Investigating the Prevalence of Metabolic Syndrome

GROUP- 12

Asha Chirumamilla

Pallavi Telu

William Teske

Likhitha Arigala

Teja Pavani Jyesta

Introduction

- Metabolic syndrome is a group of conditions that together raise the risk of coronary heart disease, diabetes, stroke, and other serious health problems. Metabolic syndrome is also called insulin resistance syndrome

- Metabolic syndrome (MetS), a major contributor to cardiovascular disease and diabetes, is considered to be among the most common public health problems worldwide(Hosseini-Esfahani et al., 2021).

Problem Statement

Research Question 1

Q1) Is there any association between various demographics (age, sex, race) and health factors like (BMI, uric acid levels, blood glucose, and HDL) with the prevalence of metabolic syndrome? This project will use statistical methods to analyze how these factors influence metabolic syndrome.

Null Hypothesis: There is no significant association between demographics and health factors with prevalence of metabolic syndrome.

Alternate Hypothesis: There is a significant association between demographic and health factors with prevalence of metabolic syndrome.

Statistical Methods : Chi-Square Test ,Correlation Analysis, Simple Logistic Regression

Research Question 2

Q2) Which demographic or health factor has the highest impact on the prevalence of metabolic syndrome among individuals with certain risk factors?

Null Hypothesis: There is no significant difference in the prevalence of metabolic syndrome among individuals with certain risk factors based on demographic or health factors.

Alternate Hypothesis: At least one demographic or health factor has a significant impact on the prevalence of metabolic syndrome among individuals with certain risk factors.

Statistical Method: Multiple Logistic Regression

Dataset

The dataset is a Prevalence of Metabolic Syndrome dataset from Kaggle contains information on individuals with metabolic syndrome.

Dataset Link: https://www.kaggle.com/datasets/antimoni/metabolic-syndrome

It comprises 2402 rows and 15 columns. Each row represents an individual, and columns include attributes such as demographic, physiological, and health factors, as well as the presence or absence of metabolic syndrome.

VARIBLES

Categorical Variables	Numerical Variables
Sex Race Metabolic Syndrome	Age BMI HDL Blood Glucose Uric Acid

Data Importing

The CSV file containing the prevalence of metabolic syndrome was uploaded to R using the “read.csv” code.

```{r}
df<-read.csv('/Users/ashac/Downloads/Metabolic Syndrome.csv')
head(df)
```

Following the dataset upload, we opted to look for number of missing values.

```{r}
#Checking the missing values
col_missing <- colSums(is.na(df))
col_missing
```

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

DATA DESCRIPTION:

We examined the first few rows of the modified dataset with the command "head(data)" and obtained a summary of the dataset using "summary(data).”

```r
head(df)
```

	seqn <int>	Age <int>	Sex <chr>	Marital <chr>	Income <int>	Race <chr>	WaistCirc <dbl>	BMI <dbl>	Albuminuria <int>
1	62161	22	Male	Single	8200	White	81.0	23.3	0
2	62164	44	Female	Married	4500	White	80.1	23.2	0
3	62169	21	Male	Single	800	Asian	69.6	20.1	0
4	62172	43	Female	Single	2000	Black	120.4	33.3	0
5	62177	51	Male	Married	NA	Asian	81.1	20.1	0
6	62178	80	Male	Widowed	300	White	112.5	28.5	0

```r
summary(Metabolic_Syndrome)
```

```
      seqn            Age            Sex                Race                BMI       
 Min.   :62161   Min.   :20.00   Length:2375        Length:2375        Min.   :13.4  
 1st Qu.:64563   1st Qu.:34.00   Class :character   Class :character   1st Qu.:24.0  
 Median :67058   Median :48.00   Mode  :character   Mode  :character   Median :27.7  
 Mean   :67028   Mean   :48.67                                         Mean   :28.7  
 3rd Qu.:69501   3rd Qu.:63.00                                         3rd Qu.:32.1  
 Max.   :71915   Max.   :80.00                                         Max.   :68.7  
    UricAcid       BloodGlucose        HDL          MetabolicSyndrome
 Min.   : 1.800   Min.   : 39.0   Min.   : 14.00   Min.   :0.000     
 1st Qu.: 4.500   1st Qu.: 92.0   1st Qu.: 43.00   1st Qu.:0.000     
 Median : 5.400   Median :100.0   Median : 51.00   Median :0.000     
 Mean   : 5.481   Mean   :108.3   Mean   : 53.36   Mean   :0.344     
 3rd Qu.: 6.400   3rd Qu.:110.0   3rd Qu.: 62.00   3rd Qu.:1.000     
 Max.   :11.300   Max.   :382.0   Max.   :156.00   Max.   :1.000     
```

DATA DESCRIPTION

We explored the dimensions of the dataset using "dim(data)," revealing the number of rows and columns.

```{r}
dim(df)
```

```
[1] 2401   15
```

We examined the structure of the dataset with "str(data)," obtaining information about the variable types and their respective attributes.

```{r}
#TYPE OF DATA
str(Metabolic_Syndrome)
```

```
'data.frame':   2375 obs. of  9 variables:
 $ seqn            : int  62161 62164 62169 62172 62177 62178 62184 62189 62195 62199 ...
 $ Age             : int  22 44 21 43 51 80 26 30 35 57 ...
 $ Sex             : chr  "Male" "Female" "Male" "Female" ...
 $ Race            : chr  "White" "White" "Asian" "Black" ...
 $ BMI             : num  23.3 23.2 20.1 33.3 20.1 28.5 22.1 22.4 28.2 28 ...
 $ UricAcid        : num  4.9 4.5 5.4 5 5 4.8 5.4 6.7 6.7 6 ...
 $ BloodGlucose    : int  92 82 107 104 95 105 87 83 94 100 ...
 $ HDL             : int  41 28 43 73 43 47 61 48 46 35 ...
 $ MetabolicSyndrome: int  0 0 0 0 0 0 0 0 0 1 ...
```

DATA CLEANING

```r
#PREPROCESSING (Only BMI has 26 null values)
Metabolic_Syndrome <- Metabolic_Syndrome[!is.na(Metabolic_Syndrome$BMI), ]
summary(Metabolic_Syndrome)


#COUNT OF NUMBER OF ROWS
nrow(Metabolic_Syndrome)
```

```
      seqn            Age            Sex                Race          
 Min.   :62161   Min.   :20.00   Length:2375        Length:2375       
 1st Qu.:64563   1st Qu.:34.00   Class :character   Class :character  
 Median :67058   Median :48.00   Mode  :character   Mode  :character  
 Mean   :67028   Mean   :48.67                                        
 3rd Qu.:69501   3rd Qu.:63.00                                        
 Max.   :71915   Max.   :80.00                                        
[1] 2375
```

```r
anyDuplicated(Metabolic_Syndrome)
```

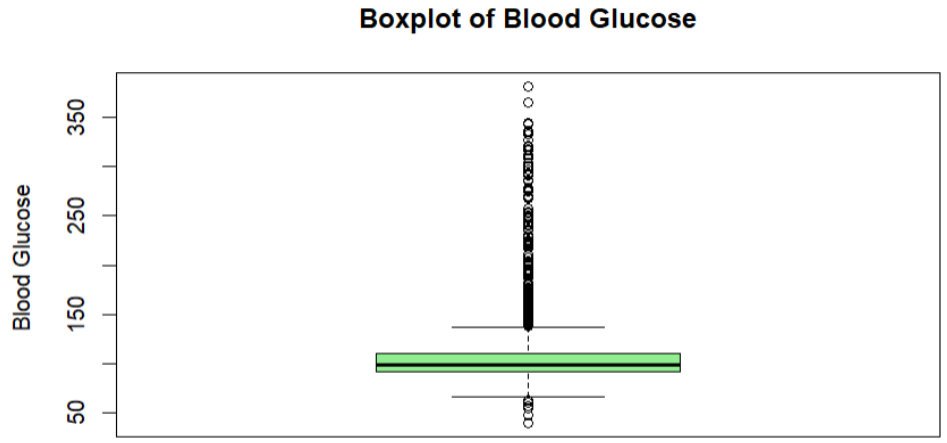
```
[1] 0
```

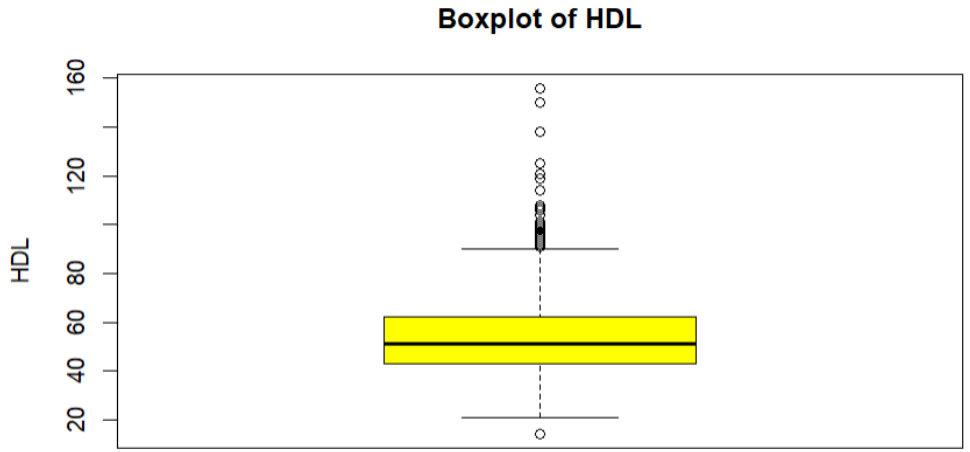
- ✓ We removed the unnecessary columns based on our objectives.
- ✓ After removing null values we resulted in 2375 rows.
- ✓ We searched for duplicate values in our dataset and determined that there were none present.

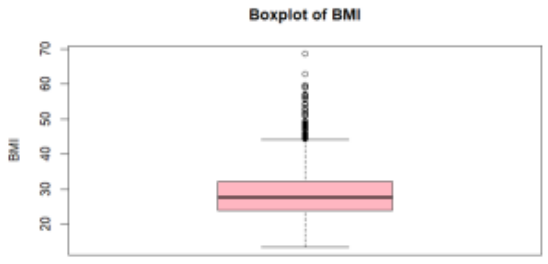
Identifying Outliers

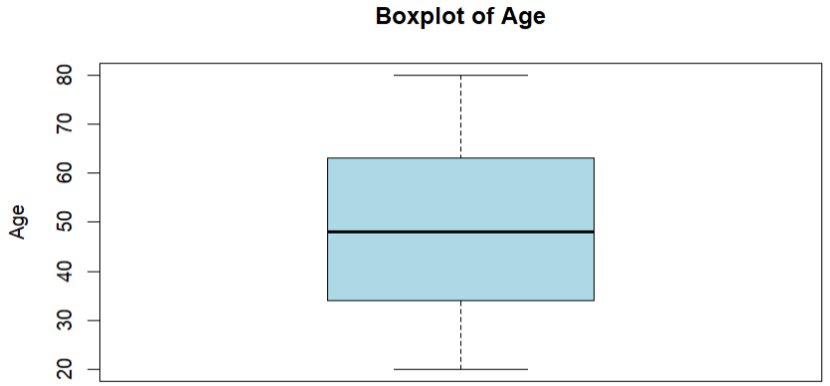
```r
# Counting the number of outliers
count_outliers <- function(data, column_name) {
  # Calculate the first and third quartiles
  q1 <- quantile(data[[column_name]], 0.25)
  q3 <- quantile(data[[column_name]], 0.75)

  # Calculate the interquartile range (IQR)
  iqr <- q3 - q1

  # Define the lower and upper bounds for outliers
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr

  # Identify outliers
  outliers <- data[[column_name]] < lower_bound | data[[column_name]] > upper_bound

  # Count the number of outliers
  num_outliers <- sum(outliers)

  # Return the count of outliers
  return(num_outliers)
}
# Count outliers for BMI
num_outliers_bmi <- count_outliers(Metabolic_Syndrome, "BMI")
cat("Number of outliers for BMI:", num_outliers_bmi, "\n")

# Count outliers for Blood Glucose Level
num_outliers_bg <- count_outliers(Metabolic_Syndrome, "BloodGlucose")
cat("Number of outliers for Blood Glucose Level:", num_outliers_bg, "\n")

num_outliers_hdl <- count_outliers(Metabolic_Syndrome, "HDL")
cat("Number of outliers for HDL:", num_outliers_hdl, "\n")

num_outliers_bg <- count_outliers(Metabolic_Syndrome, "UricAcid")
cat("Number of outliers for UricAcid:", num_outliers_bg, "\n")
```

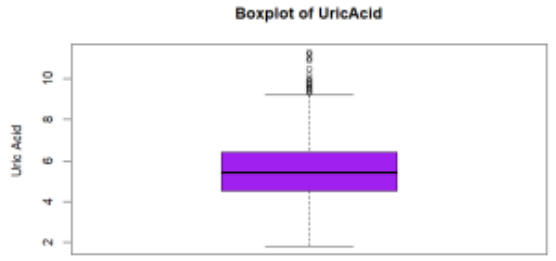
```
Number of outliers for BMI: 67
Number of outliers for Blood Glucose Level: 218
Number of outliers for HDL: 53
Number of outliers for UricAcid: 29
```


Boxplot of Blood Glucose
Blood Glucose
50
150
250
350


Boxplot of HDL
HDL
20
40
60
80
120
160


Boxplot of BMI
BMI
20
30
40
50
60
70


Boxplot of Age
Age
20
30
40
50
60
70
80


Boxplot of UricAcid
Uric Acid
2
4
6
8
10

CAPPING OUTLIERS

```r
#Capping outliers for required columns.
# Function to cap outliers for specified columns
cap_outliers <- function(data, columns) {
  for (col in columns) {
    # Calculate the first and third quartiles
    q1 <- quantile(data[[col]], 0.25)
    q3 <- quantile(data[[col]], 0.75)

    # Calculate the interquartile range (IQR)
    iqr <- q3 - q1

    # Define the lower and upper bounds for outliers
    lower_bound <- q1 - 1.5 * iqr
    upper_bound <- q3 + 1.5 * iqr

    # Cap outliers
    data[[col]][data[[col]] < lower_bound] <- lower_bound
    data[[col]][data[[col]] > upper_bound] <- upper_bound
  }

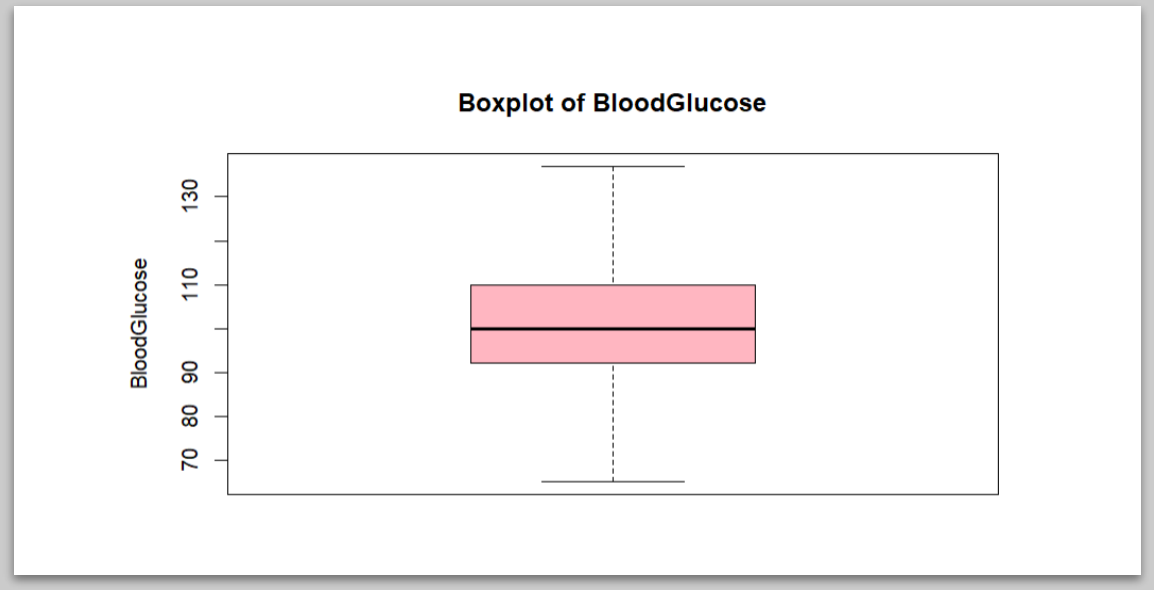
# Columns to cap outliers for (e.g., "BMI" and "BloodGlucoseLevel")
columns <- c("BMI", "BloodGlucose", "HDL", "UricAcid")

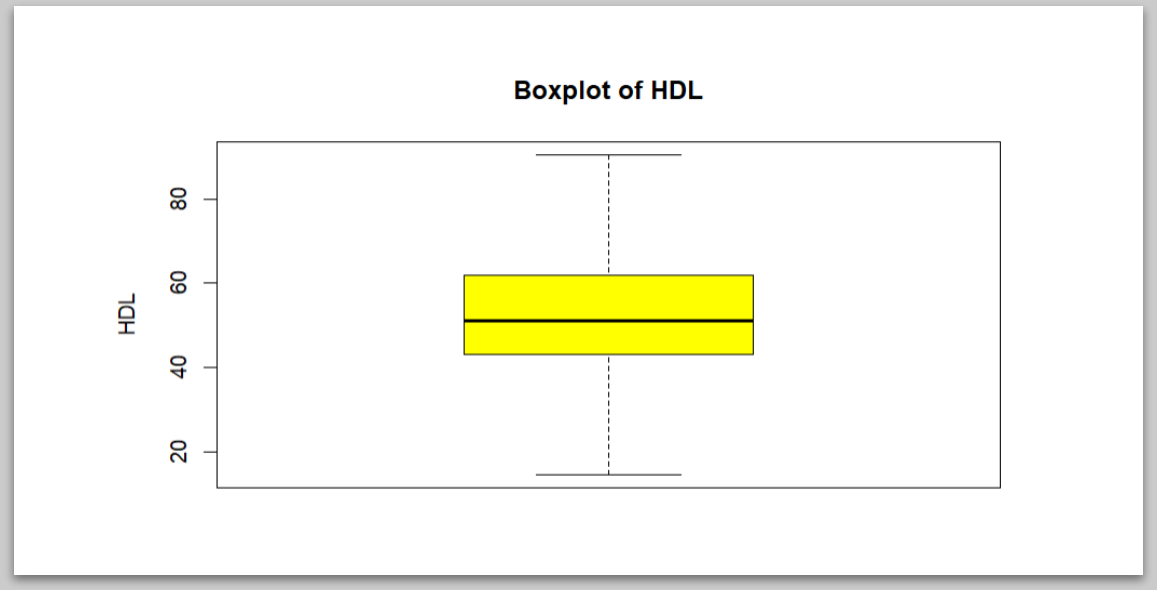
# Cap outliers for specified columns
Metabolic_Syndrome_final <- cap_outliers(Metabolic_Syndrome, columns)

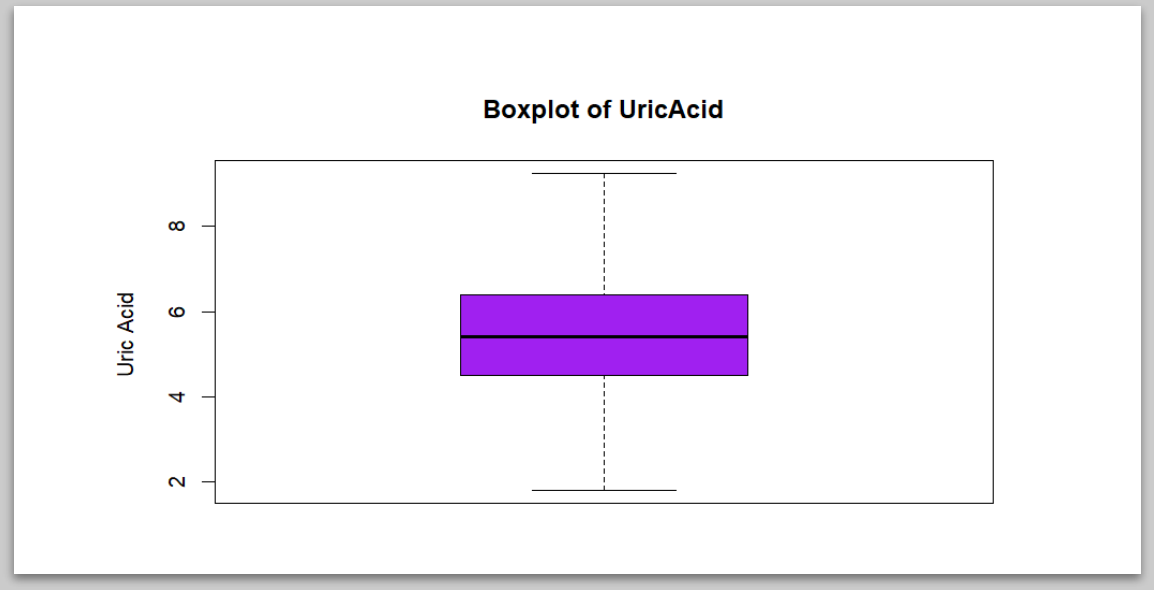
# View the updated dataset
Metabolic_Syndrome_final

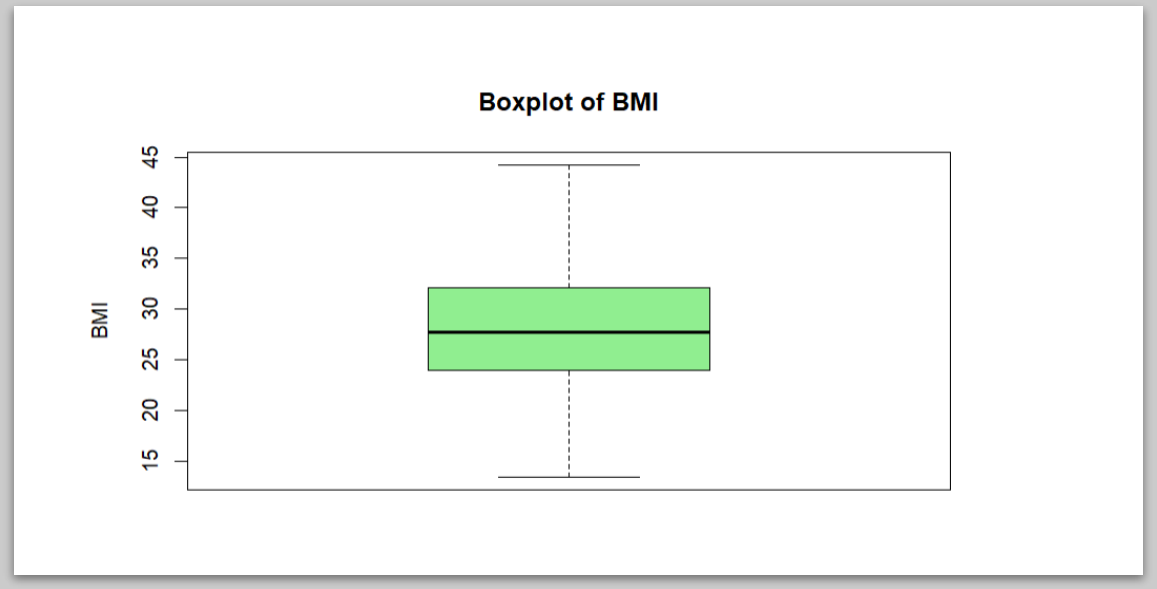
```

We done capping method for outliers in specified columns of a dataset by replacing values beyond 1.5 times the interquartile range with the nearest inner quartile boundary.

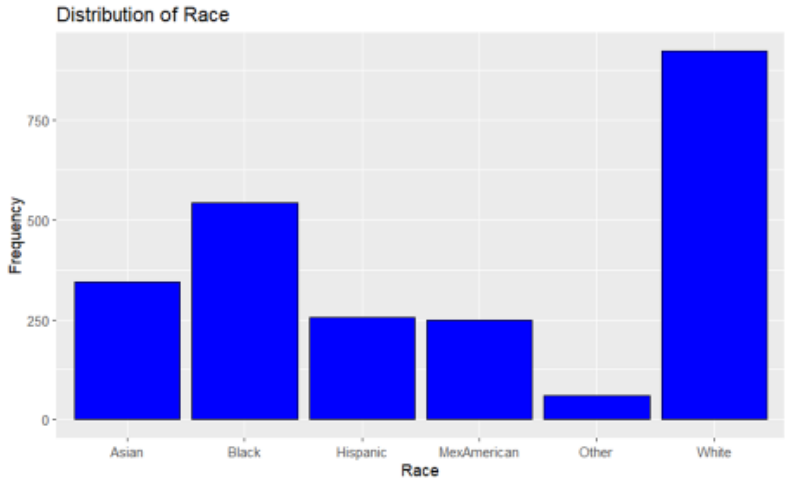

Boxplot of BloodGlucose
BloodGlucose
70
80
90
110
130

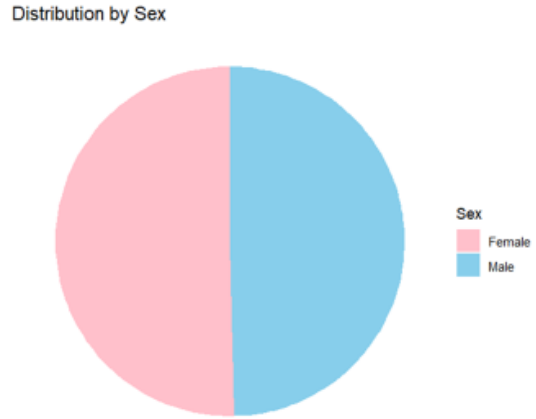

Boxplot of HDL
HDL
20
40
60
80

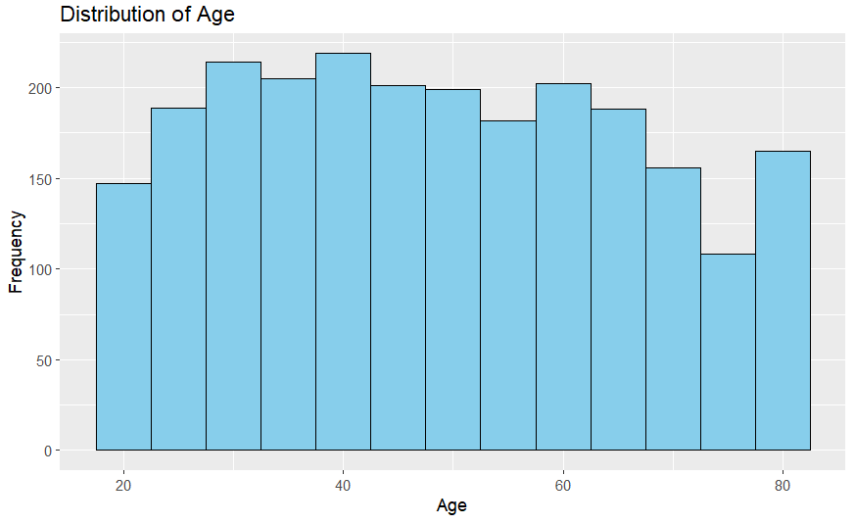

Boxplot of UricAcid
Uric Acid
2
4
6
8

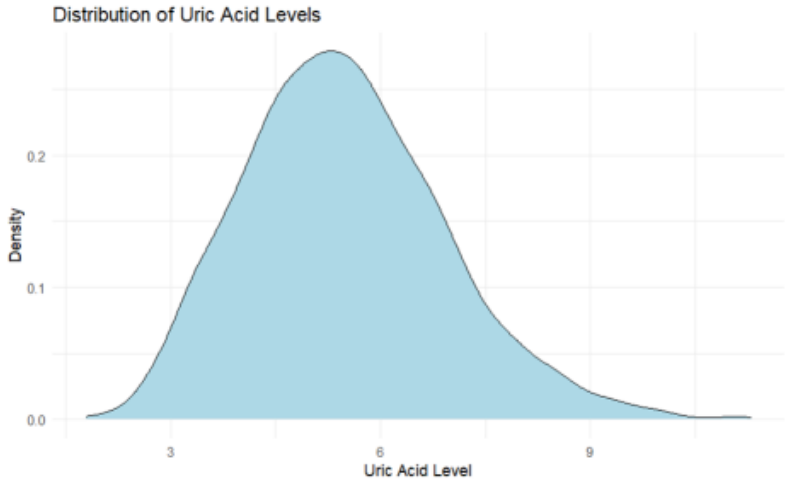

Boxplot of BMI
BMI
15
20
25
30
35
40
45

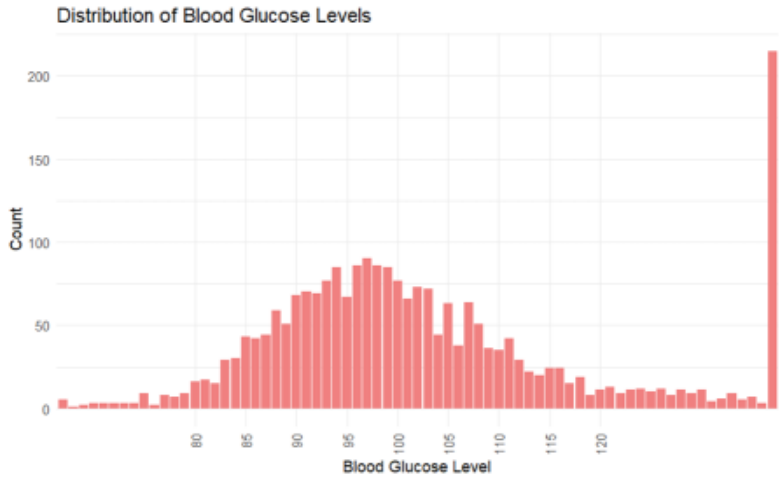
DATA VISUALIZATION


Distribution of Race
Frequency
750
500
250
0
Asian
Black
Hispanic
MexAmerican
Other
White
Race


Distribution by Sex
Sex
Female
Male

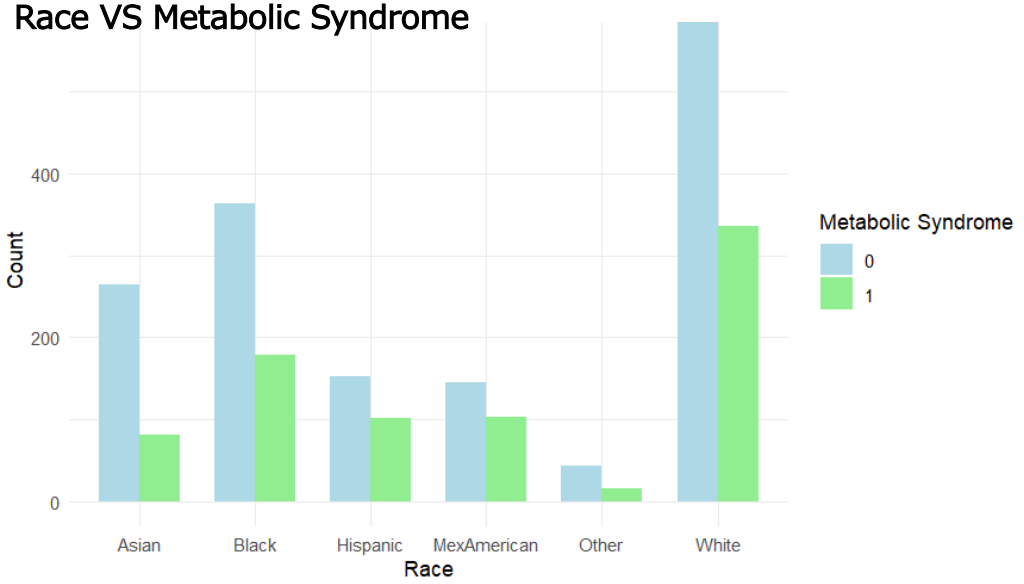

Distribution of Age
Frequency
200
150
100
50
0
20
40
60
80
Age


Distribution of Uric Acid Levels
Density
0.2
0.1
0.0
3
6
9
Uric Acid Level

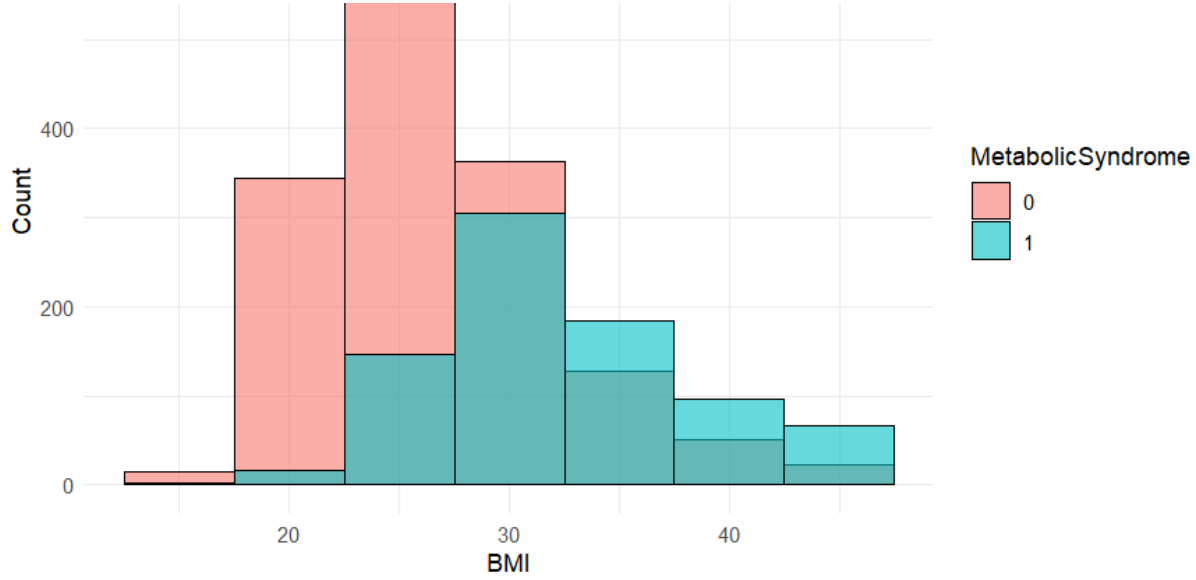

Distribution of Blood Glucose Levels
Count
200
150
100
50
0
80
85
90
95
100
105
110
115
120
Blood Glucose Level

HDL VS Metabolic Syndrome

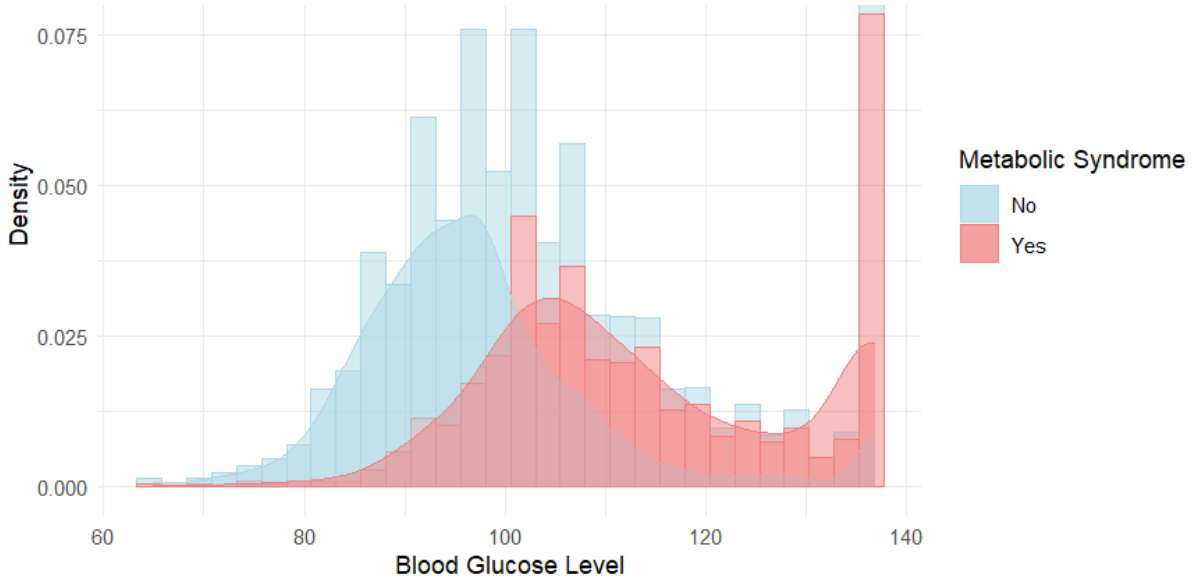
Race VS Metabolic Syndrome



BMI VS Metabolic Syndrome



BloodGlucose VS Metabolic Syndrome



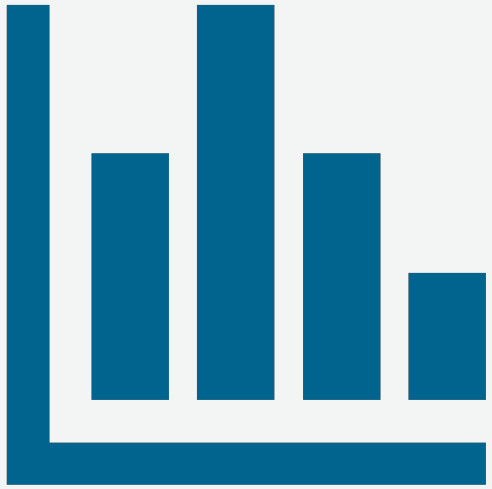
EXPLORATORY DATA ANALYSIS

Descriptive Statistics:

```
#STATISTICAL ANALYSIS
#Exploratory data Analysis

```{r}
summary(Metabolic_Syndrome_final)
```
```

```
      seqn            Age            Sex                Race                BMI        
 Min.   :62161   Min.   :20.00   Length:2375        Length:2375        Min.   :13.40  
 1st Qu.:64563   1st Qu.:34.00   Class :character   Class :character   1st Qu.:24.00  
 Median :67058   Median :48.00   Mode  :character   Mode  :character   Median :27.70  
 Mean   :67028   Mean   :48.67                                         Mean   :28.55  
 3rd Qu.:69501   3rd Qu.:63.00                                         3rd Qu.:32.10  
 Max.   :71915   Max.   :80.00                                         Max.   :44.25  
    UricAcid      BloodGlucose        HDL        MetabolicSyndrome
 Min.   :1.800   Min.   : 65.0   Min.   :14.50   Min.   :0.000     
 1st Qu.:4.500   1st Qu.: 92.0   1st Qu.:43.00   1st Qu.:0.000     
 Median :5.400   Median :100.0   Median :51.00   Median :0.000     
 Mean   :5.474   Mean   :102.9   Mean   :53.12   Mean   :0.344     
 3rd Qu.:6.400   3rd Qu.:110.0   3rd Qu.:62.00   3rd Qu.:1.000     
 Max.   :9.250   Max.   :137.0   Max.   :90.50   Max.   :1.000     
```



STATISTICAL ANALYSIS

CHECKING NORMALITY OF DATA

Is data normally distributed?

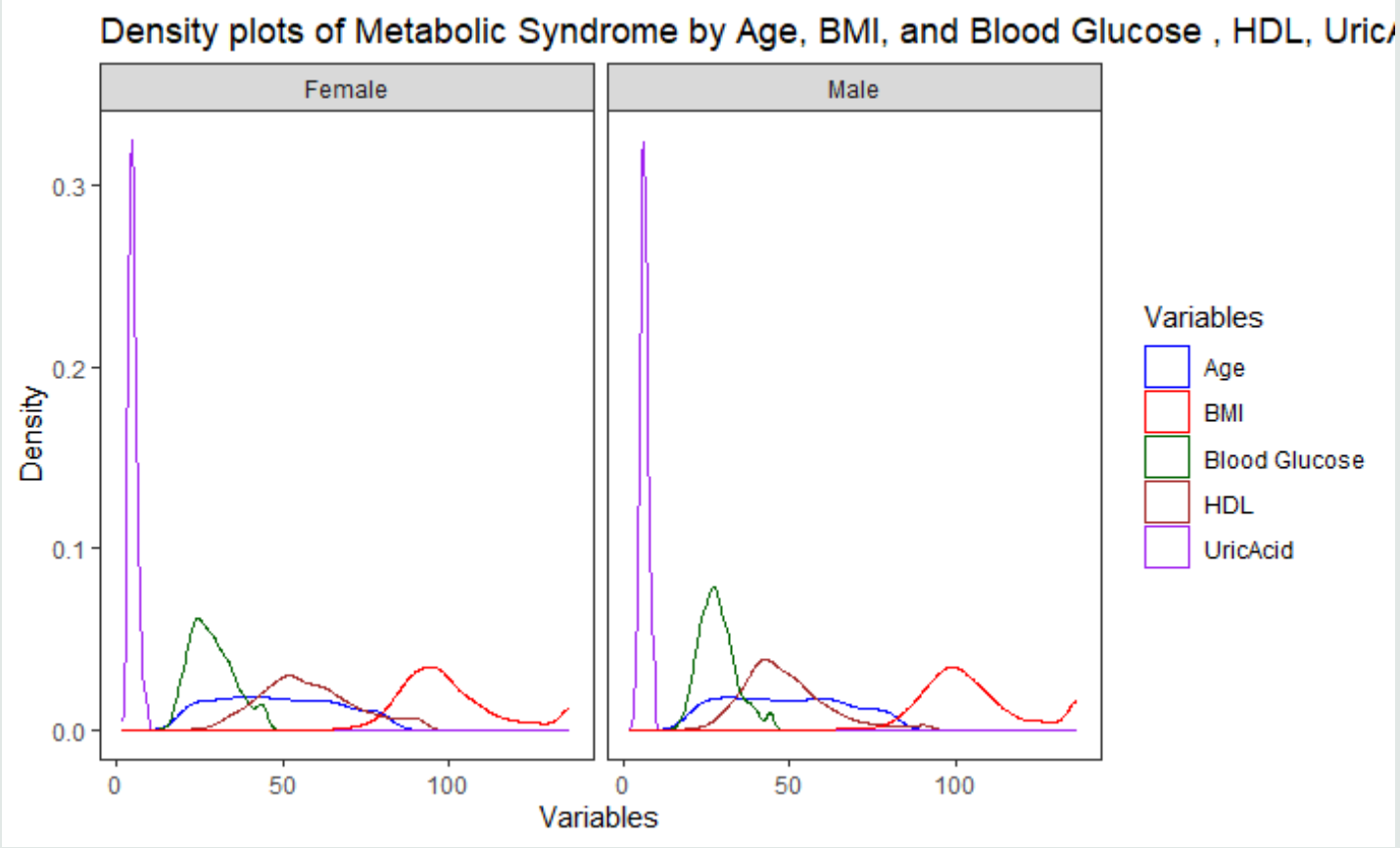
The extremely small p-values from Shapiro-Wilk tests suggest strong evidence against normality for all variables tested in the dataset "Metabolic_Syndrome_final," indicating non-normal distribution of the data.

```r
```{r}
numeric_columns = Metabolic_Syndrome_final[sapply(Metabolic_Syndrome_final,is.numeric)]
normality_results = sapply(numeric_columns,function(x) {
  shapiro.test = shapiro.test(x)
  p_value = shapiro.test$p.value
})
normality_results
```
```

```
        seqn          Age          BMI     UricAcid BloodGlucose          HDL
7.679795e-27 8.851966e-26 3.886852e-23 4.517758e-13 3.090104e-33 3.288860e-23
```

DENSITY PLOT

Based on the density plot, the distributions of Age, BMI, HDL, Uric Acid, and Blood Glucose appear to be right-skewed, suggesting non-normality in the data.



RQ1

Is there any association between various demographics (age, sex, race) and health factors like (BMI, uric acid levels, blood glucose, and HDL) with the prevalence of metabolic syndrome?

Null Hypothesis: There is no significant association between demographic and health factors with prevalence of metabolic syndrome.

Alternate Hypothesis: There is a significant association between demographic and health factors with prevalence of metabolic syndrome.

CHI-SQUARE TESTS

Is there an association between Sex and Metabolic Syndrome ?

As the observed p-value (0.3303) is more than 0.05, we fail to reject null hypothesis stating there is no significant association between Sex and Metabolic Syndrome

```r
# Create a contingency table of Sex and Metabolic Syndrome
contingency_table <- table(Metabolic_Syndrome_final$Sex,
                           Metabolic_Syndrome_final$MetabolicSyndrome)

# Perform chi-square test
chi_square_result <- chisq.test(contingency_table)

# Print the results
print(chi_square_result)

# Check for significance and provide interpretation
if (chi_square_result$p.value < 0.05) {
  cat("\nThere is a significant association between Sex and Metabolic Syndrome.\n")
} else {
  cat("\nThere is no significant association between Sex and Metabolic Syndrome.\n")
}
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0.94767, df = 1, p-value = 0.3303


There is no significant association between Sex and Metabolic Syndrome.
```

CHI-SQUARE TESTS

Is there an association between Race and Metabolic Syndrome ?

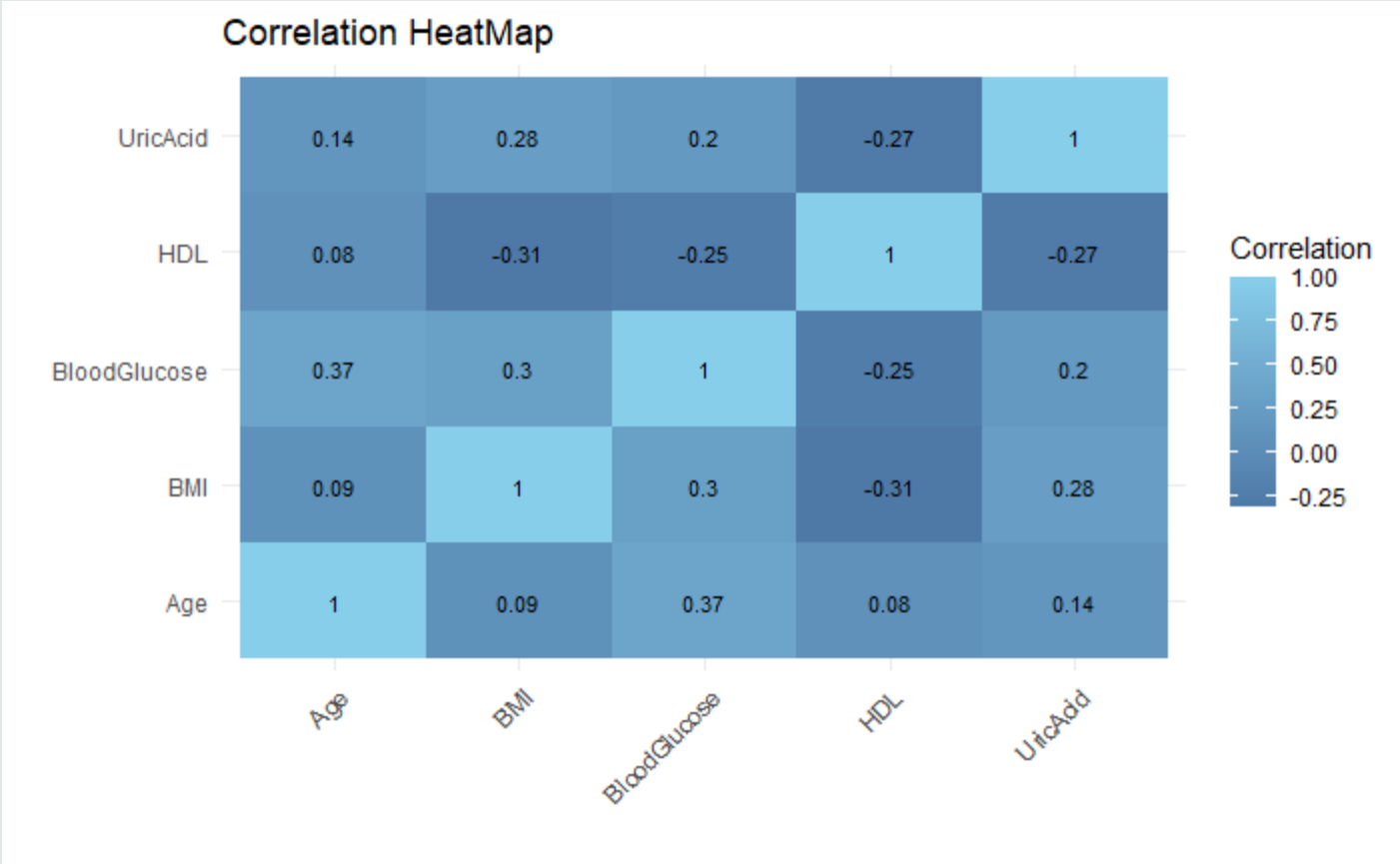
As the observed p-value (1.342e-05) Is less than 0.05, we reject null hypothesis stating that there is a significant association between Race and Metabolic Syndrome

```r
# Create a contingency table of Race and Metabolic Syndrome
contingency_table_race <- table(Metabolic_Syndrome_final$Race,
                                Metabolic_Syndrome_final$MetabolicSyndrome)

# Perform chi-square test for Race
chi_square_result_race <- chisq.test(contingency_table_race)

# Print the results for Race
print(chi_square_result_race)

# Check for significance and provide interpretation
if (chi_square_result_race$p.value < 0.05) {
  cat("\nThere is a significant association between Race and Metabolic Syndrome.\n")
} else {
  cat("\nThere is no significant association between Race and Metabolic Syndrome.\n")
}
```

```
        Pearson's Chi-squared test

data:  contingency_table_race
X-squared = 30.209, df = 5, p-value = 1.342e-05


There is a significant association between Race and Metabolic Syndrome.
```

Correlation:

Correlation heatmap that visually represents the correlation coefficients between five different variables: Age, BMI, BloodGlucose, HDL, and UricAcid.



After performing a correlation analysis on our dataset, we examined the relationships and degree of associations between the different variables under study.

- Age is positively correlated with BMI, blood glucose, and uric acid.
- Age is negatively correlated with HDL cholesterol.
- BMI is positively correlated with blood glucose and uric acid.
- BMI is negatively correlated with HDL cholesterol.
- Blood glucose is positively correlated with uric acid.
- Blood glucose is negatively correlated with HDL cholesterol.

We found a moderate positive correlation between Age and BloodGlucose ($r = 0.37$), suggesting that as age increases, blood glucose levels tend to rise as well. There was also a moderate negative correlation between BMI and HDL ($r = -0.31$), indicating an inverse relationship where higher BMI is associated with lower HDL levels.

WHY Logistic Regression?

- ✓ **Binary Outcome:** MetabolicSyndrome is a binary outcome variable (presence or absence), making logistic regression suitable for modeling such categorical data.
- ✓ **Predicting Probability:** Logistic regression estimates the probability of an event (having MetabolicSyndrome) based on predictor variables (like Age).
- ✓ **Non-Normal Data:** It does not assume normality of data, so it's appropriate even if your data is not normally distributed.
- ✓ **Interpretability:** It provides interpretable results in terms of odds ratios, helping understand the effect of Age on the likelihood of having MetabolicSyndrome.
- ✓ **Handling Continuous Predictors:** It can handle both continuous (like Age) and categorical predictors efficiently.

Age and Metabolic Syndrome

The odds ratio for Age (1.031683) suggests that for every one-year increase in age, the odds of having metabolic syndrome increase by a factor of approximately 1.03 times.

This suggests that age is positively associated with the likelihood of having metabolic syndrome.

```
Call:
glm(formula = MetabolicSyndrome ~ Age, family = binomial, data = Metabolic_Syndrome_final)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.20847    0.14105  -15.66   <2e-16 ***
Age          0.03119    0.00260   12.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 2904.1  on 2373  degrees of freedom
AIC: 2908.1

Number of Fisher Scoring iterations: 4

Odds Ratio for Age: 1.031683
For every one year increase in age, the odds of having metabolic syndrome
    increase by a factor of 1.03 times, holding all other variables constant.
```

BMI and Metabolic Syndrome

The odds ratio for BMI (1.188187) suggests that for every one-unit increase in BMI, the odds of having metabolic syndrome increase by a factor of approximately 1.19 times, or an increase of about 18.82%.

This suggests that BMI is positively associated with the likelihood of having metabolic syndrome.

```
Call:
glm(formula = MetabolicSyndrome ~ BMI, family = binomial, data = Metabolic_Syndrome_final)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.682010   0.268580  -21.16   <2e-16 ***
BMI          0.172429   0.008945   19.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 2570.8  on 2373  degrees of freedom
AIC: 2574.8

Number of Fisher Scoring iterations: 4

Odds Ratio for BMI: 1.188187
For every one unit increase in BMI, the odds of having metabolic syndrome
    increase by a factor of 1.19 times, holding all other variables constant.
```

BloodGlucose and Metabolic Syndrome

For every one-unit increase in blood glucose level, the odds of having metabolic syndrome increase by approximately 8.18%.

This suggests that individuals with higher blood glucose levels have around 8.18% higher odds of having metabolic syndrome

```
Call:
glm(formula = MetabolicSyndrome ~ BloodGlucose, family = binomial,
    data = Metabolic_Syndrome_final)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.858154   0.396141  -22.36   <2e-16 ***
BloodGlucose  0.078647   0.003759   20.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 2434.1  on 2373  degrees of freedom
AIC: 2438.1

Number of Fisher Scoring iterations: 4

Odds Ratio for Blood Glucose: 1.081822
For every one unit increase in blood glucose level, the odds of having
    metabolic syndrome increase by a factor of 1.08 times.
```

HDL and Metabolic Syndrome

For every one unit increase in HDL, the odds of having metabolic syndrome decrease by a factor of approximately 0.93 times.

The intercept implies that when HDL levels are zero, the estimated odds of having metabolic syndrome are significantly higher.

```
Call:
glm(formula = MetabolicSyndrome ~ HDL, family = binomial, data = Metabolic_Syndrome_final)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.126841   0.215954   14.48   <2e-16 ***
HDL         -0.074356   0.004325  -17.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 2653.2  on 2373  degrees of freedom
AIC: 2657.2

Number of Fisher Scoring iterations: 4

Odds Ratio for HDL: 0.9283408
For every one unit increase in HDL, the odds of having metabolic syndrome
    increase by a factor of 0.93 times.
```

UricAcid and Metabolic Syndrome

For every one unit increase in UricAcid, the odds of having metabolic syndrome increase by approximately 45.64%.

This suggests a positive association between UricAcid levels and the likelihood of developing metabolic syndrome

```
Call:
glm(formula = MetabolicSyndrome ~ UricAcid, family = binomial,
    data = Metabolic_Syndrome_final)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.74226    0.18996  -14.44   <2e-16 ***
UricAcid     0.37597    0.03261   11.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 2914.5  on 2373  degrees of freedom
AIC: 2918.5

Number of Fisher Scoring iterations: 4

Odds Ratio for UricAcid: 1.456402
For every one unit increase in UricAcid, the odds of having
    metabolic syndrome increase by a factor of 1.46 times.
```

Conclusion RQ1

- ✓ While, the chi-square test showed mixed results for the association between sex and metabolic syndrome, the correlation analysis and logistic regression results provide strong evidence of significant associations between various demographic factors (age) and health factors (BMI, blood glucose, HDL, and uric acid) with the prevalence of metabolic syndrome.
- ✓ As the p-values for age, BMI, blood glucose, HDL, and uric acid are all less than 0.05, we can conclude that there is a significant association between demographic and health factors with the prevalence of metabolic syndrome, supporting the alternate hypothesis.

WHY not Kruskal-Waliis H test or Mann-Whitney U test ?

KRUSKAL-WALLIS H TEST

- Our dependent variable is not continuous or ordinal.
- Dependent is a binary categorical variable.
- Variable of interest does not have more than 2 independent groups.

MANN-WHITNEY U TEST

- Our dependent variable is not continuous or ordinal.
- It is binary categorical variable.
- Our independent variables does not consist of 2 categorical independent groups

RQ2

Which demographic or health factor has the highest impact on the prevalence of metabolic syndrome among individuals with certain risk factors?

Null Hypothesis (H0): There is no significant difference in the prevalence of metabolic syndrome among individuals with certain risk factors based on demographic or health factors.

Alternate Hypothesis (H1): At least one demographic or health factor has a significant impact on the prevalence of metabolic syndrome among individuals with certain risk factors.

Multiple Logistic Regression

Demographics with Metabolic Syndrome

We assessed the impact of demographic variables (such as Age, Sex, and Race) on the outcome variable, metabolic syndrome.

This analysis aims to identify which demographic factors have the greatest influence on metabolic syndrome

```
Call:
glm(formula = as.formula(paste(outcome_var1, "~", paste(variables1,
    collapse = " + "))), family = binomial, data = Metabolic_Syndrome_final)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.709339   0.191045 -14.182  < 2e-16 ***
Age              0.031309   0.002645  11.836  < 2e-16 ***
SexMale          0.081120   0.089967   0.902 0.367236
RaceBlack        0.399407   0.161006   2.481 0.013113 *
RaceHispanic     0.667227   0.185622   3.595 0.000325 ***
RaceMexAmerican  0.898895   0.186159   4.829 1.37e-06 ***
RaceOther        0.286902   0.322467   0.890 0.373622
RaceWhite        0.462566   0.148862   3.107 0.001888 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 2876.0  on 2367  degrees of freedom
AIC: 2892

Number of Fisher Scoring iterations: 4
```

- RaceMexAmerican shows the highest impact on the likelihood of the outcome variable.
- Specifically, individuals identifying as MexAmerican have approximately 2.46 times higher odds of experiencing the outcome compared to the reference category, after controlling for other variables in the model.
- This effect is statistically significant with a p-value of approximately 1.37e-06.

```r
# Get summary of the model
summary_logit <- summary(logit_model1)

# Extract coefficient estimates and p-values
coefficients <- coef(logit_model1)
p_values <- summary_logit$coefficients[, 4]

# Exclude intercept (if present)
if ("(Intercept)" %in% names(coefficients)) {
  coefficients <- coefficients[-1]
  p_values <- p_values[-1]
}

# Calculate absolute coefficients
abs_coefficients <- abs(coefficients)

# Find the index of the variable with the highest absolute coefficient
highest_impact_index <- which.max(abs_coefficients)
```

```
Variable with the highest impact:  RaceMexAmerican
Coefficient estimate:  0.8988951
P-value:  1.37472e-06
```

Health Factors with Metabolic Syndrome

- ✓ We assessed the impact of clinical factor variables (such as BMI, UricAcid, BloodGlucose, HDL) on the outcome variable, metabolic syndrome.
- ✓ This analysis aims to identify which health factors have the greatest influence on metabolic syndrome

```
Call:
glm(formula = as.formula(paste(outcome_var2, "~", paste(variables2,
    collapse = " + "))), family = binomial, data = Metabolic_Syndrome_final)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.469499   0.624559  -15.16   <2e-16 ***
BMI          0.134302   0.010419   12.89   <2e-16 ***
UricAcid     0.086849   0.042366    2.05   0.0404 *
BloodGlucose 0.069759   0.004174   16.71   <2e-16 ***
HDL         -0.057337   0.005122  -11.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3057.4  on 2374  degrees of freedom
Residual deviance: 1966.3  on 2370  degrees of freedom
AIC: 1976.3

Number of Fisher Scoring iterations: 5
```

- Among all the variables BMI (Body Mass Index) demonstrates the highest impact on the likelihood of the outcome variable, metabolic syndrome.

- The coefficient estimate of 0.1343017 indicates that for every one-unit increase in BMI, the log odds of experiencing metabolic syndrome increase by approximately 0.134, holding all other variables constant.

- The extremely low p-value (2.549866e-32) suggests that this relationship is statistically significant, indicating strong evidence against the null hypothesis that there is no association between BMI and metabolic syndrome.

```r
# Get summary of the model
summary_logit2 <- summary(logit_model2)

# Extract coefficient estimates and p-values
coefficients2 <- coef(logit_model2)
p_values2 <- summary_logit2$coefficients2[, 4]

# Exclude intercept (if present)
if ("(Intercept)" %in% names(coefficients2)) {
  coefficients2 <- coefficients2[-1]
  p_values2 <- p_values2[-1]
}

# Calculate absolute coefficients
abs_coefficients2 <- abs(coefficients2)

# Find the index of the variable with the highest absolute coefficient
highest_impact_index2 <- which.max(abs_coefficients2)
```

```
Variable with the highest impact:  BMI
Coefficient estimate:  0.1343017
P-value:  2.549866e-32
```

Conclusion RQ2

- ✓ The analysis of demographic and health factors on the prevalence of metabolic syndrome among individuals shows demographic and certain health factors have a significant impact.
- ✓ Regarding demographic factors, the race/ethnicity variable "**MexAmerican**" shows the highest impact.
- ✓ Concerning health factors, **Body Mass Index** (BMI) demonstrates the highest impact on the likelihood of metabolic syndrome.
- ✓ Based on observed p-values, we accept the alternate hypothesis that at least one demographic (MexAmerican race/ethnicity) and one health factor (BMI) have a significant impact on the prevalence of metabolic syndrome among individuals with certain risk factors.

Limitations:

- ✓ Our study included 2375 rows, but a larger sample might provide more reliable results.
- ✓ The chi-square test only examines the association between two variables,like Sex and Metabolic Syndrome or Race and Metabolic Syndrome. However, there may be other confounding variables that influence this relationship, such as age, health factors etc.
- ✓ Just because there's an association between Race and MetabolicSyndrome doesn't mean one causes the other. We need more research to understand the relationship better.
- ✓ Due to the binary nature of the outcome variable (Metabolic Syndrome), we cannot create a correlation heatmap to visualize the relationships between predictor variables and the outcome variable.
- ✓ While the regression analysis identifies associations between demographic and health factors with the prevalence of metabolic syndrome, it does not establish causality.

References:

Hosseini-Esfahani, F., Alafchi, B., Cheraghi, Z., Doosti-Irani, A., Mirmiran, P., Khalili, D., & Azizi, F. (2021). Using machine learning techniques to predict factors contributing to the incidence of metabolic syndrome in Tehran: Cohort study. *JMIR Public Health and Surveillance*, *7*(9), e27304. https://doi.org/10.2196/27304

What is metabolic syndrome? | NHLBI, NIH. (2022, May 18). NHLBI, NIH. https://www.nhlbi.nih.gov/health/metabolic-syndrome

THANK YOU