Hamid Muneerulhudhakalvathi N hxm151830@utdallas.edu

ML ASSIGNMENT 3 SOLUTION

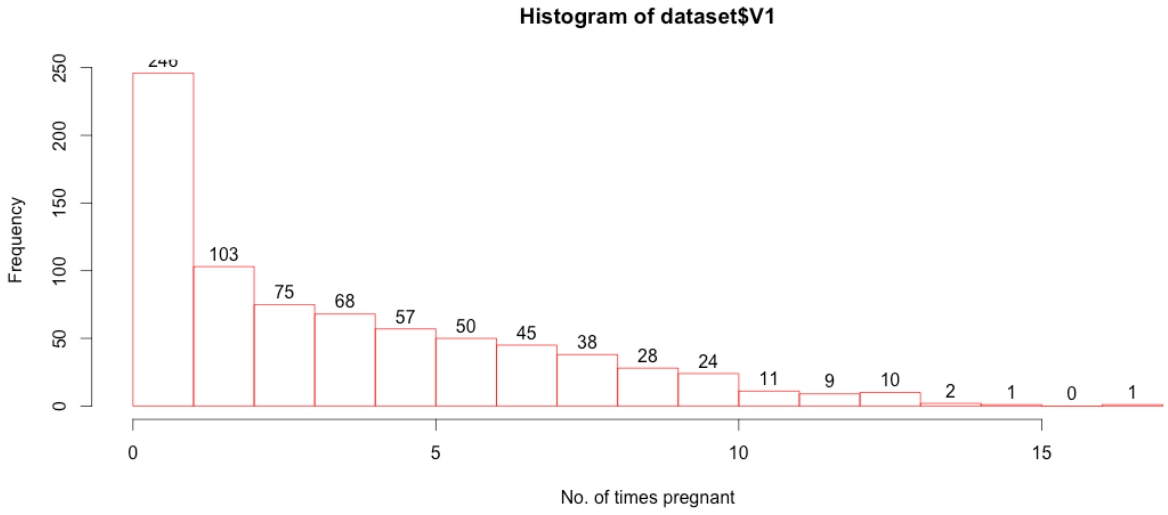
Dataset: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

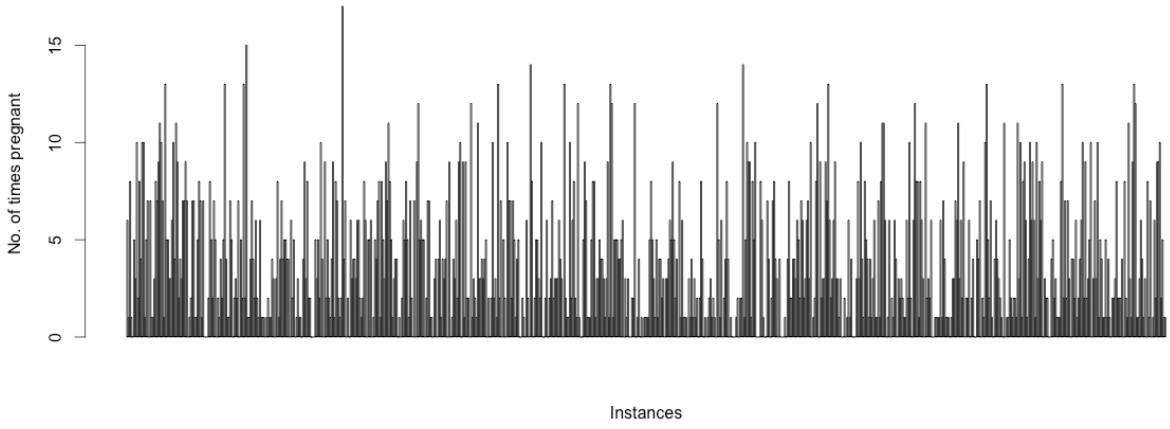
Exploratory data analysis:

1. Create the following plots:
Histogram and Barplot.

a. Number of times pregnant:

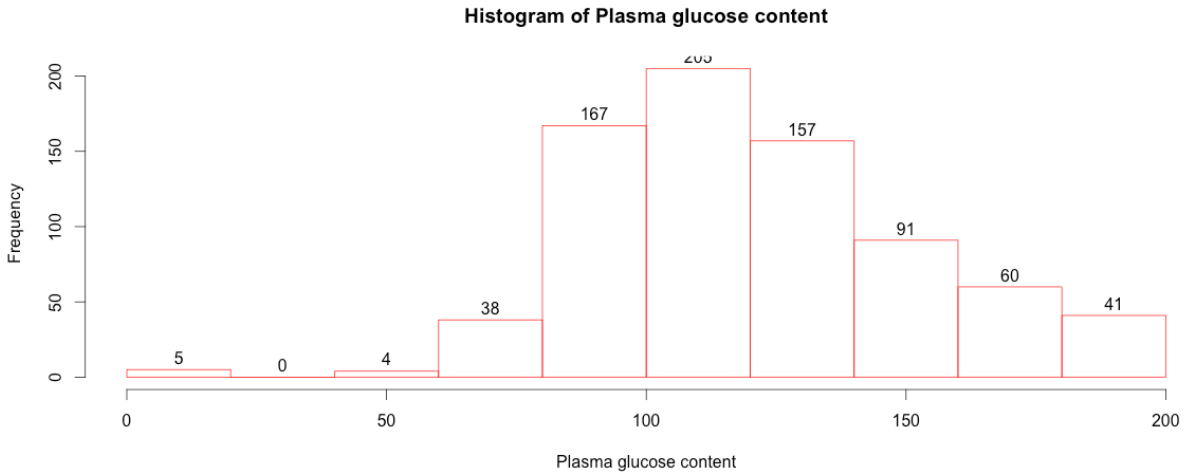
Range: 0.000000 to 17.000000
Mean: 3.845052 Median: 3.000000
No normal distribution.

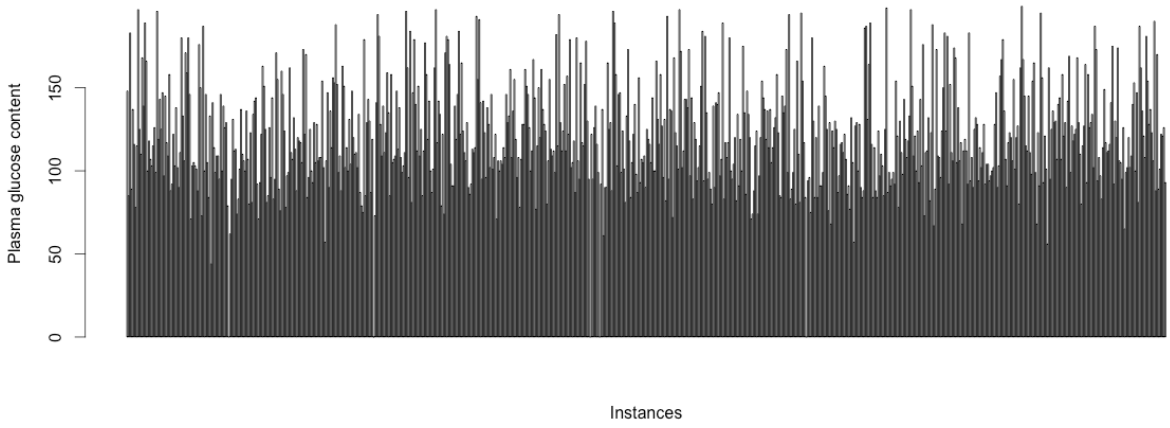




b. Plasma glucose concentration a 2 hours in an oral glucose tolerance test:

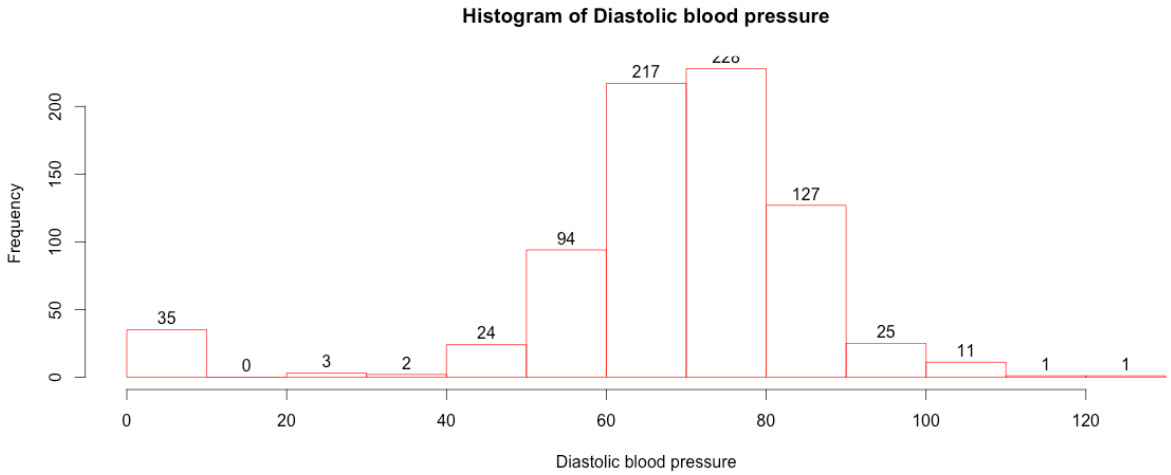
Range: 0.0000 to 199.0000
Mean: 120.8945 Median: 117.0000
Normal distribution.

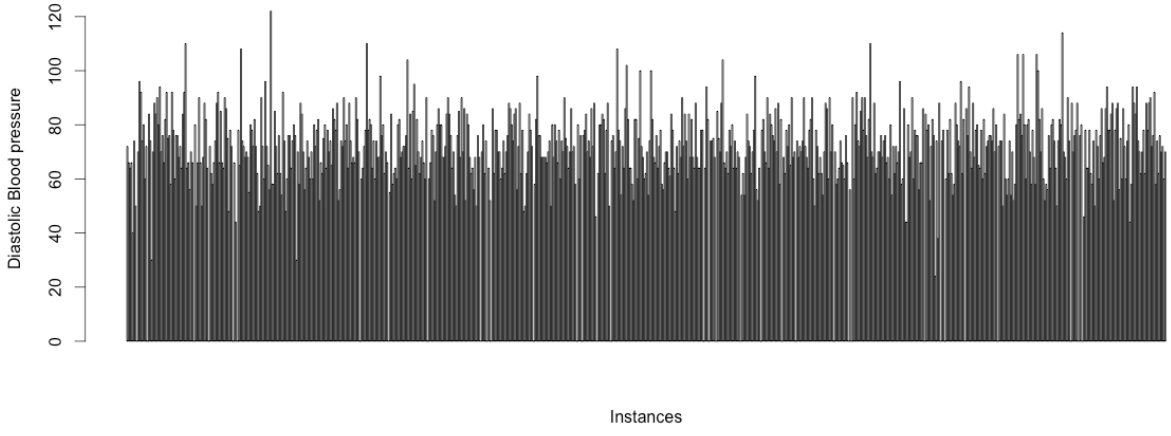




c. Diastolic blood pressure (mm Hg):

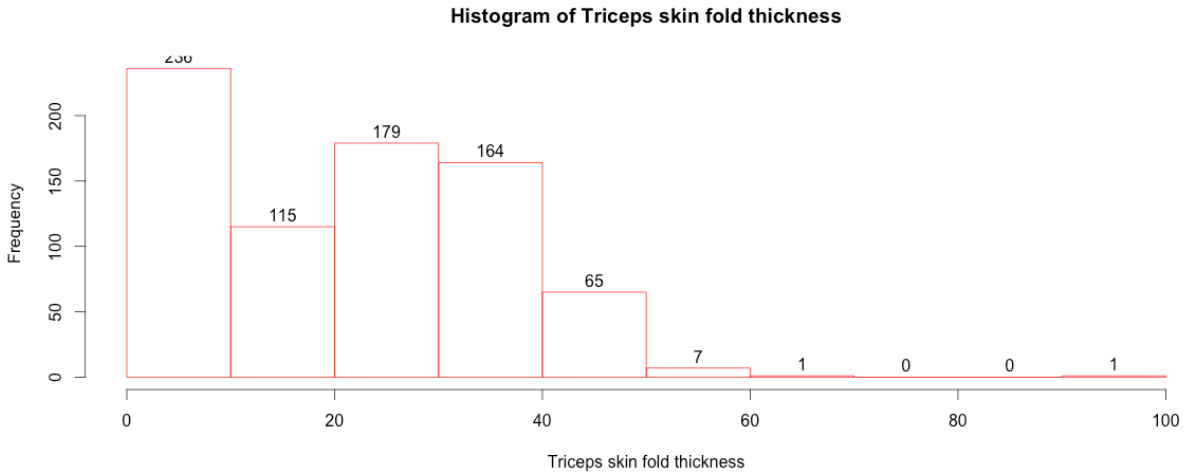
Range: 0.0000 to 122.0000
Mean: 69.10547 Median: 72.00000
Normal distribution.

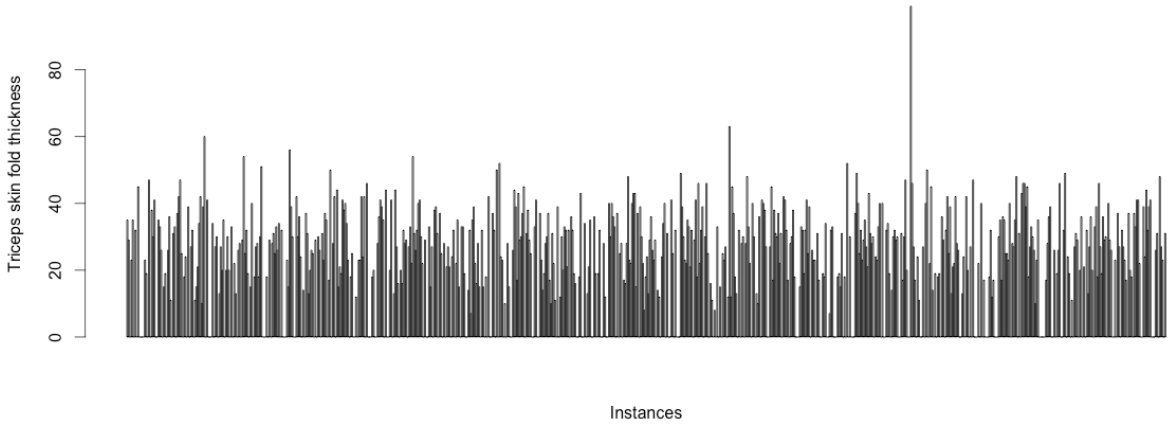




d. Triceps skin fold thickness (mm):

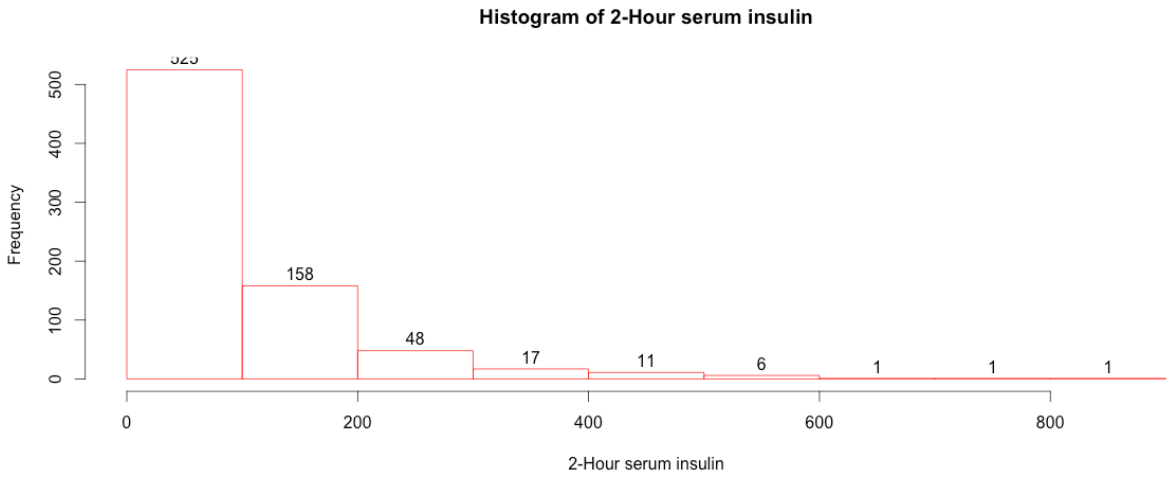
Range: 0.0000 to 99.0000
Mean: 20.53646 Median: 23.00000
No Normal distribution.

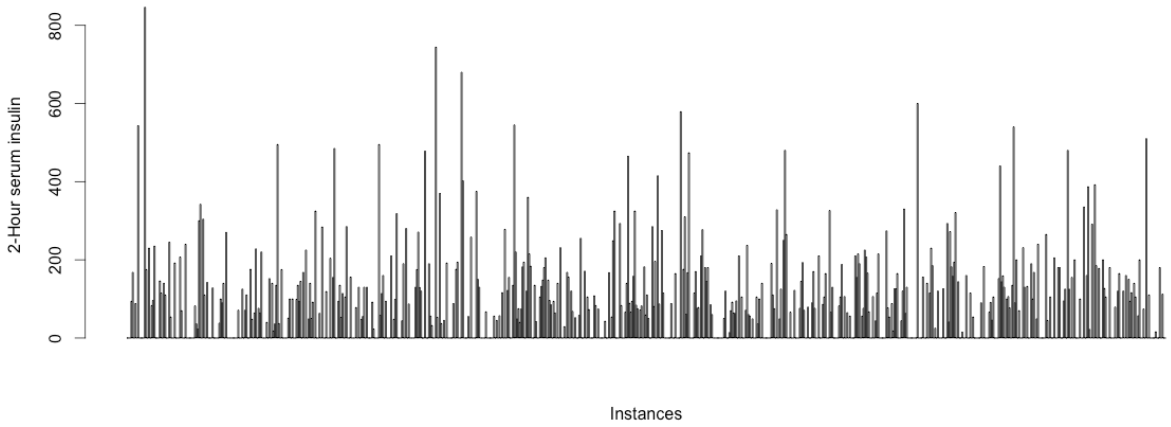




e. 2-Hour serum insulin (mu U/ml):

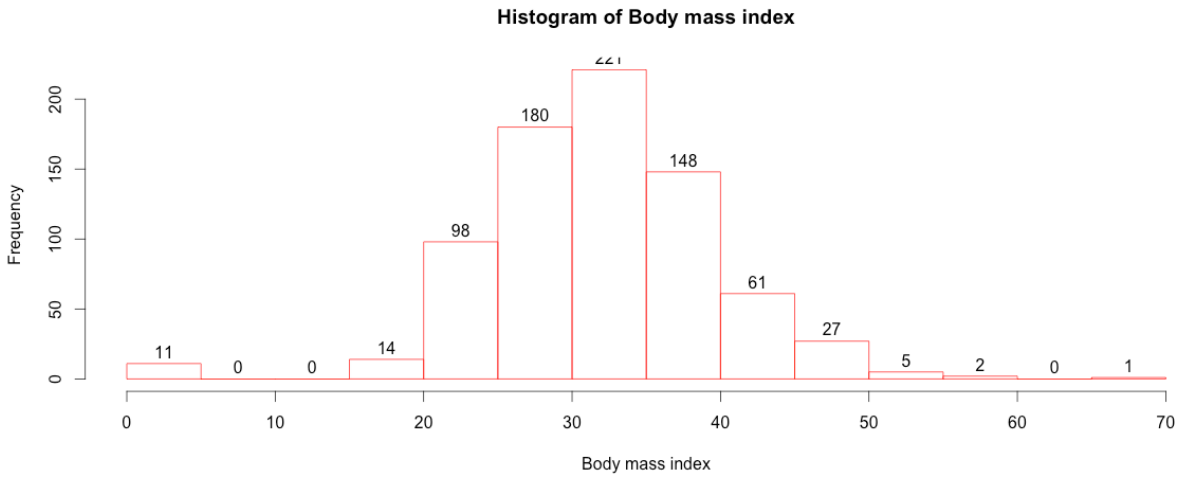
Range: 0.0000 to 846.0000
Mean: 79.79948 Median: 30.50000
No Normal distribution.

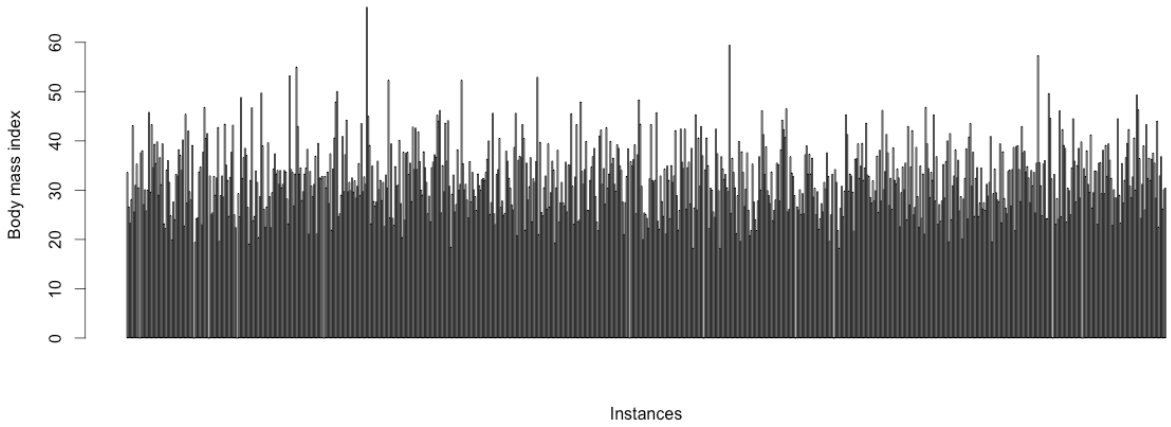




f. Body mass index (weight in kg/(height in m)^2):

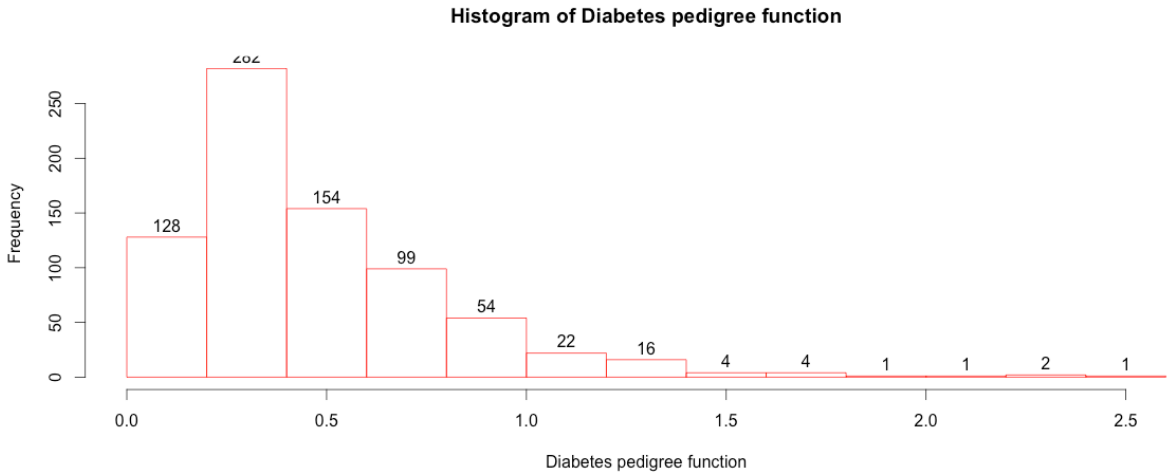
Range: 0.0000 to 67.10000
Mean: 31.99258 Median: 32.0000
Normal distribution.

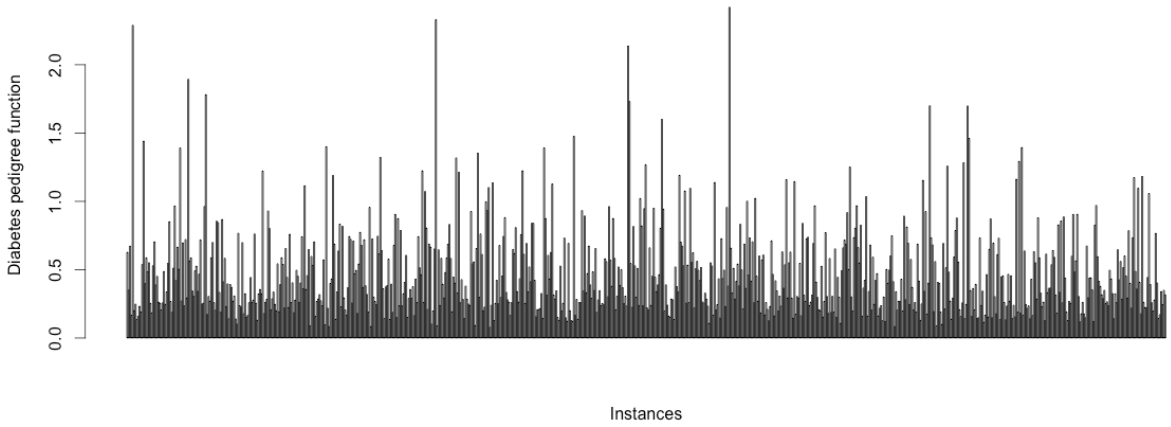




g. Diabetes pedigree function:

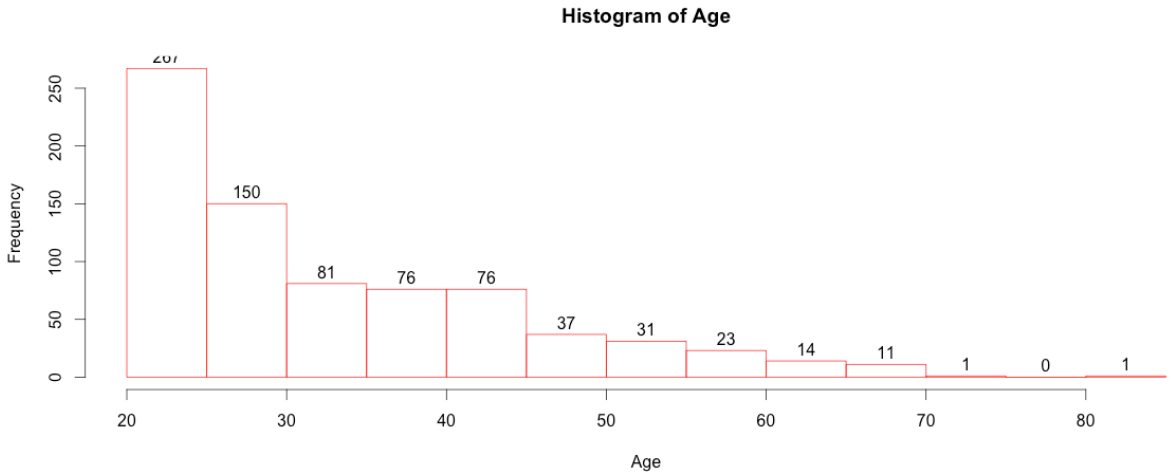
Range: 0.0780000 to 2.4200000
Mean: 0.4718763 Median: 0.3725000
No Normal distribution.

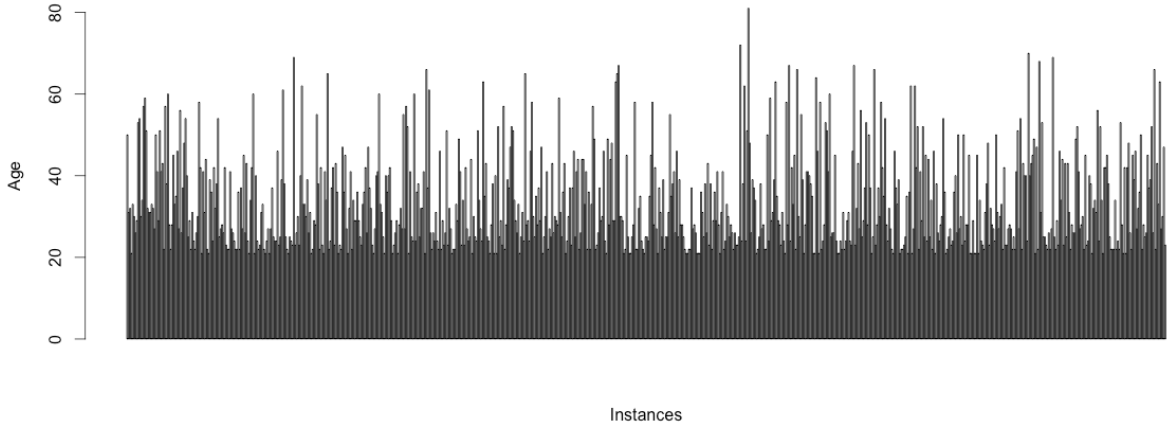




h. Age (years):

Range: 21.00000 to 81.00000
Mean: 33.24089 Median: 29.00000
No Normal distribution.





2. Find the correlation between each of the attributes and the class variable.

Correlation	Class Attribute(0 or 1)
No of times pregnant	0.221898153
Plasma glucose concentration	0.4665813983
Diastolic blood pressure (mm Hg)	0.06506835955
Triceps skin fold thickness (mm)	0.07475223192
2-Hour serum insulin (mu U/ml)	0.1305479549
Body mass index (weight in kg/(height in m)^2)	0.2926946626
Diabetes pedigree function	0.1738440657
Age (years)	0.238355983

”Plasma glucose concentration a 2 hours in an oral glucose tolerance test” has the maximum correlation with the class variable.

3. Compute the correlation between all pairs of the 8 attributes. Which two attributes have the highest mutual correlation?

The correlation between age and the number of times pregnant is maximum.
Correlation : 0.5443412284