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 aim is to explore the data to see what picture emerges.

As the data is mostly a collection of recorded values (~330,000 across the 5 tables) from a variety of sources there is, as expected, a huge amount of missing data points to deal with. It seemed wasteful to fully omit these observations, so 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 has an influence.

Methods

The main analysis in this project will assess both departure delays out of NYC and subsequent arrival delays in destination airports. Exploratory data analysis on `nycflights13` package was done subsetting out some fundamental variables in order to further investigate.

This project makes use of the dplyr package for data manipulation, ggplot2 for data visualisation along with some additional color scheme packages to optimise plots. Some spatial visualisation is done using the simple features package as well as using geographic coordinates provided in the data. This was used in a plotly graph to implement some interactive analysis.

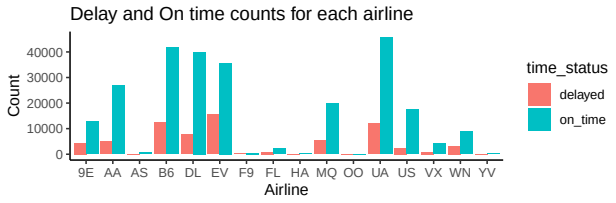
Further statistical techniques provide some detailed explanations towards justifications for delays/on-time flights including: correlation plots and comparative loess curves as well as general observations alongside visual analysis.

Airline delays and Analysis

The first plot (below, left) shows the average total delay (Arrival delay + Departure Delay) of each airline. Alaska Airlines appears to be the most punctual, more analysis needs to be done here to assess the time delays as there are more flights recorded for certain airlines and apparent outliers which may skew the average.

The plot on the right shows where the total delay is more than an hour. Alaska Airlines still comes out 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 airports that have the most delays on arrival, a larger dot represents a bigger delay. To view the interactive plotly version please follow this link: https://rpubs.com/suedge12/757742

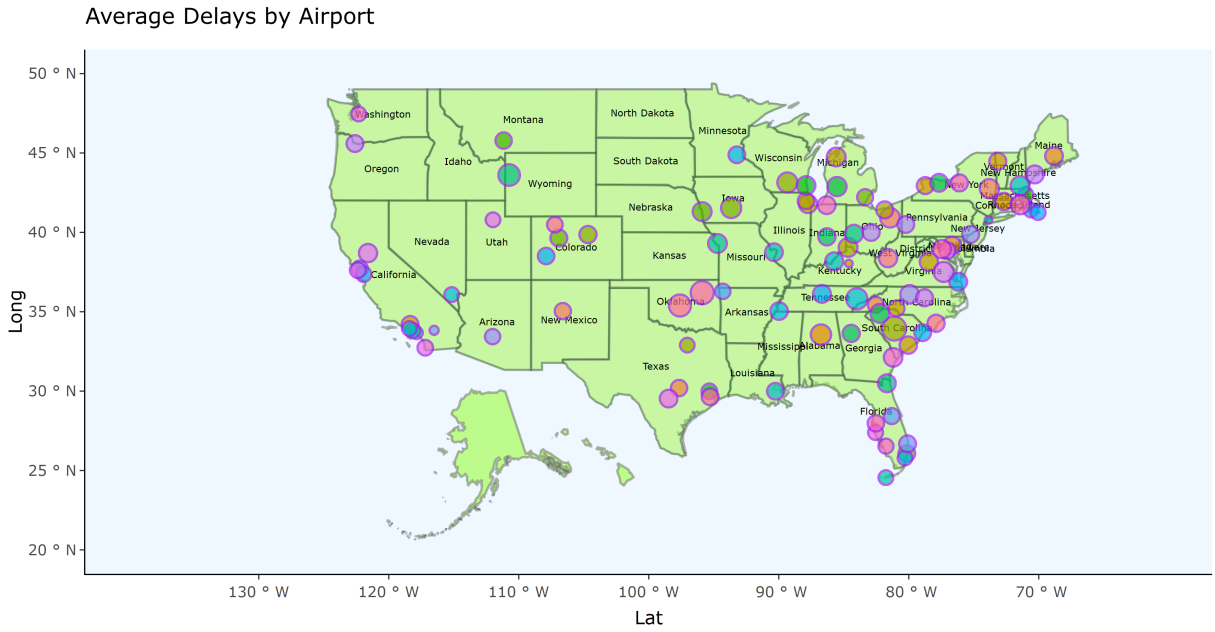


Figure 1: Flight Delays in US airports

Best airport punctuality performers

Allied to the map above, the top 6 performing airlines, when ranked by punctuality are:

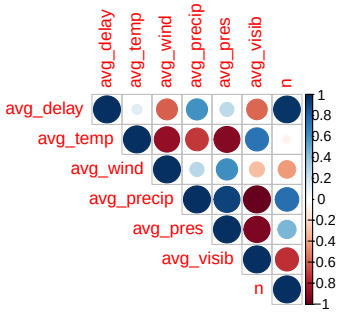
dest	avg_delay	name
LEX	-31.000000	Blue Grass
PSP	-15.666667	Palm Springs Intl
SNA	-1.087438	John Wayne Arpt Orange Co
MVY	6.604762	Martha\\'s Vineyard
HNL	7.950071	Honolulu Intl
DFW	8.925966	Dallas Fort Worth Intl

Potential Delay causes

Having looked at airports, we'll now turn to weather, congestion and planes to see what impacts these have.

Weather

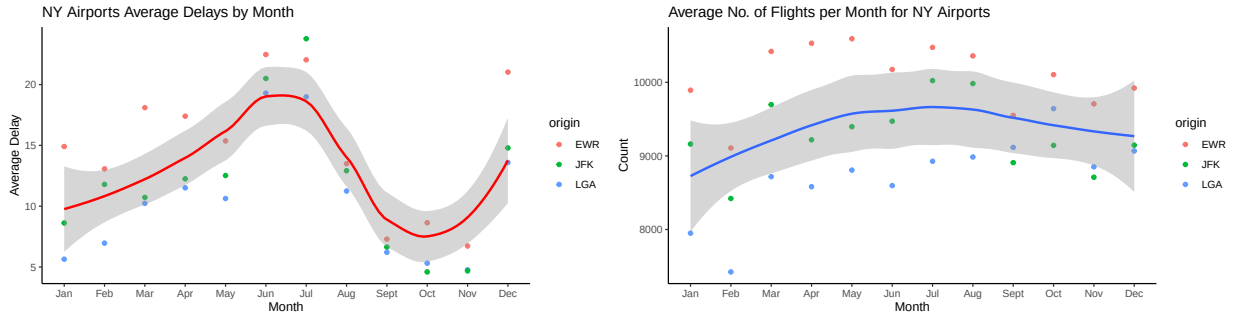
To explore weather as a factor the three Origin airports were grouped and a calculated average value for the weather data was created. It is possible to test correlations between these values and determine contributing weather factors towards departure delays from these airports. Also calculated is the correlation between the number of flights leaving each airport with the departure delays.



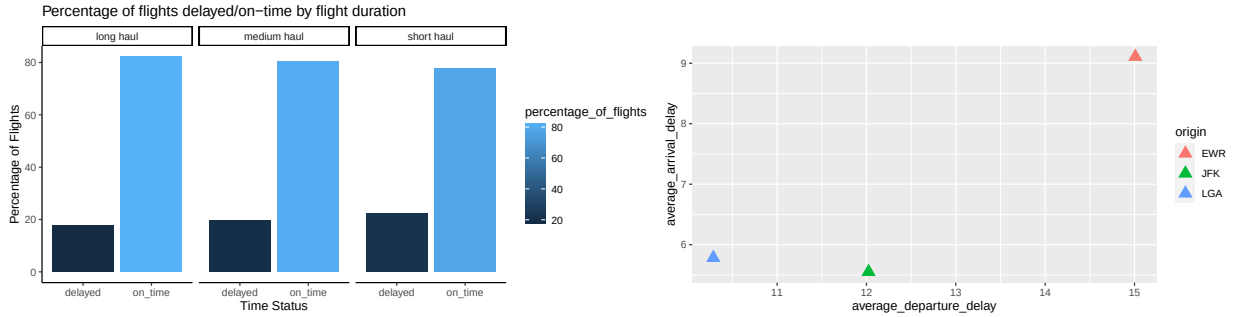
In terms of average delays per flight Newark is almost 3 minutes more delayed than JFK who in turn is about 2 minutes per flight worse off than La Guardia. A correlation plot is used in order to uncover causes of this. The plot above shows a very strong correlation between number of flights and average delays (~0.979). The weather values don't possess such strong correlations, precipitation being the next strongest (~0.61). No New York airport stands out in terms of more adverse weather conditions, this is to be expected given their proximity to each each (within 20 miles).

Congestion

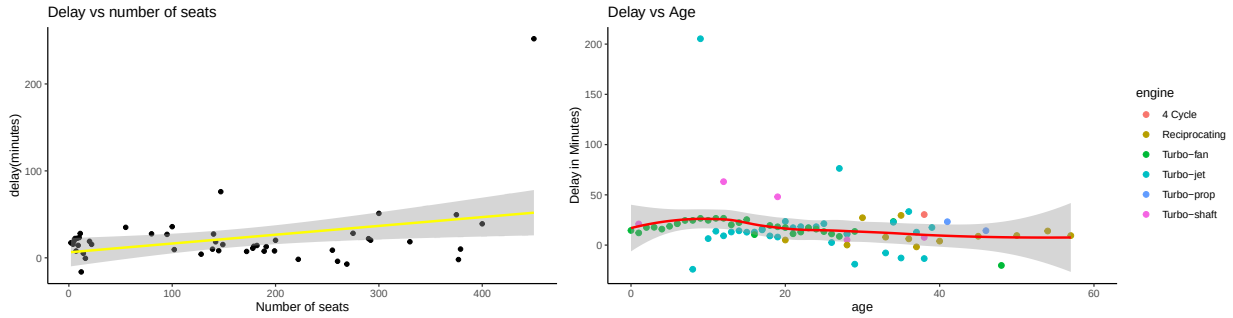
As noted above, congestion could have an adverse affect on punctuality, so here we explore that a little further. To determine this average delays per month will be compared with flight numbers per month. Delays are biggest in June and July and this coincides with high flight numbers at the three airports.



Exploring this further, flights were categorised as short, medium or long haul. Long haul flights had better punctuality, both in terms of on time arrivals and also lower numbers being delayed. Perhaps the longer flight times enabled more opportunity to make up time in the air. Looking to the individual airports, Newark appears to have the worst departure and arrivals delay record.



Turning finally to the planes themselves, does their size, or indeed age have any influence on delays? Looking at the plot on the left, as the number of seats increases, so too does the delay. Turning to age, the result is less conclusive.



Conclusion

This report has explored the `nycflights13` dataset, what we can draw 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. A smaller plane will suffer less delays. If you are flying to Blue Grass, Palm Springs or John Wayne airports you will most likely arrive ahead of time.