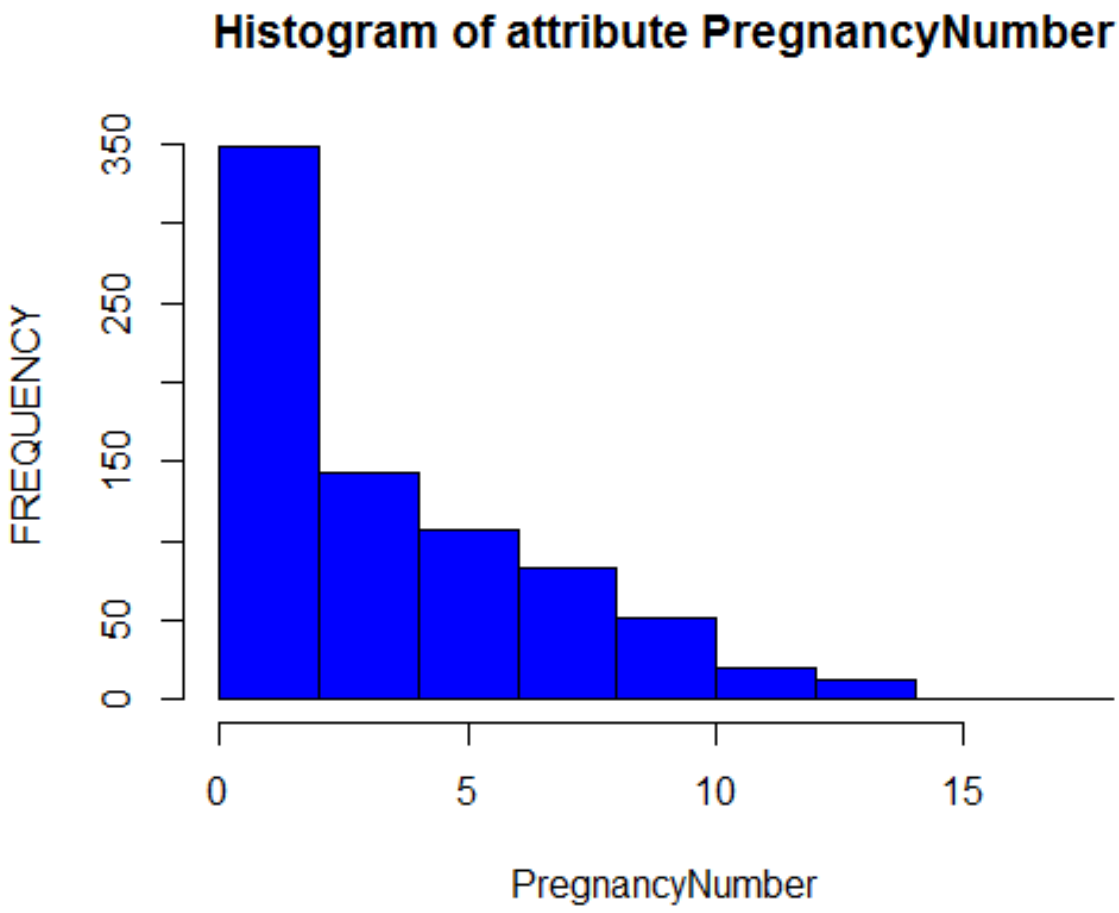
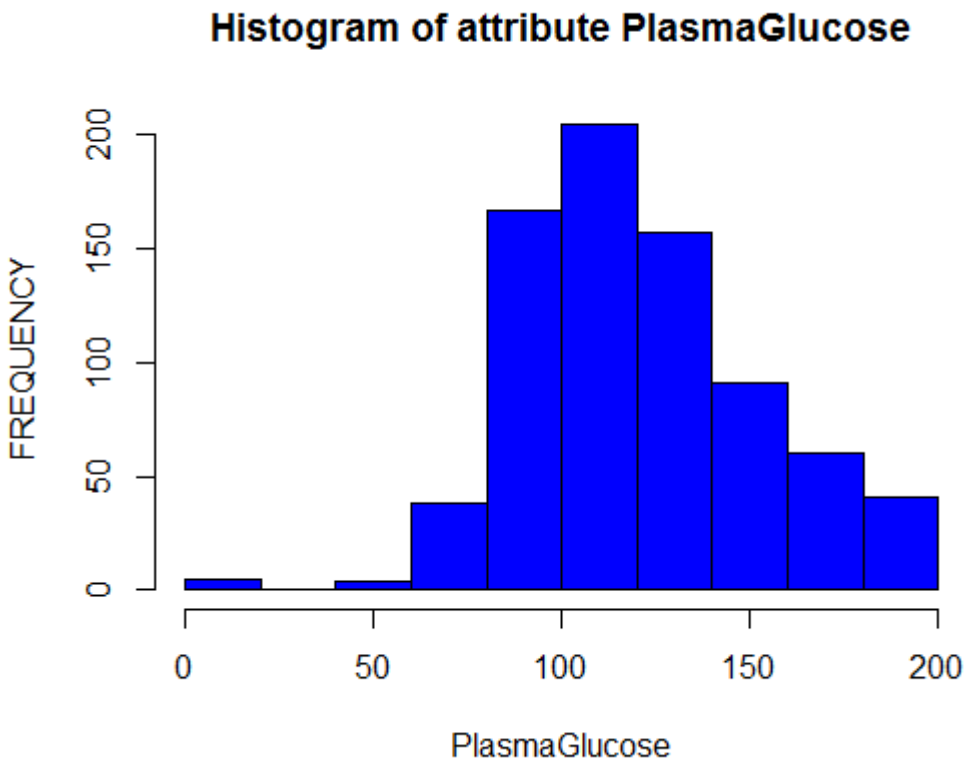
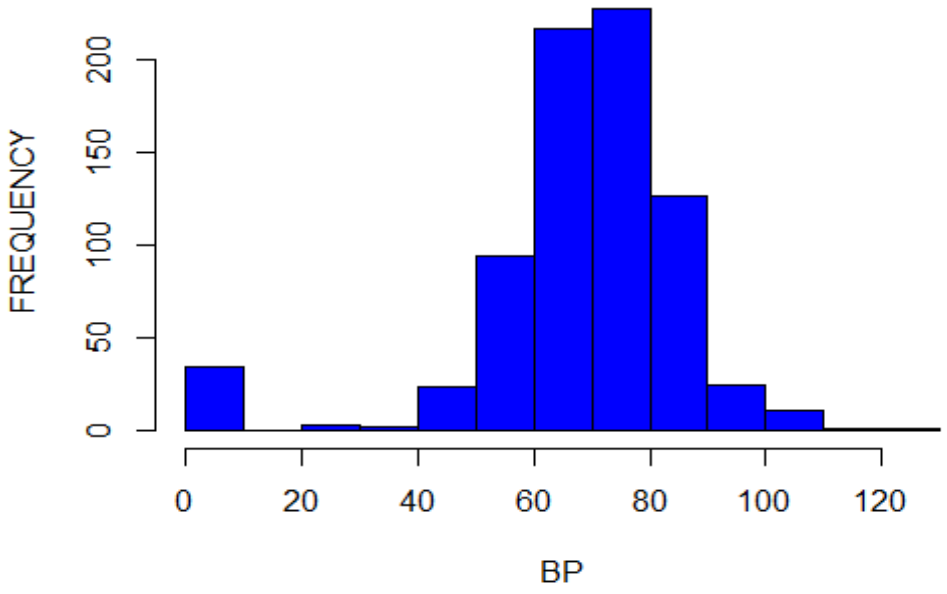
Exploratory Data Analysis:

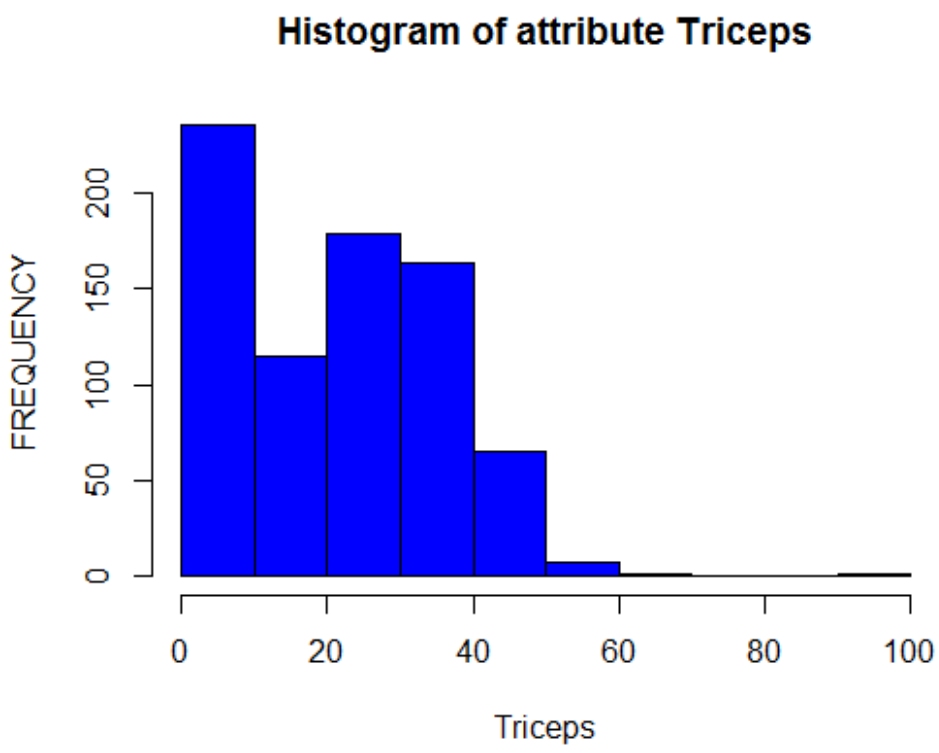
1. Create the following plots: histogram, and barplot.

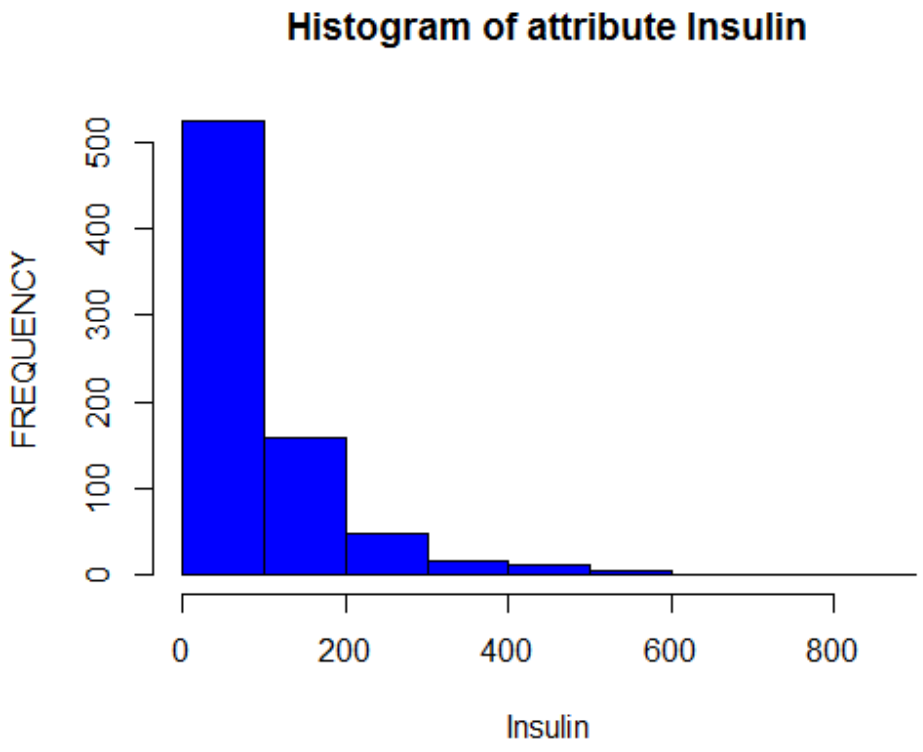
HISTOGRAM:

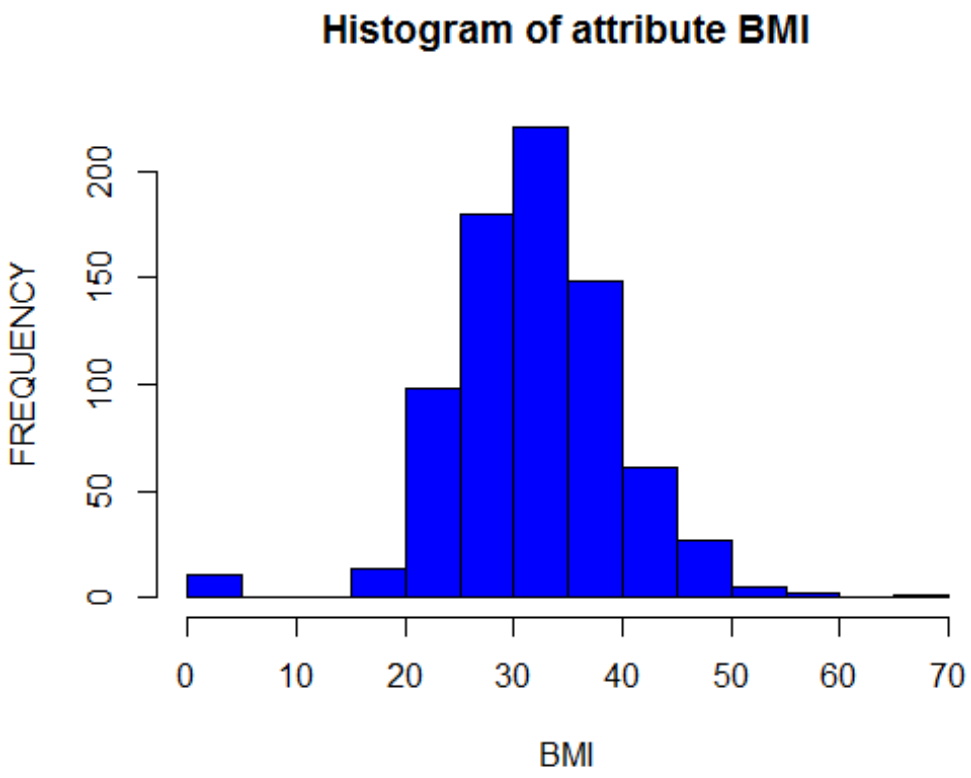


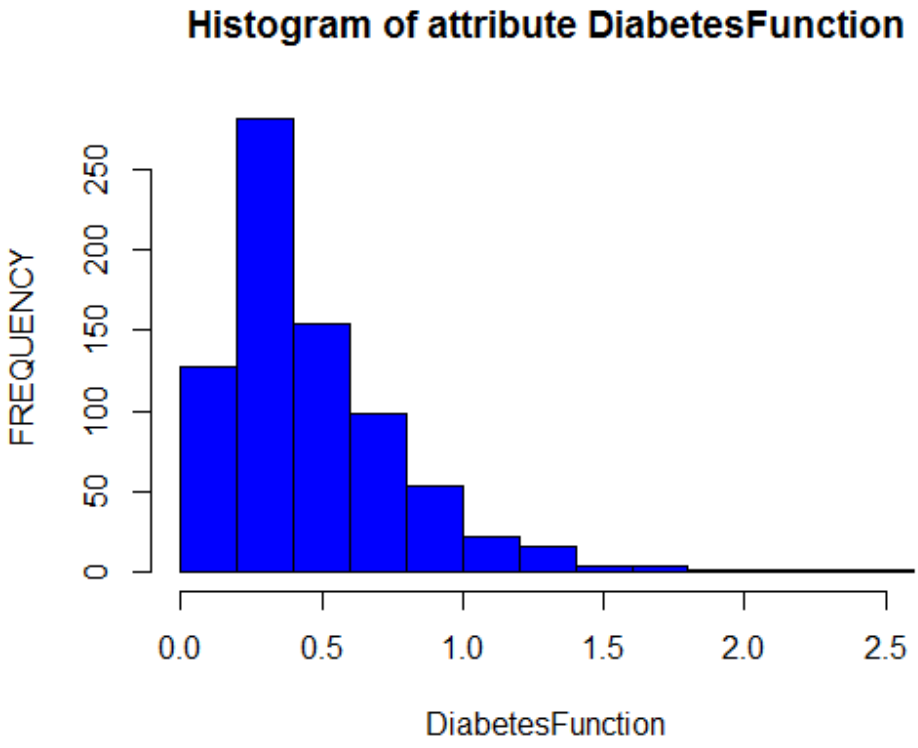

Histogram of attribute PlasmaGlucose
FREQUENCY
0
50
100
150
200
0
50
100
150
200
PlasmaGlucose

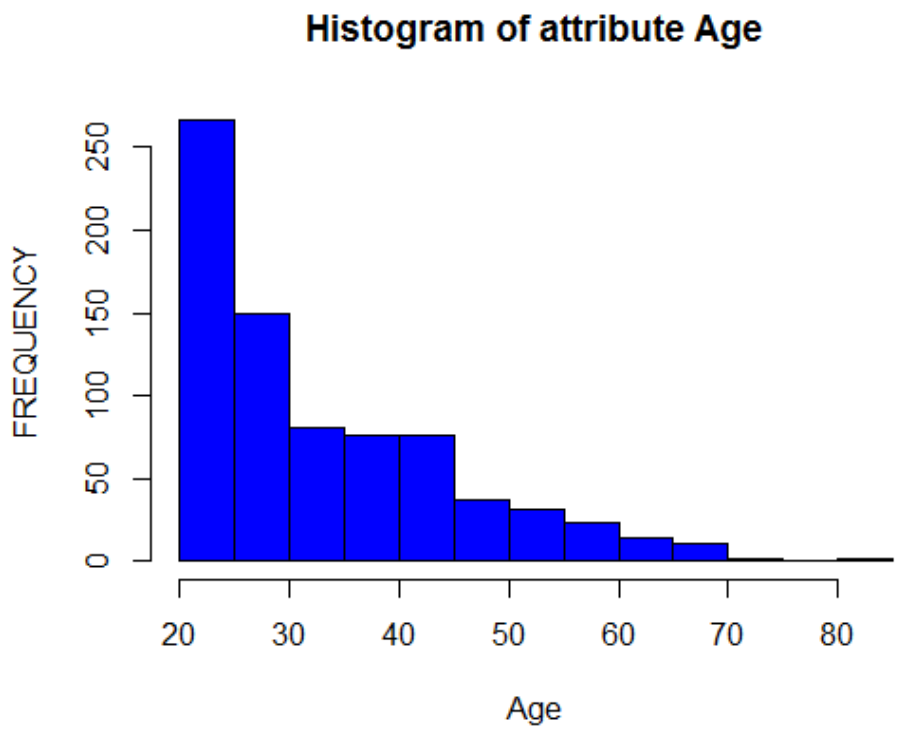

Histogram of attribute BP
FREQUENCY
0
50
100
150
200
0
20
40
60
80
100
120
BP


Histogram of attribute Triceps
FREQUENCY
0
50
100
150
200
0
20
40
60
80
100
Triceps

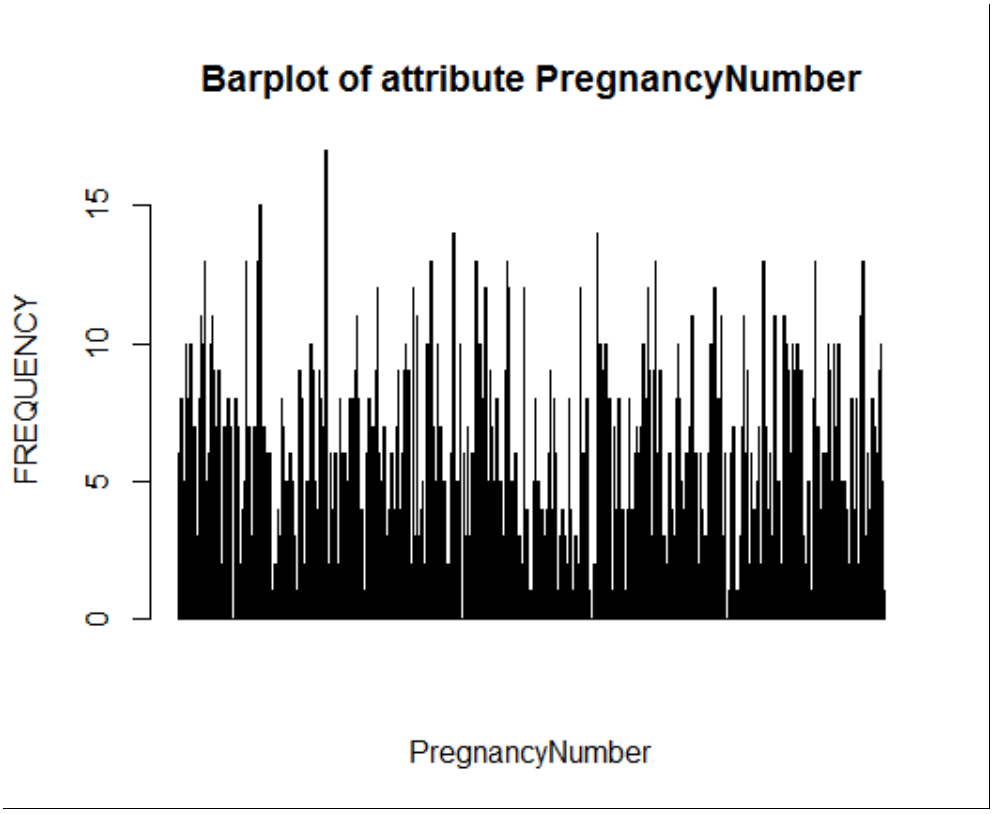

Histogram of attribute Insulin
FREQUENCY
0
100
200
300
400
500
0
200
400
600
800
Insulin

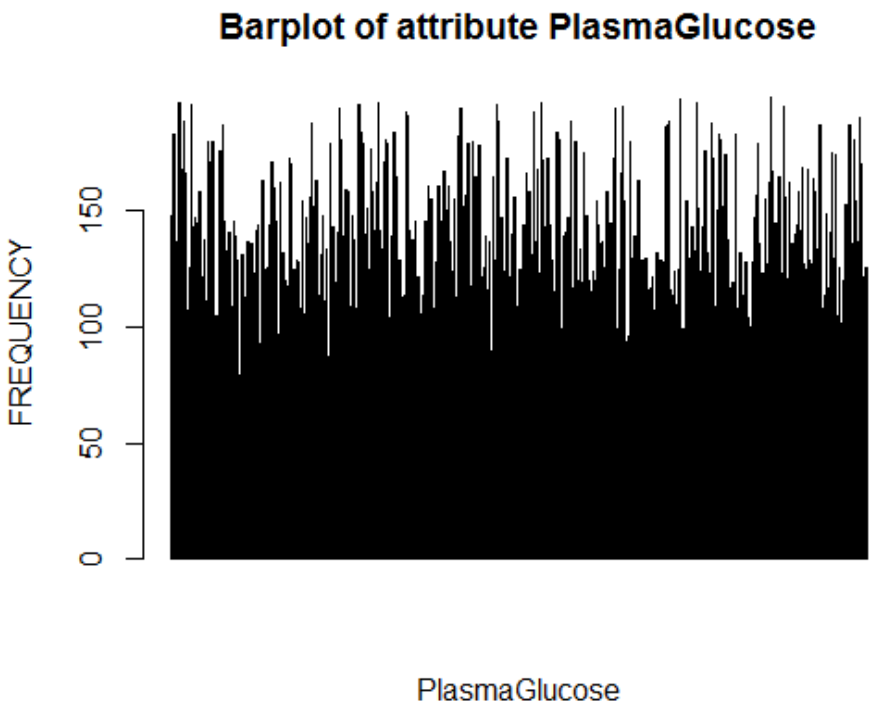

Histogram of attribute BMI
FREQUENCY
0
50
100
150
200
0
10
20
30
40
50
60
70
BMI

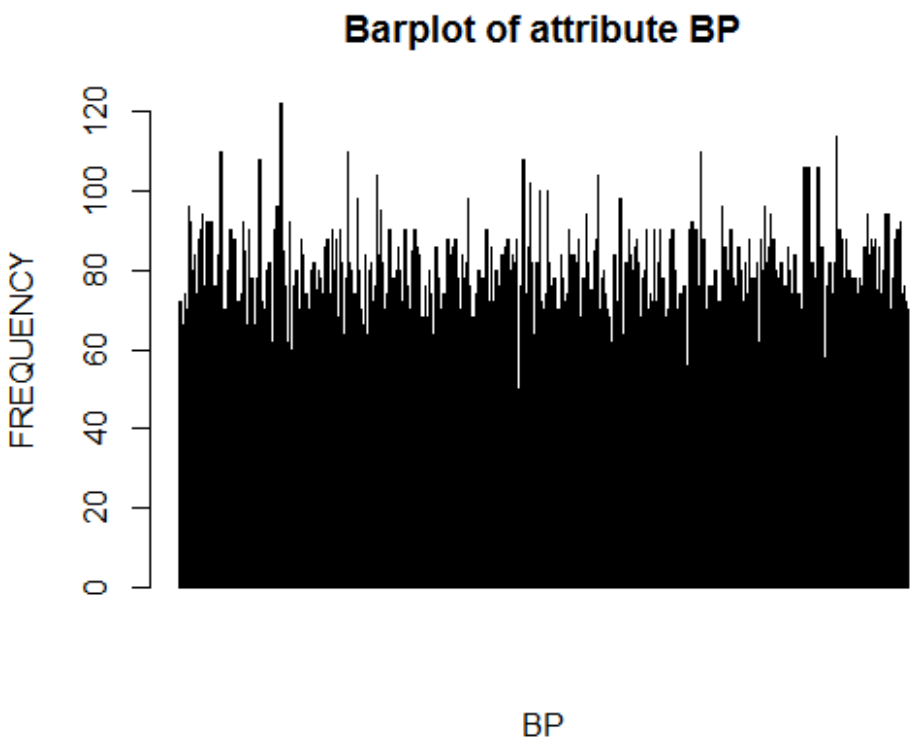

Histogram of attribute DiabetesFunction
FREQUENCY
0
50
100
150
200
250
0.0
0.5
1.0
1.5
2.0
2.5
DiabetesFunction

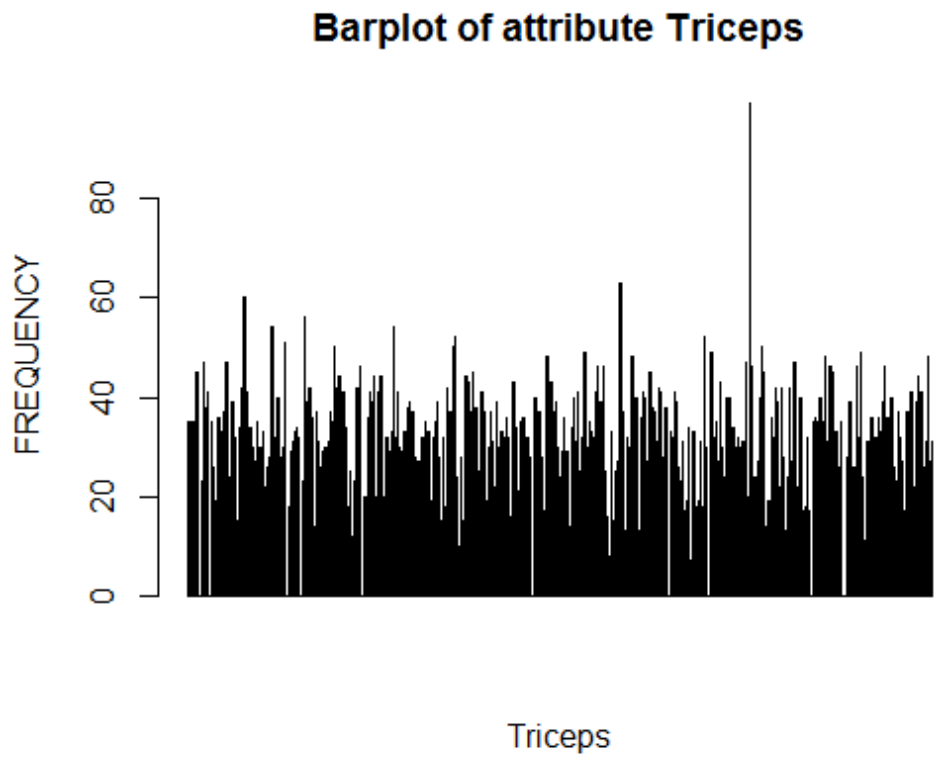

Histogram of attribute Age
FREQUENCY
0
50
100
150
200
250
20
30
40
50
60
70
80
Age

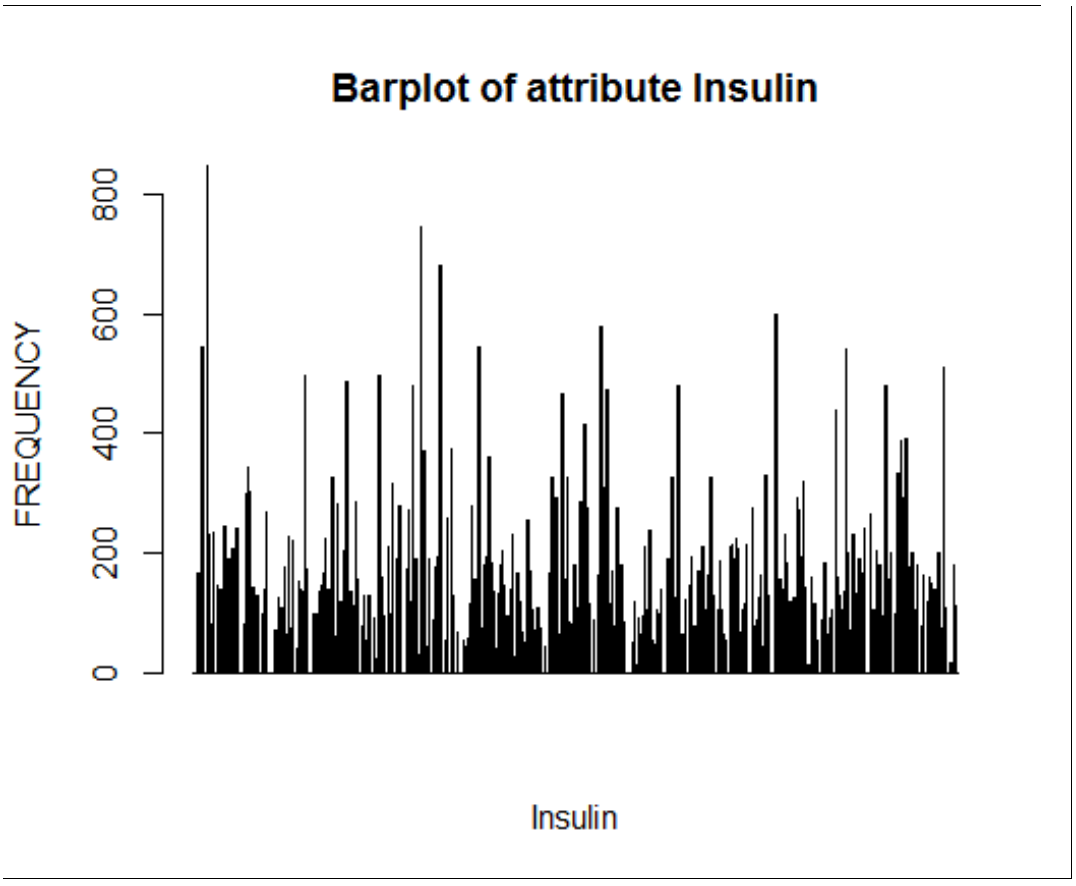
BAR PLOTS:

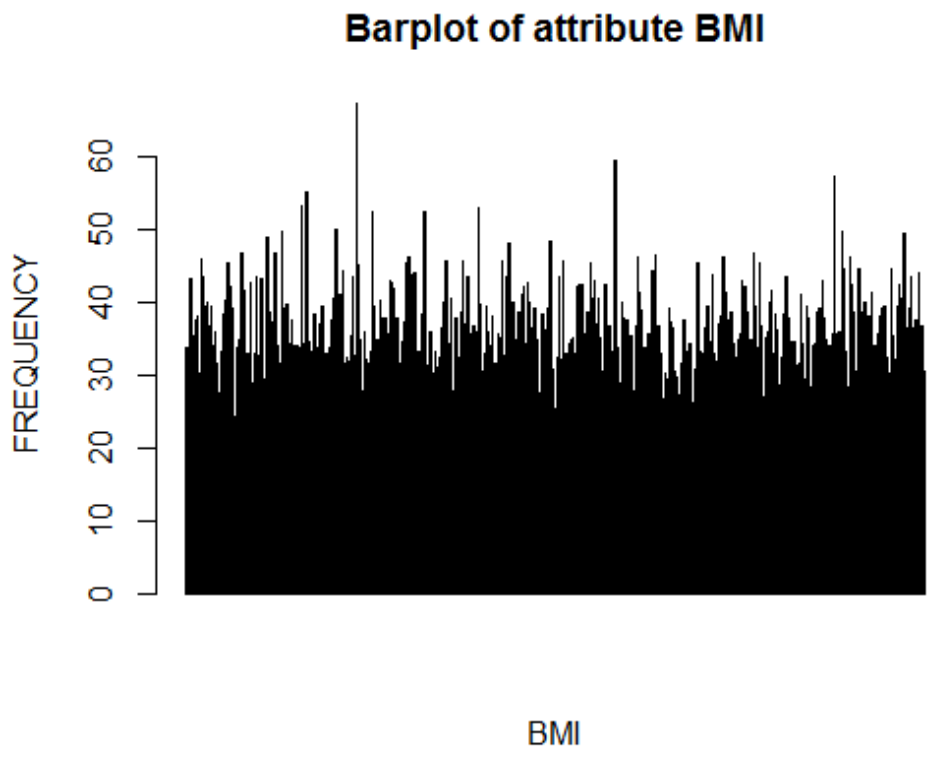

Barplot of attribute PregnancyNumber
FREQUENCY
0
5
10
15
PregnancyNumber

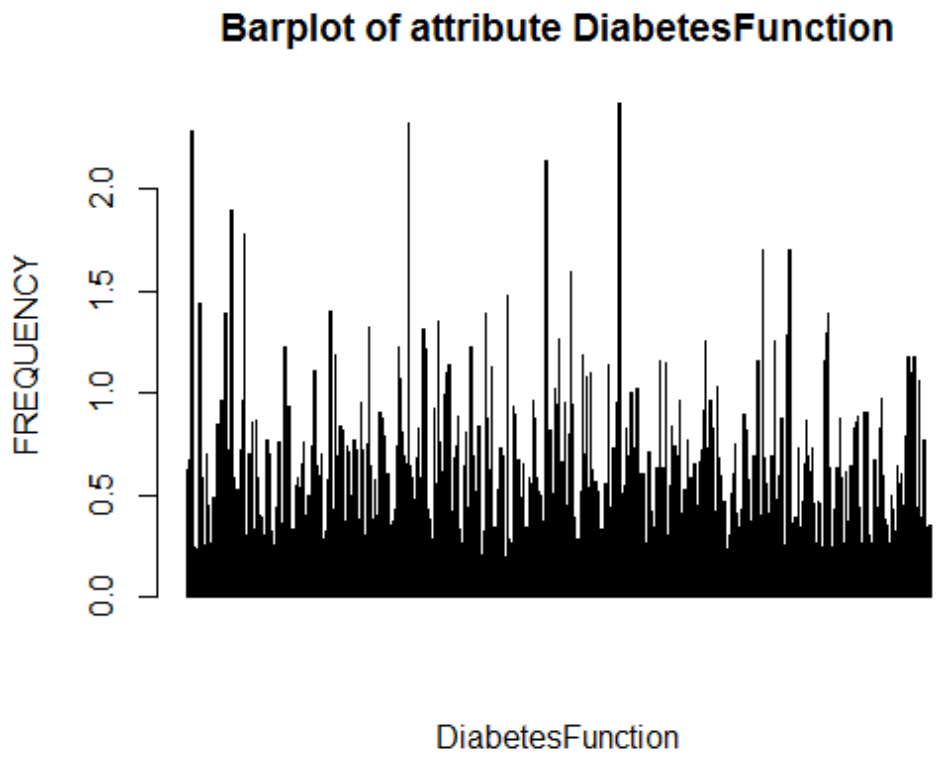

Barplot of attribute PlasmaGlucose
FREQUENCY
0
50
100
150
PlasmaGlucose

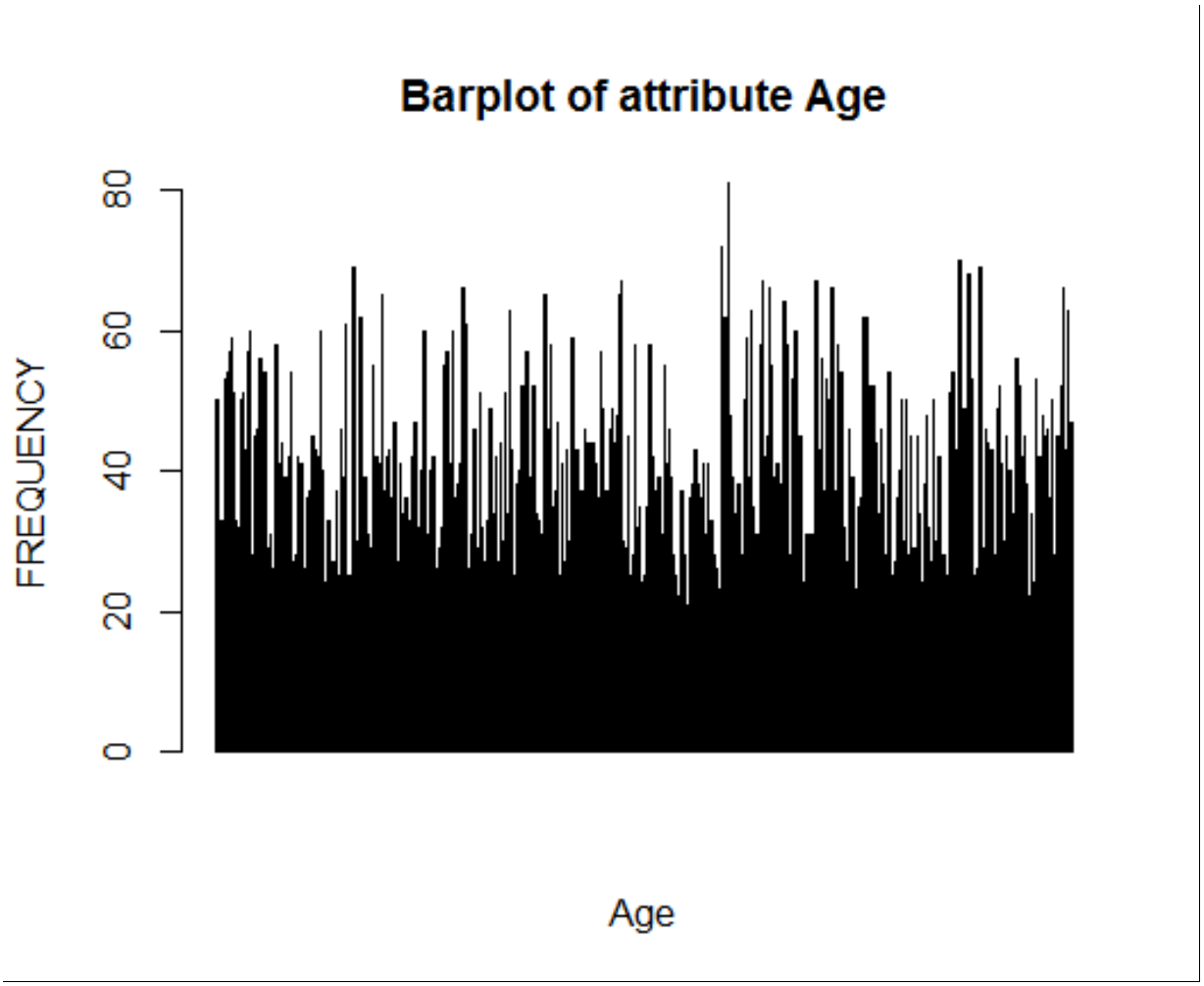

Barplot of attribute BP
FREQUENCY
0
20
40
60
80
100
120
BP


Barplot of attribute Triceps
FREQUENCY
0
20
40
60
80
Triceps


Barplot of attribute Insulin
FREQUENCY
0
200
400
600
800
Insulin


Barplot of attribute BMI
FREQUENCY
0
10
20
30
40
50
60
BMI


Barplot of attribute DiabetesFunction
FREQUENCY
0.0
0.5
1.0
1.5
2.0
DiabetesFunction



Write a short note on the distribution of the variables that you observe from the plots. Are they normally distributed?

Distribution of attributes BMI(Body Mass Index) and Plasma Glucose is normal. Also attribute BP has normal distribution with some outliers present in dataset. While distribution of other attributes Pregnancy Number, Triceps, Insulin, Diabetes Function and Age are uneven.

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

```
Correlation of attributes 1 PregnancyNumber with class is:  0.2218982
Correlation of attributes 2 PlasmaGlucose with class is:  0.4665814
Correlation of attributes 3 BP with class is:  0.06506836
Correlation of attributes 4 Triceps with class is:  0.07475223
Correlation of attributes 5 Insulin with class is:  0.130548
Correlation of attributes 6 BMI with class is:  0.2926947
Correlation of attributes 7 DiabetesFunction with class is:  0.1738441
Correlation of attributes 8 Age with class is:  0.238356
```

Which attributes seem to have a strong correlation with the output (class) variable?

Below command gives us maximum correlation with attribute:

```
> cat("\nmaximim correlation is :",max_val, "for attribute", names(data1[inde
x]))

maximim correlation is : 0.4665814 for attribute PlasmaGlucose
```

From, above result we can say that Plasma Glucose have a strong correlation with the output (class) variable.

3. Compute the correlation between all pairs of the 8 attributes.

Correlation of attributes 1 PregnancyNumber and 2 PlasmaGlucose is: 0.1294587
Correlation of attributes 1 PregnancyNumber and 3 BP is: 0.141282
Correlation of attributes 1 PregnancyNumber and 4 Triceps is: -0.08167177
Correlation of attributes 1 PregnancyNumber and 5 Insulin is: -0.07353461
Correlation of attributes 1 PregnancyNumber and 6 BMI is: 0.01768309
Correlation of attributes 1 PregnancyNumber and 7 DiabetesFunction is: -0.03352267
Correlation of attributes 1 PregnancyNumber and 8 Age is: 0.5443412
Correlation of attributes 2 PlasmaGlucose and 3 BP is: 0.1525896
Correlation of attributes 2 PlasmaGlucose and 4 Triceps is: 0.05732789
Correlation of attributes 2 PlasmaGlucose and 5 Insulin is: 0.3313571
Correlation of attributes 2 PlasmaGlucose and 6 BMI is: 0.2210711
Correlation of attributes 2 PlasmaGlucose and 7 DiabetesFunction is: 0.1373373
Correlation of attributes 2 PlasmaGlucose and 8 Age is: 0.2635143
Correlation of attributes 3 BP and 4 Triceps is: 0.2073705
Correlation of attributes 3 BP and 5 Insulin is: 0.08893338
Correlation of attributes 3 BP and 6 BMI is: 0.2818053
Correlation of attributes 3 BP and 7 DiabetesFunction is: 0.04126495
Correlation of attributes 3 BP and 8 Age is: 0.2395279
Correlation of attributes 4 Triceps and 5 Insulin is: 0.4367826
Correlation of attributes 4 Triceps and 6 BMI is: 0.3925732
Correlation of attributes 4 Triceps and 7 DiabetesFunction is: 0.1839276
Correlation of attributes 4 Triceps and 8 Age is: -0.1139703
Correlation of attributes 5 Insulin and 6 BMI is: 0.1978591
Correlation of attributes 5 Insulin and 7 DiabetesFunction is: 0.1850709
Correlation of attributes 5 Insulin and 8 Age is: -0.04216295
Correlation of attributes 6 BMI and 7 DiabetesFunction is: 0.140647
Correlation of attributes 6 BMI and 8 Age is: 0.03624187
Correlation of attributes 7 DiabetesFunction and 8 Age is: 0.03356131

Which two attributes have the highest mutual correlation?

```
> cat("\nmaximim correlation is :",max_val2, "between ", names(data1[index_1]
),"and ", names(data1[index_2]))

maximim correlation is : 0.5443412 between  PregnancyNumber and  Age
```

From above command, we can say that Pregnancy Number and Age have the highest mutual correlation.