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
```

```
library(foreach)
require("knitr")
```

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

```
#set working directory
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.

```
#build connection to the database
delay.con <- dbConnect(RSQLite::SQLite(), dbname = "AirlineDelay1980s.sqlite3")
#calculate the total rows in 1980s, which is the total number of flights in 1980s plus 1
total_80s <- dbGetQuery(delay.con,
                        "SELECT COUNT(*) FROM AirlineDelay1980s") - 1
total_80s
```

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

Based on th output, the total number of flights in the data in the 1980s is 11555122. "-1" in the code is to subtract the additional header line included in the database.

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

```
#query the daparture delay data through the SQL connection
delay_gr15_80s <- dbGetQuery(delay.con,
                             "SELECT COUNT(*) FROM AirlineDelay1980s WHERE DepDelay > 15")
delay_gr15_80s
```

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

Based on th output, the number of flights with departure delayed by more than 15 minutes in the 1980s is 1701204.

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

```
#Calculate the percentage of depature delayed by more than 15 min
delay_per_80s <- delay_gr15_80s/total_80s * 100
delay_per_80s
```

```
##   COUNT(*)
## 1 14.72251
```

Based on the output, the percentage of flights with departure delayed by more than 15 minutes during that time period is 14.72251%. I think such percentage is in a moderate delay level. On the one side, most of flights can depart on time. On the other side, there is still a certain numnber of passengers would need to wait for more than 15 minutes before the departure.

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
total_month_80s <- 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
```

```
total_month_80s
```

```
##    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
## 13        1     0
```

The table for total number of flights by month is shown above. Note that the additional last line in the table is because of the header line in the database.

(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_80s <- 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
```

```
delay_month_gr15_80s
```

```
##    COUNT(*) Month
## 1    163444     1
## 2    143415     2
## 3    156392     3
## 4     87720     4
## 5    103215     5
## 6    118894     6
## 7    117288     7
## 8    121132     8
## 9     88384     9
## 10   149727    10
## 11   180620    11
## 12   270972    12
## 13        1     0
```

The table for total number of flights by month with departure delayed by more than 15 minutes is shown above. Note that the additional last line in the table is because of the header line in the database.

(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.

```
#calculate the percentage by matrix division
deley_per_month_80s <- integer(12)
deley_per_month_80s <- delay_month_gr15_80s[1:12,1] / total_month_80s[1:12,1] * 100
#construct the table with monthes included
deley_per_month_80s <- cbind(deley_per_month_80s, delay_month_gr15_80s[1:12,2])
colnames(deley_per_month_80s) <- c("delay_percentage", "month")
deley_per_month_80s
```

```
##       delay_percentage month
##  [1,]         18.63731     1
##  [2,]         17.75476     2
##  [3,]         17.76655     3
##  [4,]         10.53151     4
##  [5,]         12.11336     5
##  [6,]         14.19474     6
##  [7,]         13.66541     7
##  [8,]         13.87769     8
##  [9,]         10.53265     9
## [10,]         11.27952    10
## [11,]         14.31805    11
## [12,]         20.71102    12
```

```
#close connection
dbDisconnect(delay.con)
```

The delay percentage by month is shown above. Comparing with the overall rate (14.72251%), note that monthes (1, 2, 3, 12) have higher delay rate while monthes (4, 5, 6, 7, 8, 9, 10, 11) have lower delay rate. Generally, the delay rates by month form a U-shape line, indicating that both at the beginning and towards the end of a year will have higher percentage of being delay by more than 15 minutes. During the middle of a year, delay rates are usually lower. The reason behind having higher rate at the beginning and at the end of a year may because: 1. people travel more during holiday seasons; 2. people need some time to warm up after the holiday season; 3. winter usually brings worse weather conditions like heavy snow or extremely low

temperature.

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.

```
#Attach the same big matrix to flight0708 using the descriptor file without creating any new large matr
flight0708 <- attach.big.matrix("air0708.desc")
#count the total number of flight using dim(), -1 to exclude the header count
total_0708 <- dim(flight0708)[1] - 1
total_0708
```

```
## [1] 14462942
```

```
#count the number of delays that are more than 15min
#note the differences between na.rm and na.omit
delay_gr15_0708 <- sum(flight0708[,"DepDelay"] > 15, na.rm=TRUE)
delay_gr15_0708
```

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

```
#calculate the percentage
deley_per_0708 <- delay_gr15_0708 / total_0708 * 100
deley_per_0708
```

```
## [1] 19.25587
```

(b) Comment on how this delay rate compares with the rate found for the 1987-1989 flights.
The delay rate for 1987-1989 is 14.72251%. The dalay rate for 2007 and 2008 is 19.25587%. Comparing these two rate, note that the delay rate has increased significantly by the time reaching 2007 and 2008.

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, use splite-apply-combine method here. Since there are only two years to split,
flight_by_year <- split(1:nrow(flight0708), flight0708[,"Year"])
names(flight_by_year) <- c("2007", "2008")

#Substract 1 in flight counting for 2007 to remove the additional number for the header line. %dopar% i
total_year_0708 <- foreach(yrInds = flight_by_year, .combine = c) %dopar% {
  length(yrInds)
}
```

```
## Warning: executing %dopar% sequentially: no parallel backend registered
```

```
total_year_0708[1] <- total_year_0708[1] - 1
total_year_0708
```

```
## [1] 7453214 7009728
```

```
#calculate the delay flight by each year. %dopar% is used here for potential parallel computing.
delay_gr15_year_0708 <- foreach(i = 1:2, .combine=c) %dopar% {
  sum(flight0708[,"DepDelay"] > 15 & flight0708[,"Year"] == 2006 + i, na.rm=TRUE)
}
delay_gr15_year_0708
```

```
## [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.

```
#compute the percentage as follow for 2007 and 2008
delay_per_year_07 <- delay_gr15_year_0708[1] / total_year_0708[1] * 100
delay_per_year_08 <- delay_gr15_year_0708[2] / total_year_0708[2] * 100
delay_per_year_07
```

```
## [1] 20.24053
```

```
delay_per_year_08
```

```
## [1] 18.20892
```

Comparing the annual rate of 2007(20.24053%) and 2008(18.20892%) with the aggregate rate (19.25587%), note that the annual rate for 2007 is higher than the aggregate and the annual rate for 2008 is lower than the aggregate. Such observation indicates that the severity of flight delay may has been moderated.

Question 5

This exercise is to compare delay rates by day of week from 1987 to 1989 with delay rates by day of week from 2007 to 2008 within the data provided.

a) Calculate the percentage of flights delayed by more than 15 minutes for each day of the week for the period from 1987 to 1989 in the data provided.

```
#build connection to the database
delay.con <- dbConnect(RSQLite::SQLite(), dbname = "AirlineDelay1980s.sqlite3")
#calcute the total flight by week
total_week_80s <- dbGetQuery(delay.con,
                             "SELECT COUNT(*), DayOfWeek FROM AirlineDelay1980s GROUP BY DayOfWeek")
```

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

```
#calculate the delay rate by week, the last addition line is due to the header line
delay_week_gr15_80s <- dbGetQuery(delay.con,
                                  "SELECT COUNT(*), DayOfWeek FROM AirlineDelay1980s WHERE DepDelay > 15 G
```

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

```
#Calculate the percentage
delay_per_week_80s <- delay_week_gr15_80s[1:7,1] / total_week_80s[1:7,1] * 100
delay_per_week_80s <- cbind(delay_per_week_80s, delay_week_gr15_80s[1:7,2])
colnames(delay_per_week_80s) <- c("delay_percentage", "DayOfWeek")
delay_per_week_80s
```

```
##      delay_percentage DayOfWeek
## [1,]         13.60409         1
## [2,]         15.00646         2
## [3,]         15.63701         3
## [4,]         16.41503         4
## [5,]         16.18207         5
## [6,]         12.81771         6
## [7,]         13.15206         7
```

```
#close connection
dbDisconnect(delay.con)
```

b) Repeat part a for 2007 and 2008 data.

```
#Calculate the total flight by week
flight_by_week <- split(1:nrow(flight0708), flight0708[,"DayOfWeek"])
names(flight_by_week) <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")

#Substract 1 in flight counting for 2007 to remove the additional number for the header line. %dopar% i
total_week_0708 <- foreach(wkInds = flight_by_week, .combine = c) %dopar% {
  length(wkInds)
}
total_week_0708[1] <- total_week_0708[1] - 1

#Calculate delay flight by week. %dopar% is used here for potential parallel computing.
delay_gr15_week_0708 <- foreach(i = 1:7, .combine=c) %dopar% {
  sum(flight0708[,"DepDelay"] > 15 & flight0708[,"DayOfWeek"] == i, na.rm=TRUE)
}

#Calculate the percentage
delay_per_week_0708 <- delay_gr15_week_0708 / total_week_0708 * 100
delay_per_week_0708 <- cbind(delay_per_week_0708, c(1,2,3,4,5,6,7))
colnames(delay_per_week_0708) <- c("delay_percentage", "DayOfWeek")
delay_per_week_0708
```

```
##      delay_percentage DayOfWeek
## [1,]         19.80156         1
## [2,]         17.16200         2
## [3,]         17.81649         3
## [4,]         20.05153         4
## [5,]         22.14811         5
## [6,]         17.23460         6
## [7,]         20.27472         7
```

c) Comment on similarities and differences in the delay rate on particular days of week between the two time periods.
For similarity, note that both time periods have a higher delay rate on DayOfWeek (4-Thursdat, 5-Friday) and a lower delay rate on DayOfWeek (6-Saturday). For differences, 0708 flight data shows a higher delay rate on DayOfWeek (1-Monday, 7-Sunday) and 80s flight data shows a lower delay rate. Also, 0708 flight data indicate a lower delay rate on DayOfWeek (2-Tuesday, 3-Wednesday) while 80s flight data shows a higher delay rate.