Pima

Brian Burrows, Eric Sellew, Yonas Shiferaw, and Kelly Yang

12/3/2018

R Markdown

```
pima <- read.csv("https://pmatheson.people.amherst.edu/Pima.dat", header=FALSE) #Loading data
colnames(pima) <- c("PRG", "PLASMA", "BP", "THICK", "INSULIN", "BODY", "PEDIGREE", "AGE", "RESPONSE") #
```

Predict the probability that individual females have diabetes

```
tally(pima$RESPONSE)
```

```
## X
##   0   1
## 500 268
```

```
268/(500+268)
```

```
## [1] 0.3489583
```

```
#Approximately 35% of the individual females in this sample have diabetes
```

```
filteredpima <- filter(pima,PLASMA>0,BP>0,THICK>0,BODY>0,INSULIN>0) #creates a dataset that does not ha
```

```
#Create a stepwise model
model <- glm(RESPONSE~., data=pima, family=binomial) %>%
  MASS::stepAIC(trace=FALSE)
summary(model)
```

```
##
## Call:
## glm(formula = RESPONSE ~ PRG + PLASMA + BP + INSULIN + BODY +
##     PEDIGREE + AGE, family = binomial, data = pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5617  -0.7286  -0.4156   0.7271   2.9297
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.4051362  0.7167033 -11.727  < 2e-16 ***
## PRG          0.1231724  0.0320688   3.841 0.000123 ***
## PLASMA       0.0351123  0.0036625   9.587  < 2e-16 ***
## BP          -0.0132136  0.0051537  -2.564 0.010350 *
## INSULIN     -0.0011570  0.0008142  -1.421 0.155275
## BODY         0.0900886  0.0144619   6.229 4.68e-10 ***
## PEDIGREE     0.9475954  0.2980063   3.180 0.001474 **
## AGE          0.0147888  0.0092897   1.592 0.111393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 723.45  on 760  degrees of freedom
## AIC: 739.45
## 
## Number of Fisher Scoring iterations: 5
```

```
#Stepwise model using the filtered dataset
filteredmodel <- glm(RESPONSE~., data=filteredpima, family=binomial) %>%
  MASS::stepAIC(trace=FALSE)
summary(filteredmodel)
```

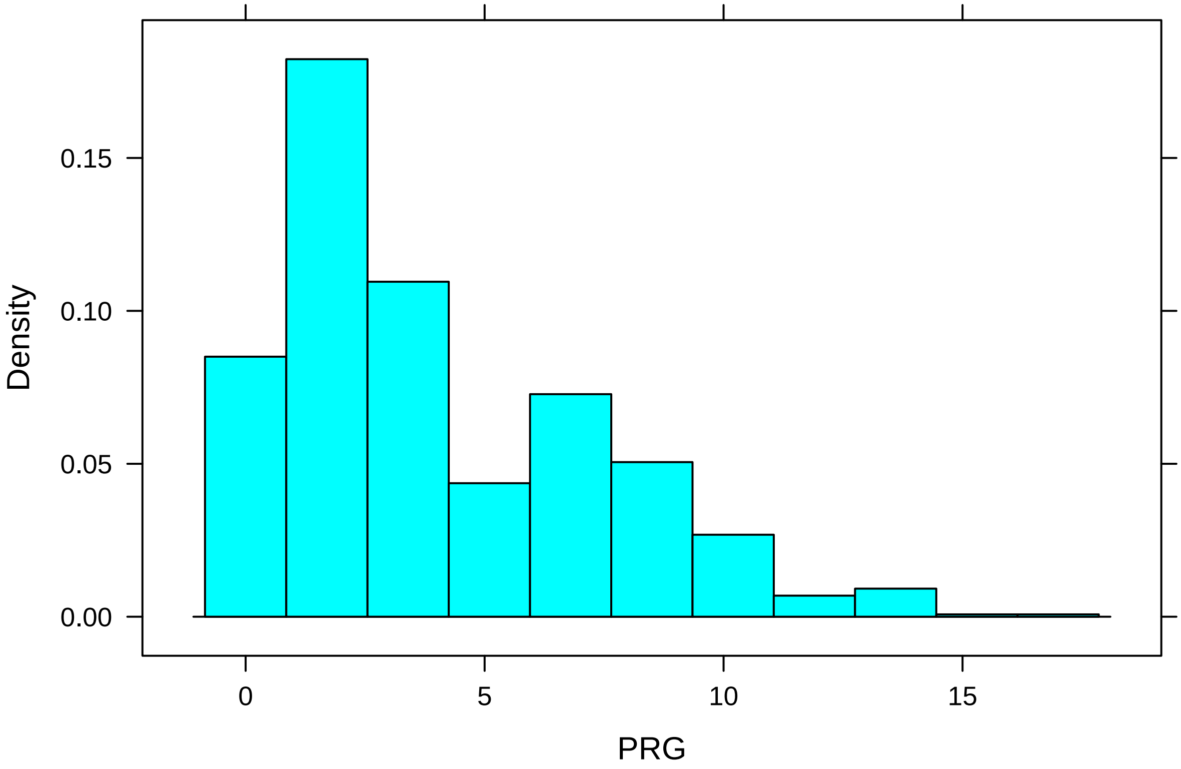
```
## 
## Call:
## glm(formula = RESPONSE ~ PRG + PLASMA + BODY + PEDIGREE + AGE, 
##     family = binomial, data = filteredpima)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.8827  -0.6535  -0.3694   0.6521   2.5814  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## PRG          0.083953   0.055031   1.526 0.127117    
## PLASMA       0.036458   0.004978   7.324 2.41e-13 ***
## BODY         0.078139   0.020605   3.792 0.000149 ***
## PEDIGREE     1.150913   0.424242   2.713 0.006670 ** 
## AGE          0.034360   0.017810   1.929 0.053692 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
## 
## Number of Fisher Scoring iterations: 5
```

EDA

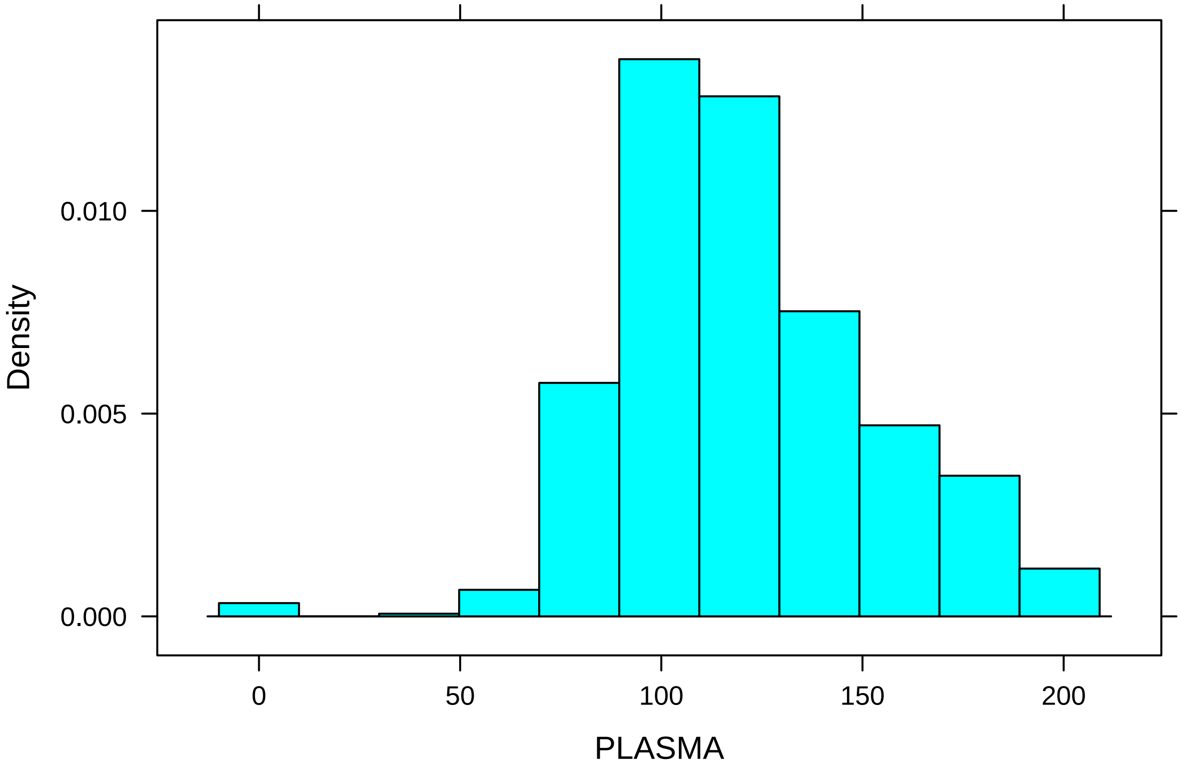
```
summary(pima)
```

```
##       PRG             PLASMA            BP             THICK      
##  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           BODY          PEDIGREE           AGE       
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
##  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
##     RESPONSE
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

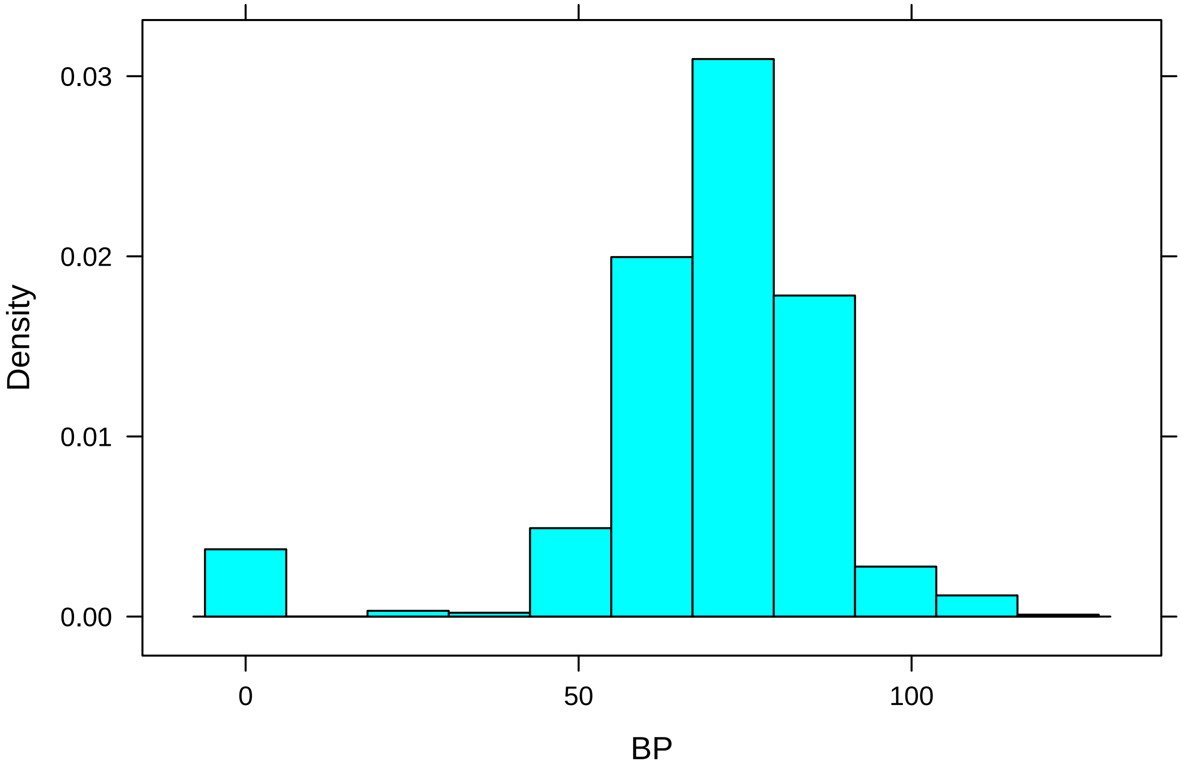
```
histogram(~PRG, data=pima) #PRG is skewed to the right
```



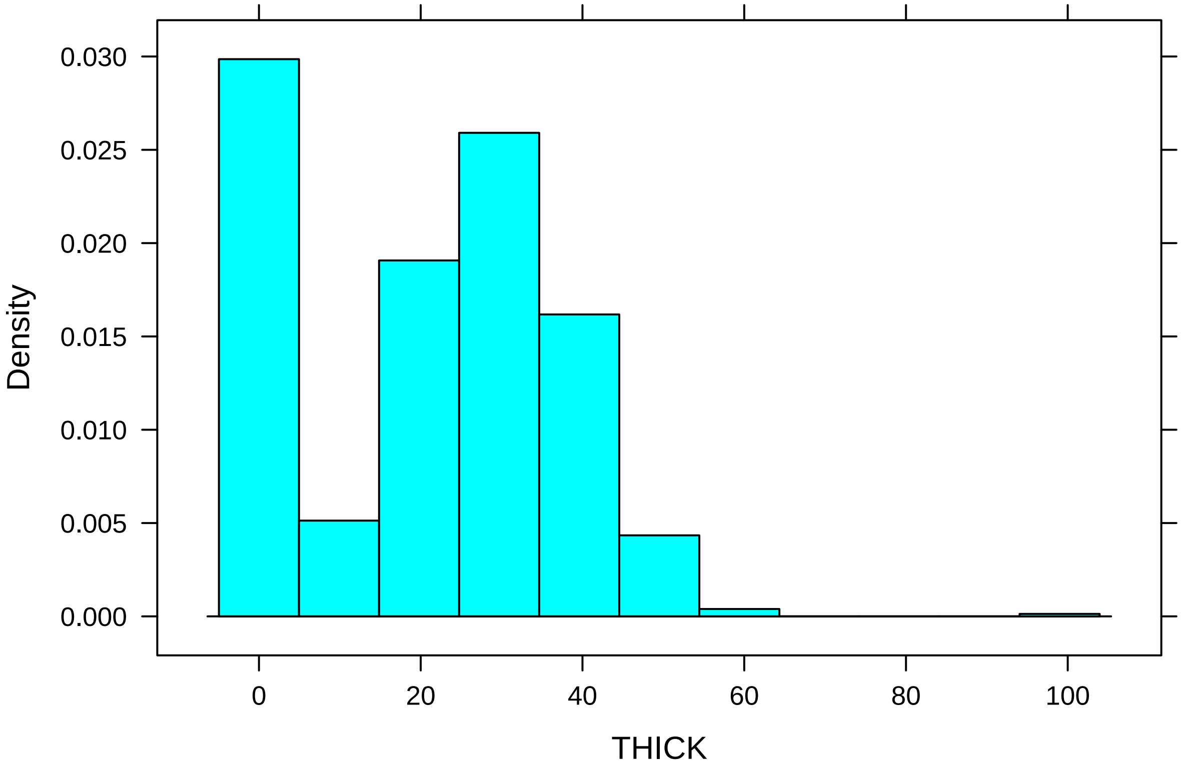
```
histogram(~PLASMA, data=pima) #rather bell-shaped
```



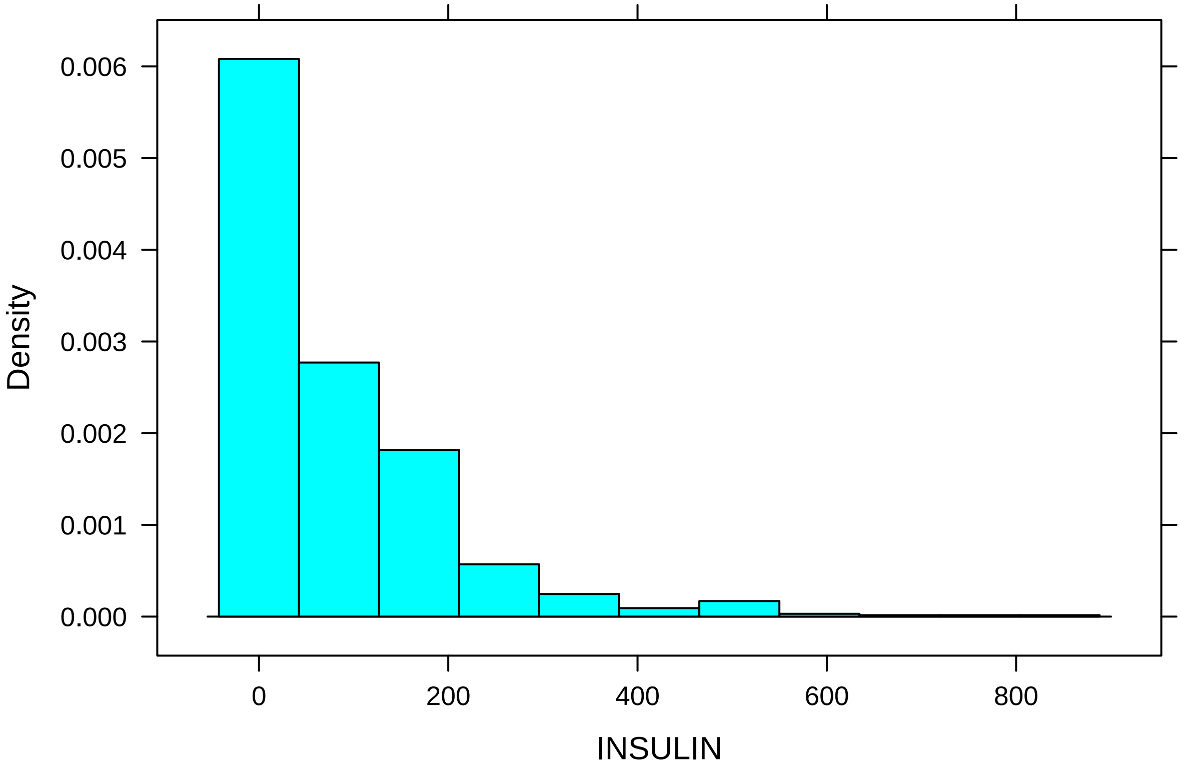
```
histogram(~BP, data=pima) #normal, but why are there so many 0s???
```



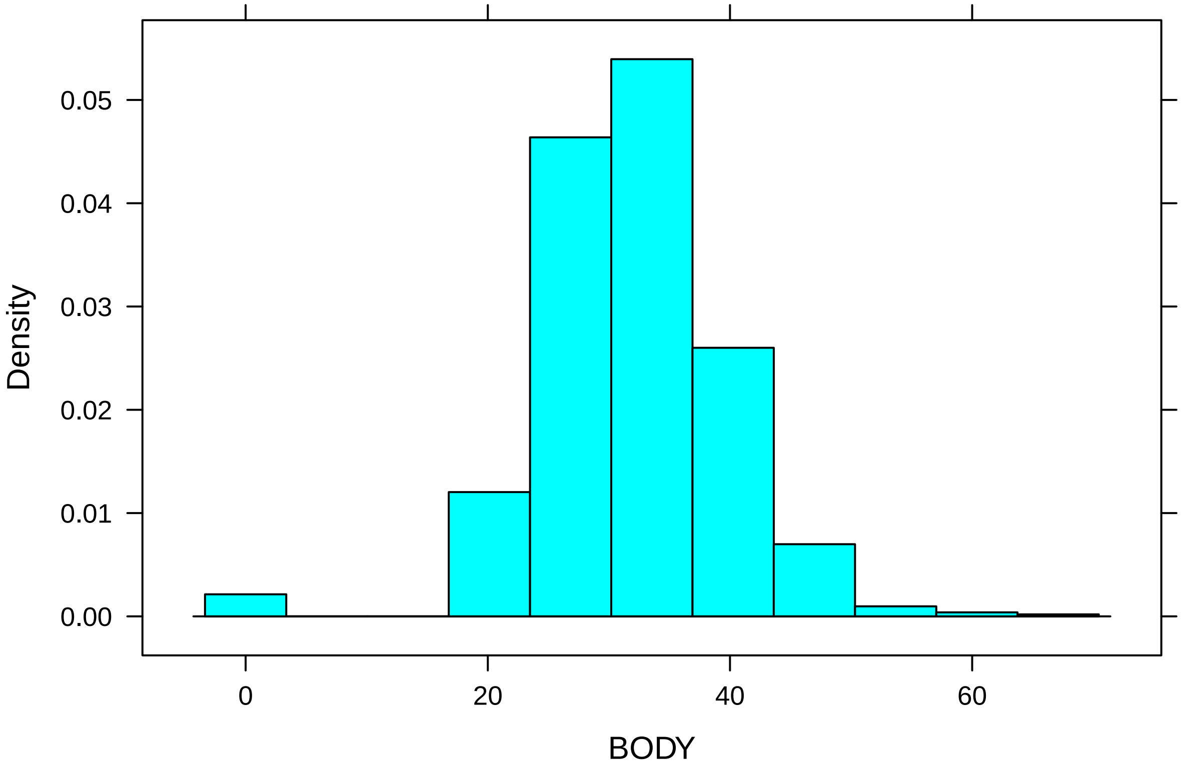
```
histogram(~THICK, data=pima) #many zeros, and one outlier at 99, but normal otherwise
```



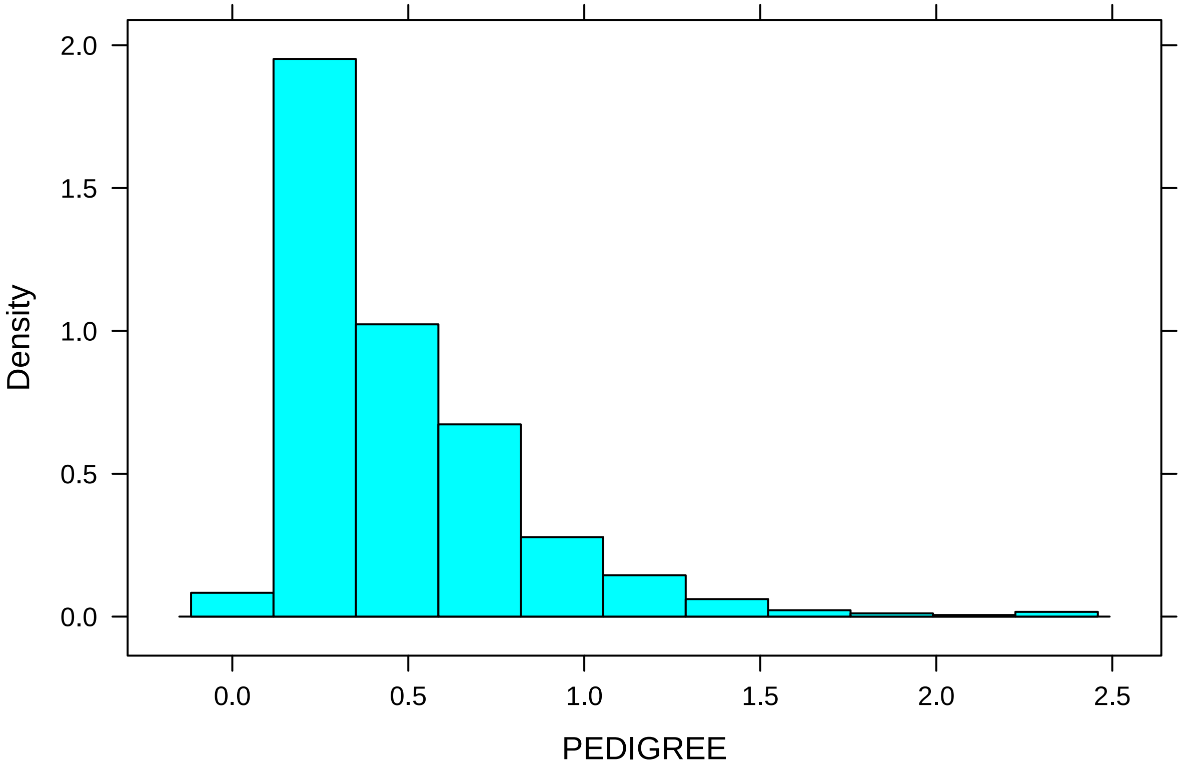
```
histogram(~INSULIN, data=pima) #skewed to the right
```



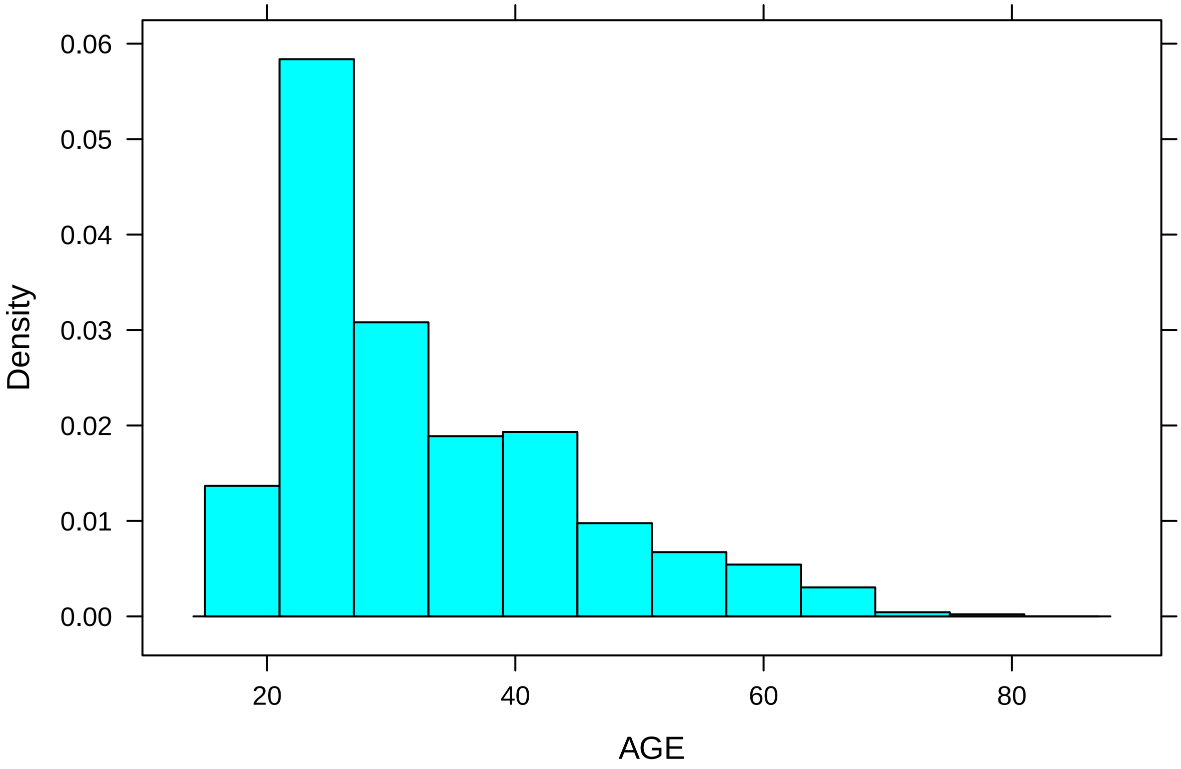
```
histogram(~BODY, data=pima) #kind of normal, slightly skewed to the right, many zeros
```



```
histogram(~PEDIGREE, data=pima) #skewed to the right
```



```
histogram(~AGE, data=pima) #skewed to the right
```

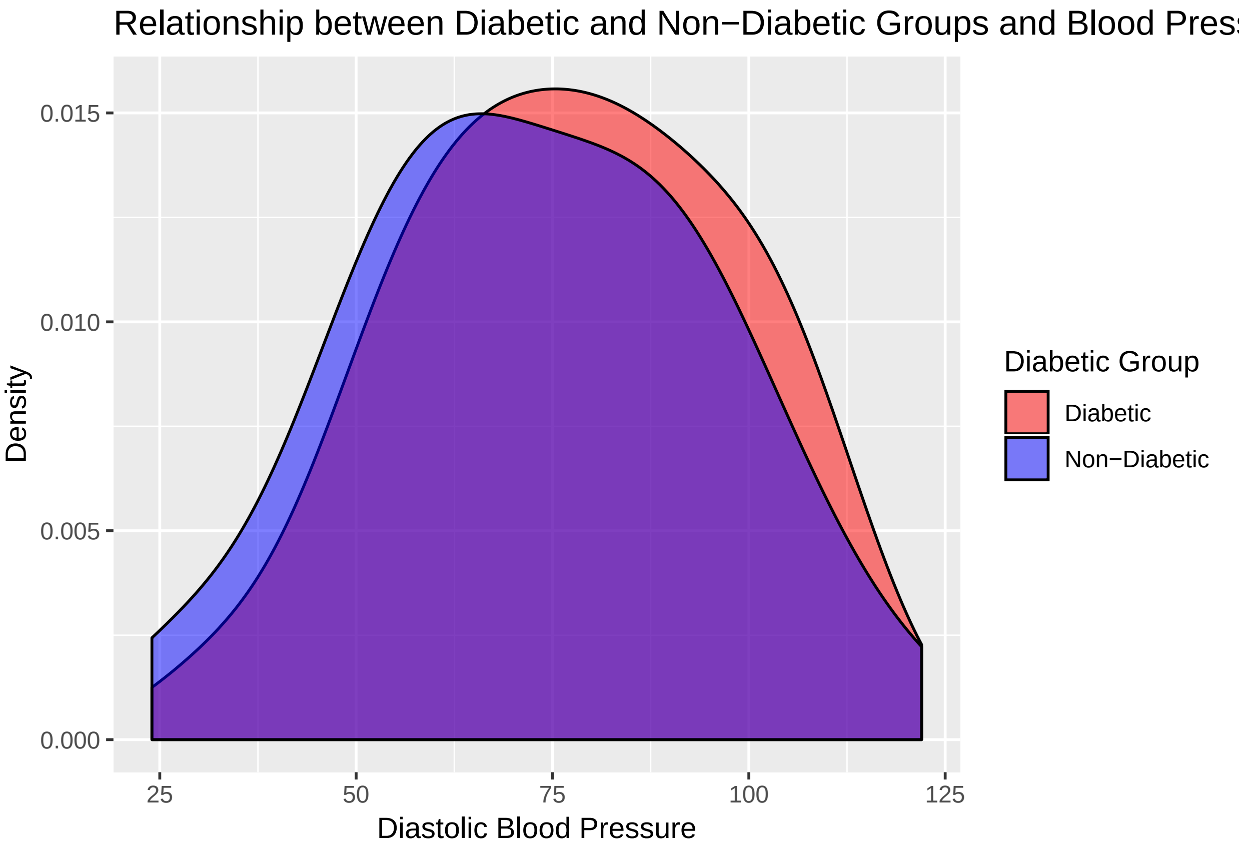


```
tally(~RESPONSE, data=pima) #500 no diabetes, 268 do have diabetes
```

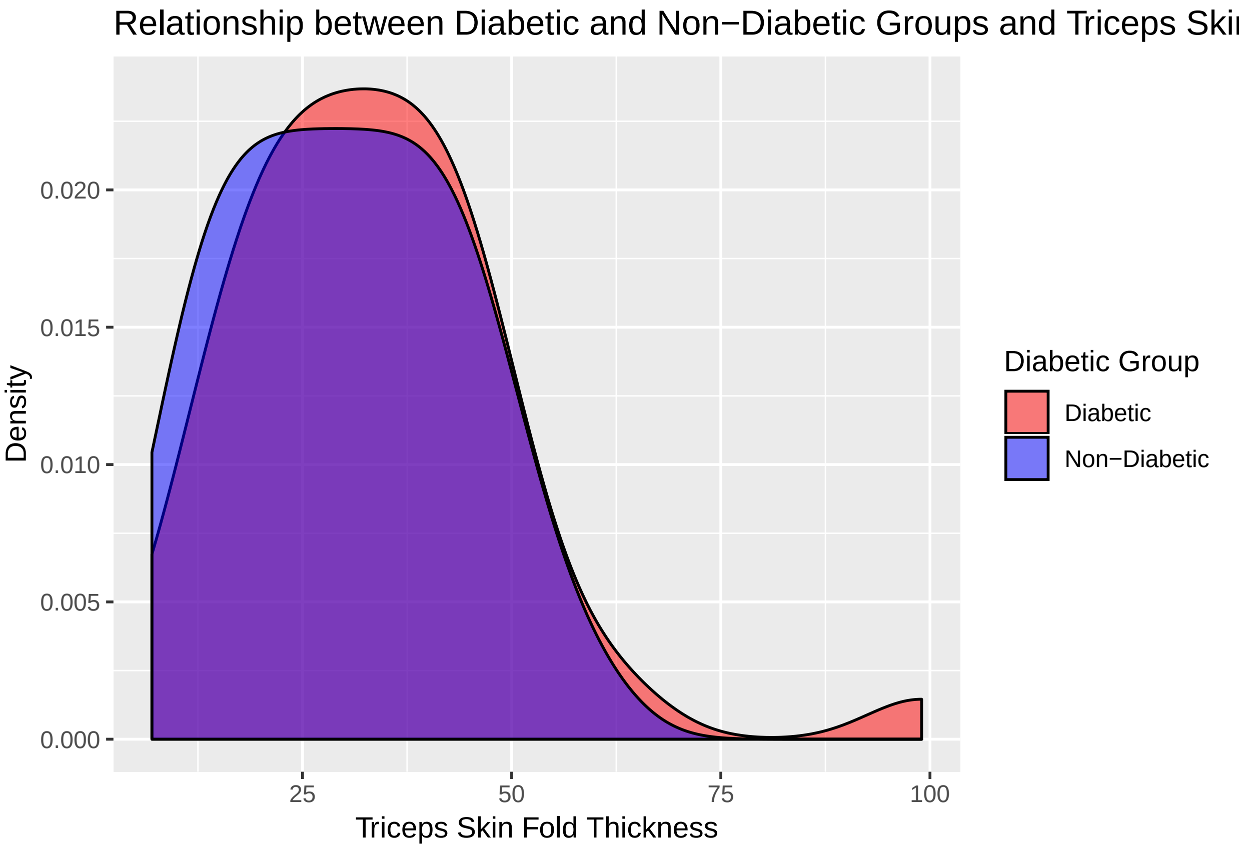
```
## RESPONSE
##   0   1
## 500 268
```

Plots

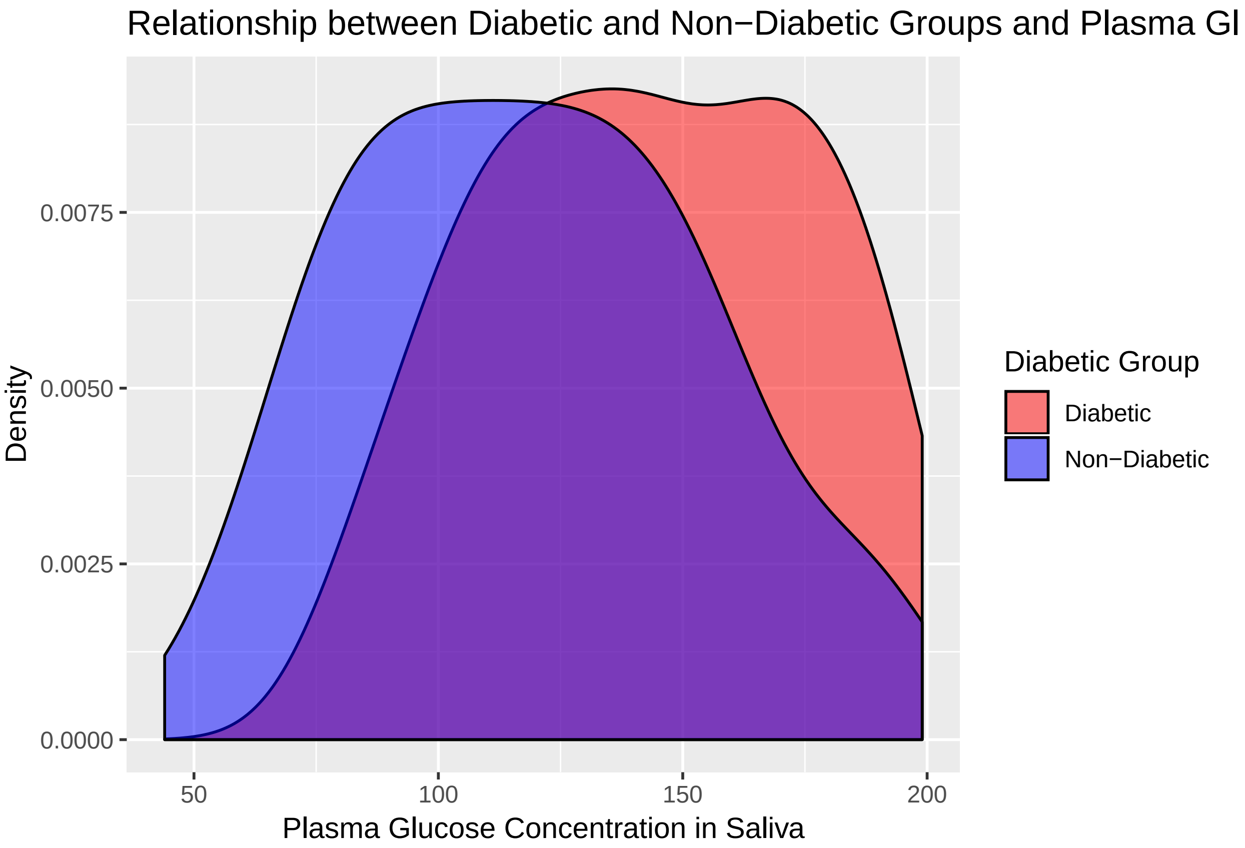
```
#Plot regarding blood pressure
pima_bp <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(BP > 0) %>%
  group_by(BP, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_bp, aes(x=BP, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Diastolic Blood Pressure") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Blood Pressure") +
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```



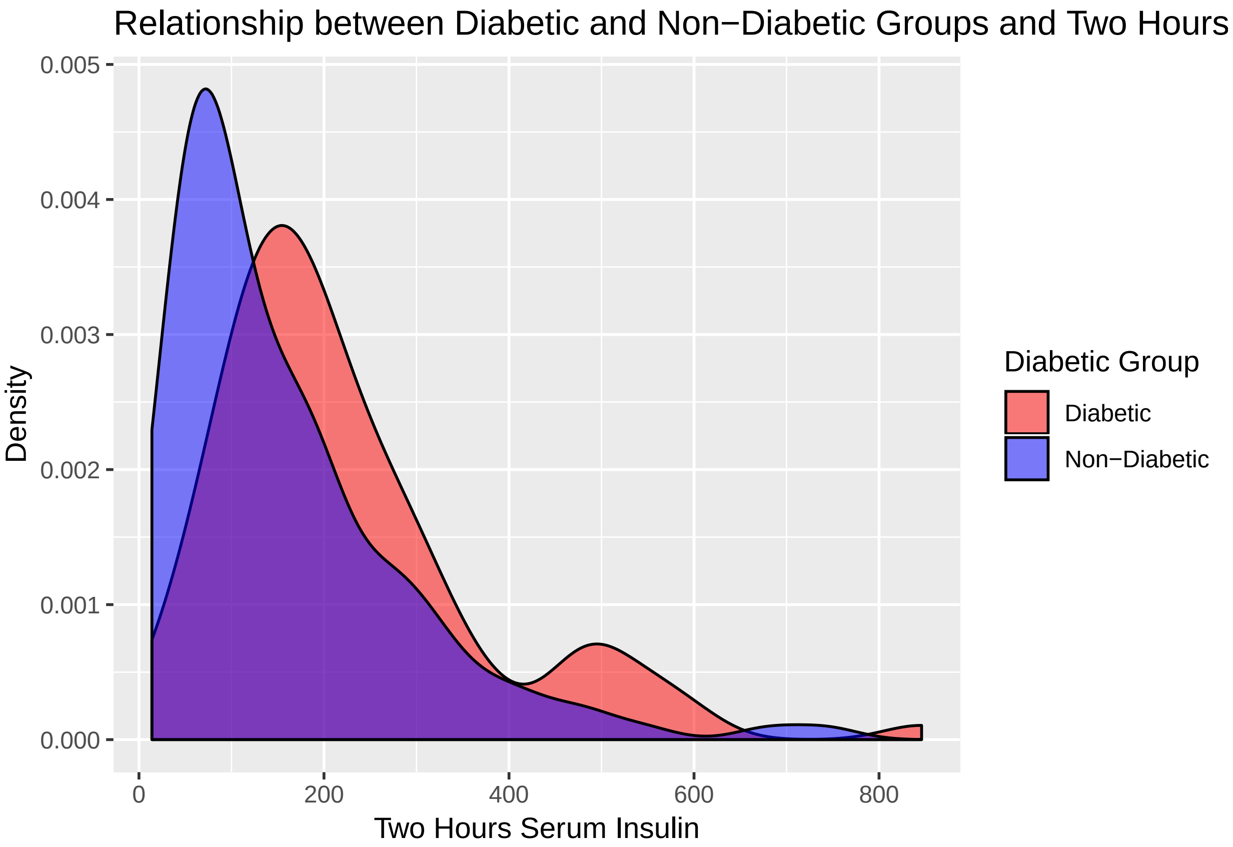
```
#Plot regarding skinfold thickness
pima_thick <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(THICK > 0) %>%
  group_by(THICK, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_thick, aes(x=THICK, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Triceps Skin Fold Thickness") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Triceps Skin Fold Thickness") +
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```



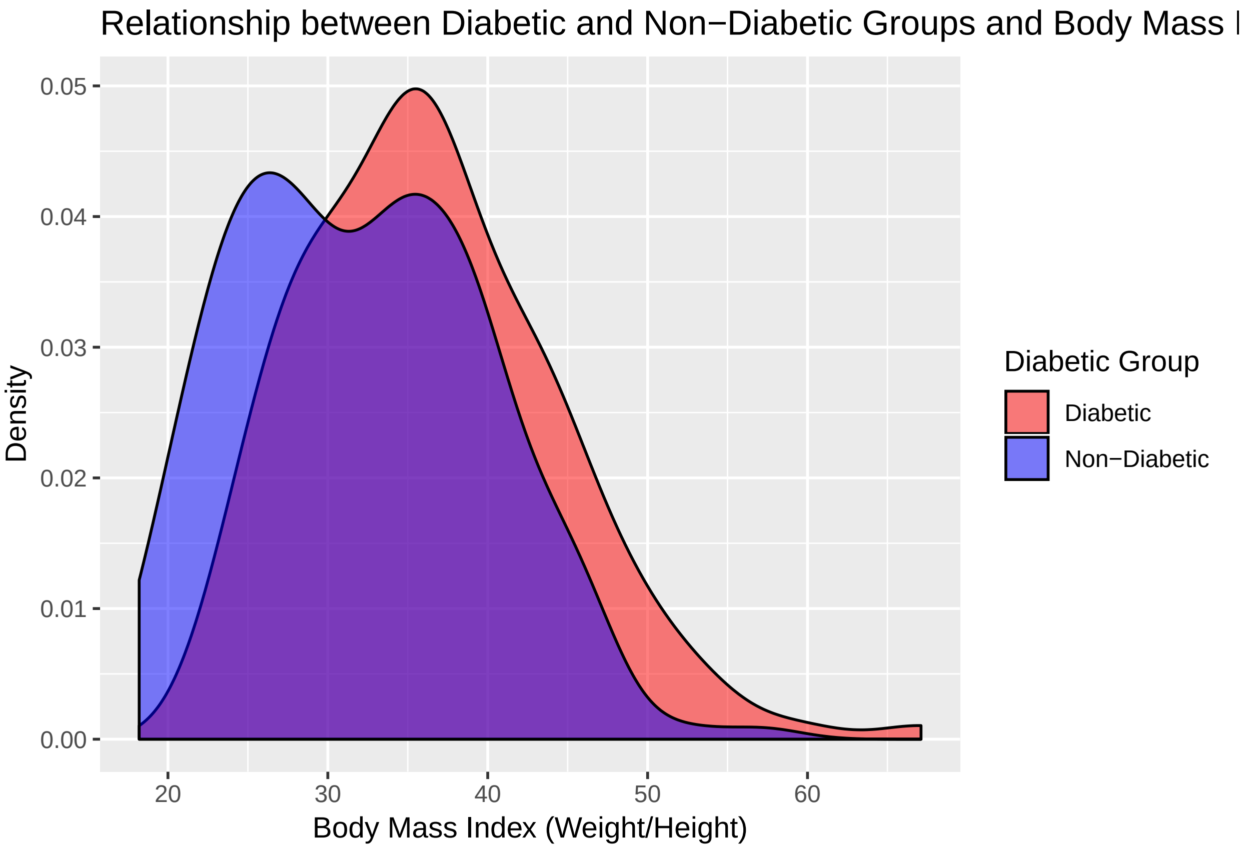
```r
#Plot regarding plasma level
pima_plasma <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(PLASMA > 0) %>%
  group_by(PLASMA, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_plasma, aes(x=PLASMA, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Plasma Glucose Concentration in Saliva") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Plasma Glucose Concentration in Sal
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```



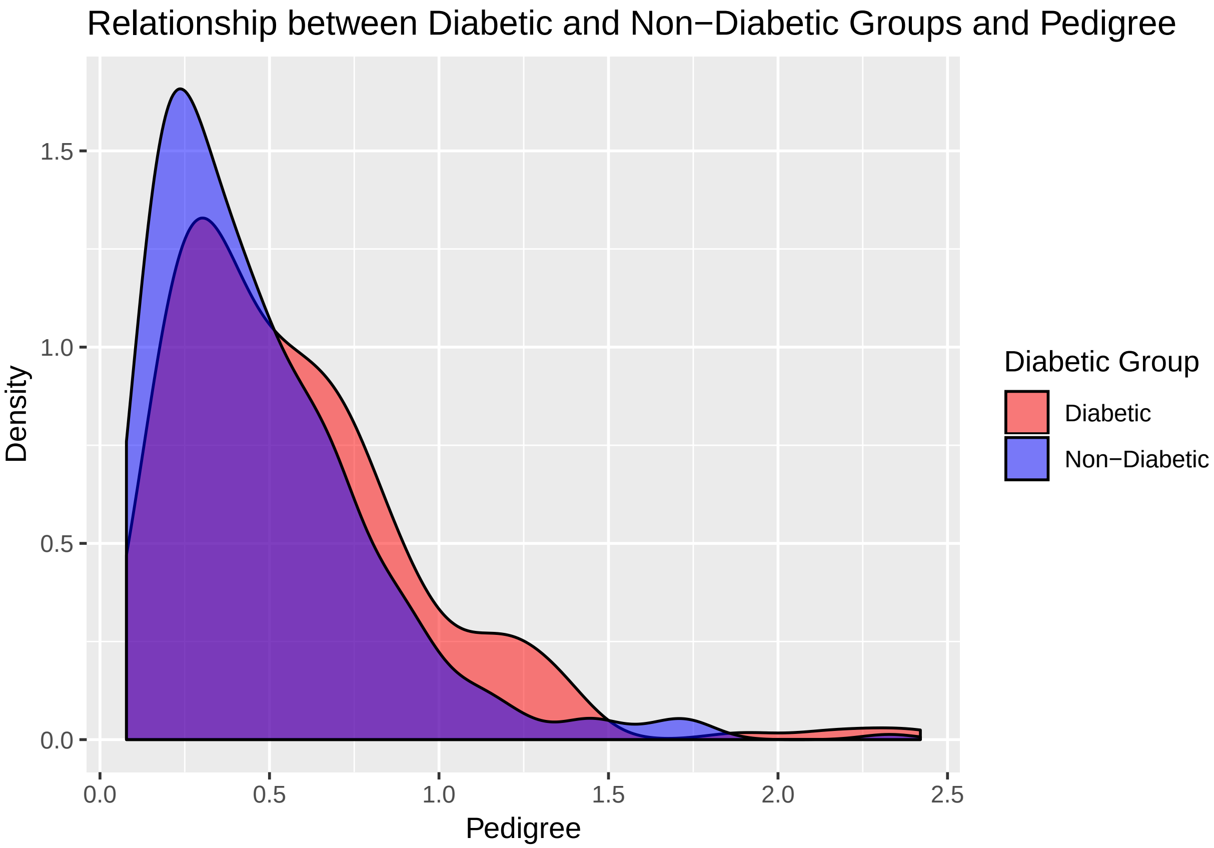
```
#Plot regarding insulin levels
pima_insulin <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(INSULIN > 0) %>%
  group_by(INSULIN, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_insulin, aes(x=INSULIN, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Two Hours Serum Insulin") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Two Hours Serum Insulin") +
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```



```
#Plot regarding BMI
pima_body <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(BODY > 0) %>%
  group_by(BODY, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_body, aes(x=BODY, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Body Mass Index (Weight/Height)") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Body Mass Index (Weight/Height)")
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```

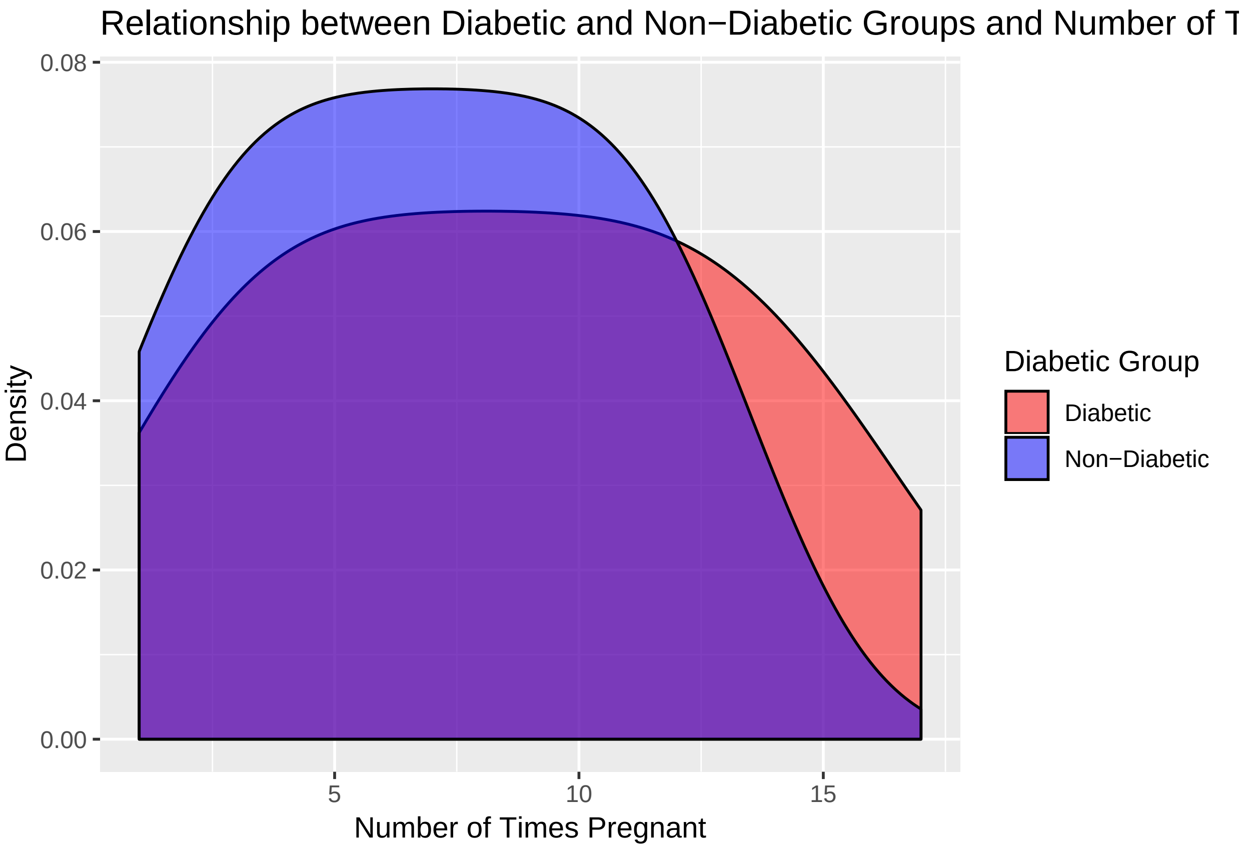


```
#Plot regarding pedigree
pima_pedigree <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(PEDIGREE > 0) %>%
  group_by(PEDIGREE, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_pedigree, aes(x=PEDIGREE, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Pedigree") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Pedigree") +
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```



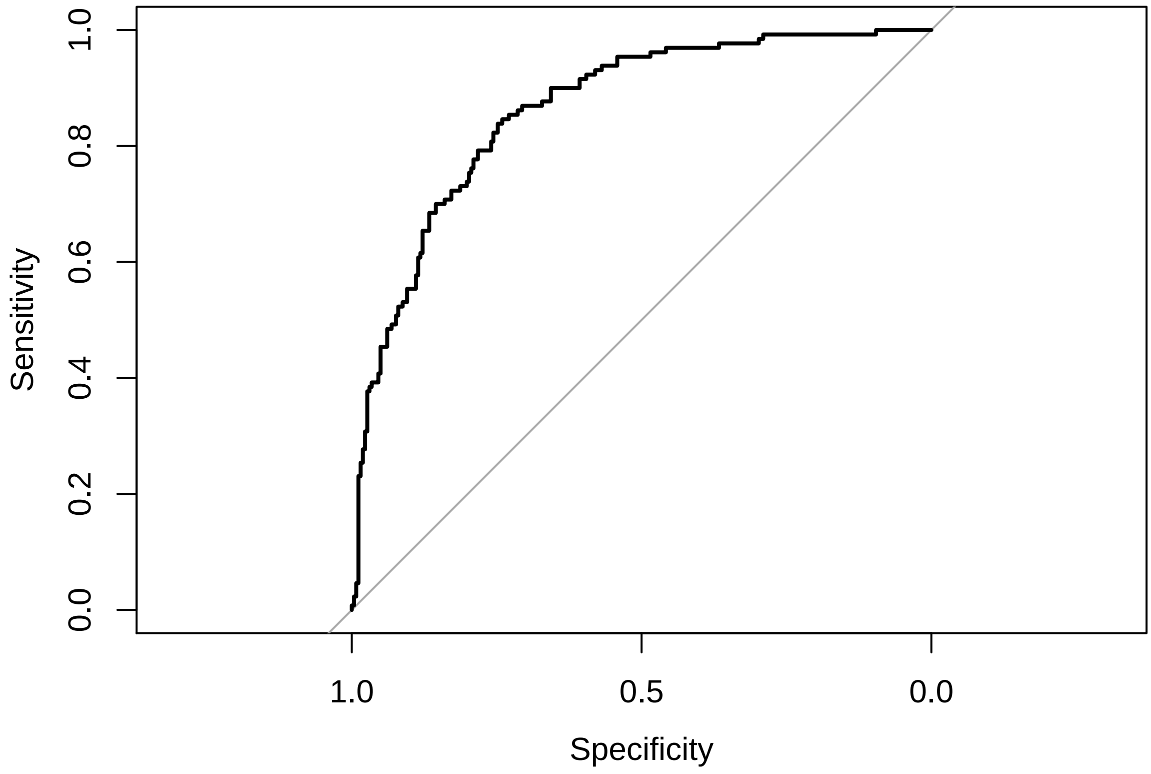
```
#Pedigree may not be overly important; however, it does seem that there may be a slight relationship
```

```
#plot regarding number of pregnancy
pima_prg <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(PRG > 0) %>%
  group_by(PRG, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_prg, aes(x=PRG, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density") +
  xlab("Number of Times Pregnant") +
  ggtitle("Relationship between Diabetic and Non-Diabetic Groups and Number of Times Pregnant") +
  scale_fill_manual("Diabetic Group", values=c("Red","Blue"))
```



ROC Curve

```
#Creates a ROC Curve and prints out the area under the curve which suggests this model is pretty strong
predictions <- predict(filteredmodel, data=filteredpima)
plot.roc(filteredpima$RESPONSE, predictions)
```



```
auc(filteredpima$RESPONSE, predictions)
```

```
## Area under the curve: 0.8631
```

Classification Tree

```
#Creates a classifcation tree that denotes the thresholds: defines no diabetes, pre-diabetes, and diabe
filteredmodelpart <- rpart(RESPONSE ~ ., data = filteredpima)
printcp(filteredmodelpart)
```

```
##
## Regression tree:
## rpart(formula = RESPONSE ~ ., data = filteredpima)
##
## Variables actually used in tree construction:
## [1] AGE      BODY     BP       INSULIN  PEDIGREE PLASMA
##
## Root node error: 86.888/392 = 0.22165
##
## n= 392
##
##         CP nsplit rel error  xerror     xstd
## 1 0.239181      0   1.00000 1.00511 0.036346
## 2 0.055005      1   0.76082 0.76797 0.049483
## 3 0.054577      2   0.70581 0.81835 0.056061
## 4 0.044501      3   0.65124 0.80197 0.057682
## 5 0.021167      5   0.56223 0.76046 0.058999
## 6 0.013939      6   0.54107 0.77132 0.063177
## 7 0.013933      8   0.51319 0.77281 0.066506
## 8 0.012276     12   0.45746 0.78618 0.067418
## 9 0.010000     13   0.44518 0.80443 0.068277
```

```
pdf("diabetes.pdf", width = 25, height = 10)
plot(as.party(filteredmodelpart))
dev.off()
```

```
## pdf
##   2
```

```
diabetes <- mutate(filteredpima, fittedtree = predict(filteredmodelpart))
```

```
#Removes age and pedigree from the classication tree (see pdf in the folder)
filteredmodelpart <- rpart(RESPONSE ~ . - AGE - PEDIGREE, data = filteredpima)
printcp(filteredmodelpart)
```

```
##
## Regression tree:
## rpart(formula = RESPONSE ~ . - AGE - PEDIGREE, data = filteredpima)
##
## Variables actually used in tree construction:
## [1] INSULIN PLASMA  PRG     THICK
##
## Root node error: 86.888/392 = 0.22165
##
## n= 392
##
##          CP nsplit rel error  xerror     xstd
## 1 0.239181      0   1.00000 1.00446 0.036315
## 2 0.055005      1   0.76082 0.77091 0.049641
## 3 0.035786      2   0.70581 0.74793 0.053488
## 4 0.031114      3   0.67003 0.75900 0.057239
## 5 0.021753      4   0.63891 0.75356 0.056017
## 6 0.018444      6   0.59541 0.76175 0.058570
## 7 0.010000      7   0.57696 0.77484 0.062781
```

```
pdf("diabetes1.pdf", width = 25, height = 12)
plot(as.party(filteredmodelpart))
dev.off()
```

```
## pdf
##   2
```

```
diabetes <- mutate(filteredpima, fittedtree = predict(filteredmodelpart))
```