Statistics 133 Midterm Exam
March 3, 2010

When I ask for an "R program", I mean one or more R commands. Try your best to make your answers general, i.e. they shouldn't depend on the specific values presented in the examples.

Total: 40 points

1. Recall the `world` data frame that we've examined in class. Here is some information about the data frame:

```
> summary(world)
      country    cont         gdp            income         literacy        military             phys
 Albania  :  1   AF:47   Min.   :  500   Min.   :  569   Min.   :12.80   Min.   :6.500e+06   Min.   :  1.132
 Algeria  :  1   AS:41   1st Qu.: 1800   1st Qu.: 2162   1st Qu.:69.10   1st Qu.:5.650e+07   1st Qu.: 21.866
 Angola   :  1   EU:34   Median : 4900   Median : 5894   Median :88.40   Median :2.343e+08   Median :125.653
 Argentina:  1   NA:15   Mean   : 9054   Mean   :10258   Mean   :80.95   Mean   :5.679e+09   Mean   :154.422
 Armenia  :  1   OC: 4   3rd Qu.:11800   3rd Qu.:14233   3rd Qu.:98.50   3rd Qu.:1.775e+09   3rd Qu.:267.853
 Australia:  1   SA:12   Max.   :55100   Max.   :63609   Max.   :99.90   Max.   :3.707e+11   Max.   :606.496
 (Other)  :147                           NA's   :    1
```

   (a) (2 points) Write an R program that willl calculate the mean of `income` and `military` for each continent.

   Solution:

```
aggregate(world[,c('income','military')],
                  list(world$cont),mean,na.rm=TRUE)
```

   Note that income had a missing value so that `na.rm=TRUE` is required.

   (b) (2 points) What is the difference between the plots produced by these two commands:

```
xyplot(literacy~gdp,groups=cont,data=world)
```

   and

```
xyplot(literacy~gdp|cont,data=world)
```

   Solution: The first command produces a single plot with different colors for each continent. The second command creates a separate panel for each continent.

   (c) (2 points) What option needed to be passed to `read.csv` to insure that countries in North America didn't have missing values for `cont`.

   Solution: `na.strings=''`

(d) (2 points) What option would be passed to `read.csv` to insure that `country` and `cont` were character variables, not factors.

Solution: `stringsAsFactors=FALSE`

(e) (2 points) Write an R program to produce a barplot showing the number of countries in each continent.

Solution:

```
barplot(table(world$cont))
```

2. Consider the following vector of values stored in a variable called `x`:

```
> x
 [1]  7 12  9 15 NA  8 14 NA  2  9 NA  8
```

(a) (2 points) Write an R program to return only the non-missing values into a vector called `y`

Solution: `x[!is.na(x)]`

(b) (2 points) Write an R program to count the number of missing values in `x`.

Solution: `sum(is.na(x))`

(c) (2 points) Write an R program to calculate the median of `x` ignoring the missing values.

Solution: `median(x,na.rm=TRUE)`

(d) (2 points) Write an R program that will replace the missing values with the value in the vector immediately before the missing value.

Solution:

```
wh = which(is.na(x))
x[wh] = x[wh - 1]
```

3. Consider a data frame called `crackers`:

```
> summary(crackers)
           Company                              Product       Grams          Calories
 Nabisco        :36   Barnum's Animal crackers   : 1   Min.   :13.00   Min.   : 50.0
 Keebler        :26   Better Cheddars            : 1   1st Qu.:15.00   1st Qu.: 70.0
 Sunshine       : 8   Better Cheddars Low Sodium : 1   Median :16.00   Median : 80.0
 Pepperidge Farm: 6   Better Cheddars Reduced Fat: 1   Mean   :22.11   Mean   :102.2
 Adrienne       : 3   Big Wheat Thins            : 1   3rd Qu.:30.00   3rd Qu.:140.0
 Dare           : 3   Bretton                    : 1   Max.   :31.00   Max.   :160.0
 (Other)        :10   (Other)                    :86   NA's   :11.00
```

(a) (2 points) Write an R program to plot `Calories` on the y-axis and `Grams` on the x-axis, using a different color for each level of `Company`.

Solution:

```
library(lattice)
xyplot(Calories ~ Grams,groups=Company,data=crackers)
```

or

```
crackers$Company = factor(crackers$Company)
mycolors = topo.colors(length(levels(crackers$Company)))
plot(crackers$Grams,crackers$Calories,
     col=mycolors[crackers$Company])
```

(b) (2 points) Write an R program to show how many observations there are for each `Company`.

Solution: `table(crackers$Company)`

(c) (2 points) Write an R program that will rearrange the rows of the data frame so that they are sorted by the value of `Calories`.

Solution: `crackers[order(crackers$Calories),]`

(d) (2 points) Write an R program that will show the row number of the observation with the maximum value for `Calories`.

Solution: `which.max(crackers$Calories)`
or
`which(crackers$Calories == max(crackers$Calories))`

4. Use regular expressions to answer the following questions:

   (a) (2 points) Consider a vector called `values`:

   ```
   > values
   [1] "$17,244.41" "$25,622.41" "19,588.41"  "$24,441.32"
   ```

   Write an R program to convert these values into proper numbers, and to calculate the sum of those numbers.

   Solution: `sum(as.numeric(gsub('[$,]','',values)))`

   (b) (2 points) Consider a vector called `nms`:

   ```
   > nms
   [1] "...Company.."  "Interest.Rate" "..Industry."   ".Year"
   ```

   Write an R program to eliminate the leading and trailing periods (`.`) from these values, but *not* periods inside the values.

   Solution:

   ```
   nms = gsub('^\\.+','',nms)
   nms = gsub('\\.+$','',nms)
   ```

   or

   ```
   sub('^\\.*(.*)\\.*$','\\1',nms)
   ```

   (c) (2 points) Consider a vector of file names, called `fnames`:

   ```
   > fnames
   [1] "dog.jpg"  "jpeg.doc"  "homework.r" "duck.jpeg" "jpeg.txt" "cat.jpg"
   ```

   Write an R program that will create a vector with all the file names that end in either `jpg` or `jpeg`.

   Solution: `grep('jpe?g$',fnames,value=TRUE)`

   (d) (2 points) Consider the following text, and call to `gregexpr`:

   ```
   > txt = 'name=Fred job=Cashier pay=$12000'
   > matches = gregexpr('(.*)=(.*)',txt)
   ```

   `matches` contains only one match. How would you modify the regular expression passed to `gregexpr` to return three matches?

   Solution: `gregexpr('([^ ]*)=([^ ]*)',txt)`

5. Suppose you find an online article that shows how to create a plot that you would like to use. When you try to follow the instructions, you see the following error:

```
> library(plotrix)
Error in library(plotrix) : there is no package called 'plotrix'
```

(a) (2 points) How would you make the `plotrix` command available on your computer?

> **Solution:** Either use the `Packages` drop-down menu in the console, or call `install.packages`.

(b) (2 points) Write an R command that would open a browser on newsgroup postings concerning the `plotrix` package.

> **Solution:** `RSiteSearch('plotrix')`

(c) (2 points) After seeing the error, one thought might be to use

```
> help(plotrix)
```

Why would the `help` command not be useful in this case?

> **Solution:** The `help` function can only find information about things that are already installed on your computer.