SAP - Druga auditorna vježba

Case study *FIFA 19 data*: Statističko zaključivanje za metričke i kategorijske podatke podatke

Tessa Bauman, Stjepan Begušić, David Bojanić, Tomislav Kovačević, Andro Merćep

17.11.2021.

Uvod

U ovim vježbama obrađujemo općenite koncepte postavljanja i testiranja statističkih hipoteza te računanja p-vrijednosti na primjeru konkretnih statističkih testova za metričke podatke (kao što su t-test, χ^2-test i F-test) te statističkih testova za kategorijske podatke (kao što su χ^2-test i Fisher-Irwinov egzaktni test).

Case study: *FIFA 19*

Kaggle.com je web stranica namjenjena natjecanju pojedinaca nad raznim zadatcima istraživačke naravi kao što su zadatci strojnoga učenja, zadatci statističke analize i slično. Osim toga, Kaggle pruža istraživačima na raspolaganje niz skupova podataka i njihovih analiza, tečajeva te dio svojih računalnih moći.

Jedan takav dataset je i FIFA dataset koji se sastoji od statistika igrača koji su sudjelovali u nogometnoj igrici FIFA od 2015. do 2020. godine. Iz opisa vidimo da su originalni podatci preuzeti sa sofifa web stranice sa koje možemo isčitati opis svakog podatka. U ovim vježbama ćemo analizirati FIFA 2019 skup podataka te ćemo se dodatno poslužiti podatcima FIFA 2020 kako bi usporedili napredak igrače kroz sezonu.

Kao što smo već naglasili, prije svakog korištenja podataka, potrebno je znati kontekst podataka. Koristit ćemo se alatima deskriptivne analize kako bi se upoznali sa našim skupom podataka i pravilno ih interpretirali.

Deskriptivna analiza

Učitajmo potrebne pakete.

```
library(dplyr)
```

Učitajmo podatke.

```
fifa19 = read.csv("players_19.csv")
dim(fifa19)
```

```
## [1] 17770   104
```

Podatci se sastoje od 17770 igrača i 104 njihovih opisa (varijabli).

Koji su nam to opisi igrača?

```
names(fifa19)
```

```
##   [1] "sofifa_id"                "player_url"
##   [3] "short_name"               "long_name"
##   [5] "age"                      "dob"
##   [7] "height_cm"                "weight_kg"
##   [9] "nationality"              "club"
```

```
##  [11] "overall"                    "potential"
##  [13] "value_eur"                  "wage_eur"
##  [15] "player_positions"           "preferred_foot"
##  [17] "international_reputation"   "weak_foot"
##  [19] "skill_moves"                "work_rate"
##  [21] "body_type"                  "real_face"
##  [23] "release_clause_eur"         "player_tags"
##  [25] "team_position"              "team_jersey_number"
##  [27] "loaned_from"                "joined"
##  [29] "contract_valid_until"       "nation_position"
##  [31] "nation_jersey_number"       "pace"
##  [33] "shooting"                   "passing"
##  [35] "dribbling"                  "defending"
##  [37] "physic"                     "gk_diving"
##  [39] "gk_handling"                "gk_kicking"
##  [41] "gk_reflexes"                "gk_speed"
##  [43] "gk_positioning"             "player_traits"
##  [45] "attacking_crossing"         "attacking_finishing"
##  [47] "attacking_heading_accuracy" "attacking_short_passing"
##  [49] "attacking_volleys"          "skill_dribbling"
##  [51] "skill_curve"                "skill_fk_accuracy"
##  [53] "skill_long_passing"         "skill_ball_control"
##  [55] "movement_acceleration"      "movement_sprint_speed"
##  [57] "movement_agility"           "movement_reactions"
##  [59] "movement_balance"           "power_shot_power"
##  [61] "power_jumping"              "power_stamina"
##  [63] "power_strength"             "power_long_shots"
##  [65] "mentality_aggression"       "mentality_interceptions"
##  [67] "mentality_positioning"      "mentality_vision"
##  [69] "mentality_penalties"        "mentality_composure"
##  [71] "defending_marking"          "defending_standing_tackle"
##  [73] "defending_sliding_tackle"   "goalkeeping_diving"
##  [75] "goalkeeping_handling"       "goalkeeping_kicking"
##  [77] "goalkeeping_positioning"    "goalkeeping_reflexes"
##  [79] "ls"                         "st"
##  [81] "rs"                         "lw"
##  [83] "lf"                         "cf"
##  [85] "rf"                         "rw"
##  [87] "lam"                        "cam"
##  [89] "ram"                        "lm"
##  [91] "lcm"                        "cm"
##  [93] "rcm"                        "rm"
##  [95] "lwb"                        "ldm"
##  [97] "cdm"                        "rdm"
##  [99] "rwb"                        "lb"
## [101] "lcb"                        "cb"
## [103] "rcb"                        "rb"
```

Igrači su opisani raznim općenitim varijablama (kao što su ime, godine, za koji klub igraju, itd.) te "nogometnim" varijablama (kao što su kvantizirane karakteristike napada, obrane, itd.).

```
View(fifa19)
```

Vidimo da nam nisu svi podatci jednako koristni. Na primjer, varijabla "player_url" je web url igrača na sofifa stranicama odakle je preuzet skup podataka. Sa takvim podatkom ne možemo raditi nikakve statističke

zaključke. Dodatno, što više varijabli imamo, to je više naš skup podataka nepregledan. Stoga ih je često poželjno izbaciti iz samog dataseta.

```
fifa19 = select(fifa19, -c("player_url","long_name","real_face","player_tags","loaned_from","player_tra
dim(fifa19)
```

```
## [1] 17770    98
```

Preostalo nam je 98 varijabli koje opisuju nogometne igrače.

Kako se ponašaju te varijable?

```
summary(fifa19)
```

```
##    sofifa_id             short_name         age               dob
##  Min.   :   164   J. Rodríguez:   10   Min.   :16.00   1992-02-29:  115
##  1st Qu.:199803   Paulinho    :    7   1st Qu.:21.00   1984-02-29:  106
##  Median :221350   J. Gómez    :    6   Median :25.00   1988-02-29:  104
##  Mean   :213798   J. Hernández:    6   Mean   :25.19   1993-01-15:   13
##  3rd Qu.:235968   J. Williams :    6   3rd Qu.:28.00   1993-03-05:   13
##  Max.   :246063   R. Williams :    6   Max.   :45.00   1991-01-08:   12
##                   (Other)     :17729                   (Other)   :17407
##    height_cm       weight_kg         nationality
##  Min.   :154.0   Min.   : 50.00   England  : 1656
##  1st Qu.:176.0   1st Qu.: 70.00   Germany  : 1191
##  Median :181.0   Median : 75.00   Spain    : 1037
##  Mean   :181.3   Mean   : 75.31   Argentina:  943
##  3rd Qu.:186.0   3rd Qu.: 80.00   France   :  892
##  Max.   :205.0   Max.   :110.00   Brazil   :  816
##                                   (Other)  :11235
##                      club          overall        potential
##  Arsenal                :   33   Min.   :47.00   Min.   :48.00
##  AS Monaco              :   33   1st Qu.:62.00   1st Qu.:67.00
##  Athletic Club de Bilbao:   33   Median :66.00   Median :71.00
##  Atlético Madrid        :   33   Mean   :66.24   Mean   :71.38
##  Borussia Dortmund      :   33   3rd Qu.:71.00   3rd Qu.:75.00
##  Bournemouth            :   33   Max.   :94.00   Max.   :95.00
##  (Other)                :17572
##    value_eur            wage_eur      player_positions preferred_foot
##  Min.   :        0   Min.   :     0   CB     :2199     Left : 4131
##  1st Qu.:   300000   1st Qu.:  1000   GK     :1986     Right:13639
##  Median :   675000   Median :  3000   ST     :1795
##  Mean   :  2440756   Mean   :  9956   CM     : 756
##  3rd Qu.:  2000000   3rd Qu.:  9000   CDM, CM: 690
##  Max.   :118500000   Max.   :565000   LB     : 632
##                                       (Other):9712
##  international_reputation   weak_foot      skill_moves            work_rate
##  Min.   :1.000            Min.   :1.000   Min.   :1.000   Medium/Medium:9621
##  1st Qu.:1.000            1st Qu.:3.000   1st Qu.:2.000   High/Medium  :3152
##  Median :1.000            Median :3.000   Median :2.000   Medium/High  :1628
##  Mean   :1.118            Mean   :2.946   Mean   :2.347   High/High    : 976
##  3rd Qu.:1.000            3rd Qu.:3.000   3rd Qu.:3.000   Medium/Low   : 814
##  Max.   :5.000            Max.   :5.000   Max.   :5.000   High/Low     : 682
##                                                           (Other)      : 897
##       body_type     release_clause_eur  team_position  team_jersey_number
##  Normal    :10410   Min.   :    13000   SUB    :7593   Min.   : 1.00
```

```
##  Lean      : 6268   1st Qu.:   525000   RES    :2928   1st Qu.: 8.00
##  Stocky    : 1085   Median :  1200000   GK     : 642   Median :17.00
##  Akinfenwa :    1   Mean   :  4645685   LCB    : 638   Mean   :19.67
##  C. Ronaldo:    1   3rd Qu.:  3500000   RCB    : 638   3rd Qu.:26.00
##  Courtois  :    1   Max.   :228100000   LB     : 558   Max.   :99.00
##  (Other)   :    4   NA's   :1513        (Other):4773   NA's   :223
##         joined      contract_valid_until nation_position nation_jersey_number
##            : 1504   Min.   :2018                :16666   Min.   : 1.00
##  2018-07-01: 1495   1st Qu.:2019         SUB    :  576   1st Qu.: 6.00
##  2017-07-01: 1122   Median :2020         GK     :   48   Median :12.00
##  2018-01-01:  618   Mean   :2020         LCB    :   48   Mean   :12.14
##  2016-07-01:  611   3rd Qu.:2021         RCB    :   48   3rd Qu.:18.00
##  2015-07-01:  361   Max.   :2026         LB     :   44   Max.   :87.00
##  (Other)   :12059   NA's   :223          (Other):  340   NA's   :16666
##       pace           shooting        passing        dribbling
##  Min.   :24.00   Min.   :15.00   Min.   :24.00   Min.   :23.00
##  1st Qu.:61.00   1st Qu.:42.00   1st Qu.:50.00   1st Qu.:57.00
##  Median :69.00   Median :54.00   Median :58.00   Median :64.00
##  Mean   :67.82   Mean   :52.25   Mean   :57.08   Mean   :62.27
##  3rd Qu.:75.00   3rd Qu.:63.00   3rd Qu.:64.00   3rd Qu.:69.00
##  Max.   :96.00   Max.   :93.00   Max.   :92.00   Max.   :96.00
##  NA's   :1986    NA's   :1986    NA's   :1986    NA's   :1986
##    defending         physic        gk_diving      gk_handling
##  Min.   :15.00   Min.   :30.00   Min.   :45.00   Min.   :43.00
##  1st Qu.:37.00   1st Qu.:59.00   1st Qu.:60.00   1st Qu.:58.00
##  Median :56.00   Median :66.00   Median :65.00   Median :63.00
##  Mean   :51.67   Mean   :64.97   Mean   :65.38   Mean   :62.93
##  3rd Qu.:65.00   3rd Qu.:72.00   3rd Qu.:71.00   3rd Qu.:68.00
##  Max.   :91.00   Max.   :89.00   Max.   :91.00   Max.   :92.00
##  NA's   :1986    NA's   :1986    NA's   :15784   NA's   :15784
##    gk_kicking     gk_reflexes       gk_speed     gk_positioning
##  Min.   :35.00   Min.   :44.00   Min.   :12.00   Min.   :38.00
##  1st Qu.:56.00   1st Qu.:60.00   1st Qu.:30.00   1st Qu.:57.00
##  Median :61.00   Median :66.00   Median :40.00   Median :63.00
##  Mean   :61.51   Mean   :66.15   Mean   :38.68   Mean   :63.09
##  3rd Qu.:66.00   3rd Qu.:72.00   3rd Qu.:46.00   3rd Qu.:69.00
##  Max.   :91.00   Max.   :94.00   Max.   :65.00   Max.   :90.00
##  NA's   :15784   NA's   :15784   NA's   :15784   NA's   :15784
##  attacking_crossing attacking_finishing attacking_heading_accuracy
##  62     :  501      58     :  441       58     :  583
##  60     :  485      60     :  412       60     :  551
##  59     :  478      62     :  404       55     :  526
##  58     :  476      65     :  392       59     :  517
##  64     :  467      59     :  385       62     :  507
##  65     :  463      64     :  363       65     :  468
##  (Other):14900      (Other):15373       (Other):14618
##  attacking_short_passing attacking_volleys skill_dribbling  skill_curve
##  64     :  785           49     :  389     64     :  669   45     :  383
##  65     :  739           48     :  371     63     :  602   58     :  377
##  68     :  718           59     :  368     65     :  594   48     :  369
##  62     :  699           45     :  358     62     :  585   60     :  359
##  66     :  699           55     :  358     66     :  584   64     :  355
##  63     :  675           52     :  353     68     :  578   49     :  352
##  (Other):13455           (Other):15573     (Other):14158   (Other):15575
```

```
##  skill_fk_accuracy skill_long_passing skill_ball_control movement_acceleration
##  42     :  469     58     :  586      65     :  754      68     :  689
##  32     :  468     63     :  562      64     :  718      69     :  664
##  40     :  457     62     :  561      62     :  675      67     :  652
##  35     :  442     60     :  560      66     :  667      66     :  568
##  34     :  425     59     :  558      63     :  658      65     :  556
##  39     :  421     65     :  518      68     :  653      74     :  543
##  (Other):15088     (Other):14425      (Other):13645      (Other):14098
##  movement_sprint_speed movement_agility movement_reactions movement_balance
##  67     :  664         68     :  553    65     :  803      68     :  585
##  69     :  661         70     :  533    64     :  789      67     :  568
##  68     :  659         69     :  528    62     :  771      66     :  565
##  66     :  614         72     :  525    60     :  769      70     :  561
##  65     :  582         66     :  524    63     :  736      65     :  554
##  72     :  558         67     :  524    58     :  693      69     :  551
##  (Other):14032         (Other):14583    (Other):13209      (Other):14386
##  power_shot_power power_jumping   power_stamina   power_strength
##  68     :  548    70     :  658   68     :  610   67     :  605
##  62     :  519    72     :  621   69     :  601   68     :  597
##  70     :  517    71     :  617   67     :  560   65     :  577
##  65     :  501    68     :  601   72     :  548   69     :  573
##  66     :  487    63     :  589   65     :  545   70     :  563
##  60     :  476    65     :  589   66     :  540   72     :  546
##  (Other):14722    (Other):14095   (Other):14366   (Other):14309
##  power_long_shots mentality_aggression mentality_interceptions
##  58     :  479    65     :  486        62     :  465
##  62     :  463    70     :  478        65     :  451
##  59     :  454    68     :  470        66     :  450
##  55     :  409    58     :  446        63     :  434
##  64     :  392    60     :  437        64     :  420
##  52     :  389    55     :  421        60     :  391
##  (Other):15184    (Other):15032        (Other):15159
##  mentality_positioning mentality_vision mentality_penalties mentality_composure
##  58     :  537         58     :  571    45     :  481       60     :  682
##  65     :  501         55     :  536    55     :  467       55     :  666
##  60     :  490         60     :  509    48     :  464       65     :  658
##  62     :  480         65     :  479    59     :  453       58     :  638
##  64     :  452         62     :  477    49     :  449       62     :  637
##  59     :  445         59     :  475    58     :  448       59     :  596
##  (Other):14865         (Other):14723    (Other):15008       (Other):13893
##  defending_marking defending_standing_tackle defending_sliding_tackle
##  60     :  533     66     :  553             64     :  535
##  62     :  515     65     :  543             62     :  516
##  65     :  487     64     :  529             63     :  503
##  64     :  441     62     :  475             65     :  472
##  58     :  425     68     :  456             60     :  434
##  55     :  412     63     :  439             14     :  416
##  (Other):14957     (Other):14775             (Other):14894
##  goalkeeping_diving goalkeeping_handling goalkeeping_kicking
##  8      :1575       10     :1597         12     :1579
##  7      :1550       11     :1568         9      :1575
##  12     :1540       12     :1567         7      :1557
##  14     :1540       7      :1556         13     :1538
##  9      :1529       14     :1537         8      :1534
```

```
##  13     :1524       8      :1519         14     :1528
##  (Other):8512       (Other):8426         (Other):8459
##  goalkeeping_positioning goalkeeping_reflexes       ls              st
##  10     :1600            11     :1565                : 1986          : 1986
##  8      :1575            10     :1562         60+2   :  683   60+2   :  683
##  7      :1541            9      :1546         59+2   :  664   59+2   :  664
##  11     :1532            7      :1535         61+2   :  664   61+2   :  664
##  9      :1531            8      :1533         58+2   :  652   58+2   :  652
##  12     :1518            14     :1523         57+2   :  646   57+2   :  646
##  (Other):8473            (Other):8506         (Other):12475   (Other):12475
##        rs              lw              lf              cf
##         : 1986          : 1986          : 1986          : 1986
##  60+2   :  683   63+2   :  712   61+2   :  685   61+2   :  685
##  59+2   :  664   61+2   :  694   59+2   :  680   59+2   :  680
##  61+2   :  664   60+2   :  680   63+2   :  677   63+2   :  677
##  58+2   :  652   64+2   :  675   60+2   :  666   60+2   :  666
##  57+2   :  646   62+2   :  646   65+2   :  664   65+2   :  664
##  (Other):12475   (Other):12377   (Other):12412   (Other):12412
##        rf              rw              lam             cam
##         : 1986          : 1986          : 1986          : 1986
##  61+2   :  685   63+2   :  712   62+2   :  698   62+2   :  698
##  59+2   :  680   61+2   :  694   63+2   :  696   63+2   :  696
##  63+2   :  677   60+2   :  680   60+2   :  695   60+2   :  695
##  60+2   :  666   64+2   :  675   61+2   :  688   61+2   :  688
##  65+2   :  664   62+2   :  646   59+2   :  671   59+2   :  671
##  (Other):12412   (Other):12377   (Other):12336   (Other):12336
##       ram             lm              lcm             cm
##         : 1986          : 1986          : 1986          : 1986
##  62+2   :  698   61+2   :  768   60+2   :  730   60+2   :  730
##  63+2   :  696   63+2   :  732   62+2   :  722   62+2   :  722
##  60+2   :  695   62+2   :  717   59+2   :  714   59+2   :  714
##  61+2   :  688   59+2   :  711   58+2   :  705   58+2   :  705
##  59+2   :  671   65+2   :  707   61+2   :  693   61+2   :  693
##  (Other):12336   (Other):12149   (Other):12220   (Other):12220
##       rcm             rm              lwb             ldm
##         : 1986          : 1986          : 1986          : 1986
##  60+2   :  730   61+2   :  768   59+2   :  684   62+2   :  636
##  62+2   :  722   63+2   :  732   61+2   :  680   61+2   :  630
##  59+2   :  714   62+2   :  717   60+2   :  641   64+2   :  566
##  58+2   :  705   59+2   :  711   62+2   :  634   59+2   :  561
##  61+2   :  693   65+2   :  707   56+2   :  616   60+2   :  560
##  (Other):12220   (Other):12149   (Other):12529   (Other):12831
##       cdm             rdm             rwb             lb
##         : 1986          : 1986          : 1986          : 1986
##  62+2   :  636   62+2   :  636   59+2   :  684   61+2   :  636
##  61+2   :  630   61+2   :  630   61+2   :  680   58+2   :  633
##  64+2   :  566   64+2   :  566   60+2   :  641   59+2   :  633
##  59+2   :  561   59+2   :  561   62+2   :  634   63+2   :  625
##  60+2   :  560   60+2   :  560   56+2   :  616   64+2   :  620
##  (Other):12831   (Other):12831   (Other):12529   (Other):12637
##       lcb             cb              rcb             rb
##         : 1986          : 1986          : 1986          : 1986
##  63+2   :  614   63+2   :  614   63+2   :  614   61+2   :  636
##  62+2   :  596   62+2   :  596   62+2   :  596   58+2   :  633
```

```
##  64+2   :  577   64+2   :  577   64+2   :  577   59+2   :  633
##  65+2   :  563   65+2   :  563   65+2   :  563   63+2   :  625
##  61+2   :  548   61+2   :  548   61+2   :  548   64+2   :  620
##  (Other):12886   (Other):12886   (Other):12886   (Other):12637
```

```
sapply(fifa19, class)
```

```
##                  sofifa_id                short_name
##                  "integer"                  "factor"
##                        age                       dob
##                  "integer"                  "factor"
##                  height_cm                weight_kg
##                  "integer"                "integer"
##                nationality                     club
##                   "factor"                  "factor"
##                    overall                potential
##                  "integer"                "integer"
##                  value_eur                 wage_eur
##                  "integer"                "integer"
##           player_positions             preferred_foot
##                   "factor"                  "factor"
##   international_reputation                weak_foot
##                  "integer"                "integer"
##                skill_moves                work_rate
##                  "integer"                  "factor"
##                  body_type         release_clause_eur
##                   "factor"                "integer"
##              team_position         team_jersey_number
##                   "factor"                "integer"
##                     joined       contract_valid_until
##                   "factor"                "integer"
##            nation_position       nation_jersey_number
##                   "factor"                "integer"
##                       pace                   shooting
##                  "integer"                "integer"
##                    passing                dribbling
##                  "integer"                "integer"
##                  defending                   physic
##                  "integer"                "integer"
##                  gk_diving              gk_handling
##                  "integer"                  "integer"
##                 gk_kicking              gk_reflexes
##                  "integer"                  "integer"
##                   gk_speed           gk_positioning
##                  "integer"                  "integer"
##         attacking_crossing        attacking_finishing
##                   "factor"                   "factor"
## attacking_heading_accuracy    attacking_short_passing
##                   "factor"                   "factor"
##          attacking_volleys           skill_dribbling
##                   "factor"                   "factor"
##                skill_curve          skill_fk_accuracy
##                   "factor"                   "factor"
##         skill_long_passing         skill_ball_control
##                   "factor"                   "factor"
```

```
##      movement_acceleration      movement_sprint_speed
##                  "factor"                  "factor"
##          movement_agility         movement_reactions
##                  "factor"                  "factor"
##          movement_balance           power_shot_power
##                  "factor"                  "factor"
##             power_jumping             power_stamina
##                  "factor"                  "factor"
##            power_strength           power_long_shots
##                  "factor"                  "factor"
##      mentality_aggression    mentality_interceptions
##                  "factor"                  "factor"
##     mentality_positioning           mentality_vision
##                  "factor"                  "factor"
##       mentality_penalties        mentality_composure
##                  "factor"                  "factor"
##         defending_marking  defending_standing_tackle
##                  "factor"                  "factor"
##   defending_sliding_tackle         goalkeeping_diving
##                  "factor"                  "factor"
##      goalkeeping_handling        goalkeeping_kicking
##                  "factor"                  "factor"
##    goalkeeping_positioning       goalkeeping_reflexes
##                  "factor"                  "factor"
##                        ls                        st
##                  "factor"                  "factor"
##                        rs                        lw
##                  "factor"                  "factor"
##                        lf                        cf
##                  "factor"                  "factor"
##                        rf                        rw
##                  "factor"                  "factor"
##                       lam                       cam
##                  "factor"                  "factor"
##                       ram                        lm
##                  "factor"                  "factor"
##                       lcm                        cm
##                  "factor"                  "factor"
##                       rcm                        rm
##                  "factor"                  "factor"
##                       lwb                       ldm
##                  "factor"                  "factor"
##                       cdm                       rdm
##                  "factor"                  "factor"
##                       rwb                        lb
##                  "factor"                  "factor"
##                       lcb                        cb
##                  "factor"                  "factor"
##                       rcb                        rb
##                  "factor"                  "factor"
```

Skup podataka se pretežito sastoji od "integer" i "factor" podataka.

```
for (col_name in names(fifa19)){
  if (sum(is.na(fifa19[,col_name])) > 0){
```

```
    cat('Ukupno nedostajućih vrijednosti za varijablu ',col_name, ': ', sum(is.na(fifa19[,col_name])),'
  }
}
```

```
## Ukupno nedostajućih vrijednosti za varijablu  release_clause_eur :  1513
## Ukupno nedostajućih vrijednosti za varijablu  team_jersey_number :  223
## Ukupno nedostajućih vrijednosti za varijablu  contract_valid_until :  223
## Ukupno nedostajućih vrijednosti za varijablu  nation_jersey_number :  16666
## Ukupno nedostajućih vrijednosti za varijablu  pace :  1986
## Ukupno nedostajućih vrijednosti za varijablu  shooting :  1986
## Ukupno nedostajućih vrijednosti za varijablu  passing :  1986
## Ukupno nedostajućih vrijednosti za varijablu  dribbling :  1986
## Ukupno nedostajućih vrijednosti za varijablu  defending :  1986
## Ukupno nedostajućih vrijednosti za varijablu  physic :  1986
## Ukupno nedostajućih vrijednosti za varijablu  gk_diving :  15784
## Ukupno nedostajućih vrijednosti za varijablu  gk_handling :  15784
## Ukupno nedostajućih vrijednosti za varijablu  gk_kicking :  15784
## Ukupno nedostajućih vrijednosti za varijablu  gk_reflexes :  15784
## Ukupno nedostajućih vrijednosti za varijablu  gk_speed :  15784
## Ukupno nedostajućih vrijednosti za varijablu  gk_positioning :  15784
```

```
cat('\n Dimenzija podataka: ', dim(fifa19))
```

```
##
##  Dimenzija podataka:  17770 98
```

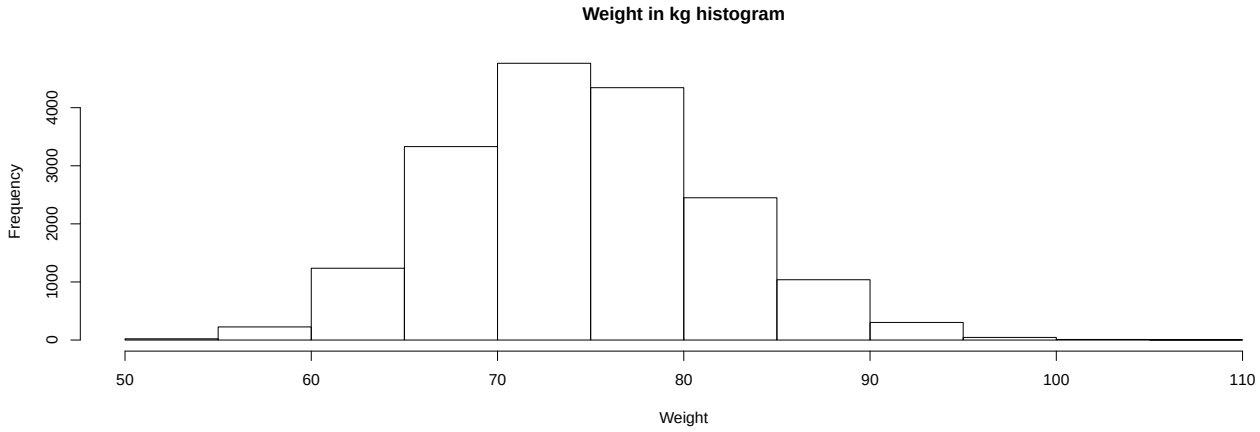
Smisleno je da varijable koji počinju sa "gk_*" nisu poznate za sve nogometne igrače obzirom da su to varijable koje opisju vratare. Dakle, takve varijable nećemo izbacivati iz skupa podataka. S druge strane, varijabla "nation_jersey_number" ima 93% nedostajećih vrijednosti što znači da nam ne daje puno informacija o igračima. Preostale varijable nemaju puno nedostajećih vrijednosti.

```
fifa19 = select(fifa19, -c("nation_jersey_number"))
dim(fifa19)
```

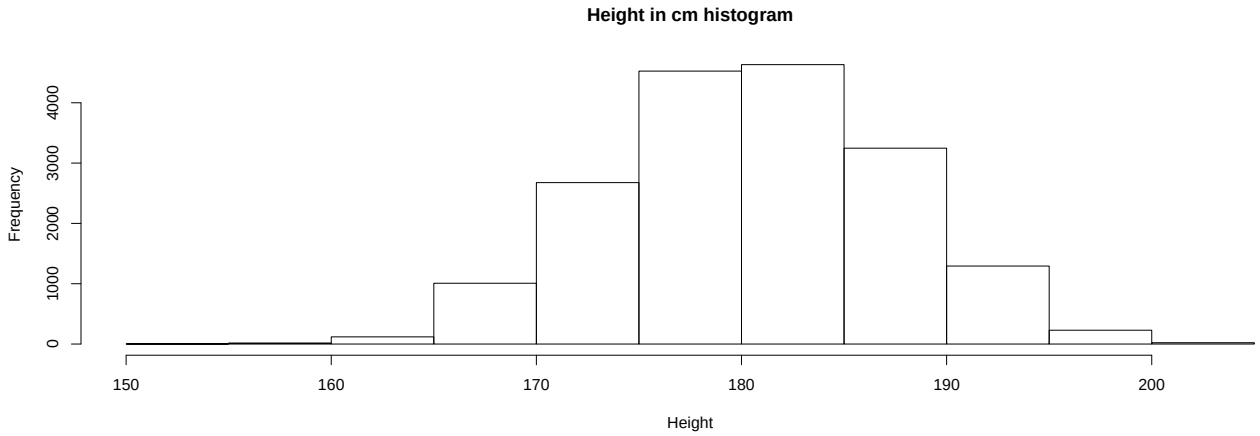
```
## [1] 17770    97
```

Promotrimo sada kako izgledaju neke od varijabli nad kojima ćemo kasnije provesti analizu. Promotrimo najprije numeričke varijable.

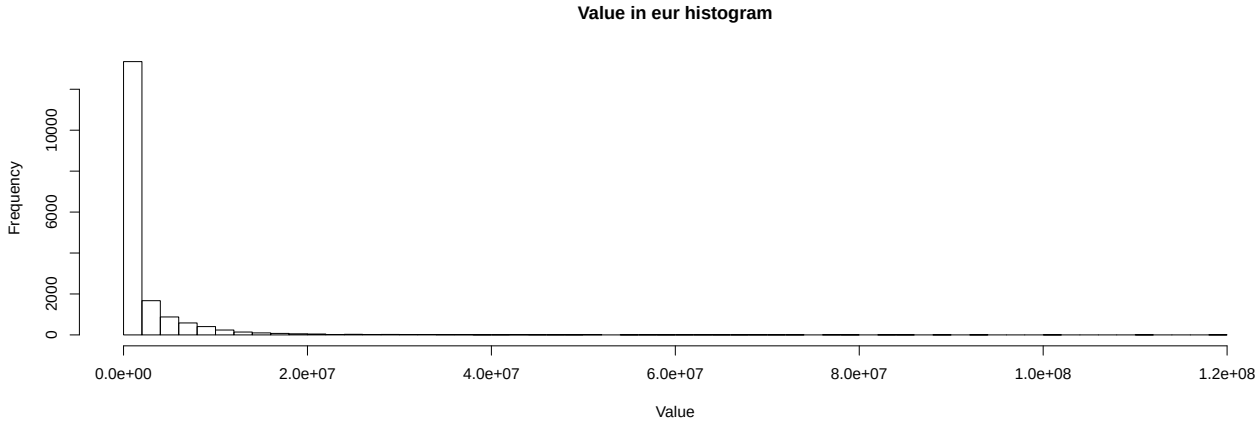
```
hist(fifa19$weight_kg,main='Weight in kg histogram', xlab='Weight', ylab='Frequency')
```




```
hist(fifa19$height_cm,main='Height in cm histogram', xlab='Height', ylab='Frequency')
```



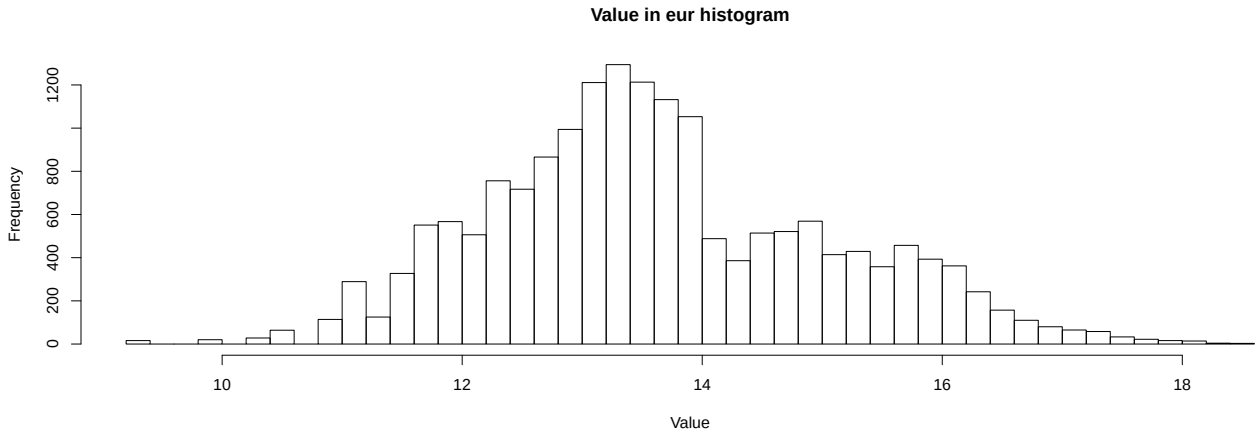

```
hist(fifa19$value_eur,main='Value in eur histogram',xlab='Value',ylab='Frequency', breaks=50)
```




```
hist(fifa19$wage_eur,main='Wage in eur histogram',xlab='Wage',ylab='Frequency', breaks=50)
```

Pokušajmo log transformacijom približiti podatke normalnoj distribuciji.

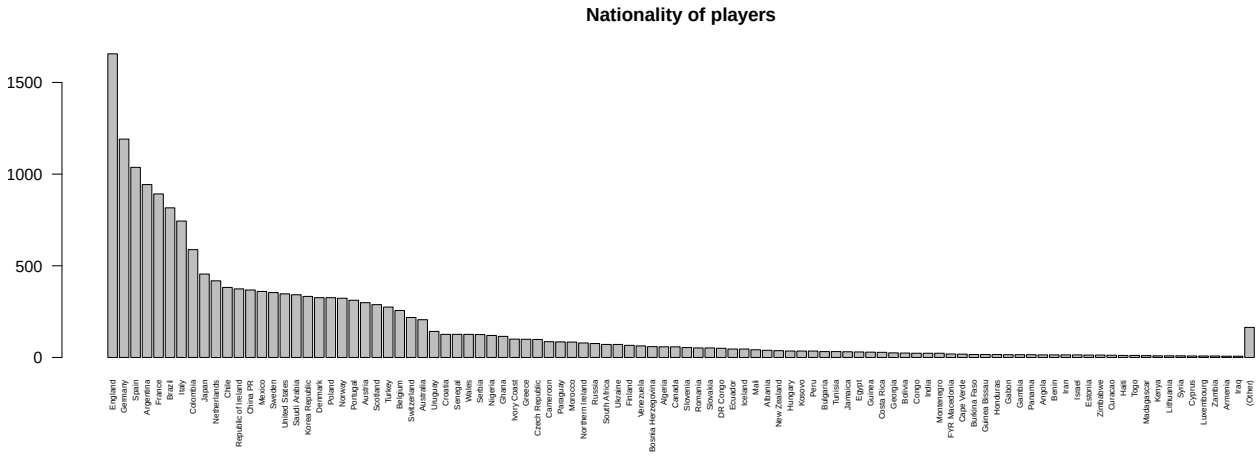
```
hist(log(fifa19$value_eur),main='Value in eur histogram',xlab='Value',ylab='Frequency', breaks=50)
```




```
hist(log(fifa19$wage_eur),main='Wage in eur histogram',xlab='Wage',ylab='Frequency', breaks=20)
```

Promotrimo sada kategoričke varijable.

```
barplot(summary(fifa19$nationality),las=2,cex.names=.5,main='Nationality of players')
```




```
print('Igračeva preferirana noga za udarce: ')
```

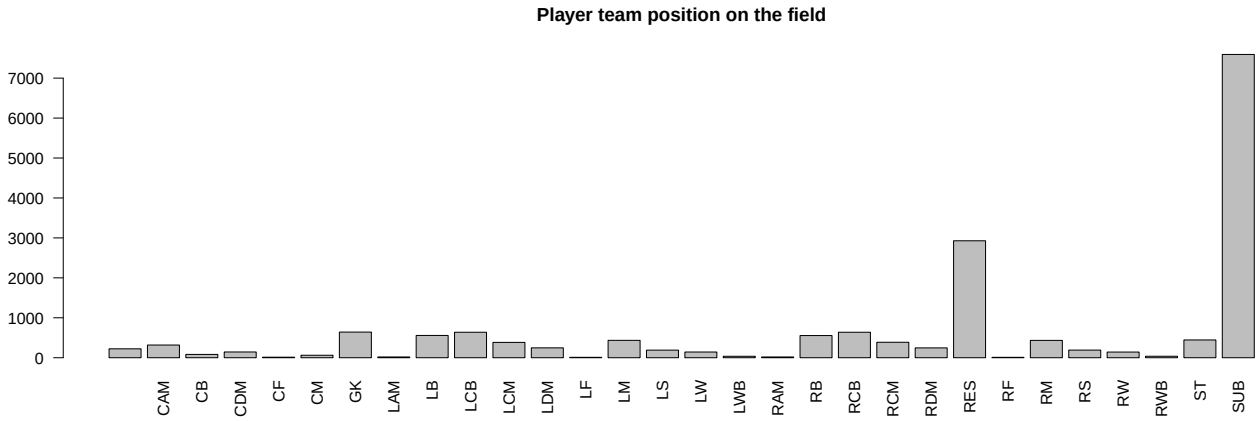
```
## [1] "Igračeva preferirana noga za udarce: "
```

```
table(fifa19$preferred_foot)
```

```
##
```

```
##  Left Right
##  4131 13639
```

```
barplot(table(fifa19$team_position),las=2,main='Player team position on the field')
```



Primijetimo da varijabla "team_position" sadrži i igrače bez pozicije.

```
table(fifa19$team_position)
```

```
##
##       CAM   CB  CDM   CF   CM   GK  LAM   LB  LCB  LCM  LDM   LF   LM   LS   LW
##  223  318   83  144   13   62  642   19  558  638  384  247   10  435  190  143
##  LWB  RAM   RB  RCB  RCM  RDM  RES   RF   RM   RS   RW  RWB   ST  SUB
##   36   19  555  638  387  246 2928   10  434  191  142   37  445 7593
```

Konkretnije, njih 223, nemaju poziciju u timu. Moramo pripaziti pri analizi takvih varijabli.

Sada kada smo dobili bolji uvid u naše podatke, možemo si postaviti zanimljiva pitanja te pokušati odgovoriti na njih koristeći razne statističke alate.

Jesu li hrvatski igrači viši od španjolskih?

```
croatian_players = fifa19[fifa19$nationality == "Croatia",]
spanish_players = fifa19[fifa19$nationality == "Spain",]
```

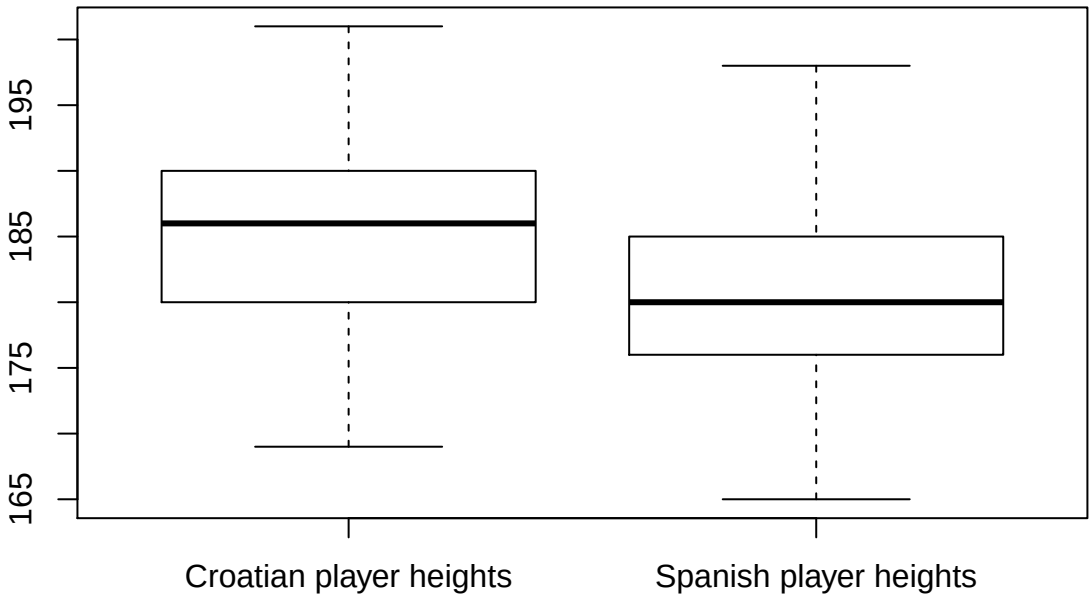
```
cat('Prosječna visina hrvatskih igrača iznosi ', mean(croatian_players$height_cm),'\n')
```

```
## Prosječna visina hrvatskih igrača iznosi  185.3095
```

```
cat('Prosječna visina španjolskih igrača iznosi ', mean(spanish_players$height_cm), '\n')
```

```
## Prosječna visina španjolskih igrača iznosi  180.6114
```

```
boxplot(croatian_players$height_cm, spanish_players$height_cm,
        names = c('Croatian player heights','Spanish player heights'),
        main='Boxplot of croatian and spanish player heights')
```



Postoje indikacije da bi hrvatski igrači trebali biti viši od španjolskih.

Ovakvo ispitivanje možemo provesti t-testom. Moraju li neke pretpostavke biti zadovoljene za naše podatke?

Testiranje jednakosti srednjih vrijednosti dvije populacije

Neka su $X_1^1, X_1^2, \ldots, X_1^{n_1}$ i $X_2^1, X_2^2, \ldots, X_2^{n_2}$ dva nezavisna slučajna uzorka koji dolaze iz normalnih distribucija s očekivanjima μ_1 i μ_2 te s nepoznatim, ali jednakim varijancama σ. Zajednička disperzija uzorka se računa kao težinska sredina disperzija S_{X_1} i S_{X_2}:

$$S_X^2 = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2].$$

Slučajna varijabla

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ima jediničnu normalnu distribuciju. Slučajna varijabla

$$W^2 = \frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{\sigma^2}$$

ima χ^2 razdiobu s $n_1 + n_2 - 2$ stupnja slobode. Zato slučajna varijabla

$$T = \frac{Z\sqrt{n_1 + n_2 - 2}}{W} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_X\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ima egzaktnu t distribuciju s $n_1 + n_2 - 2$ stupnja slobode.

Ukoliko imamo 2 nezavisno normalo distribuirana uzorka, ali ovoga puta sa različitim varijancama, tada koristimo testnu statistiku

$$T' = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}}}$$

koja ima aproksimativnu t-distribuciju sa stupnjevima slobode

$$v = \frac{(s^2_{X_1}/n_1 + s^2_{X_2}/n_2)^2}{(s^2_{X_1}/n_1)^2/(n_1 - 1) + (s^2_{X_2}/n_2)^2/(n_2 - 1)}$$

gdje je

$$s^2_{X_i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_i^j - \bar{X}_i)^2$$

za $i = 1, 2$.

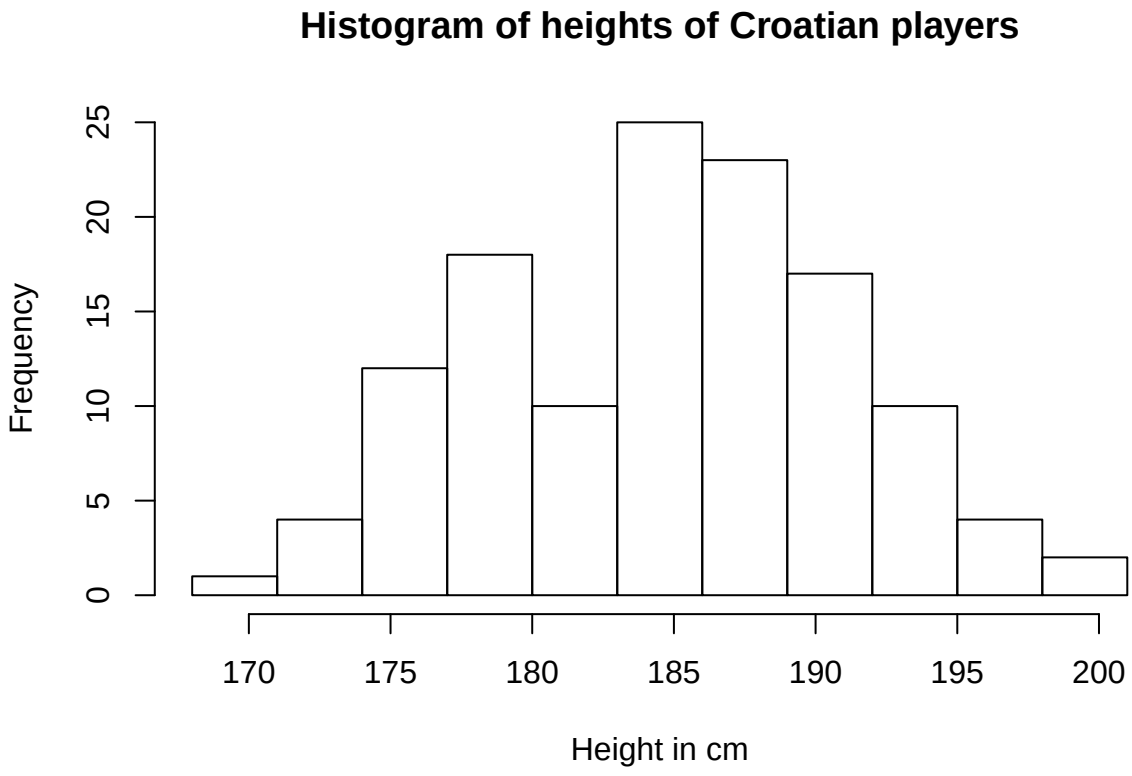
Hipoteze tada glase:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2 \quad , \quad \mu_1 > \mu_2 \quad , \quad \mu_1 \neq \mu_2$$

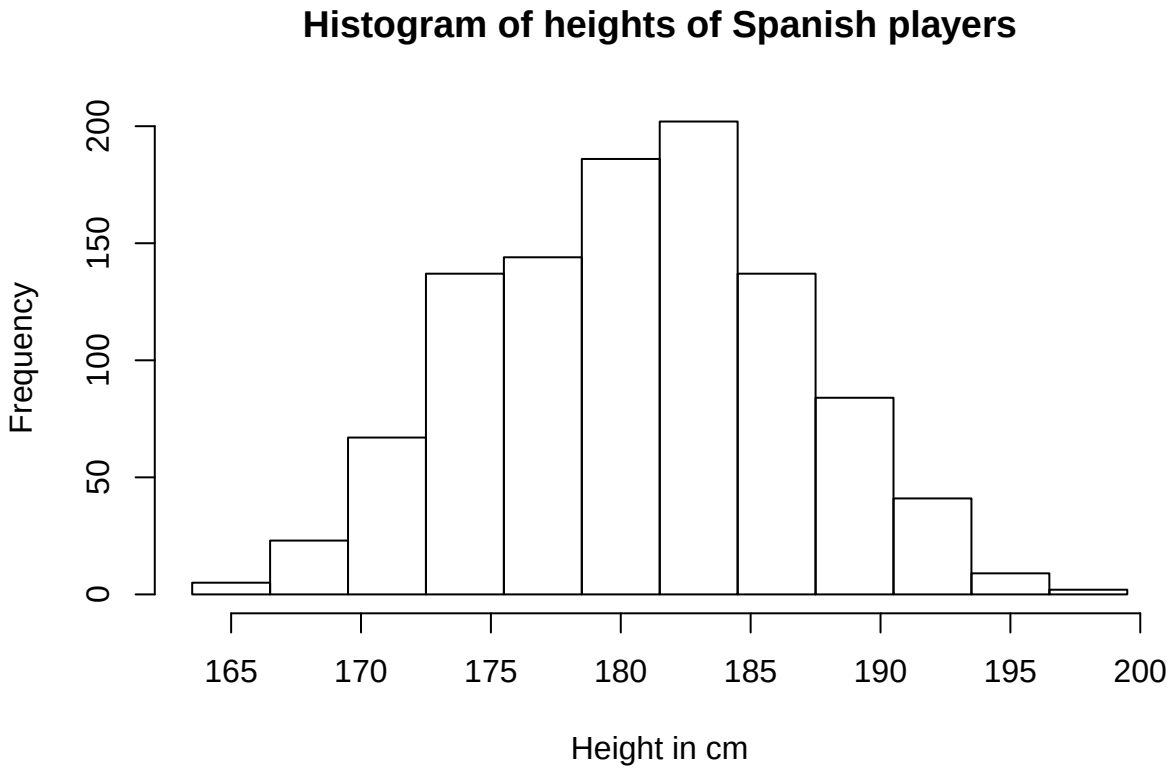
Test o jednakosti srednjih vrijednosti dvije populacije u R-u je implementiran u funkciji `t.test()`.

Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo dva uzoraka iz dvije različite zemlje, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju najčešće provjeravamo: histgoramom, qq-plotom te KS-testom (kojim provjeravamo pripadnost podataka distribuciji).

```
hist(croatian_players$height_cm,
     breaks=seq(min(croatian_players$height_cm)-1,max(croatian_players$height_cm)+1,3),
     main='Histogram of heights of Croatian players',
     xlab='Height in cm')
```

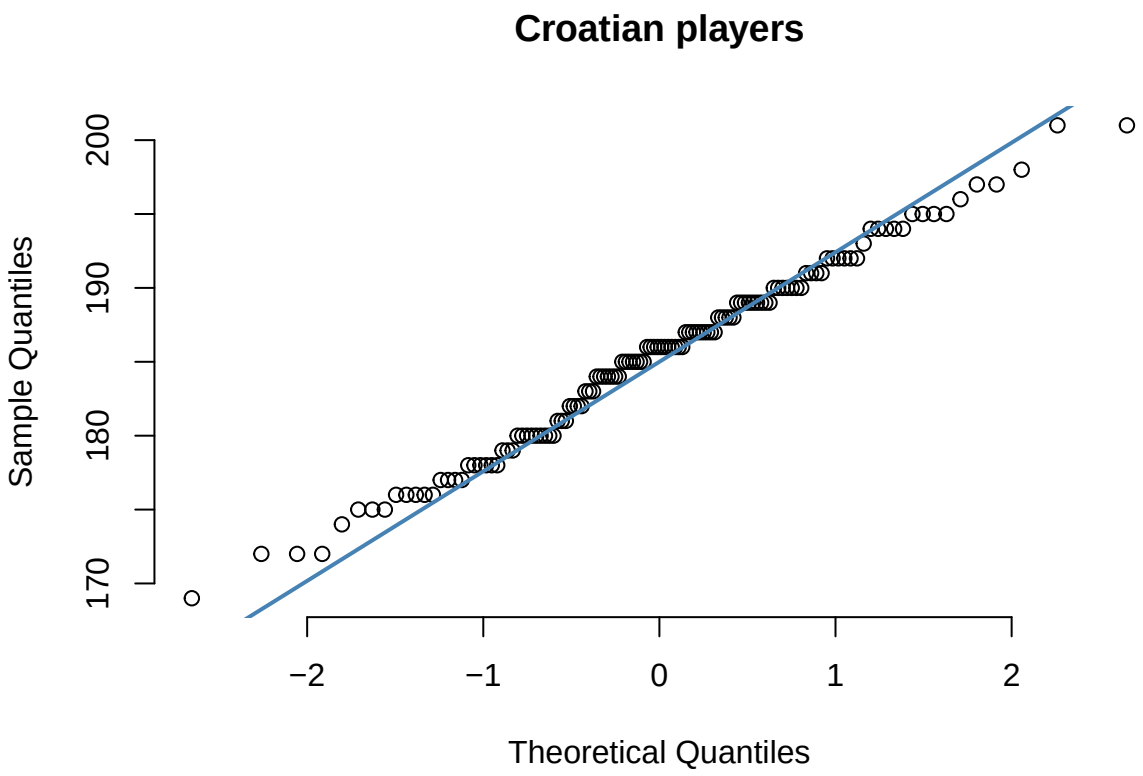


```
hist(spanish_players$height_cm,
     breaks=seq(min(spanish_players$height_cm)-1.5,max(spanish_players$height_cm)+1.5,3),
     main='Histogram of heights of Spanish players',
     xlab='Height in cm')
```

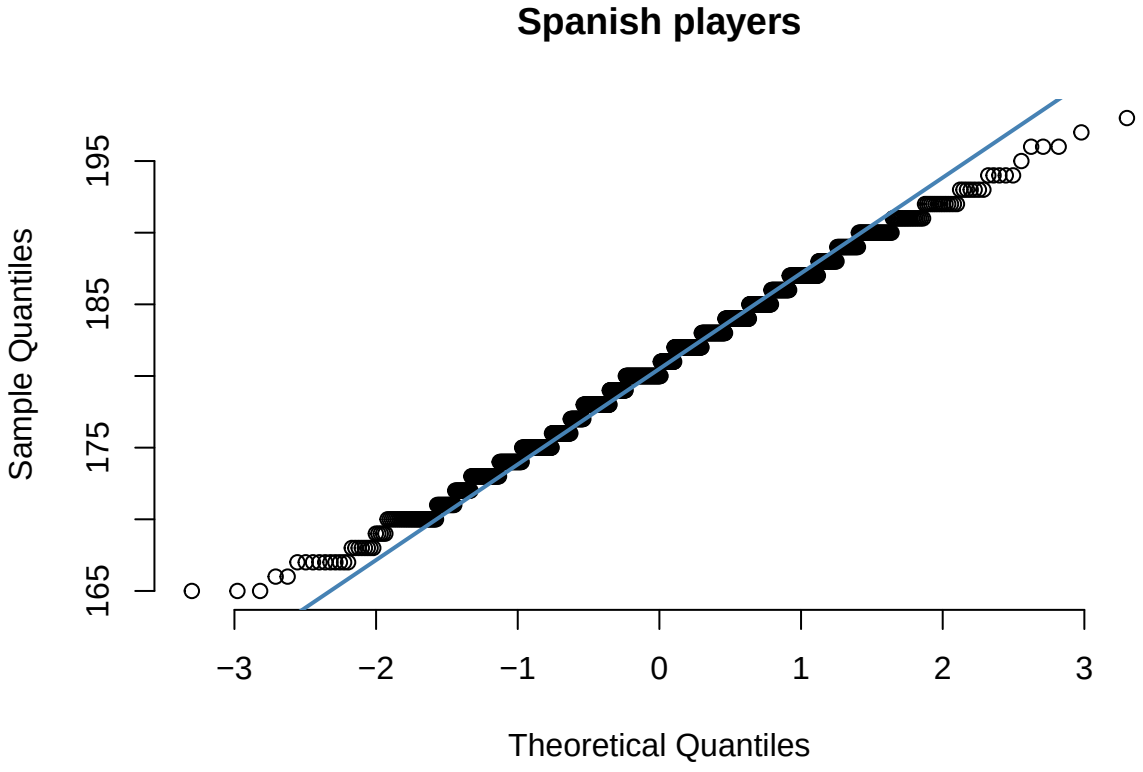


Histogrami upućuju na normalnost podataka. Normalnost možemo još provjeriti i qqplot-ovima ili testom koji ispituje normalnost.

```
qqnorm(croatian_players$height_cm, pch = 1, frame = FALSE,main='Croatian players')
qqline(croatian_players$height_cm, col = "steelblue", lwd = 2)
```




```
qqnorm(spanish_players$height_cm, pch = 1, frame = FALSE,main='Spanish players')
qqline(spanish_players$height_cm, col = "steelblue", lwd = 2)
```



Pod uvjetom da podatci zadovoljavaju sve pretpostavke možemo nastaviti sa t-testom kako bi ispitali da li su hrvatski igrači viši od španjolskih.

Koji test koristiti? Kakve su varijance danih uzoraka?

```
var(croatian_players$height_cm)
```

```
## [1] 42.96743
```

```
var(spanish_players$height_cm)
```

```
## [1] 36.67026
```

Jesu li varijance značajno različite?

Test o jednakosti varijanci

Ako imamo dva nezavisna slučajna uzorka $X_1^1, X_1^2, \dots X_1^{n_1}$ i $X_2^1, X_2^2, \dots, X_2^{n_2}$ koji dolaze iz normalnih distribucija s varijancama σ_1^2 i σ_2^2, tada slučajna varijabla

$$F = \frac{S_{X_1}^2/\sigma_1^2}{S_{X_2}^2/\sigma_2^2}$$

ima Fisherovu distribuciju s $(n_1 - 1, n_2 - 1)$ stupnjeva slobode, pri čemu vrijedi:

$$S_{X_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_1^i - \bar{X}_1)^2, \quad S_{X_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_2^i - \bar{X}_2)^2.$$

Hipoteze testa jednakosti varijanci glase:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2 \quad , \quad \sigma_1^2 > \sigma_2^2 \quad , \quad \sigma_1^2 \neq \sigma_2^2$$

U programskom paketu R test o jednakosti varijanci je implementiran u funkciji `var.test()`, koja prima uzorke iz dvije populacije čije varijance uspoređujemo.

Dakle, ispitajmo jednakost varijanci naših danih uzoraka.

```
var.test(croatian_players$height_cm, spanish_players$height_cm)
```

```
##
##  F test to compare two variances
##
## data:  croatian_players$height_cm and spanish_players$height_cm
## F = 1.1717, num df = 125, denom df = 1036, p-value = 0.2141
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9128465 1.5459734
## sample estimates:
## ratio of variances
##           1.171724
```

```
# F = testna statistika
#  num df = 125, denom df = 1036 - stupnjevi slobode
# p-value = 0.2141, vjerojatnost da dobijemo ovaj uzorak pod uvjetom da je H0 istinit
# p value usporedujemo sa Alpha (0.05)
# je li p < Alpha ?
# p > Alpha => odbacujemo hipotezu

# ## 95 percent confidence interval:
```

```
##  3.749018      Inf

# ako uzimamo razlicite uzorke onda ce varijanbca uzorka upadati u interval [3.749018      Inf] u 95% s

# testna statisika je isto sto i podjela varijanci (?)
```

p-vrijednost od 0.2141 nam govori da nećemo odbaciti hipotezu H_0 da su varijance naša dva uzorka jednaka.

Provedimo sada t-test uz pretpostavku jednakosti varijanci.

```
# Bitan je poredak kojim funkciji 't.test()' prosljedjujemo uzorke!
t.test(croatian_players$height_cm, spanish_players$height_cm, alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  croatian_players$height_cm and spanish_players$height_cm
## t = 8.1485, df = 1161, p-value = 4.728e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.749018      Inf
## sample estimates:
## mean of x mean of y
##  185.3095  180.6114
```

Zbog jako male p-vrijednost možemo odbaciti H_0 hipotezu o jednakosti prosječnih vrijednosti u korist H_1, odnosno možemo reći da su hrvatski igrači u prosjeku značajno viši od onih španjolskih.

Jesu li se španjolski igrači poboljšali u sezoni 2020 (naspram sezone 2019)?

Naš podatkovni skup sadrži opise igrača za sve sezone od 2015. do 2020., što znači da možemo promatrati njihovo ponašanje iz jedne u drugu sezonu. Primijetimo da `t-test` pretpostavlja nezavisnost uzoraka koje uspoređujemo. Ta pretpostavka je u ovakvim podatcima narušena obzirom da ispitujemo istu vrijablu za isti primjerak za dva različita vremenska trenutka. Dakle, treba nam t-test koji uzima u obzir zavisne (uparene) uzorke.

Upareni podatci

Ukoliko imamo dva mjerenja neke veličine na istom uzorku i želimo ispitati postoji li značajna razlika između mjerenja, koristimo t-test za uparene podatke, također implementiran u funkciji `t.test()`.

Ako je par $(X_1^i, X_2^i)_{i=1}^n$ uzorak dva mjerenja n primjeraka, slučajna varijabla D_i je tada dana sa:

$$D_i = X_1^i - X_2^i$$

.

Obzirom da su dva mjerenja uzeta na istoj populaciji, slučajne varijable X_1^i i X_2^i nisu nezavisne i vrijedi:

$$\sigma_D^2 = Var(D_i) = Var(X_1^i - X_2^i) = \sigma_1^2 + \sigma_2^2 - 2Cov(X_1^i, X_2^i).$$

Ako stavimo

$$X_1^i = \mu_1 + \eta_1^i, \quad X_2^i = \mu_2 + \eta_2^i,$$

dobivamo

$$\sigma_D^2 = Var(\eta_1^i) + Var(\eta_2^i) - 2Cov(\eta_1^i, \eta_2^i).$$

Testna statistika je dana sa:

$$T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}}$$

a nulta hipoteza glasi:

$$H_0 : \mu_D = d_0$$
$$H_1 : \mu_D < d_0 \quad , \quad \mu_D > d_0 \quad , \quad \mu_d \neq d0$$

note: upareni T test koristimo umjesto T testa ako su nam podaci meduzavisni (npr ako promatramo istog igraca ali iz godine u godinu)

parametar paired = TRUE,

Učitajmo podatke igrača iz sezone 2020:

```
fifa20 = read.csv("players_20.csv")
dim(fifa20)
```

```
## [1] 18278   104
```

```
names(fifa20)
```

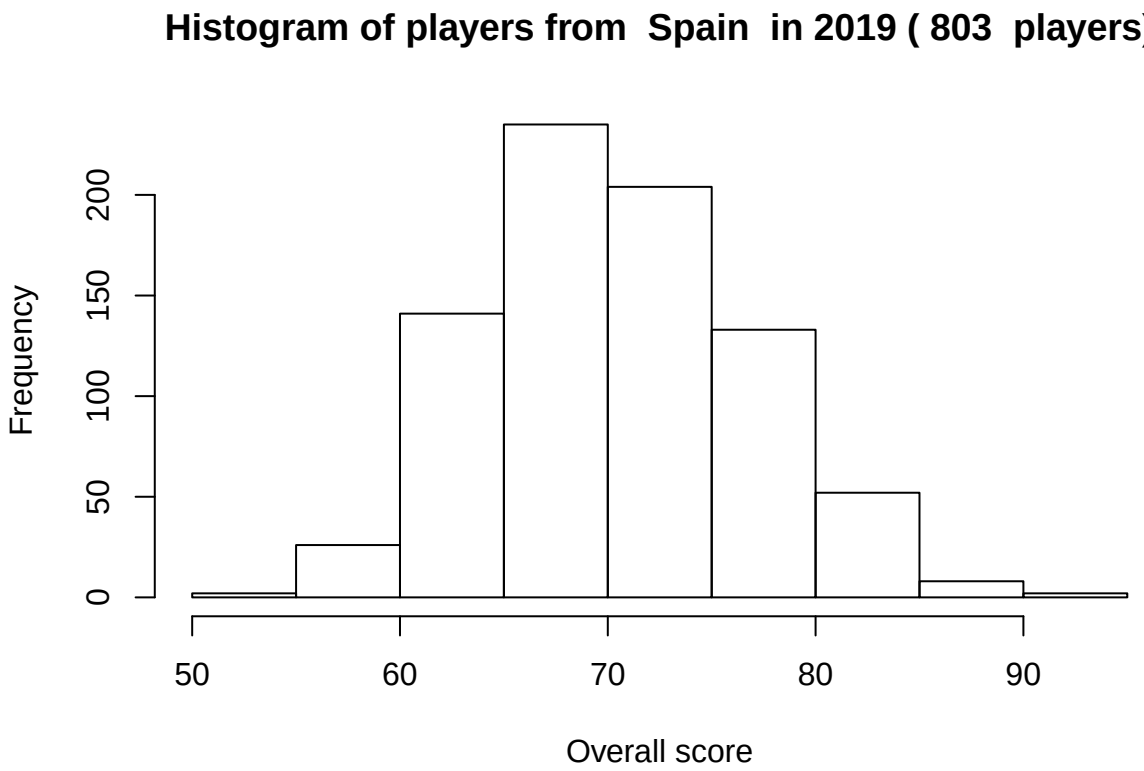
```
##   [1] "sofifa_id"                 "player_url"
##   [3] "short_name"                "long_name"
##   [5] "age"                       "dob"
##   [7] "height_cm"                 "weight_kg"
##   [9] "nationality"               "club"
##  [11] "overall"                   "potential"
##  [13] "value_eur"                 "wage_eur"
##  [15] "player_positions"          "preferred_foot"
##  [17] "international_reputation"  "weak_foot"
##  [19] "skill_moves"               "work_rate"
##  [21] "body_type"                 "real_face"
##  [23] "release_clause_eur"        "player_tags"
##  [25] "team_position"             "team_jersey_number"
##  [27] "loaned_from"               "joined"
##  [29] "contract_valid_until"      "nation_position"
##  [31] "nation_jersey_number"      "pace"
##  [33] "shooting"                  "passing"
##  [35] "dribbling"                 "defending"
##  [37] "physic"                    "gk_diving"
##  [39] "gk_handling"               "gk_kicking"
##  [41] "gk_reflexes"               "gk_speed"
##  [43] "gk_positioning"            "player_traits"
##  [45] "attacking_crossing"        "attacking_finishing"
##  [47] "attacking_heading_accuracy" "attacking_short_passing"
##  [49] "attacking_volleys"         "skill_dribbling"
##  [51] "skill_curve"               "skill_fk_accuracy"
##  [53] "skill_long_passing"        "skill_ball_control"
##  [55] "movement_acceleration"     "movement_sprint_speed"
##  [57] "movement_agility"          "movement_reactions"
##  [59] "movement_balance"          "power_shot_power"
##  [61] "power_jumping"             "power_stamina"
##  [63] "power_strength"            "power_long_shots"
##  [65] "mentality_aggression"      "mentality_interceptions"
##  [67] "mentality_positioning"     "mentality_vision"
##  [69] "mentality_penalties"       "mentality_composure"
```

```
##  [71] "defending_marking"          "defending_standing_tackle"
##  [73] "defending_sliding_tackle"   "goalkeeping_diving"
##  [75] "goalkeeping_handling"       "goalkeeping_kicking"
##  [77] "goalkeeping_positioning"    "goalkeeping_reflexes"
##  [79] "ls"                         "st"
##  [81] "rs"                         "lw"
##  [83] "lf"                         "cf"
##  [85] "rf"                         "rw"
##  [87] "lam"                        "cam"
##  [89] "ram"                        "lm"
##  [91] "lcm"                        "cm"
##  [93] "rcm"                        "rm"
##  [95] "lwb"                        "ldm"
##  [97] "cdm"                        "rdm"
##  [99] "rwb"                        "lb"
## [101] "lcb"                        "cb"
## [103] "rcb"                        "rb"
```

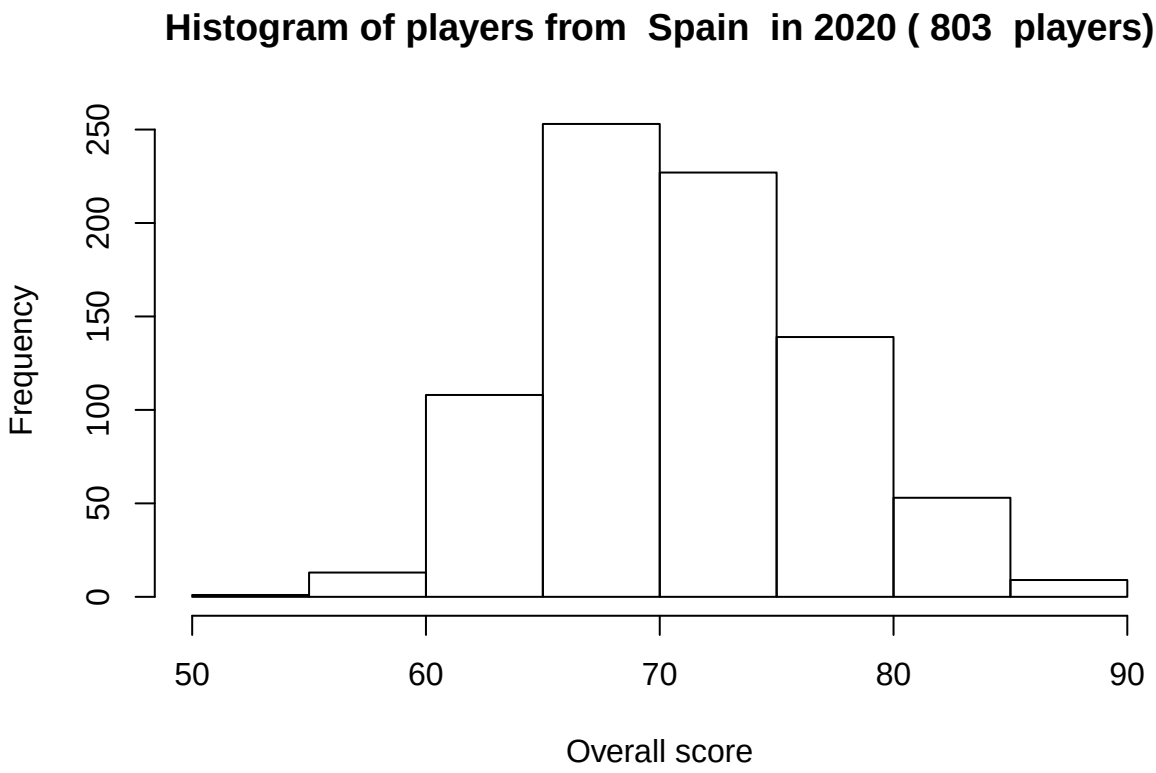
Spojimo podatke temljem jedinstvenog ključa svakog igrača i provjerimo normalnost podataka:

```
merged_df = merge(fifa19, fifa20, by="sofifa_id", suffixes = c(".19",".20"))
```

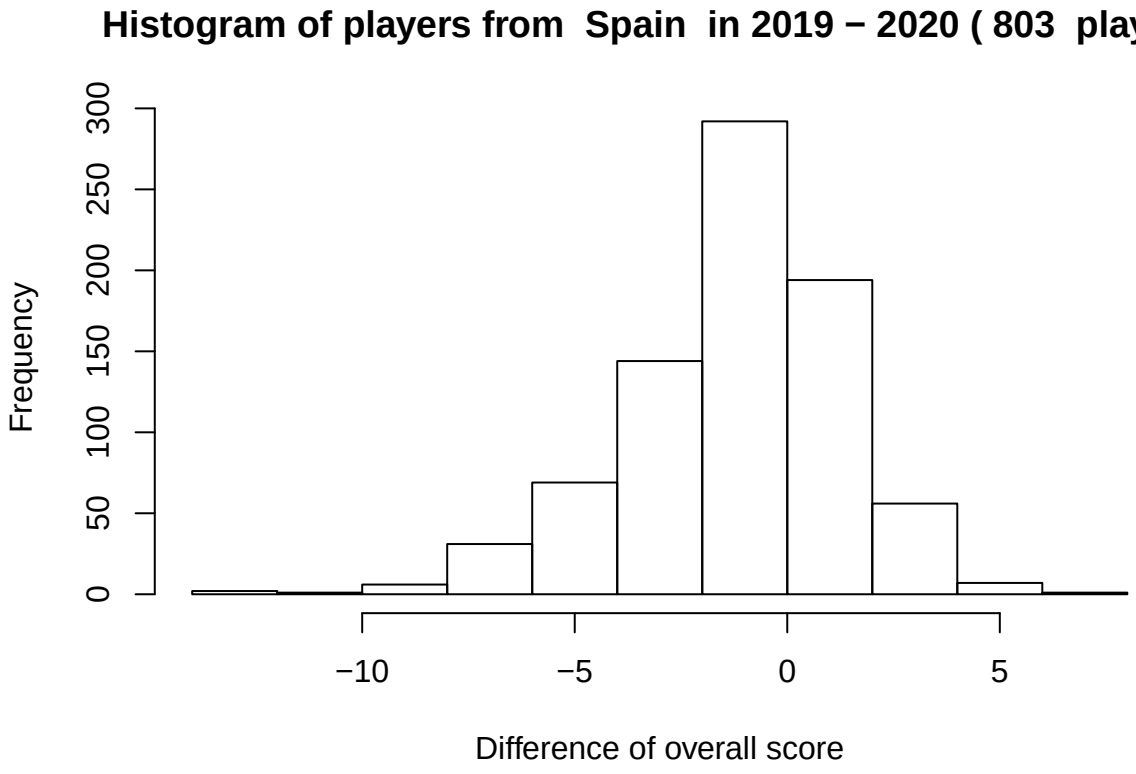
```
country = 'Spain'

len = length(merged_df[merged_df$nationality.19 == country,]$overall.19 -  merged_df[merged_df$national

hist(merged_df[merged_df$nationality.19 == country,]$overall.19,
     main=paste('Histogram of players from ',country,' in 2019 (',len,' players)'),
     xlab='Overall score')
```



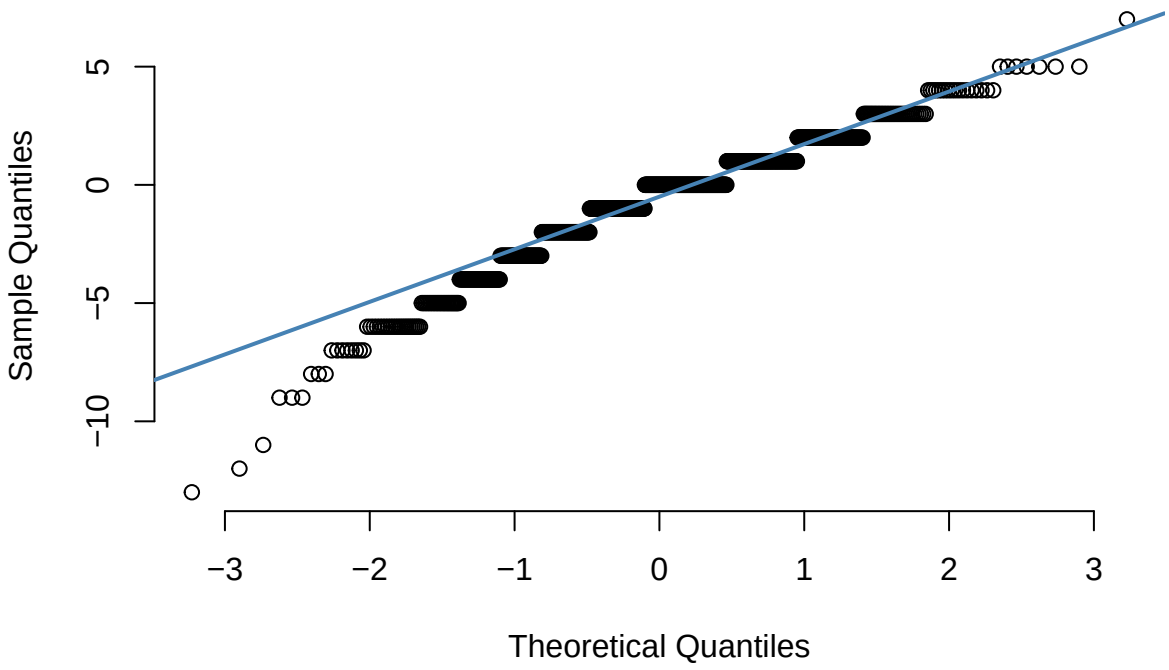
```
hist(merged_df[merged_df$nationality.19 == country,]$overall.20,
     main=paste('Histogram of players from ',country,' in 2020 (',len,' players)'),
     xlab='Overall score')
```




```
hist(merged_df[merged_df$nationality.19 == country,]$overall.19 -
       merged_df[merged_df$nationality.19 == country,]$overall.20,
     main=paste('Histogram of players from ',country,' in 2019 - 2020 (',len,' players)'),
     xlab='Difference of overall score')
```




```
qqnorm(merged_df[merged_df$nationality.19 == country,]$overall.19 - merged_df[merged_df$nationality.19 =
       pch = 1,
       frame = FALSE,
       main=paste('QQ-plot for overall score of players from',country,' (',len,'players)'))
qqline(merged_df[merged_df$nationality.19 == country,]$overall.19 - merged_df[merged_df$nationality.19 =
       col = "steelblue", lwd = 2)
```



Histogram razlika nam sugerira normalnost podataka, dok iz qq-plota vidimo malo odstupanje lijevog repa. Pod pretpostavkom da su podatci normalni, koristimo upareni t-test.

```
t.test(merged_df[merged_df$nationality.19 == country,]$overall.19,
       merged_df[merged_df$nationality.19 == country,]$overall.20,
       paired = TRUE,
       alt = "less")
```

```
##
##  Paired t-test
##
## data:  merged_df[merged_df$nationality.19 == country, ]$overall.19 and merged_df[merged_df$nationali
## t = -7.5589, df = 802, p-value = 5.557e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5366892
## sample estimates:
## mean of the differences
##              -0.6861768
```

Jako mala p-vrijednost nam ukazuje da postoji statistički značajna razlika u "overall score-u" svakog španjolskog igrača, što znači da se njihova kvaliteta promijenila iz sezone 2019 u sezonu 2020.

Napomenimo kako statistička značajnost ne nalaže nužno i praktičnost u pravom svijetu. Vidimo da je razlika u prosječnim "overall" vrijednostima (na skali od 0 do 100) koja opisjue naše igrače porasla za 0.68 iz jedne u drugu sezonu. U slučaju da se ta promjena desila zbog promjene trenera, koji je sada na primjer puno skuplji od prošloga, takva promjena ne bi nužno bila isplativa za klub. Ovakve situacije su česte u slučajevima kada imamo jako velik broj primjera (puno veći nego ovdje), stoga svakako trebamo obratiti pozornost i na smislenost zaključaka nakon testiranja.

Je li preferirana noga za udarac nezavisna od toga nalazi li pozicija igrača lijevo, desno ili centralno na terenu?

Programski paket R nudi široku podršku za rad s kategorijskim podatcima - od factor tipa podataka, do raznih statističkih alata i testova za analizu kategorijskih varijabli. Kod analize factor tipa podataka, moramo paziti koje su moguće vrijednosti (`levels()`) naše varijable, a koje se stvarno pojavljuju u podatcima.

```
levels(fifa19$preferred_foot)
```

```
## [1] "Left"  "Right"
```

```
levels(fifa19$team_position)
```

```
##  [1] ""    "CAM" "CB"  "CDM" "CF"  "CM"  "GK"  "LAM" "LB"  "LCB" "LCM" "LDM"
## [13] "LF"  "LM"  "LS"  "LW"  "LWB" "RAM" "RB"  "RCB" "RCM" "RDM" "RES" "RF"
## [25] "RM"  "RS"  "RW"  "RWB" "ST"  "SUB"
```

```
table(fifa19$preferred_foot)
```

```
##
##  Left Right
##  4131 13639
```

```
table(fifa19$team_position)
```

```
##
##       CAM   CB  CDM   CF   CM   GK  LAM   LB  LCB  LCM  LDM   LF   LM   LS   LW
##  223  318   83  144   13   62  642   19  558  638  384  247   10  435  190  143
##  LWB  RAM   RB  RCB  RCM  RDM  RES   RF   RM   RS   RW  RWB   ST  SUB
##   36   19  555  638  387  246 2928   10  434  191  142   37  445 7593
```

U ovom slučaju vidimo da se sve moguće vrijednosti pojavljuju u podatcima.

Sada možemo združiti podatke ovisno da li je njihova pozicija desno, lijevo ili na centru terena. Pozicije igrača imaju raspodjelu kao na slici.

```
knitr::include_graphics("team_positions.png")
```

Kopirajmo najprije podatke u novi data.frame kako ne bi promijenili prave vrijednosti.

```
fifa19_copy = data.frame(fifa19)
tracemem(fifa19)==tracemem(fifa19_copy)
```

```
## [1] FALSE
```

```
untracemem(fifa19_copy)
untracemem(fifa19_copy)
```

Kako bi lakše baratali sa varijablom "team_position", možemo je pretvoriti u "character" tip podatka.

```
fifa19_copy['team_position'] <- sapply(fifa19_copy['team_position'], as.character);
```

Združimo sve razrede centralnih, lijevih i desnih pozicija.

```
# ZDRUŽIMO SVE CENTRALNE POZICIJE
for (column_name in c("ST","CF","CAM","CM","CDM","CB")){
  fifa19_copy$team_position[fifa19_copy$team_position == column_name] = "Central_positions";
}

# ZDRUZIMO SVE LIJEVE POZICIJE
for (column_name in c("LS","LW","LF","LAM","LM","LCM", "LWB", "LDM", "LB", "LCB")){
```

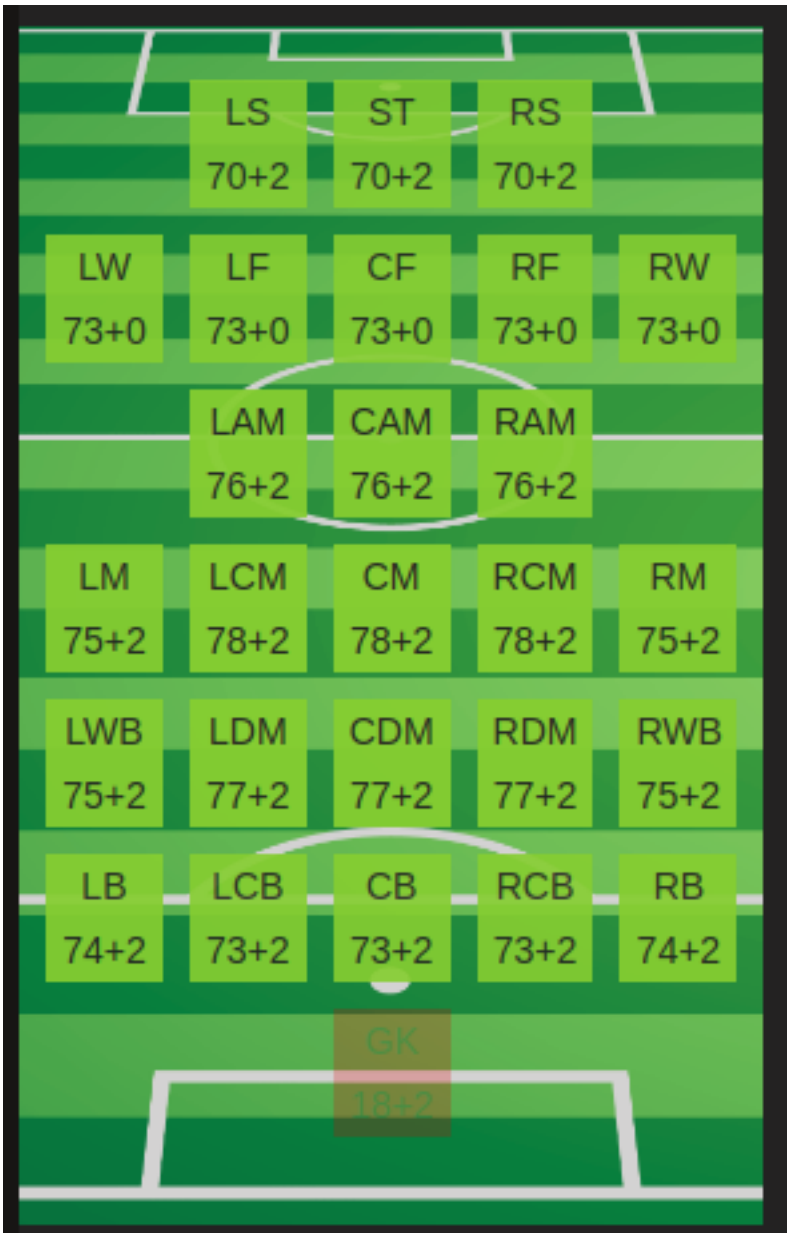


Figure 1: Klasifikacija pozicije igrača na terenu temeljem sofifa.com stranice odakle su preuzeti podaci

```
  fifa19_copy$team_position[fifa19_copy$team_position == column_name] = "Left_positions";
}

# ZDRUZIMO SVE DESNE POZICIJE
for (column_name in c("RS","RF","RW","RAM", "RCM","RM", "RDM", "RWB", "RCB", "RB")){
  fifa19_copy$team_position[fifa19_copy$team_position == column_name] = "Right_positions";
}
```

Kontingencijsku tablicu jedne kategorijske varijable moguće je dobiti pozivanjem funkcije `table()`.

```
tbl = table(fifa19_copy$team_position)
print(tbl)
```

```
##
##                   Central_positions                  GK    Left_positions
##                 223                1065                642                2660
##                 RES   Right_positions                SUB
##                2928                2659               7593
```

Izbacimo poziciju vratara i zamjenskih igrača i pogledajmo kontingencijsku tablicu varijabli pozicije i preferirane noge za udarce.

```
tbl = table(fifa19_copy[fifa19_copy$team_position == "Central_positions" | fifa19_copy$team_position ==
            fifa19_copy[fifa19_copy$team_position == "Central_positions" | fifa19_copy$team_position ==
tbl
```

```
##
##                      Left Right
##   Central_positions   188   877
##   Left_positions     1180  1480
##   Right_positions     312  2347
```

Kontingencijskoj tablici možemo dodati i sume redaka i stupaca na sljedeći način.

```
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##                      Left Right  Sum
##   Central_positions   188   877 1065
##   Left_positions     1180  1480 2660
##   Right_positions     312  2347 2659
##   Sum                1680  4704 6384
```

Test nezavisnosti χ^2 test u programskom paketu R implementiran je u funkciji `chisq.test()` koja kao ulaz prima kontingencijsku tablicu podataka koje testiramo na nezavisnost. Ispitajmo nezavisnost pozicije igrača na terenu i njegove preferirane noge za udarce.

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).

```
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',(added_margins_tbl[row_names,
    }
  }
}
```

```
## Očekivane frekvencije za razred  Left - Central_positions :  280.2632
## Očekivane frekvencije za razred  Left - Left_positions :  700
## Očekivane frekvencije za razred  Left - Right_positions :  699.7368
## Očekivane frekvencije za razred  Right - Central_positions :  784.7368
## Očekivane frekvencije za razred  Right - Left_positions :  1960
## Očekivane frekvencije za razred  Right - Right_positions :  1959.263
```

Sve očekivane frekvencije su veće od 5. Možemo nastaviti sa χ^2 testom.

```
chisq.test(tbl,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 779.5, df = 2, p-value < 2.2e-16
```

Odbacujemo H0 u korist H1 koja kaže da je pozicija igrača na terenu i njegova preferirana noga za udarce zavisna.

Još neka zanimljiva pitanja koje možete razmotriti su:

1. Je li pozicija igrača na terenu nezavisna od tjelesnog tipa igrača?
2. Jesu li braniči bolji u kontroliranju lopte ("skill_ball_control") od napadača?
3. Imaju li braniči bolje vještine u driblanju ("dribbling") od napadača?