# ST662 Group Project

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

## Introduction

This report will look at the NYC Flights 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. Our aim is to explore the data to see what picture emerges.

As the data is mostly a collection of recorded values from a variety of sources there is, as expected, a huge amount of missing data points (~330,000 across the 5 tables) 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.

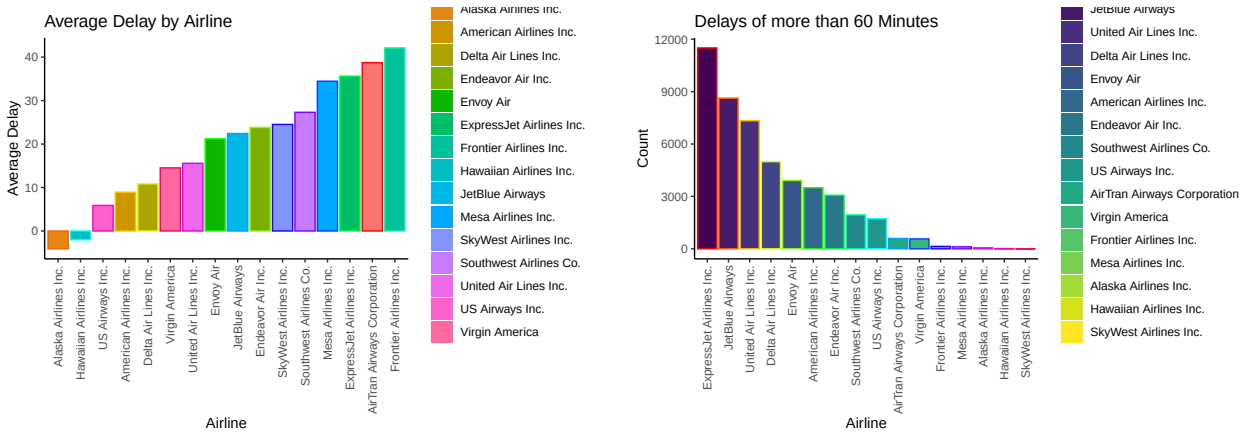
## Methods

In preparing this report we have used the following functions within R: mutate, pipe, join, facet wrap, select, group by, correlation. We've also used the latitude and longitude information from the airports dataset to generate a map of airports and their performances vis a vis delays. We have used the following packages: • library(nycflights13) • library(dplyr) • library(ggplot2) • library(plotly) • library(RColorBrewer) • library(tidyverse) • library(sf) • library(hrbrthemes) • library(viridis)

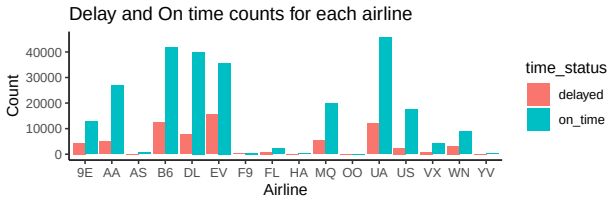
## Airline delays and Analysis

The first plot 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.




This plot shows a count of all airlines accross all three airports




## Mapping Airport Delays

This map looks at the airports that have the most delays on arrival, the larger dot represents a bigger delay (in minutes). https://rpubs.com/suedge12/757742

## Ranking of Airports

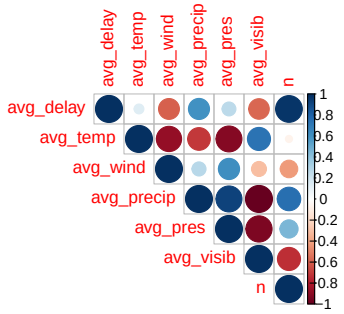
| dest | avg_delay | name |
|---|---|---|
| LEX | -31.000000 | Blue Grass |
| PSP | -15.666667 | Palm Springs Intl |
| SNA | -1.087438 | John Wayne Arpt Orange Co |
| MVY | 6.604762 | Martha\\'s Vineyard |
| HNL | 7.950071 | Honolulu Intl |
| DFW | 8.925966 | Dallas Fort Worth Intl |

## Potential Delay causes

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

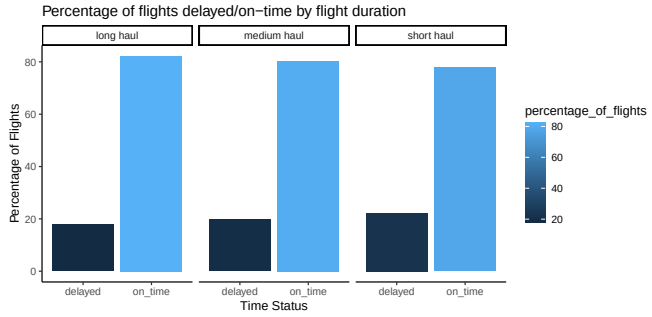
## Weather

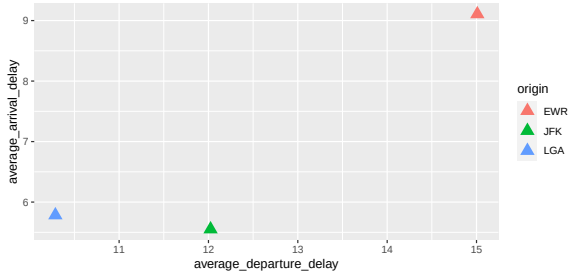
In looking at weather as a factor what we've done is group by the three Origin airports for New York and calculated the average values for the weather data. It is possible to test correlations between these values and determine contributing weather factors towards departure delays from these airports. 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, to be expected given their proximity to each each (within 20 miles).

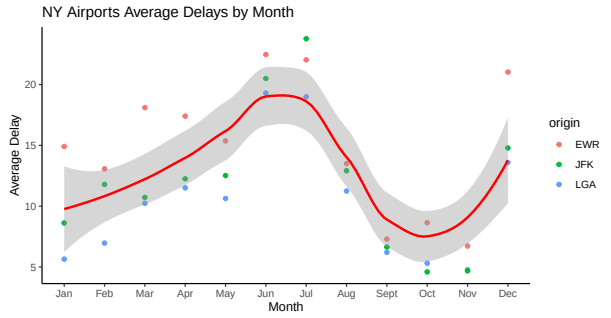
Is there an airport who's punctuality out performs?

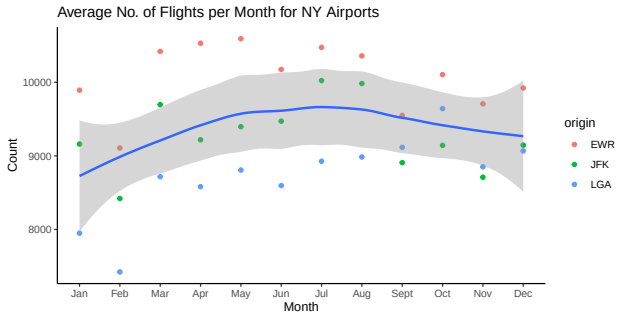



## Congestion

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




We wondered did flight time have an impact, were delays different between short haul, medium haul or long haul flights? From the analysis there is some evidence that short haul flights were less punctual.

## Planes Analysis

In considering congestion, we wondered does the plane itself have any bearing, the number of seats, the age, the manufacturer? In looking at the number of seats, we noticed that as the number of seats increased, so too did the delays, the results from aircraft age are less conclusive.