University of Stirling MATPMDA

Computing Science & Mathematics 2023

**MATPMDA MATHEMATICAL AND STATISTICAL FOUNDATIONS**

**PROJECT : AUTUMN SEMESTER 2023**

**Submission due 18th December 17:00**

**Student Number: <Student id>**

**Declaration: In submitting this project I declare that this is all my own work and I did not seek help to complete it.**

**For each project question, insert answers below.**

1. Perform an exploratory data analysis, taking care to describe the type of variables in the data set.

**Exploratory Data Analysis:**

1. The dataset provided consists of LBM (lean body mass) in kgs & BMI (body mass index), the variables are **Quantitative & Continuous,** the summary of data, as taken from the R code is as below:

Sex LBM BMI

Length:178 Min. :43.016 Min. :17.090

Class :character 1st Qu.:52.286 1st Qu.:20.154

Mode :character Median :58.484 Median :21.371

Mean :60.151 Mean :21.615

3rd Qu.:66.083 3rd Qu.:23.170

Max. :86.240 Max. :26.115

The mean LBM of the given data is 60.151, with the median of 58.484, the mean for BMI is 21.615 with median ranging around 21.371, these are measures of *central tendency*, one other measure for central tendency is *mode,* also known as the most frequently occurred value in the sample (University of Stirling, n.d.), for our data the mode is calculated at 64.417 for LBM and 22.186 for BMI.

1. **Visualizations:** For visualization of Quantitative data firstly the “Stem & Leaf plot” is used (University of Stirling, n.d.).

**BMI**

1 | 2: represents 1.2

leaf unit: 0.1

n: 178

5 17\* | 00333

10 17. | 57777

12 18\* | 22

15 18. | 999

25 19\* | 0111333344

36 19. | 55666779999

57 20\* | 000111111111111233344

69 20. | 555555777788

(27) 21\* | 000000112222223333333334444

82 21. | 7799

78 22\* | 00001111223333

64 22. | 6666777788899

51 23\* | 00111111112222

37 23. | 55666666778899

23 24\* | 003444444

14 24. | 555566

8 25\* | 44

6 25. | 5577

2 26\* | 11

**LBM**

1 | 2: represents 12

leaf unit: 1

n: 178

6 t | 333333

9 f | 555

12 s | 677

25 4. | 8888899999999

39 5\* | 00000000111111

54 t | 222222222223333

69 f | 444444444555555

85 s | 6666666666667777

(12) 5. | 888888999999

81 6\* | 00000000111

70 t | 222222233333

58 f | 444444455555

46 s | 666666777

37 6. | 8889999999

27 7\* | 0001111

20 t | 233

17 f | 5

16 s | 667

13 7. | 889

10 8\* | 001

7 t | 23

5 f | 4455

1 s | 6

The middle value for BMI’s lies on stem 21, that has 27 observations in it, and the middle value of LBM’s lie on stem>5, that has 12 observations in it.

1. **Separating the Data:** To compare data across genders, we separate the dataset for male and female and do the visualizations of the data to see the graphical representations.

Step 3, 4, 5 & 6 combined shows the *univariate analysis* of the data consisting of: descriptive statistics, measures of central tendency, measures of dispersion & visualizations of the data.

For the multivariate analysis, we will look at the scatterplots across genders, correlation coefficient & cluster analysis through stem & leaf backpack.

**Summary:**

**Male:**

Sex LBM BMI

Length:97 Min. :45.504 Min. :17.090

Class :character 1st Qu.:59.222 1st Qu.:20.124

Mode :character Median :65.481 Median :21.354

Mean :65.563 Mean :21.501

3rd Qu.:70.962 3rd Qu.:23.134

Max. :86.240 Max. :26.115

**Female:**

Sex LBM BMI

Length:81 Min. :43.016 Min. :17.090

Class :character 1st Qu.:50.549 1st Qu.:20.178

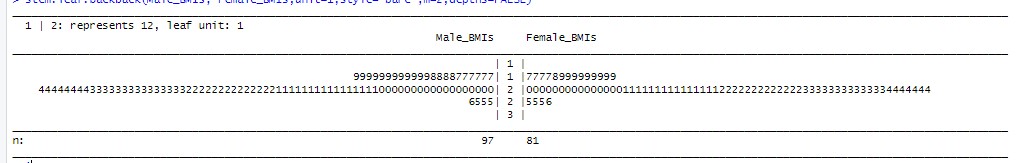
Mode :character Median :53.679 Median :21.448

Mean :53.670 Mean :21.751

3rd Qu.:56.435 3rd Qu.:23.182

Max. :65.354 Max. :26.115

**Stem & Leaf Backpack**



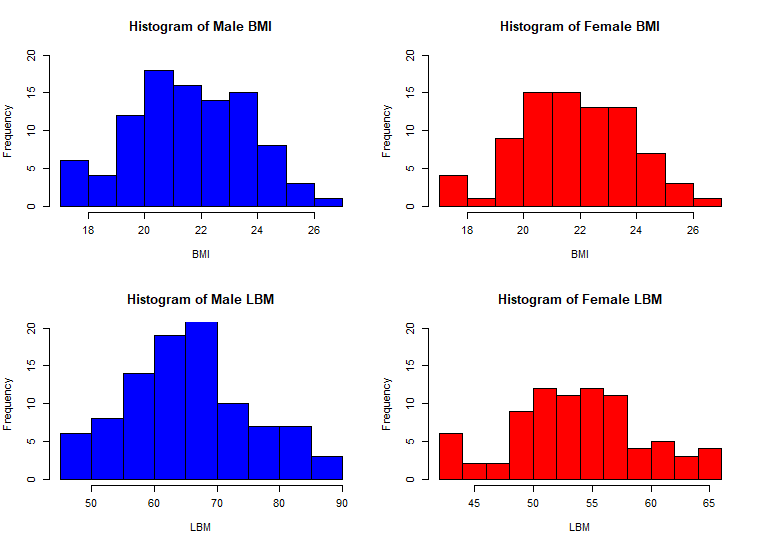
In the male BMI section, we can see a cluster of values around 25. There are also fewer values at the extremes, such as very low or very high BMIs.

In the female BMI section, there's a similar clustering around 25, but the distribution appears slightly different compared to males. Additionally, there are fewer extreme values.

**Histograms**

The second used method for visualization for quantitative data is Histogram (University of Stirling, n.d.).

Below are the histogram plots for both BMI’s and LBM’s across genders:



BMI Distribution:

Both male and female BMI distributions exhibit a roughly symmetric shape, with a peak around specific values.

The symmetry suggests that the majority of individuals fall within a certain BMI range.

Males have a broader spread, indicating greater variability in BMI values.

Females show a slightly narrower distribution, centered on a lower BMI value.

LBM Distribution:

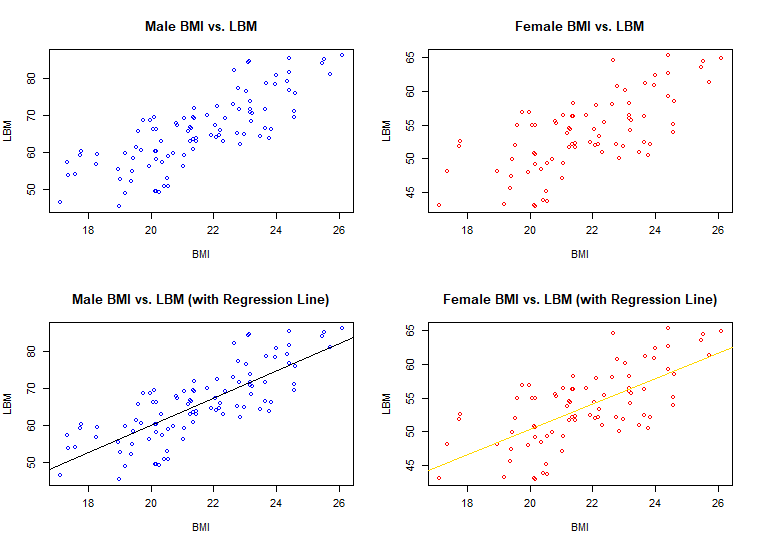
Lean Body Mass (LBM) distributions also display symmetry, but with distinct centers.

Males have higher LBM values, centered on 60.

Females exhibit lower LBM values, centered around 50.

The symmetry implies that both genders have a typical LBM range, but males tend to have leaner mass overall.

**Scatterplots:**



Regression Line Error:

The regression line represents the best-fit linear relationship between BMI and LBM.

The distance of each data point from this line indicates the error or residual.

Larger distances imply greater deviation from the predicted LBM based on BMI.

The variability in these distances highlights the imprecision of using BMI alone to predict LBM.

Scatter of Data Points:

The scatter around the regression line shows the spread or dispersion of data.

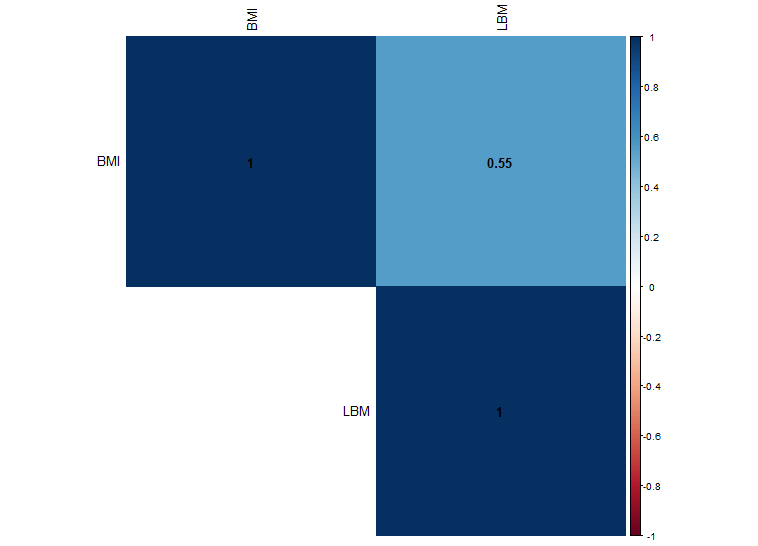
Both male and female data points exhibit variability. Female data points are relatively more scattered than that of males’.

Some points are close to the line, while others deviate significantly.

This scatter suggests that BMI is not a perfect predictor of LBM; other factors contribute to individual variations.

**Correlation Heatmap:**

Combined:



The correlation heatmap visually represents the relationships between different variables. In this case, it shows the correlation between Body Mass Index (BMI) and Lean Body Mass (LBM).

The key points from the heatmap:

Color Gradient:

The heatmap uses colors to indicate the strength and direction of correlation.

Dark red represents negative correlation (as one variable increases, the other decreases).

Dark blue represents positive correlation (both variables increase or decrease together).

Lighter shades indicate weaker correlations.

Correlation Value:

The numeric values within each cell indicate the correlation coefficient.

A value close to 1 (positive) or -1 (negative) indicates a strong relationship.

A value close to 0 suggests a weak or no linear relationship.

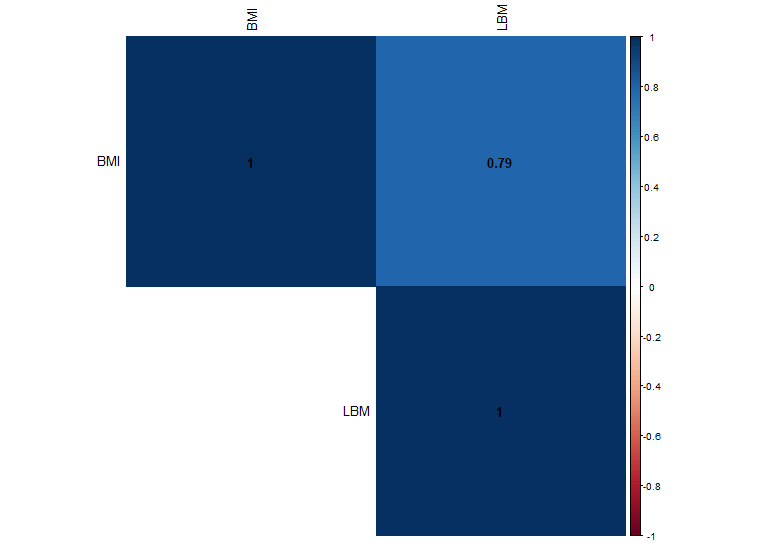
Interpretation:

The heatmap reveals that BMI and LBM have a moderate positive correlation (around 0.55).

This means that as BMI increases, LBM tends to increase as well (and vice versa).

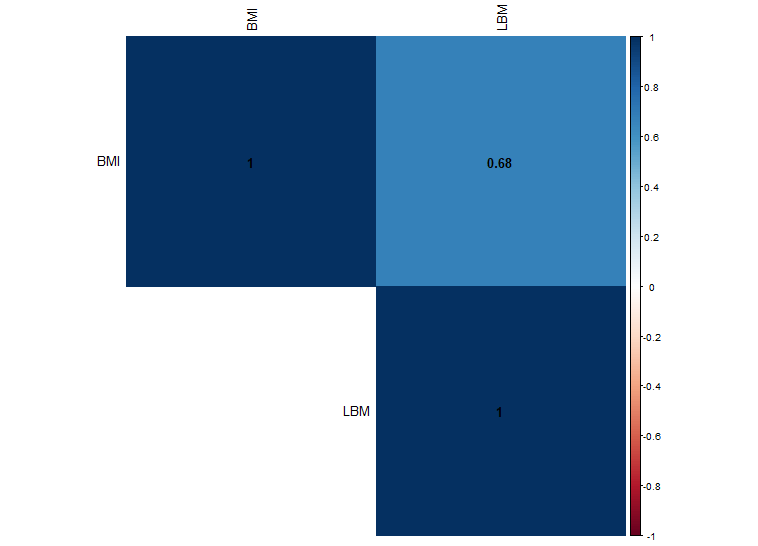
However, keep in mind that correlation does not imply causation (University of Stirling, n.d.), and other factors may influence this relationship.

**Male Dataset:**



The heatmap reveals that BMI and LBM for males have a good/strong positive correlation (around 0.79 or 79%).

**Female Dataset:**

****

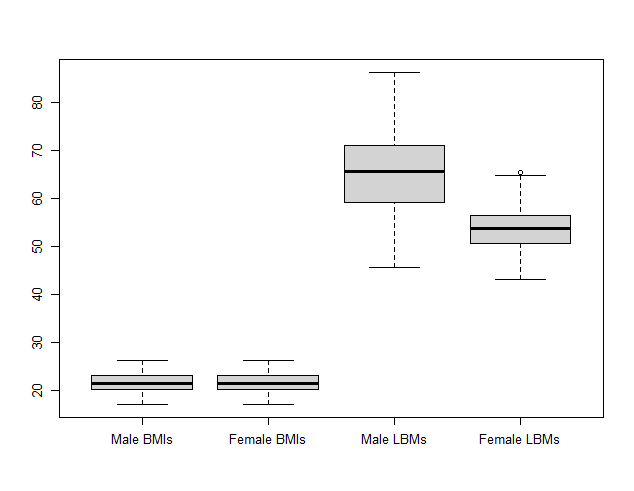
The heatmap reveals that BMI and LBM for females have a moderate positive correlation (around 0.68 or 68%).

1. **Summary Statistics**

For summary statistics of the data, i.e., finding spread and central location of the data, we will use the five number summary method (University of Stirling, n.d.).

It is possible to turn the Five Number Summary into a picture by means of a boxplot (University of Stirling, n.d.).

Below are R results:



*We can notice that there is an outlier in female LBM values*

1. **Measure of central tendency**

For central tendency, we use *mean, median & mode* (University of Stirling, n.d.), the results can be seen in summary of the data across the genders (in step 1 & 3) shown above.

1. **Measure of dispersion**

How could we measure total variation of the data?, For that we use variance or, its square root standard deviation, and resistant measure like IQR, the results are as below:

**BMI**

Male IQR: 3.01

Female IQR: 3.00

Male Standard Deviation: 2.09

Female Standard Deviation: 1.99

**LBM**

Male IQR: 11.74

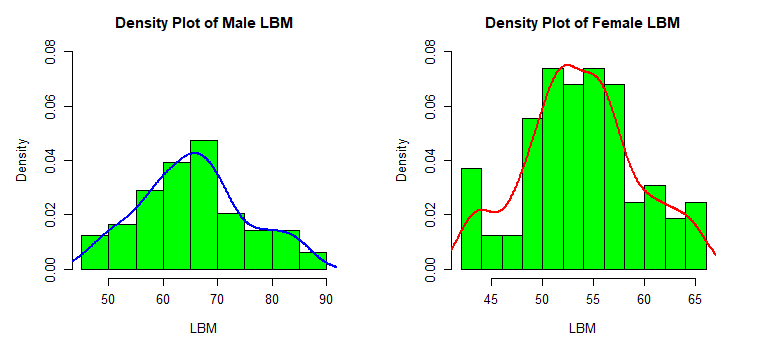
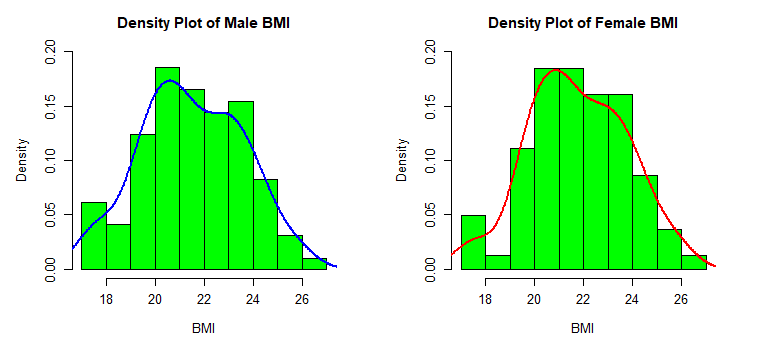
Female IQR: 5.89

Male Standard Deviation: 9.67

Female Standard Deviation: 5.46

1. **Skewness**

Checking skewness of the distribution of the given data using density plots:



**BMI**

Male Skewness: 0.01

Female Skewness: -0.02

**LBM**

Male Skewness: 0.19

Female Skewness: 0.07

*Other than Male LBMs, almost all 3 are approximately symmetric.*

2. Using an appropriate statistical test, investigate whether there is a difference in mean LBM between males and females.

We use T-test: Two Sample Independent Groups to investigate the difference between mean LBM between males and females (University of Stirling, n.d.).

The results taken from R as below:

**Normality Test:**

Our null hypothesis **H**0 is that the variables (male & female LBM) follow a normal distribution, and our alternative hypothesis **H**1 is that these variables does *not* follow normal distribution, we will conclude by comparing P-Value to the significance level of 5%

Shapiro-Wilk normality test

data: Male\_LBMs

W = 0.982, p-value = 0.212

Since the P-value (0.212) is greater than the significance level (commonly chosen as 0.05), we fail to reject the null hypothesis H0. This means that there is not enough evidence to conclude that the variable "Male\_LBMs" does not follow a normal distribution. In other words, based on the Shapiro-Wilk test, there is no strong evidence to suggest that the distribution of "Male\_LBMs" significantly deviates from a normal distribution.

Shapiro-Wilk normality test

data: Female\_LBMs

W = 0.980, p-value = 0.245

Since the p-value (0.245) is greater than the significance level (commonly chosen as 0.05), we fail to reject the null hypothesis. This means that there is not enough evidence to conclude that the variable "Female\_LBMs" does not follow a normal distribution.

**Testing means:**

Welch Two Sample t-test

data: Male\_LBMs and Female\_LBMs

t = 10.303, df = 156.04, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

9.612882 14.172960

sample estimates:

mean of x mean of y

65.56265 53.66973

Our Null hypothesis H0 is H0 =µ1 - µ2 = 0, and H1 = µ1 - µ2 ≠ 0, if the P-Value is less than 0.05, we will reject the null hypothesis and conclude that the alternative hypothesis, H1 is true.

*Hence the difference is not equal to zero, the null hypothesis is rejected at 1% level of significance, the 95% confidence interval also does not contain “zero” in it, suggesting the same conclusion.*

3. For male and female sports people separately, calculate the correlation coefficient for LBM and BMI given and comment on the relationship between LBM and BMI.

To calculate the correlation coefficient we “cor()” command, here are the results taken from R:

Male Sportspeople: 0.79

Female Sportspeople: 0.68

The relationship between LBM and BMI for male is closer to one, indicating stronger relationship (University of Stirling, n.d.), and the relationship for female variables is relatively less, and moderate. The details on correlation has also passed earlier in discussion under heatmap in

part 3.

4. We would like to investigate a model to test the relationship between LBM and BMI for male sportspeople. You must include output from R to support your findings.

Details you should include are:

(a) using your previous results comment on whether there would be any value in

including the data for females in this model.

(b) a description of the model;

(c) a summary of the fitted model with interpretation of test statistics and parameter estimates;

(d) evidence as to whether assumptions of the model have been met;

(e) conduct a formal test to question whether there is a significant linear relationship

between LBM and BMI.

1. There would not be any value in including the data for female in this model as the correlation between female BMI and LBM is lesser, this could lead to misleading results for the male dataset.
2. **DESCRIPTION OF THE MODEL:**

To test this relation, we use ***Simple linear regression model*** with the following hypothesis:

The model states that the higher body mass index (BMI) will lead to the higher value of lean body mass (LBM), we use dataset for males for this model, our explanatory (independent) variable is BMI and anticipated (dependent) variable is LBM.

1. **SUMMARY OF THE FITTED MODEL:**

Call:

lm(formula = LBM ~ BMI, data = dfmale)

Residuals:

Min 1Q Median 3Q Max

-11.989 -3.687 -0.191 4.355 13.036

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -13.201 6.293 -2.1 0.039 \*

BMI 3.663 0.291 12.6 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.96 on 95 degrees of freedom

Multiple R-squared: 0.625, Adjusted R-squared: 0.621

F-statistic: 158 on 1 and 95 DF, p-value: <2e-16

The parameter estimate for slope is 3.66, tells us that a unit increase in BMI will lead to an increase of 3.66 units in LBM, the P-Values are less 0.01, we reject the **H0** that there lies no relationship between LBM and BMI and conclude the positive relationship (**H1**), the model is statistically significant at 1% level of significance.

1. The assumptions of the model is checked by the Anderson-darling normality testing (University of Stirling, n.d.), the results are as below:

Anderson-Darling normality test

data: model$residuals

A = 0.362, p-value = 0.437

The P-value is 43.76%, the null hypothesis H0 that the distribution is normal is not

rejected.

1. By the help of T testing results, we can confirm that there is a positive relationship between LBM and BMI as the P value is less than 0.01 (significant at 1% level of significance).

5. Use the model developed in Question 4 to predict the LBM for a male whose BMI is 25.

For prediction, we use the predict() command, the result:

**LBM WHEN BMI IS 25=**

78.383

Our estimated equation is LBM = -13.2 + 3.66 BMI; -13.2 + 3.66 (25) = 78.383

6. Assess the predictive performance of the model.

To access the predictive performance of the model, we check the R-Squared value (AbdulWaheed, 2023), which is the coefficient of determination.

The R-Squared value in our model is 0.625, which implies:

* About 62.5% of the variability in Lean Body Mass (LBM) is explained by the variation in Body Mass Index (BMI) in the model.
* The remaining 37.5% of the variability in LBM is attributed to other factors not included in the model.
* A higher R-squared value indicates a better fit of the model to the data, suggesting that BMI is a moderately good predictor of LBM for male sportspeople in this case.

7. In this final section include all R code that you have used for this project verbatim. Ensure that:

* the code for each question can be easily found;
* all code is adequately commented;
* variable names are sensible.

**Question 1**

# Creating Data Frame

df <- sports

# Exploratory Data Analysis

summary(df, digits = 5)

# calculating mode

# Function to calculate the mode of a vector

calculate\_mode <- function(x) {

unique\_x <- unique(x)

freq <- tabulate(match(x, unique\_x))

mode\_value <- unique\_x[which.max(freq)]

return(mode\_value)

}

# Calculate mode for LBM

mode\_lbm <- calculate\_mode(df$LBM)

print("Mode of LBM:")

print(mode\_lbm)

# Calculate mode for BMI

mode\_bmi <- calculate\_mode(df$BMI)

print("Mode of BMI:")

print(mode\_bmi)

library('aplpack')

stem.leaf(df)

# Separating the Data

dfmale <- subset(df, sex == "male")

dffemale <- subset(df, sex == "female")

# Summaries

summary(dfmale, digits = 5)

summary(dffemale, digits = 5)

# Visualizations by Stem Leaf Backpack

Male\_BMIs <- dfmale$BMI

Female\_BMIs <- dffemale$BMI

Male\_LBMs <- dfmale$LBM

Female\_LBMs <- dffemale$LBM

stem.leaf.backback(Male\_BMIs, Female\_BMIs, unit = 1, style = "bare", m = 2, depths = FALSE)

# Histogram

# Set up the layout for the plots (2 rows, 2 columns)

par(mfrow = c(2, 2))

# Plot 1: Histogram for male BMI

hist(Male\_BMIs, main = "Histogram of Male BMI", xlab = "BMI", col = "blue", ylim = c(0, 20))

# Plot 2: Histogram for female BMI

hist(Female\_BMIs, main = "Histogram of Female BMI", xlab = "BMI", col = "red", ylim = c(0, 20))

# Plot 3: Histogram for male LBM

hist(Male\_LBMs, main = "Histogram of Male LBM", xlab = "LBM", col = "blue", ylim = c(0, 20))

# Plot 4: Histogram for female LBM

hist(Female\_LBMs, main = "Histogram of Female LBM", xlab = "LBM", col = "red", ylim = c(0, 20))

# Scatter

# Set up the layout for the plots (2 rows, 2 columns)

par(mfrow = c(2, 2))

# Plot 1: Scatterplot for male BMI vs. LBM

plot(Male\_BMIs, Male\_LBMs, main = "Male BMI vs. LBM", xlab = "BMI", ylab = "LBM", col = "blue")

# Plot 2: Scatterplot for female BMI vs. LBM

plot(Female\_BMIs, Female\_LBMs, main = "Female BMI vs. LBM", xlab = "BMI", ylab = "LBM", col = "red")

# Plot 3: Scatterplot for male BMI vs. LBM with regression line

plot(Male\_BMIs, Male\_LBMs, main = "Male BMI vs. LBM (with Regression Line)", xlab = "BMI", ylab = "LBM", col = "blue")

abline(lm(Male\_LBMs ~ Male\_BMIs), col = "black")

# Plot 4: Scatterplot for female BMI vs. LBM with regression line

plot(Female\_BMIs, Female\_LBMs, main = "Female BMI vs. LBM (with Regression Line)", xlab = "BMI", ylab = "LBM", col = "red")

abline(lm(Female\_LBMs ~ Female\_BMIs), col = "gold")

# Correlation Heatmap

install.packages("corrplot")

library(corrplot)

# Create a correlation matrix

correlation\_matrix <- cor(df[, c("BMI", "LBM")])

# Plot the correlation heatmap

corrplot(correlation\_matrix, method = "color", type = "upper", addCoef.col = "black", tl.col = "black")

# For male

# Create a correlation matrix

correlation\_matrix <- cor(dfmale[, c("BMI", "LBM")])

# Plot the correlation heatmap

corrplot(correlation\_matrix, method = "color", type = "upper", addCoef.col = "black", tl.col = "black")

# For female

# Create a correlation matrix

correlation\_matrix <- cor(dffemale[, c("BMI", "LBM")])

# Plot the correlation heatmap

corrplot(correlation\_matrix, method = "color", type = "upper", addCoef.col = "black", tl.col = "black")

# Summary Statistics

# Show a boxplot of the BMIs and LBMs across genders

boxplot(Male\_BMIs, Female\_BMIs, Male\_LBMs, Female\_LBMs,

names = c("Male BMI", "Female BMI", "Male LBM", "Female LBM"))

# Dispersion

# BMI

male\_bmi\_iqr <- diff(quantile(dfmale$bmi, c(0.25, 0.75)))

female\_bmi\_iqr <- diff(quantile(dffemale$bmi, c(0.25, 0.75)))

male\_bmi\_sd <- sd(dfmale$bmi)

female\_bmi\_sd <- sd(dffemale$bmi)

# LBM

male\_lbm\_iqr <- diff(quantile(dfmale$lbm, c(0.25, 0.75)))

female\_lbm\_iqr <- diff(quantile(dffemale$lbm, c(0.25, 0.75)))

male\_lbm\_sd <- sd(dfmale$lbm)

female\_lbm\_sd <- sd(dffemale$lbm)

# Density Plots

# Create density plot for BMI across genders

par(mfrow = c(1, 2)) # Arrange plots in a 1x2 grid

hist(dfmale$bmi, main = "Density Plot of Male BMI", xlab = "BMI", prob = TRUE, col = "green", ylim = c(0, 0.2))

lines(density(dfmale$bmi), col = "blue", lwd = 2)

hist(dffemale$bmi, main = "Density Plot of Female BMI", xlab = "BMI", prob = TRUE, col = "green", ylim = c(0, 0.07))

lines(density(dffemale$bmi), col = "red", lwd = 2)

# Create density plot for LBM across genders

par(mfrow = c(1, 2)) # Arrange plots in a 1x2 grid

hist(dfmale$lbm, main = "Density Plot of Male LBM", xlab = "LBM", prob = TRUE, col = "green", ylim = c(0, 0.07))

lines(density(dfmale$lbm), col = "blue", lwd = 2)

hist(dffemale$lbm, main = "Density Plot of Female LBM", xlab = "LBM", prob = TRUE, col = "green", ylim = c(0, 0.07))

lines(density(dffemale$lbm), col = "red", lwd = 2)

# Skewness

# Load necessary library

library(e1071)

# Calculate skewness for BMI and LBM across genders

# BMI

male\_bmi\_skewness <- skewness(Male\_BMIs)

female\_bmi\_skewness <- skewness(Female\_BMIs)

# LBM

male\_lbm\_skewness <- skewness(Male\_LBMs)

female\_lbm\_skewness <- skewness(Female\_LBMs)

**Question 2**

# Perform two-sample t-test for LBM between males and females

# normality test

# Shapiro-Wilk test for normality on male LBM

shapiro.test(Male\_LBMs)

# Shapiro-Wilk test for normality on female LBM

shapiro.test(Female\_LBMs)

t\_test\_result <- t.test(Male\_LBMs, Female\_LBMs)

# Display the results

t\_test\_result

# Calculate correlation coefficient for LBM and BMI for male sportspeople

cor\_male <- cor(Male\_LBMs, Male\_BMIs)

**Question 3**

# Calculate correlation coefficient for LBM and BMI for female sportspeople

cor\_female <- cor(Female\_LBMs, Female\_BMIs)

# Display the correlation coefficients

cat("Correlation Coefficients for LBM and BMI:")

cat("\n")

cat("Male Sportspeople: ", cor\_male, "\n")

cat("Female Sportspeople: ", cor\_female, "\n")

**Question 4**

# Regression Model Linear

model <- lm(LBM ~ BMI, data = dfmale)

sum <- summary(model)

print(sum, digits = 3)

# check assumptions

# include nortest library that has ad.test()

library("nortest")

# normality of residuals

qqnorm(model$residuals)

qqline(model$residuals)

ad.test(model$residuals)

**Question 5**

# prediction

# Predict LBM for BMI = 25

predicted\_lbm <- predict(model, newdata = data.frame(BMI = 25))

print(predicted\_lbm, digits = 5)

References.

Include here references to statistical methods you have used in the module notes, or any online resources you have used to produce this project.

# References

AbdulWaheed, 2023. *Econometrics, Applications with Eviews.* 1 ed. Karachi, Pakistan: Royal Book Company.

University of Stirling, n.d. *Computing Science and Mathematics, MATPMDA Mathematical and Statistical Foundations, Chapter 10 t-tests.* s.l.:s.n.

University of Stirling, n.d. *Computing Science and Mathematics, MATPMDA Mathematical and Statistical Foundations, Chapter 12 Correlation and Linear Regression.* s.l.:s.n.

University of Stirling, n.d. *MATPMDA Mathematical and Statistical Foundations, Chapter 8 Descriptive Statistics.* s.l.:s.n.