Descriptive Statistics

- **Mean findout**

  mean_age <- mean(Med_diabet$Age)

  print(mean_age)

  or, summary(Med_diabet$Age)

```
> print(mean_age)
[1] 33.24089
>
> summary(Med_diabet$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   24.00   29.00   33.24   41.00   81.00
```

  Similarly, median_age <- median(Med_diabet$Age)

  print(median_age)

  The mean is the average of a numeric dataset and the median is the middle value in a dataset. It is useful for understanding the central tendency of the data.

- **Mode findout:**

  install.packages("DescTools") # Install package (only once)

  library(DescTools)

```
> mode_Age <- Mode(Med_diabet$Age, na.rm = TRUE)
> print(paste("Mode of age:", mode_Age))
[1] "Mode of age: 22"
> print(mode_Age)
[1] 22
attr(,"freq")
[1] 72
```

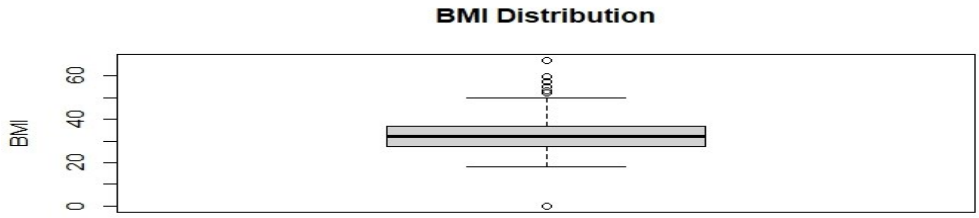
  The **mode** of the Age column is **22** (the most frequently occurring age in the dataset).

  The **frequency** of the mode is **72**, meaning the age **22 appears 72 times** in the dataset.

- **Box-Plot:**

  # Example: Boxplot for BMI

  boxplot(Med_diabet$BMI, main = "BMI Distribution", ylab = "BMI")



- **Quartile and IQR find-out:**

```
> quantiles_bp <- quantile(Med_diabet$BloodPressure, probs = c(0.25, 0.5, 0.75))
> print(quantiles_bp)
25% 50% 75%
 62  72  80
>
> # Interquartile range (IQR)
> iqr_bp <- IQR(Med_diabet$BloodPressure)
> print(iqr_bp)
[1] 18
```

- **Outlier Range find-out:**

```
> q1 <- quantile(Med_diabet$BloodPressure, probs = 0.25)
> print(q1)
25%
 62
> q3 <- quantile(Med_diabet$BloodPressure, probs = 0.75)
> print(q3)
75%
 80
>
> iqr_bp <- q3-q1
> print(iqr_bp)
75%
 18
>
> lower_bound<-q1-1.5*iqr_bp
> print(lower_bound)
25%
 35
> upper_bound<-q3+1.5*iqr_bp
> print(upper_bound)
75%
107
```

- **Range find-out :**

```
> # Example: Range of BMI (Body Mass Index)
> range_bmi <- range(Med_diabet$BMI)
> print(range_bmi)
[1]  0.0 67.1
```

- **Standard Deviation find-out :**

  sd() calculates the standard deviation, which is the square root of variance. 11.76023 means that the ages in Med_diabet$Age typically vary by 11.76 years from the average/mean.

  That is. mean+sd or mean-sd

```
> # Example: Standard deviation of age
> sd_age <- sd(Med_diabet$Age)
> print(sd_age)
[1] 11.76023
```

- **Variance find-out :**

  Variance measures how far each value in the dataset is from the mean. 1022.248 means the glucose values have a high spread from their mean. or, Standard Deviation (SD) = **√1022.248 ≈ 31.98**.This means that **glucose values** in diabetes_data typically deviate by **31.98 units** from the mean.

```
> variance_glucose <- var(Med_diabet$Glucose)
> print(variance_glucose)
[1] 1022.248
```

- **Standard Deviation, Variance, and Correlation find-out :**

```
> sd_age <- sd(diabetes_data$Age)
> print(sd_age)
[1] 12.05148
> variance_glucose <- var(diabetes_data$Glucose)
> print(variance_glucose)
[1] 700
> correlation_bmi_glucose <- cor(diabetes_data$BMI, diabetes_data$Glucose)
> print(correlation_bmi_glucose)
[1] 0.8436451
```

- **Kurtosis and Skewness find-out :**

  Skewness measures the asymmetry of the data distribution, while kurtosis measures the "tailedness" or how outliers are distributed.

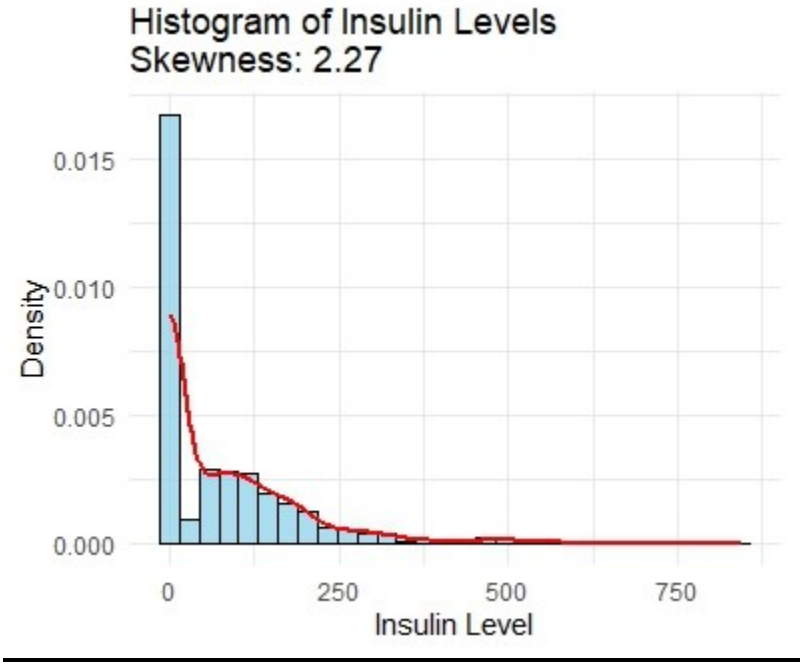
  A **positive skewness (>0)** indicates that the **distribution is right-skewed**, meaning that the tail on the **right side** of the distribution is longer or fatter.

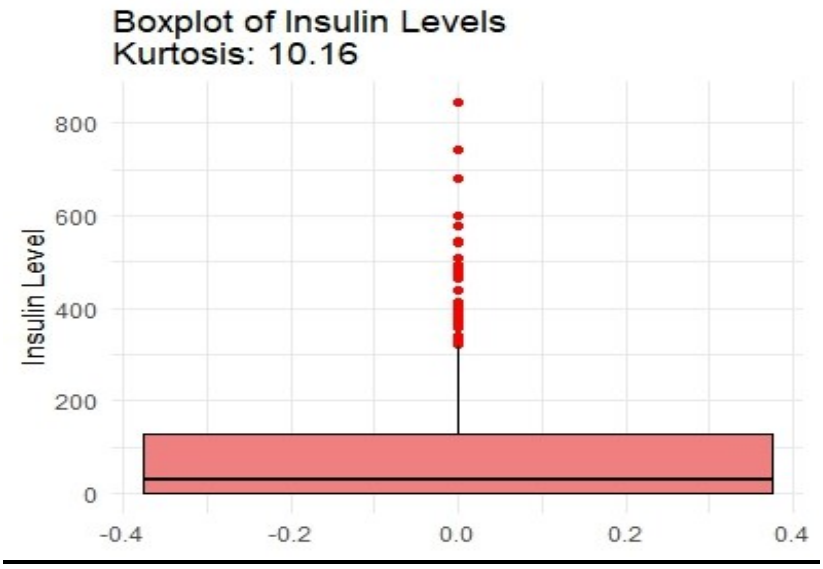
  ✓ **Right-skewed (positively skewed) distribution**
  ✓ **Mean > Median > Mode** (because extreme high values increase the mean)
  ✓ **Long right tail** due to high values.

Suppose, If you look at **income distribution**, most people earn an average income, but a few extremely high earners (like billionaires) pull the distribution to the right, making it positively skewed.

In this figure, most people have an average insulin range, but a few individuals have extremely high insulin levels, pulling the distribution to the right. This makes the distribution **positively skewed** (right-skewed).

The skewness value of **2.27** confirms this, as it is **greater than 0**, indicating a strong right-skewed distribution.





```r
# Load necessary libraries
library(ggplot2)
# Install and load the e1071 package
#install.packages("e1071")
library(e1071)
library(moments)  # For skewness and kurtosis functions

# Compute skewness and kurtosis
skewness_value <- skewness(Med_diabet$Insulin, na.rm = TRUE)
kurtosis_value <- kurtosis(Med_diabet$Insulin, na.rm = TRUE)

# Print the values
print(paste("Skewness:", skewness_value))
print(paste("Kurtosis:", kurtosis_value))

# Histogram with Density Curve# Histogram with correct kurtosis value
ggplot(Med_diabet, aes(x = Insulin)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_density(color = "red", size = 1) +
  ggtitle(paste("Histogram of Insulin Levels\nSkewness:", round(skewness_value, 2))) +
  xlab("Insulin Level") +
  ylab("Density") +
  theme_minimal()

# Boxplot to Detect Outliers
ggplot(Med_diabet, aes(y = Insulin)) +
  geom_boxplot(fill = "lightcoral", color = "black", outlier.color = "red", outlier.shape = 16) +
  ggtitle(paste("Boxplot of Insulin Levels\nKurtosis:", round(kurtosis_value, 2))) +
  ylab("Insulin Level") +
  theme_minimal()

```

Kurtosis measures the "tailedness" of the distribution. A normal distribution has kurtosis = 3. High kurtosis (>3) means the distribution has heavy tails, meaning more extreme outliers. Low kurtosis (<3) means the distribution has light tails, meaning fewer extreme values.

Interpretation of our Kurtosis = 7.133

Since 7.133 is much greater than 3,our Insulin data has very heavy tails.

This means there are many extreme values (outliers) in the dataset. our insulin data has more extreme high values than a normal distribution would expect.