Problem Set 6

Thibault Doutre, Student ID 26980469

STAT243 : Statistical Computing
University of California, Berkeley

I worked on my own.

1 Airline Database

I create a script which I can execute via EC2. The script builds a database "airline.db" and store every file into a single table, using RSQLite. In order to do it, I first download the files in a directory named "data" and create the database. In this database, I store a table called "airline" in wich I will append the data for every year.

```
#Download
url = 'http://www.stat.berkeley.edu/share/paciorek/1987-2008.csvs.tgz'
download.file(url, ".file")

# Untar
untar(".file", compressed = 'bzip2', exdir = "./data/")

# Create Database
library(RSQLite)
drv <- dbDriver("SQLite")
db <- dbConnect(drv, dbname = "airline.db")

# Create airline Table
dbSendQuery(conn = db,
            [1325 chars quoted with '"']
)
```

Then, for every year:

- Open a connection to the file with bzcat
- Get the data and store it into a variable "line"
- Create a temporary table from line
- Append this table to "airline" using INSERT
- Remove the temporary table
- Close the connection

```
# Append years into airline table
for (i in 1987:2008){
  con=pipe(paste("bzcat ",i,".csv.bz2",sep=""), open = 'r')
  lines = read.csv(con, header = TRUE)
  dbWriteTable(db, paste("y",i,sep=""),lines)
  dbSendQuery(db,paste("INSERT INTO airline SELECT * FROM y",i,sep=""))
  dbRemoveTable(db,paste("y",i,sep=""))
  close(con)
}
```

Then, I print the size of the database:

```
# Size in Gb
file.size("./airline.db")/2^30

# 9.273604
```

We can see that the database is 9 Gb big, it is less than the original CSV of 12 Gb but significantly bigger than the bzipped file of 1.7 Gb.