STAT480_Homework_1

Bin Feng

```
#include library
library(RSQLite)
library(biganalytics)
```

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

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

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

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

```
#set working directory
require("knitr")
```

```
## Loading required package: knitr
```

```
opts_knit$set(root.dir = "~/Stat480/RDataScience/AirlineDelays")
```

Question 1

This exercise is for aggregate departure delay information for flights from 1987 to 1989 in the data.
(a) Using SQL, obtain the total number of flights in the data in the 1980s.

```
#code below can be referencing to "Chapter5.R" discussed in lecture
#define delay80 as the valuable storing all flights information in the 1980s
delay.con <- dbConnect(RSQLite::SQLite(), dbname = "AirlineDelay1980s.sqlite3")
#calculate the total rows in 1980s, which is the same as the total number of flights in 1980s
dbGetQuery(delay.con, "SELECT COUNT(*) FROM AirlineDelay1980s")
```

```
##   COUNT(*)
## 1 11555123
```

```
#total number of flights in the 1980s is 11555123.
```

(b) Using SQL, obtain the number of flights with departure delayed by more than 15 minutes in the 1980s in the data.

```
#code below can be referencing to "Chapter5.R" discussed in lecture
#query the daparture delay data through the SQL connection
dbGetQuery(delay.con, "SELECT COUNT(*), DepDelay FROM AirlineDelay1980s WHERE DepDelay > 15")
```

```
##   COUNT(*) DepDelay
## 1  1701204       60
```

(c) Comment on the percentage of flights with departure delayed by more than 15 minutes during that time period.

```
#Calculate the percetage of depature delayer by more than 15 min
percentage <- 1701204/11555123
percentage
```

```
## [1] 0.1472251
```

```
#Comment!!
```

Question 2

Now we look at the similar delay information by month during that period. (Note: This is just by month, not by month and year. For instance, flights for January 1987, January 1988, and January 1989 will be aggregated together.)
(a) Obtain a table for the total number of flights in our data by month in the 1980s from the data.

```
#using SQL language FROM and GROUP BY
##how can we get rid of the header row
delay_month <- dbGetQuery(delay.con,
                          "SELECT COUNT(*), Month FROM AirlineDelay1980s GROUP BY Month")
```

```
## Warning in result_fetch(res@ptr, n = n): Column `Month`: mixed type, first
## seen values of type integer, coercing other values of type string
```

(b) In a separate table, obtain the number of flights by month with departure delayed by more than 15 minutes in the 1980s in the data.

```
delay_month_gr15 <- dbGetQuery(delay.con,
                               "SELECT COUNT(*), Month FROM AirlineDelay1980s WHERE DepDelay > 15 GROUP
```

```
## Warning in result_fetch(res@ptr, n = n): Column `Month`: mixed type, first
## seen values of type integer, coercing other values of type string
```

(c) From the results in parts a and b, programmatically calculate the percentage of flights delayed by more than 15 minutes by month of year during that time period, and comment on how the monthly rates compare to the overall rate found in exercise 1.

```
##what does it mean by programmatically
percentage_month <- integer(12)
for (i in 1:12){
  percentage_month[i] <- delay_month_gr15[i,1]/sum(delay_month[i,1])
}
#close connection
dbDisconnect(delay.con)
#Comment!!
```

Question 3

Now we look at aggregate flight data for 2007 and 2008
(a) Obtain the total number of flights in 2007 and 2008, the number of flights delayed by more than 15 minutes during that time period, and the percentage of flights delayed by more than 15 minutes during that time period.

```
flight0708 <- attach.big.matrix("air0708.desc")
#count the total number of flight using dim(), -1 to exclude the header count
total_fight <- dim(flight0708)[1] - 1
#count the number of delays that are more than 15min
#there are differences between na.rm and na.omit
delay_0708_gr15 <- sum(flight0708[,"DepDelay"] > 15, na.rm=TRUE)
#calculate the percentile
percentage_gr15 <- delay_0708_gr15/total_fight
```

(b) Comment on how this delay rate compares with the rate found for the 1987-1989 flights.

```
#comment!!!
```

Question 4

Now we look at the delay rate per year for 2007 and 2008. (a) For each year from 2007 to 2008, calculate the number of flights and the number of flights delayed by more than 15 minutes. (You should have counts for 2007 and counts for 2008.) Be sure to use efficient programming techniques.

```
# calculate by each year
flight_by_year <- split(1:nrow(flight0708), flight0708[,"Year"])
names(flight_by_year) <- c("2007", "2008")
##for checking flight_by_year$"2007"[1:10], why only shows 1,2,3,4,5...
##since we only have two section, is split-apply-combine necessary here??
##TotalCount also include header??
TotalCount <- foreach(dayInds = flight_by_year, .combine = c) %do% {
  length(dayInds)
}
TotalCount
```

```
## [1] 7453215 7009728
```

```
##how to use split-apply-combine in here??
##can we use this library??
library(foreach)
DelayCount <- foreach(i = 1:2, .combine=c) %do% {
  sum(flight0708[,"DepDelay"] > 15 & flight0708[,"Year"] == 2006 + i, na.rm=TRUE)
}
DelayCount
```

```
## [1] 1508570 1276396
```

(b) Compute the percentage of flights with departure delayed by more than 15 in each of those two years and compare the annual rates with the aggregate rate found in exercise 3.

```
rate07 <- DelayCount[1]/TotalCount[1]
rate08 <- DelayCount[2]/TotalCount[2]
```