Final Exam
Introduction to Data Science
Fall 2018

This exam has 10 questions. It is 5 pages long and is out of 67.
No electronics permitted. Closed books and notes.

NAME: _____

ID: _____

1) [10] Suppose we have a data.frame with information on 1000 vehicles (types are "compact", "sedan", "4x4", "truck") and the following line of code is run:

```
> sapply(vehicles,class)
Price          Weight         Type
"numeric"      "numeric"      "factor"
```

For each of the following, write one line of code that returns the information specified.

a) number of vehicles corresponding to each type.

b) the average price of the compact vehicles that weigh less than 1500 pounds.

c) the average price of the vehicles of each type (answer should be a vector of length 4).

d) Write R code to produce a plot with Price on the x axis and Weight on the y axis. Give the plot a title, label the axes, and make the different types of vehicles different colors.

e) Give R code that adds to the plot in d) a circle for each Type (4 circles total) located at the average Price and Weight for that vehicle type. The size of the circle should be proportional to the number of vehicles of that type and the transparency should be 0.3. The circles should be the same colors as the points in d).

2) [4] X is a vector of 1000 numbers. Values may appear more than once.

a) Split X randomly into 2 disjoint vectors of equal size, call these vectors x1 and x2.

b) Create a vector x3 that contains 100 random values from X. Then remove these values from X.

3) [4] Write a single line of R code to return a simulation of one roll of a fair six-sided die:

a) using the sample function

b) not using the sample function

4) [9] Indicate which strings contain a match to the pattern:

	"hi mabc"	"abc"	"abcd"	"abccd"	"abcabcdx"	"cab"	"abd"	"cad"
abc								
^abc								
abc.d								
abc+d								
abc?d								
abc$								
abc.*d								
abc?								
a[b?d]								

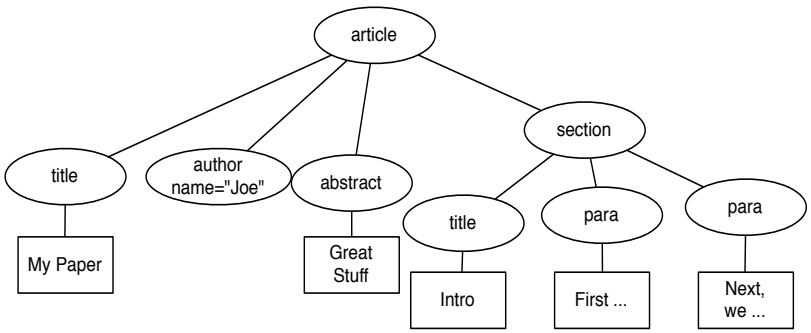
5) [6] Suppose I have an HTML file and I want to use regular expressions to find each node called `div` that has an attribute `id` with the value `USA` and extract the contents of these nodes.

a) Why would this regular expression pattern be a bad way to do this? `<div id="USA">.*</div>`

b) What would a better regular expression pattern be?

c) Write code to extract this using Xpath in R.

6) [4] From the following tree, write the corresponding (well formed) XML:



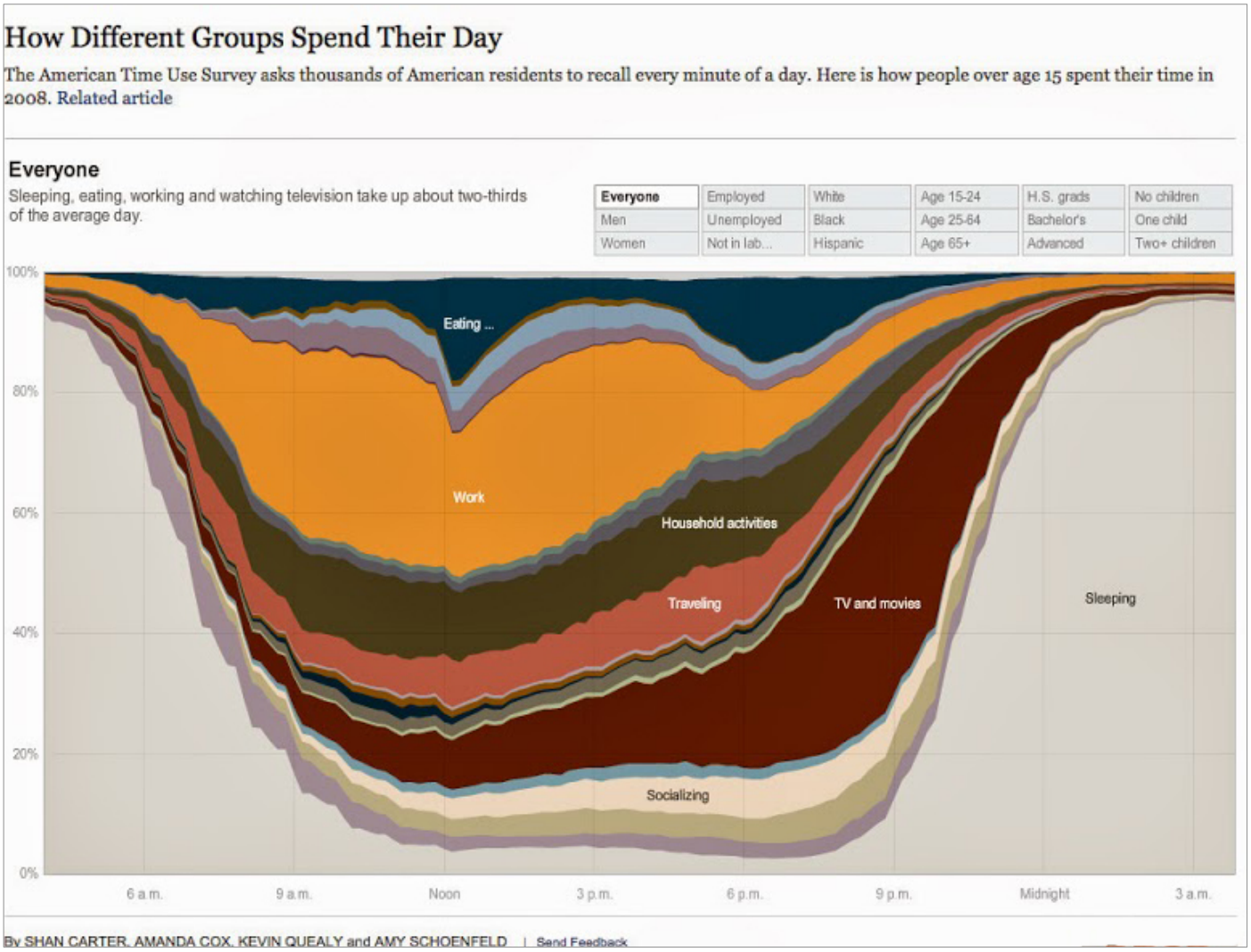
7) [8] Suppose the XML you gave in the previous question is stored in a file called `my.xml`. Give the unix (bash shell) command(s) that resolves each of the following:

a) [2] How many lines are in `my.xml`?

b) [2] Substitutes para with paragraph.

c) [4] Substitutes title, under section, with title2. Make sure title under article is unchanged.

8) [8] The following plot was published in the New York Times as a "Graphic of the Day".



a) [4] Describe (without using code) how you would represent these data in a plot.

b) [4] Give R code that implements your description in part a).

9) [8] Suppose we run the following code:

```
mat <- matrix(rnorm(25), nrow=5, ncol=5)

calculation.dataset <- function(mat){
      result <- list()
      for (j in 1:ncol(mat))
            {
                  result[j] <- sum(mat[,j])/length(mat[,j])
            }
      result <- unlist(result)
      return(result)
      }
```

a) [2] Describe `result`. What type of object is it, what dimensions, and what does it contain?

b) [3] Suppose I've already created the matrix `mat`. Write one line of code using an apply function that will create `result`.

c) [3] Again, suppose we've created `mat`. Write one line of code *without* using either a loop or apply function that will create `result`.

10) [6] Suppose you are carrying out a data science project and you have collected and merged two datasets: the vehicles dataset from question 1 and the California housing dataset discussed in class (assume there's a linking variable, e.g. addresses of vehicle owners are given). You apply the function `calculation.dataset()` from the previous question, and present `result` as your finding.

a) Describe the *Lifecycle of Data Science* for this project as completely as you can. Although it is not necessary, you may wish to refer to the Lifecycles we discussed in class.

b) Why is the Lifecycle of Data Science important?