Data

# Description of Covariates

There are several important variables that may impact momentum. We have divided these covariates into three categories: *structural* covariates, *catalytic* covariates, and *reducing* covariates. Structural covariates define the structure and natural conditions of a match. These are the rules under which the players play. Catalytic covariates are events within a match that we hypothesize spark momentum. Reducing covariates are events or rules that can slow or reset momentum.

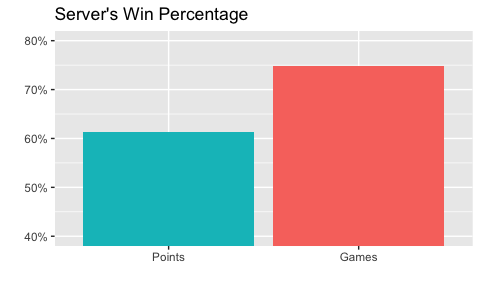
## Structural Covariates

The structural covariates define the normal conditions. As a rule, any variable that directly impacts a players’ probability of winning a point for the entire duration of a match, regardless of the score, is structural. Important structural variables include **server**, **sex**, **bet\_odds**, and **tournament**. *[Maybe include: year, age, height, handedness (player effects)]*.

One of the difficulties in momentum literature is isolating momentum effect from a mere asymmetry in player skill (Meier et al., 2020). These structural covariates help control for the non-momentum forces that impact a match.

### Server

One of the most powerful shots in the game of tennis is the serve. Players alternate service games and they hold a large advantage while serving. Over the course of our data, players won 61.41% of points while serving and 74.78% of the games while serving.

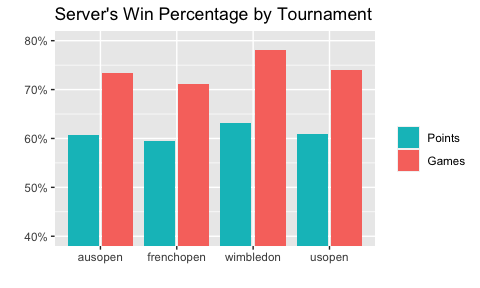


The indicator variable **server** holds a value of 1 when player 1 serves and 2 when player 2 serves.

### Tournament

Each of the four Grand Slams hold different characteristics that affect the player’s style of play. The surface of the court and common weather conditions can cause the same shot to travel at different speeds and spin rates. Generally, the conditions at Wimbledon are fastest and those at the French Open are slowest. However, these exact conditions vary from year to year. They are even impacted by the time of day.

These conditions most noticeably impact the players’ dominance on the serve.



These changes in conditions lead to a slightly different overall style of play. Rallies are shorter at Wimbledon. Serve speeds and winners are higher at Wimbledon and lower at the French Open.

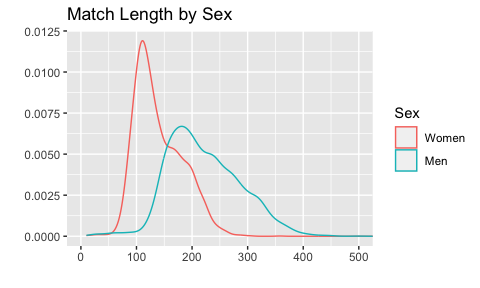
| **Tournament** | **Rally Length** | **First Serve Speed** | **Second Serve Speed** | **Ace Rate** | **Winner Rate** |
| --- | --- | --- | --- | --- | --- |
| Australian Open | 3.85 | 107.59 | 88.52 | 7.04% | 32.86% |
| French Open | 4.13 | 106.13 | 87.73 | 4.68% | 32.30% |
| Wimbledon | 3.08 | 110.62 | 92.67 | 7.69% | 32.98% |
| US Open | 3.70 | 107.13 | 87.95 | 6.87% | 31.63% |

We will use the categorical variable **tournament** to describe this condition.

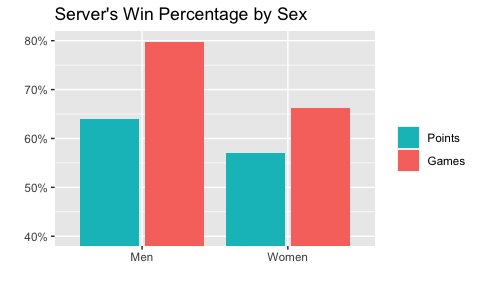
### Sex

In general, the men’s and women’s games of tennis are very similar. In the Grand Slams, men play a best-of-five set match while women play a best-of-three set match. Otherwise, the conditions are exactly same. Still, this small difference impacts our study of trends in momentum over time. Men play more points over the course of a match, so there are more opportunities to observe changes in momentum.

Men play an average of 221.6 points per match and women play an average of 139.9 points per match. Thus, while 49.1% of our data set are women’s matches (3885 matches), only 37.8% of the points are played by women (543284 points).



While the rules are nearly identical, the characteristics of the men’s and women’s games differ in the relative dominance of the serve. Men tend to win a higher proportion of points while serving than women. Women still have an advantage on their serve, but it is much lower.



The indicator variable **sex** holds a value of 0 in women’s matches and 1 in men’s matches.

### Betting Odds

Pre-match betting odds are maintained by several sportsbooks for all Grand Slam tennis matches. We selected two sportsbooks, Bet365 and Pinnacle Sports, that employ the European (Decimal) Odds method. Their format is relatively simple. The decimal odds is the value a better receives for a $1 bet. A better wagering $100 on a match with decimal odds at 1.01 wins $101 for a $1 profit, while a better wagering $100 on a match with decimal odds at 2.00 wins $200 for a $100 profit. Thus, small decimal odds are safe bets on likely winners and large decimal odds are risky bets on unlikely winners.

Fortunately, decimal odds are also easily transferred into an implied probability percentage between 0 and 1 with the equation:

Sportsbooks have a strong incentive to precisely adjust their betting odds to reflect realistic pre-match win probabilities (*cite*?), so it is safe to assume that the betting odds represent the two players’ strength entering a match.

We performed three logistic regressions to assess the predictive power of Bet 365’s betting odds, Pinnacle Sport’s betting odds, and the player’s rankings the outcome of the matches in our data set. In total, the data set includes 7711 matches (51.58% men) across 39 tournaments.

We found that a logistic

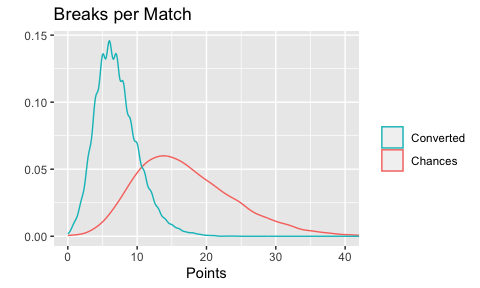
*Here, I will visualize the betting odds covariate we select and describe any transformations we need to make to the variable*

## Catalytic Covariates

The catalytic covariates are potential catalysts of momentum. We hypothesize that these decisive and influential moments in a match provide a boost to players. These variables are events that take place within the bounds play, important point characteristics, or breaks in play that may disrupt natural flows of momentum. We include **break\_converted**, **break\_saved**, **p1\_winner**, **p1\_unforced\_errors**, **p1\_distance\_run**, **rally\_count**, **interruption**, and **change\_ends**.

### Break Converted

Converting a break point is one of the keys to winning a match (Klaasen & Magnus, 1998). It has been found to cause positive momentum for the breaking player (Meier et al., 2020). The average match has 17.0 break chances and 7.1 converted break points.

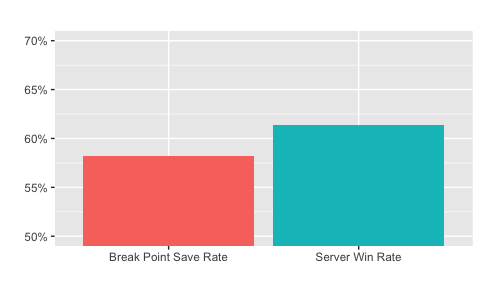


The indicator variable **break\_converted** holds a value of 1 the point a break occurs and a value of 0 otherwise.

### Break Saved

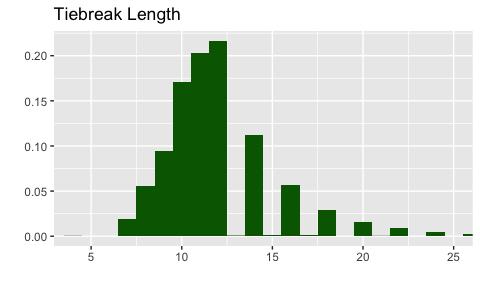
*convert to total\_break\_saved?*

Each time a player misses an opportunity to convert a break point, his or her opponent successfully saves a break point. While saving a break point does not lead to a positional or strategic advantage, we hypothesize that it could cause a recurring psychological advantage. Servers save break points about 58.19% of the time. This is notably less than their overall win rate on serve (61.14%).



### Tiebreak Victory

If two players are tied after twelve games of play in a set, they will enter into a tiebreak. Like break chances, tiebreaks have an asymmetric impact on the outcome of a tennis match. Winning a tiebreak can award a significant momentum effect on the rest of a match (Page & Coates, 2017). Tiebreaks are relatively uncommon; only 14.11% of sets end in a tiebreak. The average length of a tiebreak is 11.85 points.

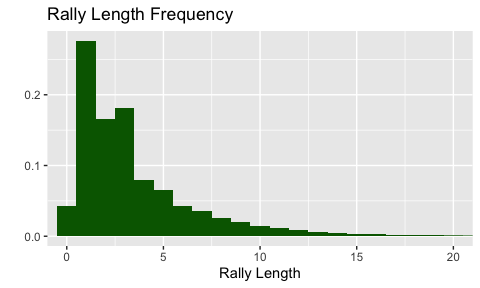


The variable **tiebreak\_victory** holds a value of 1 if player 1 wins a tiebreak, 2 if player 2 wins a tiebreak, and 0 otherwise.

### Rally Length

Not all tennis points are equal. Some last a few shots and players exert very little energy, while others are more grueling rallies that cost players significant amounts of energy. We hypothesize that as players exert more mental and physical energy during a single point, the point becomes more impactful.

The rally length adds up the total number of shots in a single point. In general, serves and made shots count but errors do not. A double fault (server misses serve) is a 0 shot rally and an unreturned serve qualifies as a 1 shot rally. A point with a serve, three completed shots, and an error has a rally length of 5 shots. The average rally length is 3.53 shots, but 67% of rallies are three shots or fewer.

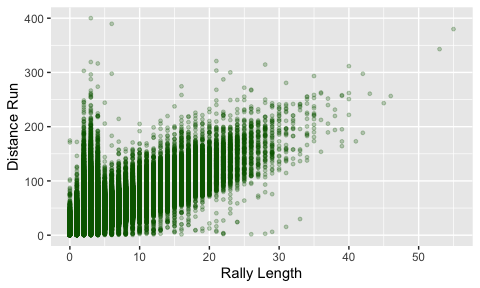


We will use the **rally\_count** variable to hold the length of rallies.

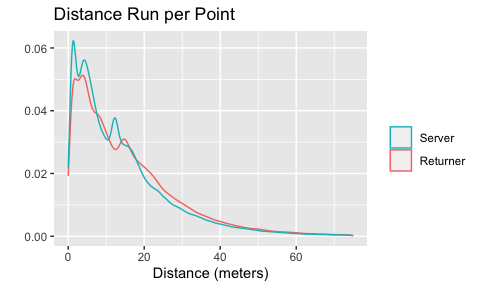
### Distance Run

The players’ distance run is a second measure of the mental and physical energy exerted on a given point. **Distance\_run** is measured in meters and is generally correlated with **rally\_count**. Below is a scatter plot of the total distance run by rally count.

The scatterplot shows that there are some accuracy issues with both variables. For example, if **rally\_count** is 0, then the distance run should also be close to 0. It is unlikely for the distance run to be close to 0 if the rally count is large. Perhaps most concerning is the cluster of points with rally lengths of 3-5 and distance runs over 100. The majority of these observations came from the US Open in 2021 and Wimbledon in 2021 and 2023.



In general, the distance run of the two players in a point is highly related. The average distance run by the server is 14.48 meters and the average distance run by the returner is 16.07 meters.



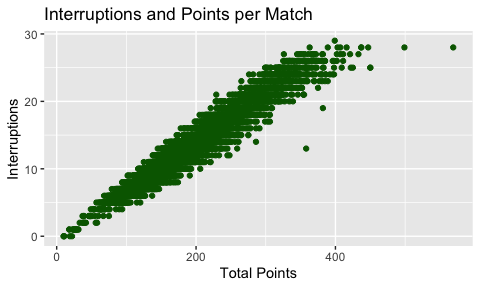
## Reducing Covariates

### Interruption

In the standard structure of a tennis match, players are afforded regular breaks from play. The players rest for approximately 90 seconds after the third, fifth, seventh, ninth, and eleventh games of a set. Injury timeouts typically take place during these natural interruptions. Players are also given a rest of 120 seconds after the conclusion of each set. In some tournaments, after the conclusion of the second set, players may choose to leave the courts for the locker room during a longer 10 minute break.

Outside of these breaks, the play is relatively continuous. Players are encouraged to begin a point within 25 seconds of the conclusion of the previous point. In recent years, tournaments have employed serve clocks to lessen stoppage of play. There is empirical evidence that interruptions dampen the effects of catalysts of momentum (Meier et al., 2020).

Interruptions are a natural part of play, so the number of interruptions in a match is highly correlated with the number of points in a match. Longer matches have more interruptions and shorter matches have fewer. On average, there are 12.54 interruptions per match.

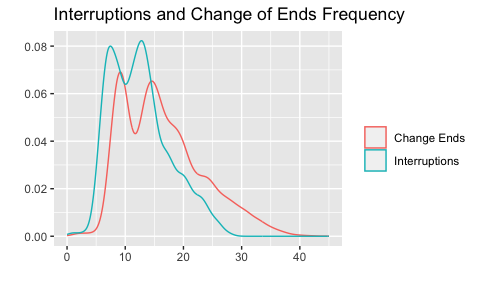


The indicator variable **interruption** holds a value of 1 the point following an interruption and a value of 0 otherwise.

### Change Ends

Players also regularly switch sides of the court. The typically switch during interruptions, but they also switch after the conclusion of the first game of a set and the sixth, twelfth, eighteenth, etc. games in a tiebreak. Players are permitted to drink and refresh, but are not allowed to sit down. Thus, these situations can serve as miniature interruptions.

The two are highly related and have bimodal distributions. This is directly related to the contrasting structure of men’s and women’s tennis. Women’s matches tend to be shorter, so most women’s matches are located in the first peak. Conversely, men’s matches are longer and are typically contained within the second peak. On average, there are 16.48 change of ends per match.



The indicator variable **change\_ends** holds a value of 1 the point following all of change of ends (including interruptions) and a value of 0 otherwise.

### New Set

The hierarchical nature of tennis creates dramatic and sudden changes in incentives. Players suffering from a large deficit in one set find themselves tied at the beginning of the next. Thus, we posit that the end of a set can reduce the effects of momentum in the previous set.

The indicator variable **last\_pt\_set** holds a value of 1 on the last point of a set.

# Exploratory Momentum Analysis

Here, we include our initial analysis of the effect of momentum.

## Set Momentum

*move to results/analysis as univariate analysis?*

In a best-of-five set match, the two players can arrive at the pivotal fifth set in several ways. We hypothesize that the sequence of victories in the first four sets impacts the odds of winning the fifth set.

Our initial descriptive analysis supports this conclusion. Players that win the fourth set win the fifth set 57.6% of the time. Players who enter the fifth set with momentum tend to win the fifth set more often. Interestingly, this momentum builds. Players who won the third and fourth set win the fifth set over 60% of the time.

Momentum in Five Set Matches

| **Sequence of Set Outcomes** | **Fifth Set Win Percentage** |
| --- | --- |
| WWLL | 39.54% |
| WLWL | 40.95% |
| LWWL | 47.06% |
| WLLW | 52.94% |
| LWLW | 59.05% |
| LLWW | 60.46% |
| Win 4th set | 57.65% |
| Win 3rd set | 50.21% |
| Win 2nd set | 47.78% |
| Win 1st set | 44.35% |

This set-by-set descriptive analysis of momentum suggests that previous results impact players, but at a diminishing rate. Results in set 4 play a much larger role in the outcome of set 5 than results in set 1.