CS614 – Prof. Berardi Name:___________________________________
HW2
Due September 30, 2018

In this assignment, you will practice creating and interpreting visualizations in R. You will use a data frame that summarizes longitudinal data from individuals wearing accelerometers over the course of several months. Each day participants were assigned a goal for moderate-to-vigorous physical activity and were rewarded if this goal was met. They were instructed to wear the accelerometer during all waking hours, but periods of non-wear were common. The data consists of demographic features and daily averages. The included variables are:

participant_id
goal_minutes: average daily goal (in minutes) throughout the study
wear: average # of minutes the accelerometer was worn each day
bout.min: average # of minutes with moderate to-vigorous physical activity each day
Age
Sex
BMI
income: classified as high/low
walk : classified as high/low

Perform the following tasks:

1. Study recruitment was based on income/walkability quadrants where the design called for roughly equal number of people from Hi Income/Hi Walk, Hi Income/Low Walk, Low Income/High Walk, and Low Income/Low Walk. Determine if the recruitment goal of equal quadrants was met by creating a new variable that combines the information in the income and walkability variables. Create a bar plot for this new variable.

2. Is the distribution of the average number of bout minutes or the average number of wear minutes more skewed? Use histograms to support your answer. Why do you think this may be? For the more skewed variable, create a second histogram of the log transform of values.

3. Use the tools we have been developing to explore the goal minutes category. you see anything interesting? Can you come up with ways to improve the information summarized in your histogram?

4. Use side-by-side box plots to determine whether the average number of bout minutes differs by sex. Back up your findings with summary statistics.

5. Eliminate the `income` and `walk` variables and the variable created in Problem 1 from the data frame. Then, using the `plot` function, create a matrix of scatter plots comparing every variable in the

data set to every other. Identify two strong relationships and provide an explanation for why they exist. Also, explain why several scatter plots have horizontal or vertical lines in the center of the figure.

6. In class when examining the TAO data, we discussed why imputing with the mean or a random value are not optimal imputation methods. Here, two superior imputation methods are suggested. Please choose one and implement it on the TAO data. Note that your procedure should be implemented separately for the 1993 and 1997 data. Provide and updated `humidity` versus `air temp` scatrer plot. Try to plot the 1993 and 1997 data together like we did in class.

Method 1: (Regression Approach): For each variable in the data frame that has missing data, create a regression model (using the `lm` function) that uses all of the other variables as predictors. This regression model can be used to simulate missing values for this variable. However, in order to not have every missing entry be assigned the same value based on the regression coefficients, introduce a noise component by adding an error term sampled from the distribution of the residual.

Method 2: (Multivariate Normal Approach): A multivariate distribution allows all missing values to be imputed at once. This can be done easily using the `norm` package. Use the `prelim.norm` function to obtain multivariate characteristics for the data and then use `em.norm` on this result to find maximum likelihood estimates for the variables. `getparam.norm` will display the output in a legible manner that should make more sense. Lastly, use `imp.norm` to impute the missing values. When using `imp.norm`, you must call the `rngseed` random number generator.