# ST2195 PROGRAMMING IN DATA SCIENCE

Carine Chua Wentian

220460376

Singapore Institute of Management

**PART 2 REPORT**

# Table of Contents

## Introduction

This report examines ten consecutive years of flight departure and arrival details, from 1995 to 2004, using datasets from Harvard Dataverse. This report seeks to address fundamental questions regarding flight delays, the impact of aircraft age on delays, and the probability of flight diversions. By leveraging statistical principles and advanced analytical tools, we aim to unravel underlying patterns, trends, and correlations within aviation data.

## Data Import and Data Frame Setup

To initiate our analysis, we imported CSV files containing airports, carriers, plane data, and yearly flight data spanning from 1995 to 2004. These datasets were organized into separate data frames: airports, carriers, plane data, and individual yearly datasets. Using the ‘planes’ identifier, we consolidated the yearly datasets into a comprehensive data frame spanning the entire decade. Within the resulting ‘planes’ data frame, we calculated a new column named “Delays” representing the total delays for each flight, derived from the sum of departure delay (“DepDelay”) and arrival delay (“ArrDelay”) for each respective flight record.

## Question (a): What are the best times and days to minimise delays each year

### a(i) – What are the best times to minimise delays each year?

In our analysis to identify optimal times to minimise flight delays, we structured flight data into meaningful categories using time binning and aggregation techniques. To ensure data quality, rows with invalid or missing time entries were removed.

We created two data frames, ‘DayTime_Dep’ and ‘DayTime_Arr’ to summarise departure and arrival flights along with delays by time of day. These were merged into a 'DayTime_df' data frame, which categorizes flights and delays into distinct time periods: ‘Morning’ (0600 to 1200), ‘Afternoon’ (1200 to 1800), ‘Evening’ (1800 to 0000), and ‘Night’ (0000 to 0600). Column headers were then renamed for clarity.

Subsequently, we computed a "Percentage Delayed" column and added it to the merged data frame. The results were visualised using a bar graph to depict the percentage of delayed flights for each time of day, effectively highlighting variation in delay rates throughout the day.

From the graph, the 'Evening' period exhibited the highest percentage of delayed flights at 52.43%, while 'Morning' had the lowest delay rate at 33.47%. Based on this analysis, we can conclude that the 'Morning' period offers the most favourable conditions for minimizing delays, followed by the 'Afternoon'.

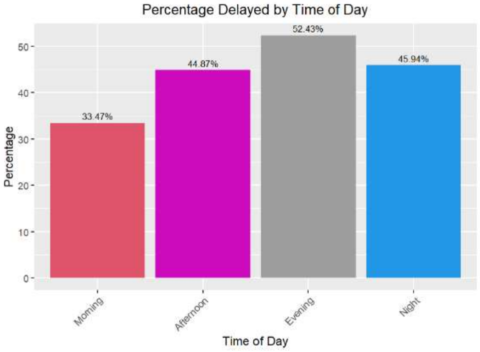


*Figure 1: R Markdown – Percentage delayed by time of day*

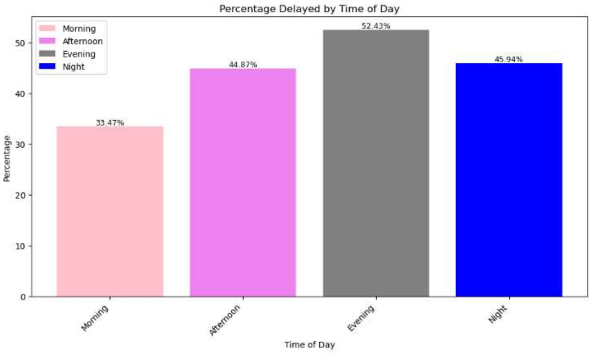
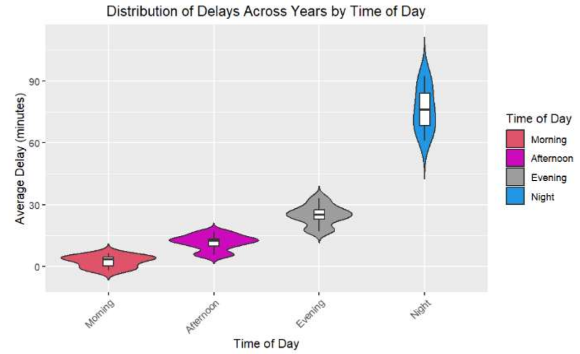


*Figure 2: Python – Percentage delayed by time of day*

We extended our analysis by creating separate data frames, ‘DelaybyDayTime_Dep’ for departure delays and ‘DelaybyDayTime_Arr’ for arrival delays, grouped by year and time of day (‘Morning’, ‘Afternoon’, ‘Evening’, ‘Night’). These data frames were then merged into ‘DayTimeDelay’ to facilitate comparative analysis. To ensure data quality, rows with invalid or missing values were removed, and column headers were renamed for readability.

A violin plot (‘DayTimeViolin’) was generated to visualise delay distributions across different times of day, providing insights in the exploration of delay patterns and trends over the years. Each violin shape on the plot represents the density distribution of delays during specific periods.

The violin plot highlights variations in delay frequency and severity across different times of day, with ‘Night’ showing a propensity for extreme delays and ‘Morning’ exhibiting the least variability in delay times. This analysis identified 'Morning' and 'Afternoon' as periods with the lowest average delays.



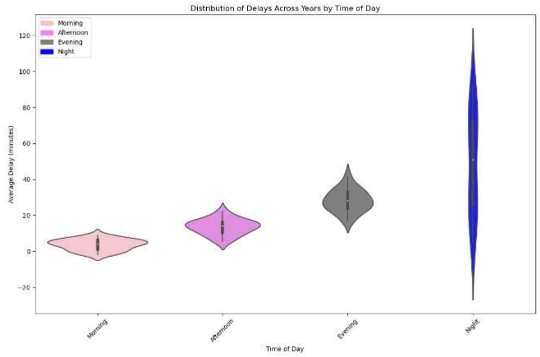


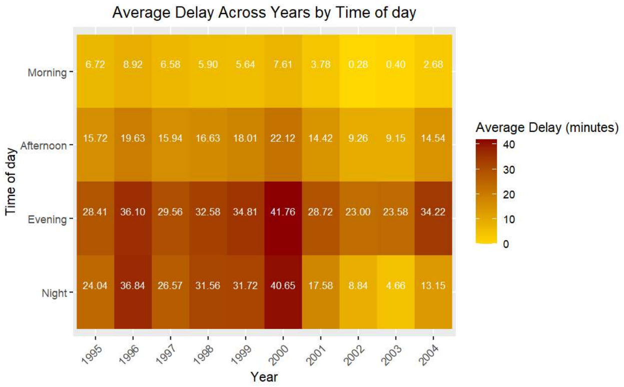
*Figure 3: R Markdown – Distribution of delays across years by time of day*

*Figure 4: Python – Distribution of delays across years by time of day*

We utilized the arrival and departure delay data frames to merge both datasets, resulting in the creation of a new ‘DayTimeDelay2’ data frame. This combined data frame was then leveraged to construct a pivot table (‘DayTimeDelay_pivot’) for comprehensive data analysis. Subsequently, ‘DayTimeDelay_pivot’ was utilised to generate a heatmap (‘HeatmapDayTime’) to examine average delays across years by time of day.

The heatmap employs a colour gradient from "gold to dark red" in R Markdown and "yellow to red" in Python, enhancing readability and visualisation of delay trends.

Key findings from the heatmap analysis reveal consistent patterns: Evening periods consistently exhibit high average delays, with the peak delay of 41.76 minutes recorded in 2000. Conversely, Morning periods consistently maintain low average delays, with the highest delay observed in 1996 at 8.92 minutes.



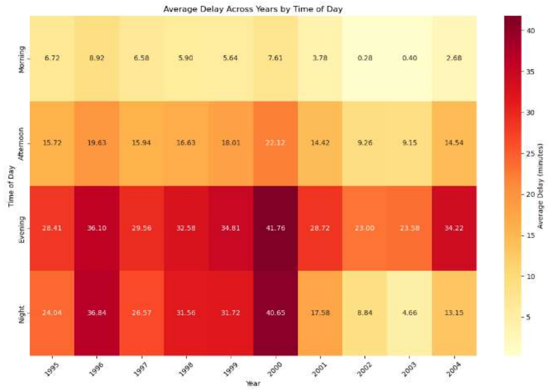


*Figure 5: R Markdown – Average delay across years by time of day*

*Figure 6: Python – Average delay across years by time of day*

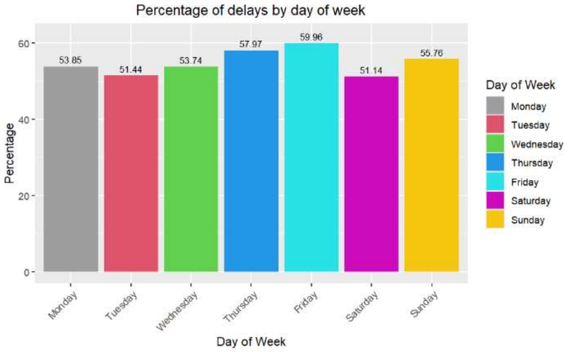
In conclusion, our analysis, incorporating bar graphs, violin plots, and heatmaps, indicates that 'Morning' (6 a.m. to 12 p.m.) consistently emerges as the optimal time to minimize delays each year, followed by 'Afternoon' (12 p.m. to 6 p.m.). To mitigate delays, it is advisable to avoid scheduling flights during 'Evening' (6 p.m. to 12 a.m.) and 'Night' (12 a.m. to 6 a.m.) periods. These insights provide actionable recommendations for scheduling flights to optimize on-time performance and minimise delays.

### a(ii) – What are the best days to minimise delays each year?

In our detailed analysis focused on identifying optimal days to minimise flight delays, we categorized flight data using data mapping and aggregation techniques, ensuring data quality by removing invalid or missing entries.

First, we created a ‘DayWeekTotal’ data frame outlining the flights and delays categorised into distinct day of week: ‘Monday’, ‘Tuesday’, ‘Wednesday’, ‘Thursday’, ‘Friday’, ‘Saturday’, ‘Sunday’. Column headers were renamed for clarity. We then computed “Percentage of flights” and “Percentage Delayed” within the merged data frame and visualized the results using a bar graph, depicting the percentage of delayed flights for each day of the week. This visualisation effectively highlights the variation in delay rates across different days of the week.

The analysis reveals that ‘Friday’ consistently experiences the highest percentage of delayed flights at 59.96%, while ‘Saturday’ shows the lowest delay percentage at 51.14%. Based on this analysis, we can conclude ‘Saturday’ would be the most favourable day to minimise delays, followed by ‘Tuesday’ which boast a 51.44% of delayed flights.



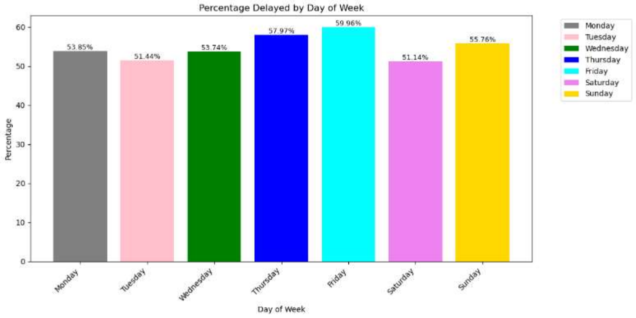


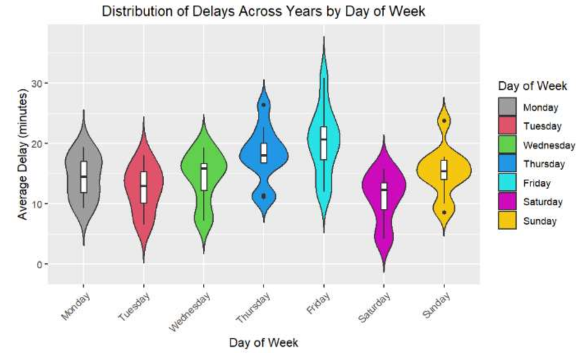
*Figure 7: R Markdown – Percentage Delayed by day of week*

*Figure 8: Python – Percentage Delayed by day of week*

We expanded our analysis by constructing a data frame, ‘DelaybyDayWeek’, which aggregates average delays for each year and day of week. The data frame served as the basis for a pivot table (‘DelaybyDayWeek_pivot’) used to facilitate deeper analysis. Subsequently, ‘DayTimeDelay_pivot’ was leveraged to generate a violin plot (‘DayWeekViolin’) to examine average delays across years by day of week.

To ensure data quality, rows with invalid or missing data were excluded prior to analysis. The violin plot provides a clear representation of delay distributions across different weekdays, offering insights into delay patterns and trends over the years. Each violin shape on the plot represents the density distribution of delays during specific periods.

Our findings reveal notable variations in delay frequency and severity across weekdays. Specifically, ‘Friday’ exhibited a propensity for extreme delays, with the highest average delay recorded at approximately 37 minutes. In contrast, ‘Saturday’ shows the least variability in delay times, boasting the lowest average delay of approximately 20 minutes, with a range widening around 10 to 15 minutes. This analysis identified ‘Saturday’ as the period with the lowest average delays.



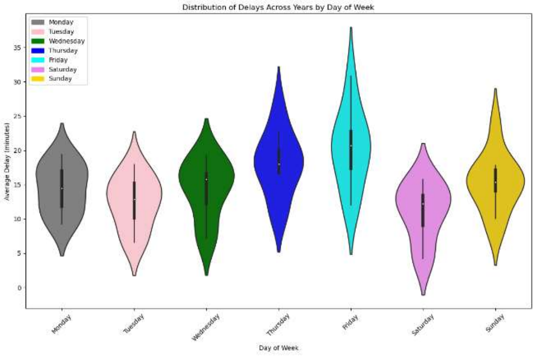
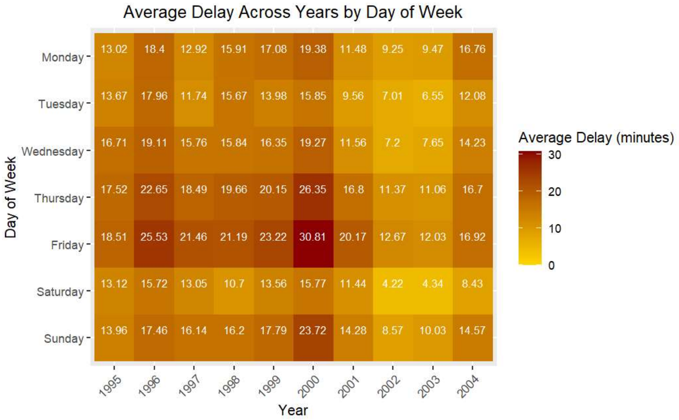


*Figure 9: R Markdown – Distribution of delays across years by day of week*

*Figure 10: Python – Distribution of delays across years by day of week*

A second pivot table (‘pivot_DelaybyDayWeek2’) was derived from the existing ‘DelaybyDayWeek’ data frame. This pivot table was used to generate a heatmap ('HeatmapDayWeek') to examine average delays across years by days of week. The heatmap employs a colour gradient from "gold to dark red" in R Markdown and "yellow to red" in Python, enhancing readability and visualisation of delay trends.

The heatmap provides a clear illustration of average delays across years by days of the week. Notably, in 2000, 'Friday' experienced the most significant delay, averaging 30.81 minutes. Conversely, 'Saturday' consistently exhibited the lowest delays, with averages of 4.22 minutes in 2002 and 4.34 minutes in 2003.



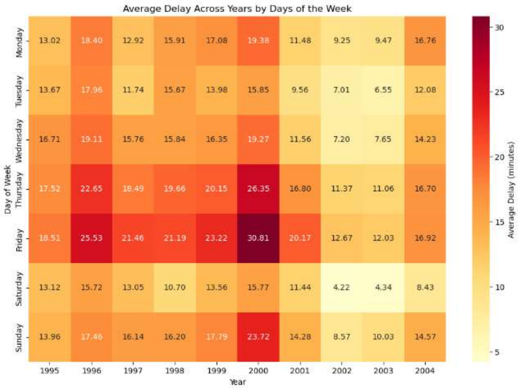


*Figure 11: R Markdown – Average delay across years by day of week*

*Figure 12: Python – Average delay across years by day of week*

In summary, our analysis, utilizing bar graphs, violin plots, and heatmaps, underscores 'Saturday' as the most favourable day for minimizing delays each year. Scheduling numerous flights on 'Friday' is ill-advised due to the consistently high delay averages observed. These insights offer actionable recommendations for flight scheduling to optimize on-time performance and mitigate delays effectively.

## Question (b): Evaluate whether older planes suffer more delays on year-to-year basis

In our analysis on evaluating whether older planes suffer more delays, we prepared and cleaned the data by creating a subset (‘planedata_subset’) of necessary columns—Tail Number (“tailnum”) and manufacturing year (“year”)—from ‘plane_data’. This subset was then merged with the 'planes' dataset to form a new data frame named ‘airplanes’. To enhance clarity, we renamed the “year” column to “PlaneYear”. Rows with missing or invalid “PlaneYear” were excluded to ensure data quality.

Next, we calculated the age of each plane (“PlaneAge”) and categorised them as ‘New’ (aged 0 to 21 years) or ‘Old’ (aged 21 years and above) under “PlaneAgeCategory” column. To identify delayed flights, we introduced a Boolean column (“Delayed_Boolean”) where 1 represents "Delayed" and 0 represents "Not Delayed". After filtering out invalid data, we computed the percentage of delayed flights and plotted a bar graph (‘PercentageAirplanes’) depicting the percentage of delays among ‘New’ planes versus ‘Old’ planes.

The analysis showed that ‘New’ planes (aged 0 to 21 years) (Paramount Business Jets, n.d.) contributed to a higher proportion of delayed flights, representing 57.16% of delays among all ‘New’ planes, compared to ‘Old’ planes (aged 21 years and above), which contributed to 47.69% of delays among all ‘Old’ planes.

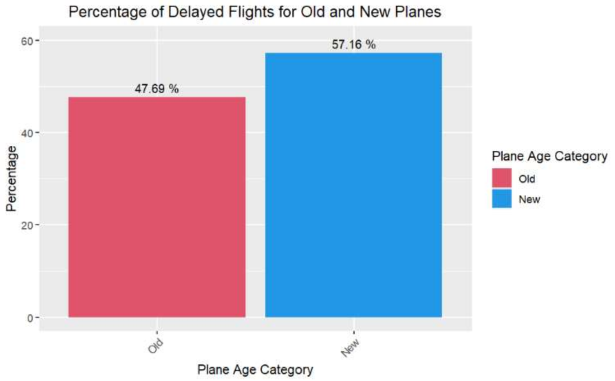


*Figure 13: R Markdown – Percentage delayed for old and new planes*

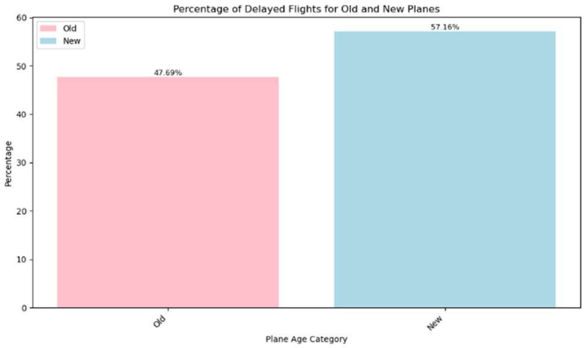


*Figure 14: Python – Percentage delayed for old and new planes*

We focused our analysis on airplane data from the years 1995 to 2004 by filtering the 'airplanes' dataset to create the 'airplanes_filtered' data frame. This refined dataset was then used to create 'DelaysPerYear', a new data frame detailing the total number of delays for each year and plane age category. From this data frame, the 'PlaneAgeDelay' line graph was produced, illustrating delay trends across years by plane age category.

The 'PlaneAgeDelay' graph provided valuable insights into delay patterns among different plane age categories. Contrary to initial expectations, newer planes ('New') generally experienced higher overall delays during this period. In 2004, 'New' planes had over two million delays, while in 2001, they had approximately 110,000 delays—the lowest during this period. In contrast, 'Old' planes exhibited their highest number of delays in 2000 (70,946 flights) and the lowest in 2004 (473 flights).

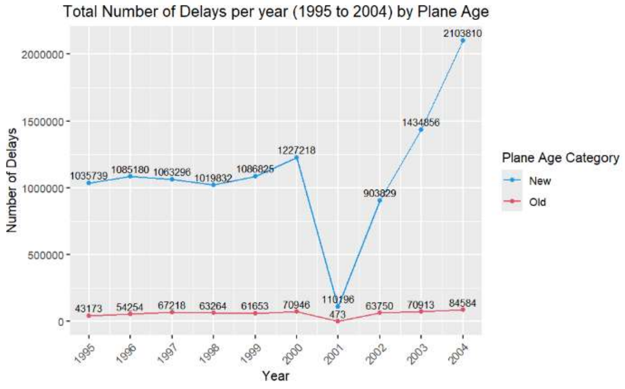


*Figure 15: R Markdown –Total number of delays for old and new planes*

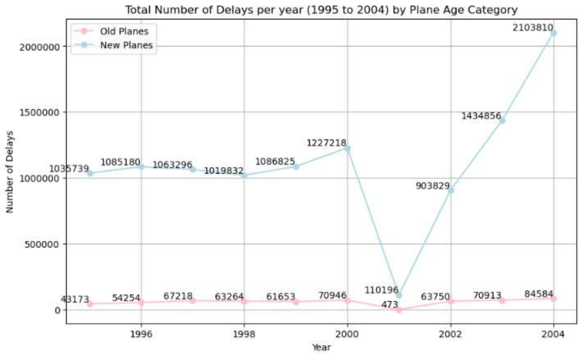


*Figure 16: Python – Total number of delays for old and new planes*

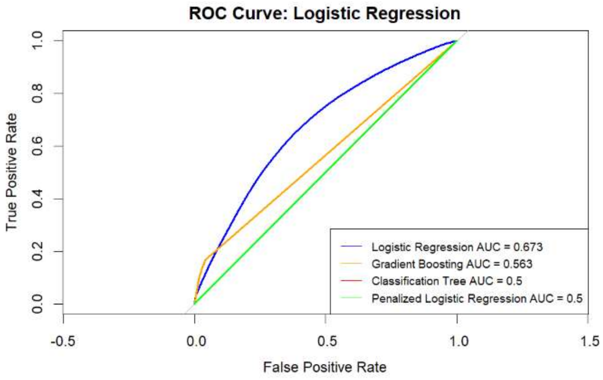
In conclusion, our analysis from 1995 to 2004 suggests that 'New' planes significantly contribute to the number of delayed flights, challenging the notion that older planes suffer more delays. However, it is important to note that older planes still make a substantial contribution (47.69% of delayed flights within the old planes category) to the total number of delays within their respective category. This nuanced understanding highlights the complex relationship between aircraft age and delays, emphasizing the need for comprehensive data analysis to draw accurate conclusions about delay patterns over time.

### Question (c): For each year, fit a logistic regression model for the probability of diverted US flights using as many features as possible from attributes of the departure date, the scheduled departure and arrival times, the coordinates and distance between departure and planned arrival airports, and the carrier. Visualize the coefficients across years.

In this analysis, we selected pertinent features (‘PlaneFeatures’) from the ‘planes’ data frame and preprocessing them through imputation and scaling for numerical features and encoding categorical features into dummy variables. Following this preprocessing step, the dataset was divided into training and testing sets.

Four classification models, logistic regression (‘PlaneFeatures_pipeLR’), gradient boosting (‘PlaneFeatures_pipegdb’), penalized logistic regression (‘PlaneFeatures_pipeplr’), and classification tree (‘PlaneFeatures_pipetree’), were trained on the training data. Probability predictions were generated for each model on the test set to compute ROC AUC values, providing a metric to evaluate the models' performance.

The resulting ROC curves illustrate the trade-off between sensitivity and specificity across decision thresholds. The logistic regression model achieved an AUC of 0.67, indicating moderate discrimination ability between diverted and non-diverted flights based on the selected features. This analysis provides valuable insights for predicting flight diversions and enhancing decision-making in aviation operations.



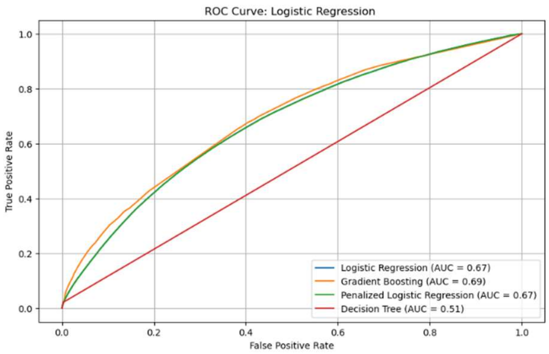
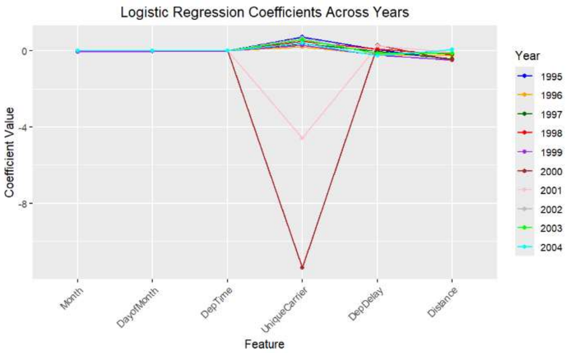


*Figure 17: R Markdown – ROC Curve: Logistic Regression*

*Figure 18: Python – ROC Curve: Logistic Regression*

We performed logistic regression analysis from 1995 to 2004 on flight characteristics (‘PlaneFeatures’) predicting flight diversion ('Diverted'). Each year had its logistic regression model fitted with preprocessing for feature scaling and encoding. Coefficients and ROC AUC scores were recorded and used to visualize temporal trends in feature importance.

From the plotted graph, it was observed that the carrier type (‘UniqueCarrier’) had the lowest coefficient, suggesting that carrier identity had a relatively minor impact on flight diversion. while flight distance consistently influenced diversion likelihood, as evidenced by consistently positive coefficient values throughout the decade.



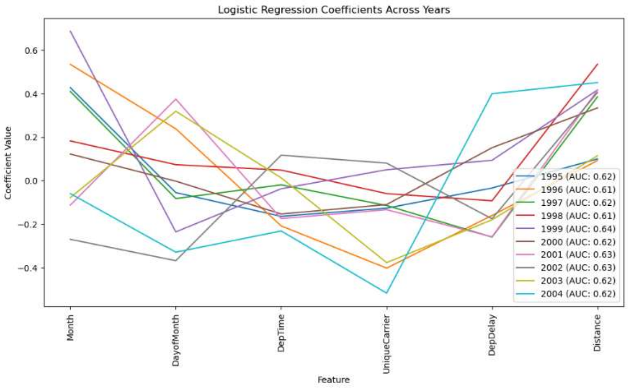


*Figure 19: R Markdown – Logistic Regression Across Years*

*Figure 20: Python – Logistic Regression Across Years*

While carrier identity seemed to have limited influence, flight distance emerged as a key determinant affecting the likelihood of flight diversion. This insight underscores the importance of considering dynamic factors over time when assessing predictive models in aviation data analysis. Further investigation into other temporal variables and model enhancements could provide deeper insights into flight diversion dynamics across different periods.

## References

Paramount Business Jets, n.d. *Is the age of an aircraft a safety factor?.* [Online]
Available at: https://www.paramountbusinessjets.com/faq/age-of-aircraft-safety-factor#:~:text=Here%20are%20the%20approximate%20ages,aircraft%20%3D%2010%20years%20or%20less
[Accessed 24 April 2024].