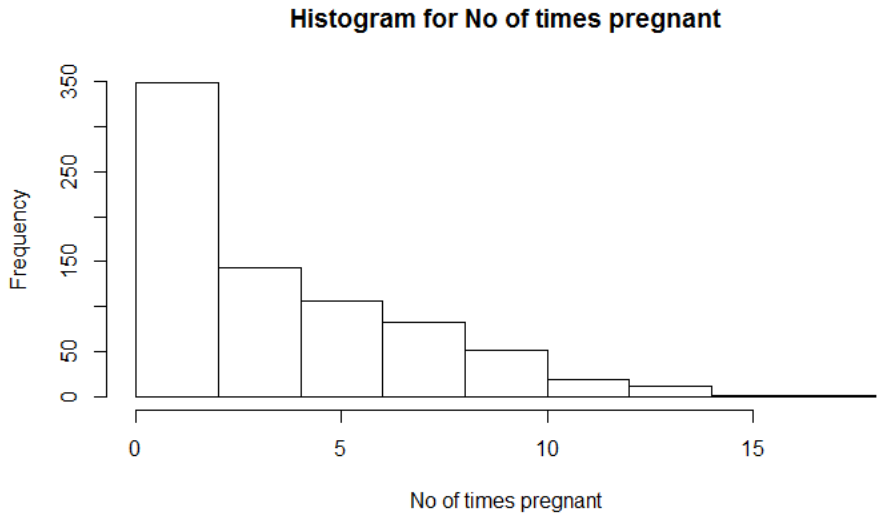
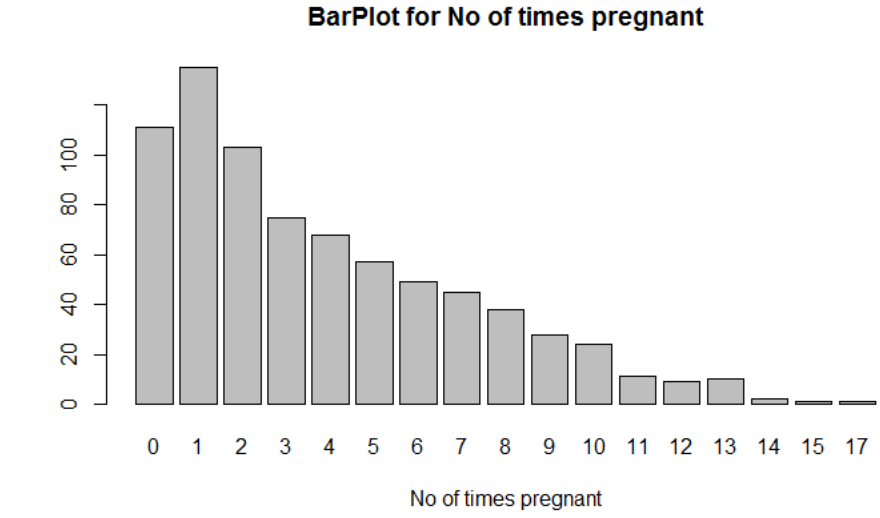
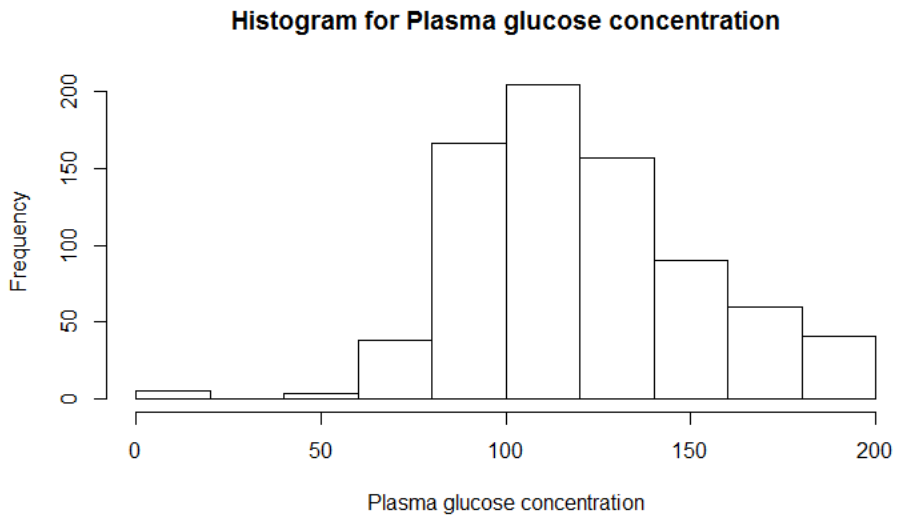
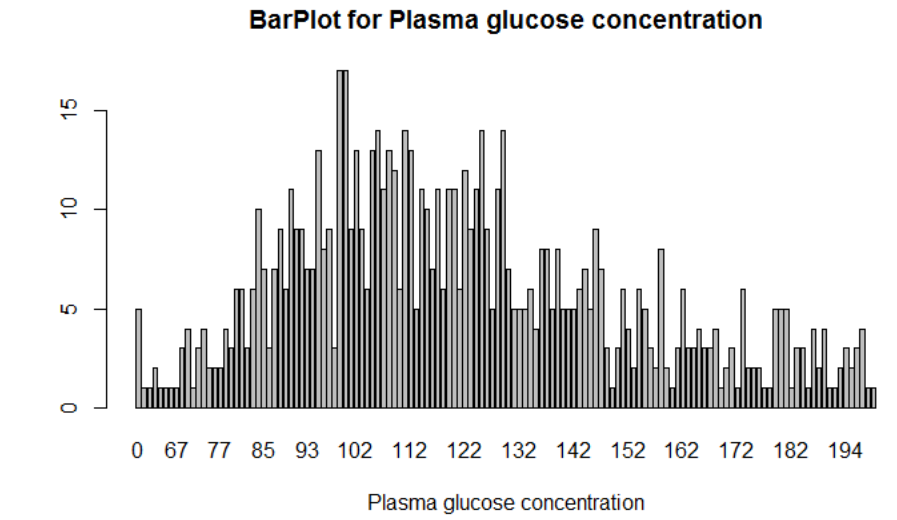
I. Exploratory Data Analysis

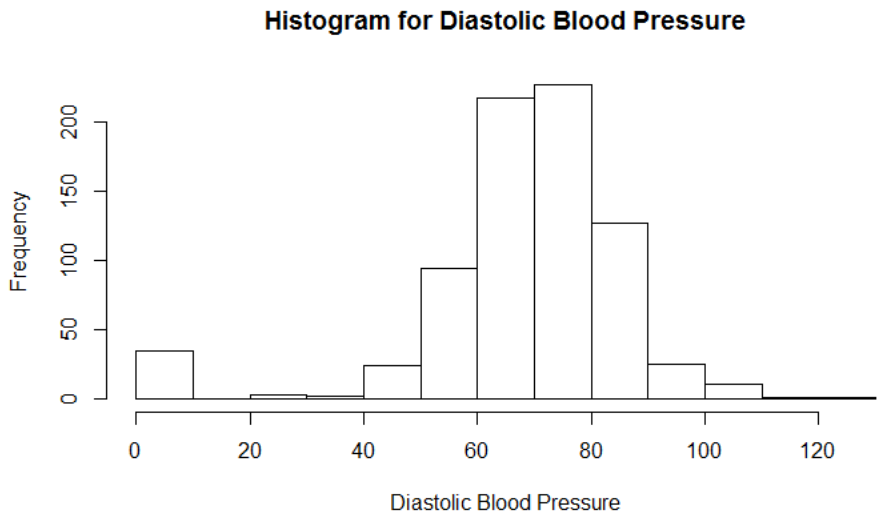
1. The Exploratorydataanalysis.R creates histogram barplots for each of the attributes in the dataset.

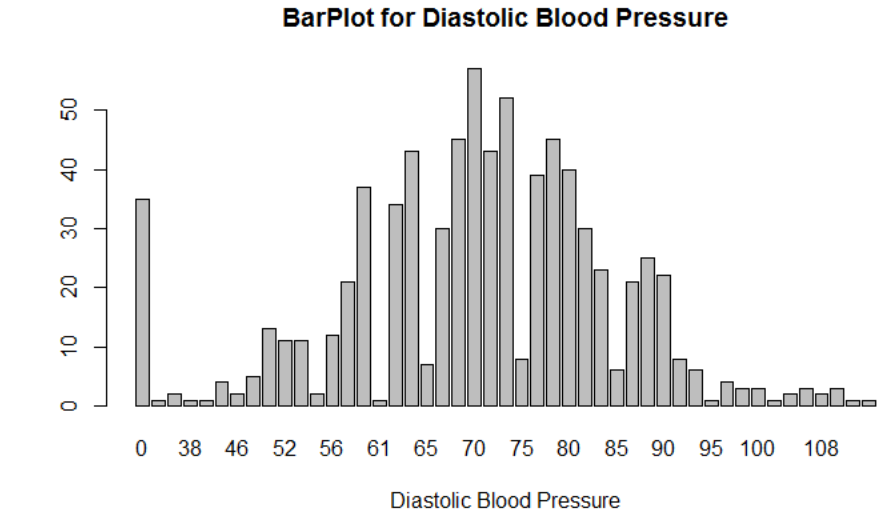


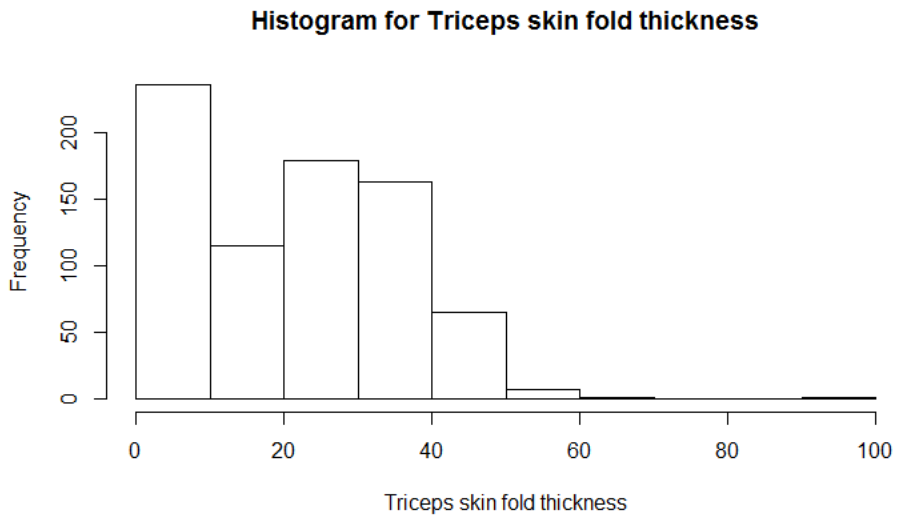


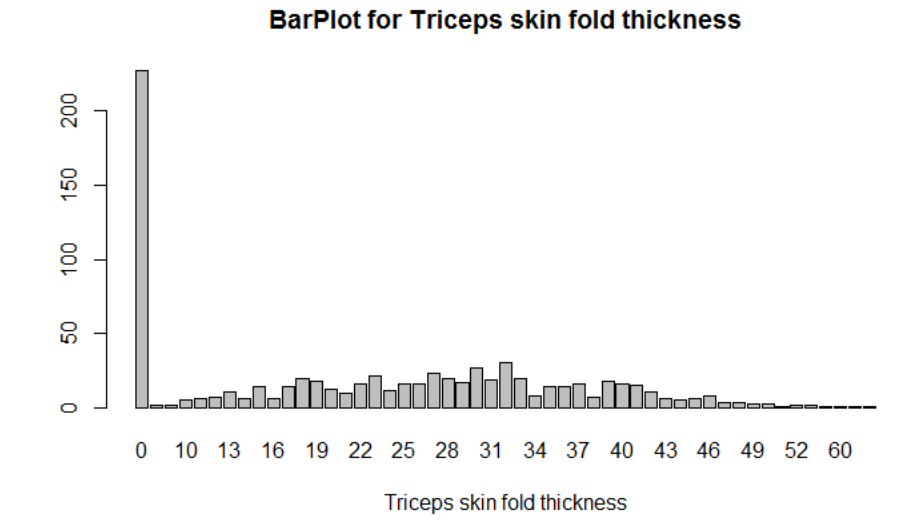

Histogram for Plasma glucose concentration
Frequency
0
50
100
150
200
Plasma glucose concentration

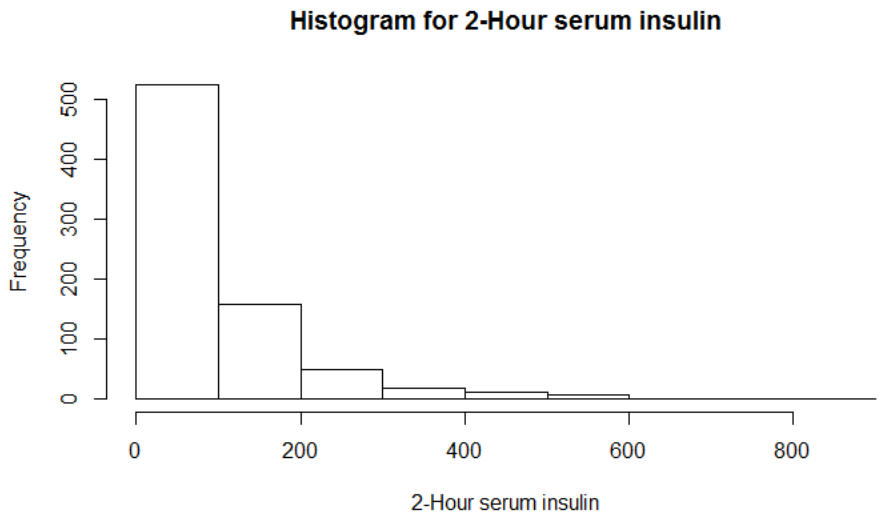

BarPlot for Plasma glucose concentration
0
5
10
15
0 67 77 85 93 102 112 122 132 142 152 162 172 182 194
Plasma glucose concentration

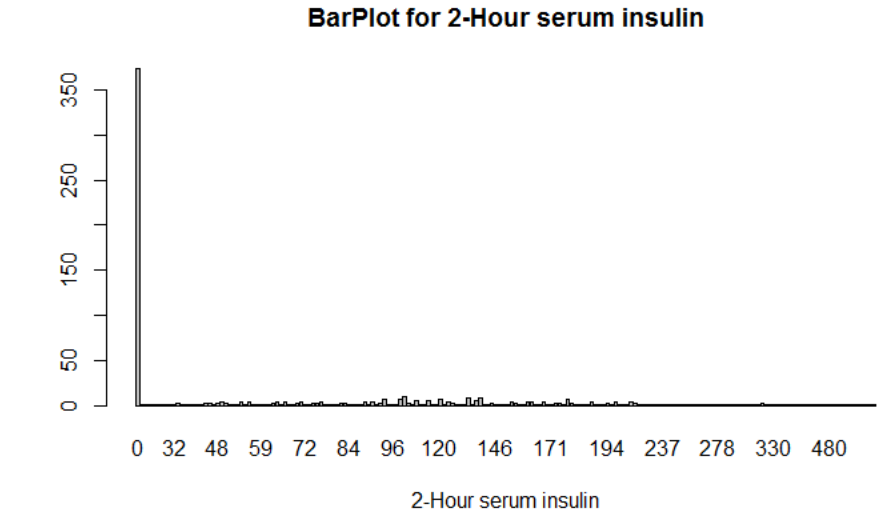

Histogram for Diastolic Blood Pressure
Frequency
0
50
100
150
200
0
20
40
60
80
100
120
Diastolic Blood Pressure

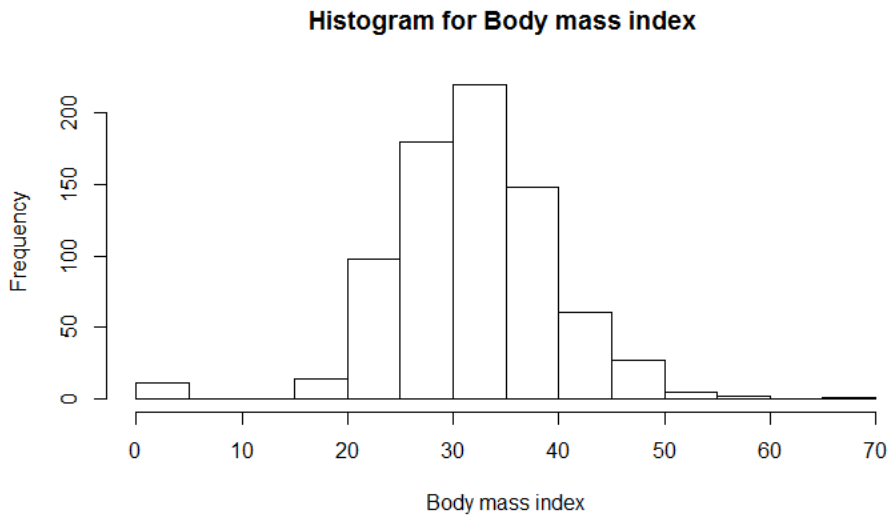

BarPlot for Diastolic Blood Pressure
0
10
20
30
40
50
0
38
46
52
56
61
65
70
75
80
85
90
95
100
108
Diastolic Blood Pressure

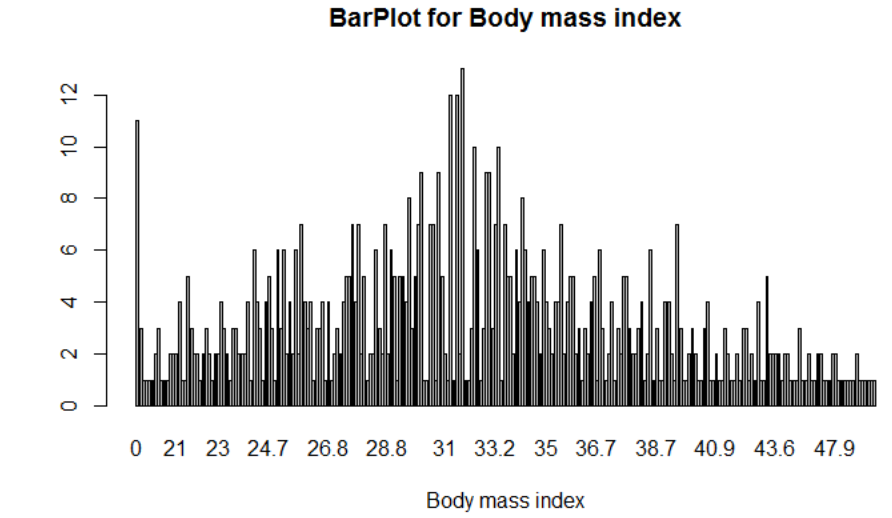

Histogram for Triceps skin fold thickness
Frequency
0
50
100
150
200
0
20
40
60
80
100
Triceps skin fold thickness

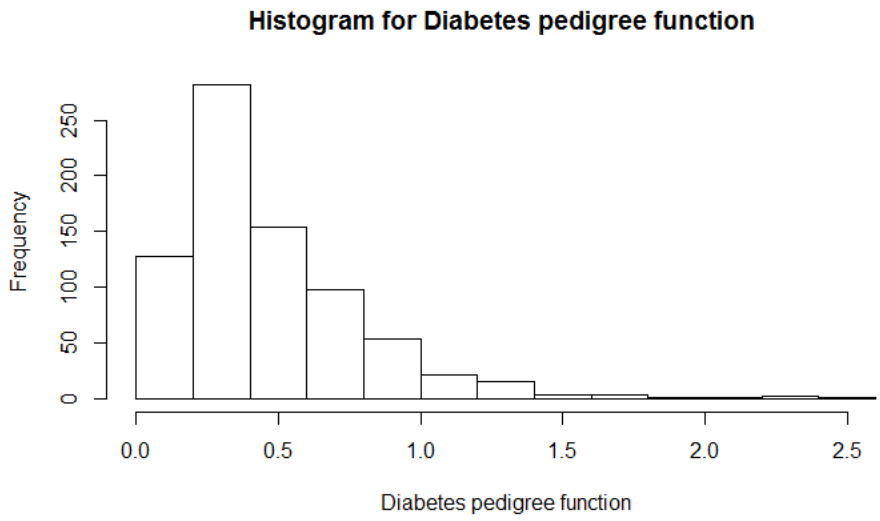

BarPlot for Triceps skin fold thickness
0
50
100
150
200
0
10
13
16
19
22
25
28
31
34
37
40
43
46
49
52
60
Triceps skin fold thickness

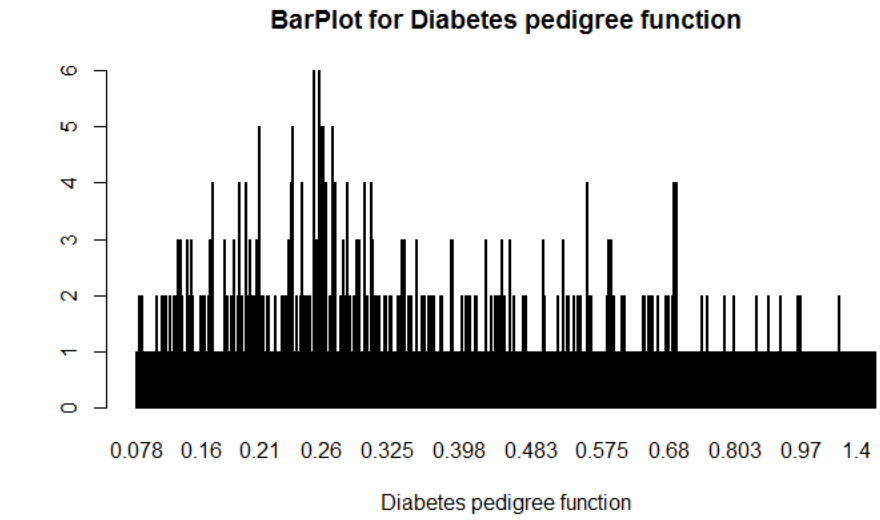

Histogram for 2-Hour serum insulin
Frequency
0
100
200
300
400
500
0
200
400
600
800
2-Hour serum insulin

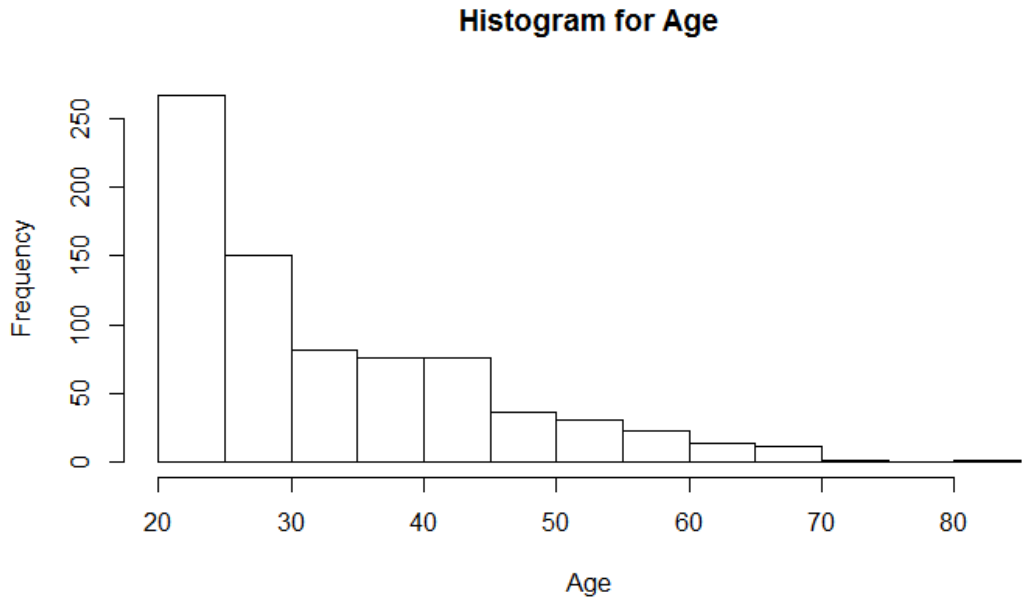

BarPlot for 2-Hour serum insulin
0
50
150
250
350
0
32
48
59
72
84
96
120
146
171
194
237
278
330
480
2-Hour serum insulin

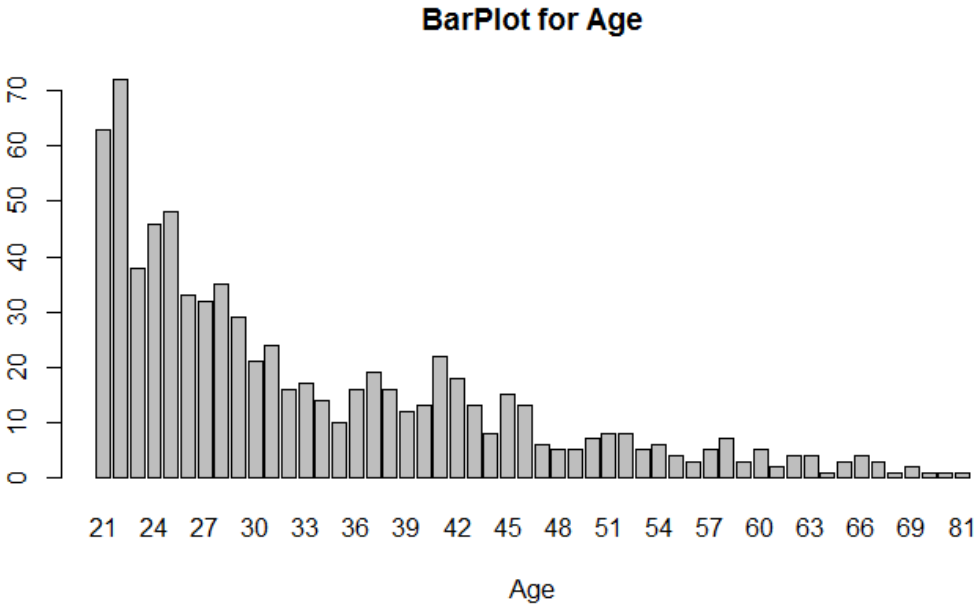

Histogram for Body mass index
Frequency
0
50
100
150
200
0
10
20
30
40
50
60
70
Body mass index


BarPlot for Body mass index
0
2
4
6
8
10
12
0
21
23
24.7
26.8
28.8
31
33.2
35
36.7
38.7
40.9
43.6
47.9
Body mass index


Histogram for Diabetes pedigree function
Frequency
0
50
100
150
200
250
0.0
0.5
1.0
1.5
2.0
2.5
Diabetes pedigree function


BarPlot for Diabetes pedigree function
0
1
2
3
4
5
6
0.078
0.16
0.21
0.26
0.325
0.398
0.483
0.575
0.68
0.803
0.97
1.4
Diabetes pedigree function





From the plots it is observed that Body Mass Index, PG Concentration and Diacolistic BP have a normal distribution. Other variables have distributions other than the normal distribution.

2. Also, the correlation between each of the attributes and the class variable is calculated in Exploratorydataanalysis.R

   Correlation between No_of_Pregnancies and Class Value : 0.22
   Correlation between Plasma Glucose Concentration and Class Value: 0.46
   Correlation between Diagnostic Blood Pressure and Class Value: 0.064
   Correlation between Triceps Skin Fold Thickness and Class Value: 0.073
   Correlation between serum insulin and Class Value: 0.13
   Correlation between Diabetes Pedigree Function and Class Value: 0.173
   Correlation between Body Mass Index and Class Value: 0.29
   Correlation between Age and Class Value: 0.23

   The PG Concentration has strong correlation with the class variable Diabetes with correlation value of 0.46

3. The highest mutually correlated attributes are NPG and AGE with mutual correlation 0.54.

II. Naïve Bayesian Classifier

The package e1071 in R is used to build the Naïve Bayesian Classifier.

Experiment	Accuracy
1	1 74.02597
2	2 75.32468
3	3 74.02597
4	4 83.11688
5	5 81.81818
6	6 80.51948
7	7 83.11688
8	8 77.92208
9	9 76.62338
10	10 80.51948

The average accuracy of the 10 experiments is 78.70

III. SVM Classifier

	Kernel	Average_Accuracy
1	Linear	78.58974
2	Poynomial	71.66667
3	Radial	75.84416
4	Sigmoid	56.79487

IV. kNN

K	Accuracy
3	69.87179
5	71.15385
7	71.28205
9	75.32468
11	71.53846