Patel’s Dataset 02 Normal Poisson

Harsh Patel

2024-09-27

```
library(readxl)

Chronic_Kidney_Disease_data <- read_excel("Chronic_Kidney_Dsease_data.xlsx")
```

I’m kicking things off by loading the readxl library, which lets me easily read my Chronic Kidney Disease dataset. This data is full of valuable information that I’m eager to dive into!

```
Chronic_Kidney_Disease_data$Hypertension <-
as.factor(Chronic_Kidney_Disease_data$FamilyHistoryHypertension)

levels(Chronic_Kidney_Disease_data$Hypertension) <- c("No", "Yes")
```

Next, I’m transforming the hypertension status into a factor! This is crucial because it helps categorize the data into “No” and “Yes,” making it so much easier to analyze the relationship between hypertension and chronic kidney disease. I can’t wait to see what insights this will reveal!

```
table_hypertension <- table(Chronic_Kidney_Disease_data$Diagnosis,
Chronic_Kidney_Disease_data$Hypertension)

print("Contingency Table for CKD and Hypertension:")

## [1] "Contingency Table for CKD and Hypertension:"

print(table_hypertension)

##
##       No  Yes
##   0   99   36
##   1 1060  464

prop_hypertension <- prop.table(table_hypertension, 2)
print("Proportion of CKD by Hypertension Status:")

## [1] "Proportion of CKD by Hypertension Status:"

print(prop_hypertension)
```

```
##
##              No        Yes
##   0 0.08541846 0.07200000
##   1 0.91458154 0.92800000
```

I’m creating a contingency table to show the relationship between CKD diagnosis and hypertension status. This table will really help visualize the frequency of each category! I’m also calculating proportions to understand how CKD cases are distributed between those with and without hypertension.

```
binom_test <- binom.test(x = sum(Chronic_Kidney_Disease_data$Hypertension ==
"Yes" & Chronic_Kidney_Disease_data$Diagnosis == "CKD"),
                         n = sum(Chronic_Kidney_Disease_data$Hypertension ==
"Yes"),
                         p = 0.5)

print("Binomial Test Results:")

## [1] "Binomial Test Results:"

print(binom_test)

##
##  Exact binomial test
##
## data:  sum(Chronic_Kidney_Disease_data$Hypertension == "Yes" &
Chronic_Kidney_Disease_data$Diagnosis == "CKD") and
sum(Chronic_Kidney_Disease_data$Hypertension == "Yes")
## number of successes = 0, number of trials = 500, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.00000000 0.00735061
## sample estimates:
## probability of success
##                      0
```

I’m now conducting a binomial test to determine if the proportion of CKD patients with a history of hypertension significantly deviates from what I would expect by chance. This test will give me essential insights into whether hypertension is a notable risk factor for CKD!

```
mean_age <- mean(Chronic_Kidney_Disease_data$Age, na.rm = TRUE)
sd_age <- sd(Chronic_Kidney_Disease_data$Age)

print(paste("Mean Age:", round(mean_age, 2)))
```

```
## [1] "Mean Age: 54.44"

print(paste("Standard Deviation of Age:", round(sd_age, 2)))

## [1] "Standard Deviation of Age: 20.55"
```

I’m diving into the ages of my patients! I’m calculating the mean and standard deviation of ages to get a better grasp of the age distribution in my dataset. Knowing the average age helps me understand the demographics of CKD patients, and I’m excited to see what the numbers reveal!

```
age_value <- 60

z_score <- (age_value - mean_age) / sd_age
print(paste("Z-score for age", age_value, ":", round(z_score, 2)))

## [1] "Z-score for age 60 : 0.27"
```

Next, I’m calculating the z-score for the age of 60! This tells me how many standard deviations this age is from the mean. Understanding this z-score helps identify if 60 is an age of concern when it comes to CKD risk.

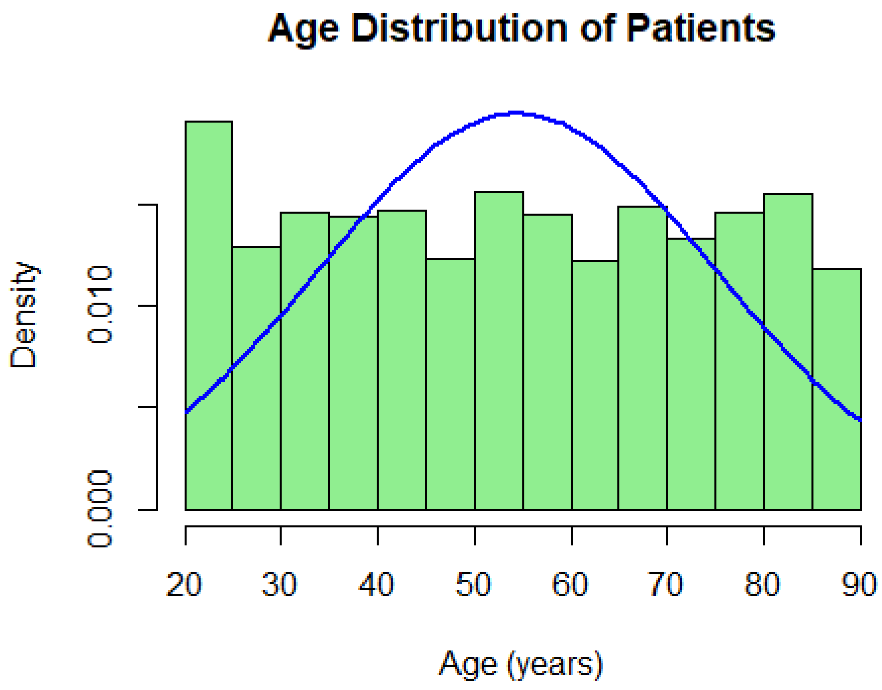
```
prob_age_over_60 <- pnorm(60, mean = mean_age, sd = sd_age, lower.tail =
FALSE)

print(paste("Probability of CKD onset for age ≥ 60:", round(prob_age_over_60,
4)))

## [1] "Probability of CKD onset for age ≥ 60: 0.3934"

prob_age_under_50 <- pnorm(50, mean = mean_age, sd = sd_age)

print(paste("Probability of CKD onset for age < 50:",
round(prob_age_under_50, 4)))

## [1] "Probability of CKD onset for age < 50: 0.4144"

z_90 <- qnorm(0.90, mean = mean_age, sd = sd_age)

print(paste("90th Percentile Age for CKD Onset:", round(z_90, 2)))

## [1] "90th Percentile Age for CKD Onset: 80.78"
```

I’m calculating the 90th percentile age for CKD onset! This helps me determine the age at which 90 percent of the patients fall below, providing a valuable threshold for assessing risk.

```
lambda_uti <- mean(Chronic_Kidney_Disease_data$UrinaryTractInfections, na.rm
= TRUE)

print(paste("Average Number of UTIs per Patient:", round(lambda_uti, 2)))

## [1] "Average Number of UTIs per Patient: 0.21"
```

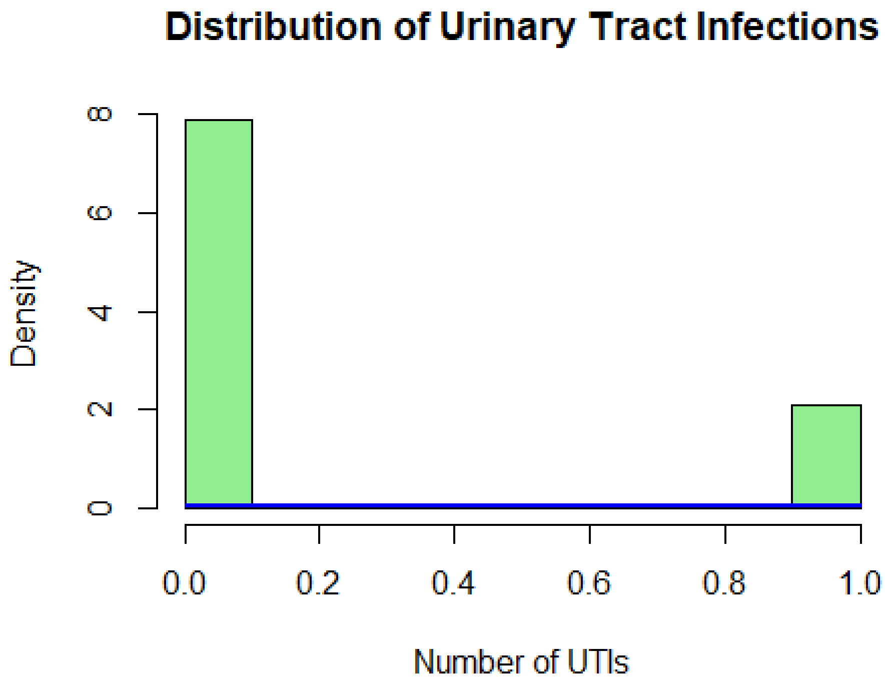
Now I’m finding the average number of urinary tract infections per patient! This information is crucial because UTIs can complicate CKD, and understanding the average helps in assessing patient care needs.

```
hist(Chronic_Kidney_Disease_data$Age, main = "Age Distribution of Patients",
xlab = "Age (years)",
     probability = TRUE, col = "lightgreen", breaks = 20)

curve(dnorm(x, mean = mean_age, sd = sd_age), add = TRUE, col = "blue", lwd =
2)
```



It’s time to visualize my findings! I’m plotting a histogram of patient ages with a normal distribution curve overlaid. This visualization helps me see trends and patterns in age distribution.

```
lambda_uti <- 4

hist(Chronic_Kidney_Disease_data$UrinaryTractInfections,
     main = "Distribution of Urinary Tract Infections",
     xlab = "Number of UTIs", breaks = 10,
     probability = TRUE, col = "lightgreen")

x_values <- 0:max(Chronic_Kidney_Disease_data$UrinaryTractInfections)

lines(x_values, dpois(x_values, lambda = lambda_uti),
      col = "blue", lwd = 2)
```



I’m visualizing the distribution of urinary tract infections! I’m plotting another histogram to understand how common UTIs are among patients. This helps me see potential health risks related to CKD.

```
prob_more_than_2_UTIs <- ppois(2, lambda = lambda_uti, lower.tail = FALSE)

print(paste("Probability of having more than 2 UTIs:",
round(prob_more_than_2_UTIs, 4)))
```

```
## [1] "Probability of having more than 2 UTIs: 0.7619"
```

Now I’m calculating the probability of patients having more than 2 UTIs! This is important for understanding infection risks and how they might impact CKD.

```
prob_fewer_than_2_UTIs <- ppois(2, lambda = lambda_uti)

print(paste("Probability of having fewer than 2 UTIs:",
round(prob_fewer_than_2_UTIs, 4)))

## [1] "Probability of having fewer than 2 UTIs: 0.2381"
```

Let’s also look at the probability of having fewer than 2 UTIs! This information will help me grasp the overall infection rates among my patients, giving me a better picture of their health.

```
prob_5_UTIs <- dpois(5, lambda = lambda_uti)

print(paste("Probability of having exactly 5 UTIs:", round(prob_5_UTIs, 4)))

## [1] "Probability of having exactly 5 UTIs: 0.1563"
```

Finally, I’m calculating the probability of a patient having exactly 5 UTIs! This precise information helps me understand specific infection rates.

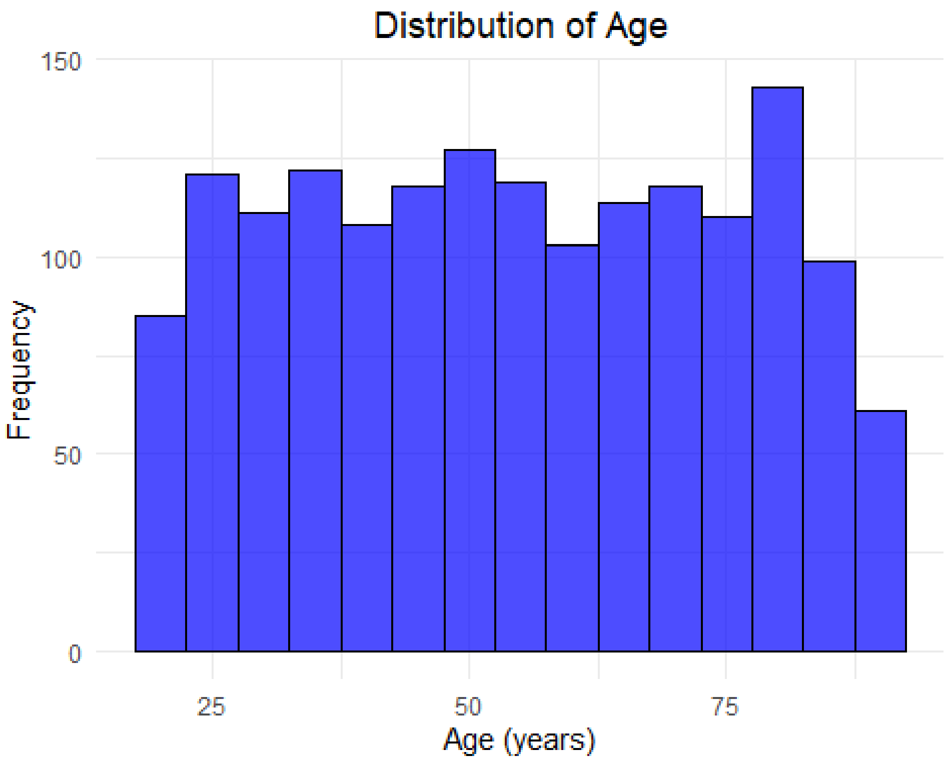
```
library(ggplot2)
library(readxl)

Chronic_Kidney_Disease_data <- read_excel("Chronic_Kidney_Dsease_data.xlsx")
```

We’re starting by loading the essential libraries and reading our dataset from an Excel file into Chronic_Kidney_Disease_data. This sets the stage for our exploration into chronic kidney disease.

```
Y <- Chronic_Kidney_Disease_data$Diagnosis
X1 <- Chronic_Kidney_Disease_data$Age
X2 <- Chronic_Kidney_Disease_data$SystolicBP
X3 <- Chronic_Kidney_Disease_data$SerumCreatinine
X4 <- Chronic_Kidney_Disease_data$HemoglobinLevels
X5 <- Chronic_Kidney_Disease_data$SerumElectrolytesSodium
```

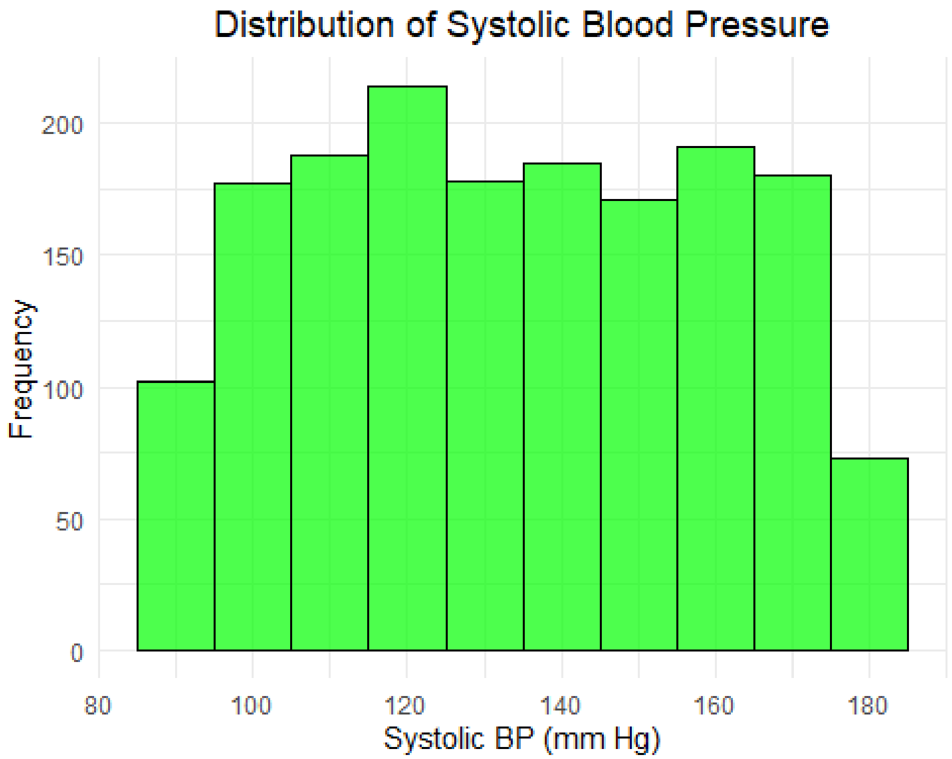
Next, we’re defining our key variables. Here, Y represents the diagnosis, while X1 through X5 are our predictors like age and blood pressure. Let’s see how these factors play into kidney health!

```r
ggplot(Chronic_Kidney_Disease_data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Age", x = "Age (years)", y = "Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



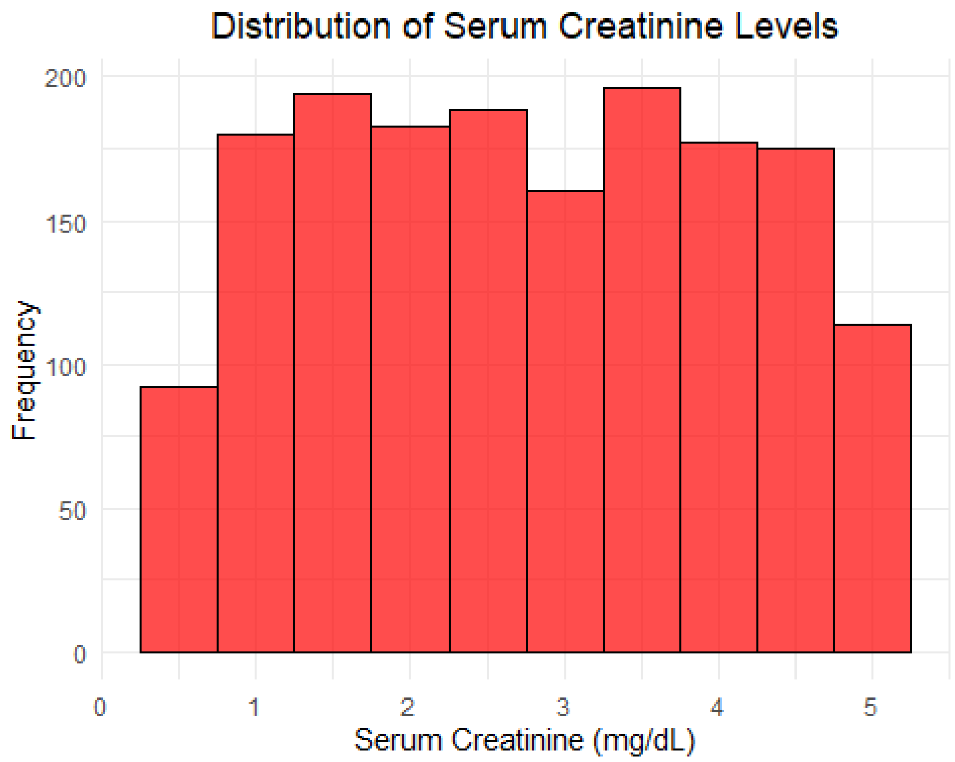
The age distribution is relatively uniform, ranging from 25 to 85 years, with a slight increase in the 70-75 year age group. This shows that chronic kidney disease affects a broad age range, not just older adults. The increase in older patients aligns with the decline in kidney function that typically occurs with age. This highlights the need for kidney health awareness and management across all adult age groups.

```r
ggplot(Chronic_Kidney_Disease_data, aes(x = SystolicBP)) +
  geom_histogram(binwidth = 10, fill = "green", color = "black", alpha = 0.7)
+
  labs(title = "Distribution of Systolic Blood Pressure", x = "Systolic BP
(mm Hg)", y = "Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



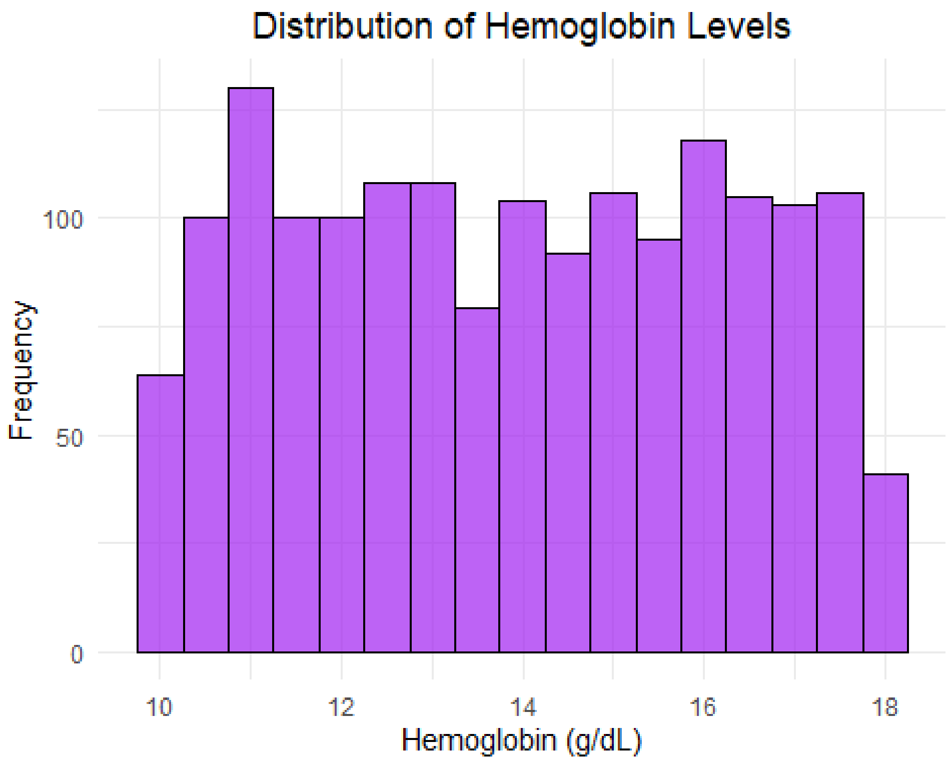
Systolic blood pressure is approximately normally distributed, with a slight right skew, ranging from 80 to 180 mm Hg. The peak around 120-130 mm Hg indicates many patients have normal or well-controlled blood pressure. However, the right skew and extended tail show that a significant number of patients have hypertension, which is common in kidney disease, highlighting the importance of managing blood pressure in this group.

```
ggplot(Chronic_Kidney_Disease_data, aes(x = SerumCreatinine)) +
  geom_histogram(binwidth = 0.5, fill = "red", color = "black", alpha = 0.7)
+
  labs(title = "Distribution of Serum Creatinine Levels", x = "Serum
Creatinine (mg/dL)", y = "Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



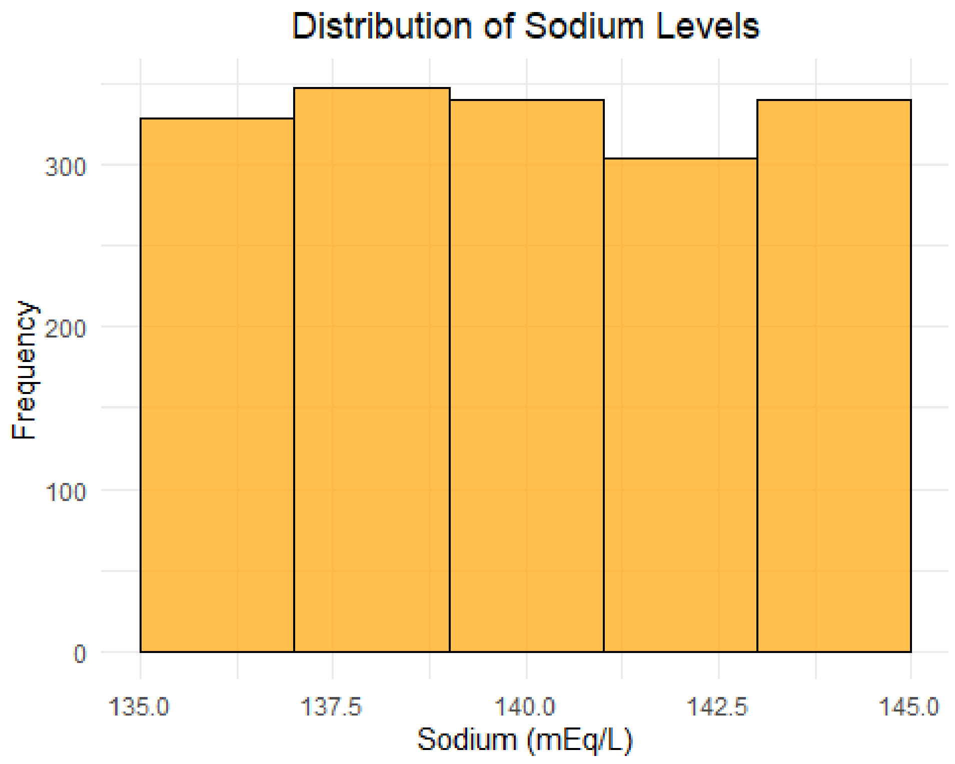
The serum creatinine levels are strongly right-skewed, ranging from 0 to 5 mg/dL. This pattern is typical for kidney disease populations. Most patients have normal to mildly elevated levels, but the long right tail indicates some patients have significantly impaired kidney function. This highlights the varying degrees of kidney dysfunction, from early to advanced chronic kidney disease.

```r
ggplot(Chronic_Kidney_Disease_data, aes(x = HemoglobinLevels)) +
  geom_histogram(binwidth = 0.5, fill = "purple", color = "black", alpha =
0.7) +
  labs(title = "Distribution of Hemoglobin Levels", x = "Hemoglobin (g/dL)",
y = "Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



The hemoglobin levels show a slight left skew, ranging from about 10 to 18 g/dL. This suggests that many patients may have lower hemoglobin levels, indicating a risk of anemia, which is common in chronic kidney disease. The peak around 11 g/dL likely represents those with mild to moderate anemia, while the peak around 16 g/dL could indicate patients with normal levels or those receiving treatment for anemia.

```
ggplot(Chronic_Kidney_Disease_data, aes(x = SerumElectrolytesSodium)) +
  geom_histogram(binwidth = 2, fill = "orange", color = "black", alpha = 0.7)
+
  labs(title = "Distribution of Sodium Levels", x = "Sodium (mEq/L)", y =
"Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



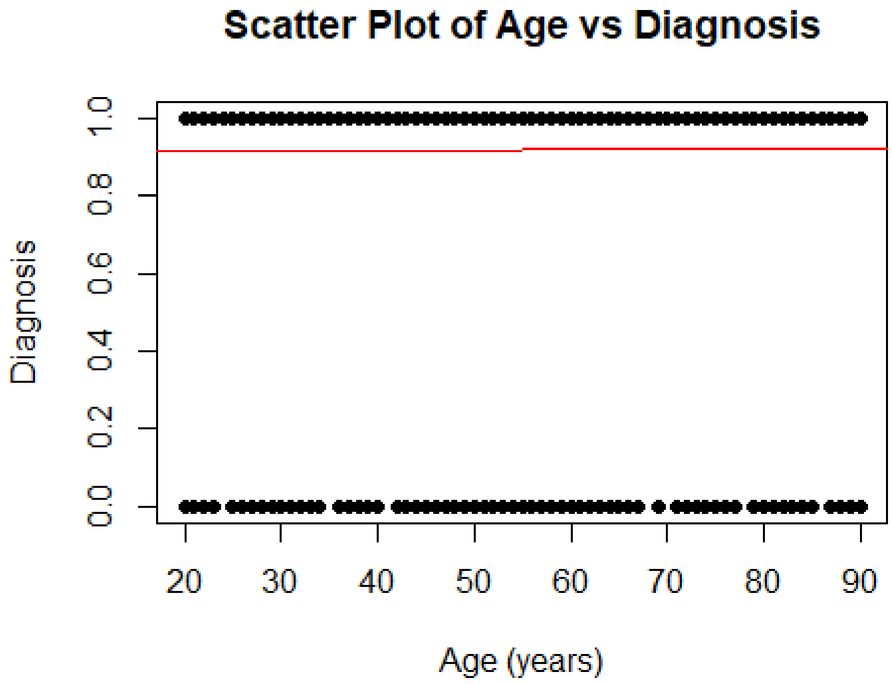
The sodium levels in the patient population are approximately normally distributed, ranging from about 135 to 145 mEq/L. This indicates that most patients have sodium levels within the normal range, suggesting that severe electrolyte imbalances are uncommon. However, the variability among patients may arise from differences in kidney function, medications, or diet.

```
correlation1 <- cor(X1, Y)
correlation2 <- cor(X2, Y)
correlation3 <- cor(X3, Y)
correlation4 <- cor(X4, Y)
correlation5 <- cor(X5, Y)

correlation_results <- data.frame(
  Variable = c("Age", "Systolic BP", "Serum Creatinine", "Hemoglobin Levels",
"Sodium Levels"),
  Correlation = c(correlation1, correlation2, correlation3, correlation4,
correlation5)
)

print(correlation_results)
```

```
##             Variable  Correlation
## 1                Age 0.0009192701
## 2        Systolic BP 0.0835281516
## 3   Serum Creatinine 0.2011245716
## 4 Hemoglobin Levels 0.0440014588
## 5      Sodium Levels 0.0323715277
```

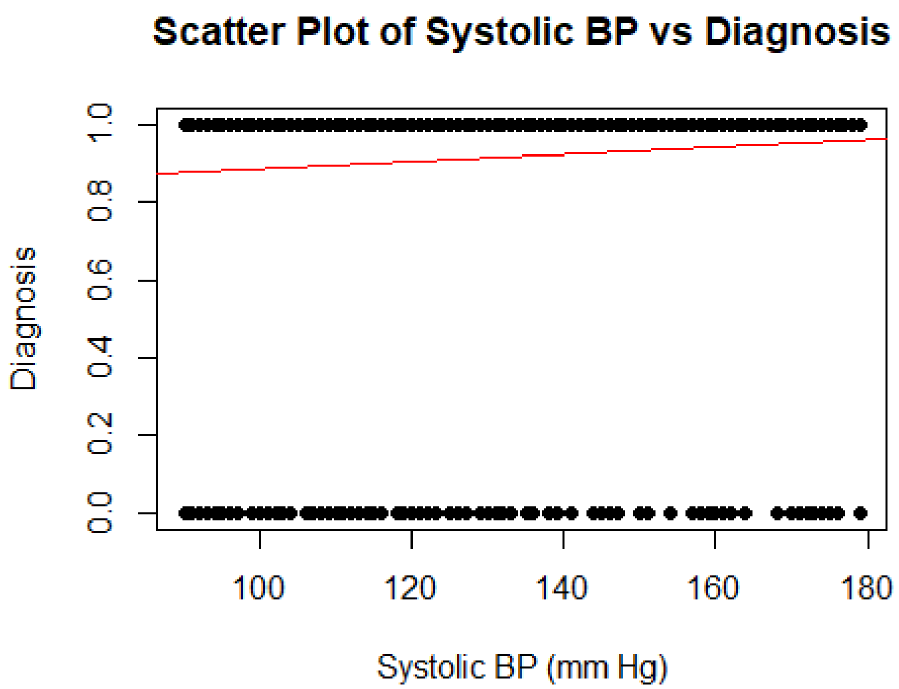
The correlation results reveal how different predictors relate to the diagnosis of chronic kidney disease (CKD). Serum Creatinine has the strongest correlation at 0.201, indicating that higher creatinine levels are linked to a greater chance of CKD diagnosis. Systolic BP has a weak correlation of 0.084, suggesting a minor association with CKD. The correlations for Age (0.001), Hemoglobin Levels (0.044), and Sodium Levels (0.032) are very low, showing little to no significant effect on CKD diagnosis. Overall, serum creatinine is the most important predictor of CKD in this analysis.

```
plot(X1, Y, main = "Scatter Plot of Age vs Diagnosis", xlab = "Age (years)",
ylab = "Diagnosis", pch = 19)
abline(lm(Y ~ X1), col = "red")
```



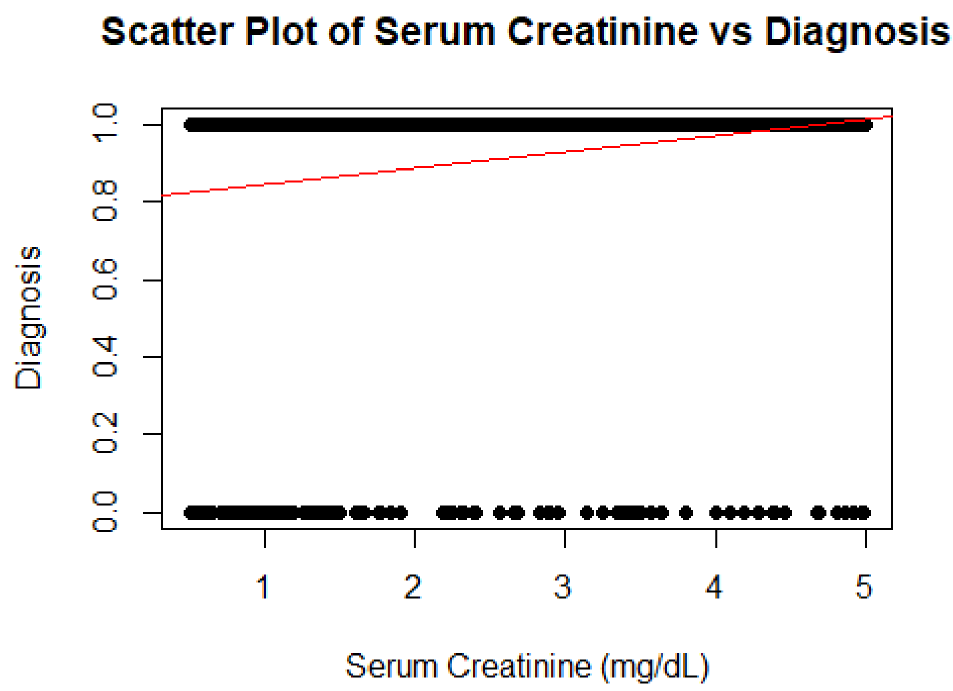
This scatter plot shows the relationship between age and CKD diagnosis. The plot reveals that all points are concentrated at 0.0 or 1.0, it indicates that the diagnosis is binary, leading to straight lines and no variation in CKD status across different ages.

```r
plot(X2, Y, main = "Scatter Plot of Systolic BP vs Diagnosis", xlab =
"Systolic BP (mm Hg)", ylab = "Diagnosis", pch = 19)
abline(lm(Y ~ X2), col = "red")
```



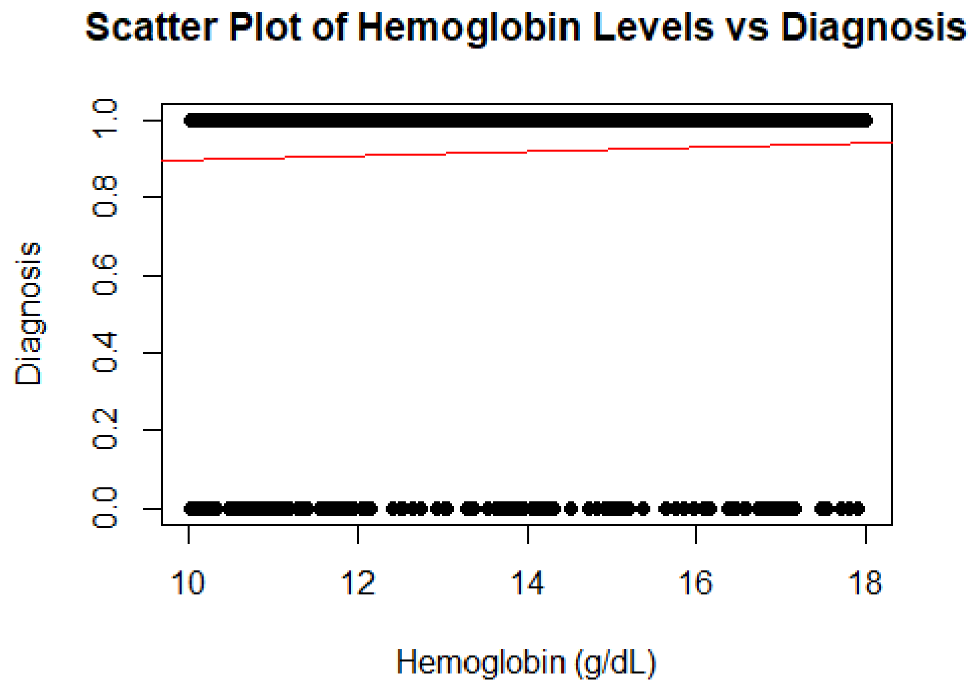
This plot illustrates the relationship between systolic blood pressure and CKD diagnosis. Similar to the age plot, if points are limited to 0.0 or 1.0, it suggests a binary diagnosis, resulting in a lack of variation and straight lines.

```r
plot(X3, Y, main = "Scatter Plot of Serum Creatinine vs Diagnosis", xlab =
"Serum Creatinine (mg/dL)", ylab = "Diagnosis", pch = 19)
abline(lm(Y ~ X3), col = "red")
```



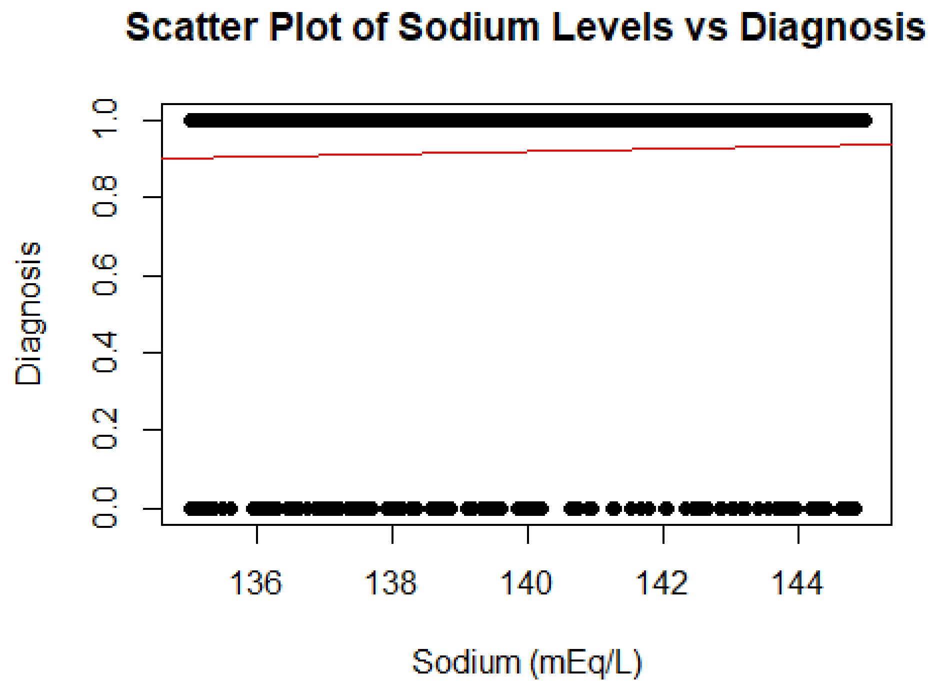
This scatter plot depicts the relationship between serum creatinine levels and CKD diagnosis. The data points cluster around 0.0 or 1.0, it indicates the binary nature of the diagnosis, making it challenging to see how creatinine levels influence CKD status.

```
plot(X4, Y, main = "Scatter Plot of Hemoglobin Levels vs Diagnosis", xlab =
"Hemoglobin (g/dL)", ylab = "Diagnosis", pch = 19)
abline(lm(Y ~ X4), col = "red")
```



This plot shows the relationship between hemoglobin levels and CKD diagnosis. A concentration of points at 0.0 or 1.0 suggests that the hemoglobin levels do not vary in their impact on the binary diagnosis, resulting in straight lines.

```
plot(X5, Y, main = "Scatter Plot of Sodium Levels vs Diagnosis", xlab =
"Sodium (mEq/L)", ylab = "Diagnosis", pch = 19)
abline(lm(Y ~ X5), col = "red")
```



This scatter plot represents the relationship between serum sodium levels and CKD diagnosis. Again, if points are primarily at 0.0 or 1.0, it indicates a binary outcome, limiting the ability to assess how sodium levels relate to CKD status, which leads to straight lines in the plot.