EXPLORING TWO DECADES OF FLIGHT DELAYS IN THE USA OF FACTORS, FAILURES, AND PREDICTIVE MODELING

200579973

Table of Content

1. Introduction

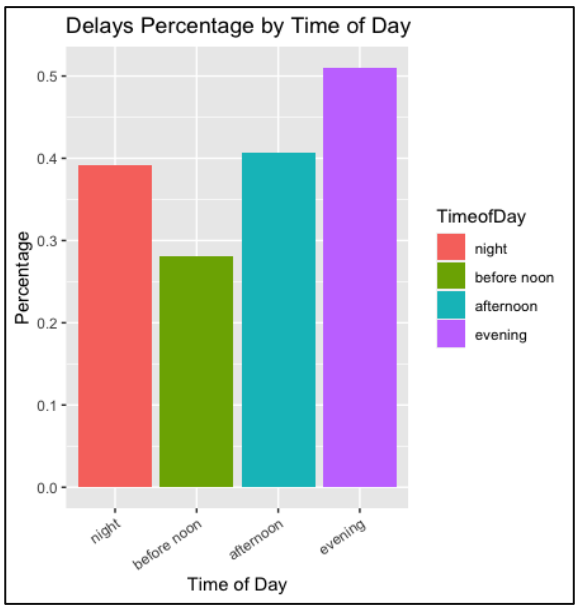
This huge dataset consist flight arrival and departure data for significant carriers in the United States from 1987 to 2008, however, we will concentrate for further analysis in the year 2004 and 2005. By narrowing down, we can acquire a more focused understanding of flight patterns, delays, and other factors that might have an impact on the airline sector during that time period by limiting the information to two years.

The examination of this dataset is crucial since flight delays and cancellations can have a major negative impact on both airlines' bottom lines and their customers' quality of life. Furthermore, being aware of the causes of these disruptions and cancellations might point out areas for improvement in air traffic control and airline operations. Thus, it can assist in informing policy development and decision-making efforts targeted at lowering the incidence and effects of flight delays and cancellations.

2. R Analysis

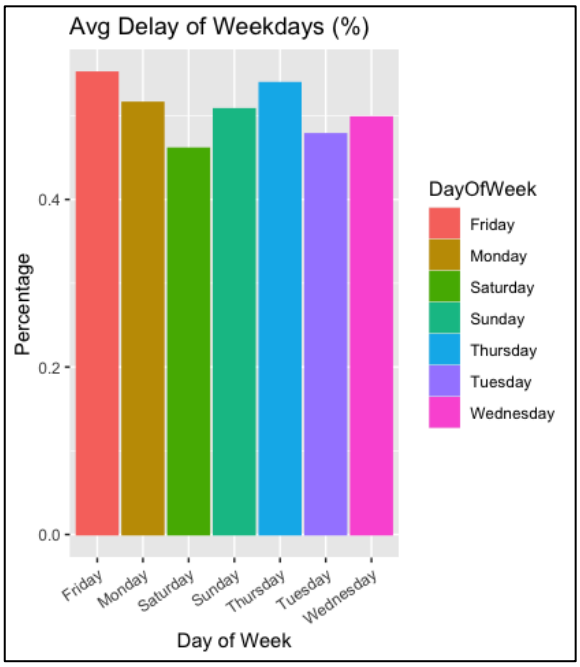
2.1 Question 1. When is the best time of day, day of the week, and time of year to fly to minimise delays?

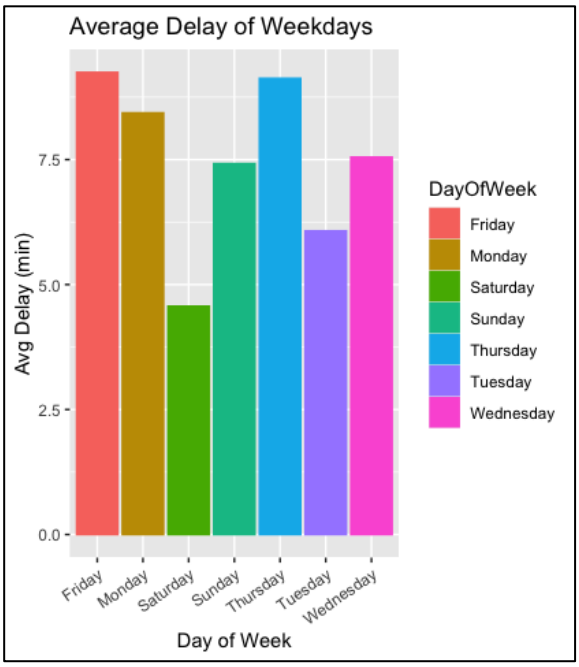
2.1.1 When is the best time of day to fly to minimise delays?



The data is split into four categories which are night, before noon, afternoon, and evening. For each time of day category, the overall arrival and departure delays are determined. 'Before noon' category has the smallest delay percentage since it has the fewest delays overall, which results in a lower denominator when computing the delay percentage with 0.280% followed by night with 0.392%. This could be due to less flights during this time scheduled to arrive or depart or less congestion early in the morning.

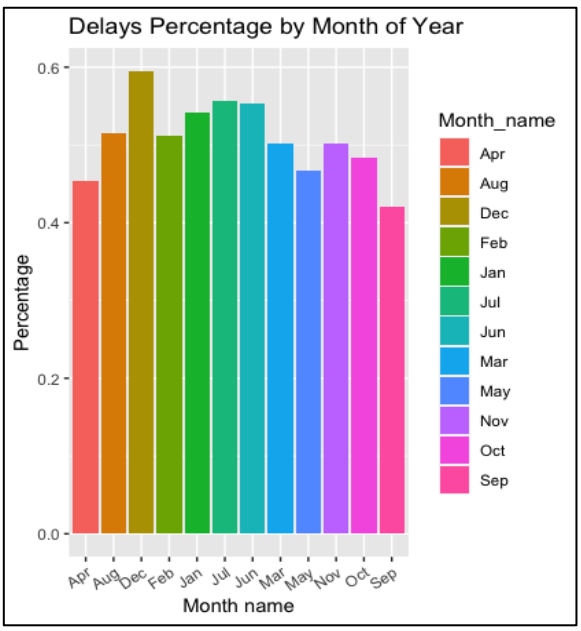
2.1.2 When is the best day of the week to fly to minimise delays?





The graphic display the delay for each day of the week (in minutes and percentage). According to the plot, flights on Saturdays experience the least amount of delays on average, followed by flights on Tuesday and Wednesday. In light of this data, it would seem that Saturdays are the best day of the week to fly with lowest delay of 4.57 minutes following 0.461 percentage. One of possible reasons that may be influenced is there are fewer business travellers flying on Saturdays. Due to their strict schedules and potential lack of flexibility, business travellers often experience more delays during the week, when airports are busier or fewer connecting flights.

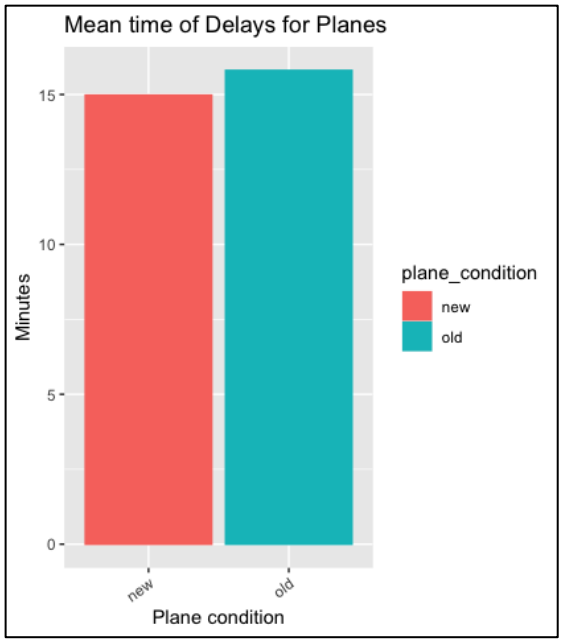
2.1.3 When is the best time of year to fly to minimise delays?

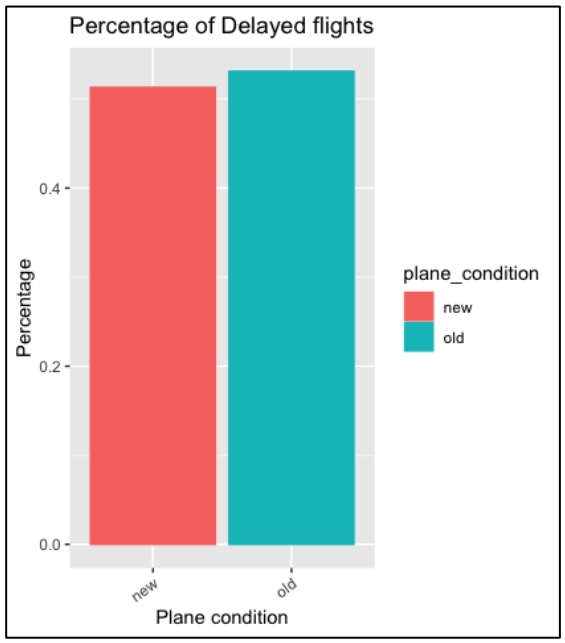


Based on the bar graph, September has the least delay among the 12 months in a year with 0.420% followed by April with 0.454%. In addition, I have also made analysis based on season in a year whereby it shows that autumn which is in September has the least average delay with 29.537 in comparison to other seasons such as spring which has average delay of 32.959.

This could be due to less travellers where September is not a peak season of travelling like summer or winter holidays, hence, there are less travellers, furthermore, less people means that airports and airlines can run their business more effectively and without delays.

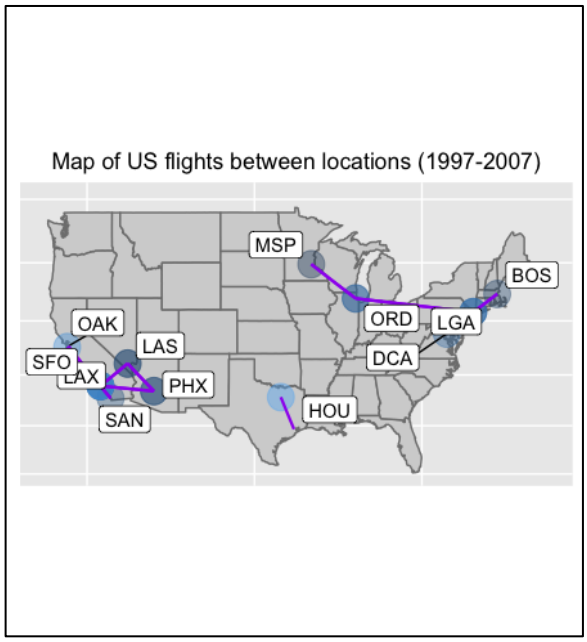
2.2 Question 2. Do older planes suffer more delays?

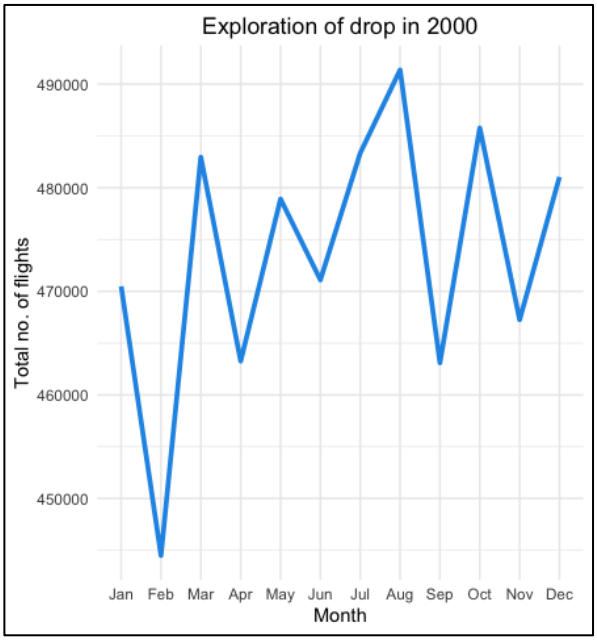




After joining two datasets, 'data_airplanes' and 'airplanes' and create a column named 'plane_condition' which classifies airplanes as new or old based on the year they were made. Using this analysis, we can compare the mean delay between new and old plane which is shown on the left side that new plane has average of 15 minutes delay whereas old has 15.8 minutes. Moreover on the right side, it shows the percentage for new and old plane condition where old one has higher percentage with 0.531% followed by new with 0.513%. However, as the difference is not significant, hence, we can't conclude that older planes suffer more delays than new planes and there can be additional factors that cause delays beside age of planes.

2.3 Question 3. How does the number of people flying between different locations change over time?



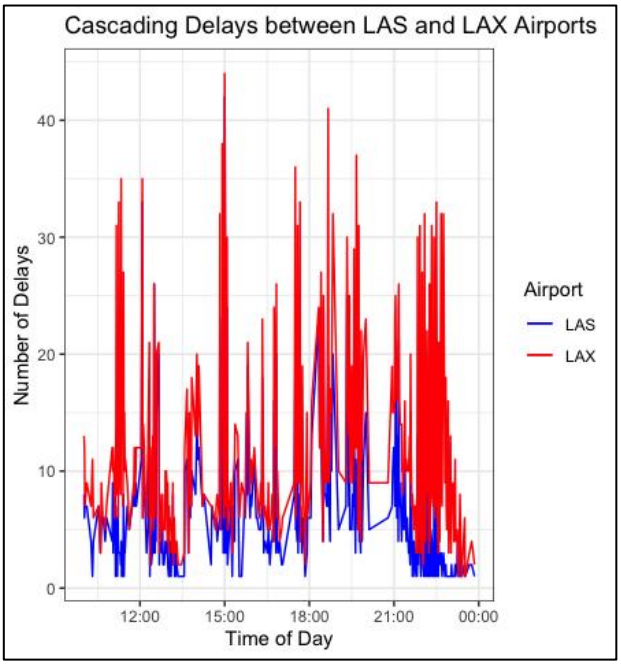


I used data from 2000-2005 and then sums the number of flights between each origin-destination pair for each year independently. The top 20 origin-destination pairs with the highest number of flights over the preceding six years are then chosen, and by adding the number of flights for each pair, they are further aggregated into 10 bigger "connections." The result is a table that lists these 10 connections together with the overall number of flights that took place over the course of the six years.

The map on the left shows flights between locations in the United States between the years 2000 and 2005 with thicker lines indicating more flights between airports and the size and colour of each point representing rank of airport . This can give information about the most popular flight paths and how passenger numbers alter over time.

On the right side, it is a line plot of exploration of drop in 2000 where selects “Origin”, “Dest”, and “Month” columns and group by month and calculate number of flights for each month. The result is there is a significant drop on February compared to other months.

2.4 Question 4. Can you detect cascading failures as delays in one airport create delays in others?



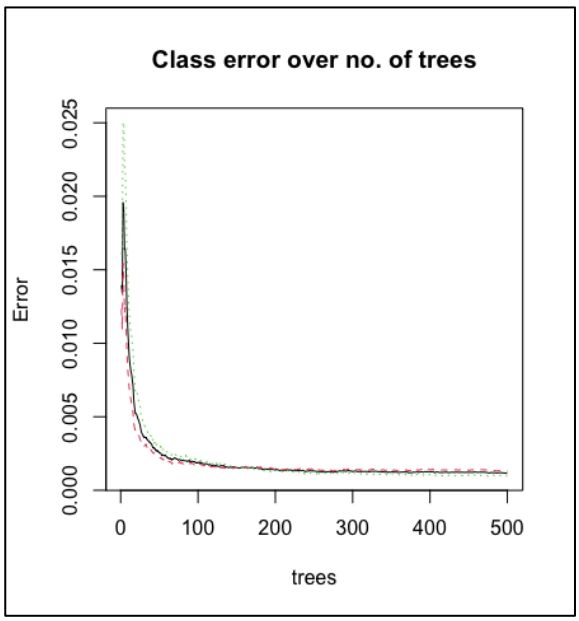
The analysis here is made to detect cascading failures as delays in one airport create delays in others, specifically between LAS and LAX.

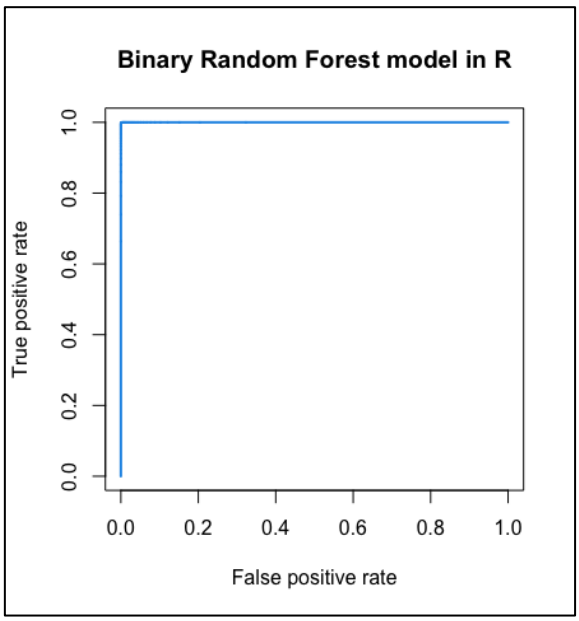
As a result, there were 368 cascading delay failures from LAS to LAX and back, which is 46.4 % of all flights on this route.

According to this analysis, delays at one airport (LAS) may cause delays at another airport (LAX) and then back again, impacting nearly half of all flights on this route.

The graph is used to show the trend of delays over time at the two airports (LAS and LAX) and shows the number of delays at the LAS and LAX airports over time, furthermore, as delays in one airport could potentially affect the other airport and cause cascading failures. The blue line represents the number of delays at LAS airport, and the red line represents the number of delays at LAX airport. According to the plot, there is a sharp increase delays at LAX shortly after increase delays at LAS as this could be an indication of cascading failures.

2.5 Question 5. Use the available variables to construct a model that predict delays.





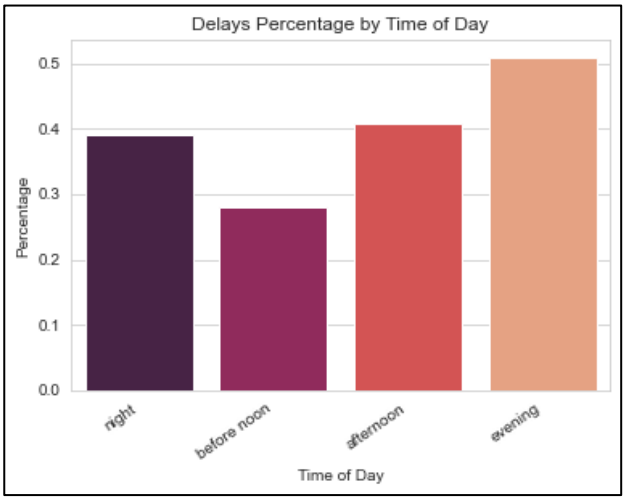
This model uses Random Forest to the flight data in order to predict whether a flight will be delayed or not. A binary target variable (delay binary) is created to prepare the data and is set to 1 if the delay is larger than 0 (i.e., the flight is delayed), and 0 otherwise. The model is then tested on the remaining 20% of the data after being trained on the remaining 80%. Graph on the left depicts the link between the classification error rate and the number of trees used in the random forest model. The plot makes it easier to decide how many trees should be used in the random forest model. Finding the number of trees that produces the lowest classification error rate is the objective. There is a threshold of diminishing returns where adding more trees does not considerably enhance the performance of the model, despite the fact that, generally, the classification error rate lowers as the number of trees increases. The output "Random Forest model prediction: 0.99912" indicates that the random forest model accurately forecasted the target variable with a 99.912% accuracy on the test set. In other words, for 99.912% of the flights in the test set, the model accurately predicted whether the flight would be delayed or not.

Meanwhile graph on the right, it shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at different classification thresholds.

3. Python Analysis

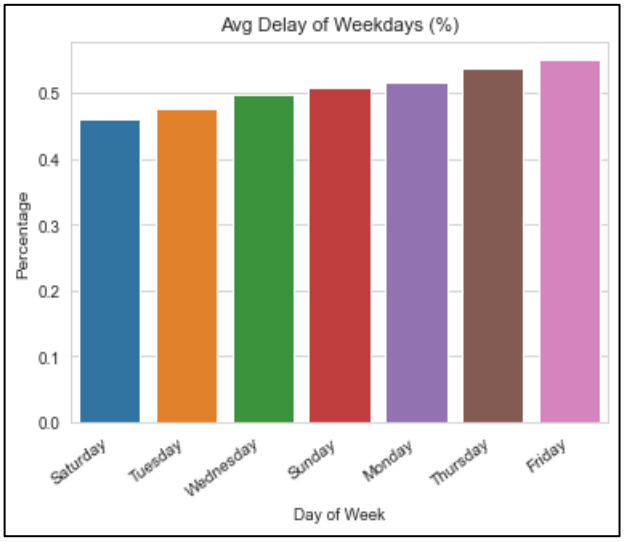
3.1 Question 1. When is the best time of day, day of the week, and time of year to fly to minimise delays?

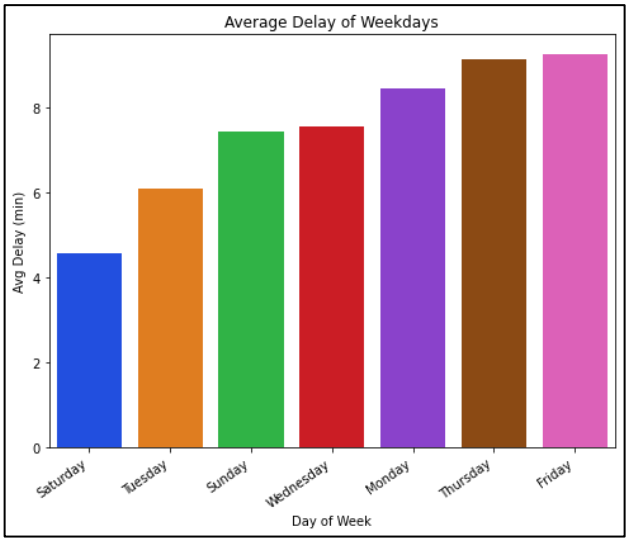
3.1.1 When is the best time of day to fly to minimise delays?



Based on the graph, it shows that best time of day divided by four categories is 'before noon' as the percentage of delay is 0.28% with average delay is 1.30. Hence, it is better for passengers to fly before noon or the next possible timing is at night to minimize the delay around 0.392%.

3.1.2 When is the best day of the week to fly to minimise delays?

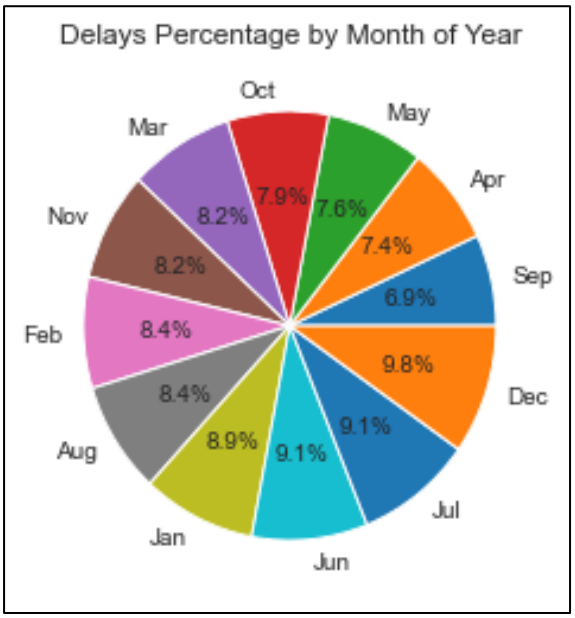




	DayOfWeek	Flights	Flights_perc	Delay_perc	Delays	Delays_perc
0	Friday	2072701	14.825586	8.179010	1143472	0.551682
1	Thursday	2064554	14.767312	7.961973	1113129	0.539162
2	Monday	2048953	14.655722	7.556611	1056457	0.515608
3	Wednesday	2034281	14.550776	7.248275	1013350	0.498137
4	Tuesday	2028319	14.508131	6.935785	969662	0.478062
5	Sunday	1943817	13.903706	7.062346	987356	0.507947
6	Saturday	1787942	12.788766	5.894289	824055	0.460896

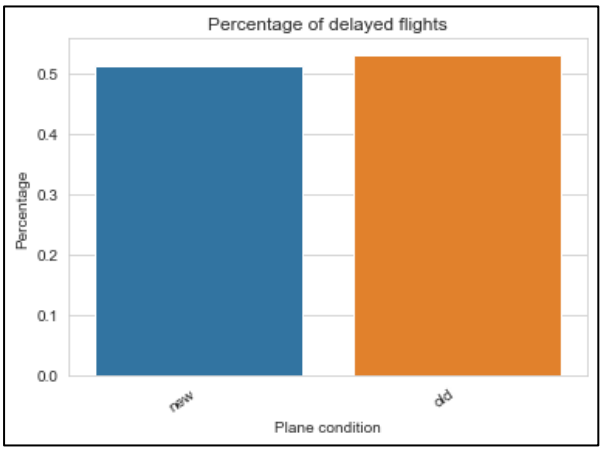
Based on the graph, it shows the ascending for both percentage and mean that Saturday has the lowest delay among other days in a week with 0.460% indicating that compared to other days of the week, Saturdays tend to have more timely flights. The next best timing is on Tuesday with 0.478%. Thus, passenger now has more options to go both on weekdays which is Tuesday and weekends on Saturday.

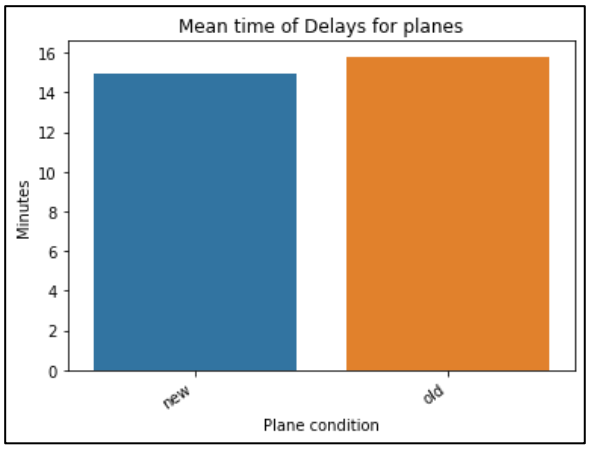
3.1.3 When is the best time of year to fly to minimise delays?



Here, I grouped the best time of year by months in a year, hence, it is easier to see the timing.
For the total number of flights and delays, respectively, in each month, this code generates two sets of data. By combining the two sets of information, it determines the proportion of delayed flights for each month.
This code generates a pie chart that displays the percentage of aircraft delays by month of the year. The percentage of delays for each month was obtained after grouping the data by month. The pie chart will display which months have the lowest percentage of delays and which have the most once the resulting data is sorted by the percentage of delays in ascending order.
Thus based on this pie chart, to find best time of year to get the minimum delays is on September with 6.9%, following by April of 7.4%, and May with 7.6%.

3.2 Question 2. Do older planes suffer more delays?

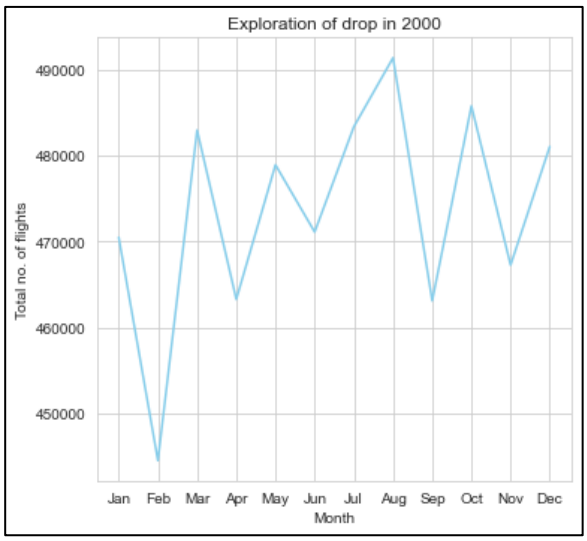




	plane_condition	Flights	Flights_perc	Delays	Delays_perc	Total_avg_delay
0	new	8146989	87.486522	4179778	0.513046	14.976610
1	old	1165290	12.513478	618822	0.531045	15.801298

The analysis includes integrating the flight and airplane data sets based on tail number and matching the variable names between the two. Depending on whether the plane was built before or after 1987, a new column is added to show the state of the aircraft. The analysis then determines the quantity and proportion of flights that are delayed for both old and new aircraft, as well as the mean delay duration for each circumstance. The percentage of delayed flights and mean delay duration are both higher for older planes, which indicates that they encounter more delays than newer aircraft. Where the average delay for older planes is 15.8% whereas new plane is 14,97%. However, even though both mean and percentage are higher for older planes, it doesn't necessarily mean that older planes suffer more delays as we need to consider from other variables.

3.3 Question 3. How does the number of people flying between different locations change over time?



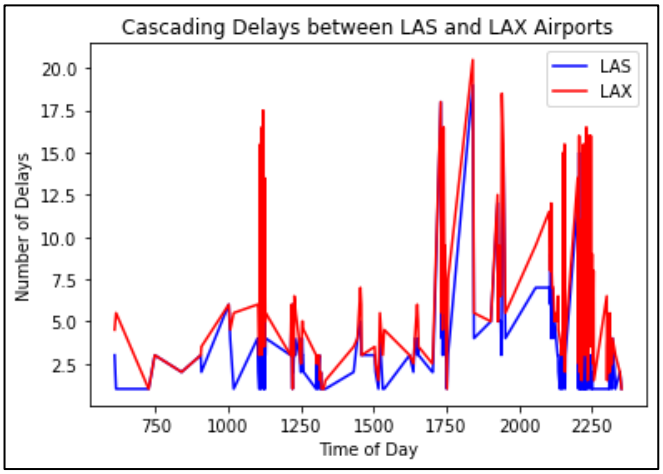
It reads in flight data for six years between different airports, groups the data by origin and destination airports, then counts the number of flights between each airport pair in each year.

It allows us to observe how the quantity of travellers between various areas changes over time. The busiest airport pairings and connected airports can be found. Also, we can check for any patterns in the frequency of flights between particular airport pairs over the span of the previous six years. Also, we can determine which airlines are most likely to operate between two places, which is valuable for evaluating the airline sector.

The line plot displays an overall rising trend, which suggests that as time has gone on, more people are taking flights to get from one place to another. It can be seen that there was a sharp drop on February below 450,000 and September around 460,000. And it just started to fluctuate until there was an increased in December above 480,000. The drop might be due to some aspects such as natural disasters, economic downturn, and etc.

3.4 Question 4. Can you detect cascading failures as delays in one airport create delays in others?

	Origin	Dest	y2000	y2001	y2002	y2003	y2004	y2005
0	LAX	LAS	17745.0	16165.0	12733.0	11262.0	12494.0	12536.0
1	PHX	LAX	17041.0	14515.0	11066.0	10498.0	10422.0	10819.0
2	LAS	LAX	16899.0	15788.0	12526.0	11177.0	12487.0	12519.0
3	LAX	PHX	16652.0	14149.0	11027.0	10518.0	10469.0	10820.0
4	LAX	SFO	16021.0	13244.0	9832.0	8487.0	9049.0	8427.0

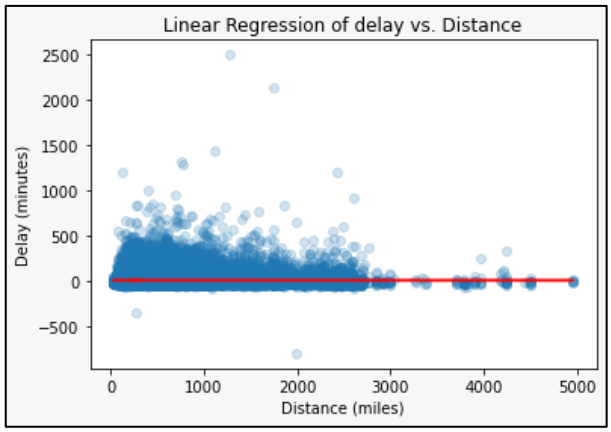


The table shows the Origin and Destination airports across years in 2000 to 2005. It examining the frequency of cascading delays from LAX (Los Angeles) to LAS (Las Vegas) and vice versa. It restricts the data to only show aircraft that left from LAS and arrived at LAX, and it gives the DepDelay and ArrDelay columns binary values depending on whether or not the delays are positive. The filtered data is then further filtered to include only planes with a positive arrival delay. The final analysis is total number of failures caused by cascading delays from LAS to LAX and back, as well as the proportion of all flights on this route that were affected by such delays.

Regarding the plot, it displays the patterns in delays for the airports of LAS and LAX. The number of delays is represented on the y-axis, and the time of day is represented on the x-axis. The red line shows the trend in delays at LAX airport, whereas the blue line shows delays at LAS airport. Which line corresponds to which airport is indicated by the description on the right side of the graphic. The plot makes it easier to see how both airports' delay trends and how they relate to one another. When the number of delay of airport LAX increased, it will also affect airport LAS and both the trend of line increased. In conclusion, the two airports' delays are related to one another, and delays at one airport might have a big effect on the other airport.

In addition, there were 181 cascading delay failures from LAS to LAX and back, which is 16.5 % of all flights on this route. Thus, it means 16.5% of all flights on the Las Vegas (LAS) to Los Angeles (LAX) and back route had cascading delay issues. The investigation discovered a total of 181 flights that had cascading delays, which means that these flights had delays that resulted in a cascade effect on following flights. This emphasizes how interconnected and dependent planes and airports are, as well as how interruptions in one area of the system could have a significant impact on the entire system.

3.5 Question 5. Use the available variables to construct a model that predict delays.



The first model uses Linear Regression because the objective is to predict the amount of delay (in minutes) based on the distance of the flight. The model will forecast a continuous output rather than a binary result in the absence of a threshold. This means that instead of just forecasting if there will be a delay or not, the model will use the input features to forecast the length of the delay for a certain trip. Instead of using a classification mode in this situation, you would use a regression model. The predictor variable is the distance of the flight, which is a continuous variable, and the outcome variable is the amount of delay,

which is also a continuous variable. Subsequently, a model that forecasts delays based on distance is created using linear regression.
A linear regression model's RMSE value of 61.92247 indicates that the model's predictions are typically 61.92247 minutes off from the actual delay time. This figure suggests that the model's accuracy could be increased and that its performance is not very good.

If the LR model's accuracy within 10 minutes is 0.1559, only 15.59% of the time does the model accurately forecast that the delay will be within 10 minutes of the actual delay. Its result is similarly very inadequate, indicating that the model is not well suited to forecast flight delays.

4. Conclusion

For question 1, In summary, before noon and then at night are the best times of day to fly and avoid delays. The day of the week with the fewest flight delays is Saturday, followed by Tuesday and Wednesday. The least amount of delays occur in September, then April, then autumn, which has the lowest average delay. These results imply that fewer travellers and less traffic during these times and periods may result in lesser waits.

For question 2, The data reveals that older planes had greater percentages of delayed flights and longer average delays, but the difference is not large enough to draw the conclusion that older planes experience more delays than newer ones. In addition to the age of the aircraft, other variables may potentially cause delays.

For question 3, In order to monitor changes in the number of passengers flying between different places over time and determine the busiest airport combinations, related airports, and flight frequency patterns, the research looks at flight data from the previous six years. It also covers the top airlines that fly between cities and looks into why there were decreased flights in February 2000. A list of the top 10 airport connections and a map of flights connecting US cities from 2000 to 2005 are included in the results.

For question 4, Between the airports of LAS and LAX, the research finds cascading delay issues. The graph illustrates the pattern of delays at the two airports and demonstrates how delays at one airport can impact delays at the other. Cascading delays affected a total of 181 flights, or 16.5% of all flights on this route. According to the analysis, delays at LAS may result in delays at LAX and vice versa, impacting over half of all flights.

For question 5, In light of the data available, the Random Forest model outperforms the Linear Regression model in forecasting flight delays. The Random Forest model is simpler to understand and apply in reality because it predicts a binary outcome and has a greater accuracy (99.912%). The provided graphs demonstrate that it is also a classification model, which is better suitable for this situation, and that it performs well. As a result, it is recommended that the Random Forest model be used to forecast flight delays.