Problem Set 6

github: stephaniesookim

November 3, 2014

/raggedright

Number 1

```
library(RSQLite)
database <- dbConnect(SQLite(), dbname = "Airline_db.sqlite")
years <- 1987:2008

system.time({

# download each file in the url by year
for (i in 1:length(years)) {
      url <- paste("http://www.stat.berkeley.edu/share/paciorek/", years[i], ".csv.bz2", sep="")
      filename <- paste(years[i], ".csv.bz2", sep="")
      download.file(url, filename)

# read the table using the bzfile function and save it to a variable "data"
# I used the colClasses argument to control the type for the fields in the table
      data <- read.table(bzfile(filename), sep = ",", head = TRUE,
                                 colClasses = c('integer', rep('factor', 3), rep('integer', 4),
                                                'factor', 'integer', 'factor', rep('integer',5),
                                                'factor', 'factor', rep('integer', 4), 'factor',
                                                rep('integer', 6)))

# I replaced the missing values (N/A) in DepDelay column with 0
      data$DepDelay[is.na(data$DepDelay)] <- 0

# create a new database by writing a table using dbWriteTable function
      dbWriteTable(conn = database, name = "data", value = data,
                   row.names=FALSE, header=TRUE, append=TRUE,
                   colClasses = c('integer', rep('factor', 3), rep('integer', 4),
                                  'factor', 'integer', 'factor', rep('integer',5),
                                  'factor', 'factor', rep('integer', 4), 'factor',
                                   rep('integer', 6)))
        }
})

# now go back to Ubuntu on terminal and test for the size of the database
# ls -l Airline_db.sqlite
# -rw-r--r-- 1 ubuntu ubuntu 9631064064
# The database file is 9631064064 bytes
```

Number 2

Number 3

Number 4

```
# download the zipped from the url
wget http://www.stat.berkeley.edu/share/paciorek/1987-2008.csvs.tgz

# unzip the file using the bzip2 function
bzip2 -d $(tar zxvf 1987-2008.csvs.tgz)

# extract the SFO/OAK flights, save it to 4.txt, and measure the time it takes to run the code
time awk -F ',' -v OFS=',' '{ if ($17=="SFO" || $17=="OAK") print $0;}' *.csv > 4.txt

# real 4m1.146s
# user 3m51.658s
# sys  0m6.209s
```