Statistics 133 Midterm Exam

March 2, 2011

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. 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 the positions of the missing values in x.

Solution:

```
which(is.na(x))
```

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

Solution:

```
sum(!is.na(x))
```

(c) (2 points) Write an R program to replace the missing values in x with the mean of the non-missing values in x.

Solution:

```
x[is.na(x)] = mean(x,na.rm=TRUE)
```

(d) (2 points) Write an R *function* that, when passed a vector, will return a vector with the missing values in the vector replaced by the mean of the non-missing values of the vector.

Solution:

```
replacena = function(x){
  x[is.na(x)] = mean(x,na.rm=TRUE)
  x
}
```

2. Consider a data frame called `cars`:

```
> summary(cars)
   Country                         Car           MPG            Weight         Horsepower
 France : 1   AMC Concord D/L       : 1   Min.   :15.50   Min.   :1.915   Min.   : 65.0
 Germany: 5   AMC Spirit            : 1   1st Qu.:18.52   1st Qu.:2.208   1st Qu.: 78.5
 Italy  : 1   Audi 5000             : 1   Median :24.25   Median :2.685   Median :100.0
 Japan  : 7   BMW 320i              : 1   Mean   :24.76   Mean   :2.863   Mean   :101.7
 Sweden : 2   Buick Century Special: 1   3rd Qu.:30.38   3rd Qu.:3.410   3rd Qu.:123.8
 U.S.   :22   Buick Estate Wagon    : 1   Max.   :37.30   Max.   :4.360   Max.   :155.0
              (Other)               :32
```

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

Solution:

```
library(lattice)
xyplot(MPG~Horsepower,group=Country,data=cars)
```

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

Solution:

```
cars[order(cars$Horsepower),]
```

(c) (2 points) Write an R program that will show the row number of the observation with the with the highest ratio of MPG to weight.

Solution:

```
which.max(cars$MPG / cars$weight)
```

3. Consider a vector called `book`, each element of which contains the text of one sentence of a book. For the purposes of this question, consider a word as text separated from other text by one or more blanks.

   (a) (2 points) Write an R program to find the average number of characters in each sentence including the blanks, and another program to find the average number of characters in each sentence not including the blanks.

   Solution:

   ```
   mean(nchar(book))
   mean(nchar(gsub(’ ’,’’,book)))
   ```

   (b) (2 points) Write an R program to find the average number of words in each line of the book.

   Solution:

   ```
   words = strsplit(book,’ +’)
   mean(sapply(words,length))
   ```

   (c) (2 points) Write an R program to find the line in the book with the most characters.

   Solution:

   ```
   book[which.max(nchar(book))]
   ```

4. Consider a data frame called `wine`, which contains information about the chemical composition of different types of wines. Here is some information about the data frame:

```
Type      Alcohol         Malic.Acid        Proline
A:36   Min.   :11.03   Min.   :0.740   Min.   : 278.0
B:46   1st Qu.:12.36   1st Qu.:1.597   1st Qu.: 500.5
C:35   Median :13.05   Median :1.845   Median : 673.5
D:31   Mean   :13.00   Mean   :2.298   Mean   : 746.9
E:30   3rd Qu.:13.68   3rd Qu.:3.030   3rd Qu.: 985.0
       Max.   :14.83   Max.   :5.510   Max.   :1680.0
                       NA’s   :2.000
```

   (a) (2 points) Write an R program that willl calculate the median of `Alcohol` and `Malic.Acid` for each Type of wine.

   Solution:

   ```
   aggregate(wine[,c(’Alcohol’,’Malic.Acid’)],wine[’Type’],
               median,na.rm=TRUE)
   ```

(b) (2 points) Write an R program to count the number of observations with Alcohol greater than 13 and Proline less than 650.

Solution:

```
sum(wine$Alcohol > 13 & wine$Proline < 650)
```

(c) (2 points) If you were reading this data from a comma-separated file, what option would be passed to `read.csv` to insure that `Type` was read as a character variable, not a factor?

Solution:

```
stringsAsFactors=FALSE
```

(d) (2 points) Write an R program to produce a barplot showing the number of wines of each type in the data frame.

Solution:

```
barplot(table(wine$Type))
```

5. Consider the following vector:

```
> text = c(’cat 122’,’dog 213’,’721 chicken’,’fish 42’,’893 duck’)
```

Use regular expressions to answer the following questions:

(a) (2 points) Write an R program to create a vector like `text`, with the number in each element appearing before the animal name.

Solution:

```
sub(’([a-z]+) ([0-9]+)’,’\\2 \\1’,text)
```

(b) (2 points) Write an R program to create a vector containing just the animal names in `text`.

Solution:

```
gsub(’[0-9 ]’,’’,text)
```

(c) (2 points) Write an R program to produce a vector containing the position of the blank in each element of `text`.

Solution:

```
unlist(gregexpr(’ ’,text))
```

(d) (2 points) Write an R program to remove the first three characters in each of the elements of `text`

Solution:

```
sub(’^...’,’’,text)
```

6. Consider a data frame called `stock`. Here are the first few lines of the data frame:

```
> head(stock,n=3)
        Date Price
1 2011-02-25  1.44
2 2011-02-24  1.39
3 2011-02-23  1.44
```

Suppose you tried to plot `Price` versus `Date` and saw the following:

```
> plot(stock$Date,stock$Price)
Error in plot.window(...) : need finite ’xlim’ values
In addition: Warning messages:
1: In xy.coords(x, y, xlabel, ylabel, log) : NAs introduced by coercion
2: In min(x) : no non-missing arguments to min; returning Inf
3: In max(x) : no non-missing arguments to max; returning -Inf
```

(a) (2 points) What would you do to fix the problem, and get a meaningful plot?

Solution:

```
stock$Date = as.Date(stock$Date)
plot(stock$Date,stock$Price)
```

(b) (2 points) What would the class of the `stock$Date` variable be in order to cause the error message regarding “no non-missing arguments” to `min` and `max`.

Solution: It would have to be character, because a factor would produce a plot with lots of little lines (and no error message).