Stat 243
Problem set 2

A. Strand, ID: 540441, GitHub: AndreasStrand

September 21, 2015

1 Problem

(a) To create the random sample with the desired columns a combination of Bash and R was used. First random columns were drawn in Bash, then the data was loaded in to Rand the desired columns were stored in a data frame.

(b) The elapsed time for using read.csv() to load the csv-file into a data frame was 1.092. Another way of doing this is using readLines(). This function is faster in loading the data into R, with a elapsed time of 0.49. The downside of readLines() is that the structure is just a vector of character strings representing each line, and this is not desirable when handling the data after.

(c) Instead of loading all columns in to R, the desirable columns could be obtained in bash using cut. Then there would be less copies made of the data. Also the randomization process could be made faster by extracting random indices instead of sorting the array randomly and then pulling out the first 10000 rows.

(d) For this part two tests were run. One checking the number of unique values in each column and the other checking for number of NAs in each column. There was several columns with the same amount of NAs as another column. This means that if a product has a piece of information that is not available, then it would be likely to lack other information as well.

```
########## Bash code ###########
# Creating work folder, downloading and unzipping data
mkdir ps1
cd ps1
wget www.stat.berkeley.edu/share/paciorek/ss13hus.csv.bz2
mv*.zip hhdata.zip
unzip hhdata.zip

# Extracting 10000 random lines, while keeping the header
head -n1 ss13hus.csv >smpl.csv
sort -R hhdata.csv | head -n 10000 >>smpl.csv

########### R code #############
# From the preparations in bash 10000 random lines were selected in addition
# to the header. Now the relevant columns are extracted.
# Comparing runtime of read.csv() and read.lines()
cols = c("ST", "NP", "BDSP", "BLD", "RMSP", "TEN", "FINCP",
 "FPARC", "HHL", "NOC", "MV", "VEH", "YBL")
setwd("~/Documents/ps2")
readcsvTime = system.time(hhdata <- read.csv(file="smpl.csv", header=TRUE, sep=","))
hhCols = hhdata[, cols]
readlinesTime = system.time(hhdata <- readLines("smpl.csv"))

# Checking unique values in each column
sapply(hhCols, function(x) length(unique(x)))
# Checking for elements that are not NA
sapply(hhCols, function(x) sum(!is.na(x)))
```