete information.

1. Fix logic error

文本

AI 生成的内容可能不正确。

Figure 2.2.5.5: Validation of Logical Flight Records

Records with inconsistent or illogical time combinations were filtered and removed. For example, cancelled flights (CANCELLED = 1) that still had recorded departure times were excluded, and all delay or air time fields for cancelled flights were set to NA.

This process guaranteed internal data integrity, aligning flight status with timing information.

### 2.2.6 Check for duplicate records

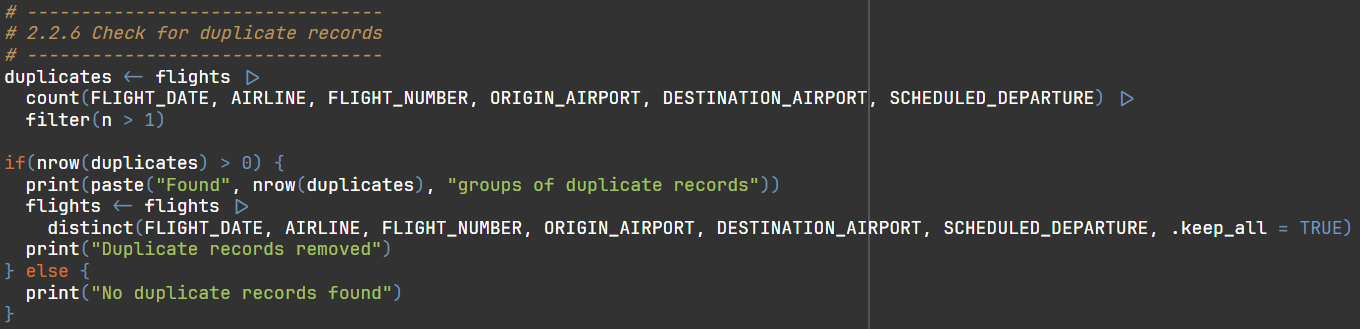


Figure 2.2.6.1: Code of Detection and Removal of Duplicate Flights

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.6.2: Detection and Removal of Duplicate Flights

Duplicates were detected using a combination of identifying fields (FLIGHT\_DATE, AIRLINE, FLIGHT\_NUMBER, ORIGIN\_AIRPORT, DESTINATION\_AIRPORT, SCHEDULED\_DEPARTURE). Duplicate records were removed using distinct() to ensure each flight appeared only once in the dataset.

### 2.2.7 Check for Outlier records

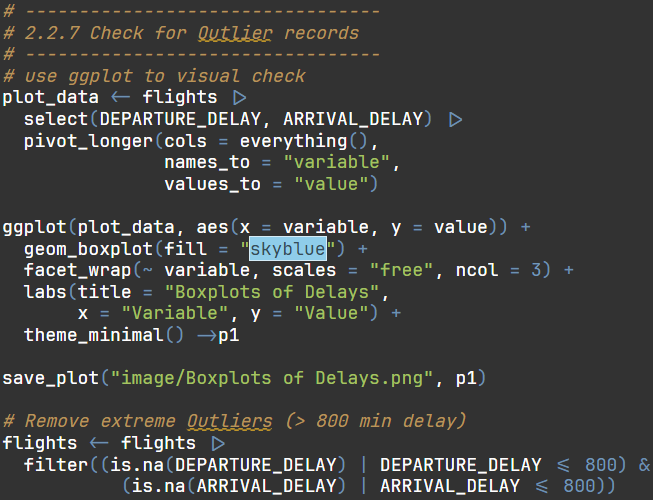


Figure 2.2.7.1: Code of Boxplots of Departure and Arrival Delays

**Output:**

图表, 箱线图

AI 生成的内容可能不正确。

Figure 2.2.7.2: Boxplots of Departure and Arrival Delays

Boxplots were created to visually inspect extreme delay values in DEPARTURE\_DELAY and ARRIVAL\_DELAY. Outliers exceeding 800 minutes were removed since they represented unrealistic or corrupted data. This process improved the accuracy of subsequent statistical models.

### 2.2.8 Identify abnormal airport codes

1. Find out if ORIGIN\_AIRPORT or DESTINATION\_AIRPORT code is numeric

文本

AI 生成的内容可能不正确。

Figure 2.2.8.1: Code of Detection of Numeric Airport Codes

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.8.2: Detection of Numeric Airport Codes

Flights containing numeric-only airport codes were identified using the str\_detect() function. The total number of flights with numeric airport codes is 8325. Since valid IATA airport codes consist of alphabetic characters (e.g., LAX, JFK)

1. Delete these abnormal airport codes

文本

AI 生成的内容可能不正确。

Figure 2.2.8.3: Removed abnormal Airport Codes

All flights with numeric-only airport codes were excluded from the dataset to maintain consistency and accuracy in airport identification.

### 2.2.9 Factorize airports and group rare ones

1. Count total flights per airport (origin + destination)

文本

AI 生成的内容可能不正确。

Figure 2.2.9.1: Code of Count Total of airport

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.9.2: Total of airport

Total flight counts for each airport were calculated by combining both ORIGIN\_AIRPORT and DESTINATION\_AIRPORT variables. This allowed identification of major airports with the highest flight volumes.

1. Print out major airports(flights>2000) list

文本

AI 生成的内容可能不正确。

Figure 2.2.9.3: Code of select of Major Airports (>2000 Flights)

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.2.9.4: List of Major Airports (>2000 Flights)

1. Recode airports code and convert airports code to factor

文本

AI 生成的内容可能不正确。

Figure 2.2.9.5: Reclassification of Airports

Airports with fewer than 2,000 flights were grouped under a new category called "Others". Both origin and destination airports were then converted into factor variables to improve categorical data consistency during analysis and modelling.

## 2.3 Data Verification

After cleaning, a verification step was performed to ensure dataset consistency.

1. Preliminary view of data

形状

AI 生成的内容可能不正确。

Figure 2.3.1: Code of Preliminary view of Cleaned Flight Dataset

**Output:**

1. Head of flights

文本

AI 生成的内容可能不正确。

Figure 2.3.2: Head of Cleaned Flight Dataset

head(flights) was used to view the first few rows of the cleaned dataset.

1. Summary of flights

图片包含 日历

AI 生成的内容可能不正确。

Figure 2.3.3: Summary Statistics of Cleaned Dataset

summary(flights) provided descriptive summaries of numerical variables, confirming appropriate value ranges and no abnormal outliers.

1. Structure of flights

图片包含 图形用户界面

AI 生成的内容可能不正确。

Figure 2.3.4: Structure of Cleaned Dataset

str(flights) verified the correct data types for each column, ensuring numeric and factor variables were properly defined.

1. Basic data overview

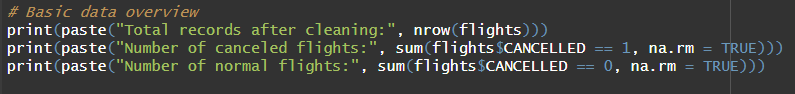


Figure 2.3.5: Code of Basic Data Overview

**Output:** 文本

AI 生成的内容可能不正确。

Figure 2.3.5: Basic Data Overview

The total number of data row after cleaning was printed, along with counts of cancelled and normal flights. This provided an overview of the dataset size and ensured that the number of valid flights matched expected totals after filtering.

1. Logical check: departure must happen before arrival

文本

AI 生成的内容可能不正确。

Figure 2.3.6: Code of Logical Time Validation Check

**Output:**

文本

AI 生成的内容可能不正确。

Figure 2.3.7: Logical Time Validation Check

All records were checked to confirm that DEPARTURE\_TIME occurred before ARRIVAL\_TIME. Flights violating this rule were identified and removed, guaranteeing temporal consistency across all remaining records.

1. Final missing value check

文本

AI 生成的内容可能不正确。

Figure 2.3.8: Code of Final Missing Value Report

**Output:**

图形用户界面

AI 生成的内容可能不正确。

Figure 2.3.9: Final Missing Value Report

A final missing value summary was generated to confirm that NAs only existed in valid cases like cancelled flights. Besides, confirm that no critical analysis variable contained missing data.

# 3.0 Analysis

## 3.1 Objective 1: To investigate the relationship between departure delay and arrival delay, including the impact of additional flight factors (e.g., distance, air time and flight date), and explore their overall correlation and trend. – Chan Min Huey (TP083261)

### 3.1.1 Analysis 1-1: To analyse the distribution of departure delay and arrival delay in order to understand their basic characteristics.

|  |  |
| --- | --- |
| Type | Descriptive Analysis |
| Independent Variable | Departure Delay (Continuous data) |
| Dependent Variable | Arrival Delay (Continuous data) |
| Analysis Technique Used | Summary Statistics (Mean, Median, Standard Deviation, Range) |
| Visualisation | Faceted frequency polygon and Combined Boxplot with Jittered Points |

1. **Summary Statistics (Mean, Median, Standard Deviation, Summary)**

**A screenshot of a computer code

AI-generated content may be incorrect.**

*Figure 3.1.1: Steps to perform Summary Statistics (Mean, Median, Standard Deviation, Range)*

**Output ：**

**A screenshot of a computer code

AI-generated content may be incorrect.**

*Figure 3.1.1.2: Resu*lt ofSummary Statistics (Mean, Median, Standard Deviation, Range)

The mean shows that departure delay (9.63 minutes) has longer delay than arrival delay (4.81 minutes). The median for arrival delay and departure delay is -5 and -1 respectively, which means at least half of the flights departed or arrived earlier. The standard deviation of arrival delay is 38.6 and departure delay is 36.3, indicate high variability in both delay type. Lastly, the range of arrival delay is -74 to 797 while departure delay has a range of -42 to 796.

1. **Visualisation**
2. **Faceted Frequency Polygon**

A computer screen shot of a program

AI-generated content may be incorrect.

*Figure 3.1.1.3: Step of plotting faceted frequency polygon*

**Output:**

A graph of a graph showing a line of different times

AI-generated content may be incorrect.

*Figure 3.1.1.4: Result of plotting faceted frequency polygon*

Figure 3.1.4 shows a faceted frequency polygon that display the frequency distribution of arrival delay and departure delay. Both frequencies are right-skewed distributions, which means that most flight arrive/ depart earlier or on-time. In arrival delay, the most count of flights delay is about -5 which means majority of flights arrive earlier. While departure delay, the most count of flights delay is about -1 which means majority of flights are on time. Arrival delay shows a larger distribution while departure delay has more concentrated distribution.

1. **Combined Boxplot with jittered points**

**A screenshot of a computer program

AI-generated content may be incorrect.**

*Figure 3.1.1.5: Step of plotting Combined Boxplot with Jittered Points*

**Output:**

A screen shot of a graph

AI-generated content may be incorrect.

*Figure 3.1.1.6: Result of plotting Combined Boxplot with Jittered Points*

The combined boxplot with jittered points in Figure 3.1.1.6 highlights the central tendency, extreme delays, and variability. Both arrival delay and departure delay have median below 0 (negative value), indicates that at least half of all the flights are early or on time. However, mean values (shown as red diamonds in the figure) are higher due to extreme delays. Most flights fall within the 0–99minute range (orange), with few severe ranges (blue). Departure delays have more extreme outliers, whereas arrival delays show a higher variability within the normal ranges.

**Additional analysis: Does flight distance affects the distribution?**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

***A screen shot of a graph

AI-generated content may be incorrect.***

The boxplot with jittered points is to explore how flight distance influence delay distributions (Distance group: long-haul, medium-haul and short haul). Long-haul distance group have fewer outliers and concentrated distribution, showing more stable operations. Medium-haul has moderate delay volatility, while short-haul distance group have the widest delay spread and more frequent outliers, which shows greater instability. In conclusion, the results show that shorter routes are more prone to unstable and extreme delays, while longer routes have greater operational stability.

### 3.1.2 Analysis 1-2: To examine the relationship between departure delay and arrival delay.

|  |  |
| --- | --- |
| Type | Diagnostic Analysis |
| Independent Variable | Departure Delay (Continuous data) |
| Dependent Variable | Arrival Delay (Continuous data) |
| Analysis Technique Used | Correlation Analysis (Pearson Correlation) with hypothesis testing |
| Visualisation | 2D Binned Heatmap with Regression Line, Correlation Heatmap |

1. **Correlation Analysis (Pearson Correlation) with hypothesis testing**

**A computer code with text

AI-generated content may be incorrect.**

*Figure 3.1.7: Steps to perform Correlation Analysis (Pearson Correlation) with hypothesis testing*

**Output:**

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.1.8: Result of Correlation Analysis (Pearson Correlation) with hypothesis testing*

The pearson correlation between departure delay and arrival delay is very strong because their correlation is 0.943, which is close to 1. In pearson correlation, 0.943 represent a strong positive linear relationship, meaning that if a flight experiences a departure delay, the probability of an arrival delay is extremely high. The extreme low p-value,2.2e-16 (less than 0.001), confirms that the relationship between departure delay and arrival delay is statistically significant and highly unlikely to have occurred by chance.

1. **Visualisation**

**2D Binned Heatmap with Regression Line**

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.1.9: Steps of plotting 2D Binned Heatmap with Regression Line*

**Output:**

**A graph with a red line

AI-generated content may be incorrect.**

*Figure 3.1.10: Result of plotting 2D Binned Heatmap with Regression Line*

Figure 3.1.10 shows the 2D Binned Heatmap with Regression Line for departure and arrival delays. Most of the flights point gather at the origin point, which means that majority of flights only experience minor delays or no delay for both departure delay and arrival delay. The regression line’s pronounced upward slope reflects a strong positive linear relationship between departure delays and arrival delays (r=0.94, p<0.001), confirms that departure delay strongly affect arrival delays. Although there is some outliers in the upper-right, but they don’t affect overall linear association.

### 3.1.3 Analysis 1-3: To evaluate whether schedule depature with flight factors like flight distance, and air time significantly predict arrival delay.

|  |  |
| --- | --- |
| Type | Predictive Analysis |
| Independent Variables | Schedule departure, Distance, Air Time (Continuous data) |
| Dependent Variable | Arrival Delay (Continuous data) |
| Analysis Technique Used | Multiple Linear Regression |
| Visualisation | Residuals vs Fitted Plot, Observed vs Predicted Plot |

1. **Multiple Linear Regression**

**A screenshot of a computer program

AI-generated content may be incorrect.**

*Figure 3.1.11: Step to perform multiple linear regression*

**Output :**

**A screenshot of a computer program

AI-generated content may be incorrect.**

*Figure 3.1.12: Result of multiple linear regression*

The multiple linear regression shows that departure delay is the strongest predictor of arrival delay (Estimate = 1.006, p <0.001), which means that each minute additional in departure delay increases approximately one minutes with arrival delay. Flight distance (Distance) has a small negative effect (Estimate = -0.04, p<0.001) on arrival delays, suggest that longer flights can recover time. Air time has a small positive effect (Estimate = 0.33, p<0.001) with arrival delays, showing longer flights can slightly increase arrival delay. The model explains 90.2% of the variance in arrival delay (R-squared =0.902).

1. **Visualisation**
2. **Residual vs Fitted Plot**

**A screen shot of a computer code

AI-generated content may be incorrect.**

*Figure 3.1.13: Step of plotting residual vs fitted plot*

**Output :**

**A graph with a number of colored dots

AI-generated content may be incorrect.**

*Figure 3.1.14: Result of plotting residual vs fitted plot*

Figure 3.1.14 shows the residual vs fitted plot for the multiple regression model. Most of the residual scattered around zero, showing the model fits the overall. The gradient colour shows overestimated (yellow), accurate (green), and underestimated (red). Although there are some outliers, but most of the flights data points fall within a stable residual range, which means that this regression model provides reliable predictions. The RSME for this model is 12.08 minutes average.

1. **Observed vs Predicted Plot**

**A screen shot of a computer program

AI-generated content may be incorrect.**

*Figure 3.1.15: Step of plotting observed vs predicted plot*

**Output :**

**A green line with red dots

AI-generated content may be incorrect.**

*Figure 3.1.16: Result of plotting observed vs predicted plot*

The observed vs predicted plot in Figure 3.1.1.16 compares the predicted arrival delay with the actual arrival delay. Most of the data points align closely with the 45 ° red dotted line, showing that the model prediction mostly is accurate. Those data points that do not align with the dotted lines are prediction errors, where the model over or underestimated the arrival delay. This confirms that departure delay, together with distance and air time, provides strong predictive power for arrival delays.

### 3.1.4 Analysis 1-4: To provide recommendations on managing different departure delay ranges and evaluate whether weekday/weekend patterns influence arrival delays.

|  |  |
| --- | --- |
| **Type** | Prescriptive Analysis |
| **Independent Variables** | Departure Delay (Continuous data)  Flight date (continuous data) |
| **Dependent Variable** | Arrival Delay (Continuous data) |
| **Analysis Technique Used** | Group Comparison (ANOVA / t-test with descriptive summaries) |
| **Visualisation** | Violin Plot with Boxplot Overlay, Bar Plot with Error Bars (Mean + 95% CI) |

1. **Group Comparison (ANOVA and descriptive summaries)**

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.1.17: Step to perform Group Comparison (ANOVA and descriptive summaries)*

1. **T-test**

**A black text on a white background

AI-generated content may be incorrect.**

*Figure 3.1.18: Step to perform T-test*

**Output:** **A screenshot of a computer code

AI-generated content may be incorrect.**

*Figure 3.1.19: Result of perform T-test*

The t-test shows weekType (weekdays or weekends) has a statistically significant difference in arrival delays (p<0.001). However, the mean difference between weekdays and weekends is only about 2 minutes, which is negligible relative to the overall delay scale.This indicates that week type has no meaningful practical impacts on arrival delays.

1. **Visualisation**
2. **Violin Plot with Boxplot Overlay**

**A computer screen shot of a program

AI-generated content may be incorrect.**

*Figure 3.1.20: Step of plotting Violin Plot with Boxplot Overlay*

**Output:**

**A graph of a schedule

AI-generated content may be incorrect.**

*Figure 3.1.21: Result of plotting Violin Plot with Boxplot Overlay*

Figure 3.1.21 shows violin plot with boxplot overlay comparing with arrival delays across five departure delay groups, separated by weekdays(magenta) and weekends (blue). Early departure group (<0) consistently arrive early, while slight delays (0-29) have little impact on arrival delay. From moderate delay (30-59) to heavy delay (60-119), delay variability increases. In severe delay (120+), arrival delays can exceed several hundred minutes caused by extreme departure delays. Across all group, weekday and weekend patterns are similar, confirms that departure delay drives arrival delay outcome, rather than week type.

1. **Bar Plot with Error Bars (Mean + 95% CI)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

*Figure 3.1.22: Step of plotting Bar Plot with Error Bars (Mean + 95% CI)*

**Output:**

**A graph with blue and purple bars

AI-generated content may be incorrect.**

*Figure 3.1.23: Result of plotting Bar Plot with Error Bars (Mean + 95% CI)*

Figure 2.1.23 shows a bar plot with error bar comparing mean arrival delay across five departure delay groups, separated by weekday(magenta) and weekend (blue). The bar shows the average arrival delay for each group, while the black error bars represent the 95% CI (confidence intervals), which shows the range within which the true mean is expected to fall. Mean arrival delay increases with larger departure delays: early (<0) and slight (0-29) groups shows nearly 0 delays, while moderate (30-59), heavy (60-119) and serve (120+) show higher delays. Overlapping error bars (95%CI) show similar weekday-weekend patterns and confirm no significant difference, showing a stable and reliable trend.

### 3.1.5 Extra Features

* Additional variables: distance –grouped by DistanceGroup in analysis 1-1 to determine whether distance affect delay distribution and assist departure delay on analysis 1-3 for stronger prediction
* Additional variables: airtime – to assist departure delay on analysis 1-3 for stronger prediction
* Additional variables: flight date –group by WeekType to assist delay group management on analysis 1-4

### 3.1.6 Conclusion

This analysis highlights the relationship between departure delay and arrival delay, including the impact of additional flight factors: distance, airtime and flight date. Among these 4 variables, departure delay is the most significant factor causing arrival delay. While factors like distance and airtime play secondary roles (significant, less meaningful impact) and flight date have no meaning impact on arrival delays. Both delay distribution is right-skewed, with most flights are early but a high mean due to outliers. A strong positive correlation (r=0.943, p=0.001) shows that late departures lead to late arrivals. Regression analysis confirms that departure delay as the strongest predictor, following by distance (-0.04) and airtime (0.03). Weekend versus weekday comparisons show no significant difference. Overall, for this analysis, it clearly highlights the central role of departure delay in determining overall flight punctuality.

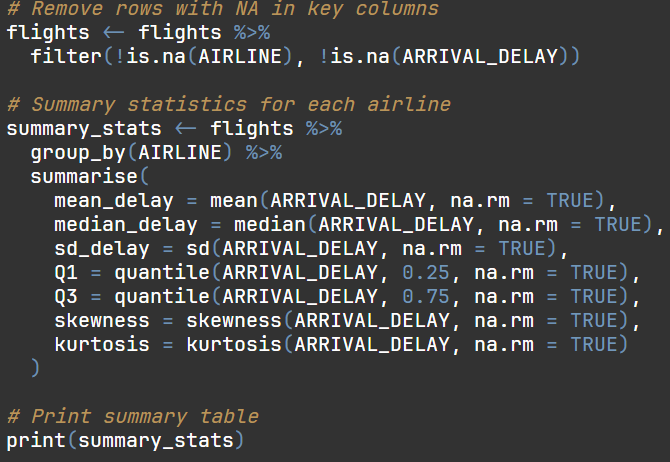
### 3.1.7 Recommendation

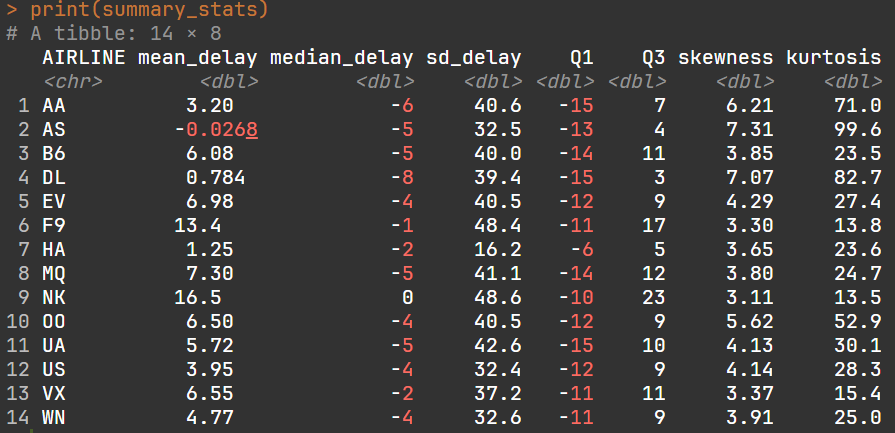
Airlines should focus on reducing departure delays to minimize arrival delays. For example, aircraft could reduce delay time through efficient ground operations and early intervention to the flights that are at risk with severe delays. Next, adjusting short-haul flights schedules also helps to minimize the arrival delay because long-haul flights can sometimes recover lost time. Overall, departure delay management and proactive planning are important for improving.

## 3.2 Objective 2: To evaluate the performance of different airlines in managing flight delays – Kang Hong Qian (TP081205)

### 3.2.1 Analysis 2-1: Flight Delay Distribution by Airline

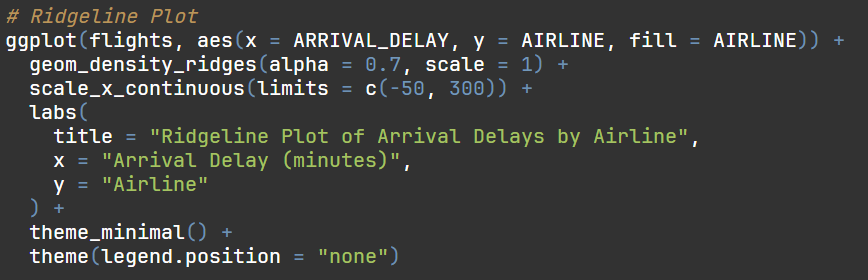
|  |  |
| --- | --- |
| **Type** | Descriptive Analysis |
| **Independent Variable** | Airline |
| **Dependent Variable** | Arrival Delay (minutes) |
| **Analysis Technique Used** | Summary Statistics and Distribution |
| **Visualisation** | Violin Plot with Boxplot Overlay + Median Points |



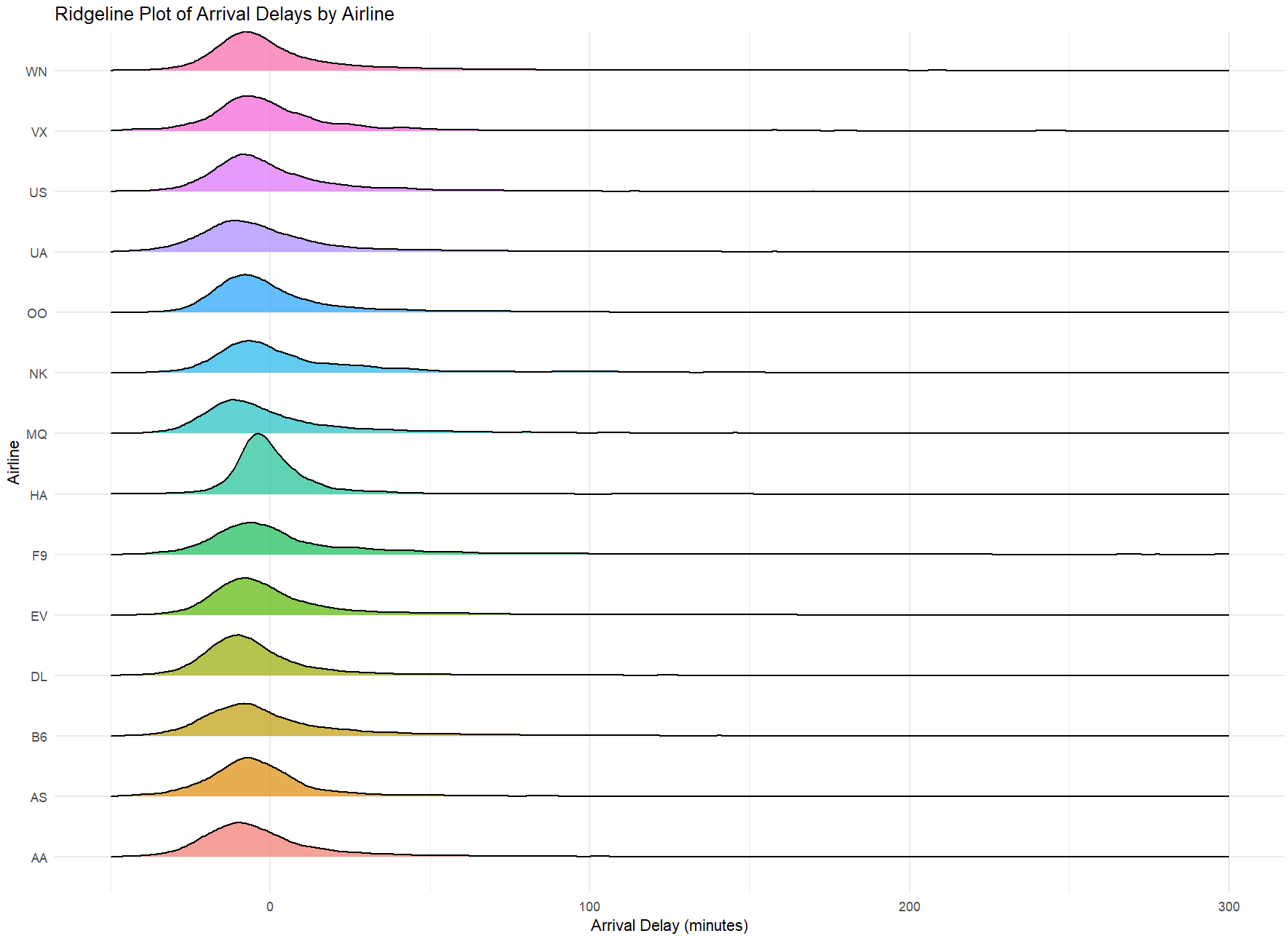


The summary statistics provide a quantitative description of airline arrival delay. Spirit Airlines (NK) had the highest median delay of about 16.5 minutes and the highest variability, while American Airlines (AA), Delta Air Lines (DL), and United Airlines (UA) had earlier median arrival and more stable performance.

**Ridgeline Plot**

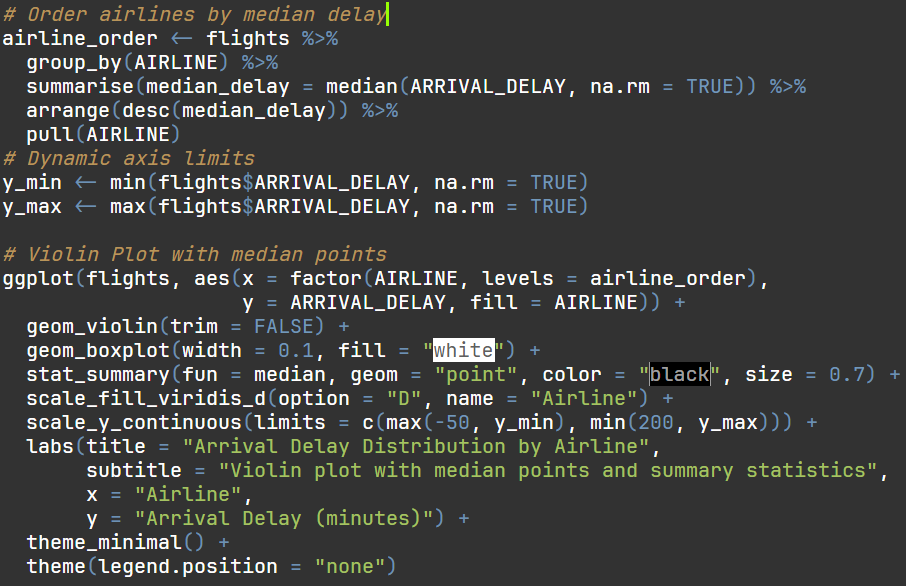


**Output:**

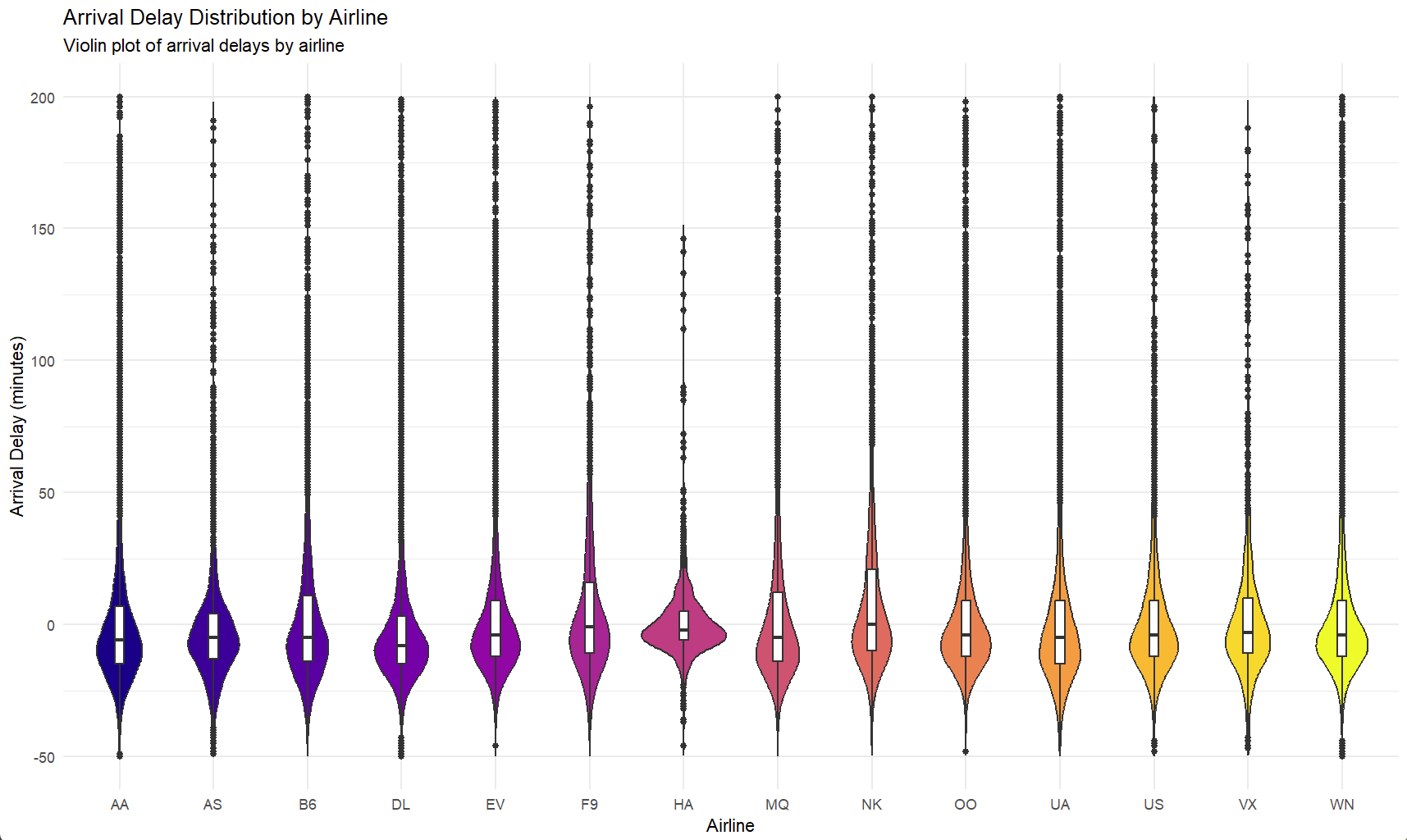


The Ridgeline Plot presents the complete distribution of arrival delay for each carrier. Spirit Airlines (NK) presents the widest and highest right-skewed distribution, meaning frequent as well as sporadic extreme delays. Comparably, major US carriers like American Airlines (AA), Delta Air Lines (DL), and United Airlines (UA) have slim distributions clustered around zero, implying steadier on-time performance. Ridgeline ordering, coupled with sorting by median delay, enables straightforward visual inspection of central tendency and variability across carriers, as well as which have greater or more erratic delays.

**Violin Plot**



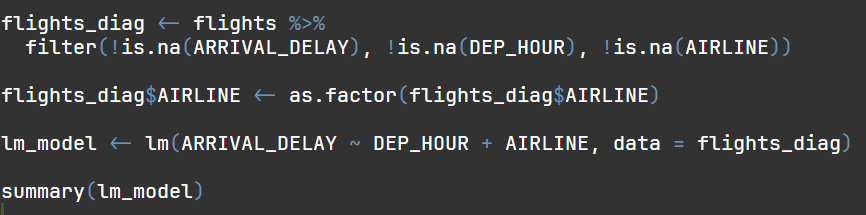
**Output:**

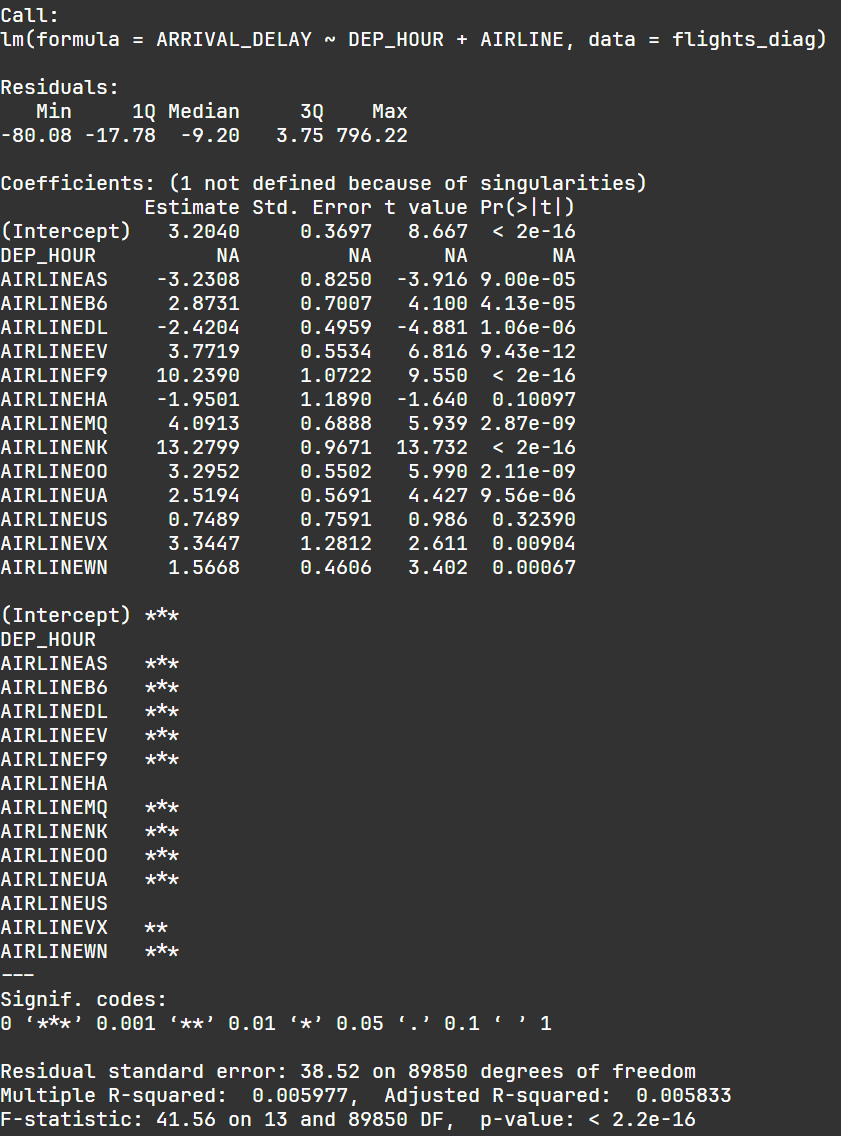


The Violin Plot presents the complete distribution of arrival delay per airline. Spirit Airlines (NK) had the highest average delay of approximately 16.5 minutes, implying inefficient delay management. Subsequently, NK still exhibited high variability in delay. Conversely, major-domestic carriers, including American Airlines (AA), Delta Air Lines (DL), and United Airlines (UA), indicated negative median delays, implying flights tended to arrive on schedule or slightly ahead, implying efficient scheduling. Low-cost carriers had wide distributions, indicating increased variability and inconsistent on-time performance.

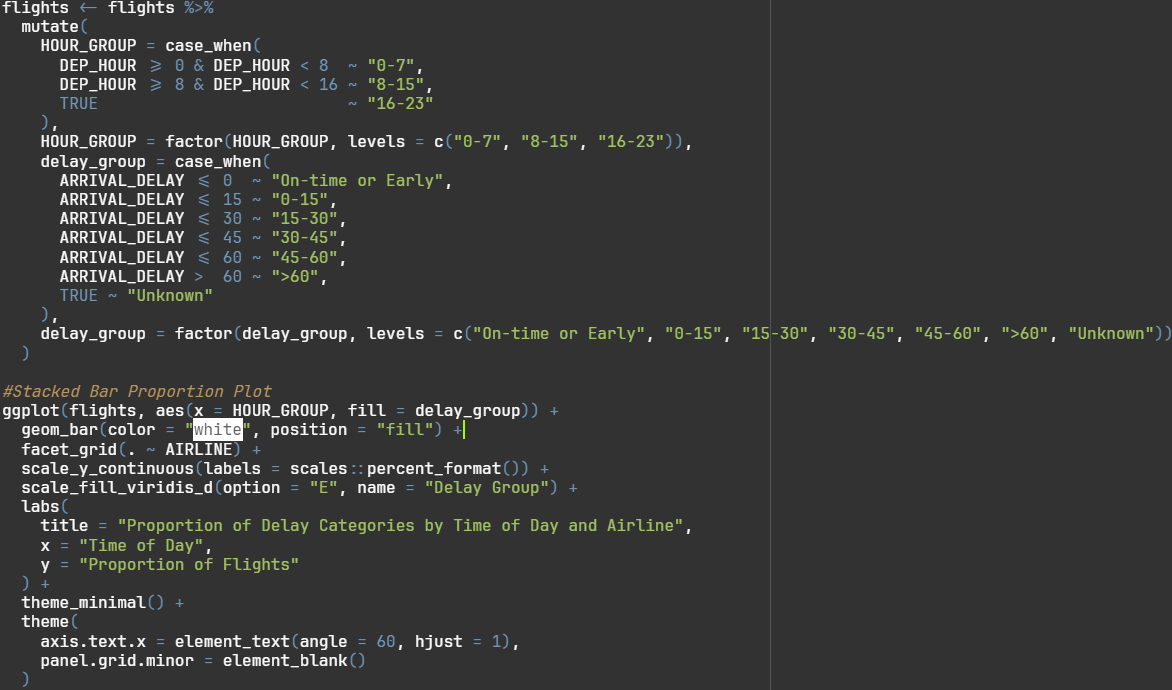
### 3.2.2 Analysis 2-2: Impact of Departure Time and Airline on Delay

|  |  |
| --- | --- |
| **Type** | Diagnostic Analysis |
| **Independent Variable** | Scheduled Departure Hour, Airline |
| **Dependent Variable** | Arrival Delay (minutes) |
| **Analysis Technique Used** | Multiple Linear Regression + Categorical Grouping |
| **Visualisation** | Stacked Bar Chart (by Airline and Delay Group) |

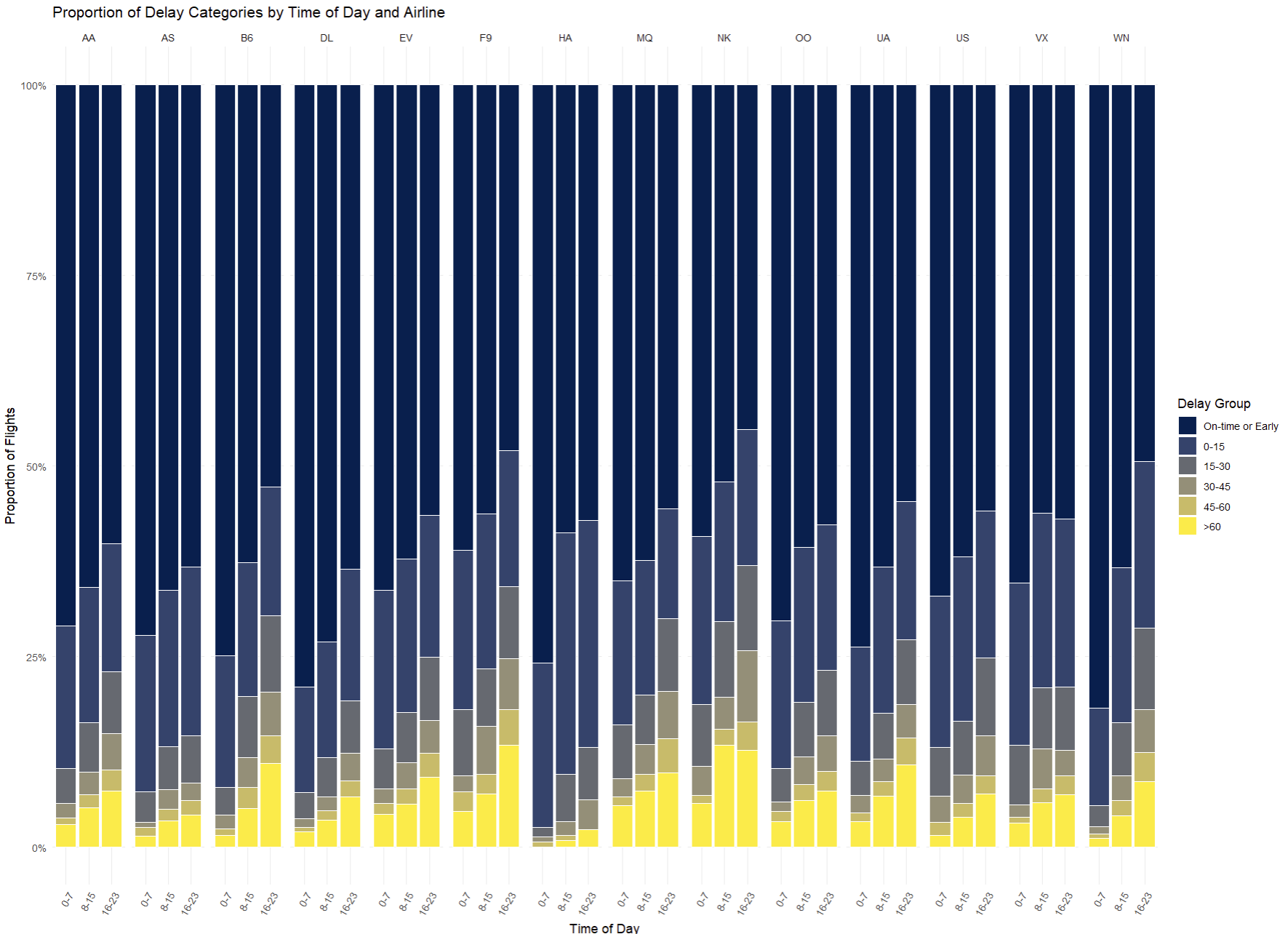




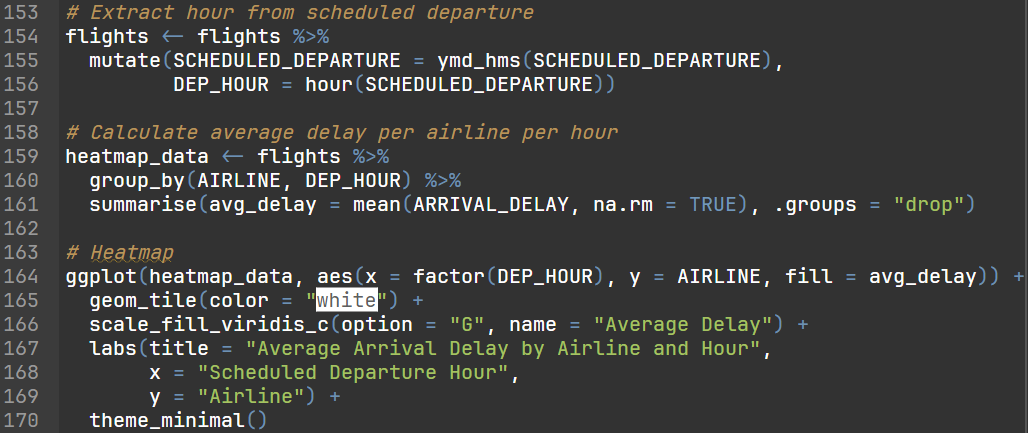
The multiple regression model, estimating arrival delay as a function of departure hour and airline, supports these insights. Airlines like Spirit Airlines (NK) had much larger delays, and airlines like American Airlines (AA), Delta Air Lines (DL), and United Airlines (UA) had smaller, more regular delays. Even though the R-squared on the model is small (about 0.006), such that there are many other factors that compose delay, the coefficients indicate a definite impact of airline selection on arrival delay. This statistical evidence reinforces the stacked bar plot, verifying that time of day as well as airline identity is a constituent part of delay patterns.

**Stacked Bar Proportion Chart** 

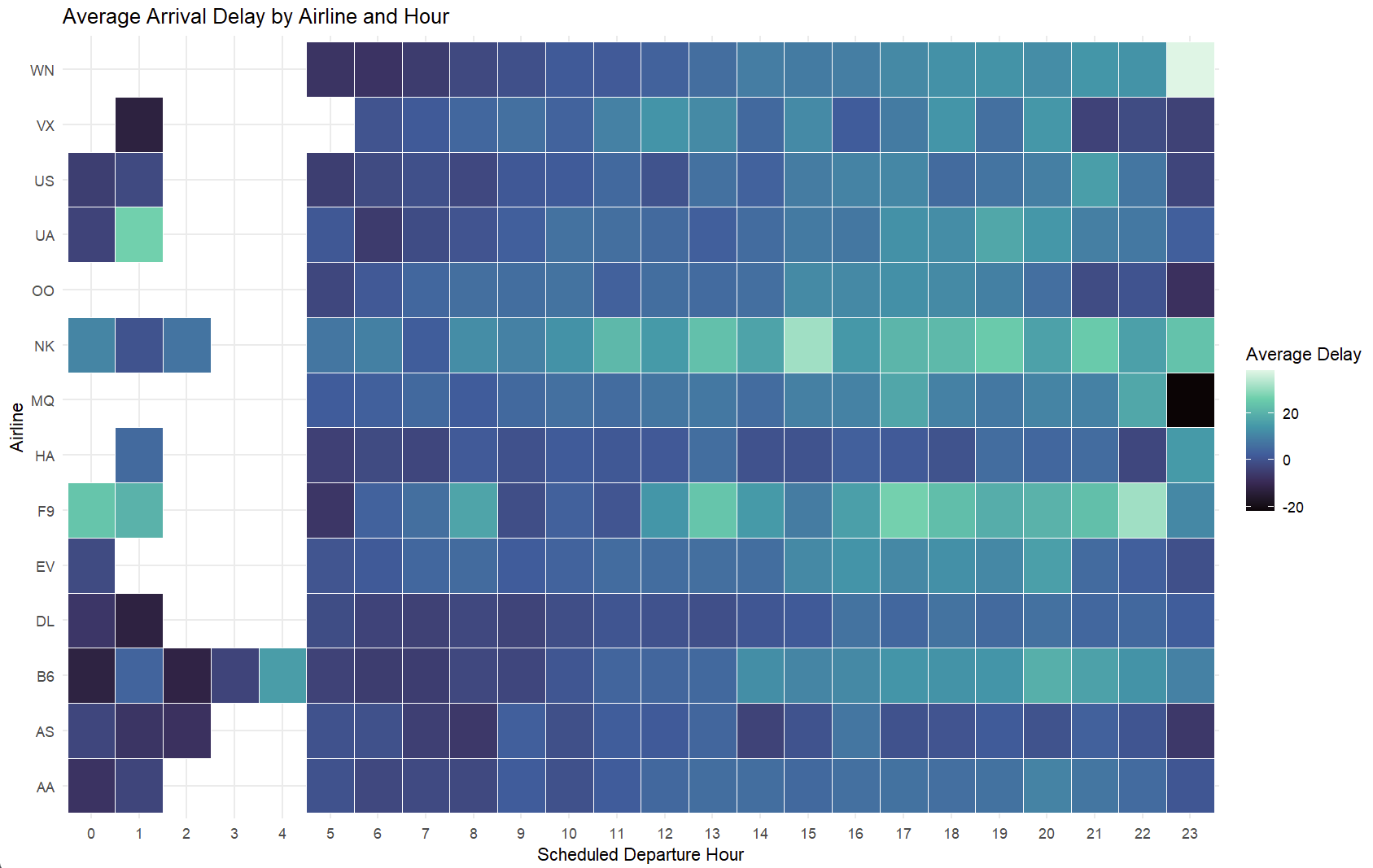
**Output:**



The choice of a stacked bar proportion chart is to display the relative proportions of flight delays across different times of the day for each airline. From the chart, it is evident that on-time or short-delay flights (dark blue sections) make up a larger proportion in the morning (0–7 and 8–15 hours), whereas long delays (yellow sections) become more common during late-day flights (16–23 hours). The visualization shows that some airlines, like NK and F9, have significantly higher rates of severe delays, which last more than 60 minutes. This highlights the differences in how carriers manage delays.

**Heatmap：**  


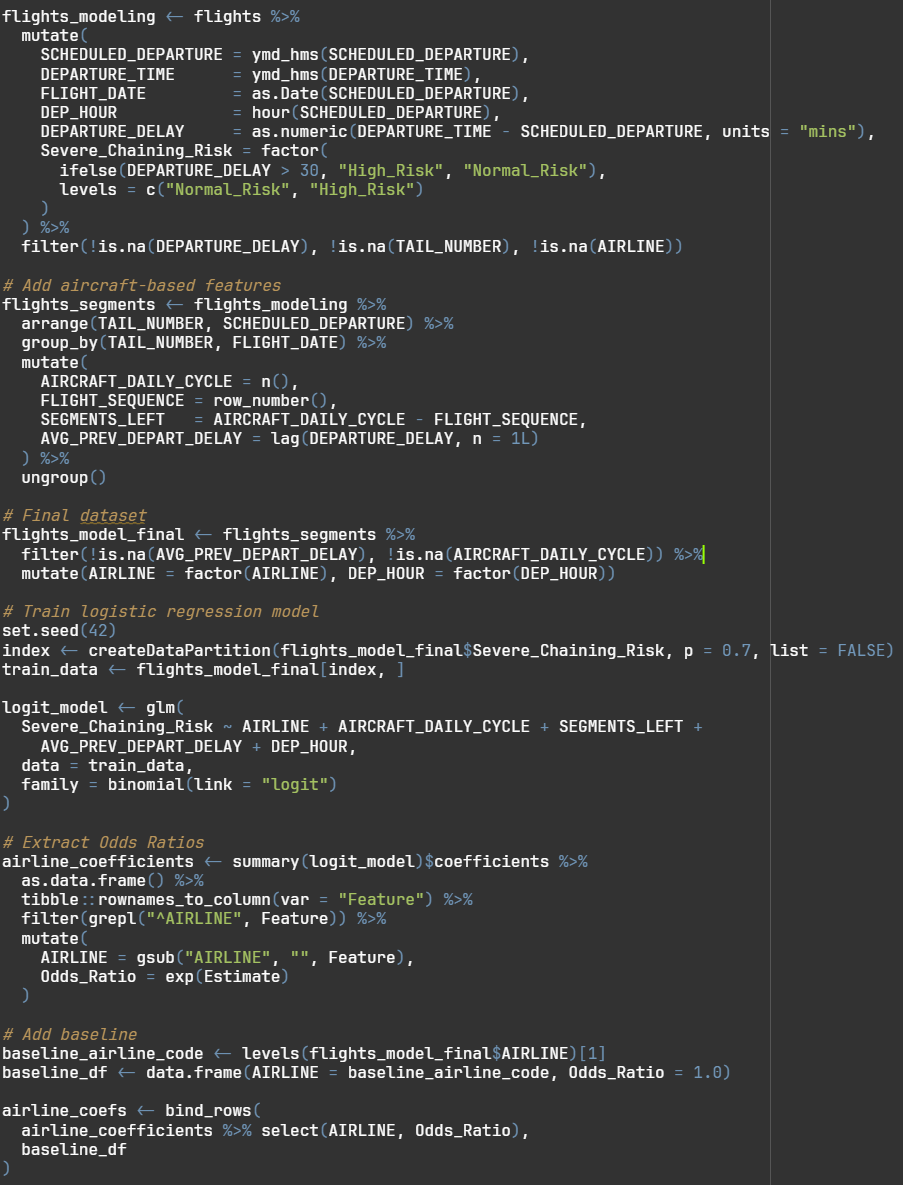
**Output：**

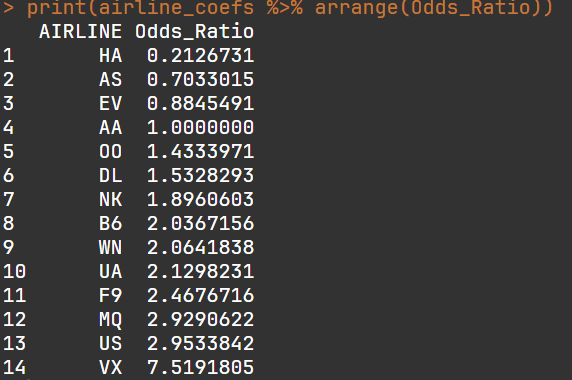


The use of a heatmap is to indicate the mean arrival delay per hour of the day by airline. Plotting on the x-axis the schedule departure hour and on the y-axis the airline enables the straightforward comparative observation of performance over time. Analysis results show that delays are generally lower in the early morning hours and rise in the busy periods, implying that congestion and operational issues increase as the day goes by.

### 3.2.3 Analysis 2-3: Assessing Chaining Delay Risk Using Logistic Regression

|  |  |
| --- | --- |
| **Type** | Predictive Analysis |
| **Independent Variables** | Airline, Departure Hour, Aircraft Daily Cycle, Previous Departure Delay, Segments Left |
| **Dependent Variable** | Severe Chaining Risk (High Risk vs Normal Risk) |
| **Analysis Technique Used** | Logistic Regression with Operational and Temporal Predictors |
| **Visualisation** | Bar Chart of Odds Ratio (Performance Ranking by Airline) |





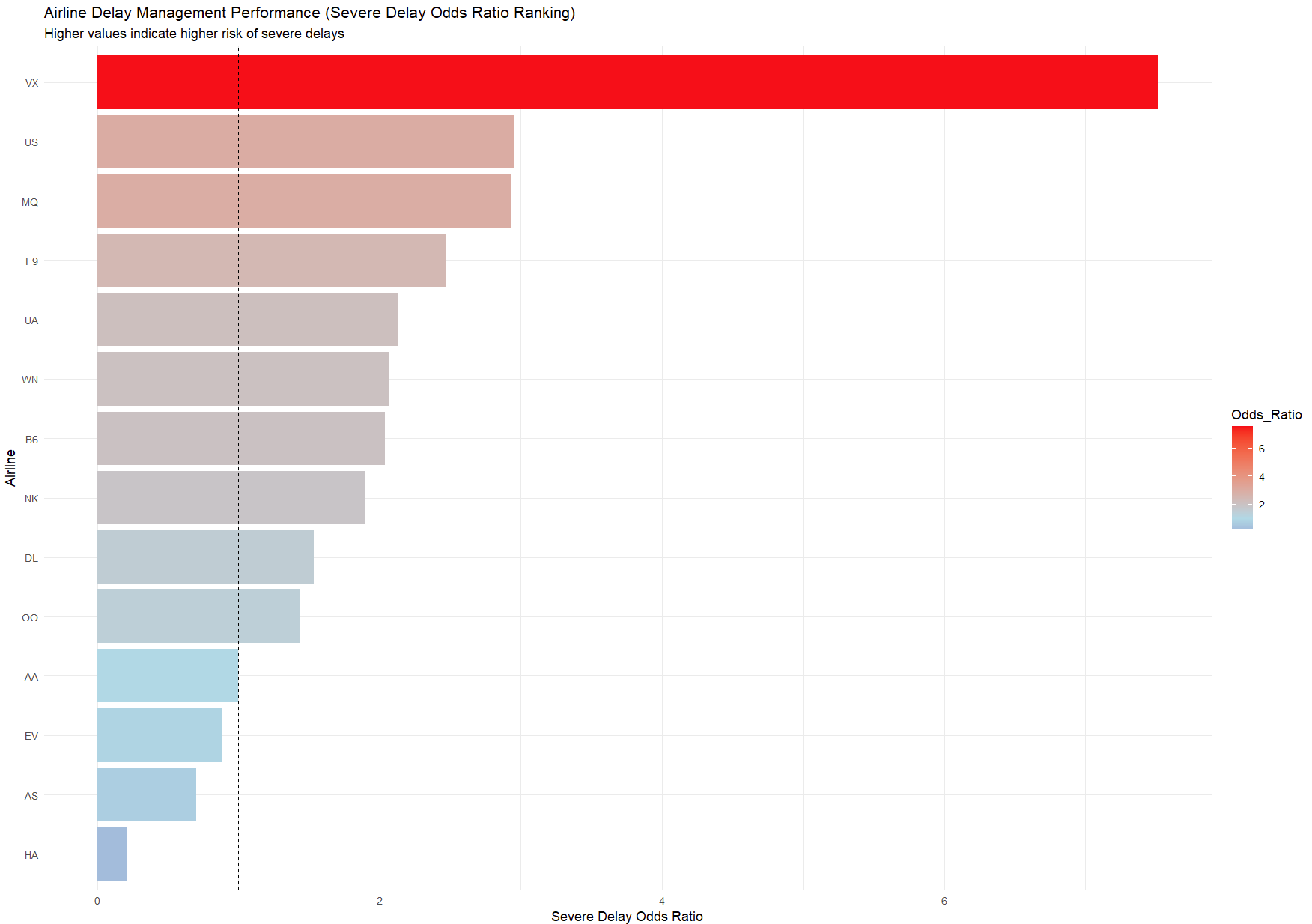
A logistic regression model was created to predict the chance of severe chaining delays. These are cases where a flight’s departure delay goes beyond 30 minutes. The results indicate that past departure delay is the strongest predictor. When one delay happens with an airplane, the subsequent flights.

The odds ratios from the model provide a performance ranking between each airline. Airlines whose odds ratios were greater than 1.0 had a greater likelihood of high-severity chaining delay. This indicates lower efficiency in managing delays. On the other hand, airlines with odds ratios below 1.0 showed better resilience in turn-around times.

**Bar Chart**



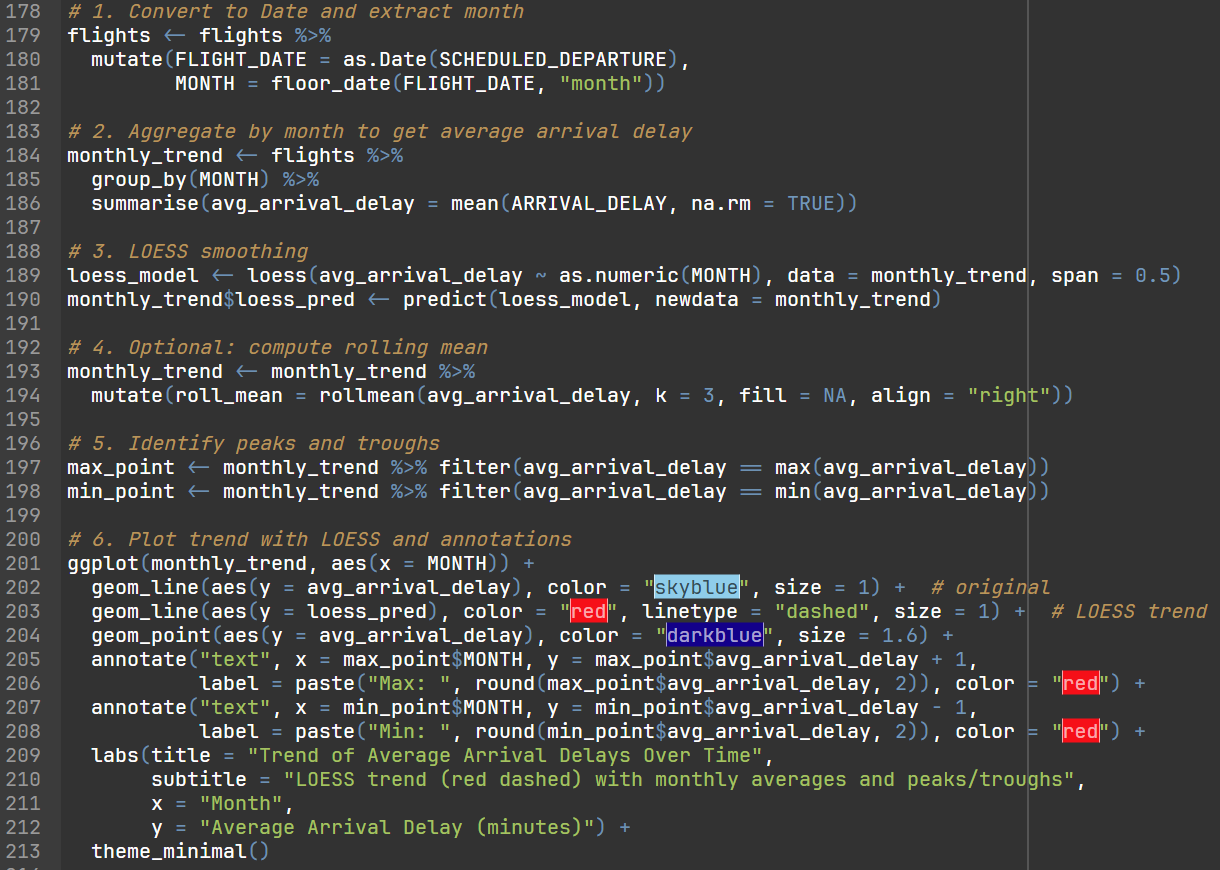
**Output:**



The visualisation, a bar chart of odds ratios, clearly highlights performance differences among airlines. From a managerial perspective, this predictive insight helps identify which carriers should improve turnaround management, add schedule buffers, or enhance delay recovery strategies to minimize cascading disruptions across their operations.

### 3.2.4 Analysis 2-4: Identifying Monthly and Seasonal Delay Patterns

|  |  |
| --- | --- |
| **Type** | Trend / Prescriptive Analysis |
| **Independent Variable** | Month |
| **Dependent Variable** | Average Arrival Delay (minutes) |
| **Analysis Technique Used** | LOESS Smoothing on Monthly Average Arrival Delay |
| **Visualisation** | Line Chart with LOESS Curve, Rolling Average, and Annotated Peaks/Minima |
| **Additional Feature** | LOESS smoothing highlights underlying trends: rolling mean and annotated max/min months provide actionable insights |



**Output：**



The time-based change in average delay arrival is illustrated with the use of a line chart. LOESS smoothing is utilized to emphasize the trend, minimizing the effect from short-term variability, and a rolling mean is also used as a further basis on which local patterns can be identified. Months with the highest delay and least delay are highlighted through the use of annotations such that the seasonality or cyclical patterns are emphasized.

The preliminary estimate indicates significant variability in the monthly average delay. February had the highest peak delay over the sample period (~9.14 minutes), and September had the lowest (~ -1.05 minutes, or flights arriving ahead of schedule often). It is reasonable that these extremes reflect seasonally varying factors like weather, holiday demand, and operational constraints.

### 3.2.5 Extra Features

1. **Additional Variable**: **DEP\_HOUR** – from SCHEDULED\_DEPARTURE to examine the effect of departure time on arrival delay in separate analyses (Analysis 2–3).
2. **Additional Variable**: **HOUR\_GROUP** – coded from DEP\_HOUR to categorize flights into morning, afternoon, and evening time periods, allowing visual comparison in delay proportion analysis (Analysis 2).
3. **Additional Variable**: **delay\_group** – extracted from ARRIVAL\_DELAY in order to categorize severity levels of delay, increasing interpretability in diagnostic visualization (Analysis 2).
4. **Additional Variable**: **Severe Chaining Risk** – created to categorize flights into “High Risk” or “Normal” delay groups, enabling predictive modelling of chaining delays (Analysis 3).

### 3.2.6 Conclusion

The impact of departure time, airline, and operational factors on aircraft delays is examined in this investigation. The results indicate a unique pattern: delays increasingly rise during the day, with evening flights experiencing longer and more erratic delays than morning flights.  
Arrival delays are highly impacted by both the airline and the departure hour, according to the regression study. Hawaiian (HA) and Delta (DL) were timelier than Spirit Airlines (NK), which had the longest delays.

As the stacked bar chart illustrates, early in the day, on-time arrivals predominate, whereas late flights frequently have delays of more than 60 minutes. Additionally, logistic regression indicates that the greatest impact on new delays is caused by past departure delays, with airlines like as Virgin America (VX) and Spirit (NK) exhibiting higher chain reactions. On the other hand, Hawaiian (HA) exhibits proficient turnaround management.

Finally, the trend observation via LOESS smoothing points out the presence of monthly variability, with the delay peak occurring in February and the least delay time in September. In combination, these outcomes indicate that the timing of operations, management practices by the airlines, and the day-of-month cycles all have significant contributions towards flight punctuality.

### 3.2.7 Recommendation

Departure punctuality and schedule operations should be given importance by airlines to reduce cumulative and chain delays. Ground coordination can be strengthened, and rotation plans for the planes adjusted to stop delay escalation during peak periods. Buffer time between flights can be included and real-time monitoring enhanced as specific remedies for carriers having high risk of chaining like VX and NK.

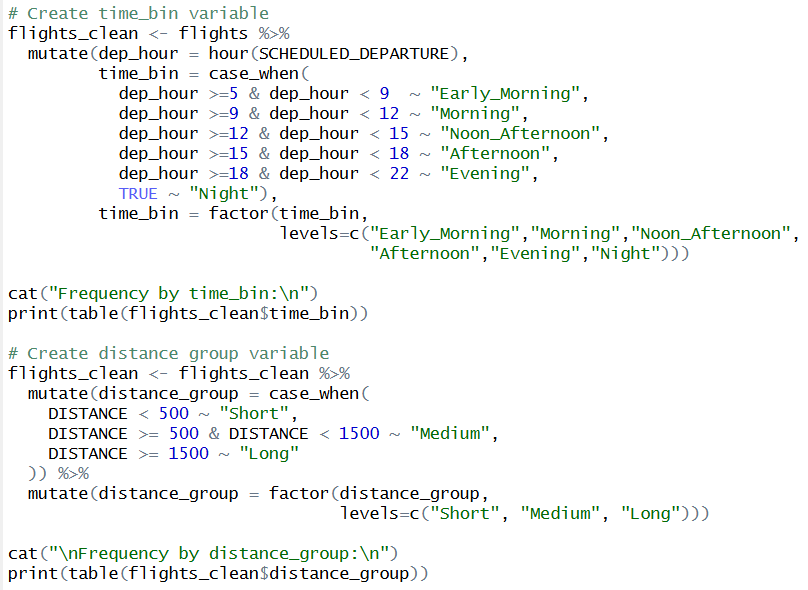
Also, studying season delay patterns allows the airlines to invest more efforts in the high-delay months such as June and February, without jeopardizing the efficiency in the low-delay months such as September. On-going performance monitoring, and the integration of predictive modelling, can also assist the airlines in predicting potential disruptions as well as boosting overall on-time performance.

## 3.3 Objective 3: To analyse how various factors , including time of day, day of week, flight distance, and departure delay, influence flight arrival delays, and to identify the relationships that contribute to overall delay performance. – Gong Yee Cheng (TP081910)

### 3.3.1 Analysis 3-1: Assessing Time-of-Day and Flight Distance Interaction Effects on Flight Arrival Delays using Two-Way ANOVA

|  |  |
| --- | --- |
| Type | Diagnostic Analysis |
| Independent Variable | time\_bin and distance\_group |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Two-Way ANOVA (analysis of variance with interaction term) |
| Visualisation | Combined boxplot showing arrival delay by time and distance, heatmap of mean arrival delay across time–distance combinations |

**Create time\_bin variable and create distance\_group variable**



**Output:**



A close-up of a computer code

AI-generated content may be incorrect.

1. **Two-way ANOVA (Time x Distance)**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A screenshot of a computer

AI-generated content may be incorrect.

1. **Visualisation**
   1. **Boxplot with both factors**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A graph of different colored lines

AI-generated content may be incorrect.

* 1. **Mean Arrival Delay Heatmap**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A chart of different shades of red

AI-generated content may be incorrect.

This analysis uses Two-Way ANOVA to analyse how time of day and flight distance affect arrival delays. The flights departing in the afternoon and evening generally have longer delayed. The interaction effect shows that time of day impact depends on distance, long-haul flights tend to be more delayed later in the day. Boxplot and heatmap display these time distance delay patterns, providing better scheduling and delay management insight.

### 3.3.2 Analysis 3-2: Exploring Weekly Delay Trends through Descriptive Statistical Profiling

|  |  |
| --- | --- |
| Type | Descriptive Analysis |
| Independent Variable | DAY\_OF\_WEEK (1 = Monday ... 7 = Sunday) |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Mean and median summary; bar chart of average delays per weekday. |
| Visualisation | Barplot showing mean ARRIVAL\_DELAY by weekday; optional boxplot for variability. |

1. **Mean and Median Summary**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A screenshot of a computer code

AI-generated content may be incorrect.

1. **Visualisation**

**Barplot of Mean arrival delay by day of week**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A graph of a graph

AI-generated content may be incorrect.

This descriptive analysis examines how average flight delays vary by weekday using the wday() function. Mean and median delays were summarized and visualized in a bar chart. Results show higher delays on Fridays and Sundays, while midweek flights are usually more punctual, reflecting typical travel demand patterns across the week.

### 3.3.3 Analysis 3-3: Examining the Relationship between Flight Distance and Arrival Delay using Spearman Rank Correlation

|  |  |
| --- | --- |
| Type | Correlation Analysis |
| Independent Variable | DISTANCE (continuous) |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Spearman rank correlation |
| Visualisation | Scatter plot of ARRIVAL\_DELAY vs DISTANCE with smooth trend line. |

1. **Spearman rank correlation**

A computer screen shot of a red and black text

AI-generated content may be incorrect.

**Output:**

A white background with black text

AI-generated content may be incorrect.

1. **Scatter plot of arrival delay vs flight distance**

A computer screen shot of a red and black text

AI-generated content may be incorrect.

**Output:**

A graph of flight distance

AI-generated content may be incorrect.

Spearman’s rank correlation was used to analyse the relationship between the flight distance and arrival delay. The results show a weak negative correlation and suggesting longer flights tend to have slightly shorter the delays, but the effect is very low. The scatter plot shows that flight distance has limited influence on delay duration.

### 3.3.4 Analysis 3-4: Evaluating the Predictive Influence of Departure Delay on Arrival Delay through Linear Regression Modelling

|  |  |
| --- | --- |
| Type | Predictive Analysis |
| Independent Variable | DEPARTURE\_DELAY (continuous) |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Simple Linear Regression |
| Visualisation | Scatter plot of ARRIVAL\_DELAY vs DEPARTURE\_DELAY with regression line. |

1. **Simple linear regression: Arrival Delay ~ Departure Delay**

A computer code with text

AI-generated content may be incorrect.

**Output:**

A screenshot of a computer error

AI-generated content may be incorrect.

1. **Visualisation**

**Scatter plot with regression line**

A computer code with text

AI-generated content may be incorrect.

**Output:**

A graph with a line and dots

AI-generated content may be incorrect.

This section uses simple linear regression model to evaluate how departure delay affects arrival delay. Using the lm() function, the model can estimate how much arrival delay increases for every additional minute of departure delay. The results show a strong and statistically significant positive relationship, meaning late departures almost always lead to late arrivals. The scatter plot with the fitted regression line visually confirms this trend, demonstrating that departure delay is a major predictor of arrival performance and important factor for improving operational punctuality.

### 3.3.5 Extra Features

* Additional variable creation: Custom grouping of time\_bin (based on scheduled departure hour) and distance\_group (based on flight distance).
* Interaction analysis: Combined Time of Day × Flight Distance using Two-Way ANOVA to assess interaction effects on arrival delays.

### 3.3.6 Conclusion

Overall, the analysis illustrates how time and operational factor that affect flight arrival delays. The outcomes indicate that afternoon and evening flights are significantly delayed, and departure delays are significantly correlated with delayed arrivals. Flight distance then further conditions these effects, such that medium and long-haul flights are particularly responsive to time-of-day effects. These outcomes indicate that flight delays are not due to one cause alone but instead occur due to the interaction between multiple flight schedule and operational conditions. Overall, this analysis shows the need for proactive delay management and strategic of flight scheduling, offering useful insights towards airline on-time performance improvement and flight operation optimization.

### 3.3.7 Recommendation:

Airline should be aimed at lessening departure delays to keep overall arrival delays short. Smooth ground operations, early intervention for flights at risk, and improved coordination during peak periods are some of the ways whereby delay escalation may be avoided. Short-haul flight rescheduling during peak periods may also enhance on-time performance, given that long-haul flights recover lost hours during flight.

## 3.4 Objective 4: To analyse the propagation mechanism of flight delays, with a focus on the relationship between Late Aircraft Delay and Arrival Delay and find out the key drivers that amplify delay propagation as well as potential intervention points for mitigation. – Yap Li Shan (TP080968)

### 3.4.1 Analysis 4-1: Identify the primary attributable factors to flight delays and analysing the localized delay risk profile across the Top 15 busiest origin airports.

|  |  |
| --- | --- |
| Type | Descriptive Analysis |
| Independent Variable | LATE\_AIRCRAFT\_DELAY (Primary independent variable)  AIR\_SYSTEM\_DELAY, SECURITY\_DELAY, AIRLINE\_DELAY, WEATHER\_DELAY  ORIGIN\_AIRPORT |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Summary Statistics, Comparative Descriptive Analysis, |
| Visualisation | Horizontal Bar Chart, Stacked Bar Chart, Heatmap |

1. Horizontal Bar Chart: Total Delay Contribution by Attributable Factor

文本

AI 生成的内容可能不正确。

Figure 3.4.1.1: Code of Total Delay Contribution by Attributable Factor

Output:

图表, 条形图

AI 生成的内容可能不正确。

Figure 3.4.1.2: Total Delay Contribution by Attributable Factor

In this analysis, the contribution of various attributable factors to flight arrival delays was depicted in this analysis using a horizontal bar chart. According to the findings, the primary reason for delays is LATE\_AIRCRAFT\_DELAY, which accounts for 404,675 minutes overall, or 40% of all delay minutes. This demonstrates the substantial domino effect brought on by earlier flights' delayed arrival. The second largest factor is AIRLINE\_DELAY, which accounts for 322,069 minutes overall, or 31.8% of all delay minutes. AIRLINE\_DELAY is mostly caused by operational problems unique to the airline, like inefficient ground handling, crew scheduling, or maintenance. With 232,831 minutes, or 23% of all delay minutes, AIR\_SYSTEM\_DELAY comes in third place and reflects the effects of traffic management restrictions and airspace congestion. In contrast, WEATHER\_DELAY and SECURITY\_DELAY play relatively minor roles in various attributable factors, accounting for 5.1% and 0.1%.

1. Stacked Bar Chart: Delay Contribution Structure for Top 15 Busiest Origin Airports

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

文本

AI 生成的内容可能不正确。

Figure 3.4.1.3: Code of Delay Contribution Structure for Top 15 Busiest Airports

Output:

图表

AI 生成的内容可能不正确。

Figure 3.4.1.4: Delay Contribution Structure for Top 15 Busiest Airports

In this analysis, we observe clear structural variations in delay attribution across airports. LATE\_AIRCRAFT\_DELAY and AIRLINE\_DELAY dominate overall, with ripple effects especially strong at SFO and LAX (over 45%), while airline-related scheduling issues are most critical at ATL and PHX (≈38%). AIR\_SYSTEM\_DELAY are more prominent at BOS and LGA (≈30%), reflecting heavy traffic control in these regions. WEATHER\_DELAY though generally minor, still reach 6–10% at ORD, DFW, and ATL, pointing to regional weather impacts. In contrast, SECURITY\_DELAY remain negligible (<1%) across all airports.

1. Heatmap: Delay Attribution Heatmap for Top 15 Busiest Origin Airports

文本

AI 生成的内容可能不正确。

Figure 3.4.1.5: Code of Delay Attribution Heatmap for Top 15 Origin Airports

Output:

表格

AI 生成的内容可能不正确。

Figure 3.4.1.6: Delay Attribution Heatmap for Top 15 Origin Airports

The heatmap is offers a more intuitive visual comparison compared to the previous chart. Firstly, it reveals that SFO has the highest proportion of LATE\_AIRCRAFT\_DELAY at 50.2%. Indicating its sensitivity to delay propagation is greater than other airports.

Secondly, BOS exhibits an AIR\_SYSTEM\_DELAY rate of 30.3%, markedly higher in all airports. This carries out the operational pressures stemming from congested airspace in the region.

Moreover, the heatmap reveals regional variations in WEATHER\_DELAY. For instance, ORD (10.4%), DFW (8.6%), and ATL (8.5%) exhibit pronounced WEATHER\_DELAY, whereas airports like LAX (1.5%) and DEN (2.2%) remain largely unaffected. Compared to stacked bar charts, heatmaps help to identify which airports are most vulnerable to specific delay types.

### 3.4.2 Analysis 4-2: To diagnose how ground operational efficiency (measured by TAXI\_OUT) moderates the relationship between LATE\_AIRCRAFT\_DELAY and ARRIVAL\_DELAY.

|  |  |
| --- | --- |
| Type | Diagnostic Analysis |
| Independent Variable | LATE\_AIRCRAFT\_DELAY (Primary independent variable)  AIR\_SYSTEM\_DELAY  SECURITY\_DELAY  AIRLINE\_DELAY  WEATHER\_DELAY  ORIGIN\_AIRPORT  TAXI\_OUT (Moderator variable) |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Multiple Linear Regression with Interaction Term, Model Comparison |
| Visualisation | Interaction Plot: |

1. Regression model comparison
2. Model\_1: Validate the structural relationship between the five delay attribution categories and ARRIVAL\_DELAY

图形用户界面, 文本, 应用程序, 电子邮件

AI 生成的内容可能不正确。

Figure 3.4.2.1: Code of Model\_1

Output:

文本, 表格

AI 生成的内容可能不正确。

Figure 3.4.2.2: Delay attribution contributes to ARRIVAL\_DELAY

The regression analysis results for Model 1 demonstrate an extremely high explanatory power (R2 =0.9945), confirming that the five attributed delay factors almost entirely account for ARRIVAL\_DELAY. The coefficients of all predictors are highly significant (p<0.001) and are very close to 1.0. This outcome validates the structural design of the dataset, where attributed delays are the defining components of the arrival delay measure. Although we already confirming the statistical necessity of these factors, Model 1 lacks diagnostic value regarding operational efficiency. Thus, Model 2 will focus on identifying the specific operational factors that moderate how this delay is propagated.

1. Model\_2: Diagnose whether TAXI\_OUT can moderate the impact of LATE\_AIRCRAFT\_DELAY on arrive delay

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

Figure 3.4.2.3: Code of Model\_2

Output:

文本

AI 生成的内容可能不正确。

Figure 3.4.2.4: Summary of Model\_2

In model 2, we observe that (LATE\_AIRCRAFT\_DELAY\_c:TAXI\_OUT\_c) exhibits high statistical significance (p=0.000603<0.001). This indicates that TAXI\_OUT exerts a significant moderating effect on LATE\_AIRCRAFT\_DELAY.

The coefficient for the interaction term in model 2 is positive (1.915e-03), indicating that if TAXI\_OUT time increases, the negative impact of LATE\_AIRCRAFT\_DELAY on ARRIVAL\_DELAY is amplified. In other words, when inbound flights are already delayed, long TAXI\_TIME will exacerbate the propagation of delays. This finding operationally explains why LATE\_AIRCRAFT\_DELAY proves difficult to manage and establishes optimising TAXI\_OUT as a key intervention lever for mitigating delay propagation risks.

1. Visualisation
2. Interaction Plot: model\_2

图形用户界面, 文本

AI 生成的内容可能不正确。

Figure 3.4.2.5: Code of Interaction Plot

Output:

图表, 散点图

AI 生成的内容可能不正确。

Figure 3.4.2.1: Moderating Effect of TAXI\_OUT on Delay Propagation

Figure 4.2.3 visualises the significant interaction term (p=0.000603) in Model 2, providing strong evidence that TAXI\_OUT exerts a significant amplifying regulatory effect on LATE\_AIRCRAFT\_DELAY. The steepest slope of the dark blue solid line, representing low efficiency (‘High TAXI\_OUT’), indicates that during the longest ground waiting times (lowest operational efficiency), each minute of LATE\_AIRCRAFT\_DELAY exerts the strongest intensity effect on the final ARRIVAL\_DELAY. Conversely, the light blue dotted line representing high efficiency (‘Low TAXI\_OUT’) exhibits the gentlest slope, demonstrating that efficient operations can more effectively absorb delays.

### 3.4.3 Analysis 4-3: To develop a pre-departure predictive model by focused on pre-flight observable factors (ORIGIN\_AIRPORT, AIRLINE, MONTH) to identify inherent flight delay risks before take-off for preventive intervention.

|  |  |
| --- | --- |
| Type | Predictive Analysis |
| Independent Variable | ORIGIN\_AIRPORT, DESTINATION\_AIRPORT, AIRLINE  MONTH, DAY\_OF\_WEEK |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Random Forest Regression Model, Performance Evaluation using RMSE and R², Variable Importance Analysis |
| Visualisation | Scatter plot, %IncMSE plot |

1. Random Forest

文本

AI 生成的内容可能不正确。

Figure 3.4.3.1: Code of Randon Forest

Output:

图片包含 文本

AI 生成的内容可能不正确。

Figure 3.4.3.2: Output of Randon Forest

When only limited to pre-departure observable factors, the random forest model's predictive power was extremely low, the result of RMSE is 52 minutes and R2 is just 0.02. This outcome highlighted the inherent difficulty of predicting precise arrival delay durations using only schedule and location-based information. It validates that most delays are driven by unpredictable and dynamic factors such as weather, air traffic control limitations, and late aircraft delay at previous.

1. Visualisation
2. Scatter plot: Prediction arrival delay vs Actual arrival delay Graph

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

Figure 3.4.3.3: Code of Scatter Plot

Output:

图表, 折线图

AI 生成的内容可能不正确。

Figure 3.4.3.4: Plot of Random Forest Model Performance

1. %IncMSE plot: Variable Importance in Predicting Arrival Delay

文本

AI 生成的内容可能不正确。

文本

AI 生成的内容可能不正确。

Figure 3.4.3.5: Code of %IncMSE plot

Output:

图表, 漏斗图

AI 生成的内容可能不正确。

Figure 3.4.3.6: Variable Importance for Predicting Arrival Delays

Although the random forest model exhibited limited predictive accuracy, the variable importance ranking chart (Figure 4.3.2) quantifies the actual contribution of all variables to delay prediction within the nonlinear model. The findings show that ORIGIN\_AIRPORT was the most influential pre-departure factor, followed by AIRLINE and MONTH. This suggests that structural characteristics of origin airports such as congestion patterns, ground handling capacity, and scheduling efficiency are play a relatively stronger role in flight delay management. Thus, improving operational efficiency at origin airports may yield significant benefits in reducing arrival delays.

### 3.4.4 Analysis 4-4: Quantify each airport's average total contribution in minutes across the two core risk factors and identify the top 15 origin airports with the highest minute.

|  |  |
| --- | --- |
| Type | Prescriptive Analysis |
| Independent Variable | AIRLINE\_DELAY  LATE\_AIRCRAFT\_DELAY  ORIGIN\_AIRPORT |
| Dependent Variable | ARRIVAL\_DELAY |
| Analysis Technique Used | Descriptive Aggregation and Ranking Analysis, Comparative Evaluation |
| Visualisation | Lollipop Chart |

1. Visualisation
2. Lollipop Chart: Origin Airports should be notices to improve (Have Highest Average Core Delay Minutes)

文本, 应用程序

AI 生成的内容可能不正确。

Figure 3.4.4.1: Code of Lollipop Chart: Origin Airports should be notices to improve

Output:

图表, 散点图

AI 生成的内容可能不正确。

Figure 3.4.4.2: Top 15 Origin Airports with the Highest Average Core Risk Delays in Minutes

As observed in Figure 4.4.1, Newark Liberty International Airport (EWR) recorded the highest average core delay minutes at 30 minutes. This highlights its facing most serious situation in flight chain effects and operational management. Following closely are Chicago Midway International Airport (MDW, 29 minutes), LaGuardia Airport (LGA, 29 minutes), Orlando International Airport (MCO, 27 minutes) and John F. Kennedy International Airport (JFK, 27 minutes).

These findings provide clear prioritisation for improving airport-level operational efficiency, such as enhancing gate turnover processes, ground handling coordination, and air traffic flow management. By targeting these high-delay airports can achieve the most significant reductions in propagated arrival delays.

### 3.4.5 Extra Features

- TAXI\_OUT: Used as a moderator variable in Analysis 4-2 to assess ground operational efficiency and examine how longer taxi-out times amplify delay propagation.

- ORIGIN\_AIRPORT and DESTINATION\_AIRPORT: Include in Analyses 4-1, 4-3, and 4-4 to capture airport-level heterogeneity. These variables make it possible to investigate regional delay risk profiles, patterns of airport traffic, and structural variations at the route level that impact the spread of delays.

- DAY\_OF\_WEEK and MONTH: Include in Analysis 4-3 to capture seasonal or weekday/weekend variations and identify high-delay periods for preventive planning. These variables help in identifying cyclical fluctuations that could support proactive resource planning and pre-departure risk identification.

- WEATHER\_DELAY, AIR\_SYSTEM\_DELAY, SECURITY\_DELAY and AIRLINE\_DELAY: Include in Analyses 4-1 and 4-2 to ensure that the model captures all major attributable factors influencing total arrival delay composition.

### 3.4.6 Conclusion

This analysis demonstrates that the main cause of subsequent ARRIVAL\_DELAY is delays from earlier flights, highlighting a significant delay propagation relationship between ARRIVAL\_DELAY and LATE\_AIRCRAFT\_DELAY. Among all variable, LATE\_AIRCRAFT\_DELAY is the most significant contributor, followed by AIRLINE\_DELAY and AIR\_SYSTEM\_DELAY, while WEATHER\_DELAY and SECURITY\_DELAY have comparatively minor impacts. The moderating analysis indicates that longer TAXI\_OUT times will amplify delay propagation effects. Although pre-departure factors such as ORIGIN\_AIRPORT, AIRLINE, and MONTH exhibit limited predictive power, they still help identify high-risk airports. Overall, this objective demonstrates that improving operational efficiency and reducing LATE\_AIRCRAFT\_DELAY are key to mitigating cascading arrival delays.

### 3.4.7 Recommendation

- Enhance turnaround efficiency: Streamline ground operations such as gate assignment, refuelling, and baggage handling to reduce LATE\_AIRCRAFT\_DELAY accumulation.

- Optimize taxi-out management: Implement dynamic taxi routing and real-time coordination with air traffic control to reduce TAXI\_OUT duration.

- Prioritize high-risk airports: Focus operational improvement programs at key delay hubs such as EWR, MDW, and LGA, where average core delay minutes are highest.

# 4.0 Conclusion

## 4.1 Overall Discussion

In this assignment, we use R programming to carry out statistical, diagnostic, and predictive analytics to analyze the major factors of flight delays and airline performance of "Flight Delay Prediction and Airline Reliability Assessment,". For all goals, the results always indicated that the departure delay was the leading cause of arrival delay, with a significant positive association. Multiple regression analysis also confirmed that for every one-minute increase in departure delay, arrival delays also increased by approximately one minute, highlighting the key role of time in determining on-time performance. However, the impact of flight distance and flight time, while minor, is tangible; for example, longer flights can partially offset time losses. Time-of-day and airline identity analyses uncovered patterns of operation, afternoon and evening flights experience longer delays, whereas some carriers like Spirit Airlines (NK) and Virgin America (VX) have greater variability and chaining risks. Furthermore, late aircraft delay was identified as the most influential propagation factor accounting for around 40% of total delay minutes. Interaction analysis demonstrated that inefficient ground taxi out time magnifies the delay propagation and emphasizing the operational importance of turnaround efficiency.

Overall, the integrated analysis provides a comprehensive understanding of how multiple factors like temporal, operational, and airline-based, interact to affect flight reliability. These insights provide a solid basis for strategic delay mitigation and predictive scheduling in the aviation industry.

## 4.2 Recommendation

To improve overall on time performance, airlines operators should prioritize the preparations before departure and intervene on the high-risk flights that often directly lead to arrival delays. Improving ground operations by reducing the taxi times and increasing boarding gate turnaround efficiency can also help minimize the aircraft delays. Furthermore, by using data-driven flight scheduling through predictive analytics, airlines can predict prone to delay periods, such as late afternoons or high-demand months like February and adjust flight schedules or crew rotations accordingly. The high-risk airports with consistently high average core delay minutes, such as EWR, MDW, and LGA, we can have some investments to upgrade the infrastructure and streamline operations. Another way, seasonal resource planning can be implemented to ensure improvements, allocating more operational resources during months with higher delay rates while maintaining efficiency levels during periods with lower delays.

## 4.3 Limitation and Future Direction

This assignment has several limitations, primarily due to the static nature of the dataset, which lacks real-time variables such as weather conditions, air traffic congestion, and crew changes that could cause flight delays. The predictive models, especially the random forest using only pre-departure features, demonstrated low accuracy, indicating that static and limited variables are insufficient to capture the complexity of delay dynamics. Moreover, the dataset covers only a short time period and excludes external environmental or seasonal factors, restricting the generalizability of the findings. For future work, incorporating dynamic and contextual data, such as live weather updates, air traffic information and aircraft maintenance records that could improve model performance and practical applicability. Expanding the dataset to multiple years and applying advanced machine learning models like Gradient Boosting or Deep Learning would also enhance the predictive accuracy and having a better understanding of non-linear interactions. Additionally, integrating external data sources and developing real-time decision-support systems also can help airlines to predict, prevent and manage flight more efficiently.

# 5.0 References

Kaggle Datasets (Feb 2017). *2015 Flight Delays and Cancellations.*

<https://www.kaggle.com/datasets/uslbdot/flight-delays>

FabienDaniel (Sep 2017). *Predicting flight delays [Tutorial].*

<https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial>

‌

# 6.0 Workload Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NAME  PART | CHAN MIN HUEY  **TP083261** | KANG HONG QIAN  **TP081205** | GONG YEE CHENG **TP081910** | YAP LI SHAN  **TP080968** |
| Introduction | ✔ |  |  |  |
| Data Preparation |  | ✔ |  | ✔ |
| Analysis | | | | |
| Objective 1 | ✔ |  |  |  |
| Objective 2 |  | ✔ |  |  |
| Objective 3 |  |  | ✔ |  |
| Objective 4 |  |  |  | ✔ |
|  | | | | |
| Conclusion |  |  | ✔ |  |
| Signature | 形状  描述已自动生成 |  |  | 图片包含 游戏机, 物体  描述已自动生成 |
| Workload % | 25% | 25% | 25% | 25% |

Total Word Count: 7838 words