Exploratory Data Analysis: Mental Health of Athletes

Ravenna Miller

2025-06-27

Introduction

Mental health is a growing concern in the world of athletics, where performance pressure, training load, and stress may influence psychological well-being. This project explores a simulated dataset designed to reflect the mental health experiences of athletes.

Variables Under Study: - `mental_health_score`: A continuous variable ranging from 0–100, where higher values indicate better self-reported mental health. - `training_hours`: A continuous variable measuring how many hours an athlete trains per week. - `stress_level`: A categorical variable with three levels — `Low`, `Moderate`, `High` — indicating self-reported psychological stress.

Research Question:
What is the relationship between training hours and mental health? Does self-reported stress modify this relationship? We are interested in whether athletes with higher stress or more intense training tend to report better or worse mental health.

Univariate Exploration

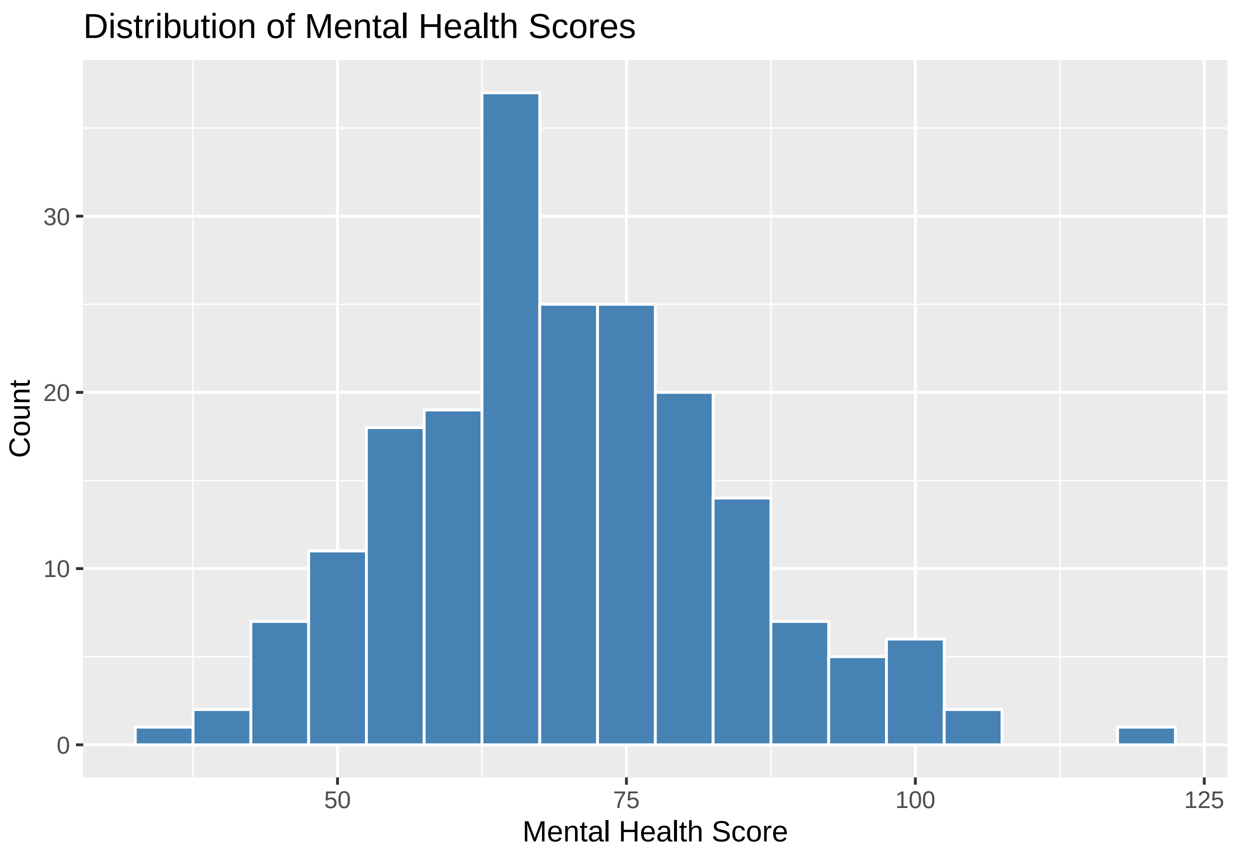
We begin by describing each variable individually using summary statistics and visualizations.

Mental Health Score

```r
summary(athlete_data$mental_health_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.36   60.61   69.12   69.87   78.53  118.62
```

```r
ggplot(athlete_data, aes(x = mental_health_score)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Mental Health Scores",
       x = "Mental Health Score", y = "Count")
```



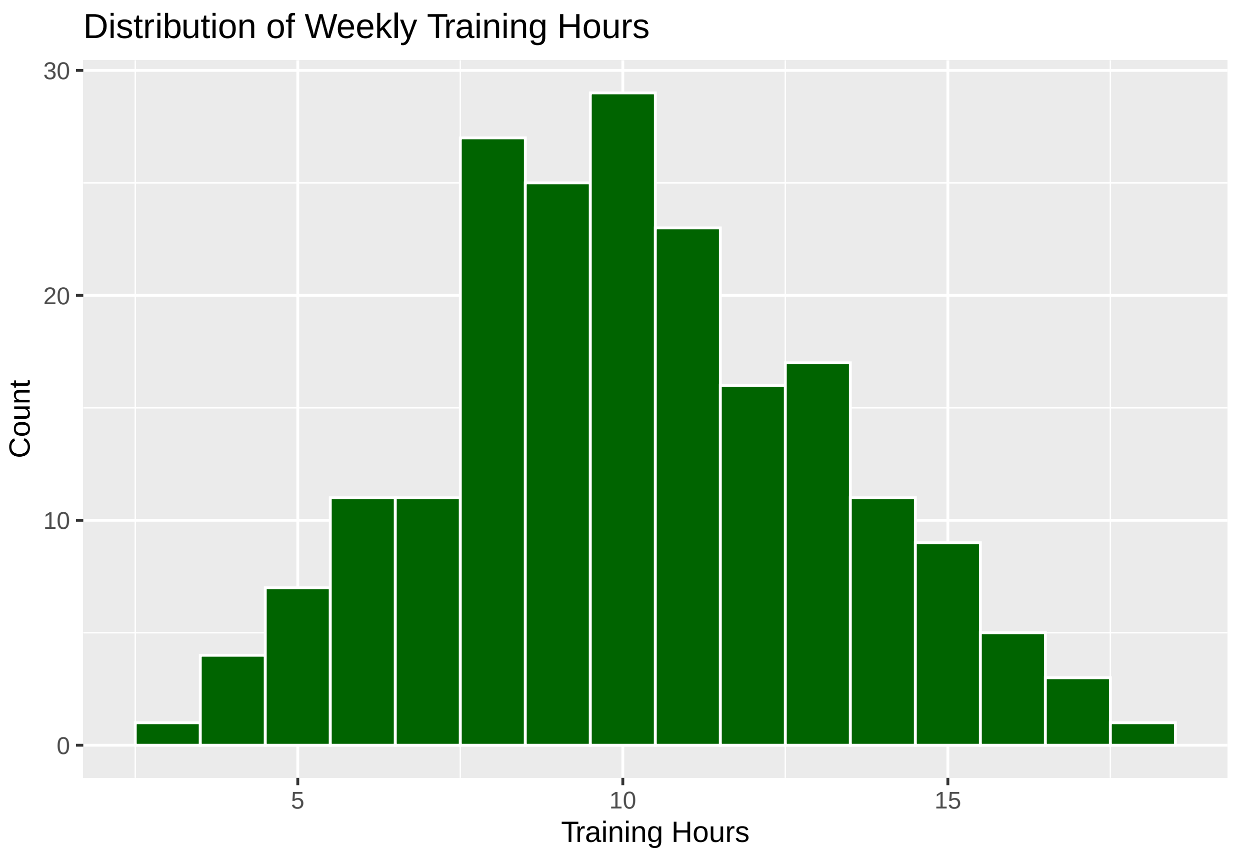
Description:
The average mental health score is approximately 70 with a standard deviation of 15. The distribution is roughly symmetric, indicating a fairly balanced spread of responses across athletes, with most scoring between 55 and 85.

Training Hours

```
summary(athlete_data$training_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.602   8.228  10.068  10.126  12.144  17.714
```

```
ggplot(athlete_data, aes(x = training_hours)) +
  geom_histogram(binwidth = 1, fill = "darkgreen", color = "white") +
  labs(title = "Distribution of Weekly Training Hours",
       x = "Training Hours", y = "Count")
```



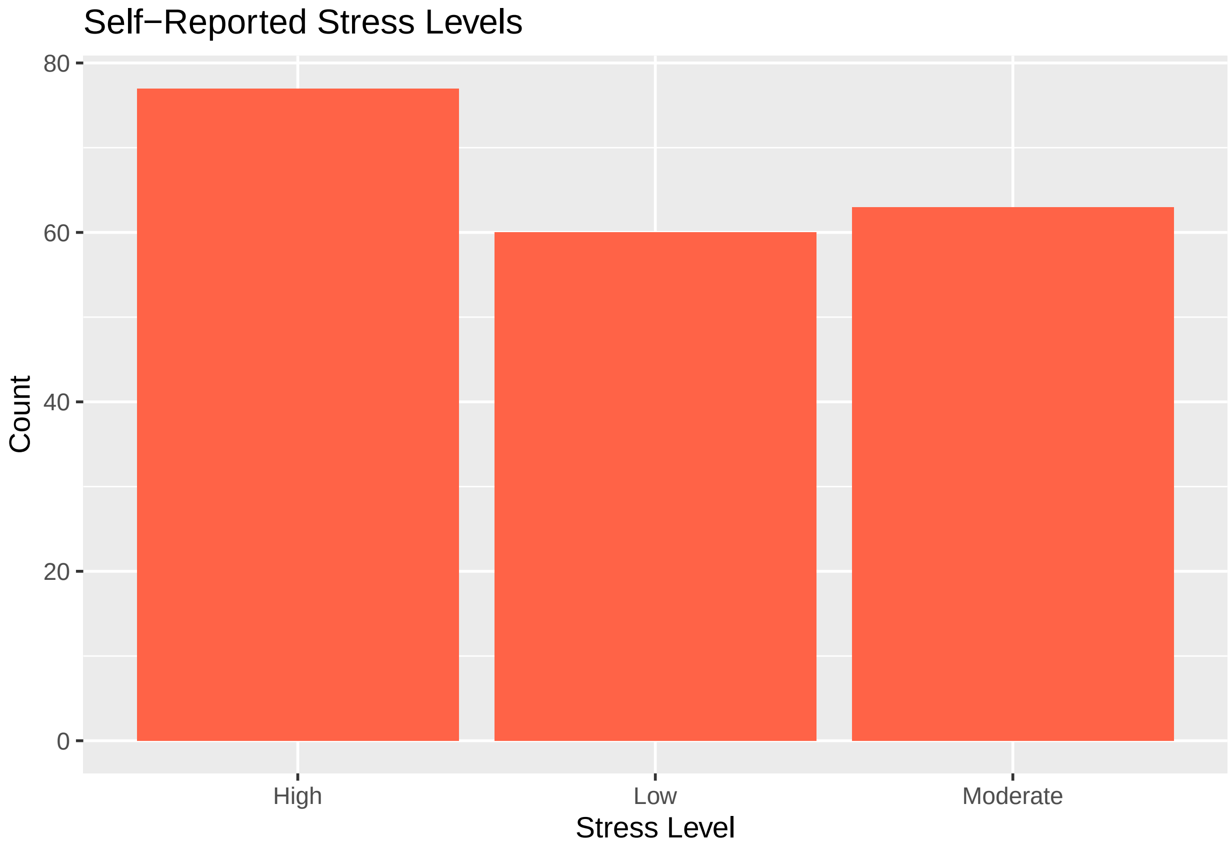
Description:
Training hours average around 10 per week with a standard deviation of about 3 hours. Most athletes train between 7 to 13 hours weekly. The distribution is slightly skewed right, suggesting a few athletes train substantially more than average.

Stress Level

```
table(athlete_data$stress_level)
```

```
##
##     High      Low Moderate
##       77       60       63
```

```
ggplot(athlete_data, aes(x = stress_level)) +
  geom_bar(fill = "tomato") +
  labs(title = "Self-Reported Stress Levels", x = "Stress Level", y = "Count")
```



Description:
Stress levels are fairly evenly distributed among Low, Moderate, and High categories, indicating a balanced representation of psychological stress within the sample.

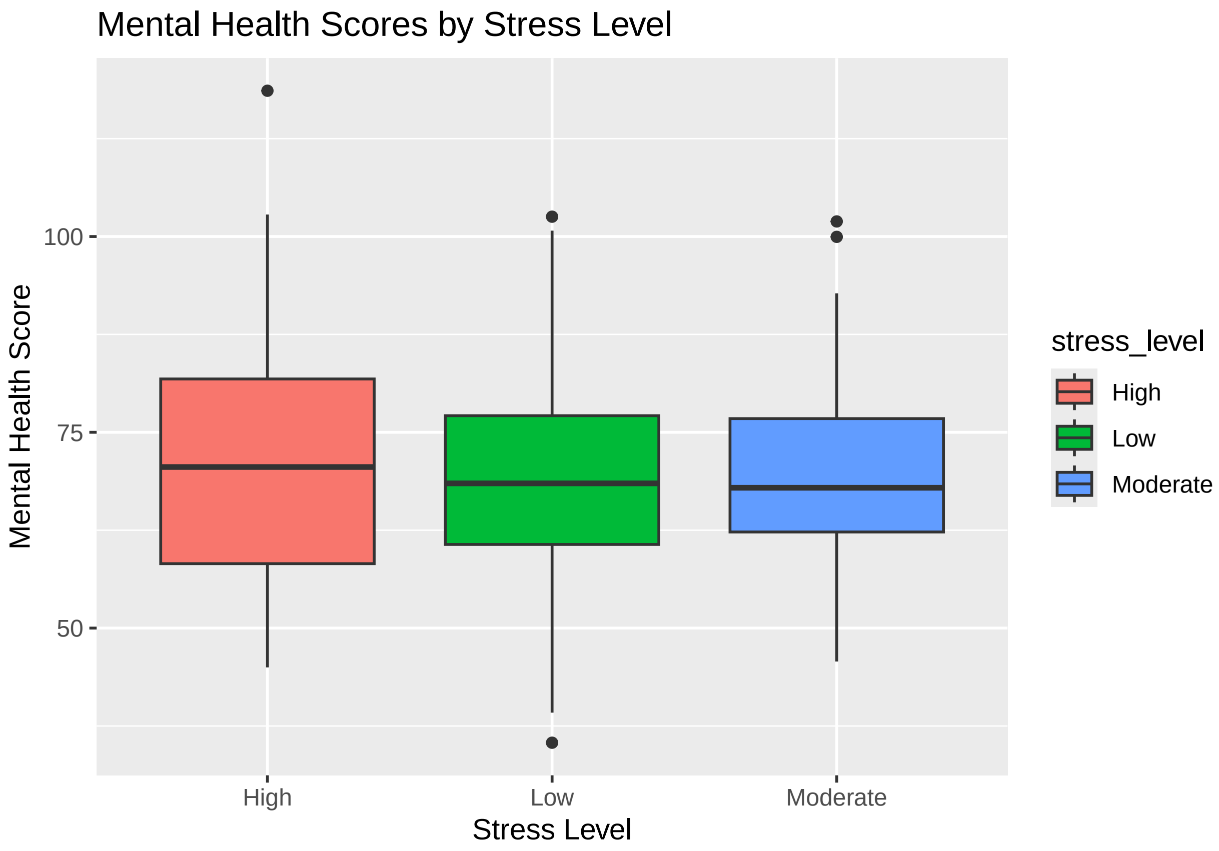
Bivariate Exploration

Mental Health Score by Stress Level

```
athlete_data %>%
  group_by(stress_level) %>%
  summarize(mean_mh = mean(mental_health_score),
            sd_mh = sd(mental_health_score),
            n = n())
```

```
## # A tibble: 3 x 4
##   stress_level mean_mh sd_mh     n
##   <chr>          <dbl> <dbl> <int>
## 1 High            70.9  15.8    77
## 2 Low             68.7  14.1    60
## 3 Moderate        69.8  12.1    63
```

```
ggplot(athlete_data, aes(x = stress_level, y = mental_health_score, fill = stress_level)) +
  geom_boxplot() +
  labs(title = "Mental Health Scores by Stress Level",
       x = "Stress Level", y = "Mental Health Score")
```

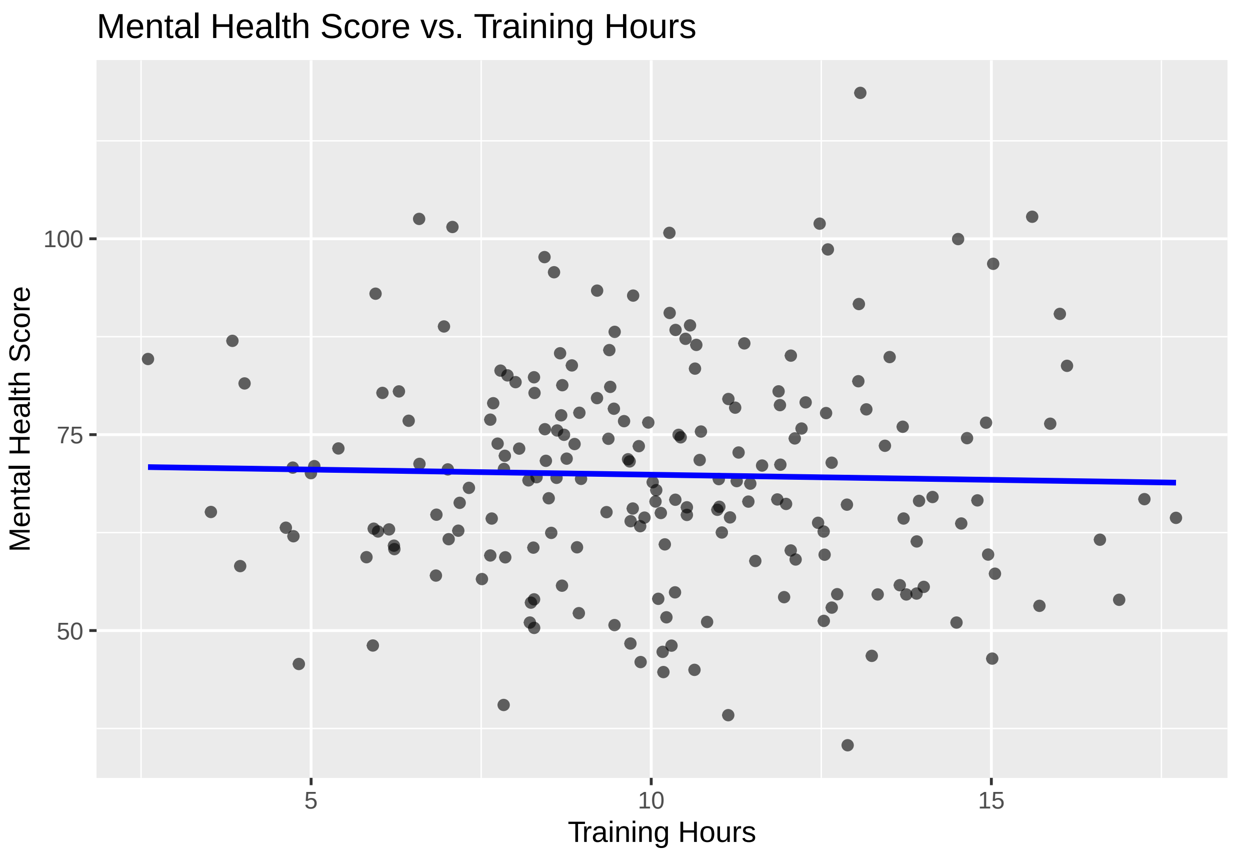


Description:
Athletes with low stress levels report significantly higher mental health scores compared to those with high stress. This trend is consistent with expectations — greater psychological stress appears to associate with reduced mental well-being.

Mental Health Score vs Training Hours

```
ggplot(athlete_data, aes(x = training_hours, y = mental_health_score)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Mental Health Score vs. Training Hours",
       x = "Training Hours", y = "Mental Health Score")
```

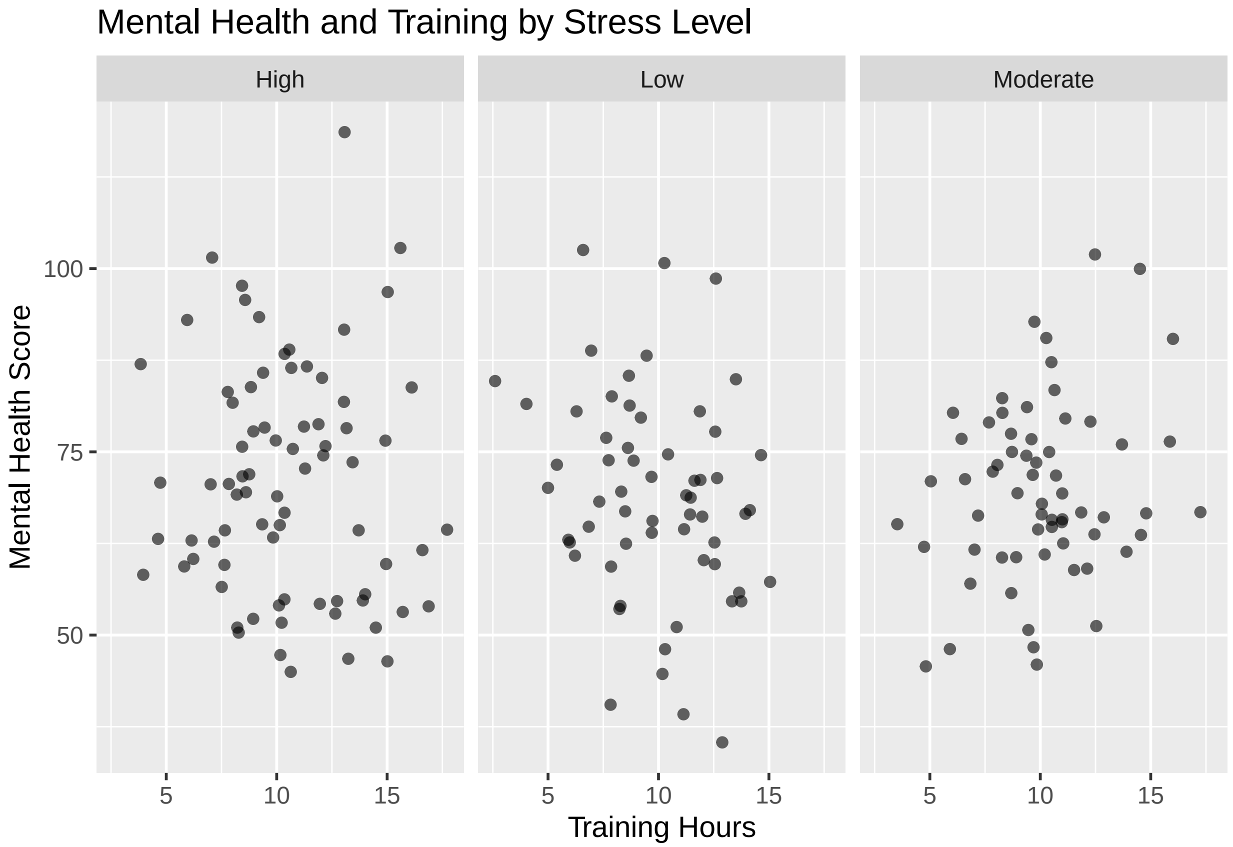
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Description:
There appears to be a mild positive relationship between training hours and mental health score. Athletes training more tend to report slightly better mental health, though variability remains high.

Panel Plot: Mental Health vs. Training by Stress Level

```r
ggplot(athlete_data, aes(x = training_hours, y = mental_health_score)) +
  geom_point(alpha = 0.6) +
  facet_wrap(~stress_level) +
  labs(title = "Mental Health and Training by Stress Level",
       x = "Training Hours", y = "Mental Health Score")
```



Description:
When separated by stress level, we see a clearer pattern: for athletes with low stress, mental health increases more consistently with training. For high-stress athletes, the relationship appears weaker or flat. This suggests stress may moderate the training-mental health relationship.

Conclusion

This exploratory data analysis revealed several key insights: - Athletes with **low stress levels** report better mental health on average. - More **training hours** tend to associate with higher mental health scores, especially among athletes with lower stress. **- Stress level may be a modifying factor** in the relationship between training and mental health.

These findings are **descriptive only** and should not be interpreted as statistically significant or causal. Further statistical analysis would be needed to confirm these patterns.