Big data analytics with R

1. Create a subset of your flights data showing flights that were not canceled and origin flight date, airline id, city, and destination city?

```
#Install the libraries and call them
library(ff)
library(ffbase)

#Load the package and create directory where data will be stored
system("mkdir ffdf")
```

We do subsetting for data management and ff, ffbase packages support subsetting of ffdf objects through the subset.ffdf() method.

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

```
#indicate the path to the newly created directory
dir_air=paste0(getwd(),"/ffdf")
dir_air
```

```
## [1] "C:/Users/Mr.Semicolon/Desktop/R 1/Student/ffdf"
```

```
options(fftempdir = dir_air)

#Now I can import airline data set
airline.ff<- read.table.ffdf(file="flights_sep_oct15.txt",
                             sep=",", VERBOSE=TRUE,
                             header=TRUE, next.rows=100000, colClasses=NA)
```

```
## read.table.ffdf 1..100000 (100000)  csv-read=0.53sec ffdf-write=0.33sec
## read.table.ffdf 100001..200000 (100000)  csv-read=0.72sec ffdf-write=0.23sec
## read.table.ffdf 200001..300000 (100000)  csv-read=0.56sec ffdf-write=0.21sec
## read.table.ffdf 300001..400000 (100000)  csv-read=0.59sec ffdf-write=0.23sec
## read.table.ffdf 400001..500000 (100000)  csv-read=0.68sec ffdf-write=0.25sec
## read.table.ffdf 500001..600000 (100000)  csv-read=0.73sec ffdf-write=0.2sec
## read.table.ffdf 600001..700000 (100000)  csv-read=0.58sec ffdf-write=0.2sec
## read.table.ffdf 700001..800000 (100000)  csv-read=0.61sec ffdf-write=0.22sec
## read.table.ffdf 800001..900000 (100000)  csv-read=0.57sec ffdf-write=0.26sec
## read.table.ffdf 900001..951111 (51111)  csv-read=0.31sec ffdf-write=0.24sec
##  csv-read=5.88sec  ffdf-write=2.37sec  TOTAL=8.25sec
```

```
#See the number of columns and rows
dim(airline.ff)
```

```
## [1] 951111     28
```

```
#subset all records were not canceled and origin flight date, airline id, city, and destination city.
airline_subset.ff <- subset.ffdf(airline.ff, CANCELLED == 0,
                        select = c(FL_DATE, AIRLINE_ID,
                                   ORIGIN_CITY_NAME,
                                   DEST_CITY_NAME))
```

```r
dim(airline_subset.ff)
```

```
## [1] 946582      4
```

2. Save the result from 1 in 4 separate files corresponding to the variables in the subset.

```r
# 4 files (one for each column) created in my ffdb directory
td <- tempfile()
save.ffdf(airline_subset.ff , overwrite=TRUE, dir=td)
dir(td)
```

By default save.ffdf() function saves the subsets to new folder called ffdb in our working directory. But here I'm using tempfile() to store my subsets.

```
## [1] "airline_subset.ff$AIRLINE_ID.ff"
## [2] "airline_subset.ff$DEST_CITY_NAME.ff"
## [3] "airline_subset.ff$FL_DATE.ff"
## [4] "airline_subset.ff$ORIGIN_CITY_NAME.ff"
```

3. Remove the airline__subset.ff file from your workspace and then navigate to the stored copy and restore it (remember to use the tab function to expand the path to the file). Show and comment on your code line by line.

```r
#Remove airline_subset.ff from my workspace
rm(airline_subset.ff)
#Load the file back by giving the path to my temporary directory(td)
load.ffdf(dir=td)
dim(airline_subset.ff)
```

```
## [1] 946582      4
```