STAT 40001/STAT 59800   **Statistical Computing**   **Fall 2020**
Homework 3

                                                    Name:
Due : October 6, 2020   PUID:

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

Q.N. 1) Generate 500 random numbers from normal distribution with mean 50 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. 2) 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.

Q.N. 3) 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. 4) The `MPV` package in R contains data set named *table.b1* related to National Football League 1976 team preference. Note that there were 28 teams in NFL in 1976.
a) How many variables are included in the data set?
b) The variable y is the number of games won per 14-game season. How many teams win 10 or more games?
c) The variable x_1 and x_2 represent the rushing yards and passing yards respectively of each team. Calculate the numerical summary(the mean, median, standard deviation etc.) of both variables.

Q.N. 5) 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. 6) The CO2 emission (metric tons per capita) provided by Millennium Development Goals Database, UN Statistics Division are attached to this assignment. The databse provides information from 1960-2014.
a) Import thye data in R
b) Select the data for USA, United Kingdom and Canada and display their emission rate over time.

Q.N. 7) 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. 8) 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) Test whether the SMS reminder helped not to miss the appointment.
f) Are female older than male? Perform the test.(Hint: xtabs function in R will be useful to create Cross-Tabulation