Name:\_Rahul Avadhoot\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Last 4 Digits of Student ID Number:\_1787\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Introduction to R for Data Analysis in the Health Sciences**

**BIOST 509**

**In-Class Exercise 2**

**Due 6:00 PM on 10/4/2017**

Instructions

Enter your numerical and/or written answers to the questions using this file or create your own doc or pdf file with your answers. Submit your answers in a Word (.doc or .docx) or pdf file to the course canvas page by the end of the lab session today. Your work will be graded as 1 (complete) if at least 70% of the questions are answered correctly (and/or an obvious “good faith” effort), or 0 (incomplete) otherwise. For this in-class exercise, also include the R commands that you used to obtain your answers, when applicable.

1. Download and read in the salary.txt dataset from the “files” folder within the course canvas page, using a function from the command line; call this data object salary. Also using a function from the command line, read in the salary.txt dataset which is available on the web at: <http://faculty.washington.edu/tathornt/Biost509/DataSets/salary.txt>; call this data object othersalary.

Verify that salary and othersalary are the same. Hint: First use the help page to learn about the ***identical()*** function (e.g., use the “?identical” command) and then use the function.

R Commands:

salary <- read.table("C:/Users/rahul/Downloads/salary.txt", header=TRUE)

othersalary <- read.table("http://faculty.washington.edu/tathornt/Biost509/DataSets/salary.txt", header=TRUE)

identical(salary,othersalary)

Answer the following questions using the salary dataset.

1. In class today, we saw that 4 entries had missing salary variables. Using the is.na() function and other commands, determine which number rows have these missing values.

Rows 4, 44, 444, and 4444 have NA values in salary column

R Command:

which(is.na(salary$salary))

1. Create a dataset with the 4 entries (rows) with missing salary data excluded, and rename the data object salary.

R Command:

salary <- salary[is.na(salary$salary),]

updated\_data <- othersalary[is.na(othersalary$salary) == FALSE,]

For the remaining questions, use the dataset *without* the 4 entries (rows) containing NAs for the salary variable (19788 rows, 1596 unique ids).

1. What is the mean salary of the entries? What proportion of the entries have a greater “salary” than the mean “salary”?

Mean salary = 4721.712

Proportion having greater than mean salary = 0.4252577

R Command:

mean(updated\_data$salary)

sum(updated\_data$salary>mean(updated\_data$salary))/length(updated\_data$salary)

1. What is the earliest “startyr”? What is the “id” of the individual with the earliest start year?

Earliest startyr = 48

id = 1535

R Commands:

min(updated\_data$startyr)

updated\_data[updated\_data$startyr == min(updated\_data$startyr),]

1. Which “rank” has the largest average “salary” (averaged over all entries)? Which “rank” has the lowest average “salary”? What is the difference between of the ranks with the highest and the lowest average salaries?

‘Full’ rank has the largest average salary (5835.906)

‘Assist’ rank has the largest average salary (3388.289)

Difference between average salaries = 2447.617

R Command:

aggregate(updated\_data$salary, list(updated\_data$rank), mean)