Metabolic Syndrome Risk Analysis

Metabolic Syndrome is a cluster of conditions, including high blood pressure, high blood sugar, excess body fat around the waist, and abnormal cholesterol or triglyceride levels, that together increase the risk of heart disease, stroke, and type 2 diabetes. Early detection is critical for prevention and better health outcomes.

This project aims to uncover key health indicators associated with Metabolic Syndrome by analyzing patient data. Through a combination of data cleaning, exploratory analysis, and risk factor modeling, we examine patterns in variables such as age, BMI, blood glucose, HDL cholesterol, and triglycerides. Patients are then grouped based on their risk levels to highlight early warning signs.

Data Source : https://data.world/informatics-edu/metabolic-syndrome-prediction Extracting Data

- The dataset is imported into R using read.csv() to begin the analysis.
- I used glimpse() to get a quick summary of the dataset's structure, including variable names, types, and sample values.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

```
data <- read.csv("/Users/akhilarachuri/Documents/R Project/Metabolic Syndrome.csv")
glimpse(data)
```

```
## Rows: 2,401
## Columns: 15
## $ seqn              <int> 62161, 62164, 62169, 62172, 62177, 62178, 62184, 621~
## $ Age               <int> 22, 44, 21, 43, 51, 80, 26, 30, 70, 35, 57, 36, 28, ~
## $ Sex               <chr> "Male", "Female", "Male", "Female", "Male", "Male", ~
## $ Marital           <chr> "Single", "Married", "Single", "Single", "Married", ~
## $ Income            <int> 8200, 4500, 800, 2000, NA, 300, 9000, 6200, 1000, 25~
## $ Race              <chr> "White", "White", "Asian", "Black", "Asian", "White"~
## $ WaistCirc         <dbl> 81.0, 80.1, 69.6, 120.4, 81.1, 112.5, 78.6, 80.2, NA~
## $ BMI               <dbl> 23.3, 23.2, 20.1, 33.3, 20.1, 28.5, 22.1, 22.4, NA, ~
## $ Albuminuria       <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1~
## $ UrAlbCr           <dbl> 3.88, 8.55, 5.07, 5.22, 8.13, 9.79, 9.21, 8.78, 45.6~
## $ UricAcid          <dbl> 4.9, 4.5, 5.4, 5.0, 5.0, 4.8, 5.4, 6.7, 5.4, 6.7, 6.~
## $ BloodGlucose      <int> 92, 82, 107, 104, 95, 105, 87, 83, 96, 94, 100, 94, ~
## $ HDL               <int> 41, 28, 43, 73, 43, 47, 61, 48, 35, 46, 35, 58, 40, ~
## $ Triglycerides     <int> 84, 56, 78, 141, 126, 100, 40, 91, 75, 86, 98, 182, ~
## $ MetabolicSyndrome <chr> "No MetSyn", "No MetSyn", "No MetSyn", "No MetSyn", ~
```

```r
data$MetabolicSyndrome <- as.factor(data$MetabolicSyndrome)
data$Marital <- as.factor(data$Marital)
data$Sex <- as.factor(data$Sex)
data$Race <- as.factor(data$Race)
data$Albuminuria <- as.factor(data$Albuminuria)
summary(data)
```

```
##       seqn            Age            Sex            Marital          Income
##  Min.   :62161   Min.   :20.00   Female:1211            : 208   Min.   : 300
##  1st Qu.:64591   1st Qu.:34.00   Male  :1190   Divorced : 242   1st Qu.:1600
##  Median :67059   Median :48.00                 Married  :1192   Median :2500
##  Mean   :67031   Mean   :48.69                 Separated:  95   Mean   :4005
##  3rd Qu.:69495   3rd Qu.:63.00                 Single   : 498   3rd Qu.:6200
##  Max.   :71915   Max.   :80.00                 Widowed  : 166   Max.   :9000
##                                                                 NA's   :117
##           Race       WaistCirc           BMI       Albuminuria
##  Asian      :349   Min.   : 56.20   Min.   :13.4   0:2089
##  Black      :548   1st Qu.: 86.67   1st Qu.:24.0   1: 254
##  Hispanic   :257   Median : 97.00   Median :27.7   2:  58
##  MexAmerican:253   Mean   : 98.31   Mean   :28.7
##  Other      : 61   3rd Qu.:107.62   3rd Qu.:32.1
##  White      :933   Max.   :176.00   Max.   :68.7
##                    NA's   :85       NA's   :26
##     UrAlbCr           UricAcid       BloodGlucose        HDL
##  Min.   :   1.40   Min.   : 1.800   Min.   : 39.0   Min.   : 14.00
##  1st Qu.:   4.45   1st Qu.: 4.500   1st Qu.: 92.0   1st Qu.: 43.00
##  Median :   7.07   Median : 5.400   Median : 99.0   Median : 51.00
##  Mean   :  43.63   Mean   : 5.489   Mean   :108.2   Mean   : 53.37
##  3rd Qu.:  13.69   3rd Qu.: 6.400   3rd Qu.:110.0   3rd Qu.: 62.00
##  Max.   :5928.00   Max.   :11.300   Max.   :382.0   Max.   :156.00
##
##  Triglycerides    MetabolicSyndrome
##  Min.   :  26.0   MetSyn   : 822
##  1st Qu.:  75.0   No MetSyn:1579
##  Median : 103.0
##  Mean   : 128.1
##  3rd Qu.: 150.0
##  Max.   :1562.0
##
```

Data Cleaning

- Checked for duplicates - found none
- Checked for null values - found 228.

```
sum(duplicated(data))
```

```
## [1] 0
```

```
sum(is.na(data) | data == "")
```

```
## [1] 436
```

Checked the sum of null values in each column

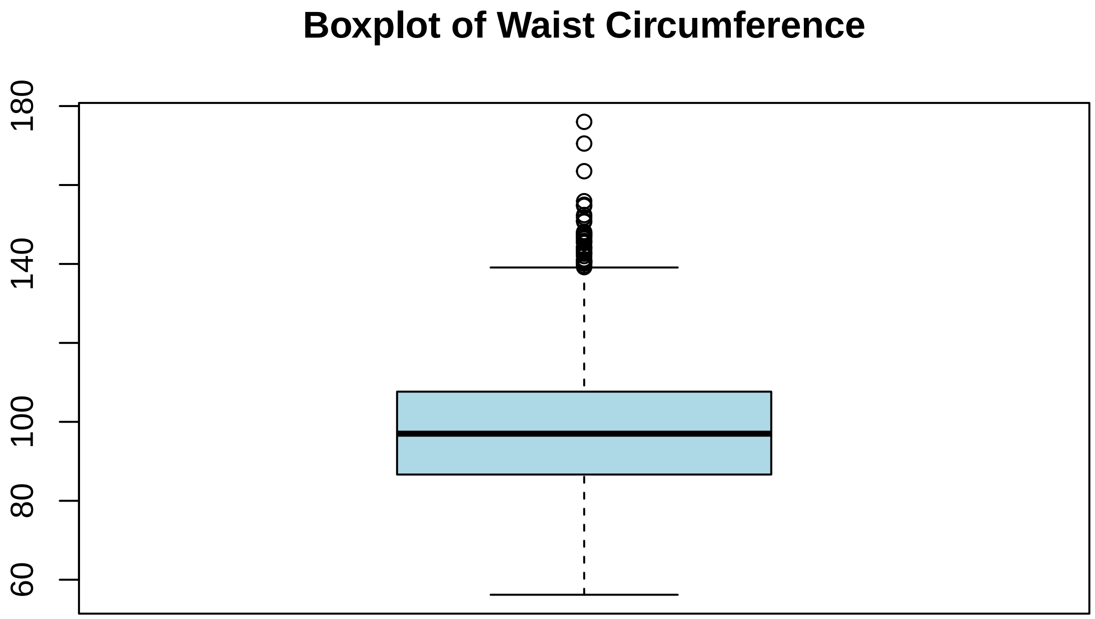
```
colSums(is.na(data) | data == "")
```

```
##               seqn               Age               Sex           Marital
##                  0                 0                 0               208
##             Income              Race         WaistCirc               BMI
##                117                 0                85                26
##        Albuminuria           UrAlbCr          UricAcid      BloodGlucose
##                  0                 0                 0                 0
##                HDL     Triglycerides MetabolicSyndrome
##                  0                 0                 0
```

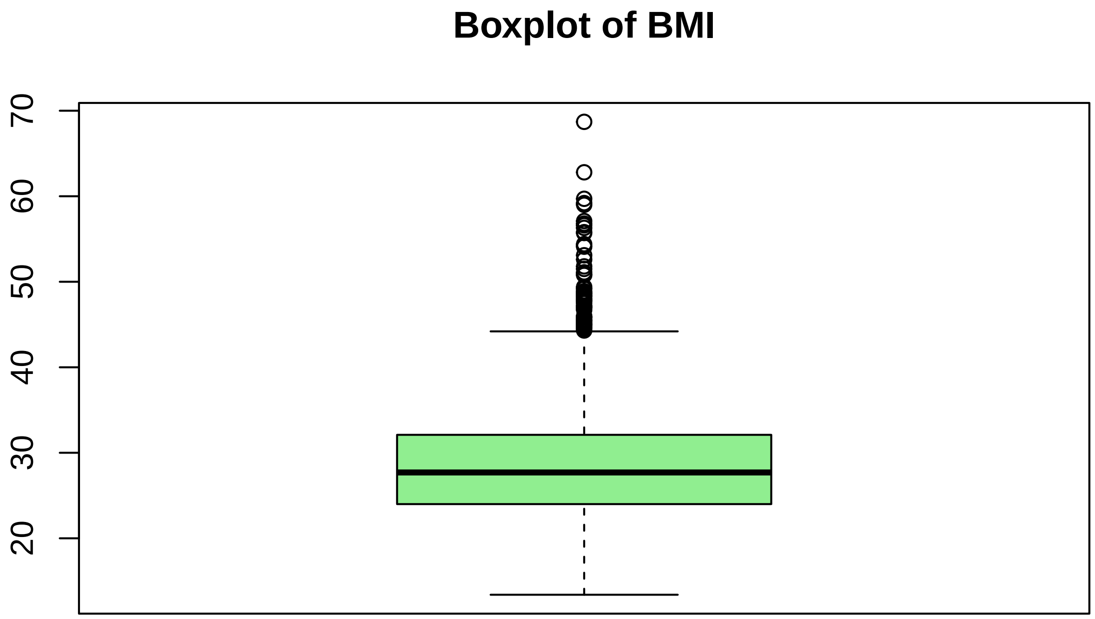
Checked for Outliers

- The **Boxplot of Waist Circumference** shows a symmetric distribution with several high-value outliers above the upper whisker, indicating some individuals have significantly larger waist sizes than the typical range.
- The **Boxplot of BMI** reveals a right-skewed distribution with many outliers on the higher end, suggesting a portion of the population has unusually high BMI values compared to the median.

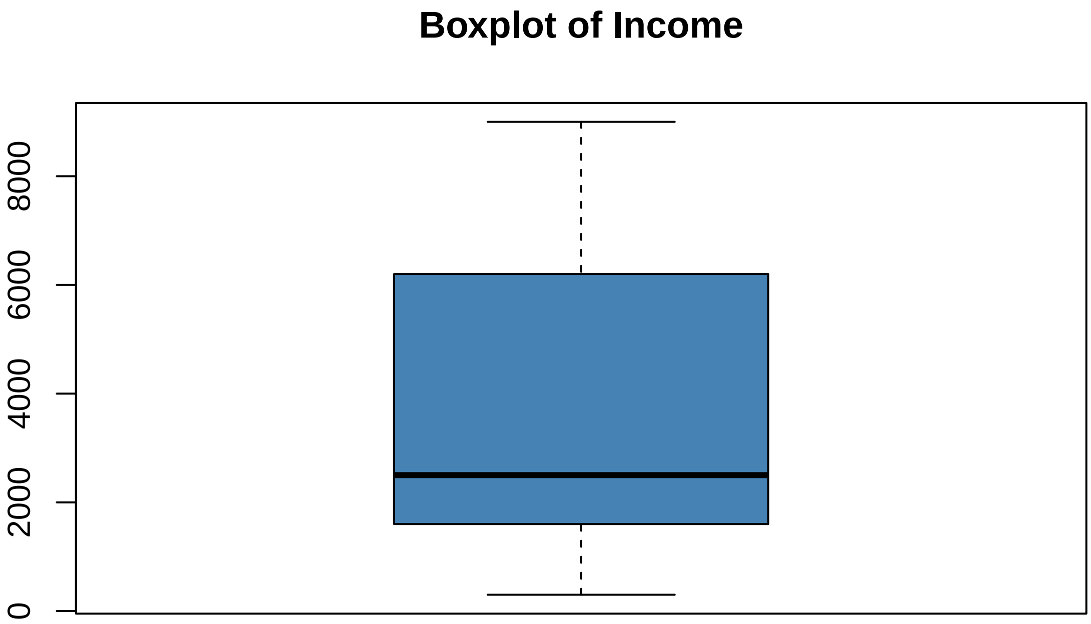
```
boxplot(data$WaistCirc, main = "Boxplot of Waist Circumference",col = "lightblue")
```



```r
boxplot(data$BMI, main = "Boxplot of BMI",col = "lightgreen")
```



```r
boxplot(data$Income, main = "Boxplot of Income",col = "steelblue")
```



Outlier Removal Using IQR Method

```
Q1_waist <- quantile(data$WaistCirc, 0.25, na.rm = TRUE)
Q3_waist <- quantile(data$WaistCirc, 0.75, na.rm = TRUE)
IQR_waist <- Q3_waist - Q1_waist
lower_waist <- Q1_waist - 1.5 * IQR_waist
upper_waist <- Q3_waist + 1.5 * IQR_waist
Q1_bmi <- quantile(data$BMI, 0.25, na.rm = TRUE)
Q3_bmi <- quantile(data$BMI, 0.75, na.rm = TRUE)
IQR_bmi <- Q3_bmi - Q1_bmi
lower_bmi <- Q1_bmi - 1.5 * IQR_bmi
upper_bmi <- Q3_bmi + 1.5 * IQR_bmi
data <- data %>%
  filter(
    WaistCirc >= lower_waist & WaistCirc <= upper_waist,
    BMI >= lower_bmi & BMI <= upper_bmi
  )
```

Handling Missing Values Using Mean and Median Imputation

```
data$Income[is.na(data$Income)] <- median(data$Income, na.rm = TRUE)
data$WaistCirc[is.na(data$WaistCirc)] <- mean(data$WaistCirc, na.rm = TRUE)
data$BMI[is.na(data$BMI)] <- mean(data$BMI, na.rm = TRUE)
```

- Again Checked for null values - found none.

```r
sum(is.na(data))
```

```
## [1] 0
```

- Removed rows with missing Marital values from the data

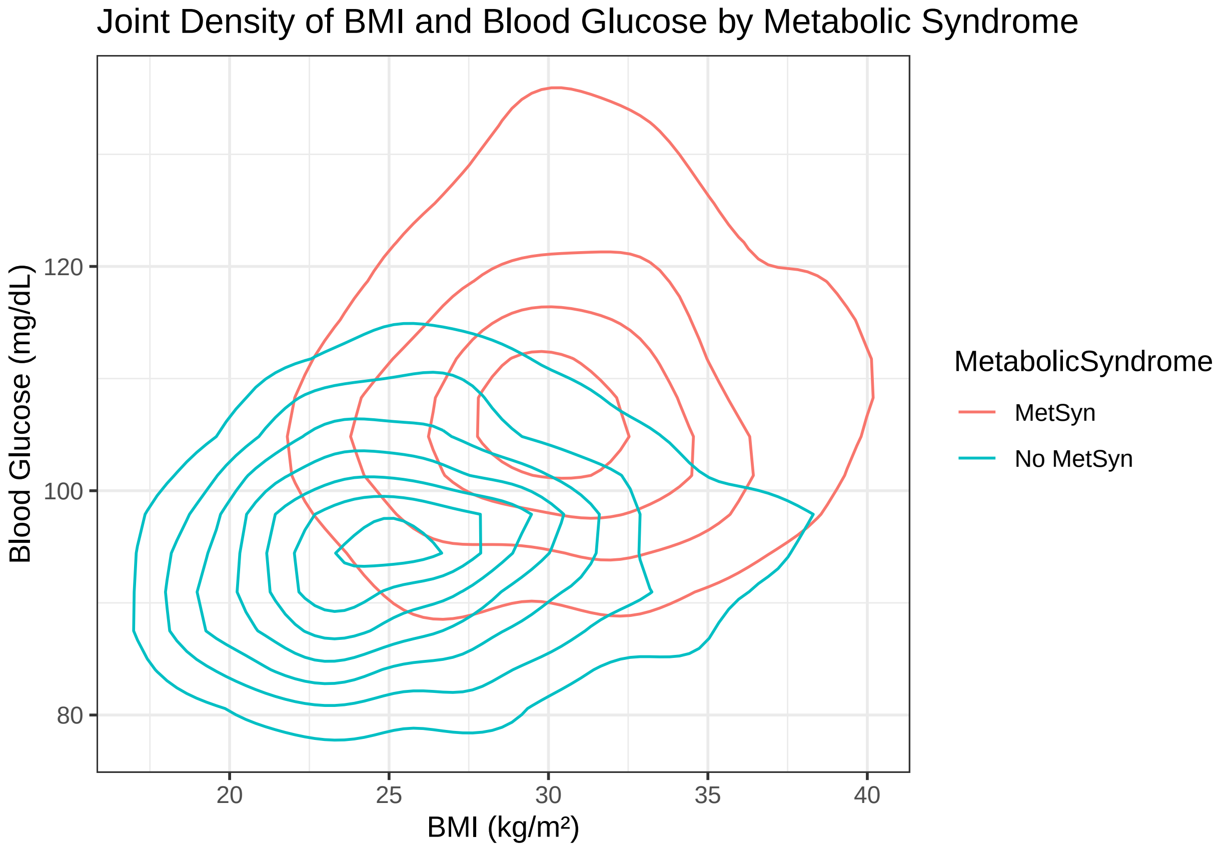
```r
data <- data %>%
  filter(!is.na(Marital), Marital != "")
```

Exploratory Data Analysis

Joint Density of BMI and Blood Glucose by Metabolic Syndrome

- The Metabolic Syndrome group (red contours) peak density lies approximately in the BMI range of 30–35 and glucose range of 100–120, which are clinical indicators of obesity and impaired glucose metabolism.
- The Non Metabolic Syndrome group (blue contours) is more concentrated around BMI 22–28 and glucose 80–100, a healthier metabolic range.

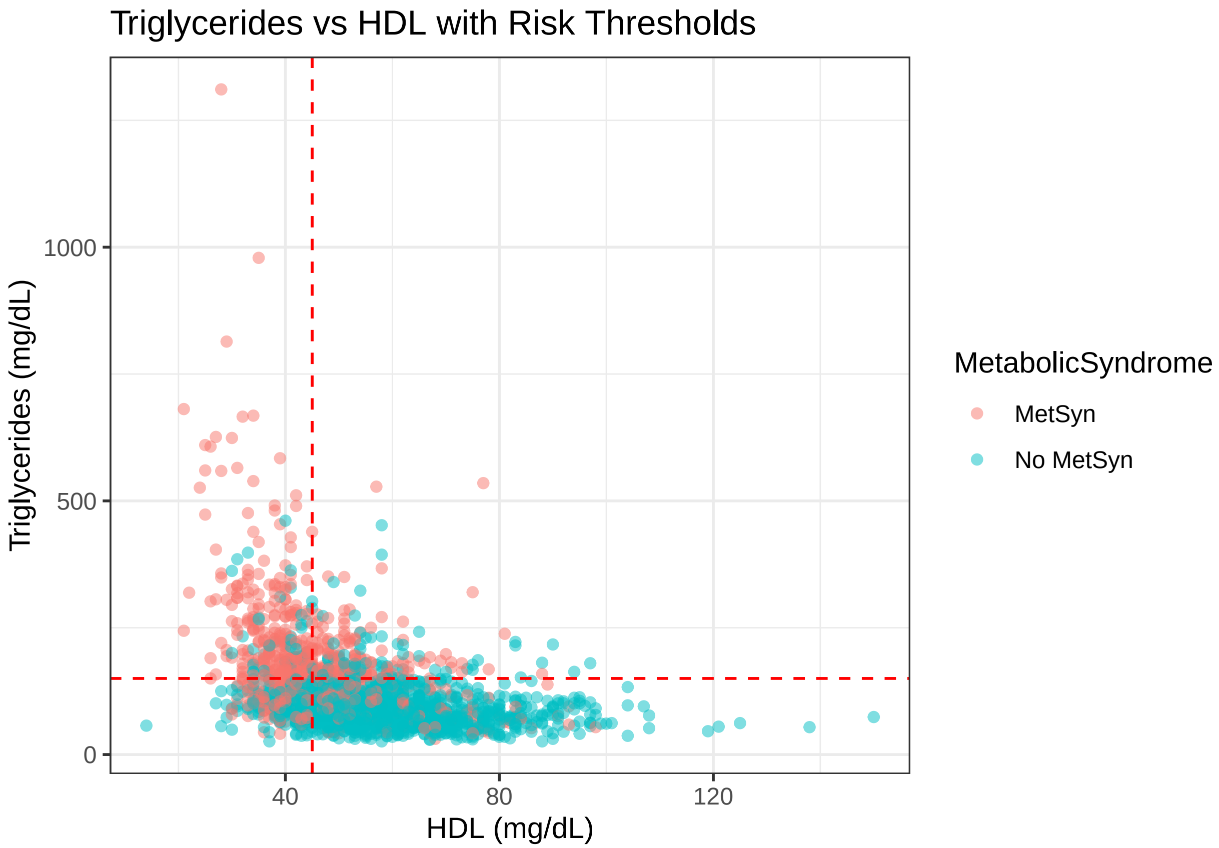
```r
ggplot(data, aes(x = BMI, y = BloodGlucose, color = MetabolicSyndrome)) +
  geom_density2d() +
  labs(x = "BMI (kg/m²)", y = "Blood Glucose (mg/dL)", title = "Joint Density of BMI and Blood Glucose 
  theme_bw()
```



Scatter Plot with Clinical Threshold Lines

- Vertical at HDL = 45 mg/dL -> clinical threshold for low HDL and Horizontal at Triglycerides = 150 mg/dL -> clinical threshold for high triglycerides
- Metabolic Syndrome is concentrated in the high-risk zone, where HDL < 45 and Triglycerides > 150.
- Non Metabolic Syndrome patients are mostly clustered in the safer bottom-right quadrant.

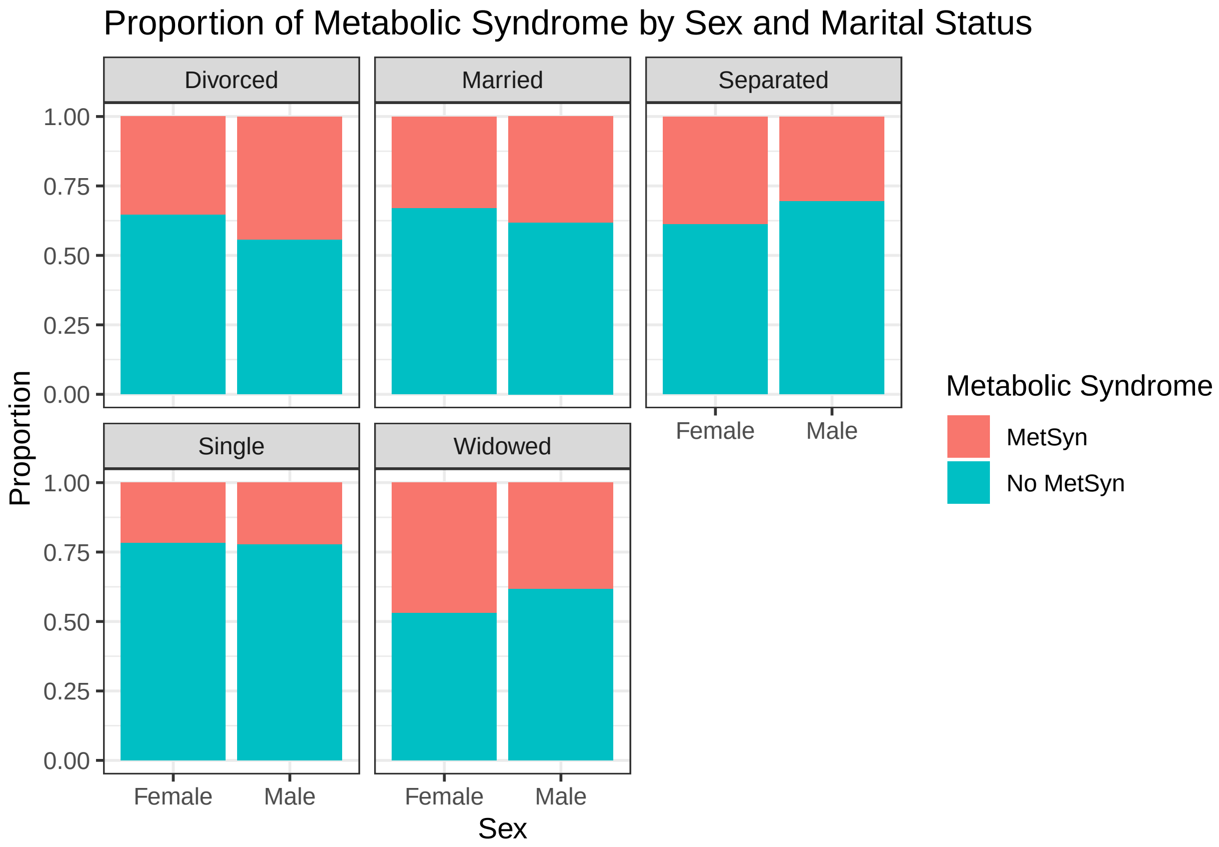
```
ggplot(data, aes(x = HDL, y = Triglycerides, color = MetabolicSyndrome)) +
  geom_point(alpha = 0.5) +
  geom_vline(xintercept = 45, linetype = "dashed", color = "red") +
  geom_hline(yintercept = 150, linetype = "dashed", color = "red") +
  labs(x = "HDL (mg/dL)", y = "Triglycerides (mg/dL)", title = "Triglycerides vs HDL with Risk Threshol
  theme_bw()
```



Metabolic Syndrome by Gender and Marital Status

- Widowed females and divorced males have highest Metabolic Syndrome rates.
- Single males/females show lowest Metabolic Syndrome proportions.

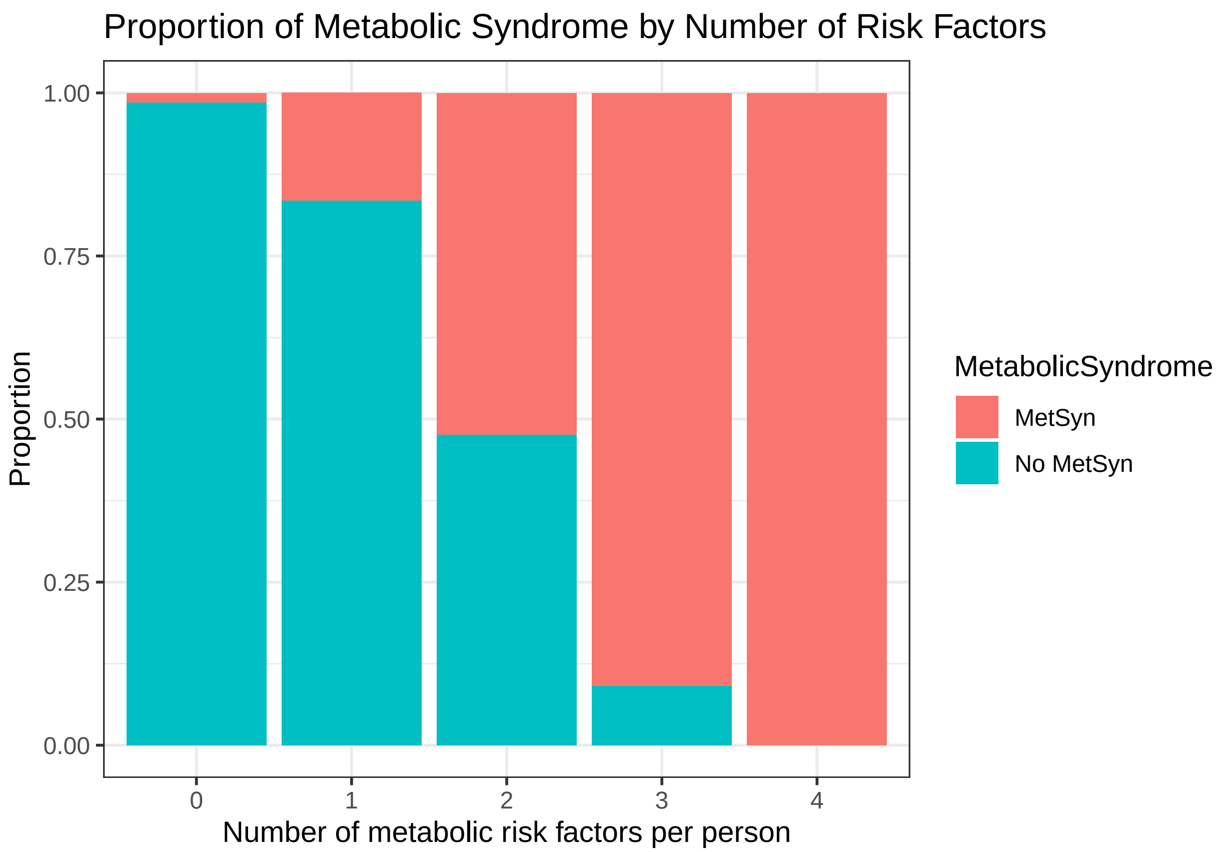
```
data$MetabolicSyndrome <- ifelse(data$MetabolicSyndrome == "MetSyn", "MetSyn", "No MetSyn")
ggplot(data %>% filter(!is.na(MetabolicSyndrome)),
       aes(x = Sex, fill = MetabolicSyndrome)) +
  geom_bar(position = "fill") +
  facet_wrap(~ Marital) +
  labs(
    title = "Proportion of Metabolic Syndrome by Sex and Marital Status",
    x = "Sex", y = "Proportion",
    fill = "Metabolic Syndrome"
  ) +
  theme_bw()
```



Metabolic Syndrome by Risk Factor Count

- As the number of risk factors increases (e.g., high BMI, low HDL, high glucose, high triglycerides), the proportion of patients with Metabolic Syndrome (red) also increases sharply.
- 0–1 risk factors: Almost all patients do not have Metabolic Syndrome
- 2 risk factors: There’s a balanced split between those with and without Metabolic Syndrome
- 3–4 risk factors: Nearly all patients have Metabolic Syndrome

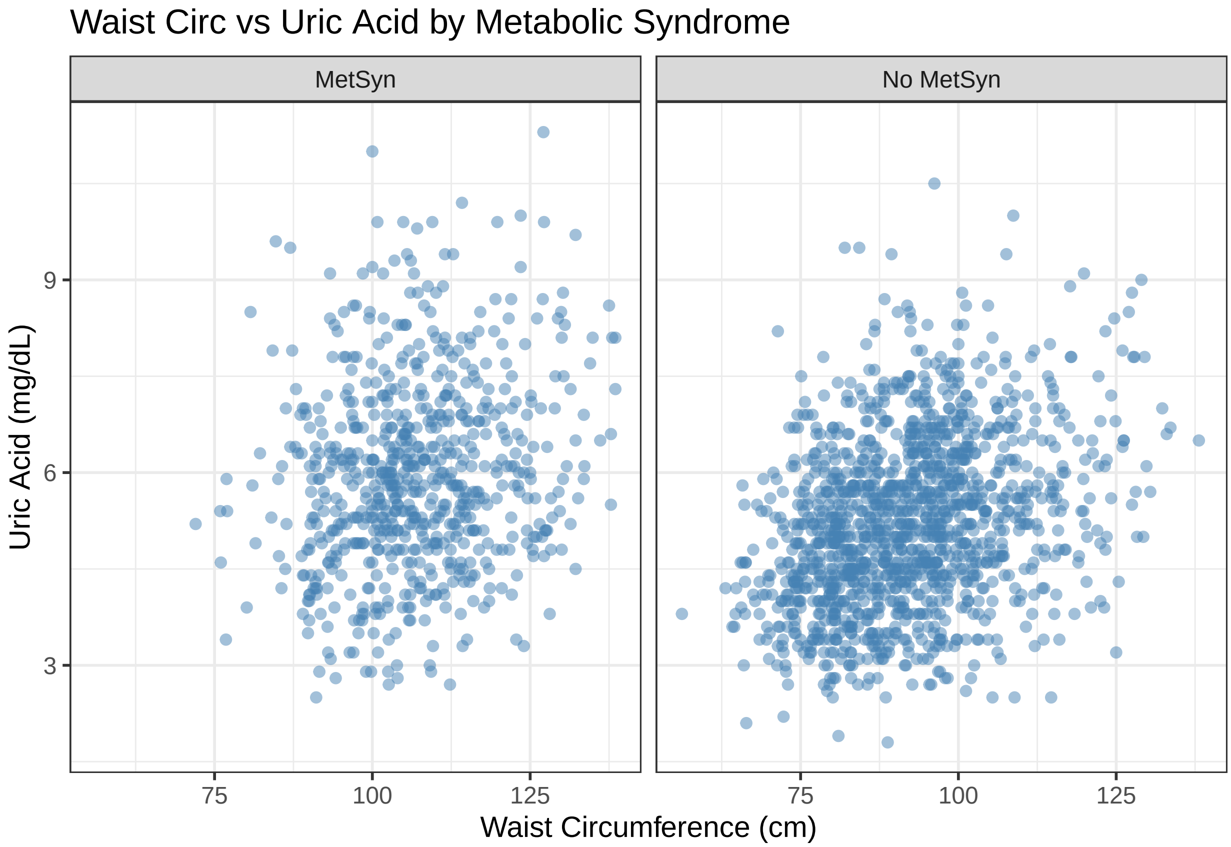
```
RiskFactors <- with(data,
  (BMI > 30) + (HDL < 45) + (BloodGlucose > 100) + (Triglycerides > 150))
ggplot(data, aes(x = factor(RiskFactors), fill = MetabolicSyndrome)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Metabolic Syndrome by Number of Risk Factors",
       x = "Number of metabolic risk factors per person", y = "Proportion") +
  theme_bw()
```



Waist Circumference vs Uric Acid by MetSyn Group

- Most Metabolic Syndrome patients often have waists over 100 cm and have uric acid levels clustered toward the upper normal range, around 5–7 mg/dL.
- This highlights a combination of abdominal obesity and a bit of raised uric acid.
- Most people without Metabolic Syndrome have waist sizes around 80–110 cm and keep their uric acid under 7 mg/dL, both below clinical risk cutoffs, showing a stable, healthy profile.
- Uric acid levels are slightly elevated in Metabolic Syndrome patients compared to Non Metabolic Syndrome, though overlap exists this indicates a potential contributing factor rather than a clear-cut discriminator.

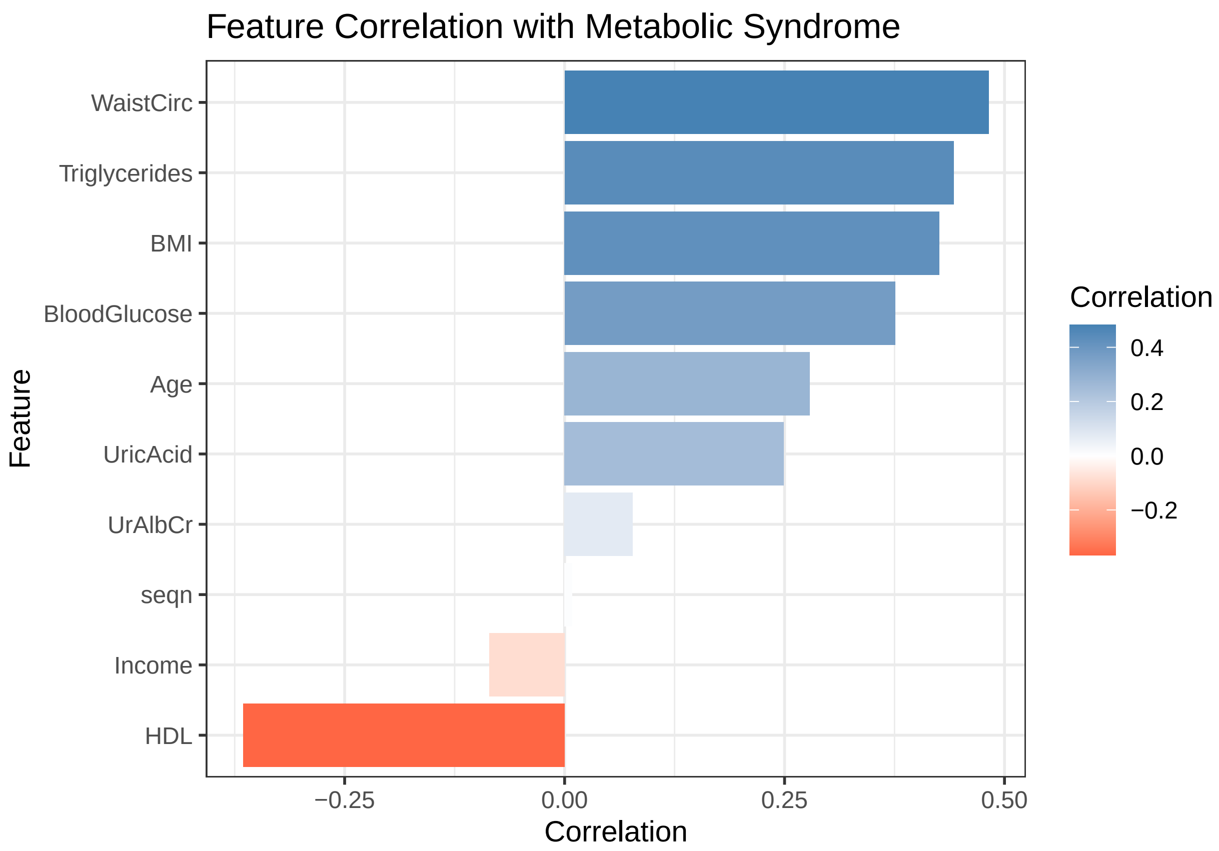
```r
ggplot(data, aes(x = WaistCirc, y = UricAcid)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  facet_wrap(~ MetabolicSyndrome) +
  labs(title = "Waist Circ vs Uric Acid by Metabolic Syndrome",
       x = "Waist Circumference (cm)", y = "Uric Acid (mg/dL)") +
  theme_bw()
```



Key Predictors of Metabolic Syndrome

- Waist Circumference, BMI, Triglycerides, and Blood Glucose show strong positive correlations with Metabolic Syndrome, indicating that higher values in these features significantly increase the likelihood of having the condition.
- HDL (good cholesterol) has a strong negative correlation, meaning lower HDL levels are closely associated with the presence of Metabolic Syndrome, making it an important factor in identifying increased risk.

```r
# Transformation of traget variable into 0 and 1
data$MetabolicSyndrome <- ifelse(data$MetabolicSyndrome == "MetSyn", 1, 0)
numeric_data <- data %>%
  select(where(is.numeric))
cor_matrix <- cor(numeric_data)
cor_target <- cor_matrix[, "MetabolicSyndrome"]
cor_df <- data.frame(
  Feature = names(cor_target),
  Correlation = cor_target)
cor_df <- cor_df %>% filter(Feature != "MetabolicSyndrome")
ggplot(cor_df, aes(x = reorder(Feature, Correlation), y = Correlation, fill = Correlation)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_gradient2(low = "red", mid = "white", high = "steelblue", midpoint = 0) +
  labs(title = "Feature Correlation with Metabolic Syndrome",
       x = "Feature", y = "Correlation") +
  theme_bw()
```



Predictive Modeling

- Based on the above Correlation plot, top features with the strongest positive and negative correlation to Metabolic Syndrome (such as BMI, Waist Circumference, Triglycerides, Blood Glucose, and HDL) were identified as the most influential variables for prediction.
- These selected features were used to train a decision tree model, which visually captured how combinations of clinical markers contribute to Metabolic Syndrome risk, enabling clear and interpretable rule-based classification.

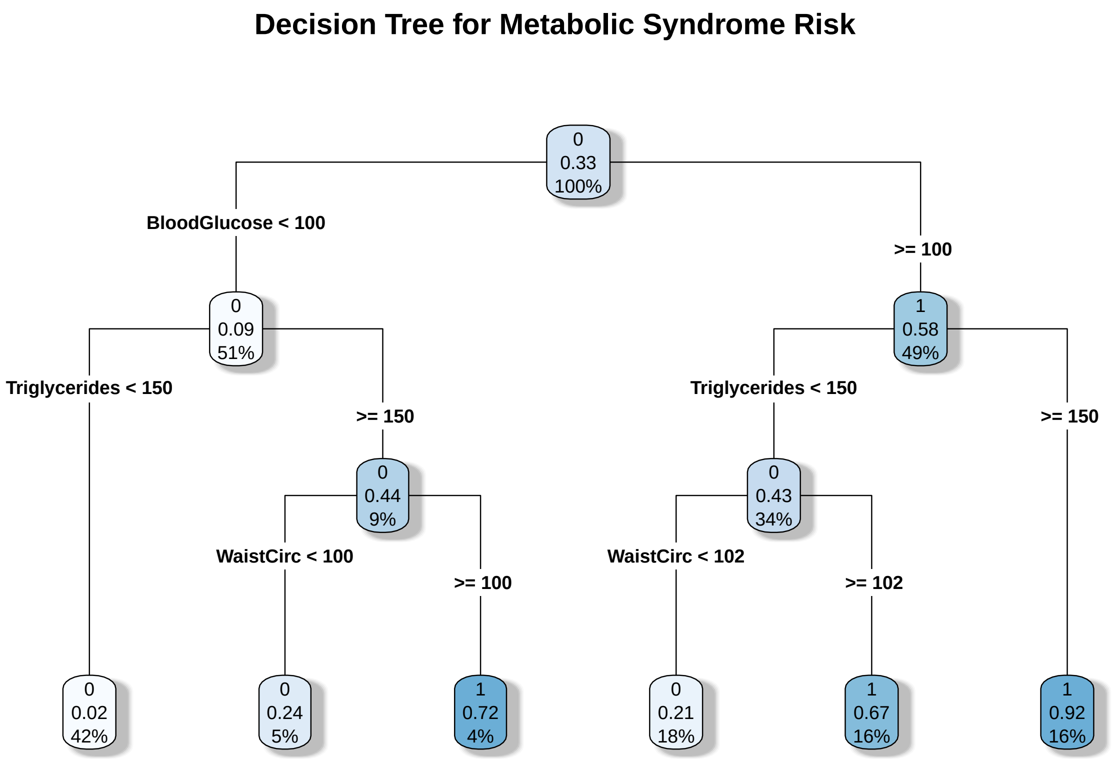
```
library(rpart)
library(rpart.plot)
library(caret)
set.seed(123)
data_split <- sample(1:nrow(data), 0.8 * nrow(data))
train <- data[data_split, ]
test <- data[-data_split, ]
fit <- rpart(MetabolicSyndrome ~ BMI + HDL + Triglycerides + BloodGlucose + WaistCirc, data = train, me
print(fit$variable.importance)
```

```
##  BloodGlucose Triglycerides     WaistCirc           BMI           HDL
##      203.8270      175.1290      136.0298      101.2249       61.5168
```

```
pred <- predict(fit, newdata = test, type = "class")
accuracy <- sum(pred == test$MetabolicSyndrome) / length(pred)
print(paste("Decision Tree Accuracy:", round(accuracy*100,2), "%"))
```

```
## [1] "Decision Tree Accuracy: 86.03 %"
```

```
rpart.plot(fit, main = "Decision Tree for Metabolic Syndrome Risk",type = 4,box.palette = "Blues",shado
```



1.Blood Glucose is the Primary Split:

- The first and most important factor in predicting Metabolic Syndrome is whether blood glucose is above or below 100 mg/dL, a clear indicator of risk.

2.High Triglycerides and Waist Circumference Increase Risk:

- Patients with high triglycerides (145 or 153) and larger waist circumference (100 or 102) are more likely to have Metabolic Syndrome, as shown by the high probabilities in the leaf nodes on the right side of the tree.

```
view(data)
```