Trabalho Final - INF-0612

Kaleb Roncatti de Souza e Nelson Gomes Brasil Júnior

28 de agosto de 2022

1 Introdução

Vivemos numa época extremamente privilegiada no que diz respeito à tecnologia e seus impactos no nosso dia-a-dia. Através do avanço e desenvolvimento dos computadores, possuímos cada vez mais disponibilidade de memória tanto a nível processamento quanto a nível armazenamento nos dispositivos que utilizamos no cotidiano. Esta grande disponibilidade de poder computacional torna cada vez mais palpável e importante o uso de dados agregados em praticamente todos os ramos que possamos imaginar. O uso de dados pode nos ajudar a entender o comportamento de um dado fenômeno, nos permitindo tirar conclusões baseadas em ocorrências passadas do mesmo. Ademais, pode-se tentar prever ou realizar projeções de possíveis comportamentos futuros do fenômeno em questão, assim como veremos adiante no curso de Mineração de Dados Complexos.

No entanto, dados puramente brutos não possuem valor atrelado no que diz respeito à análises, projeções e previsões, precisando passar por todo um processo de tratamento e análise, o qual faremos no trabalho vigente, em acordo com a bagagem adquirida ao longo da disciplina INF-0612.

Como dados de referência a serem tratados/analisados ao longo deste trabalho, utilizaremos dados climatológicos da cidade de Campinas, disponibilizados pela CEPAGRI. Demonstraremos ao longo do trabalho inúmeras visões e tratamentos dos dados em questão. Iniciaremos carregando o conjunto de dados e realizando todos os tratamentos necessários para a subsequente análise.

Este trabalho está dividido da seguinte maneira: Na Seção 2 fazemos uma análise exploratória dos dados e na Seção 3 apresentamos algumas análises interessantes sobre os dados da CEPAGRI.

2 Análise Exploratória dos dados

Nesta seção iremos apresentar uma análise exploratória dos dados da CEPAGRI obtidos em [4], de modo a darmos sequência nas análises que apresentaremos a seguir. Este passo é importante para que as análises sejam performadas de maneira mais precisa possível, removendo impurezas do conjunto de dados.

Os códigos em `R` para esta seção, estarão anexados a este trabalho.

2.1 Carregamento e contextualização

Após carregar os dados, obtivemos um dataframe com 5 colunas e 439897 linhas. As colunas representam as seguintes medidas:

- *dt* do tipo `character`, as datas e horário que os dados foram obtidos do CEPAGRI;
- *temperature* do tipo `character`, a temperatura (em $°C$) no momento *dt*;
- *wind_speed* do tipo `numeric`, a velocidade do vento (em km/h) no momento *dt*;
- *humidity* do tipo `numeric`, a umidade do ar (em %) em *dt*;
- *thermal_sensation* do tipo `numeric`, a sensação térmica em *dt*.

2.2 Transformações e tratamentos

No que diz respeito à transformações/tratamentos nos dados realizados ao longo do projeto, aplicamos os seguintes passos:

- **Modificação do tipo dos dados da coluna *dt***: Para o tipo `POSIXt`, um tipo de data na linguagem `R`. Isto foi feito para que pudéssemos filtrar as datas entre os anos 2015 e 2021.
- **Modificação do tipo dos dados da coluna *temperature***: Para o tipo `numeric`, como o próprio nome se refere, um tipo numérico para que possamos observar o sumário e calcular algumas métricas.
- **Limpeza dos dados através da remoção dos NA's e os dados com erros**: Para tal identificação, reparamos que, desde o carregamento dos dados de origem e da obtenção do seu *summary*, tínhamos um número não desprezível de *Not A Number's*, cerca de 8.1% de todo o conjunto. Desta maneira, seguimos com a visualização de tais dados e percebemos que, a única informação relevante em todo este conjunto problemático eram as datas apresentadas, isto é, todas as outras informações além da data para estes dados específicos não nos traziam nada relevante. Desta maneira, removemos estes dados.
- **Retiramos os *outliers***: Baseando-se nos valores obtidos através do *summary*, percebemos dois casos específicos se mostravam como casos aberrantes: valores de sensação térmica extremamente elevados para os padrões brasileiros ($99\,°C$) e valores de umidade igual a zero. Para ambos os casos, verificamos os histogramas (exemplificado para a umidade do ar na Figura 1) e identificamos que tais valores de fato se mostravam como valores fora das distribuições.
- **Inserimos novas colunas para complementar a análise**: Definimos colunas segmentando-se o objeto `POSIXt` em hora, dia, mês e ano. Ademais, construímos uma função que mapeia os dias do ano para a estação do ano correspondente e definimos a coluna *season* para futuros tratamentos.

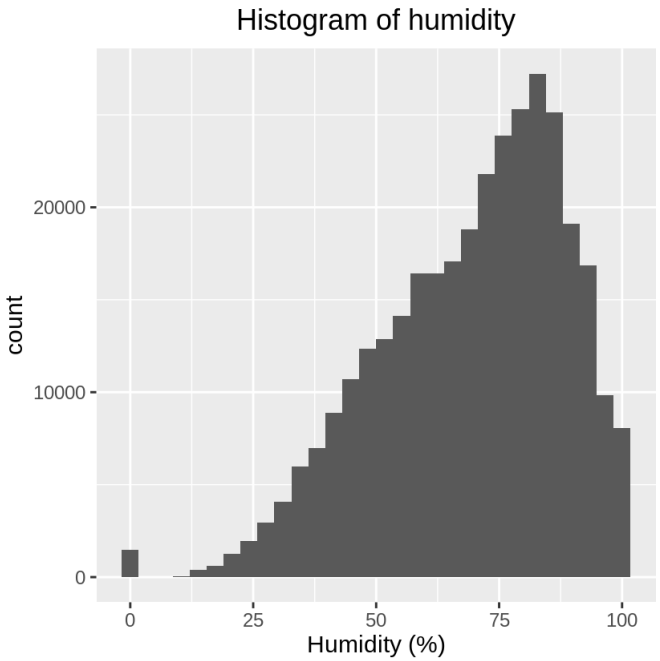


Figura 1: Umidade do ar

Após a remoção dos `NA`, outliers e filtrando apenas os anos que vamos analisar, resultamos com a distribuição de valores por ano. Note que para os anos de 2020 e 2021 tivemos uma porcentagem alta de dados nulos (veja Tabela 1), o que pode influenciar nossas análises se fizermos uma divisão por ano.

Tabela 1: Proporção de dados `NA` por ano

ano	**dados**	**dados `NA`**	%
2015	51,780	568	1.1
2016	52,195	65	0.12
2017	52,083	1,168	2.2
2018	52,217	246	0.47
2019	52,298	22	0.042
2020	51,350	20,016	39
2021	52,357	11,611	22.2

3 Análises

Com os dados tratados, e apenas com os dados que devem ser usados neste trabalho, vamos iniciar as análises.

3.1 Análise 1: Variação de temperatura ao longo do dia

Começamos analisando a variação da temperatura média ao longo do dia para cada uma das estações do ano. Então, agrupamos os dados por hora e por estação e calculamos a média e o resultado é apresentado na Figura 2. A partir desta figura, podemos tirar as seguintes conclusões:

- Temperaturas médias no verão e na primavera muito próximas no decorrer do dia, sendo a maior diferença durante a noite e madrugada;
- Temperaturas médias no inverno e outono são próximas a partir do meio-dia, sendo as madrugadas no inverno mais frias do que no outono;
- A amplitude térmica no inverno é maior do que nas outras estações, o que pode ser visto também na Figura 3

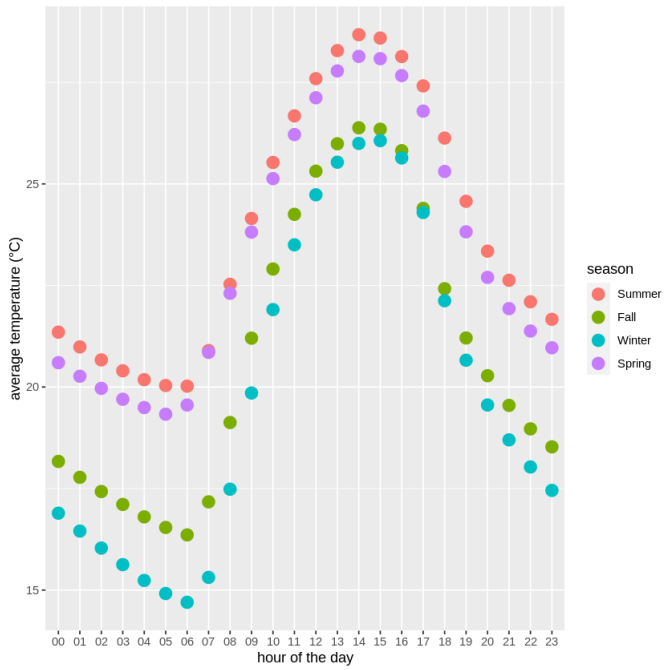


Figura 2: Temperatura média em Campinas por hora do dia e estação do ano

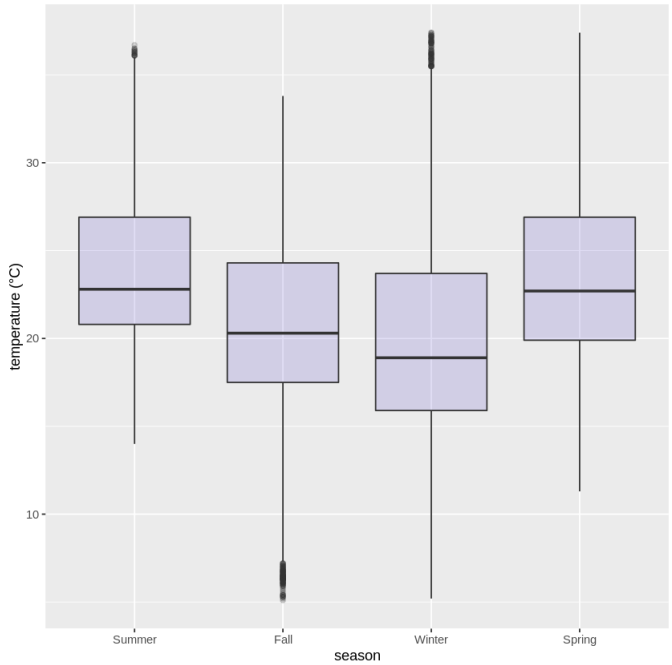


Figura 3: Temperatura em Campinas por estação do ano

3.2 Análise 2: Relação entre temperatura e umidade médias

No momento que analisávamos os dados para a retirada dos outliers, percebemos alguns valores de sensação térmica abaixo do esperado para a região de Campinas (cerca de $-8\,°C$) o que nos levou a uma investigação mais detalhada destes valores. Uma hipótese que pensamos foi de que a umidade do ar ou a velocidade dos ventos pudesse influenciar a sensação térmica e, portanto, fazer ela atingir estes valores. Uma primeira avaliação que fizemos foi olhar a correlação entre as variáveis que estamos trabalhando, e isto é mostrado na Tabela 2. Note que, como esperado, temperatura e sensação térmica possuem um grau bem elevado de correlação.

Tabela 2: Correlação entre as variáveis

	temperature	**wind speed**	**humidity**	**thermal sensation**
temperature	1.00	−0.15	−0.64	0.94
wind speed	−0.15	1.00	0.10	−0.22
humidity	−0.64	0.10	1.00	−0.54
thermal sensation	0.94	−0.22	−0.54	1.00

Umidade e temperatura tem uma correlação de $\rho = -0.64$, o que é considerado uma correlação média entre as variáveis. Resolvemos então investigar este comportamento e, para isto, decidimos explorar a variação por estação de modo a concentrar os dados em grupos que possuem maior similaridade, pois não faria sentido considerar verão e inverno na mesma categoria, por exemplo. Agrupando por estação e horário do dia e calculando a média da umidade do ar

e da temperatura, obtivemos a Figura 4. Observando pelo gráfico, é possível notar que parece existir uma relação linear entre temperatura e umidade médias por estação. É interessante notar também que, para uma mesma temperatura média, o inverno aparenta um menor valor de umidade do ar média, o que faz sentido dado que invernos campineiros costumam ser mais secos do que outras estações.

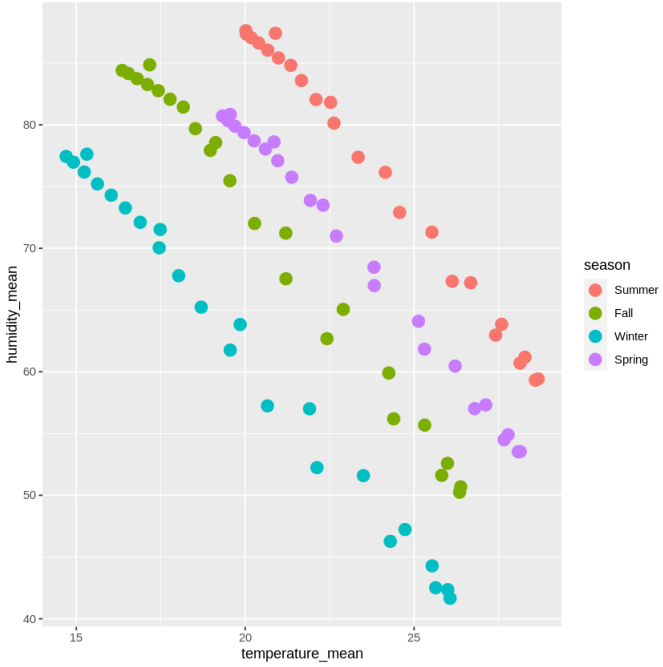


Figura 4: Relação entre temperatura e umidade médias

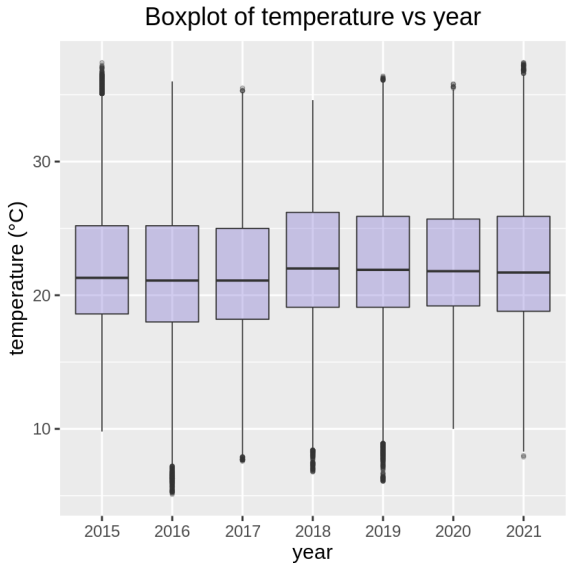
3.3 Análise 3: Variação da temperatura e sensação térmica por ano

No que diz respeito aos indicadores ao longo dos anos, partimos do pressuposto (viés/senso comum) de que haveria uma discrepância altamente relevante de temperatura ou sensação térmica para contextualizarmos a nível aquecimento global. No entanto, observando-se os *boxplots* das Figuras 5a e 5b, percebemos que as medianas de temperatura se mantiveram aproximadamente em torno do mesmo valor, e observamos um pequeno aumento na mediana da sensação térmica entre 2016 e 2018. O que é possível observar de maneira clara é que, tanto para a temperatura quanto para a sensação térmica, para os anos de 2015 até 2019 tivemos uma concentração mais relevante de valores baixos, longe da mediana que não representam outliers (menores do que 10 graus Celsius).

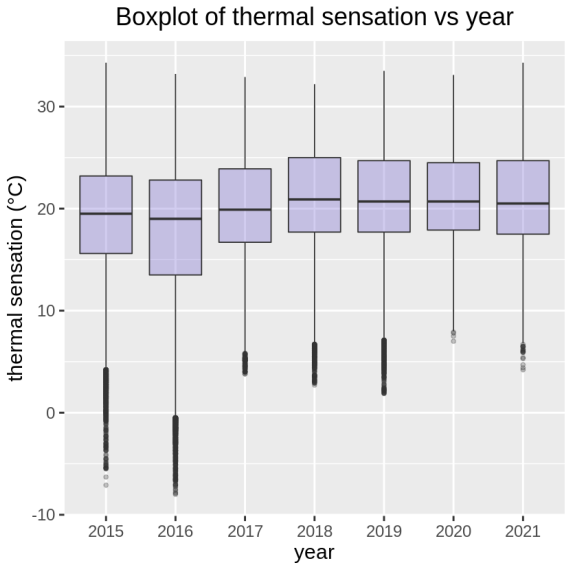
Tabela 3: Dados perdidos (NAs) em cada mês no ano de 2020

month	lost records
05	958
06	1,923
07	4,432
08	3,490
09	4,155
10	4,460
11	459
12	139

Porém, se relembrarmos da remoção de NA's que fizemos no começo do processo de tratamento de dados e analisarmos mais a fundo, olhando para cada um dos meses para os anos que mais perdemos dados (2020 e 2021, ano de 2020 mostrado na Tabela 3), veremos que os meses cujos dados foram perdidos se concentram justamente no período do inverno (meses 6 ao 9), o que poderia explicar a ausência de valores mais reduzidos para temperatura/sensação térmica nos anos em questão sem a consideração de fatores externos.



(a) Temperatura por ano



(b) Sensação térmica por ano

Figura 5: Boxplot para visualização da temperatura e sensação térmica ao decorrer dos anos

Se desejássemos obter efeitos mais claros e aparentes dos efeitos de aquecimento global, precisaríamos analisar uma timestamp um pouco mais extensa, da ordem de dezenas de anos [1].

3.4 Análise 4: Relação entre temperatura e sensação térmica

A Tabela 2 que mostra a matriz de correlação entre as variáveis nos diz que há uma correlação forte entre *temperatura* e *sensação térmica*, o que já era esperado dado que, de fato, essas duas medidas são intrinsecamente dependentes, ou seja, espera-se que dias com temperatura mais alta se tenha uma sensação térmica maior e o mesmo para dias mais frios.

A Figura 6 mostra o gráfico temperatura vs sensação térmica e marcamos cada ano com uma cor diferente, de modo a comparar o comportamento dos pontos. Pudemos então perceber a maior dispersão dos dados nos anos de 2015 e 2016, não mantendo a relação visualmente linear que percebemos nos outros anos, inclusive havendo superposição nos dados. Podemos explicar esta discrepância pela análise que fizemos na Seção 3.3 e também observando os outliers para estes anos na Figura 5b em comparação com os dados na Figura 5a. Ou seja, pelos gráficos mostrados, os anos de 2015 e 2016 tiveram um comportamento diferente, provavelmente devido a algum fenômeno climático que ocorreu nestes anos [2, 3]

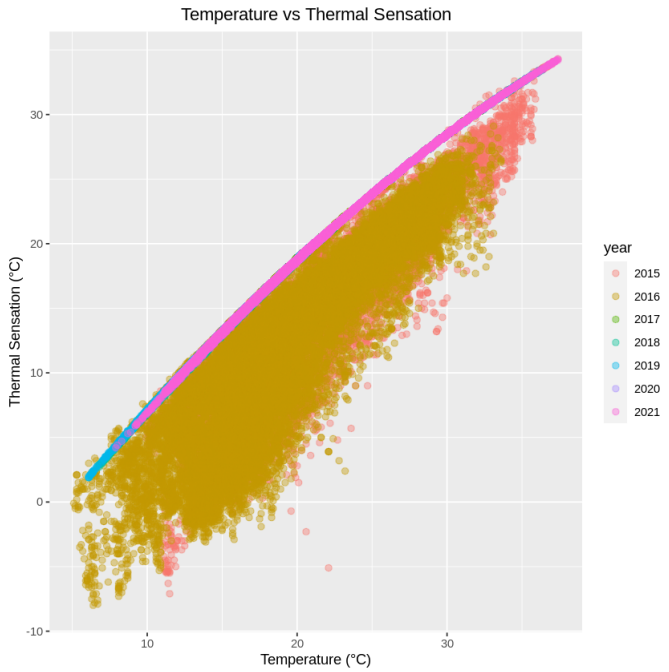


Figura 6: Relação entre temperatura e sensação térmica por ano

Para os anos de 2017 e 2021 o comportamento foi bem mais parecido, notado pela superposição na Figura 6 e também pelos pontos em preto na Figura 7. Na última figura, inclusive, podemos notar que a relação linear entre as variáveis fica evidente. O nosso próximo passo foi aproximar os pontos por uma reta para tentar encontrar qual seria a relação linear entre tais variáveis. Usamos o comando `lm` do `R`, que performa uma regressão linear para encontrar a reta que mais se ajusta aos dados minimizando a soma dos quadrados dos resíduos. Assim, encontramos a reta $T(t) = -1.85 + 1.02t$, onde t é a temperatura em graus Celsius. Na Tabela 4 apresentamos a distribuição dos erros e os coeficientes que encontramos na regressão ao usarmos o método descrito.

Tabela 4: Resultados da regressão realizada para os anos de 2017 a 2021

Resíduos:

Min	**1Q**	**Mediana**	**3Q**	**Max**
−2.4815	−0.1115	0.1065	0.2090	0.3776

Coeficientes:

	Estimate	**Std. Error**
(Intercept)	−1.848129	0.003001
temperature	1.020911	0.000132

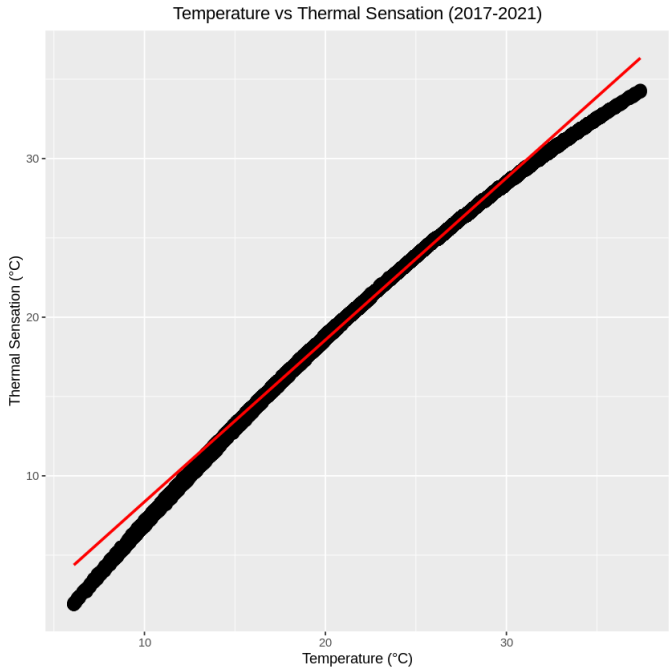


Figura 7: Regressão Linear feita a partir dos dados de temperatura vs sensação térmica entre 2017 e 2021 em Campinas

Referências

[1] BBC Brasil. `https://www.bbc.com/portuguese/geral-46424720`, 2018.

[2] G1 Campinas. `https://g1.globo.com/sp/campinas-regiao/noticia/2016/06/fenomeno-de-microexplosao-atingiu-campinas-explica-cepagri.html`, 2016.

[3] G1 Campinas. `https://g1.globo.com/sp/campinas-regiao/noticia/2016/06/campinas-registra-minima-de-53c-nesta-segunda-feira-diz-cepagri.html`, 2016.

[4] Zanoni Dias. `https://www.ic.unicamp.br/~zanoni/cepagri/cepagri.csv`, 2022.