S&DS 230 Final Project: Goal Gurus

S&DS 230 Final Project

GROUP NAME: Goal Gurus

Introduction (Background and Motivation)

Soccer, a game celebrated for its dynamic blend of offensive and defensive strategies, is rich in statistical insights encompassing various facets such as offensive/defensive capabilities, set-piece proficiency, and individual player attributes. As the application of statistical analysis continues to permeate the sporting world, most notably in sports such as baseball, the goal of our project is to harness a data-driven approach to forecast the goal-scoring potential of our favorite players and uncover trends pivotal to the evolution of this amazing game.

We acquired a comprehensive dataset from Kaggle, comprising the per-90-minute statistics of professional soccer players playing in Europe's top five leagues: the Premier League, Ligue 1 Uber Eats, Bundesliga, Serie A, and La Liga. This dataset encompasses over 2500 entries, each representing a player, alongside 124 columns representing various game statistics.

Data and Data cleaning process

Here we read in the data

The raw dataset we read in has data on 2518 professional soccer players and 124 variables relating to their 2022-2023 season soccer playing statistics, ranging from minutes played to number of goals. This dataset can be found here: **LINK**

Variables that We Use and Brief Description

There are over 124 variables, but we only use 13. Here they are listed below, along with a brief description:

- Continent (categorical): Continent of the country a player represents for their national team
- Age (continuous): Age of the player
- Position (categorical): The primary postion a player plays in - forward, midfielder, and defender
- Goals (continuous): Average number of goals scored per 90 minutes
- Assists (continuous): Average number of assists (pass to another teammate that helped score a goal) provided per 90 minutes
- PassCompletionPct (continuous): Percent of average number of completed passes (not intercepted opponent) over total pass attempts made in a game

- League (categorical): The league a player participates in for their club team
- MinutesPer90 (continuous): Average number of minutes played per game
- ShotsOnTarget (continuous): Number of shot attempts made within the vicinity of the goal
- PenaltyKicksAttempted (continuous): Average number of penalty kicks attempted per game
- CornerKicks (continuous): Average number of corner kicks taken by a player per game
- YellowCards (continuous): Average number of yellow cards (warnings given for rule-violations) given to a player per 90 minute period
- CompletedPasses (continuous): Average number of completed passes (a pass that retains possession by another teammate) made by a player per game

Data Cleaning Process Description

1. There were a lot of corrupt letters in player names and team names in the raw dataset, so we first cleaned the corrupt characters and replaced them with the correct ones. This was a tedious process, but the corrupt characters were not random; mostly special and decorated characters were corrupted, so it was relatively simple to root out using gsub.

2. Next, we discovered that there were very detailed soccer positions assigned to each player (9 in total), but we just wanted to consolidate them into three the fundamental positions: Defender, Forward, and Midfielder, so we did that by recoding them.

3. Occasionally, player names were NA, so we removed those. Also, as players transferred around during the season, the same player would occupy multiple rows, which could disrupt the analysis. As a result, we removed duplicate rows for players.

4. To make the dataframe more intuitive to read, we made the player name the row name

5. We deabbreviated the variable names to make them more understandable.

6. We also created a new variable called continents. We did this by recoding the player countries into the respective continents, and subsequently inserted that into a new variable column.

7. We had to respect the official, sponsered name of Ligue 1, so we corrected it.

8. We only actually analyzed 13 of the 124 variables, so we subsetted them out to create a new, smaller dataframe, which would be easier to work with throughout the actual analysis.

9. We also noticed that some of the pass completion percentages seemed unreasonably high or unreasonably low, so we replaced the values for the PassCompletionPct's that were either 100 (too accurate to be a reasonable statistic) or 0 (too inaccurate to be a reasonable statistic) with NA.

Issue we encountered during data cleaning: *The corrupt characters were a bit frustrating to root out and involved a process of guess and checking to correct them into their true, non-corrupt values. Additionally, we had to clean the Yellow Card variable later in two seperate*

chunks because each of the two different analyses required slightly different cleaning methodologies, which we will explain later on.

Analysis

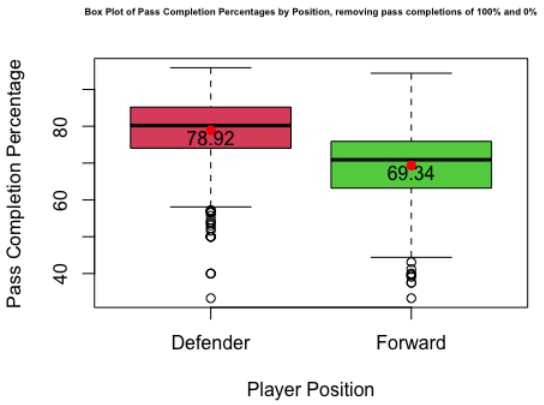
Now, we move onto the analysis portion of our project

T-Test

In the cleaned dataset, we have Pass Completion, a percentage as a variable. We're interested in seeing if there is a significant difference between the pass completion rate between forwards and defenders because intuitively, defenders should be better at passing because they're in charge of protecting the goal, and it would make more sense if they are more successful at passing (implying higher pass completion) on average than forwards.

Our null hypothesis is that the difference between mean pass completion percentages of forwards and defenders is zero. Our alternative hypothesis is that the difference between mean pass completion percentages of forwards and defenders is not zero.

Hence, we used Welch's two-sample t-test with a significance level of 0.05 to test the significance.



From boxplots, it appears that the IQR of the data between the two positions is fairly similar (implying similar variability), and that the data is not significantly skewed for either of the two positions, so we can proceed with the t-test.

```
##
##  Welch Two Sample t-test
##
## data:  PassCompletionPct by Position
## t = 19.496, df = 1171.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Defender
and group Forward is not equal to 0
## 95 percent confidence interval:
##   8.611719 10.538937
```

```
## sample estimates:
## mean in group Defender  mean in group Forward
##               78.91828               69.34295
```

Discussion of Results:

After running the t.test on pass-completion percentages between positions, we can reject the null hypothesis and can conclude that there is a statistically significant difference between the mean pass completion percent between defenders and forwards because the p value is less than 2.2e-16, which is under our significance level of 0.05. Thus, it demonstrates that the mean pass completion percent of defenders is statistically significantly greater than the mean pass completion percent of forwards.

Bootstrap Confidence Interval

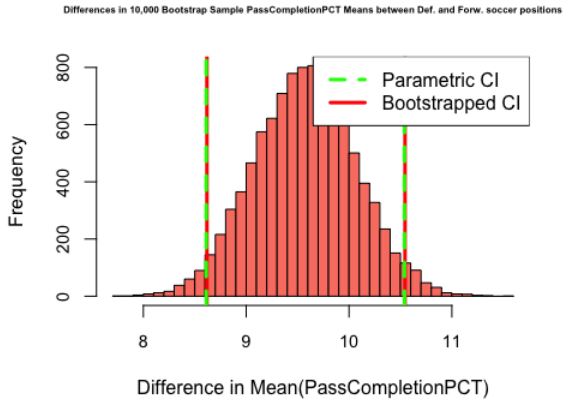
Now, we will create a non-parametric bootstrapped confidence interval for the hypothesis testing we just did, whether the difference between mean pass completion percentages of forwards and defenders is not 0.

```
##     2.5%    97.5%
##  8.61816 10.54451
```

The above is the 95% non-parametric confidence interval generated from the bootstrap.

```
## [1]  8.611719 10.538937
## attr(,"conf.level")
## [1] 0.95
```

The above is the 95% confidence interval generated from the t.test.



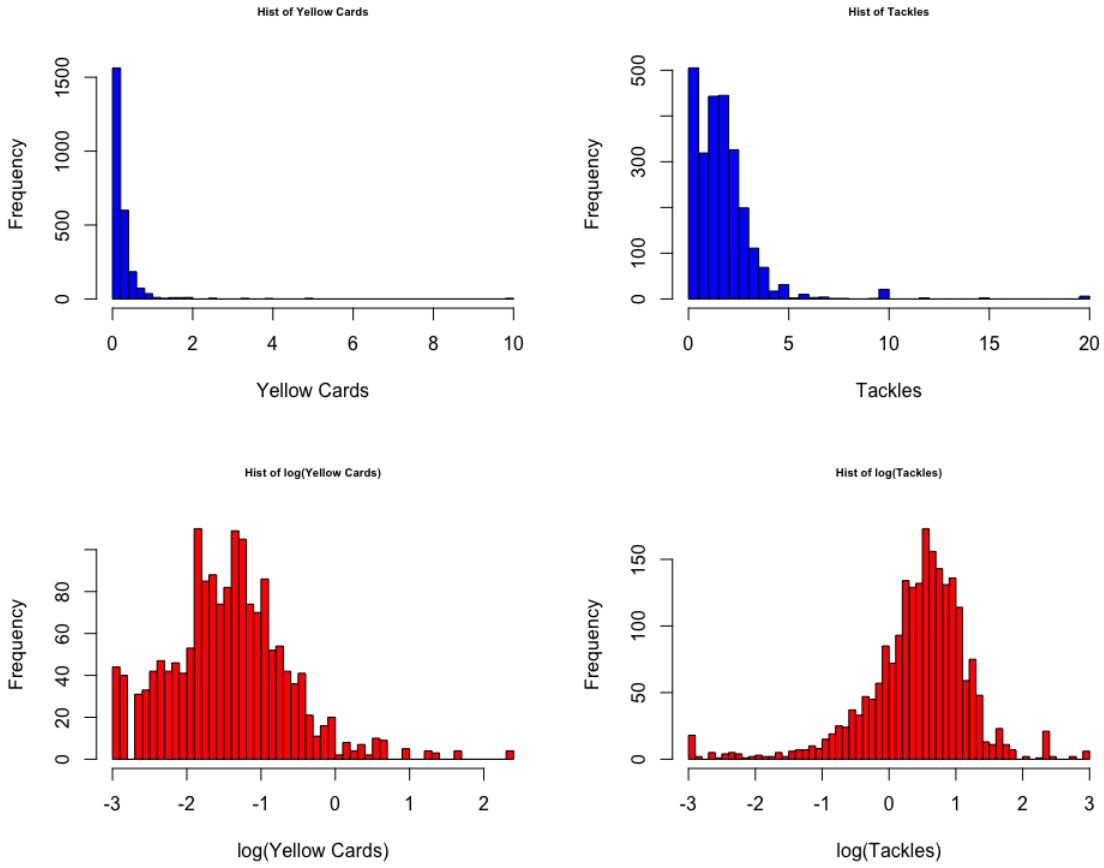
The above is the histogram where we ploted the bootrsapped mean pass completion percent differences between the positions, with the boostrapped and parametric CI superimposed as lines.

Bootstrap Discussion: *The lower bound of the Parametric CI is slightly lower than the lower bound of the Bootstrap CI, and the upper bound of the parametric CI is slightly lower than the upperbound of the bootstrapped CI, but overall, the two CI's appear fairly similar and both do not contain 0, so via both bootstrapped and parametric 95% confidence intervals, we can*

reject the null hypothesis and say that that there exists statistically significant difference between the mean(PassCompletionPct) of defenders vs the mean(PassCompletionPct) of forwards, as our 95% confidence intervals do not contain a difference of 0.

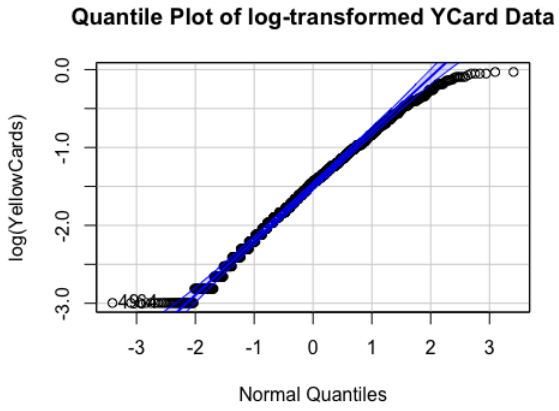
Correlation

For our correlation test, we wanted to explore the relationship between yellow cards received during a match and the number of tackles a player makes, under the hypothesis that a player who perform more tackles are associated with they receiving a greater number of yellow cards per game, because playing more rough should intuitively associate with receiving more yellow cards (warnings) from referees.



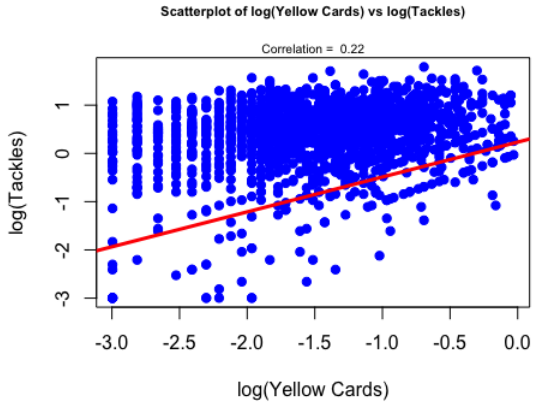
Based on the histogram of the average yellow cards and tackles per match, the data appears to be heavily left skewed and contains mostly 0s (blue graph). Therefore, we used a log transformation of both variables which helped make the distribution more normal and less skewed as shown above.

Furthermore, we removed all instances were yellow cards were zero or greater than 1 (which is unreasonable for soccer players, because it's difficult for a player to average more than 1 yellow card per game), and we then double checked the normality again with a normal quantile plot for the yellow cards below.




```
## [1] 49 64
```

From this norm-quantile plot, we can see that the it points are positive, linear and largely within the blue bounds, which suggests that the log-transformed yellow card data is normally distributed, which meets the assumptions of the correlation test. We also log transformed that tackles data to meet the assumptions of the correlation test, as seen in the previous histograms.



Discussion of Correlation Test:

Our correlation coefficient of 0.22 indicates there is a somewhat moderate positive relationship between the two variables and our slope of 0.2357503 suggests that a player performs about 0.2357503 more log(tackles) per game for each additional log(yellow card) they receive on average. The intercept of 0.7237923 suggests that a player who receives 0 yellow cards on average will perform 0.7237923 tackles per match. It should be noted we removed outliers of players with 0 tackles or 0 yellow cards as it suggests they do not play enough game time to actively count as a starter as well as any player who receive on average 1 or more yellow cards per game as it's unreasonable to commit that many fouls per game on average. The results of our test support our argument that a player that receives more tackles will tend to receive more yellow cards per game on average.

Permutation Test

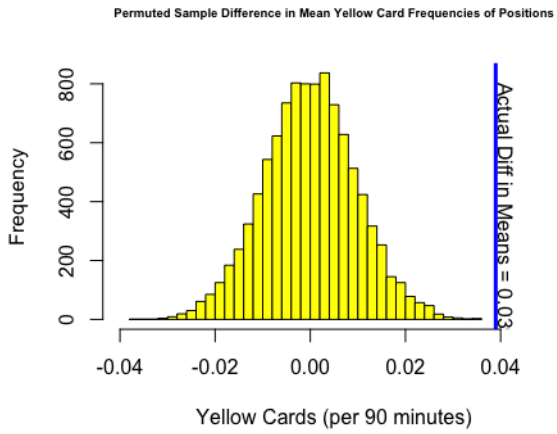
One question we're interested in exploring further is whether there is a significant difference in mean yellow card frequencies between forwards and defenders. Intuitively, defenders should have a have a higher yellow card frequency on average because they defend the goal and as a result have an incentive to play rougher to prevent the opposite team from scoring. Hence, the two samples will be defenders and forwards.

This is our null hypothesis, that the difference in mean Yellow Cards is 0 between positions.

$$H_0: \mu_{Defender} - \mu_{Forward} = 0$$

This is our alternative hypothesis:, that the difference in means is not 0 between positions.

$$H_a: \mu_{Defender} - \mu_{Forward} \neq 0$$



The above histogram shows the permuted sample difference in mean yellow cards between positions, with a line superimposed at the actual difference in mean yellow cards between positions.

Now that we have the permuted sample means, we went to calculate the extremeness of our alternative hypothesis, which we will test using one-sided p-values, following the alternative hypothesis below.

$$H_a: \mu_{Defender} - \mu_{Forward} > 0$$

Extremeness would mean obtaining a fake test statistic greater than or equal to our actual test statistic.

```
## [1] 0
```

PERMUTATION TEST RESULTS & P-VALUE DISCUSSION:

0 suggests that none of our 10000 simulations produced a difference that was more extreme than the actual difference in means. So our p-value is less than 0.0001, which is less than our significance level of 0.05. Because our permutation test's p-value is less than 0.0001, which is less than our significance level of 0.05, we reject our null hypothesis that there is no difference

between defender and forward means, and conclude that defenders have a statistically significantly greater mean yellow card frequency compared to forwards.

Multiple Regression

Description of Plan:

Our goal is to use multiple linear regression to predict the number of goals scored based on a set of predictors: + Assists to gauge whether play making is tied to goal scoring + Pass completion percentage to measure accuracy of passes and ball possession + Minutes per 90 to measure playtime + Shots on target to measure shot accuracy + Penalty kicks attempted to measure the number of penalty opportunities + Cornerkicks to measure effectiveness of set pieces + Yellow cards to see its relation to player behavior

Multiple linear regression makes the most sense for this task as it enables a player to determine statistically significant factors in a game that lead to higher goal scoring opportunities on average. To conduct our multiple linear regression, we made a dataset from soccer_data2 containing only these variables and used best subsets regression with BIC to determine the number and specific predictors to use in our model. Based on our normal quantile and residual plots, we'll make assessments for transformations and perform our final analysis of the results - contextualizing our findings and further analysis such as looking at potential collinearity.

```
##                        Goals Assists PassCompletionPct MinutesPer90
## Goals                   1.00    0.16             -0.09         0.36
## Assists                 0.16    1.00             -0.03         0.02
## PassCompletionPct      -0.09   -0.03              1.00         0.18
## MinutesPer90            0.36    0.02              0.18         1.00
## ShotsOnTarget           0.37    0.12             -0.14        -0.06
## PenaltyKicksAttempted   0.42    0.07             -0.08         0.07
## CornerKicks             0.10    0.13             -0.14         0.01
## YellowCards            -0.06   -0.04             -0.04        -0.13
##                        ShotsOnTarget PenaltyKicksAttempted CornerKicks
## Goals                           0.37                  0.42        0.10
## Assists                         0.12                  0.07        0.13
## PassCompletionPct              -0.14                 -0.08       -0.14
## MinutesPer90                   -0.06                  0.07        0.01
## ShotsOnTarget                   1.00                  0.14        0.09
## PenaltyKicksAttempted           0.14                  1.00        0.09
## CornerKicks                     0.09                  0.09        1.00
## YellowCards                    -0.02                 -0.02       -0.01
##                        YellowCards
## Goals                        -0.06
## Assists                      -0.04
## PassCompletionPct            -0.04
## MinutesPer90                 -0.13
## ShotsOnTarget                -0.02
## PenaltyKicksAttempted        -0.02
```

```
## CornerKicks                 -0.01
## YellowCards                  1.00

## [1] 5

## [1] "Assists"               "PassCompletionPct"     "MinutesPer90"
## [4] "ShotsOnTarget"         "PenaltyKicksAttempted"
```

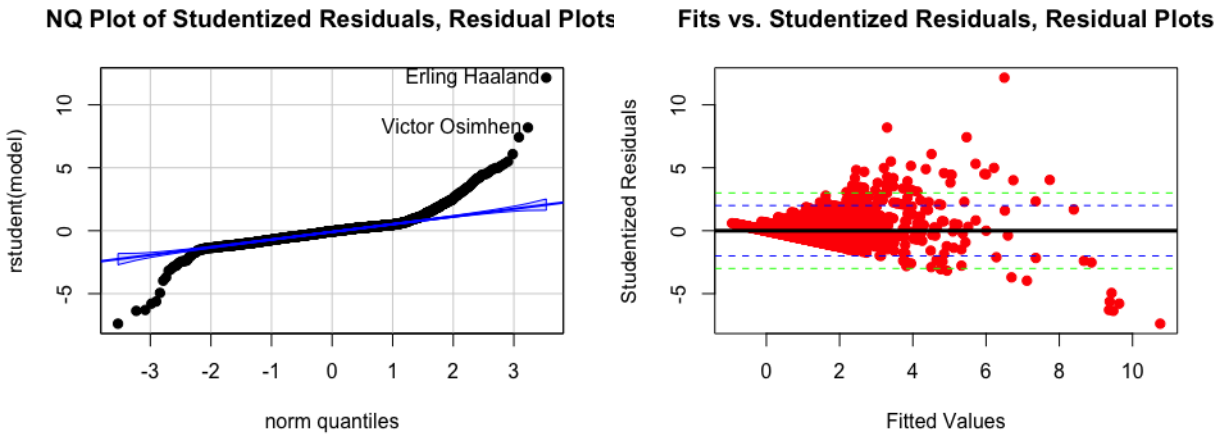
Based on our preliminary analysis, number of goals scored is positively correlated with all the predictors apart from pass completion percentage. Moreover, minutes per 90 (0.36), shots on target (0.37), and penalty kick attempts (0.42) appear to have a moderately strong relationship with goals scored. Using best subset regression, the optimal number of predictors to use for our model according to BIC (which we used to optimize the number of predictors given the large number of variables we started out with) is 5 and those predictors are: assists, pass completion percentage, minutes per 90, shots on target, and penalty kicks attempted.

```
##
## Call:
## lm(formula = Goals ~ Assists + PassCompletionPct + MinutesPer90 +
##     ShotsOnTarget + PenaltyKicksAttempted, data = soccer_regression)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7555  -0.7947  -0.0847   0.5042  18.5007
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.701164   0.250496   2.799  0.00516 **
## Assists                  0.885346   0.159157   5.563 2.95e-08 ***
## PassCompletionPct       -0.017260   0.003275  -5.270 1.48e-07 ***
## MinutesPer90             0.117893   0.005001  23.572  < 2e-16 ***
## ShotsOnTarget            1.014335   0.049188  20.622  < 2e-16 ***
## PenaltyKicksAttempted 12.611313   0.603125  20.910  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.574 on 2437 degrees of freedom
## Multiple R-squared:  0.416,  Adjusted R-squared:  0.4148
## F-statistic: 347.2 on 5 and 2437 DF,  p-value: < 2.2e-16
```

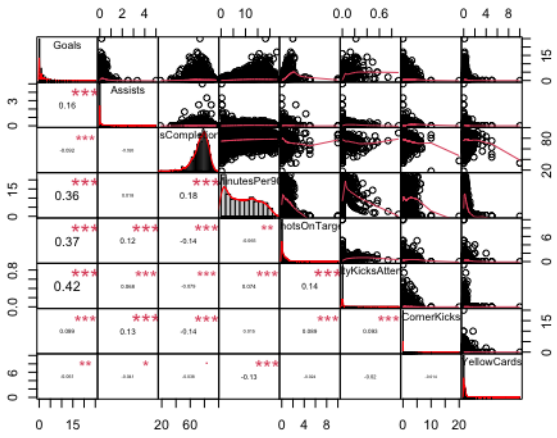
Based on the best fit model using BIC, the R-squared of 0.416 implies that 41.6% of the variability in goals scored can be explained by our model using the five predictors: assists, pass completion percentage, minutes per 90, shots on target, and penalty kicks attempted. All of our predictors have p-values less than our alpha of 0.05, which indicates that they are statistically significant predictors of goals. Moreover, based on our coefficients, for every additional assist goal scored increases by 0.885, for every one percent increase in pass completion percentage goals decreases by -0.0172, for every additional minute spent on average on the field goals increases by 0.1178, for every additional shot on target goals increases by 1.01 and for every penalty kick attempted goals scored increases by 12.6. It

should be noted that the relationship between goals scored and penalty kicks attempted is disproportionately larger due to the nature of most teams employing a designated penalty taker, which means a player who is given a penalty is likely to score many goals for this reason.

Now we check the normality and heteroskedasticity of residuals, along with any signs of colinearity between our predictors in the the current model.



The normal quantile plot suggests that our data is not normally distributed as both tails curve outside the guidelines and the relationship is not entirely linear. Moreover, the fits and studentized residual plots shows many extreme outliers (points above and below the green boundary) and strong evidence of heteroskedaticity as the variance of residuals with smaller fitted values is less than the residuals with higher fitted values.

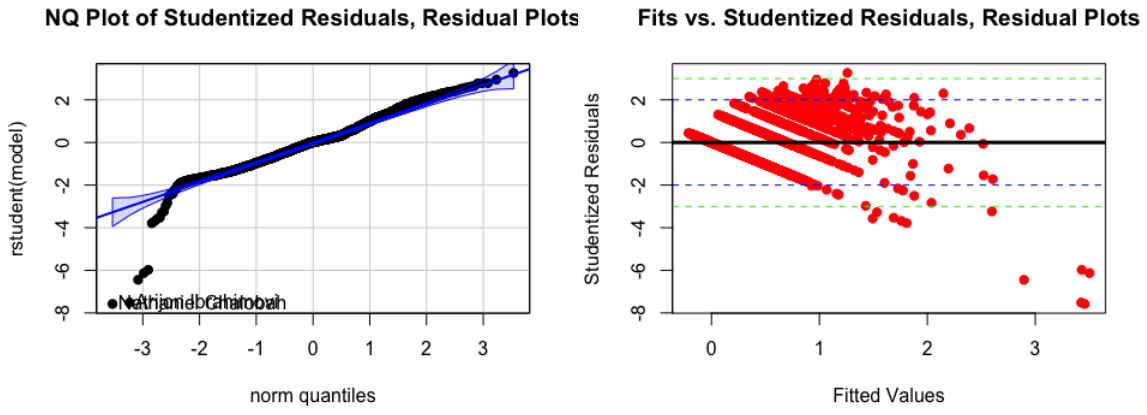


Based on the plot, it appears that the distribution for goals is heavily skewed to the right, which is likely due to the fact that a significant number of players scored 0 goals in the 2022/2023 season. Looking at the first row of plots, goals appears to have a linear relationship with all of the predictors. Between the predictors, there may be some concern for

collinearity as there appears to be a slight linear relationship between assists and some of the other predictors.

The below is basically the same multiple regression model with the same predictors as the one above, but with the one alteration that we log-transformed our response variable (goals) to make the residuals more normal to fit the assumptiosn of multiple regression.

```
##
## Call:
## lm(formula = log(Goals + 1) ~ Assists + PassCompletionPct + MinutesPer90 +
##     ShotsOnTarget + PenaltyKicksAttempted, data = soccer_regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4574 -0.3199  0.0071  0.2774  1.5743
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.243771   0.077071   3.163  0.00158 **
## Assists                  0.322522   0.048969   6.586 5.51e-11 ***
## PassCompletionPct       -0.004833   0.001008  -4.797 1.71e-06 ***
## MinutesPer90             0.041909   0.001539  27.235  < 2e-16 ***
## ShotsOnTarget            0.359614   0.015134  23.762  < 2e-16 ***
## PenaltyKicksAttempted    3.341135   0.185567  18.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4843 on 2437 degrees of freedom
## Multiple R-squared:  0.4448, Adjusted R-squared:  0.4437
## F-statistic: 390.5 on 5 and 2437 DF,  p-value: < 2.2e-16
```



Findings and Discussion of Multiple Regression Results:

In order to perform a boxcox, we had to add a factor of 0.1 to our response variable since it wouldn't accept minimum values of 0; in doing so, we received a lambda of -0.42, which when transforming the data based on the recommendation led to a normal quantile plot that was

still not normal. In contrast, after looking at the distribution of goals being heavily right skewed, we used a natural log transformation instead to test a better fit, which in turn led to a normal quantile plot that is more linear and mostly contained in the guidelines (apart from the left tail), suggesting the data is somewhat normally distributed. Moreover, the fits vs residuals plot still contains extreme outliers but exhibits less evidence of heteroskedaticity although variance is not constant across the residuals. Based on the transformed model our R-squared increased to 0.4448 which implies that 44.48% of the variability in log(goals) scored can be explained by our model using the five predictors: assists, pass completion percentage, minutes per 90, shots on target, and penalty kicks attempted. All of our predictors have p-values less than our alpha of 0.05, which indicates that they are statistically significant predictors of goals. Moreover, based on our coefficients:

- for every additional assist log(goal) scored increases by 0.322522
- for every one percent increase in pass completion percentage log(goals) decreases by -0.004833
- for every additional minute spent on average on the field log(goals) increases by 0.041909
- for every additional shot on target log(goals) increases by 0.359614
- for every penalty kick attempted log(goals) scored increases by 3.341135

Although the log transformation may make it slightly harder to interpret the data, the bottom line is that all 5 predictors of goals are significant, and (apart from pass completion percentage which is the opposite) tend to be associated with an increase in goals scored as each predictor variable increases.

Two Way ANOVA

Note: discussions of results are scattered throughout this section

For this project, we will be using two way ANOVA. Two way ANOVA is a statistical method used to compare the means of a response variable across two different independent variables (also known as factors) and to determine if there is an interaction between them.

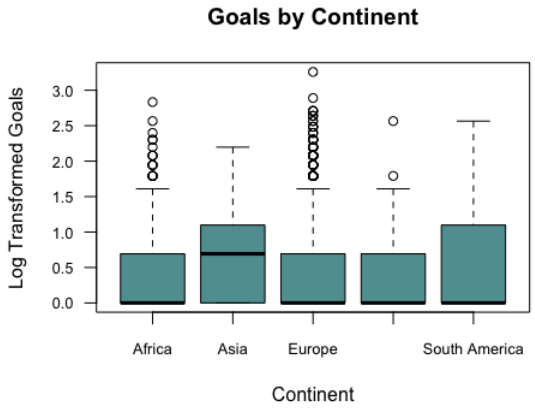
In this analysis, our continuous variable will be completed passes and we will be using League, Continent, and Position as our categorical variables. Before we begin, in order to do ANOVA, we want to ensure that all our combinations have at least one entry.

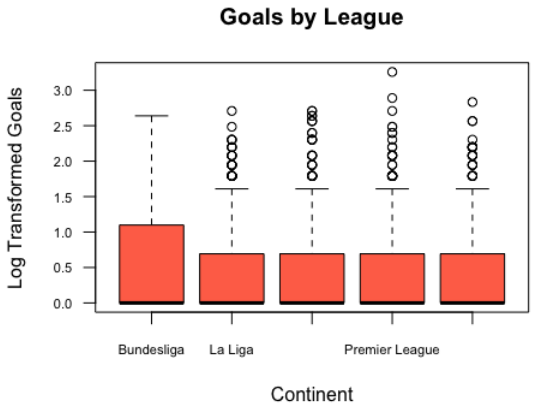
```
##
##                 Bundesliga La Liga Ligue 1 Uber Eats Premier League Serie
A
##                          0       0                 0              1
0
##   Africa                28      38               128             52
45
##   Asia                  19       6                 6              8
12
##   Europe               380     378               333            383
392
```

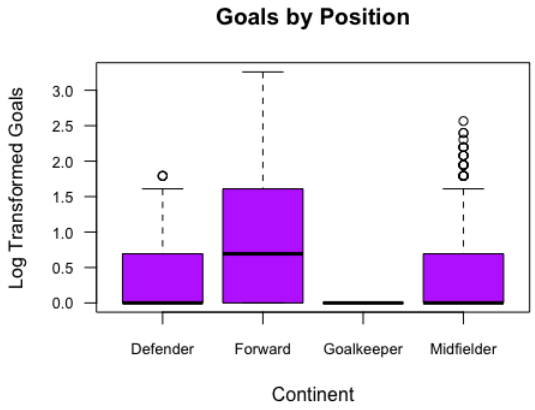
```
##   North America          9      10                10             16
6
##   Oceania                1       1                 1              3
2
##   South America         16      87                31             59
57

##
##              Bundesliga La Liga Ligue 1 Uber Eats Premier League Serie A
##   Defender          169     178                188            182     200
##   Forward           113     129                120            144     116
##   Goalkeeper         32      33                 34             29      31
##   Midfielder        139     180                167            167     167

##
##                  Defender Forward Goalkeeper Midfielder
##                         0       1          0          0
##   Africa               94      95          6         96
##   Asia                 15      16          2         18
##   Europe              687     426        137        616
##   North America        20      18          2         11
##   Oceania               2       3          0          3
##   South America        99      63         12         76
```

As we can see in our table of Continent by Position, there is a 0 between Goalkeeper and Oceania, which causes problems when fitting the model with interaction terms because it implies that there are certain combinations of factors that do not occur in the data set. Because of this issue, we have decided to remove 'Oceania' for our analysis as it has the least amount of data available. Additionally, lets remove the empty row from the 'Continents.'

Let's first create box plots to examine the relationship between goals and the categorical variables 'continent', 'league', and 'position'.

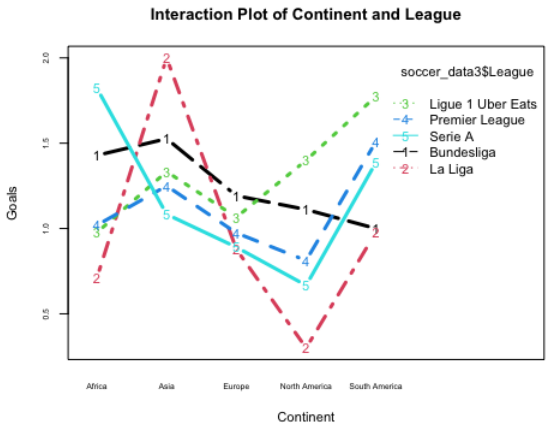


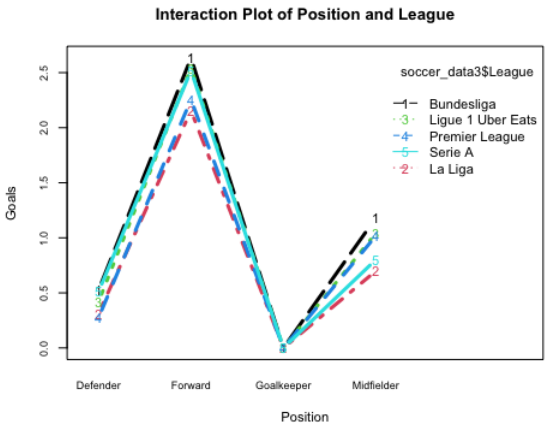


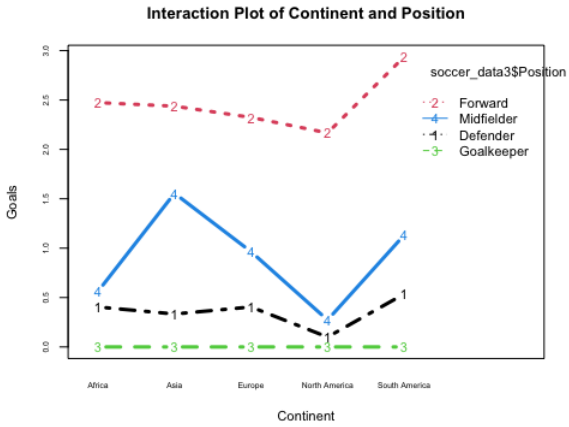


The box plot comparing Goals and Continent suggests a relationship, as the median value for goals scored by players from Asia is higher compared to other continents. This is indicated by the line within Asia's box being above zero, unlike the medians for other continents, which are at zero. Additionally, the medians for Africa, Europe, North America, and South America are approximately 0, demonstrating that scoring is rare. Most outliers also come from Africa and Europe, indicating that the highest scorers tend to come from these continents. The box plot between Goals and League is relatively similar, with the medians being very close to each other, suggesting that league might not be a good predictor for goals. Finally, there appears to be a relationship between Goals and Position Type, which is expected. Goalkeepers obviously tend not to score, while forwards tend to score the most.

Next we will create an interaction plot for each of the pairings for the categorical variables: Continent and League, Position and League, and Continent and Position. This is so that way we can see whether there are any significant relationships/interactions between the pairings.







For the interaction plot between continent and league, there does seem to be some interaction, as the lines are not moving in parallel with one another. In the interaction plot of position and league, the means between these positions appear to be roughly the same across leagues, indicating minimal interaction between position and league. For the interaction between continent and position, there also seems to be an interaction, as the lines are not moving in parallel. Based on these plots, we can assume that the interaction between position and league is the weakest, followed by continent and position, and then continent and league.

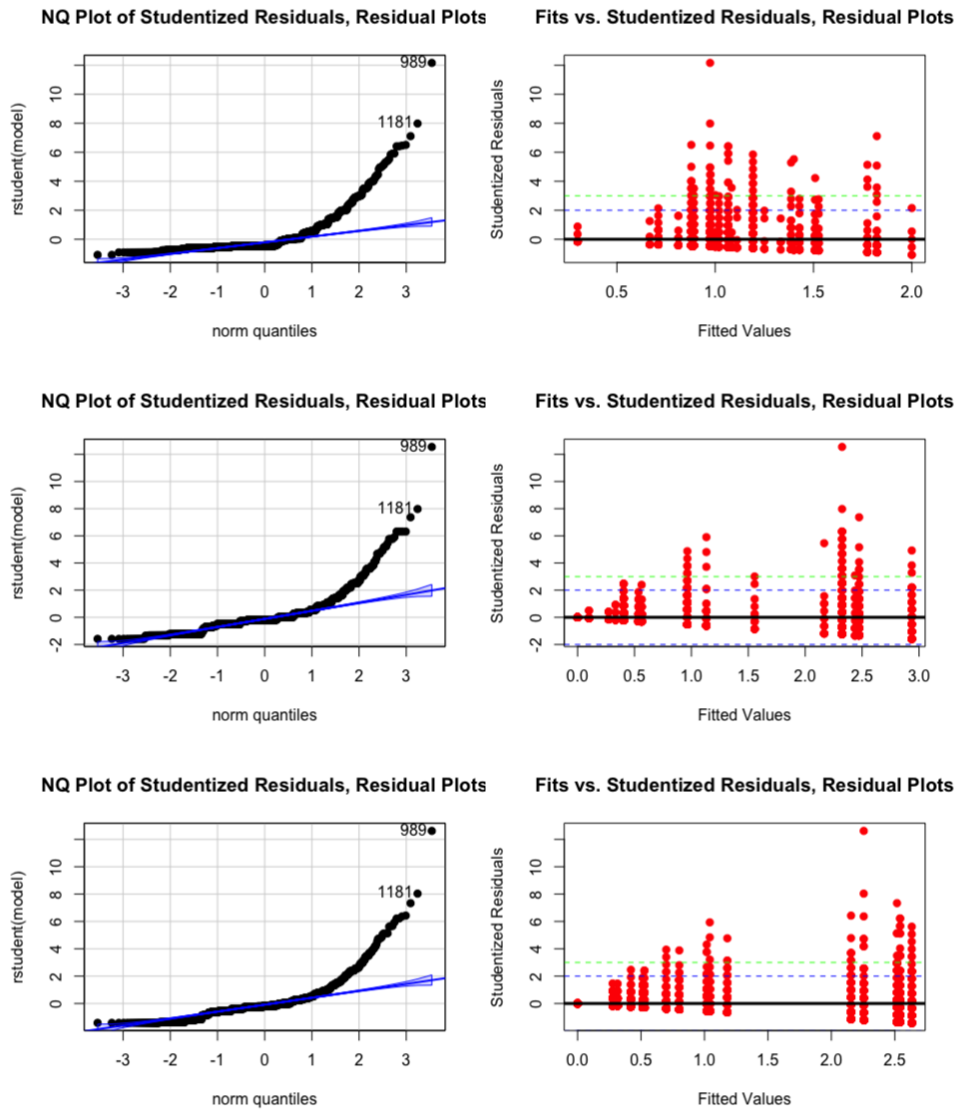
Now we will calculate three different two-way ANOVA with one for each pair of categorical variables ('Continent' and 'League', 'Position' and 'Continent', and 'Position' and 'League').

```
## Anova Table (Type III tests)
##
## Response: soccer_data3$Goals
##                                              Sum Sq   Df F value    Pr(>F)
## (Intercept)                                    57.1    1 13.7911 0.0002088
***
## soccer_data3$Continent                          4.2    4  0.2505 0.9094741
## soccer_data3$League                            34.4    4  2.0747 0.0815751
.
## soccer_data3$Continent:soccer_data3$League     58.7   16  0.8857 0.5859670
## Residuals                                   10292.4 2484
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type III tests)
##
## Response: soccer_data3$Goals
##                                              Sum Sq   Df F value   Pr(>F)
## (Intercept)                                    15.4    1  4.4145  0.03573
*
## soccer_data3$Continent                          3.6    4  0.2578  0.90503
## soccer_data3$Position                         259.6    3 24.8672 7.52e-16
***
## soccer_data3$Continent:soccer_data3$Position   27.7   12  0.6644  0.78703
## Residuals                                    8661.2 2489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type III tests)
##
## Response: soccer_data3$Goals
##                                           Sum Sq   Df F value    Pr(>F)
## (Intercept)                                 46.9    1 13.4695 0.0002476
***
## soccer_data3$Position                      353.0    3 33.8138 < 2.2e-16
***
## soccer_data3$League                          9.3    4  0.6659 0.6156441
## soccer_data3$Position:soccer_data3$League   19.6   12  0.4688 0.9335714
## Residuals                                 8661.0 2489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results are consistent with the interaction plot, in which the interaction between position and league was the lowest, followed by continent and position, and then continent and league. Upon analyzing all three tables, 'Continent' alone was not statistically significant, as the p-values of around 0.91 were all greater than 0.05, demonstrating that there are no significant differences in the number of goals across continents. Additionally, 'League' was not statistically significant, as the p-values of 0.08 and 0.61 were also greater than our critical value of 0.05. However, 'Position' was statistically significant (p-value of less than 0.05), which demonstrates that there are significant differences in the number of goals scored by players in different positions. When analyzing the interactions, though, all three p-values were much greater than 0.05. For the first table, we can conclude that the effect of the league on goals scored does not differ significantly across continents. For the second table, the effect of position on goals does not differ significantly across continents. Finally, for the third table, the effect of position on the number of goals does not differ significantly across leagues.

To see if the model assumptions are met, we will create a residual and normal quantile plot to see if there are any outliers and heteroskedascity.

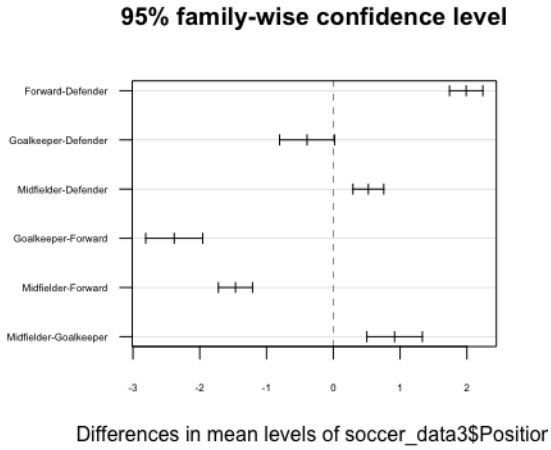


Model assumptions are not met as the normal quantile plot for all the interactions are not linear and there are outliers, heteroskedasticity, and the residuals do not seem normally distributed.

We will now perform a Tukey Honest Significant Difference test on factors of our ANOVA model. According to our two way ANOVA analysis, position is the only statistically significant factor in terms of goals scored, so we will use a Tukey HSD test to determine which levels of position differ from each other.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = soccer_data3$Goals ~ soccer_data3$Continent +
soccer_data3$Position + soccer_data3$Continent * soccer_data3$Position)
##
## $`soccer_data3$Position`
```

```
##                                diff        lwr         upr     p adj
## Forward-Defender          1.9893370  1.7396441  2.23902985 0.0000000
## Goalkeeper-Defender      -0.3926668 -0.8047007  0.01936701 0.0682981
## Midfielder-Defender       0.5246205  0.2937905  0.75545053 0.0000000
## Goalkeeper-Forward       -2.3820038 -2.8084435 -1.95556416 0.0000000
## Midfielder-Forward       -1.4647164 -1.7203751 -1.20905775 0.0000000
## Midfielder-Goalkeeper     0.9172874  0.5016112  1.33296359 0.0000001
```



All pairs (Forward-Defender, Midfielder-Defender, Goalkeeper-Forward, Midfielder-Forward, Midfielder-Goalkeeper) are statistically significant, with confidence intervals that do not include 0 and p-values below 0.05, except for the Goalkeeper-Defender pair. This means players that specialize in multiple roles tend to have a statistically significant difference in goal scoring capabilities.

ANCOVA

Note: discussions of results are scattered throughout this section

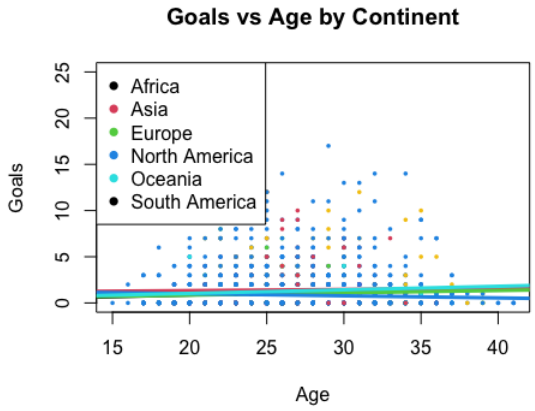
We will be using ANCOVA to predict goals based on age but fit separate lines for our categorical variable Continent. Our goal is to visually assess the predictive ability of age on goals and whether or not the continent has an impact.

```
## Anova Table (Type III tests)
##
## Response: soccer_data3$Goals
##                                          Sum Sq   Df F value Pr(>F)
## (Intercept)                                 0.0    1  0.0069 0.9340
## soccer_data3$Age                            6.7    1  1.6091 0.2047
## soccer_data3$Continent                      3.4    4  0.2046 0.9359
## soccer_data3$Age:soccer_data3$Continent     5.0    4  0.2990 0.8787
## Residuals                               10344.8 2499

##
## Call:
## lm(formula = soccer_data3$Goals ~ soccer_data3$Age *
```

```
soccer_data3$Continent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7690 -1.0494 -0.8740  0.0759 24.1010
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                        0.069559   0.839531
## soccer_data3$Age                                   0.040555   0.031971
## soccer_data3$ContinentAsia                         1.015314   2.200566
## soccer_data3$ContinentEurope                       0.278078   0.881287
## soccer_data3$ContinentNorth America                1.402791   1.711864
## soccer_data3$ContinentSouth America                0.193221   1.178611
## soccer_data3$Age:soccer_data3$ContinentAsia        -0.027944   0.084019
## soccer_data3$Age:soccer_data3$ContinentEurope      -0.015491   0.033550
## soccer_data3$Age:soccer_data3$ContinentNorth America -0.063739   0.064217
## soccer_data3$Age:soccer_data3$ContinentSouth America -0.001933   0.044217
##                                                   t value Pr(>|t|)
## (Intercept)                                         0.083    0.934
## soccer_data3$Age                                    1.269    0.205
## soccer_data3$ContinentAsia                          0.461    0.645
## soccer_data3$ContinentEurope                        0.316    0.752
## soccer_data3$ContinentNorth America                 0.819    0.413
## soccer_data3$ContinentSouth America                 0.164    0.870
## soccer_data3$Age:soccer_data3$ContinentAsia         -0.333    0.739
## soccer_data3$Age:soccer_data3$ContinentEurope       -0.462    0.644
## soccer_data3$Age:soccer_data3$ContinentNorth America  -0.993    0.321
## soccer_data3$Age:soccer_data3$ContinentSouth America  -0.044    0.965
##
## Residual standard error: 2.035 on 2499 degrees of freedom
## Multiple R-squared:  0.006671,   Adjusted R-squared:  0.003094
## F-statistic: 1.865 on 9 and 2499 DF,  p-value: 0.05278
```

Regarding the primary effects, the influence of age on goals scored was not statistically significant, with a p-value of 0.2047. This suggests that age does not markedly affect the number of goals scored. Similarly, the impact of continent alone was also not statistically significant, indicating that it does not substantially influence scoring. Additionally, the interaction effect between age and continent was not significant either, with a p-value of 0.8787. This indicates that the influence of age on goals scored is consistent across different continents. The model proved to be ineffective in explaining the variance in goals scored, as evidenced by the very low R-squared value of 0.006671, which indicates that the model accounts for only a minimal fraction of the variance in goals scored. Lastly, the wide range of residuals suggests the presence of potential outliers, which could be adversely affecting the performance of our model.

Now we are going to make a plot that shows 'Goals' predicted by age with separate colors for each continent. Then we will superimpose predicted regression lines for each continent.



From this plot, there is no obvious trend that indicates a strong relationship between age and goals scored. Additionally, there is no clear indication that players from one continent score significantly more or fewer goals than players from another.

Conclusion & Summary

Throughout the course of our analysis, we gained insights into various player characteristics and statistics that contributed to their ability to score goals and also impacted their performance in other respects. **First**, we found that the mean pass completion percent of defenders is statistically significantly greater than the mean pass completion percent of forwards using a t-test, suggesting that defenders are better at passing than forwards, subsequently generating a bootstrapped CI to corroborate that claim. **Second**, with a correlation test, we also found that there is a moderate positive correlation between the amount of yellow cards received by a player and the amount of tackles that they commited, suggesting that players who play rough also get more yellow cards. **Third**, we discovered that assists, minutes per 90, shots on target, and penalty kicks attempted are statistically significant positive predictors of the number of goals a player scores while pass completion percentage is a statistically significant negative predictor of number of goals scored, adding to our understanding of what characteristics of soccer players explain goal-scoring success. **Fourth**, with a permutation test, we concluded that defenders have a statistically significantly greater mean yellow card frequency compared to forwards, suggesting that defenders play rougher than forwards on average. **Fifth**, with ANOVA, we concluded that the interactions between our three categorical variables 'Continent', 'League,' and 'Position' for 'Goals' were not statistically significant (the effect of league on goals scored did not significantly differ across continent; the effect of position on goals did not significantly differ across continent; the effect of position on the number of goals does not significantly differ across league). **Lastly**, with ANCOVA, we found that the interaction between 'Age' and 'Continent' was not a statistically significant predictor for 'Goals', demonstrating that the influence of age on goals scored is consistent across different continents.