STA 380 Homework 2

Henry Chang, Joseph Chin, Tiffany Sung, Jeffrey Fulkerson

August 17, 2018

Problem 1: Flighs at ABIA

```
library(RColorBrewer)
library(gplots)
#setup

# Read in Data
data_raw <- read.csv("ABIA.csv")
# Clean
# --Year is the same for every row
# --Only interested in flights originating from Austin, TX
drops <- c("Year")
data <- data_raw[ data_raw$Origin == "AUS", !(names(data_raw) %in% drops)]
rm(data_raw)
```

For the purpose of this project, We decide to narrow our analysis to flights departing from AUS only.

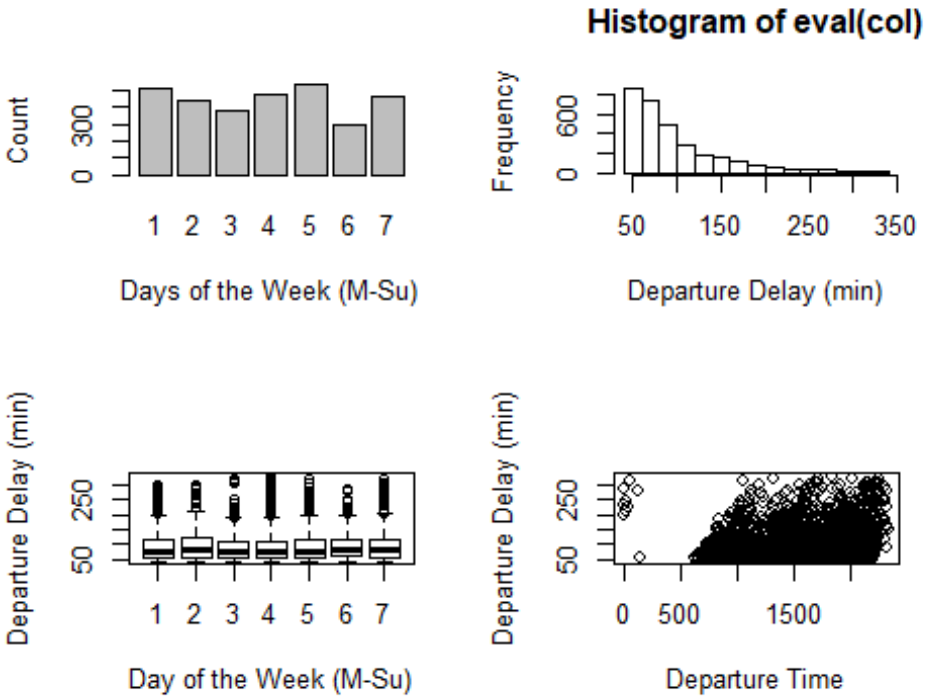
EDA

How do we minimize delay?

We believe that a delay longer than 45 minutes is significant to a business traveler; thus, we decide to analyze flights meeting this standard in Austin-Bergstrom International Airport in 2008.

```
# Plotting functions
plot_delays <- function(dcol, col, colName){
  ## This function plots delay variables
  # --Model
  threshold = 45
  delays <- data[eval(dcol) > threshold,]
  # --Clean Delays data of high outliers and NA
  delays <- delays[eval(col) < unname(quantile(eval(col), 0.99, na.rm =
TRUE)),]
  delays <- delays[! is.na(eval(col)),]
  # --Display Results
  print(round(table(delays$DayOfWeek) / dim(delays)[1] * 100.0, 2))
  print(summary(eval(col)))
  par(mfrow=c(2,2))
  barplot(table(delays$DayOfWeek), xlab="Days of the Week (M-Su)",
```

```
ylab="Count")
  hist(eval(col), xlab = colName)
  boxplot(eval(col) ~ delays$DayOfWeek, xlab = "Day of the Week (M-Su)" ,
ylab = colName)
  plot(delays$DepTime, eval(col), xlab = "Departure Time" , ylab = colName)
}

##
##     1     2     3     4     5     6     7
## 16.34 14.36 12.39 15.18 17.35  9.50 14.88
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   46.00   58.00   78.00   96.13  116.00  328.00
```



Given the preliminary exploratory data analysis, we found out that Friday on average has the worst delay statistics, with most flights having longer than 45 minute delays. Separately, the graph on the bottom right showcases that departure time also affects the departure delay time of a flight.

Delay Heatmap

Given the summary above, we would like to create several delay heatmaps to help business travelers flying out of Austin minimize delay time by choosing the optimal time, day, and Airline combination.

We create heatmaps for the top 5 airlines with the most number of flights in AUS:

1. Southwest Airlines (WN)

2. American Airlines (AA)
3. Continental Airlines (CO)
4. Mesa Airlines (YV)
5. JetBlue Airways (B6)

These heatmaps will show the possibility of a flight having a delay of more than 45 minutes, given the flight's departure time and day of week. The darker the color, the higher possibility of delay.

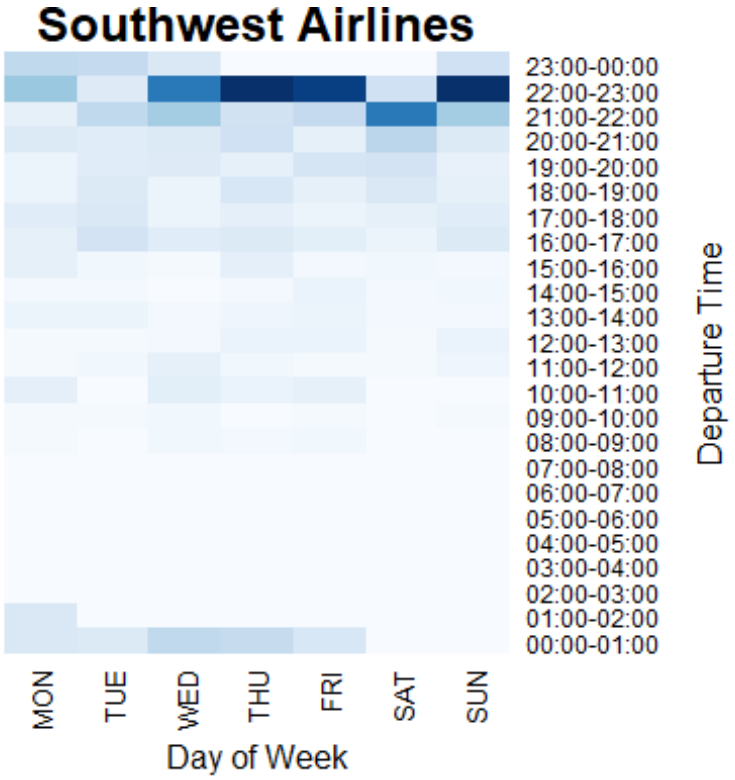
```
# CREATE A HEATMAP
# Create a Day of Week by Departure Time matrix of Average Delays (in
minutes)
# for a particular airline
airlines <- c("WN", "AA", "CO", "YV", "B6")
carrierCodeLookup <- c("Southwest Airlines",
                       "American Airlines",
                       "Continental (UA)",
                       "Mesa Airlines",
                       "JetBlue")

airlineHeatMap <- function(cc, airlineName, timeIntervalInMinutes,
thresholdInMinutes){

  # Setup
  timeInterval_converted <- timeIntervalInMinutes*(5.0/3.0) # Convertion from
base 60 (time) to base 100 (0-2400)
  threshold <- thresholdInMinutes*(5.0/3.0) # How many minutes late are we
counting, converted to base 100
  numRows <- 2400/(timeInterval_converted)
  values <- c()

  # Get Dataframe of just Carrier Code cc
  cc_data <- data[data$UniqueCarrier == cc, c("DayOfWeek", "DepTime",
"DepDelay")] # TODO Limit this to just the necessary list of columns

  interval <- seq(from=0, to=2400, by=timeInterval_converted)
  l_interval <- length(interval)

  # For each day of the week...
  for (day in c(1,2,3,4,5,6,7)){
    cc_data_day <- cc_data[cc_data$DayOfWeek == day,]

    # For each time period...
    for (t in c(2:l_interval-1)){
      cc_data_day_t <- cc_data_day[interval[t] < cc_data_day$DepTime &
cc_data_day$DepTime < interval[t+1],]

      # Get the Percent Chance of Delays * Avg. Duration of Delays in minutes
      totalFlights_t <- dim(cc_data_day_t)[1]
```

```
      delays <- cc_data_day_t[cc_data_day_t$DepDelay >= threshold,"DepDelay"]
      numDelays_t <- length(delays[!is.na(delays)])
      sumDelays_t <- sum(delays, na.rm=TRUE)

      #delayIntensity <-
(numDelays_t/totalFlights_t)*(sumDelays_t/numDelays_t)
      delayIntensity <- numDelays_t/totalFlights_t
      if (is.na(delayIntensity)){delayIntensity = 0}
      values <- c(values, delayIntensity) # TESTING: paste(day,
totalFlights_t, sep=':')
    }
  }

  # Create Matrix with Values inside
  return <- matrix(data=values, nrow=numRows, ncol=7)
}

#create datafrmae that with
WN_m <- airlineHeatMap(airlines[1],
                       carrierCodeLookup[1],
                       60,
                       45)
AA_m <- airlineHeatMap(airlines[2],
                       carrierCodeLookup[2],
                       60,
                       45)
CO_m <- airlineHeatMap(airlines[3],
                       carrierCodeLookup[3],
                       60,
                       45)
YV_m <- airlineHeatMap(airlines[4],
                       carrierCodeLookup[4],
                       60,
                       45)
B6_m <- airlineHeatMap(airlines[5],
                       carrierCodeLookup[5],
                       60,
                       45)

#setup for cleaner heatmap
day=c( "MON", "TUE" ,"WED", "THU", "FRI", "SAT", "SUN")
a= c()
for (i in c(0:23)){

  if (i < 10){
    a = c(a, paste("0",i,":00" ,sep="" ))
  }

  else{
    a = c(a, paste(i,":00" ,sep="" ))
```

```
  }

}

b= c()
for (i in c(1:23, 00)){
  if (i < 10){
    b = c(b, paste("0",i,":00" ,sep="" ))
  }
  else{
    b = c(b, paste(i,":00" ,sep="" ))
  }
}


hour = c()
for (i in c(1:24)){
  hour = c(hour, paste(a[i],b[i],sep="-" ))
}
```

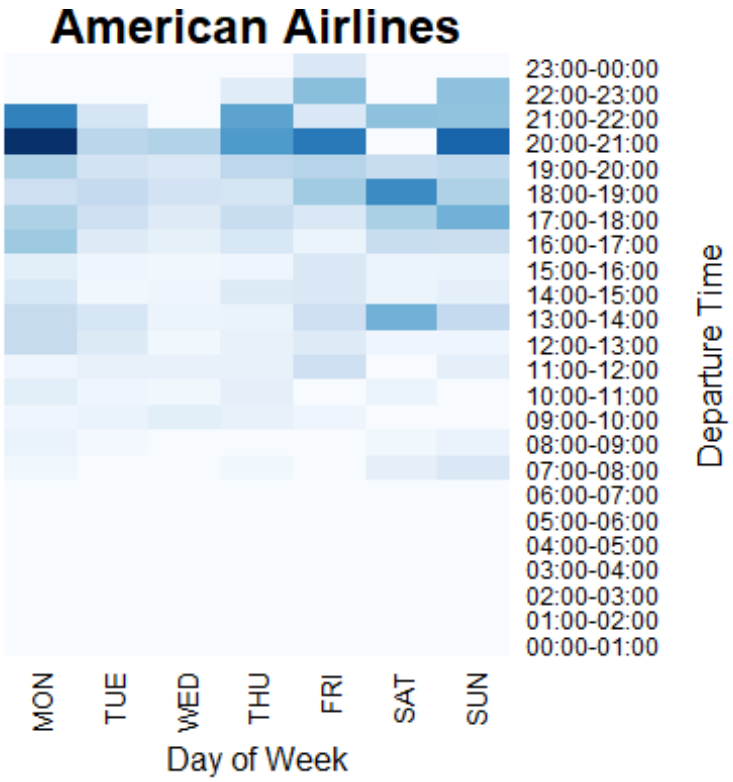
Heatmaps

```
#plot a heatmap for Southwest Airlines

heatmap(WN_m, Rowv=NA, Colv=NA,col= colorRampPalette(brewer.pal(9,
"Blues"))(100),xlab="Day of Week", ylab="Departure Time", main="Southwest
Airlines", scale = 'none', labCol = day, labRow = hour, margins = c(4,7),
cexRow=0.9,cexCol = 1)
```



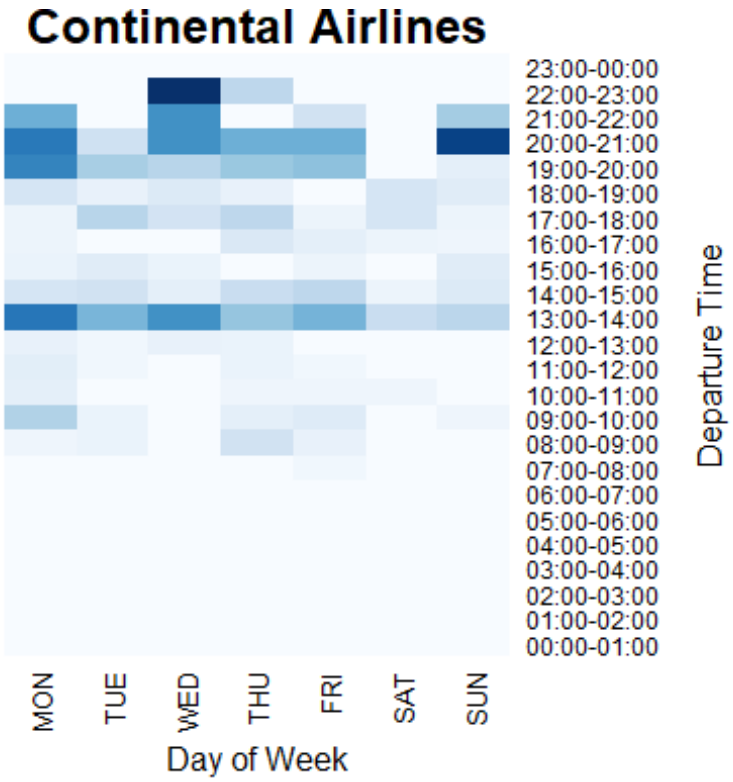
If you want to minimize delay, avoid Thursday, Friday, Sunday 22:00-23:00 flights from Southwest.

```
#plot a heatmap for American Airlines

heatmap(AA_m, Rowv=NA, Colv=NA,col= colorRampPalette(brewer.pal(9,
"Blues"))(100),xlab="Day of Week", ylab="Departure Time", main="American
Airlines", scale = 'none', labCol = day, labRow = hour, margins = c(4,7),
cexRow=0.9,cexCol = 1)
```



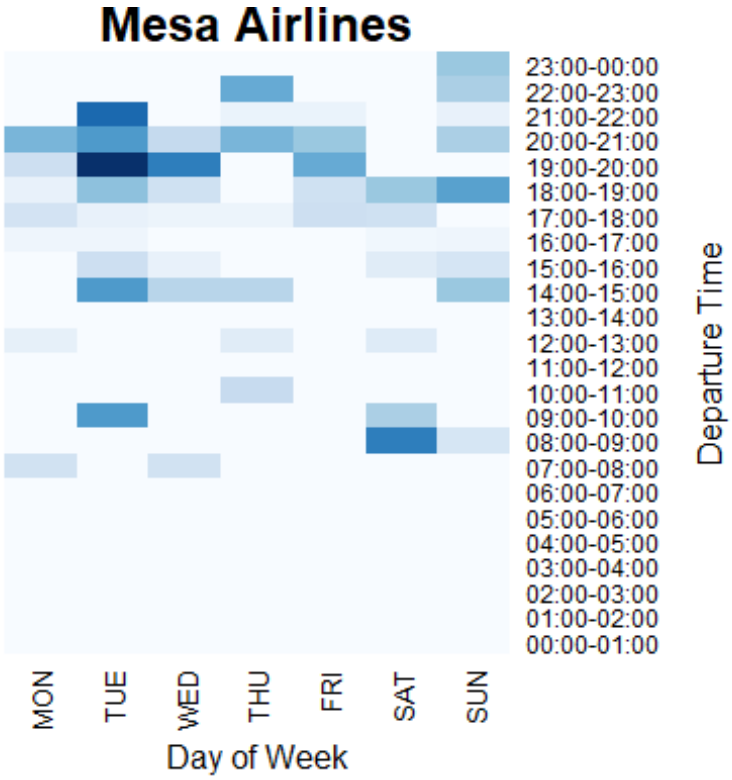
If you want to minimize delay, avoid Sunday and Monday 20:00 - 22:00 flights from American Airlines.

```
#plot a heatmap for Continental Airlines

heatmap(CO_m, Rowv=NA, Colv=NA,col= colorRampPalette(brewer.pal(9,
"Blues"))(100),xlab="Day of Week", ylab="Departure Time", main="Continental
Airlines", scale = 'none', labCol = day, labRow = hour, margins = c(4,7),
cexRow=0.9,cexCol = 1)
```



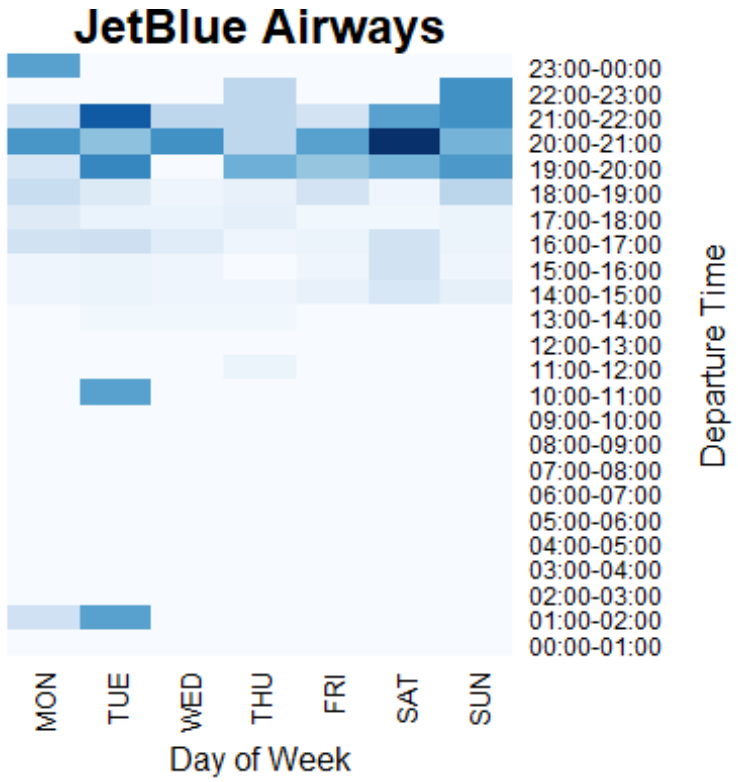
Try avoiding flights departing around 13:00-14:00 on Weekdays, as well as Monday, Wednesday, Sunday late night flights with Continental.

```
#plot a heatmap for Mesa Airlines

heatmap(YV_m, Rowv=NA, Colv=NA,col= colorRampPalette(brewer.pal(9,
"Blues"))(100),xlab="Day of Week", ylab="Departure Time", main="Mesa
Airlines", scale = 'none', labCol = day, labRow = hour, margins = c(4,7),
cexRow=0.9,cexCol = 1)
```



Mesa has several interesting delay blocks. In general, stay away from Mesa if you want to travel on a Tuesday night.

```
#plot a heatmap for Jetblue Airlines

heatmap(B6_m, Rowv=NA, Colv=NA,col= colorRampPalette(brewer.pal(9,
"Blues"))(100),xlab="Day of Week", ylab="Departure Time", main="JetBlue
Airways", scale = 'none', labCol = day, labRow = hour, margins = c(4,7),
cexRow=0.9,cexCol = 1)
```



If you want to minimize the possibility of delay, avoid Tuesday, Saturday, and Sunday 19:00-22:00 flights from Jetblue.

Conclusion

With these heatmaps, travelers will be able to minimize delay time by choosing the best day-time combination with one of the top five airlines.
