BME 590 Homework #2

In this homework, the objectives are to:

1. *Use DCC to download dataset and transfer to local directory to work in R*
2. *Get comfortable with exploratory data visualization and data processing methods in R*
3. *Work with dates in R*
4. *Get comfortable using dplyr, ggplot and corrplot packages*
5. *Discover methods for data imputation*

Submit your R Mardown AND knitted html. Please include **ALL codes, plots, and written answers in your R Markdown for full marks. Also, follow the file naming convention, as indicated in the syllabus.**

I. Get Heart Disease Data Set

1) Download Heart Disease data set from UCI Machine Learning Repository (This will not be graded)
   $ wget https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
2) Import the file into R and label the column names properly. Note you can get some necessary information from https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Print the first few lines of the dataframe including your header row.
3) The last column of the data contains diagnosis information ranging from 0 to 4. The "0" label indicates that the subject is healthy, whereas the subjects with non-zero value in this field are positively diagnosed with heart disease. Using dplyr, add a new column named **"diagnosed"** that contains binary information about whether the subject has heart disease or not (e.g. true/false, 0/1, etc). Print the first few rows of your dataframe to confirm you have successfully created the new column.

II. Explore and Visualize Heart Disease Data Set

4) Summarize the data. Your summary should include the following information below. Make sure you show all your code and your output should include the answers to the items below.
   i. number of counts by diagnosis status (use "diagnosed" column info)
   ii. number of counts by sex
   iii. average of cholesterol level by diagnosis status (use "diagnosed" column info)
   iv. min and max of the subjects' age by diagnosis status (use "diagnosed" column info)
5) Using dplyr, create a new dataframe named **"heart_df"** that contains information on the subjects' age, sex, serum cholesterol level, and maximum heart rate, as well as the new column "diagnosed".
6) Density plots visualize the data distribution. Using ggplot, draw a density function curve of the subjects' age for each of the diseased and healthy groups. What is your observation? Label your axis and legend appropriately for full credit.
7) Correlation plots are a way to visualize multivariate relationships. Using the *corrplot* package, make a correlation plot of the attributes fields found in the new dataframe "heart_df" that you created in

Question (6) above. Clearly label your axis and legend for full mark. Which factor has the strongest correlation with the diagnosis status?

8) Using ggplot, make a scatter plot of age and maximum heart rate. Label (color-code) data points for their corresponding diagnosis status and label the axes.
9) The scatter plot you generate in Question (8) above is probably not the most effective way of visualizing our data if we want to learn about the relationship between diagnosis status and age or max heart rate. Choose other types of charts (besides correlation plot) to show how age and max heart rate are related to the subjects' diagnosis status. Clearly label your chart. What is your observation? Why did you choose the chart of your choice?

III. Understand and Transform Heart Disease Data Set

10) A dataset is said to be skewed if the distribution is asymmetrical and shifted to one direction. Show ALL codes / plots / answers to the following:
    i. Plot a histogram for serum cholesterol level.
    ii. Is this dataset skewed?
    iii. Does it have positive skewness or negative skewness?
    iv. Compute the skewness using the definition from the lecture.
    v. According to the criterion introduced in the lecture, is the cholesterol level dataset moderately skewed or highly skewed?
11) Sometimes, it is necessary to remove extreme outliers (samples >3*SD from mean), where SD is the standard deviation and IQR is the inter quartile range. Show ALL codes / plots / answers to the following:
    i. Make boxplots of cholesterol level to compare the two diagnosis groups. Note that the serum cholesterol level field has some outliers.
    ii. How many of the cholesterol level datapoints are extreme outliers?
    iii. One way of minimizing the effects of the outliers is capping, also known as winsorization. Typically, you can decide a threshold (typically a specific range of percentiles) and replace all the datapoints outside of the threshold with the closest value from within the threshold. An example from Wikipedia may be a helpful demonstration: "a 90% winsorization would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile." Conduct a 90% winsorization on the cholesterol level dataset and show the new statistics summary. What is your observation?
12) Transforming data can also alleviate the effects of outliers and skewness. Show ALL codes / plots / answers to the following:
    i. Standardize the cholesterol level data.
    ii. Make a density plot of the standardized cholesterol dataset.
    iii. Show the new statistics summary on the standardized cholesterol dataset.

IV. Get Diabetes Data Set

13) Download Diabetes data set from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Diabetes in your work folder in the DCC. Show all your commands.
14) Extract the tarball in the command line/Git BASH through the DCC. (Don't need to show command)
15) Export one file from Diabetes data (data-XX) to your local directory. This represents the data for each participant. Choose any of the participants. Print all your commands.

V. Explore, Visualize, and Process Diabetes Data

16) Select the 3 pre-meal blood glucose measurements (codes 58, 60, 62). Create a table reporting any relevant summary statistics. What is your observation? You may choose to include some plots to support your explanation.
17) Use ggplot to display the 3 blood glucose measurements over time.
18) Using the 3 meal measurements from above, align the data on a daily basis, i.e. transform the data to a wider format where each line represents a day and has a measurement per-person. Comment on any generated missing data.
19) Use the *corrplot* package to make a correlation plot. What happens if you ignore the missing data?
20) One version of single imputation appropriate for longitudinal data is called "hot-deck" imputation or "last observed carried forward (locf)". Here the data are ordered (typically by time) and the last observed value is imputed into any time slots without an observed value. Using the data above perform two versions of LOCF and recalculate your correlation plots.
    i. Carry the last observation forward separately for each of the 3 categories
    ii. Carry the last observation forward from the time of day (i.e. impute glucose before lunch using glucose before breakfast).
    iii. Comment on the implications of each approach. Which approach do you think is better? *Hint: The zoo package has the function na.locf()

Make sure you show ALL COMMANDS,

WRITTEN ANSWERS, and PLOTS in R Markdown

& Knit to HTML

Submit both .Rmd & .HTML with correct naming convention