FIFA PLAYER PERFORMANCE ANALYSIS

DRASHTI PATEL(7964120)

12/01/2024

Question

The purpose of this project is to analyze the factors that influence the overall performance of FIFA players during the 2024 season. As the response variable, the Overall rating reflects the skill level, value, and current performance of a player in professional soccer.This study examines how two explanatory variables, Hits and Potential, impact the response variable.

The hits measures a player's popularity and engagement, measured by the amount of fan interactions and media mentions.The hypothesis is that Hits will boost overall performance. Higher Hits may lead to increased confidence, motivation, and recognition, all of which can improve performance on the field.Potential represents how good a player is going to be during their career. Potential is hypothesized to positively influence Overall. Teams and coaches often consider potential ratings when assessing current performance levels when assessing players' development and abilities.Using these variables, the project aims to understand how engagement and growth potential contribute to player performance, identify the strongest predictor, and provide insight into factors that drive success.

Data Set

```r
# Loading the dataset
fifa <- read.csv("FIFA.csv")

# Selecting relevant columns
fifa_subset <- fifa[,c("Name","Hits","Potential","Overall")]

# Creating the table
kable(head(fifa_subset,10),
      caption= "Table 1: FIFA Player Performance Dataset",
      booktabs = TRUE,
      align = 'c')
```

Table 1: Table 1: FIFA Player Performance Dataset

Name	Hits	Potential	Overall
Lionel Messi	299	94	94
Cristiano Ronaldo	276	93	93
Neymar Jr	186	92	92
Virgil van Dijk	127	92	91
Jan Oblak	47	93	91

Name	Hits	Potential	Overall
Kevin De Bruyne	119	91	91
Robert Lewandowski	89	91	91
Eden Hazard	66	91	91
Alisson	53	91	90
Mohamed Salah	94	90	90

Source: Mishra, A. (2024)._FIFA 2021 complete player data Data set. Kaggle._ https://www.kaggle.com/datasets/aayushmishra1512/fifa-2021-complete-player-data

Variable Definitions and Units:

1)*Name:* Every FIFA player has a unique name, which serves as their unique identifier. 2)*Hits:* Player hits are the number of interactions or engagements they receive, which represent how popular they are. An increase in hits indicates an increase in fan attention and media coverage. 3)*Potential:* A player's potential is measured on a scale of 1 to 100 based on the highest performance they are projected to achieve during their career. 4)*Overall:*Player's current performance rating, measured on a scale of 1 to 100, which serves as the response variable.

Scatterplots and Coefficient of Determination (R^2)

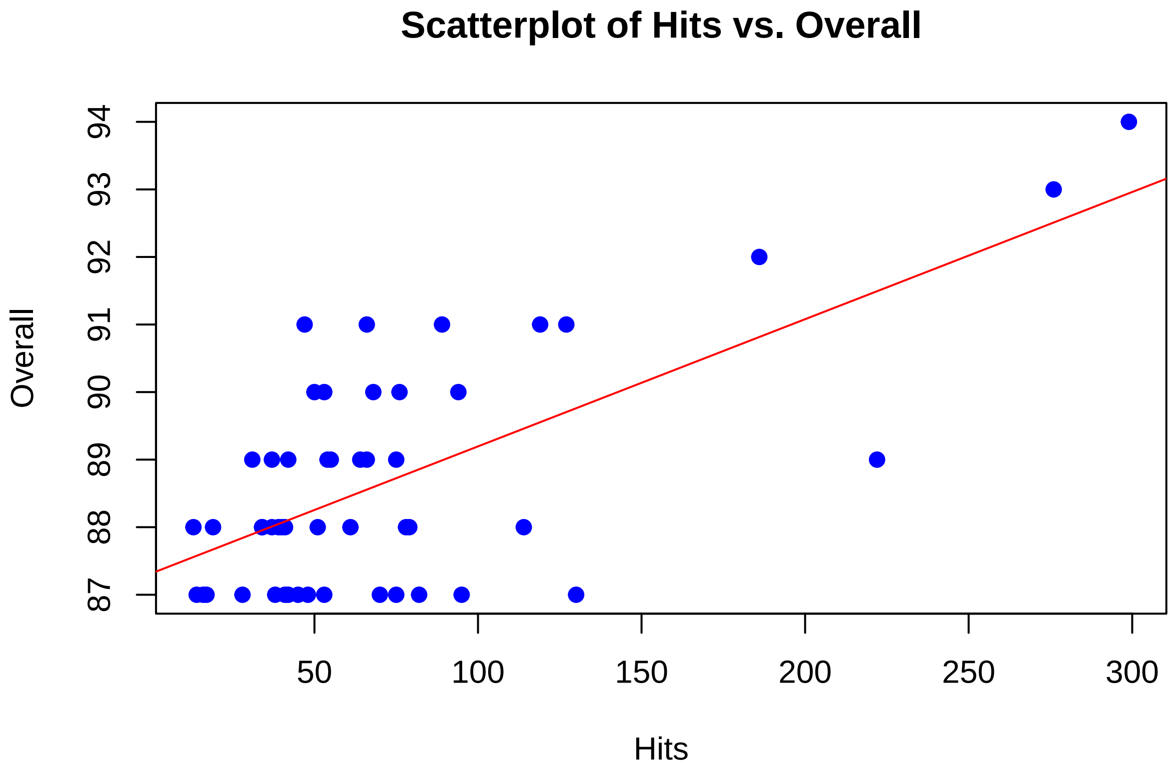
To assess the relationship between each explanatory variable and the **Overall**, scatterplots were created, and the coefficient of determination (R^2) was calculated to measure how well the variables explain changes in Overall.

1. Hits vs. Overall

The following scatterplot visualizes the relationship between **Overall** and **Hits**. The regression line added to the plot displays the trend.

```r
# Scatterplot for Hits vs. Overall
plot(fifa$Hits, fifa$Overall,
     main = "Scatterplot of Hits vs. Overall",
     xlab = "Hits",
     ylab = "Overall",
     pch = 19,
     col = "blue")

# Adding a linear regression line
abline(lm(Overall ~ Hits, data = fifa), col = "red")
```



```r
# Calculating R-squared
summary(lm(Overall ~ Hits, data = fifa))
```

```
##
## Call:
## lm(formula = Overall ~ Hits, data = fifa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76074 -0.98147  0.01853  0.97198  2.80200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.313066   0.288573 302.569  < 2e-16 ***
## Hits         0.018828   0.003075   6.123 1.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.293 on 48 degrees of freedom
## Multiple R-squared:  0.4385, Adjusted R-squared:  0.4268
## F-statistic: 37.49 on 1 and 48 DF,  p-value: 1.629e-07
```

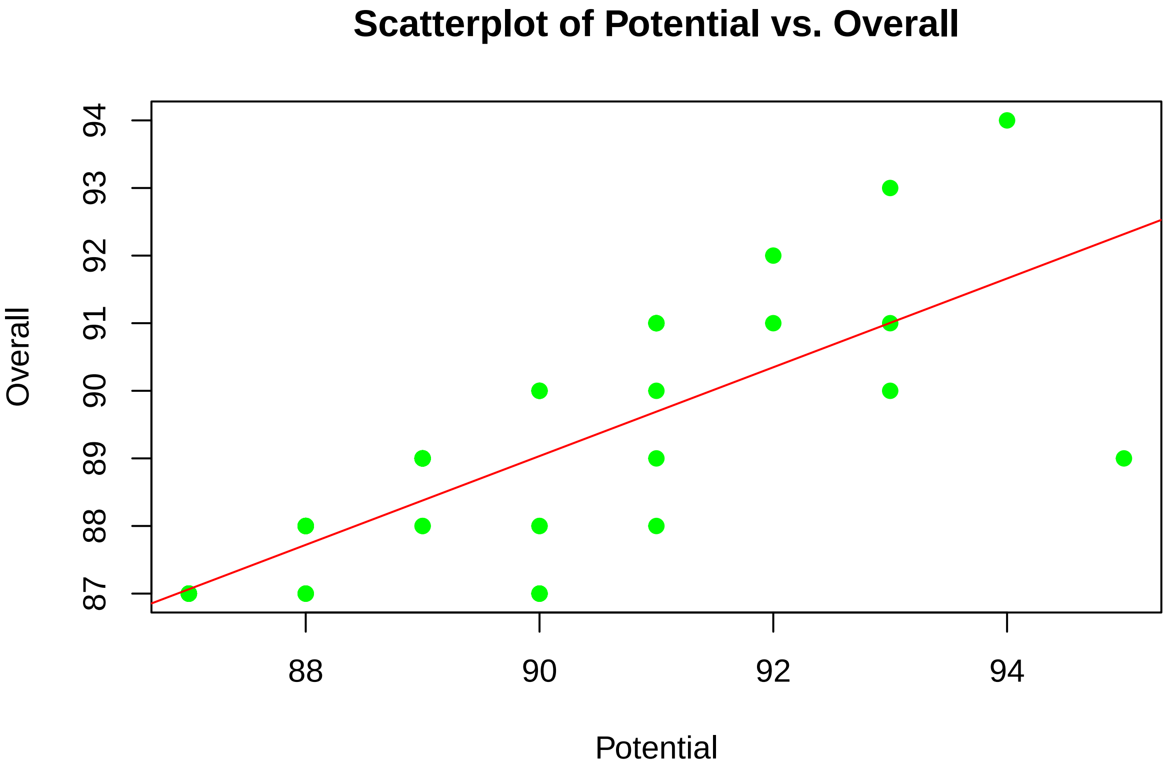
$$R^2 = 0.4385$$

This indicates that approximately 43.85% of the variability in *Overall* can be explained by *Hits*.

2. Potential vs. Overall

The following scatterplot visualizes the relationship between **Overall** and **Potential**. The regression line added to the plot displays the trend.

```
#Scatterplot for Potential vs. Overall
plot(fifa$Potential, fifa$Overall,
     main = "Scatterplot of Potential vs. Overall",
     xlab = "Potential",
     ylab = "Overall",
     pch = 19,
     col = "green")
# Adding a linear regression line
abline(lm(Overall ~ Potential, data = fifa), col = "red")
```



```
# Calculating R-squared
summary(lm(Overall ~ Potential, data = fifa))
```

```
##
## Call:
## lm(formula = Overall ~ Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3181 -0.6130  0.2788  0.6221  2.3386
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.93265    7.20616   4.154 0.000134 ***
## Potential    0.65669    0.08053   8.154 1.29e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.117 on 48 degrees of freedom
## Multiple R-squared:  0.5808, Adjusted R-squared:  0.572
## F-statistic: 66.49 on 1 and 48 DF,  p-value: 1.288e-10
```

$$R^2 = 0.5808$$

This indicates that approximately 58.08% of the variability in *Overall* can be explained by *Potential*, indicating a stronger relationship compared to *Hits.*

Summary for the above calculations: The scatterplot of Hits vs. Overall shows a moderate positive correlation (Rˆ2 =0.4385), suggesting that higher engagement leads to better performance. The Potential vs. Overall plot shows a stronger correlation (Rˆ2 =0.5808), indicating that a player’s potential is a more significant predictor of performance. Combining both variables improves the model’s explanatory power, with an adjusted Rˆ2 =0.6103, providing a more accurate view of overall performance.

Preliminary Model

To explore the relationship between the response variable (Overall) and the explanatory variables (Hits and Potential), the following models were fitted:

1) Model 1: Overall vs. Hits

```r
fifa1 <- lm(Overall ~ Hits, data = fifa)
summary(fifa1)
```

```
##
## Call:
## lm(formula = Overall ~ Hits, data = fifa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76074 -0.98147  0.01853  0.97198  2.80200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.313066   0.288573 302.569  < 2e-16 ***
## Hits         0.018828   0.003075   6.123 1.63e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.293 on 48 degrees of freedom
## Multiple R-squared:  0.4385, Adjusted R-squared:  0.4268
## F-statistic: 37.49 on 1 and 48 DF,  p-value: 1.629e-07
```

```
# R^2 = 0.4385, Adj_r^2 = 0.4268
```

The first model explores the relationship between Overall (Y) and Hits (X_1). The regression line is:

$$\hat{Y} = 87.31 + 0.01 \cdot X_1$$

Adjusted R-Squared: 0.4268
Interpretation: This model indicates that Hits (X_1) explain approximately 42.68% of the variability in Overall (Y). The positive coefficient for X_1 suggests that as a player's popularity and engagement increase, their Overall performance rating is expected to increase as well.

2) Model 2: Overall vs. Potential

```
fifa2 <- lm(Overall ~ Potential, data = fifa)
summary(fifa2)
```

```
##
## Call:
## lm(formula = Overall ~ Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3181 -0.6130  0.2788  0.6221  2.3386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.93265    7.20616   4.154 0.000134 ***
## Potential    0.65669    0.08053   8.154 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.117 on 48 degrees of freedom
## Multiple R-squared:  0.5808, Adjusted R-squared:  0.572
## F-statistic: 66.49 on 1 and 48 DF,  p-value: 1.288e-10
```

```
# R^2 = 0.5808, Adj_r^2 = 0.572
```

The second model explores the relationship between Overall (Y) and Potential (X_2). The regression line is:

$$\hat{Y} = 29.93 + 0.65 \cdot X_2$$

Adjusted R-Squared: 0.572
Interpretation: This model indicates that Potential (X_2) explain approximately 57.2% of the variability in Overall (Y). The positive coefficient for X_2 suggests that as potential rating increase, their Overall performance rating is expected to increase as well.

3) Combined Model: Overall vs. Hits + Potential

```
# Combined model R^2
fifa_combined <- lm(Overall ~ Hits + Potential, data = fifa)
summary(fifa_combined)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6311 -0.4591  0.1151  0.7563  1.6093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.723110   8.975289   4.871 1.30e-05 ***
## Hits         0.008054   0.003369   2.391   0.0209 *
## Potential    0.496000   0.102092   4.858 1.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.066 on 47 degrees of freedom
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6103
## F-statistic: 39.37 on 2 and 47 DF,  p-value: 9.048e-11
```

```
# Adjusted R^2 = 0.6103
```

The combined model includes both explanatory variables (Hits and Potential) to predict Overall. The regression line is:

$$\hat{Y} = 43.72 + 0.008 \cdot X_1 + 0.49 \cdot X_2$$

Adjusted R-Squared: 0.6103
Interpretation: The combined model explains 61.03% of the variability in Overall (Y), which is a notable improvement compared to the individual models. Both Hits(X_1) and Potential (X_2) contribute positively to predicting Overall.

4)Full Second-Order Model:

To enhance the analysis, a second-order model was fitted, including interaction and quadratic terms:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1^2 + \beta_4 \cdot X_2^2 + \beta_5 \cdot X_1 \cdot X_2$$

```
lm_second_order <- lm(Overall ~ Hits + Potential + I(Hits^2) + I(Potential^2) + Hits:Potential, data =
summary(lm_second_order)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential + I(Hits^2) + I(Potential^2) +
##     Hits:Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1952 -0.4874  0.2781  0.5165  1.5149
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.286e+02  3.417e+02  -1.840   0.0726 .
## Hits             1.776e-01  2.016e-01   0.881   0.3832
## Potential        1.532e+01  7.703e+00   1.989   0.0529 .
## I(Hits^2)        1.002e-04  5.184e-05   1.932   0.0598 .
## I(Potential^2) -8.160e-02  4.343e-02  -1.879   0.0669 .
## Hits:Potential -2.105e-03  2.327e-03  -0.905   0.3705
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9948 on 44 degrees of freedom
## Multiple R-squared:  0.6952, Adjusted R-squared:  0.6606
## F-statistic: 20.08 on 5 and 44 DF,  p-value: 2.23e-10
```

Regression Line:

$$\hat{Y} = -628.6 + 0.1776 \cdot X_1 + 15.32 \cdot X_2 + 0.0001002 \cdot X_1^2 - 0.0816 \cdot X_2^2 - 0.002105 \cdot X_1 \cdot X_2$$

Adjusted R-Squared: 0.6606 **Interpretation**: The adjusted R-squared value indicates that approximately 87.86% of the variability in Overall performance (Y) is explained by the model which includes both the quadratic and interaction terms.

ANOVA Test

To identify if at least one of the terms in the full second-order model is significant, an ANOVA test is performed:

```
anova(lm_second_order)
```

```
## Analysis of Variance Table
##
## Response: Overall
##                Df Sum Sq Mean Sq F value    Pr(>F)
## Hits            1 62.653  62.653 63.3110 4.653e-10 ***
## Potential       1 26.821  26.821 27.1026 4.869e-06 ***
## I(Hits^2)       1  1.157   1.157  1.1689  0.285525
## I(Potential^2)  1  7.897   7.897  7.9794  0.007085 **
## Hits:Potential  1  0.810   0.810  0.8186  0.370524
## Residuals      44 43.543   0.990
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

To assess the significance of the higher-order terms, an ANOVA test was performed:

1. **Level of Significance**:
   $\alpha = 0.05$
2. **Hypotheses**:

   - Null Hypothesis: $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$
     (The higher-order terms do not contribute significantly to the model.)

   - Alternative Hypothesis: H_A: At least one of $\beta_3, \beta_4, \beta_5 \neq 0$.

3. **Decision Rule**:
   Reject H_0 if p-value $\leq \alpha$.

4. **Test Statistic**:
   $F = 20.08$

5. **P-Value**:
   $p = 2.23e \times 10^{-10}$

6. **Conclusion**:
   Since $p = 2.23e \times 10^{-10} < \alpha = 0.05$, we reject H_0. Therefore, there is sufficient evidence to conclude that at least one of the higher-order terms contributes significantly to the model.

Model Refinement

#Step 1: Examine the Significance of Coefficients in the Full Model

The `summary()` function was used on the full second-order model to evaluate the significance of each coefficient. The results showed the following p-values:

- **Intercept**: $p = 0.0726$ (Not Significant)
- **Hits** (X_1): $p = 0.3832$ (Not Significant)
- **Potential** (X_2): $p = 0.0529$ (Not Significant)
- **Hits^2** (X_1^2): $p = 0.0598$ (Not Significant)
- **Potential^2** (X_2^2): $p = 0.0669$ (Not Significant)
- **Interaction** ($X_1 \cdot X_2$): $p = 0.3705$ (Not Significant)

At this stage, none of the terms in the model (X_1^2, X_2^2, or $X_1 \cdot X_2$) were statistically significant at $\alpha = 0.05$.

Step 2: Evaluate Multicollinearity Using VIF To check for potential multicollinearity, we calculated the Variance Inflation Factor (VIF) for each term in the full model using the vif() function from the car package.

```
vif(lm_second_order)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##            Hits      Potential      I(Hits^2) I(Potential^2) Hits:Potential
##      7257.69396    11536.22356       41.01236    11945.25002     8515.96735
```

Step 3: Removing X_2^2

Reason for Removal: The term X_2^2 was identified as having the largest Variance Inflation Factor (VIF) in the full model, indicating a high level of multicollinearity. To address this issue, the term was removed, and a new model was fitted.

Updated Model: The new model excludes X_2^2. The remaining terms in the model are:

- Intercept
- X_1 (Hits)
- X_2 (Potential)
- X_1^2 (Hits²)
- $X_1 \cdot X_2$ (Interaction)

```
lm_refined1 <- lm(Overall ~ Hits + Potential + I(Hits^2) + Hits:Potential, data = fifa)
summary(lm_refined1)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential + I(Hits^2) + Hits:Potential,
##     data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0356 -0.4895  0.1143  0.6747  1.6392
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.278e+01  1.622e+01   0.788   0.4349
## Hits            3.856e-01  1.731e-01   2.228   0.0310 *
## Potential       8.523e-01  1.847e-01   4.614 3.29e-05 ***
## I(Hits^2)       1.256e-04  5.143e-05   2.442   0.0186 *
## Hits:Potential -4.482e-03  2.007e-03  -2.233   0.0306 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.022 on 45 degrees of freedom
## Multiple R-squared:  0.6708, Adjusted R-squared:  0.6415
## F-statistic: 22.92 on 4 and 45 DF,  p-value: 2.237e-10
```

```
vif(lm_refined1)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = ’predictor’; see ?vif
```

```
##           Hits      Potential      I(Hits^2) Hits:Potential
##    5067.919144       6.282508      38.218873    5999.316240
```

Step 4: Removing X_1^2 (Potential²)

Reason for Removal: The term X_1^2 was identified as having the highest Variance Inflation Factor in the refined model, indicating significant multicollinearity. To address this issue, the term X_1^2 was removed, and a new model was fitted.

Updated Model: The new model excludes X_1^2. The remaining terms in the model are:

- Intercept
- X_1 (Hits)
- X_2 (Potential)
- $X_1 \cdot X_2$ (Interaction)

```
lm_refined2 <- lm(Overall ~ Hits + Potential + Hits:Potential, data = fifa)
summary(lm_refined2)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential + Hits:Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4971 -0.4413  0.1643  0.7213  1.5815
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.0550087 11.9531019   3.435 0.001267 **
## Hits            0.0449665  0.1079113   0.417 0.678837
## Potential       0.5252702  0.1339284   3.922 0.000291 ***
## Hits:Potential -0.0004002  0.0011695  -0.342 0.733731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.076 on 46 degrees of freedom
## Multiple R-squared:  0.6272, Adjusted R-squared:  0.6029
## F-statistic: 25.79 on 3 and 46 DF,  p-value: 6.153e-10
```

```
vif(lm_refined2)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##           Hits      Potential Hits:Potential
##     1777.15624        2.98035     1838.76929
```

Step 5: Removing $X_1 \cdot X_2$ (Interaction)

Reason for Removal: The interaction term $X_1 \cdot X_2$ was identified as having the highest Variance Inflation Factor (VIF = 1838.76) in the refined model, indicating significant multicollinearity. To address this issue, the term $X_1 \cdot X_2$ was removed, and a new model was fitted.

Updated Model: The new model excludes $X_1 \cdot X_2$. The remaining terms in the model are:

- Intercept
- X_1 (Hits)
- X_2 (Potential)

```r
lm_refined3 <- lm(Overall ~ Hits + Potential, data = fifa)
summary(lm_refined3)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6311 -0.4591  0.1151  0.7563  1.6093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.723110   8.975289   4.871 1.30e-05 ***
## Hits         0.008054   0.003369   2.391   0.0209 *
## Potential    0.496000   0.102092   4.858 1.36e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.066 on 47 degrees of freedom
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6103
## F-statistic: 39.37 on 2 and 47 DF,  p-value: 9.048e-11
```

```r
vif(lm_refined3)
```

```
##      Hits Potential
##  1.764983  1.764983
```

Final Refined Model

Model Summary After iterative removal of terms with high multicollinearity and insignificant p-values, the final refined model includes only the following predictors: - X_1: Hits - X_2: Potential

The refined model is as follows:

$$\hat{Y} = 43.723 + 0.008 \cdot X_1 + 0.496 \cdot X_2$$

```r
lm_final <- lm(Overall ~ Hits + Potential, data = fifa)
summary(lm_final)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6311 -0.4591  0.1151  0.7563  1.6093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.723110   8.975289   4.871 1.30e-05 ***
## Hits         0.008054   0.003369   2.391   0.0209 *
## Potential    0.496000   0.102092   4.858 1.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.066 on 47 degrees of freedom
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6103
## F-statistic: 39.37 on 2 and 47 DF,  p-value: 9.048e-11
```

Final Model Evaluation

- **Coefficients and Significance**:
  - Intercept: $\beta_0 = 43.723$, $p = 1.30 \times 10^{-5}$ (Significant)
  - X_1 (Hits): $\beta_1 = 0.008$, $p = 0.0209$ (Significant)
  - X_2 (Potential): $\beta_2 = 0.496$, $p = \times 10^{-5}$ (Significant)
- **Model Fit**:
  - Residual Standard Error: 1.066
  - Multiple R-squared: 0.6262
  - Adjusted R-squared: 0.6103
  - F-statistic: 39.37 (p-value $= 9.048 \times 10^{-11}$)
- **Variance Inflation Factors (VIF)**:
  Both predictors have acceptable VIF values, indicating no multicollinearity:
  - X_1: VIF $= 1.764983$
  - X_2: VIF $= 1.764983$

Nested F-Test Results

To check if the reduced model is a good fit compared to the full model, we perform a nested F-test. The purpose of this test is to determine whether the terms removed during the refinement process ($X_1^2, X_2^2, X_1 \cdot X_2$) are statistically insignificant.

```r
anova(lm_final, lm_second_order)
```

```
## Analysis of Variance Table
##
## Model 1: Overall ~ Hits + Potential
## Model 2: Overall ~ Hits + Potential + I(Hits^2) + I(Potential^2) + Hits:Potential
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     47 53.406
## 2     44 43.543  3    9.8633 3.3223 0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. **LEVEL OF SIGNIFICANCE**:
   $\alpha = 0.05$

2. **HYPOTHESES**:

   - Null Hypothesis (H_0): The coefficients of the removed terms are all equal to zero.

     $$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

   - Alternative Hypothesis (H_A): At least one of the coefficients of the removed terms is not equal to zero.

     $$H_A : \text{At least one of } \beta_3, \beta_4, \beta_5 \neq 0$$

3. **DECISION RULE**:
   Reject H_0 if the p-value $\leq \alpha = 0.05$.

4. **TEST STATISTIC**:
   The F-statistic was calculated as:

   $$F = 3.3223$$

5. **P-VALUE**:
   The p-value associated with the F-test is:

   $$p = 0.0282$$

6. **CONCLUSION**:
   Since $p = 0.0282$ is less than $\alpha = 0.05$, we reject H_0. This indicates that there is sufficient evidence to conclude that the removed terms $(X_1^2, X_2^2, X_1 \cdot X_2)$ contribute significantly to the model. This suggests that these terms do have an effect on the model's ability to explain Overall performance.

Final Model and Assessment

ANOVA Test for the Reduced Model

To check if the predictors in the reduced model meaningfully explain the variation in the response variable Y (Overall), we performed an ANOVA test.

```
summary(lm_final)
```

```
##
## Call:
## lm(formula = Overall ~ Hits + Potential, data = fifa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6311 -0.4591  0.1151  0.7563  1.6093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.723110   8.975289   4.871 1.30e-05 ***
## Hits         0.008054   0.003369   2.391   0.0209 *
## Potential    0.496000   0.102092   4.858 1.36e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.066 on 47 degrees of freedom
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6103
## F-statistic: 39.37 on 2 and 47 DF,  p-value: 9.048e-11
```

Test Details

1. **LEVEL OF SIGNIFICANCE**:
   $\alpha = 0.05$

2. **HYPOTHESES**:

   - Null Hypothesis (H_0): The coefficients of the predictors β_1 (Hits) and β_2 (Potential) are equal to zero.

   $$H_0 : \beta_1 = \beta_2 = 0$$

   - Alternative Hypothesis (H_A): At least one of the coefficients of the predictors is not equal to zero.

   $$H_A : \text{At least one of } \beta_1, \beta_2 \neq 0$$

3. **DECISION RULE**:
   Reject H_0 if the p-value $\leq \alpha = 0.05$.

4. **TEST STATISTIC**:
   The F-statistic for the model is:

   $$F = 39.37$$

5. **P-VALUE**:
   The p-value associated with the test is:

   $$p = 9.048 \times 10^{-11}$$

6. **CONCLUSION**:
   As $p \approx 0 \leq \alpha = 0.05$, we reject H_0. We conclude that there is sufficient evidence to support that at least one of the coefficients for the predictors β_1 (Hits) or β_2 (Potential) is non-zero and significantly contributes to explaining the variation in Overall performance.

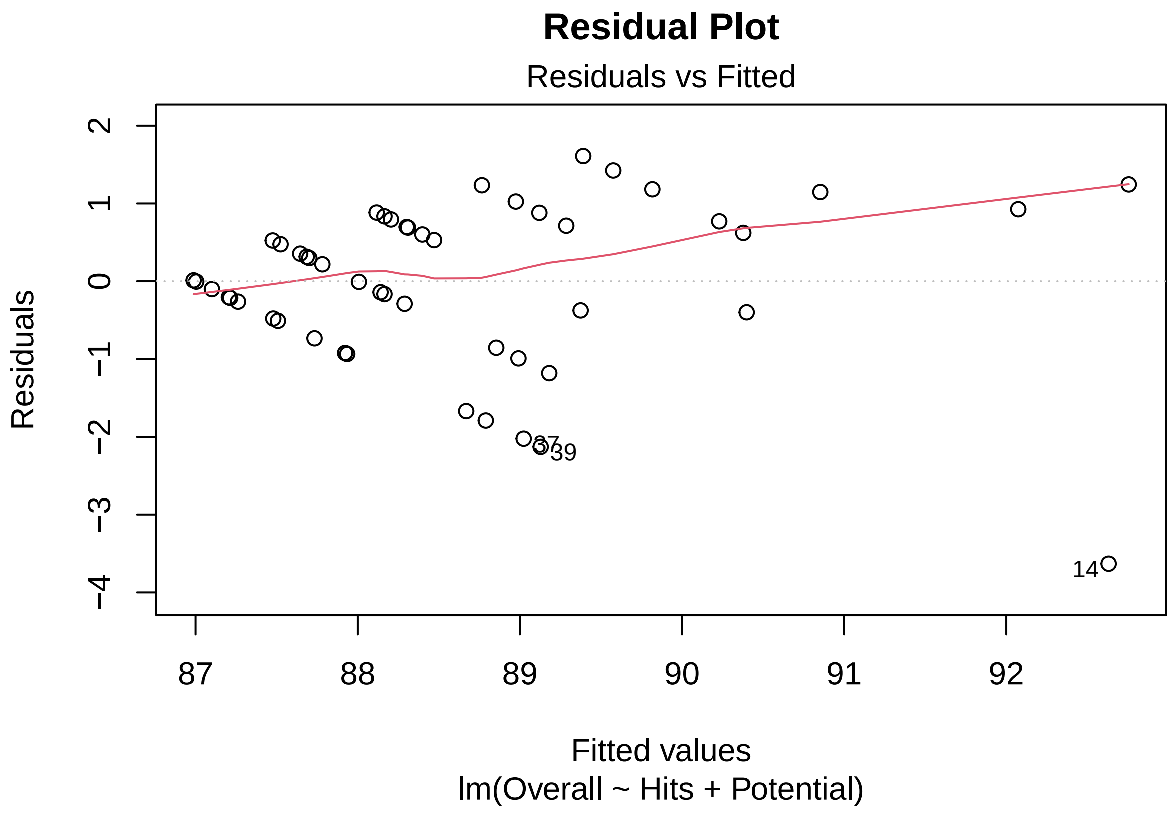
Residual and Normal Quantile Plots

To evaluate the validity of the linear regression model, residual and normal quantile plots were constructed. These diagnostic tools are essential for verifying the underlying assumptions of the model:

1. **Linearity**: The relationship between the predictors and the response variable is linear.
2. **Independence**: The residuals are independent of each other.
3. **Constant Variance**: The residuals exhibit constant variance.
4. **Normality**: The residuals follow a normal distribution.

Residual Plot A residual plot was created to check if the assumptions of the model are met. This plot displays the residuals on the vertical axis and the fitted values on the horizontal axis. If the residuals are scattered randomly, it suggests that the model's assumptions are satisfied.

```r
plot(lm_final, which = 1, main = "Residual Plot")
```



Interpretation of Residual Plot The residual plot shows the difference between the predicted values and the actual values (residuals) on the vertical axis, with the fitted values (predictions) on the horizontal axis. This plot helps us check if the model's assumptions—linearity, independence, and constant variance—are met.

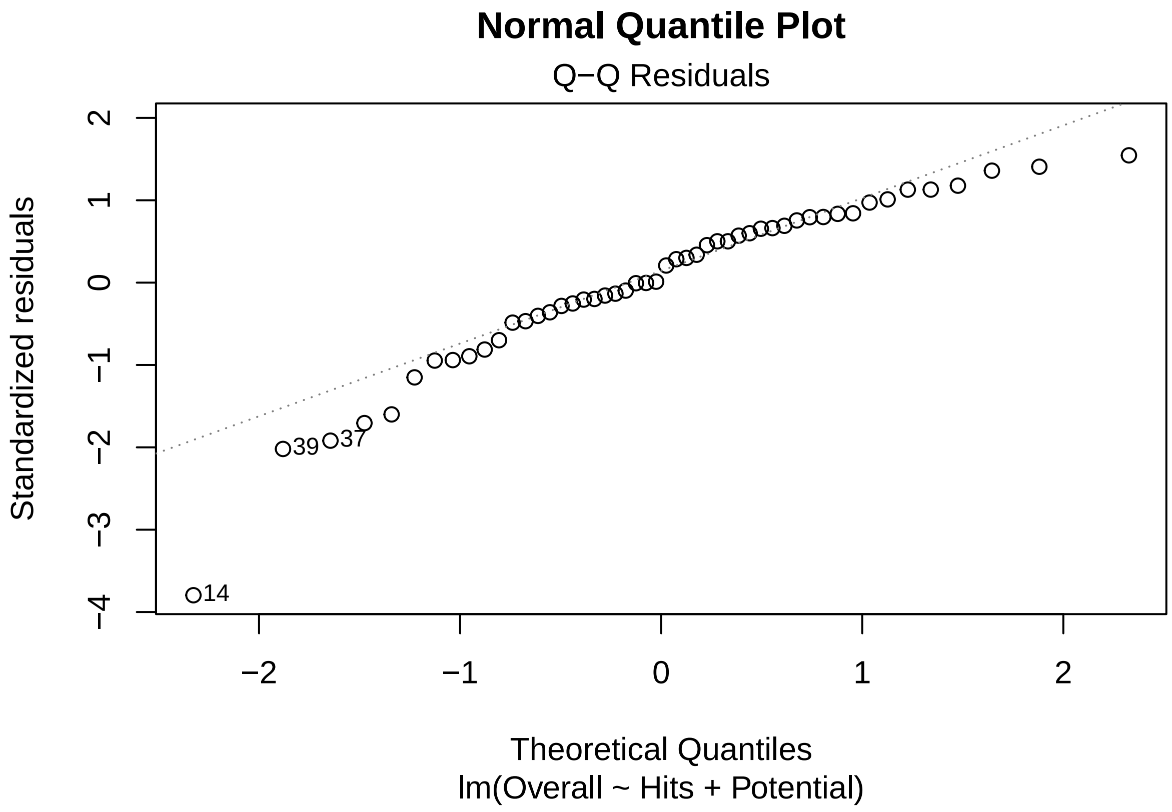
1. **Linearity**:
   The residuals are randomly scattered around the horizontal line (y = 0), with no clear patterns. This suggests that the relationship between the predictors (Hits and Potential) and the response variable (Overall performance) is linear, which is what we expect.
2. **Independence**:
   There are no obvious patterns or groupings of the residuals, which supports the idea that the residuals are independent of each other. This is another key assumption of the model.
3. **Constant Variance**:
   TThe spread of the residuals is fairly even across the range of fitted values. This shows that the variance of the residuals remains constant, which satisfies the assumption of homoscedasticity.
4. **Deviations**:
   A few outliers (e.g., points labeled as 37, 39, and 14) deviate slightly, but they do not appear to significantly violate the model assumptions.

Conclusion: The residual plot shows no strong violations of the model assumptions. The model seems to be valid in terms of linearity, independence, and constant variance, meaning the model is a good fit for the data.

Normal Quantile Plot A normal quantile plot (or Q-Q plot) was created to check if the residuals follow a normal distribution. In this plot, the expected values for a normal distribution are compared to the actual residuals from the model. If the residuals are normally distributed, the points should fall close to a straight diagonal line.

R command for the normal quantile plot:

```r
plot(lm_final, which = 2, main = "Normal Quantile Plot")
```



Interpretation of Normal Quantile Plot The normal quantile plot helps us see if the residuals follow a normal distribution by comparing the actual residuals to the expected values of a normal distribution.

1. **Normality**:
   Most of the points in the plot are close to the diagonal line, which means the residuals are roughly normally distributed.

2. **Deviations**:
   A few points at the ends of the plot (like points labeled 39 and 37) deviate slightly from the line. This suggests that there may be a mild non-normality in the extreme residuals, but these deviations are small and unlikely to have a major impact on the model.

Conclusion: The normal quantile plot shows that the residuals generally follow a normal distribution. The small deviations at the extremes are not enough to cause concern about the overall performance of the model.

Conclusion

This project aimed to explore the factors influencing the overall performance of FIFA players, specifically focusing on predicting a player's Overall performance rating based on two explanatory variables: Hits (player popularity) and Potential (projected future performance). Using a linear regression model, the relationship between these variables and overall performance was analyzed.

The final regression equation, which provides the best estimate of the relationship between the response variable Y (Overall) and the explanatory variables X_1 (Hits) and X_2 (Potential), is:

$$\hat{Y} = 43.723 + 0.008 \cdot X_1 + 0.496 \cdot X_2$$

This equation suggests the following: - For every one-unit increase in **Hits**, the predicted **Overall** increases by approximately 0.00918 units, assuming Potential remains constant. - For every one-unit increase in **Potential**, the predicted **Overall** increases by approximately 0.41052 units, assuming Hits remains constant.

The analysis showed that both Hits and Potential are statistically significant predictors of Overall performance, with p-values well below the significance level($\alpha = 0.05$). Additionally, the adjusted R^2 value of 0.6103 indicates that approximately 61.03% of the the variability in Overall performance is explained by the model.

Based on these findings, the explanatory variables—**Hits** and **Potential**—adequately predict the response variable **Overall**. The model demonstrates that both a player's popularity (Hits) and their future potential (Potential) are key factors in determining their current overall performance in FIFA.

This project provides useful insights for teams, managers, and analysts to focus on both a player's current engagement and their projected future growth as important factors in assessing player performance.