Final Project

Mohammed Saif Wasay (002815958)

Masters of Professional Studies in Analytics, Northeastern University

ALY 6010: Probability Theory and Introductory Statistics

Harpreet Sharma

June 26th, 2024

Introduction

In this project, we will look at the FIFA dataset, which contains a plethora of essential information about FIFA players that may be leveraged to acquire exciting insights about the players and their market value. The dataset contains several insights about FIFA players, which is fascinating for both gamers and football fans. The information covers 5682 players' country, weight, age, dribbling, attacking, defending, passing, shooting, and goalkeeping metrics.

First, we clean the data, which is necessary for any data analysis. To clean the data, we handle null values, delete redundant data, and ensure that all data types are as specified. Following that, we divide the dataset into subsets, one for goalkeepers and their statistics, and another for outfielders and their statistics.

We visually explored the data to determine the age and weight distribution of the players in order to better understand the data.

Figure 1: Bar Chart for Age distribution

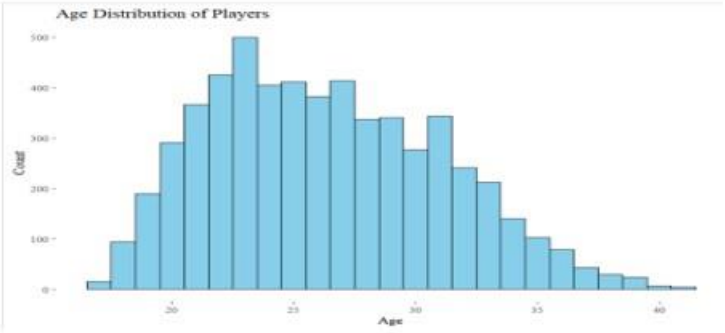
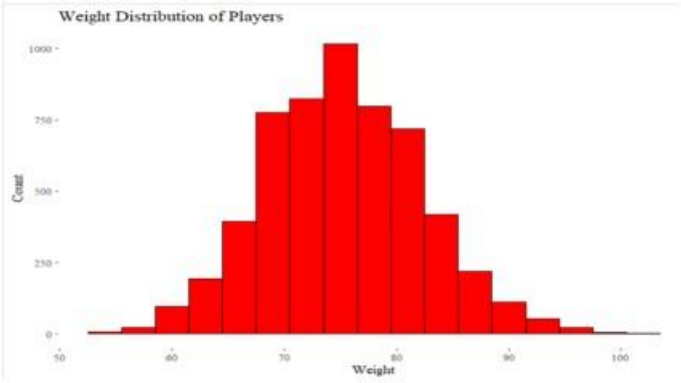


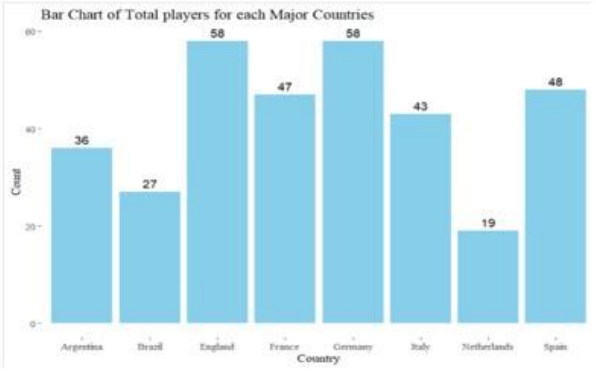
Figure 2: Bar Chart for Weight distribution



We may determine that the distribution of age and weight of the players follows a normal distribution or is close to normal from the graphs. Furthermore, we can see that the majority of the players are between the ages of 22 and 27, with the biggest number of players aged 23 (about 500).

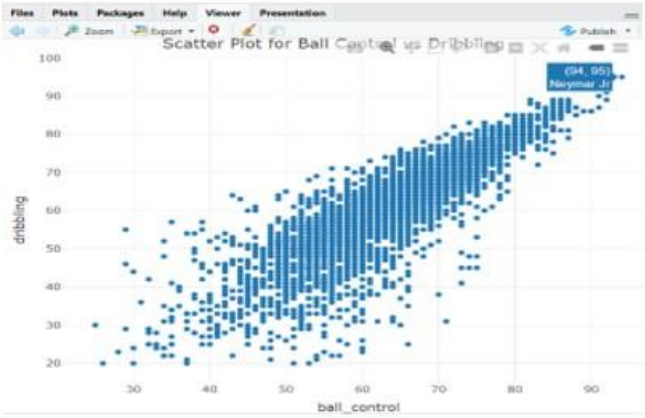
Following that, we produced a subset of all the main football nations and plotted their distribution in a bar chart. England and Germany each had 58 players, while France and Spain had 47 and 46 respectively.

Figure 3: Bar Chart of Players in Major Nations



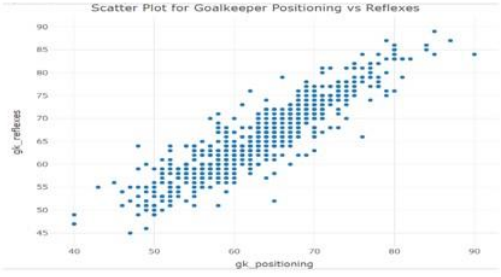
We wanted to see if players who dribbled well also had strong ball control, and we were able to validate this by visualizing these parameters in a scatter plot.

Figure 4: Scatter Plot for Ball Control vs Dribbling



We could also see that goalkeepers with strong positioning had good reflexes.

Figure 5: Scatter Plot for Positioning vs Reflexes



The data was then tested by a series of hypothesis tests. Is there a significant difference in the average weight of keepers and fielders. Is it true that outfielders are statistically shorter than goalkeepers? Is it statistically true that goalkeepers have longer careers?

Figure 6: Hypothesis testing for average weight

```
> #Hypothesis testing that average weight of outfielders is lesser than goalkeepers weight
> of_wt <- mean(outfield_data$weight)
> t.test(goalkeeper_data$weight, mu = of_wt, alternative = "greater", conf.level = .95)

        One Sample t-test

data:  goalkeeper_data$weight
t = 31.012, df = 631, p-value < 0.00000000000000022
alternative hypothesis: true mean is greater than 74.47524
95 percent confidence interval:
 81.35609      Inf
sample estimates:
mean of x
 81.74209
```

We used a One Sample t-test with a 95% confidence interval to see if the average weight of goalkeepers is greater than that of outfielders (null hypothesis: no significant difference). We strongly reject the null hypothesis with a t-value of 31.012, a p-value of 0.00000000000000022, and 631 degrees of freedom. The 95% confidence interval (infinity to 81.35609 kg) provides strong support for the alternative hypothesis, implying that the true mean weight of goalkeepers is far higher than 74.47524 kg.

Figure 7: Hypothesis testing for average height

```
> #Hypothesis testing that average height of goalkeepers is more than average outfielder height
> t.test(goalkeeper_data$height, mu = 180, alternative = "greater", conf.level = .95)

        One Sample t-test

data:  goalkeeper_data$height
t = 48.585, df = 631, p-value < 0.00000000000000022
alternative hypothesis: true mean is greater than 180
95 percent confidence interval:
 188.5298      Inf
sample estimates:
mean of x
 188.8291
```

We used a One Sample t-test at a 95% confidence interval to compare the average height of goalkeepers to a benchmark of 180 cm, with the null hypothesis of no statistical difference and the alternative hypothesis that average goalkeeper heights exceed outfielder heights. The resulting p-value, which is surprisingly small, strongly rejects the null hypothesis, indicating that such dramatic findings are unlikely if the true mean height is 180 cm. The 95% confidence interval (188.5298 to positive infinity) strengthens the conclusion that the genuine mean height exceeds 180 cm, providing strong statistical evidence to back up the claim. The extremely small p-value contribute to the strength of this finding, indicating a significant height difference for goalkeepers.

Figure 8: Hypothesis testing for Goalkeeper and Outfielder Age

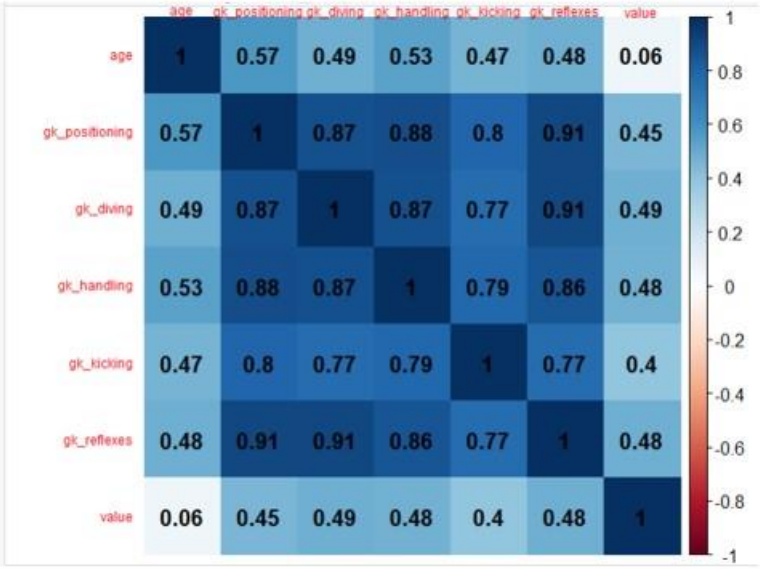
```
> #Hypothesis testing that average age of goalkeepers and outfield players
> t.test(outfield_data$age, goalkeeper_data$age, alternative = "less", var.equal = FALSE, conf.level = .95)

        Welch Two Sample t-test

data:  outfield_data$age and goalkeeper_data$age
t = -3.6102, df = 756.09, p-value = 0.0001631
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.4351041
sample estimates:
mean of x mean of y
 26.22682  27.02690
```

The null hypothesis states that there is no significant difference in ages when a Two Sample t-test at a 95% confidence interval is used to compare the average age of goalkeepers and outfielders. A relatively tiny p-value indicates a significant difference in mean ages, which enhances the evidence against the null hypothesis. The confidence interval (-Inf, -0.4351041) shows that the true difference is most likely less than -0.4351041. The statistical study clearly supports the conclusion that outfielders are younger than goalkeepers, with the modest p-value and 95% confidence interval highlighting this age difference.

Finally, based on our findings, we attempt to answer three questions: What relationship does goalkeeper positioning have with goalkeeper statistics? What outfielder statistics explain the variation in outfielder dribbling? How do a player's vision and long passing talents connect to their short passing ability? We will also construct a fit line graph and a regression model for each of the examples.

We use the numeric variables from the goalkeeper subset to create a correlation matrix for the variables in order to find a pairwise link between the goalkeeper metrics.

Figure 9: Correlation Matrix of Goalkeeper variables



According to the correlation heatmap, "gk_reflexes", "gk_diving", "gk_handling" and "gk_positioning" have a substantial positive correlation, showing a high degree of linkage between these variables. This shows that goalkeepers who are good at positioning are also good at diving, handling, and reflexes, and vice versa. The positive association is strong, meaning that as one of these characteristics improves, the other is likely to improve as well. In practice, this means that goalkeepers with higher positioning skills frequently display enhanced reaction abilities, which contribute to their total goalkeeping effectiveness.

Figure 10: Correlation Test for Goalkeeper Positioning and Reflexes

```
> #Correlation Matrix for the data
> correlation_matrix_gk <- cor(gk_num)
>
> #Correlation plot for the data
> corrplot(correlation_matrix_gk, method = "color", addCoef.col = "black", tl.cex = 0.60, tl.srt = 0)
>
> #Correlation test on the reflexes and positioning
> cor.test(gk_num$gk_positioning, gk_num$gk_reflexes, method = "pearson")

        Pearson's product-moment correlation

data:  gk_num$gk_positioning and gk_num$gk_reflexes
t = 54.4, df = 630, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8932760 0.9207918
sample estimates:
      cor
0.9080082
```

The null hypothesis holds that the true correlation between "gk_positioning" and "gk_reflexes" is zero, whereas the alternative hypothesis holds that the true correlation is greater than zero. With a p-value of 0.00000000000000022, there is substantial evidence to reject the null hypothesis. The 95% confidence interval (0.8932760 to 0.9207918) excludes zero, confirming the null hypothesis' rejection. As a result of the research, there is a strong and statistically significant positive correlation (0.908) between "gk_positioning" and "gk_reflexes" in the dataset.

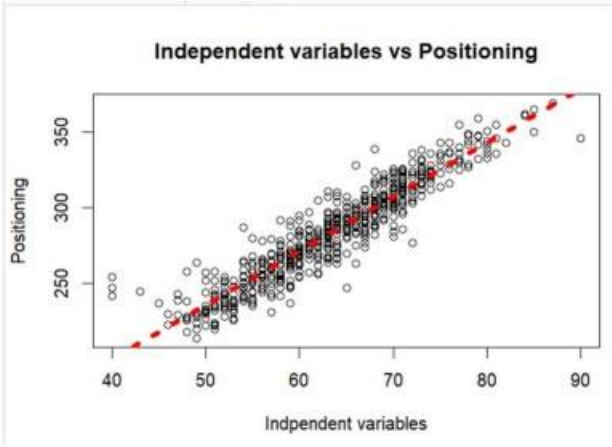
Figure 11: Fitting model to predict Goalkeeper Positioning

```
> summary(fit)

Call:
lm(formula = age + gk_diving + gk_handling + gk_kicking + gk_reflexes ~
    gk_positioning, data = gk_num)

Residuals:
    Min      1Q  Median      3Q     Max
-42.538  -7.187  -0.408   6.342  53.815

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)    57.22037    3.50203   16.34 <0.0000000000000002 ***
gk_positioning  3.57411    0.05485   65.16 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.34 on 630 degrees of freedom
Multiple R-squared:  0.8708,    Adjusted R-squared:  0.8706
F-statistic:  4246 on 1 and 630 DF,  p-value: < 0.00000000000000022
```

The linear regression model reveals a substantial and very significant link between age, gk_diving, gk_handling, gk_kicking, gk_reflexes, and "gk_positioning." Because "gk_positioning" is unlikely to be exactly zero, the intercept of 57.22037 is less easily interpretable. The coefficient for

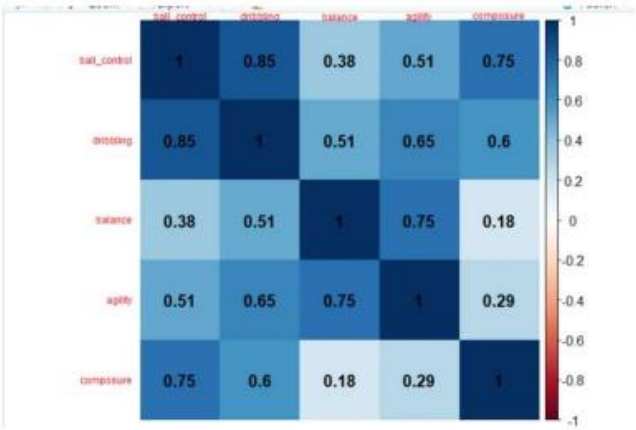
"gk_positioning" is 3.57411, suggesting that, on average, each one-unit increase in "gk_positioning" corresponds to a 3.57-unit increase in the response variables when all other factors are held constant. Both the intercept and the coefficient "gk_positioning" are very significant (p-value 0.0000000000000002). The R-square value is 0.8708, indicating that the model explains 87.08% of the variance in the response variables. The corrected R-square, which takes into consideration the number of predictors, is 0.8706. The F-statistic (4246) is very significant, confirming the model's overall relevance and predictive abilities. This implies a strong link, with "gk_positioning" serving as an important predictor in explaining the variance in the response variables.

Figure 12: Scatter Plot of Goalkeeper Positioning with regression line



A scatter plot for goalkeeper positioning with the hypothesized fit line based on a linear regression model with independent variables provides insight into their relationship. Based on the values of the independent variables, the fit line predicts the predicted value of the dependent variable ("goalkeeper_positioning").

Figure 13: Correlation Matrix for Player Control



The correlation matrix shows significant positive correlations between "ball_control" and "dribbling" (0.85), "ball_control" and "agility" (0.51), and "dribbling" and "agility" (0.65). These data show that

players with better ball control have better dribbling and agility skills. The positive connections imply that these traits coexist, underscoring the dataset's interrelated nature of ball control, dribbling, and agility.

Figure 14: Correlation test for player ball control, agility and dribbling

```
> #Correlation test on the reflexes and positioning
> cor.test(of_num$ball_control + of_num$agility, of_num$dribbling, method = "pearson")

        Pearson's product-moment correlation

data:  of_num$ball_control + of_num$agility and of_num$dribbling
t = 114.04, df = 5046, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8409053 0.8563370
sample estimates:
      cor
0.8488018
```

Pearson's product-moment correlation for "ball_control" and "agility" paired with "dribbling" is highly significant (p 0.00000000000000022). The null hypothesis is strongly rejected, implying that the true correlation is zero. The alternative hypothesis is supported, suggesting that the true correlation is not equal to zero. With a 95% confidence interval of (0.8409053, 0.8563370), the correlation coefficient is 0.8488. This indicates a strong positive association, indicating that players with greater combined "ball_control" and "agility" ratings tend to have superior "dribbling" skills.

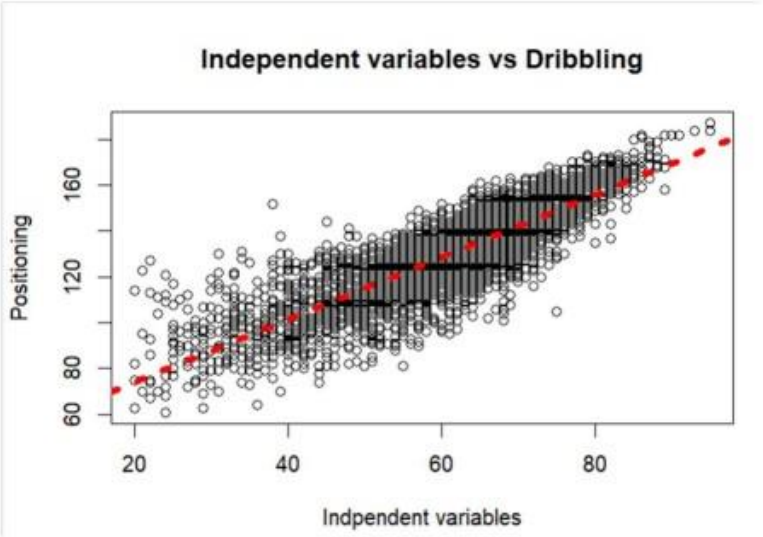
Figure 15: Regression model summary for dribbling

```
> summary(fit1)

Call:
lm(formula = ball_control + agility ~ dribbling, data = of_num)

Residuals:
    Min      1Q  Median      3Q     Max
-44.069  -5.633   0.546   6.239  53.213

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) 47.14689    0.74530   63.26 <0.0000000000000002 ***
dribbling    1.35896    0.01192  114.04 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.811 on 5046 degrees of freedom
Multiple R-squared:  0.7205,    Adjusted R-squared:  0.7204
F-statistic: 1.301e+04 on 1 and 5046 DF,  p-value: < 0.00000000000000022
```

The linear regression model, which fits "ball_control" and "agility" to "dribbling," shows a highly significant association. The intercept is 47.14689, which represents the predicted average value when "dribbling" is zero, albeit this may not be applicable in this case. The coefficient for "dribbling" is 1.35896, which means that for every one-unit increase in "dribbling," the combined scores in "ball_control" and "agility" should rise by around 1.36 units. The intercept and coefficient for "dribbling" are both statistically significant (p 0.0000000000000002). With an R-square of 0.7205, the

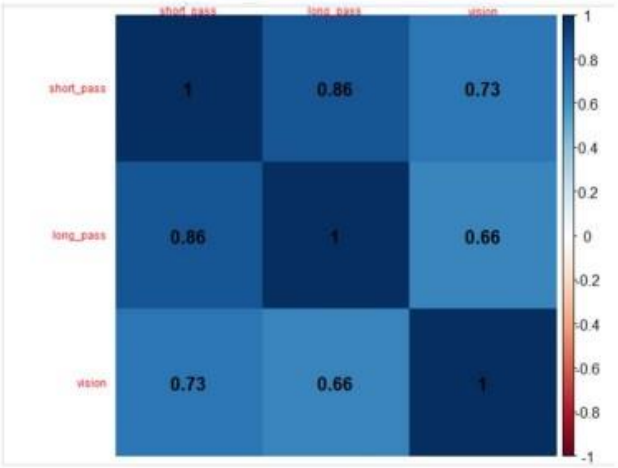
model explains 72.05% of the variance in the response variables. The F-statistic is substantial, indicating that the model is effective at predicting the total scores.

Figure 16: Scatter plot for Dribbling with model fit line



A scatter plot of player dribbling skills with the theoretical fit line based on a linear regression model involving independent variables (in this case, "ball_control" and "agility") reveals information about the relationship between these variables. Based on the values of the independent variables ("ball_control" and "agility"), the fit line predicts the expected value of the dependent variable ("dribbling").

Figure 17: Correlation Matrix for Short, Long Pass and Vision



The correlation matrix shows that there are high positive connections between the passing-related criteria. "short_pass" and "long_pass" have a strong positive correlation of 0.8596, "short_pass" and "vision" have a strong positive correlation of 0.7291, and "long_pass" and "vision" have a positive correlation of 0.6585. In the dataset, these findings imply a coherent relationship between short-

passing ability, long-passing ability, and vision. Players that score higher in one passing attribute tend to score higher in the others, highlighting a shared proficiency in various elements of passing skills.

Figure 18: Correlation test long, short pass and vision

```
> #Correlation test on the reflexes and positioning
> cor.test(of_pss$long_pass + of_pss$vision, of_pss$short_pass, method = "pearson")

        Pearson's product-moment correlation

data:  of_pss$long_pass + of_pss$vision and of_pss$short_pass
t = 124.27, df = 5046, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8612027 0.8747994
sample estimates:
      cor
0.8681639
```

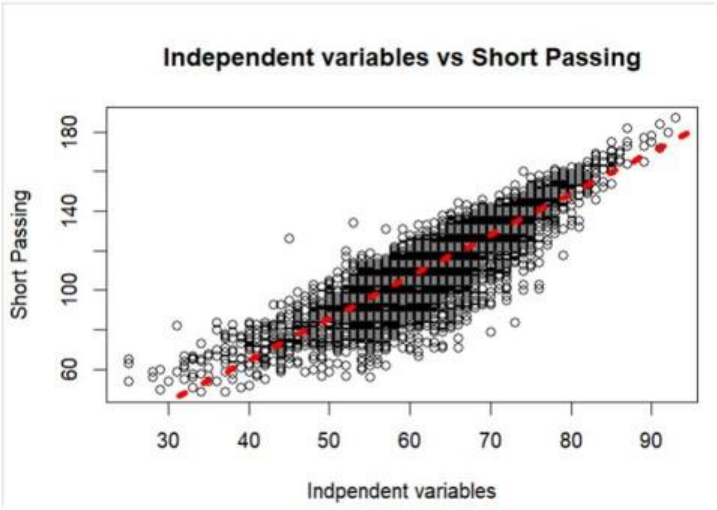
The Pearson's product-moment correlation between "long_pass" and "vision" combined and "short_pass" is highly significant ($p < 0.00000000000000022$). The null hypothesis, suggesting that the true correlation is zero, is strongly rejected. The alternative hypothesis, indicating that the true correlation is not equal to zero, is supported. The correlation coefficient is 0.8682, with a 95% confidence interval of (0.8612027, 0.8747994). This suggests a robust positive correlation, indicating that players with higher combined scores in "long_pass" and "vision" tend to demonstrate superior "short_pass" skills.

Figure 19: Regression model summary for short passing

```
> summary(fit2)

Call:
lm(formula = long_pass + vision ~ short_pass, data = of_pss)

Residuals:
    Min      1Q  Median      3Q     Max
-49.999  -5.713   2.025   7.525  50.669

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -18.95595    1.07752  -17.59 <0.0000000000000002 ***
short_pass    2.09527    0.01686  124.27 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.82 on 5046 degrees of freedom
Multiple R-squared:  0.7537,    Adjusted R-squared:  0.7537
F-statistic: 1.544e+04 on 1 and 5046 DF,  p-value: < 0.00000000000000022
```

The linear regression model, which fits "long_pass" and "vision" to "short_pass," shows a highly significant association. The intercept is -18.95595, which represents the predicted average value when "short_pass" is zero, albeit this may not be applicable in this case. The coefficient for "short_pass" is 2.09527, implying that for every one-unit rise in "short_pass," the combined scores in "long_pass" and "vision" should increase by around 2.10 units. The intercept and coefficient for "short_pass" are both statistically significant (p 0.0000000000000002). With an R-square of 0.7537, the model explains 75.37% of the variation in the response variables. The F-statistic is substantial, indicating that the model is effective at predicting the total scores.

Figure 20: Scatter plot of Short Passing with fit line



A scatter plot for player short passing skills with the theoretical fit line based on a linear regression model involving independent variables (in this case, "long_pass" and "vision") provides insights into the relationship between these variables. The fit line predicts the expected value of the dependent variable ("short_pass") based on the values of the independent variables ("long_pass" and "vision").

Conclusion

We were able to conduct thorough statistical and analytical analysis on FIFA player data. We were able to understand the data and answer a few questions based on our preliminary exploratory data analysis (EDA), such as the age of the majority of players in the dataset, which top nations had the greatest number of players, and so on. Following our EDA, we ran statistical hypothesis testing to see if there were any statistical differences in goalie and outfielder heights, weights, or ages. We may conclude from our hypothesis testing that goalkeepers are bigger and taller than outfielders, and their average age is much greater. Finally, we investigated the relationship between our player statistics and developed three regression models to predict goalkeeper placement, player dribbling, and short passing stats. We predicted with an accuracy of 87.08%, 72.05%, and 75.3%, respectively.

Citations

R Documentation, An introduction to R. Retrieved June 26th, 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation

Null hypothesis, Retrieved June 26th, 2024 from https://byjus.com/maths/null-hypothesis/

Linear Regression, Retrieved June 26th, 2024 from https://www.geeksforgeeks.org/ml-linear-regression