**Prediction of Soccer Matches**

**using**

**Data Mining Techniques**

By,

Ruthwik Nadam

Johnathon Kulich

Dang Huu Thien Nguyen (Thomas)

**Table of Contents**

[1. PROBLEM STATEMENT 2](#_Toc1025688232)

[2. INTRODUCTION 3](#_Toc632601447)

[2.1 DATA SOURCES 3](#_Toc569557533)

[2.2 DATA CLEANING AND TRANSFORMATION 4](#_Toc358140853)

[2.3 DESCRIPTIVE STATISTICS 4](#_Toc1533512952)

[3. MODELING 11](#_Toc1298371215)

[3.1 VARIABLE SELECTION 11](#_Toc848399338)

[3.2 DECISION TREE 13](#_Toc198835836)

[3.3 NAÏVE BAYES 17](#_Toc1977490057)

[3.4 CLUSTERING 18](#_Toc117979838)

[4. RESULTS 18](#_Toc1404408617)

[4.1 DECISION TREE RESULTS 19](#_Toc90066924)

[4.2 NAÏVE BAYES RESULTS 19](#_Toc653443711)

[4.3 MODEL COMPARISION 20](#_Toc1404586796)

[4.4 CLUSTERING RESULTS 20](#_Toc842882909)

[5. CONCLUSION 20](#_Toc1788184532)

[6. REFERENCES 21](#_Toc1691794454)

[APPENDIX 22](#_Toc1068383565)

# **1. PROBLEM STATEMENT**

The problem to be addressed through this project is to use RStudio to work on soccer data in order to make predictions on various areas of the game. We have different predictions that we want to make in this project. These include, but are not limit to, some basic analysis about soccer, predicting the results from betting odds, determining the significant factors for a position on the field, and team clustering. This problem is important to address because this research could help teams get an understanding of what makes players and teams the best.

The aim of this project is to identify the variables that are important determinants of predicting the best soccer players and teams. This project sets out to create a way to evaluate the sport of soccer using data mining techniques. This also will help players and teams know what the most important factors are in becoming the best or being the worst at the game. Soccer is a fast and growing sport and using this knowledge may give players an understanding of what to do to move forward.

# **2. INTRODUCTION**

## **2.1 DATA SOURCES**

The data sources that we are going to use come from the European Soccer Database on Kaggle.com. It consists of seven tables, each with different attributes. The data is from the 2008 season to the 2016 season. It includes details about team and player information as well as events that happened in the match.

1. +25,000 matches
2. +10,000 players
3. 11 European Countries with their lead championship
4. Seasons 2008 to 2016
5. Players’ and Teams' attributes, sourced from EA Sports' FIFA video game series, including the weekly updates.
6. Team line-up with squad formation (X, Y coordinates)
7. Betting odds from up to 10 providers
8. Detailed match events (goal types, possession, corner, cross, fouls, cards, etc…) for +10,000 matches

The seven tables are:

1. Player: contains 11,060 observations of 7 variables including the id (just the order), some basic information about the players (height, weight, birthday, name), player\_api\_id and player\_fifa\_api\_id are foreign keys for player and match
2. Match: contains 25,979 observations of 115 variables including the league, country, date, season, team, stage, players who played that match, betting odds, and match events
3. Country: contains 11 observations of 2 variables, which is the country name and the country\_id
4. League: contains 11 observations of 3 variables including the league id, country id and the name of the league
5. Team: contains 299 observations of 5 variables including the ids, the team's long name and the team's short name
6. Player Attributes: contains 183,978 observations of 42 variables that describe the rate of the players. Data of this table is based on the EA Sports FIFA game.
7. Team Attributes: contains 1458 observations of 25 variables that describe the rate of the team. Similarly, the data is based on the EA Sports game.

## **2.2 DATA CLEANING AND TRANSFORMATION**

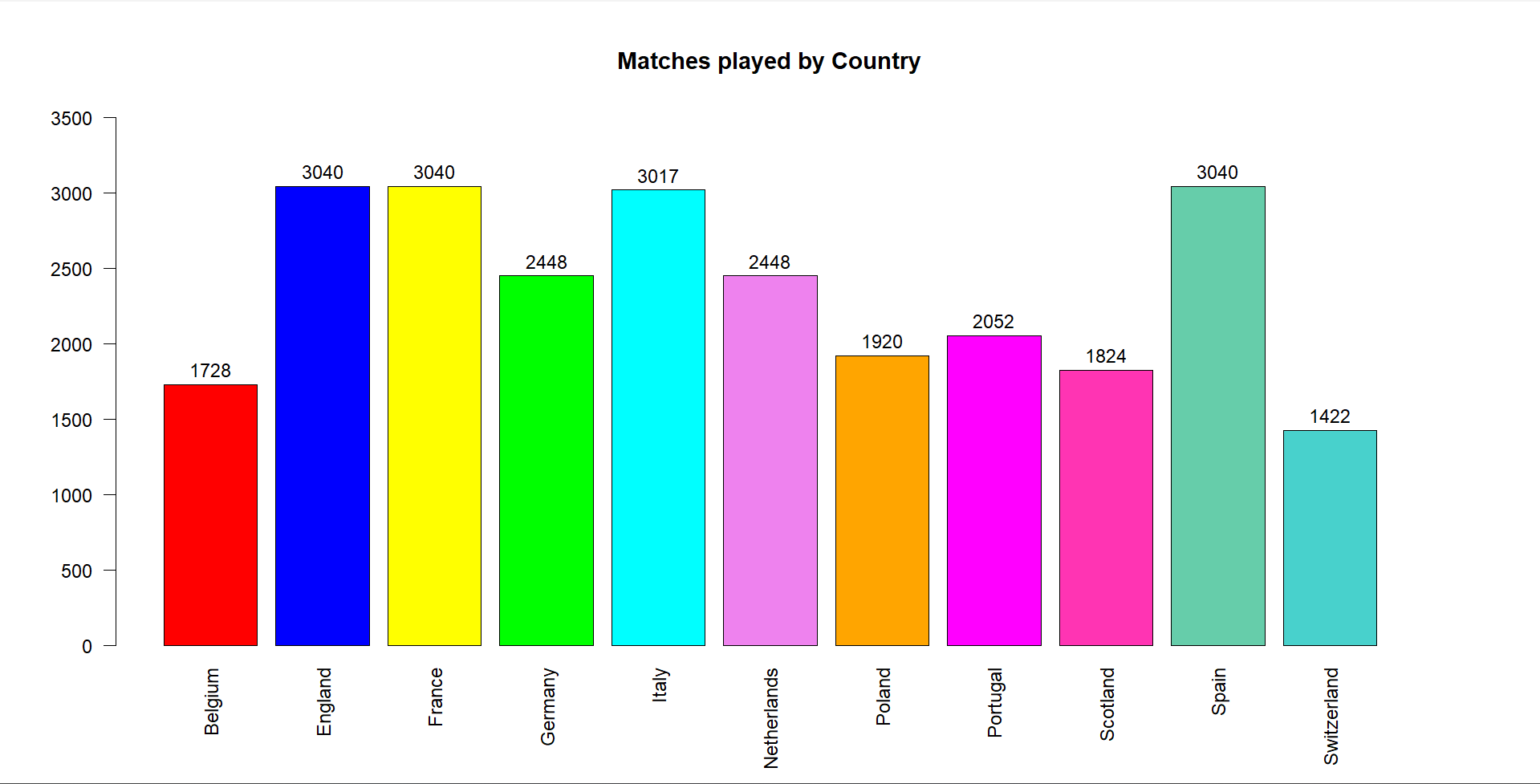
The main tables we want to use are the Match, Player\_Attributes, and Team\_Attributes. The other tables are mainly foreign keys for the three tables above (for example, in Match, it only contains the league\_id, but to know the name of the league, we need to compare it in the League table). Therefore, the values of the data from the other tables were merged into the tables above with their respective keys/identifiers. The method left\_join() from the “dplyr” package was used to merge the values to the Match dataset. A column match\_result was added to the dataset to show the result of the match, and match\_result\_who was added to show if the winner is the home team or the away team, or a draw, which will be used for some basic analysis later.

The unused columns (those that were added or duplicated due to the joining, or those that don’t have an effect in building models) are removed from the dataset and all the Null (NA) values from the dataset were omitted. There are still many variables which require more understanding on how to use them in the model and they will bring changes in the prediction.

For a good prediction model, the data must be shuffled (random). If not, when we split it into the Train set and the Test set, the Train set may only contain very few data from the last portion of the dataset (or even none of them at all), so it will not see the pattern/structures of those data, hence, it’ll only build the model based on the first portion of the dataset, and when we apply it to the Test set to predict the outcome, the accuracy will be significantly low.

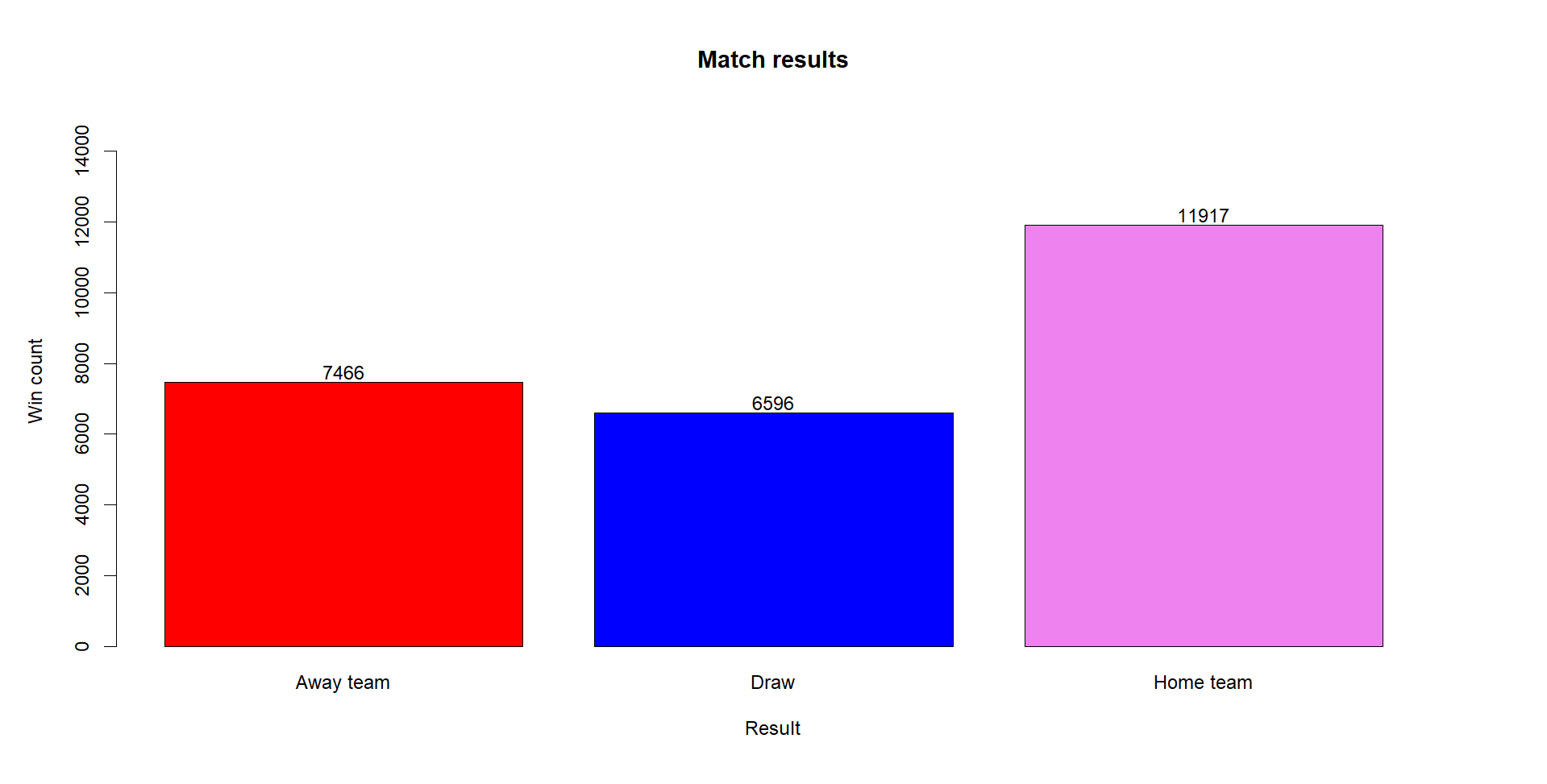
## **2.3 DESCRIPTIVE STATISTICS**

For a brief overview of the data, some plots and basic analysis have been provided. For this data, although League and Country are 2 different variables, in terms of data visualization and analysis, they will be the same. The reason for this is that this data only recorded information from the highest league of the 11 countries; in other words, each country from the data only has ONE league represented (for example, only England Premier League is presented in the data while England also has English Football Championship, English Football League One, and English football League Two,...and so on, with a total of 8 different leagues). The plot below shows the number of matches recorded for each country in our dataset.



*Figure 2.3.1. The total numbers of matches recorded for each country*

The summary() function has shown that the average number of goals scored by the home team is 1.563 and by the away team is 1.192, and the maximum number of goals scored by home and away teams stands at 9 and 6, respectively. There is a “rumor” saying that the home teams always have an advantage over the away teams and are more likely to win a match, so we are going to produce some visualization and statistical tests to check if it’s true. The plot below shows the results of the matches, and from it, we can see that the number of matches that the home teams won is about 1.6 times higher than those that the away team won, so the “rumor” may be true.

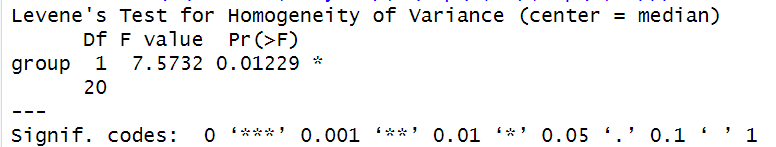


*Figure 2.3.2. The overall matches’ results*

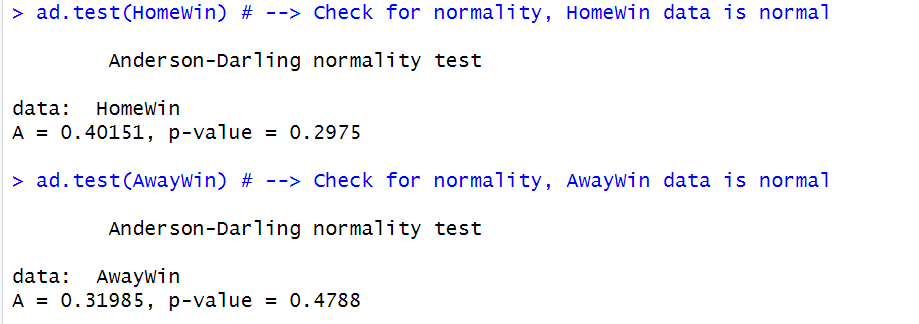
Plots are good for visualization to have a “brief” feeling of the data, but it doesn’t exactly tell if the home team wins more significantly. A paired t-test was used to determine that, along with checking for normality (Anderson-Darling test) and equal variances (Levene’s test) before performing the t-test. These checks will tell us if we need to transform the data and use a pooled variance for the t-test. Both data (home teams win and away teams win) are normal, and the population variances are not equal. The result provided a p-value of 3.506\*10-6, which indicates that home teams significantly win more than away teams, so the “rumor” is true.

**Ho : The mean number of matches that home teams win and away teams win are the same**

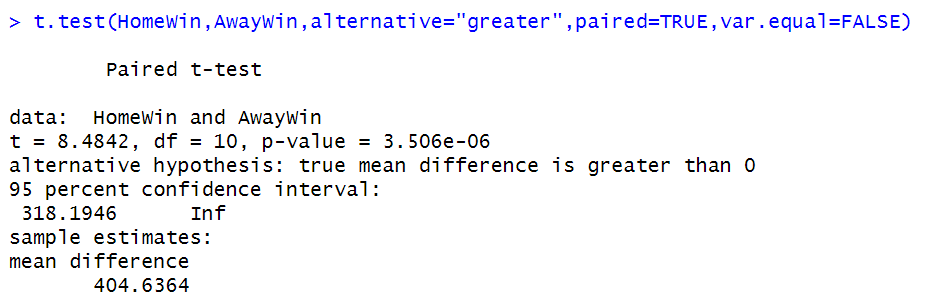
**Ha : The mean number of matches that home teams win is more than away teams win**



*Figure 2.3.3. Levene’s test for equal variances*

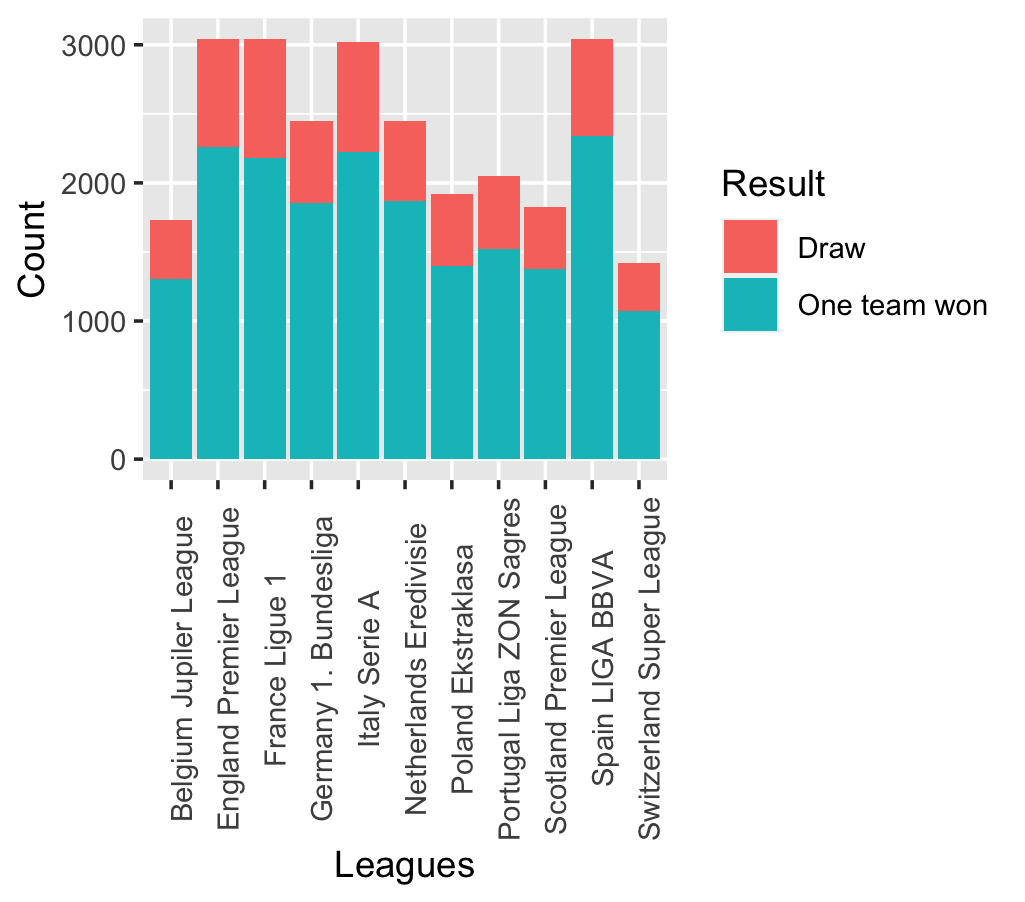


*Figure 2.3.4. Anderson-Darling normality tests*



*Figure 2.3.5. Result of the paired t-test*

Another assumption is that the England Premier Legue has more draw matches than the other leagues. This comes from the fact that the other leagues usually have a few “dominating” teams that are too powerful compared to other teams of the league (for example Bayern Munich and Borussia Dortmund at Bundesliga). The difference of team attributes(power) between teams of these leagues are usually significantly clear so matches usually end with a winner. This is said to not happen at Premier League. The plot below shows the number of matches that end with a draw or a winner for each league in the data.

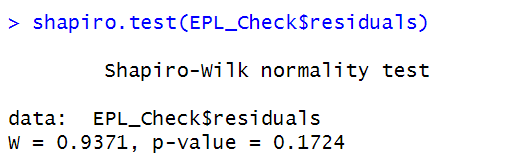


*Figure 2.3.6. The matches’ results for each league*

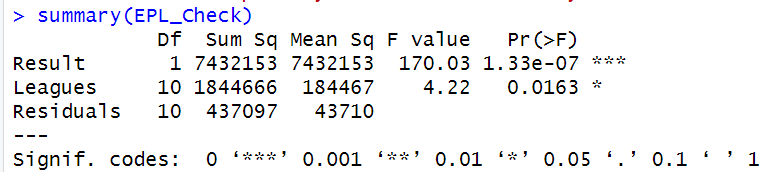
From the plot above, we can see that it looks like there is no difference between the proportions of the draw matches and one team won matches for different leagues. Once again, we will check this assumption statistically using a 2-way ANOVA test and the post-hoc Tukey comparison. By performing a Shapiro-Wilk test, the data is normally distributed.

**Ho : There is no different in the proportion of draw matches and one-team-win matches between leagues**

**Ha: There is at least one difference in the proportion of draw matches and one-team-win matches between leagues**

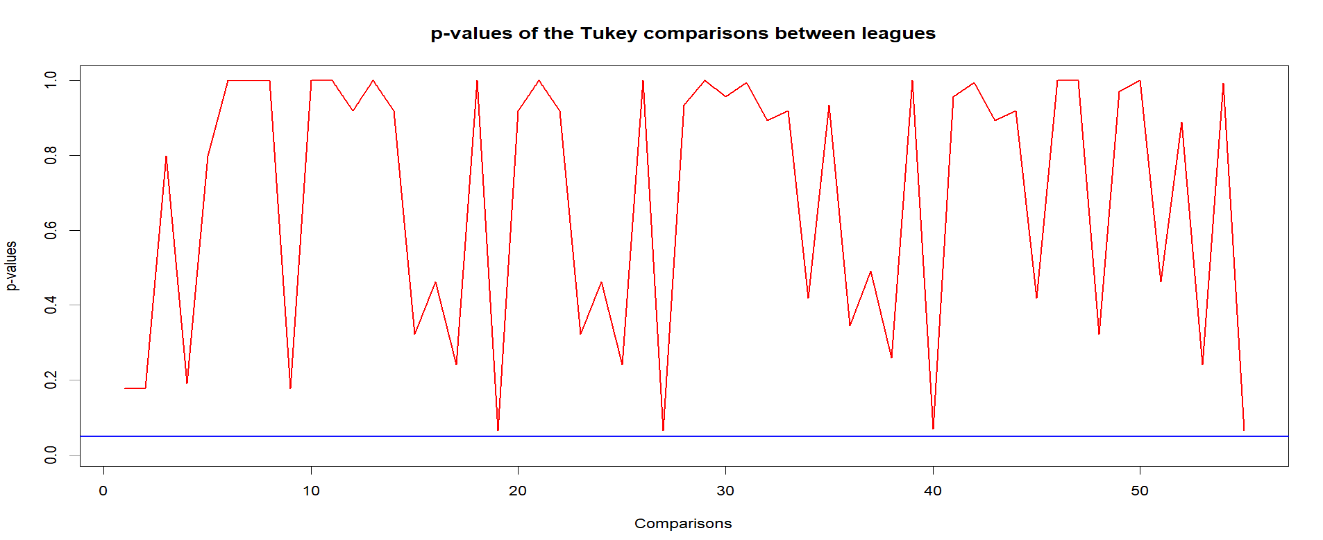


*Figure 2.3.7. Shapiro-Wilk test for normality*



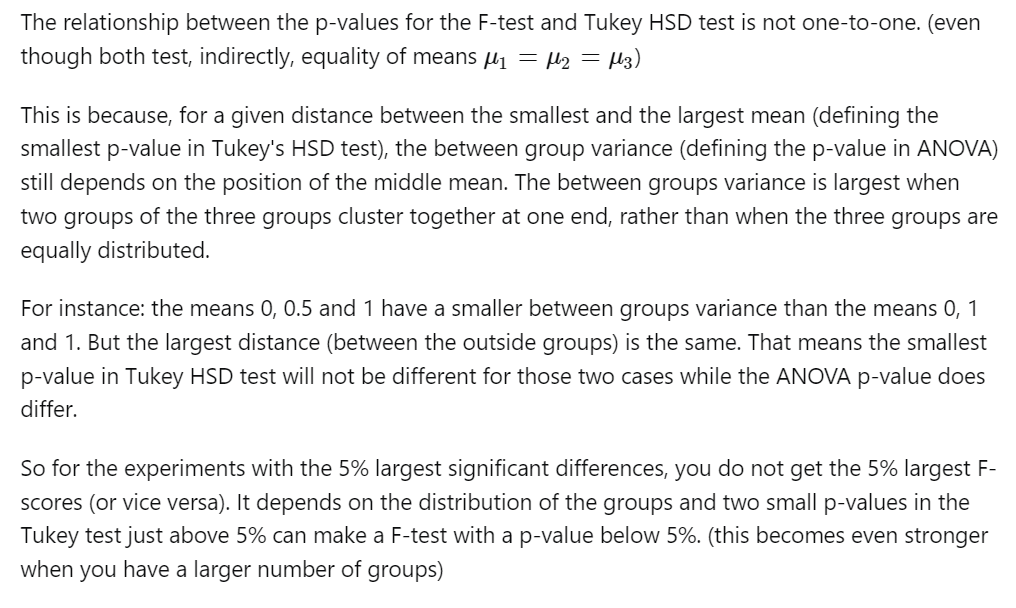
*Figure 2.3.8. Result of the 2-way ANOVA test*

An issue came across while we performed the ANOVA test and the post-hoc comparisons. As shown in the ANOVA table above, there is a significant difference between the leagues; in other words, there is at least one league that will have the proportion of draw-one team won significantly different from the others. However, with the post-hoc Tukey comparisons, none of the pairwise comparisons result in a significant difference. The specific list of all comparisons is long, hence, it will be easy to miss read a value, so we’ve created a line plot of the p-values of the pairwise comparisons instead. As shown below, none of the p-values is under the horizontal line of α (0.05). Filtering out the p-values that are less than 0.05 also resulted in an empty set.



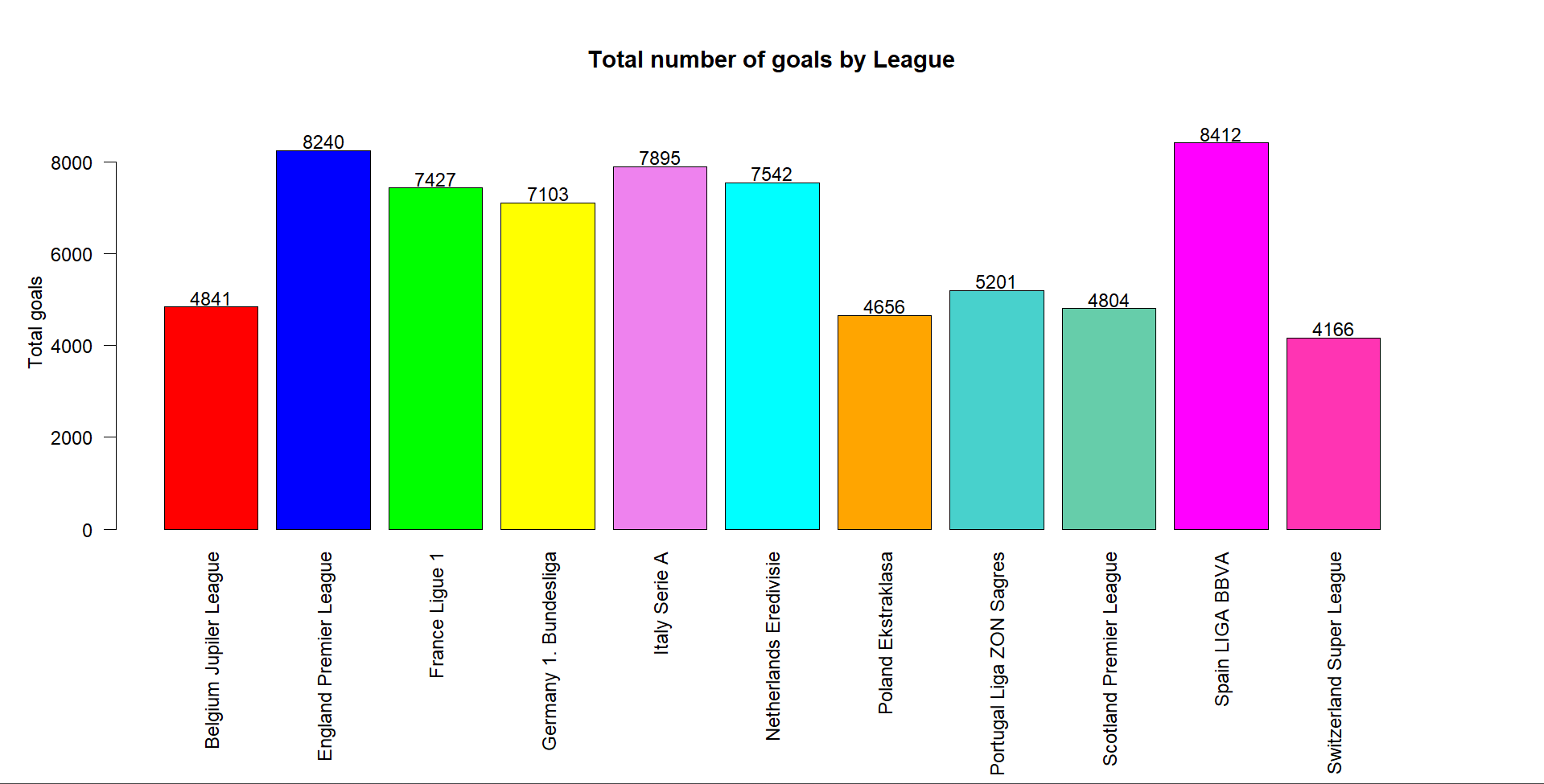
*Figure 2.3.9. The p-values of the Tukey comparisons*

We've done a quick research on the reason for the issue, and we found the explanation below on Stack Exchange that looks to be the best match for our issue. As it said, “becomes even stronger if you have a larger number of groups.” This is true for our data, as the League variable is a factor with 11 levels.



*Figure 2.3.10. The explanation about the difference in p-values of the overall ANOVA and Tukey comparisons*

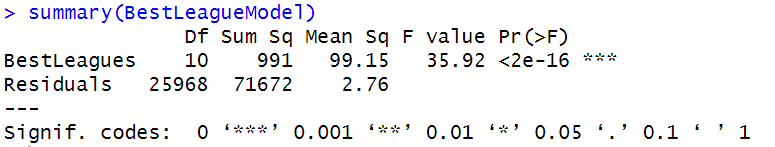
Another fact about soccer is that a large proportion of soccer watchers prefer matches that have goals instead of watching the teams and players’ performances. The total goals of each league are plotted below and a one-way ANOVA test along with Tukey comparisons were performed to see which of the leagues has the most significant goal.



*Figure 2.3.11. The number of goals recorded for each league*

**Ho : The mean number of goals are the same for all leagues**

**Ha : There is at least one difference in the mean number of goals between leagues**



*Figure 2.3.12. Result of the one-way ANOVA test*

From the ANOVA result, we know that there is a significant difference in the number of goals between different leagues, so a post-hoc pairwise Tukey comparison was performed to determine which league is different. Once again, the specific list of all pairwise comparisons is long, so we’ve filtered out the pairs that resulted in not significantly different and used the underline method to represent the result. The method follows the steps:

+ Sort the means of the goals of each league in descending order.



*Figure 2.3.13. Mean number of goals per match for each league, sorted by descending order*

+ Underline those leagues that represented insignificant differences.



*Figure 2.3.14. Pairs that don’t have a significant difference in the mean number of goals*

There are 19 out of 55 pairs that result in insignificant differences, which are shown by the underlining method below. We are counting in pairs, so an underline in 2 leagues means ONE pair. If there are more than 2 leagues in one underline, that means all pairwise comparisons (nC2) in those are insignificant. Checking the underlining below, we have 1+4C2+4C2+3C2+1+1+1=19. If we want to figure out if any 2 leagues have significantly different numbers of goals or not, we only need to check if they fall into one of the underlined groups. For example, Germany 1. Bundesliga and Spain LIGA BBVA are in the same underline, so they don’t have significantly different numbers of goals. England Premier League and France Ligue 1, on the other hand, are significantly different, and since the leagues have been put in descending order, EPL has significantly more goals than Ligue 1.

*Figure 2.3.15. Result of the underline technique*

# **3. MODELING**

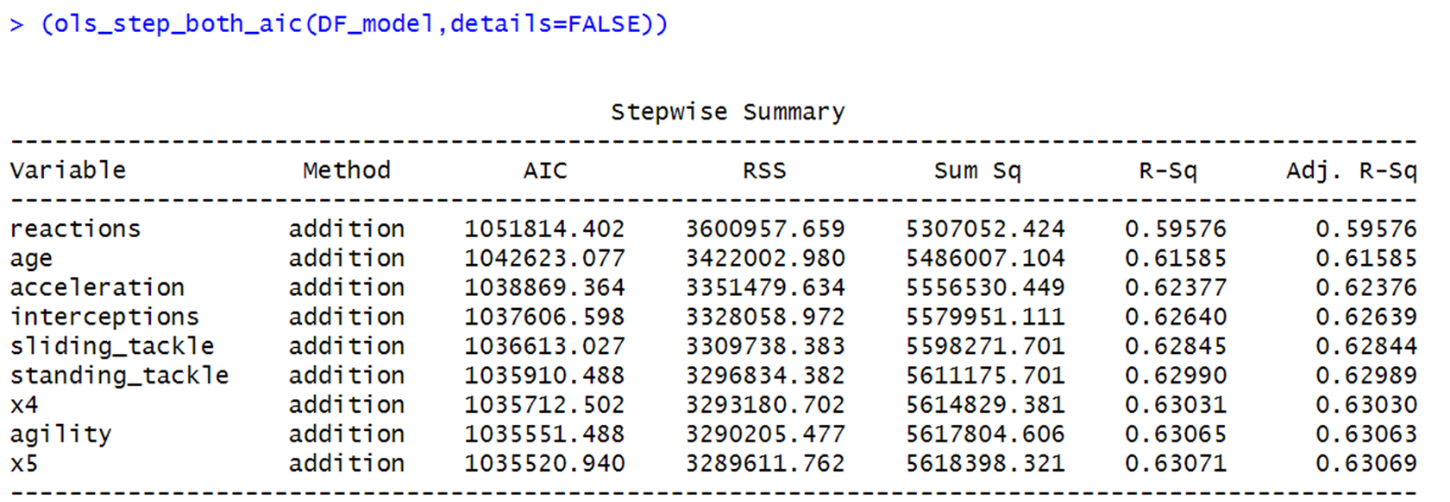
## **3.1 VARIABLE SELECTION**

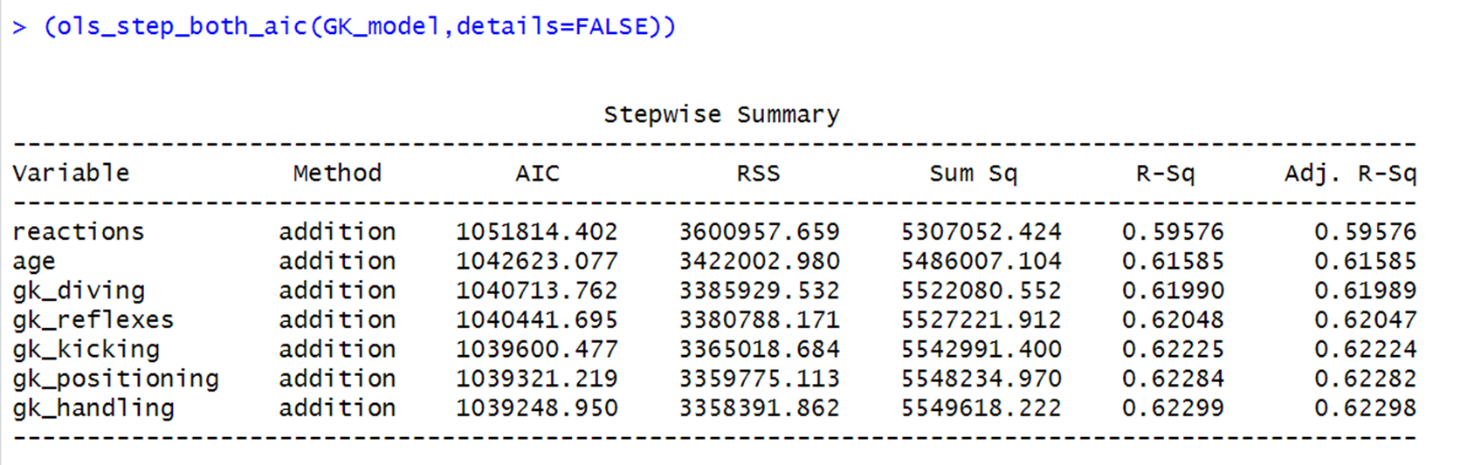
Many factors that can affect a player’s performance on the field. In this project, we’ve created 4 models for different positions: goalkeeper, defender, midfielder, and forwarder; we then performed a variable selection to see which factors contribute significantly to that position. We will use the Player\_Attributes table for this technique. The function ols\_step\_both\_aic() was used. It will perform stepwise variable selection two ways (both forward and backward) and select the variables based on AICs.

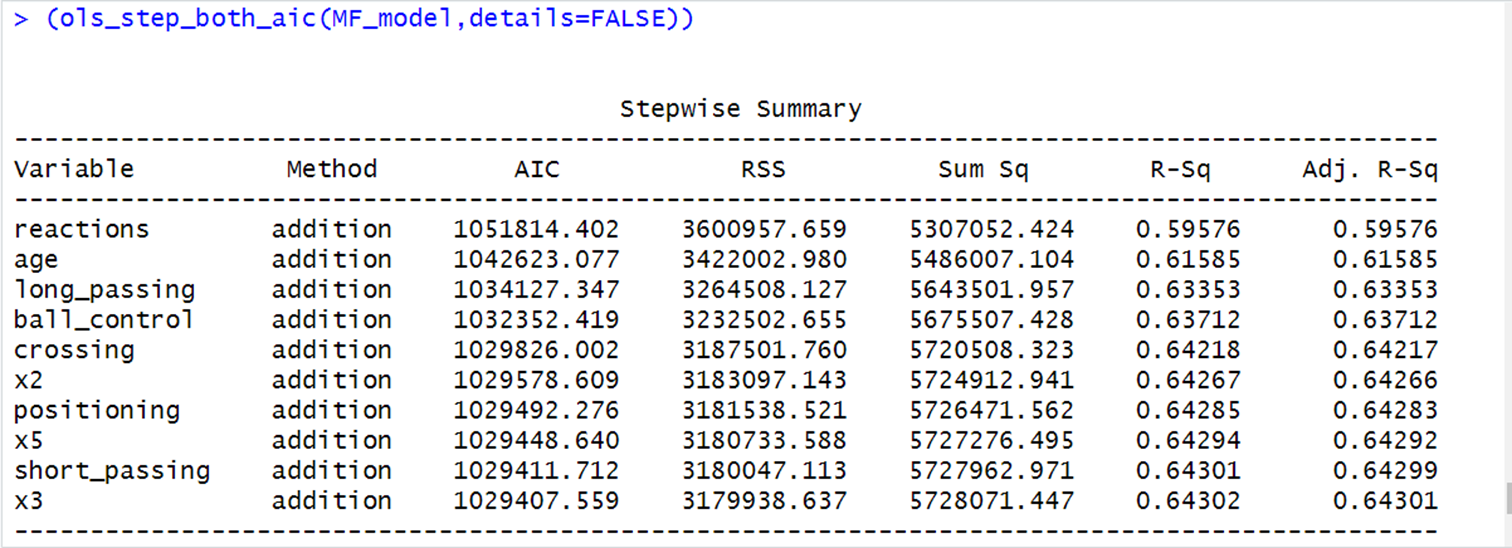
Before creating models and performing a variable selection, we need to prepare the data first. There are some categorical variables, so we need to create dichotomous (dummy) variables for them.

* x1 = 1 if preferred\_foot is the right foot, and 0 if it’s the left foot
* x2 = 1 if attacking\_work\_rate is high, 0 otherwise
* x3 = 1 if attacking\_work\_rate is medium, 0 otherwise
* x4 = 1 if defensive\_work\_rate is high, 0 otherwise
* x5 =1 if defensive\_work\_rate is medium, 0 otherwise

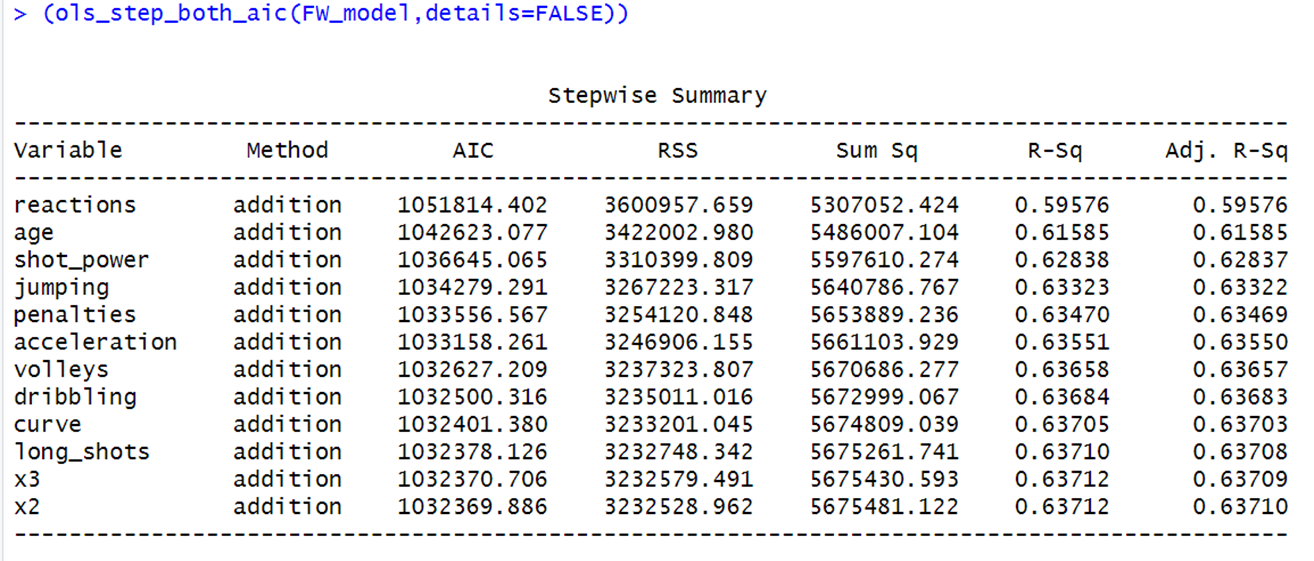
For two pairs, x2 and x3, x4 and x5, since they both refer to one variable, in variable selecting, whether the decision is dropping or keeping, we will need to do it for BOTH (i.e., we cannot drop one and keep another).

*Figure 3.1.1. The variable selection model for Defender*

*Figure 3.1.2. The variable selection model for Goalkeeper*



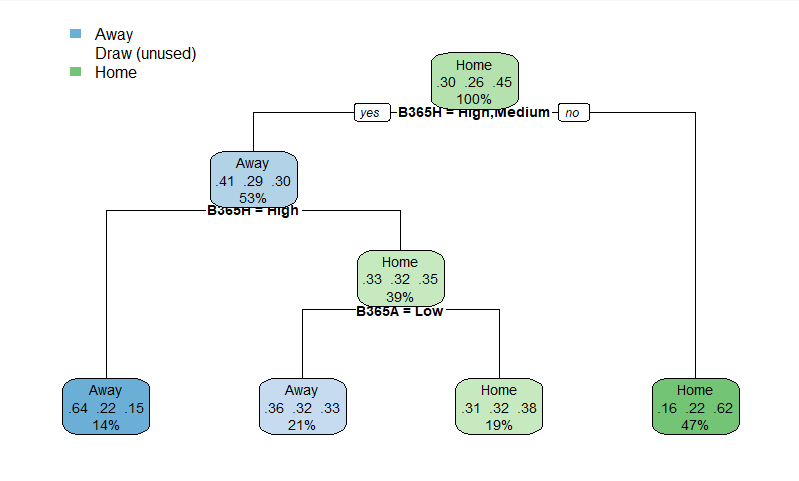
*Figure 3.1.3. The variable selection model for Midfielder*



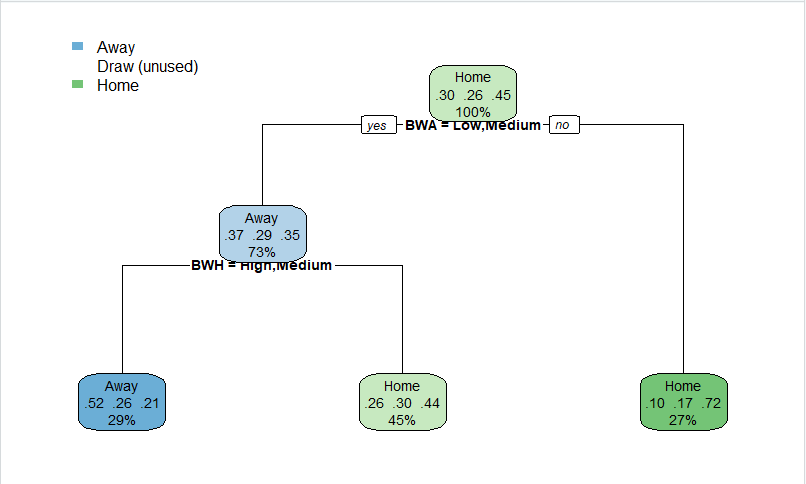
*Figure 3.1.4. The variable selection model for Forwarder*

These results give us the variable selection for a defender, goalkeeper, midfielder, and forward. The top two for each of these models are both age and reactions. This tells us that these two attributes give us the best way to overall evaluate players no matter which position. After those two attributes, we can see that there are different attributes for each grouping that have the most impact on the players' performance on the field. As mentioned earlier, we need to be careful with x2 and x3, x4 and x5, and that’s what happened in the model for Midfielder. The selection method only selected x5, but we must decide whether to keep both x4 and x5 or drop both. Since a midfielder’s work also requires a lot of defending, we will use both x4 and x5 for the model. It is also worth noting that for different positions, different variables were selected outside of age and reactions. Using these results, we hope teams can look at one player and figure out what they might value in making one player better than another.

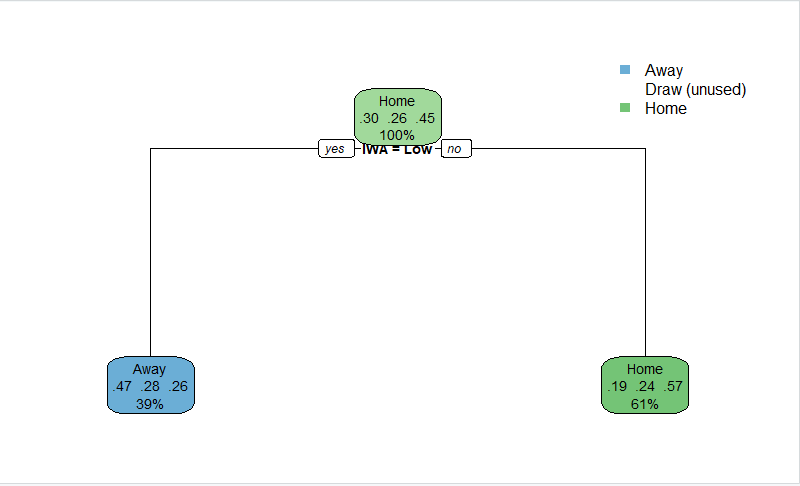
## **3.2 DECISION TREE**



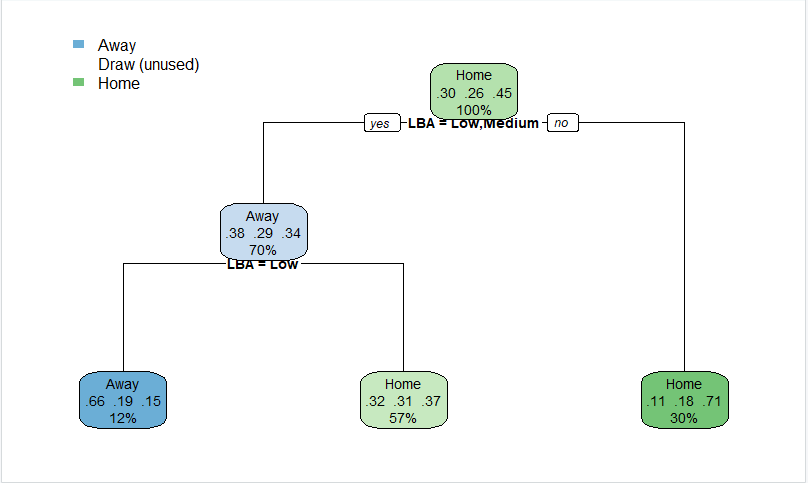
*Figure 3.2.1. The decision tree for the B365 betting provider*



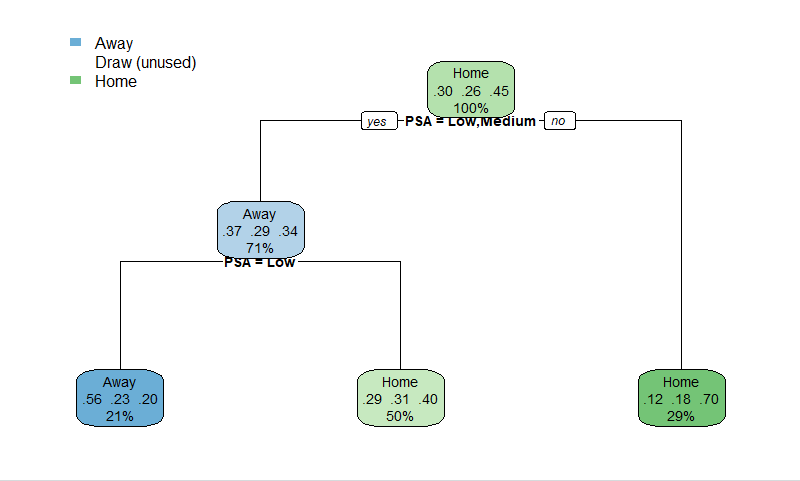
*Figure 3.2.2. The decision tree for the BW betting provider*



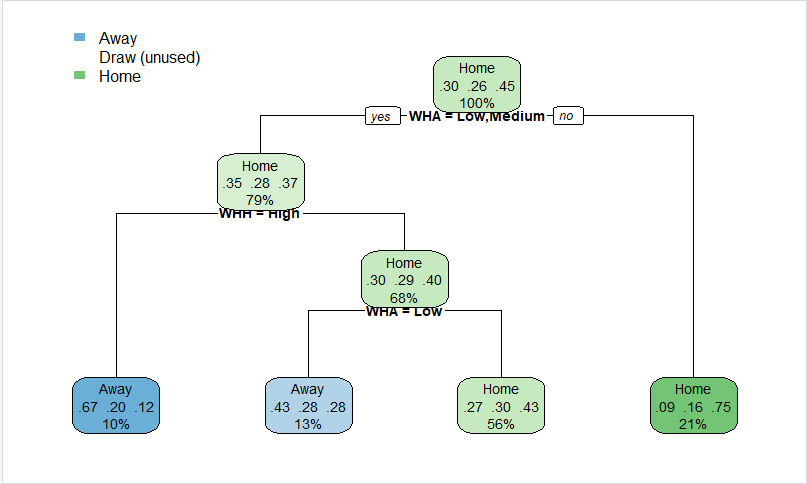
*Figure 3.2.3. The decision tree for the IW betting provider*



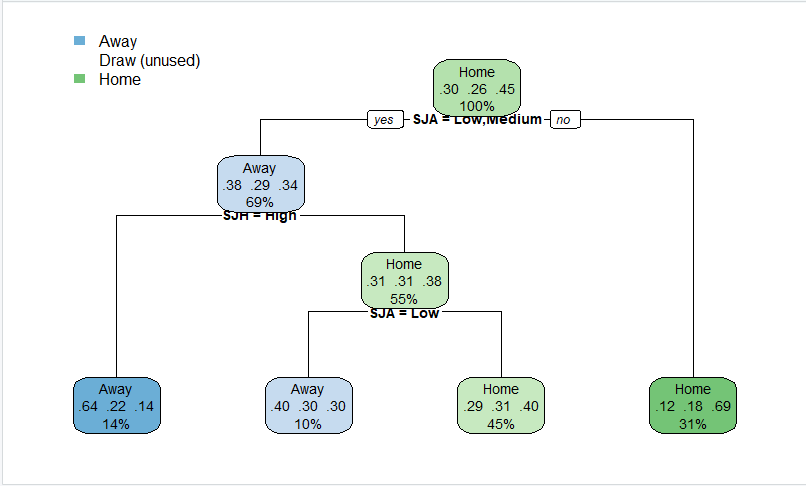
*Figure 3.2.4. The decision tree for the LB betting provider*



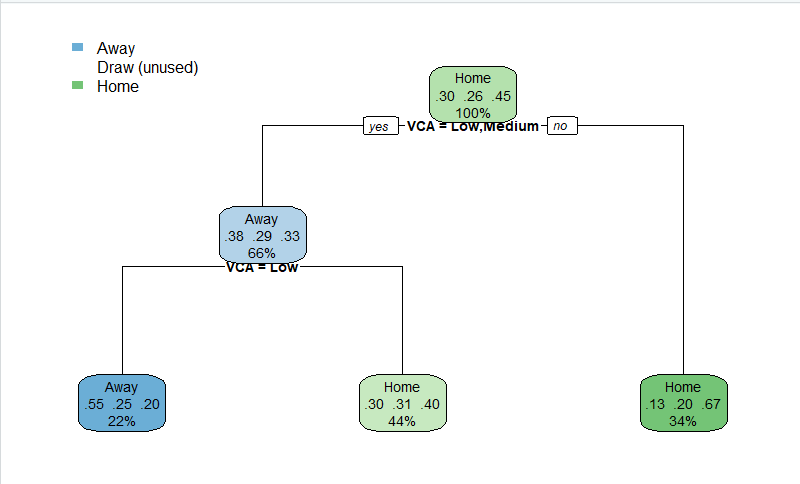
*Figure 3.2.5. The decision tree for the PS betting provider*



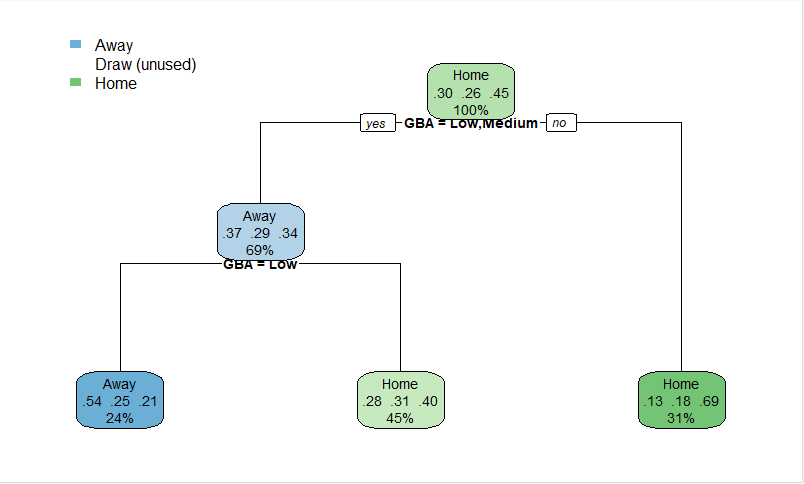
*Figure 3.2.6. The decision tree for the WH betting provider*



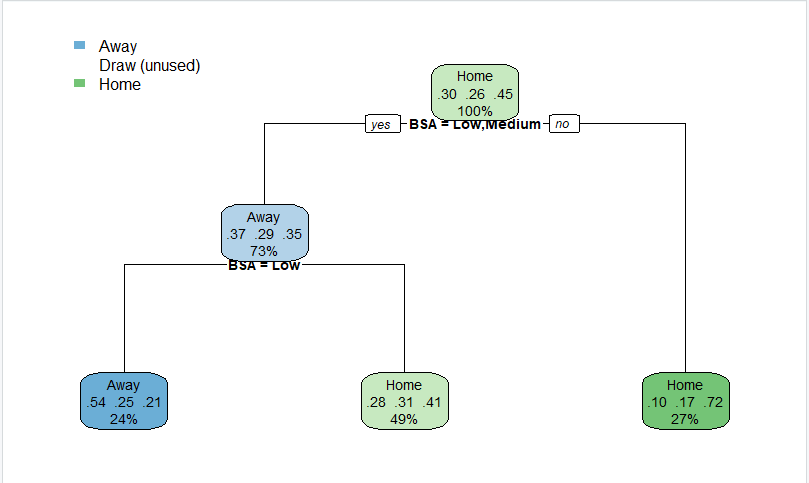
*Figure 3.2.7. The decision tree for the SJ betting provider*



*Figure 3.2.8. The decision tree for the VC betting provider*



*Figure 3.2.9. The decision tree for the GB betting provider*



*Figure 3.2.10. The decision tree for the BS betting provider*

The decision trees above give us predictions for the match results based on the betting odds. There are 10 different betting providers, and each has 3 odds, so there are 30 odds in total. With H denoting a victory for home teams, D denoting a draw, and A denoting a victory for away teams, we can make predictions on the match’s result given certain betting odds. The above gives us the results for each of the ten betting models using decision trees. After this is done, we use the Naive Bayes results to compare the accuracy of the two methods and decide which one is better. Since we have 10 different providers, the decision may be different between them, in other words, for some providers, the decision tree model will be better while the Naive Bayes will be better for others.

## **3.3 NAÏVE BAYES**

As mentioned above, we also used the Naive Bayes to compare the accuracy of each provider with the decision tree models. This is a probabilistic machine learning algorithm based on conditional probability known for its simplicity and efficiency, which has proved to be a great tool for our match results predictions. Its application involves scrutinizing historical betting odds data which encompasses home and away teams. By incorporating odds data into its analysis, this model can factor the predictions and other influences to refine its predictions. This allows it to discern patterns and relationships that contribute to more accurate predictions regarding match outcomes.

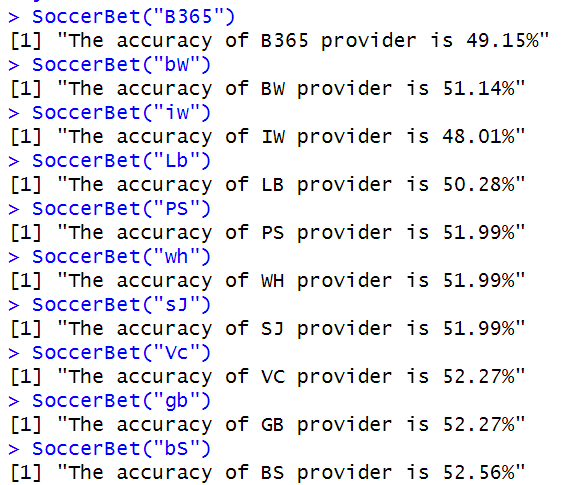
In the dynamic landscape of soccer betting, where odds are indicative of market sentiments, Naive Bayes becomes a valuable ally for enthusiasts and analysts seeking an edge. It not only considers traditional performance metrics but also adapts to the ever-changing odds landscape, providing a nuanced understanding of the evolving dynamics in the soccer betting market. This integration of Naive Bayes with betting odds underscores the synergy between sophisticated algorithms and real-time market information, offering a data-driven approach to soccer predictions that goes beyond conventional analysis.

## **3.4 CLUSTERING**

For this project, we are performing k-means clustering with the Team\_Attributes table. Before clustering, we must standardize the values with the scale() function. This is because if we leave all of them in raw data, the range of the values will be large, hence, it will make the clustering result worse. RStudio has the function kmeans() to do the technique. The result will include the clusters’ sizes, the clusters’ centers, and a vector showing which cluster each observation belongs to. One important thing is that they’re only the outputs of the final iteration. To see the clustering process, we need to use the function kmeans.ani(), which will provide us the animation process of selecting the centers, assigning each observations to the clusters, and it will repeat again and again until we get the final result.

# **4. RESULTS**

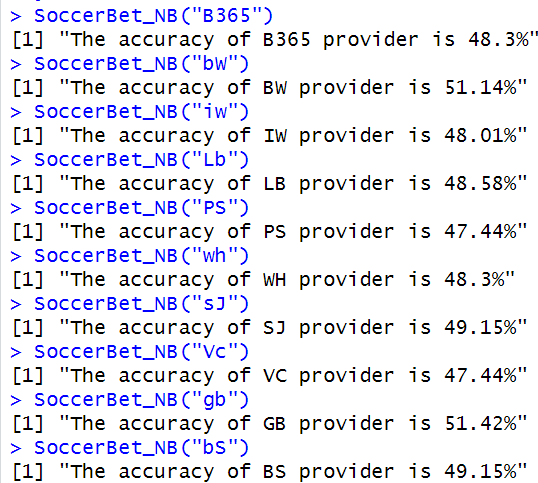
## **4.1 DECISION TREE RESULTS**



*Figure 4.1.1. The accuracy of each provider using the decision tree model*

Looking at the decision tree results, we can see that the accuracy of these is all below 55%, but given the scenario that what we are looking for is based on betting, a type of gambling, we think that they’re appropriate. From the results, the decision tree of the betting provider BS is 52.56%, which is the highest of the providers.

**4.2 NAÏVE BAYES RESULTS**



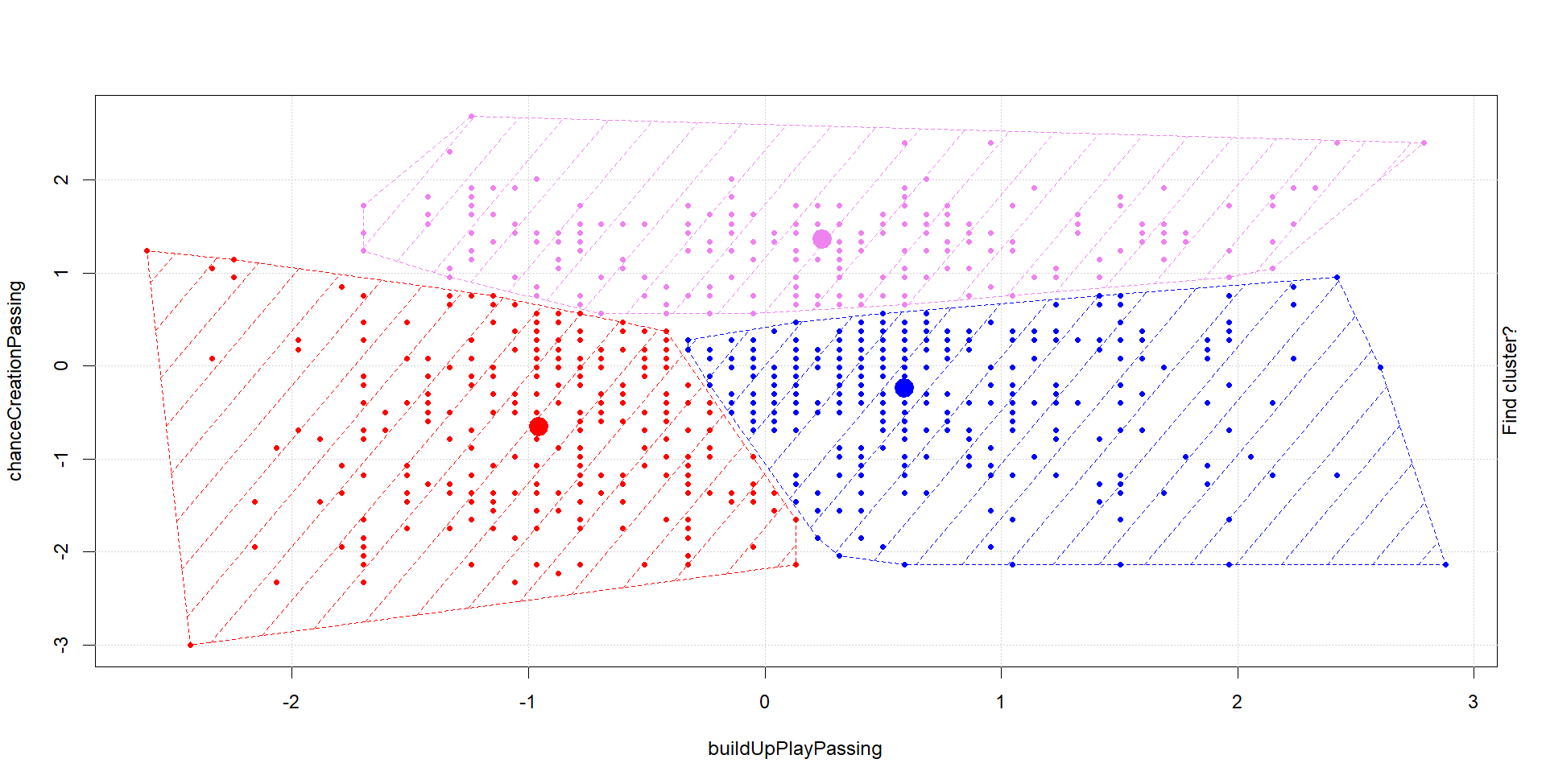
*Figure 4.1.2. The accuracy of each provider using the Naive Bayes model*

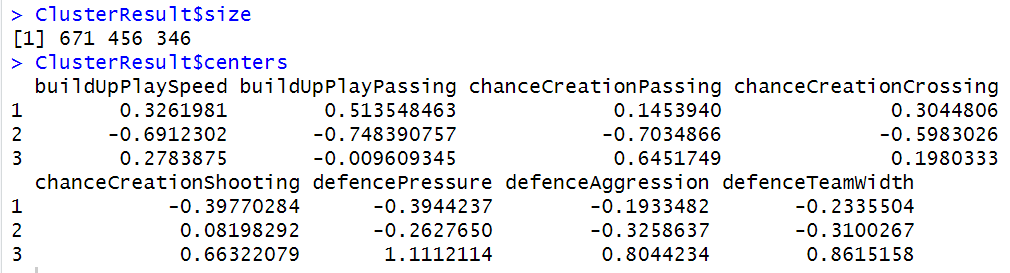
Similarly, the accuracy of all providers is below 55% but still appropriate. The provider with the highest accuracy is the GB provider at 51.42%.

**4.3 MODEL COMPARISION**

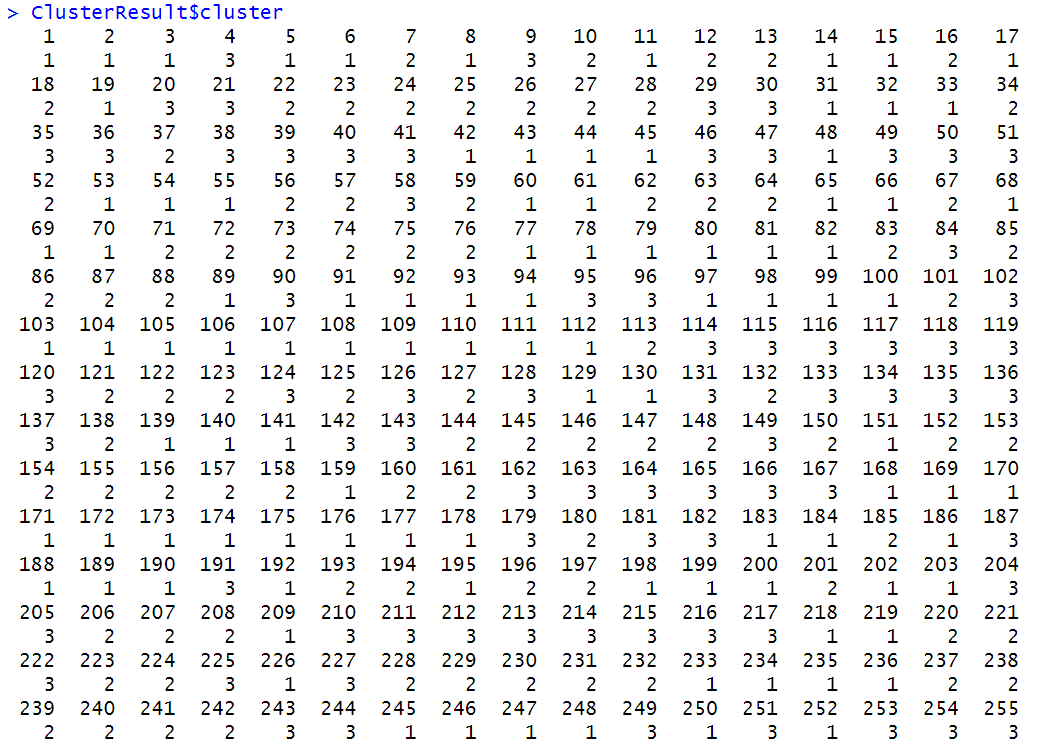
As mentioned earlier, we will compare the accuracy between the two models for each provider to determine the better one. Comparing results yields that the decision tree model is better for all providers. An interesting fact about the decision tree is that it will change every time we run the program. This is because we used shuffled data with the shuffled index being a randomly generated vector, so it won’t be the same in every run. The final decision of each tree is not likely to change as it will require a significant change in the data, but the probabilities and the accuracy will vary. This is the reason why during the in-class presentation, we had a provider that resulted in the Naive Bayes being the better model, but the result above shows that the accuracy of the decision trees models is all better than the Naive Bayes models.

## **4.4 CLUSTERING RESULTS**

*Figure 4.4.1. The clustering result of the final iteration*



*Figure 4.4.2. The sizes and centers of the clusters*



*Figure 4.4.3. The vector showing which cluster each observation belongs to. There are 1473 observations in* *total and this is only the first part of it.*

The team analysis, conducted through comprehensive cluster analysis, has yielded valuable insights into the dynamics and composition of the various groups within the team. The results indicate distinct clusters that share common characteristics, suggesting the presence of subgroups with similar working styles, communication patterns, or skill sets. This segmentation provides a nuanced understanding of team dynamics, enabling targeted interventions to enhance collaboration and overall performance. The identification of clusters also opens avenues for leveraging diverse strengths within each subgroup, fostering a more cohesive and synergistic team environment. These findings serve as a foundation for strategic decision-making and tailored interventions aimed at optimizing team effectiveness and achieving organizational goals.

# **5. CONCLUSION**

In conclusion, we want to journey through the realm of data mining to predict soccer outcomes. Our aim is to bridge the gap between theoretical knowledge and real-world implementation in the realm of soccer. Our project is to demonstrate the potential of leveraging available data to make informed predictions in the dynamic domain of soccer. By looking to leverage data mining techniques we are aiming to extract meaningful patterns, uncover trends, and enhance our understanding of the factors influencing match results. The primary goal of this project is to showcase the feasibility of employing these techniques as valuable tools for enthusiasts, analysts, and anyone interested in gaining a competitive edge in predicting soccer matches. There are still a lot more techniques that we want to use as they will provide more choices to determine the best decisions on various outcomes. Since RStudio allows us to write our own functions with function(), we also want to try modifying the processes and results of some techniques ourselves instead of just using the built-in functions. We look forward to continuing to develop this project and are excited to use these methods moving forward!

# **6. REFERENCES**

[1] The European soccer data, <https://www.kaggle.com/datasets/hugomathien/soccer>

[2] Significant Difference in ANOVA and Pairwise Tukey’s, [post hoc - How can I get a significant overall ANOVA but no significant pairwise differences with Tukey's procedure? - Cross Validated (stackexchange.com)](https://stats.stackexchange.com/questions/16665/how-can-i-get-a-significant-overall-anova-but-no-significant-pairwise-difference)

[3] Katherine Price, How Much Does Home Field Advantage Matter in Soccer Games? A Causal Inference Approach for English Premier League Analysis, [How Much Does Home Field Advantage Matter in Soccer Games? A Causal Inference Approach for English Premier League Analysis (arxiv.org)](https://arxiv.org/pdf/2205.07193.pdf)

[4] Angelo Bravo, Analysis of Individual Player Analysis and Their Effect on Winning in College Soccer, [Analysis of Individual Player Performances and Their Effect on Winning in College Soccer (smu.edu)](https://scholar.smu.edu/cgi/viewcontent.cgi?article=1176&context=datasciencereview)