

# ATP Data Analysis

2024-01-10

Instalacija potrebnih paketa.

```
# install.packages("dplyr")  
# install.packages("lubridate")  
# install.packages("ggplot2")  
# install.packages("caret")
```

Učitavanje biblioteka.

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##     filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##  
## Attaching package: 'lubridate'  
  
## The following objects are masked from 'package:base':  
##  
##     date, intersect, setdiff, union
```

```
library(ggplot2)  
library(caret)
```

```
## Loading required package: lattice
```

Učitavanje i opis podataka

```

all_matches <- data.frame()
for (year in 1991:2023) {
  file_name <- paste0("dataset/atp_matches_", year, ".csv")
  matches_year <- read.csv(file_name, stringsAsFactors = FALSE)
  all_matches <- rbind(all_matches, matches_year)
}

print(head(all_matches))

```

```

##   tourney_id tourney_name surface draw_size tourney_level tourney_date
## 1   1991-339   Adelaide    Hard         32              A    19901231
## 2   1991-339   Adelaide    Hard         32              A    19901231
## 3   1991-339   Adelaide    Hard         32              A    19901231
## 4   1991-339   Adelaide    Hard         32              A    19901231
## 5   1991-339   Adelaide    Hard         32              A    19901231
## 6   1991-339   Adelaide    Hard         32              A    19901231
##   match_num winner_id winner_seed winner_entry winner_name winner_hand
## 1         1    101723          NA              Magnus Larsson        R
## 2         2    100946          NA              Q Slobodan Zivojinovic    R
## 3         3    101234          NA              Patrik Kuhnen           R
## 4         4    101889           8              Todd Woodbridge        R
## 5         5    101274          NA              Udo Riglewski          R
## 6         6    102148          NA              Fabrice Santoro        R
##   winner_ht winner_ioc winner_age loser_id loser_seed loser_entry
## 1        193       SWE       20.7   101414          1
## 2        198       YUG       27.4   101256         NA
## 3        190       GER       24.8   101421         NA
## 4        178       AUS       19.7   101703         NA
## 5        185       GER       24.4   101843          4
## 6        178       FRA       18.0   101285         NA
##   loser_name loser_hand loser_ht loser_ioc loser_age score
## 1   Boris Becker        R    190       GER       23.1 6-4 3-6 7-6(2)
## 2   Mark Kratzmann        L    178       AUS       24.6 6-3 3-6 7-6(6)
## 3   Veli Paloheimo        R    183       FIN       23.0      6-0 6-4
## 4   Guillaume Raoux        R    180       FRA       20.8      7-6(2) 6-1
## 5   Sergi Bruguera        R    188       ESP       19.9      7-5 6-3
## 6   Thierry Champion        R    183       FRA       24.3      6-2 6-3
##   best_of round minutes w_ace w_df w_svpt w_1stIn w_1stWon w_2ndWon w_SvGms
## 1         3   R32     130      6   2     96      55      39      25      15
## 2         3   R32     119     19   4    101      56      45      25      15
## 3         3   R32      71      6   1     54      31      24      13      8
## 4         3   R32      85      2   0     60      40      30      14      9
## 5         3   R32      90      4   2     72      40      33      14     10
## 6         3   R32      88      2   1     61      45      32       4      8
##   w_bpSaved w_bpFaced l_ace l_df l_svpt l_1stIn l_1stWon l_2ndWon l_SvGms
## 1         2         4      8   3     95      62      44      23      16
## 2         9        10      8   2     84      41      35      27      15
## 3         1         1      2   2     60      37      22       6       8
## 4         3         3      3   3     74      45      30      11      10
## 5         7         8      2   2     77      41      28      15      11
## 6         7         9      1   0     62      45      20       8       9
##   l_bpSaved l_bpFaced winner_rank winner_rank_points loser_rank
## 1         6         8         56              NA         2

```

```
## 2      1      2      304      NA      75
## 3      4      8      82      NA      69
## 4      5      8      50      NA      84
## 5      4      8      88      NA      28
## 6     10     16      62      NA      59
##  loser_rank_points
## 1              NA
## 2              NA
## 3              NA
## 4              NA
## 5              NA
## 6              NA
```

TODO Opis ispisa

```
print(names(all_matches))
```

```
## [1] "tournament_id"      "tournament_name"    "surface"
## [4] "draw_size"          "tournament_level"   "tournament_date"
## [7] "match_num"          "winner_id"          "winner_seed"
## [10] "winner_entry"       "winner_name"        "winner_hand"
## [13] "winner_ht"          "winner_ioc"         "winner_age"
## [16] "loser_id"           "loser_seed"         "loser_entry"
## [19] "loser_name"         "loser_hand"         "loser_ht"
## [22] "loser_ioc"          "loser_age"          "score"
## [25] "best_of"            "round"              "minutes"
## [28] "w_ace"              "w_df"               "w_svpt"
## [31] "w_1stIn"            "w_1stWon"           "w_2ndWon"
## [34] "w_SvGms"            "w_bpSaved"          "w_bpFaced"
## [37] "l_ace"              "l_df"               "l_svpt"
## [40] "l_1stIn"            "l_1stWon"           "l_2ndWon"
## [43] "l_SvGms"            "l_bpSaved"          "l_bpFaced"
## [46] "winner_rank"        "winner_rank_points" "loser_rank"
## [49] "loser_rank_points"
```

TODO Opis ispisa

```
print(summary(all_matches))
```

```
##  tournament_id      tournament_name      surface      draw_size
## Length:104682      Length:104682      Length:104682      Min.   : 2.00
## Class :character    Class :character    Class :character    1st Qu.: 32.00
## Mode  :character    Mode  :character    Mode  :character    Median : 32.00
##                                     Mean   : 53.52
##                                     3rd Qu.: 64.00
##                                     Max.   :128.00
##
##  tournament_level    tournament_date      match_num      winner_id
## Length:104682      Min.   :19901231      Min.   : 1.00      Min.   :100284
## Class :character    1st Qu.:19971006      1st Qu.: 10.00      1st Qu.:102148
## Mode  :character    Median :20050815      Median : 24.00      Median :103602
##                                     Mean   :20058134      Mean   : 72.47      Mean   :106703
```

```

##          3rd Qu.:20140224    3rd Qu.: 73.00    3rd Qu.:104797
##          Max.    :20230828    Max.    :1701.00    Max.    :211468
##
## winner_seed winner_entry winner_name winner_hand
## Min.    : 1.00 Length:104682 Length:104682 Length:104682
## 1st Qu.: 3.00 Class :character Class :character Class :character
## Median : 5.00 Mode  :character Mode  :character Mode  :character
## Mean    : 6.92
## 3rd Qu.: 8.00
## Max.    :35.00
## NA's    :62282
## winner_ht winner_ioc winner_age loser_id
## Min.    :160.0 Length:104682 Min.    :14.30 Min.    :100282
## 1st Qu.:180.0 Class :character 1st Qu.:23.00 1st Qu.:102154
## Median :185.0 Mode  :character Median :25.50 Median :103566
## Mean    :185.7 Mean    :25.77 Mean    :106814
## 3rd Qu.:190.0 3rd Qu.:28.30 3rd Qu.:104919
## Max.    :211.0 Max.    :42.70 Max.    :212041
## NA's    :2454 NA's    :5
## loser_seed loser_entry loser_name loser_hand
## Min.    : 1.00 Length:104682 Length:104682 Length:104682
## 1st Qu.: 4.00 Class :character Class :character Class :character
## Median : 6.00 Mode  :character Mode  :character Mode  :character
## Mean    : 8.29
## 3rd Qu.:11.00
## Max.    :35.00
## NA's    :81382
## loser_ht loser_ioc loser_age score
## Min.    :160.0 Length:104682 Min.    :14.50 Length:104682
## 1st Qu.:180.0 Class :character 1st Qu.:23.00 Class :character
## Median :185.0 Mode  :character Median :25.70 Mode  :character
## Mean    :185.2 Mean    :25.88
## 3rd Qu.:190.0 3rd Qu.:28.50
## Max.    :211.0 Max.    :46.00
## NA's    :4855 NA's    :18
## best_of round minutes w_ace
## Min.    :3.000 Length:104682 Min.    : 0.0 Min.    : 0.000
## 1st Qu.:3.000 Class :character 1st Qu.: 75.0 1st Qu.: 3.000
## Median :3.000 Mode  :character Median : 96.0 Median : 5.000
## Mean    :3.441 Mean    :103.8 Mean    : 6.526
## 3rd Qu.:3.000 3rd Qu.:125.0 3rd Qu.: 9.000
## Max.    :5.000 Max.    :1146.0 Max.    :113.000
## NA's    : NA's    :13036 NA's    :10207
## w_df w_svpt w_1stIn w_1stWon
## Min.    : 0.000 Min.    : 0.00 Min.    : 0.00 Min.    : 0.00
## 1st Qu.: 1.000 1st Qu.: 56.00 1st Qu.: 34.00 1st Qu.: 26.00
## Median : 2.000 Median : 73.00 Median : 44.00 Median : 33.00
## Mean    : 2.734 Mean    : 78.13 Mean    : 47.66 Mean    : 35.93
## 3rd Qu.: 4.000 3rd Qu.: 94.00 3rd Qu.: 58.00 3rd Qu.: 43.00
## Max.    :26.000 Max.    :491.00 Max.    :361.00 Max.    :292.00
## NA's    :10207 NA's    :10207 NA's    :10207 NA's    :10207
## w_2ndWon w_SvGms w_bpSaved w_bpFaced
## Min.    : 0.00 Min.    : 0.00 Min.    : 0.000 Min.    : 0.000
## 1st Qu.:12.00 1st Qu.: 9.00 1st Qu.: 1.000 1st Qu.: 2.000

```

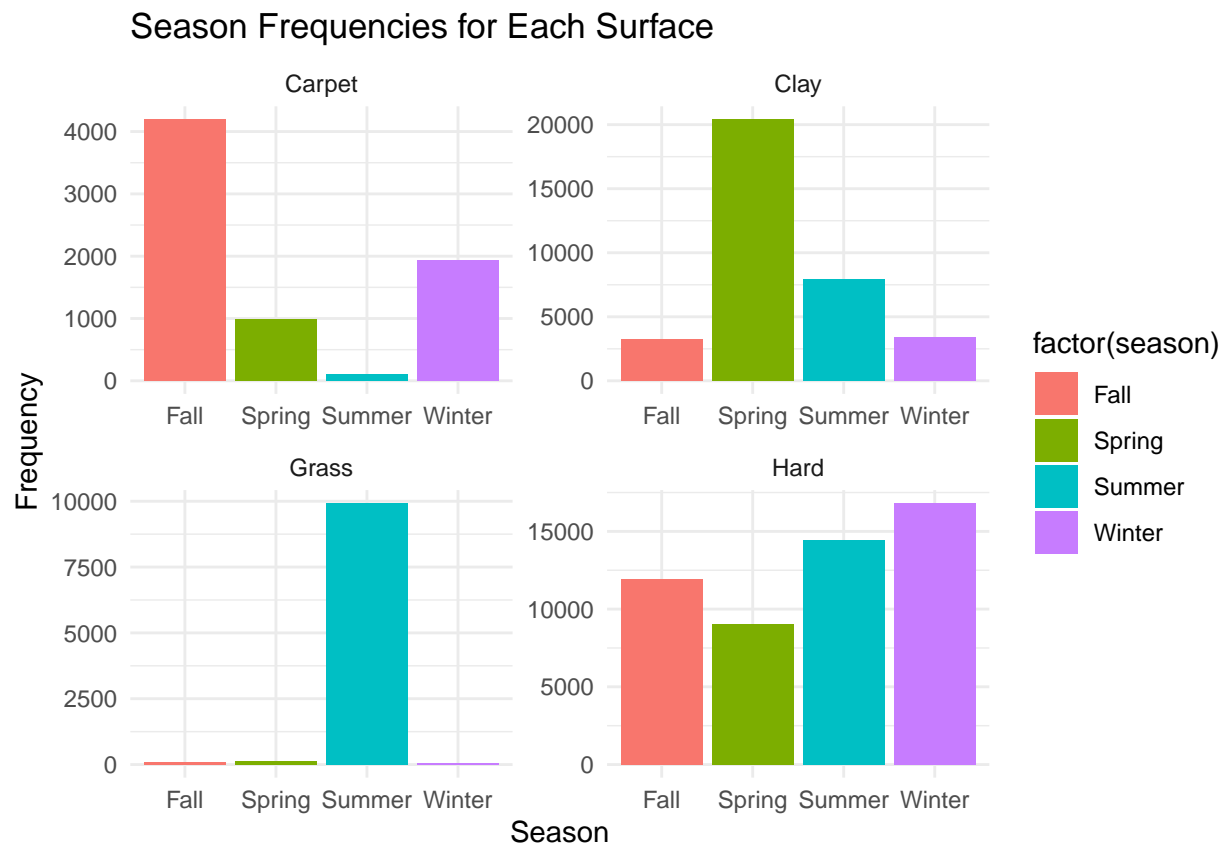
```

## Median :16.00 Median :11.00 Median : 3.000 Median : 4.000
## Mean :16.73 Mean :12.41 Mean : 3.526 Mean : 5.164
## 3rd Qu.:21.00 3rd Qu.:15.00 3rd Qu.: 5.000 3rd Qu.: 7.000
## Max. :82.00 Max. :90.00 Max. :24.000 Max. :34.000
## NA's :10207 NA's :10206 NA's :10207 NA's :10207
## l_ace l_df l_svpt l_1stIn
## Min. : 0.000 Min. : 0.000 Min. : 0.00 Min. : 0.00
## 1st Qu.: 2.000 1st Qu.: 2.000 1st Qu.: 59.00 1st Qu.: 34.00
## Median : 4.000 Median : 3.000 Median : 76.00 Median : 45.00
## Mean : 4.841 Mean : 3.485 Mean : 80.97 Mean : 48.09
## 3rd Qu.: 7.000 3rd Qu.: 5.000 3rd Qu.: 97.00 3rd Qu.: 58.00
## Max. :103.000 Max. :26.000 Max. :489.00 Max. :328.00
## NA's :10207 NA's :10207 NA's :10207 NA's :10207
## l_1stWon l_2ndWon l_SvGms l_bpSaved
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : -6.000
## 1st Qu.: 22.00 1st Qu.: 10.00 1st Qu.: 9.00 1st Qu.: 2.000
## Median : 30.00 Median : 14.00 Median :11.00 Median : 4.000
## Mean : 31.95 Mean : 14.98 Mean :12.21 Mean : 4.813
## 3rd Qu.: 40.00 3rd Qu.: 19.00 3rd Qu.:15.00 3rd Qu.: 7.000
## Max. :284.00 Max. :101.00 Max. :91.00 Max. :28.000
## NA's :10207 NA's :10207 NA's :10206 NA's :10207
## l_bpFaced winner_rank winner_rank_points loser_rank
## Min. : 0.00 Min. : 1.00 Min. : 1 Min. : 1.0
## 1st Qu.: 6.00 1st Qu.: 18.00 1st Qu.: 529 1st Qu.: 37.0
## Median : 8.00 Median : 46.00 Median : 880 Median : 70.0
## Mean : 8.74 Mean : 80.66 Mean : 1429 Mean : 119.1
## 3rd Qu.:11.00 3rd Qu.: 89.00 3rd Qu.: 1598 3rd Qu.: 119.0
## Max. :38.00 Max. :2101.00 Max. :16950 Max. :2159.0
## NA's :10207 NA's :1189 NA's :2177 NA's :2536
## loser_rank_points
## Min. : 1.0
## 1st Qu.: 395.0
## Median : 658.0
## Mean : 895.6
## 3rd Qu.: 1040.0
## Max. :16950.0
## NA's :3519

```

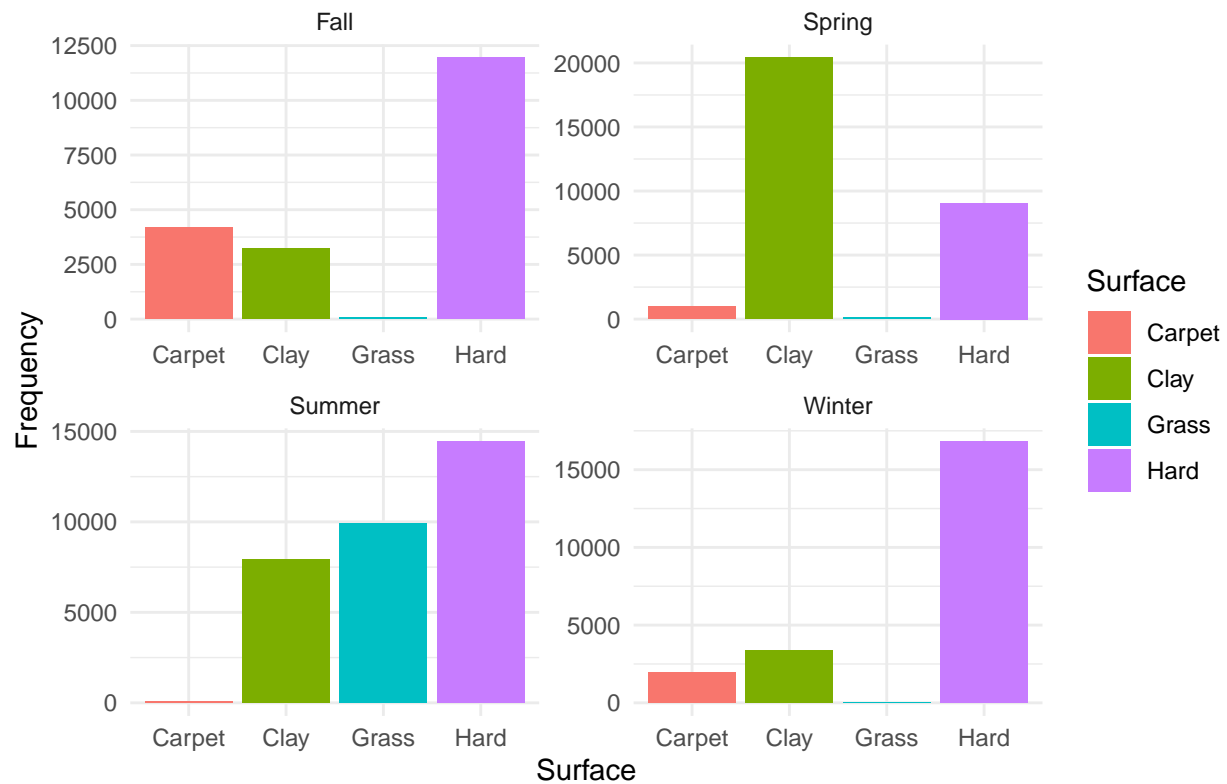
TODO Opis ispisa, možda uzet summary samo za neke značajke

Zadatak 1. Kakva je distribucija mečeva na specifičnim podlogama u različitim godišnjim dobima?



TODO Opisat histogram, šta prikazuje

## Surface Frequencies for Each Season



TODO Opisat histogram, šta prikazuje

**Zadatak 2.** Postoji li značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu?

```
##
## Welch Two Sample t-test
##
## data: open_surface_data and closed_surface_data
## t = 0.86988, df = 6800.5, p-value = 0.3844
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.05464593 0.14183246
## sample estimates:
## mean of x mean of y
## 6.221035 6.177441
```

TODO Opis ispisa

**Zadatak 3.** Ima li razlike u broju serviranih asova na različitim podlogama?

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## surface    3  754563   251521    4319 <2e-16 ***
```

```
## Residuals    94471 5501707      58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TODO Opis ispisa

**Zadatak 4. Kakva je veza između vrste terena i vjerojatnosti da će mečevi otići u peti set?**

```
##
##           FALSE TRUE
## Carpet      700  179
## Clay       5550 1240
## Grass      3471  819
## Hard       9090 2054
```

TODO Opis ispisa

```
##
## Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 3.2059, df = 3, p-value = 0.361
```

TODO Opis ispisa

**Zadatak 5. Možemo li procijeniti broj asova koje će igrač odservirati u tekućoj godini (zadnjoj dostupnoj sezoni) na temelju njegovih rezultata iz prethodnih sezona?**

```
## Warning: Using an external vector in selections was deprecated in tidysselect 1.1.0.
## i Please use 'all_of()' or 'any_of()' instead.
##   # Was:
##   data %>% select(features)
##
##   # Now:
##   data %>% select(all_of(features))
##
## See <https://tidysselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
## 'summarise()' has grouped output by 'player_id', 'year', 'winner_ht'. You can
## override using the '.groups' argument.
## 'summarise()' has grouped output by 'player_id', 'year', 'loser_ht'. You can
## override using the '.groups' argument.
```

```
## # A tibble: 7,417 x 9
## # Groups:   player_id, year, height [7,417]
```



```
##      player_id year height hand total_aces avg_1stIn avg_1stWon svpt    df
##      <int> <dbl> <int> <fct>      <int>      <dbl>      <dbl> <dbl> <int>
##  1    100284 1991    178 L         45      60.3      40.5 90.4    38
##  2    100284 1992    178 L         37      53.6      36.1 80.3    31
##  3    100284 1993    178 L          4       57       40  92.3    11
##  4    100284 1994    178 L          2       61       36  89      5
##  5    100284 1995    178 L          7       43      31.5 78.5    10
##  6    100529 1991    185 R        168      45.3      36.2 81.2    43
##  7    100529 1992    185 R         87      38.3      30.5 78.3    47
##  8    100532 1991    175 R         17       33      26.3 66      8
##  9    100581 1991    180 L        205      39.0      30.8 69.9   123
## 10    100581 1992    180 L        175      50.6      40.3 86.3   126
## # i 7,407 more rows
```

```
## # A tibble: 10,396 x 9
```

```
## # Groups:   player_id, year, height [10,396]
```

```
##      player_id year height hand total_aces avg_1stIn avg_1stWon svpt    df
##      <int> <dbl> <int> <fct>      <int>      <dbl>      <dbl> <dbl> <int>
##  1    100282 1992    180 L          0      67.5      40.5 96      5
##  2    100284 1991    178 L          9      49.2      27.1 75.6    34
##  3    100284 1992    178 L         25      57.9      33.4 90.6    46
##  4    100284 1993    178 L          4      37.4      22.2 60.4    14
##  5    100284 1994    178 L          1       56       34  87.3     3
##  6    100284 1995    178 L          3       48       29  67      2
##  7    100284 1996    178 L          3       55       30  93      2
##  8    100286 1991    168 R          0       32       18  60      2
##  9    100321 1993    193 R          0       34       14  48      0
## 10    100431 1992    178 R          8      46.5      30.5 76      4
## # i 10,386 more rows
```

```
## # A tibble: 40 x 9
```

```
## # Groups:   player_id, year, height [20]
```

```
##      player_id year height hand total_aces avg_1stIn avg_1stWon svpt    df
##      <int> <dbl> <int> <fct>      <int>      <dbl>      <dbl> <dbl> <int>
##  1    104925 2004    188 R          4       60       39  91      2
##  2    104925 2005    188 R         43      62.1      45.4 96.4    26
##  3    104925 2006    188 R        216      49.3      37   79.3    92
##  4    104925 2007    188 R       420      54.2      40.0 83.5   147
##  5    104925 2008    188 R       413      47.3      35.6 72.3   113
##  6    104925 2009    188 R       420      46.2      34.3 73.0   212
##  7    104925 2010    188 R       232      49.2      35.9 77.5   198
##  8    104925 2011    188 R       320      47.0      35.2 71.9   131
##  9    104925 2012    188 R       456      47.4      36.0 73.6   117
## 10    104925 2013    188 R       424      47.5      36.6 72.4    94
## 11    104925 2014    188 R       371      50.8      38.5 75.9    91
## 12    104925 2015    188 R       441      48.5      36.4 72.9   124
## 13    104925 2016    188 R       263      48.6      36.2 74.5   168
## 14    104925 2017    188 R       138      51.0      37.8 76.6    56
## 15    104925 2018    188 R       286      50.2      38.2 75.7   117
## 16    104925 2019    188 R       332      46.2      36.4 70.4   136
## 17    104925 2020    188 R       257      50.5      38.5 78.4   125
## 18    104925 2021    188 R       416      55.7      43.1 85.4   130
## 19    104925 2022    188 R       244      46.0      36.7 70.1    66
## 20    104925 2023    188 R       295      53.8      42.2 84.9   128
```

```
## 21    104925    2004    188 R          22    57.3      34    93.7    19
## 22    104925    2005    188 R          45    57      37.6   91.3    32
## 23    104925    2006    188 R          63    52.3     34.2   82.2    59
## 24    104925    2007    188 R          98    49      32.2   79.9    48
## 25    104925    2008    188 R          73    53.8     36.6   84.6    40
## 26    104925    2009    188 R          82    53.9     35.9   86.8    51
## 27    104925    2010    188 R          72    61.1     39.9   93.1    84
## 28    104925    2011    188 R          23    57.2     36.6   88.4    12
## 29    104925    2012    188 R          46    54      37.2   87.4    30
## 30    104925    2013    188 R          52    73.1     47.2  110.    24
## 31    104925    2014    188 R          57    60      41.5   91.4    14
## 32    104925    2015    188 R          30    60.2     39.8   91.8    11
## 33    104925    2016    188 R          38    51.8     35     82.1    20
## 34    104925    2017    188 R          31    57.8     38.6   90.1    23
## 35    104925    2018    188 R          56    57.4     38.8   87.1    35
## 36    104925    2019    188 R          60    61.4     40.3   91.3    32
## 37    104925    2020    188 R          21    45.6     31.2   72     12
## 38    104925    2021    188 R          31    56.4     39.6   92     18
## 39    104925    2022    188 R          38    69      45.2  106     22
## 40    104925    2023    188 R          15    66      41    100.    15
```

```
## 'summarise()' has grouped output by 'player_id', 'year', 'height'. You can
## override using the '.groups' argument.
```

```
## # A tibble: 20 x 9
## # Groups:   player_id, year, height [20]
##   player_id year height hand total_aces avg_1stIn avg_1stWon svpt    df
##   <int> <dbl> <int> <fct>    <int>    <dbl>    <dbl> <dbl> <int>
## 1    104925    2004    188 R         26    58.7     36.5   92.3    21
## 2    104925    2005    188 R         88    59.6     41.5   93.9    58
## 3    104925    2006    188 R        279    50.8     35.6   80.8   151
## 4    104925    2007    188 R        518    51.6     36.1   81.7   195
## 5    104925    2008    188 R        486    50.5     36.1   78.4   153
## 6    104925    2009    188 R        502    50.0     35.1   79.9   263
## 7    104925    2010    188 R        304    55.1     37.9   85.3   282
## 8    104925    2011    188 R        343    52.1     35.9   80.2   143
## 9    104925    2012    188 R        502    50.7     36.6   80.5   147
## 10   104925    2013    188 R        476    60.3     41.9   91.0   118
## 11   104925    2014    188 R        428    55.4     40.0   83.6   105
## 12   104925    2015    188 R        471    54.3     38.1   82.4   135
## 13   104925    2016    188 R        301    50.2     35.6   78.3   188
## 14   104925    2017    188 R        169    54.4     38.2   83.4    79
## 15   104925    2018    188 R        342    53.8     38.5   81.4   152
## 16   104925    2019    188 R        392    53.8     38.3   80.8   168
## 17   104925    2020    188 R        278    48.1     34.8   75.2   137
## 18   104925    2021    188 R        447    56.0     41.4   88.7   148
## 19   104925    2022    188 R        282    57.5     40.9   88.1    88
## 20   104925    2023    188 R        310    59.9     41.6   92.6   143
```

```
## # A tibble: 20 x 10
## # Groups:   player_id, year, height [20]
##   player_id year height hand total_aces avg_1stIn avg_1stWon svpt    df
##   <int> <dbl> <int> <fct>    <int>    <dbl>    <dbl> <dbl> <int>
```

```

## 1 104925 2004 188 R 26 58.7 36.5 92.3 21
## 2 104925 2005 188 R 88 59.6 41.5 93.9 58
## 3 104925 2006 188 R 279 50.8 35.6 80.8 151
## 4 104925 2007 188 R 518 51.6 36.1 81.7 195
## 5 104925 2008 188 R 486 50.5 36.1 78.4 153
## 6 104925 2009 188 R 502 50.0 35.1 79.9 263
## 7 104925 2010 188 R 304 55.1 37.9 85.3 282
## 8 104925 2011 188 R 343 52.1 35.9 80.2 143
## 9 104925 2012 188 R 502 50.7 36.6 80.5 147
## 10 104925 2013 188 R 476 60.3 41.9 91.0 118
## 11 104925 2014 188 R 428 55.4 40.0 83.6 105
## 12 104925 2015 188 R 471 54.3 38.1 82.4 135
## 13 104925 2016 188 R 301 50.2 35.6 78.3 188
## 14 104925 2017 188 R 169 54.4 38.2 83.4 79
## 15 104925 2018 188 R 342 53.8 38.5 81.4 152
## 16 104925 2019 188 R 392 53.8 38.3 80.8 168
## 17 104925 2020 188 R 278 48.1 34.8 75.2 137
## 18 104925 2021 188 R 447 56.0 41.4 88.7 148
## 19 104925 2022 188 R 282 57.5 40.9 88.1 88
## 20 104925 2023 188 R 310 59.9 41.6 92.6 143
## # i 1 more variable: aces_in_following_year <int>

## 1 2 3 4
## 415.2551 508.1003 382.2384 331.1461

```