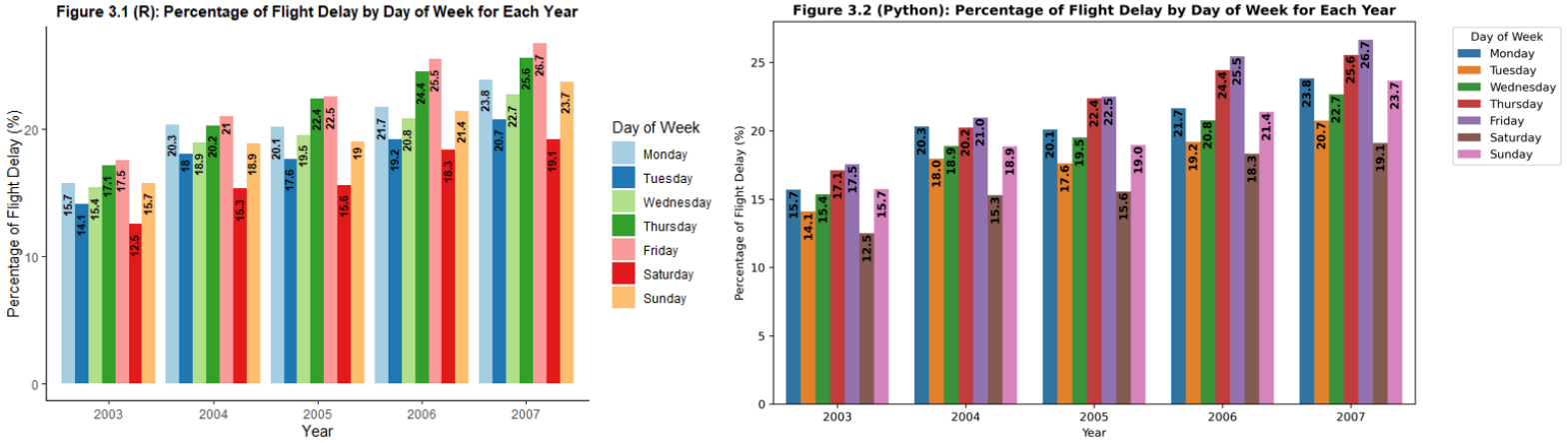
The dataset includes information on flight arrival and departure for all commercial flights in the United States from 2003 to 2007. It contains a variety of fields, such as the date, time, airline, origin and destination airports, and information related to departure and arrival delays.

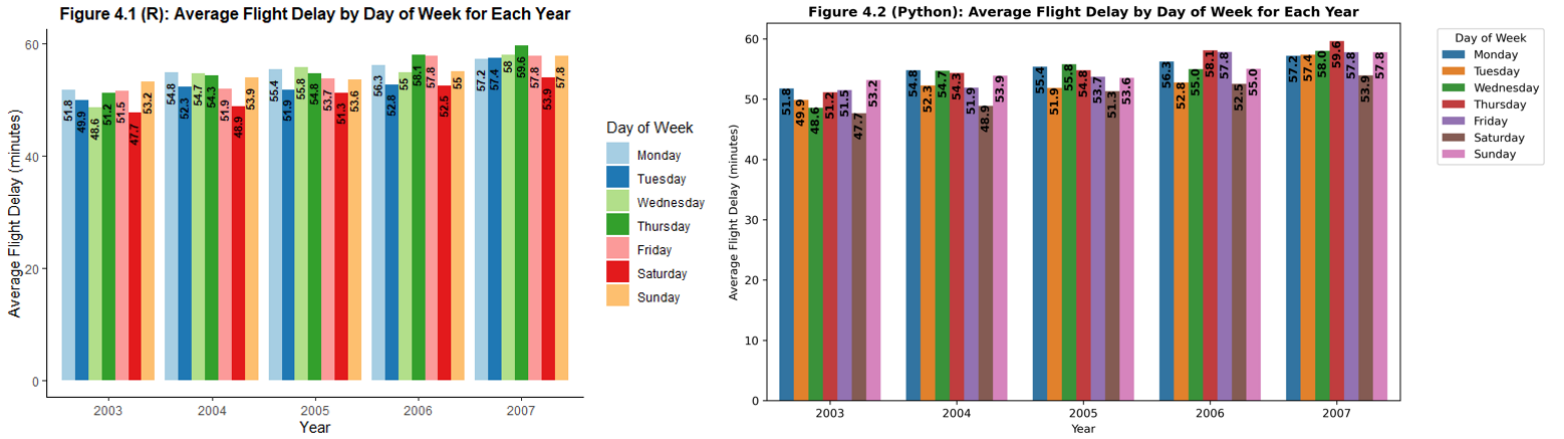
(a) What are the best times and days of the week to minimise delays each year?

(i) Best day of the week

Arrival delay is selected as the primary indicator of delay as it represents the total delay experienced by passengers, making it the most relevant metric for minimising delays. According to the Bureau of Transportation Statistics (2021), a flight that arrives more than 15 minutes is delayed. Therefore, flights with an arrival delay exceeding 15 minutes are categorized as delayed.



Figures 3.1 and 3.2 show the percentage of flights delayed by day of week from 2003 to 2007. There is an increasing trend in flight delays over the years, with notable fluctuations on different days of the week. We can see that delays tend to be more frequent on certain days such as Thursday and Friday across all years. Friday consistently recorded the highest percentage of delays from 2003 to 2007 with an increase from 17.5% to 26.7%. Meanwhile, Saturday consistently recorded the lowest percentage of delays from 2003 to 2007 with an increase from 12.6% to 19.5%. This pattern suggests that travellers should anticipate longer delays towards the end of the work week, especially on Thursday and Friday.



Figures 4.1 and 4.2 show the average flight delay by day of week from 2003 to 2007. Similarly, Saturday has the lowest average delay from 2003 to 2007 with an increase in duration from 47.7 minutes to 53.9 minutes. This suggests a decline in flights punctuality over the years with flight delays becoming more frequent and longer. Despite this increasing trend, Saturday remains the best

day with the lowest delays for 5 consecutive years. Therefore, the best day to minimize flight delays is on Saturday.

(ii) Best Time

The arrival time is used as an indicator to analyse how delays vary throughout the day based on the actual arrival time of flights. To facilitate this analysis, arrival time is divided into 8 groups, each spanning a three-hour interval: 0000–0259, 0300–0559, 0600–0859, and so on. This grouping provides insights across different times of the day and identifies trends between specific time windows and the duration of delays.

Figure 5.1 (R): Percentage of Flight Delay by Time Interval for Each Year

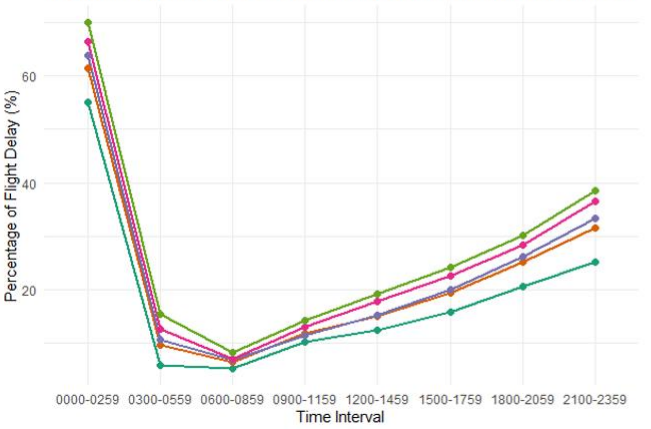
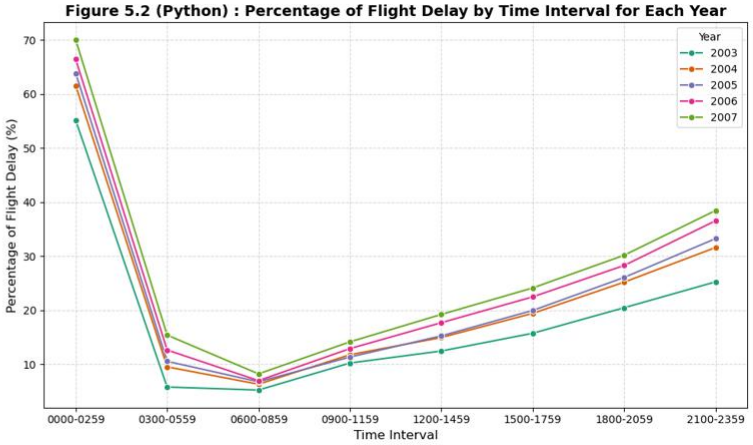


Figure 5.2 (Python) : Percentage of Flight Delay by Time Interval for Each Year



	Year	TimeInterval	DelayPercentage
0	2003	0600-0859	5.204578
1	2004	0600-0859	6.287805
2	2005	0600-0859	6.737836
3	2006	0600-0859	6.947213
4	2007	0600-0859	8.198985

Figures 5.1 and 5.2 show the percentage of flights delayed by time interval. The table on the left shows the lowest percentage of flight delays in each year. Early morning hours from 0600 to 0859 consistently show the lowest percentage of flight delay from 2003 to 2007 with an increase from 5.2% to 8.2%.

Figure 6.1 (R): Average Flight Delay by Time Interval for Each Year

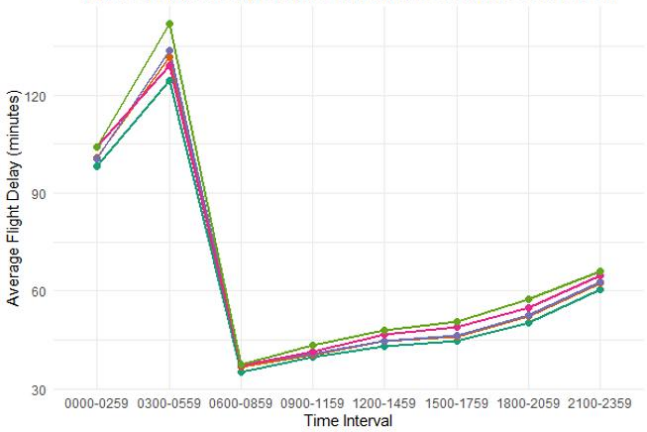
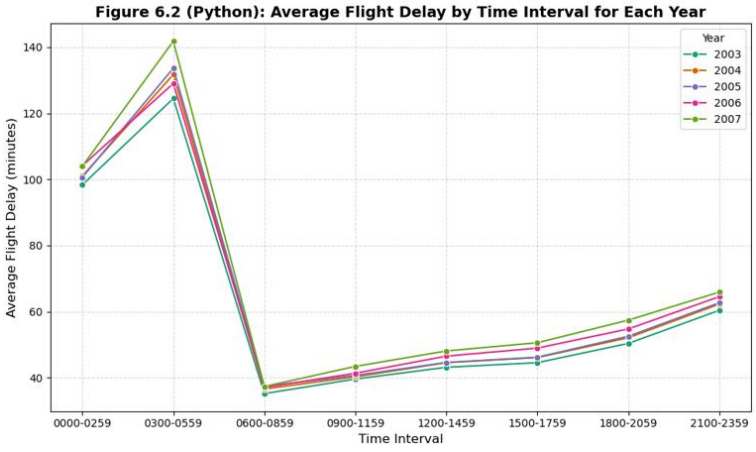


Figure 6.2 (Python): Average Flight Delay by Time Interval for Each Year



	Year	TimeInterval	AvgDelay
0	2003	0600-0859	35.196991
1	2004	0600-0859	36.594608
2	2005	0600-0859	37.383669
3	2006	0600-0859	37.062926
4	2007	0600-0859	37.341756

Figures 6.1 and 6.2 show the average delay by time interval. We observe that flight delays are generally highest in the early morning hours especially during 0300 to 0559. This indicates that flights during this period experience significantly longer delays than other times of the day. Following this peak, there is a sharp drop in average delay during the 0600 to 0859, where delays reach their lowest point from 2003 to 2007 as indicated by the table on the left, which shows the lowest average flight delay for each year. However, the average flight delay during this interval has increased from 35.2 minutes to 37.3 minutes over the years. Despite this slight rise, early mornings between 0600 and 0859 remain the best option for minimizing delays.

Conclusion

Saturday and early mornings from 0600 to 0859 are the best day and times to minimise flight delays. Travelers seeking to reduce the likelihood of delays should prioritize early morning flights and consider scheduling their trips on Saturdays. By focusing on these optimal times and days, airlines and passengers alike can enhance punctuality and improve overall travel efficiency.

(b) Evaluate whether older planes suffer more delays on a year-to-year basis.

(i) Distribution of the number of flights by plane age

We first want to find the distribution of flights based on plane age to understand how many flights the plane would take based on their age.

Figure 7.1 (R): Distribution of Number of Flights by Planes Age

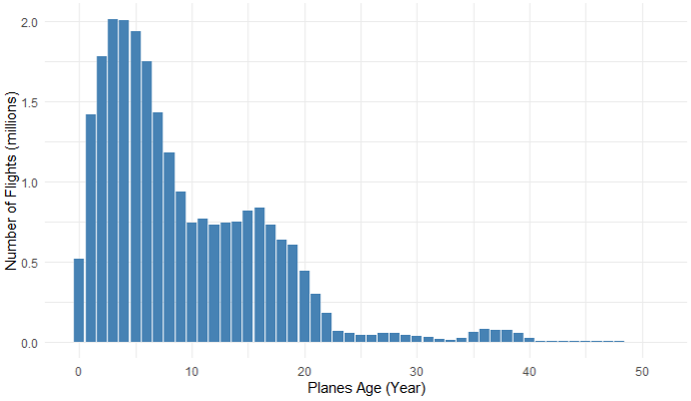
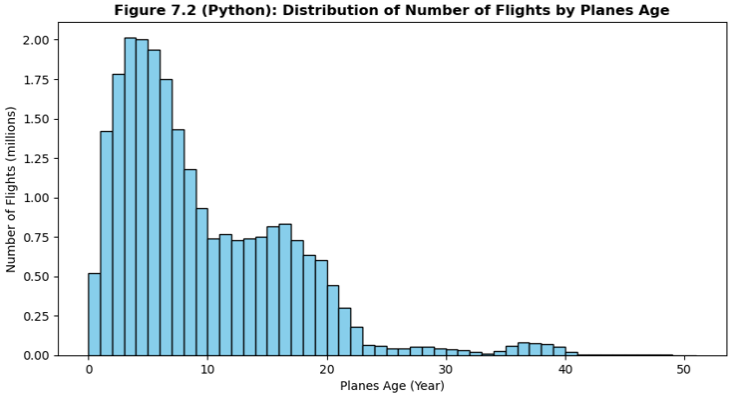


Figure 7.2 (Python): Distribution of Number of Flights by Planes Age



Figures 7.1 and 7.2 show the distribution of the number of flights by the plane's age at the time it was used to fly from 2003 to 2007. We can see that planes older than 22 years have significantly fewer flights compared to planes younger than 22 years old. Due to this smaller sample size, the average delay of planes aged more than 22 may not accurately represent the true average delay. To address this, we will analyze planes aged between 0 and 22 and planes aged more than 22 to better understand the relationship between plane age and flight delays.

(ii) Percentage of Flights Delayed

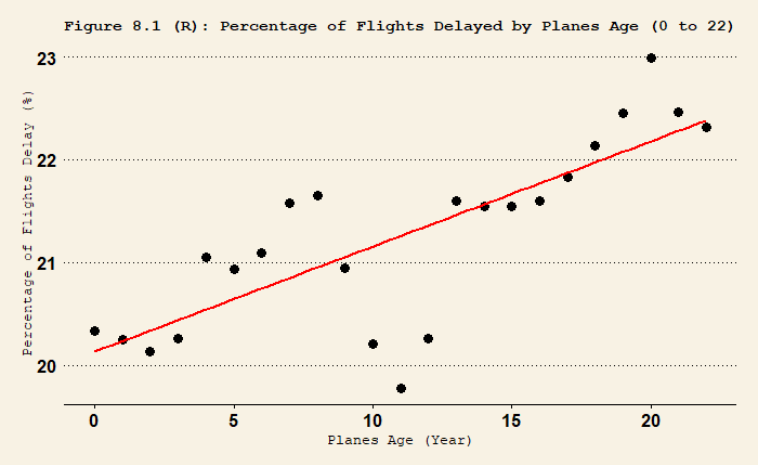
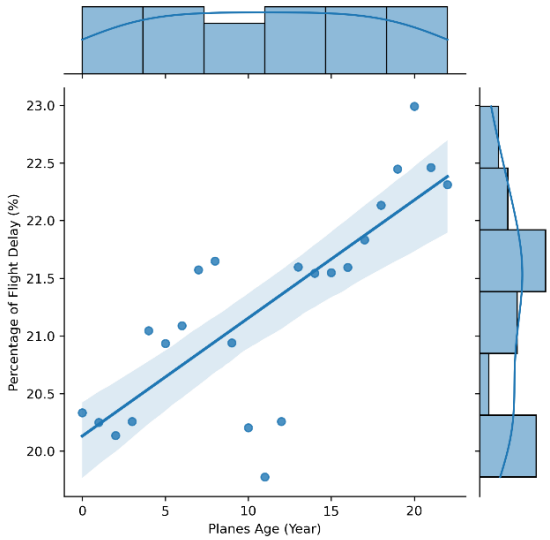


Figure 8.2 (Python): Percentage Delay based on Age of Plane (0 to 22)



Figures 8.1 and 8.2 are scatterplots on the percentage of flight delay of planes aged between 0 and 22 years. There is a positive correlation indicating that as planes get older, the percentage of delayed flights increases. Although the trend is upward-sloping, individual data points exhibit noticeable variations.

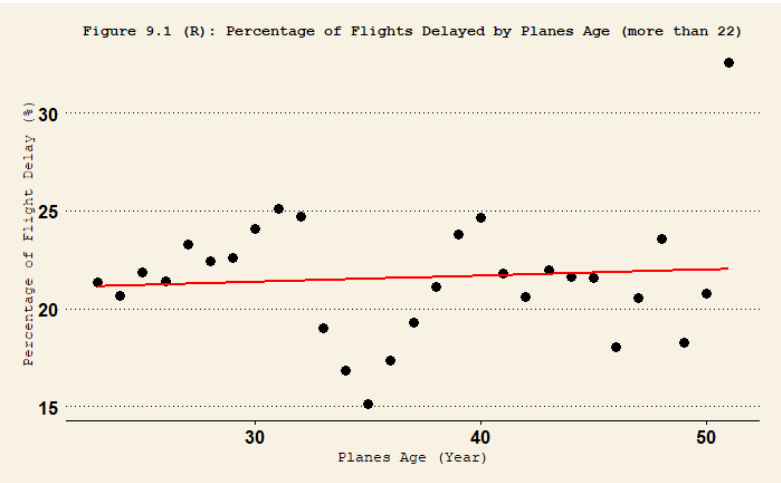
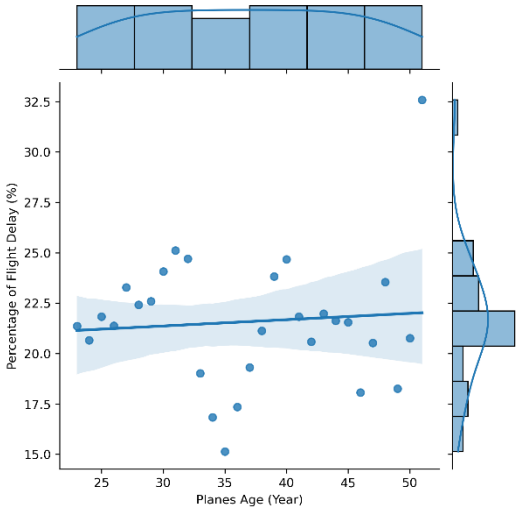


Figure 9.2 (Python): Percentage Delay based on Planes Age (more than 22)



Figures 9.1 and 9.2 are scatterplots on the percentage of flight delay of planes aged more than 22 years. There is a weak positive correlation between plane age and the percentage of delays. This suggests that older planes aged over 22 years may experience a small increase in delay percentage, but the effect is not strong. The data points are widely scattered around the trend line, suggesting that older planes do not necessarily experience higher delays at a consistent rate.

(iii) Average Flight Delay

Figure 10.1 (R): Average Flight Delay based on Planes Age (0 to 22)

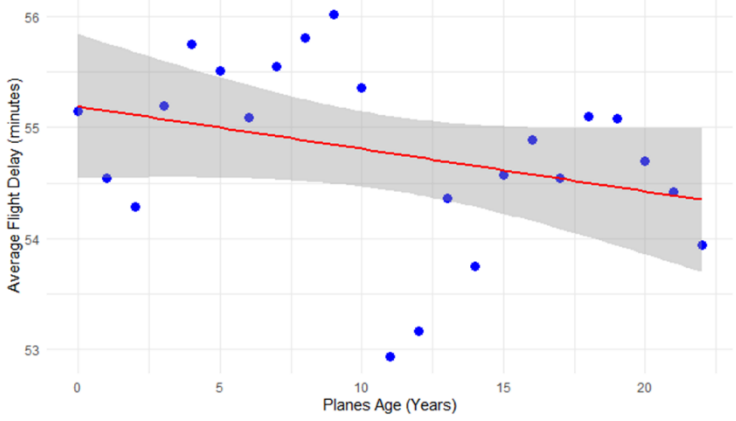
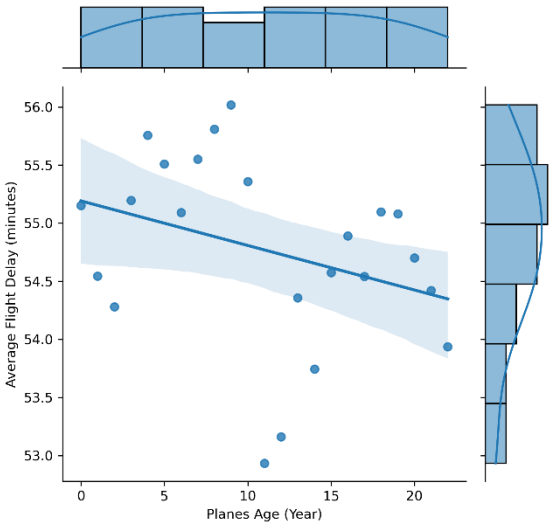


Figure 10.2 (Python): Average Flight Delay based on Plane Age (0 to 22)



Figures 10.1 and 10.2 are scatterplots showing the average flight delay of planes aged between 0 to 22, with regression lines to visualize the trend. There is a negative correlation between average flight delay and plane age between 0 and 22. The regression line slopes downward suggest that as planes age within this range, delays tend to decrease.

Figure 11.1 (R): Average Flight Delay based on Planes Age (more than 22)

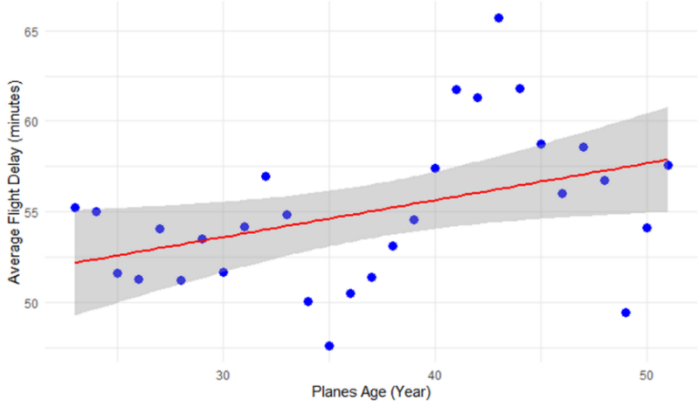
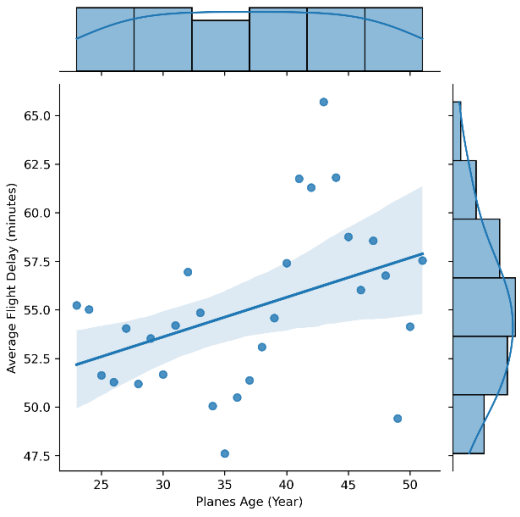


Figure 11.2 (Python): Average Flight Delay based on Planes Age (more than 22)



Figures 11.1 and 11.2 are scatterplots showing the average delay of planes aged more than 22, with regression lines to visualize the trend. There is a positive correlation between plane age and flight delay. The regression line slopes upwards, indicating that as planes get older, they tend to experience longer delays. However, the relationship is weak as indicated by the scattered data points and the wide confidence interval around the regression line.

Overall, these findings suggest that average delays decrease as planes age from 0 to 22 years but begin to increase for planes older than 22 years. However, the trend for planes older than 22 years is less reliable due to the smaller sample size. The data for planes aged 0 to 22 years is more representative and reliable, making it the focus of our yearly trend analysis.

(iii) Yearly Trend of Flight Delays

We will examine the yearly trends of flight delays for planes aged 0 to 22 years.

Figure 12.1 (R): Yearly Trend of Percentage Delay based on Planes Age (0 to 22)

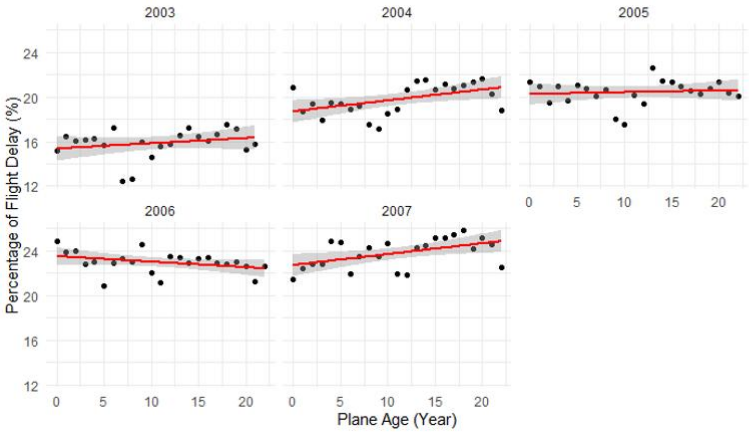
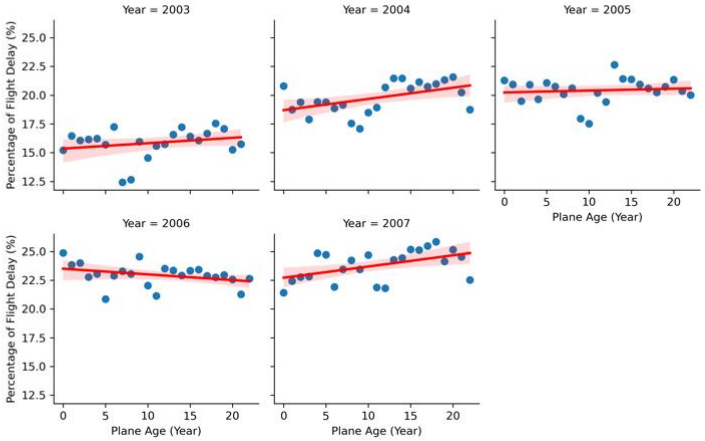


Figure 12.2 (Python): Yearly Trend of Percentage Delay based on Planes Age (0 to 22)



Figures 12.1 and 12.2 show the yearly trend of the percentage delay from 2003 to 2007. In 2003, 2004 and 2007, there is a positive correlation between plane age and the percentage of flights delayed. This suggests that in these years, older planes generally experienced a higher percentage of delays compared to younger planes. In 2005, the trend appears to be flat, indicating little relationship between plane age and percentage delay. This suggests that plane age did not impact the delay. In 2006, there is a negative correlation. This suggests that newer planes had a slightly higher percentage of delays compared to older planes.

Figure 13.1 (R): Yearly Trend of Average Delay based on Planes Age (0 to 22)

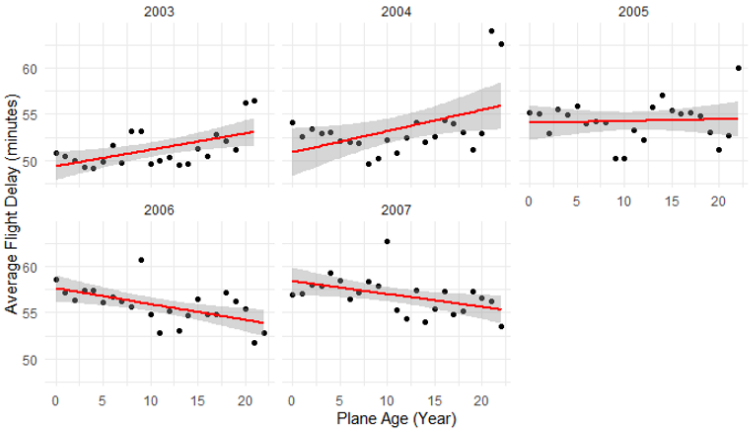
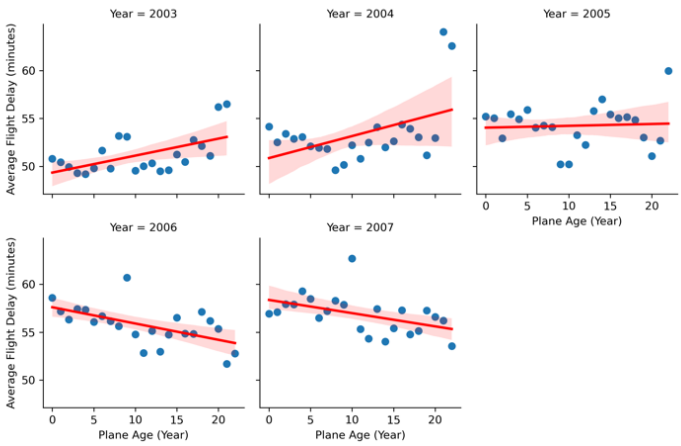


Figure 13.2 (Python): Yearly Trend of Average Delay based on Planes Age (0 to 22)



Figures 13.1 and 13.2 show the yearly trend of the average delay from 2003 to 2007. In 2003, 2004, and 2005, there is a positive correlation that suggests that older planes tend to have slightly higher average flight delays. The confidence intervals in 2004 and 2005 are slightly wider, indicating some

uncertainty in the trend. In 2006 and 2007, there is a negative correlation, where older planes exhibit shorter average delays compared to newer ones.

Conclusion

While older planes aged more than 22 experience increasing delays over time, there is a weak correlation between plane age and delays. For planes aged between 0 and 22, the percentage of delayed flights increases with age, but their average delay time does not necessarily worsen. While there is some evidence that plane age contributes to delays, the effect is not consistent across all years, indicating plane age is not the only determining factor for flight delays.

(c) Logistic regression model

The aim is to create a logistic regression model for the probability of diverted US flights from 2003 to 2007 based on various flight features such as scheduled departure and arrival times, distance and carrier. To better understand these trends, we will visualize the regression coefficients across the years. We will use mlr3 package for R and sklearn package in Python to build the model as these packages are efficient tools for machine learning.

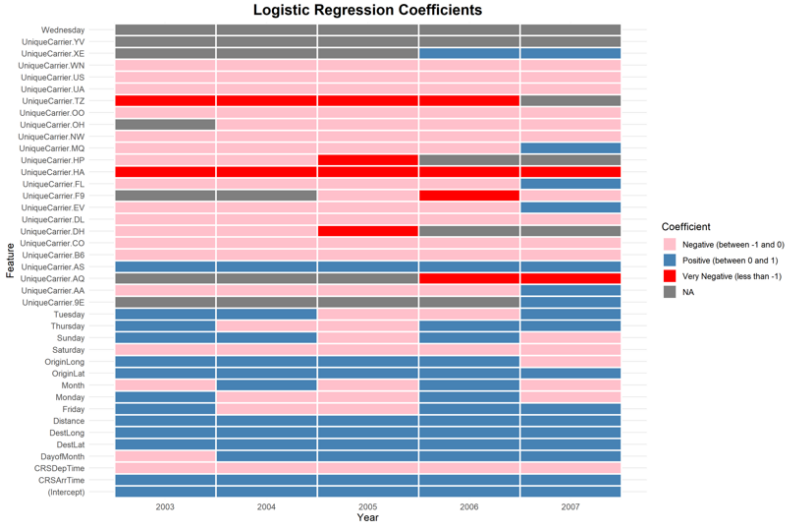


Figure 14.1 – R

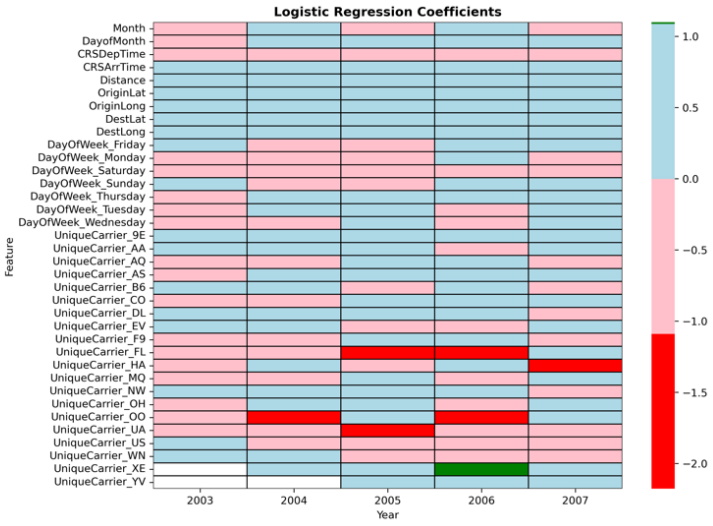


Figure 14.2 – Python

Figures 14.1 and 14.2 show a heatmap of the logistic regression coefficients over years using R and Python respectively. In Figure 14.1, the coefficient ranges from -1 to 1. We can see that some airlines such as Carrier HA show strong negative coefficients, highlighted in red. This suggests that flights operated by this airline were significantly less likely to be diverted during those periods. Meanwhile, other flight attributes such as certain days of the week show positive coefficient, highlighted in blue. Some flight attributes are highlighted in grey, which indicates insufficient data or a lack of significant correlation with diversions. On the other hand, the coefficient ranges from -2 to 1 in Figure 14.2. Similarly, Carrier HA exhibit strong negative coefficients in 2007. Carrier XE show more positive coefficients in 2006, highlighted in green. This indicates that flights operated by this carrier were more likely to be diverted during that year. Other features such as distance and time generally have more consistent positive or negative effects across years.

Conclusion

Factors such as time and airport locations appear to have less impact on the probabilities of flight diversion while airline carriers seem to have some impact on the probabilities of flight diversion.