Table Of Contents

Introduction

Description

The air transportation network, like other vital infrastructures, has a huge impact on local, national, and international economies. Every year, about 700 million passengers fly, putting the air transportation industry on the verge of collapse. The failures and inefficiencies of the air transportation system have large economic costs; flight delays cost European countries 150 billion to 200 billion Euro in 1999 alone[1].

This report examines the correlation between departure delays and cascading failures in airports, the causes of flight delays, as well as methods to avoid them. To predict future departure delays, a model will be constructed. The following questions from the Programming for Data Science (ST2195) coursework will be answered at the relevant segments:

- What is the best time of day / day of the week / time of year to fly to minimize delays?
- Do older planes suffer more delays?
- How does the number of people flying between different locations change over time?
- Can you detect cascading failures as delays in one airport create delays in others?
- Use the available variables to construct a model that predicts delays.

Data

The data in this report is obtained from the Harvard Dataverse, an open-source repository for research data, and it spans the years 2005, 2006, and 2007. Between years 2005 and 2007, 21,735,733 flights departed from the target airport, among which 416,412 flights (1.92%) were cancelled, 21,996 flights (0.1%) flights were diverted, and 8,508,330 flights (39.14%) were delayed, with the most severe delay being 2,601 minutes.

Data Processing

A portion of the data in this report has been omitted or unutilized for the following reasons:

- Lack of information
- Inexplicable records

How the data in this report have been imputed:

- Logical data cleansing
- Imputing a small fraction of missing values with mean

1 Departure Delay

Often, seasons with a higher volume of flights are more likely to experience flight delays, which could cause a great deal of inconvenience. In this segment, delay rates across different periods and times of day will be analyzed to provide recommendations on when to fly to minimize delays.

1.1 Analysis on Departure Delays

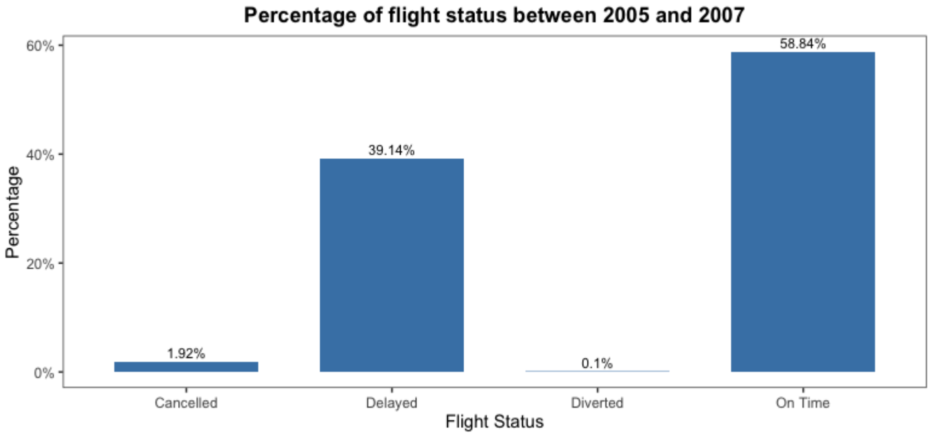


Figure 1.1

The percentages of each flight status were derived by categorizing continuous variables and dividing the sum of each category by the total number of flights between 2005 and 2007.

Based on Figure 1, the percentage of delayed flights outweighs the percentages of cancelled and diverted flight combined substantially. This percentage is also more than half of the on-time flight percentage.

Departure delays can be caused by several factors, including:

- Carrier delay
- Weather delay
- Security delay
- Late aircraft delay
- National Airspace System (NAS) delay

Notwithstanding these factors, the month, day, and hour of a departing flight can have a significant impact on the departure delay rates.

1.2 Minimizing delays

To avoid flight delays, one can fly during off-peak hours and seasons, or when the departure delay rate is at its lowest.

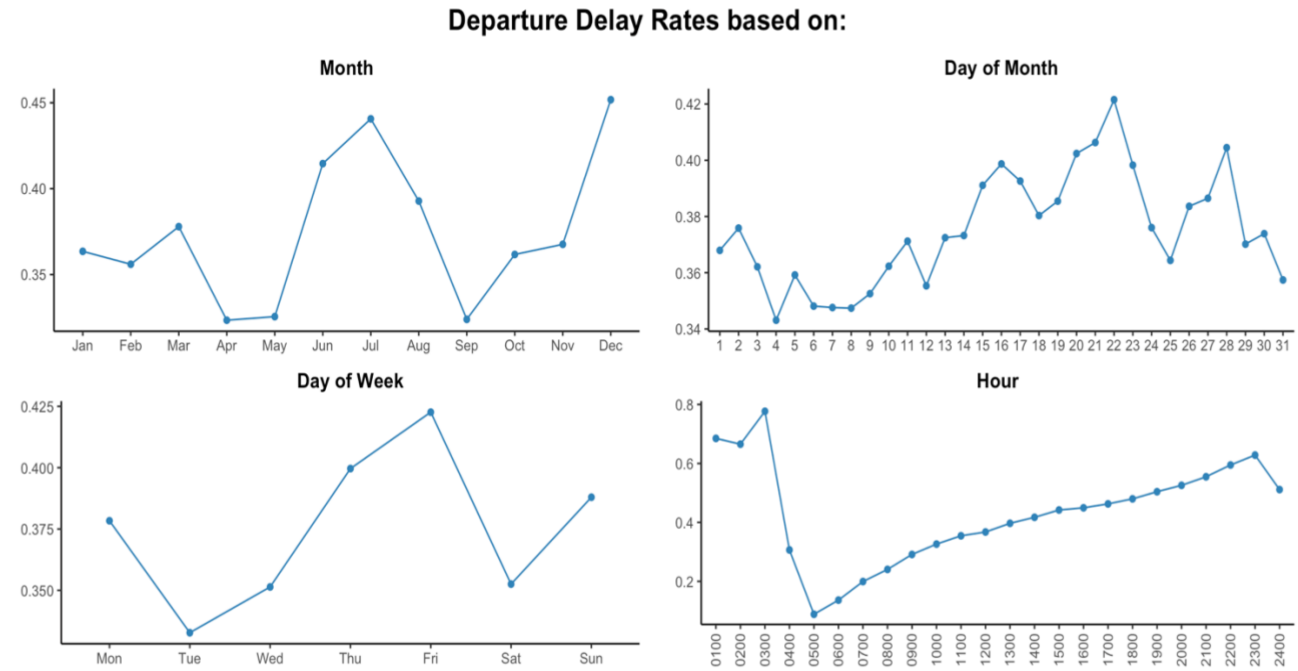


Figure 1.2

The departure delay rates were obtained by categorizing the attributes of each variable and dividing the sum of delayed flights by the total number of flights in each attribute. 0.33% of the data was omitted from the variable ‘Hour’ due to inexplicable records.

Figure 1.2 shows that the lowest departure delay rates are in April, May, and September, while the highest delay rates are in July and December. The 4th, 7th and 8th of each month have significantly lower departure delay rates as compared to the 21st, 22nd and 28th. Flying on Tuesdays and Wednesdays is preferred over Thursdays and Fridays. The best time to fly is between 05:00 and 07:00 in the morning with the worst being between 0100 and 0300.

1.3 Recommendations to minimize delays

Unquestionably, the optimum time to fly would be when the departure delay rate is at its lowest. Hence, the best time of day would be 05:00 a.m., with a delay rate of 0.088, day of the week would be Tuesdays with a delay rate of 0.333 and the best time of the year is April 4th, with delay rates of 0.343 and 0.323 respectively.

Conclusion based on findings: 4th April, or Tuesdays in April at 05:00 a.m.

2 Relationship between older planes and delays

The aircraft engine system will inevitably fail as planes age, necessitating more periodic checks and maintenance on older planes. Is this, however, implying that older planes experience more delays? In this segment, the association between the age of planes and departure delays will be explored.

2.1 Association between plane’s age and mean delay

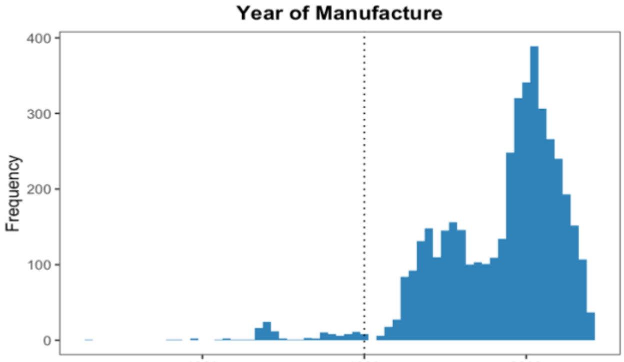


Figure 2.1

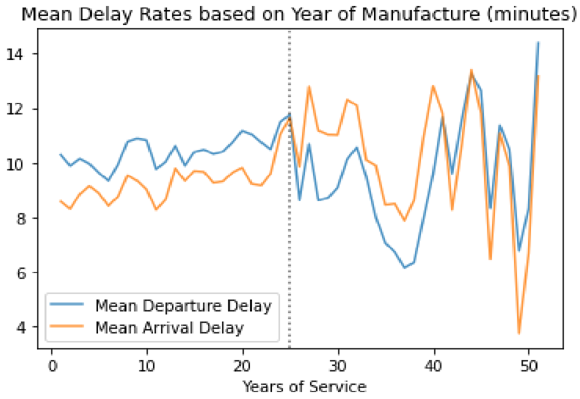


Figure 2.2

Figure 2.1 shows a histogram of all planes' year of manufacture, with most planes manufactured between 1999 and 2001. The oldest plane was built in 1946, and as the histogram shows, the proportion of planes built before 1980 is rather small. The age of planes has been segregated into 2 groups: below and over 25 years, with below 25 years representing younger planes and over 25 years representing older planes.

The mean values for each delay were derived by the years of service to assess the mean departure and arrival delays based on the years of service of a plane. These values are depicted in a line chart shown in Figure 2.2.

Younger planes have a higher mean departure delay than mean arrival delay, with both mean delay values being closely related, whereas older planes have a higher mean arrival delay than mean departure delay, with both mean delay values fluctuating over the years.

2.2 Association between plane’s age and delay factors

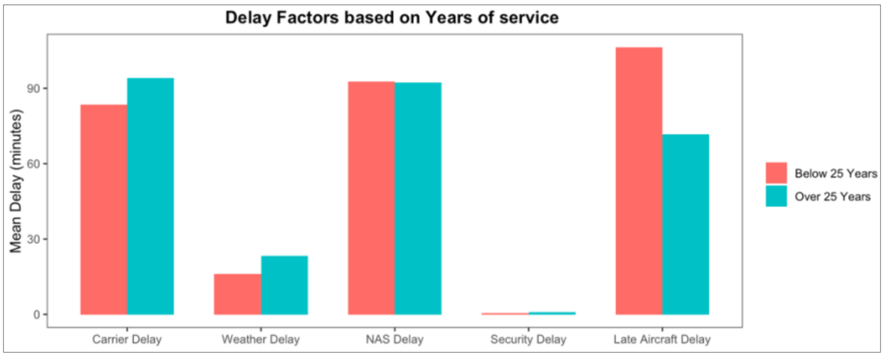


Figure 2.3

The planes' ages have been divided into two groups: below 25 years and over 25 years, as their mean delays have changed drastically after the age of 25. The mean values for each delay factor in each group are tabulated in Figure 2.3, which presents a bar chart that has been grouped by age and delay factors. Both age groups have similar delay factors, with Carrier Delay, NAS Delay, and Late Aircraft Delay being the top 3 delay factors.

Mean Delay based on the Top 3 Delay Factors

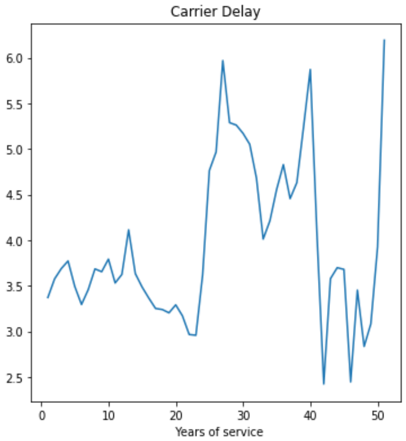


Figure 2.4

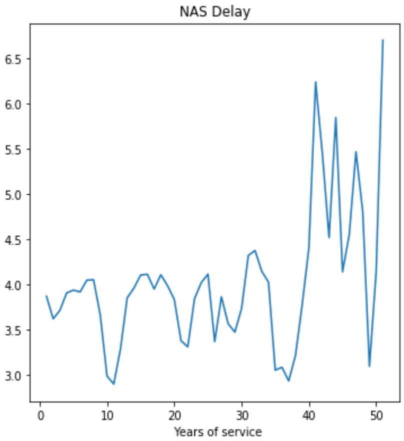


Figure 2.5

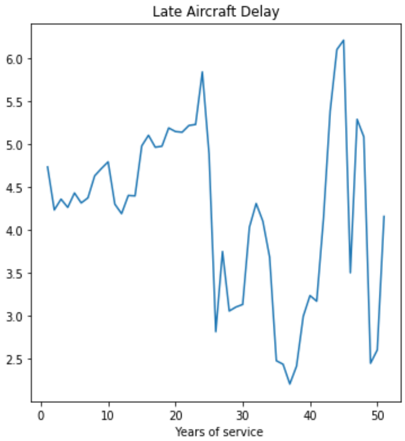


Figure 2.6

When looking at the top three delay factors across all ages, the mean delay for Carrier Delay initially peaks at 27 years old with a mean delay of 5.95, and then peaks again at 39 and 50 years old with mean delays of 4.76 and 4.40 respectively.

NAS Delay varies from 0 to 40 years old, with mean delays of 2.5 to 4.5 minutes, and peaks at 40 years old, with a mean delay of 6.91 minutes, and then peaks twice more at 44 and 47 years old, with mean delays of 4.52 and 5.46 minutes, respectively.

Lastly, Late Aircraft Delay rises and peaks at 24 years with a mean delay of 4.82 minutes, then rises again at 44 and 47 years with mean delays of 4.82 and 5.29 minutes.

2.3 Findings

In Figure 2.2, even though the mean departure and arrival delays are increasing over the years, the peaks for the mean of both delays only occur after 25 years. Almost all the peaks of mean delays in the top three delay factors, like the mean departure and arrival delays, only occur after 25 years.

Even though the mean delays for younger planes (below 25 years) are high, it does not see as many peaks as the older planes (over 25 years). Therefore, older planes do suffer more delays.

3 Air traffic

The continued global expansion in air travel is attributed to several key factors, two of which are the surge in low-cost carriers and increased infrastructure spending. As the number of low-cost carriers grew, consumers' access to economical air travel increased. In addition, airport infrastructure spending has increased in the Asia Pacific region, resulting in increased global carrying capacity. An analysis on air traffic will be conducted in this section.

3.1 Popular & Rising Origin-Destination combinations

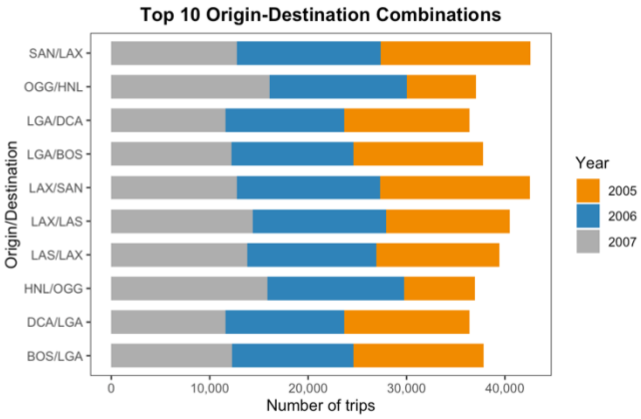


Figure 3.1

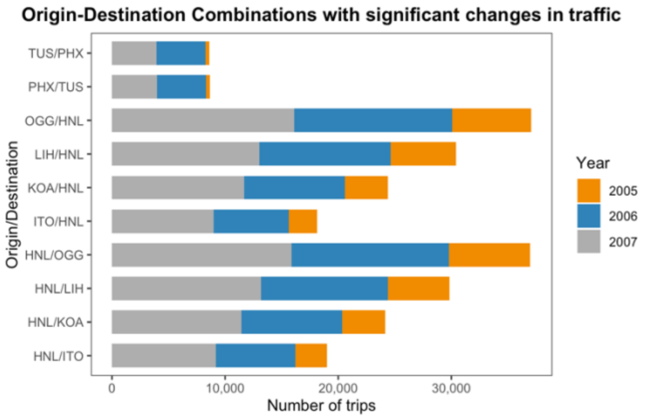


Figure 3.2

The top 10 Origin-Destination combinations were derived by comparing the total number of flights for all combinations. Figure 3.1 documents the air traffic for the top 10 pairings between 2005 and 2007. With 42,610 flights over three years, San Diego International-Lindbergh (SAN) and Los Angeles International (LAX) are the busiest airports, followed by Los Angeles International (LAX) and McCarran International (LAS) with 40,468 flights.

The change in air traffic is computed in Figure 3.2 by computing the maximum difference of all combinations, and the top 10 combinations with significant increases are indicated. With 7,183 flights in 2005 and 15,876 flights in 2007, Honolulu International (HNL) and Kahului (OGG) airports have seen more than a twofold increase, whereas Phoenix Sky Harbor International (PHX) and Tucson International (TUS) airports have seen a tenfold increase, with 365 flights in 2005 and 3,998 flights in 2007.

3.2 Relationship between traffic and airports

The top 10 origin-destination combinations' traffic remained relatively consistent from 2005 and 2007, with SAN, LAX, and LAS being the busiest airports. Meanwhile, traffic increased significantly in origin-destination combinations such as HNL, OGG, PHX, and TUS.

4 Cascading failure due to flight delays

Air traffic is an essential part of human mobility and global trade, but now it is outstripping the capacity and becoming frequently congested. Passenger airline flight delays and cancellations are prevalent and have socio-economic and environmental consequences[2,3,4]. Cascading complications may arise at each flight's subsequent destinations if the airline has more scheduled flights than it could handle and it usually occurs when a flight departs and arrives later than the scheduled times, causing delays in the subsequent scheduled flights.

4.1 Analysis on flight delays & delay rates

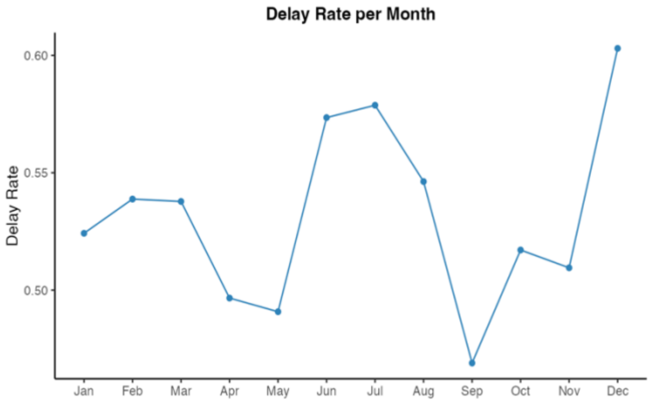


Figure 4.1

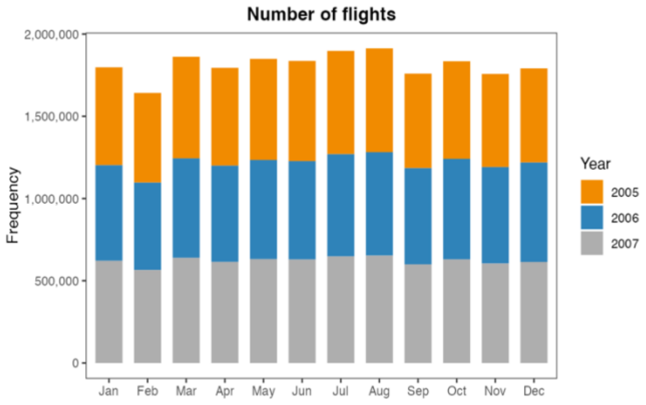


Figure 4.2

Figure 4.1 differs slightly from Figure 1.2's upper left graph as it factors in arrival delay as well, whereas Figure 1.2's graph just accounts for departure delay. The delay rates were calculated by categorizing both departure and arrival delays in the same attribute of each variable and dividing the total number of delayed flights by the total number of flights in that attribute. September has the lowest rate of delays (0.469), while December has the greatest rate of delays (0.603).

Figure 4.2 simply adds up the number of flights for each month, divides them into years, and displays them in a grouped bar chart. With 1,897,765 and 1,912,915 flights respectively, July and August have the most flights.

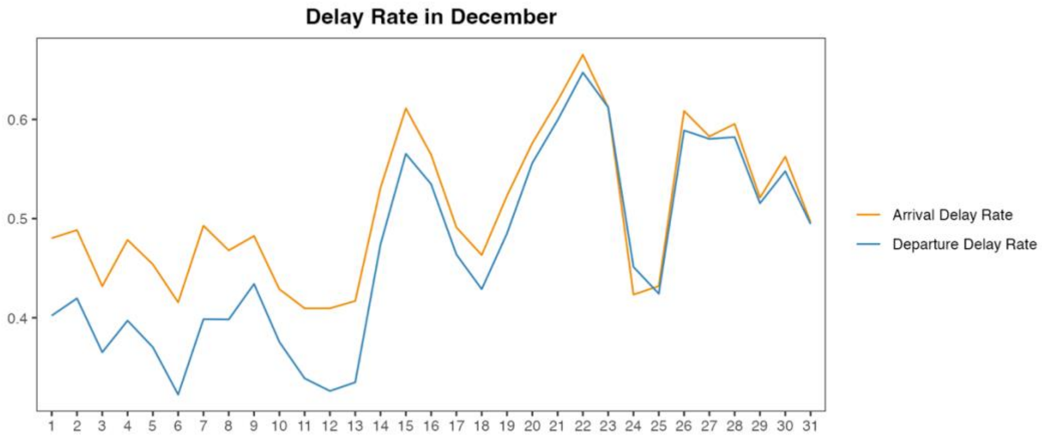


Figure 4.3

In this part, the month with the highest delay rates will be utilized to detect cascading failures as a result of aircraft delays. The month of December will be the focus of this analysis, as illustrated in Figure 4.3.

Both delay rates were computed by categorizing the days in December for departure and arrival delays and dividing the total number of delayed flights by the total number of flights in each variable. December 22nd has the highest departure and arrival delay rates of 0.647 and 0.665, while December 6th has the lowest delay rates of 0.323 and 0.416. The rates of departure and arrival delays have a similar trend, therefore they could be correlated.

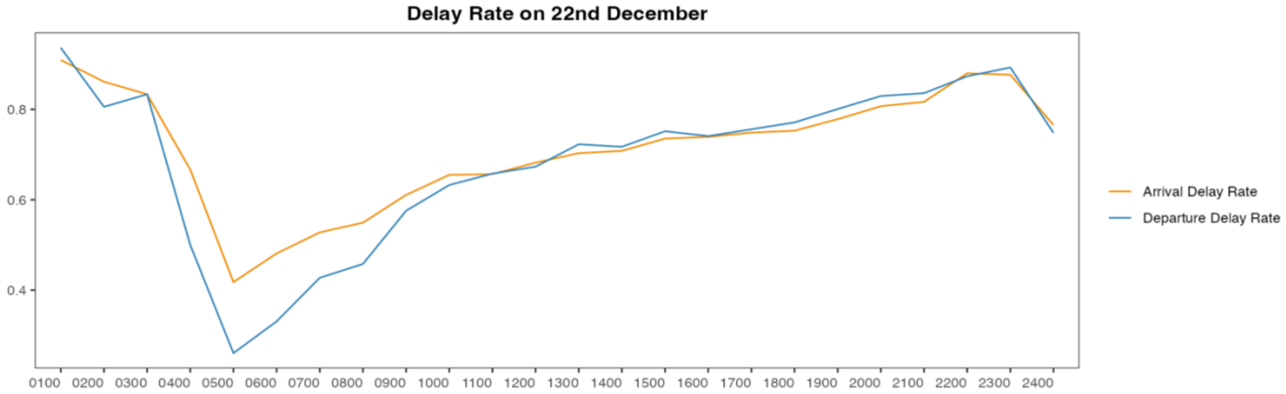
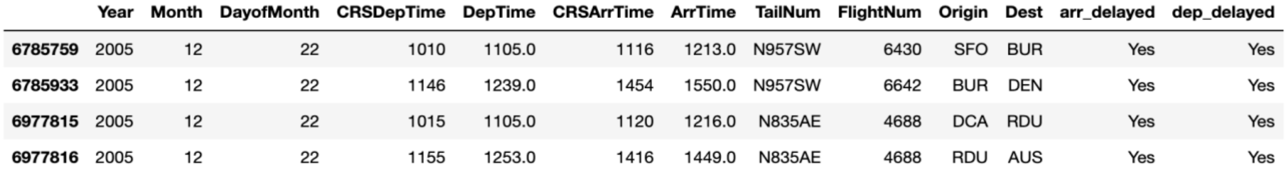


Figure 4.4

The 22nd of December will be included in this analysis since it has the highest rate of delays during the month of December. A small percentage of the departure timings on December 22nd went unutilized due to incomprehensible data, and the ensuing graph is displayed in Figure 4.4.

For cascading failures to occur, a flight's arrival time must be later than its anticipated arrival time, causing the next flight's departure to be delayed. A second degree delay occurs when the arrival time of the second flight is later than the scheduled time due to the first flight's arrival delay, which causes the departure of the third flight to be delayed. With that, cascading failures in flight delays are more likely to occur when the departure delay rates surpasses the arrival delay rates.

4.2 Identifying cascading failures



	Year	Month	DayofMonth	CRSDepTime	DepTime	CRSArrTime	ArrTime	TailNum	FlightNum	Origin	Dest	arr_delayed	dep_delayed
6785759	2005	12	22	1010	1105.0	1116	1213.0	N957SW	6430	SFO	BUR	Yes	Yes
6785933	2005	12	22	1146	1239.0	1454	1550.0	N957SW	6642	BUR	DEN	Yes	Yes
6977815	2005	12	22	1015	1105.0	1120	1216.0	N835AE	4688	DCA	RDU	Yes	Yes
6977816	2005	12	22	1155	1253.0	1416	1449.0	N835AE	4688	RDU	AUS	Yes	Yes

Figure 4.5

The records were filtered down to December 22nd between the hours of 1100 and 1300 to find cascading failures, as this is when departure delay rates exceed arrival delay rates. As illustrated in Figure 4.5, two occurrences of first degree delays have been detected, with tail numbers of N957SW and N835AE.

5 Predicting delays

As the air travels have a significant role in economy of agencies and airports, it is necessary for them to increase quality of their services. One of the important modern life challenges of airports and airline agencies is flight delay[5].

5.1 Exploratory data analysis (EDA)

The model will take into account a total of 8 factors to predict departure delays.

- Factors 1 and 2: Scheduled and actual departure times

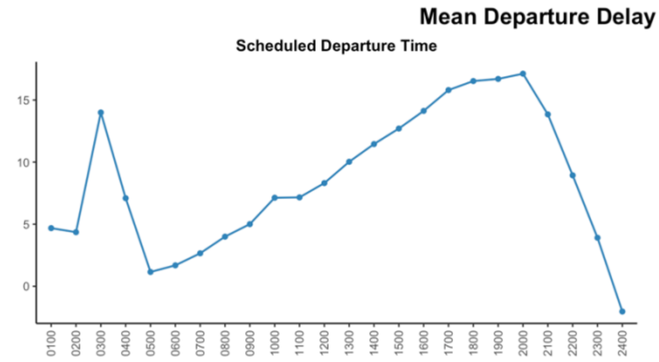


Figure 5.1

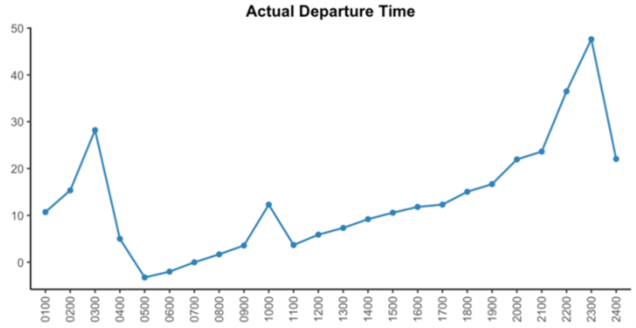


Figure 5.2

After some logical data cleansing, the single and double digits in both variables signify 00:00 or 24:00 minutes, while values over 2400 imply departure times for the next day. For example, 2945 signifies 05:45 the next day.

- Factor 3: Age of planes

Based on the findings derived in segment 2.1, it is clear that a plane's age can affect delays. For any null values present, it will be imputed with the median value.

- Factors 4 and 5: Arrival delay (mins) and IATA, or airports

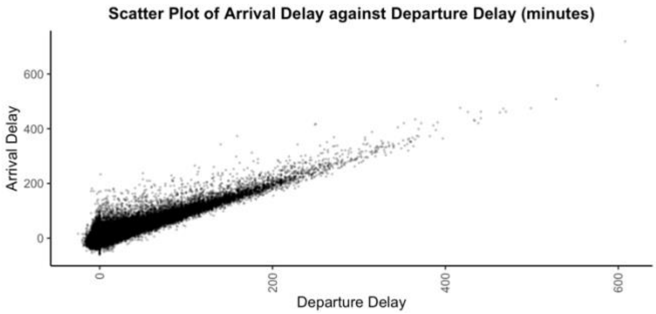


Figure 5.3

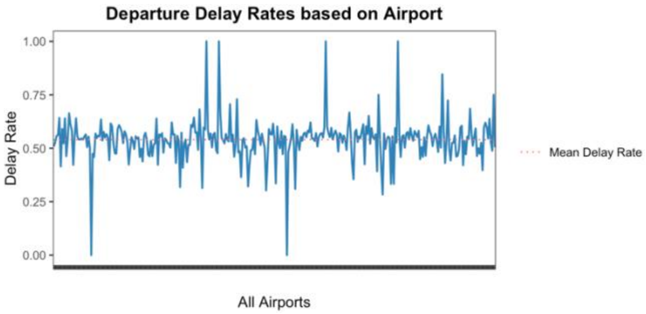


Figure 5.3

Since both delays have a positive linear correlation, arrival delay should be incorporated in the model. IATA will be factored into the model as well, as the delay rates vary per airport.

- Factors 6,7 and 8: Carrier delay, NAS delay and late aircraft delay

The top three delay variables appeared to have a higher level of importance than the other two delay factors, based on the data in section 2.1. Hence, this will be accounted for in the model.

5.2 Ridge regression, tuning hyperparameters & random forests

The mean squared error (MSE) for the first linear regression model was 69.24. For the second model, the ridge regression model computed an MSE of 69.24 after tuning of the hyperparameters, and an MSE of 69.23 before. Finally, for random forests, the model returned an MSE of 31.17. The model with the lowest MSE value, random forests, fared the best of the models.

5.3 Benchmarking

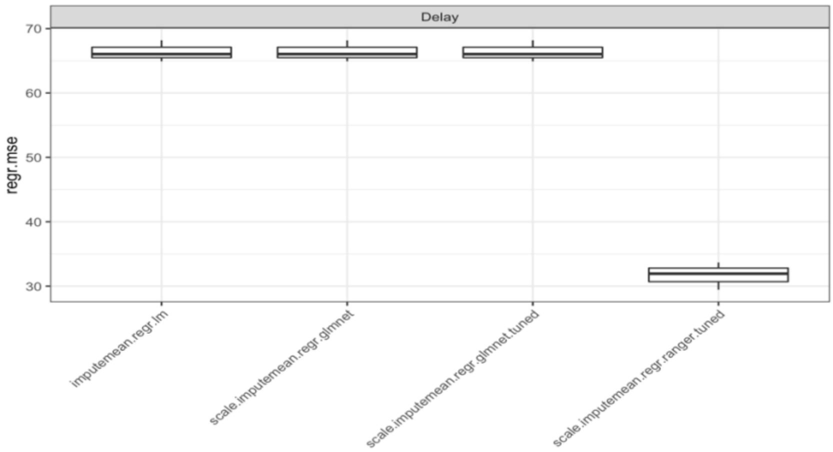


Figure 5.4

The four machine learning algorithms provided will be assessed and contrasted in terms of their ability to learn patterns in the 'benchmark' datasets used for benchmarking. The comparisons will be visualized with a box plot in Figure 5.4, and three-fold cross-validation will be employed. In the Delay dataset, the results show that random forests clearly outperform regression methods.

Conclusion

According to the conclusions of this analysis, older planes experience higher delays, and cascading failures in flights are seen when a flight's arrival time coincides with the departure time of the next flight. In addition, the optimum time to fly is around 05:00 on Tuesday mornings in April, when the delay rate is at its lowest. In terms of predicting departure delay, the random forests model showed clear dominance over the other regression models.

References

[1] European Organisation for the Safety of Air Navigation (2000) *Cost of Air Transportation Delay in Europe*, Technical Report (European Organisation for the Safety of Air Navigation, Brussels).

[2] Chen, Z., Wanke, P., Antunes, J. J. M. & Zhang, N. Chinese airline efficiency under co2 emissions and flight delays: a Stochastic network dea model. *Energy Econ.* **68**, 89–108 (2017).

[3] Song, C., Guo, J. & Zhuang, J. Analyzing passengers emotions following flight delays-a 2011–2019 case study on SKYTRAX comments. *J. Air Transp. Manag.* **89**, 101903 (2020).

[4] Gopalakrishnan, K. & Balakrishnan, H. Control and optimization of air traffic networks. *Annu. Rev. Control Robot. Auton. Syst.* **4**, 397–424 (2021).

[5] Yazdi, M.F., Kamel, S.R., Chabok, S.J.M. *et al.* Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *J Big Data* **7,** 106 (2020).