Lab 7: Split-Apply-Combine

36-350, Statistical Computing

Friday, 28 October 2011

Agenda: Use `plyr`.

Instructions: Save all your answers in a single plain text file (Word files will not be graded), and upload it to Blackboard, using the page for this assignment. (There is no general digital dropbox any more.) When a question asks you to do something, give the command you use to do it. For questions which ask you to explain, write a short explanation in coherent, complete sentences. (You will be graded on your written explanation, not what you might say to the TA.)

Preamble Enter the following commands in R.

```
library(plyr)
library(ggplot2)
cars <- read.csv('http://www.stat.cmu.edu/~cshalizi/statcomp/labs/07/epafuel2010.csv')
```

This will load the required packages and a data frame named `cars`.

Data Each row of `cars` gives information provided by the EPA on 2010 models of a car. The columns in `cars` are several variables including fuel efficiency (`mpg.city`, `mpg.hwy`, ...), engine size (`displ`), number of cylinders (`cylinders`), country of origin (`country`), etc ... Use the functions `names` and `head` to get a feel for the organization of the data.

1. Make a scatter plot of combined MPG (`mpg.combined`) and engine displacement (`displ`) using the command:

   ```
   qplot(x = displ, y = mpg.combined, data = cars)
   ```

   Do the points follow some pattern? (5)

2. Which car has a combined MPG of about 60 and 0 engine displacement? How is zero possible? (5)

3. Read the help for the function `summarize`. Use `ddply` and `summarize` to compute the average combined MPG (`mpg.combined`) for each vehicle class (`vehicle.class`) and to store the results in a column named in a data frame with columns `vehicle.class` and `mpg.combined.avg`. Which vehicle class has the highest average? (20)

4. Extract the subset of `cars` corresponding to compact cars (`vehicle.class` is equal to `Compact Cars`) and store it in a data frame named `cars.compact`. Compute a summary (as in the previous problem) of the average combined MPG for each country (`country`). Which countries have the highest and lowest average combined MPGs? (20)

5. Plot combined MPG and engine displacement separately for each country using the command

   ```
   qplot(x = displ, y = mpg.combined, data = cars, facets = ~ country)
   ```

   Explain what you see. (5)

6. Let's fit a simple linear regression model to explain MPG as a function of engine displacement for each country. Use `ddply` and `lm` to return a data frame with columns `country` and `slope` giving the country, and the estimated slope in the linear model for that country. You can use the `coef` function to extract the coefficient corresponding to slope. Example:

```
m <- lm(y ~ x)
coef(m)[2]
```

or more succinctly,

```
coef(lm(y ~ x))[2]
```

Explain why the slopes are different? (20)

7. Transformations of the data are sometimes useful for fitting linear models. Let's try a log transform of the engine displacement. Read the help for the function `transform` and then use it to create a new data frame containing all of the columns of `cars` with a new column named `log.displ` that is `log(displ)`. Name this data frame `cars2` then repeat the plot from Problem 5 but with `log.displ` instead of `displ`. Explain the difference between the two plots. What happened to the car with 0 engine displacement? (15)

8. Repeat Problem 6, but with `log.displ` instead of `displ`. (10 and bonus karma for removing the outlier first.)

Bonus This command shows you a plot of what you accomplished:

```
qplot(x = log.displ, y = mpg.combined, data = subset(cars2, displ > 0),
      facets = ~ country) + stat_smooth(method = 'lm')
```