Turning 50-50s into an Advantage

Characterizing How Nearby Players Affect Duels

Team Stat Sistahs (Taylor and Carly Venenciano)

Acknowledgements

The sources below were used for this project:

- Link to GitHub: https://github.com/tvenen/WSD2022
- UEFA.com link to publicly accessible data: https://www.uefa.com/uefaeuro/history/seasons/2020/statistics/
- StatsBomb data: https://github.com/statsbomb/StatsBombR

We would also like to thank Women in Sports Data Hackathon team for giving us an opportunity to participate in this event!

Introduction + Background

Research questions:

- How is a 1v1 duel outcome affected by the number of nearby players?
- Is there an optimal strategy for placing teammates nearby a 1v1 duel (i.e., transforming a 50-50 battle into a more favorable percentage)?
- Does a team's particular strategy for nearby players (+2, +1, 0, -1, or -2) in a 1v1 duel affect their performance in the tournament?

Important Definitions:

- 1v1 duel: a 50-50 contest between two players of opposing sides in the match
- Nearby players: teammates or opponents who are within 5 meters of the location of the duel, defined by the actor's location
- Types of nearby player duel situtations:
  - +2: two more teammates than opponents
  - +1: one more teammate than opponent
  - 0: equal number of teammates and opponents
  - -1: one less teammate than opponent
  - -2: two less teammates than opponents

We investigated the "duel" type as it occurs 1.37% of the time in all events, in comparison to "50/50" which occurs 0.05% of the time. There are approximately 20-30 duels in one game. Duel outcomes of "won", "success in play", and "success out" were set as a "win", and the "lost in play" and "lost out" were set as a "loss". We chose to exclude aerial duels as their duel outcome was not recorded, so our findings only relate to ground tackles. The duel win percentages for -2, -1, 0, 1, and 2 ('min2', 'min1', 'zero', 'plus1', and 'plus2') are calculated by dividing the number of "win" outcomes by the total amount of duels. The development_code.Rmd file in the GitHub shows the processes used to calculate the duel win percentages. We then manually input the 'max_value_of3' and 'min_value_of3' columns based on the maximum or minimum win percentage type (either -1, 0, or +1) for both the by_country and by_match spreadsheets. We only looked at -1, 0, and +1 as there was limited data for +2 and -2. Data analysis is separated into two sections: by country and by match. There are 24 countries (teams) who played in the tournament, and 51 matches with two teams playing in each for a total of 102 observations in the match section.

Purpose and Motivation

In my past three years of collegiate soccer, my team has always set the same collective goal to win the first 50-50. We thought this would set the pace and intensity level of the game and thus help us to win the game. Winning 50-50s inspired our project for Women in Sports Data.

Summary of Results (expanded on in Full Results section)

A framework for investigating duel win percentage in various scenarios (-2, -1, 0, +1, +2) was developed. Data analysis on duel win percentages and other metrics of success in the tournament and individual matches was completed. The most notable result of the data analysis suggests that teams with a higher duel win percentage when there are an equal number of teammates and opponents nearby, as opposed to when there is one more teammate than opponent nearby, are overall more successful in the tournament.

When to use this recommendation

The framework can be used by data analysts to further investigate the impact of duels on a team's success in tournaments and in individual matches. The result (teams with a higher duel win percentage when there are an equal number of teammates and opponents nearby, as opposed to when there is one more teammate than opponent nearby, are overall more successful in the tournament) can be directly applied in trainings and tournaments as a team strategy. Teams should focus on improving their duel win rate when there are an even number of teammates and opponents (i.e., training more often in a 2v2 situation as opposed to a 1v2 or 2v1). By improving in this even-numbers duel, teams can expect to have more success in the tournament.

Difficulties + Challenges Faced

The main difficulty we faced was deciding how to determine a team's success in the tournament. There was no metric in the StatsBomb data nor in general soccer culture that directly measured how well a team did in a tournament considering both outcome and importance of match should be weighted appropriately. For example, a team that won one game in the group stage and advanced to the quarter-final round did better in the tournament than a team that won all 3 games in the group stage but lost in the round of 16, but a standard points metric does not measure this tournament-advancement success accurately. We ended up developing the 'perc_wstand', but further development could be useful in producing a better metric. Another challenge we faced was collaborating effectively as a team in a remote environment with different time zones and schedules. Our team ended up only being two of the initial three, and one person ended up doing most of the work for the project.

Further extensions

We used a set distance of 5 meters to define nearby players, but a further extension to this project could look at distances of 3 or 7 meters. It may be interesting to see if duel win percentages increase or decrease depending on how close the nearby players are. Another extension is whether the number of nearby teammates or opponents affect duel win percentages, as opposed to the difference in teammates and opponents which we investigated. Lastly, aerial duels (if their outcome was recorded) could be investigated to see if there are similar duel win percentages to the ground tackles.

Full Results

The full spreadsheets (located in the GitHub) are read into this data analysis file below:

```
stats_by_country <- read.csv("by_country_maxmin.csv")
stats_by_match <- read.csv("by_match_maxmin.csv")
```

By Country

In this section, the ‘stats_by_country’ spreadsheet is statistically analyzed. Matches, goals, possession, and number of matches won (‘match_won’) was found online at UEFA.com. From those values, goals per game and percentage of matches won (‘perc_match_won’) were calculated. Points (with data from UEFA.com) is calculated in the typical soccer fashion: 3 points for a win, 1 point for a tie, and 0 points for a loss. Point percentage (‘perc_point’) is a normalized quantity where each country’s points is simply divided by the maximum (17, Italy and England). ‘Standing’ is determined by a team’s progress in the tournament: 5 points for champion, 4 points for finalist, 3 points for semifinalist until 0 for exit at group stage. ‘Wstand’ is created by adding ‘points’ and ‘standing’, and ‘perc_wstand’ is normalized by dividing each country’s value by the maximum (Italy, 22). Thus, ‘perc_wstand’ best accounts for how well a team performed because it considers a team’s progression in the tournament. The max and min values (-1, 0, or +1) were manually inputted as stated above. Inconclusive results were found between duel win percentage and success in the tournament, and significant results were found between a team’s best duel type and their success in the tournament.

Does the win percentage for a certain duel type (-2, -1, 0, +1, +2) have an effect on how well a team did in the tournament?

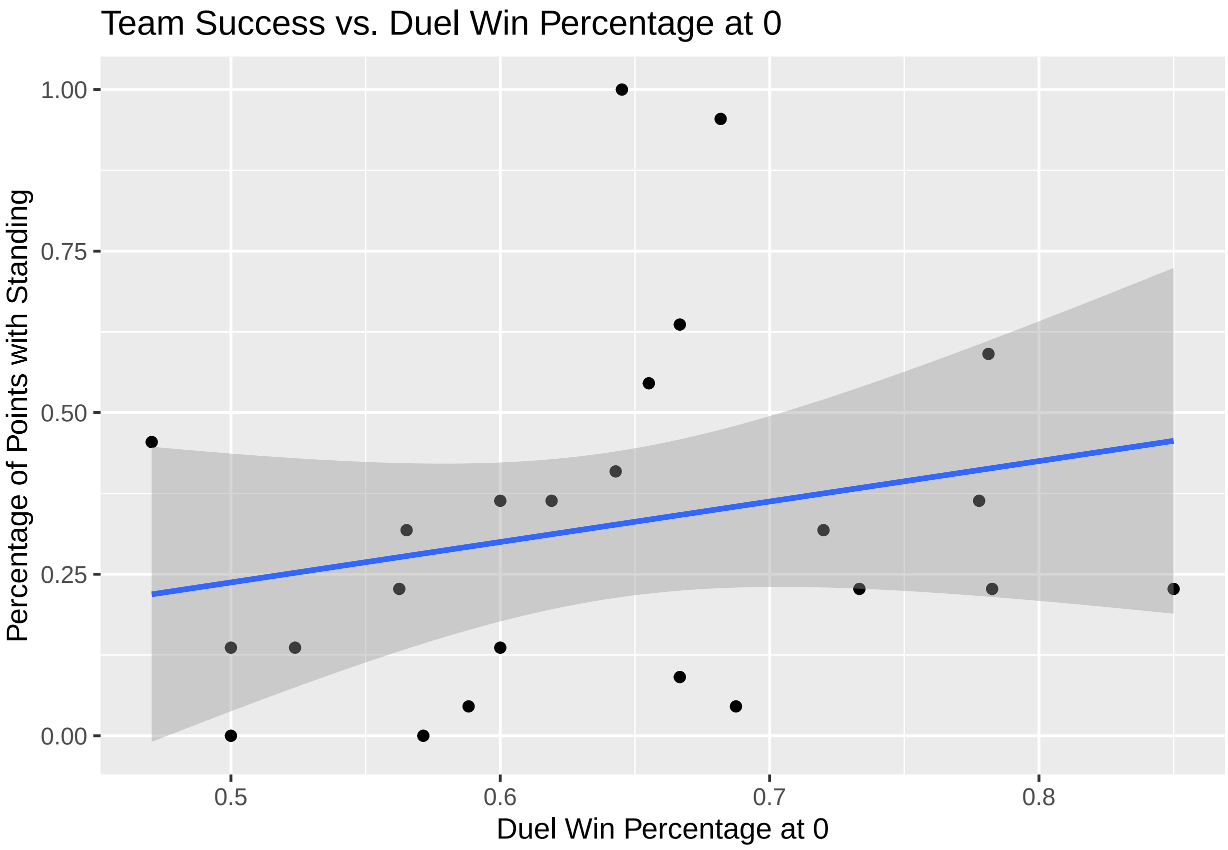
Linear models were fit to each duel type’s win percentages and the team’s success measured by ‘perc_wstand’. One example is shown below with the 0 duel situation win percentages and ‘perc_wstand’:

```
my_data = stats_by_country[!is.na(stats_by_country$zero), ]

x_val = my_data$zero
y_val = my_data$perc_wstand

ggplot(my_data, aes(x = x_val, y = y_val)) + geom_point() + geom_smooth(method = 'lm') + xlab('Duel Win
```

```
## `geom_smooth()` using formula 'y ~ x'
```

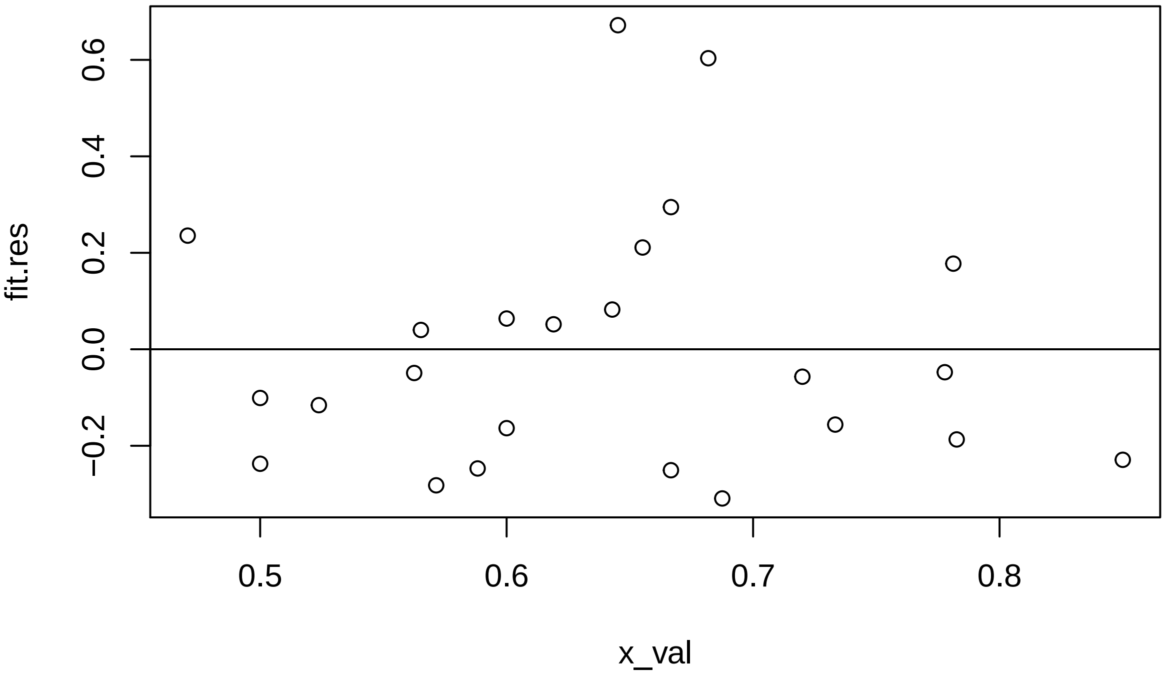


```
lm_fit <- lm(y_val ~ x_val, data = my_data)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = y_val ~ x_val, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30921 -0.19748 -0.05299  0.10617  0.67184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0757     0.3633  -0.208    0.837
## x_val         0.6260     0.5601   1.118    0.276
##
## Residual standard error: 0.2675 on 22 degrees of freedom
## Multiple R-squared:  0.05373,    Adjusted R-squared:  0.01072
## F-statistic: 1.249 on 1 and 22 DF,  p-value: 0.2758
```

```
fit.res = resid(lm_fit)

 plot(x_val, fit.res) + abline(0, 0)
```



```
## integer(0)
```

The residual plot shows that the residuals are larger in the positive region than the negative one, and the adjusted R^2 value is 0.01072. With these observations and a visual inspection of the linear model on the data in the graph above, the linear model is not perfect. This trend of imperfect linear models is consistent when looking at the -2, -1, and +1 data compared to 'perc_wstand.' The +2 data had binary values (0 or 1) so was not used. The p-value of the slope, which measures how much 'perc_wstand' changes with an increase in duel win percentage at the 0 situation, is also 0.276 which is inconclusive. The statistics are listed below from each of the models:

Duel Type	Slope	Intercept	Adjusted R^2	p-value of slope
-2	0.053	0.314	-0.050	0.821
-1	0.257	0.185	-0.033	0.610
0	0.626	-0.076	0.011	0.276
+1	-0.252	0.490	0.006	0.301

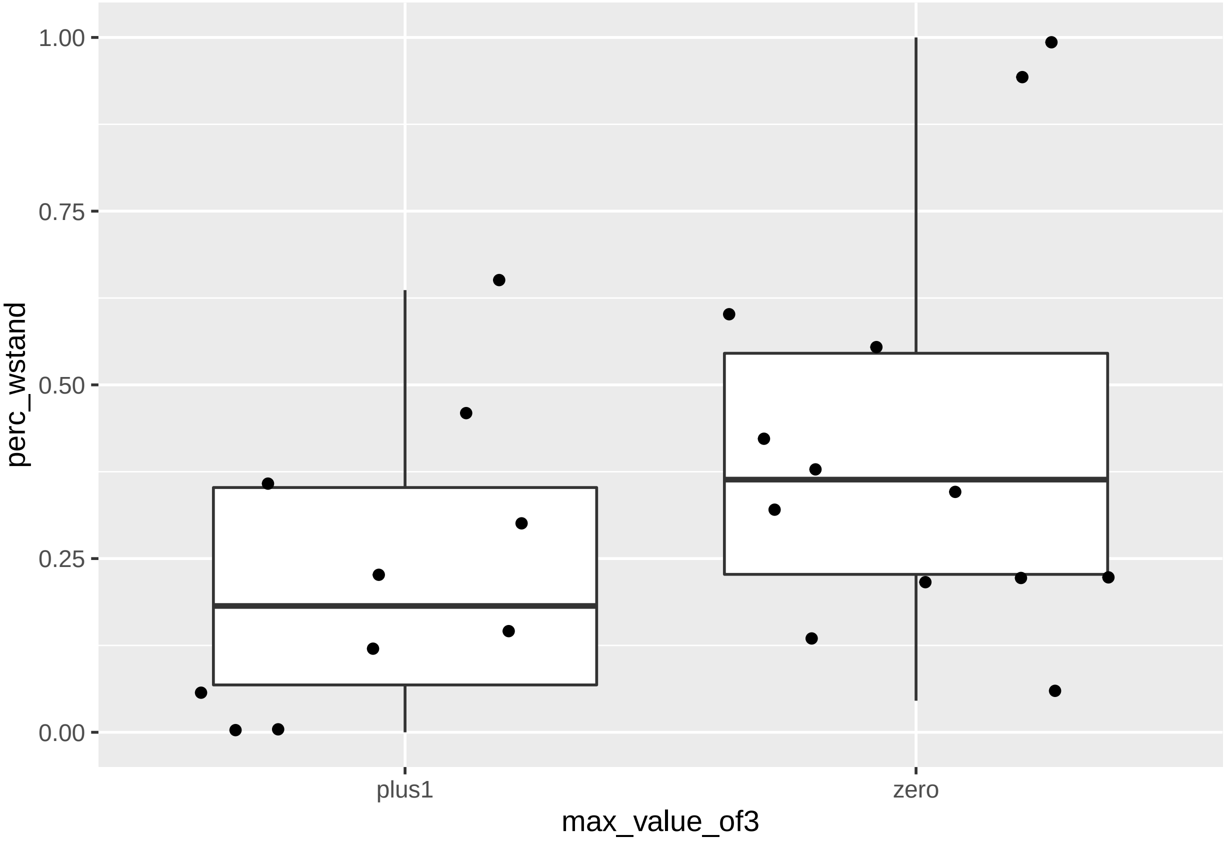
Simply looking at the slopes, it seems that the 0 duels (teammate to opponent ratio is 1:1) have the largest effect on how well a team did in the tournament. Thus, the higher a team's win percentage in a 0 situation throughout the tournament, the better they did overall. Less of an effect is seen in the -1 and -2 as they have smaller slopes, and interestingly, the +1 duels had a negative effect on how teams did in the tournament. These results should be interpreted with caution as the linear fits are not great; however with more data, there is a chance that these slopes become significant.

Does a team's best duel type (-1, 0, +1) affect how well they do in the tournament?

Only one team (Hungary) had -1 as their highest duel win percentage thus analysis was limited to comparing teams that had their best win percentage in the +1 or 0 situation. T-tests between 'max_value_of3' and possession, percentage of matches won, and goals per game for each team yielded large p-values. Therefore, there was no real significance between a team's best duel type and possession, match win percentage, or goals per game throughout the tournament.

However, using the code below, significant p-values (less than 0.1) were found between best duel type (0 or +1) and ‘perc_wstand’, ‘standing’, and ‘perc_points.’ As ‘perc_wstand’ best accounts for a team’s success in the tournament as explained previously, those results are below.

```
stats_by_country %>%
  filter(max_value_of3 == 'zero' | max_value_of3 == 'plus1') %>%
  ggplot(aes(x=max_value_of3, y=perc_wstand)) + geom_boxplot() + geom_jitter()
```



In the above box plot, it is evident that the mean of ‘perc_wstand’ of the teams with a better 0 duel win percentage is higher than those with a better +1 percentage. The following T-test confirms this:

```
mydata_h = stats_by_country %>%
  filter(max_value_of3 == 'zero' | max_value_of3 == 'plus1')

t <- t.test(mydata_h$perc_wstand ~ mydata_h$max_value_of3, alternative = 'less')
t
```

```
##
##  Welch Two Sample t-test
##
## data:  mydata_h$perc_wstand by mydata_h$max_value_of3
## t = -1.7617, df = 20.952, p-value = 0.04636
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.00426217
## sample estimates:
```

```
## mean in group plus1  mean in group zero
##           0.2318182           0.4160839
```

Thus with a p-value of 0.0436, there is strong evidence that the true difference in means of 'perc_wstand' between the +1 and 0 teams is less than zero. The two-sided p-value is 0.09271 which is also moderately significant. In terms of soccer strategy, this means that teams with a higher duel win percentage when there are an equal number of teammates and opponents nearby as opposed to when there is one more teammate than opponent nearby are more successful in the tournament.

By Match

In this section, the 'stats_by_match' spreadsheet is statistically analyzed. Goals for and against are from the StatsBomb data. Goal difference is calculated by taking the difference, and 'outcome' (win, loss, tie) is determined. The -2, -1, 0, +1, and +2 columns are calculated as explained above. No significant results were found as the by match data is limited by the number of duels recorded in one match and many of the duel win percentages are 0 or 1. However, the framework below is included in the project to provide a basis for future investigation with more matches and more duels.

Does a certain duel type (-2,-1 0, +1, +2) have an effect on a team's goal differential in a match?

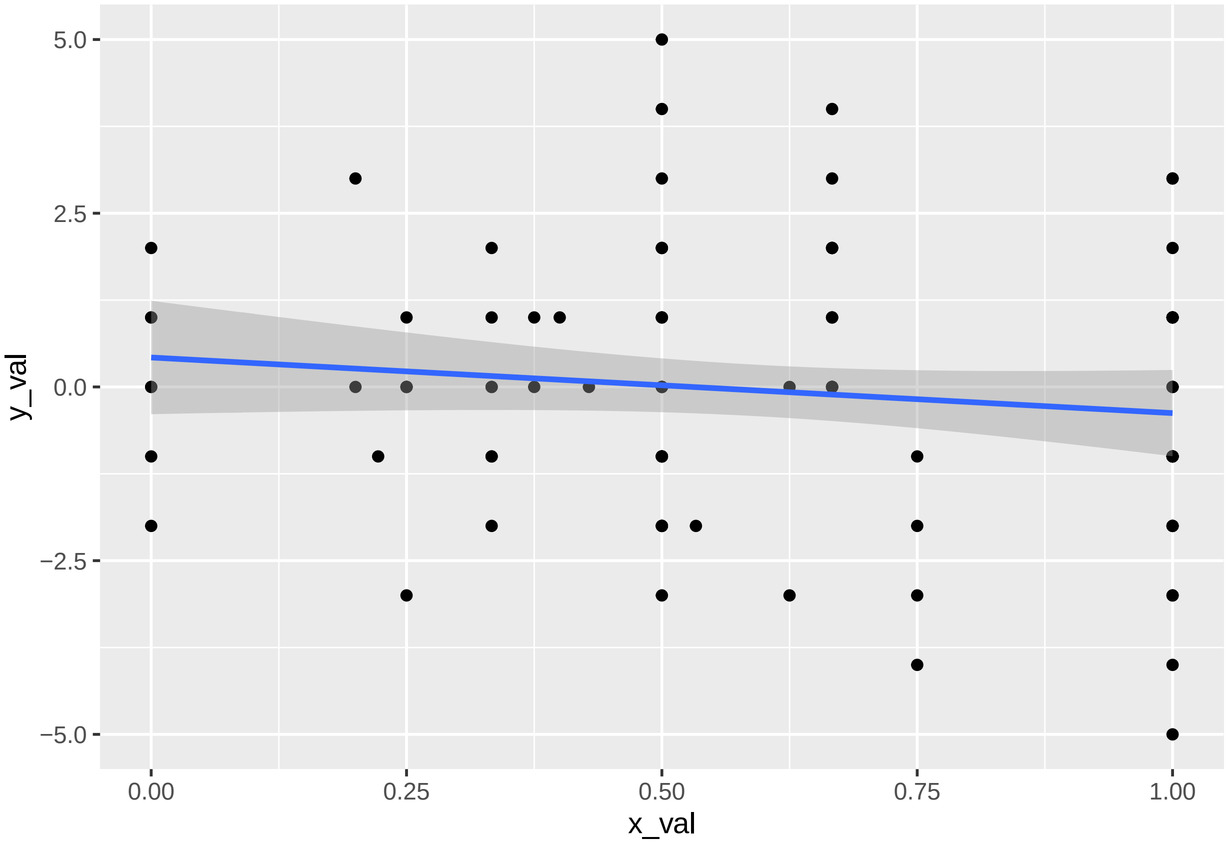
The code below was run on each -2, -1, 0, +1, and +2 data to see if the win percentages of a specific duel had an effect on a team's goal differential in a match. No significant slopes were found for the five duel types.

```
my_data = stats_by_match[!is.na(stats_by_match$zero), ]

x_val = my_data$zero
y_val = my_data$goal_diff

ggplot(my_data, aes(x = x_val, y = y_val)) + geom_point() + geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



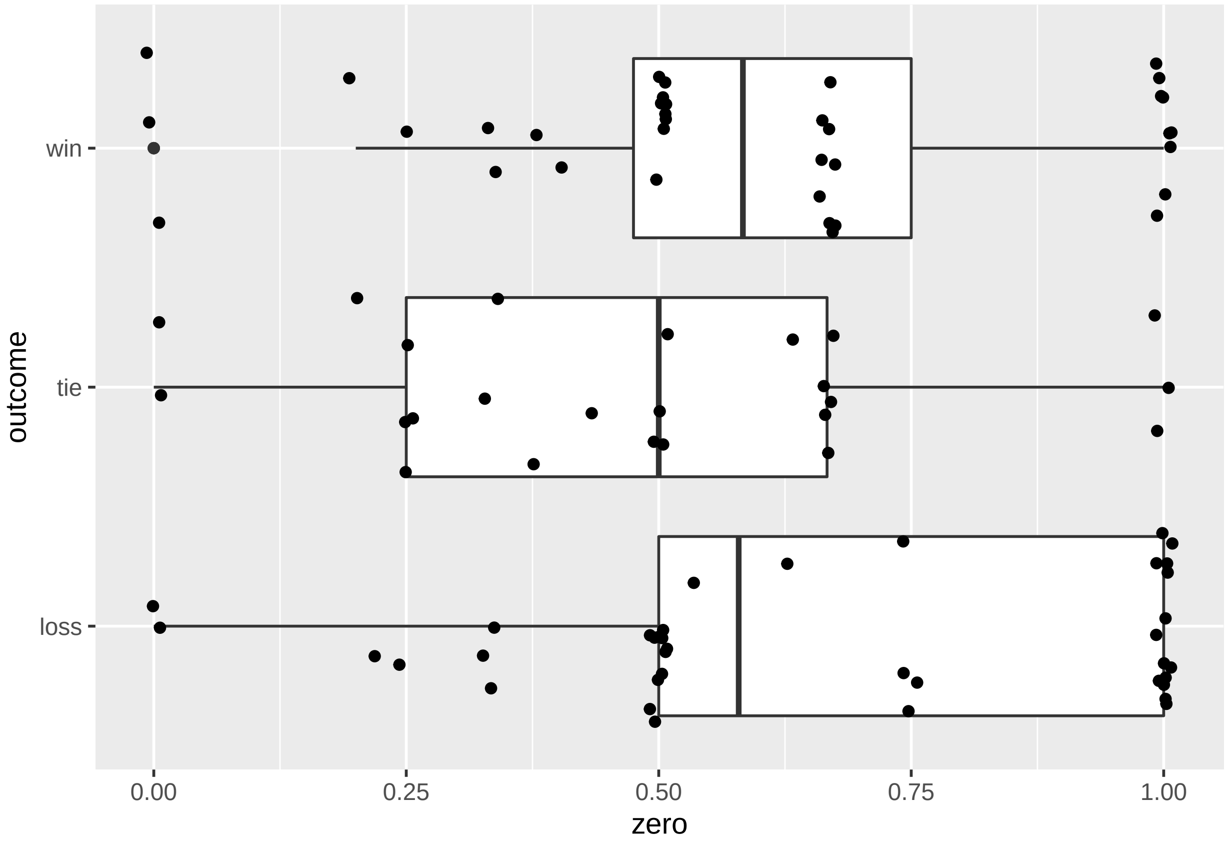
Does a certain duel type (-2,-1 0, +1, +2) have an effect on if a team won a game (outcome)?

The code below was run on each -2, -1, 0, +1, and +2 data to see if the win percentages of a specific duel had an effect on the outcome of the game.

```
stats_by_match %>%
  ggplot(aes(x=zero, y=outcome)) + geom_boxplot() + geom_jitter()
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



Due to the large range of values, no significant relationship was found in a T-test between any of the duel types and the outcome of the game.

Does a team's best duel type (-1, 0, +1) have an effect on if a team won a match (outcome)?

A Chi-squared test is run on the 'max_value_of3' and 'outcome' variables.

```
dat = stats_by_match
table(dat$max_value_of3, dat$outcome)
```

```
##
##         loss tie win
##   min1     8   6  10
##   plus1    7   9   8
##   zero    10   5   9
```

```
test <- chisq.test(table(dat$max_value_of3, dat$outcome))
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(dat$max_value_of3, dat$outcome)
## X-squared = 2.0822, df = 4, p-value = 0.7206
```

With a p-value of 0.7206, there seems to be inconclusive evidence on whether a team's best duel type affects the outcome of the game.