# Assignment 10

library(microbenchmark)  
  
  
#1a.  
d <- read.csv('usnews\_r.csv')  
  
f <- function(df)  
{   
 length(which(df$Instructional.expenditure.per.student > df$Out.of.state.tuition))  
}  
  
g <- function(df)  
{  
 c <- 0  
 for(i in 1:dim(df)[1])  
 {  
 if(any(is.na(df[i,c('Instructional.expenditure.per.student', 'Out.of.state.tuition')])))  
 {  
 next   
 }  
 if(df[i,'Instructional.expenditure.per.student'] > df[i,'Out.of.state.tuition'])  
 {  
 c = c + 1  
 }  
 }  
 c  
}  
  
#1b.  
system.time(for(i in 1:1000)f(d))

## user system elapsed   
## 0.03 0.00 0.04

system.time(for(i in 1:1000)g(d))

## user system elapsed   
## 188.01 0.03 188.22

#1c.  
h <- function(df)  
{  
 nums <- sapply(df, is.numeric)  
 apply(df[,nums], 2, FUN=mean, na.rm=TRUE)  
}   
  
i <- function(df)  
{  
 mymean <- function(x, ...)  
 {  
 sum(x, ...) / sum(!is.na(x))  
 }  
 nums <- sapply(df, is.numeric)  
 apply(df[,nums], 2, FUN=mymean, na.rm=TRUE)  
}  
  
j <- function(df)  
{  
 for(i in 1:dim(df)[2])  
 {  
 if(!is.numeric(df[,i]))   
 {  
 next  
 }   
 else   
 {  
 c <- 0  
 t <- 0  
 for (x in df[,i])  
 {  
 if(is.na(x))  
 {  
 next  
 }  
 else  
 {  
 c <- c + 1  
 t <- t + x   
 }  
 }  
 print(t / c)  
 }  
 }  
}  
  
#1d.  
microbenchmark(h(d), i(d), j(d))

## Unit: milliseconds  
## expr min lq mean median uq max neval  
## h(d) 4.204373 4.381013 4.655329 4.488958 4.820906 6.189651 100  
## i(d) 4.268799 4.439678 4.726864 4.725332 4.910718 5.975038 100  
## j(d) 52.022592 52.830698 54.744238 53.713471 55.455337 94.839000 100

### 1b.

Using the combination of length and which is much faster (0.03s) than iterating using control flow (188.04s).

### 1c.

I would expect the first formula to be the fastest. This is because it avoids explicit iteration and also uses a built-in function which, where possible, is implemented in a lower-level language. I would expect the second formula to be close but slightly slower. Finally, I would expect the last formula to be quite slow because we are explicitly iterating over the entire data frame.

### 1d.

The results confirm my hypothesis. However, it is suprising how close formula 1 is to formula 2. Comparing the medians, formula 2 only took 1.5% longer than formula 1. I was also surprised how long formula 3 took. Again comparing the medians, formula 3 took 1054% longer than formula 1!