Diabetes Dataset

Surya Bhosale

30/03/2020

Importing Required Library:

```
library(ggplot2)
library(astsa)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(e1071)
```

Loading Dataset:

```
diabetes <- read.csv("~/Desktop/diabetes.csv")
View(diabetes)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/
## Resources/modules/R_de.so'' had status 69
```

```
summary(diabetes)
```

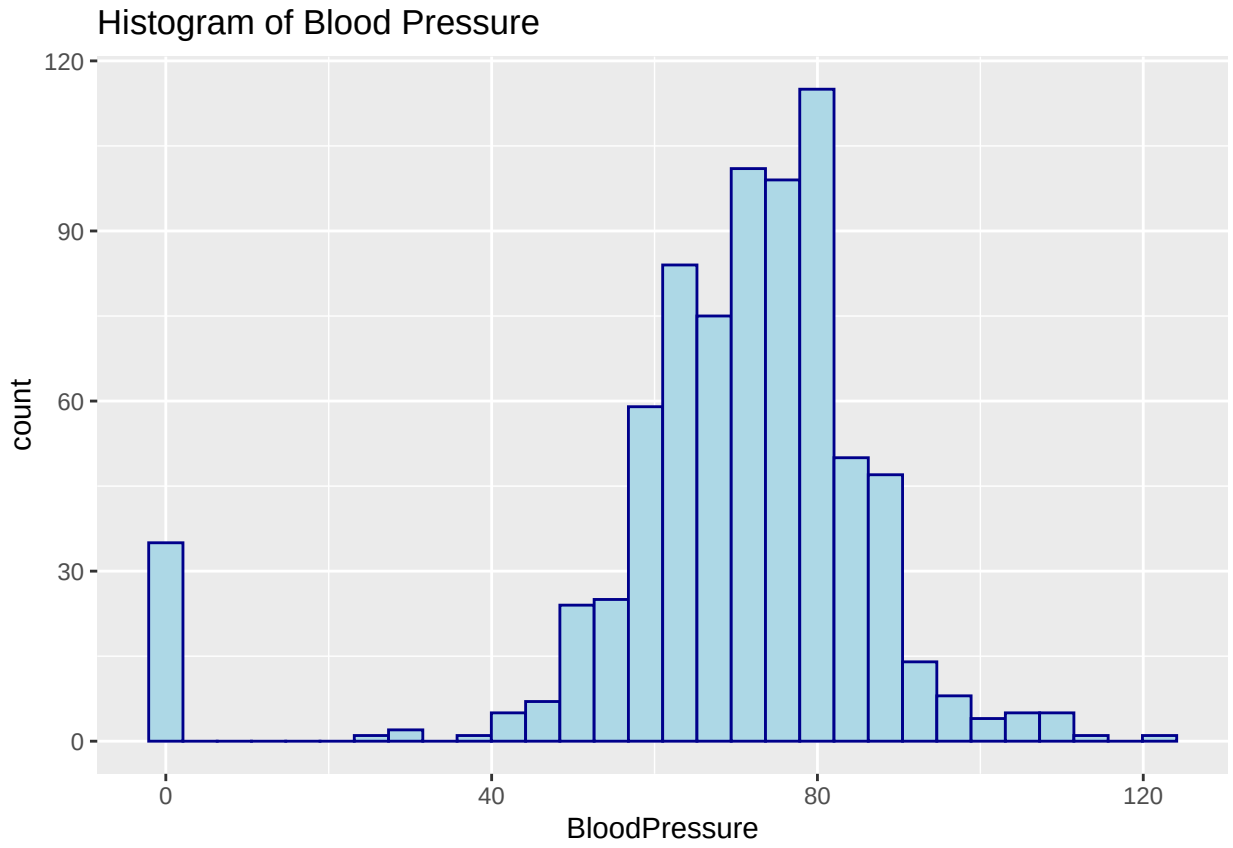
```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :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           BMI        DiabetesPedigreeFunction      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
##     Outcome
##  Min.   :0.000
```

```
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

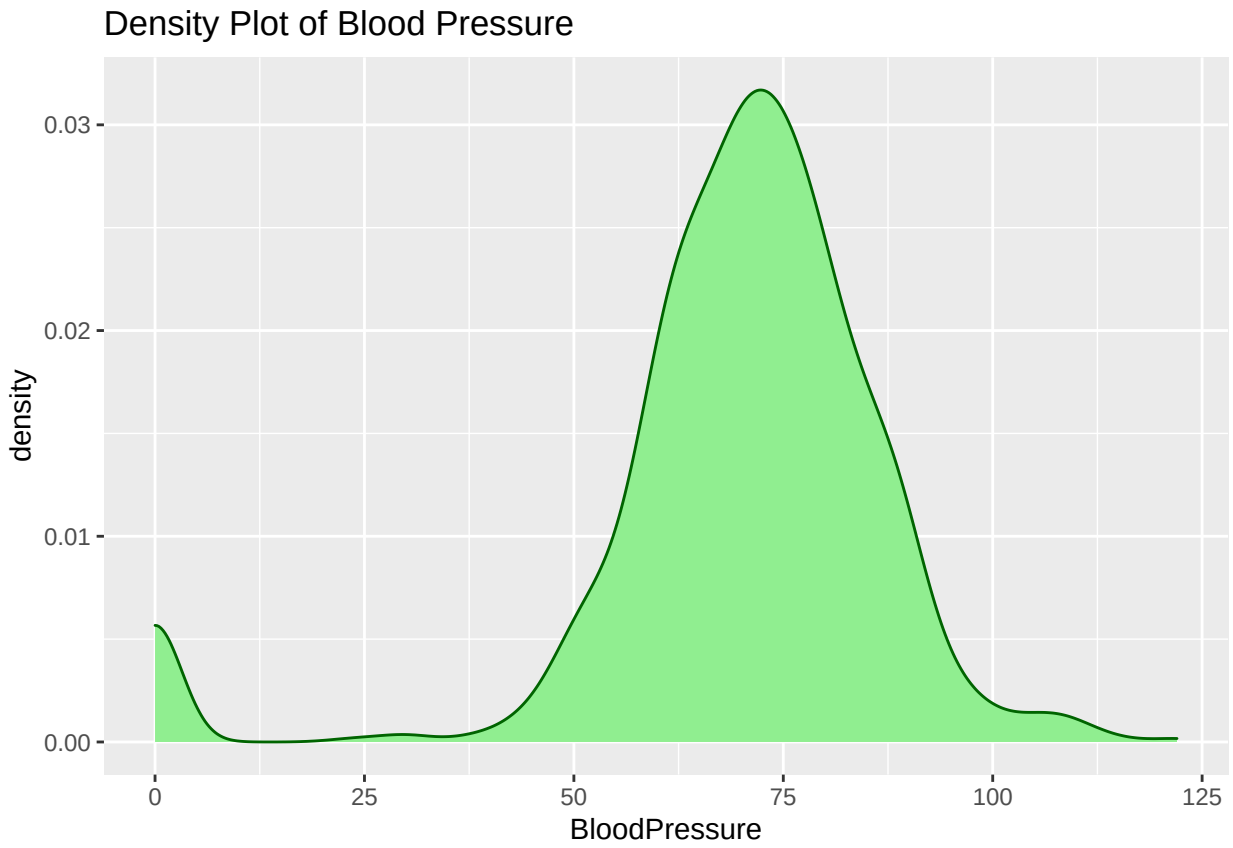
Data Visulatisation For BP:

```
ggplot(diabetes, aes(x=BloodPressure)) +
  geom_histogram(color = "darkblue", fill = "lightblue") +
  labs(title = "Histogram of Blood Pressure")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

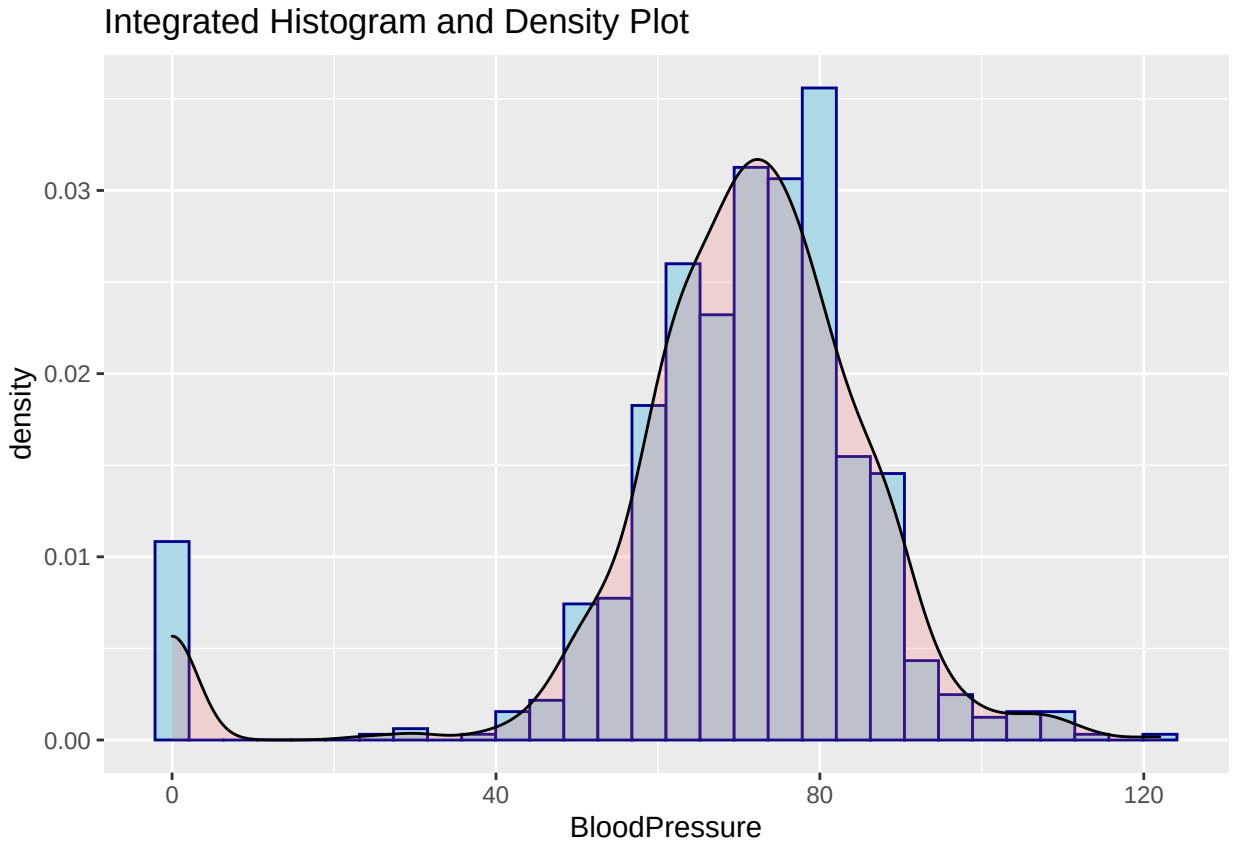



```
ggplot(diabetes, aes(x=BloodPressure)) +
  geom_density(color = "darkgreen", fill = "lightgreen") +
  labs(title = "Density Plot of Blood Pressure")
```

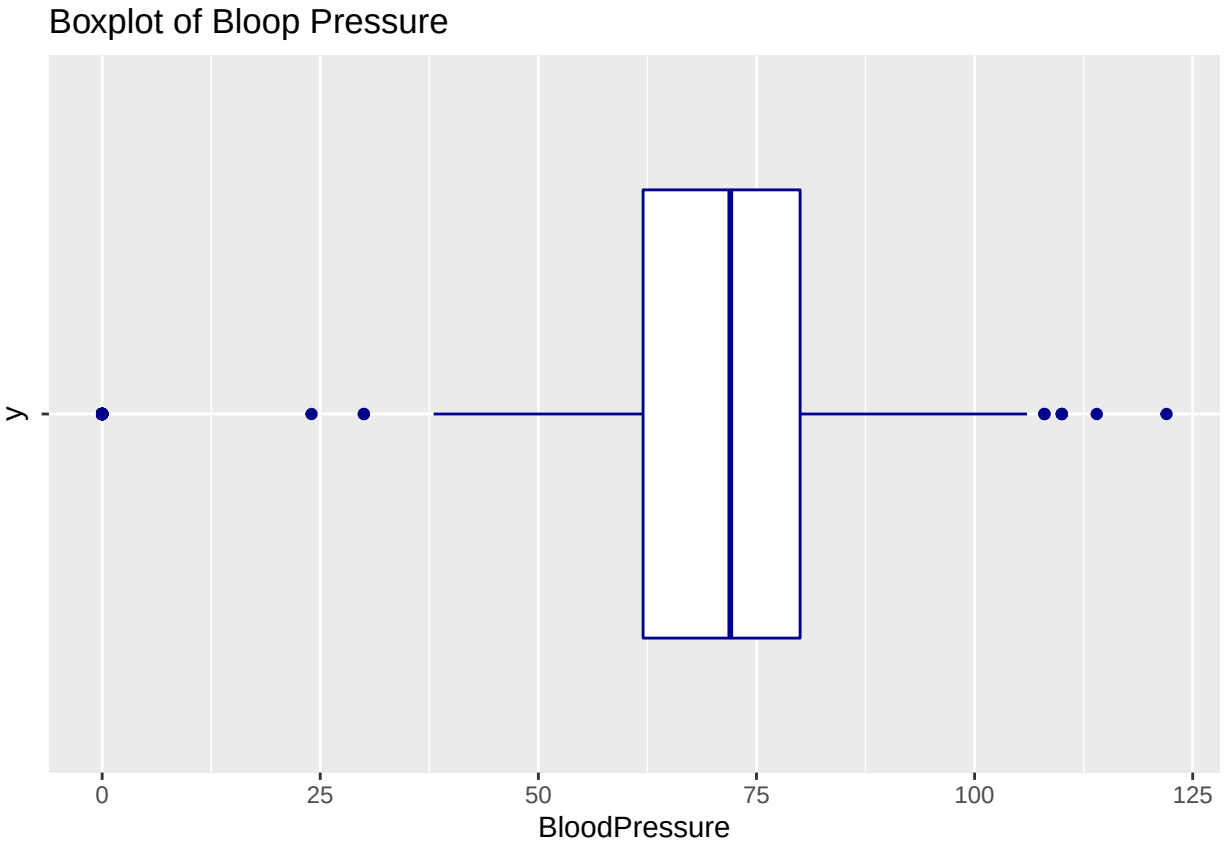



```
ggplot(diabetes, aes(x=BloodPressure)) +
  geom_histogram(aes(y=..density..), colour="darkblue", fill="lightblue")+
  geom_density(alpha=.2, fill="#FF6666") +
  labs(title = "Integrated Histogram and Density Plot")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(diabetes, aes(x = BloodPressure, y = "")) +
  geom_boxplot( color = "darkblue") + labs(title = "Boxplot of Bloop Pressure")
```




```
skewness(diabetes$BloodPressure)
```

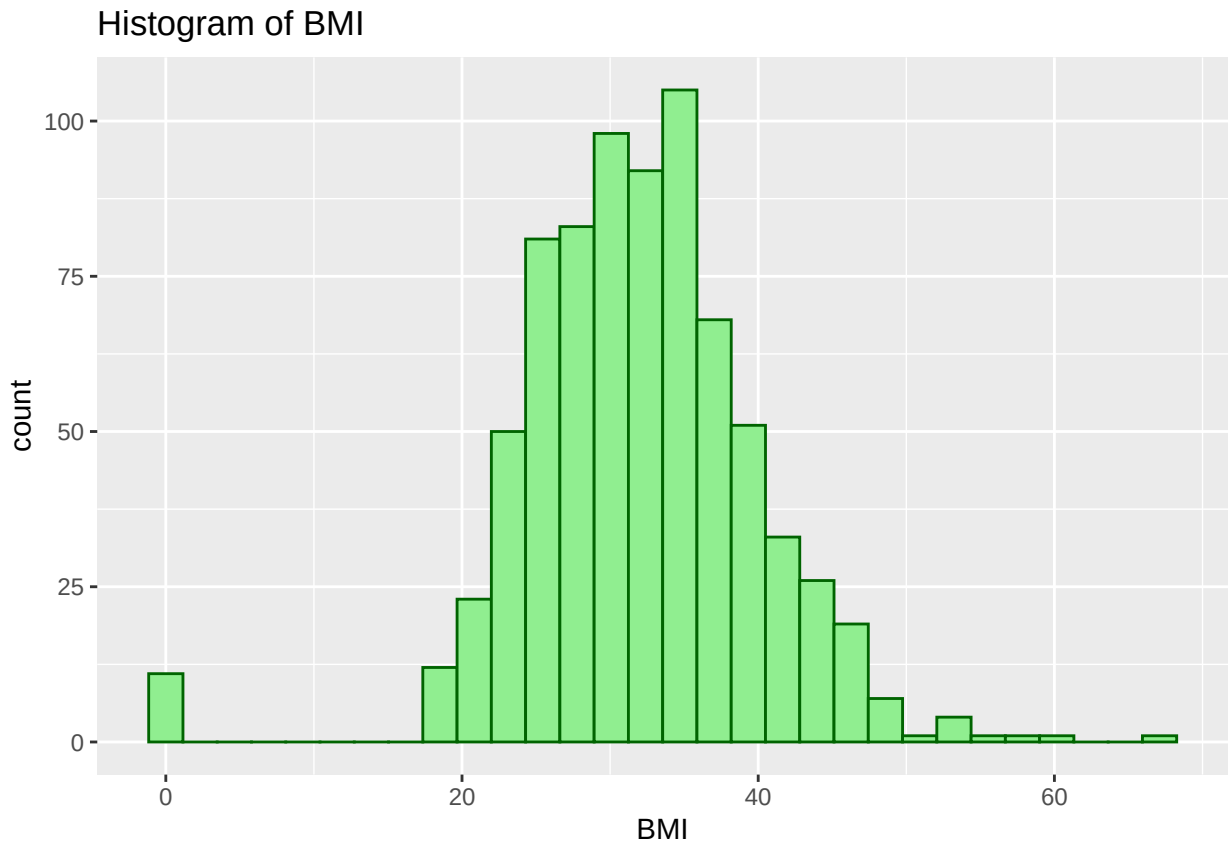
```
## [1] -1.836413
```

Comment: The modal value for blood pressure in the diabetes dataset appears to be near the 80 mark. the density plot in the shows that the distribution is left skewed, with extreme values to the left.There are 35 values with BP as 0. The skewness test resulted in the value of -1.836. This substantiates that the parameter is left skewed. However, this must be an error as it is impossible unless the indivdual is deceased. The scope of this assignment does not ask for rectifying the discrepency, therefore leaving abstaining from initiating correction. The boxplot aids in uncovering the median blodo pressure which is 72.

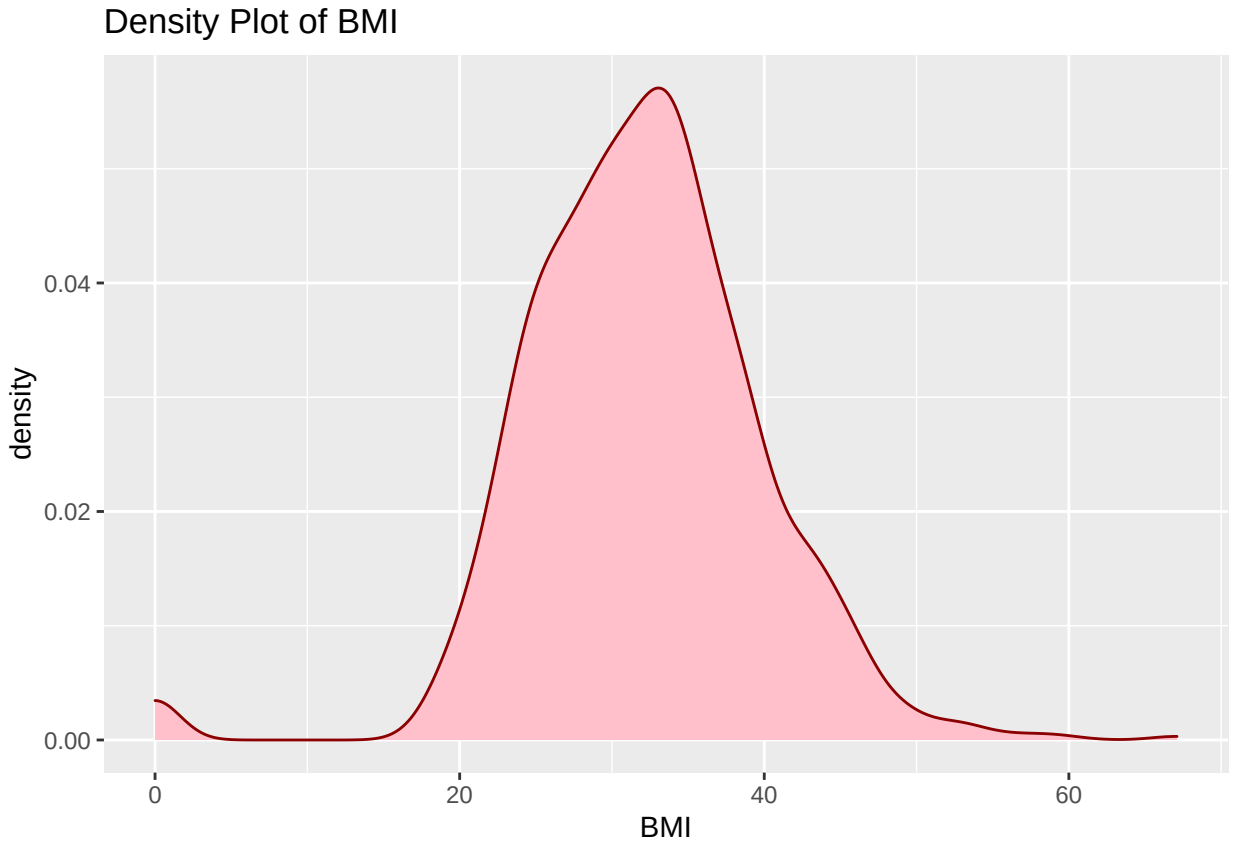
Data Visulatisation For BMI:

```
ggplot(diabetes, aes(x=BMI)) +
  geom_histogram(color = "darkgreen", fill = "lightgreen") +
  labs(title = "Histogram of BMI")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

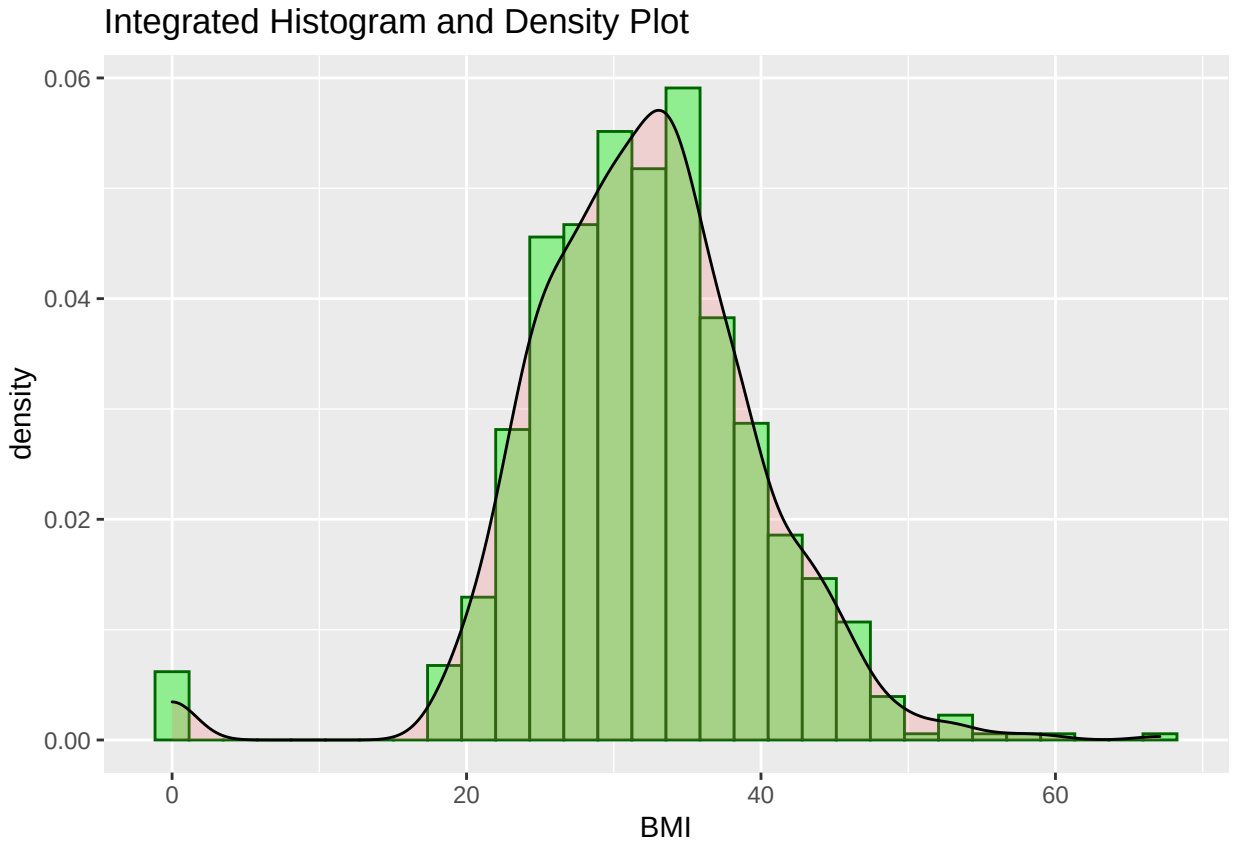



```
ggplot(diabetes, aes(x=BMI)) +
  geom_density(color = "darkred", fill = "pink") +
  labs(title = "Density Plot of BMI")
```

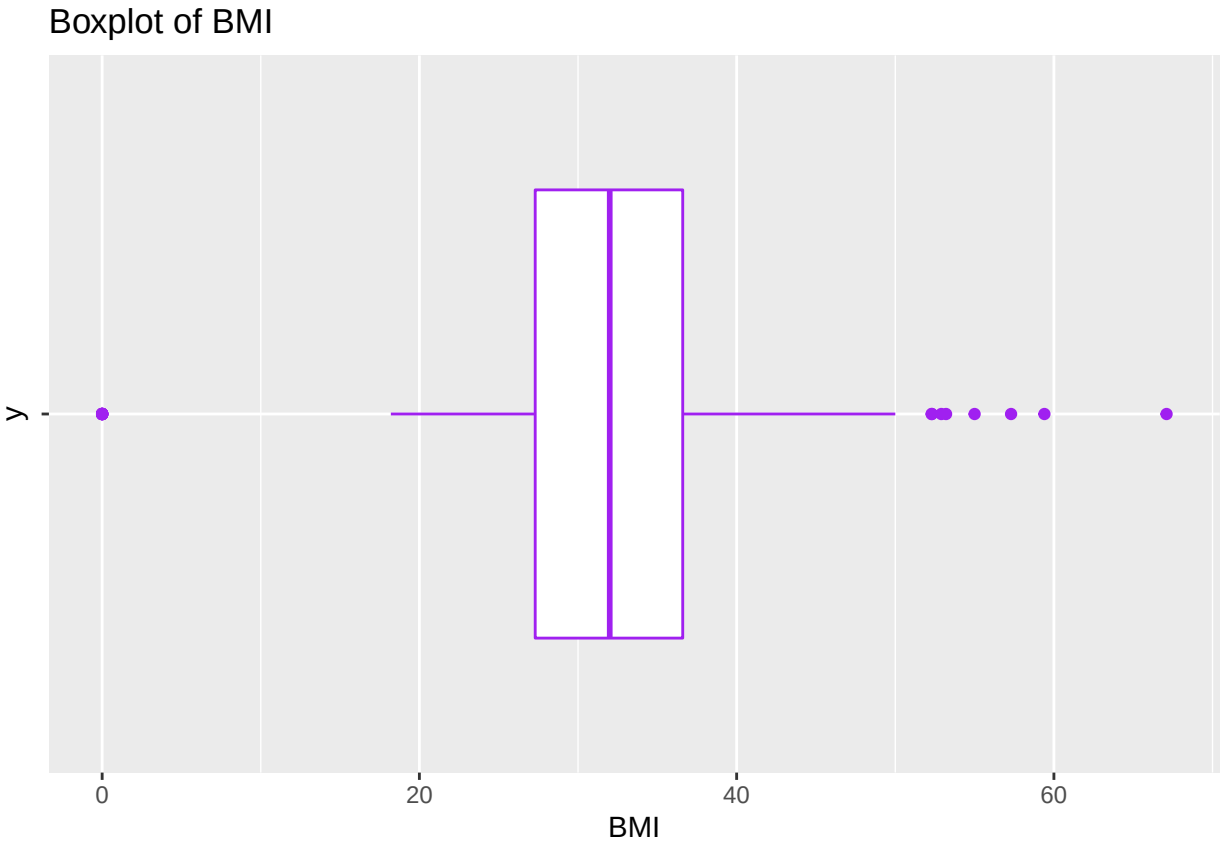



```
ggplot(diabetes, aes(x=BMI)) +
  geom_histogram(aes(y=..density..), colour="darkgreen", fill="lightgreen")+
  geom_density(alpha=.2, fill="#FF6666") +
  labs(title = "Integrated Histogram and Density Plot")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(diabetes, aes(x = BMI, y = "")) +
  geom_boxplot( color = "purple") +
  labs(title = "Boxplot of BMI")
```




```
kurtosis(diabetes$BMI)
```

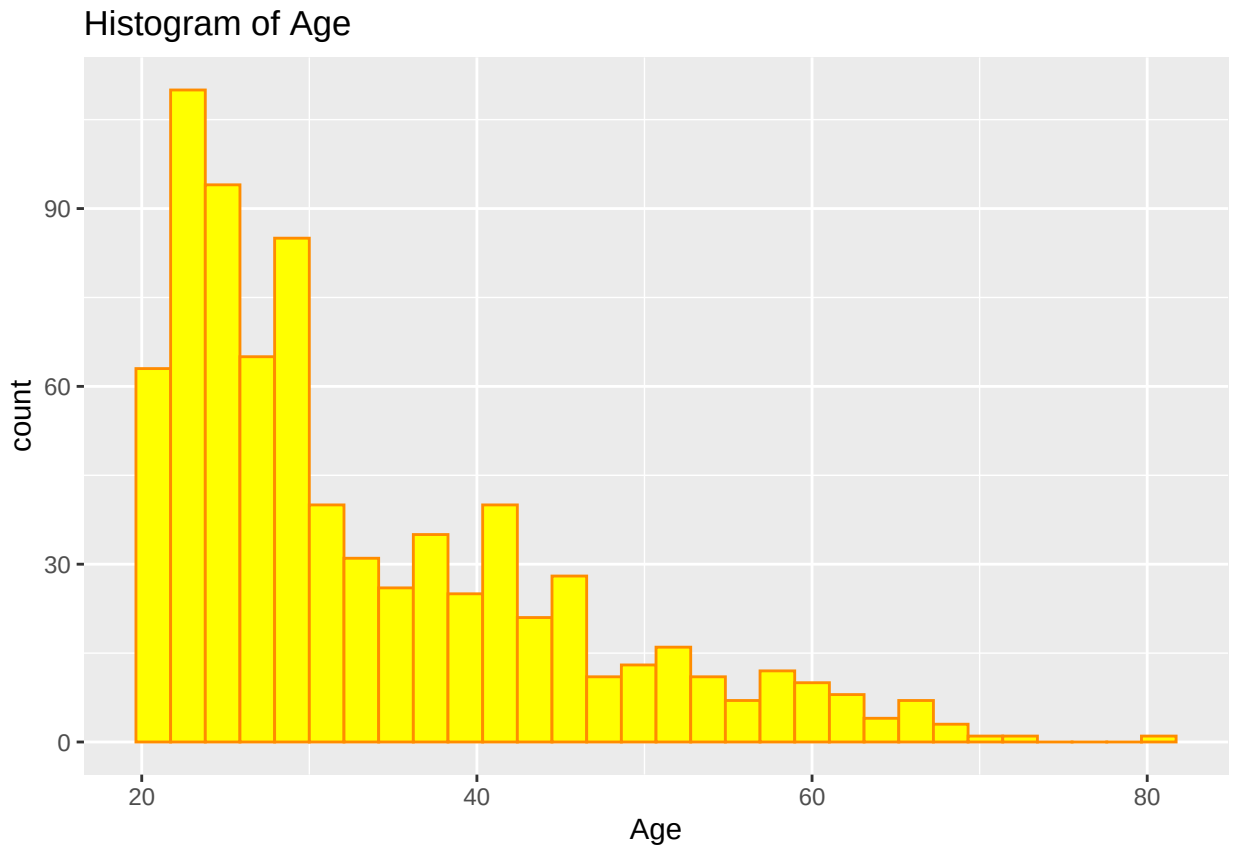
```
## [1] 3.244963
```

Comment: The histogoram of BMI depicts that the modal value of BMI in this dataset is in the range of 36-38. The density plot appears to be leopkurtic. This claim is supported by th ekurtosis test which resulted in the value of 3.244 (greater than 3), therefore we may assume that the distribution does not abide the norms of normal distribution. This is due to the high concentration of data in the range of 30-36. The boxplot aids in the noting that the median BMI is 32.

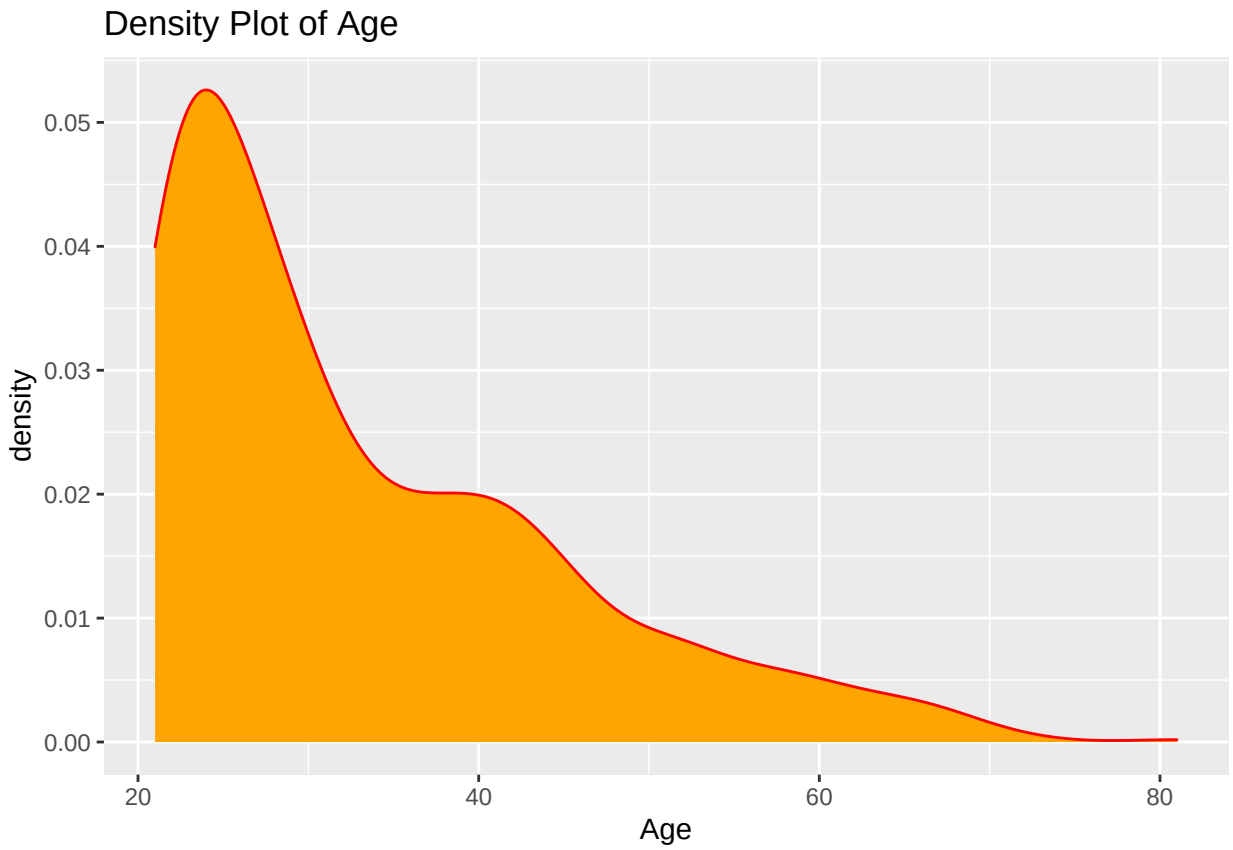
Data Visulatisation For Age:

```
ggplot(diabetes, aes(x=Age)) +
  geom_histogram(color = "darkorange", fill = "yellow") +
  labs(title = "Histogram of Age")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

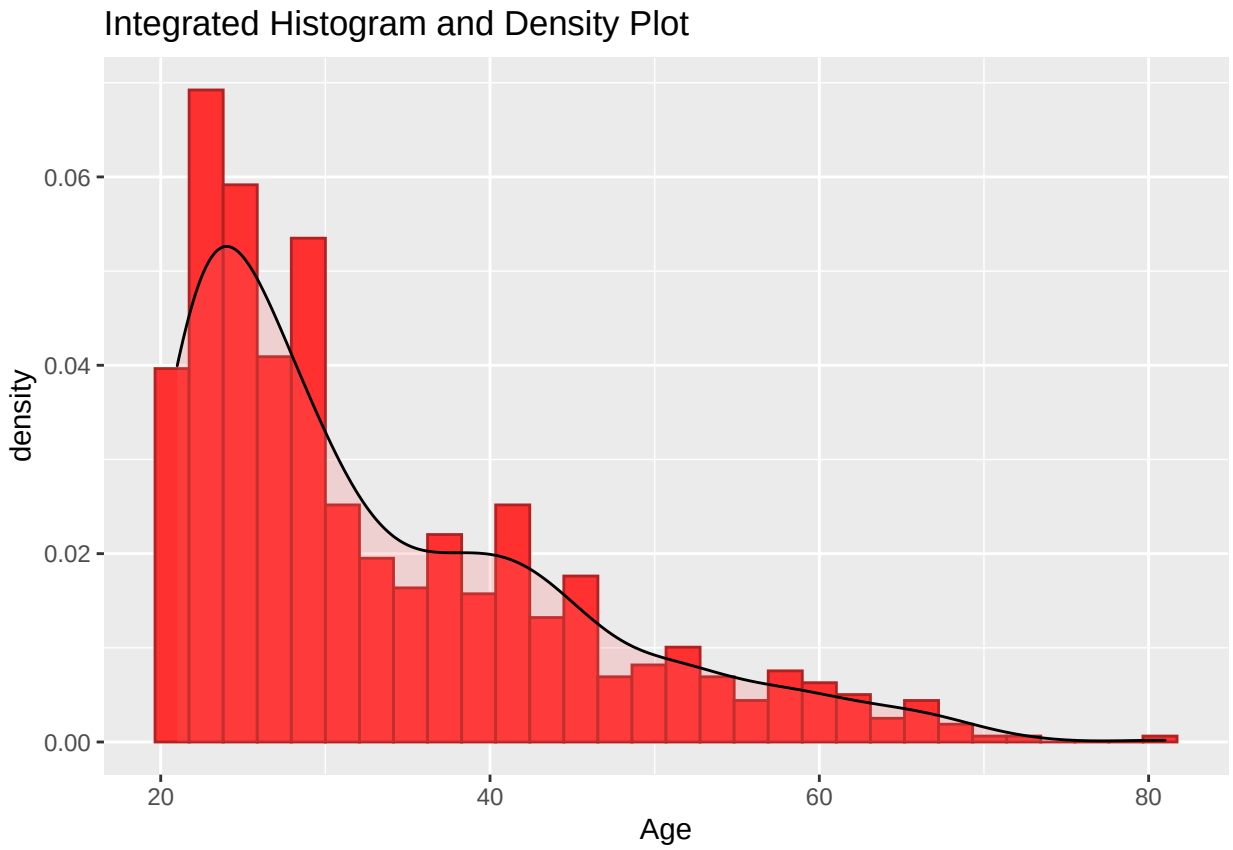



```
ggplot(diabetes, aes(x=Age)) +
  geom_density(color = "red", fill = "orange") +
  labs(title = "Density Plot of Age")
```

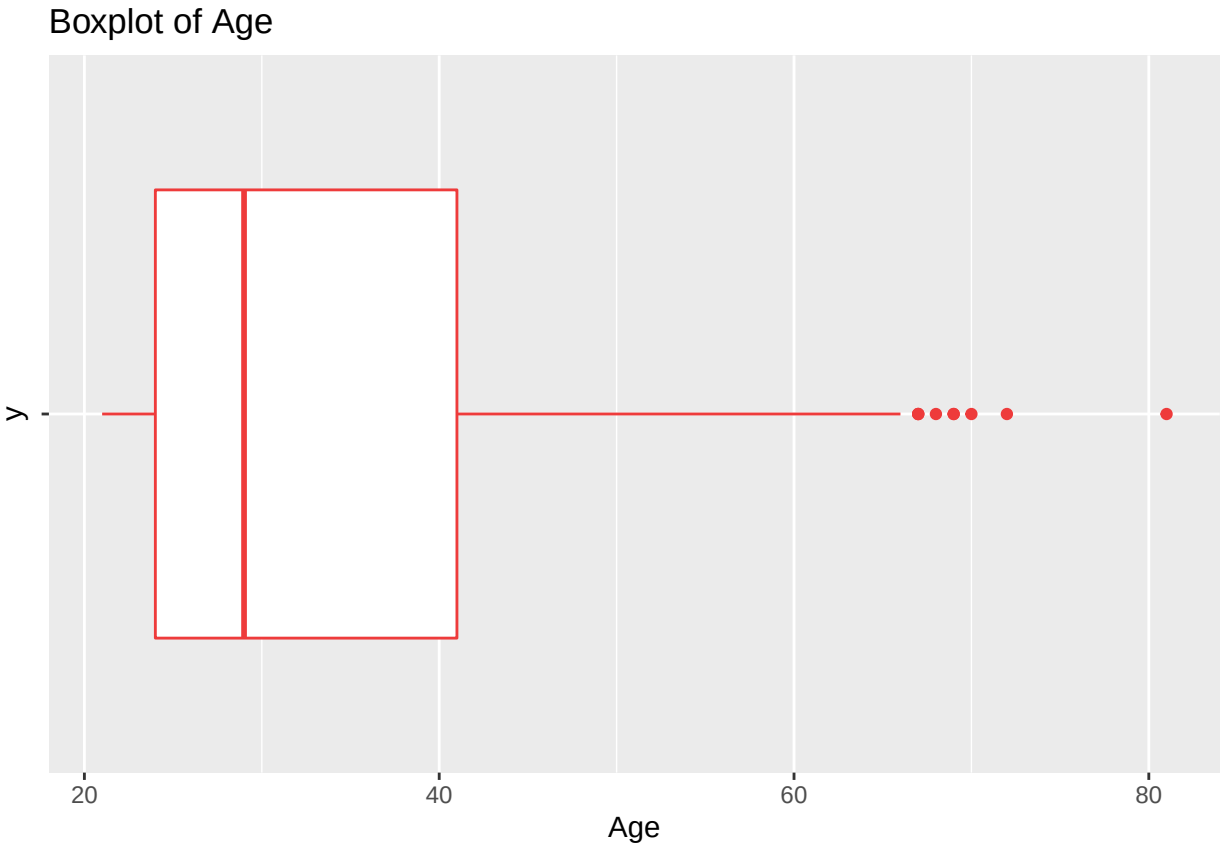



```
ggplot(diabetes, aes(x=Age)) +
  geom_histogram(aes(y=..density..), colour="firebrick", fill="firebrick1")+
  geom_density(alpha=.2, fill="#FF6666") +
  labs(title = "Integrated Histogram and Density Plot")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(diabetes, aes(x = Age, y = "")) +
  geom_boxplot( color = "brown2") +
  labs(title = "Boxplot of Age")
```




```
skewness(diabetes$Age)
```

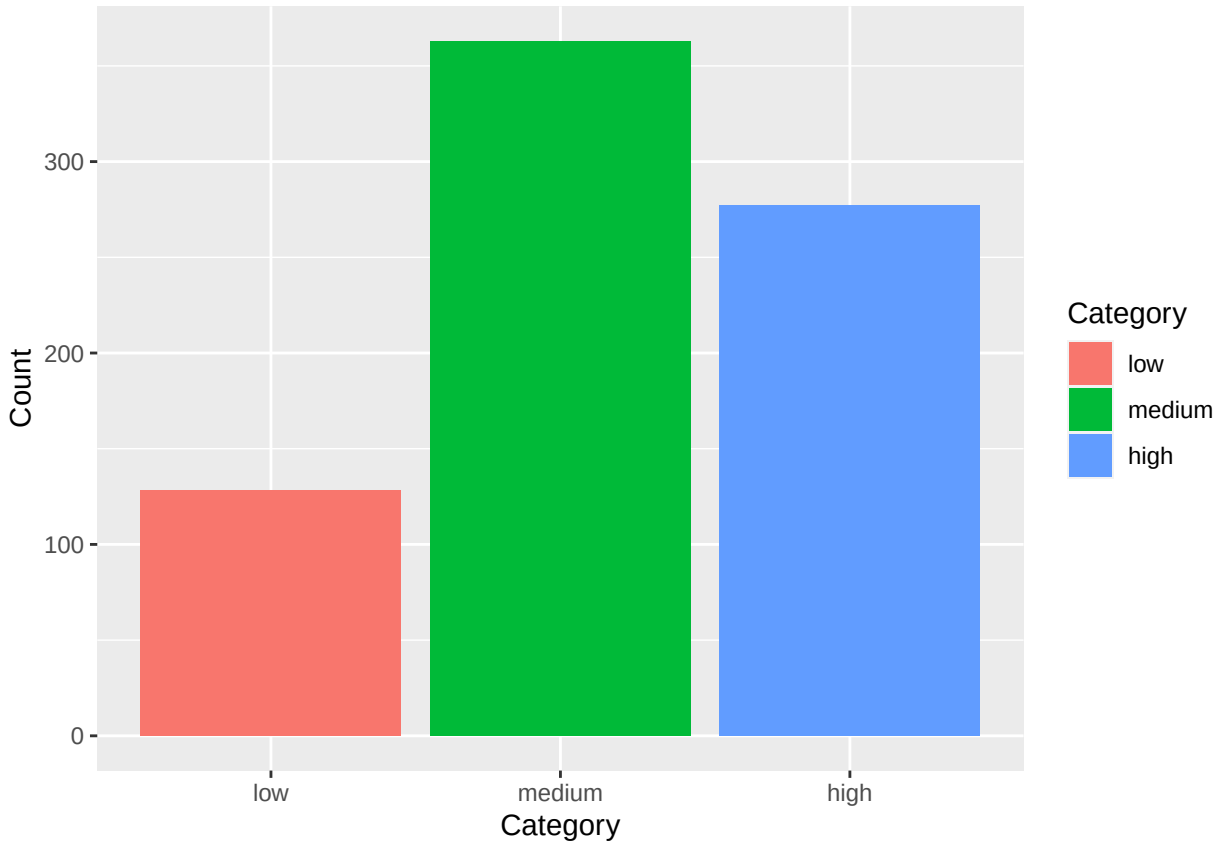
```
## [1] 1.125188
```

Comment: The model value of the parameter age is 23-24, and the histogram as well as the density plot aid in the concluding that the distribution is right skewed with extreme value to the right of the distribution. The skewness test returns the value of 1.25188, this aids in further substantiation of the the distribution being right-skewed. The histogram aids in concudinf that the median age is 29 years old. With the youngest person being 21 and the oldest being 81 years old.

Barplot for Diabetes

```
diabetes$diabcategory = cut(diabetes$DiabetesPedigreeFunction, breaks = c(-Inf,0.2,0.5,Inf), labels = c
counttype = diabetes %>%
  group_by(diabetes$diabcategory)%>%
  summarise(count = n())
colnames(counttype) = c("Category", "Count")
pctdiab = round((counttype$Count/sum(counttype$Count)*100), digits = 2)


ggplot(counttype, aes(factor(Category), Count, fill = Category)) + geom_col() + xlab("Category")
```



Comment: A new variable has been created to classify the people into categories of Low, Medium, and high. Then another variable was created to count the number of people classified under each category. This was used to create the barplot above. The category with most number of people is 'medium' with the second highest being the 'high' category and the one with the least being the 'low; category.

Checking for Nomaility:

```
shapiro.test(diabetes$BloodPressure)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diabetes$BloodPressure
## W = 0.81892, p-value < 2.2e-16
```

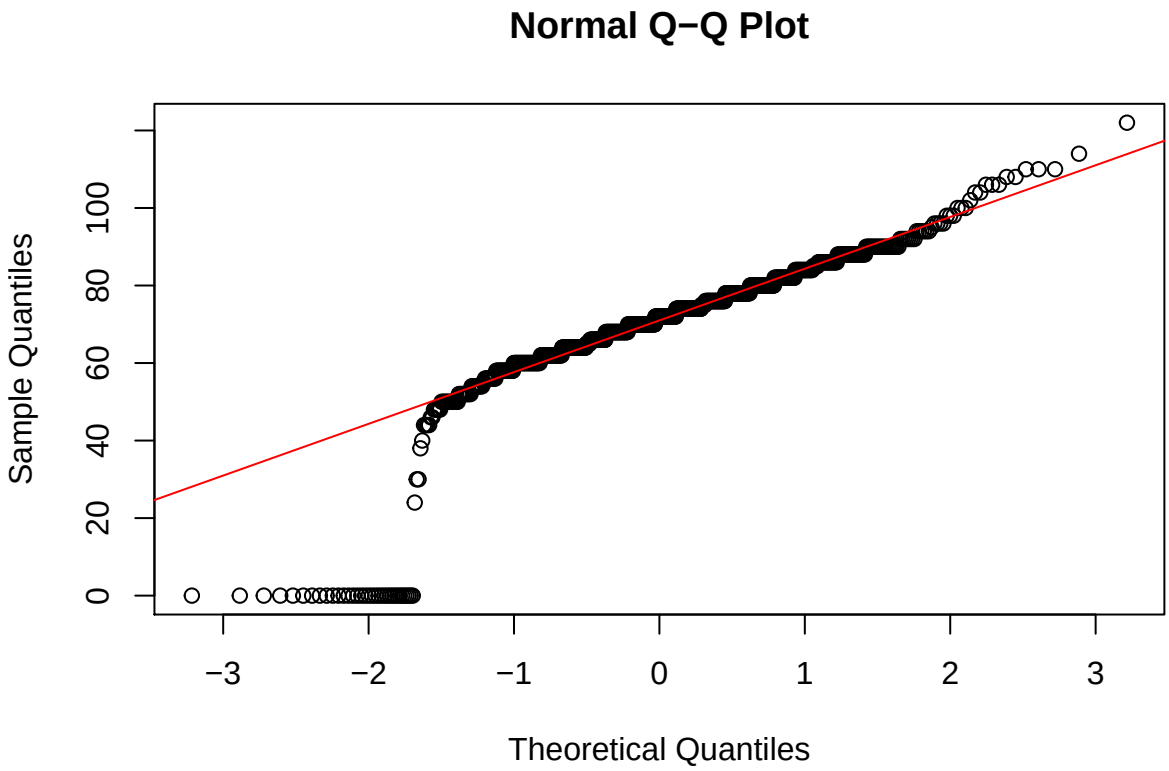
```
shapiro.test(diabetes$BMI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diabetes$BMI
## W = 0.94999, p-value = 1.842e-15
```

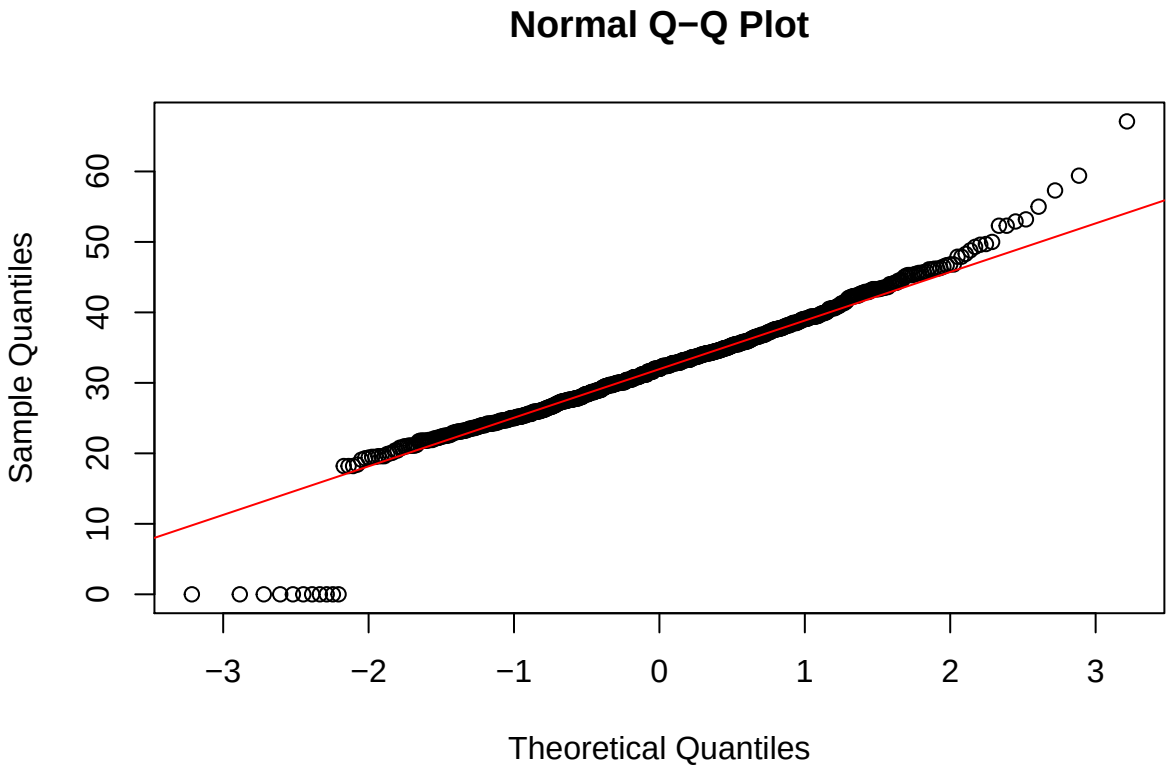
```
shapiro.test(diabetes$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diabetes$Age
## W = 0.87477, p-value < 2.2e-16
```

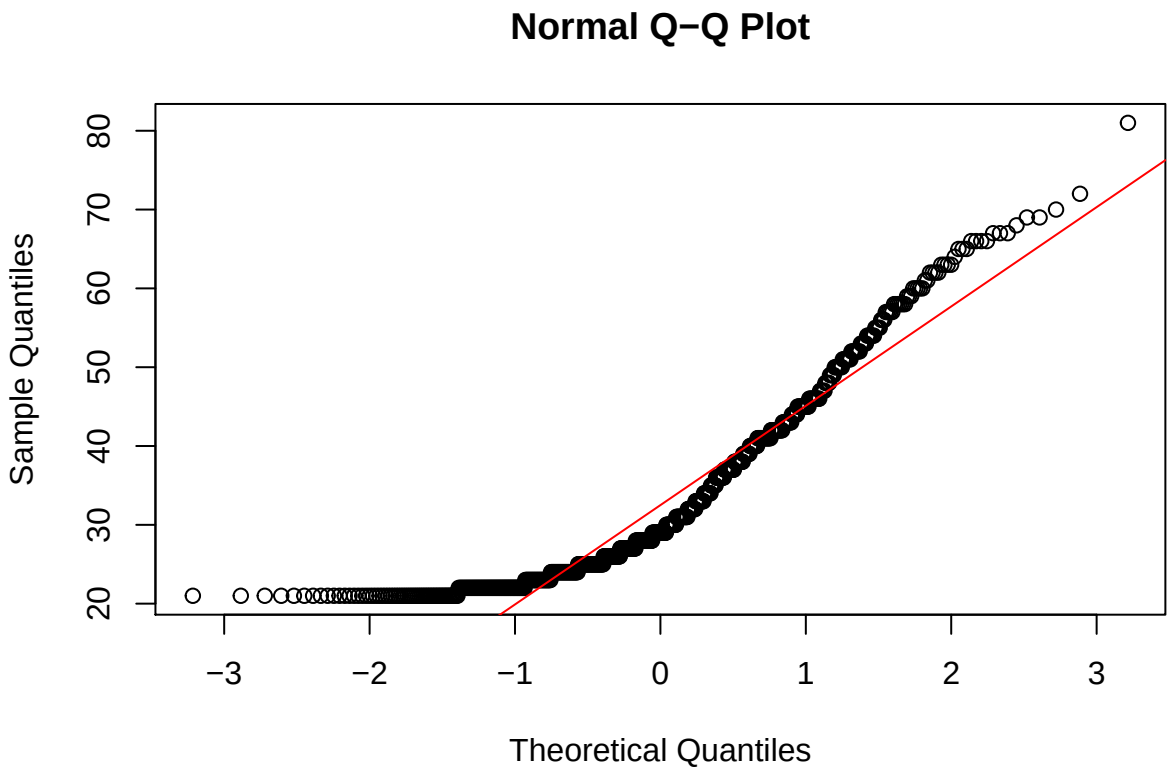
```r
qqnorm(diabetes$BloodPressure);qqline(diabetes$BloodPressure, col = 2)
```




```r
qqnorm(diabetes$BMI);qqline(diabetes$BMI, col = 2)
```

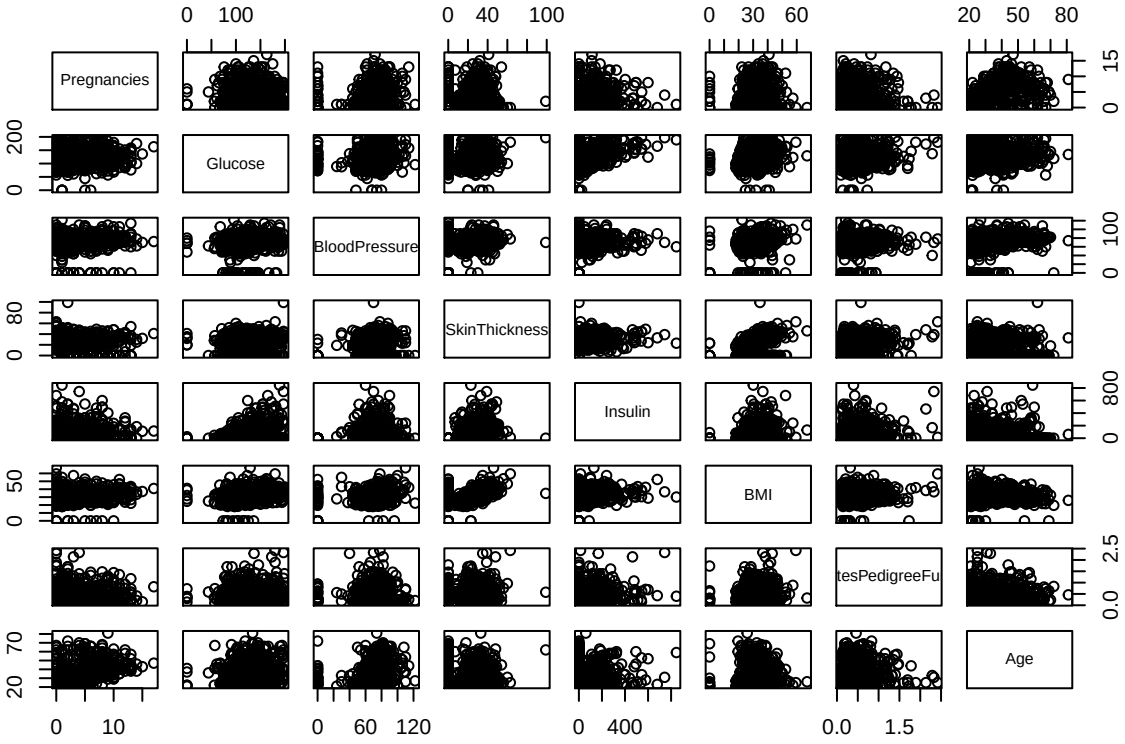


```
qqnorm(diabetes$Age);qqline(diabetes$Age, col = 2)
```



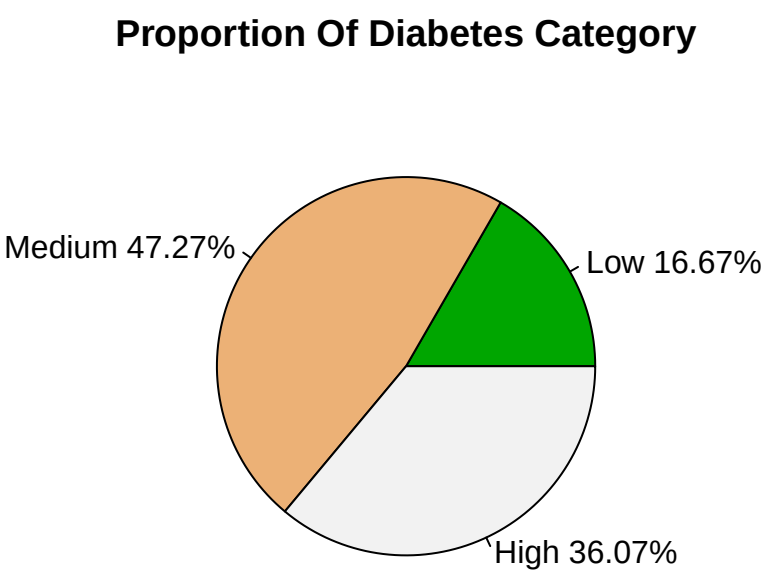
Along with the tests performed at every individual point of analysis of Bloop Pressure, BMI, and Age, the shapiro test proves that neither parameters abide to the norms of normal distribution. As the p value of each of the parameters is below 0.05 we conclude that they do not significantly follow normal distribution. The three qqnorm and qqline plots show that distribution to have longer tails and deviations from the qline which supports deviatin from norms of normal distribution.

```
plot(~Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFunction +
```



Comment: Pregnancies do not have a direct and directional visual relationship with any of the parameters except insulin. The relationship between the two parameters is weakly negative. As the number of pregnencies increses the average insulin appears to drop. There appears to be a positive relationship between glucose and inuslin. There seems to be a weak positive relationship between Age and Pregnencies. There appears to be a higher number of pregnancies as age increases. Also the BMI appears to increase with increased skin thickness.

```
labelsdiab = c("Low", "Medium", "High")
labelsdiab = paste(labelsdiab,pctdiab)
labelsdiab = paste(labelsdiab,"%", sep = "")
pie(counttype$Count, labels = labelsdiab, main = "Proportion Of Diabetes Category" ,col = terrain.colors
```



Comment: This chart was constructed using the the variable "pctdiab". It takes the prportion of the people in each category. We can refer this graph to the total count by category bar graph in we contructed earlier. The results remain the same with

'Medium' category dominating with 47.27% and the smallest proportion being 'Low' of 16.67%.

Q2:

```
Production = c(435,563,647,534,653,576,674,675,785,758,435,785,658,578,847,654,887,675,785,867)
wheatprod = data_frame(Production)
```

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

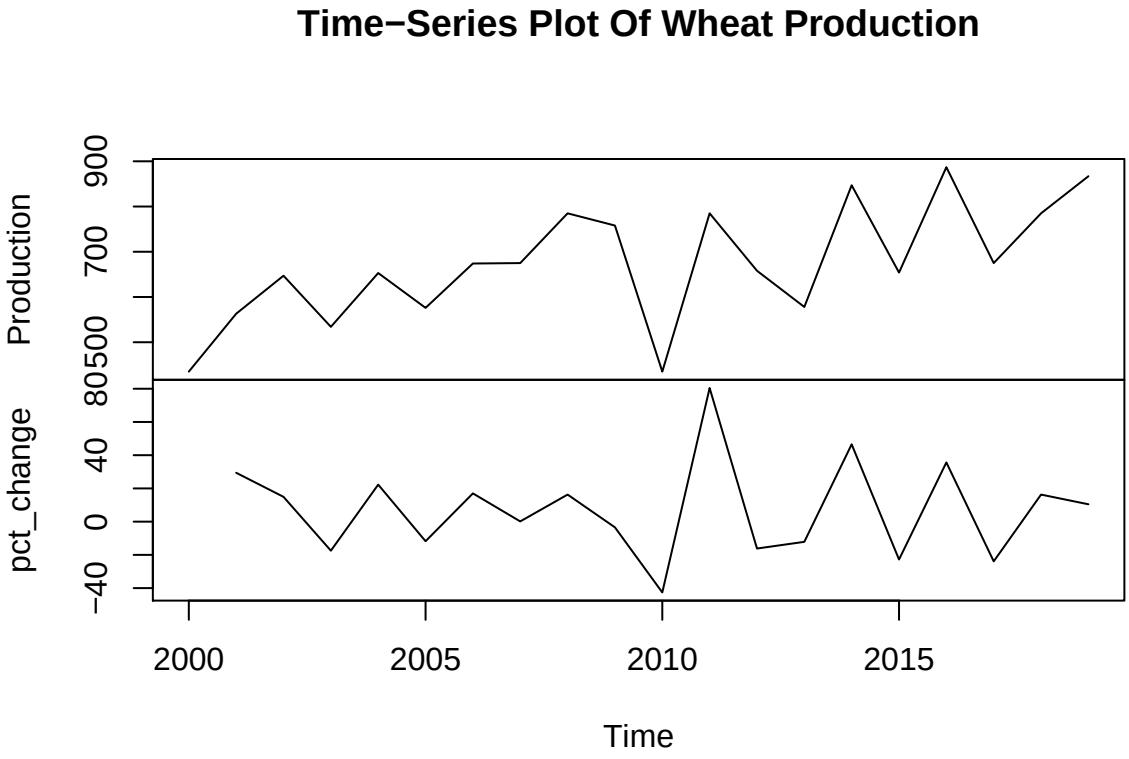
```
n = nrow(wheatprod)
for(i in 2:n) {
  wheatprod$pct_change[i] <- (wheatprod$Production[i]-wheatprod$Production[i-1])/wheatprod$Production[i
}
```

```
## Warning: Unknown or uninitialised column: 'pct_change'.
```

```
wheatprod$pct_change = wheatprod$pct_change * 100
tswheatprod = ts(wheatprod, start = c(2000), end = c(2019))
tswheatprod
```

```
## Time Series:
## Start = 2000
## End = 2019
## Frequency = 1
##      Production pct_change
## 2000        435         NA
## 2001        563  29.425287
## 2002        647  14.920071
## 2003        534 -17.465224
## 2004        653  22.284644
## 2005        576 -11.791730
## 2006        674  17.013889
## 2007        675   0.148368
## 2008        785  16.296296
## 2009        758  -3.439490
## 2010        435 -42.612137
## 2011        785  80.459770
## 2012        658 -16.178344
## 2013        578 -12.158055
## 2014        847  46.539792
## 2015        654 -22.786305
## 2016        887  35.626911
## 2017        675 -23.900789
## 2018        785  16.296296
## 2019        867  10.445860
```

```
plot(tswheatprod, main = "Time-Series Plot Of Wheat Production")
```



Comment:

The data above was plotted using the astsa library. We notice the wheat production expands from 2000 to 2010, after which it plumets in 2011 by 42.61% to recover by 80.45%. The biggest jump in production followed by the year 2014, in which there was a 46.53% jump.