STAT480 Homework1

Chenz Zhang, NetID chenziz2

1/26/2019

Contents

Question 1

```
library(RSQLite)

setwd("~/Stat480/RDataScience/AirlineDelays")
```

1.1

```
delay.con <- dbConnect(RSQLite::SQLite(),
                       dbname = "AirlineDelay1980s.sqlite3")

num.total <- dbGetQuery(delay.con,
                        "SELECT COUNT(*) FROM AirlineDelay1980s WHERE Year")

num.total
```

```
##   COUNT(*)
## 1 11555122
```

1.2

```
num.delay15 <- dbGetQuery(delay.con,
                          "SELECT COUNT(*) FROM AirlineDelay1980s
                          WHERE DepDelay>15 AND DepDelay != 'NA'
                          AND DepDelay != 'DepDelay'")

num.delay15
```

```
##   COUNT(*)
## 1  1557190
```

1.3

```
percent.delay15 <- num.delay15/num.total

percent.delay15
```

```
##    COUNT(*)
## 1 0.1347619
```

The delay rate for flights from 1987 to 1989 is relatively low, comparing with the recent delay rate for flights in Bejing International Airport.

Question 2

Method1 by SQL

2.1

```
table.total <- dbGetQuery(delay.con, "SELECT COUNT(*), Month FROM AirlineDelay1980s
                          WHERE Month != 'Month' GROUP BY Month")

table.total
```

```
##    COUNT(*) Month
## 1    876972     1
## 2    807755     2
## 3    880261     3
## 4    832929     4
## 5    852076     5
## 6    837592     6
## 7    858284     7
## 8    872854     8
## 9    839143     9
## 10  1327424    10
## 11  1261485    11
## 12  1308347    12
```

2.2

```
table.delay <- dbGetQuery(delay.con, "SELECT COUNT(*), Month FROM AirlineDelay1980s
                          WHERE DepDelay > 15 AND DepDelay != 'DepDelay' AND
                          DepDelay != 'NA' GROUP BY Month")

table.delay
```

```
##    COUNT(*) Month
## 1    140649     1
## 2    127986     2
## 3    130411     3
## 4     83220     4
## 5     98065     5
## 6    113969     6
## 7    111585     7
## 8    115962     8
## 9     81151     9
## 10   140209    10
## 11   168415    11
## 12   245568    12
```

2.3

```
percent.delay <- cbind(table.delay[1]/table.total[1],table.total[2])

percent.delay
```

```
##       COUNT(*) Month
## 1  0.16038026     1
## 2  0.15844656     2
## 3  0.14815038     3
## 4  0.09991248     4
## 5  0.11508950     5
## 6  0.13606744     6
## 7  0.13000941     7
## 8  0.13285383     8
## 9  0.09670700     9
## 10 0.10562488    10
## 11 0.13350535    11
## 12 0.18769333    12
```

```
dbDisconnect(delay.con)
```

Method2 by big.matrix

2.1

```
library(biganalytics)
```

```
## Loading required package: bigmemory
```

```
## Loading required package: foreach
```

```
## Loading required package: biglm
```

```
## Loading required package: DBI
```

```
x <- read.big.matrix("AirlineData1980s.csv", header = TRUE,
                     backingfile = "air1980s.bin",
                     descriptorfile = "air1980s.desc",
                     type = "integer")
x <- attach.big.matrix("air1980s.desc")

library(foreach)
monthCount <- foreach(i = 1:12, .combine = c) %do%{
  sum(x[,"Month"] == i)
}

monthCount
```

```
##  [1]  876972  807755  880261  832929  852076  837592  858284  872854
##  [9]  839143 1327424 1261485 1308347
```

2.2

```
dow <- split(1:nrow(x),x[,"Month"])

delay.monthCount <- foreach(monthInds = dow, .combine = c) %do% {
  sum(x[monthInds,"DepDelay"] > 15, na.rm = TRUE)
}

delay.monthCount
```

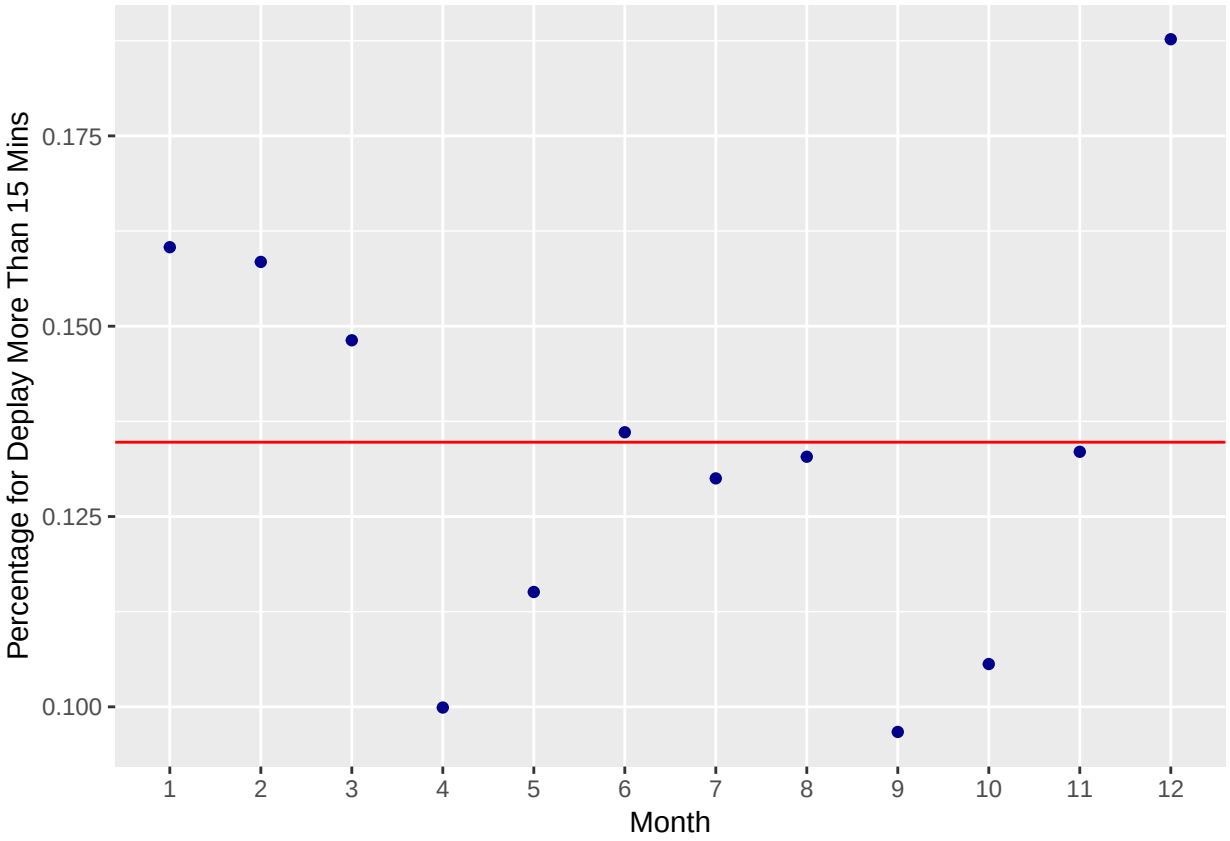
```
##  [1] 140649 127986 130411  83220  98065 113969 111585 115962  81151 140209
## [11] 168415 245568
```

2.3

```
percent.monthdelay15 <- delay.monthCount/monthCount
percent.monthdelay15
```

```
##  [1] 0.16038026 0.15844656 0.14815038 0.09991248 0.11508950 0.13606744
##  [7] 0.13000941 0.13285383 0.09670700 0.10562488 0.13350535 0.18769333
```

```
df <- data.frame(x2 = percent.monthdelay15, Y = factor(seq(1,12,1)))

library(ggplot2)
plot2 <- ggplot(data = df, aes(x = Y, x2)) +
         geom_point(color = "darkblue") +
         xlab("Month") + ylab("Percentage for Deplay More Than 15 Mins")
plot2 <- plot2 + geom_hline(yintercept = percent.delay15[[1]], color = "red")
plot2
```



- The red line in this graph is the overall rate found in exercise 1.
- The darkblue points in this graph are percentage of flights delayed by more than 15 mins by month of year during 1980s.
- From this graph, we can see that delay rates in Jan, Feb, Mar, Jun and Dec are higher than overall delay rate. The rest rates are lower than the overall delay rate.

Question3

3.1

```
y <- attach.big.matrix("air0708.desc")

total0708 <- sum(y[,"Year"] == 2007) + sum(y[,"Year"] == 2008)
total0708
```

```
## [1] 14462943
```

```
delay0708 <- sum(y[,"DepDelay"] > 15, na.rm = TRUE)
delay0708
```

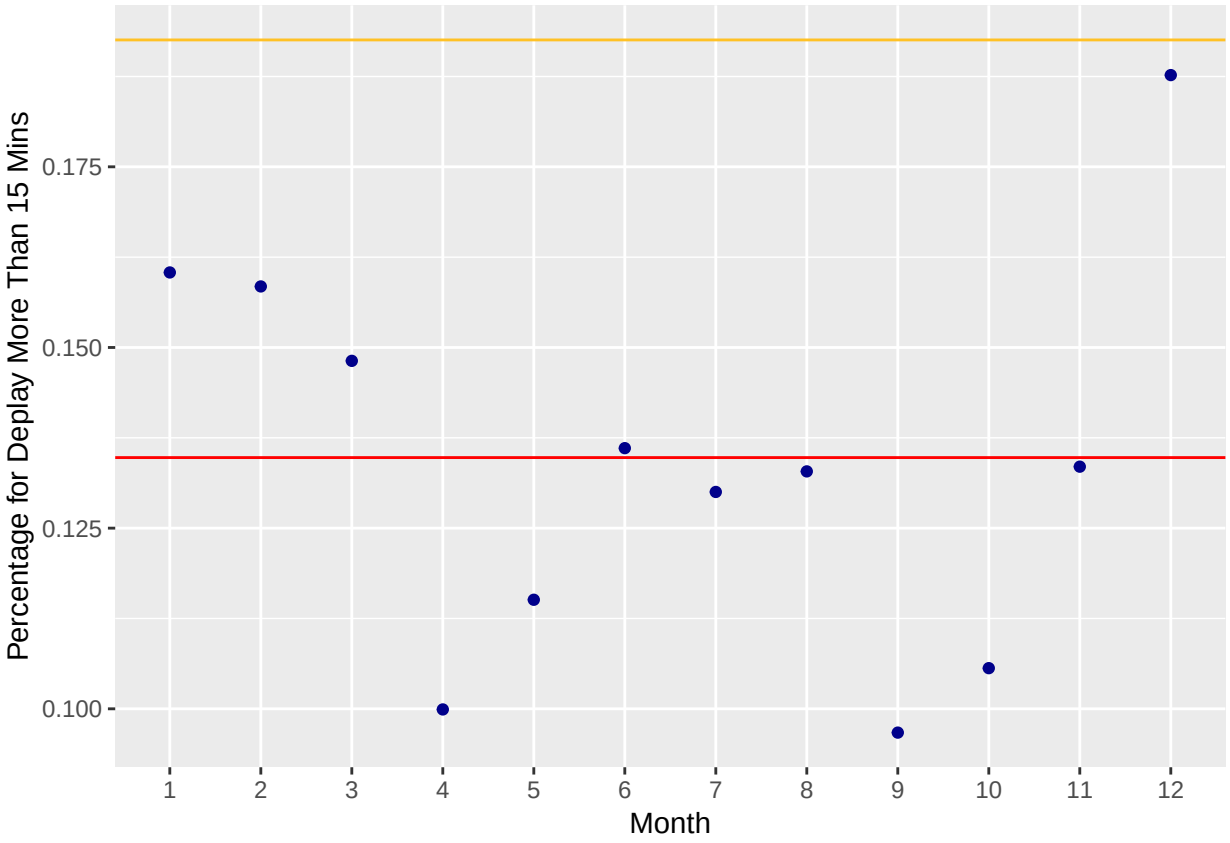
```
## [1] 2784966
```

```
percent.delay0708 <- delay0708/total0708
percent.delay0708
```

```
## [1] 0.1925587
```

3.2

```
plot3 <- plot2 + geom_hline(yintercept = percent.delay0708, color = "goldenrod1")
plot3
```



- The golden line indicate the aggregate percentage of flights delayed by more than 15 mins during 2007 and 2008.
- We can conclude that the aggregate delay rate for 2007-2008 is higher than that of 1987-1989.

Question 4

4.1

```
dow4 <- split(1:nrow(y),y[,"Year"])

yearCount <- matrix(foreach(yearInds = dow4, .combine = c) %do% {
  c(nrow(y[yearInds,]),sum(y[yearInds,"DepDelay"] > 15, na.rm = TRUE))
},2,2)

rownames(yearCount) <- c("Flights","Delays>15mins")
```

```
colnames(yearCount) <- c("2007","2008")
yearCount
```

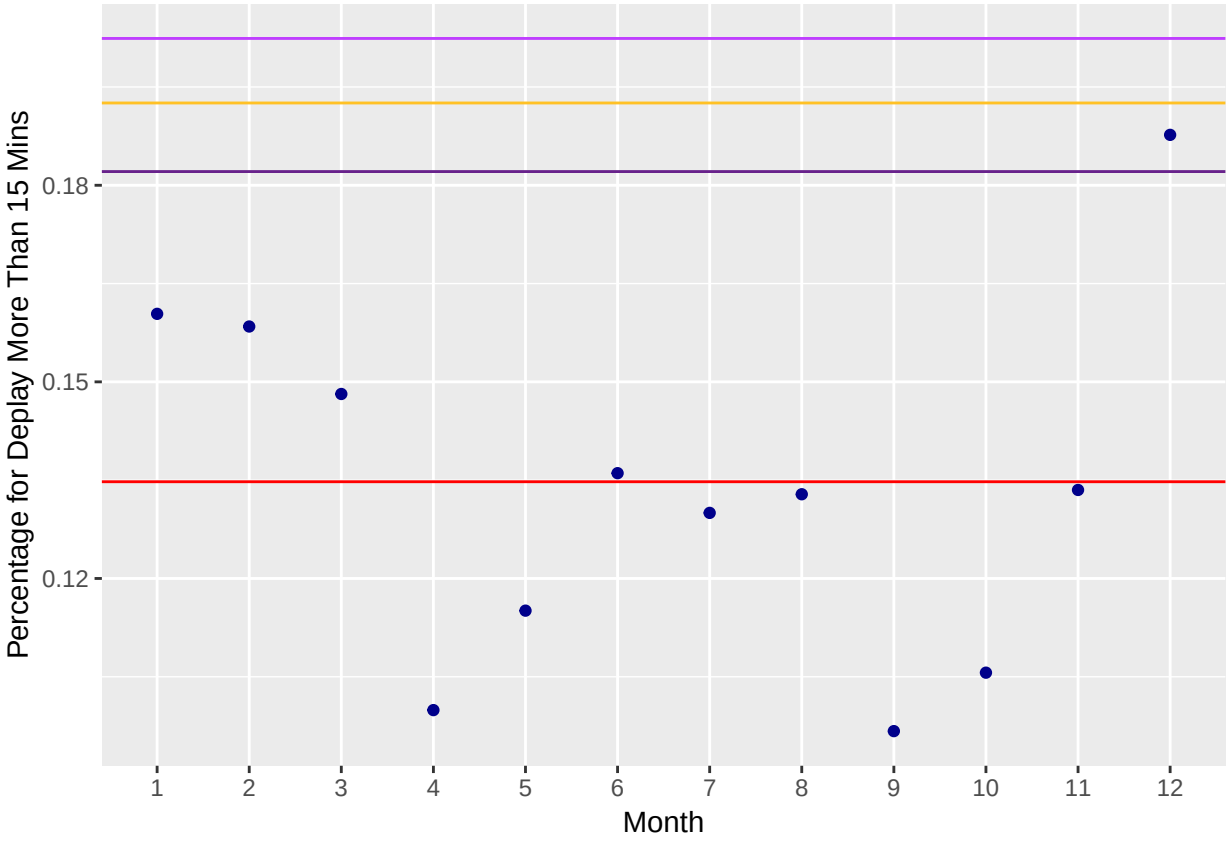
```
##                   2007    2008
## Flights        7453215 7009728
## Delays>15mins 1508570 1276396
```

4.2

```
percent.yeardelay0708 <- (yearCount[2,]/yearCount[1,])
percent.yeardelay0708
```

```
##      2007      2008
## 0.2024053 0.1820892
```

```
plot4 <- plot3 + geom_hline(yintercept = percent.yeardelay0708,
                            color = c("darkorchid1","darkorchid4"))
plot4
```



- The lighter darkorchid line is delay rate from 2007 and the darker darkorchid is delay rate from 2008.
- We can conclude that:
  - Delay rate in 2007 is higher than the aggregate delay rate we got from exercise 3.
  - Delay rate in 2008 is lower than the aggregate delay rate we got from exercise 3.

Question 5

5.1

```
dow5.1 <- split(1:nrow(x),x[,"DayOfWeek"])

percent.weekdelay1980s <- foreach( DayWeekIn1980= dow5.1, .combine = c) %do% {
  sum(x[DayWeekIn1980,"DepDelay"] > 15, na.rm = TRUE)/nrow(x[DayWeekIn1980, ])
}
percent.weekdelay1980s
```

```
## [1] 0.1216016 0.1377796 0.1452894 0.1520203 0.1489151 0.1149401 0.1203486
```

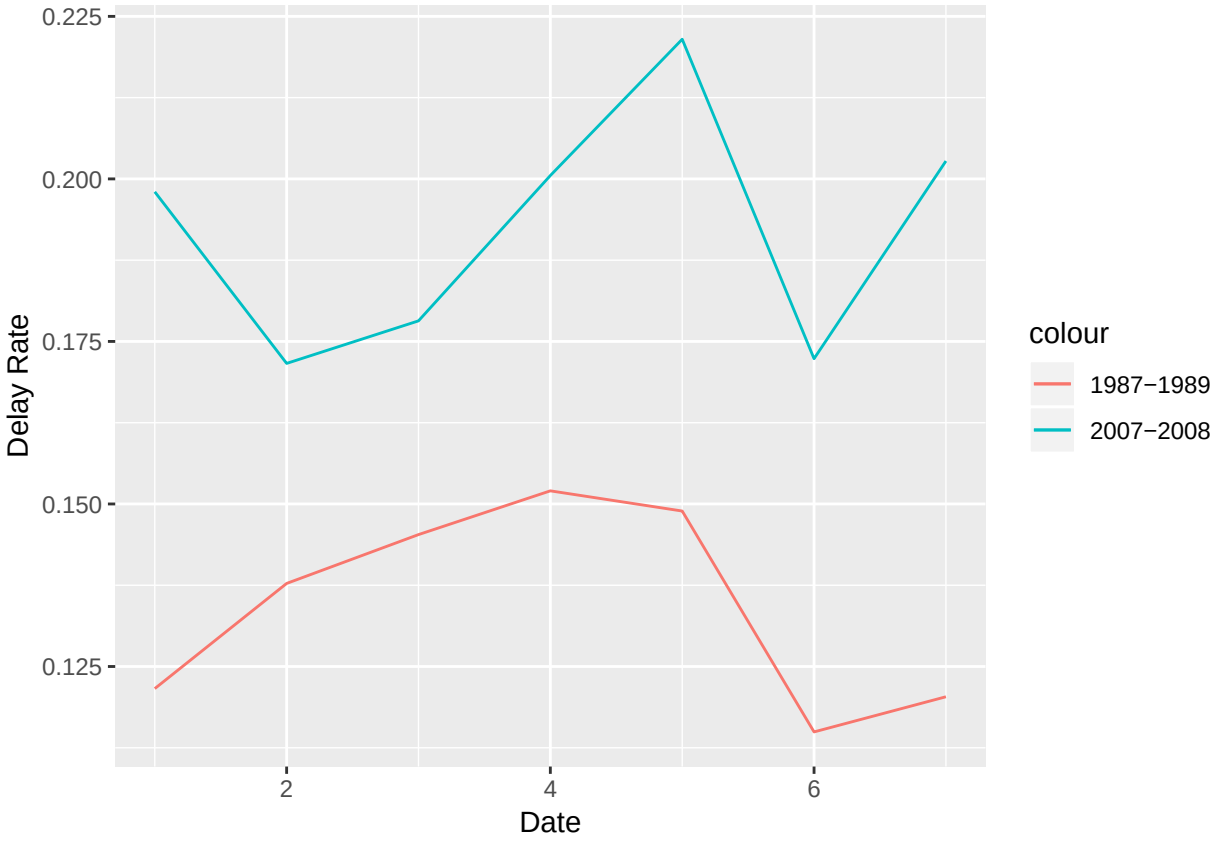
5.2

```
dow5.2 <- split(1:nrow(y),y[,"DayOfWeek"])

percent.weekdelay0708 <- foreach(DayWeekIn0708= dow5.2, .combine = c) %do% {
  sum(y[DayWeekIn0708,"DepDelay"] > 15, na.rm = TRUE)/nrow(y[DayWeekIn0708, ])
}
percent.weekdelay0708
```

```
## [1] 0.1980155 0.1716200 0.1781649 0.2005153 0.2214811 0.1723460 0.2027472
```

5.3

```
df5 <- data.frame(Date = 1:7, x1 = percent.weekdelay1980s, x2 = percent.weekdelay0708)
plot5 <- ggplot(df5, aes(Date)) +
         geom_line(aes(y = x1, colour = "1987-1989")) +
         geom_line(aes(y = x2, colour = "2007-2008")) +
         ylab("Delay Rate")

plot5
```



- Delay rate in 1987-1989 is lower than that in 2007-2008 for each day of week.