COMPUTATIONAL STATISTICS AND PROBABILITY

(AIM-5002-1)

“Final Report and Analysis Plan”

NAME: Ravisankar Chengannagari
BANNER ID: 800765235

SUBMITTED TO
Jochen G. Raimann sir

ABSTRACT

Obesity is a worldwide public health concern, posing risks to individual health and burdening healthcare systems. This study investigates the relationship between creatinine and obesity across different age groups using data from the National Health and Nutrition Examination Survey (NHANES). The primary indicator of obesity is Body Mass Index (BMI), which creatinine offers insights into muscle metabolism. By knowing and analyzing these demographics, health related, lifestyle factors, we aim to understand how creatinine may be associated with obesity. Our analysis suggest that higher creatinine may correlate with low prevalence of obesity.

INTRODUCTION

Obesity is a widespread health issue affecting peoples worldwide. It is characterized by excessive body fat accumulation. It causes significant risks to the people overall health. It is also connected with various chronic conditions like type two diabetes, heart disease, and certain cancers. Understanding the factors cause to obesity is helpful for developing required prevention and intervention techniques.

BMI (Body Mass Index) works as an indicator of obesity. it is calculated with the help of person's height and weight. We can say if a person's BMI is less than 30 then he/she is classified as normal and if a person's BMI is greater than or equal to 30 then he/she classified as obese. There are so many factors also play roles in obesity, genetic and environmental factors, lifestyle factors like diet, smoking habits, physical activity also changes weight status.

A waste product called creatinine which is produced by muscle metabolism, it is filtered out of the blood by the kidneys. Hight levels of creatinine may specify impaired kidney function. There is some kind of potential link between obesity and creatinine levels because in weight regulation the muscles metabolism plays an important role. By finding this link we can get some insights into the underlying mechanisms of obesity.

In this study, we aim to explore the relationship between creatinine levels and obesity using from National Health and Nutrition Examination Survey (NHANES). We seek to understand how the creatinine may be associated with obesity by looking into lifestyle demographic factors, smoking, and health related factors. By performing statistical analyses and hypothesis testing, we try to uncover the potential links between creatinine and obesity.

Additionally, we propose to conduct subgroup analyses across different age groups to assess if the association between creatinine levels and obesity persists across different age demographics. By exploring this age patterns, we aim to give insights into the relationship between creatinine levels and obesity across the lifespan.

METHODS

DATA SOURCE:

We got the required data from the National Health and Nutrition Examination Survey (NHANES) dataset. It is a survey data conducted by Centers for Disease Control and Prevention (CDC) which contains demographic information, health related information and physiological measurements of people who participated in the survey.

DATA CLEANING AND PREPARATION:

We will focus on the required attributes like age, gender, race, ethnicity, creatinine levels, Body Mass Index, physical activity, smoking information, and diabetic information. Initially, there is no attribute called creatinine in NHANES dataset. But we can create it with the help of Urine volume and flow. We also need to handle the missing values by imputation or exclusion by knowing the nature of the attributes and extend of missingness. There are so many missing values of smoking information, we can't be able to remove it because it is one of the main attributes in our study. So, we used mice library to prevent eliminating so many rows.

DESCRIPTIVE ANALYSIS:

We will calculate descriptive analysis to summarize the characteristics of study population including the distribution of creatinine levels and BMI, and prevalence of obesity. We will apply exploratory data analysis techniques like box plots, histograms, scatterplots, etc. to visualize the distribution of creatinine levels and BMI across different subgroups.

BIVARIATE ANALYSIS:

We will explore the relationship between BMI and creatinine levels using correlation analysis and scatter plots. Additionally, we will investigate potential association between creatinine levels, BMI, and other demographic and lifestyle factors through bivariate analyses such as t-tests.

MULTIVARIATE REGRESSION ANALYSIS:

We will construct multivariate regression models to assess the relationship between creatinine levels and BMI while adjusting for confounding factors, like age, gender, physical activity, ethnicity, and diabetes status. Regression coefficients and their significance levels were examined to identify significant predictors of creatinine levels.

STRATIFIED ANALYSIS:

We will conduct stratified analyses to explore variations in the relationship between creatinine levels and BMI across different subgroups defined by demographic and lifestyle characteristics. Specific regression models were fitted to assess the relationship between creatinine levels and BMI within each subgroup.

STATISTICAL ANALYSIS:

We will perform statistical analysis including t-tests, regression analyses using appropriate statistical packages. The significance level was set at $p < 0.05$ to find out statistical significance,

RESULTS

DESCRIPTIVE ANALYSIS:

We performed descriptive analysis included a total of 10,000 participants from the NHANES dataset, with a mean age of 36.74 years. The distribution of BMI revealed that approximately 27.66% of the population had a BMI greater than or equal to 30 which indicating obesity, while 30.22% had a BMI between 25 and 30 which indicating overweight.
The mean of creatine levels of participants found to be 0.1668. Based on the 75th percentile threshold, approximately 25% of participants were classified as having high creatinine levels.

BIVARIATE ANALYSIS:

The bivariate analysis revealed a weak negative correlation between creatinine levels and BMI ($r = -0.0025$, $p > 0.05$). It is suggesting that there is no linear relationship between BMI and creatinine variables. Scatter plots also supported these results and showed no clear pattern of relationship between BMI and creatinine levels.

MULTIVARIATE REGRESSION ANALYSIS:

The logistic regression model examining the relationship between creatinine levels and obesity adjusted for age, race, and gender, revealed several significant findings. There was no significant relationship between creatinine levels and odds of obesity (beta = -0.1108, $p = 0.206$). However, age remained a significant predictor of obesity with older individuals having higher odds of obesity (beta = 0.0264, $p < 0.001$).

STRATIFIED ANALYSIS:

We conducted stratified analysis on gender subgroups revealed variations in the relationship between creatinine levels and BMI. In the male subgroup, the regression model indicated a positive relationship between creatinine levels and BMI, with coefficient of 0.001416 which is less than 0.05 (p-value). In the female subgroup, we can observe a negative association with a coefficient of -0.001114. These relationships being statistically significant, they are both very weak and suggesting minimal impact of BMI on creatinine levels within each subgroup of gender.

Additionally, we conducted stratified analysis to find variations in the relationship between creatinine levels and obesity across different age groups. We performed two-sample test to compare creatinine levels between obese and non-obese individuals within each age group. For Middle-Aged Adults, got a significant difference ($t = -2.5095$, df = 3450.4, p-value = 0.01213) and mean creatinine level was 0.179872 for obese individuals and 0.203941 for non-obese individuals, with a 95% confidence interval of [-0.0442873598, -0.005264295]. For Young Adults, got a significant difference in creatinine levels between obese and non-obese individuals ($t = -0.97326$, df = 1046.9, p-value = 0.3306) and mean creatinine level was 0.1510731 for obese and 0.1606936 for non-obese individuals, with a 95% confidence interval of [-0.029016867, 0.009775837]. For Older Adults, revealed a significant difference in creatinine levels between obese and non-obese ($t = -2.6689$, df = 1350.2, p-value = 0.007702) and mean creatinine level was 0.1023153 for obese and 0.1305413 for non-obese, with a 95% confidence interval of [-0.048973248, -0.007478655].

DISCUSSION

In our study, we dig into the relationship between obesity and creatinine levels and how this relationship varies across different ethnicities and age groups. Our results offer valuable insights into the complex interplay between these factors and their implications for efforts of public health.

One of the findings is the relationship between obesity and creatinine levels. Creatinine is a waste product produced by muscle metabolism and has been linked to obesity. Our study confirms this relationship and showing that persons with higher creatinine levels tend to have lower prevalence of obesity.

Our analysis shows patterns in age-specific in the relationship between obesity and creatinine levels and found that this relationship persists across different age groups with variations in strength. For example, in older adults and middle-aged adults, higher creatinine levels were consistently related with lower prevalence of obesity. However, in young adults, this association was less pronounced which indicates differences in underlying mechanisms driving obesity across age groups.

We also observed differences in the relationship between creatinine levels and obesity across different ethnic groups. Some ethnicities showed stronger relationships between creatinine levels and obesity, others exhibited weaker relationships.

CONCLUSION

In this study, our analysis of the relationship between creatinine levels and obesity using the data of National Health and Nutrition Examination Survey (NHANES) yielded many important insights.

We observed a negative relationship between creatinine levels and obesity across the entire study population. Some people with higher creatinine levels tend to have lower BMI values which indicating a potential protective effect against obesity.

We revealed variations in stratified variations in the relationship between obese and creatinine levels across different age groups. While the negative association persisted in most age strata, some nuances were observed. Older Adults and Middle-aged adults have the strongest inverse relationship between obesity and creatinine levels. It is showing the importance of age-specific considerations in obesity research and interventions.

We uncovered gender and ethnicity differences in the relationship between creatinine levels and obesity. Gender did not influence the association, certain ethnic groups showed varying degrees of association. It suggests potential ethnic disparities in obesity risk factors.

These results have important implications for public health strategies aimed at combating the obesity epidemic. Finding creatinine levels as a potential biomarker for obesity risk, healthcare professionals can better target interventions and resources towards people having high risk, mainly in middle aged and older aged people.

In conclusion, our study contributes to a good understanding of the relationship between obesity and creatinine levels, highlighting age-specific patterns and ethnic disparities.

REFERENCES

World Health Organization (WHO) --- Obesity and Overweight --- https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

National Health and Nutrition Examination Survey (NHANES) --- https://www.cdc.gov/nchs/nhanes/index.htm

World Health Organization (WHO) --- Physical Activity and Adults --- https://www.who.int/news-room/fact-sheets/detail/physical-activity

CODE AND RESULTS

```
#remove all the objects currently stored in the environment
rm(list = ls())

#loads NHANES package
library(NHANES)

data(NHANES)

#creating object with NHANES data
nhanes_data <- NHANES::NHANES

#can see the few rows of NHANES data
head(nhanes_data)
```

```
## # A tibble: 6 × 76

##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
MaritalStatus HHIncome

##   <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
<fct>         <fct>

## 1 51624 2009_10  male      34 " 30-39"        409 White <NA>  High School
Married       25000-349…

## 2 51624 2009_10  male      34 " 30-39"        409 White <NA>  High School
Married       25000-349…

## 3 51624 2009_10  male      34 " 30-39"        409 White <NA>  High School
Married       25000-349…

## 4 51625 2009_10  male       4 " 0-9"           49 Other <NA>  <NA>
<NA>          20000-249…

## 5 51630 2009_10  female    49 " 40-49"        596 White <NA>  Some College
LivePartner   35000-449…

## 6 51638 2009_10  male       9 " 0-9"          115 White <NA>  <NA>
<NA>          75000-999…

## # ℹ 65 more variables: HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>,
HomeOwn <fct>,

## #   Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>, Height <dbl>,
BMI <dbl>,

## #   BMICatUnder20yrs <fct>, BMI_WHO <fct>, Pulse <int>, BPSysAve <int>, BP
DiaAve <int>,

## #   BPSys1 <int>, BPDia1 <int>, BPSys2 <int>, BPDia2 <int>, BPSys3 <int>,
BPDia3 <int>,
```

```
## #   Testosterone <dbl>, DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,
## #   UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>, HealthGen <fct>,
## #   DaysPhysHlthBad <int>, DaysMentHlthBad <int>, LittleInterest <fct>, Depressed <fct>, …
```

#each variable summary statistics
summary(nhanes_data)

```
##        ID           SurveyYr       Gender         Age          AgeDecade    AgeMonths
##  Min.   :51624   2009_10:5000   female:5020   Min.   : 0.00    40-49 :1398   Min.   :  0.0
##  1st Qu.:56904   2011_12:5000   male  :4980   1st Qu.:17.00    0-9   :1391   1st Qu.:199.0
##  Median :62160                                Median :36.00    10-19 :1374   Median :418.0
##  Mean   :61945                                Mean   :36.74    20-29 :1356   Mean   :420.1
##  3rd Qu.:67039                                3rd Qu.:54.00    30-39 :1338   3rd Qu.:624.0
##  Max.   :71915                                Max.   :80.00   (Other):2810   Max.   :959.0
##                                                                NA's   : 333   NA's   :5038
##       Race1           Race3               Education         MaritalStatus        HHIncome
##  Black   :1197   Asian   : 288   8th Grade     : 451   Divorced    : 707   more 99999 :2220
##  Hispanic: 610   Black   : 589   9 - 11th Grade: 888   LivePartner : 560   75000-99999:1084
##  Mexican :1015   Hispanic: 350   High School   :1517   Married     :3945   25000-34999: 958
##  White   :6372   Mexican : 480   Some College  :2267   NeverMarried:1380   35000-44999: 863
##  Other   : 806   White   :3135   College Grad  :2098   Separated   : 183   45000-54999: 784
##                  Other   : 158   NA's          :2779   Widowed     : 456   (Other)    :3280
##                  NA's    :5000                         NA's        :2769   NA's       : 811
```

```
##   HHIncomeMid        Poverty        HomeRooms       HomeOwn              Wo
rk          Weight
##  Min.   :  2500   Min.   :0.000   Min.   : 1.000   Own  :6425   Looking
: 311   Min.   :  2.80
##  1st Qu.: 30000   1st Qu.:1.240   1st Qu.: 5.000   Rent :3287   NotWorking
:2847   1st Qu.: 56.10
##  Median : 50000   Median :2.700   Median : 6.000   Other: 225   Working
:4613   Median : 72.70
##  Mean   : 57206   Mean   :2.802   Mean   : 6.249   NA's :  63   NA's
:2229   Mean   : 70.98
##  3rd Qu.: 87500   3rd Qu.:4.710   3rd Qu.: 8.000
3rd Qu.: 88.90
##  Max.   :100000   Max.   :5.000   Max.   :13.000
Max.   :230.70
##  NA's   :811      NA's   :726     NA's   :69
NA's   :78
##      Length          HeadCirc         Height           BMI            BMICa
tUnder20yrs
##  Min.   : 47.10   Min.   :34.20   Min.   : 83.6   Min.   :12.88   UnderWei
ght:  55
##  1st Qu.: 75.70   1st Qu.:39.58   1st Qu.:156.8   1st Qu.:21.58   NormWeig
ht : 805
##  Median : 87.00   Median :41.45   Median :166.0   Median :25.98   OverWeig
ht : 193
##  Mean   : 85.02   Mean   :41.18   Mean   :161.9   Mean   :26.66   Obese
: 221
##  3rd Qu.: 96.10   3rd Qu.:42.92   3rd Qu.:174.5   3rd Qu.:30.89   NA's
:8726
##  Max.   :112.20   Max.   :45.40   Max.   :200.4   Max.   :81.25
##  NA's   :9457     NA's   :9912    NA's   :353     NA's   :366
##          BMI_WHO          Pulse           BPSysAve        BPDiaAve
BPSys1
##  12.0_18.5   :1277   Min.   : 40.00   Min.   : 76.0   Min.   :  0.00   Min
.   : 72.0
##  18.5_to_24.9:2911   1st Qu.: 64.00   1st Qu.:106.0   1st Qu.: 61.00   1st
Qu.:106.0
##  25.0_to_29.9:2664   Median : 72.00   Median :116.0   Median : 69.00   Med
ian :116.0
##  30.0_plus   :2751   Mean   : 73.56   Mean   :118.2   Mean   : 67.48   Mea
n   :119.1
##  NA's        : 397   3rd Qu.: 82.00   3rd Qu.:127.0   3rd Qu.: 76.00   3rd
Qu.:128.0
```

```
##                      Max.   :136.00   Max.   :226.0   Max.   :116.00   Max
.   :232.0
##                      NA's   :1437     NA's   :1449    NA's   :1449     NA'
s   :1763
##      BPDia1           BPSys2          BPDia2           BPSys3          BPD
ia3
##  Min.   :  0.00   Min.   : 76.0   Min.   :  0.00   Min.   : 76.0   Min.
:  0.0
##  1st Qu.: 62.00   1st Qu.:106.0   1st Qu.: 60.00   1st Qu.:106.0   1st Qu.
: 60.0
##  Median : 70.00   Median :116.0   Median : 68.00   Median :116.0   Median
: 68.0
##  Mean   : 68.28   Mean   :118.5   Mean   : 67.66   Mean   :117.9   Mean
: 67.3
##  3rd Qu.: 76.00   3rd Qu.:128.0   3rd Qu.: 76.00   3rd Qu.:126.0   3rd Qu.
: 76.0
##  Max.   :118.00   Max.   :226.0   Max.   :118.00   Max.   :226.0   Max.
:116.0
##  NA's   :1763     NA's   :1647    NA's   :1647     NA's   :1635    NA's
:1635
##   Testosterone       DirectChol       TotChol         UrineVol1       Urin
eFlow1
##  Min.   :   0.25   Min.   :0.390   Min.   : 1.530   Min.   :  0.0   Min.
: 0.0000
##  1st Qu.:  17.70   1st Qu.:1.090   1st Qu.: 4.110   1st Qu.: 50.0   1st Qu
.: 0.4030
##  Median :  43.82   Median :1.290   Median : 4.780   Median : 94.0   Median
: 0.6990
##  Mean   : 197.90   Mean   :1.365   Mean   : 4.879   Mean   :118.5   Mean
: 0.9793
##  3rd Qu.: 362.41   3rd Qu.:1.580   3rd Qu.: 5.530   3rd Qu.:164.0   3rd Qu
.: 1.2210
##  Max.   :1795.60   Max.   :4.030   Max.   :13.650   Max.   :510.0   Max.
:17.1670
##  NA's   :5874      NA's   :1526    NA's   :1526     NA's   :987     NA's
:1603
##    UrineVol2       UrineFlow2     Diabetes     DiabetesAge        HealthGe
n    DaysPhysHlthBad
##  Min.   :  0.0   Min.   : 0.000   No  :9098   Min.   : 1.00   Excellent: 8
78   Min.   : 0.000
##  1st Qu.: 52.0   1st Qu.: 0.475   Yes : 760   1st Qu.:40.00   Vgood    :25
08   1st Qu.: 0.000
```

```
##  Median : 95.0   Median : 0.760   NA's: 142   Median :50.00   Good     :29
56   Median : 0.000
##  Mean   :119.7   Mean   : 1.149               Mean   :48.42   Fair     :10
10   Mean   : 3.335
##  3rd Qu.:171.8   3rd Qu.: 1.513               3rd Qu.:58.00   Poor     : 1
87   3rd Qu.: 3.000
##  Max.   :409.0   Max.   :13.692               Max.   :80.00   NA's     :24
61   Max.   :30.000
##  NA's   :8522    NA's   :8524                 NA's   :9371
NA's   :2468
##  DaysMentHlthBad  LittleInterest   Depressed     nPregnancies       nBabie
s         Age1stBaby
##  Min.   : 0.000   None   :5103   None   :5246   Min.   : 1.000   Min.   :
0.000   Min.   :14.00
##  1st Qu.: 0.000   Several:1130   Several:1009   1st Qu.: 2.000   1st Qu.:
2.000   1st Qu.:19.00
##  Median : 0.000   Most   : 434   Most   : 418   Median : 3.000   Median :
2.000   Median :22.00
##  Mean   : 4.127   NA's   :3333   NA's   :3327   Mean   : 3.027   Mean   :
2.457   Mean   :22.65
##  3rd Qu.: 4.000                                 3rd Qu.: 4.000   3rd Qu.:
3.000   3rd Qu.:26.00
##  Max.   :30.000                                 Max.   :32.000   Max.   :1
2.000   Max.   :39.00
##  NA's   :2466                                   NA's   :7396     NA's   :7
584     NA's   :8116
##  SleepHrsNight    SleepTrouble PhysActive  PhysActiveDays       TVHrsDay
CompHrsDay
##  Min.   : 2.000   No  :5799    No  :3677   Min.   :1.000   2_hr     :1275
0_to_1_hr:1409
##  1st Qu.: 6.000   Yes :1973    Yes :4649   1st Qu.:2.000   1_hr     : 884
0_hrs    :1073
##  Median : 7.000   NA's:2228    NA's:1674   Median :3.000   3_hr     : 836
1_hr     :1030
##  Mean   : 6.928                            Mean   :3.744   0_to_1_hr: 638
2_hr     : 589
##  3rd Qu.: 8.000                            3rd Qu.:5.000   More_4_hr: 615
3_hr     : 347
##  Max.   :12.000                            Max.   :7.000   (Other)  : 611
(Other)  : 415
##  NA's   :2245                              NA's   :5337    NA's     :5141
NA's     :5137
```

```
##  TVHrsDayChild   CompHrsDayChild Alcohol12PlusYr   AlcoholDay      Alcohol
Year    SmokeNow
##  Min.   :0.000   Min.   :0.000   No  :1368       Min.   : 1.000   Min.   :
0.0   No  :1745
##  1st Qu.:1.000   1st Qu.:0.000   Yes :5212       1st Qu.: 1.000   1st Qu.:
3.0   Yes :1466
##  Median :2.000   Median :1.000   NA's:3420       Median : 2.000   Median :
24.0   NA's:6789
##  Mean   :1.939   Mean   :2.198                   Mean   : 2.914   Mean   :
75.1
##  3rd Qu.:3.000   3rd Qu.:6.000                   3rd Qu.: 3.000   3rd Qu.:
104.0
##  Max.   :6.000   Max.   :6.000                   Max.   :82.000   Max.   :
364.0
##  NA's   :9347    NA's   :9347                    NA's   :5086     NA's   :
4078
##  Smoke100         Smoke100n       SmokeAge     Marijuana   AgeFirstMarij
RegularMarij
##  No  :4024   Non-Smoker:4024   Min.   : 6.00   No  :2049   Min.   : 1.00
No  :3575
##  Yes :3211   Smoker    :3211   1st Qu.:15.00   Yes :2892   1st Qu.:15.00
Yes :1366
##  NA's:2765   NA's      :2765   Median :17.00   NA's:5059   Median :16.00
NA's:5059
##                                Mean   :17.83               Mean   :17.02
##                                3rd Qu.:19.00               3rd Qu.:19.00
##                                Max.   :72.00               Max.   :48.00
##                                NA's   :6920                NA's   :7109
##   AgeRegMarij    HardDrugs   SexEver         SexAge      SexNumPartnLife
SexNumPartYear
##  Min.   : 5.00   No  :4700   No  : 223   Min.   : 9.00   Min.   :   0.00
Min.   : 0.000
##  1st Qu.:15.00   Yes :1065   Yes :5544   1st Qu.:15.00   1st Qu.:   2.00
1st Qu.: 1.000
##  Median :17.00   NA's:4235   NA's:4233   Median :17.00   Median :   5.00
Median : 1.000
##  Mean   :17.69                           Mean   :17.43   Mean   :  15.09
Mean   : 1.342
##  3rd Qu.:19.00                           3rd Qu.:19.00   3rd Qu.:  12.00
3rd Qu.: 1.000
##  Max.   :52.00                           Max.   :50.00   Max.   :2000.00
Max.   :69.000
```

```
##  NA's   :8634                            NA's   :4460    NA's   :4275
NA's   :5072
##  SameSex          SexOrientation  PregnantNow
##  No  :5353   Bisexual    : 119   Yes     :  72
##  Yes : 415   Heterosexual:4638   No      :1573
##  NA's:4232   Homosexual  :  85   Unknown:  51
##              NA's        :5158   NA's   :8304
##
##
##
```

#creating a vector with required columns for our project
attributes <- c("ID", "Gender", "Age", "Race1", "BMI", "PhysActive", "Diabetes", "UrineVol1", "UrineFlow1", "UrineVol2", "UrineFlow2", "SmokeNow")

#creates new dataframe having only the required attributes
data <- nhanes_data[, names(nhanes_data) %in% attributes]

#shows total missing values of each column
colSums(is.na(data))

```
##         ID     Gender        Age      Race1        BMI  UrineVol1 UrineFlo
w1  UrineVol2 UrineFlow2
##          0          0          0          0        366        987       16
03       8522       8524
##   Diabetes PhysActive   SmokeNow
##        142       1674       6789
```

#finding mean of Body Mass Index
bmi_mean <- mean(data$BMI, na.rm = TRUE)
bmi_mean

```
## [1] 26.66014
```

```
#assign mean of BMI inplace of null values
data$BMI[is.na(data$BMI)] <- bmi_mean

#removing the UrineVol2 and UrineFlow2 columns because of having more null values
data <- subset(data, select = -c(UrineVol2, UrineFlow2))

#loads mice package
library(mice)

#set seed for reproducibility
set.seed(123)

#creates multiple imputations using Predictive Mean Matching
impute <- mice(data, method = 'pmm', m = 5)
```

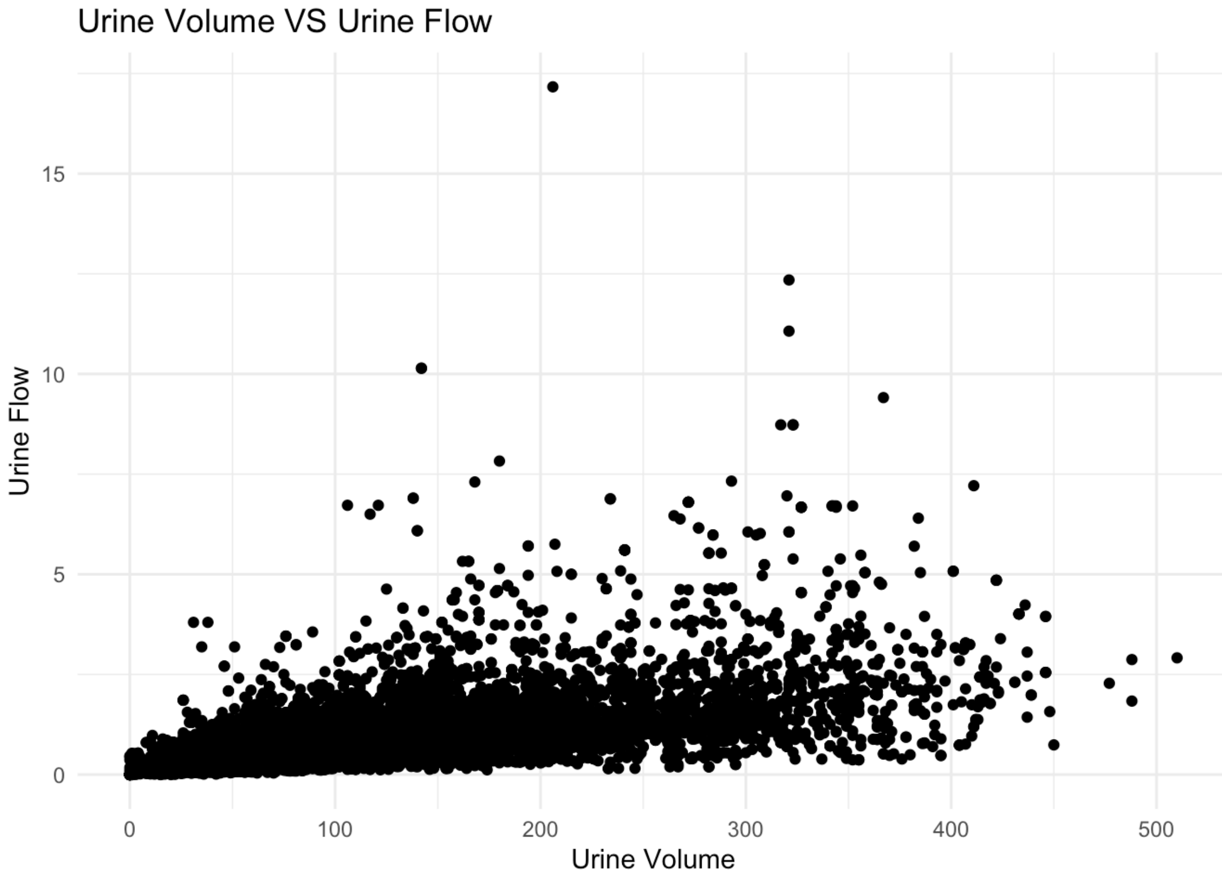
```
##
##  iter imp variable
##   1   1  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   1   2  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   1   3  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   1   4  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   1   5  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   2   1  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   2   2  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   2   3  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   2   4  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   2   5  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   3   1  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   3   2  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   3   3  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   3   4  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   3   5  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   4   1  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   4   2  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   4   3  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   4   4  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
```

```
##   4   5  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   5   1  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   5   2  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   5   3  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   5   4  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
##   5   5  UrineVol1  UrineFlow1  Diabetes  PhysActive  SmokeNow
```

```
#replaces missing values
data <- complete(impute)

#shows total missing values of each column
colSums(is.na(data))
```

```
##          ID     Gender        Age      Race1        BMI  UrineVol1 UrineFlo
w1   Diabetes PhysActive
##           0          0          0          0          0          0
0          0          0
##   SmokeNow
##           0
```

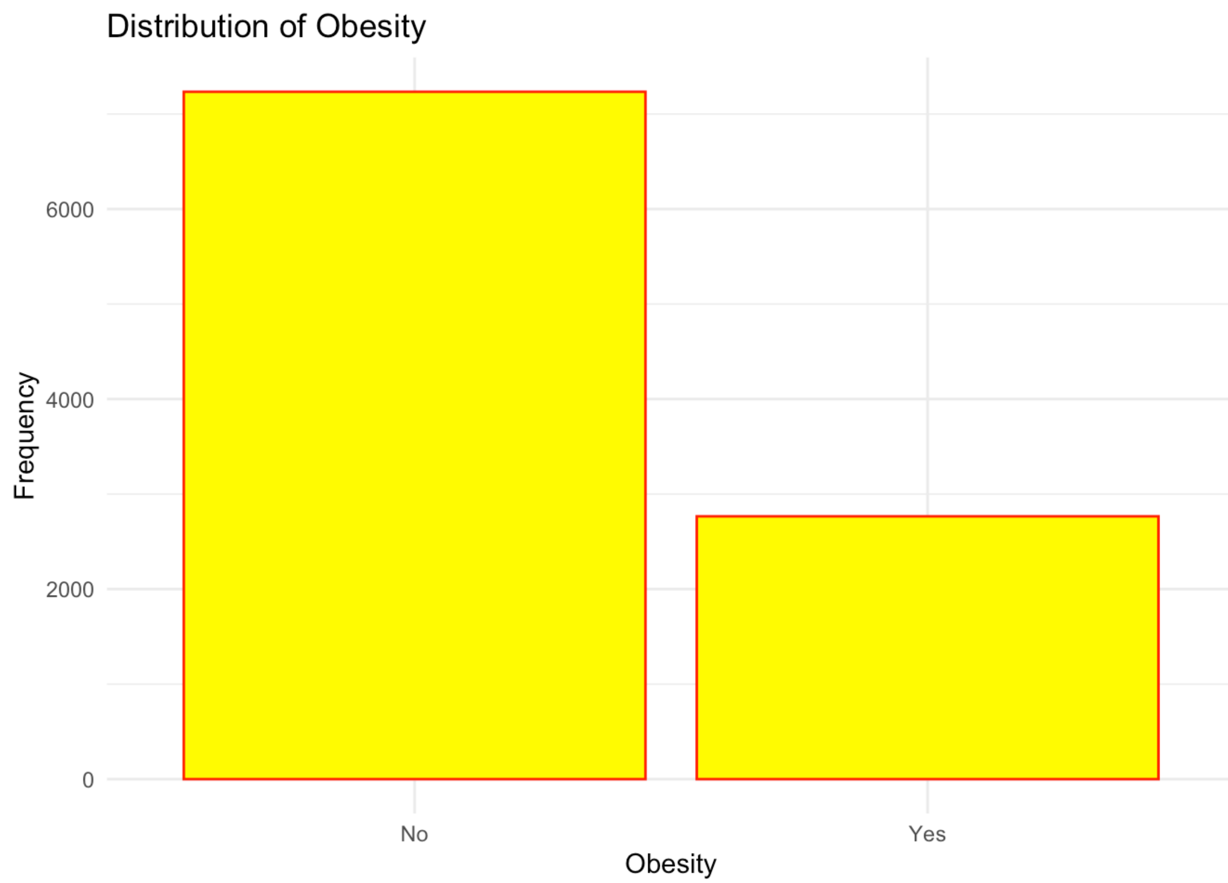
```
#removes rows having null values
data <- na.omit(data)

#loads ggplot2 package
library(ggplot2)

#plots scatter plot of UrineVol1 and UrineFlow1
ggplot(data, aes(x = UrineVol1, y = UrineFlow1)) +
  geom_point() +
  labs(title = "Urine Volume VS Urine Flow",
       x = "Urine Volume",
       y = "Urine Flow") +
  theme_minimal()
```



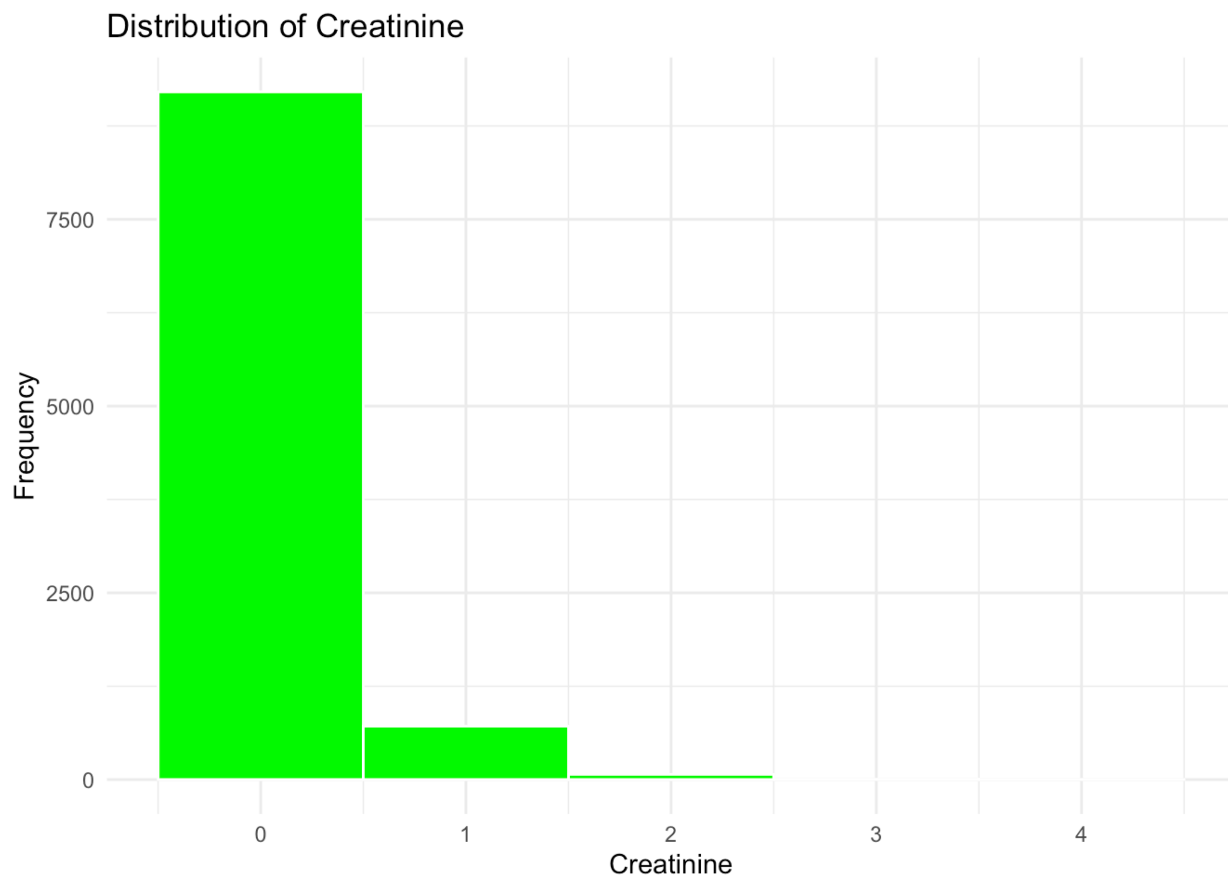

```
#function returns creatinine value
cal_creatinine <- function(vol, flow) {
  return(vol * flow / 1000)
}

#creating column with creatinine values
data$Creatinine <- cal_creatinine(data$UrineVol1, data$UrineFlow1)

#creating Obesity values
data <- data %>%
  mutate(Obesity = ifelse(BMI >= 30, "Yes", "No"))

#each variable summary statistics of required attributes data
summary(data)
```

```
##        ID           Gender          Age              Race1          BMI        
UrineVol1    
##  Min.   :51624   female:5020   Min.   : 0.00   Black   :1197   Min.   :12.
88   Min.   :  0.0  
##  1st Qu.:56904   male  :4980   1st Qu.:17.00   Hispanic: 610   1st Qu.:21.
80   1st Qu.: 50.0  
##  Median :62160                 Median :36.00   Mexican :1015   Median :26.
30   Median : 94.0  
##  Mean   :61945                 Mean   :36.74   White   :6372   Mean   :26.
66   Mean   :119.2  
##  3rd Qu.:67039                 3rd Qu.:54.00   Other   : 806   3rd Qu.:30.
60   3rd Qu.:166.0  
##  Max.   :71915                 Max.   :80.00                   Max.   :81.
25   Max.   :510.0  
##    UrineFlow1      Diabetes   PhysActive SmokeNow     Creatinine        Ob
esity          
##  Min.   : 0.0000   No :9238   No :4021   No :4415   Min.   :0.00000   Leng
th:10000         
##  1st Qu.: 0.4000   Yes: 762   Yes:5979   Yes:5585   1st Qu.:0.02094   Clas
s :character  
##  Median : 0.6935                                    Median :0.06422   Mode  
:character  
##  Mean   : 0.9740                                    Mean   :0.16681
##  3rd Qu.: 1.2250                                    3rd Qu.:0.19136
##  Max.   :17.1670                                    Max.   :3.96307
```

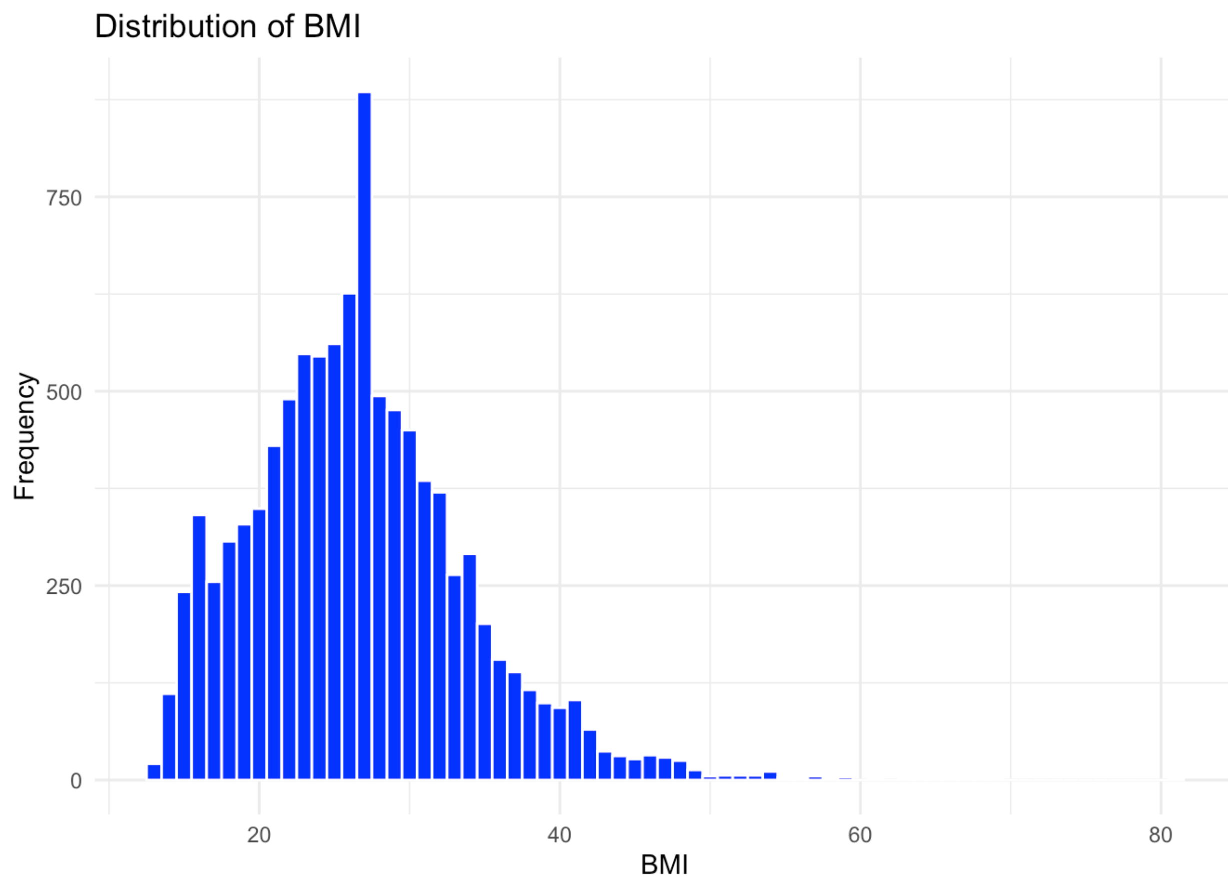
```
#plots distribution of Obesity
ggplot(data, aes(x = Obesity)) +
  geom_bar(fill = "yellow", color = "red") +
  labs(title = "Distribution of Obesity",
       x = "Obesity",
       y = "Frequency") +
  theme_minimal()
```



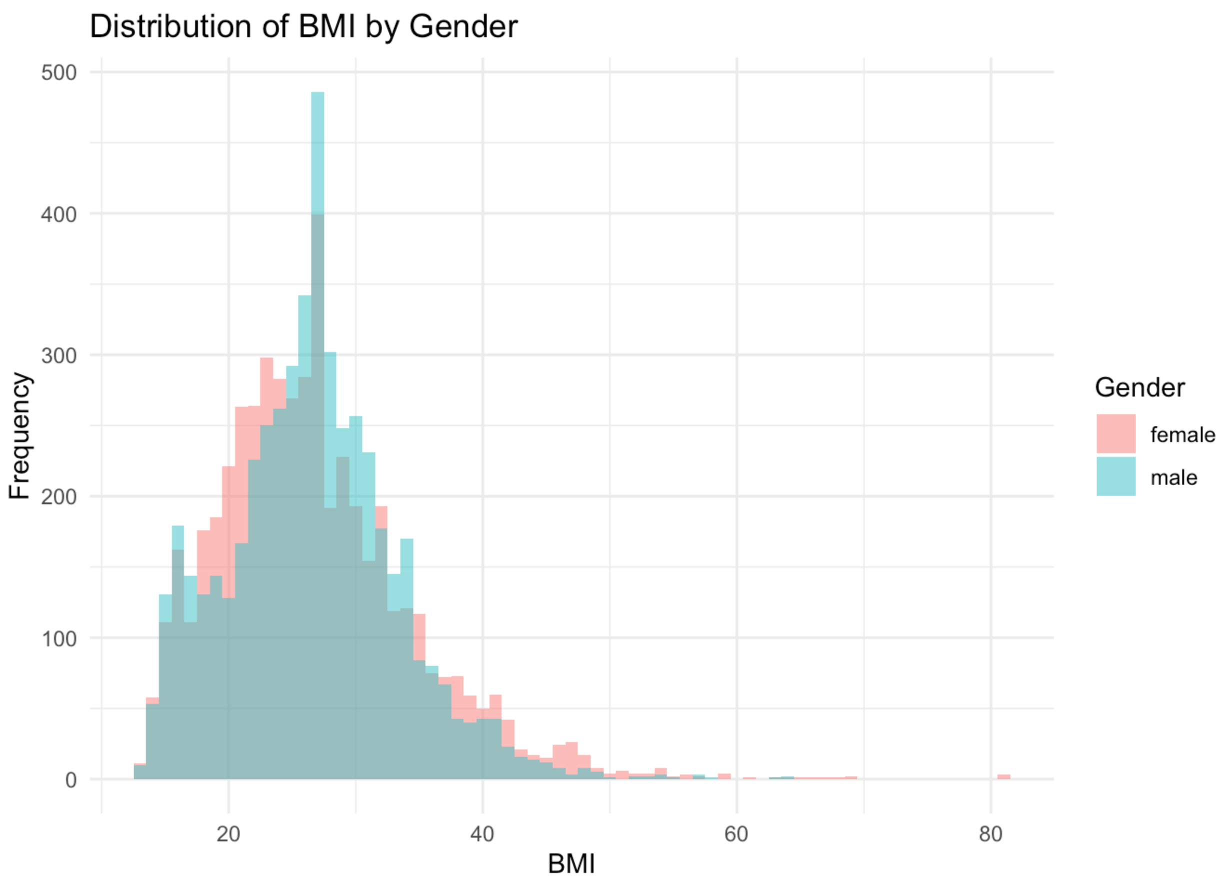

```
#shows total missing values of each column
colSums(is.na(data))

#plots distribution of creatinine
ggplot(data, aes(x = Creatinine)) +
  geom_histogram(binwidth = 1, fill = "green", color = "white") +
  labs(title = "Distribution of Creatinine",
       x = "Creatinine",
       y = "Frequency") +
  theme_minimal()
```



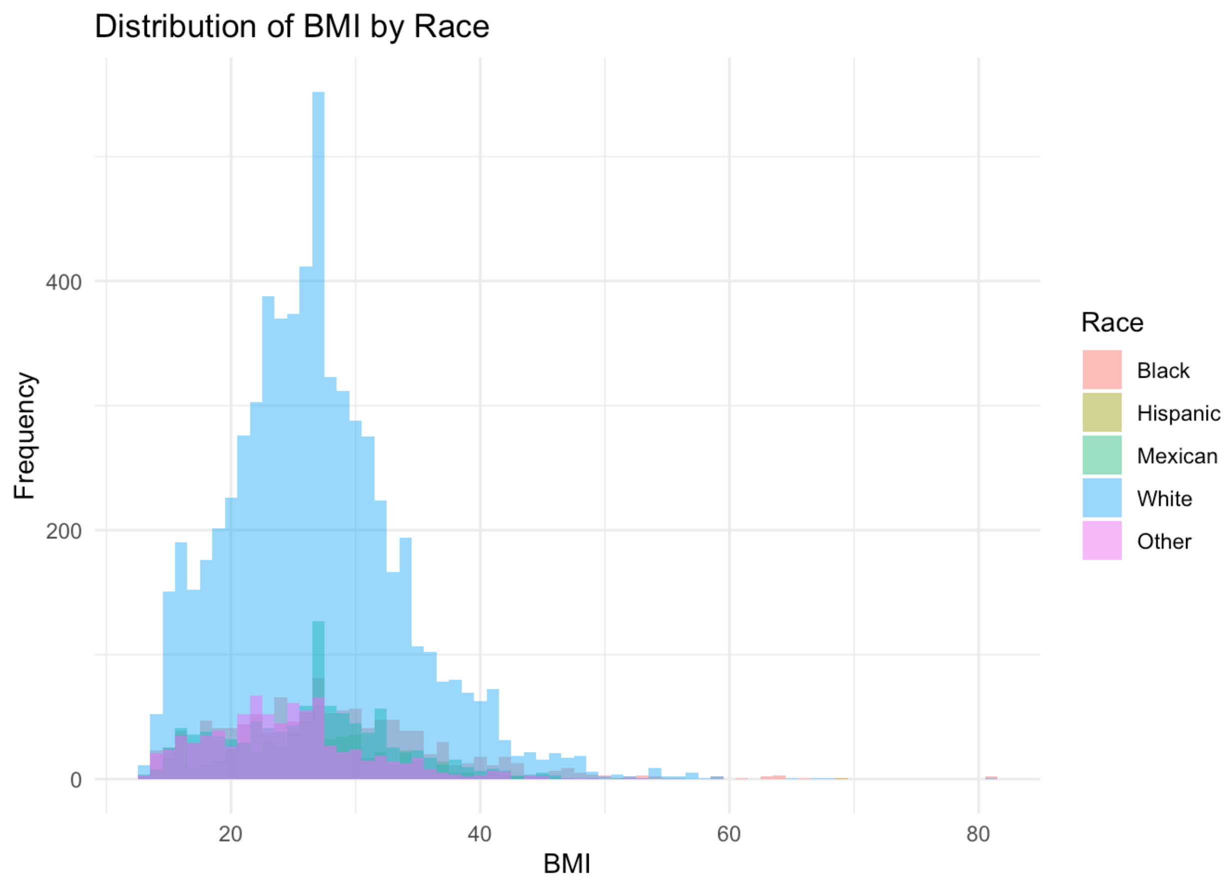

```
#plots distribution of BMI
ggplot(data, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  labs(title = "Distribution of BMI",
       x = "BMI",
       y = "Frequency") +
  theme_minimal()
```



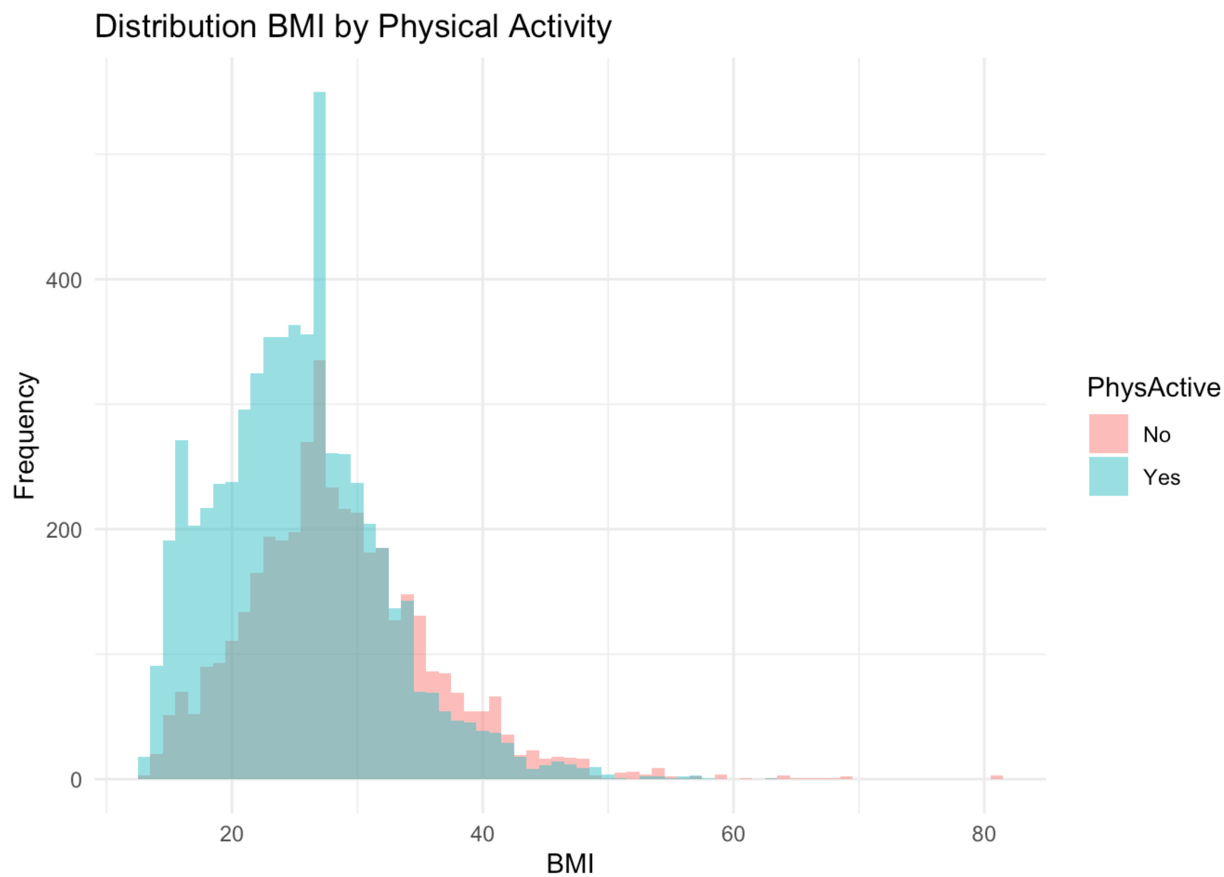

```
#plots distribution of BMI by Gender
ggplot(data, aes(x = BMI, fill = Gender)) +
  geom_histogram(binwidth = 1, position = "identity", alpha = 0.5) +
  labs(title = "Distribution of BMI by Gender",
       x = "BMI",
       y = "Frequency",
       fill = "Gender") +
  theme_minimal()
```



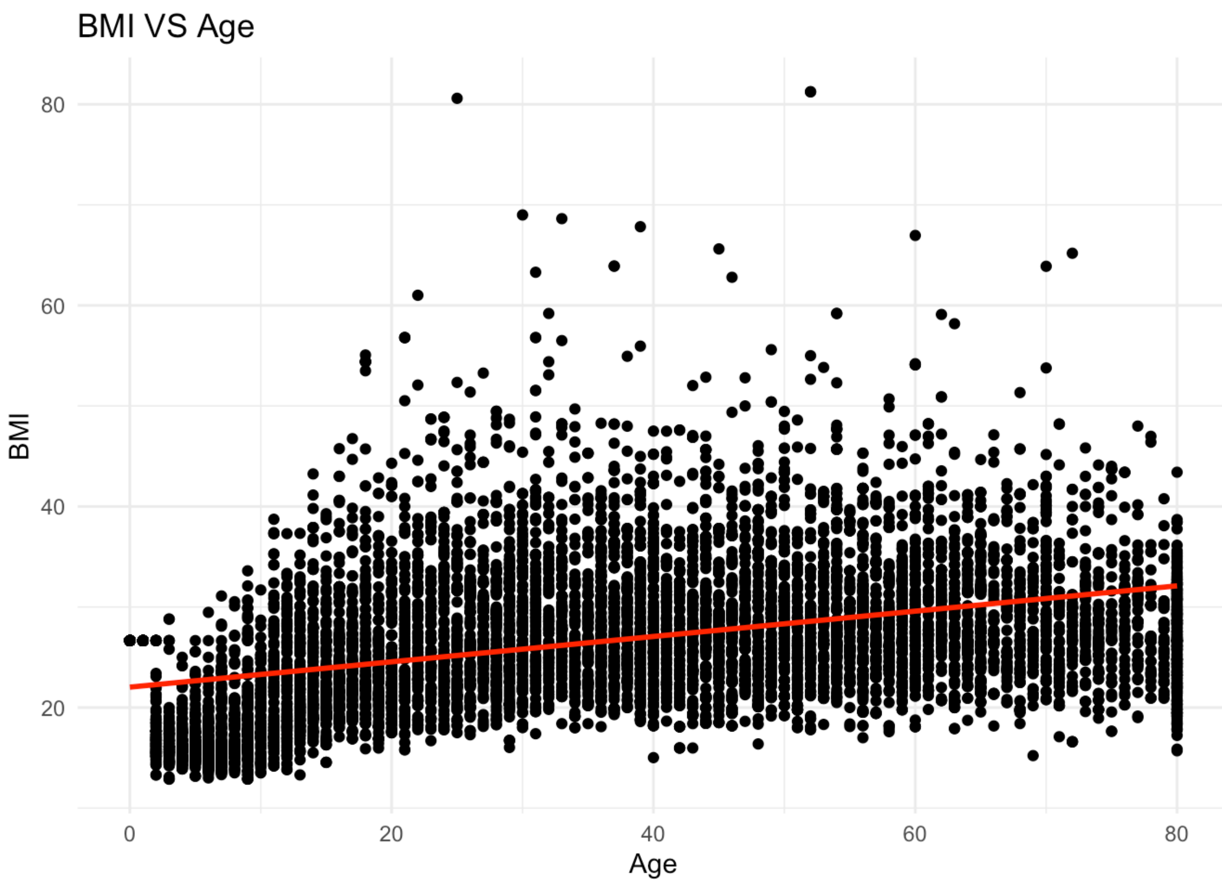

```
#plots distribuition of BMI by Race
ggplot(data, aes(x = BMI, fill = Race1)) +
  geom_histogram(binwidth = 1, position = "identity", alpha = 0.5) +
  labs(title = "Distribution of BMI by Race",
       x = "BMI",
       y = "Frequency",
       fill = "Race") +
  theme_minimal()
```



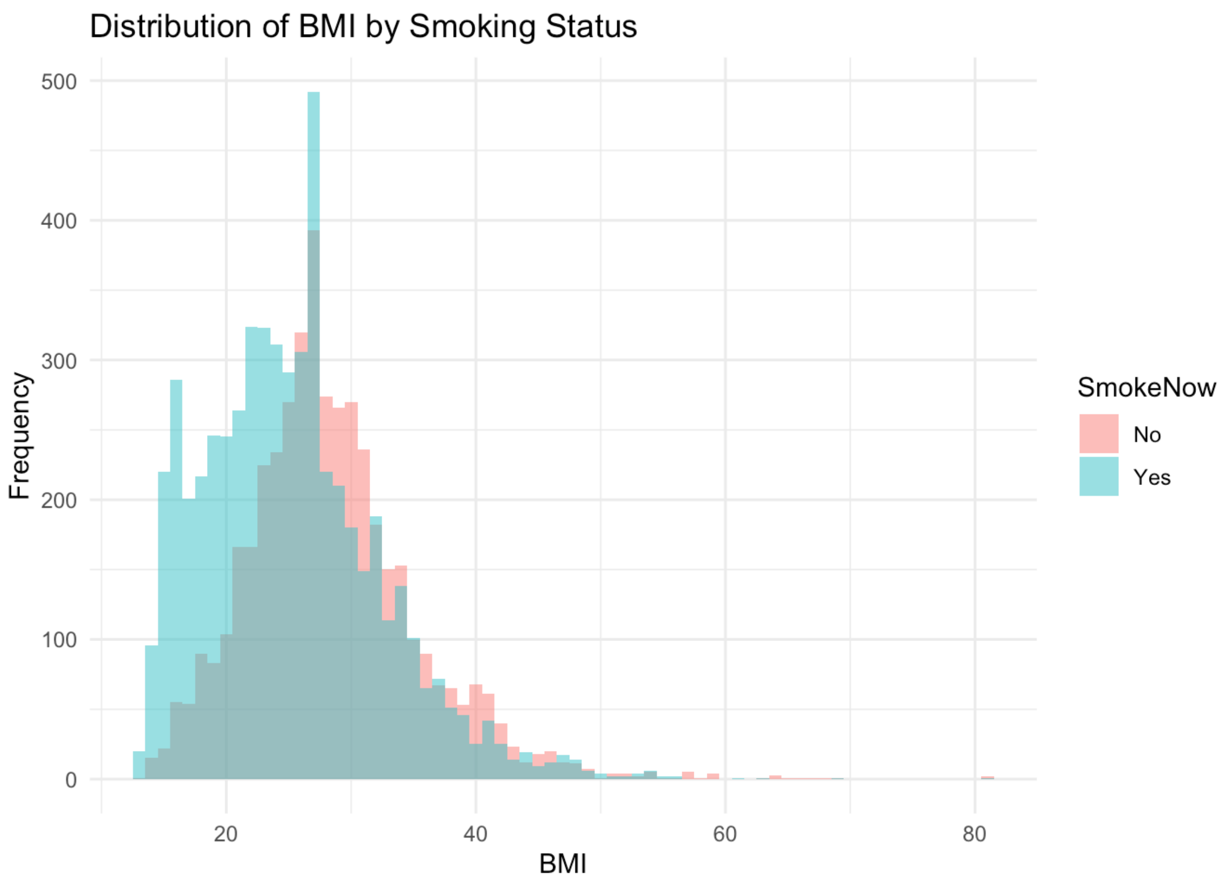

```
#plots distribution of BMI by PhyActice
ggplot(data, aes(x = BMI, fill = PhysActive)) +
  geom_histogram(binwidth = 1, position = "identity", alpha = 0.5) +
  labs(title = "Distribution BMI by Physical Activity",
       x = "BMI",
       y = "Frequency",
       fill = "PhysActive") +
  theme_minimal()
```




```
#plots distribution of BMI and Age
ggplot(data, aes(x = Age, y = BMI)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "BMI VS Age",
       x = "Age",
       y = "BMI") +
  theme_minimal()
```




```
#plots distribution of BMI by smoking status
ggplot(data, aes(x = BMI, fill = SmokeNow)) +
  geom_histogram(binwidth = 1, position = "identity", alpha = 0.5) +
  labs(title = "Distribution of BMI by Smoking Status",
       x = "BMI",
       y = "Frequency",
       fill = "SmokeNow") +
  theme_minimal()
```



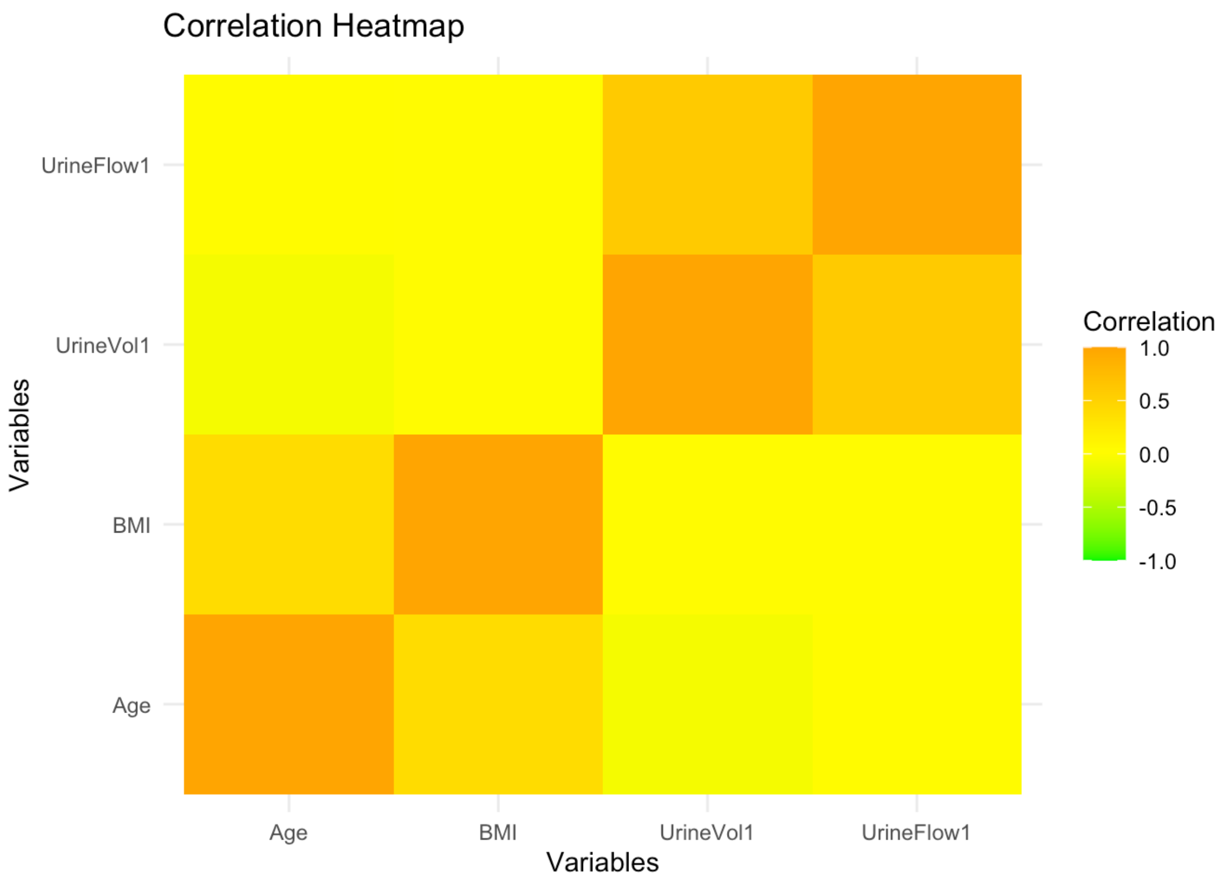
#calculates correlation matrix between the variables Age, BMI, UrineVol1, UrineFlow1
corr_mat <- cor(data[, c("Age", "BMI", "UrineVol1", "UrineFlow1")])
corr_mat

```
##                   Age        BMI   UrineVol1 UrineFlow1
## Age        1.00000000 0.38961061 -0.06861398 0.03339869
## BMI        0.38961061 1.00000000  0.01661327 0.01059941
## UrineVol1 -0.06861398 0.01661327  1.00000000 0.59148715
## UrineFlow1  0.03339869 0.01059941  0.59148715 1.00000000
```

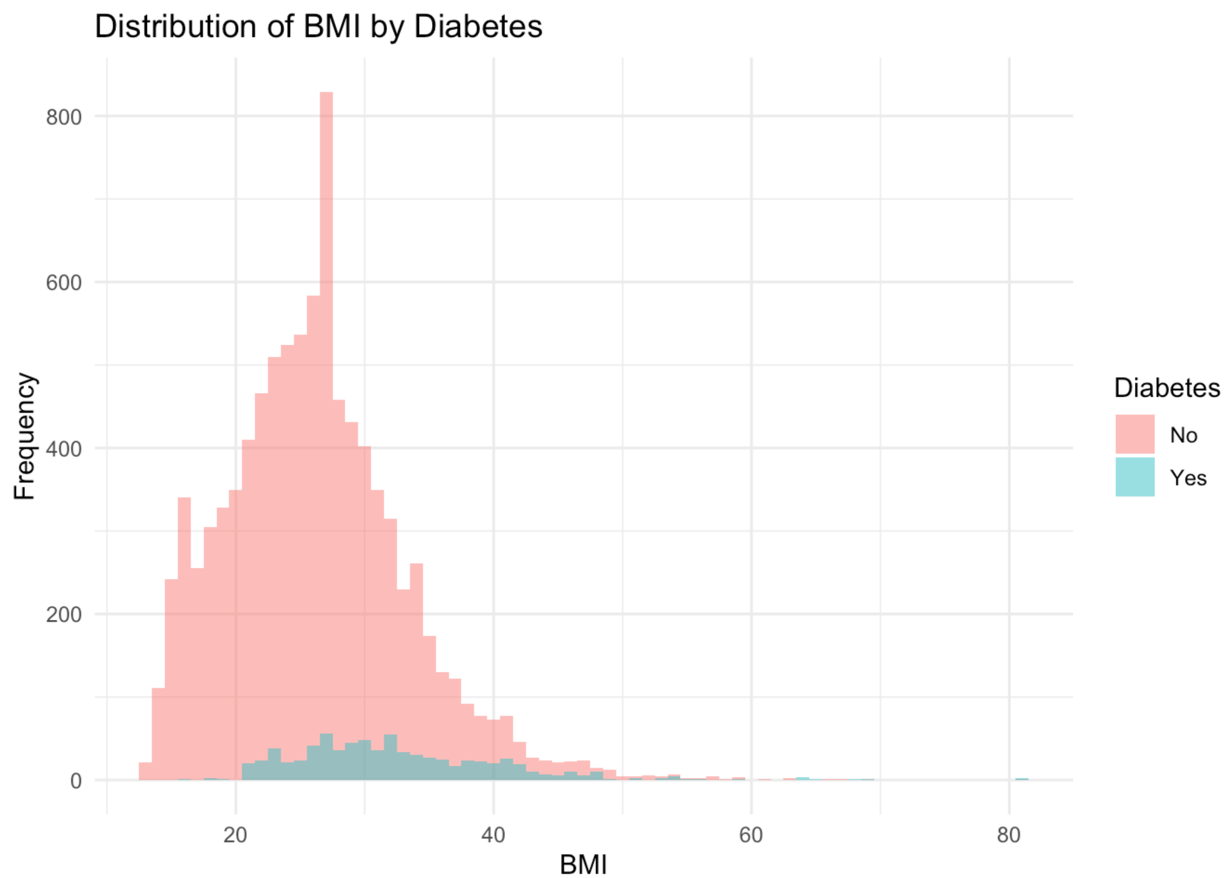
#loads reshape2 package
library(reshape2)

#plots correlation matrix of Age, BMI, UrineVol1, UrineFlow1
ggplot(data = melt(corr_mat), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +

```
  scale_fill_gradient2(low = "green", mid = "yellow", high = "orange",
                       midpoint = 0, limits = c(-1,1),
                       name="Correlation") +
  labs(title = "Correlation Heatmap",
       x = "Variables",
       y = "Variables") +
  theme_minimal()
```



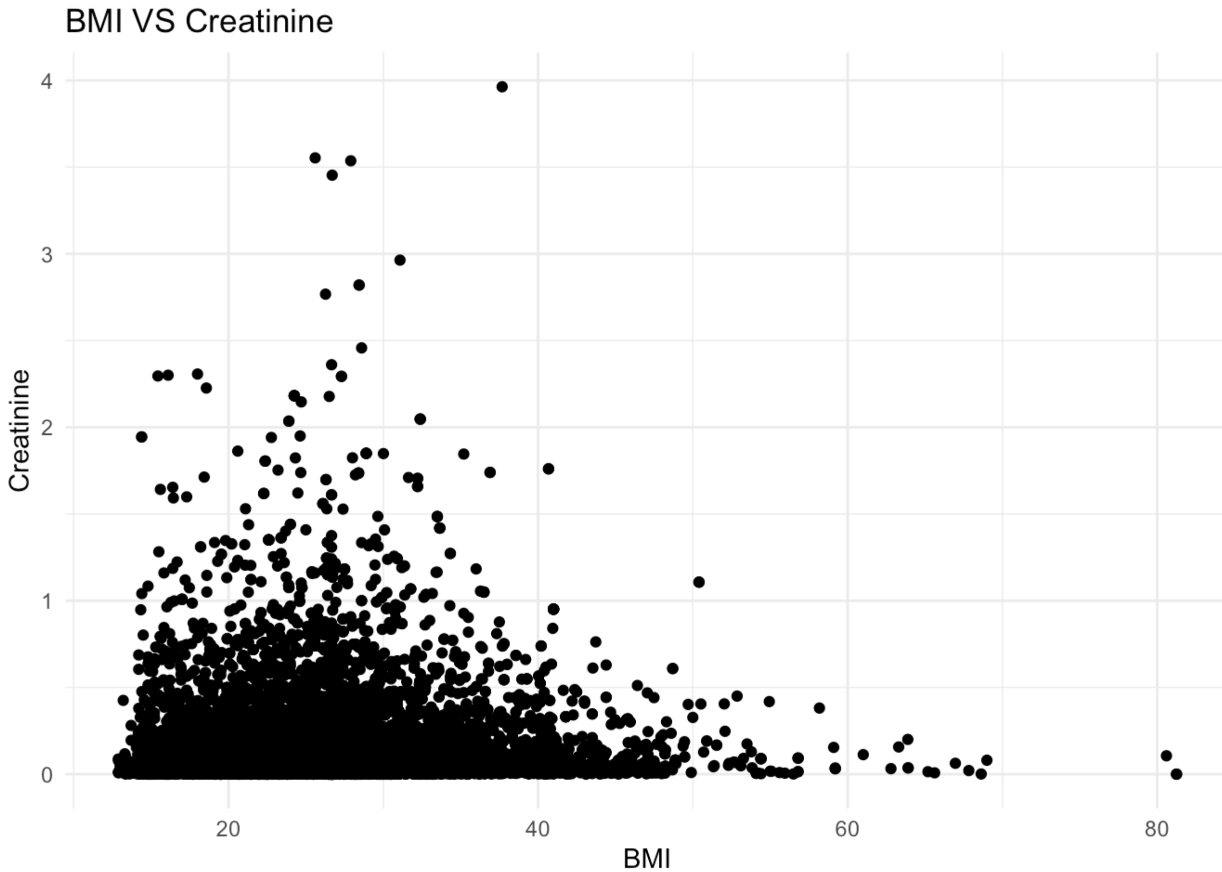

```
#plots distribution of BMI by Diabetes
ggplot(data, aes(x = BMI, fill = Diabetes)) +
  geom_histogram(binwidth = 1, position = "identity", alpha = 0.5) +
  labs(title = "Distribution of BMI by Diabetes",
       x = "BMI",
       y = "Frequency",
       fill = "Diabetes") +
  theme_minimal()
```



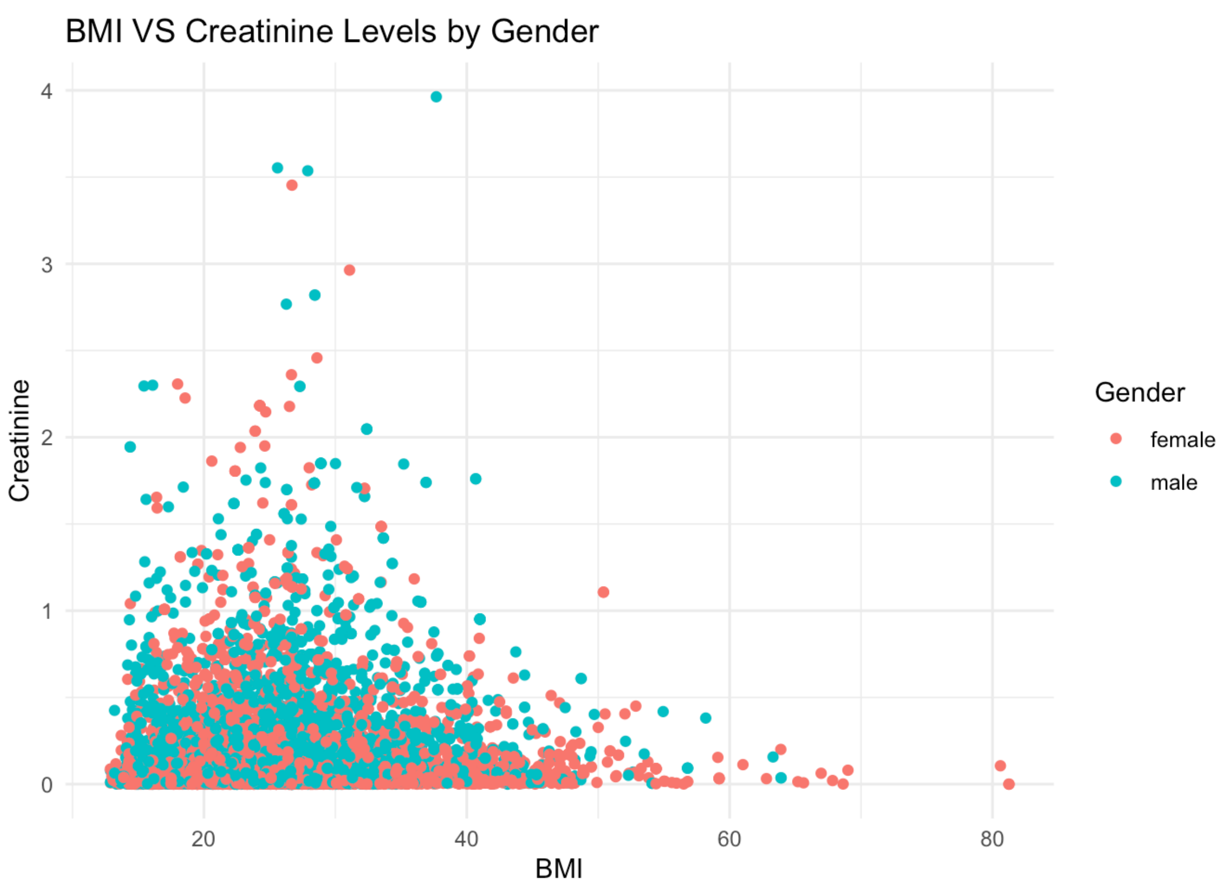

```
#calculates correlation between BMI and Creatinine
corr_mat <- cor(data$BMI, data$Creatinine)
corr_mat
```

```
## [1] -0.002472583
```

```
#plots distribution of BMI and Creatinine
ggplot(data, aes(x = BMI, y = Creatinine)) +
  geom_point() +
  labs(title = "BMI VS Creatinine",
       x = "BMI",
       y = "Creatinine") +
  theme_minimal()
```




```
#plots distribution of BMI VS Creatinine by Gender
ggplot(data, aes(x = BMI, y = Creatinine, color = Gender)) +
  geom_point() +
  labs(title = "BMI VS Creatinine Levels by Gender",
       x = "BMI",
       y = "Creatinine",
       color = "Gender") +
  theme_minimal()
```



#fits linear regression model Creatinine as response variable and BMI, Age, Gender, Race1, PhyActive as predictor variables
model <- lm(Creatinine ~ BMI + Age + Gender + Race1 + PhysActive, data = data)
model

```
##
## Call:
## lm(formula = Creatinine ~ BMI + Age + Gender + Race1 + PhysActive,
##     data = data)
##
## Coefficients:
##   (Intercept)           BMI           Age     Gendermale  Race1Hispanic
Race1Mexican
##     0.0792620     0.0006605    -0.0001182     0.0499709      0.0276650
0.0084645
##    Race1White      Race1Other  PhysActiveYes
```

```
##     0.0421146      0.0632144      0.0249557
```

#shows summary of linear regression model
summary(model)

```
##
## Call:
## lm(formula = Creatinine ~ BMI + Age + Gender + Race1 + PhysActive,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2312 -0.1380 -0.0924  0.0206  3.7743
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0792620  0.0147745   5.365 8.29e-08 ***
## BMI            0.0006605  0.0004187   1.577   0.1148
## Age           -0.0001182  0.0001406  -0.841   0.4005
## Gendermale     0.0499709  0.0055150   9.061  < 2e-16 ***
## Race1Hispanic  0.0276650  0.0137000   2.019   0.0435 *
## Race1Mexican   0.0084645  0.0117868   0.718   0.4727
## Race1White     0.0421146  0.0087540   4.811 1.52e-06 ***
## Race1Other     0.0632144  0.0126143   5.011 5.50e-07 ***
## PhysActiveYes  0.0249557  0.0059499   4.194 2.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2751 on 9991 degrees of freedom
## Multiple R-squared:  0.01475,    Adjusted R-squared:  0.01396
## F-statistic: 18.69 on 8 and 9991 DF,  p-value: < 2.2e-16
```

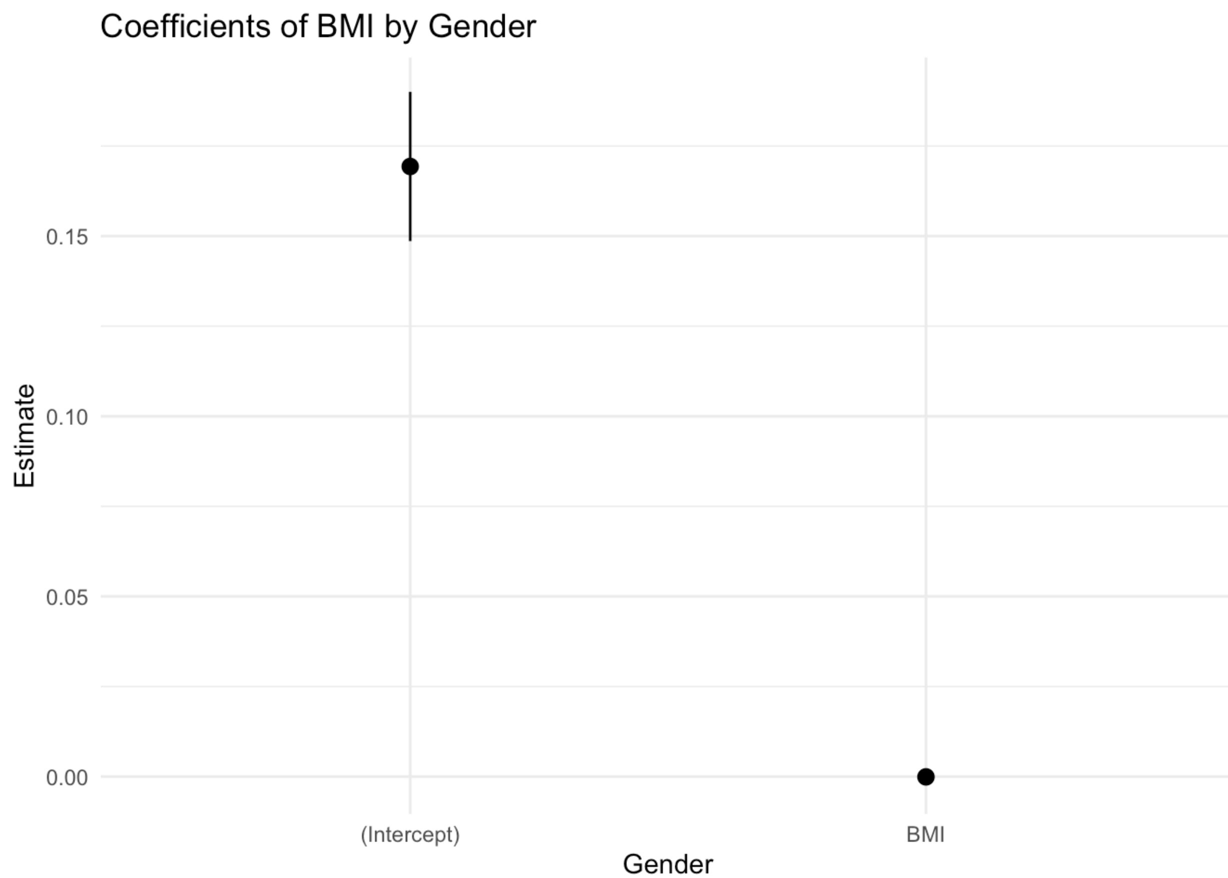
#fits linear regression model Creatine as reponse variable and BMI as predictor variable
strat_analysis <- lm(Creatinine ~ BMI, data = data)
strat_analysis

```
##
## Call:
## lm(formula = Creatinine ~ BMI, data = data)
##
## Coefficients:
## (Intercept)          BMI
##   0.1693335   -0.0000946
```

#provides coefficients, p-values, etc,... from the summary of the model
strat_results <- coef(summary(strat_analysis))
strat_results

```
##                  Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  1.693335e-01 0.010570120 16.0200211 4.736337e-57
## BMI         -9.459654e-05 0.000382619 -0.2472343 8.047320e-01
```

#plots the distribution of coefficients of BMI by Gender
ggplot(data = data.frame(strat_results), aes(x = row.names(strat_results), y = Estimate, ymin = Estimate - 1.96 * Std..Error, ymax = Estimate + 1.96 * Std..Error)) +
 geom_pointrange() +
 labs(title = "Coefficients of BMI by Gender",
   x = "Gender",
   y = "Estimate") +
 theme_minimal()



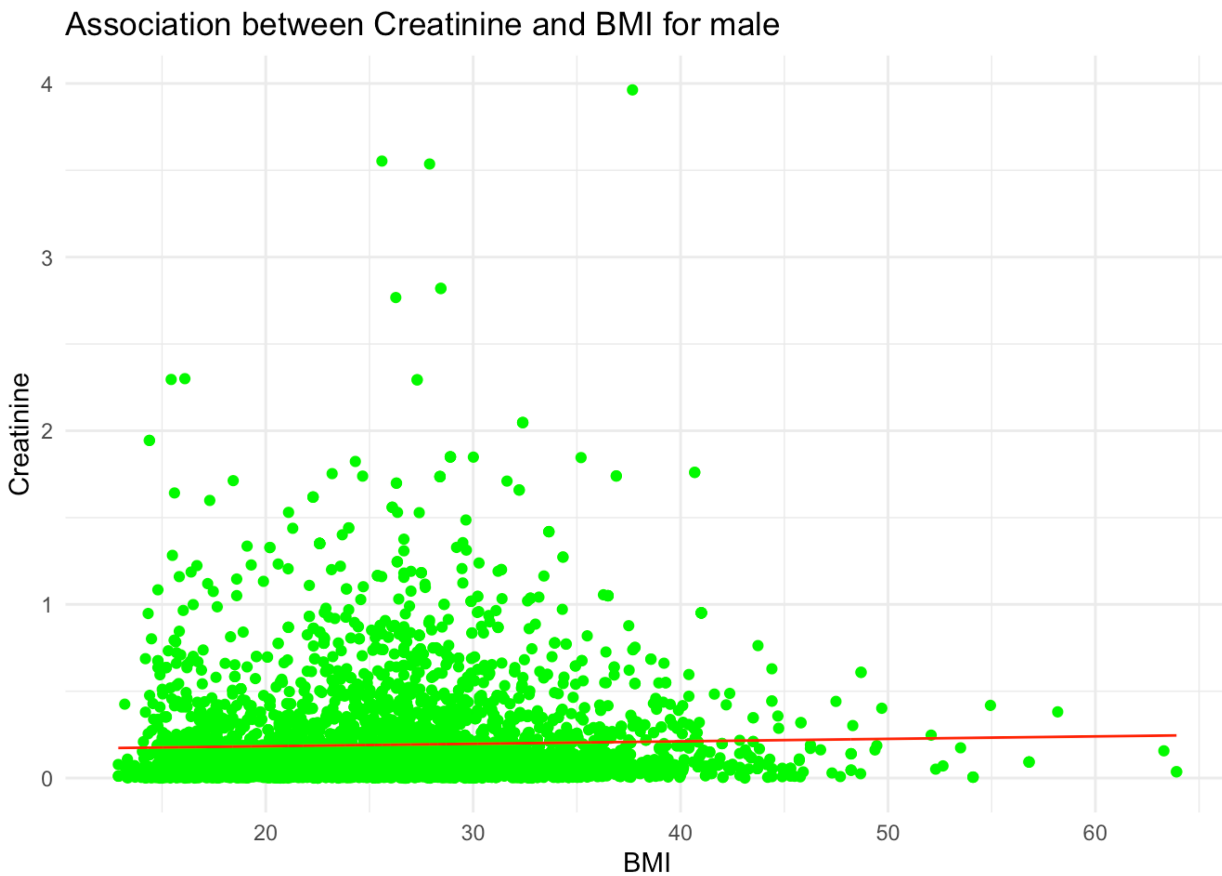

```
#creates subgroups of gender column
subgroups <- unique(data[["Gender"]])

#list to store regression model results for each subgroup
strat_models <- list()

#fit regression models for each subgroup
for (s in subgroups) {
  subgroup_data <- filter(data, !!as.name("Gender") == s)
  model <- lm(Creatinine ~ BMI, data = subgroup_data)
  strat_models[[s]] <- model
}
strat_models
```

```
## $male
##
## Call:
## lm(formula = Creatinine ~ BMI, data = subgroup_data)
##
## Coefficients:
## (Intercept)          BMI
##    0.154702     0.001416
##
##
## $female
##
## Call:
## lm(formula = Creatinine ~ BMI, data = subgroup_data)
##
## Coefficients:
## (Intercept)          BMI
##    0.171356    -0.001114
```

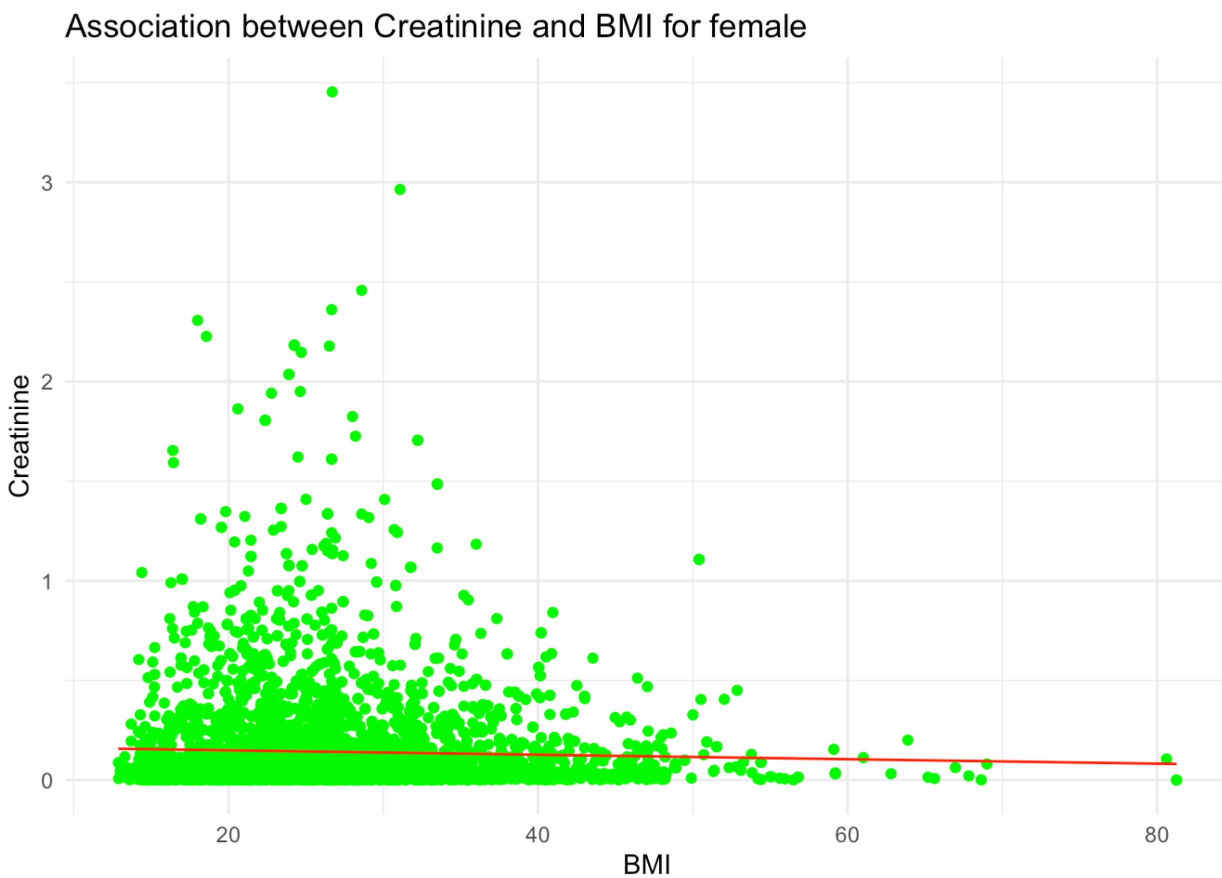
```
#iterates over each subgroup
lapply(names(strat_models), function(subgroup) {

  #extract data for the current subgroup
  subgroup_data <- filter(data, !!as.name("Gender") == subgroup)

  #checks the variable 'Creatinine' exists in the subgroup
  if ("Creatinine" %in% colnames(subgroup_data)) {

    #predicts creatinine using the regression model
    subgroup_data$Predicted_Creatinine <- predict(strat_models[[subgroup]], newdata =
subgroup_data)
    #plot a scatter plot of association between creatinine and BMI for each subgroup
    ggplot(data = subgroup_data, aes(x = BMI, y = Creatinine)) +
      geom_point(color = "green") +
      geom_line(aes(y = Predicted_Creatinine), color = "red") +
      labs(title = paste("Association between Creatinine and BMI for", subgroup),
```

```
      x = "BMI",
      y = "Creatinine") +
    theme_minimal()

  } else {
    NULL
  }
})
```

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

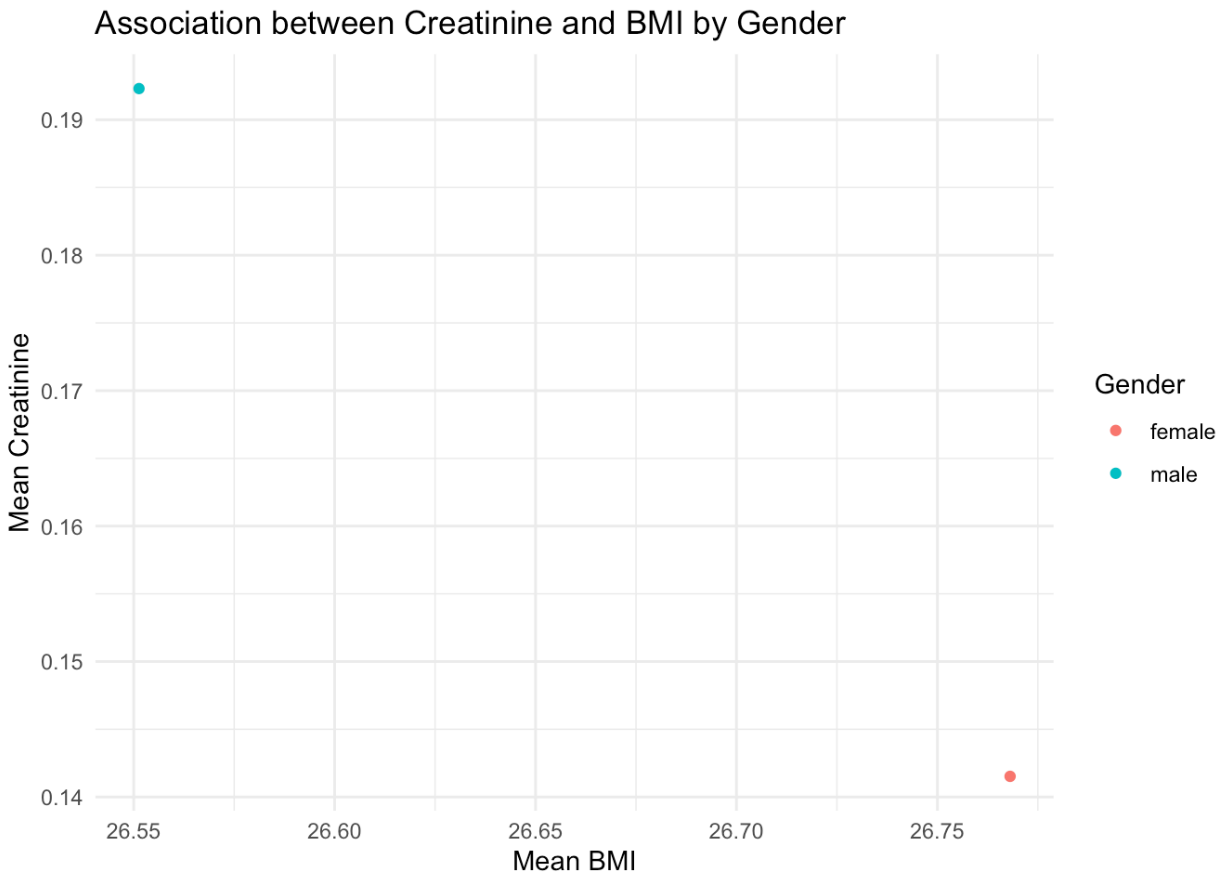



```
## [[2]]
```




```
#loads dplyr package
library(dplyr)

#group data of gender
grouped_data <- data %>%
  group_by(Gender)

#calculate summary statistics for Creatinine and BMI within each group
summary_stats <- grouped_data %>%
  summarize(
    Mean_Creatinine = mean(Creatinine, na.rm = TRUE),
    Mean_BMI = mean(BMI, na.rm = TRUE)
  )

#plots association between Creatinine and BMI for each subgroup
ggplot(summary_stats, aes(x = Mean_BMI, y = Mean_Creatinine, color = Gender)) +
  geom_point() +
  labs(title = "Association between Creatinine and BMI by Gender",
```

```
    x = "Mean BMI",
    y = "Mean Creatinine") +
 theme_minimal()
```




```
#if obesity value is yes then it will rewrite it as 1 otherwise 0
data$Obesity <- ifelse(data$Obesity == "Yes", 1, 0)

#fits logistic regression model response variable is Obesity and predictor variable is Creatinine
model <- glm(Obesity ~ Creatinine, data = data, family = binomial)
model
```

```
##
## Call:  glm(formula = Obesity ~ Creatinine, family = binomial, data = data)
##
## Coefficients:
## (Intercept)   Creatinine
##     -0.9265      -0.2137
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
## Null Deviance:       11790
## Residual Deviance: 11790     AIC: 11790
```

#provides summary statistics of the model
summary(model)

```
##
## Call:
## glm(formula = Obesity ~ Creatinine, family = binomial, data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.92654    0.02620 -35.363   <2e-16 ***
## Creatinine  -0.21371    0.08574  -2.492   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11794  on 9999  degrees of freedom
## Residual deviance: 11788  on 9998  degrees of freedom
## AIC: 11792
##
## Number of Fisher Scoring iterations: 4
```

#fits logistic regression model response var is Obesity snd predictor variables are Creatinine, Age, Gender, Race1
model <- glm(Obesity ~ Creatinine + Age + Gender + Race1, data = data, family = binomial)
model

```
##
## Call:  glm(formula = Obesity ~ Creatinine + Age + Gender + Race1, family = 
binomial,
##     data = data)
##
## Coefficients:
##   (Intercept)     Creatinine            Age     Gendermale  Race1Hispanic
Race1Mexican
##     -1.447836      -0.110823       0.026444       0.005209      -0.503161
-0.302378
##    Race1White     Race1Other
##     -0.632645      -1.128704
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9992 Residual
## Null Deviance:       11790
## Residual Deviance: 11050     AIC: 11070
```

#provides summary statistics of the model
summary(model)

```
##
## Call:
## glm(formula = Obesity ~ Creatinine + Age + Gender + Race1, family = binomi
al,
##     data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.447836   0.078453 -18.455  < 2e-16 ***
## Creatinine    -0.110823   0.087574  -1.265  0.20570
```

```
## Age            0.026444   0.001096  24.135  < 2e-16 ***
## Gendermale     0.005209   0.046751   0.111  0.91129
## Race1Hispanic -0.503161   0.113816  -4.421 9.83e-06 ***
## Race1Mexican  -0.302378   0.095652  -3.161  0.00157 **
## Race1White    -0.632645   0.069304  -9.129  < 2e-16 ***
## Race1Other    -1.128704   0.115892  -9.739  < 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: 11794  on 9999  degrees of freedom
## Residual deviance: 11051  on 9992  degrees of freedom
## AIC: 11067
##
## Number of Fisher Scoring iterations: 4
```

```
#subset data having obesity value 1
obese <- subset(data, Obesity == 1)

#subset data having obesity value 0
non_obese <- subset(data, Obesity == 0)

#t-test for creatinine with obese and without obese
t_test_result <- t.test(obese$Creatinine, non_obese$Creatinine)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  obese$Creatinine and non_obese$Creatinine
## t = -2.6363, df = 5599.4, p-value = 0.008405
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.026979448 -0.003966988
## sample estimates:
## mean of x mean of y
## 0.1556183 0.1710915
```

```
#categorizing ages into three groups
data$AgeGroup <- cut(data$Age, breaks = c(0, 30, 60, max(data$Age)),
                     labels = c("Young Adults", "Middle-Aged Adults", "Older Adults"),
                     include.lowest = TRUE)

#storing unique values of AgeGroup
age_groups <- unique(data$AgeGroup)

#stores t-test results
t_test_results <- list()

#iterating each unique value of AgeGroup
for (age_group in age_groups) {

  #subset the AgeGroup of different categories
  age_group_data <- subset(data, AgeGroup == age_group)
  #subset the data with obese(1)
  obese_age <- subset(age_group_data, Obesity == 1)
  #subset the data without obese(0)
  non_obese_age <- subset(age_group_data, Obesity == 0)
  #t-test on with obese and without obese
  t_test_result_age <- t.test(obese_age$Creatinine, non_obese_age$Creatinine)
  #intializing the test results
  t_test_results[[age_group]] <- t_test_result_age
```

}

#can see the results
t_test_results

```
## $`Middle-Aged Adults`
##
##  Welch Two Sample t-test
##
## data:  obese_age$Creatinine and non_obese_age$Creatinine
## t = -2.5095, df = 3450.4, p-value = 0.01213
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.042873598 -0.005264295
## sample estimates:
## mean of x mean of y
##  0.179872  0.203941
##
##
## $`Young Adults`
##
##  Welch Two Sample t-test
##
## data:  obese_age$Creatinine and non_obese_age$Creatinine
## t = -0.97326, df = 1046.9, p-value = 0.3306
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.029016867  0.009775837
## sample estimates:
## mean of x mean of y
## 0.1510731 0.1606936
##
##
## $`Older Adults`
##
```

```
##  Welch Two Sample t-test
##
## data:  obese_age$Creatinine and non_obese_age$Creatinine
## t = -2.6689, df = 1350.2, p-value = 0.007702
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.048973248 -0.007478655
## sample estimates:
## mean of x mean of y
## 0.1023153 0.1305413
```