**Sheridan College Institute of Technology and Advanced Learning**

**Faculty of Applied Science and Technology**

**Laboratory 2**

**Introduction to R**

Student Name (s) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Student Number \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_\_\_\_\_\_\_\_\_\_

**Objectives:**

* To practice using some library functions in R
* To learn how to visualize the given data

**Instructions:**

This lab relates to a data base called *College* dataset which can be found in the file *College.csv* (provided in the last lab). It contains a number of variables for 777 different colleges in the US. Specifically, the variables are:

1. Using R, and specifically, the *read.csv()* function, read the data into R, and name the loaded data set, *college.*
2. Look at the data using the *fix()* function. Notice that the first column is in fact just the name of the university, we don’t want R to treat this as data, as they are just labels. Therefore, enter the following commands:
   1. *row.names(college) = college[,1]*
   2. *fix(college)*

You should see now that there is a row.names() column with the name of each university recorded. Now, let us **eliminate the first column from the dataset, as we do not require it in our analysis**. Do that as follows: *college = college[,-1]*

*fix(college)*

**Now that the required “cleaning” of the data has been done, it is ready for analysis, please complete the following series of exercises:**

1. ***Create a new qualitative variable, called Elite by binning the Top10perc variable: divide the universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school exceeds 50%.***
2. **Now, use the *summary()* function to see how many Elite universities there are. Use the *plot()* function to produce side-by-side boxplots of *Outstate* vs. *Elite.***
3. **Finally, use the *hist()* function to produce histograms with variable number of bins for a few of the quantitative variables.**

**Evaluation Rubric for the Lab Assignment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Objective** | **Needs Improvement and Unsatisfactory**  **(0-40%)** | **Satisfactory (40-60%)** | **Good (80%)** | **Excellent (100%)** |
| **R code writing and Correct run (40%)** | * Completed less than 50% of the requirements.   + Does not execute due to errors.   No testing has been completed.   * Insufficient use of R code | * Completed 50-60% of the requirements. * Executes without errors. * Some testing has been completed, but did not cover all cases. | * Completed 70-90% of the requirements. * Wrote enough R code to solve the problem * Executes without errors. * Thorough testing has been completed; all possible test inputs were examined | * Wrote detailed and clear R code with comments to solve the problem |
| **Visualization (20%)** | * Do not provide visualization | * Provide at least one visualization | * Provide at least one visualization with explanation | * Provide more than one visualization with explanations of each |
| **Result Analysis (20%)** | * Did not provide sufficient analysis of results | * Provided enough analysis to understand the results | * Provided good analysis to understand the results | * Provided detailed analysis that shows thorough understanding of the results |
| **Clarity and documentation**  **(10%)** | * No code documentation included. * No comments on main sections/subsections of code | * Basic documentation has been completed * Purpose is noted for each function /method | * Clearly documented. * Specific purpose is noted for each function and control structure. | * Clearly and effectively documented including descriptions of all variables. * Specific purpose is noted for each section, control structure, input and output |
| **Delivery (10%)** | * Not delivered in correct format (not submitted online, not in correct MS Word/PDF/ZIP format...) * No name, date, or document title included | * Delivered on time, and in correct format (Word/ PDF/Zip etc) * Includes name, date, and assignment title. | * Completed between 70-90% of the requirements. * Delivered on time, in correct format (Word/PDF/ZIP etc) * Includes name, date, and assignment title. | * Completed between 90- 100% of the requirements. * Delivered on time, and in correct format (Word/PDF/ZIP etc.) * Includes name, date, and assignment title. |