Trabalho final de Análise de Dados (INF0612)

Leonardo Cesar Silva dos Santos, Fernando Augusto Cardoso Candalaft

Março 2024

Contents

1 Introdução

O estudo em questão consiste em tratar e analisar os dados dados climatológicos da cidade de Campinas, disponíveis em "https://www.ic.unicamp.br/ zanoni/cepagri/cepagri.csv". O intervalo de tempo dos dados a ser considerado compreende os dados entre 1-Jan-2015 e 31-Dez-2023.

2 Tratamento dos dados

O tratamento pré análise dos dados foi feito sobre uma tabela com a seguinte estrutura:

Table 1: Dados Meteorológicos

Data e Hora	**Temp. (°C)**	**Vel. Vento (km/h)**	**Umidade (%)**	**Sensação Térmica (°C)**
02/03/2014-19:08	23.7	59.3	77.1	22.6
02/03/2014-19:18	23.4	59.1	77.9	22.3
02/03/2014-19:28	23.2	56.7	78.9	22.1
02/03/2014-19:38	23.0	55.4	79.2	21.9
02/03/2014-19:48	22.8	52.6	79.7	21.7
02/03/2014-19:58	22.6	62.6	80.7	21.5

2.1 Renomeando as colunas

A primeira etapa consistiu em renomear as colunas de maneira intuitiva, como mostrado na tabela acima.

2.2 Identificação de valores nulos

O primeiro processo de fato relevante que foi feito sobre a base de dados foi a identificação de valores nulos assim como o tipo de cada coluna, uma vez que ambas as informações são de extrema importância para qualquer análise. A tabela com tais informações é dada logo abaixo.

Table 2: Descrição dos dados

Data e Hora	**Temp. (°C)**	**Vel. Vento (km/h)**	**Umidade (%)**	**Sensação Térmica (°C)**
Length:518094	Length:518094	Min. :0.00	Min. :0.00	Min. :-8.20
Class :character	Class :character	1st Qu.:10.50	1st Qu.:56.30	1st Qu.:16.60
Mode :character	Mode :character	Median :19.10	Median :72.50	Median :19.90
		Mean :22.77	Mean :68.96	Mean :19.86
		3rd Qu.:31.00	3rd Qu.:83.40	3rd Qu.:23.90
		Max. :143.60	Max. :100.00	Max. :99.90
		NA's :40864	NA's :40864	NA's :74270

Usando a tabela de descrição dos dados como referência foi possível notar a necessidade de tratamento da coluna *vl temp*, uma vez que a mesma deveria ser numérica mas está listada como categórica. Deste modo, convertemos a mesma para o tipo numérico usando o auxílio da função *as.numeric* do R que introduz valores nulos (NA) por coerção. Após isso, removemos também as linhas com valores nulos usando como referencia as colunas *vl temp*, *vl vel vento* e *vl umidade*. Logo em seguida, verificamos novamente a descrição dos dados.

Da tabela acima é possível que nosso valor máximo de temperatura é 38.1 graus Celsius enquanto que temos uma sensação térmica de 99.9 graus Celsius, o que indica que temos algum tipo de erro associado a essa informação. Avaliando os valores associados ao máximo de sensação térmica observada (99.9) é possível ver

Table 3: Descrição dos dados após remoção de valores nulos

Data e Hora	Temp. (°C)	Vel. Vento (km/h)	Umidade (%)	Sensação Térmica (°C)
Length:477230	Min.:4.60	Min.:0.00	Min.:0.00	Min.:-8.20
Class:character	1st Qu.:18.60	1st Qu.:10.50	1st Qu.:56.30	1st Qu.:16.60
Mode:character	Median:21.50	Median:19.10	Median:72.50	Median:19.90
	Mean:21.99	Mean:22.77	Mean:68.96	Mean:19.86
	3rd Qu.:25.50	3rd Qu.:31.00	3rd Qu.:83.40	3rd Qu.:23.90
	Max.:38.10	Max.:143.60	Max.:100.00	Max.:99.90
				NA's:33406

que temos um erro na coleta, de modo que substituímos esses valores por nulos (NA) e olhamos novamente a descrição dos dados.

Table 4: Dados Meteorológicos

Data e Hora	Temp. (°C)	Vel. Vento (km/h)	Umidade (%)	Sensação Térmica (°C)
Length:477230	Min.:4.60	Min.:0.00	Min.:0.00	Min.:-8.20
Class:character	1st Qu.:18.60	1st Qu.:10.50	1st Qu.:56.30	1st Qu.:16.60
Mode:character	Median:21.50	Median:19.10	Median:72.50	Median:19.90
	Mean:21.99	Mean:22.77	Mean:68.96	Mean:19.84
	3rd Qu.:25.50	3rd Qu.:31.00	3rd Qu.:83.40	3rd Qu.:23.90
	Max.:38.10	Max.:143.60	Max.:100.00	Max.:34.70
				NA's:33548

Após isso é possível ver que os valores de sensação térmica se ajustaram. Nessa coluna ainda temos valores nulos, mas que serão tratados posteriormente de acordo com análise, uma vez que a eliminação dos mesmos neste momento pode acarretar na perda de informação de outras colunas.

2.3 Dados duplicados

Para auxiliar na identificação de dados duplicados convertemos nossa coluna *vl dhora* para o tipo adequado (*POSIXct*) e extraímos informações como hora, dia, mês e ano. Como o intervalo de coleta dos dados é de 10 minutos e isso pode trazer ruídos indesejados para a análise, resolvemos olhar somente para os intervalos de 1 hora e identificar valores duplicados e possíveis erros de coleta (isto é, valores iguais de temperatura em um mesmo dia para intervalos consecutivos de tempo) para esses intervalos.

Antes de qualquer remoção de dados duplicados ou consecutivos nesta etapa, tínhamos um conjunto de dados com 477230 linhas. Após a remoção de valores duplicados levando em consideração as colunas *vl temp*, *vl vel vento*, *vl umidade*, *vl stermica*, *cd hora*, *cd dia*, *cd ano* e *cd mes* obtivemos um novo conjunto de dados com 444672 linhas. Finalmente, após a remoção de valores consecutivos de temperatura retornamos uma tabela com 444242 linhas.

3 Análise dos dados

Após as etapas da seção anterior obtivemos a seguinte descrição dos nossos dados:

Table 5: Dados Meteorológicos

Temp. (°C)	Vel. Vento (km/h)	Umidade (%)	Sensação Térmica (°C)	Hora	Dia	Ano	Mês
Min.:4.6	Min.:0.00	Min.:0.00	Min.:-8.20	Min.:0.00	Min.:1.00	Min.:2014	Min.:1.000
1st Qu.:18.5	1st Qu.:10.30	1st Qu.:56.00	1st Qu.:16.40	1st Qu.:6.00	1st Qu.:8.00	1st Qu.:2016	1st Qu.:3.000
Median:21.4	Median:18.80	Median:72.60	Median:19.80	Median:12.00	Median:16.00	Median:2018	Median:6.000
Mean:21.9	Mean:22.56	Mean:69.05	Mean:19.71	Mean:11.61	Mean:15.71	Mean:2018	Mean:6.364
3rd Qu.:25.4	3rd Qu.:30.80	3rd Qu.:83.50	3rd Qu.:23.70	3rd Qu.:18.00	3rd Qu.:23.00	3rd Qu.:2021	3rd Qu.:9.000
Max.:38.1	Max.:143.60	Max.:100.00	Max.:34.70	Max.:23.00	Max.:31.00	Max.:2024	Max.:12.000
							NA's:27666

3.1 Avaliação dos dados entre 2018 e 2023

Para termos uma noção geral sobre o comportamento dos nossos dados ao longo dos anos calculamos algumas informações dos mesmos como média, máximo, mínimo e mediana ao longo dos respectivos meses. Olhamos especificamente, num primeiro momento, o intervalo de 2018 a 2023. Observando a Figura 1 de evolução da

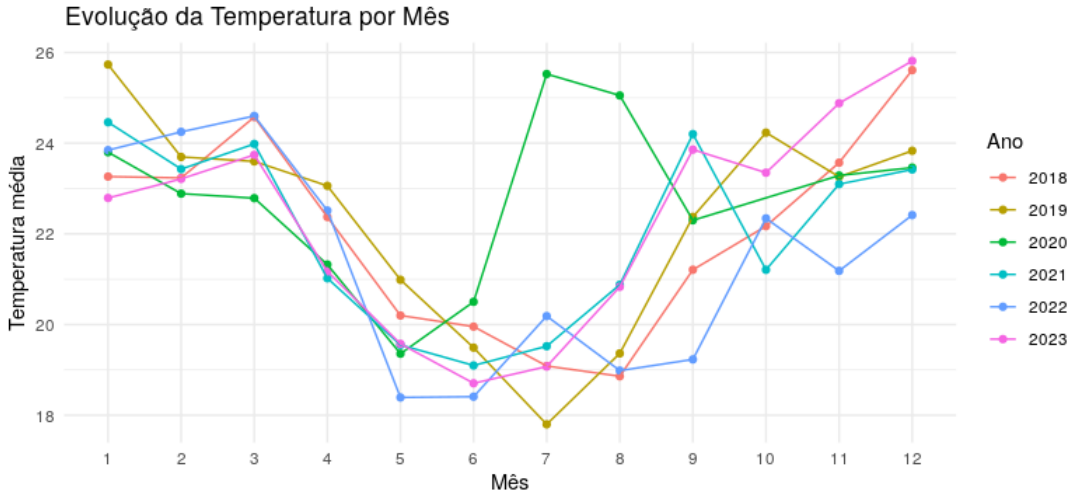

Figure 1: Evolução da temperatura média - 2018 a 2023.

temperatura média ao longo dos anos e meses é possível ver que em todos os anos temos a mesma tendência, exceto em 2020, que possui picos de temperatura nos meses 7 e 8 (Julho e Agosto).

Porém, se olhamos a Figura 2 de temperaturas máximas observadas para todos os anos, é possível notar que em 2020 tivemos temperaturas relativamente baixas até Setembro, o que talvez esteja relacionado com o fato de que neste ano estávamos enfrentando um período de pandemia, e isso talvez tenha influenciado de alguma forma, uma vez que ao menos as emissões de gases diminuíram, e tais efeitos se relacionam com a temperatura. Agora olhando para a relação entre temperatura e umidade (Figura 3), vemos que outro fator que pode ter ajudado nos picos de temperatura média no ano de 2020 é a baixa umidade no mesmo período. O que faz sentido, dado que a umidade do ar está relacionada a ocorrência de chuvas. Novamente, períodos com baixa umidade indicam uma temperatura media maior e consequentemente uma sensação térmica mais elevada, em particular o ano de 2020. Por outro lado, o ano de 2019 foi o ano com percepção de temperatura mais baixa dado que seguiu a tendência de outros anos, mas tendo pico de baixa temperatura média no mês 7 e um período de 12 meses de estabilidade com relação a umidade e velocidade do vento (Figura 4).

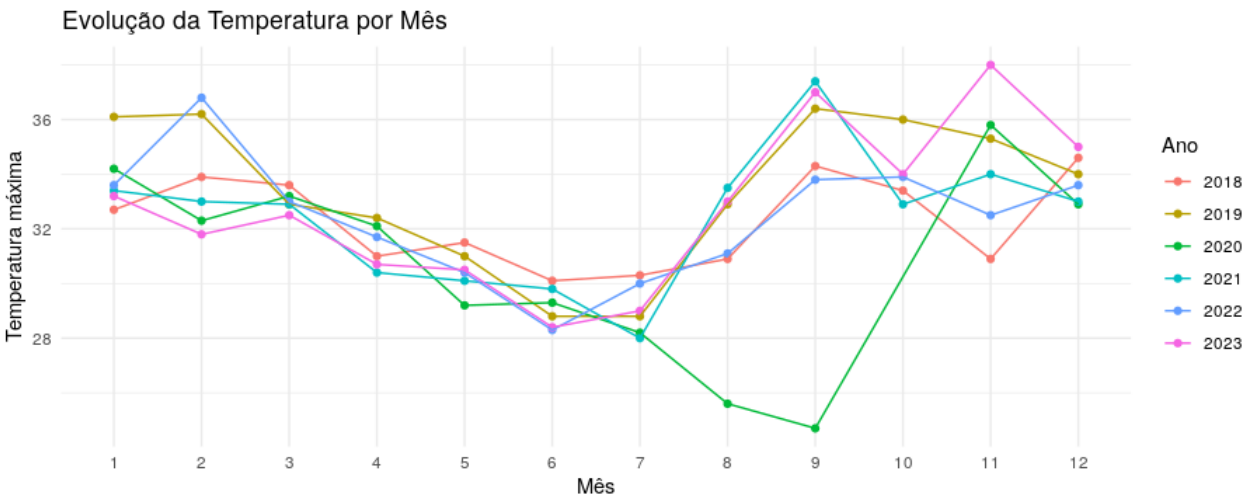

Figure 2: Evolução da temperatura máxima - 2018 a 2023.

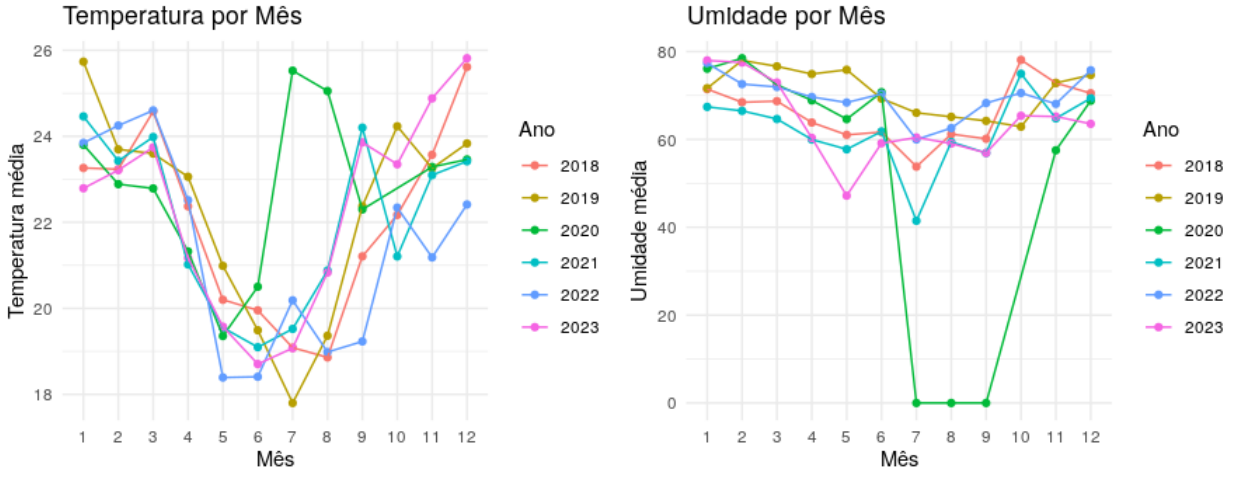

Figure 3: Evolução da temperatura média e umidade média - 2018 a 2023.

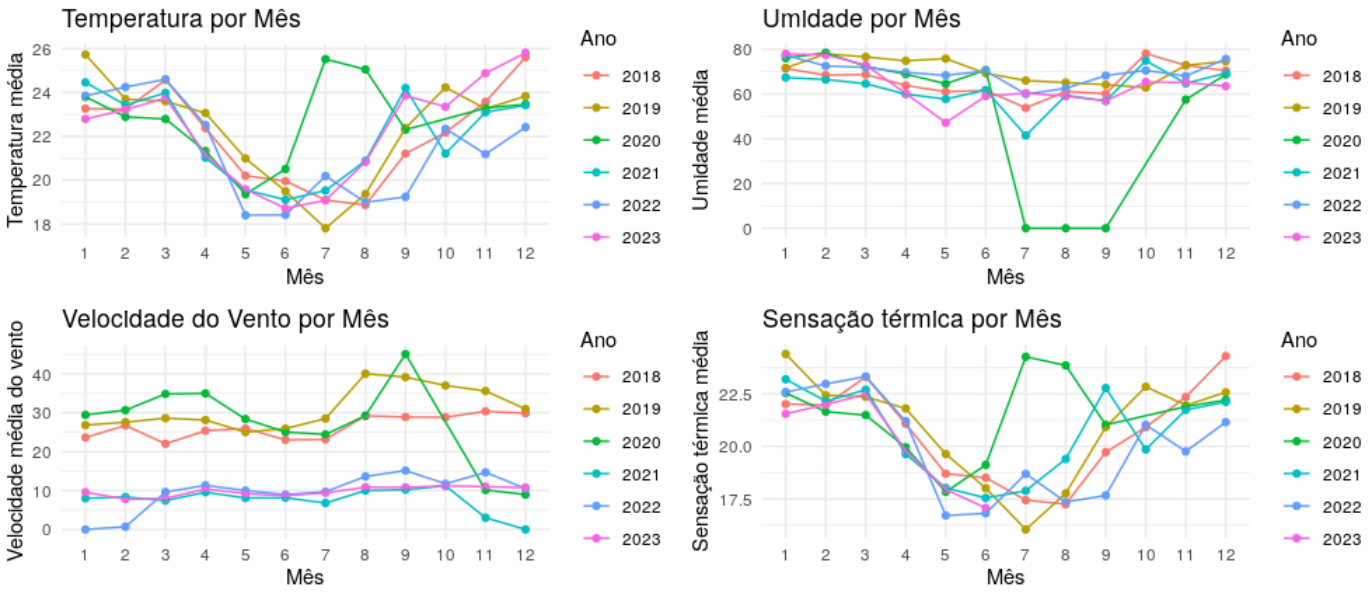

Figure 4: Comparação das medidas observadas ao longo do meses - 2018 a 2023.

3.2 Avaliação dos anos 2019 e 2023

Os últimos anos não tem sido nada comuns, uma vez que passamos por um longo período de pandemia. Dito isto, é interessante avaliarmos se houve algum tipo de comportamento diferente entre os períodos "pré" e "pós" pandemia, neste caso, os anos de 20219 e 2023.

Pelo gráfico da Figura 5 de distribuição de temperaturas para os anos de 2019 e 2023 não é possível determinar de forma clara se houve ou não uma mudança de comportamento na variável avaliada.

Se olhamos agora a diferença entre os valores de temperatura média e umidade média no ano de 2023 contra o ano de 2019 (Figura 6) fica claro que a umidade em 2019 foi muito mais expressiva do que em 2023.

Table 6: Diferenças de Temperatura e Umidade por Mês - 2023 contra 2019

Mês	Diferença na Temp. Média	Diferença na Temp. Máxima	Diferença na Temp. Mínima	Diferença na Umidade Média	Diferença na Umidade Máxima
1	-2.94	-2.9	-0.5	6.33	-6.1
2	-0.48	-4.4	2.3	-0.48	-6.4
3	0.14	-0.4	1.0	-3.68	-4.5
4	-1.89	-1.7	-1.9	-14.55	-5.1
5	-1.41	-0.5	-1.8	-28.66	-23.3
6	-0.79	-0.4	2.1	-10.15	-17.3
7	1.28	0.2	1.9	-5.63	-13.7
8	1.47	0.1	2.6	-6.09	-10.6
9	1.48	0.6	-1.6	-7.29	-14.5
10	-0.88	-2.0	-0.5	2.50	5.2
11	1.62	2.7	-0.6	-7.65	3.3
12	1.98	1.0	0.7	-11.12	2.0

Podemos avaliar também o comportamento da distribuição de temperaturas e umidades ao longo dos anos e especificamente de 2019 contra 2023 para verificar se seguem um comportamento semelhante (Figura 7).

A distribuição de umidade de 2019 contra 2023 é claramente diferente, mas a distribuição de temperaturas para os mesmos anos é semelhante. Dado tal semelhança, suspeitamos que ao menos a variável temperatura nestes anos tenha tido pouca variação. Uma medida que nos ajuda a enfatizar tal suspeita é a área sob a curva de cada ano. Para 2019 temos uma área associada a temperatura dada por 1.000616 e para 2023 temos uma área associada dada por 1.000767. Comparando as distribuições e os valores de área vemos que a temperatura nestes anos foi semelhante.

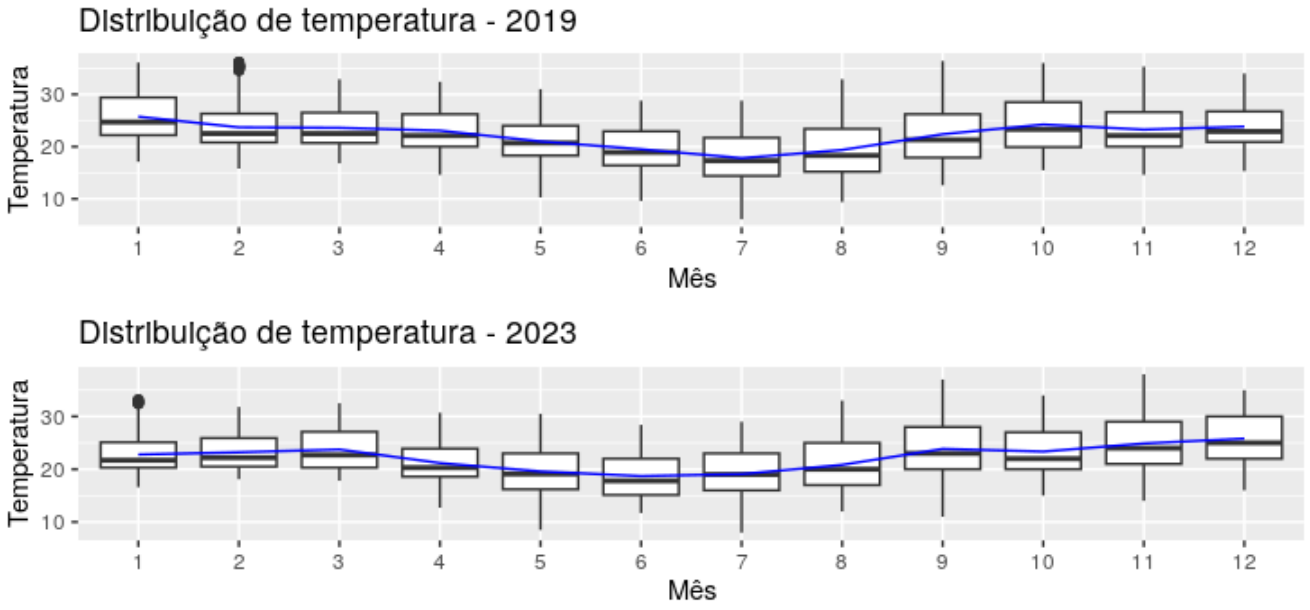

Figure 5: Distribuição de temperatura ao longo dos meses - 2019 x 2023.

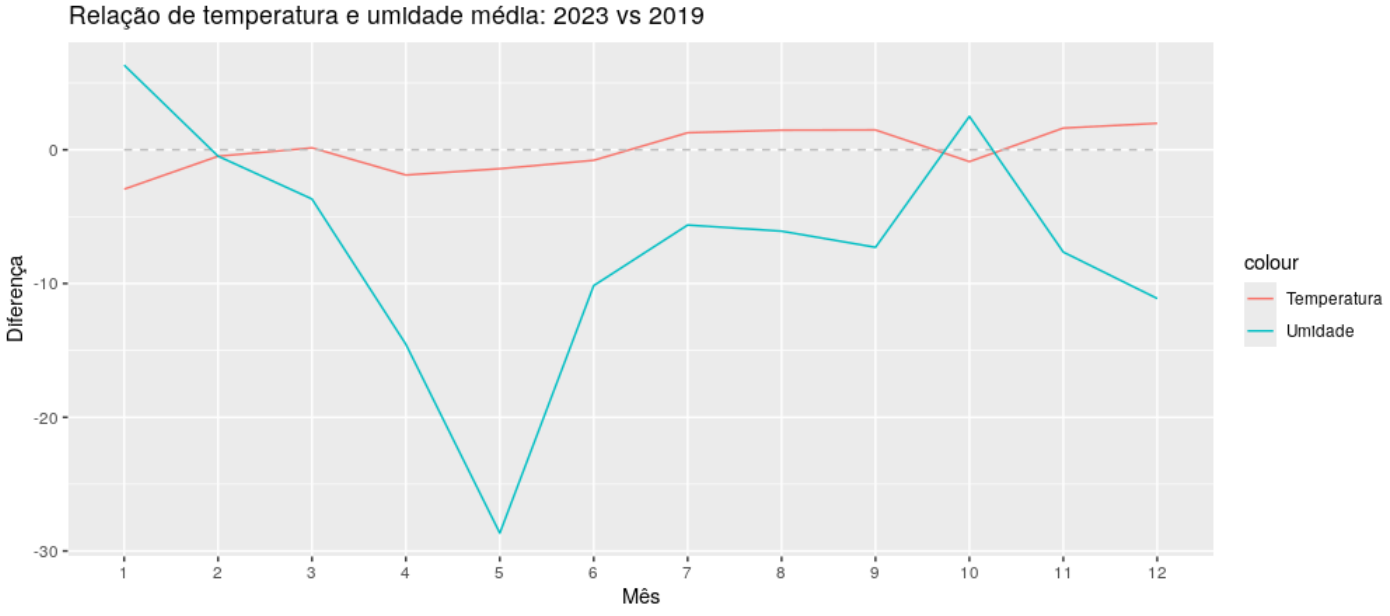

Figure 6: Diferença entre 2023 e 2019.

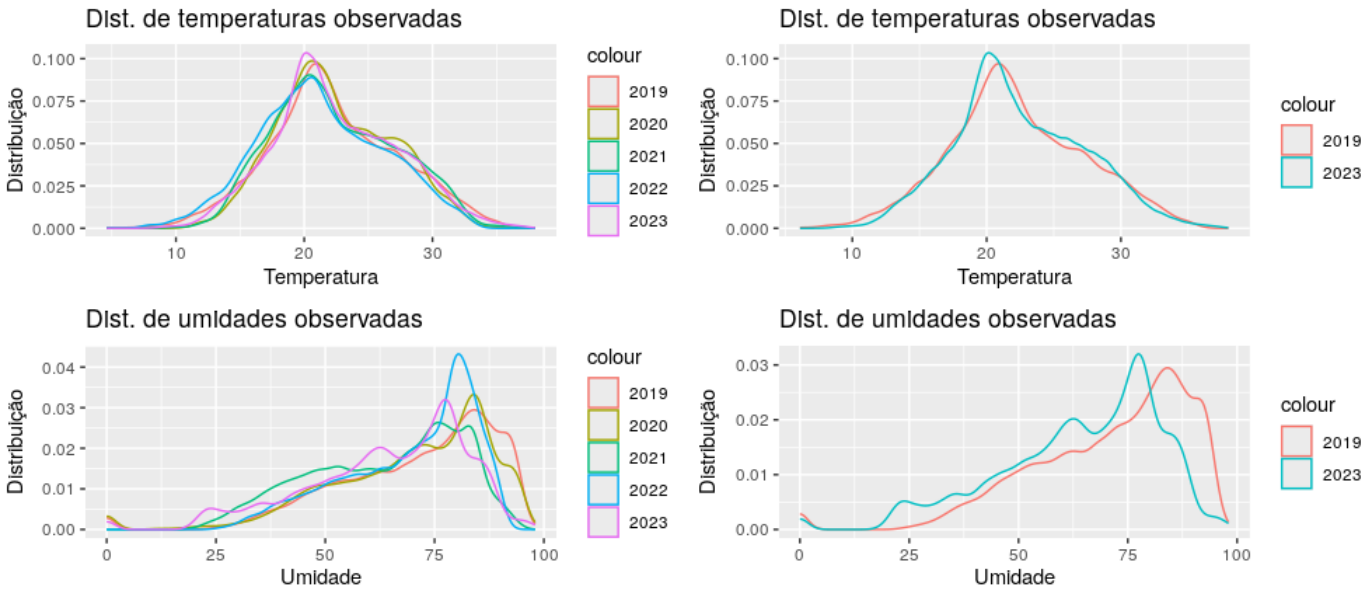

Figure 7: Distribuição de temperaturas e umidades.

4 Conclusão

Todos os anos seguem um comportamento similar, principalmente em termos de linha de tendência da variável temperatura. Porém, em termos de umidade temos algumas diferenças. Olhando especificamente os anos de 2019 e 2023 foi possível notar que em 2023 tivemos menos umidade se comparado a 2019.