STAT 40001/STAT 59800 **Statistical Computing** **Fall 2020**

Homework 2

Name:

Due : September 24, 2020 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output with ERRORS:). Some of the questions below can be answered with very little or no programming. However, write R code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) Table 1 and Table 2 below are the test scores of 10 students in Test 1 and Test 2

Name	Test 1
Ana	56
Brian	78
Cathy	87
Dough	89
John	95
Lucas	98
Marcus	59
Nabin	78
William	87
Zoe	98

Table 1: Test 1 Scores

Name	Test2
Ana	86
Brian	67
Cathy	78
Dough	89
John	87
Lucas	67
Marcus	94
Nabin	78
William	81
Zoe	83

Table 2: test 2 scores

a) Use ***merge(.,.)*** to create a single table containing the student's test 1 and test 2 scores.
b) How many students did better in the second test?
c) How many students did better in the first test?
d) How many students have the same score in both tests?
e) Calculate the average and standard deviation of both tests.

Q.N. 2) The dataset related to health insurance customers is provided in custdata.tsv. Here, "tsv"stands for tab-separated values.
a) Import the data in R
b) Using "ggplot" package display the age distributions of the customers.
c) Display the marital status of all the customers using bar graph.
d) How many customers are from the state of Indiana?

Q.N. 3) Access the data from url `http://www.stat.berkeley.edu/users/statlabs/data/vote.data` and store the information in an object named **vote** using the function *read.table()*. This includes the 1988 Stockton Primary Exit Poll Survey:
a) How many variables are included in the survey? Please print the variables.
b) One of the variable included is the voter's race. Note that following code are used.

$$\mathbf{0 = missing, 1 = White, 2 = Hispanic, 3 = Black, 4 = Asian, 5 = Other}$$

Display the distribution of the voter's race graphically.

Q.N. 4) This dataset (YouthRisk, provided with this assignment) is derived from the 2007 Youth Risk Behavior Surveillance System (YRBSS), which is an annual survey conducted by the Centers for Disease Control and Prevention (CDC) to monitor the prevalence of health-risk youth behaviors. This dataset focuses on whether or not youths have recently (in past 30 days) ridden with a drunk driver. The description of the variables is provided at
`https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/YouthRisk.html`
a) Import the data in R and determine its dimension.
b) Is there any missing value? If so please remove the missing values from the data set.
c) Display the age distribution of the individuals based on gender using Parallel boxplot.
d) Display the grade(Year in high school) distribution using a pie chart.

Q.N. 5) Consider the following data regarding the average spending on health care per person for various countries published in scholastic update in 2001.

Country	Amount (**$**)
United Kingdom	1992
Czech Republic	1106
Italy	2212
Germany	2808
France	2561
Canada	2792
U.S.	4887

Construct a piechart to display the information.

Q.N. 6) The data frame ***College*** from the ***ISLR*** package has Statistics for a large number of US Colleges from the 1995 issue of US News and World Report.
a) How many variables are included in the survey? Please print the name of the variables.
b) The variable "Outstate" provide the Out-of-state tuition fee and the variable "Private" is a factor with levels No and Yes indicating private or public university. Create a side-by-side boxplots of out of state tuition fee for private and public University.
c) The variable "PhD" provide the Percentage of faculty with Ph.D. and the variable "Private" is a factor with levels No and Yes indicating private or public university. Create a side-by-side boxplots of the percentage of faculty with PhD for private and public University.

Q.N. 7) The number of visits to a website on each day by visitors is recorded. If a user accesses the site after 30 minutes of inactivity, that will be logged as a new visit. The data is available in the Brightspace as "website traffic".
a) Create a chart that shows the variability in website traffic for each day of the week.
b) Calculate the numerical summary of the website traffic data for each day of the week.