Lab 8

Ankita Shankhdhar

5 Novembor 2015

Question 1

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.1.3
```

```
setwd("/Users/ankitashankhdhar/Documents/Grad 2nd yr/Comp Stats/Lab 8 ")
data <- read.csv("bnames.csv")
name.perc <- ddply(data,.(name),summarize,sum.perc=sum(percent))
subset.boy <- subset(data, sex=="boy")
subset.girl <-subset(data, sex=="girl")
name_boy.perc <- ddply(subset.boy,.(name),summarize,sum.perc=sum(percent))
name_girl.perc <- ddply(subset.girl,.(name),summarize,sum.perc=sum(percent))
name_boy.perc <- name_boy.perc[order(- name_boy.perc$sum.perc),]
name_boy.perc[1:10,]
```

```
##         name sum.perc
## 1843    John 5.299585
## 1707   James 4.574991
## 3358 William 4.409453
## 2825  Robert 3.821662
## 590  Charles 2.518147
## 2415 Michael 2.366102
## 1872  Joseph 2.292487
## 821    David 2.159018
## 1365  George 2.096747
## 3133  Thomas 1.901267
```

```
name_girl.perc <- name_girl.perc[order(- name_girl.perc$sum.perc),]
name_girl.perc[1:10,]
```

```
##           name sum.perc
## 2730      Mary 4.511860
## 1196 Elizabeth 1.392100
## 2642  Margaret 1.360965
## 1582     Helen 1.234222
## 246       Anna 1.195867
## 1098   Dorothy 1.065111
## 398    Barbara 1.001579
## 3128  Patricia 0.999798
## 3338      Ruth 0.942272
## 2392     Linda 0.837364
```

The resulting data frame for boys is 3437 by 2 and the resulting data fram for girls is 4018 by 2.

Question 2

```
#function takes in data and then orders by percent
popular.names<-function(data){
  yearnew<- data[order(- data$percent),]
  top<-yearnew$name[1:5]
  #order them by name and set them in a data frame
  val <- data.frame(name.1=top[1],name.2=top[2],name.3=top[3],name.4=top[4],name.5=top[5])
  return(val)
}
#test for one year and one sex
year1880.boy<- subset(data, (year=="1880")& (sex=="boy"))
popular.names(year1880.boy)
```

```
##   name.1  name.2 name.3  name.4 name.5
## 1   John William  James Charles George
```

```
#gets all the popular names every year for both genders
popular.all <- ddply(data,.(year,sex),popular.names)
head(popular.all)
```

```
##   year  sex name.1  name.2 name.3    name.4   name.5
## 1 1880  boy   John William  James   Charles   George
## 2 1880 girl   Mary    Anna   Emma Elizabeth   Minnie
## 3 1881  boy   John William  James    George  Charles
## 4 1881 girl   Mary    Anna   Emma Elizabeth Margaret
## 5 1882  boy   John William  James    George  Charles
## 6 1882 girl   Mary    Anna   Emma Elizabeth   Minnie
```

```
tail(popular.all)
```

```
##     year  sex name.1   name.2  name.3   name.4  name.5
## 253 2006  boy  Jacob  Michael  Joshua    Ethan Matthew
## 254 2006 girl  Emily     Emma Madison Isabella     Ava
## 255 2007  boy  Jacob  Michael   Ethan   Joshua  Daniel
## 256 2007 girl  Emily Isabella    Emma      Ava Madison
## 257 2008  boy  Jacob  Michael   Ethan   Joshua  Daniel
## 258 2008 girl   Emma Isabella   Emily  Madison     Ava
```

Question 3

```
# reads in the csv file and stored in the data
data <- read.csv("http://people.math.umass.edu/~jstauden/bnames.csv")

# function finds a linear model fit for percent and the year of the data
lm.fit <- function(temp){
    fit <- lm(percent~year,data=temp)
    # returns the fitted coefficients
    return(data.frame(int=fit$coef[1],slope=fit$coef[2],
    n=dim(temp)[1]))
}
# finds the fit for each distinct name and sex
inc.dec <- ddply(data,.(name,sex),lm.fit)
# only takes the sample that was larger than 100
# Example. There were 129 Aarons as boys
inc.dec <- subset(inc.dec,n>100)
# if the slope is within the outer boundaries of the quantiles then keep those names
inc.dec <- subset(inc.dec,(slope>quantile(slope,p=0.99,na.rm=T))|
(slope<quantile(slope,p=0.01,na.rm=T)))
```

Question 4

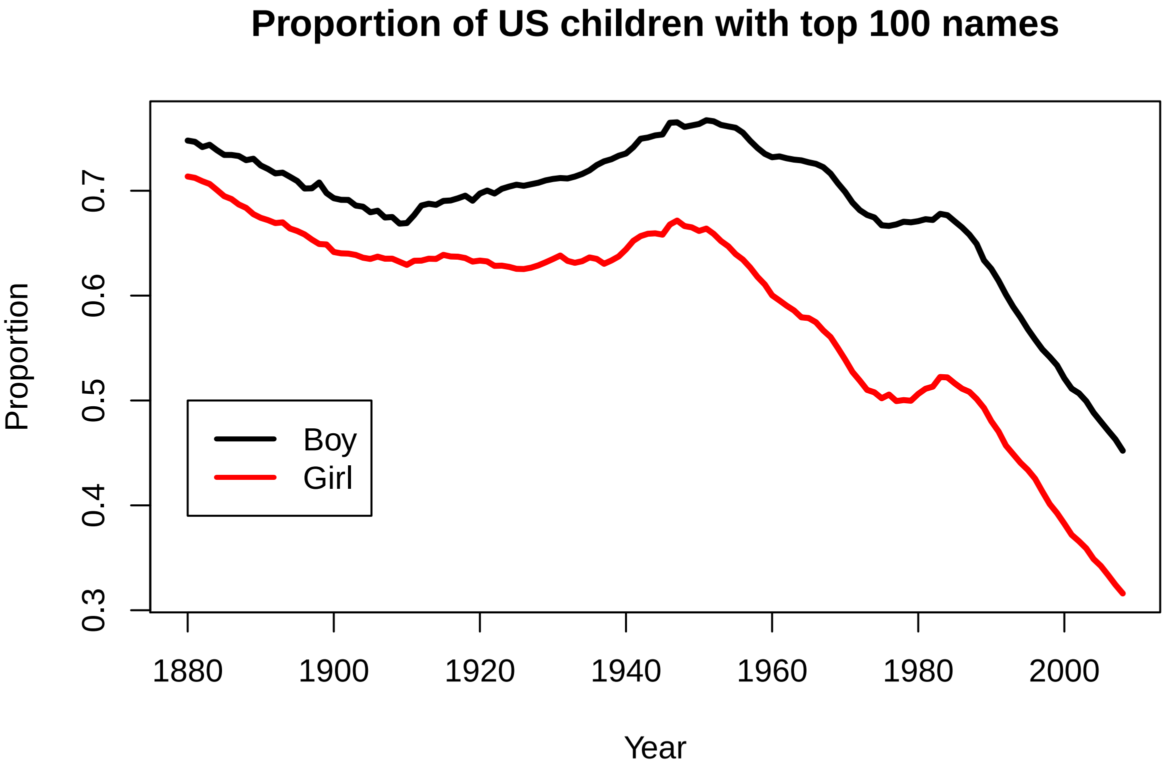
```
smaller.data <- merge(data,inc.dec)

plot.all <- function(ndata,xlims,ylims){
    fit <-lm(percent~year,data=ndata)
    plot(ndata$year,ndata$percent,xlab="year",ylab="Percent",main = paste(ndata$name[1]), xlim = xlims,
    abline(fit)
}
# have a matrix of 5 rows and 4 columns with all of them together
pdf("plots.pdf", onefile = T)
par(mfrow = c(4, 4), cex = 0.5)
d_ply(smaller.data, .(name), plot.all, xlim = range(smaller.data$year), ylim = range(smaller.data$percen

# make the plots so we can look at the pdf
dev.off()
```

```
## pdf
##   2
```

Question 5

```
# function that takes in data and orders the data by percent
# keeps the top 100 of the sorted data and sums ths percent
# output is the sum of percent
top100.names<-function(data){
  year.gender<- data[order(- data$percent),]
  year.gender<-year.gender[1:100,]
  sum.perc<-sum(year.gender$percent)
  return(sum.perc)
}
# perform the same on the data that is separated by year and sex
top.names<-ddply(data,.(year,sex),top100.names)
# get the boy one
top.names.boy<-subset(top.names,(sex=="boy"))
# get the girl one
```

```
top.names.girl<-subset(top.names,(sex=="girl"))
# plot them
plot(top.names$year,top.names$V1,
     xlab="Year",ylab="Proportion",type="n",
     main="Proportion of US children with top 100 names")
lines(top.names.boy$year,top.names.boy$V1,lwd=3)
lines(top.names.girl$year,top.names.girl$V1,col="red",lwd=3)
# add the legend
legend(1880,0.5,c("Boy","Girl"),lty=c(1,1),lwd=c(2.5,2.5),col=c("black","red")) # puts text in the lege
```


Proportion of US children with top 100 names
Proportion
Year
Boy
Girl