CATCHING FLIGHTS

A study of the efficiency of commercial flights within the USA.

Prepared by Nur Aisha binte S Burhan

UOL SN: [redacted]

Table Of Contents

1 Introduction

Commercial flights in the US are a common mode of transportation for both domestic and international travel. The US has one of the world's largest aviation industries, with a vast network of airports and airlines. In 2022, the contribution of commercial aviation to the US GDP amounted to $1.25 trillion, which is approximately 5% of the total GDP (A4A, -).

This report revolves around the analysis of the punctuality of commercial flights within the United States. Flight delays can be caused by a variety of factors such as bad weather conditions, technical problems with the aircraft, airport congestion, air traffic control issues, or other unforeseen circumstances. By observing the results from our findings and the trends of the data, we hope to unveil the patterns of delay, which could subsequently assist us in finding, and understanding, the root cause of the problem.

We take data from the 2009 ASA Statistical Computing and Graphics Data Expo, consisting of flight arrival and departure details[1] for all commercial flights on major carriers within the USA (Harvard Dataverse, 2008). This is a large dataset; hence we have strategically chosen data from years 1989-1990, 1994-1996, 2000-2002 and 2006-2007 to use for this report to ensure a reliable and accurate analysis as we compare the data over time. Throughout this report, A delayed flight is defined as a flight that departs or arrives more than 15 minutes later than its scheduled time.

The analysis of this report is done using R programming and Python programming. In simple terms, programming is the process of giving instructions to a computer to perform specific tasks. R and Python are both popular programming languages used for data science, machine learning, and statistical analysis. They are also both open-source languages, which means they are free to use and have large communities of developers contributing to their development.

The next part of this report entails a comprehensive explanation of the observations and findings gathered from the analysis done with the use of both programming languages.

[1] The definition of the different abbreviations in the Airline dataset can be found here: https://www.stat.purdue.edu/~lfindsen/stat350/airline2008_dataset_definition.pdf

2 Findings

2.1 Period with minimal delays

We define optimal by the period with the highest frequency of early and/or punctual flights, i.e., where delays are at minimum.

We search for the optimal month, day and time with the R.

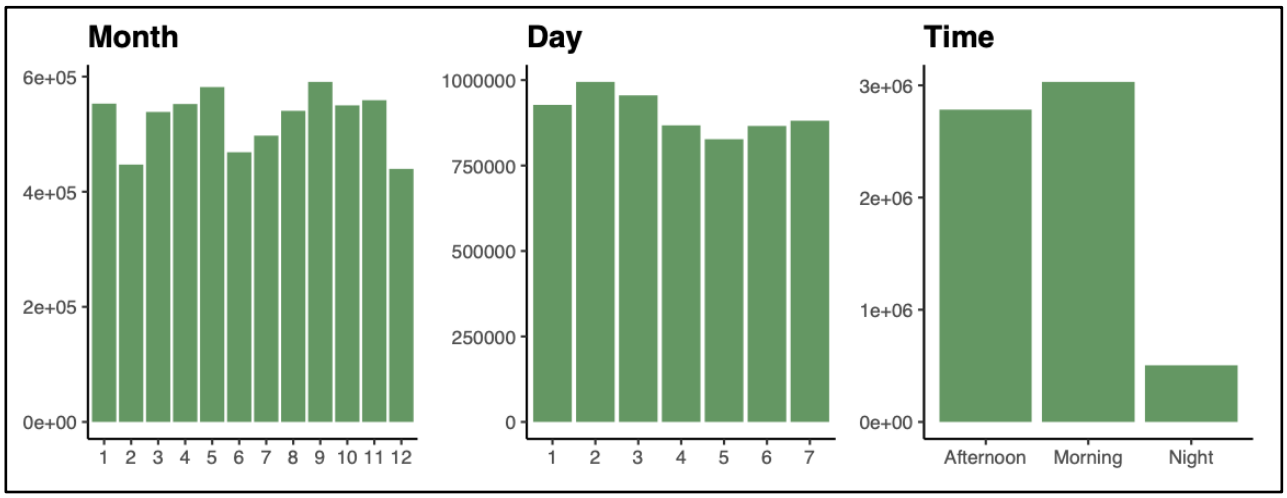

figure 2.1.1: frequency of non-delayed flights across the month, day, and time, respectively

From the figure above, we can observe that the optimal month, day and time is September, Tuesday and Morning flights respectively. The least optimal month, i.e., the month that has the highest number of delayed flights is December. This may be supported by the fact that December is a month of festivities. The least optimal day to fly are on Fridays which are typically popular days for travel as most passengers look forward to weekend getaways. High demand of flights could put pressure on airlines and the air traffic control system, resulting in more delays. The least optimal time are night flights. Higher delays at night may be due to the reduced visibility, especially during inclement weather conditions like fog, rain or snow.

We determine the optimal schedule with the use of Python. The schedule here refers to the month, day and time of flight. Figure 2.1.2 shows the top 3 common optimal flight schedule. The best time to fly to minimize delay is on a Thursday in November, at 6am. The schedules differ in month and day, but the time (6 am flight) remains constant throughout. The airspace is less likely to be crowded in the morning as all the previous day's flights have long landed (Levy, 2008). Furthermore, weather conditions are often more favorable in the morning, as thunderstorms are more common in the afternoon and night (NASA, -).

	Month	DayOfWeek	CRSDepTime	Frequency
0	11	4	600	2199
1	10	2	600	2156
2	5	3	600	2134

Figure 2.1.2: top 3 most common optimal schedule

The findings in figure 2.1.2 can be further supported by figure 1; months 5, 10 and 11 are among the highest frequency of non-delayed flights. The same can be said for the 'Day' figure as well.

2.2 Efficiency of old planes

Does the age of planes contribute to the duration of flight delays? In this section we study the efficiency of old planes. Before we begin, it is crucial to note that we do not have sufficient data to fully analyse the efficiency of older planes as there is no specific column in the available data to determine whether the aircraft is older or newer. Therefore, we proceed in the assumption that the aircrafts are renewed each year.

We examine the trend of the number of delays over time with the use of R programming.

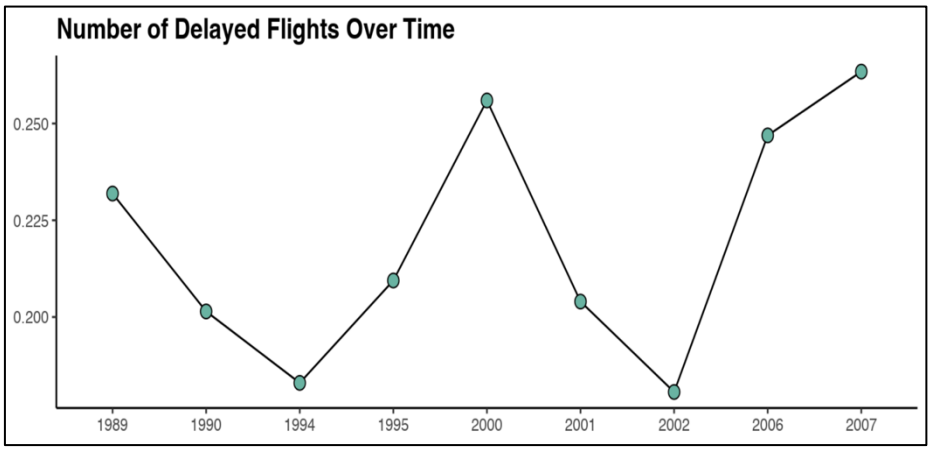

Figure 2.2.1 shows a fluctuation in the number of delayed flights over the years. The fall in flight delays between 1989 and 1994 may be due to the implementation of the Airport Noise and Capacity Act of 1990, introduced to reduce noise levels by mandating airports to develop and implement noise reduction programs (LAWA, -). This law created new ways for airports to plan and manage their operations, and provided funding for improvements that could have been used for the development of new technologies and equipment to increase airport productivity and efficiency. Comparing the periods with high frequency of delayed flights, the proportion ratio increased over the years, i.e., about 23% of the recorded flights in 1989 were delayed as compared to 2007, where close to 30% of its recorded flights were delayed. With this, it can be inferred that older planes do not necessarily suffer more delays.

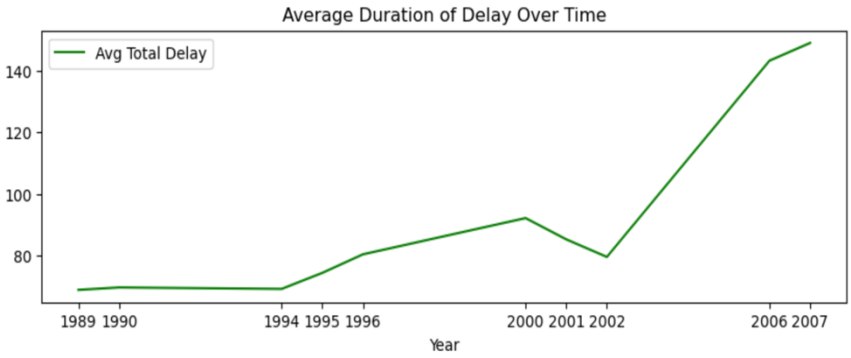

Figure 2.2.2

Using Python, we analyse the efficiency of old planes by comparing the average delay duration over time. Figure 2.2.2 shows an overall increase in the average duration of delay over the years, apart from the period between the years 2000 and 2002. Flights in 2007 has an average delay of roughly 2 hours and 30 minutes, which is significantly longer than the average delay in 1989 of roughly 1 hour. The upward trend in the average duration of delay from 1989 to 2007 suggests that older planes experienced shorter delays as compared to newer ones, thus were more efficient.

A significant increase in volume of air traffic a prominent reason for the rise in delays over time. According to the Bureau of Transportation Statistics (BTS), in 1989, there were approximately 379 million domestic passengers, and by 2007, that number had increased to over 740 million passengers. This represents a more than 95% increase in the number of domestic passengers over that 18-year period (BTS, -).

2.3 Flight destinations

In this section we investigate how the number of people flying between different locations change over time.

In R, we compare the frequency of flight movements across major airports over time. Flight movements of an airport refers to the number of flights that arrives at and departs from the airport. Based on the data we are using for this analysis, each year has a different total number of observations. Thus, for accuracy purposes, we calculate the proportion of the frequency of the flight movements relative to the total observations. We acknowledge that the data frame stores information collected across, roughly, 300 US airports. Therefore, we filtered the data to only include the top 30 airports based on their overall movement. In doing so, we can identify the major airports within the US.

As seen in figure 2.3.1, ATL airport has the overall highest airport movements and that the number of planes arriving at and departing from ATL airport has relatively increased over the years. DFW airport, however, experience a fall in airport movements over the years. This may be due to a change in preference for other airports. For example, the IAH airport, which also has a high overall airport movement, is about 225 miles away from the DFW airport (DFW Airport, -). The non-colored tiles represent the data that were not recorded.

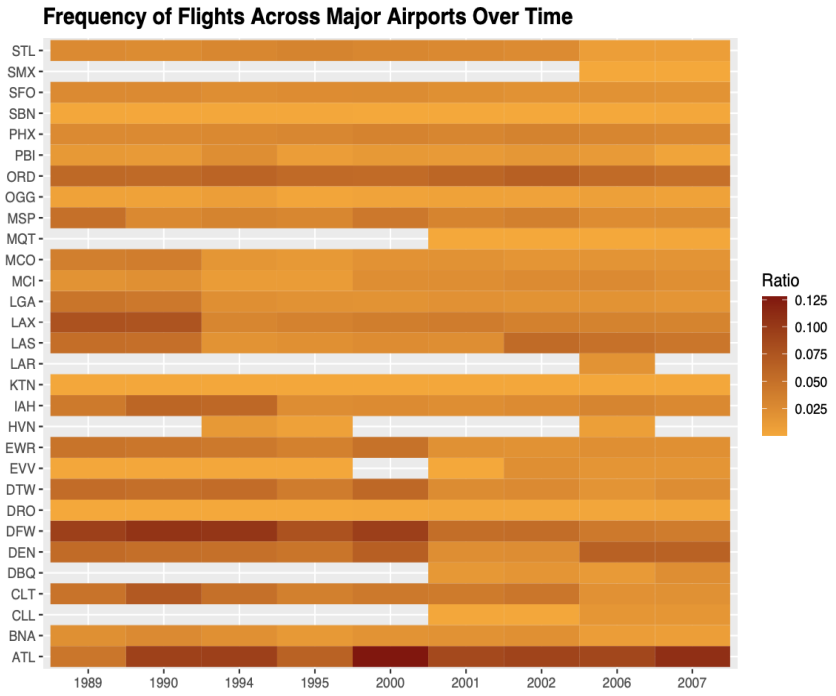

Figure 2.3.1

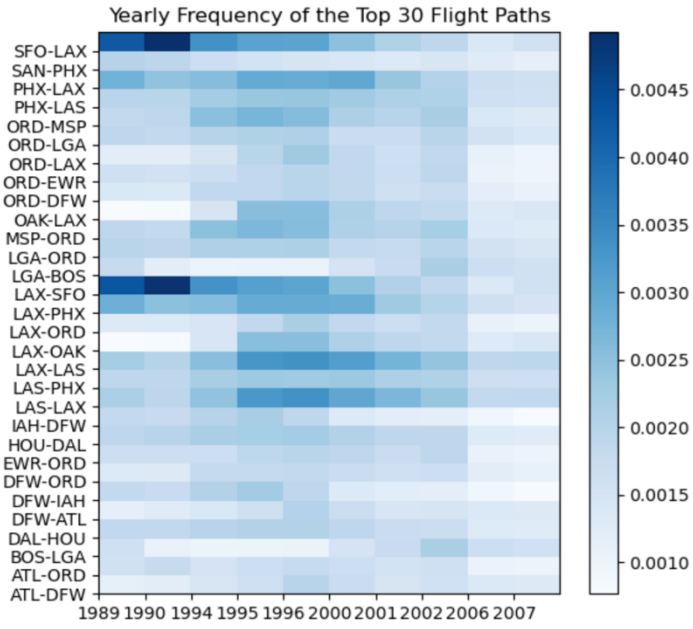

Figure 2.3.2

In Python, we compare the frequency of flight paths over time. There are about 37026 different flight paths recorded in the data. Therefore, we filtered the data to only include the top 30 flight paths based on their total frequency. In doing so, we can identify the popular flight paths within the US. We have also taken the proportion of flight paths relative to the total observations.

Figure 2.3.2 shows that the most popular flights are flights taken to-and-fro the LAX and SFO airport. However, the proportion of flights decreases over the years, which can also be seen across other flight paths. This could have resulted from the introduction of new airports in the region. The BTS stated that there were 5,431 new airports from 1989 to 2007 (BTS, -). Hence, passengers have a wider range of airports to choose from.

2.4 Cascading failures

Cascading failures in the aviation industry can occur when delays in one airport create delays in other airports, resulting in a domino effect that can cause widespread disruption of air travel. In this section, we will demonstrate 2 examples taken from the 2007 data. The first was analyzed in R, the second in Python.

Figure 2.4.1 shows that overall, a slight delay in flight-1355's arrival at SMF airport led to a more than proportionate delay as it landed at GEG airport. The cascading failure began when the flight experienced a delay on its first layover at SMF airport. The plane departed 15 minutes later than its estimated time of departure to PDX airport. Flight-1355 then arrived around the time it was supposed to depart for GEG airport. The layover in PDX airport, which was initially estimated to take about 15 minutes, took about 40 minutes instead. This eventually caused the flight to land in GEG airport about 40 minutes after its estimated time of arrival.

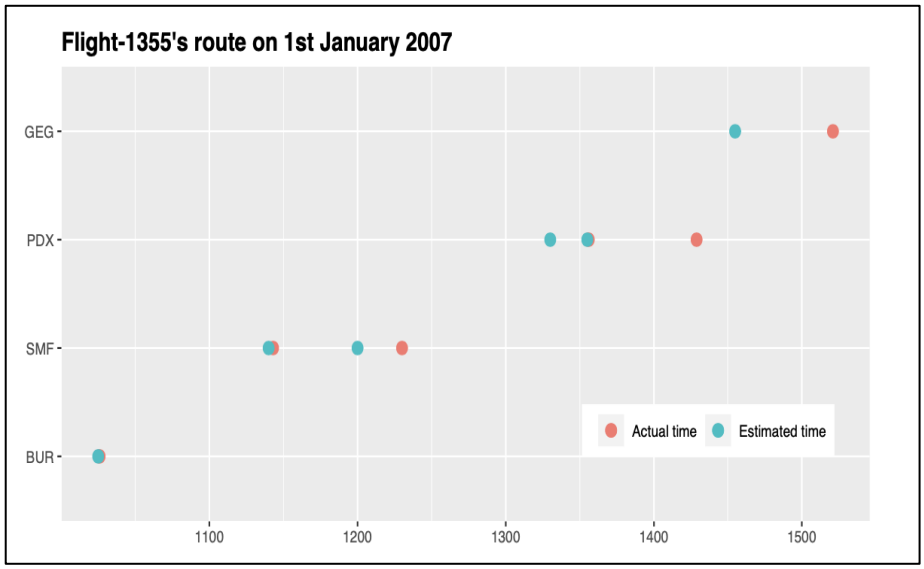

Figure 2.4.1

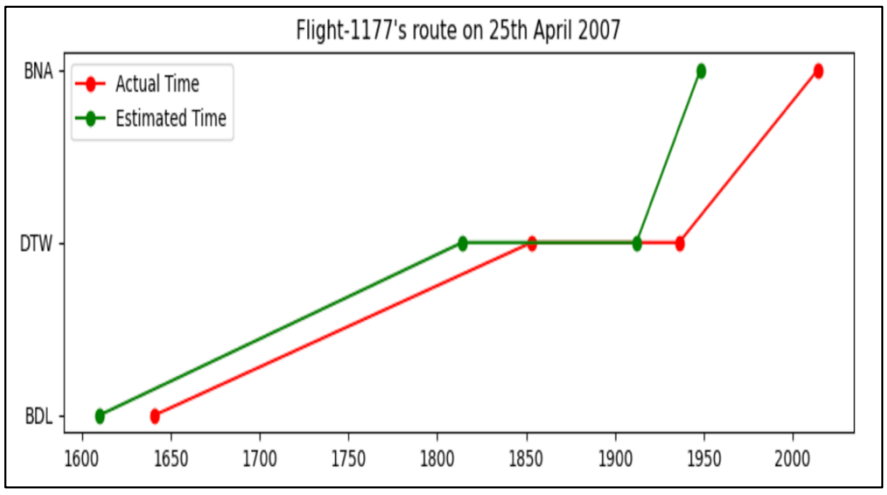

Figure 2.4.2

We move onto the second example. Figure 2.4.2 shows that flight-1177 departed BDL airport 40 minutes after the estimated time to depart. The plane landed at DTW airport roughly 40 minutes after it was estimated to land. The estimated and actual duration for the layover in DTW airport took about the same time, which is about an hour. This eventually caused the flight to land in BNA airport about an hour after its estimated time of arrival. Therefore, it can be observed that a delay in flight-1177's departure from BDL airport led to a more than proportionate increase in delay as it landed at BNA airport, just like in the first example. The cascading failure in this example, began when the flight experienced a delay on when it first departed.

The BTS reported that the most common cause of flight delays in 2007 was late-arriving aircrafts which accounted for 32% of all delays. This suggests that cascading failures are the biggest contributors to flight delays. The second most common cause stem from air traffic control issues (BTS, -).

3 Conclusion

To summarize, our analysis of commercial flight efficiency in the US has yielded noteworthy insights. On average, around 80% of flights within the country arrive on schedule. We have identified various factors, including the manufactured year of plane, airline, and flight schedule, such as the month and day, that can influence flight efficiency. We have also discovered patterns and trends between airports and how the frequency of their flight paths varies.

These discoveries were made possible with the use of R and Python programming softwares. By utilizing their capabilities, we were able to efficiently and accurately process and analyse extensive sets of data, ultimately leading to a better understanding of the trends and patterns related to flight delays. Our findings highlight the potential of R and Python as powerful tools for data analysis and emphasize their role in informed decision-making and the optimization of the aviation industry's efficiency. The various libraries and packages (composed of compiled code, data, and R functions) has helped us import and manipulate data to our desired output. The data visualization, formed through functions stored in R's ggplot2 library, allowed us to create informative graphs to communicate our analysis. Similarly, Python has been useful for data analysis, which involved importing, cleaning, and manipulating datasets using valuable libraries like Pandas and NumPy. Visualizations and charts achieved by using libraries like matplotlib and seaborn, helped us to better understand our findings as well.

To conclude, the study of commercial flight efficiency in the US with the application of R and Python programming underlines the significance of data analysis and informed decision-making in the aviation sector. It is hoped that with these findings, airports and airlines can foster a collaborative effort to enhance passengers' flying experience, mitigate financial losses attributed to delays, and boost the overall effectiveness of the US commercial aviation industry.

4 References

(A4A, -), “Economic Impact Of Commercial Aviation”: https://www.airlines.org/impact/#:~:text=Commercial%20aviation%20drives%205%25%20of,from%20more%20than%20220%20countries.

(Harvard Dataverse, 2008), “Data Expo 2009: Airline on time data”: https://doi.org/10.7910/DVN/HG7NV7

(Levy, 2008), “Best And Worst Times To Fly”: https://www.forbes.com/2008/10/14/travel-airports-time-forbeslife-cx_fl_1014travel.html?sh=1b8fefa41cda

(NASA, -), “Is there a specific time of day that a thunderstorm is most likely to occur?”: https://gpm.nasa.gov/resources/faq/there-specific-time-day-thunderstorm-most-likely-occur#:~:text=A%20thunderstorm%20is%20formed%20when,usually%20the%20afternoon%20or%20evening.

(BTS, -), “Passengers”: https://www.transtats.bts.gov/Data_Elements.aspx?Data=1

(LAWA, -), “The Airport Noise and Capacity Act Of 1990 (“Anca”)”: https://www.lawa.org/-/media/lawa-web/noise-management/files/airport_noise_and_capacity_act_of_1990.ashx

(DFW Airport, -), “Nearby Airports to Dallas Fort Worth”: https://www.ifly.com/dallas-fort-worth-international-airport/closest-airports

(BTS, -), “Number of U.S. Airports”: https://www.bts.gov/content/number-us-airportsa

(BTS, -), “Airline On-Time Statistics and Delay Causes”: https://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp?pn=1