Exploring Airline Delays

Auren Ferguson

07 October 2017

Summary

This document explores US airline flights, specifically delays for the year 2008.

EDA

High Level EDA

Ultimately, we would like to be able predict if a flight will be delayed **before the flight**, i.e, we want to avoid data leakage. The target variable will be `ArrDelay`, this is the time in minutes that a flight was delayed by. There was 8387 NA values in `ArrDelay`, these rows were removed as the data set has over 1.9 million rows. Figure 1 shows a density plot of `ArrDelay` showing the majority of delays are less than 100 minutes (extreme values are capped in the plot).

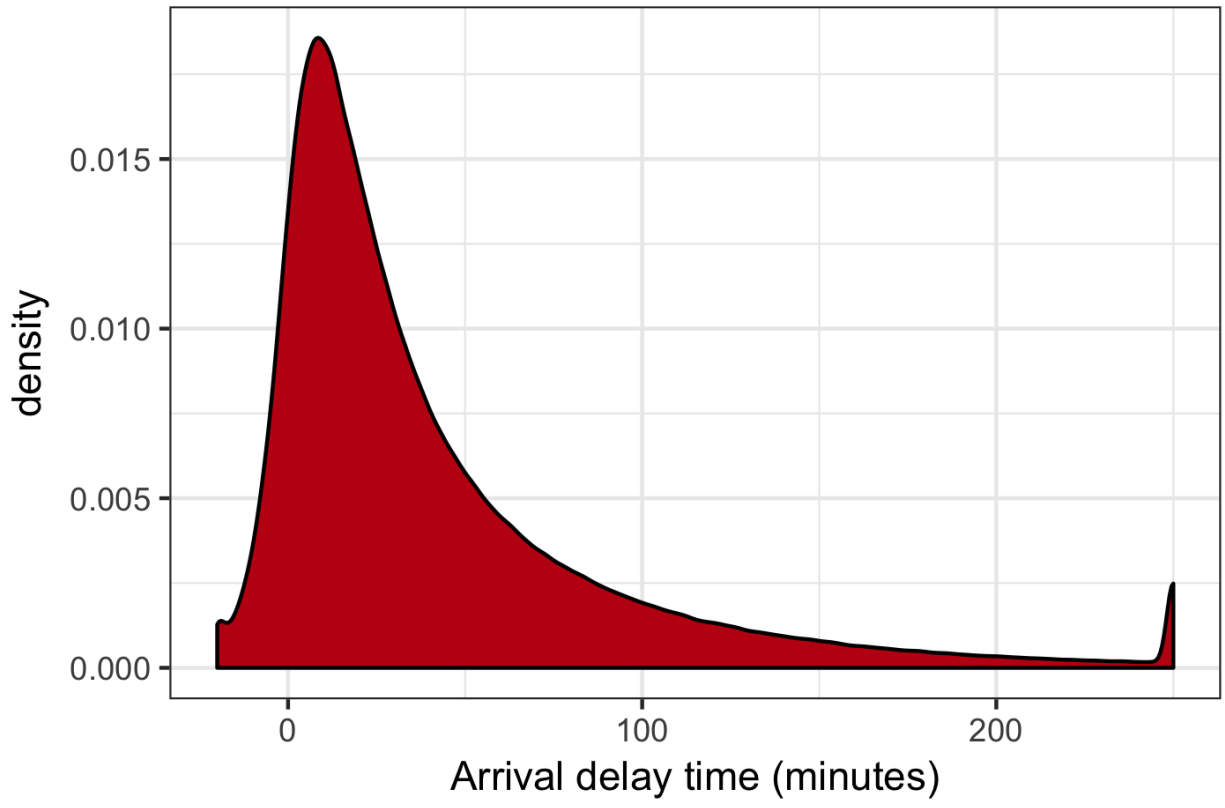

Figure 1: Density plot of flight delay time

To create a binary delayed indicator, a sensible value of `ArrDelay` had to be picked. This was done in a rather arbitrary manner based upon what I feel constitutes a flight delay, i.e. I don't consider a flight that arrives 5 minutes past its scheduled time to be delayed. I picked a time of 40 minutes. Figure 2 shows a histogram of flight delays, surprisingly, approx 37% of flights are delayed were delayed for 40 minutes or more.

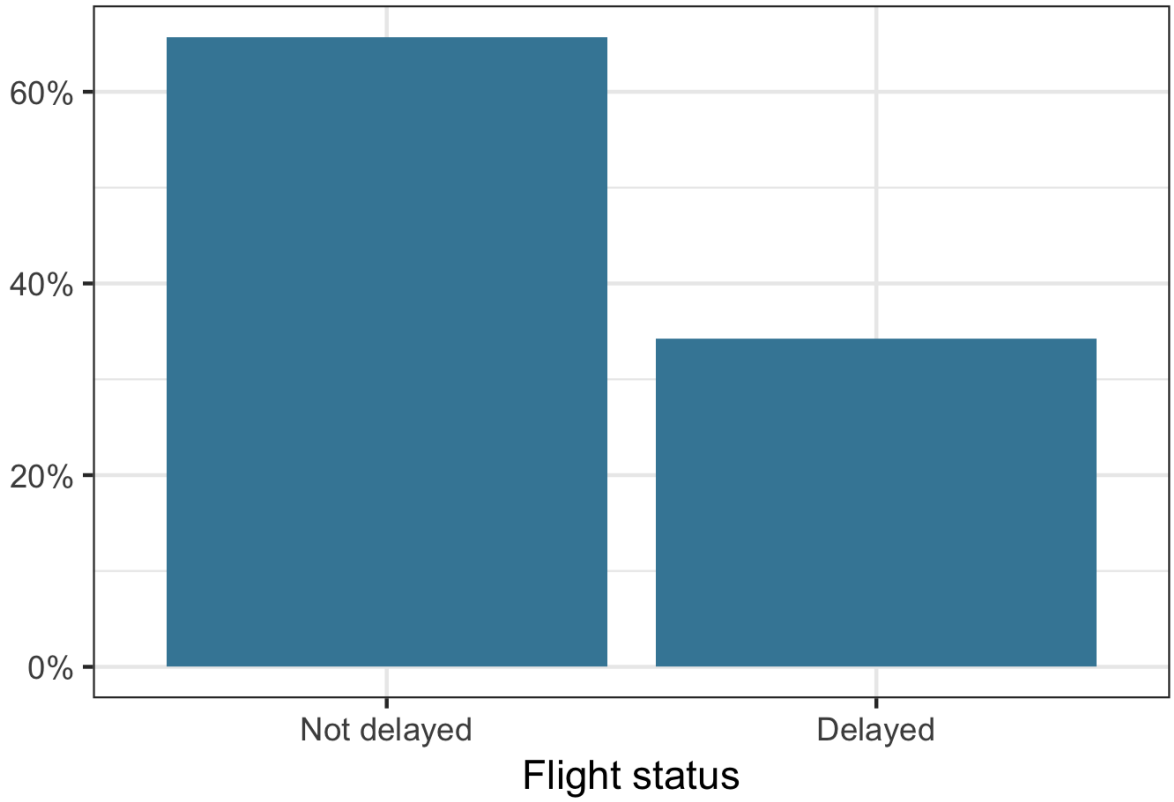

Figure 2: Histogram of flight delays

More Indepth EDA

This section looks deeping into the data to extract insights that may be of benefit to consumers and airlines.

What times of the year are the worst for travelling

Figure 3 shows average delay times (monthly adjusted) which shows several spikes in delay times including ones in June and towards the end of the year. Figure 4 is zoomed in version of figure 3 which shows spikes in delays that correspond to Thanksgiving and Christmas, these are times when lots of people will be travelling and so delays can be expected.

Which airlines are the most punctual

Figure 5 shows how each airline's delayed rate compared to the mean delayed rate. The best airlines to travel with include: South West, Frontier Airlines and US Airways. While the worst airlines include: Jet Blue, American Airlines and United Airlines.

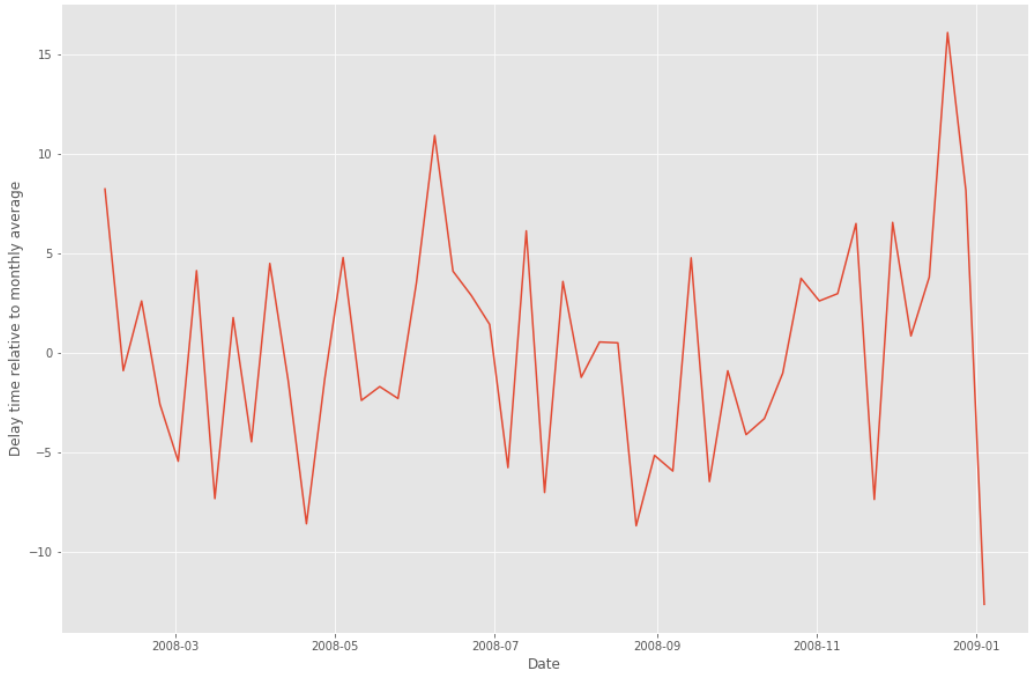

Figure 3: Delay times as a function of time (monthly adjusted)

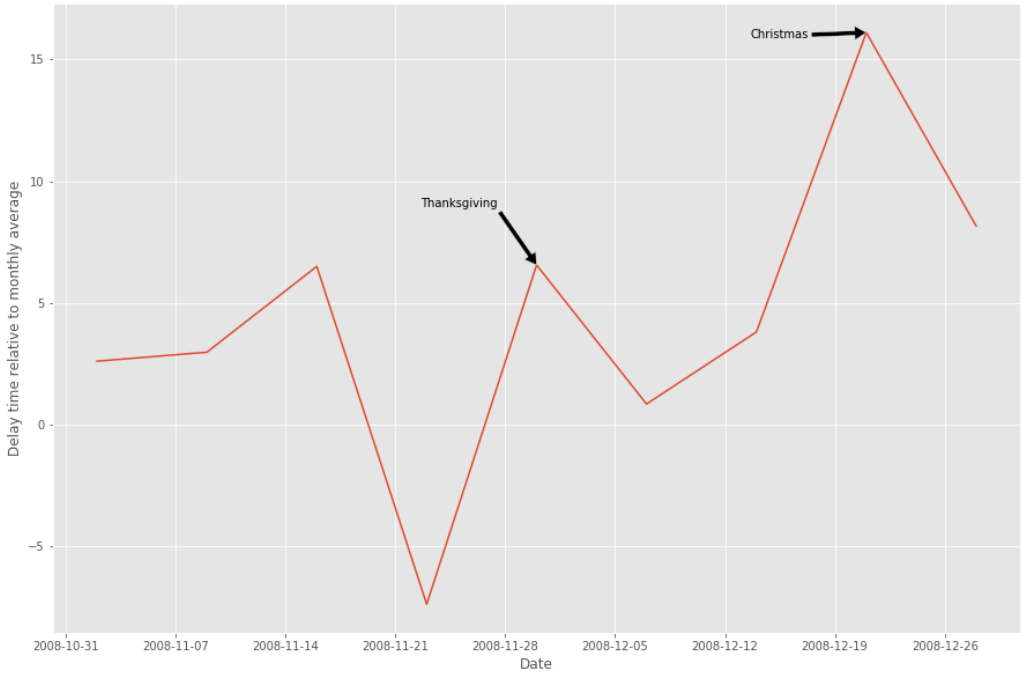

Figure 4: Delay times as a function of time (monthly adjusted) for the American holiday period

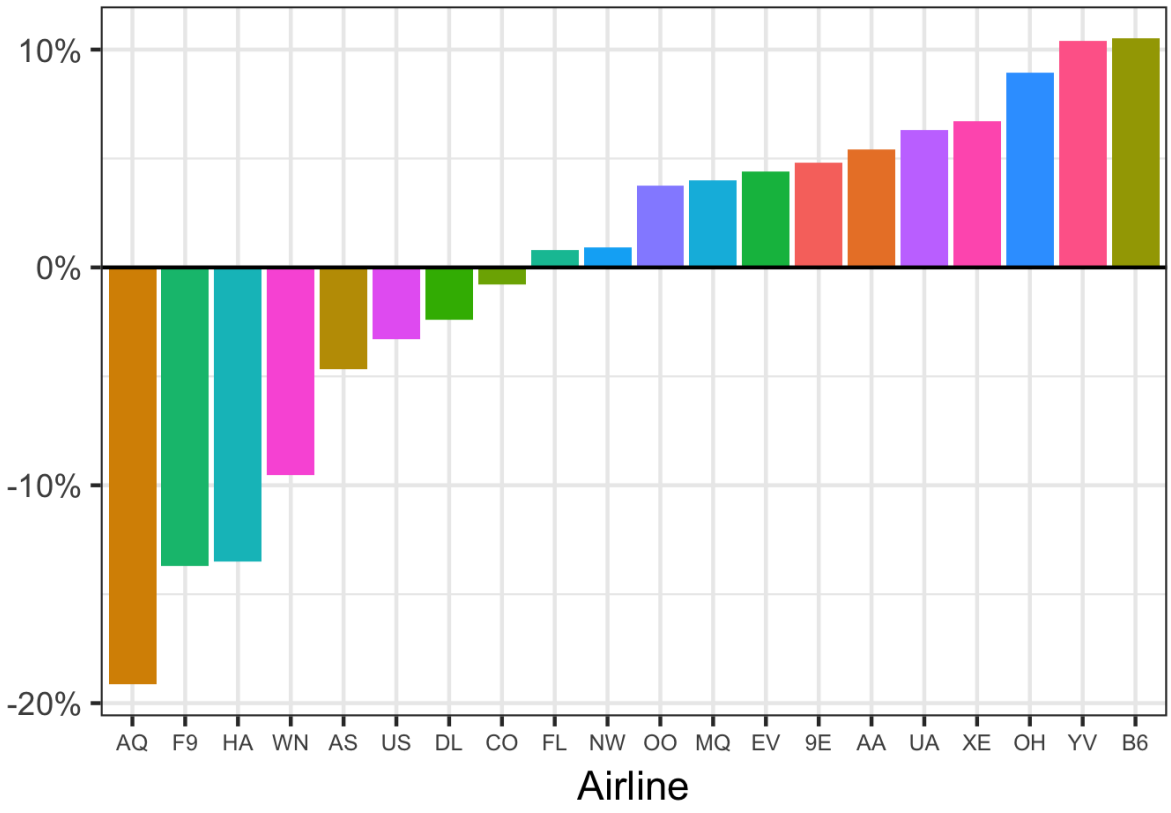

Figure 5: Best and worst airlines in terms of punctuality (y-axis is the differnce between airline delay rate and global delay rate)

The best and worst large airports

Figure 6 shows the how the 20 largest airports (>8000 flights) perform. Some of the best airports include: Phoenix, Las Vegas and LA. While some of the worst airports are: JFK, laguardia and O'Hare (Chicago). This is good to know but usually one doesn't have too much choice in airport unless you are in a large city such as New York.

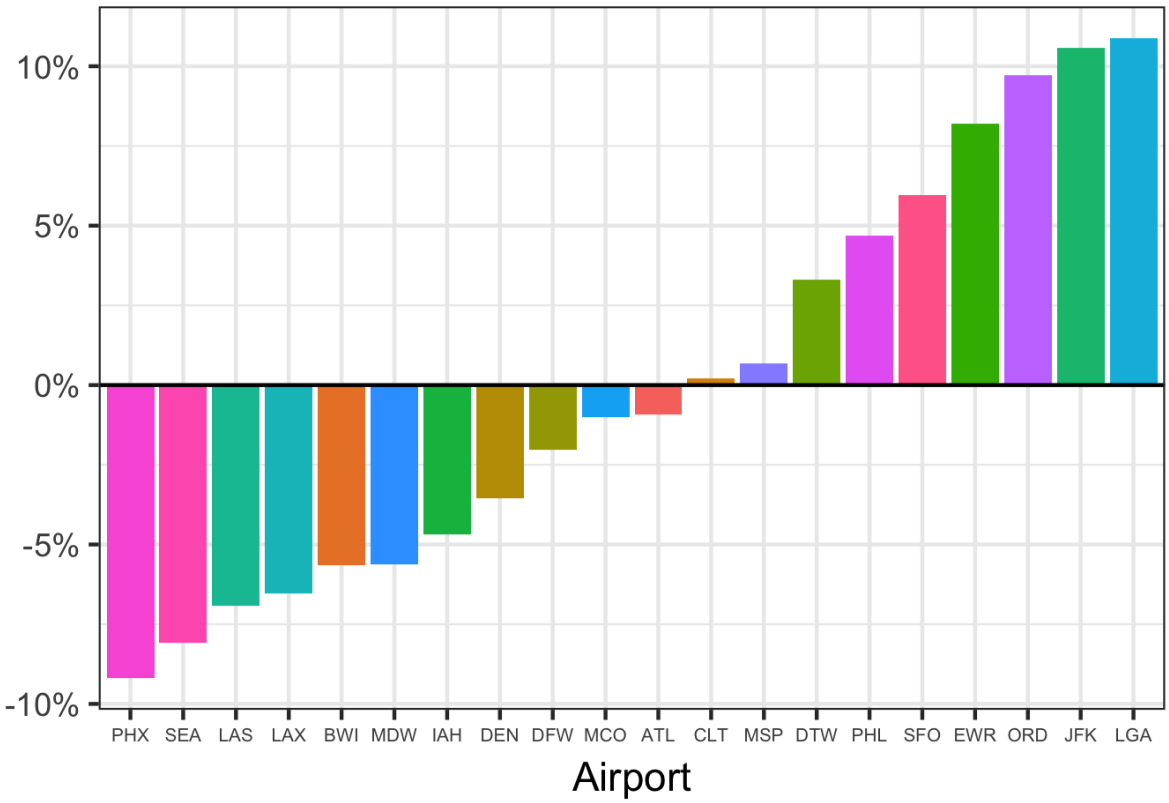

Figure 6: Best and worst large airports in terms of punctuality (y-axis is the differnce between airport delay rate and global delay rate)

Busiest routes

The delay rate of the 20 busiest routes (>2900 flights) are shown in figure 7. Some of the best routes include: Phoenix-LA, Houston-Dallas and Atlanta-Orlando while some of the worst routes are: laguardia-O'Hare, Atlanta-Washington State and LA-San Francisco.

Factors that cause delays

The largest factors that cause delays are shown in figure 8. Somewhat to my surprise, weather plays a small role in delay reasons with the majority of delays stemming from carrier problems and air traffic control. This is somewhere airlines could target to improve outcomes.

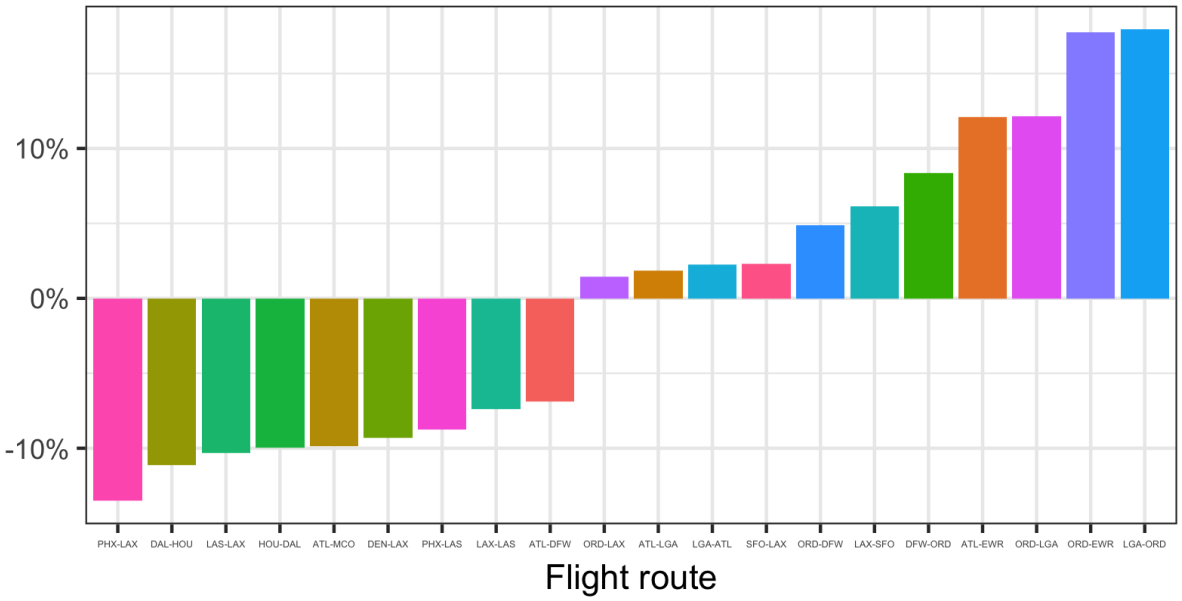

Figure 7: Best and worst routes in terms of punctuality (y-axis is the differnce between route delay rate and global delay rate)

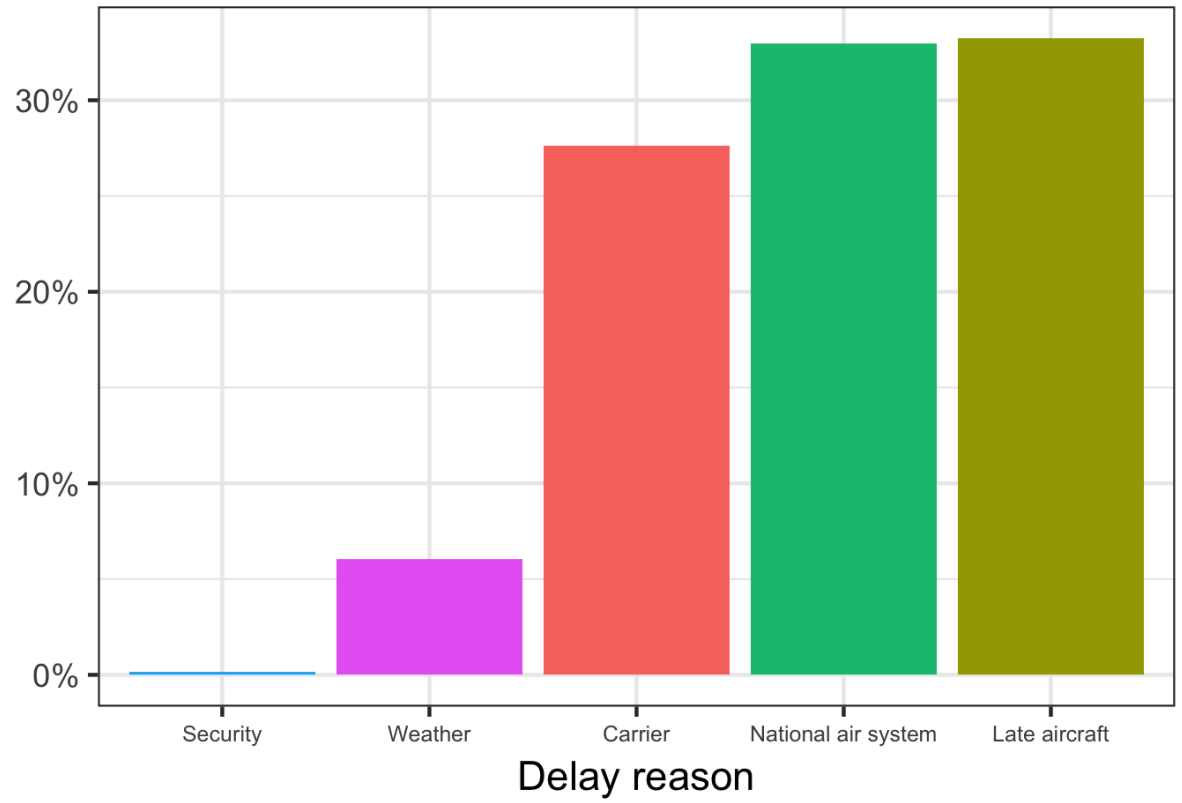

Figure 8: Contributers to flight delays

Clustering analysis

k-means clustering was undertaken to see if any that could be partitioned into preferably 2 distinct groups, delayed and not delayed flights. Some data preprocessing had to be undertaken before running the model. These included:

- Removing zero variance variables
- Remove variables that uses actual flight information, i.e. actual flight time
- only using the hour from scheduled depart and arrive time
- taking only continious variables
- scaling the data

The variables that were input into the model were:

- Scheduled departue time
- Scheduled arrival time
- Estimated flight length
- Flight distance

It would of been nice to have more continious variables to be able to model with.

At first, 2 cluster centers were used as input into the model as ideally it would be able to seperate delayed and on time flights. The results of this are shown in table 1. As we can see, the clustering wasn't very succesful as both clusters have a similar percentage of delayed flights.

Table 1: Results of k = 2 when compared to delayed/not delayed target (%)

	Cluster 1	Cluster 2
0	68	65
1	32	35

In order to see if this could be improved upon, k-means clustering was done for 1 to 5 centers and the results are shown in the scree plot in figure 9. There is a distinct elbow in the within group sum of squares at k = 3, perhaps this could be more related to delayed or on time flights. Similar to table 1, the results for k = 3 are shown in table 2

Table 2: Results of k = 3 when compared to delayed/not delayed target (%)

	Cluster 1	Cluster 2	Cluster 3
0	68	69	62
1	32	31	38

We can see from table 2 that cluster 3 has a noticible increase in delays compared to the other 2. This could be a helpful variable in our final model for predicting if a flight will be delayed but having 3 clusters for a binary outcome isn't very intuative. The input variables into the model (shown above) have a high correlation and therefore, reducing the number of dimensions using PCA may be of some benefit as it will remove the correlation (which will be helpful for the logistic regression model) and may improve outcomes in terms of have 2 distinct clusters. The results of the PCA are shown in figure 10 which show that the first two principle components account for over 90% of the variance of the input variables. k-means clustering was ran on the first two principle components with k=2 and the results are shown in table 3.

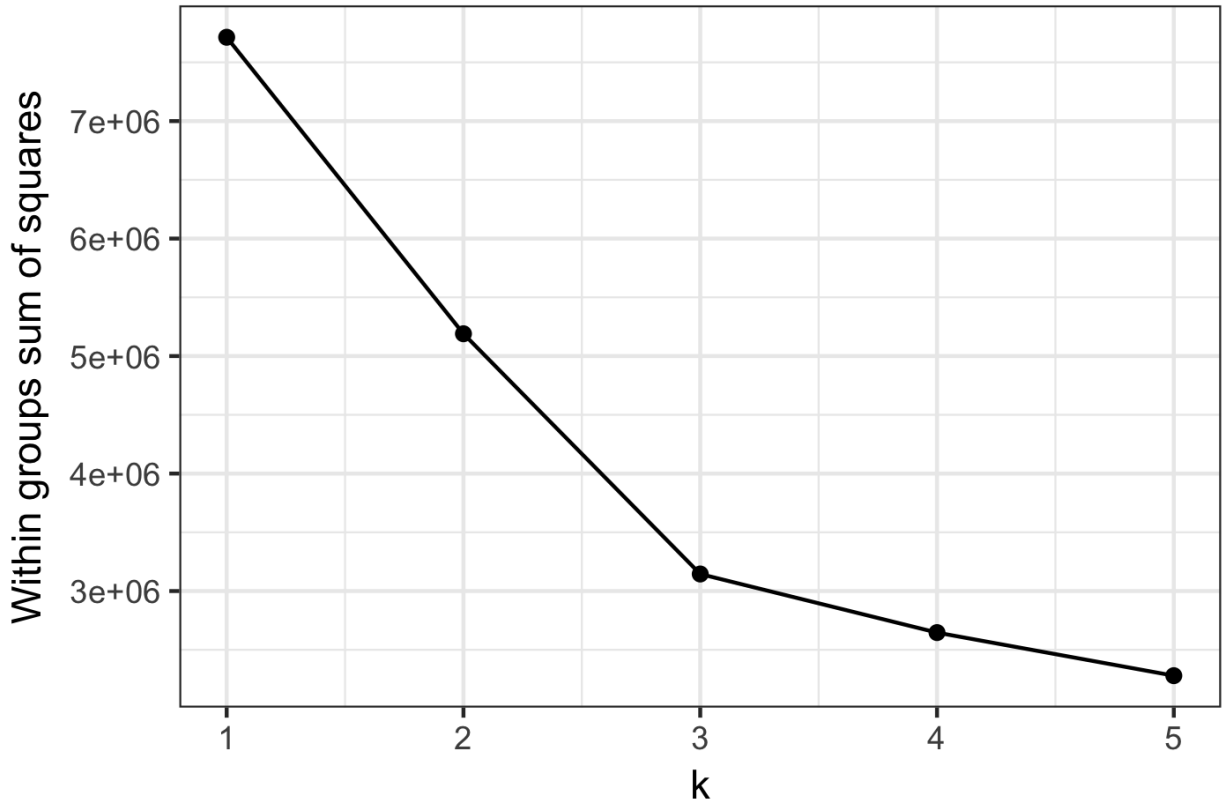

Figure 9: Scree plot for cluster centers between 1 and 5

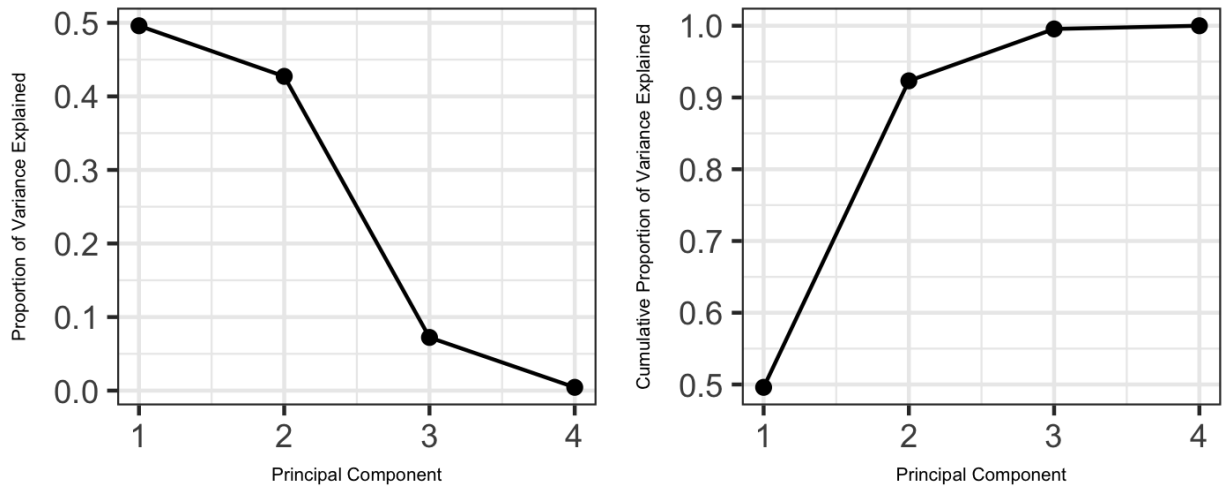

Figure 10: a) Explained variance per principle component and b) cumulative variance with number of components

Table 3: Results of k = 2 performed on the first 2 princple components when compared to delayed/not delayed target (%)

	Cluster 1	Cluster 2
0	70	63
1	30	37

From table 3 we can see we have two distinct groups with a 7% increase in delays in cluster 2 compared to cluster 1. While not an overwhelming seperation between groups it may be a predictive attribute in the logistic model. This can be seen in figure 11 which shows the the two clusters and whether the flight has been delayed or not. pca1, pca2 and assigned cluster are to be included for further modelling.

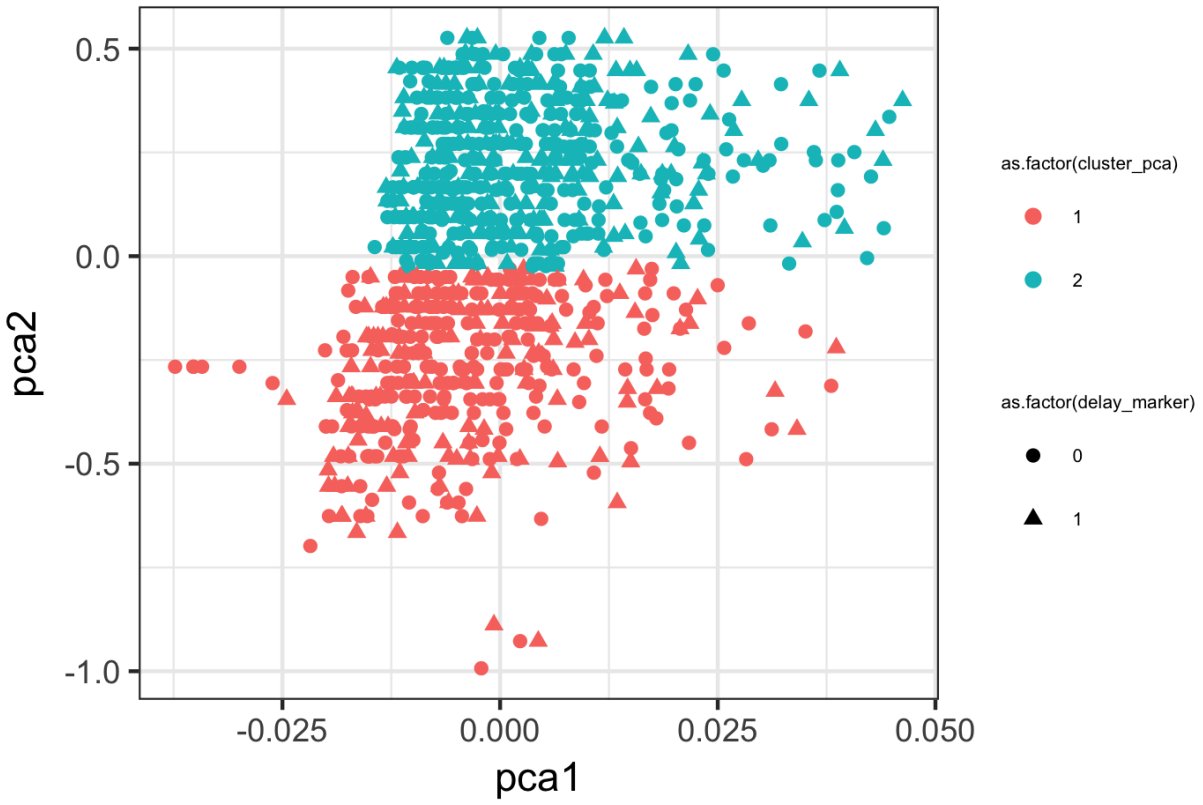

Figure 11: Scatter plot of pca1 vs pca2 showing separation of clusters and flight status

Logistic regression model

Using the input variables and the variables that were engineered above, a logistic regression model is used to predict if a flight will be delayed or not. before modelling, a preprocessing stage must be done. This included:

- Spliting the data into training and test sets (80/20 split)
- Reducing the number of factor levels for variables: `Origin`, `Dest`, `UniqueCarrier`
  - This was done by scoring each origin, etc and labelling them good, medium, bad etc
- In a similar manner, month and day of week were grouped together into good, medium, bad etc

The final inputs into the model were:

- Month when flight happend
- Day of week when flight happend
- pca1
- pca2
- assigned cluster (removed in final model)
- Airline
- Flight origin
- Flight destination

A variable correlation plot was done to ensure that none of the input variables were highly correlated to improve model performance and is shown in figure 12. Generally, I allow a correlation cofficient of 0.4, although there is no hard rule on this and very much person dependant.

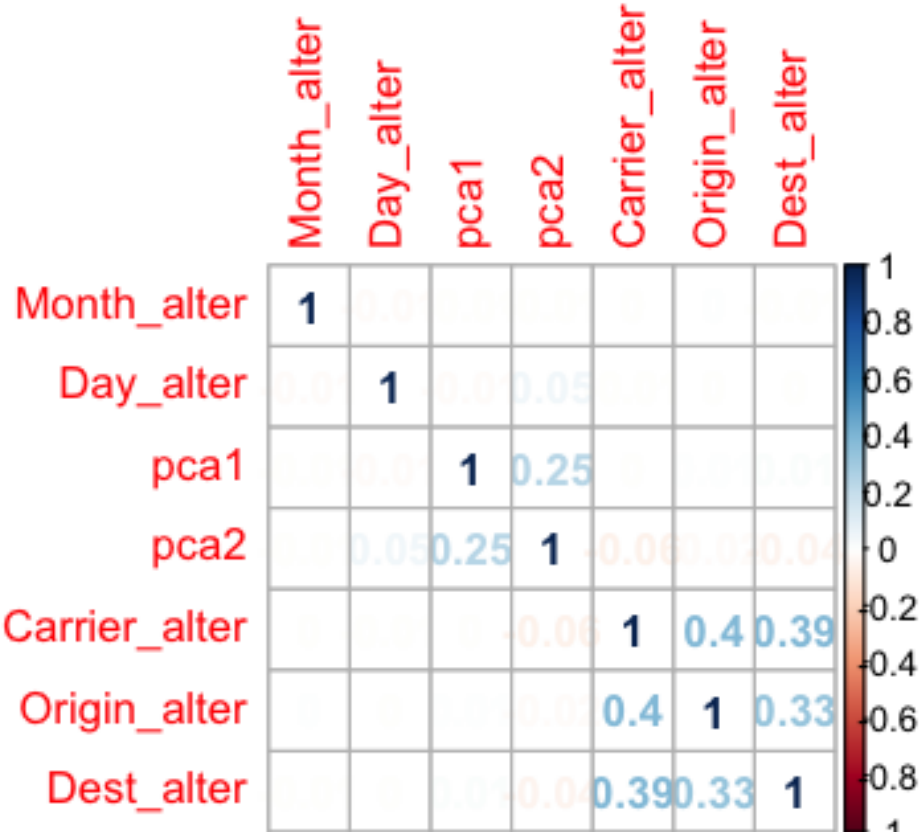

Figure 12: Correlation plot of input variables into the model

Ten fold cross-validation was done when training model to help prevent overfitting. The probobility cutoff that decides whether or not the prediction is assigned as delayed or not was approx 0.34.

Model performance on test data

The confusion matrix for the model on test data is shown in table 4.

Table 4: Confustion matrix on test data

Prediction	Delayed	Not delayed
Delayed	40	14
Not delayed	26	20