Analiza UFC borbi

Martin Bakač Adrian Golem Martin Josip Kocijan Jure Rajčić

Sažetak

Izvješće sadržava jasno objašnjene koncepte statističke analize podataka u programskom jeziku R primijenjene na postojeće skupove podataka zadane u .csv formatu. Prikupljeni podatci predstavljaju informacije o UFC borbama (i borcima koji su u njima sudjelovali) u razdoblju od 1993. - 2021. godine. Borbama su pridodane značajke poput trajanja u rundama, sudca borbe te datuma i lokacije održavanja, dok su borcima pridružene značajke kao što su visina, težina, dužina ruke ili stav borca. Dodatno je poznat i pobjednik svake borbe (i način pobjede). Izvještaj se sastoji od Deskriptivne analize i Analize podataka. Deskriptivna analiza dio je izvještaja koji uključuje korištenje različitih statističkih metoda kako bi se prikazala ključna obilježja podataka, s ciljem dobivanja bolje predodžbe o podacima i razumijevanja obilježja podataka prije nego li se koristi bilo kojom drugom metodom statističke analize. Analiza podataka dio je izvještaja kojoj je cilj utvrditi postoji li veza između dužine ruke boraca i vjerojatnosti završetka borbe nokautom, te da li postoji razlika u trajanju mečeva između pojedinih kategorija. Analiza također istražuje je li trajanje borbi za titulu duže od ostalih borbi u natjecanju te mogu li dostupne značajke predvidjeti pobjednika. Proučavaju se i vjerojatnosti pobjede crvenih boraca u mečevima. Rezultati analize pružaju zanimljive spoznaje koje nam mogu biti korisne u budućim istraživanjima.

Sadržaj

Deskriptivna analiza

Podaci o borbama

Učitavanje podataka o svim borbama

```
file <- "./total_fight_data.csv"
fight_data <- read.csv(file, sep = ";")
```

Prikaz dimenzija tablice

```
dim(fight_data)
```

```
## [1] 6012   41
```

Prikaz imena stupaca

```
names(fight_data)
```

```
##  [1] "R_fighter"       "B_fighter"       "R_KD"            "B_KD"
##  [5] "R_SIG_STR."      "B_SIG_STR."      "R_SIG_STR_pct"   "B_SIG_STR_pct"
##  [9] "R_TOTAL_STR."    "B_TOTAL_STR."    "R_TD"            "B_TD"
## [13] "R_TD_pct"        "B_TD_pct"        "R_SUB_ATT"       "B_SUB_ATT"
## [17] "R_REV"           "B_REV"           "R_CTRL"          "B_CTRL"
## [21] "R_HEAD"          "B_HEAD"          "R_BODY"          "B_BODY"
## [25] "R_LEG"           "B_LEG"           "R_DISTANCE"      "B_DISTANCE"
## [29] "R_CLINCH"        "B_CLINCH"        "R_GROUND"        "B_GROUND"
## [33] "win_by"          "last_round"      "last_round_time" "Format"
## [37] "Referee"         "date"            "location"        "Fight_type"
## [41] "Winner"
```

Prikaz prvih 6 redaka i prvih 6 stupaca tablice

```
head(fight_data[1:6])
```

```
##         R_fighter         B_fighter R_KD B_KD R_SIG_STR. B_SIG_STR.
## 1    Adrian Yanez     Gustavo Lopez    2    0  41 of 103   23 of 51
## 2    Trevin Giles     Roman Dolidze    0    0   27 of 57   32 of 67
## 3     Tai Tuivasa   Harry Hunsucker    1    0   14 of 18     2 of 6
## 4   Cheyanne Buys Montserrat Conejo    0    0   31 of 65   15 of 41
## 5   Marion Reneau     Macy Chiasson    0    0   30 of 63  51 of 138
## 6 Leonardo Santos      Grant Dawson    0    0   30 of 67   46 of 84
```

Prikaz zadnjih 6 redaka i zadnjih 6 stupaca tablice

```
tail(fight_data[35:40])
```

```
##      last_round_time        Format       Referee           date
## 6007            6:41 No Time Limit John McCarthy March 11, 1994
## 6008            9:51 No Time Limit John McCarthy March 11, 1994
## 6009            2:50 No Time Limit John McCarthy March 11, 1994
## 6010           12:13 No Time Limit John McCarthy March 11, 1994
## 6011            0:58 No Time Limit John McCarthy March 11, 1994
## 6012            0:20 No Time Limit John McCarthy March 11, 1994
##                   location       Fight_type
## 6007 Denver, Colorado, USA Open Weight Bout
## 6008 Denver, Colorado, USA Open Weight Bout
## 6009 Denver, Colorado, USA Open Weight Bout
## 6010 Denver, Colorado, USA Open Weight Bout
## 6011 Denver, Colorado, USA Open Weight Bout
```

```
## 6012 Denver, Colorado, USA Open Weight Bout
```

Prikaz sažetka svih stupaca

```
summary(fight_data)
```

```
##   R_fighter          B_fighter             R_KD             B_KD
##  Length:6012        Length:6012        Min.   :0.0000   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.0000
##                                        Mean   :0.2498   Mean   :0.1798
##                                        3rd Qu.:0.0000   3rd Qu.:0.0000
##                                        Max.   :5.0000   Max.   :4.0000
##   R_SIG_STR.         B_SIG_STR.        R_SIG_STR_pct      B_SIG_STR_pct
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  R_TOTAL_STR.       B_TOTAL_STR.           R_TD               B_TD
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    R_TD_pct           B_TD_pct           R_SUB_ATT         B_SUB_ATT
##  Length:6012        Length:6012        Min.   : 0.0000   Min.   :0.000
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.:0.000
##  Mode  :character   Mode  :character   Median : 0.0000   Median :0.000
##                                        Mean   : 0.4814   Mean   :0.344
##                                        3rd Qu.: 1.0000   3rd Qu.:0.000
##                                        Max.   :10.0000   Max.   :7.000
##      R_REV             B_REV           R_CTRL             B_CTRL
##  Min.   :0.0000   Min.   :0.0000   Length:6012        Length:6012
##  1st Qu.:0.0000   1st Qu.:0.0000   Class :character   Class :character
##  Median :0.0000   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :0.1377   Mean   :0.1354
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :5.0000   Max.   :3.0000
##     R_HEAD             B_HEAD             R_BODY             B_BODY
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     R_LEG              B_LEG            R_DISTANCE         B_DISTANCE
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##    R_CLINCH           B_CLINCH           R_GROUND           B_GROUND
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     win_by            last_round    last_round_time       Format
##  Length:6012        Min.   :1.000   Length:6012        Length:6012
##  Class :character   1st Qu.:1.000   Class :character   Class :character
##  Mode  :character   Median :3.000   Mode  :character   Mode  :character
##                     Mean   :2.317
##                     3rd Qu.:3.000
##                     Max.   :5.000
##    Referee              date             location          Fight_type
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Winner
##  Length:6012
##  Class :character
##  Mode  :character
##
##
##
```

Prikaz numeričkih stupaca

```
cat(colnames(fight_data %>%
    select(where(is.numeric))), fill = TRUE)
```

```
## R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
```

Prikaz stupaca koji sadrže znakove

```
cat(colnames(fight_data %>%
    select(where(is.character))), fill = TRUE)
```

```
## R_fighter B_fighter R_SIG_STR. B_SIG_STR. R_SIG_STR_pct B_SIG_STR_pct
## R_TOTAL_STR. B_TOTAL_STR. R_TD B_TD R_TD_pct B_TD_pct R_CTRL B_CTRL R_HEAD
## B_HEAD R_BODY B_BODY R_LEG B_LEG R_DISTANCE B_DISTANCE R_CLINCH B_CLINCH
## R_GROUND B_GROUND win_by last_round_time Format Referee date location
## Fight_type Winner
```

Prikaz broja jedinstvenih vrijednosti u numeričkim stupcima

```
unique_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(n_distinct))
print(unique_values)
##   R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
## 1    6    5        11         8     6     4          5
```

Prikaz broja jedinstvenih vrijednosti u znakovnim stupcima

```
unique_values <- fight_data %>%
    select(where(is.character)) %>%
    summarise_all(list(n_distinct))
print(unique_values)
##   R_fighter B_fighter R_SIG_STR. B_SIG_STR. R_SIG_STR_pct B_SIG_STR_pct
## 1      1514      1987       3467       3308            97            91
##   R_TOTAL_STR. B_TOTAL_STR. R_TD B_TD R_TD_pct B_TD_pct R_CTRL B_CTRL R_HEAD
## 1         4213         3963  162  159       70       64    748    625   2994
##   B_HEAD R_BODY B_BODY R_LEG B_LEG R_DISTANCE B_DISTANCE R_CLINCH B_CLINCH
## 1   2802    593    573   424   391       2748       2753      498      490
##   R_GROUND B_GROUND win_by last_round_time Format Referee date location
## 1      686      501     10             336     19     205  550      166
##   Fight_type Winner
## 1        109   1438
```

Prikaz aritmetičke sredine u numeričkim stupcima

```
mean_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(mean))
print(mean_values)
##        R_KD      B_KD R_SUB_ATT B_SUB_ATT     R_REV     B_REV last_round
## 1 0.2498337 0.1798071 0.4813706 0.3439787 0.1377246 0.1353959     2.3167
```

Prikaz standardne devijacije u numeričkim stupcima

```
sd_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(sd))
print(sd_values)
##        R_KD      B_KD R_SUB_ATT B_SUB_ATT     R_REV     B_REV last_round
## 1 0.5234081 0.4561323  0.924078 0.7918076 0.4222958 0.4147079   1.008284
```

Prikaz najmanje i najveće vrijednosti u numeričkim stupcima

```
min_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(min))
print(min_values)
##   R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
## 1    0    0         0         0     0     0          1
max_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(max))
print(max_values)
##   R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
## 1    5    4        10         7     5     3          5
```

Prikaz prvog, drugog i trećeg kvartila u numeričkim stupcima

```
first_quartile <- function(x) quantile(x, probs = c(0.25))
first_quartile_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(first_quartile))
print(first_quartile_values)
##   R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
## 1    0    0         0         0     0     0          1
```

```r
median_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(median))
print(median_values)
##   R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
## 1    0    0         0         0     0     0          3
third_quartile <- function(x) quantile(x, probs = c(0.75))
third_quartile_values <- fight_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(third_quartile))
print(third_quartile_values)
##   R_KD B_KD R_SUB_ATT B_SUB_ATT R_REV B_REV last_round
## 1    0    0         1         0     0     0          3
```

Podaci o borcima

Učitavanje podataka o svim borcima

```
file <- "./fighter_details.csv"
fighter_data <- read.csv(file)
```

Prikaz dimenzija tablice

```
dim(fighter_data)
```

```
## [1] 3596   14
```

Prikaz imena stupaca

```
names(fighter_data)
```

```
##  [1] "fighter_name" "Height"       "Weight"       "Reach"        "Stance"
##  [6] "DOB"          "SLpM"         "Str_Acc"      "SApM"         "Str_Def"
## [11] "TD_Avg"       "TD_Acc"       "TD_Def"       "Sub_Avg"
```

Prikaz prvih 6 redaka tablice

```
head(fighter_data)
```

```
##          fighter_name Height   Weight Reach   Stance          DOB SLpM Str_Acc
## 1           Tom Aaron        155 lbs.               Jul 13, 1978 0.00      0%
## 2          Papy Abedi 5' 11" 185 lbs.       Southpaw Jun 30, 1978 2.80     55%
## 3 Shamil Abdurakhimov  6' 3" 235 lbs.   76" Orthodox Sep 02, 1981 2.45     44%
## 4        Danny Abbadi 5' 11" 155 lbs.       Orthodox Jul 03, 1983 3.29     38%
## 5        Hiroyuki Abe  5' 6" 145 lbs.       Orthodox              1.71     36%
## 6       Ricardo Abreu 5' 11" 185 lbs.       Orthodox Apr 27, 1984 3.79     31%
##   SApM Str_Def TD_Avg TD_Acc TD_Def Sub_Avg
## 1 0.00      0%   0.00     0%     0%     0.0
## 2 3.15     48%   3.47    57%    50%     1.3
## 3 2.45     58%   1.23    24%    47%     0.2
## 4 4.41     57%   0.00     0%    77%     0.0
## 5 3.11     63%   0.00     0%    33%     0.0
## 6 3.98     68%   2.13    42%   100%     0.7
```

Prikaz zadnjih 6 redaka tablice

```
tail(fighter_data)
```

```
##         fighter_name Height   Weight Reach   Stance          DOB SLpM Str_Acc
## 3591 Carlos Zevallos  6' 0" 205 lbs.       Orthodox              4.36     65%
## 3592   Zhang Tiequan  5' 8" 155 lbs.   69" Orthodox Jul 25, 1978 1.23     36%
## 3593     Alex Zuniga        145 lbs.                             0.00      0%
## 3594   George Zuniga  5' 9" 185 lbs.                             7.64     38%
## 3595    Allan Zuniga  5' 7" 155 lbs.   70" Orthodox Apr 04, 1992 3.93     52%
## 3596  Virgil Zwicker  6' 2" 205 lbs.   74"          Jun 26, 1982 3.34     48%
##      SApM Str_Def TD_Avg TD_Acc TD_Def Sub_Avg
## 3591 2.28     68%   0.00     0%   100%     0.0
## 3592 2.14     51%   1.95    58%    75%     3.4
## 3593 0.00      0%   0.00     0%     0%     0.0
## 3594 5.45     37%   0.00     0%   100%     0.0
## 3595 1.80     61%   0.00     0%    57%     1.0
## 3596 4.87     39%   1.31    30%    50%     0.0
```

Prikaz sažetka svih stupaca

```
summary(fighter_data)
```

```
##  fighter_name          Height             Weight             Reach
##  Length:3596        Length:3596        Length:3596        Length:3596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Stance              DOB                 SLpM         Str_Acc
##  Length:3596        Length:3596        Min.   : 0.00   Length:3596
##  Class :character   Class :character   1st Qu.: 0.70   Class :character
##  Mode  :character   Mode  :character   Median : 2.17   Mode  :character
##                                        Mean   : 2.28
##                                        3rd Qu.: 3.42
##                                        Max.   :19.91
##       SApM          Str_Def              TD_Avg          TD_Acc
##  Min.   : 0.000   Length:3596        Min.   : 0.000   Length:3596
##  1st Qu.: 1.400   Class :character   1st Qu.: 0.000   Class :character
##  Median : 2.760   Mode  :character   Median : 0.510   Mode  :character
##  Mean   : 2.983                      Mean   : 1.211
##  3rd Qu.: 4.003                      3rd Qu.: 1.885
##  Max.   :52.500                      Max.   :32.140
##     TD_Def             Sub_Avg
##  Length:3596        Min.   : 0.0000
##  Class :character   1st Qu.: 0.0000
##  Mode  :character   Median : 0.0000
##                     Mean   : 0.6367
##                     3rd Qu.: 0.8000
##                     Max.   :21.9000
```

Prikaz numeričkih stupaca

```
cat(colnames(fighter_data %>%
    select(where(is.numeric))), fill = TRUE)
```

```
## SLpM SApM TD_Avg Sub_Avg
```

Prikaz stupaca koji sadrže znakove

```
cat(colnames(fighter_data %>%
    select(where(is.character))), fill = TRUE)
```

```
## fighter_name Height Weight Reach Stance DOB Str_Acc Str_Def TD_Acc TD_Def
```

Prikaz broja jedinstvenih vrijednosti u numeričkim stupcima

```
unique_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(n_distinct))
print(unique_values)
##   SLpM SApM TD_Avg Sub_Avg
## 1  650  759    527      95
```

Prikaz broja jedinstvenih vrijednosti u znakovnim stupcima

```
unique_values <- fighter_data %>%
    select(where(is.character)) %>%
```

```
    summarise_all(list(n_distinct))
print(unique_values)
##   fighter_name Height Weight Reach Stance  DOB Str_Acc Str_Def TD_Acc TD_Def
## 1         3596     27    113    28      6 2439      85      85     83     93
```

Prikaz aritmetičke sredine u numeričkim stupcima

```
mean_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(mean))
print(mean_values)
##       SLpM     SApM   TD_Avg   Sub_Avg
## 1 2.279633 2.982948 1.211243 0.6367075
```

Prikaz standardne devijacije u numeričkim stupcima

```
sd_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(sd))
print(sd_values)
##       SLpM     SApM  TD_Avg  Sub_Avg
## 1 1.901956 2.814008 1.91402 1.566843
```

Prikaz najmanje i najveće vrijednosti u numeričkim stupcima

```
min_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(min))
print(min_values)
##   SLpM SApM TD_Avg Sub_Avg
## 1    0    0      0       0
max_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(max))
print(max_values)
##    SLpM SApM TD_Avg Sub_Avg
## 1 19.91 52.5  32.14    21.9
```

Prikaz prvog, drugog i trećeg kvartila u numeričkim stupcima

```
first_quartile <- function(x) quantile(x, probs = c(0.25))
first_quartile_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(first_quartile))
print(first_quartile_values)
##   SLpM SApM TD_Avg Sub_Avg
## 1  0.7  1.4      0       0
median_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(median))
print(median_values)
##   SLpM SApM TD_Avg Sub_Avg
## 1 2.17 2.76   0.51       0
third_quartile <- function(x) quantile(x, probs = c(0.75))
third_quartile_values <- fighter_data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(third_quartile))
```

```
print(third_quartile_values)
##   SLpM   SApM TD_Avg Sub_Avg
## 1 3.42 4.0025  1.885     0.8
```

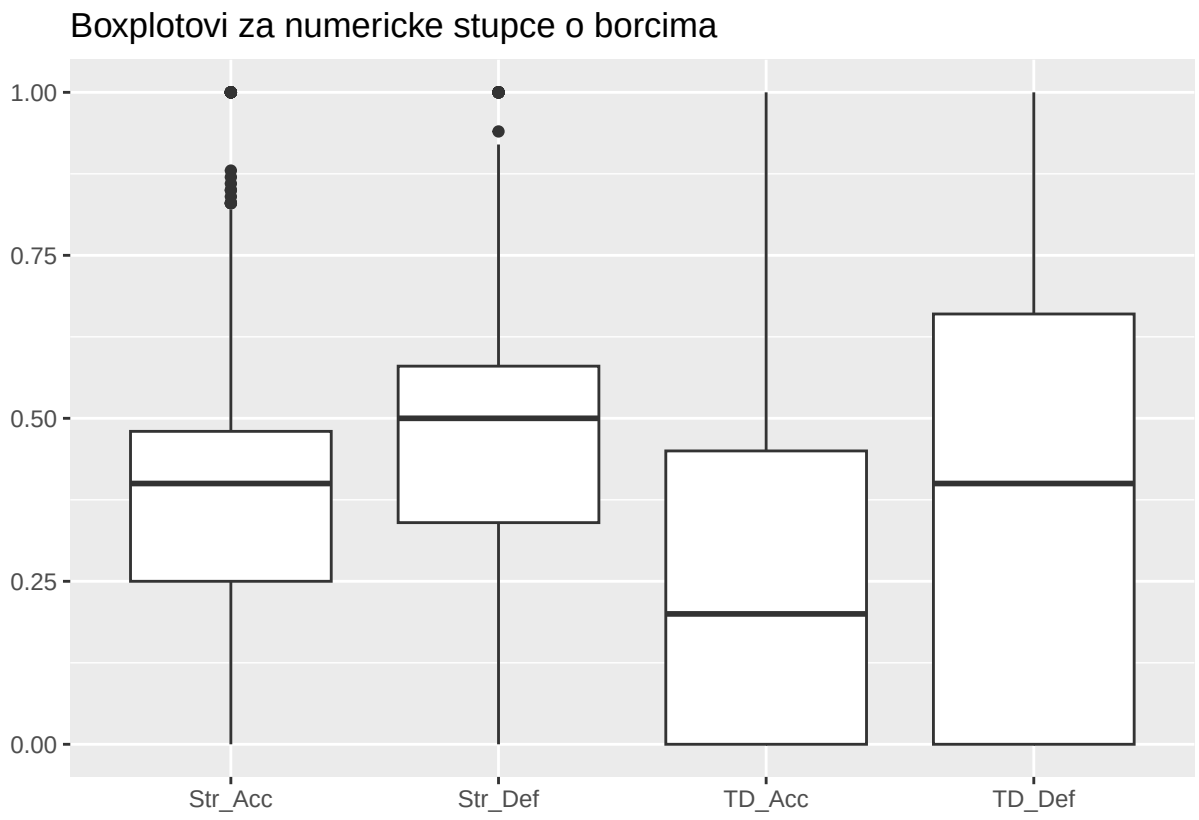
Kako bismo izračunali aritmetičku sredinu, standardnu devijaciju, najmanju i najveću vrijednost, te kvartile, za znakovne stupce moramo ih prvo pretvoriti u numeričke. To ćemo učiniti za stupce `Height`, `Str_Acc`, `Str_Def`, `TD_Acc`, `TD_Def`, `Weight`, `Reach` i `Age`.

```
fighter_data$Height <- as.numeric(str_extract(fighter_data$Height, "[0-9]+")) *
  12 + as.numeric(substring(
    str_extract(fighter_data$Height, "[0-9]+(?:\")"), 1,
    nchar(str_extract(fighter_data$Height, "[0-9]+(?:\")")) - 1
  ))
fighter_data$Str_Acc <- as.numeric(gsub("[^0-9]", "", fighter_data$Str_Acc)) /
  100
fighter_data$Str_Def <- as.numeric(gsub("[^0-9]", "", fighter_data$Str_Def)) /
  100
fighter_data$TD_Acc <- as.numeric(gsub("[^0-9]", "", fighter_data$TD_Acc)) / 100
fighter_data$TD_Def <- as.numeric(gsub("[^0-9]", "", fighter_data$TD_Def)) / 100
fighter_data$Weight <- as.numeric(gsub("[^0-9]", "", fighter_data$Weight))
fighter_data$Reach <- as.numeric(gsub("[^0-9]", "", fighter_data$Reach))
fighter_data$YOB <- as.numeric(
  substring(
    str_extract(fighter_data$DOB, ", [0-9]+"),
    3, nchar(str_extract(fighter_data$DOB, ", [0-9]+"))
  )
)
fighter_data$Age <- 2023 - fighter_data$YOB
fighter_data <- fighter_data[, -c(15)]
```

Ispišimo sada sažetke za novostvorene stupce

```
summary(fighter_data[, c(2, 3, 4, 8, 10, 12, 13, 15)])
##      Height          Weight        Reach          Str_Acc
##  Min.   :60.00   Min.   :105   Min.   :58.00   Min.   :0.0000
##  1st Qu.:68.00   1st Qu.:145   1st Qu.:69.00   1st Qu.:0.2500
##  Median :70.00   Median :170   Median :72.00   Median :0.4000
##  Mean   :70.31   Mean   :173   Mean   :71.83   Mean   :0.3447
##  3rd Qu.:73.00   3rd Qu.:185   3rd Qu.:75.00   3rd Qu.:0.4800
##  Max.   :89.00   Max.   :770   Max.   :84.00   Max.   :1.0000
##  NA's   :263     NA's   :74    NA's   :1912
##     Str_Def           TD_Acc           TD_Def            Age
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :23.00
##  1st Qu.:0.3400   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:34.00
##  Median :0.5000   Median :0.2000   Median :0.4000   Median :39.00
##  Mean   :0.4232   Mean   :0.2603   Mean   :0.3783   Mean   :39.31
##  3rd Qu.:0.5800   3rd Qu.:0.4500   3rd Qu.:0.6600   3rd Qu.:44.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :80.00
##                                                     NA's   :739
```

```
fighter_data %>%
    select(Str_Acc, Str_Def, TD_Acc, TD_Def) %>%
    gather() %>%
    ggplot(aes(x = key, y = value)) + geom_boxplot() + labs(x = "", y = "", title = "Boxplotovi za nume
```


Boxplotovi za numericke stupce o borcima
1.00
0.75
0.50
0.25
0.00
Str_Acc
Str_Def
TD_Acc
TD_Def

Analiza podataka

Možemo li očekivati završetak borbe nokautom ovisno o razlici u dužini ruku izmedu boraca?

Učitamo podatke, provjerimo dimenzije, stupce, glavu i sažetak.

```
fighter <- read.csv("fighter_details.csv") %>%
    select("fighter_name", "Reach")
dim(fighter)
## [1] 3596    2
head(fighter)
##          fighter_name Reach
## 1           Tom Aaron
## 2          Papy Abedi
## 3 Shamil Abdurakhimov   76"
## 4        Danny Abbadi
## 5        Hiroyuki Abe
## 6       Ricardo Abreu
summary(fighter)
##  fighter_name          Reach
##  Length:3596        Length:3596
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

Pripremimo podatke za daljnju analizu. Uklonimo znakove inča iz stupca `Reach` i pretvorimo stupac u numerički.

```
fighter[["Reach"]] <- as.numeric(gsub("\"", "", fighter[["Reach"]]))
head(fighter)
##          fighter_name Reach
## 1           Tom Aaron    NA
## 2          Papy Abedi    NA
## 3 Shamil Abdurakhimov    76
## 4        Danny Abbadi    NA
## 5        Hiroyuki Abe    NA
## 6       Ricardo Abreu    NA
```

Uklonimo redove s nedostajućim vrijednostima.

```
fighter <- subset(fighter, !is.na(fighter[["Reach"]]))
head(fighter)
##           fighter_name Reach
## 3  Shamil Abdurakhimov    76
## 7           Daichi Abe    71
## 9        Klidson Abreu    74
## 12          Juan Adams    80
## 13       Anthony Adams    76
## 15     Israel Adesanya    80
is.numeric(fighter[["Reach"]])
## [1] TRUE
```

Učitajmo samo potrebne stupce iz skupa podataka.

```
match <- read.csv("total_fight_data.csv", sep = ";") %>%
    select("R_fighter", "B_fighter", "win_by")
```

Spojimo oba skupa podataka kako bismo dobili dužinu ruku svakog borca u svakoj borbi.

```
merged <- match
for (s in c("R", "B")) {
    merged <- merge(merged, fighter, by.x = sprintf("%s_fighter", s), by.y = "fighter_name")
    colnames(merged)[colnames(merged) == "Reach"] <- sprintf("%s_Reach", s)
}
head(merged)
##        B_fighter     R_fighter                  win_by R_Reach B_Reach
## 1     Aalon Cruz    Uros Medic                  KO/TKO      71      78
## 2 Aaron Phillips    Jack Shore              Submission      71      71
## 3 Aaron Phillips   Sam Sicilia    Decision - Unanimous      67      71
## 4    Aaron Riley Tony Ferguson TKO - Doctor's Stoppage      76      69
## 5    Aaron Riley  Ross Pearson TKO - Doctor's Stoppage      69      69
## 6    Aaron Riley Robbie Lawler    Decision - Unanimous      74      69
```

Izračunajmo razliku u dužini ruku između boraca.

```
diff <- merged["R_Reach"] - merged["B_Reach"]
# da smo koristili apsolutne vrijednosti dobili bismo preklopljenu normalnu distribuciju
colnames(diff) <- "diff"
```

Spojimo stupac s razlikom u dužini ruku s prethodno spojenim skupom podataka.

```
merged <- cbind(merged, diff)
summary(merged)
##   B_fighter          R_fighter            win_by            R_Reach
##  Length:4964        Length:4964        Length:4964        Min.   :60.00
##  Class :character   Class :character   Class :character   1st Qu.:70.00
##  Mode  :character   Mode  :character   Mode  :character   Median :72.00
##                                                           Mean   :72.19
##                                                           3rd Qu.:75.00
##                                                           Max.   :84.00
##     B_Reach           diff
##  Min.   :58.00   Min.   :-12.00000
##  1st Qu.:70.00   1st Qu.: -2.00000
##  Median :72.00   Median :  0.00000
##  Mean   :72.15   Mean   :  0.03989
##  3rd Qu.:75.00   3rd Qu.:  2.00000
##  Max.   :84.00   Max.   : 13.00000
head(merged)
##        B_fighter     R_fighter                  win_by R_Reach B_Reach diff
## 1     Aalon Cruz    Uros Medic                  KO/TKO      71      78   -7
## 2 Aaron Phillips    Jack Shore              Submission      71      71    0
## 3 Aaron Phillips   Sam Sicilia    Decision - Unanimous      67      71   -4
## 4    Aaron Riley Tony Ferguson TKO - Doctor's Stoppage      76      69    7
## 5    Aaron Riley  Ross Pearson TKO - Doctor's Stoppage      69      69    0
## 6    Aaron Riley Robbie Lawler    Decision - Unanimous      74      69    5
```

Kako bismo mogli provjeriti ovisnost o nokautu, potrebno je prevesti stupac `win_by` u binarnu varijablu. Učinimo to tako da zamijenimo vrijednosti `KO/TKO` i `TKO - Doctor's Stoppage` s `Yes` i sve ostale vrijednosti s `No`. Binarna varijabla `knockout` će nam pomoći u daljnjem analiziranju, a stupac `win_by` ćemo ukloniti.

```
unique(merged$win_by)
## [1] "KO/TKO"                  "Submission"
## [3] "Decision - Unanimous"    "TKO - Doctor's Stoppage"
## [5] "DQ"                      "Could Not Continue"
## [7] "Decision - Split"        "Decision - Majority"
```

```
## [9] "Overturned"
yes <- "Yes"
no <- "No"
merged$win_by <- revalue(merged$win_by, c(
  "KO/TKO" = yes,
  "TKO - Doctor's Stoppage" = yes, "Decision - Unanimous" = no,
  "Submission" = no, "DQ" = no, "Could Not Continue" = no,
  "Decision - Split" = no, "Decision - Majority" = no,
  "Overturned" = no
))
unique(merged$win_by)
## [1] "Yes" "No"
colnames(merged)[colnames(merged) == "win_by"] <- "knockout"
```

Izdvajamo borbe koje su završile nokautom.

```
knockout <- subset(merged, knockout == "Yes")
head(knockout)
##     B_fighter      R_fighter knockout R_Reach B_Reach diff
## 1  Aalon Cruz     Uros Medic      Yes      71      78   -7
## 4 Aaron Riley  Tony Ferguson      Yes      76      69    7
## 5 Aaron Riley   Ross Pearson      Yes      69      69    0
## 7 Aaron Riley   Shane Nelson      Yes      70      69    1
## 8 Aaron Riley Spencer Fisher      Yes      68      69   -1
## 9  Aaron Rosa  James Te Huna      Yes      75      77   -2
```

Izdvajamo borbe koje nisu završile nokautom.

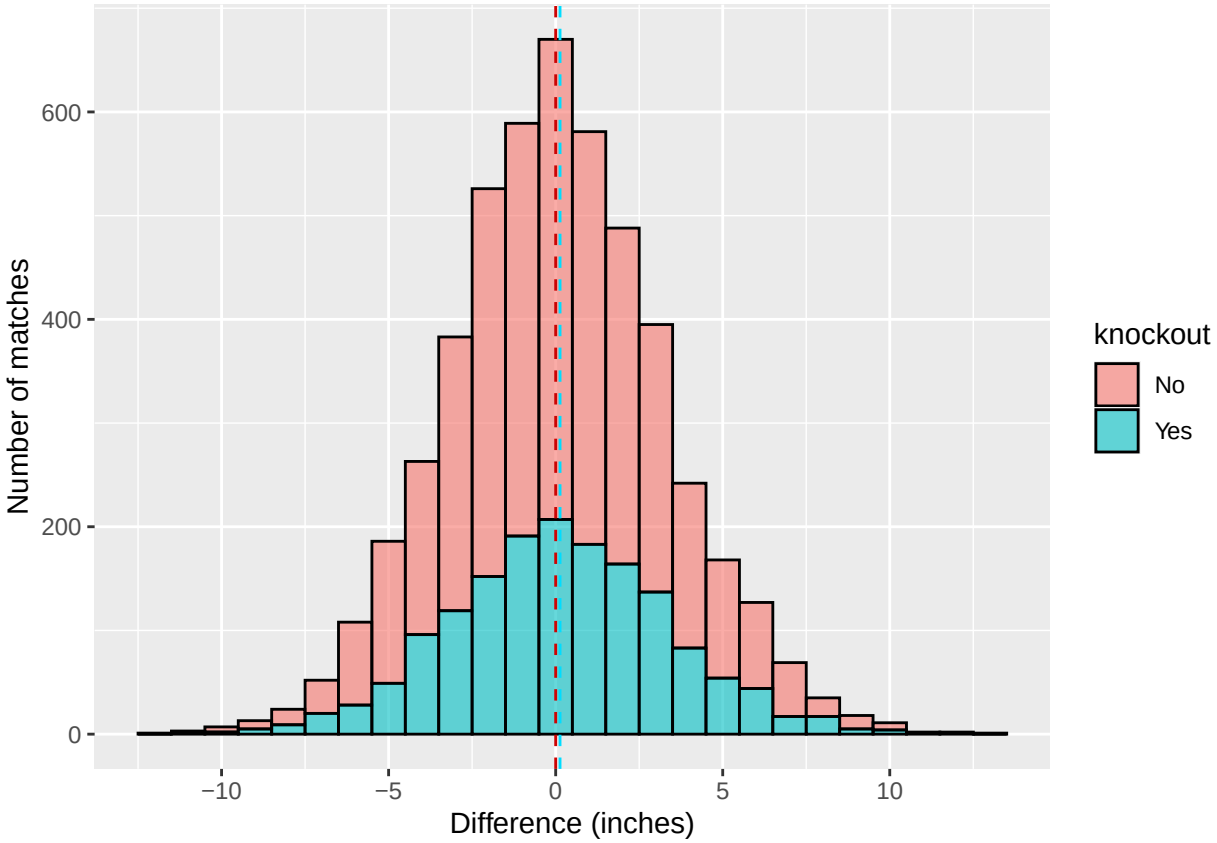
```
no_knockout <- subset(merged, knockout == "No")
head(no_knockout)
##         B_fighter     R_fighter knockout R_Reach B_Reach diff
## 2  Aaron Phillips    Jack Shore       No      71      71    0
## 3  Aaron Phillips   Sam Sicilia       No      67      71   -4
## 6     Aaron Riley Robbie Lawler       No      74      69    5
## 12  Aaron Simpson  Brad Tavares       No      74      73    1
## 13  Aaron Simpson Mario Miranda       No      74      73    1
## 16  Abel Trujillo Tony Ferguson       No      76      70    6
```

Izračunajmo srednju vrijednost razlike u dužini ruku za borbe koje su završile nokautom i borbe koje nisu završile nokautom.

```
m1 <- round(mean(knockout[["diff"]]), 2)
m2 <- round(mean(no_knockout[["diff"]]), 2)
```

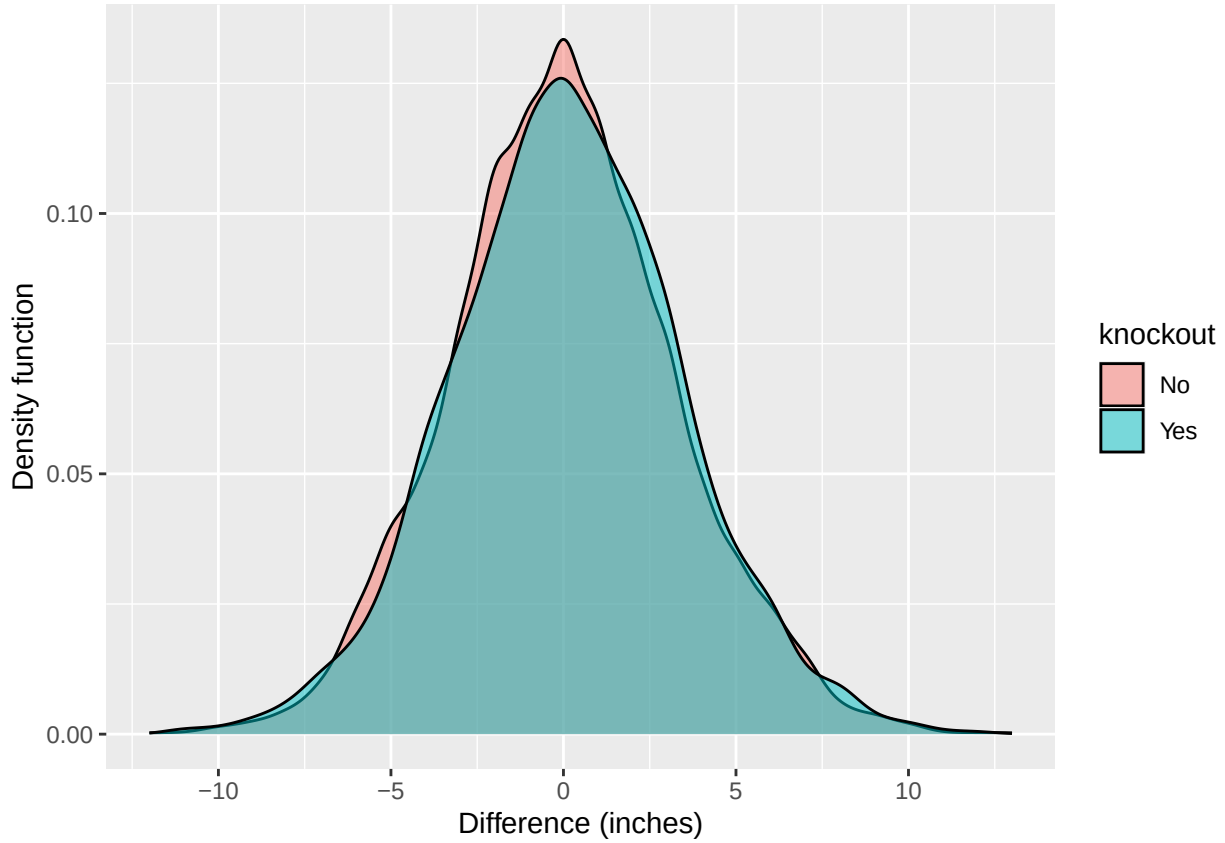
Sada možemo prikazati histogram razlike u dužini ruku za borbe koje su završile nokautom i borbe koje nisu završile nokautom.

```
merged %>%
  ggplot(aes(x = diff, fill = knockout)) +
  geom_histogram(binwidth = 1, alpha = 0.6, color = "black") +
  geom_vline(aes(xintercept = m1), color = "#00ddff", linetype = "dashed") +
  geom_vline(aes(xintercept = m2), color = "#cc0404", linetype = "dashed") +
  labs(
    x = "Difference (inches)",
    y = "Number of matches"
  )
```



Prikažimo funkciju gustoće za rezultat borbe ovisno o razlici u dužini ruku boraca.

```
ggplot(merged, aes(x = diff, fill = knockout)) +
  geom_density(alpha = 0.5) +
  labs(
    x = "Difference (inches)",
    y = "Density function"
  )
```



Možemo zaključiti da su srednje vrijednosti razlike u dužini ruku za borbe koje su završile nokautom i borbe koje nisu završile nokautom gotovo jednake. Također, možemo zaključiti da je distribucija rezultata borbe gotovo jednaka za borbe koje su završile nokautom i borbe koje nisu završile nokautom. Možemo zaključiti da razlika u dužini ruku boraca nema utjecaja na rezultat borbe, ali ćemo to provjeriti statističkim testom.

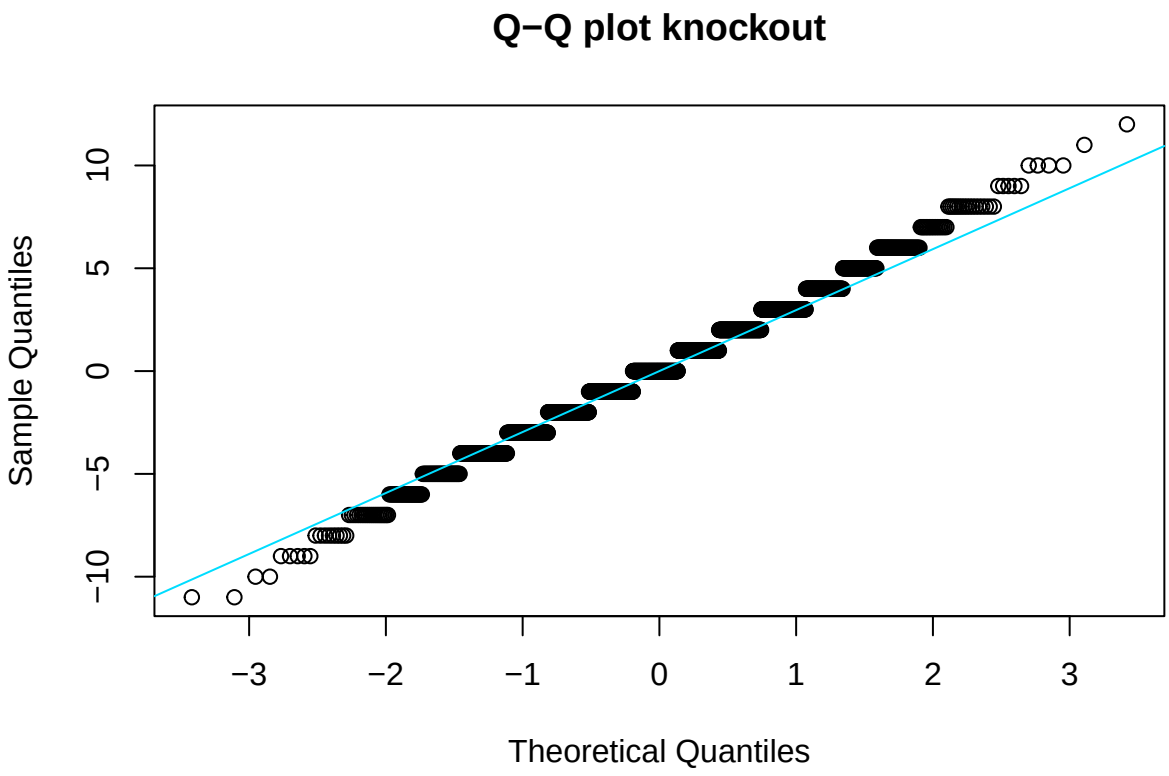
Pripremimo podatke za T test. Izdvajamo stupac s razlikom u dužini ruku za borbe koje su završile nokautom i borbe koje nisu završile nokautom.

```
dataset1 <- knockout[["diff"]]
dataset2 <- no_knockout[["diff"]]
```

T-test zahtijeva da su podaci normalno distribuirani, pa ćemo provjeriti normalnost podataka. Normalnost se može provjeriti na više načina. U sljedećim koracima provjerit ćemo normalnost na dva načina: - vizualno (qqnorm) - kvantitativnim odlukama, testovima kao što su: Lilliefors, Kolmogorov-Smirnov, Anderson-Darling test ....

Testiramo normalnost za `dataset1`

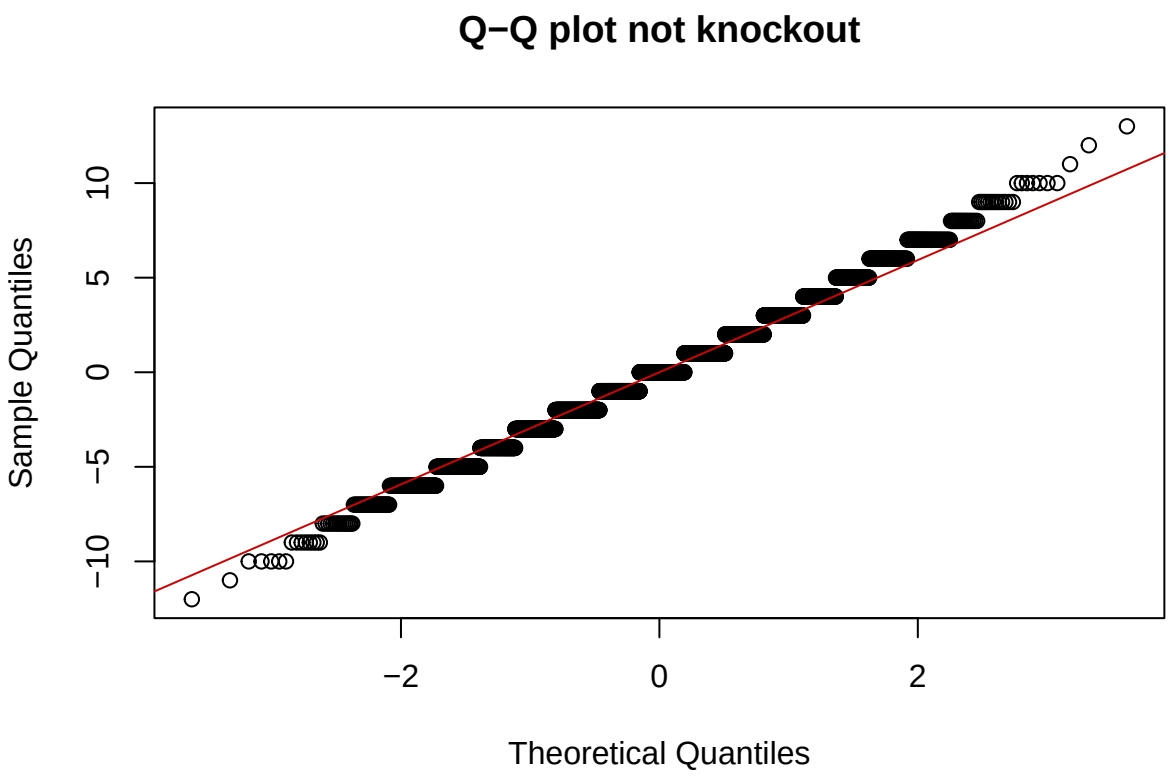
```
qqnorm(dataset1, main = "Q-Q plot knockout")
qqline(dataset1, col = "#00ddff")
```




```
lillie.test(dataset1)["p.value"]
## $p.value
## [1] 1.521733e-19
```

Testiramo normalnost za `dataset2`

```
qqnorm(dataset2, main = "Q-Q plot not knockout")
qqline(dataset2, col = "#cc0404")
```




```
lillie.test(dataset2)["p.value"]
## $p.value
## [1] 2.269751e-52
```

Lilliefors test i dalje daje nisku p-vrijednost što može ukazivati na to da podaci nisu normalno distribuirani, iako izgledaju normalno kada se vizualiziraju. Ovo objašnjavamo činjenicom kako je Lilliefors test bolji za manje skupove podataka te se njegova interpretacija kako podatci iz skupa nisu normalno distribuirani ne prihvaća. Isto vrijedi i za sljedeće testove (sintaksa kojom bi ih izvrsavali je prizana u nastavku).

Kolmogorov-Smirnov test:

```
dataset1_without_ties <- dataset1 + rnorm(length(dataset1), 0, 1e-10)
ks.test(dataset1_without_ties, "pnorm",
  mean = mean(dataset1_without_ties),
  sd = sd(dataset1_without_ties), alternative = "two.sided"
)
```

Anderson-Darling test:

```
ad.test(dataset1)
```

Ili bilo koji drugi test, uzimamo u obzir činjenicu da mnoge statističke metode, kao što su t-test, ne ovise o pretpostavci normalnosti za velike veličine uzorka. Centralni granični teorem kaže da čak i ako populacija nije normalna, srednje će vrijednosti slučajnih uzoraka dovoljne veličine biti približno normalno distribuirane ako distribucija nije previše zakrivljena (vidimo iz histograma da to nije slučaj). To znači da s povećanjem veličine uzorka t-test postaje sve otporniji na odstupanja od normalnosti u populaciji.

Nastavljamo s T testom.

Postavimo funkciju za ispisivanje rezultata testova.

```r
decision <- function(p_value, alpha = 0.05) {
    if (p_value < alpha) {
        cat("We reject the H0 hypothesis in favor of the H1 hypothesis")
    } else {
        cat("We fail to reject the H0 hypothesis")
    }
}
```

T-test zahtijeva informacije o varijanci dvaju skupova podataka, pa ćemo provjeriti jednakost varijanci F testom.

F-test je osjetljiv na nenormalnost podataka, no naši podatci na histogramu daju dobru naznaku normalnosti, te također koristimo činjenicu da se F-test temelji na omjeru dviju varijanca, koji će se stabilizirati s povećanjem veličine uzorka.

Izvodimo F test.

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

```r
cat("variances ", var(dataset1), var(dataset2))
## variances  10.9132 10.51675
mat_f_test <- var.test(dataset1, dataset2, alternative = "two.sided")
mat_f_test
##
##  F test to compare two variances
##
## data:  dataset1 and dataset2
## F = 1.0377, num df = 1589, denom df = 3373, p-value = 0.386
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.954365 1.129771
## sample estimates:
## ratio of variances
##           1.037698
decision(mat_f_test$p.value)
## We fail to reject the H0 hypothesis
```

Ne možemo odbaciti hipotezu H_0. Tvrdimo da su varijance dvaju skupova podataka gotovo jednake.

Provedimo T test. Za obje skupine podataka izvodimo nezavisni dvostrani test s jednakim varijancama.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

```r
mat_t_test <- t.test(dataset1, dataset2, alt = "two.sided", var.equal = TRUE)
mat_t_test["p.value"]
## $p.value
## [1] 0.193173
decision(mat_t_test$p.value)
## We fail to reject the H0 hypothesis
```

Ponovo ne možemo odbaciti hipotezu H_0. Ne možemo tvrditi da je razlika između dužine ruku dva borca utjecala na to da se borba završi s ili bez nokauta.

Razlikuje li se trajanje mečeva (u sekundama) između pojedinih kategorija?

Učitajmo podatke.

```
data <- read.csv("./total_fight_data.csv", sep = ";", header = TRUE)
```

Stvori novi stupac sa trajanjem borbe u sekundama.

```
data <- data %>% mutate(total_fight_time = (last_round - 1) * 5 * 60 +
  as.numeric(format(strptime(last_round_time, "%M:%S"), "%M")) * 60 +
  as.numeric(format(strptime(last_round_time, "%M:%S"), "%S")))
head(data)
##         R_fighter         B_fighter R_KD B_KD R_SIG_STR. B_SIG_STR.
## 1    Adrian Yanez     Gustavo Lopez    2    0  41 of 103   23 of 51
## 2    Trevin Giles     Roman Dolidze    0    0   27 of 57   32 of 67
## 3     Tai Tuivasa   Harry Hunsucker    1    0   14 of 18     2 of 6
## 4   Cheyanne Buys Montserrat Conejo    0    0   31 of 65   15 of 41
## 5   Marion Reneau     Macy Chiasson    0    0   30 of 63  51 of 138
## 6 Leonardo Santos      Grant Dawson    0    0   30 of 67   46 of 84
##   R_SIG_STR_pct B_SIG_STR_pct R_TOTAL_STR. B_TOTAL_STR.   R_TD    B_TD R_TD_pct
## 1           39%           45%    41 of 103     23 of 51 0 of 0  0 of 1      ---
## 2           47%           47%     43 of 73    75 of 110 1 of 2  1 of 3      50%
## 3           77%           33%     14 of 18       2 of 6 0 of 0  0 of 0      ---
## 4           47%           36%     49 of 87   136 of 168 0 of 0  4 of 5      ---
## 5           47%           36%     59 of 93    92 of 184 2 of 4  1 of 1      50%
## 6           44%           54%    74 of 115    75 of 132 1 of 2 1 of 13      50%
##   B_TD_pct R_SUB_ATT B_SUB_ATT R_REV B_REV R_CTRL B_CTRL   R_HEAD    B_HEAD
## 1       0%         0         0     0     0   0:03   0:00 32 of 83  14 of 40
## 2      33%         1         2     0     1   1:15   4:15 22 of 51  10 of 37
## 3      ---         0         0     0     0   0:10   0:00 10 of 14    1 of 5
## 4      80%         0         2     3     1   1:04   9:53 26 of 60  10 of 35
## 5     100%         0         0     0     1   2:15   3:48 14 of 40 29 of 110
## 6       7%         0         0     0     0   1:21   8:18 14 of 45  16 of 48
##    R_BODY   B_BODY    R_LEG    B_LEG R_DISTANCE B_DISTANCE R_CLINCH B_CLINCH
## 1 8 of 19   5 of 7   1 of 1   4 of 4  41 of 102   23 of 51   0 of 0   0 of 0
## 2  4 of 4  7 of 14   1 of 2 15 of 16   15 of 42   28 of 59   4 of 5   3 of 6
## 3  0 of 0   0 of 0   4 of 4   1 of 1    9 of 10     2 of 6   0 of 0   0 of 0
## 4  5 of 5   0 of 1   0 of 0   5 of 5   26 of 56   15 of 41   2 of 2   0 of 0
## 5 7 of 13 15 of 20  9 of 10   7 of 8   25 of 54  36 of 119   5 of 9 13 of 15
## 6 6 of 10 23 of 27 10 of 12   7 of 9   28 of 65   33 of 68   2 of 2  9 of 11
##   R_GROUND B_GROUND               win_by last_round last_round_time
## 1   0 of 1   0 of 0               KO/TKO          3           0:27
## 2  8 of 10   1 of 2 Decision - Unanimous          3           5:00
## 3   5 of 8   0 of 0               KO/TKO          1           0:49
## 4   3 of 7   0 of 0 Decision - Unanimous          3           5:00
## 5   0 of 0   2 of 4 Decision - Unanimous          3           5:00
## 6   0 of 0   4 of 5               KO/TKO          3           4:59
##          Format       Referee           date               location
## 1 3 Rnd (5-5-5) Chris Tognoni March 20, 2021 Las Vegas, Nevada, USA
## 2 3 Rnd (5-5-5)     Herb Dean March 20, 2021 Las Vegas, Nevada, USA
## 3 3 Rnd (5-5-5)     Herb Dean March 20, 2021 Las Vegas, Nevada, USA
## 4 3 Rnd (5-5-5)    Mark Smith March 20, 2021 Las Vegas, Nevada, USA
## 5 3 Rnd (5-5-5)    Mark Smith March 20, 2021 Las Vegas, Nevada, USA
## 6 3 Rnd (5-5-5) Chris Tognoni March 20, 2021 Las Vegas, Nevada, USA
##                 Fight_type            Winner total_fight_time
```

```
## 1         Bantamweight Bout      Adrian Yanez              627
## 2         Middleweight Bout      Trevin Giles              900
## 3          Heavyweight Bout       Tai Tuivasa               49
## 4  Women's Strawweight Bout Montserrat Conejo              900
## 5 Women's Bantamweight Bout     Macy Chiasson              900
## 6          Lightweight Bout      Grant Dawson              899
```

Očisti stupac s podacima o kategorijama tako da sadrži samo ime kategorije.

```
data$weight_class <- str_match(
  tolower(data$Fight_type),
  paste(
    "(strawweight|flyweight|bantamweight|featherweight|lightweight|",
    "welterweight|middleweight|light heavyweight|heavyweight)"
  )
)[, 1]
unique(data$weight_class)
```

```
##  [1] "bantamweight"      "middleweight"      "heavyweight"
##  [4] "strawweight"       "lightweight"       NA
##  [7] "flyweight"         "light heavyweight" "featherweight"
## [10] " welterweight"
```

Potrebno je još izbacit NA vrijednosti. Radi se o podatcima kao što su Catchweight borbe (borbe koje ne pripadaju pravoj kategoriji).

```
data <- data %>%
    filter(!is.na(weight_class))
unique(data$weight_class)
## [1] "bantamweight"      "middleweight"      "heavyweight"
## [4] "strawweight"       "lightweight"       "flyweight"
## [7] "light heavyweight" "featherweight"     " welterweight"
```

Kolmogorov-Smirnov test normalne distribucije podataka

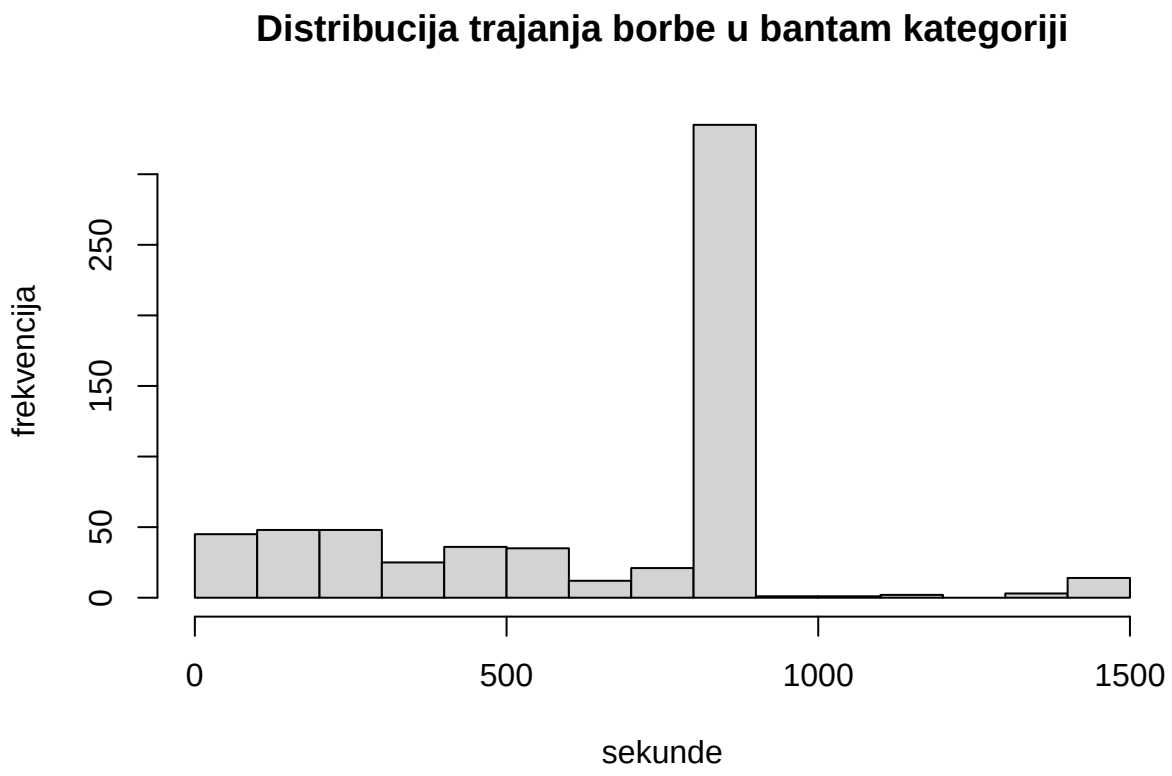
```
lillie.test(data$total_fight_time)
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$total_fight_time
## D = 0.24019, p-value < 2.2e-16
lillie.test(data$total_fight_time[data$weight_class == "middleweight"])
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$total_fight_time[data$weight_class == "middleweight"]
## D = 0.20365, p-value < 2.2e-16
```

p-vrijednost predstavlja vjerojatnost uzorkovanja jednake ili veće D vrijednosti ako su podaci normalno distribuirani. Jako mala p-vrijednost ukazuje na to da je vrlo mala vjerojatnost da smo dobili tako velik D ako su podatci normalno distribuirani.

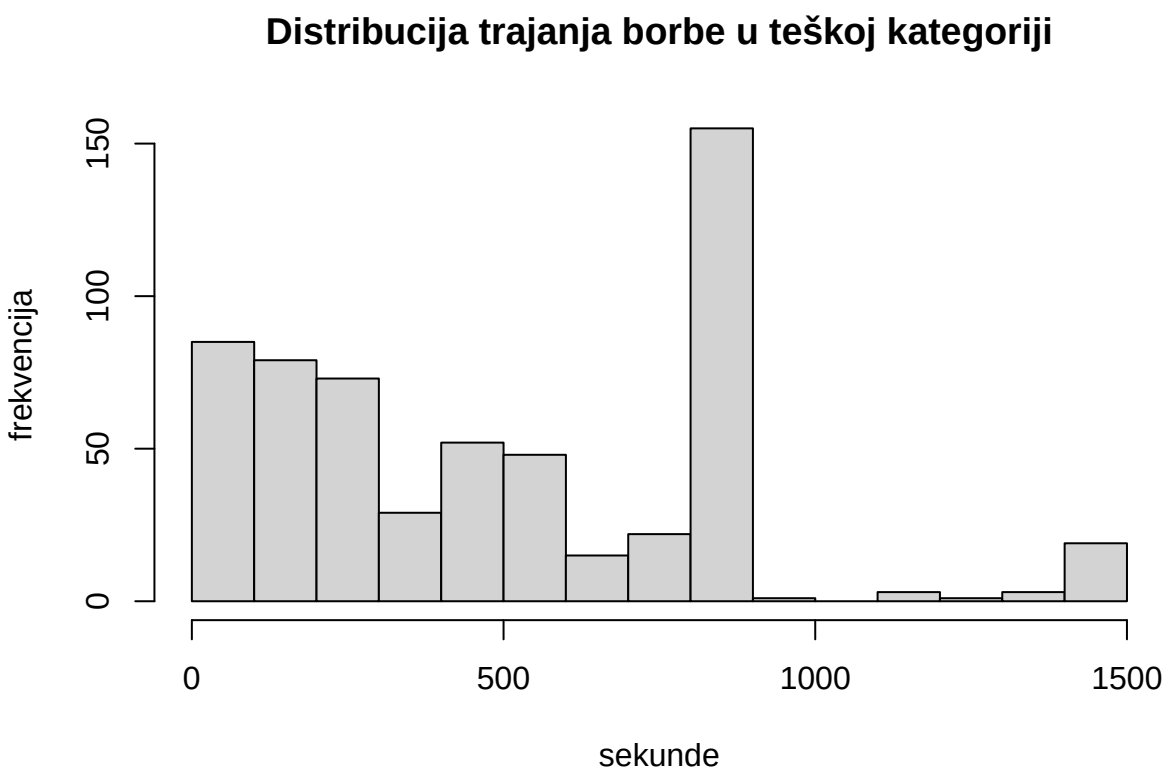
Iz tog razloga odbacujemo H_0 (normalna distribucija podataka) i zaključujemo da podatci nisu normalno distribuirani.

Histogram bantam i teške kategorije (puno više prekida u ranim trenutcima borbe u teškoj)

```
hist(data$total_fight_time[data$weight_class == "bantamweight"],
  main = "Distribucija trajanja borbe u bantam kategoriji",
  xlab = "sekunde", ylab = "frekvencija"
)
```




```
hist(data$total_fight_time[data$weight_class == "heavyweight"],
  main = "Distribucija trajanja borbe u teškoj kategoriji",
  xlab = "sekunde", ylab = "frekvencija"
)
```



Bartlett test homogenosti varijanci Definiramo hipoteze:

H_0 : varijance svih grupa su jednake

H_1 : varijance grupa nisu jednake

```
bartlett.test(total_fight_time ~ weight_class, data = data)
##
##  Bartlett test of homogeneity of variances
##
## data:  total_fight_time by weight_class
## Bartlett's K-squared = 36.549, df = 8, p-value = 1.393e-05
```

Ako gledamo na razini značajnosti 0.05, možemo odbaciti hipotezu H_0 i zaključiti da varijance nisu jednake

Kruskal-Wallis test Parametarski testovi zahtjevaju ispunjenje određenih pretpostavki o distribuciji populacije dok neparametarski testovi nemaju takve pretpostavke

ANOVA je parametarski test: prepostavlja homogenost varijanci svih grupa i normalnost distribucije reziduala, što naravno ne vrijedi uvijek.

Kruskal-Wallisov test je neparametarska alternativa (jednofaktorskoj) analizi varijance.

Koristimo ga kad ne vrijede pretpostavke o normalnosti distribucije podataka i jednakosti varijanci.

Definiramo hipoteze:

H_0 : medijani distribucija svih uzoraka su jednaki

H_1 : barem dva medijana nisu jednaka

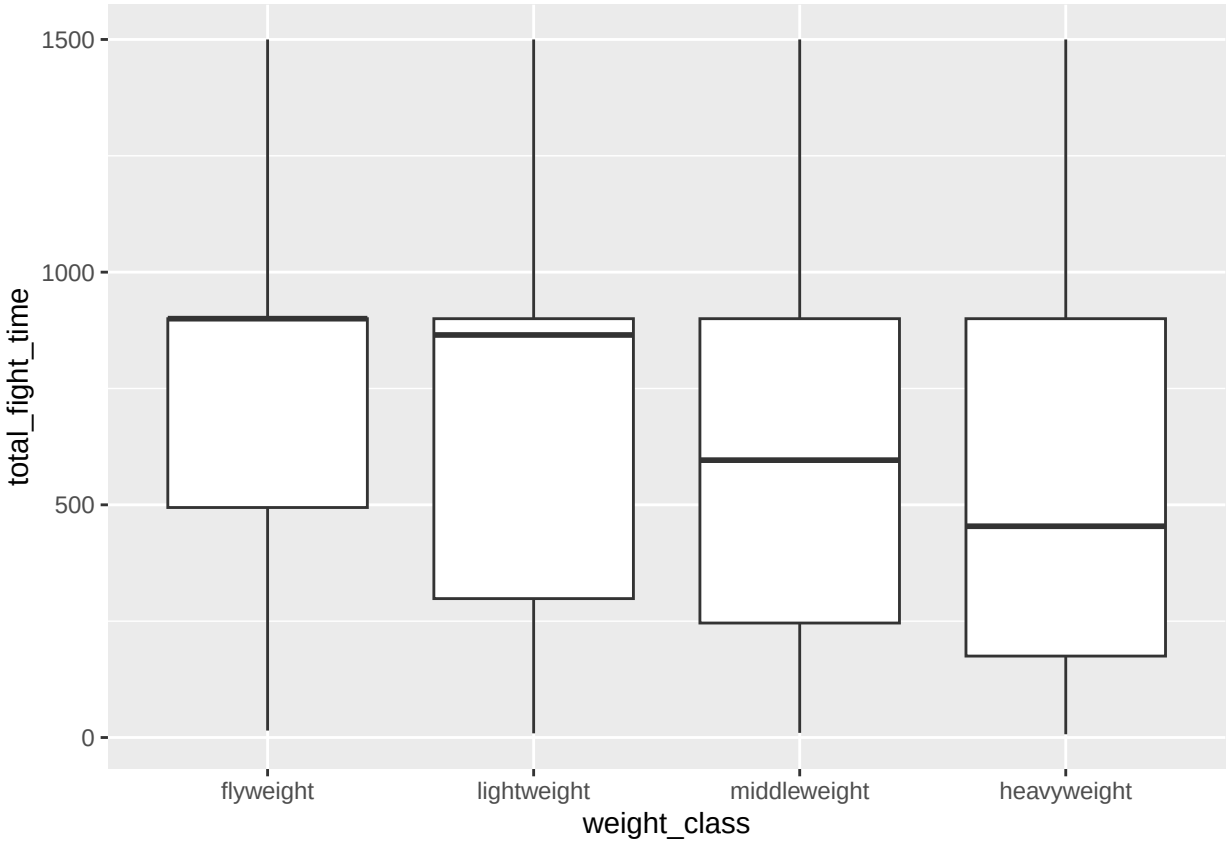
```
kruskal.test(data$total_fight_time ~ data$weight_class, data = data)
##
##  Kruskal-Wallis rank sum test
##
## data:  data$total_fight_time by data$weight_class
## Kruskal-Wallis chi-squared = 212.21, df = 8, p-value < 2.2e-16
```

S obzirom na jako malu p-vrijednost, možemo zaključiti da je test pronašao statistički značajnu razliku među grupama.

Iz tog razloga, možemo odbaciti hipotezu H_0 koja govori da su medijani svih grupa jednaki.

Boxplot trajanja borbi po kategorijama:

```
ggplot(data = data) +
  geom_boxplot(aes(x = weight_class, y = total_fight_time)) +
  scale_x_discrete(limits = c("flyweight", "lightweight", "middleweight", "heavyweight"))
```



Traju li (u rundama) borbe za titulu duže od ostalih borbi u natjecanju?

S obzirom na to da imamo kategorijske podatke i gledamo postoje li razlike u značajkama (broju rundi) između kategorija (borba za titulu ili ne), χ^2 test je najprikladniji za takve podatke. χ^2 test je parametarski test koji se koristi za testiranje hipoteza o kategorijama.

Učitaj podatke i izdvoji potrebne stupce.

```
file <- "./total_fight_data.csv"
data <- read.csv(file, sep = ";") %>%
    select("Fight_type", "last_round")
head(data)
##                  Fight_type last_round
## 1         Bantamweight Bout          3
## 2         Middleweight Bout          3
## 3          Heavyweight Bout          1
## 4  Women's Strawweight Bout          3
## 5 Women's Bantamweight Bout          3
## 6          Lightweight Bout          3
```

Je li neki borba bila za titulu saznajemo iz prisutnosti stringa `Title Bout` u `Fight_type` stupcu.

```
summary(data[0:1], maxsum = 20)
##   Fight_type
##  Length:6012
##  Class :character
##  Mode  :character
```

Broj rundi saznajemo iz `last_round` stupca.

```
summary(data$last_round)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   3.000   2.317   3.000   5.000
```

Ako stupac za neku porbu sadrži traženi niz, pišemo `yes` u novi `title` stupac, a u suprotnom pišemo `no`.

```
data$title <- ifelse(grepl("Title Bout", data$Fight_type), "yes", "no")
head(data)
##                  Fight_type last_round title
## 1         Bantamweight Bout          3    no
## 2         Middleweight Bout          3    no
## 3          Heavyweight Bout          1    no
## 4  Women's Strawweight Bout          3    no
## 5 Women's Bantamweight Bout          3    no
## 6          Lightweight Bout          3    no
```

Sada možemo iz tablice maknuti `Fight_type` stupac pošto smo izvadili sve potrebne značajke kako bismo podatke podijelili na kategorije prema broju rundi i prema tome je li borba bila za titulu.

```
data <- data %>%
    select("last_round", "title")
head(data)
##   last_round title
## 1          3    no
## 2          3    no
## 3          1    no
## 4          3    no
## 5          3    no
## 6          3    no
```

Iz tablice vidimo da je broj rundi u borbi za titulu u prosjeku veći od ostalih borbi.

```
data %>%
    group_by_all() %>%
    count()
##    last_round title freq
## 1           1    no 1675
## 2           1   yes  100
## 3           2    no  931
## 4           2   yes   58
## 5           3    no 2948
## 6           3   yes   68
## 7           4    no   15
## 8           4   yes   18
## 9           5    no   78
## 10          5   yes  121
```

Generirano tablicu zavisnosti s navedenim kategorijama

```
table <- table(data$last_round, data$title)
table
##
##       no  yes
##   1 1675  100
##   2  931   58
##   3 2948   68
##   4   15   18
##   5   78  121
```

Postavljamo hipoteze i p-vrijednost:

H_0 : broj rundi ne ovisi o tome je li borba za titulu

H_1 : broj rundi ovisi o tome je li borba za titulu

Odradimo χ^2 test:

```
chisq.test(table)
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 1259.1, df = 4, p-value < 2.2e-16
```

S obzirom na jako malu p-vrijednost, možemo zaključiti da je test pronašao statistički značajnu razliku među grupama. Iz tog razloga, možemo odbaciti hipotezu H_0 koja govori da broj rundi ne ovisi o tome je li borba za titulu.

Mogu li dostupne značajke predvidjeti pobjednika?

Učitavamo podatke.

```
total_fight_data_file <- "D:\\docs\\faks\\sap\\total_fight_data.csv"
total_fight_data <- read.csv(total_fight_data_file, sep = ";")

fighter_details <- "D:\\docs\\faks\\sap\\fighter_details.csv"
fighter_details <- read.csv(fighter_details)
```

Koristit ćemo logističku regresiju za predviđanje pobjednika borbi u UFC-u. Koristit ćemo tablicu `fighter_details` za dobivanje detalja boraca i koristit ćemo ih kao regresore. Stupci koji nas zanimaju su `Height`, `Weight`, `Reach`, `Stance`, `SLpM`, `Str_Acc`, `SApM`, `Str_Def`, `TD_Avg`, `TD_Acc`, `TD_Def` te `Sub_Avg`. Koristit ćemo stupce `R_fighter` i `B_fighter` za dobivanje detalja boraca iz tablice `fighter_details` koje ćemo koristiti kao regresore. Koristit ćemo stupac `Winner` za dobivanje stvarnog pobjednika borbe i koristit ćemo ga kao zavisnu varijablu. To ćemo učiniti tako da ćemo stvoriti novi stupac `Winner_binary` koji će biti 1 ako je borac u istom retku pobjednik i 0 ako je gubitnik. Ovaj stupac zato predstavlja zavisnu varijablu. Prvo ćemo iz tablice `total_fight_data` izvući samo stupce koji nas zanimaju. To su `R_fighter`, `B_fighter`, `date` te `Winner`.

```
fight_data <- total_fight_data %>%
    select(R_fighter, B_fighter, date, Winner)
```

Sada ćemo svaki redak raspodijeliti u dva redka. Prvi redak sadržavat će ime crvenog borca u stupcu `fighter` i podatak o tome je li pobjedio ili nije. Drugi pak će redak sadržavati ime plavog borca u stupcu `fighter` i podatak o tome je li on pobjedio.

```
detach("package:plyr", unload = TRUE)
length(fight_data$R_fighter)
## [1] 6012
fight_data <- fight_data %>%
    mutate(Winner_binary = ifelse(Winner == R_fighter, 1, 0)) %>%
    select(R_fighter, date, Winner_binary) %>%
    rename(fighter = R_fighter) %>%
    bind_rows(fight_data %>%
        mutate(Winner_binary = ifelse(Winner == B_fighter, 1, 0)) %>%
        select(B_fighter, date, Winner_binary) %>%
        rename(fighter = B_fighter))
length(fight_data$fighter)
## [1] 12024
```

Izvucimo detalje boraca iz tablice `fighter_details` i spojimo ih s tablicom `fight_data` po imenu borca.

```
fight_data <- fight_data %>%
    left_join(fighter_details, by = c(fighter = "fighter_name"))
```

Uklonimo stupce koji nam ne trebaju i redove koji sadrže nedostajuće vrijednosti.

```
fight_data <- fight_data %>%
    na.omit()
head(fight_data)
##           fighter           date Winner_binary Height   Weight Reach   Stance
## 1    Adrian Yanez March 20, 2021             1  5' 7" 135 lbs.   70" Orthodox
## 2    Trevin Giles March 20, 2021             1  6' 0" 185 lbs.   74" Orthodox
## 3     Tai Tuivasa March 20, 2021             1  6' 2" 264 lbs.   75" Southpaw
## 4   Cheyanne Buys March 20, 2021             0  5' 3" 115 lbs.   63"   Switch
## 5   Marion Reneau March 20, 2021             0  5' 6" 135 lbs.   68" Orthodox
## 6 Leonardo Santos March 20, 2021             0  6' 0" 155 lbs.   75" Orthodox
```

```
##            DOB SLpM Str_Acc SApM Str_Def TD_Avg TD_Acc TD_Def Sub_Avg
## 1 Nov 29, 1993 4.69     44% 2.31     55%   0.00     0%   100%     0.0
## 2 Aug 06, 1992 3.26     56% 1.88     62%   1.37    80%    79%     0.3
## 3 Mar 16, 1993 4.38     50% 3.44     50%   0.00     0%    46%     0.0
## 4 Jun 25, 1995 4.10     53% 1.67     65%   0.00     0%    60%     0.0
## 5 Jun 20, 1977 3.29     41% 3.37     61%   0.66    63%    50%     0.8
## 6 Feb 05, 1980 2.65     44% 2.77     58%   1.07    29%    89%     0.3
```

U regresiji ne možemo koristiti kategoričke varijable, pa ćemo ih pretvoriti u numeričke. Visine boraca zadane su u stopama i inčima, pa ćemo ih pretvoriti u inče. Vrijednosti koje su zadane kao postotci pretvorit ćemo u decimalne vrijednosti.

```
fight_data$Height <- as.numeric(str_extract(fight_data$Height, "[0-9]+")) *
  12 + as.numeric(substring(
    str_extract(fight_data$Height, "[0-9]+(?:\")"),
    1, nchar(str_extract(fight_data$Height, "[0-9]+(?:\")")) - 1
  ))
fight_data$SLpM <- as.numeric(fight_data$SLpM)
fight_data$Str_Acc <- as.numeric(gsub("[^0-9]", "", fight_data$Str_Acc)) / 100
fight_data$SApM <- as.numeric(fight_data$SApM)
fight_data$Str_Def <- as.numeric(gsub("[^0-9]", "", fight_data$Str_Def)) / 100
fight_data$TD_Avg <- as.numeric(fight_data$TD_Avg)
fight_data$TD_Acc <- as.numeric(gsub("[^0-9]", "", fight_data$TD_Acc)) / 100
fight_data$TD_Def <- as.numeric(gsub("[^0-9]", "", fight_data$TD_Def)) / 100
fight_data$Sub_Avg <- as.numeric(fight_data$Sub_Avg)
fight_data$Weight <- as.numeric(gsub("[^0-9]", "", fight_data$Weight))
fight_data$Reach <- as.numeric(gsub("[^0-9]", "", fight_data$Reach))
```

Sada još pretvorimo stupac `DOB` (datum rođenja borca) u dob borca u godinama u trenutku borbe.

```
fight_data$DOB <- as.numeric(substring(str_extract(
  fight_data$DOB,
  ", [0-9]+"
), 3, nchar(str_extract(fight_data$DOB, ", [0-9]+"))))
fight_data$DOB <- as.numeric(substring(str_extract(
  fight_data$date,
  ", [0-9]+"
), 3, nchar(str_extract(fight_data$date, ", [0-9]+")))) -
  fight_data$DOB
```

Sada moramo još na neki način ubaciti stupac `Stance` u regresiju. To činimo tako da ga pretvorimo u numeričke vrijednosti. Za svaku vrijednost stupca `Stance` stvorit ćemo novi stupac. Vrijednost tog stupca bit će 1 ako borac ima tu vrijednost stupca `Stance` i 0 ako nema.

```
fight_data$Orthodox <- ifelse(fight_data$Stance == "Orthodox", 1, 0)
fight_data$Southpaw <- ifelse(fight_data$Stance == "Southpaw", 1, 0)
fight_data$Switch <- ifelse(fight_data$Stance == "Switch", 1, 0)
fight_data$Open_Stance <- ifelse(fight_data$Stance == "Open Stance", 1, 0)
fight_data$Sideways <- ifelse(fight_data$Stance == "Sideways", 1, 0)
```

Možemo ukloniti stupce `Stance` i `date` jer nam više ne trebaju.

```
fight_data <- fight_data %>%
    select(-c(Stance, date))
head(fight_data)
##           fighter Winner_binary Height Weight Reach DOB SLpM Str_Acc SApM
```

```
## 1    Adrian Yanez             1     67    135    70  28 4.69    0.44 2.31
## 2    Trevin Giles             1     72    185    74  29 3.26    0.56 1.88
## 3     Tai Tuivasa             1     74    264    75  28 4.38    0.50 3.44
## 4   Cheyanne Buys             0     63    115    63  26 4.10    0.53 1.67
## 5   Marion Reneau             0     66    135    68  44 3.29    0.41 3.37
## 6 Leonardo Santos             0     72    155    75  41 2.65    0.44 2.77
##   Str_Def TD_Avg TD_Acc TD_Def Sub_Avg Orthodox Southpaw Switch Open_Stance
## 1    0.55   0.00   0.00   1.00     0.0        1        0      0           0
## 2    0.62   1.37   0.80   0.79     0.3        1        0      0           0
## 3    0.50   0.00   0.00   0.46     0.0        0        1      0           0
## 4    0.65   0.00   0.00   0.60     0.0        0        0      1           0
## 5    0.61   0.66   0.63   0.50     0.8        1        0      0           0
## 6    0.58   1.07   0.29   0.89     0.3        1        0      0           0
##   Sideways
## 1        0
## 2        0
## 3        0
## 4        0
## 5        0
## 6        0
```

Pripremimo izlazni vektor za regresiju. U niz r2s ćemo spremiti vrijednosti R^2 za svaku regresiju, a u niz ps ćemo spremiti p-vrijednosti za svaku regresiju.

```
r2s <- c()
ps <- c()
```

Sada ćemo napraviti regresiju za svaki stupac u `fight_data` osim za `Winner` i `fighter`. Za svaku regresiju ćemo spremiti vrijednost R^2 i p-vrijednost za koeficijente regresije. Na kraju ćemo sve to spremiti u podatkovni okvir te ga ispisati sortiranog po vrijednostima R^2.

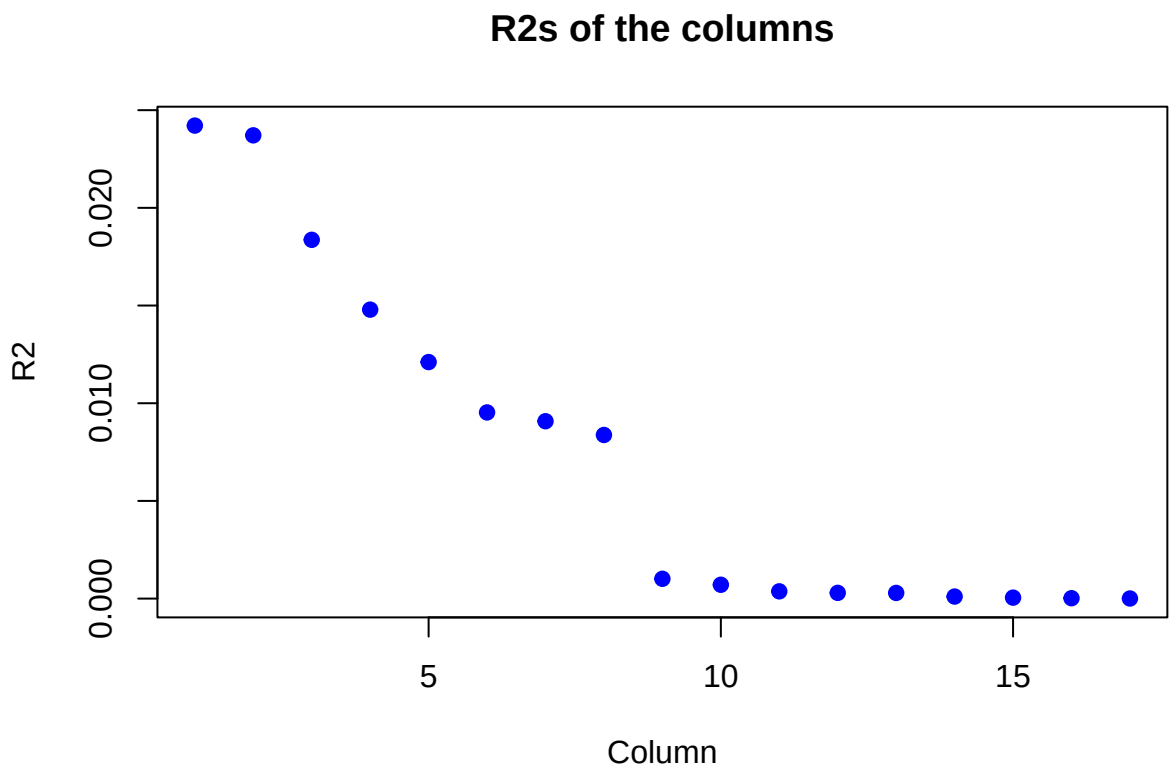
```
for (i in 3:ncol(fight_data)) {
    model <- lm(fight_data$Winner_binary ~ fight_data[[i]])
    r2s <- c(r2s, summary(model)$r.squared)
    ps <- c(ps, summary(model)$coefficients[2, 4])
}
df <- data.frame(colnames(fight_data)[3:ncol(fight_data)], r2s, ps)
df <- df[order(df[, 2], decreasing = TRUE), ]
df
##    colnames.fight_data..3.ncol.fight_data..          r2s           ps
## 5                                      SLpM 2.421088e-02 4.866871e-66
## 11                                   TD_Def 2.371218e-02 1.060638e-64
## 8                                   Str_Def 1.836399e-02 2.199907e-50
## 6                                   Str_Acc 1.479078e-02 7.484920e-41
## 10                                   TD_Acc 1.210644e-02 1.047759e-33
## 7                                      SApM 9.528528e-03 7.496232e-27
## 9                                    TD_Avg 9.078344e-03 1.178783e-25
## 4                                       DOB 8.375784e-03 2.356324e-23
## 3                                     Reach 1.013911e-03 9.725005e-04
## 14                                 Southpaw 7.086784e-04 3.508021e-03
## 12                                  Sub_Avg 3.713177e-04 3.460368e-02
## 1                                    Height 2.928434e-04 6.074847e-02
## 13                                 Orthodox 2.895576e-04 6.206143e-02
## 15                                   Switch 1.045970e-04 2.621289e-01
## 17                                 Sideways 4.956630e-05 4.401559e-01
```

```
## 16                              Open_Stance 2.066426e-05 6.181912e-01
## 2                                   Weight 5.716842e-06 7.932861e-01
```

Vidimo da je najbolji regresor za predviđanje pobjednika borbe `SLpM` (broj udaraca u minuti). Regresori koji se nalaze pri vrhu tablice su slični `SLpMu`. Svi su to regresori koji se odnose na udarce. To je logično jer je borba UFC-a borba udaraca. Svi ostali regresori dosta su slabiji od ovih. To je također logično jer su svi ostali regresori neki osobni detalji o borcu koji se ne mogu izračunati iz njegovih udaraca.

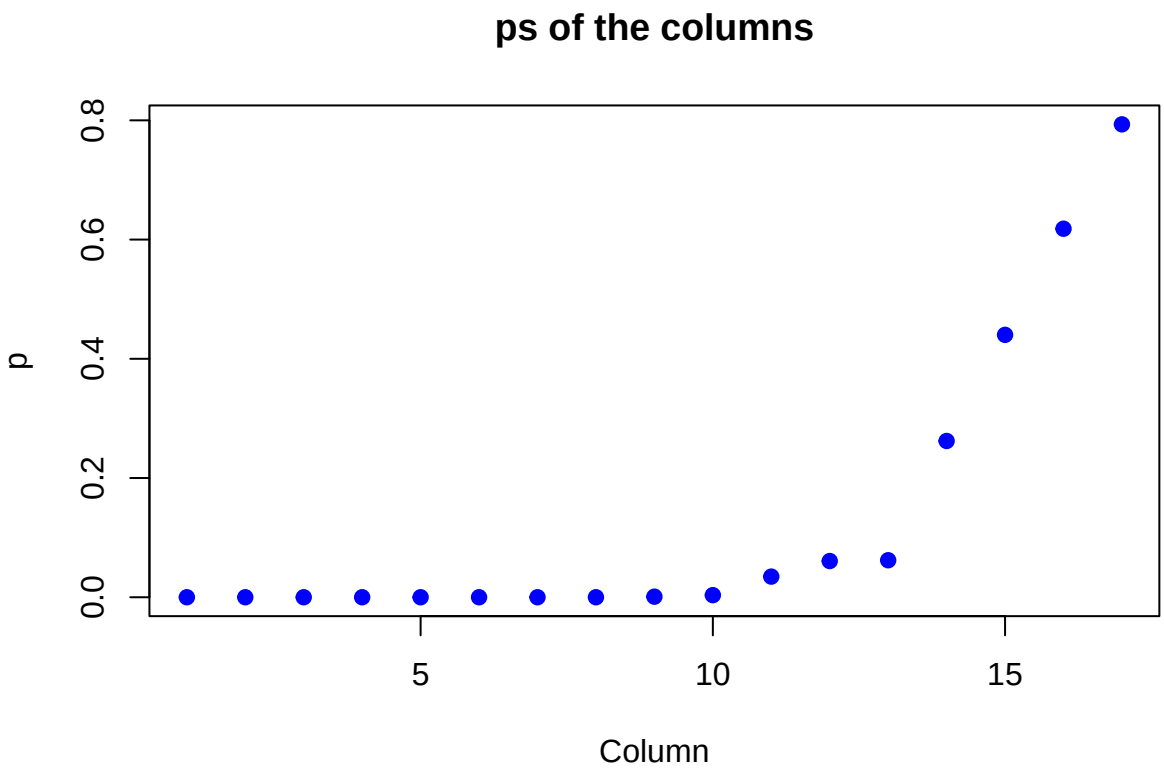
Nacrtajmo graf R^2 vrijednosti za svaki stupac.

```
plot(df[, 2], xlab = "Column", ylab = "R2", main = "R2s of the columns", pch = 19, col = "blue")
```



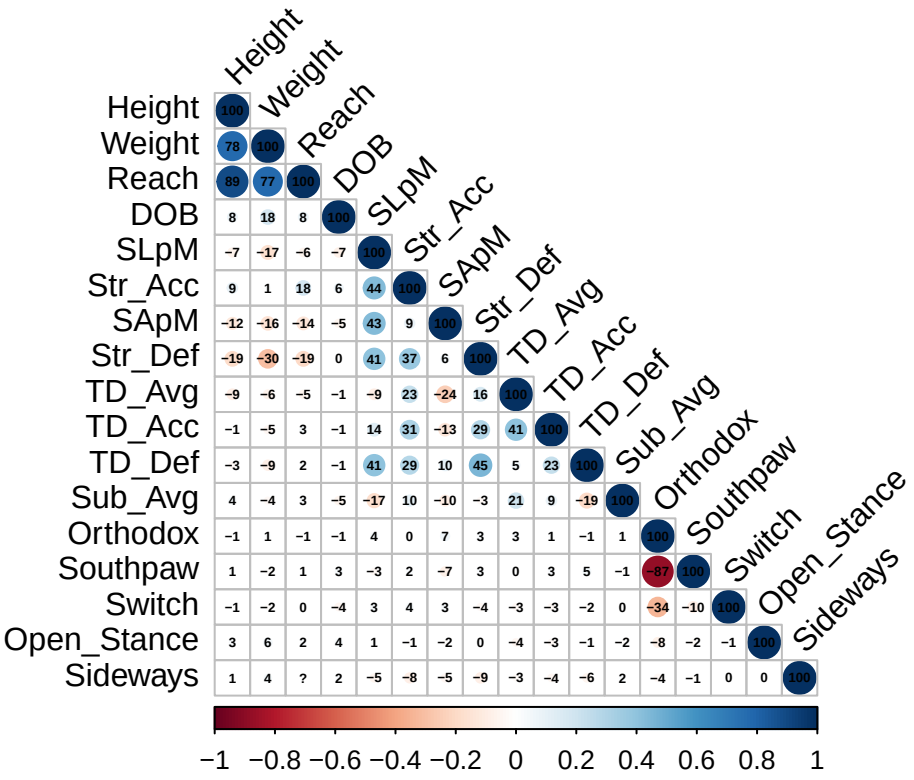
Također nacrtajmo graf p-vrijednosti za svaki stupac.

```
plot(df[, 3], xlab = "Column", ylab = "p", main = "ps of the columns", pch = 19, col = "blue")
```



Još možemo nacrtati matricu korelacija.

```
corrplot(cor(fight_data[, 3:ncol(fight_data)], use = "pairwise.complete.obs"),
  method = "circle", type = "lower", tl.col = "black", tl.srt = 45,
  addCoef.col = "black", addCoefasPercent = TRUE, number.cex = 0.4
)
```



Primjećujemo da strogo pozitivno međusobno koreliraju visina, širina i dužina ruke. S druge strane strogo negativno koreliraju `Stance` vrijednosti `Orthodox` te `Southpaw`. To je potpuno očekivano s obzirom na činjenicu da su `Orthodox` i `Southpaw` suprotni borilački stavovi. `Orthodox` jest klasičan osnovni borilački stav dešnjaka, dok je `Southpaw` klasičan osnovni borilački stav ljevaka. `Sideways` slabo korelira s prethodno navedenim stavovima zbog toga što `Sideways` može biti i `Orthodox` i `Southpaw`. `Switch` je borilački stav koji se koristi kada borac želi promijeniti borilački stav u borbi. `Open Stance` je borilački stav koji se koristi kada borac želi biti fleksibilniji u borbi. `Switch` i `Open Stance` su vrlo slični, ali `Switch` je češće korišten od `Open Stance` pa je to i razlog zašto su `Switch` i `Open Stance` slabije korelirani s ostalim stavovima.

Ima li crveni borac (često prvak) veću vjerojatnost pobjede u mečevima?

Učitajmo podatke i provjerimo dimenzije, stupce, glavu i sažetak.

```
data <- read.csv("total_fight_data.csv", sep = ";") %>%
    select("R_fighter", "B_fighter", "Winner")
```

Dodajmo novi stupac "R_Won" koji će biti popunjen sa 1 ako je crveni borac pobijedio i 0 ako nije.

```
data <- data %>%
    mutate(R_won = ifelse(R_fighter == Winner, 1, 0))
```

Ispišimo glavu i sažetak.

```
head(data)
##         R_fighter         B_fighter            Winner R_won
## 1    Adrian Yanez     Gustavo Lopez      Adrian Yanez     1
## 2    Trevin Giles     Roman Dolidze      Trevin Giles     1
## 3     Tai Tuivasa   Harry Hunsucker       Tai Tuivasa     1
## 4   Cheyanne Buys Montserrat Conejo Montserrat Conejo     0
## 5   Marion Reneau     Macy Chiasson     Macy Chiasson     0
## 6 Leonardo Santos      Grant Dawson      Grant Dawson     0
summary(data)
##   R_fighter          B_fighter           Winner              R_won
##  Length:6012        Length:6012        Length:6012        Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.0000
##                                                           Mean   :0.6618
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
```

Prebrojimo koliko je puta crveni borac pobijedio.

```
r_won <- sum(data$R_won == 1)
```

Prebrojimo koliko je puta plavi borac pobijedio.

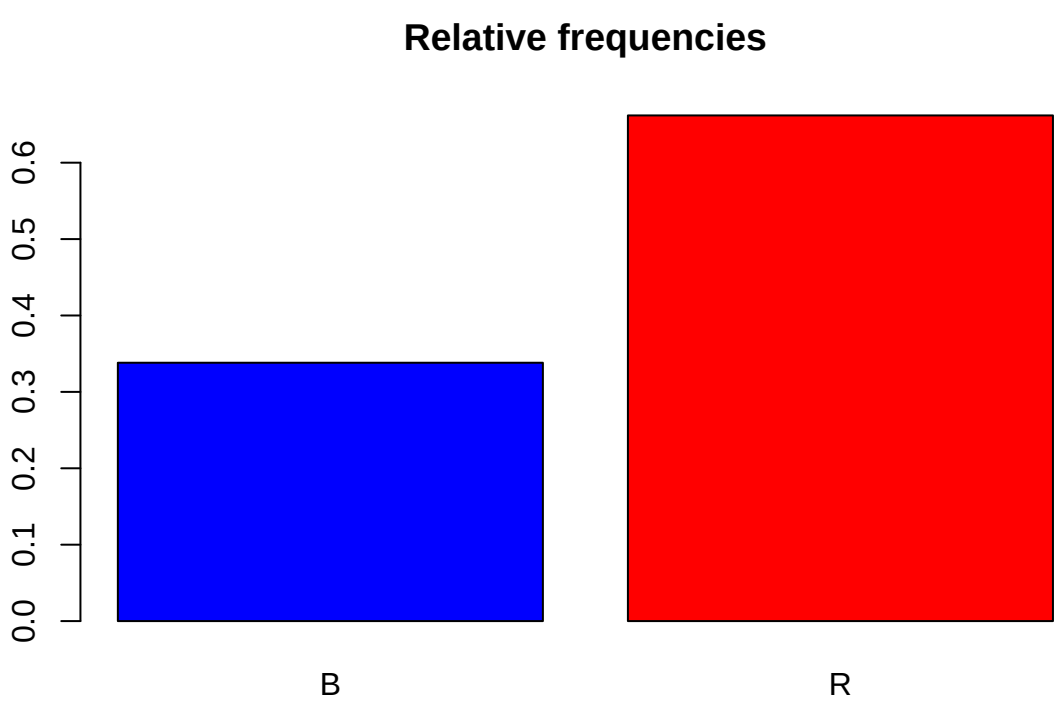
```
b_won <- sum(data$R_won == 0)
```

Prikažimo koliko je puta crveni i plavi borac pobijedio.

```
c(r_won, b_won)
## [1] 3979 2033
```

Prikazujemo relativnu frekvenciju pobjeda crvenog i plavog borca.

```
barplot(table(data$R_won) / nrow(data),
  main = "Relative frequencies",
  col = c("blue", "red"),
  border = "black",
  names.arg = c("B", "R")
)
```



Dohvatimo broj redaka u stupcu "R_won".

```
n <- length(data$R_won)
```

Provodimo binomni test. Ako je crveni borac izabran nasumično, onda bi borac u crvenom kutu bio pobjednik u 50 % mečeva.

$$H_0 : p = 0.5$$

$$H_1 : p > 0.5$$

```
alpha <- 0.05
p_value <- binom.test(r_won, n,
  p = 0.5, alternative = "greater",
  conf.level = 1 - alpha
)["p.value"]
p_value
## $p.value
## [1] 1.18179e-141
if (p_value < alpha) {
  cat("We reject the H0 hypothesis in favor of the H1 hypothesis")
} else {
  cat("We fail to reject the H0 hypothesis")
}
## We reject the H0 hypothesis in favor of the H1 hypothesis
```

Statistički zaključujemo da borac u crvenom ima vjerojatnost pobjede veću od 50 % te da ga možemo smatrati favoritom meča.