|  |
| --- |
| IST707 DATA MINING |
|  |

|  |  |
| --- | --- |
| Date Submitted  4/5/2020 | Group Project NFL Play by Play |
| Professor | Jeremy Bolton |
| Submitted by: | Beverlyn Tucker, Kyle Welch and Gilbert Guyah |

**INTRODUCTION**

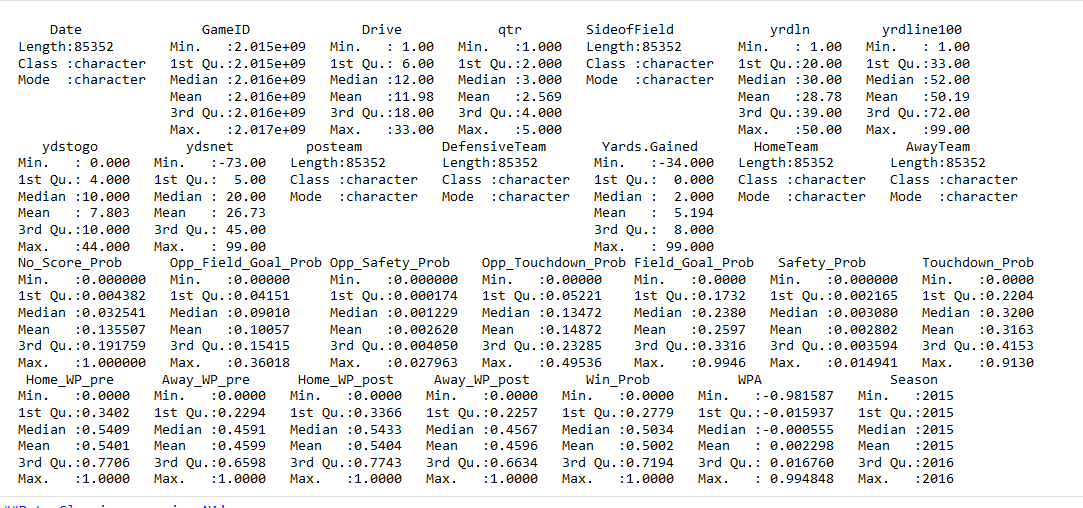
The NFL lacked any advanced statistical data that was available publicly, because of this the Sports Analytics Club from Carnegie Mellon created an R package, nflscrapR, that compiles NFL play by play data. Using the NFL’s API, the R package scrapes, cleans, parses, and output datasets, an example of this is the 2009-2016 NFL play by play dataset that was made available on Kaggle by the group from Carnegie Mellon. The dataset has 356,768 rows and over 100 columns. Each play is broken down into great detail containing information on the game situation, players involved, results, and advanced metrics such as expected point and win probability values.

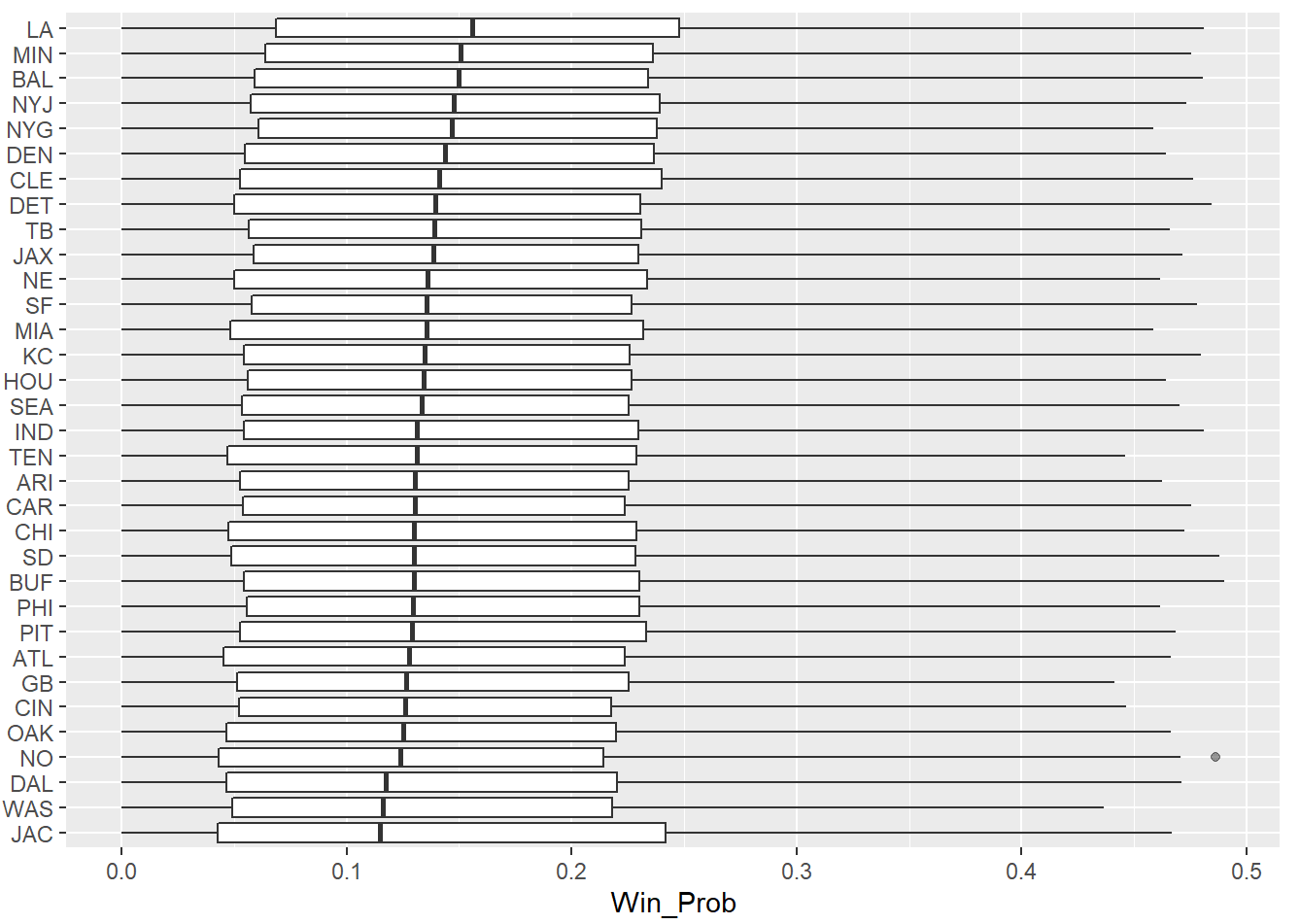
Due to the overwhelmingly large size of the data set, it was determined that the scope of this project would primarily only cover the 2016 NFL season, this limited the data set to over 80,000 rows of play by play data. Similarly, there was a vast number of variables or columns for each row of play by play data, which allowed a variety of ways for the data to be approached. The determination was made to reduce the number of columns to 16 or about 15% of the columns in the original data set. Examples of some of the chosen columns include winning probability, play type, touchdown probability, etc. Due to this data not being released officially by the NFL, but a third party, the database included many NAs that made a lot of the individual game data unusable for this analysis. The size of the dataset, coupled with the number of games that were unusable because of NAS, made the data severely challenging to approach, and this analysis was limited by the team’s computer’s processing abilities.

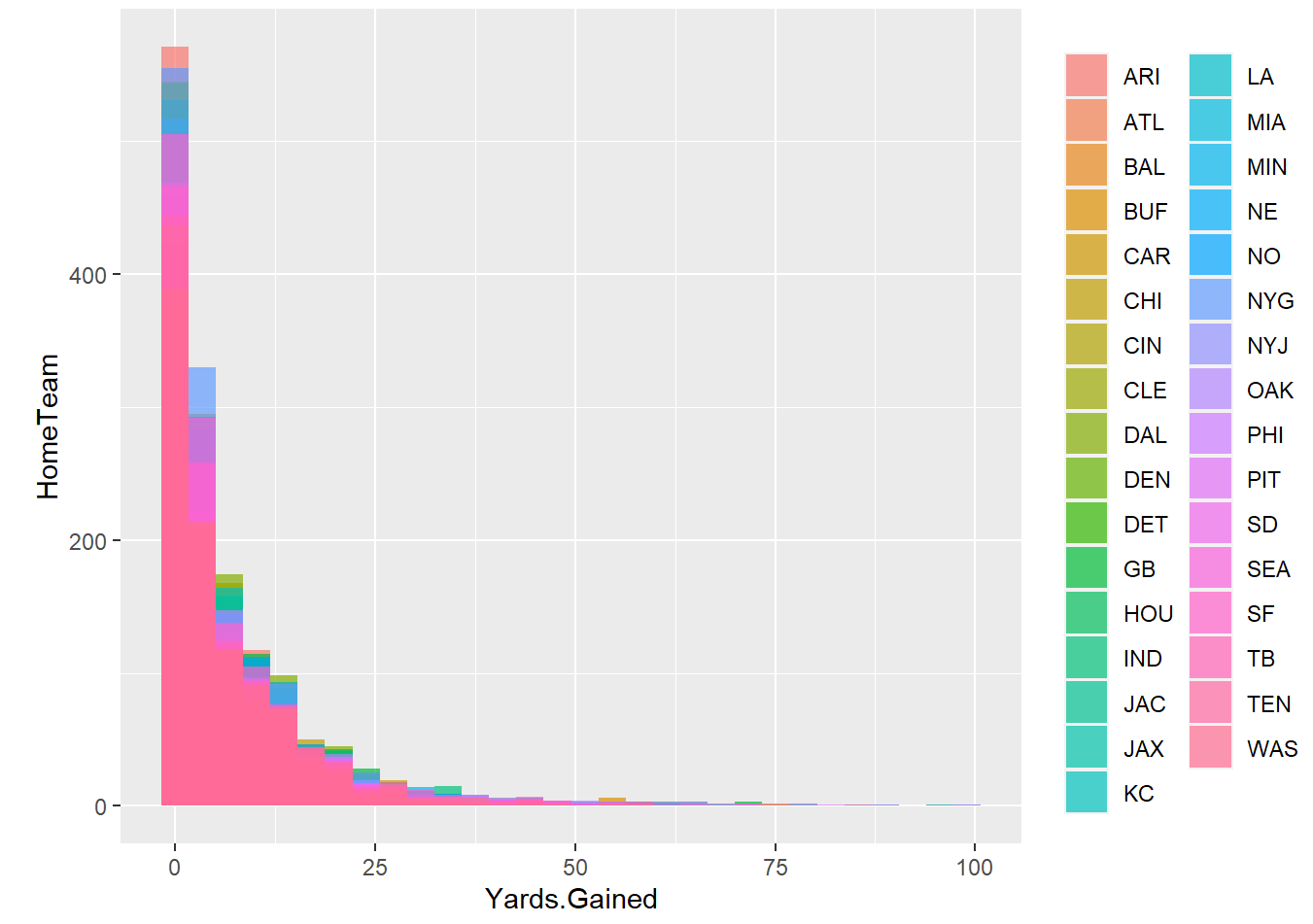
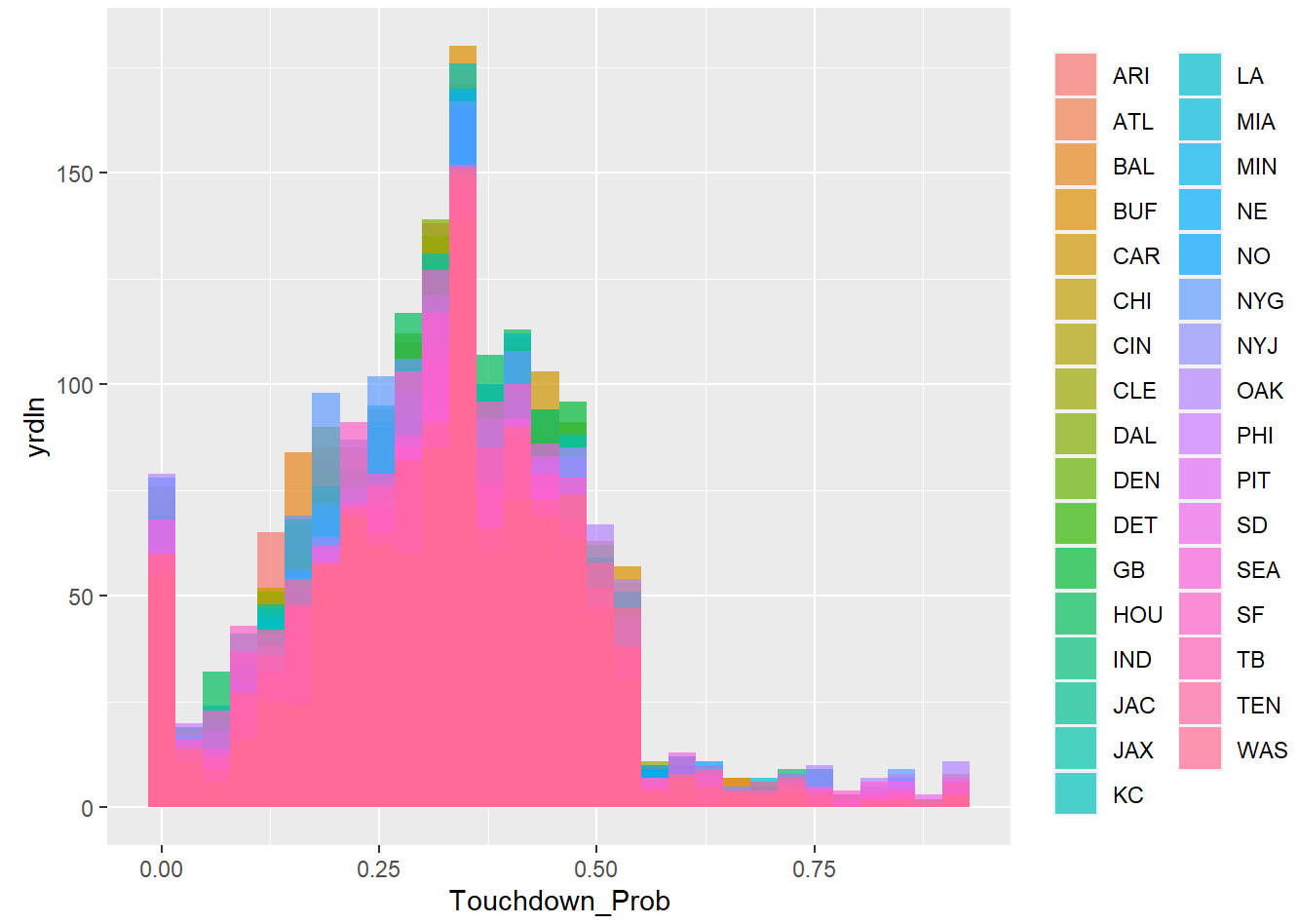
The goal of this project is to analyze the NFL based on historical data and create models to predict future team performance and game outcomes. The majority of analysis primarily focused around the 2016 NFL season; however, some year by year trend data was also created, specifically for the analysis of NFL conferences. To analyze the dataset, these models, such as random forest, linear, kNN, and decision tree, were used. Additionally, data analysis techniques, such as clustering and data visualization, were utilized.

**BASIC ANALYSIS**

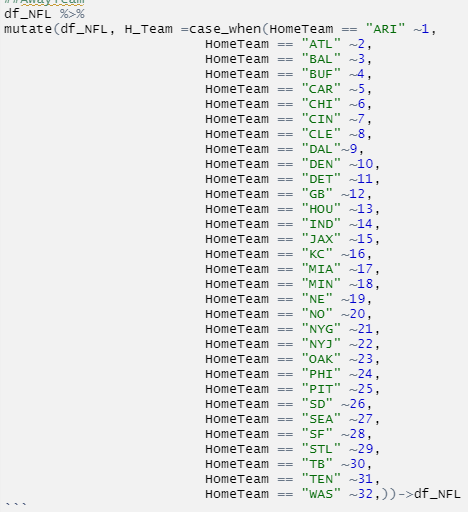
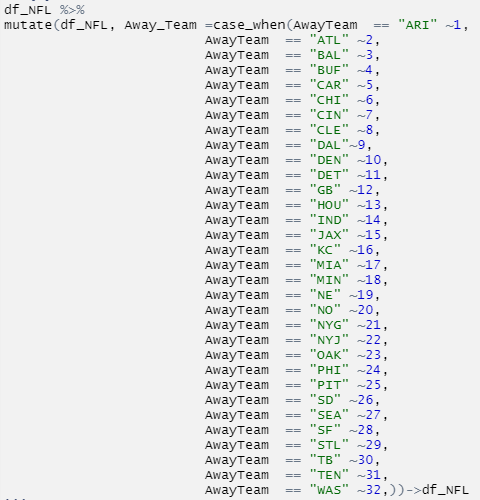
Analyze the data set using the summary function on the data to understand the distribution of data by calculating descriptive statistics of numerical and visualize using graphs (histograms and box plots)



****

** **

Discretized Home Team And Away Team Ready For Prediction

****

The response variables were converted factors and the data split into training and test sets. The models that will be used – K-nearest neighbors, decision tree, random forest, and Linear Model

Models will be trained using the caret package; teams and play type will be used as the feature selection for each model predictor. Cross-validation is used to measure model performance and prediction accuracy.

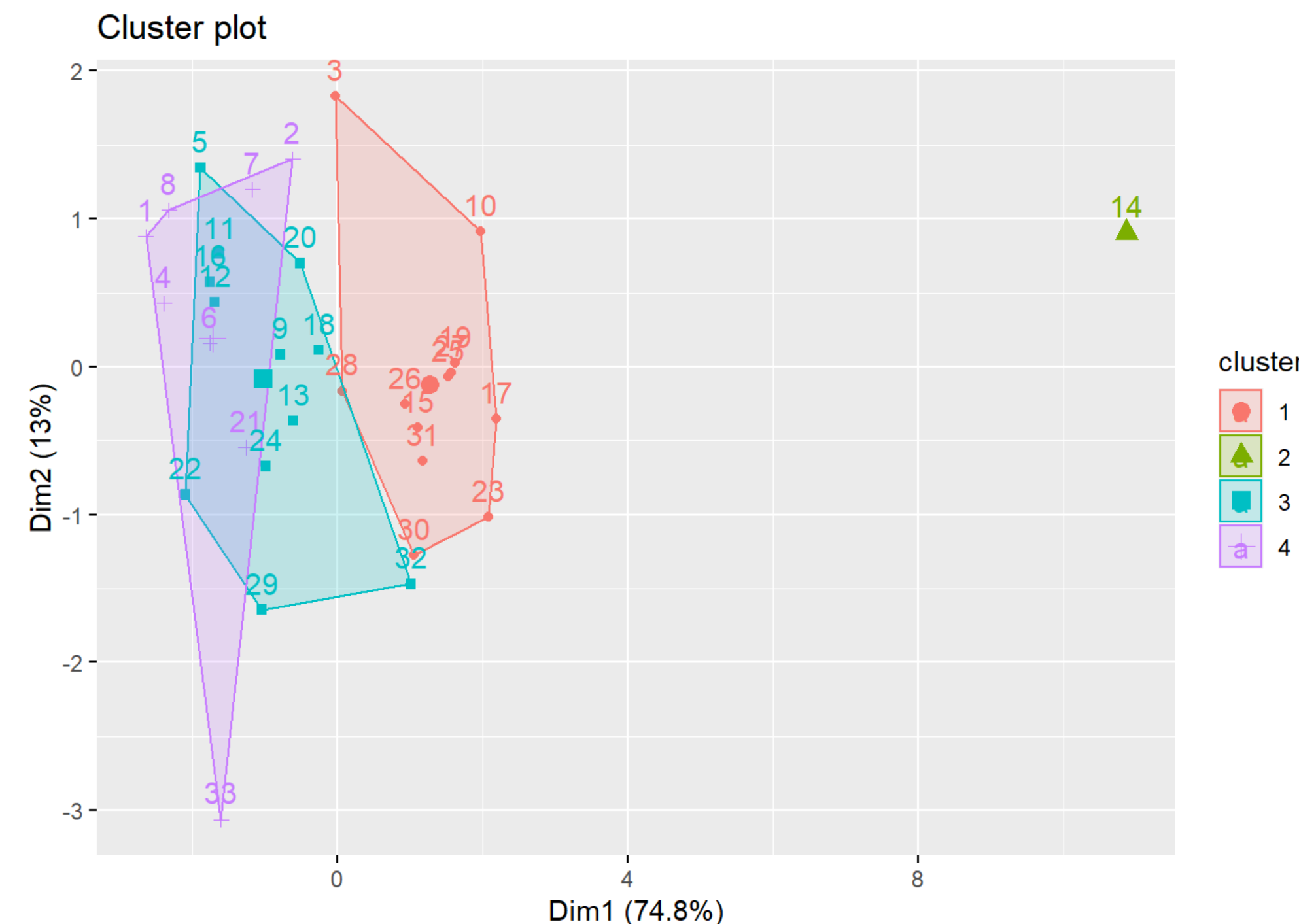
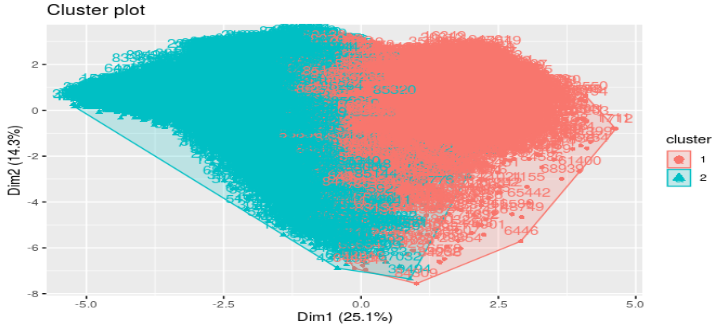
· KNN algorithm assumes that similar things exist in close proximity. Use of this algorithm will determine the teams with similar win probability (WPA). K-nearest neighbors 3 kernels will be used and for distance 2 is the Euclidean distance.

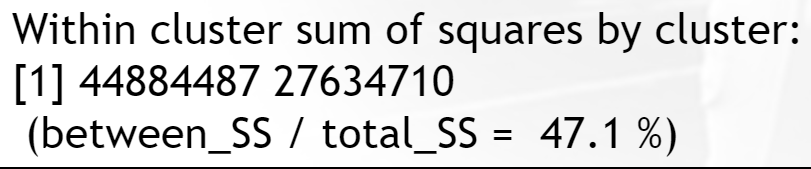
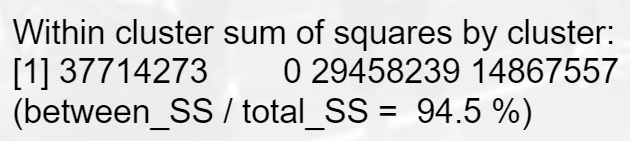
· Decision Tree - a maxdepth = 1, method = class were selected for this supervised machine learning algorithm, based on the series of sequential decisions made to determine a specific result

· Random Forest - a mtry 2 and ntree 500 were selected to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree

· Linear Model - a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

**K-means**

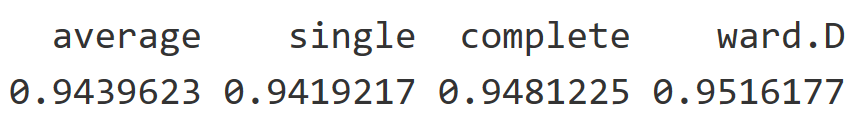
****

** **

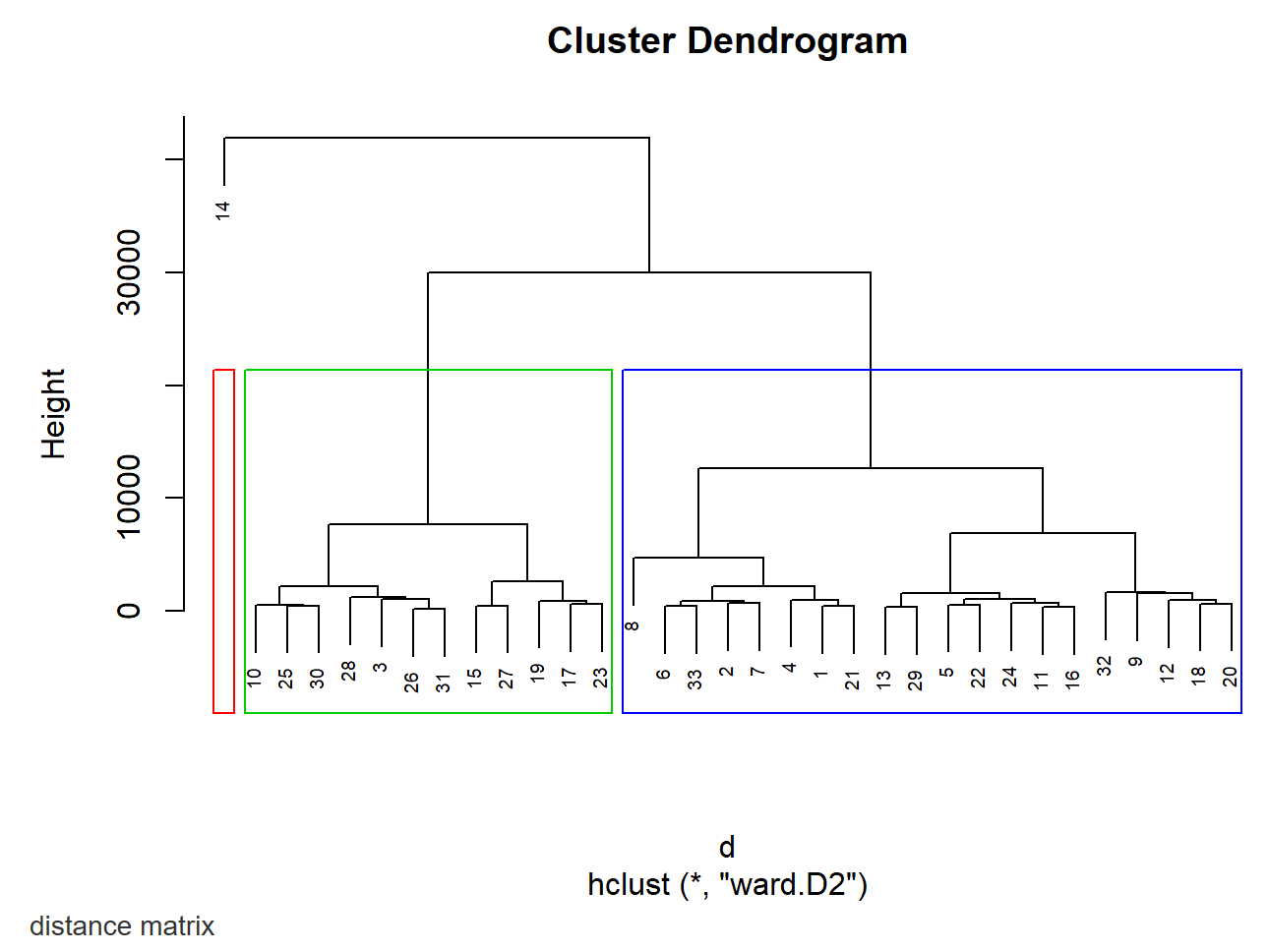
**What is the Best method for Clustering**

**m <- c( "average", "single", "complete", "ward")**

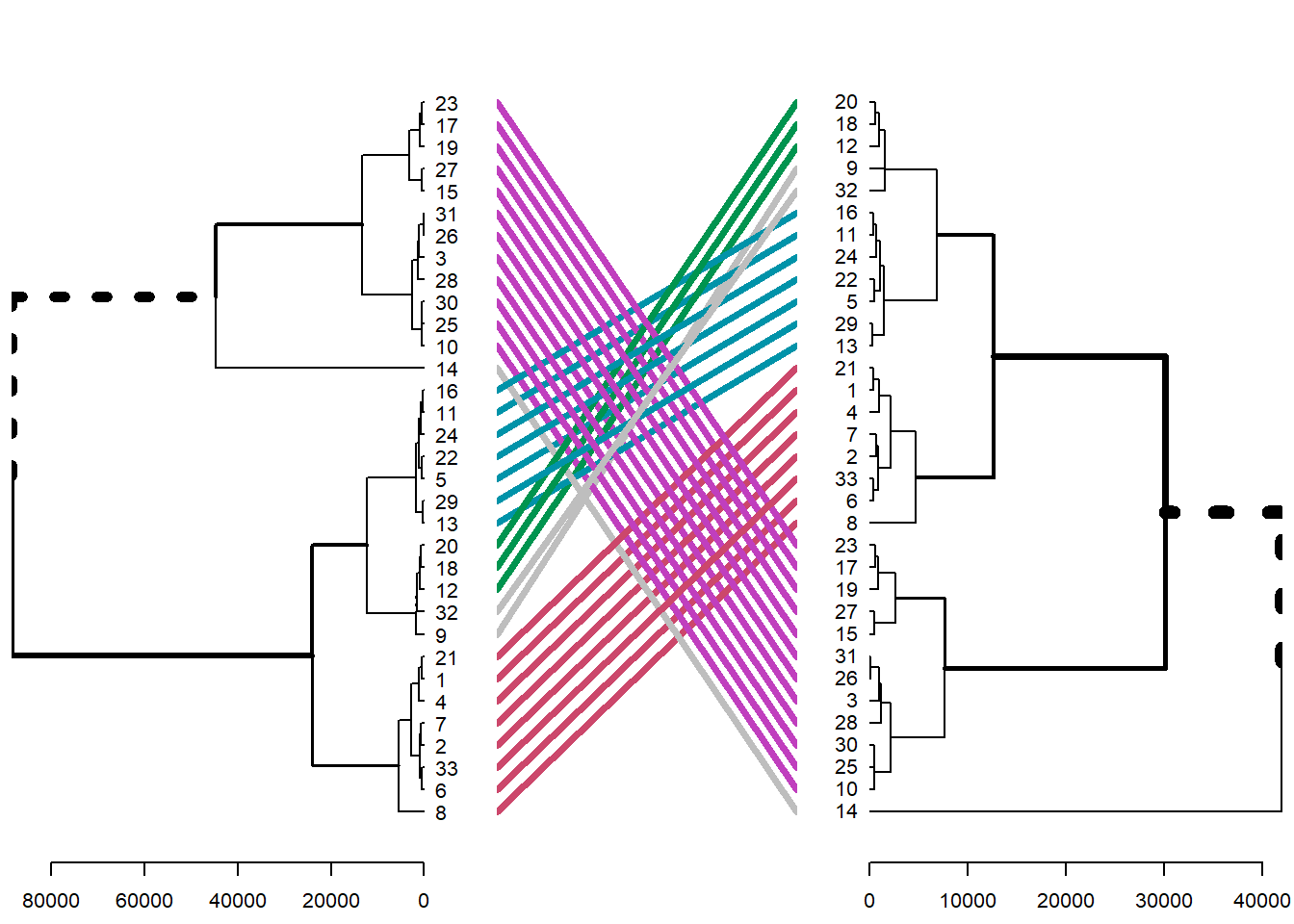
**names(m) <- c( "average", "single", "complete", "ward.D")**

****

Based on four choices of clustering, Ward.D has the best probability of 95%.

****

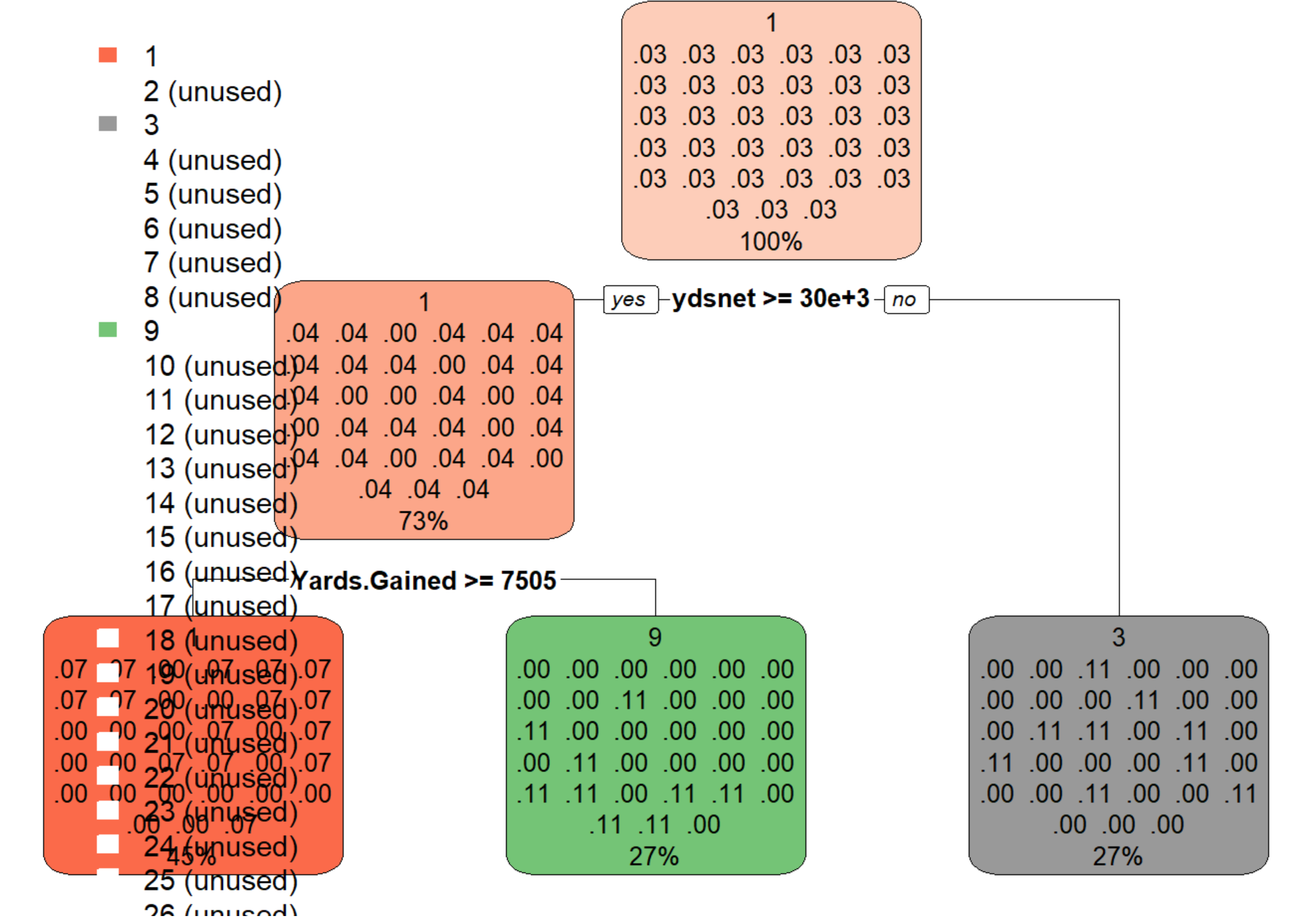
A dendrogram is a diagram representing a tree. This diagrammatic representation is frequently used in different contexts: in hierarchical clustering. The cluster Dendrogram with 3 clusters. The first height is around 250 and most of the data is similar in the first height.

****

A tanglegram is a pair of trees on the same set of leaves with matching leaves in the two trees joined by an edge. The trees above match each other so there were no new data introduced to the trees.

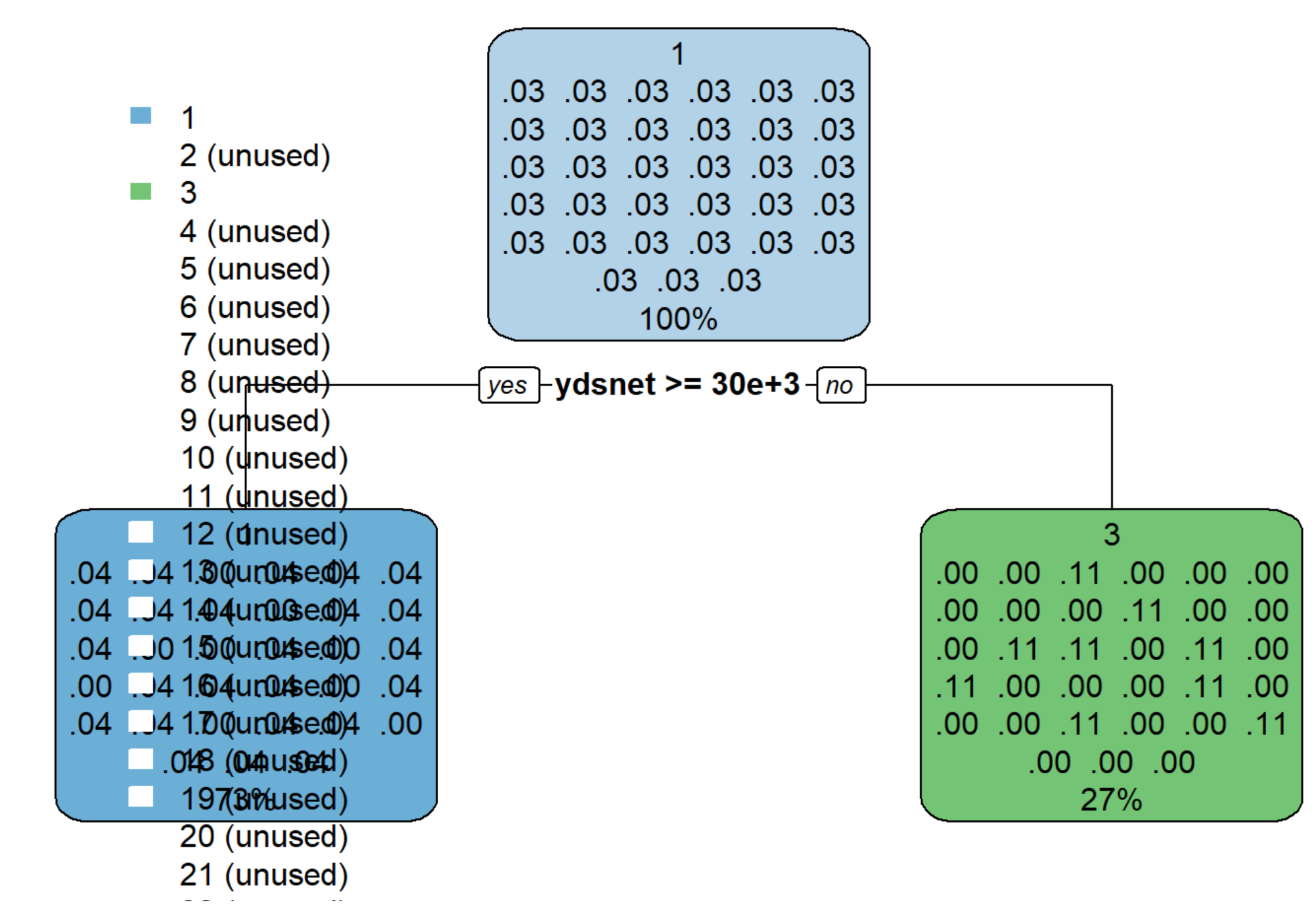
**MACHINE LEARNING MODELS**

**Decision Tree**

****

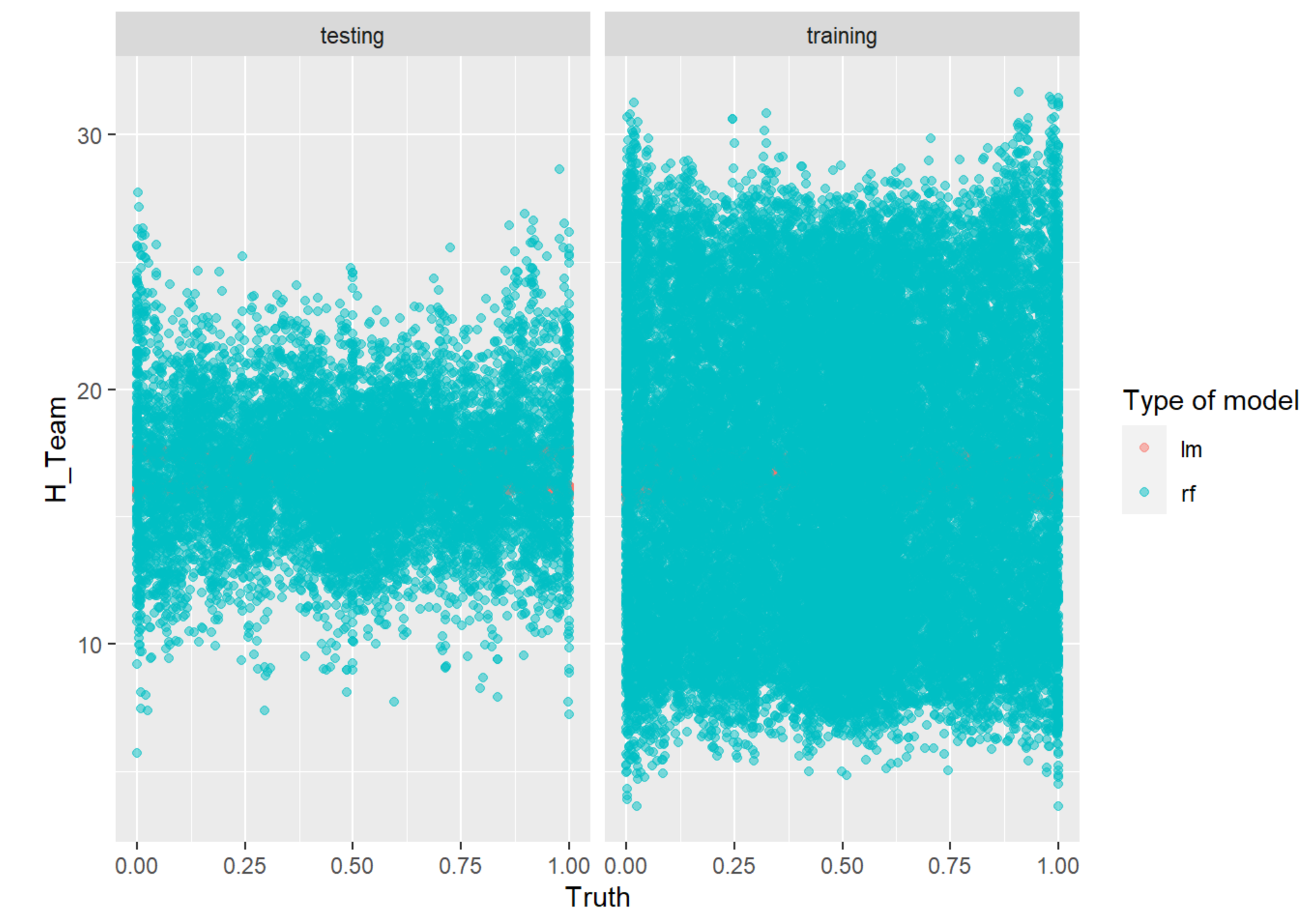
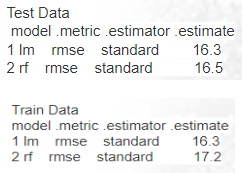
Decision tree shows Team 1, which is Arizona, has a 93% winning probability in season 2005.

**Decision Tuning**

****

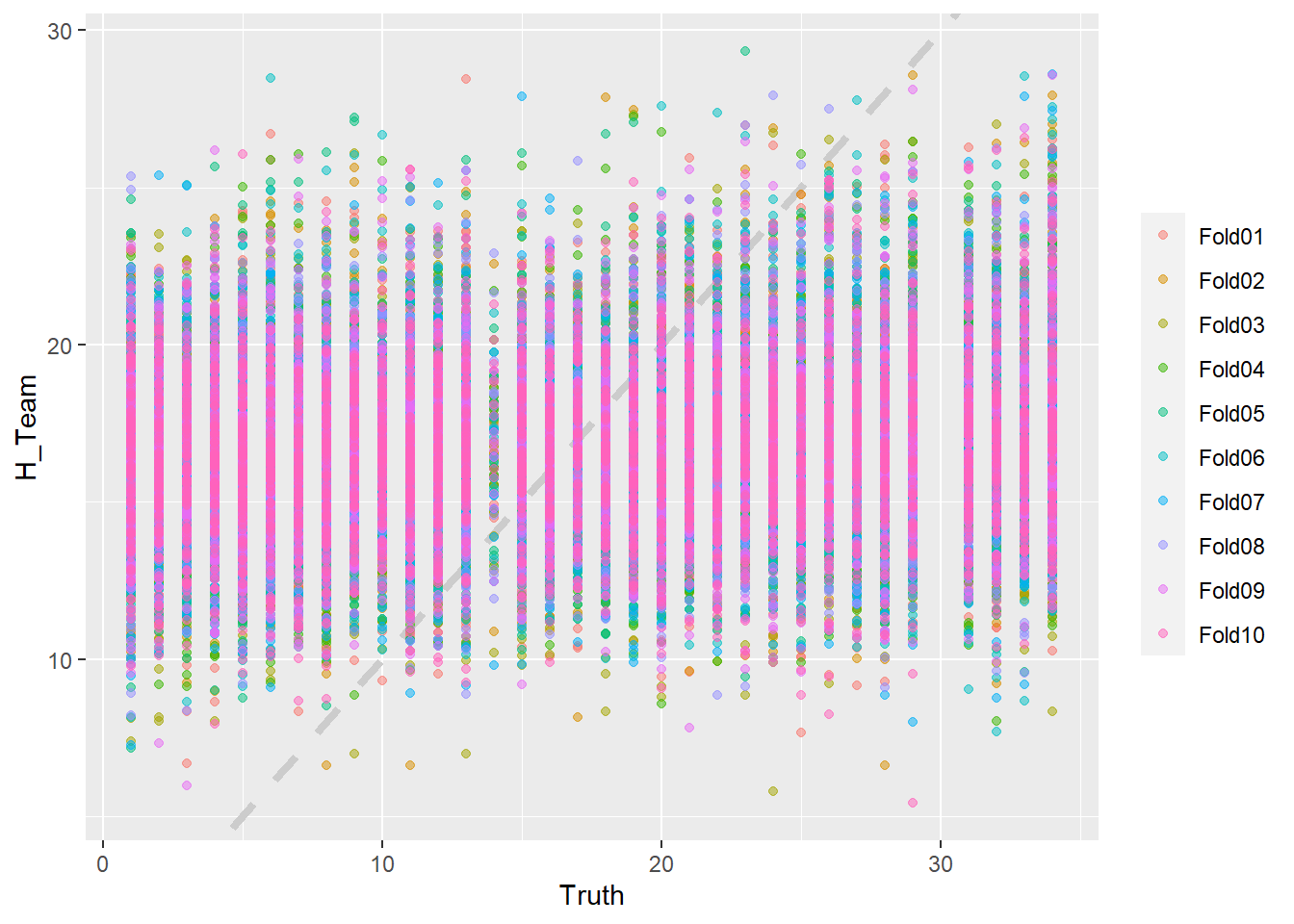
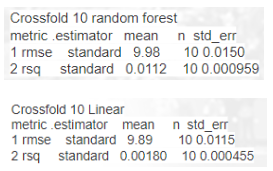
After tuning the model, the result is consistent, which also increases the winning probability of Team 1, Arizona, with a 96% winning probability in season 2005.

**Random Forest and Linear Regression**

** **

## Both models are not performing very well; it requires cross-validation to optimize the results.

**Random Forest and Linear Model fold-10 cross validation**

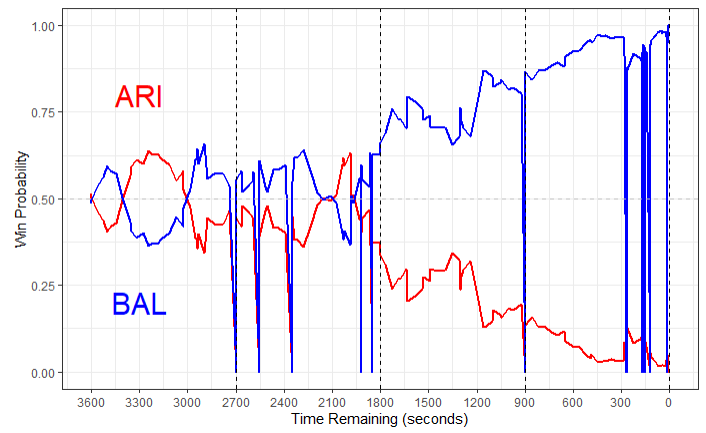
** **

## After performing fold-10 cross-validation, both models significantly improved.

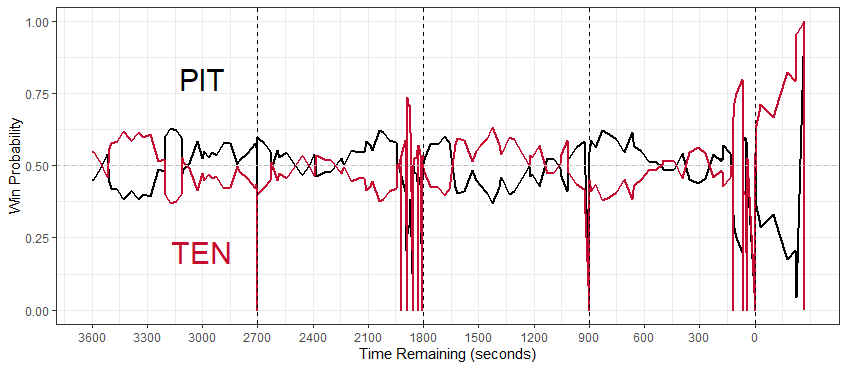
**Modeling the Probability of Winning an NFL Game**

The graph depicts the 60 minutes or 3600 seconds journey between teams by analyzing the game id, play type, play by play in second minutes, field location, various probabilities determine the game winner. The Win Probability model will try to mirror the motion of the game.NFL.com was used to valid the prediction and provide a final score.

**Team Probability Chart**

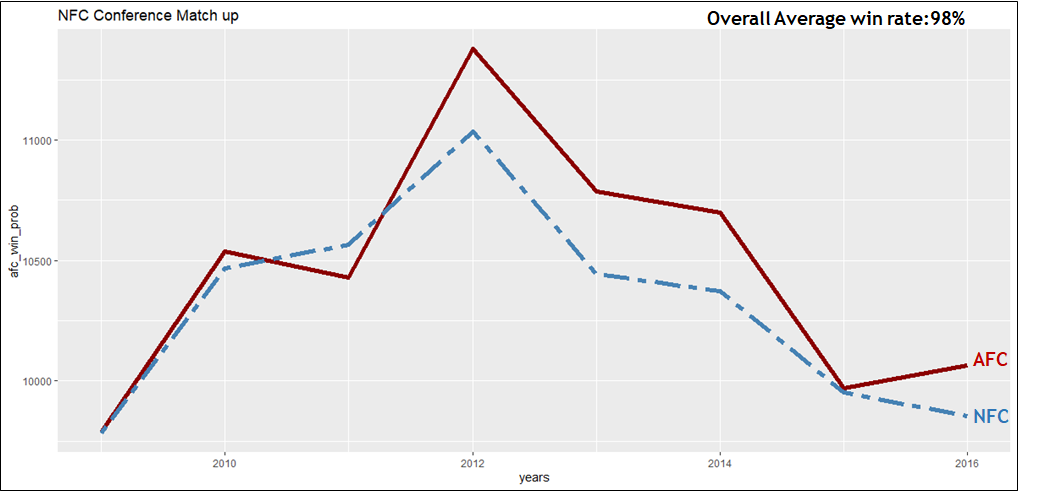
****

**Win Probability Chart**

****

**Conference Match-UP**

The NFL is divided into two conferences, NFC and AFC. However, to analyze the best conference, each team was assigned to a conference with the associated winning probability (WPA) for each NFL season. The plot depicts conference matchups from 2009 to 2016 season, as a result the AFC won more games than the NFC.

****

**CONCLUSION**

In conclusion, this project set out to use NFL play by play data to train models and build predictions of NFL team performances and individual NFL game outcomes. Requires to performed Additional analyses to compare the success of the two NFL conferences, AFC and NFC. Due to the sheer amount of data included in the play by play dataset, it was especially challenging to approach and stretched the team’s processing powers to its limits. The four models represented in this project to varying levels of success were kNN, Decision Tree, Random Forest, and Linear Model. The models put together for this project accomplished high confidence predictions, the NFL game outcome model, in particular, was able to mirror some actual NFL to a high degree.

The amount of data limited the analysis that was able to be produced. For example, another model that was considered but was ultimately scrapped was predicting the play an NFL team would run given their field position and the amount of time remaining in the game. Additionally, the amount of data and NAs meant a lot of the time spent on this project was on cleaning the data and waiting for computers to process the models. At the same time, Carnegie Mellon’s R package and resulting NFL database are very useful. It would have been extremely beneficial for this project if the NFL released an official database of their own that eliminated the NAS that was contained within Carnegie Mellon’s version of the database. These changes would have dramatically improved the team’s capacity and speed while working with the dataset.

**REFERENCES**

Linear Regression. (2019). Retrieved from https://www.geeksforgeeks.org/ml-linear-regression/

maksimhorowitz . (2019). nflscrapR. Retrieved from https://github.com/maksimhorowitz/nflscrapR

Detailed NFL Play-by-Play Data 2009-2018. (2019). Retrieved from https://www.kaggle.com /maxhorowitz/nflplaybyplay2009to2016

Understanding random Forest. (2019). Retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Introduction to Data Mining, [P. Tan](https://www.bibsonomy.org/person/14937668769a7962cb48bf7195dcad387/author/0), [M. Steinbach](https://www.bibsonomy.org/person/14937668769a7962cb48bf7195dcad387/author/1), and [V. Kumar](https://www.bibsonomy.org/person/14937668769a7962cb48bf7195dcad387/author/2). Addison Wesley, US ed edition, (May 2005)

Lantz, B. (2015). *Machine Learning with R* (Second Edition). Birmingham , UK: Packt Publishing Ltd.