Exploring the Association between Tennis Match Statistics and Match Winners

Hannah Long

2022-12-15

Introduction

Tennis is one of the most popular sports around the world. The game is very dynamic and has a lot of moving parts: serving, receiving, baseline play, and net play. Each match consists of many points, and thus a lot of data can be collected about each player's performance over the course of an entire match. The goal of this analysis is to understand which match and/or player statistics have the strongest associations with winning a match.

The dataset we are analyzing contains match statistics from 943 men's and women's matches which occurred in one of four grand slam tournaments in 2013: the Australian Open, the French Open, Wimbledon, and the US Open. Each observation is an individual match. The response variable is a binary indicator: a value of 1 indicates that player 1 won the match, and a value of 0 indicates that player 1 lost the match (the player numbers are arbitrarily assigned and correspond to which player name is listed first for each observation). The match statistics recorded for each observation are First Serve Percentage, First Serves Won, Second Serve Percentage, Second Serves Won, Aces, Double Faults, Winners, Unforced Errors, Break Points Created, Break Points Won, Net Points Attempted, Net Points Won, Total Points Won, Results for each set, and Final Number of Games Won (recorded for each player). The Australian Open and the US Open are both played on hard court, the French Open is played on clay, and Wimbledon is played on grass.

Our primary research question for this analysis is: Which tennis match statistics are most strongly associated with winning, and how does that vary by gender and/or court surface? Understanding the answer to this question will provide both spectators and players of tennis with a much deeper appreciation for the most important aspects of the game. Moreover, it can help players decide what aspects of tennis to focus on the most, both in training and in competitive matches.

Methodology

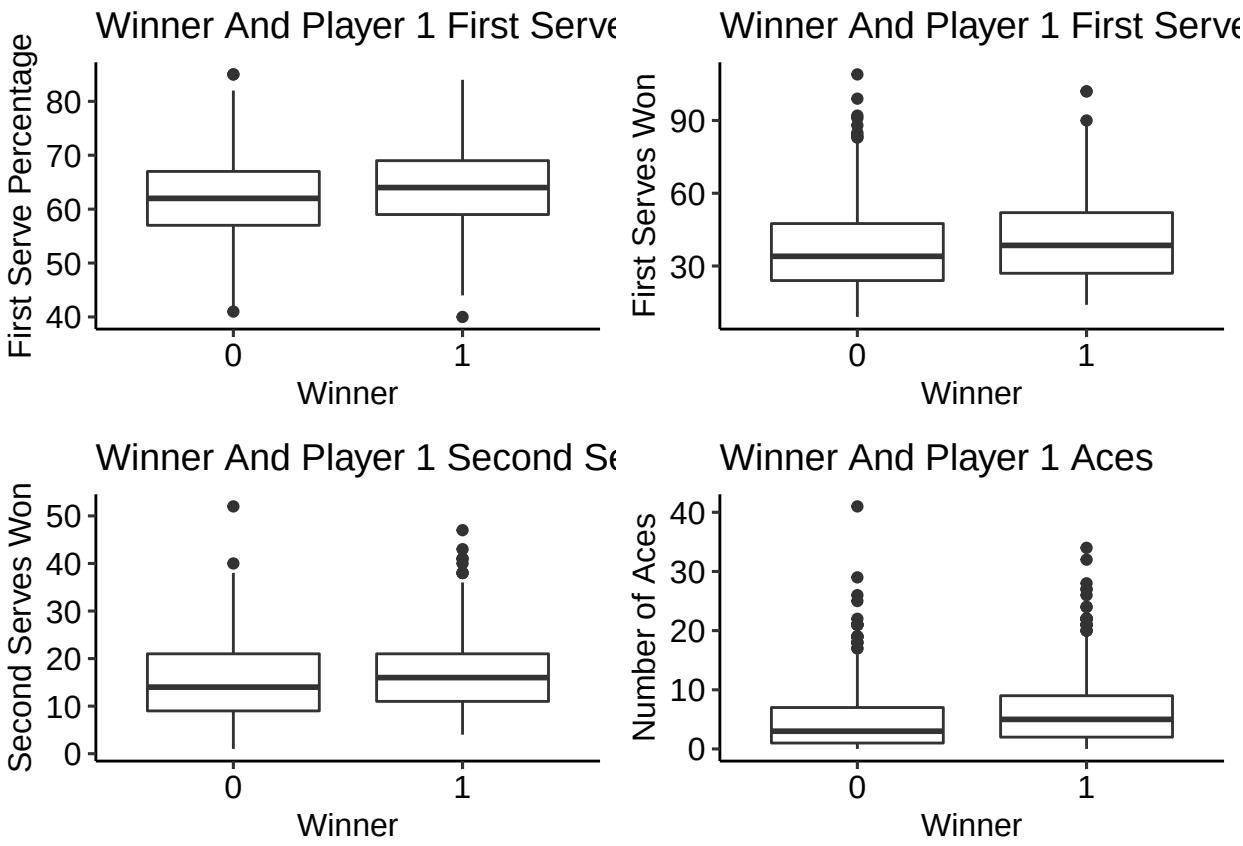
Data

After examining the data, I first selected only for relevant match statistics: player names were removed because I am not considering any background information about specific players in my analysis, and the results at the end of each set and the total games won were removed because I am only interested in purely how players' performance during the points is associated with the final winner.

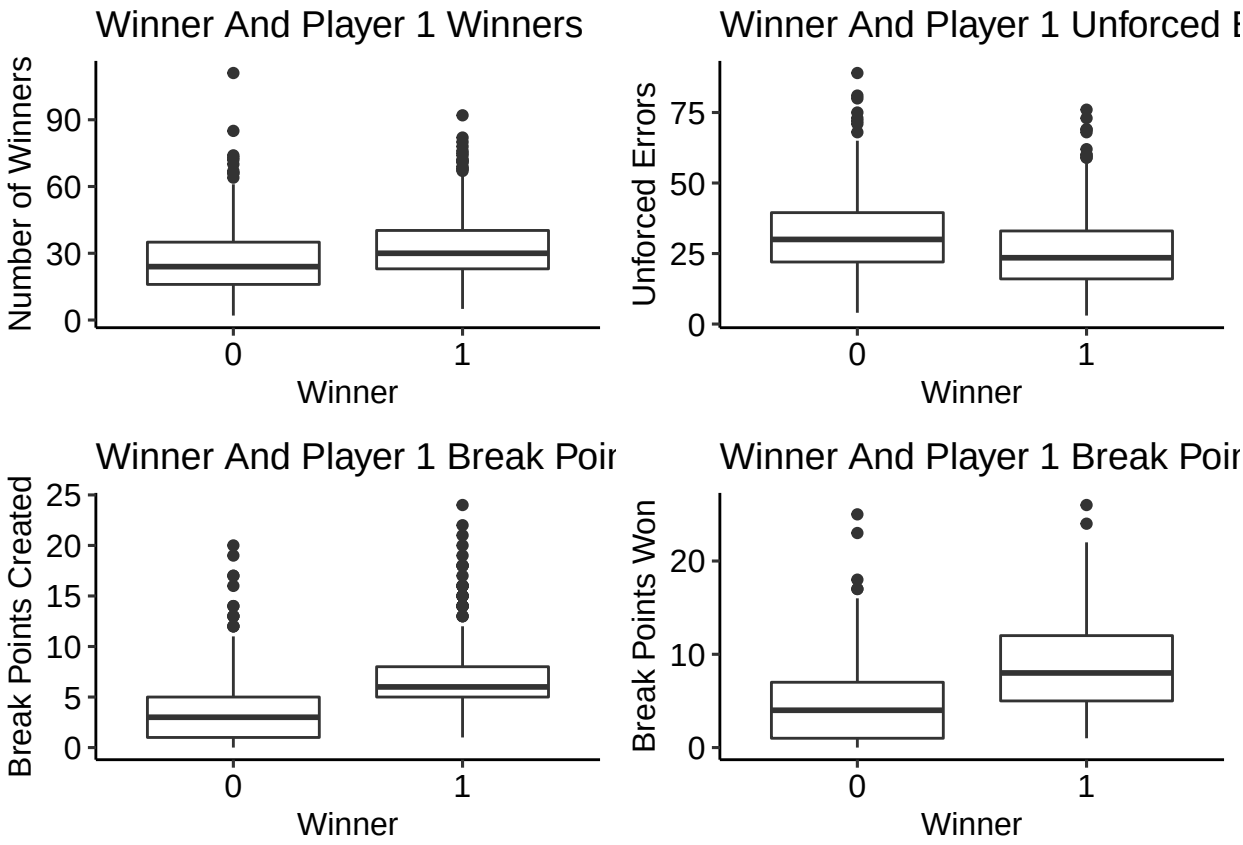
I also noticed that virtually every observation from the US Open had some missing data. Thus, I completely removed data from the US Open from my analysis and only analyzed data from the other three grand slam tournaments (one of each surface). In the remaining data, 74 out of 741 observations had some missingness with no clear pattern. Because this was a relatively small number of observations with no clear pattern of missingness, I dropped the observations with missing data from my analysis. However, it is worth noting I have not verified the assumption that the missing data is missing completely at random.

To visually investigate the associations between match statistics and match winners, I created bar plots comparing player 1 statistics when player 1 won and when player 1 lost.

$`1`


Winner And Player 1 First Serve
First Serve Percentage
80
70
60
50
40
0
1
Winner
Winner And Player 1 First Serve
First Serves Won
90
60
30
0
1
Winner
Winner And Player 1 Second Se
Second Serves Won
50
40
30
20
10
0
0
1
Winner
Winner And Player 1 Aces
Number of Aces
40
30
20
10
0
0
1
Winner

##
$`2`




```
##
## attr(,"class")
## [1] "list"      "ggarrange"
```

The box plots show that Break Points Created and Break Points Won by player 1 have the biggest difference between matches which player 1 wins and loses.

See Appendix A for boxplots of the remaining predictor variables, which I chose not to include because they did not show any differences as large as those shown above.

Modeling

With a set of predictors and a binary response variable, I started my analysis with logistic regression modeling.

Evaluating the correlation of all the predictor variables showed that Second Serve Percentage is perfectly negatively correlated with First Serve Percentage for both players (as shown in Appendix B). Therefore, I removed Second Serve Percentage entirely from my models.

I began with building a logistic regression model from the full dataset. Then, I created one logistic regression model only with data from matches between women, and another logistic regression model only with data from matches between men. Finally, I made three separate logistic regression models for each of the three different tournaments (Australian Open, French Open, and Wimbledon).

I followed the same framework of analysis to build boosted classification trees. As a result, I have built separate boosted trees for the full dataset, the data from matches between women, the data from matches between men, the data from Australian Open matches, the data from French Open matches, and the data from Wimbledon matches,

By building two different types of models on many different subsets of data, I hope to illuminate some insight

on which match statistics are most strongly associated with winning and how that might change for different genders and different court surfaces.

Results

Overall Logistic Regression Model

Table 1: Overall Logistic Regression Model

Variable	Coefficient	P-Value
Intercept	0.464	0.02
First Serve Percentage (Player 1)	-0.001	0.5
First Serve Percentage (Player 2)	0.002	0.36
First Serves Won (Player 1)	0.021	<0.01
First Serves Won (Player 2)	-0.021	<0.01
Second Serves Won (Player 1)	0.016	<0.01
Second Serves Won (Player 2)	-0.016	<0.01
Number of Aces (Player 1)	0.000	0.95
Number of Aces (Player 2)	-0.003	0.29
Number of Double Faults (Player 1)	0.001	0.84
Number of Double Faults (Player 2)	0.002	0.7
Number of Winners (Player 1)	0.003	0.07
Number of Winners (Player 2)	-0.004	0.04
Number of Unforced Errors (Player 1)	-0.005	<0.01
Number of Unforced Errors (Player 2)	0.004	0.02
Break Points Created (Player 1)	0.032	<0.01
Break Points Created (Player 2)	-0.032	<0.01
Break Points Won (Player 1)	0.024	<0.01
Break Points Won (Player 2)	-0.013	<0.01
Net Points Attempted (Player 1)	0.005	0.03
Net Points Attempted (Player 2)	-0.001	0.55
Net Points Won (Player 1)	-0.004	0.05
Net Points Won (Player 2)	0.003	0.25

The overall logistic regression model shows that Break Points Created has the strongest association with winning, followed by First Serves Won and Break Points Won.

Specifically, every one-unit increase in number of break points created by player 1 is associated with an expected increase of approximately 0.032 in the log-odds of player 1 winning the match, holding all other variables in the model constant. Similarly, every one-unit increase in number of break points created by player 2 is associated with an expected decrease of approximately 0.032 in the log-odds of player 1 winning the match, holding all other variables in the model constant.

Every one-unit increase in number of first serves won by player 1 is associated with an expected increase of approximately 0.021 in the log-odds of player 1 winning the match, holding all other variables in the model constant. Similarly, every one-unit increase in number of first serves won by player 2 is associated with an expected decrease of approximately 0.021 in the log-odds of player 1 winning the match, holding all other variables in the model constant.

Every one-unit increase in number of break points won by player 1 is associated with an expected increase of approximately 0.024 in the log-odds of player 1 winning the match, holding all other variables in the model constant. Every one-unit increase in number of break points won by player 2 is associated with an expected decrease of approximately 0.013 in the log-odds of player 1 winning the match, holding all other variables in the model constant.

Logistic Regression Model for Women

Table 2: Women Logistic Regression Model

Variable	Coefficient	P-Value
Intercept	-12.753	0.3
First Serve Percentage (Player 1)	0.064	0.66
First Serve Percentage (Player 2)	0.027	0.82
First Serves Won (Player 1)	0.720	<0.01
First Serves Won (Player 2)	-0.435	<0.01
Second Serves Won (Player 1)	0.494	0.02
Second Serves Won (Player 2)	-0.396	0.05
Number of Aces (Player 1)	0.168	0.44
Number of Aces (Player 2)	0.025	0.89
Number of Double Faults (Player 1)	-0.007	0.98
Number of Double Faults (Player 2)	0.376	0.09
Number of Winners (Player 1)	0.037	0.68
Number of Winners (Player 2)	-0.376	<0.01
Number of Unforced Errors (Player 1)	-0.281	0.01
Number of Unforced Errors (Player 2)	0.129	0.12
Break Points Created (Player 1)	0.853	<0.01
Break Points Created (Player 2)	-0.461	0.02
Break Points Won (Player 1)	0.793	<0.01
Break Points Won (Player 2)	0.048	0.77
Net Points Attempted (Player 1)	-0.026	0.85
Net Points Attempted (Player 2)	0.014	0.91
Net Points Won (Player 1)	0.015	0.92
Net Points Won (Player 2)	-0.011	0.95

The logistic regression model for women shows that Break Points Created has the strongest association with winning, followed by First Serves Won and Break Points Won. These factors are the same factors as the overall logistic regression model, but the magnitude of the coefficients is much larger. This difference in magnitude between the coefficients of the two different models is likely because the model for only women uses much less data as it is only a subset of the original data. Note that majority of the coefficients in this model are statistically insignificant.

Logistic Regression Model for Men

Table 3: Men Logistic Regression Model

Variable	Coefficient	P-Value
Intercept	-12.753	0.3
First Serve Percentage (Player 1)	0.064	0.66
First Serve Percentage (Player 2)	0.027	0.82
First Serves Won (Player 1)	0.720	<0.01
First Serves Won (Player 2)	-0.435	<0.01
Second Serves Won (Player 1)	0.494	0.02
Second Serves Won (Player 2)	-0.396	0.05
Number of Aces (Player 1)	0.168	0.44
Number of Aces (Player 2)	0.025	0.89
Number of Double Faults (Player 1)	-0.007	0.98
Number of Double Faults (Player 2)	0.376	0.09

Variable	Coefficient	P-Value
Number of Winners (Player 1)	0.037	0.68
Number of Winners (Player 2)	-0.376	<0.01
Number of Unforced Errors (Player 1)	-0.281	0.01
Number of Unforced Errors (Player 2)	0.129	0.12
Break Points Created (Player 1)	0.853	<0.01
Break Points Created (Player 2)	-0.461	0.02
Break Points Won (Player 1)	0.793	<0.01
Break Points Won (Player 2)	0.048	0.77
Net Points Attempted (Player 1)	-0.026	0.85
Net Points Attempted (Player 2)	0.014	0.91
Net Points Won (Player 1)	0.015	0.92
Net Points Won (Player 2)	-0.011	0.95

The logistic regression model for men shows that Break Points Created has the strongest association with winning, followed by First Serves Won and Break Points Won. These factors are the same factors as the overall logistic regression model, but the magnitude of the coefficients is much larger. This difference in magnitude between the coefficients of the two different models is likely because the model for only men uses much less data as it is only a subset of the original data. The magnitudes of the coefficients of the logistic model for men are relatively similar to the magnitudes of the coefficients of the logistic model for women. Thus, these models fail to show meaningful differences between men and women in the associations of match statistics and match winners. Note that majority of the coefficients in this model are statistically insignificant.

Logistic Regression Model for Australian Open

Table 4: Australian Open Logistic Regression Model

Variable	Coefficient	P-Value
Intercept	255.026	1.00
First Serve Percentage (Player 1)	0.877	1.00
First Serve Percentage (Player 2)	-6.707	1.00
First Serves Won (Player 1)	12.339	1.00
First Serves Won (Player 2)	1.738	1.00
Second Serves Won (Player 1)	15.422	1.00
Second Serves Won (Player 2)	-16.629	0.99
Number of Aces (Player 1)	-6.734	1.00
Number of Aces (Player 2)	3.323	1.00
Number of Double Faults (Player 1)	-14.678	1.00
Number of Double Faults (Player 2)	27.924	1.00
Number of Winners (Player 1)	-4.853	1.00
Number of Winners (Player 2)	-3.526	1.00
Number of Unforced Errors (Player 1)	-2.160	1.00
Number of Unforced Errors (Player 2)	-7.375	1.00
Break Points Created (Player 1)	143.264	0.99
Break Points Created (Player 2)	-52.284	1.00
Break Points Won (Player 1)	-14.732	1.00
Break Points Won (Player 2)	-9.857	1.00
Net Points Attempted (Player 1)	31.414	1.00
Net Points Attempted (Player 2)	4.164	1.00
Net Points Won (Player 1)	-20.884	1.00
Net Points Won (Player 2)	-2.120	1.00

The logistic regression model for the Australian Open shows that Break Points Created has the strongest association with winning, followed by Number of Double Faults and Second Serves Won. While the first strongest factor is the same as the overall logistic regression model, the following two are not. In addition, the magnitude of the coefficients is much larger. This difference in magnitude between the coefficients of the two different models is likely because the model for only the Australian Open uses much less data than any of the above models as it is only a subset of the original data. Many of the coefficients go against intuition (for example, the model shows that player 1 winning net points is associated with a decrease in the log-odds of player 1 winning on average and holding all other variables in the model constant). Note that majority of the coefficients in this model are statistically insignificant. Considering all the above, the very small sample size for only Australian Open data severely reduces the accuracy of interpretations based on this logistic regression model.

Logistic Regression Model for French Open

Table 5: French Open Logistic Regression Model

Variable	Coefficient	P-Value
Intercept	-964.171	1.00
First Serve Percentage (Player 1)	3.703	1.00
First Serve Percentage (Player 2)	9.695	1.00
First Serves Won (Player 1)	0.625	1.00
First Serves Won (Player 2)	4.050	1.00
Second Serves Won (Player 1)	-1.639	1.00
Second Serves Won (Player 2)	10.612	1.00
Number of Aces (Player 1)	-6.179	1.00
Number of Aces (Player 2)	-21.651	1.00
Number of Double Faults (Player 1)	-0.292	1.00
Number of Double Faults (Player 2)	18.488	1.00
Number of Winners (Player 1)	-1.338	1.00
Number of Winners (Player 2)	-3.655	1.00
Number of Unforced Errors (Player 1)	-2.401	1.00
Number of Unforced Errors (Player 2)	-2.403	1.00
Break Points Created (Player 1)	183.061	0.99
Break Points Created (Player 2)	-156.871	0.99
Break Points Won (Player 1)	-20.890	1.00
Break Points Won (Player 2)	21.511	1.00
Net Points Attempted (Player 1)	-10.422	1.00
Net Points Attempted (Player 2)	-23.948	1.00
Net Points Won (Player 1)	2.543	1.00
Net Points Won (Player 2)	23.241	1.00

The logistic regression model for the French Open shows that Break Points Created has the strongest association with winning, followed by Number of Double Faults. These factors are the same as the factors most strongly associated with winning in the Australian Open model. However, like the Australian Open model, the magnitudes of the coefficients of this model are much larger than those of the overall logistical model. This difference is likely because the model for only the French Open uses much less data than any of the above models as it is only a subset of the original data. The rest of the coefficients go against intuition (for example, the model shows that player 1 winning net points is associated with a decrease in the log-odds of player 1 winning on average and holding all other variables in the model constant). Note that all of the coefficients in this model are statistically insignificant. Considering all the above, the very small sample size for only French Open data severely reduces the accuracy of interpretations based on this logistic regression model.

Logistic Regression Model for Wimbledon

Table 6: Wimbledon Logistic Regression Model

Variable	Coefficient	P-Value
Intercept	-0.628	0.96
First Serve Percentage (Player 1)	-0.118	0.59
First Serve Percentage (Player 2)	0.109	0.63
First Serves Won (Player 1)	0.289	0.08
First Serves Won (Player 2)	-0.322	0.05
Second Serves Won (Player 1)	-0.005	0.98
Second Serves Won (Player 2)	-0.107	0.67
Number of Aces (Player 1)	-0.198	0.32
Number of Aces (Player 2)	0.191	0.36
Number of Double Faults (Player 1)	0.303	0.46
Number of Double Faults (Player 2)	0.380	0.36
Number of Winners (Player 1)	0.085	0.58
Number of Winners (Player 2)	0.062	0.60
Number of Unforced Errors (Player 1)	-0.103	0.64
Number of Unforced Errors (Player 2)	0.029	0.79
Break Points Created (Player 1)	-0.289	0.39
Break Points Created (Player 2)	0.293	0.40
Break Points Won (Player 1)	4.010	0.01
Break Points Won (Player 2)	-3.746	0.01
Net Points Attempted (Player 1)	0.104	0.63
Net Points Attempted (Player 2)	0.136	0.49
Net Points Won (Player 1)	-0.109	0.74
Net Points Won (Player 2)	-0.303	0.35

The logistic regression model for Wimbledon shows that Break Points Won has the strongest association with winning, followed by First Serves Won. These factors are different from the factors most strongly associated with winning in the previous models. However, like the Australian Open and French Open models, the magnitudes of the coefficients of this model are somewhat larger than those of the overall logistical model. This difference is likely because the model for only Wimbledon uses much less data than any of the above models as it is only a subset of the original data. The rest of the coefficients go against intuition (for example, the model shows that player 1 winning net points is associated with a decrease in the log-odds of player 1 winning on average and holding all other variables in the model constant). Note that almost all of the coefficients in this model are statistically insignificant (except Break Points Won). Considering all the above, the very small sample size for only Wimbledon data severely reduces the accuracy of interpretations based on this logistic regression model.

Overall Boosted Classification Tree

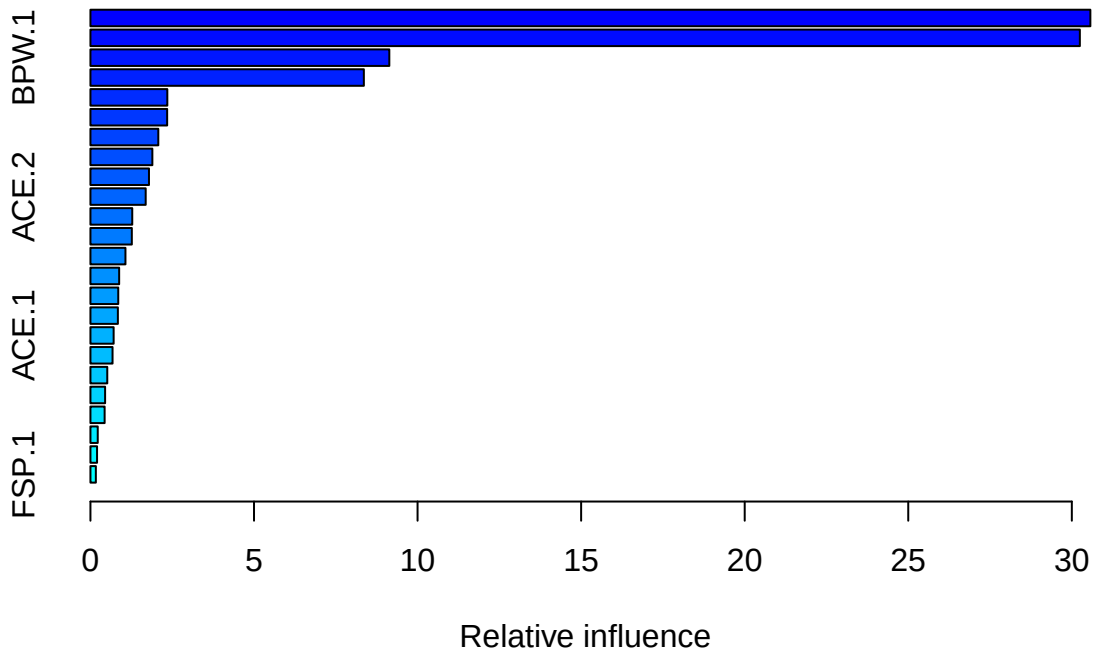


Table 7: Overall Classification Tree Relative Variable Influence

Variable	Relative Influence
Break Points Created (Player 2)	30.567
Break Points Created (Player 1)	30.245
Break Points Won (Player 1)	9.137
Break Points Won (Player 2)	8.358
Number of Unforced Errors (Player 2)	2.350
First Serves Won (Player 1)	2.345
Number of Unforced Errors (Player 1)	2.075
First Serves Won (Player 2)	1.893
Number of Winners (Player 2)	1.789
Number of Aces (Player 2)	1.689
Second Serves Won (Player 2)	1.278
Second Serves Won (Player 1)	1.265
Number of Winners (Player 1)	1.070
Net Points Attempted (Player 1)	0.878
Number of Double Faults (Player 2)	0.850
Number of Double Faults (Player 1)	0.839
Number of Aces (Player 1)	0.709
Net Points Won (Player 2)	0.676
First Serve Percentage (Player 2)	0.514
Net Points Won (Player 1)	0.448
Net Points Attempted (Player 2)	0.434

Variable	Relative Influence
NA	0.223
NA	0.202
First Serve Percentage (Player 1)	0.166

The overall boosted classification tree shows that Break Points Created has a much greater relative influence in the tree than any of the other predictor variables. This evaluation is in agreement with the insight derived from the overall logistic regression model.

Boosted Classification Tree for Women

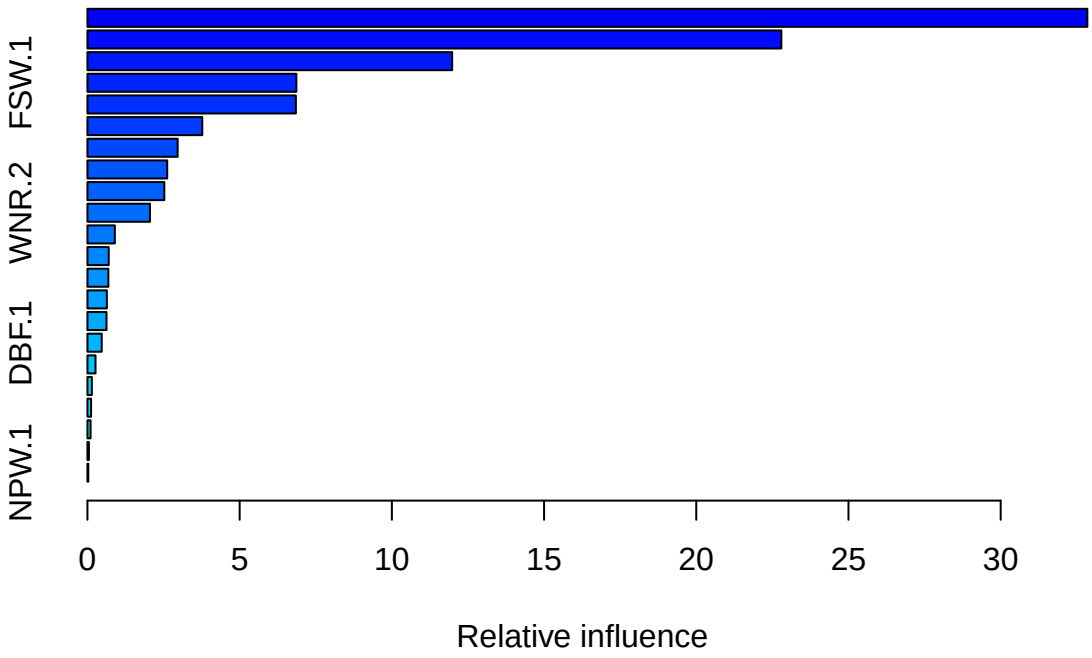


Table 8: Women Classification Tree Relative Variable Influence

Variable	Relative Influence
Break Points Created (Player 1)	32.847
Break Points Created (Player 2)	22.795
Break Points Won (Player 1)	11.981
First Serves Won (Player 1)	6.863
Break Points Won (Player 2)	6.847
Second Serves Won (Player 1)	3.772
First Serves Won (Player 2)	2.963
Number of Unforced Errors (Player 1)	2.620
Number of Unforced Errors (Player 2)	2.525
Number of Winners (Player 2)	2.054

Variable	Relative Influence
Number of Aces (Player 2)	0.900
First Serve Percentage (Player 1)	0.699
Second Serves Won (Player 2)	0.688
Number of Winners (Player 1)	0.638
Number of Aces (Player 1)	0.627
Number of Double Faults (Player 1)	0.469
Net Points Attempted (Player 2)	0.265
Net Points Won (Player 2)	0.145
Net Points Attempted (Player 1)	0.120
First Serve Percentage (Player 2)	0.104
Number of Double Faults (Player 2)	0.049
Net Points Won (Player 1)	0.028

The boosted classification tree for women shows that Break Points Created has a much greater relative influence in the tree than any of the other predictor variables. This evaluation is in agreement with the insight derived from the logistic regression model for women.

Boosted Classification Tree for Men

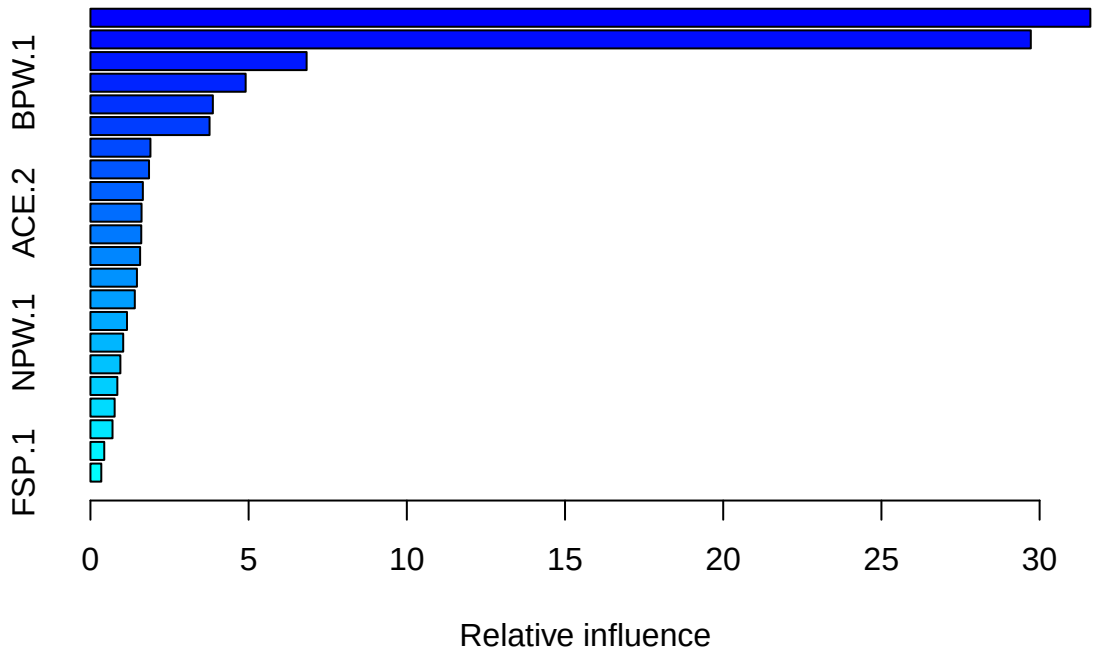


Table 9: Men Classification Tree Relative Variable Influence

Variable	Relative Influence
Break Points Created (Player 1)	31.604

Variable	Relative Influence
Break Points Created (Player 2)	29.722
Break Points Won (Player 2)	6.833
Break Points Won (Player 1)	4.906
First Serves Won (Player 1)	3.869
Number of Winners (Player 2)	3.765
Number of Unforced Errors (Player 1)	1.896
Second Serves Won (Player 2)	1.852
Net Points Won (Player 2)	1.657
Number of Aces (Player 2)	1.615
Net Points Attempted (Player 2)	1.607
First Serves Won (Player 2)	1.571
Number of Winners (Player 1)	1.470
Number of Unforced Errors (Player 2)	1.403
Net Points Attempted (Player 1)	1.155
Net Points Won (Player 1)	1.036
Number of Double Faults (Player 1)	0.947
Number of Double Faults (Player 2)	0.851
Second Serves Won (Player 1)	0.765
Number of Aces (Player 1)	0.698
First Serve Percentage (Player 2)	0.437
First Serve Percentage (Player 1)	0.343

The boosted classification tree for men shows that Break Points Created has a much greater relative influence in the tree than any of the other predictor variables. This evaluation is in agreement with the insight derived from the logistic regression model for women.

Boosted Classification Tree for Australian Open

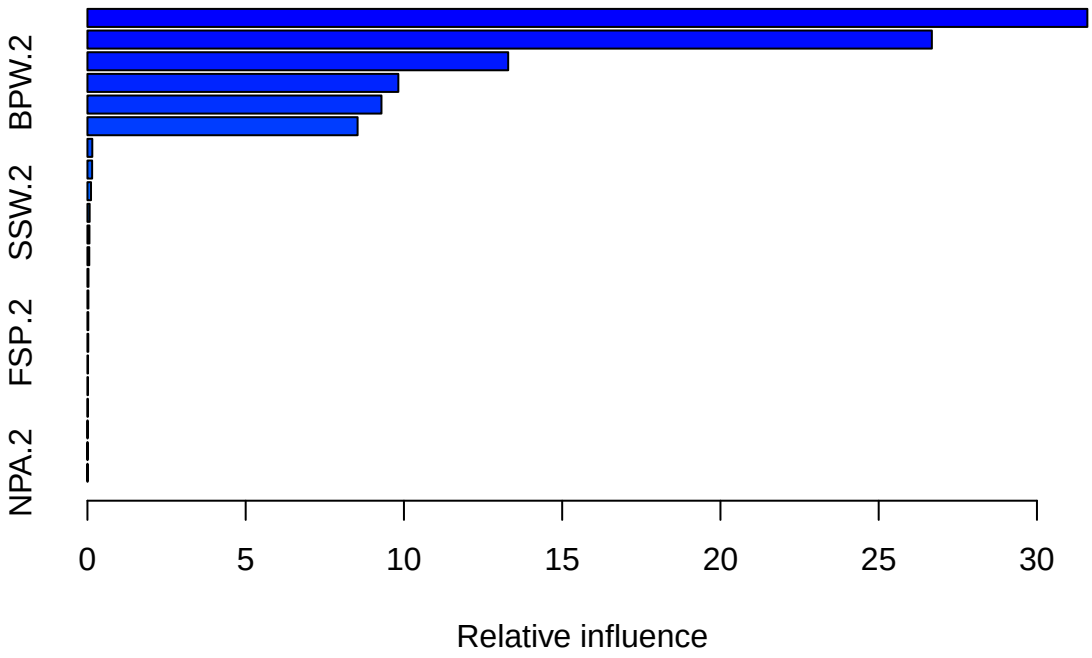


Table 10: Australian Open Classification Tree Relative Variable Influence

Variable	Relative Influence
Break Points Created (Player 1)	31.593
Number of Double Faults (Player 1)	26.679
Break Points Created (Player 2)	13.294
Break Points Won (Player 2)	9.823
Break Points Won (Player 1)	9.289
Number of Double Faults (Player 2)	8.535
Number of Unforced Errors (Player 2)	0.153
Number of Aces (Player 1)	0.148
Number of Winners (Player 1)	0.113
Second Serves Won (Player 2)	0.070
First Serves Won (Player 2)	0.063
Net Points Won (Player 1)	0.058
Number of Winners (Player 2)	0.033
Number of Aces (Player 2)	0.030
First Serves Won (Player 1)	0.027
First Serve Percentage (Player 2)	0.026
Number of Unforced Errors (Player 1)	0.019
Net Points Attempted (Player 1)	0.017
Second Serves Won (Player 1)	0.013
Net Points Won (Player 2)	0.006

Variable	Relative Influence
First Serve Percentage (Player 1)	0.004
Net Points Attempted (Player 2)	0.004

The boosted classification tree for the Australian Open shows that Break Points Created and Number of Double Faults have a much greater relative influence in the tree than any of the other predictor variables. This evaluation is in agreement with the insight derived from the logistic regression model for the Australian Open.

Boosted Classification Tree for French Open

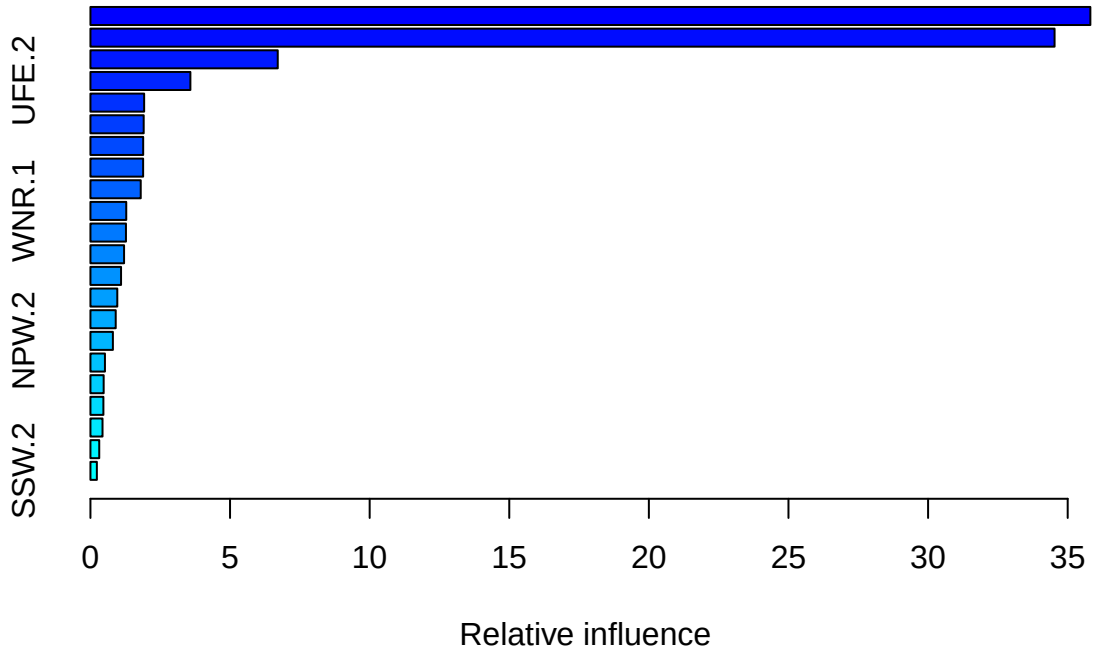


Table 11: French Open Classification Tree Relative Variable Influence

Variable	Relative Influence
Break Points Created (Player 1)	35.812
Break Points Created (Player 2)	34.530
Break Points Won (Player 1)	6.708
Number of Unforced Errors (Player 2)	3.580
Number of Double Faults (Player 2)	1.924
Number of Aces (Player 2)	1.907
First Serves Won (Player 2)	1.890
Number of Aces (Player 1)	1.887
Number of Unforced Errors (Player 1)	1.801

Variable	Relative Influence
Number of Winners (Player 1)	1.283
First Serve Percentage (Player 2)	1.273
First Serves Won (Player 1)	1.204
Number of Winners (Player 2)	1.096
Net Points Won (Player 1)	0.962
First Serve Percentage (Player 1)	0.906
Net Points Won (Player 2)	0.804
Break Points Won (Player 2)	0.519
Number of Double Faults (Player 1)	0.474
Net Points Attempted (Player 1)	0.464
Net Points Attempted (Player 2)	0.432
Second Serves Won (Player 1)	0.315
Second Serves Won (Player 2)	0.231

The boosted classification tree for the French Open shows that Break Points Created has a much greater relative influence in the tree than any of the other predictor variables. This evaluation is in agreement with the insight derived from the logistic regression model for the French Open.

Boosted Classification Tree for Wimbledon

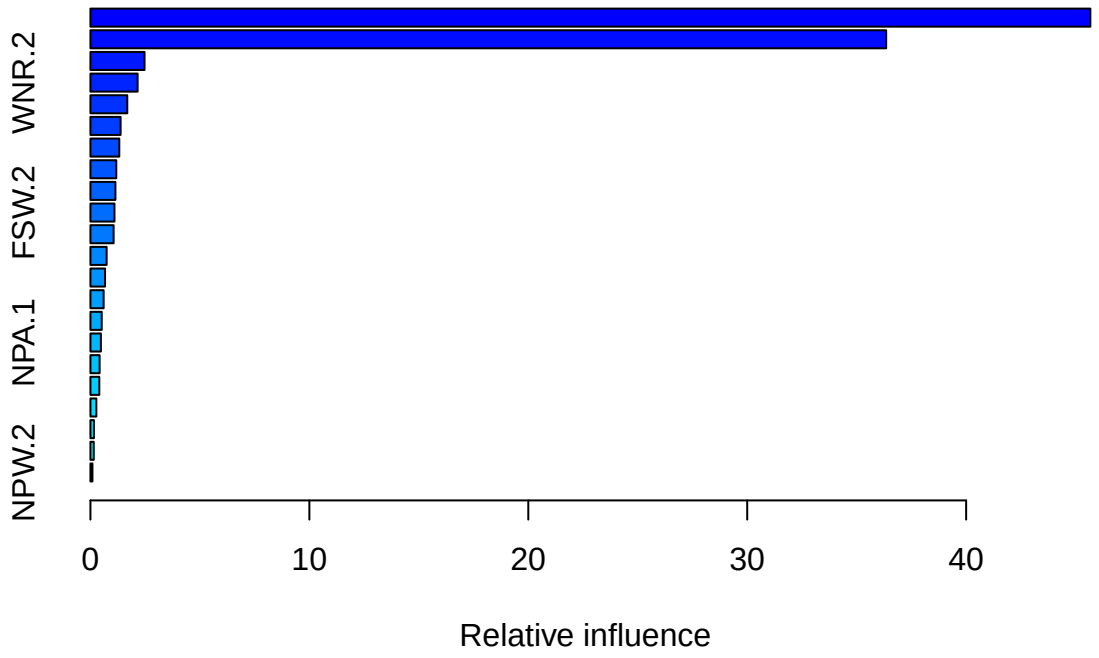


Table 12: Wimbledon Classification Tree Relative Variable Influence

Variable	Relative Influence
Break Points Won (Player 1)	45.673
Break Points Won (Player 2)	36.346
Number of Unforced Errors (Player 1)	2.471
Number of Winners (Player 2)	2.155
Number of Unforced Errors (Player 2)	1.682
Second Serves Won (Player 1)	1.377
Break Points Created (Player 1)	1.318
First Serves Won (Player 1)	1.180
Number of Double Faults (Player 1)	1.141
First Serves Won (Player 2)	1.097
Net Points Won (Player 1)	1.059
Break Points Created (Player 2)	0.740
First Serve Percentage (Player 2)	0.668
Second Serves Won (Player 2)	0.603
Net Points Attempted (Player 2)	0.515
Net Points Attempted (Player 1)	0.477
Number of Aces (Player 1)	0.420
First Serve Percentage (Player 1)	0.405
Number of Winners (Player 1)	0.270
Number of Double Faults (Player 2)	0.160
Number of Aces (Player 2)	0.151
Net Points Won (Player 2)	0.091

The boosted classification tree for Wimbledon shows that Break Points Created has a much greater relative influence in the tree than any of the other predictor variables. This evaluation is not in agreement with the insight derived from the logistic regression model for Wimbledon.

Discussion

After synthesizing the information from all the models built, it is clear that Break Points Created has the strongest association with winning a tennis match. Break points are points in a game when the receiver only needs one more point to win the game. Thus, not all points are created equal. It is not necessarily a player's performance throughout an entire match, but rather their performance during pivotal points, which is most strongly associated with winning the match. Interestingly, it is not winning break points which is most strongly associated with winning; it is simply creating break points which is most strongly associated with winning. So, creating more break points is associated with higher odds of winning.

Therefore, tennis players should focus most on setting themselves up for break points. In training, this could mean especially practicing return of serve. In competitive matches, this could mean paying extra attention to the opponent's serve and putting extra effort into the point right before a potential break point.

This analysis is inconclusive for determining how associations between match statistics and winning changes based on gender or court surface. There are not enough observations to fit accurate models on subsets of the data based on gender or tournament. If I had more time, I would like to collect a lot more data in order to fit those more specific models. From there, I could also look at subsets by both gender and surface simultaneously (ie. matches between women on clay courts).

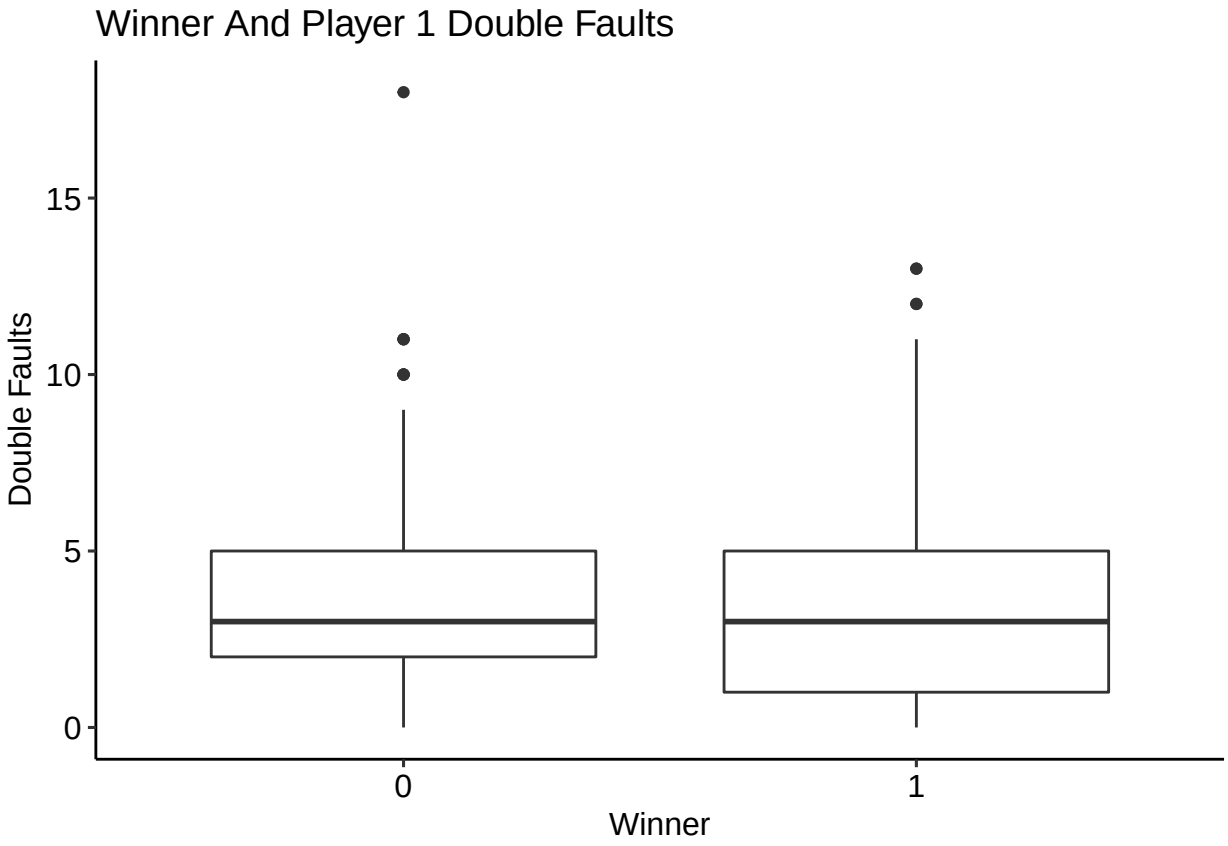
The dataset I used is from 2013, so the analysis I have conducted here is potentially outdated and might yield different results if I used more recent data. In addition, this dataset did not include background information on player demographics, skill level, ranking, etc. While I thought the exclusion of those types of variables was beneficial to my analysis, it is possible that accounting for these variables would change my results. Also,

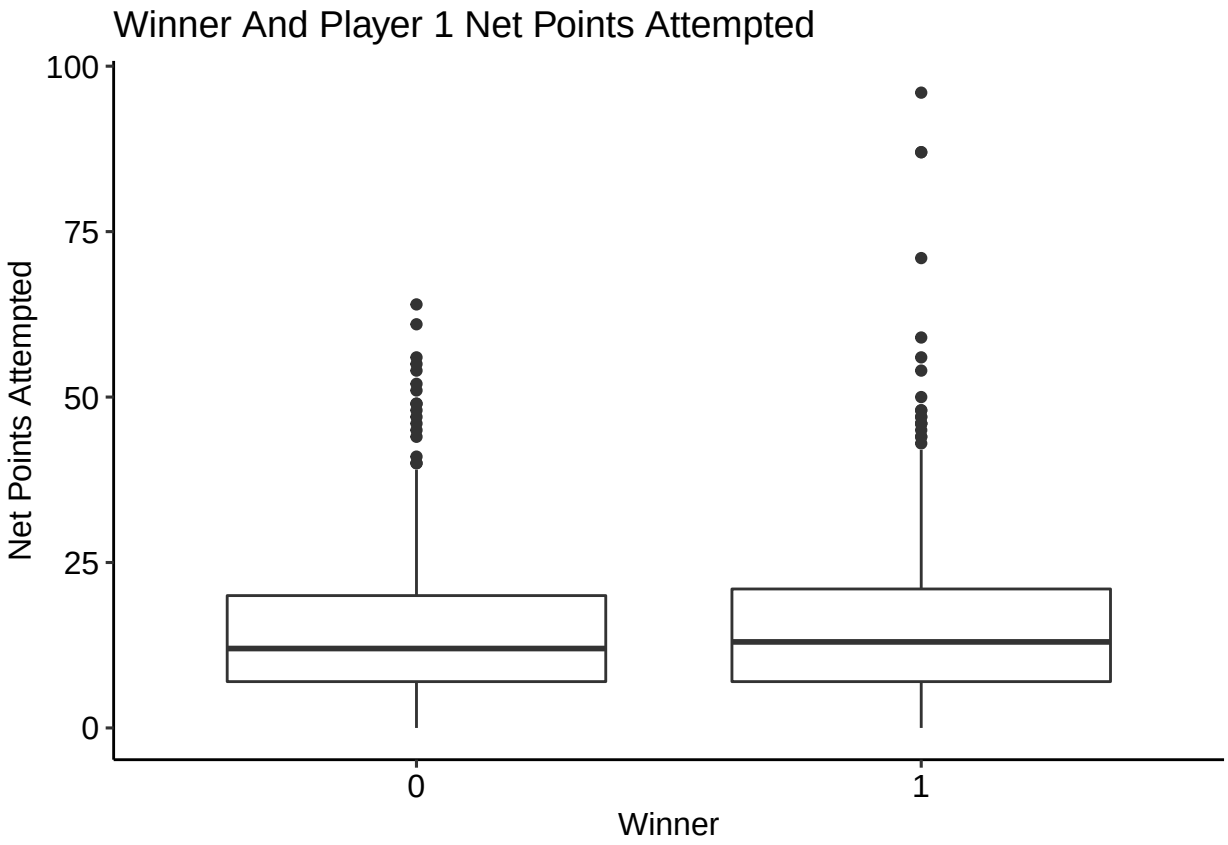
I dropped observations with missingness without verifying whether they were missing completely at random, which is a potential limitation on the accuracy of my analysis as well.

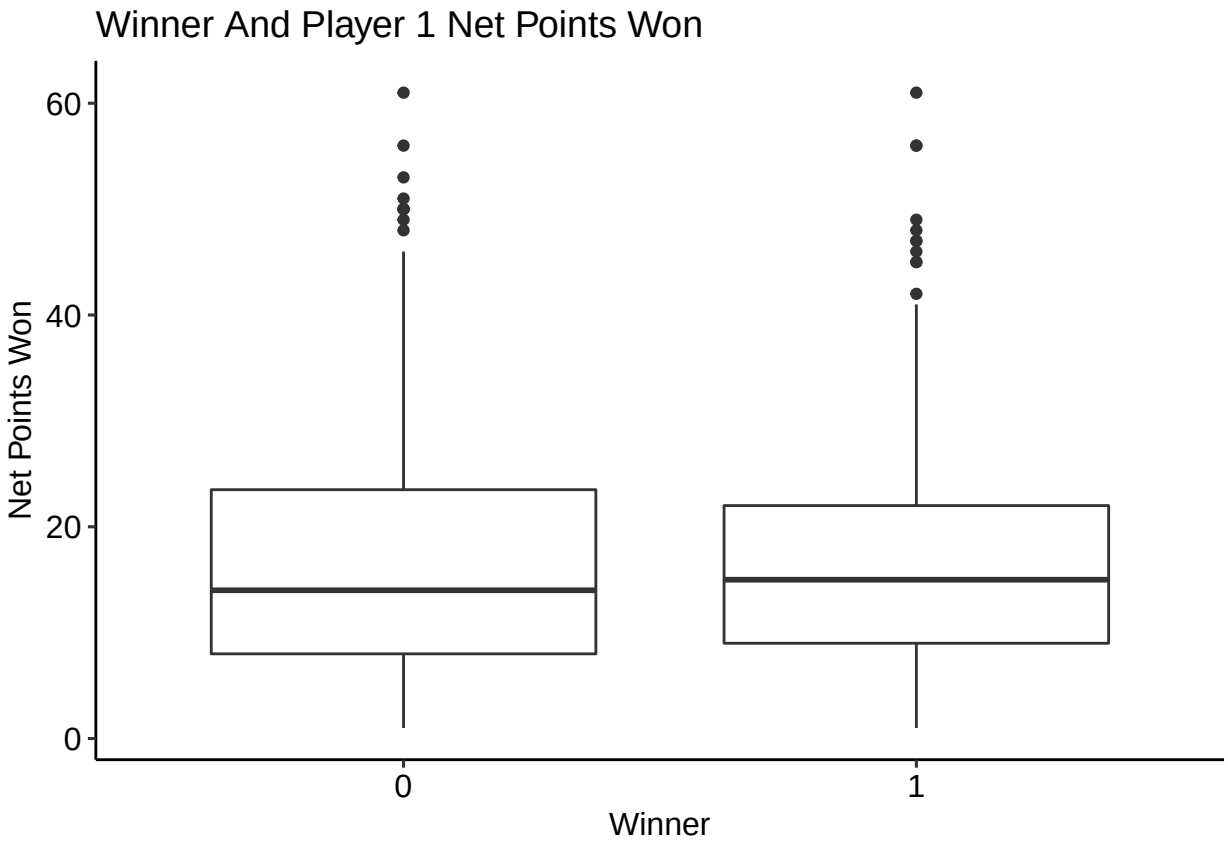
In the logistic regression modeling, note that I did not remove variables with high multicollinearity and that many of the variables were statistically insignificant.

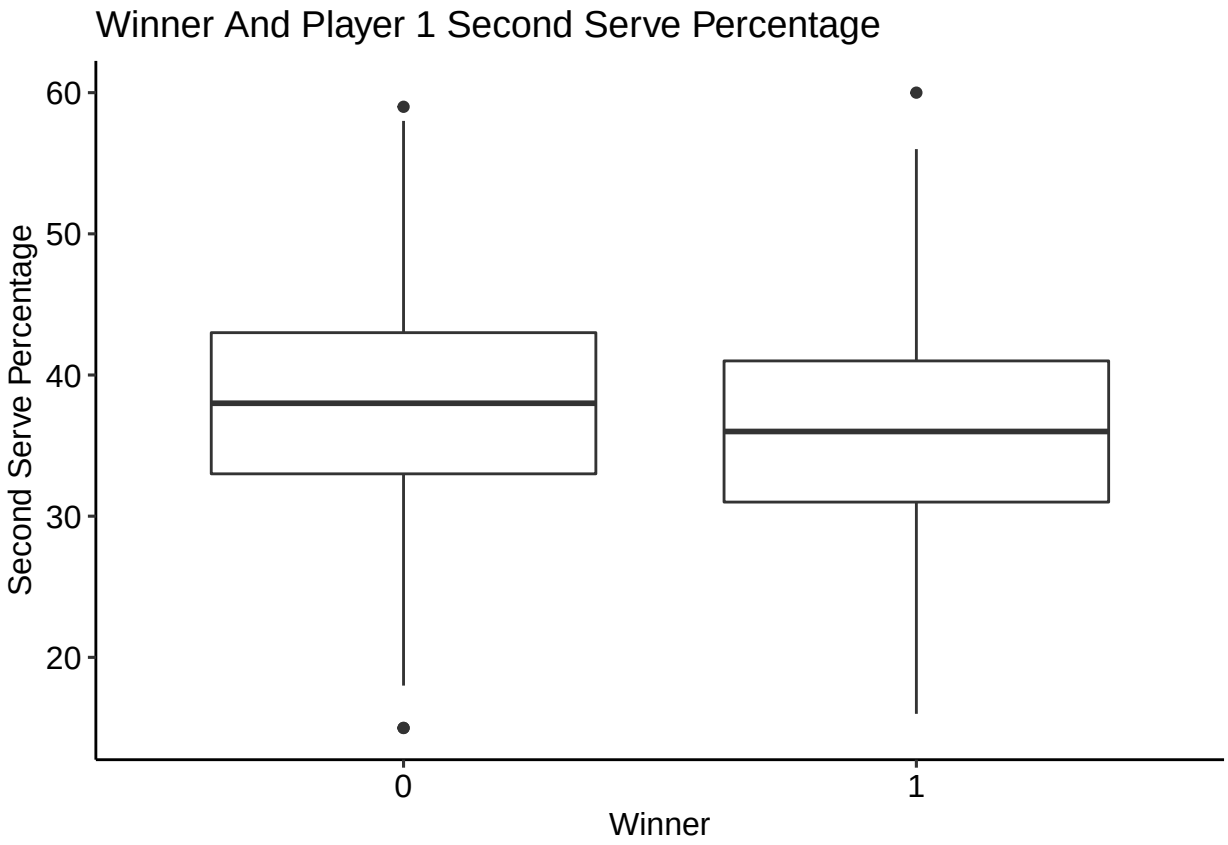
Appendix

Appendix A




Winner And Player 1 Net Points Attempted
Net Points Attempted
100
75
50
25
0
0
1
Winner


Winner And Player 1 Net Points Won
Net Points Won
60
40
20
0
0
1
Winner



Appendix B

```
##              FSP.1        FSP.2       FSW.1       FSW.2        SSP.1
## FSP.1  1.000000000  0.069236672  0.20519290  0.01732612 -1.000000000
## FSP.2  0.069236672  1.000000000  0.02647702  0.23787588 -0.069236672
## FSW.1  0.205192903  0.026477020  1.00000000  0.86550426 -0.205192903
## FSW.2  0.017326125  0.237875879  0.86550426  1.00000000 -0.017326125
## SSP.1 -1.000000000 -0.069236672 -0.20519290 -0.01732612  1.000000000
## SSP.2 -0.069236672 -1.000000000 -0.02647702 -0.23787588  0.069236672
## SSW.1 -0.358258637 -0.017989622  0.71284806  0.76677742  0.358258637
## SSW.2 -0.014847107 -0.328421436  0.77338549  0.71288529  0.014847107
## ACE.1 -0.069796998  0.007789595  0.62511821  0.56219141  0.069796998
## ACE.2 -0.007838096  0.014721986  0.57741884  0.68359337  0.007838096
## DBF.1 -0.303834348  0.029285176  0.28079521  0.30472619  0.303834348
## DBF.2 -0.087462058 -0.356327727  0.19804417  0.16709620  0.087462058
## WNR.1 -0.036118336  0.085550941  0.79674404  0.75015592  0.036118336
## WNR.2  0.065128085  0.015430079  0.73898203  0.81290964 -0.065128085
## UFE.1 -0.171009538  0.064349403  0.50234103  0.55731982  0.171009538
## UFE.2  0.013721539 -0.164998407  0.51021271  0.43922687 -0.013721539
## BPC.1  0.065141705 -0.013869657  0.23288095  0.12751960 -0.065141705
## BPC.2 -0.028616702  0.093913845  0.12175389  0.23236200  0.028616702
## BPW.1 -0.047031826 -0.124665727  0.17943862  0.08442571  0.047031826
## BPW.2 -0.118377871  0.022515349  0.13643027  0.22162957  0.118377871
## NPA.1  0.047513892  0.083971962  0.54821388  0.54062028 -0.047513892
## NPA.2  0.082835066  0.086388219  0.54896058  0.56813675 -0.082835066
## NPW.1 -0.015180338  0.053842178  0.61778334  0.60624905  0.015180338
```

```
## NPW.2  0.057183893  0.018026865  0.63850717  0.64725724 -0.057183893
##              SSP.2       SSW.1       SSW.2        ACE.1        ACE.2
## FSP.1 -0.069236672 -0.35825864 -0.01484711 -0.069796998 -0.007838096
## FSP.2 -1.000000000 -0.01798962 -0.32842144  0.007789595  0.014721986
## FSW.1 -0.026477020  0.71284806  0.77338549  0.625118209  0.577418845
## FSW.2 -0.237875879  0.76677742  0.71288529  0.562191408  0.683593370
## SSP.1  0.069236672  0.35825864  0.01484711  0.069796998  0.007838096
## SSP.2  1.000000000  0.01798962  0.32842144 -0.007789595 -0.014721986
## SSW.1  0.017989622  1.00000000  0.67262384  0.544528756  0.527116371
## SSW.2  0.328421436  0.67262384  1.00000000  0.499352980  0.521146877
## ACE.1 -0.007789595  0.54452876  0.49935298  1.000000000  0.444735528
## ACE.2 -0.014721986  0.52711637  0.52114688  0.444735528  1.000000000
## DBF.1 -0.029285176  0.35254120  0.26934857  0.208345437  0.164374713
## DBF.2  0.356327727  0.24688234  0.25369552  0.134827243  0.146705955
## WNR.1 -0.085550941  0.71979150  0.61069778  0.757264922  0.462027445
## WNR.2 -0.015430079  0.62012595  0.71093100  0.363771125  0.715952721
## UFE.1 -0.064349403  0.52327490  0.53472220  0.249984196  0.175434325
## UFE.2  0.164998407  0.51485136  0.48441792  0.137081442  0.206106819
## BPC.1  0.013869657  0.18274121  0.09809433  0.172447162  0.021665033
## BPC.2 -0.093913845  0.06589093  0.18109702 -0.088607069  0.211845597
## BPW.1  0.124665727  0.23418967  0.14164837  0.066597562 -0.132703511
## BPW.2 -0.022515349  0.18461714  0.22985206 -0.138465515 -0.006053911
## NPA.1 -0.083971962  0.44400363  0.44322360  0.424437985  0.404043426
## NPA.2 -0.086388219  0.43660236  0.45225573  0.313579394  0.428697216
## NPW.1 -0.053842178  0.55543517  0.55109280  0.410605970  0.332626518
## NPW.2 -0.018026865  0.56730001  0.58029678  0.308888526  0.411862181
##             DBF.1       DBF.2        WNR.1       WNR.2       UFE.1        UFE.2
## FSP.1 -0.30383435 -0.08746206 -0.036118336  0.06512808 -0.17100954  0.01372154
## FSP.2  0.02928518 -0.35632773  0.085550941  0.01543008  0.06434940 -0.16499841
## FSW.1  0.28079521  0.19804417  0.796744042  0.73898203  0.50234103  0.51021271
## FSW.2  0.30472619  0.16709620  0.750155916  0.81290964  0.55731982  0.43922687
## SSP.1  0.30383435  0.08746206  0.036118336 -0.06512808  0.17100954 -0.01372154
## SSP.2 -0.02928518  0.35632773 -0.085550941 -0.01543008 -0.06434940  0.16499841
## SSW.1  0.35254120  0.24688234  0.719791498  0.62012595  0.52327490  0.51485136
## SSW.2  0.26934857  0.25369552  0.610697783  0.71093100  0.53472220  0.48441792
## ACE.1  0.20834544  0.13482724  0.757264922  0.36377113  0.24998420  0.13708144
## ACE.2  0.16437471  0.14670595  0.462027445  0.71595272  0.17543432  0.20610682
## DBF.1  1.00000000  0.15343384  0.282672704  0.20103683  0.47525485  0.19096225
## DBF.2  0.15343384  1.00000000  0.119082790  0.22923944  0.16564150  0.49749179
## WNR.1  0.28267270  0.11908279  1.000000000  0.48741541  0.49624347  0.25101495
## WNR.2  0.20103683  0.22923944  0.487415407  1.00000000  0.34747388  0.50913114
## UFE.1  0.47525485  0.16564150  0.496243469  0.34747388  1.00000000  0.47163120
## UFE.2  0.19096225  0.49749179  0.251014955  0.50913114  0.47163120  1.00000000
## BPC.1  0.07490079  0.19028815  0.274979748  0.07625509 -0.07421620  0.16068369
## BPC.2  0.18562536 -0.00984233  0.005414015  0.33998218  0.17454969 -0.05404401
## BPW.1  0.10628574  0.31932085  0.254156197  0.03965773  0.32361394  0.56656054
## BPW.2  0.27278305  0.09518417  0.042757011  0.29630665  0.61687905  0.34478301
## NPA.1  0.13717369 -0.02929878  0.586876713  0.42714512  0.14152850 -0.03506010
## NPA.2  0.07628768  0.06774053  0.440125572  0.59009316  0.06004704  0.17644282
## NPW.1  0.19682970  0.02438604  0.665445378  0.48160597  0.45899576  0.18429684
## NPW.2  0.12546887  0.19403140  0.504827185  0.68988834  0.31627622  0.49203130
##             BPC.1        BPC.2       BPW.1        BPW.2       NPA.1       NPA.2
## FSP.1  0.06514170 -0.028616702 -0.04703183 -0.118377871  0.04751389  0.08283507
## FSP.2 -0.01386966  0.093913845 -0.12466573  0.022515349  0.08397196  0.08638822
```

```
## FSW.1  0.23288095  0.121753894  0.17943862  0.136430274  0.54821388  0.54896058
## FSW.2  0.12751960  0.232361998  0.08442571  0.221629573  0.54062028  0.56813675
## SSP.1 -0.06514170  0.028616702  0.04703183  0.118377871 -0.04751389 -0.08283507
## SSP.2  0.01386966 -0.093913845  0.12466573 -0.022515349 -0.08397196 -0.08638822
## SSW.1  0.18274121  0.065890931  0.23418967  0.184617143  0.44400363  0.43660236
## SSW.2  0.09809433  0.181097018  0.14164837  0.229852065  0.44322360  0.45225573
## ACE.1  0.17244716 -0.088607069  0.06659756 -0.138465515  0.42443799  0.31357939
## ACE.2  0.02166503  0.211845597 -0.13270351 -0.006053911  0.40404343  0.42869722
## DBF.1  0.07490079  0.185625357  0.10628574  0.272783046  0.13717369  0.07628768
## DBF.2  0.19028815 -0.009842330  0.31932085  0.095184166 -0.02929878  0.06774053
## WNR.1  0.27497975  0.005414015  0.25415620  0.042757011  0.58687671  0.44012557
## WNR.2  0.07625509  0.339982177  0.03965773  0.296306650  0.42714512  0.59009316
## UFE.1 -0.07421620  0.174549687  0.32361394  0.616879047  0.14152850  0.06004704
## UFE.2  0.16068369 -0.054044008  0.56656054  0.344783008 -0.03506010  0.17644282
## BPC.1  1.00000000  0.254106506  0.10615134 -0.319422025  0.42245369  0.41777464
## BPC.2  0.25410651  1.000000000 -0.32365896  0.170882959  0.34267734  0.39841732
## BPW.1  0.10615134 -0.323658962  1.00000000  0.264845802 -0.16683530 -0.16072088
## BPW.2 -0.31942202  0.170882959  0.26484580  1.000000000 -0.19312428 -0.10748119
## NPA.1  0.42245369  0.342677336 -0.16683530 -0.193124284  1.00000000  0.63332221
## NPA.2  0.41777464  0.398417315 -0.16072088 -0.107481194  0.63332221  1.00000000
## NPW.1  0.09859423  0.061294468  0.16258475  0.210302595  0.73123023  0.39429669
## NPW.2  0.11009950  0.126985395  0.23555154  0.233778847  0.37965186  0.75553637
##             NPW.1       NPW.2
## FSP.1 -0.01518034  0.05718389
## FSP.2  0.05384218  0.01802687
## FSW.1  0.61778334  0.63850717
## FSW.2  0.60624905  0.64725724
## SSP.1  0.01518034 -0.05718389
## SSP.2 -0.05384218 -0.01802687
## SSW.1  0.55543517  0.56730001
## SSW.2  0.55109280  0.58029678
## ACE.1  0.41060597  0.30888853
## ACE.2  0.33262652  0.41186218
## DBF.1  0.19682970  0.12546887
## DBF.2  0.02438604  0.19403140
## WNR.1  0.66544538  0.50482718
## WNR.2  0.48160597  0.68988834
## UFE.1  0.45899576  0.31627622
## UFE.2  0.18429684  0.49203130
## BPC.1  0.09859423  0.11009950
## BPC.2  0.06129447  0.12698540
## BPW.1  0.16258475  0.23555154
## BPW.2  0.21030259  0.23377885
## NPA.1  0.73123023  0.37965186
## NPA.2  0.39429669  0.75553637
## NPW.1  1.00000000  0.52402330
## NPW.2  0.52402330  1.00000000
```