ATP Data Analysis

2024-01-10

Instalacija potrebnih paketa.

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

Učitavanje biblioteka.

```
library(dplyr)
library(lubridate)
library(ggplot2)
library(caret)
library(nortest)
```

Učitavanje i opis podataka

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

dim(all_matches)
```

```
## [1] 104682     49
```

Skup podataka sadrži informacije o 104682 teniska meča održana od 1991. do 2023. godine uključivo. Svaki meč opisan je s 49 ispod navedenih značajki:

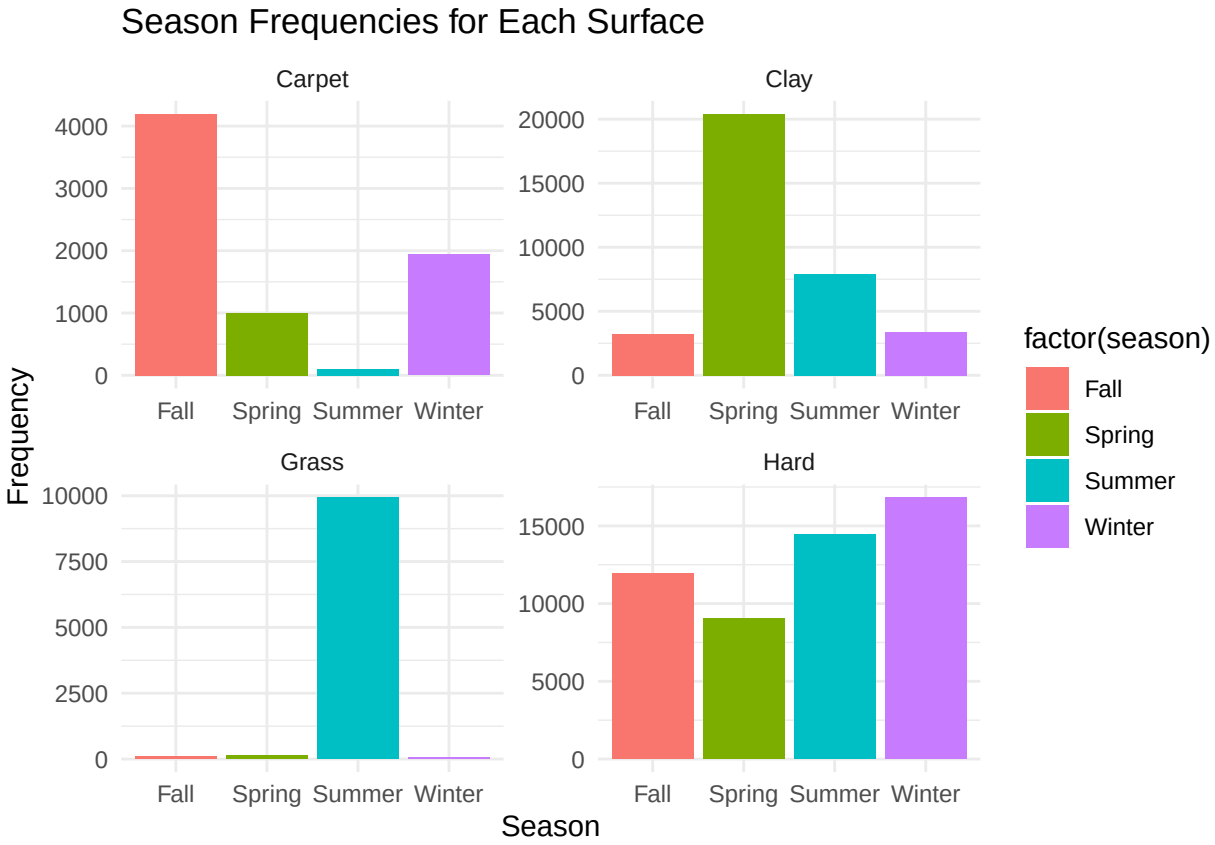
```
names(all_matches)
```

```
##  [1] "tourney_id"         "tourney_name"       "surface"
##  [4] "draw_size"          "tourney_level"      "tourney_date"
##  [7] "match_num"          "winner_id"          "winner_seed"
## [10] "winner_entry"       "winner_name"        "winner_hand"
## [13] "winner_ht"          "winner_ioc"         "winner_age"
## [16] "loser_id"           "loser_seed"         "loser_entry"
## [19] "loser_name"         "loser_hand"         "loser_ht"
## [22] "loser_ioc"          "loser_age"          "score"
## [25] "best_of"            "round"              "minutes"
```

```
## [28] "w_ace"              "w_df"               "w_svpt"
## [31] "w_1stIn"            "w_1stWon"           "w_2ndWon"
## [34] "w_SvGms"            "w_bpSaved"          "w_bpFaced"
## [37] "l_ace"              "l_df"               "l_svpt"
## [40] "l_1stIn"            "l_1stWon"           "l_2ndWon"
## [43] "l_SvGms"            "l_bpSaved"          "l_bpFaced"
## [46] "winner_rank"        "winner_rank_points" "loser_rank"
## [49] "loser_rank_points"
```

TODO Opis ispisa, možda uzet summary samo za neke značajke

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



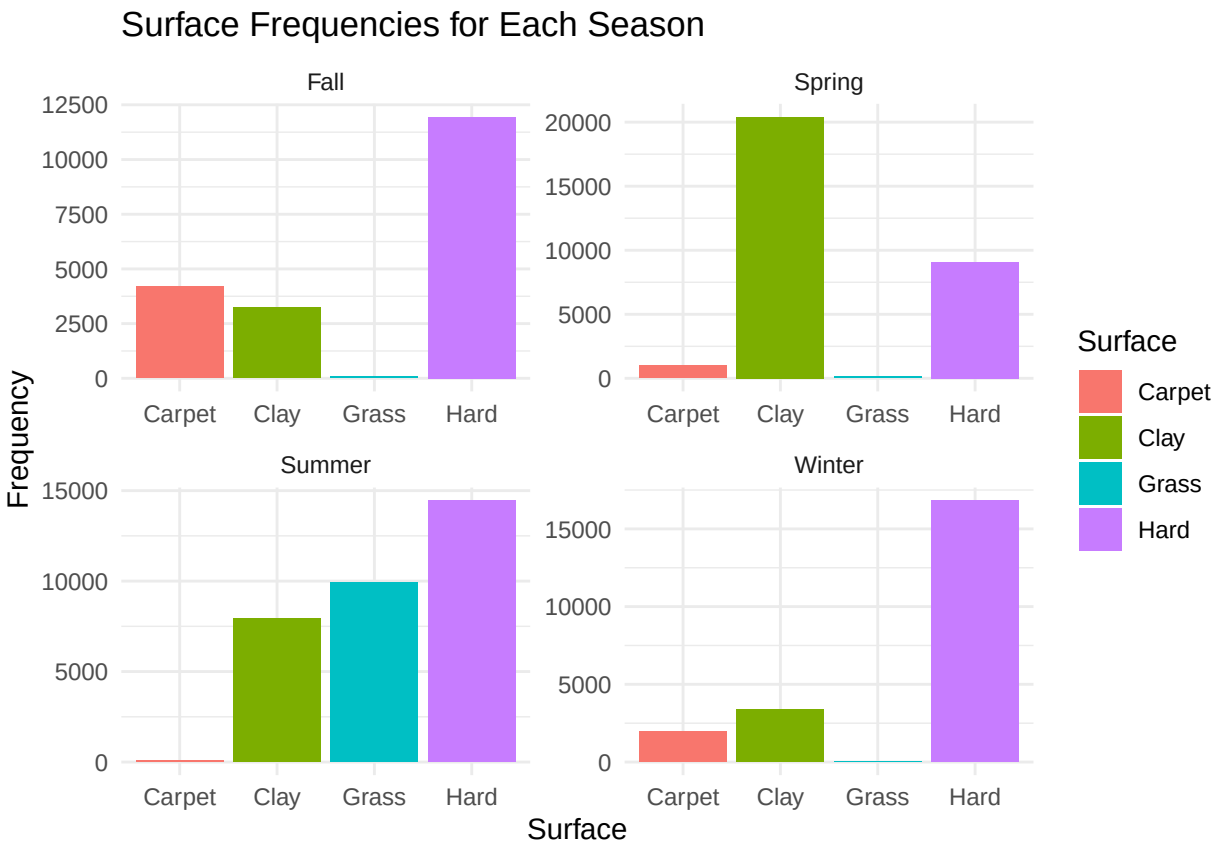
U prvom histogramu prikazana je raspodjela teniskih mečeva prema godišnjim dobima na podlozi od tepiha. Podloga od tepiha najmanje je korištena podloga za igranje mečeva. Najčešće se podloga od tepiha koristila u jesen, dosta rjeđe zimi, zatim na proljeće, a najmanje se mečeva na podlozi od tepiha igra na ljeto.

Sljedeći histogram predstavlja raspodjelu mečeva prema godišnjim dobima na zemljanoj podlozi. Mečevi na zemlji najčešće se igraju u proljetnom dijelu sezone. Dosta manje mečeva igra se na ljeto zatim otprilike podjednako na jesen i zimi.

Treći histogram opisuje distribuciju teniskih mečeva prema godišnjim dobima na travi. Teniski mečevi na travi igraju se uglavnom ljeti, a svega nekoliko mečeva igra se u preostalim godišnjim dobima.

U posljednjem histogramu promatrana je raspodjela mečeva prema godišnjim dobima na tvrdoj podlozi. Sveukupno najviše mečeva igra se na tvrdoj podlozi te je raspodjela prema godišnjim dobima manje izražena

nego kod drugih podloga. Najviše mečeva na tvrdoj podlozi održava se zimi, zatim u ljeto pa na jesen te najmanje u proljetnom dijelu sezone.



Prvi histogram prikazuje raspodjelu mečeva prema podlogama u jesen. Uvjerljivo najviše mečeva u jesen održava se na tvrdoj podlozi. Dosta manje mečeva igra se na podlozi od tepiha, a nešto malo manje na zemlji. Najmanje mečeva u jesenskom dijelu sezone igra se na travi.

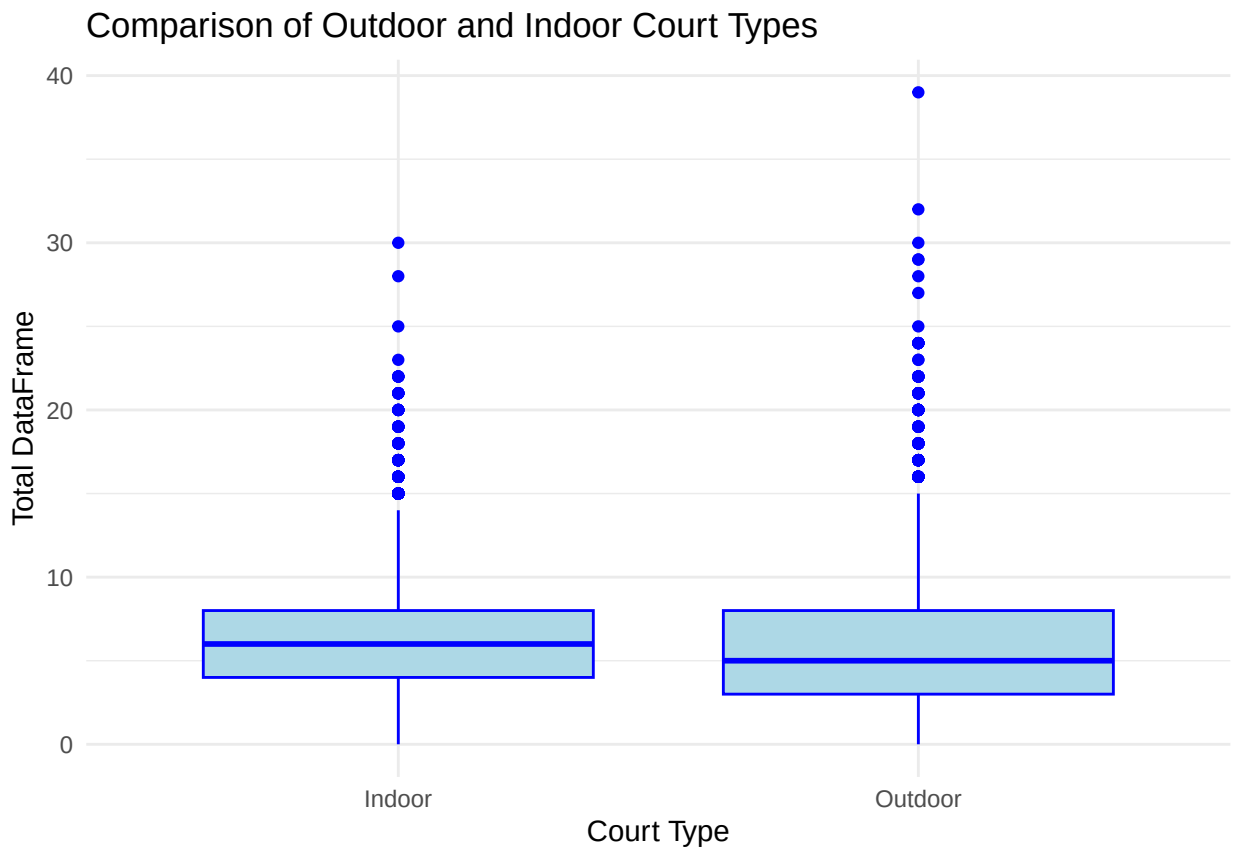
Idući histogram prikazuje raspodjelu mečeva prema podlogama u proljeće. U proljetnom dijelu sezone uvjerljivo najviše teniskih mečeva igra se na podlozi od zemlje. Više od dvostruko manje mečeva održava se na tvrdoj podlozi. Jako malo mečeva održava se na podlozi od tepiha, a još manje na travi.

U trećem histogramu promatramo raspodjelu mečeva prema podlogama tijekom ljeta. Najviše mečeva održava se na tvrdoj podlozi, zatim na travi pa na podlozi od zemlje. Svega nekoliko mečeva igra se na podlozi od tepiha.

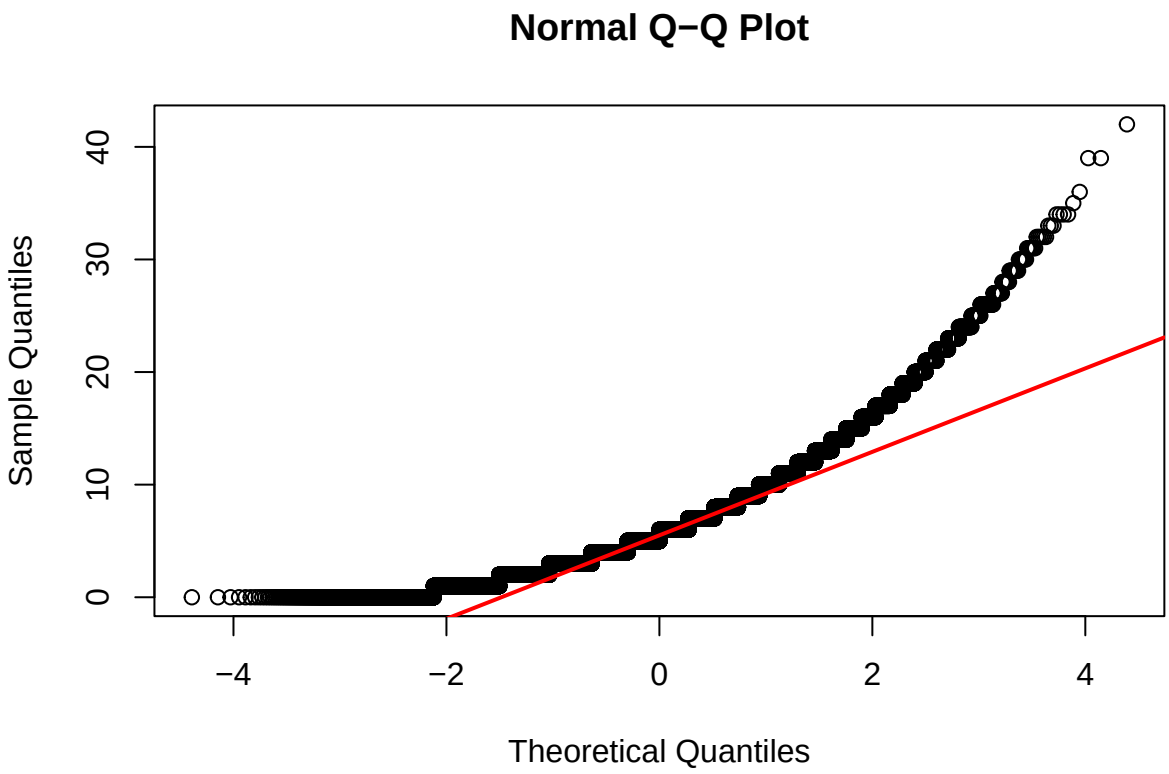
Zadnji histogram opisuje raspodjelu mečeva prema podlogama zimi. Tijekom zime prednjače mečevi na tvrdoj podlozi. Dosta manje mečeva igra se na zemlji, zatim na podlozi od tepiha te najmanje na travi.

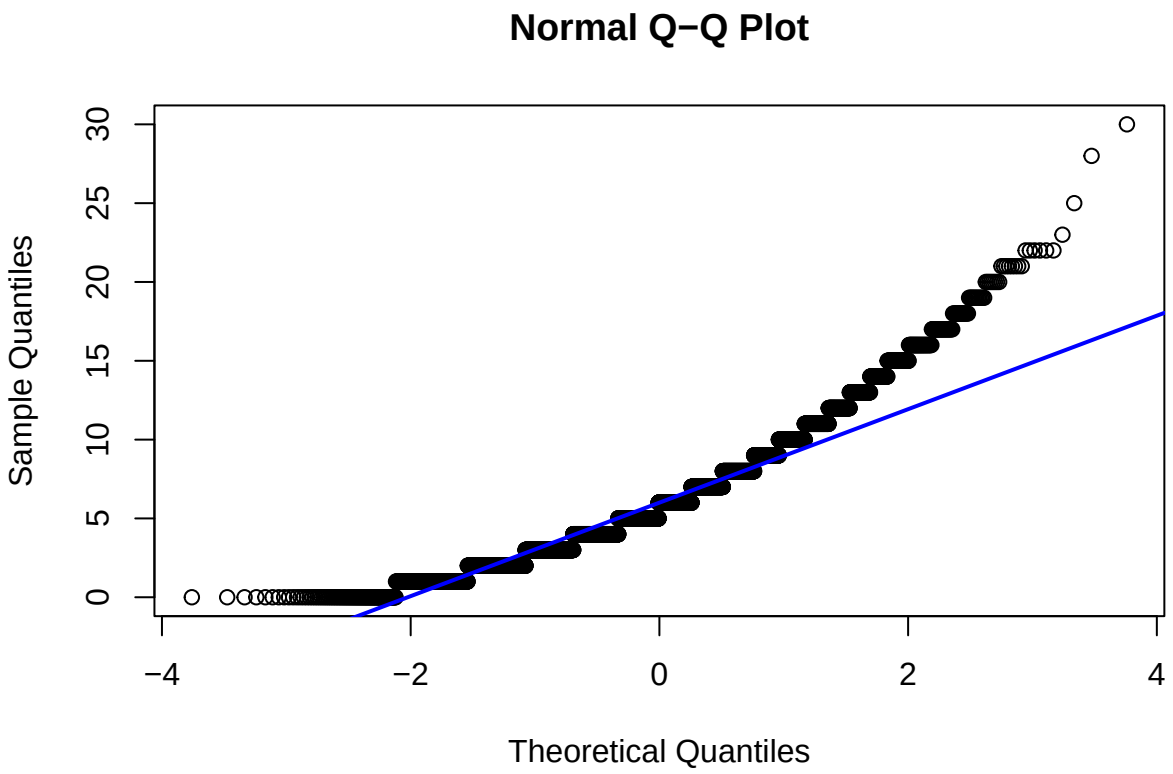
Zadatak 2. Postoji li značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu?

Prvo provjeravamo ukazuje li boxplot za moguću značajnu razliku.



Grafički prikaz ukazuje na moguću razliku između prosječnog broja dvostrukih pogrešaka između mečeva odigranih na otvorenom i zatvorenom. Kako bismo provjerili možemo li prihvatiti nultu hipotezu koja pretpostavlja da nema razlike, provest ćemo t-test. Najprije moramo provjeriti pretpostavke o normalnoj distribuciji i homogenosti varijanci. Normalnu distribuciju prvo provjeravamo pomoću qq-plota, a zatim i Lilliefors testom.


Normal Q−Q Plot
Sample Quantiles
0
10
20
30
40
−4
−2
0
2
4
Theoretical Quantiles



Iz qq-plota vidljivo je da distribucije nisu normalne niti za mečeve na otvorenom niti na zatvorenom jer postoji značajno odstupanje repa. Zatim provodimo Lilliefors test:

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  open_surface_data
## D = 0.12974, p-value < 2.2e-16
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  closed_surface_data
## D = 0.12216, p-value < 2.2e-16
```

Za oba skupa podataka (otvoreni teren i zatvoreni teren), rezultati testova normalnosti (Lilliefors test) pokazuju da podaci nisu normalno distribuirani (p-vrijednosti su manje od 0.05) što se moglo zaključiti i iz grafova. To znači da distribucija podataka odstupa od normalne distribucije.

Provedimo F-test za provjeru homogenosti varijanci:

```
##
##  F test to compare two variances
##
## data:  open_surface_data and closed_surface_data
## F = 1.1441, num df = 88596, denom df = 5877, p-value = 4.316e-12
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.101871 1.187308
## sample estimates:
## ratio of variances
##           1.144146
```

F-test za usporedbu varijanci pokazuje da postoji značajna razlika u varijancama između otvorenog terena i zatvorenog terena (p-vrijednost < 0.05). Kako pretpostavke za t-test nisu zadovoljene koristimo Wilcoxonov test:

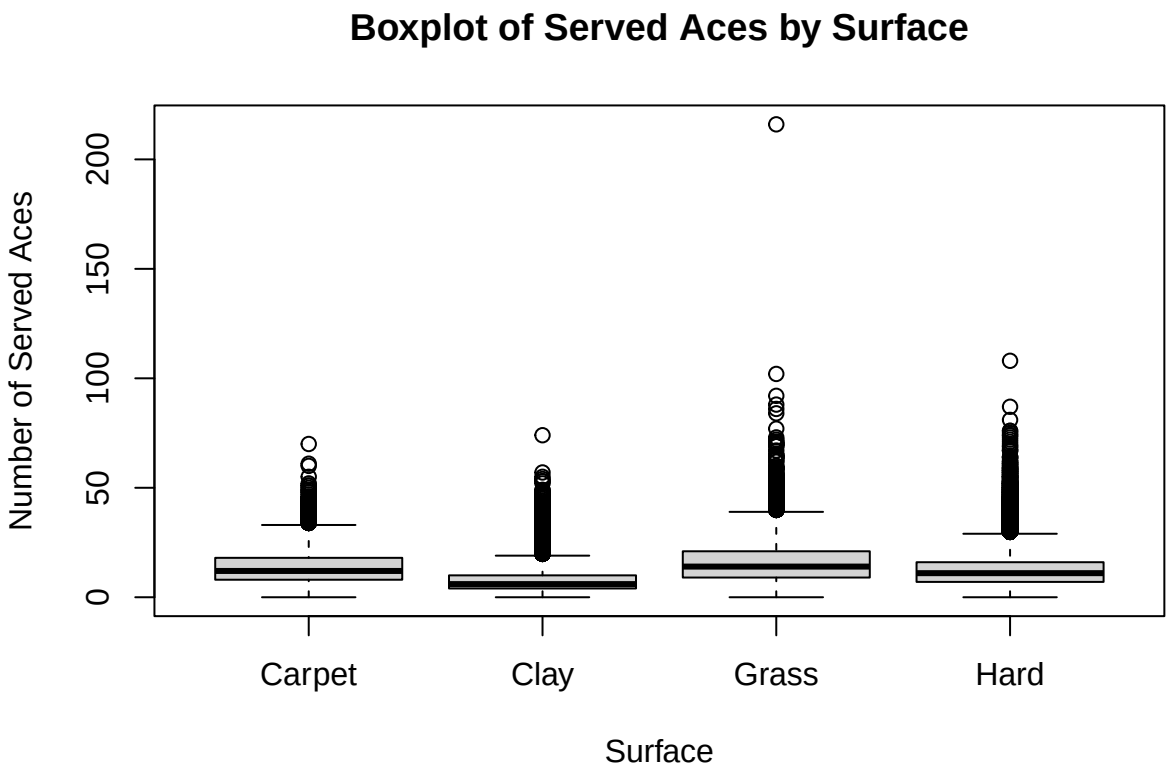
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  open_surface_data and closed_surface_data
## W = 258377269, p-value = 0.3191
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon rang-sum test ne pokazuje značajnu razliku u srednjim vrijednostima (medijanama) između otvorenog i zatvorenog terena (p-vrijednost $= 0.3191$, dakle ne možemo odbaciti nultu hipotezu).

Na temelju ovih rezultata, možemo zaključiti da nema značajne razlike u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom terenu i mečeva odigranih na zatvorenom terenu.

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

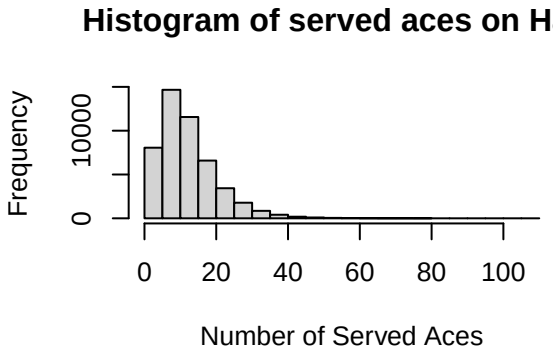
Provjerimo za početak postoje li lako uočljive razlike u broju serviranih asova na različitim podlogama pomoću grafičkog prikaza.

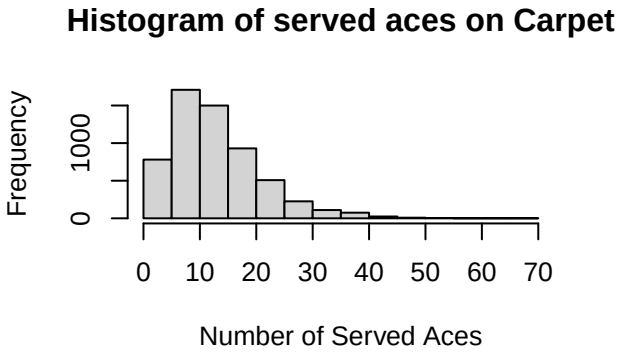


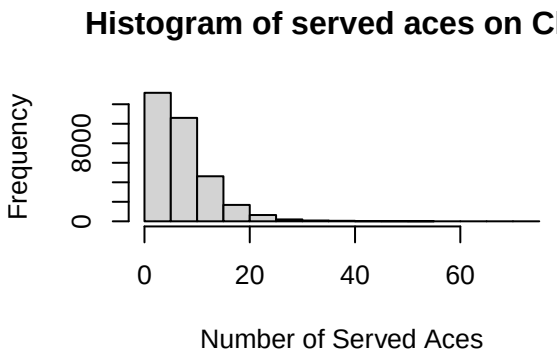
Boxplot ukazuje na to da postoje razlike u broju asova s obzirom na podlogu.

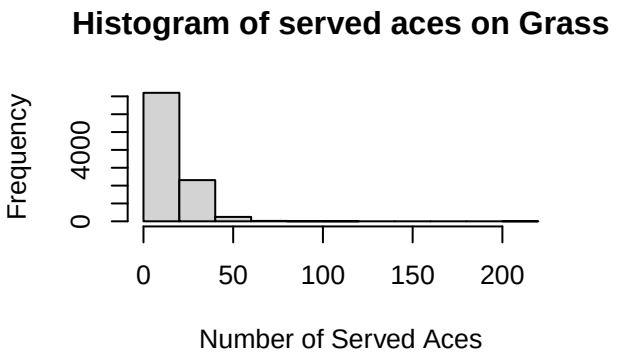
Nulta hipoteza jest da nema razlike u broju serviranih asova na različitim podlogama. Može li se odbaciti ista provjerit ćemo ANOVA testom. ANOVA analizira razliku srednje vrijednosti između više od dvije grupe. Kako bi taj test mogao biti korišten najprije moramo provjeriti pretpostavke:

1. provjera normalne distribucije









Histogrami ukazuju na to da distrubucija serviranih asova nije normalna niti na jednoj od podloga. Normalna distribucija još se provjerava i Lilliefors testom:

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  t3$aces[t3$surface == "Hard"]
## D = 0.11436, p-value < 2.2e-16

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  t3$aces[t3$surface == "Carpet"]
## D = 0.10864, p-value < 2.2e-16

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  t3$aces[t3$surface == "Clay"]
## D = 0.13505, p-value < 2.2e-16

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  t3$aces[t3$surface == "Grass"]
## D = 0.10802, p-value < 2.2e-16
```

Za svaku od 4 podloge p-vrijednost je manja od 0.05 zbog čega odbacujemo pretpostavku da je distribucija normalna.

2. provjera homogenosti varijanci Homogenost varijanci provjerava se Bartletttovim testom:

```
##
##  Bartlett test of homogeneity of variances
##
## data:  t3$aces by t3$surface
## Bartlett's K-squared = 7049.2, df = 3, p-value < 2.2e-16
```

P-vrijednost u Bartlettovom testu manja je od kritične vrijednosti od 0.05 čime se zaključuje da homogenost varijanci nije zadovoljena.

Isto se vidi i u ispisu varijance za svaku od podloga.

```
## [1] 65.28138
```

```
## [1] 63.26659
```

```
## [1] 31.9019
```

```
## [1] 104.5289
```

Kako niti jedna od pretpostavki nije zadovoljena koristit ćemo Kruskal-Wallis, neparametarsku alternativu ANOVA testu.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  aces by surface
## Kruskal-Wallis chi-squared = 13657, df = 3, p-value < 2.2e-16
```

Nakon provedenog testa dobivamo p-vrijednost manju od 0.05 što znači da možemo odbaciti nultu hipotezu u korist prve, odnosno da postoji razlika u broj serviranih asova u odnosu na podlogu, što intuitivno ima smisla jer loptica ne odskače jednako od svih podloga.

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

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

TODO Opis ispisa

Kontingencijskoj tablici dodajemo sume redaka i stupaca:

```
##
##          FALSE  TRUE   Sum
##   Carpet   700   179   879
##   Clay    5550  1240  6790
##   Grass   3471   819  4290
##   Hard    9090  2054 11144
##   Sum    18811  4292 23103
```

TODO Opis ispisa

Pretpostavka testa je da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5 (`chisq.test()` pretpostavlja da je ovaj uvjet zadovoljen stoga je prije provođenja testa potrebno to provjeriti):

```
## Očekivane frekvencije za razred  FALSE - Carpet :  715.7022
## Očekivane frekvencije za razred  FALSE - Clay :  5528.576
## Očekivane frekvencije za razred  FALSE - Grass :  3493.018
## Očekivane frekvencije za razred  FALSE - Hard :  9073.704
## Očekivane frekvencije za razred  TRUE - Carpet :  163.2978
## Očekivane frekvencije za razred  TRUE - Clay :  1261.424
## Očekivane frekvencije za razred  TRUE - Grass :  796.9822
## Očekivane frekvencije za razred  TRUE - Hard :  2070.296
```

Sve očekivane frekvencije su veće od 5, nastavljamo sa χ^2 testom.

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

TODO Opis ispisa

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

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

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

```
## # A tibble: 7,417 x 9
## # Groups:   player_id, year, height [7,417]
##    player_id  year height hand  total_aces avg_1stIn avg_1stWon  svpt    df
##        <int> <dbl>  <int> <fct>      <int>     <dbl>      <dbl> <dbl> <int>
##  1    100284  1991    178 L             45      60.3       40.5  90.4    38
##  2    100284  1992    178 L             37      53.6       36.1  80.3    31
##  3    100284  1993    178 L              4      57         40    92.3    11
##  4    100284  1994    178 L              2      61         36    89       5
##  5    100284  1995    178 L              7      43         31.5  78.5    10
##  6    100529  1991    185 R            168      45.3       36.2  81.2    43
##  7    100529  1992    185 R             87      38.3       30.5  78.3    47
##  8    100532  1991    175 R             17      33         26.3  66       8
##  9    100581  1991    180 L            205      39.0       30.8  69.9   123
## 10    100581  1992    180 L            175      50.6       40.3  86.3   126
## # i 7,407 more rows

## # A tibble: 10,396 x 9
## # Groups:   player_id, year, height [10,396]
##    player_id  year height hand  total_aces avg_1stIn avg_1stWon  svpt    df
##        <int> <dbl>  <int> <fct>      <int>     <dbl>      <dbl> <dbl> <int>
##  1    100282  1992    180 L              0      67.5       40.5  96       5
##  2    100284  1991    178 L              9      49.2       27.1  75.6    34
##  3    100284  1992    178 L             25      57.9       33.4  90.6    46
##  4    100284  1993    178 L              4      37.4       22.2  60.4    14
##  5    100284  1994    178 L              1      56         34    87.3     3
##  6    100284  1995    178 L              3      48         29    67       2
##  7    100284  1996    178 L              3      55         30    93       2
##  8    100286  1991    168 R              0      32         18    60       2
##  9    100321  1993    193 R              0      34         14    48       0
## 10    100431  1992    178 R              8      46.5       30.5  76       4
## # i 10,386 more rows

## # A tibble: 40 x 9
## # Groups:   player_id, year, height [20]
##    player_id  year height hand  total_aces avg_1stIn avg_1stWon  svpt    df
##        <int> <dbl>  <int> <fct>      <int>     <dbl>      <dbl> <dbl> <int>
##  1    104925  2004    188 R              4      60         39    91       2
##  2    104925  2005    188 R             43      62.1       45.4  96.4    26
##  3    104925  2006    188 R            216      49.3       37    79.3    92
##  4    104925  2007    188 R            420      54.2       40.0  83.5   147
##  5    104925  2008    188 R            413      47.3       35.6  72.3   113
##  6    104925  2009    188 R            420      46.2       34.3  73.0   212
##  7    104925  2010    188 R            232      49.2       35.9  77.5   198
##  8    104925  2011    188 R            320      47.0       35.2  71.9   131
##  9    104925  2012    188 R            456      47.4       36.0  73.6   117
## 10    104925  2013    188 R            424      47.5       36.6  72.4    94
## 11    104925  2014    188 R            371      50.8       38.5  75.9    91
## 12    104925  2015    188 R            441      48.5       36.4  72.9   124
## 13    104925  2016    188 R            263      48.6       36.2  74.5   168
## 14    104925  2017    188 R            138      51.0       37.8  76.6    56
## 15    104925  2018    188 R            286      50.2       38.2  75.7   117
## 16    104925  2019    188 R            332      46.2       36.4  70.4   136
## 17    104925  2020    188 R            257      50.5       38.5  78.4   125
## 18    104925  2021    188 R            416      55.7       43.1  85.4   130
```

```
## 19    104925  2022    188 R           244      46.0       36.7  70.1    66
## 20    104925  2023    188 R           295      53.8       42.2  84.9   128
## 21    104925  2004    188 R            22      57.3       34    93.7    19
## 22    104925  2005    188 R            45      57         37.6  91.3    32
## 23    104925  2006    188 R            63      52.3       34.2  82.2    59
## 24    104925  2007    188 R            98      49         32.2  79.9    48
## 25    104925  2008    188 R            73      53.8       36.6  84.6    40
## 26    104925  2009    188 R            82      53.9       35.9  86.8    51
## 27    104925  2010    188 R            72      61.1       39.9  93.1    84
## 28    104925  2011    188 R            23      57.2       36.6  88.4    12
## 29    104925  2012    188 R            46      54         37.2  87.4    30
## 30    104925  2013    188 R            52      73.1       47.2 110.     24
## 31    104925  2014    188 R            57      60         41.5  91.4    14
## 32    104925  2015    188 R            30      60.2       39.8  91.8    11
## 33    104925  2016    188 R            38      51.8       35    82.1    20
## 34    104925  2017    188 R            31      57.8       38.6  90.1    23
## 35    104925  2018    188 R            56      57.4       38.8  87.1    35
## 36    104925  2019    188 R            60      61.4       40.3  91.3    32
## 37    104925  2020    188 R            21      45.6       31.2  72      12
## 38    104925  2021    188 R            31      56.4       39.6  92      18
## 39    104925  2022    188 R            38      69         45.2 106      22
## 40    104925  2023    188 R            15      66         41   100.     15

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

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

## # A tibble: 20 x 10
## # Groups:   player_id, year, height [20]
```

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

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