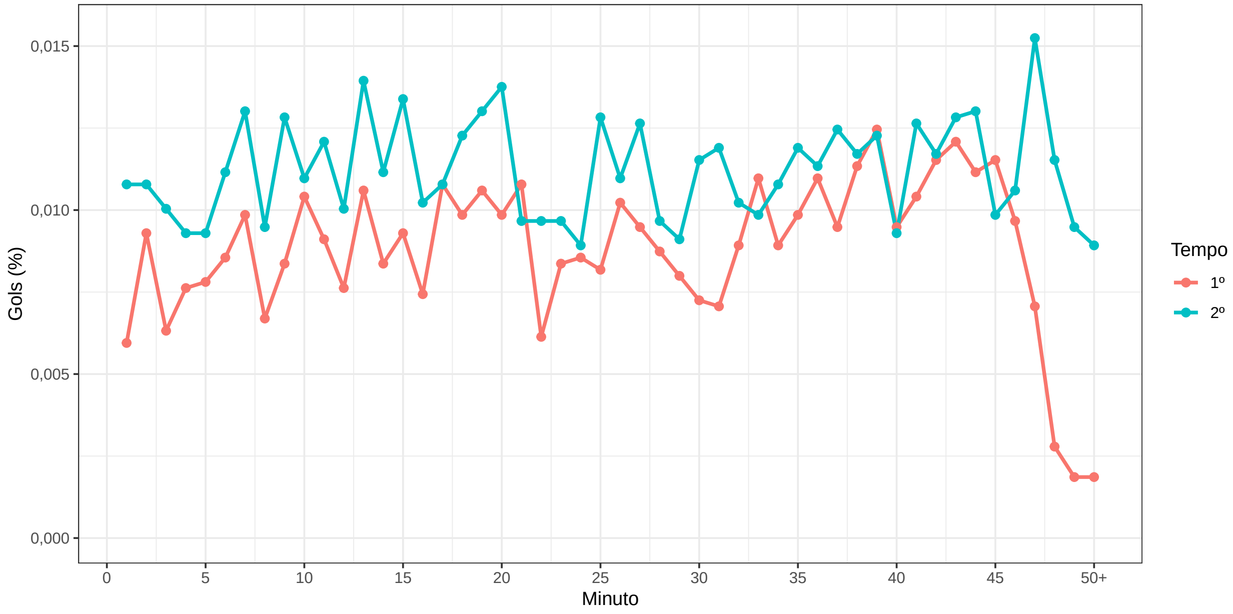
EDA Série A 2015-2020

```
options(OutDec = ",")

library(dplyr)
library(ggplot2)

load("scrape/data/goals.RData")
load("scrape/data/results.RData")
load("scrape/data/reds.RData")

resultados = results %>%
  rename(Ano = Season,
         Jogo = Match,
         Placar_1 = Score_Home,
         Placar_2 = Score_Away,
         Acréscimos_1 = Stoppage_Time_1,
         Acréscimos_2 = Stoppage_Time_2)

goals$Team[which(goals$Team == 1)] = "Mandante"
goals$Team[which(goals$Team == 2)] = "Visitante"
goals$Half[which(goals$Half == 1)] = "1º"
goals$Half[which(goals$Half == 2)] = "2º"
gols = goals %>%
  rename(Ano = Season,
         Jogo = Match,
         Placar_1 = Score_Home,
         Placar_2 = Score_Away,
         Acréscimo = Stoppage_Time,
         Minuto = Minute,
         Time = Team,
         Tempo = Half) %>%
  mutate(Time = as.factor(Time),
         Tempo = as.factor(Tempo))

reds$Team[which(reds$Team == 1)] = "Mandante"
reds$Team[which(reds$Team == 2)] = "Visitante"
reds$Half[which(reds$Half == 1)] = "1º"
reds$Half[which(reds$Half == 2)] = "2º"
reds = reds %>%
  rename(Ano = Season,
         Jogo = Match,
         Placar_1 = Score_Home,
         Placar_2 = Score_Away,
         Acréscimo = Stoppage_Time,
         Minuto = Minute,
         Time = Team,
         Tempo = Half) %>%
  mutate(Time = as.factor(Time),
         Tempo = as.factor(Tempo))
```

Gols por minuto

```
gols$Acréscimo[which(is.na(gols$Acréscimo))] = 0

gols = gols %>%
  mutate(Minuto = Minuto + Acréscimo)

gols$Minuto[which(gols$Minuto > 50)] = 50

tmp = gols %>%
  count(Minuto, Tempo) %>%
  mutate(p = n/nrow(gols))

tmp %>%
  ggplot(aes(x = Minuto, y = p, col = Tempo)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Gols (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 50, by = 5),
                     labels = c(seq(from = 0, to = 45, by = 5), "50+")) +
  ylim(0, 0.0155)
```

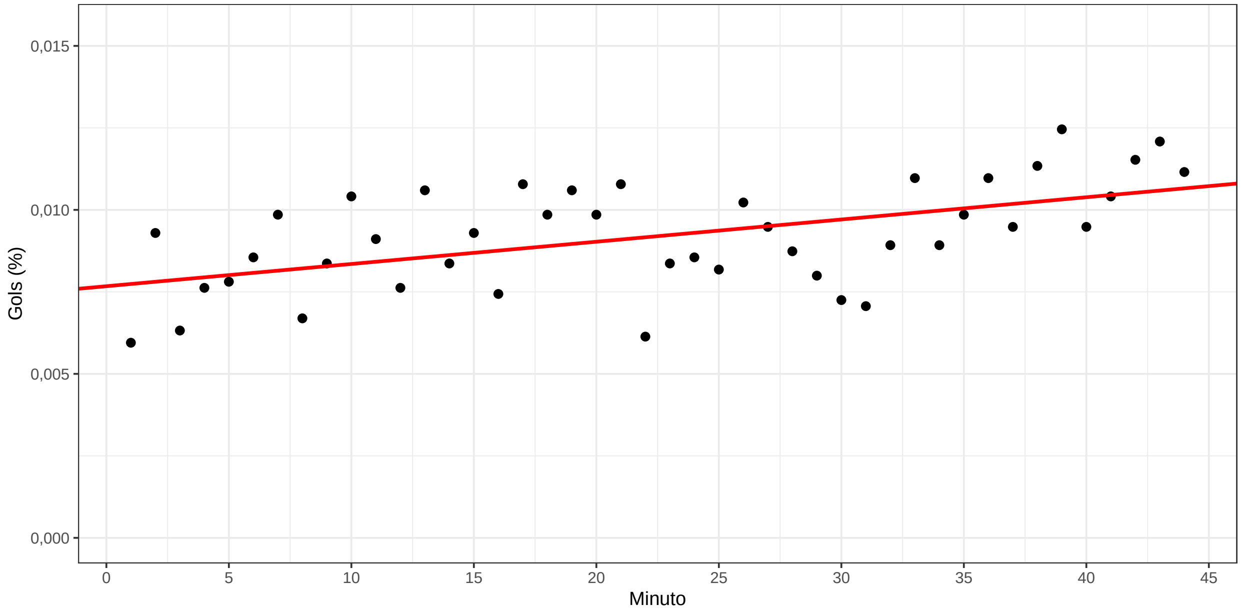


```
t1 = tmp %>%
  filter(Minuto < 45, Tempo == "1º")

lm1 = lm(p ~ Minuto, data = t1)
```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3,029e-03 -9,076e-04 -3,214e-05  1,019e-03  2,138e-03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7,672e-03  4,231e-04  18,132  < 2e-16 ***
## Minuto      6,783e-05  1,638e-05   4,141 0,000163 ***
## ---
## Signif. codes:  0 ’***’ 0,001 ’**’ 0,01 ’*’ 0,05 ’.’ 0,1 ’ ’ 1
##
## Residual standard error: 0,00138 on 42 degrees of freedom
## Multiple R-squared:   0,29,  Adjusted R-squared:  0,273
## F-statistic: 17,15 on 1 and 42 DF,  p-value: 0,0001629
```

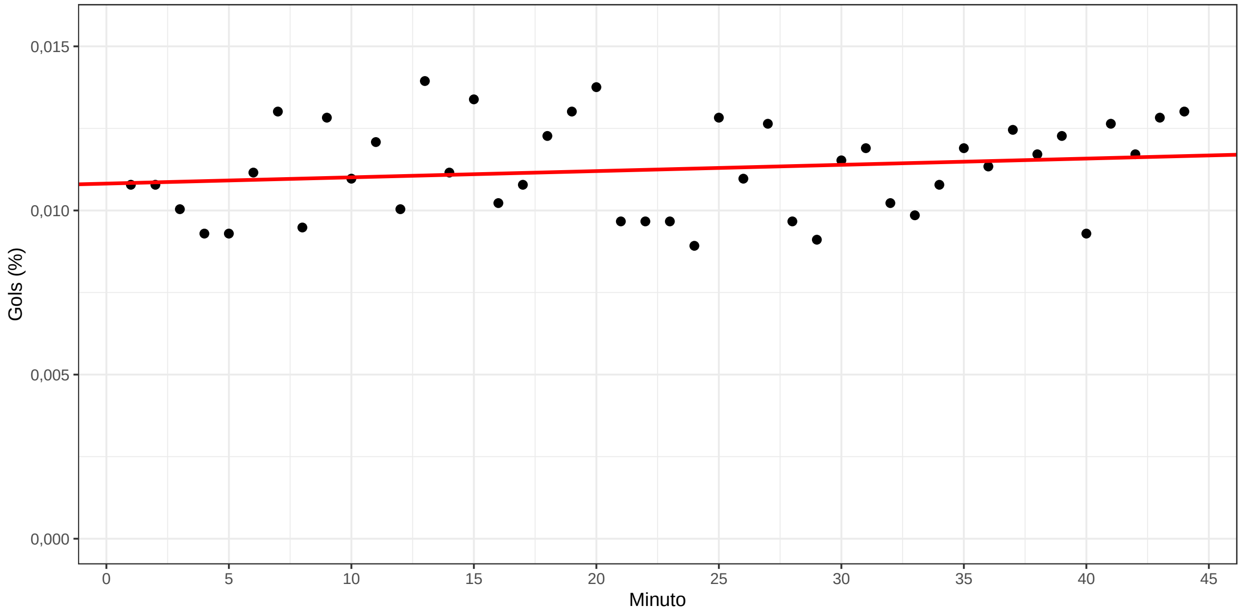
```
t1 %>%
  ggplot(aes(x = Minuto, y = p)) +
  geom_point(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Gols (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 45, by = 5)) +
  ylim(0, 0.0155) +
  geom_abline(intercept = lm1$coefficients[1], slope = lm1$coefficients[2],
              col = "red", size = 1)
```



```r
t2 = tmp %>%
  filter(Minuto < 45, Tempo == "2º")

lm2 = lm(p ~ Minuto, data = t2)

summary(lm2)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2,352e-03 -1,275e-03  1,411e-05  1,069e-03  2,877e-03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1,082e-02  4,349e-04  24,879   <2e-16 ***
## Minuto      1,905e-05  1,683e-05   1,132    0,264
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,001418 on 42 degrees of freedom
## Multiple R-squared:  0,0296, Adjusted R-squared:  0,006491
## F-statistic: 1,281 on 1 and 42 DF,  p-value: 0,2641
```

```r
t2 %>%
  ggplot(aes(x = Minuto, y = p)) +
  geom_point(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Gols (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 45, by = 5)) +
  ylim(0, 0.0155) +
  geom_abline(intercept = lm2$coefficients[1], slope = lm2$coefficients[2],
              col = "red", size = 1)
```

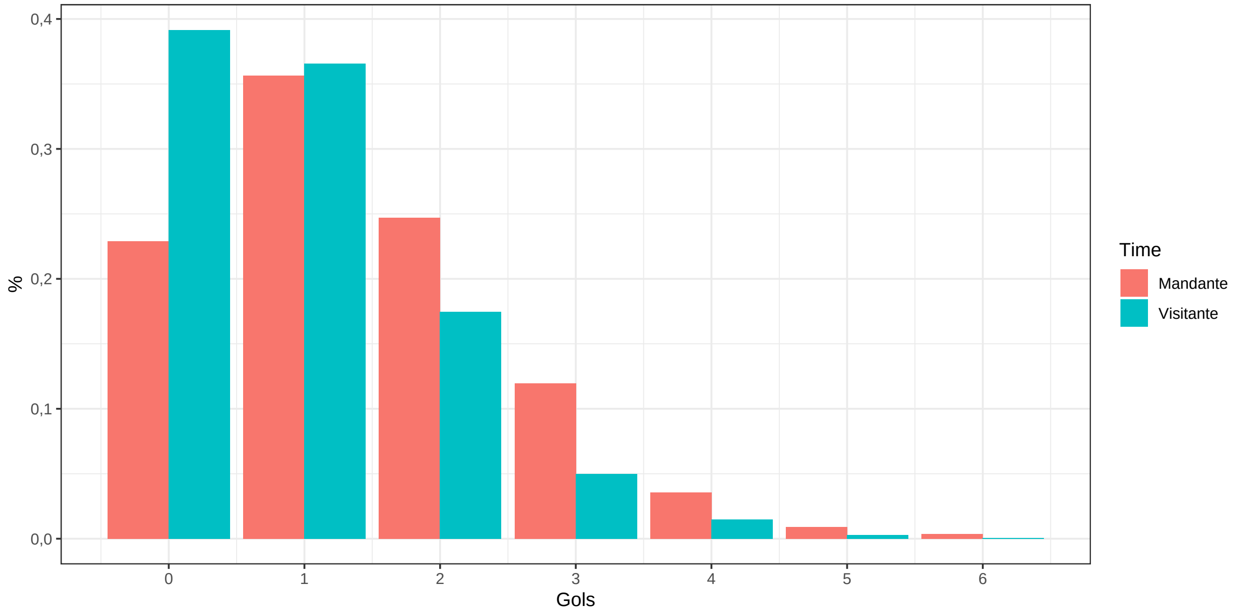


Placares mais comuns

```
resultados %>%
  count(Placar_1, Placar_2) %>%
  arrange(desc(n))
```

```
## # A tibble: 37 x 3
##    Placar_1 Placar_2     n
##       <int>    <int> <int>
##  1        1        0   340
##  2        1        1   282
##  3        2        1   226
##  4        2        0   208
##  5        0        0   202
##  6        0        1   192
##  7        1        2   139
##  8        3        0   104
##  9        2        2   101
## 10        3        1    91
## # ... with 27 more rows
```

```
mandante = resultados %>%
  count(Placar_1) %>%
  na.omit() %>%
  mutate(Time = "Mandante") %>%
  rename(Placar = Placar_1)

visitante = resultados %>%
  count(Placar_2) %>%
  na.omit() %>%
  mutate(Time = "Visitante") %>%
  rename(Placar = Placar_2)

tmp = rbind(mandante, visitante) %>%
  mutate(p = n/(nrow(resultados) - 1))

tmp %>%
  ggplot(aes(fill = Time, y = p, x = Placar)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw() +
  xlab("Gols") +
  ylab("%") +
  scale_x_continuous(breaks = 0:6)
```

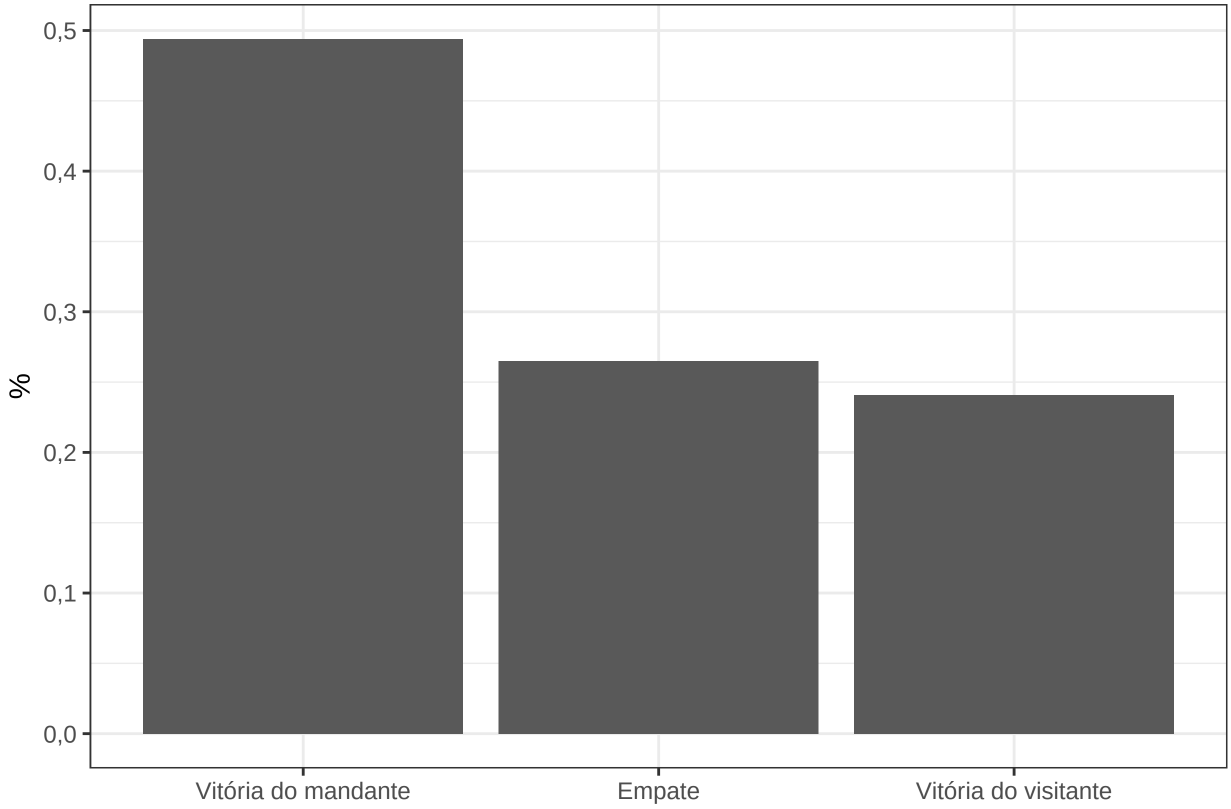


Resultados

```
tmp = resultados %>%
  mutate(resultado = ifelse(Placar_1 == Placar_2, "Empate",
                            ifelse(Placar_1 > Placar_2, "Vitória do mandante",
                                   "Vitória do visitante"))) %>%
  count(resultado) %>%
  arrange(desc(n)) %>%
  mutate(p = n/(nrow(resultados)))
tmp
```

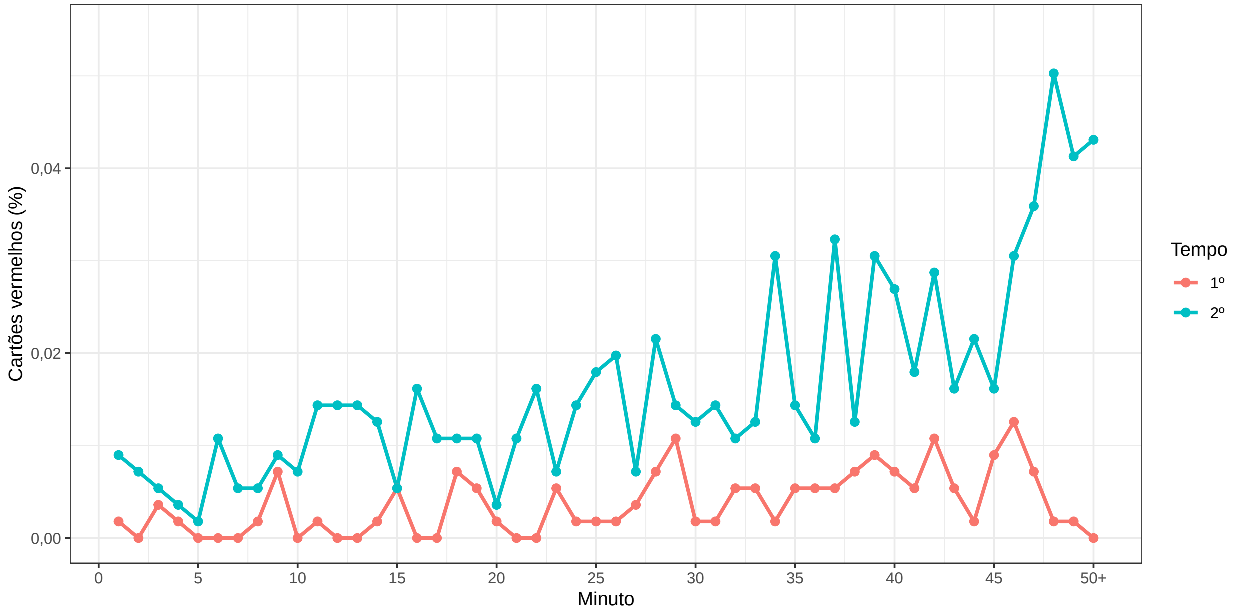
```
## # A tibble: 3 x 3
##   resultado                n     p
##   <chr>                <int> <dbl>
## 1 Vitória do mandante   1126 0.494
## 2 Empate                 604 0.265
## 3 Vitória do visitante   549 0.241
```

```
tmp %>%
  mutate() %>%
  na.omit() %>%
  mutate(resultado = factor(resultado, levels = c("Vitória do mandante",
                                                  "Empate", "Vitória do visitante"))) %>%
  ggplot(aes(x = resultado, y = p)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw() +
  xlab("") +
  ylab("%")
```



Cartões vermelhos por minuto

```
reds$Acréscimo[which(is.na(reds$Acréscimo))] = 0

reds = reds %>%
  mutate(Minuto = Minuto + Acréscimo)

reds$Minuto[which(reds$Minuto > 50)] = 50

tib_zeros = tibble(Minuto = c(1:50, 1:50),
                   Tempo = c(rep("1º", 50), rep("2º", 50)), n = 0L)
complete_zeros <- function(tib_count) {
  tib_count %>%
    full_join(tib_zeros, by = c("Minuto", "Tempo", "n")) %>%
    group_by(Minuto, Tempo) %>%
    summarise(n = sum(n))
}

tmp = reds %>%
  count(Minuto, Tempo) %>%
  complete_zeros() %>%
  mutate(p = n/nrow(reds))

tmp %>%
  ggplot(aes(x = Minuto, y = p, col = Tempo)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Cartões vermelhos (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 50, by = 5),
                     labels = c(seq(from = 0, to = 45, by = 5), "50+")) +
  ylim(0, 0.055)
```

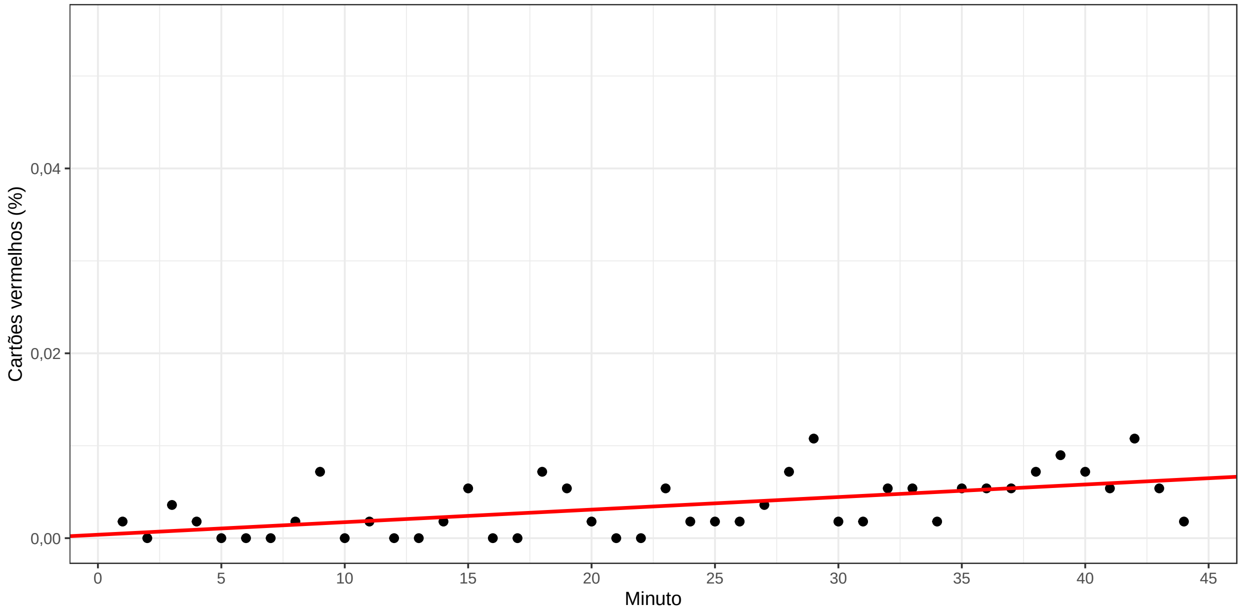


```
t1 = tmp %>%
  filter(Minuto < 45, Tempo == "1º")

lm1 = lm(p ~ Minuto, data = t1)

summary(lm1)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t1)
##
```

```
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -0,0045536 -0,0019790 -0,0004627  0,0014439  0,0064613 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3,701e-04  7,925e-04   0,467    0,643    
## Minuto      1,359e-04  3,068e-05   4,430 6,62e-05 ***
## ---
## Signif. codes:  0 ’***’ 0,001 ’**’ 0,01 ’*’ 0,05 ’.’ 0,1 ’ ’ 1
## 
## Residual standard error: 0,002584 on 42 degrees of freedom
## Multiple R-squared:  0,3184, Adjusted R-squared:  0,3022 
## F-statistic: 19,62 on 1 and 42 DF,  p-value: 6,619e-05
```

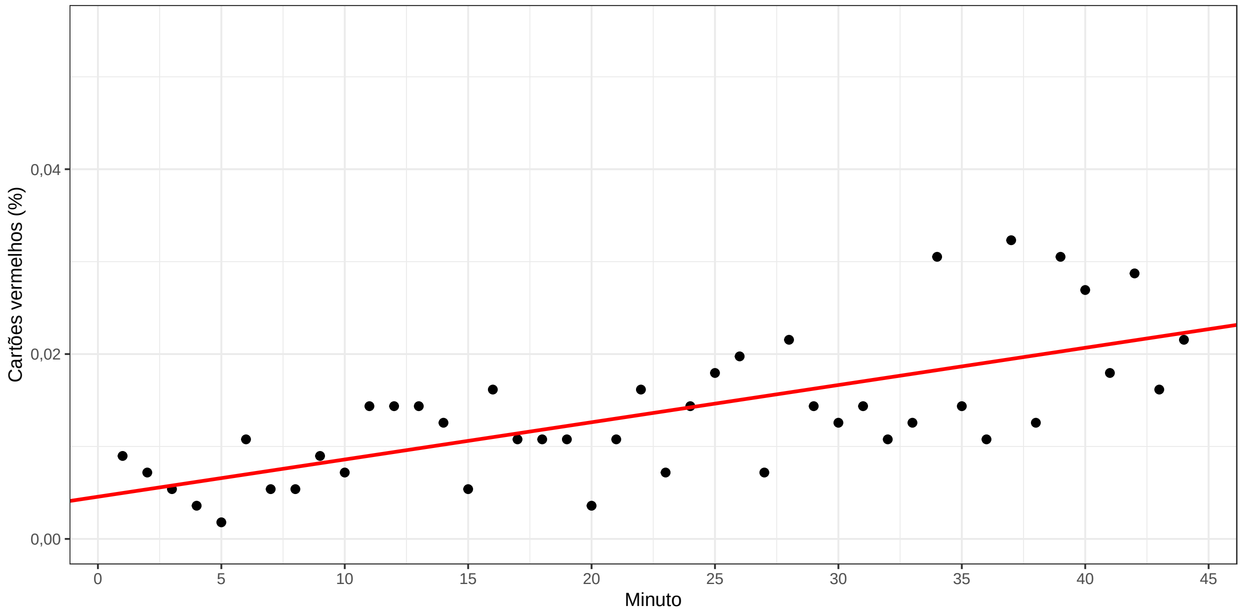
```
t1 %>%
  ggplot(aes(x = Minuto, y = p)) +
  geom_point(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Cartões vermelhos (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 45, by = 5)) +
  ylim(0, 0.055) +
  geom_abline(intercept = lm1$coefficients[1], slope = lm1$coefficients[2],
              col = "red", size = 1)
```



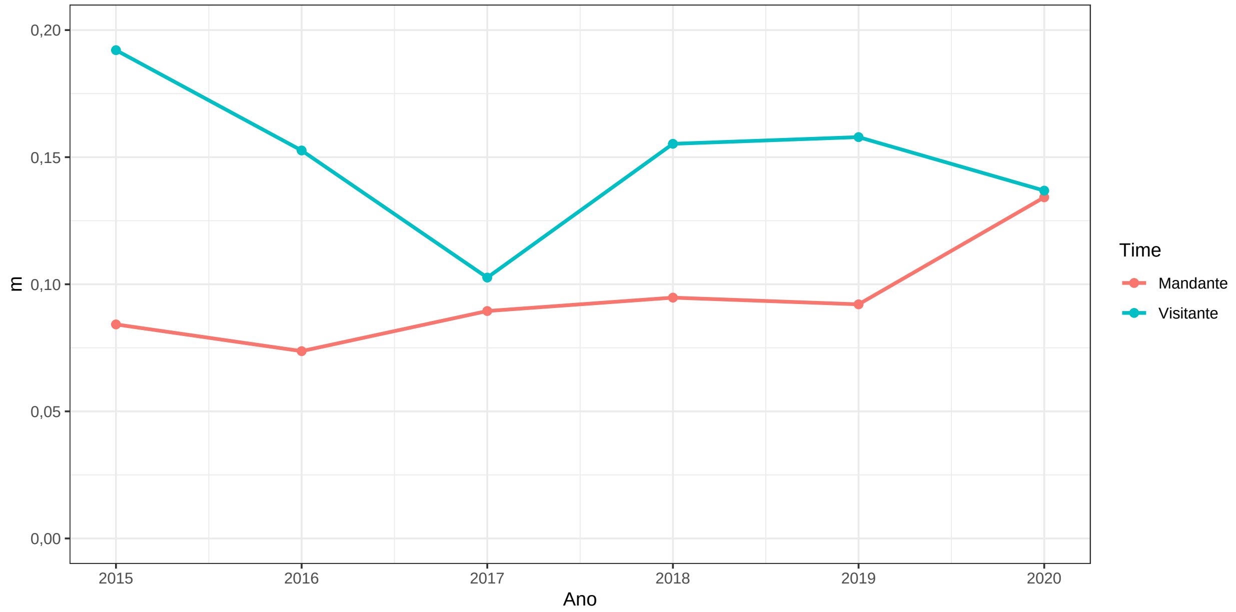
```
t2 = tmp %>%
  filter(Minuto < 45, Tempo == "2º")

lm2 = lm(p ~ Minuto, data = t2)

summary(lm2)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0,0090304 -0,0041369 -0,0008944  0,0041475  0,0128466
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4,564e-03  1,694e-03   2,694   0,0101 *
## Minuto      4,028e-04  6,556e-05   6,144 2,46e-07 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,005523 on 42 degrees of freedom
## Multiple R-squared:  0,4734, Adjusted R-squared:  0,4608
## F-statistic: 37,75 on 1 and 42 DF,  p-value: 2,463e-07
```

```
t2 %>%
  ggplot(aes(x = Minuto, y = p)) +
  geom_point(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Cartões vermelhos (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 45, by = 5)) +
  ylim(0, 0.055) +
  geom_abline(intercept = lm2$coefficients[1], slope = lm2$coefficients[2],
              col = "red", size = 1)
```

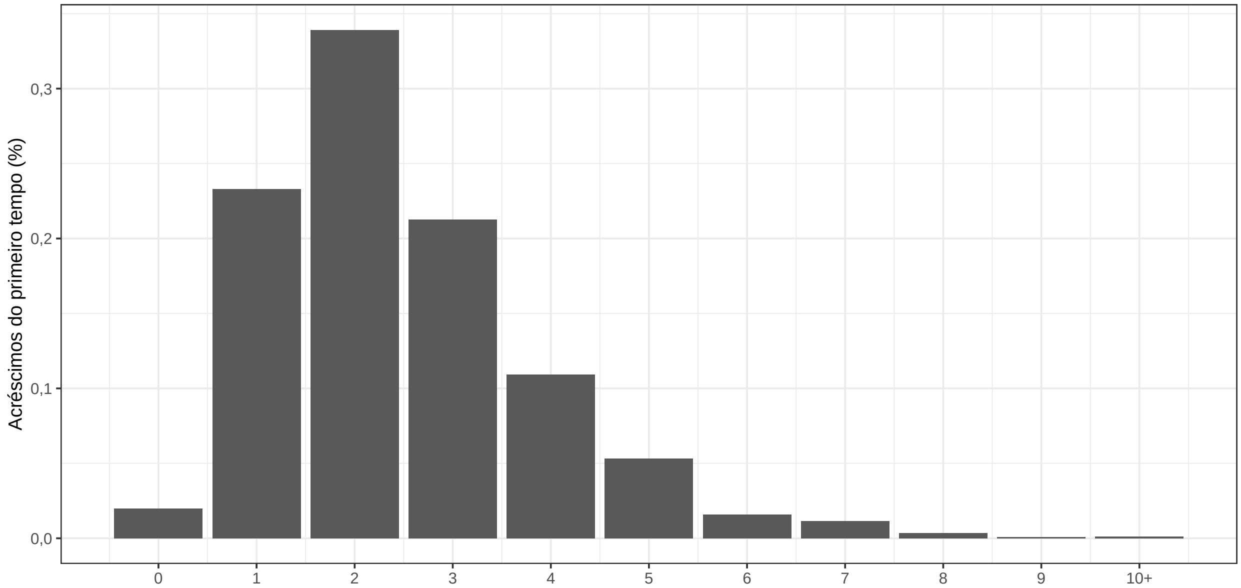


```
reds %>%
  count(Ano, Time) %>%
  mutate(m = n/380) %>%
  ggplot(aes(x = Ano, y = m, col = Time)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  scale_x_continuous(breaks = 2015:2020) +
  ylim(0, 0.2)
```

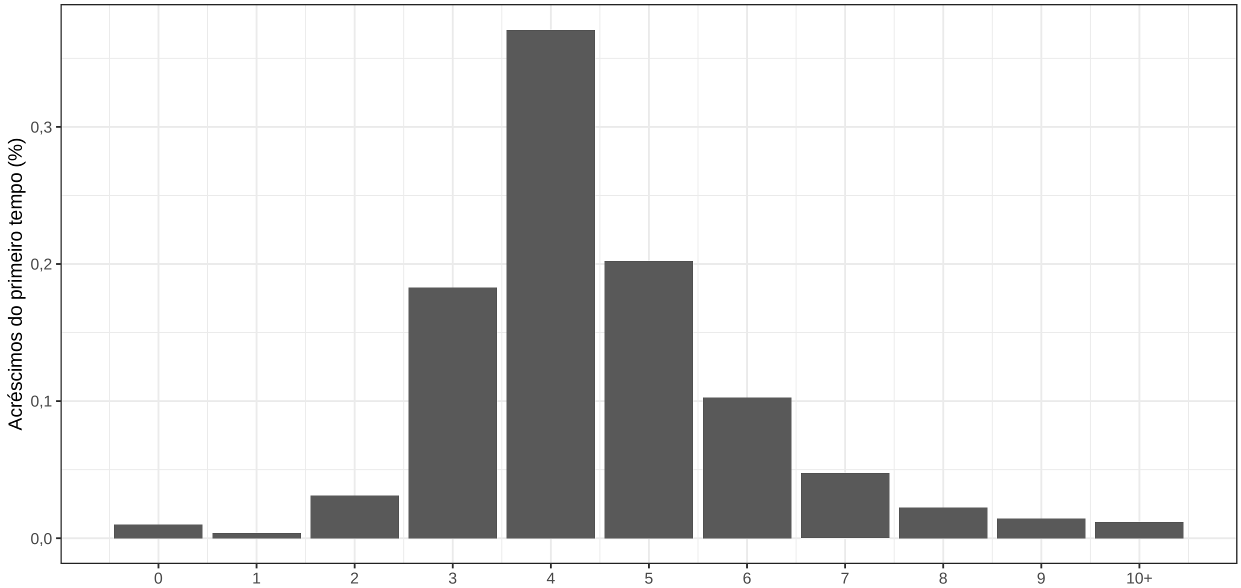


Acréscimos

```
resultados$Acréscimos_1[which(resultados$Acréscimos_1 > 10)] = 10
resultados %>%
  count(Acréscimos_1) %>%
  mutate(p = n/nrow(resultados)) %>%
  ggplot(aes(x = Acréscimos_1, y = p)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw() +
  xlab("") +
  ylab("Acréscimos do primeiro tempo (%)") +
  scale_x_continuous(breaks = 0:10,
                     labels = c(0:9, "10+"))
```



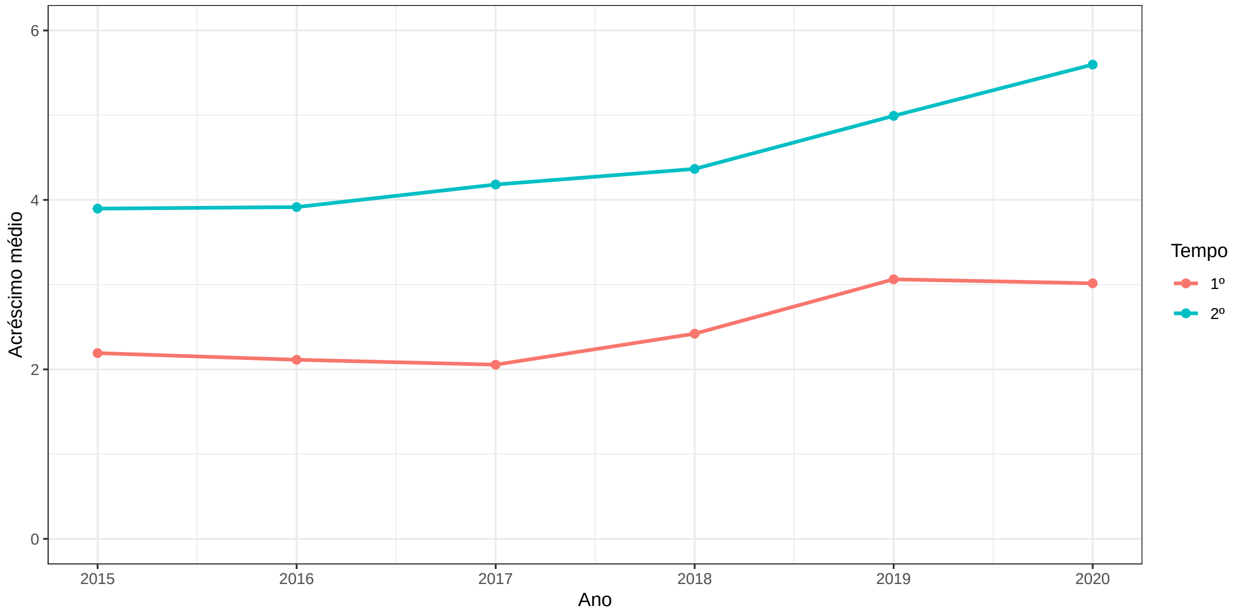
```r
resultados$Acréscimos_2[which(resultados$Acréscimos_2 > 10)] = 10
resultados %>%
  count(Acréscimos_2) %>%
  mutate(p = n/nrow(resultados)) %>%
  ggplot(aes(x = Acréscimos_2, y = p)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw() +
  xlab("") +
  ylab("Acréscimos do primeiro tempo (%)") +
  scale_x_continuous(breaks = 0:10,
                     labels = c(0:9, "10+"))
```



Acréscimo médio por ano

```
medias = results %>%
  rename(Ano = Season) %>%
  group_by(Ano) %>%
  summarise(Acréscimos_1 = mean(Stoppage_Time_1),
            Acréscimos_2 = mean(Stoppage_Time_2))
```

```
tibble(Tempo = c(rep("1º", nrow(medias)), rep("2º", nrow(medias))),
       Ano = rep(medias$Ano, 2),
       Média = c(medias$Acréscimos_1, medias$Acréscimos_2)) %>%
  ggplot(aes(x = Ano, y = Média, col = Tempo)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Acréscimo médio") +
  ylim(0, 6)
```


Acréscimo médio
Ano
Tempo
1º
2º