Patel’s Week 2 Dataset Intro to Data

Harsh Patel

2024-10-11

```r
library(readxl)

Chronic_Kidney_Disease_data <-
read_excel("C:/Users/hpate/Downloads/Chronic_Kidney_Dsease_data.xlsx")
```

In this section, I load the readxl library, which allows me to read Excel files in R. I then read the Excel file containing the chronic kidney disease data from the specified path and store it in a variable called Chronic_Kidney_Disease_data.

```r
selected_data <- Chronic_Kidney_Disease_data[c("Age", "BMI", "SystolicBP",
"DiastolicBP", "SerumCreatinine")]
```

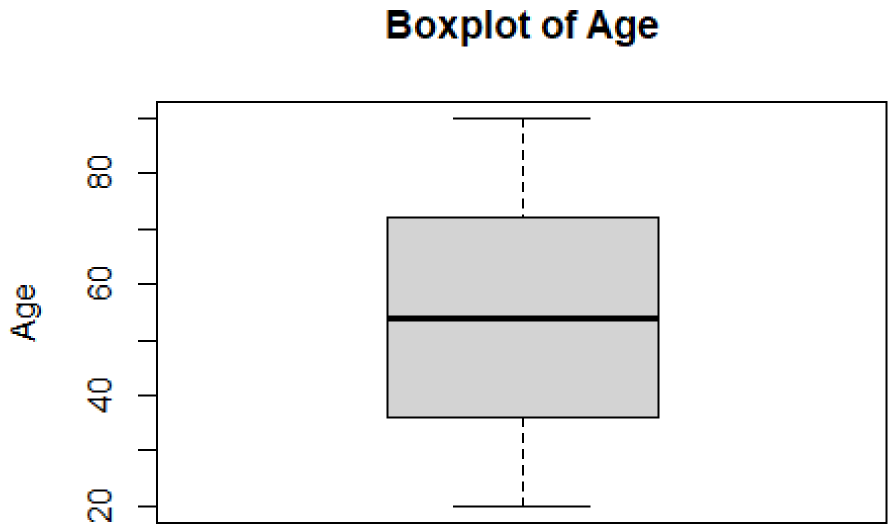
Next, I select specific columns from the Chronic_Kidney_Disease_data dataset. I focus on the variables “Age,” “BMI,” “SystolicBP,” “DiastolicBP,” and “SerumCreatinine,” and I store this subset in a new variable named selected_data. This helps me concentrate on the variables that are most relevant for my analysis.

```r
age_mean <- mean(selected_data$Age, na.rm = TRUE)
age_sd <- sd(selected_data$Age, na.rm = TRUE)
age_median <- median(selected_data$Age, na.rm = TRUE)
age_IQR <- IQR(selected_data$Age, na.rm = TRUE)

bmi_mean <- mean(selected_data$BMI, na.rm = TRUE)
bmi_sd <- sd(selected_data$BMI, na.rm = TRUE)
bmi_median <- median(selected_data$BMI, na.rm = TRUE)
bmi_IQR <- IQR(selected_data$BMI, na.rm = TRUE)

systolic_bp_mean <- mean(selected_data$SystolicBP, na.rm = TRUE)
systolic_bp_sd <- sd(selected_data$SystolicBP, na.rm = TRUE)
systolic_bp_median <- median(selected_data$SystolicBP, na.rm = TRUE)
systolic_bp_IQR <- IQR(selected_data$SystolicBP, na.rm = TRUE)

diastolic_bp_mean <- mean(selected_data$DiastolicBP, na.rm = TRUE)
diastolic_bp_sd <- sd(selected_data$DiastolicBP, na.rm = TRUE)
diastolic_bp_median <- median(selected_data$DiastolicBP, na.rm = TRUE)
```

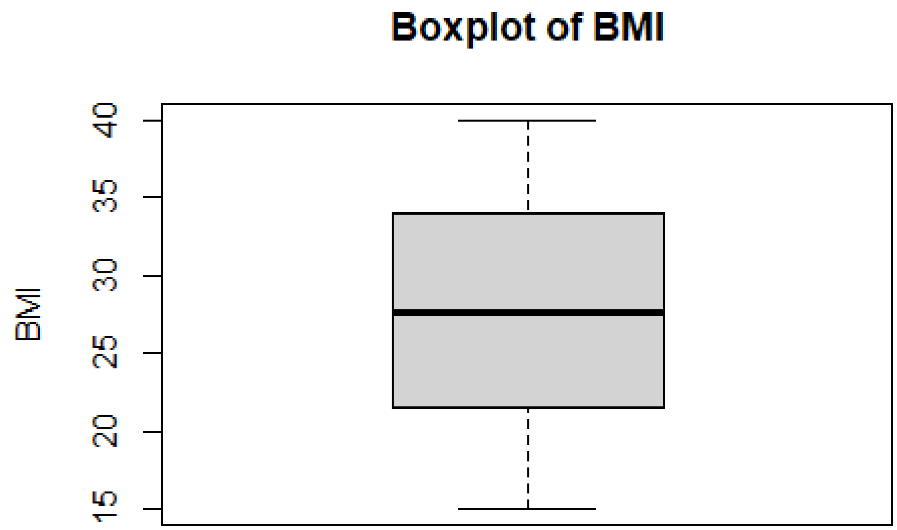
```r
diastolic_bp_IQR <- IQR(selected_data$DiastolicBP, na.rm = TRUE)

serum_creatinine_mean <- mean(selected_data$SerumCreatinine, na.rm = TRUE)
serum_creatinine_sd <- sd(selected_data$SerumCreatinine, na.rm = TRUE)
serum_creatinine_median <- median(selected_data$SerumCreatinine, na.rm =
TRUE)
serum_creatinine_IQR <- IQR(selected_data$SerumCreatinine, na.rm = TRUE)
```

In this part, I calculate summary statistics for the selected variables in selected_data. For each variable (Age, BMI, Systolic BP, Diastolic BP, and Serum Creatinine), I compute the mean, standard deviation, median, and interquartile range (IQR). By using the na.rm = TRUE argument, I ensure that any missing values are excluded from my calculations. These statistics provide valuable insight into the distribution and central tendency of the data.

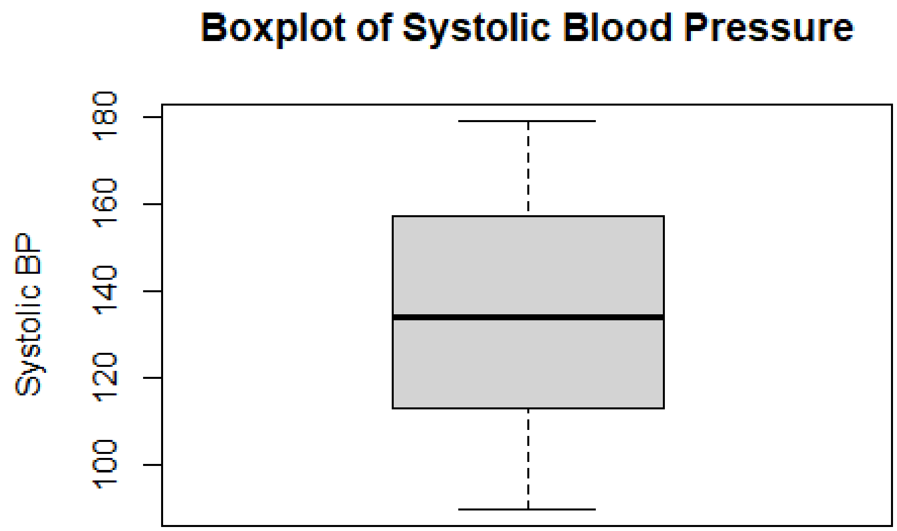
```r
boxplot(selected_data$Age, main = "Boxplot of Age", ylab = "Age")
```



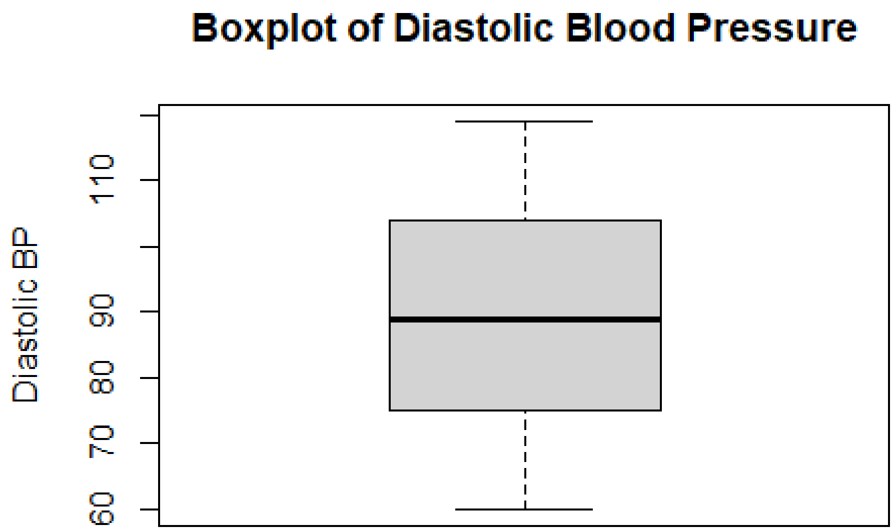
```r
boxplot(selected_data$BMI, main = "Boxplot of BMI", ylab = "BMI")
```



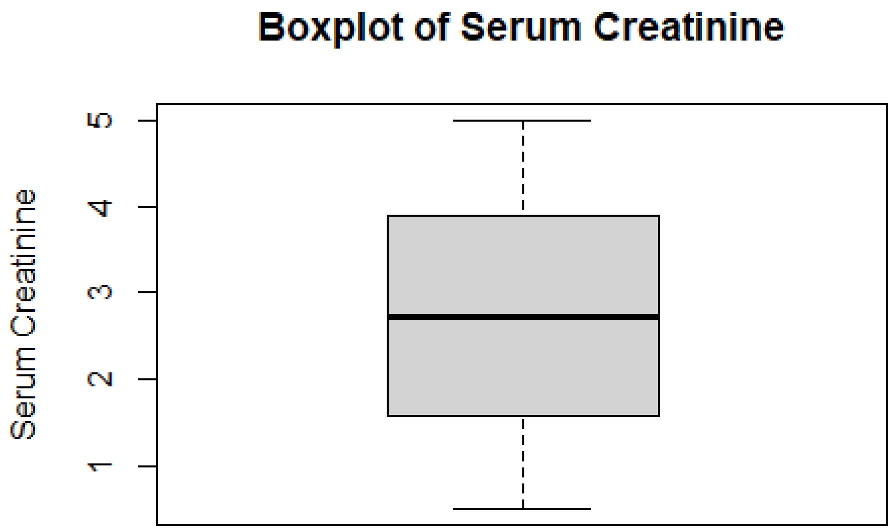
```r
boxplot(selected_data$SystolicBP, main = "Boxplot of Systolic Blood 
Pressure", ylab = "Systolic BP")
```



```
boxplot(selected_data$DiastolicBP, main = "Boxplot of Diastolic Blood
Pressure", ylab = "Diastolic BP")
```

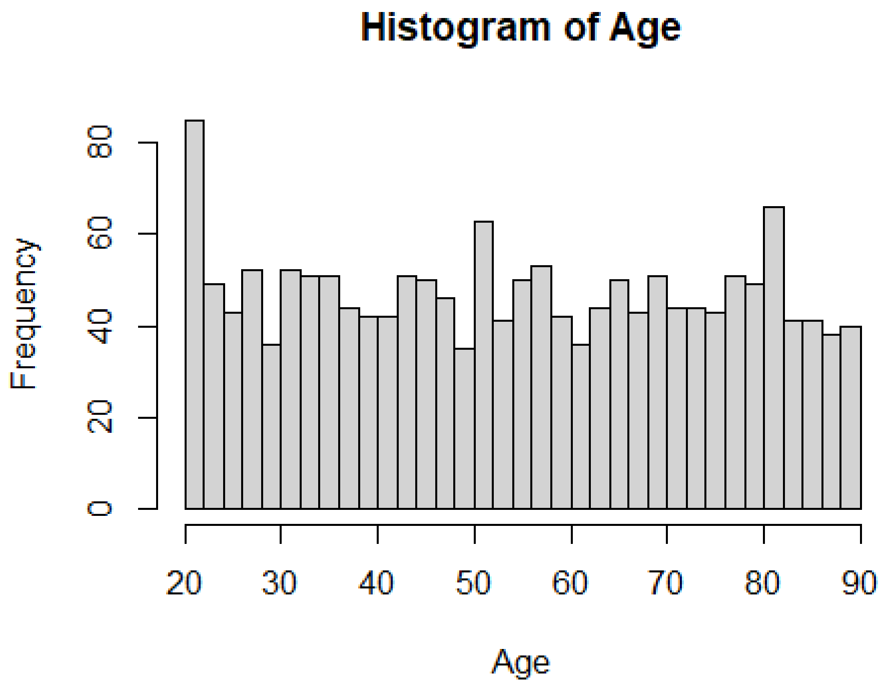


```
boxplot(selected_data$SerumCreatinine, main = "Boxplot of Serum Creatinine",
ylab = "Serum Creatinine")
```

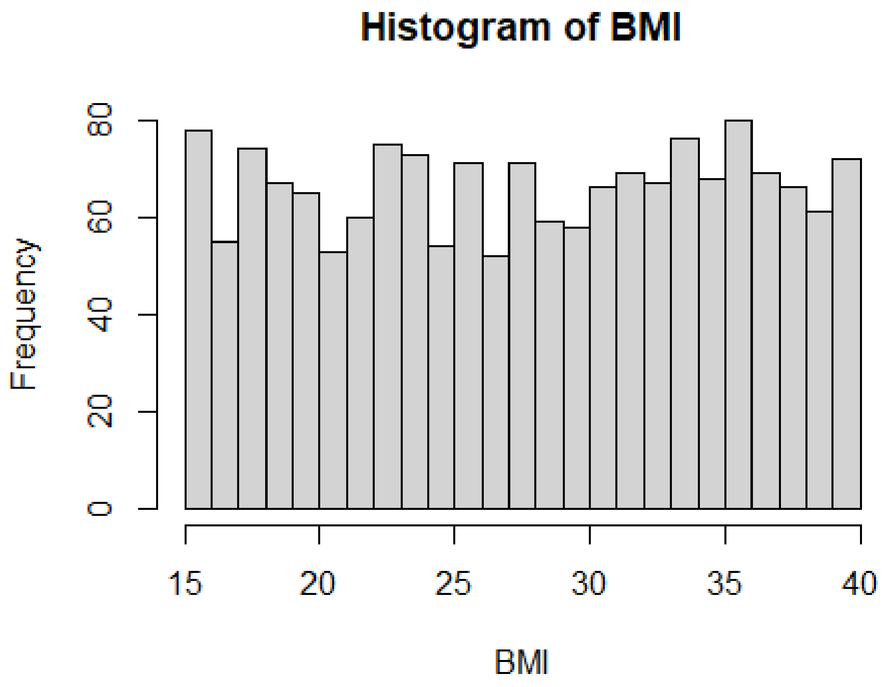


Here, I create boxplots for each of the selected variables: Age, BMI, Systolic Blood Pressure, Diastolic Blood Pressure, and Serum Creatinine. These boxplots help me visualize the distribution of the data, highlighting the median, quartiles, and any potential outliers. I ensure each plot is labeled with a title and a y-axis label for better understanding.

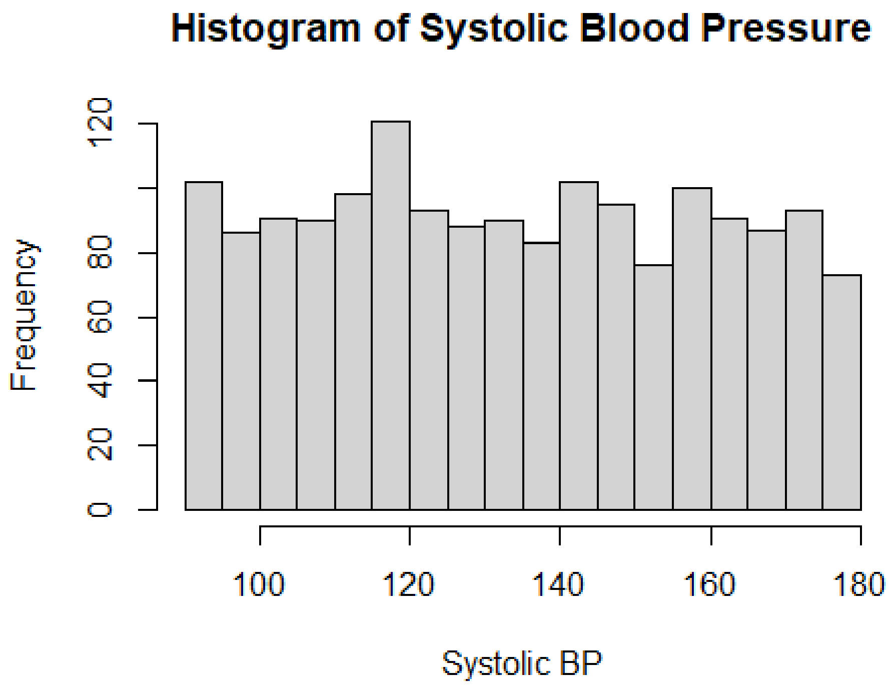
```
hist(selected_data$Age, main = "Histogram of Age", xlab = "Age", breaks = 30)
```



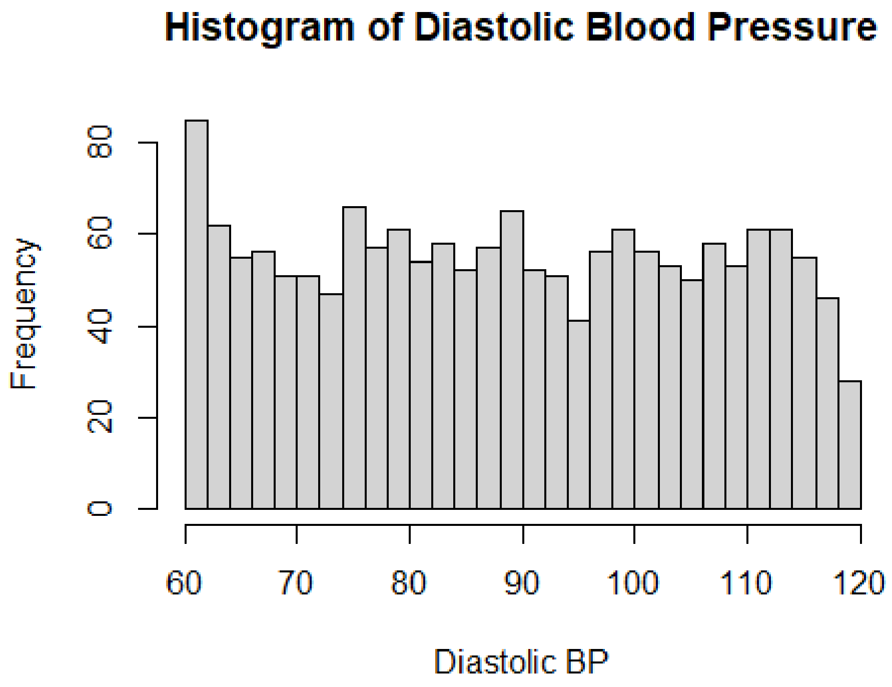
```r
hist(selected_data$BMI, main = "Histogram of BMI", xlab = "BMI", breaks = 30)
```



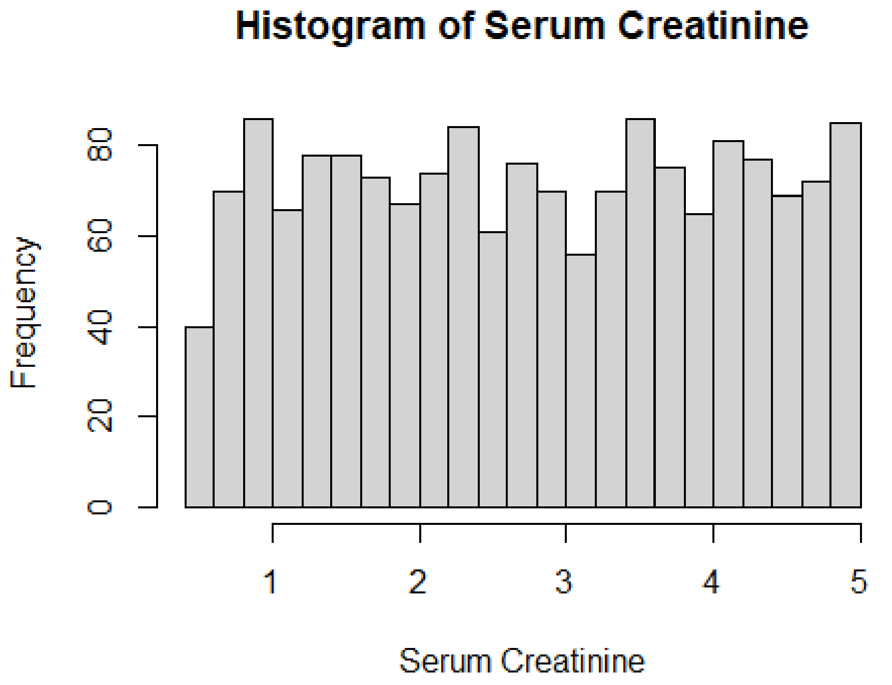
```r
hist(selected_data$SystolicBP, main = "Histogram of Systolic Blood Pressure",
xlab = "Systolic BP", breaks = 30)
```



```r
hist(selected_data$DiastolicBP, main = "Histogram of Diastolic Blood
Pressure", xlab = "Diastolic BP", breaks = 30)
```

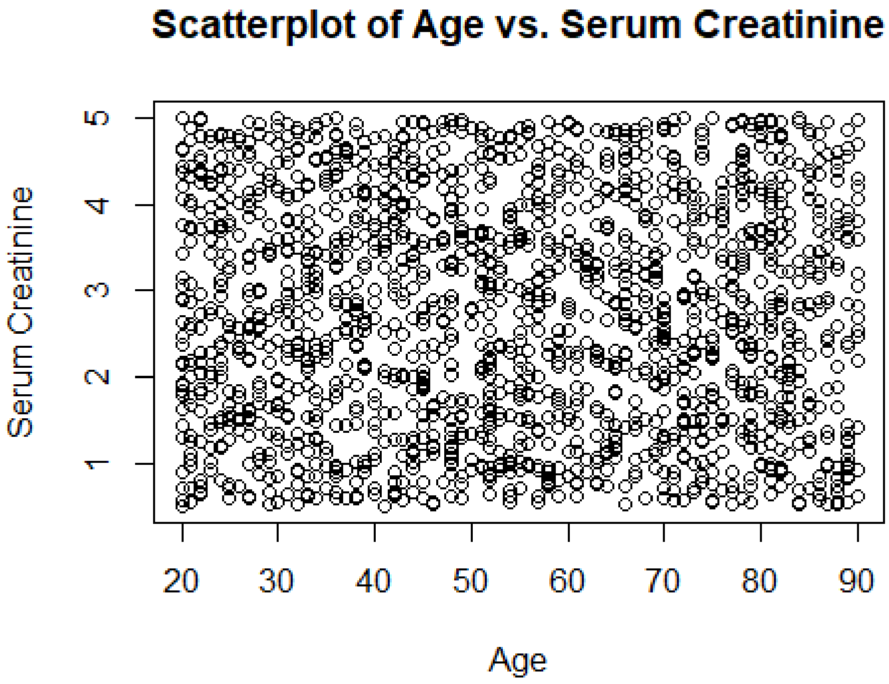


```
hist(selected_data$SerumCreatinine, main = "Histogram of Serum Creatinine",
xlab = "Serum Creatinine", breaks = 30)
```

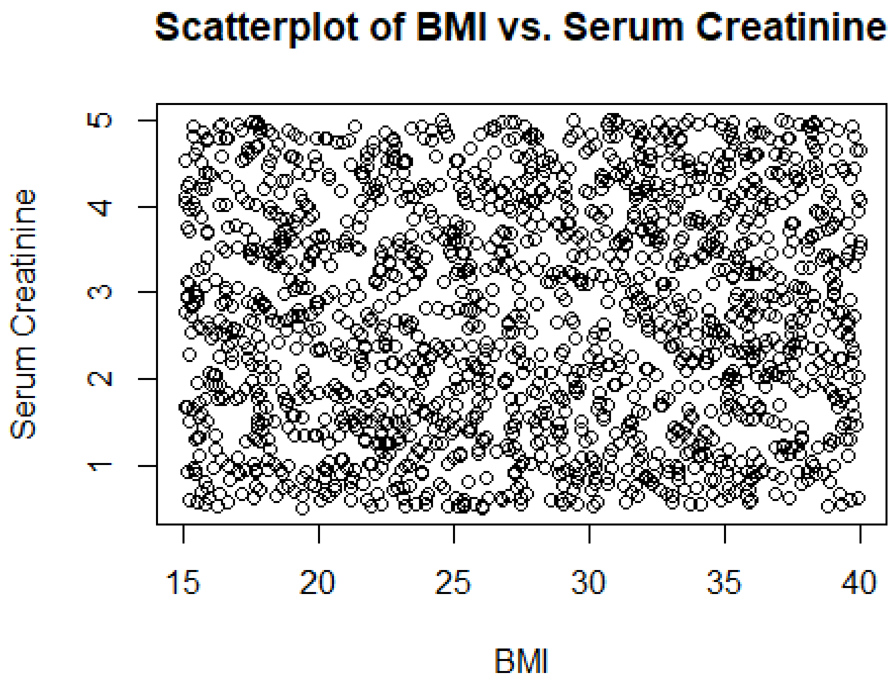


In this section, I generate histograms for each variable in selected_data. Histograms allow me to visually represent the distribution of the data by showing the frequency of values within specified intervals (or bins). By setting breaks = 30, I create a detailed view of how the data is distributed across different ranges.

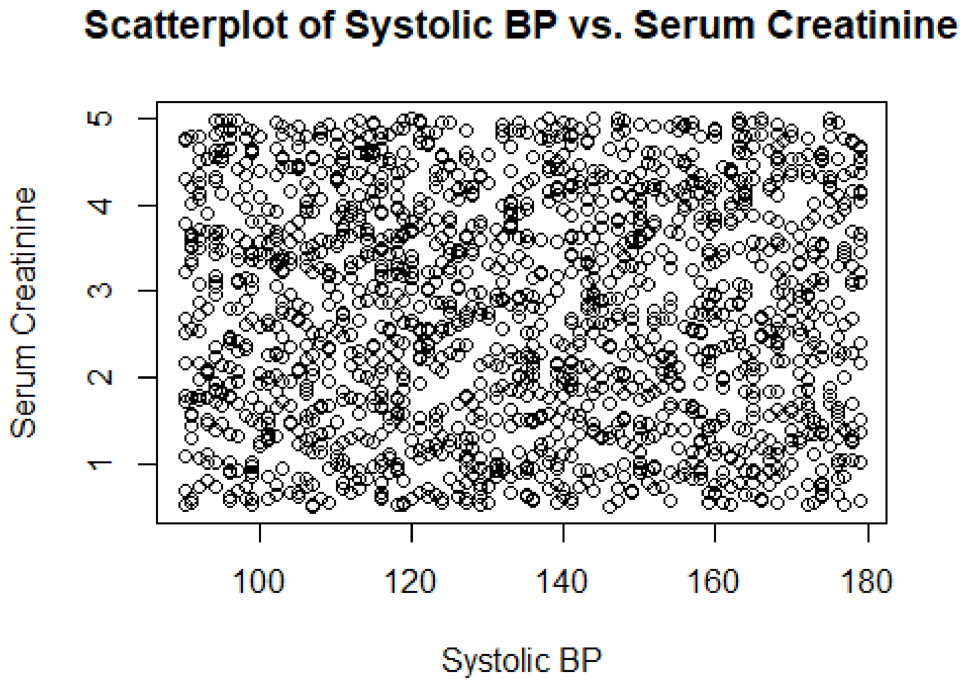
```
plot(selected_data$Age, selected_data$SerumCreatinine, main = "Scatterplot of
Age vs. Serum Creatinine", xlab = "Age", ylab = "Serum Creatinine")
```



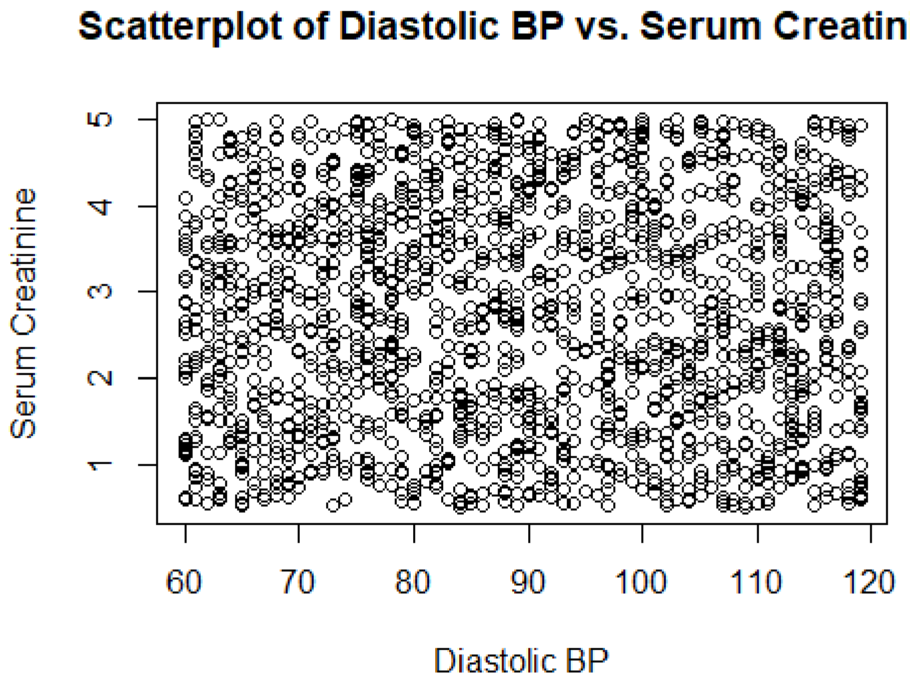
```
plot(selected_data$BMI, selected_data$SerumCreatinine, main = "Scatterplot of
BMI vs. Serum Creatinine", xlab = "BMI", ylab = "Serum Creatinine")
```



```r
plot(selected_data$SystolicBP, selected_data$SerumCreatinine, main =
"Scatterplot of Systolic BP vs. Serum Creatinine", xlab = "Systolic BP", ylab
= "Serum Creatinine")
```



```r
plot(selected_data$DiastolicBP, selected_data$SerumCreatinine, main =
"Scatterplot of Diastolic BP vs. Serum Creatinine", xlab = "Diastolic BP",
ylab = "Serum Creatinine")
```



In this part, I create scatterplots to visualize the relationship between Serum Creatinine and each of the other variables: Age, BMI, Systolic BP, and Diastolic BP. Scatterplots are useful for assessing the correlation and potential patterns between two continuous variables. I ensure that each plot is clearly labeled with appropriate titles and axis labels.

```r
summary_stats <- data.frame(
  Statistic = c("Mean", "Standard Deviation", "Median", "IQR"),
  Age = c(age_mean, age_sd, age_median, age_IQR),
  BMI = c(bmi_mean, bmi_sd, bmi_median, bmi_IQR),
  Systolic_BP = c(systolic_bp_mean, systolic_bp_sd, systolic_bp_median,
systolic_bp_IQR),
  Diastolic_BP = c(diastolic_bp_mean, diastolic_bp_sd, diastolic_bp_median,
diastolic_bp_IQR),
  Serum_Creatinine = c(serum_creatinine_mean, serum_creatinine_sd,
serum_creatinine_median, serum_creatinine_IQR)
)
print(summary_stats)
```

```
##            Statistic      Age      BMI Systolic_BP Diastolic_BP
## 1               Mean 54.44123 27.62005   134.39241     89.31344
## 2 Standard Deviation 20.54976  7.28867    25.76779     17.35448
## 3             Median 54.00000 27.65208   134.00000     89.00000
## 4                IQR 36.00000 12.54440    44.00000     29.00000
##   Serum_Creatinine
## 1         2.753198
## 2         1.317168
## 3         2.732006
## 4         2.306575
```

In the last part of my code, I create a data frame called summary_stats that summarizes important statistics for Age, BMI, Systolic Blood Pressure (Systolic BP), Diastolic Blood Pressure (Diastolic BP), and Serum Creatinine. This summary helps me understand the data better. For example, the average age of the participants is about 54.44 years, with some variability indicated by a standard deviation of 20.55 years. The average BMI is 27.62, which suggests that the participants are mostly overweight. The average systolic BP is 134.39 mmHg and the diastolic BP is 89.31 mmHg, which could indicate high blood pressure in some people. Lastly, the mean serum creatinine level is 2.75 mg/dL, which may show a decline in kidney function. Overall, this summary gives me a clear view of the participants’ health metrics.