Note: Output Designed for HTML

Tristan Macy

7/7/2021

## Introduction

This project will focus upon the *nycflights13* package, which contains a variety of datasets that contain information regarding flights originating from New York City. The full dataset and its documentation can be found on <https://www.rdocumentation.org/packages/nycflights13/versions/1.0.1>. This analysis will rely heavily on the flights and airports datasets to draw meaningful conclusions regarding travel and tourism. Questions this analysis will seek to answer include:

* How do departure delays (median and average) compare across the flight dataset?
* What destinations are the most popular among each origin airport?
* From which airport do the most cancellations originate from?
* Does the data include any unique travel destinations for summer months?
* Is there a relationship between departure delays and variables such as wind speed and time of day.

## New York City Airports

The flights dataset utilizes special codes to denote each airport. To begin the analysis we will use the airports dataset to find the names of all of the airports within New York City using their individual codes.

Airport Codes

Airport Names

EWR

Newark Liberty Intl

JFK

John F Kennedy Intl

LGA

La Guardia

We will now seek to compare Newark Liberty International AirportJohn F Kennedy International Airport and La Guardia airport to see how there average and median departure delays compare to each other.

## Departure Delays

Airport Codes

Average Departure Delays

Median Departure Delays

EWR

15.10795

-1

JFK

12.11216

-1

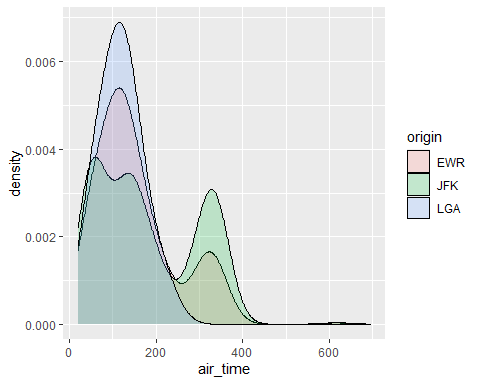
LGA

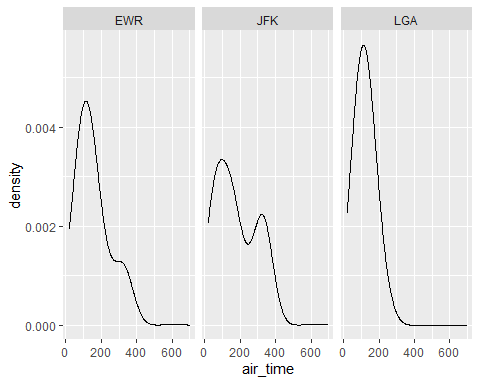
10.34688

-3

We see that the highest average delay is within **EWR** with 15.10795, followed by **JFK** with 12.11216, and **LGA** with 10.34688 minutes. The highest median delay for each airport are actually early departures with the earliest median being **LGA** at -3 with **JFK** and **EWR** sharing a -1.

## Air Time Distributions

Given that La Guardia is not an international aiport, we will see if the distribution of air time is reflective of smaller overall air times for this airport, 



As expected, La Guardia airport fields the majority of its flights under 200 minutes. The international airports have a similiar spike spikes between 300 and 400 minutes. The JFK airport appears to carry out the most of the longer duration flights (>300 minutes) between the three. Despite this, the majority of flights from all 3 airports are shorter in duration.

## Top Destinations

Next we will determine the largest amounts of flights going to destination airports and compare their amount by their origin airports.

Airport Code

Airport Name

ATL

Hartsfield Jackson Atlanta Intl

BOS

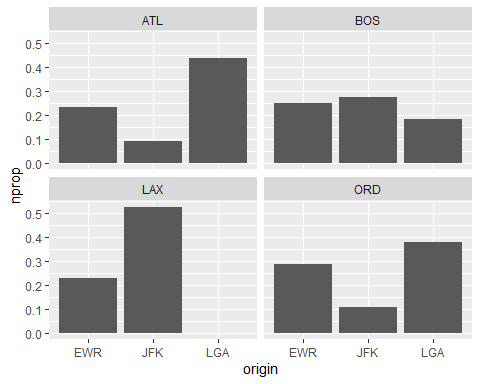
General Edward Lawrence Logan Intl

LAX

Los Angeles Intl

ORD

Chicago Ohare Intl



Each panel represents the proportion of flights taken from the three NYC airports to one of the top four destinations in 2013. Interesting to note that there were no flights originating from LGA to LAX, likely due to its non-international status.

### Summer-Only Destinations

While such flights were flown across the year, we will now see if there were any flights only taken within the summer months to isolate tourist destinations.

Airport Code

Airport Name

ANC

Ted Stevens Anchorage Intl

LGA

La Guardia

TVC

Cherry Capital Airport

These three destinations in Alaska, New York and Michigan were only travelled to within the 2013 summer months.

## High Altitude Destinations

These are the number of flights taken with altitudes of over 6000 ft in 2013.

Destination

Flights From NYC

EGE

213

HDN

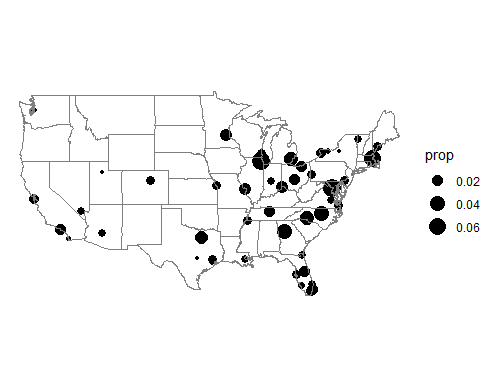
15

JAC

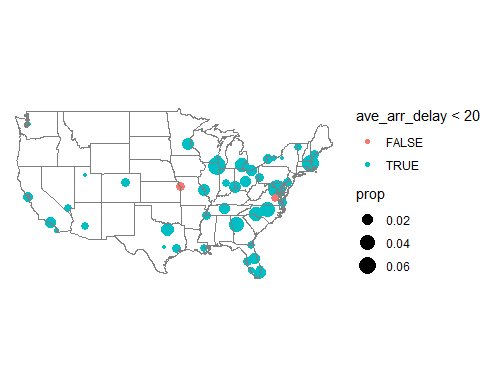
25

## Cancellations and Destination Location

Since the data contains the names of various names of cities we can use a U.S map to visually look at each airport’s proportion of cancellations.



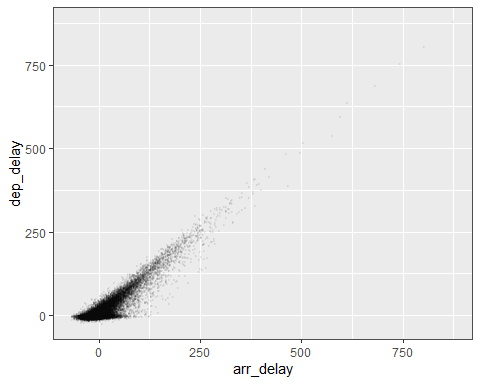
The majority of the airports are located either in the midwest or east of the U.S. Notably large proportion of cancellations occur within the Chicago area, the Massachussets area and within the area surrrounding Washington D.C. Less populated cities appear to have less cancellation proportions compared to that of the larger populated ones.

Lets see if what we can learn if we add a color dimension with *ave\_arrival delays* showing low and large. 

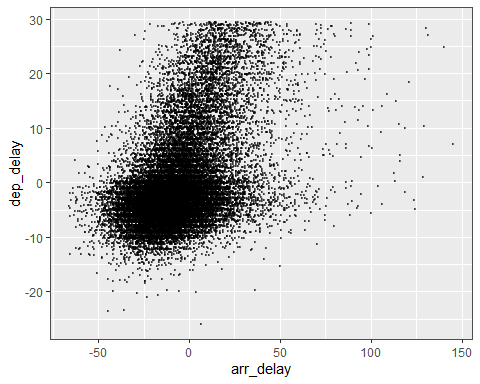
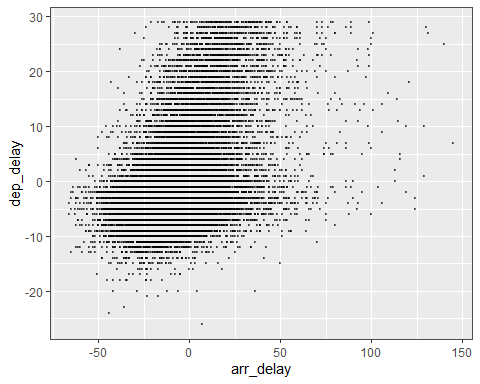
It appears that all but two aiports (MCI and RIC) maintain average *ave\_arrival delays* below 20 minutes. Color provides an easier way to see differences in *ave\_arrival delays* within clusters. It however seems difficult to distinguish how many airports are within clusters of data due to overlapping.

## Arrival and Departure Delays

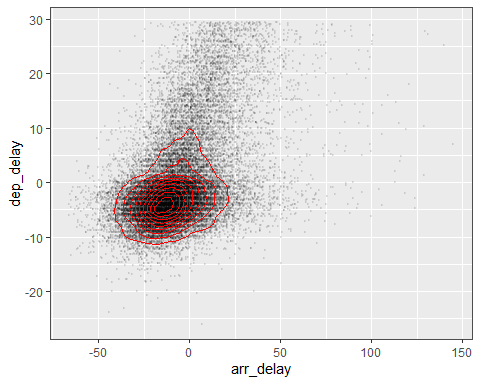
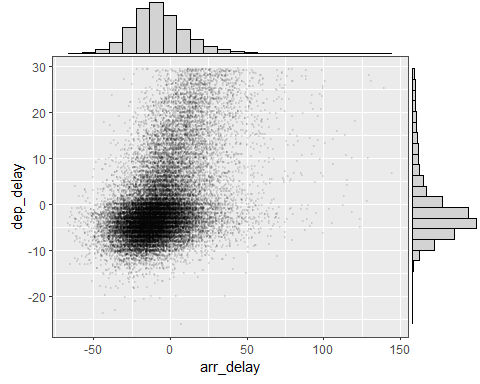
The next part of the investigation will see if there is an associations between Arrival Delays and Departure Delays within airports.



It appears that there is a clear positive relationship between *arr\_delay* and *dep\_delay*, however there appears to be outlier values that make it difficult to avoid overplotting due to the resulting increase in the graphs size. The vast majority of the data appears to be closer to 0.



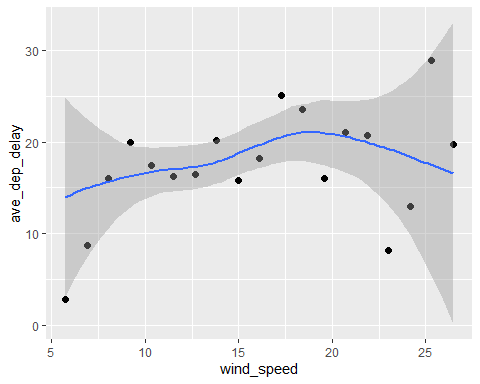
Analyzing the values at *dep\_delay* < 30 we can see that the majority of the values for *dep\_delays* and *arr\_delays* are actually early departures and early delays. This is easier to see in this graph because we had zoomed in on the data points. The addition of jittering also seems to highlight the concentration of values that have negative dep and arr delays.



The addition of the marginal density plot and the contour plot confirms our suspicions that the majority of the values are lower and even negative. We are now able to observe that there appears to be a skew in the data that trails towards the higher values.

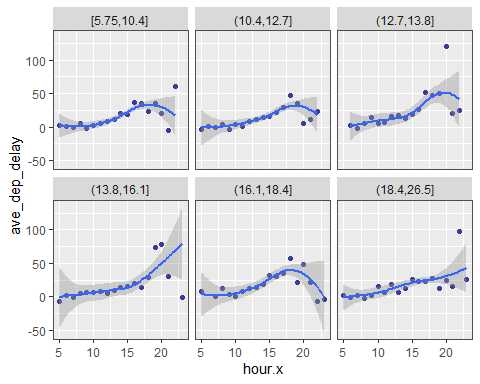
## Wind Speed, Time of Day, and Departure Delays

Is Departure Delay also impacted by the given Wind Speed within the data?



We see a slight increase in *ave\_dep\_delay* as *wind\_speed* increases from the graph.

What about the difference combinations of Time of Day and Wind Speed ? Let us use a coplot of average departure delay against hour, conditioned on wind\_speed,



From the graphs we can see that as *hour* increases, the *ave\_dep\_delay* slightly increases. From the looks of the different *wind\_speed* bins there does not appear to be much difference between the bins.