Analysis

Jestin George

2025-04-29

Final Report on the The Impact of Dietary Intake on Blood Pressure using NHANES Dataset

Blood pressure is a key indicator of cardiovascular health, and dietary intake plays a critical role in its regulation. Components like sodium, calcium, magnesium, total fat, saturated fat, cholesterol, alcohol, and sugar have varying impacts on blood pressure. While excess sodium is linked to hypertension, calcium and magnesium may help lower blood pressure. High fat, alcohol, and sugar intake are associated with cardiovascular issues that can raise blood pressure. This analysis, using NHANES data, examines how these nutrients influence systolic and diastolic blood pressure, providing insights for dietary strategies to manage hypertension.

Study Aims

Primary Study Aim

The main objective is to use the NHANES dataset to examine the relationship between nutrient intake (sodium, calcium, magnesium, total fat, saturated fat, cholesterol, alcohol, and sugar) and blood pressure (systolic and diastolic).

Secondary Study Aims

• To evaluate how factors such as age and bmi influence the association between nutrient intake and blood pressure.

• To explore whether specific nutrients, like calcium and magnesium, have protective effects on blood pressure across different population groups.

Loading necessary libraries

```
library(ppcor)
```

```
## Loading required package: MASS
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(ggplot2)
library(haven)
library(nhanesA)
```

Data Preparation steps:

1. Loading NHANES Dataset:

* The Demographic, Examination, Dietary and basic data has already been downloaded directly from the NHANES website source. This dataset includes detailed information on the nutrient intake, health-related information and basic details collected from a representative sample of the US population

2. Data Collection & Preparation

The fields that are extracted from various datasets are:

- BP Systolic
- BP Diastolic
- Sodium
- Magnesium
- Calcium
- Sugar
- TFat
- SFat
- Cholestrol
- Age
- BMI
- Gender

Haven library has been made use of in order to extract the necessary data from .xpt files.

3. Data Cleaning & Preprocessing:

* To make sure that the data is well managed and to help with the stability of the further analyses the records which included null values were omitted. Also we have only considered individuals who are older than 18 for our analyses.

Insights

- While data extraction focusing on fields that are important for the planned analyses are very important.
- Proper data cleaning and data preprocessing is critical for an accurate analyses.

```
#Data Loading
diet1 <- read_xpt("DR1TOT_L.xpt")
diet2 <- read_xpt("DR2TOT_L.xpt")
examination <- read_xpt("BPXO_L.xpt")
demographic <- read_xpt("DEMO_L.xpt")
basic <- read_xpt("BMX_L.xpt")

## DATA COLLECTION

# Create a new data frame with SEQN and necessary fields
sample_diet <- data.frame(SEQN = diet1$SEQN, SODIUM1 = diet1$DR1TSODI, SODIUM2 = di
et2$DR2TSODI, CALCIUM1 = diet1$DR1TCALC, CALCIUM2 = diet2$DR2TCALC, MAGNESIUM1 = di
et1$DR1TMAGN, MAGNESIUM2 = diet2$DR2TMAGN, TFAT1 = diet1$DR1TTFAT, TFAT2 = diet2$DR
2TTFAT, SFAT1 = diet1$DR1TSFAT, SFAT2 = diet2$DR2TSFAT, CHOLESTROL1 = diet1$DR1TCHO
L, CHOLESTROL2 = diet2$DR2TCHOL, SUGAR1 = diet1$DR1TSUGR, SUGAR2 = diet2$DR2TSUGR)
sample_examination <- data.frame(SEQN = examination$SEQN, SYSTOLIC1 = examination$B
PXOSY1, SYSTOLIC2 = examination$BPXOSY2, SYSTOLIC3 = examination$BPXOSY3, DIASTOLIC
1 = examination$BPXODI1, DIASTOLIC2 = examination$BPXODI2, DIASTOLIC3 = examination
$BPXODI3)
sample_demo <- data.frame(SEQN = demographic$SEQN, GENDER = demographic$RIAGENDR, A
GE = demographic$RIDAGEYR)
sample_basic <- data.frame(SEQN = basic$SEQN, HEIGHT = basic$BMXHT, WEIGHT = basic
$BMXWT)

# Filter sample_diet to retain only rows with SEQNO present in sample_examination
filtered_sample <- sample_diet %>%
  filter(SEQN %in% sample_examination$SEQN)

# Merge the filtered SAMPLE with sample_examination based on SEQNO
final_sample <- merge(filtered_sample, sample_examination, by = "SEQN")

# Filter sample_demo to retain only rows with SEQNO present in final_sample
filtered_sample_demo <- sample_demo %>%
  filter(SEQN %in% final_sample$SEQN)

# Merge the filtered SAMPLE with final_sample based on SEQNO
final_sample <- merge(filtered_sample_demo, final_sample, by = "SEQN")

# Filter sample_basic to retain only rows with SEQNO present in final_sample
filtered_sample_basic <- sample_basic %>%
  filter(SEQN %in% final_sample$SEQN)

# Merge the filtered SAMPLE with final_sample based on SEQNO
final_sample <- merge(filtered_sample_basic, final_sample, by = "SEQN")

## DATA CLEANING
#Removing records with null values
# data <- final_sample[!is.na(final_sample$SODIUM1) & !is.na(final_sample$SODIUM2),
]
# data <- data[!is.na(data$SYSTOLIC1), ]
# data <- data[!is.na(data$HEIGHT) & !is.na(data$WEIGHT), ]
data <- na.omit(final_sample)
```

```
# Remove records where AGE is less than 18
data <- data[data$AGE >= 18, ]

# DATA PREPARATION
data$SODIUM <- rowMeans(data[c("SODIUM1", "SODIUM2")], na.rm = TRUE)
data$CALCIUM <- rowMeans(data[c("CALCIUM1", "CALCIUM2")], na.rm = TRUE)
data$MAGNESIUM <- rowMeans(data[c("MAGNESIUM1", "MAGNESIUM2")], na.rm = TRUE)
data$TFAT <- rowMeans(data[c("TFAT1", "TFAT2")], na.rm = TRUE)
data$SFAT <- rowMeans(data[c("SFAT1", "SFAT2")], na.rm = TRUE)
data$CHOLESTROL <- rowMeans(data[c("CHOLESTROL1", "CHOLESTROL2")], na.rm = TRUE)
data$SUGAR <- rowMeans(data[c("SUGAR1", "SUGAR2")], na.rm = TRUE)
data$SYSTOLIC <- rowMeans(data[c("SYSTOLIC1", "SYSTOLIC2", "SYSTOLIC3")], na.rm = T
RUE)
data$DIASTOLIC <- rowMeans(data[c("DIASTOLIC1", "DIASTOLIC2", "DIASTOLIC3")], na.rm
= TRUE)
data$BMI <- data$WEIGHT / ((data$HEIGHT / 100)^2)

#Final data
final_data <- data.frame(SEQN = data$SEQN, AGE = data$AGE, GENDER = data$GENDER, BM
I = data$BMI, SODIUM = data$SODIUM,
                         MAGNESIUM = data$MAGNESIUM, CALCIUM = data$CALCIUM, SUGAR
= data$SUGAR, TFAT = data$TFAT, SFAT = data$SFAT,
                         CHOLESTROL = data$CHOLESTROL, SYSTOLIC = data$SYSTOLIC, DI
ASTOLIC = data$DIASTOLIC)

# Categorizing variables
# Blood Pressure Categories (normal, elevated, hypertension) for systolic and diast
olic
final_data$BPSys_Category <- cut(final_data$SYSTOLIC,
                                 breaks = c(-Inf, 120, 129, 139, 180, Inf),
                                 labels = c("Normal", "Elevated", "Stage 1 Hyperten
sion", "Stage 2 Hypertension", "Hypertensive Crisis"))

final_data$BPDia_Category <- cut(final_data$DIASTOLIC,
                                 breaks = c(-Inf, 80, 89, 99, 119, Inf),
                                 labels = c("Normal", "Elevated", "Stage 1 Hyperten
sion", "Stage 2 Hypertension", "Hypertensive Crisis"))

final_data$Age_Category <- cut(final_data$AGE,
                                 breaks = c(-Inf, 30, 50, Inf),
                                 labels = c("Young Adults", "Middle Aged Adults",
"Elderly"))
```

Descriptive Statistics

- Initially the **Frequency table** of the categorical variables are printed.
- The summary including the **mean, median, minimum value, maximum value and standard deviation** of the continuous variables are displayed.
- Further category-vise summary for BP Systolic and BP Diastolic were found out.

- Finally, anova tests were carried out comparing categories within BP Systolic and BP Diastolic.

```
##DESCRIPTIVE STATISTICS

#Overview of data
str(final_data)
```

```
## 'data.frame':    4282 obs. of  16 variables:
##  $ SEQN          : num  130378 130379 130380 130386 130387 ...
##  $ AGE           : num  43 66 44 34 68 59 74 51 67 26 ...
##  $ GENDER        : num  1 1 2 1 2 1 2 1 2 1 ...
##  $ BMI           : num  27 33.5 29.7 30.2 42.6 ...
##  $ SODIUM        : num  2144 4536 2934 3520 7098 ...
##  $ MAGNESIUM     : num  208 508 353 209 316 ...
##  $ CALCIUM       : num  430 726 988 636 605 ...
##  $ SUGAR         : num  22.1 113.5 110.9 49.1 210.1 ...
##  $ TFAT          : num  47.4 73.2 55.7 120.7 184.1 ...
##  $ SFAT          : num  15.1 17.2 18.4 42.7 44.3 ...
##  $ CHOLESTROL    : num  291 161 404 728 333 ...
##  $ SYSTOLIC      : num  133 117 109 115 141 ...
##  $ DIASTOLIC     : num  96 78.7 78.3 73.7 76 ...
##  $ BPSys_Category: Factor w/ 5 levels "Normal","Elevated",..: 3 1 1 1 4 3 4 1 4
1 ...
##  $ BPDia_Category: Factor w/ 5 levels "Normal","Elevated",..: 3 1 1 1 1 1 1 1 3
1 ...
##  $ Age_Category  : Factor w/ 3 levels "Young Adults",..: 2 3 2 2 3 3 3 3 3 1 ...
```

```
#Frequency distribution table
gender_dist <- table(final_data$GENDER)
bpsys_dist <- table(final_data$BPSys_Category)
bpdia_dist <- table(final_data$BPDia_Category)

# Print frequency tables
print(gender_dist)
```

```
##
##    1    2
## 1888 2394
```

```
print(bpsys_dist)
```

```
##
##               Normal             Elevated Stage 1 Hypertension
##                 2128                  833                  637
## Stage 2 Hypertension  Hypertensive Crisis
##                  653                   31
```

```
print(bpdia_dist)
```

```
##
##               Normal            Elevated Stage 1 Hypertension
##                 3140                 767                  294
## Stage 2 Hypertension  Hypertensive Crisis
##                   78                   3
```

```
# Summary statistics for continuous variables (Blood pressure, Nutrient intake)
summary_stats <- final_data %>%
  select(SYSTOLIC, DIASTOLIC, SODIUM, MAGNESIUM, CALCIUM, SUGAR, TFAT, SFAT, CHOLES
TROL, BMI) %>%
  summary()

# Calculate standard deviations
sd_stats <- final_data %>%
  select(SYSTOLIC, DIASTOLIC, SODIUM, MAGNESIUM, CALCIUM, SUGAR, TFAT, SFAT, CHOLES
TROL, BMI) %>%
  summarise_all(~sd(.))

# Print summary statistics and standard deviations
print(summary_stats)
```

```
##     SYSTOLIC        DIASTOLIC          SODIUM        MAGNESIUM
##  Min.   : 74.67   Min.   : 34.00   Min.   :  183   Min.   :  21.5
##  1st Qu.:110.33   1st Qu.: 67.00   1st Qu.: 2184   1st Qu.: 202.0
##  Median :120.33   Median : 73.67   Median : 2873   Median : 266.5
##  Mean   :122.59   Mean   : 74.34   Mean   : 3087   Mean   : 290.8
##  3rd Qu.:132.58   3rd Qu.: 80.92   3rd Qu.: 3778   3rd Qu.: 351.5
##  Max.   :211.00   Max.   :139.00   Max.   :16574   Max.   :1812.0
##     CALCIUM           SUGAR             TFAT              SFAT
##  Min.   :  52.5   Min.   :  0.66   Min.   :  5.535   Min.   :  1.64
##  1st Qu.: 563.0   1st Qu.: 56.22   1st Qu.: 57.580   1st Qu.: 17.31
##  Median : 795.2   Median : 83.70   Median : 79.010   Median : 24.61
##  Mean   : 875.5   Mean   : 95.28   Mean   : 84.266   Mean   : 26.96
##  3rd Qu.:1085.8   3rd Qu.:121.77   3rd Qu.:104.603   3rd Qu.: 34.09
##  Max.   :5941.5   Max.   :835.10   Max.   :377.005   Max.   :143.77
##    CHOLESTROL          BMI
##  Min.   :   0.0   Min.   :11.12
##  1st Qu.: 159.5   1st Qu.:24.70
##  Median : 265.2   Median :28.50
##  Mean   : 315.3   Mean   :29.74
##  3rd Qu.: 419.0   3rd Qu.:33.58
##  Max.   :1870.5   Max.   :68.90
```

```
print(sd_stats)
```

```
##   SYSTOLIC DIASTOLIC   SODIUM MAGNESIUM  CALCIUM    SUGAR     TFAT     SFAT
## 1 17.84511  10.79795 1341.015  133.3176 455.0184 58.23796 38.93838 13.73875
##   CHOLESTROL      BMI
## 1   213.0714 7.187911
```

```
# Descriptive statistics for systolic and diastolic blood pressure by category
bpsys_desc <- aggregate(SYSTOLIC ~ BPSys_Category, data = final_data,
                        FUN = function(x) c(mean = mean(x), sd = sd(x), median = me
dian(x), min = min(x), max = max(x)))

bpdia_desc <- aggregate(DIASTOLIC ~ BPDia_Category, data = final_data,
                        FUN = function(x) c(mean = mean(x), sd = sd(x), median = me
dian(x), min = min(x), max = max(x)))

# Print descriptive statistics
print("Descriptive statistics for Systolic Blood Pressure by Category:")
```

```
## [1] "Descriptive statistics for Systolic Blood Pressure by Category:"
```

```
print(bpsys_desc)
```

```
##         BPSys_Category SYSTOLIC.mean SYSTOLIC.sd SYSTOLIC.median SYSTOLIC.min
## 1               Normal    108.818452    7.962745      110.000000    74.666667
## 2             Elevated    124.499000    2.611668      124.333333   120.333333
## 3 Stage 1 Hypertension    133.787546    2.936292      133.666667   129.333333
## 4 Stage 2 Hypertension    150.809086    9.625413      148.666667   139.333333
## 5  Hypertensive Crisis    191.677419    8.933700      191.000000   181.000000
##   SYSTOLIC.max
## 1   120.000000
## 2   129.000000
## 3   139.000000
## 4   179.333333
## 5   211.000000
```

```
print("Descriptive statistics for Diastolic Blood Pressure by Category:")
```

```
## [1] "Descriptive statistics for Diastolic Blood Pressure by Category:"
```

```
print(bpdia_desc)
```

```
##         BPDia_Category DIASTOLIC.mean DIASTOLIC.sd DIASTOLIC.median
## 1               Normal      69.410403     7.021098        70.333333
## 2             Elevated      83.982182     2.470857        83.666667
## 3 Stage 1 Hypertension      92.995465     2.734764        92.666667
## 4 Stage 2 Hypertension     105.440171     5.266490       103.666667
## 5  Hypertensive Crisis     131.888889     6.710964       131.000000
##   DIASTOLIC.min DIASTOLIC.max
## 1     34.000000     80.000000
## 2     80.333333     89.000000
## 3     89.333333     99.000000
## 4     99.333333    117.333333
## 5    125.666667    139.000000
```

```
# Perform one-way ANOVA for systolic blood pressure categories
bpsys_anova_result <- aov(SYSTOLIC ~ BPSys_Category, data = final_data)

# Perform one-way ANOVA for diastolic blood pressure categories
bpdia_anova_result <- aov(DIASTOLIC ~ BPDia_Category, data = final_data)

# Print the ANOVA results
print("Systolic Blood Pressure ANOVA Result:")
```

```
## [1] "Systolic Blood Pressure ANOVA Result:"
```

```
summary(bpsys_anova_result)
```

```
##                  Df  Sum Sq Mean Sq F value Pr(>F)
## BPSys_Category    4 1154453  288613    5911 <2e-16 ***
## Residuals      4277  208823      49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("Diastolic Blood Pressure ANOVA Result:")
```

```
## [1] "Diastolic Blood Pressure ANOVA Result:"
```

```
summary(bpdia_anova_result)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## BPDia_Category    4 335313   83828    2188 <2e-16 ***
## Residuals      4277 163833      38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Insights

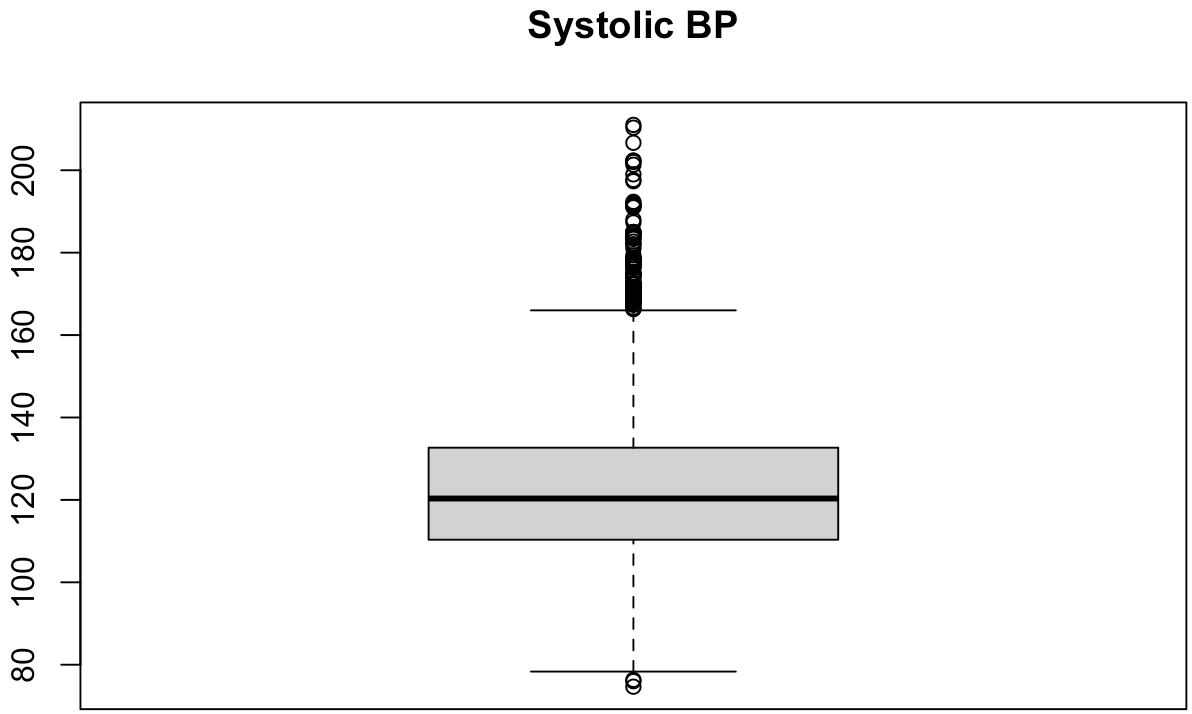
- With the summaries of continuous as well as categorical variables, the central tendencies which included mean, median, etc… of the dataset were figured out.
- With the ANOVA test carried out within the BP Systolic & BP Diastolic categories, it was confirmed that the Blood pressure differed substantially across categories.

Exploratory Data Analysis

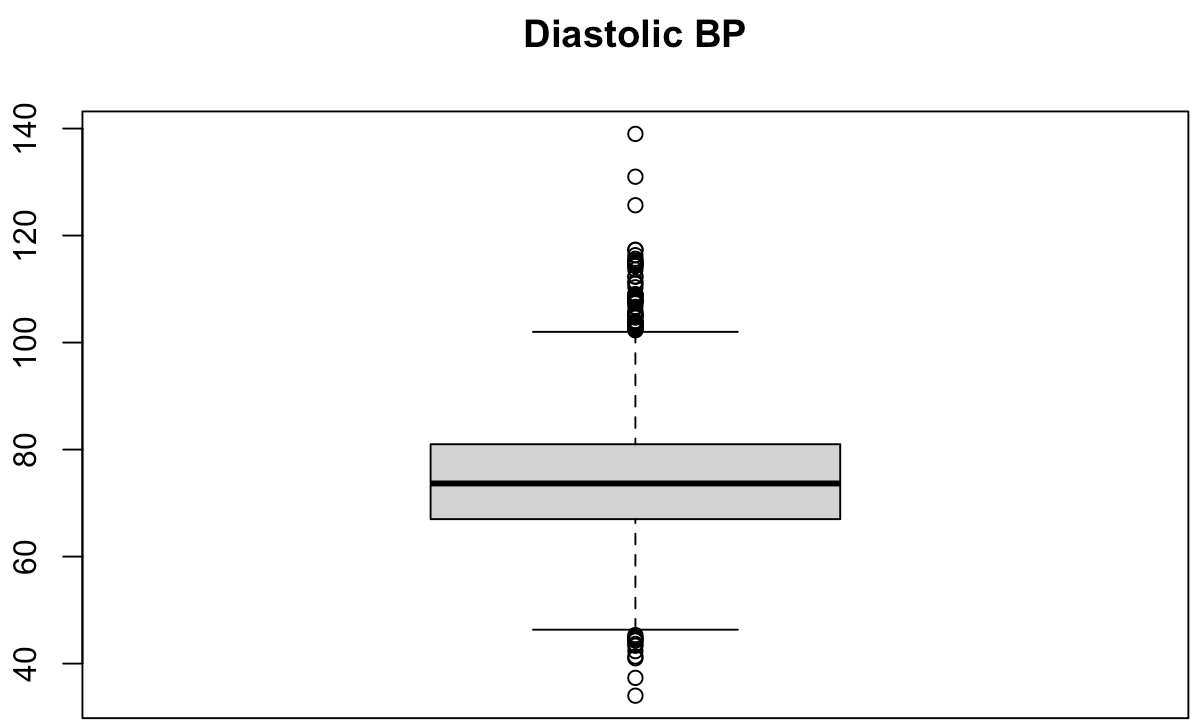
- Box plots were plotted in order to visualize the distribution of BP Systolic and BP Diastolic
- Histograms were made use to plot the distribution of the nutrients that are considered.
- Further again box plots were made use to plot Demographic data over Systolic BP & Diastolic BP.

Finally scatter plot were used to plot the dietary data over Systolic BP & Diastolic BP.

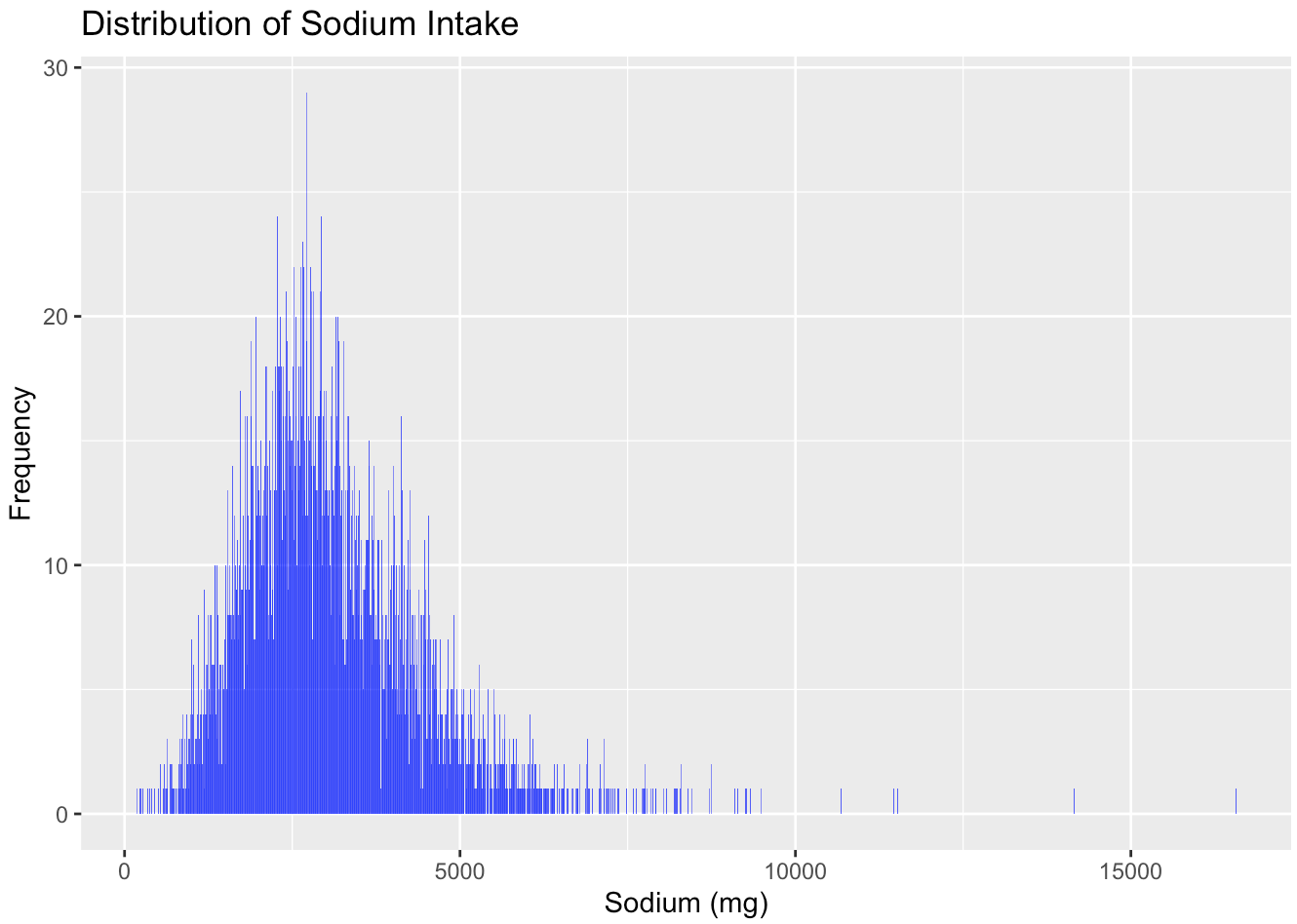
```
##EXPLORATORY DATA ANALYSIS

#Boxplots for examination data
boxplot(final_data$SYSTOLIC, main = "Systolic BP")
```



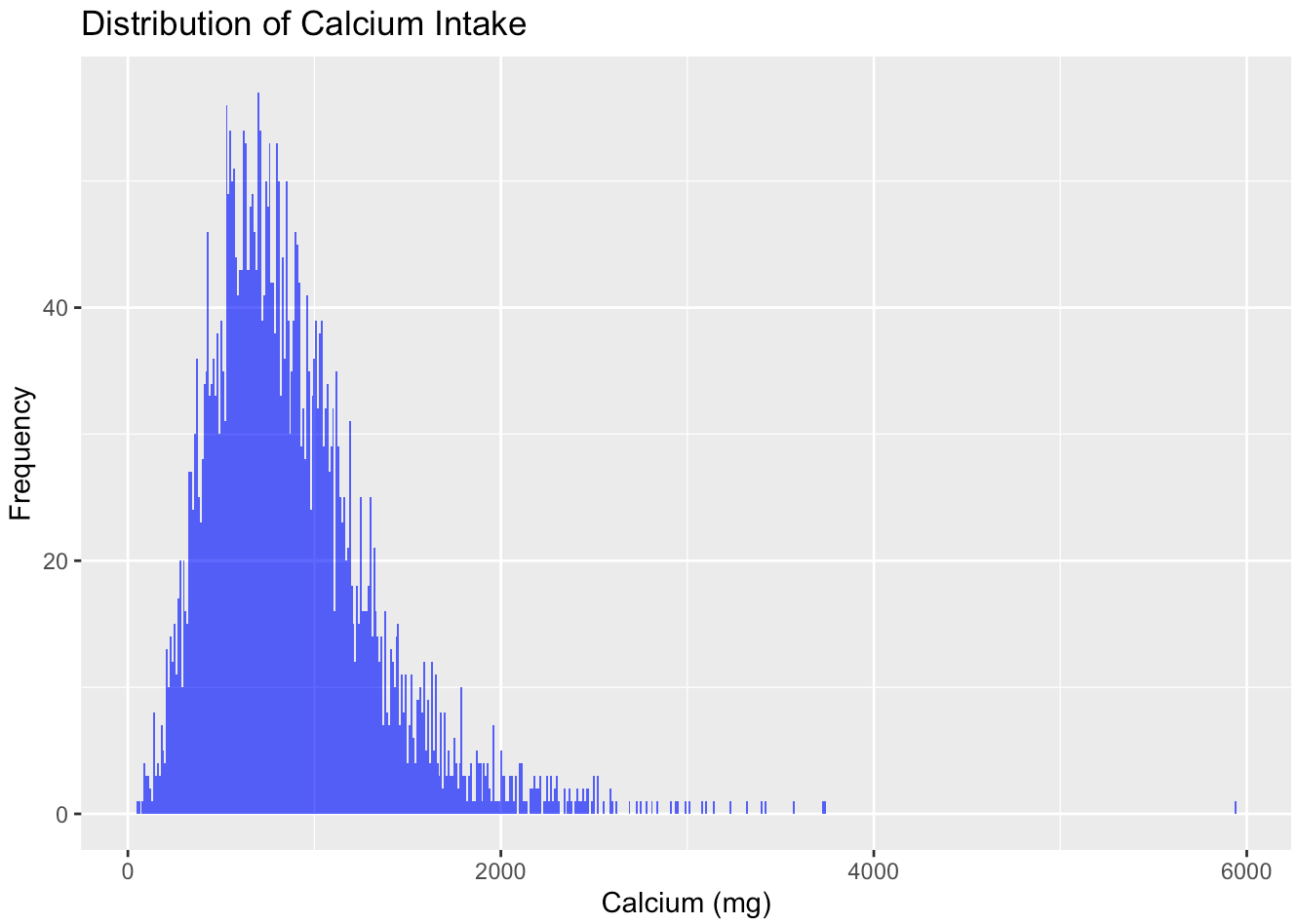

```
boxplot(final_data$DIASTOLIC, main = "Diastolic BP")
```



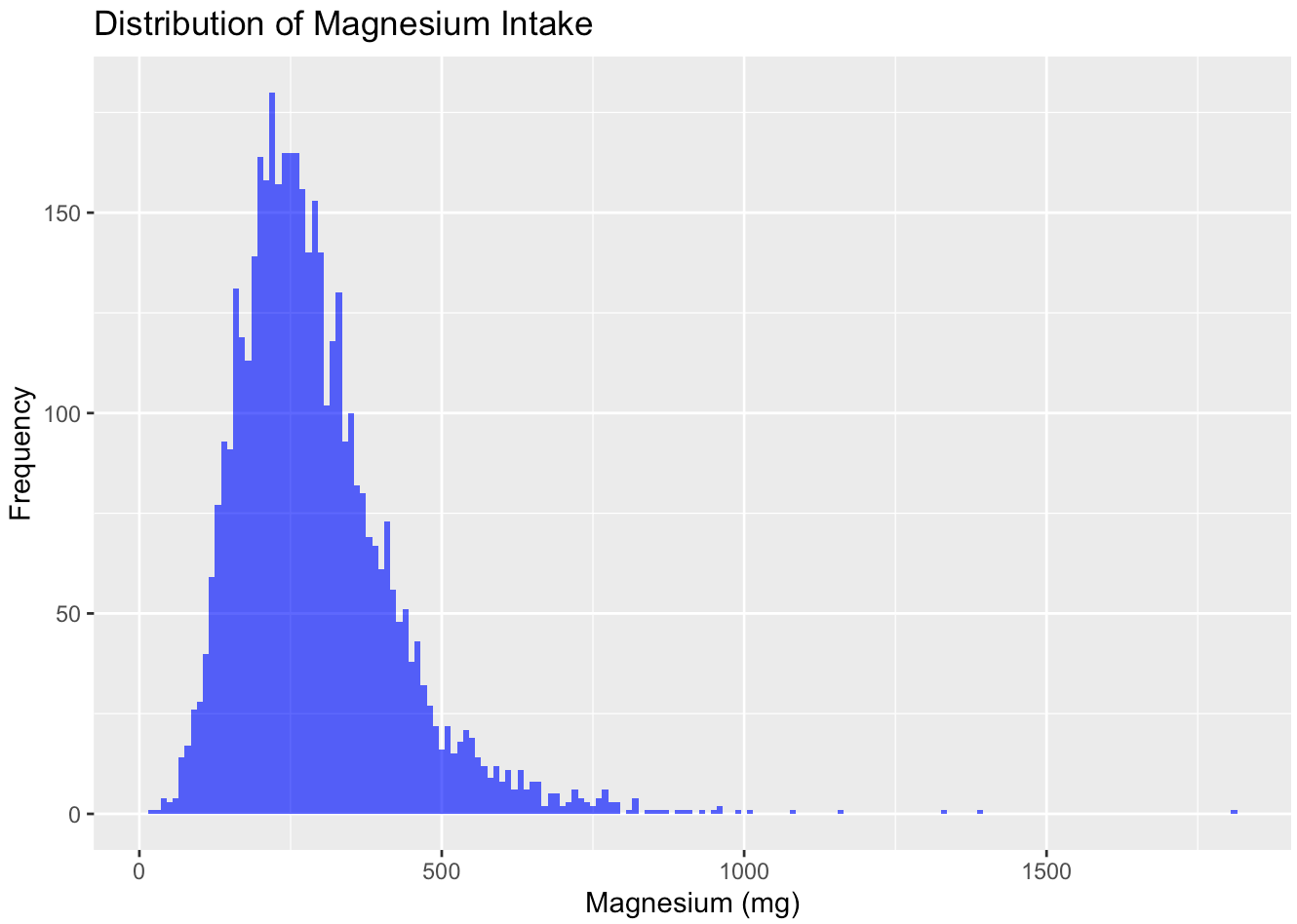

```
# Histograms for nutrient data
ggplot(final_data, aes(x = SODIUM)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Sodium Intake", x = "Sodium (mg)", y = "Frequency")
```



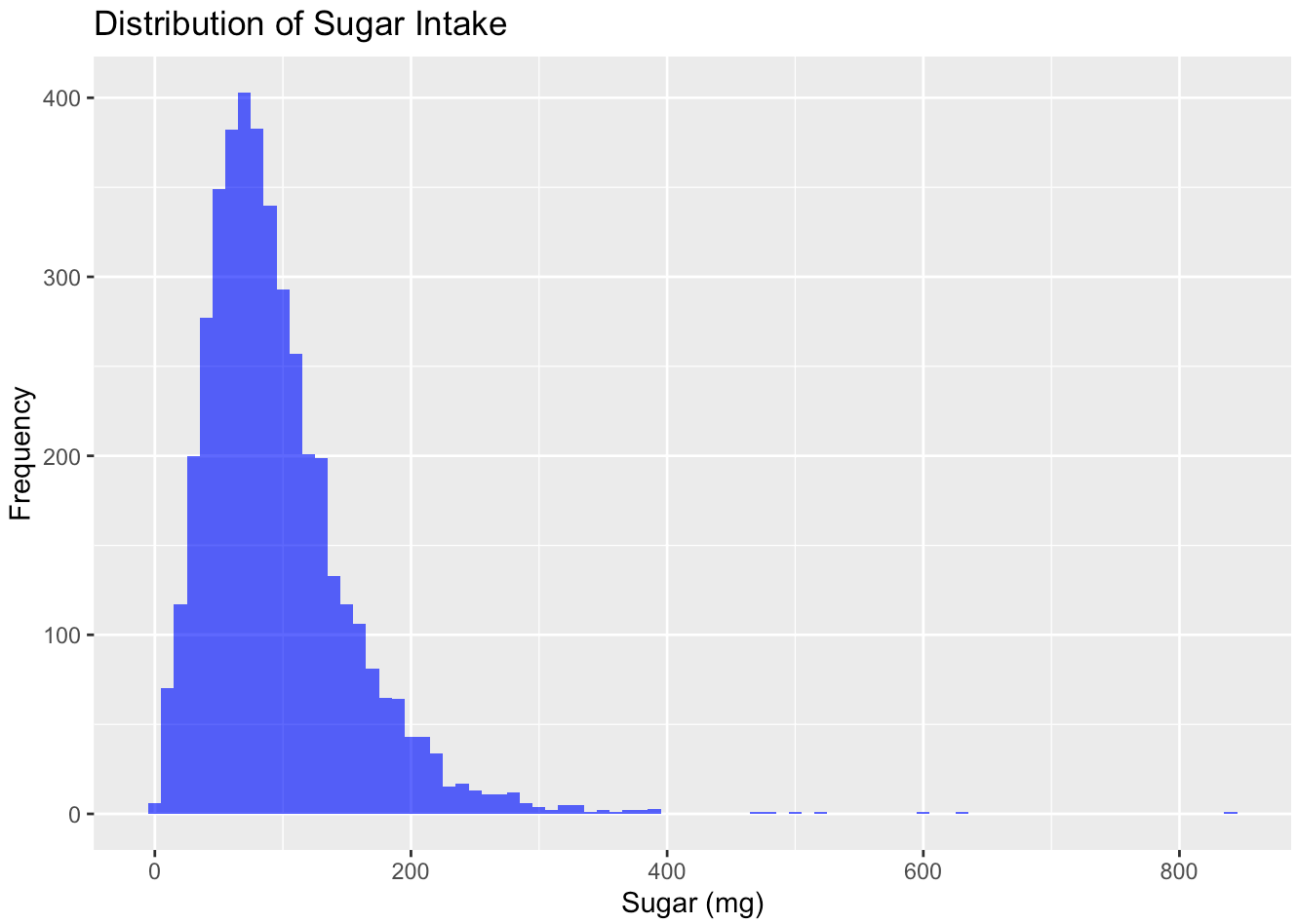

```
ggplot(final_data, aes(x = CALCIUM)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Calcium Intake", x = "Calcium (mg)", y = "Frequency")
```



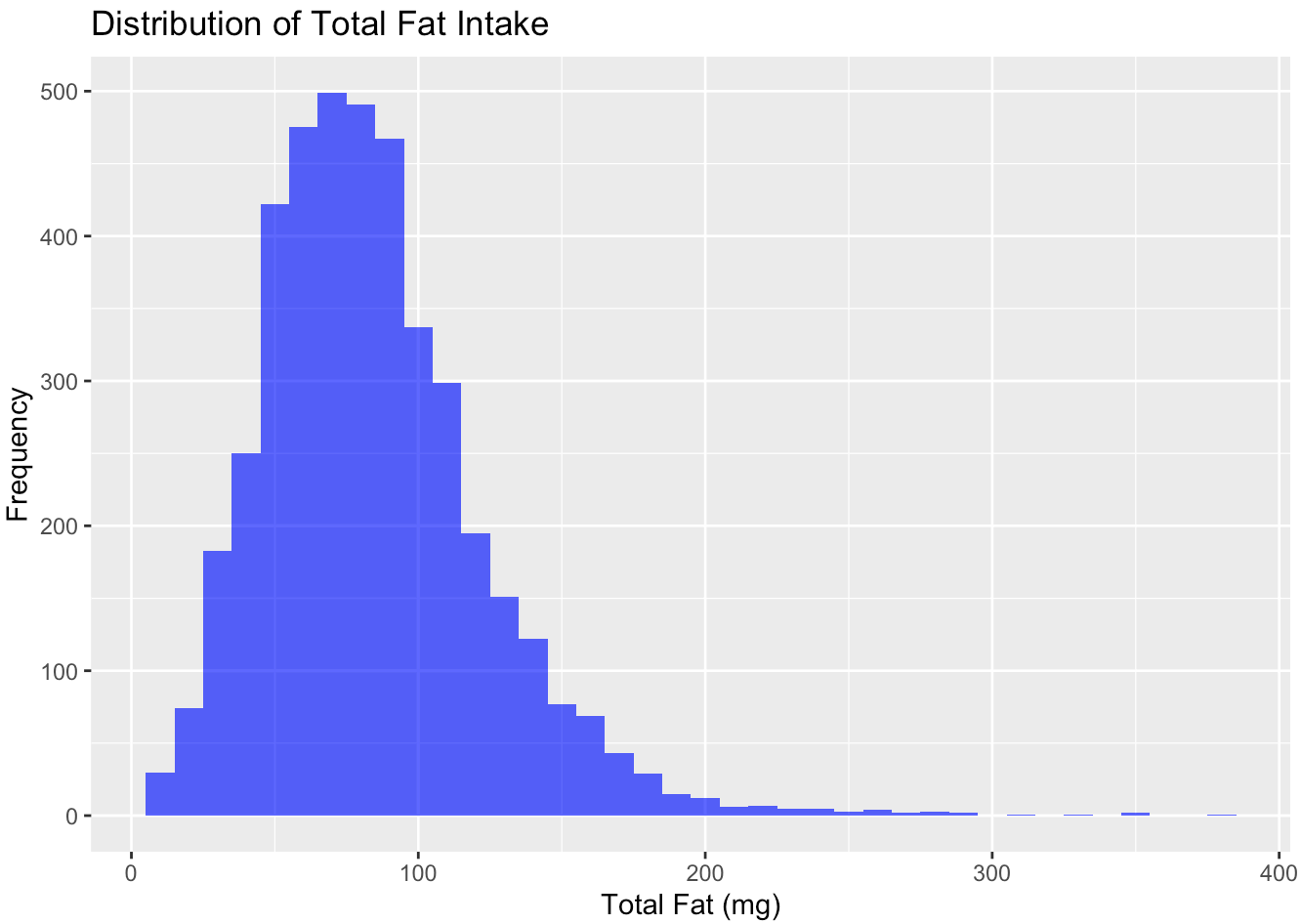

```
ggplot(final_data, aes(x = MAGNESIUM)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Magnesium Intake", x = "Magnesium (mg)", y = "Frequenc
y")
```



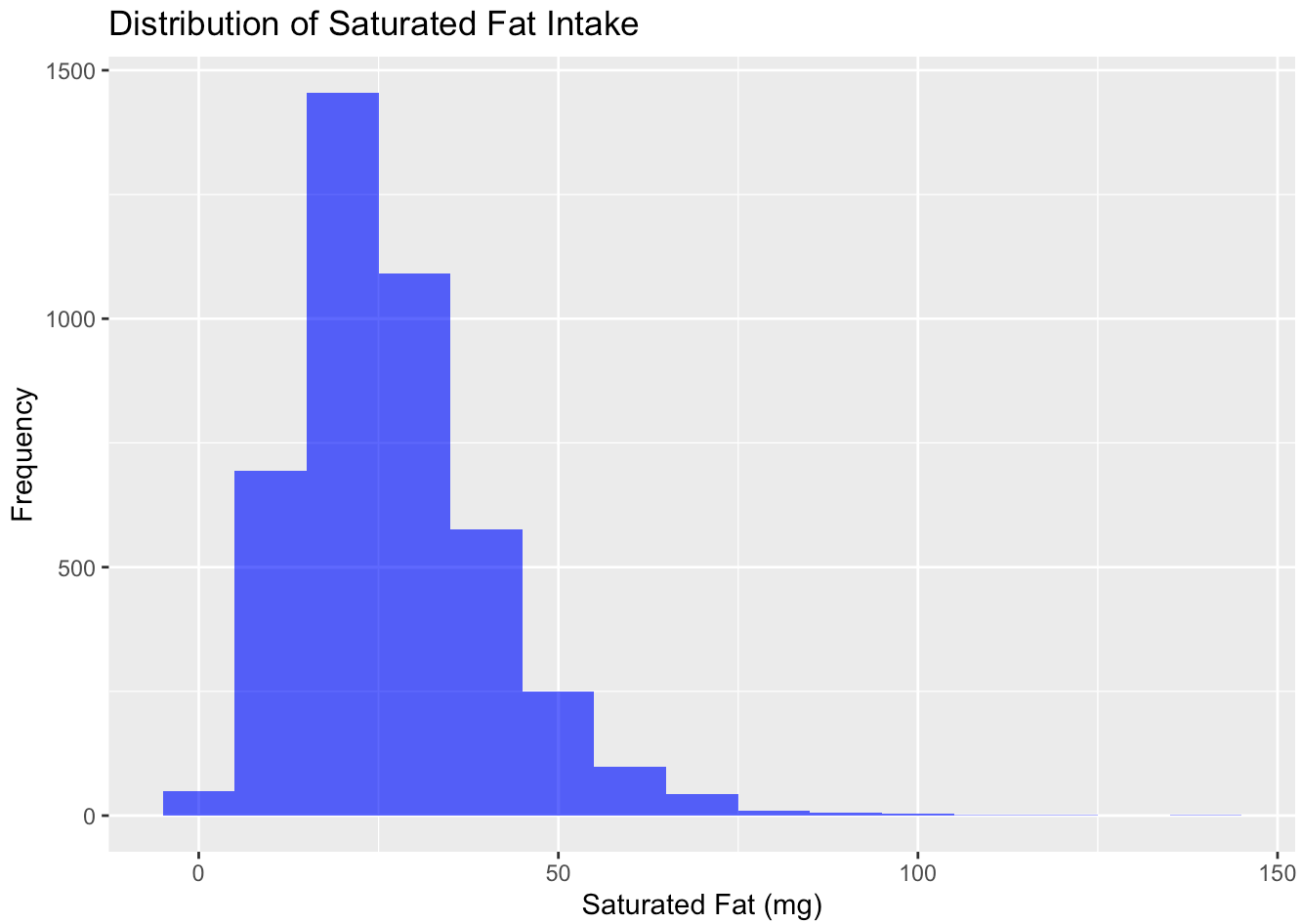

```
ggplot(final_data, aes(x = SUGAR)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Sugar Intake", x = "Sugar (mg)", y = "Frequency")
```



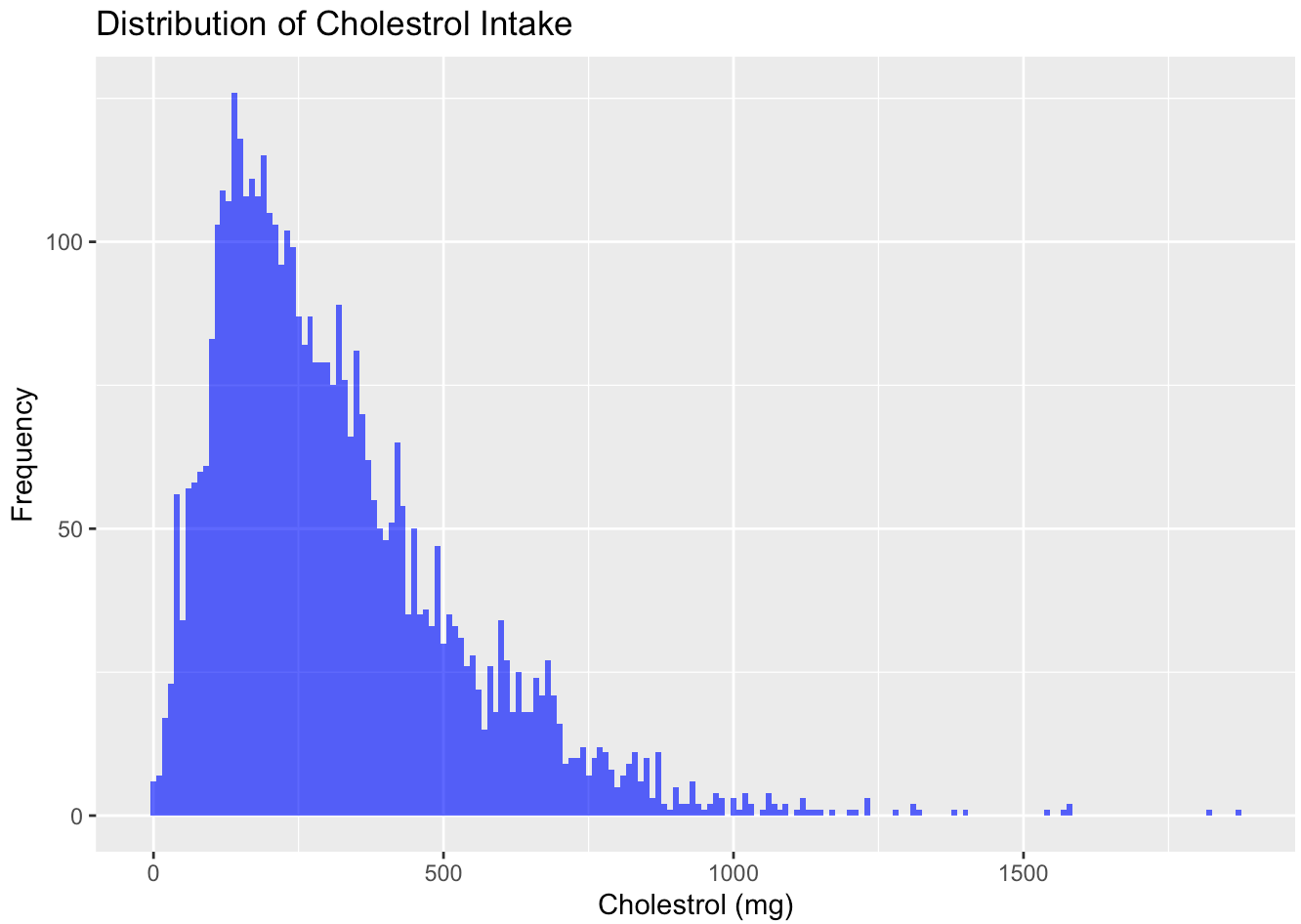

```
ggplot(final_data, aes(x = TFAT)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Total Fat Intake", x = "Total Fat (mg)", y = "Frequenc
y")
```



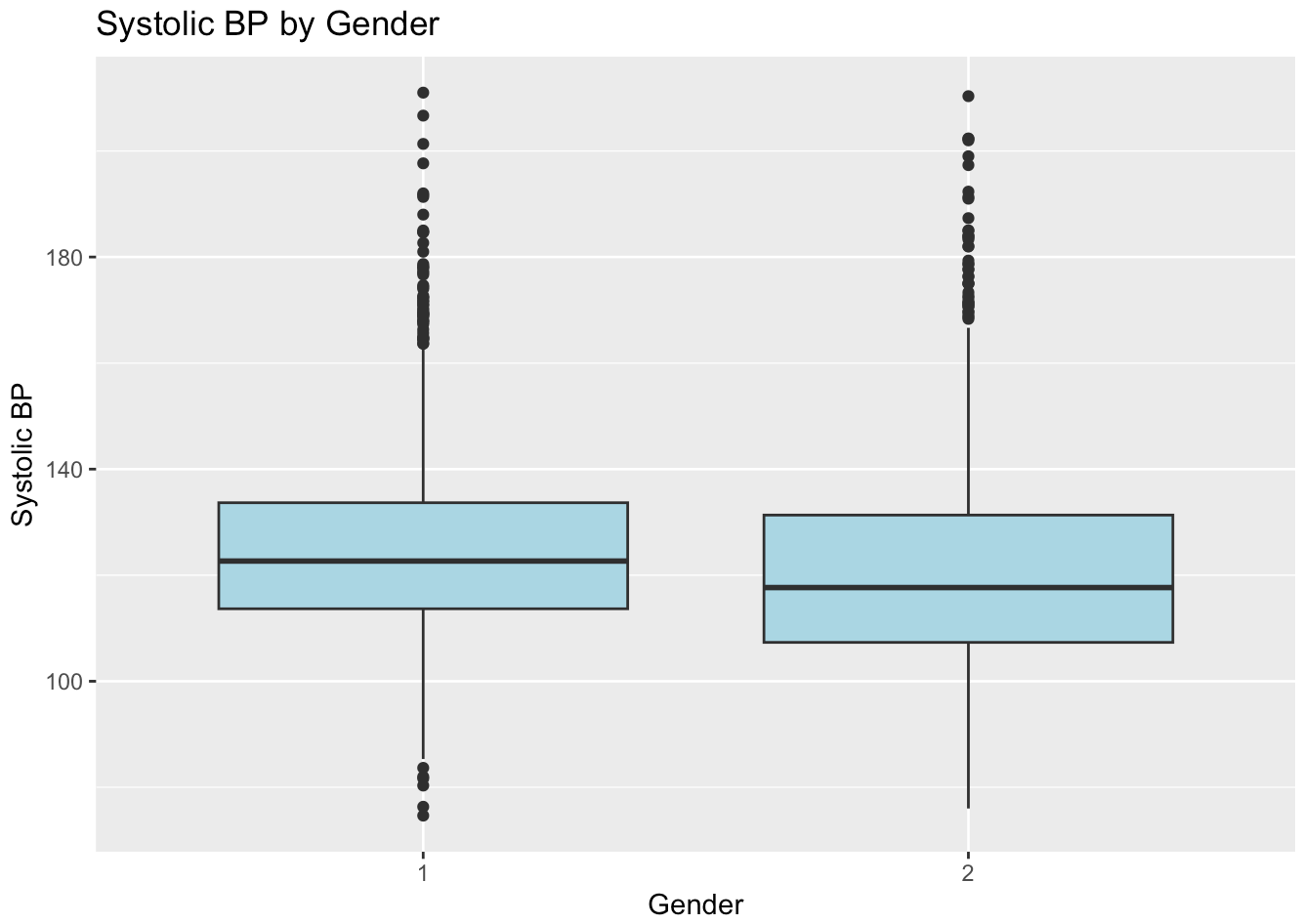

```
ggplot(final_data, aes(x = SFAT)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Saturated Fat Intake", x = "Saturated Fat (mg)", y = "Fr
equency")
```



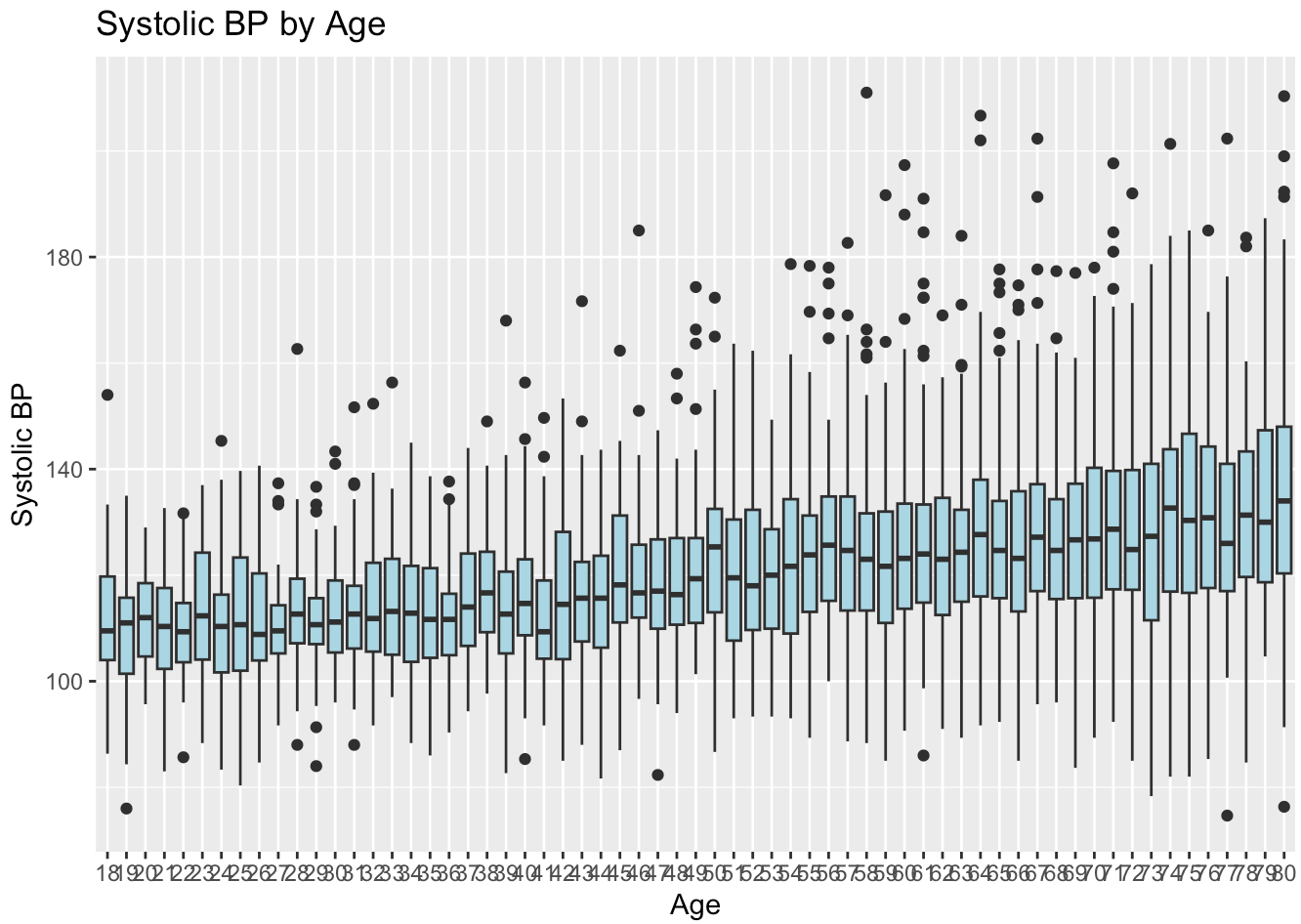

```
ggplot(final_data, aes(x = CHOLESTROL)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Cholestrol Intake", x = "Cholestrol (mg)", y = "Frequenc
y")
```



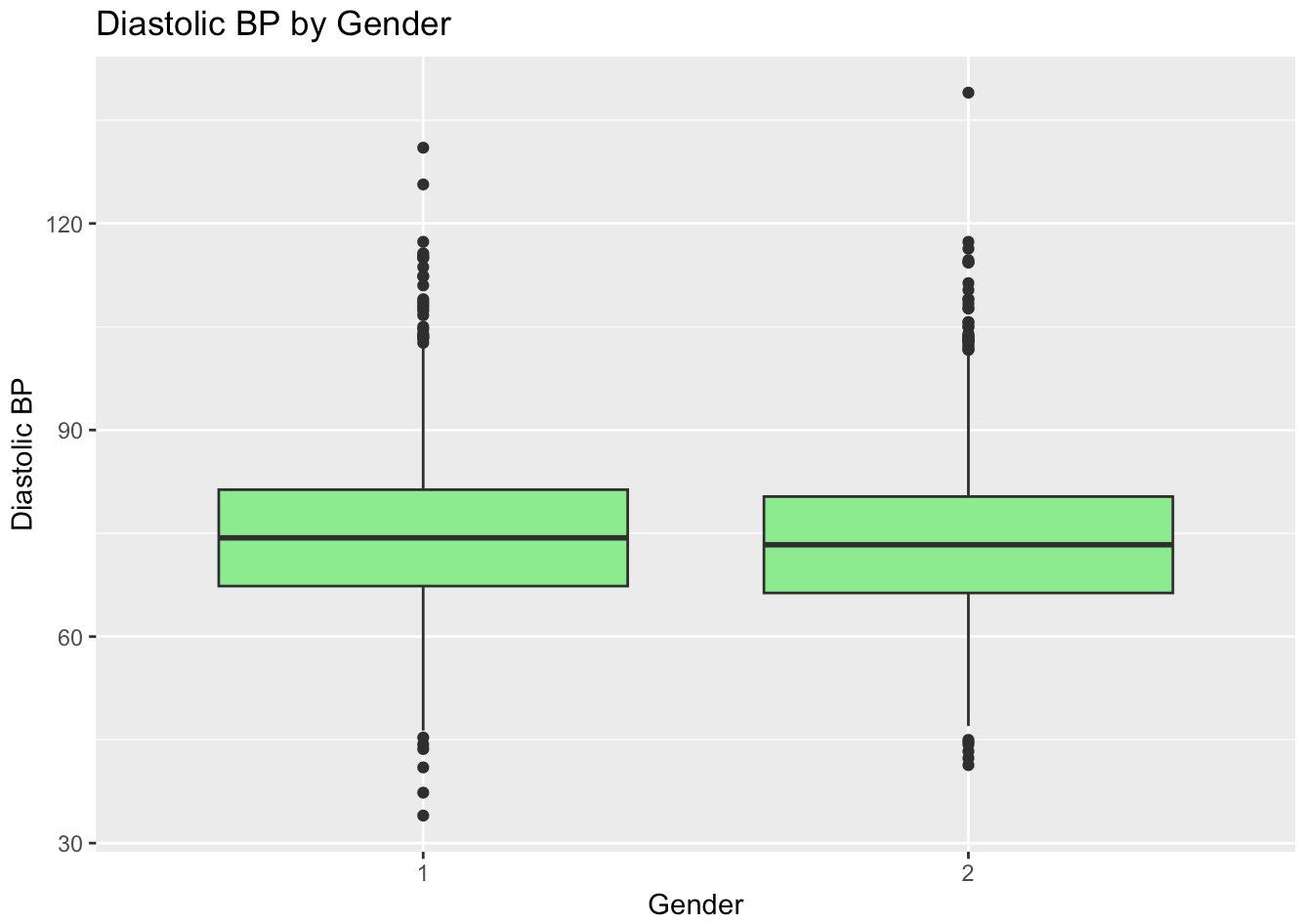

```
# Plots for demographic data over Systolic BP and Diastolic BP
ggplot(final_data, aes(x = as.factor(GENDER), y = SYSTOLIC)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Systolic BP by Gender", x = "Gender", y = "Systolic BP")
```



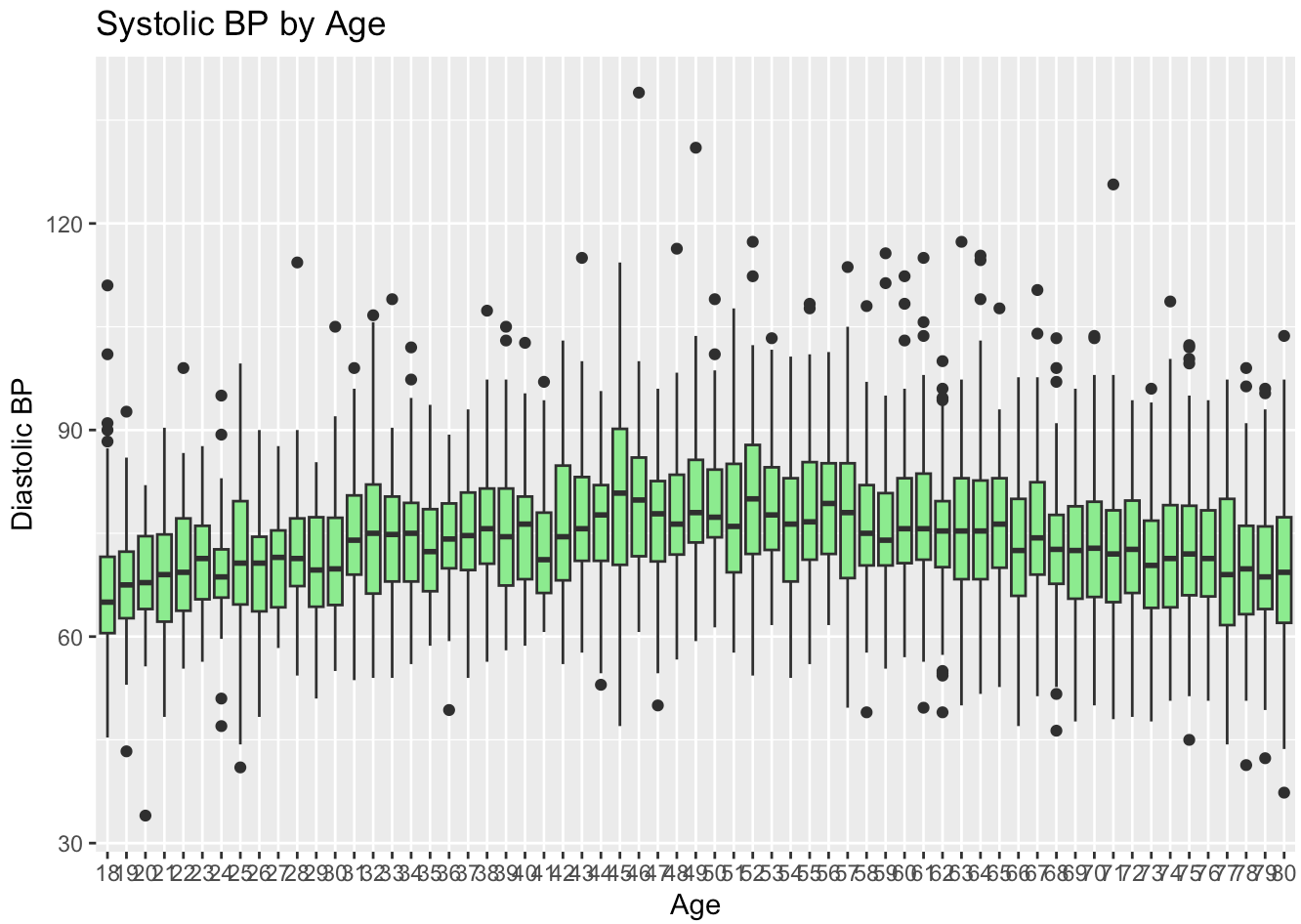

```
ggplot(final_data, aes(x = as.factor(AGE), y = SYSTOLIC)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Systolic BP by Age", x = "Age", y = "Systolic BP")
```




```
ggplot(final_data, aes(x = as.factor(GENDER), y = DIASTOLIC)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Diastolic BP by Gender", x = "Gender", y = "Diastolic BP")
```

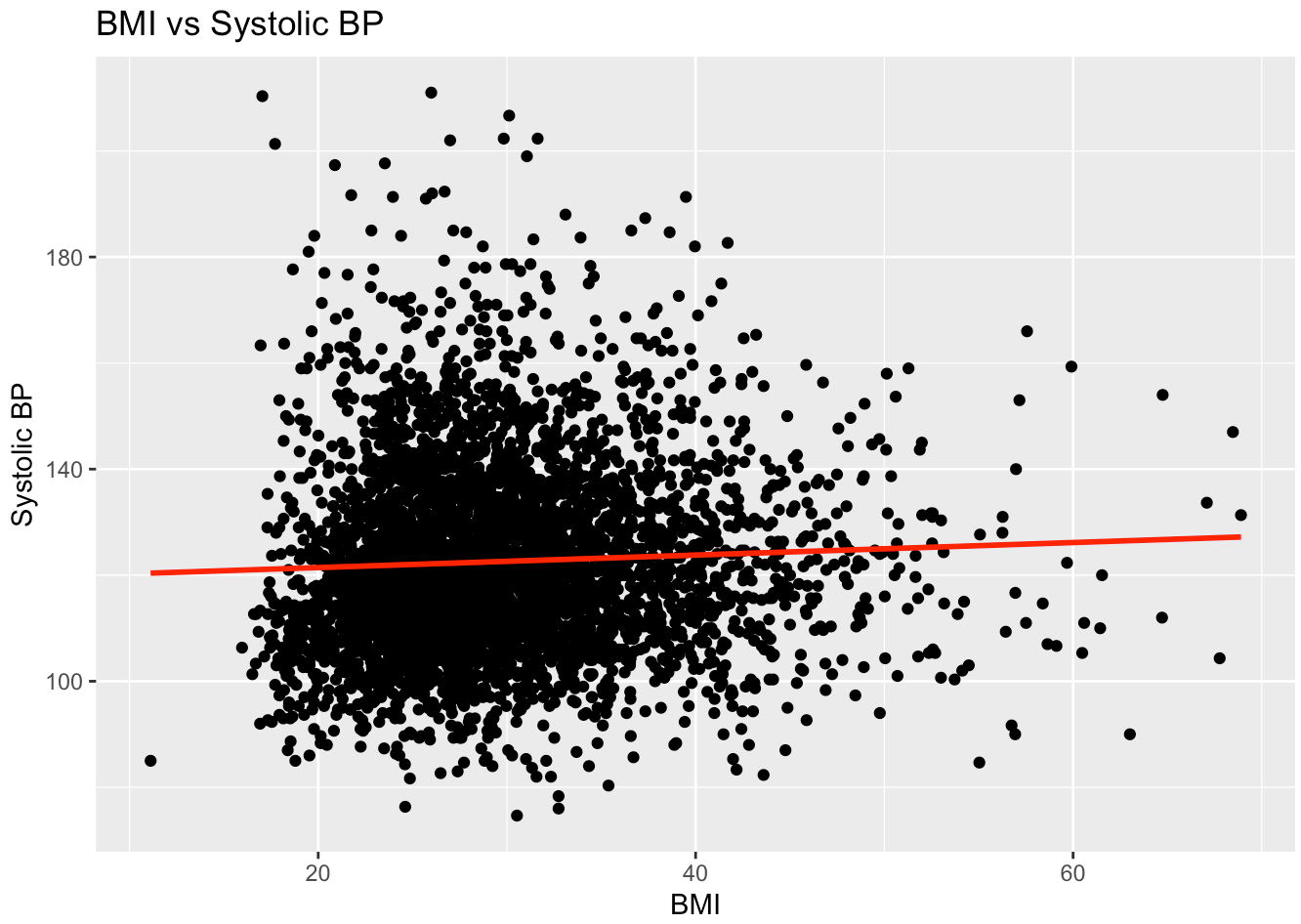



```
ggplot(final_data, aes(x = as.factor(AGE), y = DIASTOLIC)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Systolic BP by Age", x = "Age", y = "Diastolic BP")
```



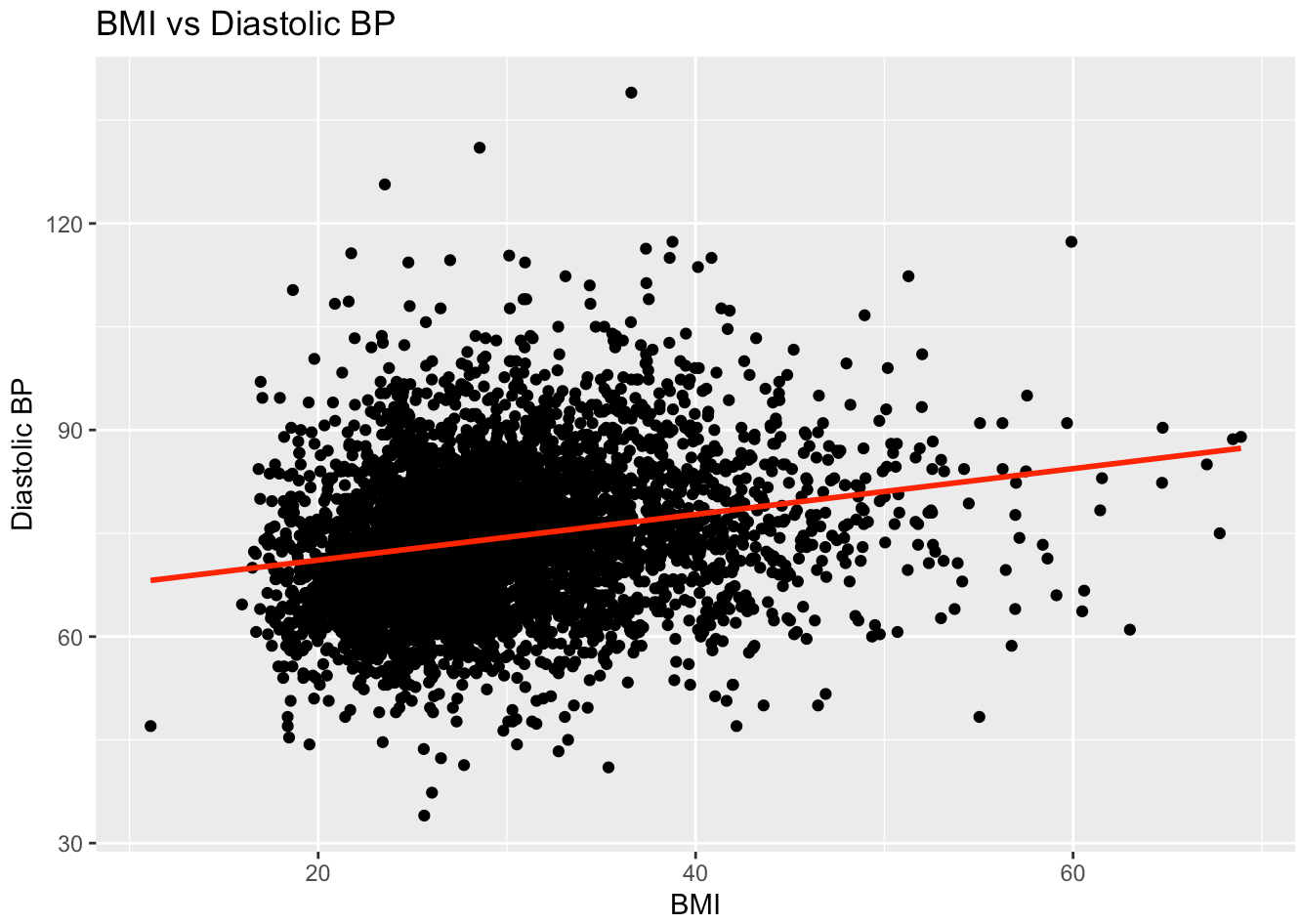

```
ggplot(final_data, aes(x = BMI, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "BMI vs Systolic BP", x = "BMI", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



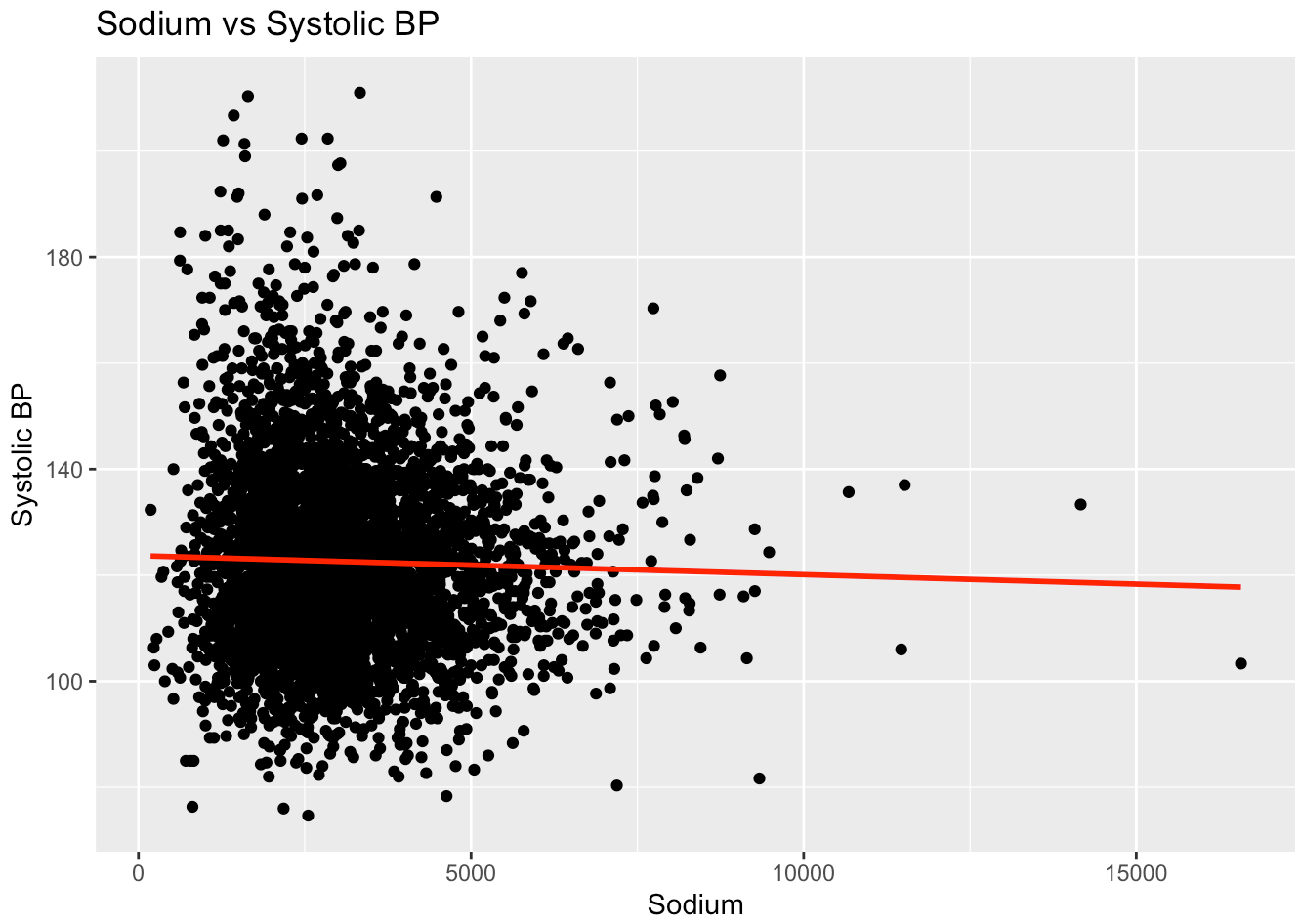

```
ggplot(final_data, aes(x = BMI, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "BMI vs Diastolic BP", x = "BMI", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



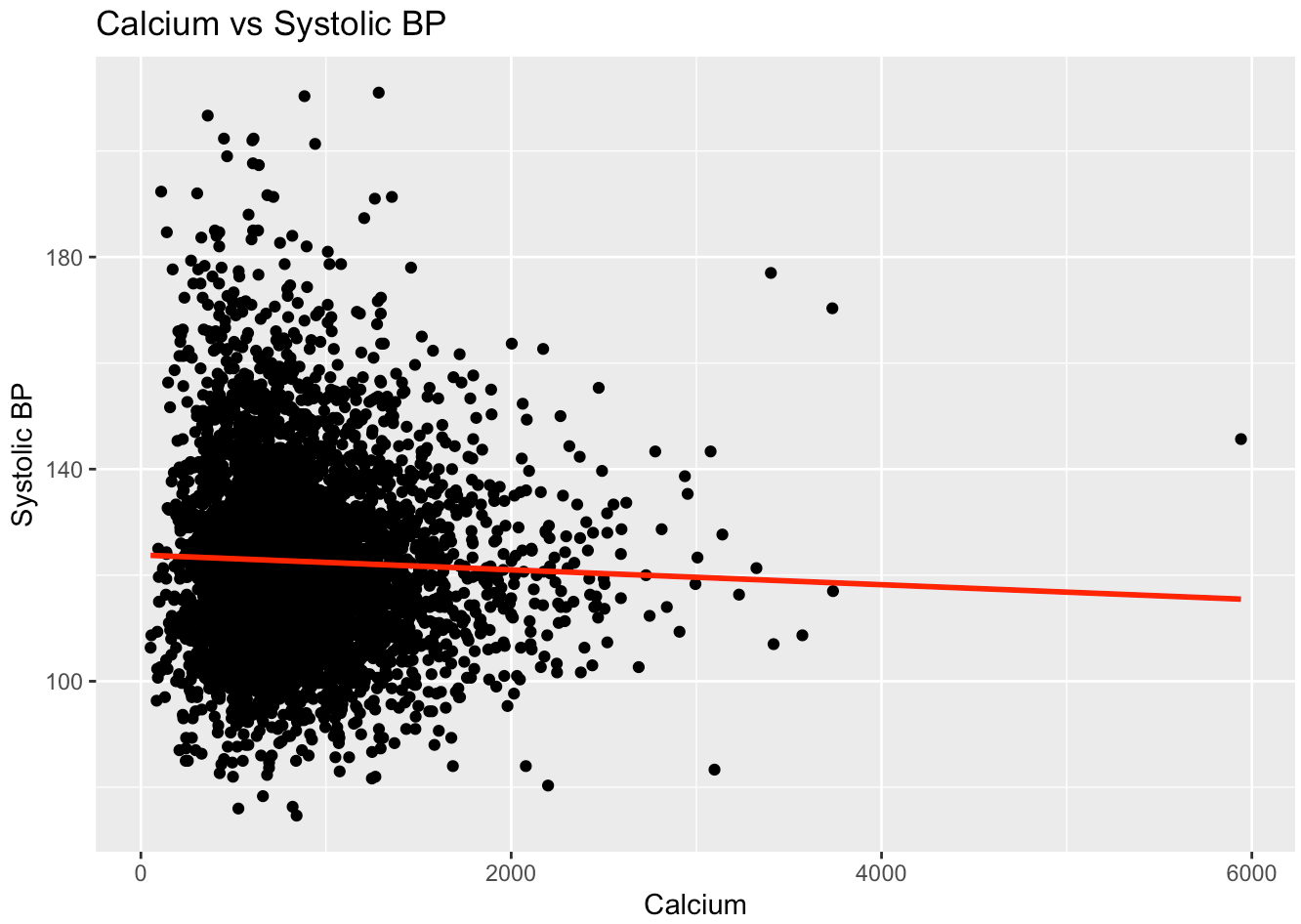

```
#Plots for dietary data over Systolic BP & Diastolic BP
ggplot(final_data, aes(x = SODIUM, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Sodium vs Systolic BP", x = "Sodium", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



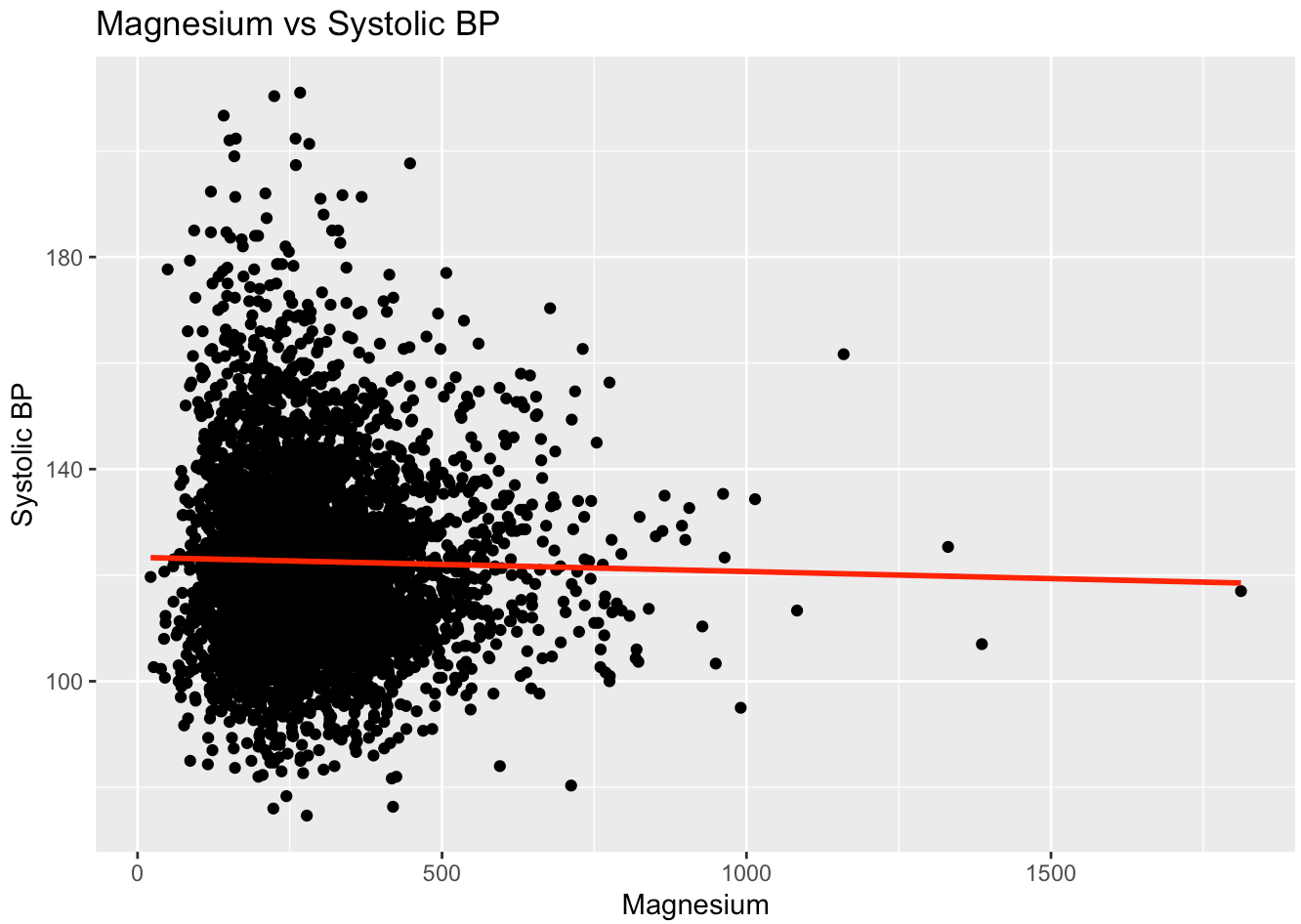

```
ggplot(final_data, aes(x = CALCIUM, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Calcium vs Systolic BP", x = "Calcium", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



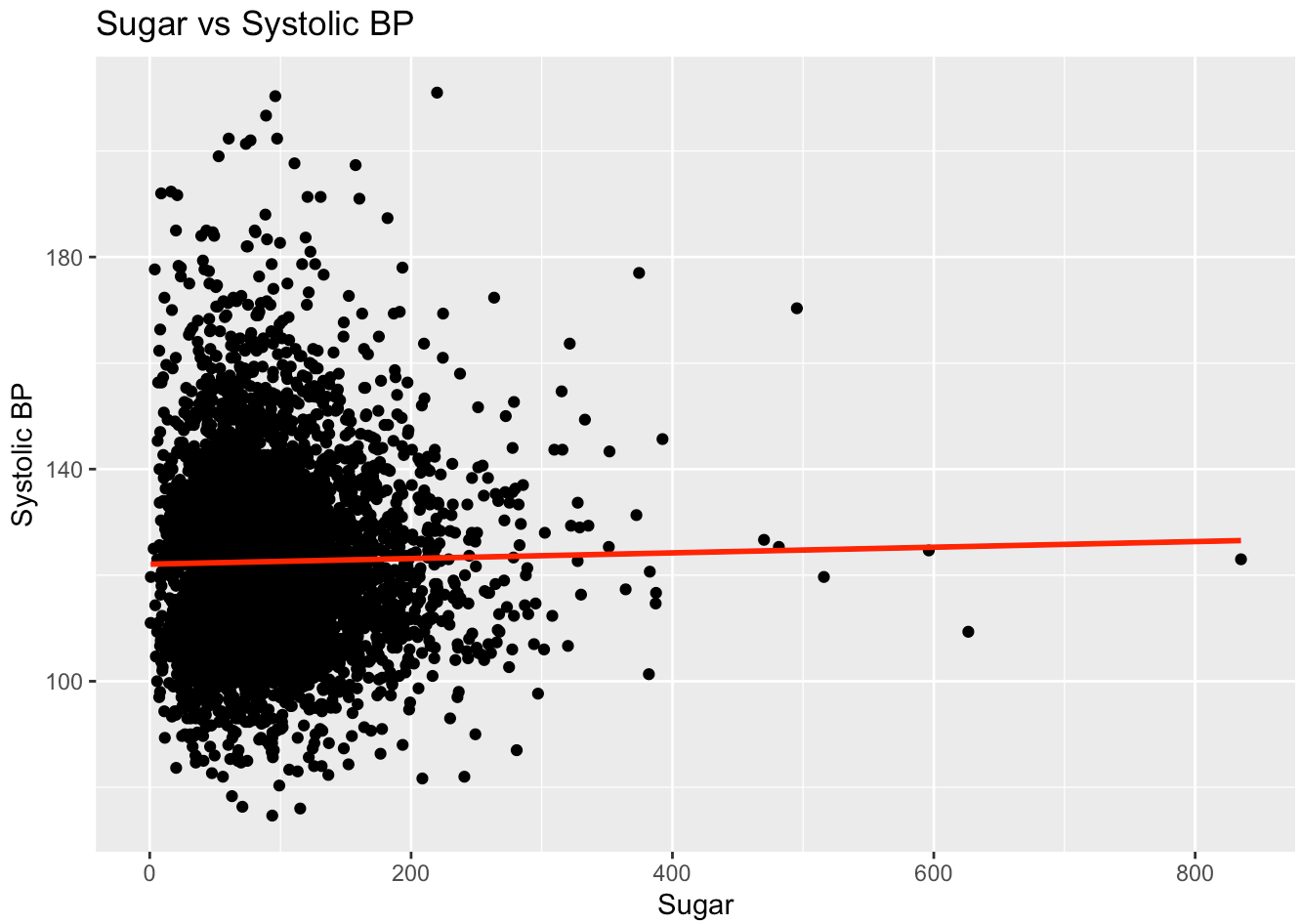

```
ggplot(final_data, aes(x = MAGNESIUM, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Magnesium vs Systolic BP", x = "Magnesium", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



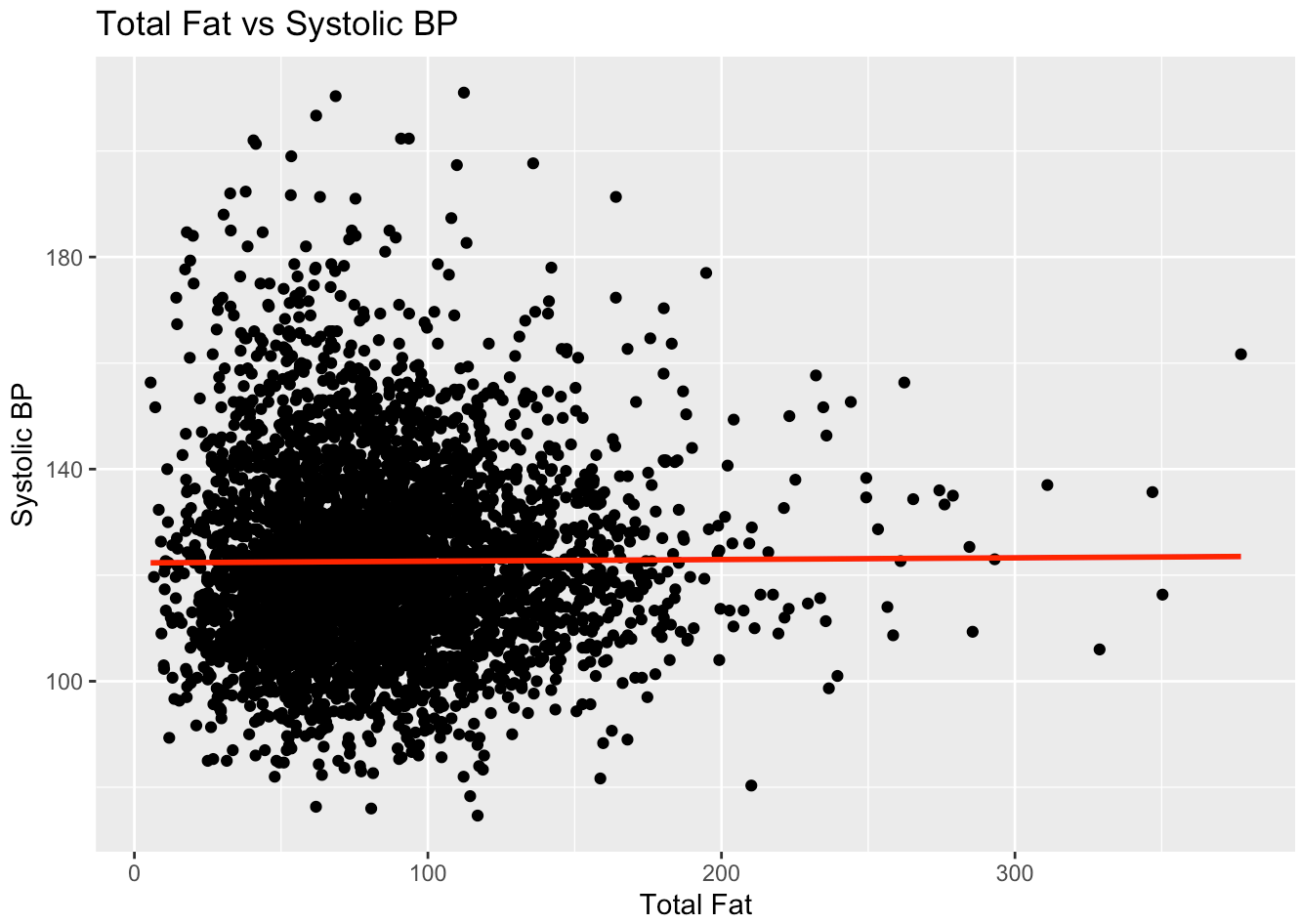

```
ggplot(final_data, aes(x = SUGAR, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Sugar vs Systolic BP", x = "Sugar", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



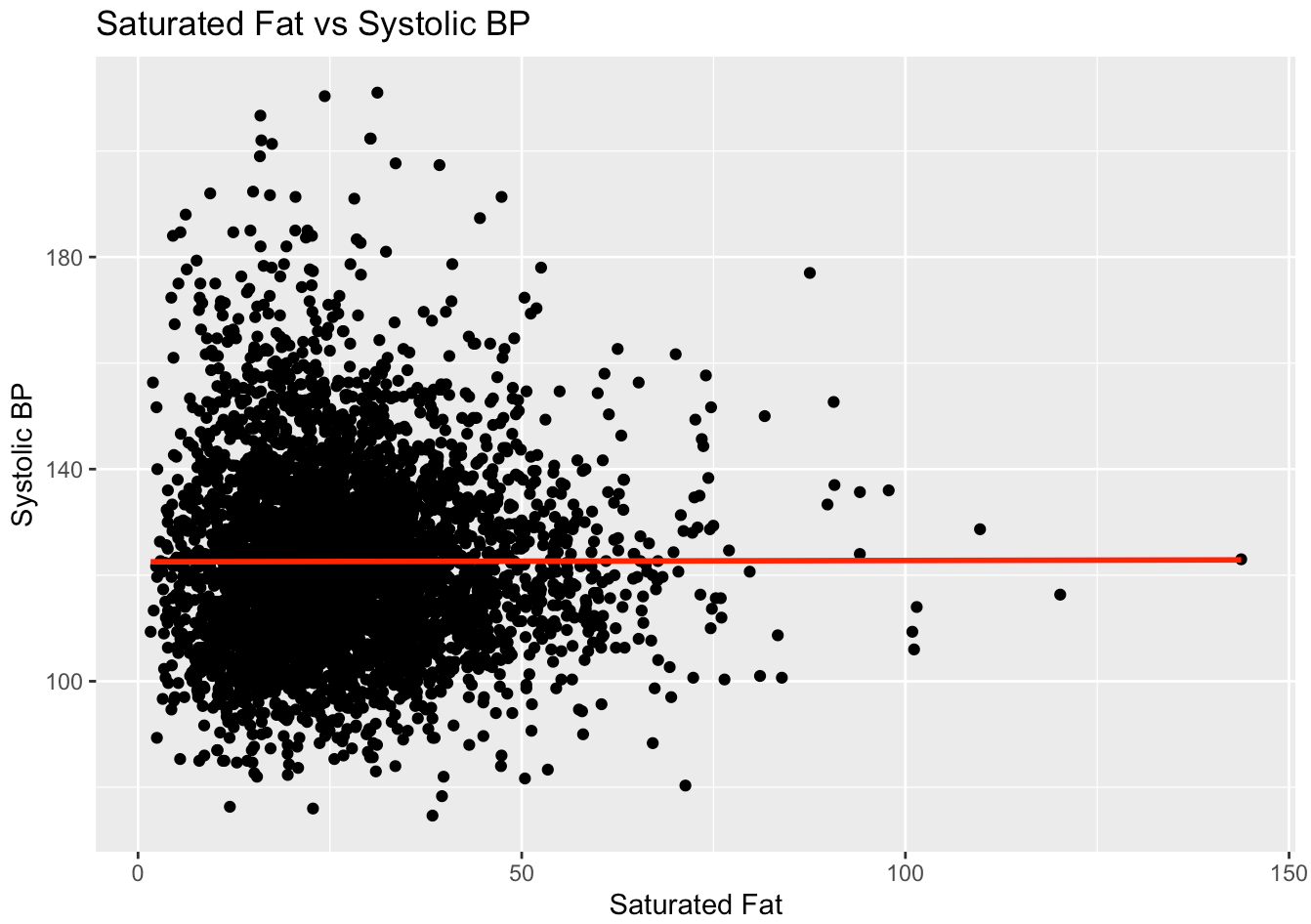

```
ggplot(final_data, aes(x = TFAT, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Total Fat vs Systolic BP", x = "Total Fat", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



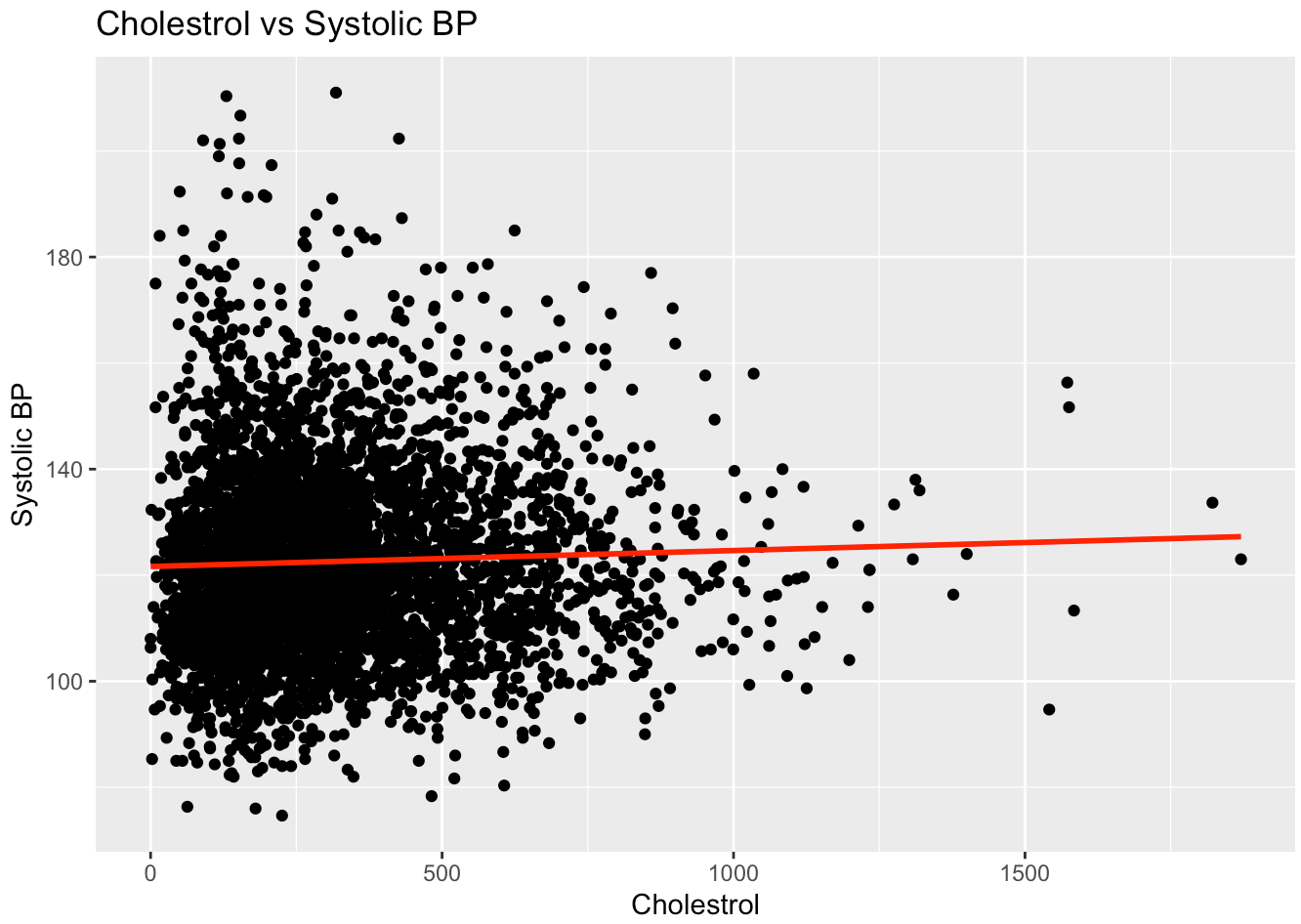

```
ggplot(final_data, aes(x = SFAT, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Saturated Fat vs Systolic BP", x = "Saturated Fat", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



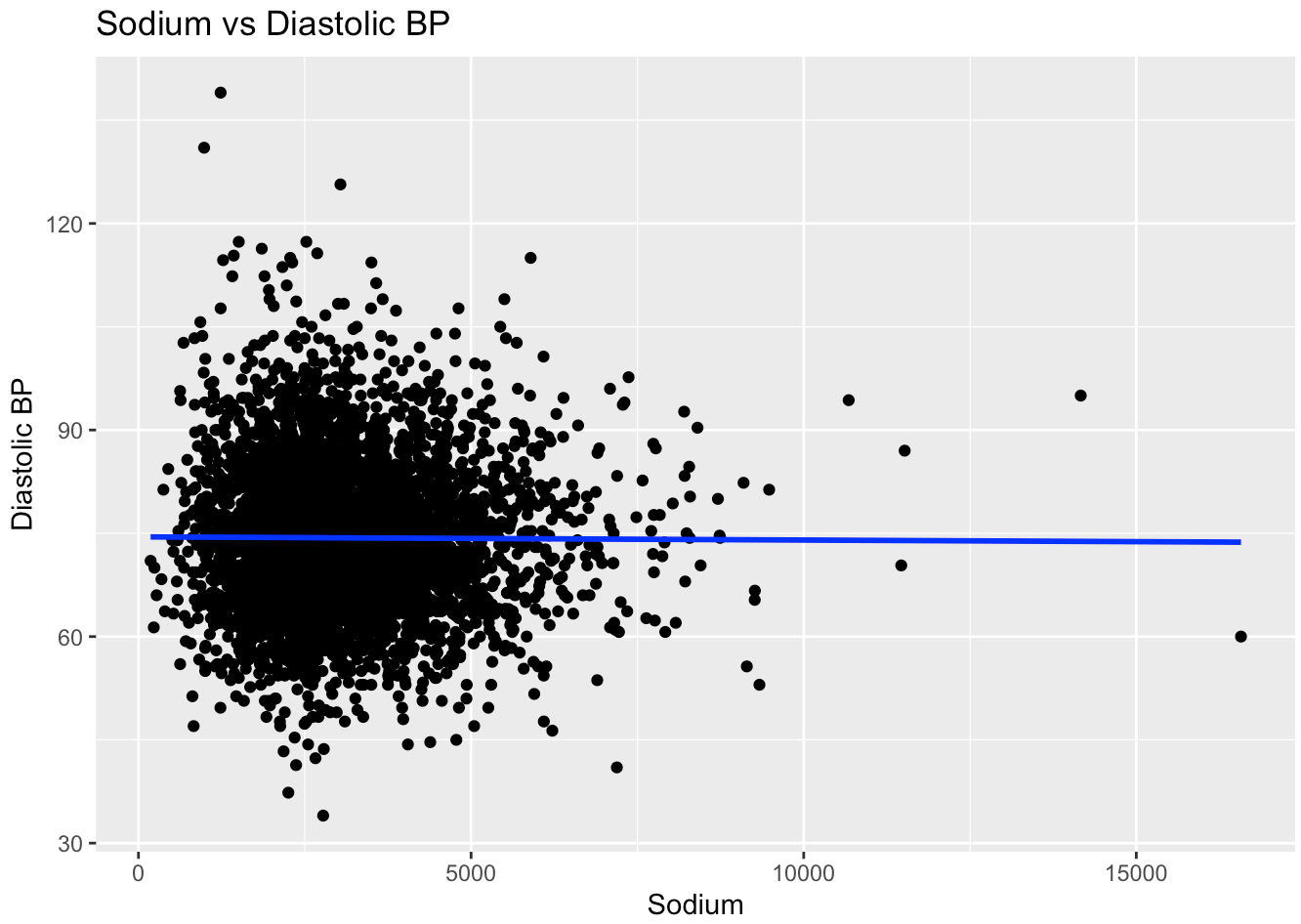

```
ggplot(final_data, aes(x = CHOLESTROL, y = SYSTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Cholestrol vs Systolic BP", x = "Cholestrol", y = "Systolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



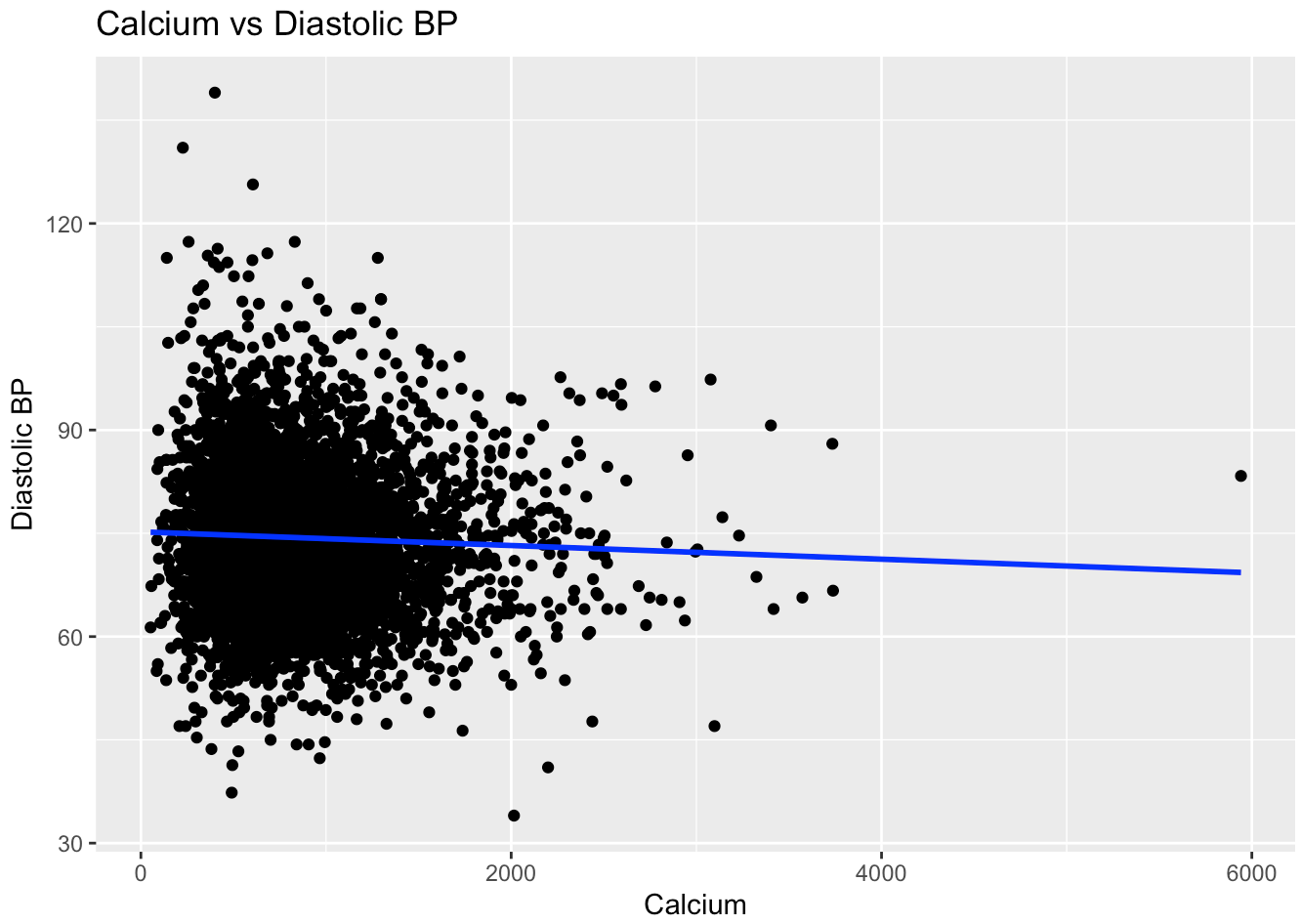

```
ggplot(final_data, aes(x = SODIUM, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Sodium vs Diastolic BP", x = "Sodium", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



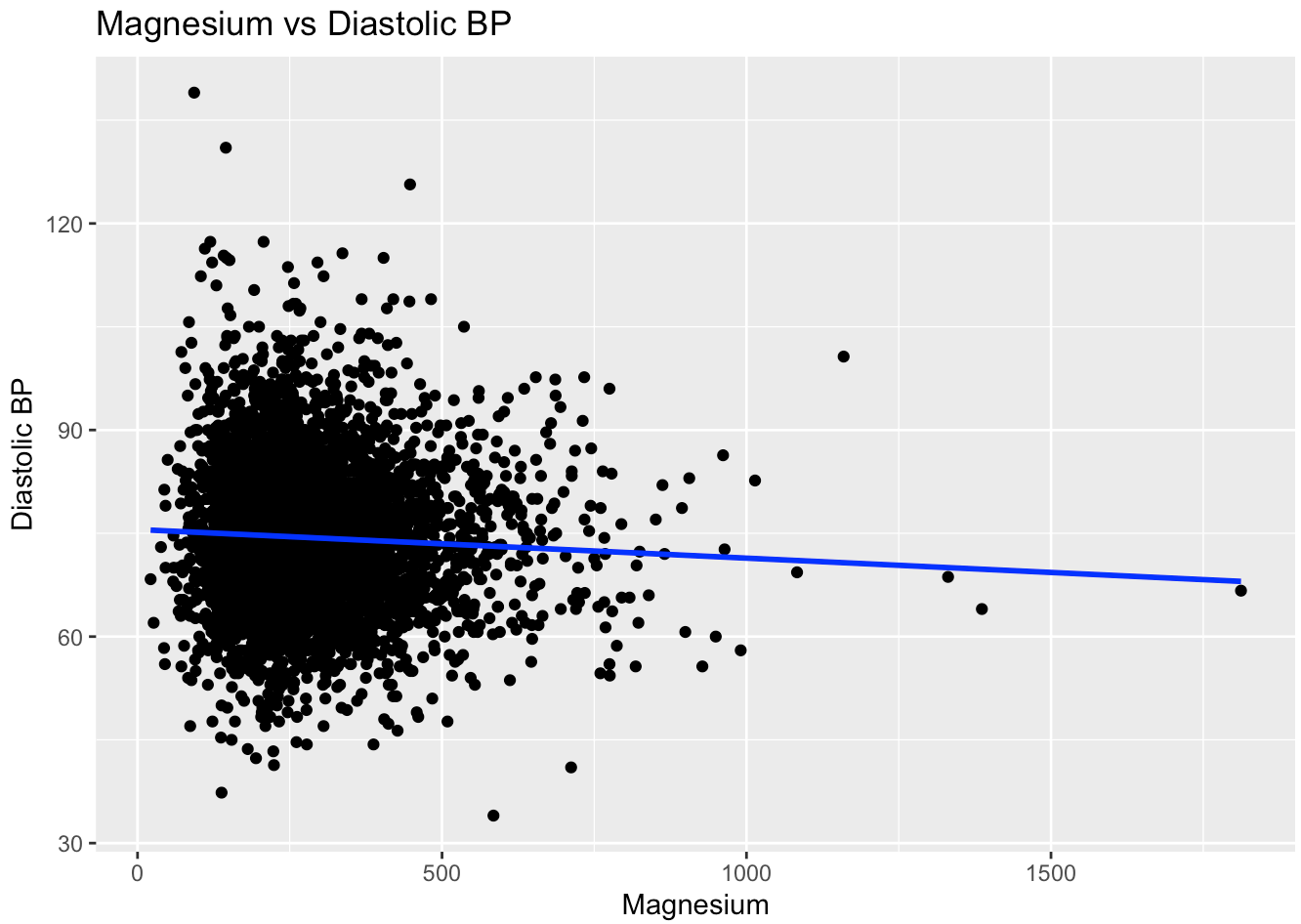

```
ggplot(final_data, aes(x = CALCIUM, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Calcium vs Diastolic BP", x = "Calcium", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



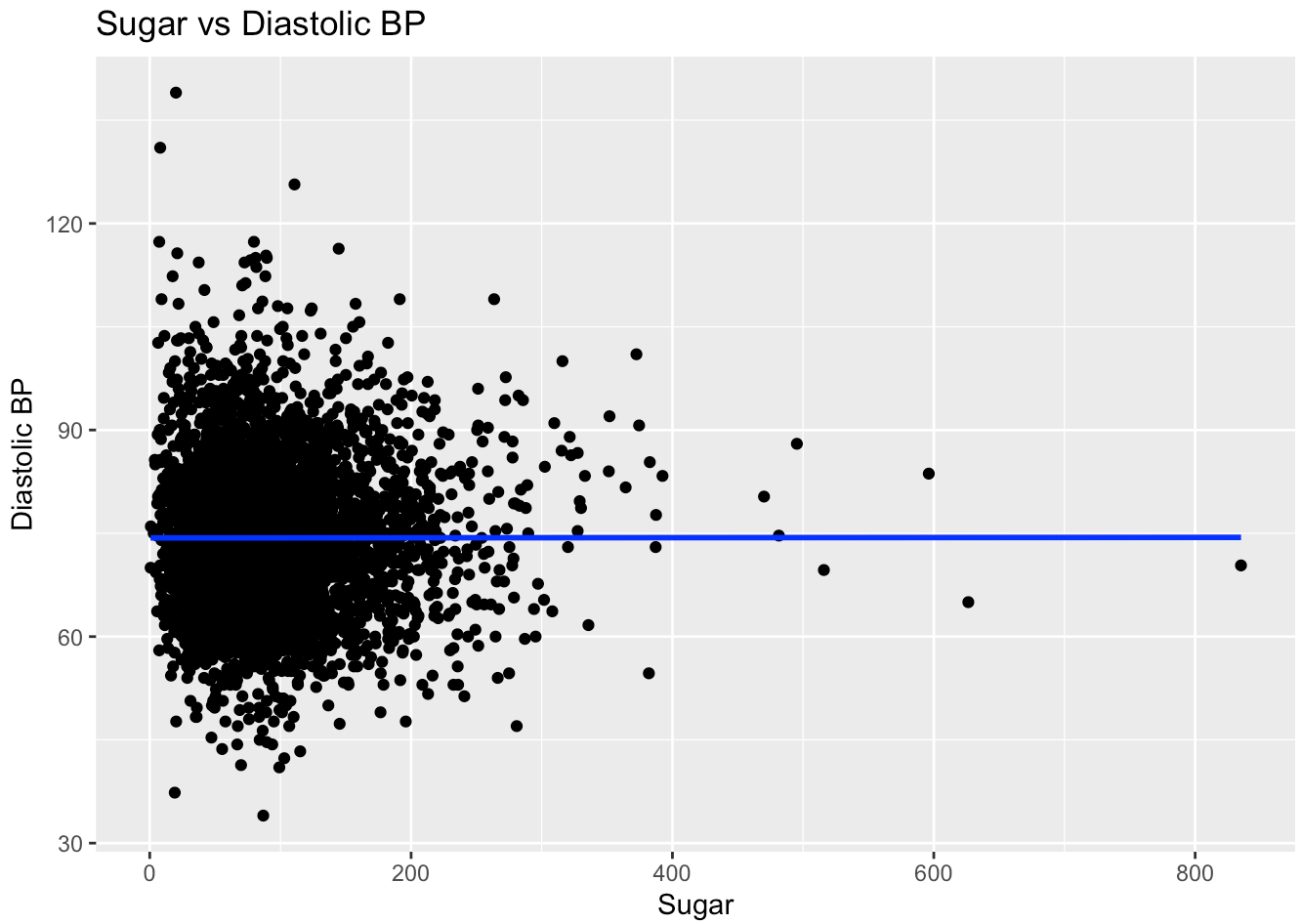

```
ggplot(final_data, aes(x = MAGNESIUM, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Magnesium vs Diastolic BP", x = "Magnesium", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



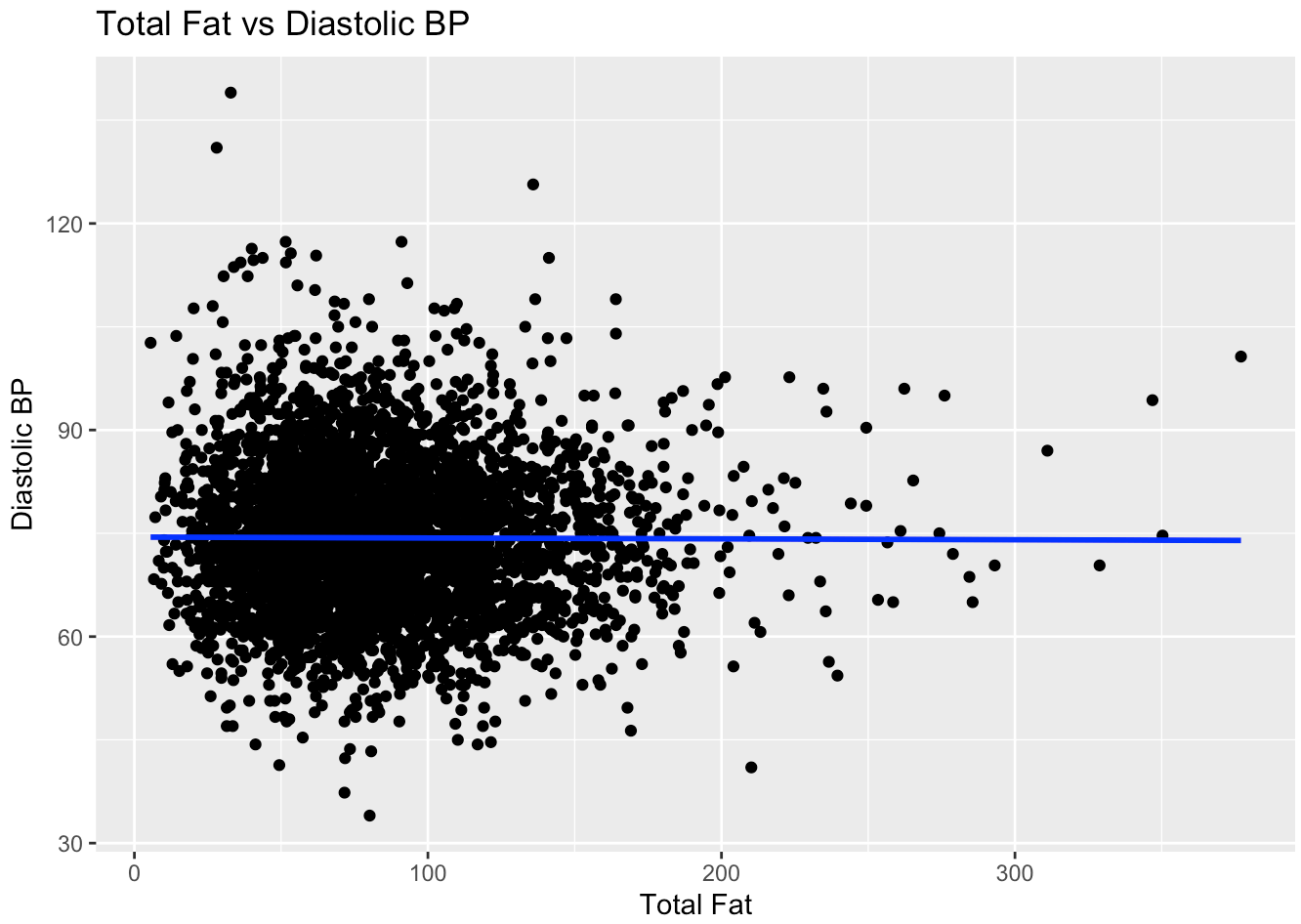

```
ggplot(final_data, aes(x = SUGAR, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Sugar vs Diastolic BP", x = "Sugar", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



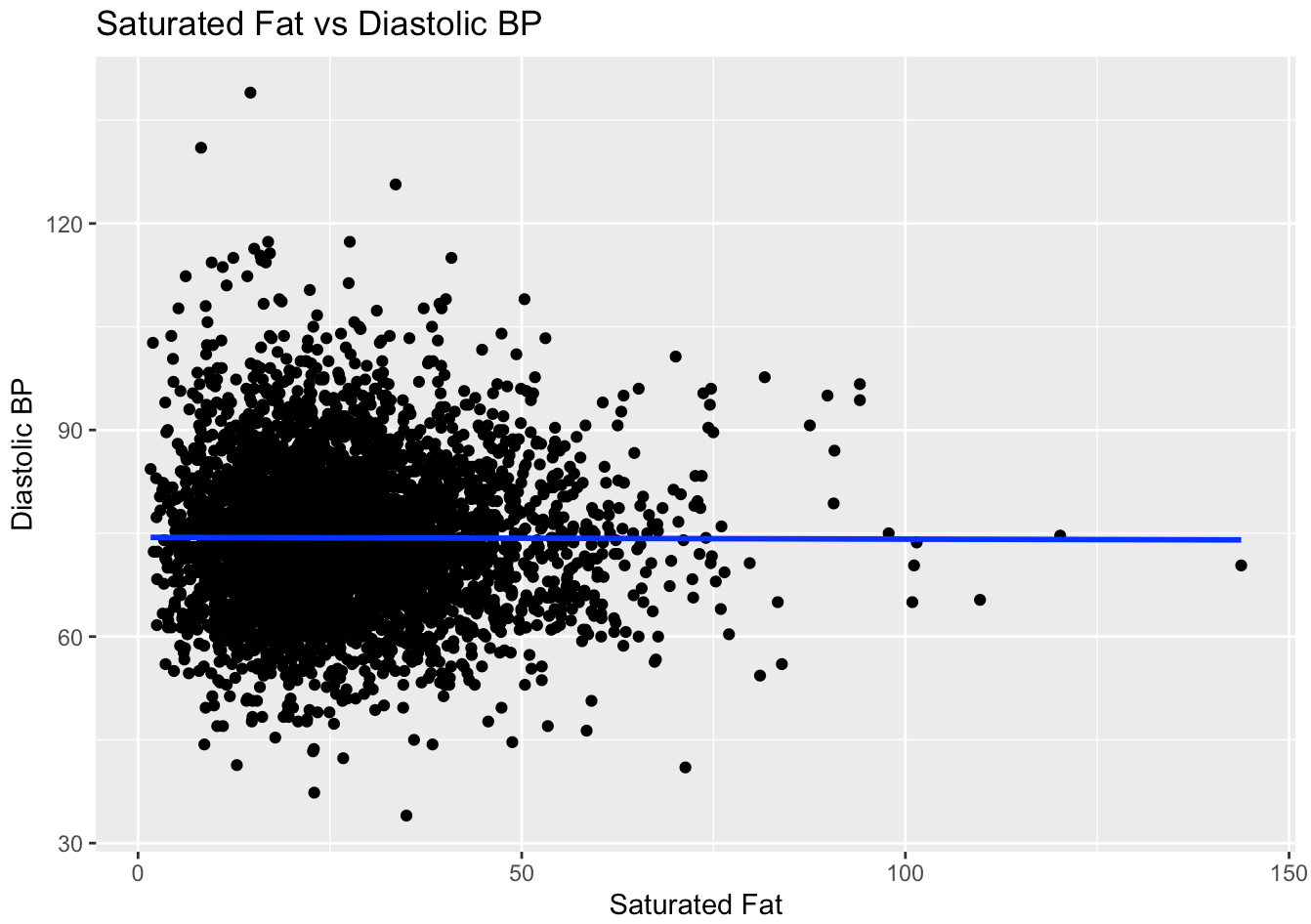

```
ggplot(final_data, aes(x = TFAT, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Total Fat vs Diastolic BP", x = "Total Fat", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



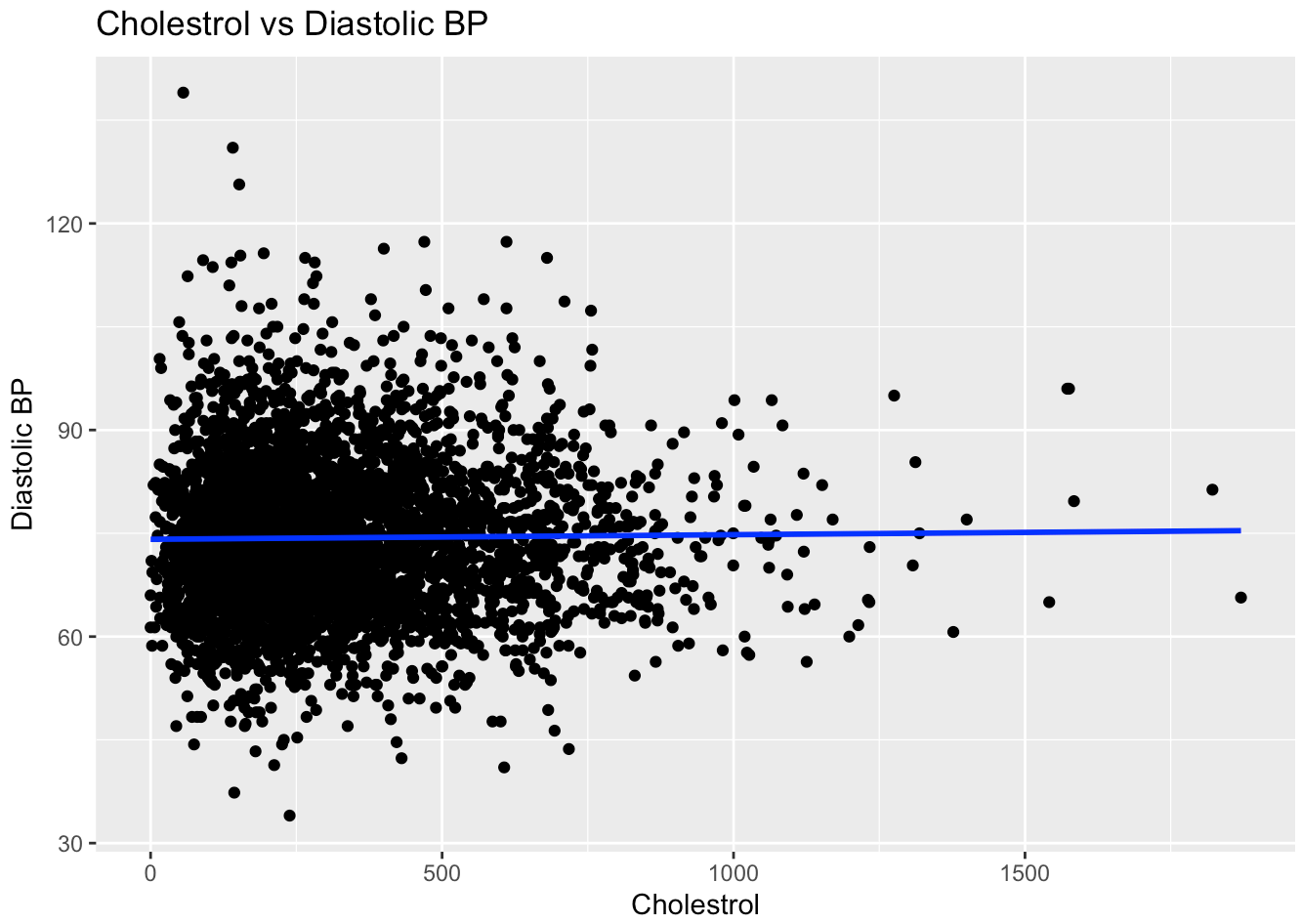

```
ggplot(final_data, aes(x = SFAT, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Saturated Fat vs Diastolic BP", x = "Saturated Fat", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(final_data, aes(x = CHOLESTROL, y = DIASTOLIC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Cholestrol vs Diastolic BP", x = "Cholestrol", y = "Diastolic BP")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Insights

- With plotted graphs we were able to visualize the effects of nutrients on Systolic BP and Diastolic BP. The scatter plot plotted between nutrients and BP helped in understanding the relation between both the factors
- From the box plots plotted between Age and Systolic BP, conveys that with age the chances of having higher Systolic BP are more. Whereas for Diastolic BP middle aged people tend to show higher values.

Correlation Analysis

In this analysis we have implemented correlation matrix, correlation plot and Pairwise correlation tests.

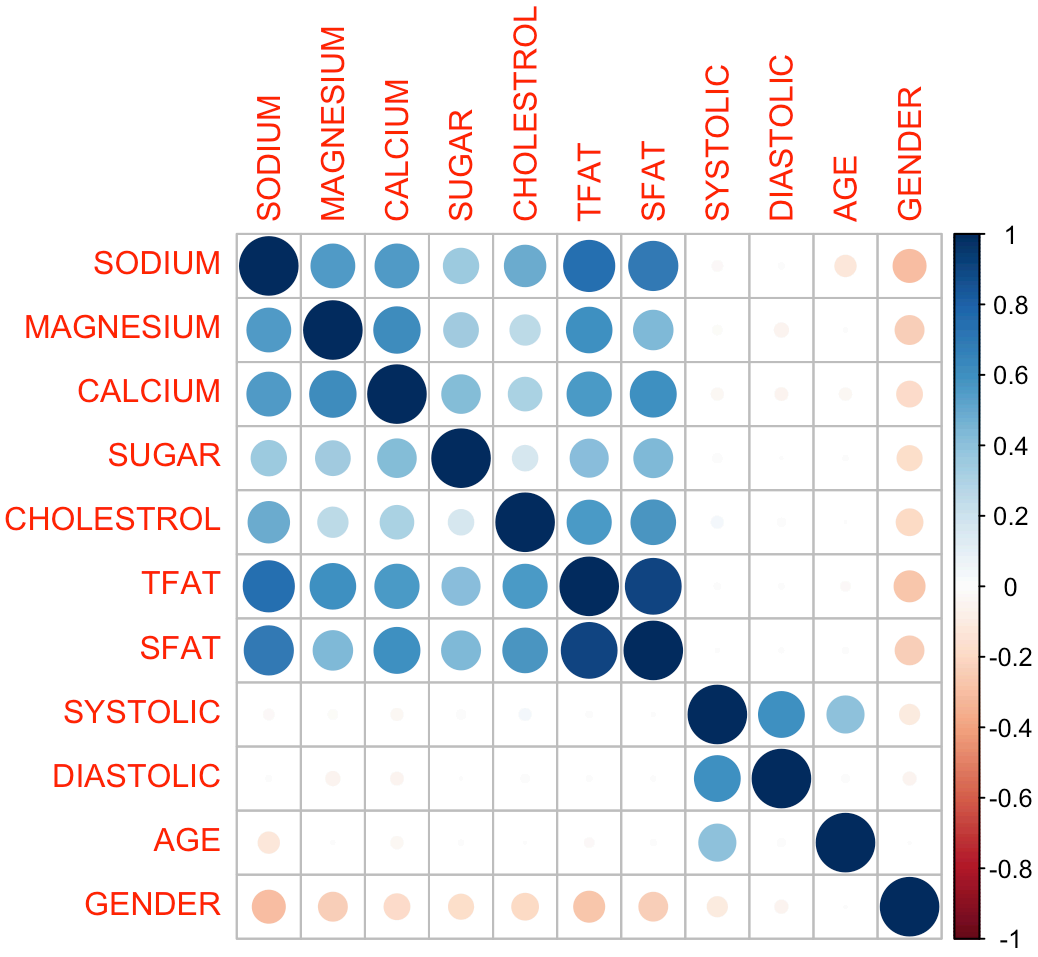
Correlation matrix & Correlation plot

It creates a matrix and plot depicting the strength and direction of the linear relationship between all pairs of following variables :

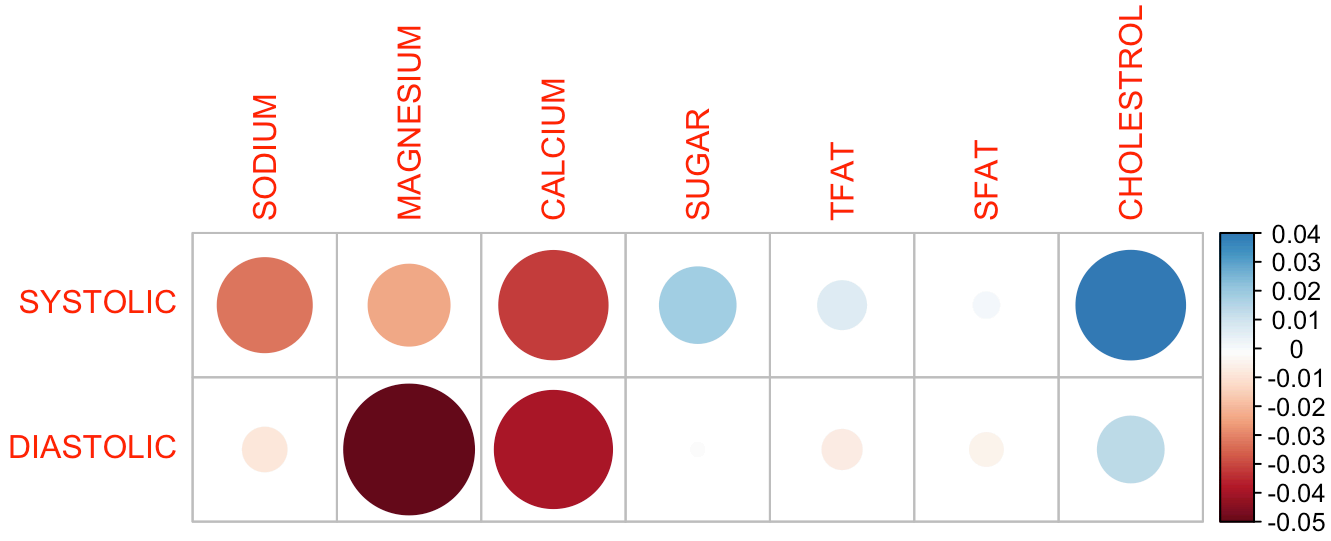
- BP Systolic
- BP Diastolic
- Sodium
- Magnesium
- Calcium

- Sugar
- TFat
- SFat
- Cholestrol
- Age
- Gender

```
## CORRELATION ANALYSIS

# Correlation Matrix
cor_matrix <- cor(final_data %>% select(SODIUM, MAGNESIUM, CALCIUM, SUGAR, CHOLESTROL, T
FAT, SFAT, SYSTOLIC, DIASTOLIC, AGE, GENDER))
corrplot(cor_matrix, method = "circle")
```




```
# Rows: systolic and diastolic, Columns: independent variables
subset_cor <- cor_matrix[c("SYSTOLIC", "DIASTOLIC"), c("SODIUM", "MAGNESIUM", "CALCIUM",
"SUGAR", "TFAT", "SFAT", "CHOLESTROL")]

# Visualize correlations
corrplot(as.matrix(subset_cor), method = "circle", is.corr = FALSE)
```



Pairwise Correlation Tests

The Pearson's correlation test measures the strength and direction of the linear relationship between two continuous variables.

```
cor_test_sys_sodium <- cor.test(final_data$SYSTOLIC, final_data$SODIUM, method = "pearso
n")
print("CORRELATION SYSOLIC X SOIDUM (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X SOIDUM (PERASON)"
```

```
print(cor_test_sys_sodium)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$SODIUM
## t = -1.7625, df = 4280, p-value = 0.07805
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.056839104  0.003024407
## sample estimates:
##         cor
## -0.02693149
```

```
cor_test_dia_sodium <- cor.test(final_data$DIASTOLIC, final_data$SODIUM, method = "pears
on")
print("CORRELATION DIASTOLIC X SOIDUM (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X SOIDUM (PERASON)"
```

```
print(cor_test_dia_sodium)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$SODIUM
## t = -0.3789, df = 4280, p-value = 0.7048
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03573884  0.02416607
## sample estimates:
##          cor
## -0.005791579
```

```
cor_test_sys_magnesium <- cor.test(final_data$SYSTOLIC, final_data$MAGNESIUM, method =
"pearson")
print("CORRELATION SYSOLIC X MAGNESIUM (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X MAGNESIUM (PERASON)"
```

```
print(cor_test_sys_magnesium)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$MAGNESIUM
## t = -1.3043, df = 4280, p-value = 0.1922
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04985589  0.01002726
## sample estimates:
##         cor
## -0.01993219
```

```
cor_test_dia_magnesium <- cor.test(final_data$DIASTOLIC, final_data$MAGNESIUM, method =
"pearson")
print("CORRELATION DIASTOLIC X MAGNESIUM (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X MAGNESIUM (PERASON)"
```

```
print(cor_test_dia_magnesium)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$MAGNESIUM
## t = -3.3626, df = 4280, p-value = 0.0007788
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08116032 -0.02141111
## sample estimates:
##         cor
## -0.05133165
```

```
cor_test_sys_calcium <- cor.test(final_data$SYSTOLIC, final_data$CALCIUM, method = "pear
son")
print("CORRELATION SYSOLIC X CALCIUM (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X CALCIUM (PERASON)"
```

```
print(cor_test_sys_calcium)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$CALCIUM
## t = -2.3392, df = 4280, p-value = 0.01937
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.065615377 -0.005784874
## sample estimates:
##         cor
## -0.03573214
```

```
cor_test_dia_calcium <- cor.test(final_data$DIASTOLIC, final_data$CALCIUM, method = "pea
rson")
print("CORRELATION DIASTOLIC X CALCIUM (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X CALCIUM (PERASON)"
```

```
print(cor_test_dia_calcium)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$CALCIUM
## t = -2.7373, df = 4280, p-value = 0.00622
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07166815 -0.01186583
## sample estimates:
##         cor
## -0.04180443
```

```
cor_test_sys_sugar <- cor.test(final_data$SYSTOLIC, final_data$SUGAR, method = "pearso
n")
print("CORRELATION SYSOLIC X SUGAR (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X SUGAR (PERASON)"
```

```
print(cor_test_sys_sugar)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$SUGAR
## t = 1.1381, df = 4280, p-value = 0.2551
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01256598  0.04732283
## sample estimates:
##       cor
## 0.01739402
```

```
cor_test_dia_sugar <- cor.test(final_data$DIASTOLIC, final_data$SUGAR, method = "pearso
n")
print("CORRELATION DIASTOLIC X SUGAR (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X SUGAR (PERASON)"
```

```
print(cor_test_dia_sugar)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$SUGAR
## t = 0.030774, df = 4280, p-value = 0.9755
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02948348  0.03042343
## sample estimates:
##          cor
## 0.0004703978
```

```
cor_test_sys_tfat <- cor.test(final_data$SYSTOLIC, final_data$TFAT, method = "pearson")
print("CORRELATION SYSOLIC X TFAT (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X TFAT (PERASON)"
```

```
print(cor_test_sys_tfat)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$TFAT
## t = 0.45658, df = 4280, p-value = 0.648
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02297945  0.03692456
## sample estimates:
##         cor
## 0.006978817
```

```
cor_test_dia_tfat <- cor.test(final_data$DIASTOLIC, final_data$TFAT, method = "pearson")
print("CORRELATION DIASTOLIC X TFAT (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X TFAT (PERASON)"
```

```
print(cor_test_dia_tfat)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$TFAT
## t = -0.30436, df = 4280, p-value = 0.7609
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03460089  0.02530474
## sample estimates:
##          cor
## -0.004652246
```

```
cor_test_sys_sfat <- cor.test(final_data$SYSTOLIC, final_data$SFAT, method = "pearson")
print("CORRELATION SYSOLIC X SFAT (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X SFAT (PERASON)"
```

```
print(cor_test_sys_sfat)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$SFAT
## t = 0.12798, df = 4280, p-value = 0.8982
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02799888  0.03190781
## sample estimates:
##         cor
## 0.001956219
```

```
cor_test_dia_sfat <- cor.test(final_data$DIASTOLIC, final_data$SFAT, method = "pearson")
print("CORRELATION DIASTOLIC X SFAT (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X SFAT (PERASON)"
```

```
print(cor_test_dia_sfat)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$SFAT
## t = -0.2124, df = 4280, p-value = 0.8318
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03319683  0.02670946
## sample estimates:
##        cor
## -0.0032466
```

```
cor_test_sys_cholestrol <- cor.test(final_data$SYSTOLIC, final_data$CHOLESTROL, method =
"pearson")
print("CORRELATION SYSOLIC X CHOLESTROL (PERASON)")
```

```
## [1] "CORRELATION SYSOLIC X CHOLESTROL (PERASON)"
```

```
print(cor_test_sys_cholestrol)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$SYSTOLIC and final_data$CHOLESTROL
## t = 2.3522, df = 4280, p-value = 0.01871
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.005984535 0.065814182
## sample estimates:
##       cor
## 0.03593156
```

```
cor_test_dia_cholestrol <- cor.test(final_data$DIASTOLIC, final_data$CHOLESTROL, method
= "pearson")
print("CORRELATION DIASTOLIC X CHOLESTROL (PERASON)")
```

```
## [1] "CORRELATION DIASTOLIC X CHOLESTROL (PERASON)"
```

```
print(cor_test_dia_cholestrol)
```

```
##
##  Pearson's product-moment correlation
##
## data:  final_data$DIASTOLIC and final_data$CHOLESTROL
## t = 0.86144, df = 4280, p-value = 0.389
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01679370  0.04310285
## sample estimates:
##       cor
## 0.01316639
```

Insights

- With the correlation matrix and correlation plot the positive relation between cholesterol and sugar with Systolic BP and Diastolic BP, as well as the negative relation between calcium and magnesium with Systolic BP and Diastolic BP were observed.
- With the correlation tests some of the points we were able to identify are:
    - Statistically significant negative relation of calcium with both Systolic BP as well as Diastolic BP with $p<0.05$
    - Negative correlation of magnesium with diastolic BP
    - Positive relation of Cholesterol with Systolic BP.

Implementation of Regression model

Basic linear regression model

The implemented linear regression model constructs a linear regression to predict the Systolic BP and Diastolic BP based on the following predictors:

- Sodium
- Magnesium
- Calcium
- Sugar
- TFat
- SFat
- Cholestrol
- Age
- BMI

Visualizing residuals

Here the differences between the both Systolic BP and Diastolic BP, from their actual values and predicted values are plotted, residuals.

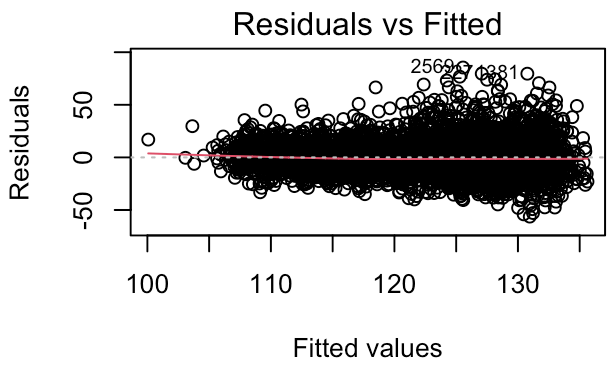
Multicolinearity

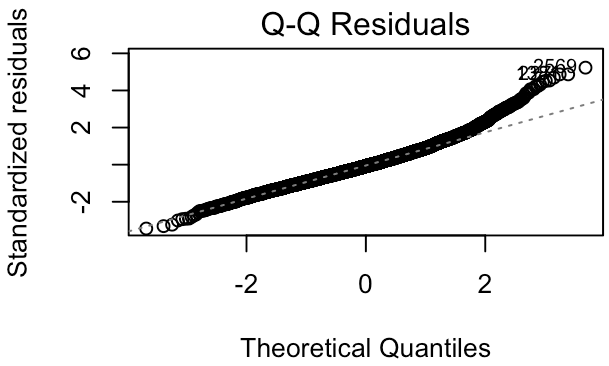
With this, the high correlation between two or more predictors will also be considered.

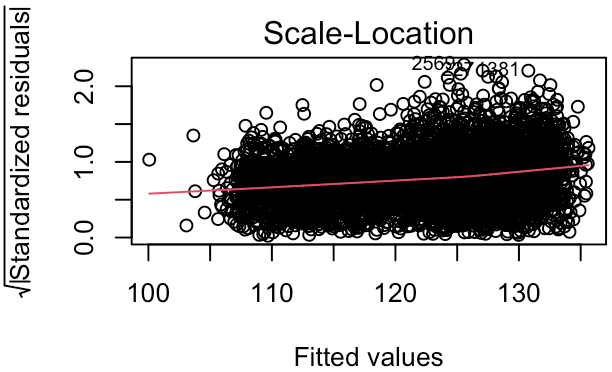
```
## REGRESSION MODEL

lm_sys <- lm(SYSTOLIC ~ SODIUM + MAGNESIUM + CALCIUM + SUGAR + SFAT + TFAT + CHOLES
TROL + AGE + BMI, data = final_data)
summary(lm_sys)
```

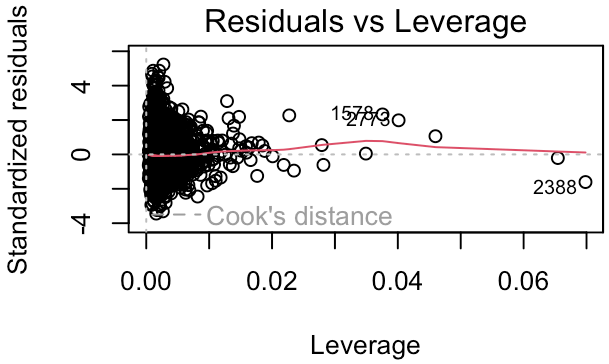
```
##
## Call:
## lm(formula = SYSTOLIC ~ SODIUM + MAGNESIUM + CALCIUM + SUGAR +
##     SFAT + TFAT + CHOLESTROL + AGE + BMI, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.291 -10.573  -1.226   9.127  85.446
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.1987076  1.5123158  64.933   <2e-16 ***
## SODIUM       0.0004738  0.0003034   1.562   0.1184
## MAGNESIUM   -0.0068797  0.0029972  -2.295   0.0218 *
## CALCIUM     -0.0008900  0.0008498  -1.047   0.2950
## SUGAR        0.0107485  0.0049321   2.179   0.0294 *
## SFAT        -0.1317863  0.0513728  -2.565   0.0103 *
## TFAT         0.0390635  0.0193246   2.021   0.0433 *
## CHOLESTROL   0.0035935  0.0014684   2.447   0.0144 *
## AGE          0.3994260  0.0143764  27.783   <2e-16 ***
## BMI          0.0696387  0.0353483   1.970   0.0489 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.36 on 4272 degrees of freedom
## Multiple R-squared:  0.1608, Adjusted R-squared:  0.1591
## F-statistic: 90.98 on 9 and 4272 DF,  p-value: < 2.2e-16
```

```
#Plots
par(mfrow=c(2,2))
plot(lm_sys)
```





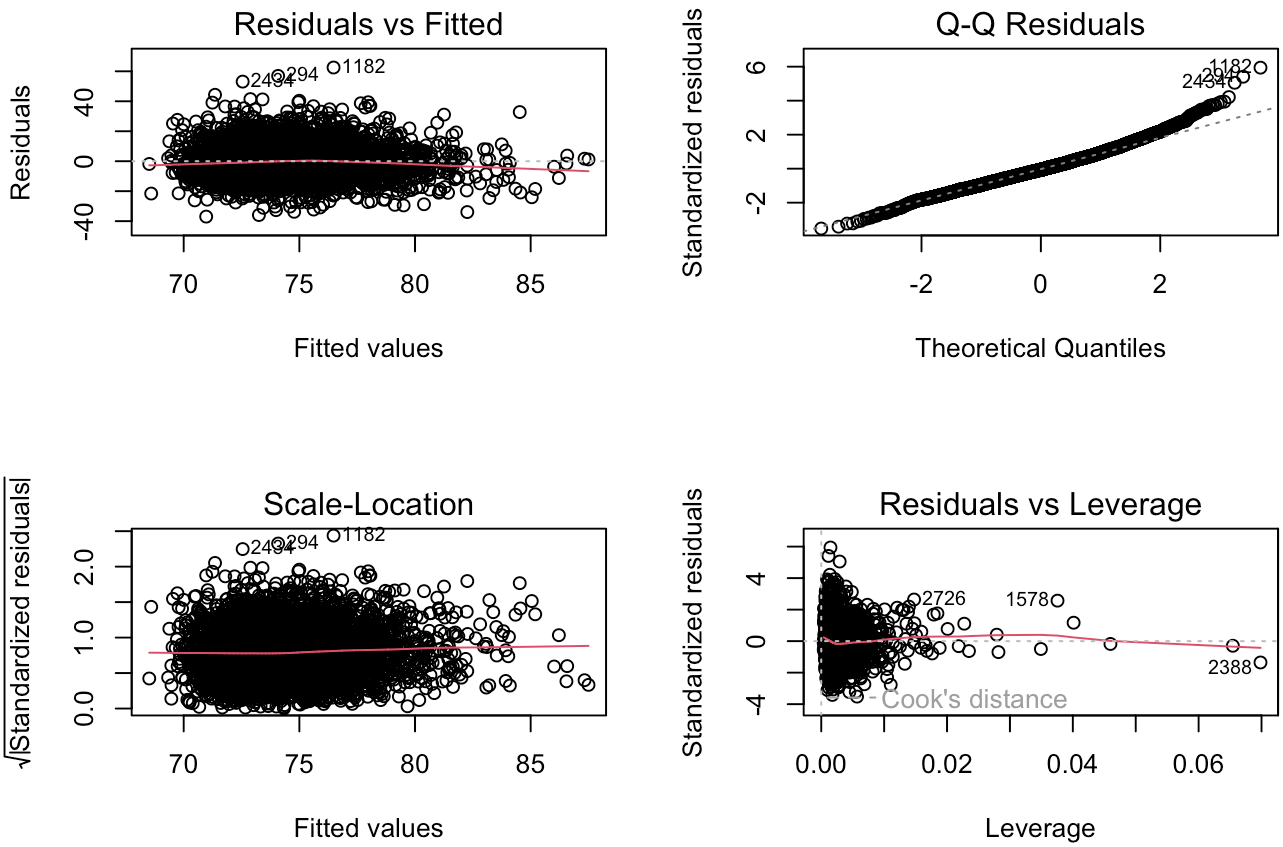





```
lm_dia <- lm(DIASTOLIC ~ SODIUM + MAGNESIUM + CALCIUM + SUGAR + SFAT + TFAT + CHOLE
STROL + AGE + BMI, data = final_data)
summary(lm_dia)
```

```
##
## Call:
## lm(formula = DIASTOLIC ~ SODIUM + MAGNESIUM + CALCIUM + SUGAR +
##     SFAT + TFAT + CHOLESTROL + AGE + BMI, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.972  -6.803  -0.664   6.268  62.521
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.476e+01  9.731e-01  66.556   <2e-16 ***
## SODIUM       5.509e-05  1.952e-04   0.282   0.7778
## MAGNESIUM   -1.612e-03  1.929e-03  -0.836   0.4033
## CALCIUM     -9.496e-04  5.468e-04  -1.737   0.0825 .
## SUGAR        4.054e-03  3.174e-03   1.278   0.2015
## SFAT        -2.464e-02  3.306e-02  -0.745   0.4561
## TFAT         9.089e-03  1.243e-02   0.731   0.4648
## CHOLESTROL   1.075e-03  9.448e-04   1.138   0.2551
## AGE          2.764e-03  9.250e-03   0.299   0.7651
## BMI          3.271e-01  2.274e-02  14.381   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.53 on 4272 degrees of freedom
## Multiple R-squared:  0.05109,    Adjusted R-squared:  0.04909
## F-statistic: 25.56 on 9 and 4272 DF,  p-value: < 2.2e-16
```

```
#Plots
par(mfrow=c(2,2))
plot(lm_dia)
```




```
#Multicolinearity check
vif(lm_sys)
```

```
##     SODIUM  MAGNESIUM    CALCIUM      SUGAR       SFAT       TFAT CHOLESTROL
##   2.645501   2.552351   2.389975   1.318951   7.963592   9.051602   1.564906
##        AGE        BMI
##   1.039665   1.032026
```

```
vif(lm_dia)
```

```
##     SODIUM  MAGNESIUM    CALCIUM      SUGAR       SFAT       TFAT CHOLESTROL
##   2.645501   2.552351   2.389975   1.318951   7.963592   9.051602   1.564906
##        AGE        BMI
##   1.039665   1.032026
```

Insights

- The regression model plotted with Systolic BP revealed that, Magnesium, Sugar, TFAT and Cholesterol has statistically significant positive effect on Systolic BP.
- Also the model revealed that factors such as Age and BMI has a strong effect on Systolic BP.

- The regression model plotted with Diastolic BP reveals that BMI has a strong positive relation with Diastolic BP.

Advanced Regression Model

To investigate more complex relationships between variables, a second, more advanced linear model is developed that incorporates interaction effects.

The interaction effects account for how the impact of one predictor on the outcome may change depending on the levels of another predictor

This model includes interaction terms involving predictors with Age and BMI to examine these relationships

```
## ADVANCED REGRESSION MODEL

lm_sys_adv <- lm(SYSTOLIC ~ SODIUM * AGE * BMI + MAGNESIUM * AGE * BMI + CALCIUM * AGE *
BMI + SUGAR * AGE * BMI + SFAT * AGE * BMI + TFAT * AGE * BMI + CHOLESTROL * AGE * BMI +
AGE + BMI, data = final_data)
summary(lm_sys_adv)
```

```
##
## Call:
## lm(formula = SYSTOLIC ~ SODIUM * AGE * BMI + MAGNESIUM * AGE *
##     BMI + CALCIUM * AGE * BMI + SUGAR * AGE * BMI + SFAT * AGE *
##     BMI + TFAT * AGE * BMI + CHOLESTROL * AGE * BMI + AGE + BMI,
##     data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.594 -10.672  -1.305   9.005  85.961
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.501e+01  9.467e+00   8.980  < 2e-16 ***
## SODIUM              5.825e-05  4.078e-03   0.014 0.988603
## AGE                 6.691e-01  1.766e-01   3.789 0.000153 ***
## BMI                 1.782e-01  3.199e-01   0.557 0.577562
## MAGNESIUM          -6.437e-02  4.112e-02  -1.565 0.117616
## CALCIUM             1.764e-02  1.199e-02   1.471 0.141333
## SUGAR               1.208e-02  5.952e-02   0.203 0.839250
## SFAT               -9.019e-02  7.011e-01  -0.129 0.897639
## TFAT                1.241e-01  2.749e-01   0.451 0.651715
## CHOLESTROL         -8.483e-03  1.936e-02  -0.438 0.661232
## SODIUM:AGE          2.435e-05  7.712e-05   0.316 0.752196
## SODIUM:BMI          8.310e-05  1.396e-04   0.595 0.551620
## AGE:BMI            -2.841e-03  5.968e-03  -0.476 0.634039
## AGE:MAGNESIUM       1.140e-03  7.651e-04   1.490 0.136196
## BMI:MAGNESIUM       2.579e-03  1.482e-03   1.740 0.081985 .
## AGE:CALCIUM        -3.518e-04  2.190e-04  -1.606 0.108295
## BMI:CALCIUM        -7.633e-04  4.080e-04  -1.871 0.061427 .
## AGE:SUGAR          -3.625e-04  1.189e-03  -0.305 0.760445
## BMI:SUGAR          -8.885e-07  1.938e-03   0.000 0.999634
## AGE:SFAT            5.995e-03  1.273e-02   0.471 0.637806
## BMI:SFAT            6.836e-03  2.394e-02   0.285 0.775285
## AGE:TFAT           -3.483e-03  4.990e-03  -0.698 0.485167
## BMI:TFAT           -3.929e-03  9.584e-03  -0.410 0.681893
## AGE:CHOLESTROL     -2.609e-05  3.670e-04  -0.071 0.943341
## BMI:CHOLESTROL      2.498e-04  6.771e-04   0.369 0.712196
## SODIUM:AGE:BMI     -2.163e-06  2.623e-06  -0.825 0.409690
## AGE:BMI:MAGNESIUM  -5.033e-05  2.741e-05  -1.837 0.066347 .
## AGE:BMI:CALCIUM     1.432e-05  7.457e-06   1.920 0.054949 .
## AGE:BMI:SUGAR       1.113e-05  3.901e-05   0.285 0.775404
## AGE:BMI:SFAT       -3.487e-04  4.344e-04  -0.803 0.422198
## AGE:BMI:TFAT        1.387e-04  1.730e-04   0.802 0.422763
## AGE:BMI:CHOLESTROL  3.748e-06  1.264e-05   0.297 0.766843
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.32 on 4250 degrees of freedom
## Multiple R-squared:   0.17,  Adjusted R-squared:  0.164
## F-statistic: 28.09 on 31 and 4250 DF,  p-value: < 2.2e-16
```

```
lm_dia_adv <- lm(DIASTOLIC ~ SODIUM * AGE * BMI + MAGNESIUM * AGE * BMI + CALCIUM * AGE
* BMI + SUGAR * AGE * BMI + SFAT * AGE * BMI + TFAT * AGE * BMI + CHOLESTROL * AGE * BMI
+ AGE + BMI, data = final_data)
summary(lm_dia_adv)
```

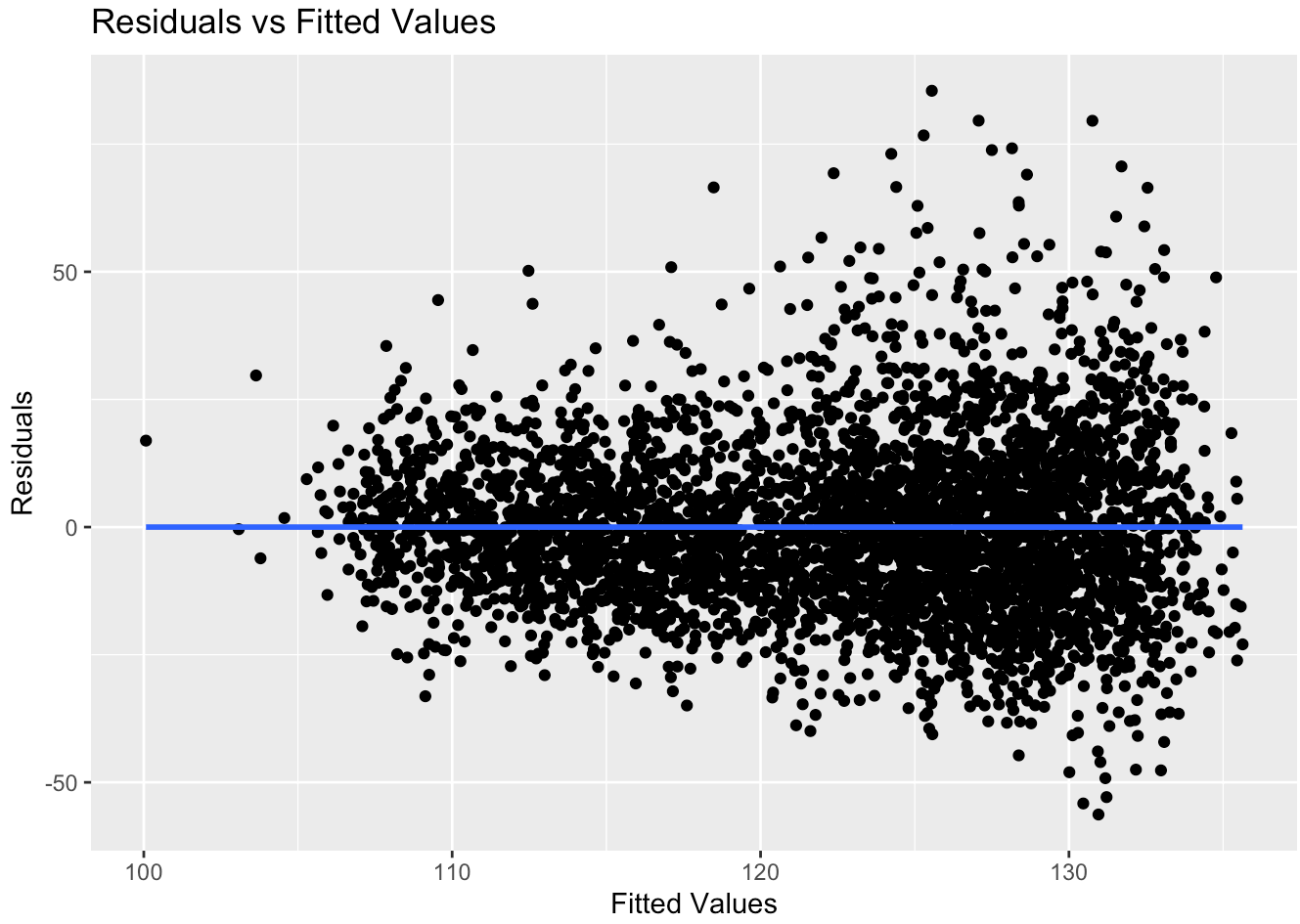
```
##
## Call:
## lm(formula = DIASTOLIC ~ SODIUM * AGE * BMI + MAGNESIUM * AGE *
##     BMI + CALCIUM * AGE * BMI + SUGAR * AGE * BMI + SFAT * AGE *
##     BMI + TFAT * AGE * BMI + CHOLESTROL * AGE * BMI + AGE + BMI,
##     data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.082  -6.707  -0.643   6.120  62.426
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.063e+01  6.071e+00   9.987   <2e-16 ***
## SODIUM             -1.751e-03  2.615e-03  -0.670   0.5032
## AGE                 7.199e-02  1.133e-01   0.636   0.5251
## BMI                 4.546e-01  2.051e-01   2.216   0.0267 *
## MAGNESIUM          -4.244e-02  2.637e-02  -1.609   0.1077
## CALCIUM             9.535e-03  7.690e-03   1.240   0.2150
## SUGAR              -2.430e-02  3.817e-02  -0.636   0.5245
## SFAT                1.601e-01  4.496e-01   0.356   0.7217
## TFAT                4.791e-02  1.763e-01   0.272   0.7859
## CHOLESTROL         -2.525e-02  1.241e-02  -2.034   0.0420 *
## SODIUM:AGE          6.134e-05  4.946e-05   1.240   0.2149
## SODIUM:BMI          6.475e-05  8.951e-05   0.723   0.4695
## AGE:BMI            -2.177e-03  3.827e-03  -0.569   0.5694
## AGE:MAGNESIUM       6.884e-04  4.906e-04   1.403   0.1607
## BMI:MAGNESIUM       1.692e-03  9.506e-04   1.779   0.0752 .
## AGE:CALCIUM        -1.538e-04  1.405e-04  -1.095   0.2735
## BMI:CALCIUM        -5.147e-04  2.617e-04  -1.967   0.0493 *
## AGE:SUGAR           2.180e-04  7.624e-04   0.286   0.7750
## BMI:SUGAR           7.749e-04  1.243e-03   0.623   0.5331
## AGE:SFAT           -4.820e-03  8.166e-03  -0.590   0.5551
## BMI:SFAT            3.909e-04  1.536e-02   0.025   0.9797
## AGE:TFAT           -4.224e-04  3.200e-03  -0.132   0.8950
## BMI:TFAT           -2.444e-03  6.147e-03  -0.398   0.6909
## AGE:CHOLESTROL      3.701e-04  2.354e-04   1.572   0.1159
## BMI:CHOLESTROL      9.250e-04  4.342e-04   2.130   0.0332 *
## SODIUM:AGE:BMI     -2.114e-06  1.682e-06  -1.257   0.2090
## AGE:BMI:MAGNESIUM  -2.911e-05  1.758e-05  -1.656   0.0977 .
## AGE:BMI:CALCIUM     8.060e-06  4.782e-06   1.685   0.0920 .
## AGE:BMI:SUGAR      -4.473e-06  2.502e-05  -0.179   0.8581
## AGE:BMI:SFAT        4.748e-05  2.786e-04   0.170   0.8647
## AGE:BMI:TFAT        3.441e-05  1.110e-04   0.310   0.7565
## AGE:BMI:CHOLESTROL -1.292e-05  8.107e-06  -1.594   0.1110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.46 on 4250 degrees of freedom
## Multiple R-squared:  0.0677, Adjusted R-squared:  0.0609
## F-statistic: 9.956 on 31 and 4250 DF,  p-value: < 2.2e-16
```

Insights

- Age is a significant factor, having a positive association with Systolic BP.
- BMI:Calcium and BMI:Cholesterol showcased how the effects of it on Diastolic BP depends on BMI levels.

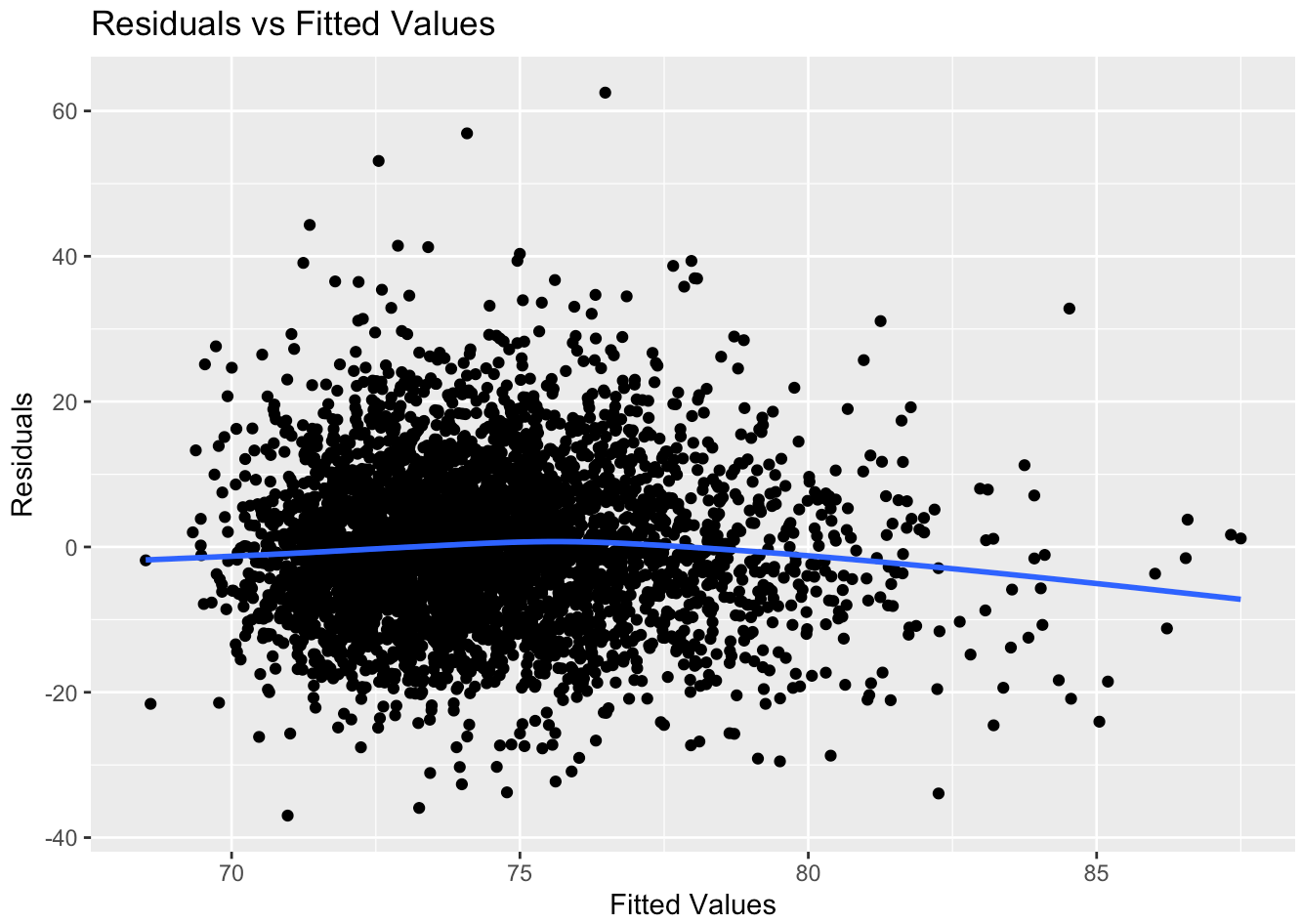
```
ggplot(lm_sys, aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(title = "Residuals vs Fitted Values", x = "Fitted Values", y = "Residuals")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



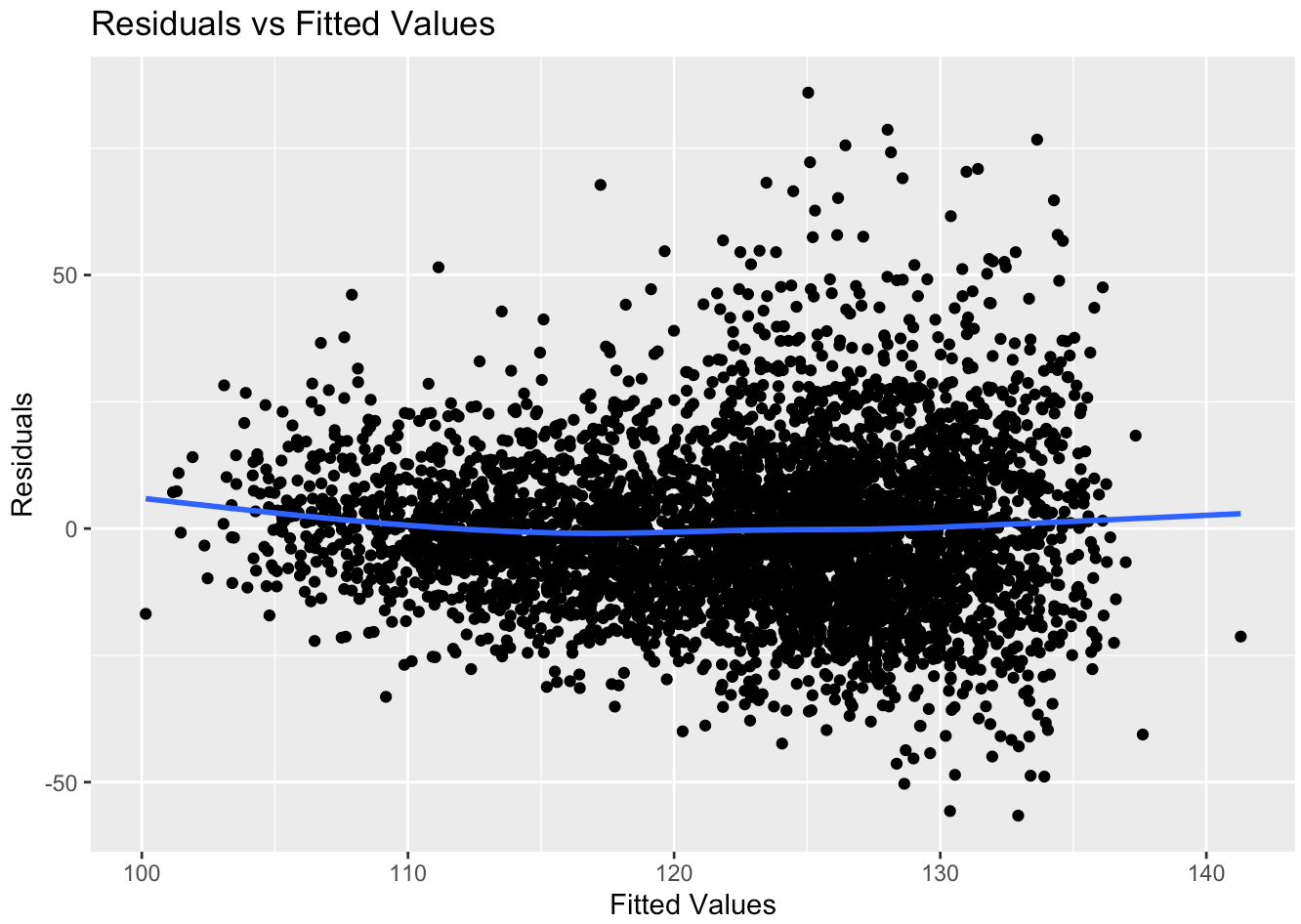

```
ggplot(lm_dia, aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(title = "Residuals vs Fitted Values", x = "Fitted Values", y = "Residuals")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



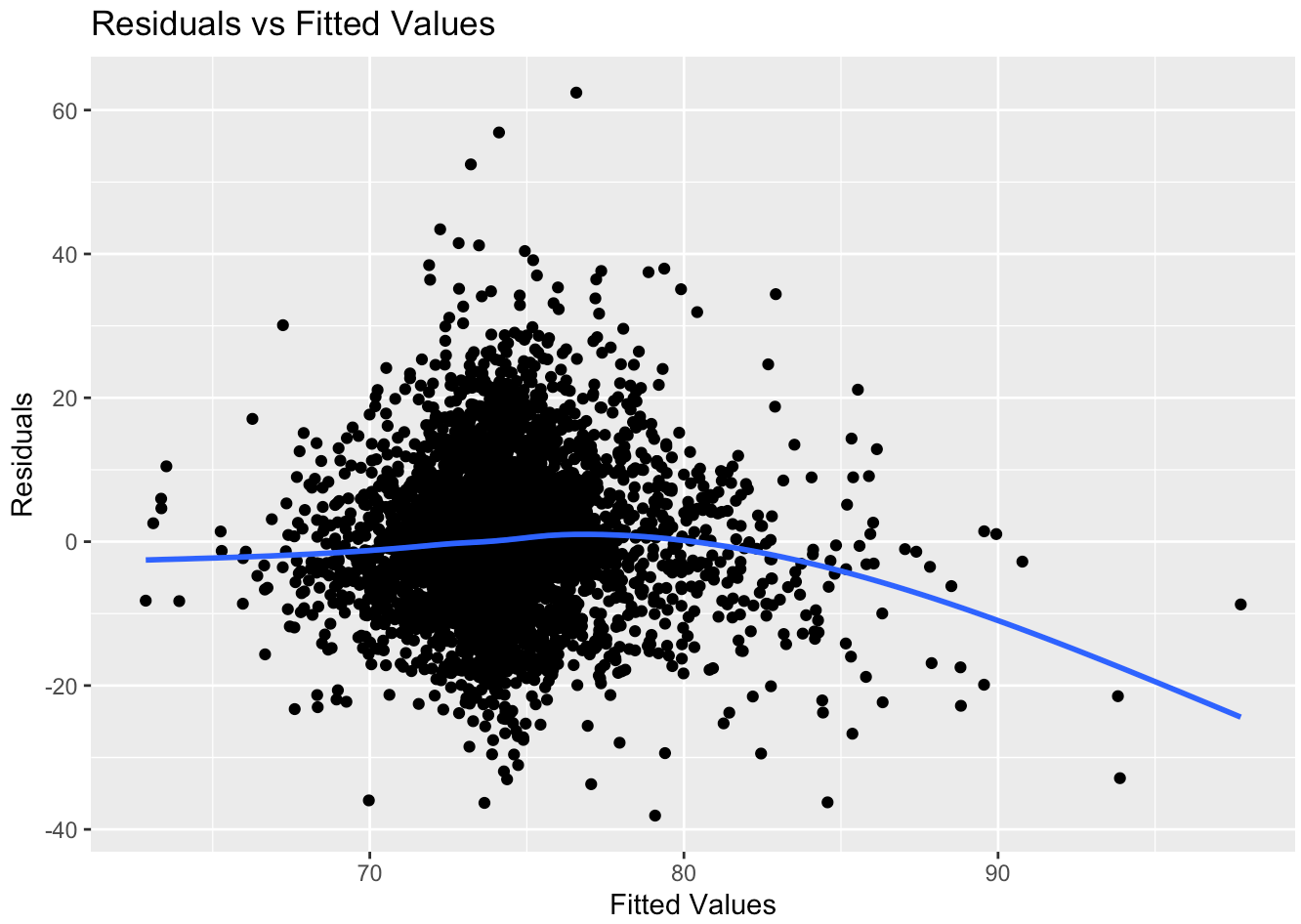

```
ggplot(lm_sys_adv, aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(title = "Residuals vs Fitted Values", x = "Fitted Values", y = "Residuals")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
ggplot(lm_dia_adv, aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(title = "Residuals vs Fitted Values", x = "Fitted Values", y = "Residuals")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Correlation test on the effect of Sodium on BP Age Category Wise

```
# Perform correlation test by age category
cor_test_results <- by(
  final_data,
  final_data$Age_Category,
  function(subset_data) {
    cor.test(subset_data$SYSTOLIC, subset_data$SODIUM, method = "pearson")
  }
)

# Print results for each age category
print("CORRELATION BETWEEN SYSTOLIC AND SODIUM (PEARSON) BY AGE CATEGORY")
```

```
## [1] "CORRELATION BETWEEN SYSTOLIC AND SODIUM (PEARSON) BY AGE CATEGORY"
```

```
for (age_category in names(cor_test_results)) {
  cat("\nAge Category:", age_category, "\n")
  print(cor_test_results[[age_category]])
}
```

```
##
## Age Category: Young Adults
##
##  Pearson's product-moment correlation
##
## data:  subset_data$SYSTOLIC and subset_data$SODIUM
## t = 2.5416, df = 570, p-value = 0.0113
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02408596 0.18622391
## sample estimates:
##       cor
## 0.1058585
##
##
## Age Category: Middle Aged Adults
##
##  Pearson's product-moment correlation
##
## data:  subset_data$SYSTOLIC and subset_data$SODIUM
## t = 3.9604, df = 1021, p-value = 8.002e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0621799 0.1829160
## sample estimates:
##      cor
## 0.123003
##
##
## Age Category: Elderly
##
##  Pearson's product-moment correlation
##
## data:  subset_data$SYSTOLIC and subset_data$SODIUM
## t = -1.7951, df = 2685, p-value = 0.07275
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.072341782  0.003195184
## sample estimates:
##         cor
## -0.03462275
```

Insights

- In young adults a weak positive correlation was observed between Sodium and Systolic BP
- In middle aged adults, a highly significant strong evidence of positive relationship between sodium and Systolic BP was noted.

Results

Descriptive Statistics

- The code provide summary tables that depicts key characteristics of the variables in the dataset like:
    - Means
    - Median
    - Quadrant, etc..
- We have also displayed the frequence tables of categorical variables
- Finally, we made use of ANOVA tests to confirm the BP distribution across the categories.

Exploratory Data Analysis

- The scatter plots between nutrients and blood pressure (both systolic and diastolic) helped visualize the relationship between nutrient intake and blood pressure levels, providing insights into how different nutrients may affect BP.
- The box plots comparing age and blood pressure indicate that the likelihood of having higher systolic blood pressure increases with age. In contrast, middle-aged individuals tend to exhibit higher values for diastolic blood pressure.

Correlation Analysis

- The correlation matrix and plot revealed a positive relationship between cholesterol and sugar with both systolic and diastolic blood pressure, as well as a negative relationship between calcium and magnesium with both types of blood pressure.
- The correlation tests highlighted several key findings:
    - A statistically significant negative relationship between calcium and both systolic and diastolic blood pressure ($p < 0.05$).
    - A negative correlation between magnesium and diastolic blood pressure.
    - A positive relationship between cholesterol and systolic blood pressure.

Regression Analysis

- The regression model for systolic blood pressure indicated that magnesium, sugar, TFAT, and cholesterol have a statistically significant positive effect on systolic BP.
- Additionally, the model revealed that age and BMI have a strong influence on systolic BP.
- The regression model for diastolic blood pressure highlighted that BMI has a strong positive relationship with diastolic BP.

Advanced Regression Analysis

- Age is a significant factor, showing a positive association with systolic blood pressure.

- The interactions between BMI and calcium, as well as BMI and cholesterol, demonstrated how these factors' effects on diastolic blood pressure vary depending on BMI levels.

Correlation test on the effect of Sodium on BP, Age category wise

- A weak positive correlation between sodium and systolic blood pressure was found in young adults.
- A strong and statistically significant positive relationship between sodium and systolic blood pressure was observed in middle-aged adults.

Future Scope

- **Longitudinal Studies**: Conducting longitudinal studies to track the changes in BP over time in relation to nutrient intake would provide deeper insights into causal relationships and help in formulating long-term dietary guidelines for managing blood pressure.
- **Personalized Health Recommendations**: Development of AI-based tools for personalized nutrition advice could be pursued, leveraging the findings to design individualized dietary plans for managing and preventing hypertension, particularly focusing on the age-based differences identified in this study.

Key Takeaways

- Nutrients such as Magnesium, Sugar, TFAT, and Cholesterol have a significant impact on Systolic and Diastolic blood pressure, with age and BMI being key factors influencing BP levels across different age groups.
- Correlation and regression analyses revealed varying relationships between nutrients and BP in different age categories, with a strong positive association between Sodium and Systolic BP in middle-aged adults, and significant effects of Calcium and Magnesium on both Systolic and Diastolic BP.