Premier League R Analysis

Goal:

My analysis aims to investigate the relationship between the age of the players and the goals-per-game ratio compared to younger players, along with the number of games started by Premier League players and their performance, as measured by the 'Goals-per-game' ratio.

Dataset:

The dataset contains the statistics of Premier League players, including some important factors their age, number of games played, number of games started, goals, scoring frequency, and goal conversion rates.

Models Used:

Model for Hypothesis 1 (Age and Performance): ANOVA and Tukey's Honest Significant Difference (HSD) post-hoc test were used to compare goals per game ratios across different age groups.
Model for Hypothesis 2 (Playing Time and Performance): Linear regression was used to assess the relationship between the number of games a player starts and their goals per game ratio.

Data Source:

Taken from ‘https://www.kaggle.com/datasets/jackhan9811/the-premier-league-yearly-dataset-from-18192021’ and further simplified to contain relevant and remove missing data.

Analysis Hypothesis and Results from R-Output:

Hypothesis 1: Older players have a lower goals-per-game ratio compared to younger players..

```
> summary(anova_result)
              Df Sum Sq Mean Sq F value   Pr(>F)    
AgeGroup       3  0.291 0.09704   6.151 0.000376 ***
Residuals   1274 20.099 0.01578                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = Goals.per.game ~ AgeGroup, data = data)
$AgeGroup
                    diff          lwr         upr     p adj
25-30-<25    0.022606065  0.001665917  0.04354621 0.0284420
31-35-<25   -0.005268684 -0.030293448  0.01975608 0.9487995
>35-<25     -0.046836536 -0.105245445  0.01157237 0.1659978
31-35-25-30 -0.027874749 -0.051723628 -0.00402587 0.0143062
>35-25-30   -0.069442601 -0.127357460 -0.01152774 0.0111909
>35-31-35   -0.041567852 -0.101081547  0.01794584 0.2753282
```

The ANOVA and subsequent Tukey's HSD indicated significant differences in goals-per-game ratios between some age groups, suggesting a decline in scoring as players age

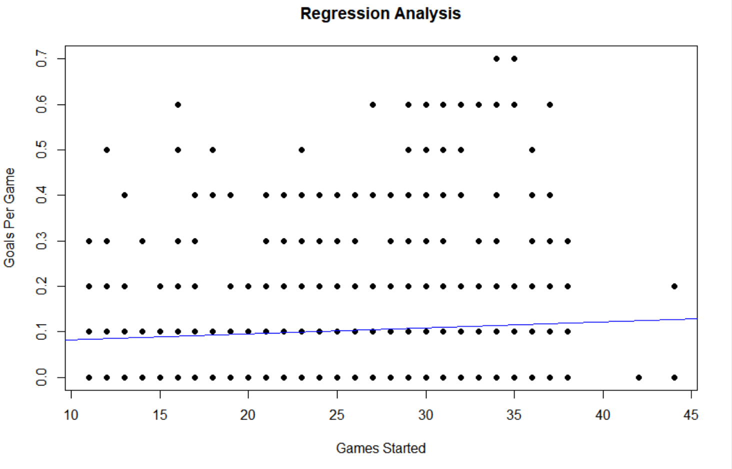
Hypothesis 2: Players who start more matches have a higher 'Goals per game' ratio compared to those who are often substitutes. The linear regression analysis showed a positive relationship between the number of games started and the goals-per-game ratio, indicating that players who start more tend to score more goals per game.

```
> summary(model)
Call:
lm(formula = Goals.per.game ~ Started, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.12671 -0.10084 -0.01636  0.01468  0.58622 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0697958  0.0148110   4.712 2.86e-06 ***
Started     0.0012936  0.0005726   2.259   0.0241 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1373 on 851 degrees of freedom
Multiple R-squared:  0.005961, Adjusted R-squared:  0.004792 
F-statistic: 5.103 on 1 and 851 DF,  p-value: 0.02414
```



The linear regression analysis showed a positive relationship between the number of games started and the goals-per-game ratio, indicating that players who start more tend to score more goals per game.

<u>Conclusion:</u>

<u>Hypothesis 1:</u>The hypothesis that older players have a lower goals per game ratio compared to younger players is partially supported.

The peak performance in terms of 'Goals per game' appears to be in the 25-30 age group, which is significantly higher than the younger group (<25) and the older group (>35).

These results indicate that mid-career players tend to have a higher goal-scoring ratio.

<u>Hypothesis 2:</u> The initial hypothesis is true with some caveats; The regression analysis supports this hypothesis to a small extent; there is a statistically significant positive relationship between the number of games started and the 'Goals per game' ratio. But the effect size is small, and the low R-squared value indicates that 'Started' is not a strong predictor of 'Goals per game'.