Exercise 2

Lio, Lio, Zhu, Thelakkat

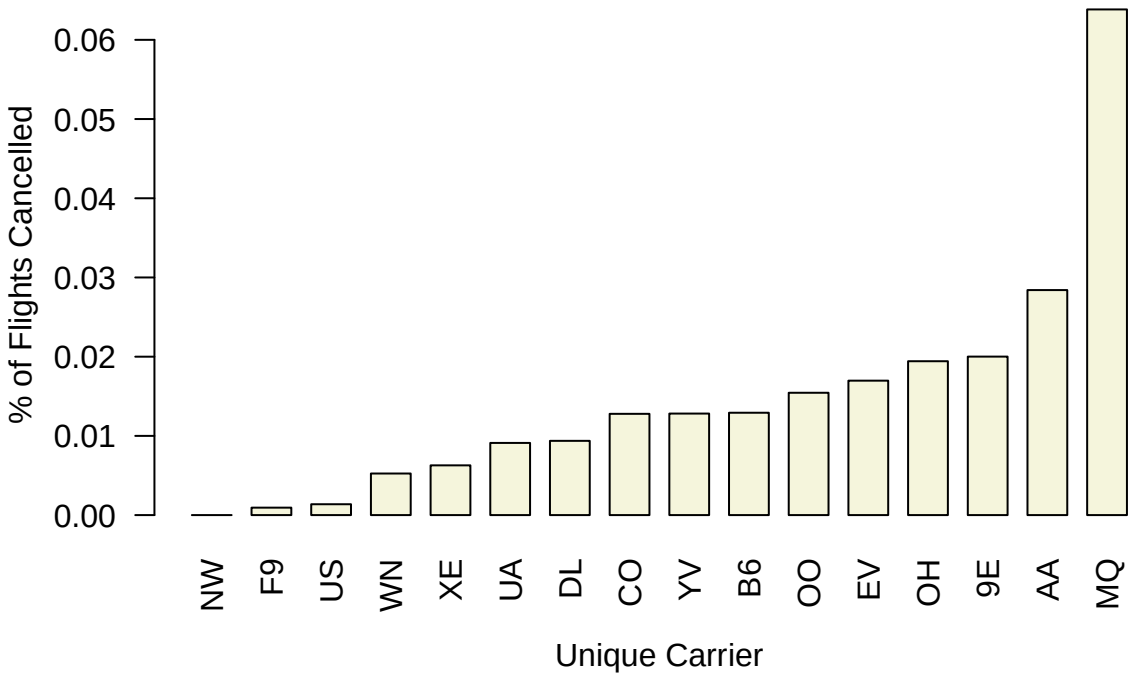
August 13, 2017

Flights at ABIA

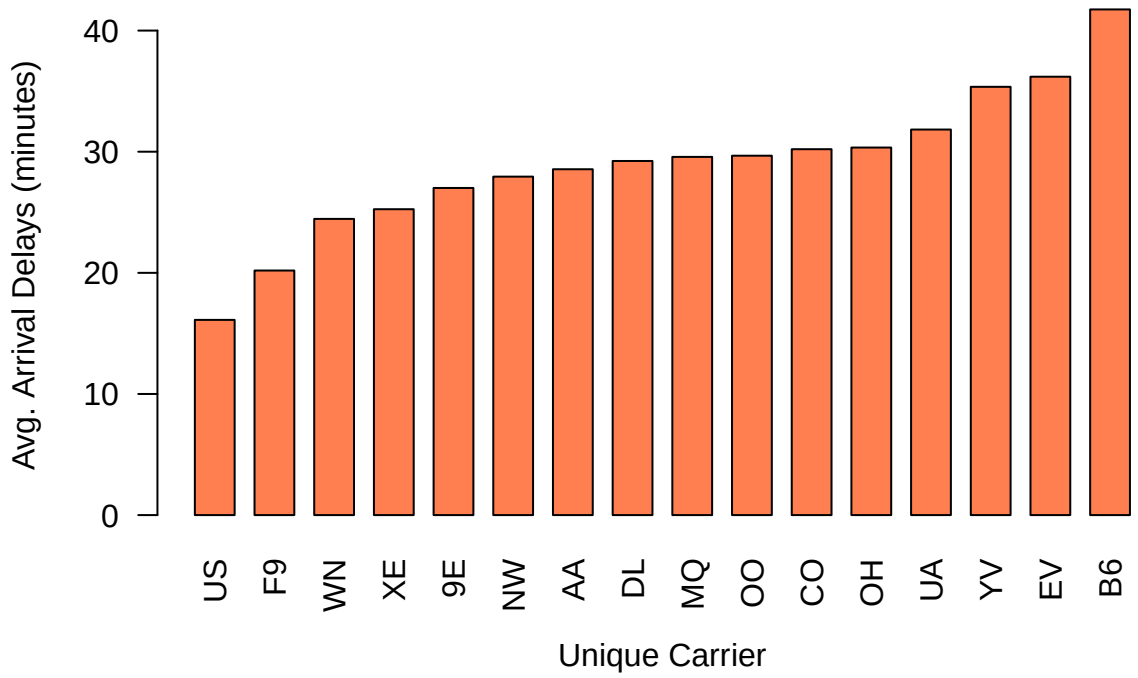
"Your task is to create a figure, or set of related figures, that tell an interesting story about flights into and out of Austin. You can annotate the figure and briefly describe it, but strive to make it as stand-alone as possible. It shouldn't need many, many paragraphs to convey its meaning. Rather, the figure should speak for itself as far as possible."

For our first section of exploratory data analysis, we decided to focus on airlines to see if we could draw any insights about which Airlines were more reliable in terms of delays and cancellations. We then looked into average arrival and average departure delay times (in minutes) each airline had when flying into or out of Austin.

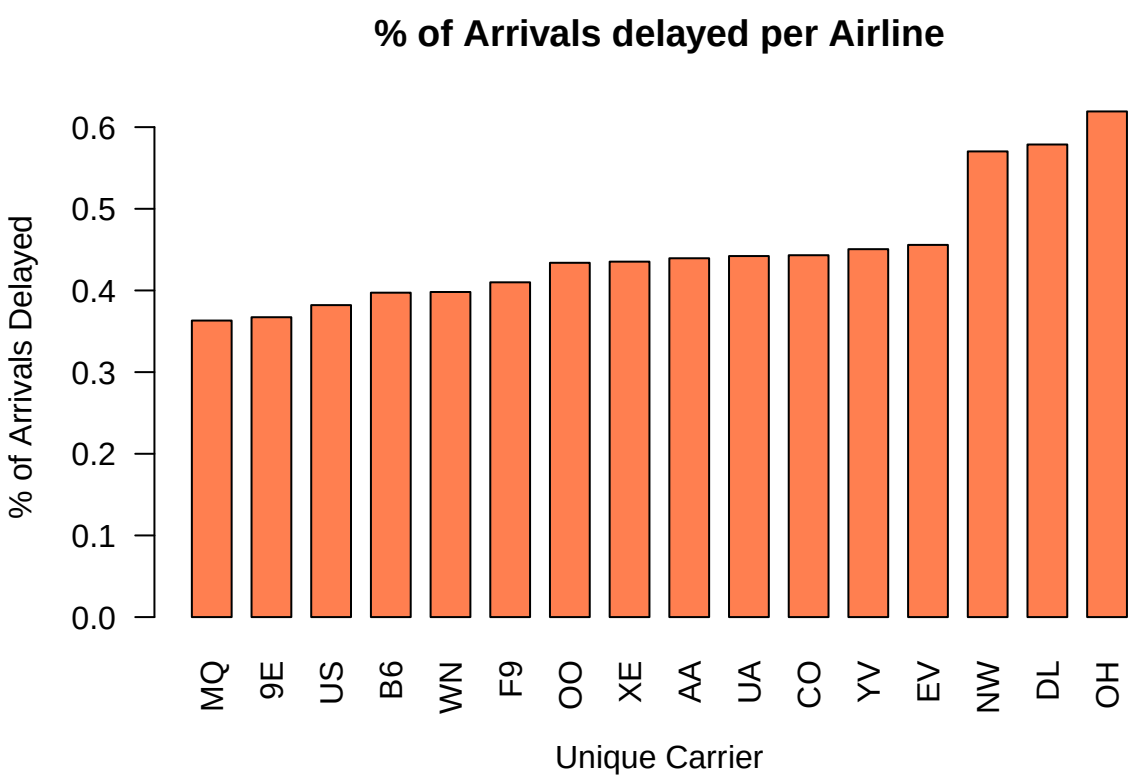
```
airlinedata = read.csv("ABIA.csv")
#calculating percentage of flights cancelled for each airline

dfcancel = data.frame(aggregate(airlinedata$Cancelled~ airlinedata$UniqueCarrier,
                                airlinedata ,sum))

df = data.frame(aggregate(airlinedata$FlightNum~ airlinedata$UniqueCarrier,
                          airlinedata, length))

finaldf = merge(dfcancel, df)
finaldf = within(finaldf, percent <- airlinedata.Cancelled/airlinedata.FlightNum)
finaldf = finaldf[order(finaldf$percent),]

barplot(finaldf$percent, names = finaldf$arrivaldelays.UniqueCarrier,
        xlab = "Unique Carrier", ylab = "% of Flights Cancelled",
        main = "% of Flights Cancelled per Airline", las=2, space=.5, col='beige',
        names.arg=c("NW", "F9", "US",'WN', "XE", "UA", "DL","CO", "YV",
                    "B6","OO", "EV", "OH", "9E", "AA", "MQ"))
```



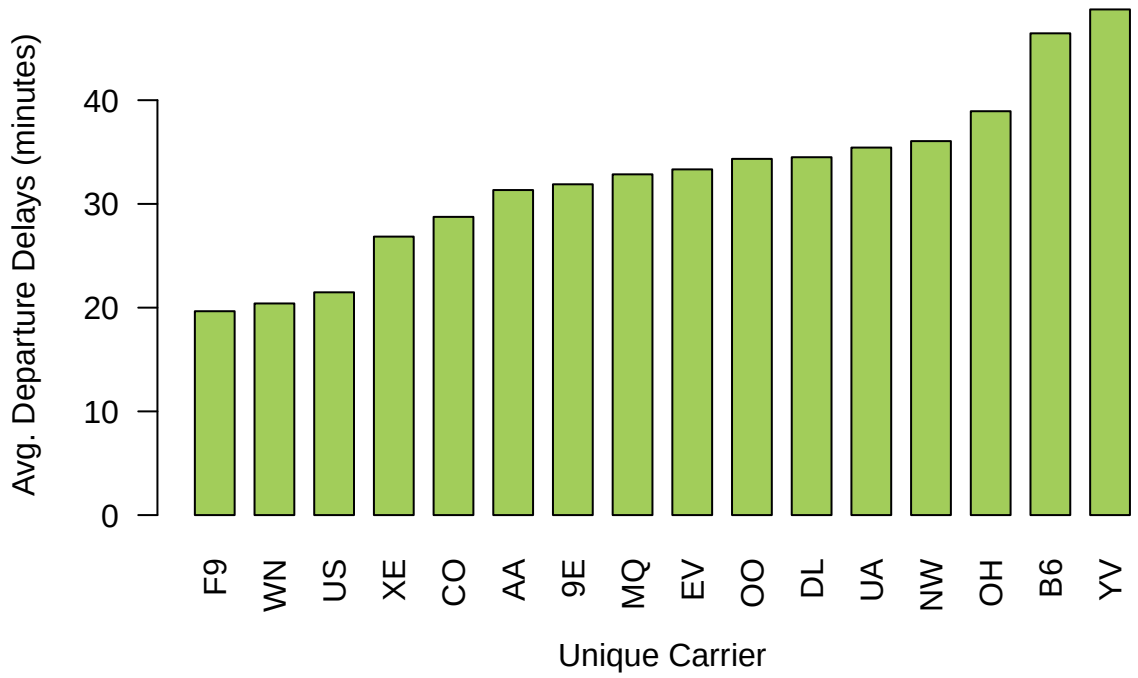

```
arrivaldelays = airlinedata[which(airlinedata[,15]>0),]
df3 = data.frame(aggregate(arrivaldelays$ArrDelay~ arrivaldelays$UniqueCarrier,
                           arrivaldelays, mean))
df4 = df3[order(df3$arrivaldelays.ArrDelay),]
fix(df4)

barplot(df4$arrivaldelays.ArrDelay, names = df4$arrivaldelays.UniqueCarrier,
        xlab = "Unique Carrier", ylab = "Avg. Arrival Delays (minutes)",
        main = "Avg. Arrival Delay times per Airline", las=2, space=.5, col='coral')
```



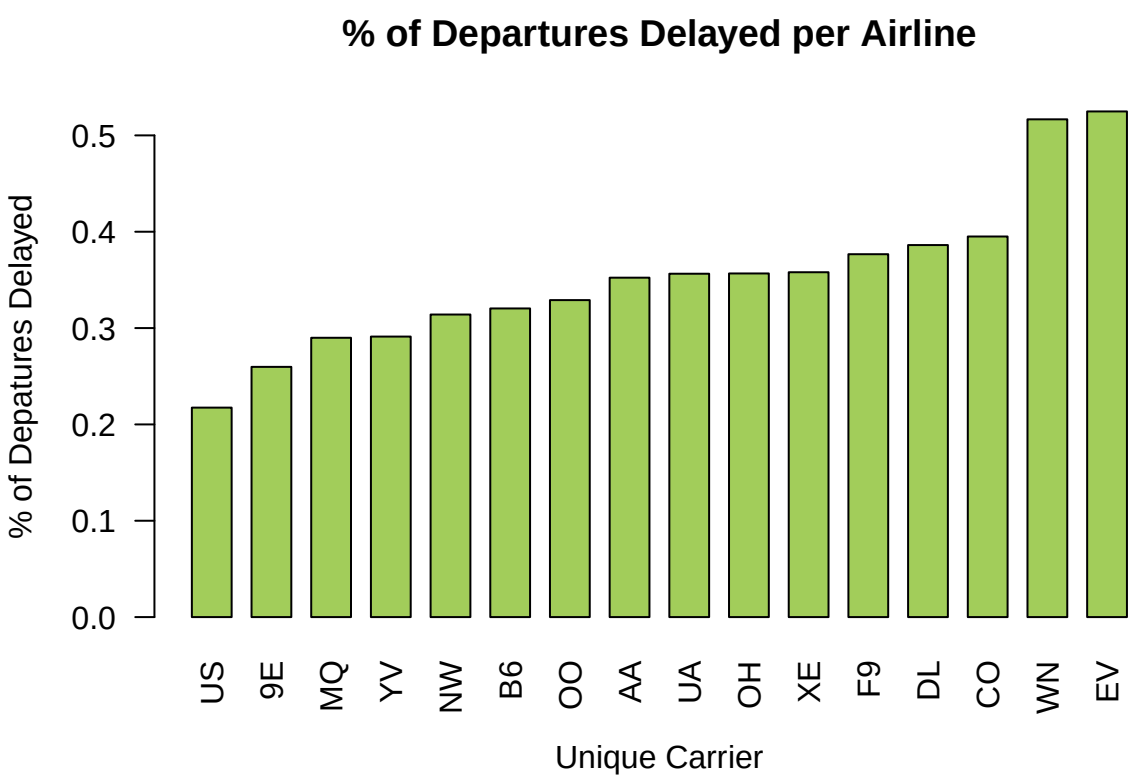

```
#calculating percentage of flights delayed for each airline

df33 = data.frame(aggregate(arrivaldelays$ArrDelay~ arrivaldelays$UniqueCarrier,
                            arrivaldelays, length))
df = data.frame(aggregate(airlinedata$FlightNum~ airlinedata$UniqueCarrier,
                          airlinedata, length))

finaldf = merge(df33, df, by.x="arrivaldelays.UniqueCarrier", by.y="airlinedata.UniqueCarrier")
finaldf = within(finaldf, percent <- arrivaldelays.ArrDelay/airlinedata.FlightNum)
finaldf = finaldf[order(finaldf$percent),]
barplot(finaldf$percent, names = finaldf$arrivaldelays.UniqueCarrier,
        xlab = "Unique Carrier", ylab = "% of Arrivals Delayed",
        main = "% of Arrivals delayed per Airline", las=2, space=.5, col='coral')
```




```
departdelays = airlinedata[which(airlinedata[,16]>0),]
df5 = data.frame(aggregate(departdelays$DepDelay~ departdelays$UniqueCarrier,
                           departdelays, mean))

df6 = df5[order(df5$departdelays.DepDelay),]

barplot(df6$departdelays.DepDelay, names = df6$departdelays.UniqueCarrier,
        xlab = "Unique Carrier", ylab = "Avg. Departure Delays (minutes)",
        main = "Avg. Departure Delay times per Airline", las=2, space=.5,
        col='darkolivegreen3')
```

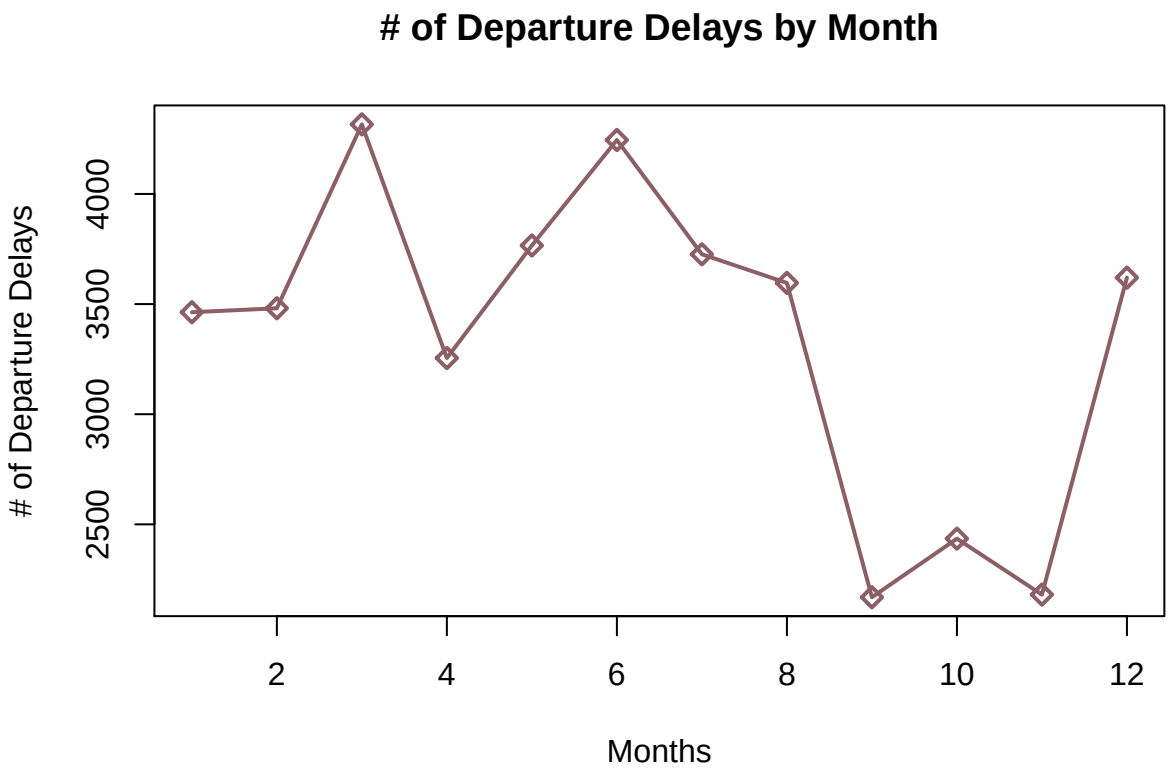



```
#calculating percentage of flights delayed for each airline

df55 = data.frame(aggregate(departdelays$DepDelay~ departdelays$UniqueCarrier,
                            departdelays, length))

df = data.frame(aggregate(airlinedata$FlightNum~ airlinedata$UniqueCarrier, airlinedata, length))

finaldf = merge(df55, df, by.x="departdelays.UniqueCarrier", by.y="airlinedata.UniqueCarrier")
finaldf = within(finaldf, percent <- departdelays.DepDelay/airlinedata.FlightNum)
finaldf = finaldf[order(finaldf$percent),]
barplot(finaldf$percent, names = finaldf$departdelays.UniqueCarrier,
        xlab = "Unique Carrier", ylab = "% of Depatures Delayed",
        main = "% of Departures Delayed per Airline", las=2, space=.5, col='darkolivegreen3')
```

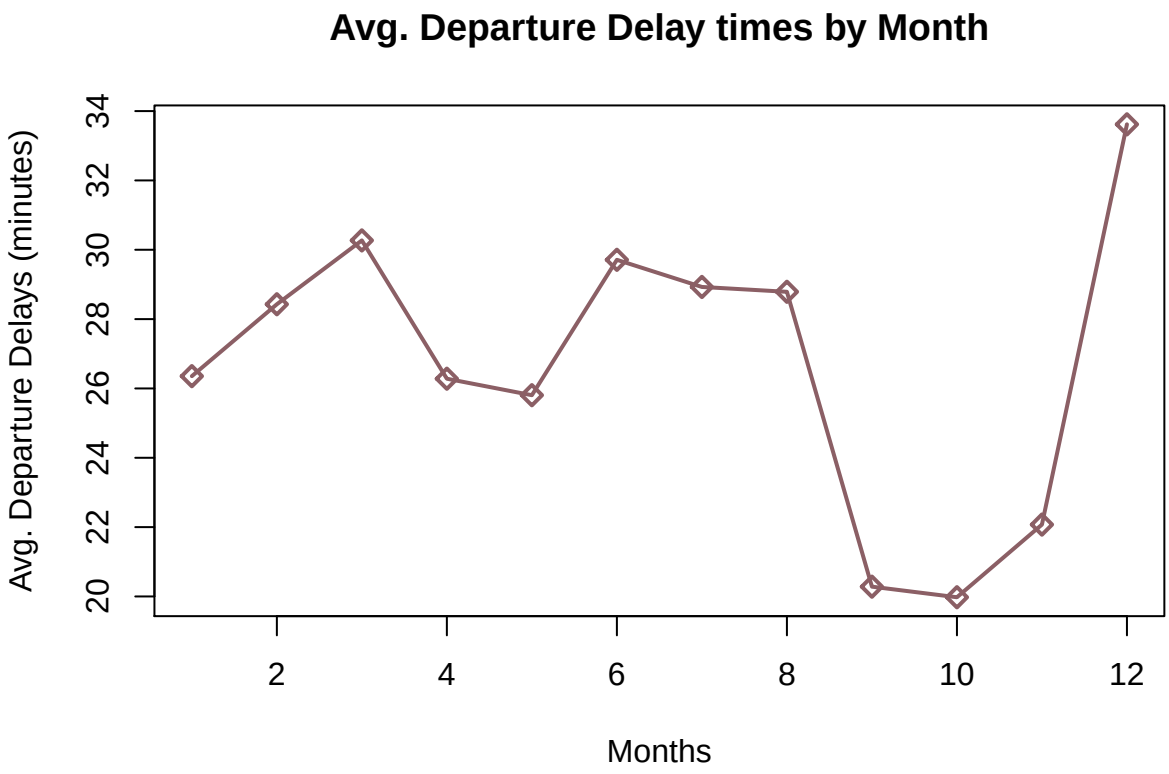


For our next part of the analysis we focused more on which dates (time, days, months) of the year were the most reliable to fly on. We used a subset of the data, using only the rows where Departure Delay was greater than 0 (i.e. showing a departure delay took place). We did this becuase the departure delay variable focused on people in Austin, who would be flying out of Austin.

```
delays = airlinedata[which(airlinedata[,16]>0),]

dfm = data.frame(aggregate(delays$DepDelay~ delays$Month, delays, length))

plot(dfm$delays.Month, dfm$delays.DepDelay,
        xlab = "Months", ylab = "# of Departure Delays",
        main = "# of Departure Delays by Month", type='o',
        col='lightpink4', lwd=2, pch=5)
```

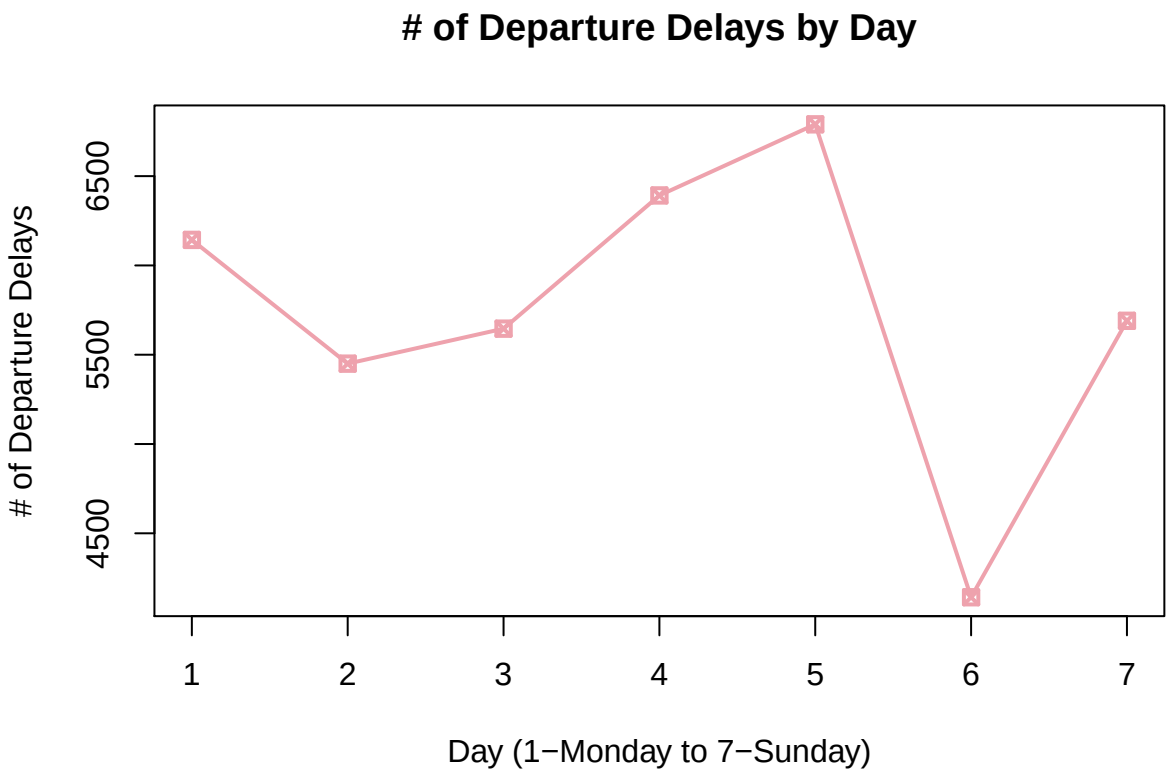



```
dfmm = data.frame(aggregate(delays$DepDelay~ delays$Month, delays, mean))


plot(dfmm$delays.Month, dfmm$delays.DepDelay,
        xlab = "Months", ylab = "Avg. Departure Delays (minutes)",
        main = "Avg. Departure Delay times by Month", type='o',
        col='lightpink4', lwd=2, pch=5)
```

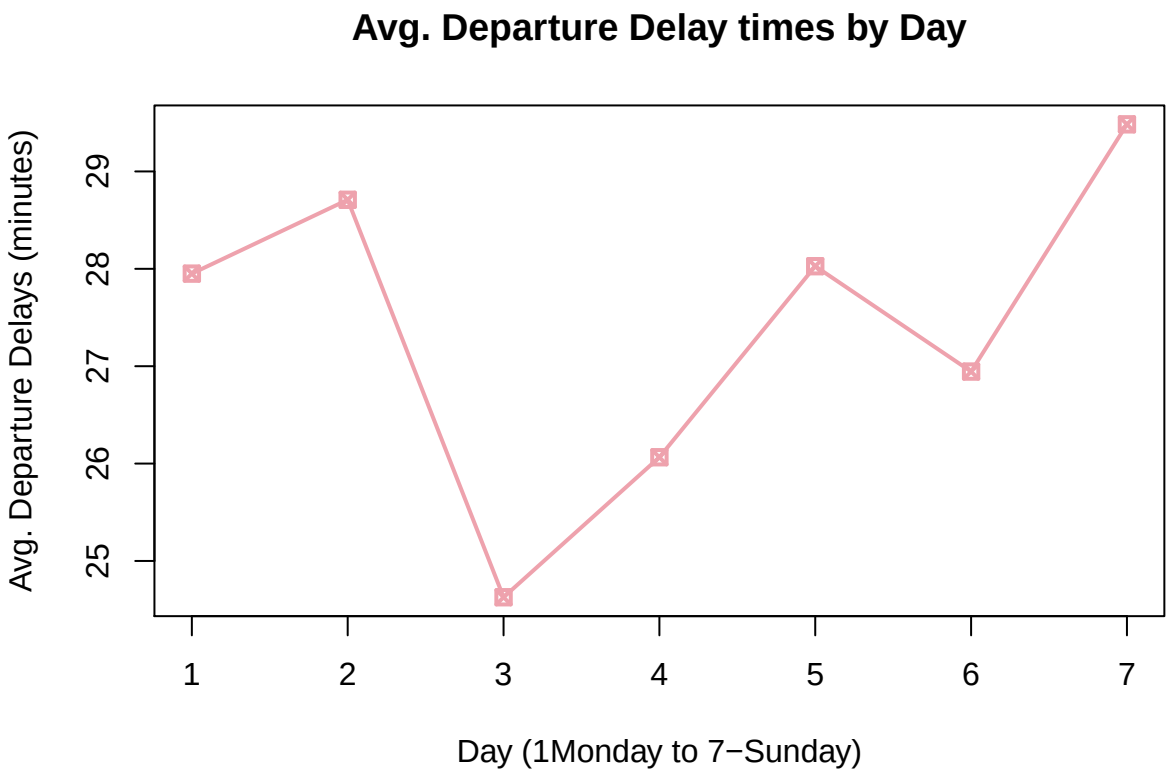


The highest amount of departure delays happened in March and June, but as you can tell from the plots above, the month with the longest departure delays (on average) is in Decemember.

```
delays = airlinedata[which(airlinedata[,16]>0),]

dfday = data.frame(aggregate(delays$DepDelay~ delays$DayOfWeek, delays, length))

plot(dfday$delays.DayOfWeek, dfday$delays.DepDelay,
        xlab = "Day (1-Monday to 7-Sunday)", ylab = "# of Departure Delays",
        main = "# of Departure Delays by Day", type='o',
        col='lightpink2', lwd=2, pch=7)
```

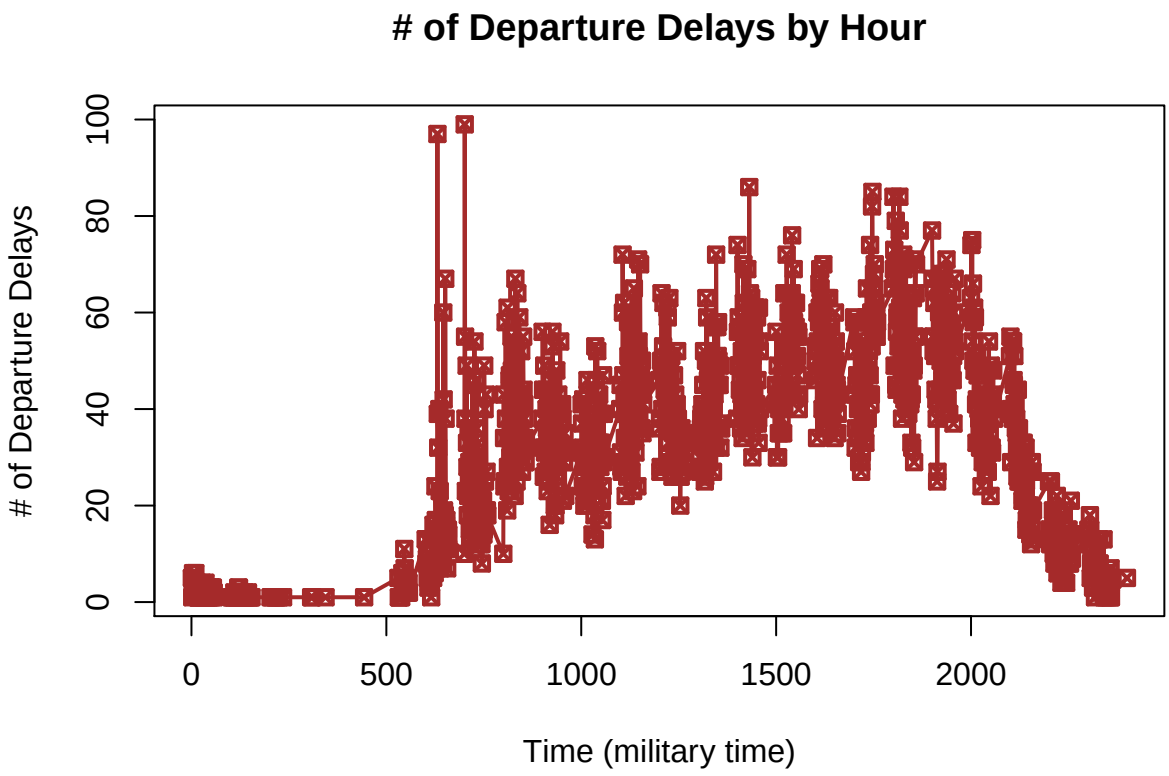



```
dfmdays = data.frame(aggregate(delays$DepDelay~ delays$DayOfWeek, delays, mean))


plot(dfmdays$delays.DayOfWeek, dfmdays$delays.DepDelay,
        xlab = "Day (1Monday to 7-Sunday)", ylab = "Avg. Departure Delays (minutes)",
        main = "Avg. Departure Delay times by Day", type='o',
        col='lightpink2', lwd=2, pch=7)
```

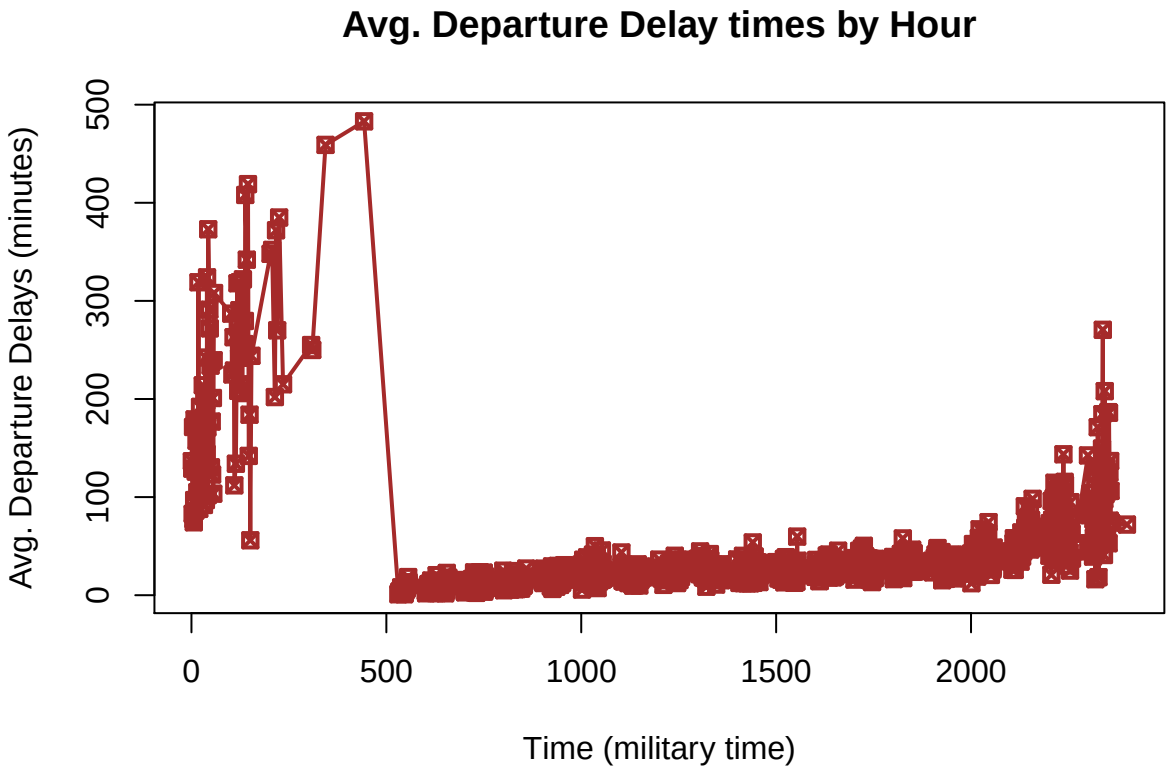


Although Friday had the highest amount of delays, the delays were not necessarily the longest. The longest delays, on average, happened on Sunday.

```
delays = airlinedata[which(airlinedata[,16]>0),]

dfhour = data.frame(aggregate(delays$DepDelay~ delays$DepTime, delays, length))


plot(dfhour$delays.DepTime, dfhour$delays.DepDelay,
        xlab = "Time (military time)", ylab = "# of Departure Delays",
        main = "# of Departure Delays by Hour", type='o',
        col='brown', lwd=2, pch=7)
```




```
dfmhours = data.frame(aggregate(delays$DepDelay~ delays$DepTime, delays, mean))


plot(dfmhours$delays.DepTime, dfmhours$delays.DepDelay,
        xlab = "Time (military time)", ylab = "Avg. Departure Delays (minutes)",
        main = "Avg. Departure Delay times by Hour", type='o',
        col='brown', lwd=2, pch=7)
```



Lastly, it seems like the most delays happen in the middle of the day between 05:00 and 20:00. In contrast, barely any delays occur betwen 0:00 and 5:00, but when they do, they are very long.