Transfermarkt

```
library(dplyr)
library(ggplot2)

options(scipen = 999)

load("data/transfermarkt.RData")
```

```
glimpse(transfermarkt)
```

```
## Rows: 7,030
## Columns: 36
## $ ano_campeonato           <int> 2003, 2003, 2003, 2003, 2003, 2003, 2003, 200~
## $ data                     <date> 2003-03-29, 2003-03-29, 2003-03-30, 2003-03-~
## $ horario                  <chr> "08:00", "08:00", "09:00", "09:00", "09:00", ~
## $ rodada                   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, ~
## $ estadio                  <chr> "Estádio Brinco de Ouro da Princesa", "Arena ~
## $ arbitro                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ publico                  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ publico_max              <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ time_man                 <chr> "Guarani", "Atlético-PR", "Flamengo", "Goiás ~
## $ time_vis                 <chr> "Vasco da Gama", "Grêmio", "Coritiba FC", "Pa~
## $ tecnico_man              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ tecnico_vis              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ colocacao_man            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ colocacao_vis            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ valor_equipe_titular_man <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ valor_equipe_titular_vis <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ idade_media_titular_man  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ idade_media_titular_vis  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ gols_man                 <int> 4, 2, 1, 2, 1, 2, 2, 0, 2, 1, 2, 0, 1, 0, 0, ~
## $ gols_vis                 <int> 2, 0, 1, 2, 1, 0, 2, 0, 2, 1, 2, 3, 1, 0, 1, ~
## $ gols_1_tempo_man         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ gols_1_tempo_vis         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ escanteios_man           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ escanteios_vis           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ faltas_man               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ faltas_vis               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ chutes_bola_parada_man   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ chutes_bola_parada_vis   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ defesas_man              <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ defesas_vis              <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ impedimentos_man         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ impedimentos_vis         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ chutes_man               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ chutes_vis               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
## $ chutes_fora_man          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ chutes_fora_vis          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
transfermarkt = transfermarkt %>%
  filter(ano_campeonato >= 2015,
         !is.na(gols_man))
```

```
transfermarkt %>%
  filter(time_man == "Flamengo") %>%
  select(data, valor_equipe_titular_man)
```

```
## # A tibble: 114 x 2
##    data       valor_equipe_titular_man
##    <date>                        <int>
##  1 2015-05-17                 26000000
##  2 2015-05-31                 30500000
##  3 2015-06-06                 35500000
##  4 2015-06-20                 28250000
##  5 2015-07-05                 22250000
##  6 2015-07-12                 24000000
##  7 2015-07-18                 26000000
##  8 2015-08-02                 30750000
##  9 2015-08-13                 22000000
## 10 2015-08-23                 28500000
## # ... with 104 more rows
```

```
sum(is.na(transfermarkt$valor_equipe_titular_man))/nrow(transfermarkt)
```

```
## [1] 0.04607284
```

```
sum(is.na(transfermarkt$valor_equipe_titular_vis))/nrow(transfermarkt)
```

```
## [1] 0.04607284
```

Existem partidas com missing nos valores de mercado, repetir os valores da última partida não missing ou usar a média da temporada até então?

Acho que usar a média da temporada dá menos problema pois evita problema de jogos com time titular após jogos com time reserva.

```
teams = sort(unique(transfermarkt$time_man))
teams
```

```
##  [1] "América-MG"     "Athletico-PR"   "Atlético-GO"    "Atlético-MG"
##  [5] "Atlético-PR"    "Avaí FC"        "Botafogo"       "Ceará SC"
##  [9] "Chapecoense"    "Corinthians"    "Coritiba FC"    "Cruzeiro"
## [13] "CSA"            "EC Bahia"       "EC Vitória"     "Figueirense FC"
## [17] "Flamengo"       "Fluminense"     "Fortaleza"      "Goiás EC"
## [21] "Grêmio"         "Internacional"  "Joinville-SC"   "Palmeiras"
## [25] "Paraná"         "Ponte Preta"    "RB Bragantino"  "Santa Cruz"
## [29] "Santos FC"      "São Paulo"      "Sport Recife"   "Vasco da Gama"
```

Existem “Athletico-PR” e “Atlético-PR”, vou padronizar.

```
transfermarkt$time_man[which(transfermarkt$time_man == "Atlético-PR")] = "Athletico-PR"
transfermarkt$time_vis[which(transfermarkt$time_vis == "Atlético-PR")] = "Athletico-PR"
```

```
team = NULL
season = NULL
mean_value = NULL

for(t in teams) {
  for(s in 2015:2020) {
    man = transfermarkt %>%
      filter(time_man == t,
             ano_campeonato == s) %>%
      .$valor_equipe_titular_man
    vis = transfermarkt %>%
      filter(time_vis == t,
             ano_campeonato == s) %>%
      .$valor_equipe_titular_vis
    all = c(man, vis)
    team = c(team, t)
    season = c(season, s)
    mean_value = c(mean_value, mean(all, na.rm = TRUE))
  }
}

mean_values = tibble(team, season, mean_value) %>%
  filter(!is.na(mean_value))
```

```
na_man = which(is.na(transfermarkt$valor_equipe_titular_man))
na_vis = which(is.na(transfermarkt$valor_equipe_titular_vis))
sum(na_man == na_vis) / length(na_man == na_vis)
```

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

Imputando os missings.

```
for(i in 1:length(na_man)) {
  man = transfermarkt$time_man[na_man[i]]
  vis = transfermarkt$time_vis[na_man[i]]
  ano = transfermarkt$ano_campeonato[na_man[i]]
  mean_man = mean_values %>%
    filter(season == ano,
           team == man) %>%
    .$mean_value
  mean_vis = mean_values %>%
    filter(season == ano,
           team == vis) %>%
    .$mean_value
  transfermarkt$valor_equipe_titular_man[na_man[i]] = mean_man
  transfermarkt$valor_equipe_titular_vis[na_man[i]] = mean_vis
}
```

Plotando a quantidade de gols de acordo com a (log) diferença do mandante/visitante.

```
transfermarkt = transfermarkt %>%
  mutate(dif_valor = valor_equipe_titular_man - valor_equipe_titular_vis,
         dif_log_valor = log(valor_equipe_titular_man) - log(valor_equipe_titular_vis))
```

```
transfermarkt %>%
  arrange(desc(dif_valor)) %>%
  select(ano_campeonato, time_man, time_vis, valor_equipe_titular_man,
         valor_equipe_titular_vis, dif_valor)
```

```
## # A tibble: 2,279 x 6
##    ano_campeonato time_man time_vis  valor_equipe_ti~ valor_equipe_ti~ dif_valor
##             <int> <chr>    <chr>               <dbl>            <dbl>     <dbl>
##  1           2019 Grêmio   EC Bahia          90500000         13650000  76850000
##  2           2019 Grêmio   CSA               82500000          7300000  75200000
##  3           2019 Flamengo Ceará SC          81000000          7800000  73200000
##  4           2019 Flamengo CSA               80000000          8700000  71300000
##  5           2019 Grêmio   Avaí FC           72900000          4500000  68400000
##  6           2018 Grêmio   América-~         72650000          7050000  65600000
##  7           2020 Flamengo Fluminen~         74650000         10050000  64600000
##  8           2019 Flamengo Vasco da~         83000000         19050000  63950000
##  9           2019 Grêmio   Botafogo          82500000         18700000  63800000
## 10           2020 Flamengo EC Bahia          71150000          7600000  63550000
## # ... with 2,269 more rows
```

```
transfermarkt %>%
  arrange(dif_valor) %>%
  select(ano_campeonato, time_man, time_vis, valor_equipe_titular_man,
         valor_equipe_titular_vis, dif_valor)
```

```
## # A tibble: 2,279 x 6
##    ano_campeonato time_man  time_vis valor_equipe_ti~ valor_equipe_ti~ dif_valor
##             <int> <chr>     <chr>               <dbl>            <dbl>     <dbl>
##  1           2020 Sport Re~ Flamengo          7950000         75050000 -67100000
##  2           2020 Coritiba~ Flamengo          6630000         71500000 -64870000
##  3           2019 Fortaleza Flamengo          8850000         73650000 -64800000
##  4           2019 Chapecoe~ Grêmio           12150000         76800000 -64650000
##  5           2019 Athletic~ Grêmio           17600000         81150000 -63550000
##  6           2020 Fluminen~ Flamengo          9750000         71800000 -62050000
##  7           2019 Botafogo  Flamengo         14900000         76000000 -61100000
##  8           2018 Fluminen~ Flamengo         13650000         74500000 -60850000
##  9           2019 Cruzeiro  Flamengo         17600000         78300000 -60700000
## 10           2019 CSA       Santos ~          8100000         68000000 -59900000
## # ... with 2,269 more rows
```

```
transfermarkt %>%
  arrange(valor_equipe_titular_man) %>%
  select(ano_campeonato, time_man, time_vis, valor_equipe_titular_man,
         valor_equipe_titular_vis, dif_valor)
```

```
## # A tibble: 2,279 x 6
```

```
##    ano_campeonato time_man time_vis  valor_equipe_ti~ valor_equipe_ti~ dif_valor
##             <int> <chr>    <chr>               <dbl>            <dbl>     <dbl>
##  1           2018 Paraná   Palmeiras         3050000         44950000 -41900000
##  2           2018 Paraná   Internac~         3100000         18400000 -15300000
##  3           2018 Paraná   Atlético~         3350000         16250000 -12900000
##  4           2018 Paraná   EC Vitór~         3550000          7950000  -4400000
##  5           2018 Paraná   Corinthi~         3730000         31500000 -27770000
##  6           2020 Goiás EC Atlético~         3850000         27900000 -24050000
##  7           2019 Avaí FC  Fortaleza         3950000          8100000  -4150000
##  8           2020 Botafogo São Paulo         3980000         39200000 -35220000
##  9           2019 Avaí FC  Vasco da~         4150000         13550000  -9400000
## 10           2018 Paraná   Vasco da~         4400000         14000000  -9600000
## # ... with 2,269 more rows
```

```
transfermarkt %>%
  arrange(valor_equipe_titular_vis) %>%
  select(ano_campeonato, time_man, time_vis, valor_equipe_titular_man,
         valor_equipe_titular_vis, dif_valor)
```

```
## # A tibble: 2,279 x 6
##    ano_campeonato time_man  time_vis valor_equipe_ti~ valor_equipe_ti~ dif_valor
##             <int> <chr>     <chr>              <dbl>            <dbl>     <dbl>
##  1           2017 Atlético~ Grêmio          28000000           600000  27400000
##  2           2019 Goiás EC  Grêmio          10400000          2800000   7600000
##  3           2020 Atlético~ Coritib~         5730000          2930000   2800000
##  4           2018 Palmeiras EC Vitó~        43450000          3280000  40170000
##  5           2018 Botafogo  Paraná          12150000          3300000   8850000
##  6           2020 Ceará SC  Botafogo         6450000          3430000   3020000
##  7           2018 EC Bahia  Paraná          11400000          3450000   7950000
##  8           2018 Cruzeiro  Paraná          29400000          3450000  25950000
##  9           2018 Ceará SC  Paraná           7080000          3500000   3580000
## 10           2016 São Paulo Santa C~        36500000          3700000  32800000
## # ... with 2,269 more rows
```

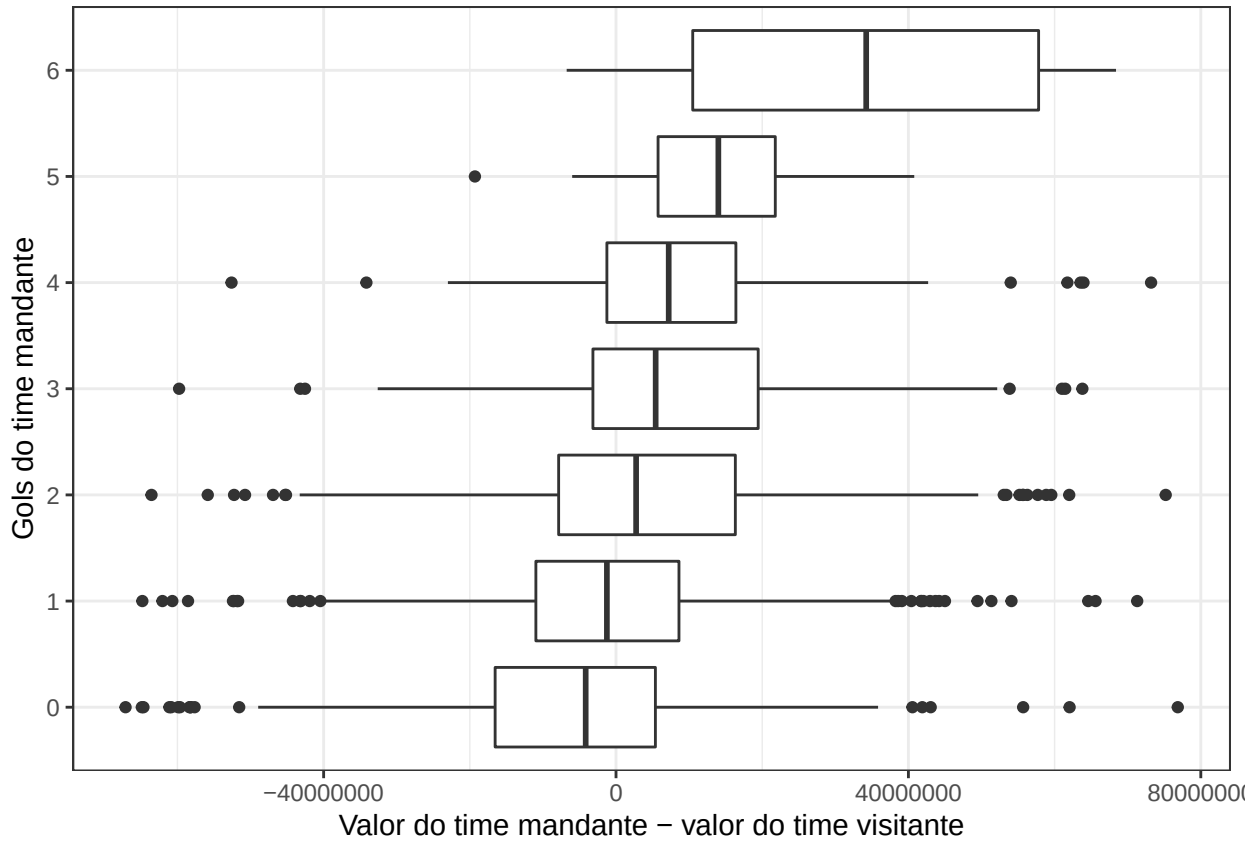
```
transfermarkt %>%
  arrange(desc(valor_equipe_titular_man)) %>%
  select(ano_campeonato, time_man, time_vis, valor_equipe_titular_man,
         valor_equipe_titular_vis, dif_valor)
```

```
## # A tibble: 2,279 x 6
##    ano_campeonato time_man time_vis  valor_equipe_ti~ valor_equipe_ti~ dif_valor
##             <int> <chr>    <chr>               <dbl>            <dbl>     <dbl>
##  1           2019 Grêmio   EC Bahia          90500000         13650000  76850000
##  2           2019 Flamengo Vasco da~         83000000         19050000  63950000
##  3           2019 Grêmio   Internac~         82900000         25200000  57700000
##  4           2019 Grêmio   Botafogo          82500000         18700000  63800000
##  5           2019 Grêmio   CSA               82500000          7300000  75200000
##  6           2020 Flamengo Corinthi~         81250000         22400000  58850000
##  7           2019 Flamengo Ceará SC          81000000          7800000  73200000
##  8           2019 Flamengo Palmeiras         80300000         54000000  26300000
##  9           2019 Flamengo Santos FC         80300000         30850000  49450000
## 10           2019 Flamengo Internac~         80300000         30400000  49900000
## # ... with 2,269 more rows
```

```
transfermarkt %>%
  arrange(desc(valor_equipe_titular_vis)) %>%
  select(ano_campeonato, time_man, time_vis, valor_equipe_titular_man,
         valor_equipe_titular_vis, dif_valor)
```

```
## # A tibble: 2,279 x 6
##    ano_campeonato time_man  time_vis valor_equipe_ti~ valor_equipe_ti~ dif_valor
##             <int> <chr>     <chr>               <dbl>            <dbl>     <dbl>
##  1           2019 Santos FC Flamengo         38300000         90900000 -52600000
##  2           2019 Palmeiras Flamengo         51750000         89400000 -37650000
##  3           2019 Palmeiras Grêmio           60750000         82500000 -21750000
##  4           2019 Athletic~ Grêmio           17600000         81150000 -63550000
##  5           2020 São Paulo Flamengo         28300000         79050000 -50750000
##  6           2019 Cruzeiro  Grêmio           20100000         78650000 -58550000
##  7           2019 Cruzeiro  Flamengo         17600000         78300000 -60700000
##  8           2019 Chapecoe~ Grêmio           12150000         76800000 -64650000
##  9           2019 Botafogo  Flamengo         14900000         76000000 -61100000
## 10           2020 Sport Re~ Flamengo          7950000         75050000 -67100000
## # ... with 2,269 more rows
```

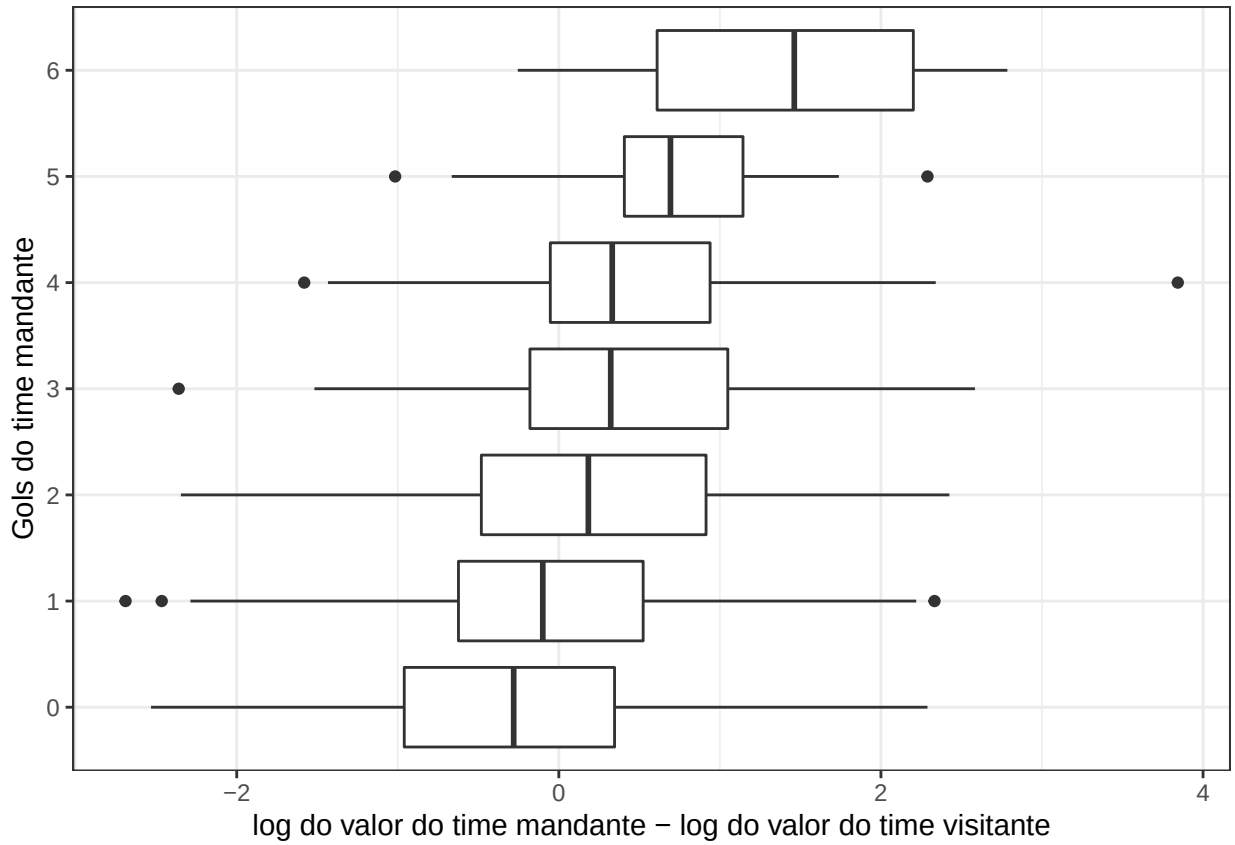
```
p = transfermarkt %>%
  ggplot(aes(x = factor(gols_man), y = dif_valor)) +
  geom_boxplot() +
  theme_bw() +
  xlab("Gols do time mandante") +
  ylab("Valor do time mandante - valor do time visitante") +
  coord_flip()
p
```




```
ggsave(filename = paste0("plots/mandante_sem_log.png"),
       plot = p, width = 10, height = 5, dpi = 1000)
```

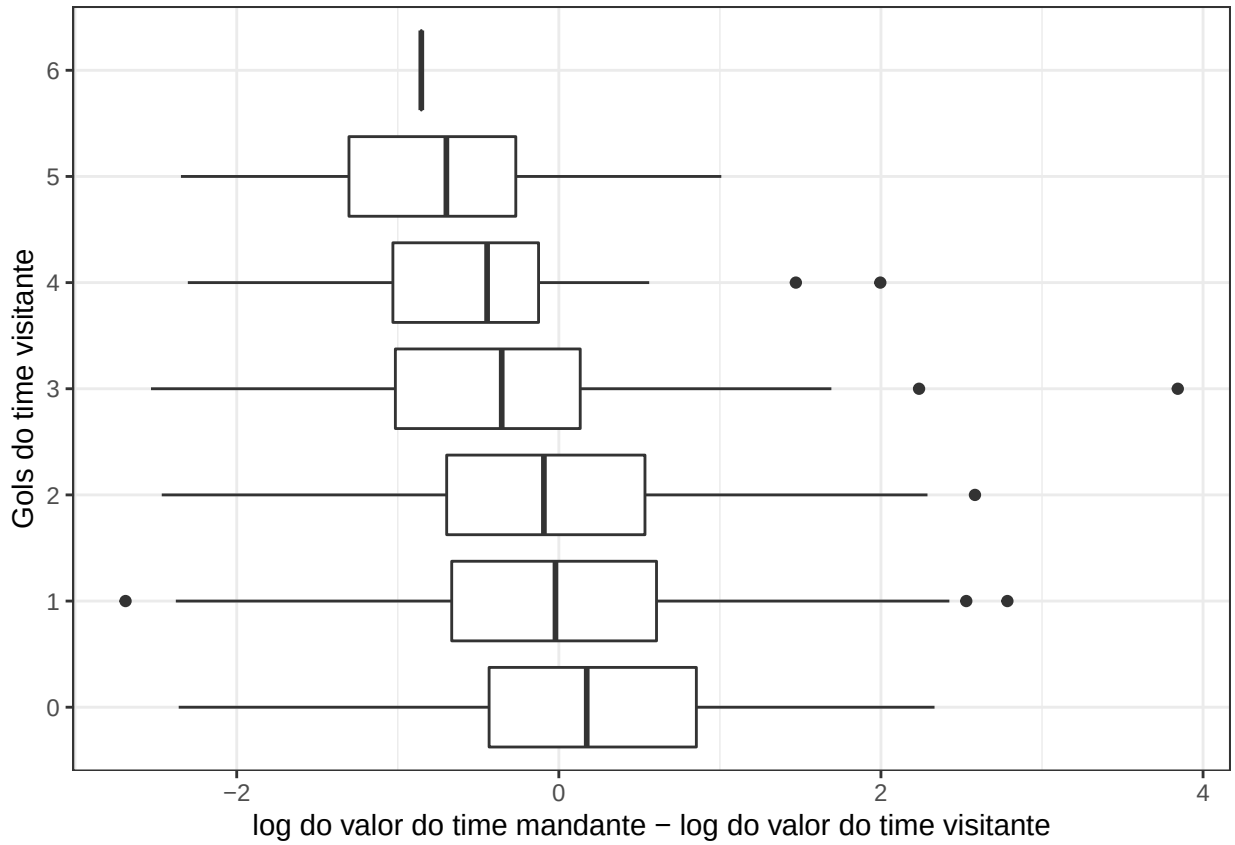
```
p = transfermarkt %>%
  ggplot(aes(x = factor(gols_vis), y = dif_valor)) +
  geom_boxplot() +
  theme_bw() +
  xlab("Gols do time visitante") +
  ylab("Valor do time mandante - valor do time visitante") +
  coord_flip()
ggsave(filename = paste0("plots/visitante_sem_log.png"),
       plot = p, width = 10, height = 5, dpi = 1000)
```

```
p = transfermarkt %>%
  ggplot(aes(x = factor(gols_man), y = dif_log_valor)) +
  geom_boxplot() +
  theme_bw() +
  xlab("Gols do time mandante") +
  ylab("log do valor do time mandante - log do valor do time visitante") +
  coord_flip()
p
```




```
ggsave(filename = paste0("plots/mandante_com_log.png"),
       plot = p, width = 10, height = 5, dpi = 1000)
```

```
p = transfermarkt %>%
  ggplot(aes(x = factor(gols_vis), y = dif_log_valor)) +
  geom_boxplot() +
  theme_bw() +
  xlab("Gols do time visitante") +
  ylab("log do valor do time mandante - log do valor do time visitante") +
  coord_flip()
p
```




```
ggsave(filename = paste0("plots/visitante_com_log.png"),
       plot = p, width = 10, height = 5, dpi = 1000)
```

```
cor(transfermarkt$gols_man, transfermarkt$dif_valor)
```

```
## [1] 0.2508633
```

```
cor(transfermarkt$gols_vis, transfermarkt$dif_valor)
```

```
## [1] -0.1689147
```

```
cor(transfermarkt$gols_man, transfermarkt$dif_log_valor)
```

```
## [1] 0.2657704
```

```
cor(transfermarkt$gols_vis, transfermarkt$dif_log_valor)
```

```
## [1] -0.1690221
```

```
cor(transfermarkt$gols_man, transfermarkt$dif_valor, method = "kendall")
```

```
## [1] 0.1842327
```

```
cor(transfermarkt$gols_vis, transfermarkt$dif_valor, method = "kendall")
```

```
## [1] -0.1242121
```

```
cor(transfermarkt$gols_man, transfermarkt$dif_log_valor, method = "kendall")
```

```
## [1] 0.1902347
```

```
cor(transfermarkt$gols_vis, transfermarkt$dif_log_valor, method = "kendall")
```

```
## [1] -0.1247808
```

Preparando o data set para juntar com o que temos.

```
load("~/GitHub/soccer-live-predictions/soccer-live-predictions/scrape/data/results2.RData")
```

```
transfermarkt = transfermarkt %>%
  rename(Season = ano_campeonato,
         Home_Team = time_man,
         Away_Team = time_vis,
         Value_Home = valor_equipe_titular_man,
         Value_Away = valor_equipe_titular_vis,
         Dif_Value = dif_valor,
         Dif_Log_Value = dif_log_valor) %>%
  select(Season, Home_Team, Away_Team, Value_Home, Value_Away,
         Dif_Value, Dif_Log_Value, rodada)
```

```
sort(unique(transfermarkt$Home_Team))
```

```
##  [1] "América-MG"     "Athletico-PR"   "Atlético-GO"    "Atlético-MG"
##  [5] "Avaí FC"        "Botafogo"       "Ceará SC"       "Chapecoense"
##  [9] "Corinthians"    "Coritiba FC"    "Cruzeiro"       "CSA"
## [13] "EC Bahia"       "EC Vitória"     "Figueirense FC" "Flamengo"
## [17] "Fluminense"     "Fortaleza"      "Goiás EC"       "Grêmio"
## [21] "Internacional"  "Joinville-SC"   "Palmeiras"      "Paraná"
## [25] "Ponte Preta"    "RB Bragantino"  "Santa Cruz"     "Santos FC"
## [29] "São Paulo"      "Sport Recife"   "Vasco da Gama"
```

```
sort(unique(results$Home_Team))
```

```
##  [1] "América-MG"          "Athletico-PR"        "Atlético-GO"
##  [4] "Atlético-MG"         "Avaí"                "Bahia"
##  [7] "Botafogo"            "Ceará"               "Chapecoense"
## [10] "Corinthians"         "Coritiba"            "Cruzeiro"
## [13] "Csa"                 "Figueirense"         "Flamengo"
## [16] "Fluminense"          "Fortaleza"           "Goiás"
## [19] "Grêmio"              "Internacional"       "Joinville"
## [22] "Palmeiras"           "Paraná"              "Ponte Preta"
## [25] "Red Bull Bragantino" "Santa Cruz"          "Santos"
## [28] "São Paulo"           "Sport"               "Vasco da Gama"
## [31] "Vitória"
```

```
transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Avaí FC")] =
  "Avaí"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Avaí FC")] =
  "Avaí"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Ceará SC")] =
  "Ceará"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Ceará SC")] =
  "Ceará"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Coritiba FC")] =
  "Coritiba"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Coritiba FC")] =
  "Coritiba"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "CSA")] =
  "Csa"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "CSA")] =
  "Csa"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "EC Bahia")] =
  "Bahia"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "EC Bahia")] =
  "Bahia"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "EC Vitória")] =
  "Vitória"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "EC Vitória")] =
  "Vitória"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Figueirense FC")] =
  "Figueirense"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Figueirense FC")] =
  "Figueirense"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Goiás EC")] =
  "Goiás"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Goiás EC")] =
  "Goiás"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Joinville-SC")] =
  "Joinville"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Joinville-SC")] =
  "Joinville"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "RB Bragantino")] =
  "Red Bull Bragantino"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "RB Bragantino")] =
  "Red Bull Bragantino"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Santos FC")] =
  "Santos"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Santos FC")] =
```

```
  "Santos"

transfermarkt$Home_Team[which(transfermarkt$Home_Team == "Sport Recife")] =
  "Sport"
transfermarkt$Away_Team[which(transfermarkt$Away_Team == "Sport Recife")] =
  "Sport"
```

```
results = results %>%
  mutate(rodada = ceiling(results$Match/10))
results$rodada[1020] = 28 # https://conteudo.cbf.com.br/sumulas/2017/142261se.pdf
results$rodada[1030] = 27 # https://conteudo.cbf.com.br/sumulas/2017/142271se.pdf
```

```
results = left_join(results, transfermarkt)
```

```
## Joining, by = c("Season", "Home_Team", "Away_Team", "rodada")
```

```
sum(is.na(results))
```

```
## [1] 0
```

```
save(results, file = "data/results.RData")
```