STAT 382: Project 1 — Medical Data Analysis

Amir Syrtlanov

October 24, 2025

Project 1 Report

This document contains all Tasks (1–9) for STAT 382 Project 1 using `medicaldata1.csv` .

Task 1: R Markdown Setup and Data Import

I will knit this R Markdown file to an **HTML document**. To create the **PDF** for submission, I will open the HTML in my web browser and use **File → Print → Save as PDF**.

The **`.Rmd`** file is the *source* with code and explanations. The **PDF** is the *final report* showing formatted text, results, tables, and graphs (with code hidden when appropriate).

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE, fig.width = 7, fig.height = 4.5)


medicaldata <- read.csv("medicaldata1.csv")


head(medicaldata)
```

```
##   Age Height Weight Gallstone_Status Comorbidity CAD Diabetes_Mellitus
## 1  57    170  143.5                0           0   0                 0
## 2  53    150  108.4                1           1   0                 1
## 3  54    155   80.2                1           0   0                 0
## 4  63    150   69.0                1           0   0                 0
## 5  73    160  104.6                1           1   0                 1
## 6  56    158  100.5                0           0   0                 0
##   Total_Body_Water Total_Body_Fat_Ratio Lean_Mass Body_Protein_Content
## 1             66.2                42.30     57.70                 7.99
## 2             40.0                50.92     48.99                 9.05
## 3             32.8                46.13     53.87                 9.63
## 4             29.9                42.46     57.39                10.49
## 5             52.7                34.20     65.77                11.31
## 6             35.9                49.80     50.25                11.41
##   Hepatic_Fat_Accumulation Glucose Triglyceride
## 1                        3     118          112
## 2                        2     119           97
## 3                        2     111          104
## 4                        2     106           96
## 5                        3     230          208
## 6                        3     129           81
```

Task 2: Convert Categorical Variables to Factors

```
medicaldata$Gallstone_Status <- factor(medicaldata$Gallstone_Status,
levels = c(0, 1),
labels = c("Yes", "No"))



medicaldata$Comorbidity <- factor(medicaldata$Comorbidity,
levels = c(0,1,2,3),
labels = c("None","One","Two","Three+"),
ordered = TRUE)



medicaldata$CAD <- factor(medicaldata$CAD,
levels = c(0,1),
labels = c("No","Yes"))



medicaldata$Diabetes_Mellitus <- factor(medicaldata$Diabetes_Mellitus,
levels = c(0,1),
labels = c("No","Yes"))



medicaldata$Hepatic_Fat_Accumulation <- factor(medicaldata$Hepatic_Fat_Accumulation,
levels = c(0,1,2,3),
labels = c("None","Mild","Moderate","Severe"),
ordered = TRUE)

# Quick structure check

str(medicaldata)
```

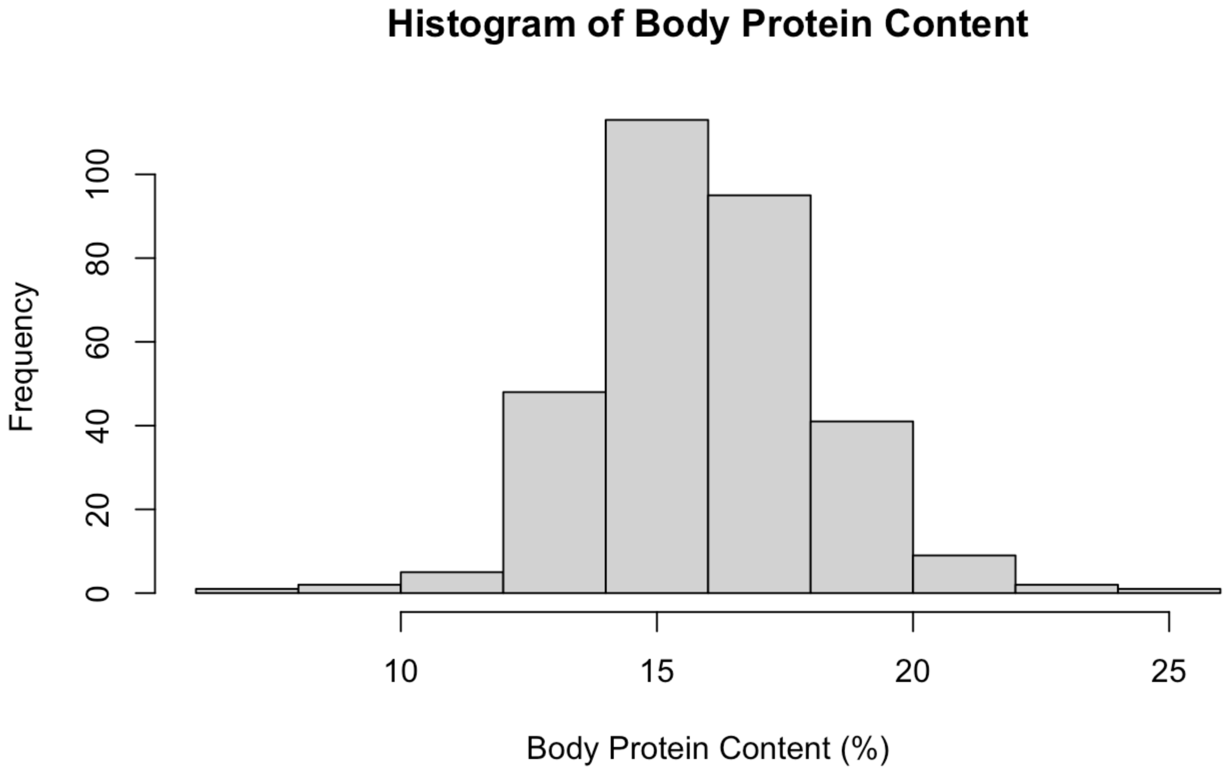
```
## 'data.frame':    317 obs. of  14 variables:
##  $ Age                     : int  57 53 54 63 73 56 48 59 34 46 ...
##  $ Height                  : int  170 150 155 150 160 158 162 147 155 165 ...
##  $ Weight                  : num  143.5 108.4 80.2 69 104.6 ...
##  $ Gallstone_Status        : Factor w/ 2 levels "Yes","No": 1 2 2 2 2 1 2 2 1 2 ...
##  $ Comorbidity             : Ord.factor w/ 4 levels "None"<"One"<"Two"<..: 1 2 1 1 2
1 1 1 1 1 ...
##  $ CAD                     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Diabetes_Mellitus       : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 1 1 1 ...
##  $ Total_Body_Water        : num  66.2 40 32.8 29.9 52.7 35.9 44 34.2 35.1 38.4 ...
##  $ Total_Body_Fat_Ratio    : num  42.3 50.9 46.1 42.5 34.2 ...
##  $ Lean_Mass               : num  57.7 49 53.9 57.4 65.8 ...
##  $ Body_Protein_Content    : num  7.99 9.05 9.63 10.49 11.31 ...
##  $ Hepatic_Fat_Accumulation: Ord.factor w/ 4 levels "None"<"Mild"<..: 4 3 3 3 4 4 3 3
4 1 ...
##  $ Glucose                 : num  118 119 111 106 230 129 99 97 111 91 ...
##  $ Triglyceride            : num  112 97 104 96 208 81 89 106 93 122 ...
```

Task 3: Quantitative Variable — Body Protein Content

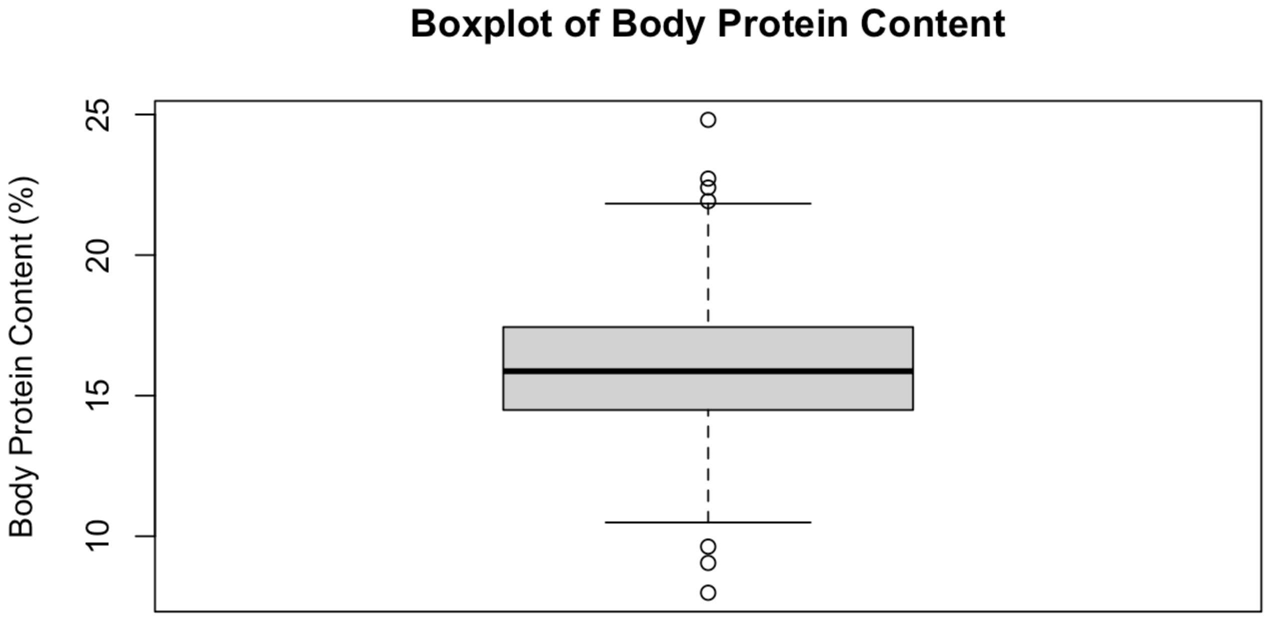
```
# Missing values

sum(is.na(medicaldata$Body_Protein_Content))
```

```
## [1] 0
```

```
# Histogram

hist(medicaldata$Body_Protein_Content,
main = "Histogram of Body Protein Content",
xlab = "Body Protein Content (%)")
```



```
# Boxplot

boxplot(medicaldata$Body_Protein_Content,
main = "Boxplot of Body Protein Content",
ylab = "Body Protein Content (%)")
```



```
# Descriptive statistics

bpc_summary <- summary(medicaldata$Body_Protein_Content)
bpc_sd <- sd(medicaldata$Body_Protein_Content, na.rm = TRUE)
list(Summary = bpc_summary, SD = bpc_sd)
```

```
## $Summary
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    7.99   14.49   15.87   15.97   17.44   24.81 
## 
## $SD
## [1] 2.267927
```

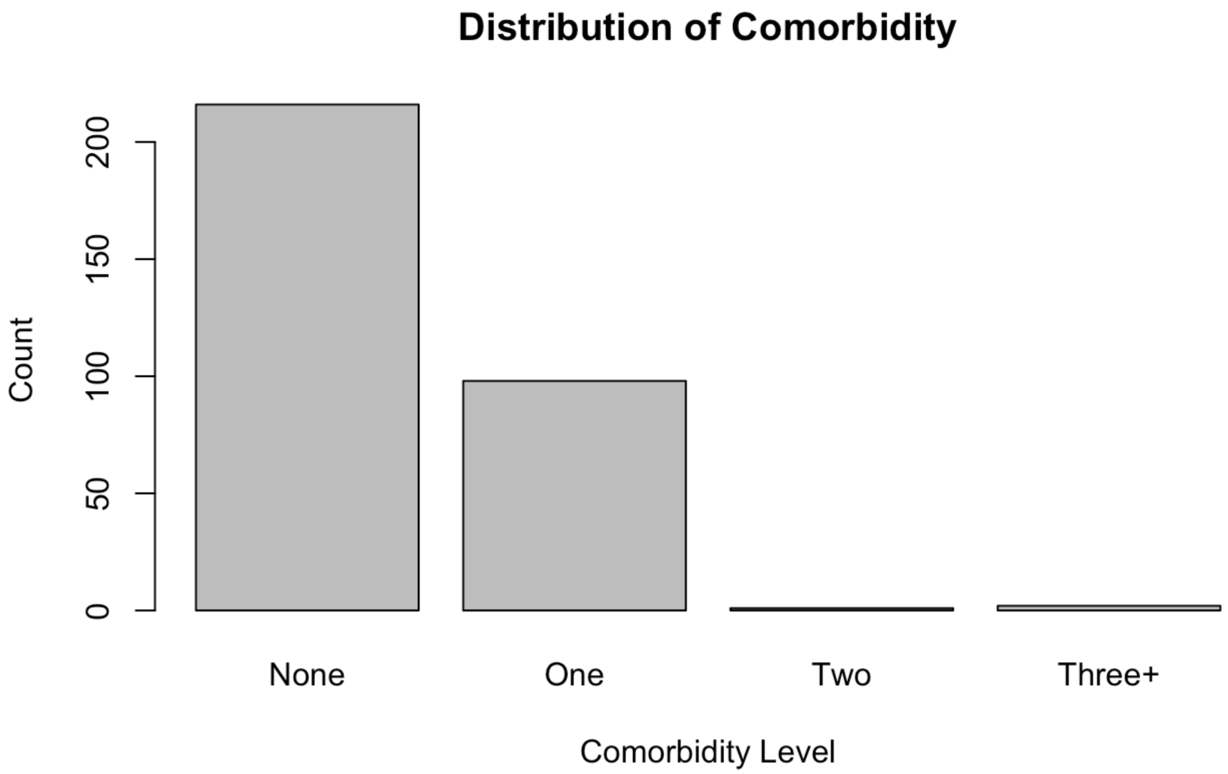
```
#Interpretation:
#The histogram shows Body Protein Content to be approximately symmetric with mean approximately 16% and range approximately 8% to 25%.
#It also indicates the presence of some minor outliers on the higher side, reflecting slightly right-skewed data.
#Overall, the variation is moderate (SD ≈ 2.3%), thus most of the participants have similar protein levels.
```

Task 4: Categorical Variable — Comorbidity

```
comorb_tbl <- table(medicaldata$Comorbidity)
comorb_tbl
```

```
##
##   None    One    Two Three+
##    216     98      1      2
```

```
# Bar plot

barplot(comorb_tbl,
main = "Distribution of Comorbidity",
xlab = "Comorbidity Level",
ylab = "Count")
```



```
#Interpretation:
#The bar chart and frequency table both show that the majority of participants have no c
omorbidities, with fewer and fewer as more comorbidities there are.
#This indicates that the majority of the population in the dataset is very healthy, whil
e there are few individuals with more than two comorbidities.
#Distribution is heavily right-skewed, highlighting that high levels of comorbidity are
rare among this sample.
```

Task 5: Gallstone Status vs Hepatic Fat

Accumulation

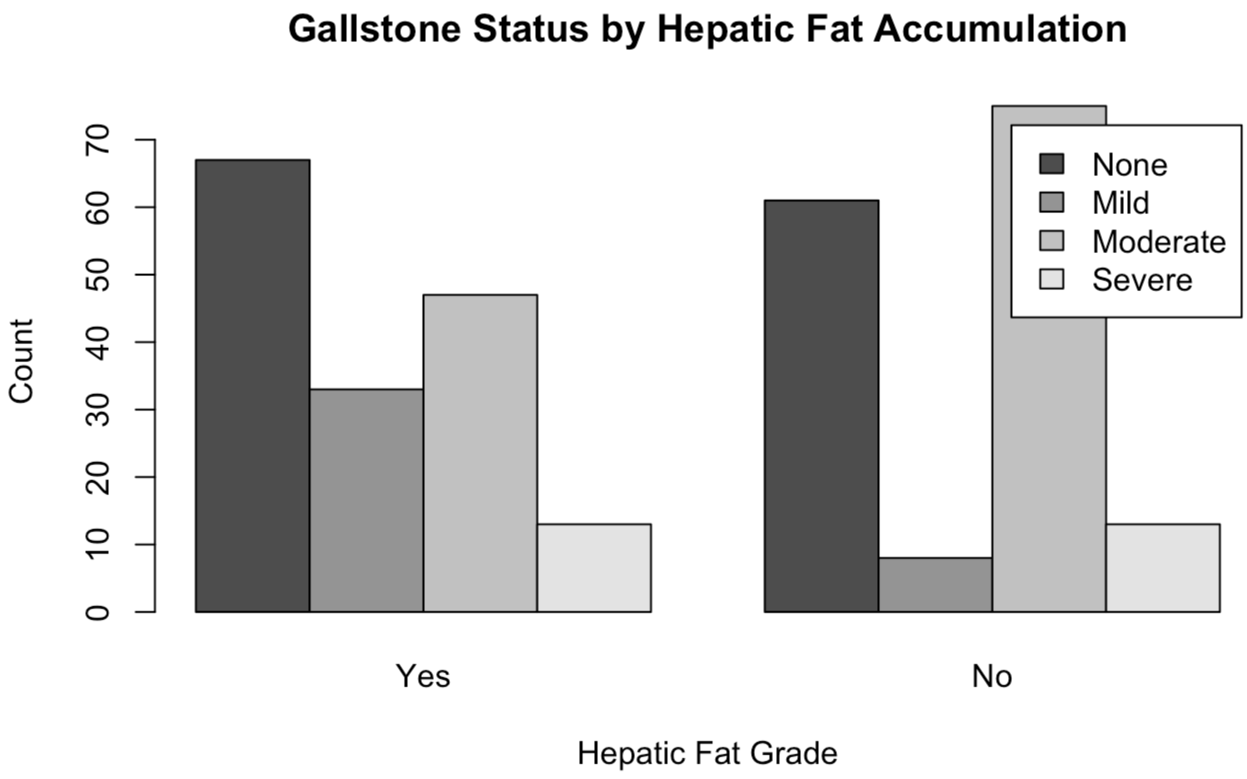
```
tab_GH <- table(Hepatic_Fat = medicaldata$Hepatic_Fat_Accumulation,
Gallstones = medicaldata$Gallstone_Status)
tab_GH
```

```
##             Gallstones
## Hepatic_Fat Yes No
##    None      67 61
##    Mild      33  8
##    Moderate  47 75
##    Severe    13 13
```

```
# Row-wise (conditional on hepatic fat grade) proportions

prop_GH <- prop.table(tab_GH, margin = 1)
prop_GH
```

```
##             Gallstones
## Hepatic_Fat       Yes        No
##    None     0.5234375 0.4765625
##    Mild     0.8048780 0.1951220
##    Moderate 0.3852459 0.6147541
##    Severe   0.5000000 0.5000000
```

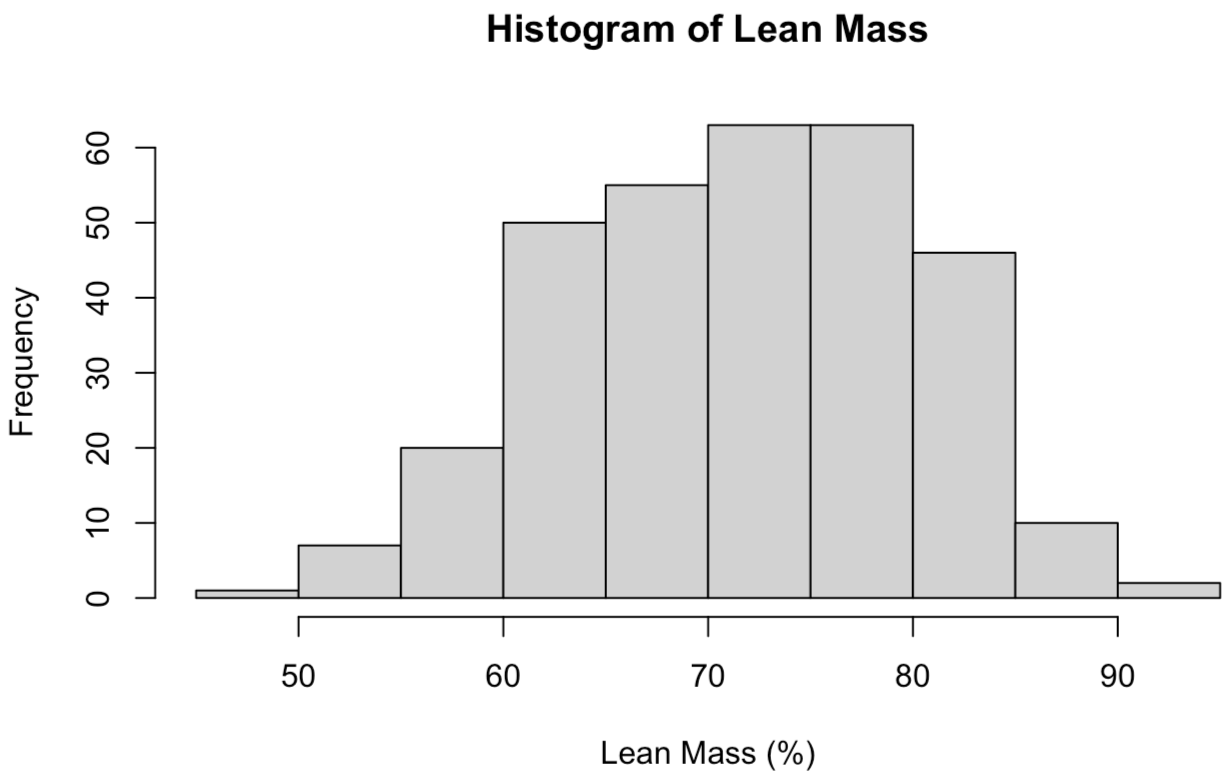
```
# Side-by-side barplot of counts

barplot(tab_GH, beside = TRUE, legend = TRUE,
main = "Gallstone Status by Hepatic Fat Accumulation",
xlab = "Hepatic Fat Grade", ylab = "Count")
```


Gallstone Status by Hepatic Fat Accumulation
Count
0
10
20
30
40
50
60
70
Yes
No
Hepatic Fat Grade
None
Mild
Moderate
Severe

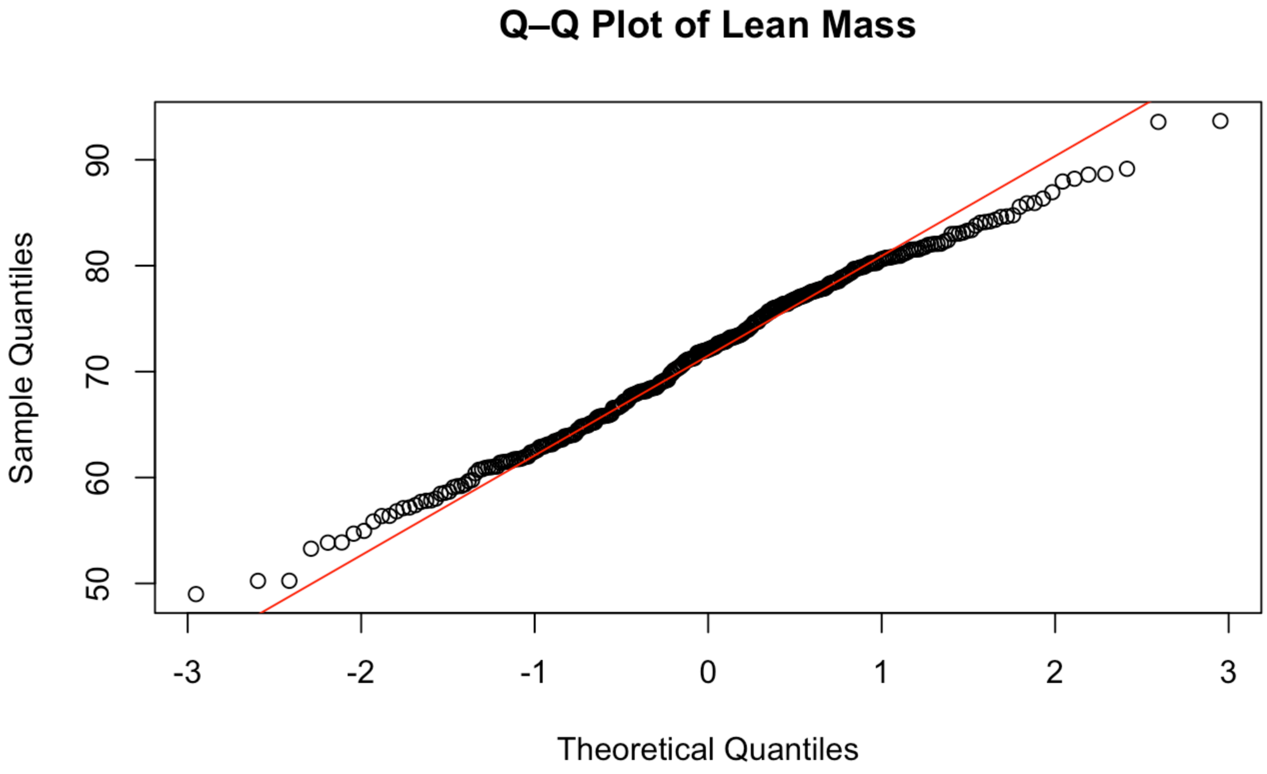
```
# Moment-based estimators without external packages

sample_skewness <- function(x) {
x <- x[is.finite(x)]
n <- length(x); m <- mean(x); s <- sd(x)
if (n < 3 || s == 0) return(NA_real_)
sum(((x - m)/s)^3) * (n / ((n - 1)*(n - 2)))
}

sample_kurtosis <- function(x) {

# Returns kurtosis where normal ≈ 3

x <- x[is.finite(x)]
n <- length(x); m <- mean(x); s <- sd(x)
if (n < 4 || s == 0) return(NA_real_)
num <- sum(((x - m)/s)^4) * (n*(n+1)) / ((n-1)*(n-2)*(n-3))
adj <- 3 * ((n-1)^2) / ((n-2)*(n-3))
num - adj + 3
}



#Interpretation:
#The bar chart and table illustrate the fact that gallstones are most common among participants with mild hepatic fat content.
#As the intensity of hepatic fat increases, gallstone prevalence is not correspondingly increased, showing merely a weak association.
#Overall, the visual data indicate that gallstone status and hepatic fat accumulation are not very closely related.
```

Task 6: Normality of Lean Mass

```
lm <- medicaldata$Lean_Mass

# Histogram

hist(lm, main = "Histogram of Lean Mass", xlab = "Lean Mass (%)")
```



```
# Q–Q plot

qqnorm(lm, main = "Q–Q Plot of Lean Mass")
qqline(lm, col = "red")
```



```
skew_lm <- sample_skewness(lm)
kurt_lm <- sample_kurtosis(lm)
list(Skewness = skew_lm, Kurtosis = kurt_lm)
```

```
## $Skewness
## [1] -0.1185574
##
## $Kurtosis
## [1] 2.508279
```

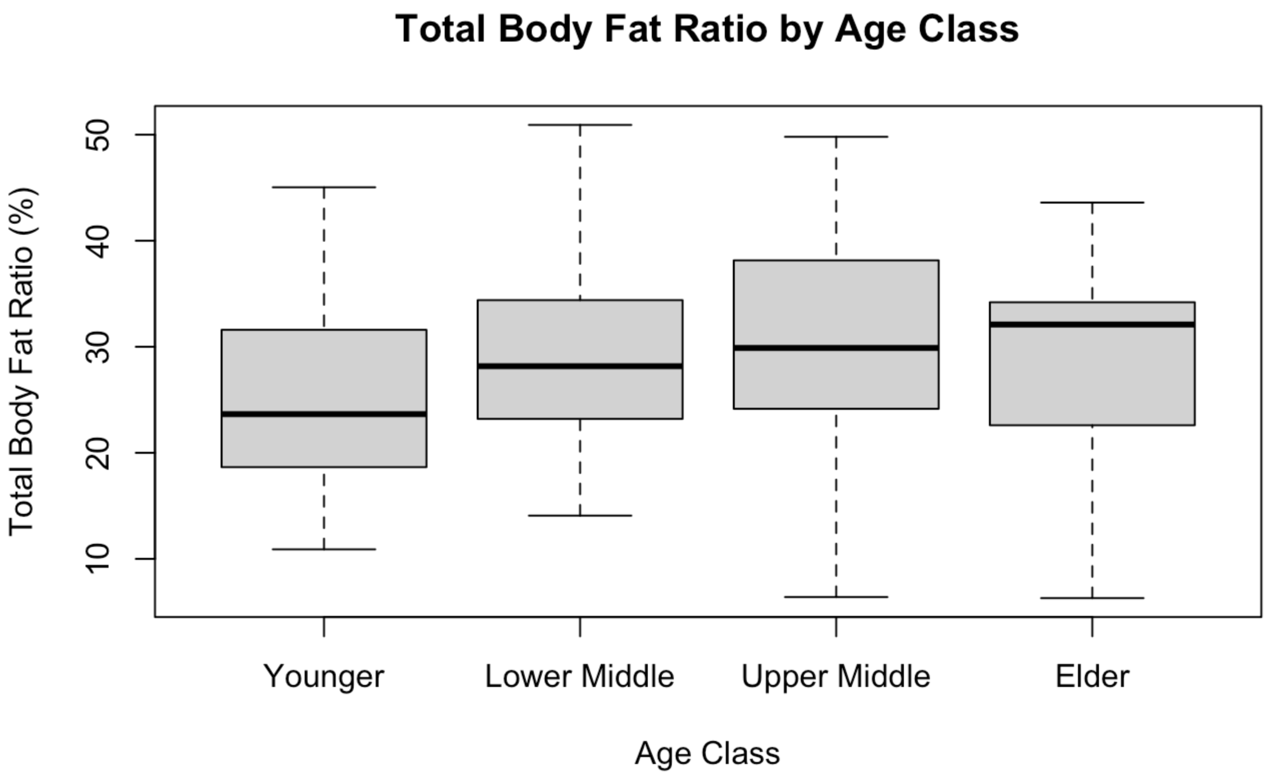
```
shapiro.test(lm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm
## W = 0.99178, p-value = 0.07566
```

```
#Interpretation:
#Q-Q plot and histogram show Lean Mass to be nearly bell-shaped with little deviation from the diagonal line, pointing towards near-normality.
#Both the skewness and kurtosis values are close to those of the normal distribution, and the Shapiro-Wilk test p-value is greater than 0.03.
#We cannot, therefore, reject the null hypothesis and conclude that Lean Mass can be approximated as normally distributed for this sample.
```

Task 7: Age Class and Total Body Fat Ratio

```
medicaldata$Age_Class <- cut(
medicaldata$Age,
breaks = c(-Inf, 40, 55, 65, Inf),
labels = c("Younger", "Lower Middle", "Upper Middle", "Elder"),
ordered_result = TRUE
)

# Boxplots

boxplot(Total_Body_Fat_Ratio ~ Age_Class, data = medicaldata,
main = "Total Body Fat Ratio by Age Class",
xlab = "Age Class", ylab = "Total Body Fat Ratio (%)")
```



```
# Group summaries: mean and sd per age class

aggregate(Total_Body_Fat_Ratio ~ Age_Class, data = medicaldata,
FUN = function(x) c(mean = mean(x), sd = sd(x)))
```

```
##      Age_Class Total_Body_Fat_Ratio.mean Total_Body_Fat_Ratio.sd
## 1      Younger                 25.192841                8.231202
## 2 Lower Middle                 29.099787                7.583341
## 3 Upper Middle                 30.559552                9.214016
## 4        Elder                 28.670952                9.535348
```

```
#Interpretation:
#The boxplot indicates that Total Body Fat Ratio tends to rise with age up to the "Upper
Middle" group, but then levels off slightly for "Elder" participants.
#The summary statistics confirm that the older groups of patients have a greater mean bo
dy fat percentage than the younger groups of patients.
#This reflects a negative weak correlation between age and ratio of body fat, that is, a
s age goes on, body fat rises but tends to stabilize later in life.
```

Task 8: Hypothesis Test — Mean Lean Mass > 70

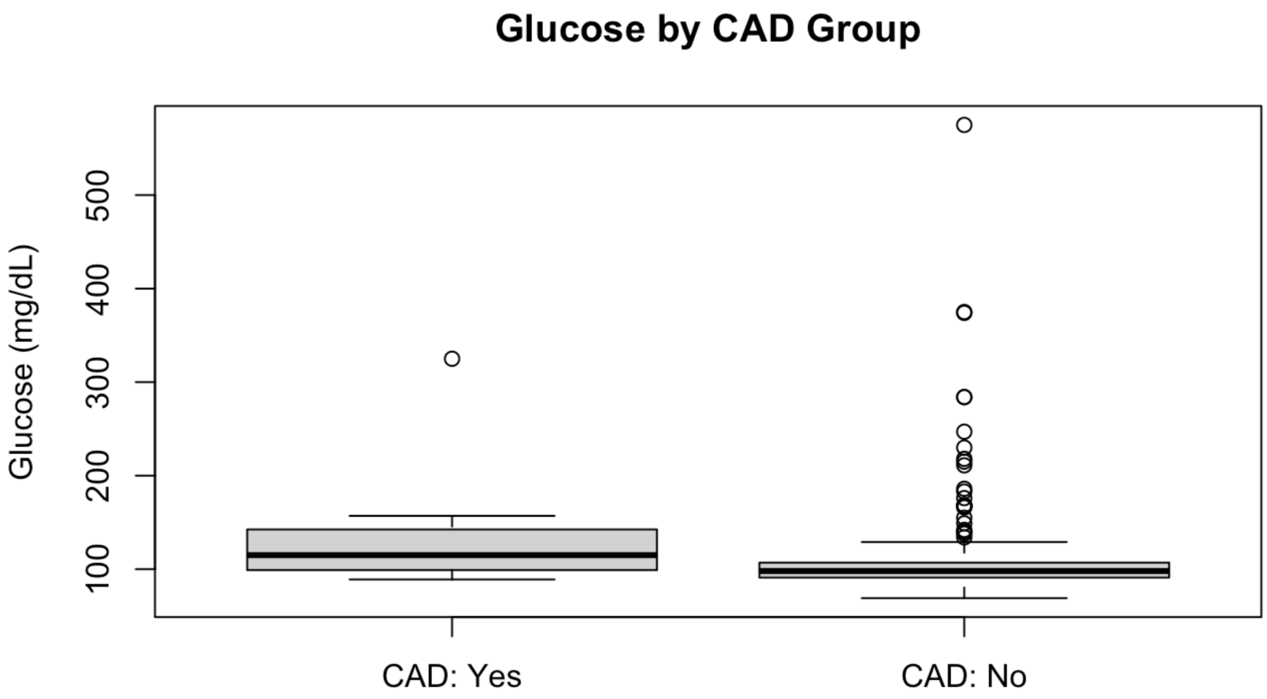
```
t_res_lm <- t.test(medicaldata$Lean_Mass,
mu = 70,
alternative = "greater",
conf.level = 0.94)
t_res_lm
```

```
##
##  One Sample t-test
##
## data:  medicaldata$Lean_Mass
## t = 3.4041, df = 316, p-value = 0.0003747
## alternative hypothesis: true mean is greater than 70
## 94 percent confidence interval:
##  70.87668      Inf
## sample estimates:
## mean of x
##  71.61741
```

```
#Interpretation:
#One-sample t-test for Lean Mass gave a p-value smaller than the significance level of 0.06, i.e., the sample mean is significantly greater than 70%.
#The 94% confidence interval for the mean is entirely above 70%, contributing to this conclusion.
#Therefore, we reject the null hypothesis and deduce that, on average, participants have a mean Lean Mass percentage greater than 70%.
```

Task 9: Hypothesis Test — Glucose by CAD Status

```
glucose_yes <- medicaldata$Glucose[medicaldata$CAD == "Yes"]
glucose_no  <- medicaldata$Glucose[medicaldata$CAD == "No"]

# Quick visual

boxplot(glucose_yes, glucose_no, names = c("CAD: Yes", "CAD: No"),
main = "Glucose by CAD Group", ylab = "Glucose (mg/dL)")
```



```
# Variance test (info only)

var.test(glucose_yes, glucose_no)
```

```
##
##  F test to compare two variances
##
## data:  glucose_yes and glucose_no
## F = 2.3002, num df = 10, denom df = 305, p-value = 0.02581
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.100282 7.139052
## sample estimates:
## ratio of variances
##            2.30021
```

```
t_two <- t.test(glucose_yes, glucose_no,
alternative = "two.sided",
var.equal = FALSE,
conf.level = 0.99)  # 99% CI to match alpha = 0.01
t_two
```

```
## 
##  Welch Two Sample t-test
## 
## data:  glucose_yes and glucose_no
## t = 1.4289, df = 10.315, p-value = 0.1826
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -34.72758  92.46751
## sample estimates:
## mean of x mean of y 
##  136.5455  107.6755
```

```
#Interpretation:
#The boxplot shows that the glucose levels are very similar for CAD and non-CAD participants.
#The variance test indicates unequal variances, so the Welch two-sample t-test was used.
#The p-value found is greater than the significance level 0.01, and thus we fail to reject the null hypothesis.
#In this data, there is no notable difference in the mean glucose level among CAD and non-CAD groups.
```