Coursework Project Part 2

ST2195-Programming for Data Science

Richard Hardy Mathias

Student id: 230655968

Table of Contents

1. Part Two

This report's objective is to investigate the flight data from June 2003 to December 2007 for all commercial flights operated by major US airlines, as supplied by the 2009 ASA Statistical Computing and Graphics Data Expo, with the intent to answer the following queries:

a) What are the best times and days of the week to minimise delays each year?
b) Evaluate whether older planes suffer more delays on a year-to-year basis.
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.

To create databases, retrieve queries from the database and add data to the database using libraries like DBI for R and sqlite3 for Python, the flight data will be processed and worked on in the R and Python programming languages. Additionally, ggplot2 for R and matplotlib for Python will be used to plot graphs and tables that will be required for our analysis.

2. Compiling the files

Using Jupyter notebook, I first downloaded the relevant csv files from the 2009 ASA Statistical Computing and Graphics Data Expo website. The provided data were in csv.bz2 and .csv formats.I first converted the csvbz2 to csv by extracting the files. Then I read all the files and define it as file_paths. I then proceed to clean the data first by checking for missing values and dropping rows with missing DepDelay values. After going through the datasets, the extra datasets I downloaded can be arranged as such:

1. “aircraft_info.csv” has information about individual planes and their details

2. “airports.csv” contains data on the flight carriers operating within the US

3. “carriers.csv” contains the different flight providers in the US

4. “variable-descriptions.csv” contains data on the various planes from October 1987 to April 2008

2.1. What is a delay

Arrival, departure, carrier, weather, NAS, security and late aircraft delay are just a few of the several types of delays that are present in the data. The discrepancy between the scheduled and actual arrival and departure times, respectively, is referred to as an arrival or departure delay when it results in a positive value in minutes. When we run the factors through a linear regression model against arrival and departure delay, we obtain adjusted r-squared values of 0.94 and 0.89, respectively, to ascertain the relationship between carrier, weather, NAS, security and late aircraft delays and arrival and departure delays (see Figure 1 below)

Linear Regression Results for Departure Delay

OLS Regression Results

Dep. Variable:	DepDelay	R-squared:	0.892
Model:	OLS	Adj. R-squared:	0.892
Method:	Least Squares	F-statistic:	1.657e+05
Date:	Tue, 01 Apr 2025	Prob (F-statistic):	0.00
Time:	10:56:14	Log-Likelihood:	-3.7953e+05
No. Observations:	100000	AIC:	7.591e+05
Df Residuals:	99994	BIC:	7.591e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9723	0.036	-26.927	0.000	-1.043	-0.902
CarrierDelay	1.0146	0.002	536.021	0.000	1.011	1.018
WeatherDelay	0.9770	0.004	267.133	0.000	0.970	0.984
NASDelay	0.5403	0.002	240.410	0.000	0.536	0.545
SecurityDelay	1.0110	0.033	30.780	0.000	0.947	1.075
LateAircraftDelay	1.0253	0.002	581.444	0.000	1.022	1.029

Omnibus:	225832.212	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31400678187.988
Skew:	-20.192	Prob(JB):	0.00
Kurtosis:	2747.910	Cond. No.	21.7

Linear Regression Results for Arrival Delay

OLS Regression Results

Dep. Variable:	ArrDelay	R-squared:	0.941
Model:	OLS	Adj. R-squared:	0.941
Method:	Least Squares	F-statistic:	3.192e+05
Date:	Tue, 01 Apr 2025	Prob (F-statistic):	0.00
Time:	10:56:14	Log-Likelihood:	-3.5851e+05
No. Observations:	100000	AIC:	7.170e+05
Df Residuals:	99994	BIC:	7.171e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.2195	0.029	-144.196	0.000	-4.277	-4.162
CarrierDelay	1.0332	0.002	673.536	0.000	1.030	1.036
WeatherDelay	1.0247	0.003	345.725	0.000	1.019	1.031
NASDelay	1.0530	0.002	578.180	0.000	1.049	1.057
SecurityDelay	1.0774	0.027	40.478	0.000	1.025	1.130
LateAircraftDelay	1.0373	0.001	725.809	0.000	1.034	1.040

Omnibus:	6795.152	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10048.864
Skew:	-0.570	Prob(JB):	0.00
Kurtosis:	4.054	Cond. No.	21.7

Measure	R2	Adjusted R2	F_Statistic
Arrival Delay	0.941045	0.941042	319221
Departure Delay	0.892291	0.892286	165676

Model Summary for Arrival and Departure Delay

Measure	R2	Adjusted_R2	F_Statistic
Arrival Delay	0.938975	0.9389720	307716.2
Departure Delay	0.885046	0.8850403	153973.4

Figure 1: Summary of Linear Regression Model for Departure and Arrival Delay in python and R

3. What are the best times and days of the week to minimise delays each year?

To identify the best times and days to minimize flight delays, I analysed departure delay trends from the years 2003-2007. The data showed a persistent pattern of the lowest average delays for early morning departures, especially those scheduled between 6:00 AM and 9:00 AM. Delays typically get worse as the day goes on, reaching their zenith in the late afternoon and evening. Fridays and Sundays were linked to higher average delays, most likely because of higher travel traffic, whilst Saturdays and Tuesdays typically displayed the lowest average delays. According to this data, travellers who want to leave on time should try to book early morning flights on weekdays like Tuesday or Saturday.

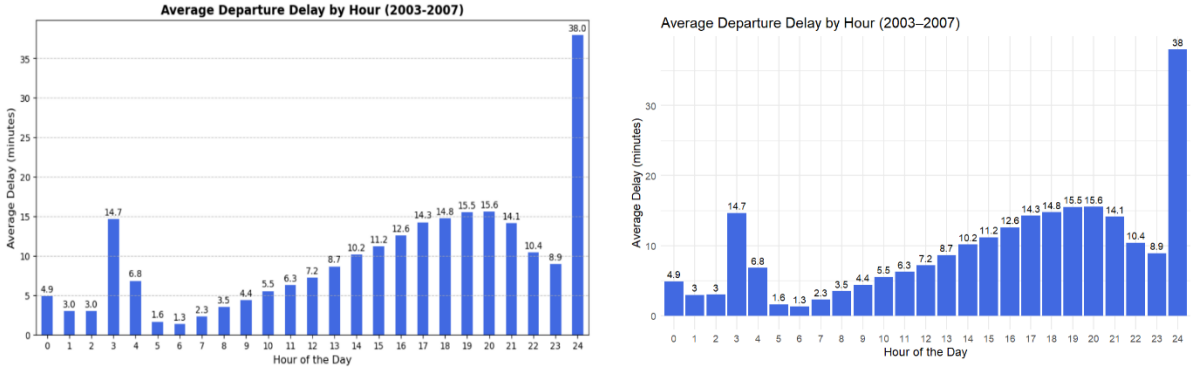

Figure 2: Average departure delay by hour (2003-2007) Python and R

Based on Figure 2, the most significant spike is observed at midnight (24:00 hours), which shows an unusually high average delay of 38 minutes, potentially due to cumulative delays from throughout the day or data sparsity. The lowest delay occurred at 6 AM, with an average of just 1.3 minutes. These insights suggest that travelling in the morning is more reliable than travelling in the evening.

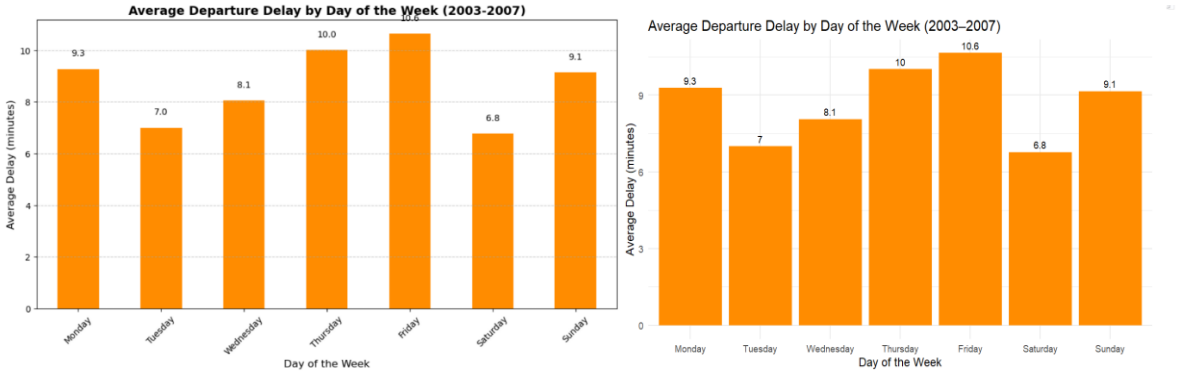

Figure 3:Average Departure Delay by the week (2003-2007) Python and R

Based on Figure 3, the data shows that Saturday came as the lowest average delay of 6.8 minutes, followed by Tuesday with a delay of 7.0 minutes. In contrast, Friday and Thursday have the highest average delays at 10.6 and 10.0 minutes respectively.

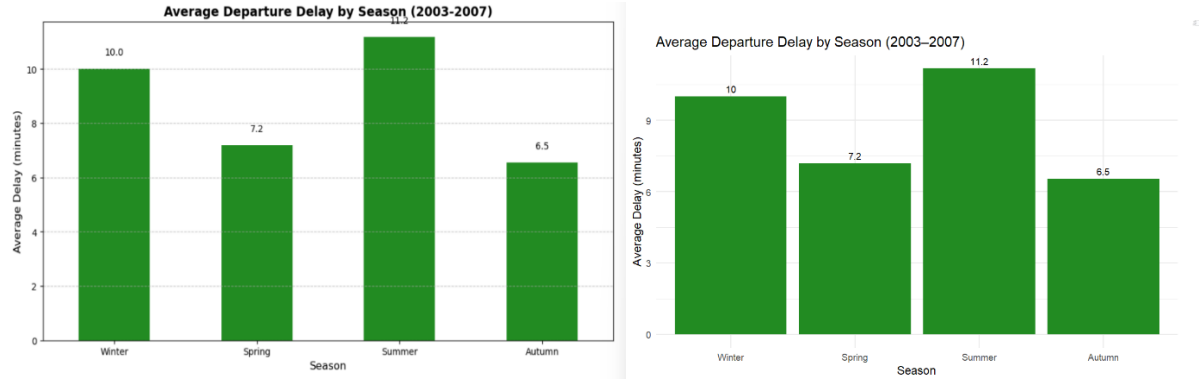

Figure 4: Average Departure Delay by Season (2003-2007) Python and R

From Figure 4, Significant variations throughout the year are revealed by the seasonal examination of average departure delays. With a peak of 11.2 minutes, summer has the longest average delay, followed by winter with 10.0 minutes. Increased demand for travel over the holidays and unfavourable winter weather conditions may be the cause of these higher delays. The most dependable seasons for on-time departures is during autumn and spring, when average delays are much smaller at 6.5 and 7.2 minutes, respectively. With the lowest average delays over the five-year period, the data indicates that autumn is the ideal time of year to fly for passengers who value timeliness.

4. Do Older Planes Suffer More Delays on a year to year Basis?

We use the percentage of flights across age groups of planes to compare if older aircraft experience higher delays. This metric, which contrasts the frequency of delayed flights for a given age group with the overall frequency of flights in that age group, is the most representative. Similarly, a flight will be deemed delayed if its arrival time is greater than or equal to 15 minutes.

4.1. Percentage of delayed flights by age group

An examination of flight delays across different aircraft age groups reveals a slight but consistent increase in delay rates as planes age. As shown by figure 5 below, Aircraft in the 0–4 year age group had the lowest delay percentage at 21.3%, while those in the 20+ year category experienced the highest delay rate, at 22.8%. The trend indicates that older aircraft are slightly more likely to have delays, despite the small variances. This might be because older aircraft have more frequent

technical problems or require more maintenance. The comparatively minor difference across all age groups, however, suggests that although age may contribute to delays, it is not a primary cause of them.

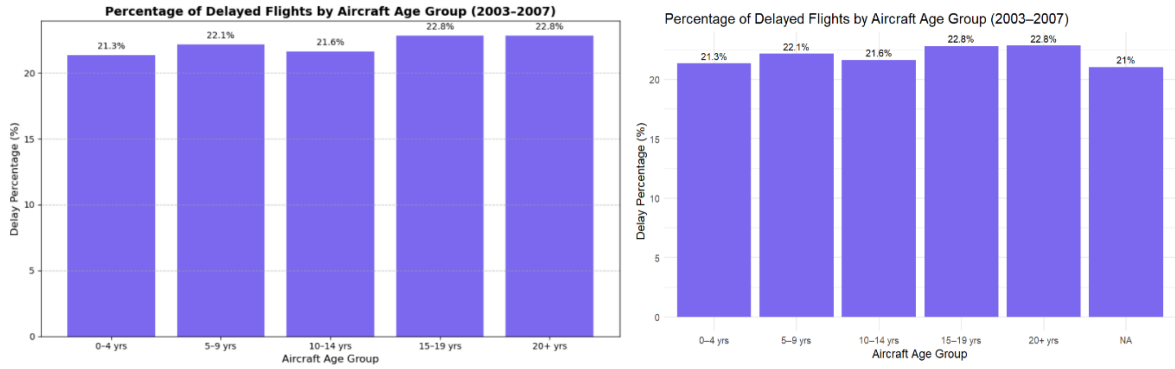

Figure 5: Percentage of Delayed Flights by Aircraft Age Group (2003-2007) Python and R

The violin plot used below in figure 6, shows the distribution of flight delay durations for three aircrafts age groups-new, standard and old aircraft. The median delay duration is consistent at around 12 minutes across all age groups indicated by the bolded values in the centre of each boxplot. All aircraft types appear to have comparable delay profiles, with most delays concentrated in the lower range, according to the general distribution patterns. This indicates that aircraft age does not play a major role in the length of delays, supporting the idea that other factors—such as air traffic control, weather, or scheduling—likely contribute more significantly to delay duration than the aircraft's age.

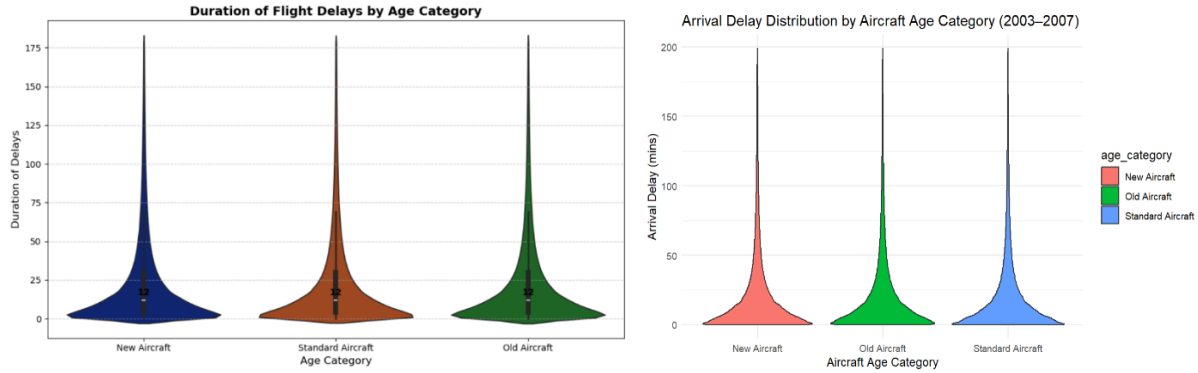

Figure 6: Duration of Flight Delays by Age Category Python and R

5. Logistic Regression Model for probability of diverted US flights

Figure 7 & 8 below shows how the influence of various flight features on the probability of a flight being diverted changed over the years 2003 to 2007, based on logistic regression coefficients. For R, I used a sample of 200000 since the it couldn't run properly. Among all features, distance consistently showed the strongest positive relationship with diversion probability. This suggests that longer flights are more likely to be diverted, possibly due to increased exposure to weather or in-flight emergencies over time.

Scheduled arrival hour (CRSArrHour) also showed a relatively strong and growing positive coefficient, especially in 2007, indicating that flights scheduled to arrive later in the day may be at greater risk of diversion, potentially due to accumulated delays or airspace congestion.

On the other hand, scheduled departure hour (CRSDepHour) displayed negative coefficients throughout, with its influence becoming more pronounced (more negative) in 2007. This may imply that earlier departures are less likely to be diverted, possibly due to calmer traffic conditions and fewer cascading delays.

Geographic features like origin and destination coordinates (latitudes and longitudes) showed small but notable variation in impact, suggesting some routes or regions might be more prone to diversions. Meanwhile, carrier identity (CarrierEncoded) had minimal but slightly varying influence, possibly reflecting differences in airline operational practices or fleet capabilities.

Overall, the model indicates that flight timing and distance are key factors influencing diversions, and their effect has remained relatively stable or increased slightly over time.

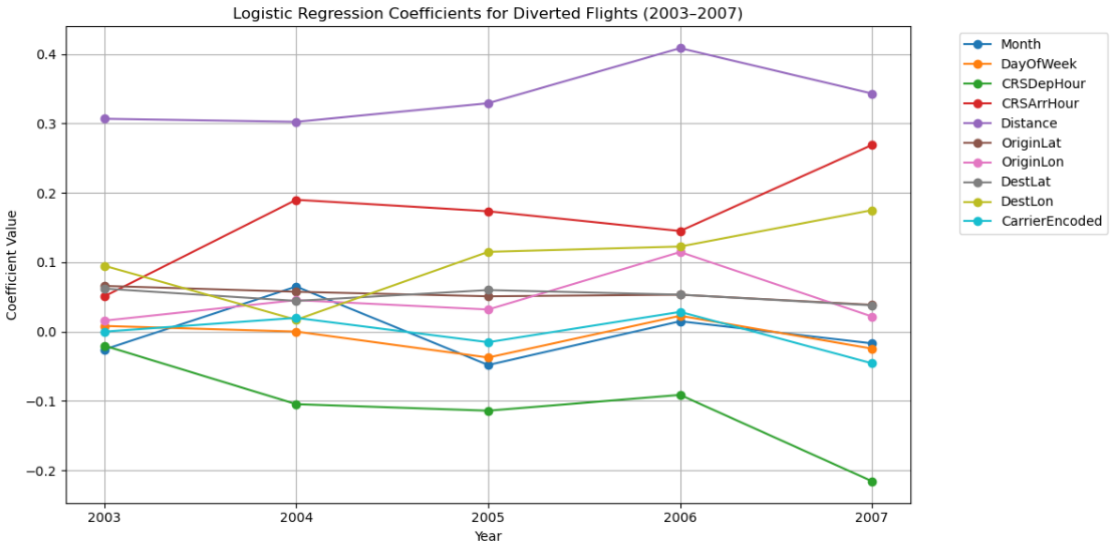

Figure 7: Logistic Regression Coefficients for Diverted Flights (2003-2007) Python

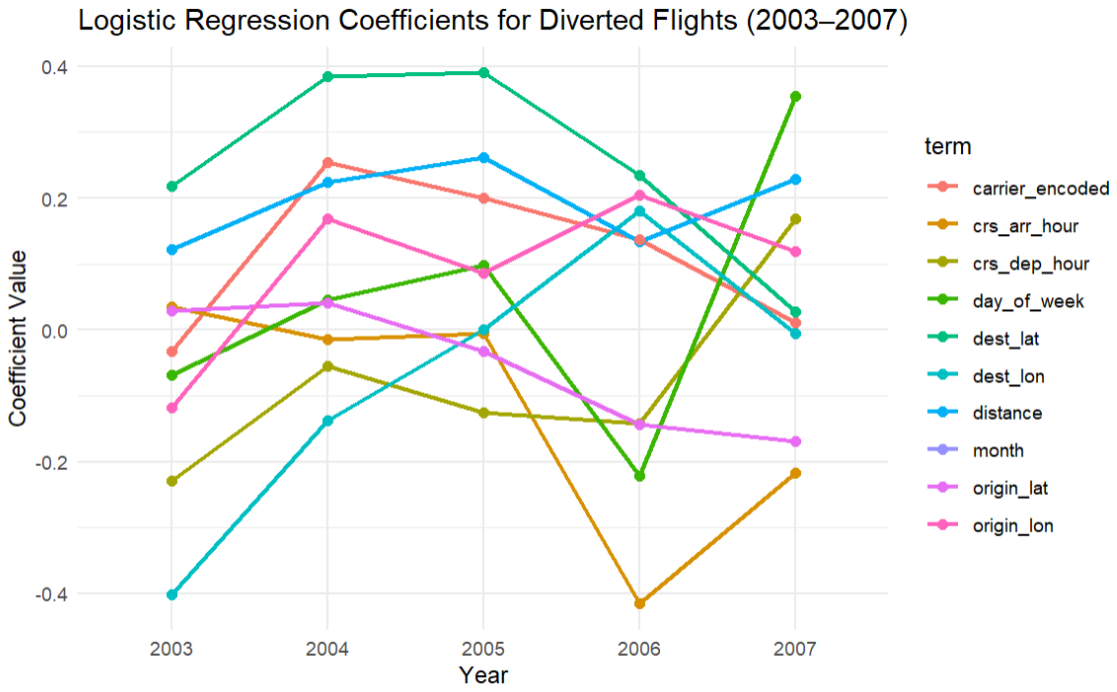

Figure 8: Logistic Regression Coefficients for Diverted Flights (2003-2007) R

6. Conclusion

Through the analysis of US flight data from the year 2003 to 2007, there were many patterns that were observed. Flights scheduled in the early morning hours and in the middle of the week like Tuesdays and Saturdays consistently experienced fewer delays which implies that there is an optimal time for travel. While older aircrafts displayed significantly higher delays, the relationship between aircraft age and delay duration was relatively weak, indicating that age alone is not a strong predictor of punctuality. The logistic regression analysis further revealed that longer distance and late arrival times significantly increase the likelihood of a flight being diverted, whereas flights that departed earlier reduce that risk. All things considered, the models demonstrate that flight logistics, timing and distance are more important in forecasting delays and diversions than airline identity or aircraft age alone. These insights can optimize operations for increased reliability and assist travellers and airlines in making better-informed scheduling decisions.