Assignment 1a

Use missing_data.csv and perform data pre-processing

a. identify rows and columns we can remove
b. impute data – try 2 different methods

Assignment 1b

Data - blood_sample.xls

There are three discrete variables, sex (1=male; 2=female), laboratory (lab 1 to lab 6), and age group (1=65-69; 2=70-74; 3=75-79; 4=80-84; 5=85-89).

The other variables are continuous. They are age of patient, alkaline phosphatase (IU / L), raw calcium (mg / dL), inorganic phosphorus (mg / dL), calcium (mmol / L), and phosphorus (mmol / L)

Tasks:

1. Check the validity of the data. Determine whether it contains observations that do not make sense. Flag of the observations. Assume that there are no extreme values in this data. Possibility of human error while data entry is high. Use other observations to correct the data.
2. Data also contains duplicate values
3. Perform a summary analysis of the continuous variables on clean dataset. Report the mean, median, standard deviation, min and max. summarize the discrete variables also
4. Construct side by side box plots on the six continuous variables with the factor variable as sex. Then construct side by side box plots of the six continuous variables with the factor variable as laboratory.
5. Summarize your findings. Plot the graphs and give inferences
6. Is there any difference exists in levels for male and female?