ST662 Group Project

Susan Edgeworth, Pengyu Yang, Jack Francis Hickey, Aaron John Doyle, James Doherty Ferris

Introduction

This report will look at the `nycflights13` dataset in R, this package contains data on flights into and out of the three main airports, Newark, JFK and La Guardia, serving New York for every day in 2013.

The data is a collection of recorded values (~330,000 across 5 tables) from a variety of sources there are a huge amount of missing data points to deal with. Missing data is dealt with on a case-by-case as the analysis is being done.

In generating visualisations we have joined data from the airlines, airports, planes and weather data frames to seek insights. First we explore airlines to look at their punctuality. This leads us on to explore delays and what factors have influence, we map airports with punctuality data, before looking at weather and finally turning to the planes themselves to see if plane size or age has an influence.

Methods

This project will assess both departure delays and arrival delays. Exploratory data analysis on `nycflights13` package was done subsetting out some fundamental variables in order to further investigate.

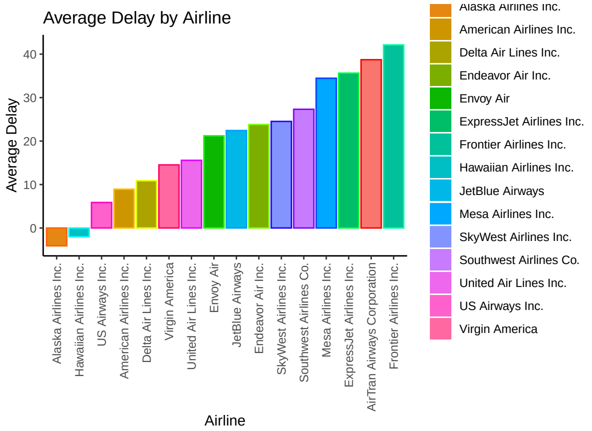
The dplyr package was used for data manipulation, ggplot2 for data visualisation and some color scheme packages to optimise plots. Spatial visualisation used the simple features package and geographic coordinates provided in the data to generate an interactive map using plotly.

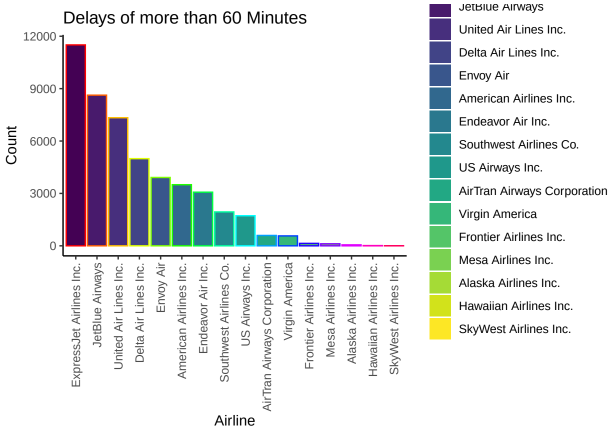
Further statistical techniques used to assess delays/on-time flights included: correlation plots, comparative loess curves, and general observations alongside visual analysis.

Airline delays and Analysis

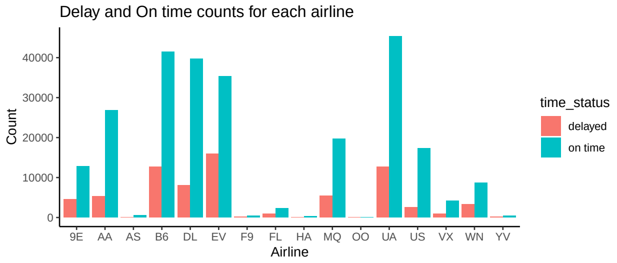
The first plot (below, left) shows the average total delay (Arrival delay + Departure Delay) of each airline. Alaska Airlines is the most punctual. A question arises, does operating more flights increase the liklihood of delays?

By filtering the data to asses delays of over 60 minutes the plot on the right shows Alaska Airlines is still in the top three with 56 flights that had over an hour delay, compared to ExpressJet Airlines with 11503 flights with over an hour delay.



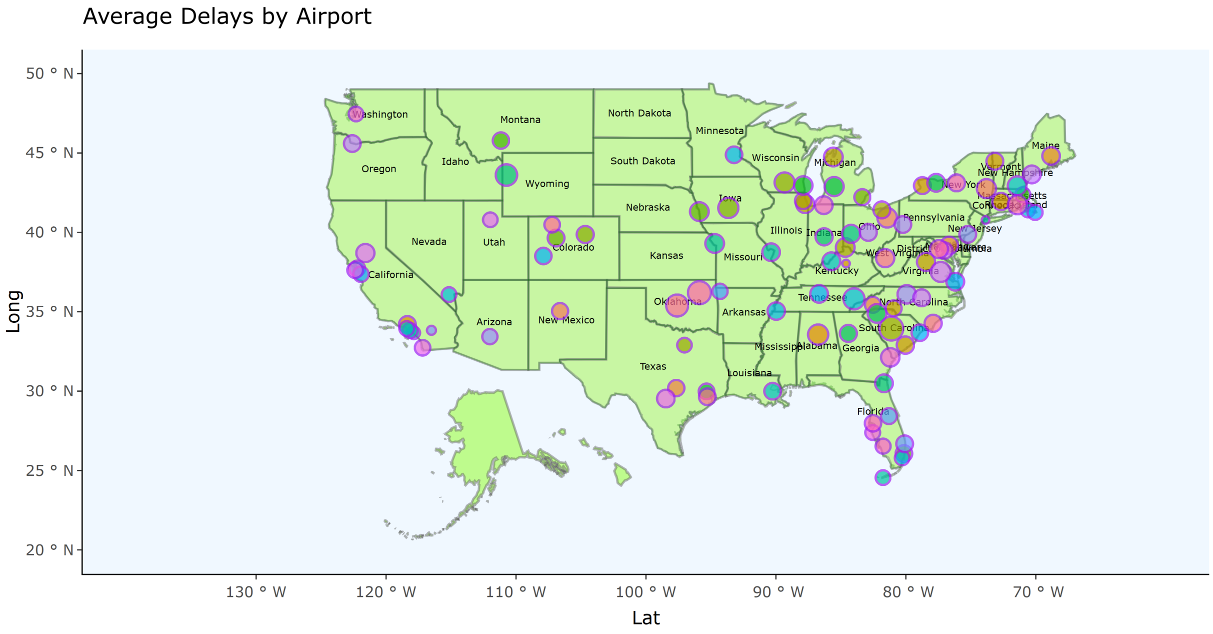


The plot below shows the overall picture for punctuality across all airlines in the three airports serving NY.



Mapping Airport Delays

To further explore this data, a map was created by joining the latitude and longitude information from the airports dataset. The map shows all airports in the US, a larger dot represents a bigger delay. To view the interactive plotly version click: https://rpubs.com/suedge12/757742

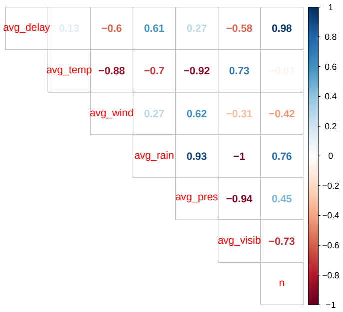


Potential Delay causes

In order to further analyse flights delays, the impacts of weather, general flight congestion and the variety of planes were considered.

Weather vs congestion

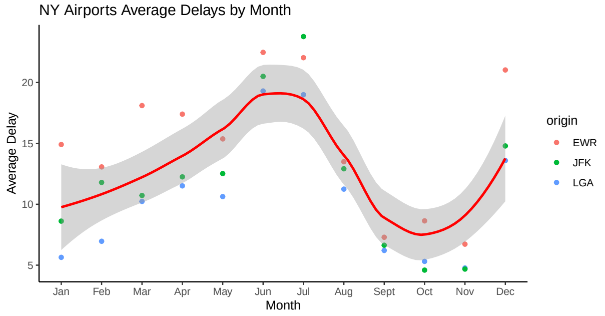
A correlation plot is used to test the contribution weather and congestion have on average delays, and each other. Average values were calculated to provide a general overview of these variables.

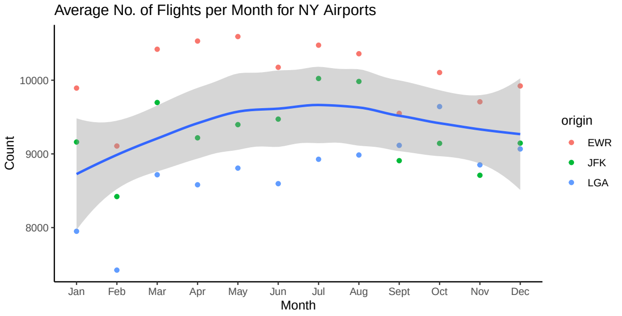


The correlation plot shows a very strong relationship (0.98) between average delays and number of flights, suggesting congestion can be a big part in delays. Some weather values are linked with delays: wind (-0.6), rain (0.61) and visibility (-0.58), these factors could play a part. The number of flights is the strongest contributor. Temperature has little to no correlation with the number of flights (-0.07), suggesting the New York airports are adept at dealing with all weathers.

Congestion

Congestion seems to have an adverse affect on punctuality, so this will be explored further. Analysis will compare average delays per month with the number of flights per month leaving New York airports.

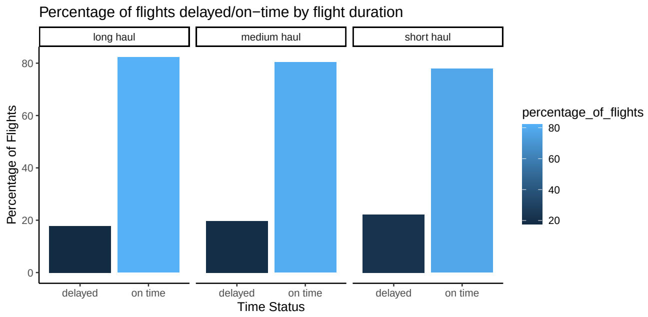


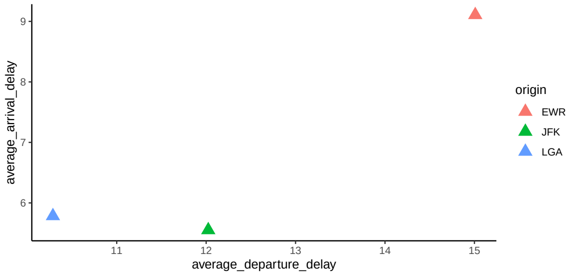


Largest delays occur in the busiest months in terms of flights (June and July) with the trends matching for the rest of the year except November and December where delays increase. Presumably with New York being a 'commuter city' the holiday season is to blame here (Thanksgiving and Christmas) with people going home. So although the number of flights leaving New York isn't an exceptionally high number, people traveling in these months all travel around the same few days, causing major congestion delays in clusters.

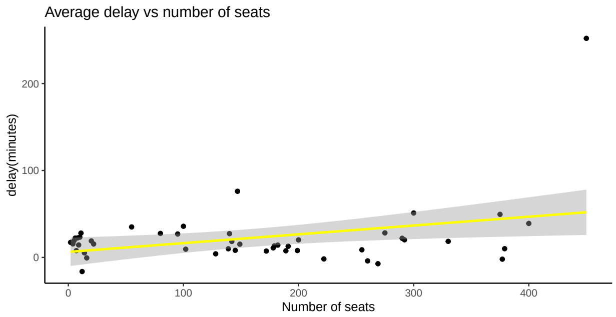
To uncover whether any time is made-up by pilots along journeys flights were categorized as short (0-3 hours), medium (3-6 hours) or long haul (6+hours) and delays were analysed to find disparaties. Long haul flights had better punctuality, both in terms of on time arrivals and also lower numbers being delayed.

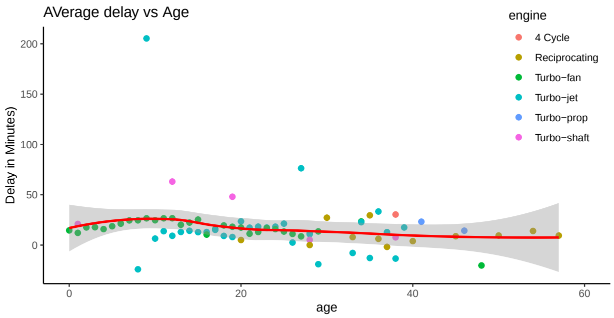
Perhaps longer flight times allow an opportunity to make up time in the air. Looking at specific airports, Newark has the worst departure and arrivals delay record.





Turning finally to the planes themselves, does their size, or age have an influence on delays? The data shows that as the number of seats increases, so too do the delays. Prompting the question is it the slowness of passengers boarding/disembarking being underestimated by airlines thus causing a backlog? Or is the larger amounts of baggage slower to get on the planes hereby slowing the process? It may be a combination of these, but would need further examination. When the analysis looks at age, the result is less conclusive.





Conclusion

This report has explored the `nycflights13` dataset, what can be drawn from this is the following: Newark is the busiest airport, it suffers more departure and arrival delays than JFK or La Guardia. Weather should not be a major concern as wind, rain, humidity don't appear to influence punctuality as much as congestion does. No NYC airport stands out in terms of more adverse weather conditions, this is to be expected given their proximity to each each (within 20 miles). Furthermore it seems a smaller plane will suffer less delays opening the question of are passengers to blame. Finally, if you are flying from NYC to Blue Grass, Palm Springs or John Wayne airports you will most likely arrive ahead of time.