# SA2 Applied Multivariate Part 1

2024-12-13

## Contents

## The Dataset:

```
head(rehab_data_df)
```

```
##   ID   program physical_health psychological_wellbeing
## 1  1 Program A           64.40                   72.95
## 2  2 Program A           67.70                   69.39
## 3  3 Program A           85.59                   66.91
## 4  4 Program A           70.71                   59.98
## 5  5 Program A           71.29                   75.89
## 6  6 Program A           87.15                   60.20
```

```
colnames(rehab_data_df)
```

```
## [1] "ID"                     "program"
## [3] "physical_health"        "psychological_wellbeing"
```

**Description:**

The dataset used 'rehab_data' is the data that contains the participant's ID, physical_health, program, and psychological_wellbeing.

# Check assumptions: MANOVA

## Adequate Sample Size

```
table(rehab_data_df$program)
```

```
##
## Program A Program B Program C
##        30        30        30
```

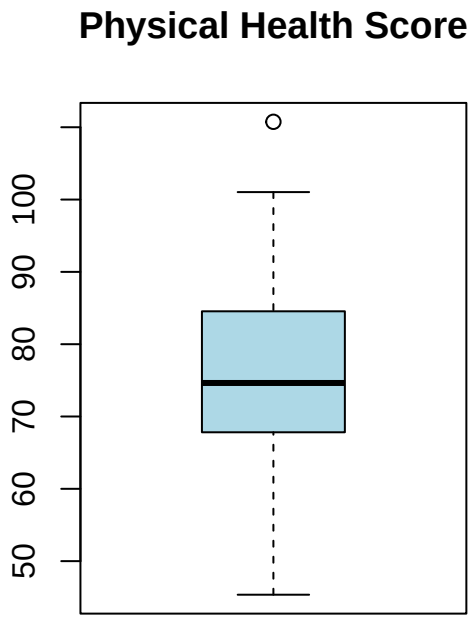
**Findings:**

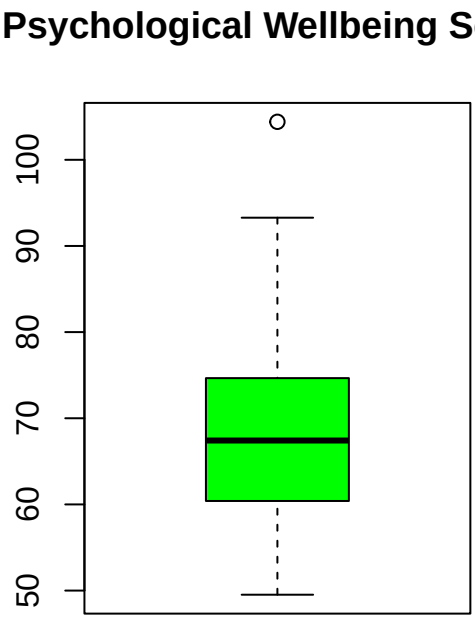
According to this, we have 3 groups namely A,B, and C. While the indicated group are the number of population of participants in each group.

## Absence of Univariate or Multivariate Outliers

```
par(mfrow = c(1, 2))

boxplot(rehab_data_df$physical_health, main = "Physical Health Score", col = "lightblue")
boxplot(rehab_data_df$psychological_wellbeing, main = "Psychological Wellbeing Score", col = "green")
```



```r
par(mfrow = c(1, 1))
```

**Findings:**

The boxplots demonstrate that the distributions of the **Psychological Wellbeing Score** and the **Physical Health Score** are very similar, with the majority of the data falling between the 50th and 90th percentiles. The dots above the boxes show the presence of outliers. Furthermore, each variable displays a balanced range of values, and the score distribution appears to be consistent between the two variables.

## Shapiro-Wilk Tormality Test

```r
shapiro_physical_health
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rehab_data_df$physical_health
## W = 0.99483, p-value = 0.9809
```

```r
shapiro_psychological_wellbeing
```

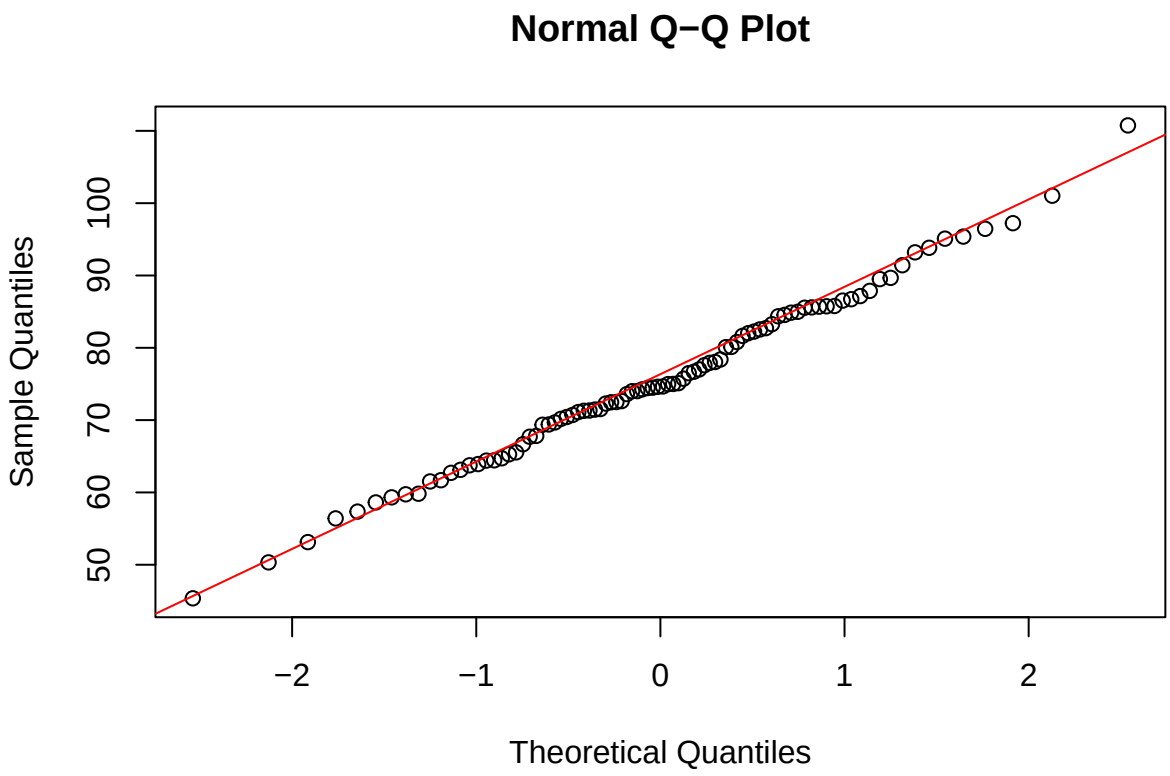
```
##
##  Shapiro-Wilk normality test
##
## data:  rehab_data_df$psychological_wellbeing
## W = 0.97055, p-value = 0.03868
```

**Findings:**

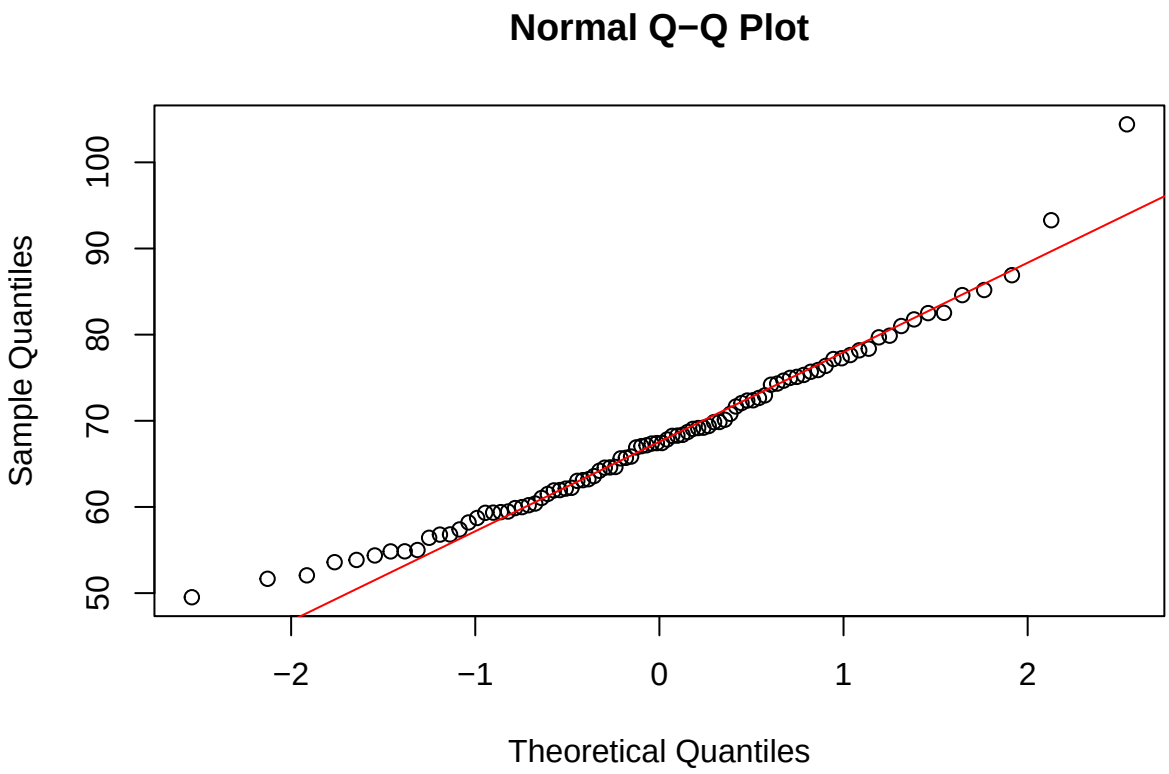
The Shapiro-Wilk test for **Physical Health Scores** indicates a p-value of 0.9809, suggesting that the data follows a normal distribution. However, for **Psychological Wellbeing Scores**, the test yields a p-value of 0.03868, which is below the standard significance level of 0.05, indicating a deviation from normality. These results imply that further analyses, such as MANOVA, should carefully consider the non-normality of psychological wellbeing scores or use robust methods.

### Q-Q Plots for Visual of Normality

```
# Q-Q plot for physical_health
qqnorm(rehab_data_df$physical_health)
qqline(rehab_data_df$physical_health, col = "red")
```




```
# Q-Q plot for psychological_wellbeing
qqnorm(rehab_data_df$psychological_wellbeing)
qqline(rehab_data_df$psychological_wellbeing, col = "red")
```




**Findings:**

As we can see, the normality in the two Q-Q plots aligns well with the red line for **Physical Health Scores**, indicating that the data is approximately normally distributed. However, the Q-Q plot for **Psychological Wellbeing Scores** shows slight deviations from the red line, particularly at the tails, further supporting the Shapiro-Wilk test results that suggest non-normality. This confirms the need to either transform the data or use non-parametric methods for further analysis involving psychological wellbeing scores.

## Shapiro-Wilk test in Each Group

```
results
```

```
## $`Program A`
## $`Program A`$physical_health
##
##  Shapiro-Wilk normality test
##
## data:  program_data$physical_health
## W = 0.97893, p-value = 0.7964
##
##
## $`Program A`$psychological_wellbeing
##
```

```
##  Shapiro-Wilk normality test
##
## data:  program_data$psychological_wellbeing
## W = 0.97104, p-value = 0.568
##
##
##
## $`Program B`
## $`Program B`$physical_health
##
##  Shapiro-Wilk normality test
##
## data:  program_data$physical_health
## W = 0.98663, p-value = 0.9615
##
##
## $`Program B`$psychological_wellbeing
##
##  Shapiro-Wilk normality test
##
## data:  program_data$psychological_wellbeing
## W = 0.95621, p-value = 0.2471
##
##
##
## $`Program C`
## $`Program C`$physical_health
##
##  Shapiro-Wilk normality test
##
## data:  program_data$physical_health
## W = 0.98085, p-value = 0.8476
##
##
## $`Program C`$psychological_wellbeing
##
##  Shapiro-Wilk normality test
##
## data:  program_data$psychological_wellbeing
## W = 0.90916, p-value = 0.01417
```

**Findings:**

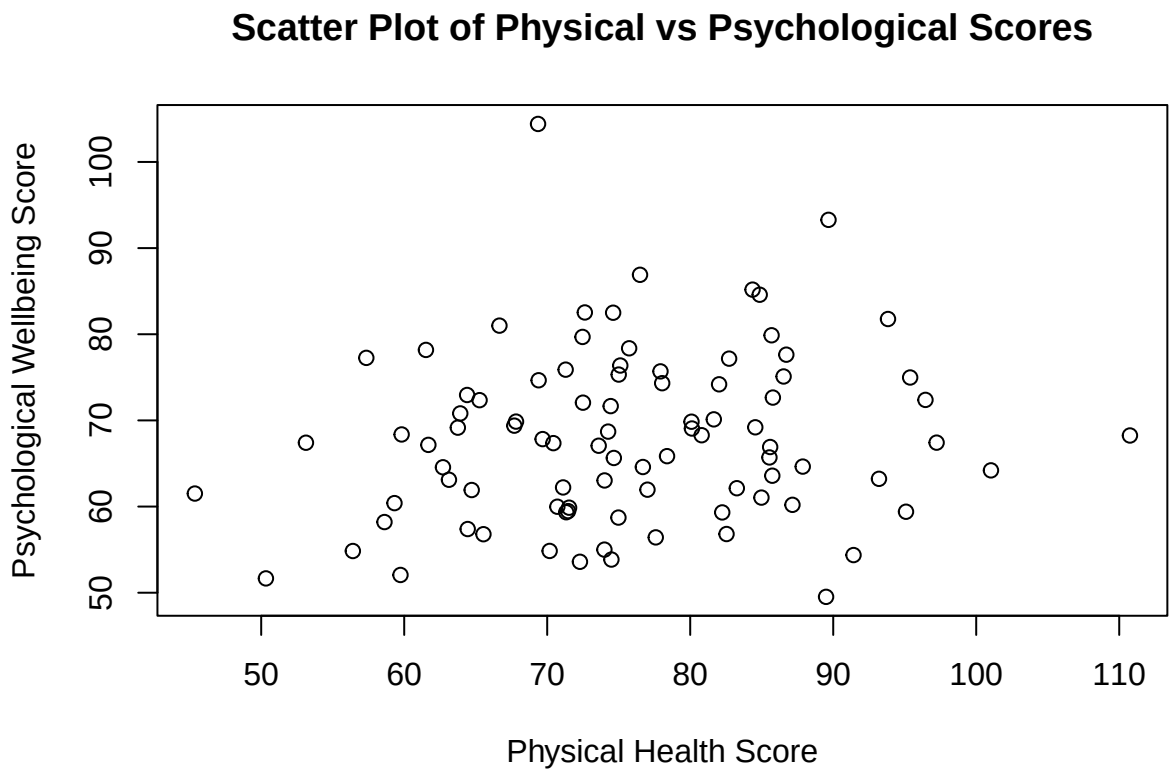
1. For **Program A** and **Program B**, both physical health and psychological wellbeing scores show no evidence of deviation from normality, as all p-values are greater than 0.05.

2. For **Program C**, physical health scores are approximately normally distributed (p-value = 0.8476), but psychological wellbeing scores show significant deviation from normality (p-value = 0.01417).

3. Overall, most of the data appear to meet the normality assumption, except for psychological wellbeing scores in Program C, which may require further adjustments or non-parametric analysis.

### Linearity

```
plot(rehab_data_df$physical_health, rehab_data_df$psychological_wellbeing,
     xlab = "Physical Health Score", ylab = "Psychological Wellbeing Score",
     main = "Scatter Plot of Physical vs Psychological Scores")
```




**Findings:**

The scatter plot shows the relationship between physical health scores and psychological wellbeing scores, with points scattered across the plot. Additionally, the data points exhibit variability, with some outliers at both high and low ranges of scores, indicating that other factors might influence these scores.

### Homogeneity of Variances and Covariance

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
leveneTest(physical_health ~ program, data = rehab_data_df)
```

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

```
leveneTest(psychological_wellbeing ~ program, data = rehab_data_df)
```

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

```
library(biotools)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## ---
## biotools version 4.2
```

```
boxM(rehab_data_df[, c("physical_health", "psychological_wellbeing")], rehab_data_df$program)
```

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

**Findings:**

The results of **Levene's Test for Homogeneity of Variance** show that the variances across groups are not significantly different for both tests ($F(2, 87) = 1.0087$, $p = 0.3689$ and $F(2, 87) = 1.6497$, $p = 0.1981$). Similarly, **Box's M-test for Homogeneity of Covariance Matrices** indicates no significant differences in covariance matrices across groups (Chi-Sq $= 6.6998$, df $= 6$, $p = 0.3495$). These results suggest that the assumption of homogeneity of variance and covariance is **satisfied** for the data.

## Fitting MANOVA Model

```
# Fit MANOVA model
manova_model <- manova(cbind(physical_health, psychological_wellbeing) ~ program, data = rehab_data_df)

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
```

**Findings:**

The **MANOVA results** using Pillai's trace indicate that the effect of the program on the combination of physical health and psychological wellbeing scores is **statistically significant** (Pillai = 0.28046, $F(4, 174) = 7.0948$, $p < 0.001$). This suggests that the program has a significant impact on the two dependent variables.

## The ANOVA and Post-Hoc Tests

```
# Perform ANOVA for each dependent variable
anova_physical <- aov(physical_health ~ program, data = rehab_data_df)
summary(anova_physical)
```

```
##             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
```

```
anova_psychological <- aov(psychological_wellbeing ~ program, data = rehab_data_df)
summary(anova_psychological)
```

```
##             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
```

```
# Tukey's HSD for pairwise comparisons
tukey_physical <- TukeyHSD(anova_physical)
tukey_psychological <- TukeyHSD(anova_psychological)

# Display Tukey's HSD results
tukey_physical
```

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

```
tukey_psychological
```

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

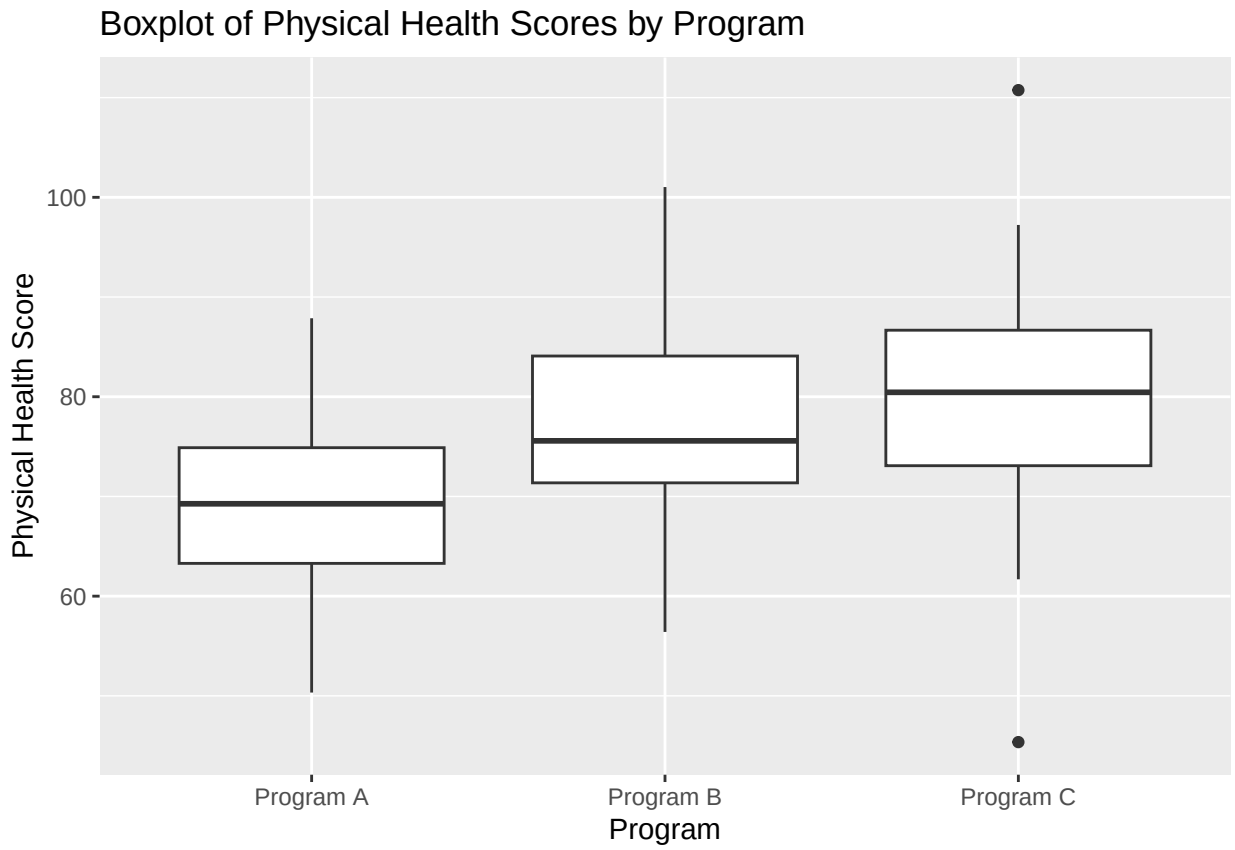
**Findings:**

The ANOVA results reveal that the **program significantly affects both physical health** ($F(2, 87) = 7.595$, $p < 0.001$) and **psychological wellbeing** ($F(2, 87) = 8.607$, $p < 0.001$). Tukey's HSD post-hoc tests show that **Program C significantly improves both physical health and psychological wellbeing** compared to Program A, with p-values of 0.0008 and 0.0005, respectively. Additionally, **Program C outperforms Program B in psychological wellbeing** ($p = 0.008$), but there is no significant difference in physical health between these two programs.

# Additional Visualizations

## Boxplot of Physical Health Scores by Program

```
library(ggplot2)
ggplot(rehab_data_df, aes(x = program, y = physical_health)) +
  geom_boxplot() +
  ggtitle("Boxplot of Physical Health Scores by Program") +
  xlab("Program") + ylab("Physical Health Score")
```

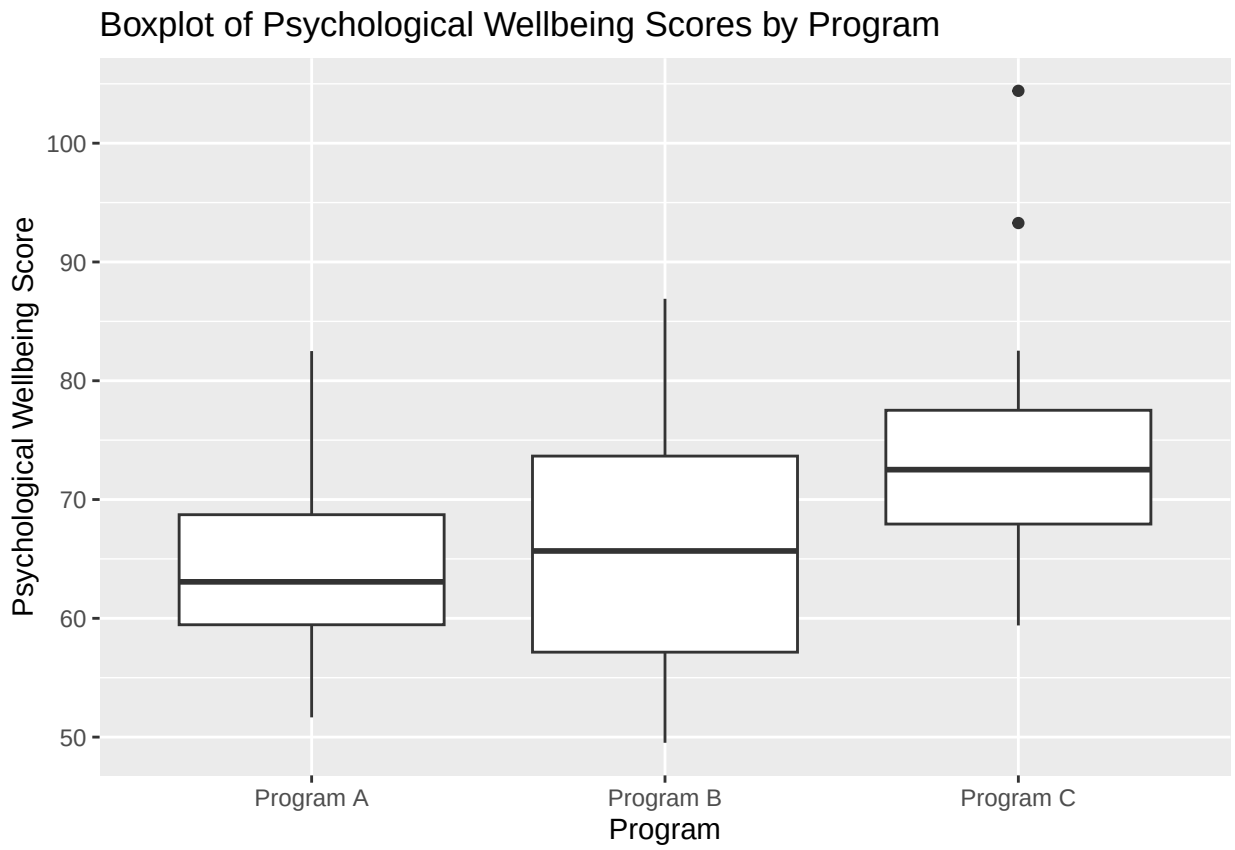



**Findings:**

The boxplot illustrates the distribution of physical health scores across the three programs (A, B, and C). **Program C exhibits the highest median score**, followed by Program B, while **Program A has the lowest median score** with a narrower interquartile range. There are a few outliers observed in both Program B and Program C, indicating variability in individual responses within these programs.

### Boxplot of Psychological Wellbeing Scores by Program

```
ggplot(rehab_data_df, aes(x = program, y = psychological_wellbeing)) +
  geom_boxplot() +
  ggtitle("Boxplot of Psychological Wellbeing Scores by Program") +
  xlab("Program") + ylab("Psychological Wellbeing Score")
```

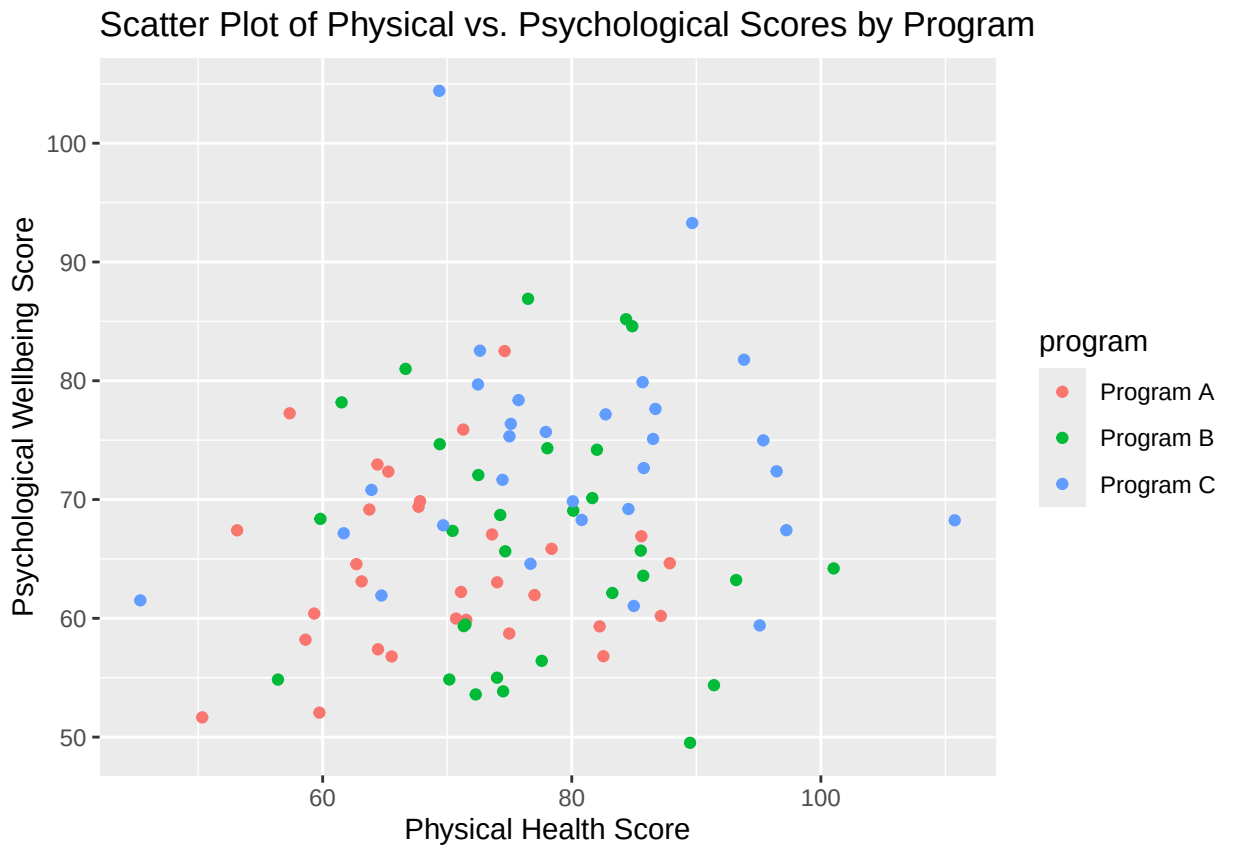



**Findings:**

Here now in our **Boxplot of Psychological Wellbeing Scores by Program**, just like on a previous plot Program C also seems to show an oulier indicating variability. However, Program C maintain still as the highest program apart the 2.

### Scatter Plot of Physical vs. Psychological Scores by Program

```
ggplot(rehab_data_df, aes(x = physical_health, y = psychological_wellbeing, color = program)) +
  geom_point() +
  ggtitle("Scatter Plot of Physical vs. Psychological Scores by Program") +
  xlab("Physical Health Score") + ylab("Psychological Wellbeing Score")
```




**Findings:**

Using the scatter plot, it is more clear here that **Program C contains a scatter variables** in which what indicate a greater outlier on our previous outlines.