This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The aim here is to use the Principal Component Analysis (PCA) and/or Partial Least Squares Regression (PLS) to find whether the existence possible relations between the variables of the dataset or not.

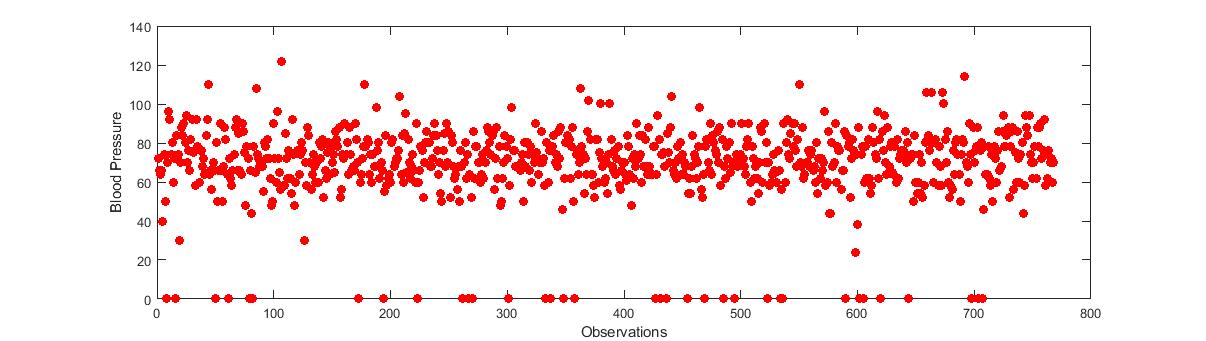
The datasets consists of several medical predictor variables, which are:

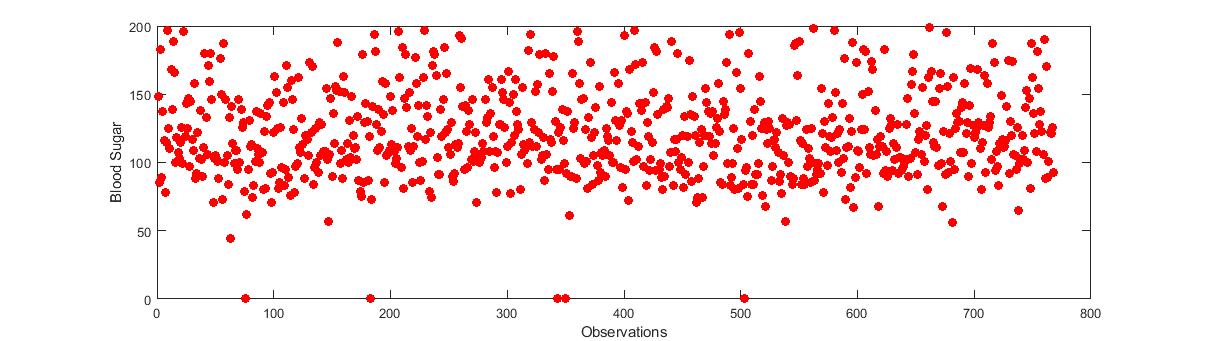
1. The number of pregnancies the patient has had.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test (Blood Sugar).
3. Diastolic blood pressure, i.e., the number on the bottom (Blood Pressure).
4. Body mass index (Mass).
5. Diabetes pedigree function (pedigree).
6. Age.
7. The presence of diabetes (Diabetes).

After plotting each variable separately, we found that there are three variables with missing values. The variables are:

1. Blood Sugar.
2. Blood Pressure.
3. Mass.

Figure (1) bellow shows those missing value, where there is no human being with zero value of Blood Sugar, Blood Pressure, or Mass.





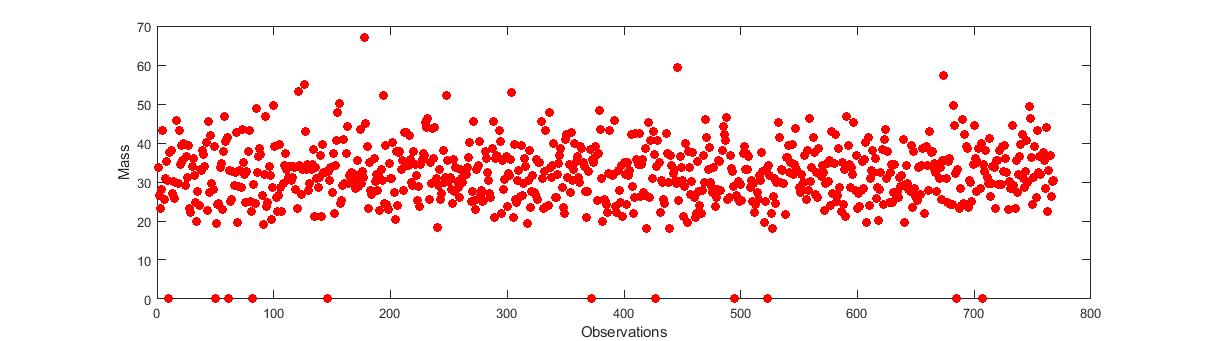


Figure (1): The three variables with missing values.

After removing the rows with those missing value, we are going to see the Principal Component Number that we are going to use