INSPER INSTITUTO DE ENSINO E PESQUISA

APS 1 (ANÁLISE DE DADOS) - ESTATÍSTICA I

Bárbara Naves Barbiero;

Bruno Weber Maurer;

Maria Eduarda Miranda;

Victor Manoel Alves.

|  |
| --- |
| São Paulo - SP  2021 |

Bárbara Naves Barbiero, Bruno Weber Maurer, MARIA EDUARDA MIRANDA E Victor Manoel Alves.

APS (ANÁLISE DE DADOS) - ESTATÍSTICA I

Atividade Prática Supervisionada apresentado ao curso de Administração e Ciências Econômicas do Insper, como requisito parcial para a Aprovação na matéria.

|  |
| --- |
| São Paulo - SP  2021 |

Lista de ilustrações

[Gráfico 1 — Quantidade de Resultados por Temporada (2012-2020). 6](#_30j0zll)

[Gráfico 2 — Distribuição da Média de Gols da Casa e do Visitante por Temporada (2012-2020). 7](#_1fob9te)

[Gráfico 3 — Distribuição de resultados nos dias mais quentes e mais frios de 2012 e 2014 8](#_3znysh7)

[Gráfico 4 — Distribuição de resultados nos dias mais quentes e mais frios de 2016 e 2018 9](#_2et92p0)

[Gráfico 5 — Distribuição da média do total de gols por temporada e clima 10](#_tyjcwt)

[Gráfico 6 — Média da soma de gols por temporada no frio por período 11](#_3dy6vkm)

[Gráfico 7 — Distribuição da soma de gols por período do dia (1) 11](#_1t3h5sf)

[Gráfico 8 — Distribuição da soma de gols por período do dia (2) 12](#_4d34og8)

[Gráfico 9 — Distribuição da soma de gols por período do dia 13](#_2s8eyo1)

[Gráfico 10 — Soma dos gols por estado e clima em 2016 e 2018 14](#_17dp8vu)

[Gráfico 11 — Soma dos gols por estado e clima em 2012 e 2014. 14](#_3rdcrjn)

[Gráfico 12 — Distribuição da quantidade total de gols por estado e período (1) 15](#_26in1rg)

[Gráfico 13 — Distribuição da quantidade total de gols por estado e período (2) 16](#_lnxbz9)

[Gráfico 14 — Distribuição das partidas por estado 16](#_35nkun2)

**Sumário**

[1. INTRODUÇÃO 4](#_Toc68885376)

[2. DESENVOLVIMENTO 5](#_Toc68885377)

[2.1 AS EQUIPES MANDANTES FORAM MAIS PREJUDICADAS EM 2020? 6](#_Toc68885378)

[2.2 O CLIMA, EM ESPECIAL O CALOR, REDUZ A QUALIDADE DOS JOGOS? 8](#_Toc68885379)

[2.3 EM QUAL CENÁRIO É MAIS PROVÁVEL DE OCORRER UM JOGO EMOCIONANTE? 16](#_Toc68885380)

1. **INTRODUÇÃO**

Para realizar a análise dos dados presentes na base de dados nominada “BRA2” foi utilizado a linguagem de programação estatística R e os conhecimentos de análise de dados adquiridos durante as aulas e monitorias. Além do R base, foram utilizados pacotes adicionais, como “*tidyverse*”, “*gridExtra*”, “*janitor*” e outros pacotes estéticos.

A base de dados “BRA2” é constituída por 14 variáveis e 3420 entradas (linhas). Em toda a análise, foi utilizado 10 das 14 variáveis. Além disso, para a melhor compreensão deste trabalho, é necessário definir as seguintes variáveis:

país: o país no qual a partida foi realizada.

estado: o estado no qual a partida foi realizada.

temporada: a temporada em que a partida foi realizada.

data: a data da partida.

período: o período do dia em que a partida foi realizada (Noite ou Dia).

casa: o time mandante da partida.

visitante: o time visitante da partida.

golcasa: a quantidade de gols realizados pelo mandante na partida.

golvisitante: a quantidade de gols realizados pelo visitante na partida.

total: a soma de golcasa e golvisitante.

res: o resultado da partida (E = empate, C = vitória do mandante, V = vitória do visitante).

pc: probabilidade de vitória do mandante.

pe: probabilidade de empate

pv: probabilidade de vitória do visitante.

1. **DESENVOLVIMENTO**

Antes de mais nada, foi necessário importar os dados com o pacote *readxl*, além disso, a base de dados foi manipulada com os pacotes do pacote *tidyverse* para a obtenção de tabelas, gráficos e outros *datasets,* sendo também manipulados com o pacote *janitor* para que o nome das variáveis fosse de maiúsculas para minúsculas.

Em seguida, da base de dados original foram criados diversos objetos, sejam *datasets* de um corte de variáveis, *dataset* para a contagem de frequência, objetos gráficos ou vetores.

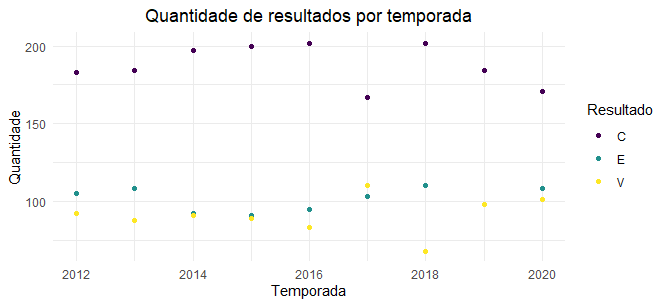
No que tange à criação dos gráficos, quando utilizados para comparação, todos foram transformados para o mesmo tamanho do eixo y para que não houvesse algum tipo de confusão na visualização. Os gráficos de barras, colunas e histogramas são por padrão agrupados um em cima do outro, para evitar interpretações erradas todos eles foram colocados na mesma escala. Por exemplo, se há um gráfico dividido em duas categorias, a contagem de cada categoria encontra-se entre o valor mínimo da contagem entre as duas e o valor máximo entre as duas categorias. Por motivos de simplificação, foram escolhidos alguns anos para serem expostos no relatório. É importante ressaltar que o ano de 2020 foi fora do padrão das outras temporadas, apresentando diferente distribuição de jogos durante os meses e estações.

Quando analisados por time visitante e temporada ou time mandante e temporada, filtrando apenas um time, nota-se que a quantidade de jogos é de 38 por ano e time. Sendo assim, é possível concluir que a base de dados representa, provavelmente, o Campeonato Brasileiro de Futebol (Brasileirão).

* 1. AS EQUIPES MANDANTES FORAM MAIS PREJUDICADAS EM 2020?

Para que essa possível responder tal questionamento, foram criados *datasets* com as medidas de dispersão e posição média dos resultados por ano, medidas de dispersão e posição de vitórias do visitante por temporada e por time (mandante ou visitante), quantidade total de cada resultado por temporada, frequência absoluta e relativa de cada resultado por ano e por time mandante. Assim, sendo possível realizar análises gráficas e obter melhores respostas para a pergunta.

Gráfico 1 — Quantidade de Resultados por Temporada (2012-2020).



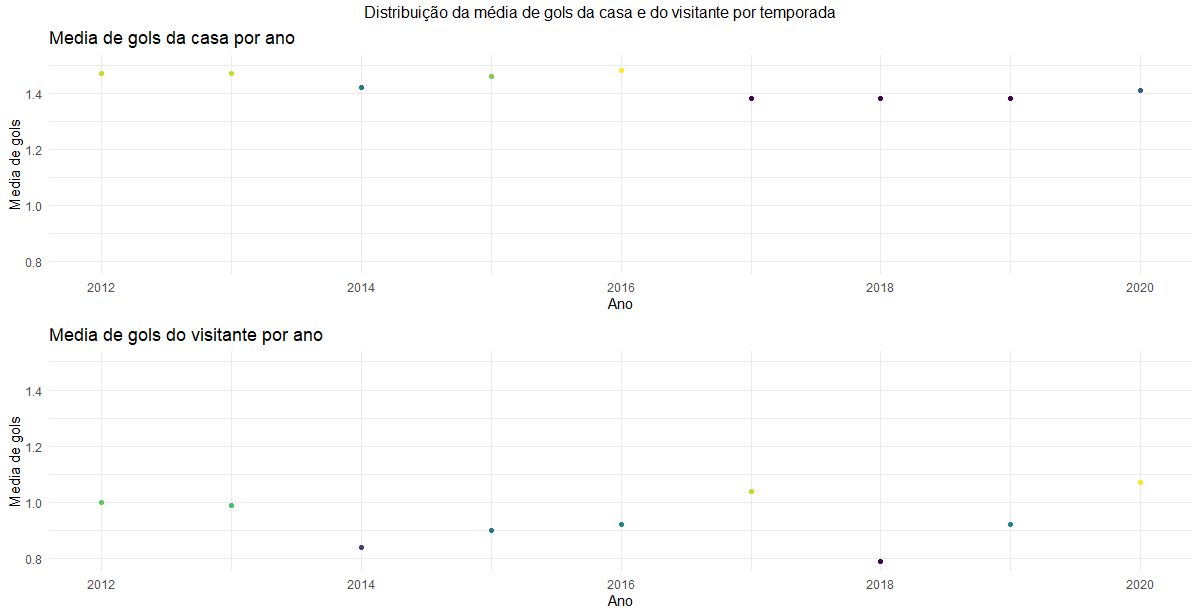
Fonte: elaborado pelos autores (2021)

Este gráfico representa os resultados dos jogos dos times (mandantes e visitantes) ao longo do tempo (período de 2012 a 2020). Sendo “C” a quantidade de vitórias do time mandante, é possível perceber que, em relação aos anos anteriores, as vitórias foram realmente menores (em uma faixa de 175 vitórias em 2020 versus uma faixa média de 200 vitórias entre 2014 a 2019). Além disso, em consequência da queda de vitórias dos mandantes, o número de vitórias dos visitantes (V) teve um aumento drástico em relação a 2018, mas em relação a 2019 o aumento foi bem menor, devido ao aumento de empates entre os times desde 2016.

 Assim, pode-se afirmar que, segundo o gráfico analisado acima, as equipes mandantes diminuíram sua performance em 2020. É possível perceber, inclusive, que desde 2014 os times mandantes estavam em uma crescente de vitórias (exceto em 2017), mas foram diminuindo ao longo do tempo até o último ano analisado (2020), obtendo uma das quantidades de vitórias mais baixas do período.

 Além de analisar os resultados gerais dos jogos, é importante observar a média de gols ao ano das equipes, para, dessa forma, perceber se ocorreu de fato uma mudança na performance dos jogadores, resultando em prejuízos ou não aos times. Assim, como demonstra o gráfico abaixo.

Gráfico 2 — Distribuição da Média de Gols da Casa e do Visitante por Temporada (2012-2020).



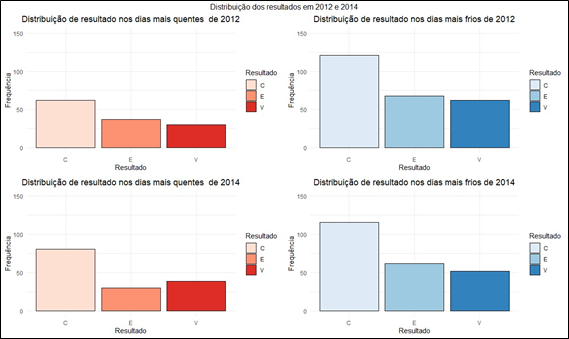
Fonte: elaborado pelos autores (2021).

Ao observar a média de gols dos times da casa em 2020 em relação a média de gols dos visitantes em 2020, pode-se perceber que os times visitantes tiveram, aparentemente, uma performance muito maior em relação aos gols do time mandante. Contudo, ao analisar numericamente os gráficos, a média de gols do time mandante (1.41) é maior que a do time visitante (próxima de 1,05). Além disso, fazendo uma análise da taxa de crescimento de 2018 até 2020, a média de gols dos times visitantes cresceu mais do que dos times mandantes. Essa taxa foi medida com a fórmula ((1.41-1.38) /1.38) para o time mandante, resultando em aproximadamente 0,02 e para os times visitantes: ((1,05-0.8) /0.8), resultando em aproximadamente 0,3.

Sendo assim, devido ao fato dos crescimentos de gols e vitórias dos times mandantes terem crescido menos que dos times visitantes, pode-se afirmar que em 2020 as equipes mandantes foram mais prejudicadas.

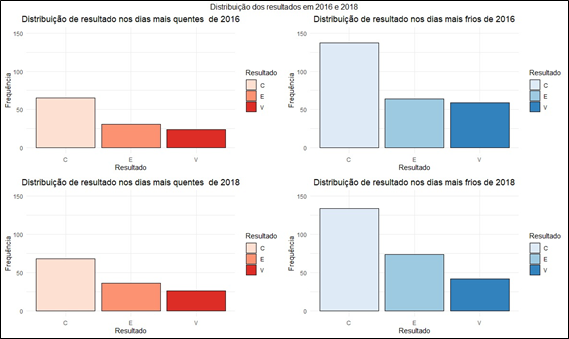
* 1. O CLIMA, EM ESPECIAL O CALOR, REDUZ A QUALIDADE DOS JOGOS?

Gráfico 3 — Distribuição de resultados nos dias mais quentes e mais frios de 2012 e 2014



Fonte: elaborado pelos autores (2021)

Gráfico 4 — Distribuição de resultados nos dias mais quentes e mais frios de 2016 e 2018.

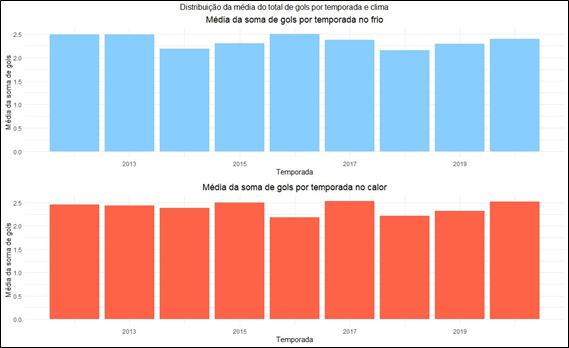


Fonte: elaborado pelos autores (2021)

Inicialmente, é importante considerar que os períodos descritos como “mais frios” e “mais quentes” no gráfico acima apontam, respectivamente, as partidas que ocorreram durante os meses de inverno/outono e primavera/verão. E, portanto, as próximas vezes em que esses termos forem utilizados na pesquisa, se referem à descrição especificada.

         Sob esse viés, de acordo com os gráficos 3 e 4, é notório que a maior frequência de qualquer resultado, seja ele vitória do time da casa, visitante ou empate, ocorrem durante os períodos frios do ano. Dessa forma, primeiramente é possível supor que a maioria dos jogos ocorrem nas estações mais frias, tornando-se necessária uma análise mais profunda sobre quantidade de dias na estação e gols que ocorreram, para chegar a alguma conclusão sobre a estação das partidas realmente impactar ou não o resultado dos jogos. Como foi realizado no gráfico abaixo.

Gráfico 5 — Distribuição da média do total de gols por temporada e clima

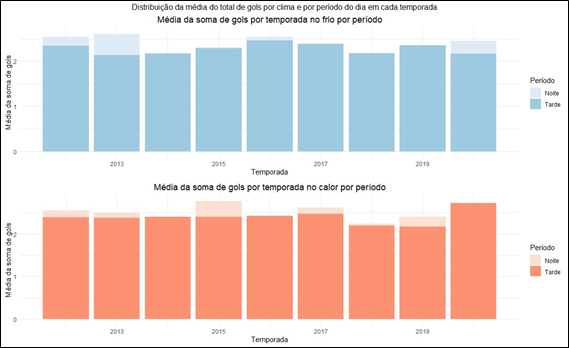


Fonte: elaborado pelos autores (2021).

Agora, a partir dos dados do gráfico 5, podemos elaborar uma resposta sobre a influência da estação no resultado dos jogos, através do cálculo da média de soma de gols por temporada em cada período. Assim, durante os períodos analisados anteriormente, de 2012, 2014, 2016 e 2018, percebe-se que não podemos chegar em uma análise demasiada útil, afinal, houve 2 anos em que a média foi maior no calor (2014 e 2018) e 2 anos em que a média foi maior no frio (2012 e 2016).

         Dessa maneira, para uma maior categorização de resultados, será utilizado os dados de todos os anos informados, e, dessa forma, é notável a ocorrência de 6 anos em que a média de gols foi maior durante a temporada no calor, com apenas 3 anos na temporada no frio. Logo, o resultado obtido com a análise é de que apesar das maiores médias de gols por estações ocorrerem de forma “aleatória”, normalmente a temporada no calor tem melhores resultados, por esse motivo, o torcedor terá mais chances de ver partidas mais concorridas durante os meses de verão e primavera.

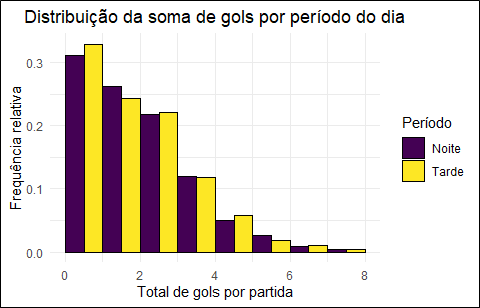
Gráfico 6 — Média da soma de gols por temporada no frio por período.



Fonte: elaborado pelos autores (2021).

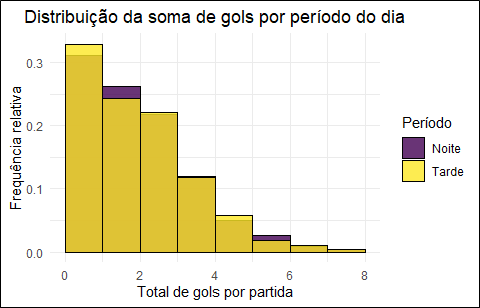
Além disso, avaliando os dados do gráfico 6, também podemos notar que a média de gols, independente do período que ocorre (frio ou calor), é, no mínimo, igual ou maior à noite.

Gráfico 7 — Distribuição da soma de gols por período do dia (1)



Fonte: elaborado pelos autores (2021).

Gráfico 8 — Distribuição da soma de gols por período do dia (2)



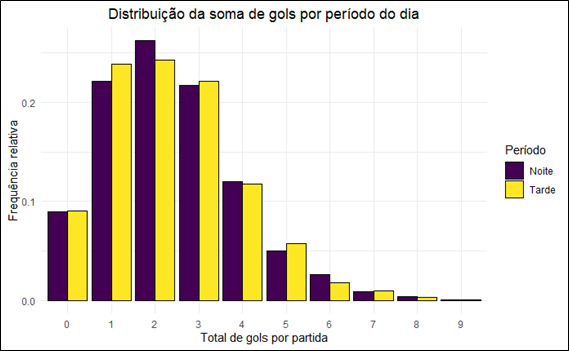
Fonte: elaborado pelos autores (2021).

Em seguida, realizando uma análise geral sobre os períodos do dia em que ocorrem as partidas, no gráfico 7 e 8, é possível notar que os jogos que ocorrem à noite possuem maior chance de ocorrer, pelo menos, de zero a um gol do que partidas à tarde. Todavia, tais critérios praticamente se equiparam quando tratamos sobre jogos com soma de 2 a 3, 3 a 4, 6 a 7 e 7 a 8 gols.

         Ainda que seja o caso, e a diferença da soma de gols por período do dia seja pequena, no geral é legítimo afirmar que jogos que ocorrem à noite geralmente tem um resultado melhor (pois há uma frequência menor de zero a um gol do que à tarde). Assim, tal conclusão pode decorrer de diferentes fatos, primeiramente, pode-se supor que partidas que ocorrem à noite tem maior audiência, afinal, a maioria das pessoas já terminaram seu turno de trabalho, logo, os jogos mais importantes, com times e jogadores mais experientes e capacitados, ocorrem justamente nesse horário mais tardio, pois há maior possibilidade de arrecadar valor e visualizações.

         Isso reafirma a conclusão que chegamos através do gráfico 6, ademais, segue o gráfico 7 para visualizarmos melhor a frequência de gols em cada período:

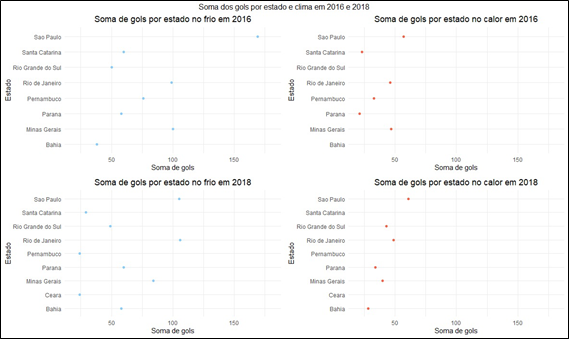
Gráfico 9 — Distribuição da soma de gols por período do dia



Fonte: Os autores (2021)

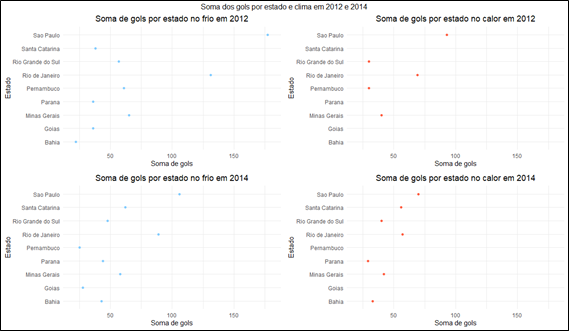
Utilizando o gráfico 7 em conjunto com os gráficos 6 e 5, podemos notar que a frequência relativa por soma de gols, ainda indica um melhor resultado, em média, para as partidas que ocorrem à noite. Portanto, confirma-se a hipótese de que jogos disputados após o período da tarde sejam mais disputados, e, portanto, uma melhor alternativa para o torcedor assistir um jogo.

Gráfico 10 — Soma dos gols por estado e clima em 2016 e 2018



Fonte: Os autores (2021)

Gráfico 11 — Soma dos gols por estado e clima em 2012 e 2014.



Fonte: Os autores (2021)

Ademais, analisando os gráficos 10 e 11, pode-se perceber que independente da influência de temperatura, usualmente, o estado em que mais ocorrem gols é São Paulo, seguido pelo Rio de Janeiro e Minas Gerais, logo, o torcedor que busca participar de uma partida “movimentada” deve assisti-la em alguma dessas regiões. Entretanto, os dados também podem significar que mais partidas ocorrem em tais locais, assim, uma segunda conclusão que podemos chegar, é a de frequência de partidas em cada região, logo, sabemos que normalmente ocorrem mais jogos em São Paulo, dessa forma, caso um fã busque assistir algo presencialmente, terá mais chances de conseguir um ingresso nessa área.

Gráfico 12 — Distribuição da quantidade total de gols por estado e período (1)

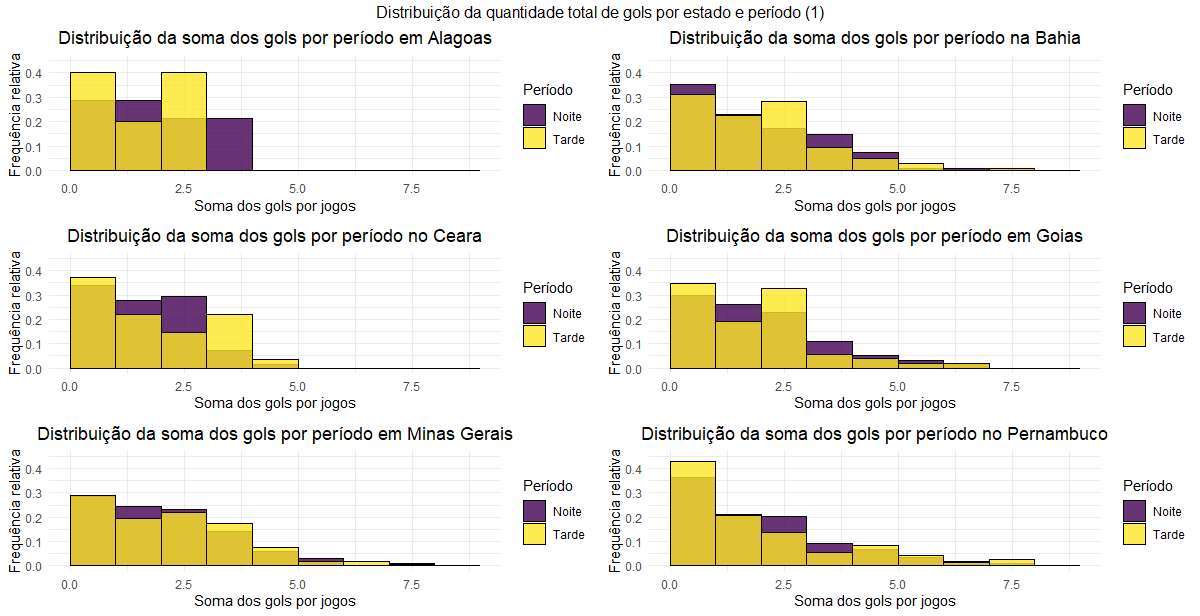
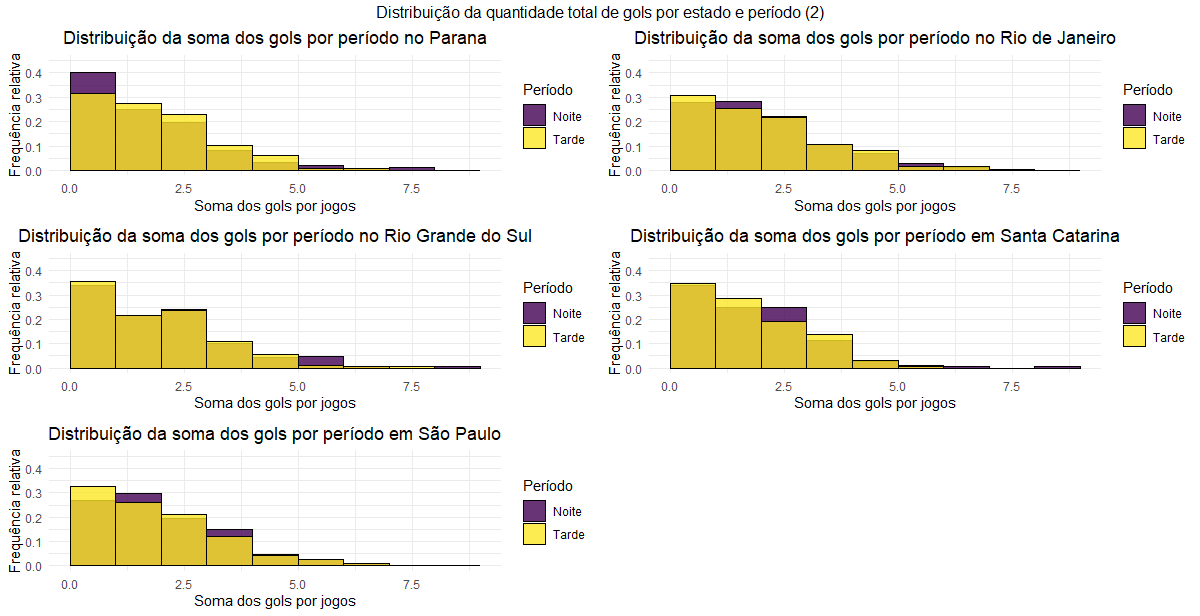
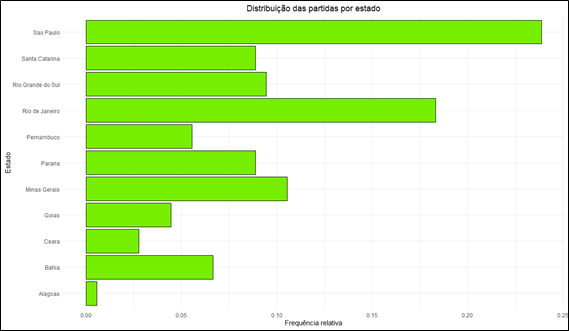
Fonte: Os autores (2021)

Gráfico 13 — Distribuição da quantidade total de gols por estado e período (2)



Fonte: Os autores (2021)

Gráfico 14 — Distribuição das partidas por estado



Fonte: elaborado pelos autores (2021).

### EM QUAL CENÁRIO É MAIS PROVÁVEL DE OCORRER UM JOGO EMOCIONANTE?

Com as análises feitas anteriormente chegou-se a conclusão de que o cenário mais propício para um jogo ser emocionante, do ponto de vista de um torcedor que visa o maior número de gols na partida, seria em estações quentes, como verão e primavera, assim como visto, principalmente, a partir do Gráfico 3. Tratando sobre o período do dia entre manhã, tarde ou noite, o melhor resultado é obtido no período noturno, o que justifica os jogos mais importantes serem realizados nessa faixa de horário. Por fim, em relação a localidade os jogos "mais emocionantes" são encontrados no estado de São Paulo, seguidos por Rio de Janeiro e Minas Gerais, ou seja, todos da região Sudeste. Todas as condições citadas são as mais favoráveis ao jogador que busca um jogo repleto de emoções.

REFERÊNCIAS

WICKHAM, Hadley; GROLEMUND, Garrett. **R for Data Science**: Import, Tidy, Transform, Visualize, and Model Data. "O'Reilly Media, Inc.", v. 1, 2016.