Term Project

Group 1

1/25/2022

```
library(knitr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(mlbench)
library(caret)
```

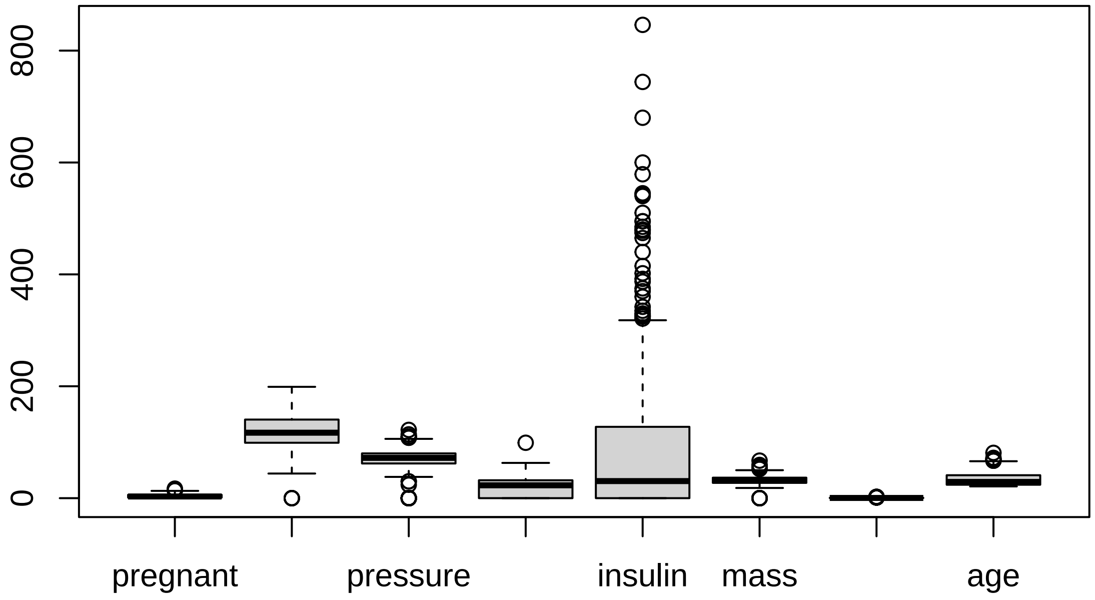
```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

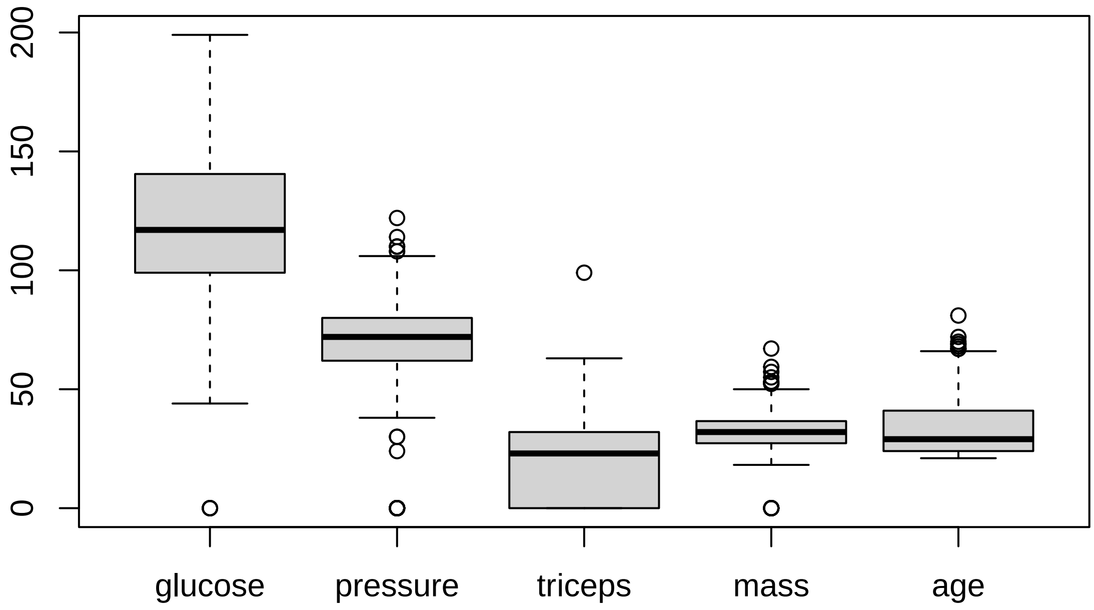
```
library(e1071)
library(ggplot2)

data(PimaIndiansDiabetes)

summary(PimaIndiansDiabetes)
```

```
##     pregnant         glucose         pressure         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           mass          pedigree           age        diabetes
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   neg:500
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   pos:268
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

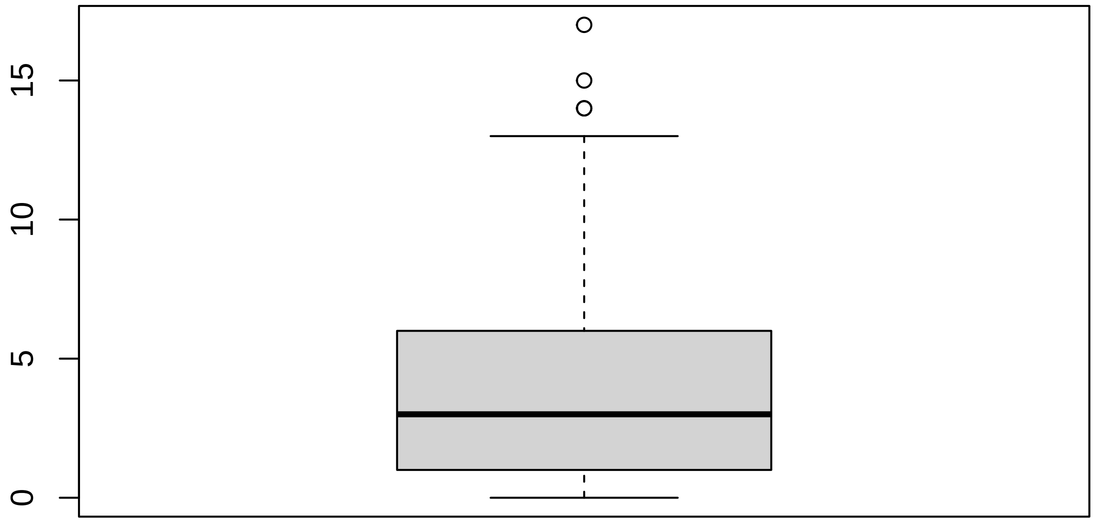
```
predictors <- PimaIndiansDiabetes[ , -(9)]

boxplot(predictors)
```



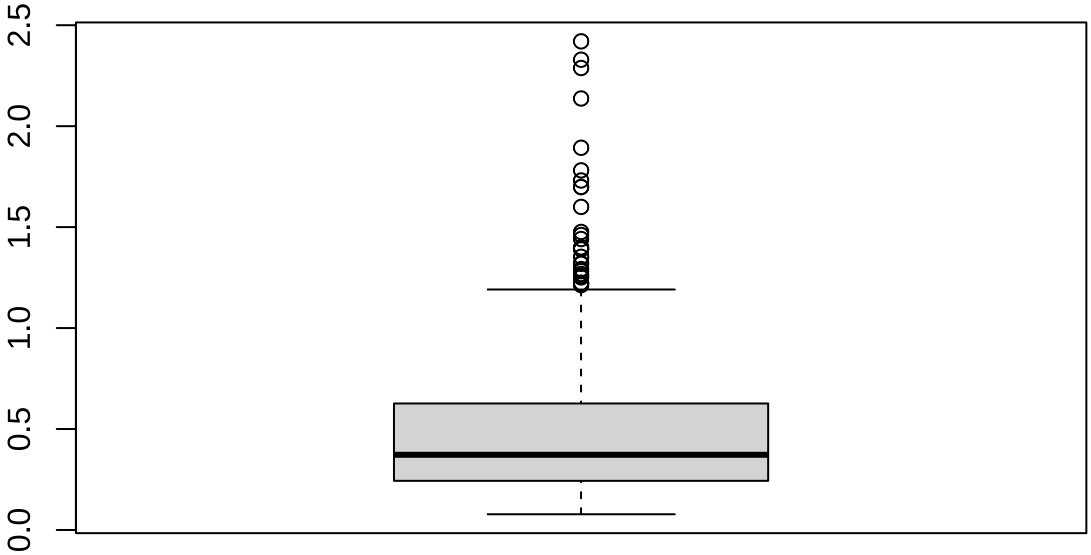
```
boxplot(predictors[,-c(5, 1, 7)]) # Glucose looks normal, Blood pressure normal but with outliers, skin
```



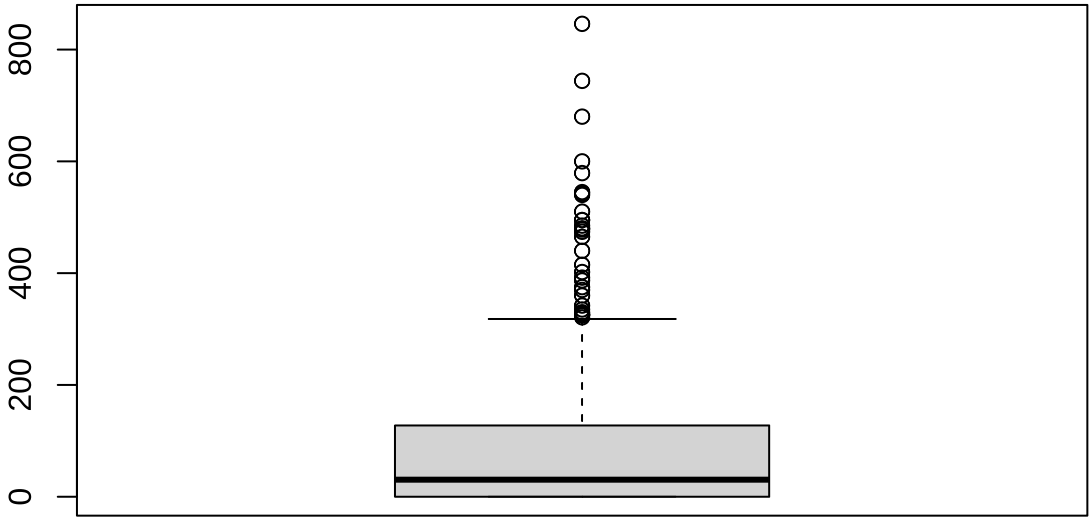
```
boxplot(predictors[,c(1)]) # Skewed positive
```



```
boxplot(predictors[,(7)]) # Heavily Skewed positive
```



```
boxplot(predictors[,(5)]) # Heavily Skewed positive
```



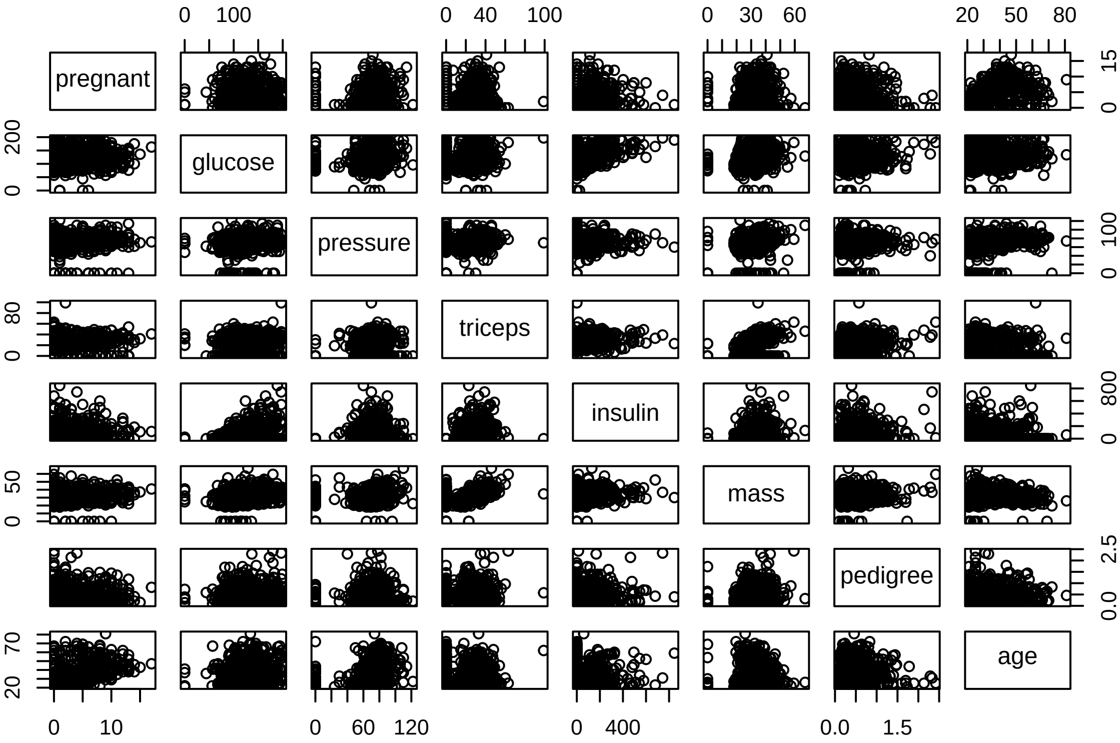
Just for visual review right now. Numerical analysis of skewness and outliers below.

```
# no near zero variance predictors
print(nearZeroVar(predictors))
```

```
## integer(0)
```

No near Zero predictors… clear from the visual inspection but good to have a mathematical confirmation.

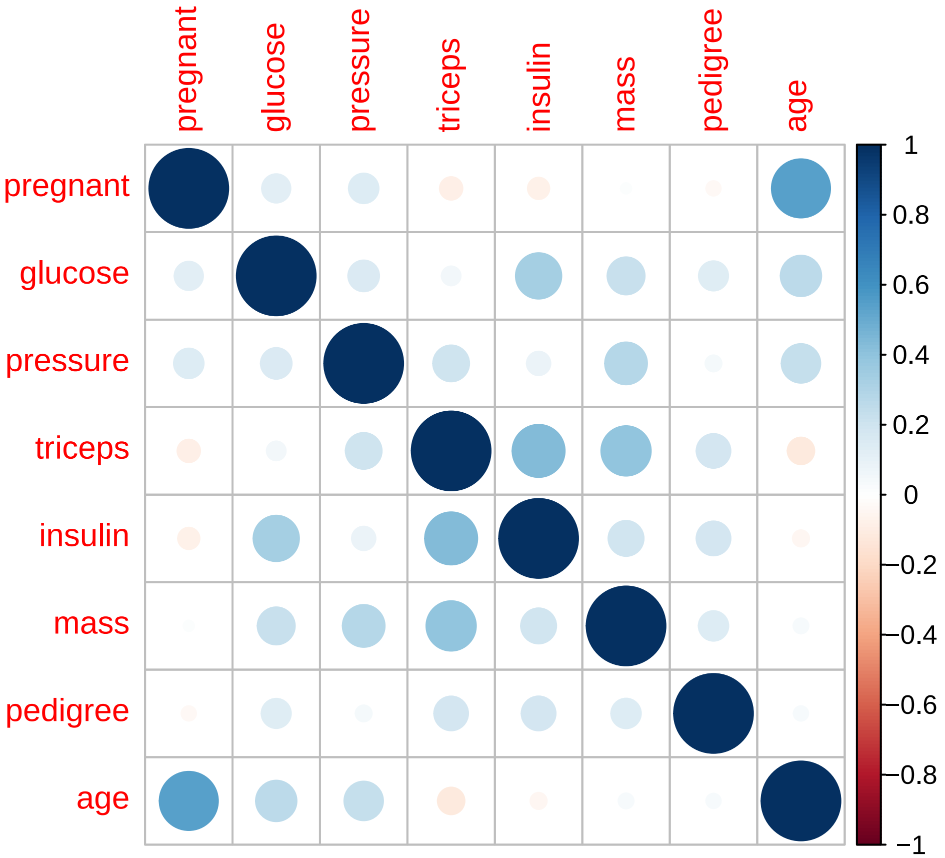
```
pairs(predictors)
```



```
cor( predictors )
```

```
##             pregnant    glucose   pressure     triceps     insulin       mass
## pregnant  1.00000000 0.12945867 0.14128198 -0.08167177 -0.07353461 0.01768309
## glucose   0.12945867 1.00000000 0.15258959  0.05732789  0.33135711 0.22107107
## pressure  0.14128198 0.15258959 1.00000000  0.20737054  0.08893338 0.28180529
## triceps  -0.08167177 0.05732789 0.20737054  1.00000000  0.43678257 0.39257320
## insulin  -0.07353461 0.33135711 0.08893338  0.43678257  1.00000000 0.19785906
## mass      0.01768309 0.22107107 0.28180529  0.39257320  0.19785906 1.00000000
## pedigree -0.03352267 0.13733730 0.04126495  0.18392757  0.18507093 0.14064695
## age       0.54434123 0.26351432 0.23952795 -0.11397026 -0.04216295 0.03624187
##             pedigree         age
## pregnant -0.03352267  0.54434123
## glucose   0.13733730  0.26351432
## pressure  0.04126495  0.23952795
## triceps   0.18392757 -0.11397026
## insulin   0.18507093 -0.04216295
## mass      0.14064695  0.03624187
## pedigree  1.00000000  0.03356131
## age       0.03356131  1.00000000
```

```
# Use the "corrplot" command:
corrplot( cor( predictors ))
```

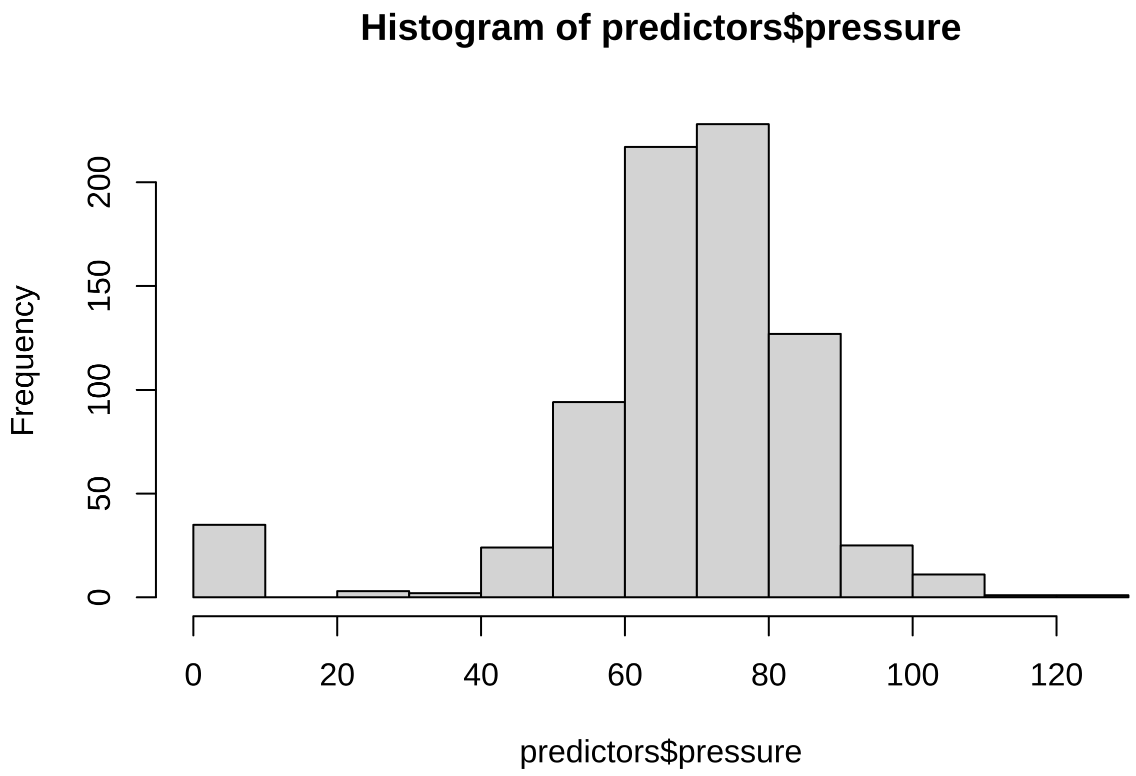


None of the predictors are significantly correlated. Age and Pregnancy are somewhat correlated as is to be expected.

```
Skewness <- apply( predictors, 2, skewness )

Outliers <- c()
SkewnessQ <- c()
for (i in 1:ncol(predictors)) {
  BoxPlot = boxplot(predictors[,i], plot=FALSE)
  if (length(BoxPlot$out) > 0) {
    Outliers = append(Outliers, "Yes")}
  else {
    Outliers = append(Outliers, "No")}
  if (abs(Skewness[i]) < .5) {
    SkewnessQ = append(SkewnessQ, "None")}
  else if (abs(Skewness[i]) >= .5 & (abs(Skewness[i]) < 1)){
    SkewnessQ = append(SkewnessQ, "Moderate")}
  else {
    SkewnessQ = append(SkewnessQ, "High")
  }

}

characteristics = data.frame(Skewness, SkewnessQ, Outliers)

kable(characteristics, format = "markdown", col.names = c("Skewness", "Skewness Level", "Contains Outli
```

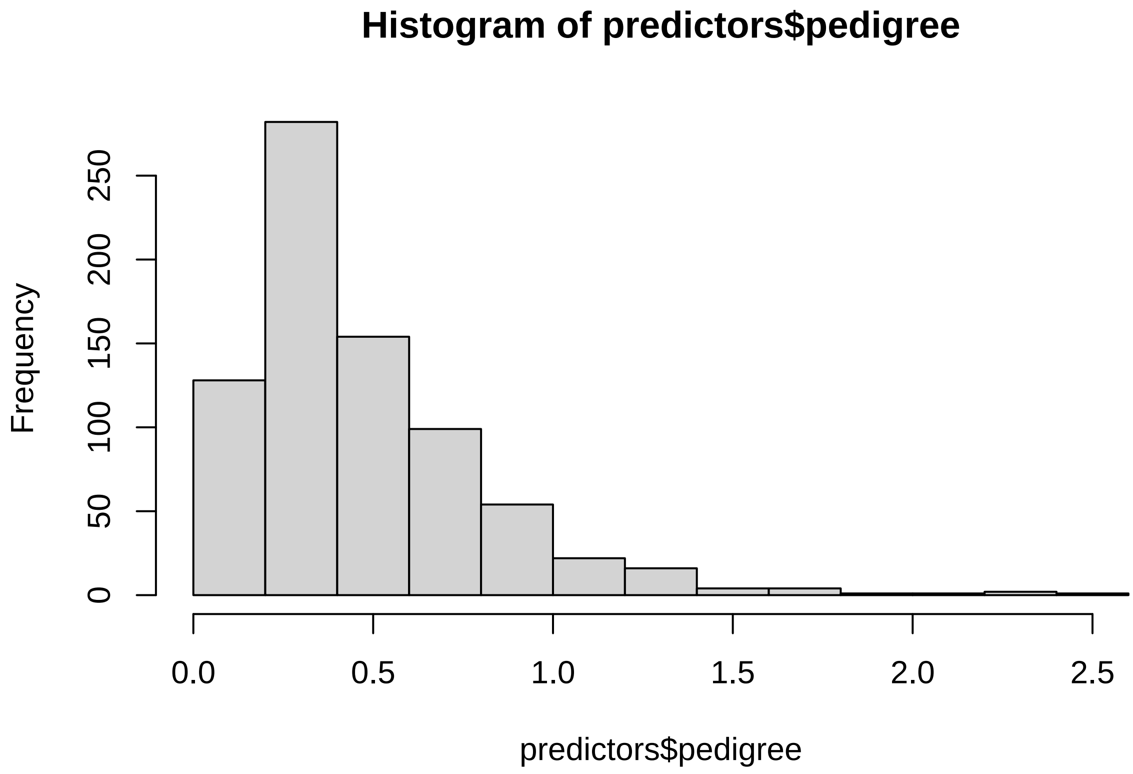
	Skewness	Skewness Level	Contains Outliers
pregnant	0.8981549	Moderate	Yes
glucose	0.1730754	None	Yes
pressure	-1.8364126	High	Yes
triceps	0.1089456	None	Yes
insulin	2.2633826	High	Yes
mass	-0.4273073	None	Yes
pedigree	1.9124179	High	Yes
age	1.1251880	High	Yes

```
hist(predictors$pressure)
```

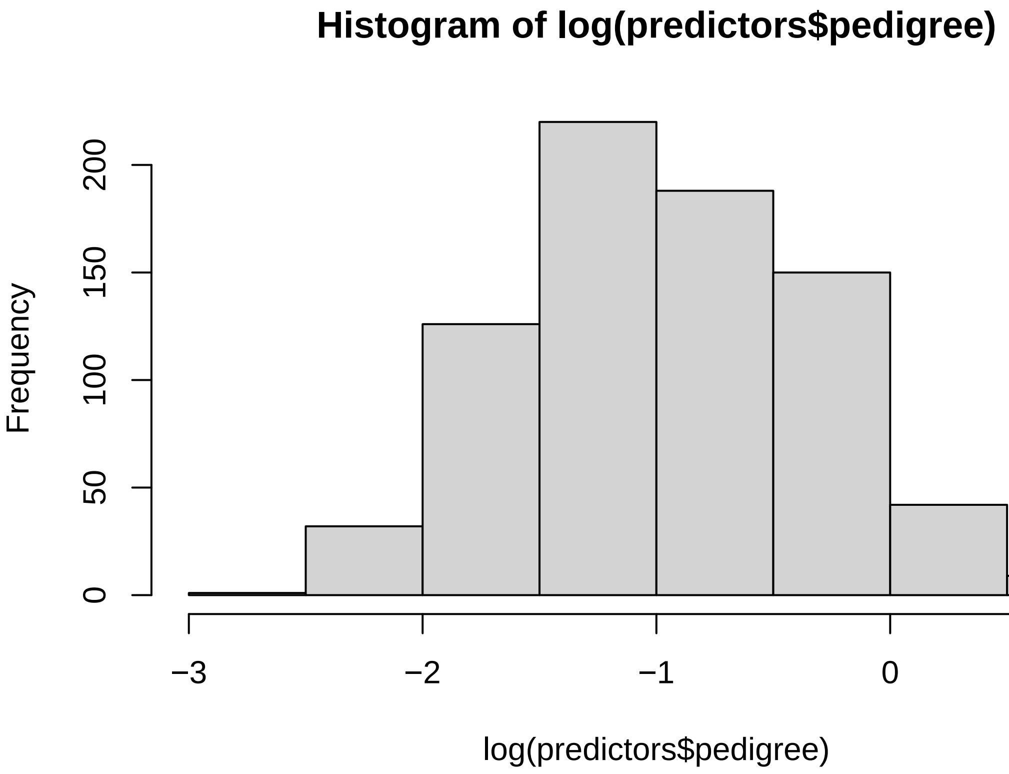


Looks like if we took care of the 0 values this would be a pretty normal distribution

```
hist(predictors$pedigree)
```

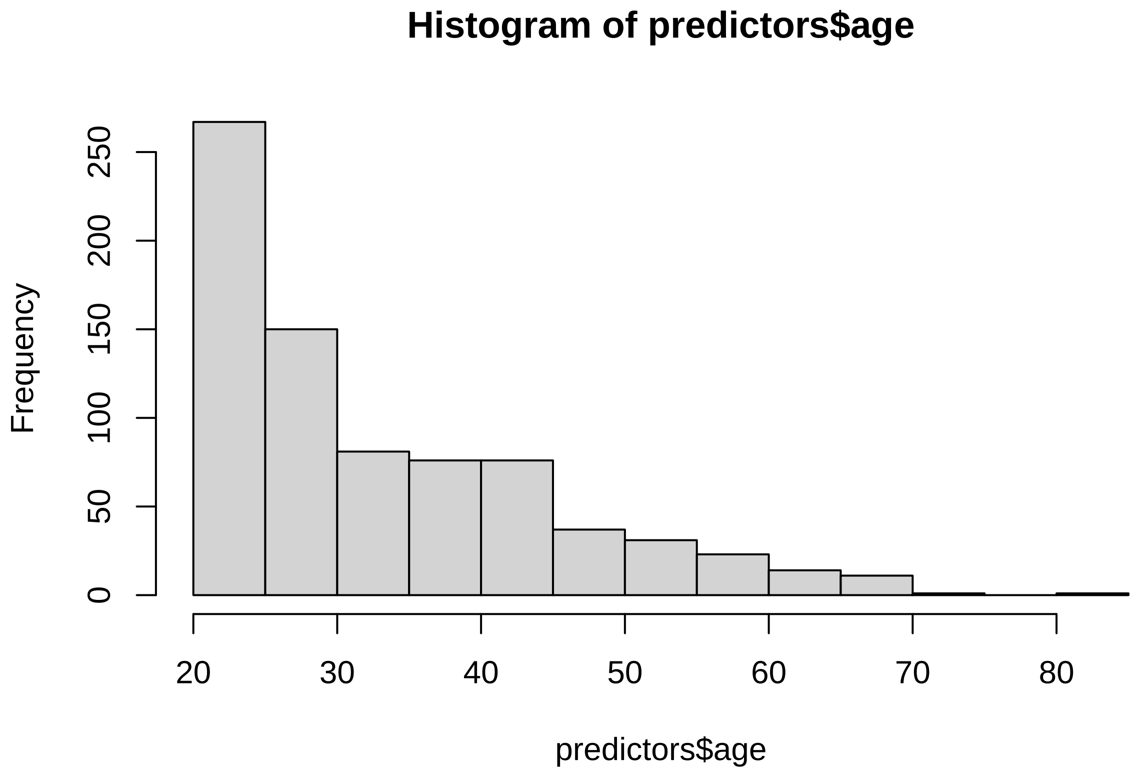


```
hist(log(predictors$pedigree))
```



Looks like taking the log of this would make a normal distribution

```
hist(predictors$age)
```



Not sure what transformation can make this more normal…

There are significant outliers on all the predictors and some are heavily skewed.

```
predictorPP <- preProcess(predictors, c("BoxCox", "center", "scale"))

predictorPP$method$BoxCox
```

```
## [1] "pedigree" "age"
```

```
predictorPP$bc$pedigree
```

```
## Box-Cox Transformation
##
## 768 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0780  0.2437  0.3725  0.4719  0.6262  2.4200
##
## Largest/Smallest: 31
## Sample Skewness: 1.91
##
## Estimated Lambda: -0.1
## With fudge factor, Lambda = 0 will be used for transformations
```

```
predictorPP$bc$age
```

```
## Box-Cox Transformation
```

```
## 
## 768 data points used to estimate Lambda
## 
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   21.00   24.00   29.00   33.24   41.00   81.00 
## 
## Largest/Smallest: 3.86 
## Sample Skewness: 1.13 
## 
## Estimated Lambda: -1.1
```

BoxCox results are difficult to interpret. I understand if I had one predictor that the lambda value is the power on the outcome but in this case we have multiple predictors and the outcome is categorical. Does that mean the lambda is the power of the predictor? I need more investigation.

```
Pimapca <- prcomp(predictors,center = TRUE, scale. = TRUE)

summary(Pimapca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4472 1.3158 1.0147 0.9357 0.87312 0.82621 0.64793
## Proportion of Variance 0.2618 0.2164 0.1287 0.1094 0.09529 0.08533 0.05248
## Cumulative Proportion  0.2618 0.4782 0.6069 0.7163 0.81164 0.89697 0.94944
##                            PC8
## Standard deviation     0.63597
## Proportion of Variance 0.05056
## Cumulative Proportion  1.00000
```

```
Pimapca
```

```
## Standard deviations (1, .., p=8):
## [1] 1.4471973 1.3157546 1.0147068 0.9356971 0.8731234 0.8262133 0.6479322
## [8] 0.6359733
## 
## Rotation (n x k) = (8 x 8):
##                 PC1        PC2         PC3         PC4        PC5          PC6
## pregnant -0.1284321  0.5937858 -0.01308692  0.08069115 -0.4756057  0.193598168
## glucose  -0.3930826  0.1740291  0.46792282 -0.40432871  0.4663280  0.094161756
## pressure -0.3600026  0.1838921 -0.53549442  0.05598649  0.3279531 -0.634115895
## triceps  -0.4398243 -0.3319653 -0.23767380  0.03797608 -0.4878621  0.009589438
## insulin  -0.4350262 -0.2507811  0.33670893 -0.34994376 -0.3469348 -0.270650609
## mass     -0.4519413 -0.1009598 -0.36186463  0.05364595  0.2532038  0.685372179
## pedigree -0.2706114 -0.1220690  0.43318905  0.83368010  0.1198105 -0.085784088
## age      -0.1980271  0.6205885  0.07524755  0.07120060 -0.1092900 -0.033357170
##                  PC7          PC8
## pregnant  0.58879003  0.117840984
## glucose   0.06015291  0.450355256
## pressure  0.19211793 -0.011295538
## triceps  -0.28221253  0.566283799
## insulin   0.13200992 -0.548621381
## mass      0.03536644 -0.341517637
## pedigree  0.08609107 -0.008258731
## age      -0.71208542 -0.211661979
```