NYC Flights 2013 Analysis

Introduction

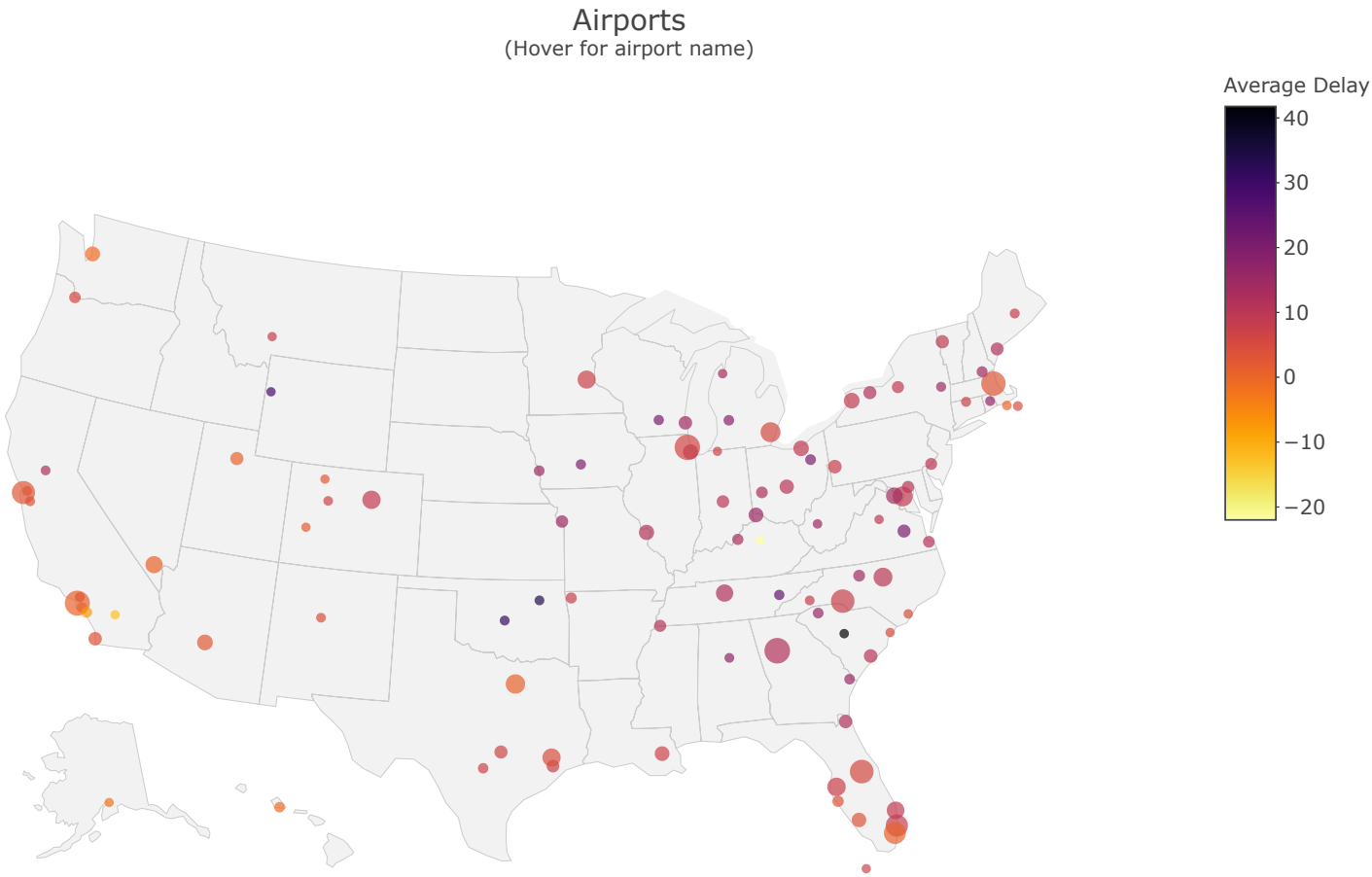
This report aims to outline the factors, both passenger controllable (e.g. the airline used) and uncontrollable (e.g. weather), that most influenced whether or not a flight would arrive late, along with the extent of this delay. The goal here was to allow potential passengers to make “smarter” booking decisions in order to minimise the possibility of experiencing delays, while also allowing them to better expect potential delays.

To do this the report first outlines the details of the underlying data used. Following this, details of delays by airline will be discussed. This will include basic details of which airline experiences the most delays, but will also detail how severe these delays can be and their variation by airline. Next, a discussion of any other factors in the dataset that contribute to delays will be conducted. Finally, the results of the two models built to predict delays and the accuracy of their predictions will be presented.

Datasets

As the report’s title suggests, the data being considered relates to domestic US flights originating from New York City’s 3 airports (Newark Liberty International, John F Kennedy International and La Guardia) in 2013. This data is contained in the R ‘nycflights13’ package, which contains 5 separate datasets. An overview of what is included in these is detailed in Appendix A.

The most fundamental of these is the “flights” dataset. This contains the details of all flights, including the route, all schedule and timing details, along with the aircraft tail number for identifying planes. The map below shows the destinations of these flights with the size identifying the destination’s popularity and the colour identifying the average arrival delay.



Evidently from the above plot there are delays, in some cases quite significant delays. In fact, analysis of the data shows that approximately 2 out of every 7 flights are more than 10 minutes late. This report aims to highlight the factor that may be causing or contributing to these delays.

However, in order to complete this analysis the flights dataset alone is insufficient. This is in part because some the data is encoded, making it meaningless, while there are also other factors that need to be considered. In order to understand the encoded variables, the data is combined with the “airlines”, “airports”, and “planes” datasets.

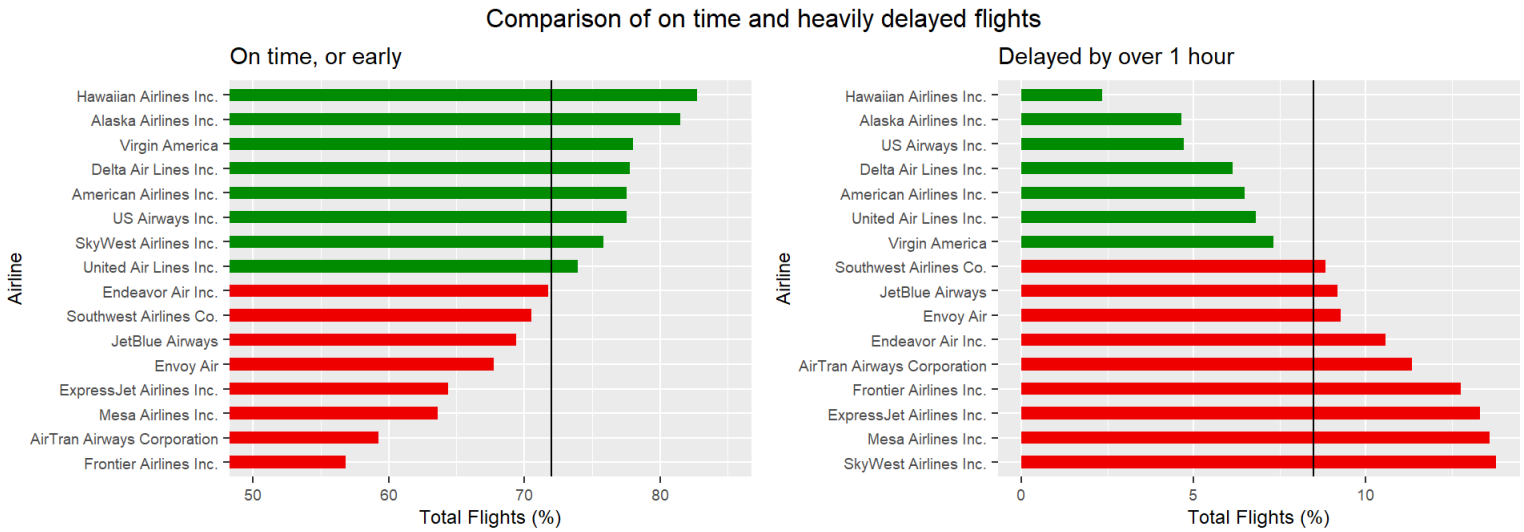
The “weather” dataset provides hourly readings for each of the airports and, since the weather could have an impact on a plane’s ability to take off, this was combined with the flight details.

As with all datasets none of these will be perfectly clean or error free. Consequently work was done to tidy the data and correct any errors that were identified. Details of the changes made are included in Appendix B.

Airline Delays

One of the most fundamental choices any passenger will make in booking a flight will be deciding which airline to fly with. In total there are 16 different airlines operating flights from NYC's major airports. However, the flight volumes vary dramatically. Some, such as United Air Lines or JetBlue Airways, operate tens of thousands of flights while others, such as SkyWest Airlines, only operate several dozen flights. In fact, the top 5 carriers account for nearly 75% of all flights out of New York airports.

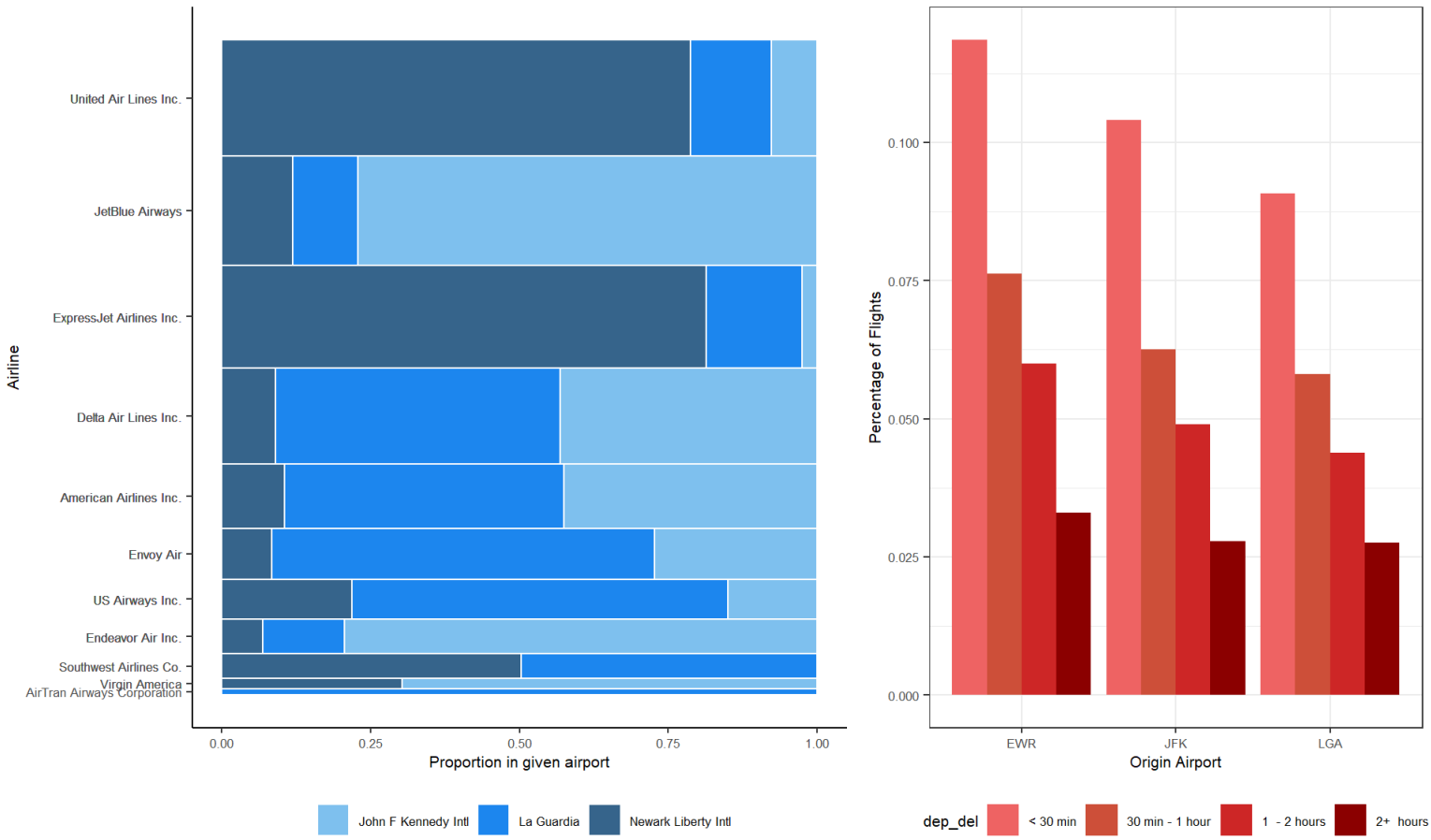
The plot below shows how the different airlines perform in terms of timeliness along with showing the proportion of serious delays they experience.



This plot suggests that based on the proportion of flights on time and severely delayed, Hawaiian Airlines and Delta Air Lines are reliable while Frontier Airlines and ExpressJet Airlines should be avoided. However, further investigations are made to determine whether this is a true reflection of these airlines' performance or simply a result of other factors.

Airport Delays

As mentioned above there are 3 departure airports that were considered. The plots below show how the flight volume of the different airlines varied between these airports and also the proportion of delays from each airport.

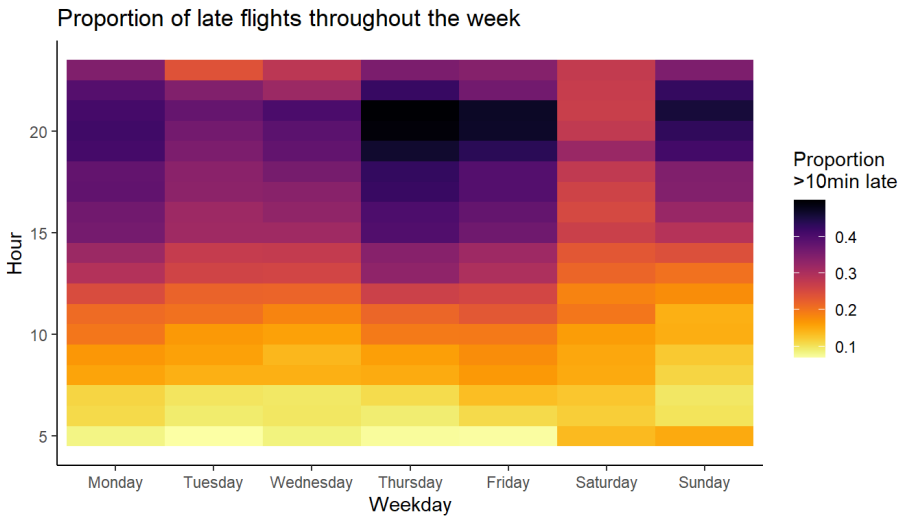


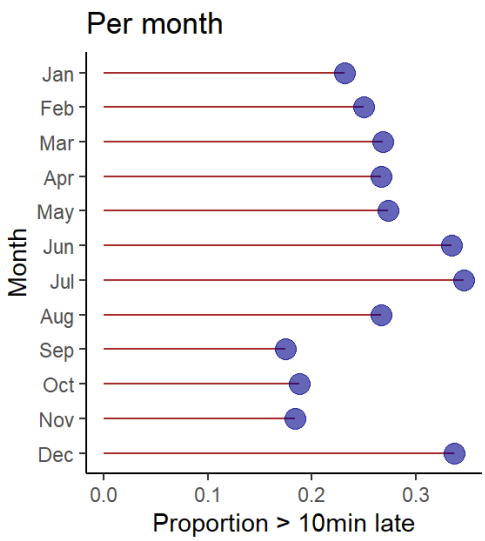
This shows that more flights are delayed in Newark than in JFK which, in turn, has more delays than La Guardia. We also see that both United Air Lines and ExpressJet Airlines operate a significant volume of their flights from Newark, with volumes that are significantly higher than those of the other carriers. This might incline us to believe that these should, similar to the airport they are operating from, show relatively poor performance in terms of delays. This, however, is not universally the case as while ExpressJet Airlines does show relatively poor performance the same is not true of United Air Lines. In fact, United Airlines shows better than average timeliness performance. This shows that there are either other factors contributing to the delays or the airports' performance is in fact driven by the airlines' performance rather than contributing to them.

Time Delays

The plots below show how flight performance varies with different time factors. In particular, it shows that delays are more common in the evenings with Thursdays and Sundays being noticeably worse. It also shows that timeliness is generally better for early morning flights. It can be observed that Saturday flights suffer the lowest levels of delays throughout the evening.

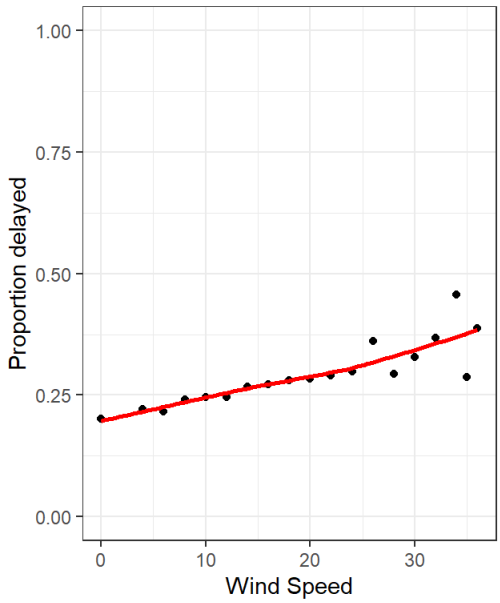
In terms of the months, the Summer months of June and July are when delays are most common with a much lower volume of delays in September through to November.

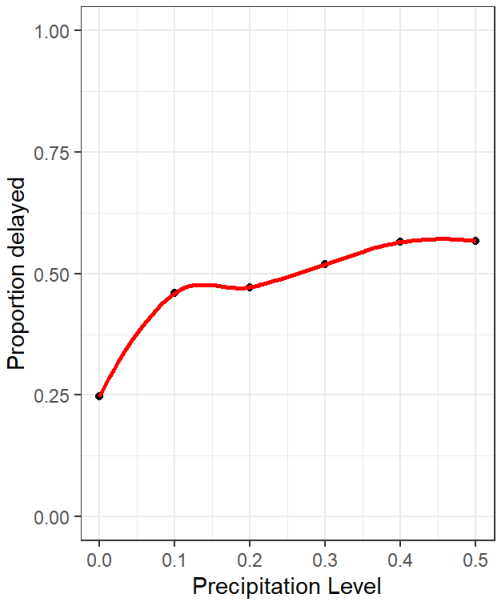


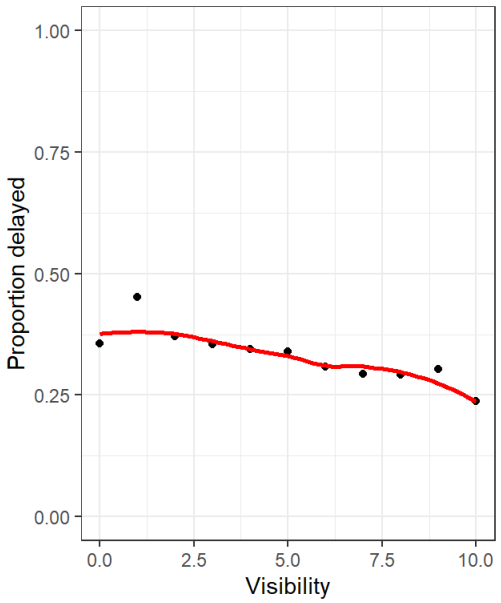


Weather Delays

The final factors that impacted flight timeliness under consideration were those involving weather conditions. The analysis identified 6 of the weather-related factors impacted flight timeliness. The plots below summarise how the 3 most influential weather events impact flight delays. In particular, as either the wind speed or precipitation increases so does the proportion of delays. Conversely, an increase in visibility reduces the volume of delays.







While these are useful in terms of better understanding why a flight might be delayed, they are, given the limitations of meteorological forecasting, unlikely to help potential passengers when booking flights. Yet they may prove useful in allowing them to anticipate delays and better plan their personal schedule.

Arrival and Departure Delay Models

Flight delays can be very frustrating to passengers and costly to airline companies. Flight delays are not easy to understand as they arise from multiple reasons. In order to help determine the prevalent reasons, two models were designed.

Diving further into the roots, an investigation was made into the departure delay by modeling it with predictors from the “weather”, “flights”, and “plane” datasets. Summarizing the results, it was found that the departure delay depends the most on the departure hour. A one hour increase in departure hour, where all the meteorological factors along with the plane’s age are fixed, increases the odds of getting a departure delay of more than 15 minutes by approximately 15%. Additionally, an increase of 1 Mile per hour in the wind speed, where all the other predictors are fixed, increases the odds of getting a departure delay of more than 15 minutes by 2.7%. Some other significant predictors for predicting departure delay are temperature, pressure, humidity, and plane’s age. Overall, the model can predict whether a flight would be delayed or not while leaving from the origin with an 80% accuracy.

The arrival delay model aimed to generate predicted arrival delays by fitting a linear model using some of the important predictors. The results from the summary output show that all weather features are significant to the likelihood of flight delays. The model also showed that the quarter of the year in which the flight is taken will affect the delay with changing from quarter 1 to quarter 2, all else being equal, will increase the estimated flight delays by 3 minutes. Similarly, choosing to fly with Delta Air Lines Inc over AirTrans Airways decreases the estimated arrival delay by 6 minutes. The model also showed a near perfect correlation between departure and arrival delays with a minutes departure delay giving rise to a similar minute delay in the estimated arrival time, again all else being equal.

A summary of these models can be found in Appendices C & D.

Summary

The report considered US internal flights for 2013 operating out of New York City’s 3 major airports with the aim of identifying the factors that are contributing to (approx) 2 out of every 7 flights being late. The hope is that by identifying these factors it would better allow potential passengers to reduce their chances of being delayed.

These included
Airlines: Delta, US AIrways and Hawaiian Airlines all show good timeliness performance with a high proprtion of on time arrivals and low level of severe delays while ExpressJet and Frontier airline both show the opposite with poor timeliness performance.
Airports: Newark shows a greater volume of delays than JFK while La Guardia shows the lowest proportion.
Time: Delays are more prevalent in the evenings with Thursdays and Sundays having the highest percentage of delays and weekday mornings having the lowest percentage. Summer has the highest proportion of delays while late Autumn and early Winter have the lowest proportion.
Weather: Stronger winds and heavier precipitation increase delays, high visibility reduces delays.

Conclusion

While this has shown the contributing factors to delays, it was not established which of them were causations and which were simply consequences, e.g. are airlines experiencing delays because airports are poorly operated or is poor airport performance the result of the inefficiency of the airlines that use them? This effectively highlights a limitation of both the dataset provided and the “cold” analysis process. There are numerous other factors both qualitative (e.g. pilot experience) and quantitative (e.g. plane turn around times) that can contribute to delays that are not included and may simply not exist. A lack of domain specific knowledge limits how much information can actually be obtained.

While it is impossible not to appreciate that this information is useful it is only part of the story. In order to fully establish which choices a potential passenger should make, a more detailed analysis will need to be completed. This, however, provides a good starting point and will, even without additional work, prove useful in helping passengers make “smarter” decisions.

Appendix A – Dataset overview

Airlines: This is a mapping data set contacting details of the airline's carrier code along with its full name.

Airports: This contains the mapping of the FAA airport code to its full name along with providing geographic location and time zone details.

Flights: This contains the details of all flights, including the origin and destination along with the distance between them, the departure and arrival times, both scheduled and actual, along with the flight time, the operating airline and aircraft tail number.

Planes: This dataset provides a mapping between an aircraft's tail number and the aircraft's details. These details include make and model of the aircraft, year of manufacture, passenger capacity, number and type of engines and speed.

Weather: This provides hourly meteorological reading taken at each of the 3 destination airports for all of 2013. The readings taken include temperature, dew point, humidity, wind (speed and direction), precipitation levels, pressure, and visibility.

Further details of the data sets can be found here: https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf (https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf)

Appendix B – details of data corrections

Flights

- A number of flights showed no departure or arrival time. A stop check of these showed that these flights were canceled and so an additional variable was added to the dataset to highlight this.
- A number of flights showed a departure time but no arrival time. A spot check showed that these flights were diverted and as above an additional variable was added to the dataset to highlight this
- A number of flights had both departure and arrival times but did not show an arrival delay. Investigations showed that these flights were diverted but ultimately arrived at their destinations, the arrival delay was calculated and added while a marker was added to reflect that a diversion had taken place.

Weather

- A number of observations were missing or their values were anomalous (these anomalies were absolute, e.g. a wind speed of over 1,000, or relative, e.g. a temperature dropping by c. 44 in an hour and then rising by c. 44 in the next)
- Where these anomalies arose the value was linearly imputed based on the surrounding values.
- Weather data ends at 18:00:00 on the 30th of December 2013. We could not impute weather data for flights that left after this time.

Airports

- A number of airports had incorrect geographical co-ordinates. These were imputed using online sources.
- A number of airports were missing. These were imputed using online sources.

Further details can be found in "Imputation.R"

Appendix C – departure delay model details

This is a bionimal generalised linear model designed to idenify the probability of being delayed by more than 15 minutes. Outlines below are the predictors used in the model along with how they affect the prediction.

Departure Hour: An additional hour in departure time after 5am, keeping all other predictors fixed, changes the odds of getting a departure delay of more than 15 minutes by a factor of 1.148665 which is 14.87% increase.

Wind Speed: A 1 mph increase in wind speed, keeping all other predictors fixed, changes the odds of getting a departure delay of more than 15 minutes by a factor of 1.026618 which accounts for a 2.7% increase.

Pressure: A 1 millibar increase in pressure, keeping all other predictors fixed, changes the odds of getting a departure delay of more than 15 minutes by a factor of 0.9782305 which is 0.22% decrease.

Temperature: An increase of 1 degree Farenheit in temperature, keeping all other predictors fixed, changes the odds of getting a departure delay of more than15 minutes by a factor of 1.005187 which is 0.05% increase.

Humidity: A unit increase in relative humidity, keeping all other predictors fixed, changes the odds of getting a departure delay of more than 15 minutes by a factor of 1.014443 which is 1.4% increase.

Plane's Age: A unit increase in plane's age, keeping all other predictors fixed, changes the odds of getting a departure delay of more than 15 minutes by a factor of 0.9957949 which is 0.04% decrease.

Seats: A unit increase in seats on a plane, keeping all other predictors fixed, changes the odds of getting a departure delay of more than 15 minutes by a factor of 0.9981417 which is 0.02% decrease.

On testing this model showed accuracy of 78.86% for the training dataset and accuracy of 78.93% for the testing dataset.

Appendix D - arrival delay model details

Even though,k-cross-validation results have a high accuracy rate (92.34%) and MSE value of 299.80. Modeling assumptions had failed in this experiment.

Thus, multiple linear regression failed to predict the delay of US flights. A future analysis using non-parametric methods may be conducted to carry out the estimation of delays flights departing from NYC, for instance decision trees, random forests can be used in this matter.

Summary of key predictors

Note: Not all the models predictors are outlined here just the most numerically significant ones

All else being equal each 1 mile increase in visibility range will reduce the arrival delay be c. 1 minute and 13 seconds.

All else being equal by flying with the below Airline over AirTrans Airway will give the respective change in the estimated arrival delay

Airline	Delay
Alaska Airlines	-13.4 minutes
American Airlines	-7.3 minutes
Delta Air Lines	-6.5 minutes
Endeavor Air	-7.3 minutes
Envoy Air	37.1 minutes
ExpressJet Airlines	-4.4 minutes
Frontier Airlines	1.3 minutes
Hawaiian Airlines	-11.84 minutes
JetBlue Airways	-2.4 minutes
Mesa Airlines	-5.1 minutes
SkyWest Airlines	8.1 minutes
Southwest Airlines	-7.8 minutes
United Air Lines	-7.5 minutes
US Airways	-1.8 minutes
Virgin America	-8.9 minutes

All else being equal by flying with the below Airport rather than JFK will give the respective drop in the estimated arrival delay

Airport	Delay
La Guardia	25 seconds
Newark Liberty Intl	30 seconds

All else being equal by flying in the below quarter rather than Q1 will give the respective drop in the estimated arrival delay

Quarter	Delay
Q2	3.4 minutes
Q3	3.4 minutes
Q4	2.9 minutes

Further details and unused plots can be found in “compilation.rmd”