NYC Flights Lab

Ryan Cheng

2023-10-27

Introduction

This lab explores and visualizes data about flights departing from NYC airports, those being LaGuardia, JFK, and Newark. The data set used in this lab comes from OpenIntro – https://www.openintro.org/data/index.php?data=nycflights – and is comprised of data for 32,735 randomly chosen flights in 2013. Departure and arrival times are relative to the local time zone, that being Eastern Time. Departure and arrival delays are recorded in terms of minutes, with negative values representing earliness. Time in the air is recorded in minutes and distance flown in miles.

This lab will attempt to answer the following questions in hopes of providing travelers with information on when to travel, where to travel from, and what to travel on.

- Which airline has the smallest arrival delay? What about departure delay?
- Which airport has the smallest departure delay?
- What time(s) of the year are departure and arrival delays largest? What about time(s) of the month? Or of the day?

Observations and Analysis

First, let's import the data set. Here's how the first six rows look:

```
nyc.flights.data = read.csv("nycflights.csv")
head(nyc.flights.data)
```

```
##   year month day dep_time dep_delay arr_time arr_delay carrier tailnum flight
## 1 2013     6  30      940        15     1216        -4      VX  N626VA    407
## 2 2013     5   7     1657        -3     2104        10      DL  N3760C    329
## 3 2013    12   8      859        -1     1238        11      DL  N712TW    422
## 4 2013     5  14     1841        -4     2122       -34      DL  N914DL   2391
## 5 2013     7  21     1102        -3     1230        -8      9E  N823AY   3652
## 6 2013     1   1     1817        -3     2008         3      AA  N3AXAA    353
##   origin dest air_time distance hour minute
## 1    JFK  LAX      313     2475    9     40
## 2    JFK  SJU      216     1598   16     57
## 3    JFK  LAX      376     2475    8     59
## 4    JFK  TPA      135     1005   18     41
## 5    LGA  ORF       50      296   11      2
## 6    LGA  ORD      138      733   18     17
```

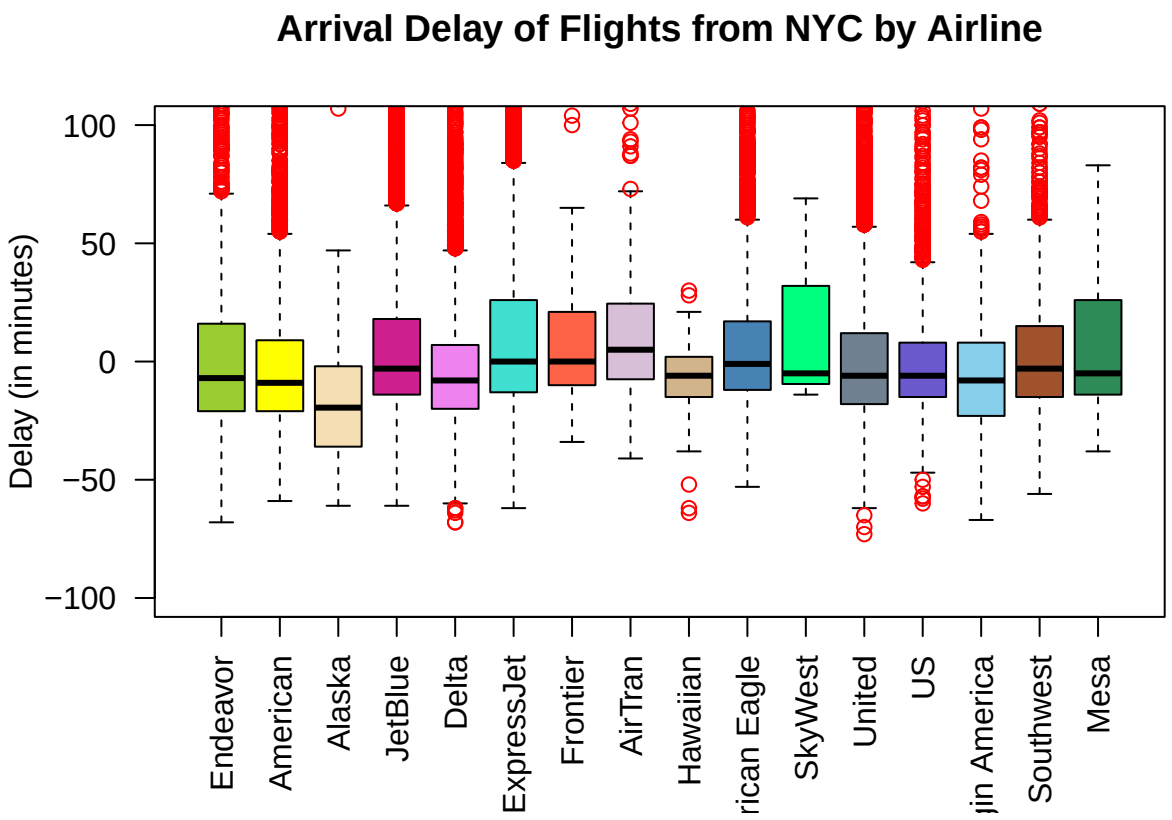
What to Travel On

In the following section, we will explore what airline is best to travel on if you want to minimize departure and/or arrival delays.

Let's first change the class of the variable `carrier` from character to factor. Then let's redefine the levels for easier understanding in the future.

```
nyc.flights.data$carrier = as.factor(nyc.flights.data$carrier)
levels(nyc.flights.data$carrier) = c("Endeavor", "American", "Alaska", "JetBlue", "Delta", "ExpressJet"
```

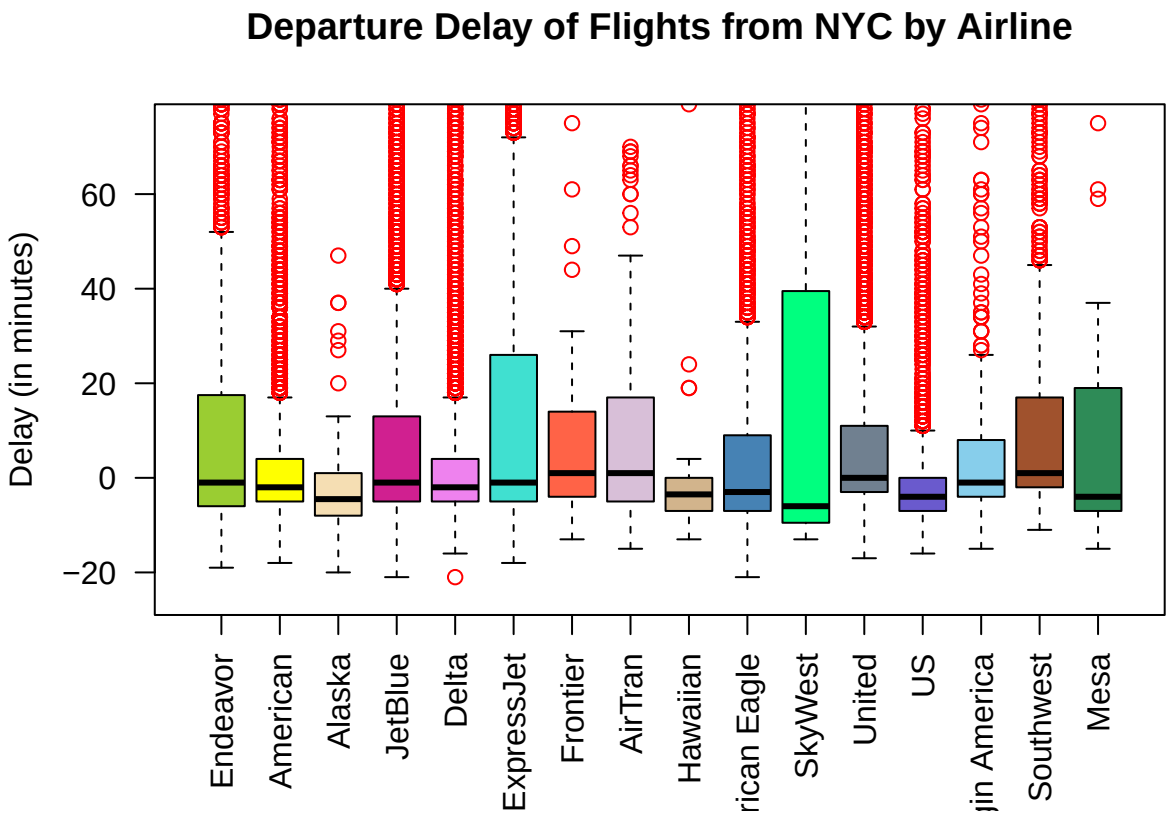
```
boxplot(nyc.flights.data$arr_delay~nyc.flights.data$carrier,
        xlab = "",
        ylab = "Delay (in minutes)",
        ylim = c(-100, 100),
        main = "Arrival Delay of Flights from NYC by Airline",
        col = c("yellowgreen", "yellow", "wheat", "violetred", "violet", "turquoise", "tomato", "thistl
        outcol = "red",
        las = 2
        )
```



This is a comparative boxplot comparing arrival delay in minutes with airline. Alaska has the lowest median delay, while AirTran has the highest median delay. In this case, the median, or the middle value, is a better measure than the mean because it is not affected by outliers. As we can see, there are many outliers with high arrival delays that would increase the mean. However, because the number of these outliers is small in comparison to the number of total flights, the delay of your next flight is probably closer to the median.

```
boxplot(nyc.flights.data$dep_delay~nyc.flights.data$carrier,
        xlab = "",
        ylab = "Delay (in minutes)",
        ylim = c(-25, 75),
        main = "Departure Delay of Flights from NYC by Airline",
        col = c("yellowgreen", "yellow", "wheat", "violetred", "violet", "turquoise", "tomato", "thistl
        outcol = "red",
        las = 2
```

```
)
```



Again, this is a comparative boxplot, this time comparing departure delay in minutes with airline. SkyWest has the lowest median value while Frontier, AirTran, and Southwest share the highest median value.

While these boxplots tell us which airline has the lowest and highest median delays, what if you want to know what the probability of having a flight with a huge delay with a certain airline is? In both boxplots, the median values between airlines are relatively close; as a traveler, you might not care too much about those extra 5-10 minutes. What really matters is if you will have to wait a really long time.

Let's create two new variables in the data frame that checks whether a flight's arrival and departure delays were greater than 60 minutes, respectively.

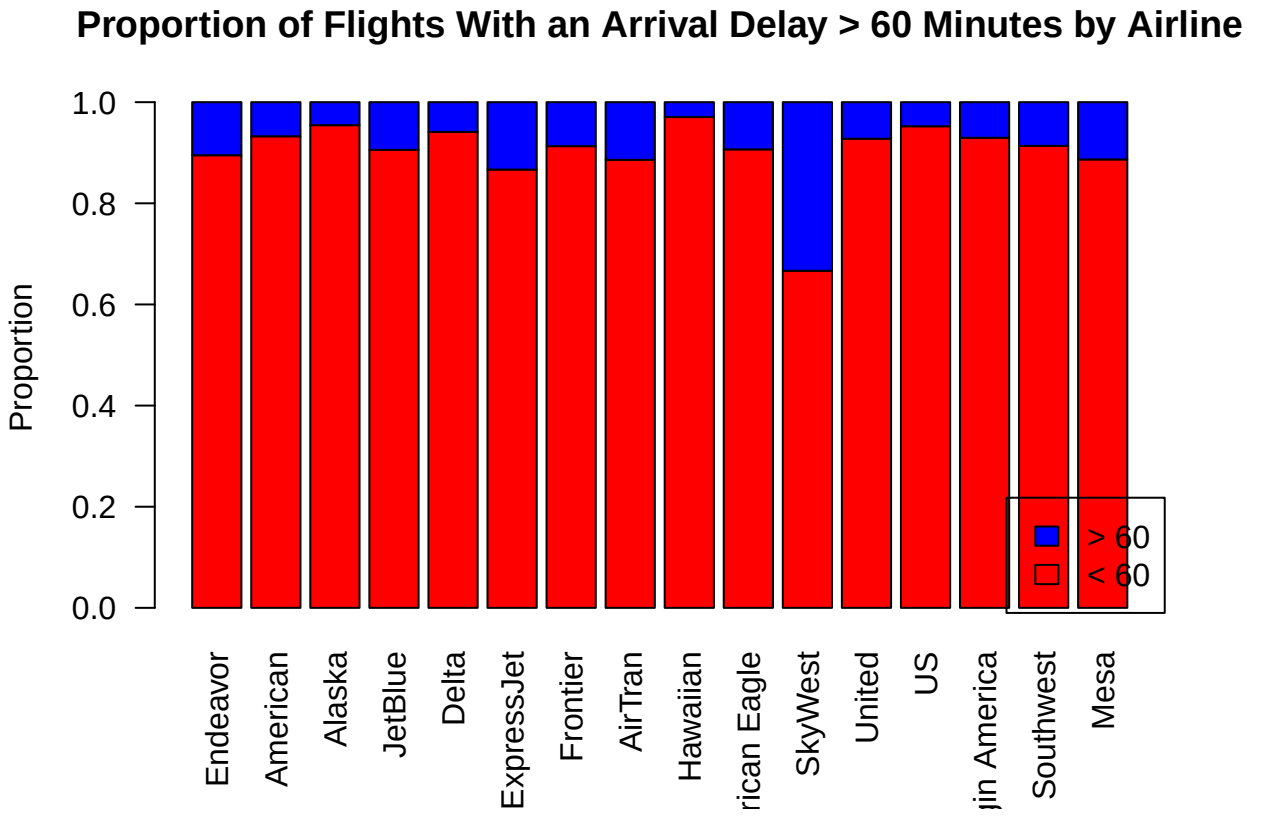
```
nyc.flights.data$arr_delay_above_hour = nyc.flights.data$arr_delay > 60
nyc.flights.data$dep_delay_above_hour = nyc.flights.data$dep_delay > 60
```

And let's also change the class of these variables so that we can rename TRUE and FALSE to > 60 and < 60; it will make legends on graphs more understandable in the future.

```
nyc.flights.data$arr_delay_above_hour = as.factor(nyc.flights.data$arr_delay_above_hour)
levels(nyc.flights.data$arr_delay_above_hour) = c("< 60", "> 60")

nyc.flights.data$dep_delay_above_hour = as.factor(nyc.flights.data$dep_delay_above_hour)
levels(nyc.flights.data$dep_delay_above_hour) = c("< 60", "> 60")
```

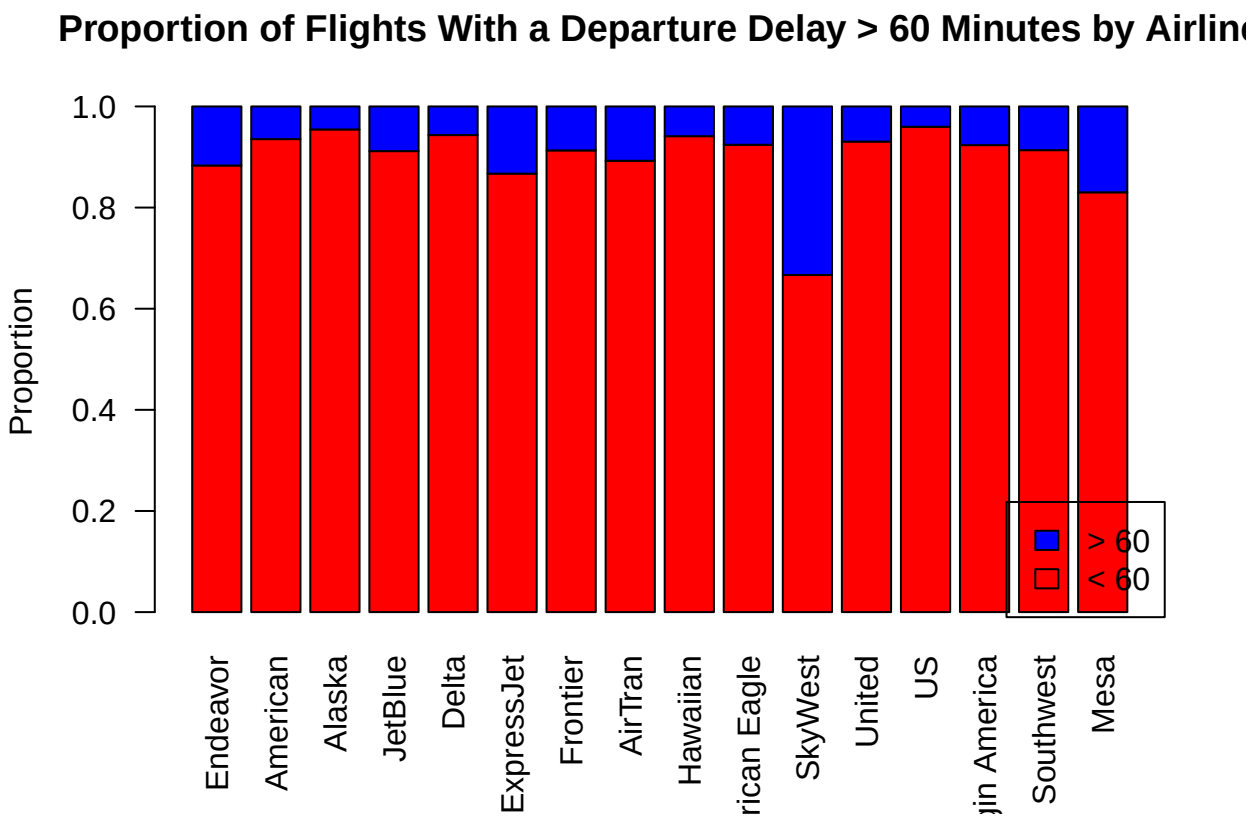
```
barplot(prop.table(table(nyc.flights.data$arr_delay_above_hour, nyc.flights.data$carrier), 2),
        las = 2,
        xlab = "",
        ylab = "Proportion",
        main = "Proportion of Flights With an Arrival Delay > 60 Minutes by Airline",
```

```
        legend.text = TRUE,
        col = c("red", "blue"),
        args.legend = list(x = "bottomright", cex = 1))
```



This is a proportional stacked barplot comparing flights with an arrival delay > 60 minutes with airlines. Hawaiian has the smallest proportion, meaning that a good majority of its flights arrive within 60 minutes of the ETA, and SkyWest the biggest, meaning that a lot of its flights arrive an hour or ore after the ETA. So next time you fly, try to choose Hawaiian and definitely avoid SkyWest!

```
barplot(prop.table(table(nyc.flights.data$dep_delay_above_hour, nyc.flights.data$carrier), 2),
        las = 2,
        xlab = "",
        ylab = "Proportion",
        main = "Proportion of Flights With a Departure Delay > 60 Minutes by Airline",
        legend.text = TRUE,
        col = c("red", "blue"),
        args.legend = list(x = "bottomright", cex = 1))
```



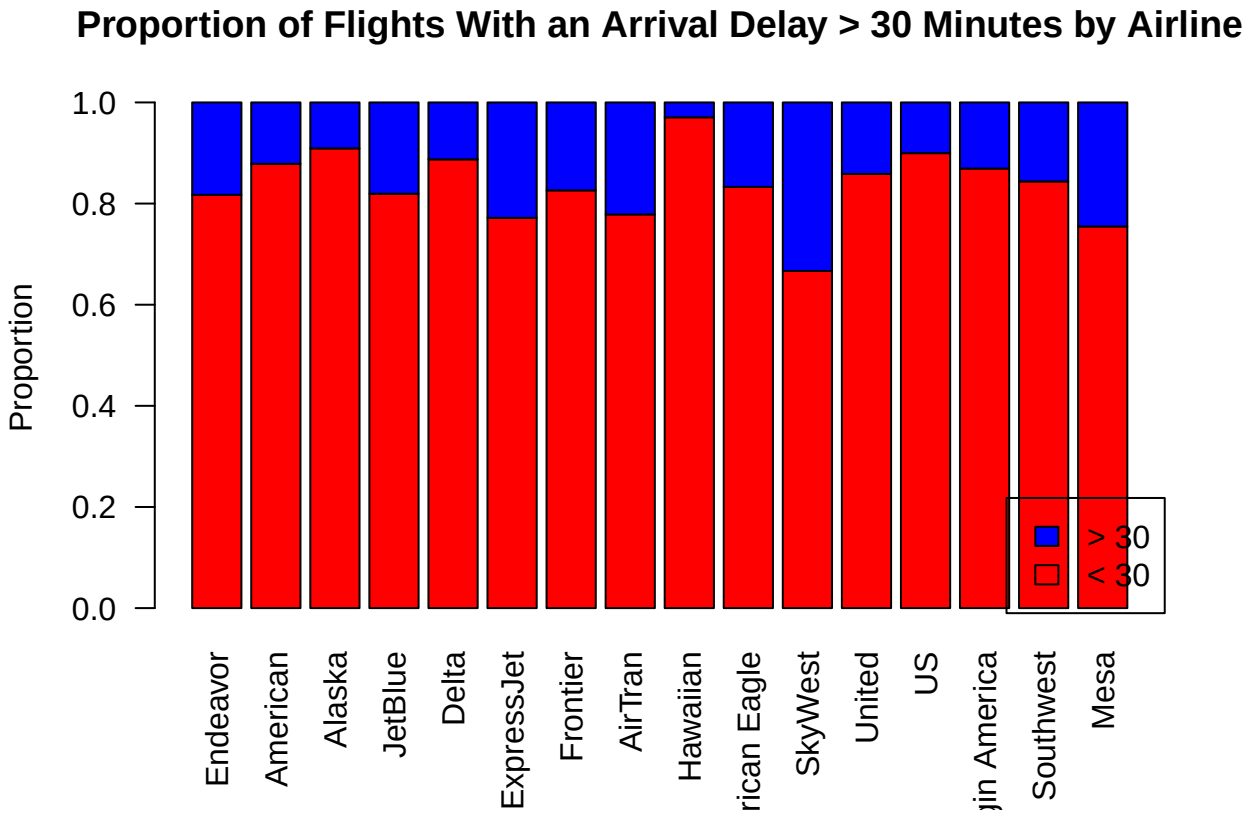
This proportional stacked barplot compares departure delays greater than 60 minutes with airlines. While not as important as arrival delay – a plane can depart late and still arrive on time – it can save a lot of traveler frustration. Here, US has the smallest proportion and SkyWest once again has the biggest proportion. Really, try to avoid SkyWest for an all around smoother travel experience!

But maybe sixty minutes is too long. Maybe you've got to catch a meeting right after you touch down, and so anything above thirty minutes is unacceptable. Or maybe you've got kids who simply can't bear to wait and are extremely restless. Let's see how the best and worst airlines change in response.

We'll create two new variables in the data frame; this time, to check if flight arrival and departure delays are greater than 30 minutes instead of 60. And we'll also rename TRUE and FALSE to >30 and <30 as we did earlier.

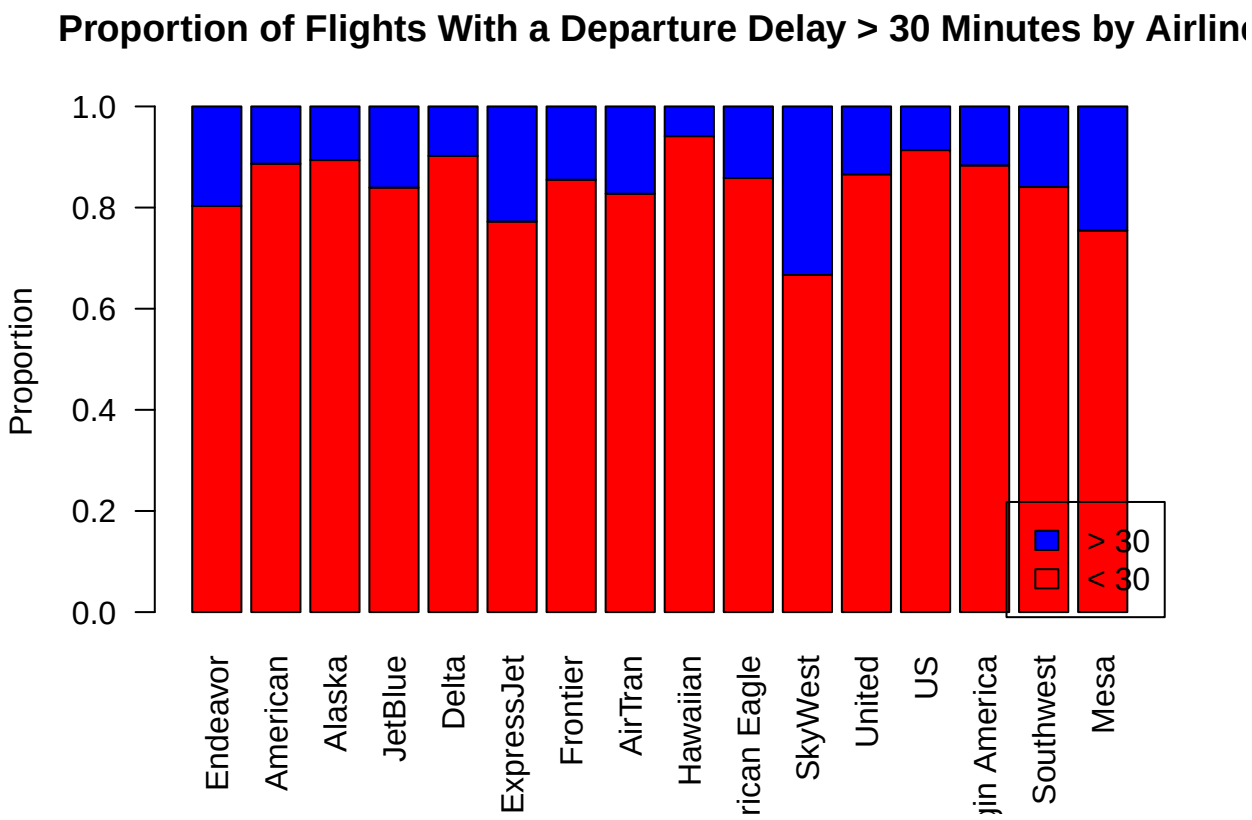
```
nyc.flights.data$arr_delay_above_half_hour = nyc.flights.data$arr_delay > 30
nyc.flights.data$dep_delay_above_half_hour = nyc.flights.data$dep_delay > 30

nyc.flights.data$arr_delay_above_half_hour = as.factor(nyc.flights.data$arr_delay_above_half_hour)
levels(nyc.flights.data$arr_delay_above_half_hour) = c("< 30", "> 30")

nyc.flights.data$dep_delay_above_half_hour = as.factor(nyc.flights.data$dep_delay_above_half_hour)
levels(nyc.flights.data$dep_delay_above_half_hour) = c("< 30", "> 30")
```

```
barplot(prop.table(table(nyc.flights.data$arr_delay_above_half_hour, nyc.flights.data$carrier), 2),
        las = 2,
        xlab = "",
        ylab = "Proportion",
        main = "Proportion of Flights With an Arrival Delay > 30 Minutes by Airline",
        legend.text = TRUE,
        col = c("red", "blue"),
        args.legend = list(x = "bottomright", cex = 1))
```



The airline with the smallest proportion of flights with delays over 30 minutes is still Hawaiian and the airline with the biggest proportion is still SkyWest. What's interesting to note, however, is that while the proportion of flights with delays over 30 minutes compared to 60 minutes increased for almost every airline, Hawaiian's stays practically the same. It seems as though Hawaiian does a really good job with getting its passengers to their destinations on time!

```
barplot(prop.table(table(nyc.flights.data$dep_delay_above_half_hour, nyc.flights.data$carrier), 2),
        las = 2,
        xlab = "",
        ylab = "Proportion",
        main = "Proportion of Flights With a Departure Delay > 30 Minutes by Airline",
        legend.text = TRUE,
        col = c("red", "blue"),
        args.legend = list(x = "bottomright", cex = 1))
```



SkyWest once again has the biggest proportion of flights with some sort of departure delay. This is only further evidence to avoid it. In a twist, however, Hawaiian snatches from US the smallest proportion of flights with a departure delay greater than 30 minutes.

In short, try to fly Hawaiian and try not to fly SkyWest.

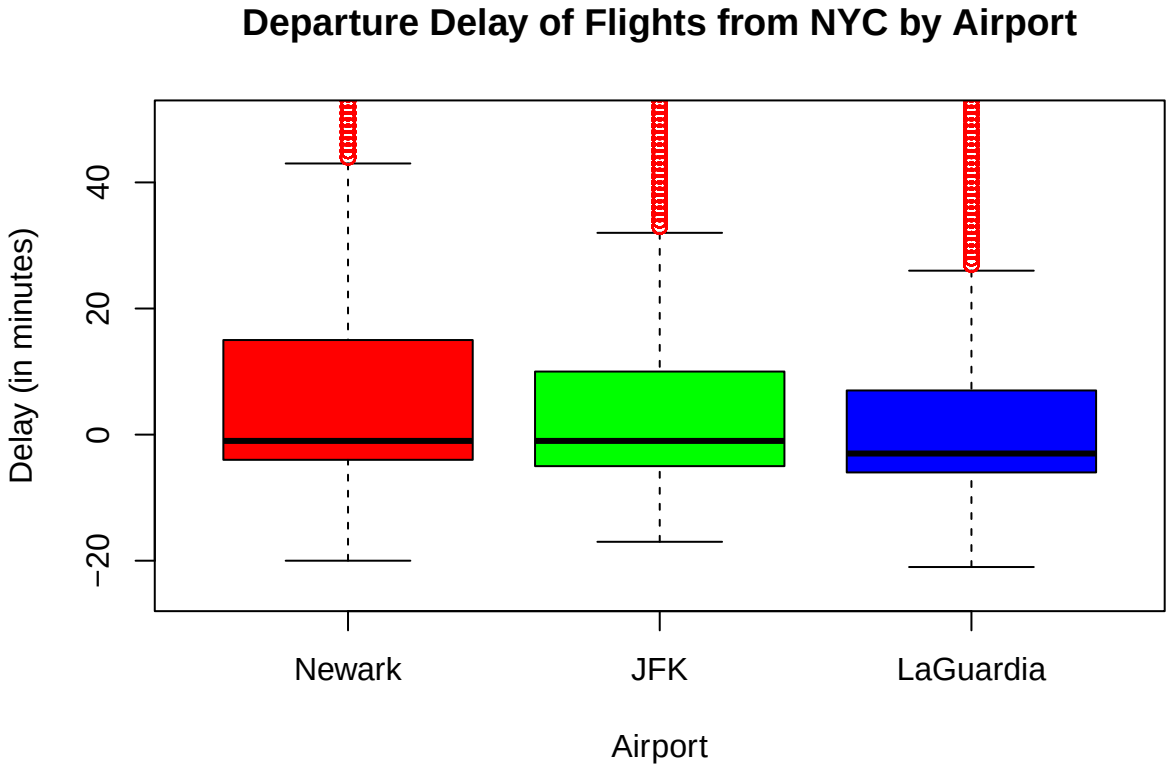
Where to Travel From

Now let's take a look at which one of the three NYC airports – JFK, LaGuardia, and Newark – has the smallest departure delay.

First, let's change the variable `origin` to be a factor so that we can rename its levels to be more understandable than raw airport codes.

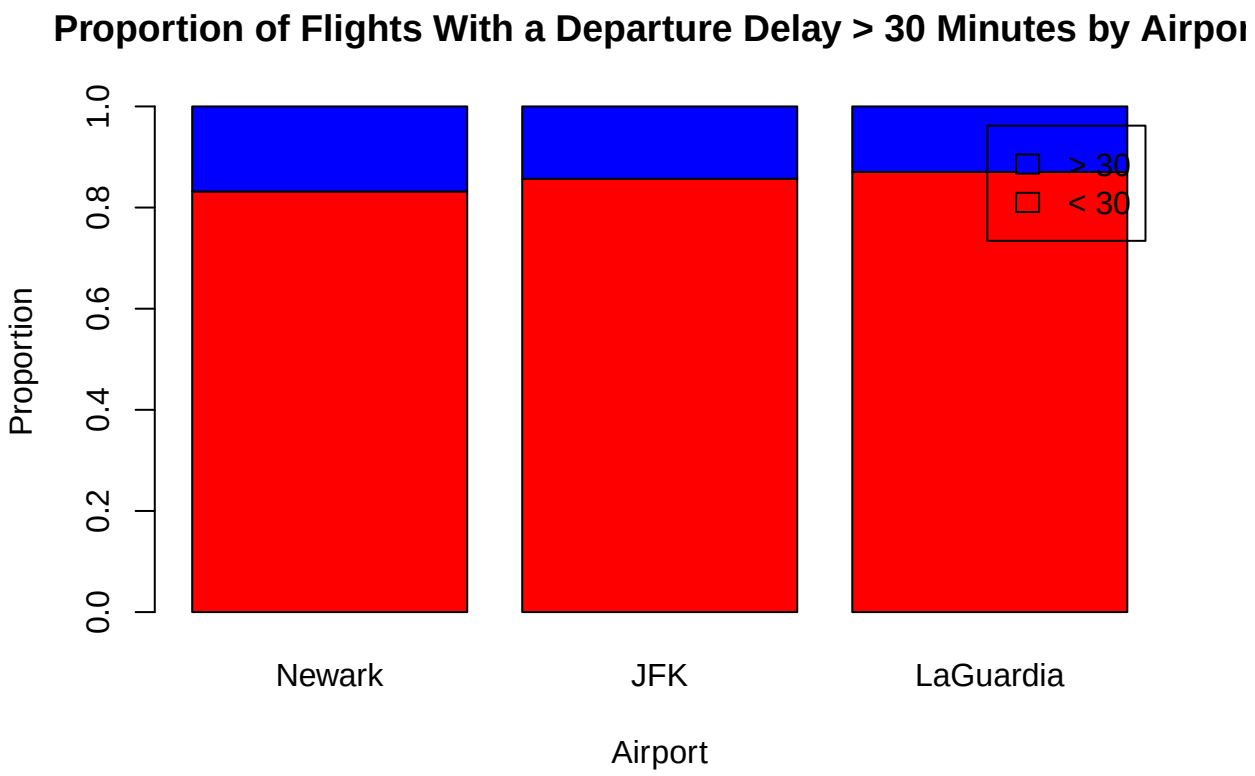
```
nyc.flights.data$origin = as.factor(nyc.flights.data$origin)
levels(nyc.flights.data$origin) = c("Newark", "JFK", "LaGuardia")
```

```
boxplot(nyc.flights.data$dep_delay~nyc.flights.data$origin,
        xlab = "Airport",
        ylab = "Delay (in minutes)",
        main = "Departure Delay of Flights from NYC by Airport",
        col = c("red", "green", "blue"),
        outcol = "red",
        ylim = c(-25, 50))
```



This is a comparative boxplot comparing departure delay of flights with NYC airports. LaGuardia has the lowest median as well as the smallest and lowest IQR. Newark and JFK have very similar medians although Newark has a higher top IQR bound, suggesting more of its delay times above the median are higher than those of JFK. In short, the order (from worst to best) of NYC airports by departure delay is: Newark, JFK, LaGuardia. Again, however, we should consider which airport has "big" ($>$ 30 minutes) delays most often.

```
barplot(prop.table(table(nyc.flights.data$dep_delay_above_half_hour, nyc.flights.data$origin), 2),
        xlab = "Airport",
        ylab = "Proportion",
        main = "Proportion of Flights With a Departure Delay > 30 Minutes by Airport",
        col = c("red", "blue"),
        legend.text = TRUE,
        )
```



In this proportional stacked barplot, we can see that LaGuardia has the smallest and Newark the biggest proportion of flights with a departure delay greater than 30 minutes, corroborating our previous ranking of the NYC airports.

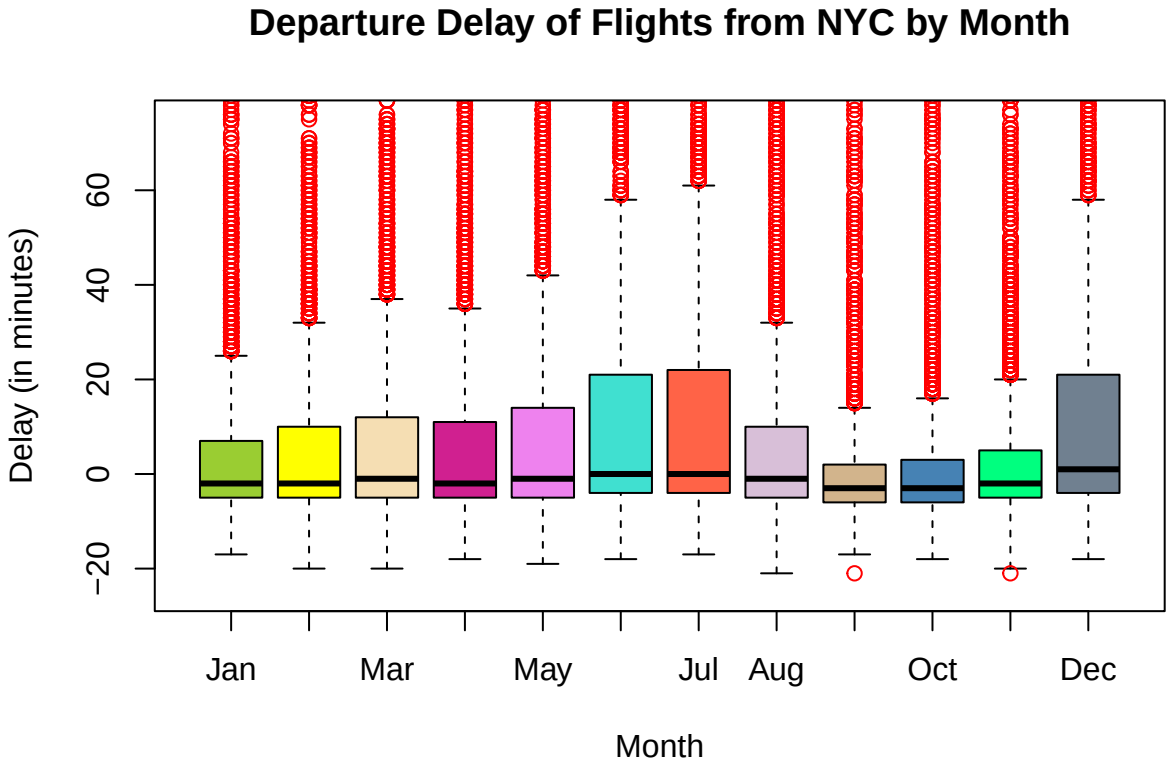
When to Travel

Now let's take a look at when to travel: which months, which weeks, and what times of day.

First, let's change the variable `month` to be a factor so that we can rename them from numbers to abbreviations; again, for future viewing simplicity.

```
nyc.flights.data$month = as.factor(nyc.flights.data$month)
levels(nyc.flights.data$month) = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct"
```

```
boxplot(nyc.flights.data$dep_delay~nyc.flights.data$month,
        xlab = "Month",
        ylab = "Delay (in minutes)",
        ylim = c(-25, 75),
        main = "Departure Delay of Flights from NYC by Month",
        col = c("yellowgreen", "yellow", "wheat", "violetred", "violet", "turquoise", "tomato", "thistl
        outcol = "red"
        )
```

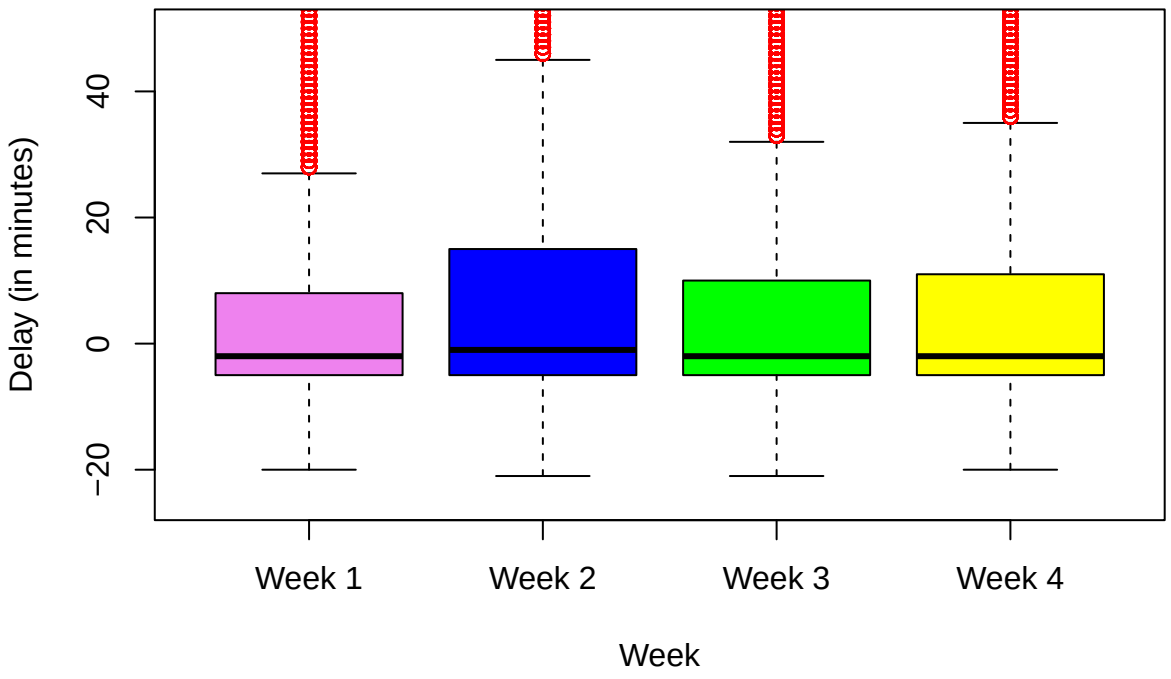


This is a comparative boxplot comparing departure delay of flights in minutes with the months of the year. The medians look very close to each other, but there are noticeable differences in the IQRs, with June, July, and December all having larger IQRs than the rest of the months. In particular, the bottom bound of the IQR remains the same while the top increases by a significant amount, suggesting that more flights above the median had higher delay times. This phenomenon makes sense; June, July, and December are all travel-heavy months, with summer and winter vacations taking place during these times. In contrast, September appears to be the least busiest time of the year, with not only the lowest median departure delay, but also the smallest IQR, meaning departure delays between the 25th and 75th percentile are very close to each other.

Now let's take a look at what the best times of the month to travel are. Using weeks to determine times of the month, we'll assume that days beyond the 28th belong to the 4th week, when in reality that may not be the case.

```
breaks.for.day = c(0, 7, 14, 21, 32)
nyc.flights.data$week = cut(nyc.flights.data$day, breaks = breaks.for.day, right = TRUE)
levels(nyc.flights.data$week) = c("Week 1", "Week 2", "Week 3", "Week 4")
```

```
boxplot(nyc.flights.data$dep_delay~nyc.flights.data$week,
        ylim = c(-25, 50),
        main = "Departure Delay of Flights by Week",
        ylab = "Delay (in minutes)",
        xlab = "Week",
        col = c("violet", "blue", "green", "yellow"),
        outcol = "red")
```

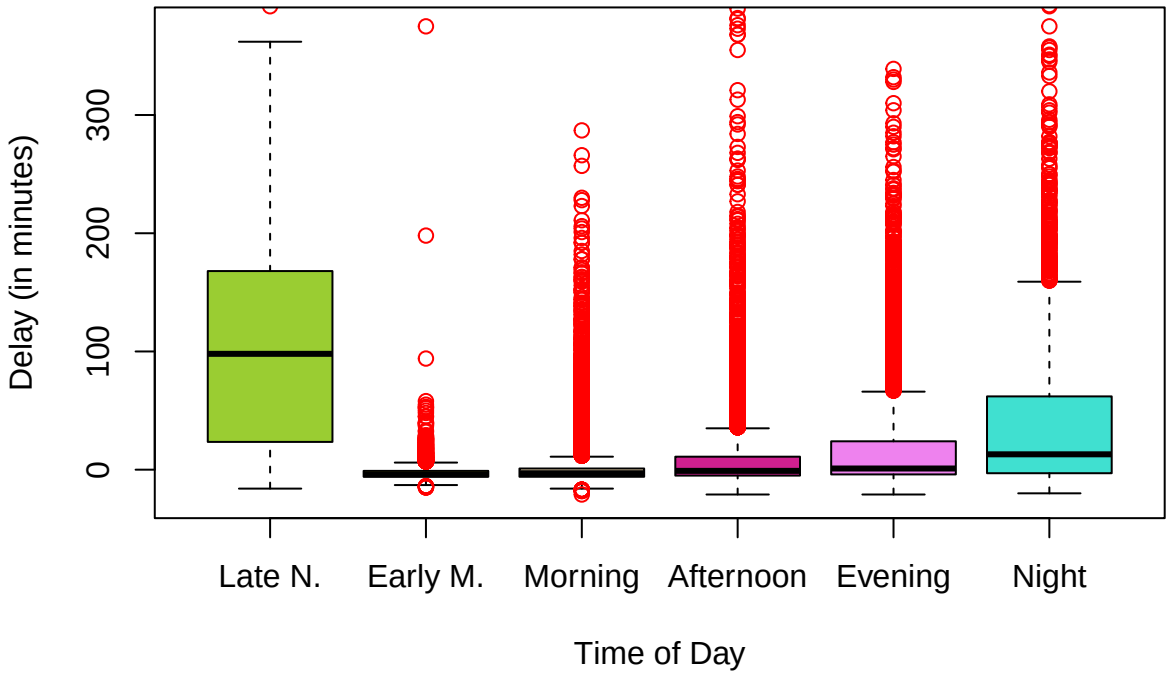


This comparative boxplot shows that the median departure delays among all 4 weeks are relatively the same. However, it's interesting to note that Week 2 has a bigger IQR and a greater upper bound, suggesting more flights with departure delays above the median had greater delays than the other weeks. I would have expected Weeks 1 & 4 to be the busiest as that's when people who travel for work most likely travel, so this graph surprised me.

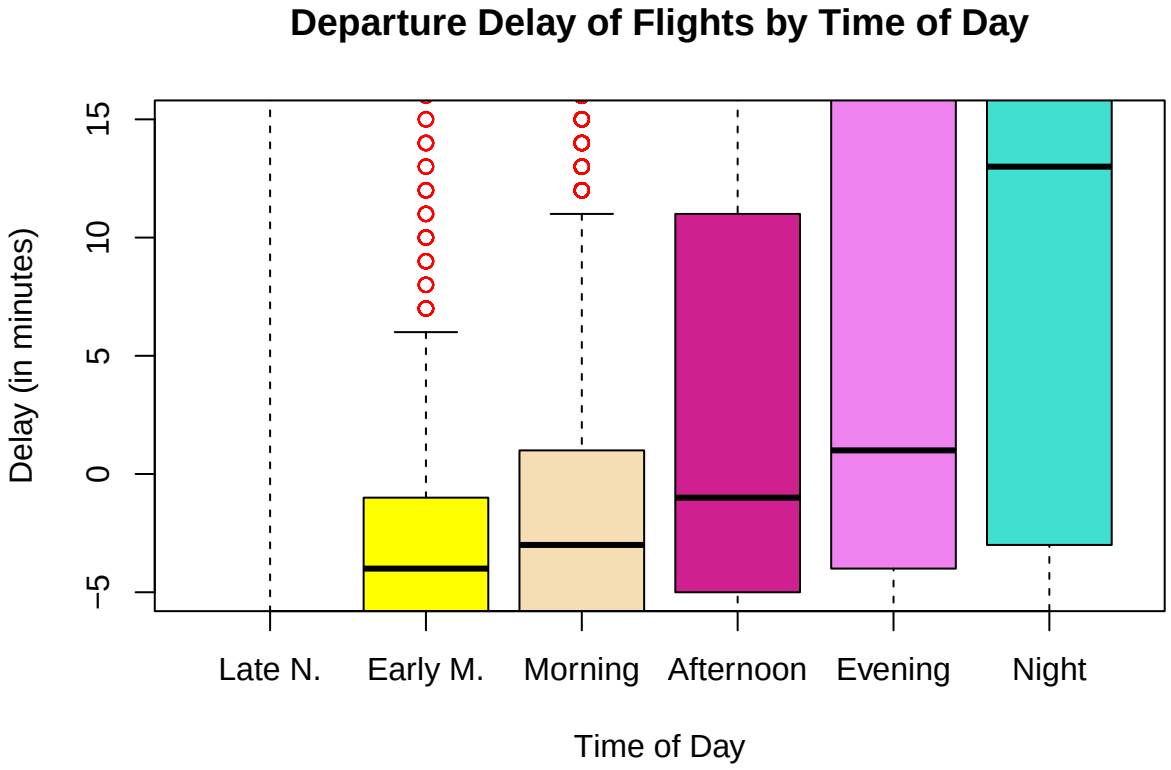
Now let's take a look at what times of day were busiest. Let's group times into six categories: 11 PM - 3 AM = late night, 3 AM - 7 AM = early morning, 7 AM - 12 PM = morning, 12 PM - 5 PM = afternoon, 5 PM - 8 PM = evening, 8 PM - 11 PM = night.

```
breaks.for.time = c(0, 300, 700, 1200, 1700, 2000, 2300, 2500)

nyc.flights.data$dep_time_of_day = cut(nyc.flights.data$dep_time, breaks = breaks.for.time, right = FALS

levels(nyc.flights.data$dep_time_of_day) = c("Late N.", "Early M.", "Morning", "Afternoon", "Evening",
```

```
boxplot(nyc.flights.data$dep_delay~nyc.flights.data$dep_time_of_day,
        ylim = c(-25, 375),
        main = "Departure Delay of Flights by Time of Day",
        ylab = "Delay (in minutes)",
        xlab = "Time of Day",
        col = c("yellowgreen", "yellow", "wheat", "violetred", "violet", "turquoise"),
        outcol = "red")
```




```
boxplot(nyc.flights.data$dep_delay~nyc.flights.data$dep_time_of_day,
        ylim = c(-5, 15),
        main = "Departure Delay of Flights by Time of Day",
        ylab = "Delay (in minutes)",
        xlab = "Time of Day",
        col = c("yellowgreen", "yellow", "wheat", "violetred", "violet", "turquoise"),
        outcol = "red")
```



This comparative boxplot compares departure delay with departure time of day. I was honestly quite surprised by this graph; I originally predicted that flights in the afternoon would have the largest departure delays as that's when I thought most flights would take place. To see that it was actually late night flights and the magnitude of their delay was very shocking. I have a hypothesis as to why this is. Since departure time records the actual time and not the estimated time of departure, perhaps many of these late night flights were flights meant to have departed earlier in the day, such as in the afternoon, evening, or night, and got delayed until the late night, thus being recorded as a flight during such. And to exacerbate this effect, there are less regular, non-delayed, flights during the late night, meaning these delayed flights from earlier in the day have a bigger effect.

Aside from this late night observation, however, early morning flights have the lowest median departure delay. So if you have to book an important flight, plan it for the early morning; it might be a pain to wake up that early, but at least you won't have to be stuck waiting.

Conclusion

In this lab, I was able to analyze the best times to fly, the best airlines to fly on, and which one of the three NYC airports to fly from. I was not aware how good Hawaiian Airlines was! I was also surprised that flights during the late night have pretty big departure delays; I thought they would have some of the smallest. I was also surprised that LaGuardia had the smallest departure delays out of the three NYC airports; granted, this data is from 2013, but I've heard some less than good things about LaGuardia.

In a future lab, I'd like to explore more recent data, since some of the airlines in this data set are now defunct. I'd also like to explore data where the day of the week is given; I wonder which days are busiest? Mondays and Fridays? It'd also be interesting to know where the flights came from and if the distance flown (from previous destination to NYC airport) had an effect on departure delays.