STAT 40001/STAT 50001 **Statistical Computing** **Fall 2022**

Homework 2

Name:

Due : September 29, 2022 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 in addition to their major field of study and class standing

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

Table 1: Test 1 Scores

Name	Class	Test 2
Ana	Junior	86
Brian	Junior	67
Cathy	Senior	78
Dough	Junior	89
John	Senior	87
Lucas	Senior	67
Marcus	Junior	94
Nabin	Senior	78
William	Senior	81
Zoe	Junior	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) Display the age distributions of the customers using a histogram.
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) Generate 500 random numbers from normal distribution with mean 10 and variance 25. How many observations are within one, two and there standard deviations from the mean? Compare your findings with the empirical rule.
According to the empirical rule **68%**, **95%** *and* **99.7%** *data reside within one, two and three standard deviation of the mean. Does your data meet this rule?*

Q.N. 6) FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. The data provided in the link below contains determinations of FEV on children ages 6-22 who were seen in the Childhood Respiratory Disease Study in 1980 in East Boston, Massachusetts. The data are part of a larger study to follow the change in pulmonary function over time in children.
ID - ID number
Age - years FEV - litres
Height - inches
Sex - Male or Female

`http://www.statsci.org/data/general/fev.txt`
a) Import the data in R. How many children are included in this study?
b) Display the FEV of Male and Female children.
c) Test the hypothesis whether there is a difference in FEV for male and female.
d) Construct a 95% confidence interval for the difference in the mean FEV for male and female students.

Q.N. 7) The employee satisfaction in any job depends on several factors including the salary. The attached data (Employee Satisfaction) provides information about 15000 employees.
a) Import the data in R
b) Display the satisfaction scores for low, medium and high salary employees.
c) Test whether the job satisfaction level for high earning employees is significantly different from the low earning employees.

Q.N. 8) Results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions is provided in the data frame `PlantGrowth` in the R dataset.
a) How many observations are recorded in the data set?
b) What is the mean of each of the control and treatment conditions?
c) Test the hypothesis whether there is a significance difference between the treatment 1 and treatment 2.

Q.N. 9) The babies data frame in the `UsingR` packages has a collection of variables taken for each new mother in a Child and Health Development Study. The variable age contains the mom's age and the variable `dage` contains the dad's age for several babies. Do a significance test of the null hypothesis of equal ages against a one-sided alternative that dads are older.

Q.N. 10) A person makes a doctor appointment, receives all the instructions and doesn't show up for appointment, Who to blame? Data set containing some information including the age, gender are provided in the data set (Noshow). (Other variable names are self-explanatory).
a) Import the data in R and identify its dimension
b) Print the variables included in the dataset.
c) Display the Age distribution by gender creating parallel box plot.
d) Test whether female are more likely to miss the appointment than male.
e) Are female older than male? Perform the test.
(Hint: xtabs function in R will be useful to create Cross-Tabulation in part(d))