* * * Računarstvo u društvenim naukama * * *

Predmet: Analiza i vizuelizacija podataka
Datum: 11.april 2020. god.

Podaci o bazi

Baza podataka koja je korišćena u ovom radu preuzeta je sa sajta kaggle.com i može joj se pristupiti preko sledećeg linka:
https://www.kaggle.com/pablodroca/atp-tennis-matches-20002019#atp_matches_2017.csv.

Originalno su u bazi sakupljeni podaci o svim ATP (Asocijacija teniskih profesionalaca) mečevima odigranim od 2000. do 2017. godine, ali za potrebe ovog rada iskorišćen je deo baze sa podacima o mečevima iz 2017. godine. Prilikom pretrage baza, bilo je i onih koje su još novijeg datuma i pružaju informacije čak i o 2019. godini, ali s obzirom na to da je 2019. godine uveden novi sistem Davis cup takmičenja što je uticalo na promenu planova i učešća na turnirima nekih tenisera, a kako je 2018. godina obeležena povredama mnogih trenutno vodećih svetskih igrača tenisa, odlučila sam se da ovu skromnu analizu sprovedem nad podacima iz 2017. godine.

Proveravamo u kom smo radnom direktorijumu i učitavamo bazu

```
getwd()
atp_matches_2017<-read.csv("atp_matches_2017.csv", stringsAsFactors = FALSE)
tenis_baza<-atp_matches_2017
```

Upoznavanje sa osnovnim karakteristikama baze i sređivanje za potrebe istraživanja

```
ncol(tenis_baza)
nrow(tenis_baza)
head(tenis_baza,3)
tail(tenis_baza,25)
str(tenis_baza)
```

```
ncol(tenis_baza)
[1] 32

> nrow(tenis_baza)
[1] 2886
```

```
> head(tenis_baza,3)

tourney_id tourney_name tourney_date surface winner_id loser_id        score best_of
round
1   2017-M020      Brisbane      20170102     Hard     104678    106415        6-4 7-5
3    R32
2  2017-M020     Brisbane     20170102    Hard    106378   124014 7-6(4) 7-6(6)       3
R32
3  2017-M020     Brisbane     20170102    Hard    106298   104468 7-6(6) 7-6(4)       3
R32
  minutes  w_ace  w_df  w_svpt  w_1stIn  w_1stWon  w_2ndWon  w_SvGms  w_bpSaved  w_bpFaced
l_ace l_df
1       91     11     5      64       45        35         6       11           1           3
0    1
2      130     11     2      83       48        37        19       12           2           3
11    3
3      125      7     2     102       52        37        24       12           8          12
1    4
  l_svpt   l_1stIn   l_1stWon   l_2ndWon   l_SvGms   l_bpSaved   l_bpFaced   winner_rank
winner_rank_points
1      82        53         33         13        11           6          10            29
1385
2     113        67         39         27        12           9          10            45
1001
3      76        42         29         16        12           0           4            15
2156
  loser_rank loser_rank_points
1        100               604
2        141               443
3         25              1585
> tail(tenis_baza,25)
                       tourney_id               tourney_name tourney_date surface
winner_id
2862                     2017-0352             Paris Masters     20171030    Hard
103898
2863                     2017-0352             Paris Masters     20171030    Hard
104545
2864                     2017-0352             Paris Masters     20171030    Hard
105936
2865                     2017-0352             Paris Masters     20171030    Hard
106058
2866                     2017-0352             Paris Masters     20171030    Hard
105936
2867                     2017-0352             Paris Masters     20171030    Hard
106058
2868                     2017-0605               Tour Finals     20171113    Hard
105676
2869                     2017-0605               Tour Finals     20171113    Hard
106058
2870                     2017-0605               Tour Finals     20171113    Hard
106058
2871                     2017-0605               Tour Finals     20171113    Hard
100644
2872                     2017-0605               Tour Finals     20171113    Hard
103819
```

```
2873                        2017-0605                Tour Finals     20171113    Hard
103819
2874                        2017-0605                Tour Finals     20171113    Hard
103819
2875                        2017-0605                Tour Finals     20171113    Hard
105777
2876                        2017-0605                Tour Finals     20171113    Hard
105676
2877                        2017-0605                Tour Finals     20171113    Hard
105777
2878                        2017-0605                Tour Finals     20171113    Hard
105777
2879                        2017-0605                Tour Finals     20171113    Hard
106233
2880                        2017-0605                Tour Finals     20171113    Hard
105676
2881                        2017-0605                Tour Finals     20171113    Hard
105777
2882                        2017-0605                Tour Finals     20171113    Hard
105777
2883 2017-M-DC-2017-WG-M-BEL-FRA-01 Davis Cup WG F: BEL vs FRA     20171124    Hard
105676
2884 2017-M-DC-2017-WG-M-BEL-FRA-01 Davis Cup WG F: BEL vs FRA     20171124    Hard
104542
2885 2017-M-DC-2017-WG-M-BEL-FRA-01 Davis Cup WG F: BEL vs FRA     20171124    Hard
105676
2886 2017-M-DC-2017-WG-M-BEL-FRA-01 Davis Cup WG F: BEL vs FRA     20171124    Hard
106298
     loser_id                score best_of round minutes w_ace w_df w_svpt w_1stIn
w_1stWon
2862   105227         7-6(5) 7-5        3     QF      123     12     3      86        52
37
2863   105223     6-4 6-7(5) 6-4        3     QF      130     15     4      87        60
51
2864   104745                W/O        3     QF       NA     NA    NA      NA        NA
NA
2865   103898            7-5 6-2        3     SF       78      6     1      52        34
26
2866   104545 6-4 6-7(2) 7-6(5)        3     SF      148     12     2     116        77
61
2867   105936        5-7 6-4 6-1        3      F      118      9     3      72        39
31
2868   104745 7-6(5) 6-7(4) 6-4        3     RR       NA     NA    NA      NA        NA
NA
2869   105227     5-7 6-2 7-6(4)        3     RR       NA      9     2     104        61
46
2870   100644        6-4 1-6 6-4        3     RR      113      5     4      82        56
40
2871   105227        6-4 3-6 6-4        3     RR       NA      9     2      94        52
42
2872   106058         6-4 7-6(4)        3     RR       NA      5     1      63        40
36
2873   105227     6-7(5) 6-4 6-1        3     RR      114     17     2      93        64
49
```

2874	100644	7-6(6) 5-7 6-1	3	RR	NA	6	2	98	52	43		
2875	105676	6-0 6-2	3	RR	NA	2	1	47	27	23		
2876	106233	6-4 6-1	3	RR	71	2	5	58	38	28		
2877	106233	6-3 5-7 7-5	3	RR	NA	5	5	97	52	40		
2878	105807	6-1 6-1	3	RR	60	3	2	36	20	19		
2879	105807	6-3 3-6 6-4	3	RR	NA	12	2	79	41	37		
2880	103819	2-6 6-3 6-4	3	SF	105	7	1	93	55	40		
2881	106058	4-6 6-0 6-3	3	SF	119	3	5	92	59	40		
2882	105676	7-5 4-6 6-3	3	F	150	5	6	103	67	47		
2883	106298	7-5 6-3 6-1	3	RR	119	12	0	74	48	40		
2884	104327	6-3 6-2 6-1	5	RR	106	12	2	64	33	31		
2885	104542	7-6(5) 6-3 6-2	3	RR	164	1	0	122	71	47		
2886	104327	6-3 6-1 6-0	5	RR	94	7	1	57	38	34		

	w_2ndWon	w_SvGms	w_bpSaved	w_bpFaced	l_ace	l_df	l_svpt	l_1stIn	l_1stWon	l_2ndWon	l_SvGms
2862	15	12	1	4	9	1	81	48	32	15	12
2863	18	16	0	0	10	2	90	66	55	12	16
2864	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2865	12	10	3	5	3	5	59	34	24	8	10
2866	23	17	5	5	31	2	92	74	63	12	17
2867	20	14	1	4	5	5	105	72	43	13	15
2868	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2869	23	16	4	7	13	4	101	64	49	18	16
2870	8	13	10	14	5	8	89	59	39	14	14
2871	20	14	4	6	5	1	80	50	37	17	15
2872	15	11	0	0	11	6	86	52	36	20	11
2873	19	15	1	1	12	4	99	50	35	27	14
2874	25	15	9	11	8	8	107	67	49	17	16

2875	11	7	1	1	2	5	57	32	20	5
7										
2876	9	8	5	6	1	6	49	21	14	10
9										
2877	25	17	0	2	8	6	93	44	35	25
16										
2878	9	7	0	0	0	2	48	34	16	4
7										
2879	15	14	0	3	2	1	104	67	41	18
14										
2880	22	14	9	11	7	3	77	47	35	17
13										
2881	17	13	4	6	3	5	76	47	29	13
12										
2882	18	16	7	10	10	5	107	54	42	22
15										
2883	17	14	0	0	6	2	95	53	33	21
14										
2884	19	12	0	0	1	2	81	54	30	12
12										
2885	33	15	10	10	10	3	88	54	44	13
14										
2886	11	11	0	0	1	0	87	53	29	11
11										

	l_bpSaved	l_bpFaced	winner_rank	winner_rank_points	loser_rank	loser_rank_points
2862	3	7	83	634	5	4185
2863	0	2	14	2505	17	2435
2864	NA	NA	77	681	1	10465
2865	4	9	22	1945	83	634
2866	0	1	77	681	14	2505
2867	2	8	22	1945	77	681
2868	NA	NA	8	2975	1	10645
2869	6	10	9	2765	5	3805
2870	6	10	9	2765	3	4410
2871	1	4	3	4410	5	3805
2872	5	6	2	9005	9	2765
2873	6	9	2	9005	5	3805
2874	4	8	2	9005	3	4410
2875	6	11	6	3650	8	2975
2876	2	7	8	2975	4	3815
2877	5	8	6	3650	4	3815
2878	5	10	6	3650	10	2615
2879	9	13	4	3815	10	2615
2880	1	3	8	2975	2	9005
2881	3	8	6	3650	9	2765
2882	11	15	6	3650	8	2975
2883	6	11	7	3775	18	2235
2884	5	11	15	2320	76	667
2885	7	10	7	3775	15	2320
2886	4	11	18	2235	76	667

```
> str(tenis_baza)
'data.frame':  2886 obs. of  32 variables:
 $ tourney_id        : chr  "2017-M020" "2017-M020" "2017-M020" "2017-M020" ...
 $ tourney_name      : chr  "Brisbane" "Brisbane" "Brisbane" "Brisbane" ...
 $ tourney_date       : int  20170102 20170102 20170102 20170102 20170102 20170102
20170102 20170102 20170102 20170102 ...
 $ surface           : chr  "Hard" "Hard" "Hard" "Hard" ...
 $ winner_id         : int  104678 106378 106298 111577 111442 103970 105777 103917
105032 104745 ...
 $ loser_id          : int  106415 124014 104468 104180 111200 106071 105449 103565
105732 105238 ...
 $ score             : chr  "6-4 7-5" "7-6(4) 7-6(6)" "7-6(6) 7-6(4)" "6-4 6-4" ...
 $ best_of           : int  3 3 3 3 3 3 3 3 3 ...
 $ round             : chr  "R32" "R32" "R32" "R32" ...
 $ minutes           : num  91 130 125 75 90 83 75 120 130 74 ...
 $ w_ace             : num  11 11 7 12 1 3 3 13 20 2 ...
 $ w_df              : num  5 2 2 2 0 5 0 3 5 1 ...
 $ w_svpt            : num  64 83 102 55 46 52 48 86 106 52 ...
 $ w_1stIn           : num  45 48 52 33 28 30 25 55 59 36 ...
 $ w_1stWon          : num  35 37 37 27 26 25 23 42 45 26 ...
 $ w_2ndWon          : num  6 19 24 13 6 12 12 17 28 10 ...
 $ w_SvGms           : num  11 12 12 10 8 10 9 15 17 9 ...
 $ w_bpSaved         : num  1 2 8 0 1 2 0 2 8 2 ...
 $ w_bpFaced         : num  3 3 12 1 2 3 1 4 10 3 ...
 $ l_ace             : num  0 11 1 10 1 5 2 9 9 2 ...
 $ l_df              : num  1 3 4 2 6 8 3 8 5 2 ...
 $ l_svpt            : num  82 113 76 58 74 74 57 94 87 53 ...
 $ l_1stIn           : num  53 67 42 37 43 44 36 46 48 28 ...
 $ l_1stWon          : num  33 39 29 27 23 29 21 29 33 20 ...
 $ l_2ndWon          : num  13 27 16 7 13 12 7 30 24 7 ...
 $ l_SvGms           : num  11 12 12 10 9 11 8 14 16 9 ...
 $ l_bpSaved         : num  6 9 0 2 10 5 4 9 4 0 ...
 $ l_bpFaced         : num  10 10 4 5 15 9 8 12 7 4 ...
 $ winner_rank       : num  29 45 15 105 79 21 17 39 180 9 ...
 $ winner_rank_points: num  1385 1001 2156 570 689 ...
 $ loser_rank        : num  100 141 25 34 160 26 33 54 78 62 ...
 $ loser_rank_points : num  604 443 1585 1255 372 ...



>
```

Vidimo da baza pruža podatke o 2.886 mečeva odigranih tokom 2017. godine. Podaci su prikazani kroz 32 varijable: identifikacioni broj turnira, naziv turnira, podloga na kojoj se turnir igra, identifikacioni broj pobednika i poraženog, rezultat, broj odigranih setova, runda turnira, trajanje meča (izraženo u minutima), broj asova pobednika i poraženog, broj duplih grešaka pobednika i poraženog, broj servisa pobednika i poraženog, broj ubačenih prvih servisa pobednika i poraženog, broj osvojenih poena na prvi servis pobednika i poraženog, broj osvojenih poena na drugi servis pobednika i poraženog, broj odserviranih gemova pobednika i poraženog, broj spasenih brejk lopti pobednika i poraženog, broj brejk lopti sa kojim se suočio pobednik odnosno poraženi, rang pobednika i poraženog, broj poena na rang listi pobednika i

poraženog (podaci za pobednika i poraženog su predstavljeni posebnim varijablama). S obzirom na to da je cilj istraživanja provera da li se na osnovu neke od varijabli koje nudi ova baza može predvideti rang tenisera, jasno je da nam sve gorespomenute varijable nisu neophodne i da ih stoga ne bi bilo loše radi bolje preglednosti odmah eliminisati:

```
tenismoja<- tenis_baza[, !(names(tenis_baza) %in% c('tourney_id','tourney_date',
                                                    'winner_id','loser_id',
                                                    'best_of','winner_rank_points',
                                                    'loser_rank_points'))]
```

Iako bi se sa rezultatom poziva funkcije *head()* sasvim dovoljno upoznali sa bazom, značajnije je možda pogledati kraj baze jer, budući da preko 10 godina pratim ovaj sport, znam po rasporedu turnira da ću tako sigurno videti kako su se prilikom pravljenja baze izborili sa završnim masters turnirom i finalom, samim tim i ostalim rundama, Davis cup takmičenja (iz ovog razloga sam od R-a zatražila veći broj podataka o mečevima s kraja baze nego o onima sa početka).
Iako su podaci o Davis cup mečevima prilično popunjeni, s obzirom na to da se učešćem ili pobedom u ovoj vrsti takmičenja ne stiču poeni, pa time oni ne mogu uticati ni na rang igrača, odlučila sam da podatke o ovim mečevima eliminišem. To, pak, nije slučaj sa završnim masters turnirom, gde se, iako je sistem takmičenja drugačiji u odnosu na ostale turnire (u pitanju je takmičenje po grupama, zato u koloni za rundu stoji RR tj. round robin), poeni dobijaju, te iz tog razloga ove mečeva ipak nisam izostavila. Ono što mi je međutim prvo skrenulo pažnju jeste oznaka W/O koja stoji u koloni za rezultat jednog od mečeva. U pitanju je skraćenica od engleske reči walkover, pa stoga nije ni čudno da, pošto meč nije odigran (iz bilo kojeg razloga, najčešće predaje igrača pred početak meča), nemamo podatke za taj meč. Ipak, u kontekstu našeg istraživanja, mečevi za koje nemamo podatke o varijablama koje se odnose na učinke igrača u toku meča, nisu relevantni, stoga je i njih potrebno zanemariti. Osim predaje pre meča, postoji i predaja u toku samog meča (najčešće usled povrede jednog od igrača). Za ove mečeve rezultati su uredno zabeleženi u bazi, ali lično smatram da oni nisu adekvatni jer se ne odnose na ceo meč (predaja može biti i posle prvog gema, i posle prvog seta ili na samom kraju meča) i nisu za mešanje sa onim podacima koji su u tom smislu celoviti. Ovo je razlog što ću ukloniti i sve one mečeve gde se u koloni za rezultat javlja RET (od engleskog retirement).
Sve goreopisano je sprovedeno u sledećim kodovima:

```
teniswalkover<-tenismoja[!(tenismoja$score =="W/O" |
                           tenismoja$score =="Walkover") , ]
tenisdavisrr <-teniswalkover[!grepl("Davis Cup", teniswalkover$tourney_name),]
tenisret<-tenisdavisrr[!grepl("RET", tenisdavisrr$score),]
```

Naposletku, sledi provera da li imamo NA vrednosti. Iako NA vrednosti inicijalno nema, za neke mečeve jednostavno nisu pruženi svi ili pojedini podaci i kao nekompletni smatram da je i njih neophodno eliminisati. Najpre su svi prazni stringovi zamenjeni sa vrednošću NA, a potom uz pomoć funkcije *na.omit()* uklonjeni iz baze, što potvrđuje i krajnji zbir ovih vrednosti čiji je rezultat 0.

```
tenismissing<- replace(tenisret, tenisret== "", NA)
tenis <- na.omit(tenismissing)
sum(is.na(tenis))
```

Napomena: Mečevi za koje se za neku od varijabli javlja vrednost 0 nisu brisani, jer je sasvim normalno da npr. igrač u toku meča ne napravi ni jednu duplu grešku i slično, stoga se to ne tretira kao nedostajući podatak.

Ostaje još samo da pre početka analize, bacimo pogled na sređenu bazu, koja nosi naziv *tenis*.

```
summary(tenis)
str(tenis)
```

```
> summary(tenis)
 tourney_name         surface             score              round              minutes
 Length:2495        Length:2495        Length:2495        Length:2495        Min.   :
36.0
 Class :character   Class :character   Class :character   Class :character   1st Qu.:
79.0
 Mode   :character     Mode   :character     Mode   :character     Mode   :character
Median :102.0

Mean   :108.8
                                                                                3rd
Qu.:132.0

Max.   :987.0
     w_ace             w_df            w_svpt          w_1stIn         w_1stWon
 Min.   : 0.000   Min.   : 0.000   Min.   : 32.00   Min.   : 16.0   Min.   : 12.00
 1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.: 58.00   1st Qu.: 36.0   1st Qu.: 28.00
 Median : 6.000   Median : 2.000   Median : 75.00   Median : 46.0   Median : 35.00
 Mean   : 7.383   Mean   : 2.809   Mean   : 79.89   Mean   : 49.3   Mean   : 37.53
 3rd Qu.:10.000   3rd Qu.: 4.000   3rd Qu.: 96.50   3rd Qu.: 59.5   3rd Qu.: 45.00
 Max.   :75.000   Max.   :14.000   Max.   :213.00   Max.   :153.0   Max.   :127.00
    w_2ndWon        w_SvGms        w_bpSaved        w_bpFaced          l_ace
 Min.   : 2.00   Min.   : 7.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:12.00   1st Qu.:10.00   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 2.000
 Median :16.00   Median :12.00   Median : 3.000   Median : 4.000   Median : 4.000
 Mean   :17.04   Mean   :12.85   Mean   : 3.514   Mean   : 5.063   Mean   : 5.654
 3rd Qu.:21.00   3rd Qu.:15.00   3rd Qu.: 5.000   3rd Qu.: 7.000   3rd Qu.: 8.000
 Max.   :51.00   Max.   :42.00   Max.   :20.000   Max.   :22.000   Max.   :46.000
      l_df            l_svpt          l_1stIn         l_1stWon        l_2ndWon
 Min.   : 0.000   Min.   : 32.00   Min.   : 14.00   Min.   :  4.0   Min.   : 2.00
 1st Qu.: 2.000   1st Qu.: 61.00   1st Qu.: 36.00   1st Qu.: 24.0   1st Qu.:10.00
 Median : 3.000   Median : 78.00   Median : 46.00   Median : 31.0   Median :14.00
 Mean   : 3.505   Mean   : 82.68   Mean   : 49.55   Mean   : 33.6   Mean   :15.44
 3rd Qu.: 5.000   3rd Qu.: 98.50   3rd Qu.: 59.00   3rd Qu.: 41.0   3rd Qu.:20.00
 Max.   :17.000   Max.   :237.00   Max.   :157.00   Max.   :125.0   Max.   :49.00
    l_SvGms        l_bpSaved       l_bpFaced       winner_rank      loser_rank
 Min.   : 6.00   Min.   : 0.00   Min.   : 0.000   Min.   :  1.00   Min.   :   1.00
 1st Qu.:10.00   1st Qu.: 2.00   1st Qu.: 5.000   1st Qu.: 17.00   1st Qu.:  35.00
 Median :12.00   Median : 4.00   Median : 8.000   Median : 40.00   Median :  63.00
```

```
 Mean   :12.61   Mean   : 4.72   Mean   : 8.448   Mean   : 60.18   Mean   :  92.39
 3rd Qu.:15.00   3rd Qu.: 7.00   3rd Qu.:11.000   3rd Qu.: 77.00   3rd Qu.: 101.50
 Max.   :41.00   Max.   :24.00   Max.   :30.000   Max.   :993.00   Max.   :1360.00
> str(tenis)
'data.frame':  2495 obs. of  25 variables:
 $ tourney_name: chr  "Brisbane" "Brisbane" "Brisbane" "Brisbane" ...
 $ surface     : chr  "Hard" "Hard" "Hard" "Hard" ...
 $ score       : chr  "6-4 7-5" "7-6(4) 7-6(6)" "7-6(6) 7-6(4)" "6-4 6-4" ...
 $ round       : chr  "R32" "R32" "R32" "R32" ...
 $ minutes     : num  91 130 125 75 90 83 75 120 130 74 ...
 $ w_ace       : num  11 11 7 12 1 3 3 13 20 2 ...
 $ w_df        : num  5 2 2 2 0 5 0 3 5 1 ...
 $ w_svpt      : num  64 83 102 55 46 52 48 86 106 52 ...
 $ w_1stIn     : num  45 48 52 33 28 30 25 55 59 36 ...
 $ w_1stWon    : num  35 37 37 27 26 25 23 42 45 26 ...
 $ w_2ndWon    : num  6 19 24 13 6 12 12 17 28 10 ...
 $ w_SvGms     : num  11 12 12 10 8 10 9 15 17 9 ...
 $ w_bpSaved   : num  1 2 8 0 1 2 0 2 8 2 ...
 $ w_bpFaced   : num  3 3 12 1 2 3 1 4 10 3 ...
 $ l_ace       : num  0 11 1 10 1 5 2 9 9 2 ...
 $ l_df        : num  1 3 4 2 6 8 3 8 5 2 ...
 $ l_svpt      : num  82 113 76 58 74 74 57 94 87 53 ...
 $ l_1stIn     : num  53 67 42 37 43 44 36 46 48 28 ...
 $ l_1stWon    : num  33 39 29 27 23 29 21 29 33 20 ...
 $ l_2ndWon    : num  13 27 16 7 13 12 7 30 24 7 ...
 $ l_SvGms     : num  11 12 12 10 9 11 8 14 16 9 ...
 $ l_bpSaved   : num  6 9 0 2 10 5 4 9 4 0 ...
 $ l_bpFaced   : num  10 10 4 5 15 9 8 12 7 4 ...
 $ winner_rank : num  29 45 15 105 79 21 17 39 180 9 ...
 $ loser_rank  : num  100 141 25 34 160 26 33 54 78 62 ...
 - attr(*, "na.action")= 'omit' Named int  88 110 111 133 307 545 821 822 825 829 ...
  ..- attr(*, "names")= chr  "90" "113" "114" "139" ...
```

Vidimo da je broj kolona tj. varijabli smanjen (sada je 25, umesto 32 koliko je bio na početku), baš kao i broj mečeva koji dolaze u obzir za opservaciju, sada on iznosi 2.495, što je i dalje pozamašan uzorak, pa ovo smanjenje neće uticati na relevantnost dobijenih rezultata analize, dok nam s druge strane pruža sigurnost da su naši podaci kompletni, bez nedostataka i spremni za analizu.

Pre same analize, bilo mi je zanimljivo da proverim par inicijalnih stvari koje su se dale videti iz same funkcije *summary()*. Možemo videti da se u jednom meču poraženi najviše suočio sa čak 30 brejk lopti. Nije bilo teško zaključiti da i nije baš najbolje rangiran:

```
which(tenis$l_bpFaced==max(tenis$l_bpFaced))
[1] 1244
tenis[1244,'loser_rank']
[1] 470
```

Interesantno je videti i na kojoj to podlozi je pobednik uspeo da odservira najviše, čak 75 asova. Koliko je taj meč trajao? Mora da se igrao u nedogled u petom setu…

```
which(tenis$w_ace==max(tenis$w_ace))
tenis[178,c('surface','minutes','score')]

surface    minutes      score
Hard       314          6-7(6) 3-6 7-5 6-2 22-20
```

Igralo se preko pet sati, pa nije ni čudo da se igrač toliko naservirao asova (iako bih pre očekivala da se to desilo na travi kao bržoj podlozi)!

Istraživačka pitanja

1. Da li postoji veza između broja asova i duplih grešaka, kao i broja ubačenih prvih servisa sa dužinom trajanja meča?
Pretpostavka je da veza postoji sa ubačenim prvim servisom, jer neubačeni prvi, automatski podrazumeva serviranje i drugog servisa, što oduzima vreme.

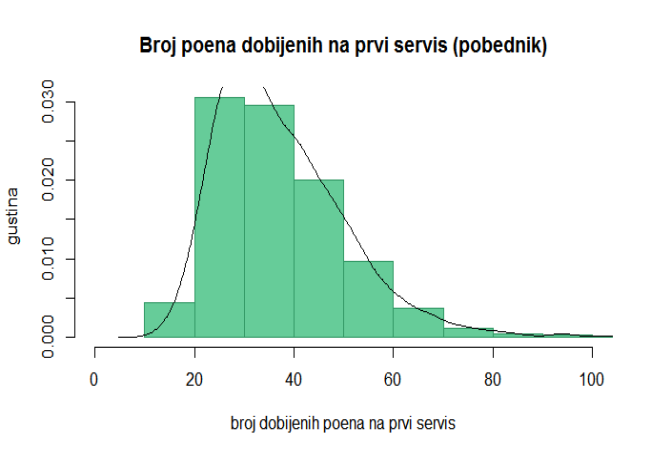
2. Da li postoji razlika u broju asova, duplih grešaka, broju ubačenih prvih servisa, osvojenih poena na prvi i drugih servisa, broju spasenih brejk lopti i onih sa kojima se pobednik suočio, a kada se porede različita kola turnira?
Pretpostavka je da se bar između nekih kola turnira, u nekim od pomenutih od brojeva, javlja razlika, pre svega zbog kvaliteta suparnika i samim tim kvaliteta igre na početku i kraju turnira.

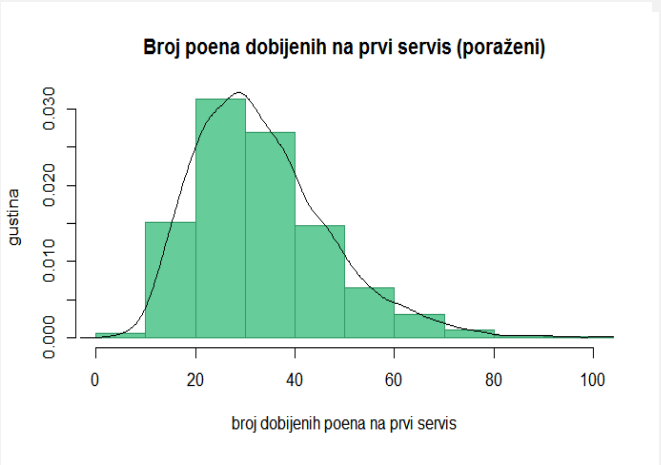
3. Da li se na osnovu broja asova, duplih grešaka, ubačenih prvih servisa, osvojenih poena na prvi odnosno drugi servis i broja spasenih brejk lopti može predvideti rang tog igrača?
Pretpostavka je da može, ali u jako maloj meri. Ipak je u tenisu dosta i psihičke i taktičke igre koja nema veze niti sa rangom niti sa statistikom.

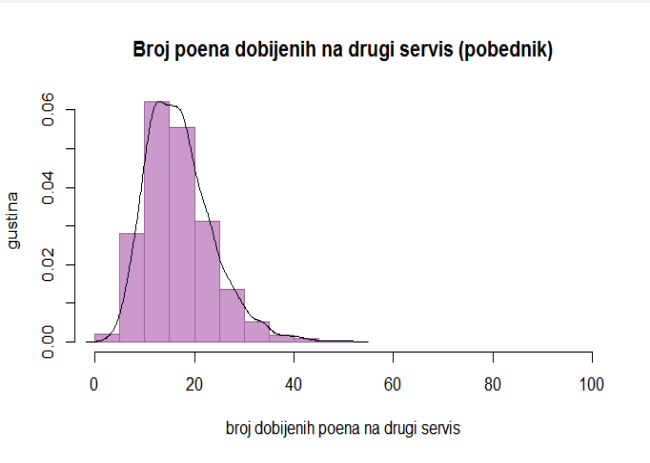
Provera normalnosti raspodele

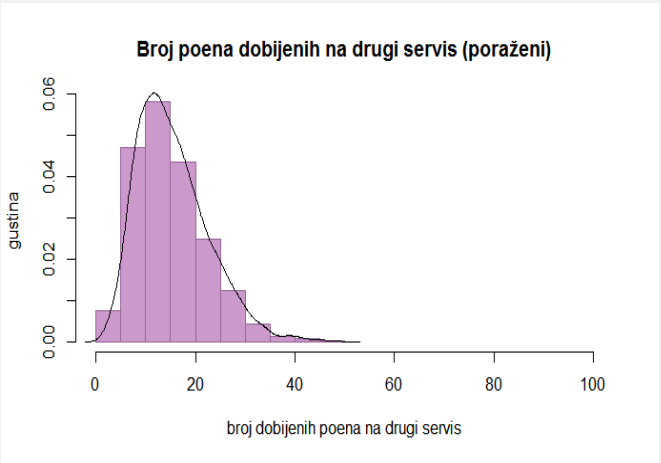
Pre nego započnemo analize, neophodno je najpre proveriti da li varijable koje ulaze u analizu imaju normalnu raspodelu ili ne. Isključivo od ovih rezultata zavisiće koje testove ćemo upotrebiti, da li one parametarske (ukoliko je raspodela normalna) ili njihove neparametarske alternative (u slučaju da odbacujemo nultu hipotezu i donosimo zaključak da je alternativna hipoteza prihvatljivija tj. da raspodela nije normalna). Normalnost možemo proveriti najpre vizuelno, pomoću histograma. Sledi primer koda koji je korišćen za pravljenje histograma svih varijabli koje ulaze u analizu (samo su menjani nazivi labela i varijabli, ostalo je potpuno istovetno), kao i dobijene vizualizacije:

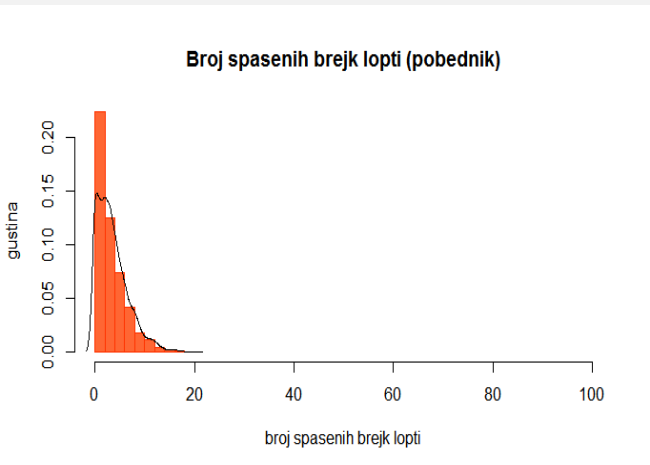
```
hist(tenis$w_ace, main ="Broj asova (pobednik)" ,
     xlab = "broj asova", ylab = "gustina",border ="#993300",col = "#FF9933",
     xlim = c(0,100),freq = FALSE)
lines(density(tenis$w_ace))
```

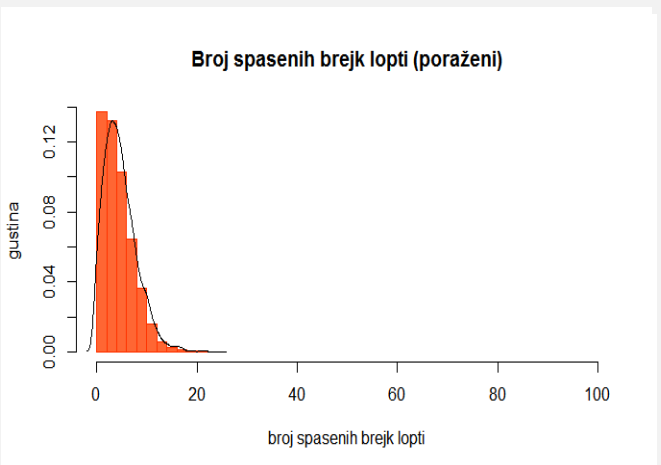

Broj poena dobijenih na prvi servis (pobednik)
gustina
0.000
0.010
0.020
0.030
0
20
40
60
80
100
broj dobijenih poena na prvi servis

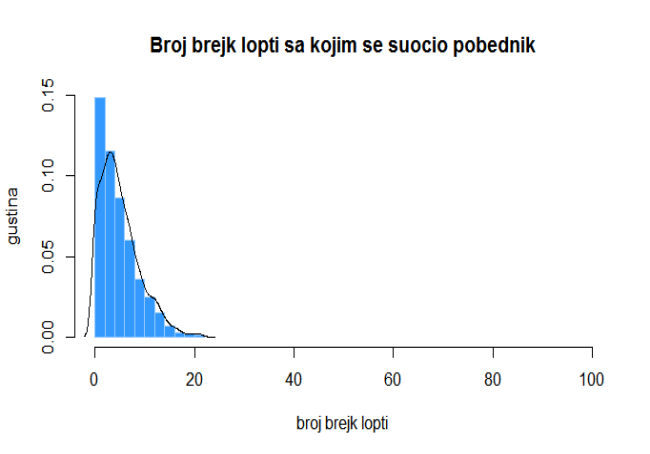

Broj poena dobijenih na prvi servis (poraženi)
gustina
0.000
0.010
0.020
0.030
0
20
40
60
80
100
broj dobijenih poena na prvi servis

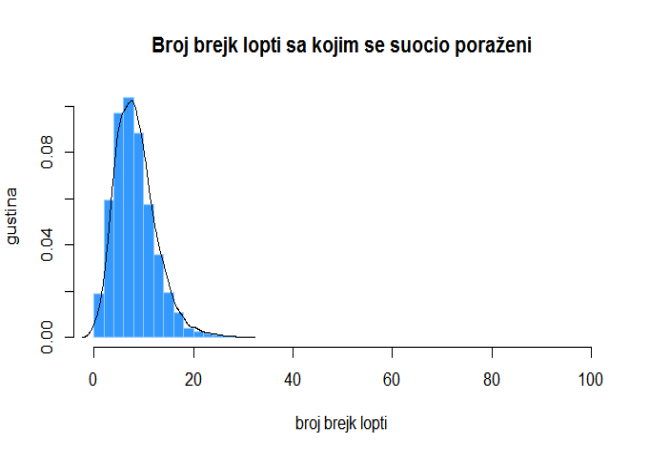

Broj poena dobijenih na drugi servis (pobednik)
gustina
0.00
0.02
0.04
0.06
0
20
40
60
80
100
broj dobijenih poena na drugi servis

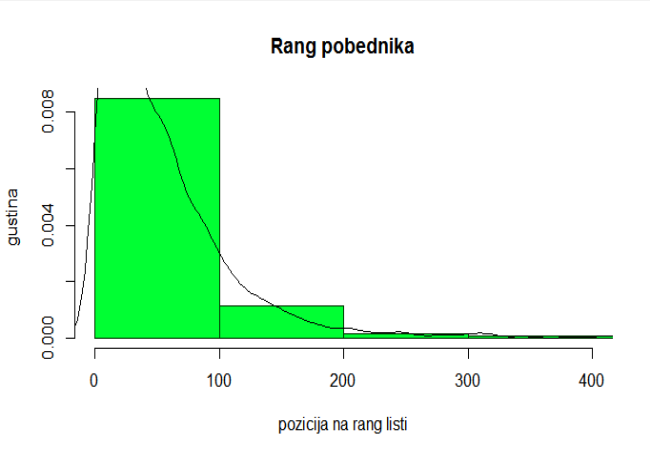

Broj poena dobijenih na drugi servis (poraženi)
gustina
0.00
0.02
0.04
0.06
0
20
40
60
80
100
broj dobijenih poena na drugi servis

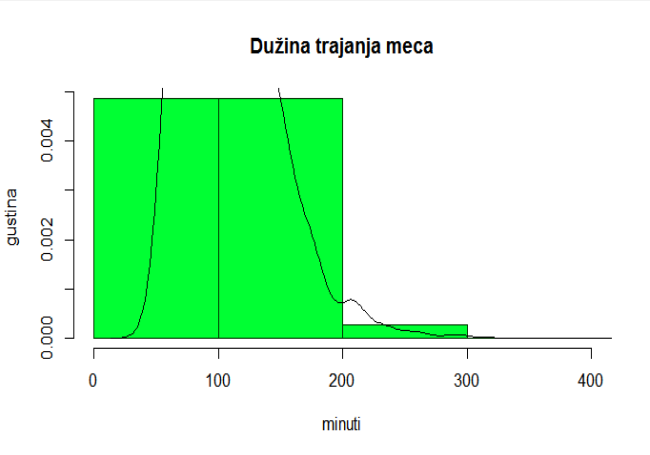

Broj spasenih brejk lopti (pobednik)
gustina
0.00
0.05
0.10
0.15
0.20
0
20
40
60
80
100
broj spasenih brejk lopti


Broj spasenih brejk lopti (poraženi)
gustina
0.00
0.04
0.08
0.12
0
20
40
60
80
100
broj spasenih brejk lopti









Kao što se iz priloženih histograma da zaključiti, sve varijable prilično odstupaju od normalne raspodele. To jedino, kako se čini, nije slučaj sa varijablama *broj ubačenih prvih servisa pobednika* i *broj ubačenih prvih servisa poraženog*. Ipak, kako oko nekad ume da vara, najsigurnije je date zaključke proveriti i uz pomoć testova. Iako je Kolmogorov-Smirnov test možda poznatiji, Šapirov test je jači i stoga sam se odlučila za njega. Slede rezultati testa za sve varijable za koje su i pravljeni histogrami:

```
> shapiro.test(tenis$w_ace)

        Shapiro-Wilk normality test

data:  tenis$w_ace
W = 0.86314, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$l_ace)

        Shapiro-Wilk normality test

data:  tenis$l_ace
```

```
W = 0.83066, p-value < 2.2e-16

> #nije normalna
> shapiro.test(tenis$w_df)

        Shapiro-Wilk normality test

data:  tenis$w_df
W = 0.90156, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$l_df)

        Shapiro-Wilk normality test

data:  tenis$l_df
W = 0.92694, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$w_1stIn)

        Shapiro-Wilk normality test

data:  tenis$w_1stIn
W = 0.93854, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$l_1stIn)

        Shapiro-Wilk normality test

data:  tenis$l_1stIn
W = 0.93896, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$w_1stWon)

        Shapiro-Wilk normality test

data:  tenis$w_1stWon
W = 0.93066, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$l_1stWon)

        Shapiro-Wilk normality test

data:  tenis$l_1stWon
W = 0.94404, p-value < 2.2e-16
```

```
> #nije normalna

> shapiro.test(tenis$w_2ndWon)

        Shapiro-Wilk normality test

data:  tenis$w_2ndWon
W = 0.95737, p-value < 2.2e-16

> #nije normalna


> shapiro.test(tenis$l_2ndWon)

        Shapiro-Wilk normality test

data:  tenis$l_2ndWon
W = 0.95422, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$w_bpSaved)

        Shapiro-Wilk normality test

data:  tenis$w_bpSaved
W = 0.89335, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$l_bpSaved)

        Shapiro-Wilk normality test

data:  tenis$l_bpSaved
W = 0.93101, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$w_bpFaced)

        Shapiro-Wilk normality test

data:  tenis$w_bpFaced
W = 0.91821, p-value < 2.2e-16

> #nije normalna

> shapiro.test(tenis$l_bpFaced)

        Shapiro-Wilk normality test

data:  tenis$l_bpFaced
W = 0.9573, p-value < 2.2e-16
```

```
> #nije normalna
>
> shapiro.test(tenis$minutes)

        Shapiro-Wilk normality test

data:  tenis$minutes
W = 0.82787, p-value < 2.2e-16

> #nije normalna
> shapiro.test(tenis$winner_rank)

        Shapiro-Wilk normality test

data:  tenis$winner_rank
W = 0.60949, p-value < 2.2e-16

> #nije normalna
```

Rezultati testa potkrepljuju pretpostavke iznete nakon pregleda histograma. Jedino varijable *broj ubačenih prvih servisa pobednika* i *broj ubačenih prvih servisa poraženog* imaju normalnu raspodelu.

Prvo istraživačko pitanje

1. Da li postoji veza između broja asova i duplih grešaka, kao i broja ubačenih prvih servisa sa dužinom trajanja meča?

Za odgovor na ovo pitanje ćemo se poslužiti testom korelacije. Kako nisu sve varijable koje koreliramo sa normalnom raspodelom, odlučila sam se za Spirmanovu korelaciju, ne Pirsonovu. Nezgoda sa testom korelacije (kada niste u mogućnosti da instalirate neophodne pakete) je što on sam po sebi ne daje i *p vrednost*, koja nam je neophodna da bismo videli da li je veza statistički značajna. Stoga smo formirali pomoćnu tabelu u koju smo ubacili sve varijable koje ubacujemo i u korelaciju. Potom smo napravili funkciju koja uzima po dve varijable iz pomoćne tabele, vrši korelaciju i ispisuje p vrednost (uz pomoć *expand.grid()* ćemo naterati R da korelira svaku varijablu sa svakom), i ovu funkciju primenili na našu pomoćnu tabelu. Rezultate smo predstavili u matrici.

Napomena: Kako će isti slučaj biti u daljim analizama, ponovo ćemo koristiti isti princip i istu funkciju da dođemo do p vrednosti.

#kod

```
cor(tenis[,5],tenis[,c(6,7,9,15,16,18)],method = "spearman")
korelacija_pomocna<-data.frame(tenis$minutes, tenis$w_ace, tenis$w_df, tenis$l_ace,
                               tenis$l_df,tenis$w_1stIn, tenis$l_1stIn)
svaka_sa_svakom_cor = expand.grid(names(korelacija_pomocna),
                                  names(korelacija_pomocna))
funkcija_za_p = function(col_name1, col_name2, data_frame) {
  cor.test(data_frame[[col_name1]], data_frame[[col_name2]])$p.value
pvrednosti_kor <- mapply(funkcija_za_p,
                     col_name1 = svaka_sa_svakom_cor[[1]],
                     col_name2 = svaka_sa_svakom_cor[[2]],
                     MoreArgs = list(data_frame = korelacija_pomocna))

matrix(pvrednosti_kor, 7, 7, dimnames = list(names(korelacija_pomocna),
                                             names(korelacija_pomocna)))
```

#rezultati

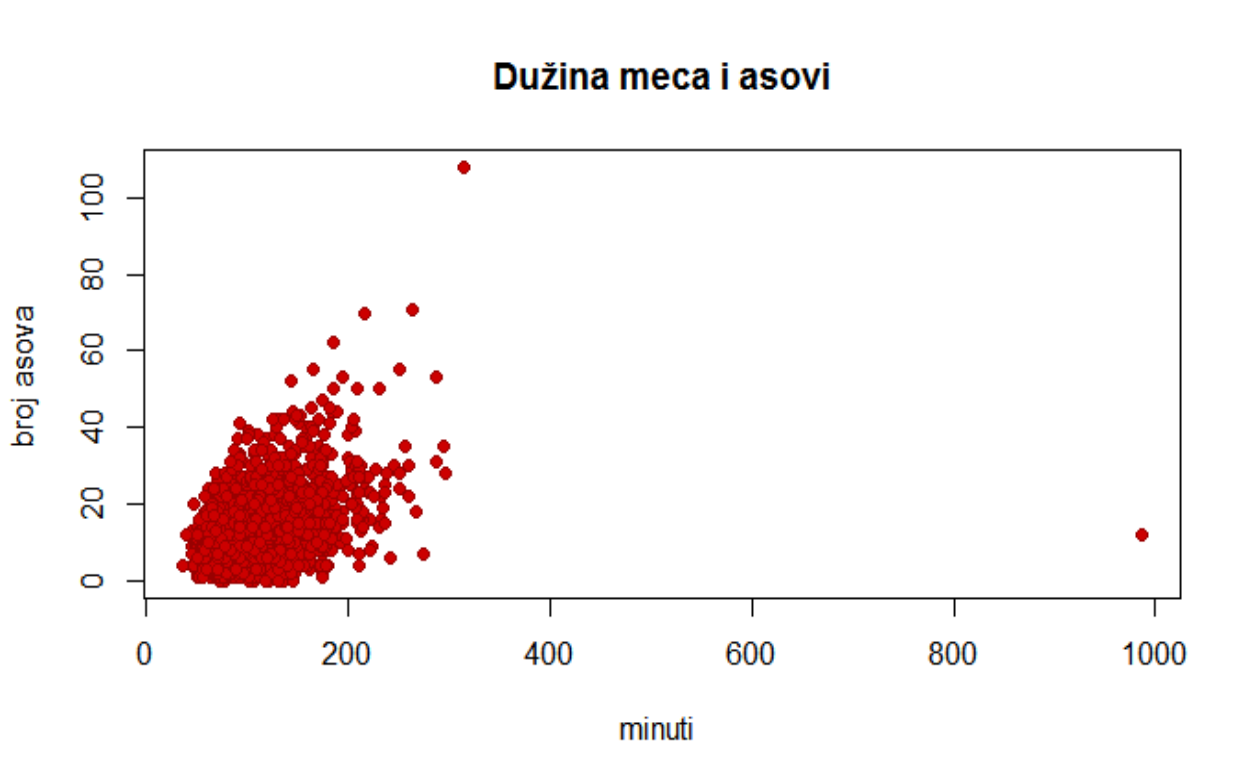
```
cor(tenis[,5],tenis[,c(6,7,9,15,16,18)],method = "spearman")
         w_ace      w_df   w_1stIn     l_ace      l_df   l_1stIn
[1,] 0.2362736 0.3939676 0.8729846 0.4031716 0.2861095 0.8454377
```

```
              tenis.minutes  tenis.w_ace   tenis.w_df   tenis.l_ace   tenis.l_df
tenis.minutes  0.000000e+00 1.791502e-45 1.392409e-86  9.896777e-93 6.800619e-57
tenis.w_ace    1.791502e-45 0.000000e+00 1.226156e-35  1.886197e-44 6.499186e-06
tenis.w_df     1.392409e-86 1.226156e-35 0.000000e+00  3.166564e-25 1.158061e-36
tenis.l_ace    9.896777e-93 1.886197e-44 3.166564e-25  0.000000e+00 3.630961e-43
tenis.l_df     6.800619e-57 6.499186e-06 1.158061e-36  3.630961e-43 0.000000e+00
tenis.w_1stIn  0.000000e+00 1.373679e-71 5.440326e-75 7.999133e-136 5.403268e-52
tenis.l_1stIn  0.000000e+00 2.877974e-92 1.355724e-65 1.403963e-120 1.820467e-36
              tenis.w_1stIn tenis.l_1stIn
tenis.minutes  0.000000e+00  0.000000e+00
tenis.w_ace    1.373679e-71  2.877974e-92
tenis.w_df     5.440326e-75  1.355724e-65
tenis.l_ace   7.999133e-136 1.403963e-120
tenis.l_df     5.403268e-52  1.820467e-36
tenis.w_1stIn  0.000000e+00  0.000000e+00
tenis.l_1stIn  0.000000e+00  0.000000e+00
```

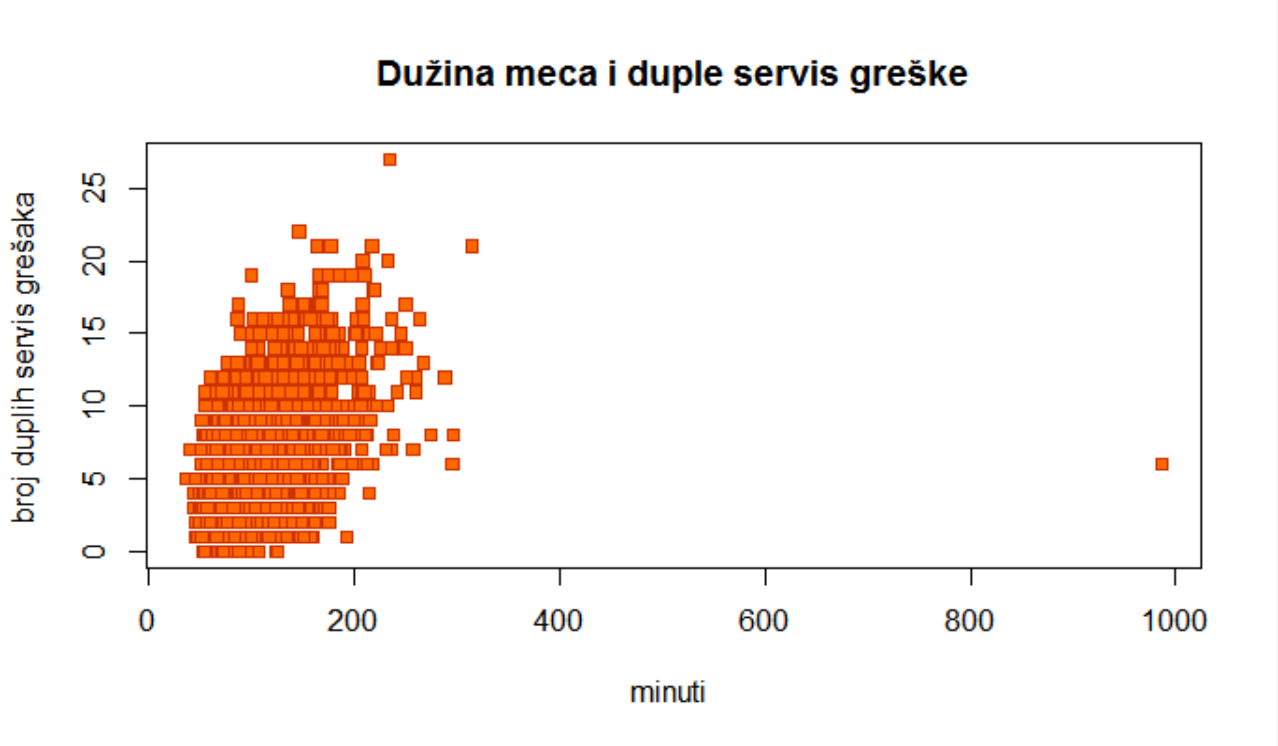
Iz prve tabele možemo videti da korelacija slabe ili eventualno umerene jačine postoji sa brojem asova i duplih grešaka, ali, kao što je i bilo i očekivati, korelacija je jaka i pozitivna (0.8) kada je reč o ubačenom prvom servisu. Druga tabela nam daje *p vrednosti* korelacije svake varijable sa svakom, pri čemu možemo videti da je korelacija varijable *dužina trajanja meča* statistički značajna sa svim ostalim sa kojima je korelirana (*p* je čak daleko ispod 0.01, iako će se u ovom radu uzimati 95%-ni interval poverenja za odlučivanje o odbacivanju nulte hipoteze).

Sledi grafički prikaz korelacije broja asova, broja duplih grešaka i broja ubačenih prvih servisa (svake varijable ponaosob) sa dužinom trajanja meča:

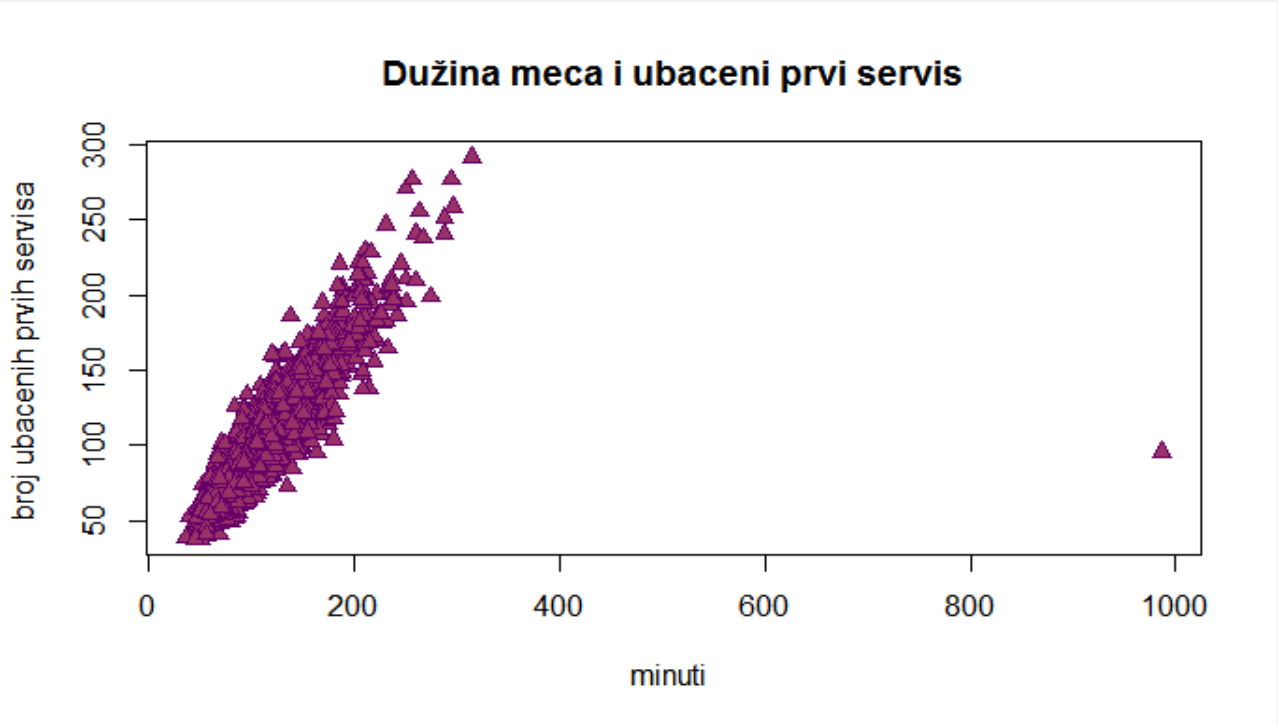
```
plot(tenis$minutes,tenis$w_ace+tenis$l_ace , main="Dužina meča i asovi",
     xlab="minuti ", ylab="broj asova ", type="p",pch=21,col='#990000',bg='#CC0000')
```




```
plot(tenis$minutes,tenis$w_df+tenis$l_df , main="Dužina meča i duple servis greške",
     xlab="minuti ", ylab="broj duplih servis grešaka ", type="p",pch=22,
     col='#CC3300',bg='#FF6600')
```

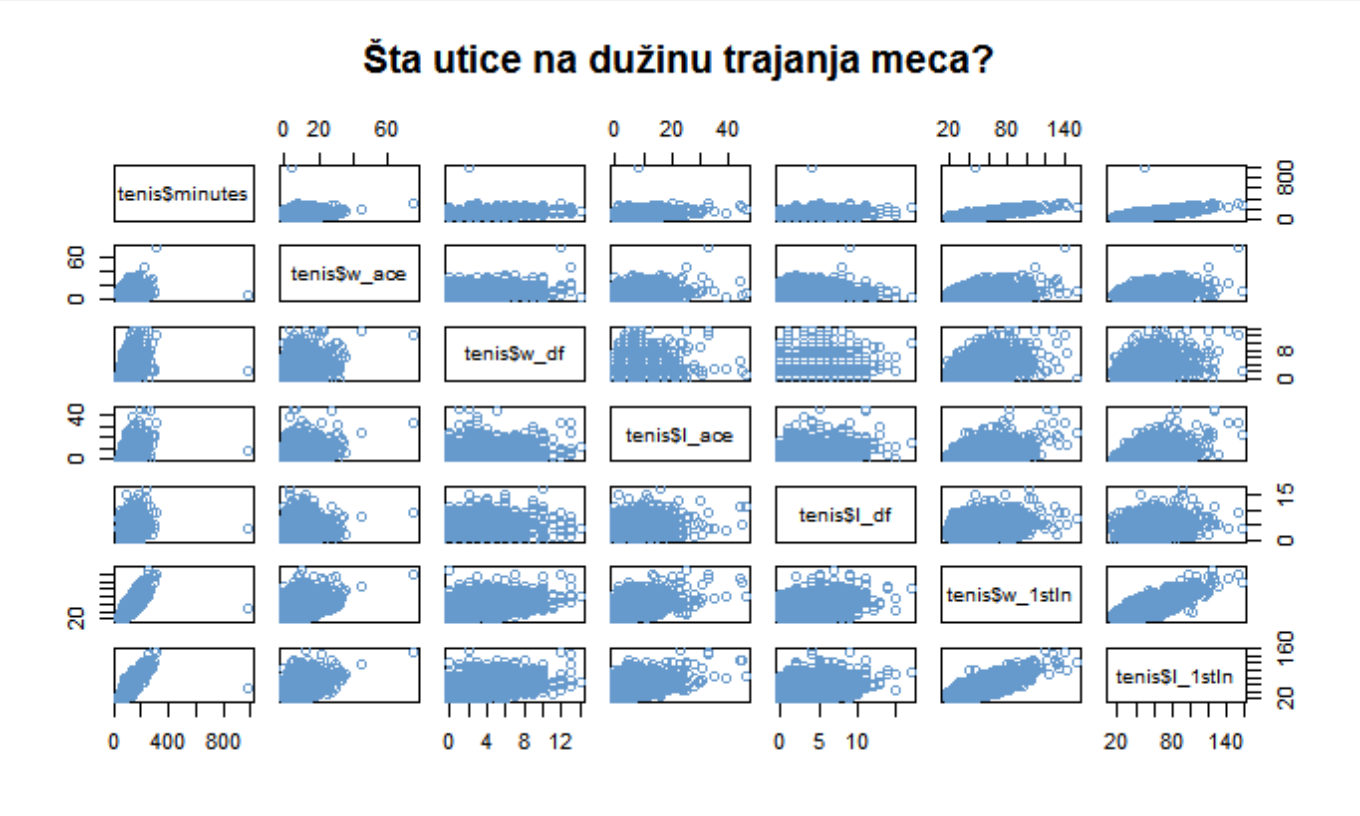



```
plot(tenis$minutes,tenis$w_1stIn+tenis$l_1stIn  ,  main="Dužina  meča  i  ubačeni  prvi
servis",
     xlab="minuti ", ylab="broj ubačenih prvih servisa ", type="p",pch=24,
     col='#660066',bg='#993366')
```



Zbirno bi to izgledalo ovako:

```
pairs(~tenis$minutes+tenis$w_ace+tenis$w_df+tenis$l_ace+tenis$l_df+tenis$w_1stIn+
      tenis$l_1stIn,data=tenis, main="Šta utiče na dužinu trajanja meča?",
col='#6699CC')
```



Grafički prikaz jasno potvrđuje rezultate analize i nesumnjivo pokazuje jaku korelaciju jedino sa brojem ubačenih prvih servisa (trouglići/ kružići se grupišu ka jednoj liniji sa smerom na gore, što ukazuje na pozitivnu linaernu povezanost ovih varijabli - što je veći broj ubačenih servisa to je i dužina trajanja meča veća).

Rezultati potvrđuju prvobitnu pretpostavku i donekle opravdavaju uvođenje vremenskog ograničenja (25 sekundi) između dva servisa/ poena (iako moja malenkost smatra da je ova mera, uvedena na US open-u 2018. godine, vrlo surova prema igračima).

Drugo istraživačko pitanje

2. Da li postoji razlika u broju asova, duplih grešaka, broju ubačenih prvih servisa, osvojenih poena na prvi i drugih servisa, broju spasenih brejk lopti i onih sa kojima se pobednik suočio, a kada se porede različita kola turnira?

Za potrebe ovog istraživačkog pitanja iskoristićemo test poređenja grupa, jer su kola turnira svojevrsne grupe po kojima poredimo navedene varijable. Najpre ćemo kola pretvoriti u faktore, a potom i izmeniti redosled „nivoa" jer onaj koji je R napravio nije odgovarajući (Napomena: RR je stavljeno čak posle finala, jer jedino gde je u bazi ostavljen tj. nije izbrisan ovaj sistem takmičenja jeste završni masters turnir koji okuplja samo 8 najboljih tenisera u tom trenutku, te se kvalitet smatra izuzetnim).

```
roundfactor <- factor(tenis$round)
levels(roundfactor)
```

```
> levels(roundfactor)
[1] "F"    "QF"   "R128" "R16"  "R32"  "R64"  "RR"   "SF"
```

```
roundfactor_ok <- factor(x=tenis$round, levels = c('R128', 'R64', 'R32','R16',
                                                   'QF','SF','F','RR'))
levels(roundfactor_ok)
```

```
> levels(roundfactor_ok)
[1] "R128" "R64"  "R32"  "R16"  "QF"   "SF"   "F"    "RR"
```

Provera normalnosti varijabli je pokazala da jedino broj ubačenih prvih servisa ima normalnu raspodelu. Zato ćemo na ovu varijablu primeniti parametarski test za poređenje grupa tzv. ANOVA test.

```
round_anova<- aov(tenis$w_1stIn ~ roundfactor_ok, data = tenis)
summary(round_anova)
```

```
summary(round_anova)
                Df Sum Sq Mean Sq F value Pr(>F)
roundfactor_ok    7  64784    9255   29.97 <2e-16 ***
Residuals      2487 768048     309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dati rezultat ANOVA testa nam pokazuje da postoji statistički značajna razlika između broja ubačenih prvih servisa i kola turnira, ali, kao i kod svih testova poređenja grupa, ne znamo između kojih to tačno grupa/ kola turnira postoji razlika, pa je neophodno sprovesti tzv. posthoc testove. Ovde će to biti Tukejev test. Slede rezultati.

```
TukeyHSD(round_anova)
```

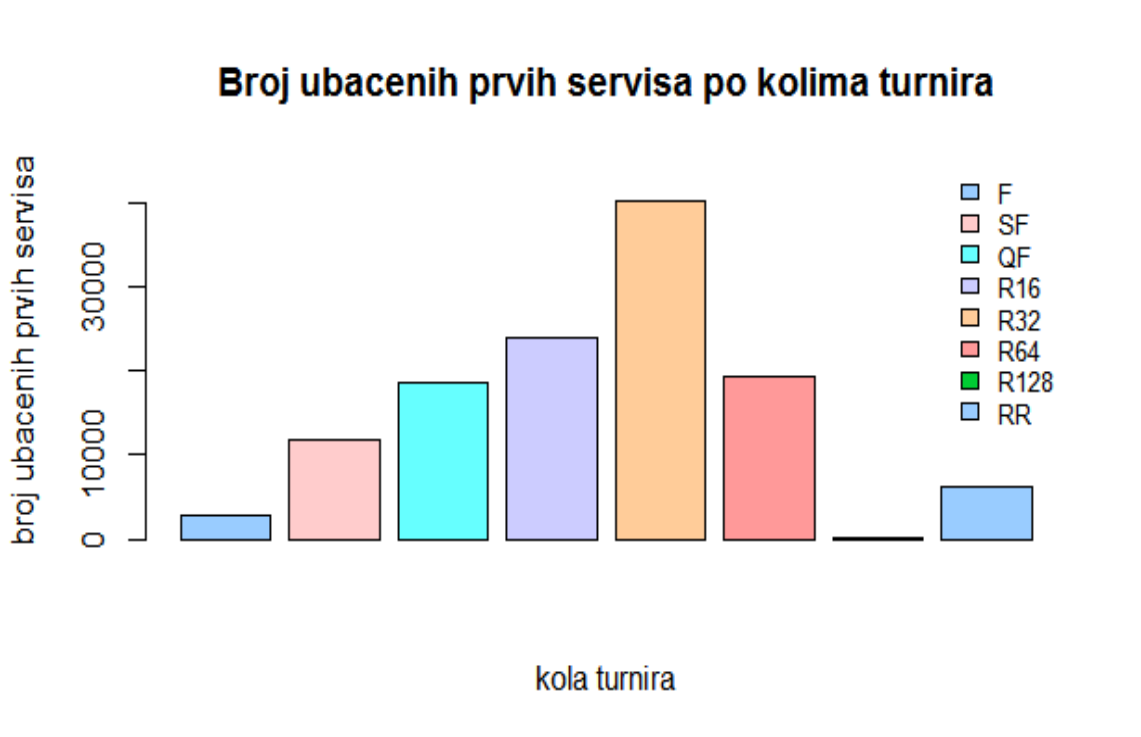
```
TukeyHSD(round_anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = tenis$w_1stIn ~ roundfactor_ok, data = tenis)

$roundfactor_ok
                diff        lwr         upr     p adj
R64-R128 -10.8886472 -15.013037  -6.7642576 0.0000000
R32-R128 -15.9270520 -19.498883 -12.3552207 0.0000000
R16-R128 -15.4446758 -19.324852 -11.5645001 0.0000000
QF-R128  -15.0200000 -19.585069 -10.4549308 0.0000000
SF-R128  -13.5907087 -19.234203  -7.9472142 0.0000000
F-R128   -16.5615873 -23.949454  -9.1737204 0.0000000
RR-R128  -17.7600000 -44.591368   9.0713685 0.4766089
R32-R64   -5.0384048  -8.328275  -1.7485349 0.0000967
R16-R64   -4.5560286  -8.178325  -0.9337325 0.0034842
QF-R64    -4.1313528  -8.479355   0.2166493 0.0765918
SF-R64    -2.7020614  -8.171458   2.7673348 0.8084503
F-R64     -5.6729401 -12.928686   1.5828054 0.2557556
RR-R64    -6.8713528 -33.666643  19.9239378 0.9942296
R16-R32    0.4823762  -2.495613   3.4603651 0.9997002
QF-R32     0.9070520  -2.920808   4.7349119 0.9964674
SF-R32     2.3363434  -2.729393   7.4020796 0.8579872
F-R32     -0.6345353  -7.591060   6.3219891 0.9999939
RR-R32    -1.8329480 -28.548767  24.8828711 0.9999992
QF-R16     0.4246758  -3.692398   4.5417496 0.9999859
SF-R16     1.8539672  -3.433704   7.1416386 0.9641329
F-R16     -1.1169115  -8.236674   6.0028510 0.9997578
RR-R16    -2.3153242 -29.074113  24.4434646 0.9999958
SF-QF      1.4292913  -4.379630   7.2382123 0.9955315
F-QF      -1.5415873  -9.056580   5.9734051 0.9985890
RR-QF     -2.7400000 -29.606650  24.1266497 0.9999869
F-SF      -2.9708786 -11.185760   5.2440032 0.9575736
RR-SF     -4.1692913 -31.240051  22.9014685 0.9997858
RR-F      -1.1984127 -28.685821  26.2889951 1.0000000
```

Statistički značajna i izrazito jaka razlika se javlja između prvog kola gren slem turnira i svih ostalih rundi, kako je runda bliža finalu to je razlika veća i tako sve do najveće razlike koja se javlja sa brojem ubačenih prvih servisa na završnom masters turniru (što je itekako razumljivo kada se u obzir uzme koji igrači mogu da se nađu u prvom kolu nekog gren slema odnosno na završnom turniru u Londonu).

Kako ostale varijable koje smo želeli da ispitamo u kontekstu kola turnira nemaju normalnu raspodelu, za njih ćemo koristiti neparametarsku alternativu ANOVA testa - Kruskal-Volisov test, a kao posthoc test koji će nam detaljnije prikazati razlike između pojedinačnih grupa koristićemo Vilkoksonov test. Slede rezultati oba testa, za svaku varijablu ponaosob.

Grafički prikaz svih varijabli (pa i ubačenog prvog servisa) obuhvatao je najpre agregaciju, sabiranje vrednosti date varijable za svako kolo ponaosob, a potom predstavljanje uz pomoć barplot-a i dodavanje legende. Sledi kod i rezultat plotovanja.

```
tenis3<-tenis
moje_boje=c("#99CCFF",  "#FFCCCC", "#66FFFF","#CCCCFF","#FFCC99","#FF9999",
            "#00CC33")

ubaceniprvi<-aggregate(tenis3$w_1stIn, by=list(Category=tenis3$round), FUN=sum)

barplot(height = ubaceniprvi$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Broj ubačenih prvih servisa po kolima
turnira",
        xlab = "kola turnira",ylab = "broj ubačenih prvih servisa",ylim = c(0,45000))
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



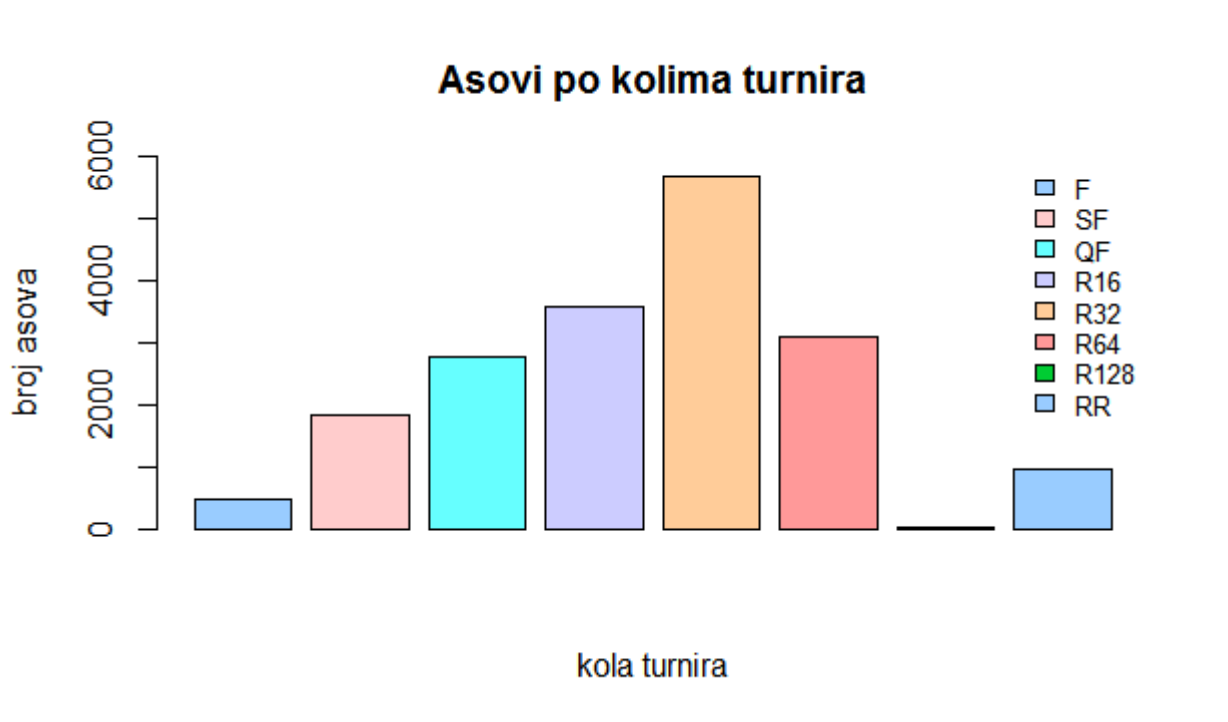
Broj asova

```
kruskal.test (tenis$w_ace ~ roundfactor_ok, data = tenis)
pairwise.wilcox.test(tenis$w_ace, roundfactor_ok,
                     p.adjust.method = "BH")
```

```
kruskal.test (tenis$w_ace ~ roundfactor_ok, data = tenis)

        Kruskal-Wallis rank sum test

data:  tenis$w_ace by roundfactor_ok
Kruskal-Wallis chi-squared = 38.22, df = 7, p-value = 2.752e-06

> pairwise.wilcox.test(tenis$w_ace, roundfactor_ok,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  tenis$w_ace and roundfactor_ok

    R128    R64     R32     R16     QF      SF      F
R64 0.15461 -       -       -       -       -       -
R32 3.4e-07 0.00551 -       -       -       -       -
R16 0.00013 0.09859 0.70202 -       -       -       -
QF  0.01154 0.57346 0.31721 0.70202 -       -       -
SF  0.09859 0.70202 0.41322 0.70202 0.93879 -       -
F   0.30662 0.70202 0.70202 0.70202 0.96637 0.96637 -
RR  0.70202 0.70202 0.79800 0.79800 0.70202 0.70202 0.70202

P value adjustment method: BH
```

Grafički prikaz ukupnog broja asova prikazanog po kolima

```
asovi<-aggregate(tenis3$w_ace, by=list(Category=tenis3$round), FUN=sum)
barplot(height = asovi$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Asovi po kolima turnira",
        xlab = "kola turnira",ylab = "broj asova",ylim = c(0,6000))
# Add legend
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



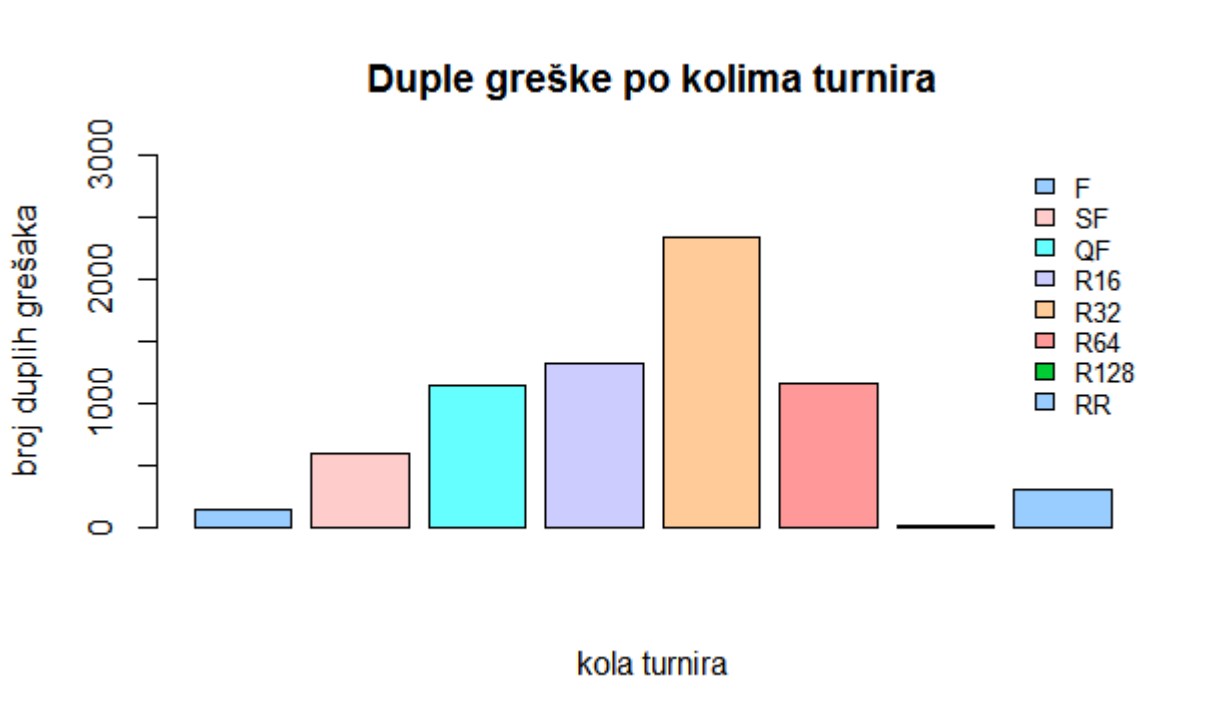
Broj duplih grešaka

```
kruskal.test (tenis$w_df ~ roundfactor_ok, data = tenis)
pairwise.wilcox.test(tenis$w_df, roundfactor_ok,
                     p.adjust.method = "BH")
```

```
Kruskal-Wallis rank sum test

data:  tenis$w_df by roundfactor_ok
Kruskal-Wallis chi-squared = 80.619, df = 7, p-value = 1.03e-14

> pairwise.wilcox.test(tenis$w_df, roundfactor_ok,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  tenis$w_df and roundfactor_ok

    R128    R64    R32    R16    QF     SF     F
R64 0.0007  -      -      -      -      -      -
R32 1.2e-09 0.0842 -      -      -      -      -
R16 6.2e-11 0.0110 0.2358 -      -      -      -
QF  6.2e-11 0.0011 0.0258 0.2358 -      -      -
SF  2.1e-06 0.0495 0.3140 0.7466 0.5585 -      -
F   6.2e-07 0.0024 0.0132 0.0495 0.2358 0.1529 -
RR  0.8309  0.6980 0.4748 0.3452 0.2878 0.3452 0.1898

P value adjustment method: BH
```

Grafički prikaz ukupnog broja duplih grešaka po kolima

```
duple<-aggregate(tenis3$w_df, by=list(Category=tenis3$round), FUN=sum)
barplot(height = duple$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Duple greške po kolima turnira",
        xlab = "kola turnira",ylab = "broj duplih grešaka",ylim = c(0,3000))
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



Broj poena osvojenih na prvi servis

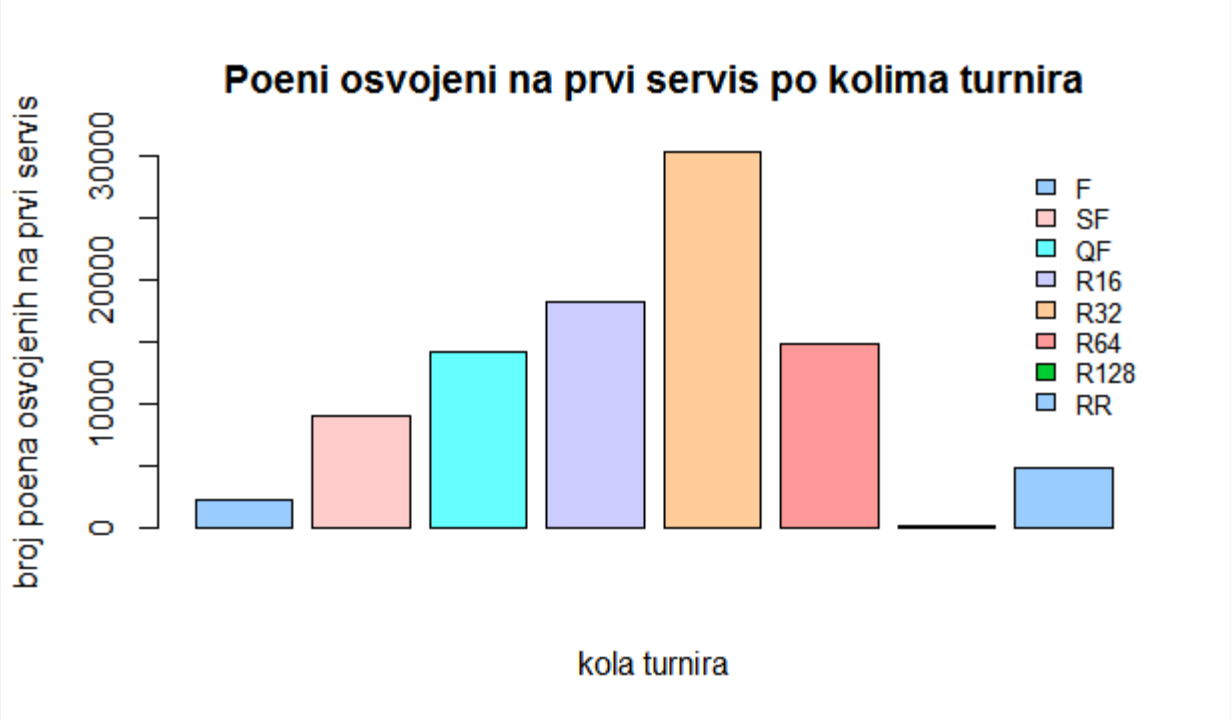
```
kruskal.test (tenis$w_1stWon ~ roundfactor_ok, data = tenis)
pairwise.wilcox.test(tenis$w_1stWon, roundfactor_ok,
                     p.adjust.method = "BH")
```

```
pairwise.wilcox.test(tenis$w_1stWon, roundfactor_ok,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  tenis$w_1stWon and roundfactor_ok
```

```
    R128    R64     R32    R16    QF     SF     F
R64 6.6e-12 -       -      -      -      -      -
R32 < 2e-16 7.9e-05 -      -      -      -      -
R16 < 2e-16 0.0034  0.6678 -      -      -      -
QF  < 2e-16 0.0450  0.6028 0.9217 -      -      -
SF  1.0e-10 0.5047  0.2754 0.5661 0.7667 -      -
F   3.4e-09 0.2818  0.8330 0.9217 0.9882 0.8538 -
RR  0.2754  0.8108  0.9217 0.9217 0.9217 0.9217 0.9217

P value adjustment method: BH
```

Grafički prikaz ukupnog broja osvojenih poena na prvi servis

```
osvojeniprvi<-aggregate(tenis3$w_1stWon, by=list(Category=tenis3$round), FUN=sum)
barplot(height = osvojeniprvi$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Poeni osvojeni na prvi servis po kolima
turnira",
        xlab = "kola turnira",ylab = "broj poena osvojenih na prvi servis",ylim =
c(0,30000))
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



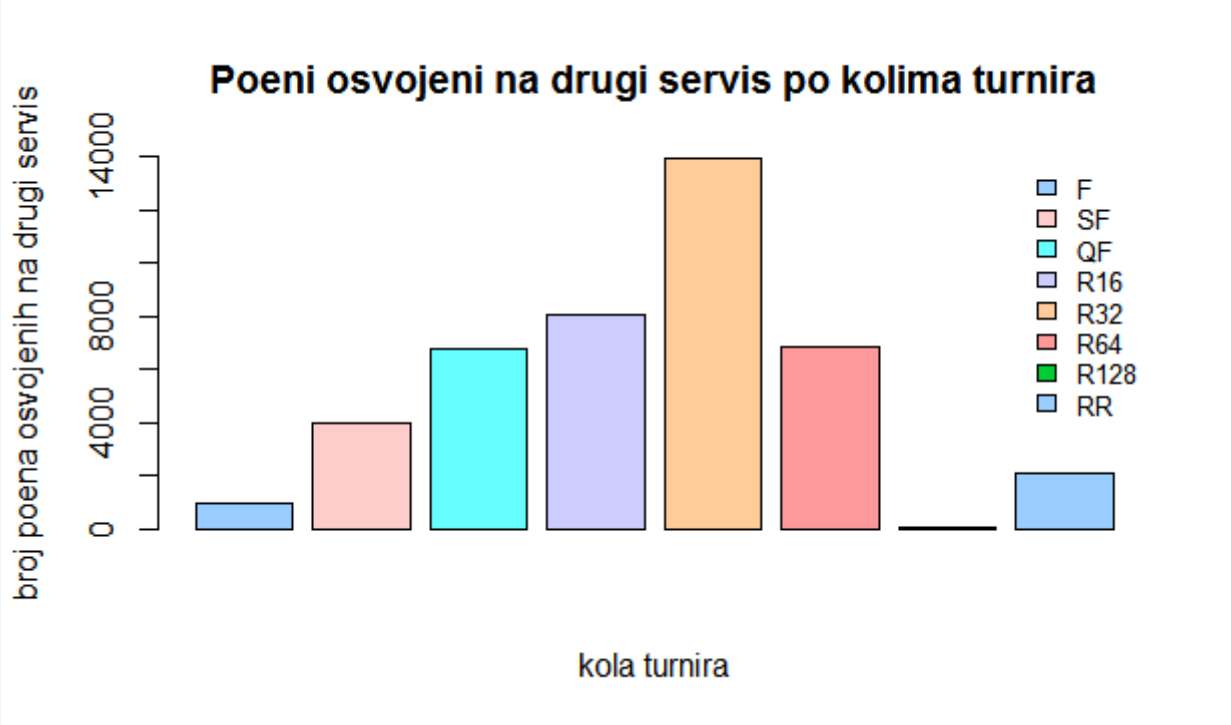
Broj poena osvojenih na drugi servis

```
kruskal.test (tenis$w_2ndWon ~ roundfactor_ok, data = tenis)
pairwise.wilcox.test(tenis$w_2ndWon, roundfactor_ok,
                     p.adjust.method = "BH")
```

```
kruskal.test (tenis$w_2ndWon ~ roundfactor_ok, data = tenis)

        Kruskal-Wallis rank sum test

data:  tenis$w_2ndWon by roundfactor_ok
Kruskal-Wallis chi-squared = 218.16, df = 7, p-value < 2.2e-16

> pairwise.wilcox.test(tenis$w_2ndWon, roundfactor_ok,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  tenis$w_2ndWon and roundfactor_ok

    R128    R64     R32     R16     QF      SF      F
R64 4.7e-14 -       -       -       -       -       -
R32 < 2e-16 4.2e-05 -       -       -       -       -
R16 < 2e-16 5.6e-06 0.30976 -       -       -       -
QF  < 2e-16 8.1e-05 0.30976 0.85379 -       -       -
SF  1.8e-15 0.01421 0.89785 0.66609 0.57139 -       -
F   7.0e-13 0.00069 0.11794 0.23017 0.30976 0.16074 -
RR  0.01421 0.07232 0.13943 0.14605 0.15643 0.12365 0.15947

P value adjustment method: BH
```

Grafički prikaz ukupnog broja osvojenih poena na prvi servis

```
osvojenidrugi<-aggregate(tenis3$w_2ndWon, by=list(Category=tenis3$round), FUN=sum)
barplot(height = osvojenidrugi$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Poeni osvojeni na drugi servis po
kolima turnira",
        xlab = "kola turnira",ylab = "broj poena osvojenih na drugi servis",ylim =
c(0,14000))
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



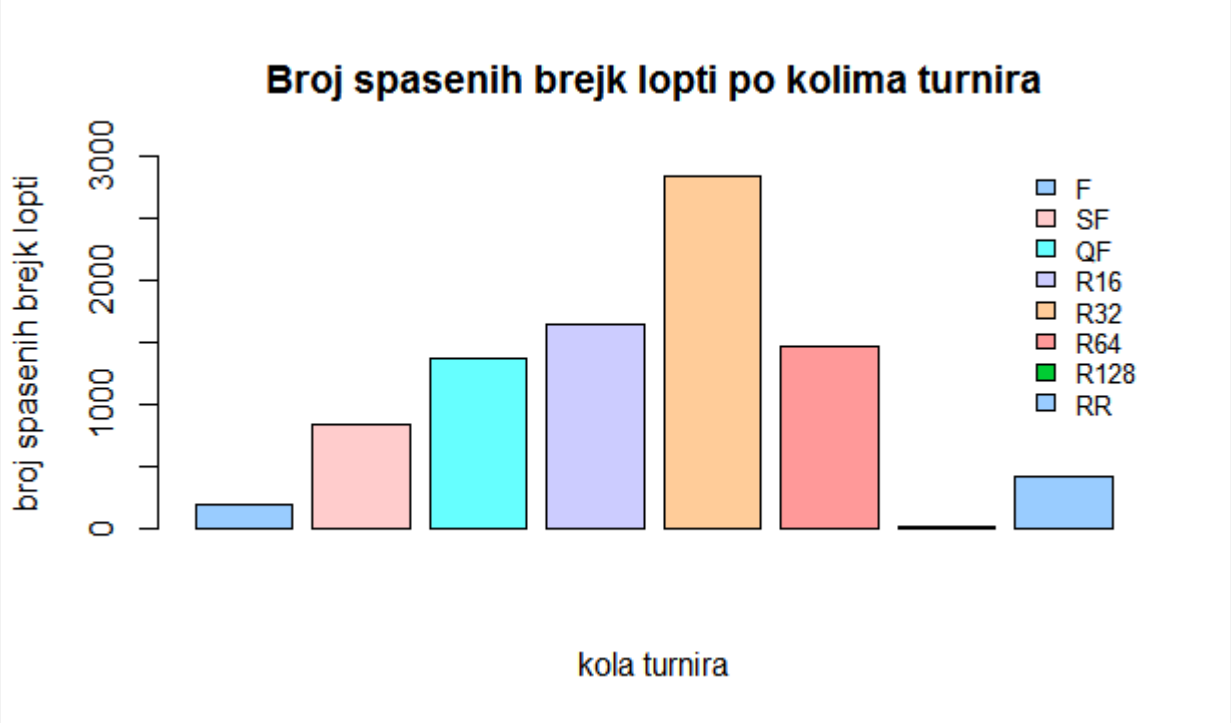
Broj spasenih brejk lopti

```
kruskal.test (tenis$l_bpSaved ~ roundfactor_ok, data = tenis)
pairwise.wilcox.test(tenis$l_bpSaved, roundfactor_ok,
                     p.adjust.method = "BH")
```

```
kruskal.test (tenis$l_bpSaved ~ roundfactor_ok, data = tenis)

        Kruskal-Wallis rank sum test

data:  tenis$l_bpSaved by roundfactor_ok
Kruskal-Wallis chi-squared = 84.26, df = 7, p-value = 1.858e-15

> pairwise.wilcox.test(tenis$l_bpSaved, roundfactor_ok,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  tenis$l_bpSaved and roundfactor_ok

    R128    R64     R32     R16     QF      SF      F
R64 0.00010 -       -       -       -       -       -
R32 9.5e-12 0.06518 -       -       -       -       -
R16 3.6e-15 0.00014 0.01795 -       -       -       -
QF  2.3e-09 0.01822 0.39564 0.59469 -       -       -
SF  0.00020 0.55946 0.85511 0.17817 0.55946 -       -
F   0.00014 0.15497 0.55946 0.85511 0.88149 0.55946 -
RR  0.59469 0.88149 0.73901 0.59469 0.64331 0.85511 0.59469

P value adjustment method: BH
```

Grafički prikaz ukupnog broja spasenih brejk lopti po kolima

```
brejkspasene<-aggregate(tenis3$w_bpSaved, by=list(Category=tenis3$round), FUN=sum)
barplot(height = brejkspasene$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Broj spasenih brejk lopti po kolima
turnira",
        xlab = "kola turnira",ylab = "broj spasenih brejk lopti",ylim = c(0,3000))
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



Broj brejk lopti sa kojim se pobednik suočio

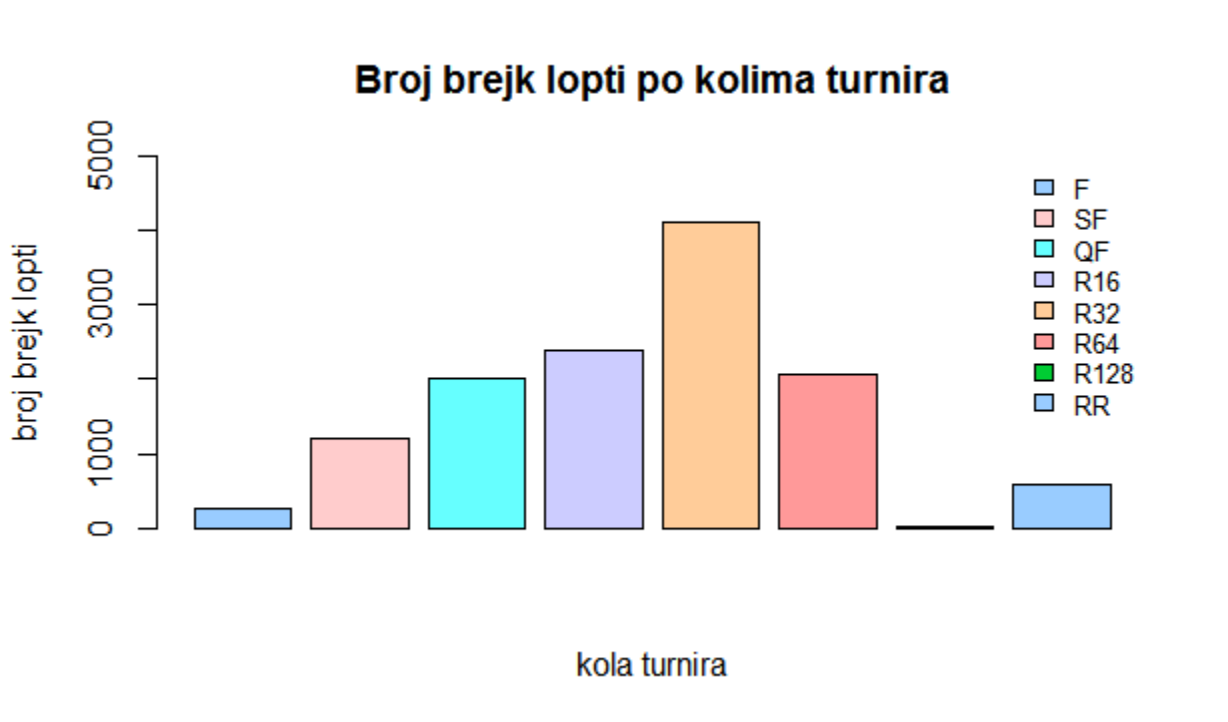
```
kruskal.test (tenis$w_bpFaced ~ roundfactor_ok, data = tenis)
pairwise.wilcox.test(tenis$w_bpFaced, roundfactor_ok,
                     p.adjust.method = "BH")
```

```
kruskal.test (tenis$w_bpFaced ~ roundfactor_ok, data = tenis)

        Kruskal-Wallis rank sum test

data:  tenis$w_bpFaced by roundfactor_ok
Kruskal-Wallis chi-squared = 58.894, df = 7, p-value = 2.509e-10
```

```
> pairwise.wilcox.test(tenis$w_bpFaced, roundfactor_ok,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  tenis$w_bpFaced and roundfactor_ok

    R128    R64    R32    R16    QF     SF     F
R64 0.0018  -      -      -      -      -      -
R32 4.4e-10 0.0363 -      -      -      -      -
R16 3.3e-09 0.0398 0.9894 -      -      -      -
QF  5.1e-06 0.2185 0.9894 0.9894 -      -      -
SF  7.9e-05 0.2526 0.9894 0.9894 0.9894 -      -
F   7.9e-05 0.0530 0.3831 0.3847 0.3838 0.4453 -
RR  0.7876  0.9894 0.9894 0.9894 0.9894 0.9894 0.9894

P value adjustment method: BH
```

Grafički prikaz ukupnog broja brejk lopti po kolima

```
brejksuocene<-aggregate(tenis3$w_bpFaced, by=list(Category=tenis3$round), FUN=sum)
barplot(height = brejksuocene$x,col = moje_boje,
        beside = TRUE, axis.lty="solid",main="Broj brejk lopti po kolima turnira",
        xlab = "kola turnira",ylab = "broj brejk lopti",ylim = c(0,5000))
legend("topright", legend = c('F', 'SF', 'QF','R16','R32','R64','R128','RR'),
       fill = moje_boje, box.lty = 0, cex = 0.8)
```



Kada pregledamo sve prethodne rezultate možemo uočiti da su svi gotovo identični i potvrđuju i zaključak donesen nakon analize rezultata prve varijable (*broj ubačenih prvih servisa*), kao i pretpostavku iznesenu pre analize. Statistički značajna razlika postoji između prvog kola gren slema i svih ostalih rundi turnira posmatrano po svim varijablama. Trend koji se uočava jeste da se ta razlika kod svih varijabli javlja i između šesnaestine finala (R32) i drugih kola (posebno drugog kola i osmine finala), što verovatno ukazuje na veliku ,,selekciju" koja se događa u ovom delu turnira (zanimljivo je i da ovo kolo po graficima ubedljivo prednjači u ukupnom broju svih testiranih varijabli). Interesantno je i da se jedino kod broj duplih grešaka javlja statistički značajna razlika finala sa većinom drugih kola, što se opet može objasniti pritiskom koji igrači osećaju zbog težine meča, kao i većim rizikom na koji moraju da igraju u mečevima takvog značaja.

Treće istraživačko pitanje

3. Da li se na osnovu broja asova, duplih grešaka, ubačenih prvih servisa, osvojenih poena na prvi odnosno drugi servis i broja spasenih brejk lopti može predvideti rang tog igrača?

Kako su histogrami pokazali neznatne razlike u raspodeli varijabli kod pobednika i poraženog, potpuno je sigurno zaključiti da će rezultati biti gotovo pa identični ukoliko budemo analizu primenjivali posebno na podatke u vezi sa pobednikom meča, posebno na one u vezi sa poraženim, stoga je odlučeno da se u test višestruke regresije, koji ćemo koristiti za ovu analizu, ubace varijable koje se odnose na pobednika. Test regresije se nadovezuje na test korelacije, a pravljenje korelacijske matrice je ujedno i prvi korak u sprovođenju regresije. U nastavku sledi kod kojim je dobijena korelacijska matrica i odgovarajuće *p vrednosti.*

matrica

```
regresija_winner_pomocna <- data.frame(tenis$winner_rank, tenis$w_ace, tenis$w_df,
                                       tenis$w_1stIn, tenis$w_1stWon,
                                       tenis$w_2ndWon,tenis$w_bpSaved)
regresija_winner<- cor(as.matrix(regresija_winner_pomocna), method = "spearman")
regresija_winner
```

```
regresija_winner
                 tenis.winner_rank tenis.w_ace tenis.w_df tenis.w_1stIn
tenis.winner_rank        1.00000000 -0.07533664 0.07633786    0.05201448
tenis.w_ace             -0.07533664  1.00000000 0.19652054    0.27129830
tenis.w_df               0.07633786  0.19652054 1.00000000    0.33444070
tenis.w_1stIn            0.05201448  0.27129830 0.33444070    1.00000000
tenis.w_1stWon           0.01183864  0.40427946 0.33679564    0.95529306
tenis.w_2ndWon           0.04619329  0.25191373 0.40081106    0.53161828
tenis.w_bpSaved          0.09781962  0.03272512 0.33345077    0.57209463
```

```
                 tenis.w_1stWon tenis.w_2ndWon tenis.w_bpSaved
tenis.winner_rank     0.01183864     0.04619329      0.09781962
tenis.w_ace           0.40427946     0.25191373      0.03272512
tenis.w_df            0.33679564     0.40081106      0.33345077
tenis.w_1stIn         0.95529306     0.53161828      0.57209463
tenis.w_1stWon        1.00000000     0.49814760      0.47122944
tenis.w_2ndWon        0.49814760     1.00000000      0.42425663
tenis.w_bpSaved       0.47122944     0.42425663      1.00000000
```

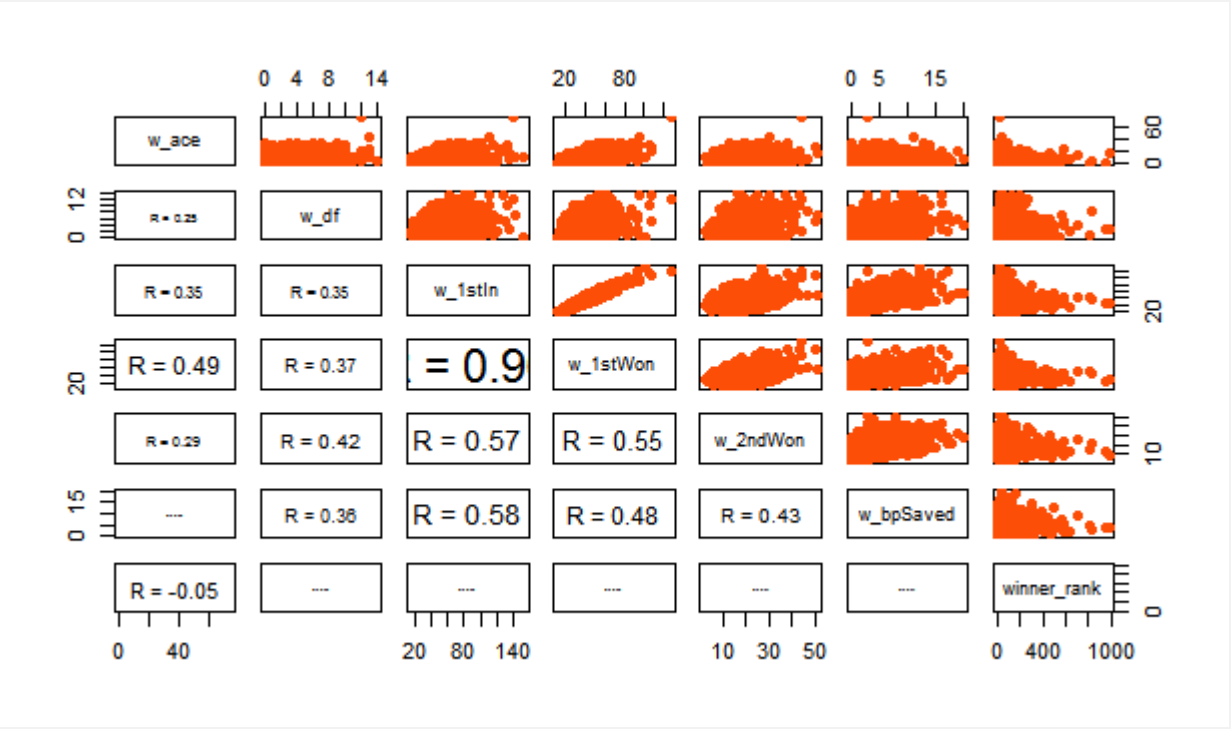
 # p vrednosti

```
svaka_sa_svakom = expand.grid(names(regresija_winner_pomocna),
                              names(regresija_winner_pomocna))
funkcija_za_p = function(col_name1, col_name2, data_frame) {
  cor.test(data_frame[[col_name1]], data_frame[[col_name2]])$p.value
}
pvrednosti <- mapply(funkcija_za_p,
                     col_name1 = svaka_sa_svakom[[1]],
                     col_name2 = svaka_sa_svakom[[2]],
                     MoreArgs = list(data_frame = regresija_winner_pomocna))
matrix(pvrednosti, 7, 7, dimnames = list(names(regresija_winner_pomocna),
                                         names(regresija_winner_pomocna)))
```

```
                  tenis.winner_rank   tenis.w_ace    tenis.w_df tenis.w_1stIn
tenis.winner_rank      0.0000000000  9.121360e-03  2.020282e-03  1.463016e-02
tenis.w_ace            0.0091213603  0.000000e+00  1.226156e-35  1.373679e-71
tenis.w_df             0.0020202822  1.226156e-35  0.000000e+00  5.440326e-75
tenis.w_1stIn          0.0146301640  1.373679e-71  5.440326e-75  0.000000e+00
tenis.w_1stWon         0.2086826285 7.018105e-150  1.319222e-79  0.000000e+00
tenis.w_2ndWon         0.5669629141  5.180971e-48 2.127475e-109 9.707888e-212
tenis.w_bpSaved        0.0002472916  1.002280e-02  3.457605e-75 1.151119e-226
                  tenis.w_1stWon tenis.w_2ndWon tenis.w_bpSaved
tenis.winner_rank   2.086826e-01   5.669629e-01    2.472916e-04
tenis.w_ace        7.018105e-150   5.180971e-48    1.002280e-02
tenis.w_df          1.319222e-79  2.127475e-109    3.457605e-75
tenis.w_1stIn       0.000000e+00  9.707888e-212   1.151119e-226
tenis.w_1stWon      0.000000e+00  8.819332e-196   3.732091e-145
tenis.w_2ndWon     8.819332e-196   0.000000e+00   4.405234e-113
tenis.w_bpSaved    3.732091e-145  4.405234e-113    0.000000e+00
```

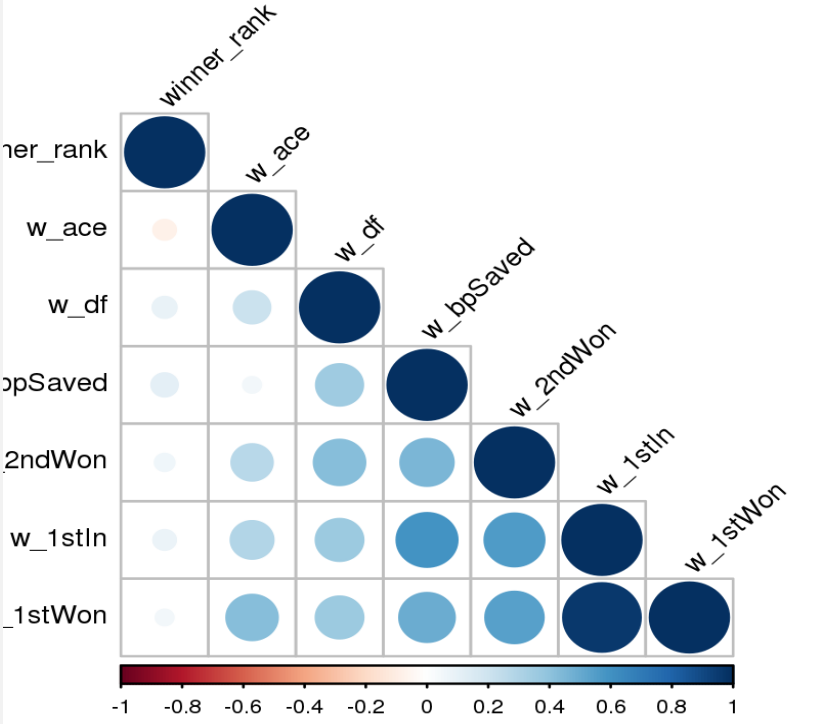
Vizuelizacija korelacijske matrice: Napravljene su dve funkcije, jedna za gornji i jedna za donji deo korelacijske matrice. Funkcija će u donjem delu matrice beležiti rezultat korelacije, zaokružene na dve decimale, što je veća korelacija veća su i slova (ostali argumenti su prepisani iz R dokumentacije, kojom sam se i vodila pri sastavljanju koda, i neophodni su da bi se željeni efekat postigao, dok je veličina teksta dobijena isprobavanjem). Na kraju funkcija *pairs()* pravi konačni grafički prikaz matrice.

```
korelacija_dole <- function(x, y){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y), digits=2)
  txt <- paste0("R = ", r)
  cex.cor <- 1.5/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

korelacija_gore<-function(x, y){
  points(x,y, pch = 19, col = "#FC4E07")
}

pairs(tenis[,c(6,7,9,10,11,13,24)],
      lower.panel = korelacija_dole,
      upper.panel = korelacija_gore)
```



U nemogućnosti pristupanja paketima namenjenim za grafičko prikazivanje u R-u, online corrplot alatka za pravljenje vizuelnih prikaza poslužila je i da napravim još jedan način vizuelizacije (iako je u pitanju najjednostavniji kod, bez ikakvih dodatnih argumenata, što je i za očekivati za online verziju ovog paketa).

```
corrplot(regresija_winner, method="circle", type="lower")
```



Grafički rezultati vizuelizovali su pomalo neočekivane rezultate analize (date pre vizuelizacije). Iako lepi za oko, jasno stavljaju do znanja da varijable iskorišćene u analizi nisu relevantne, ili bolje rečeno, dovoljne kako bismo mogli na osnovu njih predvideti rang pobednika, a u opštem slučaju i igrača. Korelacija sa brojem osvojenih poena na prvi i drugi servis nije značajna, dok sa drugim varijablama iako *p vrednosti* ukazuju na statističku značajnost, veza je izuzetno slaba, gotovo zanemarljiva (reda veličine 0.09, 0.07, 0.05 i slično). Od ovih je „najveća“ sa brojem spasenih brejk lopti, što jeste odlika velikih igrača da u najbitnijim trenucima odigraju najbolje, a interesantno je da, koliko god da je mala, korelacija sa brojem asova je negativna. Ukoliko se fanovi tenisa zamisle, i videće da broj asova nije najvatrenije, ili bolje rečeno jedino, oružje najboljih tenisera. Karlović, Raonić, Izner, Anderson i teniseri koji svoju igru zasnivaju na servisu, iako izuzetno dobro plasirani, ipak zauzimaju nešto udaljenije pozicije od teniskog trona. Za trenutak sam pomislila da uzorak možda nije dovoljno veliki za istraživanje ovog tipa, pa sam jednostavnim kopiranjem koda, postupak ponovila nad bazom koja čuva podatke o mečevima od 1968. do 2018. godine, ali rezultat je gotovo identičan, razlikuje se svega u neku decimalu. Pri ovakvim rezultatima, nema sumnje da je regresioni model koji bi objasnio najveći procenat varijabiliteta ranga igrača bio onaj u koji bismo uključili sve varijable koje imaju statistički značajnu vezu sa rangom (što bi značilo isključiti iz modela broj osvojenih poena na prvi i drugi servis):

```
model_pobednik <- lm(tenis$winner_rank ~ tenis$w_ace + tenis$w_df + tenis$w_1stIn +
                       tenis$w_bpSaved)
summary(model_pobednik)
```

```
summary(model_pobednik)

Call:
lm(formula = tenis$winner_rank ~ tenis$w_ace + tenis$w_df + tenis$w_1stIn +
    tenis$w_bpSaved)

Residuals:
   Min     1Q Median     3Q    Max
-78.12 -41.66 -18.65  13.84 942.39

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      52.3034     4.5462  11.505  < 2e-16 ***
tenis$w_ace      -1.0787     0.2971  -3.630 0.000289 ***
tenis$w_df        1.8821     0.7464   2.521 0.011749 *
tenis$w_1stIn     0.1487     0.1129   1.317 0.188016
tenis$w_bpSaved   0.9175     0.6357   1.443 0.149069
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.62 on 2490 degrees of freedom
Multiple R-squared:  0.01206,  Adjusted R-squared:  0.01048
F-statistic: 7.602 on 4 and 2490 DF,  p-value: 4.384e-06
```

Iako je model statistički značajan, objašnjava svega 1.2% varijabiliteta, što je apsolutno zanemarljivo u kontekstu nekog ozbiljnijeg proučavanja. Možemo videti i da broj ubačenih prvih servisa najmanje doprinosi rezultatima, stoga možemo isprobati model bez njega:

```
summary(model_pobednik)
model_pobednik_asduplebrejk<-lm(tenis$winner_rank ~ tenis$w_ace + tenis$w_df +
                                  tenis$w_bpSaved)
summary(model_pobednik_asduplebrejk)
```

```
summary(model_pobednik_asduplebrejk)

Call:
lm(formula = tenis$winner_rank ~ tenis$w_ace + tenis$w_df + tenis$w_bpSaved)

Residuals:
   Min     1Q Median     3Q    Max
-79.39 -41.57 -18.81  14.70 939.84
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     56.6435     3.1319  18.086  < 2e-16 ***
tenis$w_ace     -0.9375     0.2772  -3.383 0.000729 ***
tenis$w_df       1.9917     0.7419   2.685 0.007309 **
tenis$w_bpSaved  1.3842     0.5279   2.622 0.008793 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.63 on 2491 degrees of freedom
Multiple R-squared:  0.01138,  Adjusted R-squared:  0.01019
F-statistic: 9.555 on 3 and 2491 DF,  p-value: 2.844e-06
```

Kako je i ovaj model značajan, a objašnjava skoro isto (1.1%) varijabiliteta kao i prvi model, možda se čak može i uzeti kao bolji, jer sa varijablom manje, dolazi do gotovo iste uspešnosti (napomena: dodatno izbacivanje varijabli iz modela drastično smanjuje objašnjenost varijabiliteta stoga to nije dalje činjeno).
U nastavku je spojeni prikaz sve tri varijable sa rangom koji predstavlja svojevrsnu vizuelizaciju ovog „najidelanijeg“ regresionog modela, a i po njemu se može videti potpuna nelinearnost zvezdica na slici čime se ukazuje na nepostojanje ikakve značajne veze.

```
plot(tenis$winner_rank,tenis$w_ace+ tenis$w_df+tenis$w_bpSaved,
     main="Regresioni model",
     xlab="rang ", ylab="asovi, duple greške i spasene brejk lopte ", type="p",pch=8,
     col='#660066',bg='#993366')
```

Pretpostavka jeste bila da se slabo može predvideti rang igrača na osnovu učinka u konkretnom meču, ali možda jeste ipak malo iznenađujaća ovako mala veza tog učinka i statističkih brojeva sa samim rangom. Ono što sam sigurna da bi bilo interesantno jeste da se statistika o riternu (procentu vraćenih servisa, osvojenih poena na protivnikov prvi/drugi servis) uključi u regresionu analizu. Verujem da bi procenat objašnjenog varijabiliteta bio nešto veći. S druge strane, ovakav rezultat ne bi trebalo da čudi, jer igra u tenisu i ishod meča zavisi od toliko mnogo, brojevima statistike nemerljivih, komponenti da je jednostavno nemoguće izvoditi takve prenagljene zaključke iz prostih brojeva. Psiha, forma, umor, utreniranost, vremenski uslovi, ili prosto kakav vam je dan, ne kakav rang na ATP listi, samo su neki od faktora koji utiču na igru i čine je takvom da ona zavisi od datog trenutka. Ali uostalom to je čar igre, čar sporta.

Kraj analize

Analizu sprovela: Olivera Golijanin
GitHub profil: https://github.com/OliveraGolijanin