R Notebook

Code

Hide

```
#nejake kniznice
library(dplyr)
library(tidyverse)
library(data.table)
library(corrplot)
library(RColorBrewer)
library(rcompanion)
library(lawstat)
library(moments)
library(forcats)
library(car)
library(ggpubr)
```

Dáta

V tejto sekcii bude opísaný základný zdroj dát + postup, akým sme získali náš unikátny dataset - bolo potrebné vykonať viacero krokov, ktoré zahŕňali spájanie viacerých datasetov, elimináciu niektorých atribútov, deduplikáciu a na záver vytváranie úplne nových relevantných atribútov na základe hodnôt niektorých atribútov pôvodného datasetu.

Tabuľka game

Tabuľka obsahuje základné dáta jednotlivých zápasov. Neobsahuje však detailné informácie o tímoch alebo tímových štatistikách konkrétneho zápasu. Taktiež neobsahuje čitateľné identifikátory tímov (t.j. ich názvy, obsahuje iba ID) Pomocou atribútov game_id, away_team_id a home_team_id však dokážeme namapovať iné tabuľky tak, aby sme dostali jeden dataset so všetkými pre nás relevantnými hodnotami.

Hide

```
#setwd(dir)
path <- ('data/game.csv')
game <- fread(path)
head(game)
```

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>
2016020045	20162017	R	2016-10-19 00:30:00	4	16	4
2017020812	20172018	R	2018-02-07 00:00:00	24	7	4
2015020314	20152016	R	2015-11-24 01:00:00	21	52	4
2015020849	20152016	R	2016-02-17 00:00:00	52	12	1

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>
2017020586	20172018	R	2017-12-30 03:00:00	20	24	1
2016020610	20162017	R	2017-01-10 00:30:00	15	8	4

6 rows | 1-8 of 15 columns

Niektoré atribúty sú z pohľadu štatistiky nezaujímavé a teda ich môžeme odstrániť pred spájaním tabuliek. Napríklad údaje o časovej zóne, alebo identifikátory štadiónov a pododobne. Pri týchto atribútoch predpokladáme, že pre naše riešenie zaujímavé žiadnym spôsobom nebudú - samozrejme, využitie by sa určite nájsť dalo, no aj vzhľadom na predpokladaný počet atribútov nemáme potrebu tieto atribúty v datasete ponechať.

Hide

```
game <- subset(game, select=-c(venue_link,venue_time_zone_id, venue_time_zone_offset,
venue_time_zone_tz))
head(game)
```

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>
2016020045	20162017	R	2016-10-19 00:30:00	4	16	4
2017020812	20172018	R	2018-02-07 00:00:00	24	7	4
2015020314	20152016	R	2015-11-24 01:00:00	21	52	4
2015020849	20152016	R	2016-02-17 00:00:00	52	12	1
2017020586	20172018	R	2017-12-30 03:00:00	20	24	1
2016020610	20162017	R	2017-01-10 00:30:00	15	8	4

6 rows | 1-8 of 11 columns

Hide

```
dim(game)
```

```
[1] 26305    11
```

Hide

```
dim(game[game$season==20192020,])
```

```
[1] 2425   11
```

Hide

```
dim(subset(game, season==20192020 & type=='P'))
```

```
[1] 231  11
```

Prvotnou analýzou sme zistili, že v minulej sezóne sa odohralo približne 2500 zápasov, z čoho bolo +-10 percent (231 zápasov) v playoff. Percentuálny pomer zápasov regulárnej a vyraďovacej časti nie je vhodný pre rozdelenie na trénovací a testovací dataset (tam je odporúčaný pomer 70:30, resp. 80:20). Náš prvotný predpoklad, že môžeme dataset trénovať na regulárnej sezóne a následne testovať na zápasoch play-off teda nebude ideálny - vhodnejší bude možno prístup rozdelenia podľa sezón, kedy na zápasoch starších sezón model natrénujeme a testovať ho budeme na novšej sezóne.

Počet hier v minulej sezóne bol omnoho vyšší ako je zvykom, tak sme skontrolovali existenciu duplikátnych záznamov v tabuľke. Z nasledujúceho výpisu sme odhalili 2570 duplikátov v celej tabuľke.

Hide

```
game[duplicated(game)]
```

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>
2019020001	20192020	R	2019-10-02 23:00:00	9	10	3
2019020002	20192020	R	2019-10-03 00:00:00	15	19	3
2019020003	20192020	R	2019-10-03 02:00:00	23	22	2
2019020004	20192020	R	2019-10-03 02:30:00	28	54	1
2019020005	20192020	R	2019-10-03 23:00:00	13	14	2
2019020006	20192020	R	2019-10-03 23:00:00	52	3	4
2019020007	20192020	R	2019-10-03 23:00:00	7	5	3
2019020008	20192020	R	2019-10-03 23:00:00	8	12	3
2019020009	20192020	R	2019-10-04 00:00:00	30	18	2
2019020010	20192020	R	2019-10-04 00:30:00	6	25	2

1-10 of 2,570 rows | 1-8 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Aby sme si overili existenciu duplikátov, tak sme si vypísali náhodnú hru z vyššie vygenerovanej tabuľky duplikátov.

Hide

```
game[game$game_id==2019020369]
```

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>	home_g <int>
2019020369	20192020	R	2019-11-26	9	29	0	

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>	home_g <int>
2019020369	20192020	R	2019-11-26	9	29	0	

2 rows | 1-8 of 11 columns

Duplikáty sme eliminovali využitím funkcie distinct, ktorá zachová jedinečné záznamy. Elimináciu duplikátov sme overili opäť použitím funkcie duplicated, ktorá vrátila 0 záznamov. Počet záznamov v tabuľke klesol o 2570, aktuálny počet záznamov je teda 23735.

Hide

```
game <- distinct(game)
game[duplicated(game)]
```

0 rows | 1-9 of 11 columns

Hide

```
dim(game)
```

```
[1] 23735    11
```

Tabuľka game_teams_stats

Druhá tabuľka obsahuje špecifické štatistiky tímov k jednotlivým zápasom. Problémom je, že jeden záznam obsahuje štatistiky iba jedného tímu. Teda máme pre jeden záznam zápasu z tabuľky games dva záznamy v tabuľke game_teams_stats.

Hide

```
path <- ('data/game_teams_stats.csv')
game_teams_stats <- fread(path)
head(game_teams_stats)
```

game_id <int>	team_id <int>	H… <chr>	won <lgl>	settled_in <chr>	head_coach <chr>	goals <int>	shots <int>	hits <int>	pin <int>
2016020045	4	away	FALSE	REG	Dave Hakstol	4	27	30	6
2016020045	16	home	TRUE	REG	Joel Quenneville	7	28	20	8
2017020812	24	away	TRUE	OT	Randy Carlyle	4	34	16	6
2017020812	7	home	FALSE	OT	Phil Housley	3	33	17	8
2015020314	21	away	TRUE	REG	Patrick Roy	4	29	17	9
2015020314	52	home	FALSE	REG	Paul Maurice	1	21	22	11

6 rows | 1-10 of 17 columns

Vyhodili sme atribúty, ktoré nám nepomôžu pri štatistickom vyhodnocovaní a vypísali sme si informácie o datasete. Všimli sme si, že údajov je viac ako dva-krát viac oproti záznamom v tabuľke game. To znamená, že niektoré záznamy sú duplicitné, alebo chybné.

Hide

```
game_teams_stats <- subset(game_teams_stats, select=-c(head_coach, startRinkSide, goa
ls))
head(game_teams_stats)
```

game_id <int>	team_id <int>	H… <chr>	won <lgl>	settled_in <chr>	sh… <int>	hits <int>	pir <int>	powerPlayOpportunities <int>	power
2016020045	4	away	FALSE	REG	27	30	6	4	
2016020045	16	home	TRUE	REG	28	20	8	3	
2017020812	24	away	TRUE	OT	34	16	6	3	
2017020812	7	home	FALSE	OT	33	17	8	2	
2015020314	21	away	TRUE	REG	29	17	9	3	
2015020314	52	home	FALSE	REG	21	22	11	2	

6 rows | 1-10 of 14 columns

Hide

```
dim(game_teams_stats)
```

```
[1] 52610    14
```

Podľa unikátnych identifikátorov zápasov vidíme, že ich je rovnako ako v tabuľke game. To znamená, že v tabuľke game_teams_stats máme duplikáty, ktoré budeme musieť odstrániť.

Hide

```
length(unique(game_teams_stats$game_id))
```

```
[1] 23735
```

Tabuľka team_info

Táto tabuľka obsahuje iba základné infromácie o tímoch, ktoré nesúvisia so sezónami a zápasmi. Je však potrebná pre doplnenie mien tímov k jednotlivým hrám - tým spôsobom bude možné priradenie názvov tímov k záznamom a možné prípadné overenie zmysluplnosti hodnôt atribútov podľa reálnych výsledkov napr. na NHL portáli.

Hide

```
path <- ('data/team_info.csv')
team_info <- fread(path)
head(team_info)
```

team_id <int>	franchiseId <int>	shortName <chr>	teamName <chr>	abbreviation <chr>	link <chr>
1	23	New Jersey	Devils	NJD	/api/v1/teams/1
4	16	Philadelphia	Flyers	PHI	/api/v1/teams/4
26	14	Los Angeles	Kings	LAK	/api/v1/teams/26
14	31	Tampa Bay	Lightning	TBL	/api/v1/teams/14
6	6	Boston	Bruins	BOS	/api/v1/teams/6
3	10	NY Rangers	Rangers	NYR	/api/v1/teams/3

6 rows

Z tejto tabuľky nám stačí extrahovať atribúty team_id, podľa ktorého spojíme tabuľky a abbreviation, ktorý nám doplní skratky názvov tímov ku zápasom - plné mená tímov potrebné nie sú, ako aj franchiseID alebo link na API.

Hide

```
team_info <- subset(team_info, select=-c(franchiseId, link, shortName, teamName))
head(team_info)
```

team_id <int>	abbreviation <chr>
1	NJD
4	PHI
26	LAK
14	TBL
6	BOS
3	NYR

6 rows

Hide

```
dim(team_info)
```

```
[1] 33  2
```

Spájanie tabuliek

V tejto sekcii vykonáme spojenie predom opisovanej trojice tabuliek, resp. troch datasetov do jedného, finálneho datasetu. Ten bude základom pre náš projekt, pričom na ňom budeme vykonávať EDA, MEDA, štatistické učenie, dokazovanie hypotéz a podobne.

Hide

```
game_teams_stats <- left_join(game_teams_stats, team_info, "team_id")
head(game_teams_stats)
```

game_id <int>	team_id <int>	H… <chr>	won <lgl>	settled_in <chr>	sh… <int>	hits <int>	pir <int>	powerPlayOpportunities <int>	power
2016020045	4	away	FALSE	REG	27	30	6	4	
2016020045	16	home	TRUE	REG	28	20	8	3	
2017020812	24	away	TRUE	OT	34	16	6	3	
2017020812	7	home	FALSE	OT	33	17	8	2	
2015020314	21	away	TRUE	REG	29	17	9	3	
2015020314	52	home	FALSE	REG	21	22	11	2	

6 rows | 1-10 of 15 columns

Hide

```
dim(game_teams_stats)
```

```
[1] 52610    15
```

Unikátny počet ID tímov v datasete je 37 (výpis nižšie)

Hide

```
length(unique(game_teams_stats$team_id))
```

```
[1] 37
```

Reálny počet tímov v NHL je však 31, čiže dataset obsahuje tímy navyše - táto skutočnosť je jednoducho vysvetliteľná, keďže dataset obsahuje zápasy od sezóny 2000, v ktorej hrávali tímy ktoré v dnešnej dobe už neexistujú (napr. Atlanta Trashers). Nápodobne, niektoré dnešné tímy v minulosti ešte neexistovali (napr. Las Vegas Golden Knights).

Hide

```
df_h <- game_teams_stats[(game_teams_stats$HoA == 'home'),]
head(df_h)
```

game_id <int>	team_id <int>	H… <chr>	won <lgl>	settled_in <chr>	sh… <int>	hits <int>	pir <int>	powerPlayOpportunities <int>	power
2016020045	16	home	TRUE	REG	28	20	8	3	
2017020812	7	home	FALSE	OT	33	17	8	2	
2015020314	52	home	FALSE	REG	21	22	11	2	
2015020849	12	home	TRUE	REG	29	16	8	5	
2017020586	24	home	TRUE	REG	41	15	13	6	
2016020610	8	home	FALSE	REG	23	27	4	4	

6 rows | 1-10 of 15 columns

Hide

```
dim(df_h)
```

```
[1] 26305    15
```

Hide

```
df_h <- distinct(df_h)
unique(df_h$HoA)
```

```
[1] "home"
```

Hide

```
df_a <- game_teams_stats[(game_teams_stats$HoA == 'away'),]
head(df_a)
```

game_id <int>	team_id <int>	H… <chr>	won <lgl>	settled_in <chr>	sh… <int>	hits <int>	pir <int>	powerPlayOpportunities <int>	power
2016020045	4	away	FALSE	REG	27	30	6	4	
2017020812	24	away	TRUE	OT	34	16	6	3	
2015020314	21	away	TRUE	REG	29	17	9	3	
2015020849	52	away	FALSE	REG	21	21	10	4	
2017020586	20	away	FALSE	REG	23	20	19	3	
2016020610	15	away	TRUE	REG	39	19	8	2	

6 rows | 1-10 of 15 columns

Hide

```
dim(df_a)
```

```
[1] 26305    15
```

Hide

```
df_a <- distinct(df_a)
unique(df_a$HoA)
```

```
[1] "away"
```

Hide

```
df <- left_join(df_a, df_h, "game_id", suffix = c(".away", ".home"))
head(df)
```

game_id <int>	team_id.away <int>	HoA.a… <chr>	won.a… <lgl>	settled_in.away <chr>	shots.away <int>	hits.away <int>	pim.
2016020045	4	away	FALSE	REG	27	30	
2017020812	24	away	TRUE	OT	34	16	
2015020314	21	away	TRUE	REG	29	17	
2015020849	52	away	FALSE	REG	21	21	
2017020586	20	away	FALSE	REG	23	20	
2016020610	15	away	TRUE	REG	39	19	

6 rows | 1-8 of 29 columns

Hide

```
dim(df)
```

```
[1] 23735    29
```

Hide

```
df <- left_join(game, df, "game_id")
head(df)
```

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>
2016020045	20162017	R	2016-10-19 00:30:00	4	16	4
2017020812	20172018	R	2018-02-07 00:00:00	24	7	4

game_id <int>	season <int>	t… <chr>	date_time_GMT <S3: POSIXct>	away_team_id <int>	home_team… <int>	away_goals <int>
2015020314	20152016	R	2015-11-24 01:00:00	21	52	4
2015020849	20152016	R	2016-02-17 00:00:00	52	12	1
2017020586	20172018	R	2017-12-30 03:00:00	20	24	1
2016020610	20162017	R	2017-01-10 00:30:00	15	8	4

6 rows | 1-8 of 39 columns

Hide

```
dim(df)
```

```
[1] 23735    39
```

Hide

```
dim(df[df$season==20192020,])
```

```
[1] 1215   39
```

Hide

```
dim(subset(df, season==20192020 & type=='P'))
```

```
[1] 118  39
```

Hide

```
dim(df[df$season==20182019,])
```

```
[1] 1360   39
```

Hide

```
dim(subset(df, season==20182019 & type=='P'))
```

```
[1] 87 39
```

Hide

```
dim(df[df$season==20172018,])
```

```
[1] 1363   39
```

Hide

```
dim(subset(df, season==20172018 & type=='P'))
```

```
[1] 92 39
```

Hide

```
dim(df[df$season==20162017,])
```

```
[1] 1323   39
```

Hide

```
dim(subset(df, season==20162017 & type=='P'))
```

```
[1] 93 39
```

Hide

```
glimpse(df)
```

```
Rows: 23,735
Columns: 39
$ game_id                      <int> 2016020045, 2017020812, 2015020314, 2015020849, 2
017020586, ~
$ season                       <int> 20162017, 20172018, 20152016, 20152016, 20172018,
20162017, ~
$ type                         <chr> "R", "R", "R", "R", "R", "R", "R", "R", "R", "R",
"R", "R", ~
$ date_time_GMT                <dttm> 2016-10-19 00:30:00, 2018-02-07 00:00:00, 2015-1
1-24 01:00:~
$ away_team_id                 <int> 4, 24, 21, 52, 20, 15, 10, 23, 29, 22, 52, 2, 2,
52, 52, 16,~
$ home_team_id                 <int> 16, 7, 52, 12, 24, 8, 26, 24, 21, 5, 25, 26, 53,
53, 4, 1, 2~
$ away_goals                   <int> 4, 4, 4, 1, 1, 4, 1, 1, 0, 3, 4, 2, 2, 3, 2, 3,
3, 5, 7, 6, ~
$ home_goals                   <int> 7, 3, 1, 2, 2, 1, 2, 4, 2, 2, 1, 4, 3, 2, 5, 2,
4, 4, 3, 1, ~
$ outcome                      <chr> "home win REG", "away win OT", "away win REG", "h
ome win REG~
$ home_rink_side_start         <chr> "right", "left", "right", "right", "left", "righ
t", "right",~
$ venue                        <chr> "United Center", "KeyBank Center", "MTS Centre",
"PNC Arena"~
$ team_id.away                 <int> 4, 24, 21, 52, 20, 15, 10, 23, 29, 22, 52, 2, 2,
52, 52, 16,~
$ HoA.away                     <chr> "away", "away", "away", "away", "away", "away", "
away", "awa~
$ won.away                     <lgl> FALSE, TRUE, TRUE, FALSE, FALSE, TRUE, FALSE, FAL
SE, FALSE, ~
$ settled_in.away              <chr> "REG", "OT", "REG", "REG", "REG", "REG", "REG", "
REG", "REG"~
$ shots.away                   <int> 27, 34, 29, 21, 23, 39, 26, 20, 34, 36, 26, 24, 3
4, 27, 32, ~
$ hits.away                    <int> 30, 16, 17, 21, 20, 19, 24, 11, 12, 31, 22, 35, 2
3, 21, 32, ~
$ pim.away                     <int> 6, 6, 9, 10, 19, 8, 19, 27, 8, 8, 8, 4, 4, 7, 2,
8, 6, 6, 4,~
$ powerPlayOpportunities.away <int> 4, 3, 3, 4, 3, 2, 4, 5, 3, 4, 3, 2, 2, 2, 1, 5,
4, 3, 1, 4, ~
$ powerPlayGoals.away          <int> 2, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 2,
1, 0, 0, 1, ~
$ faceOffWinPercentage.away    <dbl> 50.9, 43.8, 45.7, 31.4, 54.7, 46.6, 56.9, 47.5, 5
7.6, 50.6, ~
$ giveaways.away               <int> 12, 7, 13, 4, 10, 8, 6, 13, 3, 5, 3, 12, 5, 3, 6,
5, 10, 8, ~
$ takeaways.away               <int> 9, 4, 5, 14, 4, 5, 2, 5, 8, 2, 2, 2, 3, 4, 11, 5,
3, 2, 8, 2~
$ blocked.away                 <int> 11, 14, 20, 16, 7, 24, 7, 15, 14, 17, 15, 17, 15,
11, 21, 6,~
$ abbreviation.away            <chr> "PHI", "ANA", "COL", "WPG", "CGY", "WSH", "TOR",
```

```
"VAN", "CBJ~
$ team_id.home                <int> 16, 7, 52, 12, 24, 8, 26, 24, 21, 5, 25, 26, 53,
53, 4, 1, 2~
$ HoA.home                    <chr> "home", "home", "home", "home", "home", "home", "
home", "hom~
$ won.home                    <lgl> TRUE, FALSE, FALSE, TRUE, TRUE, FALSE, TRUE, TRU
E, TRUE, FAL~
$ settled_in.home             <chr> "REG", "OT", "REG", "REG", "REG", "REG", "REG", "
REG", "REG"~
$ shots.home                  <int> 28, 33, 21, 29, 41, 23, 41, 35, 32, 41, 34, 28, 3
3, 25, 22, ~
$ hits.home                   <int> 20, 17, 22, 16, 15, 27, 23, 10, 20, 36, 22, 30, 2
5, 35, 32, ~
$ pim.home                    <int> 8, 8, 11, 8, 13, 4, 10, 15, 8, 8, 6, 4, 4, 9, 2,
10, 8, 6, 2~
$ powerPlayOpportunities.home <int> 3, 2, 2, 5, 6, 4, 2, 6, 3, 4, 4, 2, 2, 1, 1, 4,
3, 3, 2, 3, ~
$ powerPlayGoals.home         <int> 2, 1, 0, 2, 1, 1, 0, 3, 1, 1, 1, 0, 0, 0, 0, 1,
1, 1, 1, 0, ~
$ faceOffWinPercentage.home   <dbl> 49.1, 56.2, 54.3, 68.6, 45.3, 53.4, 43.1, 52.5, 4
2.4, 49.4, ~
$ giveaways.home              <int> 16, 5, 13, 12, 13, 12, 12, 11, 5, 10, 13, 9, 3,
9, 11, 7, 9,~
$ takeaways.home              <int> 8, 6, 7, 11, 4, 7, 2, 6, 5, 5, 2, 2, 4, 9, 10, 7,
12, 6, 17,~
$ blocked.home                <int> 9, 14, 9, 13, 21, 18, 14, 12, 13, 12, 14, 19, 8,
7, 20, 16, ~
$ abbreviation.home           <chr> "CHI", "BUF", "WPG", "CAR", "ANA", "MTL", "LAK",
"ANA", "COL~
```

Následne môžeme zo spojených datasetov odstrániť nepodstatné atribúty (atribútov je aj tak veľké množstvo, čize odstránenie nie dôležitých atribútov nám uľahčí prácu s datasetom). Taktiež si vytvoríme nový atribút reprezentujúci percentuálnu úspešnost zákrokov domáceho a hosťujúceho brankára, ktorú získame ako podiel striel na bránu - počet gólov vydelené celkovým počtom striel na bránu).

Hide

```
df <- subset(df, select=-c(game_id, date_time_GMT, outcome, home_rink_side_start, ven
ue, away_team_id, home_team_id, settled_in.away, HoA.away, HoA.home))
df <- rename(df, 'settled_in' = 'settled_in.home')
df <- rename(df, 'goals.away' = 'away_goals')
df <- rename(df, 'goals.home' = 'home_goals')
df$save_percentage.home = round((df$shots.away-df$goals.away)/df$shots.away, 4)
df$save_percentage.away = round((df$shots.home-df$goals.home)/df$shots.home, 4)
head(df)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20162017	R	4	7	4	FALSE	27	30	

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	R	4	3	24	TRUE	34	16	
20152016	R	4	1	21	TRUE	29	17	
20152016	R	1	2	52	FALSE	21	21	
20172018	R	1	2	20	FALSE	23	20	
20162017	R	4	1	15	TRUE	39	19	

6 rows | 1-9 of 31 columns

Hide

```
dim(df)
```

```
[1] 23735    31
```

Atribúty HoA.home a Hoa.away boli redundantné, keďže informáciu o domácom a hosťujúcom tíme máme zahrnutú v atribúte team_id.home a team_id.away.

Prieskumná analýza datasetu

Kedže cieľom práce je práca na predikčnom modeli, bude vhodné nepracovať s celým datasetom ale iba poslednými sezónami. Dôvôdom je častá obmena kádrov tímov NHL, čo môže mať za následok drasticky odlíšne výkony tímov medzi sezónami. Práca s veľkým množstvom sezón bude mať pravdepodobne za následok nízku presnosť modelu a neprehľadnosť grafov. Dátovú množinu sme teda zmenšili na posledných 5 sezón, čo nám poskytne presnejšie výsledky pre aktuálne pravidlá zámorského hokeja.

Hide

```
df <- df[df$season >= 20152016]
print(df)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20162017	R	4	7	4	FALSE	27	30	
20172018	R	4	3	24	TRUE	34	16	
20152016	R	4	1	21	TRUE	29	17	
20152016	R	1	2	52	FALSE	21	21	
20172018	R	1	2	20	FALSE	23	20	
20162017	R	4	1	15	TRUE	39	19	
20152016	R	1	2	10	FALSE	26	24	

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	R	1	4	23	FALSE	20	11	
20172018	R	0	2	29	FALSE	34	12	
20152016	R	3	2	22	TRUE	36	31	

1-10 of 6,582 rows | 1-9 of 31 columns Previous **1** 2 3 4 5 6 … 100 Next

charakteristika dát, ich rozdelenie a identifikácia problémov (+ riešenie).

Hide

```
glimpse(df)
```

```
Rows: 6,582
Columns: 31
$ season                      <int> 20162017, 20172018, 20152016, 20152016, 20172018,
20162017, ~
$ type                        <chr> "R", "R", "R", "R", "R", "R", "R", "R", "R", "R",
"R", "R", ~
$ goals.away                  <int> 4, 4, 4, 1, 1, 4, 1, 1, 0, 3, 4, 2, 2, 3, 2, 3,
3, 5, 7, 6, ~
$ goals.home                  <int> 7, 3, 1, 2, 2, 1, 2, 4, 2, 2, 1, 4, 3, 2, 5, 2,
4, 4, 3, 1, ~
$ team_id.away                <int> 4, 24, 21, 52, 20, 15, 10, 23, 29, 22, 52, 2, 2,
52, 52, 16,~
$ won.away                    <lgl> FALSE, TRUE, TRUE, FALSE, FALSE, TRUE, FALSE, FAL
SE, FALSE, ~
$ shots.away                  <int> 27, 34, 29, 21, 23, 39, 26, 20, 34, 36, 26, 24, 3
4, 27, 32, ~
$ hits.away                   <int> 30, 16, 17, 21, 20, 19, 24, 11, 12, 31, 22, 35, 2
3, 21, 32, ~
$ pim.away                    <int> 6, 6, 9, 10, 19, 8, 19, 27, 8, 8, 8, 4, 4, 7, 2,
8, 6, 6, 4,~
$ powerPlayOpportunities.away <int> 4, 3, 3, 4, 3, 2, 4, 5, 3, 4, 3, 2, 2, 2, 1, 5,
4, 3, 1, 4, ~
$ powerPlayGoals.away         <int> 2, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 2,
1, 0, 0, 1, ~
$ faceOffWinPercentage.away   <dbl> 50.9, 43.8, 45.7, 31.4, 54.7, 46.6, 56.9, 47.5, 5
7.6, 50.6, ~
$ giveaways.away              <int> 12, 7, 13, 4, 10, 8, 6, 13, 3, 5, 3, 12, 5, 3, 6,
5, 10, 8, ~
$ takeaways.away              <int> 9, 4, 5, 14, 4, 5, 2, 5, 8, 2, 2, 2, 3, 4, 11, 5,
3, 2, 8, 2~
$ blocked.away                <int> 11, 14, 20, 16, 7, 24, 7, 15, 14, 17, 15, 17, 15,
11, 21, 6,~
$ abbreviation.away           <chr> "PHI", "ANA", "COL", "WPG", "CGY", "WSH", "TOR",
"VAN", "CBJ~
$ team_id.home                <int> 16, 7, 52, 12, 24, 8, 26, 24, 21, 5, 25, 26, 53,
53, 4, 1, 2~
$ won.home                    <lgl> TRUE, FALSE, FALSE, TRUE, TRUE, FALSE, TRUE, TRU
E, TRUE, FAL~
$ settled_in                  <chr> "REG", "OT", "REG", "REG", "REG", "REG", "REG", "
REG", "REG"~
$ shots.home                  <int> 28, 33, 21, 29, 41, 23, 41, 35, 32, 41, 34, 28, 3
3, 25, 22, ~
$ hits.home                   <int> 20, 17, 22, 16, 15, 27, 23, 10, 20, 36, 22, 30, 2
5, 35, 32, ~
$ pim.home                    <int> 8, 8, 11, 8, 13, 4, 10, 15, 8, 8, 6, 4, 4, 9, 2,
10, 8, 6, 2~
$ powerPlayOpportunities.home <int> 3, 2, 2, 5, 6, 4, 2, 6, 3, 4, 4, 2, 2, 1, 1, 4,
3, 3, 2, 3, ~
$ powerPlayGoals.home         <int> 2, 1, 0, 2, 1, 1, 0, 3, 1, 1, 1, 0, 0, 0, 0, 1,
1, 1, 1, 0, ~
$ faceOffWinPercentage.home   <dbl> 49.1, 56.2, 54.3, 68.6, 45.3, 53.4, 43.1, 52.5, 4
```

```
2.4, 49.4, ~
$ giveaways.home              <int> 16, 5, 13, 12, 13, 12, 12, 11, 5, 10, 13, 9, 3,
9, 11, 7, 9,~
$ takeaways.home              <int> 8, 6, 7, 11, 4, 7, 2, 6, 5, 5, 2, 2, 4, 9, 10, 7,
12, 6, 17,~
$ blocked.home                <int> 9, 14, 9, 13, 21, 18, 14, 12, 13, 12, 14, 19, 8,
7, 20, 16, ~
$ abbreviation.home           <chr> "CHI", "BUF", "WPG", "CAR", "ANA", "MTL", "LAK",
"ANA", "COL~
$ save_percentage.home        <dbl> 0.8519, 0.8824, 0.8621, 0.9524, 0.9565, 0.8974,
0.9615, 0.95~
$ save_percentage.away        <dbl> 0.7500, 0.9091, 0.9524, 0.9310, 0.9512, 0.9565,
0.9512, 0.88~
```

V datasete máme 31 atribútov. Vyšší počet je ich najmä z dôvodu potreby štatistík o hre z pohľadu oboch tímov ktoré odohrali zápas (preto máme počet atribútov cca zdvojnásobený). V tejto sekcii práce sa budeme venovať analýze jednotlivých atribútov, resp. dvojíc atribútov.

Atribút season

charakteristika: atribút reprezentuje sezónu, v ktorej sa odohral NHL zápas. Dataset obsahuje sezóny od roku 2015-2020.

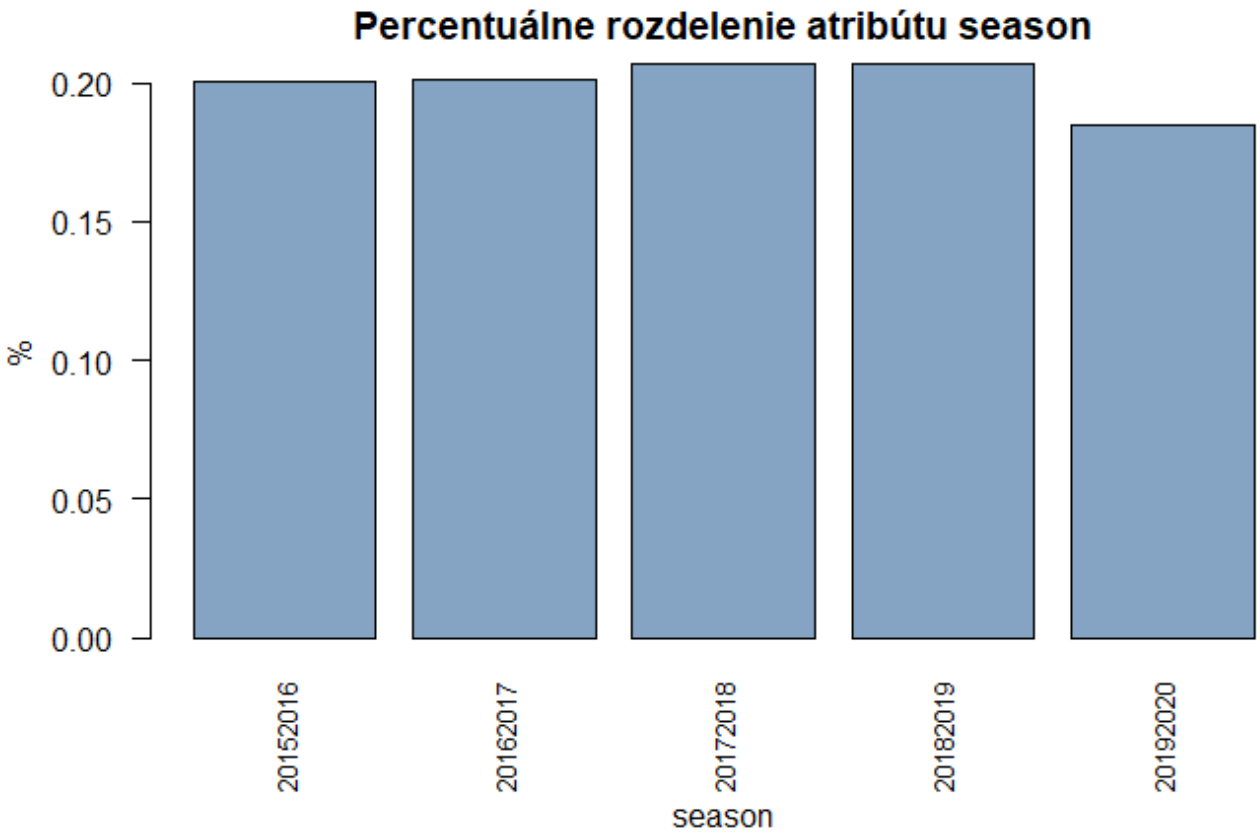
Hide

```
unique(df$season)
```

```
[1] 20162017 20172018 20152016 20192020 20182019
```

Hide

```
barplot(prop.table(table(df$season)), las=2, cex.names=.8, col=rgb(0.2,0.4,0.6,0.6),
ylab="%", xlab="", main="Percentuálne rozdelenie atribútu season")
mtext(text = "season", side = 1,line = 4)
```



Hide

```
table(df$season)
```

```
20152016 20162017 20172018 20182019 20192020
    1321     1323     1363     1360     1215
```

Z grafu môžeme vidieť, že dataset obsahuje 5 zvolených sezón. Rozdelenie počtu hier v sezónach je percentuálne približne rovnaké (+- 3%), pričom najmenej odohraných hier bolo v poslednej sezóne - pravdepodobne zapríčinené pandemickou situáciou. Posledná sezóna v datasete je ročník 2019/2020, keďže aktuálna sezóna 2020/2021 ešte nie je ukončená a teda nie je súčasťou datasetu.

Atribút type

charakteristika: atribút reprezentuje typ odohraného zápasu. Nadobúda tri základné hodnoty:

- R - regular, označuje regulárne zápasy ktoré sa hrajú počas NHL sezóny
- P - playoff, označuje zápasy vyraďovacej časti NHL, hrajú sa na konci sezóny. Tejto časti sa zúčastňujú iba najlepšie tímy.
- A - ?, predpokladáme že exhibičné zápasy

Hide

```
cat("Unikátne hodnoty: \n")
```

```
Unikátne hodnoty:
```

Hide

```
unique(df$type)
```

```
[1] "R" "P" "A"
```

Hide

```
cat(" \n Percentuálny podiel: \n")
```

```

 Percentuálny podiel:
```

Hide

```
cat("R = ", (100/nrow(df))*nrow(df[df$type=='R']), "%\n")
```

```
R =  92.61623 %
```

Hide

```
cat("P = ", (100/nrow(df))*nrow(df[df$type=='P']), "%\n")
```

```
P =  7.307809 %
```

Hide

```
cat("A = ", (100/nrow(df))*nrow(df[df$type=='A']), "%\n")
```

```
A =  0.07596475 %
```

Hide

```
cat(" \n Počet výskytov:")
```

```

 Počet výskytov:
```

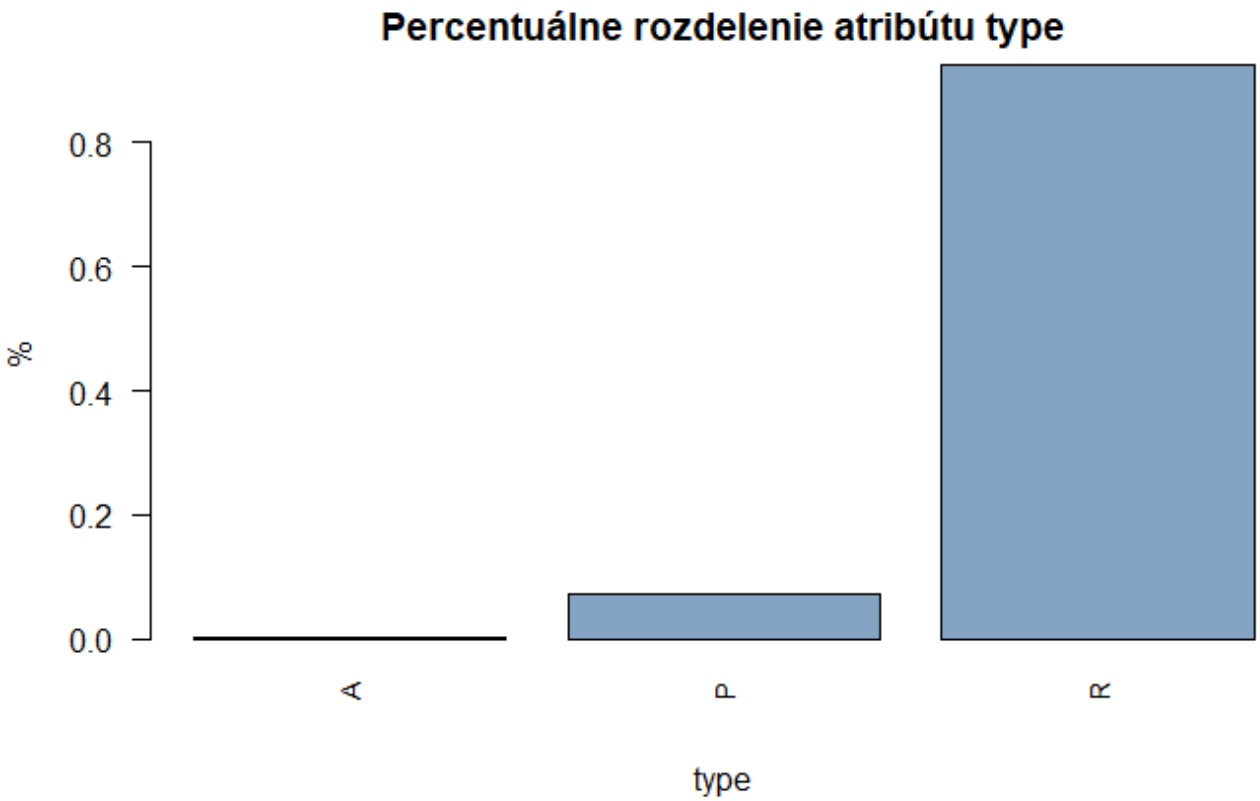
Hide

```
table(df$type)
```

```
  A    P    R
  5  481 6096
```

Hide

```
barplot(prop.table(table(df$type)), las=2, cex.names=.8, col=rgb(0.2,0.4,0.6,0.6), yl
ab="%", xlab="type", main="Percentuálne rozdelenie atribútu type")
```



Z početností jednotlivých atribútov môžeme vidieť obrovskú prevahu zápasov regulárnej časti sezóny, čo aj zodpovedá skutočnosti. Použitie regulárnej sezóny na realizáciu trénovacieho datasetu a playoff na realizáciu testovacieho datatsetu nie je vhodné, pretože percentuálny podiel 92:7 nesedí s odporúčaným pomerom 70:30, alebo 80:20. Z tohoto dôvodu bude potrebné dodatočné prerozdelenie zápasov regulárnej sezóny tak, aby bol tento pomer splnený - bude riešené neskôr v práci. Hodnotu NA nemá žiaden atribút.

Hodnotu ‘A’ nadobúda len 5 záznamov, v žiadnom z nich sa nejedná o zápasy tímov NHL (abbreviation = NA).

Hide

```
subset(df, type=='A')
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20192020	A	5	9	88	FALSE	16	0	
20192020	A	10	5	90	TRUE	28	0	

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20192020	A	4	5	87	FALSE	11	0	
20182019	A	7	4	88	TRUE	26	0	
20182019	A	10	5	88	TRUE	22	0	

5 rows | 1-9 of 31 columns

V zdroji datasetu nie je vysvetlené, čo hodnota “A” v atribúte type reprezentuje. Predpokladáme, že sa jedná o exhibičné zápasy, pretože atribúty tímov nie sú namapované na žiadne NHL tímy.

Atribúty goals.home a goals.away

charakteristika: atribút reprezentuje počet gólov z pohľadu domáceho a hosťovského tímu.

Nejedná sa o spojitý atribút - počet gólov môže nadobúdať iba celé hodnoty, jedná sa teda o diskrétny atribút. Už z podstaty atribútov **goals.gome** a **goals.away** vieme povedať, že sa nebude jednať o normálne rozdelenie, ale o distribúciu naklonenú vpravo (skewed right) - toto vieme poznamenať ešte pred vizualizáciou prostredníctvom histogramov. Taktiež bude pravdepodobne obsahovať vychýlené hodnoty, keďže sa zvyknú vyskytovať zápasy s nadmerným počtom gólov. Reálny priemer sa však zvykne pohybovať okolo hodnoty 2 góly pre tím - predpokladáme, že v datasete tomu tak bude tiež. Na začiatku si teda oba atribúty vizualizujeme prostredníctvom krabicových grafov a histogramov, aby sme mohli overiť naše iniciálne predpoklady.

Hide

```
cat("Základná štatistika pre goals.home", "\n")
```

```
Základná štatistika pre goals.home
```

Hide

```
summary(df$goals.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    2.00    3.00    3.01    4.00   10.00 
```

Hide

```
cat("\n","Základná štatistika pre goals.away", "\n")
```

```

 Základná štatistika pre goals.away
```

Hide

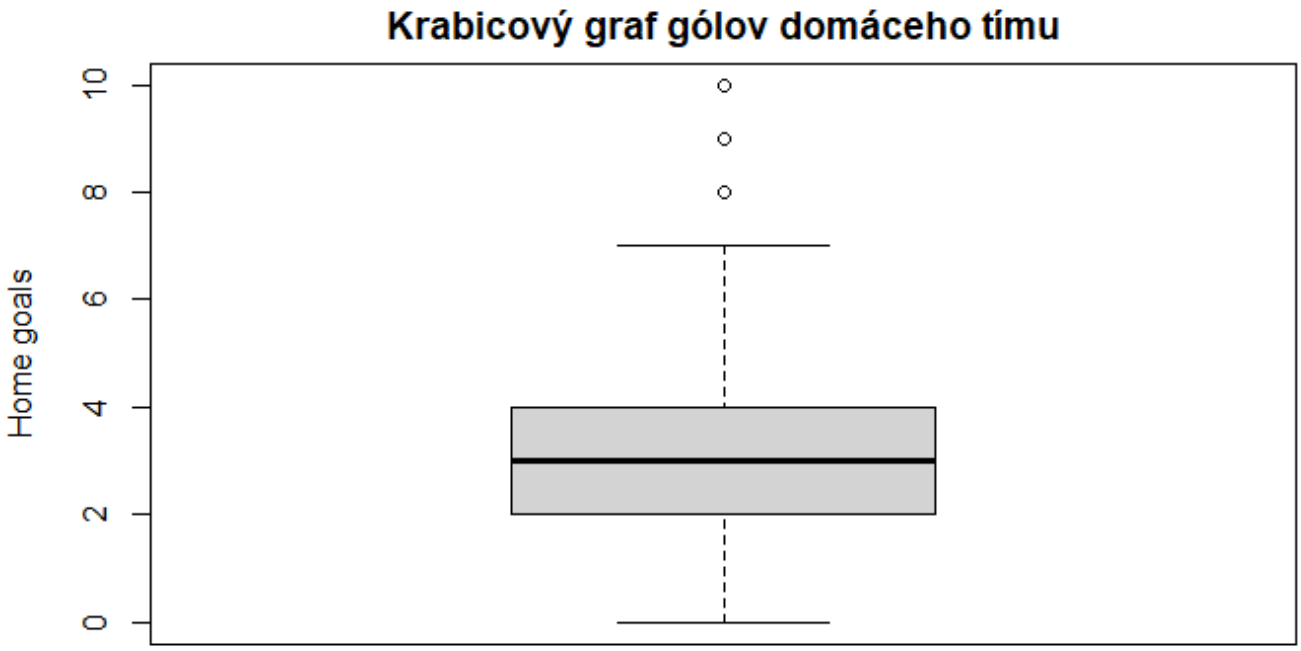
```
summary(df$goals.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   2.000   3.000   2.749   4.000  10.000 
```

Zo základnej štatistiky vieme povedať, že hodnoty sú koncentrované okolo počtu gólov 3 - konkrétne na základu priemeru a mediánu (menej citlivý na vychýlené hodnoty). Maximálny počet gólov pre domáce aj hosťujúce tímy je v oboch prípadoch 10. Podľa tretieho kvantilu = 4 vieme povedať, že hodnota 10 pôsobí vychýlene. Najvyššia koncentrácia vyzerá bude v pri hodnotách 2,3,4.

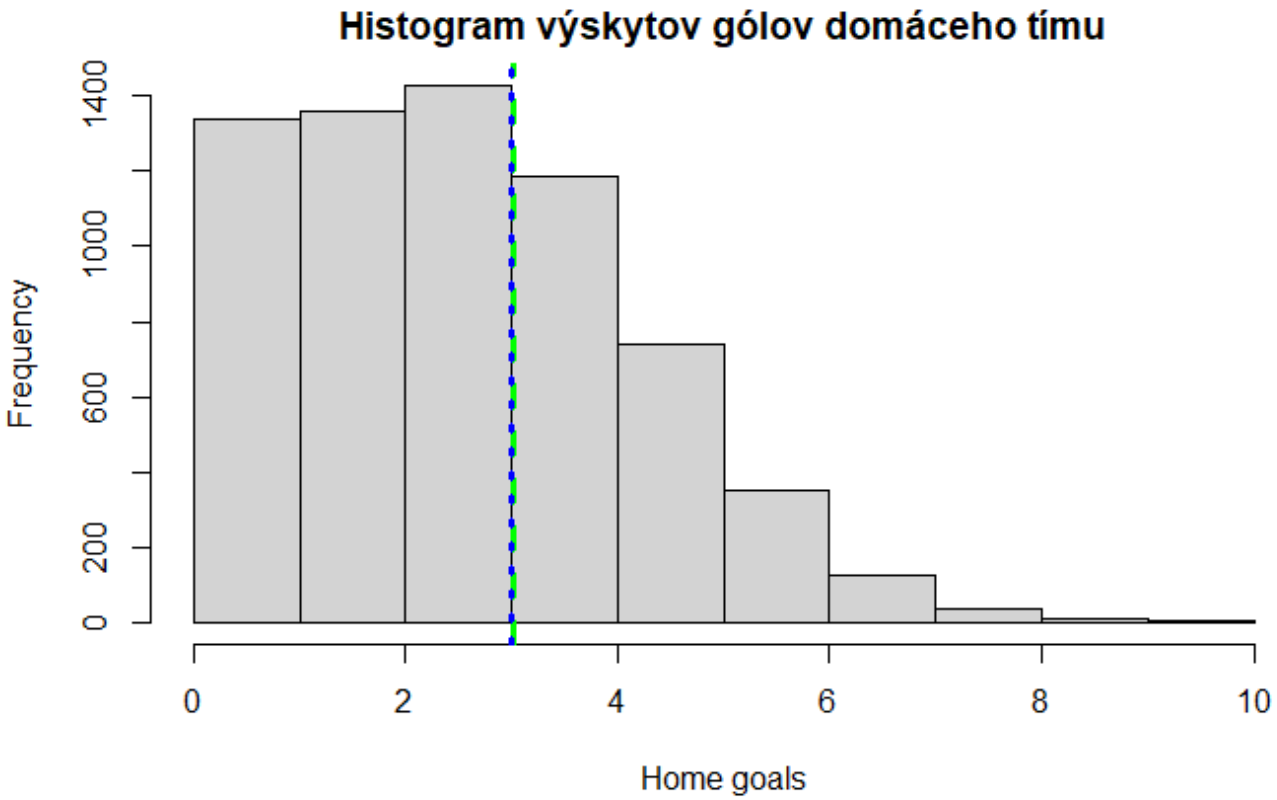
Hide

```
boxplot(df$goals.home, ylab="Home goals", xlab="", main="Krabicový graf gólov domáceh
o tímu")
```



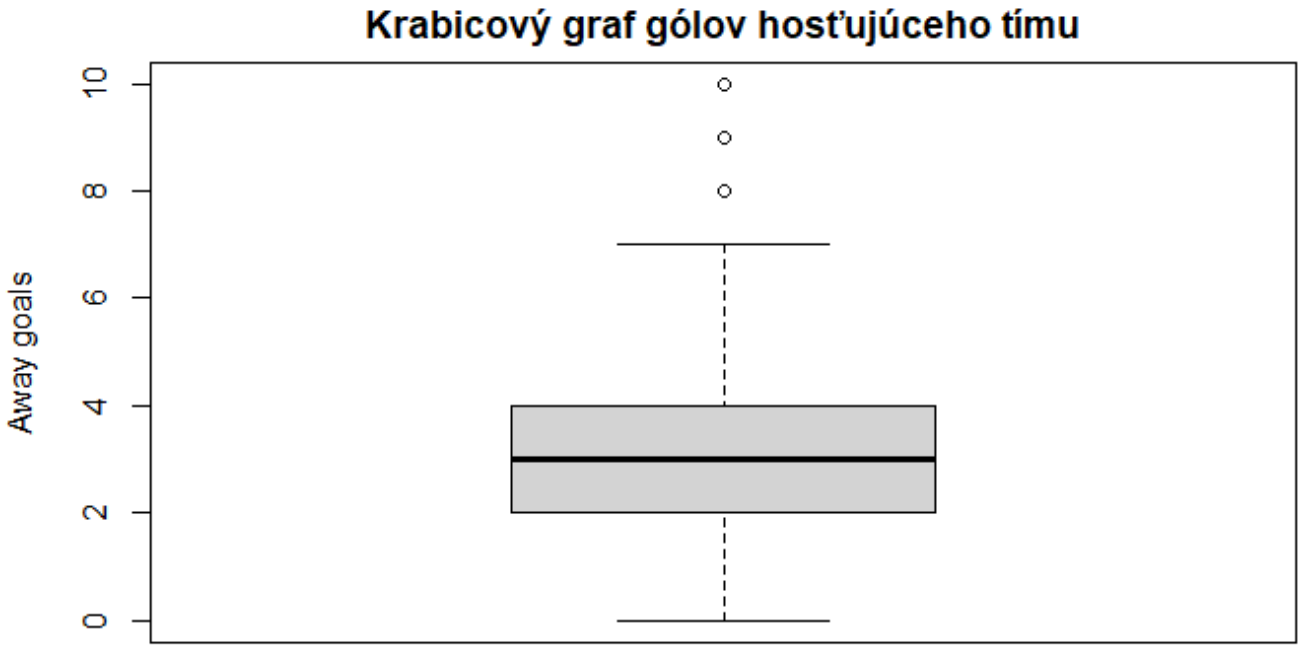
Hide

```
hist(df$goals.home, xlab="Home goals", main="Histogram výskytov gólov domáceho tímu")
abline(v = c(mean(df$goals.home),  median(df$goals.home)), col=c("green", "blue"), lt
y=c(2,3), lwd=c(3,3))
```



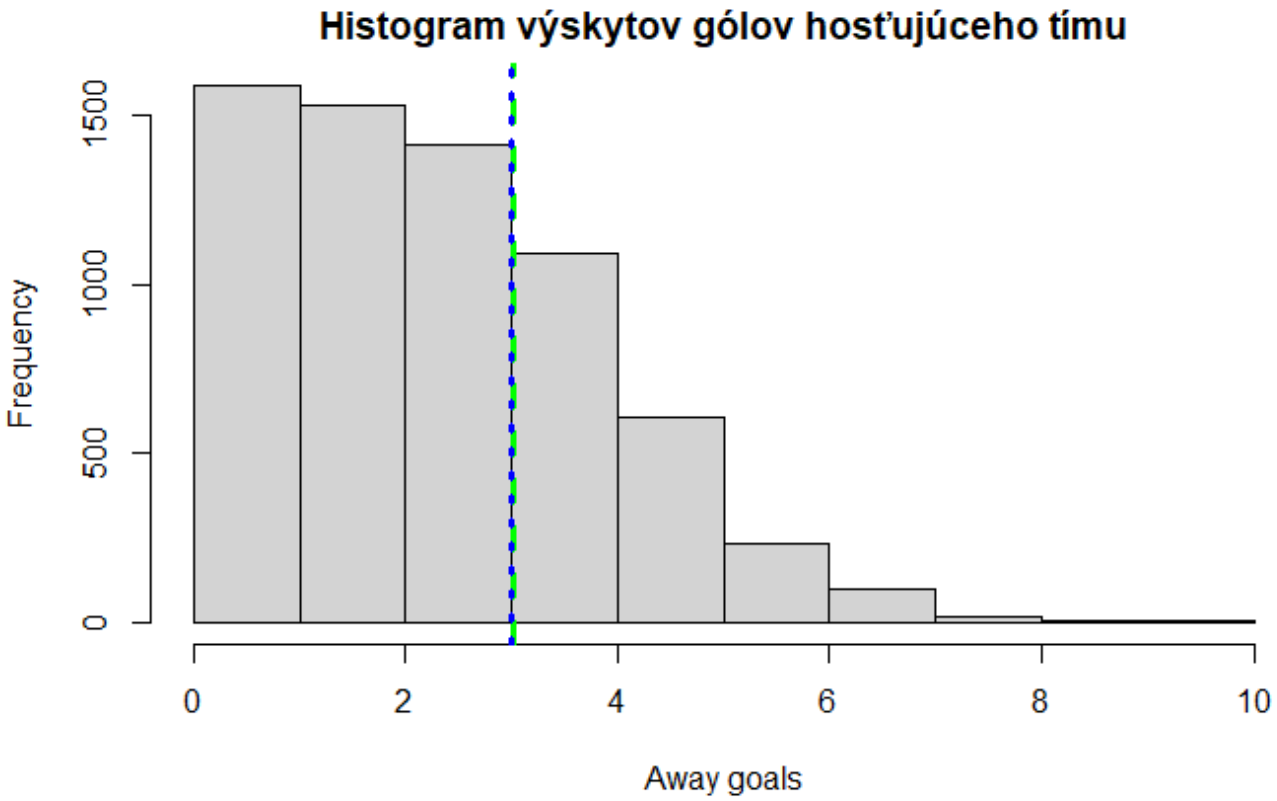
Hide

```
boxplot(df$goals.away, ylab="Away goals", xlab="", main="Krabicový graf gólov hosťujú
ceho tímu")
```



Hide

```
hist(df$goals.away, xlab="Away goals", main="Histogram výskytov gólov hosťujúceho tím
u")
abline(v = c(mean(df$goals.home),  median(df$goals.home)), col=c("green", "blue"), lt
y=c(2,3), lwd=c(3,3))
```



Dodatočne ešte overíme, či neexistujú chýbajúce hodnoty týchto atribútov v niektorých záznamoch:

Hide

```
cat("Počet záznamov kde je goals.home NA", nrow(subset(df, is.na(goals.home)==TRUE)),
'\n')
```

```
Počet záznamov kde je goals.home NA 0
```

Hide

```
cat("Počet záznamov kde je goals.away NA",nrow(subset(df, is.na(goals.away)==TRU
E)),'\n')
```

```
Počet záznamov kde je goals.away NA 0
```

Ani jeden záznam nenadobúda pri atribútoch goals.home a goals.away NA hodnoty, preto nebude potrebné riešiť voľbu stratégie dopĺňania týchto hodnôt.

Už z vizualizácií atribútov vidíme, že predpoklady boli korektné. Histogramy oboch atribútov sú veľmi podobné -

obe majú distribúciu naklonenú vpravo, určite sa nebude jednať o normálne rozdelenie - z tohoto dôvodu netreba ani vykonávať test normality. Z histogramov možno vyčítať, že domáci tím strelí v priemere viac gólov ako hosťujúci tím - zatiaľ čo pri hosťujúcich tímoch nadobúdajú záznamy najčastejšie hodnotu 0 gólov, pri domácich tímoch je maximálna frekvencia pri hodnote 2 góly. Z tejto skutočnosti môžeme vyvodiť hypotézu, že domáce prostredie môže mať reálny vplyv na priemerný počet strelených gólov, resp. vonkajšie prostredie znižuje priemerný počet strelených gólov. Pri krabicových grafoch možno sledovať úplnú identitu týchto grafov pre oba atribúty, čo je zaujímavý jav, keďže histogramy sú mierne odlišné. Väčšina hodnôt sa pohybuje v rozmedzí 2-4 góly, čiže náš počiatočný predpoklad bol čiastočne korektný. Vychýlené hodnoty sú v oboch prípadoch počet 8, 9 a 10 gólov. Pre overenie si tieto maximálne hodnoty pre oba atribúty vypíšeme, pričom zaujímavý bude najmä počet záznamov, ktorý tieto hodnoty nadobúda.

Hide

```
subset(df, goals.home >= 8)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	R	3	8	22	FALSE	29	32	
20152016	R	1	8	22	FALSE	31	18	
20172018	R	4	8	5	FALSE	26	28	
20172018	R	3	8	24	FALSE	34	21	
20172018	R	1	10	17	FALSE	23	36	
20162017	R	4	8	19	FALSE	37	11	
20162017	R	7	8	15	FALSE	28	41	
20162017	R	2	8	25	FALSE	30	18	
20162017	R	6	8	23	FALSE	34	15	
20162017	R	1	10	21	FALSE	16	35	

1-10 of 48 rows | 1-9 of 31 columns Previous **1** 2 3 4 5 Next

Hide

```
cat("Počet vychýlených hodnôt atribútu goals.home podľa krabicového grafu:", nrow(sub
set(df, goals.home >= 8)))
```

```
Počet vychýlených hodnôt atribútu goals.home podľa krabicového grafu: 48
```

Hide

```
subset(df, goals.away >= 8)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	R	8	3	1	TRUE	28	34	
20162017	R	8	4	12	TRUE	33	20	
20172018	R	8	3	8	TRUE	29	24	
20172018	R	8	2	16	TRUE	43	14	
20152016	R	9	2	26	TRUE	57	30	
20172018	R	8	5	14	TRUE	38	8	
20172018	P	8	5	5	TRUE	28	34	
20192020	R	8	1	6	TRUE	24	19	
20192020	R	8	3	18	TRUE	24	24	
20192020	R	8	5	30	TRUE	33	22	

1-10 of 23 rows | 1-9 of 31 columns Previous **1** 2 3 Next

Hide

```
cat("Počet vychýlených hodnôt atribútu goals.away podľa krabicového grafu:", nrow(sub
set(df, goals.away >= 8)))
```

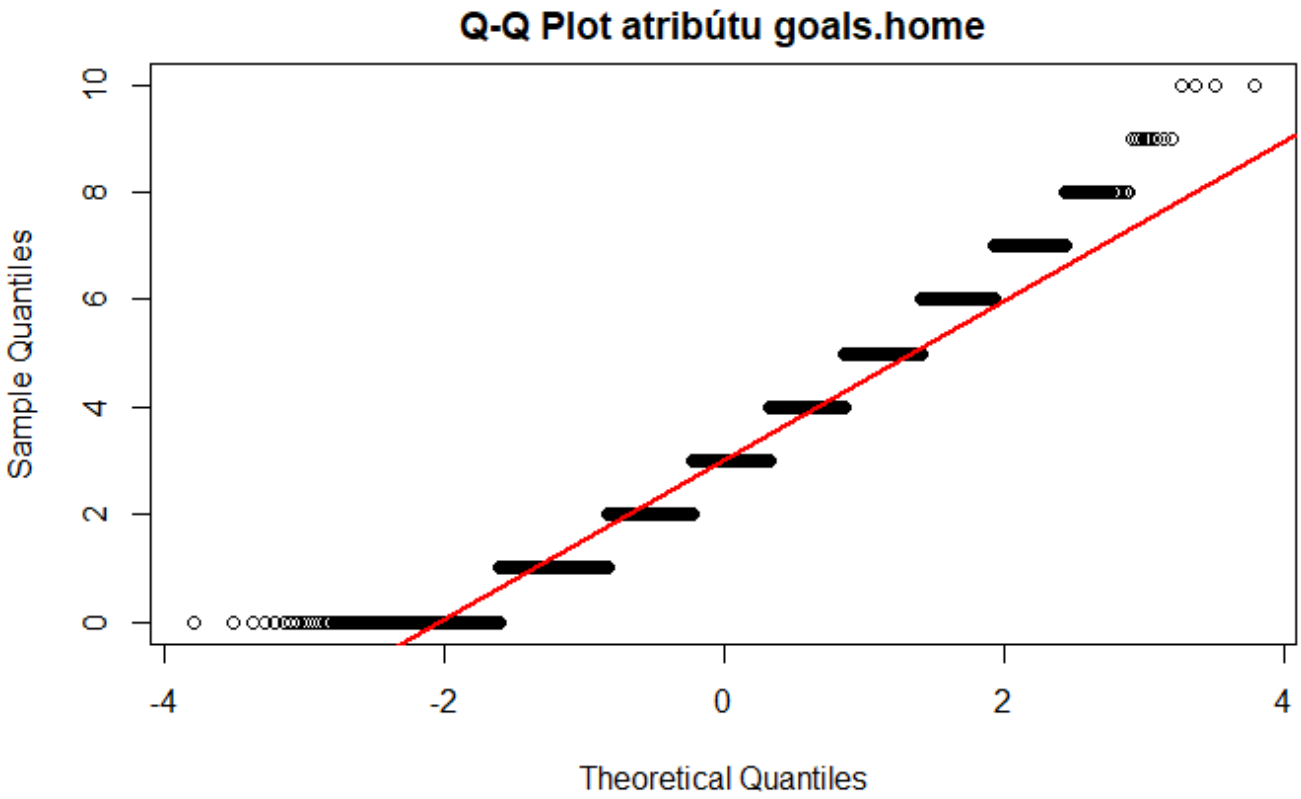
```
Počet vychýlených hodnôt atribútu goals.away podľa krabicového grafu: 23
```

Z výpisov vidíme, že počet zápasov v ktorých domáci tím strelil 8 a viac gólov je 48, zatiaľ čo počet zápasov v ktorých vonkajší tím strelil 8 a viac gólov je 23 - tento fakt len podporuje hypotézu, že domáce prostredie má reálny vplyv na množstvo strelených gólov. Krabicový graf tieto hodnoty označuje ako vychýlené, no keďže obsahujú dôležité informácie a jedná sa o reálne hodnoty (nie napr. chyby senzorov), tieto hodnoty nie je vhodné odstraňovať a ani normovať, resp. zrážať na nižšiu hodnotu.

V analýze atribútu budeme pokračovat prostredníctvom QQplotu, ktorý zobrazuje odchylku od teoretického normálneho rozdelenia.

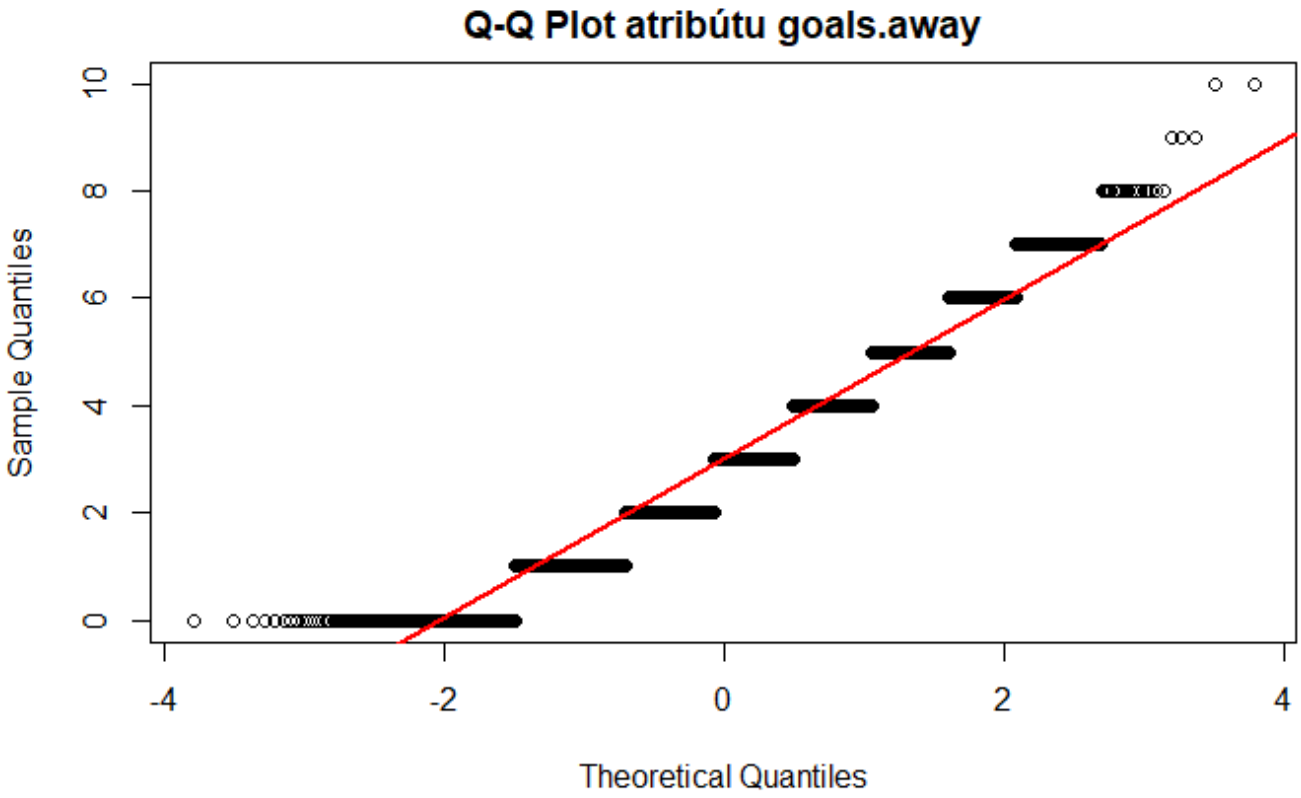
Hide

```
qqnorm(df$goals.home, main="Q-Q Plot atribútu goals.home")
qqline(df$goals.home, col = "red", lwd = 2)
```



Hide

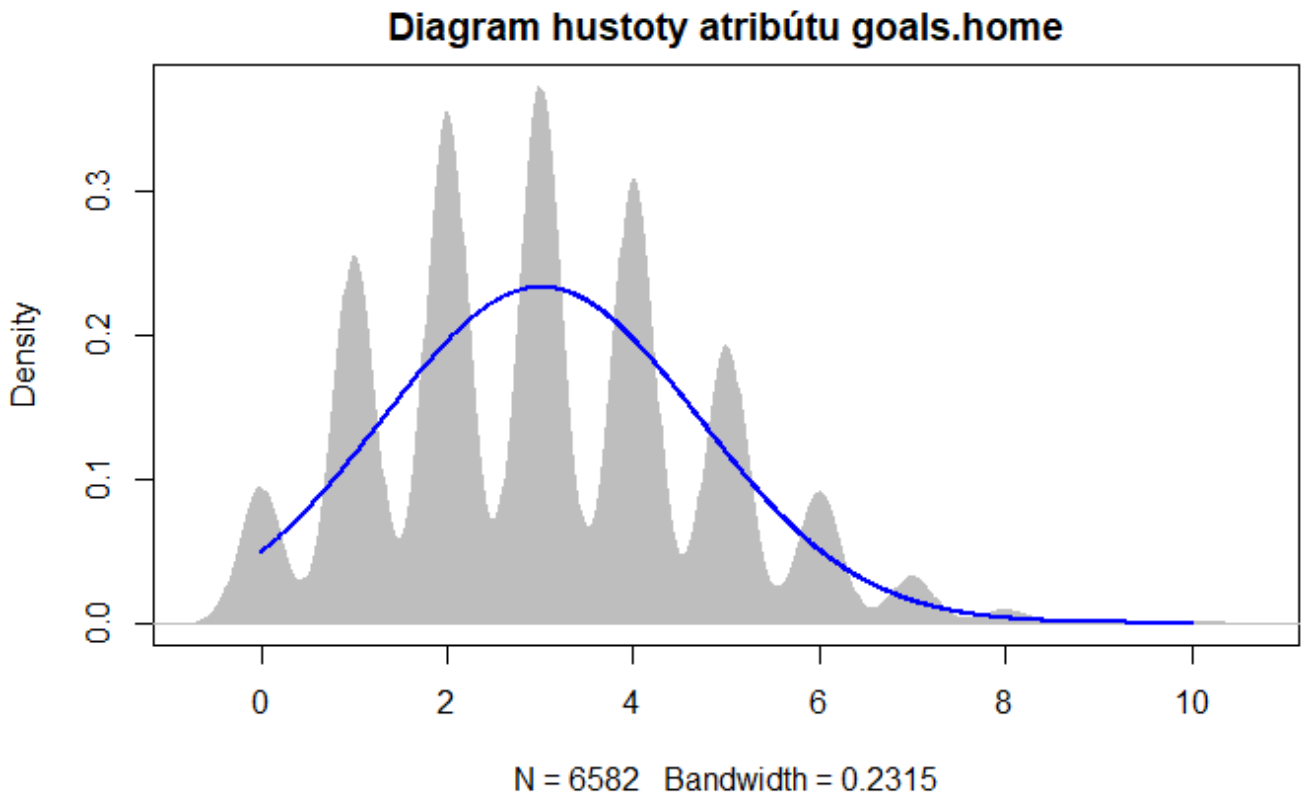
```
qqnorm(df$goals.away, main="Q-Q Plot atribútu goals.away")
qqline(df$goals.away, col = "red", lwd = 2)
```



Ako už bolo spomínane v úvode analýzy týchto atribútov, nejedná sa o spojité hodnoty - táto skutočnost je viditeľná aj na grafe, kedy hodnoty atribútu **goals.home** nie sú spojité - toto správanie je očakávané, keďže počty gólov musia byť celé čisla. Z grafov môžeme vidieť, že počet gólov sa vychyľuje od teoretického normálneho rozdelenia.

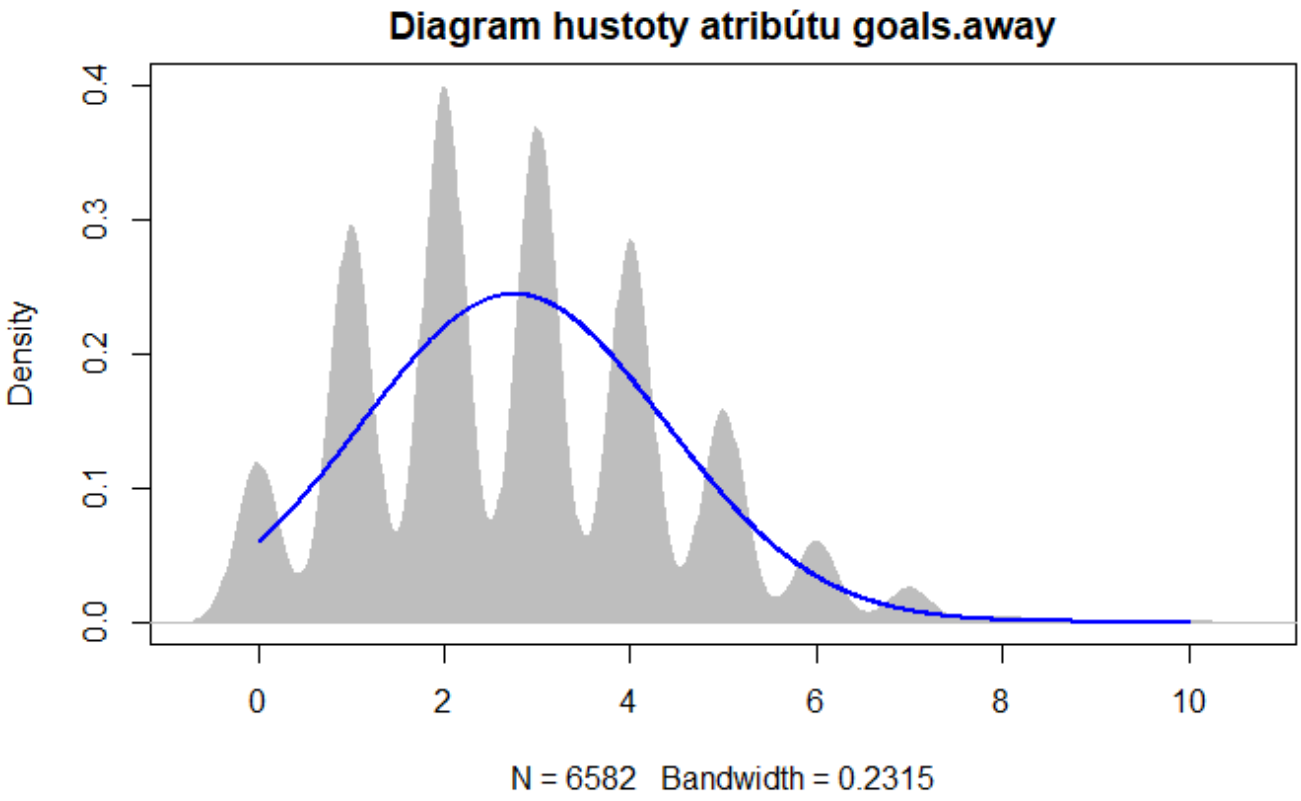
Hide

```
plotNormalDensity(df$goals.home, main = "Diagram hustoty atribútu goals.home")
```



Hide

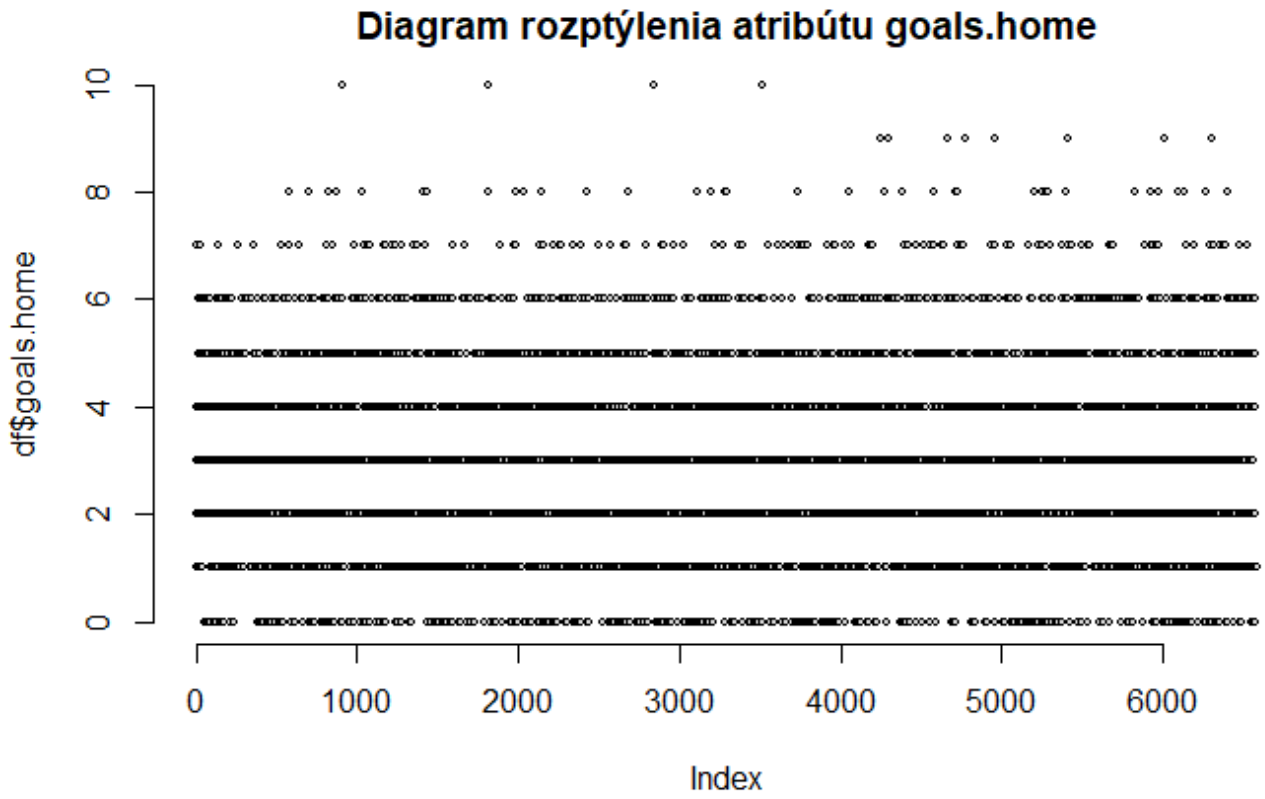
```
plotNormalDensity(df$goals.away, main = "Diagram hustoty atribútu goals.away")
```



Z oboch diagramov hustoty môžeme vidieť už predom spomínaný fakt - atribúty **goals.home** a **goals.away** nepochádzajú z normálneho rozdelenia - výrazne vychýlené od krivky hustoty normálneho rozdelenia sú najmä hodnoty v intervale od (1-5 gólov) na x-ovej osi. Kostrbatosť grafu spôsobuje fakt, že dáta nie sú spojité.

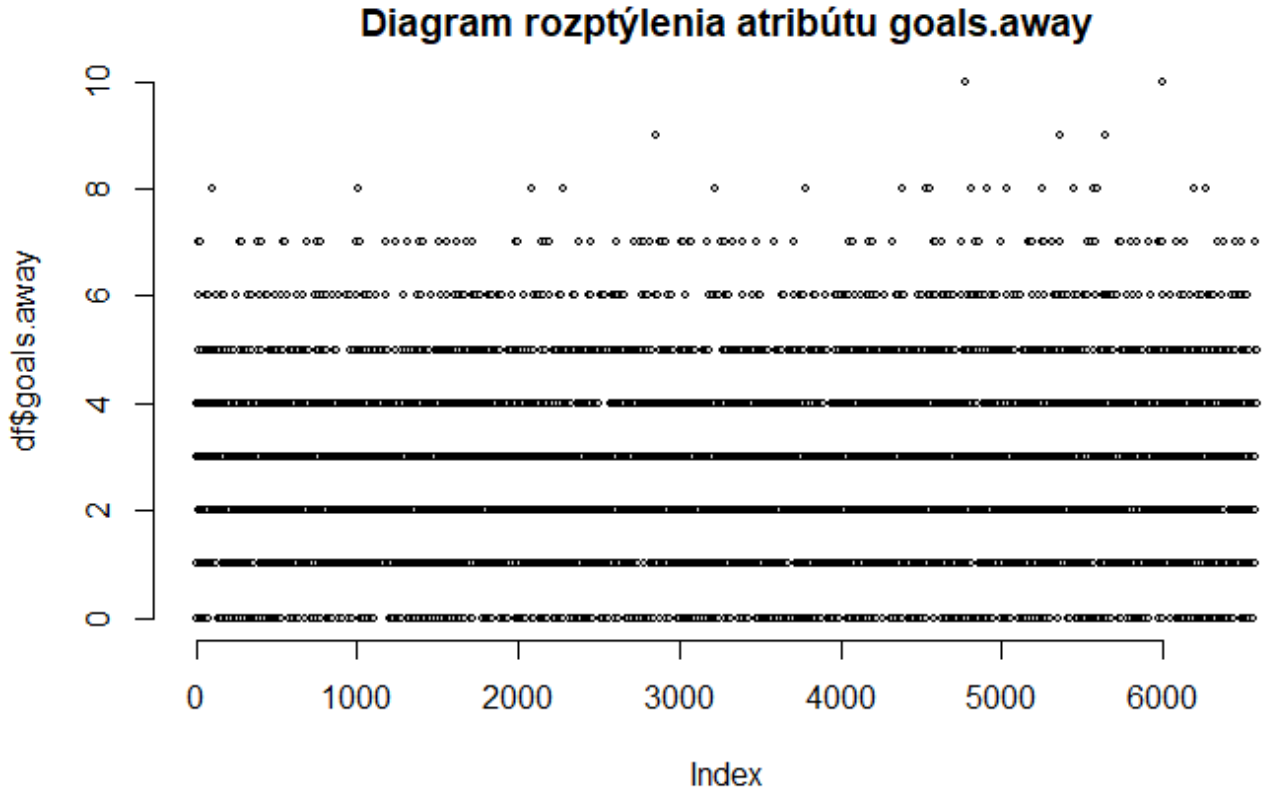
Hide

```
plot(x=df$goals.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = F
ALSE,main = "Diagram rozptýlenia atribútu goals.home")
```



Hide

```
plot(df$goals.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FAL
SE, main = "Diagram rozptýlenia atribútu goals.away")
```



Z grafov rozptýlenia vidieť, že podobne ako bolo viditeľné ak v krabicovom grafe, vychýlené hodnoty sú pri počte gólov >=8 - vychýlenosť idnikuje menšia hustota záznamov nadobúdajúcich hodnoty atribútov **goals.home** a **goals.away** >= 8 (na grafe viditeľné ako pomerne výrazný pokles “bodiek”). Ako už však bolo spomínané, odstránenie alebo úprava týchto hodnôt by so sebou niesla riziko straty dôležitých informácií - preto bude vhodné tieto hodnoty ponechať, keďže sa jedná o reálne údaje. Z grafu nie je príliš dobre vidieť, pre aký počet gólov je najväčšia koncentrácia záznamov (príliš veľa záznamov spôsobuje nízku čitateľnosť grafu) - už z krabicového grafu však vieme, že najvyššia koncentrácia je v rozmedzí 2-4 gólov, čiže táto informácia už nie je kľúčová.

Hide

```
#TU NEJAKY GRAF ESTE
```

Zhrnutie: atribúty **goals.home** a **goals.away** sú diskrétne atribúty, ktoré nepochádzajú z normálneho rozdelenia (viditeľné na histogramoch, QQ-plotoch a grafoch hustoty), ich distribúcia je podľa očakávaní naklonená vpravo. Z grafov rozptýlenia a krabicových grafov vidieť, že za vychýlené hodnoty možno pri oboch atribútoch považovať počty gólov väčšie rovné ako 8 - z dôvodu významnosti tejto informácie však tieto hodnoty neplánujeme nijako upravovať. Oba atribúty možno z hľadiska plánovaných hypotéz považovať za kľúčové.

Atribúty team_id.home a team_id.away

charakteristika: Identifikačné číslo NHL tímu v rámci datasetu.

Aktuálne sa v NHL nachádza 31 tímov, pričom sa tento počet od sezóny 2015/2016 nezmenil. Overíme, či sa v datatsete sedí počet tímov a pozrieme sa aj na počet hier každého tímu, ktorý sa môže meniť pri účasti vo vyraďovacej fáze.

Hide

```
length(table(df$team_id.home))
```

```
[1] 34
```

Hide

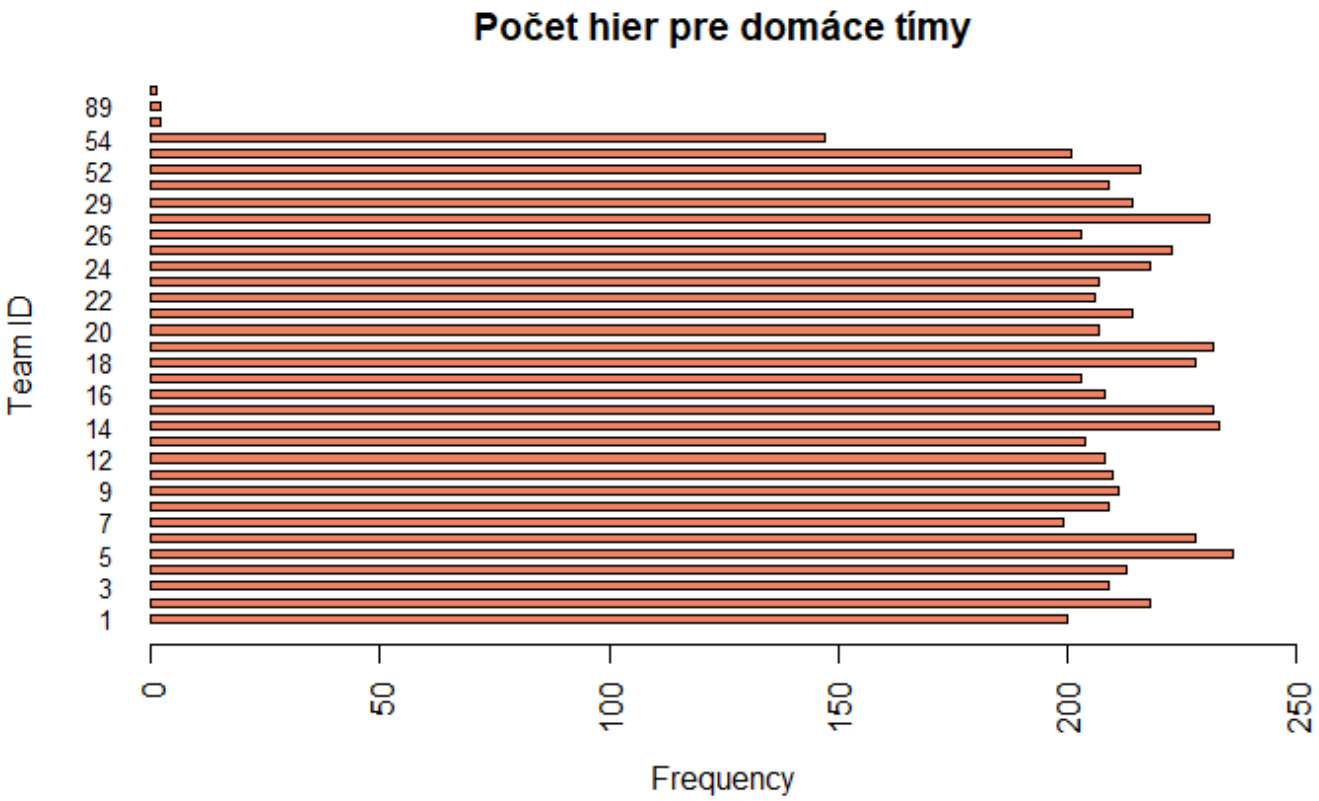
```
table(df$team_id.home)
```

```
  1   2   3   4   5   6   7   8   9  10  12  13  14  15  16  17  18  19  20  21  22
23  24  25
200 218 209 213 236 228 199 209 211 210 208 204 233 232 208 203 228 232 207 214 206 2
07 218 223
 26  28  29  30  52  53  54  87  89  90
203 231 214 209 216 201 147   2   2   1
```

Hide

```
barplot(table(df$team_id.home), las=2, cex.names=.8, xlab="Frequency", ylab="Team I
D", main="Počet hier pre domáce tímy", horiz=TRUE, xlim=c(0,250), space=c(1,1,1,1), c
ol=c("salmon2"))
```

```
longer object length is not a multiple of shorter object lengthlonger object length i
s not a multiple of shorter object length
```



Hide

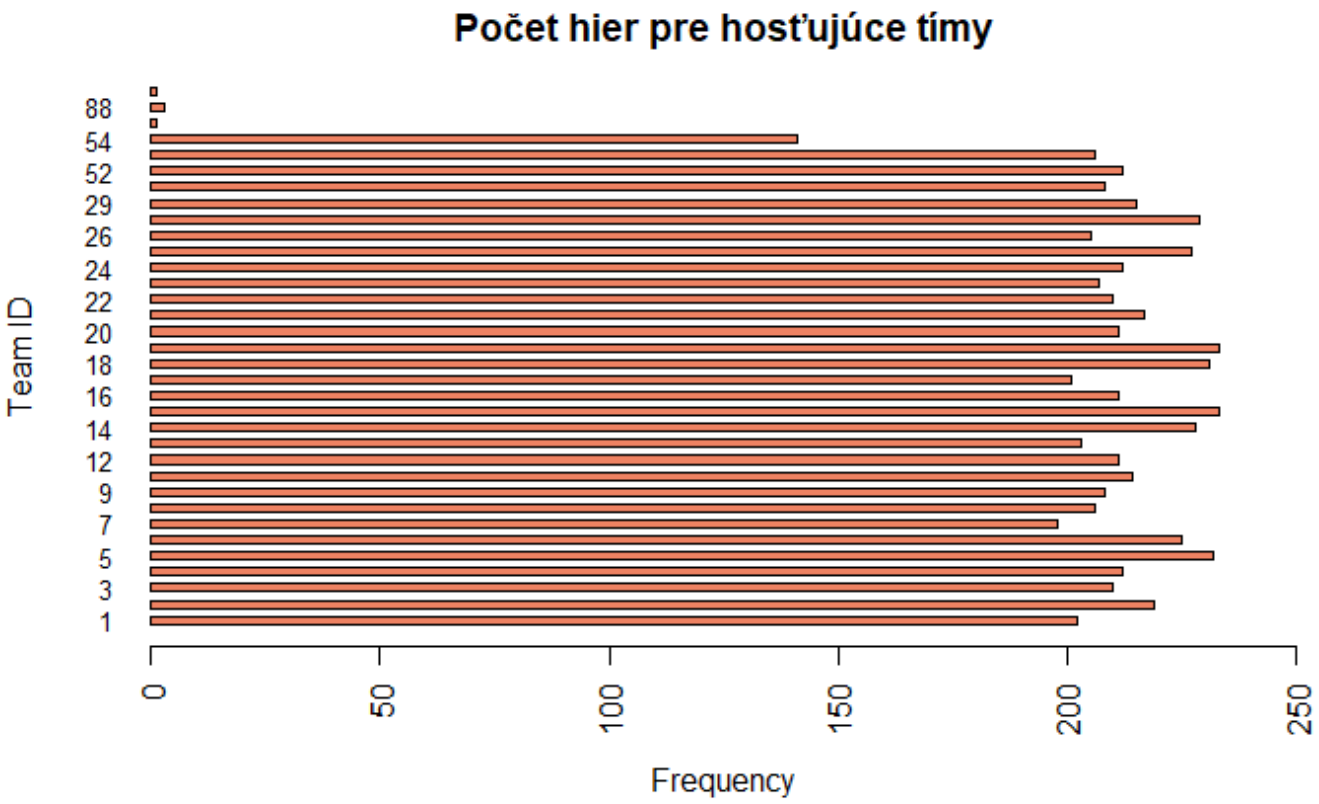
```
table(df$team_id.away)
```

```
  1   2   3   4   5   6   7   8   9  10  12  13  14  15  16  17  18  19  20  21  22
23  24  25
202 219 210 212 232 225 198 206 208 214 211 203 228 233 211 201 231 233 211 217 210 2
07 212 227
 26  28  29  30  52  53  54  87  88  90
205 229 215 208 212 206 141   1   3   1
```

Hide

```
barplot(table(df$team_id.away), las=2, cex.names=.8, xlab="Frequency", ylab="Team I
D", main="Počet hier pre hosťujúce tímy", horiz=TRUE, xlim=c(0,250), space=c(1,1,1,
1), col=c("salmon2"))
```

```
longer object length is not a multiple of shorter object lengthlonger object length i
s not a multiple of shorter object length
```



Zistili sme, že v datasete je viac ako 31 tímov. Taktiež z grafov môžeme vidieť, že niektoré tímy odohrali iba pár zápasov. Tieto zápasy si vypíšeme nižšie. Ostatné tímy majú približne rovnaký počet zápasov ~(200-230).

Hide

```
df[df$team_id.away >= 55]
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20192020	A	5	9	88	FALSE	16	0	
20192020	A	10	5	90	TRUE	28	0	
20192020	A	4	5	87	FALSE	11	0	
20182019	A	7	4	88	TRUE	26	0	
20182019	A	10	5	88	TRUE	22	0	

5 rows | 1-9 of 31 columns

Z výpisu môžeme vidieť, že tímy s nízkym počtom zápasov sú exhibičné tímy, ktoré sme úvadzali vyššie - nejedná sa o tímy NHL. Tieto záznamy teda bude potrebné odstrániť.

Atribúty won.home a won.away

charakteristika: jedná sa o boolean atribúty, vyjadrujú ktorý tím v zápase zvíťazil. V drtivej väčšine prípadov sú opačné, t.j TRUE FALSE alebo FALSE TRUE. Môže však nastať aj situácia v ktorej majú oba hodnotu FALSE - keďže remízy v týchto sezónach neboli povolené, tieto hodnoty by sa reálne v datasete vyskytovať pri legitímnych záznamoch nemali.

Hide

```
unique(df$won.home)
```

```
[1]  TRUE FALSE
```

Hide

```
unique(df$won.away)
```

```
[1] FALSE  TRUE
```

Hide

```
subset(df, won.home==TRUE & won.away==TRUE)
```

0 rows | 1-9 of 31 columns

Hide

```
subset(df, won.home==FALSE & won.away==FALSE)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	0	0	
20172018	P	0	0	28	FALSE	*NA*	*NA*	
20172018	P	0	0	26	FALSE	0	0	
20172018	P	0	0	52	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	0	0	
20172018	P	0	0	15	FALSE	0	0	
20172018	P	0	0	5	FALSE	0	0	
20162017	P	0	0	20	FALSE	0	0	
20162017	P	0	0	10	FALSE	0	0	

1-10 of 14 rows | 1-9 of 31 columns Previous **1** 2 Next

Hide

```
verify <- subset(df, won.home==FALSE & won.away==FALSE)
```

Výsledok zápasu TRUE TRUE nie je možný. Preto je potrebné skontrolovať, či sa v datasete tento výsledok nenachádza (ak áno je potrebné ho opraviť). Po kontrole s vysužitím funkcie unique sme zistili, že TRUE TRUE sa v datasete nachádza 0x. FALSE FALSE sa vŠak v datasete nachádza 14x - po pohľade na dáta vieme povedať, že sa nebude jednať o remízy - pre záznamy chýbajú dáta, pravdepodobne sa bude jednať o odložené alebo zrušené zápasy. Žiaden z týchto zápasov však pre naše riešenie nemá pridanú hodnotu - keďže chýbajú všetky relevantné štatistické atribúty (resp. ich hodnota je 0), tieto zápasy môžeme úplne bez problémov z datasetu pri čistení odstrániť - nestratíme totižto žiadnu štatisticky cennú informáciu.

Hide

```
unique(verify$season)
```

```
[1] 20172018 20162017
```

Hide

```
table(is.na(df$won.home))
```

```

FALSE
 6582
```

Hide

```
table(is.na(df$won.away))
```

```

FALSE
 6582
```

Atribúty won.home a won.away neobsahujú žiadne prázdne hodnoty, preto nie je potrebné voliť stratégie ich nahrádzania.

Atribúty shots.home a shots.away

charakteristika: atribút reprezentuje počet striel na bránu z pohľadu domáceho a hosťovského tímu.

Pri atribútoch overíme, či neobsahujú chýbajúce hodnoty:

Hide

```
cat("Počet záznamov kde je shots.home NA", nrow(subset(df, is.na(shots.home)==TRUE)),
'\n')
```

```
Počet záznamov kde je shots.home NA 4
```

Hide

```
cat("Počet záznamov kde je shots.away NA",nrow(subset(df, is.na(shots.away)==TRU
E)),'\n')
```

```
Počet záznamov kde je shots.away NA 4
```

Vidíme, že v štyroch záznamoch nadobúdajú atribúty shots.home a shots.away hodnotu NA. Na tieto záznamy sa môžeme bližšie pozrieť.

Hide

```
subset(df, is.na(shots.home)==TRUE)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	*NA*	*NA*	
20172018	P	0	0	52	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	*NA*	*NA*	
20162017	P	0	0	3	FALSE	*NA*	*NA*	

4 rows | 1-9 of 31 columns

Hide

```
subset(df, is.na(shots.away)==TRUE)
```

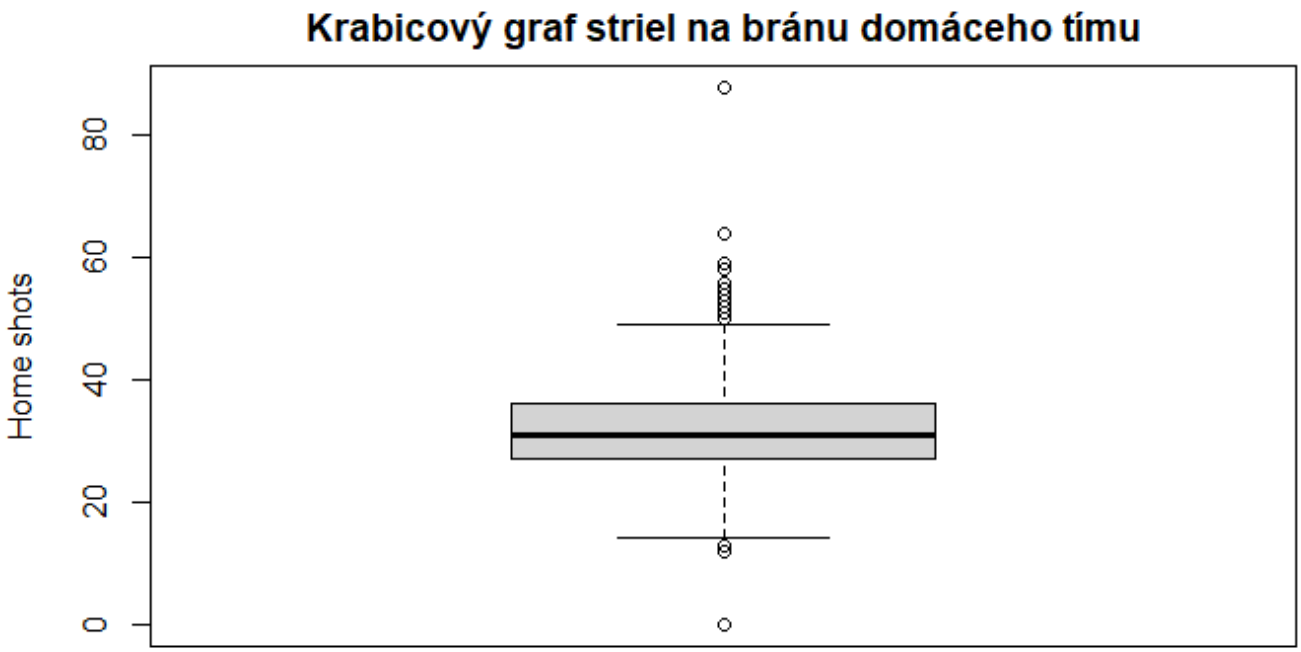
season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	*NA*	*NA*	
20172018	P	0	0	52	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	*NA*	*NA*	
20162017	P	0	0	3	FALSE	*NA*	*NA*	

4 rows | 1-9 of 31 columns

Z výpisu vidíme, že v oboch prípadoch sa jedná o zápasy play-off medzi tímami NHL, ktoré vŠak neobsahujú žiade štatistické informácie (takmer všetko je NA) - tieto záznamy bude vhodné odstrániť, keďže neobsahujú žiadne relevantné informácie, pravdepodobne sa jedná o zrušené zápasy v play-off. Taktiež je vhodné podotknúť, že tieto 4 zápasy sú rovnaké pre oba atribúty, t.j. po ich odstránení budú odstránené NA hodnoty pre oba atribúty - bude riešené vo fázi čistenia dát.

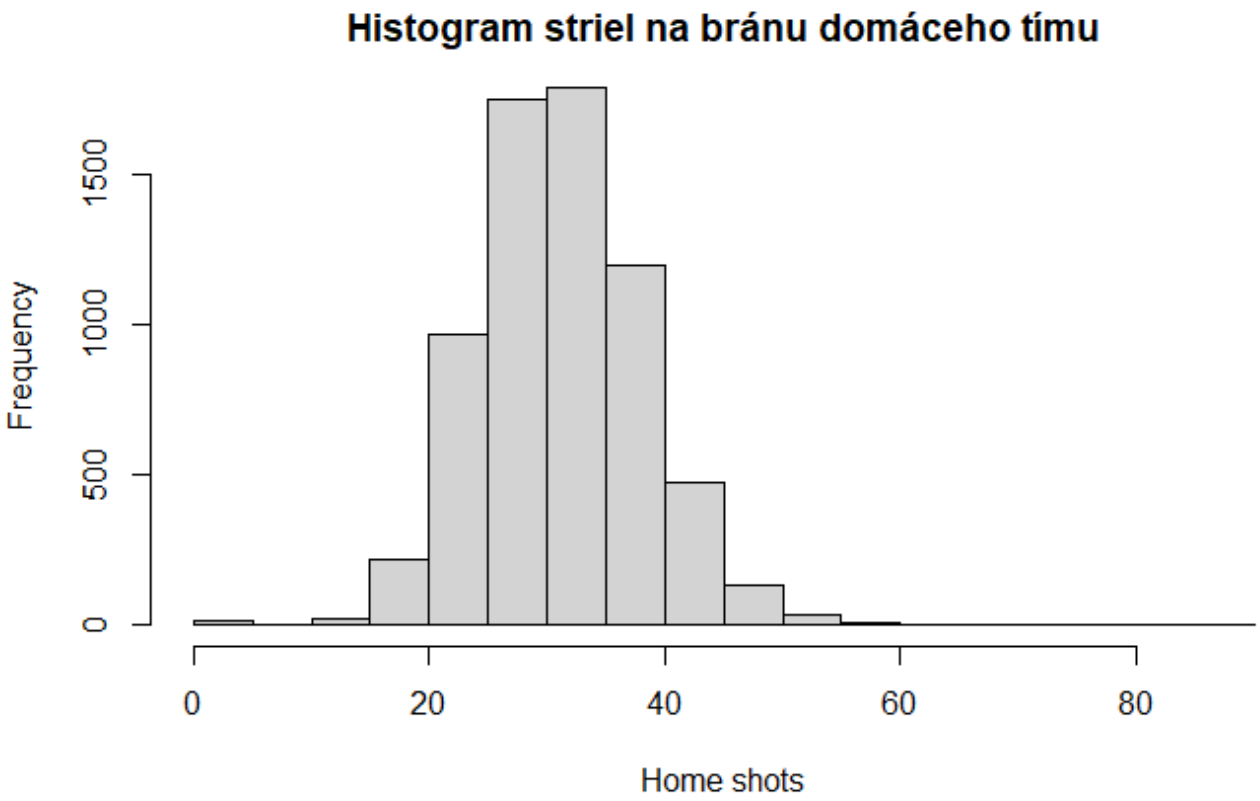
Hide

```
boxplot(df$shots.home, ylab="Home shots", xlab="", main="Krabicový graf striel na brá
nu domáceho tímu")
```



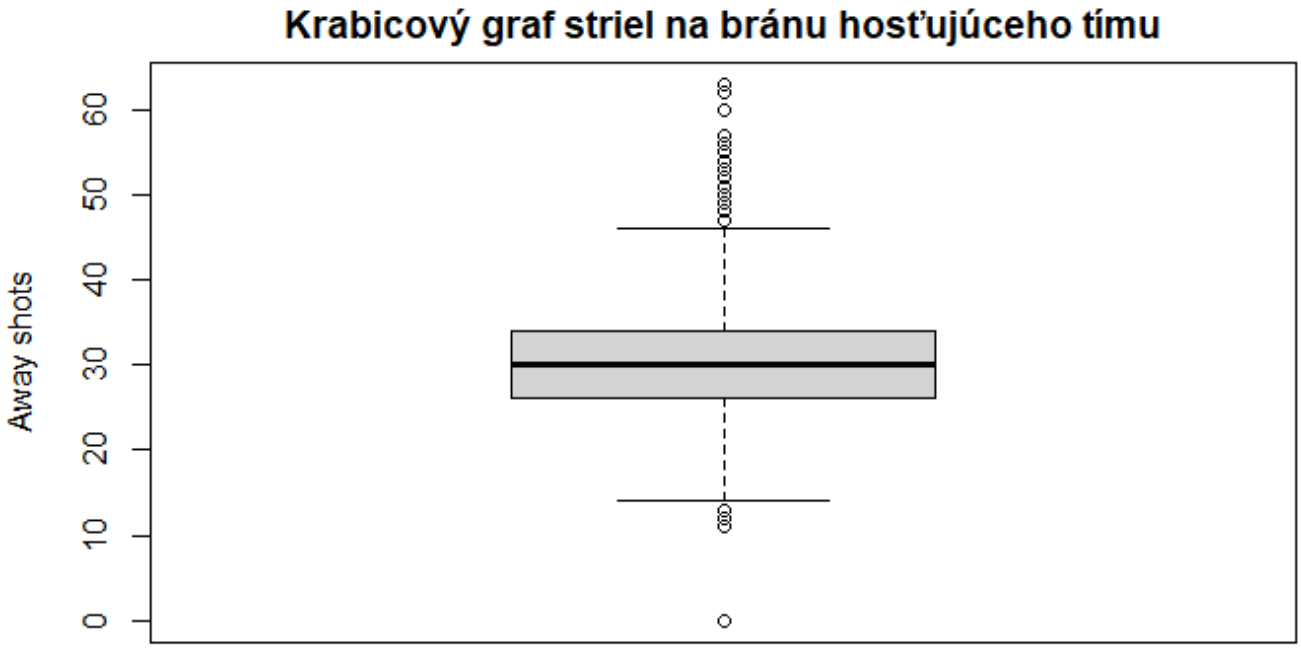
Hide

```
hist(df$shots.home, xlab="Home shots", main="Histogram striel na bránu domáceho tím
u")
```



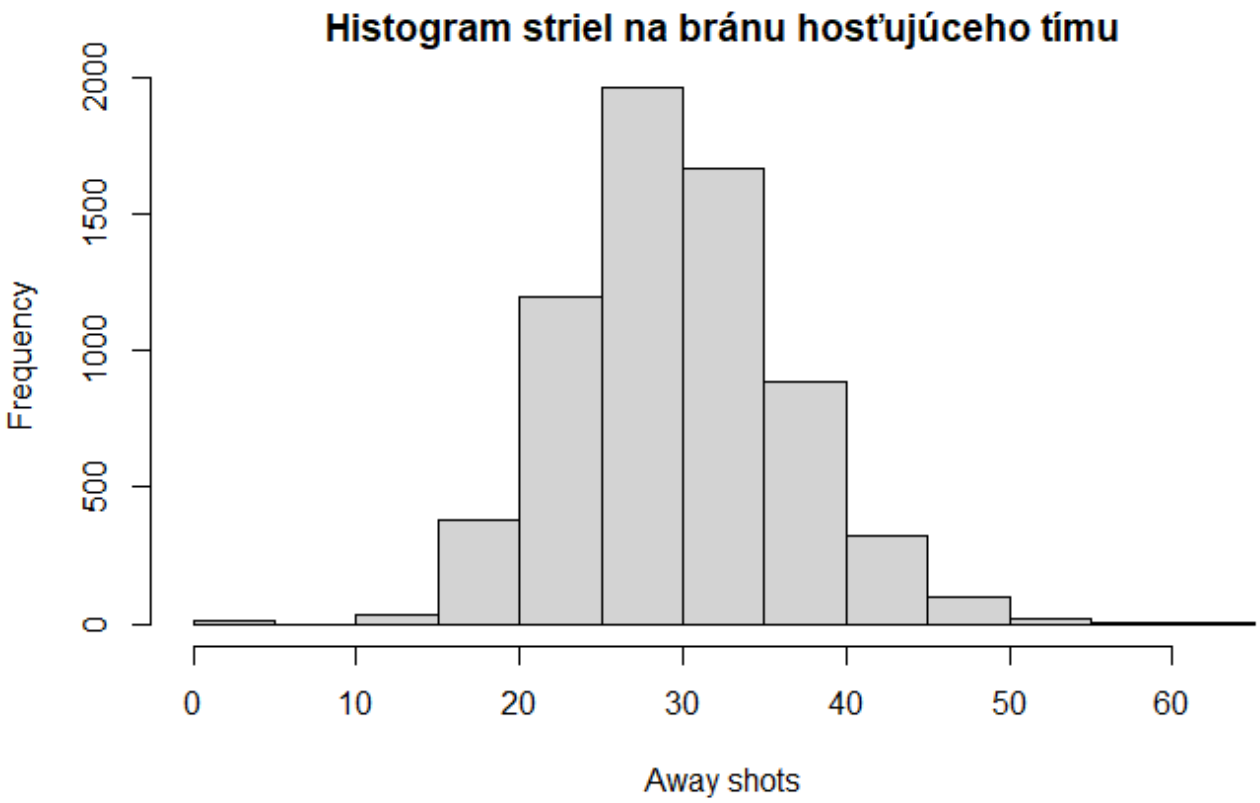
Hide

```
boxplot(df$shots.away, ylab="Away shots", xlab="", main="Krabicový graf striel na brá
nu hosťujúceho tímu")
```



Hide

```
hist(df$shots.away, xlab="Away shots", main="Histogram striel na bránu hosťujúceho tí
mu")
```



Z histogramov možno vidieť, že distribúcia atribútov shots.home a shots.away pripomína normálnu distribúciu. Taktiež možno sledovať odľahlú hodnotu - konkrétne počet striel 0. Ako sme už zistili, jedná sa o chýbajúce záznamy, ktoré budú neskôr odstránené. Predpokladáme teda, že táto odľahlá hodnota bude po fáze čistenia dát eliminovaná. Ohľadom hodnôt atribútov vieme povedať, že aj pri **shots.home** aj **shots.away** nadobúdajú atribúty veľmi podobné hodnoty - najväčšia koncentrácia záznamov je okolo 30 striel. Pri prvom pohľade na histogram však vyzerá, že hosťujúce tímy mávajú v priemere o niečo málo viac striel na bránu ako domáce, čo je zaujímavé, keďže domáce tímy mávajú priemerne viac gólov na zápas (ako bolo opísané v analýze atribútov **goals.home** a **goals.away**).

Pri krabicových grafoch je zaujímavý najmä atribút **shots.home** - konkrétne pomerne výrazne odľahlá hodnota okolo 80 striel na bránu (najbližšia druhá je okolo 60 striel). Pri atribúte **shots.away** - sú všetky "odľahlé" hodnoty združené okolo hodnoty 60 striel. Počet 80+ striel teda na prvý pohľad pôsobí podozrivo - tento záznam si vypíšeme aby sme zistili, či sa môže jednať o chybu merania. Odľahlá hodnota je taktiež hodnota 0, dôvod jej výskytu + postup riešenia sme však už riešili pri opise histogramu.

Hide

```
subset(df, df$shots.home >= 80)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>

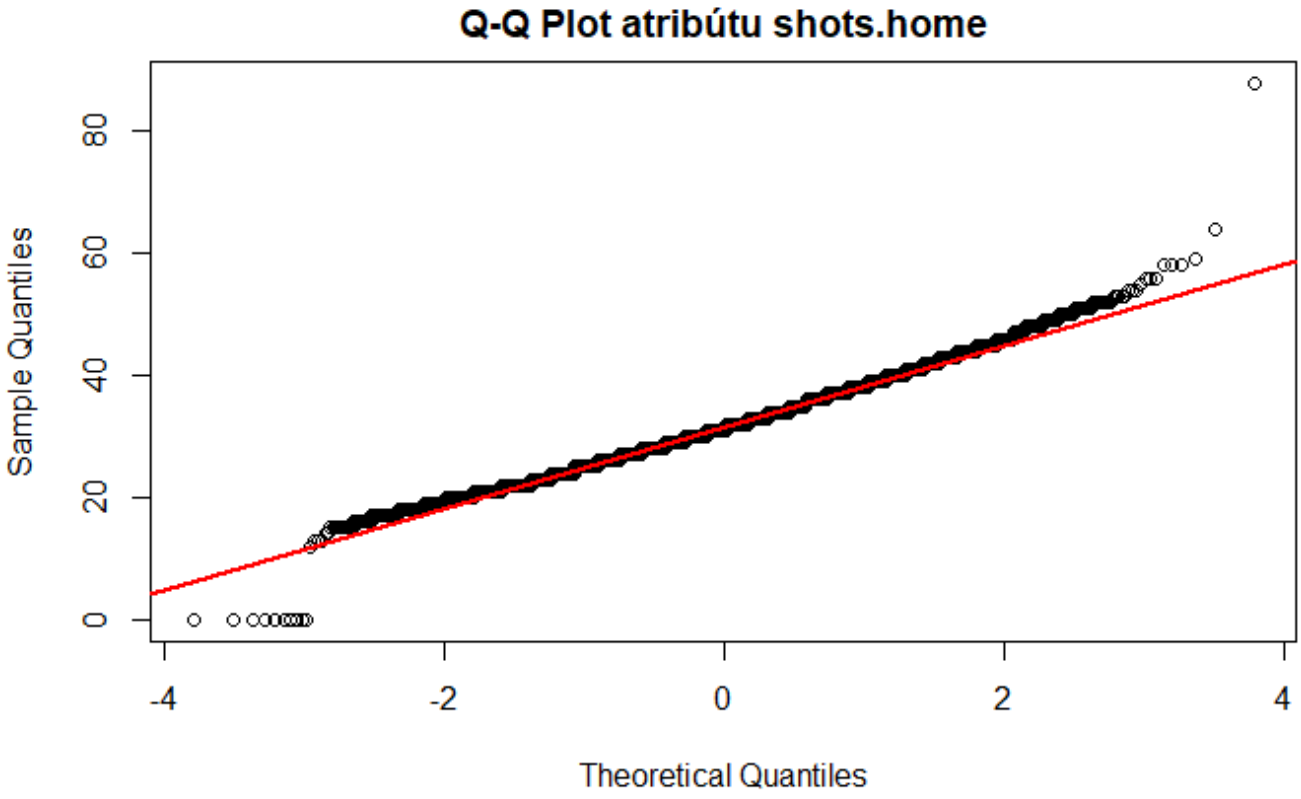
season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20192020	P	2	3	29	FALSE	63	46	

1 row | 1-9 of 31 columns

Z výpisu záznamu vidíme, že sa jedná o zápas play-off medzi Columbus Blue Jackets a Tamba Bay Lightning, kedy bol počet striel na jednej strane 88 a na druhej 63. Zápas bol taktiež ukončený v predĺžení - keďže v play-off je možný nekonečný počet predĺžení až pokiaľ nie je jasný výsledok zápasu (remíza vo vyraďovacích zápasoch nie je povolená) môžeme predpokladať, že sa jedná o zápas ktorý bol predĺžovaný viac krát. Počty striel sú teda legitímne a nejedná sa o chyby merania. Jedná sa však o veľmi unikátnu situáciu, ktorá môže potenciálne negatívne ovplyvniť výsledky predikčného modelu. Z tohoto dôvôdu je vhodné uvažovať nad normovaním tejto hodnoty.

Hide

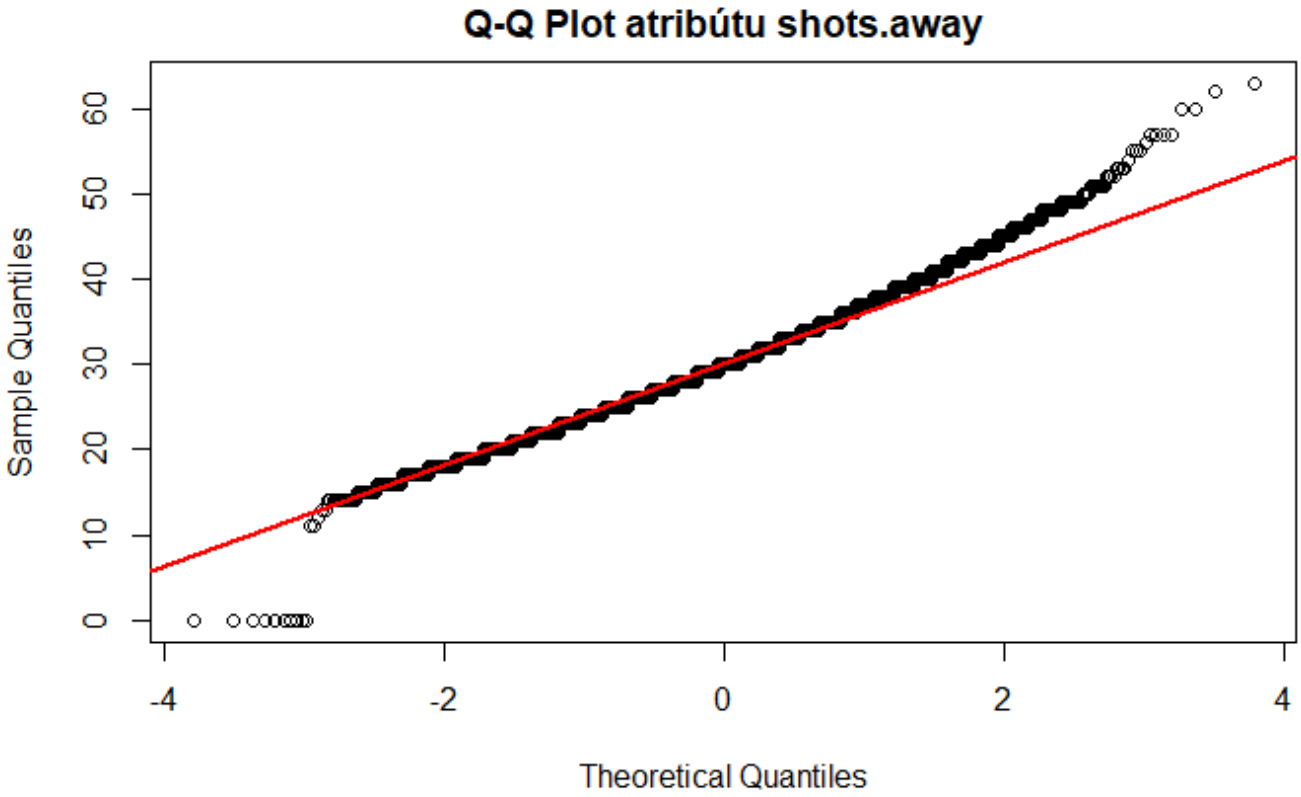
```
qqnorm(df$shots.home, main="Q-Q Plot atribútu shots.home")
qqline(df$shots.home, col = "red", lwd = 2)
```



Z QQ-plotu atribútu **shots.home** môžeme vidieť, že sa pomerne pekne drží teoretickej krivky normálneho rozdelenia. Síce sa jedná o diskrétnu celočíselnú hodnotu, graf nie je taký kostrbatý ako pri atribútoch **goals.home** a **goals.away** - je tomu tak preto, že atribút shots dosahuje širší interval hodnôt a tým pádom na grafe skoky medzi hodnotami nie je vidieť. Okolo počtu gólov 3 sa začína rozdelenie mierne odchylovať od krivky normálneho rozdelenia. Za zmienku stoja aj vychýlené hodnoty v 0 (už opísané aj zdôvodnené) a výrazne vychýlená hodnota 80 - jedná sa o reálnu hodnotu, no nad jej normalizáciou budeme uvažovať keďže situáciu, v ktorej hodnota bola nadobudnutá je možno považovať za extremálnu.

Hide

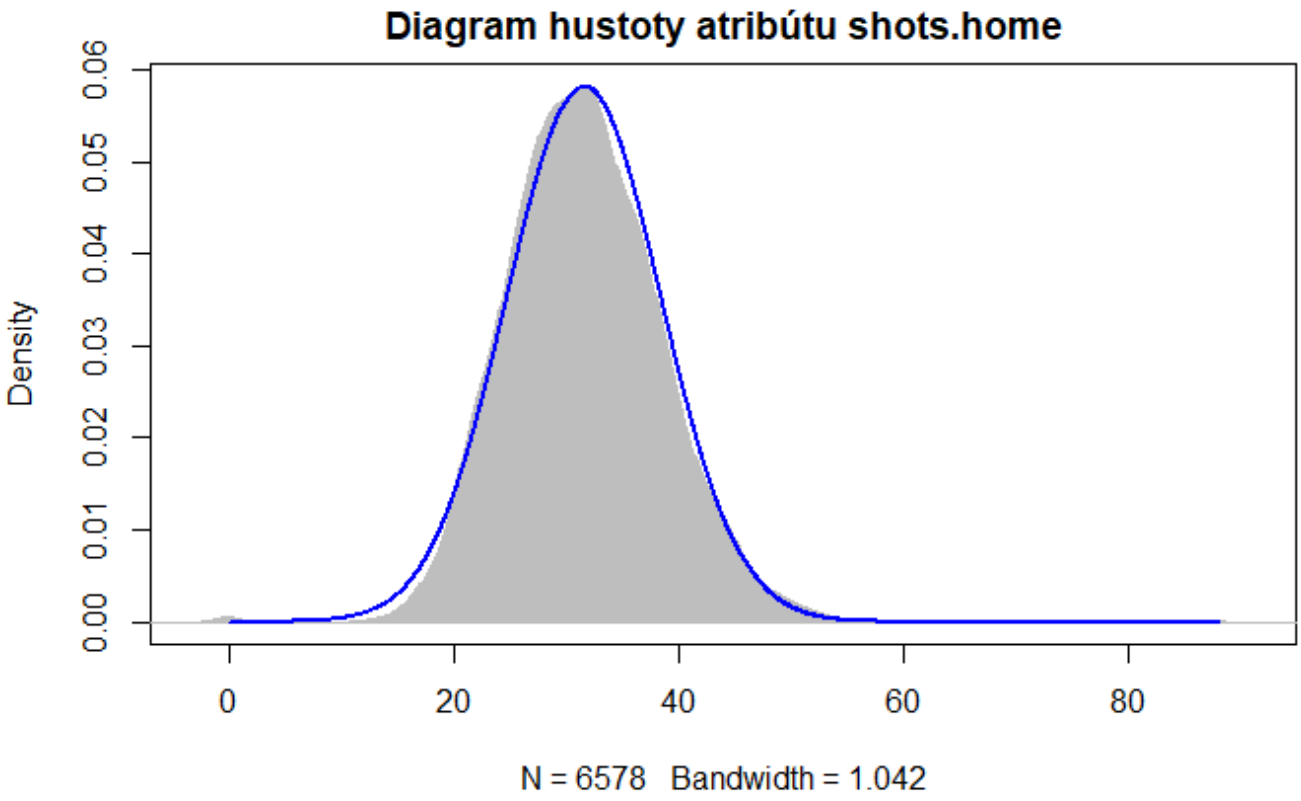
```
qqnorm(df$shots.away, main="Q-Q Plot atribútu shots.away")
qqline(df$shots.away, col = "red", lwd = 2)
```



Pri QQ-plote atribútu **shots.away** vieme poznamenať v podstate to isté ako pri atribúte **shots.home**. Rozdielom však je pomerne prudšie vychýlenie od teoretickej krivky normálneho rozdelenia pre pri počte striel väčšom rovnom ako 40. Výrazne vychýlená hodnota ako v predošlom grafe tu však nie je, všetky nadobúdajú hodnoty okolo 60 (bolo viditeľné už z krabicového grafu). Vyzerá to však tak, že distribúcia normálna nebude a bude potrebná normalizácia atribútu tak, aby sme normálne rozdelenie dosiahli.

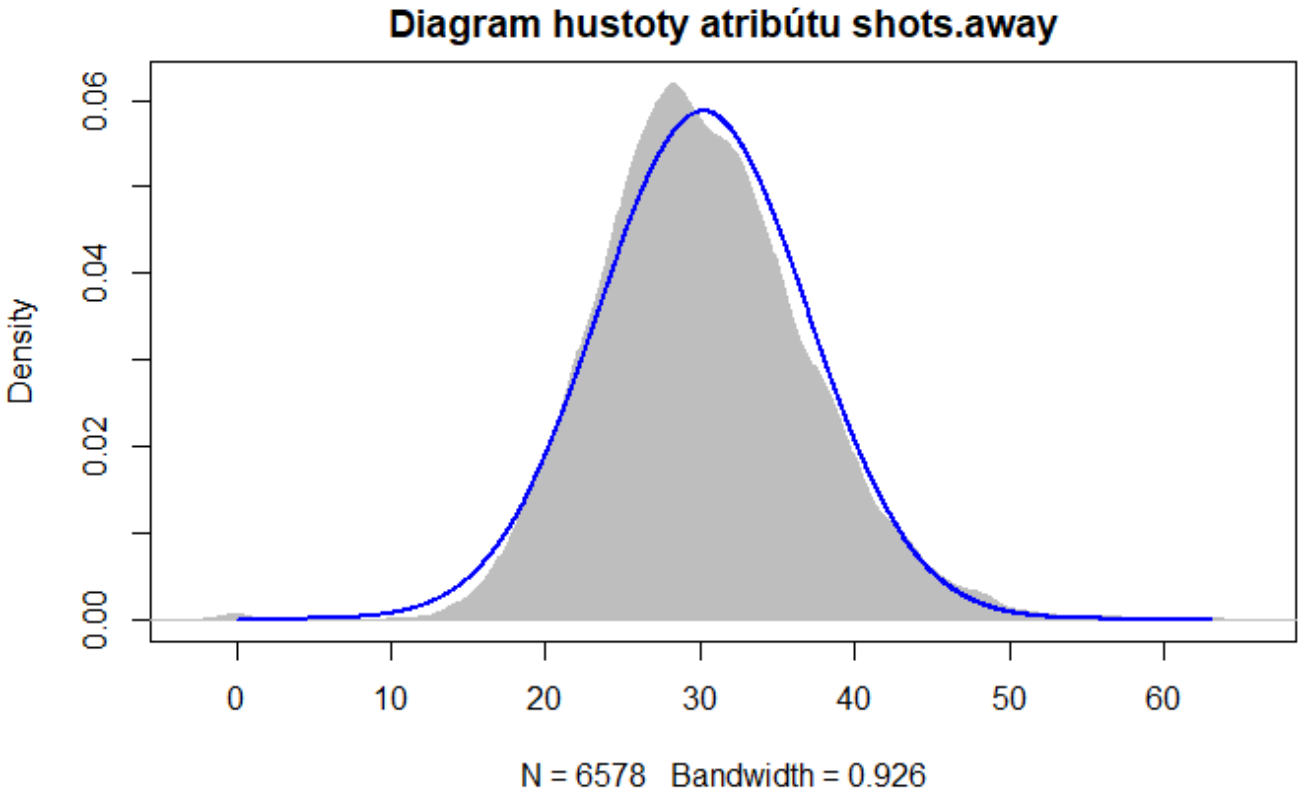
Hide

```
plotNormalDensity(df$shots.home, main = "Diagram hustoty atribútu shots.home")
```



Hide

```
plotNormalDensity(df$shots.away, main = "Diagram hustoty atribútu shots.away")
```



Grafy hustoty indikujú takmer ideálnu hustotu atribútov **shots.home** a **shots.away** - vyzerá to tak, že atribúty budú mať normálnu distribúciu. Z predošlých grafov však máme dôvod o tejto skutočnosti pochybovať, bude teda pravdepodobne potrebné vykonať test normality na reprezentatívnej vzorke dát o veľkosti 1000.

Hide

```
plot(x=df$shots.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = F
ALSE,main = "Diagram rozptýlenia atribútu shots.home")
```

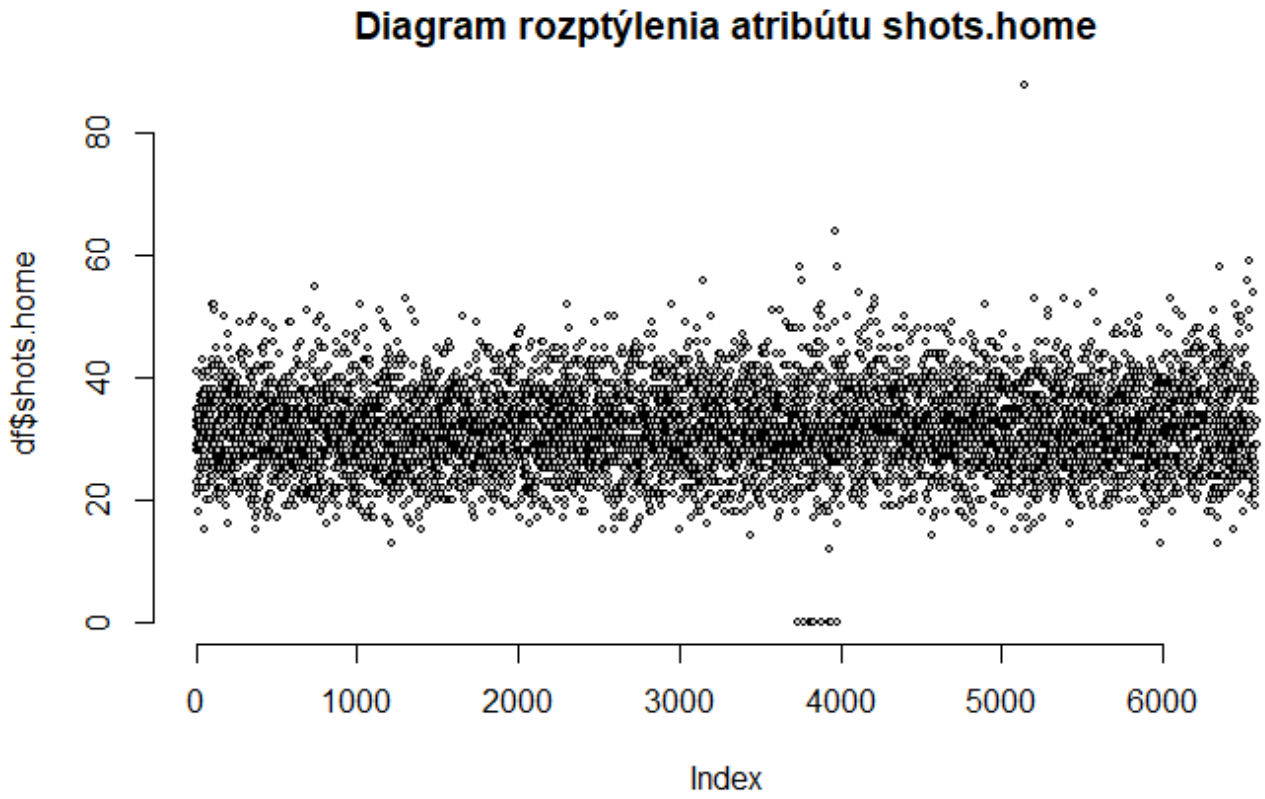


Diagram rozptýlenia atribútu **shots.home** indikuje primánrne zoskupenie hodnôt v intervale od 20-40 striel na bránu, rovnako ako krabicový graf. Opätovne môžeme vidieť extrémnu vychýlenosť hodnoty 88 striel na bránu. Hodnoty 0 sú chýbajúce záznamy, vyjadrovať sa k nim teda nie je potrebné.

Hide

```
plot(x=df$shots.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = F
ALSE,main = "Diagram rozptýlenia atribútu shots.away")
```

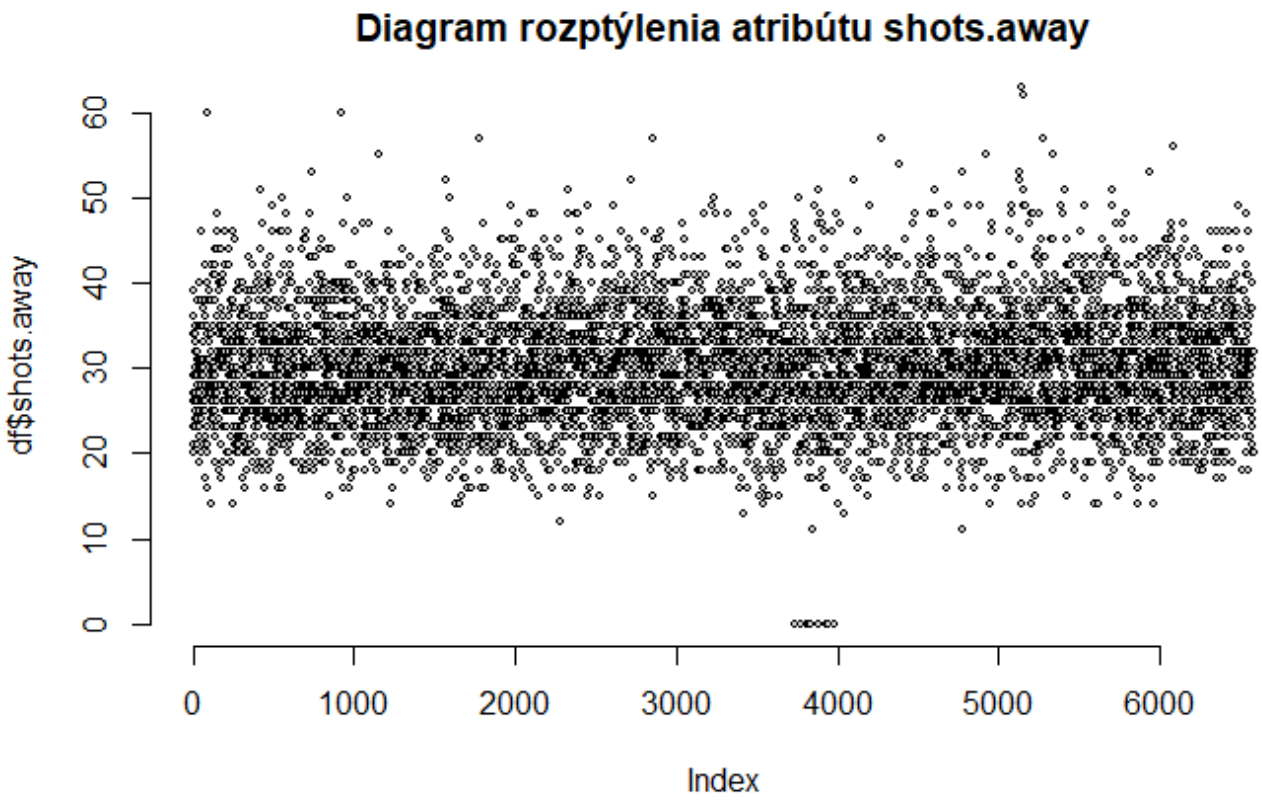
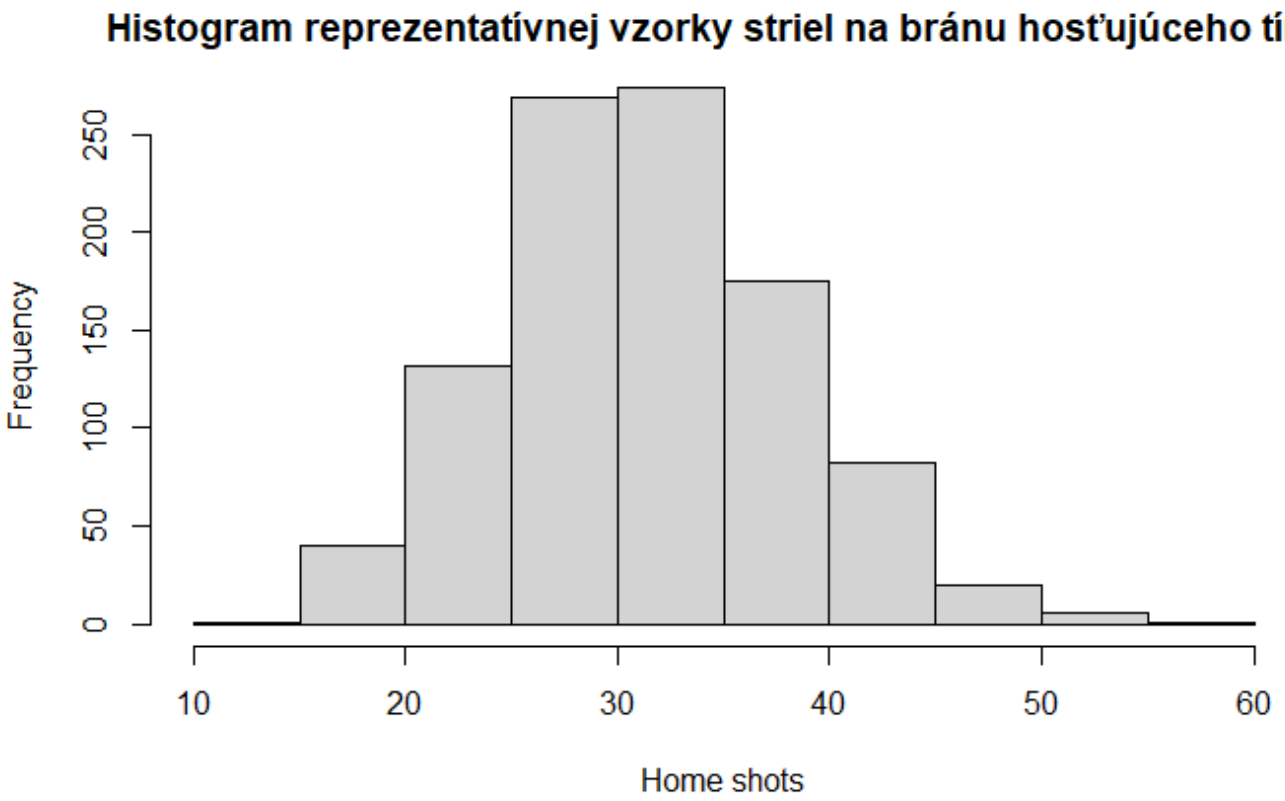


Diagram rozptýlenia atribútu **shots.away** indikuje opäť primárne zoskupenie hodnôt v intervale od 20-40 striel na bránu, rovnako ako predošlý graf rozptýlenia. Neexistuje extrémne vychýlená hodnota, tým pádom je os Y grafu posunutá. Údaje získané z grafov však vyzerajú v oboch atribútoch veľmi podobne.

Grafy indikujú signifikantnú podobnosť s normálnym rozdelením. Bude preto vhodné vykonať Shapiro-Wilkov test normality, pre ktorý je potrebné vybrať reprezentatívnu vzorku o veľkosti 1000.

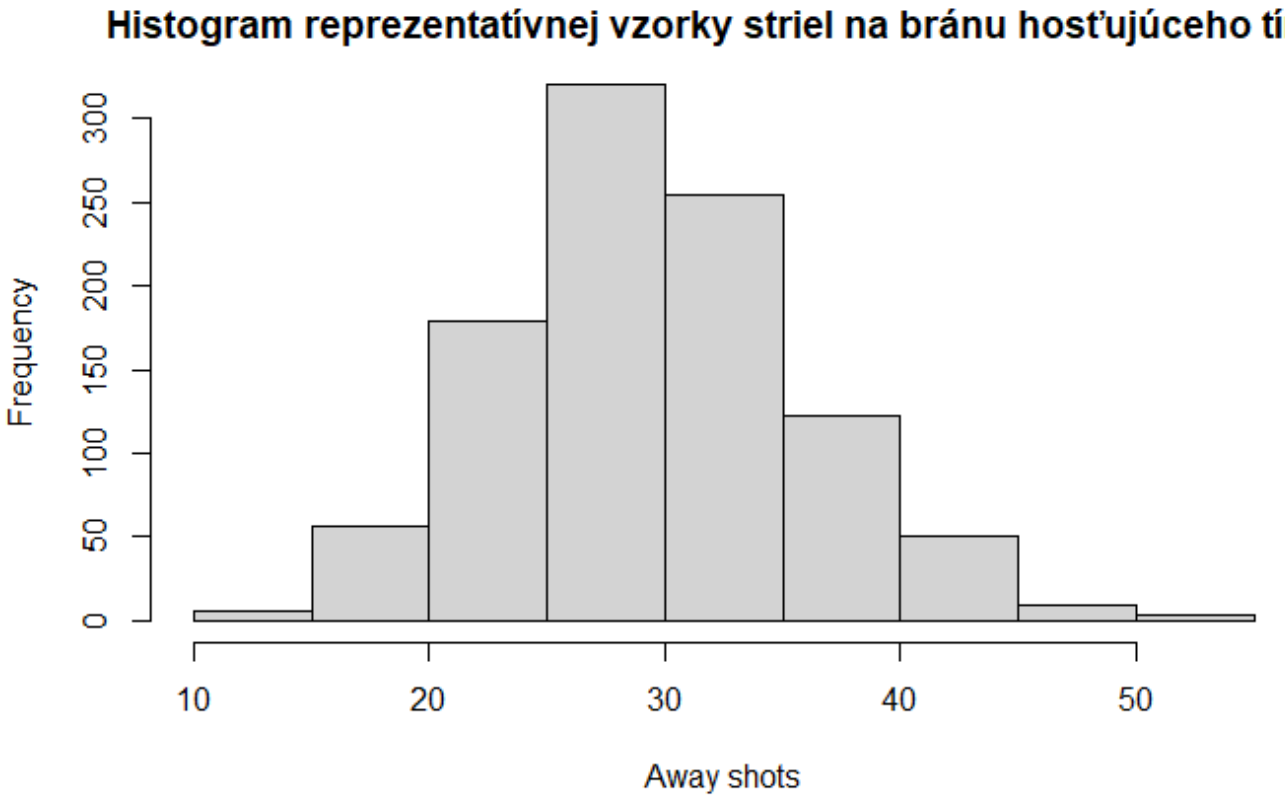
Hide

```
sample <- sample_n(df, 1000)
hist(sample$shots.home, xlab="Home shots", main="Histogram reprezentatívnej vzorky st
riel na bránu hosťujúceho tímu")
```



Hide

```
hist(sample$shots.away, xlab="Away shots", main="Histogram reprezentatívnej vzorky st
riel na bránu hosťujúceho tímu")
```



Hide

```
cat("Štatistika shots.home\n")
```

```
Štatistika shots.home
```

Hide

```
summary(df$shots.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   27.00   31.00   31.63   36.00   88.00       4 
```

Hide

```
cat("Štatistika vzorky shots.home\n")
```

```
Štatistika vzorky shots.home
```

Hide

```
summary(sample$shots.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.00   27.00   32.00   31.83   36.00   58.00 
```

Hide

```
cat("Štatistika shots.away\n")
```

```
Štatistika shots.away
```

Hide

```
summary(df$shots.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   26.00   30.00   30.19   34.00   63.00       4 
```

Hide

```
cat("Štatistika vzorky shots.away\n")
```

```
Štatistika vzorky shots.away
```

Hide

```
summary(sample$shots.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  11.00   26.00   29.00   29.96   34.00   51.00 
```

Zo základnej deskriptívnej štatistiky možno vidieť veľmi podobnú varianciu hodnôt vo vzorkách (dolný-horný kvantil, priemer a medián), preto ich považujeme za reprezentatívne (aj keď neobsahujú všetky extremálne hodnoty).

Následne vykonáme už predom spomínaný Shapiro-Wilkov test normality s hladinou p = 0.05.

Test normality atribútu **shots.home**: nulová hypotéza: dáta pochádzajú z normálneho rozdelenia

Hide

```
shapiro.test(sample$shots.home)
```

```

    Shapiro-Wilk normality test

data:  sample$shots.home
W = 0.9925, p-value = 5.913e-05
```

Keďže p < 0.05, nulovú hypotézu zamietame - dáta nepochádzajú z normálneho rozdelenia

Test normality atribútu **shots.away**: nulová hypotéza: dáta pochádzajú z normálneho rozdelenia

Hide

```
shapiro.test(sample$shots.away)
```

```

    Shapiro-Wilk normality test

data:  sample$shots.away
W = 0.99162, p-value = 1.861e-05
```

Keďže p < 0.05, nulovú hypotézu zamietame - dáta nepochádzajú z normálneho rozdelenia.

Test normality bol v pri oboch atribútoch zamietnutý, preto vieme, že ani jeden atribút nepochádza z normálneho rozdelenia - pravdepodobná je mierna asymteria distribúcií atribútov.

Zhrnutie: atribúty **shots.home** a **shots.away** sú diskrétne atribúty, ktoré nepochádzajú z normálneho rozdelenia (dokázané Shapiro-Wilkovým testom). Hodnota **shots.home** dosahuje výrazne vychýlenú hodnotu 88, pri ktorej však bola zistená jej validita (t.j. jedná sa o valídnu hodnotu a nie chybu senzora). Túto hodnotu pravdepodobne bude vhodné normalizovať napr. pomocou horného kvantilu, keďže je naozaj extremálna. Atribúty **shots.home** a **shots.away** sú na grafov veľmi podobné, dosahujú miernu asymetriu. Vychýlené hodnoty 0 sa vyskytujú v oboch atribútoch, bolo však zistené, že sa jedná o chýbajúce záznamy (zrušené/presunuté zápasy) - tie budú vyriešené pri fáze čistenia dát.

Atribúty hits.home a hits.away

charakteristika: atribút obsahuje hodnoty počtu narazení domáceho a hosťujúceho tímu. Narazenia, resp. osobné súboje vznikajú počas hry a pripočítavajú sa na stranu tímu, ktorý narazenie inicioval. Preto sa hodnoty medzi súperiacimi tímami nemusia rovnať.

Hide

```
summary(df$hits.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   18.00   23.00   23.43   28.00   80.00       4 
```

Hide

```
summary(df$hits.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   16.00   21.00   22.44   27.00   80.00       4 
```

Zo zhrnutia môžeme vidieť, že hodnoty môžu byť pravdivé, resp. neobsahujú čísla, ktoré by logicky nemohli nastať. Taktiež vidíme, že v štyroch záznamoch nám chýbajú hodnoty tohoto atribútu.

Hide

```
df[is.na(df$hits.home)]
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	*NA*	*NA*	
20172018	P	0	0	52	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	*NA*	*NA*	
20162017	P	0	0	3	FALSE	*NA*	*NA*	

4 rows | 1-9 of 31 columns

Hide

```
df[is.na(df$hits.away)]
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	*NA*	*NA*	
20172018	P	0	0	52	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	*NA*	*NA*	

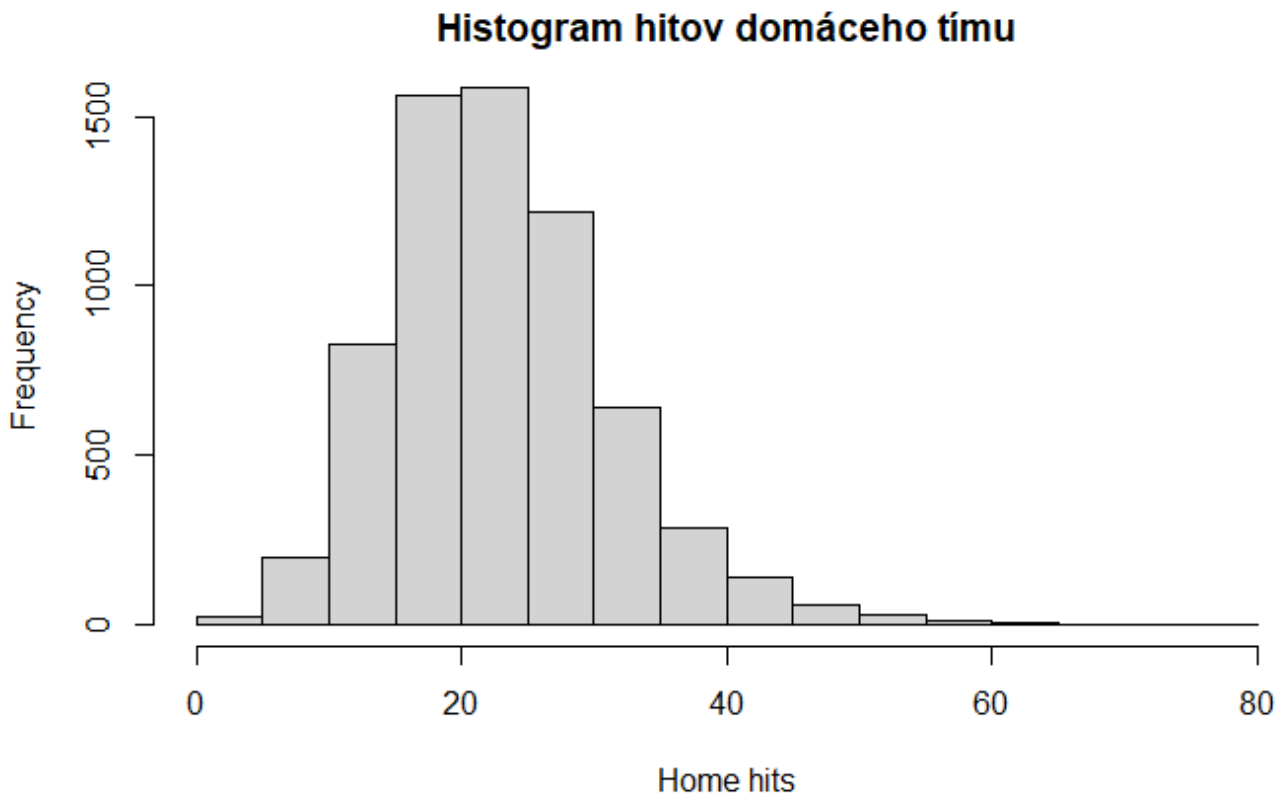
season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20162017	P	0	0	3	FALSE	*NA*	*NA*	

4 rows | 1-9 of 31 columns

Väčšina zo záznamov s chýbajúcimi hodnotami, je z jednej sezóny. Vyzerá to ako preložené, alebo zrušené zápasy a budeme sa na to musieť pozrieť.

Hide

```
hist(df$hits.home, xlab="Home hits", main="Histogram hitov domáceho tímu")
```



Hide

```
hist(df$hits.away, xlab="Away hits", main="Histogram hitov hosťujúceho tímu")
```

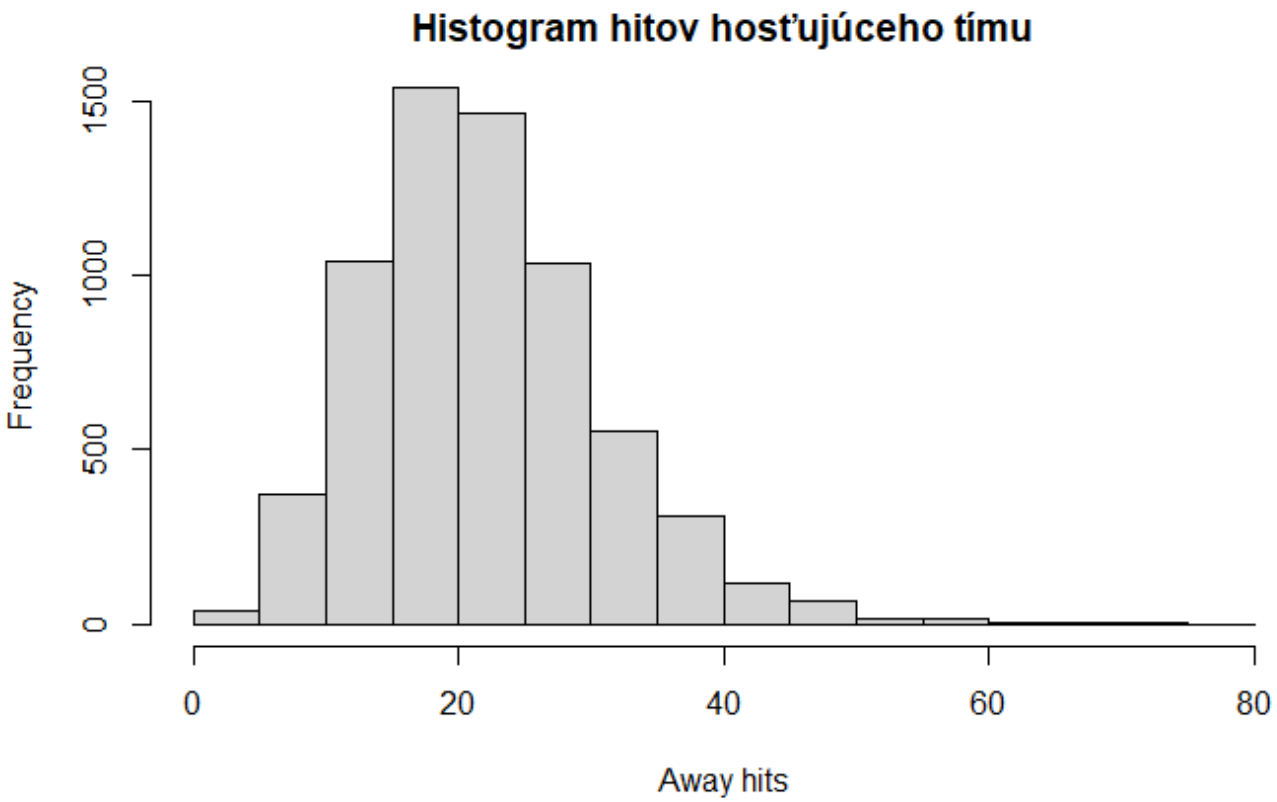
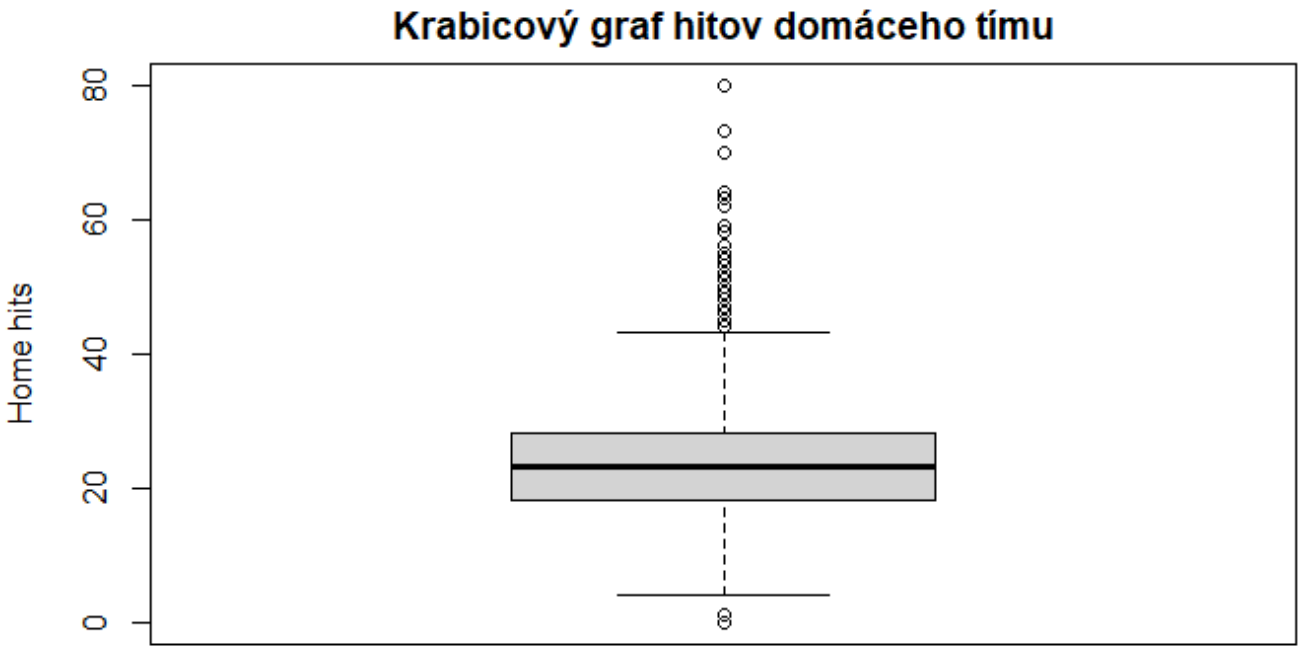


Diagram pripomína normálne rozdelenie hodnôt, ale obsahuje niekoľko vychýlených hodnôt, ktoré mohli nastať z dôvodu dlhších zápasov, napr. predĺženia. Pre tento atribút bude vhodné vykonať Shapiro-Wilkov test normality, aby sme sa uistili, že rozdelenie normálne naozaj nebude.

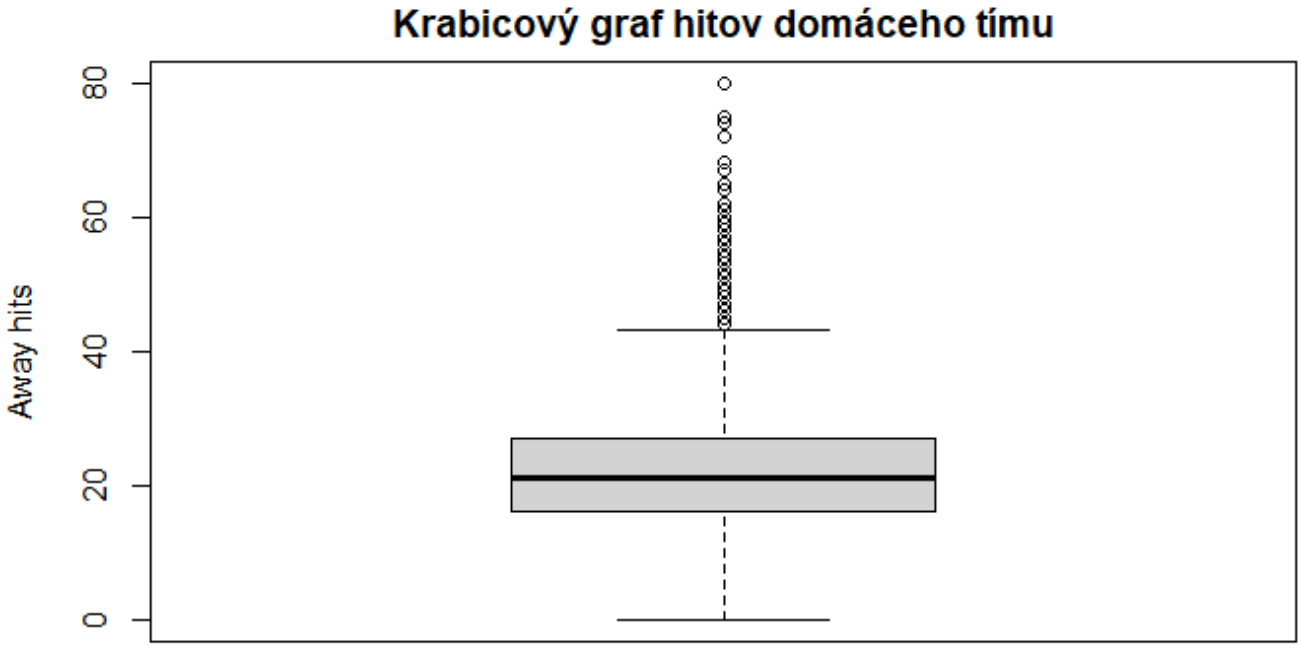
Hide

```
boxplot(df$hits.home,  ylab="Home hits", main="Krabicový graf hitov domáceho tímu")
```



Hide

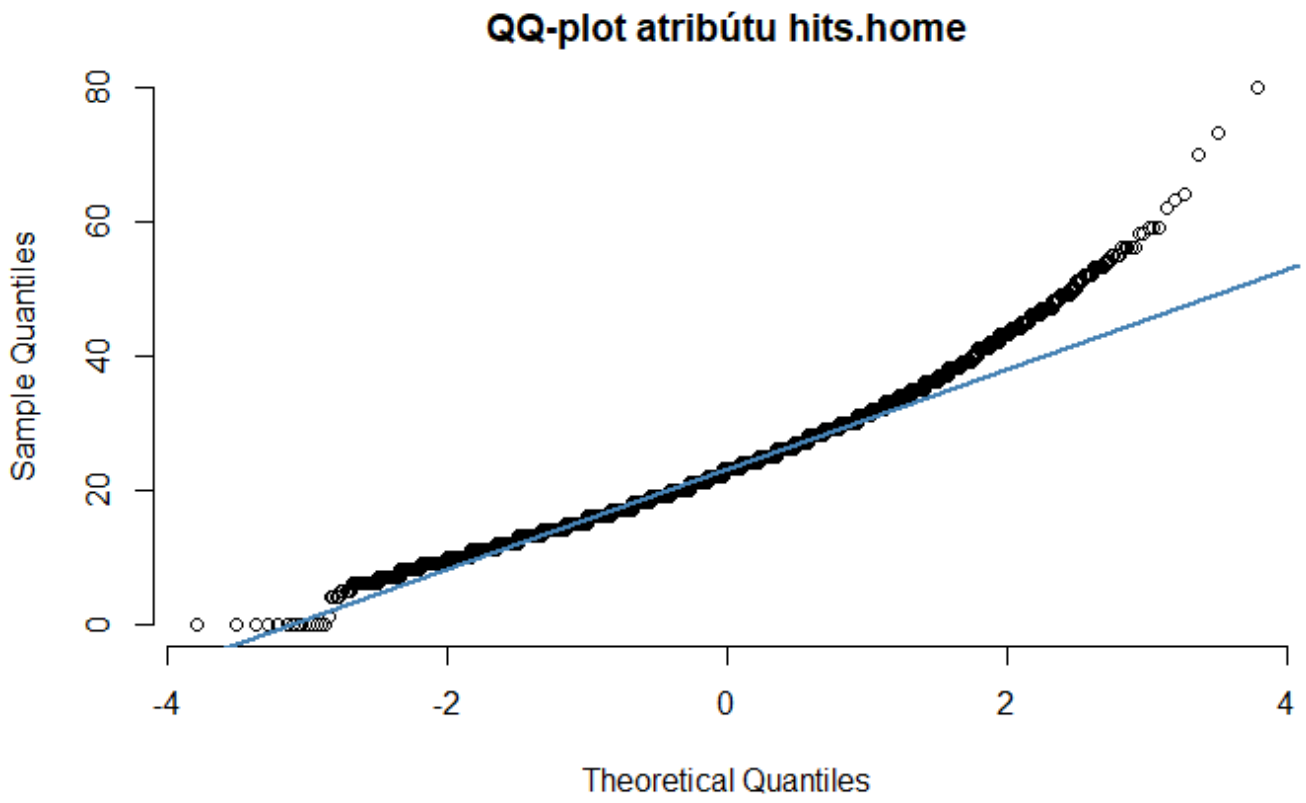
```
boxplot(df$hits.away, ylab="Away hits", main="Krabicový graf hitov domáceho tímu")
```



Vidíme, že niektoré hodnoty sú veľmi vychýlené a mali by sme ich zmenšiť. Takéto vychýlené hodnoty nastávajú pri netradičných zápasoch, alebo derby, ktré sa konajú na zamrznutých jazerách. Preto môžeme hodnoty znížiť, alebo zmeniť na najčastejšiu hodnotu. Hodnoty ale nie sú extrémne vychýlené, v intervale medzi 50-80 je ich pomerne dosť a sú aj relatívne blízko seba).

Hide

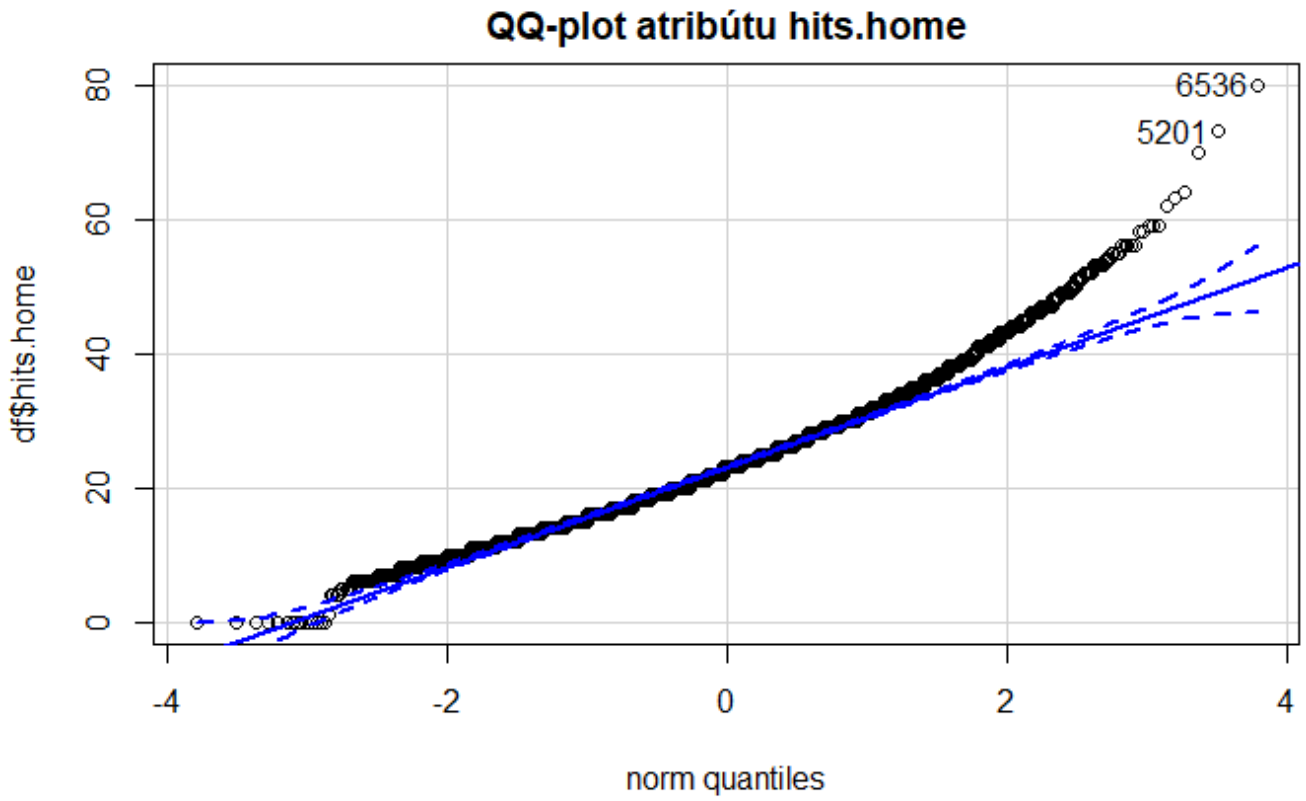
```
qqnorm(df$hits.home, pch = 1, frame = FALSE, main="QQ-plot atribútu hits.home")
qqline(df$hits.home, col = "steelblue", lwd = 2)
```



Hide

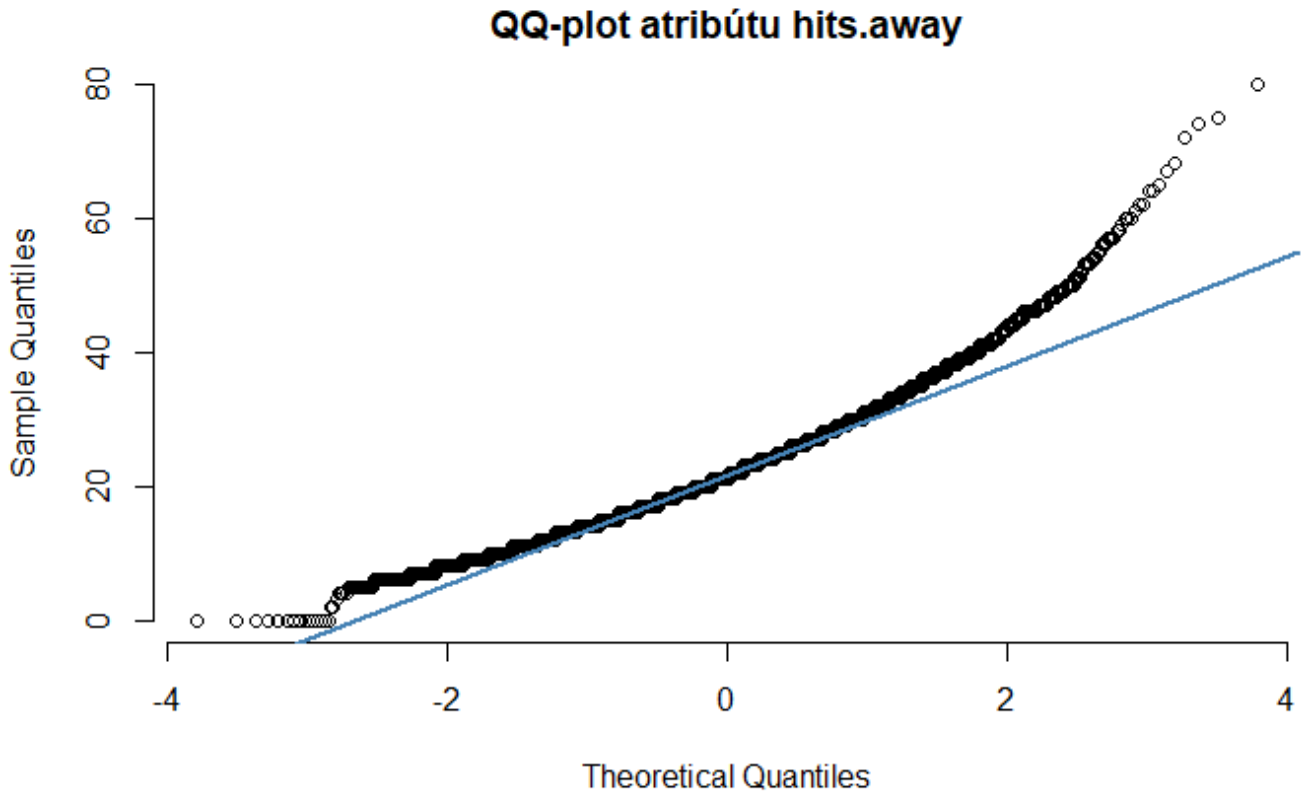
```
qqPlot(df$hits.home, main="QQ-plot atribútu hits.home")
```

```
[1] 6536 5201
```



Hide

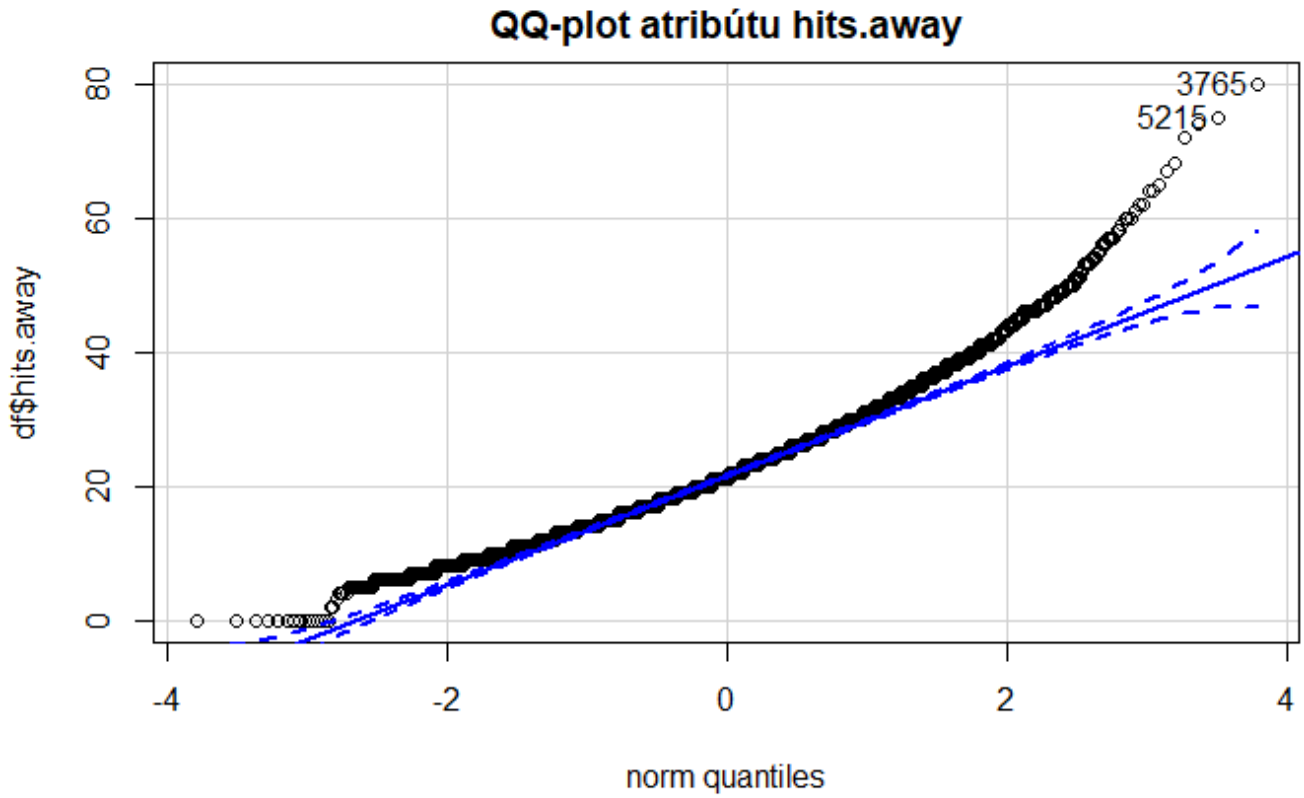
```
qqnorm(df$hits.away, pch = 1, frame = FALSE, main="QQ-plot atribútu hits.away")
qqline(df$hits.away, col = "steelblue", lwd = 2)
```



Hide

```
qqPlot(df$hits.away, main="QQ-plot atribútu hits.away")
```

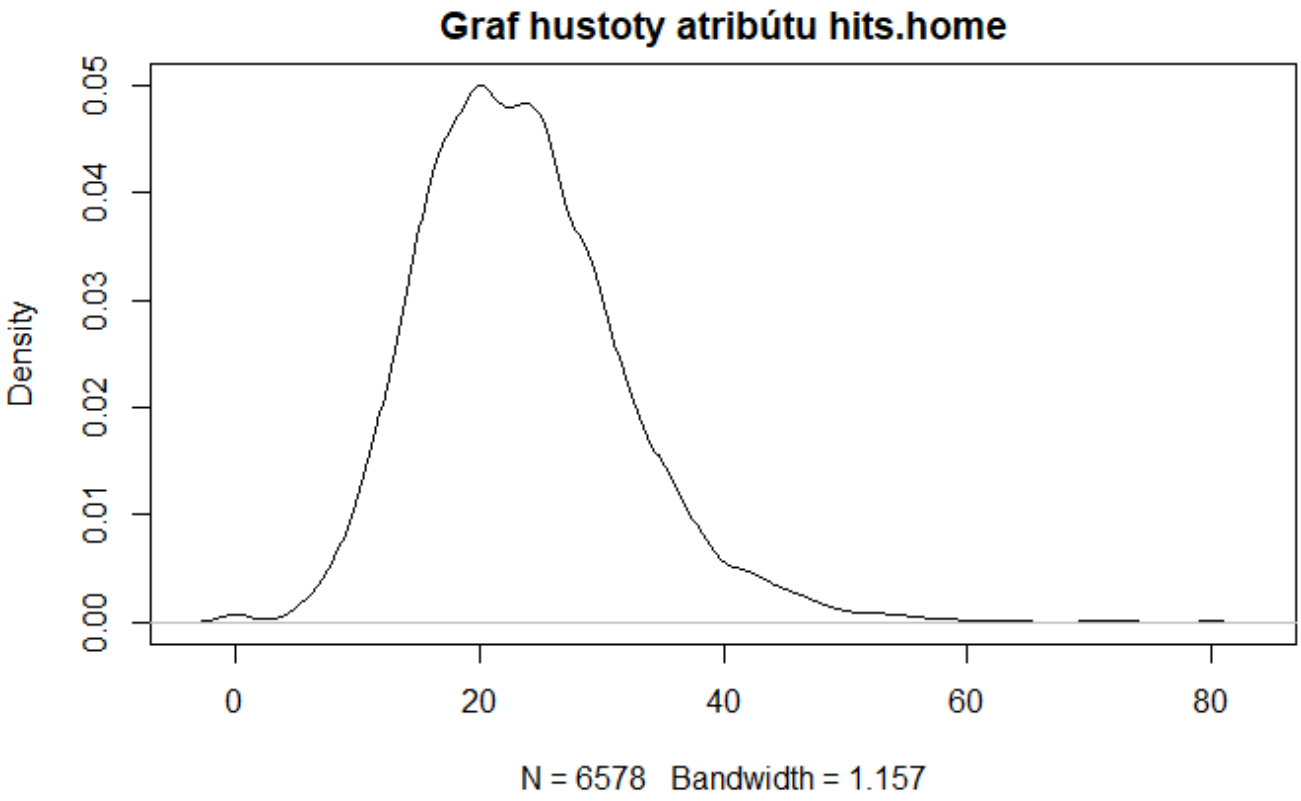
```
[1] 3765 5215
```



Z grafov môžeme vidieť, že hodnoty sa približujú normálnemu rozdeleniu, ale pri vyšších hodnotách sú odchylky vyššie. Môžeme analyzovať tieto hodnoty a skúsiť ich normalizovať.

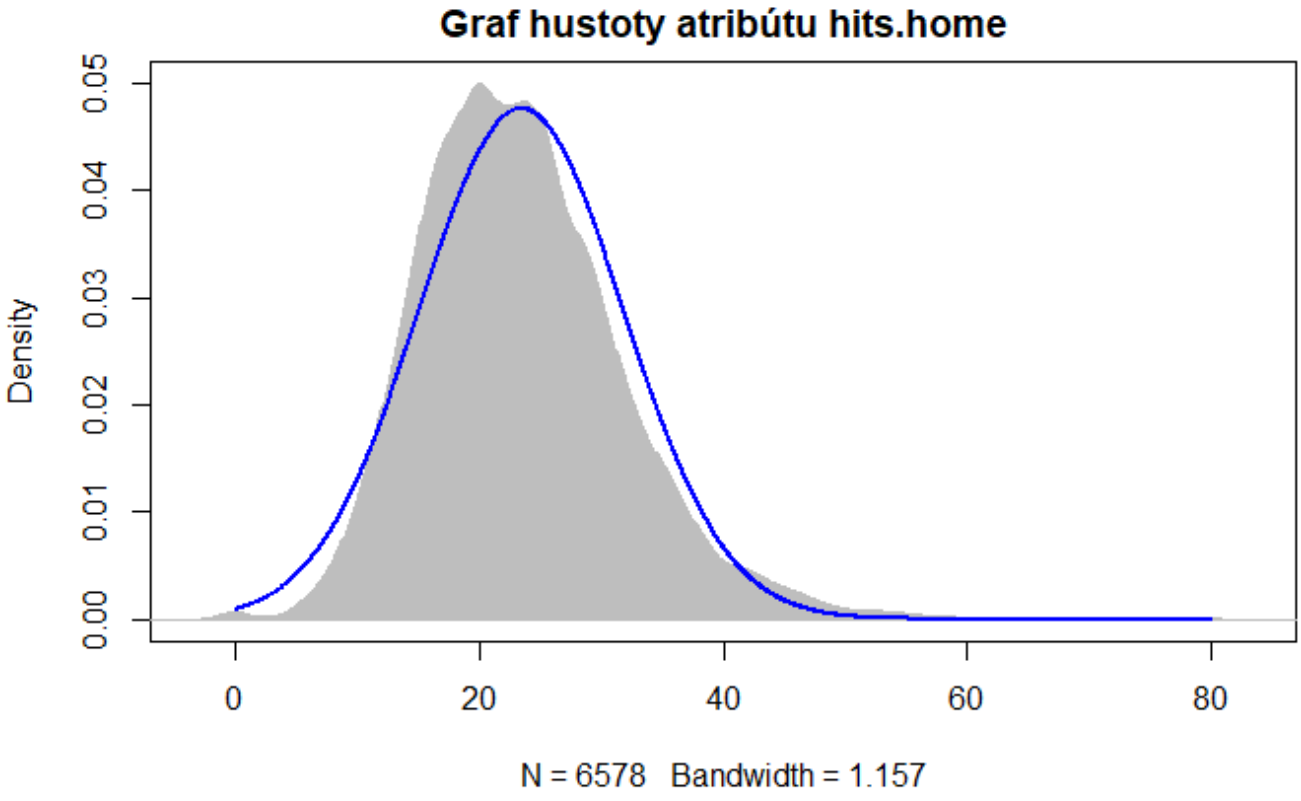
Hide

```
plot(density(na.omit(df$hits.home)), main="Graf hustoty atribútu hits.home")
```



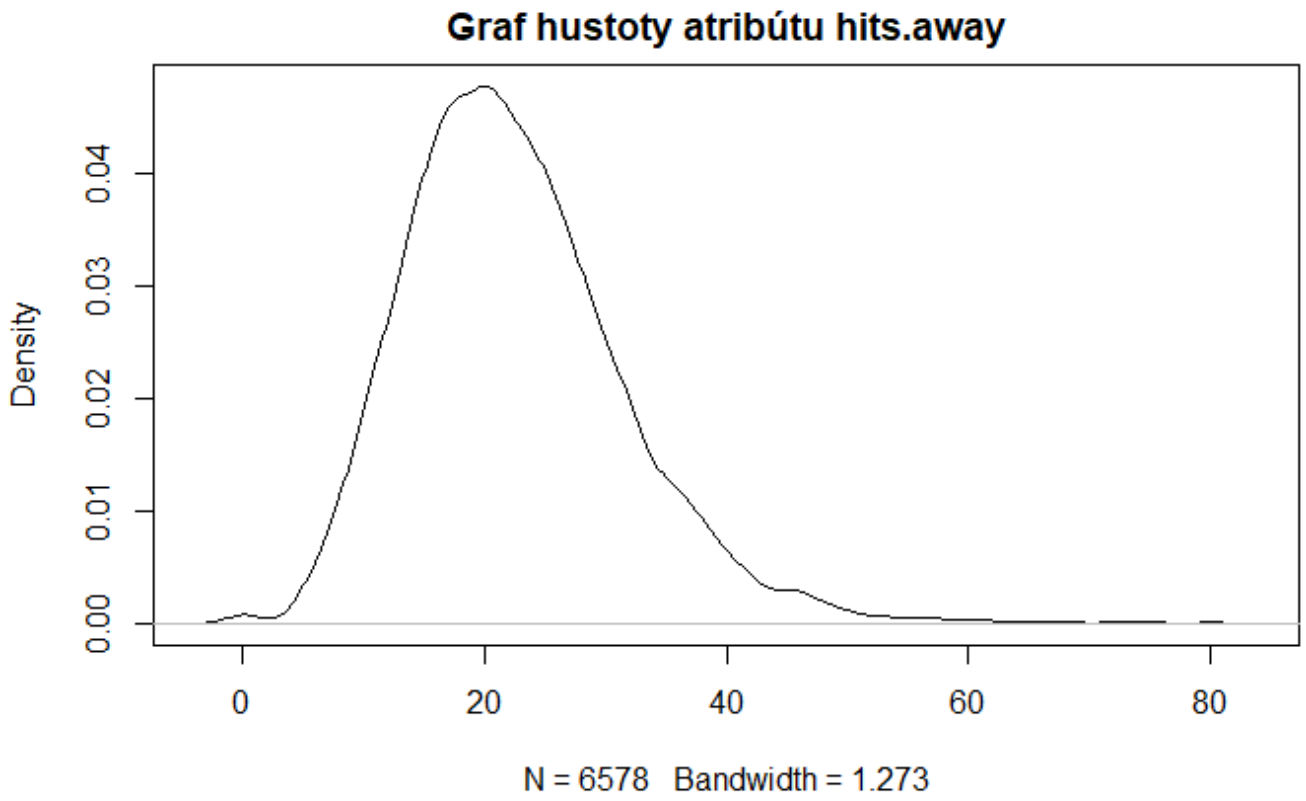
Hide

```
plotNormalDensity(df$hits.home, main="Graf hustoty atribútu hits.home")
```



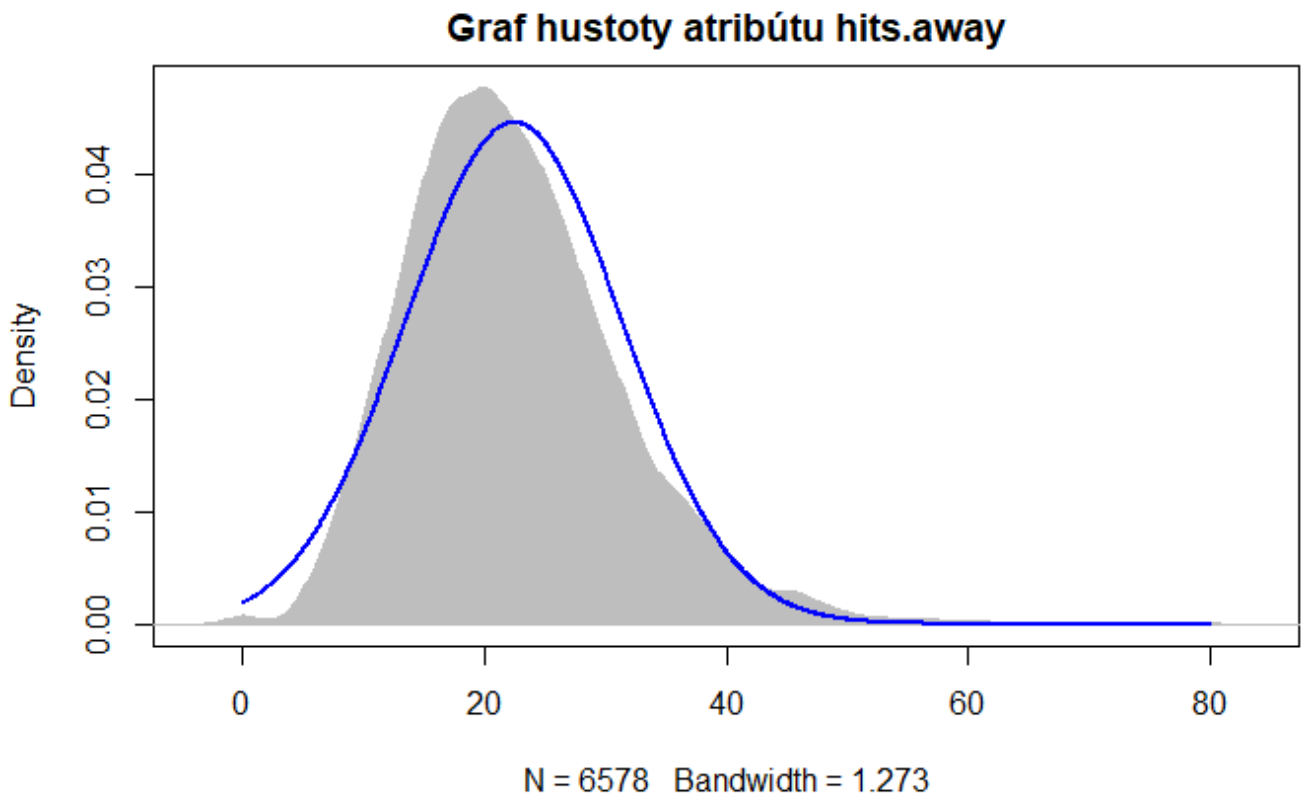
Hide

```
plot(density(na.omit(df$hits.away)), main="Graf hustoty atribútu hits.away")
```



Hide

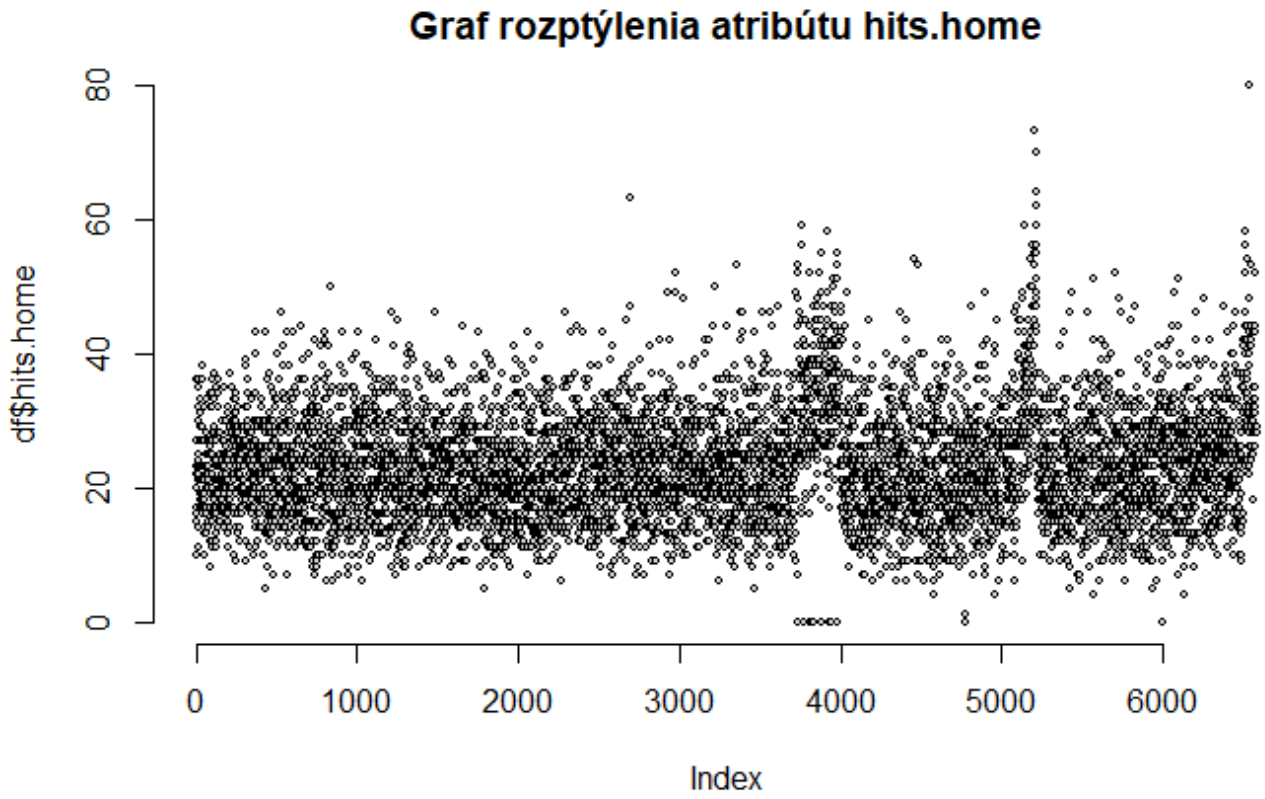
```
plotNormalDensity(df$hits.away, main="Graf hustoty atribútu hits.away")
```



Rozdelenie hodnôt veľmi pripomína normálne rozdelenie, pričom je voči krivke posunuté. Pre tento atribút bude potrebné spraviť test normality, aby sme si boli istý o aké rozdelenie sa jedná.

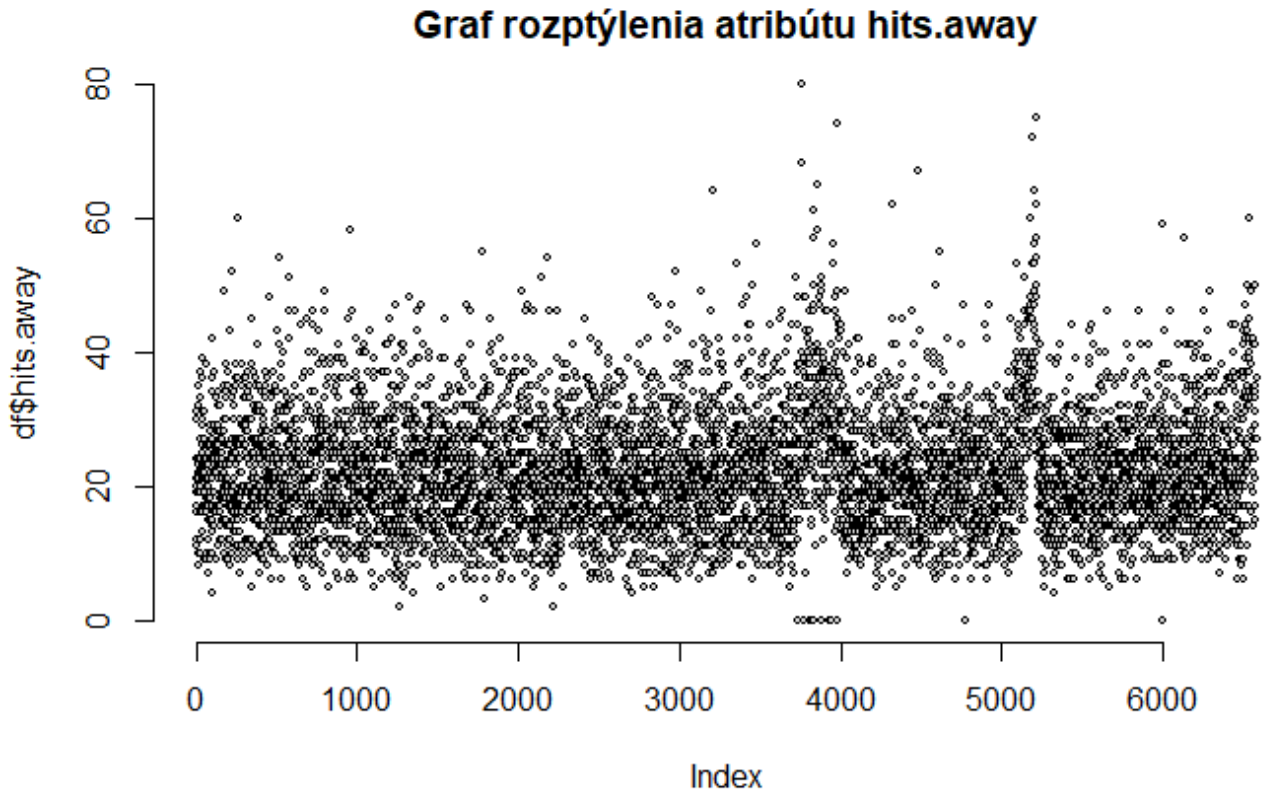
Hide

```
plot(df$hits.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FALS
E, main="Graf rozptýlenia atribútu hits.home")
```



Hide

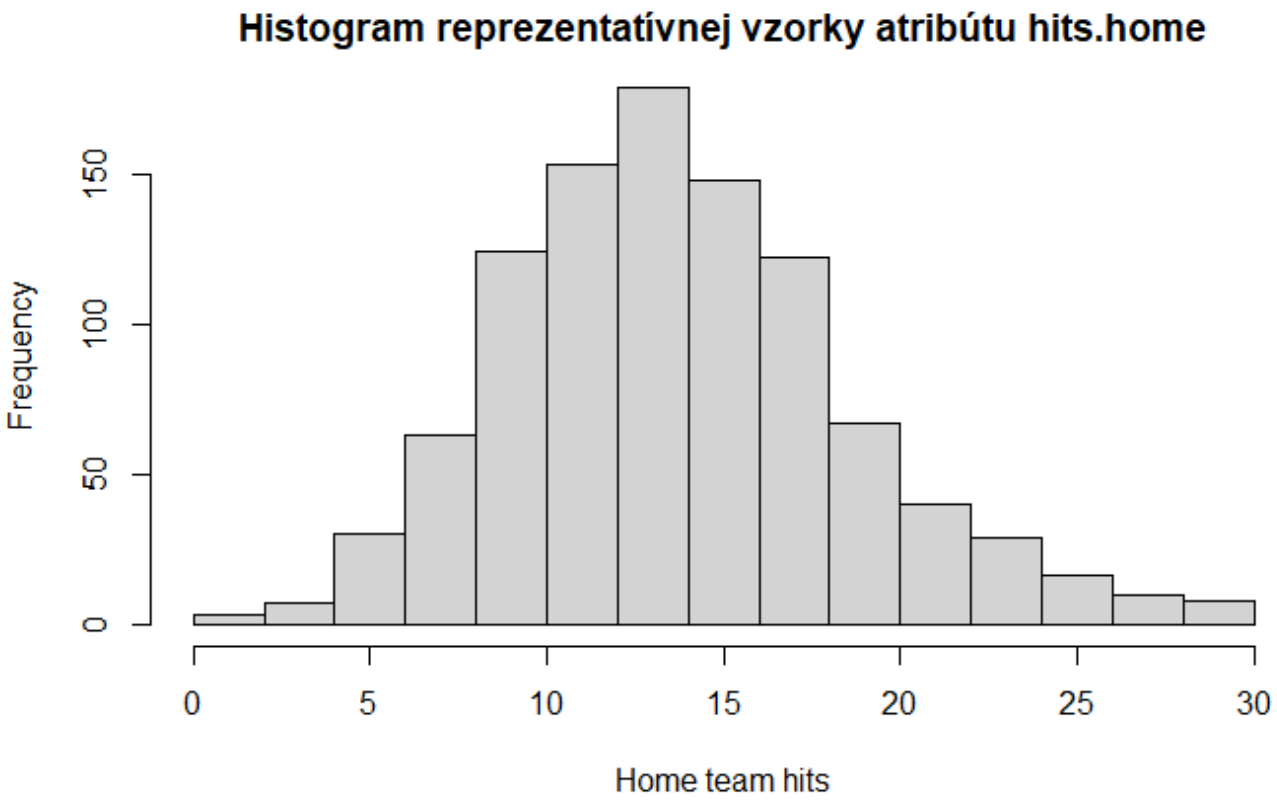
```
plot(df$hits.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FALS
E, main="Graf rozptýlenia atribútu hits.away")
```



Z grafu hustoty vidíme, že hodnoty sú rozmanité, čo logicky sedí k zápasom NHL. No taktiež vidíme, že niektoré hodnoty sú veľmi riedke, hlavne najväčšie hodnoty a najmenšie.

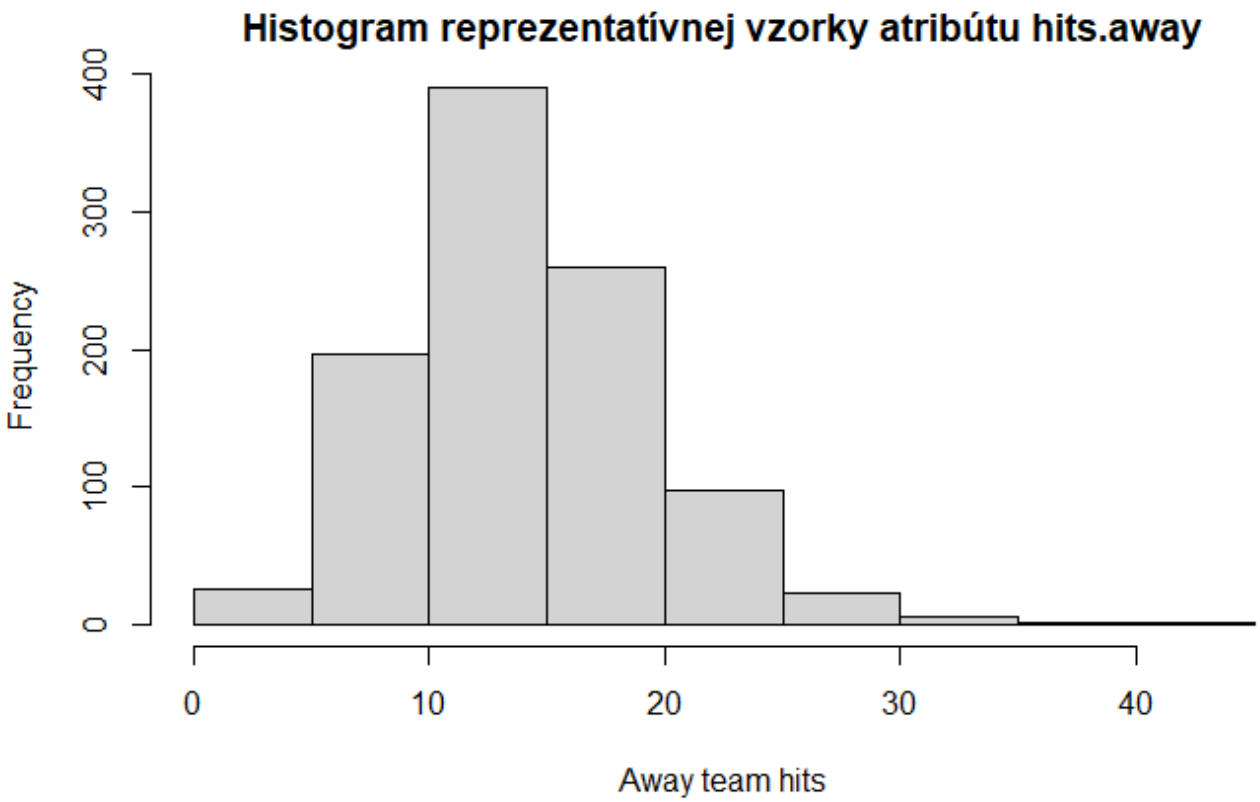
Hide

```
sample <- sample_n(df, 1000)
hist(sample$blocked.home, xlab="Home team hits", main="Histogram reprezentatívnej vzo
rky atribútu hits.home")
```



Hide

```
hist(sample$blocked.away, xlab="Away team hits", main="Histogram reprezentatívnej vzo
rky atribútu hits.away")
```



Hide

```
cat("Štatistika hits.home\n")
```

```
Štatistika hits.home
```

Hide

```
summary(df$hits.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   18.00   23.00   23.43   28.00   80.00       4 
```

Hide

```
cat("Štatistika vzorky hits.home\n")
```

```
Štatistika vzorky hits.home
```

Hide

```
summary(sample$hits.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   18.00   22.00   23.28   28.00   73.00       1 
```

Hide

```
cat("Štatistika hits.away\n")
```

```
Štatistika hits.away
```

Hide

```
summary(df$hits.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   16.00   21.00   22.44   27.00   80.00       4
```

Hide

```
cat("Štatistika vzorky hits.away\n")
```

```
Štatistika vzorky hits.away
```

Hide

```
summary(sample$hits.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   16.00   21.00   22.22   27.00   64.00       1
```

Vzorka je na základe interpretácie základnej štatistiky reprezentatívna, vykonáme test normality s hladinou p = 0.05.

Hide

```
shapiro.test(sample$hits.home)
```

```

    Shapiro-Wilk normality test

data:  sample$hits.home
W = 0.95693, p-value < 2.2e-16
```

Hide

```
shapiro.test(sample$hits.away)
```

```
    Shapiro-Wilk normality test

data:  sample$hits.away
W = 0.97076, p-value = 2.686e-13
```

V oboch prípadoch zamietame nulovú hypotézu, rozdelenie z ktorého pochádzajú atribúty **hits.home* a** hits.away** nie je normálne.

Atribúty pim.home a pim.away

charakteristika: atribút hovorí o počte trestných minút pre jednotlivé tímy. Keďže sú najčastejšie vylúčenia sankciované dvoma minútami, tak sa budú párne hodnoty vyskytovať častejšie. No poznáme aj dlhšie tresty ako 2+2, 5, 5+2, alebo aj netradičné tresty. Vylúčenia do konca zápasu za nešportové správanie sa do tejto štatistiky nezarátavajú - jedná sa teda o reálnu minutáž, ktorú strávi hráč na trestnej lavici a jeho tím je tým pádom oslabený (osobné tresty neoslabujú tím počas celej dĺžky trestu).

Hide

```
summary(df$pim.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   4.000   8.000   8.606  10.000  78.000       4
```

Hide

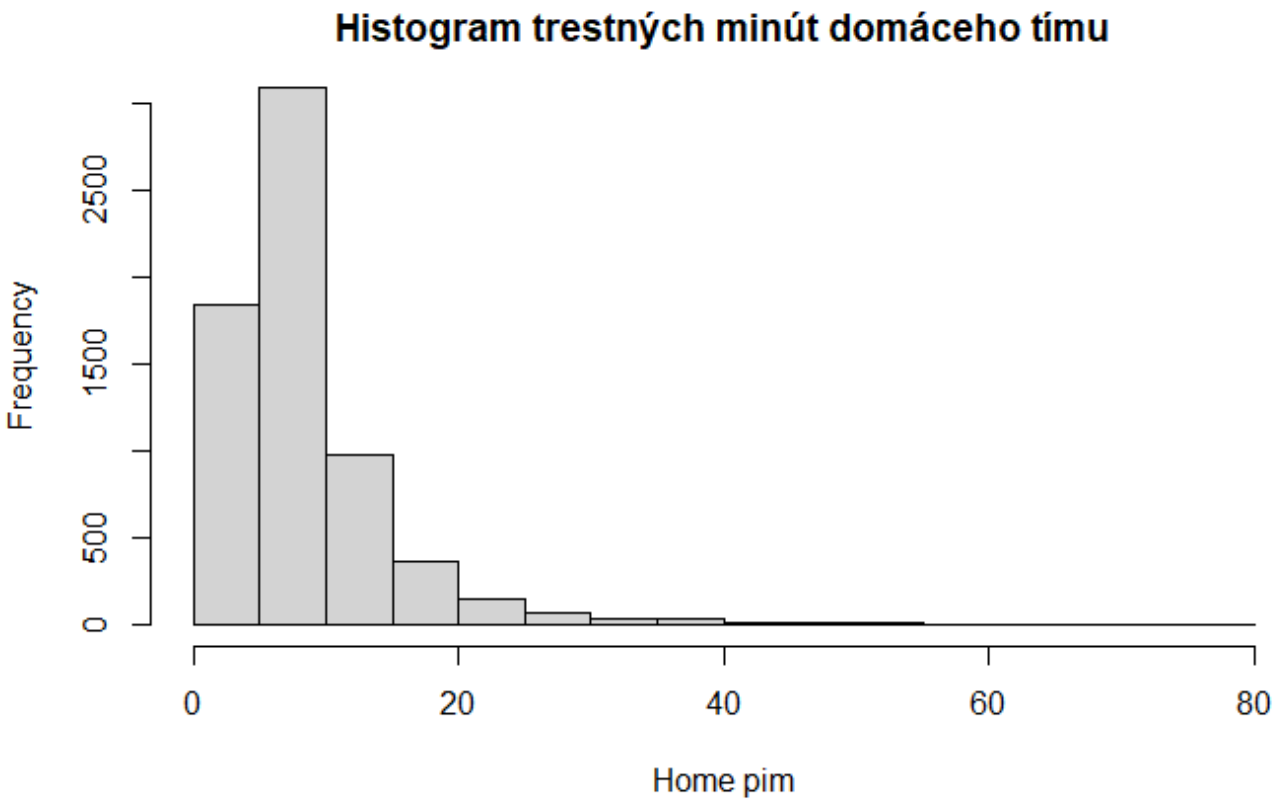
```
summary(df$pim.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   6.000   8.000   9.287  11.000  96.000       4
```

Zo zhodnotenia vidíme, že hodnoty sú priateľné a dávajú logicky zmysel, ale maximálne hodnoty sú veľmi vychýlené a záznamy majú chýbajúcu hodnotu pre tento atribút. Tatkiež sa vyskytujú už známe chýbajúce 4 hodnoty (NA).

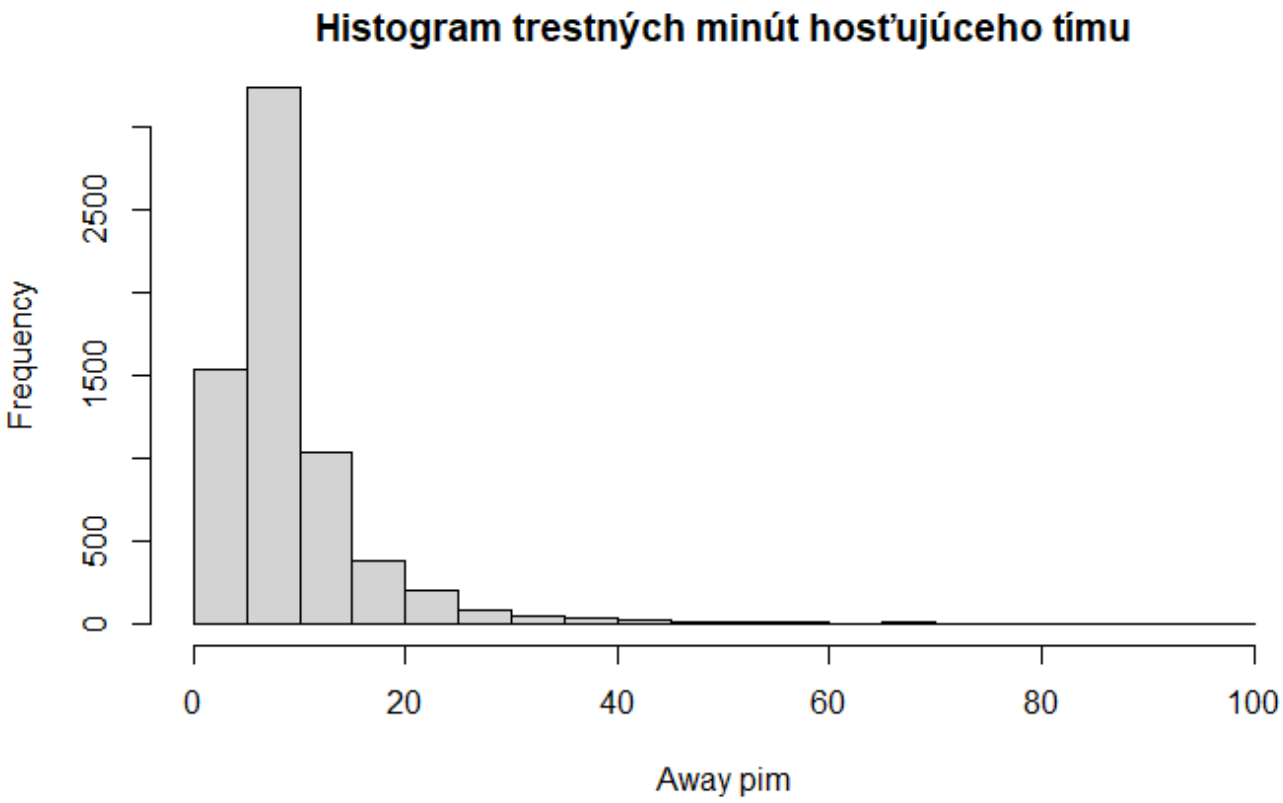
Hide

```
hist(df$pim.home, xlab="Home pim", main="Histogram trestných minút domáceho tímu")
```



Hide

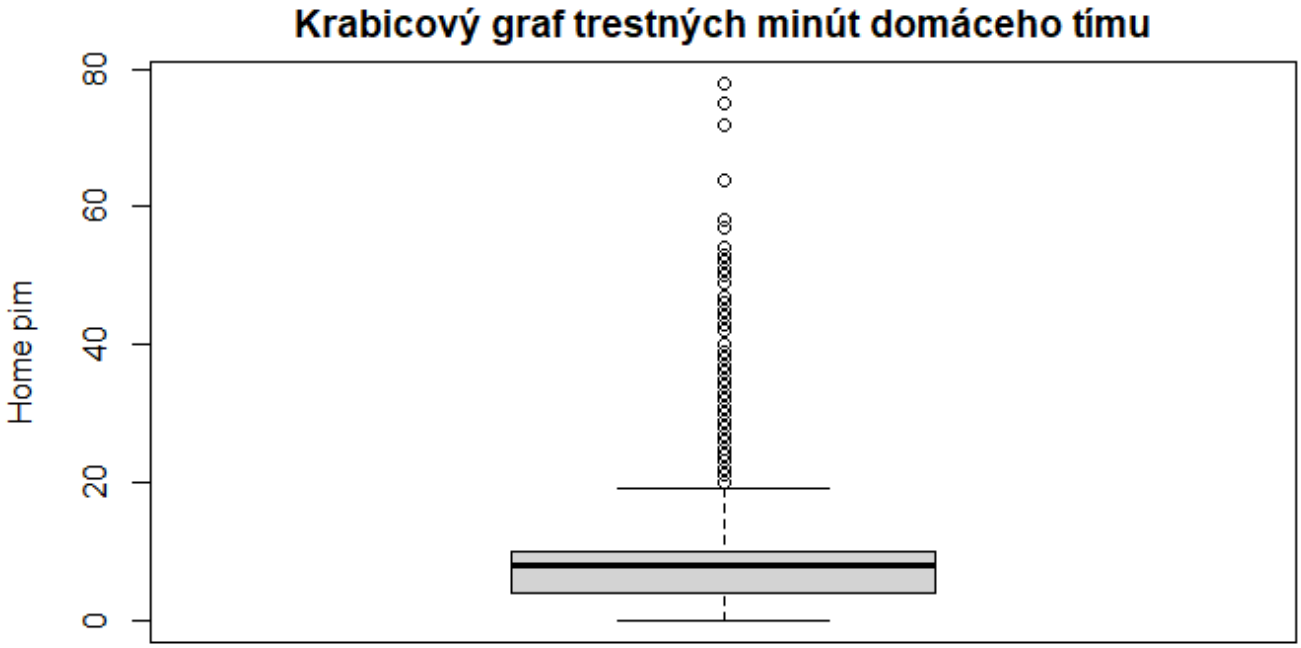
```
hist(df$pim.away,  xlab="Away pim", main="Histogram trestných minút hosťujúceho tím
u")
```



Podľa grafu neobsahuje atribút normálne rozdelenie. Hodnoty skôr pripomínajú long-tail rozdelenie.

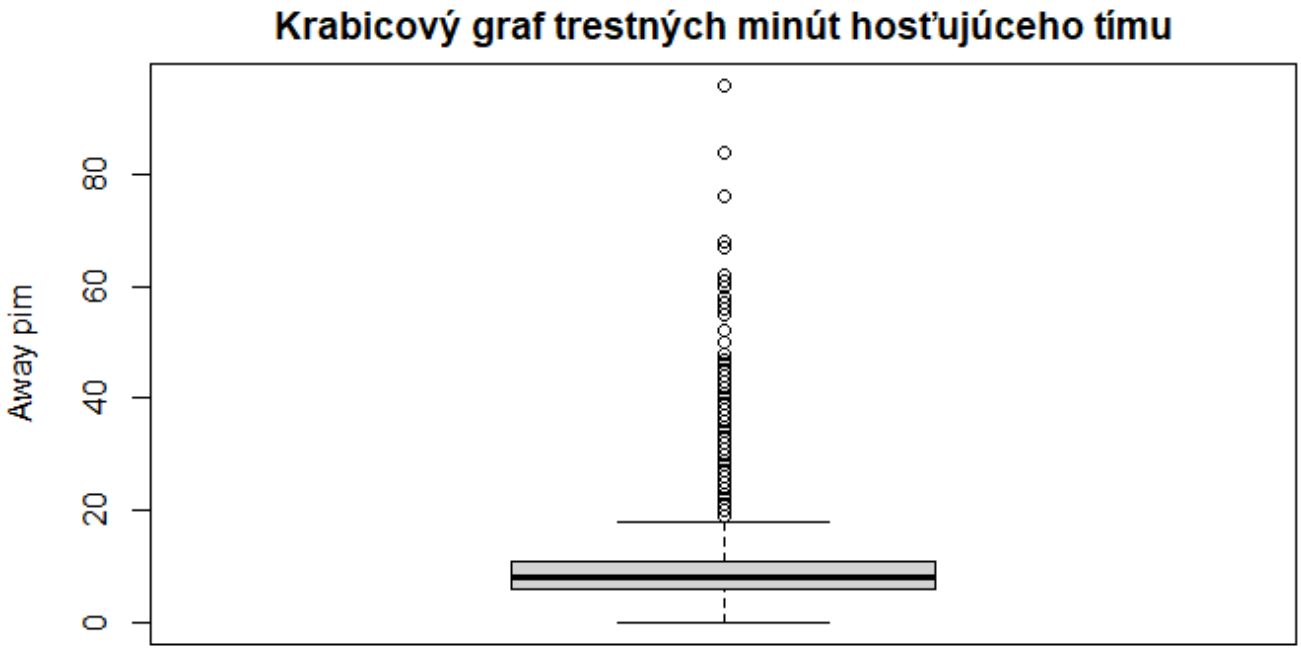
Hide

```
boxplot(df$pim.home, ylab="Home pim", main="Krabicový graf trestných minút domáceho t
ímu")
```



Hide

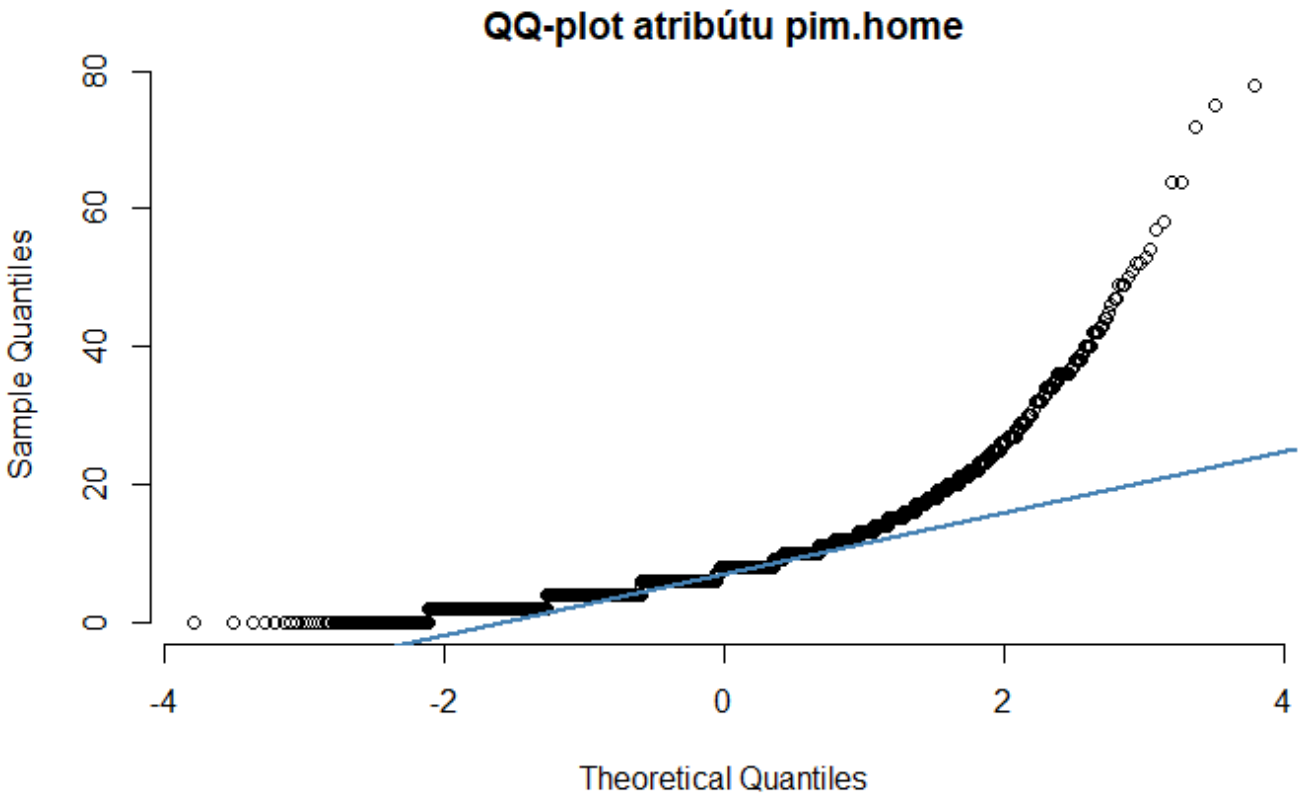
```
boxplot(df$pim.away, ylab="Away pim", main="Krabicový graf trestných minút hosťujúceh
o tímu")
```



Krabicový graf nám iba potvrdil, že atribút obsahuje viacero vychýlených hodnôt, na ktoré sa budeme musieť pozrieť a navrhnúť možnú úpravu. Predbežne odporúčame normalizáciu atribútov pokiaľ sú väčšie rovné ako 60, no túto skutočnosť ešte overíme neskôr na grafe rozptylu.

Hide

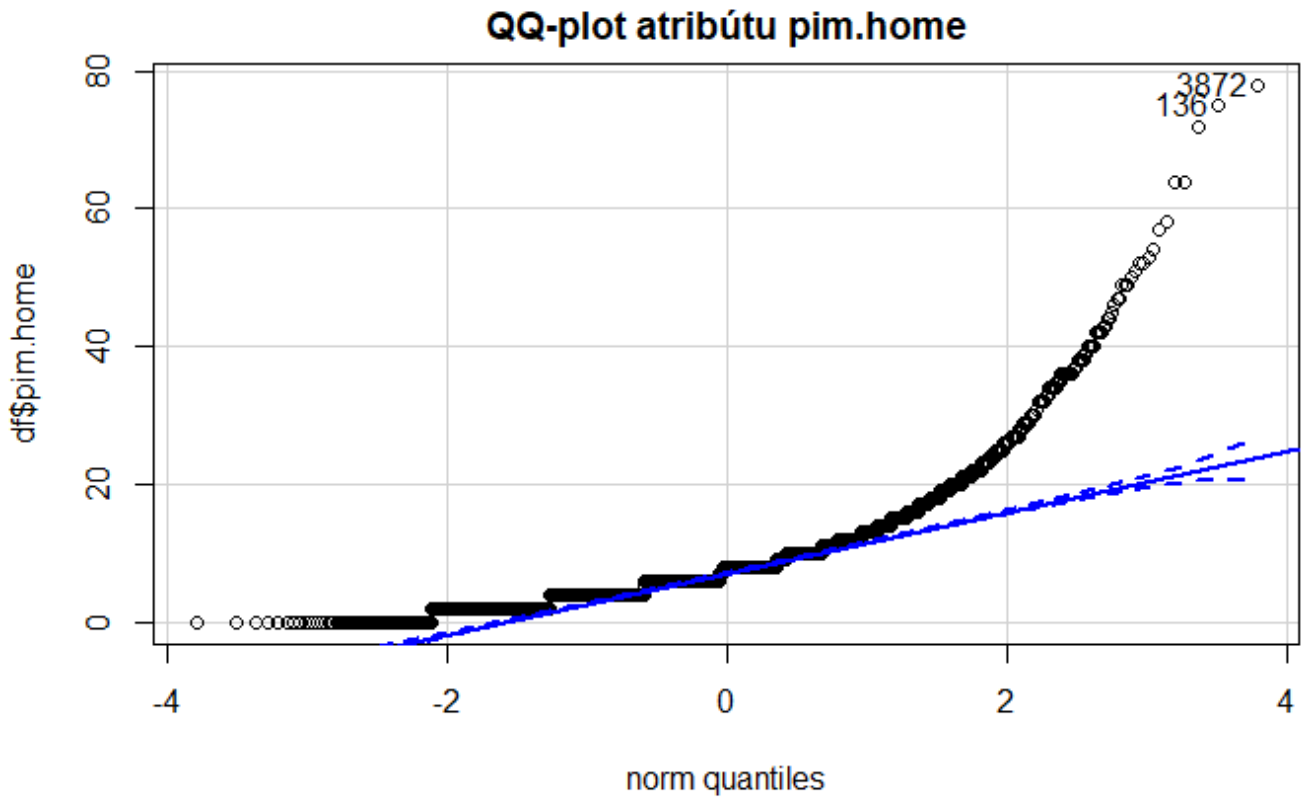
```
qqnorm(df$pim.home, pch = 1, frame = FALSE, main="QQ-plot atribútu pim.home")
qqline(df$pim.home, col = "steelblue", lwd = 2)
```



Hide

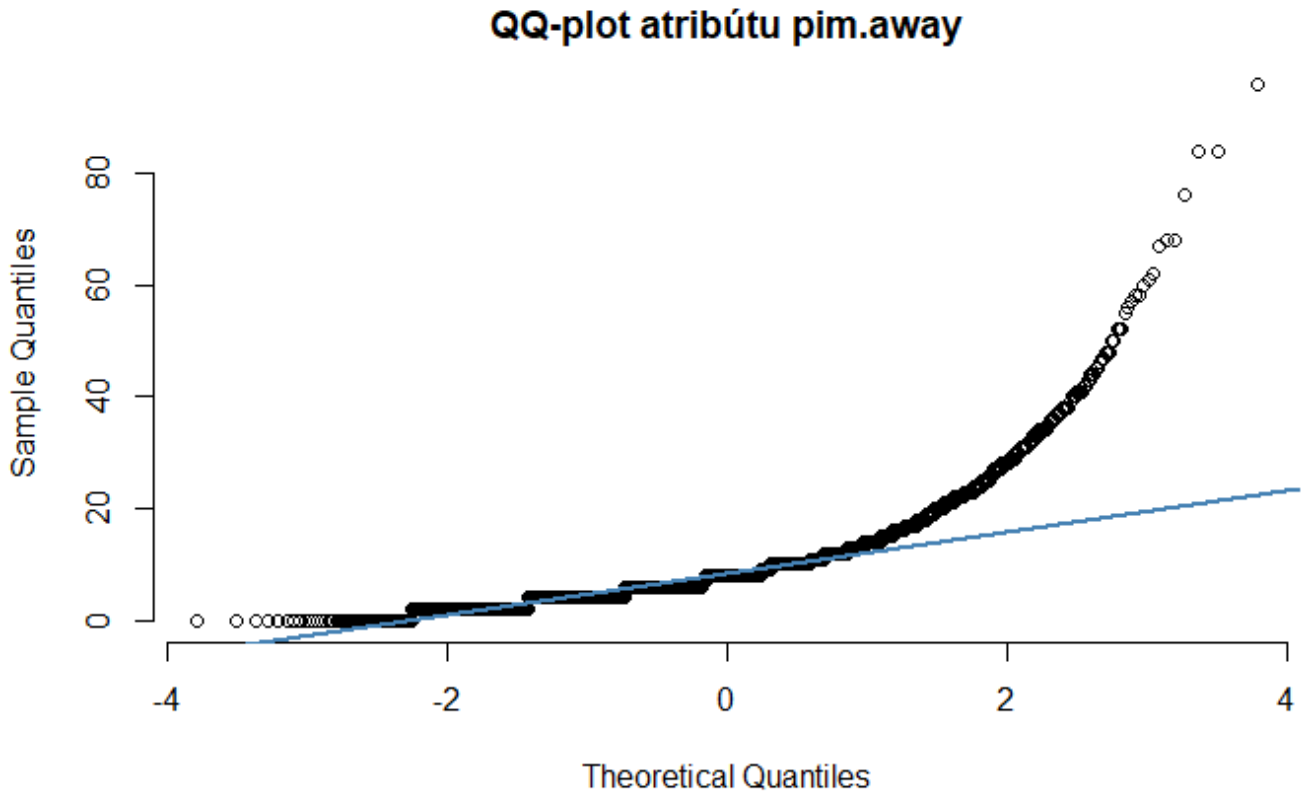
```
qqPlot(df$pim.home, main="QQ-plot atribútu pim.home")
```

```
[1] 3872  136
```



Hide

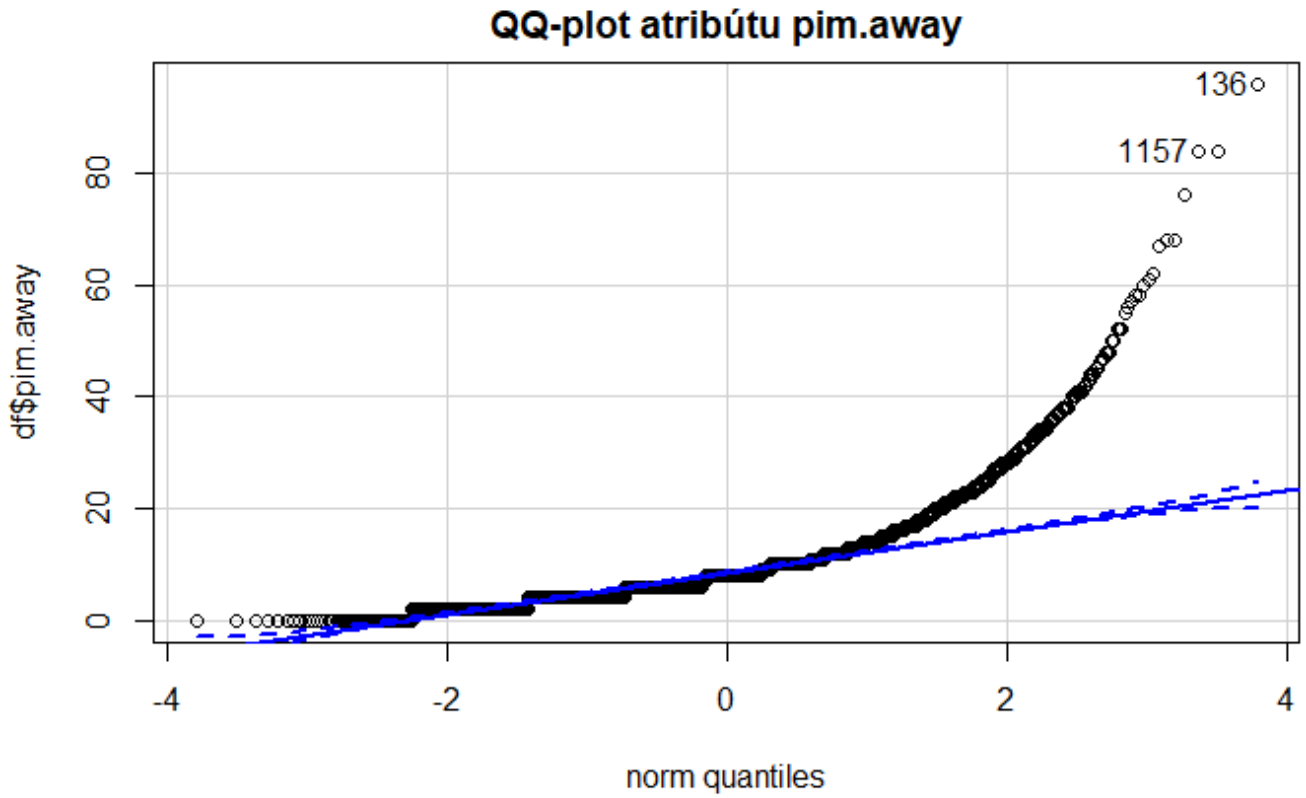
```
qqnorm(df$pim.away, pch = 1, frame = FALSE, main="QQ-plot atribútu pim.away")
qqline(df$pim.away, col = "steelblue", lwd = 2)
```



Hide

```
qqPlot(df$pim.away, main="QQ-plot atribútu pim.away")
```

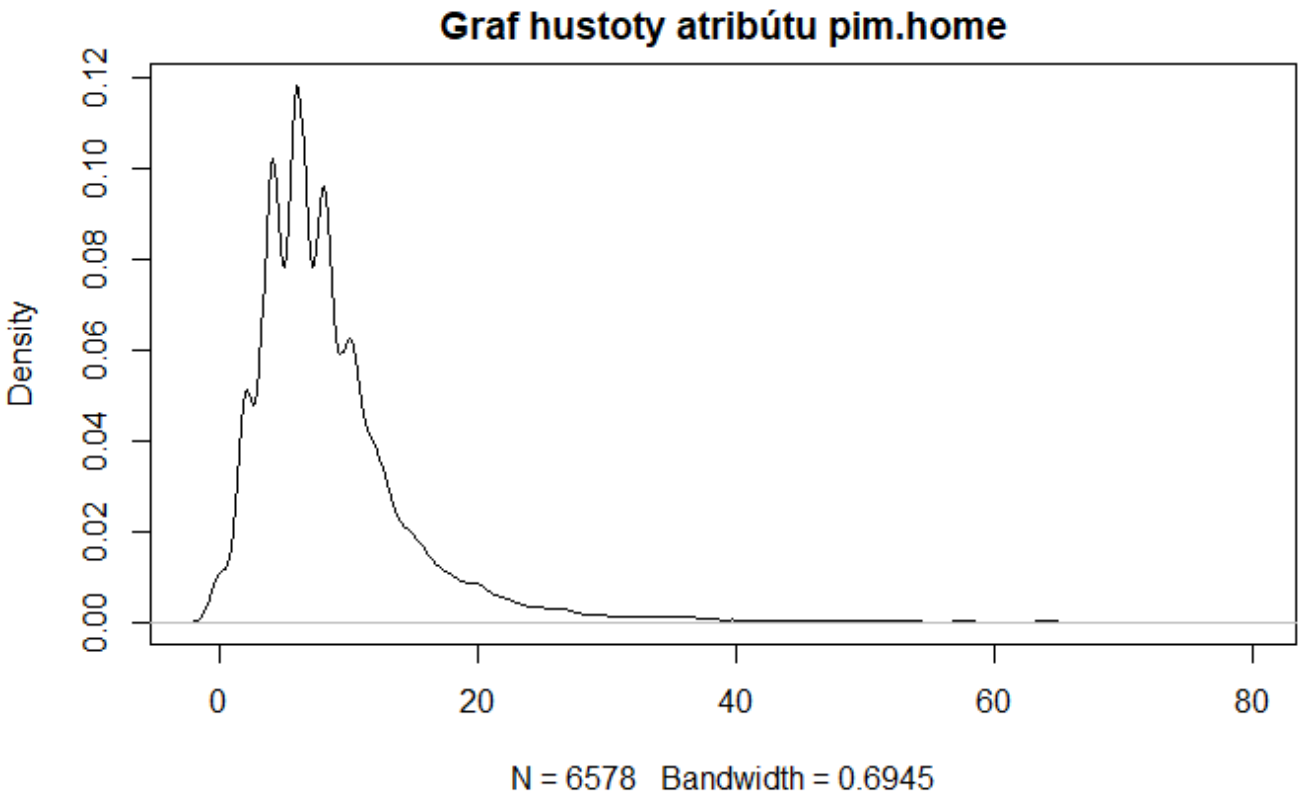
```
[1]  136 1157
```



Q-Q grafy nám dokázali, že sa nejedná o normálne rozdelenie a hodnoty sa mu ani nepribližujú. Test normality teda nebude potrebný.

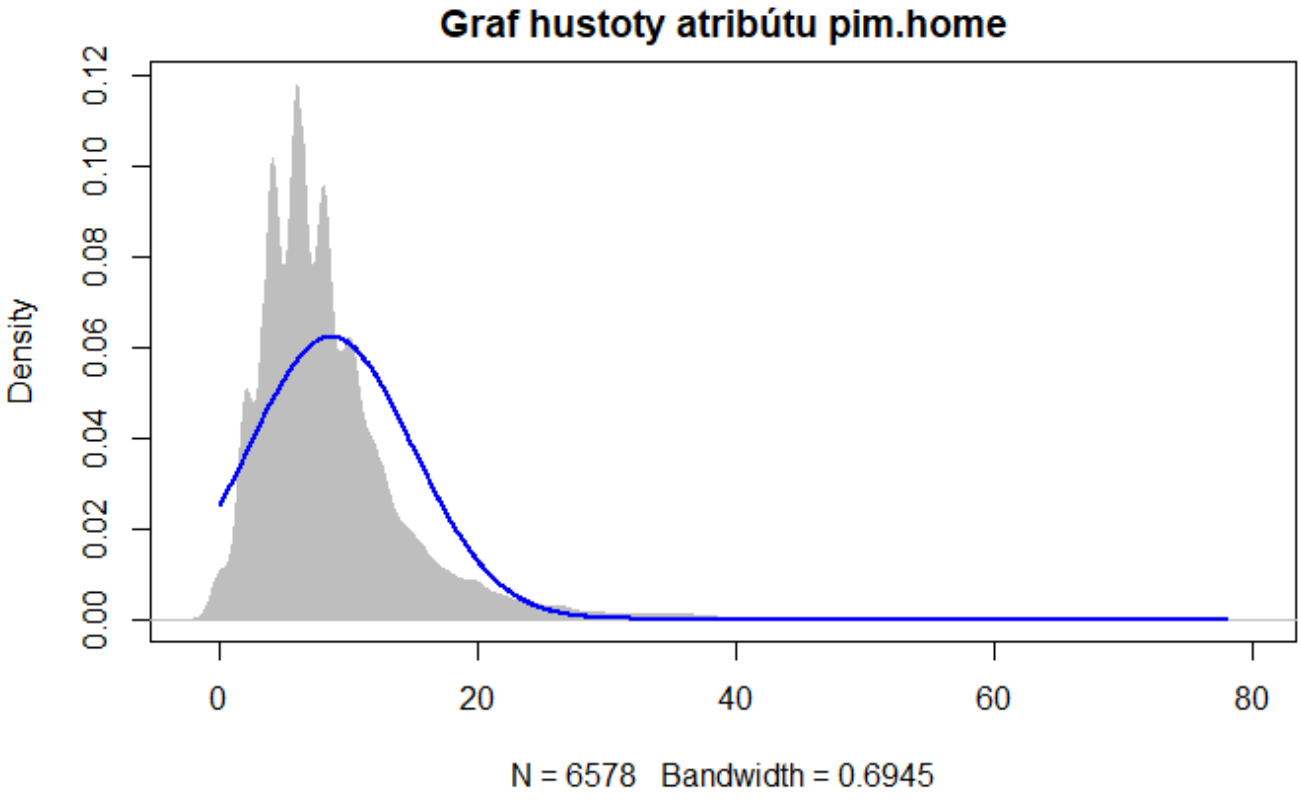
Hide

```
plot(density(na.omit(df$pim.home)), main="Graf hustoty atribútu pim.home")
```



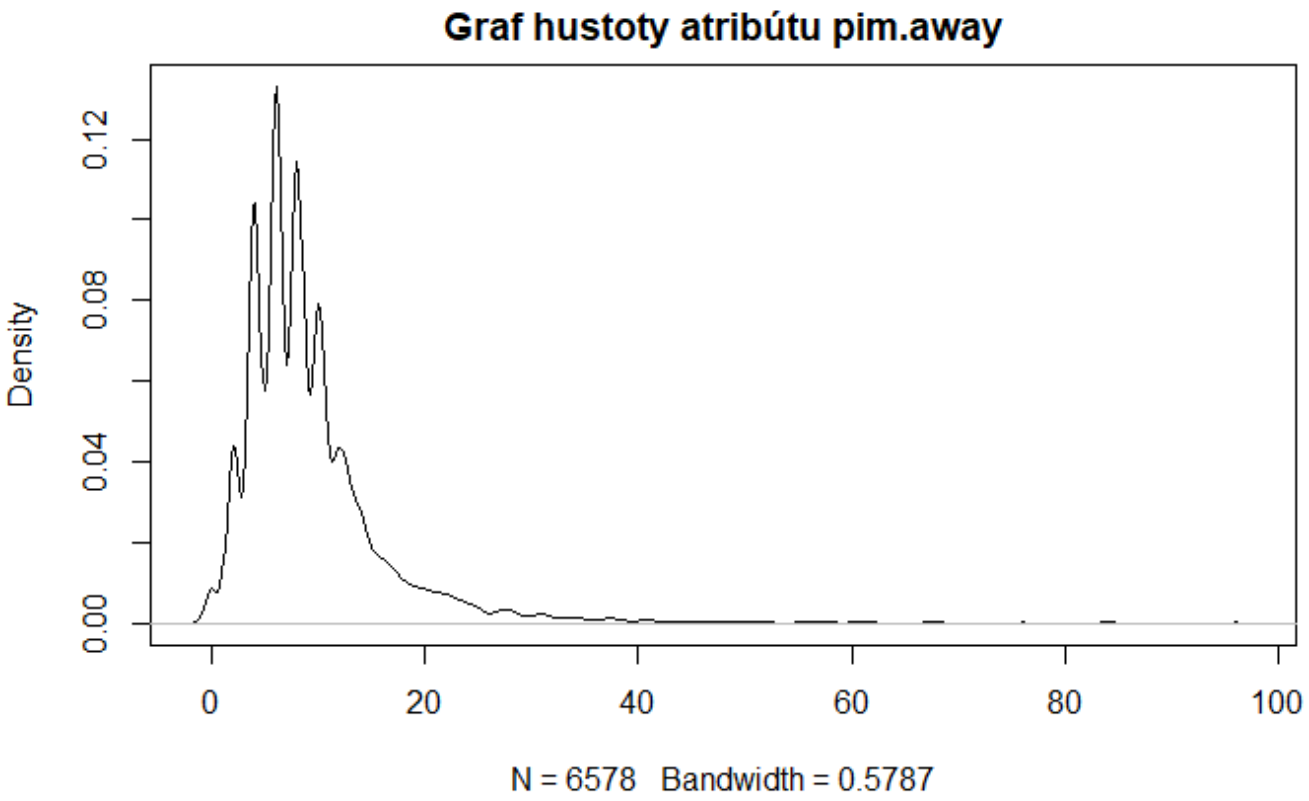
Hide

```
plotNormalDensity(df$pim.home, main="Graf hustoty atribútu pim.home")
```



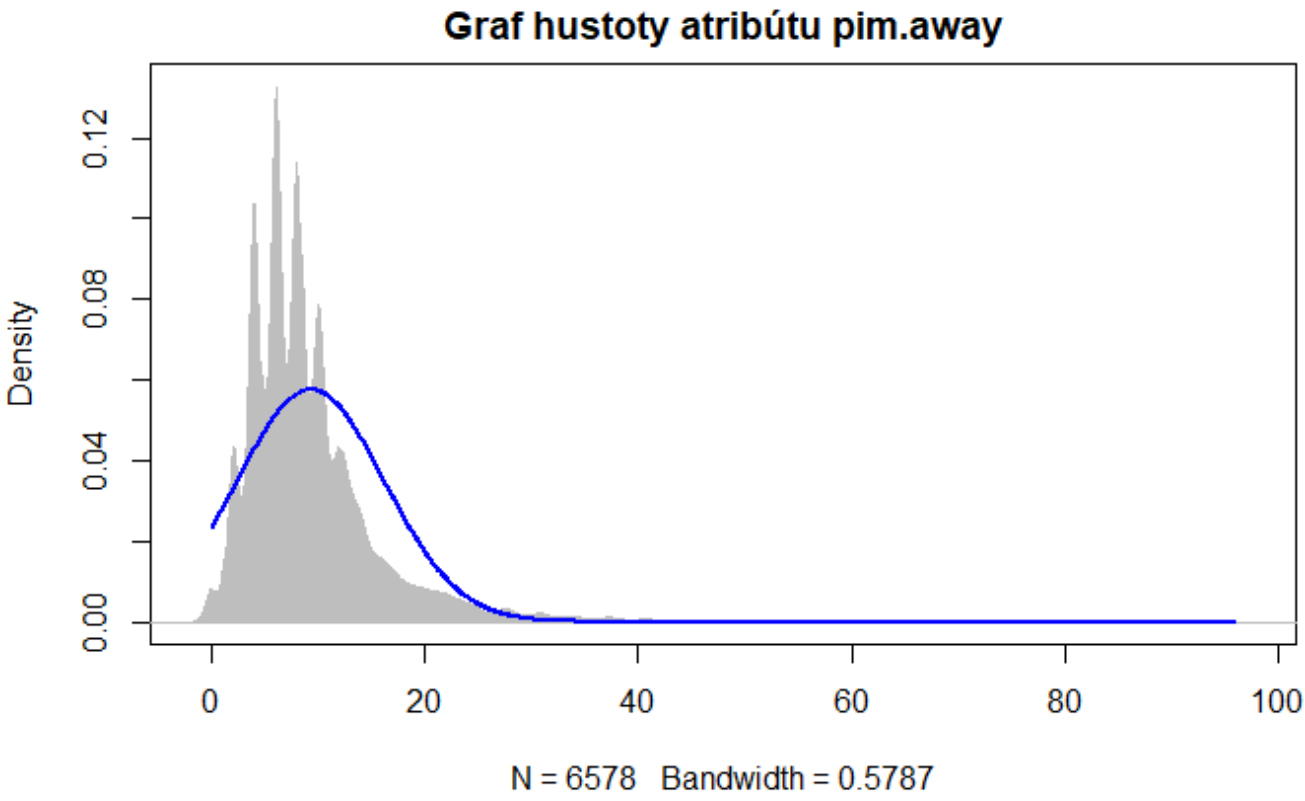
Hide

```
plot(density(na.omit(df$pim.away)), main="Graf hustoty atribútu pim.away")
```



Hide

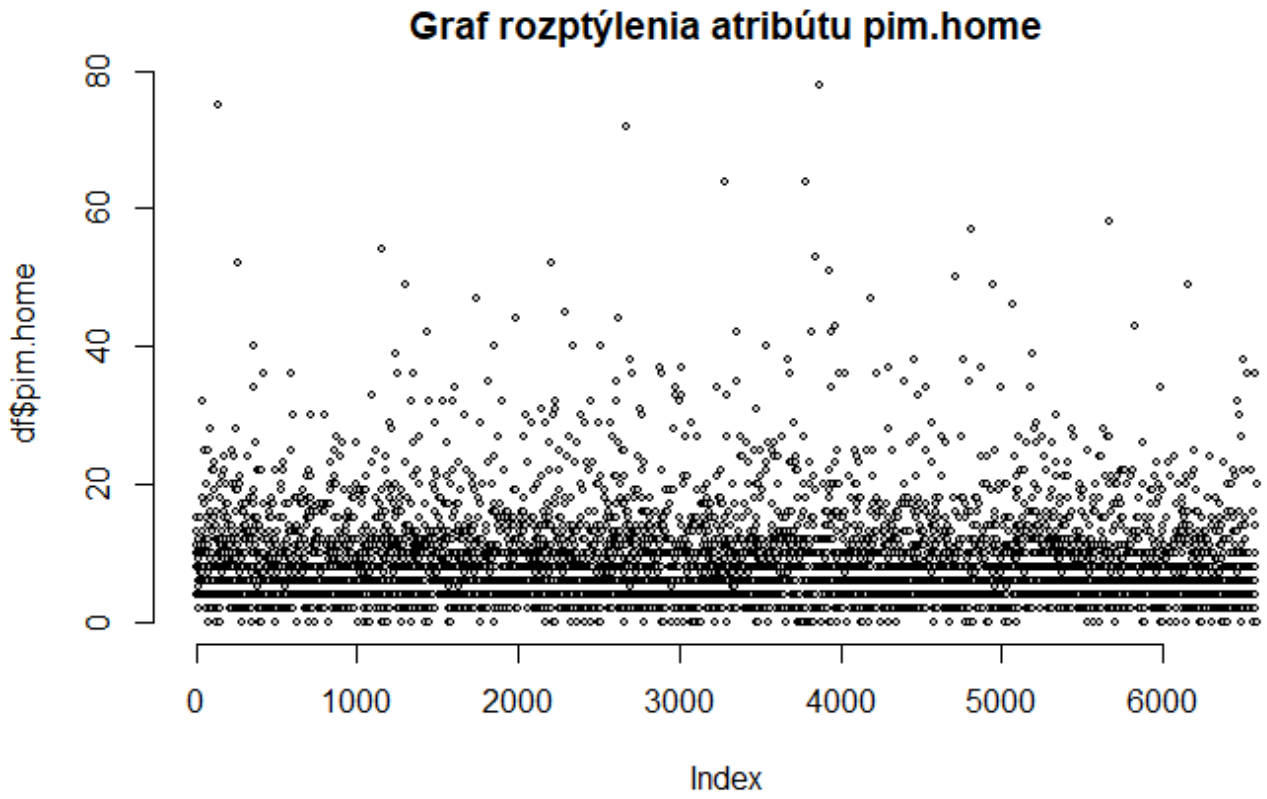
```
plotNormalDensity(df$pim.away, main="Graf hustoty atribútu pim.away")
```



Atribúty majú veľmi nezvyčajú hustotu rozdelenia hodnôt, čo je zapríčinené hlavne z dvôdou dvojminutových trestov. Hustota nepripomína normálne rozdelenie v podstate na žiadnom intervale, čize o normálne rozdelenie sa na základe grafov konať určite nebude.

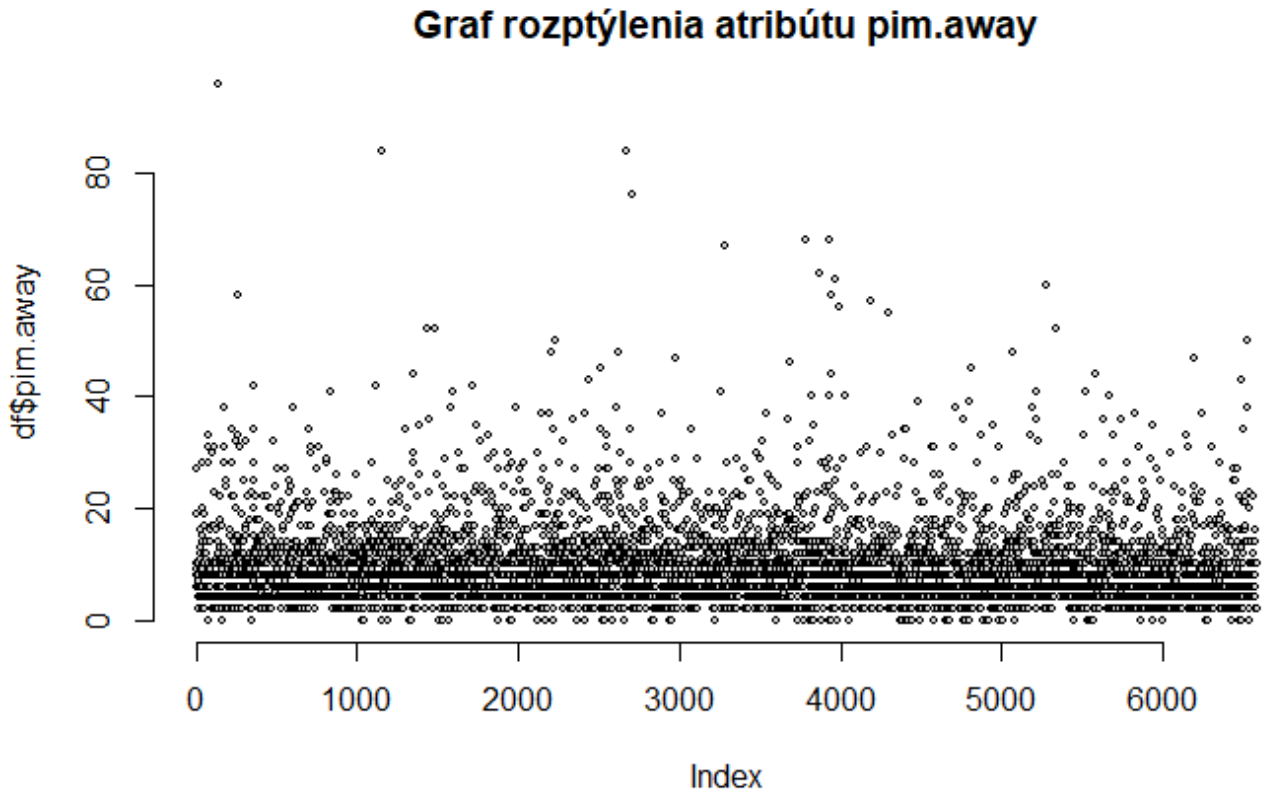
Hide

```
plot(df$pim.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FALS
E, main="Graf rozptýlenia atribútu pim.home")
```



Hide

```
plot(df$pim.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FALS
E, main="Graf rozptýlenia atribútu pim.away")
```



Graf rozptylu dokazuje teóriu, že počty trestných minút sú naozaj riedke pri hodnotách väčších ako 40. 40 minút však ešte možno považovať na “normálne” hodnoty, no 60+ sú už naozaj extrémne. Z riedkosti grafu pre oba atribúty pri y>= 60 vieme povedať, že hodnoty sú extrémne a môžeme ich znormalizovať napr. nahradením za horný kvantil (75%).

Zhrnutie: atribúty **pim.home** a **pim.away** sú spojité atribúty, ktoré pochádzajú z iného ako normálneho rozdelenia (podobné normálnemu no asymetrické). Obsahujú pomerne veľa vychýlených hodnôt najma pre hodnotu atribútov >= 60, preto bude vhodné ich normalizovať vo fáze čistenia dát. Pravdepodobne sa jednalo o vyhrotené zápasy, kedy bolo rozdaných naozaj veľa trestov. Takéto zápasy síce sú reálne, no taktiež sú veľmi zriedkavé a hrozí, že by mohli negatívne skresliť niektoré výsledky našej práce. JEdna hodnota v atribúte **pim.away** dokonca obsahuje jednu výrazne vychýlenú hodnotu (> 80).

Atribúty powerPlayOpportunities.home a powerPlayOpportunities.away

charakteristika: atribút hovorí o gólových príležitostiach počas presilových hier. Tento atribút sa odvíja od počtu presilovkových minút a teda očakávame medzi nimi koreláciu.

Hide

```
cat("Základná štatistika atribútu powerPlayOpportunities.home:\n ")
```

```
Základná štatistika atribútu powerPlayOpportunities.home:
```

Hide

```
summary(df$powerPlayOpportunities.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.000   2.000   3.000   3.168   4.000  10.000       4 
```

Hide

```
cat("\nZákladná štatistika atribútu powerPlayOpportunities.away:\n ")
```

```

Základná štatistika atribútu powerPlayOpportunities.away:
```

Hide

```
summary(df$powerPlayOpportunities.away)
```

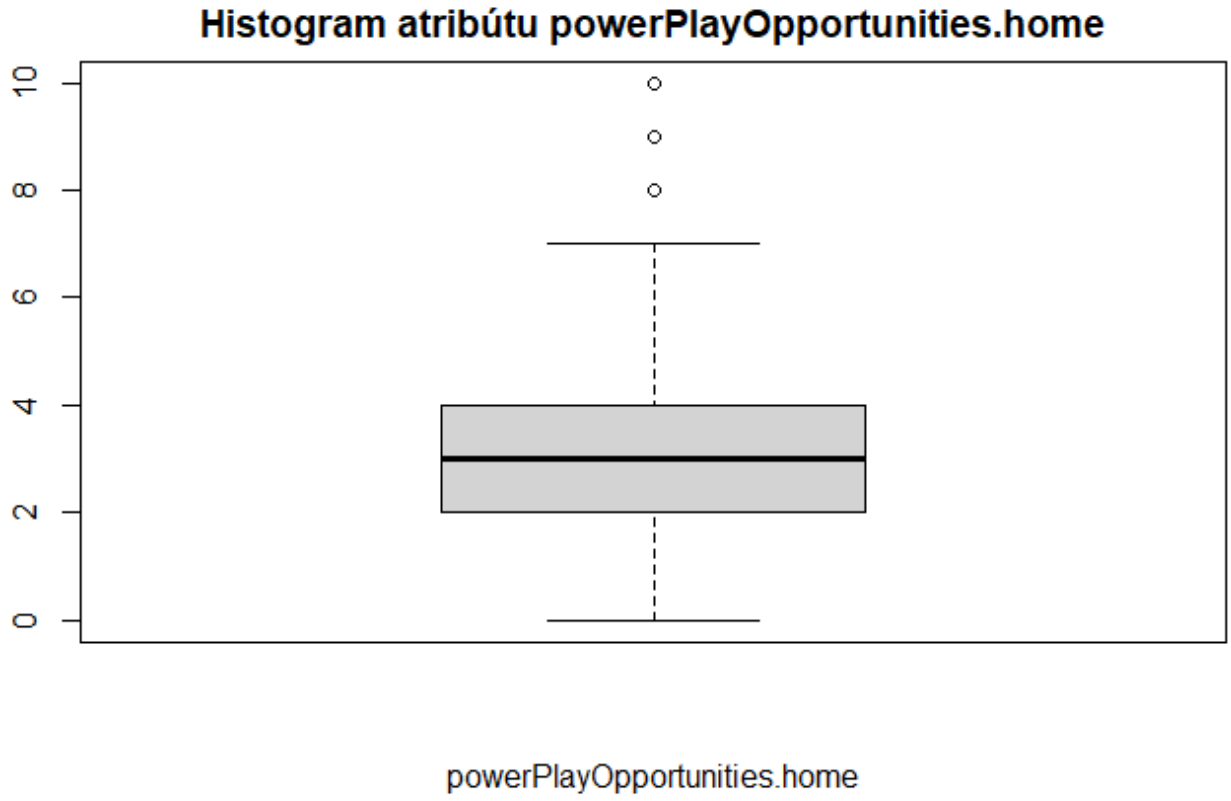
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.000   2.000   3.000   2.873   4.000   9.000       4 
```

Opätovne máme 4 NA hodnoty, tie však už boli opisované skôr. Väčšina dát bude zhluknutých okolo hodoty 3

pri oboch atribútoch, pričom vychýlené sú až okolo 10 až 9. Minimálna hodnota 0 je úplne valídna, pravdepodobne tím nehral v zápase žiadnu presilovku.

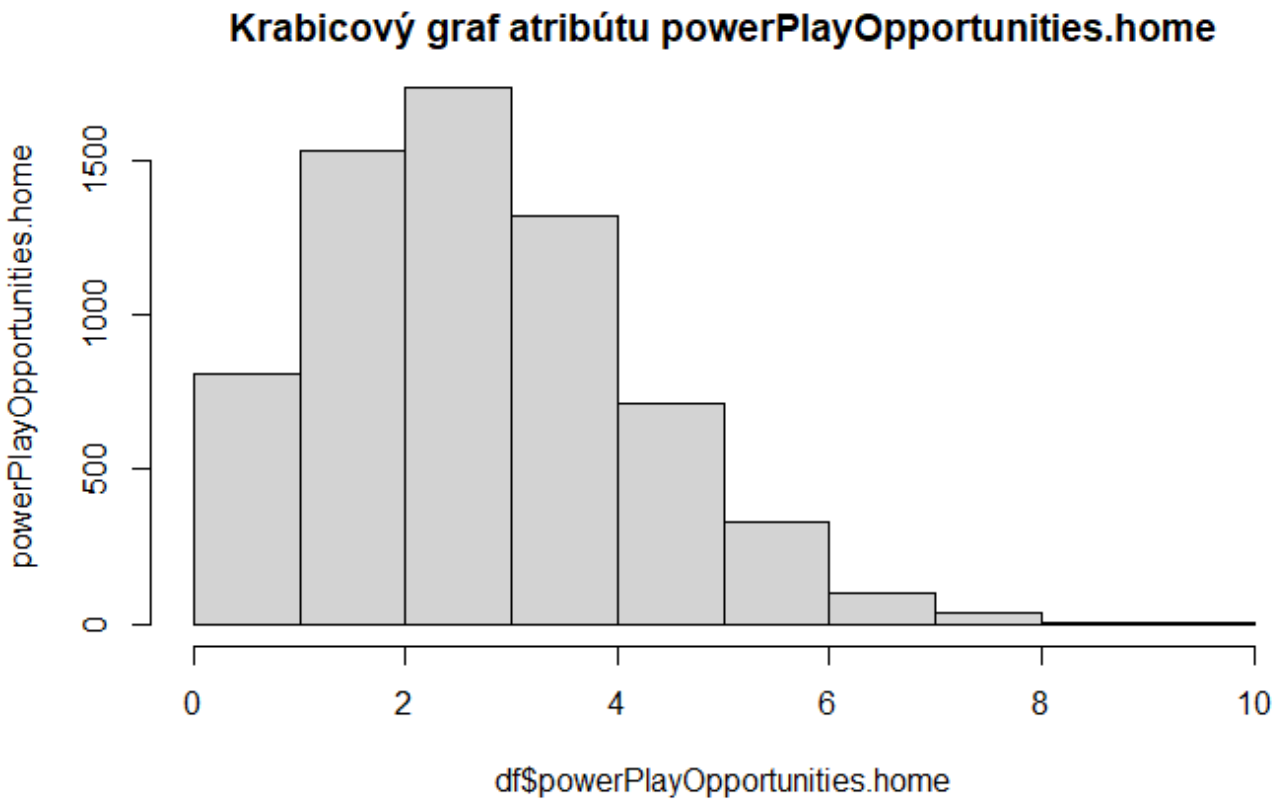
Hide

```
boxplot(df$powerPlayOpportunities.home, xlab = "powerPlayOpportunities.home", main="H
istogram atribútu powerPlayOpportunities.home")
```



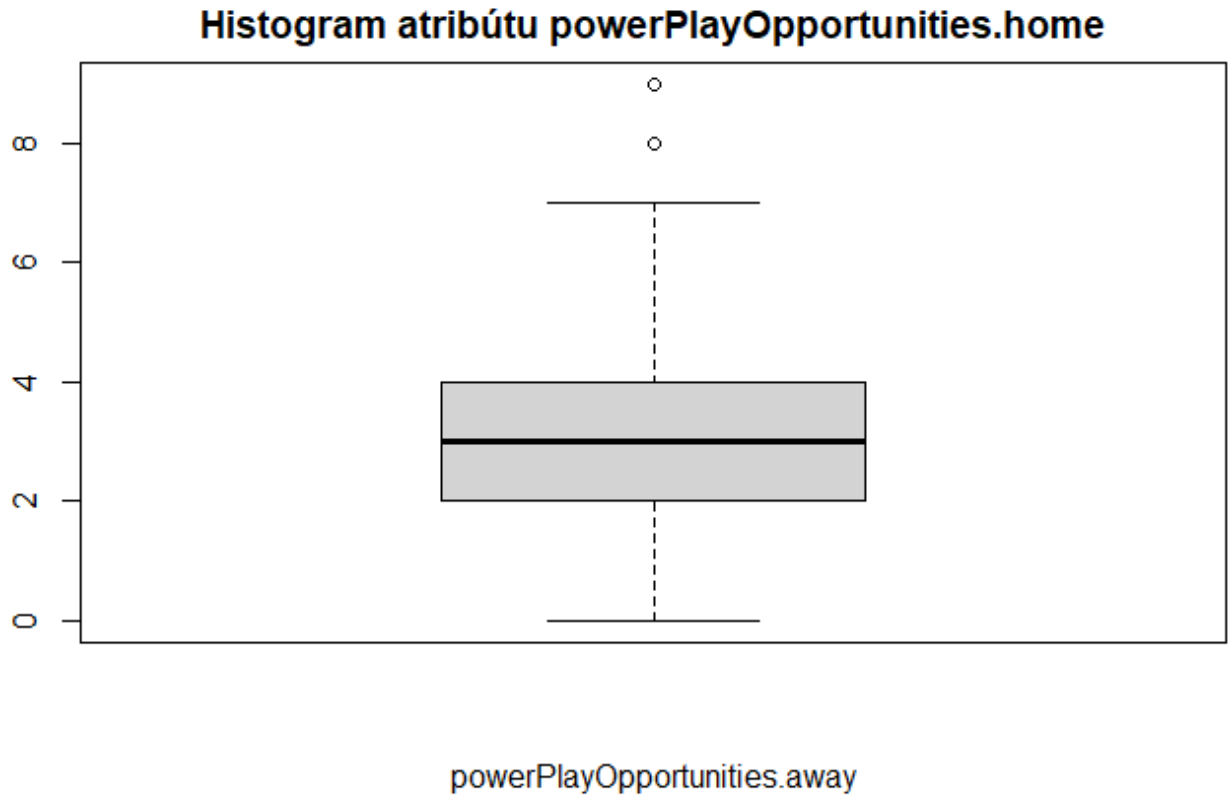
Hide

```
hist(df$powerPlayOpportunities.home, ylab = "powerPlayOpportunities.home", main="Krab
icový graf atribútu powerPlayOpportunities.home")
```



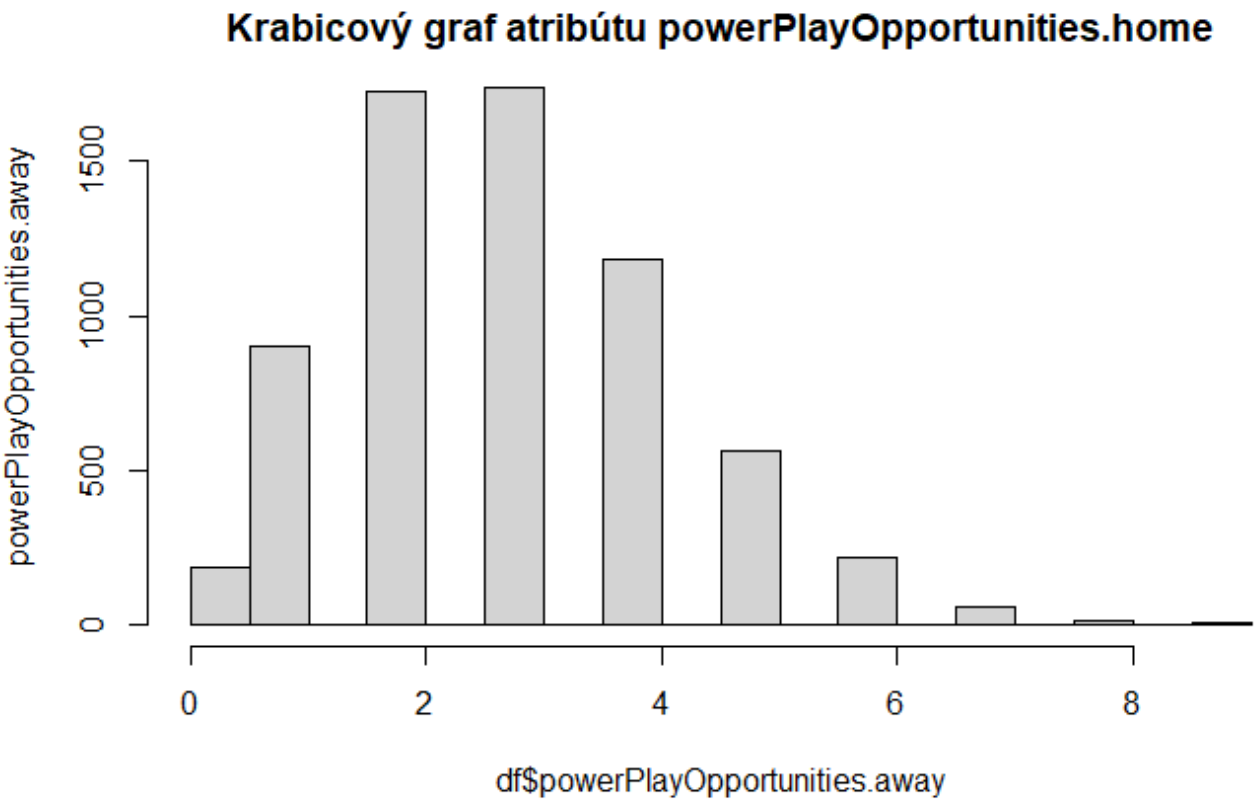
Hide

```
boxplot(df$powerPlayOpportunities.away, xlab = "powerPlayOpportunities.away", main="H
istogram atribútu powerPlayOpportunities.home")
```



Hide

```
hist(df$powerPlayOpportunities.away, ylab = "powerPlayOpportunities.away", main="Krab
icový graf atribútu powerPlayOpportunities.home")
```



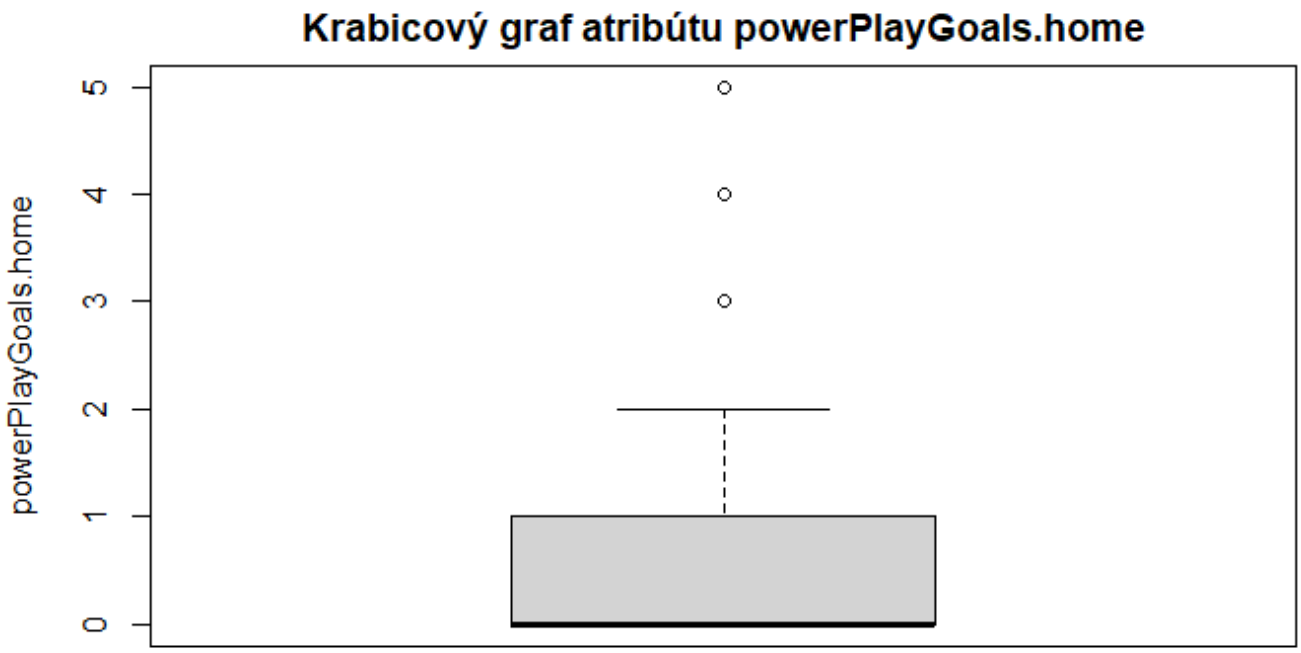
Z grafov môžeme vidieť, hodnoty atribútu pripomínajú normálne rozdelenie s malým počtom vychýlených hodnôt. Tieto hodnoty nebude potrebné riešiť, keďže sú v rámci normy a tatkiež je ich veľmi malé množstvo, čize by nemali reálne skreslovať výsledky. Z hľadiska štatistiky sa pre našu prácu taktieŽ jedná o menej zaujímavý atribút, keďže na ňom neplánujeme postaviť žiadnu z pracovných hypotéz.

Atribúty powerPlayGoals.home a powerPlayGoals.away

charakteristika: atribút reprezentuje počet presilovkových gólov pre oba tímy. Presilovkové góly sú v NHL časté, ale veľké počty presilovkových góolov sú nezvyčajné. Väčšinou je úspešnosť presiloviek okolo 20%, čo znamená jeden gól z piatich presilových hier.

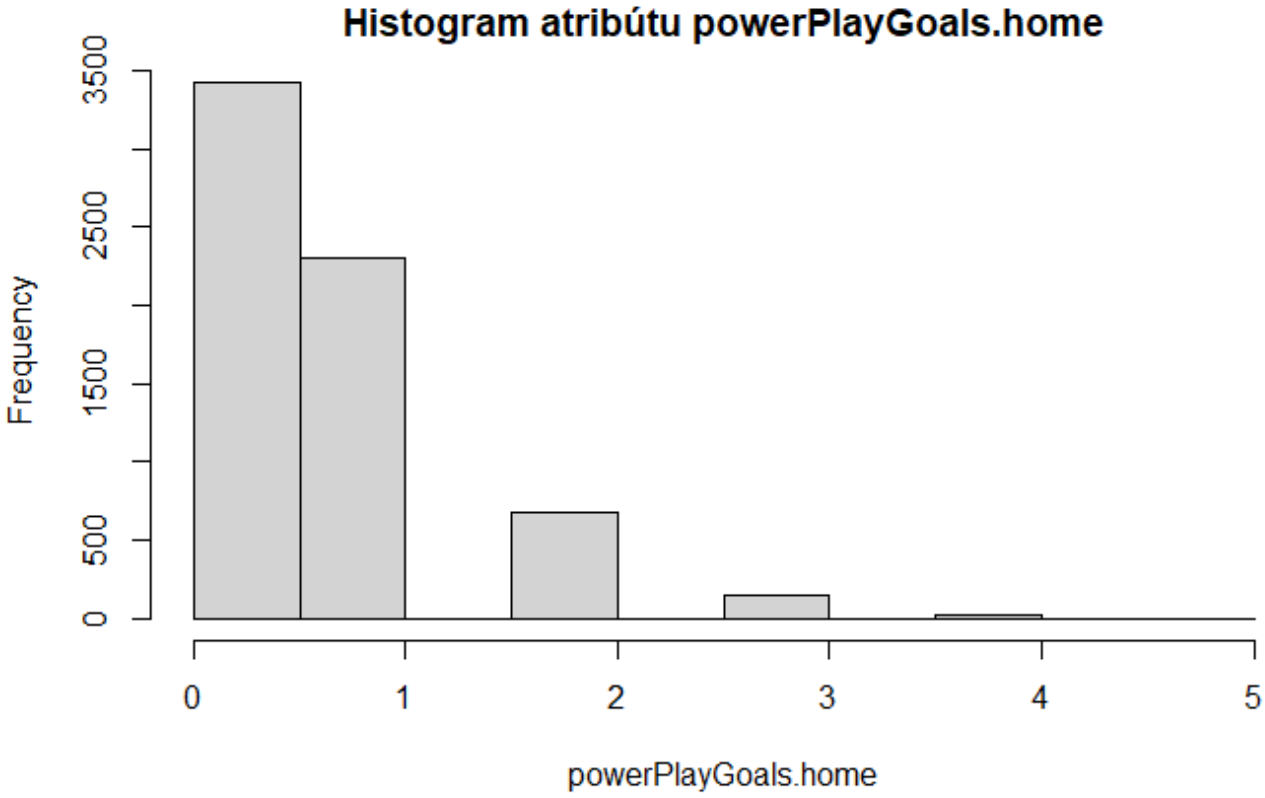
Hide

```
boxplot(df$powerPlayGoals.home, ylab = "powerPlayGoals.home", main="Krabicový graf at
ribútu powerPlayGoals.home")
```



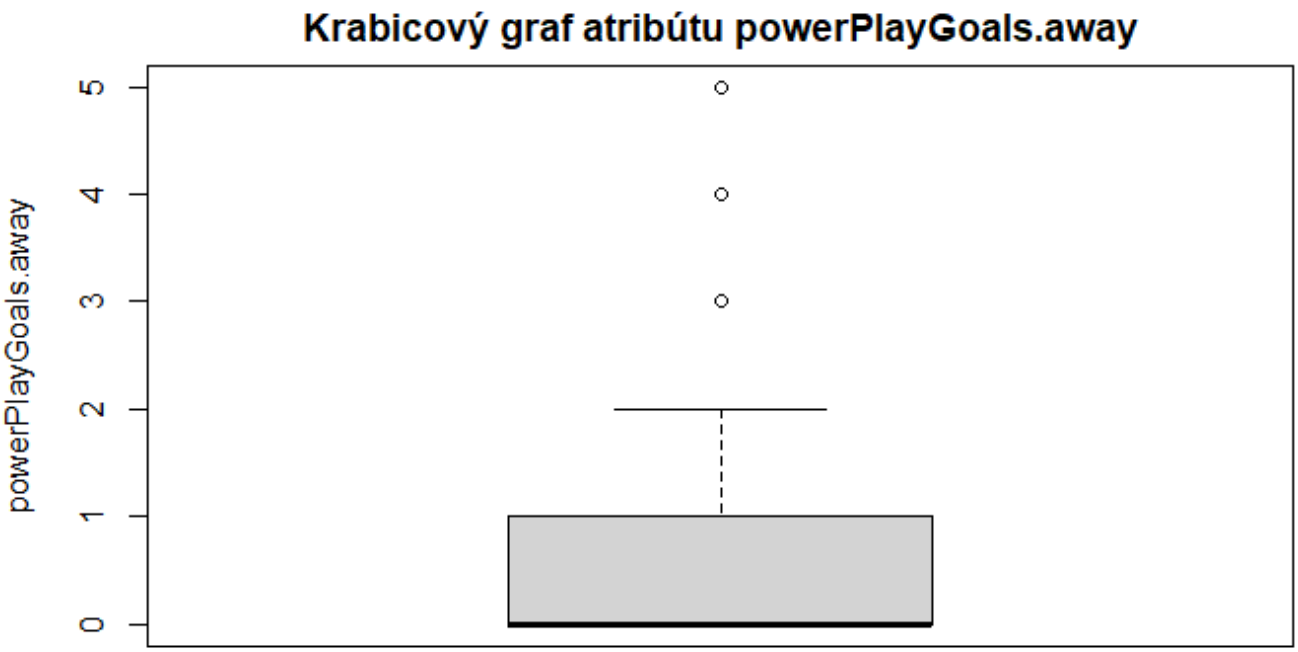
Hide

```
hist(df$powerPlayGoals.home, xlab = "powerPlayGoals.home", main="Histogram atribútu p
owerPlayGoals.home")
```



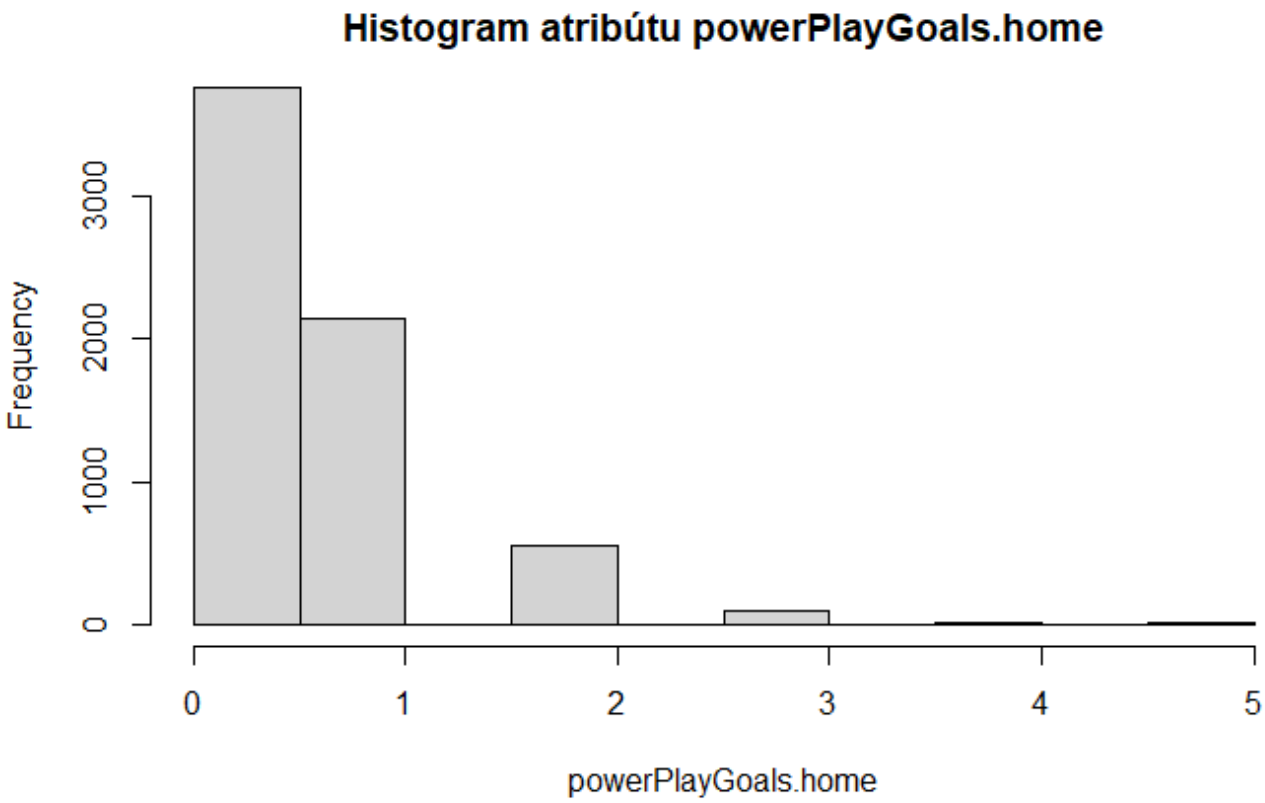
Hide

```
boxplot(df$powerPlayGoals.away, ylab = "powerPlayGoals.away", main="Krabicový graf at
ribútu powerPlayGoals.away")
```



Hide

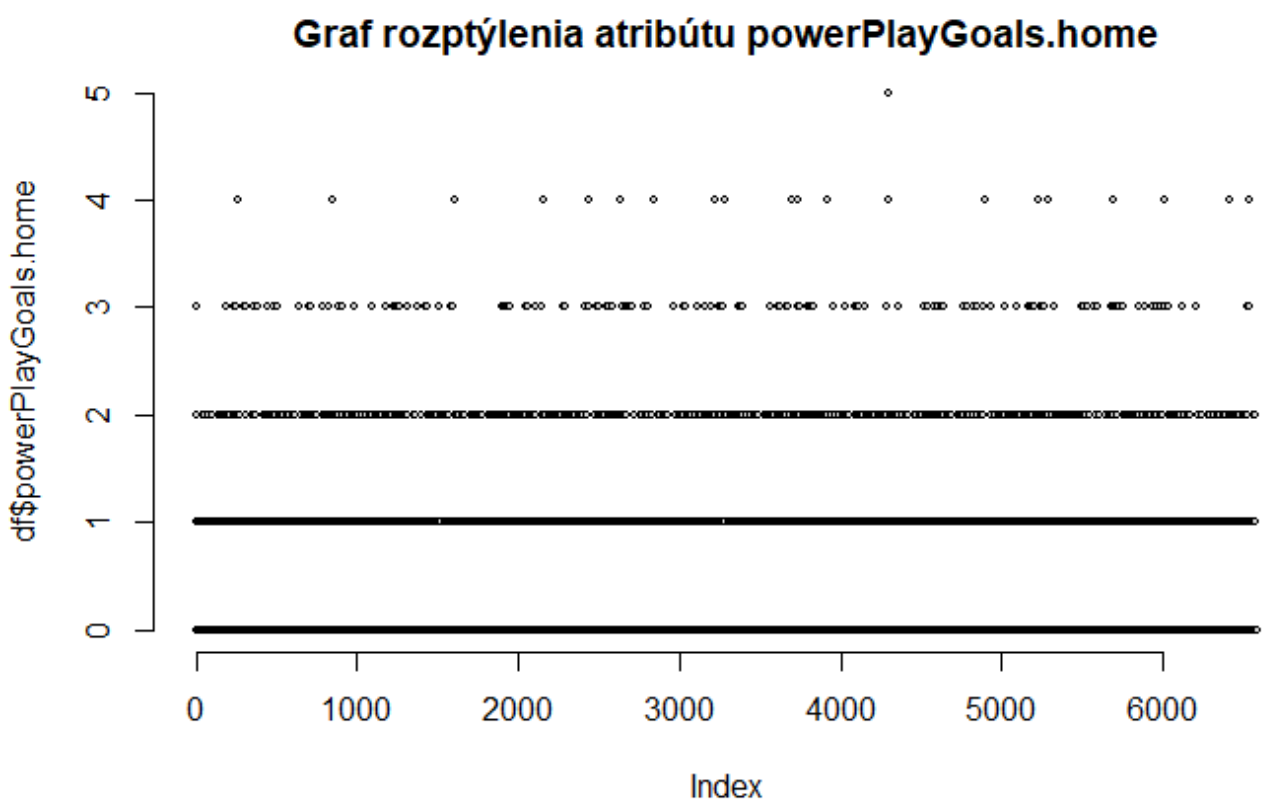
```
hist(df$powerPlayGoals.away, xlab = "powerPlayGoals.home", main="Histogram atribútu p
owerPlayGoals.home")
```



Grafy nám odhalili naklonené rozdelenia vpravo s malým počtom vychýlených hodnôt. Krabicové grafy sú pre oba atribúty totožné, čo je zaujímavé. Vyskytuje sa malé množstvo vychýlených hodnôt, konkrétne pri počte gólov v presilovke = 3, 4, 5. Vykreslíme si dodatočne graf rozptýlenia aby sme lepšie zhodnotili, či sú hodnoty naozaj extremálne.

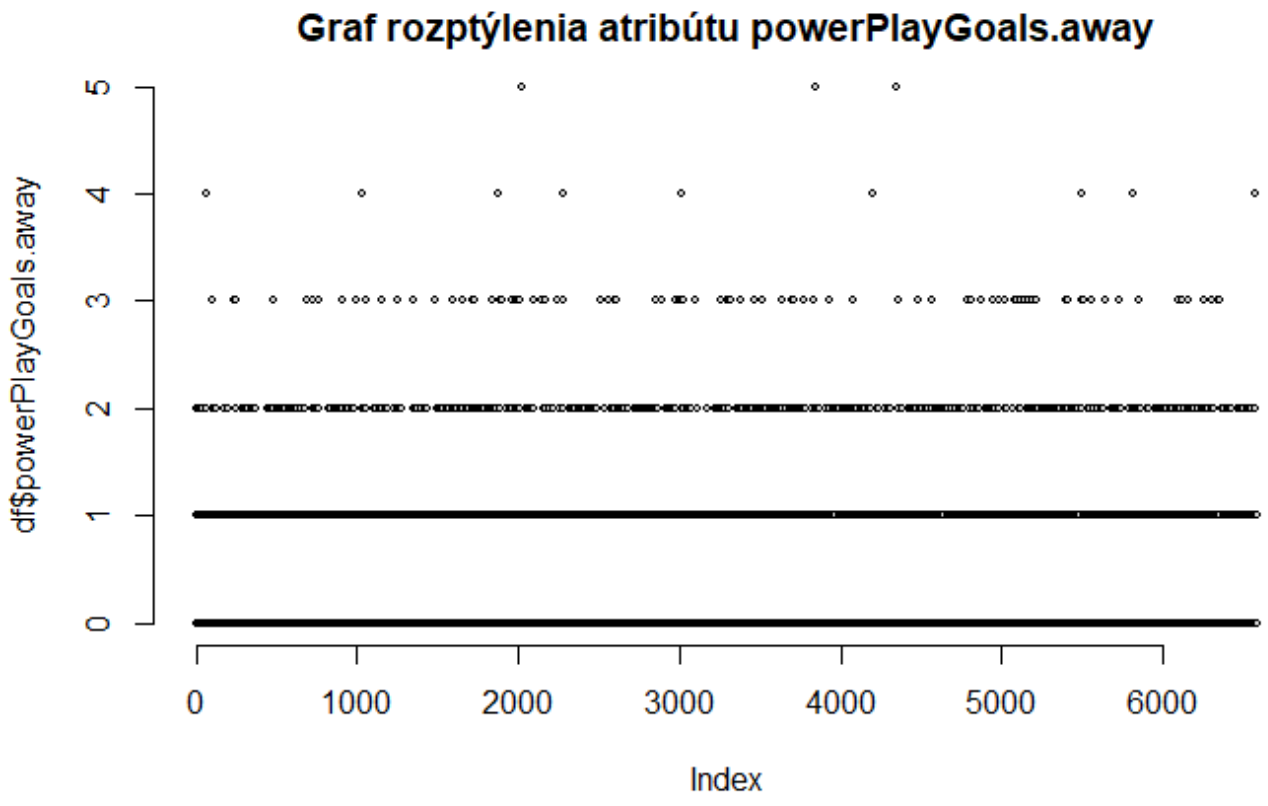
Hide

```
plot(df$powerPlayGoals.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, fr
ame = FALSE, main="Graf rozptýlenia atribútu powerPlayGoals.home")
```



Hide

```
plot(df$powerPlayGoals.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, fr
ame = FALSE, main="Graf rozptýlenia atribútu powerPlayGoals.away")
```



Hide

```
subset(df, powerPlayGoals.home >= 5)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20192020	R	3	9	3	FALSE	23	22	

1 row | 1-9 of 31 columns

Hide

```
subset(df, powerPlayGoals.away >= 5)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	R	5	4	21	TRUE	34	27	
20152016	P	6	1	15	TRUE	27	26	
20192020	R	6	3	23	TRUE	23	5	

3 rows | 1-9 of 31 columns

Nad jej normalizáciou ale uvažovať neplánujeme, keďže sa jedná o reálnu hodnotu ktorá nesie informáciu o efektívnom NHL tíme v presilovkách. Vo všetkých prípadoch sa však jedná o odliŠné tímy - Colorado, Washington, Vancouver a Tamba Bay - nebude sa pravdepodobne jednať o tím efektívny v presilovkách. Samozrejme, dodatočná analýza týchto hodnôt by bola vhodná, no vzhľadom na to, že tento atribút nepovažujeme za kľúčový z hľadiska hypotéz, sa budeme sústrediť radšej na analýzu iných atribútov.

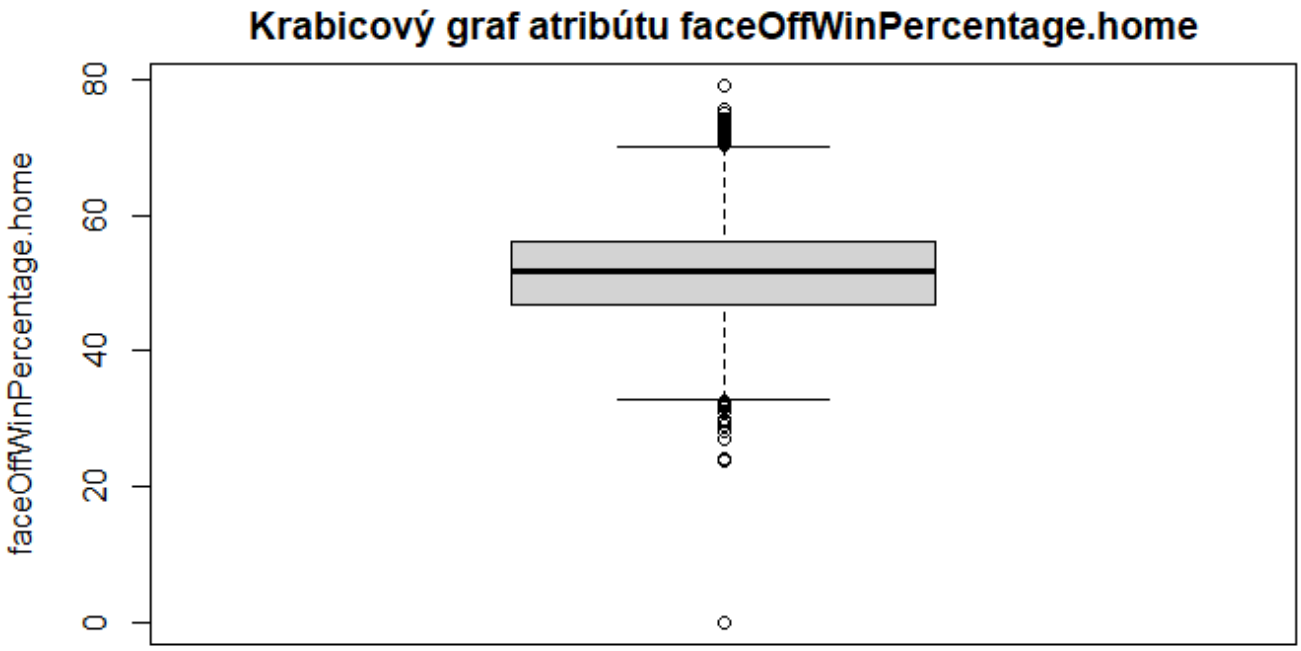
Vieme povedať, že hodnota 5 gólov sa vyskytuje iba v pár zápasoch, preto ju môžeme považovať za vychýlenú od normy. Keďže však tento atribút neplánujeme špecificky používať pri riešení projektu, normalizáciu neplánujeme vykonávať. V prípade zmeny názoru túto skutočnosť v budúcnosti prehodnotíme.

Atribúty faceOffWinPercentage.home a faceOffWinPercentage.away

charakteristika: atribút hovorí o percentuálnom počte vyhratých vhadzovaní. Spolu tvorí atribút pre domácich a hostí 100% vhadovaní.

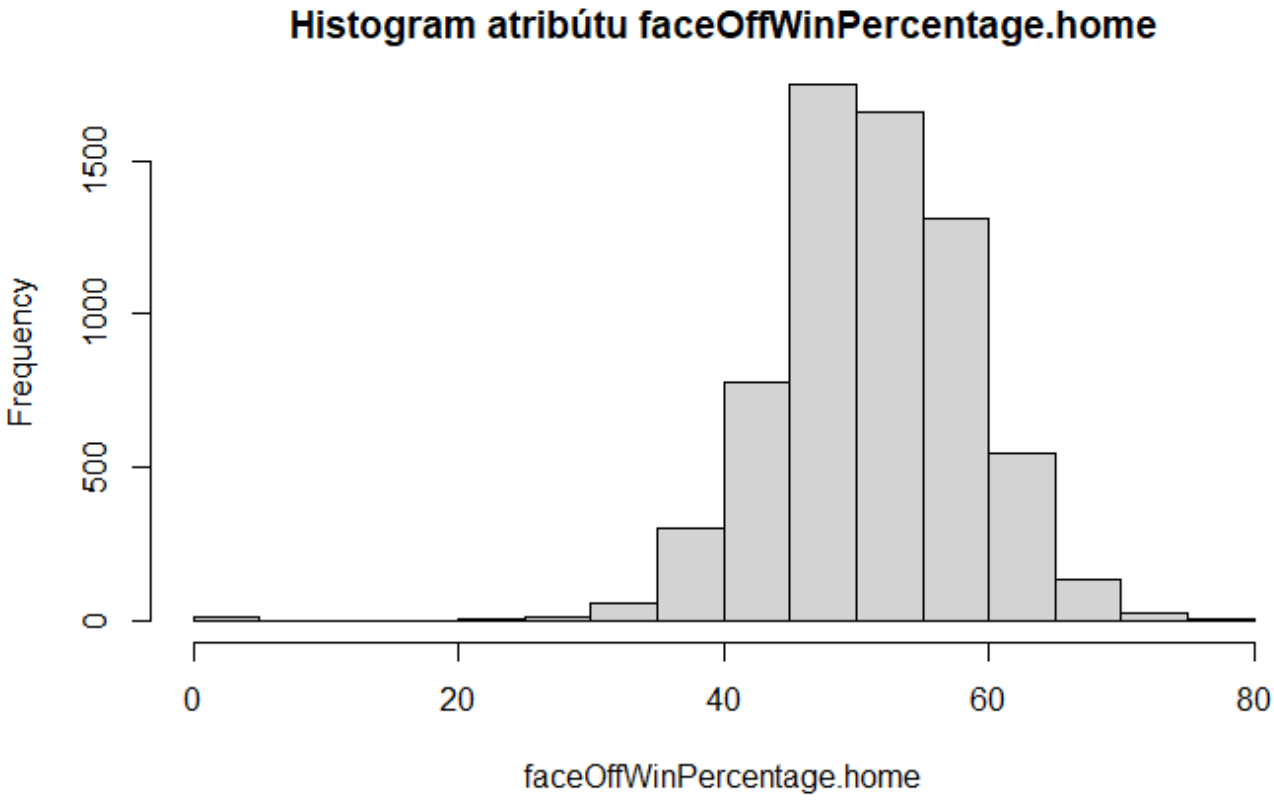
Hide

```
boxplot(df$faceOffWinPercentage.home, ylab= "faceOffWinPercentage.home", main="Krabic
ový graf atribútu faceOffWinPercentage.home")
```



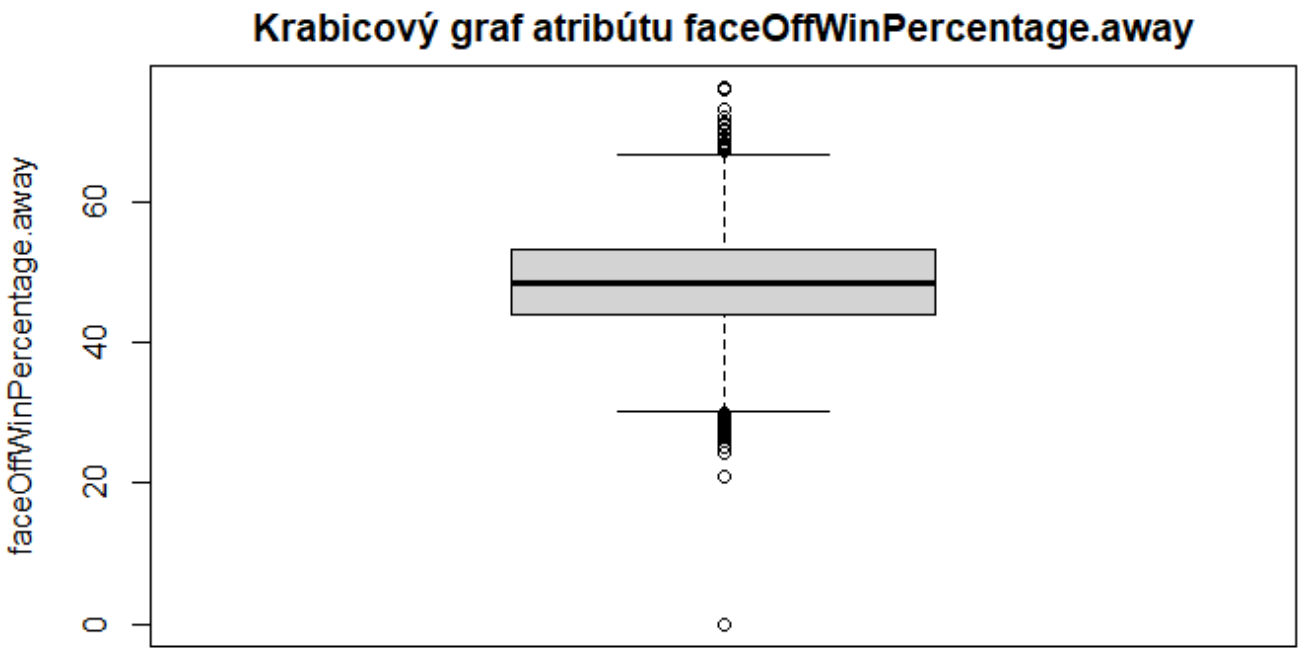
Hide

```
hist(df$faceOffWinPercentage.home, xlab="faceOffWinPercentage.home", main="Histogram
atribútu faceOffWinPercentage.home")
```



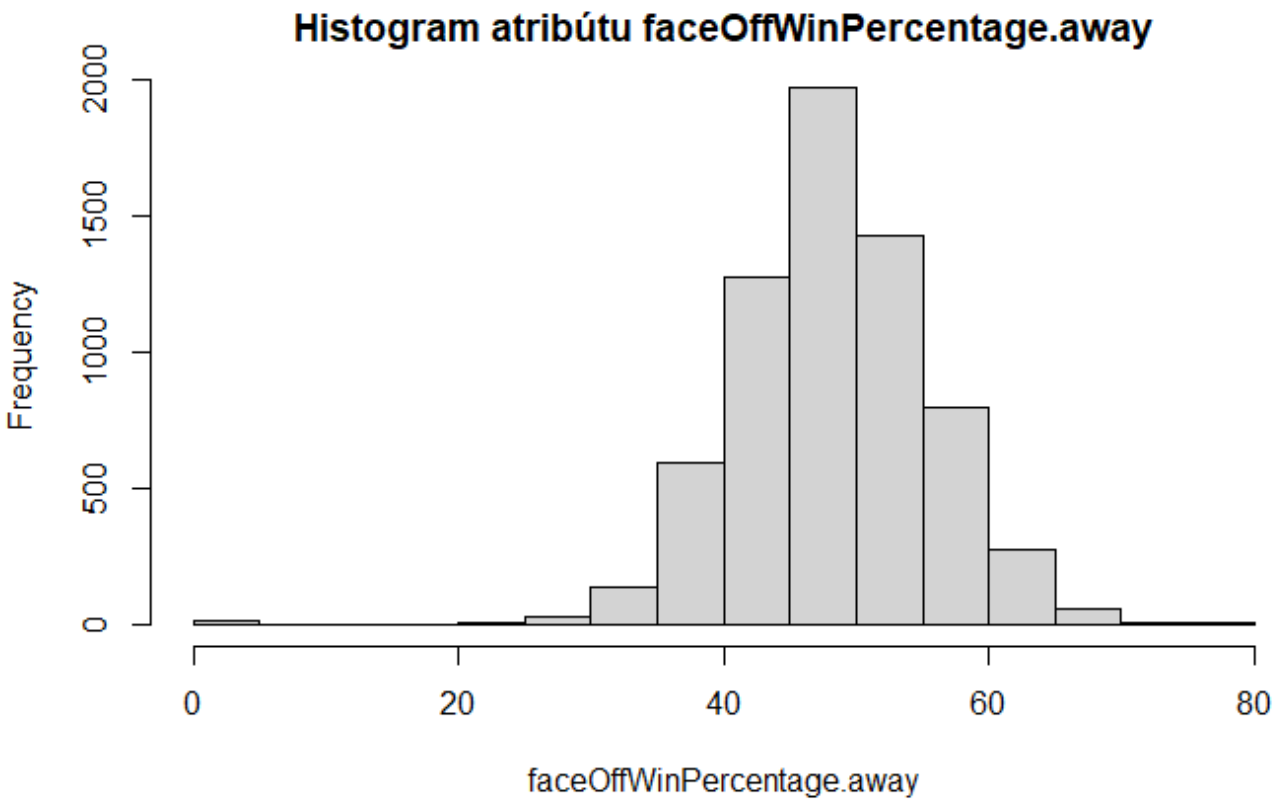
Hide

```
boxplot(df$faceOffWinPercentage.away, ylab="faceOffWinPercentage.away", main="Krabicový graf atribútu faceOffWinPercentage.away")
```



Hide

```
hist(df$faceOffWinPercentage.away, xlab= "faceOffWinPercentage.away", main="Histogram atribútu faceOffWinPercentage.away")
```



Atribút sa veľmi podobá normálnemu rozdeleniu, pričom obsahuje zopár vychýlených hôdnot, ako napr. 0, ktorá logicky nedáva zmysel a budeme sa na ňu musieť pozrieť.

Hide

```
subset(df, faceOffWinPercentage.home == 0)
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <
20172018	P	0	0	28	FALSE	0	0	
20172018	P	0	0	26	FALSE	0	0	
20172018	P	0	0	54	FALSE	0	0	
20172018	P	0	0	15	FALSE	0	0	
20172018	P	0	0	5	FALSE	0	0	
20162017	P	0	0	20	FALSE	0	0	
20162017	P	0	0	10	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	

1-10 of 10 rows | 1-9 of 31 columns

Hide

```
subset(df, faceOffWinPercentage.away == 0)
```

season <int>	t... <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a... <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	0	0	
20172018	P	0	0	26	FALSE	0	0	
20172018	P	0	0	54	FALSE	0	0	
20172018	P	0	0	15	FALSE	0	0	
20172018	P	0	0	5	FALSE	0	0	
20162017	P	0	0	20	FALSE	0	0	
20162017	P	0	0	10	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	

1-10 of 10 rows | 1-9 of 31 columns

V oboch prípadoch (pri oboch atribútoch) sa jedná o chýbajúce hodnoty záznamov s typom tbc, t.j. preložených alebo zrušených zápasov. Tieto zápasy budú v čistení odstránené, t.j. bude vyriešený problém s touto vychýlenou hodnotou. Ešte môžeme overiť, či súčet atribútov faceOff pre domáci a hosťujúci tím dávajú naozaj hodnotu 100%.

Hide

```
unique(df$faceOffWinPercentage.away + df$faceOffWinPercentage.home == 100)
```

```
[1]  TRUE FALSE    NA
```

Vidíme, ze súčet pravdepodobností nie je vždy rovný 100. Tatkiež dosahuje hodnotu NA v niektorých záznamoch, ale to je očakávané keďže dáta ešte neboli vyčistené. Môžeme sa ale bližšie pozrieť na záznamy, ktoré nedosahujú súčet 100.

Hide

```
subset(df, df$faceOffWinPercentage.away + df$faceOffWinPercentage.home != 100)
```

season <int>	t... <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a... <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	0	0	
20172018	P	0	0	26	FALSE	0	0	

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	54	FALSE	0	0	
20172018	P	0	0	15	FALSE	0	0	
20172018	P	0	0	5	FALSE	0	0	
20162017	P	0	0	20	FALSE	0	0	
20162017	P	0	0	10	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	
20162017	P	0	0	18	FALSE	0	0	

1-10 of 11 rows | 1-9 of 31 columns Previous **1** 2 Next

Jedná sa o našich klasických 10 nadbytočných záznamov. Zaujímavé však je, že sa tu naskytuje aj 11-ty záznam z reálneho zápasu medzi San Jose Sharks a Montreal Canadiens.

Hide

```
38.8 + 61.3
```

```
[1] 100.1
```

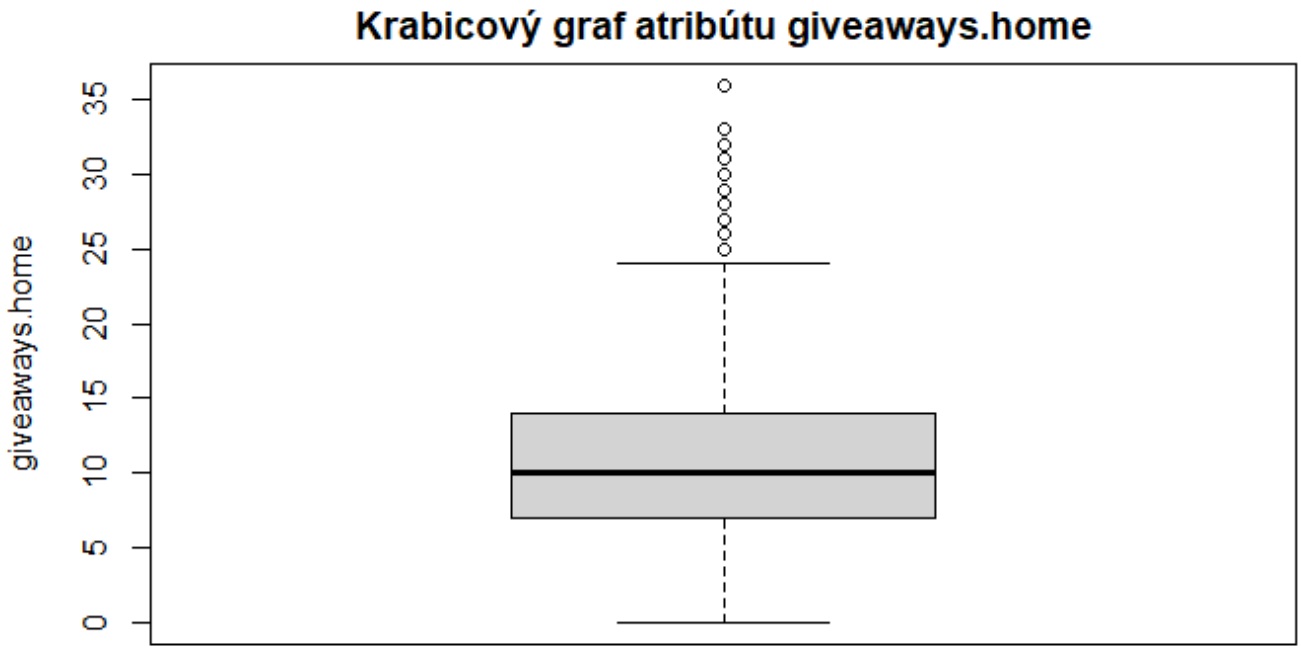
Tento zápas presahuje maximálnu úspešnosť vhadzovaní o 0.1%. Toto sa pravdepodobne stalo kvôli zaokrúhľovaniu úspešností vhadzovaní jednotlivých tímov. Nejedná sa však o nejakú kritickú chybu, reálne to na dáta a výstup projektu nebude mať žiaden dopad a preto tento problém riešiť nebude potrebné (dalo by sa prípadne odpočítať od každej pravdepodobnosti -0.05 tak, aby bola výsledná hodnota 100).

Atribúty giveaways.home a giveaways.away

charakteristika: atribút reprezentujúci odovzdané puky súperiacemu tímu bez súboja. Takýchto situácií sa v zápasoch vyskytuje viacero a nie sú nezvyčajným javom zápasu NHL.

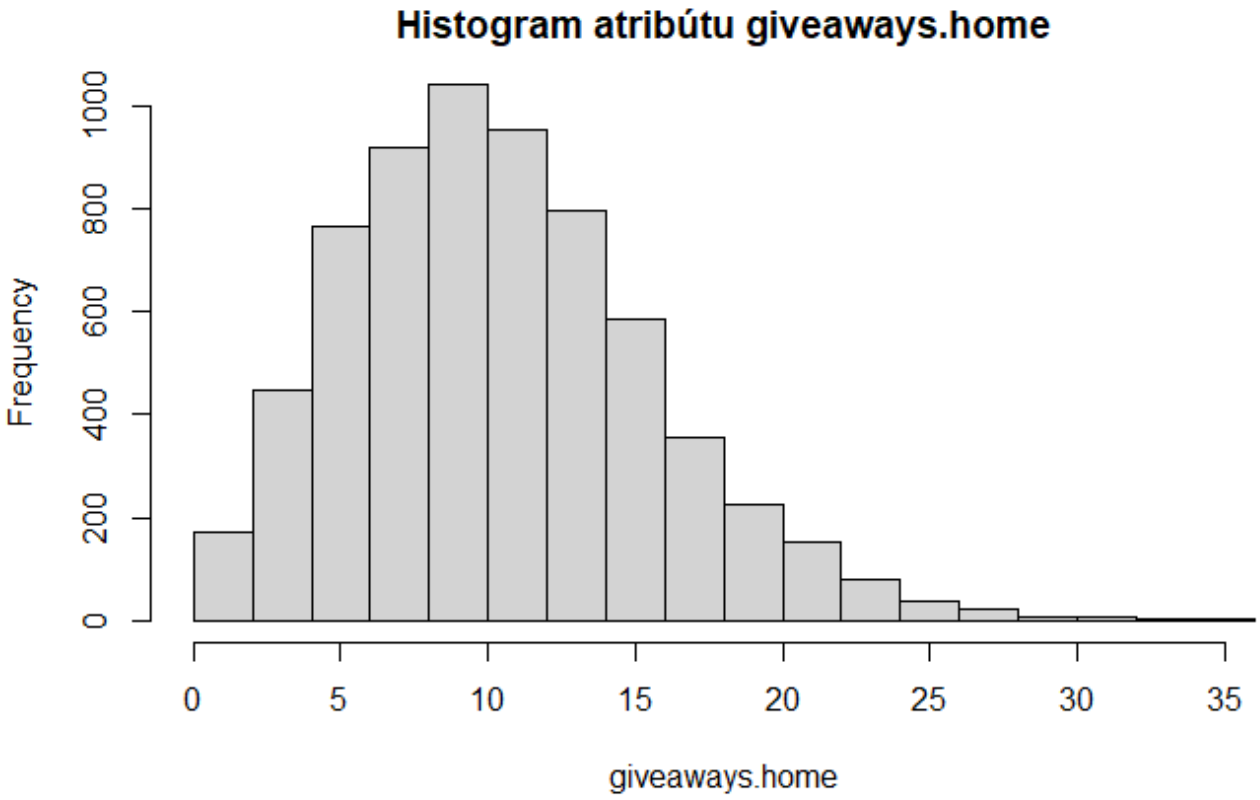
Hide

```
boxplot(df$giveaways.home, ylab = "giveaways.home", main="Krabicový graf atribútu giv
eaways.home")
```



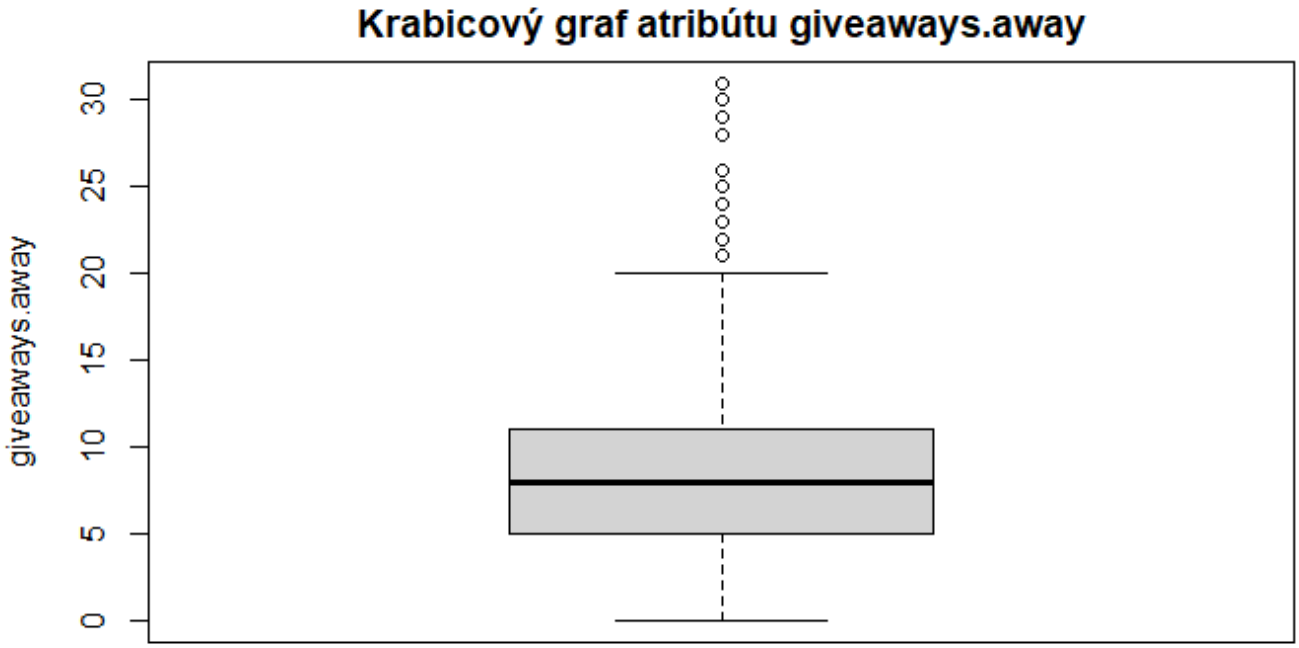
Hide

```
hist(df$giveaways.home, xlab = "giveaways.home", main="Histogram atribútu giveaways.h
ome")
```



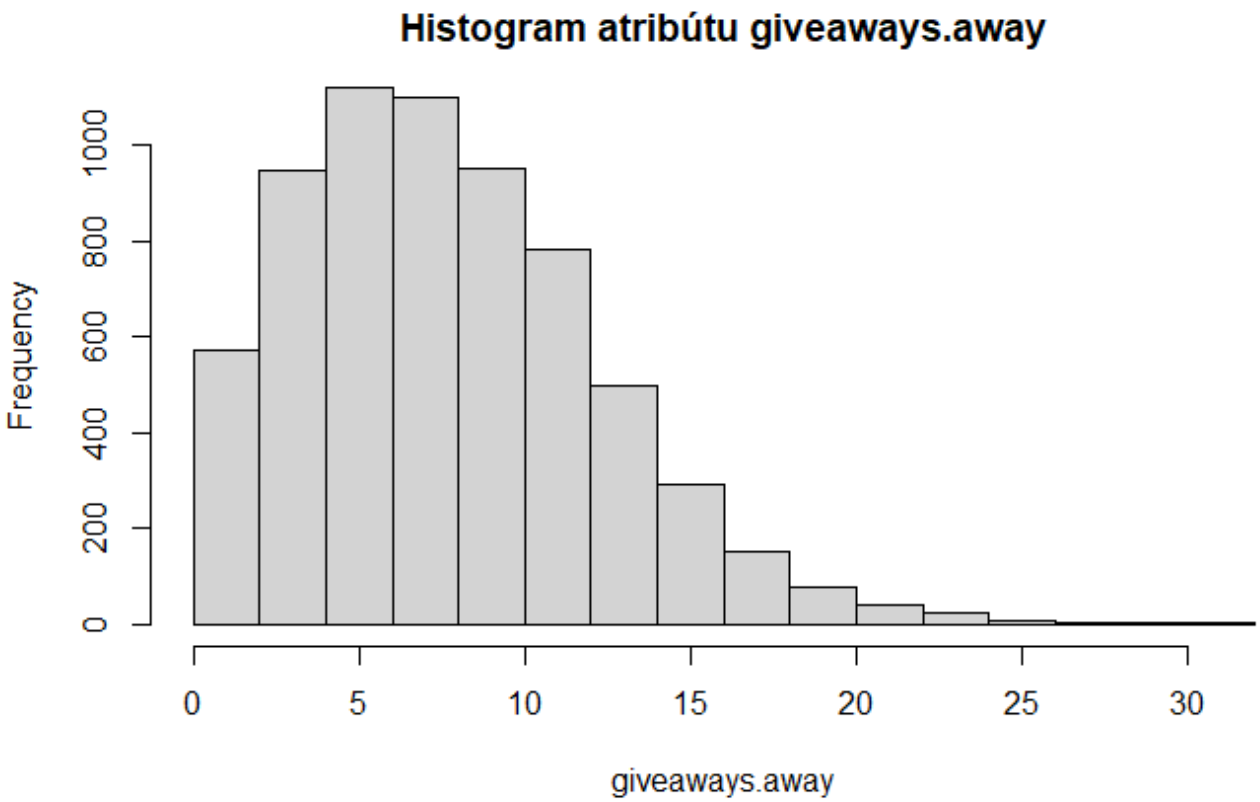
Hide

```
boxplot(df$giveaways.away, ylab = "giveaways.away", main="Krabicový graf atribútu giv
eaways.away")
```



Hide

```
hist(df$giveaways.away, xlab = "giveaways.away", main="Histogram atribútu giveaways.a
way")
```



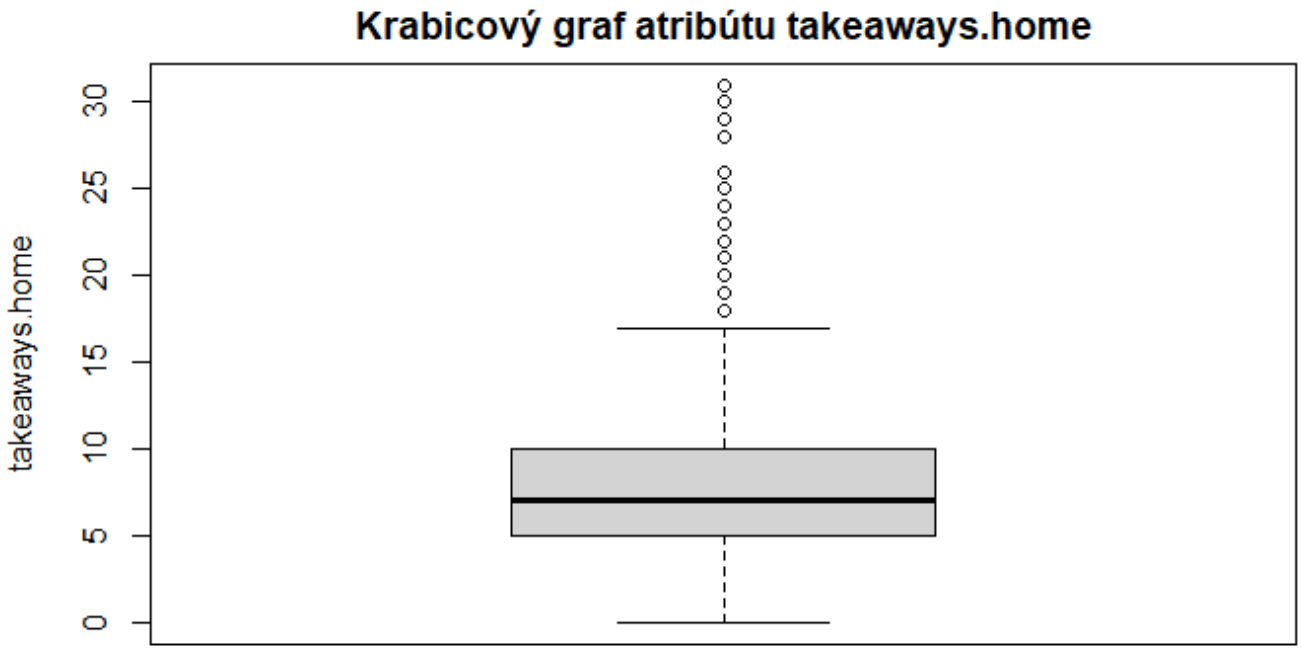
Hodnoty jemne pripomínajú normálne rozdelenie, no reálne sa bude jednať o distribúciu naklonenú vpravo. Taktiež je tam veľa vychýlených hodnôt, ktoré však nemusia znamenať chyby merania. Atribút nebudeme analyzovať podrobnejšie z dôvodu, že pre našu prácu ho nepovažujeme za kľúčový.

Atribúty takeaways.home a takeaways.away

charakteristika: atribút, ktorý reprezentuje počet odobraných pukov súperiacemu tímu. Tento atribút nie je komplementom vyššie spomenutého atribútu odovzdania pukov.

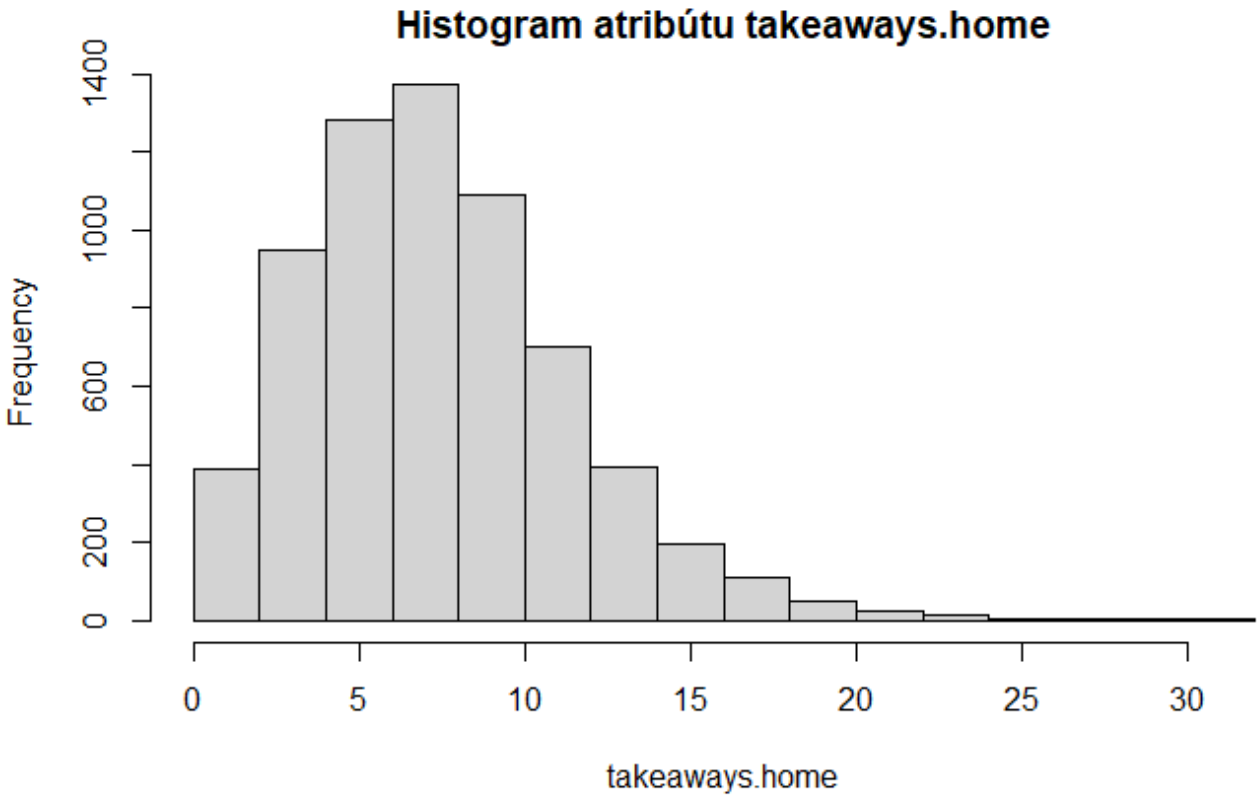
Hide

```
boxplot(df$takeaways.home, ylab = "takeaways.home", main="Krabicový graf atribútu tak
eaways.home")
```



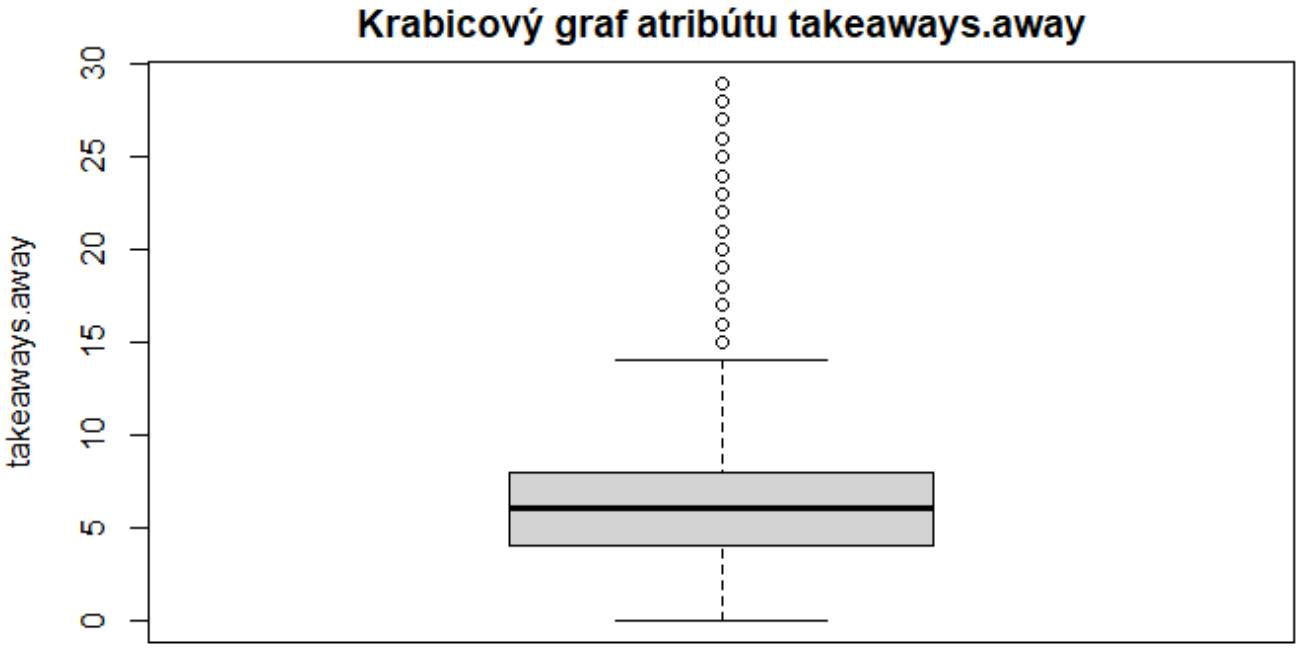
Hide

```
hist(df$takeaways.home, xlab = "takeaways.home", main="Histogram atribútu takeaways.h
ome")
```



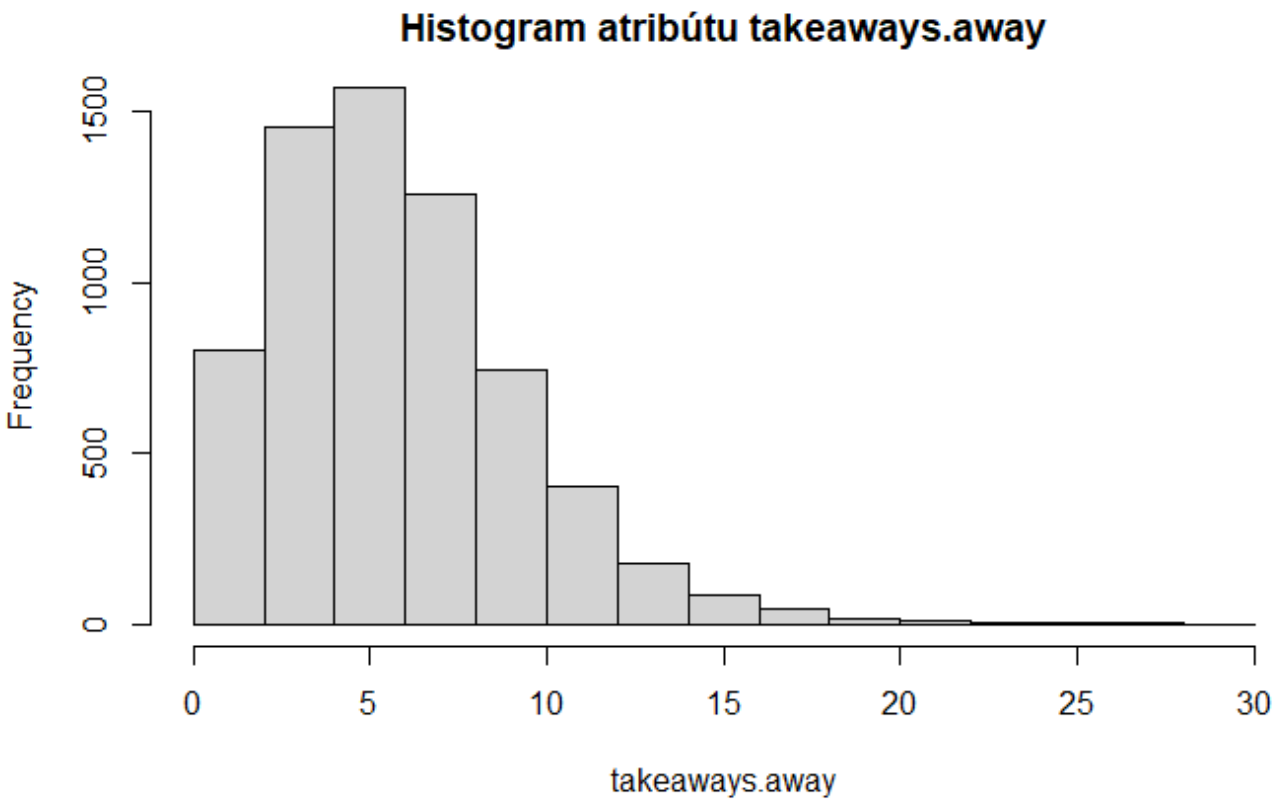
Hide

```
boxplot(df$takeaways.away, ylab = "takeaways.away", main="Krabicový graf atribútu tak
eaways.away")
```



Hide

```
hist(df$takeaways.away, xlab = "takeaways.away", main="Histogram atribútu takeaways.a
way")
```



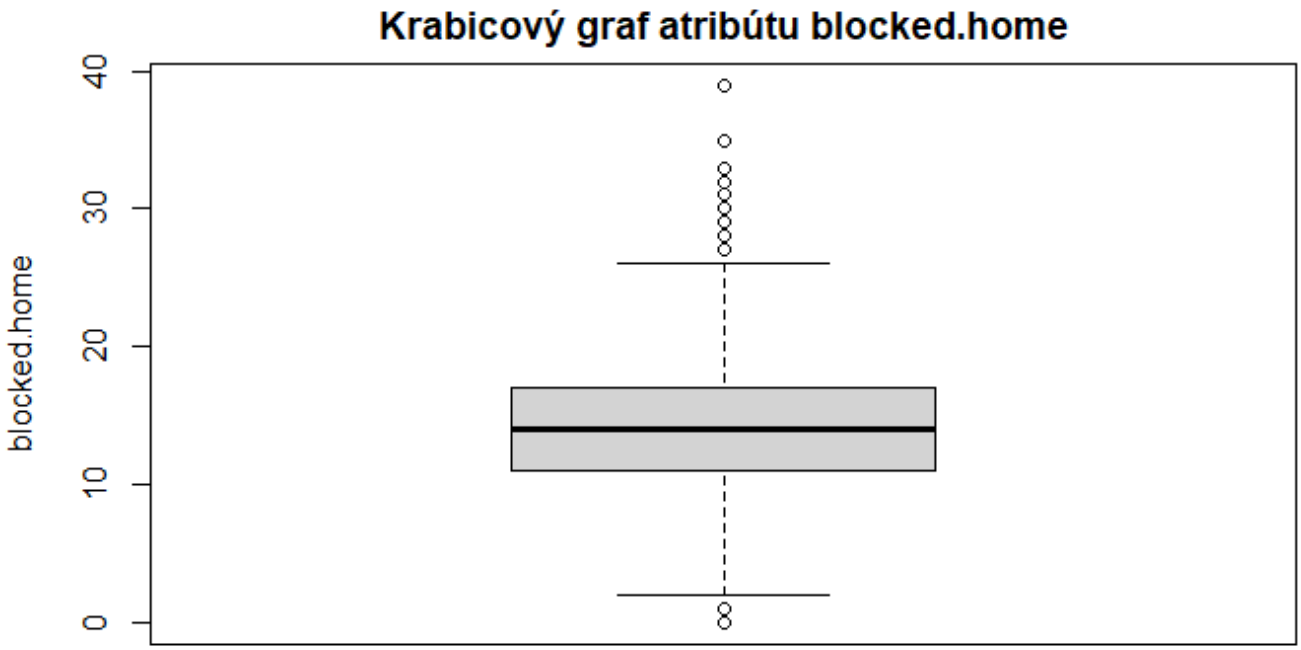
Na histograme môžeme vidieť, že sa jedná o distribúciu naklonenú vpravo. Taktiež je tam dosť vychýlených hodnôt, ktoré však opäť nemusia (a pravdepodobne ani nebudú) znamenať chyby merania. Atribút nebudeme analyzovať podrobnejšie z dôvodu, že pre našu prácu ho nepovažujeme za kľúčový (má maximálne len situačné využitie).

Atribúty blocked.home a blocked.away

charakteristika: atribút súčtu zablokovaných striel na bránu. hráčmi, inými ako brankár. To znamená, že hráč zablokuje strelu na bránu hokejkou, korčuľou, alebo telom a puk sa vôbec nedostane k brankárovi. Zabkovania strely nie sú nezvyčajné v NHL, ale nevyskytujú sa vo veľkých počtoch, ako napr. strely na bránku.

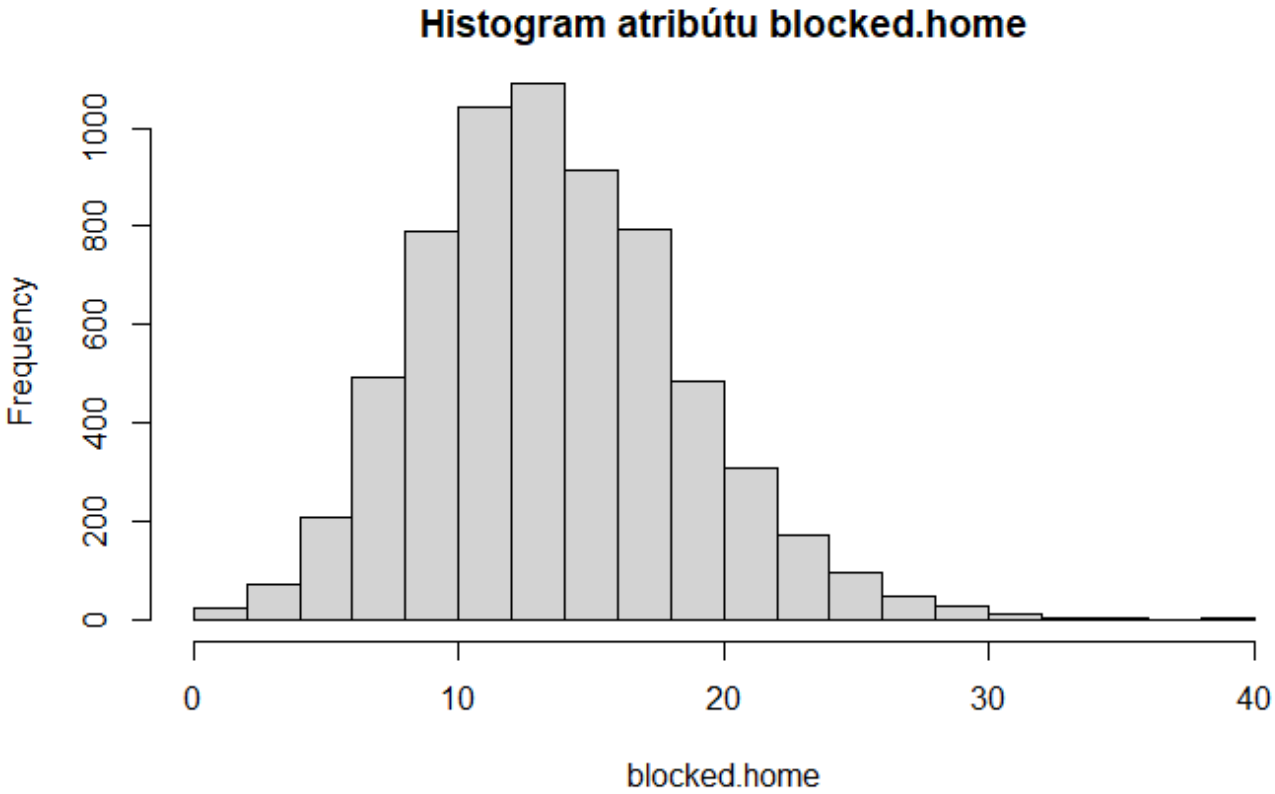
Hide

```
boxplot(df$blocked.home, ylab="blocked.home", main="Krabicový graf atribútu blocked.h
ome")
```



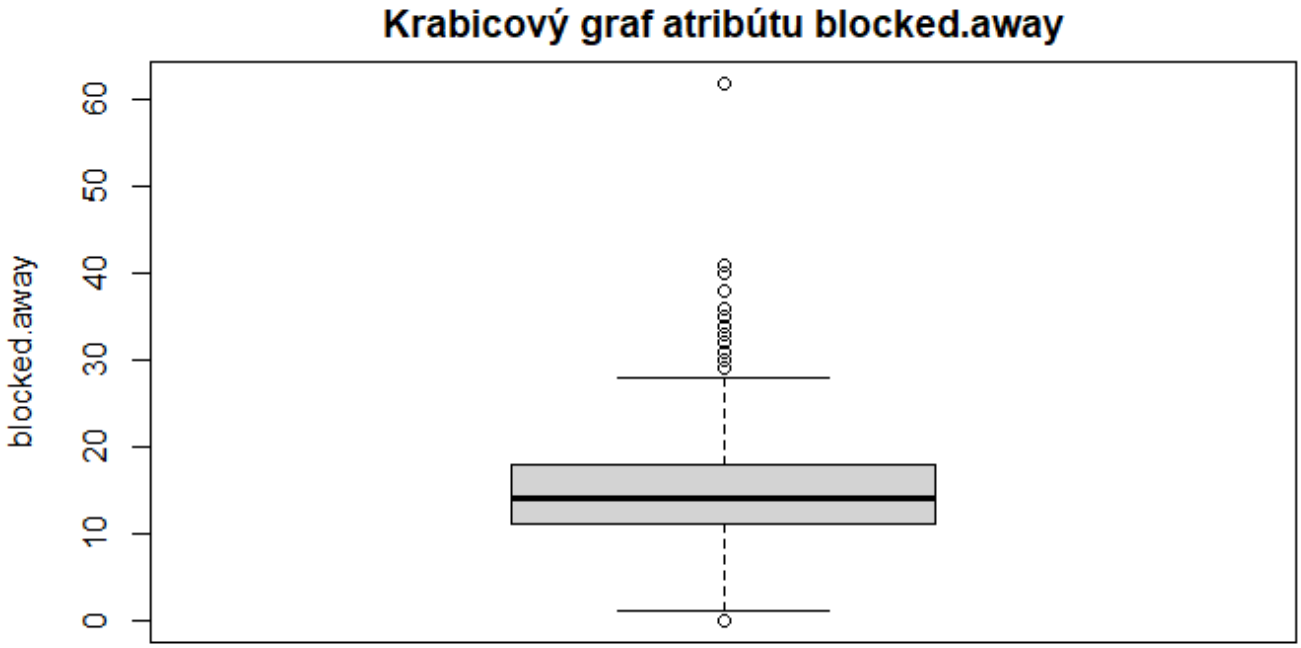
Hide

```
hist(df$blocked.home, xlab="blocked.home", main="Histogram atribútu blocked.home")
```



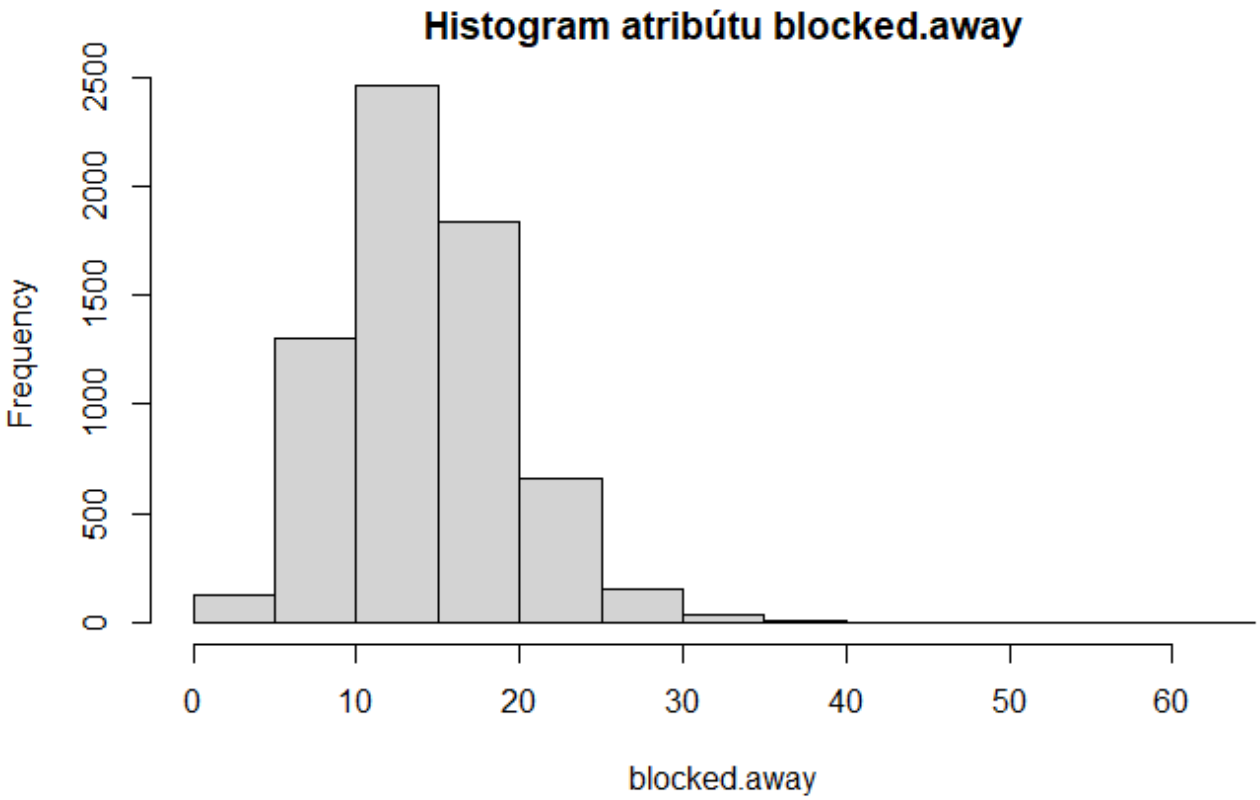
Hide

```
boxplot(df$blocked.away, ylab="blocked.away", main="Krabicový graf atribútu blocked.away")
```



Hide

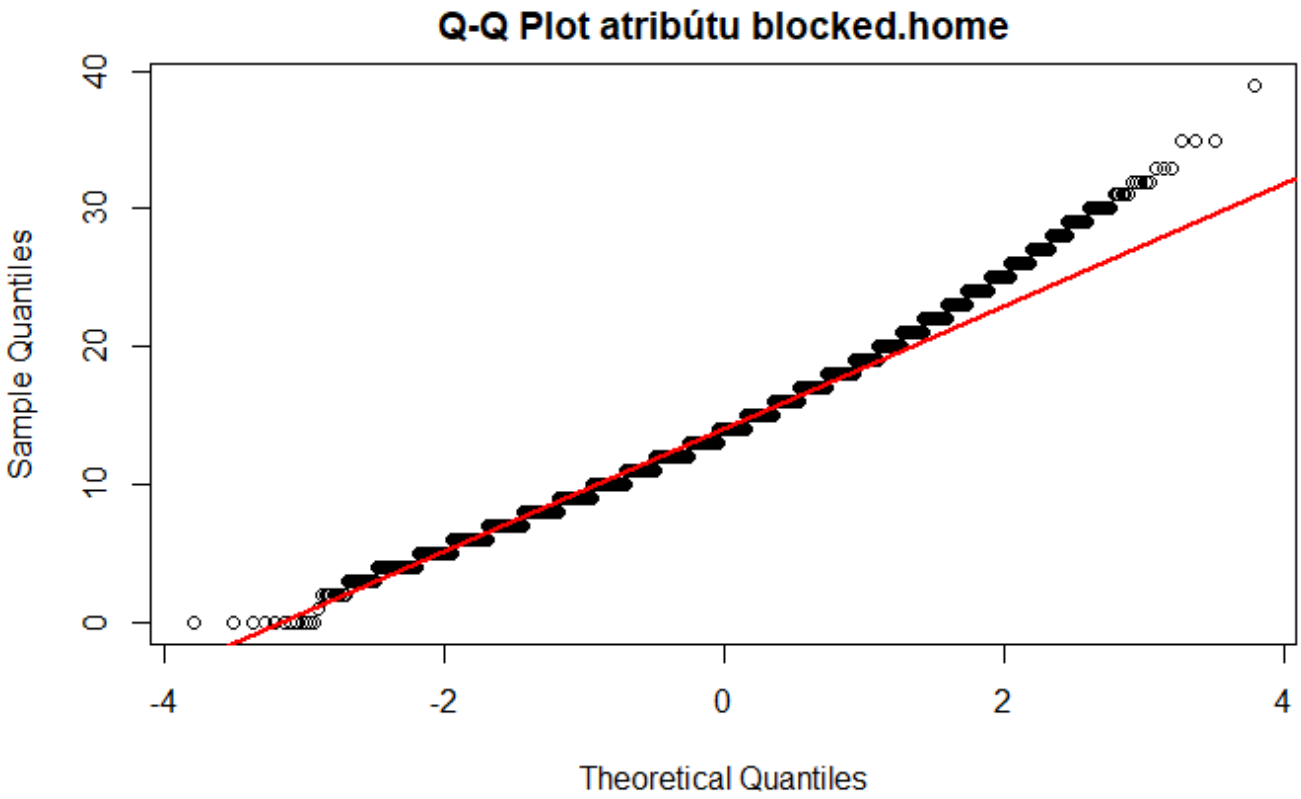
```
hist(df$blocked.away, xlab="blocked.away", main="Histogram atribútu blocked.away")
```



Hodnoty atribútu veľmi pripomínajú normálne rozdeleni, ale vidíme aj zopár vychýlených hodnôt, čo môže byť z dôvodu dlhší hier, napr. kvôli predĺženiu. Atribút **blocked.away** obsahuje jednu výrazne vychýlenú hodnotu, ktorú bude vhodné riešiť pri čistení dát. Distribúcia pripomína normálne rozdelenie, nakreslíme si QQ-plot a pozrieme sa na zhodu s krivkou teoretického normálneho rozdelenia.

Hide

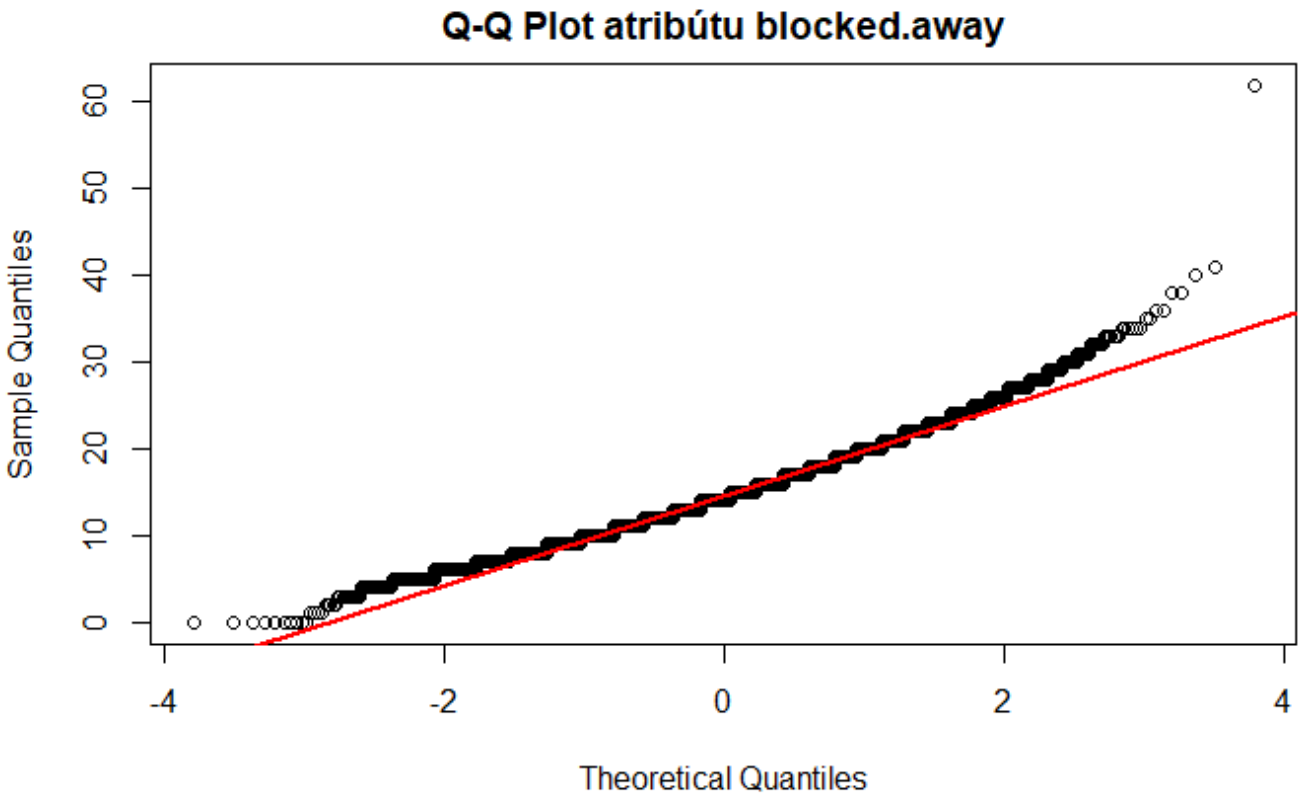
```
qqnorm(df$blocked.home, main="Q-Q Plot atribútu blocked.home")
qqline(df$blocked.home, col = "red", lwd = 2)
```



Atribút sa odkláňa od normálnej krivky pri hodnotách >= 20.

Hide

```
qqnorm(df$blocked.away, main="Q-Q Plot atribútu blocked.away")
qqline(df$blocked.away, col = "red", lwd = 2)
```

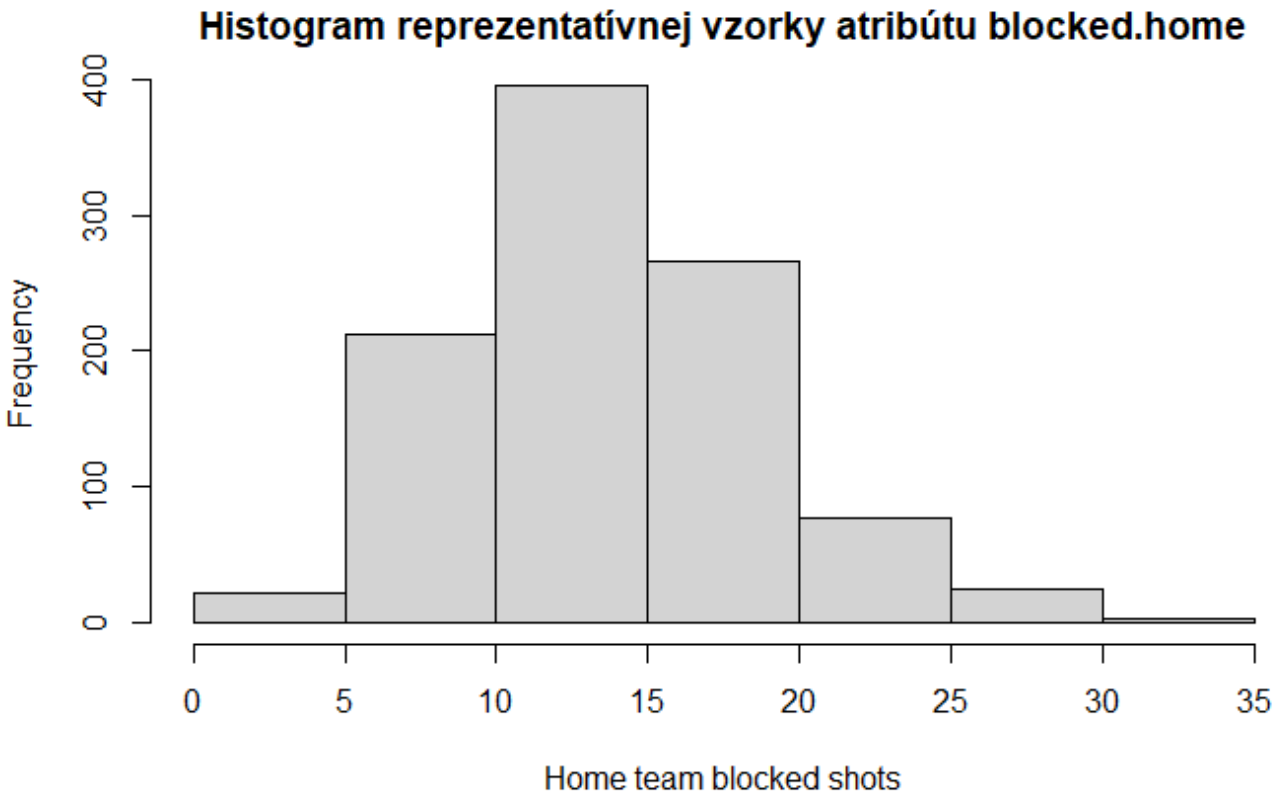


Atribút sa odkláňa od normálnej krivky pri hodnotách >= 20. Obsahuje jednu výrazne vychýlenú hodnotu (>60). Túto hodnotu bude vhodné redukovať tak, aby nehrozilo skreslenie výsledkov. Vykonáme ešte Shapiro-Wilkov test normality aby sme sa uistili, že atribúty nepochádzajú z normálneho rozdelenia.

Grafy indikujú signifikantnú podobnosť s normálnym rozdelením. Bude preto vhodné vykonať Shapiro-Wilkov test normality, pre ktorý je potrebné vybrať reprezentatívnu vzorku o veľkosti 1000.

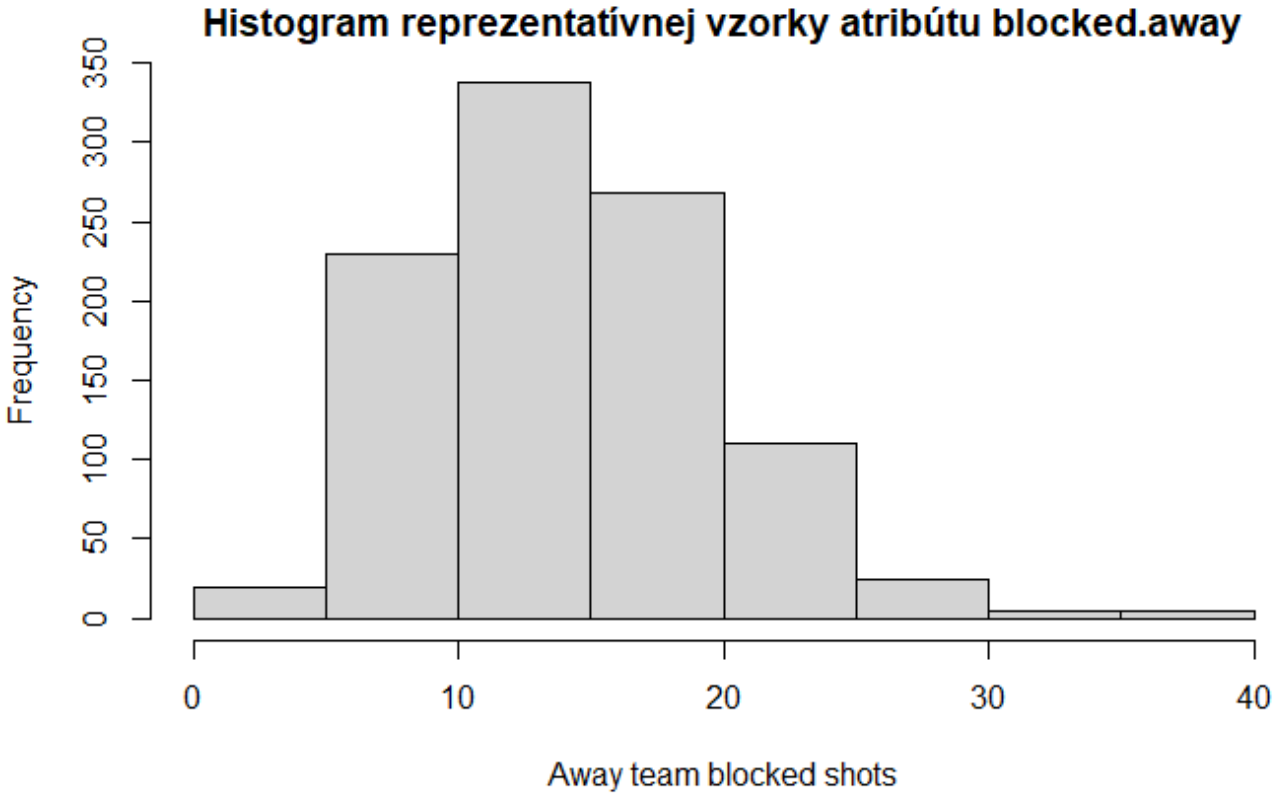
Hide

```
sample <- sample_n(df, 1000)
hist(sample$blocked.home, xlab="Home team blocked shots", main="Histogram reprezentat
ívnej vzorky atribútu blocked.home")
```



Hide

```
hist(sample$blocked.away, xlab="Away team blocked shots", main="Histogram reprezentat
ívnej vzorky atribútu blocked.away")
```



Hide

```
cat("Štatistika blocked.home\n")
```

```
Štatistika blocked.home
```

Hide

```
summary(df$blocked.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   11.00   14.00   14.09   17.00   39.00       4
```

Hide

```
cat("Štatistika vzorky blocked.home\n")
```

```
Štatistika vzorky blocked.home
```

Hide

```
summary(sample$blocked.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   11.00   14.00   14.25   18.00   35.00       1
```

Hide

```
cat("Štatistika blocked.away\n")
```

```
Štatistika blocked.away
```

Hide

```
summary(df$blocked.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   11.00   14.00   14.68   18.00   62.00       4
```

Hide

```
cat("Štatistika vzorky blcoked.away\n")
```

```
Štatistika vzorky blcoked.away
```

Hide

```
summary(sample$blocked.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   11.00   14.00   14.67   18.00   38.00       1
```

Vzorka má podobnú varianciu ako pôvodný dataset, preto na nej vykonáme test normality s hladinou p = 0.05.

Hide

```
shapiro.test(sample$blocked.home)
```

```

    Shapiro-Wilk normality test

data:  sample$blocked.home
W = 0.98171, p-value = 7.07e-10
```

Hide

```
shapiro.test(sample$blocked.away)
```

```

    Shapiro-Wilk normality test

data:  sample$blocked.away
W = 0.97375, p-value = 1.849e-12
```

Test v oboch prípadoch zamietame, atribút nepochádza z normálneho rozdelenia. Vychýlenú hodnotu v atribúte **blocked.away** bude vhodné pri čistení dát riešiť.

Atribúty abbreviation.home a abbreviation.away

charakteristika: atribút, ktoréh hodnoty obsahujú iba tri písmená a reprezentuje skratkú tímov., ktoré prosti sebe nastúpili v zápase. Pre nami vybrané sezóny bolo v NHL 31 tímov a teda má tento atribút 31 unikátnych hodnôt.

Tento atribút ma iba čisto informatívny charakter, môžeme akurát overiť či sa počet skratiek tímov reálne rovná počtu tímov patriacich do súťaže NHL (čiže 31 tímov).

Hide

```
length(unique(df$abbreviation.away))
```

```
[1] 32
```

Hide

```
length(unique(df$abbreviation.home))
```

```
[1] 32
```

Zistili sme, že v oboch atribútoch je o jednu hodnotu navyše. Pozrieme sa teda na hodnoty a skúsime nájsť vinníka.

Hide

```
unique(df$abbreviation.away)
```

```
 [1] "PHI" "ANA" "COL" "WPG" "CGY" "WSH" "TOR" "VAN" "CBJ" "EDM" "NYI" "CHI" "PIT" "T
BL" "BOS"
[16] "NJD" "OTT" "MTL" "LAK" "FLA" "CAR" "SJS" "MIN" "NYR" "DAL" "DET" "BUF" "ARI" "S
TL" "NSH"
[31] "VGK" NA
```

PHI - Philadelphia Flyers

ANA - Annaheim Ducks

COL - Colorado Avalanche

WPG - Winnipeg Jets

CGY - Calgary Flames

WSH - Washington Capitals

TOR - Toronto Maple Leafs

VAN - Vancouver Canucks

CBJ - Columbus Blue Jackets

EDM - Edmonton Oilers

NYI - New York Islanders

CHI - Chicago Blackhawks

PIT - Pittsburg Penguins

TBL - Tamba Bay Lightning

BOS - Boston Bruins

NJD - New Jersey Devils

OTT - Ottawa Senators

MTL - Montreal Canadiens

LAK - Los Angeles Kings

FLA - Florida Panthers

CAR - Carolina Hurricanes

SJS - San Jose Sharks

MIN - Minnesota Wild

NYR - New York Rangers

DAL - Dallas Stars

DET - Detroit Red Wings

BUF - Buffalo Sabres

ARI - Arizone Coyotes

STL - Saint Louis Blues

NSH - Nashville Predators

VGK - Las Vegas Golden Knights

NA - hodnota ktorú nadobúda abbreviation pri type A - zápasy mimo NHL

Všetky hodnoty boli popísané a taktiež bol vysvetlený o jeden väčší počet - ten bude pri čistení dát odstránený. Tento atribút však z hľadiska štatistiky neprináša žiadnu dodatočnú hodnotu, slúži iba na priradenie zápasov k reálnym názvom tímov dodatočne ako k iba im ID-čkam.

Atribúty settled_in

charakteristika: atribút, ktorý ukazuje ako skončil zápas. V NHL môže zápas skončiť viacerými spôsobmi: v riadnom hracom čase, v predĺžení, po nájazdoch (v starších sezónach), alebo z netradičných dôvodov (roztopenie ľadu, výpadok elektrického prúsu, …).

Hide

```
unique(df$settled_in)
```

```
[1] "REG" "OT"  "tbc"
```

Tento atribút nadobúda 3 hodnoty - REG, OT a tbc. REG - rozhodnutý zápas v riadnom hraciom čase OT - zápas rozhodnutý v predĺžení tbc - pravdepodobne zrušený zápas

Keďže si niesme istý, čo reprezentuje hodnota atribútu “tbc”, vypíŠeme si záznamy ktoré ju obsahujú.

Hide

```
subset(df, settled_in=='tbc')
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	FALSE	0	0	
20172018	P	0	0	28	FALSE	*NA*	*NA*	
20172018	P	0	0	26	FALSE	0	0	
20172018	P	0	0	52	FALSE	*NA*	*NA*	

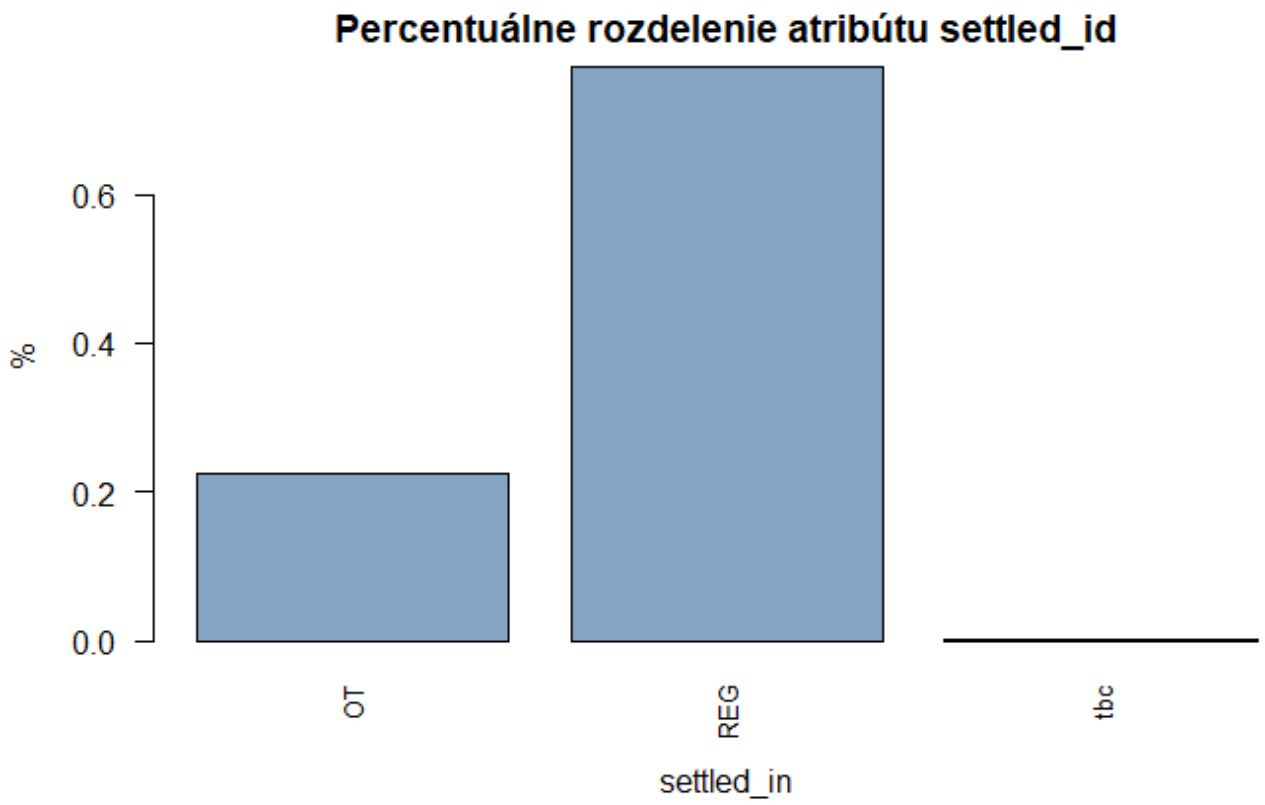
season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <lgl>	shots.away <int>	hits.away <int>	pim.a <
20172018	P	0	0	54	FALSE	*NA*	*NA*	
20172018	P	0	0	54	FALSE	0	0	
20172018	P	0	0	15	FALSE	0	0	
20172018	P	0	0	5	FALSE	0	0	
20162017	P	0	0	20	FALSE	0	0	
20162017	P	0	0	10	FALSE	0	0	

1-10 of 19 rows | 1-9 of 31 columns Previous **1** 2 Next

V drtivej väčšine prípadov sa jedná o zrušené zápasy - vyskytujú sa tu však aj zápasy, ktorých štatistiky nadobúdajú hodnoty odlišné od 0 - vo všetkých prípadoch majú tieto zápasy však type = A, čo sme už v predošlej analýze atribútu type usúdili, že sa nejedná o valídne NHL zápasy a preto budú z datasetu pri čistení odstránené. Vieme povedať, že všetky zápasy ktoré nadobúdajú hodnotu tbc budú teda odstránené - ani jeden totižto neprináša pridanú hodnotu (0-vé hodnoty relevantných štatistík alebo zápasy mimo súťaže NHL).

Hide

```
barplot(prop.table(table(df$settled_in)), las=2, cex.names=.8, col=rgb(0.2,0.4,0.6,0.
6), ylab="%", xlab="settled_in", main="Percentuálne rozdelenie atribútu settled_id")
```



Z grafu vidíme, že cca 20% zápasov je ukončených mimo riadneho hracieho času (v predĺžení). Percentuálny pomer hodnoty tbc je zanedbateľný. Okolo 80% zápasov je tým pádom rozhodnutých v riadnom hracom čase.

Pri párovej analýze môže byť dodatočne zaujímave pozrieť sa, koľko zápasov play-off sa končí v predĺžení oproti zápasom regulárnej sezóny (predpokladáme že v play-off bývajú zápasy vyrovnanejšie a tým pádom je častejšie po riadnom hracom čase remízový stav).

Keďže tento atribút reálne nadobúda iba 2 hodnoty, bude možné ho pri fáze transformácie dát konvertovať na binárny atribút - ten je jednoduchšie strojovo spracovateľný a je moŽné sledovať aj závislosti s inými numerickými atribútmi napr. pomocou korelácií alebo logistickej regresie.

Atribúty save_percentage.home a save_percentage.away

charakteristika: Tento atribút reprezentuje úspešnosť brankára v precentách. Úspešnosť brankára nie je súčasťou pôvodného datasetu - atribút sme si vytvorili a údaje doplnili na základe hodnôt z iných atribútov.

Hide

```
summary(df$save_percentage.home)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.5455  0.8696  0.9130  0.9058  0.9487  1.0000      14 
```

Hide

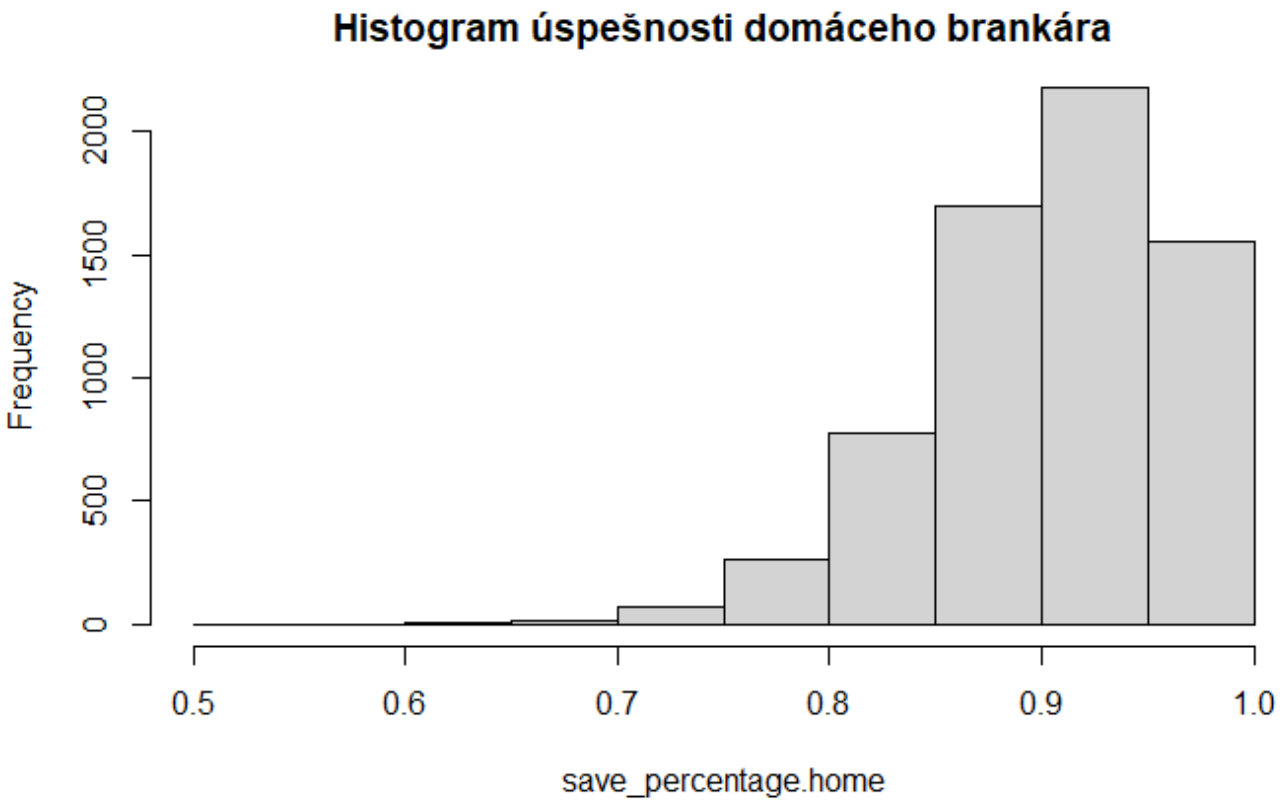
```
summary(df$save_percentage.away)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.5833  0.8649  0.9070  0.9017  0.9444  1.0000      14 
```

Hodnoty úspešnosti brankárov sme vypočítali sami a poskytujú veľmi podobné priemerné hodnoty, ako uvádza oficiálna nHL stránka. Môžeme ale vidieť, že 14 záznamov má tento atribút nevyplnený. Tento problém budeme riešiť pri čistení dát, prípadne následne môŽeme analýzu opätovne vykonať a pozrieť sa či sa niečo nezmení - aj keď vieme, že 14 záznamov je veľmi málo a na celkový výsledok analýzy pravdepodobne budú mať iba veľmi minimálny vplyv.

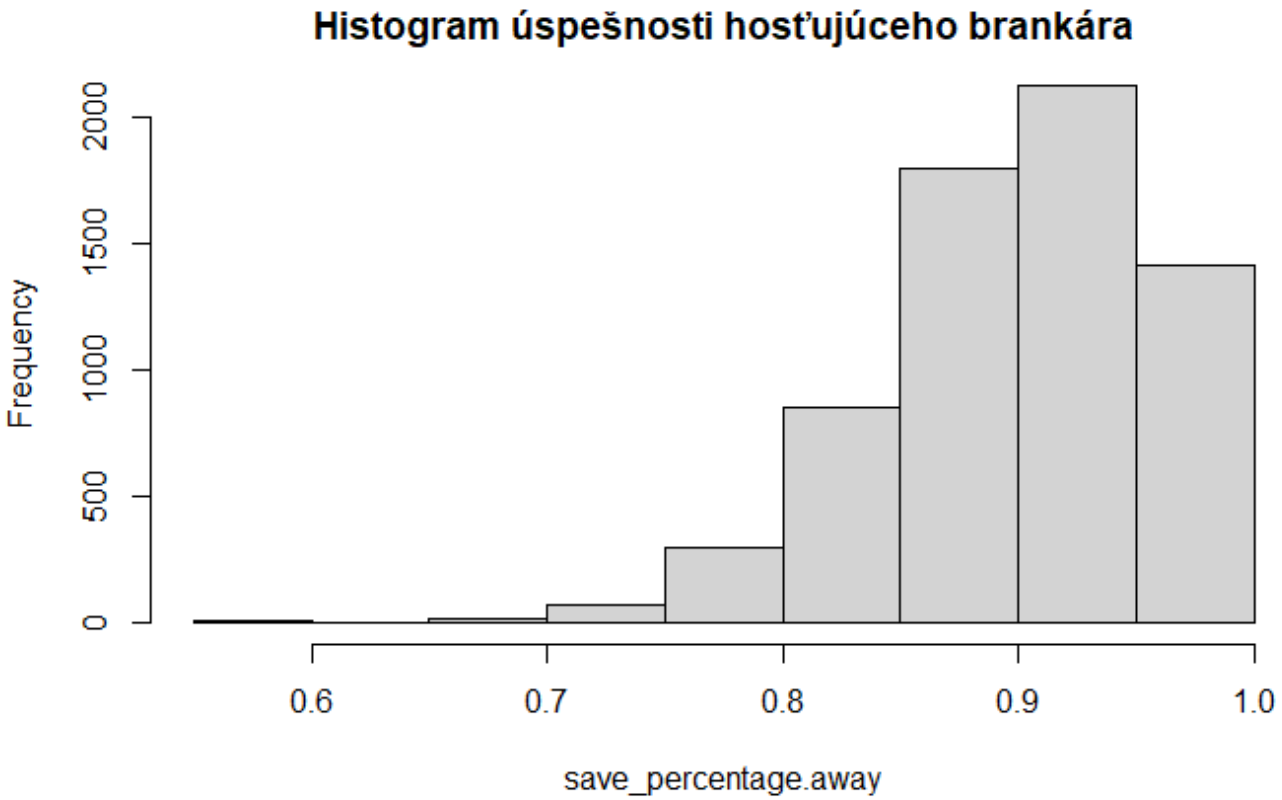
Hide

```
hist(df$save_percentage.home, xlab="save_percentage.home", main="Histogram úspešnosti
domáceho brankára")
abline(v = c(mean(df$save_percentage.home),  median(df$save_percentage.home)), col=c
("green", "blue"), lty=c(2,3), lwd=c(3,3))
```



Hide

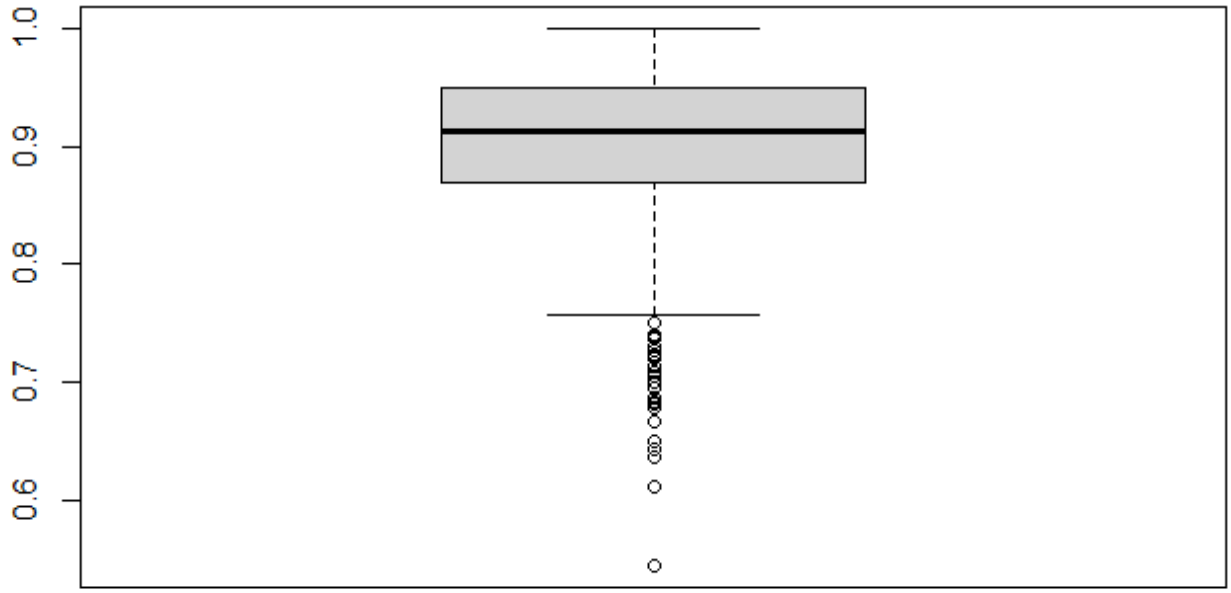
```
hist(df$save_percentage.away, xlab="save_percentage.away", main="Histogram úspešnosti
hosťujúceho brankára")
abline(v = c(mean(df$save_percentage.home),  median(df$save_percentage.home)), col=c
("green", "blue"), lty=c(2,3), lwd=c(3,3))
```



Graf nám ukazuje, že hodnoty netvoria normálne rozdelenie. Skôr sú naklonené vľavo s malým počtom vychýlených hodnôt (okolo hodnoty 0.6) - 60% úspešnosť zákrokov je síce naozaj nízka, no jedná sa o reálne hodnoty ktoré sa v NHL a celkovo pri hokeji zriedkavo vysknú.

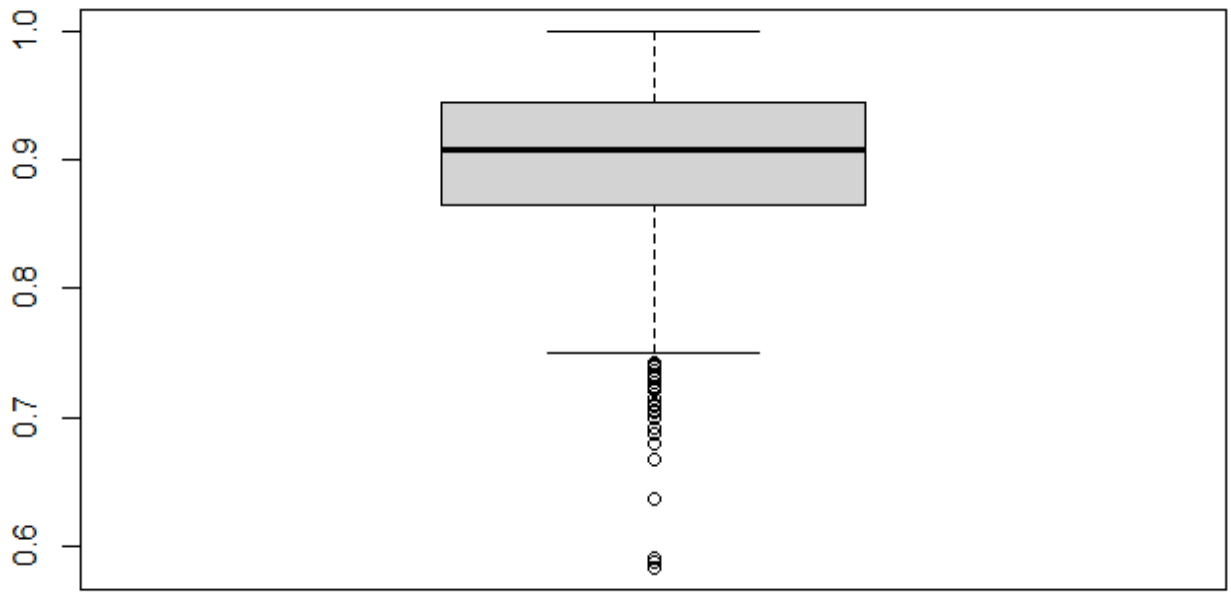
Hide

```
boxplot(df$save_percentage.home)
```



Hide

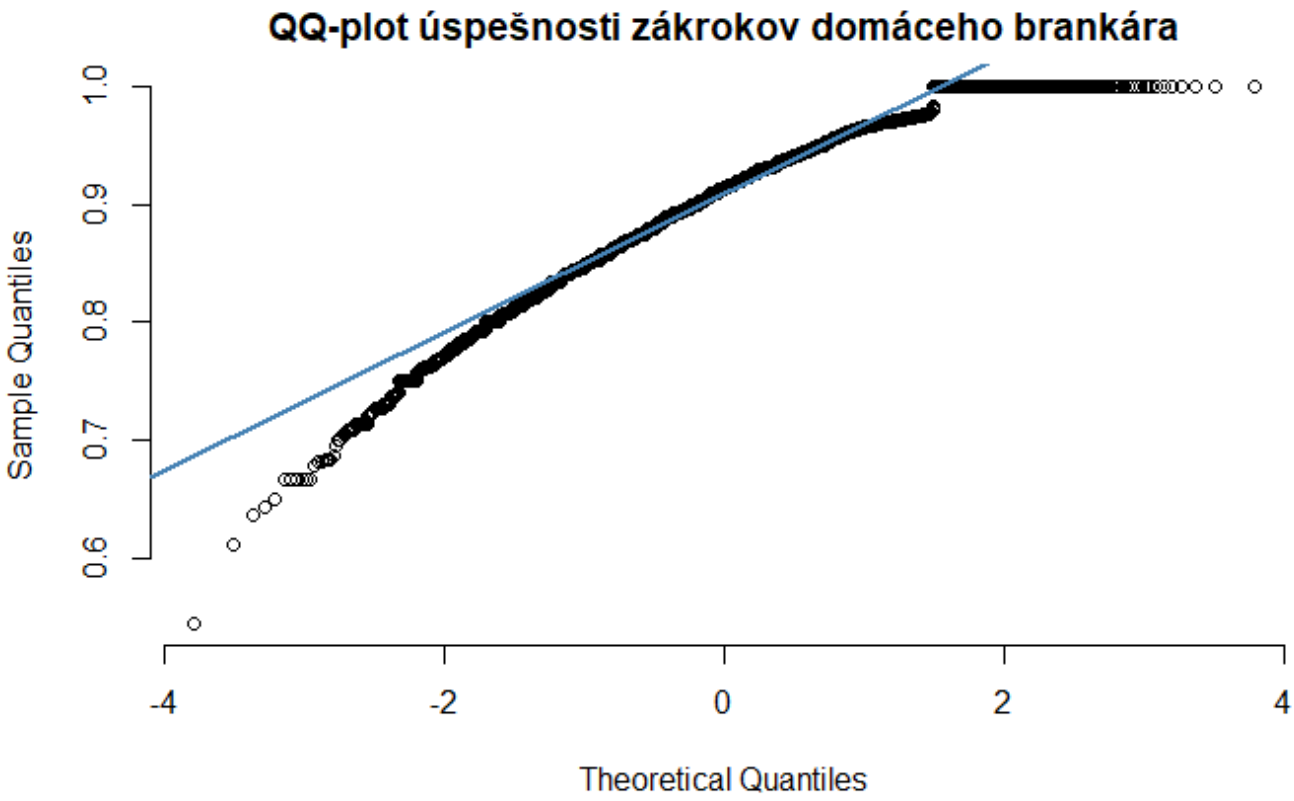
```
boxplot(df$save_percentage.away)
```



Vychýlené hodnoty sú zo spodného kvantilu a nie je ich veľa. Môžeme sa na ne pozrieť pri čistení dát a zhodnotiť ich úpravu, no tá bude aj vzhľadom na predošlé konštatovanie o reálnom výskyte týchto údajov veľmi nepravdepodobná - predsa len nám tieto hodnoty poskytujú dôležitú štatistickú informáciu o výkone brankárov tímov, ktorú potenciálnou úpravou nechceme stratiť.

Hide

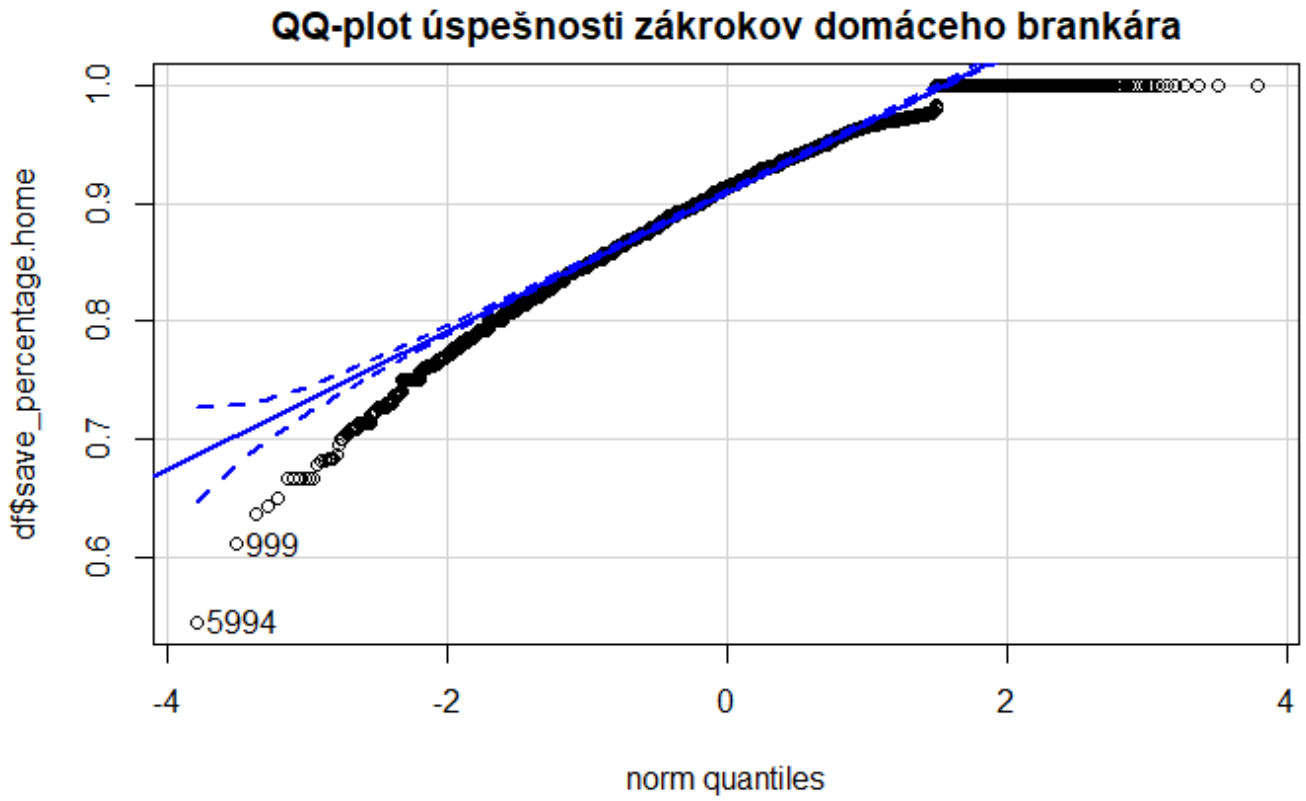
```
qqnorm(df$save_percentage.home, pch = 1, frame = FALSE, main = "QQ-plot úspešnosti zá
krokov domáceho brankára")
qqline(df$save_percentage.home, col = "steelblue", lwd = 2)
```



Hide

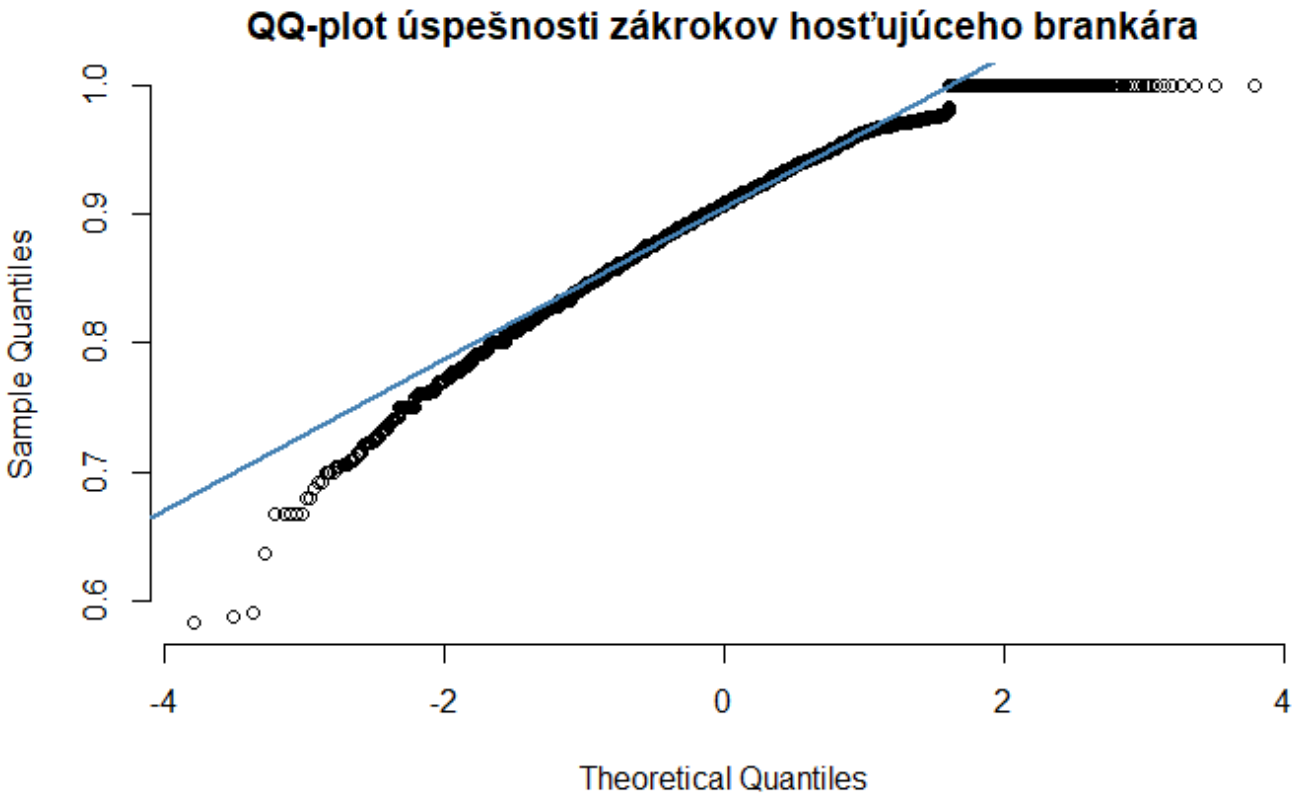
```
qqPlot(df$save_percentage.home, main = "QQ-plot úspešnosti zákrokov domáceho brankár
a")
```

```
[1] 5994  999
```



Hide

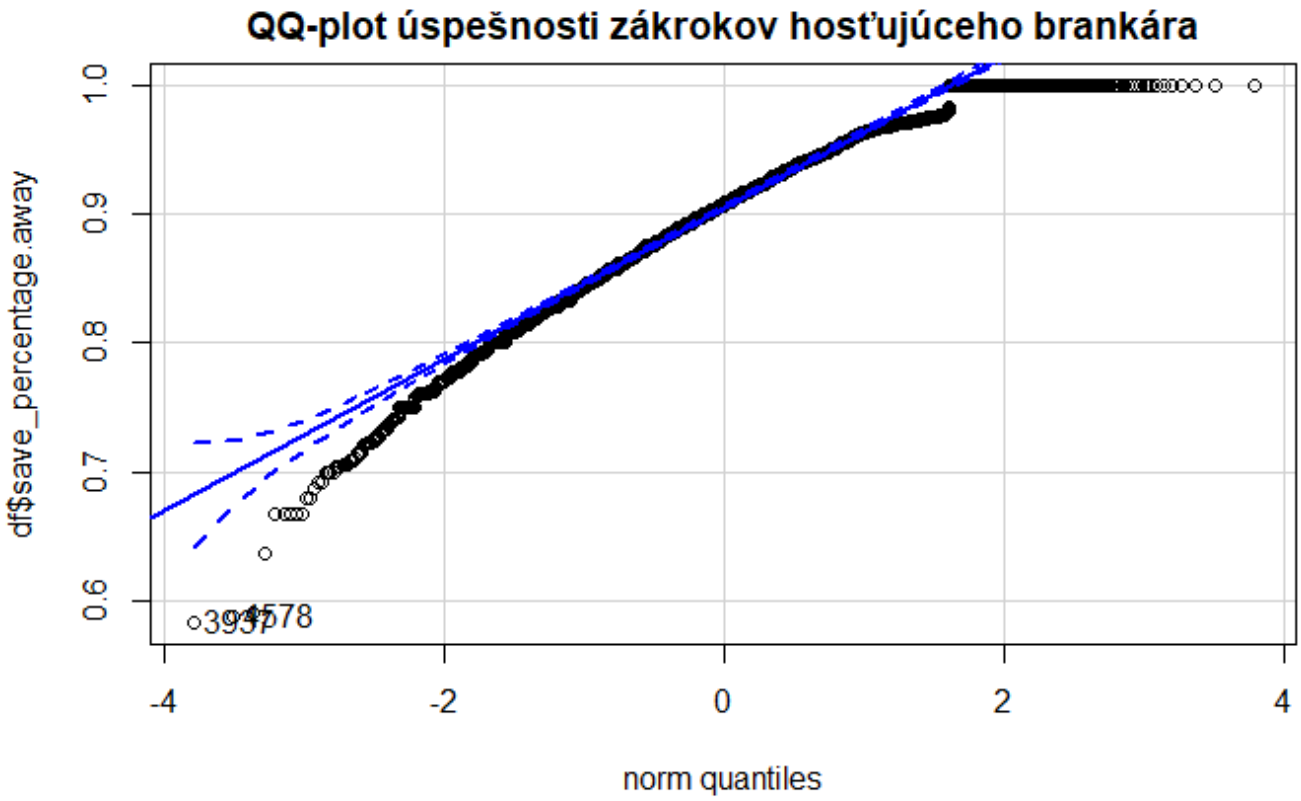
```
qqnorm(df$save_percentage.away, pch = 1, frame = FALSE,main = "QQ-plot úspešnosti zák
rokov hosťujúceho brankára")
qqline(df$save_percentage.away, col = "steelblue", lwd = 2)
```



Hide

```
qqPlot(df$save_percentage.away, main = "QQ-plot úspešnosti zákrokov hosťujúceho brank
ára")
```

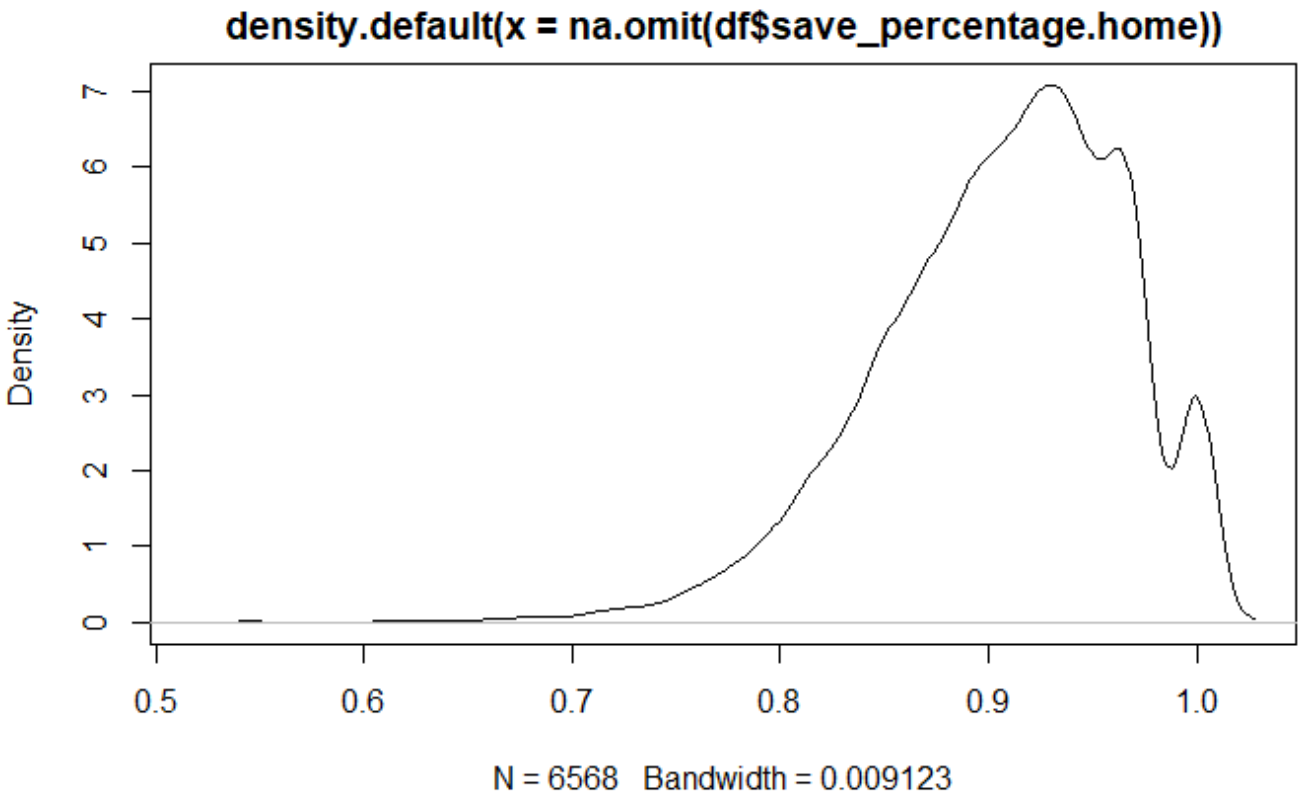
```
[1] 3937 4578
```



Grafy nám ukazujú, že sa nejedná o normálne rozdelenie, pretože vrchné a spodné hodnoty majú vysokú odchylku. Na určitom intervale síce hodnoty pripomínajú normálne rozdelenie (napr. pre **save_percentage.away** aj **save_percentage.home** interval 0.8 až cca 0.96).

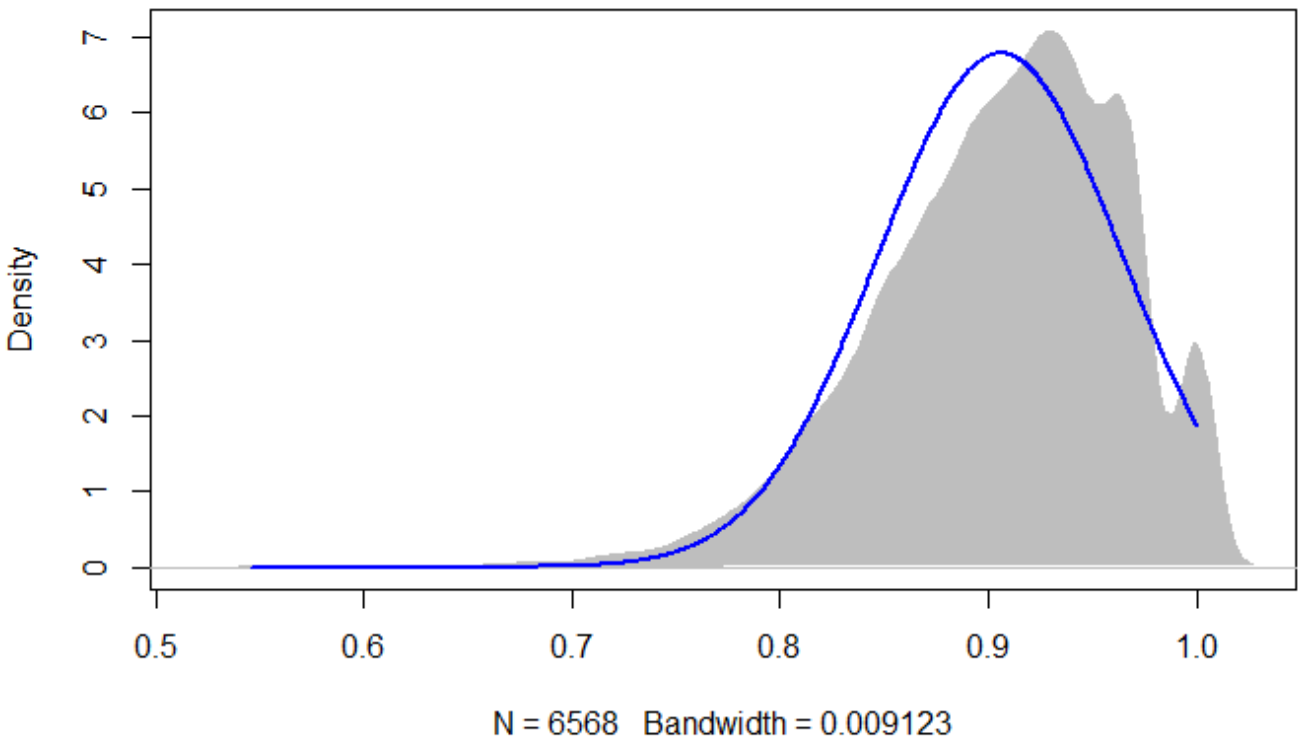
Hide

```
plot(density(na.omit(df$save_percentage.home)))
```



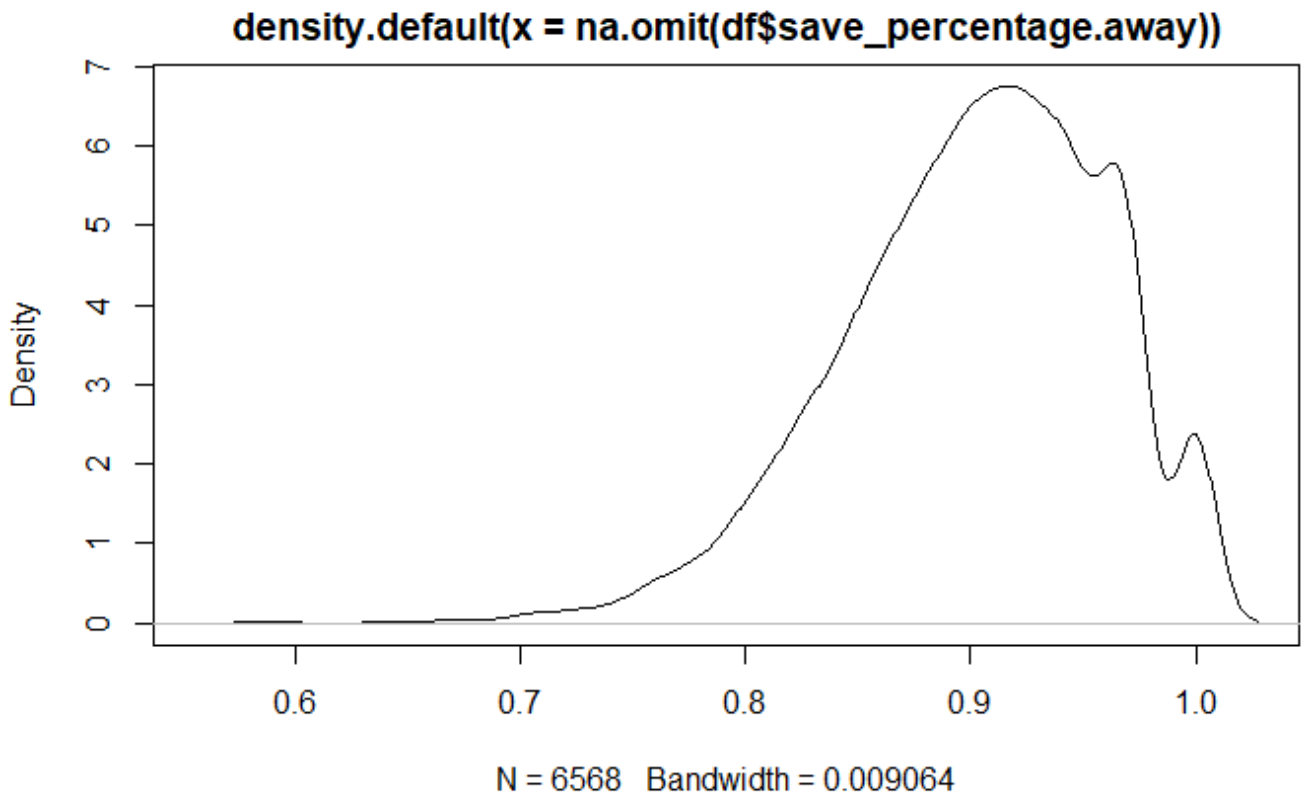
Hide

```
plotNormalDensity(df$save_percentage.home)
```



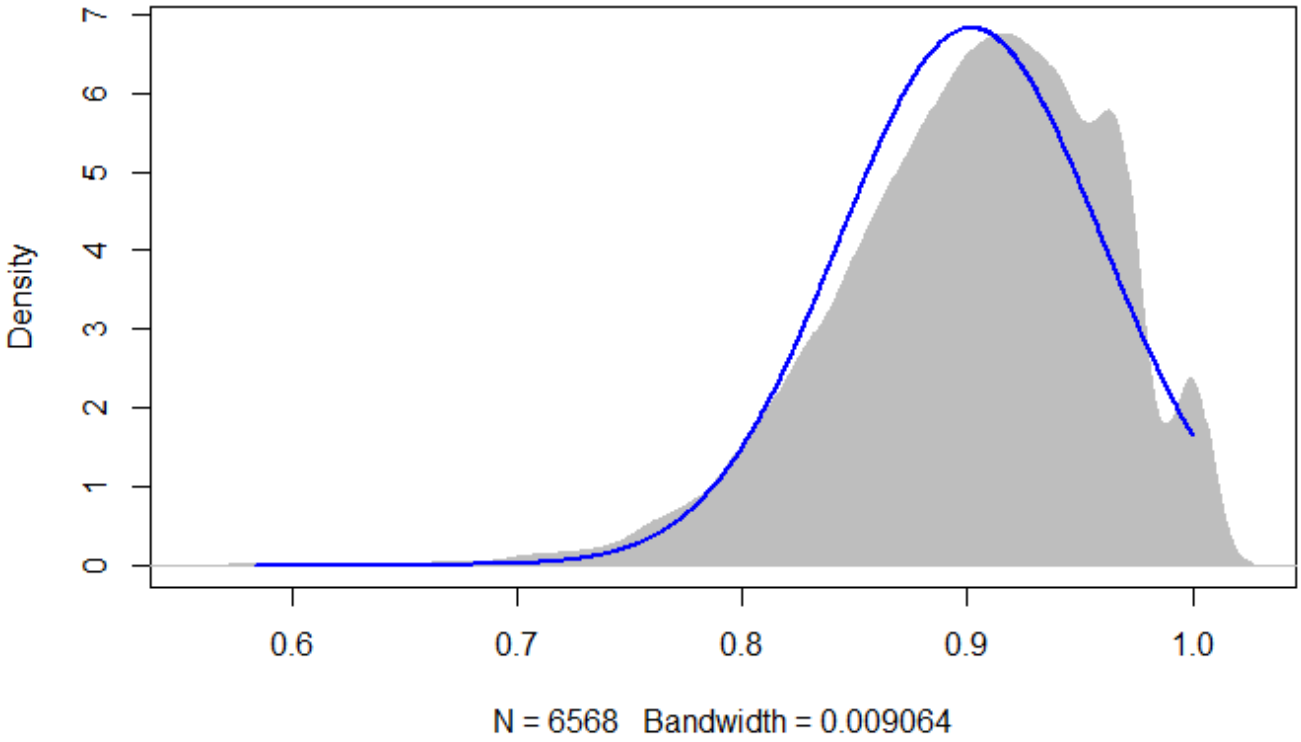
Hide

```
plot(density(na.omit(df$save_percentage.away)))
```



Hide

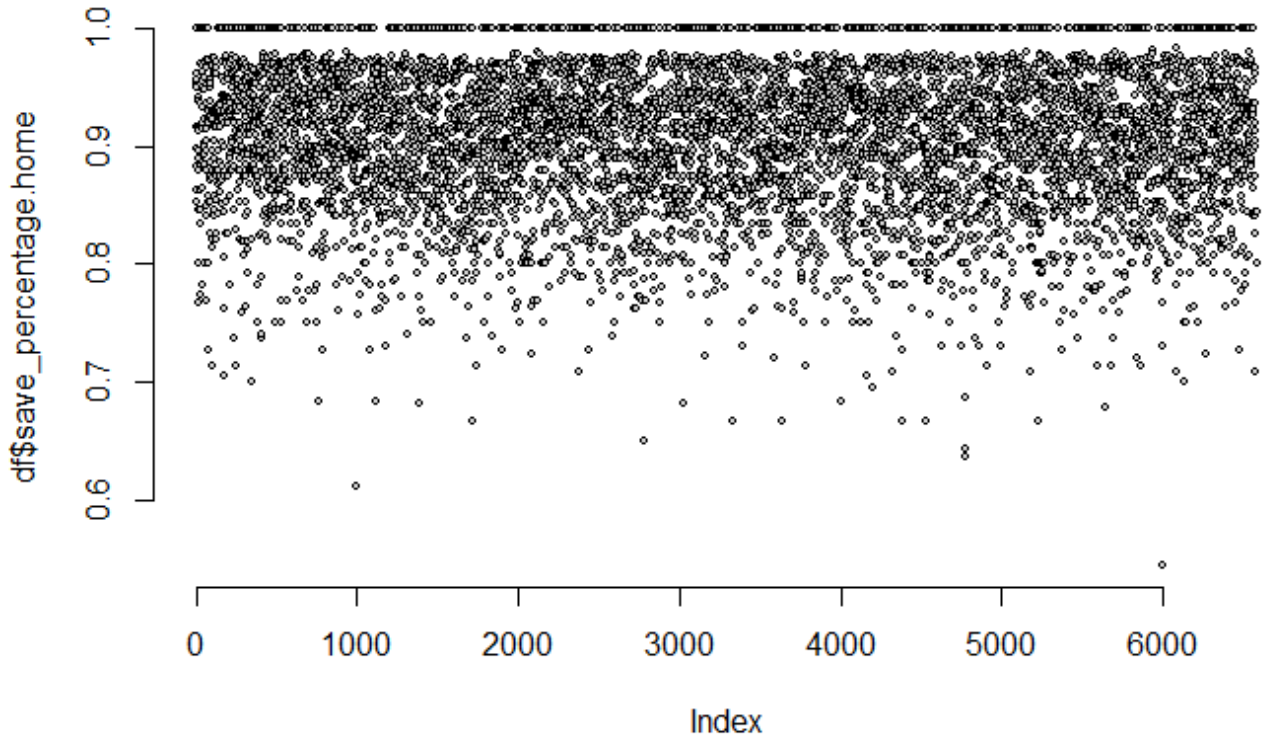
```
plotNormalDensity(df$save_percentage.away)
```



Grafy hustoty s krivkou normálneho rozdelenia nám ukazujp, že hodnoty sa veľmi podobajú normálnemu rozdeleniu, ale obsahujú viacero lokálnych vrcholov, ktoré sú celkom nezvyčajné.

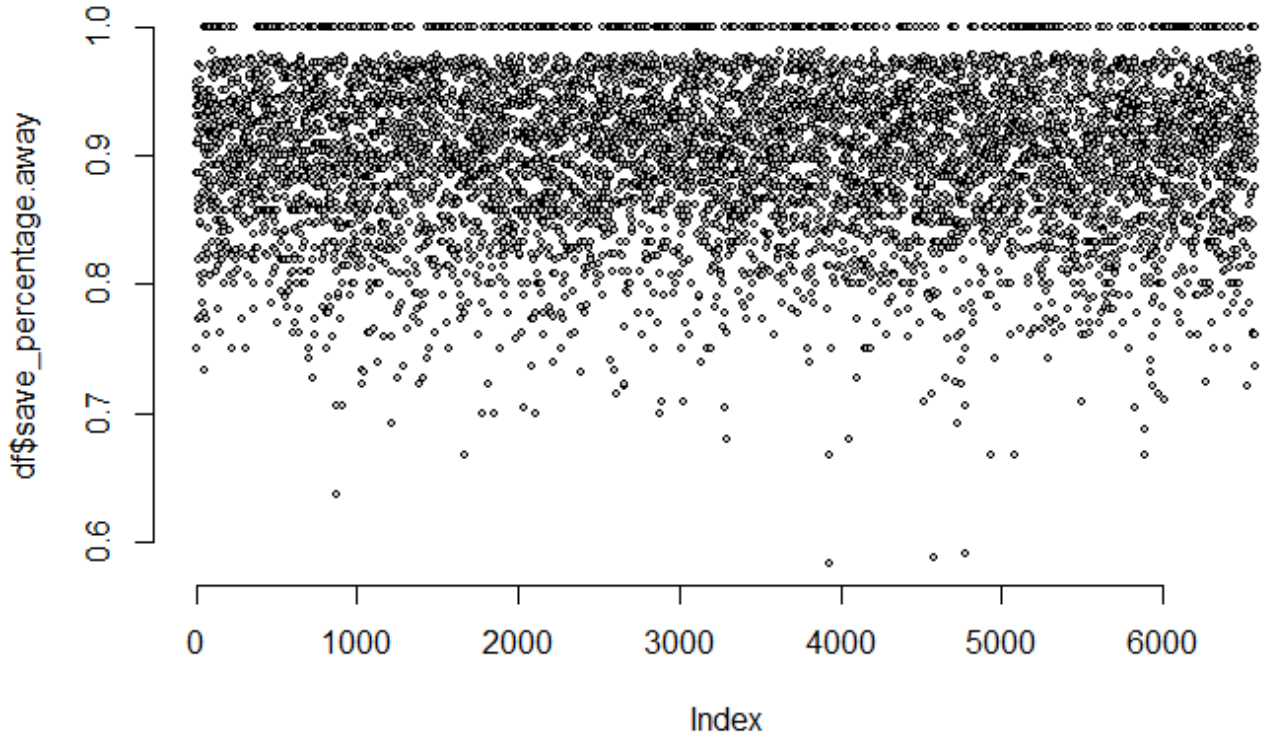
Hide

```
plot(df$save_percentage.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, f
rame = FALSE)
```



Hide

```
plot(df$save_percentage.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, f
rame = FALSE)
```



Graf hustoty hodnôt nám ukazuje, že medzi čistými výkonmi brankárov a inkasovaným aspoň jedným gólom je veľká medzera, čo znamená, že góly padajú aj pri menšom počte striel na bránu. Taktiež v niektorých zápasoch majú brankári veľmi majú percentuálnu úspešnosť, čo môže byť spôsobené napríklad zranením pri zákroku (alebo veľmi zlým dňom pre brankára), z ktorého padol gól. Medzera medzi 100% úspešnosťou a zvyšnými hodnotami je spôsobená aj počtom striel - pokiaľ brankár inkasuje gól a má mať percentuálnu úspešnosť veľmi blízku číslu 100, muselo by naňho v zápase byť vystrelených naozaj veľa striel - preto sa väčšinou percentuálne úspešnosti pri najlepších brankároch pohybujú “iba” okolo hodnoty 96%. 99% pravdepodobnosť je preto takmer nereálna.

Test normality nebude potrebný, keďže grafy jednoznačne ukazujú, že distribúcia atribútov **save_percentage** nebude pochádzať z normálneho rozdelenie ale z rozdelenia s distribúciou naklonenou vľavo.

Hide

```
skewness(df$save_percentage.home)
```

```
[1] NA
```

Hide

```
skewness(df$save_percentage.away)
```

```
[1] NA
```

Šikmosť nie je možné vypočítať, keďže atribút ešte obsahuje NA hodnoty. Tie budú odstranené pri čistení dát.

Zhrnutie: Atribúty **save_percentage.home** a **save_percentage.away** sú spojité atribúty reprezentujúceho percentuálnu úspešnosť brankára (vyjadrené pomocou desatin. čísla), ktoré majú distribúciu naklonenú vĺavo.

Párová analýza

V tejto časti chceme identifikovať možné vzťahy medzi jednotlivými atribútmi prostredníctvom výpočtu korelácií, respektíve korelačnej matice. Nakoľko niektoré atribúty obsahujú nenumerické hodnoty, pre prvotnú identifikáciu možných vzťahov medzi atribútmi použijeme iba atribúty s numerickými hodnotami. Čo sa týka nenumerických atribútov **won.home** a **won.away**, tie aktuálne obsahujú pravdivostné hodnoty TRUE alebo FALSe a teda ich môžeme zmeniť na numerické (nahradením TRUE za 1 a FALSE za 0).

Hide

```
df$won.home[df$won.home == "TRUE"] <- "1"
df$won.home[df$won.home == "FALSE"] <- "0"
df$won.away[df$won.away == "TRUE"] <- "1"
df$won.away[df$won.away == "FALSE"] <- "0"
class(df$won.home) = "numeric"
class(df$won.away) = "numeric"
unique(df$won.home)
```

```
[1] 1 0
```

Hide

```
unique(df$won.away)
```

```
[1] 0 1
```

Pre vytvorenie korelačnej matice potrebujeme numerické atribúty. Vytvoríme si pomocný dataset, ktorý bude obsahovať iba numerické atribúty pôvodného datasetu.

Hide

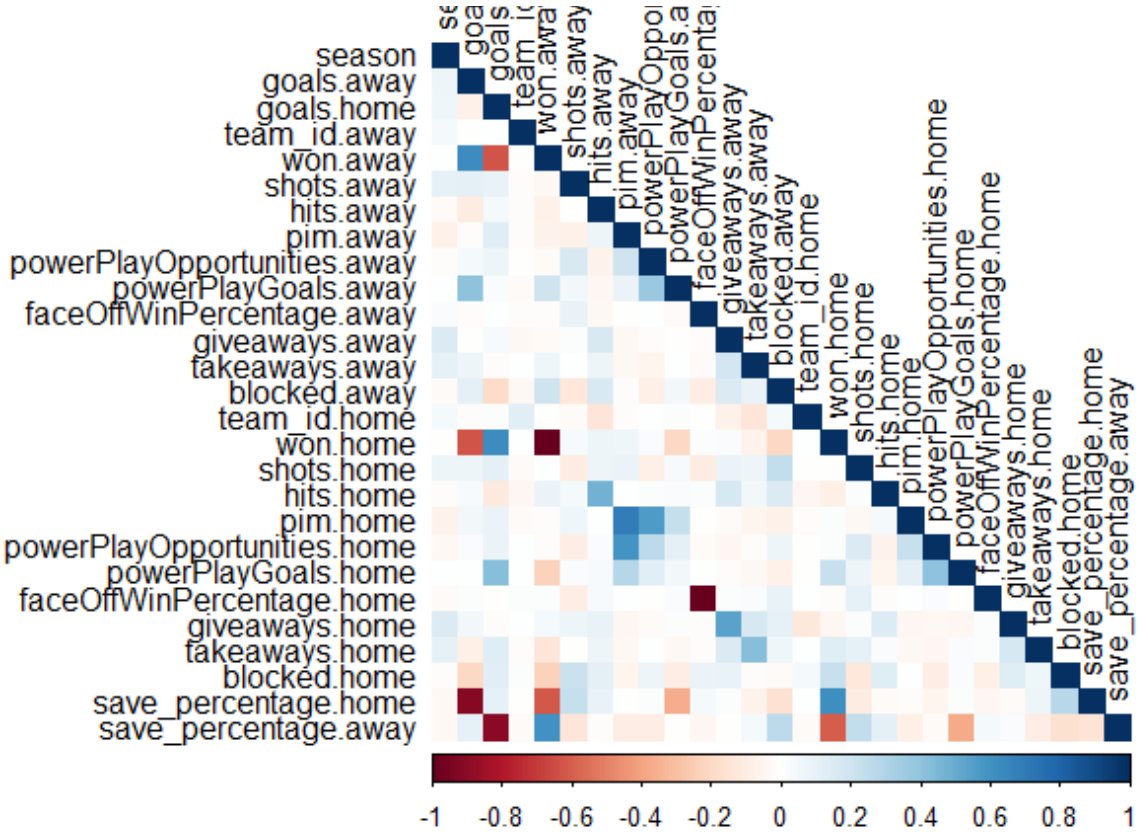
```
df_numeric <- subset(df, select=-c(type, abbreviation.away, settled_in, abbreviation.
home))
sapply(df_numeric, is.numeric)
```

```
                    season                  goals.away                  goals.home 
                      TRUE                        TRUE                        TRUE 
             team_id.away                   won.away                  shots.away 
                      TRUE                        TRUE                        TRUE 
                hits.away                  pim.away powerPlayOpportunities.away 
                      TRUE                        TRUE                        TRUE 
      powerPlayGoals.away   faceOffWinPercentage.away              giveaways.away 
                      TRUE                        TRUE                        TRUE 
          takeaways.away               blocked.away                team_id.home 
                      TRUE                        TRUE                        TRUE 
                won.home                  shots.home                   hits.home 
                      TRUE                        TRUE                        TRUE 
                pim.home powerPlayOpportunities.home        powerPlayGoals.home 
                      TRUE                        TRUE                        TRUE 
 faceOffWinPercentage.home             giveaways.home             takeaways.home 
                      TRUE                        TRUE                        TRUE 
            blocked.home       save_percentage.home       save_percentage.away 
                      TRUE                        TRUE                        TRUE 
```

Z výpisu vyššie môžeme vidieť, že všetky atribúty pomocného datasetu **df_numeric** sú numerické. Môžeme vytvoriť korelačnú maticu.

Hide

```
corrplot(cor(df_numeric, use="complete.obs"), type="lower", method="color", tl.col="b
lack")
```



Z korelačnej matice môžeme vidieť, že najväčšia korelácia (záporná) je medzi atribútmi **won.home** a **won.away**, čo je pochopiteľné, pretože keď jeden tím vyhrá, druhý musí prehrať. Podobné záporné korelácie môžeme vidieť medzi atribútmi **faceOffWinPercentage.home** a **faceOffWinPercentage.away**, ktoré hovoria o vyhratých vhadzovaniach. Tiež si môžeme všimnúť vysoké korelácie medzi atribútmi **save_percentage.home** a **goals.away**, **save_percentage.home** a **won.away** alebo **save_percentage.home** a **won.home** (podobne aj pre atribút **save_percentage.away** s niektorými atribútmi domáceho tímu). Tieto korelácie značia napríklad o tom ako úspešnoť brankára domáceho tímu ovplyvňuje počet gólov hosťujúceho tímu alebo jeho šancu na výhru (podobne to platí aj naopak). Ďalšie korelácie sú medzi atribútmi **won.home** a **goals.home**:

Hide

```
cor(df$save_percentage.away, df$goals.home)
```

```
[1] NA
```

Hide

```
cor(df$save_percentage.away, df$won.home)
```

```
[1] NA
```

Hide

```
cor(df$save_percentage.home, df$goals.away)
```

```
[1] NA
```

Hide

```
cor(df$save_percentage.home, df$won.away)
```

```
[1] NA
```

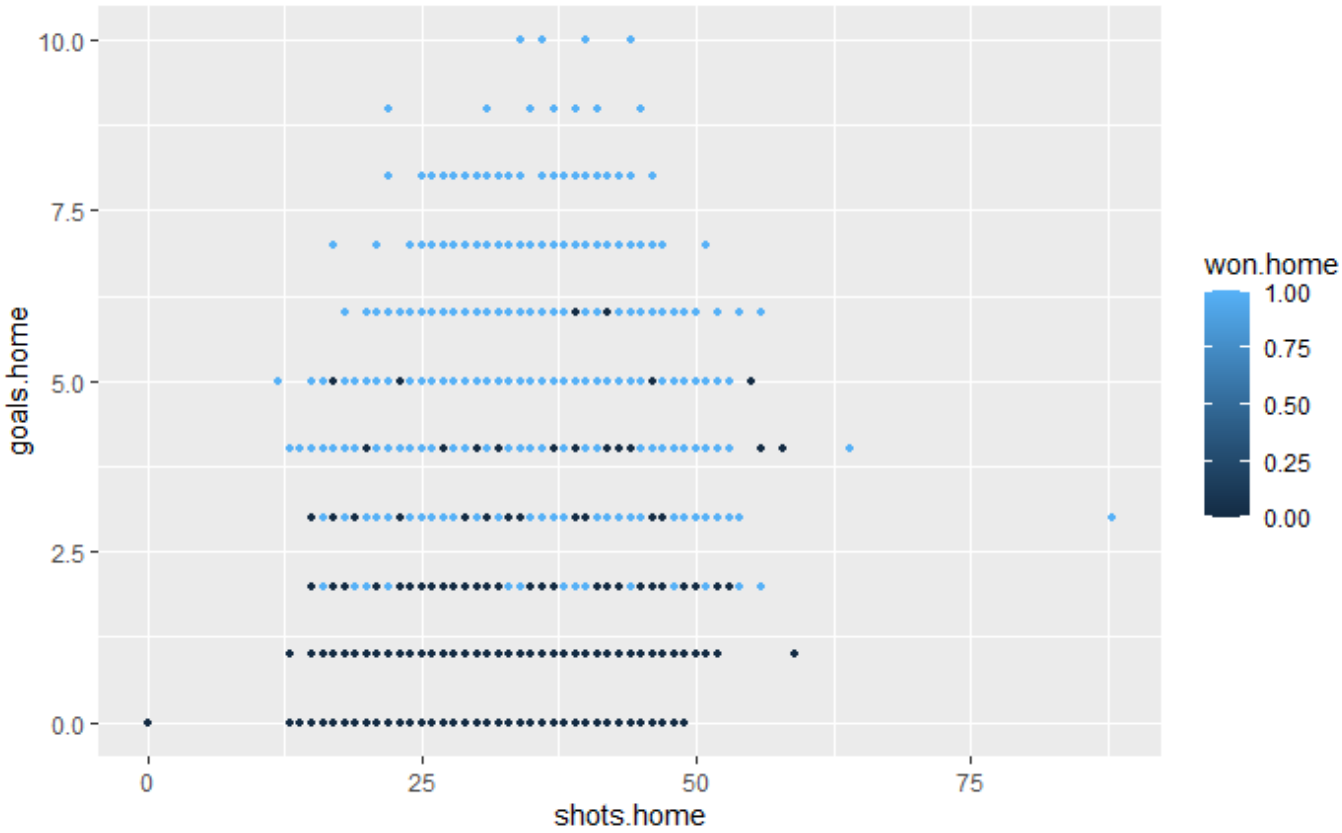
Hide

```
cor(df$won.home, df$goals.home)
```

```
[1] 0.6289938
```

Hide

```
ggplot(df, aes(x=shots.home, y=goals.home, color=won.home)) + geom_point(size=1)
```



Korelácia týchto dvoch atribútov je tiež logická. Môžeme z nej vyvodiť, že čím viac gólov tím strelí, tým je väčšia šanca, že zápas vyhrá. Podobne to platí aj pre atribúty **won.away** a **goals.away**:

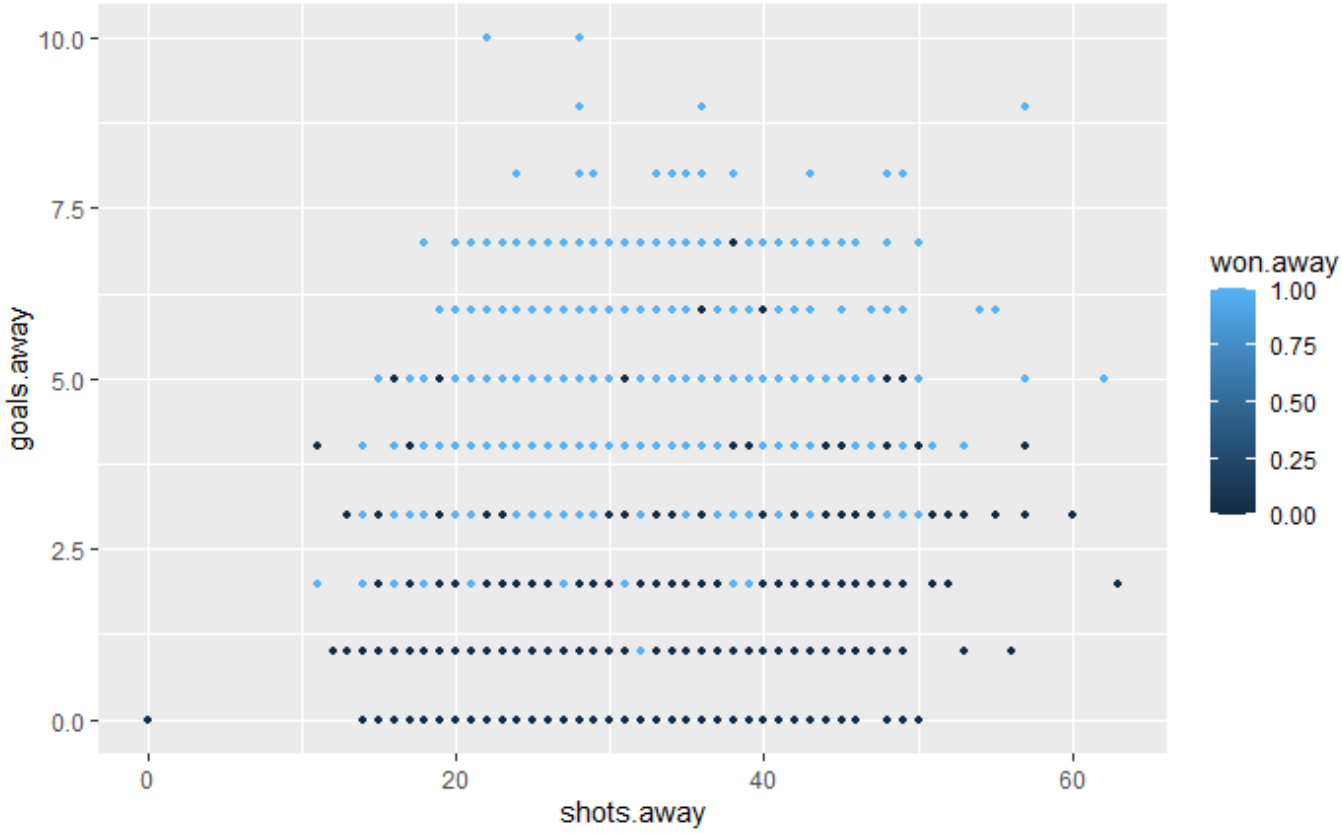
Hide

```
cor(df$won.away, df$goals.away)
```

```
[1] 0.6295663
```

Hide

```
ggplot(df, aes(x=shots.away, y=goals.away, color=won.away)) + geom_point(size=1)
```



Rovnako môžeme sledovať aj koreláciu medzi atribútmi **won.away** a **goals.home** alebo **won.home** a **goals.away**. V tomto prípade však budú korelácie záporné:

Hide

```
cor(df$won.away, df$goals.home)
```

```
[1] -0.6216961
```

Hide

```
cor(df$won.home, df$goals.away)
```

```
[1] -0.6221667
```

Ďalšiu koreláciu môžeme pozorovať medzi atribútmi **pim.home** a **pim.away**. Atribúty hovoria o trestných minutách tímov. Určitá korelácia je pochopiteľná, pretože ak sa napríklad dvaja hráči pobijú, idú obaja na trestnú lavičku. V takomto prípade dostávajú trestné minúty oba tíma zároveň, čoho dôsledkom je zvýšená korelácia týchto atribútov. Atribúty však obsahujú N/A hodnoty, takže ich pre znázornenie korelácie budeme ignorovať.

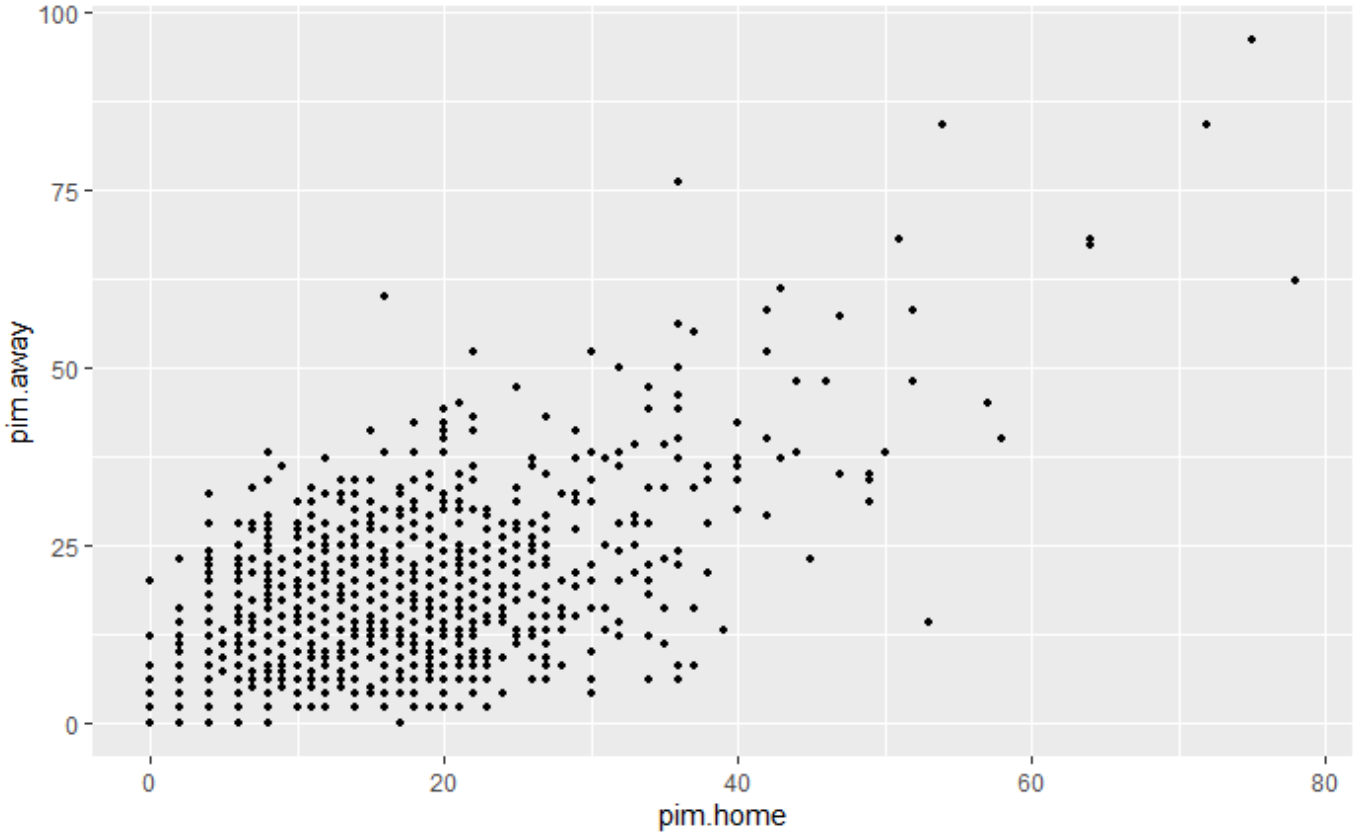
Hide

```
cor(df$pim.home, df$pim.away, use="complete.obs")
```

```
[1] 0.6993652
```

Hide

```
ggplot(df, aes(x=pim.home, y=pim.away)) + geom_point(size=1)
```



Trestné minúty jedného tímu znamenajú presilovku druhého tímu. Z toho tiež vyplýva korelácia medzi atribútmi **pim.home** s **powerPlayOpportunities.away** a naopak **pim.away** a **powerPlayOpportunities.home**:

Hide

```
cor(df$pim.home, df$powerPlayOpportunities.away, use="complete.obs")
```

```
[1] 0.5669678
```

Hide

```
cor(df$pim.away, df$powerPlayOpportunities.home, use="complete.obs")
```

```
[1] 0.5926063
```

Z vyššie uvedených korelácií medzi jednotlivými atribútmi môžeme vytvoriť určité pracovné hypotézy, ktoré

budeme v ďalších fázach projektu overovať.

Vhodné by bolo zobraziť aj úspešnosť brankárov pomocou grafov, ale nie je to zatiaľ možné, pretože atribút obsahuje aj iné/nenumerické hodnoty - NaN. Tento problém plánujeme vyriešiť pri čistení dát, po ktorom môžeme vykonať novú analýzu tohoto atribútu.

Čistenie a úprava dát

V tejto kapitole sa venujeme čisteniu a úprave dát. Niektoré úpravy sme však spravili už pred analýzou. Tieto úpravy, ako napríklad deduplikáciu dát, bolo potrebné spraviť hneď na začiatku projektu, pretože sa týkali vytvárania datasetu, ktorý v projekte používame. Dataset sme vytvorili spojením 3 tabuliek. Okrem deduplikácie dát sme tiež odstránili niektoré atribúty a iné sme naopak pridali, respektíve vytvorili (ako napríklad vytvorenie atribútu úspešnosti brankára, ktorá sa počíta na základe gólov a striel oponenta). Cieľom tejto kapitoly je, aby všetky záznamy obsahovali údaje pre všetky atribúty. Je teda potrebné odstrániť všetky N/A hodnoty. Z analýzy atribútu **type** vieme, že náš dataset obsahuje výsledky niektorých zápasov, ktoré nespadajú do súťaže NHL. Pre začiatok sa pozrieme, aké atribúty obsahujú N/A hodnoty a následne odstránime dané zápasy, ktoré nepatria do NHL (takéto zápasy majú hodnotu atribútu **type** rovnú ‘A’).

V tejto sekcii sa teda sústredíme najmä na náhradu NA hodnôt, elimináciu zbytočných záznamov (bez štatistického prínosu), normalizáciu vychýlených hodnôt identifikovaných pri prieskumnej analýze a na záver transformácii formátu niektorých atribútov (z reťazcov na numerické hodnoty) prostredníctvom tzv. one-hot encoding.

Po odstránení N/A hodnôt budeme môcť odstrániť vychýlené hodnoty atribútov **shots.home**, **pim.home**, **pim.away** a **blocked.away** využitím horného kvantilu.

Ako prvé skontrolujeme, ktoré všetky atribúty obsahujú N/A hodnoty.

Hide

```
cbind(lapply(lapply(df, is.na), sum))
```

```
                            [,1]
season                      0
type                        0
goals.away                  0
goals.home                  0
team_id.away                0
won.away                    0
shots.away                  4
hits.away                   4
pim.away                    4
powerPlayOpportunities.away 4
powerPlayGoals.away         4
faceOffWinPercentage.away   4
giveaways.away              4
takeaways.away              4
blocked.away                4
abbreviation.away           5
team_id.home                0
won.home                    0
settled_in                  0
shots.home                  4
hits.home                   4
pim.home                    4
powerPlayOpportunities.home 4
powerPlayGoals.home         4
faceOffWinPercentage.home   4
giveaways.home              4
takeaways.home              4
blocked.home                4
abbreviation.home           5
save_percentage.home        14
save_percentage.away        14
```

Ako bolo spomenuté, niektoré zápasy napatria do súťaže NHL. Niektoré atribúty môžu obsahovať N/A hodnoty práve z tohto dôvodu. Najskôr odstránime spomínané zápasy.

Hide

```
table(df$type)
```

```
   A    P    R
   5  481 6096
```

Hide

```
df = df[!(df$type == 'A')]
table(df$type)
```

```
   P    R
 481 6096
```

Ako môžeme vidieť z výpisu vyššie, dataset už neobsahuje žiade zápasy, ktoré by mali hodnotu atribútu **type** rovnú ‘A’. Teraz sa pozrieme, či to nejakým spôsobom ovplyvnilo počet N/A hodnôt ostatných atribútov.

Hide

```
cbind(lapply(lapply(df, is.na), sum))
```

```
                            [,1]
season                         0
type                           0
goals.away                     0
goals.home                     0
team_id.away                   0
won.away                       0
shots.away                     4
hits.away                      4
pim.away                       4
powerPlayOpportunities.away    4
powerPlayGoals.away            4
faceOffWinPercentage.away      4
giveaways.away                 4
takeaways.away                 4
blocked.away                   4
abbreviation.away              0
team_id.home                   0
won.home                       0
settled_in                     0
shots.home                     4
hits.home                      4
pim.home                       4
powerPlayOpportunities.home    4
powerPlayGoals.home            4
faceOffWinPercentage.home      4
giveaways.home                 4
takeaways.home                 4
blocked.home                   4
abbreviation.home              0
save_percentage.home          14
save_percentage.away          14
```

Môžeme si všimnúť, že zmizli všetky N/A hodnoty atribútov **abbrevation.away** a **abbreviation.home**. Teraz začneme postupne prechádzať ostatné atribúty obsahujúce N/A hodnoty, pričom navrhneme spôsob akým ich nahradíme.

Hide

```
df[is.na(df$shots.away)]
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <dbl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	0	*NA*	*NA*	
20172018	P	0	0	52	0	*NA*	*NA*	
20172018	P	0	0	54	0	*NA*	*NA*	
20162017	P	0	0	3	0	*NA*	*NA*	

4 rows | 1-9 of 31 columns

Z výpisu záznamov, ktoré obsahujú N/A hodnotu v atribúte **shots_away** si môžeme všimnúť, že aj ďalšie atribúty majú N/A hodnoty. Môžeme predpokladať, že všetky N/A hodnoty (okrem atribútov **save_percentage.home** a **save_percentage.away**) v našom datasete sú obsiahnuté v práve týchto 4 konkrétnych zápasoch. Vypíšeme si čísla tímov jednotlivých atribútov, ktoré obsahujú N/A hodnoty. Ak sa čísla tímov budú zhodovať, znamená to, že všetky N/A hodnoty sú obsiahnuté v spomínaných 4 zápasoch.

Hide

```
df$team_id.away[is.na(df$shots.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$hits.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$pim.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$powerPlayOpportunities.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$powerPlayGoals.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$faceOffWinPercentage.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$giveaways.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$takeaways.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$blocked.away)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$shots.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$hits.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$pim.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$powerPlayOpportunities.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$powerPlayGoals.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$faceOffWinPercentage.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$giveaways.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$takeaways.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.away[is.na(df$blocked.home)]
```

```
[1] 28 52 54  3
```

Hide

```
df$team_id.home[is.na(df$shots.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$hits.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$pim.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$powerPlayOpportunities.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$powerPlayGoals.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$faceOffWinPercentage.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$giveaways.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$takeaways.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$blocked.away)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$shots.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$hits.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$pim.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$powerPlayOpportunities.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$powerPlayGoals.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$faceOffWinPercentage.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$giveaways.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$takeaways.home)]
```

```
[1] 24 54 52  9
```

Hide

```
df$team_id.home[is.na(df$blocked.home)]
```

```
[1] 24 54 52  9
```

Z výpisov môžeme vidieť, že sa čísla tímov pre všetky atribúty obsahujúce N/A hodnoty opakujú. Z toho vyplýva, že všetky N/A hodnoty v našom datasete sú z práve týchto 4 konkrétnych zápasov. Tieto zápasy pre nás teda nemajú žiadny prínos, pretože im chýbajú hodnoty veľkého množstva atribútov. Z tohto dôvodu nebudeme N/A hodnoty nijakým spôsobom nahrádzať a záznamy o daných zápasoch z datesetu jednoducho vymažeme.

Hide

```
df = df[!(is.na(df$shots.away))]
cbind(lapply(lapply(df, is.na), sum))
```

```
                           [,1]
season                     0
type                       0
goals.away                 0
goals.home                 0
team_id.away               0
won.away                   0
shots.away                 0
hits.away                  0
pim.away                   0
powerPlayOpportunities.away 0
powerPlayGoals.away        0
faceOffWinPercentage.away  0
giveaways.away             0
takeaways.away             0
blocked.away               0
abbreviation.away          0
team_id.home               0
won.home                   0
settled_in                 0
shots.home                 0
hits.home                  0
pim.home                   0
powerPlayOpportunities.home 0
powerPlayGoals.home        0
faceOffWinPercentage.home  0
giveaways.home             0
takeaways.home             0
blocked.home               0
abbreviation.home          0
save_percentage.home       10
save_percentage.away       10
```

Ako môžeme vidieť z výpisu vyššie, po odstránení záznamov, ktoré obsahujú N/A hodnoty v atribúte

shots.away, zmizli všetky N/A hodnoty z celého datasetu okrem atribútov **save_percentage.home** a **save_percentage.away**. Teraz skontrolujeme tieto dva atribúty.

Hide

```
df[is.na(df$save_percentage.away)]
```

season <int>	t… <chr>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <dbl>	shots.away <int>	hits.away <int>	pim.a <int>
20172018	P	0	0	28	0	0	0	
20172018	P	0	0	26	0	0	0	
20172018	P	0	0	54	0	0	0	
20172018	P	0	0	15	0	0	0	
20172018	P	0	0	5	0	0	0	
20162017	P	0	0	20	0	0	0	
20162017	P	0	0	10	0	0	0	
20162017	P	0	0	18	0	0	0	
20162017	P	0	0	18	0	0	0	
20162017	P	0	0	18	0	0	0	

1-10 of 10 rows | 1-9 of 31 columns

Z výpisu vyššie môžeme vidieť, že 10 N/A hodnôt v **save_percentage.home** je pri rovnakých zápasoch ako pri atribúte **save_percentage.away**. Tiež si môžeme všimnúť, že tieto hodnoty vznikli z dôvodu delenia nulou (pri výpočte save_percentage vyhádzame z rozdielu počtu striel a počtu gólov, pričom výsledok ešte delíme počtom striel, t. j. v prípade, že počet striel je 0, výsledkom je NaN, pretože nulou nedelíme). Tiež si môžeme všimnúť, že všetky ostatné atribúty v týchto zápasoch majú hodnotu 0. Aj keby sme nahradili save_percentage 100-percentnou úspešnosťou, nemali by nás tieto zápasy žiadnu výpovednú hodnotu. Tieto zápasy teda môžeme odstrániť.

Hide

```
df = df[!(is.na(df$save_percentage.away))]
```

Po odstránení spomínaných zápasov overíme, že dataset už neobsahuje žiadne N/A hodnoty.

Hide

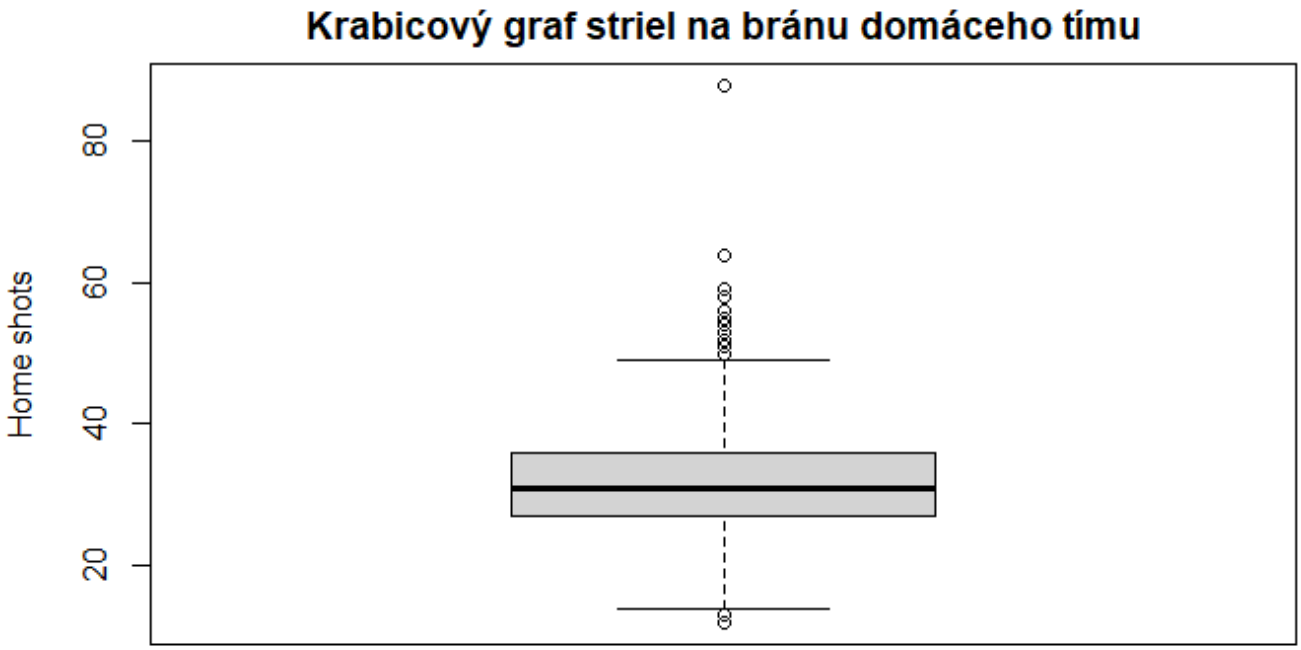
```
cbind(lapply(lapply(df, is.na), sum))
```

```
                            [,1]
season                      0
type                        0
goals.away                  0
goals.home                  0
team_id.away                0
won.away                    0
shots.away                  0
hits.away                   0
pim.away                    0
powerPlayOpportunities.away 0
powerPlayGoals.away         0
faceOffWinPercentage.away   0
giveaways.away              0
takeaways.away              0
blocked.away                0
abbreviation.away           0
team_id.home                0
won.home                    0
settled_in                  0
shots.home                  0
hits.home                   0
pim.home                    0
powerPlayOpportunities.home 0
powerPlayGoals.home         0
faceOffWinPercentage.home   0
giveaways.home              0
takeaways.home              0
blocked.home                0
abbreviation.home           0
save_percentage.home        0
save_percentage.away        0
```

Ako môžeme vidieť z výpisu vyššie, dataset už neobsahuje žiadne N/A hodnoty. Môžeme teda prejsť na odstraňovanie vychýlených atribútov **shots.home**, **pim.home**, **pim.away** a **blocked.away**.

Hide

```
boxplot(df$shots.home, ylab="Home shots", xlab="", main="Krabicový graf striel na brá
nu domáceho tímu")
```



Vychýlenú hodnotu nahradíme 75% kvantilom počtu striel domácich tímov v zápasoch, ktoré sa dohrali v predĺžení.

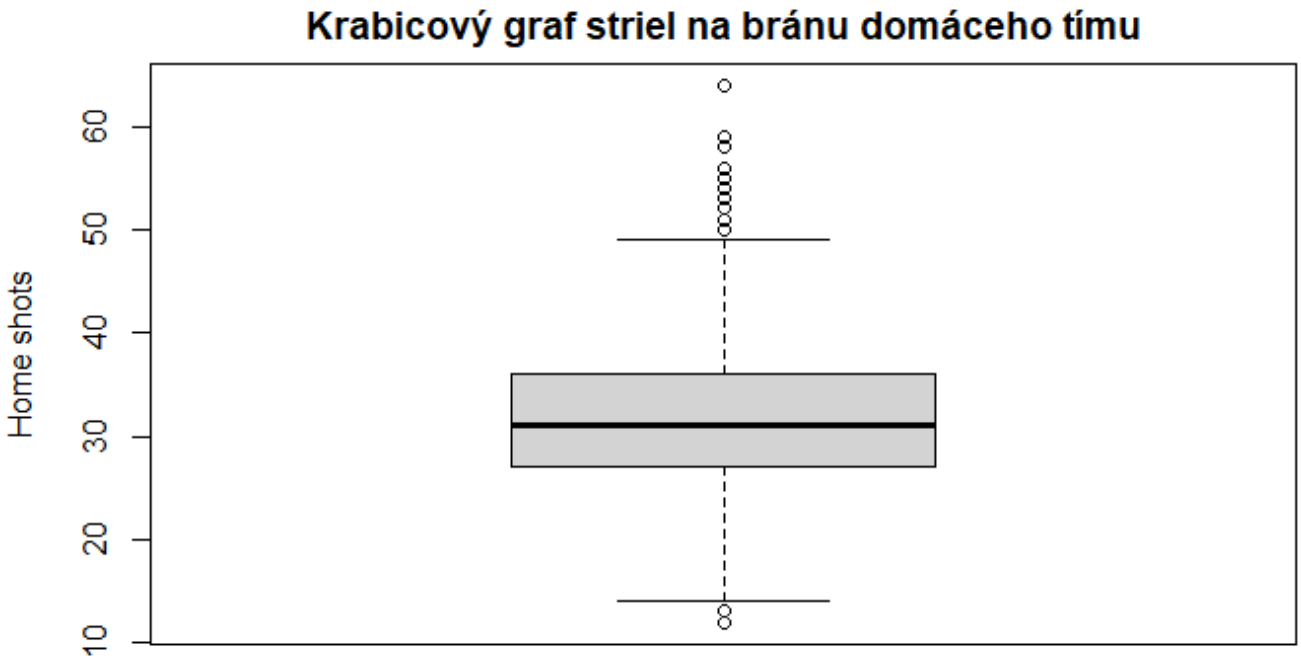
Hide

```
df$shots.home[df$shots.home > 70] = quantile(df$shots.home[df$settled_in == 'OT'], 0.
75)
```

Teraz skontrolujeme, či sa vychýlená hodnota nahradila.

Hide

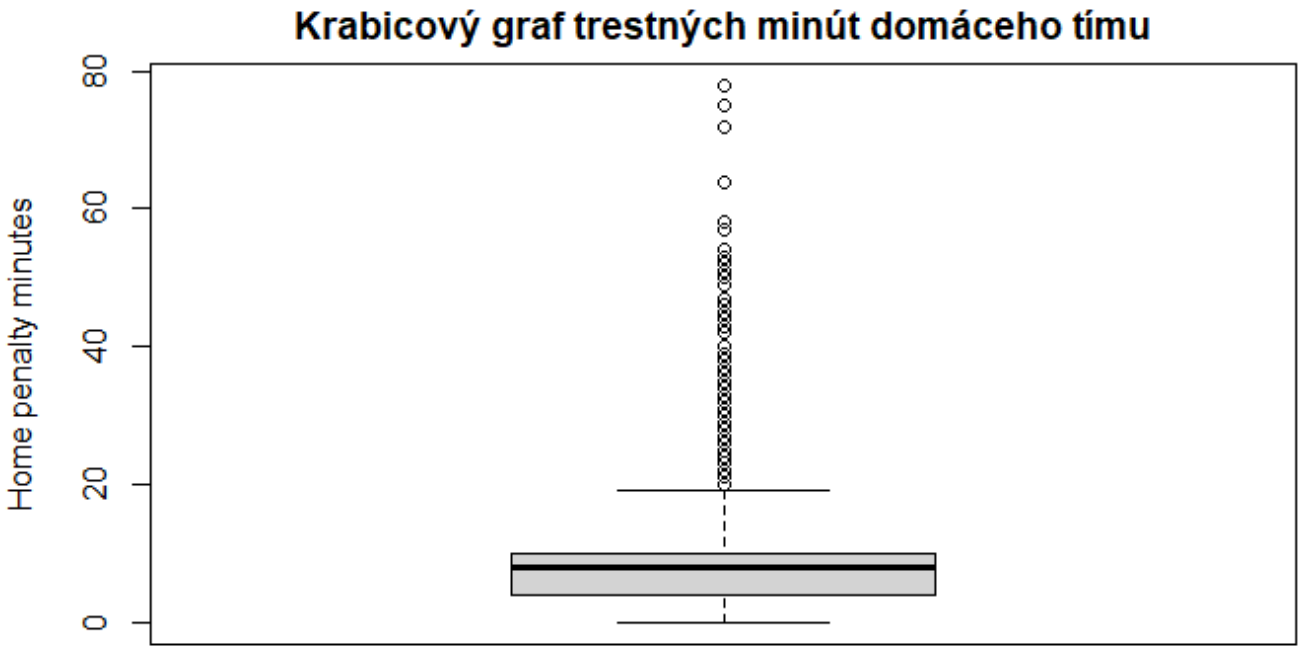
```
boxplot(df$shots.home, ylab="Home shots", xlab="", main="Krabicový graf striel na brá
nu domáceho tímu")
```



Na krabicovom grafe vyššie môžeme vidieť, že sme vychýlenú hodnotu zrazili. Podobným spôsobom nahradíme aj vychýlené hodnoty vo zvyšných atribútoch.

Hide

```
boxplot(df$pim.home, ylab="Home penalty minutes", xlab="", main="Krabicový graf trest
ných minút domáceho tímu")
```



V prípade atribútu **pim.home** odstránime hodnoty väčšie ako 60.

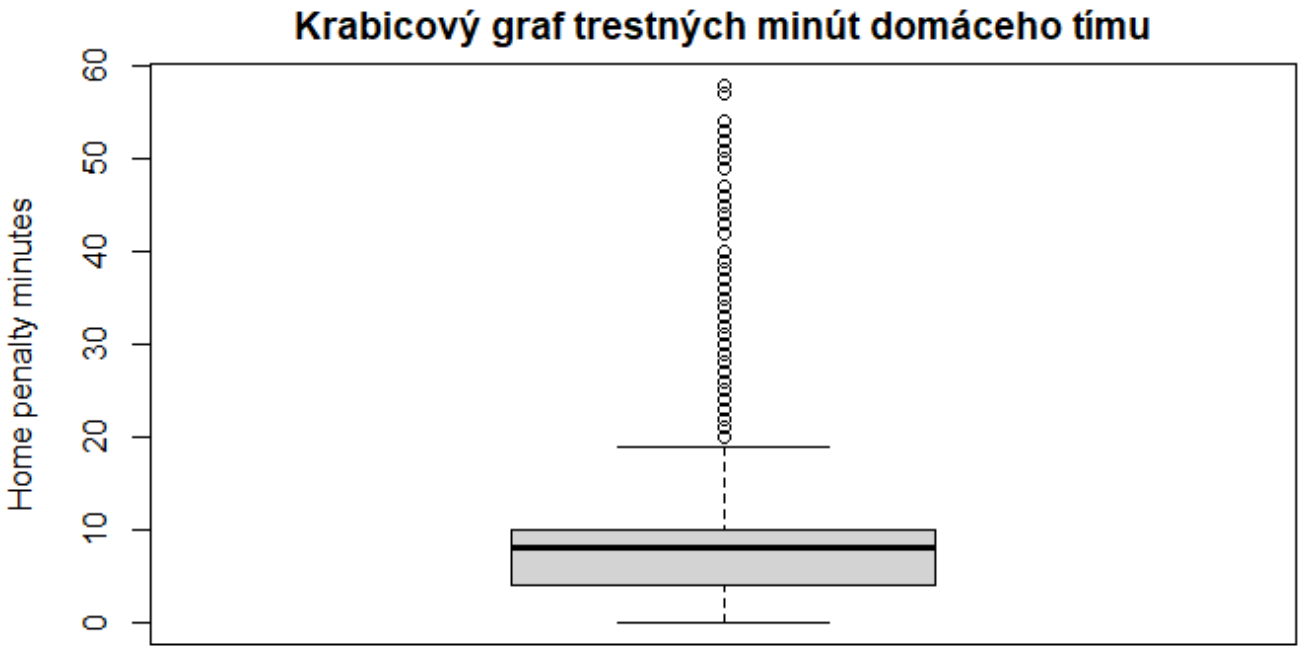
Hide

```
df$pim.home[df$pim.home > 60] = quantile(df$pim.home, 0.75)
```

Odstránenie hodnôt skontrolujeme.

Hide

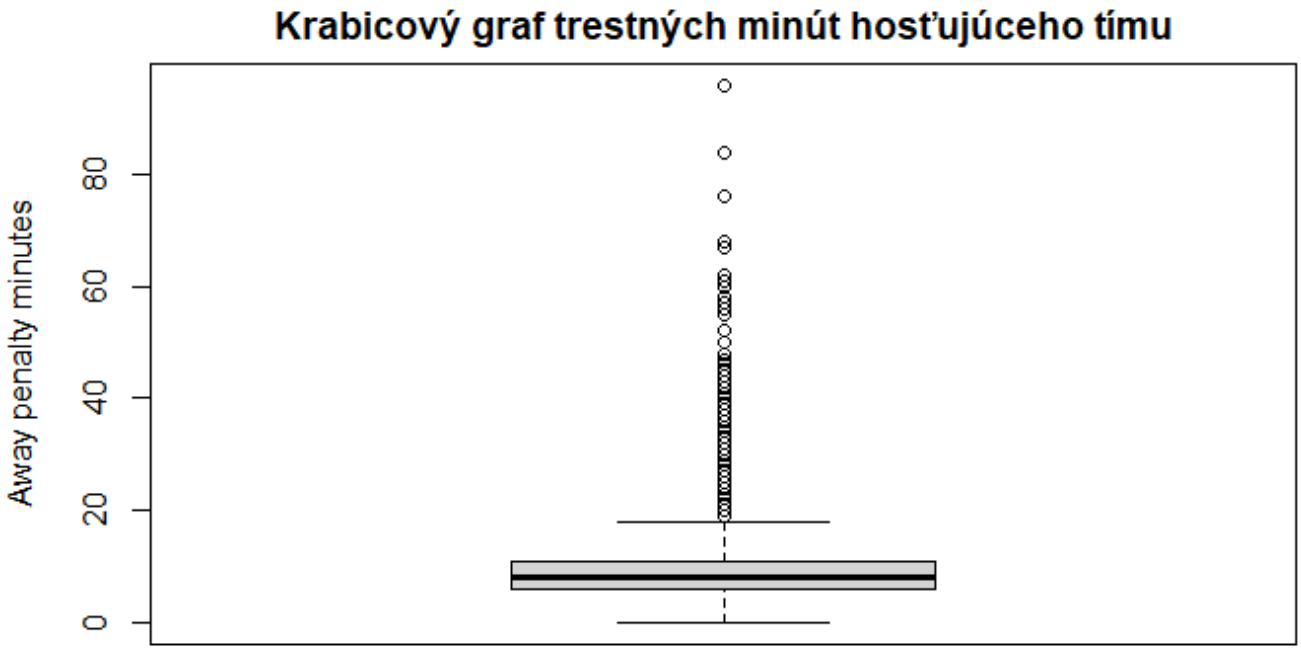
```
boxplot(df$pim.home, ylab="Home penalty minutes", xlab="", main="Krabicový graf trest
ných minút domáceho tímu")
```



Môžeme vidieť, že atribút **pim.home** už neobsahuje hodnoty väčšie ako 60.

Hide

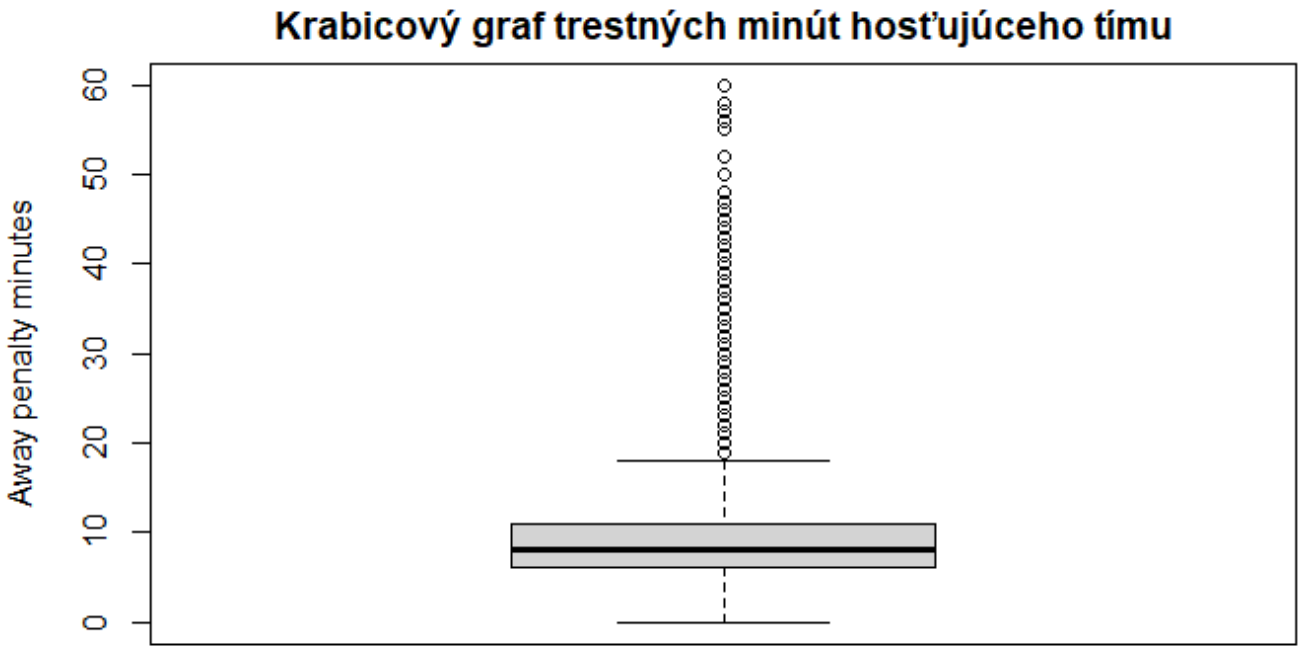
```
boxplot(df$pim.away, ylab="Away penalty minutes", xlab="", main="Krabicový graf trest
ných minút hosťujúceho tímu")
```



Podobne ako pri **pim.home** odstránime aj vychýlené hodnoty atribútu **pim.away** (tiež hodnoty väčšie ako 60) a hodnoty tohto atribútu následne skontrolujeme.

Hide

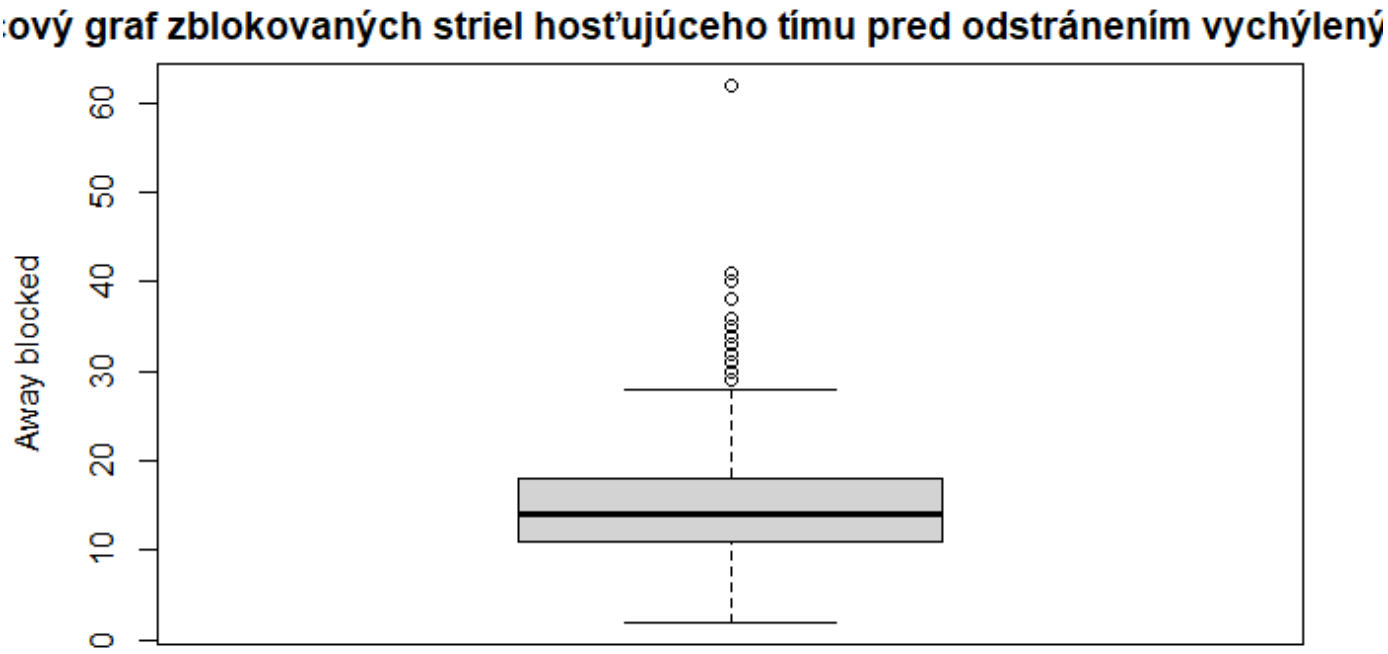
```
df$pim.away[df$pim.away > 60] = quantile(df$pim.away, 0.75)
boxplot(df$pim.away, ylab="Away penalty minutes", xlab="", main="Krabicový graf trest
ných minút hosťujúceho tímu")
```



Posledným atribútom s vychýlenými hodnotami je atribút **blocked.away**. Aj v tomto prípade odstránime hodnoty väčšie ako 60 použitím horného kvantilu. Nakoniec hodnoty opät skontrolujeme pomocou krabicového grafu.

Hide

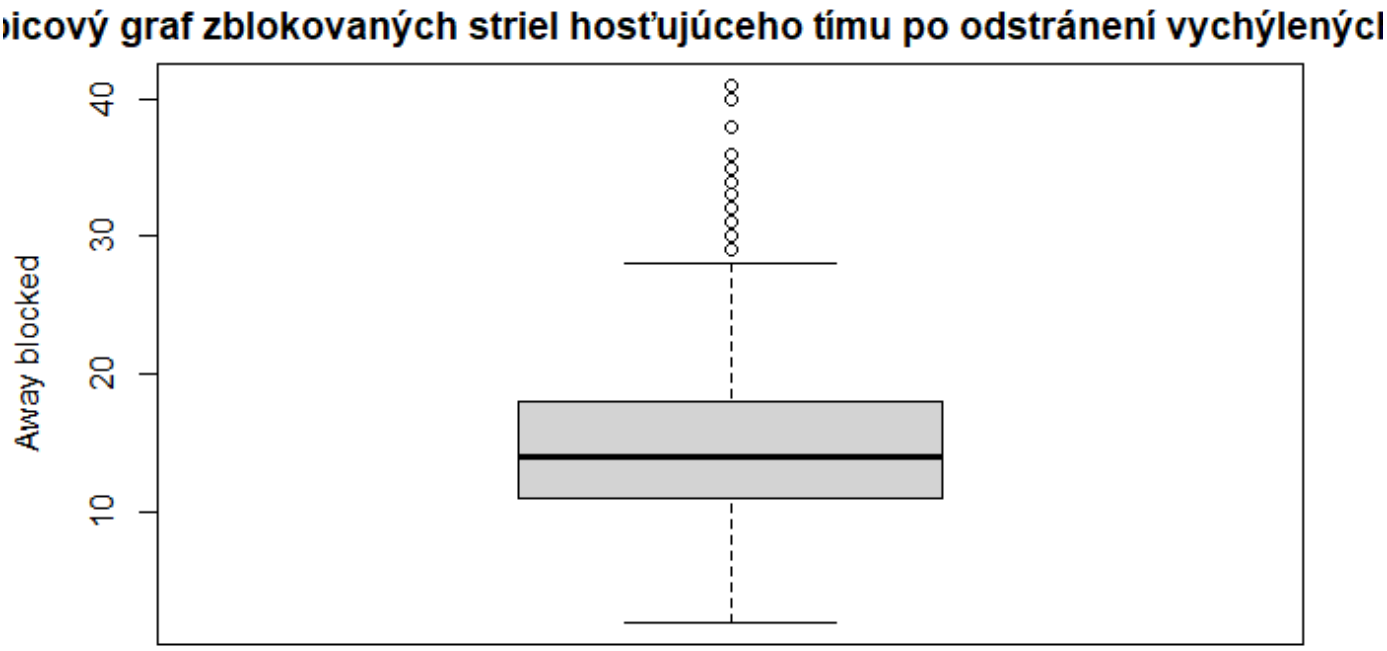
```
boxplot(df$blocked.away, ylab="Away blocked", xlab="", main="Krabicový graf zblokovan
ých striel hosťujúceho tímu pred odstránením vychýlených hodnôt")
```



Hide

```
df$blocked.away[df$blocked.away > 60] = quantile(df$blocked.away, 0.75)

boxplot(df$blocked.away, ylab="Away blocked", xlab="", main="Krabicový graf zblokovan
ých striel hosťujúceho tímu po odstránení vychýlených hodnôt")
```



V tomto momente náš dataset už neobsahuje žiadne vychýlené hodnoty a ani N/A hodnoty.

Príprava dát pre tréningový a trénovací dataset

V rámci čistenia dát je potrebné pripraviť dataset, aby bolo možné použiť ho pre strojové učenie, zhlukovú analýzu a Bayesovú štatistiku, pričom treba dataset rozdeliť na tréningový a testovací. Aby bolo možné ho použiť pre tieto účely, je potrebné, aby mal numerické dáta. Náš dataset obsahuje 4 nenumerické atribúty: **type**, **settled_in**, **abbreviation.home** a **abbreviation.away**. Atribúty abbreviation nie sú zo štatistického hľadiska dôležité, sú to len skratky jednotlivých tímov. Atribúty **type** a **settled_in** však hovoria o type zápasu (či bol zápas v rámci regulárnej sezóny alebo v rámci vyraďovacej časti) a či bol zápas ukončený v riadnom hracom čase alebo v predĺžení. Oba atribúty teda nadobúdajú práve dve hodnoty a môžeme ich zmeniť na numerické (0 alebo 1). Ako prvé zmeníme atribúty **type**:

Hide

```
unique(df$type)
```

```
[1] "R" "P"
```

Ako môžeme vidieť, atribút **type** obsahuje dve hodnoty - “R”, ak bol zápas odohraný v rámci regulárnej sezóny, alebo “P”, ak bol zápas odohraný v rámci vyraďovacej časti. Tento atribút môžeme nahradiť za **isPlayOff**, pričom tento atribút bude obsahovať hodnotu 0, ak bol zápas odohraný v rámci sezóny (teda pre hodnoty “R”) a 1, ak bol odohraný v rámci vyraďovacej časti (teda pre hodnoty “P”).

Hide

```
df$isPlayOff <- ifelse(df$type == "P", 1, 0)
```

Po nahradení môžeme skontrolovať, či boli hodnoty dosadené správne.

Hide

```
unique(df$isPlayOff[df$type == "R"])
```

```
[1] 0
```

Hide

```
unique(df$isPlayOff[df$type == "P"])
```

```
[1] 1
```

Môžeme vidieť, že pre hodnoty “R” sme dosadili do atribútu **isPlayOff** hodnoty 0, a pre “P” hodnoty 1. Teraz môžeme atribút **type** odstrániť.

Hide

```
df <- subset(df, select=-c(type))
df
```

season <int>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <dbl>	shots.away <int>	hits.away <int>	pim.aw… <dbl>
20162017	4	7	4	0	27	30	6
20172018	4	3	24	1	34	16	6
20152016	4	1	21	1	29	17	9
20152016	1	2	52	0	21	21	10
20172018	1	2	20	0	23	20	19
20162017	4	1	15	1	39	19	8
20152016	1	2	10	0	26	24	19
20172018	1	4	23	0	20	11	27
20172018	0	2	29	0	34	12	8
20152016	3	2	22	1	36	31	8

1-10 of 6,563 rows | 1-8 of 31 columns Previous **1** 2 3 4 5 6 … 100 Next

Podobným spôsobom odstránime aj atribút **settled_in**. Najskôr skontrolujeme, aké hodnoty nadobúda.

Hide

```
unique(df$settled_in)
```

```
[1] "REG" "OT"
```

Na základe nadobúdaných hodnôt môžeme opäť vytvoriť nový atribút **isOverTime**, pričom tento atribút bude mať hodnotu 1 pre “OT” v atribúte **settled_in** a 0 pre “REG”.

Hide

```
df$isOverTime <- ifelse(df$settled_in == "OT", 1, 0)
```

Opäť skontrolujeme dosadenie hodnôt do nového atribútu.

Hide

```
unique(df$isOverTime[df$settled_in == "REG"])
```

```
[1] 0
```

Hide

```
unique(df$isOverTime[df$settled_in == "OT"])
```

```
[1] 1
```

Hodnoty boli dosadené správne a teda môžeme atribút **settled_in** vymazať.

Hide

```
df <- subset(df, select=-c(settled_in))
df
```

season <int>	goals.away <int>	goals.home <int>	team_id.away <int>	won.a… <dbl>	shots.away <int>	hits.away <int>	pim.aw… <dbl>
20162017	4	7	4	0	27	30	6
20172018	4	3	24	1	34	16	6
20152016	4	1	21	1	29	17	9
20152016	1	2	52	0	21	21	10
20172018	1	2	20	0	23	20	19
20162017	4	1	15	1	39	19	8
20152016	1	2	10	0	26	24	19
20172018	1	4	23	0	20	11	27
20172018	0	2	29	0	34	12	8
20152016	3	2	22	1	36	31	8

1-10 of 6,563 rows | 1-8 of 31 columns Previous **1** 2 3 4 5 6 … 100 Next

Fázu čistenia dát môžeme ukončiť vytvorením trénovacieho a testovacieho datasetu. Rozhodli sme rozdeliť dataset na základe sezón, pričom trénovací dataset budú tvoriť 4 sezóny a testovací dataset jedna (posledná) sezóna. Percentuálne tento pomer vychádza na približne 80%:20%.

Hide

```
df_train = df[!(df$season == "20192020")]
df_test = df[df$season == "20192020"]
nrow(df_train)
```

```
[1] 5351
```

Hide

```
nrow(df_test)
```

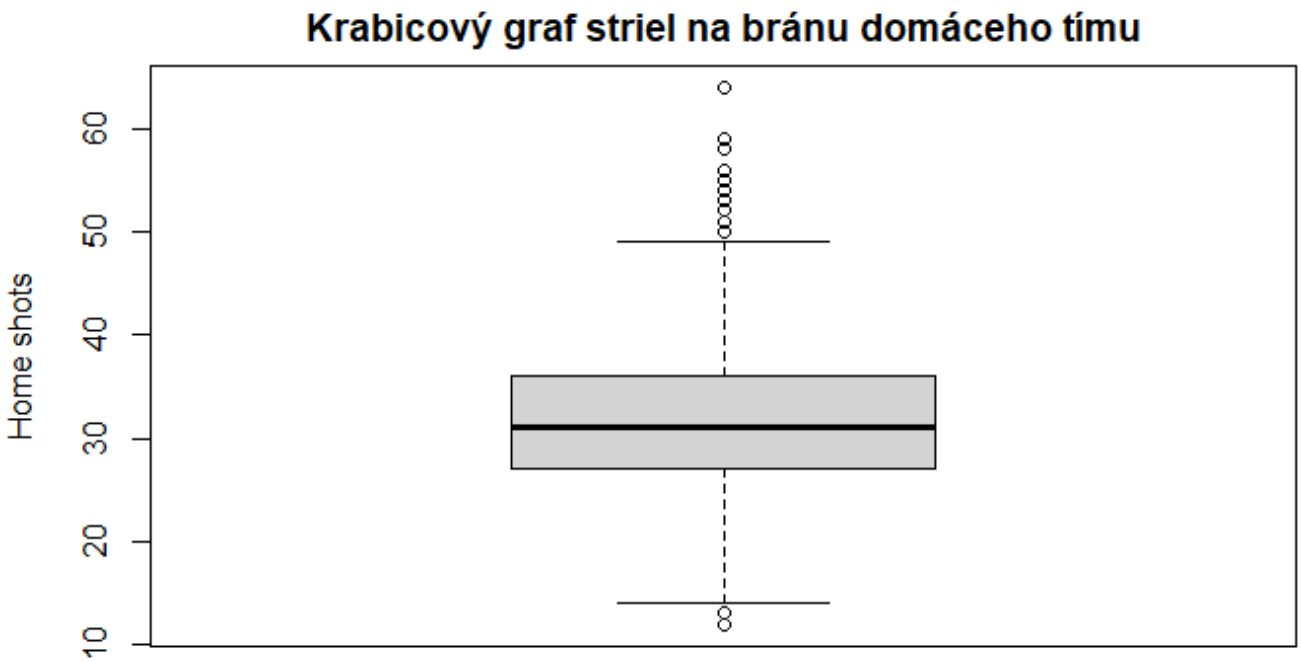
```
[1] 1212
```

Prieskumná analýza po čistení dát

Keďže sme pri čistení dát zmenili dva atribúty na numerické, môžeme opäť vykonať párovú analýzu, pretože je možné, že medzi atribútmi vznikli nové korelácie. Podobne je vhodné vykonať prieskumnú analýzu zmenených atribútov.

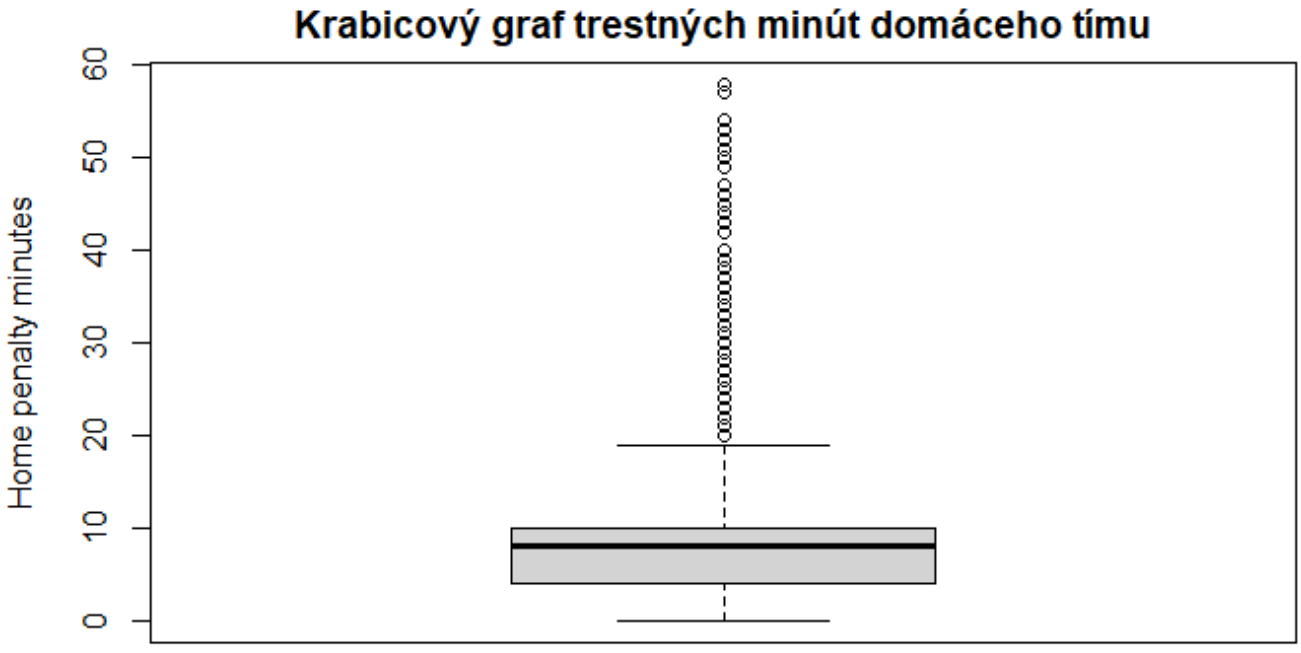
Hide

```
boxplot(df$shots.home, ylab="Home shots", xlab="", main="Krabicový graf striel na brá
nu domáceho tímu")
```



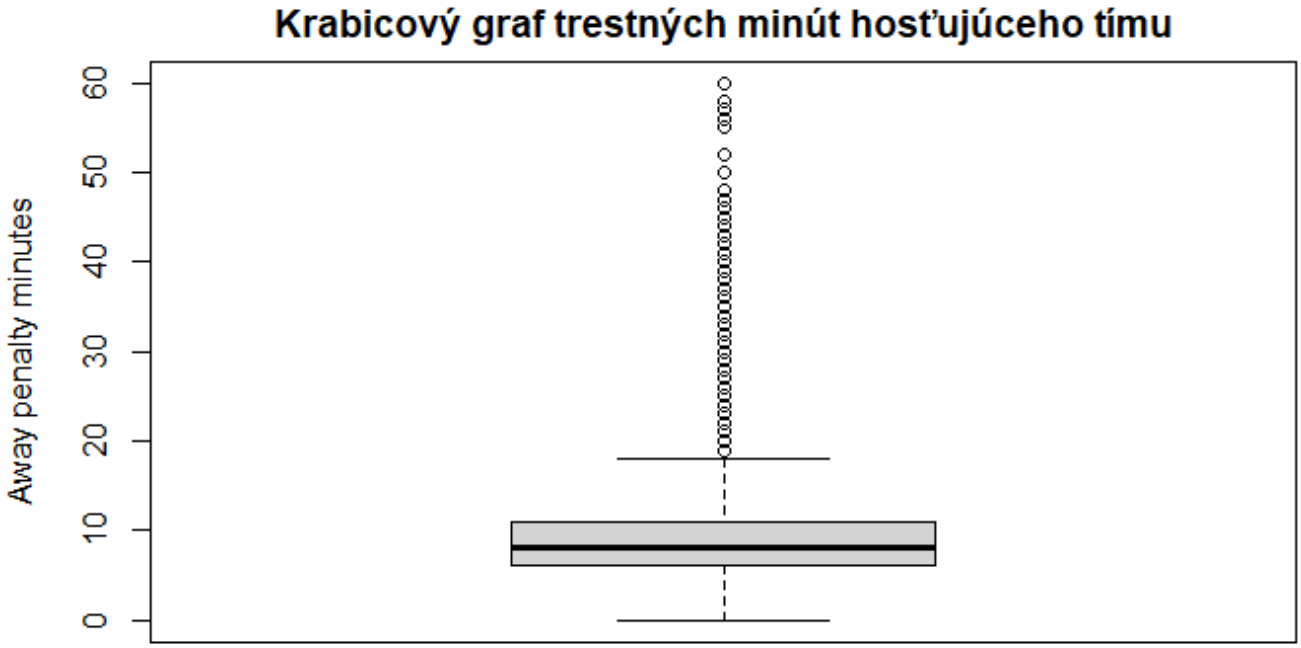
Hide

```
boxplot(df$pim.home, ylab="Home penalty minutes", xlab="", main="Krabicový graf trest
ných minút domáceho tímu")
```



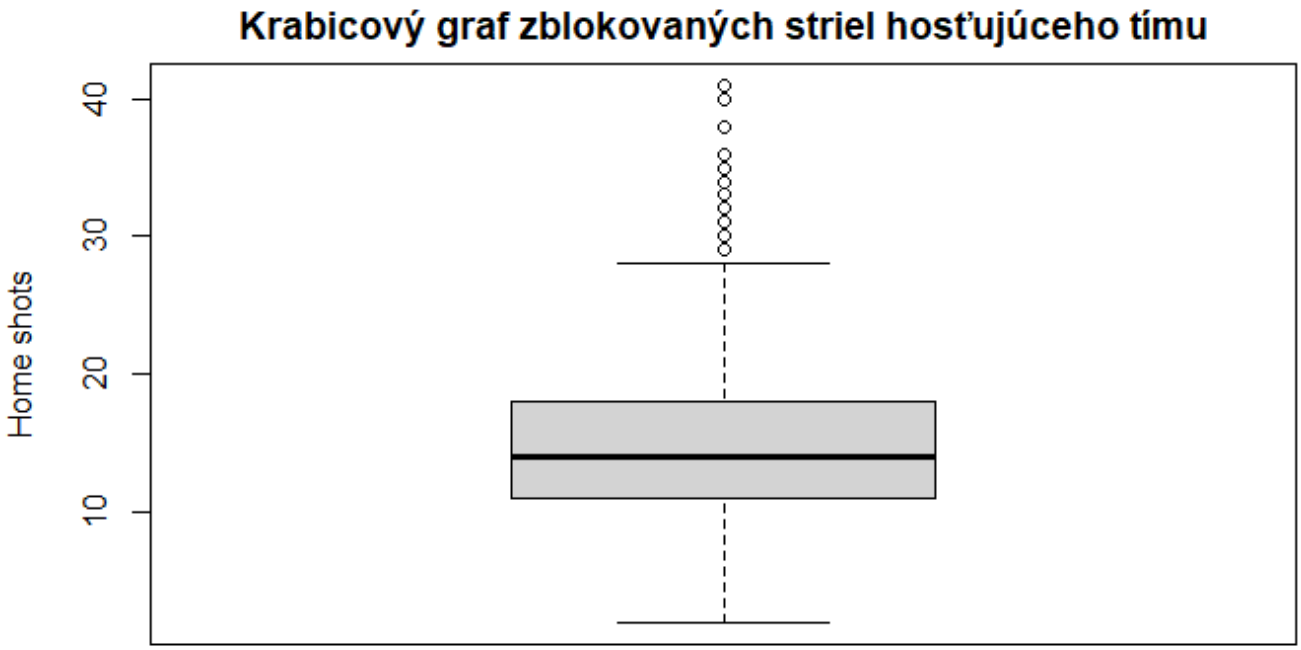
Hide

```
boxplot(df$pim.away, ylab="Away penalty minutes", xlab="", main="Krabicový graf trest
ných minút hosťujúceho tímu")
```



Hide

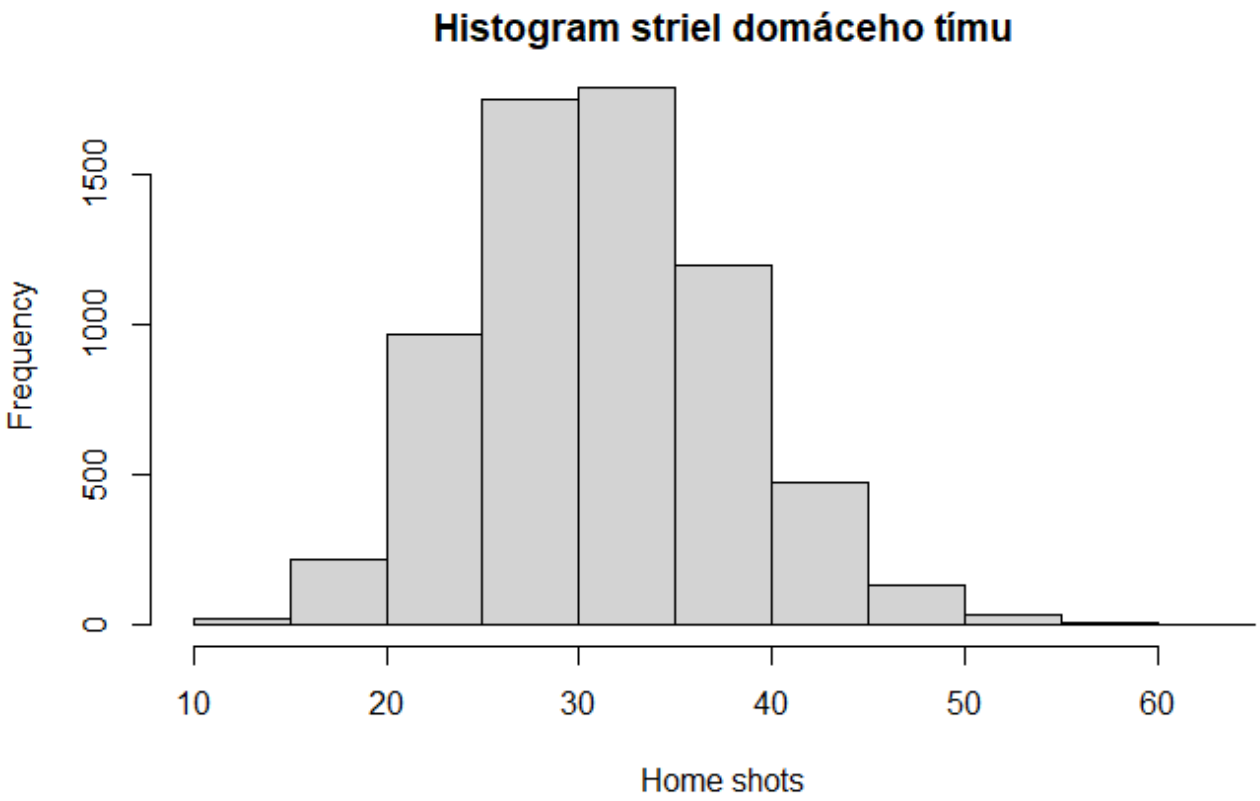
```
boxplot(df$blocked.away, ylab="Home shots", xlab="", main="Krabicový graf zblokovanýc
h striel hosťujúceho tímu")
```



Z krabicových grafov jednotlivých atribútov môžeme vidieť, že žiadny z atribútov neobsahuje extrémne vychýlené hodnoty. Pre zobrazenie týchto hodnôt môžeme vytvoriť aj histogramy, čím skontrolujeme ako sa zmenilo rozdelenie atribútov.

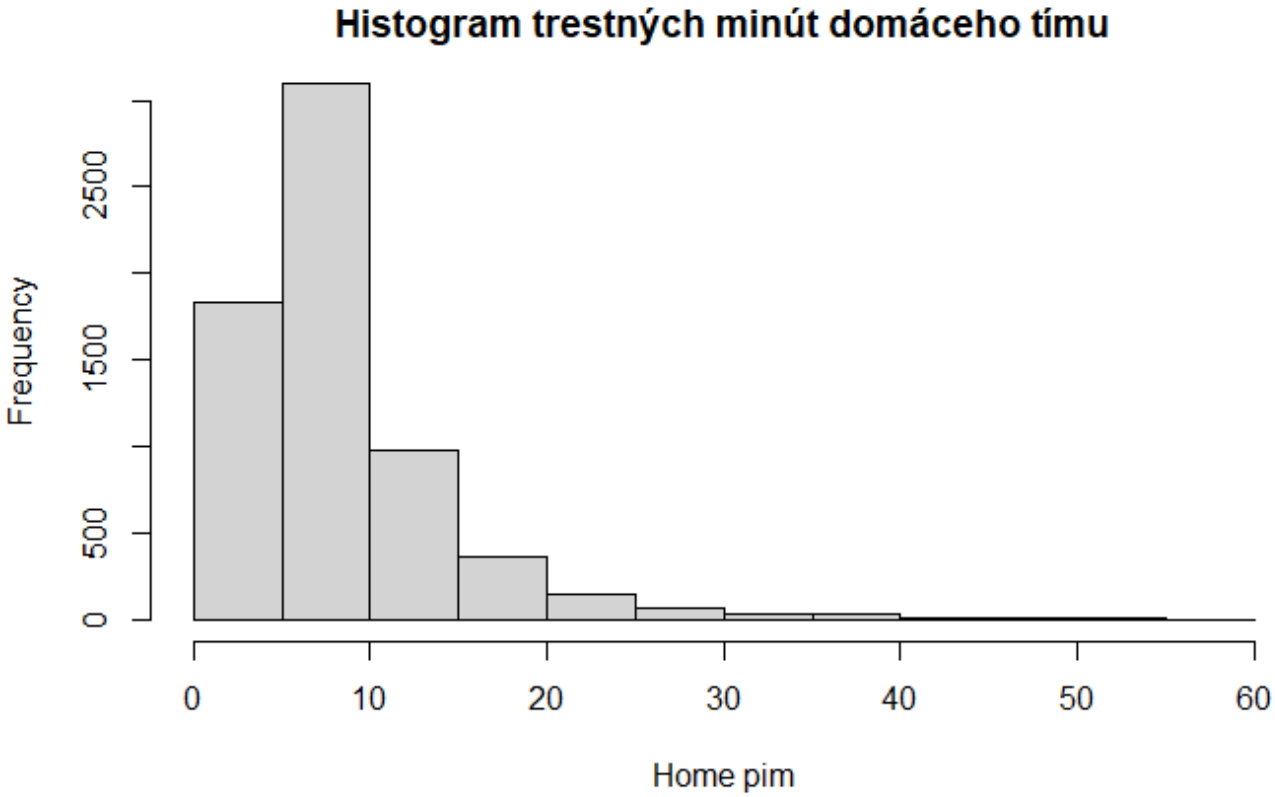
Hide

```
hist(df$shots.home, xlab="Home shots", main="Histogram striel domáceho tímu")
```



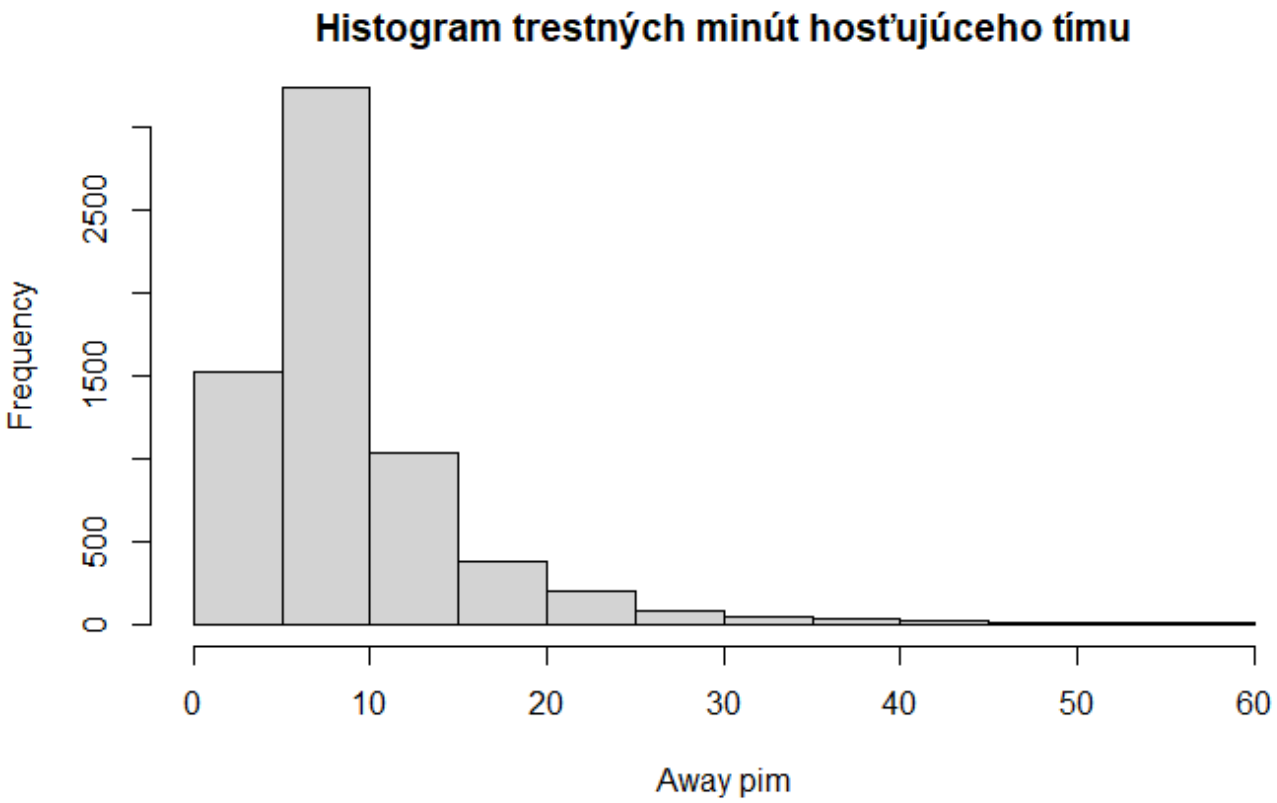
Hide

```
hist(df$pim.home, xlab="Home pim", main="Histogram trestných minút domáceho tímu")
```



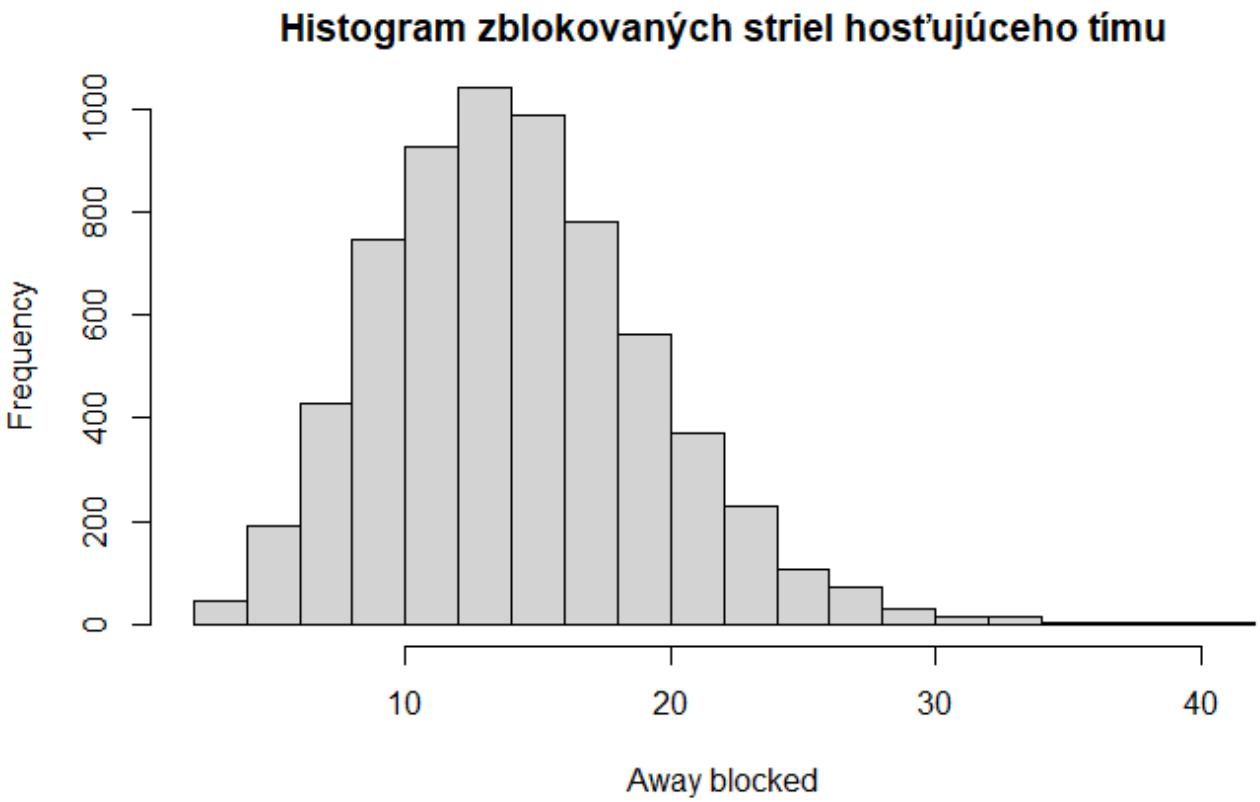
Hide

```
hist(df$pim.away,  xlab="Away pim", main="Histogram trestných minút hosťujúceho tím
u")
```



Hide

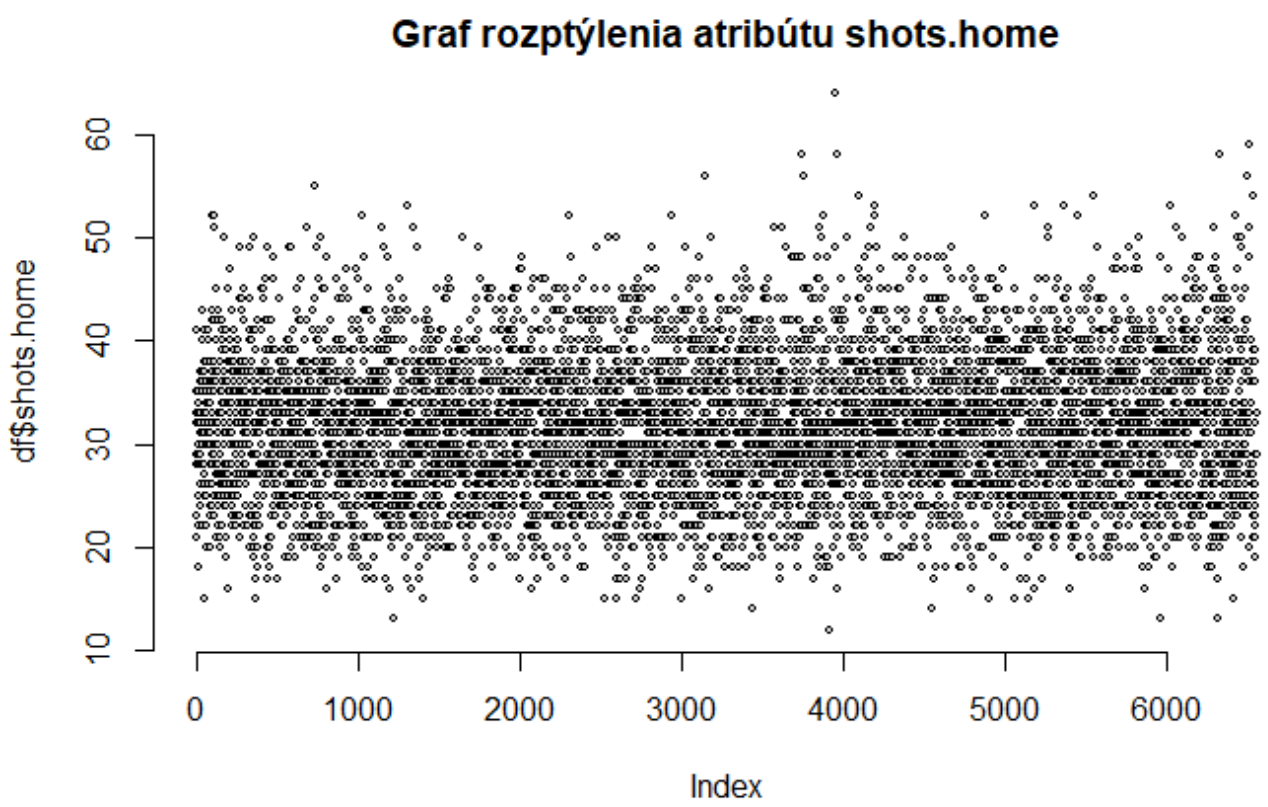
```
hist(df$blocked.away, xlab="Away blocked", main="Histogram zblokovaných striel hosťuj
úceho tímu")
```



V prípade **shots.home** tento atribút už začína pripomínať normálne rozdelenie. V prípade ostatných atribútov bude potrebné hodnoty znormalizovať.

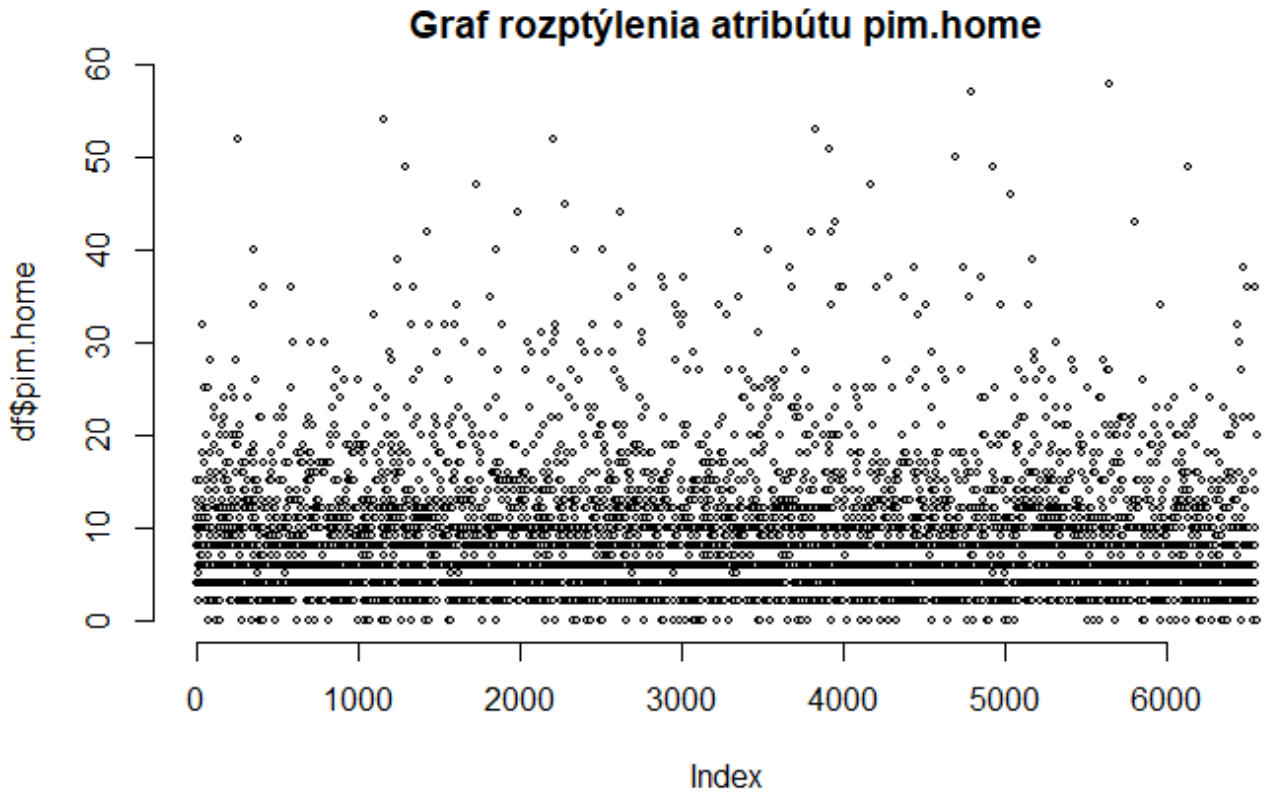
Hide

```
plot(x=df$shots.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = F
ALSE,main = "Graf rozptýlenia atribútu shots.home")
```



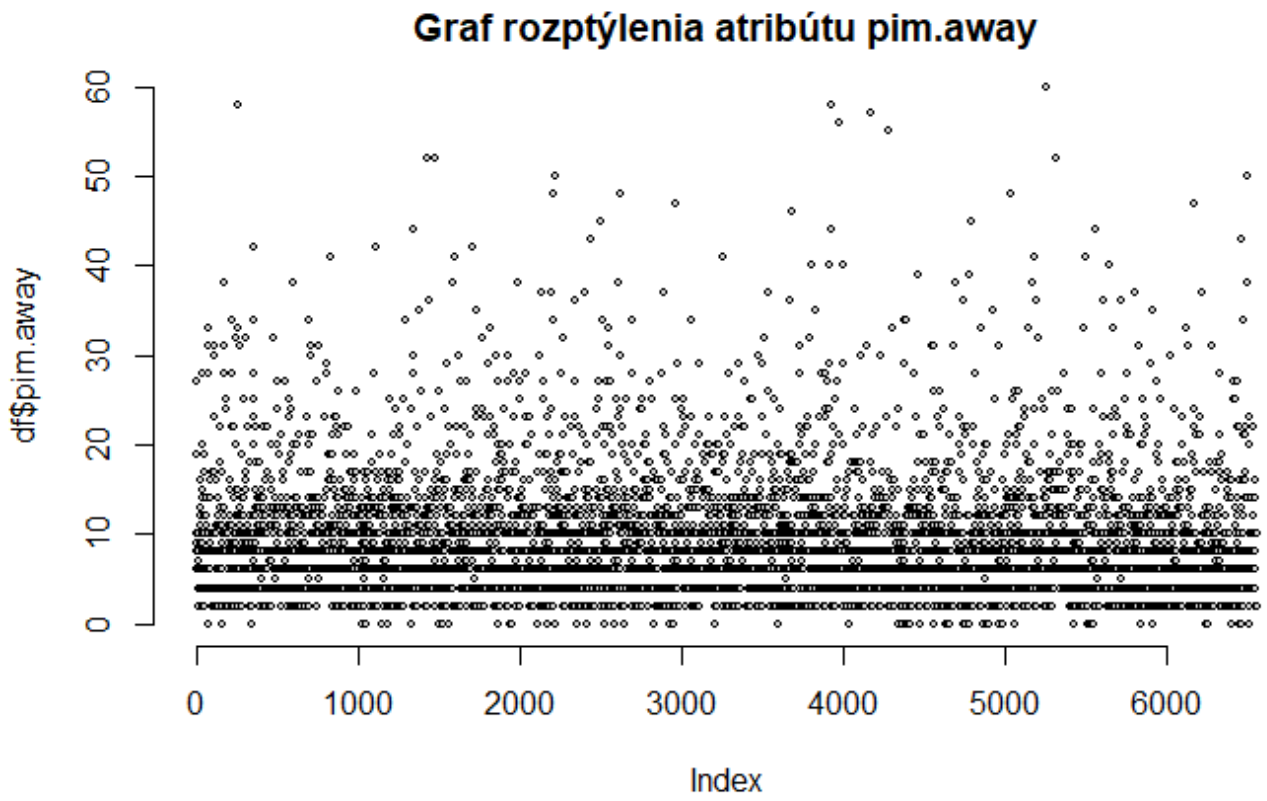
Hide

```
plot(df$pim.home, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FALS
E, main="Graf rozptýlenia atribútu pim.home")
```



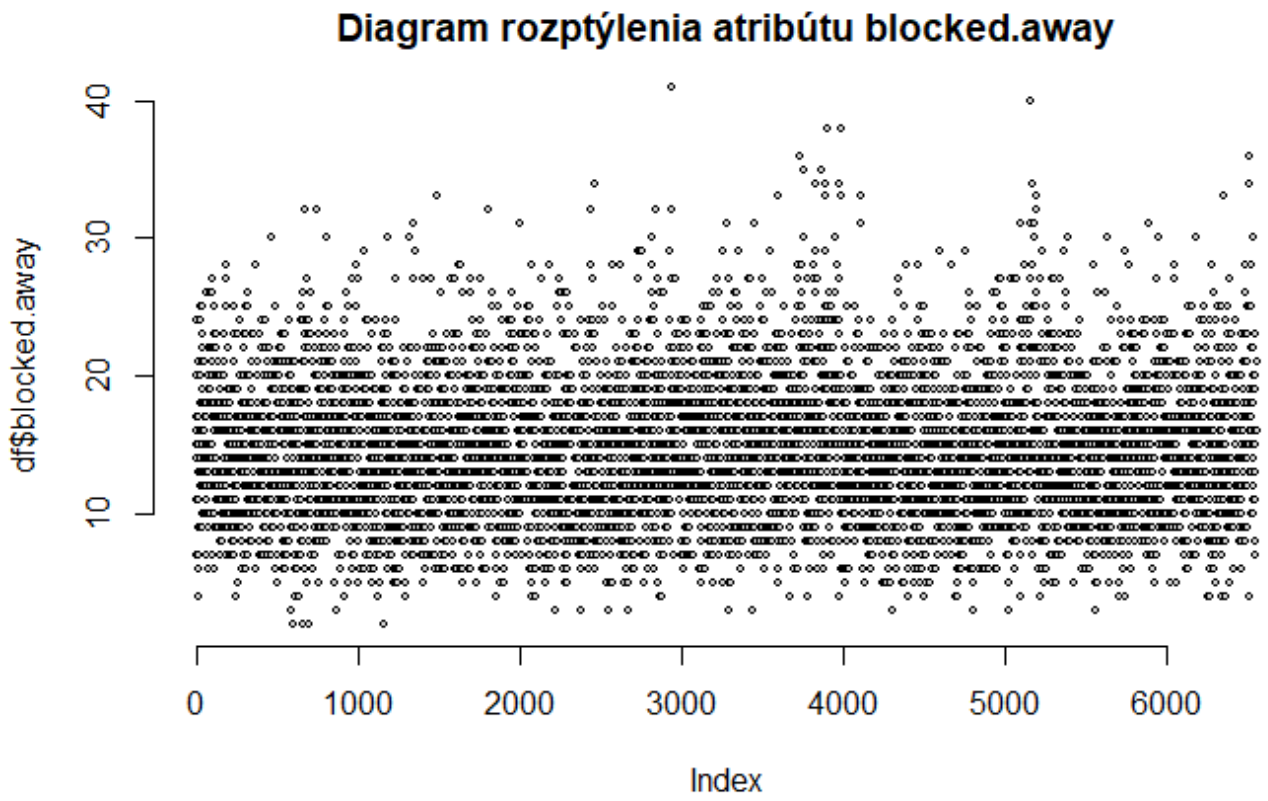
Hide

```
plot(df$pim.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame = FALS
E, main="Graf rozptýlenia atribútu pim.away")
```



Hide

```
plot(x=df$blocked.away, pch = 21, bg = "lightgray", col = "black", cex = 0.5, frame =
FALSE,main = "Diagram rozptýlenia atribútu blocked.away")
```

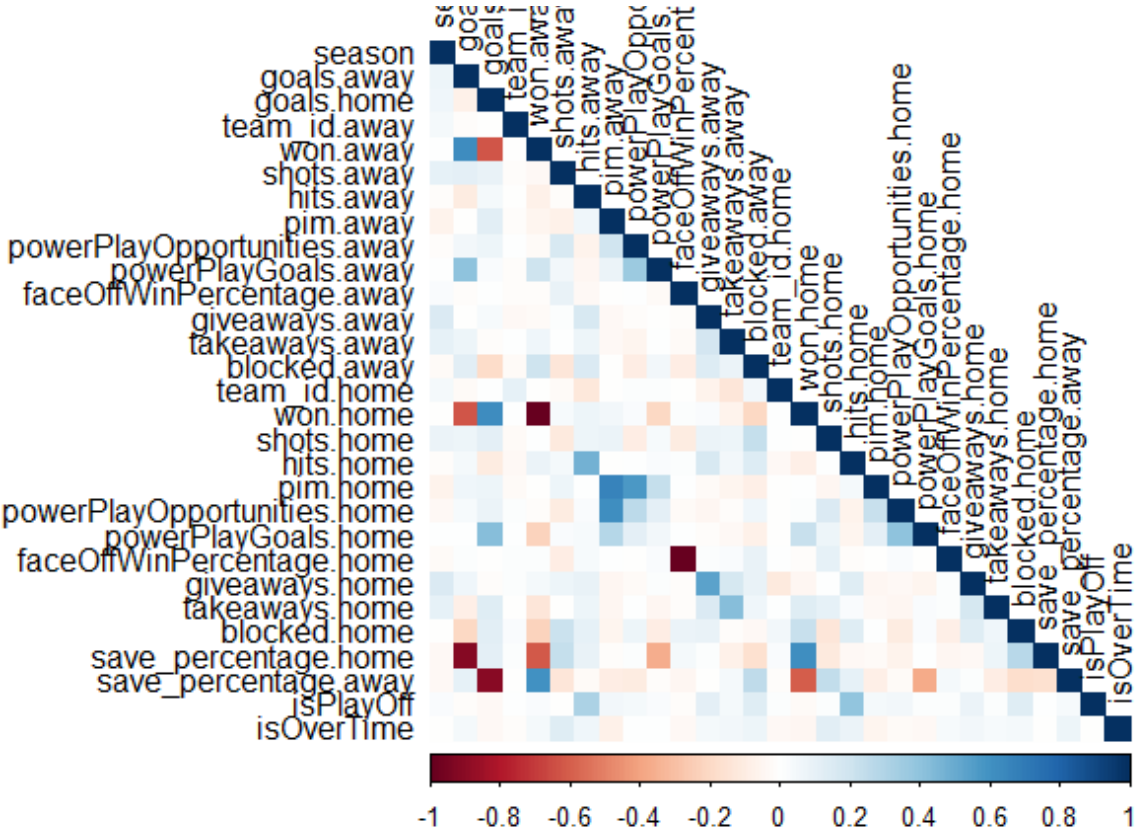


Na grafoch rozptýlenia jednotlivých atribútov môžeme stále pozorovať niektoré vychýlené hodnoty, o ktoré však z hľadiska vytvorenia trénovacieho datasetu nechceme prísť.

Párová analýza po čistení dát

Hide

```
df_numeric <- subset(df, select=-c(abbreviation.away, abbreviation.home))
corrplot(cor(df_numeric, use="complete.obs"), type="lower", method="color", tl.col="b
lack")
```



Ako môžeme vidieť na korelačnej matici, ani jeden z nových atribútov **isPlayOff** alebo **isOverTime** nemá vysokú koreláciu s ostatnými atribútmi. Korelácie týchto dvoch atribútov sú pre nás teda zanedbateľné.

Štatistické učenie, zhlukové analýzy a nachádzanie vnútorných vzorcov v dátach

Hypotézy

Počas fázy prieskumnej analýzy sme si stanovili 5 pracovných hypotéz, ktoré môžu byť vzhľadom na dáta zaujímavé. Tieto hypotézy sú nasledovné:

1. Tím, ktorý v zápase strelí viac ako 3 góly (vrátane), pravdepodobne vyhrá.
2. Čím je väčší súhrnný počet striel na bránu, tým viac gólov padne v zápase.
3. Čím je v zápase viac hitov, tým je zápas ostrejší a tým pádom rozhodcovia udelia tímom súhrnne viac trestných minút.
4. Čím má brankár vyššiu úspešnosť zákrokov, tým s väčšou pravdepodobnosťou jeho tím vyhrá.
5. Tím, ktorý hrá v domácom prostredí strelí viac gólov v zápase ako hosťujúci tím.

Vzťahy medzi atribútmi obshiahnutými v hypotézach bolo možné vidieť už pri prieskumnej analýze jednotlivých atribútov (EDA), ako aj pri párovej analýze (MEDA). Keďže závislosti existujú, predpokladáme, že väčšina hypotéz bude postavená dobre. Tieto hypotézy budú bližšie riešené vo fázi o štatistickom učení. V prvom rade však bude potrebné dáta vyčistit a upraviť takým spôsobom, aby boli použiteľné pre dodatočnú analýzu prostredníctvom metód štatistického učenia. Niektoré atribúty bude taktiež potrebné normalizovať - napr. logaritmom, odmocninou.

Hypotéza 1.

Tím, ktorý v zápase strelí viac ako 3 góly (vrátane), pravdepodobne vyhrá.

Atribúty pre túto hypotézu nie sú spojité. To znamená, že nemôžeme mať polovicu góla, alebo inú desatinnú hodnotu. **Atribúty:** goals.home, goals.away, won.home, won.away

Hide

```
par(mfrow=c(2,1))
par(mar=c(0,5,3,3))
hist(df$goals.home[df$won.home == 1], main="Histogram počtu gólov podľa výsledku zápa
su (home)" , ylab="Win", xlab="", ylim=c(0,1600), xlim=c(0,11) , xaxt="n", las=1 , co
l="slateblue1", breaks=8)
par(mar=c(5,5,0,3))
```

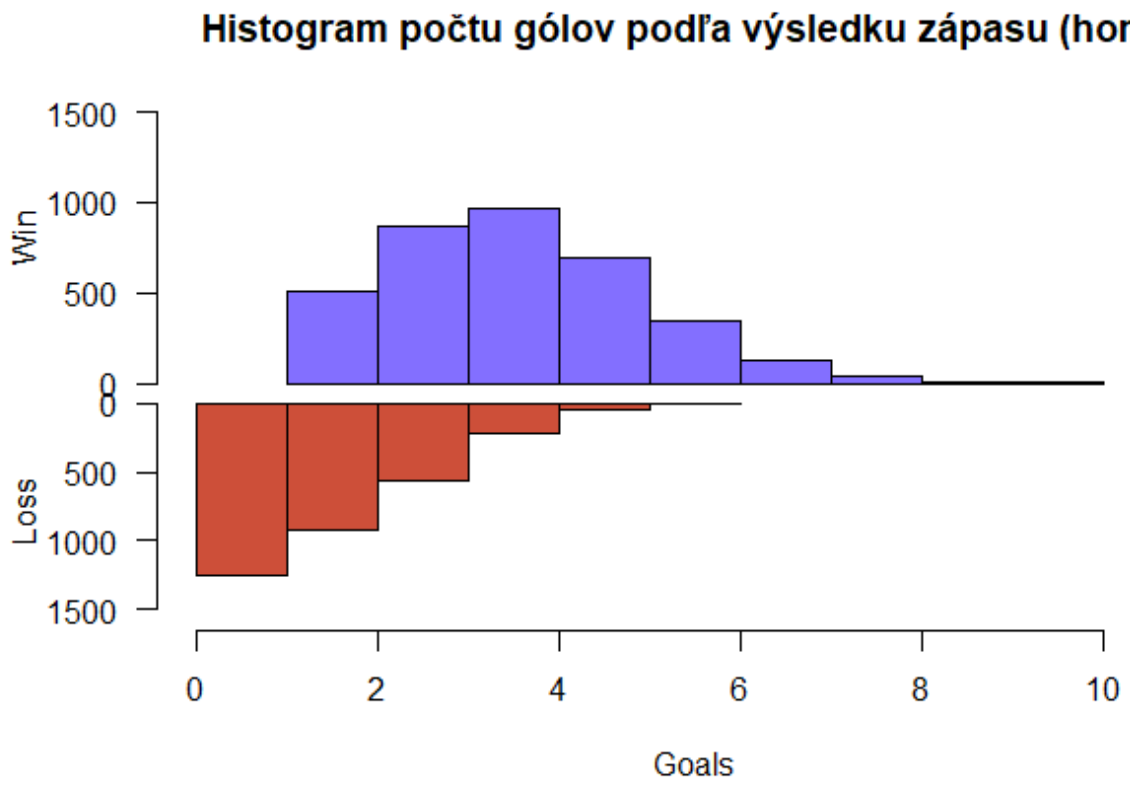
Hide

```
hist(df$goals.home[df$won.home == 0], main="" , ylab="Loss", xlab="Goals", ylim=c(160
0,0), xlim=c(0,11), las=1 , col="tomato3"  , breaks=8)

par(mfrow=c(2,1))
```

Hide

```
par(mar=c(0,5,3,3))
```

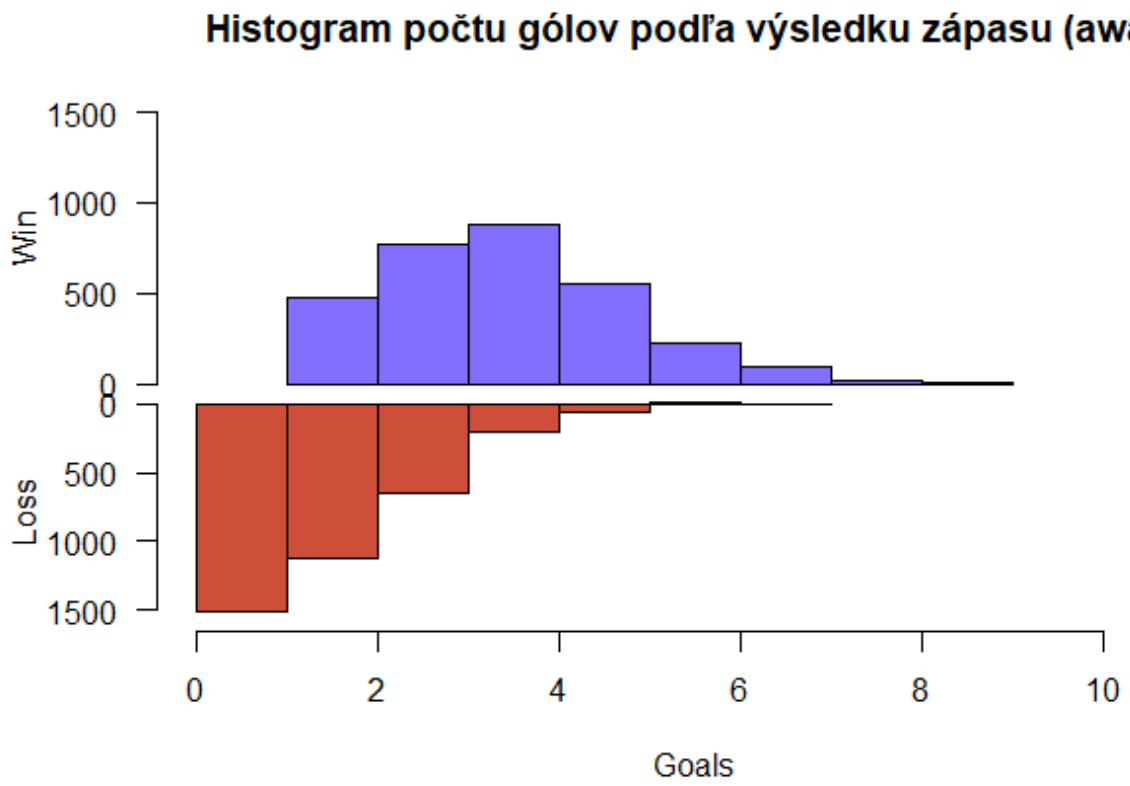


Hide

```
hist(df$goals.away[df$won.away == 1], main="Histogram počtu gólov podľa výsledku zápa
su (away)" , ylab="Win", xlab="", ylim=c(0,1600), xlim=c(0,11) , xaxt="n", las=1 , co
l="slateblue1", breaks=8)
par(mar=c(5,5,0,3))
```

Hide

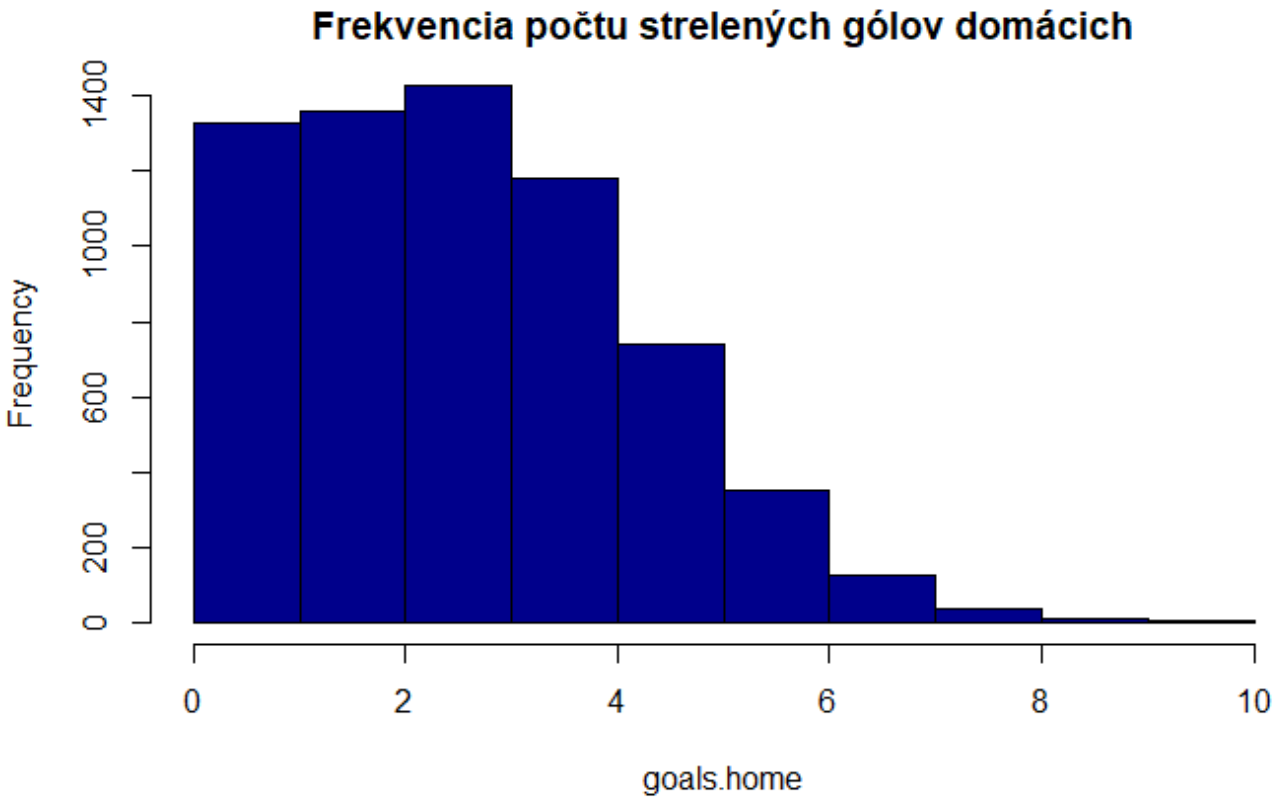
```
hist(df$goals.away[df$won.away == 0], main="" , ylab="Loss", xlab="Goals", ylim=c(160
0,0), xlim=c(0,11), las=1 , col="tomato3"  , breaks=6)
```



Z grafov môžeme vidieť, že ten zlom výhry nastáva práve pri troch góloch. Samozrejme to neznamená, že ak tím skóruje tri, alebo viac krát, tak vyhrá.

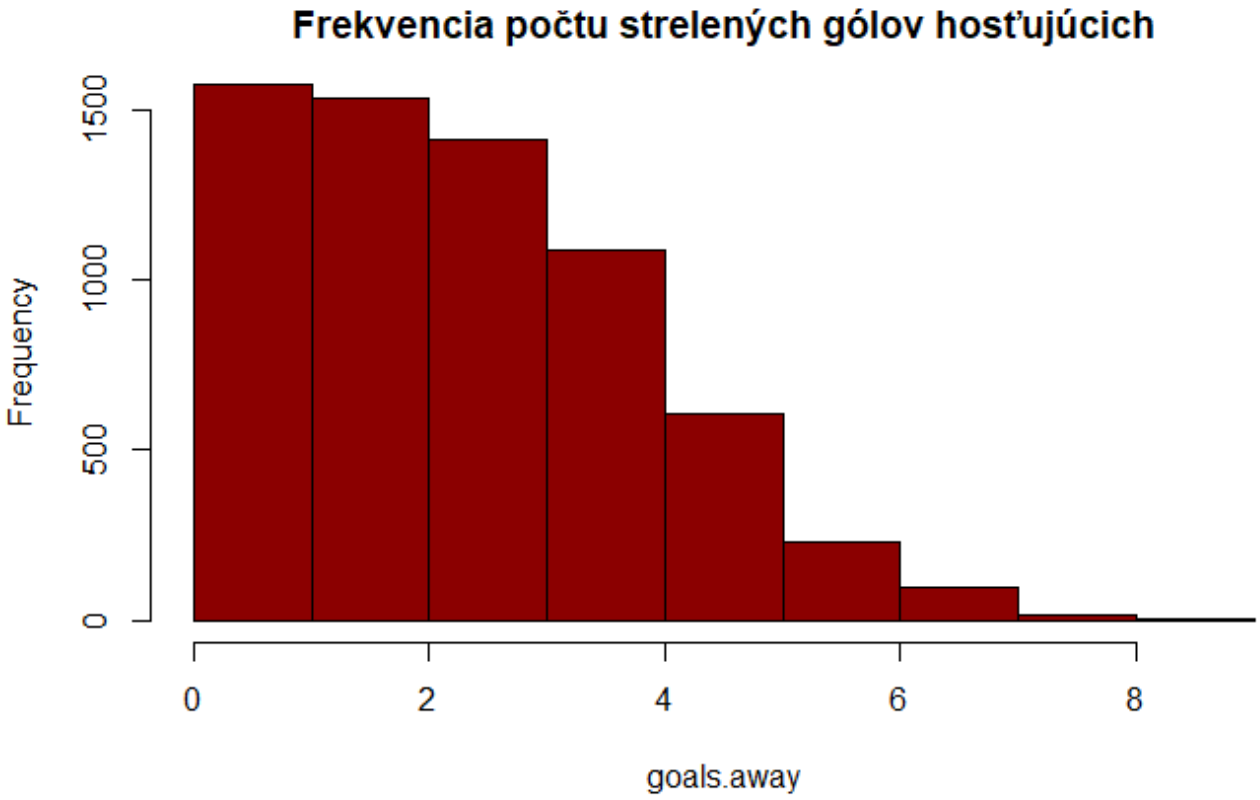
Hide

```
hist(df$goals.home, xlab="goals.home", col="darkblue", main="Frekvencia počtu strelen
ých gólov domácich")
```



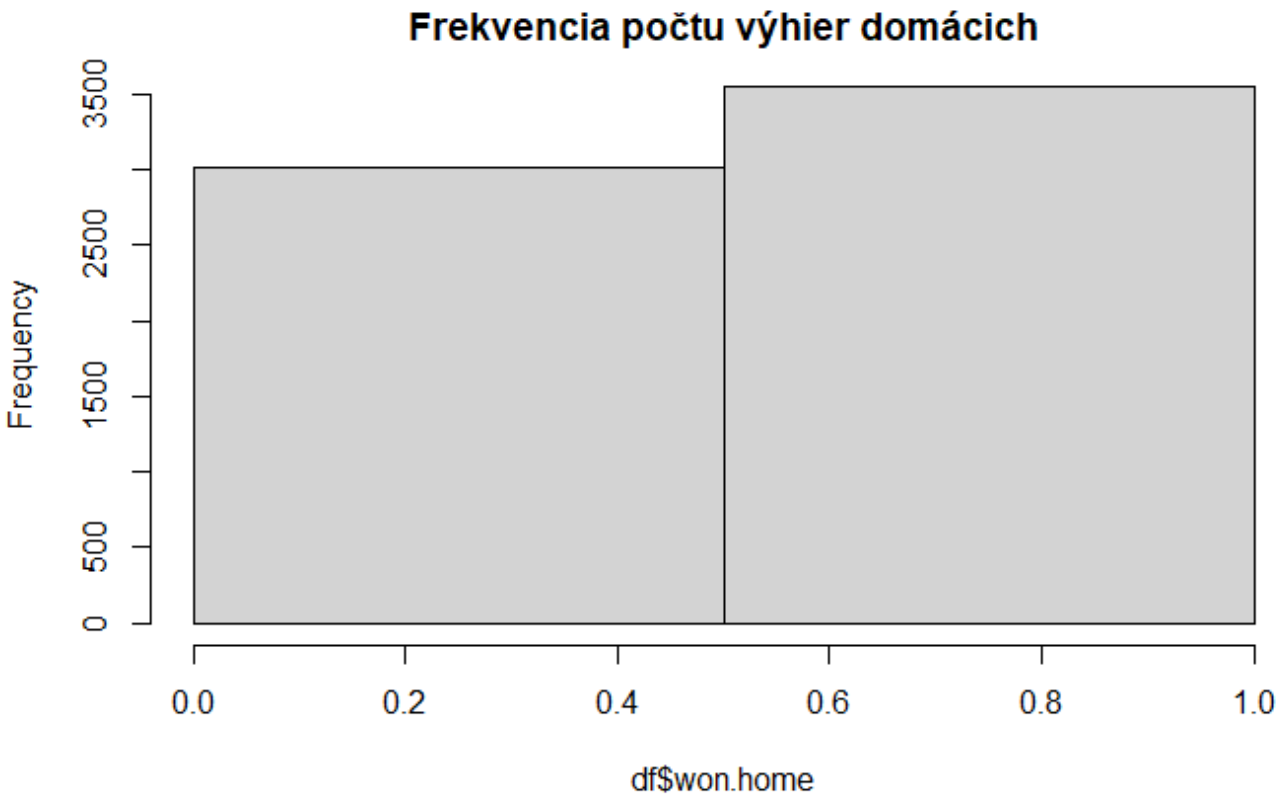
Hide

```
hist(df$goals.away, breaks = 10, xlab="goals.away", col="darkred", main="Frekvencia p
očtu strelených gólov hosťujúcich")
```



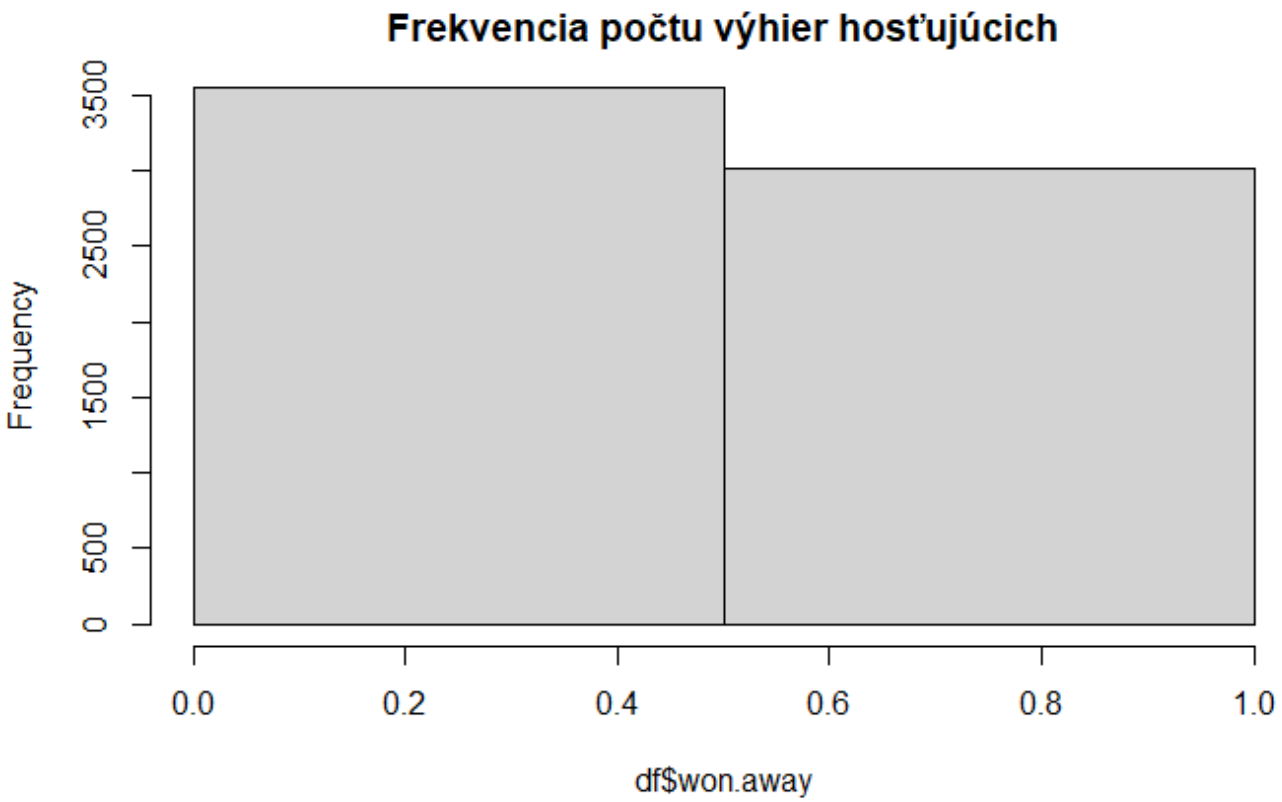
Hide

```
hist(df$won.home, breaks = 2, main="Frekvencia počtu výhier domácich")
```



Hide

```
hist(df$won.away, breaks = 2, main="Frekvencia počtu výhier hosťujúcich")
```



Z histogramov vidíme, že domáce tímy strielajú viac gólov ako hosťujúce. Taktiež hostujúce tímy najčastejšie strelia nula gólov. Vidíme aj, že tímy, ktoré hrajú doma vyhrávajú častejšie, bez ohľadu na sterelené góly.

Hide

```
shapiro_df <- sample_n(df, 5000)
skewness(shapiro_df$goals.home)
```

```
[1] 0.3745005
```

Hide

```
shapiro.test(shapiro_df$goals.home)
```

```

    Shapiro-Wilk normality test

data:  shapiro_df$goals.home
W = 0.95702, p-value < 2.2e-16
```

Hide

```
skewness(shapiro_df$goals.away)
```

```
[1] 0.4216546
```

Hide

```
shapiro.test(shapiro_df$goals.away)
```

```

    Shapiro-Wilk normality test

data:  shapiro_df$goals.away
W = 0.95163, p-value < 2.2e-16
```

Zo Shapiro-Wilkovho testu môžeme vidieť, že ani jeden z atribútov goals.home a goals.away nemá normálne rozdelenie. Pretože v oboch prípadoch je hodnota p < 0.05.

Hide

```
model_glm = glm(won.home ~ goals.home, data = df, family = "binomial")
coef(model_glm)
```

```
(Intercept)  goals.home
  -3.462533    1.275336
```

Hide

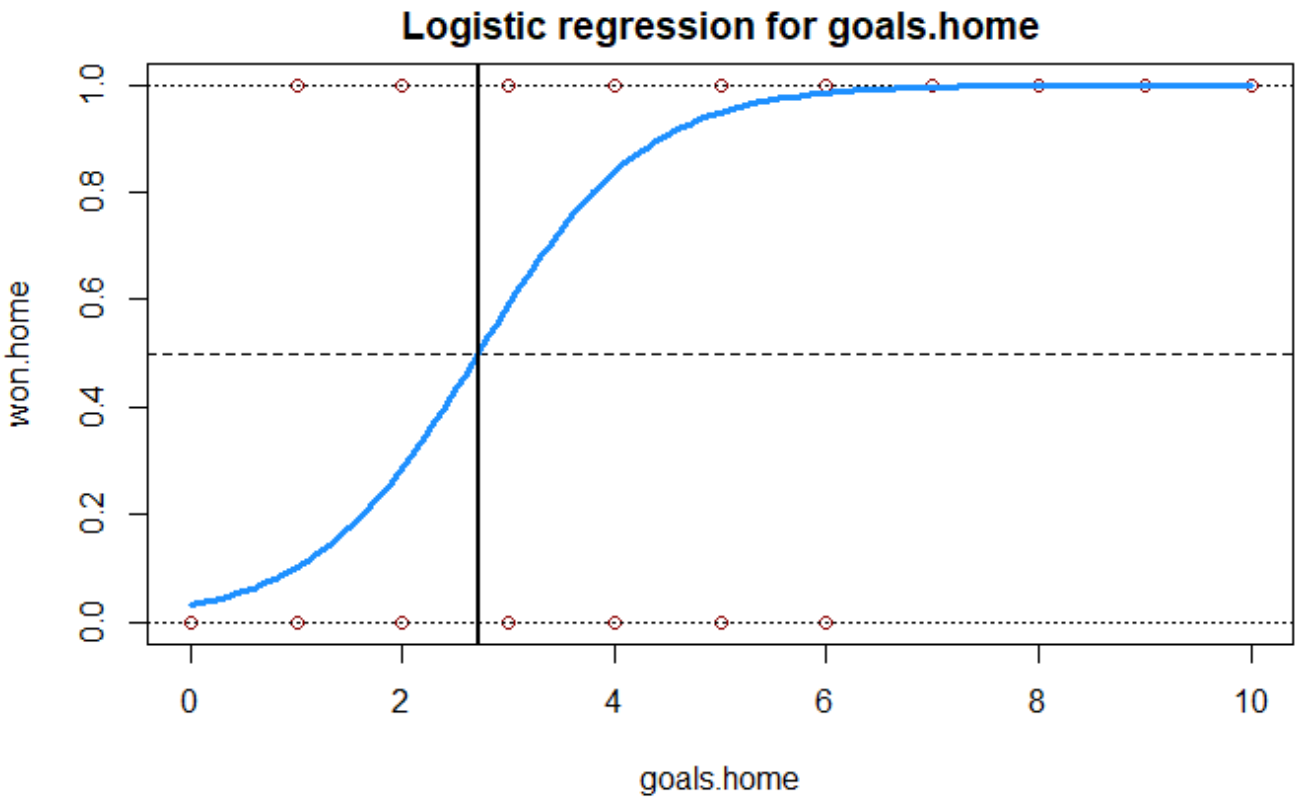
```
plot(won.home ~ goals.home, data = df,
     col = "darkred",
     main = "Logistic regression for goals.home")
abline(h = 0, lty = 3)
```

Hide

```
abline(h = 1, lty = 3)
abline(h = 0.5, lty = 2)
```

Hide

```
curve(predict(model_glm, data.frame( goals.home = x), type = "response"), add = TRUE,
lwd = 3, col = "dodgerblue")
abline(v = -coef(model_glm)[1] / coef(model_glm)[2], lwd = 2)
```



Hide

```
model_glm = glm(won.away ~ goals.away, data = df, family = "binomial")
coef(model_glm)
```

```
(Intercept)  goals.away
  -3.701000    1.277584
```

Hide

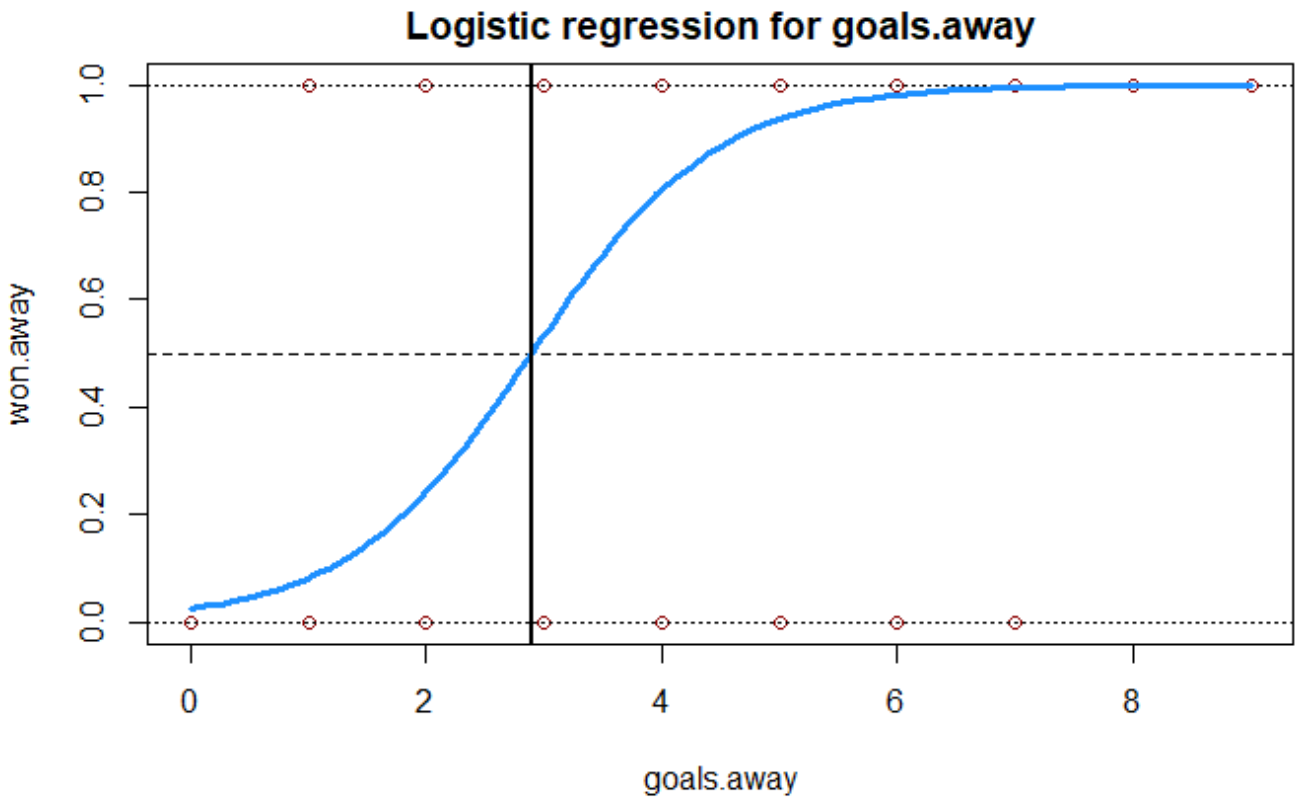
```
plot(won.away ~ goals.away, data = df,
     col = "darkred",
     main = "Logistic regression for goals.away")
abline(h = 0, lty = 3)
```

Hide

```
abline(h = 1, lty = 3)
abline(h = 0.5, lty = 2)
```

Hide

```
curve(predict(model_glm, data.frame( goals.away = x), type = "response"), add = TRUE,
lwd = 3, col = "dodgerblue")
abline(v = -coef(model_glm)[1] / coef(model_glm)[2], lwd = 2)
```



Podľa regresie môžeme vidieť, že domácemu tímu stačí často streliť menej gólov na výhru, ako hosťujúcemu. Z reálneho hladiska by sme však obe čísla zaokrúhlili na celé číslo 3, pretože góly nemôžeme reprezentovať desatinnými číslami. Logistický regresný model teda potvrdzuje našu hypotézu: Tím, ktorý v zápase strelí viac ako 3 góly (vrátane), pravdepodobne vyhrá.

Hypotéza 2.

Čím je väčší súhrnný počet striel na bránu, tým viac gólov padne v zápase.

Pracujeme s novými atribútmi, ktoré sme si vypočítali súčtom všetkých gólov v hre a súčtom všetkých striel v hre. **Atribúty:** shots.home, goals.home, shots.away, goals.away

Hide

```
par(mfrow=c(2,1))
par(mar=c(0,5,3,3))
hist(df$goals.home, ylab="Domáci", xlab="", ylim=c(0,1550), xlim=c(0,9), xaxt="n", la
s=1 , col="slateblue1", breaks=8, main="Počet strelených gólov pre oba tímy")
par(mar=c(5,5,0,3))
```

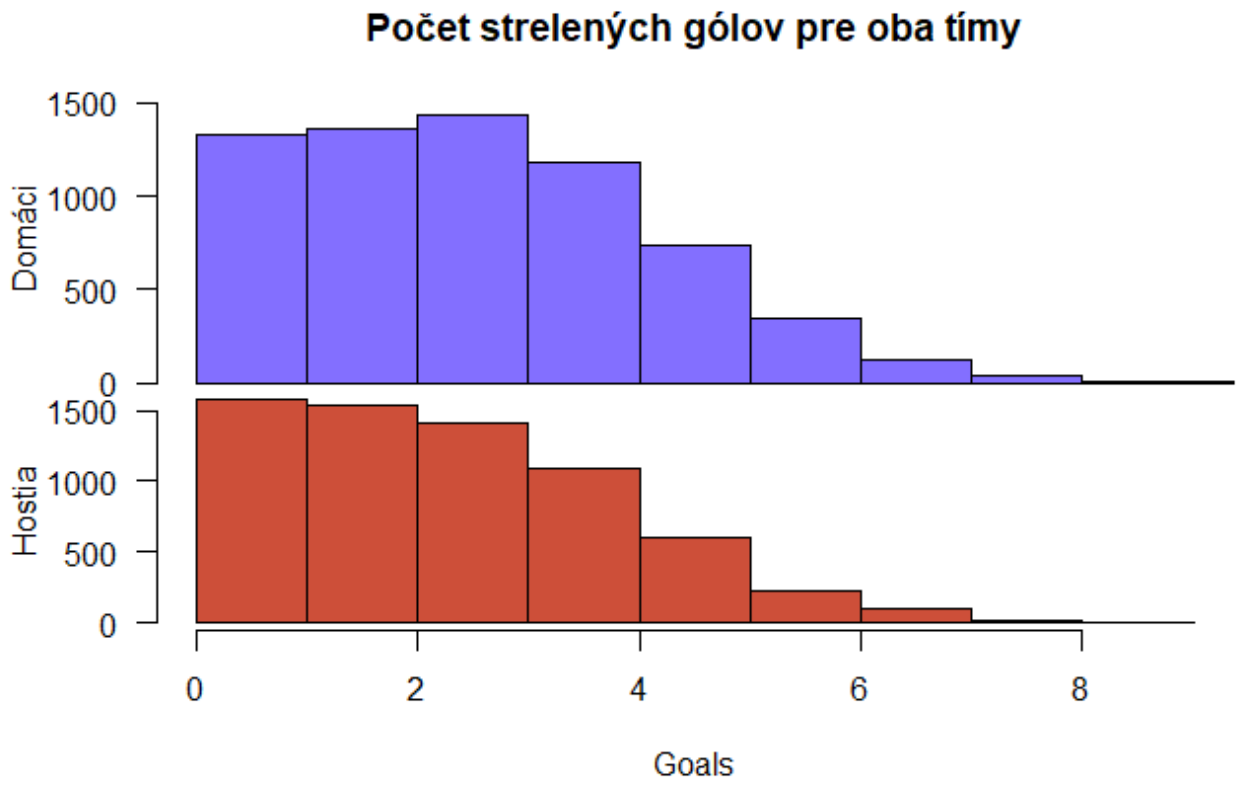
Hide

```
hist(df$goals.away, main="" , ylab="Hostia", xlab="Goals", ylim=c(0,1550), xlim=c(0,
9), las=1 , col="tomato3"  , breaks=8)

par(mfrow=c(2,1))
```

Hide

```
par(mar=c(0,5,3,3))
```

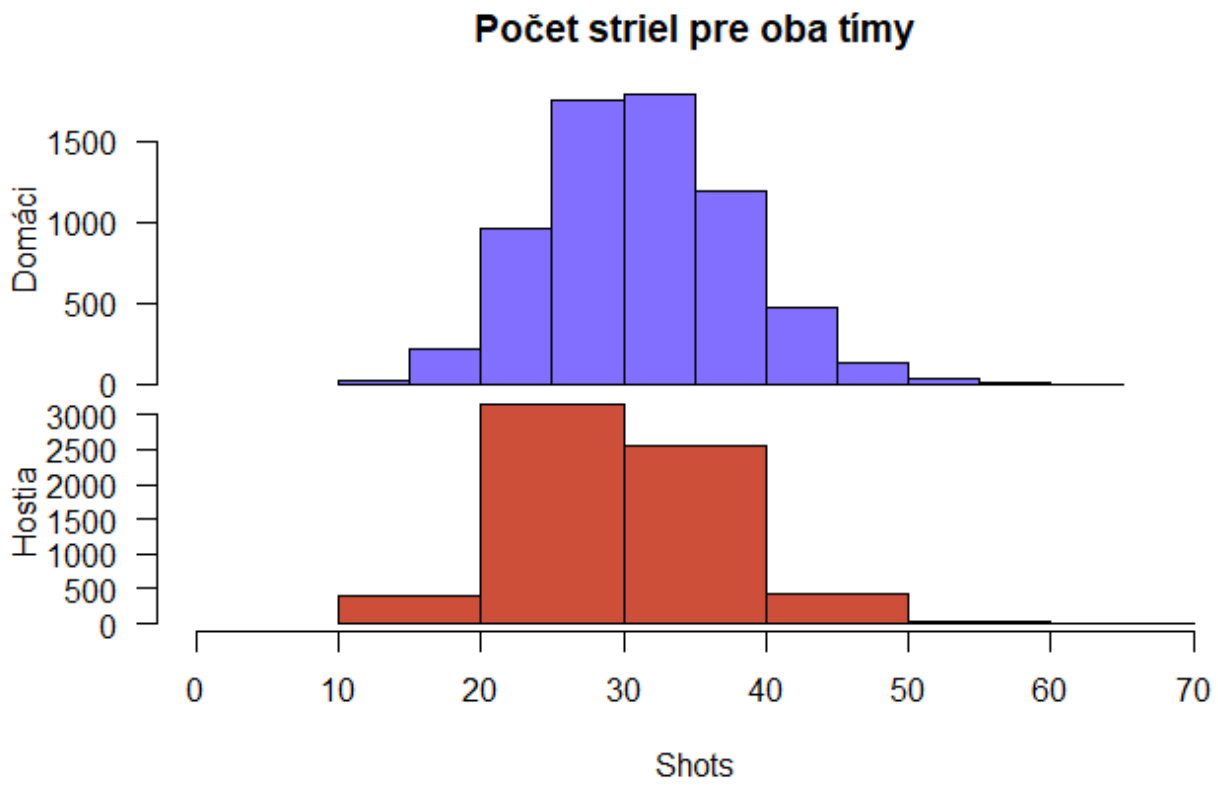


Hide

```
hist(df$shots.home, ylab="Domáci", xlab="", ylim=, xlim=c(0,70), xaxt="n", las=1 , col="slateblue1", breaks=8, main="Počet striel pre oba tímy")
par(mar=c(5,5,0,3))
```

Hide

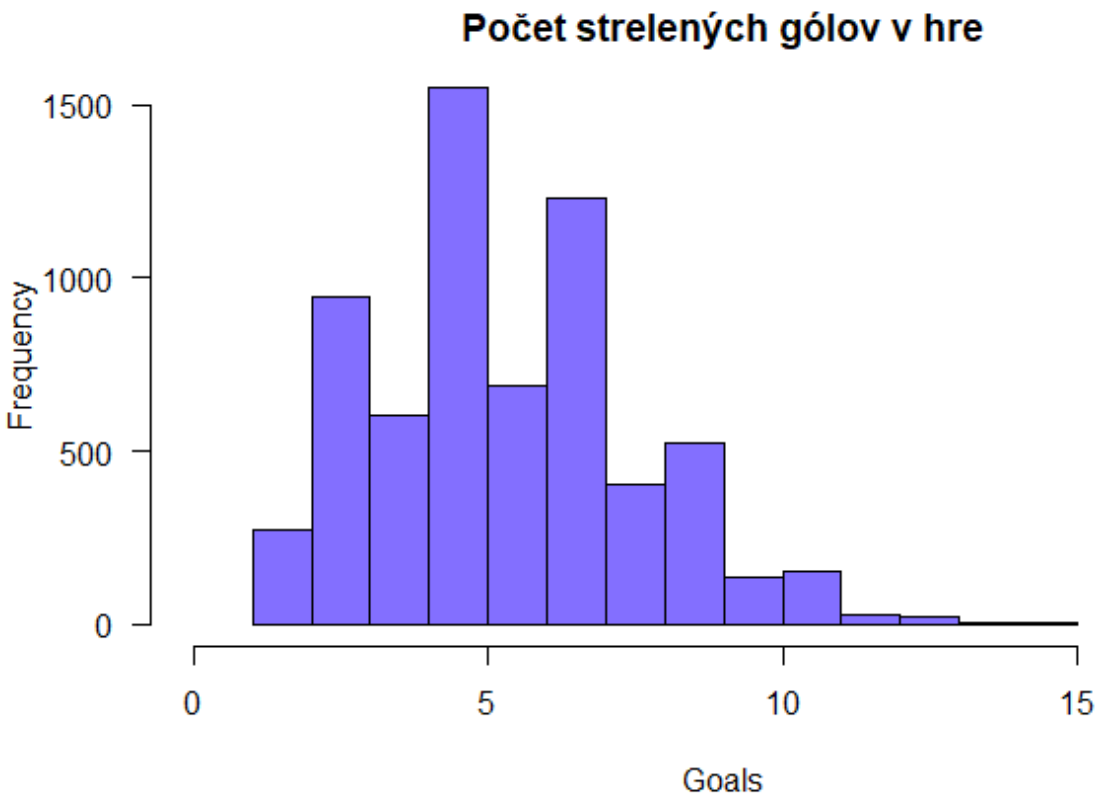
```
hist(df$shots.away, main="" , ylab="Hostia", xlab="Shots", ylim=, xlim=c(0,70), las=1 , col="tomato3"  , breaks=6)
```



Vidíme, že v domácom prostredí stielajú tímy viac gólov. Viac gólov je výsledkom väčšieho počtu striel na bránku.

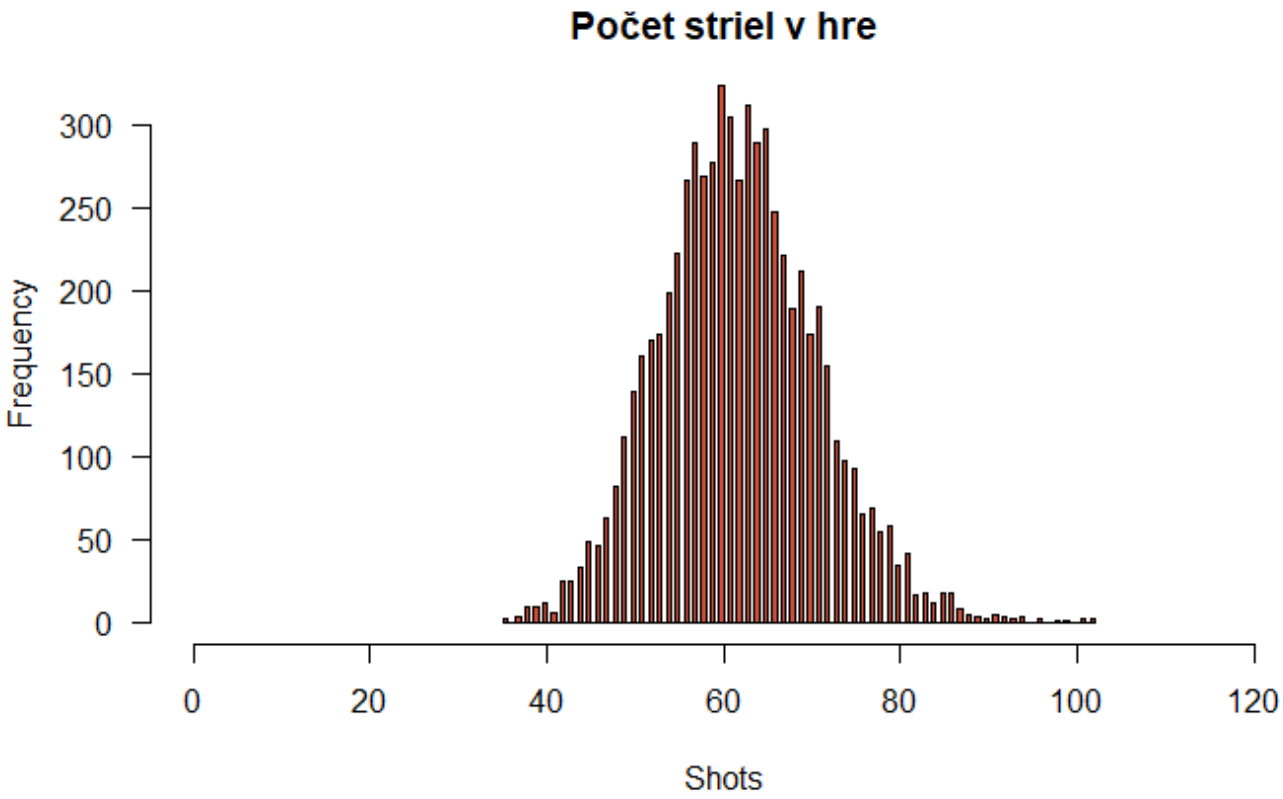
Hide

```
hist(df$goals.home+df$goals.away, ylab=, xlab="Goals", ylim=, xlim=c(0,18), las=1 , c
ol="slateblue1", breaks=12, main="Počet strelených gólov v hre")
```



Hide

```
hist(df$shots.home+df$shots.away, main="Počet striel v hre" , ylab=, xlab="Shots", yl
im=, xlim=c(0,120), las=1 , col="tomato3"  , breaks=120)
```



Hide

```
shapiro_df <- sample_n(df, 5000)
skewness(shapiro_df$shots.home+shapiro_df$shots.away)
```

```
[1] 0.2948271
```

Hide

```
shapiro.test(shapiro_df$shots.home+shapiro_df$shots.away)
```

```

    Shapiro-Wilk normality test

data:  shapiro_df$shots.home + shapiro_df$shots.away
W = 0.99391, p-value = 9.6e-14
```

Hide

```
skewness(shapiro_df$goals.home+shapiro_df$goals.away)
```

```
[1] 0.3803054
```

Hide

```
shapiro.test(shapiro_df$goals.home+shapiro_df$goals.away)
```

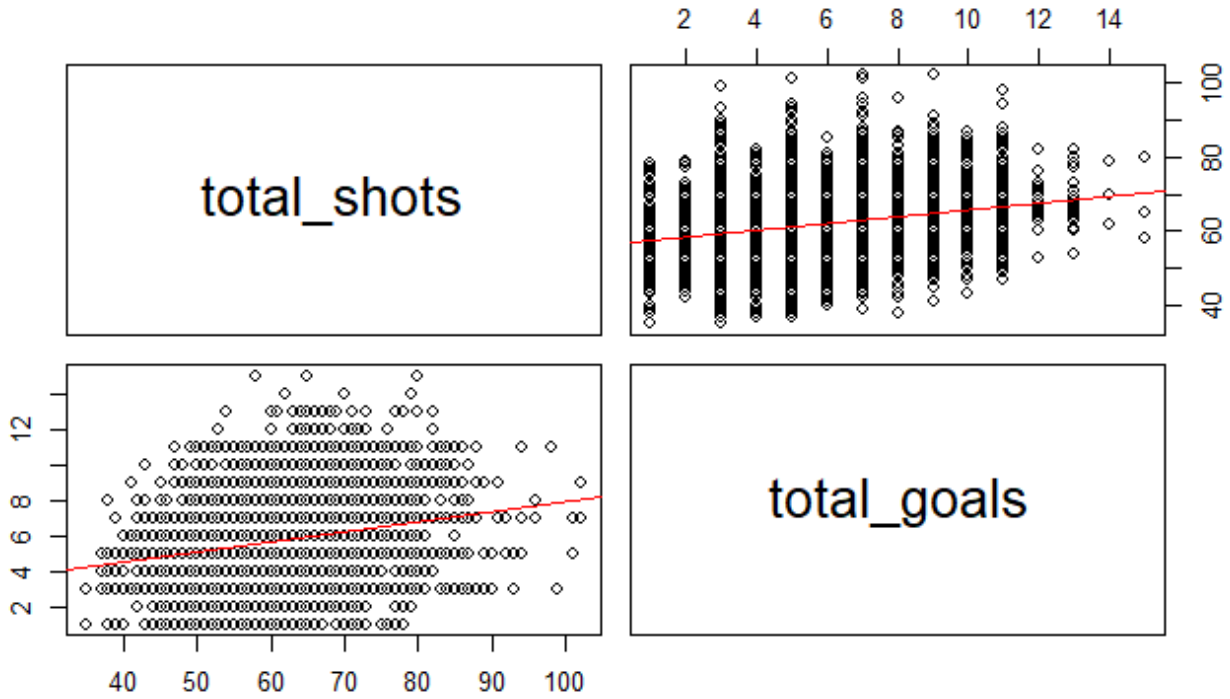
```
    Shapiro-Wilk normality test

data:  shapiro_df$goals.home + shapiro_df$goals.away
W = 0.96561, p-value < 2.2e-16
```

Podľa výsledkov Shapiro-Wilkovho testu nie je ani jedna z hodnôt všetkých gólov a všetkých striel v zápase s normálnym rozdelením.

Hide

```
total_shots <- df$shots.home+df$shots.away
total_goals <- df$goals.home+df$goals.away
pairs(~ total_shots + total_goals, panel=function(x,y){
  points(x,y)
  abline(lm(y~x), col='red')
})
```



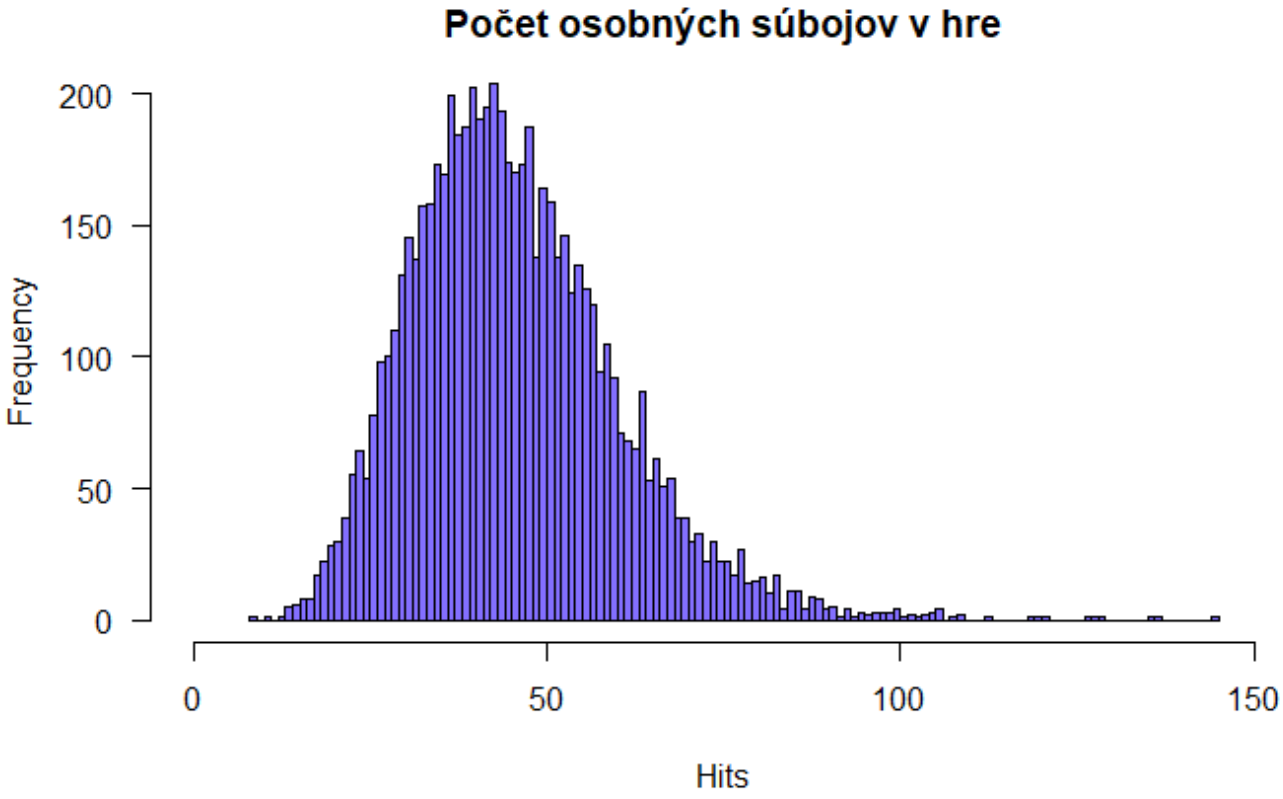
Hypotéza 3.

Čím je v zápase viac hitov, tým je zápas ostrejší a tým pádom rozhodcovia udelia tímom súhrnne viac trestných minút.

Pri tejto hypotéze pracujeme s diskrétnymi hodnotami. Taktiež sme pracovali s hodnotami súčtu atribútov.
Atribúty: hits.home, pim.home, hits.away, pim.away

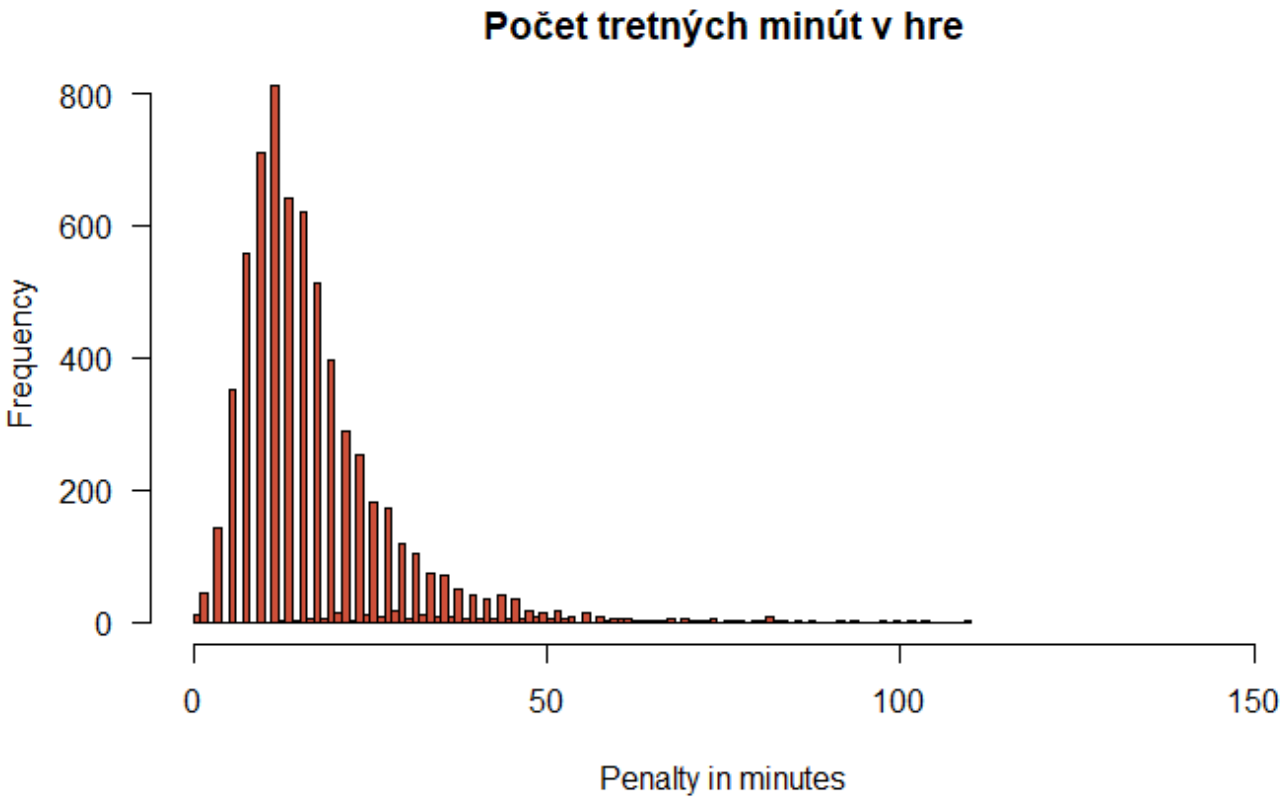
Hide

```
hist(df$hits.home+df$hits.away, ylab=, xlab="Hits", ylim=, xlim=c(0,150), las=1 , col
="slateblue1", breaks=120, main="Počet osobných súbojov v hre")
```



Hide

```
hist(df$pim.home+df$pim.away, main="Počet tretných minút v hre" , ylab=, xlab="Penalt
y in minutes", ylim=, xlim=c(0,150), las=1 , col="tomato3"  , breaks=120)
```



Hide

```
shapiro_df <- sample_n(df, 5000)
skewness(shapiro_df$hits.home+shapiro_df$hits.away)
```

```
[1] 0.9278663
```

Hide

```
shapiro.test(shapiro_df$hits.home+shapiro_df$hits.away)
```

```

    Shapiro-Wilk normality test

data:  shapiro_df$hits.home + shapiro_df$hits.away
W = 0.96003, p-value < 2.2e-16
```

Hide

```
skewness(shapiro_df$pim.home+shapiro_df$pim.away)
```

```
[1] 2.266907
```

Hide

```
shapiro.test(shapiro_df$pim.home+shapiro_df$pim.away)
```

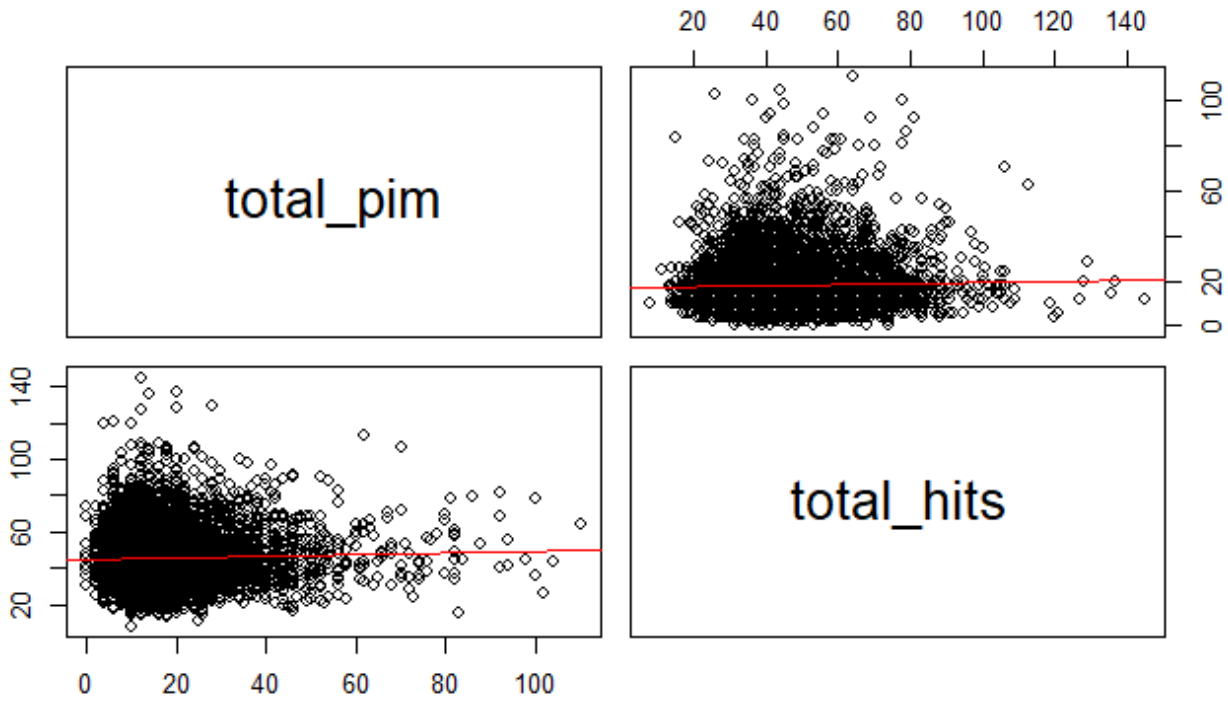
```
    Shapiro-Wilk normality test

data:  shapiro_df$pim.home + shapiro_df$pim.away
W = 0.81766, p-value < 2.2e-16
```

Oba testy normality nám dokázali, že ani súčet všetkých osobných súbojov v hre a ani súčet všetkých presilovkových minút nemajú normálne rozdelenia.

Hide

```
total_pim <- df$pim.home+df$pim.away
total_hits <- df$hits.home+df$hits.away
pairs(~ total_pim + total_hits, panel=function(x,y){
  points(x,y)
  abline(lm(y~x), col='red')
})
```



Hypotéza 4.

Čím má brankár vyššiu úspešnosť zákrokov, tým s väčšou pravdepodobnosťou jeho tím vyhrá.

Atribúty save_percentage boli vypočítané ako pomer počtu zákrokov brankára a počet všetkých striel oponenta. Pri tejto hypotéze pracujeme so spojitými atribútmi - save_percentage a kategorickými atribútmi - won. **Atribúty:** save_percentage.home, save_percentage.away, won.home, won.away

Hide

```
par(mfrow=c(2,1))
par(mar=c(0,5,3,3))
hist(df$save_percentage.home[df$won.home == 1], main="Úspešnosť brankára (home)" , yl
ab="Win", xlab="", ylim=c(0,1500), xlim=c(0.55,1) , xaxt="n", las=1 , col="slateblue
1", breaks=12)
par(mar=c(5,5,0,3))
```

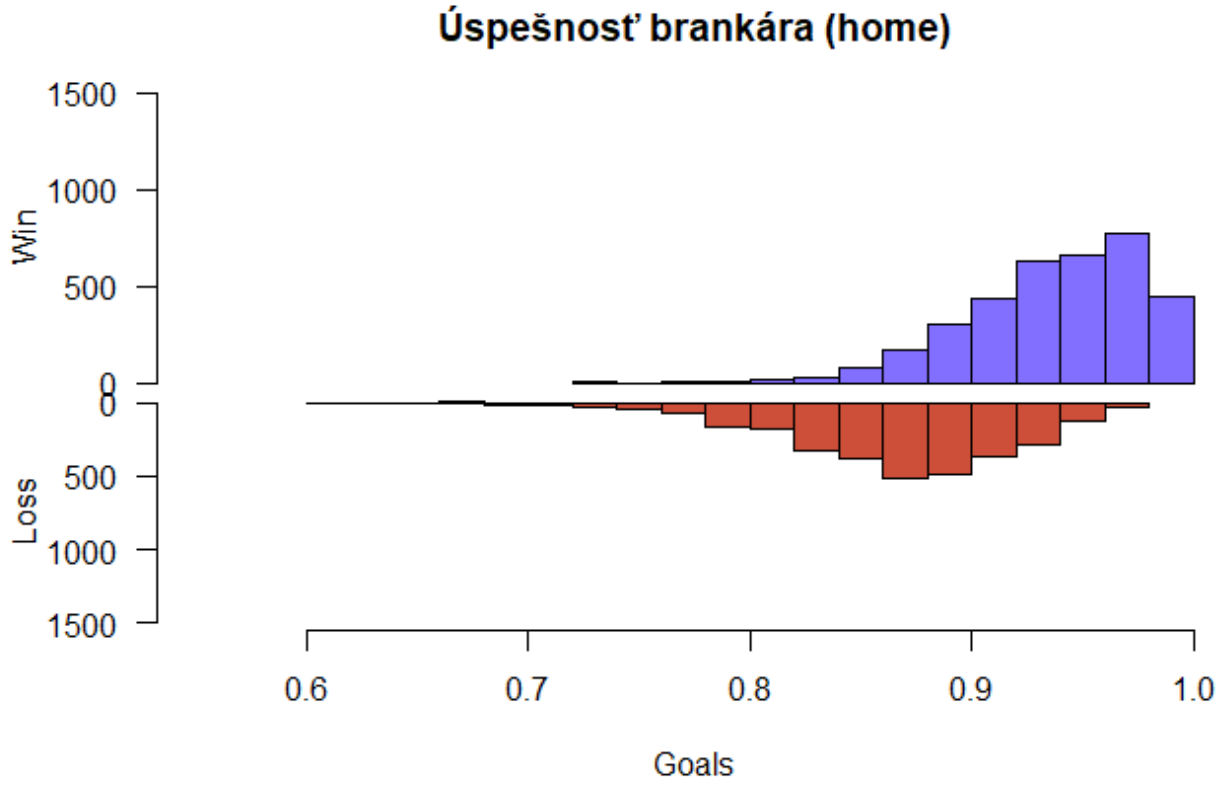
Hide

```
hist(df$save_percentage.home[df$won.home == 0], main="" , ylab="Loss", xlab="Goals",
ylim=c(1500,0), xlim=c(0.55,1), las=1 , col="tomato3"  , breaks=16)

par(mfrow=c(2,1))
```

Hide

```
par(mar=c(0,5,3,3))
```

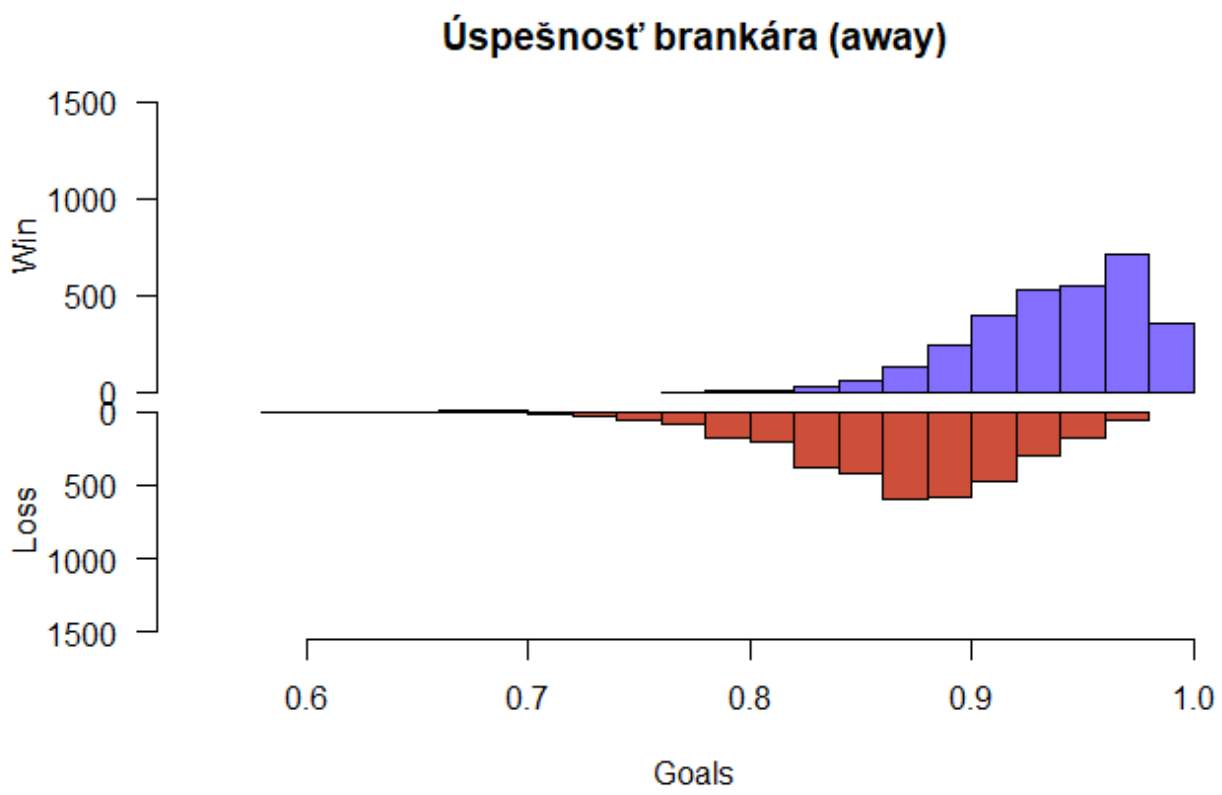


Hide

```
hist(df$save_percentage.away[df$won.away == 1], main="Úspešnosť brankára (away)" , yl
ab="Win", xlab="", ylim=c(0,1500), xlim=c(0.55,1), xaxt="n", las=1 , col="slateblue
1", breaks=12)
par(mar=c(5,5,0,3))
```

Hide

```
hist(df$save_percentage.away[df$won.away == 0], main="" , ylab="Loss", xlab="Goals",
ylim=c(1500,0), xlim=c(0.55,1), las=1 , col="tomato3"  , breaks=16)
```



Z grafu vidíme, že v domácom prostredí majú brankári väčšiu úspešnosť zákrokov.

Hide

```
shapiro_df <- sample_n(df, 5000)
skewness(shapiro_df$save_percentage.home)
```

```
[1] -0.6226289
```

Hide

```
shapiro.test(shapiro_df$save_percentage.home)
```

```

    Shapiro-Wilk normality test

data:  shapiro_df$save_percentage.home
W = 0.97007, p-value < 2.2e-16
```

Hide

```
skewness(shapiro_df$save_percentage.away)
```

```
[1] -0.619993
```

Hide

```
shapiro.test(shapiro_df$save_percentage.away)
```

```

    Shapiro-Wilk normality test

data:  shapiro_df$save_percentage.away
W = 0.97238, p-value < 2.2e-16
```

Z testov vidíme, že atribút save_percentage nemá normálne rozdelenie, ani pri domácom, ani pri hosťujúcom tíme.

Hypotéza 5.

Tím, ktorý hrá v domácom prostredí strelí viac gólov v zápase ako hosťujúci tím.

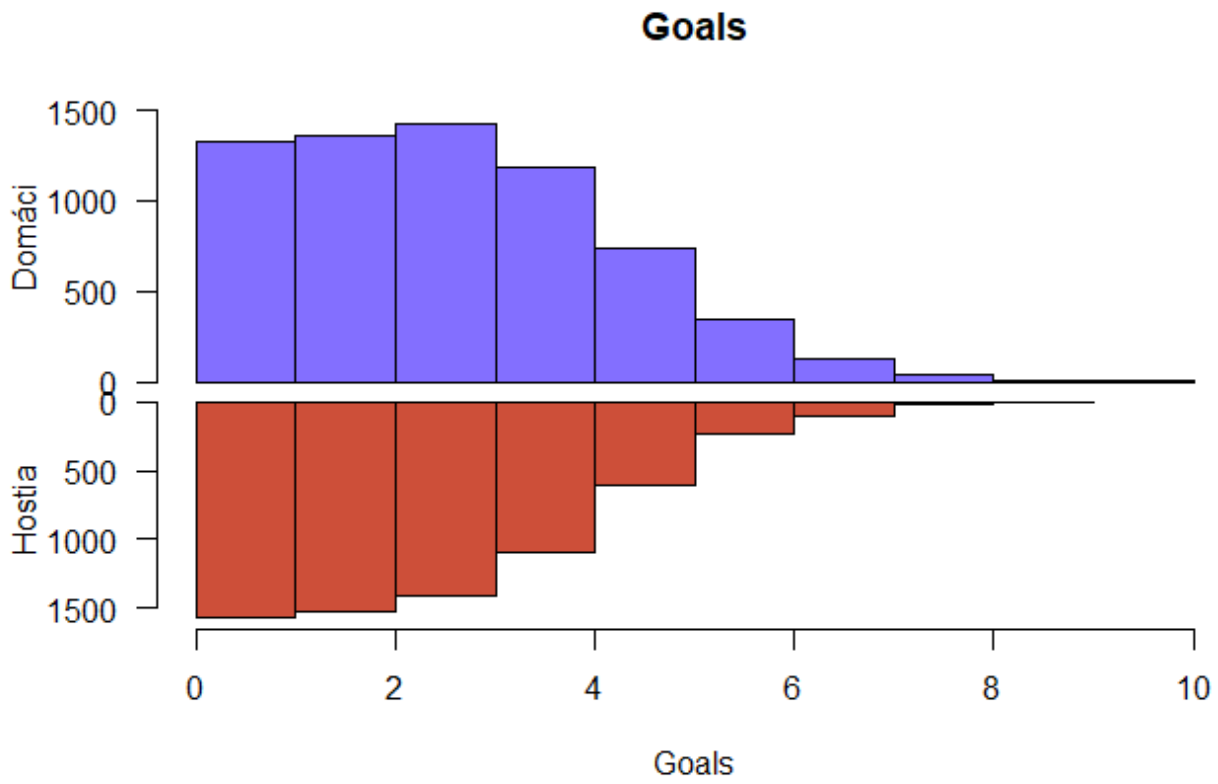
Atribúty: goals.home, goals.away

Hide

```
par(mfrow=c(2,1))
par(mar=c(0,5,3,3))
hist(df$goals.home, main="Goals" , ylab="Domáci", xlab="", ylim=c(0,1600), xlim=c(0,1
0), xaxt="n", las=1 , col="slateblue1", breaks=12)
par(mar=c(5,5,0,3))
```

Hide

```
hist(df$goals.away, main="" , ylab="Hostia", xlab="Goals", ylim=c(1600,0), xlim=c(0,1
0), las=1 , col="tomato3"  , breaks=12)
```



Už z grafu vidíme, že domáce tímy strielajú viac gólov, ako hosťujúce. Testo normality gólov sme robili vyššie a pre ani jeden z týchto atribútov nevišiel pozitívne, resp. oba atribúty nemajú normálne rozdelenie.