Appendix 1 Reading Data into R

Main.R###

```
setwd("~/Dropbox/Statistics/STA242/Assignment1")
source("readFile.R")
source("Clean Data.R")
library(stringr)

setwd("~/Dropbox/Statistics/STA242/Assignment1/data")
filenames = list.files()

results=list()
results=sapply(filenames,readFile)

Filenum=length(filenames)

#Get HOMETOWN from HOMETOWNETIM

c=unlist(sapply(c(1:Filenum),HOMETOWNNETTIMfile))
for (i in c)
{
  results[[i]]$HOMETOWN=HOMETWONTIME(i)
}

#Get DIV/TOT column
c=unlist(sapply(c(1:Filenum),NonDIVTOTfile))
for (i in c)
{
  results[[i]]$`DIV/TOT`=NA
}

#Get GUIDLINE from NETTIME

for (i in c(11,12,23,24))
{
  results[[i]]$GUIDLINE=SpecGuidline(i)
}

c=unlist(sapply(c(1:Filenum),NonGUIDLINEfile))
for (i in c)
{
  results[[i]]$GUIDLINE=NA
}

#Special Case for data class
results[[11]]$PLACE=as.numeric(gsub(intToUtf8(0xA0),"",results[[11]]$PLACE))
results[[16]]$AG=as.numeric(gsub("XX","",results[[16]]$AG))

#Get Useful Data
UsefulData=lapply(results,DataExtract)
UsefulData=do.call(rbind, UsefulData)

write.table(UsefulData,file="~/Dropbox/Statistics/STA242/Assignment1/UsefulData")
```

readFile.R###

```
FindHeader=
function(x)
{
  txt=readLines(file(x))
  k=match("=",substr(txt,1,1))
  txt[k]=gsub(' ','x',txt[k])
  txt[k]=gsub("(\\w)\\b","\\U\\1",txt[k], perl=TRUE)
  txt[k]=gsub ('x','=',txt[k])
  txt[k]=gsub ('X',' ',txt[k])
  txt[k]=strsplit(txt[k], " ")
  w=nchar(txt[k][[1]])+1
  if(is.na(k)){Header=character()} else
```

```
  {Header=read.fwf(x,widths=w,skip = k-2,n=1,comment.char = '',stringsAsFactors=FALSE)}
  Header=gsub(" ","",Header)
  Header=gsub(intToUtf8(0xA0),"",Header)
  Header=toupper(Header)
  Header=gsub("GUNTIM","TIME",Header)
  Header=gsub("GUN","TIME",Header)
  list(txt=txt,k=k,w=w,Header=Header)
}

FindBody=
function(t)
{
  pattern="([0-9]+):([0-9]+)"
  rows=grep(pattern,t)
  Body=t[rows]
}


readFile=
function(x)
  {
    pattern="([a-zA-Z]+)([0-9]+)+([a-zA-Z]+)+([[:punct:]])+([0-9]+)"
    GENDER=gsub(pattern,"\\1",x)
    YEAR=as.numeric(gsub(pattern,"\\5",x))

    txt=FindHeader(x)$txt

    k=FindHeader(x)$k
    w=FindHeader(x)$w
    Header=unlist(FindHeader(x)$Header)
    if(is.na(k)){
      if (GENDER=="men"){
        y=gsub(pattern,"women\\2\\3\\4\\5",x)
        w=FindHeader(y)$w
        Header=unlist(FindHeader(y)$Header)
                    } else
                     {
                     y=gsub(pattern,"men\\2\\3\\4\\5",x)
                     w=FindHeader(y)$w
                     Header=unlist(FindHeader(y)$Header)
                      }
                    }
    Body=textConnection (unlist(FindBody(txt)))
    data=read.fwf(Body,widths=w,comment.char = '',stringsAsFactors=FALSE)
    close(Body)

    names(data)=Header
    data$GENDER=GENDER
    data$YEAR=YEAR
    data
}
```

###CleanData.R###

```
#man_2006/woman_2006 get HOMETOWN from HOMETOWNTIM

HOMETOWNNETTIMfile=
  function(i)
  {
    if(any(grepl("HOMETOWNNETTIM",names(results[[i]]))))
    {True=i}
  }

HOMETWONTIME=
function(i)
{
  HOMETOWNNETTIM=results[[i]]$HOMETOWNNETTIM
  pattern = "([a-zA-Z]+?)\\s+([0-9]+)+(:)+([0-9]+)+(:)+([0-9]+)"
  HOMETOWN= gsub(pattern, "\\1", HOMETOWNNETTIM)
  pattern = "([a-zA-Z]+?)\\s+([0-9]+)+(:)+([0-9]+)"
```

```
  HOMETOWN= gsub(pattern, "\\1", HOMETOWN)
  pattern = "([0-9]+)+(:)+([0-9]+)+(:)+([0-9]+)"
  HOMETOWN= gsub(pattern, "", HOMETOWN)
  HOMETOWN=gsub("^ ", "", HOMETOWN)
  HOMETOWN=gsub(" $", "", HOMETOWN)
  HOMETOWN
}

#get GUIDLINE from NETTIME
SpecGuidline=
  function(i)
  {
    pattern="\\s+([0-9]+[0-9]+[[:punct:]]+[0-9]+)"
    T=gsub(pattern,"0:\\1",results[[i]]$NETTIM)
    pattern="([0-9]+[[:punct:]]+[0-9]+[[:punct:]]+[0-9]+)+(.)"
    GUIDLINE=gsub(pattern,"\\2",T)
    GUIDLINE
  }

# Identify the data without DIV/TOT column

NonGUIDLINEfile=
  function(i)
  {
    if(!any(grepl("GUIDLINE",names(results[[i]]))))
    {True=i}
  }

# Identify the data without DIV/TOT column

NonDIVTOTfile=
  function(i)
  {
    if(!any(grepl("DIV/TOT",names(results[[i]]))))
    {True=i}
  }

#Get useful data for analysis
DataExtract=
  function(r)
  {
    r[c("YEAR","NAME","GENDER","AG","HOMETOWN","DIV/TOT","PLACE","TIME","GUIDLINE")]
  }

setwd("~/Dropbox/Statistics/STA242/Assignment1")
UsefulData=read.table("UsefulData",stringsAsFactors = FALSE)

#Get DIV&TOT
DIV.TOT=UsefulData$`DIV.TOT`[!is.na(UsefulData$`DIV.TOT`)]
DIV.TOT=gsub(intToUtf8(0xA0),"",DIV.TOT)
DIV.TOT=strsplit(DIV.TOT, "/")
UsefulData$DIV[!is.na(UsefulData$`DIV.TOT`)]=as.numeric(sapply(DIV.TOT, function(x){x[1]}))
UsefulData$TOT[!is.na(UsefulData$`DIV.TOT`)]=as.numeric(sapply(DIV.TOT, function(x){x[2]}))
remove(DIV.TOT)


#Get BirthYear
UsefulData$BirthYear=UsefulData$YEAR-UsefulData$AG

#Change Time Format

T=gsub("$"," ",UsefulData$TIME)
T2=gsub("$"," ",UsefulData$TIME[is.na(UsefulData$GUIDLINE)])
pattern="\\s+([0-9]+[0-9]+[[:punct:]]+[0-9]+)"
T=gsub(pattern,"0:\\1",T)
T2=gsub(pattern,"0:\\1",T2)
pattern="([0-9]+[[:punct:]]+[0-9]+[[:punct:]]+[0-9]+)+(.)"
TIME=as.difftime(gsub(pattern,"\\1",T), units = "mins")
UsefulData$`Time(mins)`=as.numeric(TIME)
UsefulData$GUIDLINE[is.na(UsefulData$GUIDLINE)]=gsub(pattern,"\\2",T2)
```

```
UsefulData=UsefulData[!is.na(UsefulData$`Time(mins)`),]
remove(T)
remove(T2)
remove(TIME)

# Get Hometown State/Country
library(stringr)
CS=str_trim(UsefulData$HOMETOWN, side = "both")
pattern="([[:blank:]])+([A-Z])+([A-Z])"
STATE=str_trim(str_extract(CS,pattern),side = "both")
state.abb[51]="DC"
STATE_CODE=match(STATE,state.abb)
STATE= state.abb[STATE_CODE]
UsefulData$STATE=STATE
UsefulData$STATE_CODE=STATE_CODE
UsefulData$CITY=tolower(gsub(pattern,"",str_trim(UsefulData$HOMETOWN, side = "both")))
remove(CS)
remove(STATE)
remove(STATE_CODE)
remove(state.abb)

#Set Id Frequency Experience
UsefulData$NAME=tolower(str_trim(UsefulData$NAME, side = "both"))
UsefulData$Identify=paste(UsefulData$NAME,UsefulData$BirthYear,UsefulData$HOMETOWN,UsefulData$GENDER)
UsefulData$Id=unclass(factor(UsefulData$Identify))
frequency=data.frame(table(UsefulData$Id))
UsefulData$Frequency=frequency$Freq[UsefulData$Id]

frequency=UsefulData[c("Identify","Frequency")]
frequency=frequency[order(frequency$Identify),]
Experience=
  function(i)
  {
    experience=frequency[frequency$Frequency==i,]
    rep(c(0:(i-1)),times=dim(experience)[1]/i)
  }
UsefulData=UsefulData[order(UsefulData$Identify),]
UsefulData=UsefulData[order(UsefulData$Frequency),]
UsefulData$Experience=unlist(lapply(c(1:max(UsefulData$Frequency)),Experience))

remove(frequency)

#GENDER
UsefulData$GENDER[UsefulData$GENDER == "men"] = 0
UsefulData$GENDER[UsefulData$GENDER == "women"] = 1
UsefulData$GENDER=as.numeric(UsefulData$GENDER)

#Generate DataforAnalysis

DataforAnalysis=UsefulData[c("YEAR","Identify","AG","BirthYear","GENDER","Frequency","Experience","STATE","STATE_CODE","CIT
Y","TOT","GUIDLINE","Time(mins)","PLACE")]
write.table(DataforAnalysis,"~/Dropbox/Statistics/STA242/Assignment1/DataforAnalysis")
```