Table of Contents

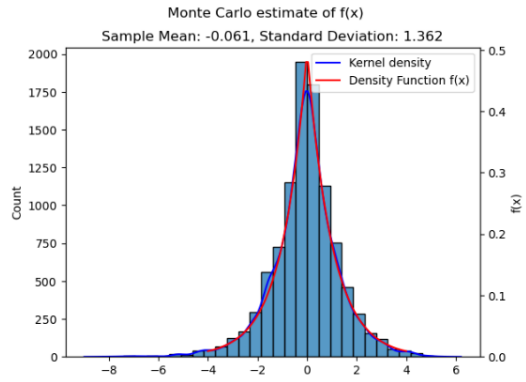
1. Question 1

Below is the visualization of the answers for part 1, in both Python and R.

Part A:

Figure 1 below shows the histogram of 10,000 randomly generated number using Monte Carlo Metropolis Random walk algorithm generated from Python and R respectively. The initial value for this algorithm I chose was 1. It is shown that the Monte Carlo estimate for the given probability density function is quite close to the actual density function itself.

Output from Python (seaborn):



Output from R (ggplot2):

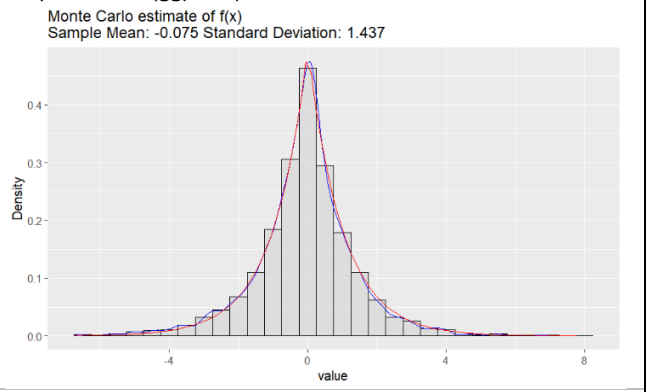
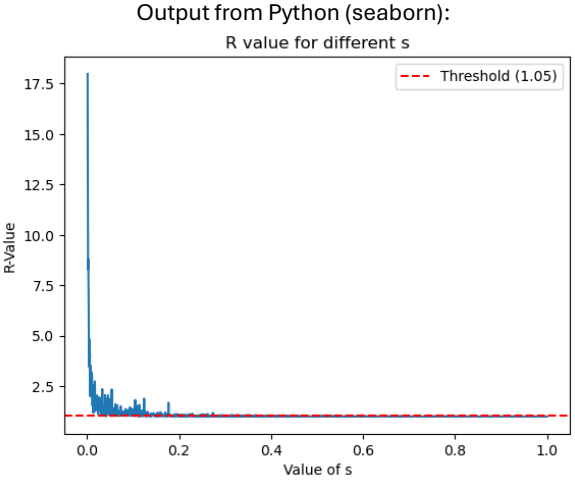


Figure 1, *Histogram and density output from Python (left) and from R (right). For both of these histograms, the red line indicates the actual density function and the blue line indicates the kernel density or the Monte Carlo estimate of the actual density function.*

Part B

Figure 2 shows the outputs for part B, which is a line plot of the computed R values for different values of Standard Deviation ranging from 0.001 to 1. Each R value was computed using 4 chains generated with the same algorithm, with four randomly generated initial values between 0 and 1, and sample size of 2,000 each. It is shown that as the Standard Deviation increases to 1, the R-Value rapidly drops to and stays at a value below 1.05, which indicates the point at which the estimate reaches a convergence. When R-value is below 1.05, this implies the variance between chains are small, meaning each chains starts to resemble each other, and the distribution of these chains becomes a good estimate of the target distribution.

Output from Python (seaborn):



Output from R (ggplot2):

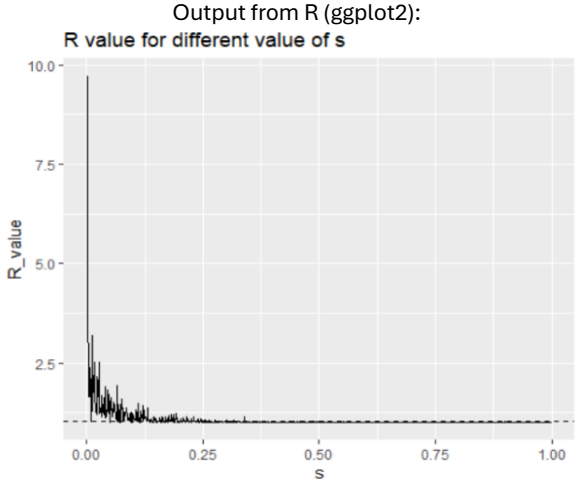


Figure 2, *Lineplot of R values of Standard Deviation ranging from 0.001 to 1 generated from Python (left) and from R (right). Dashed line indicates R value threshold of 1.05.*

2. Question 2

For question 2, only five years of consecutive flight data (2003–2007) were used instead of the requested ten years due to hardware memory limitations. Data is retrieved from the Harvard Dataverse at: https://doi.org/10.7910/DVN/HG7NV7

Part A

In part A, we are interested in finding the best days of the week and time to minimize delays of each year. Since we are only interested in flight delays for flights that took of, I have excluded cancelled flights from the analysis. There are six features related to flight delays: *'ArrDelay', 'DepDelay', 'CarrierDelay', 'WeatherDelay', 'NASDelay', 'SecurityDelay',* and *'LateAircraftDelay'*. However, there are some missing values in these features, and some negative values in '*NASDelay*', *'ArrDelay'* and '*DepDelay'*. It could be the case that the negative values meant that the flight arrived or departed earlier than scheduled, since information were not given, I assumed the missing values meant no delay (0) and negative values meant early arrival or departure, the latter case of observations was excluded for this analysis. I created a new variable, 'TotalDelay', which represents the sum of all types of delays. I then grouped the data by *'Year',* '*DayOfWeek'*, and *'CRSDepTime'*, and calculated the mean of the total delay for each group. Figure 3 presents a grid of scatterplots of the average total delay by scheduled departure time ('*CRSDepTime'*) for each year and day of the week.

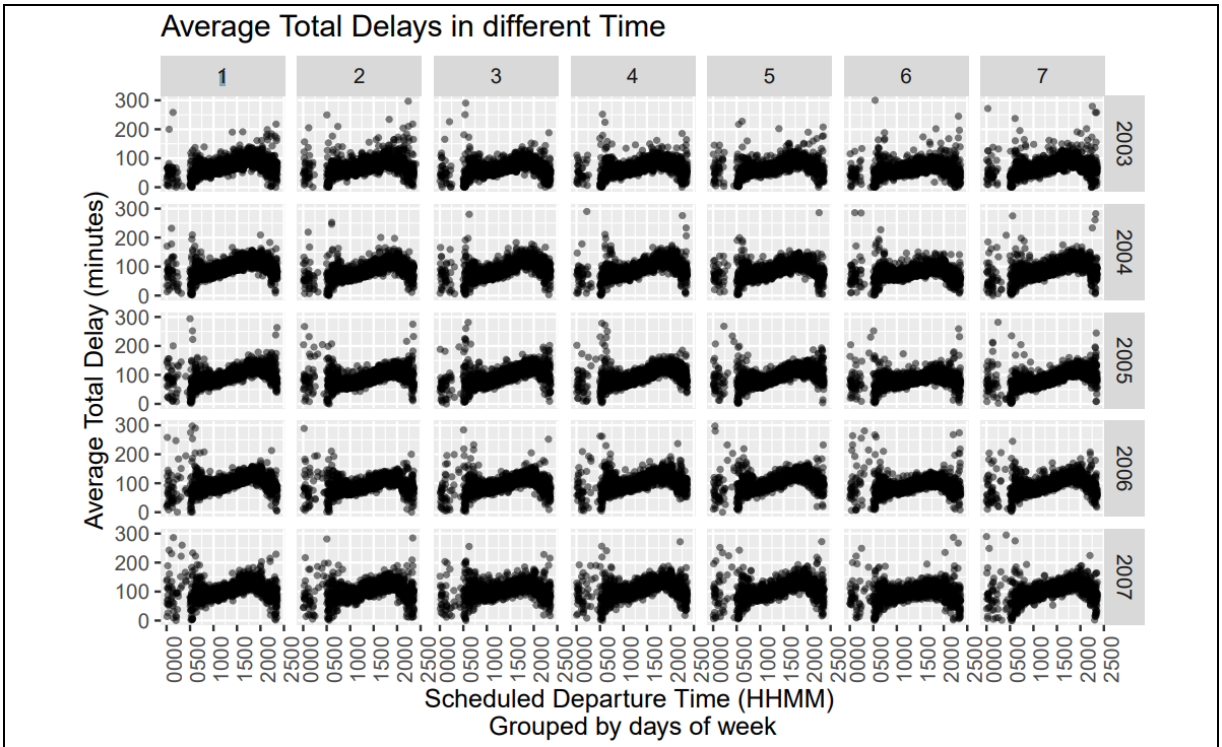


Figure 3, *Grid of scatterplots of average total delay by scheduled departure time ('CRSDepTime') for each year and day of the week. The x-axis tick labels on the top side represents day of week (1-7).*

The upside-down U-shape observed in these scatterplots suggests that there are two points in time for each day where the average total delay is minimized. Note that the Y-axis is limited to show points with average total delay up to 300 minutes only. I then proceeded to find which time results in the lowest average total delay for each day of the week in each year, and what the average total delay for the whole day is for each day of the week. To summarize my findings, I used the same plot from figure 3 and added the overall average of total delay for the whole day as an orange line, and the scheduled departure time that results in the lowest average total delay for that day as a red line, shown in figure 4.

To answer the question for this part, the red points along with the vertical red lines in each of the plots from figure 4 shows which point in the scheduled departure time for each day of the week results in the lowest average total delay. It is shown that the time between 11:00pm and 5:30am have consistently resulted in near-zero lowest average total delays for all days of the week in all years. The orange line representing the average delays in the whole day for different weekdays shows that Saturdays have the lowest average total delay out of all weekdays for years 2003 to 2007.

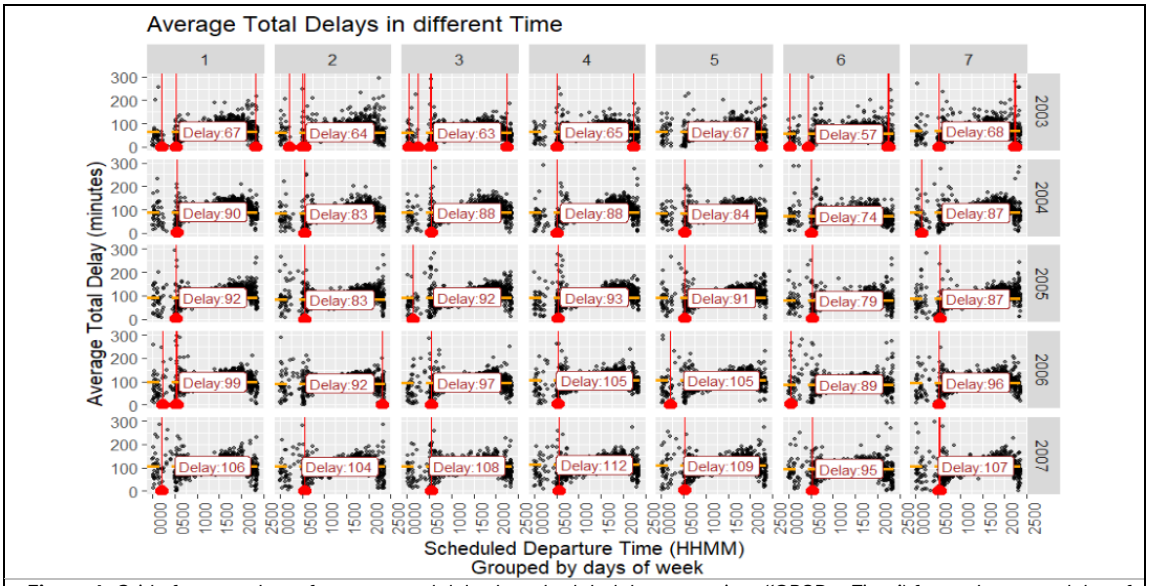


Figure 4, *Grid of scatterplots of average total delay by scheduled departure time ('CRSDepTime') for each year and day of the week. Red line represents lowest average total delay for that day, orange line represents average of the whole day. The x-axis tick labels on the top side represents day of week (1-7).*

Part B

In this part, we aim to determine if older planes experience more delays each year. To address this, I combined the planes dataset with the flight data using the tail number columns as keys. To ensure both datasets were complete, I used an inner join. I excluded 549 plane records which have missing data across all columns except '*tailnum'*, 3 planes with 'None' as the engine type, and 1 plane with an invalid status, as these were considered corrupt data. Additionally, I removed planes with *'year'* recorded as '0000' or 'None'. Given the lack of documentation, I assumed the *'year'* column indicates the plane's manufacture year and *'issue_date'* indicates its official entry into service.

Next, I calculated the plane's age at departure (*'PlaneAge'*) by subtracting the plane's manufacture year from the flight's departure year. Observations with *'PlaneAge'* below zero, indicating a future manufacture date relative to the flight departure year, were excluded. Most of the planes have aged under 10 years as of their departure year. There are not many planes over 20 years old. I then created a line plot to visualize the average total delay for each plane age from 2003 to 2007 (Figure 5). The line plot shows some upward trend in average total delays as plane age increases. Additionally, spikes that are either upward or downward in average total delay are observed at approximately every 10-year interval (ages 10, 20, 30, 40, and 50) across the years 2003-2007. To explore this further, I grouped PlaneAge into 11 categories and created annual box plots (Figure 6) to compare the distributions of total delays across different age groups. Figure 6 also includes the correlation coefficient between the variables. These coefficients stay close to zero, and the box plots show similar median values and spread across all age groups and years. This suggests that there is no strong linear relationship between the two variables. The spikes in average total delays at 10-year intervals are likely due to a few extreme delay values in specific age groups, while the median delays stay mostly flat. In summary, while plane age doesn't have a clear linear connection to delays, extreme delays in older planes (e.g., 10, 20, or 30 years old) can occasionally push up the average, even though the overall pattern remains stable.

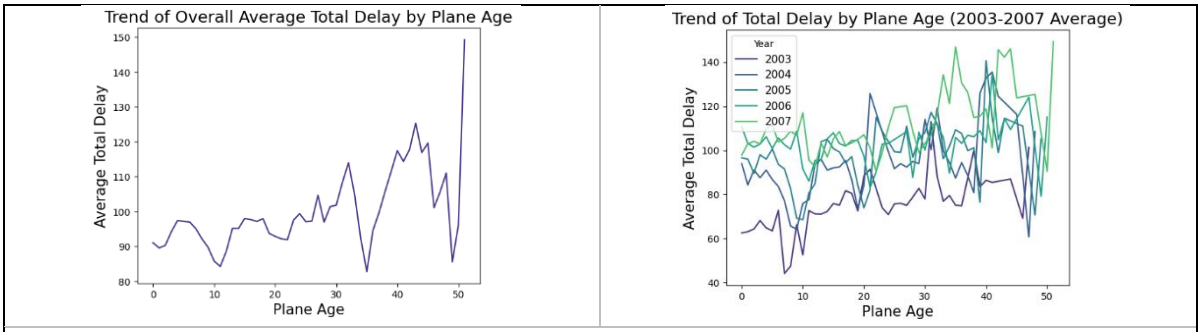


Figure 5, Left: *Average total delay over all of 2003-2007 for each plane age.* **Right:** *Average total delay for each year individually, for each plane age.*

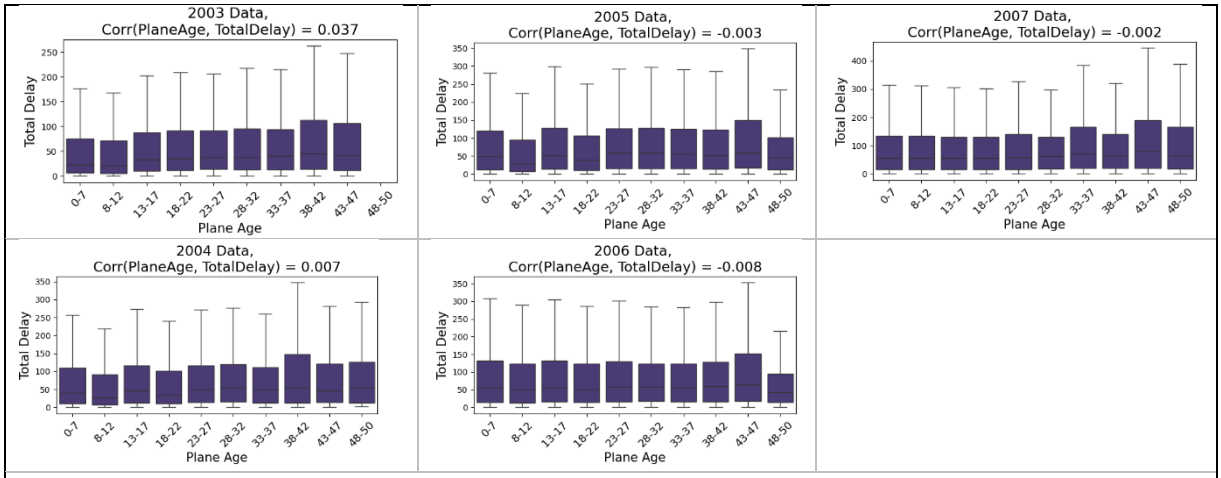


Figure 6, *Boxplots of total delay by different categories of plane age across all years. The correlation coefficient between 'PlaneAge' and 'TotalDelay' for that year is included at the top of each plot. Outliers were not included in these plots.*

Part C

In this part, the objective was to predict the probability of diverted flights using a logistic regression model with various features. The analysis began by merging flight data with airport data, using *'Origin'* and *'Dest'* from the flight data and '*iata'* from the airport data as common keys for joining. Carrier descriptions were also merged from the carriers dataset to obtain carrier names. Data cleaning involved removing observations with negative '*CRSElapsedTime'* and excluding features that lacked data for diverted flights (*'ArrTime', 'ActualElapsedTime', 'AirTime'*). Additionally, since it was mentioned in the carrier description that "America West Airlines Inc." and "US Airways Inc." have merged in September 2005, flights from these carriers were standardized under the name "US/America West merged" for flight records after this date, while their original carrier names were retained for earlier flights, this conditional renaming process required the creation of a datetime variable which is derived by combining *the 'Year', 'Month', 'DayofMonth',* and *'DepTime'* variables from the flight dataset.

2.1.1. Exploratory data analysis

Exploratory data analysis revealed that diverted flights consistently covered greater distances on average than non-diverted flights across all years. Peak travel months, such as June and July ("hot" months) and December and January ("cold" months), saw higher flight volumes and increased diversion rates. The highest percentage of diverted flights typically occurred in June or July, with Sundays also showing a higher tendency for flight diversions. The percentage of diverted flights for each month each year could be visualized in Figure 7. I have created indicator variables to indicate June, July, August flights as *"HotMonth"* and December, January flights as *"ColdMonth"*.

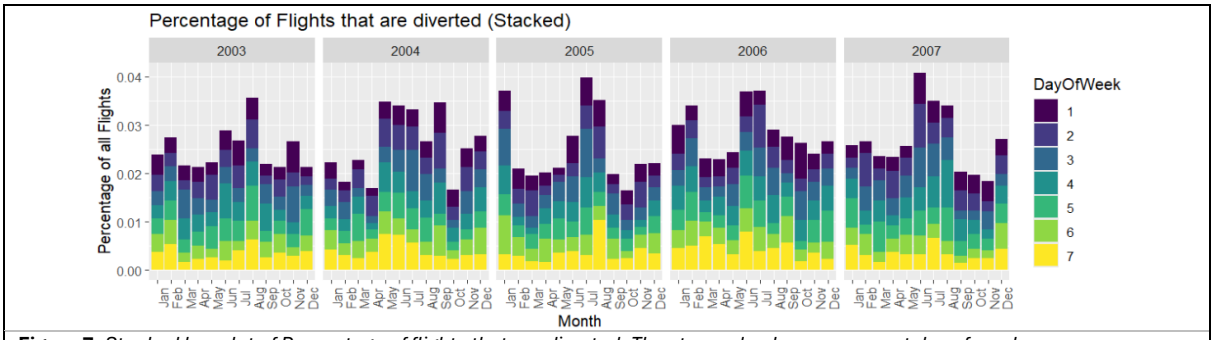


Figure 7, *Stacked bar plot of Percentage of flights that are diverted. The stepped colours represent day of week.*

Nearly all flights had the following characteristics: the majority were "multi-engine" aircraft with "Turbo-Fan" or "Turbo-Jet" engines, and almost all flights were corporate flights. However, when analysing the percentage of diverted flights across different aircraft and flight types, the less common types (e.g., "Turbo-Shaft" engine, partnership flights) showed a higher but inconsistent tendency for flight diversions. The percentage of diverted flights within each of these types are shown in Figure 8**.** To highlight these groups and reduce data redundancy, I created indicator variables: '*TurboEngine'* for flights with Turbo-Fan or Turbo-Jet engines, *'Corpo'* for corporate flights, and *'MultiEngine'* for multi-engine aircraft. These indicator variables would be used instead of their original categorical variables.

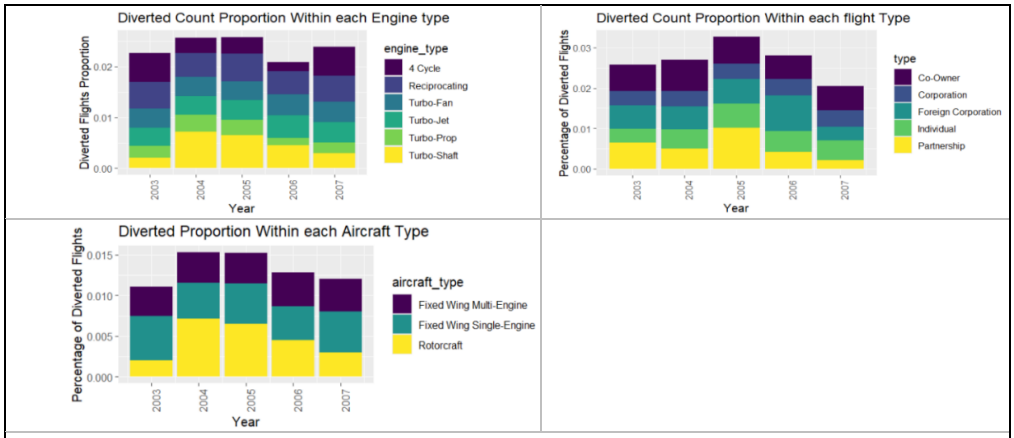


Figure 8, *Percentage of diverted flights within different aircraft and flight types.*

Data shows that most carriers have diversion rate below 0.4%, while "JetBlue Airways" and "Express Jet Airlines" have well above 0.5% diversion rate in 2006-2007. These two carriers were also quite often on the top of the list in terms of diverted flight percentage from 2003-2007. On the other hand, "Hawaiian Airlines" have consistently been least prone to flight diversions across the years. The percentage of diverted flights by carrier across 2004-2007 is shown in **Figure 9**.

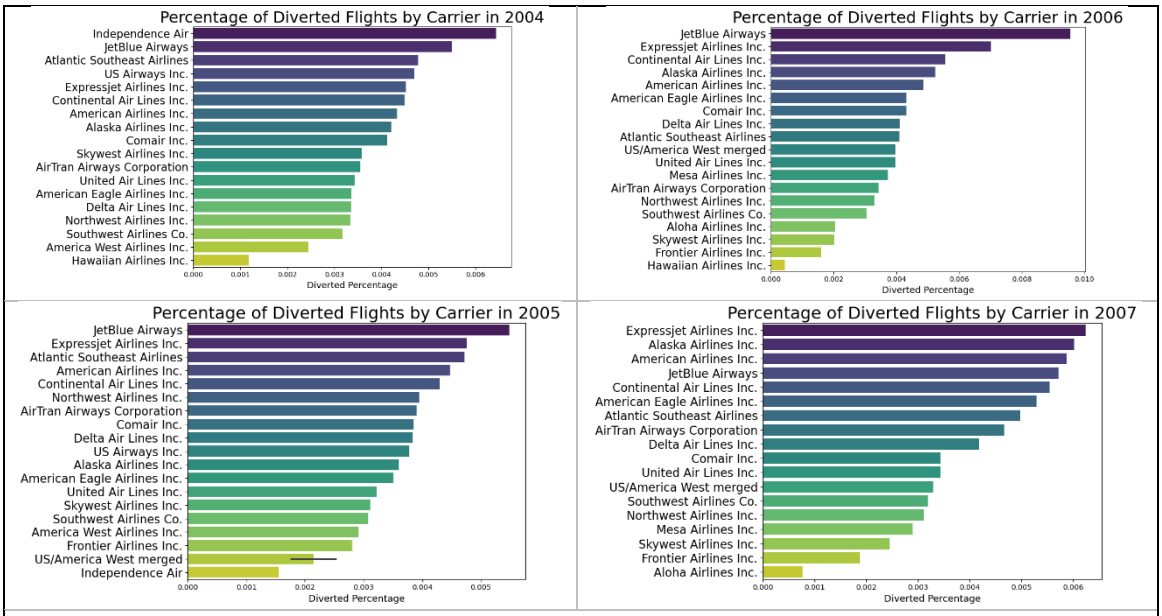


Figure 9, *Percentage of diverted flights by carrier across 2004-2007.*

Lastly, a correlation matrix of most numerical features was computed and visualized using a heatmap plot (**Figure 10**). Darker colours indicate stronger pairwise correlation, it is shown that all numerical features showed no clear correlation with the target variable 'Diverted'.

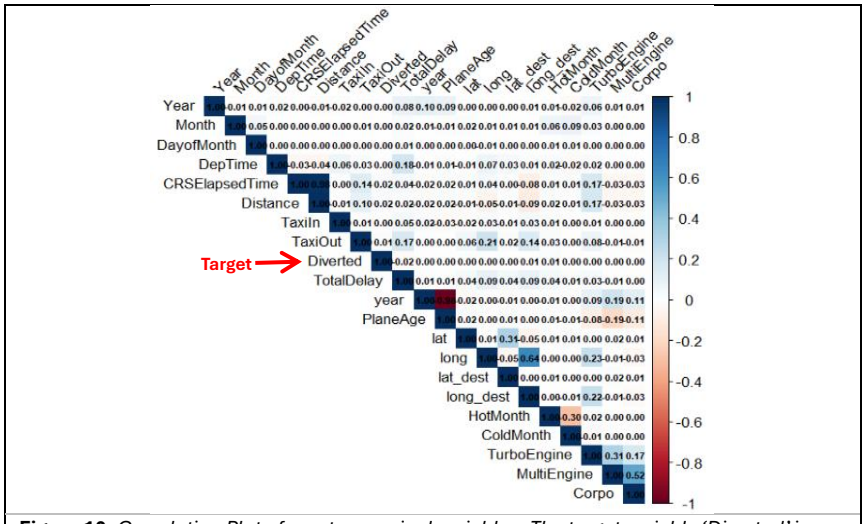


Figure 10, *Correlation Plot of most numerical variables. The target variable 'Diverted' is pointed at with a red arrow.*

2.1.2. Logistic Regression Model Training

To fit the logistic regression model, I have chosen a subset of data containing the chosen variables based on previous exploratory data analysis. The data was split into training (80%) and testing (20%) sets for each year separately. Remaining missing values within continuous variables were imputed with their mean, and missing values within categorical variables were imputed with their mode. The factor variable *"CarrierName"* was converted into indicator variables for each carrier. The logistic regression model was then trained on the training partition for each year, and its coefficients were plotted in a horizontal bar chart to visualize the impact of various features on the probability of diversion (**Figure 11**).

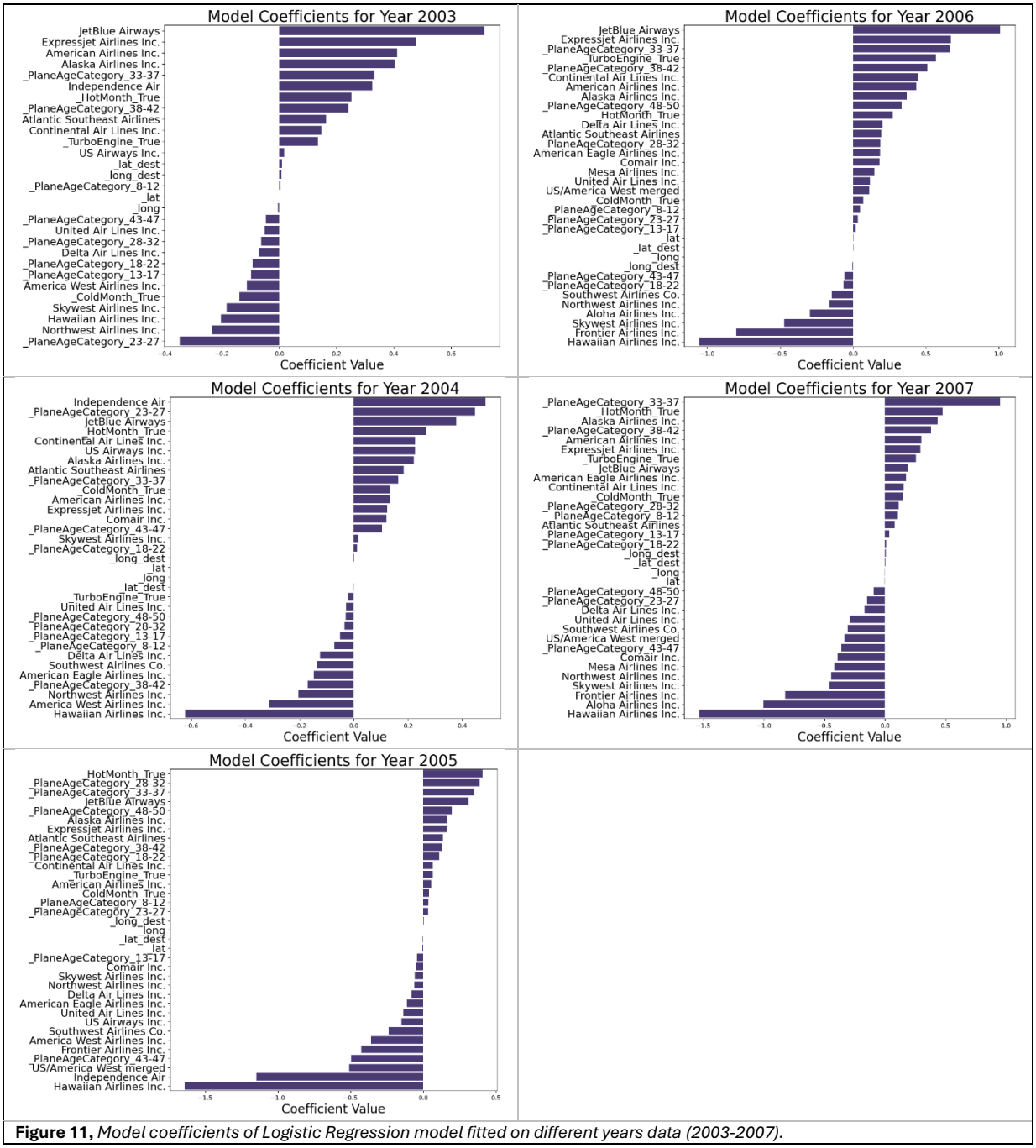


Figure 11, *Model coefficients of Logistic Regression model fitted on different years data (2003-2007).*

2.1.3. Interpreting Model coefficients

The plots in **Figure 11** shows that the *"HotMonth"* indicator (for flights taken in June or July) consistently exhibits a positive coefficient across all years. This suggests that flights during these months are more likely to be diverted compared to other months. Similarly, the *"PlaneAgeCategory"* indicator, representing the age range of the aircraft, consistently showed a strong positive coefficient for planes that are at least 23 years old each year. This indicates that older planes are more prone to diversion compared to newer ones. In addition, several flight carriers display significant negative coefficients, indicating that flights operated by these carriers are less likely to be diverted compared to the reference carrier (represented by the excluded *"CarrierName"* dummy variable), while positive coefficients suggest a higher likelihood of

diversion. A summary table in Figure 12 highlights the top two carriers most likely to divert and the top two least likely to divert each year. These findings were consistent with the diversion percentage results presented in Figure 9 .

Year	*Carriers most likely to divert (Top 2)*	*Carriers least likely to divert (Top 2)*
2003	○ JetBlue Airways ○ Expressjet Airlines	○ Northwest Airlines ○ Hawaiian Airlines
2004	○ Independence Air ○ JetBlue Alrways	○ Hawaiian Airlines ○ America West Airlines
2005	○ JetBlue Airways ○ Alaska Airlines	○ Hawaiian Airlines ○ Independence Airlines
2006	○ JetBlue Airways ○ Express Jet Airlines	○ Hawaiian Airlines ○ Frontier Airlines
2007	○ Alaska Airlines ○ American Airlines	○ Hawaiian Airlines ○ Aloha Airlines

Figure 12 *Top two carriers most likely to divert and the top two carriers least likely to divert each year, based on the coefficient values from a logistic regression model.*

It can be observed that "JetBlue Airways" had one of the highest positive coefficients among all carriers from 2003 to 2006, suggesting that it was more prone to diversions during these years. In contrast, "Hawaiian Airlines" consistently had one of the lowest coefficient values from 2003 to 2007, indicating that it was significantly less likely to experience diversions compared to other carriers.

2.1.4.Model performance

The model's performance was assessed using sensitivity, specificity, precision, and recall. The threshold of predicted probability to be classified as the positive class was adjusted to 0.004, this threshold was computed based on the proportion of diverted flights. The results revealed that the models' effectiveness does not fluctuate much across different fitted years of data, generally exhibiting nearly zero precision and low recall and sensitivity, while specificity remained moderate, the area under the ROC curve remained at around 59% for all years.

Despite the weak predictive performance of the model, the coefficients for different features provided valuable insights on flight performance. Future improvements could come from exploring additional features or employing more sophisticated machine learning models and techniques to address class imbalance and enhance model performance. Figure 13 illustrates the performance metrics and ROC curve of the model across various years.

Year of Model	2003	2004	2005	2006	2007
Precision	*0.01*	*0.00*	*0.01*	*0.01*	*0.01*
Recall	*0.45*	*0.44*	*0.48*	*0.57*	*0.62*
Sensitivity	*0.45*	*0.44*	*0.48*	*0.57*	*0.62*
Specificity	*0.70*	*0.65*	*0.64*	*0.56*	*0.55*

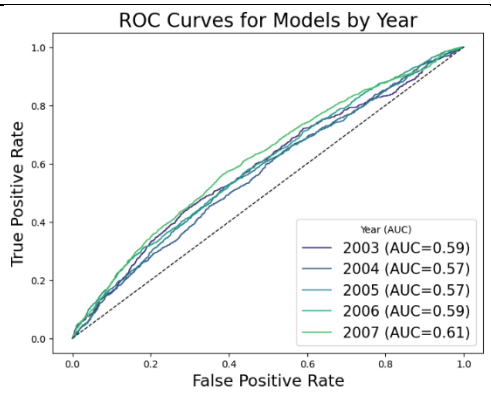


Figure 13, Left: *Model performance of logistic regression model fitted on different years data.* **Right:** *ROC Curves for each model fitted on different years of data.*

3. Reference

2008, "Data Expo 2009: Airline on time data", https://doi.org/10.7910/DVN/HG7NV7, Harvard Dataverse, V1