Estimating Win Probabilities for College Football Teams Ranked in the AP Poll- Sandbox for getting graphs

Ryan Morgan

November 30, 2017

Contents

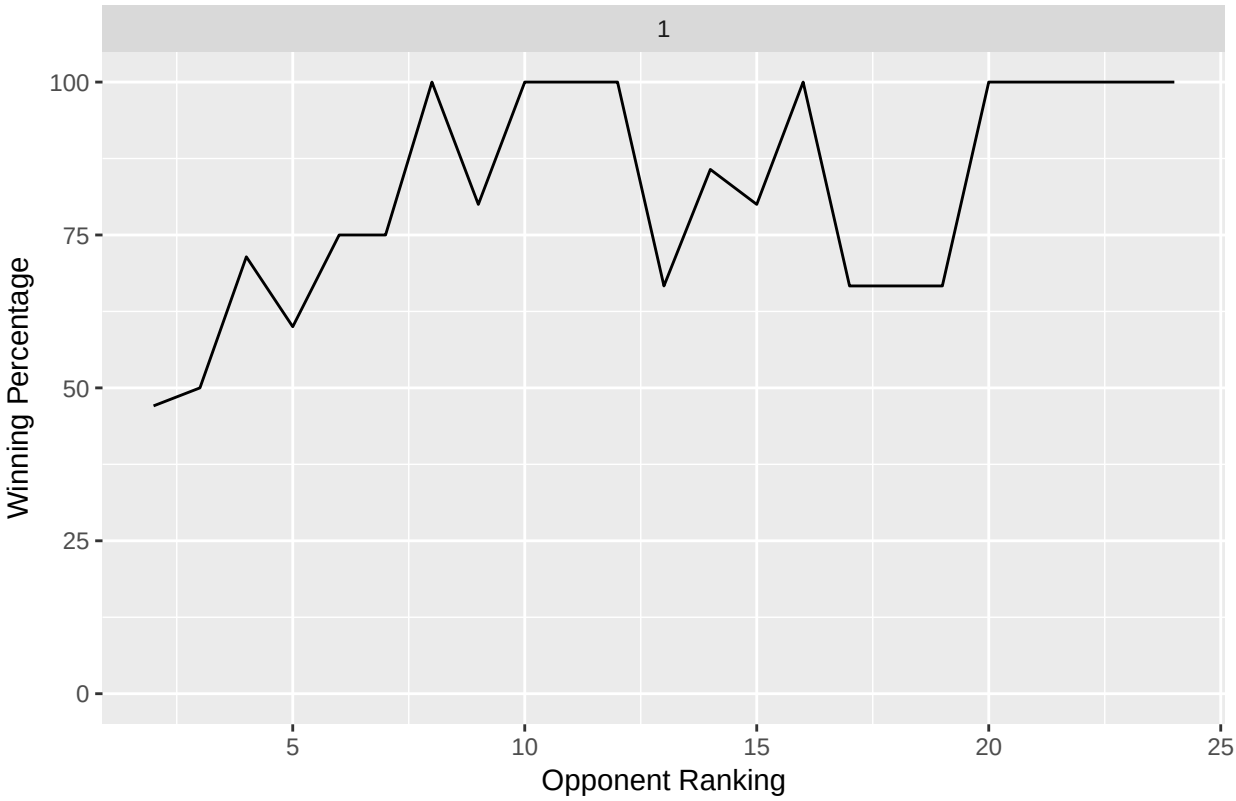


Figure 1: Winning percentage for different AP Poll matchups since 1989. The horizontal axis of the graph is the Opponent ranking, the vertical axis is the Team's winning percentage, and the facets are the Team ranking. For instance, the top left plot shows the winning percentage for teams ranked first in the AP Poll. Moving along the horizontal axis shows how the winning percentage varies as the Opponent's ranking changes.

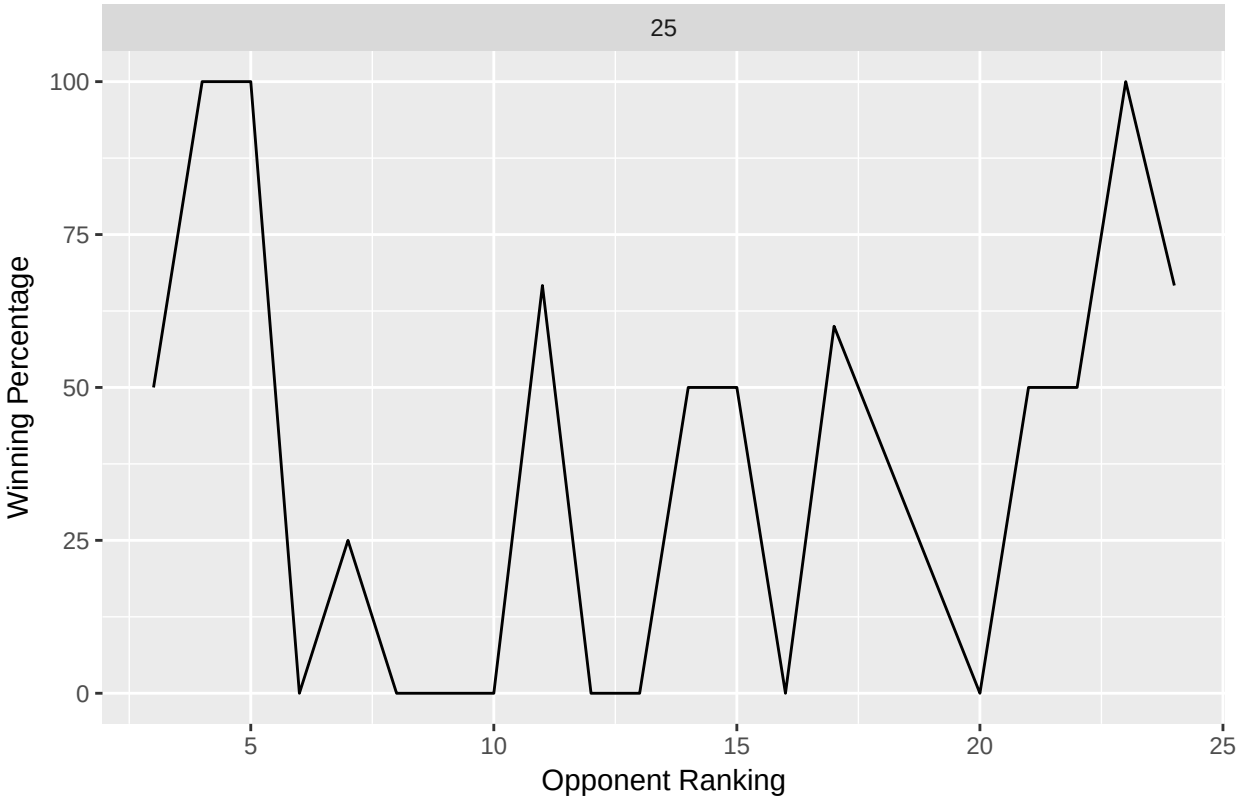


Figure 2: Winning percentage for different AP Poll matchups since 1989. The horizontal axis of the graph is the Opponent ranking, the vertical axis is the Team's winning percentage, and the facets are the Team ranking. For instance, the top left plot shows the winning percentage for teams ranked first in the AP Poll. Moving along the horizontal axis shows how the winning percentage varies as the Opponent's ranking changes.

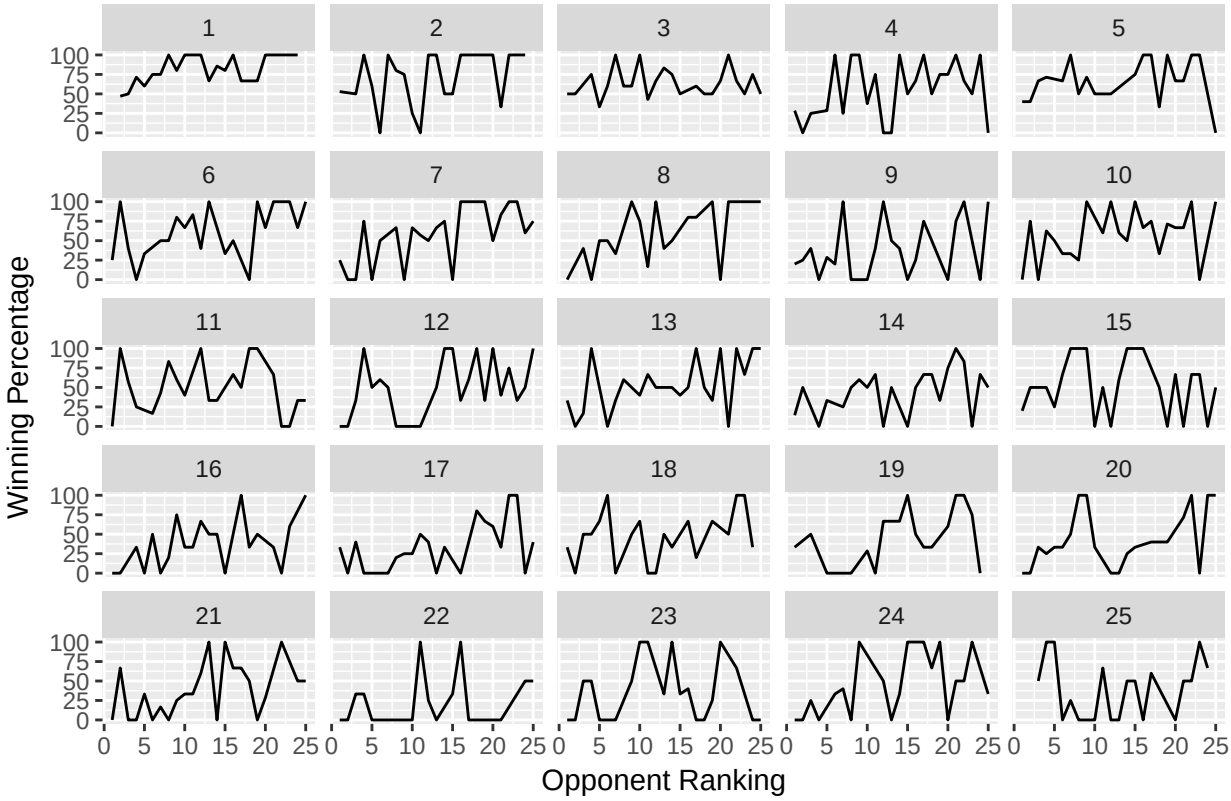
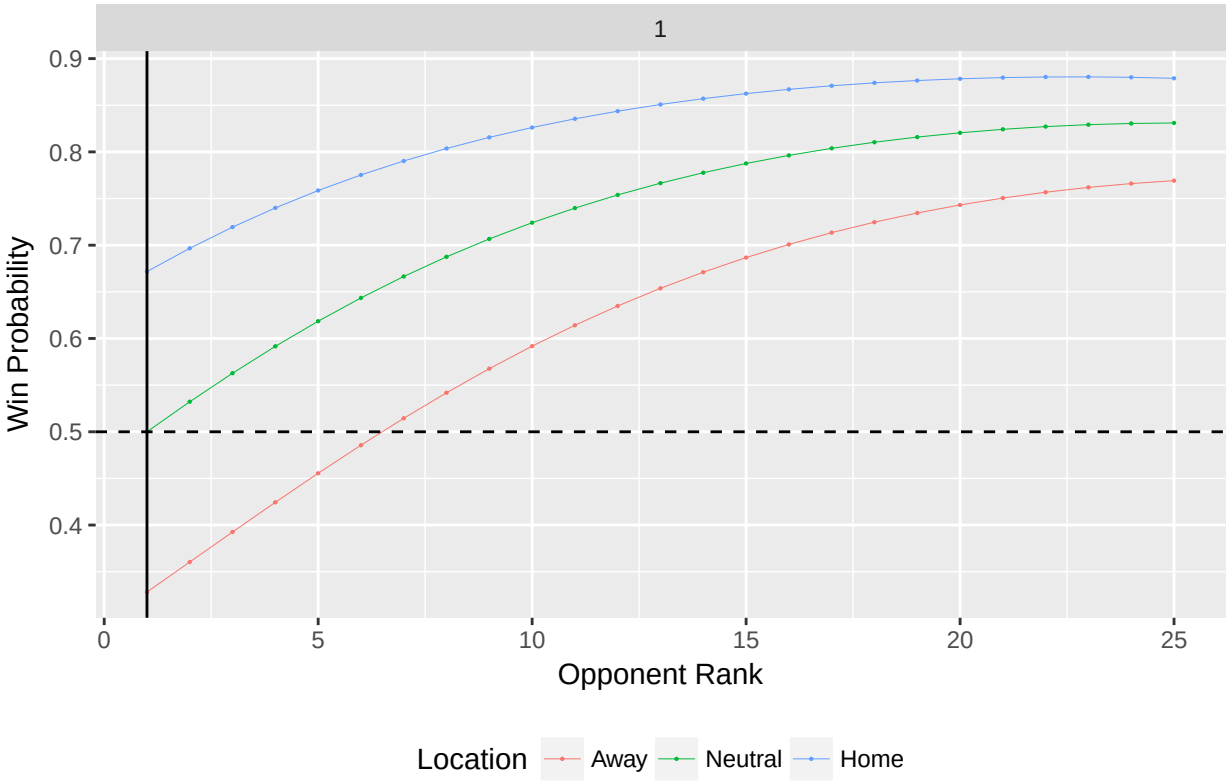
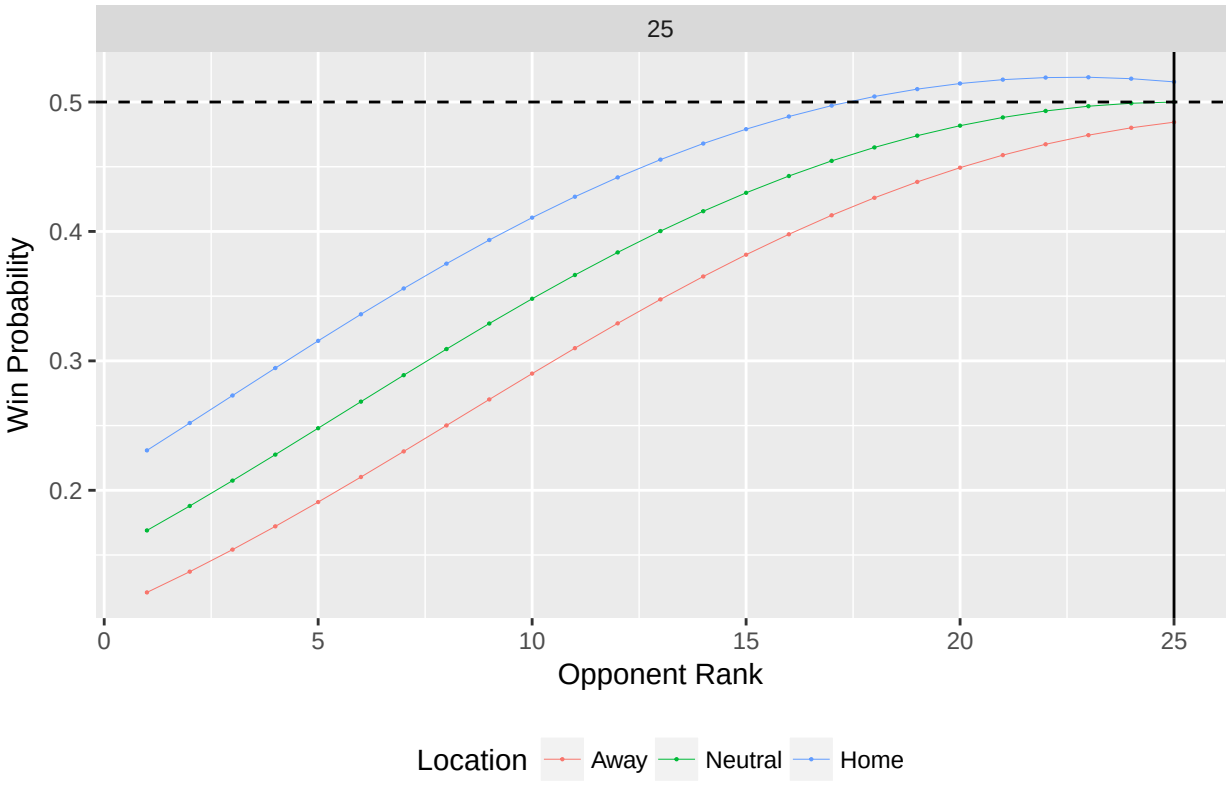
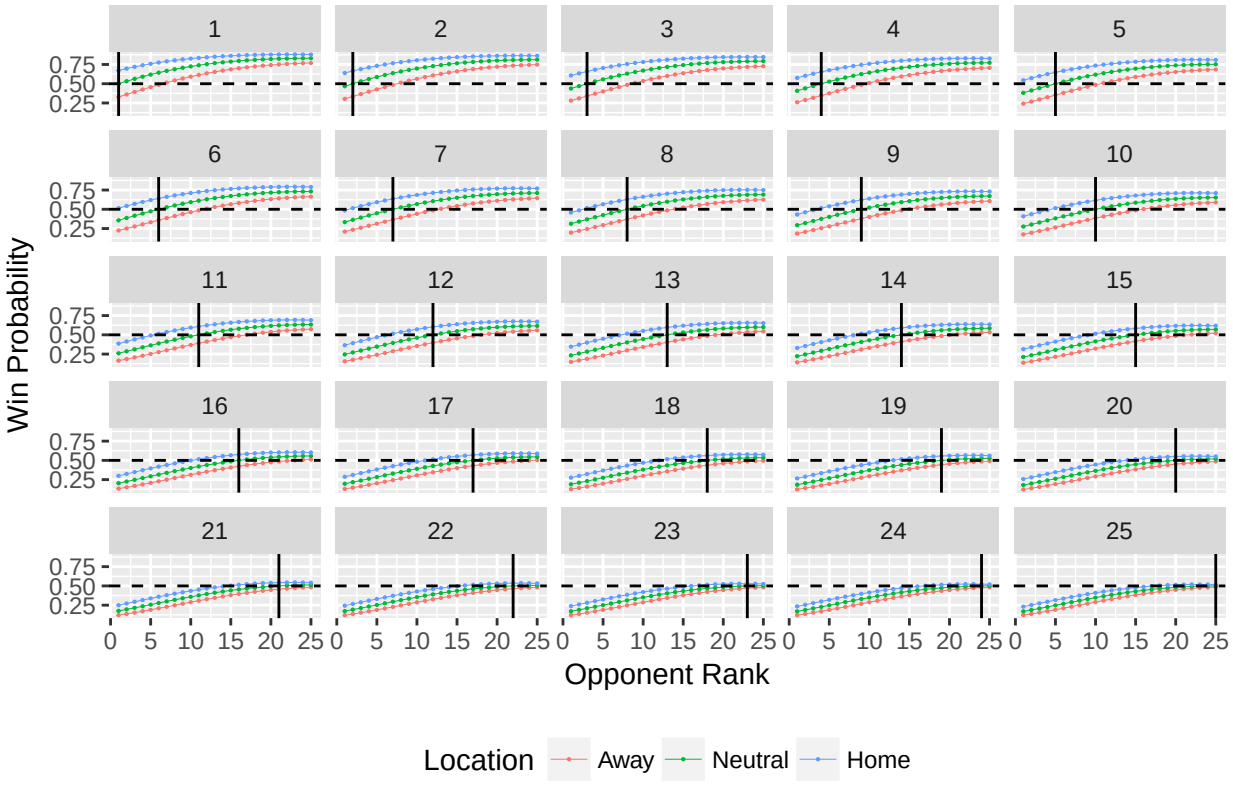


Figure 3: Winning percentage for different AP Poll matchups since 1989. The horizontal axis of the graph is the Opponent ranking, the vertical axis is the Team's winning percentage, and the facets are the Team ranking. For instance, the top left plot shows the winning percentage for teams ranked first in the AP Poll. Moving along the horizontal axis shows how the winning percentage varies as the Opponent's ranking changes.

1 Introduction




25
0.5
0.4
0.3
0.2
Win Probability
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home



2 Data Set Used

3 Generalized Linear Model Method

3.1 GLM Using Logit Link

$$Y_i = \begin{cases} 0 & \text{Team lost game } i \\ 1 & \text{Team won game } i \end{cases}$$

$X_{1,i}$: Location of Game i, -1 for away, 0 for neutral, and 1 for home

$X_{2,i}$: Difference in Team and Opponent Rank (Team minus Opponent)

$X_{3,i}$: Average of Team and Opponent Rank

$$P(Y_i = 1 | X_{1,i}, X_{2,i}, X_{3,i}) = \pi(X_{1,i}, X_{2,i}, X_{3,i})$$

$$Y_i | X_{1,i}, X_{2,i}, X_{3,i} \sim \text{Bernoulli}(\pi(X_{1,i}, X_{2,i}, X_{3,i}))$$

$$\log\left(\frac{\pi(X_{1,i}, X_{2,i}, X_{3,i})}{1 - \pi(X_{1,i}, X_{2,i}, X_{3,i})}\right) = \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i}$$

$$\pi(X_{1,i}, X_{2,i}, X_{3,i}) = \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i}])}$$

This model accounts for our "symmetry" restriction. To get the explanatory variables from the Opponent perspective, we can just take -1 times the explanatory variables from the Team perspective. This makes it so the probability a Team wins is one minus the probability the Team loses (which is the same as one minus the probability the Opponent wins). To see this, consider the following example. Suppose we want to estimate the win probability of a Team ranked 3rd playing at home against an Opponent ranked 5th. In this example, $X_1 = 1$, $X_2 = 3 - 5 = -2$, and $X_3 = \frac{3+5}{2} = 4$. Therefore, we would estimate the win probability of the Team to be

$$\pi(X_1 = 1, X_2 = -2, X_3 = 4) = \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4) = 1 - \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$= \frac{\exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}{\exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])} - \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2}$$

$$= \frac{\exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$= \frac{\exp([\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])^{-1}}{1 + \exp([\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])^{-1}}$$

$$= \frac{1}{1 + \exp([\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$= \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot -1 + \beta_2 \cdot 2 + \beta_3 \cdot -4 + \beta_4 \cdot 8 + \beta_5 \cdot 2])}$$

$$= \pi(X_1 = -1, X_2 = 2, X_3 = 4)$$

$\pi(X_1 = 1, X_2 = -2, X_3 = 4)$ gives the Team's win probability for the given game and $1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4)$ gives the *Opponent's* win probability for the given game. By constructing the model in this way, we guarantee that
$1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4) = \pi(X_1 = -1, X_2 = 2, X_3 = 4)$.

Furthermore, if two evenly ranked teams were playing each other at a neutral site, neither team should be favored. Our model allows for this, because if the Team has the same rank as the Opponent and the game was played at a neutral site, $X_1 = 0$ and $X_2 = 0$, making it so the explanatory variables and interactions in our model are equal to 0. This causes the estimated win

probability to be 0.50 because

$$\pi(X_1 = 0, X_2 = 0, X_3) = \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0 + \beta_5 \cdot 0])}$$

$$= \frac{1}{1 + \exp(0)}$$

$$= \frac{1}{2} .$$

Allowing for the "symmetry" and forcing two evenly matched teams to have win probabilities of .5 was the reasoning behind only including the location of the game, the difference in ranks, and interactions involving location or difference in ranks in our model.

The only interactions that were found to be significant were the interaction between location and average rank, the interaction between difference in rank and average rank, and the interaction between location and difference in ranks. Other variables that were explored but not found to be significant were the game number, if the game was a bowl game or not, and if the game was a conference game or not.

From the described model construction, we have five parameters to estimate corresponding to five variables constructed from three variables: X_1, X_2, and X_3. By considering all combinations of the presence or absence of the product terms $X_1 \cdot X_3$, $X_2 \cdot X_3$, and $X_1 \cdot X_2$, we arrive at the eight different sets of explanatory variables presented in Table 1. Table 2 displays the names and descriptions of the three variables used in the construction of the variables in Table 1.

Table 1: Sets of variables considered in our analyses.

Variable Set	Explanatory Variables Used	Parameters Estimated
1	$X_1, X_2, X_1 \cdot X_3, X_2 \cdot X_3, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$
2	$X_1, X_2, X_1 \cdot X_3, X_2 \cdot X_3$	$\beta_1, \beta_2, \beta_3, \beta_4$
3	$X_1, X_2, X_1 \cdot X_3, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_3, \beta_5$
4	$X_1, X_2, X_1 \cdot X_3$	$\beta_1, \beta_2, \beta_3$
5	$X_1, X_2, X_2 \cdot X_3, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_4, \beta_5$
6	$X_1, X_2, X_2 \cdot X_3$	$\beta_1, \beta_2, \beta_4$

Variable Set	Explanatory Variables Used	Parameters Estimated
7	$X_1, X_2, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_5$
8	X_1, X_2	β_1, β_2

Table 2: Variable Names and Descriptions.

Variable Name	Variable Description
X_1	**numLocation:** The location of the game from the Team perspective. -1 denotes Away, 0 denotes Neutral, 1 denotes Home
X_2	**DiffRanks:** The difference in Team Rank and Opponent Rank (Team minus Opponent)
X_3	**AvgRank:** The average of the Team Rank and Opponent Rank

Table 3: Win probabilities for a number 1 ranked Team playing at home agains an Opponent ranked 25 for six different seasons. Probabilities were found using GLMs with a logit link and each of the eight variable sets in Table 1.

	1989	1990	1991	2014	2015	2016
Set 1	0.882	0.889	0.883	0.879	0.880	0.879
Set 2	0.877	0.882	0.876	0.878	0.880	0.874
Set 3	0.885	0.891	0.887	0.882	0.882	0.881
Set 4	0.879	0.883	0.878	0.879	0.881	0.875
Set 5	0.881	0.887	0.882	0.877	0.878	0.878
Set 6	0.877	0.882	0.876	0.877	0.880	0.874
Set 7	0.885	0.891	0.887	0.882	0.882	0.881
Set 8	0.880	0.884	0.879	0.880	0.882	0.877

Table 4: Estimated win probabilities for a Team ranked 24 playing on the road against an Opponent ranked 25 for six different seasons. Probabilities were found using GLMs with a logit link. A probability was found for each variable set.

	1989	1990	1991	2014	2015	2016
Set 1	0.468	0.483	0.486	0.494	0.496	0.487
Set 2	0.469	0.484	0.487	0.494	0.496	0.488
Set 3	0.480	0.497	0.496	0.506	0.510	0.498
Set 4	0.480	0.497	0.497	0.507	0.510	0.498
Set 5	0.402	0.403	0.404	0.410	0.401	0.406
Set 6	0.402	0.403	0.405	0.410	0.401	0.406
Set 7	0.417	0.421	0.420	0.426	0.420	0.420
Set 8	0.418	0.422	0.421	0.426	0.420	0.420

To assess the GLMs, we compared AIC values and the negative log likelihood losses. For each season, variable set 2 had the lowest AIC. The AIC values for the 2012 through 2016 season are displayed in Table 5. In addition to comparing AICs, we also compared negative log likelihood losses. Because each variable set is used to construct probability estimates for each season, we can compare the estimated win probabilities to what actually happened during that season to assess which GLM performed the best. The equation used for calculating the negative log likelihood loss is $-1 \cdot \sum_{i=1}^{n}[Y_i \cdot \log(\hat{\pi}_i) + (1 - Y_i) \cdot \log(1 - \hat{\pi}_i)]$, where Y_i represents the results of game i (0: loss, 1: win) and $\hat{\pi}_i$ represents the estimated win probability for game i. Low win probabilities for losses and high win probabilities for wins make small contributions to the loss, while high win probabilities for losses and low probabilities for wins make large contributions to the loss. We calculated the negative log likelihood loss for each variable set for each season, and then summed over all seasons to get a single loss value for each variable set. The sums of the negative log likelihood losses can be found in Table 6. Variable set 2 had the lowest sum of losses.

With the lowest AIC and the lowest sum of losses, it seems that using variable set 2

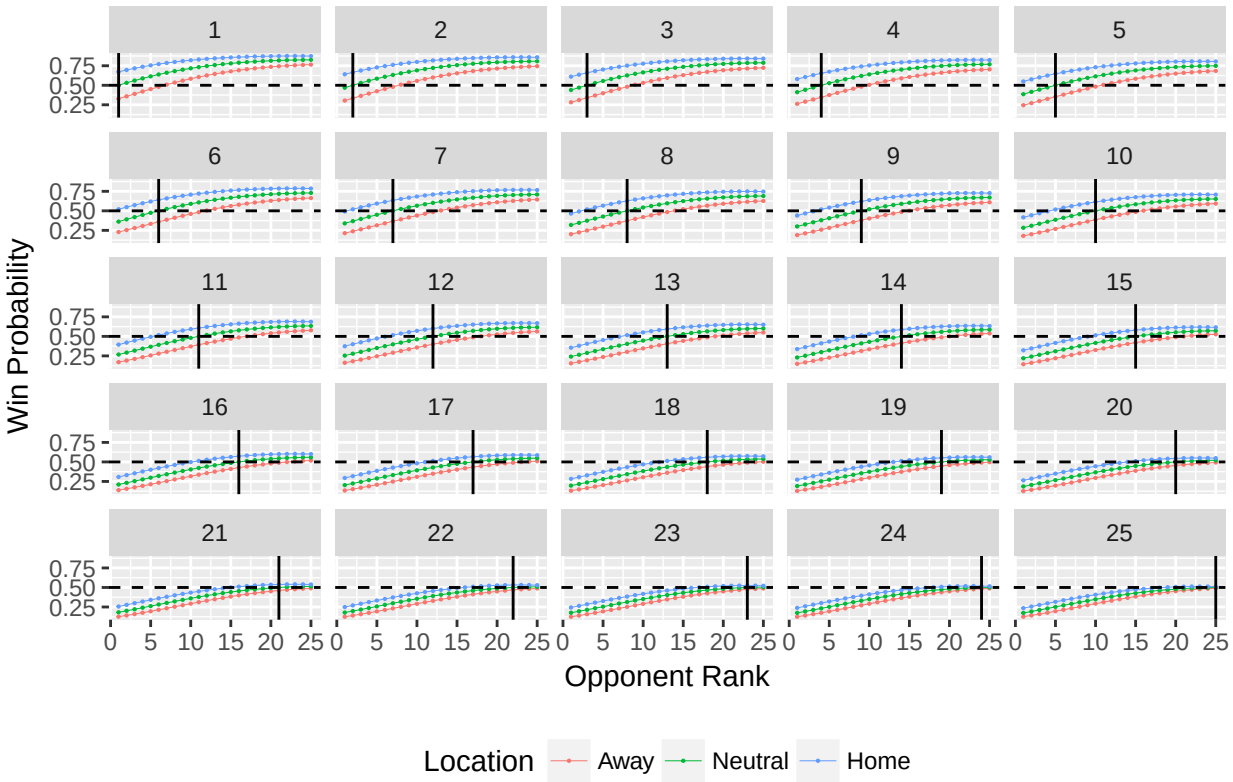


Figure 4: Estimated win probabilities for the 2016 season. Probabilities were estimated using a generalized linear model (logit link) with variable set 2. The horizontal axis of the graph is the Opponent ranking, the vertical axis is the Team's estimated win probability, and the facets are the Team ranking. The win probabilities are colored by location (from the Team's perspective). The vertical lines in each facet are where the Opponent ranking is equal to the Team ranking. This makes it easy to see if the Team is ranked higher than the Opponent. The horizontal dashed lines are at win probability equal to 0.50. This makes it easy to see if the Team is favored (a win probability higher than 0.50) or if the Opponent is favored (a Team win probability less than 0.50).

provides the best estimates of win probabilities when using a GLM with a logit link. This means the "best" model uses the game location (X_1), the difference in Team and Opponent ranks (X_2), the interaction between location and average of the Team and Opponent ranks ($X_1 \cdot X_3$), and the interaction between difference in Team and Opponent ranks and the average of the Team and Opponent ranks ($X_2 \cdot X_3$). The parameter estimates did not show a lot of variation across seasons. For the 2016 model, the estimate for β_1 is 0.746 (p-value of .0000107), the estimate for β_2 is -0.131 (p-value of .0000000296), the estimate for β_3 is -0.028 (p-value of .02129), and the estimate for β_4 is .0051 (p-value of .00368).

In context, this means as the difference in ranks increases by one (holding location and average rank constant), the odds of winning decreases by $\exp(-.131 + .0051 \cdot X_3)$. We can see this in Figure 4, where win probabilities decreases as the difference in ranks increases. From this interpretation, we can see that the difference in ranks has more of an effect when the average rank is lower. We see this in Figure 4, where an increase in difference in ranks causes the predicted win probabilities to decrease more rapidly when the average rank is 5, compared to average ranks of 10, 15, or 20. This means that the difference between a Team ranked 4 and a Team ranked 6 (difference in ranks equal to 2, average rank equal to 5) has more of an impact on win probability than the difference between a Team ranked 19 and a Team ranked 21 (difference in ranks equal to 2, average rank equal to 20).

As average rank increases (holding location and difference in ranks constant), the estimated probability of winning gets closer to 0.50. We can see this in Figure 4 by looking at the difference in line heights between the four facets. When holding the difference in ranks and location constant, the estimated win probabilities for games with higher average ranks are closer to 0.50. This tells us that when holding location and difference in ranks constant, games with higher average ranks are more difficult to predict (win probabilities closer to 0.50).

As location changes from away to neutral (meaning X_1 changes from -1 to 0) or from neutral to home (meaning X_1 changes from 0 to 1) while holding Team and Opponent rank constant, the odds of winning increases by $\exp(.746 - .028 \cdot X_3)$. This tells us that a Team's win probability is at its highest when playing at home and its lowest when playing on the road (which is to be expected). This also tells us that the home field advantage is greater for lower ranked teams.

As the average rank increases, the improved probability of winning due to location (moving the game from an away to a neutral site or from a neutral site to home) gets smaller. We see this in Figure 4, where the distance between the blue, green, and red lines is greater for the lower average ranks.

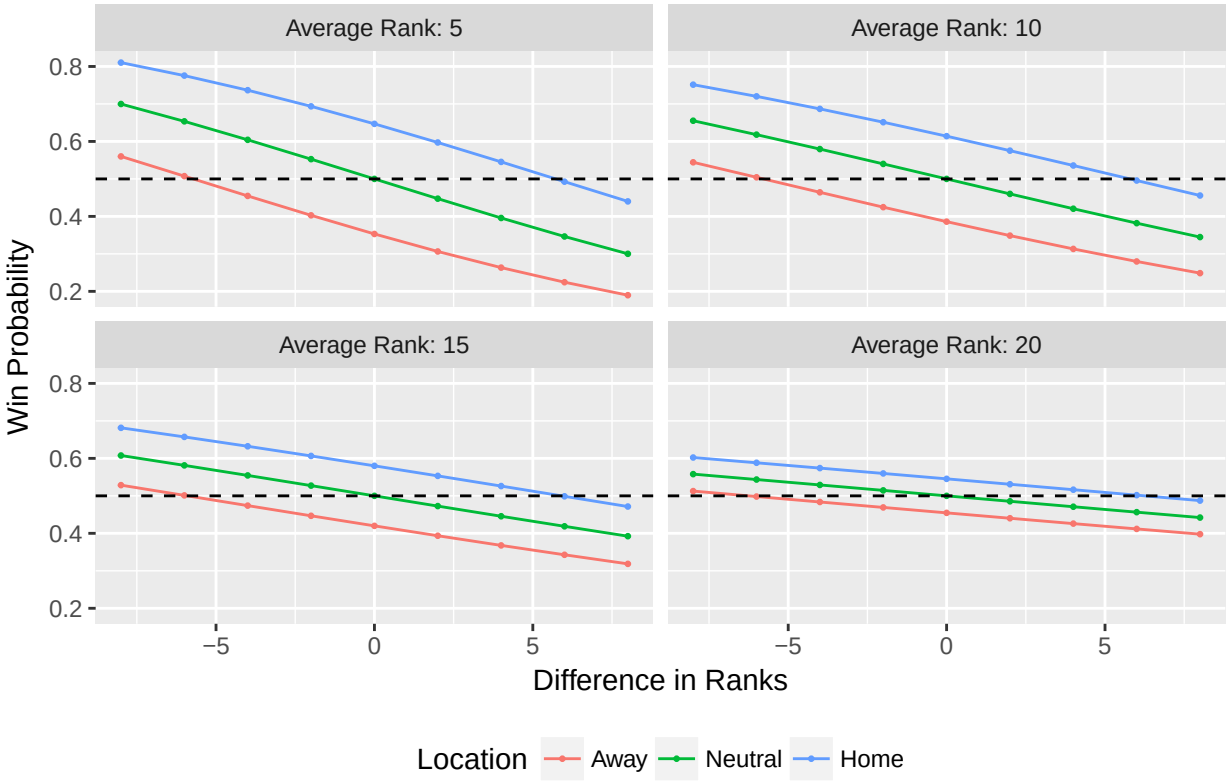


Figure 5: Estimated win probabilities for the 2016 season. Probabilities were estimated using a generalized linear model (logit link) with variable set 2. The horizontal axis of the graph is the difference in ranks (taking Team rank minus Opponent rank), the vertical axis is the Team's estimated win probability, and the facets are four different values of the average rank (values of 5, 10, 15, and 20 are shown). The win probabilities are colored by location (from the Team's perspective). The horizontal dashed lines are at win probability equal to 0.50. This makes it easy to see if the Team is favored (a win probability higher than 0.50) or if the Opponent is favored (a Team win probability less than 0.50).

Table 5: AIC values for each variable set for the 2012-2016 seasons. Variable set 2 had the lowest AIC for each season.

Variable Set	2012 AICs	2013 AICs	2014 AICs	2015 AICs	2016 AICs
Set 1	1826.751	1826.959	1815.593	1815.728	1831.611
Set 2	1824.752	1825.119	1813.600	1813.728	1829.679
Set 3	1838.469	1837.537	1823.824	1827.215	1838.071
Set 4	1836.493	1835.776	1821.867	1825.222	1836.173
Set 5	1831.511	1829.631	1819.368	1820.887	1834.970
Set 6	1829.511	1827.753	1817.368	1818.900	1833.014
Set 7	1842.071	1839.398	1826.841	1831.321	1840.755
Set 8	1840.079	1837.587	1824.855	1829.321	1838.824

Table 6: Sum of negative log likelihood losses for each parameter set, sorted by the sum of losses.

Variable Set	Sum of Losses
Set 2	947.5
Set 1	948.8
Set 6	949.4
Set 5	950.8
Set 4	951.9
Set 3	953.2
Set 8	953.4
Set 7	954.8

3.2 GLM Using Probit Link

$$Y_i = \begin{cases} 0 & \text{Team lost game } i \\ 1 & \text{Team won game } i \end{cases}$$

$X_{1,i}$: Location of Game i, -1 for away, 0 for neutral, and 1 for home

$X_{2,i}$: Difference in Team and Opponent Rank (Team minus Opponent)

$X_{3,i}$: Average of Team and Opponent Rank

$$P(Y_i = 1|X_{1,i}, X_{2,i}, X_{3,i}) = \pi(X_{1,i}, X_{2,i}, X_{3,i})$$

$$Y_i|X_{1,i}, X_{2,i}, X_{3,i} \sim \text{Bernoulli}(\pi(X_{1,i}, X_{2,i}, X_{3,i}))$$

$$\Phi^{-1}[\pi(X_{1,i}, X_{2,i}, X_{3,i})] = \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i}$$

$$\pi(X_{1,i}, X_{2,i}, X_{3,i}) = \Phi(\beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i})$$

Constructing a GLM using a probit link still retains the necessary "symmetry" property. To see this, we can return to the example from before. Suppose we want to estimate the win probability of a Team ranked 3rd playing at home against an Opponent ranked 5th. In this example, $X_1 = 1$, $X_2 = 3 - 5 = -2$, and $X_3 = \frac{3+5}{2} = 4$. Therefore, we would estimate the win probability of the Team to be

$$\pi(X_1 = 1, X_2 = -2, X_3 = 4) = \Phi(\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2) \ .$$

$\pi(X_1 = 1, X_2 = -2, X_3 = 4)$ gives the Team's win probability for the given game and $1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4)$ gives the *Opponent's* win probability for the given game. By constructing the model with a probit link instead of a logit link, we still guarantee that $1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4) = \pi(X_1 = -1, X_2 = 2, X_3 = 4)$.

A probit link also still retains the property of neither team being favored if two evenly ranked teams were playing each other at a neutral site. If a Team had the same rank as its Opponent and was playing at a neutral site, $X_1 = 0$ and $X_2 = 0$, forcing the estimated win

probability to be

$$\pi(X_1 = 0, X_2 = 0, X_3) = \Phi(\beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0 + \beta_5 \cdot 0)$$

$$= \Phi(0)$$

$$= 0.50 \ .$$

```
all_predictions %>% filter(Season==2016) %>% ggplot(aes(x=Opponent_Rank, y=Model2_GLM_Probit,
```

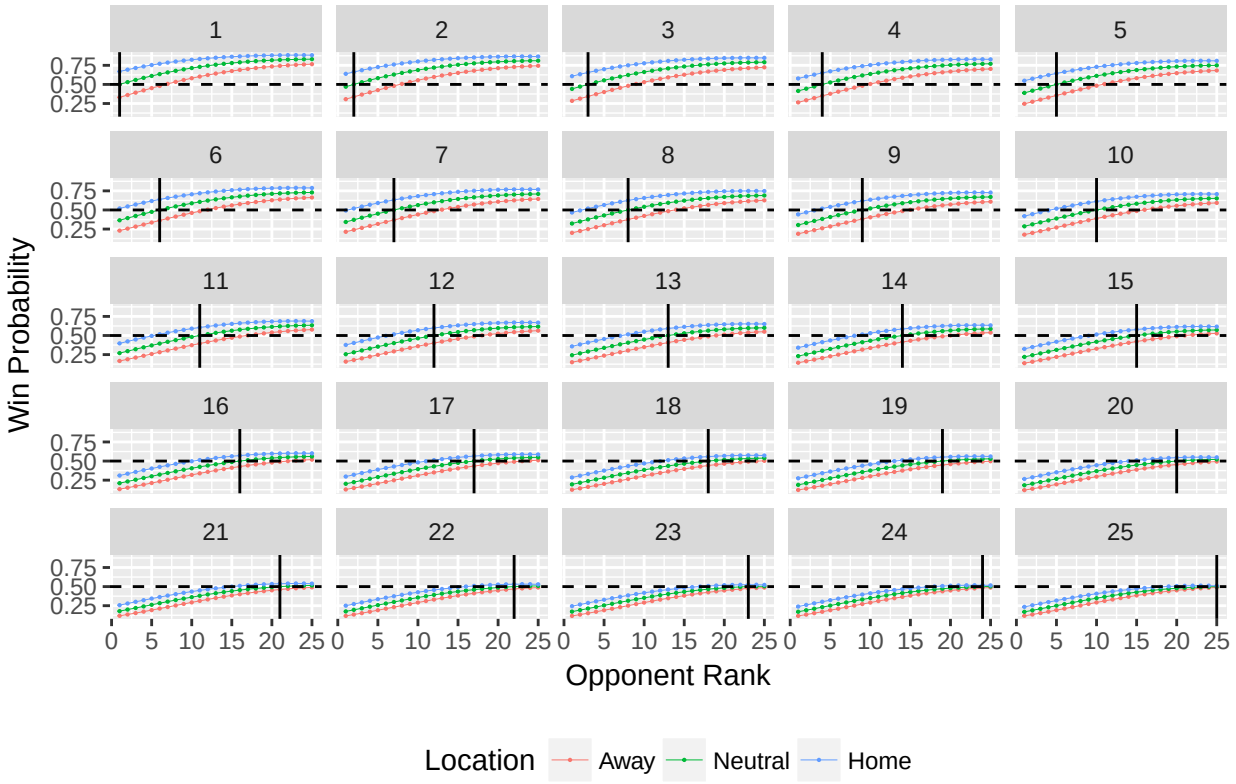


Table 7: Estimated win probabilities for a number 1 ranked Team playing at home versus an Opponent ranked 25 for six different seasons.

	1989	1990	1991	2014	2015	2016
Set 1	0.892	0.899	0.893	0.889	0.889	0.888
Set 2	0.886	0.891	0.884	0.886	0.888	0.883
Set 3	0.896	0.902	0.897	0.893	0.893	0.891
Set 4	0.888	0.893	0.887	0.889	0.891	0.885
Set 5	0.892	0.898	0.893	0.887	0.887	0.888
Set 6	0.886	0.891	0.885	0.886	0.889	0.883
Set 7	0.896	0.902	0.897	0.892	0.892	0.891
Set 8	0.889	0.894	0.888	0.889	0.892	0.886

Table 8: AIC values for each variable set for the 2012-2016 seasons.

Variable Set	2012 AICs	2013 AICs	2014 AICs	2015 AICs	2016 AICs
Set 1	1826.514	1826.866	1815.501	1815.603	1831.416
Set 2	1824.520	1825.067	1813.519	1813.604	1829.508
Set 3	1838.087	1837.157	1823.473	1826.843	1837.717
Set 4	1836.120	1835.424	1821.531	1824.856	1835.840
Set 5	1831.096	1829.352	1819.116	1820.587	1834.648
Set 6	1829.096	1827.507	1817.119	1818.591	1832.710
Set 7	1841.988	1839.218	1826.725	1831.268	1840.602
Set 8	1840.001	1837.430	1824.748	1829.268	1838.689

Table 9: Comparing AICs for variable set 2 between a GLM using a logit link and a GLM using a probit link.

Link Used	2012 AICs	2013 AICs	2014 AICs	2015 AICs	2016 AICs
Logit Link	1824.752	1825.119	1813.600	1813.728	1829.679
Probit Link	1824.520	1825.067	1813.519	1813.604	1829.508

Table 10: Comparing the sum of losses between GLMs that use a probit link and GLMs that use a logit link. Variable set 2 resulted in the lowest loss for both links.

Variable Set	Probit Link Losses	Logit Link Losses
Set 2	947.5	947.5
Set 1	948.8	948.8
Set 6	949.3	949.4
Set 5	950.7	950.8
Set 4	951.8	951.9
Set 3	953.1	953.2
Set 8	953.4	953.4
Set 7	954.8	954.8

4 Random Forest Method

$$-1 \cdot \sum_{i=1}^{n} [Y_i \cdot \log(OOB_i) + (1 - Y_i) \cdot \log(1 - OOB_i)] \ .$$

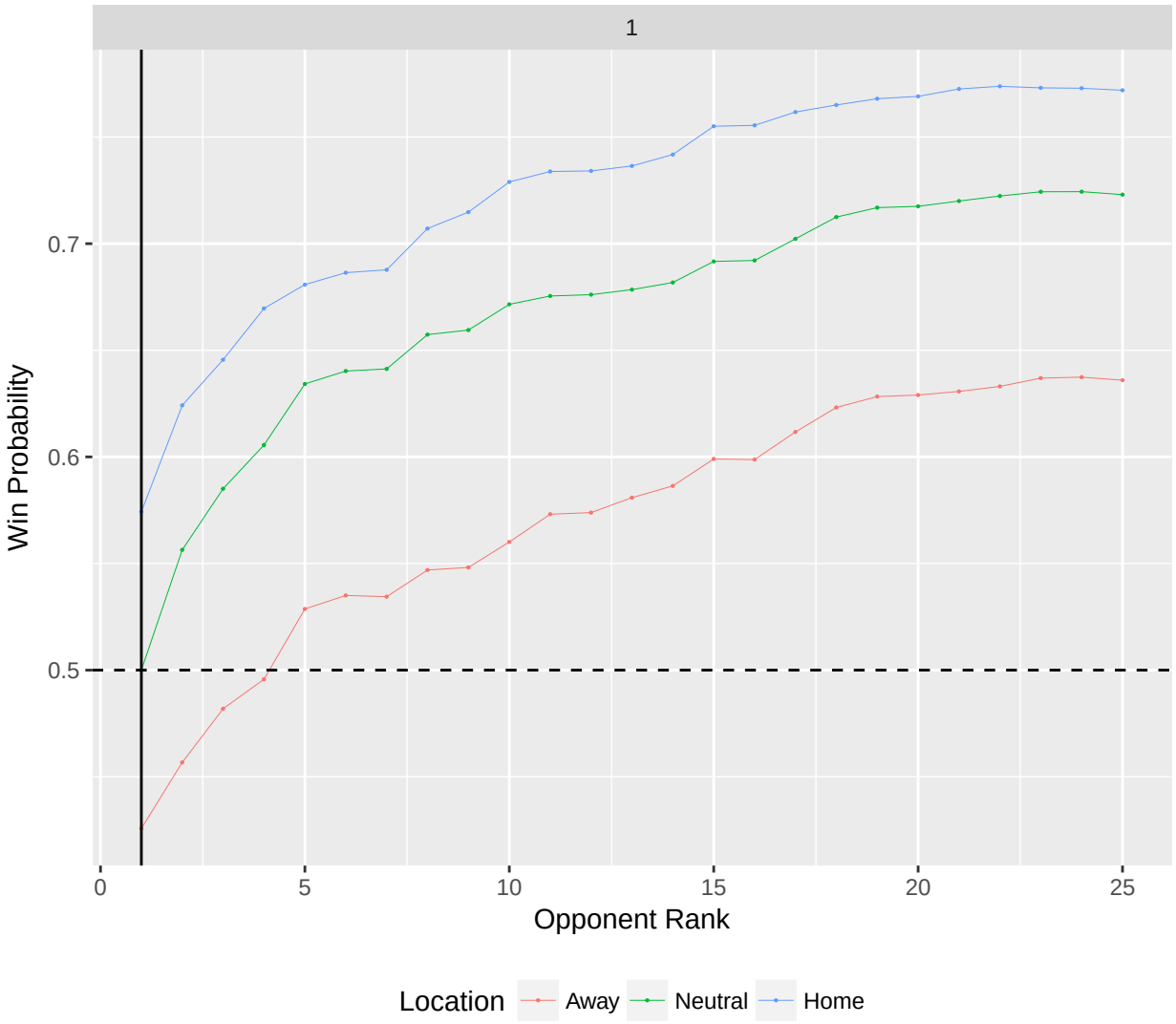
Table 11: OOB losses for random forests constructed using variable set 1 for the 2016 season.

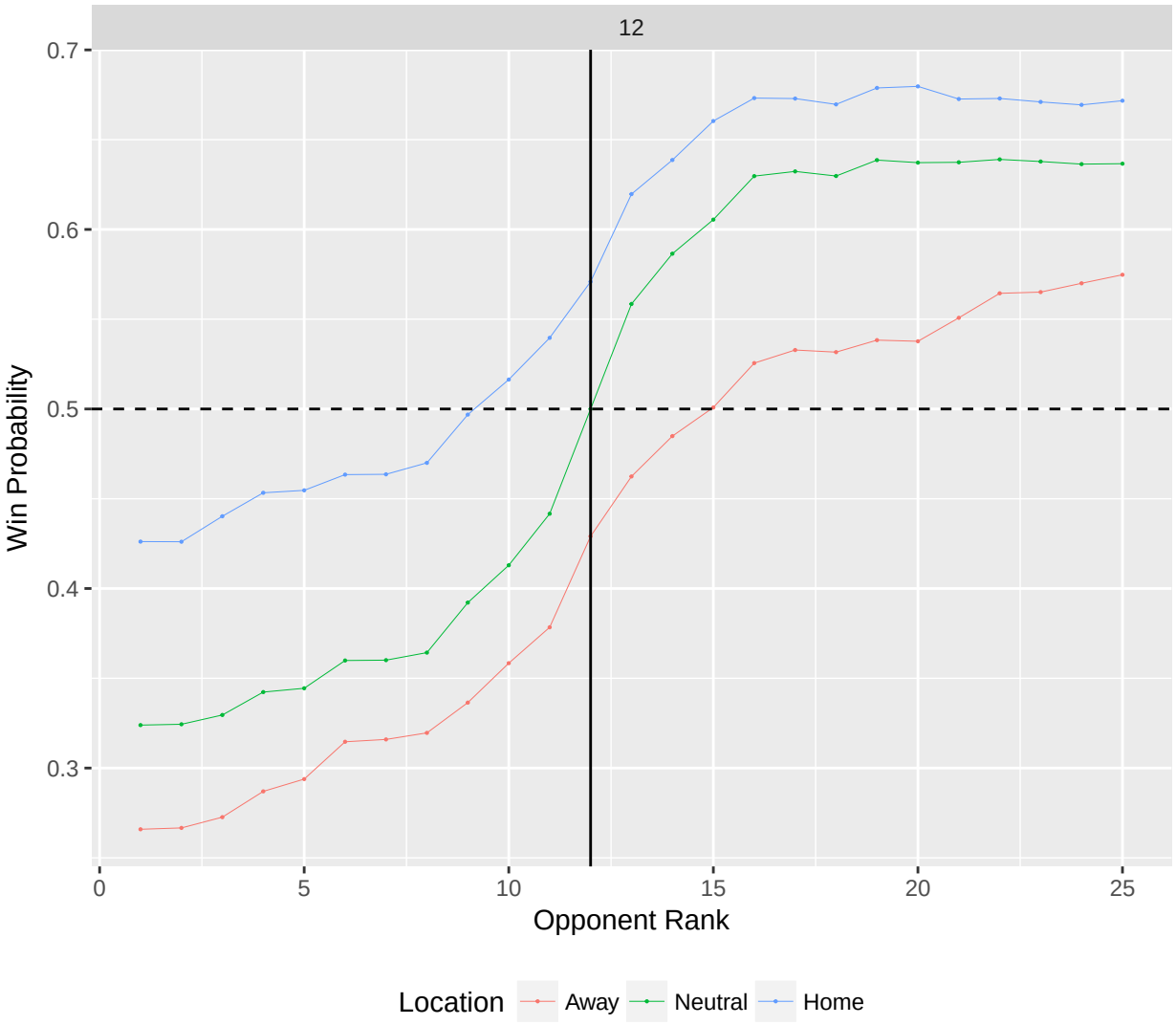
mtry	nodesize	OOB Loss
2	230	928.5
2	240	928.6
2	210	928.8
2	220	928.9
2	200	929.0
2	190	929.1
2	170	929.6
2	180	929.6
2	150	929.8
2	250	929.8
2	160	930.1
2	140	930.5
2	130	931.4
1	130	932.4
1	140	932.4
1	150	932.4
1	180	932.4
2	120	932.5
1	160	932.5
1	120	932.8
1	170	932.8
1	200	932.8
1	190	933.2
1	240	933.7
1	210	934.2

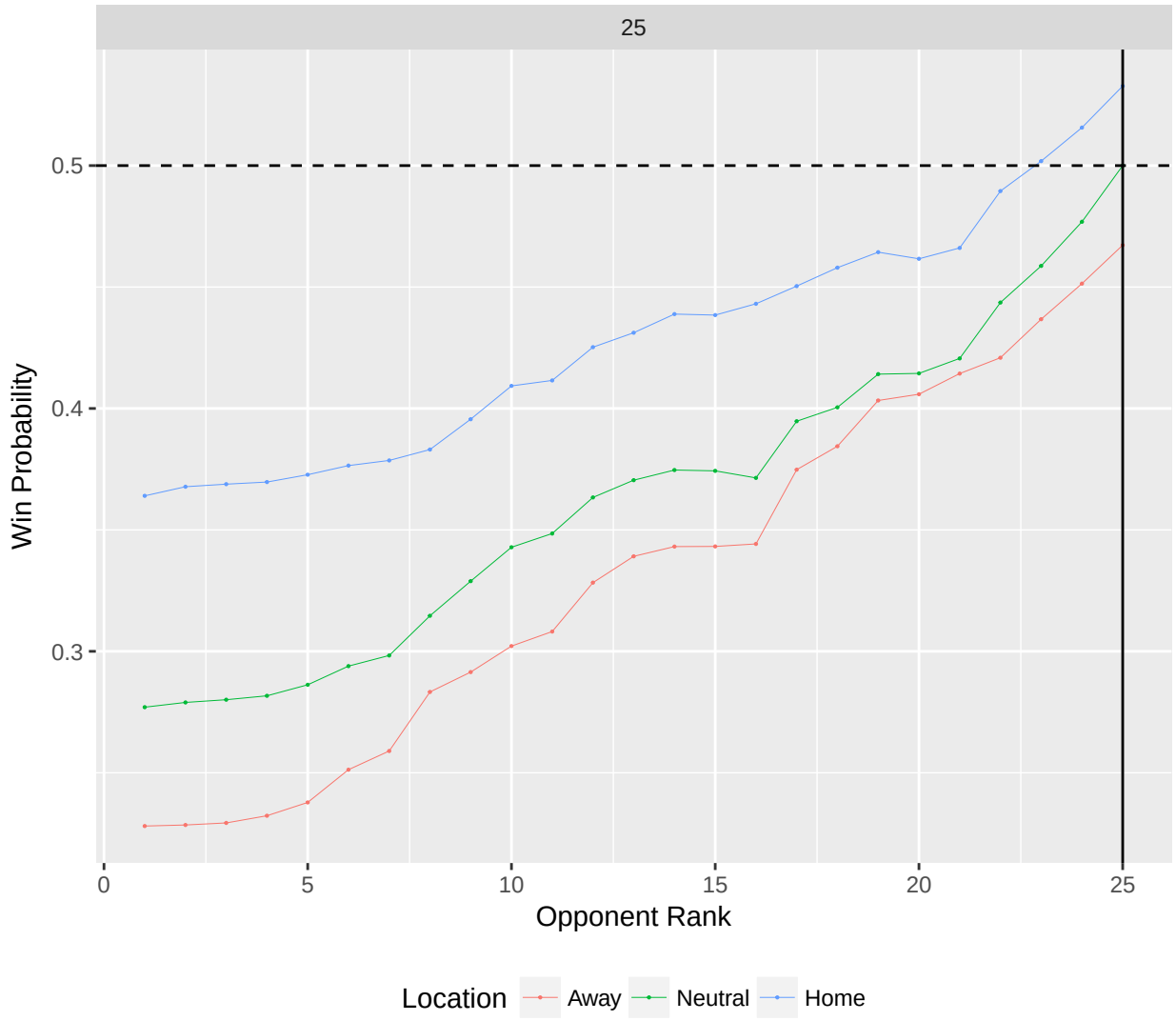
mtry	nodesize	OOB Loss
1	220	934.2
1	230	934.3
1	250	935.8

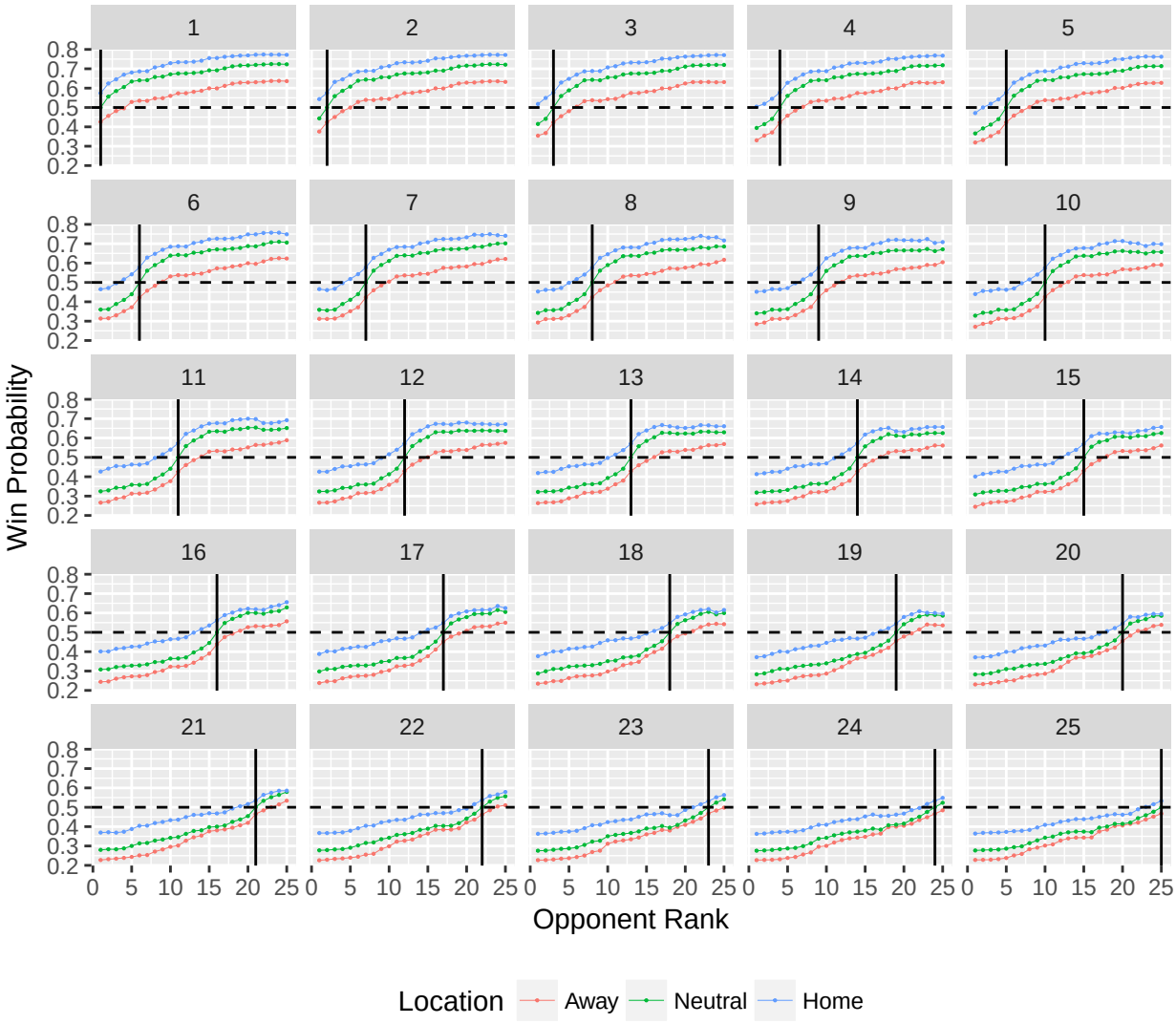
Table 12: Estimated win probabilities for a Team ranked 25 playing at home against an Opponent ranked 5 in the 2016 season.

Variable Set	GLM Logit	GLM Probit	Random Forest
Set 1	0.309	0.309	0.373
Set 2	0.316	0.317	0.373
Set 3	0.263	0.262	0.372
Set 4	0.271	0.271	0.375
Set 5	0.324	0.325	0.370
Set 6	0.330	0.332	0.375
Set 7	0.277	0.277	0.350
Set 8	0.284	0.285	0.350


1
Win Probability
0.7
0.6
0.5
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


12
0.7
0.6
0.5
0.4
0.3
Win Probability
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


25
0.5
0.4
0.3
Win Probability
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home



To assess the random forest estimations, we compared negative log likelihood losses. As before, the equation used for calculating the negative log likelihood loss is $-1 \cdot \sum_{i=1}^{n}[Y_i \cdot \log(\hat{\pi}_i) + (1 - Y_i) \cdot \log(1 - \hat{\pi}_i)]$, where Y_i represents the results of game i (0: loss, 1: win) and $\hat{\pi}_i$ represents the estimated win probability for game i. We calculated the negative log likelihood loss for each variable set for each season, and then summed over all seasons to get a single loss value for each variable set. The sums of the negative log likelihood losses for the random forest estimations, along with the sums of the losses for the GLMs, can be found in Table 13. Variable set 6, which included location, difference in ranks, and the interaction between difference in ranks and average rank, resulted in the lowest loss for the random forests. This is different than

the GLMs, which had the lowest loss when using variable set 2. It is also interesting to note that the lowest random forest loss is still higher than the highest GLM loss. This seems to imply that using a GLM will result in better predictions than a random forest.

Table 13: Comparing the sum of losses between GLMs that use a probit link, GLMs that use a logit link, and random forests. Variable set 2 resulted in the lowest loss for the GLMs, while variable set 6 resulted in the lowest loss for the random forest estimations.

Variable Set	Probit Link Losses	Logit Link Losses	Random Forest Losses
Set 1	948.8	948.8	958.3
Set 2	947.5	947.5	958.5
Set 3	953.1	953.2	958.8
Set 4	951.8	951.9	958.7
Set 5	950.7	950.8	958.1
Set 6	949.3	949.4	958.0
Set 7	954.8	954.8	958.7
Set 8	953.4	953.4	959.0

5 Multiple Linear Regression Method

In addition to using GLMs and random forests to estimate win probabilities, we also used multiple linear regression (MLR) models to predict point differentials instead of estimating win probabilities. The predicted point differentials could then be modified to predict a winner, in that a positive predicted point differential indicates a team is predicted to win and a negative predicted point differential indicates a team is predicted to lose. The MLR models were structured in the following way:

$$Y_i = \text{Points scored in game } i - \text{Opponent points scored in game } i$$

$$X_{1,i} : \text{Location of game i, -1 for away, 0 for neutral, and 1 for home}$$

$$X_{2,i} : \text{Difference in Team and Opponent Rank (Team minus Opponent)}$$

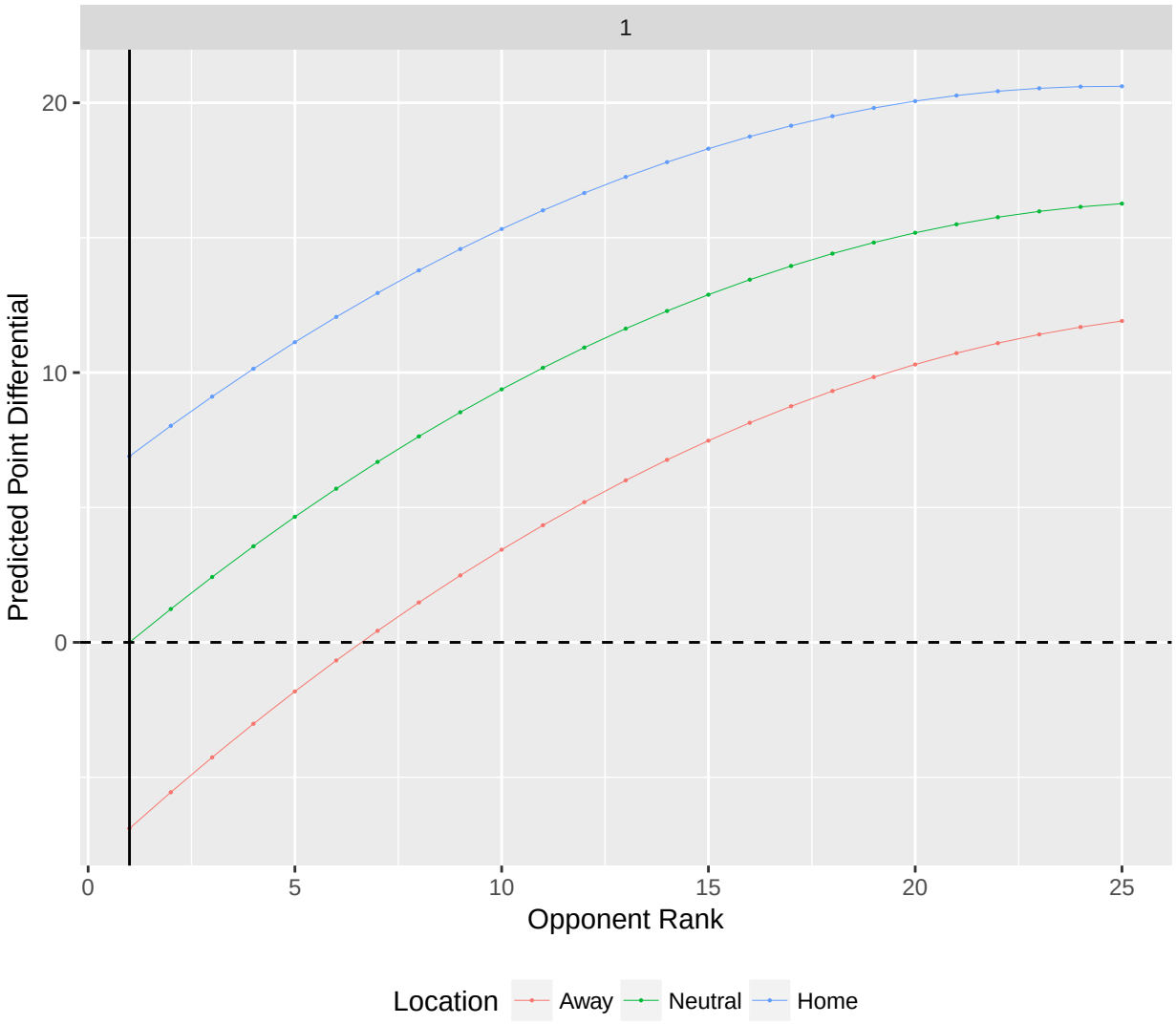
$$X_{3,i} : \text{Average of Team and Opponent Rank}$$

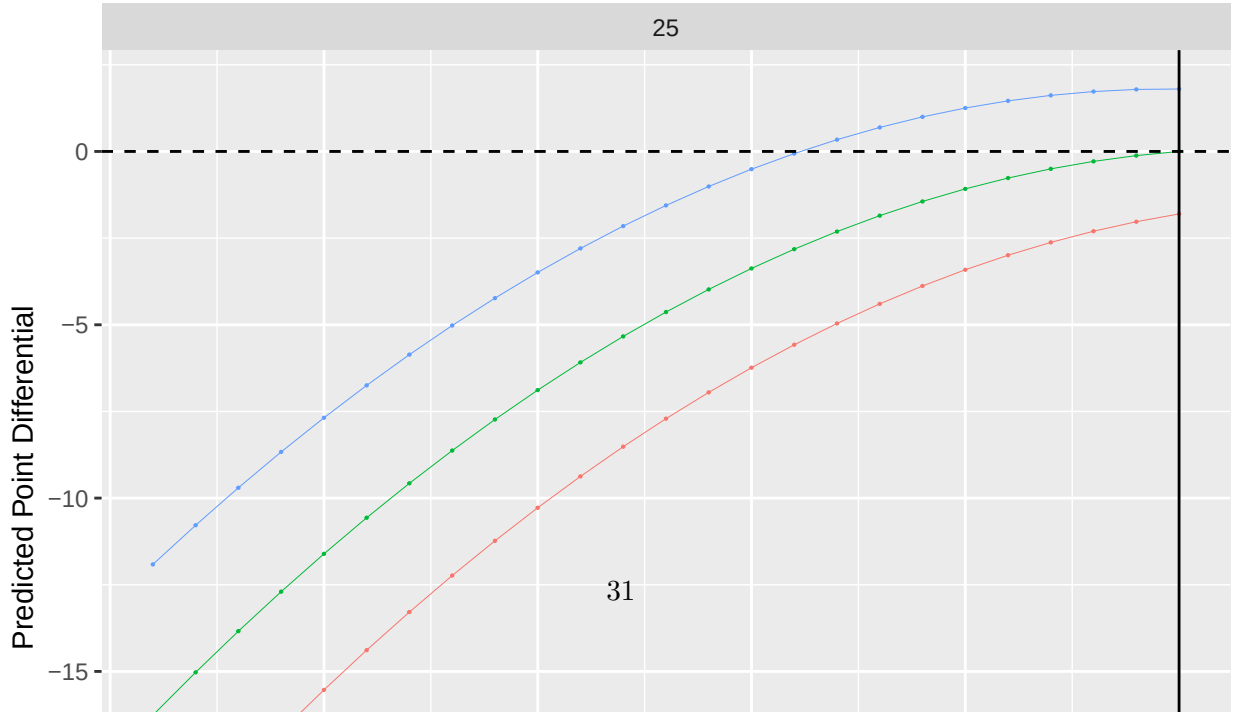
$$Y_i = \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i} + \epsilon_i$$

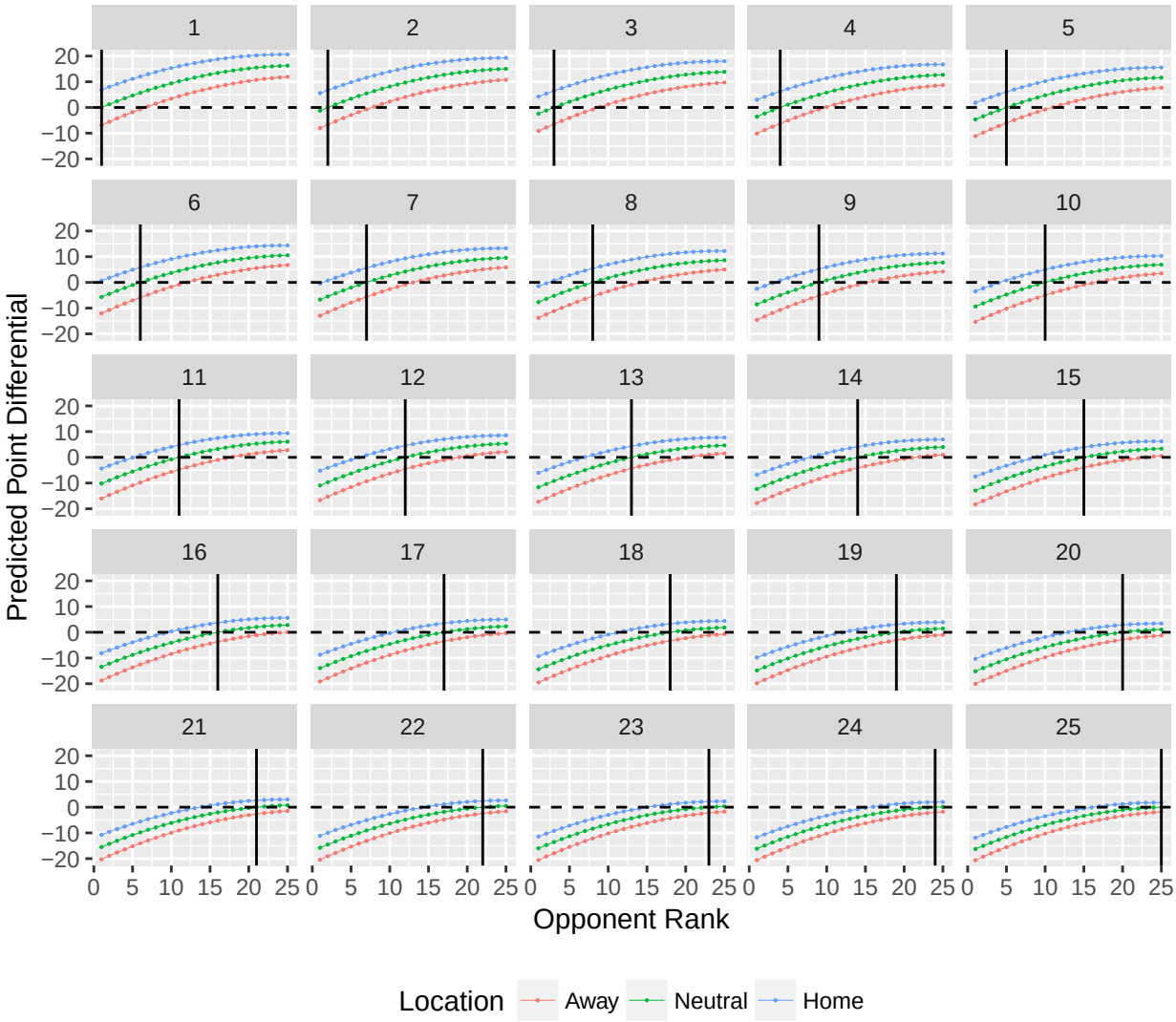
$$\epsilon_i \sim N(0, \sigma_e^2) \ .$$

Table 14: Win probability estimates and point differential predictions for a Team ranked 25 playing at home against an Opponent ranked 5 in the 2016 season.

Variable Set	GLM Logit	GLM Probit	Random Forest	MLR
Set 1	0.309	0.309	0.373	-7.57
Set 2	0.316	0.317	0.373	-7.68
Set 3	0.263	0.262	0.372	-9.84
Set 4	0.271	0.271	0.375	-9.90
Set 5	0.324	0.325	0.370	-6.97
Set 6	0.330	0.332	0.375	-7.13
Set 7	0.277	0.277	0.350	-9.24
Set 8	0.284	0.285	0.350	-9.35


1
20
10
0
Predicted Point Differential
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


25
0
−5
−10
−15
Predicted Point Differential



We computed the mean square error (MSE) to assess our MLR models by comparing the predicted point differentials to what actually happened during that season. The equation used for computing the MSE was $\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$, where Y_i represents the point differential of game i, $\hat{Y}_i$ represents the predicted point differential for game i, and n represents the number of games in a given season. The MSEs for five individual seasons are displayed in Table 15. For each variable set, we calculated the MSE for each season and then found a weighted average (weighted by number of games in a season) of the MSEs across all seasons. The averages of the MSEs can be seen in Table 16. Variable set 8, which included just location and difference in ranks, resulted in the lowest average MSE over all seasons. This is different than with the GLMs and random forests, where the

"best" variable sets were sets 2 and 6.

Table 15: Comparing MSEs for the eight variable sets for the 2012 through 2016 seasons.

Variable Set	2012 MSE	2013 MSE	2014 MSE	2015 MSE	2016 MSE
Set 1	6.7	5.9	6.7	7.7	8.2
Set 2	6.6	5.8	6.6	7.5	8.0
Set 3	6.6	5.7	6.6	7.5	8.4
Set 4	6.5	5.6	6.5	7.4	8.3
Set 5	6.5	5.9	6.6	7.4	8.0
Set 6	6.4	5.8	6.5	7.2	7.9
Set 7	6.4	5.7	6.5	7.2	8.3
Set 8	6.3	5.5	6.4	7.1	8.1

Table 16: MSEs for each variable set, averaged over all seasons.

Variable Set	Average of MSEs
Set 1	6.30
Set 2	6.17
Set 3	6.22
Set 4	6.09
Set 5	6.18
Set 6	6.05
Set 7	6.11
Set 8	5.99

6 Comparing the 32 Methods

6.1 Methods Summary

6.2 Comparing Prediction Accuracy

Table 17: Three example games from the 2016 season.

	Game 1	Game 2	Game 3
Team Rank	22	10	8
Opponent Rank	18	17	2
Location	Neutral	Neutral	Home
Result	Loss	Win	Loss
Point Differential	-9	21	-7

Table 18: Predictions by each of the 32 methods for the three games described in Table 17.

Method	Variable Set	Game 1	Game 2	Game 3
GLM Logit	1	0.471	0.607	0.490
GLM Logit	2	0.471	0.607	0.493
GLM Logit	3	0.435	0.613	0.545
GLM Logit	4	0.435	0.613	0.547
GLM Logit	5	0.470	0.607	0.440
GLM Logit	6	0.470	0.607	0.442
GLM Logit	7	0.435	0.613	0.496
GLM Logit	8	0.435	0.613	0.498
GLM Probit	1	0.471	0.606	0.491
GLM Probit	2	0.471	0.606	0.493
GLM Probit	3	0.436	0.611	0.545

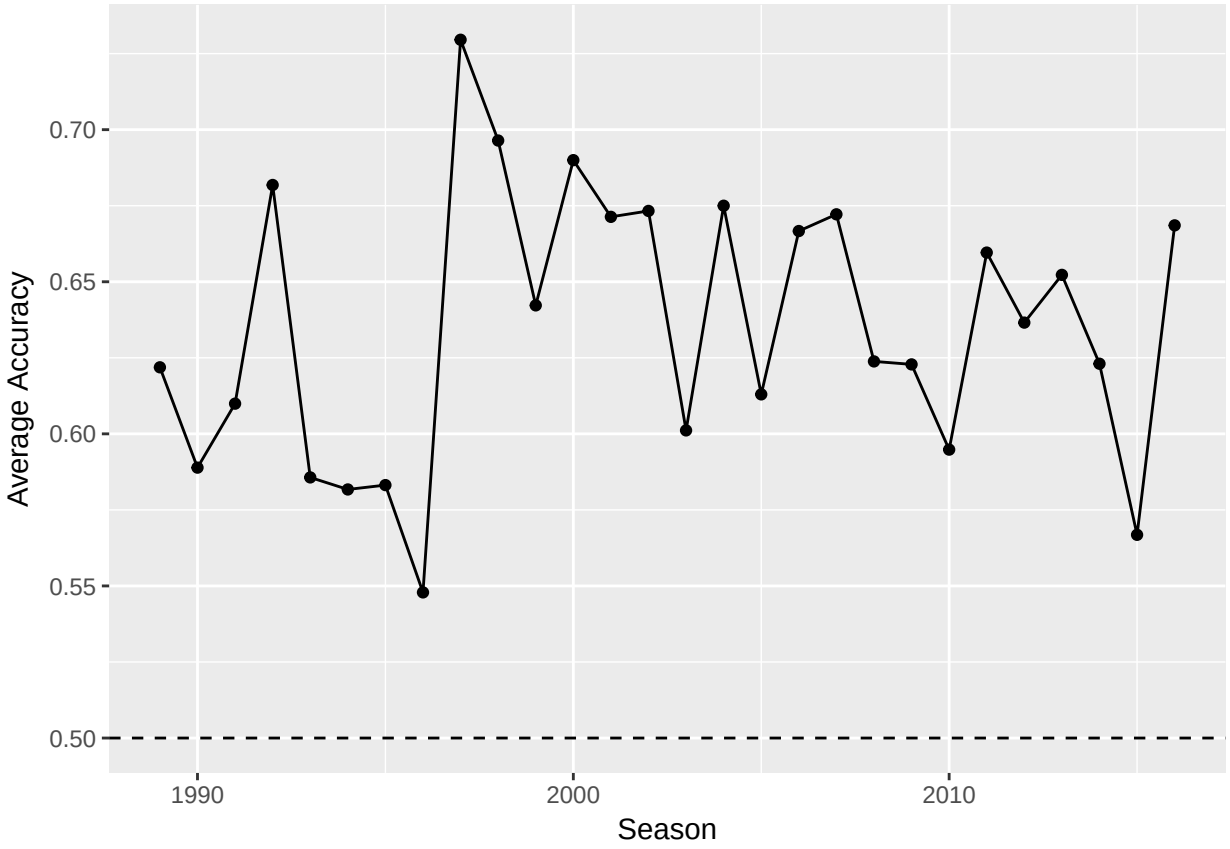
Method	Variable Set	Game 1	Game 2	Game 3
GLM Probit	4	0.436	0.611	0.548
GLM Probit	5	0.470	0.605	0.441
GLM Probit	6	0.470	0.605	0.443
GLM Probit	7	0.436	0.611	0.496
GLM Probit	8	0.436	0.611	0.498
Random Forest	1	0.402	0.646	0.461
Random Forest	2	0.405	0.649	0.461
Random Forest	3	0.405	0.647	0.458
Random Forest	4	0.402	0.649	0.461
Random Forest	5	0.403	0.649	0.461
Random Forest	6	0.405	0.647	0.460
Random Forest	7	0.372	0.654	0.471
Random Forest	8	0.373	0.653	0.472
MLR	1	-1.350	4.575	-0.322
MLR	2	-1.350	4.573	-0.349
MLR	3	-2.764	4.836	1.981
MLR	4	-2.763	4.835	1.965
MLR	5	-1.335	4.575	-1.923
MLR	6	-1.334	4.572	-1.967
MLR	7	-2.765	4.840	0.355
MLR	8	-2.764	4.837	0.323

Table 19: Accuracies of each method for the 2014, 2015, and 2016 seasons. Accuracies ranged between .5344 to .6964.

Method	Variable Set	2014 Accuracy	2015 Accuracy	2016 Accuracy
GLM Logit	1	0.6094	0.5517	0.6964
GLM Logit	2	0.6094	0.5345	0.6964

Method	Variable Set	2014 Accuracy	2015 Accuracy	2016 Accuracy
GLM Logit	3	0.5938	0.5517	0.6786
GLM Logit	4	0.6406	0.5862	0.6429
GLM Logit	5	0.6094	0.5517	0.6964
GLM Logit	6	0.6094	0.5345	0.6964
GLM Logit	7	0.5938	0.5690	0.6786
GLM Logit	8	0.6406	0.5862	0.6429
GLM Probit	1	0.6406	0.5690	0.6429
GLM Probit	2	0.6406	0.5862	0.6429
GLM Probit	3	0.6406	0.5862	0.6786
GLM Probit	4	0.6406	0.5862	0.6429
GLM Probit	5	0.6406	0.5690	0.6429
GLM Probit	6	0.6406	0.5690	0.6429
GLM Probit	7	0.6406	0.5862	0.6786
GLM Probit	8	0.6406	0.5862	0.6429
Random Forest	1	0.6250	0.5862	0.6786
Random Forest	2	0.6250	0.5862	0.6786
Random Forest	3	0.5938	0.5517	0.6786
Random Forest	4	0.6406	0.5862	0.6429
Random Forest	5	0.6250	0.5862	0.6786
Random Forest	6	0.6250	0.5862	0.6786
Random Forest	7	0.5938	0.5517	0.6786
Random Forest	8	0.6406	0.5862	0.6250
MLR	1	0.6094	0.5345	0.6964
MLR	2	0.6094	0.5345	0.6964
MLR	3	0.6094	0.5517	0.6607
MLR	4	0.6406	0.5862	0.6429
MLR	5	0.6094	0.5345	0.6964
MLR	6	0.6094	0.5345	0.6964

Method	Variable Set	2014 Accuracy	2015 Accuracy	2016 Accuracy
MLR	7	0.6094	0.5517	0.6607
MLR	8	0.6406	0.5862	0.6607



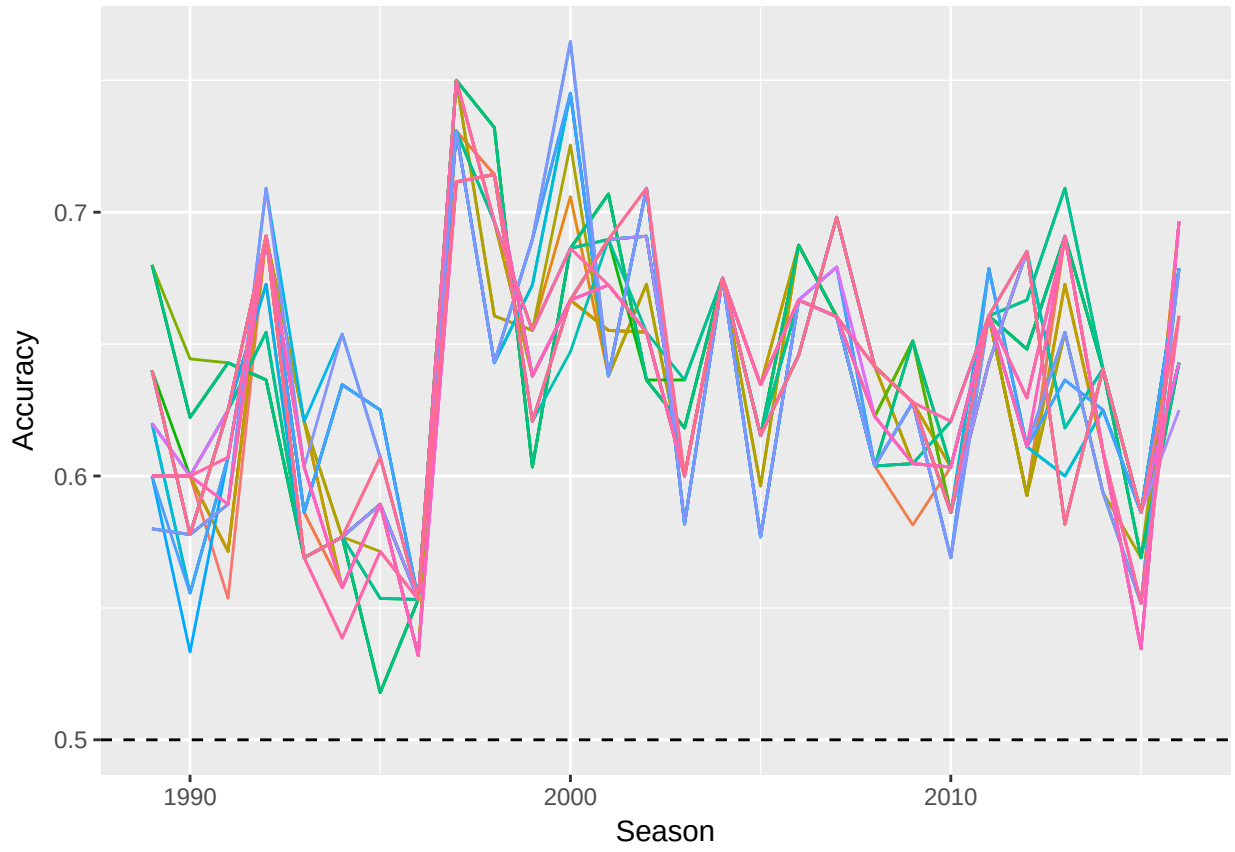


Table 20: Total Accuracies for each of the 32 Methods.

Method	Variable Set	Total Accuracy
GLM Probit	3	0.6392
GLM Probit	7	0.6392
GLM Probit	1	0.6386
GLM Probit	2	0.6386
GLM Probit	5	0.6379
GLM Probit	6	0.6379
MLR	8	0.6379
GLM Probit	4	0.6372
MLR	4	0.6372
GLM Logit	8	0.6359
GLM Probit	8	0.6359

Method	Variable Set	Total Accuracy
Random Forest	6	0.6359
Random Forest	4	0.6352
Random Forest	5	0.6352
MLR	1	0.6352
MLR	2	0.6352
GLM Logit	4	0.6345
Random Forest	2	0.6345
Random Forest	3	0.6345
Random Forest	8	0.6345
MLR	3	0.6345
MLR	7	0.6345
Random Forest	1	0.6339
Random Forest	7	0.6339
GLM Logit	5	0.6332
GLM Logit	6	0.6325
MLR	5	0.6319
MLR	6	0.6319
GLM Logit	7	0.6312
GLM Logit	3	0.6305
GLM Logit	1	0.6292
GLM Logit	2	0.6285

Accuracy Stuff:

```
season_specific_predictions %>% ggplot(aes(x=Model2_GLM_Logit))+geom_histogram(bins=60)+xlab("
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

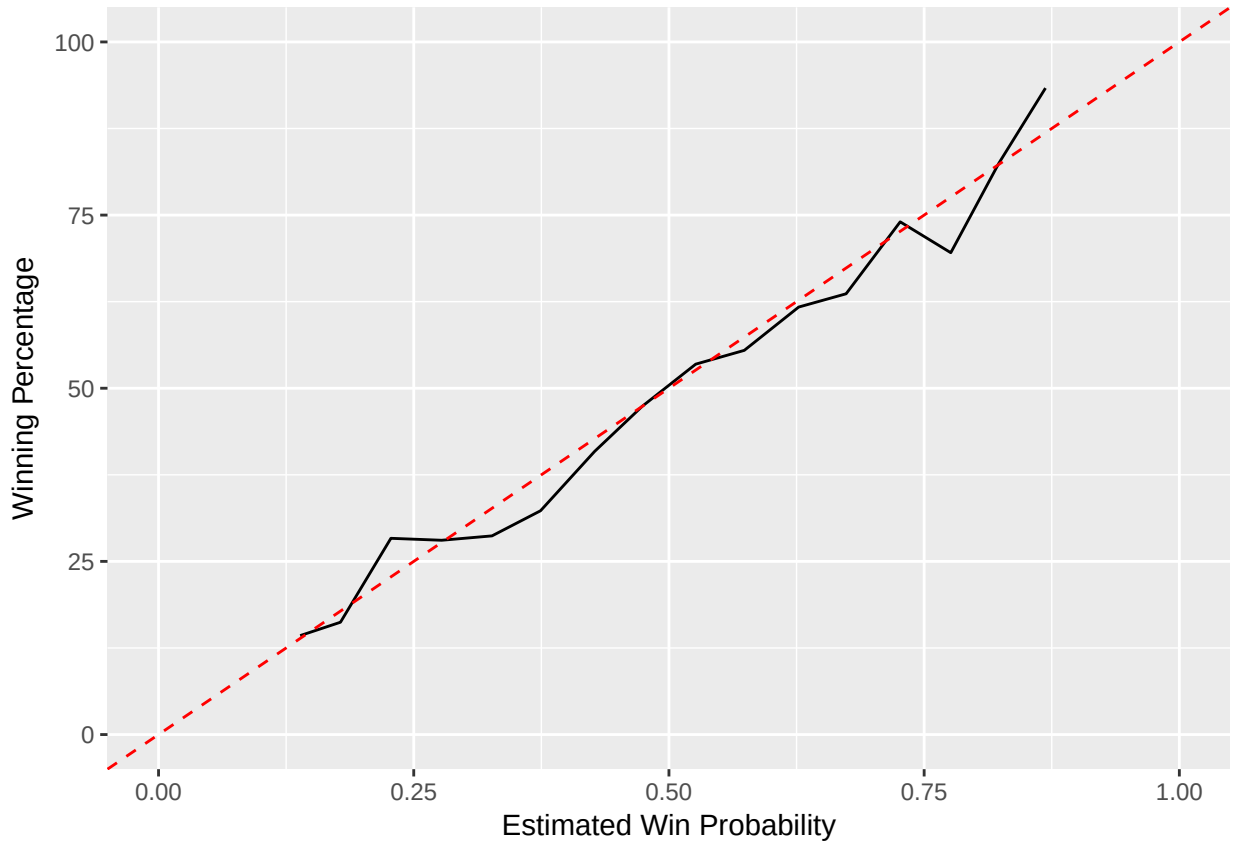
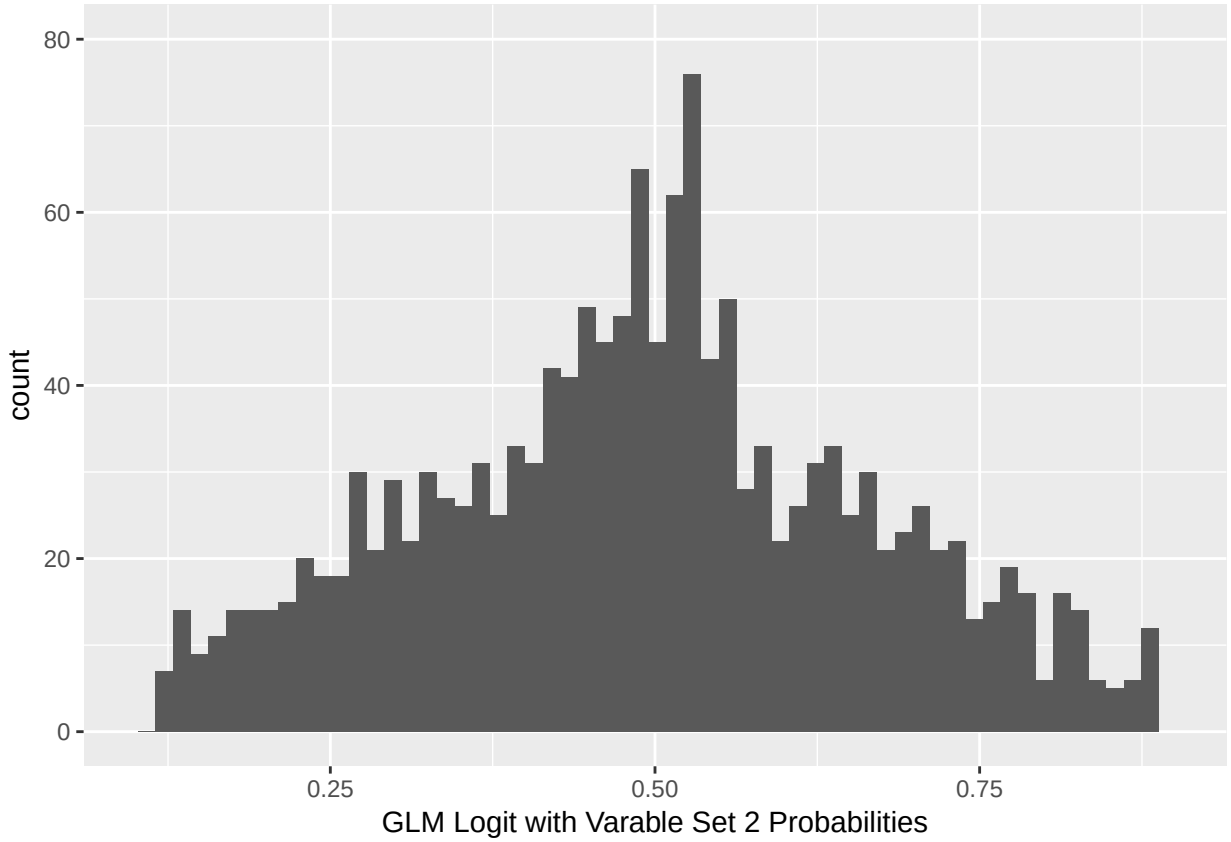
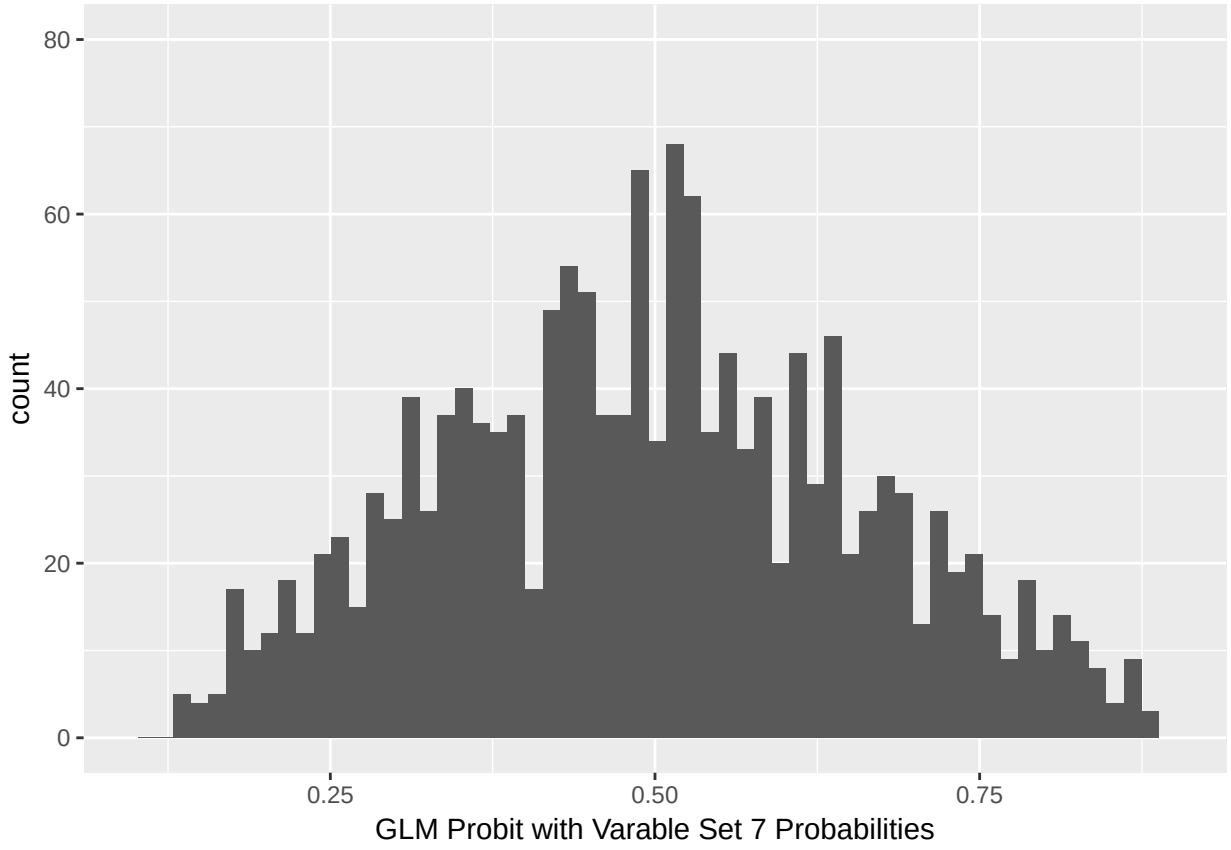


Figure 6: Relationship between estimated win probability and actual winning percentage for predictions generated by the GLM probit method using variable set 7. The dashed red line is along the line where estimated win probability is equal to actual winning percentage.

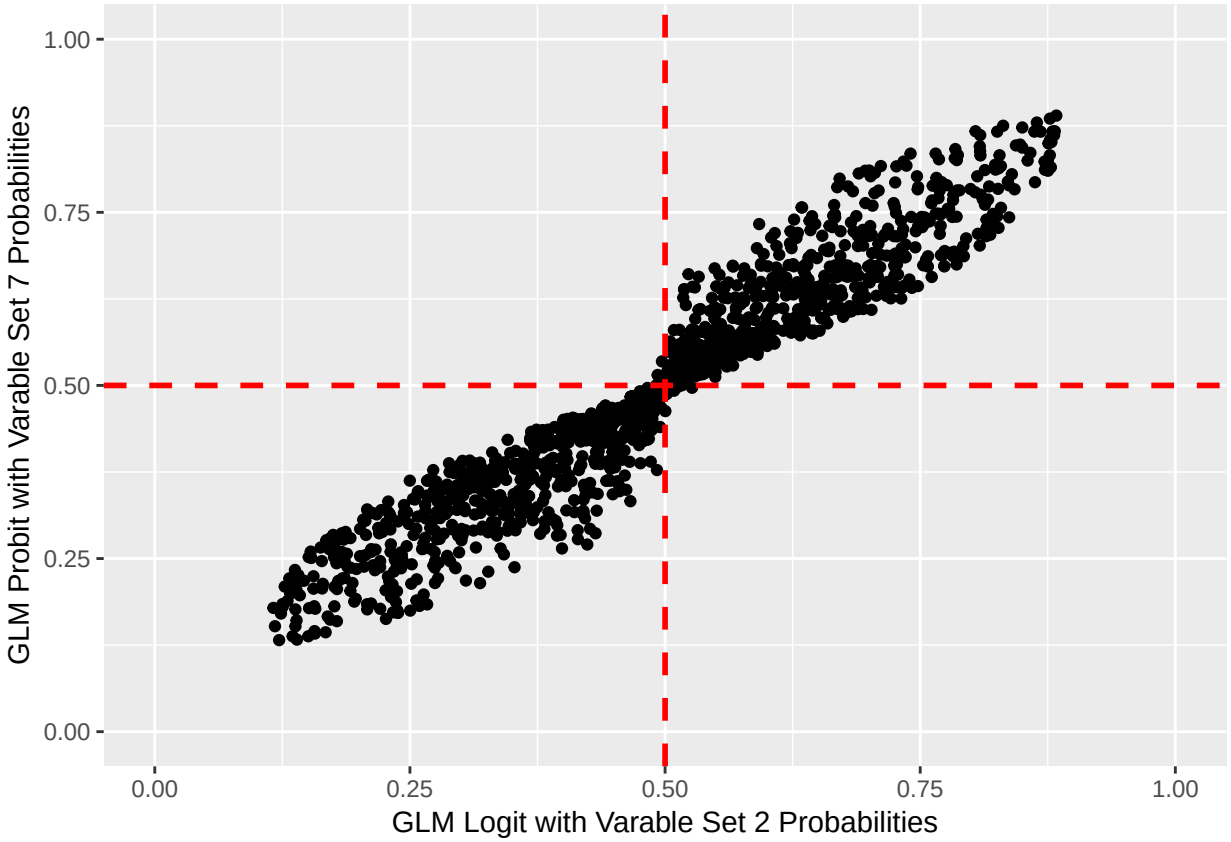



```
season_specific_predictions %>% ggplot(aes(x=Model7_GLM_Probit))+geom_histogram(bins=60)+xlab(
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```




```
season_specific_predictions %>% ggplot(aes(x=Model2_GLM_Logit, y=Model7_GLM_Probit))+geom_point
```



7 Conclusion and Future Work

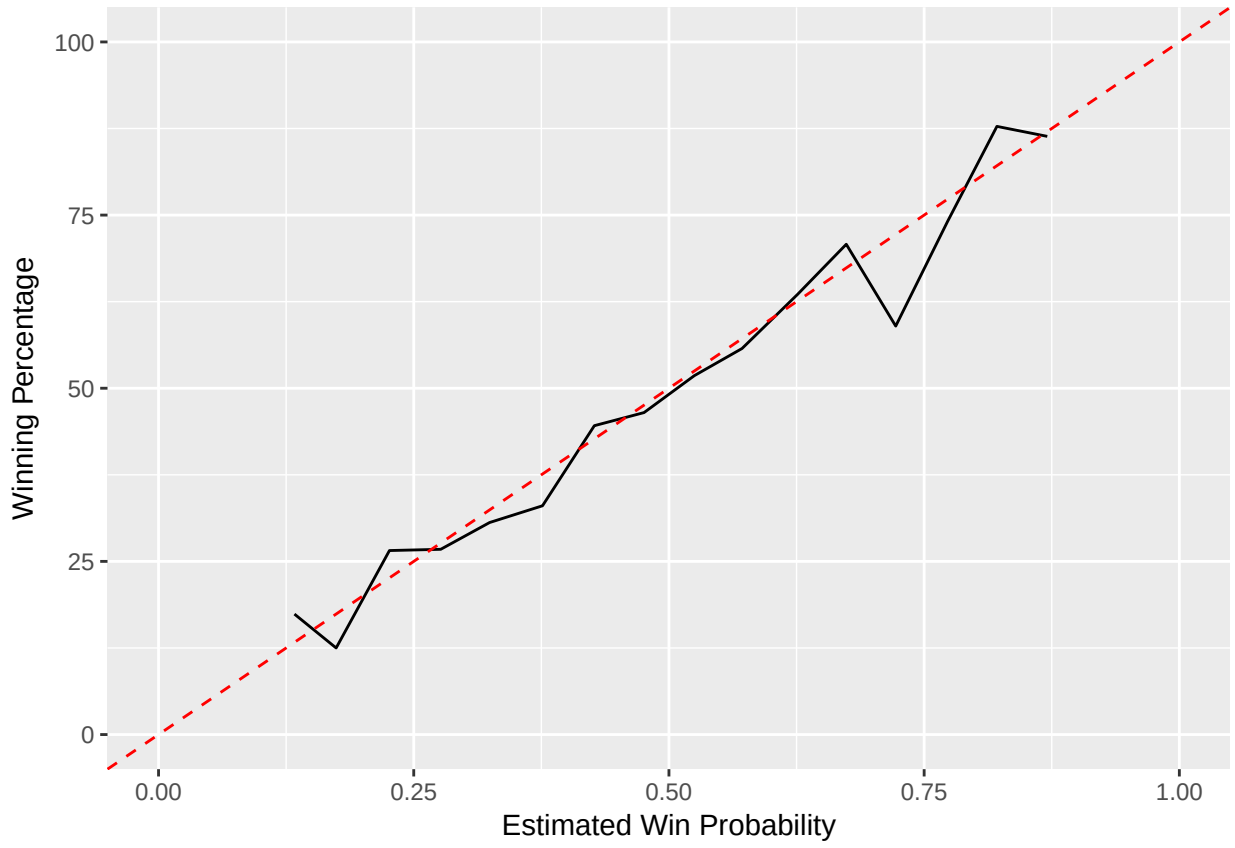


Figure 7: Relationship between estimated win probability and actual winning percentage for predictions generated by the GLM logit method using variable set 2. The dashed red line is along the line where estimated win probability is equal to actual winning percentage.

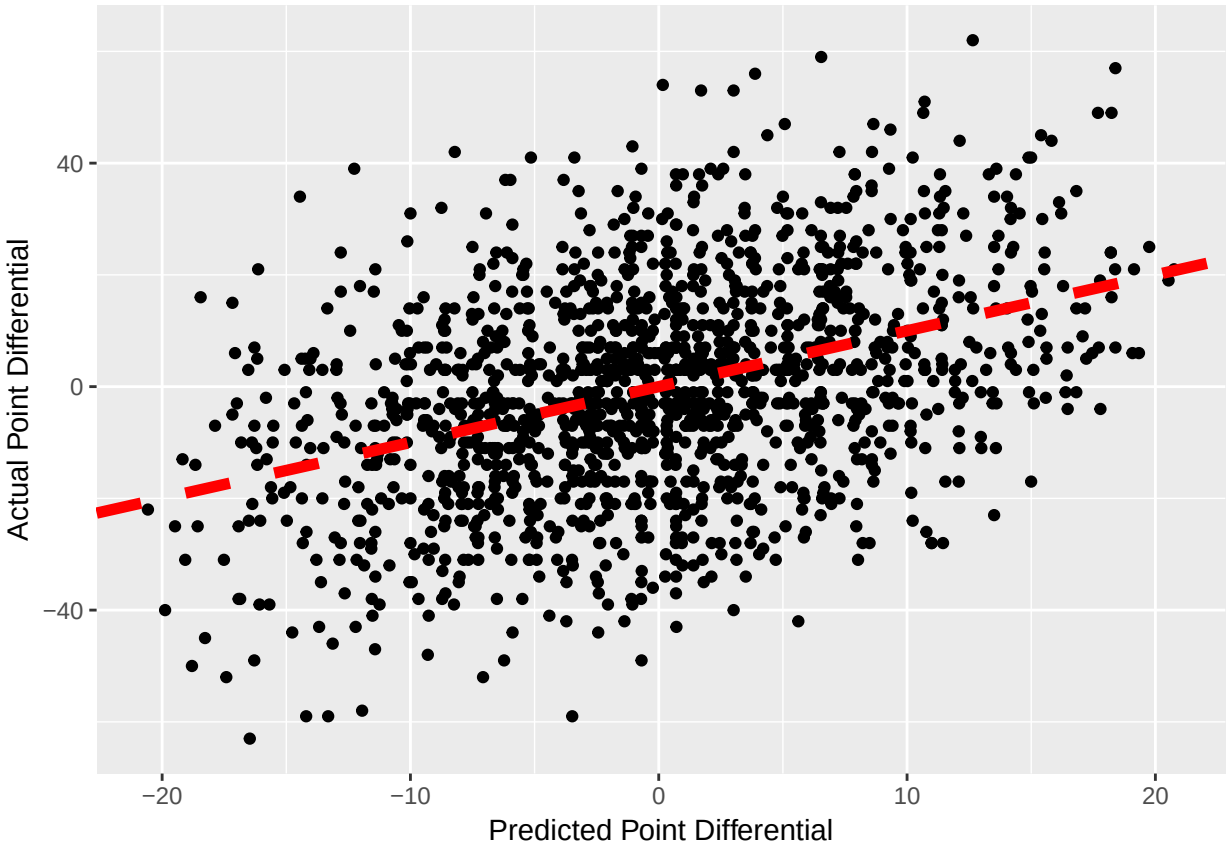


Figure 8: Relationship between predicted point differential and actual point differential for predictions generated by the MLR method using variable set 8. The dashed red line is along the line where predicted point differential is equal to actual point differential.

8 Appendix

8.1 Data Set Introduction

8.2 Total_Team_History

8.2.1 File Description

8.2.2 Example Graphs

The Total_Team_History CSV file can be used to compare total history between teams. For example, Figure 11 compares the difference in total wins between a select number of teams. Figure 12 displays the relationship between a team's all time win total and the number of seasons the team has finished ranked in the AP Poll.

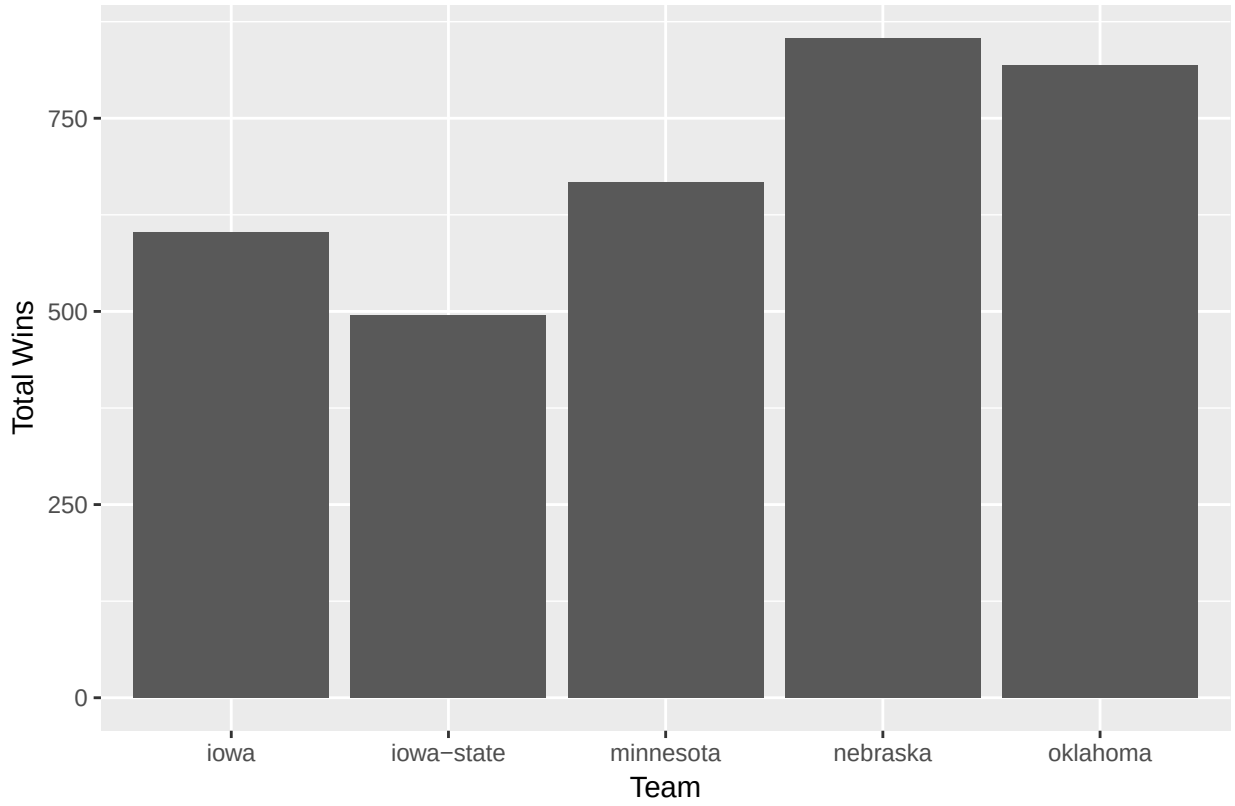


Figure 9: Comparing all time wins between the Nebraska, Iowa, Iowa State, Oklahoma, and Minnesota football programs.

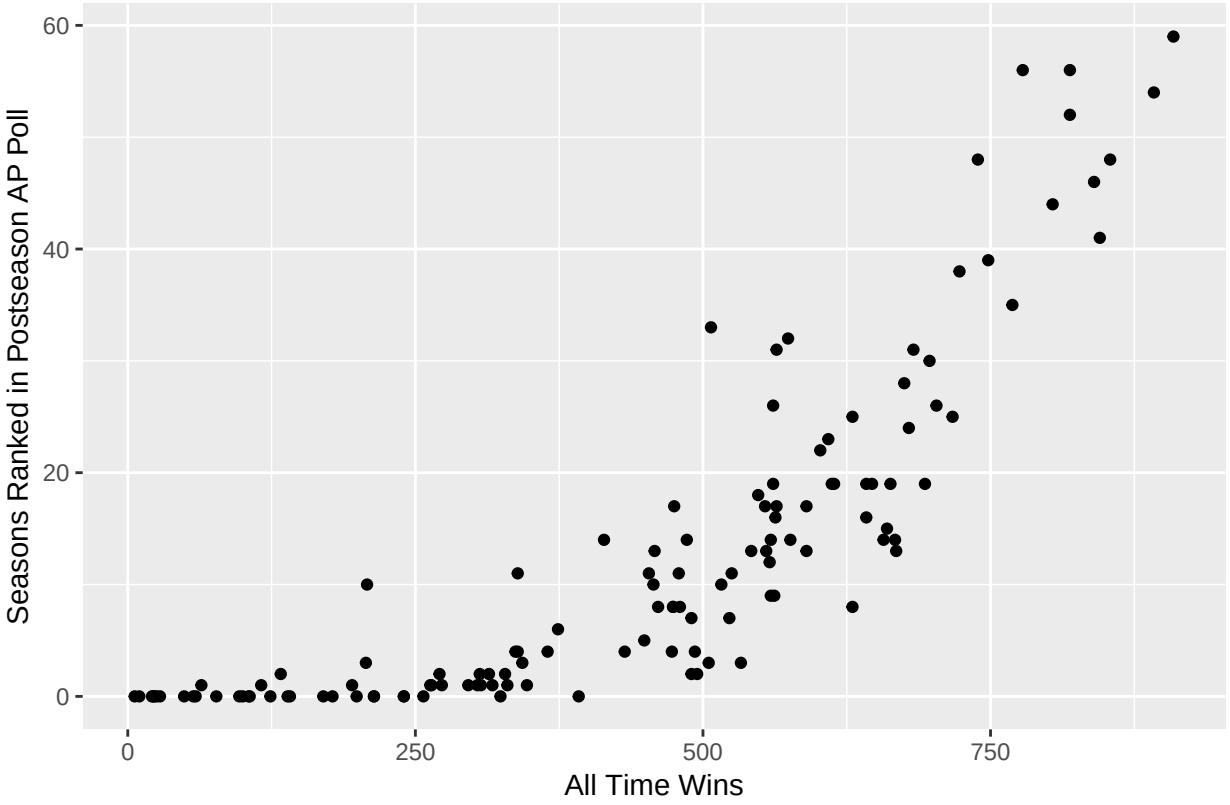


Figure 10: Comparing all time wins to number of seasons ranked in the Postseason AP Poll for teams active in 2016. Programs with more all time wins tend to have more seasons in which they finished ranked.

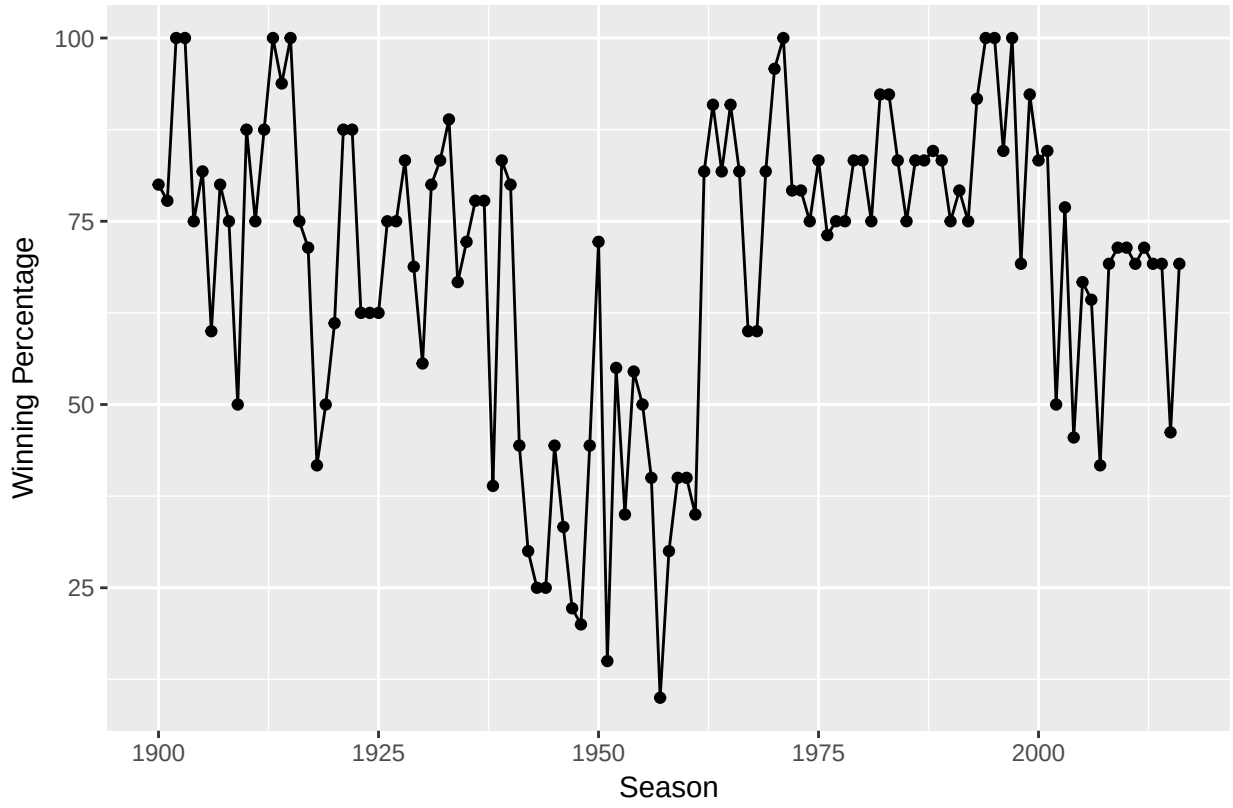


Figure 11: Nebraska winning percentage by season.

8.3 Individual_Season_Results

8.3.1 File Description

8.3.2 Example Graphs

The Individual_Season_Results CSV file has more information than the Total_Team_History file, in that the Individual_Season_Results file allows users to look at each season. For example, Figure 13 displays how Nebraska's winning percentage has varied from season to season. This data set also allows for comparing variables across multiple schools. For example, Figure 14 displays how three SEC teams' simple rating system (SRS) is related to the team's winning percentage. This data set also makes it simple to compare conferences. Figure 15 display the number of bowl wins the SEC, ACC, Big 12, Big Ten, and Pac 10/12 conferences have had per season since 2000.

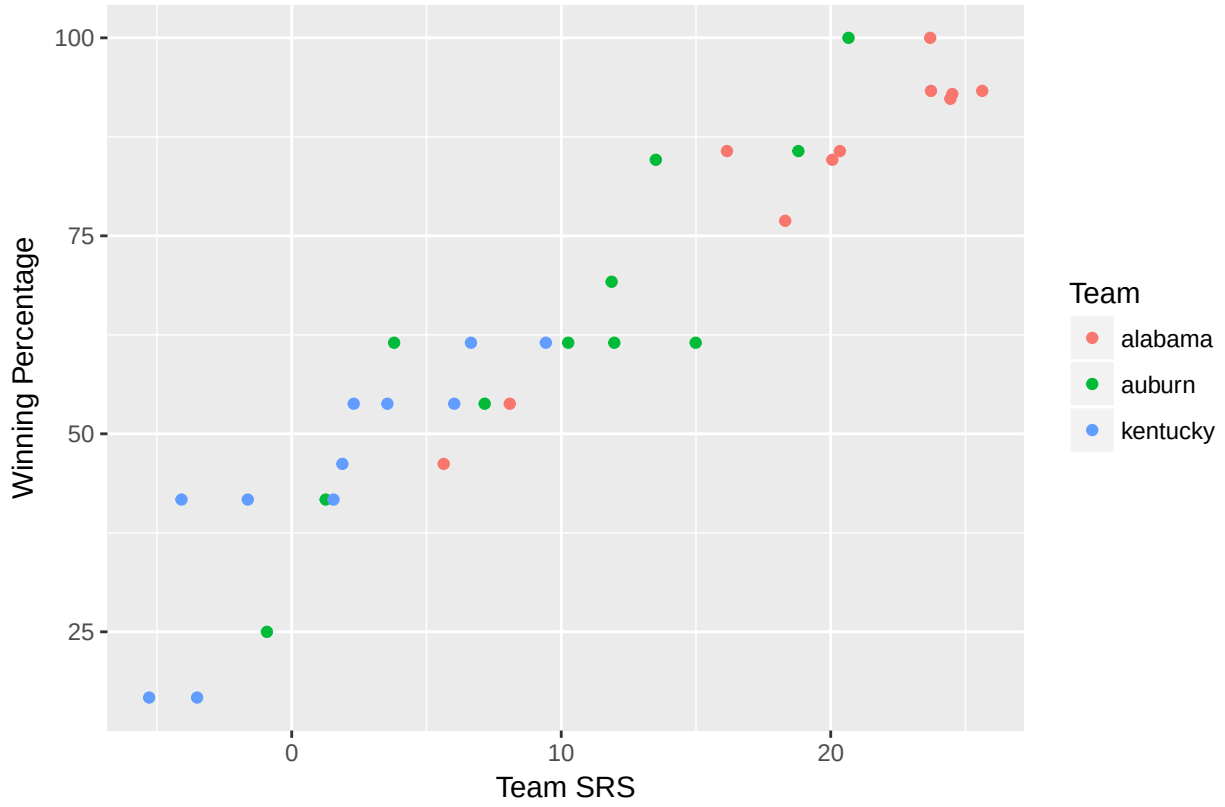


Figure 12: Relationship between SRS (Simple Rating System) and winning percentage for three SEC teams (2005-2016). The teams included show how some teams often have higher SRS and winning percentages (Alabama), some teams often have lower SRS and winning percentages (Kentucky), and some teams vary dramatically in SRS and winning percentage (Auburn).

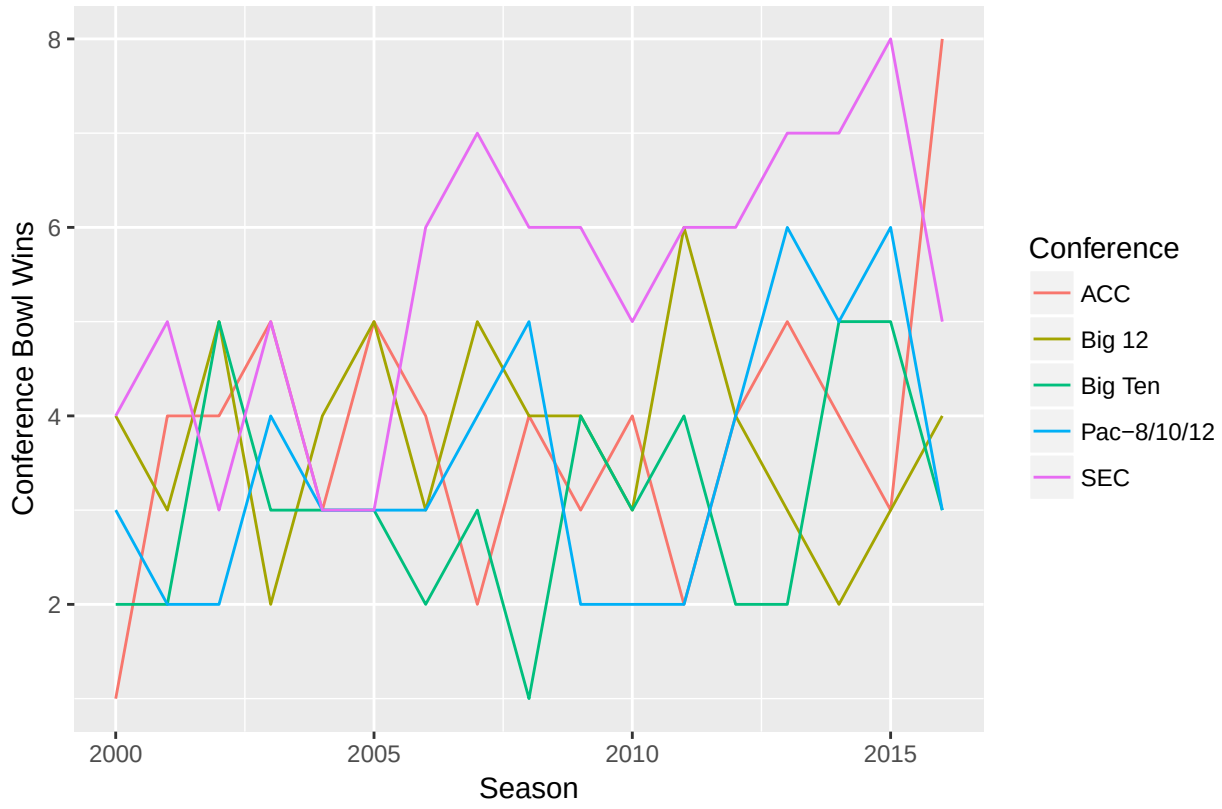


Figure 13: How a conference's number of bowl wins varies from season to season. This graph seems to confirm the reputation the SEC has for being the most successful conference in bowl games (despite the sharp drop off in 2016).

8.4 Season_Averages

8.4.1 File Description

8.4.2 Example Graphs

The Season_Averages file gives a more detailed look at a team's season, as it provides a look at offensive and defensive averages instead of just the winning percentage or total number of wins. For example, Figure 16 shows how Oregon's defensive rushing yards per game has varied since the 1950's. One could also explore how two variables may be related to each other within a particular conference. For instance, Figure 17 shows the relationship between Pac 10/12 team's offensive passing touchdowns (per game) and offensive completion percentages since 2000. This data set makes it easy to look at how teams and conferences differ from each other and change from year to year in more detailed ways. For example, Figure 18 displays side-by-side boxplots for the average passing yards per game for teams in the ACC and the Big 12.

8.5 Game_Results

8.5.1 File Description

8.5.2 Example Graphs

The Game_Results CSV file gives information on every single game played by an FBS football team. This data set makes it simple to observe the distribution of points scored in various games. For example, Figure 19 shows a scatterplot of Nebraska's points scored and Nebraska's points given up in conference games since 2000. Figure 20 shows a similar graph, but compares two teams. It is also easy to look at how many points are generally needed to win a football game. Figure 21 shows what the winning percentage is for a given number of points scored in a game (since 1990). Figure 22 shows how the winning percentage by points scored can vary between conferences. It is also easy to look at what conferences records are against other conferences. For example, Figure 23 shows the SEC's winning percentage in out of conference games. Figure 24 compares the winning percentage in out of conference games between the ACC, Big 12, Big Ten, Pac 8/10/12, and SEC.

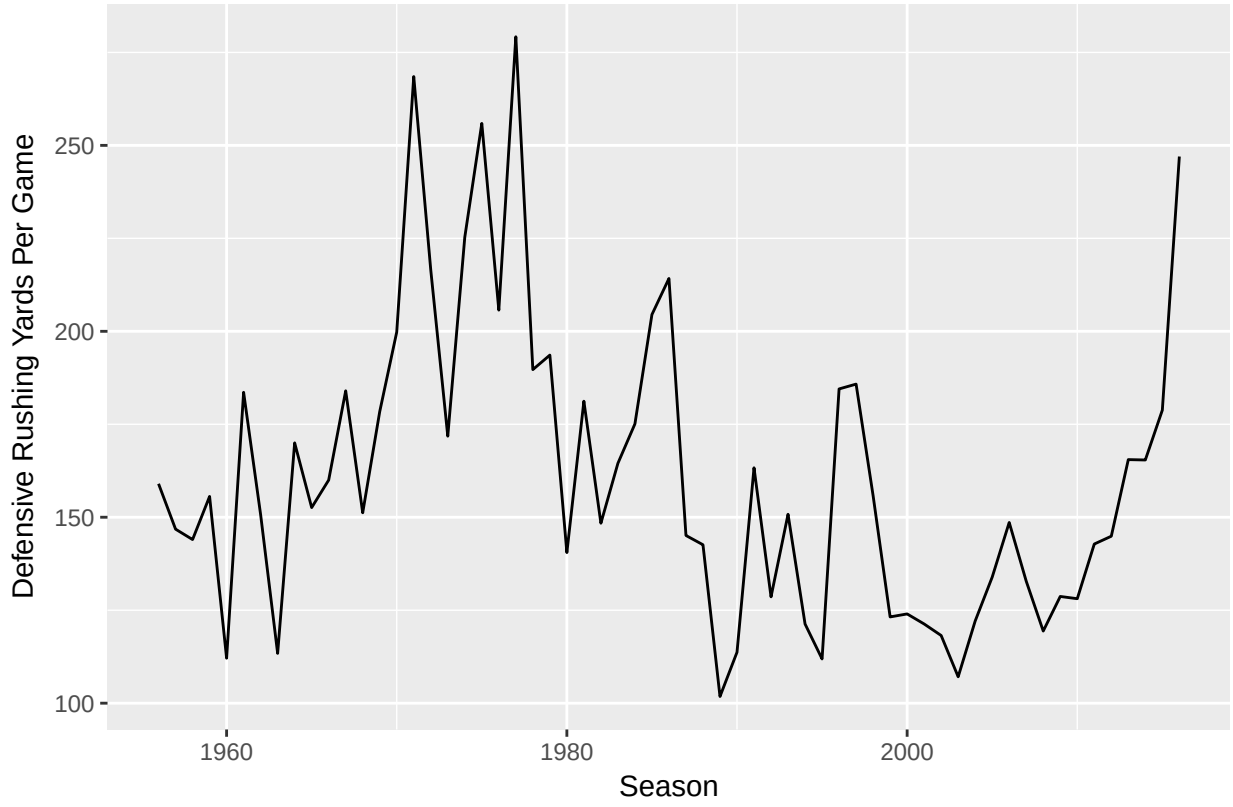


Figure 14: Oregon's defensive rushing yards per game. Something that is particularly interesting is the recent rise in rushing yards allowed by Oregon's defenses.

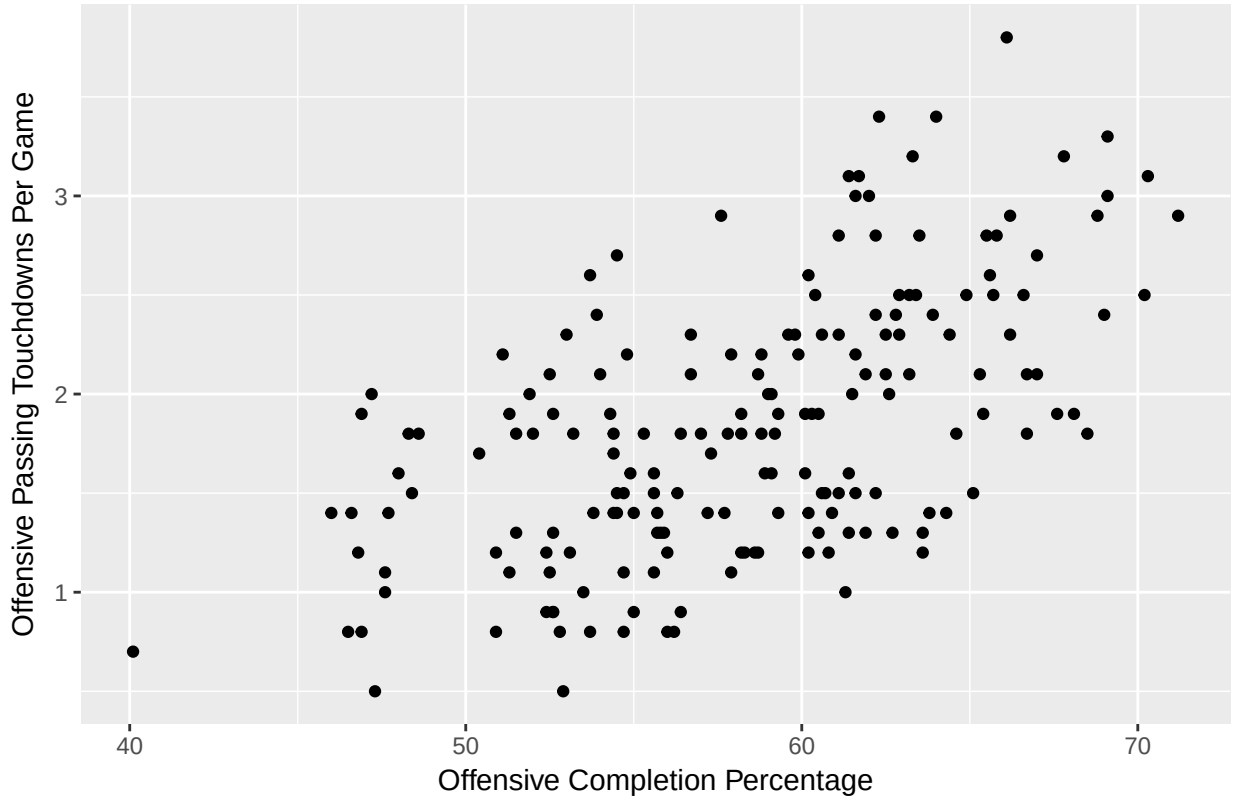


Figure 15: Relationship between Pac 10/12 Teams' (2000-2016) completion percentages and passing touchdowns per game. In a trend that should be expected, teams that have higher completion percentages tend to throw more touchdowns.

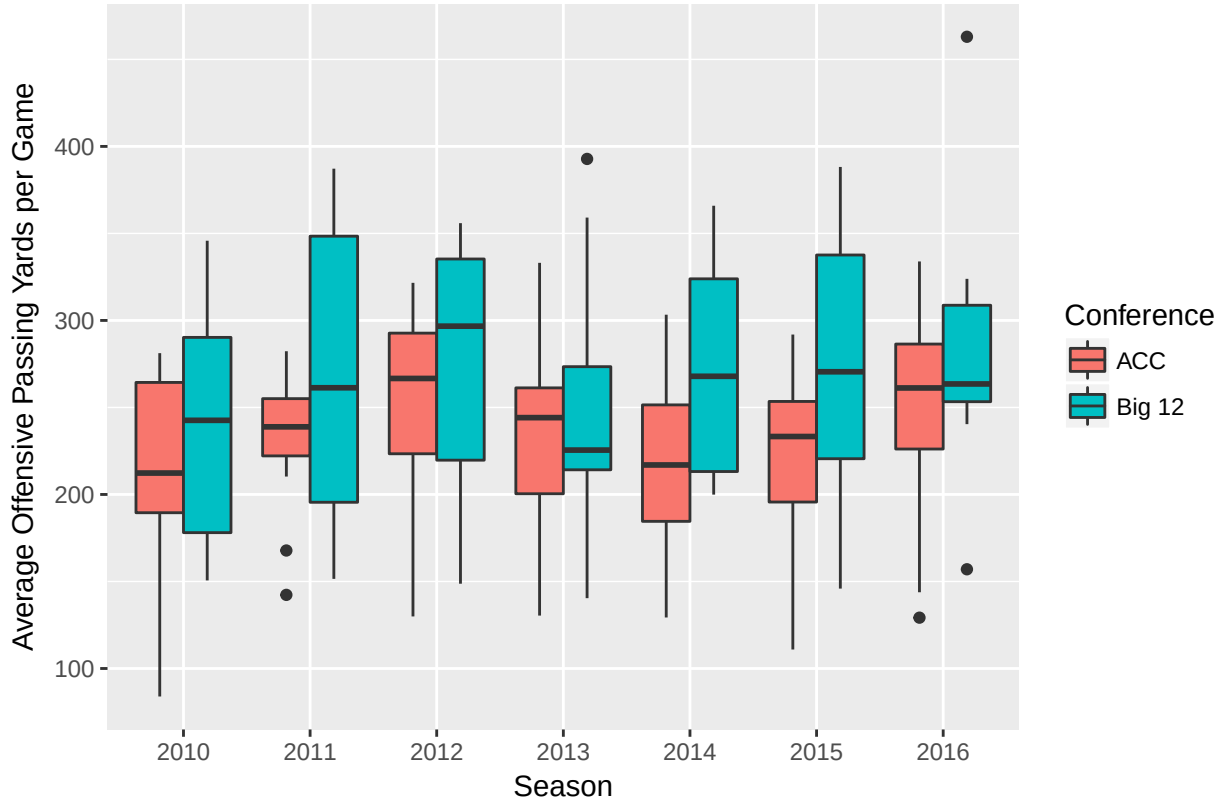


Figure 16: This graph shows how average offensive passing yards per game differs between the ACC and Big 12, as well as how the averages have changed from season to season.

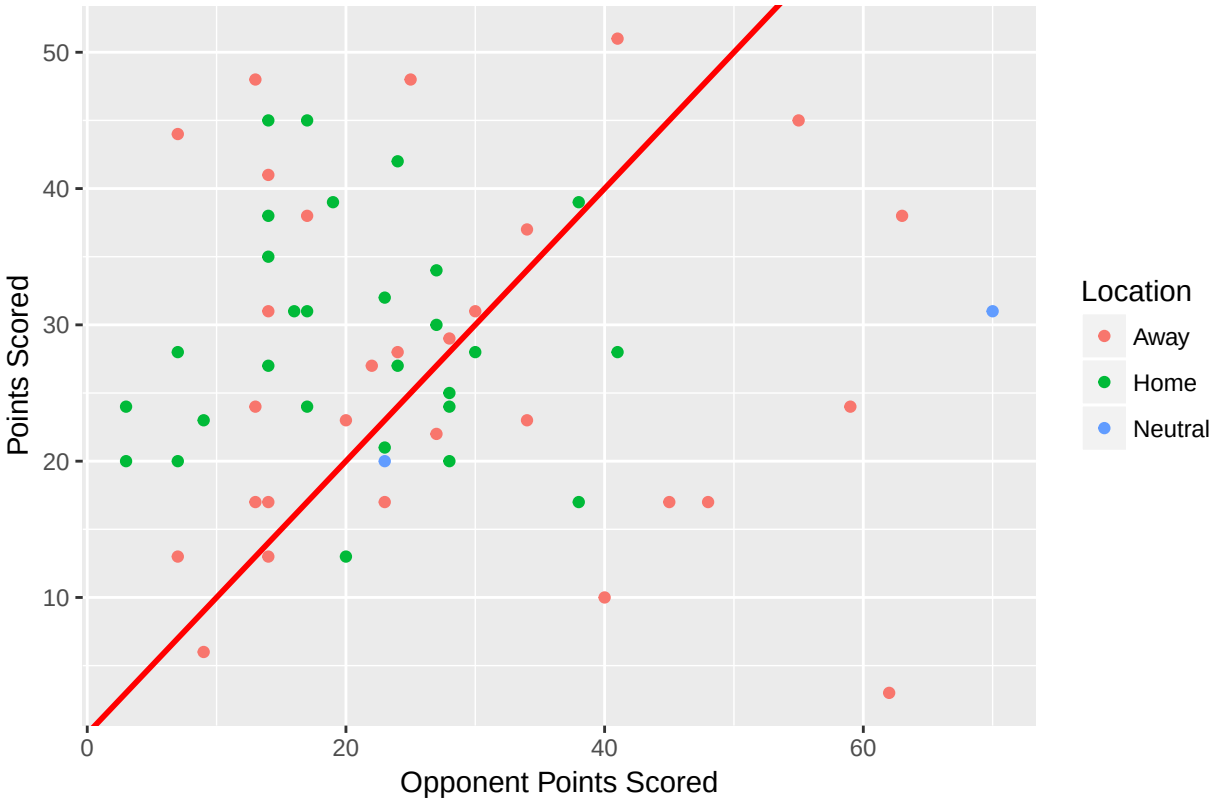


Figure 17: Scatterplot of Nebraska's points scored and Nebraska's points given up in conference games since 2000. Points above the line signify a win (where the points scored are greater than points given up) while points below the line signify a loss. The points are colored by location.

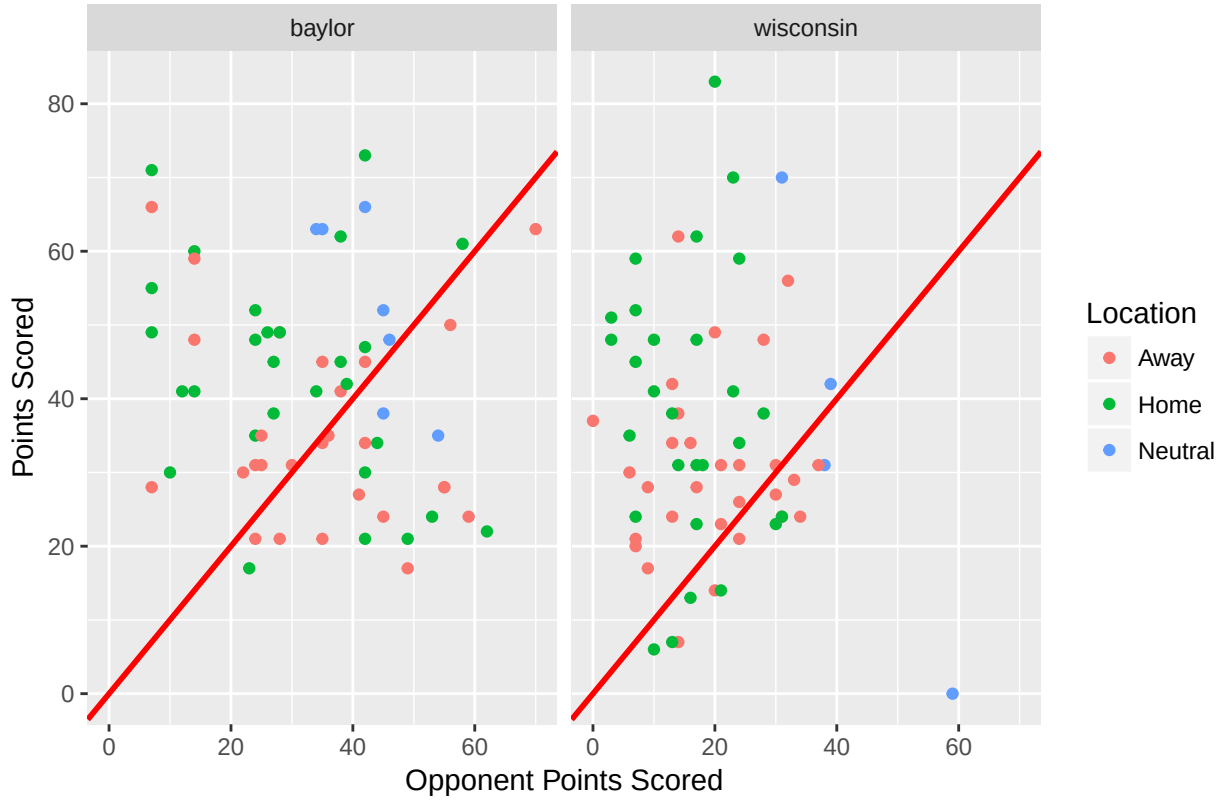


Figure 18: Scatterplots of team's points scored and team's points given up in conference games since 2000. This display makes it easy to get a quick snapshot of common scores for different teams. For example, Wisconsin's defense rarely gives up more than 40 points, and most losses generally have close scores. This is different than Baylor's football team, which gives up more than 40 points quite frequently and has losses that are not always close games.

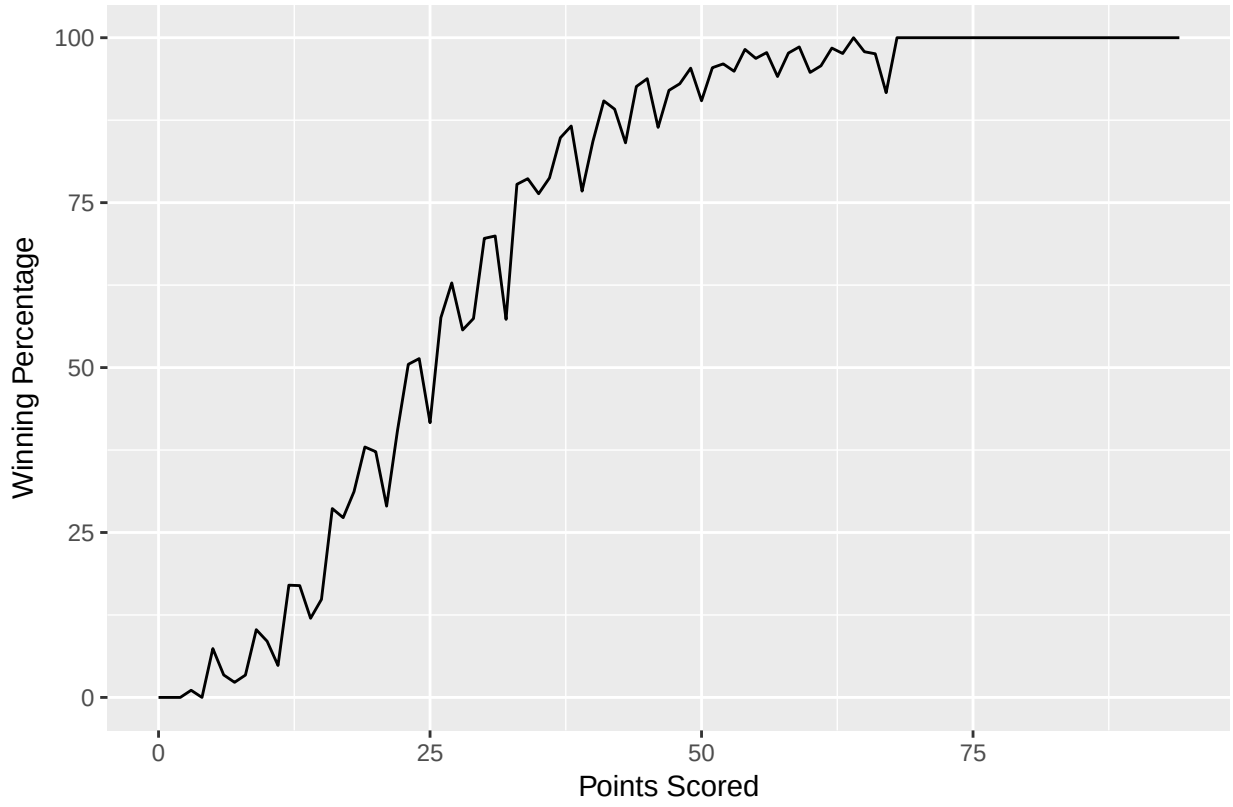


Figure 19: Winning percentage by points scored in games since 1990. The general trend seems to be that teams that score less than 25 points are more likely to lose, while teams that score more than 25 points are more likely to win.

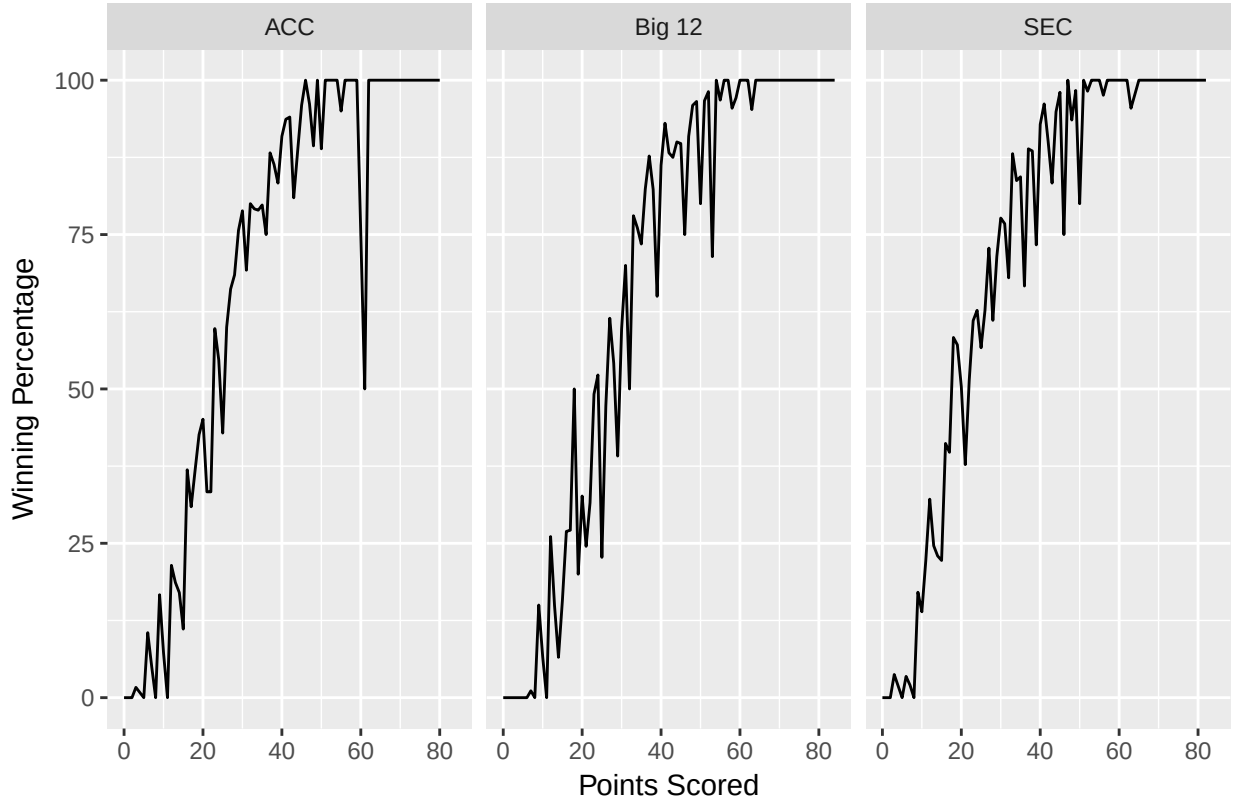


Figure 20: Winning percentage by points scored for ACC, Big 12, and SEC teams since 1990.

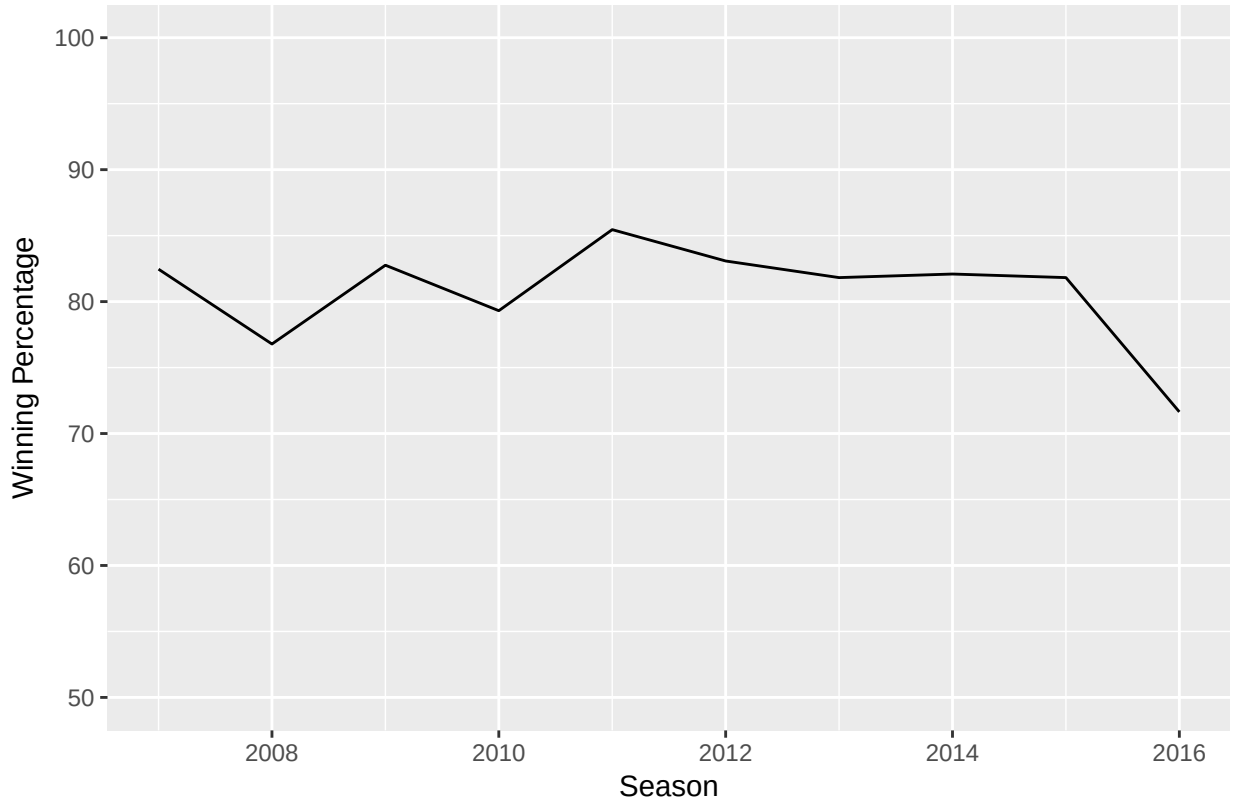


Figure 21: SEC winning percentage in non-conference games. This graph shows that the SEC has had a winning record (a winning percentage above 50 percent) in non-conference games over the past ten seasons, with a fairly sharp decline in the 2016 season.

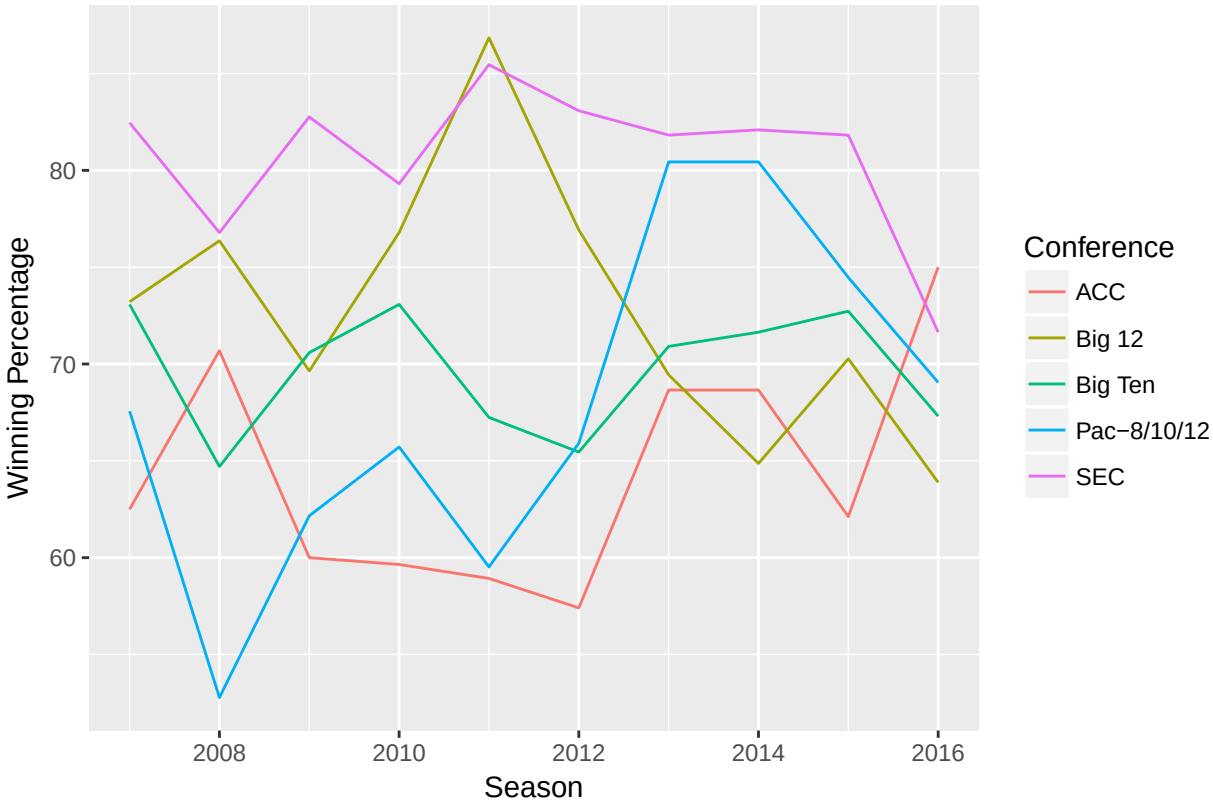


Figure 22: Comparing conference winning percentages in non-conference games. As seen in this graph, the SEC has had the best winning percentage in non-conference games except for two seasons: 2011 (when the Big 12 had the best winning percentage) and 2016 (when the ACC had the best winning percentage).

8.6 Game_Logs

8.6.1 File Description

8.6.2 Example Graphs

The Game_Logs data set allows for a more detailed look at each individual game. This allows for more detailed visualizations. One of the easier things to visualize is how certain variables may influence a team's winning percentage. Figure 25 shows how winning percentage varies by penalty yards, Figure 26 shows how winning percentage varies by turnover margin, and Figure 27 shows how winning percentage varies by offensive yards per play. Another interesting analysis someone can do is look at hexagonal heatmaps to observe the relationship between variables. For instance, Figure 28 shows the relationship between offensive yards per play and points scored. It is also possible to look at the relationship between variables across weeks. As an example, Figure 29 shows the relationship between a defense's yards given up per rush and the previous week's rushing attempts against that defense.

8.7 Data Set Conclusion

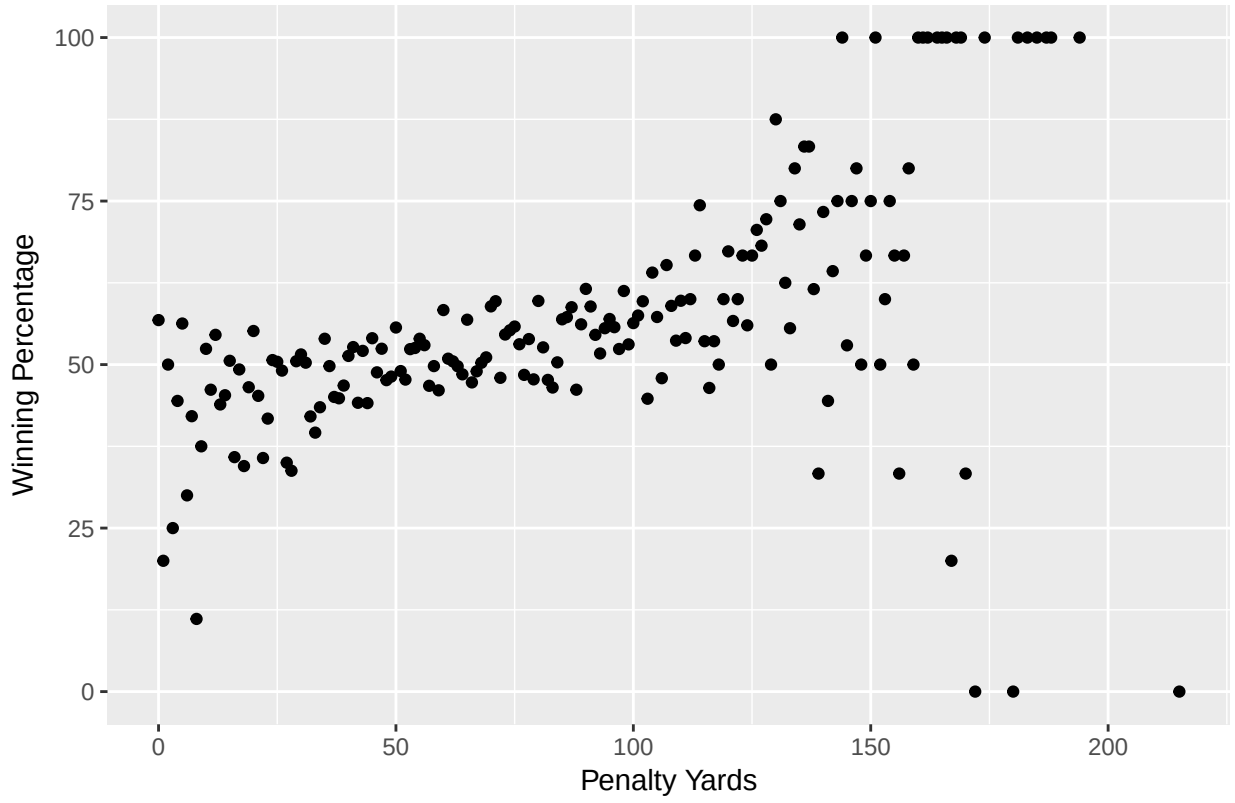


Figure 23: A scatterplot displaying how a team's penalty yards is related to winning percentage. It is not surprising that as teams gain more yards via penalty, the winning percentage increases. What is surprising is how flat the relationship appears to be (gaining 125 yards via penalty does not seem to drastically increase winning percentage from gaining only 50 yards via penalty).

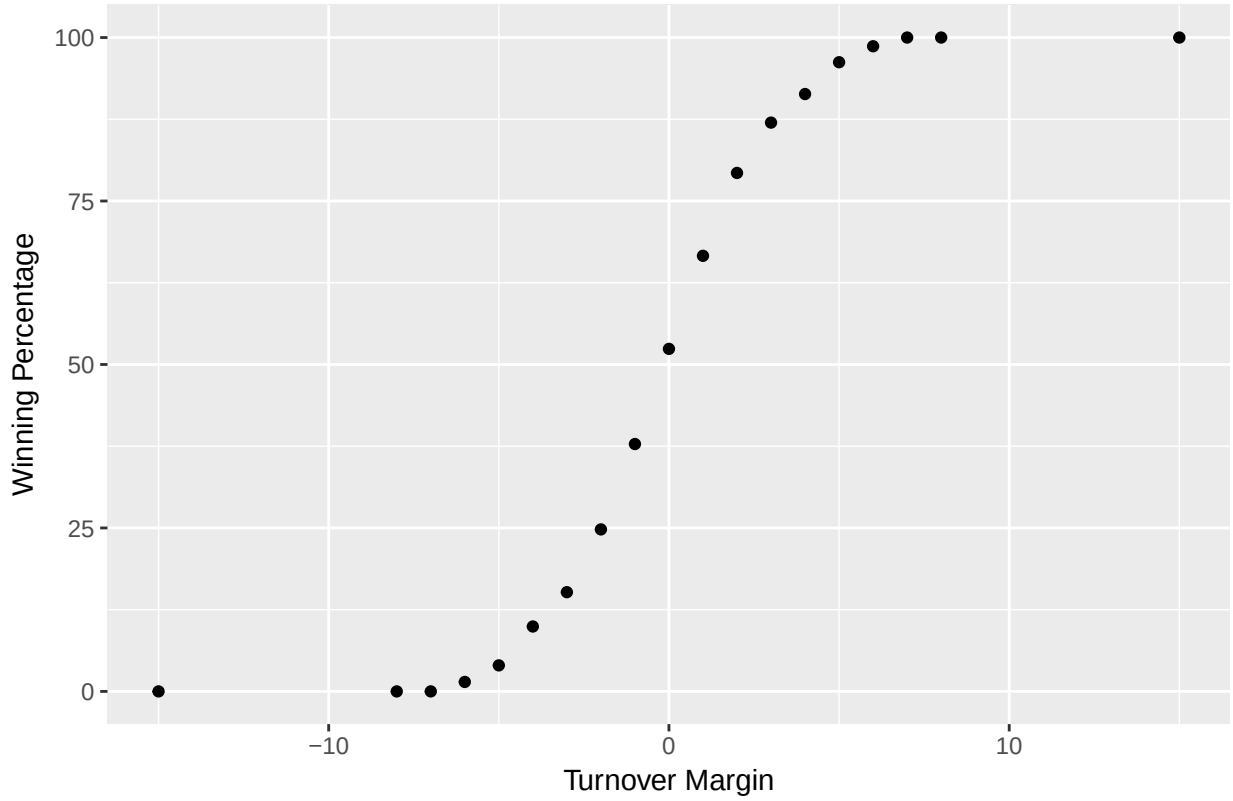


Figure 24: A scatterplot displaying how a team's turnover margin is related to winning percentage. Turnover margin is found by taking the Opponent's turnovers minus the Team's turnovers. It is not a surprise that as a Team commits less turnovers than their Opponent (resulting in a positive turnover margin), the winning percentage increases.

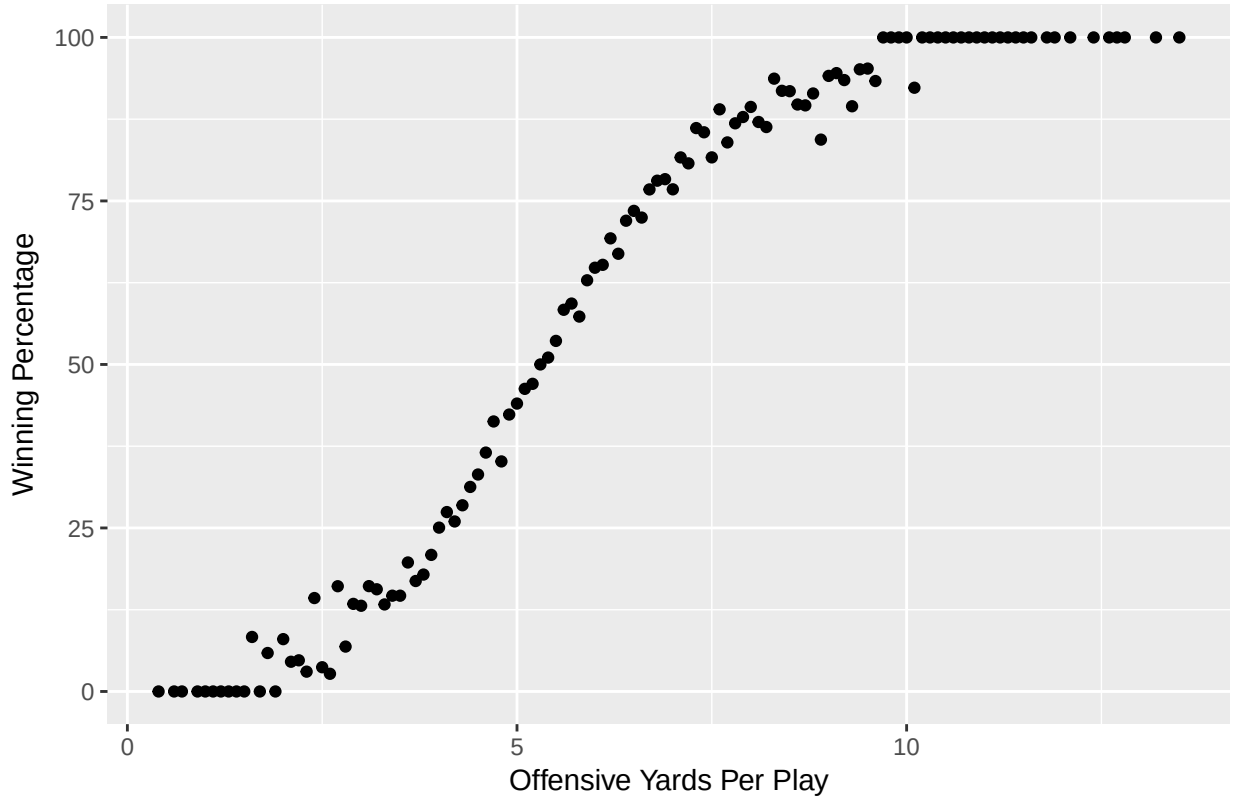


Figure 25: A scatterplot displaying how a team's offensive yards per play is related to winning percentage. It is not surprising that the more yards a team gains per play, the higher the winning percentage is. What is a little surprising is how strong the trend appears to be.

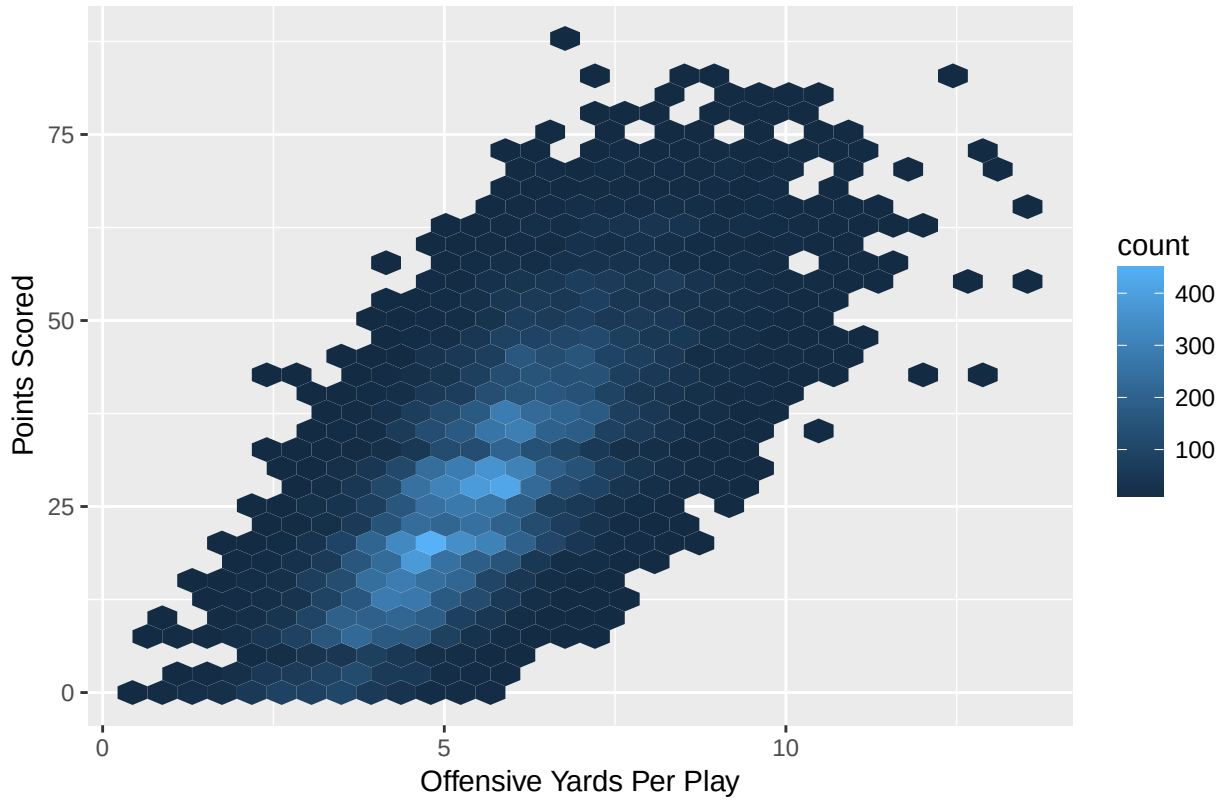


Figure 26: A graph showing the relationship between a team's offensive yards per play and points scored. Unsurprisingly, offenses that gain more yards per play tend to score more points.

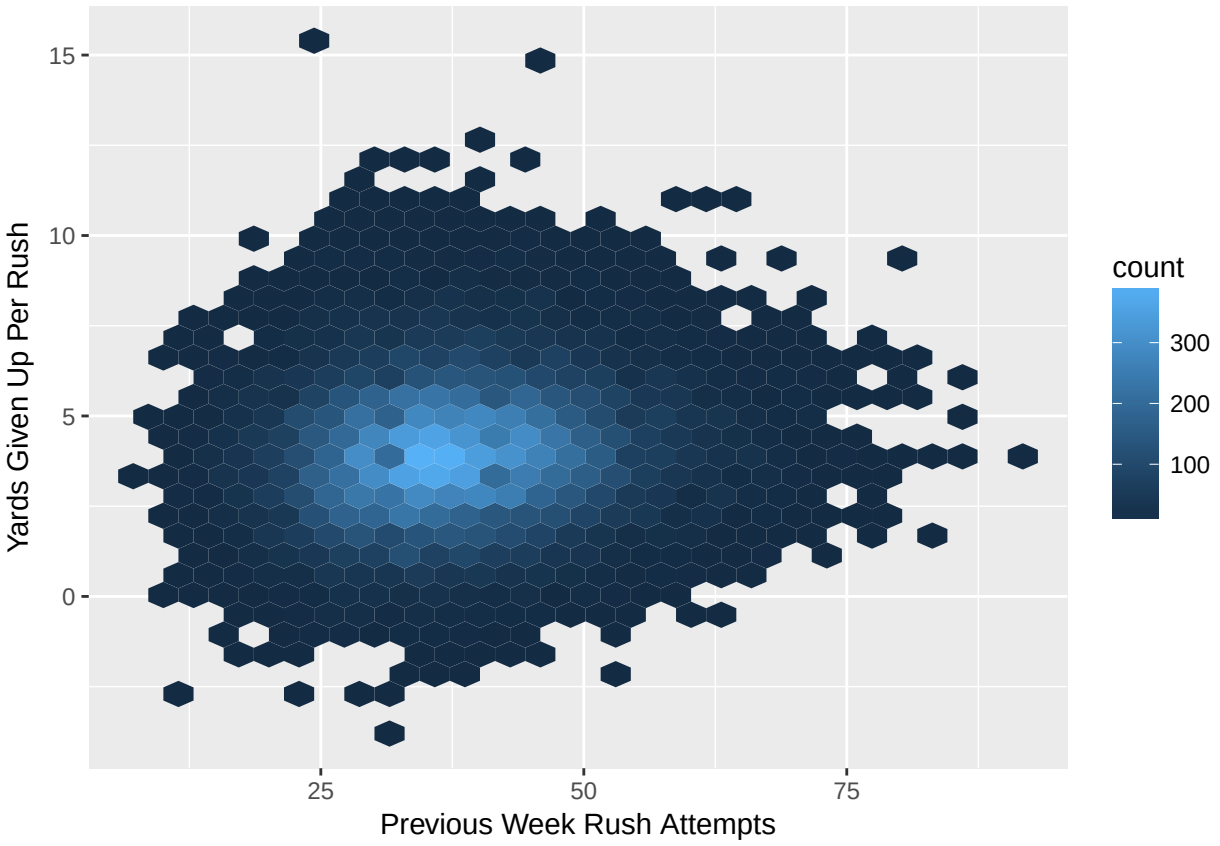


Figure 27: Relationship between a defense's yards given up per rush in a given week and the number of rushing attempts against that defense in the previous week. Commentators often claim that a defense that had many rush attempts the previous week may be less successful in stopping the run the following week. If this was true, one would expect to see a relationship between the previous week's defensive rushing attempts and the yards given up per rush the following week. The plot above does not seem to show an obviously strong relationship between these two variables.