Programming Bootcamp 2015: Lab 8

Josh Burdick & Sarah Middleton

Last Updated: June 26, 2015

1 Indexing (3pts)

Given these R definitions:

```
g = c(1,3,7,9,11.4)
h = c(1,3,5)
```

what would the output for these expressions be?

Code	Predicted Output	Actual Output
`g[1]`		
`g[0]`		
`g[h]`		
`g[ h[3] ]`		
`g[ c(2,3) ]`		
`g[ h[ c(2,3) ] ]`		
`g[ g >= 3 ]`		
`g[ g >= 3 & g <= 9 ]`		
`mean(g) < mean(h)`		
`sum( h <= mean(h) )`		

2 Data frame munging (9pts)

For this problem, load the file `gene_expr2.txt` into R as a data frame called `tab`. You will probably want to keep it in the "wide" format, but it's up to you.

(a) **(2 pt)** Which tissue had the highest average expression? What was that average?

(b) **(2 pt)** Which tissue had the highest maximum expression? What was that maximum?

(c) **(2 pt)** Filter `tab` so that it only includes genes which had an average expression (across the tissues) of at least 500. How many genes had at least this expression?

(d) **(3 pt)** The `order` function, given a vector, returns the sorted order of a vector (in terms of indices). For example:

```
> a = c(1,26,4,8,7)
> order(a)
[1] 1 3 5 4 2
> a[ order(a) ]
[1]  1  4  7  8 26
```

So `order(a)` returns a vector that, if used to index into the original vector `a`, would produce a sorted vector. Use this to write code which sorts `tab` by increasing average gene expression, and stores it in a data table.

3 Data frame subsetting (3pts)

For this problem, load the file `gene_expr2.txt` into R as a data frame called `tab`. You can decide if you prefer wide or long format for each question.

(a) **(1 pt)** Create a data frame called `tabSubset1` that contains only data pertaining to gene4, gene8, and gene18

(b) **(1 pt)** Create a data frame called `tabSubset2` that contains only data pertaining to tissue3 and tissue4

(c) **(1 pt)** Create a data frame called `tabSubset3` that contains only data pertaining to gene4/gene8/gene18 and tissue3/tissue4

4 Graphing (6pts)

For this problem, use the data frame subsets you created in the previous problem.

(a) **(1 pt)** Using `tabSubset1`, create a boxplot comparing the ranges of expression of the three genes (i.e. each gene should be represented by a box, which shows its expression across the tissues).

(b) **(1 pt)** Using `tabSubset2`, create a density plot comparing the distribution of expression values of the two tissues (i.e. each tissue should be represented by a separate density curve).

(c) **(1 pt)** Create a heatmap of the expression values in `tabSubset2`. Change the colors to something cool. :)

(d) **(3 pt)** Go to the ggplot2 docs page (`http://docs.ggplot2.org/current/`) and pick a graph type that we didn't go over. See if you can display the data in `tab` in one of these graph formats by following the examples. You are free to take subsets/use summary statistics if you want – try to create a graph that actually shows something meaningful about the data.