**Module 1 Exercises**

1. (30 points) The [UC Irvine Machine Learning Repository](http://archive.ics.uci.edu/ml/datasets/HCV+data) contains a data set related to glass identification. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe. The data can be accessed via:

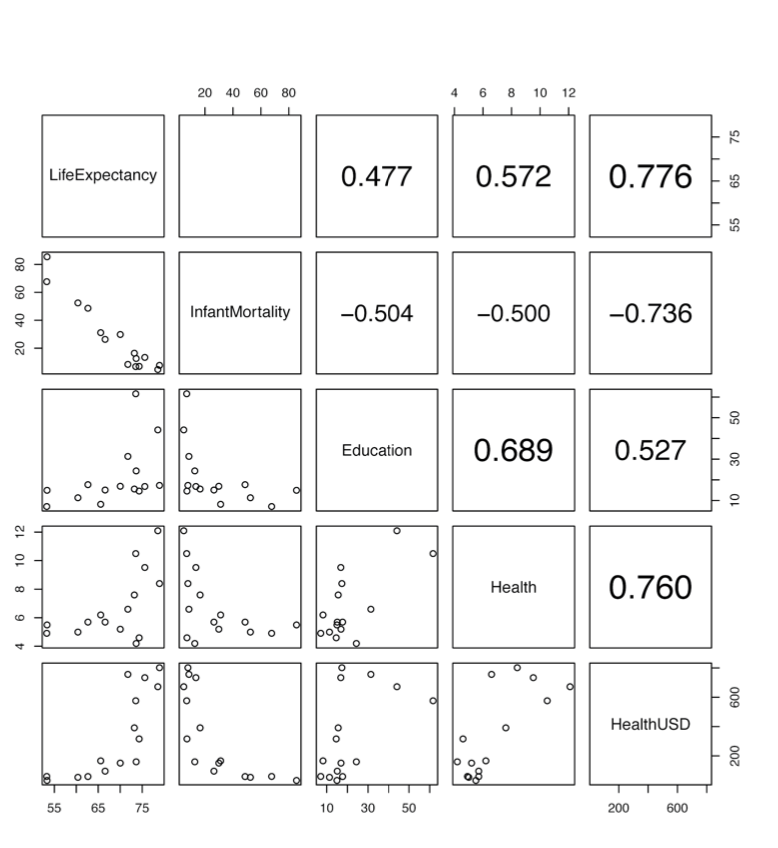
library(mlbench)

data(Glass)

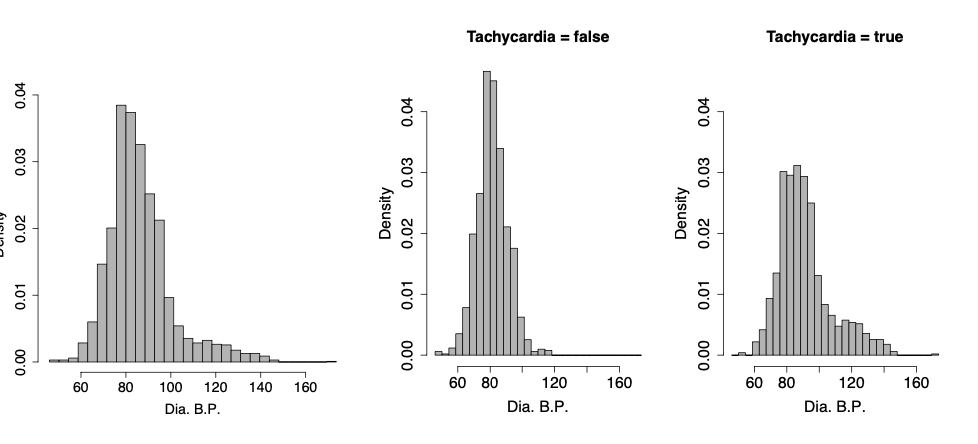
1. Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors.
2. Do there appear to be any outliers in the data? Are any predictors skewed?
3. Are there any relevant transformations of one or more predictors that

might improve the classification model?

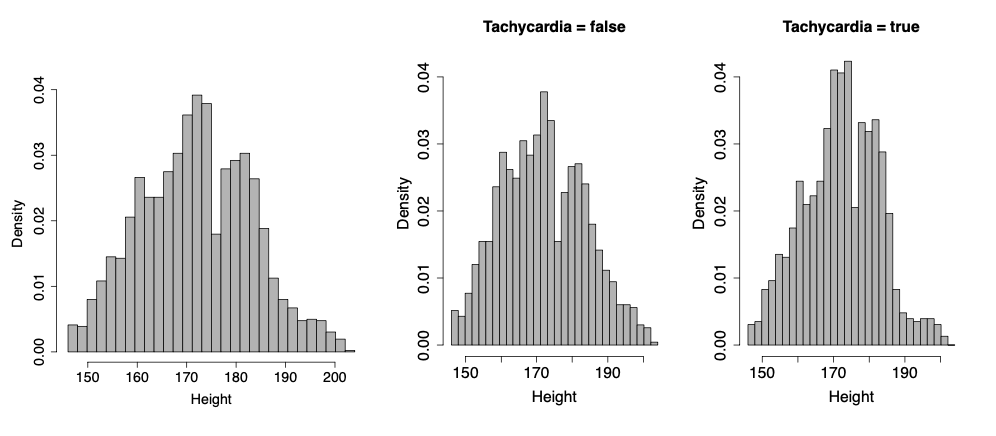
1. (20 points) The image below shows a scatter plot matrix of the continuous features of a dataset. Discuss the relationships between the features in the dataset that this scatter plot highlights.



1. (10 points) Discuss the relationships shown in each visualizations:
   1. The visualization below illustrates the relationship between Diastolic BP & Tachycardia, left most plot has data where Tachycardia = true & false.



* 1. The visualization below illustrates the relationship between Diastolic BP & Tachycardia, left most plot has data where Tachycardia = true & false.



1. (30 points) Use the data at the [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/datasets/HCV+data) web page (or download the “hcvdat0.csv” file in Blackboard) and pick the numeric predictors (exclude columns X & Age) to perform the following analysis in R:
   1. Are there any missing data in the predictors? If yes, summarize the missing data by each predictor.
   2. Are there any predictors with skewed distributions?
   3. Plot histograms of all predictors to observe skewness visually
   4. Compute skewness using the skewness function from the e1071 package.
   5. Apply box cox transformations to the data then recompute the skewness metrics and report the differences and does box cox transformation help?
   6. Plot histograms of transformed predictors to observe changes to skewness visually.