APPENDIX

Code ▾

Data Description:

Pregnancies - To express the Number of pregnancies

Glucose - To express the Glucose level in blood

BloodPressure - To express the Blood pressure measurement

SkinThickness - To express the thickness of the skin

Insulin - To express the Insulin level in blood

BMI - To express the Body mass index

DiabetesPedigreeFunction - To express the Diabetes percentage

Age - To express the age

Outcome - To express the final result 1 is Yes and 0 is No

Hide

```
diabetes <- read.csv("diabetes.csv")
diabetes
```

Pregnancies <int>	Gluc... <int>	BloodPressure <int>	SkinThickness <int>	Insulin <int>	BMI <dbl>	DiabetesPedigreeFunc <
6	148	72	35	0	33.6	0
1	85	66	29	0	26.6	0
8	183	64	0	0	23.3	0
1	89	66	23	94	28.1	0
0	137	40	35	168	43.1	2
5	116	74	0	0	25.6	0
3	78	50	32	88	31.0	0
10	115	0	0	0	35.3	0
2	197	70	45	543	30.5	0
8	125	96	0	0	0.0	0

1-10 of 768 rows | 1-7 of 9 columns Previous **1** 2 3 4 5 6 ... 77 Next

Hide

```
# Check for missing values
any(is.na(diabetes))
```

```
[1] FALSE
```

Hide

```
colSums(is.na(diabetes)) # Check for missing values in each column
```

```
              Pregnancies                  Glucose            BloodPressure             SkinThi
ckness                  Insulin                      BMI 
                       0                        0                        0 
0                        0                        0 
DiabetesPedigreeFunction                      Age                  Outcome 
                       0                        0                        0 
```

Hide

```
# Check for duplicated rows
diabetes[duplicated(diabetes), ]
```

0 rows | 1-7 of 9 columns

Hide

```
#Display the list structure
str(diabetes)
```

```
'data.frame':	768 obs. of  9 variables:
 $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

Hide

```
summary(diabetes)
```

```
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin           BMI       
DiabetesPedigreeFunction      Age           Outcome     
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 
0.00   Min.   :0.0780           Min.   :21.00   Min.   :0.000  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:2
7.30   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :3
2.00   Median :0.3725           Median :29.00   Median :0.000  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :3
1.99   Mean   :0.4719           Mean   :33.24   Mean   :0.349  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:3
6.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :6
7.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000  
```

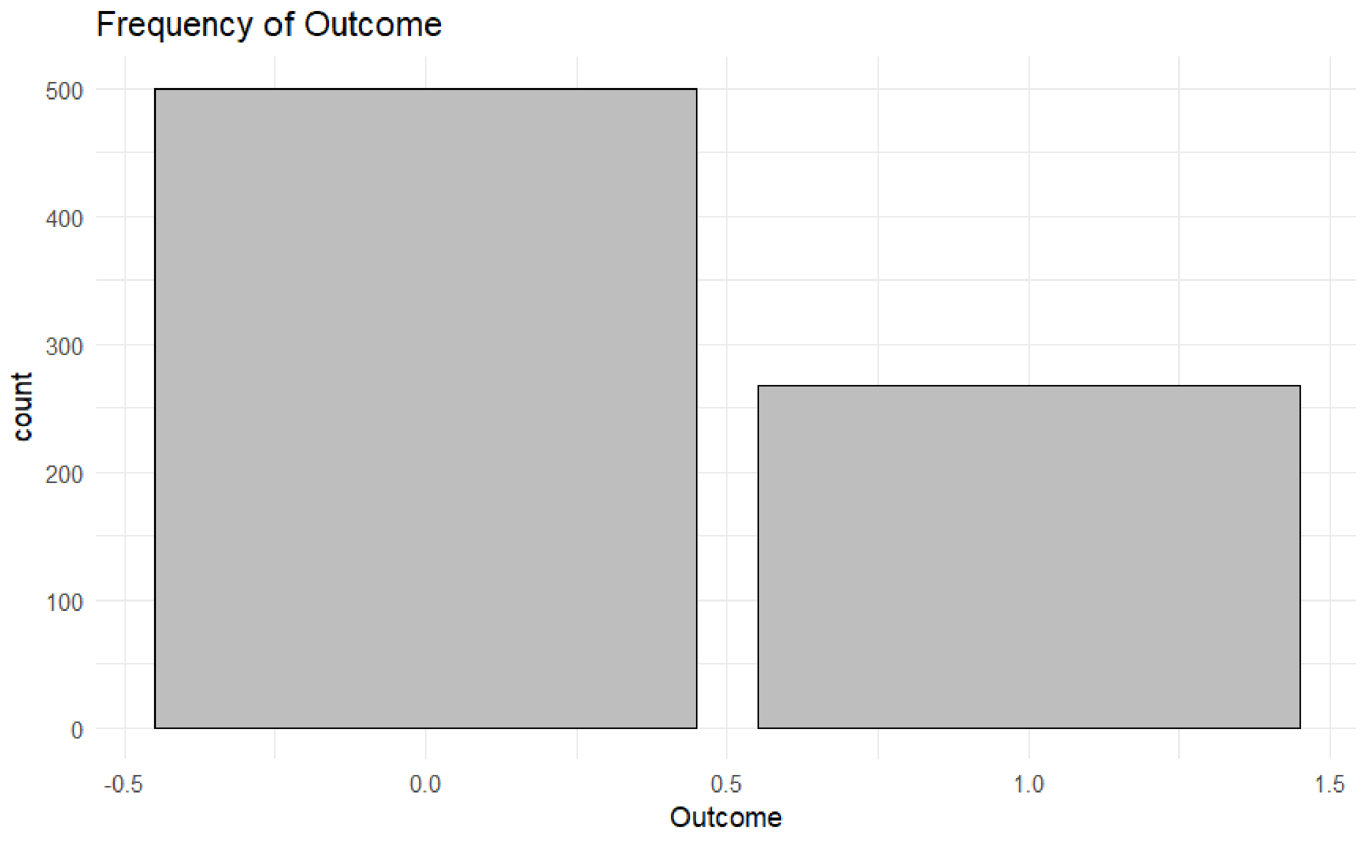
Hide

```
library(ggplot2)
```

```
Warning: package ‘ggplot2’ was built under R version 4.3.3
```

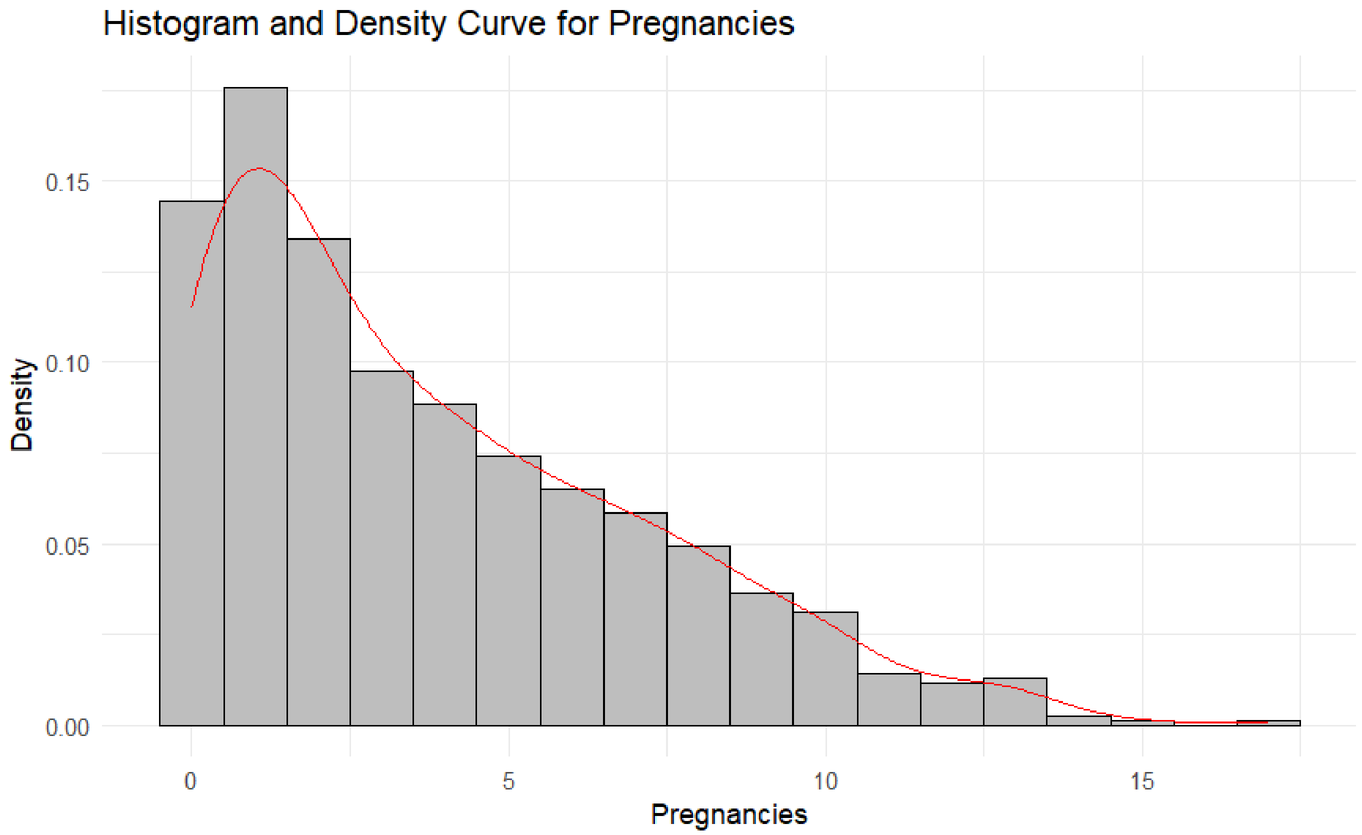
Hide

```
# Function to create barplot for categorical variable
create_barplot <- function(data, variable, title) {
  ggplot(data, aes(x = {{ variable }})) +
    geom_bar(fill = "gray", color = "black") +
    labs(title = title, x = deparse(substitute(variable))) +
    theme_minimal()
}

# Function to create histogram and density curve
create_histogram_density_plot <- function(data, variable, title) {
  ggplot(data, aes(x = {{ variable }})) +
    geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "gray", color = "blac
k") +
    geom_density(color = "red") +
    labs(title = title, x = deparse(substitute(variable)), y = "Density") +
    theme_minimal()
}

# Barplots for categorical variables
create_barplot(diabetes, Outcome, "Frequency of Outcome")
```



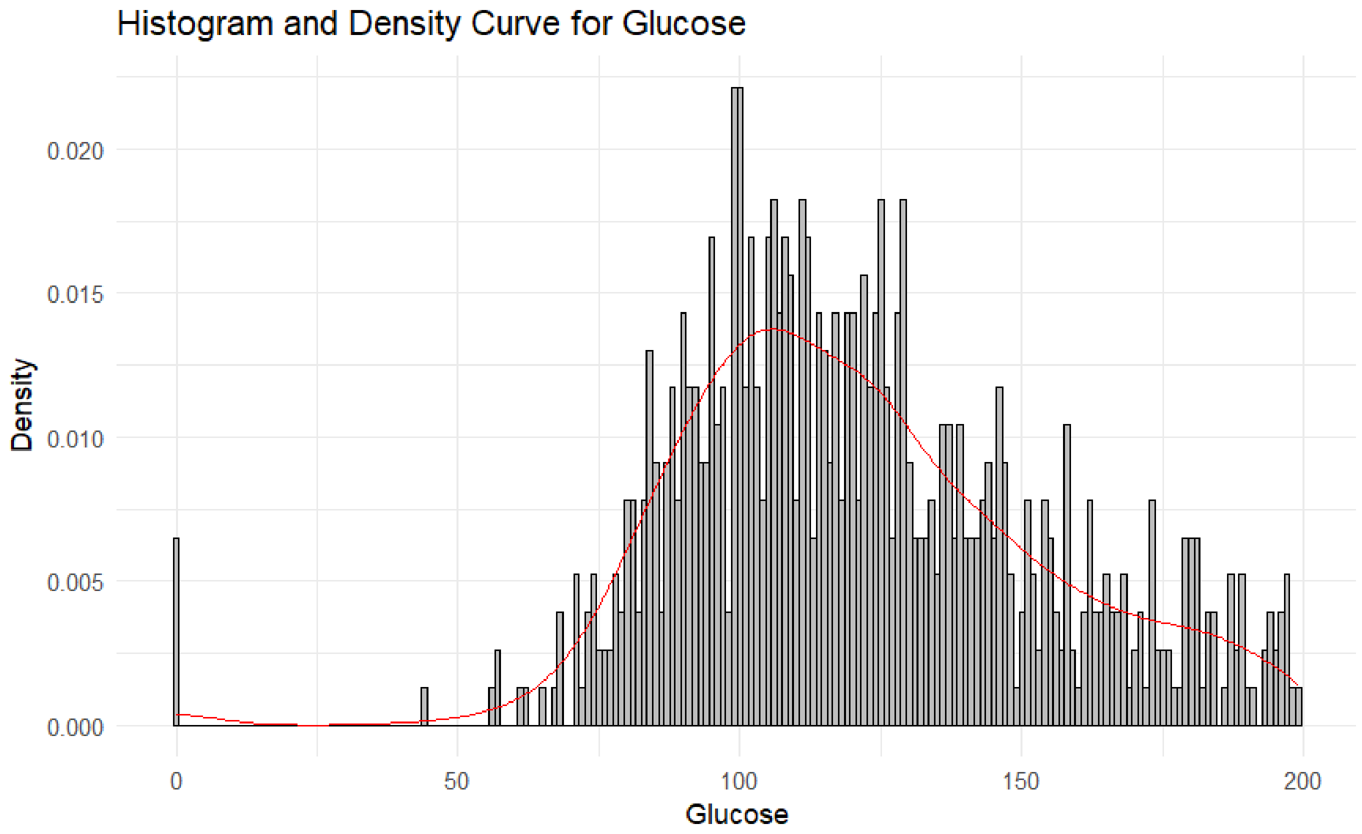
Hide

```
# Apply the function to each numeric variable
create_histogram_density_plot(diabetes, Pregnancies, "Histogram and Density Curve for Pregnancies")
```


Histogram and Density Curve for Pregnancies

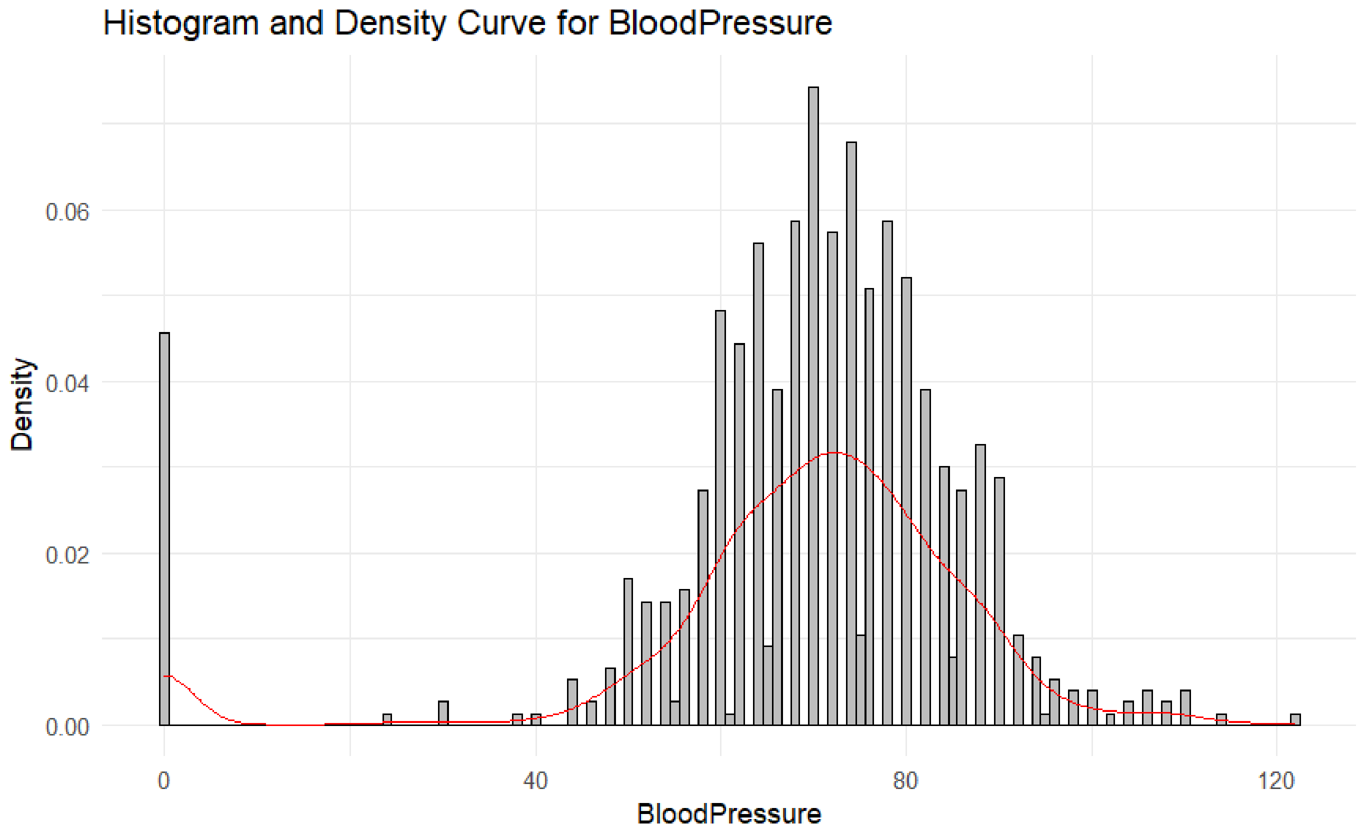
Hide

```
create_histogram_density_plot(diabetes, Glucose, "Histogram and Density Curve for Glucose")
```


Histogram and Density Curve for Glucose

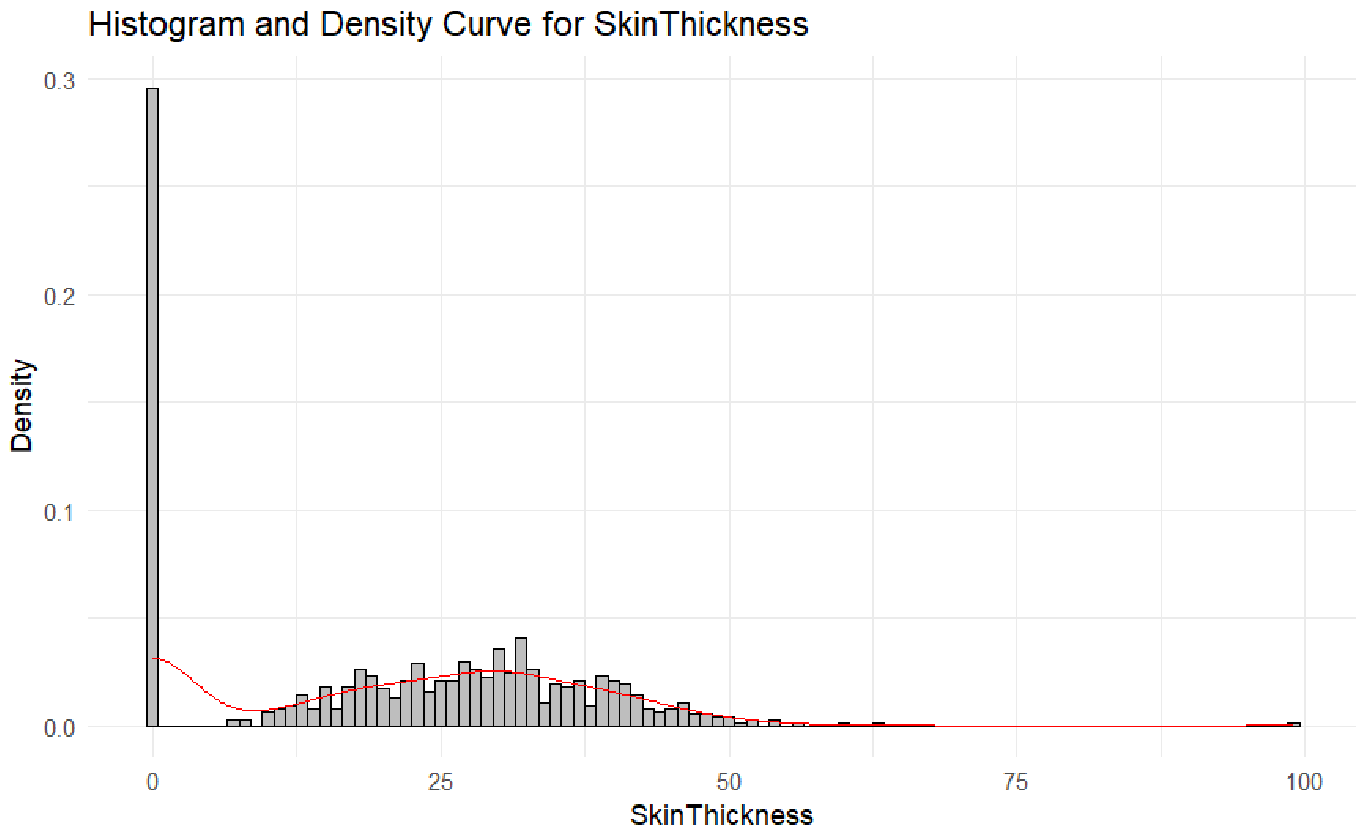
Hide

```
create_histogram_density_plot(diabetes, BloodPressure, "Histogram and Density Curve for BloodPressure")
```



Hide

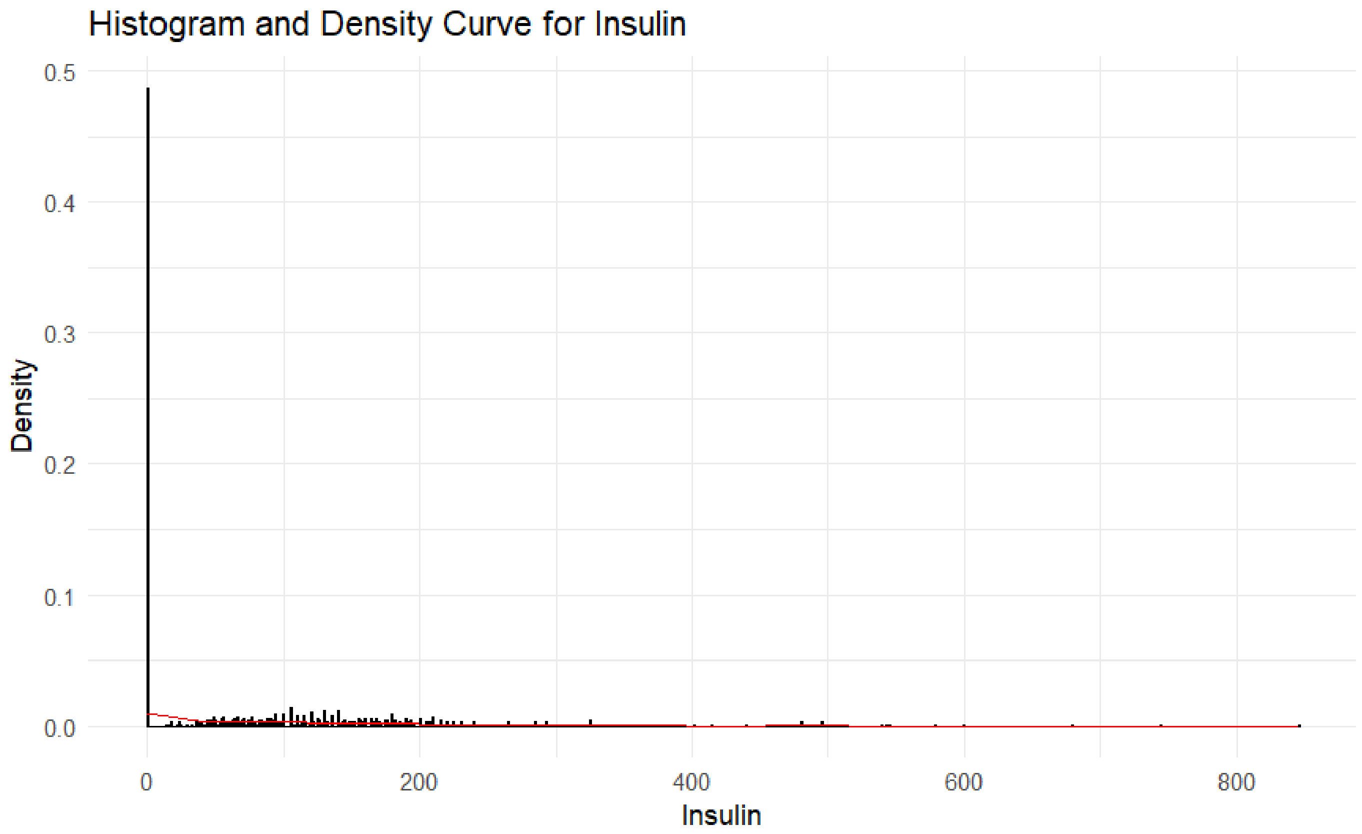
```
create_histogram_density_plot(diabetes, SkinThickness, "Histogram and Density Curve for SkinThickness")
```



Histogram and Density Curve for SkinThickness

Hide

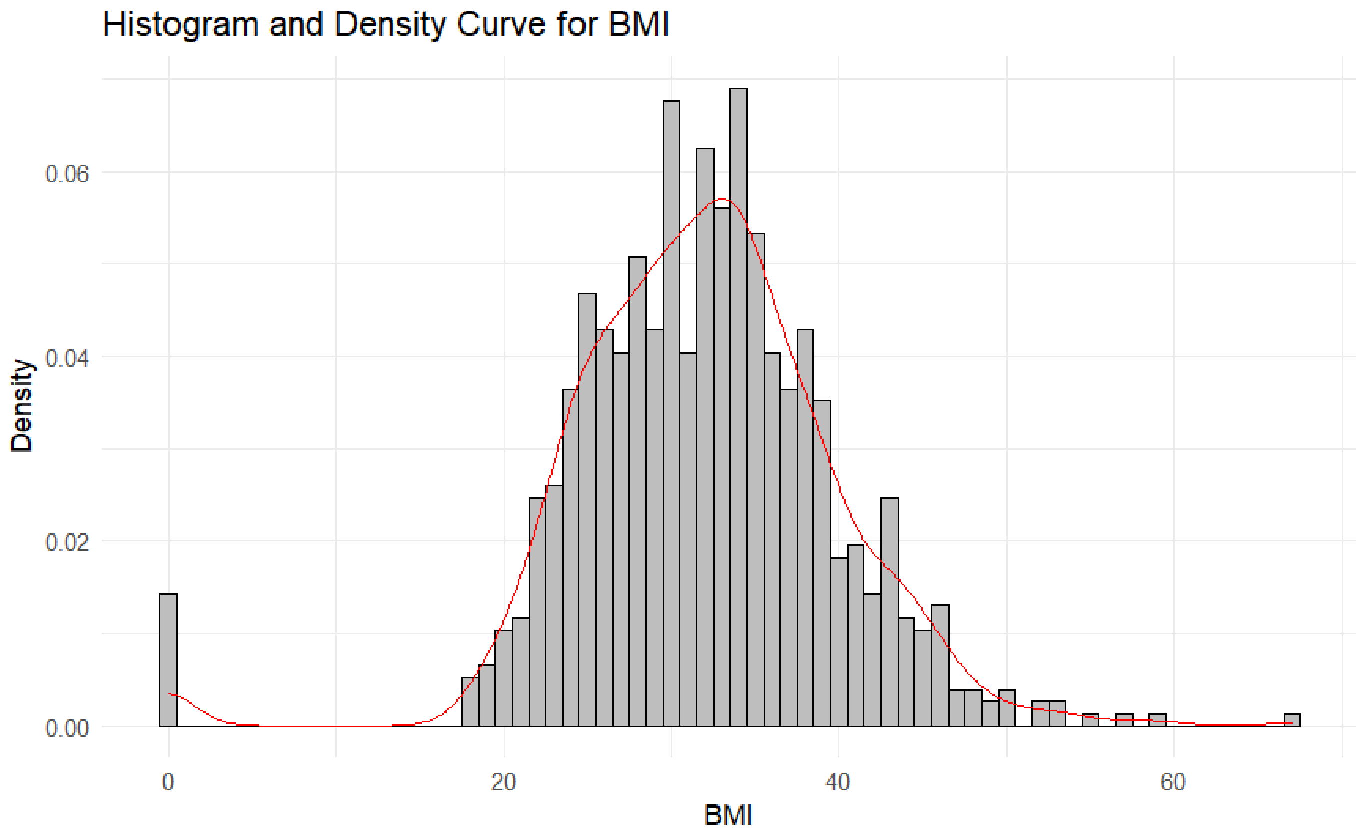
```
create_histogram_density_plot(diabetes, Insulin, "Histogram and Density Curve for Insulin")
```



Histogram and Density Curve for Insulin

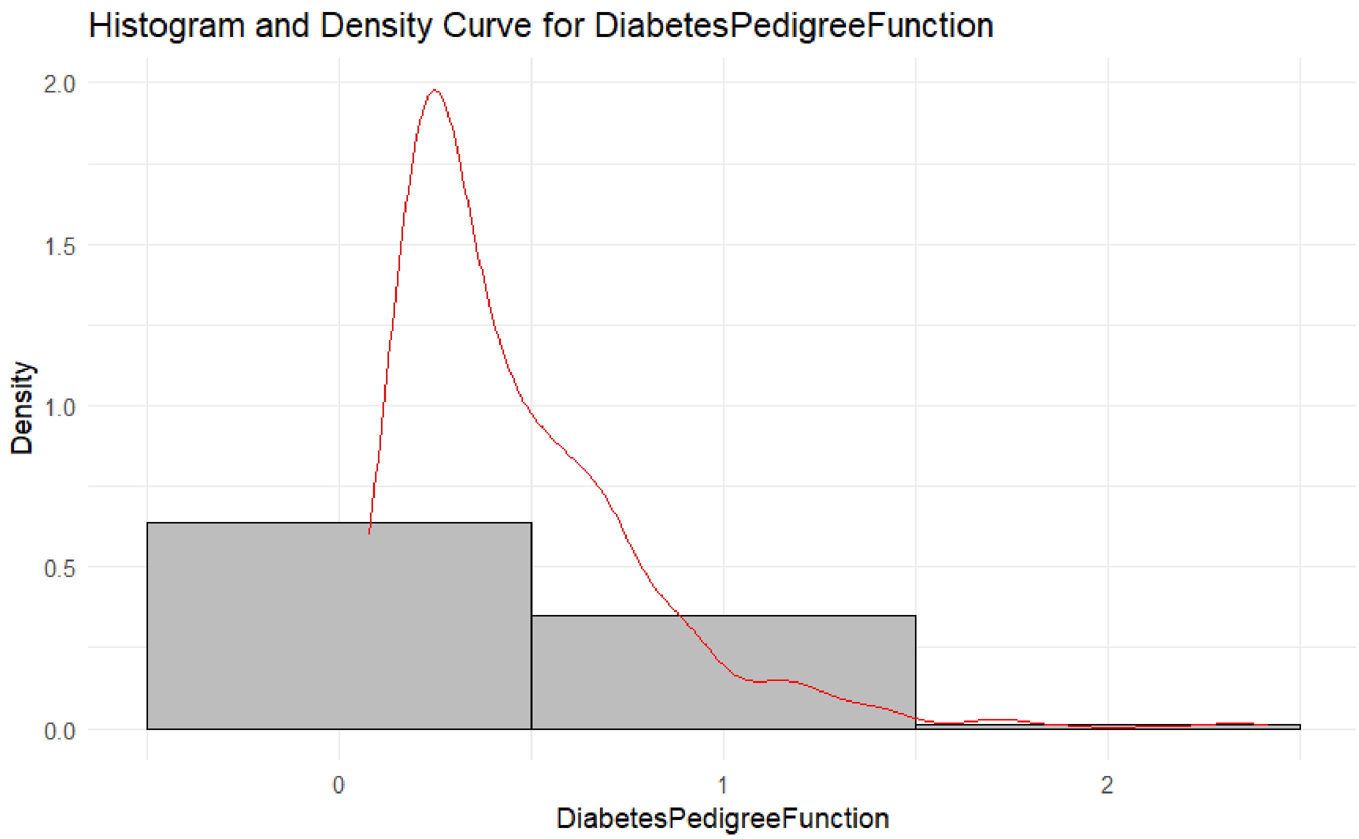
Hide

```
create_histogram_density_plot(diabetes, BMI, "Histogram and Density Curve for BMI")
```



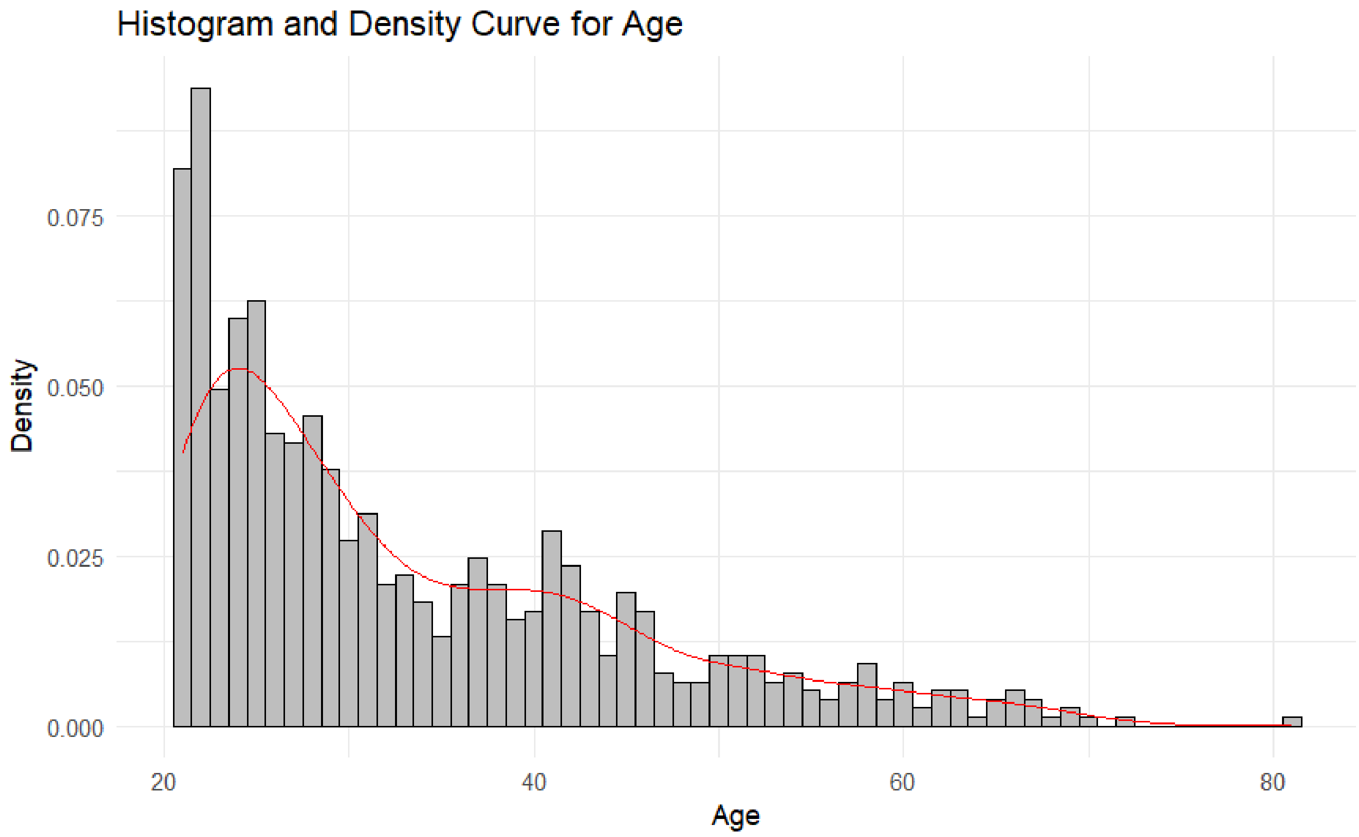
Hide

```
create_histogram_density_plot(diabetes, DiabetesPedigreeFunction, "Histogram and Density Curv
e for DiabetesPedigreeFunction")
```



Hide

```
create_histogram_density_plot(diabetes, Age, "Histogram and Density Curve for Age")
```



Hide

```
# Correlation data
datacorr <- cor(diabetes)
round(datacorr, 2)
```

```
                         Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI Diabete
sPedigreeFunction   Age Outcome
Pregnancies                     1.00    0.13          0.14         -0.08   -0.07 0.02
-0.03  0.54    0.22
Glucose                         0.13    1.00          0.15          0.06    0.33 0.22
0.14  0.26    0.47
BloodPressure                   0.14    0.15          1.00          0.21    0.09 0.28
0.04  0.24    0.07
SkinThickness                  -0.08    0.06          0.21          1.00    0.44 0.39
0.18 -0.11    0.07
Insulin                        -0.07    0.33          0.09          0.44    1.00 0.20
0.19 -0.04    0.13
BMI                             0.02    0.22          0.28          0.39    0.20 1.00
0.14  0.04    0.29
DiabetesPedigreeFunction       -0.03    0.14          0.04          0.18    0.19 0.14
1.00  0.03    0.17
Age                             0.54    0.26          0.24         -0.11   -0.04 0.04
0.03  1.00    0.24
Outcome                         0.22    0.47          0.07          0.07    0.13 0.29
0.17  0.24    1.00
```

Hide

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

Hide

```
# Calculate the correlation matrix
correlation_matrix <- cor(diabetes)
corrplot(correlation_matrix, method = "color", addCoef.col = "black")
# Save the plot as a PNG file
png("correlation_plot.png", width = 800, height = 800)
```

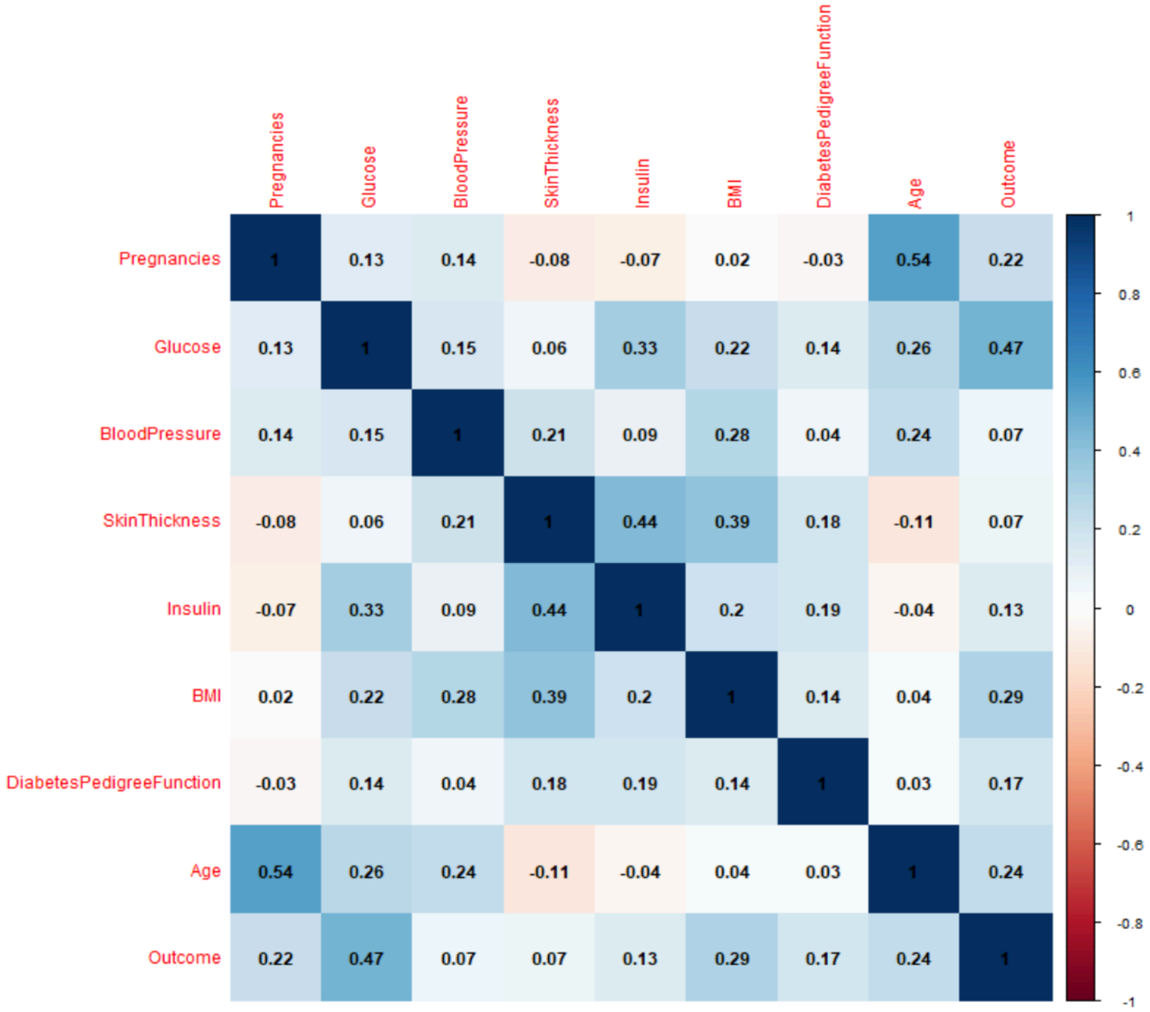
Hide

```
corrplot(correlation_matrix, method = "color", addCoef.col = "black")
dev.off()  # Close the PNG device
```

```
png
  2
```

Hide

```
library("png")
pp <- readPNG("correlation_plot.png")
plot.new()
rasterImage(pp,0,0,1,1)
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	-0.08	0.06	0.21	1	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1	0.2	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.2	1	0.14	0.04	0.29
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1

Hide

```
model <- glm(formula = Outcome ~ ., family = binomial(link="logit"), data = diabetes)
summary(model)
```

```

Call:
glm(formula = Outcome ~ ., family = binomial(link = "logit"), 
    data = diabetes)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.4046964  0.7166359 -11.728  < 2e-16 ***
Pregnancies               0.1231823  0.0320776   3.840 0.000123 ***
Glucose                   0.0351637  0.0037087   9.481  < 2e-16 ***
BloodPressure            -0.0132955  0.0052336  -2.540 0.011072 *  
SkinThickness             0.0006190  0.0068994   0.090 0.928515    
Insulin                  -0.0011917  0.0009012  -1.322 0.186065    
BMI                       0.0897010  0.0150876   5.945 2.76e-09 ***
DiabetesPedigreeFunction  0.9451797  0.2991475   3.160 0.001580 ** 
Age                       0.0148690  0.0093348   1.593 0.111192    
---
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 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

Hide

```
library(car)
```

```
Loading required package: carData
```

Hide

```
# Calculate VIF
vif_values <- car::vif(model)

# Print the VIF values
print(vif_values)
```

```
             Pregnancies                  Glucose            BloodPressure            SkinThi
ckness                  Insulin                      BMI 
                1.408434                 1.214367                 1.175283                 1.
522040                 1.467918                 1.220416 
DiabetesPedigreeFunction                      Age 
                1.034318                 1.502069 
```

Hide

```
anova(model, test="Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Outcome

Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                       767     993.48              
Pregnancies               1   37.274       766     956.21 1.026e-09 ***
Glucose                   1  171.260       765     784.95 < 2.2e-16 ***
BloodPressure             1    0.888       764     784.06 0.3460418    
SkinThickness             1    3.999       763     780.06 0.0455212 *  
Insulin                   1    1.972       762     778.09 0.1602210    
BMI                       1   41.243       761     736.85 1.344e-10 ***
DiabetesPedigreeFunction  1   10.880       760     725.97 0.0009719 ***
Age                       1    2.522       759     723.45 0.1122535    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
model2 <- glm(formula = Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction, family = binomial(link="logit"), data = diabetes)
summary(model2)
```

```


Call:
glm(formula = Outcome ~ Pregnancies + Glucose + SkinThickness + 
    BMI + DiabetesPedigreeFunction, family = binomial(link = "logit"), 
    data = diabetes)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.471433   0.661394 -12.808  < 2e-16 ***
Pregnancies               0.139757   0.027159   5.146 2.66e-07 ***
Glucose                   0.033790   0.003343  10.107  < 2e-16 ***
SkinThickness            -0.006715   0.006012  -1.117   0.2640    
BMI                       0.083831   0.014800   5.664 1.48e-08 ***
DiabetesPedigreeFunction  0.944013   0.295557   3.194   0.0014 ** 
---
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: 733.06  on 762  degrees of freedom
AIC: 745.06

Number of Fisher Scoring iterations: 5
```

Hide

```
library(car)

# Calculate VIF
vif_values <- car::vif(model2)

# Print the VIF values
print(vif_values)
```

```
             Pregnancies                  Glucose            SkinThickness                      
BMI DiabetesPedigreeFunction 
                1.026442                 1.002796                 1.182071                 1.
164280                 1.027884 
```

Hide

```
anova(model2, test="Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Outcome

Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                       767     993.48              
Pregnancies               1   37.274       766     956.21 1.026e-09 ***
Glucose                   1  171.260       765     784.95 < 2.2e-16 ***
SkinThickness             1    3.030       764     781.92  0.081761 .  
BMI                       1   38.321       763     743.60 6.003e-10 ***
DiabetesPedigreeFunction  1   10.541       762     733.06  0.001167 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
model3 <- glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction, fam
ily = binomial(link="logit"), data = diabetes)
summary(model3)
```

```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction, 
    family = binomial(link = "logit"), data = diabetes)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.415851   0.656908 -12.811  < 2e-16 ***
Pregnancies               0.141926   0.027105   5.236 1.64e-07 ***
Glucose                   0.033826   0.003345  10.112  < 2e-16 ***
BMI                       0.078097   0.013771   5.671 1.42e-08 ***
DiabetesPedigreeFunction  0.901294   0.291696   3.090    0.002 ** 
---
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: 734.31  on 763  degrees of freedom
AIC: 744.31

Number of Fisher Scoring iterations: 5
```

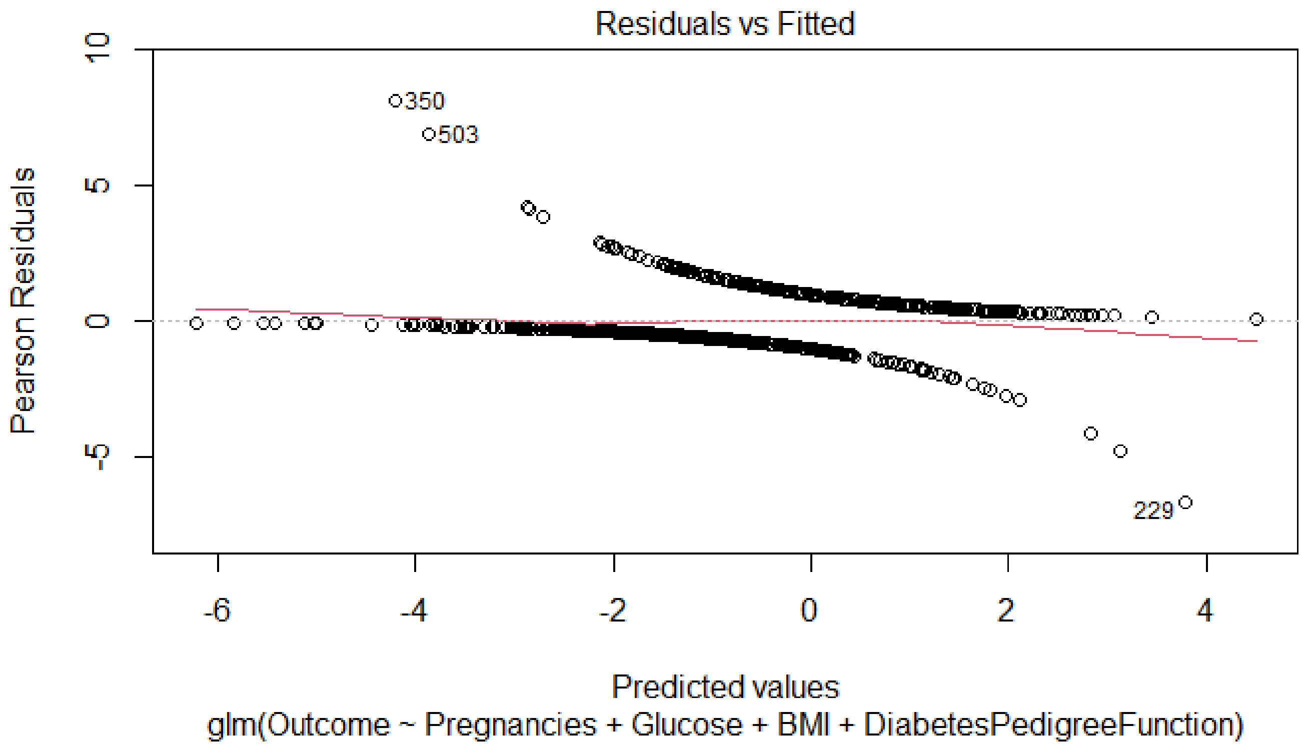
Hide

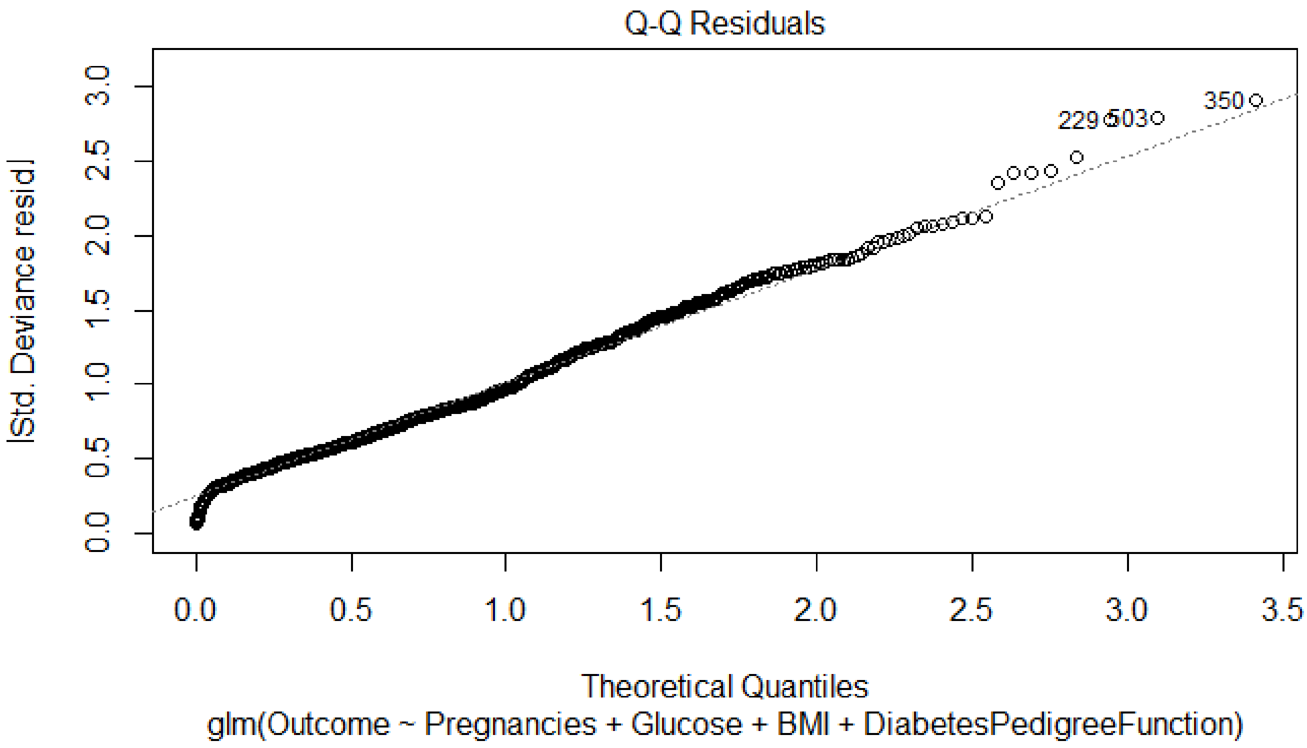
```
library(car)

# Calculate VIF
vif_values <- car::vif(model3)

# Print the VIF values
print(vif_values)
```

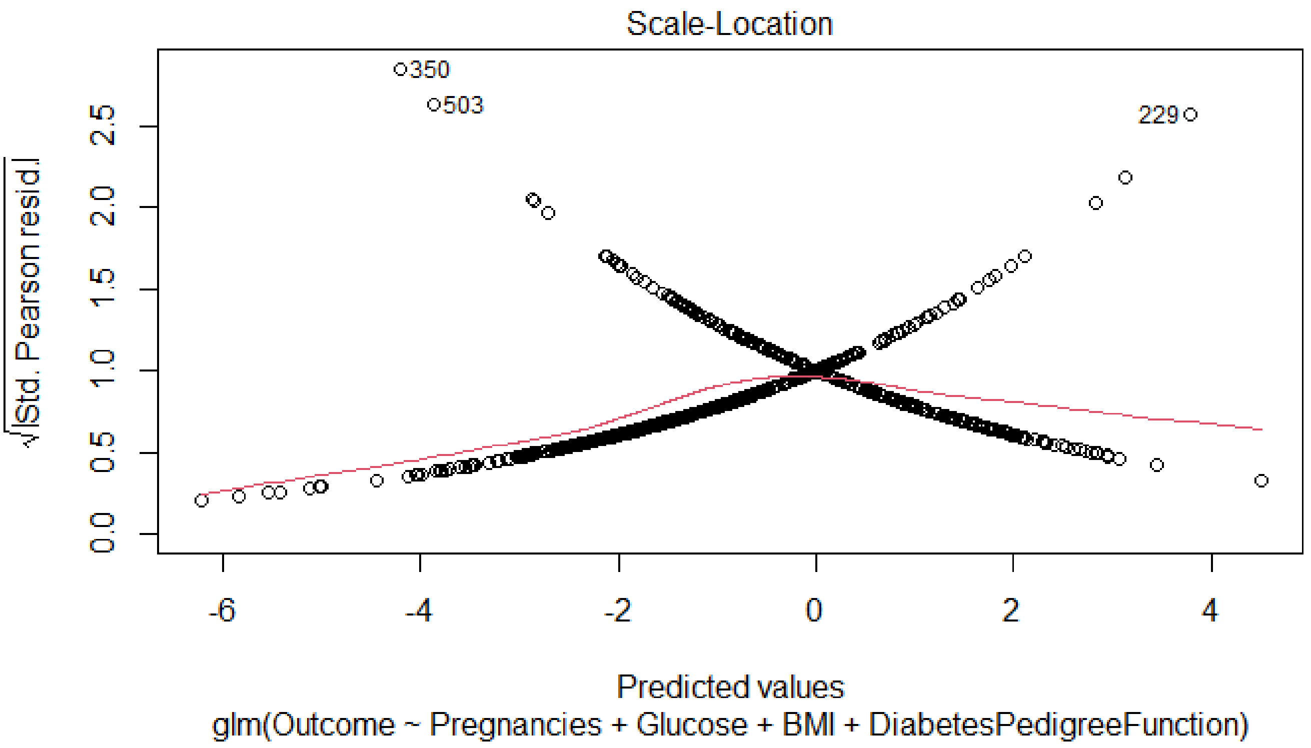
```
             Pregnancies                  Glucose                      BMI DiabetesPedigreeFu
nction 
                1.022292                 1.002622                 1.018577                 1.
009126 
```

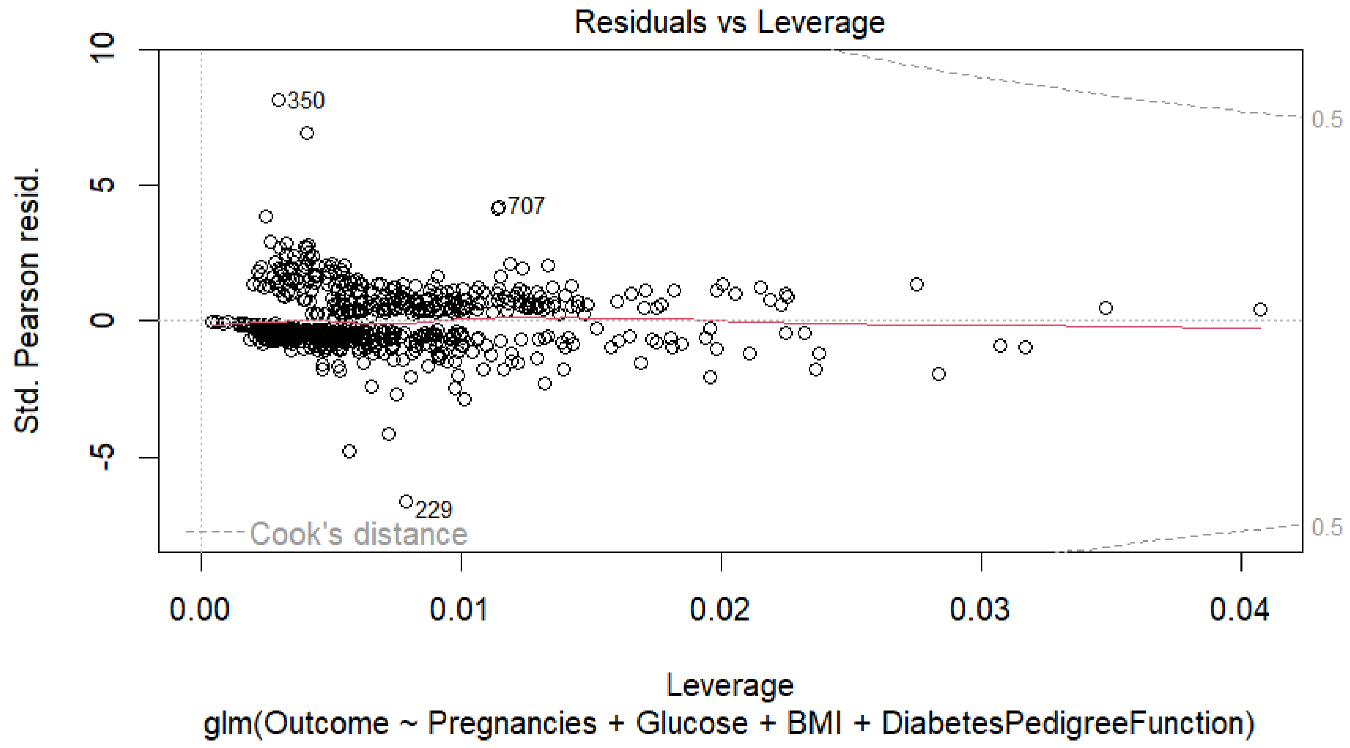
Hide

```
plot(model3)
```


Residuals vs Fitted
Pearson Residuals
10
5
0
-5
350
503
229
-6
-4
-2
0
2
4
Predicted values
glm(Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction)


Q-Q Residuals
|Std. Deviance resid|
3.0
2.5
2.0
1.5
1.0
0.5
0.0
229
503
350
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Theoretical Quantiles
glm(Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction)





Hide

```
residuals <- residuals(model3)
shapiro.test(residuals)
```

```
    Shapiro-Wilk normality test

data:  residuals
W = 0.93146, p-value < 2.2e-16
```

Hide

```
library(lmtest)
```

```
Loading required package: zoo

Attaching package: ‘zoo’

The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric
```

Hide

```
#perform Breusch-Pagan Test
bptest(model3)
```

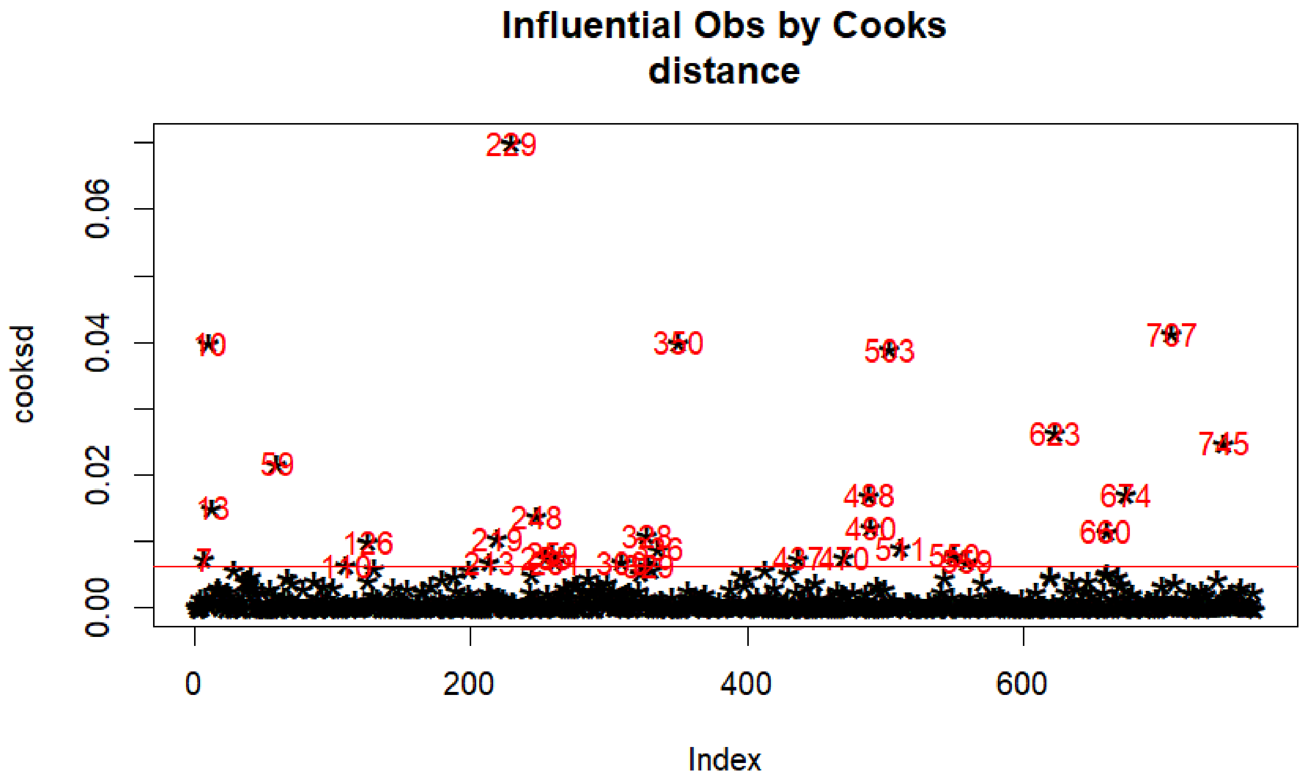
```
    studentized Breusch-Pagan test

data:  model3
BP = 37.983, df = 4, p-value = 1.129e-07
```

Hide

```
#Cook distance
cooksd <- cooks.distance(model3)
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks
distance") # plot cook's distance
abline(h = 4*mean(cooksd, na.rm=T), col="red") # add cutoff line
```

Hide

```
text(x=1:length(cooksd)+1, y=cooksd,
labels=ifelse(cooksd>4*mean(cooksd,
na.rm=T),names(cooksd),""), col="red") #add labels
```



Hide

```
# influential row numbers
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd,na.rm=T))])
influential
```

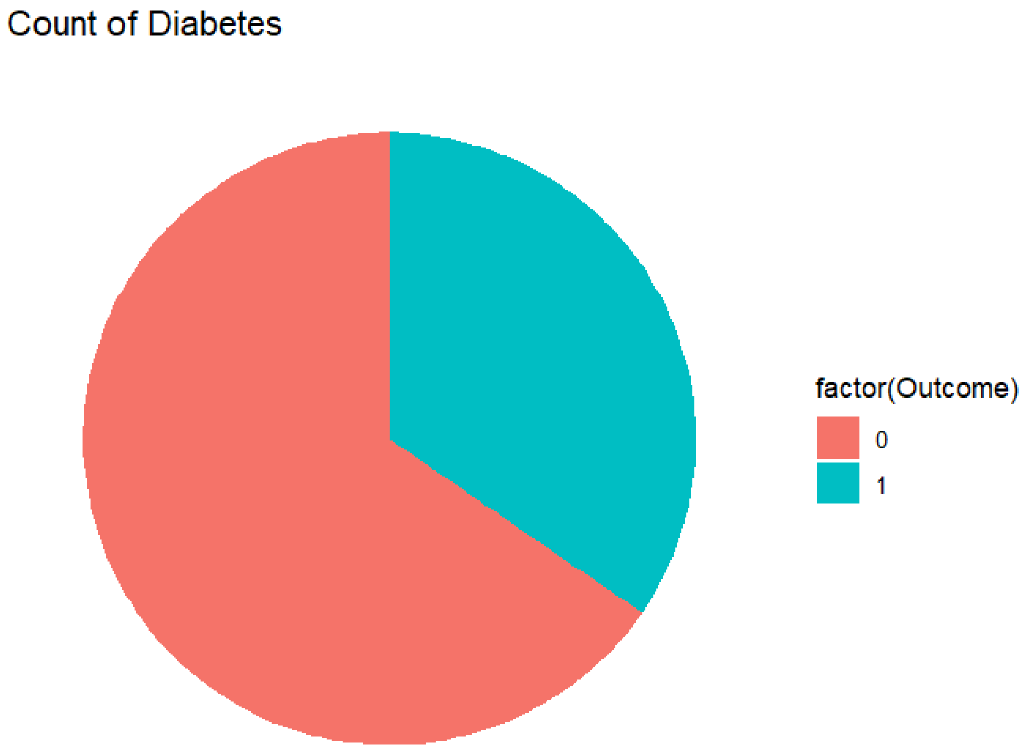
```
 [1]   7  10  13  59 110 126 213 219 229 248 255 259 261 309 328 329 336 350 437 470 488 490
503 511 550 559 623 660 674 707 745
```

Hide

```
diabetes <- diabetes[-influential, ]
```

Hide

```
# Create a pie chart
p <- ggplot(diabetes, aes(x = "", fill = factor(Outcome))) +
  geom_bar(width = 1, stat = "count") +
  coord_polar("y") +
  labs(title = "Count of Diabetes", x = NULL, y = NULL) +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank())

# Show the plot
print(p)
```



Hide

```
library(tidyverse)
```

```
Warning: package ‘tidyverse’ was built under R version 4.3.3
Warning: package ‘tidyr’ was built under R version 4.3.3
Warning: package ‘readr’ was built under R version 4.3.3
Warning: package ‘dplyr’ was built under R version 4.3.3
```

```
── Attaching core tidyverse packages ──────────────────────────────────────────────────────
──────────────────────────────────────────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ lubridate 1.9.3     ✓ tibble    3.2.1
✓ purrr     1.0.2     ✓ tidyr     1.3.1
── Conflicts ──────────────────────────────────────────────────────────────────────────────
────────────────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
✖ dplyr::recode() masks car::recode()
✖ purrr::some()   masks car::some()
ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts t
o become errors
```

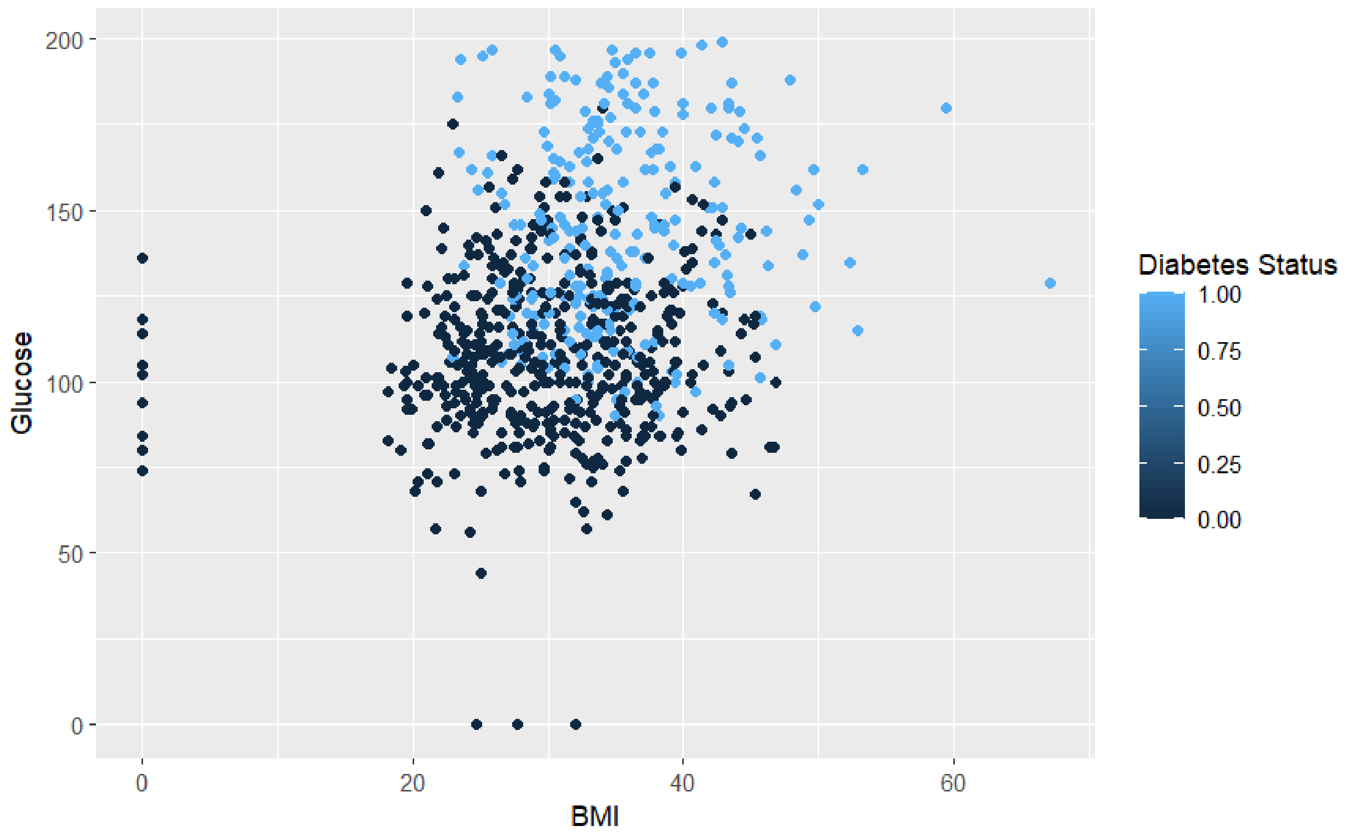
Hide

```
# Load data
data(diabetes)
```

```
Warning in data(diabetes) : data set ‘diabetes’ not found
```

Hide

```
ggplot(diabetes, aes(x = BMI, y = Glucose , color = Outcome)) +
  geom_point() +
  labs(x = "BMI", y = "Glucose", color = "Diabetes Status")
```



Hide

```
#install.packages("heatmaply")
```

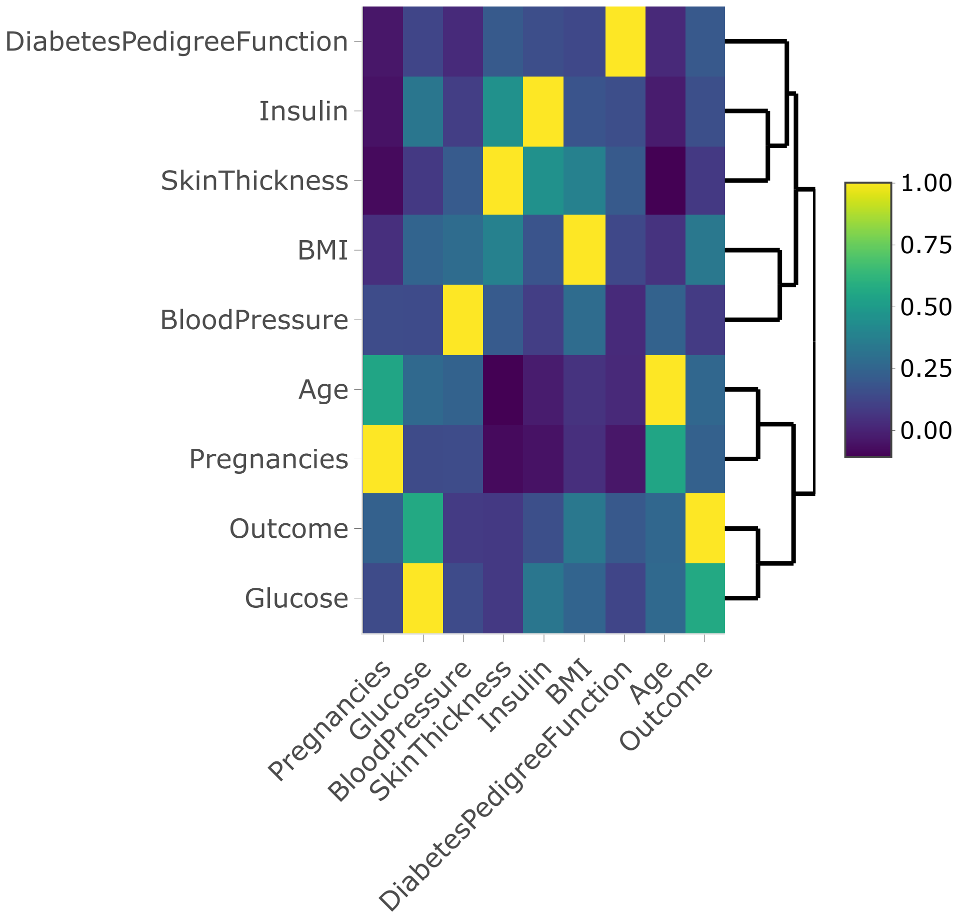
Hide

```
# Load the necessary libraries
library(pairsD3)
library(Hmisc)
library(heatmaply)

# Interactive pairs plot
pairsD3::shinypairs(diabetes, group = "Outcome")
```

```
`shiny::dataTableOutput()` is deprecated as of shiny 1.8.1.
Please use `DT::DTOutput()` instead.
See <https://rstudio.github.io/DT/shiny.html> for more information.

Listening on http://127.0.0.1:6390
`shiny::renderDataTable()` is deprecated as of shiny 1.8.1.
Please use `DT::renderDT()` instead.
See <https://rstudio.github.io/DT/shiny.html> for more information.
```

Hide

```
# Heatmap of correlations
heatmaply::heatmaply(cor(diabetes), Colv = NA)
```



Hide

```
# Summary statistics
Hmisc::describe(diabetes, digits = 1)
```

```
diabetes

 9  Variables      737  Observations
--------------------------------------------------------------------------------------------
--------
Pregnancies
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0       17    0.986        4        4        0        0        1        3
6
     .90      .95
       9       10

Value          0     1     2     3     4     5     6     7     8     9    10    11    12    1
3    14
Frequency    105   133   102    71    66    55    47    44    36    28    21    10     6
9     2
Proportion 0.142 0.180 0.138 0.096 0.090 0.075 0.064 0.060 0.049 0.038 0.028 0.014 0.008 0.01
2 0.003

Value         15    17
Frequency      1     1
Proportion 0.001 0.001

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------------------
--------
Glucose
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0      135        1      121       34       79       86       99      117
139
     .90      .95
     165      180

lowest :   0  44  56  57  61, highest: 195 196 197 198 199
--------------------------------------------------------------------------------------------
--------
BloodPressure
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0       46    0.998       69       19       40       54       62       72
80
     .90      .95
      88       90

lowest :   0  24  30  38  40, highest: 106 108 110 114 122
--------------------------------------------------------------------------------------------
--------
SkinThickness
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0       51    0.973       20       18        0        0        0       23
32
     .90      .95
```

```
      40       44

lowest :  0  7  8 10 11, highest: 54 56 60 63 99
--------------------------------------------------------------------------------------------
--------
Insulin
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0      181    0.884       78      105        0        0        0       29
126
     .90      .95
     207      291

lowest :   0  14  15  16  18, highest: 543 545 579 600 846
--------------------------------------------------------------------------------------------
--------
BMI
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0      241        1       32        8       22       24       27       32
36
     .90      .95
      41       44

lowest : 0    18.2 18.4 19.1 19.3, highest: 52.3 52.9 53.2 59.4 67.1
--------------------------------------------------------------------------------------------
--------
DiabetesPedigreeFunction
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0      500        1      0.5      0.3      0.1      0.2      0.2      0.4
0.6
     .90      .95
     0.9      1.1

lowest : 0.078 0.084 0.085 0.088 0.089, highest: 1.731 1.893 2.137 2.288 2.42
--------------------------------------------------------------------------------------------
--------
Age
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75
     737        0       52    0.997       33       13       21       22       24       29
40
     .90      .95
      51       58

lowest : 21 22 23 24 25, highest: 68 69 70 72 81
--------------------------------------------------------------------------------------------
--------
Outcome
       n  missing distinct     Info      Sum     Mean      Gmd
     737        0        2    0.679      255      0.3      0.5

--------------------------------------------------------------------------------------------
--------
```