Lab 8: PLYR!

A dataset bnames.csv is at http://people.math.umass.edu/~jstauden/bnames.csv. It contains the 1000 most popular male and female baby names in the US, from 1880 to 2008. There are 258,000 records (1000 * 2 * 129) but only four variables: year, name, sex, and percent. (Data from Prof. Wickham seminar / workshop.)

1. Build a data frame that has two columns: name, and the sum of the percentages for each name. Make one data frame for boys and one for girls. Sort the data frame so that the most popular names are at the top. How big is the resulting data frame? Only show the first 10 rows.

2. Find the 5 most popular names by gender for each year. Your result should be a data.frame with two rows for each year. The columns should be year,sex,1st.name,...,5th.name. (Please see below for an example of the first 4 lines of the desired output.) Suggested way to do this:

a. Use subset to create a temporary dataset with one year and one sex.

b. Write a function that acts on that temporary dataset and returns a data frame with 5 columns (the top 5 names in that year) and one row. (The rank() function might be useful.)

c. Use ddply() to apply the function to the whole dataset.

(Only show the 1st 5 and last 5 rows of results.)

3. Please add comments to the code below to describe what it does.

```
data <- read.csv("http://people.math.umass.edu/~jstauden/bnames.csv")

lm.fit <- function(temp)
{
  fit <- lm(percent~year,data=temp)
  return(data.frame(int=fit$coef[1],slope=fit$coef[2],
    n=dim(temp)[1]))
}

inc.dec <- ddply(data,.(name,sex),lm.fit)
inc.dec <- subset(inc.dec,n>100)
inc.dec <- subset(inc.dec,(slope>quantile(slope,p=0.99,na.rm=T))|
          (slope<quantile(slope,p=0.01,na.rm=T)))
```

4. The data.frame inc.dec above has 16 rows. For each of those names, make a scatterplot with year on the x-axis and percent on the y-axis. Label the plot with the name, and use abline() to add the least squares regression line. Keep the ranges for the x and y axes the same for each plot. Put the name on top of each plot. Put the plots in a 4 x 4 grid. (hint: use smaller.data <- merge(data,inc.dec), write a function for plotting, and use d_ply().)

5. Create a plot that shows (by year and gender) the proportion of US children who have a name in the top 100. Your plot should have proportion on the y-axis, year on the x-axis, and two lines, one for each gender. Please include a legend too.