General Instructions

15 questions. The first 12 are fill-in-the-blank/short-answer type questions, while the final three require you to manipulate a dataset. You should have 4 Total files

For problems 13-15, assume that your script and the data file are both located in the current working directory. Therefore, when you read a file in, there is no need to provide a path, just the filename. For example:

myData <- read.csv(“inputFile.csv”, header=TRUE) CORRECT

myDaya <- read.csv(“C:\WSUDocs\Desktop\inputFile.csv”, header=TRUE) WRONG

When you are asked to write a function, all the information that function needs to operate should be included in the parameters. Thus, there should be no user input involved in the function itself.

For questions which ask you to produce a plot, you may use either base R plotting or ggplot2, whichever you prefer. Your plots should have appropriately labeled axes.

1. Given a vector x:

x <- c(4,6,5,7,10,9,4,15)

What R command could I use to find out how many entries in X are less than 7?

1. Complete the following R command to generate a vector of the integers from 1:10

x <- 1\_\_\_\_\_\_

1. What is the default separator character in the paste() function?
2. A categorical variable with a fixed number of levels is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. \_\_\_\_\_\_\_\_\_\_\_\_\_ changes the class of a variable, if possible.
4. What function do you use in R to add a column to a matrix or data frame?
5. Given the vector y:

y <- c(101,85,97,102,76,89,95,94,90,80,82,75,103,100,79,69)

What R command could I use to replace any value greater than 100 with 100?

1. Write a short R function called getNames that accepts two parameters called namesVector and excludeCharacter and returns all the entries in namesVector that **DO NOT** contain excludeCharacter.
2. Would the following data be considered WIDE or LONG?

|  |  |  |
| --- | --- | --- |
| Control Treatment | Preheated Treatment | Prechilled Treatment |
| 6.1 | 6.3 | 7.1 |
| 5.9 | 6.2 | 8.2 |
| 5.8 | 5.8 | 7.3 |
| 5.4 | 6.3 | 6.9 |

1. Given a matrix x, what R command would I type to return all the columns in row 5?
2. Given two vectors:

x <- c(3,2,4)

y <- c(1,2)

The command z <- x\*y will produce a vector z that contains (3,4,4). Explain why this happens.

1. Write a short R function called replaceMean that accepts a parameter called numberData and returns a vector where any missing data has been replaced with the mean of data. For example

x <- c(1.2, 7.9, 3.4, NA, 4.2, 9.1, NA)

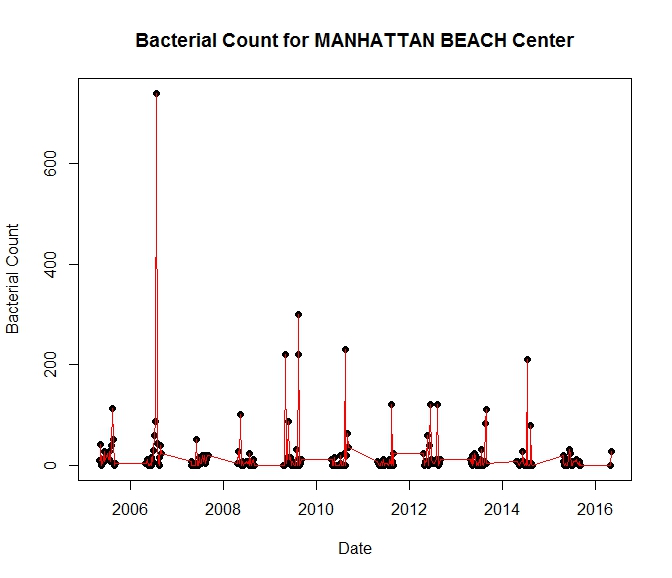
z <- replaceMean(x)

z would now contain (1.2, 7.9, 3.4, 5.16, 4.2, 9.1, 5.16)

1. I have provided you with a dataset in the file MetroMedian.csv. The file contains the median price per square foot of housing in each of the nation’s largest 557 metropolitan areas from April 1996 through December 2016. There are a number of missing entries where the data was not available.
   1. Read this file into a dataframe called metro
   2. This data is not correctly formatted. Produce a dataframe called tidyMetro that is suitable for analysis.
   3. For the entire data set, what is the mean value for the STATE of New York? Show both the R command(s) and the result
   4. Write an R function that accepts two parameters called valueFrame and searchRegion and returns the mean value for all entries for that region.
2. I have provided you with a dataset in the file BeachWaterQualiy.csv. The file contains the bacterial count results from New York beaches from May 2005 to May 2016.
   1. Read this file into a dataframe called beaches
   2. When there was no detectable level of bacteria, the Results field was left blank. These fields were read into the dataframe as NA. Write a single line of R to replace any NA values in the Results column with 0 (zero).
   3. The sample dates recorded in this dataset are not suitable for ordering the data chronologically. Add a column to the beaches dataframe called new.date that is suitable for sorting.
   4. Write an R function called beachPlot that accepts three parameters called beachData, beachName, and sampleLocation. Your function should produce a line plot of the sample results for the named beach and sample location. For example, if I were to call

beachPlot(beaches, “MANHATTAN BEACH”, “Center”)

the function would return a plot that looked similar to



You may assume the dataframe being passed into the function has a column with a sortable date, but you cannot assume that the dataframe is chronologically sorted.

1. I have provided you with a dataset in the file insight.csv. This is the mileage data (again) for my Honda Insight.
   1. Read this file into a dataframe called mileage
   2. Produce and label a scatterplot with blue points that puts Average Temperature on the x-axis and MPG on the y-axis
   3. Add a blue line that fits a linear model
   4. Add red points the put Average Temperature on the x-axis and the “car.said” mileage on the y-axis
   5. Add a red line that fits a linear model
   6. Add a legend that tells me which points are “Measured MPG” and which points are “Car Reported MPG”. You can specify the location of the legend or let the user select it, whichever you wish