StatR 501
Homework 2

Instructions: All homework must be typed. For these and subsequent homeworks, report code as appropriate for the question. Export plots as needed from R and incorporate them into your document. Upload the completed homework assignment into the course webpage drop-box. Much of this homework builds on the code in the computer lab. Please follow the same naming convention as described in hw1.

1. (a) Read the file lab2.html (under modules of the course site)

   (b) Read Ch. 5 Statistical Inference in *Probability and Statistics – The Science of Uncertainty*.

   (c) Read Ch. 1 Introduction and Preliminaries, Ch. 2 Simple Manipulations; Numbers and Vectors, Ch. 5 Arrays and Matrices, and Ch. 6 Lists and Data Frames in *An Introduction to R*.

2. More grade manipulation: Suppose that I am teaching a graduate statistics course with only 5 students. Below is a list of their scores in this class.

Exam	Alice	Bruno	Carl	Dolores	Ebenezer
quiz 1	89	95	75	82	90
quiz 2	87	90	60	0	90
midterm	84	92	72	88	96
final	78	76	58	68	80

   (a) Given the following numeric vector, use the matrix function to create the above matrix in R:

```
c(89, 87, 84, 78, 95, 90, 92, 76, 75, 60, 71, 58, 82, 0, 88, 68, 90, 90, 96, 80)
```

   (b) Given the following numeric vector, use the matrix function to create the above matrix in R. Do not use the transpose function. Instead, look for an argument in the matrix help file that will be useful here.

```
c(89, 95, 75, 82, 90, 87, 90, 60, 0, 90, 84, 92, 71, 88, 96, 78, 76, 58, 68, 80)
```

   (c) Given the following numeric vectors, use a function to generate the above matrix in R:

```
c(89, 95, 75, 82, 90) c(87, 90,
60, 0, 90) c(84, 92, 71, 88, 96)
c(78, 76, 58, 68, 80)
```

(d) Take one of the matrices you generated above and assign it to the object ClassScores. (note that all of the matrices you generate above should be identical, so it doesn't matter which one you use). Give the ClassScores matrix the same row and column names printed above in the table.

(e) Alice's grades are given by ClassScores[,"Alice"]. How would you extract just Alice's quiz scores (quiz 1 and quiz 2)?

(f) Create a new length-5 vector that containes the average quiz scores for each student (average for quiz 1 and quiz 2), call it QuizAverage. Do this by taking the mean of the vector you created above (ie, for Alice) and repeat this manually four more times (one for each remaining student). Wouldn't it be nice if there was a way to automate this?

(g) Look at the help file for the colMeans function. How could you use this function on a subset of ClassScores to get the same result you calculated manually above?

3. **Analysis of table tennis appeal:** Download and load into R the StudentSurvey.csv data from the website. Among the columns in the undergraduate student data are two categorical columns: Sex and Pingpong. The latter is a response to the question: *"How much do you enjoy playing table tennis, on a scale of 1 (not at all) to 5 (it is basically an obsession)?"*

(a) Produce a 2×5 table (ie 2 rows and 5 columns), tabulating the counts of male and female students in each category of pingpong enjoyment.

(b) Produce a barplot from that table object. Specify the arguments beside=TRUE, and col=rep(c(3,5),times=5). What do these arguments do? Experiment and/or read the help file for barplot to find out.

(c) Add a legend using the following code:

```
legend("topright", fill = c(3, 5), legend = c("Female", "Male"))
```

4. **Analysis of global patterns:** Download the CountryData.csv file and load it into R. This data file contains information on population (×1000), area (1000 km^2), literacy rate, per capita GDP ($1000), birth rate (number of births per 1000 people per year), percentage of land covered by water, and a classification by continental region.[1]

(a) Read the csv file into R, creating an object called CountryData.(Note: There is no need to present anything for this problem).

(b) The following line of code uses order() to extract a vector of the ten poorest countries:

[1] The source of these data are Wikipedia, e.g.: List of countries by birth rate, List of countries by area, List of countries by GDP. See sources within these articles for more details.

```
Poorest <- CountryData$Country[order(CountryData$GDP)][1:10]
```

Using this code as a template, separately create a vector of 10 countries for each of the following: lowest and highest GDP per capita, the highest and lowest birth rates, and the lowest literacy. Comment on any patterns that you identify in these columns.

(c) add a new column to the CountryData data.frame called density defined as area divided population. You can assign a new column that's a ratio of existing columns using the following syntax:

```
x$newcol <- x$a/x$b
```

(d) Identify the 10 countries with the highest and lowest densities, respectively, and present two data.frames that include their population, area and percentage of water coverage. (Hint: You'll definitely want to use the order function here.)

(e) Using the lab as a model, create an overlapping *frequency* histogram of birth rates in Europe, Asia, and Africa in three different, transparent colors. Add a legend to the plot identifying the continents. Make sure that the axes are appropriately labeled and the plot has a meaningful title. Experiment with the bin widths to find one that you feel best illustrates the patterns.

(f) Create a *density* histogram of the same data.

(g) Summarize the patterns in these distribution, commenting on the center, the spread, and the modality (i.e. number of humps).