R Manipulations

Step 1) Get both CSV files in R as dataframes. (CSV files must be present in R working directory)

```
counties <- read.csv('SAPS2016_CTY31.csv',header=TRUE,stringsAsFactors=FALSE)

eds <- read.csv('SAPS2016_ED3409.csv',header=TRUE,stringsAsFactors=FALSE)
```

Step 2) Keep required columns in dataframes.

```
counties <- counties[,c('GUID','GEOGDESC','T1_1AGETT','T15_2_Y','T15_2_N','T15_2_NS','T15_3_B',
'T15_3_OTH','T15_3_N','T15_3_NS','T15_3_T')]

eds <- eds[,c('GUID','GEOGDESC','T1_1AGETT','T15_2_Y','T15_2_N','T15_2_NS','T15_3_B','T15_3_OTH'
,'T15_3_N','T15_3_NS','T15_3_T')]
```

Step 3) Give meaningful names to columns starting with ‘T’.

```
colnames(counties) <- c('GUID','GEOGDESC', 'Population',  'PC_Yes',  'PC_No', 'PC_NS', 'Int_Broa
dband', 'Int_Other', 'Int_No', 'Int_NS', 'Households')

colnames(eds) <- c('GUID','GEOGDESC', 'Population',  'PC_Yes',  'PC_No', 'PC_NS', 'Int_Broadban
d', 'Int_Other', 'Int_No', 'Int_NS', 'Households')
```

Step 4) Removing thousands’ separator and quotes from all numeric columns in both dataframes so that we can use them in calculations.

```
counties$Population <- as.numeric(gsub(',', '', counties$Population))
counties$PC_Yes <- as.numeric(gsub(',', '', counties$PC_Yes))
counties$PC_No <- as.numeric(gsub(',', '', counties$PC_No))
counties$PC_NS <- as.numeric(gsub(',', '', counties$PC_NS))
counties$Int_Broadband <- as.numeric(gsub(',', '', counties$Int_Broadband))
counties$Int_Other <- as.numeric(gsub(',', '', counties$Int_Other))
counties$Int_No <- as.numeric(gsub(',', '', counties$Int_No))
counties$Int_NS <- as.numeric(gsub(',', '', counties$Int_NS))
counties$Households  <- as.numeric(gsub(',', '', counties$Households))

eds$Population <- as.numeric(gsub(',', '', eds$Population))
eds$PC_Yes <- as.numeric(gsub(',', '', eds$PC_Yes))
eds$PC_No <- as.numeric(gsub(',', '', eds$PC_No))
eds$PC_NS <- as.numeric(gsub(',', '', eds$PC_NS))
eds$Int_Broadband <- as.numeric(gsub(',', '', eds$Int_Broadband))
eds$Int_Other <- as.numeric(gsub(',', '', eds$Int_Other))
eds$Int_No <- as.numeric(gsub(',', '', eds$Int_No))
eds$Int_NS <- as.numeric(gsub(',', '', eds$Int_NS))
eds$Households  <- as.numeric(gsub(',', '', eds$Households))
```

Step 5) For analysis purpose we will add 2 new columns ‘Prop_PC’ & ‘Prop_Broadband’ to both dataframes to indicate proportion of households with PC and Broadband Internet

connection. We will use mutate function from dplyr package for this purpose.

```
suppressMessages(library(dplyr))

counties <- mutate(counties, Prop_PC = (PC_Yes/Households)*100)
counties <- mutate(counties, Prop_Broadband = (Int_Broadband /Households)*100)

eds <- mutate(eds, Prop_PC = (PC_Yes/Households)*100)
eds <- mutate(eds, Prop_Broadband = (Int_Broadband /Households)*100)
```

Step 6) Finally, let us export these county and ed dataframes as csv files using write.csv command as below.

```
write.csv(counties,file = "counties.csv",row.names=FALSE)

write.csv(eds,file = "eds.csv",row.names=FALSE)
```