hw1_estes

Andrew Estes

10/23/2021

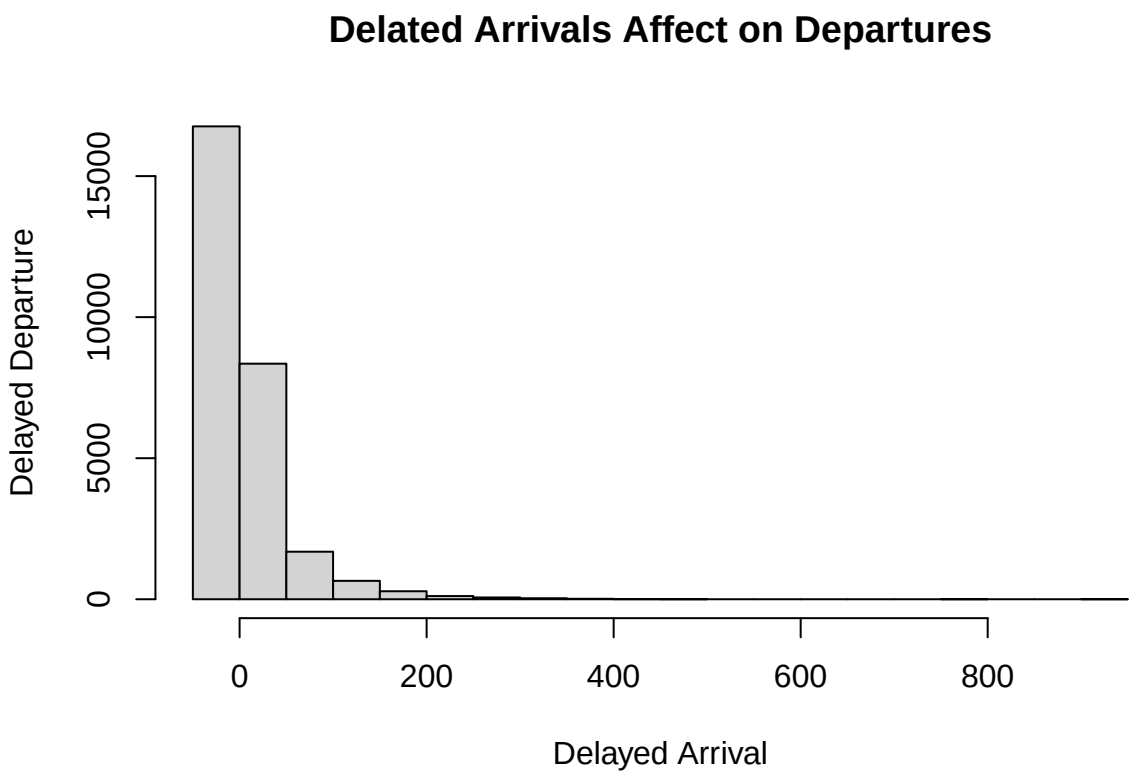
Downloading and viewing the appropriate library and functions

Below is the code to create the nycMarch variable.

```
nycMarch <- flights %>%
  filter(month==3) %>%
  subset(select = -c(origin, dest)) %>%
  arrange(desc(tailnum))
```

Here is the histogram for delayed flights in March.

```
hist(nycMarch$dep_delay,
     xlab="Delayed Arrival", ylab="Delayed Departure",
     main="Delated Arrivals Affect on Departures")
```



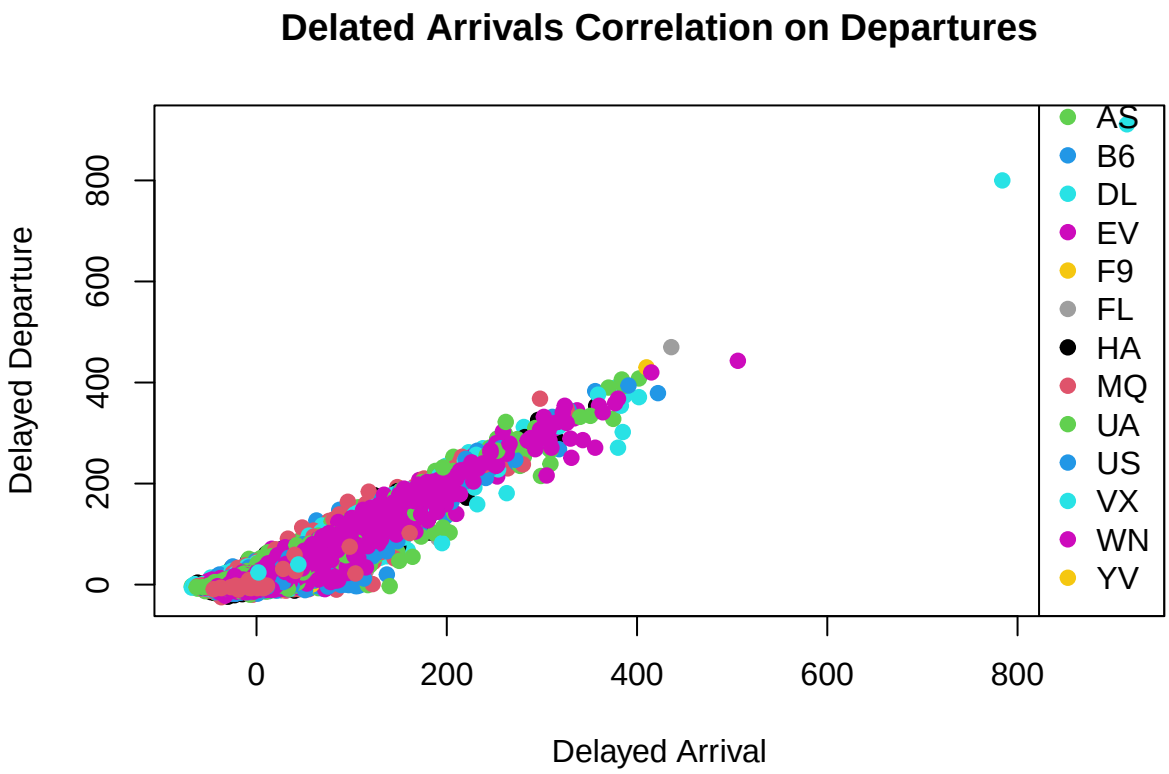
Below is the linear model we created to test the correlation between a delayed arrival and delayed departure. Unsurprisingly, there is a strong relationship with an R-Squared number of 0.85.

```
model <- lm(arr_delay ~ dep_delay, nycMarch)
summary(model)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay, data = nycMarch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.790 -10.363  -1.329   8.723 150.637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.58476    0.10667  -71.11   <2e-16 ***
## dep_delay    1.01732    0.00253  402.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.93 on 27900 degrees of freedom
##   (932 observations deleted due to missingness)
## Multiple R-squared:  0.8528, Adjusted R-squared:  0.8528
## F-statistic: 1.617e+05 on 1 and 27900 DF,  p-value: < 2.2e-16
```

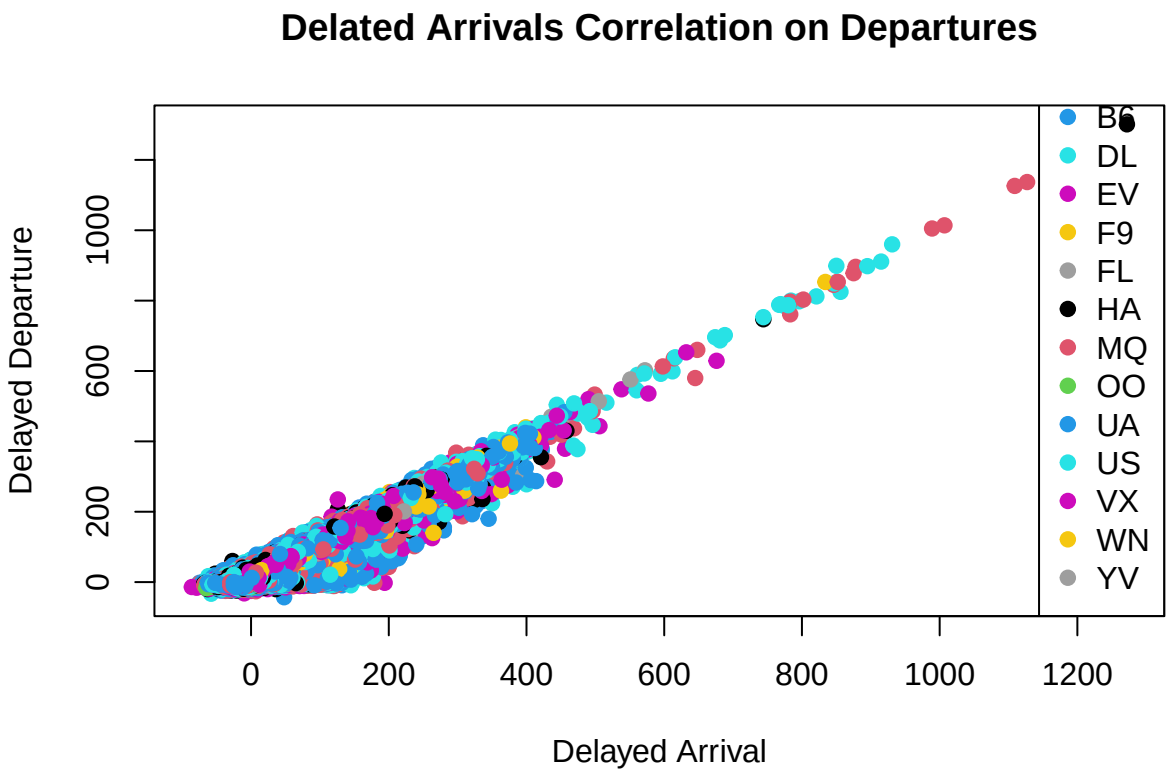
Here is the scatterplot of delayed departures in March, color coded by airline.

```
plot(nycMarch$arr_delay, nycMarch$dep_delay,
     xlab="Delayed Arrival", ylab="Delayed Departure",
     main="Delated Arrivals Correlation on Departures",
     pch = 19,
     col = factor(nycMarch$carrier))

legend("bottomright",
       legend = levels(factor(nycMarch$carrier)),
       pch = 19,
       col = factor(levels(factor(nycMarch$carrier))))
```



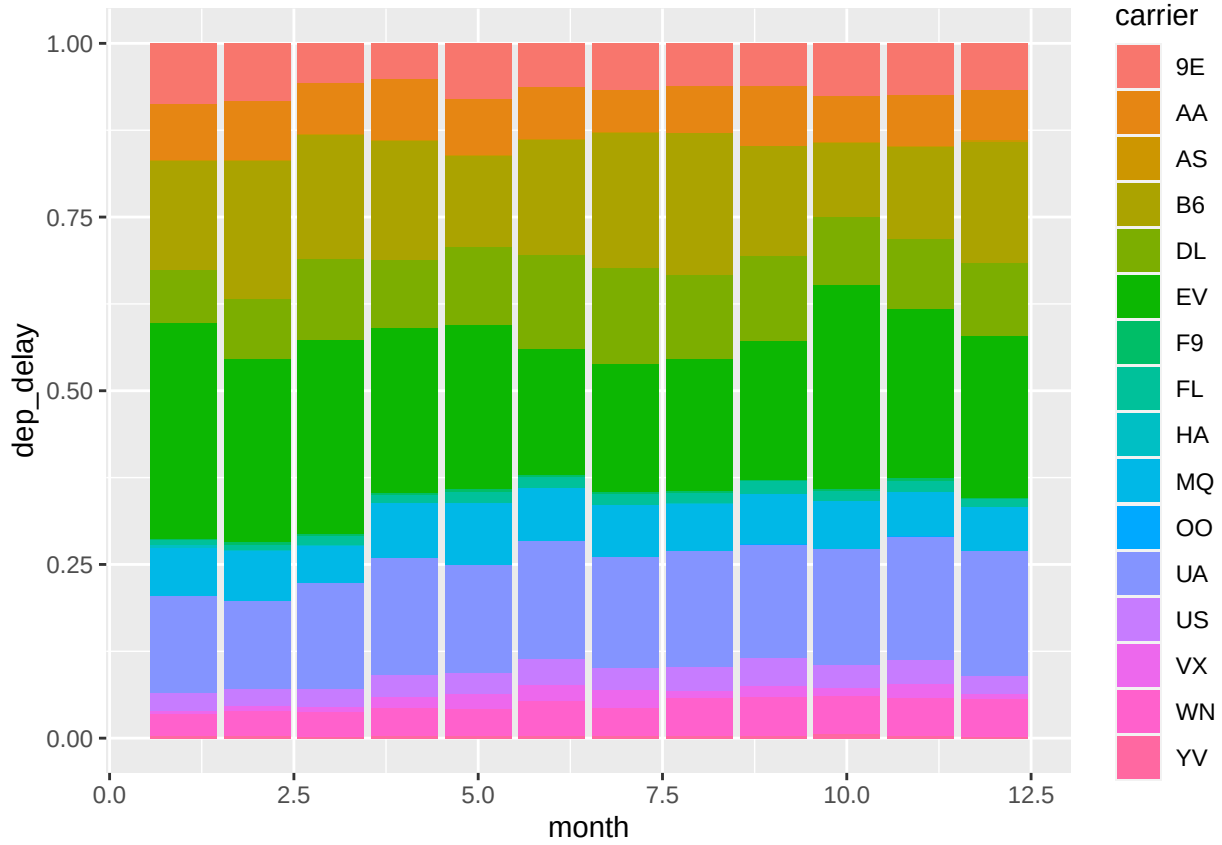
Below is a scatter plot of the entire dataset. It took an additional 18.09 seconds to run compared to just March.

```
plot(flights$arr_delay, flights$dep_delay,
     xlab="Delayed Arrival", ylab="Delayed Departure",
     main="Delated Arrivals Correlation on Departures",
     pch = 19,
     col = factor(flights$carrier))
legend("bottomright",
       legend = levels(factor(flights$carrier)),
       pch = 19,
       col = factor(levels(factor(flights$carrier))))
```



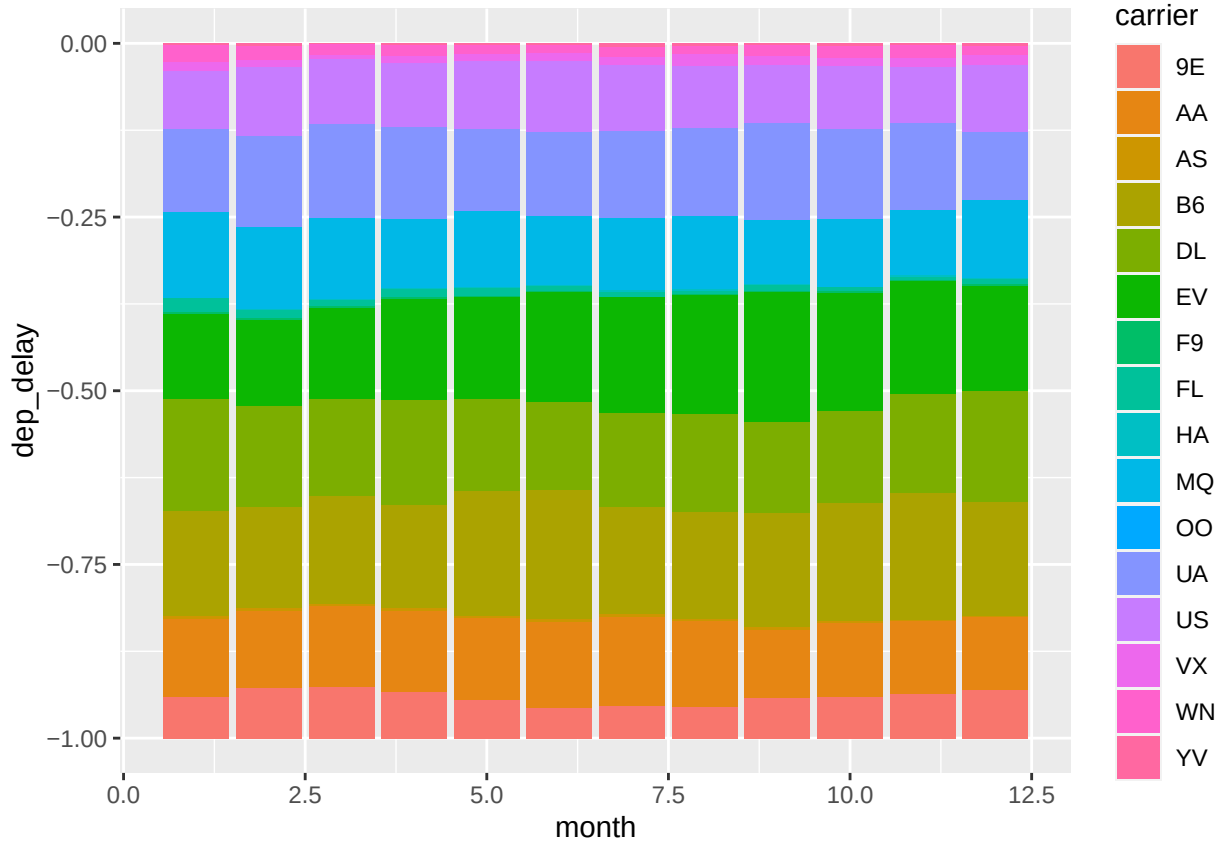
Below is a stacked barchart showing how often each airline has a delayed flight each month.This only takes into consideration flights that were delayed greater than 0 min.

```
delayedFlights <- filter(flights, dep_delay > 0)

ggplot(delayedFlights, aes(fill=carrier, y=dep_delay, x=month)) +
  geom_bar(position='fill', stat='identity')
```



On the flip side, we can see which airlines leave earlier than scheduled.

```
earlyFlights <- filter(flights, dep_delay < 0)

ggplot(earlyFlights, aes(fill=carrier, y=dep_delay, x=month)) +
  geom_bar(position='fill', stat='identity')
```

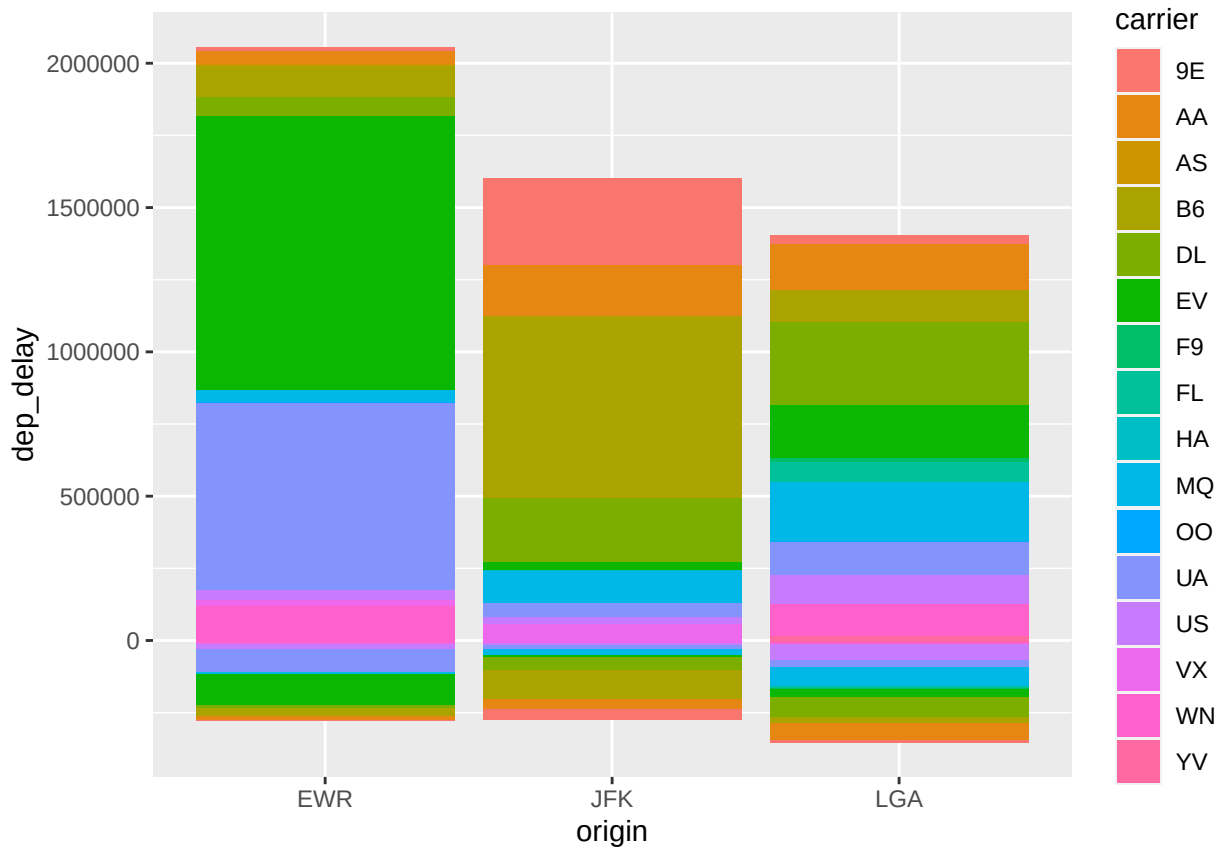


There are no decisions that can be made based off the two flights above. A proper analysis would include standardizing within the airlines prior to looking at the entire picture. For example, Delta looks to be around the 3rd most often delayed airline. And that may be true of the entire population. But they might have the most number of flights so proportionally speaking, it is entirely possible they have one of the lowest rates of delayed flights.

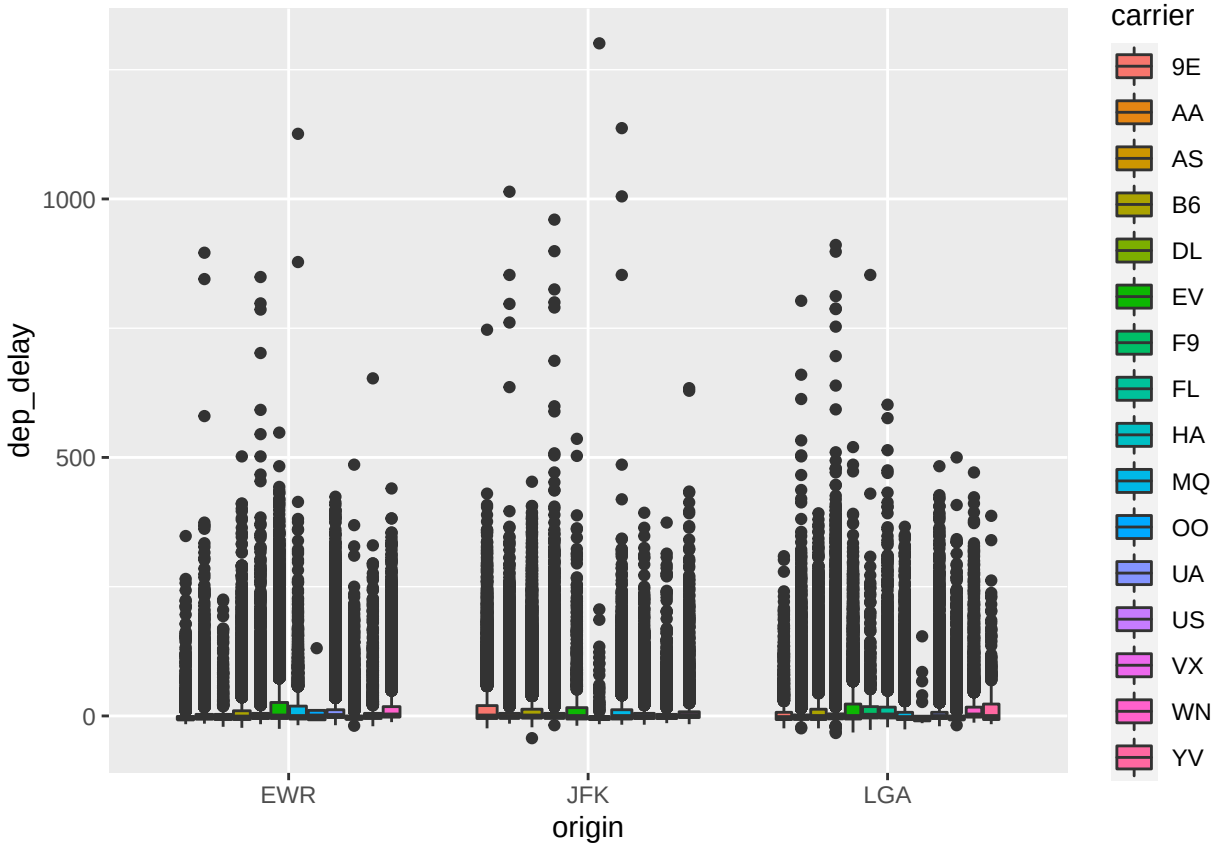
Another consideration to analyze departure delays by time per each airport. There are three airports within the NYC Flights data set and the longest departure times obviously come from Newark Liberty International (EWR). I also broke it down by airline and individual flights to see if the data provided a different outlook compared to the initial analysis. While EWR was easily the worst offender with over 2 million minutes in delays over the year, JFK had more 4x as many flights have 1000+ minute delays as EJR. This would indicate that JFK is typically more on schedule than EWR. If, however, there is a delay at JFK, it is more likely to be a longer delay than at EWR.

This will should be broken down into a percentage of flights prior to making any statements as to efficacy differences. Although based off this quick research, La Gaurdia (LGA) seems to be the most efficient, timely of the three airports in this dataset.

```
airportTime <- ggplot(data=flights) +
  geom_col(mapping=aes(x=origin, y=dep_delay, fill=carrier, na.rm=TRUE))
airportTime
```




```
airportAirline <- ggplot(flights, aes(origin, dep_delay, fill=carrier, na.rm=TRUE)) +
  geom_boxplot()
airportAirline
```


dep_delay
1000
500
0
EWR
JFK
LGA
origin
carrier
9E
AA
AS
B6
DL
EV
F9
FL
HA
MQ
OO
UA
US
VX
WN
YV