Project Report

Investigating Prevalence of Metabolic Syndrome

Asha Chirumamilla, Likhitha Arigala Pallavi Telu, Teja Pavani Jyesta, William Teske

Introduction

Metabolic syndrome is a collection of health conditions that often co-occur, raising the chances of developing heart disease, stroke, and type 2 diabetes (Mayo Clinic, 2021). Understanding the factors associated with metabolic syndrome is crucial for prevention and intervention efforts. In this report, we analyze data from the "Prevalence of Metabolic Syndrome" dataset to investigate the association between various demographics (Age, Sex, Race) and health factors (BMI, BloodGlucose, UricAcid, HDL) with prevalence of metabolic syndrome and identify the demographic or health factors with the highest impact on metabolic syndrome among individuals with certain risk factors.

Dataset Description: The dataset contains information on individuals with metabolic syndrome, comprising 2402 rows and 15 columns.

Categorical Variables	Numerical Variables
Race, Sex (Nominal) Metabolic syndrome (Binary Variable)	Age, BMI, BloodGlucose, UricAcid, HDL(Continuous variables)

Descriptive Statistics: We began our analysis by conducting descriptive statistics to gain a deeper understanding of the dataset. This involved summarizing key variables using measures of central tendency, variability, and distribution. Descriptive statistics provided insights into the characteristics of the study population, such as the average age, distribution of BMI, and prevalence of metabolic syndrome.

Data Cleaning: In the initial stages of data cleaning, we meticulously reviewed our dataset to identify and remove any unnecessary columns that did not contribute substantially to our research objectives. Following this, we conducted an assessment to identify missing values within the dataset, revealing that the "BMI" column contained a small number of null values, approximately 26 entries, representing less than 1% of the total dataset. Given the negligible proportion of missing values, we opted to address this issue by removing the corresponding rows containing null values, thereby ensuring the integrity of the dataset. Due to these data cleaning procedures, the final dataset was refined to include 2375 rows, reflecting the elimination of redundant columns and the removal of rows containing missing values in the "BMI" column. This comprehensive approach to data cleaning has enabled us in providing a solid foundation for investigating the association between demographic and health factors and the prevalence of metabolic syndrome.

Outlier Detection and Treatment: Upon plotting box plots, outliers were identified in predictor variables including BMI, HDL, blood glucose, and uric acid levels. To address this, we employed a capping technique whereby outliers were replaced with upper and lower bound values based on the outlier value. Following outlier treatment, subsequent visualization confirmed the absence of outliers, ensuring the robustness of our dataset.

Exploratory Data Analysis: As part of our data exploration process, we employed various visualization techniques to gain insights into our dataset. Predictor variables such as age, sex, blood glucose level, BMI, HDL, and uric acid levels were visualized using box plots, histograms, pie charts, and bar graphs. Notably, box plots were utilized for variables with continuous values to identify trends and distributions.

We further explored the relationship between predictor variables such as age, race, HDL, with metabolic syndrome status. Visualizations revealed significant trends, including lower HDL levels among individuals with metabolic syndrome than those without. Additionally, a higher prevalence of metabolic syndrome was observed among individuals with elevated blood glucose levels and advancing age. Furthermore, our data indicated a predominance of individuals of white race, underscoring the demographic composition of our dataset. Additionally, the demographic composition of our dataset highlighted the predominance of individuals of white race (See the Appendix below for the visualization plots).

Statistical Analysis:

Normality Assessment using the Shapiro-Wilk Test: In our data exploration process, we sought to understand the underlying distributional characteristics of our dataset by conducting the Shapiro-Wilk test for normality. The test revealed that our data deviated from a normal distribution, as evidenced by a p-value less than the conventional significance level of 0.05.

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

Based on the figure, The Shapiro-Wilk test results for the variables in our dataset – seqn, Age, BMI, Uric, Blood Glucose, and HDL – strongly suggest that none of these variables are normally distributed. Each variable has a very small p-value, far below the common significance threshold of 0.05, indicating a statistically significant rejection of the hypothesis that the data for these variables come from a normal distribution.

The Shapiro-Wilk test, commonly used for assessing normality in data sets, exhibits several notable limitations, particularly its high sensitivity to sample size and specific assumptions about

the data structure. Given the categorical nature of our response variable and the presence of non-normal data distribution, the Shapiro-Wilk test was a suitable choice for assessing normality in our dataset. Its ability to detect deviations from normality, even with non-parametric data, ensured the strength of our subsequent analyses. The Shapiro-Wilk test served as a foundational step in our data exploration process, providing insights into the distributional characteristics of our dataset. By identifying non-normal data distribution, we made informed decisions about selecting appropriate statistical techniques, enhancing the reliability of our analyses.

Chi-Square Test for Categorical Variables:

With our response variable being binary categorical, we employed the chi-square test to investigate potential associations between categorical predictor variables (sex and race) and the prevalence of metabolic syndrome. The results of the chi-square test revealed no significant association between sex and metabolic syndrome, suggesting that gender may not be a significant determinant of metabolic syndrome prevalence. In contrast, a significant association was observed between race and the occurrence of metabolic syndrome, indicating that race may play a role in predisposing individuals to metabolic syndrome.

```
        Pearson's Chi-squared test

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

Based off the Figure, given the p-value is much less than the typical significance level, you would reject their null hypothesis of no association. Thus, strong evidence suggests a statistical association between race and metabolic syndrome prevalence.

While the chi-square test is widely used for analyzing associations between categorical variables, its assumption of independence between categories may be violated in real-world datasets, potentially leading to biased results. Moreover, small sample sizes within certain categories or imbalances in cell frequencies could affect the validity of the test.

Logistic Regression Analysis:

To further explore the influence of predictor variables on the likelihood of metabolic syndrome occurrence, logistic regression analysis was conducted. Logistic regression was well-suited for our analysis, given the categorical nature of our response variable (metabolic syndrome) and the presence of non-normally distributed data. This statistical technique allows for the prediction of the probability of a binary outcome (metabolic syndrome presence or absence) based on one predictor variable. Its ability to model the probability of a binary outcome based on multiple predictor variables allowed us to explore the relationships between demographic and health

factors and metabolic syndrome prevalence. In our analysis, age, BMI, blood glucose level, HDL, and uric acid levels were included as predictor variables, with results presented in terms of odds ratios.

Model 1: Intercept of the estimated log odds of the outcome when Age is zero is –2.20847, again significant (p <2e-16). The coefficient of 0.03119 shows that for every one-year increase in age, the odds of having metabolic syndrome increased by a factor of 1.03.

Model 2: Intercept of the estimated log odds of the outcome when BMI is zero is –5.682010, again significant (p <2e-16). BMI: The coefficient of 0.172420 indicates that an increase in MBI is associated with an increase in the log odds, which is statistically significant.

Model 3: Intercept of the estimated log odds of the outcome when Blood Glucose is zero is –8.858154, again significant (p <2e-16). Blood Glucose: The coefficient of 0.078647, suggesting that an increase in Blood Glucose levels is associated with an increase in the logodds of the outcome with a very strong statistical significance.

Model 4: Intercept of the estimated log odds of the outcome when HDL is zero is 3.126841. This value is significantly different from zero. For every one-unit increase in HDL, the odds of having metabolic syndrome decreased by a factor of 0.93 (OR = 0.9283408, p-value < 2e-16).

Model 5: Intercept of the estimated log odds of the outcome when Uric Acid is zero is 1.45. This value is significantly different from zero. The coefficient of -2.74226 indicates that an increase in Uric Acid is associated with a decrease in the log odds of the outcome. The negative relationship is highly significant.

By conducting logistic regression analysis, we quantified the associations between predictor variables and the likelihood of metabolic syndrome occurrence. The odds ratios obtained provided valuable insights into the strength and direction of these associations, thereby enhancing our understanding of the factors influencing metabolic syndrome prevalence within our dataset.

Limitations: Logistic regression assumes linearity of predictor variables and independence of observations, which may not fully hold in complex datasets. Additionally, multicollinearity among predictor variables could lead to inflated standard errors and biased coefficient estimates, impacting the accuracy of the results.

Multiple Logistic Regression:

The objective of this test is to find which demographics (age, sex, race) or health factor (BMI, BloodGlucose, UricAcid, HDL) has the highest impact on the prevalence of metabolic syndrome among individuals with certain risk factors. The multiple logistic regression analysis shows that demographic and health factors on the prevalence of metabolic syndrome among individuals

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. Logistic regression is well-suited for analyzing binary or categorical outcome variables, making it appropriate for assessing the likelihood of metabolic syndrome occurrence (binary outcome). Logistic regression does not assume the normality of the data, making it suitable for analyzing non-normally distributed data, which is common in healthcare research.

Limitations: The effectiveness and reliability of multiple logistic regression analysis can be influenced by the sample size. Small sample sizes within certain demographic categories may limit the generalizability of the results. High multicollinearity among predictor variables can inflate standard errors and affect the accuracy of coefficient estimates in the regression model.

Results:

For Research Question 1, Investigating the association between demographic and health factors with metabolic syndrome. The chi-square test for categorical variables did not reveal a significant association between sex (gender) and metabolic syndrome. This suggests that gender may not be a significant determinant of metabolic syndrome prevalence in the studied population. However, a significant association was observed between race and metabolic syndrome. This finding implies that race may play a role in predisposing individuals to metabolic syndrome, although further investigation would be necessary to understand the underlying factors contributing to this association. A correlation analysis was conducted to examine the relationships among the continuous variables. The correlation heatmap showed moderate to strong positive correlations between BMI, blood glucose, and uric acid, and a moderate negative correlation between HDL and metabolic syndrome.

Logistic regression models were fitted to quantify the effects of demographic and health factors on the odds of having metabolic syndrome. The effects of demographic and health factors on the odds of having metabolic syndrome. For every one-unit increase in BMI, blood glucose, and uric acid, HDL the odds of having metabolic syndrome increased by factors of 1.19 (OR = 1.188187, p-value < 2e-16), 1.08 (OR = 1.081822, p-value < 2e-16), and 1.09 (OR = 1.091095, p-value = 0.0404), 0.93 (OR = 0.9283408, p-value < 2e-16) respectively. The results indicated that increasing age, BMI, blood glucose, and uric acid levels, HDL were associated with higher odds of having 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 by rejecting Null hypothesis.

For Research Question 2: Identifying the demographic or health factor with the highest impact on the prevalence of metabolic syndrome among individuals with certain risk factors we fitted Multiple Logistic Regression for demographic factors (age, sex, and race) and health factors (BMI, uric acid, blood glucose, and HDL) to determine their impact on the prevalence of

metabolic syndrome. Among the demographic factors, being of Mexican American descent had the highest impact on the prevalence of metabolic syndrome (coefficient estimate = 0.8988951, p-value = 1.37472e-06). Among the health factors, BMI had the highest impact (coefficient estimate = 0.1343017, p-value < 2e-16).

Conclusion:

This study investigated the association between demographic and health factors with the prevalence of metabolic syndrome. The findings revealed that there are significant associations between several demographic and health variables, including age, race, BMI, blood glucose levels, HDL cholesterol, and uric acid, with the presence of metabolic syndrome.

The data deviated from a normal distribution, necessitating the use of appropriate statistical techniques suitable for non-normally distributed data and categorical response variables. Chi-Square and Logistic regression analysis quantified the associations between predictor variables and the likelihood of metabolic syndrome occurrence, providing insights into the strength and direction of these relationships. Age also played a vital role, as the odds of having metabolic syndrome increased significantly with advancing age.

Additionally, the analysis uncovered lower levels of High-Density Lipoprotein (HDL) cholesterol in individuals with metabolic syndrome, suggesting a potential link between dyslipidemia and metabolic syndrome. Among the demographic factors, race emerged as the most influential factor, with individuals of Mexican American descent having the highest likelihood of developing metabolic syndrome compared to other racial groups. This finding highlights the potential impact of genetic and environmental factors specific to certain racial/ethnic backgrounds on the development of metabolic syndrome. Regarding health factors, BMI was identified as the variable with the highest impact on the prevalence of metabolic syndrome. This result aligns with existing evidence that obesity, as measured by BMI, is a significant risk factor for metabolic disorders, including metabolic syndrome.

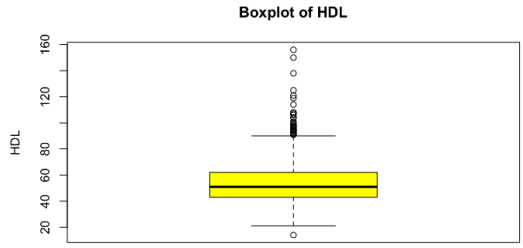
The meticulous data cleaning processes and the use of appropriate statistical methods, such as the chi-square test and logistic regression, ensured the reliability and accuracy of the research findings. The multiple logistic regression analysis provided valuable insights into the demographic and Health factors influencing metabolic syndrome prevalence.

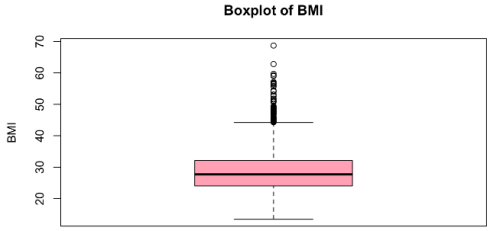
References:

Mayo Clinic. (2021, May 6). *Metabolic syndrome - Symptoms and causes*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/metabolic-syndrome/symptoms-causes/syc-20351916

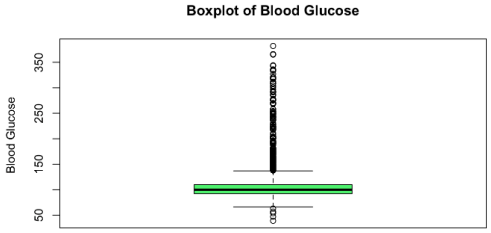
Appendix

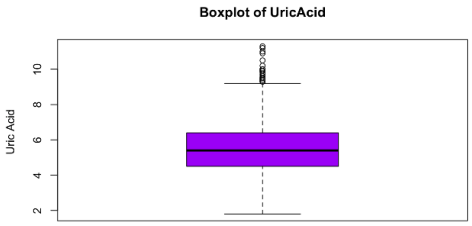
A) Identify the outliers



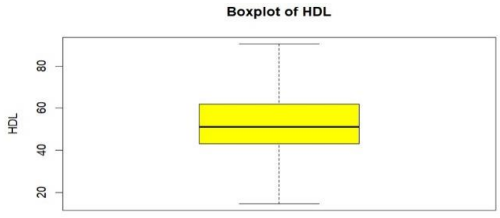


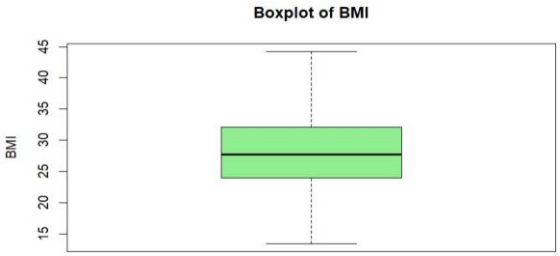
.

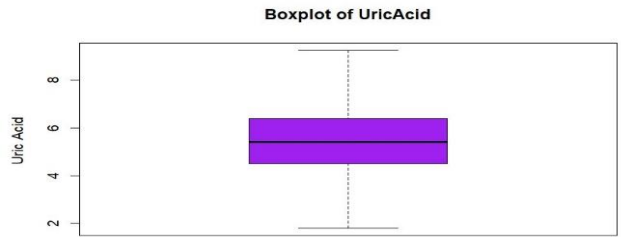


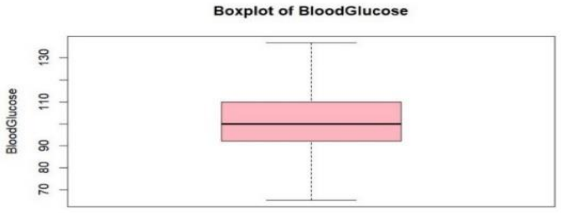


B) After Outlier Treatment









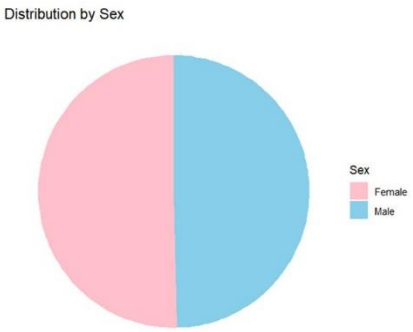
C. Descriptive Statistics:

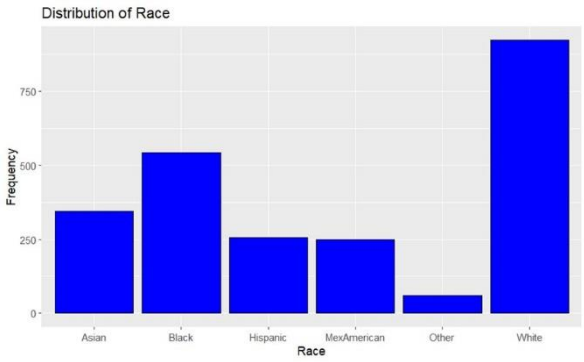
```
#STATISTICAL ANALYSIS
#Exploratory data Analysis

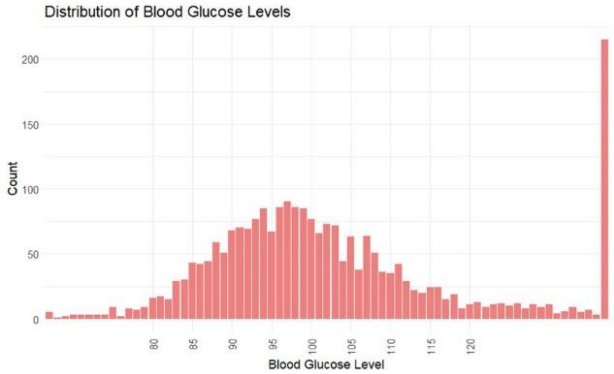
```{r}
summary(Metabolic_Syndrome_final)
```
```

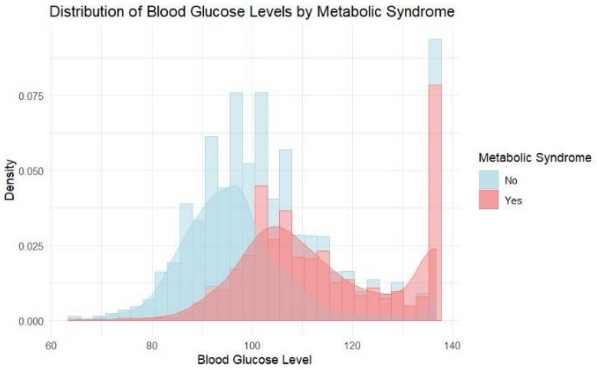
```
      seqn            Age            Sex                Race         BMI          UricAcid      BloodGlucose        HDL        MetabolicSyndrome
 Min.   :62161   Min.   :20.00   Female:1197   Asian      :345   Min.   :13.40   Min.   :1.800   Min.   : 65.0   Min.   :14.50   0:1558
 1st Qu.:64563   1st Qu.:34.00   Male   :1178   Black      :542   1st Qu.:24.00   1st Qu.:4.500   1st Qu.: 92.0   1st Qu.:43.00   1: 817
 Median :67058   Median :48.00                 Hispanic   :255   Median :27.70   Median :5.400   Median :100.0   Median :51.00
 Mean   :67028   Mean   :48.67                 MexAmerican:249   Mean   :28.55   Mean   :5.474   Mean   :102.9   Mean   :53.12
 3rd Qu.:69501   3rd Qu.:63.00                 Other      : 61   3rd Qu.:32.10   3rd Qu.:6.400   3rd Qu.:110.0   3rd Qu.:62.00
 Max.   :71915   Max.   :80.00                 White      :923   Max.   :44.25   Max.   :9.250   Max.   :137.0   Max.   :90.50
```

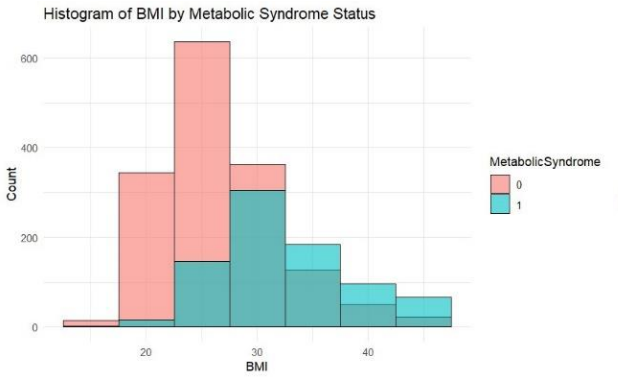
C. Data Visualizations

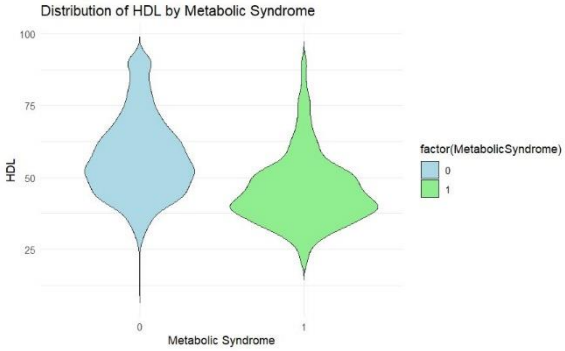


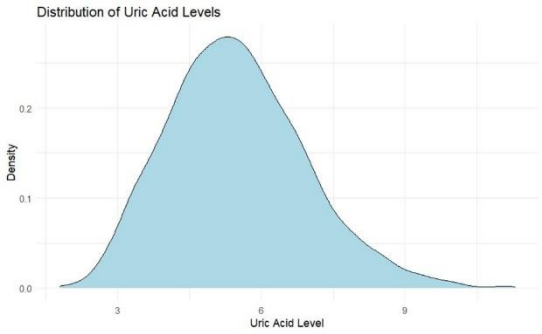


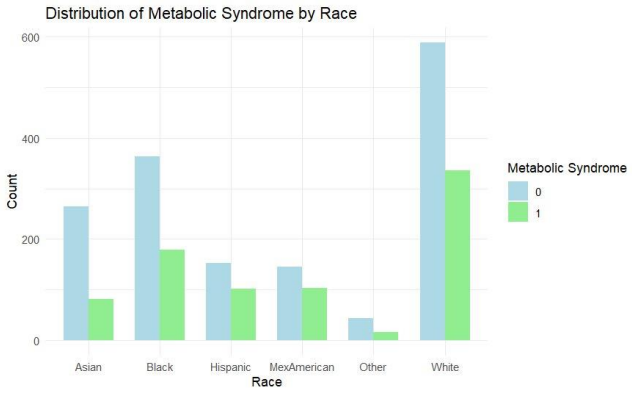


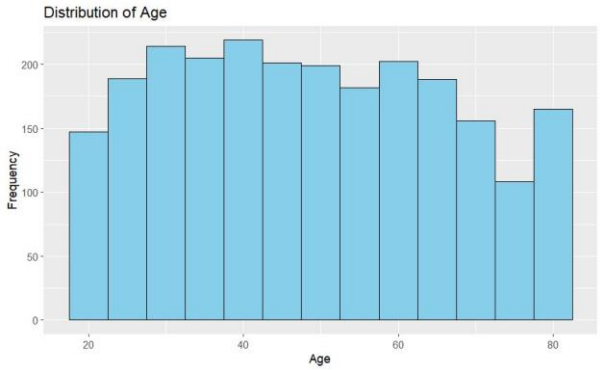




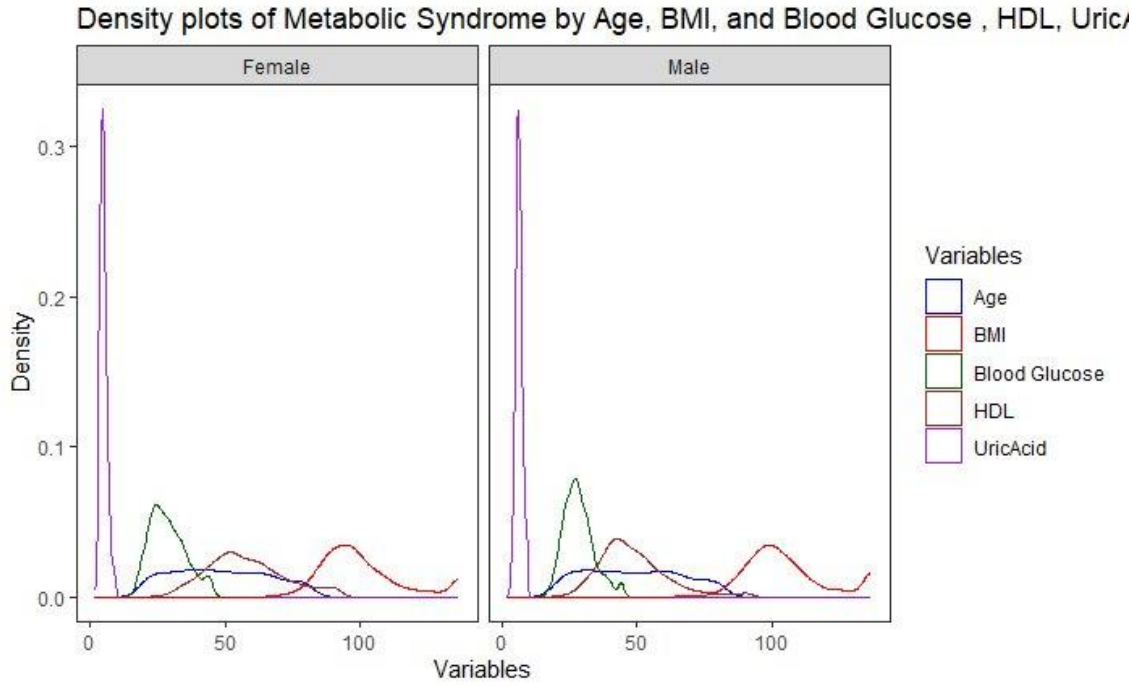




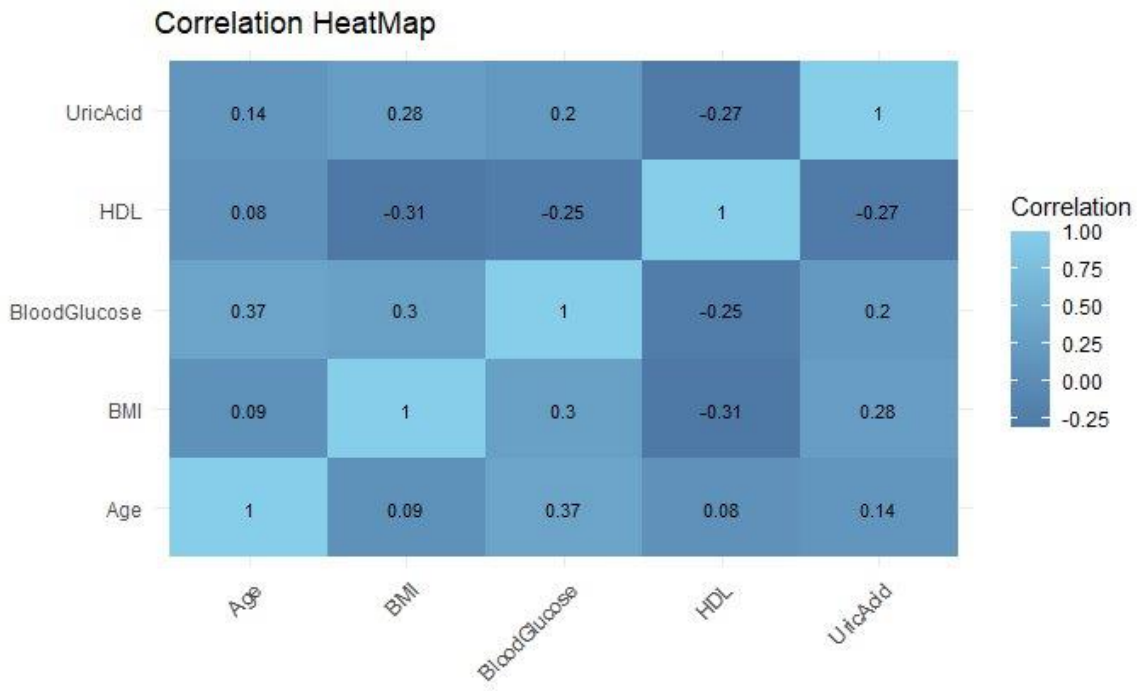




D. Checking Normality



E. Correlation Analysis



F. Simple Logistic Regression

Model 1

```
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.
```

Model 2

```
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.
```

Model 3

```
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.
```

Model 4

```
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.
```

Model 5

```
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.
```