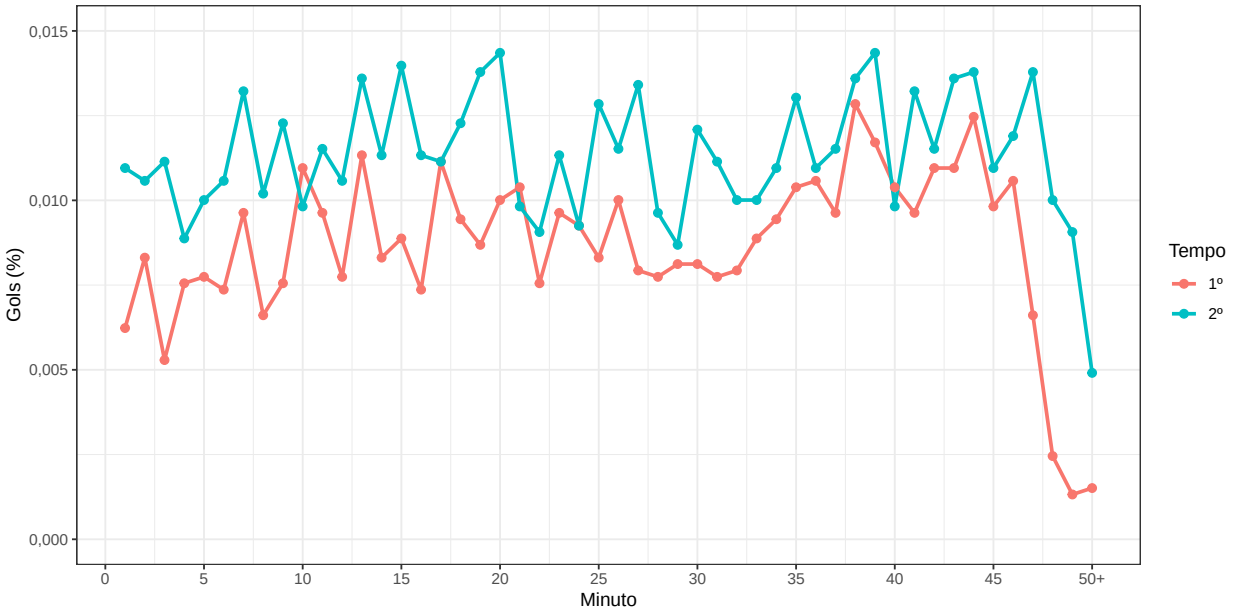
EDA Série A 2014-2019

```
options(OutDec = ",")

library(dplyr)
library(ggplot2)

load("data/gols.RData")
load("data/resultados.RData")
load("data/reds.RData")

resultados = resultados %>%
  filter(Ano >= 2014, Campeonato == "Campeonato Brasileiro Série A")

gols = gols %>%
  filter(Ano >= 2014, Campeonato == "Campeonato Brasileiro Série A")

reds = reds %>%
  filter(Campeonato == "Campeonato Brasileiro Série A")
```

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, 0.015)
```

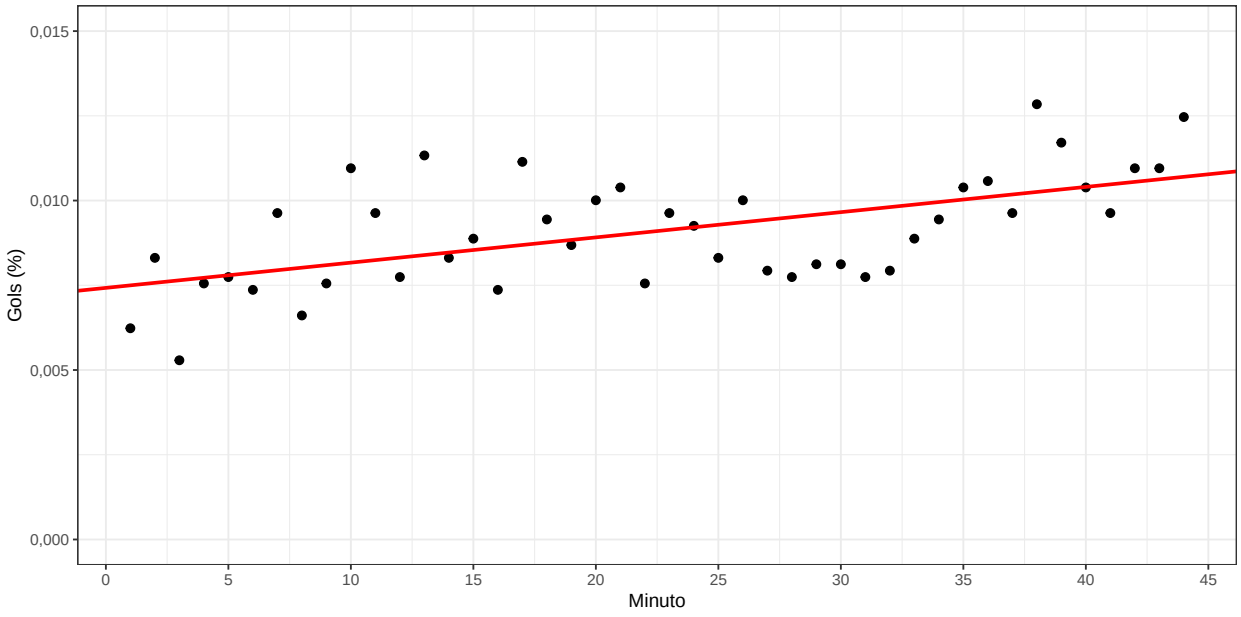



```
t1 = tmp %>%
  filter(Minuto < 45, Tempo == "1º")

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

summary(lm1)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0,0023592 -0,0010664 -0,0001022  0,0006932  0,0029395
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7,424e-03  4,207e-04  17,647  < 2e-16 ***
## Minuto      7,448e-05  1,628e-05   4,574 4,19e-05 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,001372 on 42 degrees of freedom
## Multiple R-squared:  0,3325, Adjusted R-squared:  0,3166
## F-statistic: 20,92 on 1 and 42 DF,  p-value: 4,19e-05
```

```
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, 0.015) +
  geom_abline(intercept = lm1$coefficients[1], slope = lm1$coefficients[2], col = "red", size = 1)
```

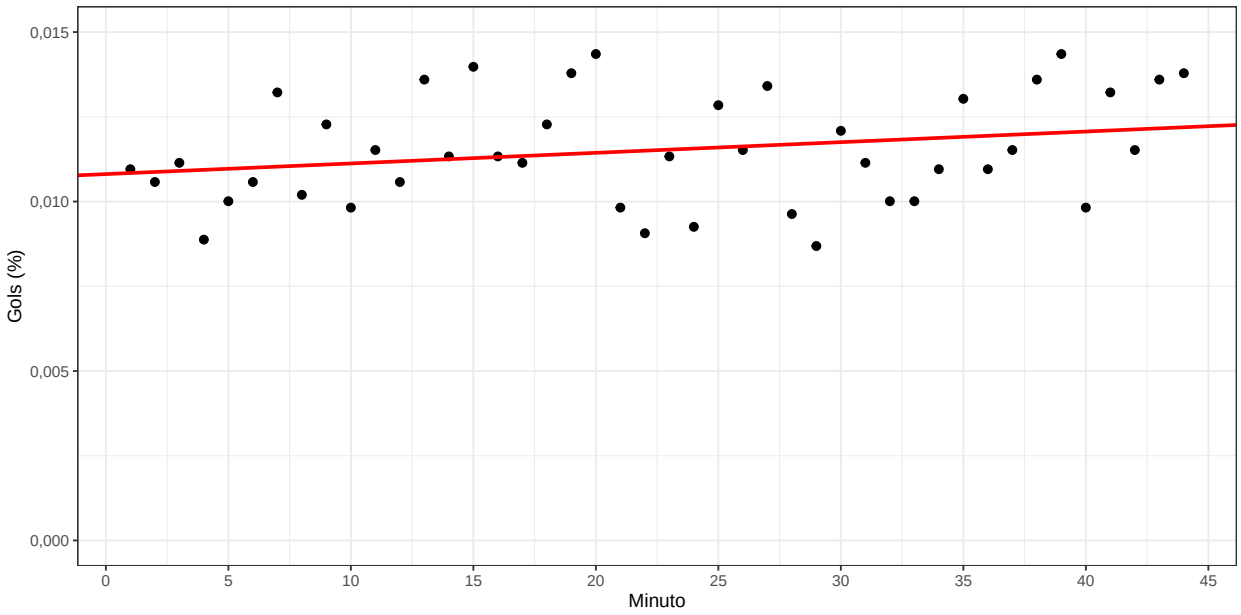



```
t2 = tmp %>%
  filter(Minuto < 45, Tempo == "2º")

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

summary(lm2)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0,0030327 -0,0009644 -0,0001529  0,0011996  0,0029156
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1,081e-02  4,834e-04  22,361   <2e-16 ***
## Minuto      3,140e-05  1,871e-05   1,678    0,101
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,001576 on 42 degrees of freedom
## Multiple R-squared:  0,06283,    Adjusted R-squared:  0,04051
## F-statistic: 2,816 on 1 and 42 DF,  p-value: 0,1008
```

```
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, 0.015) +
  geom_abline(intercept = lm2$coefficients[1], slope = lm2$coefficients[2], col = "red", size = 1)
```

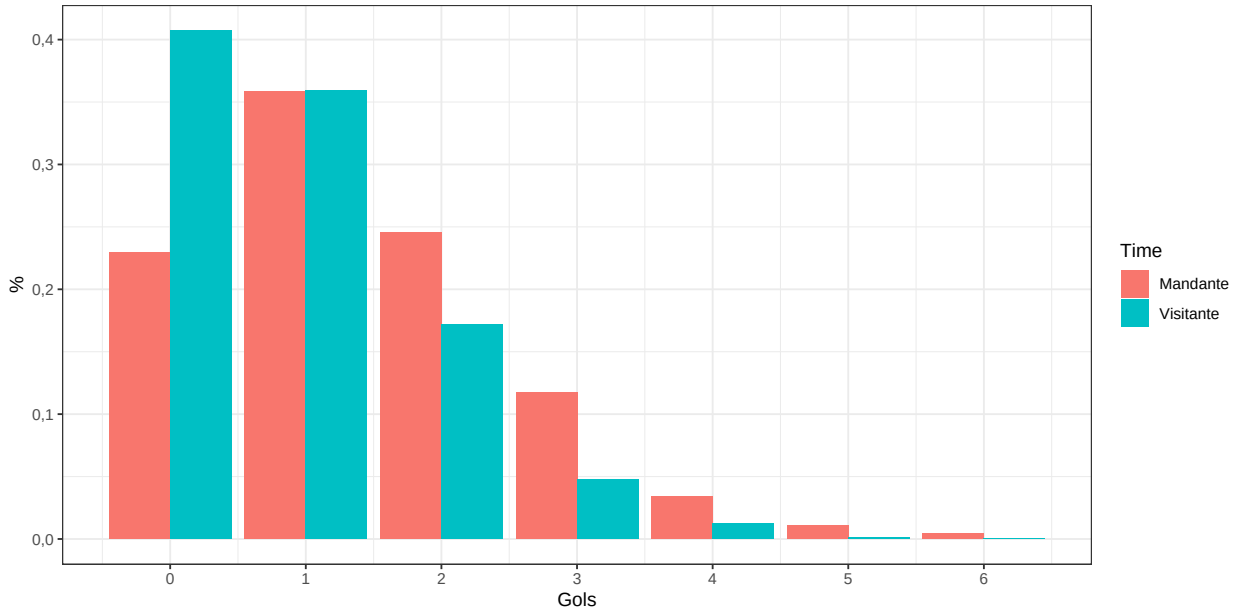


Placares mais comuns

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

```
## # A tibble: 36 x 3
##    Placar_1 Placar_2     n
##       <int>    <int> <int>
##  1        1        0   363
##  2        1        1   271
##  3        2        0   217
##  4        2        1   213
##  5        0        1   204
##  6        0        0   202
##  7        1        2   133
##  8        3        0   106
##  9        2        2   100
## 10        3        1    91
## # ... with 26 more rows
```

```r
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)
```

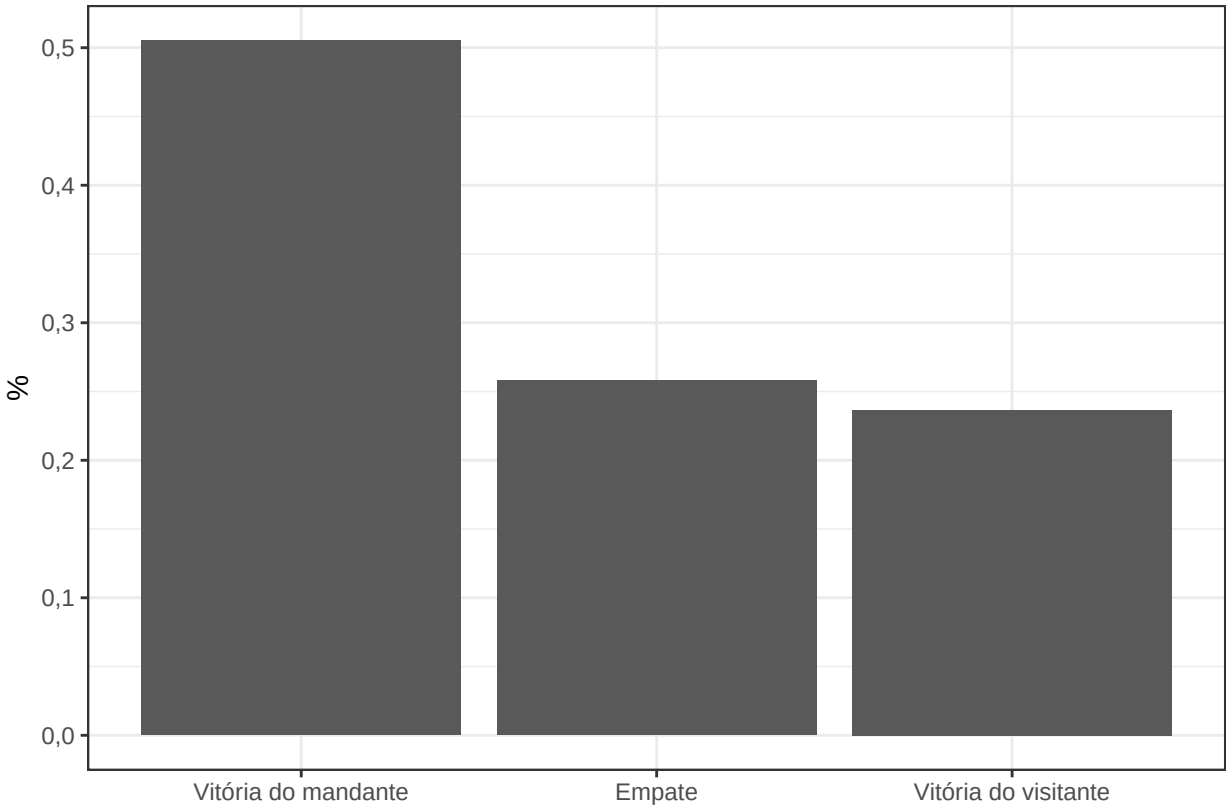


Resultados

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

```
## # A tibble: 4 x 3
##   resultado                n        p
##   <chr>                <int>    <dbl>
## 1 Vitória do mandante   1152 0.505
## 2 Empate                 588 0.258
## 3 Vitória do visitante   539 0.236
## 4 <NA>                     1 0.000439
```

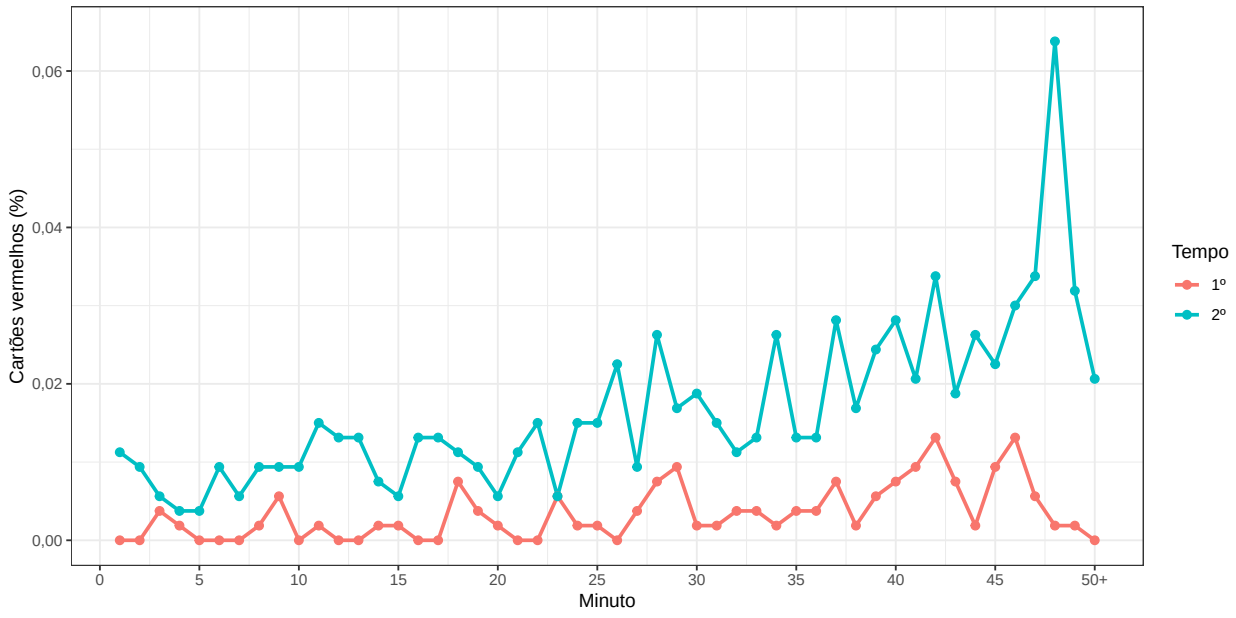
```
tmp %>%
  mutate() %>%
  na.omit() %>%
  mutate(resultado = factor(resultado, levels = c("Vitória do mandante", "Empate", "Vitória do visitante
  ggplot(aes(x = resultado, y = p)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw() +
  xlab("") +
  ylab("%")
```



Cartões vermelhos por minuto

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

reds = reds %>%
```

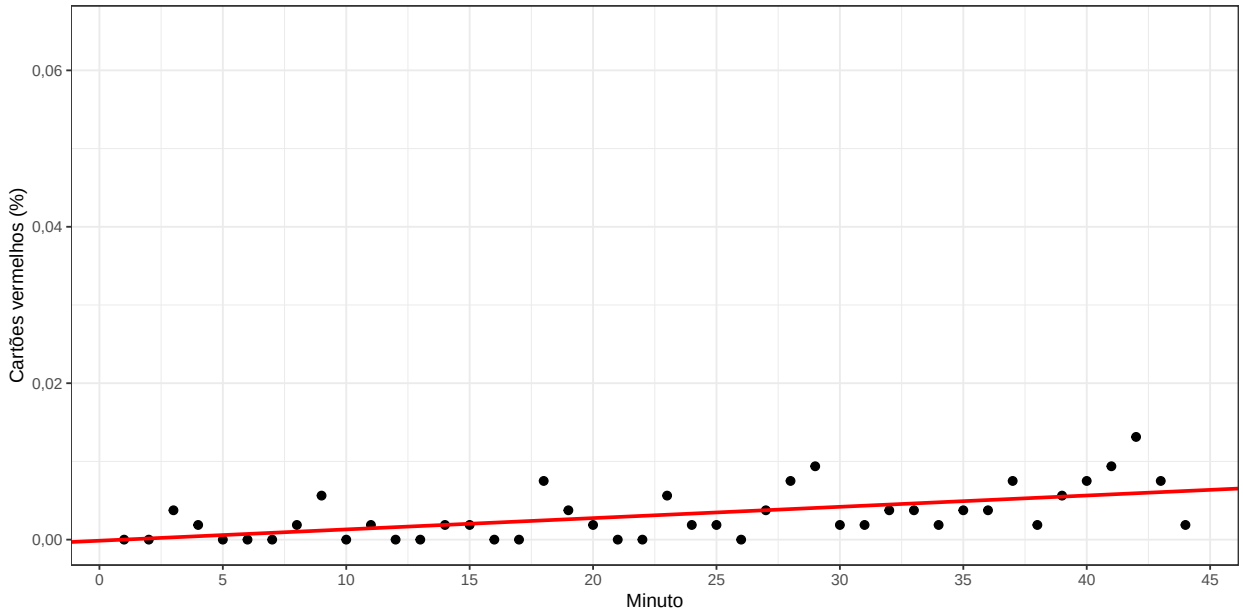
```
  mutate(Minuto = Minuto + Acréscimo)

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, 0.065)
```




```
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
## -0,004344 -0,001633 -0,000656  0,001431  0,007202
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1,388e-04  8,128e-04  -0,171    0,865
## Minuto       1,445e-04  3,146e-05   4,593 3,94e-05 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,00265 on 42 degrees of freedom
## Multiple R-squared:  0,3344, Adjusted R-squared:  0,3185
## F-statistic:  21,1 on 1 and 42 DF,  p-value: 3,939e-05
```

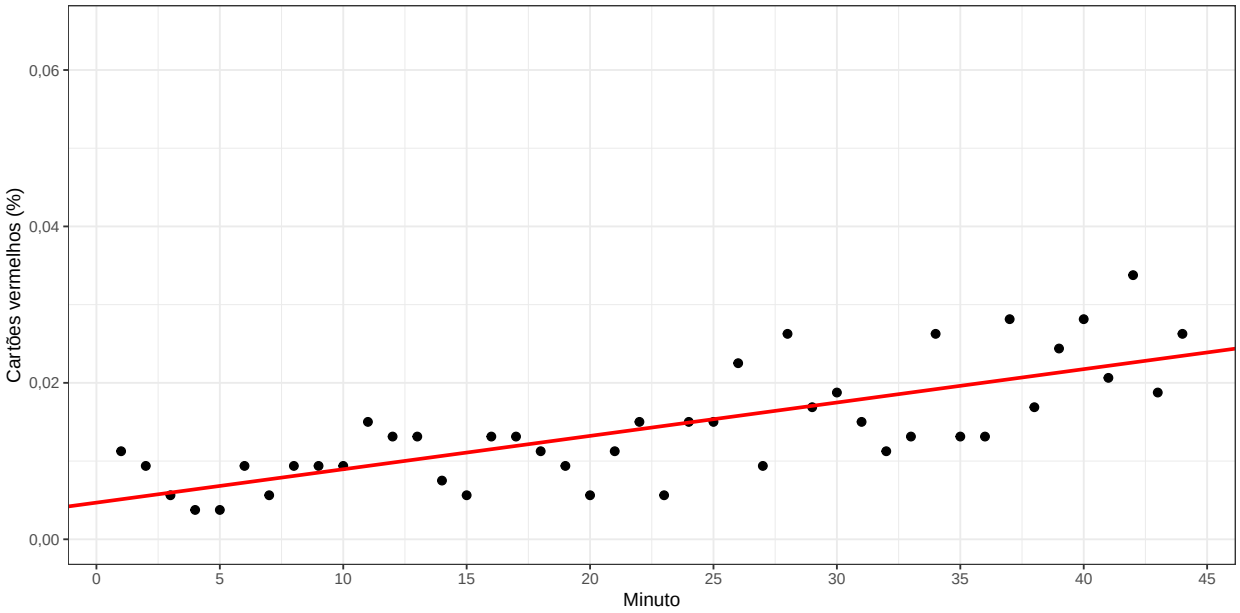
```
t1 %>%
  ggplot(aes(x = Minuto, y = p)) +
  geom_point(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  ylab("Cartões vermelhos (%)") +
  scale_x_continuous(breaks = seq(from = 0, to = 45, by = 5)) +
  ylim(0, 0.065) +
  geom_abline(intercept = lm1$coefficients[1], slope = lm1$coefficients[2], col = "red", size = 1)
```



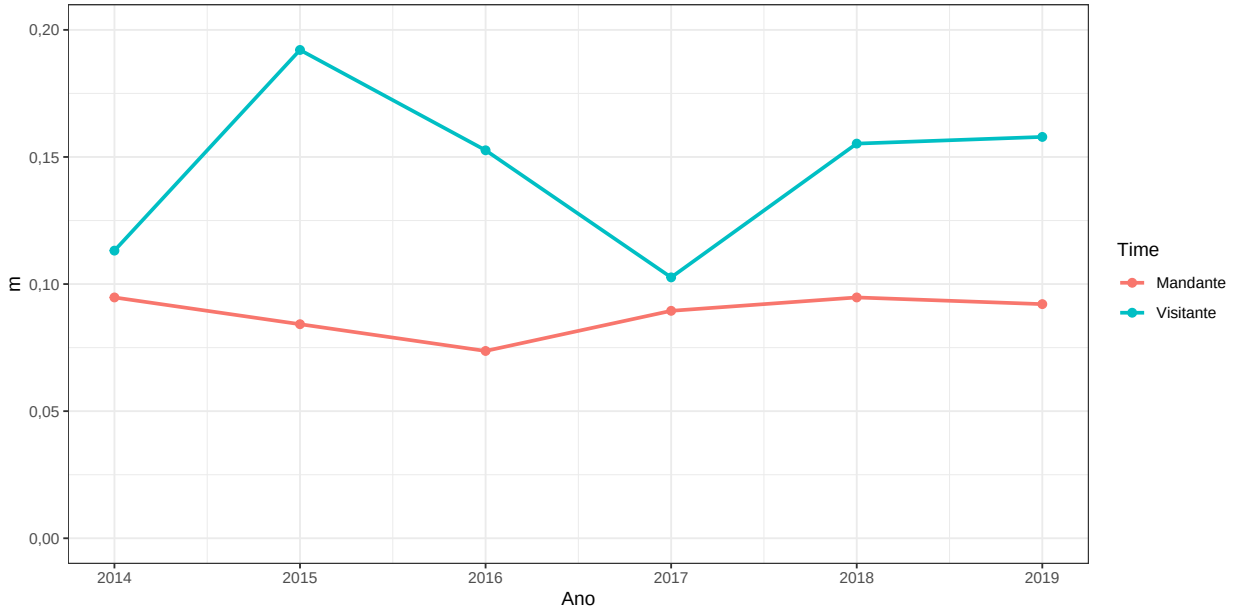
```r
t2 = tmp %>%
  filter(Minuto < 45, Tempo == "2º")

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

summary(lm2)
```

```
##
## Call:
## lm(formula = p ~ Minuto, data = t2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0,0088693 -0,0032174 -0,0000437  0,0029430  0,0111666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4,684e-03  1,516e-03   3,090  0,00355 **
## Minuto      4,267e-04  5,868e-05   7,271 5,97e-09 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,004943 on 42 degrees of freedom
## Multiple R-squared:  0,5573, Adjusted R-squared:  0,5467
## F-statistic: 52,87 on 1 and 42 DF,  p-value: 5,968e-09
```

```r
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, 0.065) +
  geom_abline(intercept = lm2$coefficients[1], slope = lm2$coefficients[2], col = "red", size = 1)
```

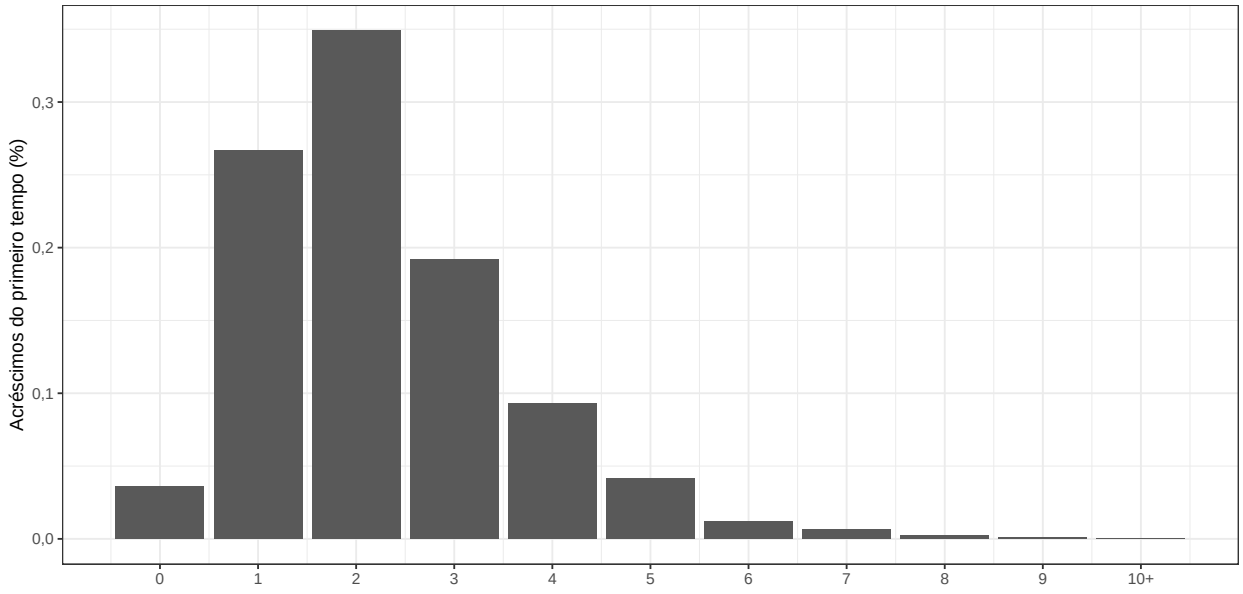



```
reds %>%
  count(Ano, Time) %>%
  mutate(m = n/380) %>%
  ggplot(aes(x = Ano, y = m, col = Time)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  scale_x_continuous(breaks = 2014:2019) +
  ylim(0, 0.2)
```

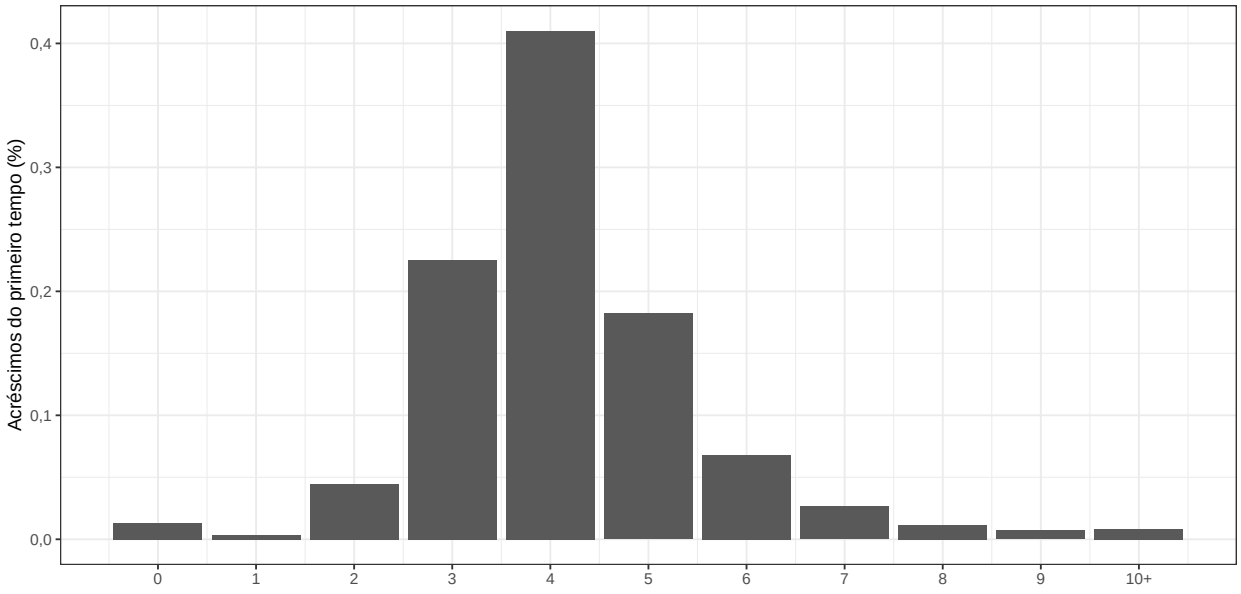


Acréscimos

```
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+"))
```



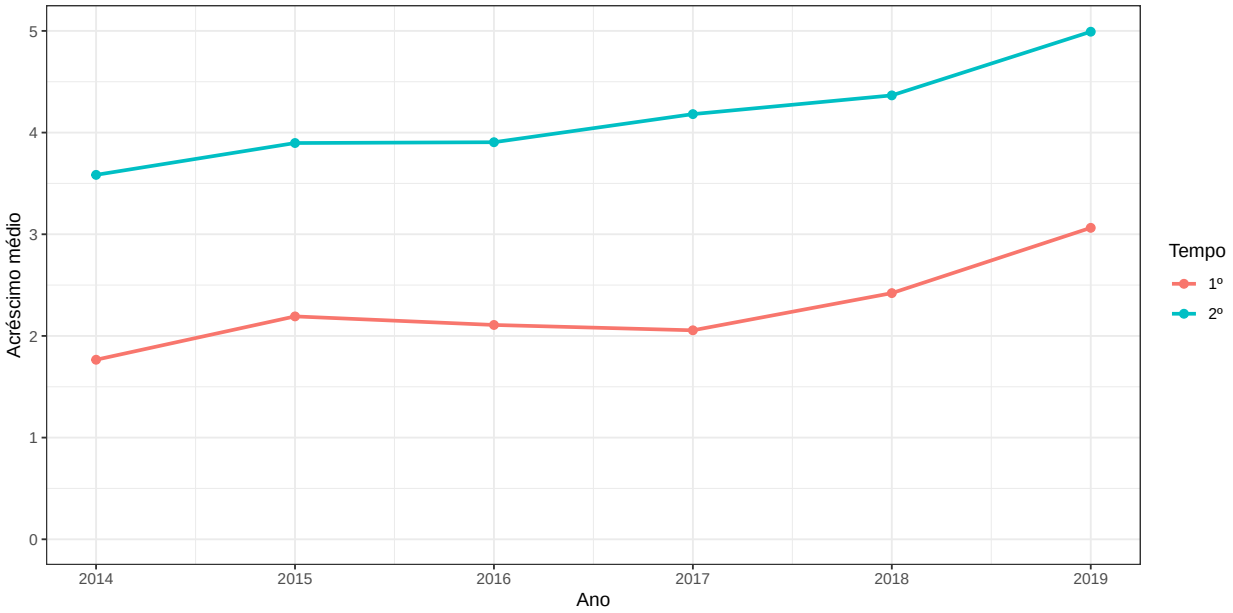

```
resultados$Acréscimos_2[which(resultados$Acréscimos_2 > 10)] = 10
resultados %>%
  count(Acréscimos_2) %>%
  mutate(p = n/nrow(resultados)) %>%
  ggplot(aes(x = Acréscimos_2, y = p)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw() +
  xlab("") +
  ylab("Acréscimos do primeiro tempo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



Acréscimo médio por ano

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

medias = resultados %>%
  filter(Ano >= 2014, Campeonato == "Campeonato Brasileiro Série A") %>%
  group_by(Ano) %>%
  summarise(Acréscimos_1 = mean(Acréscimos_1),
            Acréscimos_2 = mean(Acréscimos_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, 5)
```


Acréscimo médio
5
4
3
2
1
0
2014
2015
2016
2017
2018
2019
Ano
Tempo
1º
2º