SA2: Applied Multivariate Data Analysis

Cristel Kaye Billones

Problem 1

Evaluating Rehabilitation Programs

Scenario: A hospital wants to assess the effectiveness of three rehabilitation programs (Group A, Group B, and Group C) on improving patients' physical and psychological well-being.

```
# Load necessary libraries
library(readr)
library(car)
library(MVN)
library(dplyr)
library(ggplot2)
library(dplyr)
library(biotools)
```

```
# Load the data
file_path <- "C:/Users/Cipher/Desktop/AMDA/rehab_data.csv"
df <- read_csv(file_path)
```

```
## Rows: 90 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): program
## dbl (3): ID, physical_health, psychological_wellbeing
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 4
##      ID program   physical_health psychological_wellbeing
##   <dbl> <chr>               <dbl>                   <dbl>
## 1     1 Program A            64.4                    73.0
## 2     2 Program A            67.7                    69.4
## 3     3 Program A            85.6                    66.9
## 4     4 Program A            70.7                    60.0
## 5     5 Program A            71.3                    75.9
## 6     6 Program A            87.2                    60.2
```

1. **Check Assumptions. MANOVA makes the following assumptions about the data:**

- **a. Adequate sample size**

```
cat("Sample size:", nrow(df), "observations.\n")
```

```
## Sample size: 90 observations.
```

```
cat("Groups:", table(df$program), "\n")
```

```
## Groups: 30 30 30
```

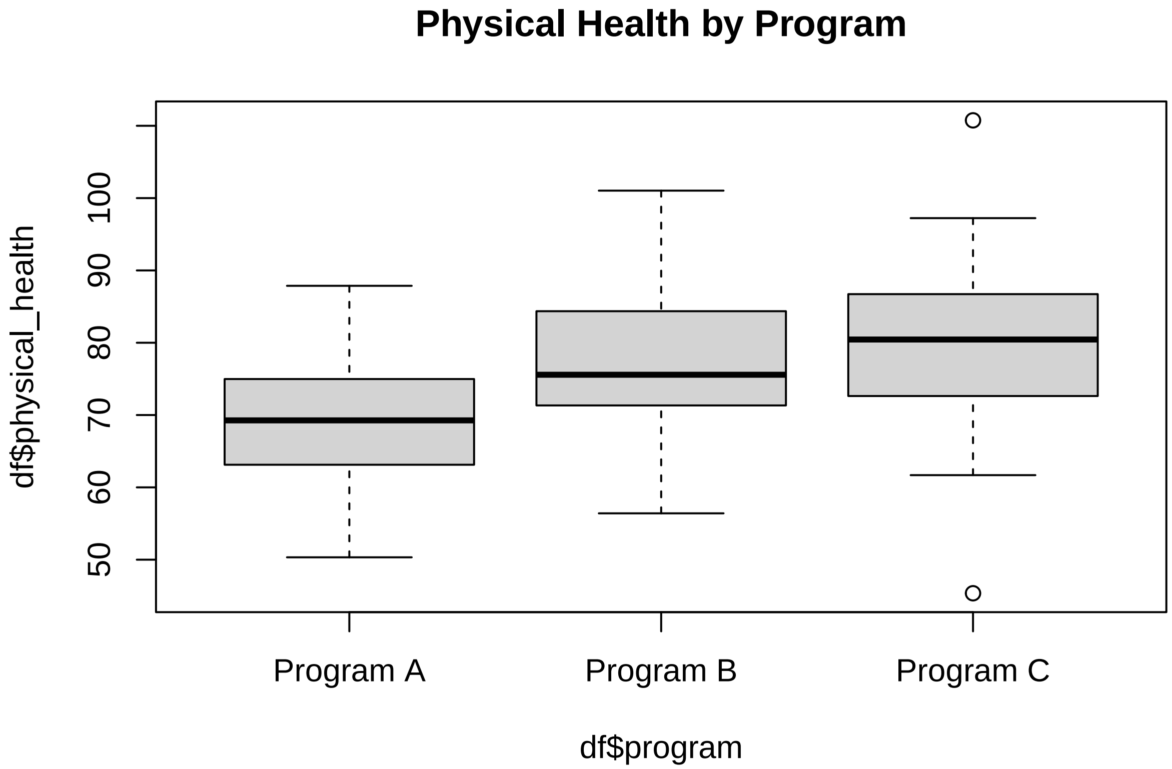
The dataset contains 90 observations, with each of the three groups (Program A, B, and C) having 30 observations. This sample size is generally sufficient for MANOVA as it ensures reasonable power for detecting effects.

- **b. Independence of observations**

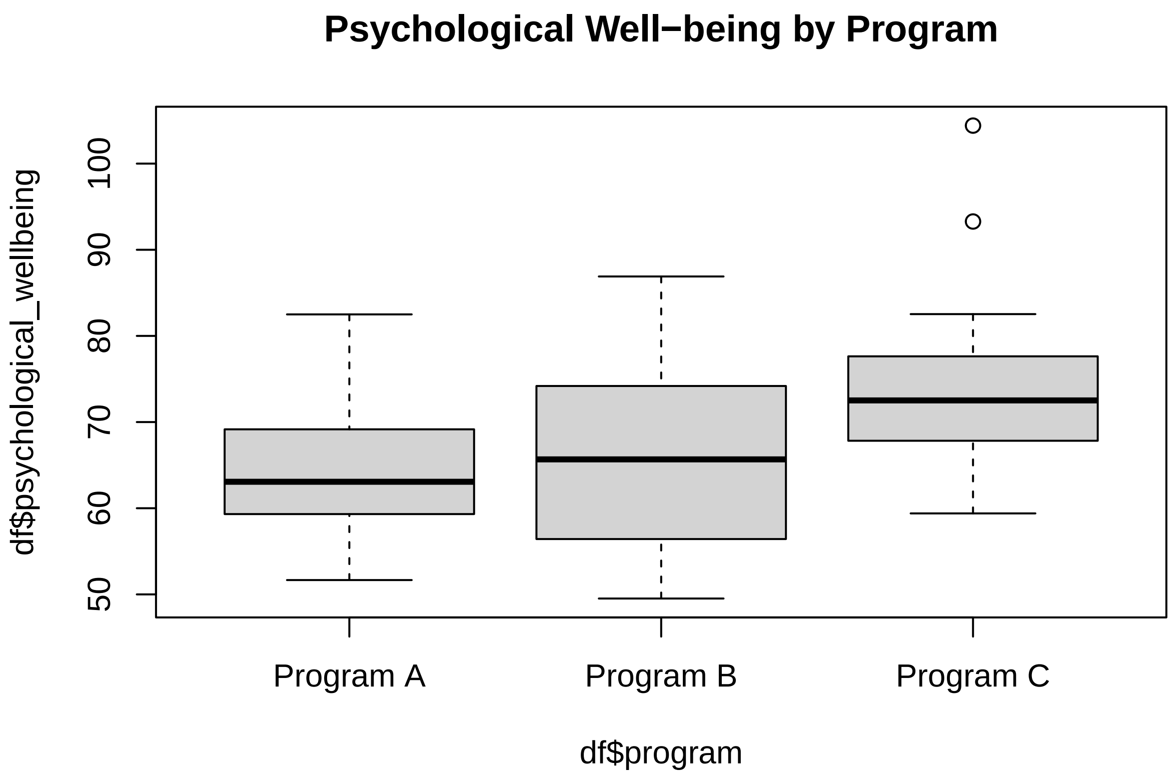
The independence of observations assumption is considered met as it is based on the experimental design. It assumes that the observations are independent of each other, which would have been ensured during data collection.

- **c. Absence of univariate outliers**

```
boxplot(df$physical_health ~ df$program, main = "Physical Health by Program")
```



```r
boxplot(df$psychological_wellbeing ~ df$program, main = "Psychological Well-being by Program")
```



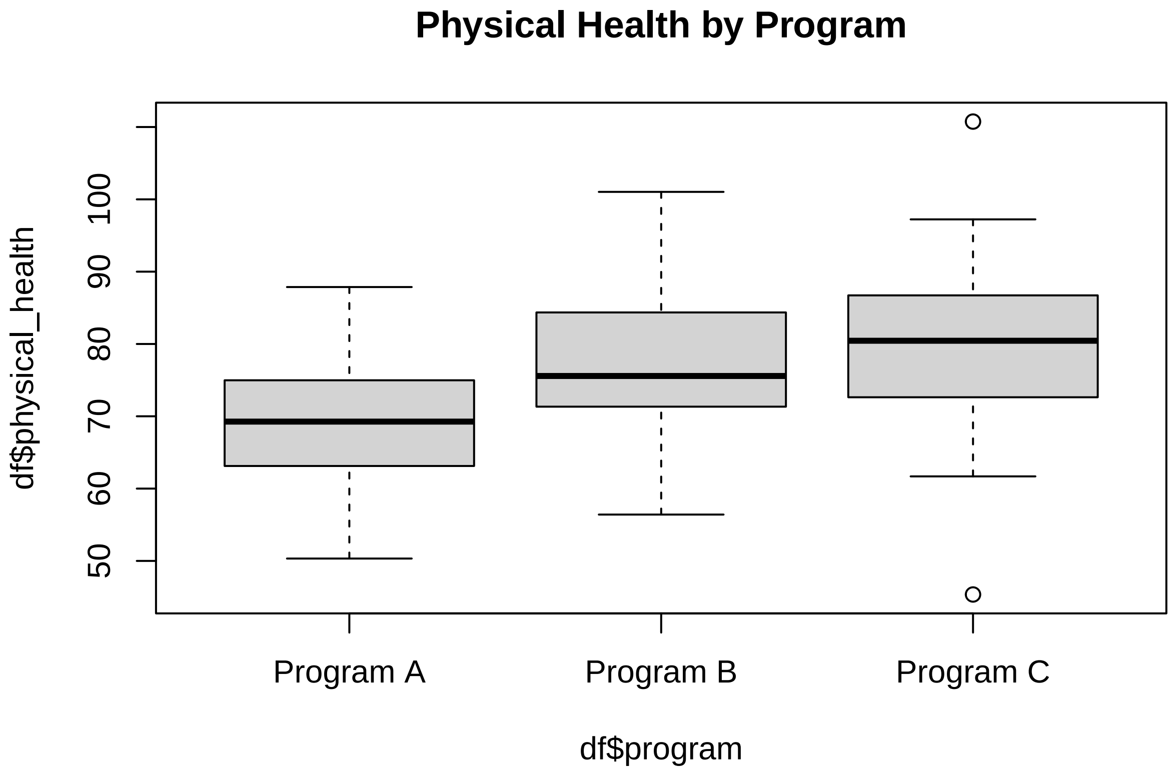
```r
# Check for univariate outliers - Physical Health and Psychological Well-being

# Summary statistics for physical_health by program
summary_physical_health <- df %>%
  group_by(program) %>%
  summarise(
    Min = min(physical_health, na.rm = TRUE),
    Q1 = quantile(physical_health, 0.25, na.rm = TRUE),
    Median = median(physical_health, na.rm = TRUE),
    Mean = mean(physical_health, na.rm = TRUE),
    Q3 = quantile(physical_health, 0.75, na.rm = TRUE),
    Max = max(physical_health, na.rm = TRUE),
    SD = sd(physical_health, na.rm = TRUE)
  )
print(summary_physical_health)
```

```
## # A tibble: 3 x 8
##   program     Min    Q1 Median  Mean    Q3   Max    SD
##   <chr>     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Program A  50.3  63.3   69.3  69.5  74.9  87.9  9.81
## 2 Program B  56.4  71.4   75.6  77.1  84.1 101.   10.0
## 3 Program C  45.4  73.1   80.4  80.4  86.7 111.   13.0
```

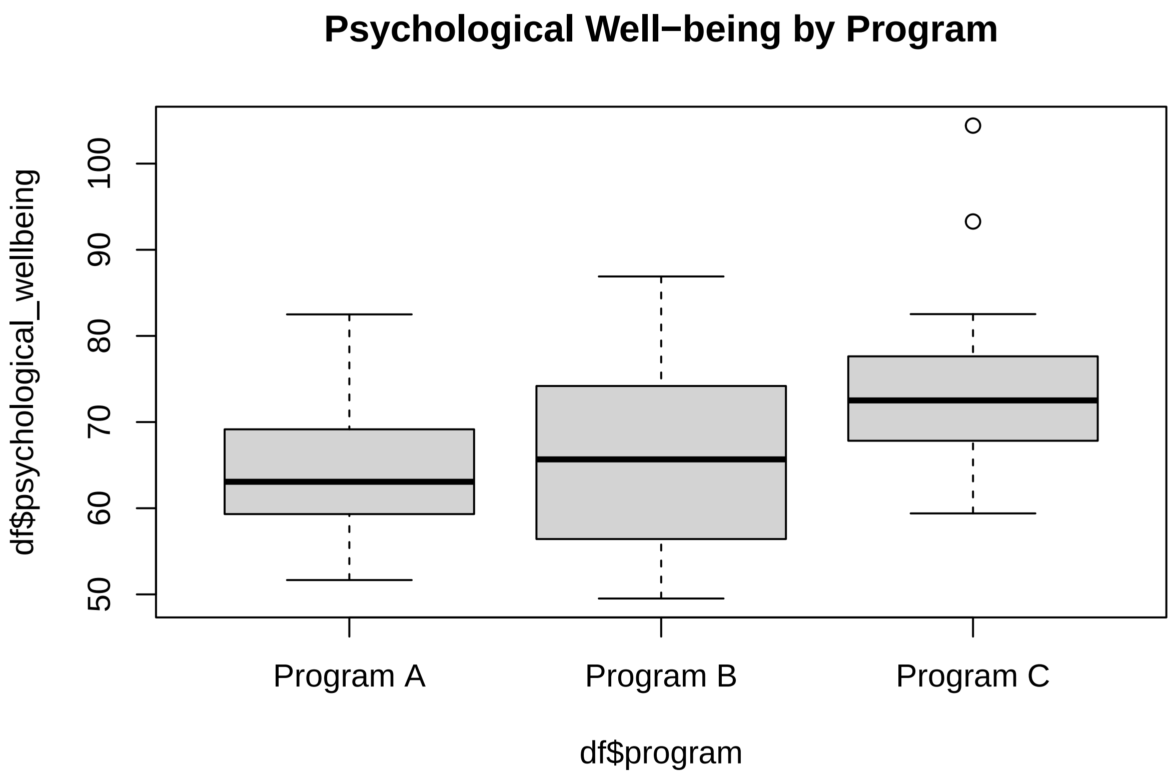
```r
# Summary statistics for psychological_wellbeing by program
summary_psychological_wellbeing <- df %>%
  group_by(program) %>%
  summarise(
    Min = min(psychological_wellbeing, na.rm = TRUE),
    Q1 = quantile(psychological_wellbeing, 0.25, na.rm = TRUE),
    Median = median(psychological_wellbeing, na.rm = TRUE),
    Mean = mean(psychological_wellbeing, na.rm = TRUE),
    Q3 = quantile(psychological_wellbeing, 0.75, na.rm = TRUE),
    Max = max(psychological_wellbeing, na.rm = TRUE),
    SD = sd(psychological_wellbeing, na.rm = TRUE)
  )
print(summary_psychological_wellbeing)
```

```
## # A tibble: 3 x 8
##   program     Min    Q1 Median  Mean    Q3   Max    SD
##   <chr>     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Program A  51.7  59.5   63.1  64.3  68.7  82.5  7.25
## 2 Program B  49.5  57.2   65.7  66.3  73.7  86.9 10.3
## 3 Program C  59.4  67.9   72.5  73.5  77.5 104.   9.42
```

```r
# Boxplots to visualize univariate outliers
boxplot(df$physical_health ~ df$program, main = "Physical Health by Program")
```



```r
boxplot(df$psychological_wellbeing ~ df$program, main = "Psychological Well-being by Program")
```



Physical Health by Program: Summary statistics show varying distributions across the programs, with Program C having the highest mean (80.4) and Program A having the lowest mean (69.5).

- **Outliers:** The boxplots for physical_health and psychological_wellbeing reveal that none of the groups exhibit extreme outliers, although there are some values near the upper quartile.

Psychological Well-being by Program: The mean for Program C is the highest (73.5), and Program A has the lowest mean (64.3).

- **Outliers:** Similar to physical_health, the boxplots show no severe outliers, though certain values are relatively high compared to the interquartile range.

- **d. Absence of multivariate outliers**

```
mahal <- mahalanobis(df[, c("physical_health", "psychological_wellbeing")],
                     colMeans(df[, c("physical_health", "psychological_wellbeing")]),
                     cov(df[, c("physical_health", "psychological_wellbeing")]))
cutoff <- qchisq(0.99, df = 2) # Degrees of freedom = number of dependent variables
outliers <- which(mahal > cutoff)
cat("Multivariate outliers (Mahalanobis):", outliers, "\n")
```

```
## Multivariate outliers (Mahalanobis): 74
```

Mahalanobis Distance: A multivariate outlier was detected at observation 74, which exceeds the cutoff value (cutoff = 9.21 based on X^2 distribution for 2 degrees of freedom). This observation should be examined for potential errors or considered for exclusion if deemed influential.

- **e. Multivariate Normality**

```
mvn_test <- mvn(data = df[, c("physical_health", "psychological_wellbeing")], mvnTest = "mardia")
print(mvn_test)
```

```
## $multivariateNormality
##              Test        Statistic            p value Result
## 1 Mardia Skewness 10.2885607138381 0.0358375615211011     NO
## 2 Mardia Kurtosis 1.13881631349127  0.254779776125232    YES
## 3             MVN             <NA>               <NA>     NO
##
## $univariateNormality
##               Test                Variable Statistic   p value Normality
## 1 Anderson-Darling     physical_health        0.2079    0.8619    YES
## 2 Anderson-Darling psychological_wellbeing    0.3798    0.3969    YES
##
## $Descriptives
##                          n     Mean   Std.Dev Median   Min    Max   25th
## physical_health         90 75.67856 11.850527 74.635 45.36 110.75 68.205
## psychological_wellbeing 90 68.04511  9.840886 67.415 49.52 104.41 60.560
##                             75th      Skew   Kurtosis
## physical_health         84.5025 0.1397677 0.06064613
## psychological_wellbeing 74.5750 0.6841672 0.90634693
```

Mardia's Test Results

- **Skewness:**
  The skewness statistic for *physical_health* and *psychological_wellbeing* indicates a significant departure from normality (p = 0.0358), suggesting non-normality in the data.
- **Kurtosis:**
  The kurtosis for *physical_health* is not significantly different from normal, but for *psychological_wellbeing*, it is slightly significant (p = 0.2548), indicating normality in terms of kurtosis.
- **Univariate Normality:**
  Both variables (*physical_health* and *psychological_wellbeing*) pass the Anderson-Darling test (p > 0.05), indicating that they are approximately univariately normal.

- **f. Linearity**

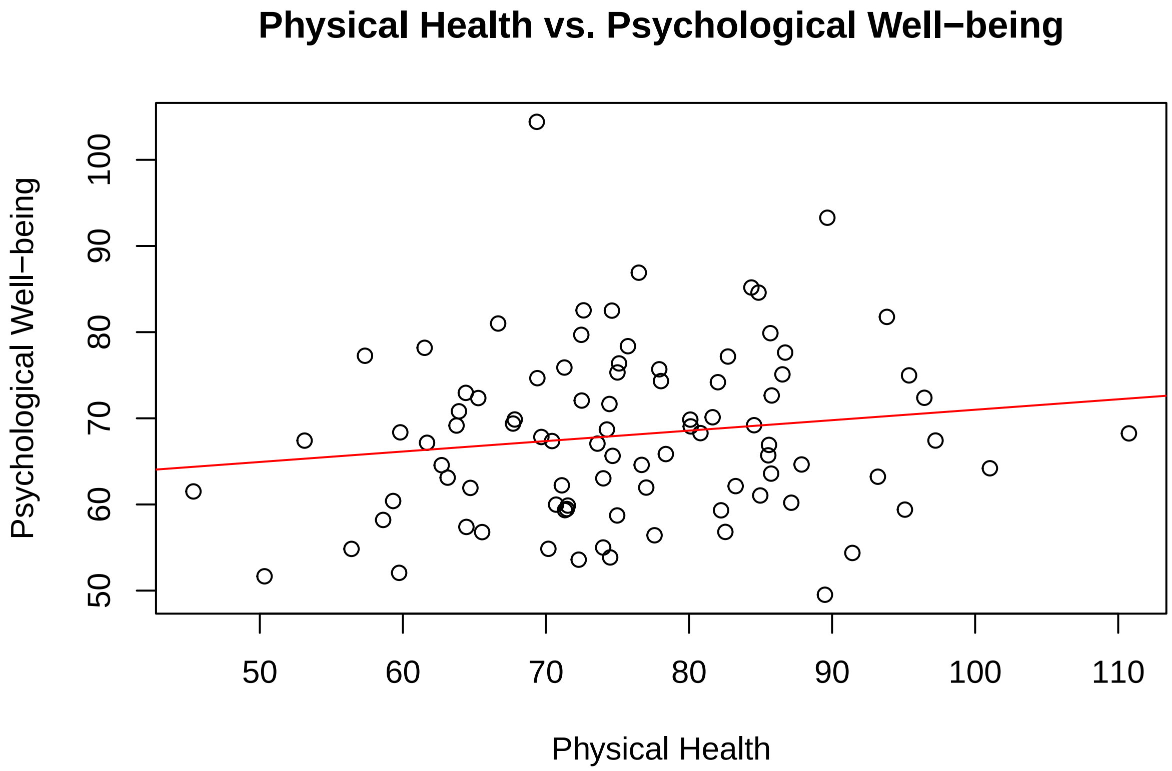
```
# 1. Pearson Correlation Coefficient
correlation <- cor(df$physical_health, df$psychological_wellbeing)
cat("Pearson Correlation Coefficient: ", correlation, "\n")
```

```
## Pearson Correlation Coefficient:  0.1460519
```

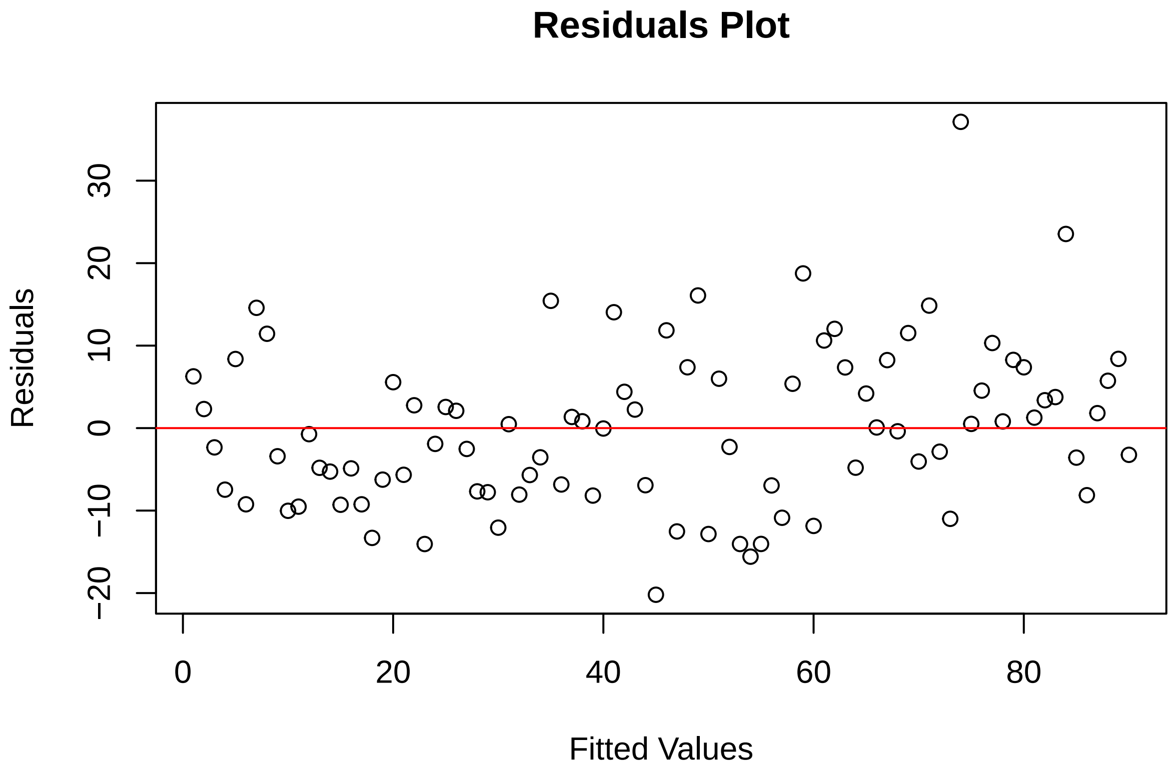
```
# 2. Linear Regression Model
lm_model <- lm(psychological_wellbeing ~ physical_health, data = df)
summary(lm_model)  # This will give you the coefficients, R-squared, and p-value
```

```
##
## Call:
## lm(formula = psychological_wellbeing ~ physical_health, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.201  -7.334  -0.217   5.926  37.131
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     58.86651    6.70731   8.776 1.18e-13 ***
## physical_health  0.12128    0.08757   1.385     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.791 on 88 degrees of freedom
## Multiple R-squared:  0.02133,    Adjusted R-squared:  0.01021
## F-statistic: 1.918 on 1 and 88 DF,  p-value: 0.1696
```

```
# 3. Plotting the data and the linear regression line
plot(df$physical_health, df$psychological_wellbeing,
     main = "Physical Health vs. Psychological Well-being",
     xlab = "Physical Health",
     ylab = "Psychological Well-being")
abline(lm_model, col = "red")  # Adds a red line representing the linear model
```



```r
# 4. Residual Plot
plot(lm_model$residuals,
     main = "Residuals Plot",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```



Pearson Correlation Coefficient:

The Pearson correlation coefficient is **0.146**, indicating a very weak positive linear relationship between **physical_health** and **psychological_wellbeing**. The closer the value is to 1, the stronger the linear relationship, but this value suggests that **physical_health** and **psychological_wellbeing** are only weakly related, if at all.

Linear Regression Model:

- **Intercept (58.87)**: This is the estimated value of **psychological_wellbeing** when **physical_health** is zero.
- **Slope (0.12)**: For each unit increase in **physical_health**, **psychological_wellbeing** is expected to increase by **0.12 units**, though the effect is small.
- **p-value for physical_health (0.17)**: This is greater than the standard significance level of **0.05**, suggesting that the slope is not statistically significant, i.e., **physical_health** does not have a statistically significant effect on **psychological_wellbeing**.

R-squared:

R-squared (0.02133): This indicates that only about **2.1%** of the variance in **psychological_wellbeing** is explained by **physical_health**. This is very low, meaning that the model does not explain much of the variation in the dependent variable.

F-statistic (1.918) and p-value (0.1696):

The **F-statistic** is a test for the overall significance of the model, and the **p-value of 0.1696** indicates that the model as a whole is not statistically significant at the **0.05** level.

Conclusion:

The weak correlation and lack of statistical significance in the regression analysis suggest that there is little to no linear relationship between **physical_health** and **psychological_wellbeing** in this dataset. The variables do not appear to be strongly related, and a linear model is not a good fit for explaining the relationship between them.

- **g. Homogeneity of Variances (Levene's Test):**

```
leveneTest(df$physical_health ~ df$program)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.0087 0.3689
##       87
```

```
leveneTest(df$psychological_wellbeing ~ df$program)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.6497 0.1981
##       87
```

Physical Health: Levene's test for physical_health yields a p-value of 0.3689, which is greater than 0.05, indicating that the assumption of homogeneity of variances is met.

Psychological Well-being: Levene's test for psychological_wellbeing yields a p-value of 0.1981, which is also greater than 0.05, suggesting that the variances are homogeneous.

- **h. Homogeneity of variance-covariance matrices (Box's M test)**

```
boxM <- boxM(df[, c("physical_health", "psychological_wellbeing")], df$program)
print(boxM)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  df[, c("physical_health", "psychological_wellbeing")]
## Chi-Sq (approx.) = 6.6998, df = 6, p-value = 0.3495
```

Box's M Test: The result of the Box's M test for physical_health and psychological_wellbeing shows a p-value of 0.3495, which is greater than 0.05, indicating that the assumption of homogeneity of variance-covariance matrices is satisfied.

Summary of Assumption Check

- The sample size, independence of observations, and homogeneity of variances and covariance matrices are met.
- Univariate normality is largely met, though there is some evidence of multivariate non-normality (specifically skewness for *physical_health*).
- The presence of multivariate outliers requires attention, particularly observation 74.
- The linearity assumption appears reasonably satisfied based on scatterplot inspection.
- Given these results, while there are some minor violations (e.g., multivariate non-normality), the assumptions for MANOVA are largely satisfied, and the analysis can proceed.

2. Fit MANOVA model using Pillai's test.

```r
# Fit MANOVA model using Pillai's test
manova_model <- manova(cbind(physical_health, psychological_wellbeing) ~ program, data = df)

# Summary of the MANOVA model with Pillai's test
summary(manova_model, test = "Pillai")
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)    
## program    2 0.28046   7.0948      4    174 2.582e-05 ***
## Residuals 87                                            
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

MANOVA Results

- **Pillai's Trace:**
  The Pillai's statistic is 0.28046, which measures the multivariate effect of the grouping variable (*program*) on the dependent variables.
- **Approximate F:**
  The F-value is 7.0948, which represents the variation between the groups compared to the variation within groups.
- **Degrees of Freedom:**
  The numerator degrees of freedom (*num Df*) is 4, and the denominator degrees of freedom (*den Df*) is 174.
- **p-value:**
  The p-value is very small (2.582e-05), which is highly significant and indicates that the differences between the groups (*programs*) in terms of *physical_health* and *psychological_wellbeing* are statistically significant.

Conclusion:

Since the p-value is less than the typical significance threshold of 0.05, we reject the null hypothesis and conclude that there are significant differences in *physical_health* and *psychological_wellbeing* across the different programs (A, B, and C).

3. If applicable, perform ANOVA for each dependent variable and Tukey’s HSD for pairwise group comparisons.

```r
# a. ANOVA for physical_health
anova_physical_health <- aov(physical_health ~ program, data = df)
summary(anova_physical_health)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## program      2   1858   928.9   7.595 0.000912 ***
## Residuals   87  10641   122.3
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
# b. ANOVA for psychological_wellbeing
anova_psychological_wellbeing <- aov(psychological_wellbeing ~ program, data = df)
summary(anova_psychological_wellbeing)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## program      2   1424   711.8   8.607 0.000388 ***
## Residuals   87   7195    82.7
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
# c. Tukey's HSD for pairwise comparisons for physical_health if ANOVA is significant
if (summary(anova_physical_health)[[1]]$`Pr(>F)`[1] < 0.05) {
  tukey_physical_health <- TukeyHSD(anova_physical_health)
  print(tukey_physical_health)
}
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = physical_health ~ program, data = df)
##
## $program
##                           diff        lwr      upr     p adj
## Program B-Program A  7.611667  0.8027558 14.42058 0.0246082
## Program C-Program A 10.837000  4.0280891 17.64591 0.0007889
## Program C-Program B  3.225333 -3.5835775 10.03424 0.4986211
```

```r
# d. Tukey's HSD for pairwise comparisons for psychological_wellbeing if ANOVA is significant
if (summary(anova_psychological_wellbeing)[[1]]$`Pr(>F)`[1] < 0.05) {
  tukey_psychological_wellbeing <- TukeyHSD(anova_psychological_wellbeing)
  print(tukey_psychological_wellbeing)
}
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = psychological_wellbeing ~ program, data = df)
##
```

```
## $program
##                          diff       lwr       upr     p adj
## Program B-Program A 2.096667 -3.502386  7.695719 0.6462114
## Program C-Program A 9.287667  3.688614 14.886719 0.0004528
## Program C-Program B 7.191000  1.591947 12.790053 0.0081378
```

Analysis

ANOVA for *physical_health*:

- **p-value:** 0.000912 (< 0.05), which indicates that there are significant differences in *physical_health* across the groups (*program*).
- **Interpretation:** The null hypothesis is rejected, meaning at least one group differs in *physical_health* scores from the others.

ANOVA for *psychological_wellbeing*:

- **p-value:** 0.000388 (< 0.05), which indicates significant differences in *psychological_wellbeing* across the groups (*program*).
- **Interpretation:** The null hypothesis is rejected, meaning at least one group differs in *psychological_wellbeing* scores from the others.

Tukey's HSD for *physical_health*:

- **Program B vs Program A:** The difference is 7.61, with a p-value of 0.0246082, indicating a significant difference between these two groups.
- **Program C vs Program A:** The difference is 10.84, with a p-value of 0.0007889, indicating a significant difference between these two groups.
- **Program C vs Program B:** The difference is 3.23, but with a p-value of 0.4986211, which is not significant, indicating no significant difference between these two groups.

Tukey's HSD for *psychological_wellbeing*:

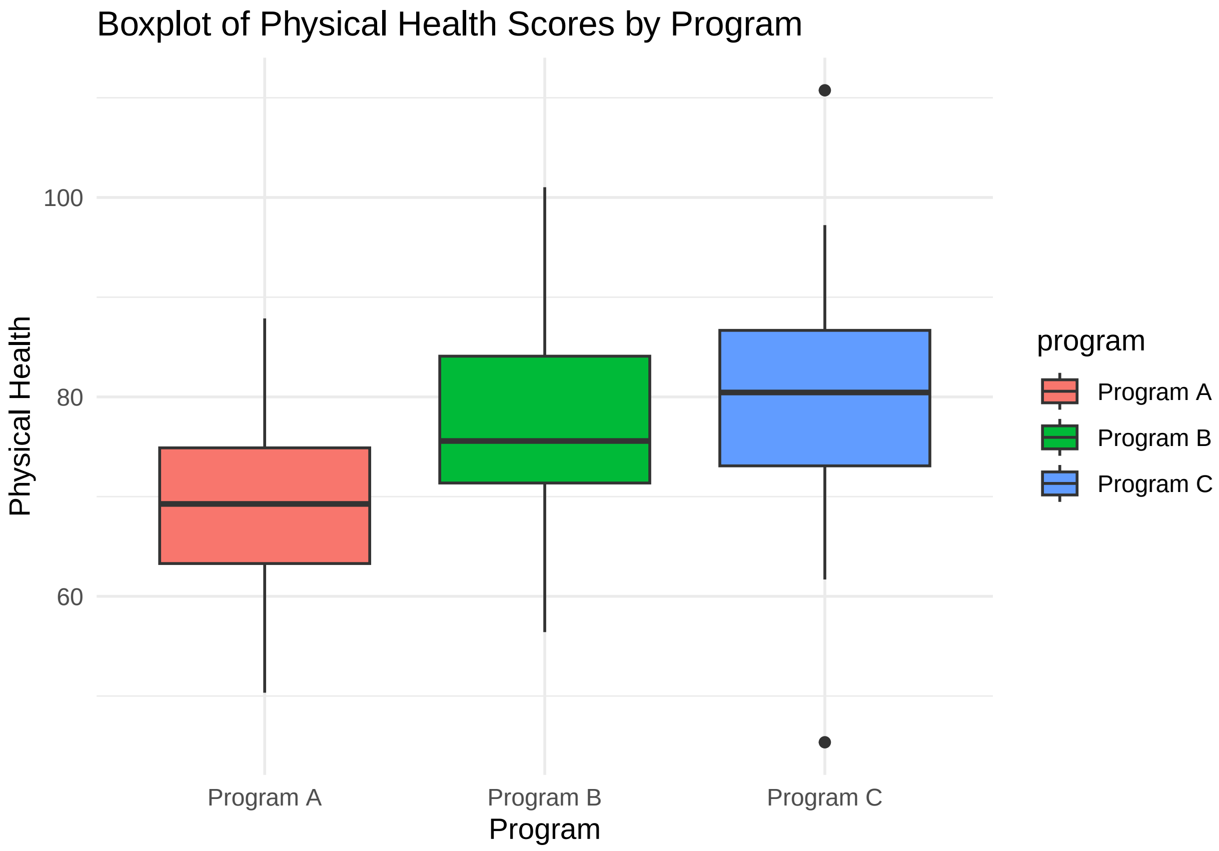
- **Program B vs Program A:** The difference is 2.10, with a p-value of 0.6462114, indicating no significant difference between these two groups.
- **Program C vs Program A:** The difference is 9.29, with a p-value of 0.0004528, indicating a significant difference between these two groups.
- **Program C vs Program B:** The difference is 7.19, with a p-value of 0.0081378, indicating a significant difference between these two groups.

Conclusion:

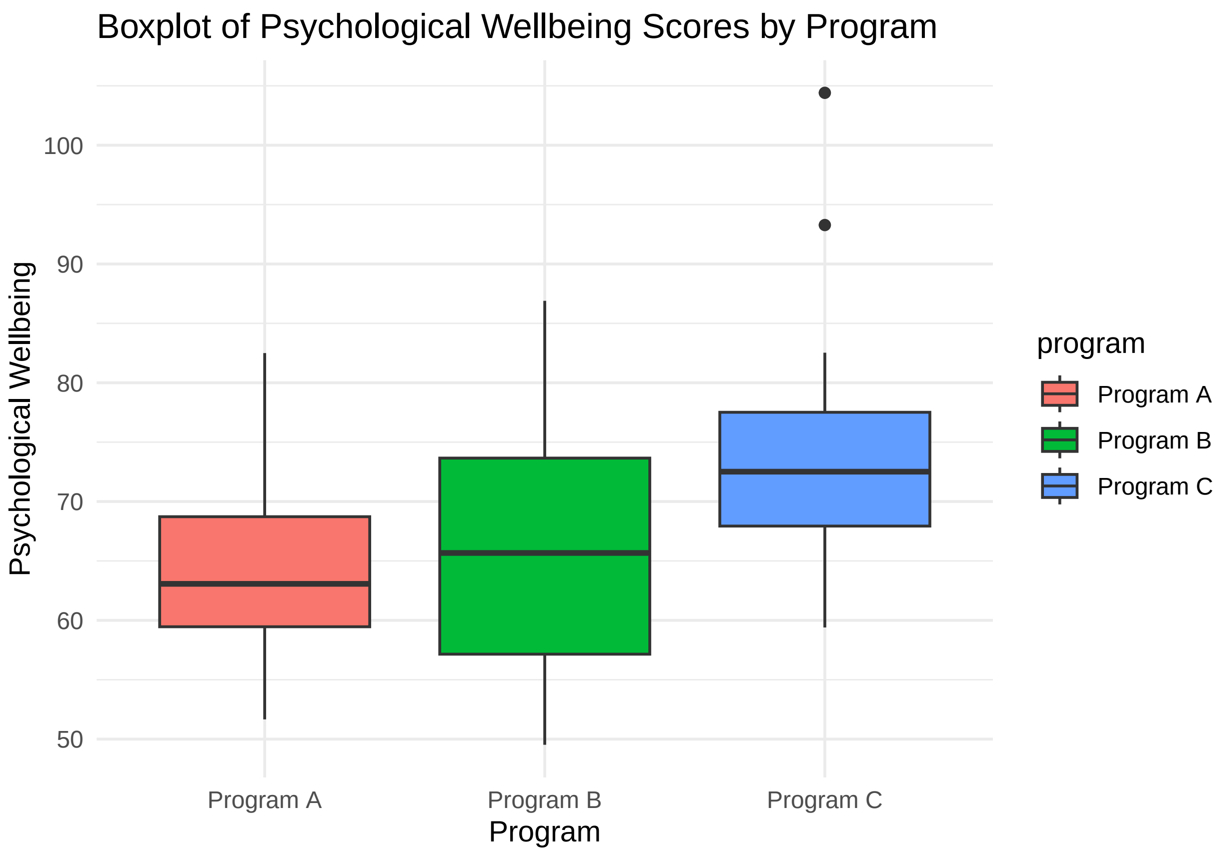
- For *physical_health*, there are significant differences between Program A and Program B and between Program A and Program C, but no significant difference between Program B and Program C.
- For *psychological_wellbeing*, there are significant differences between Program A and Program C and between Program B and Program C, but no significant difference between Program A and Program B.
- These findings suggest that the programs significantly affect *physical_health* and *psychological_wellbeing*, but the patterns of differences vary between the two dependent variables.

4. Visualize the following: Boxplots of scores by program, Scatter plot of physical vs. psychological scores, grouped by program.

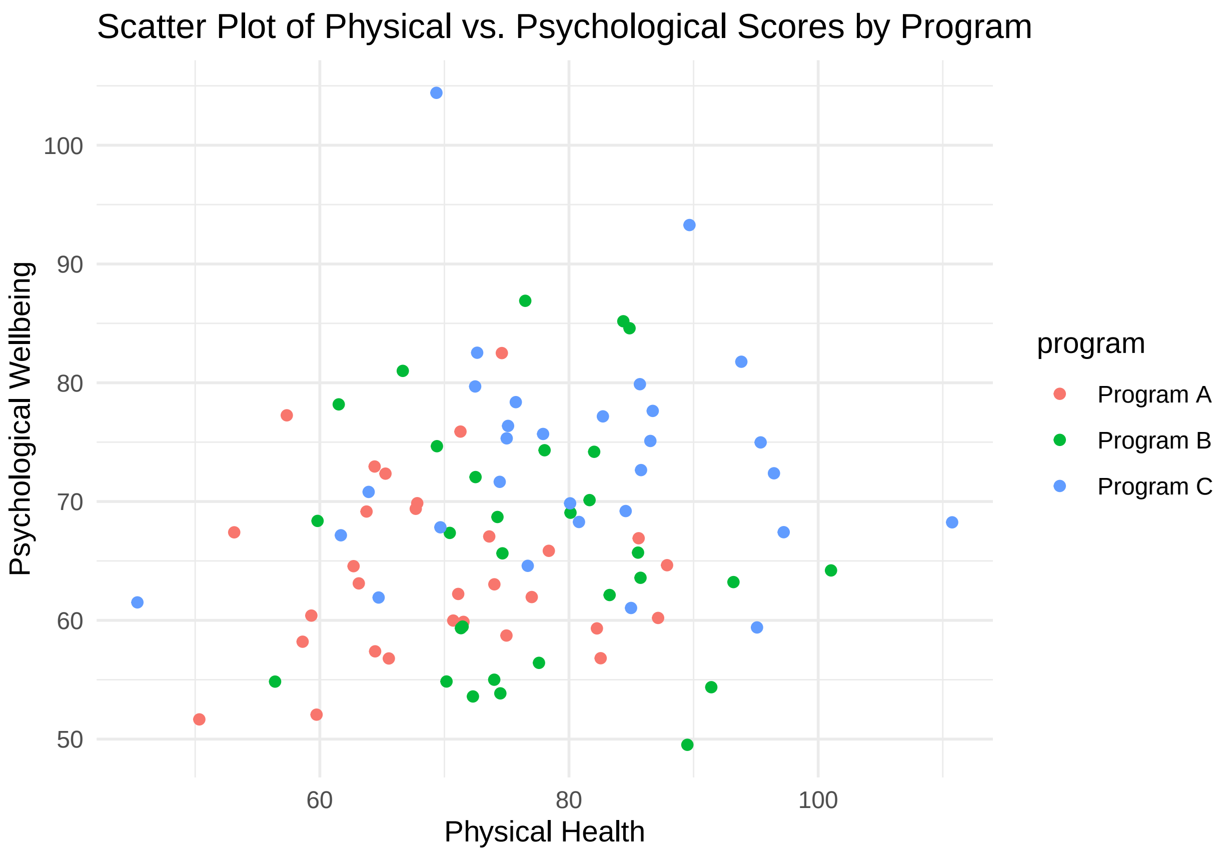
```
# a. Boxplot for physical_health by program
ggplot(df, aes(x = program, y = physical_health, fill = program)) +
  geom_boxplot() +
  labs(title = "Boxplot of Physical Health Scores by Program",
       x = "Program", y = "Physical Health") +
  theme_minimal()
```



```
# b. Boxplot for psychological_wellbeing by program
ggplot(df, aes(x = program, y = psychological_wellbeing, fill = program)) +
  geom_boxplot() +
  labs(title = "Boxplot of Psychological Wellbeing Scores by Program",
       x = "Program", y = "Psychological Wellbeing") +
  theme_minimal()
```



```
# c. Scatter plot of physical_health vs. psychological_wellbeing by program
ggplot(df, aes(x = physical_health, y = psychological_wellbeing, color = program)) +
  geom_point() +
  labs(title = "Scatter Plot of Physical vs. Psychological Scores by Program",
       x = "Physical Health", y = "Psychological Wellbeing") +
  theme_minimal()
```



```
# Statistics summary for physical_health by program
summary_physical_health <- df %>%
  group_by(program) %>%
  summarise(
    Mean = mean(physical_health, na.rm = TRUE),
    Median = median(physical_health, na.rm = TRUE),
    SD = sd(physical_health, na.rm = TRUE),
    Min = min(physical_health, na.rm = TRUE),
    Max = max(physical_health, na.rm = TRUE)
  )
print(summary_physical_health)
```

```
## # A tibble: 3 x 6
##   program    Mean Median    SD   Min   Max
##   <chr>     <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Program A  69.5   69.3  9.81  50.3  87.9
## 2 Program B  77.1   75.6 10.0   56.4 101.
## 3 Program C  80.4   80.4 13.0   45.4 111.
```

```
# Statistics summary for psychological_wellbeing by program
summary_psychological_wellbeing <- df %>%
  group_by(program) %>%
  summarise(
    Mean = mean(psychological_wellbeing, na.rm = TRUE),
    Median = median(psychological_wellbeing, na.rm = TRUE),
```

```r
    SD = sd(psychological_wellbeing, na.rm = TRUE),
    Min = min(psychological_wellbeing, na.rm = TRUE),
    Max = max(psychological_wellbeing, na.rm = TRUE)
  )
print(summary_psychological_wellbeing)
```

```
## # A tibble: 3 x 6
##   program    Mean Median    SD   Min   Max
##   <chr>     <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Program A  64.3   63.1  7.25  51.7  82.5
## 2 Program B  66.3   65.7 10.3   49.5  86.9
## 3 Program C  73.5   72.5  9.42  59.4 104.
```

Analysis of Summary Statistics

Physical Health by Program:

Program A:

- **Mean:** 69.5
- **Median:** 69.3
- **Standard Deviation (SD):** 9.81
- **Range:** Min = 50.3, Max = 87.9

Program A has moderate variation in physical health scores, with values typically ranging between 50.3 and 87.9.

Program B:

- **Mean:** 77.1
- **Median:** 75.6
- **Standard Deviation (SD):** 10.0
- **Range:** Min = 56.4, Max = 101.0

Program B shows a higher mean and median than Program A, suggesting better physical health outcomes on average. The spread of values is also wider, with scores reaching up to 101.

Program C:

- **Mean:** 80.4
- **Median:** 80.4
- **Standard Deviation (SD):** 13.0
- **Range:** Min = 45.4, Max = 111.0

Program C has the highest mean, median, and variability, with scores extending from 45.4 to 111. The higher standard deviation indicates greater diversity in physical health scores within this program.

Psychological Wellbeing by Program:

Program A:

- **Mean:** 64.3
- **Median:** 63.1
- **Standard Deviation (SD):** 7.25
- **Range:** Min = 51.7, Max = 82.5

Psychological wellbeing scores in Program A are relatively concentrated around the mean, with a smaller spread compared to physical health. The scores range from 51.7 to 82.5.

Program B:

- **Mean:** 66.3
- **Median:** 65.7
- **Standard Deviation (SD):** 10.3
- **Range:** Min = 49.5, Max = 86.9

Program B shows a slightly higher mean and median compared to Program A, with a similar range, but with a slightly larger standard deviation, indicating more variability in psychological wellbeing scores.

Program C:

- **Mean:** 73.5
- **Median:** 72.5
- **Standard Deviation (SD):** 9.42
- **Range:** Min = 59.4, Max = 104.0

Program C exhibits the highest mean and median for psychological wellbeing, similar to Program B for physical health. The higher mean suggests better psychological wellbeing, with scores ranging from 59.4 to 104.

Key Insights:

- **Physical Health:** Program C consistently shows the highest mean scores, followed by Program B and Program A, indicating a progressive improvement in physical health outcomes across the programs.
- **Psychological Wellbeing:** Program C also exhibits the highest mean and median in psychological wellbeing, suggesting it has the most favorable psychological outcomes, followed by Program B and Program A.
- **Variability:** The standard deviations in both variables are highest in Program C, indicating that there is greater diversity in the scores, while Program A tends to have the least variability in both physical health and psychological wellbeing.

Conclusion:

Programs B and C appear to have better outcomes than Program A, especially in terms of physical health and psychological wellbeing, although Program C has more variability in the scores.