Analiza UFC borbi

SAP MATER

```
# Potrebni paket i učitavanje dataseta
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
fights = read.csv("./total_fight_data.csv")
fighters = read.csv("./fighter_details.csv")
RBData = read.csv("./preprocessed_data.csv")  # Data just about red and blue fighters
```

Case study: *UFC borbe*

Analiza sportskih dogadaja (utakmica, mečeva, borbi...) kao i performansi pojedinih sportaša zadire u pore svakog dijela sportske industrije (menadžment, trening, sponzorstvo, marketing i sportske prognoze). Iz tog razloga se već duži niz godina prikupljaju podatci Ultimate Fighting Championship (UFC) borbi.

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, težine, dužina ruke ili stav borca. Dodatno je poznat i pobjednik svake borbe (i način pobjede). U nastavku se nalaze istraživača pitanja na koja će biti dani odgovori.

Općenita deskriptivna analiza

Varijable podataka su sljedeće:

```
names(fights)
```

```
##  [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"            "Year"            "City"
```

```
## [41] "State"           "Country"         "Fight_type"      "Winner"
```

```
total_fight = left_join(fights, fighters, by = c(R_fighter = "fighter_name"))
total_fight <- total_fight %>%
    rename(R_Height = "Height", R_Weight = "Weight", R_Reach = "Reach",
        R_Stance = "Stance", R_DOB = "DOB", R_SLpM = "SLpM",
        R_Str_Acc = "Str_Acc", R_SApM = "SApM", R_Str_Def = "Str_Def",
        R_TD_Avg = "TD_Avg", R_TD_Acc = "TD_Acc", R_TD_Def = "TD_Def",
        R_Sub_Avg = "Sub_Avg")
total_fight = left_join(total_fight, fighters, by = c(B_fighter = "fighter_name"))
total_fight <- total_fight %>%
    rename(B_Height = "Height", B_Weight = "Weight", B_Reach = "Reach",
        B_Stance = "Stance", B_DOB = "DOB", B_SLpM = "SLpM",
        B_Str_Acc = "Str_Acc", B_SApM = "SApM", B_Str_Def = "Str_Def",
        B_TD_Avg = "TD_Avg", B_TD_Acc = "TD_Acc", B_TD_Def = "TD_Def",
        B_Sub_Avg = "Sub_Avg")
names(total_fight)
```

```
##  [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"            "Year"            "City"
## [41] "State"           "Country"         "Fight_type"      "Winner"
## [45] "R_Height"        "R_Weight"        "R_Reach"         "R_Stance"
## [49] "R_DOB"           "R_SLpM"          "R_Str_Acc"       "R_SApM"
## [53] "R_Str_Def"       "R_TD_Avg"        "R_TD_Acc"        "R_TD_Def"
## [57] "R_Sub_Avg"       "B_Height"        "B_Weight"        "B_Reach"
## [61] "B_Stance"        "B_DOB"           "B_SLpM"          "B_Str_Acc"
## [65] "B_SApM"          "B_Str_Def"       "B_TD_Avg"        "B_TD_Acc"
## [69] "B_TD_Def"        "B_Sub_Avg"
```

Neke od važnijih varijabli su winner (pobjednik borbe koji može biti R - red ili B - blue ovisno o kutu), win_by (način pobjede, može biti KO/TKO, medicinski TKO...), težinska kategorija boraca (heavyweight, strawweight, featherweight...), itd...

Naizgled nam sve varijable daju korisne informacije i statistički su relevantne, stoga za sada neće biti uklonjena niti jedna varijabla.

Q-Q diagram za najvažnije varijable

Možda koji histogram za najvažnije podatke

Box plot za najvažnije podatke

```
summary(total_fight)
```

```
##   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                Year          City
##  Length:6012        Length:6012         Min.   :1994   Length:6012
##  Class :character   Class :character    1st Qu.:2011   Class :character
##  Mode  :character   Mode  :character    Median :2014   Mode  :character
##                                         Mean   :2013
##                                         3rd Qu.:2018
##                                         Max.   :2021
##     State              Country           Fight_type           Winner
##  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_Height           R_Weight           R_Reach            R_Stance
##  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_DOB               R_SLpM         R_Str_Acc             R_SApM
##  Length:6012        Min.   : 0.000   Length:6012        Min.   : 0.000
##  Class :character   1st Qu.: 2.420   Class :character   1st Qu.: 2.340
##  Mode  :character   Median : 3.130   Mode  :character   Median : 2.980
##                     Mean   : 3.224                      Mean   : 3.088
##                     3rd Qu.: 3.940                      3rd Qu.: 3.750
##                     Max.   :19.910                      Max.   :23.330
##   R_Str_Def            R_TD_Avg         R_TD_Acc           R_TD_Def
##  Length:6012        Min.   : 0.000   Length:6012        Length:6012
##  Class :character   1st Qu.: 0.640   Class :character   Class :character
##  Mode  :character   Median : 1.370   Mode  :character   Mode  :character
##                     Mean   : 1.625
##                     3rd Qu.: 2.380
##                     Max.   :14.190
##    R_Sub_Avg          B_Height           B_Weight           B_Reach
##  Min.   : 0.0000   Length:6012        Length:6012        Length:6012
##  1st Qu.: 0.1000   Class :character   Class :character   Class :character
##  Median : 0.5000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.6889
##  3rd Qu.: 1.0000
##  Max.   :21.9000
##    B_Stance            B_DOB               B_SLpM         B_Str_Acc
##  Length:6012        Length:6012         Min.   : 0.000   Length:6012
##  Class :character   Class :character    1st Qu.: 2.150   Class :character
##  Mode  :character   Mode  :character    Median : 3.040   Mode  :character
##                                         Mean   : 3.059
```

```
##                                      3rd Qu.: 3.880
##                                      Max.   :15.070
##      B_SApM       B_Str_Def           B_TD_Avg        B_TD_Acc
##  Min.   : 0.00   Length:6012        Min.   : 0.000   Length:6012
##  1st Qu.: 2.45   Class :character   1st Qu.: 0.480   Class :character
##  Median : 3.07   Mode  :character   Median : 1.190   Mode  :character
##  Mean   : 3.26                      Mean   : 1.462
##  3rd Qu.: 3.90                      3rd Qu.: 2.160
##  Max.   :22.50                      Max.   :13.950
##    B_TD_Def           B_Sub_Avg
##  Length:6012        Min.   : 0.0000
##  Class :character   1st Qu.: 0.0000
##  Mode  :character   Median : 0.4000
##                     Mean   : 0.6469
##                     3rd Qu.: 0.9000
##                     Max.   :16.4000
```

Iz ovog je vidljivo da postoje outlieri koji će najvjerojatnije biti izbačeni u nekom trenutku analize (npr. B_total_time_fought.seconds gdje je minimalna vrijednost 7, iako relevantna, informacija je preveliki outlier ako ju usporedimo sa medijanom ili srednjom vrijednošću).

```
sapply(fights, class)
```

```
##       R_fighter       B_fighter            R_KD            B_KD      R_SIG_STR.
##     "character"     "character"       "integer"       "integer"     "character"
##      B_SIG_STR.   R_SIG_STR_pct   B_SIG_STR_pct    R_TOTAL_STR.    B_TOTAL_STR.
##     "character"     "character"     "character"     "character"     "character"
##            R_TD            B_TD        R_TD_pct        B_TD_pct       R_SUB_ATT
##     "character"     "character"     "character"     "character"       "integer"
##       B_SUB_ATT           R_REV           B_REV          R_CTRL          B_CTRL
##       "integer"       "integer"       "integer"     "character"     "character"
##          R_HEAD          B_HEAD          R_BODY          B_BODY           R_LEG
##     "character"     "character"     "character"     "character"     "character"
##           B_LEG      R_DISTANCE      B_DISTANCE        R_CLINCH        B_CLINCH
##     "character"     "character"     "character"     "character"     "character"
##        R_GROUND        B_GROUND          win_by      last_round last_round_time
##     "character"     "character"     "character"       "integer"     "character"
##          Format         Referee            Date            Year            City
##     "character"     "character"     "character"       "integer"     "character"
##           State         Country      Fight_type          Winner
##     "character"     "character"     "character"     "character"
```

Možemo vidjeti da je su podatci većinski sastavljeni od cijelih i decimalnih brojeva, te postoji nekoliko varijabli sa tekstualnim vrijednostima.

Pretraživanje dataseta za nedostajuće podatke:

```
# daje popis svih varijabli koje imaju NA u njima
for (name in names(total_fight)) {
    if (sum(is.na(total_fight[, name])) > 0) {
        cat("Broj nedostajućih vrijednosti za varijablu ", name,
            ": ", sum(is.na(total_fight[, name])), "\n")
    }
}
```

Iz prethodnog je vidljivo da varijable R_Reach i B_Reach imaju nedostajućih podataka. Pošto su iste kasnije potrebne u testiranju, te ih stoga nećemo ukloniti, već će biti profiltrirane prije nego li će se koristiti

za testiranje.

```
# daje popis svih varijabli koje imaju stringove duljine 0
# u njima
for (name in names(total_fight)) {
    if (!sum(is.na(total_fight[, name])) > 0) {
        if (sum(nchar(total_fight[, name]) == 0) > 0) {
            cat("Broj nedostajućih vrijednosti za varijablu ",
                name, ": ", sum(nchar(total_fight[, name]) ==
                  0), "\n")
        }
    }
}
```

```
## Broj nedostajućih vrijednosti za varijablu  Referee :  32
## Broj nedostajućih vrijednosti za varijablu  Fight_type :  6
## Broj nedostajućih vrijednosti za varijablu  Winner :  618
## Broj nedostajućih vrijednosti za varijablu  R_Height :  4
## Broj nedostajućih vrijednosti za varijablu  R_Weight :  2
## Broj nedostajućih vrijednosti za varijablu  R_Reach :  406
## Broj nedostajućih vrijednosti za varijablu  R_Stance :  29
## Broj nedostajućih vrijednosti za varijablu  R_DOB :  63
## Broj nedostajućih vrijednosti za varijablu  B_Height :  10
## Broj nedostajućih vrijednosti za varijablu  B_Weight :  8
## Broj nedostajućih vrijednosti za varijablu  B_Reach :  891
## Broj nedostajućih vrijednosti za varijablu  B_Stance :  66
## Broj nedostajućih vrijednosti za varijablu  B_DOB :  172
```

Nadalje ispis iznad također pokazuje varijable sa nedostajućim podatcima, te kao i u prethodnom slučaju, ovi podatci će biti profiltrirani ovisno o potrebi.

Pitanje 1:

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

Odabir metode Ovo pitanje možemo preformulirati kao: *Postoji li zavisnost između razlike u dužini ruku boraca i završetka borbe nokautom?*
Pa tako možemo postaviti i pitanje: *Postoji li nezavisnost između razlike u dužini ruku boraca i završetka borbe nokautom?*
Prema ovim pitanjima moglo bi se zaključiti da bi trebalo provesti test nezavisnosti, no ključna informacija je da obje varijable nisu kategorijeske - razlika duljine ruku između boraca je kvantitativna varijabla, dok je ishod borbe kategorijska varijabla. Iz tog razloga test nezavisnosti nije valjana opcija za dobivanje odgovora na ovo pitanje.

Pošto test nezavisnosti nije dobra opcija, druga metoda kojom bi se moglo doći do odgovora na ovo pitanje je **analiza varijance**. Imajući samo jednu nezavisnu varijablu (razlika duljine ruku između boraca) riječ je o jednofaktorskoj analizi varijance.

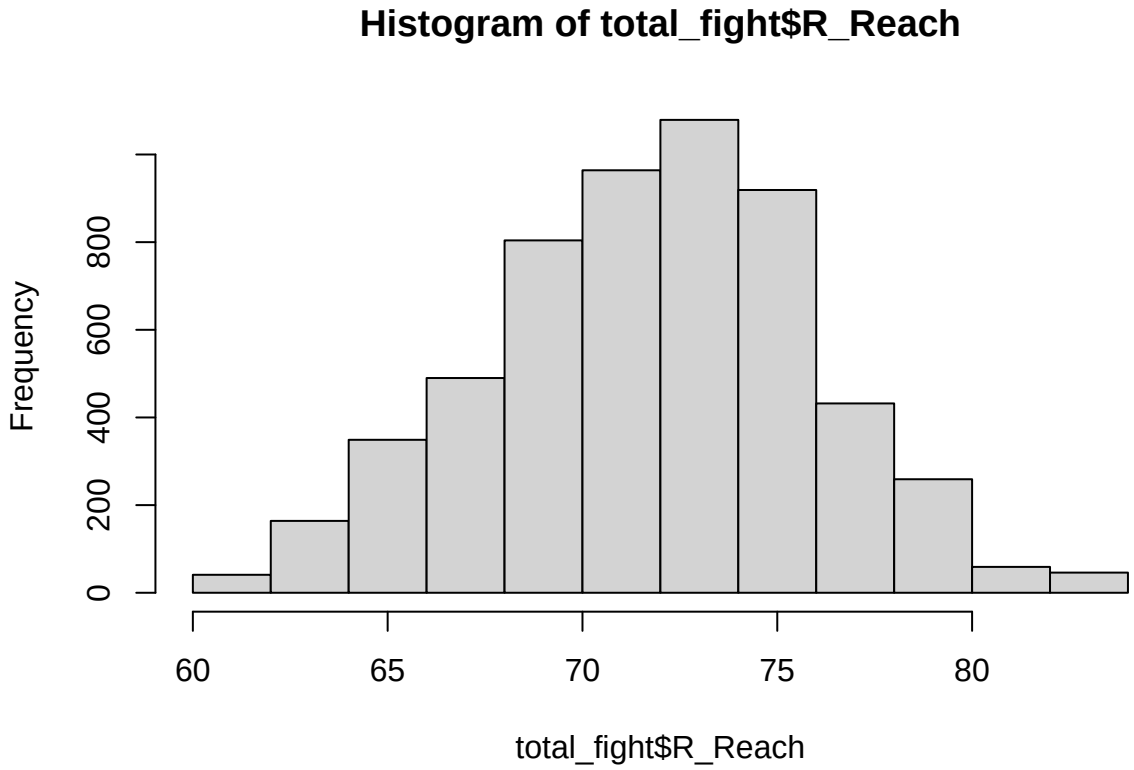
```
head(total_fight$B_Weight)
```

```
## [1] "135 lbs." "205 lbs." "241 lbs." "115 lbs." "135 lbs." "145 lbs."
```

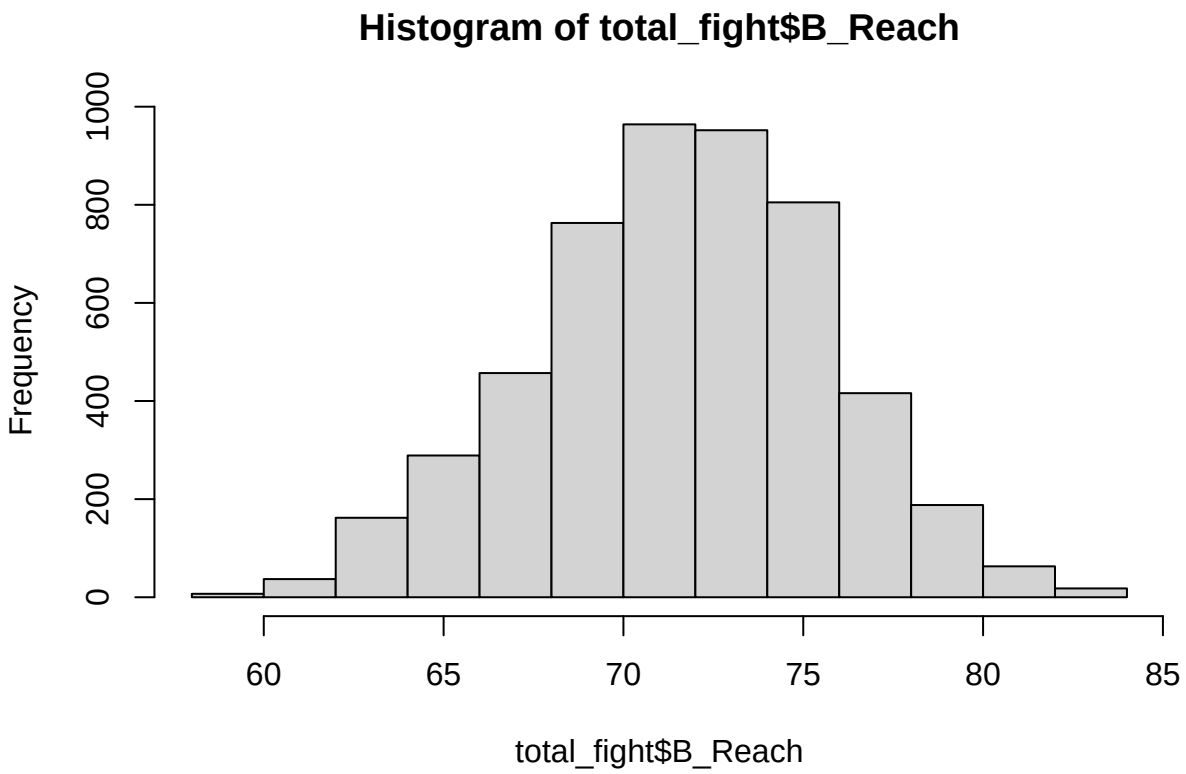
```
total_fight$R_Reach = strtoi(substr(total_fight$R_Reach, 1, 2))
total_fight$B_Reach = strtoi(substr(total_fight$B_Reach, 1, 2))
head(total_fight$R_Reach)
```

```
## [1] 70 74 75 63 68 75
```

```
hist(total_fight$R_Reach)
```




```
hist(total_fight$B_Reach)
```



Uvod

Filtriranje podataka Pošto smo već prije uočili da u ovim varijablama postoje objekti koji nemaju vrijednosti, prije nego li počnemo raditi s njima, podatci će prvo biti profiltrirani kako bi se nezadovoljavajući objekti uklonili:

```
arm_lengths <- select(total_fight, c("B_Reach", "R_Reach"))
arm_lengths <- arm_lengths[rowSums(is.na(arm_lengths)) == 0,
    ]
arm_lengths[0:20, ]
```

```
##    B_Reach R_Reach
## 1       67      70
## 2       76      74
## 3       75      75
## 4       61      63
## 5       72      68
## 6       72      75
## 7       76      71
## 8       81      77
## 9       70      75
## 11      66      68
## 12      79      77
## 13      72      74
## 14      71      69
## 15      74      75
## 16      69      64
```

```
## 17      69      70
## 18      66      71
## 19      70      72
## 20      67      67
## 21      70      75
```

```
# provjera za razliku duljine ruku
for (name in names(arm_lengths)) {
    if (sum(is.na(arm_lengths[, name])) > 0) {
        cat("Broj nedostajućih vrijednosti za varijablu ", name,
            ": ", sum(is.na(arm_lengths[, name])), "\n")
    }
}

win_methods <- select(total_fight, c("B_Reach", "R_Reach", "win_by"))
win_methods <- win_methods[rowSums(is.na(win_methods)) == 0,
    ]
win_methods <- select(win_methods, c("win_by"))
win_methods <- win_methods$win_by
win_methods[0:21]
```

```
##  [1] "KO/TKO"               "Decision - Unanimous" "KO/TKO"
##  [4] "Decision - Unanimous" "Decision - Unanimous" "KO/TKO"
##  [7] "KO/TKO"               "Decision - Unanimous" "KO/TKO"
## [10] "Decision - Split"     "KO/TKO"               "Could Not Continue"
## [13] "KO/TKO"               "Could Not Continue"   "Decision - Unanimous"
## [16] "KO/TKO"               "KO/TKO"               "Decision - Unanimous"
## [19] "Decision - Split"     "KO/TKO"               "Submission"
```

```
# provjera za razlog pobjede
for (name in names(win_methods)) {
    if (sum(is.na(win_methods[, name])) > 0) {
        cat("Broj nedostajućih vrijednosti za varijablu ", name,
            ": ", sum(is.na(win_methods[, name])), "\n")
    }
}
```

Pošto nema ispisa možemo zaključiti da su svi nezadovoljavajući objekti uklonjeni.

Sada možemo nastaviti raditi sa podatcima.

Pretpostavke

Kod provođenja testa moramo provijeriti zadovoljavaju li naši podatci preduvjete potrebne za provođenje izabrane metode. Pošto smo za odgovor na ovo pitanje odlučili provesti jednofaktorsku analizu varijance, preduvjeti su sljedeći:

Test se provodi na podatcima koji su slučajno izabrani iz populacije Koristi se pretpostavka da su podatci na kojima se provode testovi slučajno izabrani iz populacije

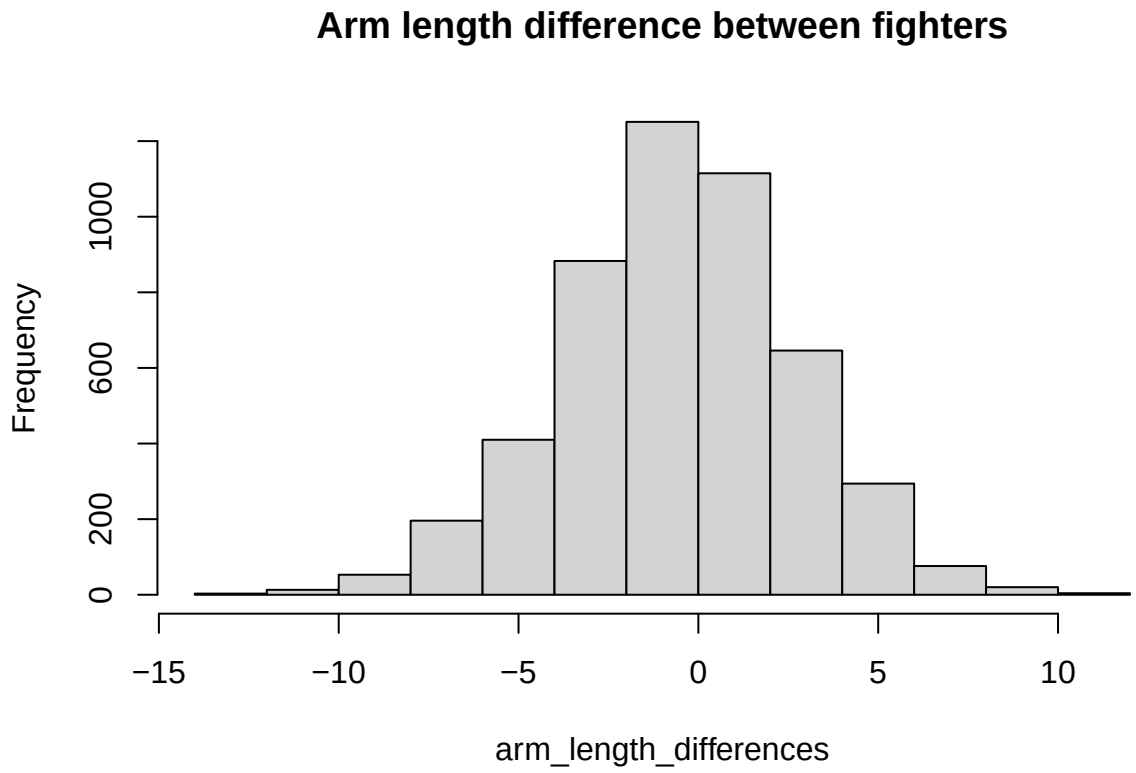
Zavisna kvantitativna varijabla mora pratiti normalnu distribuciju Apsolutna razlika dužine ruku boraca (*B_Reach* - *R_Reach*) je u ovom slučaju zavisna kvantitativna varijabla, te ćemo provjeriti zadovoljava li pretpostavku:

```
arm_length_differences <- (arm_lengths$B_Reach - arm_lengths$R_Reach)
arm_length_differences[0:76]
```

```
##  [1] -3  2  0 -2  4 -3  5  4 -5 -2  2 -2  2 -1  5 -1 -5 -2  0 -5  3 -2  0  5  4
## [26]  2  2  2  3  0  3  2  2  7 -6  1  1  3 -2  0 -2  3  0  2 -2  2 -4  0 -2 -1
## [51]  2  1 -1  0  1  2  2  0  1  3 -3 -5 -3  1 -4  2  7 -8  1  3 -8 -3  0 -2 -2
## [76] -5
```

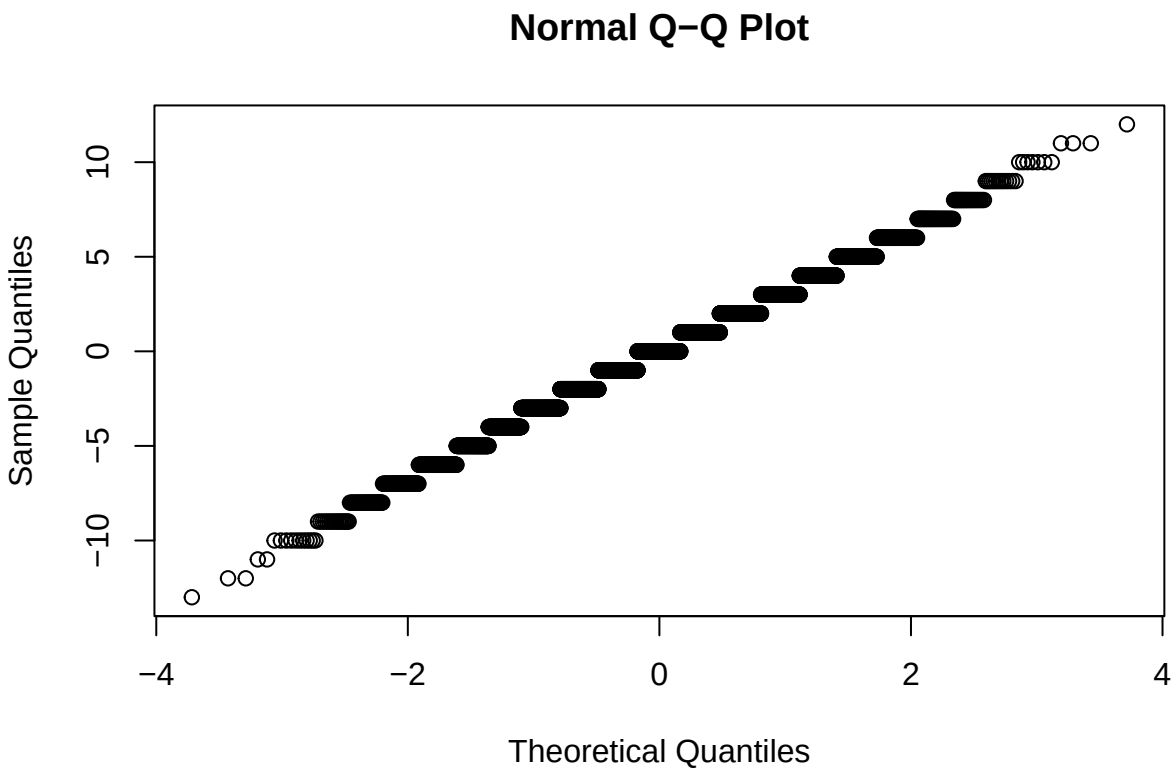
Možemo uočiti da su razlike duljina ruku cijeli brojevi, što je pomalo neočekivano, no razlog tomu je što su početne vrijednosti bile u mjernoj jedinici *inch*, te će tako ostati do završnog rezultata kako bi se očuvala točnost podataka. Pošto je sada sve u redu, možemo početi sa ispitivanjem normalnosti ove varijable. Jedan od najlakših i najefikasnijih načina provjere normalnosti je histogram.

```
hist(arm_length_differences, main = "Arm length difference between fighters")
```



Iz histograma možemo zaključiti da su podatci relativno lijepo raspoređeni po normalnoj distribuciji. Kao još jedan dijagnostički alat, možemo napraviti i Q-Q plot.

```
qqnorm(arm_length_differences)
```



Kao što možemo vidjeti, Q-Q plot također daje poprilično dobre indikacije da se radi o normalnoj distribuciji. Kao konačnu potvrdu naše pretpostavke, koristiti ćemo Lillieforsov test:

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(arm_length_differences)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  arm_length_differences
## D = 0.073975, p-value < 2.2e-16
```

```
lillie.test(sample(arm_length_differences, size = 100))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  sample(arm_length_differences, size = 100)
## D = 0.11653, p-value = 0.001919
```

Prema rezultatima Lillieforsovog testa možemo zaključiti da podatci nisu normalno distribuirani, obrnuto od naše pretpostavke, no moramo uzeti u obzir da niti Lillieforsov test ne daje sigurnost u distribuciju podataka, pošto što je skup podataka veći, to je test osjetljiviji na normalnost istih podataka. Stoga ćemo se oslanjati na robusnost ANOVA-e na normalnost pri grupama iste veličine.

Homogenost varijance u svim grupama (homoskedastičnost) Homogenost varijance kroz grupe inicijalno možemo provjeriti boxplotom:

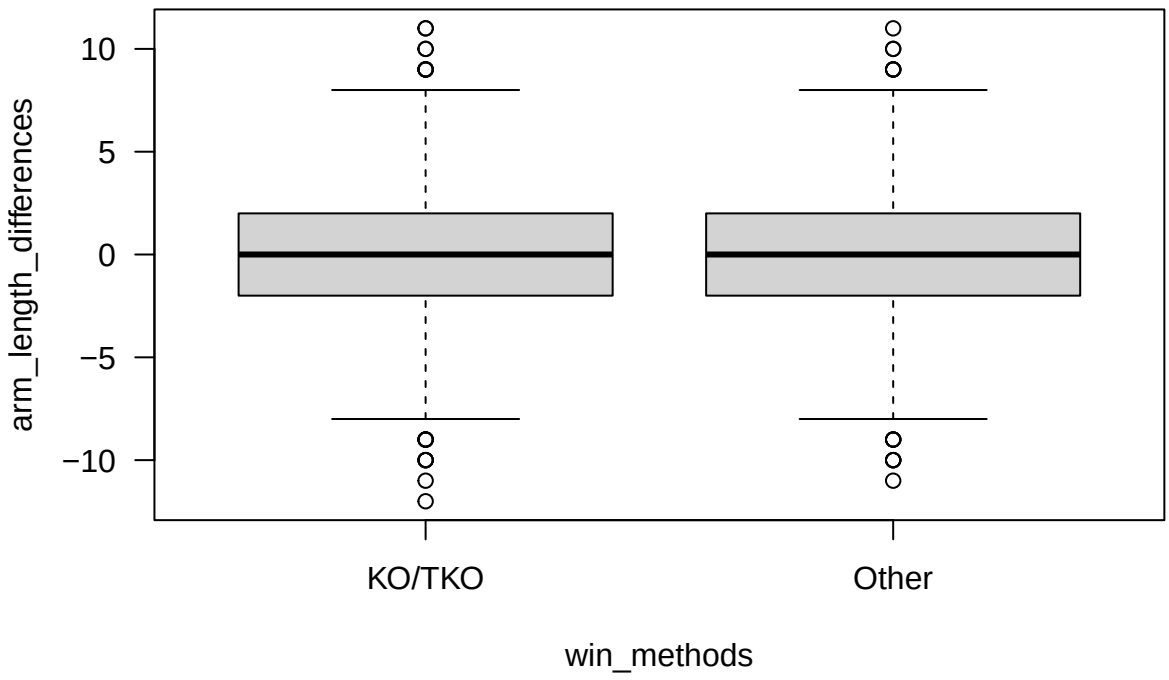
```
question_1_data <- data.frame(arm_length_differences, win_methods)
# stvaranje dviju grupa u nezavisnoj kategorijskoj
# varijabli
question_1_data$win_methods <- replace(question_1_data$win_methods,
    question_1_data$win_methods != "KO/TKO", "Other")
# length(question_1_data$win_methods[question_1_data$win_methods
# == 'KO/TKO'])
# length(question_1_data$win_methods[question_1_data$win_methods
# != 'KO/TKO']) nasumično uzimanje uzorka kako bi se
# stvorile grupe iste veličine za reproducibilnost
set.seed(8919)
other <- sample_n(tbl = question_1_data[question_1_data$win_methods !=
    "KO/TKO", ], size = 1533)
KO <- question_1_data[question_1_data$win_methods == "KO/TKO",
    ]

# grupe iste veličine
question_1_data <- rbind(KO, other)

question_1_data[0:10, ]
```

```
##    arm_length_differences win_methods
## 1                      -3      KO/TKO
## 3                       0      KO/TKO
## 6                      -3      KO/TKO
## 7                       5      KO/TKO
## 9                      -5      KO/TKO
## 11                      2      KO/TKO
## 13                      2      KO/TKO
## 16                     -1      KO/TKO
## 17                     -5      KO/TKO
## 20                     -5      KO/TKO
```

```
boxplot(arm_length_differences ~ win_methods, data = question_1_data,
    las = 1)
```



Iz boxplota možemo zaključiti da varijanca gotovo ne odstupa ovisno o razlogu pobjede, stoga smatramo da je i ovaj preduvjet za provedbu analize varijance zadovoljen. Također, možemo vidjeti kako se srednje vrijednosti ovisno o vrijednosti kategorijske varijable skoro pa ne razlikuju, što nam govori da vjerojatno nećemo imati dovoljno dokaza za odbacivanje hipoteze H0 u korist hipoteze H1.

Kao potvrdu o hipotezi za jednakost varijance napraviti ćemo Bartlettov test:

```
# H0: varijance svih grupa su iste H1: varijance barem
# dviju grupa su različite

bartlett.test(question_1_data$arm_length_differences ~ question_1_data$win_methods)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  question_1_data$arm_length_differences by question_1_data$win_methods
## Bartlett's K-squared = 1.2915, df = 1, p-value = 0.2558
```

```
var((question_1_data$arm_length_differences[question_1_data$win_methods ==
    "KO/TKO"]))
```

```
## [1] 10.96634
```

```
var((question_1_data$arm_length_differences[question_1_data$win_methods !=
    "KO/TKO"]))
```

```
## [1] 10.34751
```

U vidu dobivenih rezultata iz Bartlettovog testa ne možemo odbaciti hipotezu H0 u korist hipoteze H1 na razini signifikantnosti 0.05, drugim riječima - imamo indikacije da su varijance grupa jednake, te s time da je

preduvjet homoskedastičnosti zadovoljen.

Pošto su svi preduvjeti zadovoljeni, možemo početi sa provođenjem testiranje.

Testiranje

Provodimo jednofaktorski ANOVA test:

Hipoteze

Hipoteza H0: srednje vrijednosti obiju grupa su iste

```
# contrast_matrix <- matrix(nrow = 2, ncol = 9,
# c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
# 0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
# 1,1,1,1,1,-8,1,1,1, 0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
# 0,0,0,0,0,0,0,0,0)) colnames(contrast_matrix) <- c('Could
# Not Continue', 'Decision - Majority', 'Decision - Split',
# 'Decision - Unanimous', 'DQ', 'KO/TKO', 'Overturned',
# 'Submission', 'TKO - Doctor's Stoppage')
contrast_matrix <- matrix(nrow = 2, ncol = 1, c(0, 1))
colnames(contrast_matrix) <- c("KO/TKO")
question_1_data$win_methods <- factor(question_1_data$win_methods)
contrasts(question_1_data$win_methods) <- contrast_matrix
contrasts(question_1_data$win_methods)
```

Hipoteza H1: srednje vrijednost grupa su različite

```
##        KO/TKO
## KO/TKO      0
## Other       1
```

```
result <- aov(question_1_data$arm_length_differences ~ question_1_data$win_methods)
summary(result)
```

```
##                               Df Sum Sq Mean Sq F value Pr(>F)
## question_1_data$win_methods    1      0   0.471   0.044  0.834
## Residuals                   3064  32653  10.657
```

```
model <- lm(question_1_data$arm_length_differences ~ question_1_data$win_methods)
summary(model)
```

```
##
## Call:
## lm(formula = question_1_data$arm_length_differences ~ question_1_data$win_methods)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8565  -1.8813   0.1187   2.1187  11.1435
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.14351    0.08338  -1.721   0.0853 .
## question_1_data$win_methodsKO/TKO  0.02479    0.11791   0.210   0.8335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.264 on 3064 degrees of freedom
## Multiple R-squared:  1.442e-05,  Adjusted R-squared:  -0.0003119
## F-statistic: 0.04419 on 1 and 3064 DF,  p-value: 0.8335
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: question_1_data$arm_length_differences
##                               Df Sum Sq Mean Sq F value Pr(>F)
## question_1_data$win_methods    1      0   0.471  0.0442 0.8335
## Residuals                   3064  32653  10.657
```

Zaključak

U vidu dobivenih dokaza, ne možemo odbaciti hipotezu H0 u korist hipoteze H1 na razini signifikantnosti 0.05.

Odgovor na početno pitanje bi glasio:
"Ne, ne možemo očekivati završetak borbe nokautom ovisno o razlici duljine ruku boraca"

Pitanje 2:

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

Prije nego što možemo odgovoriti na to pitanje, trebamo izvući sve podatke iz dataseta u odgovarajućem formatu.

```
# adding total time fought in seconds
library(stringr)
last_round_vector = stringr::str_split_fixed(fights$last_round_time,
    pattern = ":", 2)

fights <- cbind(fights, total_time_in_seconds = (fights$last_round -
    1) * 5 * 60)
for (i in 1:length(fights$total_time_in_seconds)) {
    fights$total_time_in_seconds[i] = fights$total_time_in_seconds[i] +
        strtoi(last_round_vector[i], base = 10L) * 60 + strtoi((last_round_vector[i,
        ])[2], base = 10L)
}
# hist(fights$total_time_in_seconds)

# dividing fights by categories

beginFights <- fights %>%
    filter(grepl(paste("Women's|Strawweight|Flyweight|Bantamweight|Featherweight",
        "|Lightweight|Welterweight|Middleweight|Light Heavyweight|Heavyweight",
        sep = ""), Fight_type, ignore.case = TRUE))
beginFights <- cbind(beginFights, categories = gsub(paste("UFC | Title| Bout|",
    " Tournament|[0-9] |Interim |Super |Ultimate Fighter |[0-9]|Australia |",
    "Brazil |Latin America |Ultimate Japan|vs. UK |UF Nations Canada vs. |T",
    sep = ""), "", beginFights$Fight_type))
beginFights$categories <- trimws(beginFights$categories)

# all categories
validFights = beginFights
```

```
validFights$categories = factor(validFights$categories, levels = c("Women's Strawweight",
    "Women's Flyweight", "Women's Bantamweight", "Women's Featherweight",
    "Flyweight", "Bantamweight", "Featherweight", "Lightweight",
    "Welterweight", "Middleweight", "Light Heavyweight", "Heavyweight"),
    labels = c("womensStrawweight", "womensFlyweight", "womensBantamweight",
        "womensFeatherweight", "Flyweight", "Bantamweight", "Featherweight",
        "Lightweight", "Welterweight", "Middleweight", "LightHeavyweight",
        "Heavyweight"))
```

Raspodjela podataka po kategorijama

```
summary(validFights$categories)
```

```
##   womensStrawweight     womensFlyweight  womensBantamweight womensFeatherweight
##                 175                 103                 141                  16
##           Flyweight        Bantamweight       Featherweight         Lightweight
##                 208                 431                 478                1014
##        Welterweight        Middleweight    LightHeavyweight         Heavyweight
##                 984                 751                 526                 541
```

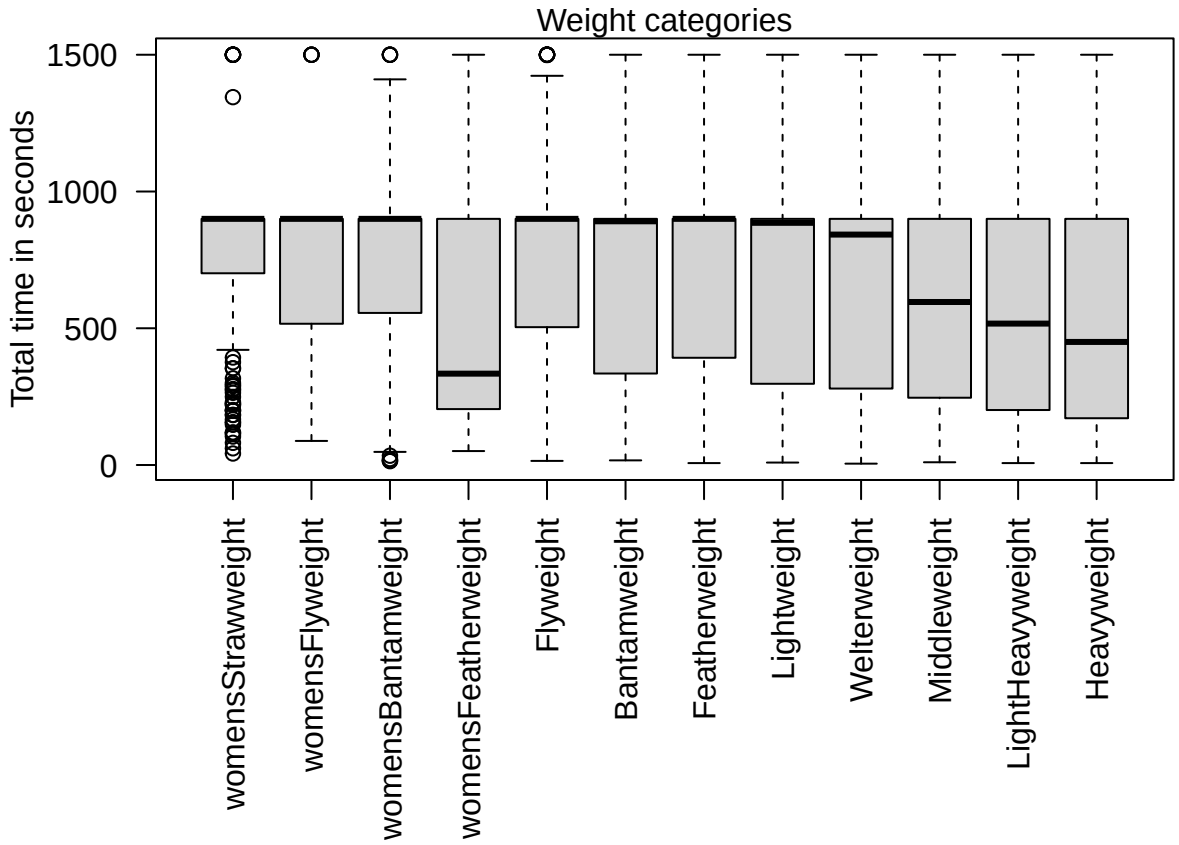
Sažetak trajanja borbi u sekundama

```
summary(validFights$total_time_in_seconds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0   273.0   782.0   626.4   900.0  1500.0
```

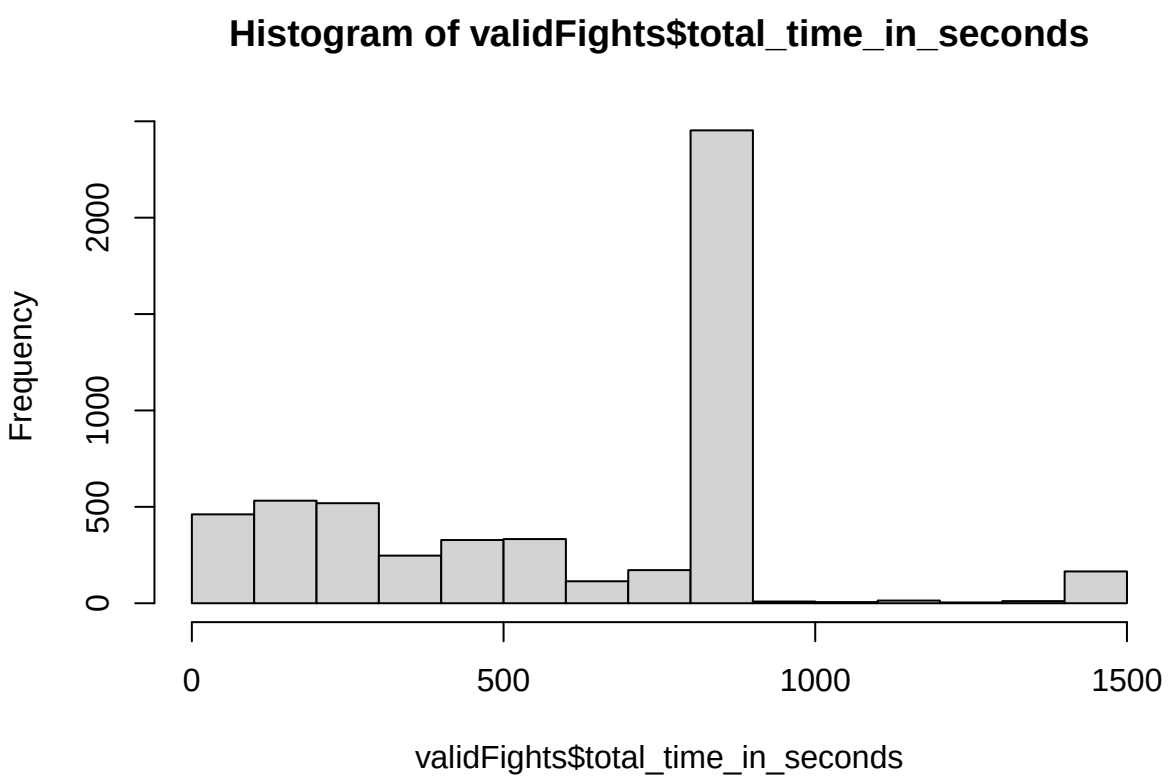
Možemo prikazati podatke u boxplotu i usporediti ih:

```
par(mar = c(10, 5, 1, 1))
boxplot(validFights$total_time_in_seconds ~ validFights$categories,
    range = 1.5, ylab = "Total time in seconds", xlab = "", las = 2)
mtext("Weight categories", side = 3)
```



Testiranje normalnosti ćemo provesti nad svim podacima, i podacima podijeljenim u grupe prema kategoriji.

```
hist(validFights$total_time_in_seconds)
```




```
require(nortest)
lillie.test(validFights$total_time_in_seconds)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  validFights$total_time_in_seconds
## D = 0.24339, p-value < 2.2e-16
```

Iz histograma svih podataka i Lillieforsovog testa nad tim podacima možemo zaključiti da podaci ne prate normalnu razdiobu, ali da budemo sigurni možemo testirati i svaki zasebnu grupu podataka. Radi sažetosti, ispitivali smo samo ako je p-vrijednost bilo koje grupe veća od $\alpha = 0.05$.

```
categoryLevels = levels(validFights$categories)
normalnostGrupe = FALSE
for (i in 1:length(categoryLevels)) {
    pVrijednost = lillie.test(validFights$total_time_in_seconds[validFights$categories ==
        categoryLevels[i]])$p.value
    if (pVrijednost >= 0.05)
        normalnostGrupe = TRUE
}

if (normalnostGrupe) {
    print(noquote("Barem jedna grupa prati normalnu razdiobu."))
} else {
    print(noquote("Nijedna grupa ne prati normalnu razdiobu."))
}
```

```
## [1] Nijedna grupa ne prati normalnu razdiobu.
```

Homogenost varijanci različitih populacija ćemo testirati Levenovim testom, jer je manje osjetljiv na odstupanje od normalnosti nego Bartlettov test.

```
library(car)
```

```
leveneTest(total_time_in_seconds ~ categories, validFights)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   11  9.7432 < 2.2e-16 ***
##       5356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podaci ne zadovoljavaju ni uvjet normalnosti ni homogenosti varijanci. Kad su veličine grupa podjednake, ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti i homogenosti varijanci ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti i homogenosti varijanci ako su veličine grupa podjednake pa možemo uezti uzorak iz svake grupe nad kojim ćemo testirati podatke. Iz podataka vidimo da je kategorija Women's Featherweight jedan red veličine manja od drugih grupa pa nju izbacujemo iz testiranja.

```
adjustedValidFights <- beginFights %>%
    filter(!grepl("Women's Featherweight", Fight_type, ignore.case = TRUE))

adjustedValidFights$categories = factor(adjustedValidFights$categories,
    levels = c("Women's Strawweight", "Women's Flyweight", "Women's Bantamweight",
        "Flyweight", "Bantamweight", "Featherweight", "Lightweight",
        "Welterweight", "Middleweight", "Light Heavyweight",
        "Heavyweight"), labels = c("womensStrawweight", "womensFlyweight",
        "womensBantamweight", "Flyweight", "Bantamweight", "Featherweight",
        "Lightweight", "Welterweight", "Middleweight", "LightHeavyweight",
        "Heavyweight"))

adjustedValidFights <- adjustedValidFights %>%
    group_by(categories) %>%
    slice_sample(n = 100)

a = aov(adjustedValidFights$total_time_in_seconds ~ adjustedValidFights$categories)
summary(a)
```

```
##                                 Df    Sum Sq Mean Sq F value   Pr(>F)
## adjustedValidFights$categories   10   7474162  747416   6.077 4.81e-09 ***
## Residuals                      1089 133942889  122996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Možemo odbaciti H0 u korist H1, time zaključujemo da postoje barem dvije grupe kojima se trajanje mečeva razlikuje.

Pitanje 3:

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

Radimo tablicu u kojoj je trajanje samo borbe za naslov prvaka kategorije. Takve borbe mogu trajati najviše 5 rundi.

```
bouts = select(fights, c("Fight_type", "last_round"))
# names(bouts)

titleFights <- fights %>%
    filter(grepl("title", Fight_type, ignore.case = TRUE))
# titleFights

head(select(titleFights, R_fighter, B_fighter, Fight_type, last_round,
    Winner))
```

```
##              R_fighter         B_fighter                            Fight_type
## 1             Petr Yan Aljamain Sterling          UFC Bantamweight Title Bout
## 2       Jan Blachowicz   Israel Adesanya     UFC Light Heavyweight Title Bout
## 3         Amanda Nunes    Megan Anderson UFC Women's Featherweight Title Bout
## 4         Kamaru Usman     Gilbert Burns          UFC Welterweight Title Bout
## 5  Deiveson Figueiredo    Brandon Moreno             UFC Flyweight Title Bout
## 6 Valentina Shevchenko     Jennifer Maia     UFC Women's Flyweight Title Bout
##   last_round               Winner
## 1          4    Aljamain Sterling
## 2          5       Jan Blachowicz
## 3          1         Amanda Nunes
## 4          3         Kamaru Usman
## 5          5
## 6          5 Valentina Shevchenko
```

```
titleFightDuration <- titleFights$last_round  #lista u koju spremamo trajanja svih title fightova
```

U ispisu vidimo da neki title fightovi nemaju pobjednika. Ako je borba završila u 5. rundi i nema pobjednika riječ je o Drawu, odnosno odluka sudaca je da su borci izjednačeni. Ako je borba završila prije 5. runde i nema pobjednika, što je iznimno rijetko, mora biti riječ o No Contestu, odnosno borba je prekinuta iz nekog razloga i nemoguće je odrediti pobjednika (npr. jednom je protivniku nanesen nenamjeran faul zbog kojeg ne može nastaviti borbu)

Računamo varijancu trajanja borbi koje su za naslov prvaka kategorije.

```
var(titleFightDuration)
```

```
## [1] 2.651129
```

Na isti način filtriramo samo borbe koje nisu za titulu šampiona. Te borbe uvijek traju najviše 3 runde, osim ako je riječ o borbi koja je main event. Tada borba može trajati do 5 rundi. Takvih je slučajeva jako malo u usporedbi s non-title borbama koje traju do 3 runde

```
nonTitleFights <- fights %>%
    filter(!grepl("title", Fight_type, ignore.case = TRUE))
# nonTitleFights

head(select(nonTitleFights, R_fighter, B_fighter, Fight_type,
    last_round, Winner))
```

```
##         R_fighter         B_fighter                Fight_type last_round
## 1    Adrian Yanez     Gustavo Lopez         Bantamweight Bout          3
## 2    Trevin Giles     Roman Dolidze         Middleweight Bout          3
## 3     Tai Tuivasa   Harry Hunsucker          Heavyweight Bout          1
## 4   Cheyanne Buys Montserrat Conejo  Women's Strawweight Bout          3
## 5   Marion Reneau     Macy Chiasson Women's Bantamweight Bout          3
## 6 Leonardo Santos      Grant Dawson          Lightweight Bout          3
```

```
##              Winner
## 1      Adrian Yanez
## 2      Trevin Giles
## 3       Tai Tuivasa
## 4 Montserrat Conejo
## 5     Macy Chiasson
## 6      Grant Dawson
```

```
nonTitleFightDuration <- nonTitleFights$last_round
```

Računamo varijancu trajanja borbi koje nisu za naslov prvaka kategorije.

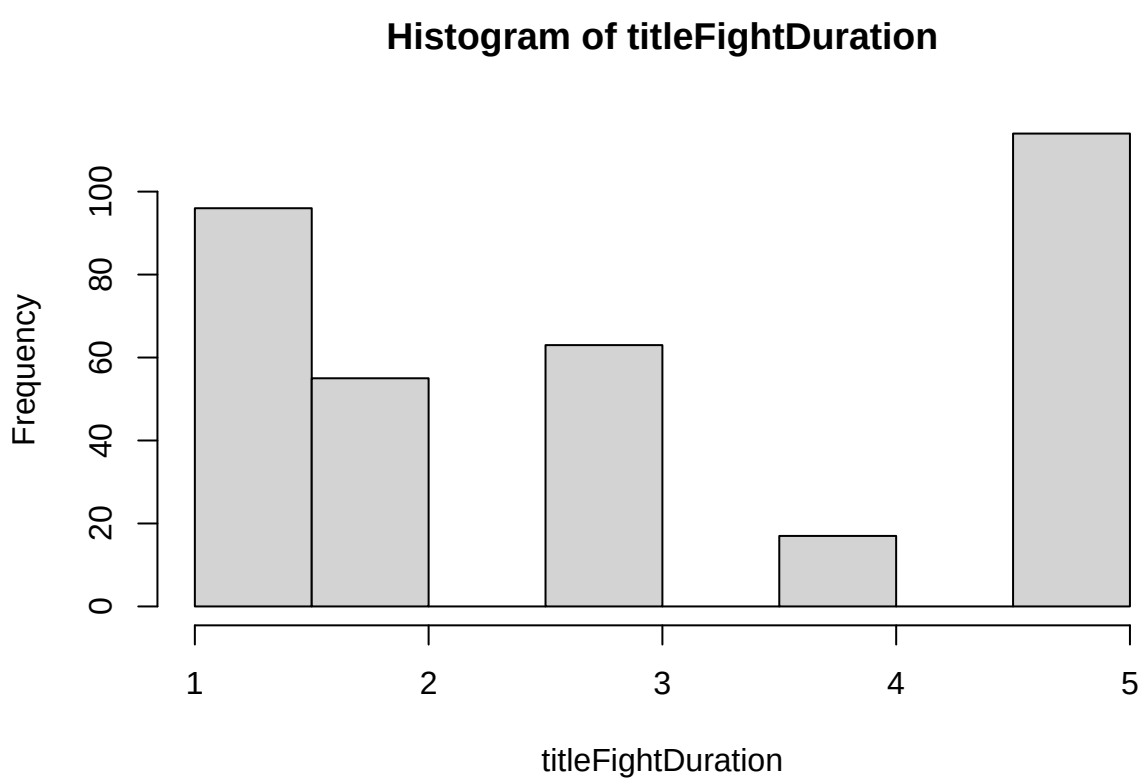
```
var(nonTitleFightDuration)
```

```
## [1] 0.8879308
```

Vidimo da borbe koje nisu za titulu šampiona imaju puno manju varijancu, što je očekivano jer većina takvih borbi može trajati maksimalno 3 runde.

Za početak radimo Q-Q plot trajanja borbi za naslov šampiona kako bismo ustvrdili radi li se o normalnoj distribuciji

```
hTitleFights <- hist(titleFightDuration)
```

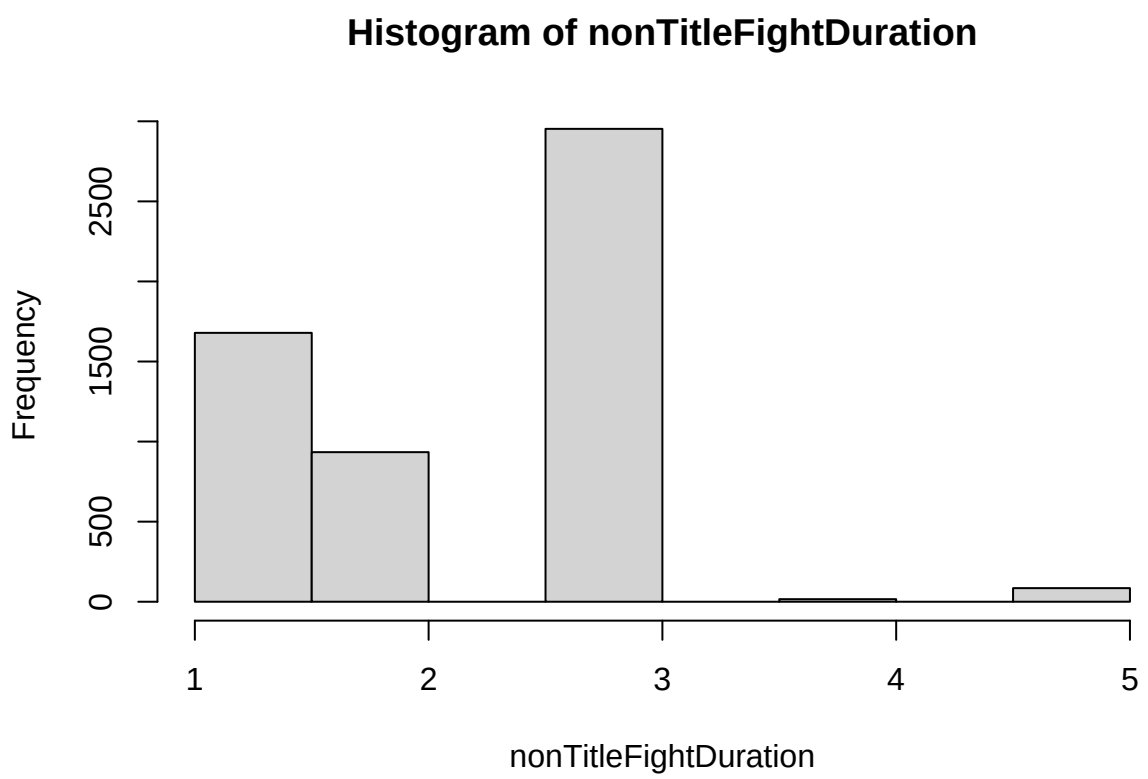



```
hTitleFights
```

```
## $breaks
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
##
## $counts
```

```
## [1]  96  55   0  63   0  17   0 114
##
## $density
## [1] 0.55652174 0.31884058 0.00000000 0.36521739 0.00000000 0.09855072 0.00000000
## [8] 0.66086957
##
## $mids
## [1] 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75
##
## $xname
## [1] "titleFightDuration"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
hNonTitleFights <- hist(nonTitleFightDuration)
```

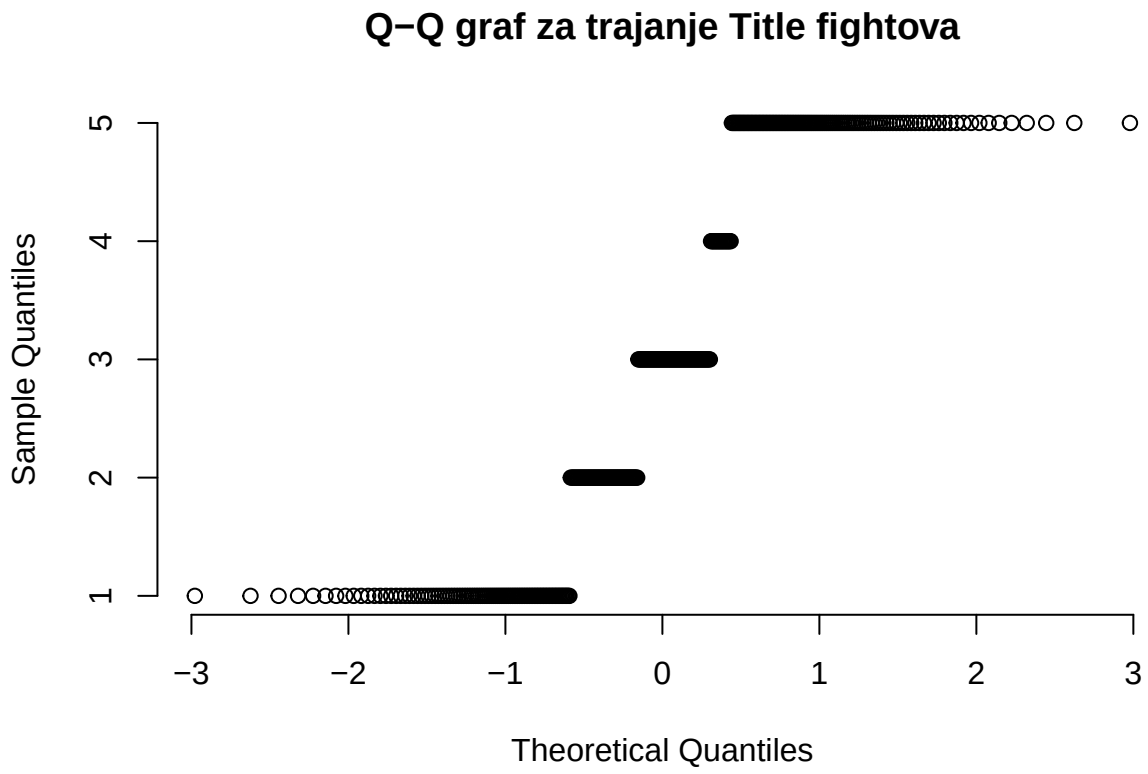



```
hNonTitleFights
```

```
## $breaks
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
##
## $counts
## [1] 1679  934    0 2953    0   16    0   85
```

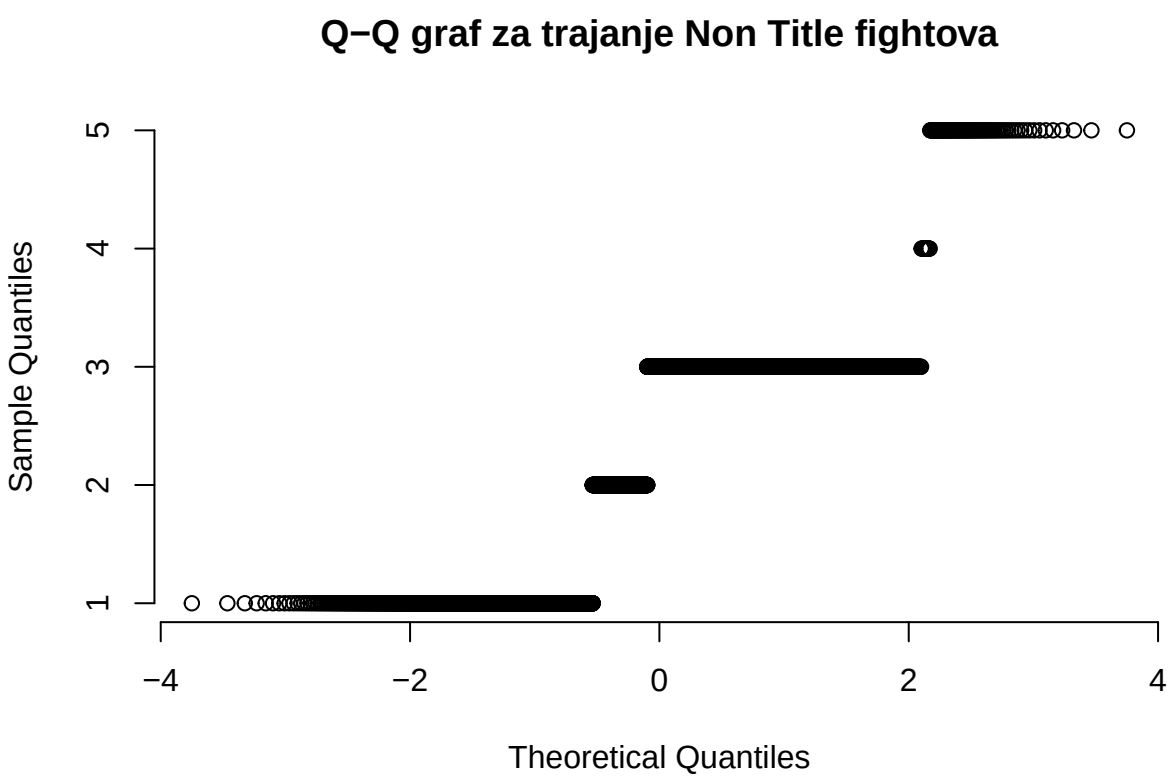
```
##
## $density
## [1] 0.592553379 0.329627669 0.000000000 1.042173990 0.000000000 0.005646727
## [7] 0.000000000 0.029998235
##
## $mids
## [1] 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75
##
## $xname
## [1] "nonTitleFightDuration"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
qqnorm(titleFightDuration, pch = 1, frame = FALSE, main = "Q-Q graf za trajanje Title fightova")
```



Vidimo da graf nimalo ne upućuje da se su trajanja borbi za titulu normalno raspodjeljena. Napravimo isti test za borbe koje nisu za naslov šampiona.

```
qqnorm(nonTitleFightDuration, pch = 1, frame = FALSE, main = "Q-Q graf za trajanje Non Title fightova")
```



Iz prethodnih grafova možemo zaključiti da trajanje niti jedne vrste borbi nema normalnu razdiobu. Pogledajmo sada kako izgledaju razdiobe na stupčastim dijagramima

```
# TO DO: dodaj bar plot sa razdiobom trajanja title fightsa
# -> koristi ggplot paket!!!
length(titleFightDuration[titleFightDuration == 5])
```

```
## [1] 114
```

```
# dodaj bar plot sa razdiobom trajanja non title fightsa
```

Još jednom vidimo da ni u jednom slučaju nije riječ o normalnoj razdiobi. Još ćemo jednom prikazati iste podatke, ali sada prikazujemo postotak svih Title fightova i Non Title fightova koji su završili u nekoj rundi, a ne konkretan broj.

```
hTitleFights$counts[hTitleFights$counts != 0]
```

```
## [1]  96  55  63  17 114
```

```
postotciZavrsetakaURundiTitle <- hTitleFights$counts[hTitleFights$counts !=
    0]/sum(hTitleFights$counts[hTitleFights$counts != 0])
postotciZavrsetakaURundiTitle
```

```
## [1] 0.27826087 0.15942029 0.18260870 0.04927536 0.33043478
```

```
hNonTitleFights$counts[hNonTitleFights$counts != 0]
```

```
## [1] 1679  934 2953   16   85
```

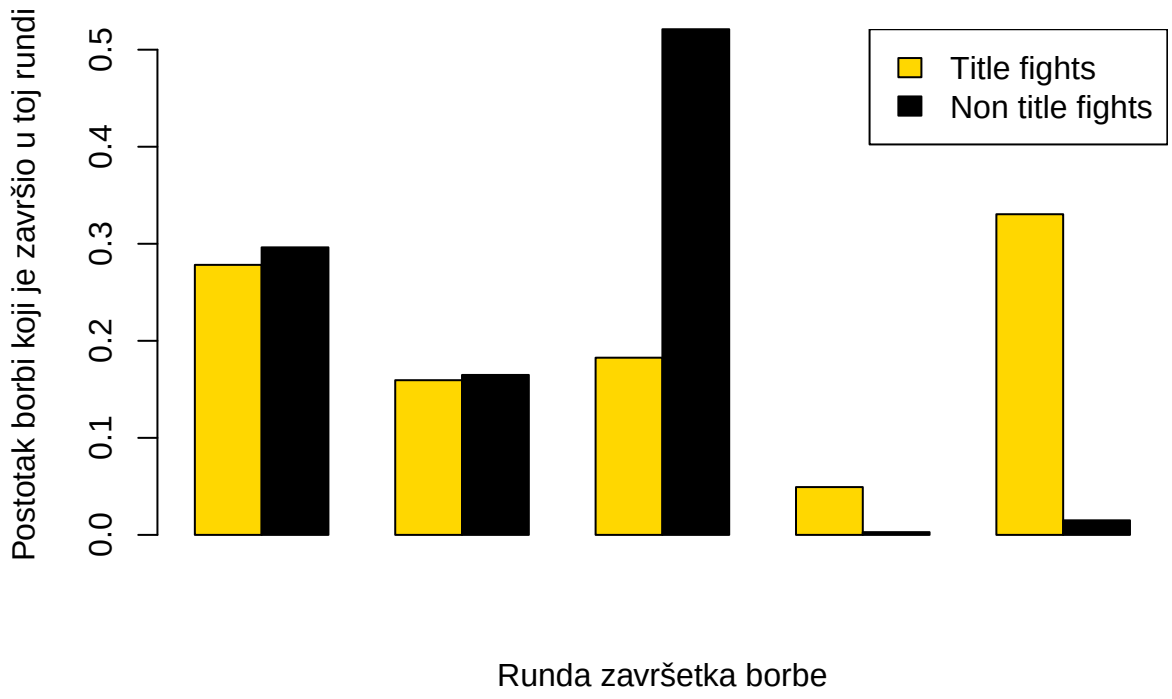
```
postotciZavrsetakaURundiNormal <- hNonTitleFights$counts[hNonTitleFights$counts !=
    0]/sum(hNonTitleFights$counts[hNonTitleFights$counts != 0])
postotciZavrsetakaURundiNormal
```

```
## [1] 0.296276690 0.164813834 0.521086995 0.002823363 0.014999118
```

```
df <- t(cbind(postotciZavrsetakaURundiTitle, postotciZavrsetakaURundiNormal))



barplot(df, beside = TRUE, xlab = "Runda završetka borbe", ylab = "Postotak borbi koji je završio u toj
    col = c("gold", "black"))
legend("topright", c("Title fights", "Non title fights"), fill = c("gold",
    "black"))
```



I opet smo se uvjerili da nije ni blizu riječ o normalnoj distribuciji.

Dakle, koristimo Wilcoxon Rank-Sum Test jer je on za razliku od T-testa ne pretpostavlja normalnost.

Ali pogledajmo svejedno kakav bi rezultat dobili da koristimo T-test.

T-test H0: titleFightDuration = nonTitleFightDuration H1: titleFightDuration > nonTitleFightDuration

```
t.test(titleFightDuration, nonTitleFightDuration, alt = "greater",
    var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  titleFightDuration and nonTitleFightDuration
## t = 8.1169, df = 358.16, p-value = 3.882e-15
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5727189       Inf
## sample estimates:
## mean of x mean of y
##  2.994203  2.275454
```

Iz T-testa bi odbacili početnu hipotezu H0 da borbe za titulu šampiona u prosjeku traju jednako kao obične borbe koje nisu za titulu šampiona jer je p-vrijednost jako jako mala.

Pogledajmo sada što možemo zaključiti iz Wilcoxon Rank-Sum Testa.

```
# napravi Wilcoxon Rank-Sum Test
```

Pitati

Naslučujući da imamo podatke o cijeloj populaciji možemo napraviti Z-test pošto možemo izračunati varijancu i standardnu devijaciju, u

slučaju da nemamo pristup cijeloj populaciji, korisitit ćemo T-test za dvije populacije sa poznatim varijancama

U oba slučaja moramo provijeriti normalnost distribucije varijabli sa histogramima i Q-Q plotovima. Vjerojatno će biti potrebno odstraniti
neke podatke pošto postoje ekstremi. Run a independency test to prove independence.

Pitanje 4:

Mogu li dostupne značajke predvidjeti pobjednika? U ovom zadatku koristimo logističku regresiju jer nas zanima ako se određeni borac sam po svojim karakteristikama može svrstati u pobjetnika ili gubitnika. Kako je rezultat (zavisna varijabla) binaran, definirat ćemo ga kao Winner = 0/1;

Nakon stvaranja modela sa svim atributima vezanih uz borca i njegova protivnika lako možemo razlučiti signifikantne i nesignifikantne regresore. Ako varijabla ima p-vrijednost veću od 0.05 ne smatramo je signifikantnom za predviđanje pobjednika.

```
# install.packages('caret')
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: ggplot2
```

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

```
# pretvaranje winner stringova 'Blue' i 'Red' tou bool
# vrijednosti kako bi se logisticka regresija mogla
# koristiti
RBData$Winner[which(RBData$Winner == "Blue")] <- "1"
RBData$Winner[which(RBData$Winner == "Red")] <- "0"
RBData$Winner <- as.numeric(RBData$Winner)

logreg.mdl = glm(formula = Winner ~ B_avg_KD + B_avg_opp_KD +
    B_avg_SIG_STR_pct + B_avg_opp_SIG_STR_pct + B_avg_TD_pct +
    B_avg_opp_TD_pct + B_avg_SUB_ATT + B_avg_opp_SUB_ATT + B_avg_REV +
    B_avg_opp_REV + B_avg_SIG_STR_att + B_avg_SIG_STR_landed +
```

```
    B_avg_opp_SIG_STR_att + B_avg_opp_SIG_STR_landed + B_avg_TOTAL_STR_att +
    B_avg_TOTAL_STR_landed + B_avg_opp_TOTAL_STR_att + B_avg_opp_TOTAL_STR_landed +
    B_avg_TD_att + B_avg_TD_landed + B_avg_opp_TD_att + B_avg_opp_TD_landed +
    B_avg_HEAD_att + B_avg_HEAD_landed + B_avg_opp_HEAD_att +
    B_avg_opp_HEAD_landed + B_avg_BODY_att + B_avg_BODY_landed +
    B_avg_opp_BODY_att + B_avg_opp_BODY_landed + B_avg_DISTANCE_att +
    B_avg_DISTANCE_landed + B_avg_opp_DISTANCE_att + B_avg_opp_DISTANCE_landed +
    B_avg_CLINCH_att + B_avg_CLINCH_landed + B_avg_opp_CLINCH_att +
    B_avg_opp_CLINCH_landed + B_total_rounds_fought + B_total_title_bouts +
    B_current_win_streak + B_current_lose_streak + B_longest_win_streak +
    B_wins + B_losses + B_Height_cms + R_Height_cms + B_Reach_cms +
    R_Reach_cms + B_avg_CONT_time_seconds + B_avg_opp_CONT_time_seconds +
    B_total_time_fought_seconds, data = RBData, family = binomial)

summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = Winner ~ B_avg_KD + B_avg_opp_KD + B_avg_SIG_STR_pct +
##     B_avg_opp_SIG_STR_pct + B_avg_TD_pct + B_avg_opp_TD_pct +
##     B_avg_SUB_ATT + B_avg_opp_SUB_ATT + B_avg_REV + B_avg_opp_REV +
##     B_avg_SIG_STR_att + B_avg_SIG_STR_landed + B_avg_opp_SIG_STR_att +
##     B_avg_opp_SIG_STR_landed + B_avg_TOTAL_STR_att + B_avg_TOTAL_STR_landed +
##     B_avg_opp_TOTAL_STR_att + B_avg_opp_TOTAL_STR_landed + B_avg_TD_att +
##     B_avg_TD_landed + B_avg_opp_TD_att + B_avg_opp_TD_landed +
##     B_avg_HEAD_att + B_avg_HEAD_landed + B_avg_opp_HEAD_att +
##     B_avg_opp_HEAD_landed + B_avg_BODY_att + B_avg_BODY_landed +
##     B_avg_opp_BODY_att + B_avg_opp_BODY_landed + B_avg_DISTANCE_att +
##     B_avg_DISTANCE_landed + B_avg_opp_DISTANCE_att + B_avg_opp_DISTANCE_landed +
##     B_avg_CLINCH_att + B_avg_CLINCH_landed + B_avg_opp_CLINCH_att +
##     B_avg_opp_CLINCH_landed + B_total_rounds_fought + B_total_title_bouts +
##     B_current_win_streak + B_current_lose_streak + B_longest_win_streak +
##     B_wins + B_losses + B_Height_cms + R_Height_cms + B_Reach_cms +
##     R_Reach_cms + B_avg_CONT_time_seconds + B_avg_opp_CONT_time_seconds +
##     B_total_time_fought_seconds, family = binomial, data = RBData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8297  -0.8835  -0.7265   1.2510   2.4882
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.778e-01  7.428e-01   0.374 0.708436
## B_avg_KD                    1.606e-02  9.479e-02   0.169 0.865426
## B_avg_opp_KD               -1.693e-01  1.139e-01  -1.486 0.137257
## B_avg_SIG_STR_pct           2.218e-01  3.656e-01   0.607 0.544162
## B_avg_opp_SIG_STR_pct       7.449e-02  3.492e-01   0.213 0.831098
## B_avg_TD_pct               -4.061e-01  1.722e-01  -2.358 0.018370 *
## B_avg_opp_TD_pct           -7.051e-01  1.850e-01  -3.812 0.000138 ***
## B_avg_SUB_ATT              -5.674e-02  5.266e-02  -1.078 0.281251
## B_avg_opp_SUB_ATT          -1.419e-01  5.997e-02  -2.367 0.017925 *
## B_avg_REV                  -2.045e-02  1.146e-01  -0.179 0.858327
## B_avg_opp_REV              -3.770e-02  1.200e-01  -0.314 0.753412
## B_avg_SIG_STR_att           4.543e-02  3.343e-02   1.359 0.174248
```

```
## B_avg_SIG_STR_landed        -5.478e-02  4.315e-02  -1.270 0.204232
## B_avg_opp_SIG_STR_att       -1.318e-01  3.715e-02  -3.547 0.000390 ***
## B_avg_opp_SIG_STR_landed     1.417e-01  4.700e-02   3.016 0.002565 **
## B_avg_TOTAL_STR_att          2.380e-03  1.434e-02   0.166 0.868214
## B_avg_TOTAL_STR_landed      -1.198e-02  1.612e-02  -0.743 0.457406
## B_avg_opp_TOTAL_STR_att      6.865e-02  1.640e-02   4.186 2.83e-05 ***
## B_avg_opp_TOTAL_STR_landed  -8.442e-02  1.854e-02  -4.554 5.27e-06 ***
## B_avg_TD_att                 5.522e-02  2.271e-02   2.432 0.015032 *
## B_avg_TD_landed              1.234e-01  5.809e-02   2.125 0.033627 *
## B_avg_opp_TD_att             1.796e-03  2.192e-02   0.082 0.934687
## B_avg_opp_TD_landed          1.458e-01  5.726e-02   2.547 0.010878 *
## B_avg_HEAD_att              -5.107e-02  2.578e-02  -1.981 0.047598 *
## B_avg_HEAD_landed            7.862e-02  3.290e-02   2.390 0.016870 *
## B_avg_opp_HEAD_att           1.055e-02  2.739e-02   0.385 0.700008
## B_avg_opp_HEAD_landed       -4.091e-03  3.424e-02  -0.119 0.904883
## B_avg_BODY_att              -2.850e-02  3.108e-02  -0.917 0.359126
## B_avg_BODY_landed            5.482e-02  3.954e-02   1.386 0.165603
## B_avg_opp_BODY_att           6.484e-02  3.422e-02   1.895 0.058101 .
## B_avg_opp_BODY_landed       -6.645e-02  4.227e-02  -1.572 0.115955
## B_avg_DISTANCE_att           3.157e-03  1.524e-02   0.207 0.835899
## B_avg_DISTANCE_landed        3.988e-03  2.270e-02   0.176 0.860551
## B_avg_opp_DISTANCE_att       4.841e-02  1.630e-02   2.969 0.002983 **
## B_avg_opp_DISTANCE_landed   -5.016e-02  2.372e-02  -2.114 0.034473 *
## B_avg_CLINCH_att             3.467e-03  2.460e-02   0.141 0.887924
## B_avg_CLINCH_landed         -5.401e-03  3.429e-02  -0.158 0.874821
## B_avg_opp_CLINCH_att         3.464e-03  2.731e-02   0.127 0.899056
## B_avg_opp_CLINCH_landed     -4.425e-03  3.770e-02  -0.117 0.906563
## B_total_rounds_fought        2.590e-03  1.405e-02   0.184 0.853720
## B_total_title_bouts         -1.937e-01  4.262e-02  -4.546 5.48e-06 ***
## B_current_win_streak         2.799e-02  3.343e-02   0.837 0.402473
## B_current_lose_streak        4.172e-02  6.095e-02   0.684 0.493680
## B_longest_win_streak         2.639e-02  4.496e-02   0.587 0.557131
## B_wins                      -2.438e-02  4.279e-02  -0.570 0.568919
## B_losses                     4.642e-02  4.689e-02   0.990 0.322192
## B_Height_cms                -3.443e-02  7.754e-03  -4.440 9.00e-06 ***
## R_Height_cms                 1.376e-03  7.709e-03   0.178 0.858360
## B_Reach_cms                  4.813e-02  6.366e-03   7.561 3.99e-14 ***
## R_Reach_cms                 -2.312e-02  6.303e-03  -3.668 0.000245 ***
## B_avg_CONT_time_seconds      5.945e-04  5.765e-04   1.031 0.302393
## B_avg_opp_CONT_time_seconds  1.490e-03  5.873e-04   2.536 0.011208 *
## B_total_time_fought_seconds  9.313e-05  3.343e-04   0.279 0.780539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7450.4  on 5901  degrees of freedom
## Residual deviance: 7034.2  on 5849  degrees of freedom
## AIC: 7140.2
##
## Number of Fisher Scoring iterations: 4
```

Usporedba modela

Sa signifikantnijim regresorima napravit ćemo samnjeni model te testirati razliku njegove i devijance početnog modela. Ukoliko devijanca nije značajno veća prihvatit ćemo smanjeni model.

```
logreg.mdl.2 = glm(Winner ~ B_avg_opp_TD_pct + B_avg_opp_SIG_STR_att +
    B_avg_opp_TOTAL_STR_att + B_avg_opp_TOTAL_STR_landed + B_total_title_bouts +
    B_Height_cms + B_Reach_cms + B_avg_opp_SIG_STR_landed + B_avg_opp_DISTANCE_att +
    R_Reach_cms + B_avg_TD_pct + B_avg_opp_SUB_ATT + B_avg_TD_att +
    B_avg_TD_landed + B_avg_opp_TD_landed + B_avg_HEAD_landed +
    B_avg_opp_DISTANCE_landed + B_avg_opp_CONT_time_seconds +
    B_avg_HEAD_att, data = RBData, family = binomial())

summary(logreg.mdl.2)
```

```
##
## Call:
## glm(formula = Winner ~ B_avg_opp_TD_pct + B_avg_opp_SIG_STR_att +
##     B_avg_opp_TOTAL_STR_att + B_avg_opp_TOTAL_STR_landed + B_total_title_bouts +
##     B_Height_cms + B_Reach_cms + B_avg_opp_SIG_STR_landed + B_avg_opp_DISTANCE_att +
##     R_Reach_cms + B_avg_TD_pct + B_avg_opp_SUB_ATT + B_avg_TD_att +
##     B_avg_TD_landed + B_avg_opp_TD_landed + B_avg_HEAD_landed +
##     B_avg_opp_DISTANCE_landed + B_avg_opp_CONT_time_seconds +
##     B_avg_HEAD_att, family = binomial(), data = RBData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9571  -0.8913  -0.7441   1.2869   2.2540
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.5081454  0.6567205   0.774 0.439071
## B_avg_opp_TD_pct            -0.7113861  0.1674422  -4.249 2.15e-05 ***
## B_avg_opp_SIG_STR_att       -0.1195069  0.0204092  -5.856 4.75e-09 ***
## B_avg_opp_TOTAL_STR_att      0.0592250  0.0149780   3.954 7.68e-05 ***
## B_avg_opp_TOTAL_STR_landed  -0.0756411  0.0170313  -4.441 8.94e-06 ***
## B_total_title_bouts         -0.1559999  0.0347453  -4.490 7.13e-06 ***
## B_Height_cms                -0.0369082  0.0075253  -4.905 9.36e-07 ***
## B_Reach_cms                  0.0514692  0.0062504   8.235  < 2e-16 ***
## B_avg_opp_SIG_STR_landed     0.1386154  0.0252026   5.500 3.80e-08 ***
## B_avg_opp_DISTANCE_att       0.0618663  0.0115417   5.360 8.31e-08 ***
## R_Reach_cms                 -0.0229279  0.0039385  -5.821 5.84e-09 ***
## B_avg_TD_pct                -0.4520491  0.1664566  -2.716 0.006613 **
## B_avg_opp_SUB_ATT           -0.1207523  0.0558293  -2.163 0.030550 *
## B_avg_TD_att                 0.0445217  0.0210792   2.112 0.034676 *
## B_avg_TD_landed              0.1246005  0.0537836   2.317 0.020520 *
## B_avg_opp_TD_landed          0.1558475  0.0430320   3.622 0.000293 ***
## B_avg_HEAD_landed            0.0100845  0.0048285   2.089 0.036750 *
## B_avg_opp_DISTANCE_landed   -0.0628479  0.0170008  -3.697 0.000218 ***
## B_avg_opp_CONT_time_seconds  0.0014172  0.0004674   3.032 0.002430 **
## B_avg_HEAD_att               0.0008153  0.0022118   0.369 0.712399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 7450.4  on 5901  degrees of freedom
## Residual deviance: 7097.1  on 5882  degrees of freedom
## AIC: 7137.1
##
## Number of Fisher Scoring iterations: 4
```

```
anova(logreg.mdl, logreg.mdl.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Winner ~ B_avg_KD + B_avg_opp_KD + B_avg_SIG_STR_pct + B_avg_opp_SIG_STR_pct +
##     B_avg_TD_pct + B_avg_opp_TD_pct + B_avg_SUB_ATT + B_avg_opp_SUB_ATT +
##     B_avg_REV + B_avg_opp_REV + B_avg_SIG_STR_att + B_avg_SIG_STR_landed +
##     B_avg_opp_SIG_STR_att + B_avg_opp_SIG_STR_landed + B_avg_TOTAL_STR_att +
##     B_avg_TOTAL_STR_landed + B_avg_opp_TOTAL_STR_att + B_avg_opp_TOTAL_STR_landed +
##     B_avg_TD_att + B_avg_TD_landed + B_avg_opp_TD_att + B_avg_opp_TD_landed +
##     B_avg_HEAD_att + B_avg_HEAD_landed + B_avg_opp_HEAD_att +
##     B_avg_opp_HEAD_landed + B_avg_BODY_att + B_avg_BODY_landed +
##     B_avg_opp_BODY_att + B_avg_opp_BODY_landed + B_avg_DISTANCE_att +
##     B_avg_DISTANCE_landed + B_avg_opp_DISTANCE_att + B_avg_opp_DISTANCE_landed +
##     B_avg_CLINCH_att + B_avg_CLINCH_landed + B_avg_opp_CLINCH_att +
##     B_avg_opp_CLINCH_landed + B_total_rounds_fought + B_total_title_bouts +
##     B_current_win_streak + B_current_lose_streak + B_longest_win_streak +
##     B_wins + B_losses + B_Height_cms + R_Height_cms + B_Reach_cms +
##     R_Reach_cms + B_avg_CONT_time_seconds + B_avg_opp_CONT_time_seconds +
##     B_total_time_fought_seconds
## Model 2: Winner ~ B_avg_opp_TD_pct + B_avg_opp_SIG_STR_att + B_avg_opp_TOTAL_STR_att +
##     B_avg_opp_TOTAL_STR_landed + B_total_title_bouts + B_Height_cms +
##     B_Reach_cms + B_avg_opp_SIG_STR_landed + B_avg_opp_DISTANCE_att +
##     R_Reach_cms + B_avg_TD_pct + B_avg_opp_SUB_ATT + B_avg_TD_att +
##     B_avg_TD_landed + B_avg_opp_TD_landed + B_avg_HEAD_landed +
##     B_avg_opp_DISTANCE_landed + B_avg_opp_CONT_time_seconds +
##     B_avg_HEAD_att
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      5849     7034.2
## 2      5882     7097.1 -33  -62.801 0.001331 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Razlika u devijancama modela je značajna, tako da nećemo moći koristiti ovaj smanjeni model kao substituciju za primarni. (p - vrijednost = 0.001331)

Izračunajmo matricu zabune za originalni model

Stvarno/Opaženo (pobjeda) Y=0 Y=1 Y=0 TN FP Y=1 FN TP

```
yHat <- logreg.mdl$fitted.values > 0.5
tab <- table(RBData$Winner, yHat)

tab
```

```
##    yHat
##     FALSE TRUE
##   0  3711  268
##   1  1614  309
```

```r
accuracy = sum(diag(tab))/sum(tab)
precision = tab[2, 2]/sum(tab[, 2])
recall = tab[2, 2]/sum(tab[2, ])
specificity = tab[1, 1]/sum(tab[, 1])

print(paste0("Točnost: ", accuracy))  # Udio točno pretostavljenih pobjednika i gubitnika  (TP+TN)/(TP+
```

```
## [1] "Točnost: 0.681125042358523"
```

```r
print(paste0("Preciznost: ", precision))  # Udio točno pretostavljenih pobjednika u onima koji su klasi
```

```
## [1] "Preciznost: 0.535528596187175"
```

```r
print(paste0("Odziv: ", recall))  # Udio točno pretostavljenih pobjednika u skupu svih koji su stvarno
```

```
## [1] "Odziv: 0.160686427457098"
```

```r
print(paste0("Specifičnost: ", specificity))  # Udio točno pretostavljenih gubitnika u skupu svih koji
```

```
## [1] "Specifičnost: 0.696901408450704"
```

Zaključak

Po danim značajkama možemo u 68% slučajeva točno pretpostaviti pobjednika i gubitnika. U značajnom broju slučajeva borca koji je pobjedio proglašavamo gubitnikom.

Možemo vidjeti da udio pravih pobjednika u skupu onih za koje smo pretpostavili da će pobjediti 53.5%, te time ne možemo sa sigurnošću zaključiti tko će pobjediti samo na osnovi danih podataka iz prijašnjih borba. Iako ove varijable mogu pripomoći predvidjeti pobjednika, to se predviđanje po danim podatcima borca ne može sa sigurnošću potvrditi.