# FIVB Beach Volleyball Historic Top 8 Teams Analysis

**Tyler Widdison**

**Dec 2019**

# 1 Project description

*Evaluation of winning trends related to court and hour for the top 8 FIVB Beach volleyball teams per gender since 2001. The concluding hypothesis is that teams ranked within the top 8 have a higher probability at winning on court 1 at 09:00. Generalized linear model is used in support of the hypothesis. Notes:* **^Court records are not kept by FIVB until 2009. ^^Hour records are not kept by FIVB until 2004**. *This project has given me the skills necessary to add new insight to other projects I work on in the future.*

# 1.1 Source of data set

This data set was webscrapped from 2 sources. First from FIVB Beach competiton database. http://www.fivb.org/EN/BeachVolleyball/Competitions/Competitions.htm (http://www.fivb.org/EN/BeachVolleyball/Competitions/Competitions.htm) Second, FIVB Beach rankings from http://www.fivb.org/Vis/Public/JS/Beach/SeasonRank.aspx? Gender=1&id=BTechPlayW&Date=20191001 (http://www.fivb.org/Vis/Public/JS/Beach/SeasonRank.aspx? Gender=1&id=BTechPlayW&Date=20191001). Numerous more links were accessed between these two links. But ultimately I was able to access the data I was looking for. I will provide the final CSV I used for this project. I did end up doing a lot more work to get this final csv than I originally planned.

^Note: The team variable is already melted. One match is listed twice. For example if it was USA vs GER, I have the 'team' be 'USA' and the opp team be 'GER', the next time this match appears the team is 'GER' and the opp is 'USA'. I did this so I can look at specific teams results. I wouldn't be able to get a true reading on how many matches 1 team played if I had kept it seperated. I call this 'The focus team.'

# 1.2 Data dictionary

| Variable Name | Description |
| --- | --- |
| *no* | Match number of specific tournament |
| *date* | Date of match |
| *time* | Time of match |
| *court* | Court match is played on |
| *result* | Result of the played match |
| *duration* | Duration of the played match |
| *tourn* | Tournament name |
| *year* | Year of tournament |
| *phase* | Phase of tournament the match was played |
| *winning_team* | Winning team of the match |
| *losing_team* | Losing team of the match |
| *gender* | Gender of teams |
| *team* | The focus team |
| *team_country* | The focus team country |
| *team_rank* | The rank of the focus teams at the end of that played year |
| *opp* | The focus teams opponent |
| *opp_country* | The focus teams opponent country |
| *matches_played* | This is always listed as '1'. Each row is one match. I wanted to be able to sum the total of matches played for my hypothesis |
| *team_match_won* | If the focus team had won the match this will be a '1' otherwise it is a '0' |
| *opp_match_won* | If the opponent team had won the match this will be a '1' otherwise it is a '0' |

| Variable Name | Description |
| --- | --- |
| *team_player_1* | First listed player in the played match for the focus team |
| *team_player_2* | Second listed player in the played match for the focus team |
| *opp_player_1* | First listed player in the played match for the opp team |
| *opp_player_2* | Second listed player in the played match for the opp team |

Each variable was considered and ultimatley I decided to do my hypothesis on my level. The main variables used were: date, court, year, gender, team, team_rank, matches_played, team_match_won, team_player_1 and team_player_2.

# 2 Preparation of Data

There was a lot of data munging I had to do in order to get to my final hypothesis. I had to change classes, column names, get rid of columns, replace strings, melt the players variable and get the hour of the time. This is a few things I did. There was more munging I originally did before I came to my final hypothesis.

## 2.1 FIVB data

I webscrapped all the data from the FIVB websites. With the amount of links (about 750) one can guess there was errors I found. I had to nix all the matches that ended in a 'injury' result. Player names were different across multiple tournaments, time of day was bad in some cases. For example: 01:00, 02:00, 04:00, 04:57 were all times I had to make correct. There were a couple of matches which still had the live score on the result. For example: LIVE (21-12, 13-15). It shouldn't have that 'LIVE' potion on the result.

## 2.2 Ranking of teams data

I webscrapped the rankings on Oct 1 of each year. After consulting with a friend we decided Oct 1 would be the final rank for that played year. Jan - Oct. The naming convention between the rankings and the FIVB schedule were different. So I had to spend some time doctoring in Excel in order to get the names, and rankings, of the player in the top 8 correct for each year. Also there were special charecters that caused some issues that I dealt with in this way:

```
unescape_html <- function(str){
  xml2::xml_text(xml2::read_html(paste0("<x>", str, "</x>")))
}
```

# 2.3 Final web scrapped data summary

```
summary(df)
```

```
##    match_no              date              time               court
##  Min.  :  1.00  Min.   :2001-04-04  Length:226126     Min.   :1.00
##  1st Qu.: 12.00  1st Qu.:2006-07-19  Class :character  1st Qu.:1.00
##  Median : 25.00  Median :2010-08-24  Mode  :character  Median :2.00
##  Mean   : 29.38  Mean   :2011-02-19                    Mean   :2.45
##  3rd Qu.: 45.00  3rd Qu.:2016-03-18                    3rd Qu.:3.00
##  Max.   :136.00  Max.   :2019-11-09                    Max.   :8.00
##  NA's   :2086                                          NA's   :105034
##     duration          tourn              year           tourn_rank
##  Min.   :  300  Length:226126     Min.   :2001  Min.   :  7.20
##  1st Qu.: 2100  Class :character  1st Qu.:2006  1st Qu.: 18.63
##  Median : 2400  Mode  :character  Median :2010  Median : 23.41
##  Mean   : 2538                    Mean   :2011  Mean   : 49.55
##  3rd Qu.: 2940                    3rd Qu.:2016  3rd Qu.: 41.50
##  Max.   :14940                    Max.   :2019  Max.   :479.09
##
##  winning_team       losing_team         gender
##  Length:226126     Length:226126     Length:226126
##  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character
##
##
##
##
##      team           team_country        team_rank
##  Length:226126     Length:226126     Min.   :   1.00
##  Class :character  Class :character  1st Qu.:  14.00
##  Mode  :character  Mode  :character  Median :  35.00
##                                      Mean   :  66.84
##                                      3rd Qu.:  84.00
##                                      Max.   :1976.00
##                                      NA's   :18
##  team_match_score   team_game_one    team_game_two    team_game_three
##  Length:226126     Min.   : 0.00    Min.   : 0.00    Min.   : 0.000
##  Class :character  1st Qu.:17.00    1st Qu.:17.00    1st Qu.: 0.000
##  Mode  :character  Median :21.00    Median :21.00    Median : 0.000
##                    Mean   :18.89    Mean   :18.86    Mean   : 4.496
##                    3rd Qu.:21.00    3rd Qu.:21.00    3rd Qu.:12.000
##                    Max.   :41.00    Max.   :42.00    Max.   :35.000
##
##      opp             opp_country       opp_match_score       opp_game_one
##  Length:226126     Length:226126     Length:226126     Min.   : 0.00
##  Class :character  Class :character  Class :character  1st Qu.:17.00
##  Mode  :character  Mode  :character  Mode  :character  Median :21.00
##                                                        Mean   :18.89
##                                                        3rd Qu.:21.00
```

```
##                                                         Max.   :41.00
##
##    opp_game_two     opp_game_three    opp_team_rank       matches_played
##  Min.    : 0.00   Min.    : 0.000   Min.    :    1.00   Min.    :1
##  1st Qu.:17.00    1st Qu.: 0.000    1st Qu.:   15.00    1st Qu.:1
##  Median :21.00    Median : 0.000    Median :   35.00    Median :1
##  Mean    :18.86   Mean    : 4.496   Mean    :   66.98   Mean    :1
##  3rd Qu.:21.00    3rd Qu.:12.000    3rd Qu.:   84.00    3rd Qu.:1
##  Max.    :42.00   Max.    :35.000   Max.    :1976.00    Max.    :1
##                                     NA's    :18
##   team_match_won opp_match_won opp_player_1       opp_player_2
##  Min.    :0.0    Min.    :0.0   Length:226126      Length:226126
##  1st Qu.:0.0     1st Qu.:0.0    Class :character   Class :character
##  Median :0.0     Median :1.0    Mode  :character   Mode  :character
##  Mean    :0.5    Mean    :0.5
##  3rd Qu.:1.0     3rd Qu.:1.0
##  Max.    :1.0    Max.    :1.0
##
##   score_diff_game_one score_diff_game_two score_diff_game_three
##  Min.    :-2.1e+01    Min.    :-2.1e+01    Min.    :-1.5e+01
##  1st Qu.:-4.0e+00     1st Qu.:-4.0e+00     1st Qu.: 0.0e+00
##  Median : 2.0e+00     Median :-2.0e+00     Median : 0.0e+00
##  Mean    : 5.3e-05    Mean    :-2.7e-05    Mean    :-2.7e-05
##  3rd Qu.: 4.0e+00     3rd Qu.: 4.0e+00     3rd Qu.: 0.0e+00
##  Max.    : 2.1e+01    Max.    : 2.1e+01    Max.    : 1.5e+01
##
##   team_final_score opp_team_final_score          variable
##  Min.    : 3.00    Min.    : 3.00       team_player_1:113063
##  1st Qu.:35.00     1st Qu.:35.00        team_player_2:113063
##  Median :42.00     Median :42.00
##  Mean    :42.24    Mean    :42.24
##  3rd Qu.:49.00     3rd Qu.:49.00
##  Max.    :85.00    Max.    :85.00
##
##     player           game_three_amount    phase                    hour
##  Length:226126      Min.    :0.0000     Length:226126      Min.    : 8.00
##  Class :character   1st Qu.:0.0000      Class :character   1st Qu.:11.00
##  Mode  :character   Median :0.0000      Mode  :character   Median :13.00
##                     Mean    :0.3311                        Mean    :13.18
##                     3rd Qu.:1.0000                         3rd Qu.:16.00
##                     Max.    :1.0000                        Max.    :23.00
##                                                            NA's    :28864
```
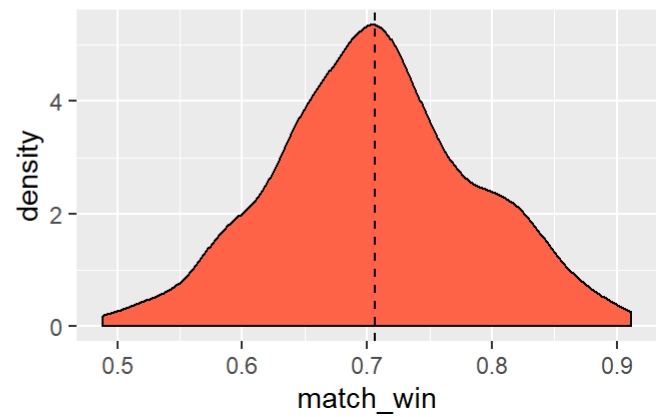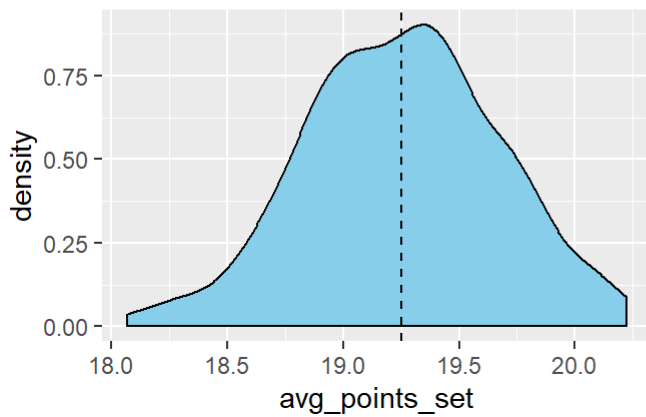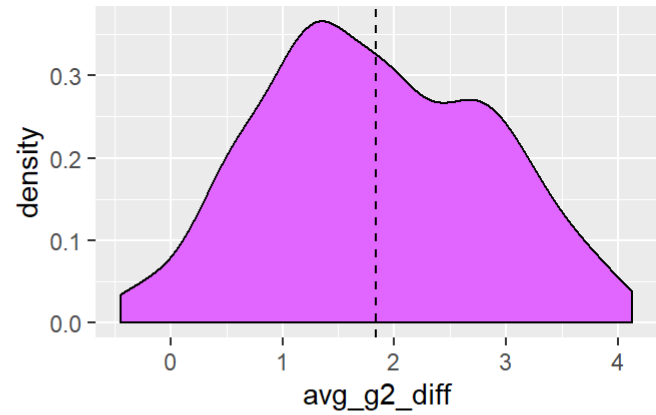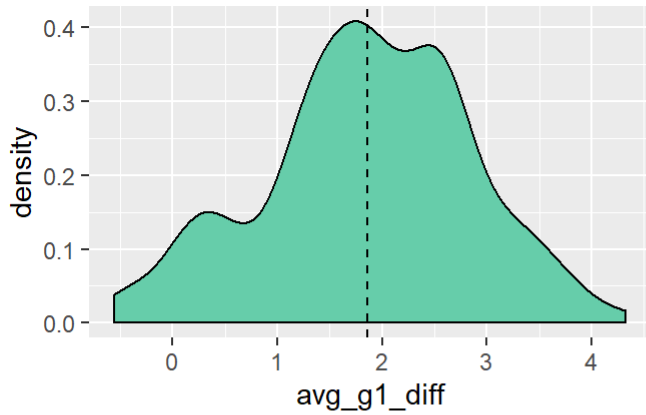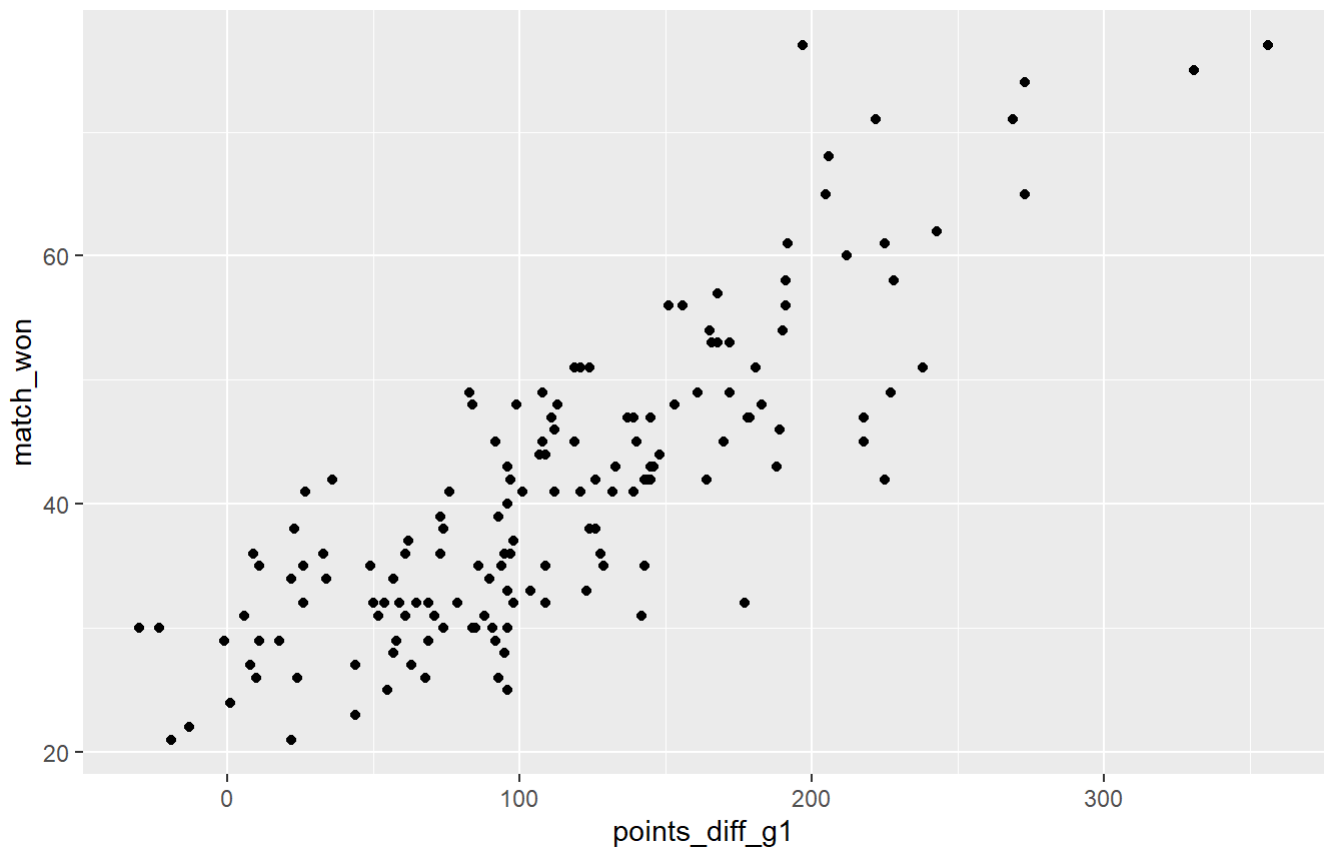
# 3 Exploring the data

I ended up spending time exploring areas I didn't dive deeper into. A big part was trying to figure out how I could use 'score_difference' and how likely a team was to win the match as a varaible for a hypothesis. Other areas I explored:
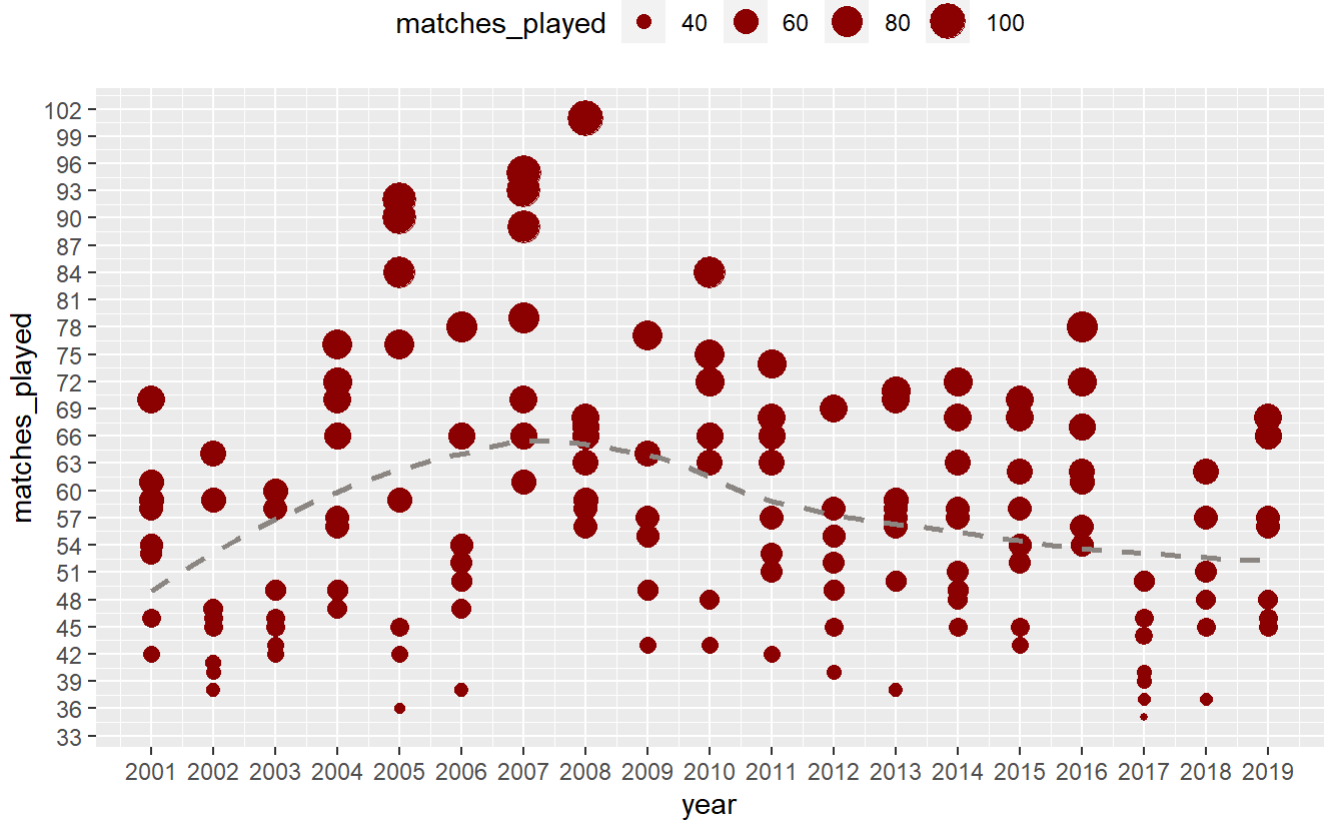
# 3.1 Places explored

- Players density charts



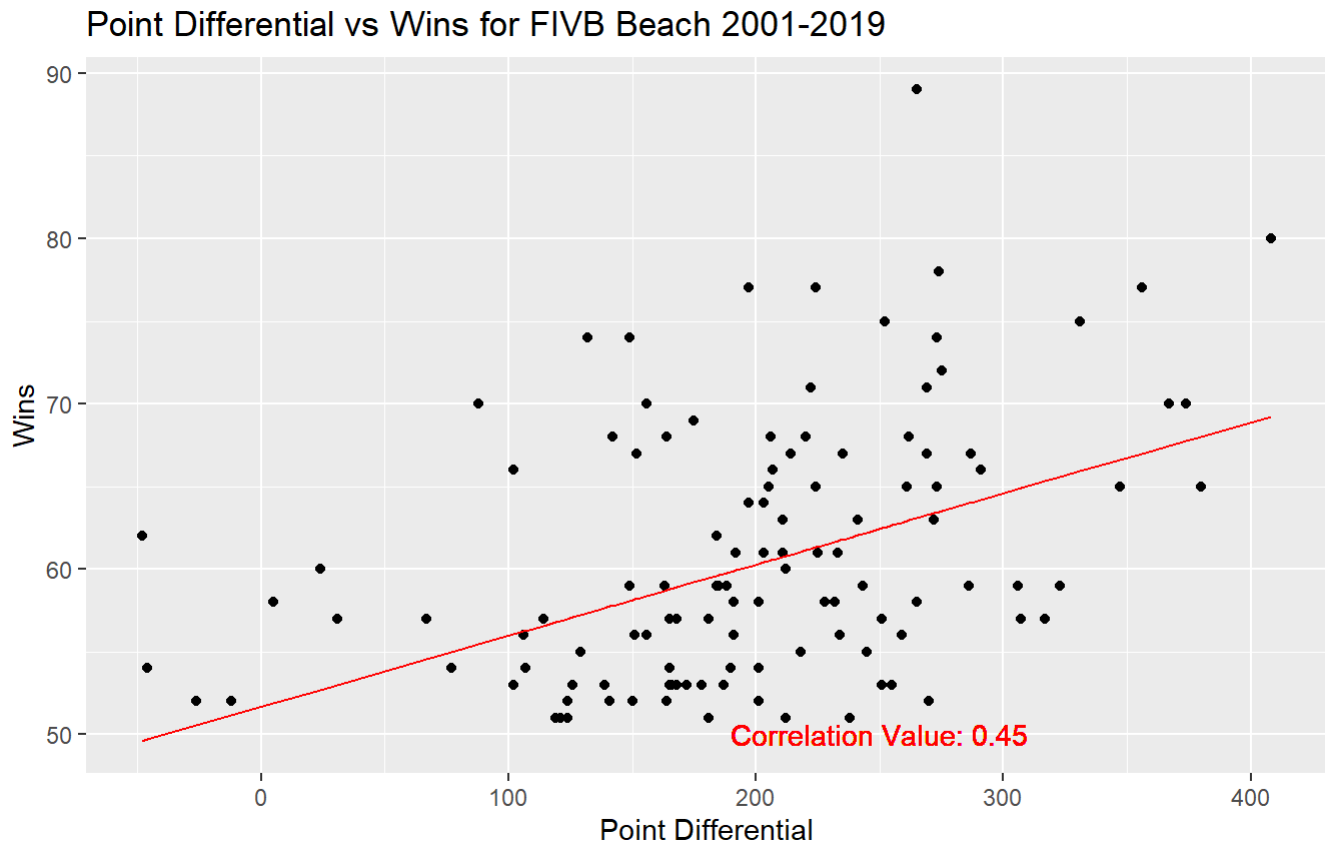- Points difference in game 1 and match won
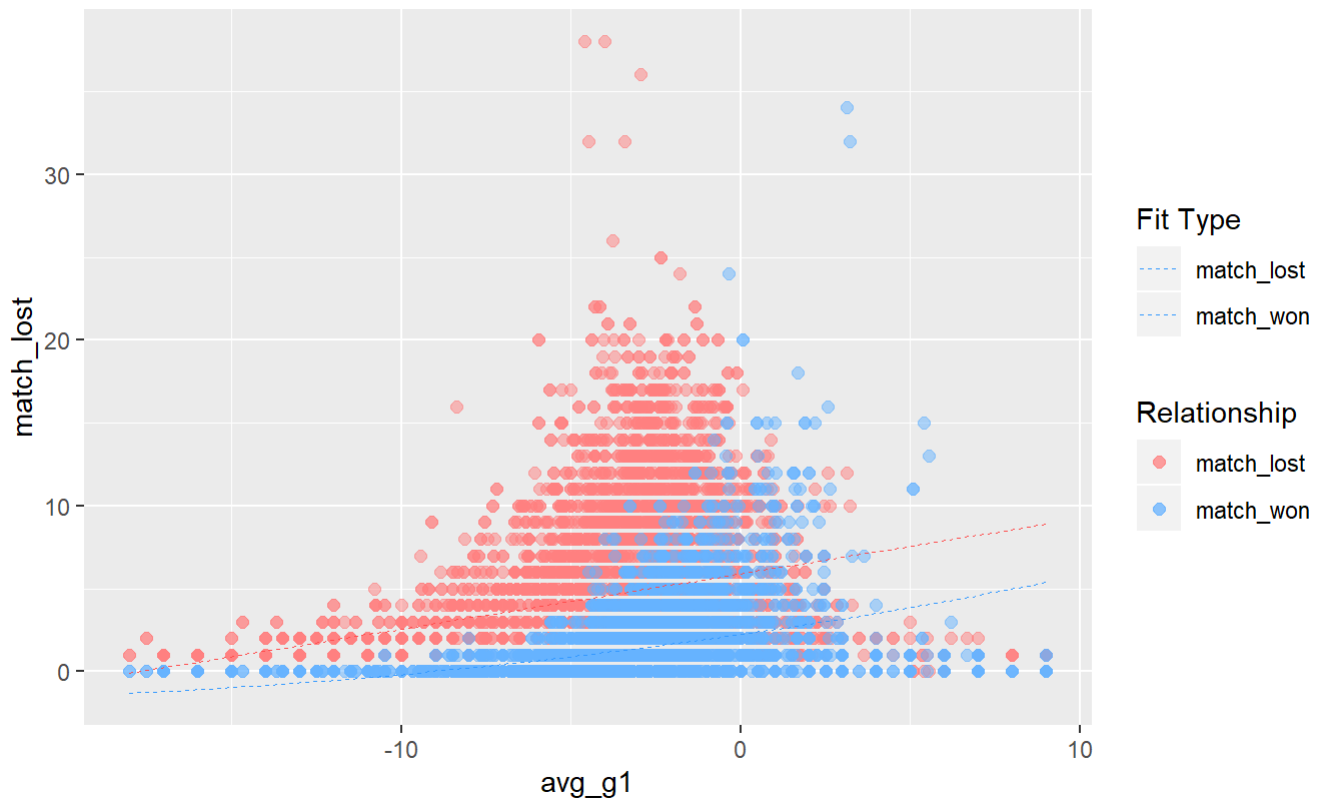
- Matches played per year for mens top 8



- Point diff and correlation value

```
##    (Intercept) points_diff_g1
##    51.67797528     0.04305242
```



Point Differential vs Wins for FIVB Beach 2001-2019

Correlation Value: 0.45

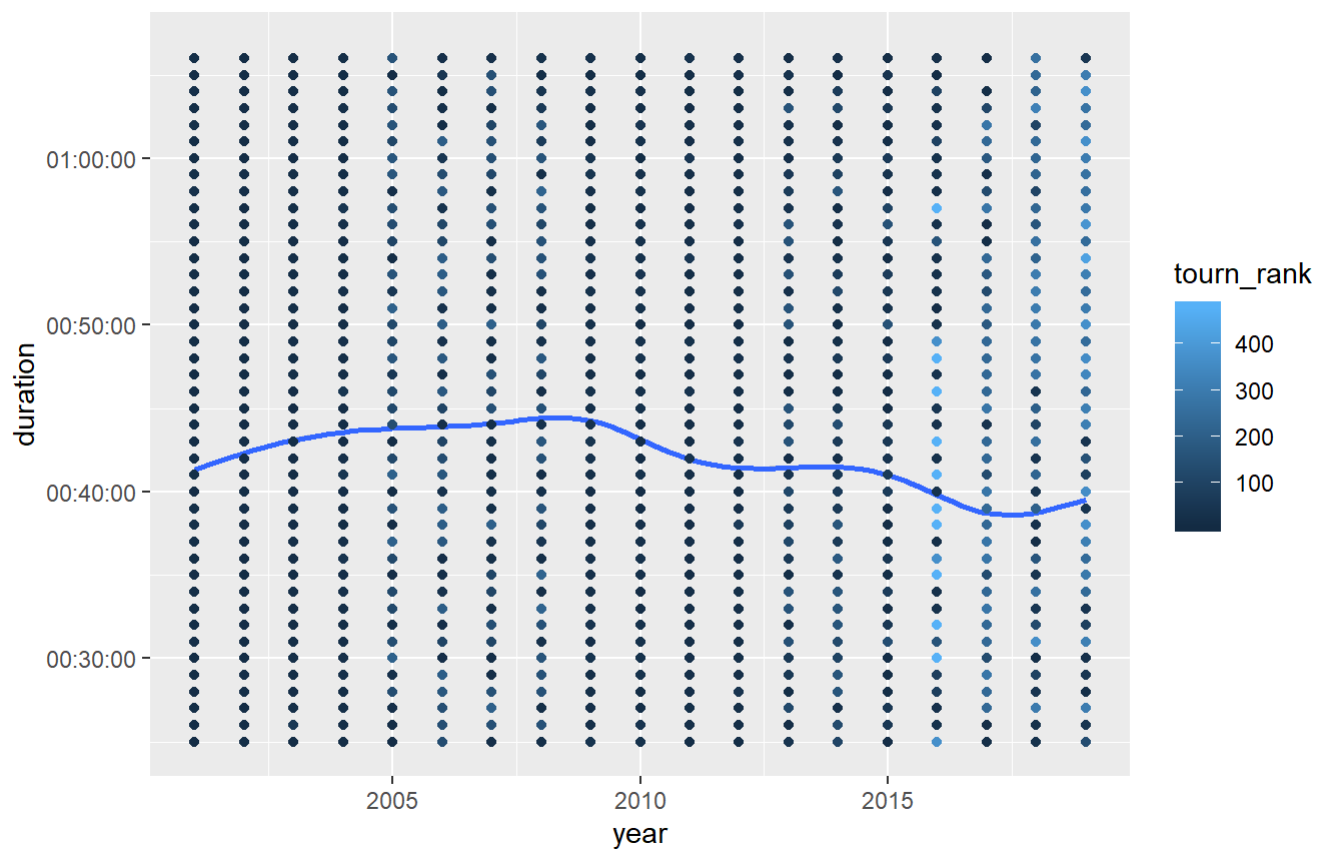- Teams ranked 20 and higher, coorlation of match wins based on average point diff in game 1

Manual Legend for Stat Smooth

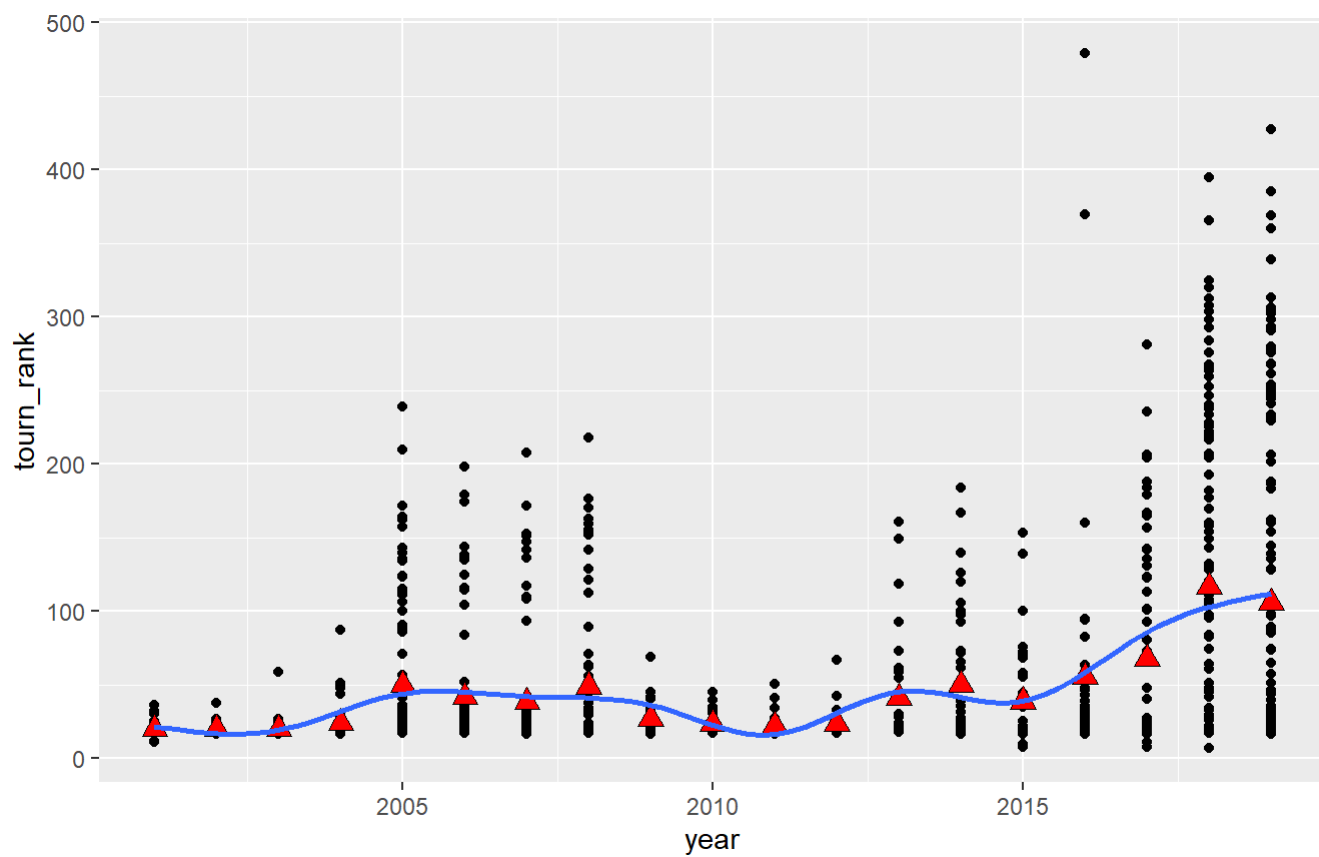- Tougher tournaments have longer matches

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 5508 rows containing missing values (geom_point).
```
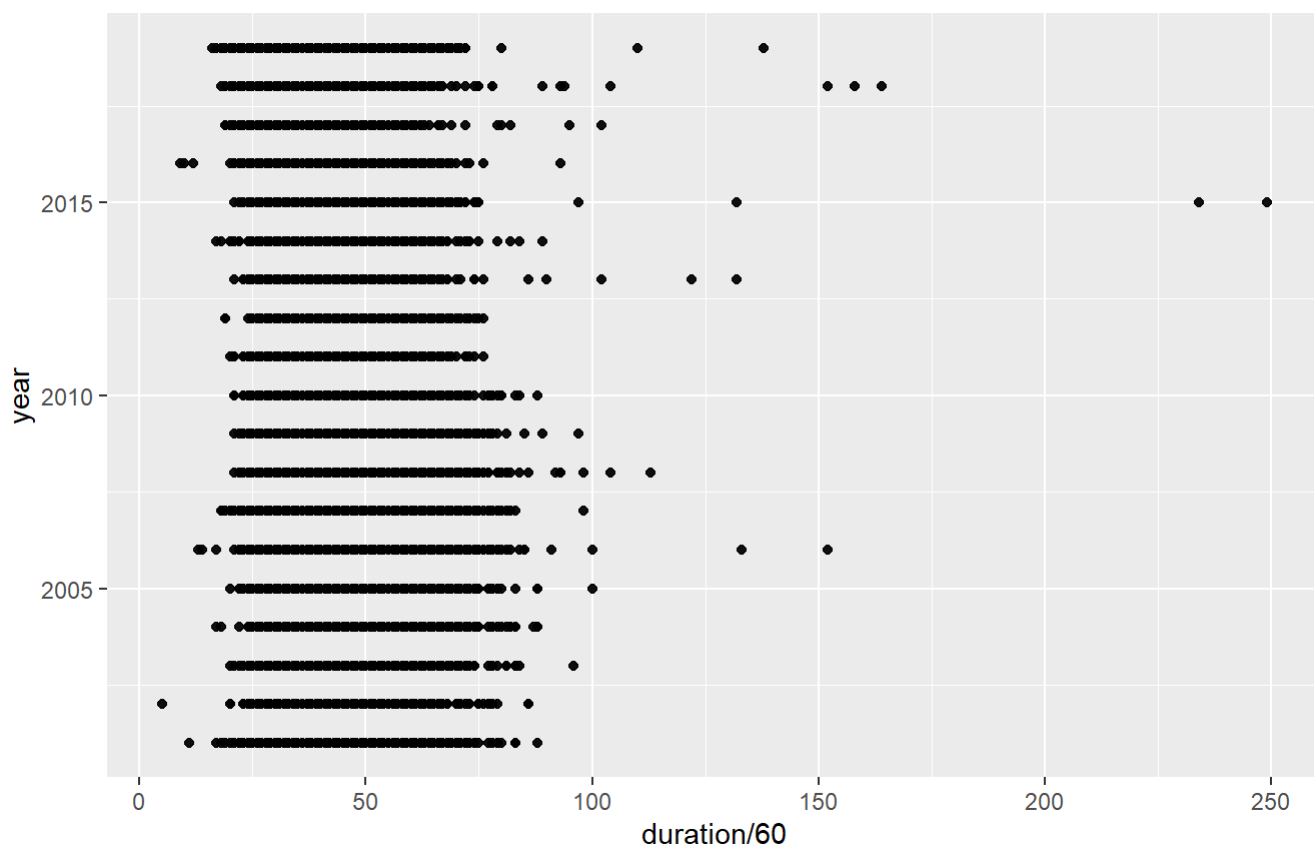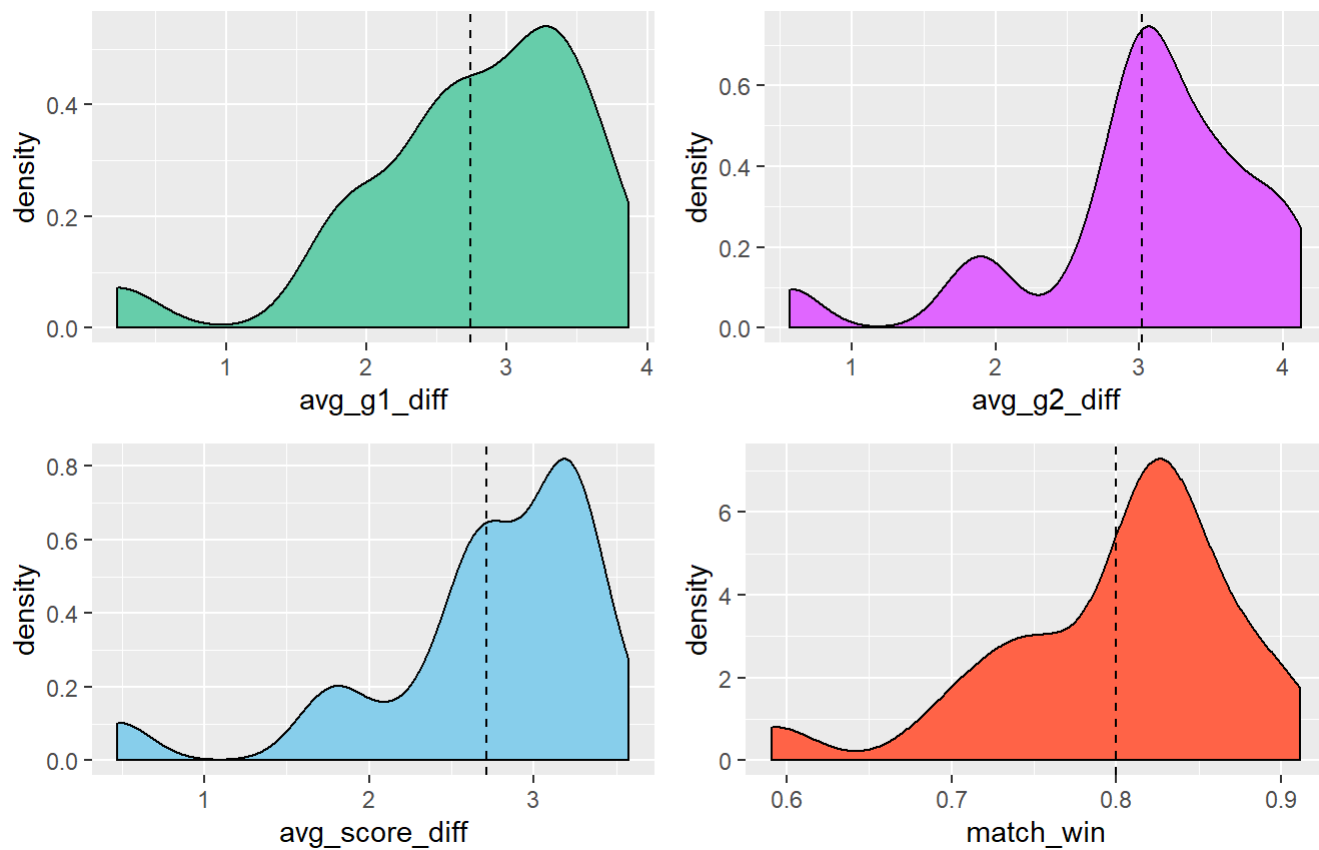
- Toughest year to play beach volleyball?

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

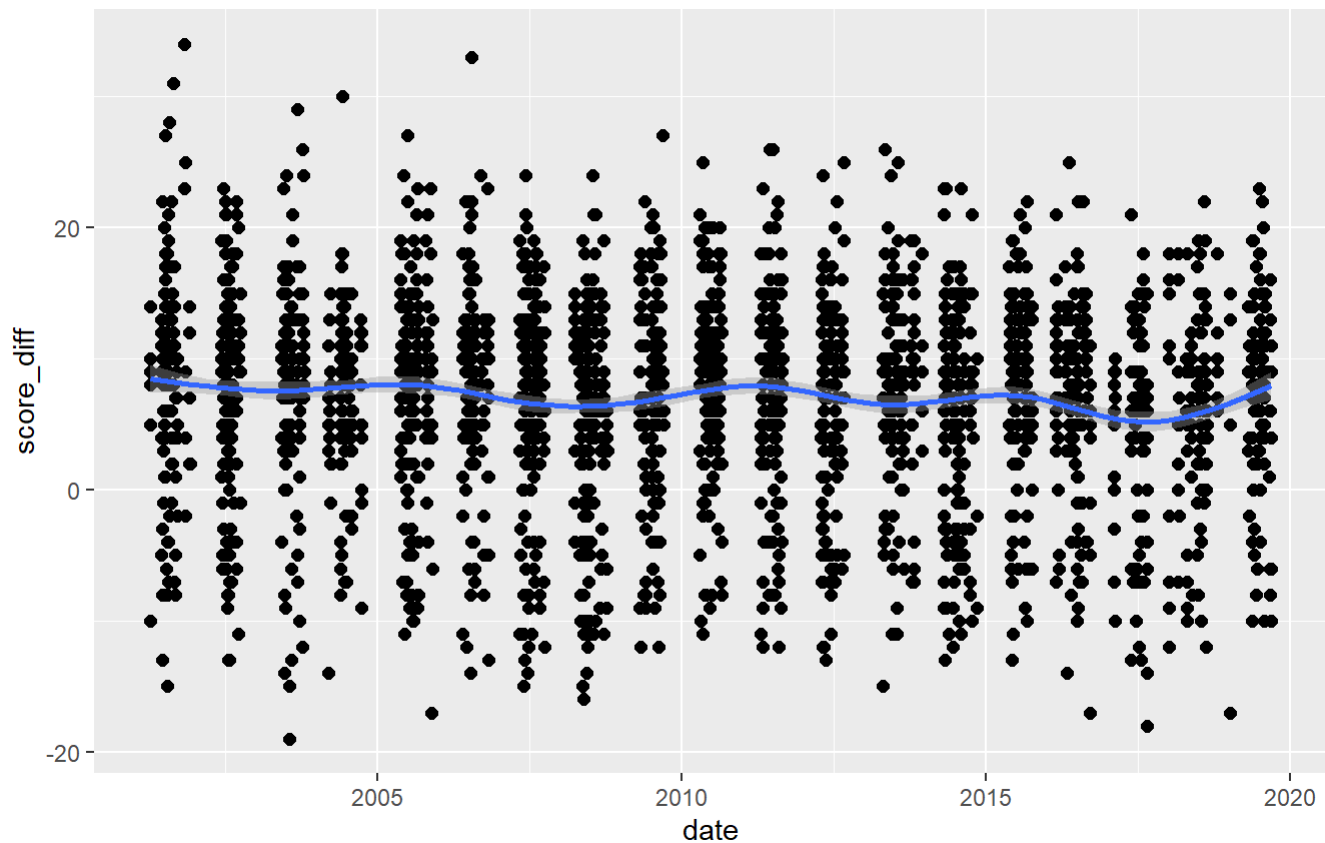- Which year avg the longest match duration?



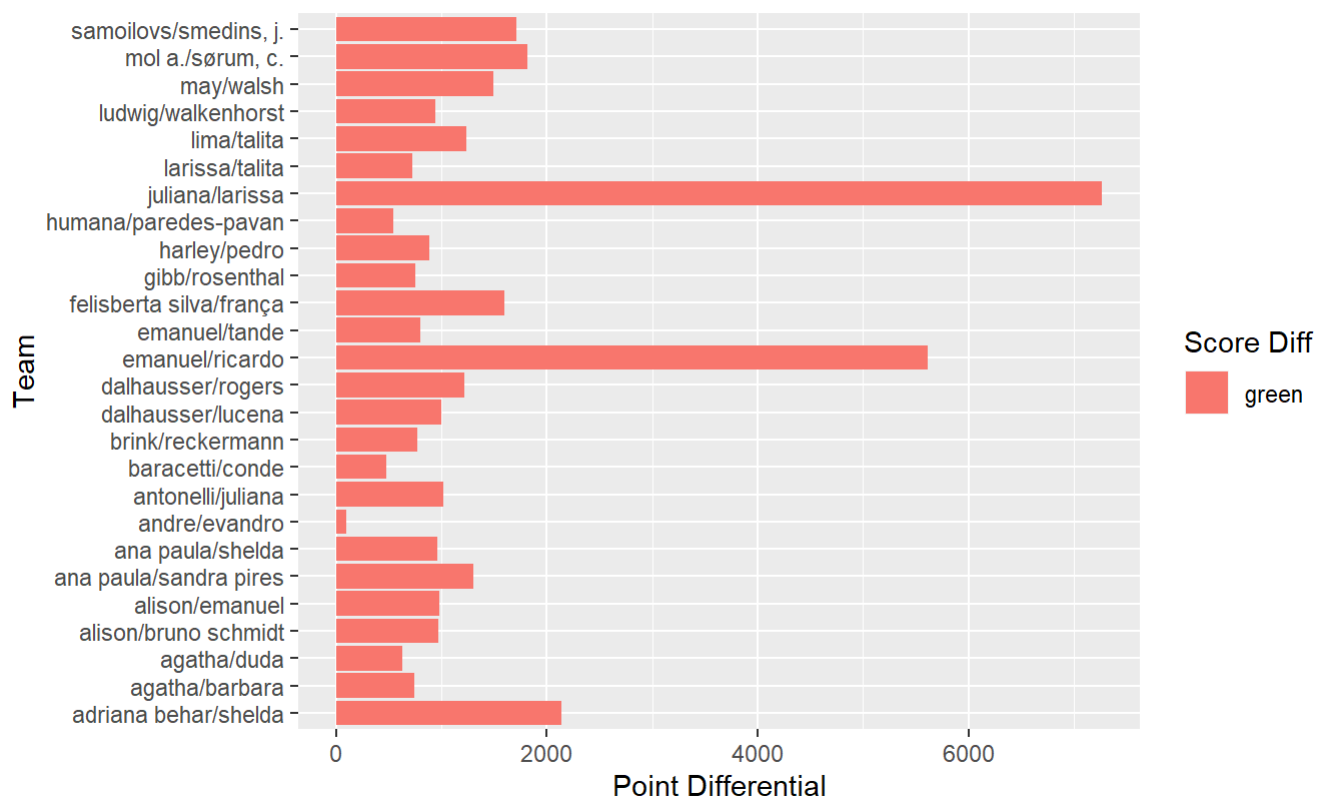- Men teams ranked 1st played how well?



- All first ranked teams score diff

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```
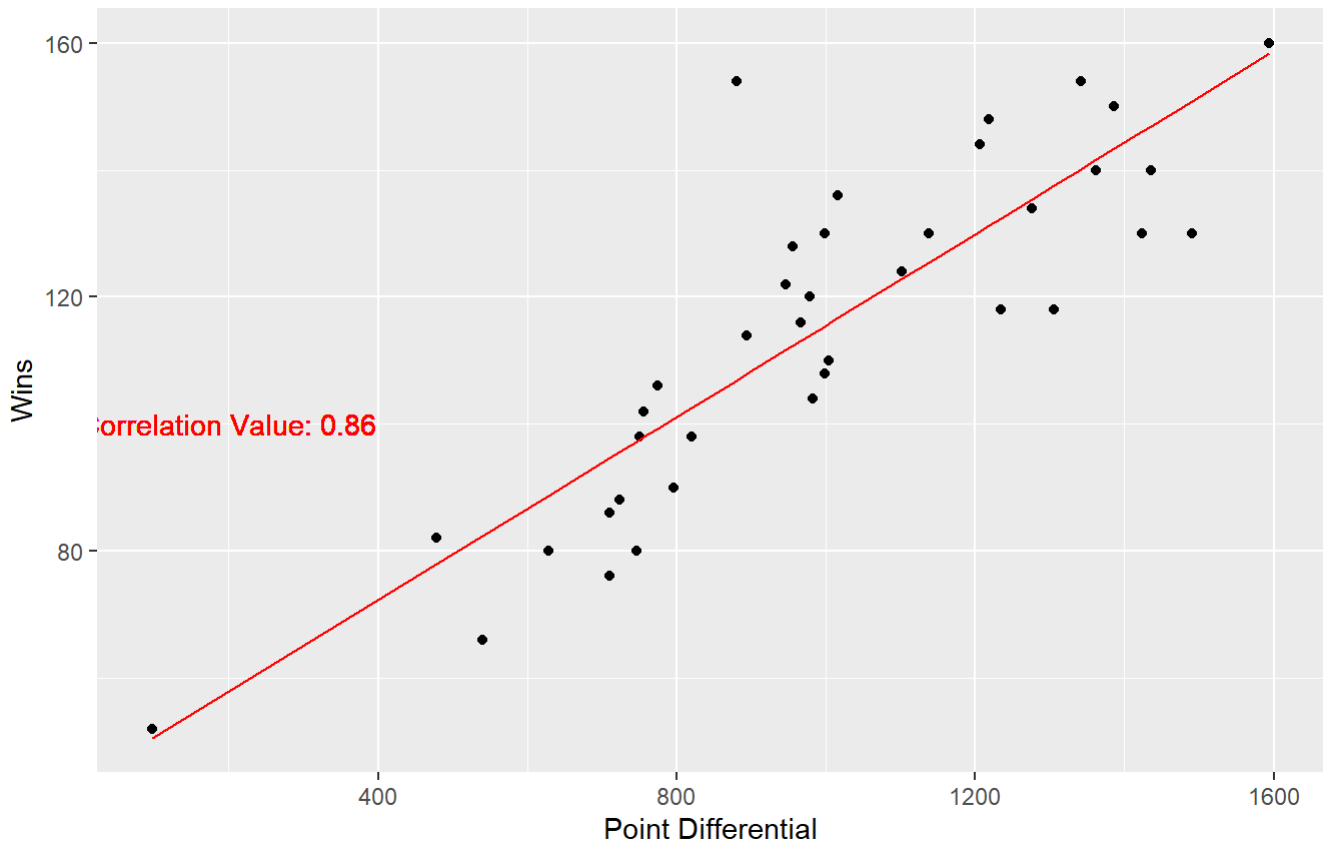


- All first place team final points difference



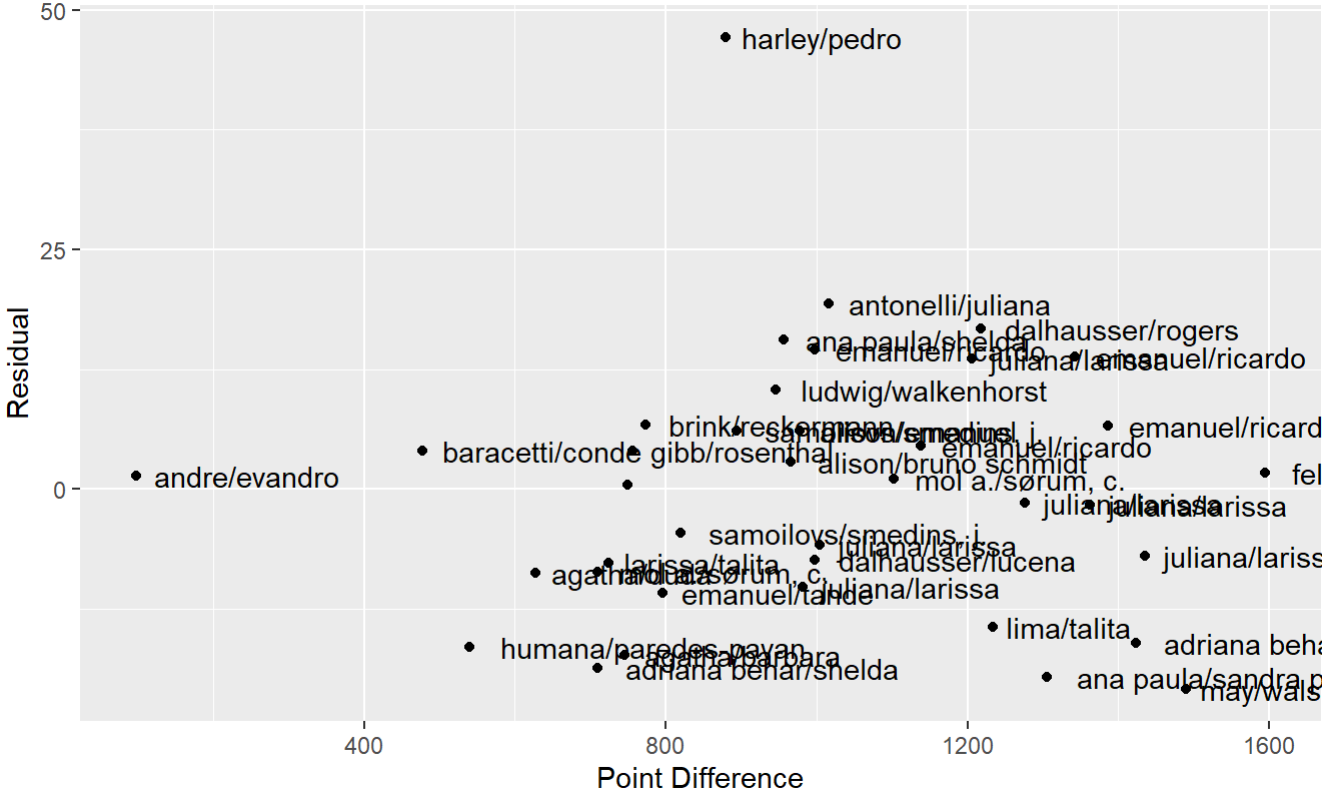FIVB First Place Team Point Differentials

- Point difference in first place teams

```
## (Intercept)  score_diff
## 43.50886944  0.07202652
```



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

## Residuals for Wins and Point Difference Model



- Getting the data put into this

```r
all_player_df <- df %>%
  dplyr::group_by(player, hour, team, team_country, tourn_rank, tourn, date, phase,
  year, match_no, court, time, team_rank, gender, opp, opp_team_rank) %>%
  dplyr::summarise(total_matches = sum(matches_played),
                   match_won = sum(team_match_won),
                   match_lost = sum(matches_played - team_match_won),
                   match_win = match_won / (match_won + match_lost),
                   total_set_threes = sum(game_three_amount),
                   total_sets_played = (total_matches + total_matches + total_set_t
  hrees),
                   rank = sum(team_rank))

all_player_df$total_matches<-suppressWarnings(as.character(all_player_df$total_matc
  hes))

all_player_df$total_matches <- str_replace(all_player_df$total_matches, '2', '1') #
  Each row is one match. When there was a team with the same name twice, widdison/w
  iddison, it would say they played in 2 matches for that team. So I did a quick sw
  itch to 1.

all_player_df$total_matches <- suppressWarnings(as.integer(all_player_df$total_matc
  hes))

df_men <- all_player_df %>%
  dplyr::filter(gender == 'm' & team_rank <= 8)

df_women <- all_player_df %>%
  dplyr::filter(gender == 'w' & team_rank <= 8)
```

# 4 Machine learning

## 4.1 Decided predictors

After some time exploring and trying things I have never done (this being my first project) I realized I wanted to keep it a bit simplar with something I would understand. That led me to using court, hour, total matches, match win % and gender as my main variables for my hypothesis. ## glm

```r
f <- df_men %>%
  dplyr::group_by(hour, court) %>%
  dplyr::summarise(total_matches = sum(total_matches),
                   match_win_perce = sum((match_won) / total_matches))


f <- na.omit(f)
f <- f %>% dplyr::filter(total_matches >= 29) #I wanted to get a sample size with m
  ore than 29 matches


formula <- match_win_perce ~hour + court



mhlm <- glm(formula, data=f)


prob <- predict(mhlm, newdata=f, type="response")
prob <- round(prob,3)
f$prob <- prob



f %>%
  ggplot(aes(court, hour)) +
  geom_point(aes(color = prob), size = 7, alpha = 0.5) +
  scale_size(range = c(.5, 9)) +
  scale_y_continuous(breaks = seq(8, 22, by = 1)) +
  labs(title = 'FIVB top 8 team probability of winning by court and hour',
       subtitle = 'Data from 2009 - 2019',
       size = 'Probability%',
       color ='Probability',
       x = 'Court',
       y = 'Hour')+
  scale_color_viridis(option="C")+
  scale_x_continuous(breaks = 1:6)
```
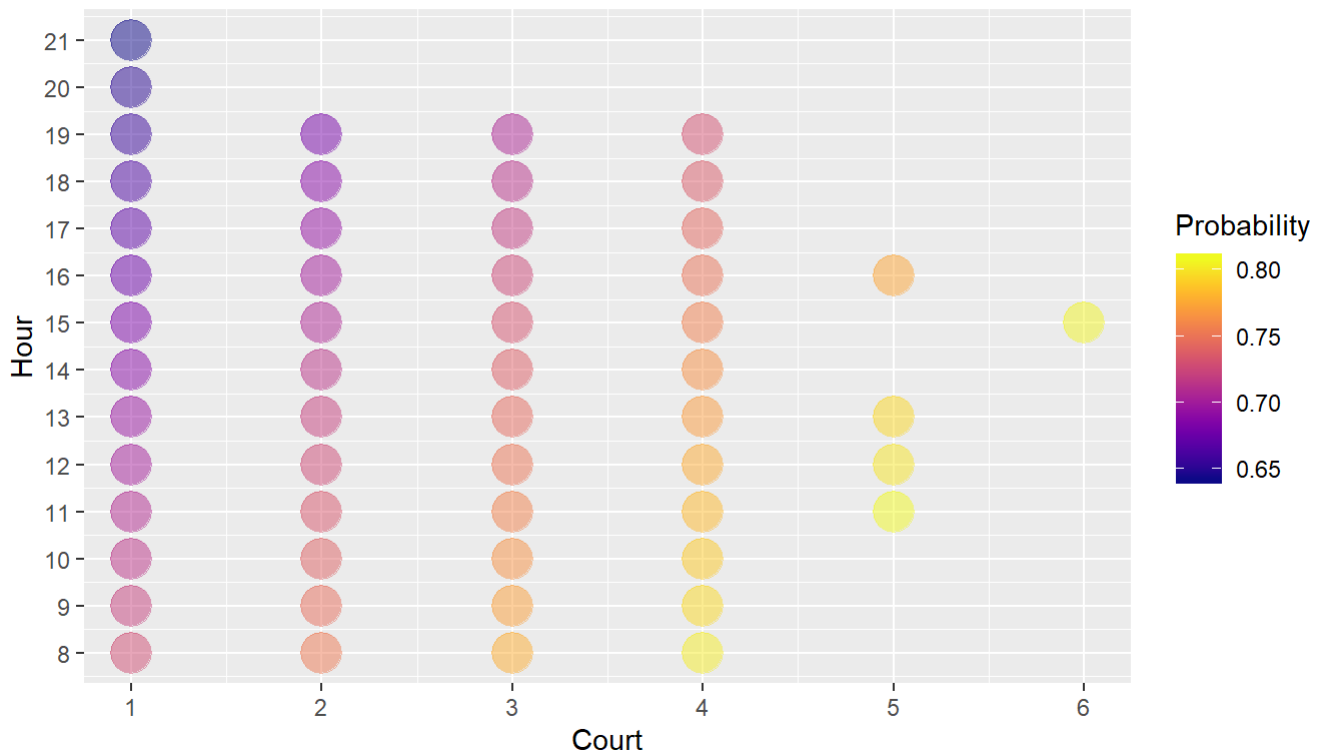
FIVB top 8 team probability of winning by court and hour
Data from 2009 - 2019

```
f$prob<-format(round(f$prob,3))
f$match_win_perce <- NULL
f$total_matches <- NULL

f<-f[order(-prob),]
colnames(f) <- c('Hour', 'Court', 'Probablity')
kable(f) %>%
  kable_styling(bootstrap_options = "striped", full_width = F, position = 'left', f
  ixed_thead = T) %>%
  scroll_box(width = '200px', height='500px')
```

| Hour | Court | Proba |
|------|-------|-------|
| 11 | 5 | 0.808 |
| 15 | 6 | 0.806 |
| 8 | 4 | 0.803 |
| 12 | 5 | 0.801 |
| 9 | 4 | 0.796 |
| 13 | 5 | 0.795 |

| Hour | Court | Proba |
| --- | --- | --- |
| 10 | 4 | 0.790 |
| 11 | 4 | 0.783 |
| 8 | 3 | 0.778 |
| 12 | 4 | 0.777 |

# 5 Conclusion

## 5.1

There is an apparent trend of teams in the top 8 winning more and having higher probility of winning on lower courts earlier in the day. And lower probility later in the day on higher courts. This trend is likely due to teams in the top 8 advancing further into tournaments and playing other teams in the top 8.