Type 2 Diabetes in Pima Native Americans

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") #
```

Prevalence of Diabetes in the Dataset

```
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)
#We noticed biological impossibilities in the data, such as blood pressures of 0. Since it is not possi
```

Probability Model of Diabetes in an Individual Pima Female

```
tally(filteredpima$RESPONSE)
```

```
## X
##   0   1
## 262 130
```

```
262/(262+130)
```

```
## [1] 0.6683673
```

```
#Approximately 67% of the individual females in the filtered dataset 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) #The stepwise logistic regression returned PLASMA, BODY, and PEDIGREE as signifi
```

```
##
## 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
```

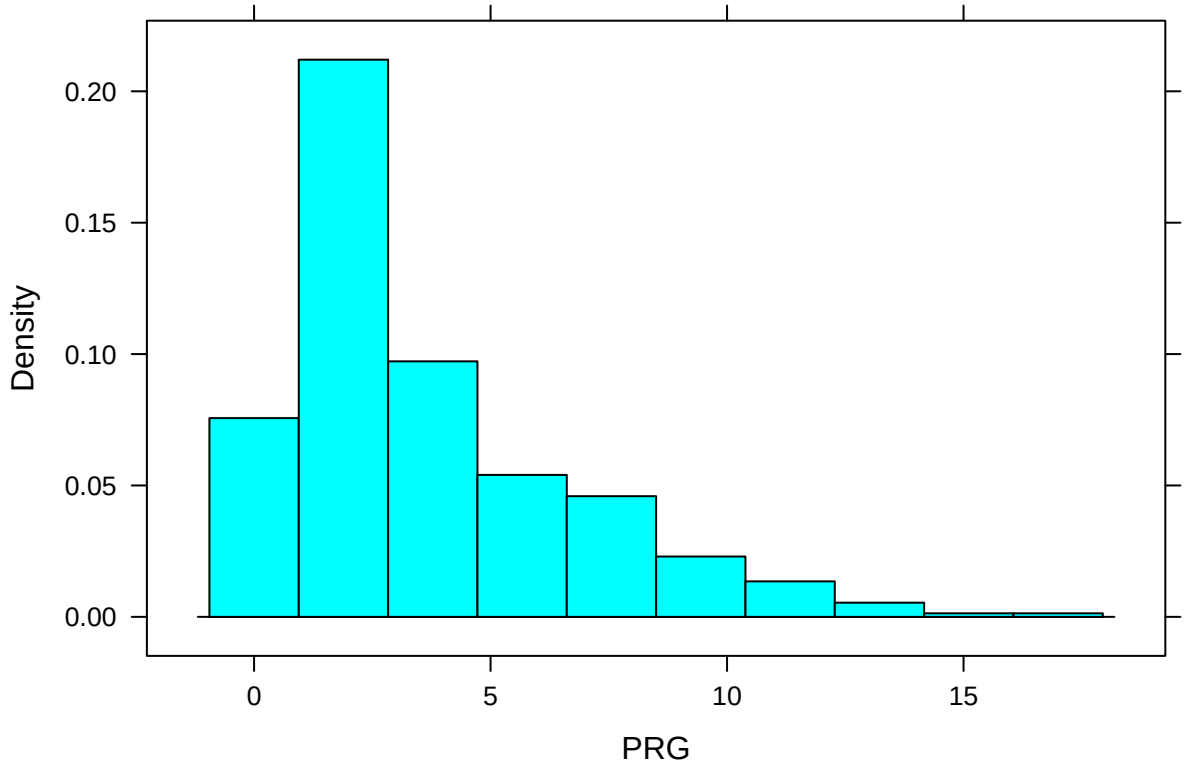
```
##
## Number of Fisher Scoring iterations: 5
```

EDA

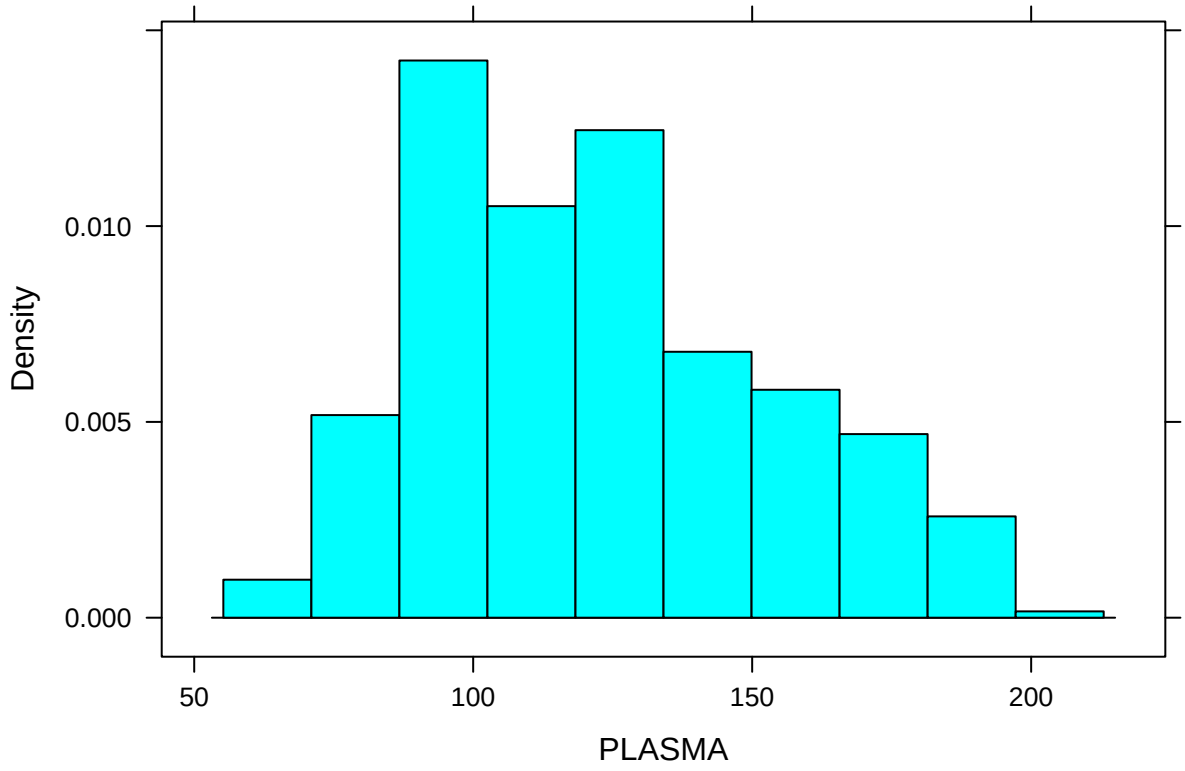
```
summary(filteredpima)
```

```
##       PRG              PLASMA             BP              THICK
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
##     INSULIN            BODY          PEDIGREE          AGE
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00
##  Median :125.50   Median :33.20   Median :0.4495   Median :27.00
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##     RESPONSE
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3316
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

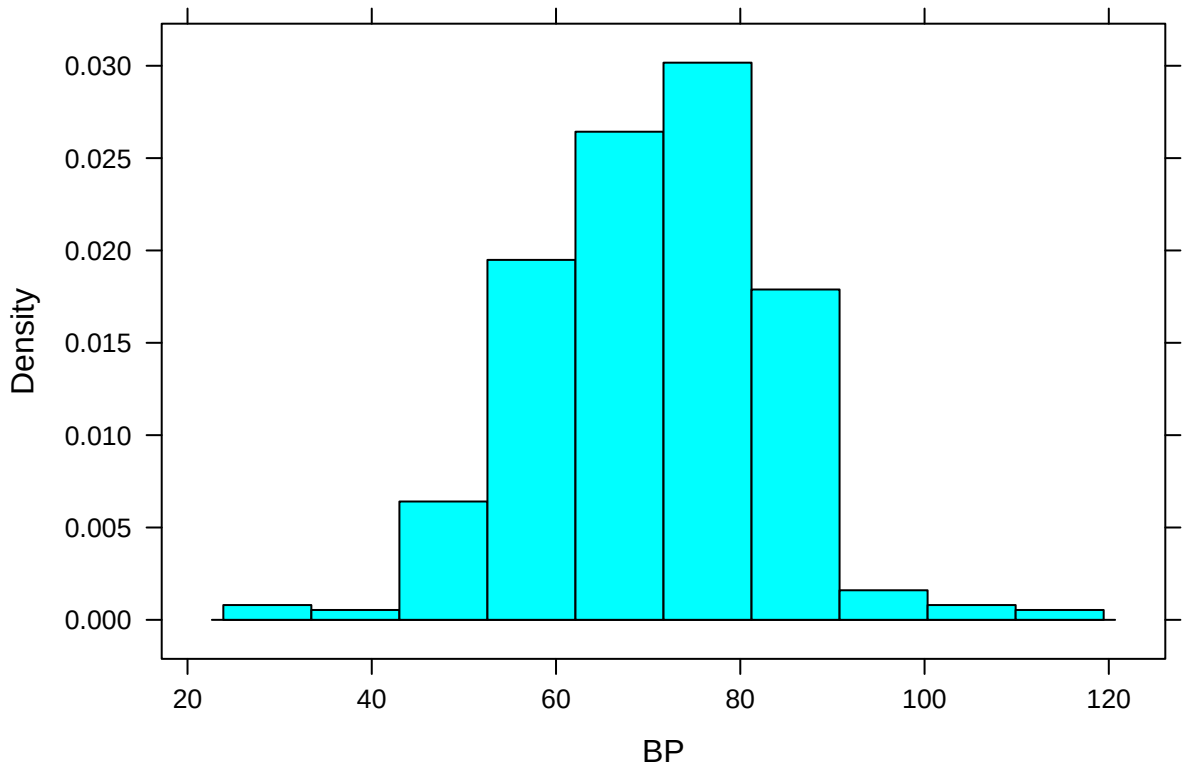
```
histogram(~PRG, data=filteredpima) #PRG is skewed to the right.
```



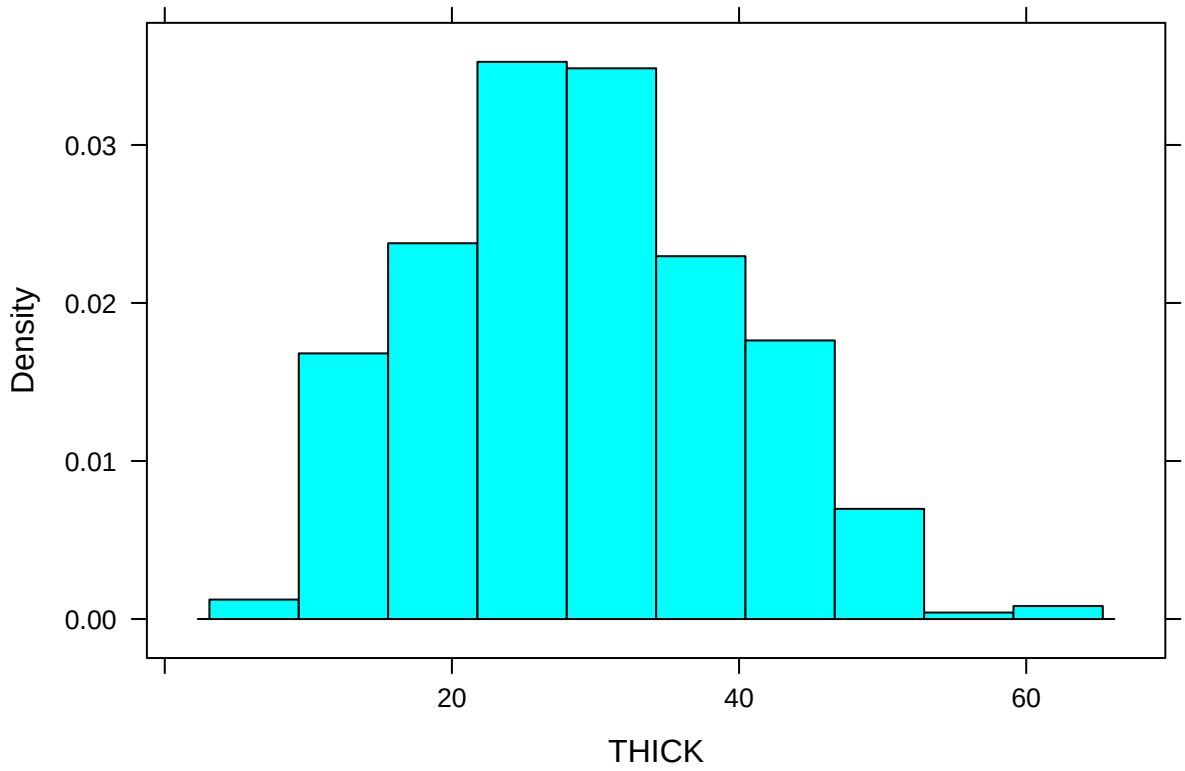

```
histogram(~PLASMA, data=filteredpima) #Follows a somewhat normal distribution.
```



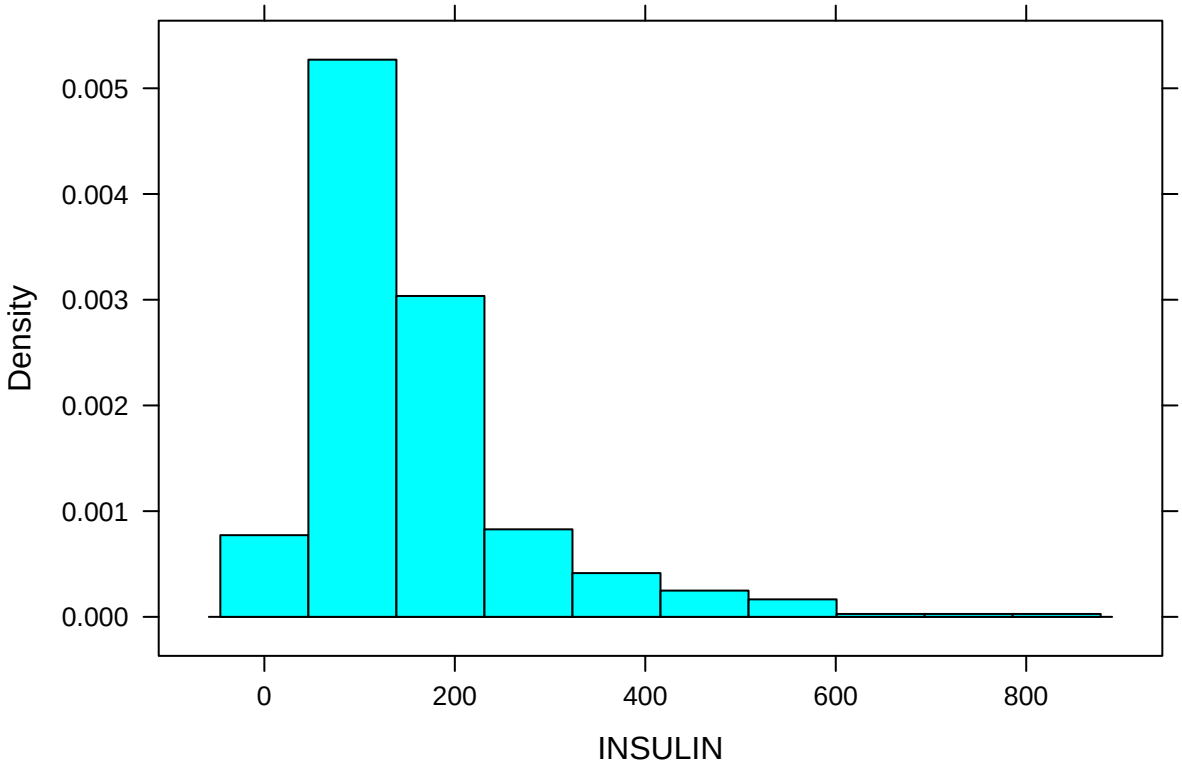

```
histogram(~BP, data=filteredpima) #Also somewhat bell-shaped.
```



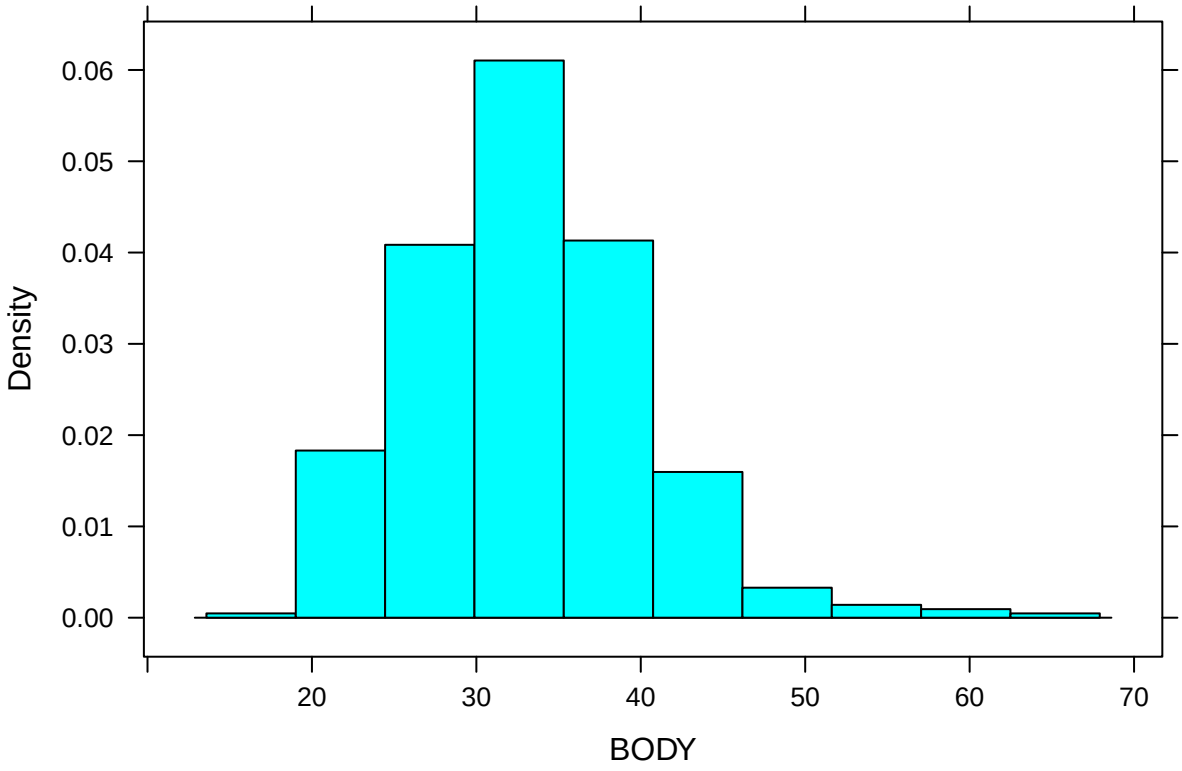

```
histogram(~THICK, data=filteredpima) #Mostly normal distribution.
```



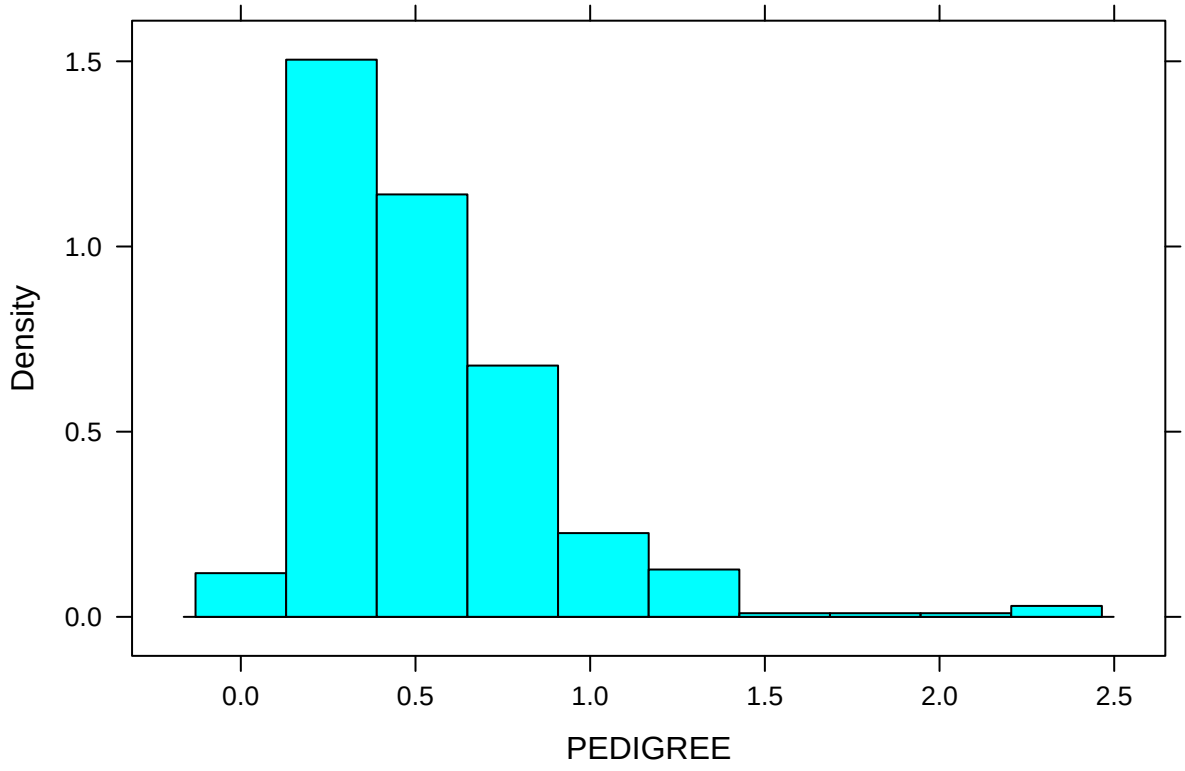

```
histogram(~INSULIN, data=filteredpima) #Skewed to the right.
```



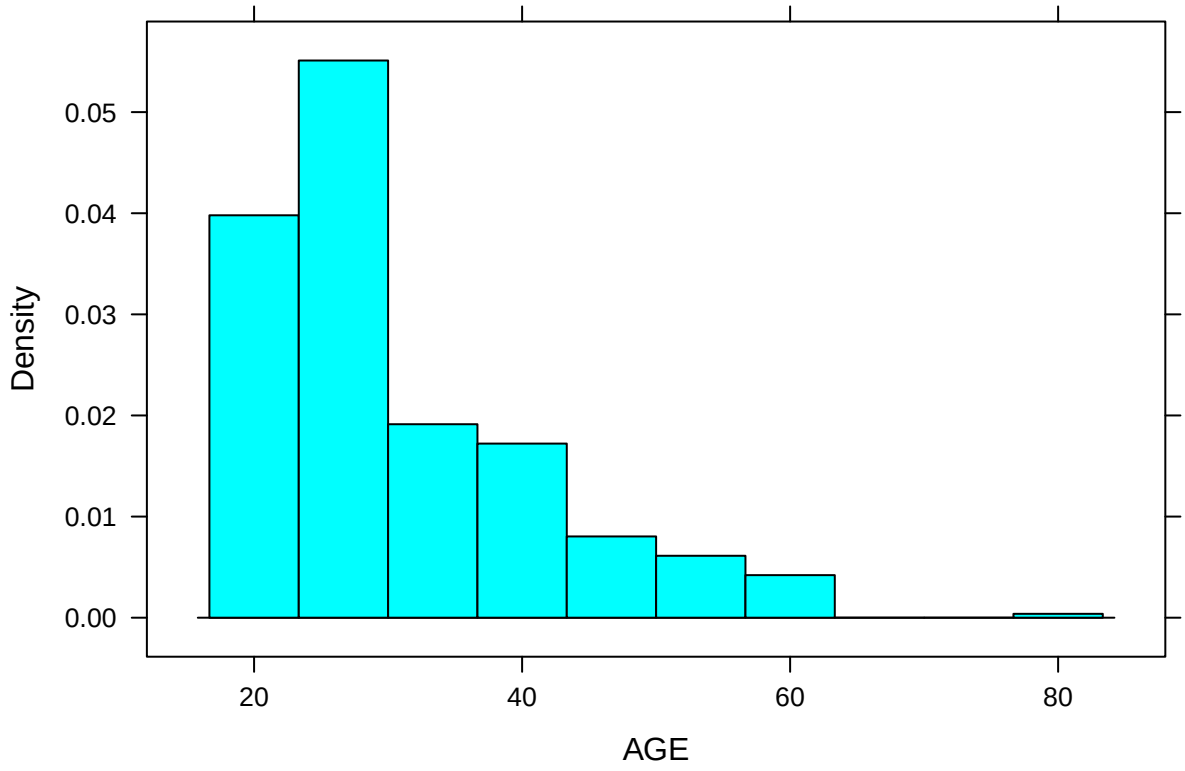

```
histogram(~BODY, data=filteredpima) #Somewhat normal distribution, but is slightly skewed to the right.
```




```
histogram(~PEDIGREE, data=filteredpima) #Skewed to the right.
```

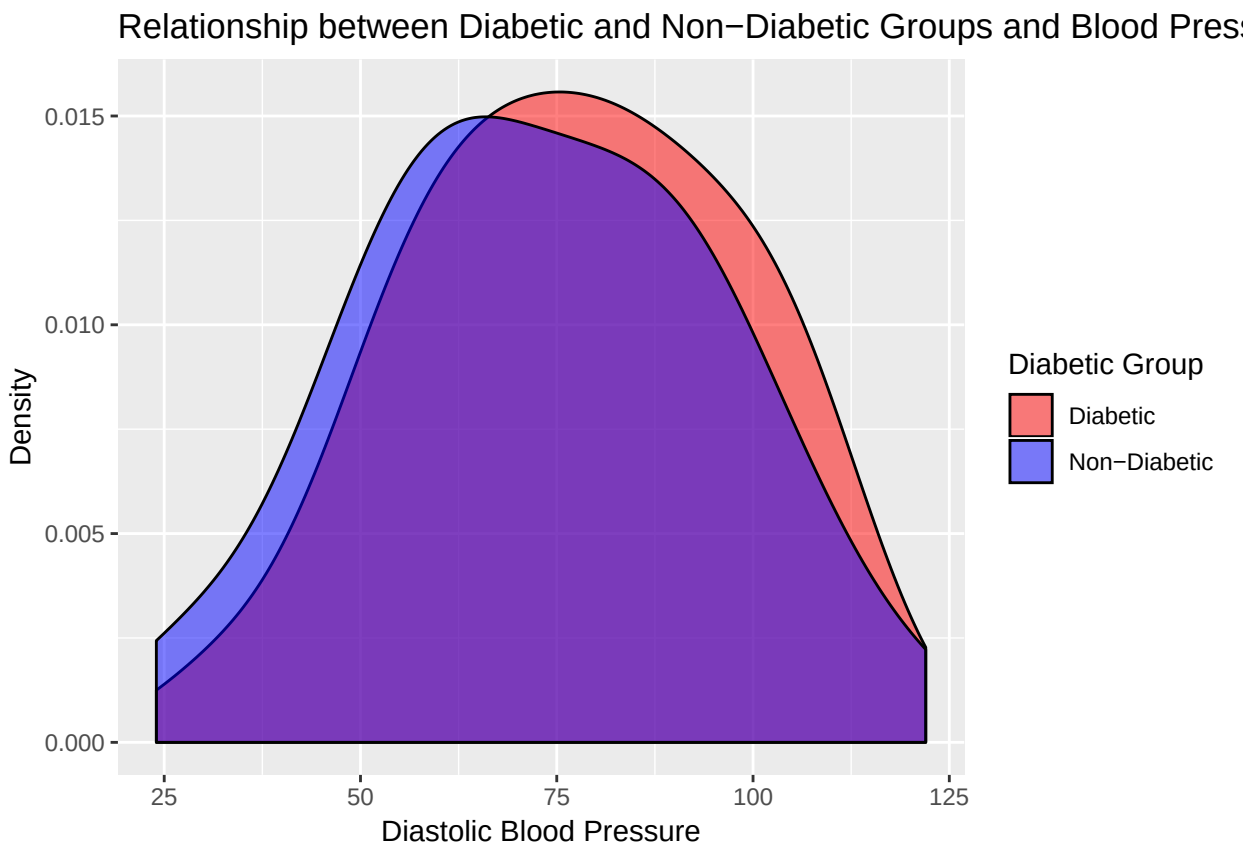



```
histogram(~AGE, data=filteredpima) #Skewed to the right.
```

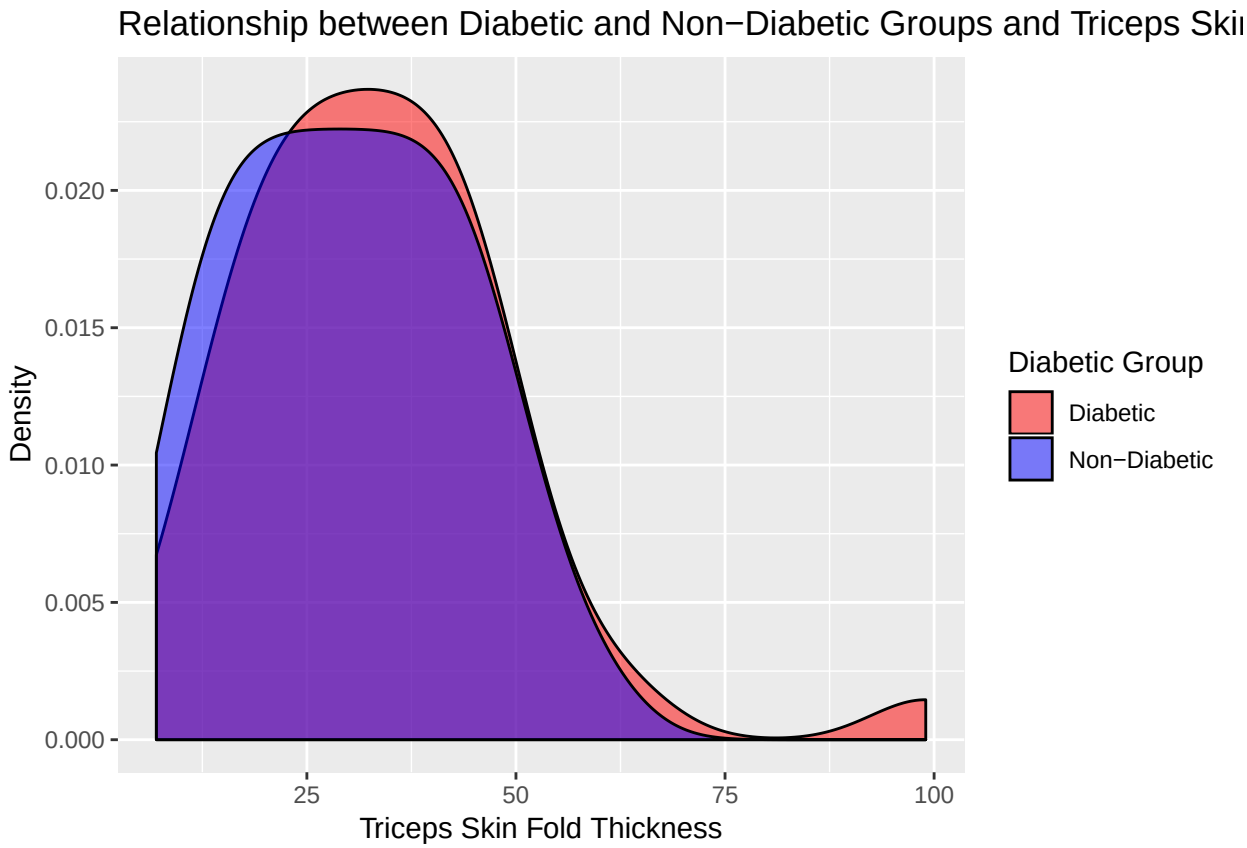


Plots

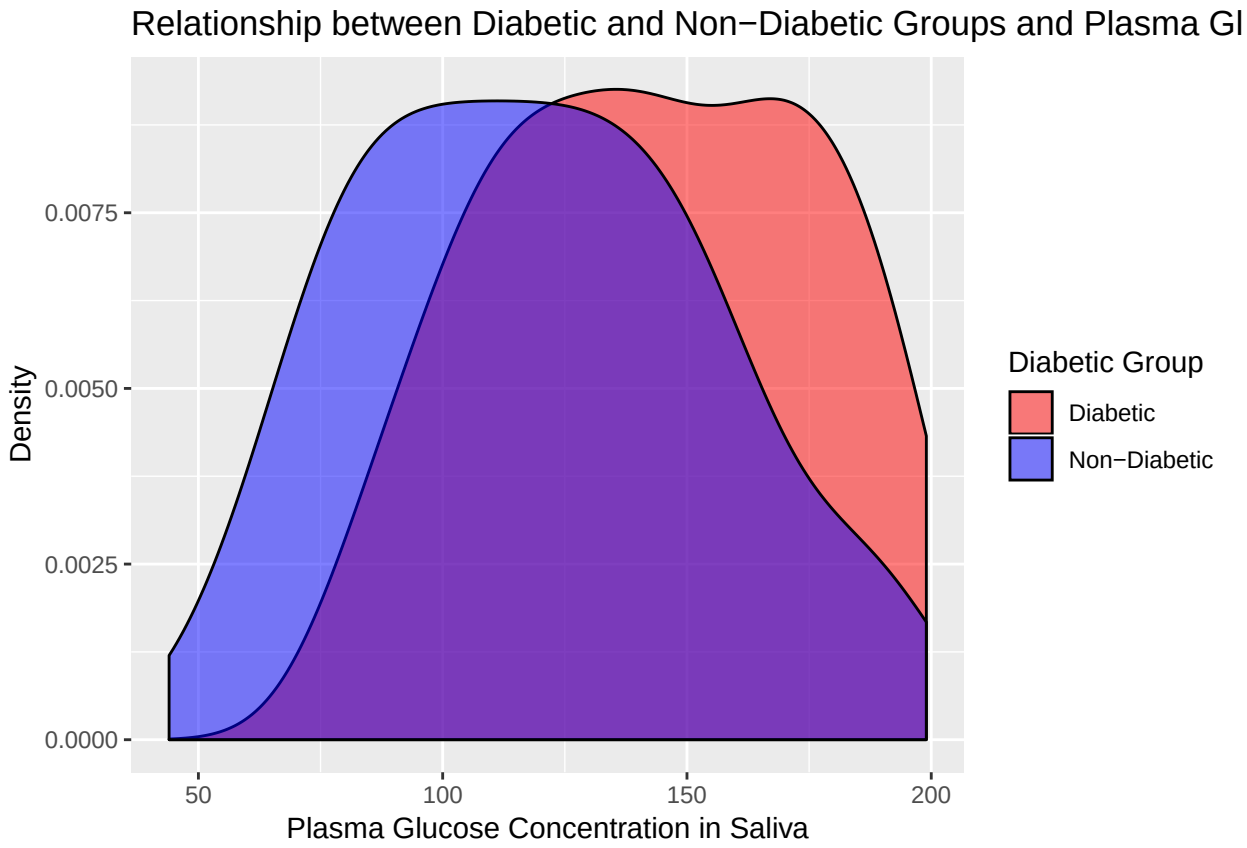
```
#Plot regarding blood pressure
pima_bp <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(BP > 0) %>%
  group_by(BP, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_bp, aes(x=BP, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  ylab("Density")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



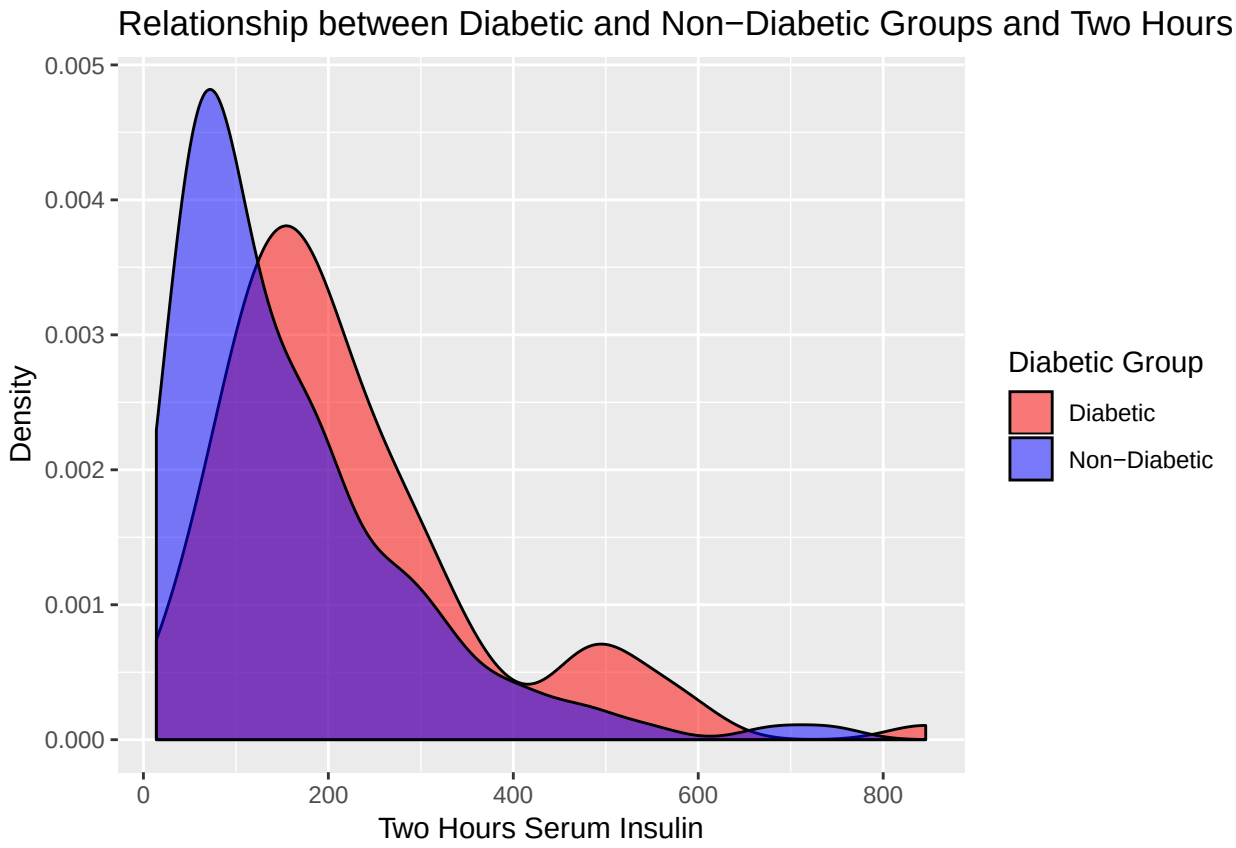

```
#Plot regarding skinfold thickness
pima_thick <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(THICK > 0) %>%
  group_by(THICK, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_thick, aes(x=THICK, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  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"))
```



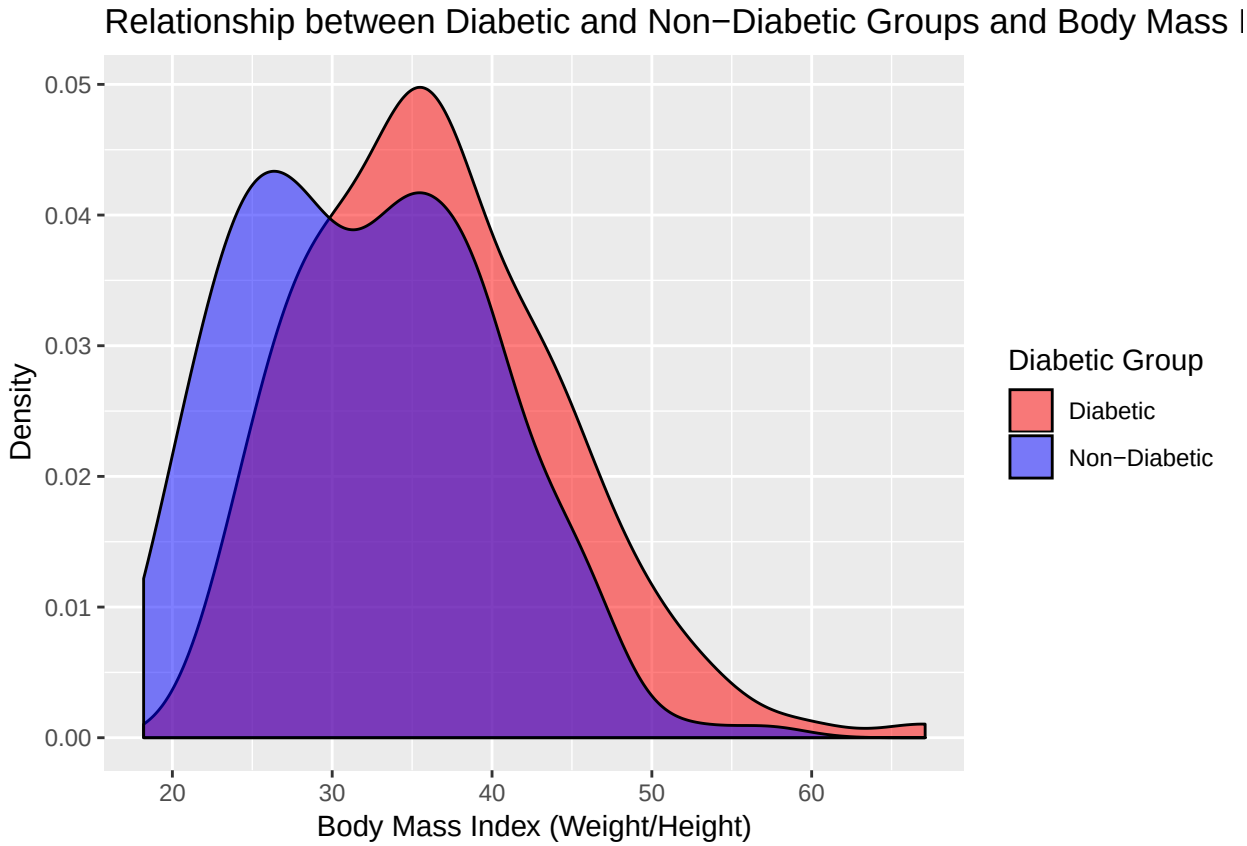

```r
#Plot regarding plasma level
pima_plasma <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(PLASMA > 0) %>%
  group_by(PLASMA, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_plasma, aes(x=PLASMA, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5)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



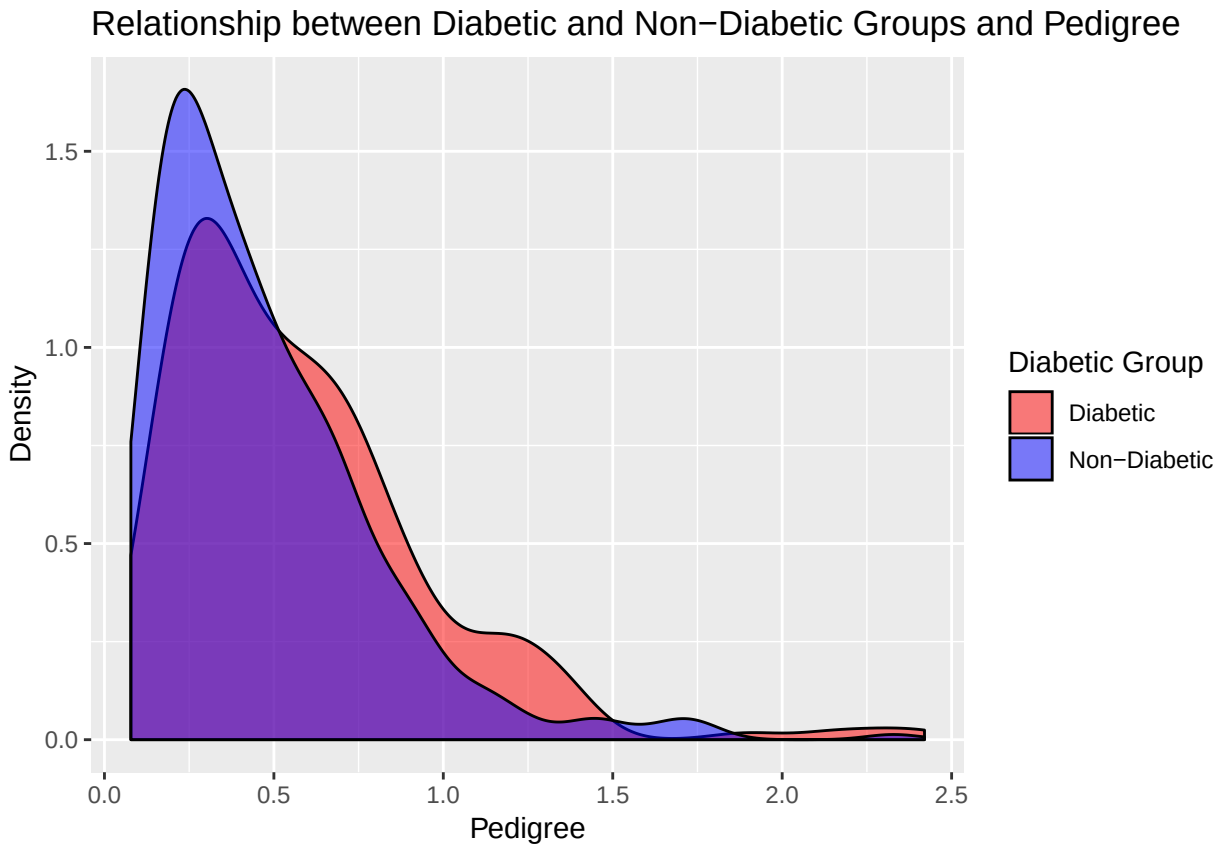

```
#Plot regarding insulin levels
pima_insulin <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(INSULIN > 0) %>%
  group_by(INSULIN, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_insulin, aes(x=INSULIN, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5) +
  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"))
```

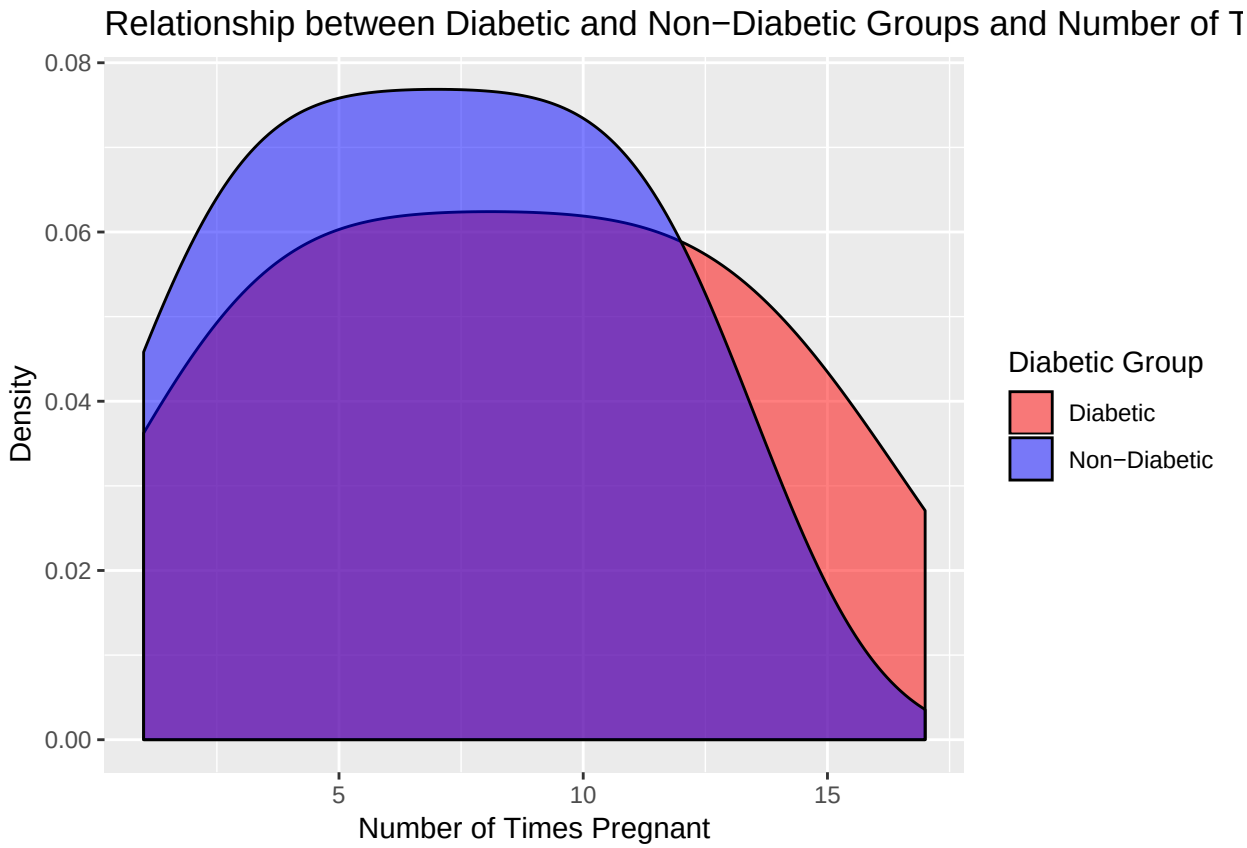



```
#Plot regarding pedigree
pima_pedigree <- pima %>%
  mutate(RESPONSE = ifelse(RESPONSE == 0, "Non-Diabetic", "Diabetic"))%>%
  filter(PEDIGREE > 0) %>%
  group_by(PEDIGREE, RESPONSE) %>%
  summarize(response_total = n())
ggplot(pima_pedigree, aes(x=PEDIGREE, fill=factor(RESPONSE))) +
  geom_density(alpha=0.5)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



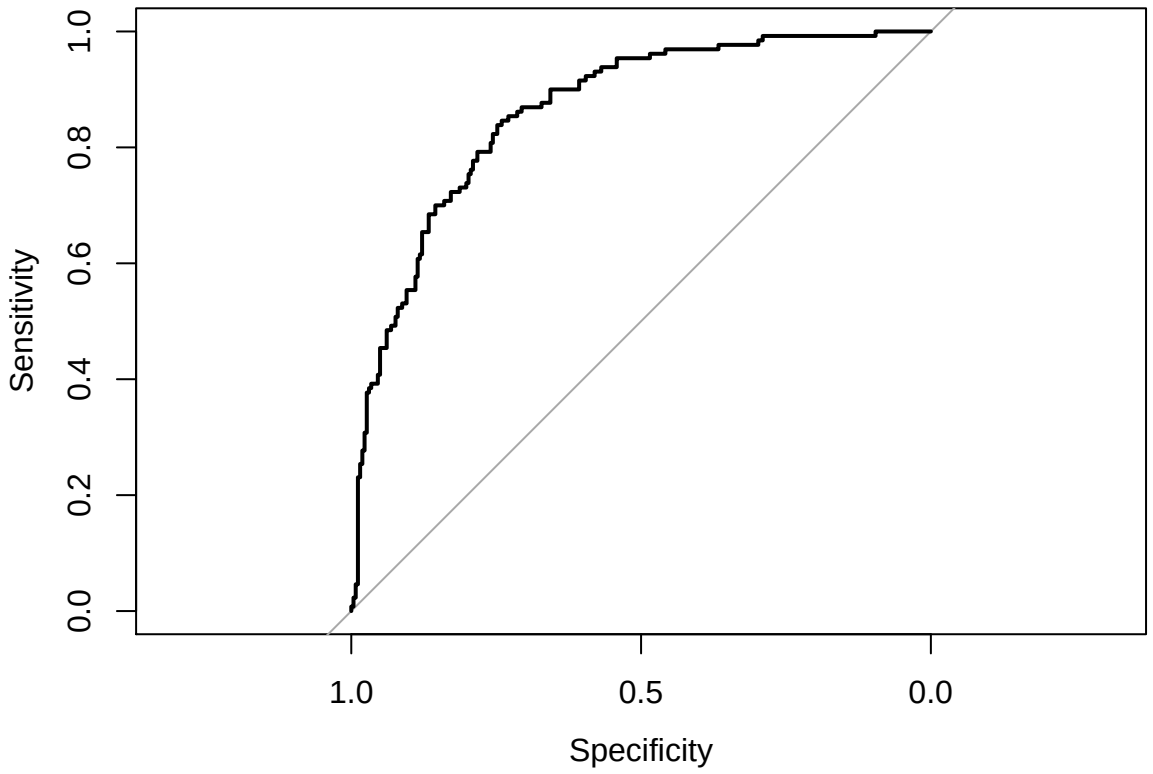

```
#Pedigree may not be overly important; however, it does seem that there may be a slight relationship
```

```
#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: An Evaluation of the Probability Model

```
#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 1.00111 0.036170
## 2 0.055005      1   0.76082 0.76373 0.049354
## 3 0.054577      2   0.70581 0.79597 0.055125
## 4 0.044501      3   0.65124 0.76355 0.057841
## 5 0.021167      5   0.56223 0.75462 0.060243
## 6 0.013939      6   0.54107 0.80901 0.064558
```

```
## 7 0.013933      8   0.51319 0.84504 0.067277
## 8 0.012276     12   0.45746 0.85412 0.067629
## 9 0.010000     13   0.44518 0.83483 0.067635
```

```
pdf("diabetes.pdf", width = 25, height = 12)
pdf("diabetes.pdf", width = 25, height = 10)
plot(as.party(filteredmodelpart))
dev.off()
```

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

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

A tree diagram is a way of representing a sequence of events and the purpose of using this diagram is to help predict if a patient has or does not have diabetes depending on certain thresholds of various physical characteristics. In this case, we start at the very top and examine an individual's plasma glucose concentration and make a move to the left or the right depending on what the individual's plasma concentration is. Then you would proceed to look at the next physical characteristic and make another move to the left or right again and again until you work your way to the bottom. The boxplots displayed on the bottom generally illustrate 2 sides: a lower risk side on the left and a higher risk side on the right

This diagram shows that plasma glucose concentration is very important because plasma shows up a total of 4 times in the tree so it suggests that plasma level is a pretty big factor in determining if an individual has diabetes.

We also created a classification tree diagram without age and pedigree as age and pedigree are uncontrollable factors.

```
#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 1.00712 0.036424
## 2 0.055005      1   0.76082 0.82624 0.052443
## 3 0.035786      2   0.70581 0.79702 0.056131
## 4 0.031114      3   0.67003 0.80348 0.057055
## 5 0.021753      4   0.63891 0.77999 0.058022
## 6 0.018444      6   0.59541 0.83971 0.063225
## 7 0.010000      7   0.57696 0.84386 0.066672
```

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

```
## pdf
```

```
##   2
```

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

Recommendations

We recommend that patients talk to their physician about their treatment choices. Namely focusing on the controllable factors such as plasma glucose concentration in saliva. Diet & exercise are cruical and they help lower levels of the most significant risk factors of diabetes (plasma, BMI, insulin, skinfold thickness, etc.).