The University of British Columbia

Irving K. Barber Faculty of Science

DATA 101

Assignment 4

Please submit your assignment as an R script file named with your last name, student number, assignment number and with the suffix R. For example, if Joe Smith, student number 87654321 hands in Assignment 4, he would name the file `Smith87654321A2.R`.

Within your answer file, include answers with your R code preceded by the `#` sign. For example, to answer the 5th question on an assignment which is "Perform the calculation $2 + 2$", you would type

```
# Question 5
2 + 2 #coding
# 4 (your answer here )
```

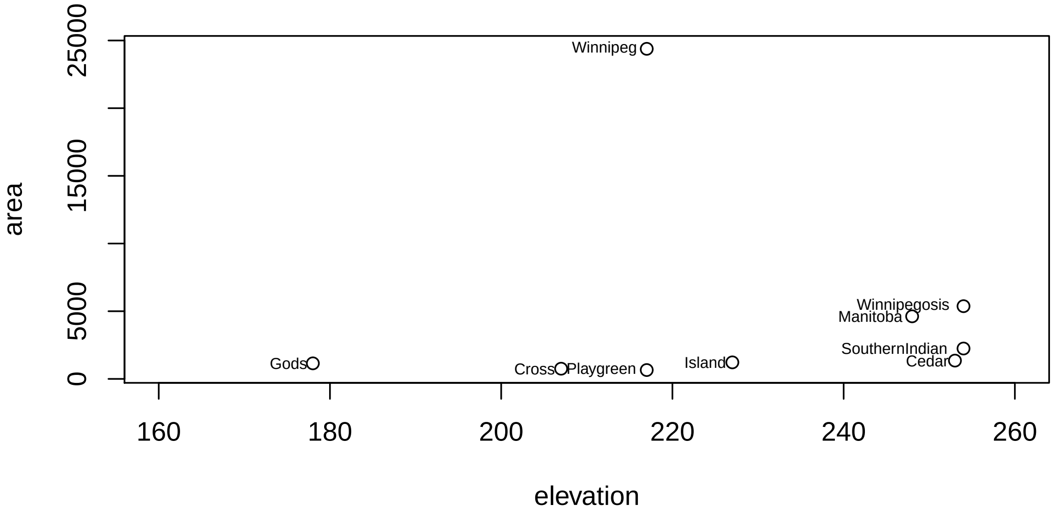
Due Date: November 6, 2020

In each question below, type the required lines of R code, together with the answer to the question.

1. If you have not already done so, install the *DAAG* package, either using the menu system in RStudio or by typing `install.packages(DAAG)` at the command prompt in R. The data frame `Manitoba.lakes` concerns the elevation and area of the largest lakes in Manitoba.

   (a) Obtain a scatter plot of area versus elevation for these lakes. Comment on any noteworthy observations. (2 points)

   (b) Use the `xlim` argument and adjust the character size when using the `text` function to add the row names to obtain a graph that appears as below: (5 points)



   You will also need to use the `adj` argument in the call to the `text` function.

   (c) Using the `lm()` function, fit the least-squares line relating area to elevation and overlay this line on your plot using the `abline()` function. (2 points)

2. Plot height versus age for the pine tree growth data in `Loblolly`, and overlay the best-fit line obtained from the `lm()` function. (5 points)

3. Do you think that the straight line is the best way to represent these data? Explain briefly. (2 points)

4. The `Loblolly` pine data actually contains age and height measurements for a number of different trees, grown from different seed sources. The code below attempts to fit straight lines to the height and age data for each different tree (represented by the factor `Seed`). The slopes and intercepts are obtained for each line.

   ```
   seed <- unique(Loblolly$Seed)
   n <- length(seed)
   slopes <- numeric(n)
   intercepts <- numeric(n)
   for (i in 1:n) {
     lob.lm <- lm(height ~ age, data = subset(Loblolly, Seed==seed[i]))
     slopes[i] <- coef(lob.lm)[2]
     intercepts[i] <- coef(lob.lm)[1]
   }
   ```

   Construct another scatter plot of height versus age for all of the trees, and use a `for` loop to overlay all `n` of the lines corresponding to the different slopes and intercepts obtained with the above code. (10 points)

5. Refer to the previous exercise. Obtain side-by-side histograms(in a 1 by 2 layout) of the slopes and intercepts of the lines that were obtained with the code in the previous question. (4 points)

6. Refer to the previous two exercises. Obtain a scatter plot of the slopes versus the intercepts, overlaying the best-fit line.

   If you had to describe a "typical" line which relates height to age for these kinds of trees, what would you say is a typical slope and what would you say is a typical intercept? (4 points)