**UNIVERSIDADE ESTADUAL DE CAMPINAS**

ESCOLA DE EXTENSÃO DA UNICAMP

INSTITUTO DE COMPUTAÇÃO

**CURSO DE APERFEIÇOAMENTO EM MINERAÇÃO DE DADOS COMPLEXOS**

**ANÁLISE DE DADOS**

TRABALHO FINAL

Guilherme Ramos Gouveia

Paola São Thiago da Cunha

Marina Abichabki Pivato

CAMPINAS

2020

# INTRODUÇÃO

Como trabalho final da disciplina de Análise de dados do curso de aperfeiçoamento em Mineração de Dados Complexos pela Escola de Extensão da Unicamp (EXTECAMP) foi proposta uma análise dos dados climatológicos da Cidade de Campinas no intervalo de 01/01/2015 a 31/12/2019, utilizando a linguagem R e o R Studio como ferramentas de análise e também, os dados obtidos através do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura da Unicamp (CEPAGRI/UNICAMP), disponibilizados no endereço  **https://www.ic.unicamp.br/~zanoni/cepagri/cepagri**.

Ao longo deste relatório são apresentados os métodos utilizados para o processamento e tratamento dos dados, as dificuldades encontradas com a amostra em questão, assim como a exibição dos resultados da análise exploratória dos dados de forma gráfica e quantitativa.

Todos os trechos de código apresentados neste trabalho fazem referência aos códigos do arquivo **inf0612\_trabalho\_final.R** do repositório <https://github.com/grgouveia/inf0612-trabalho-final> no GitHub.

# TRATAMENTO DOS DADOS

Coletamos os dados abrindo uma conexão com o endereço onde os dados estão armazenados, de acordo com o trecho de código 1.1.

Devido à existência de anomalias nos dados capturados, foi necessário realizar o tratamento dos dados para que a análise não apresentasse resultados distorcidos e portanto, não prejudicasse a compreensão final.

* 1. **OBSERVAÇÕES AUSENTES**

Observações ausentes podem surgir desde a perda de informação bem como a falta de resposta durante a coleta. Nos dados coletados, existem diversas informações ausentes, portanto marcamos essas informações com uma constante lógica indicadora de valor ausente (**NA)**  durante o tratamento dos dados, para depois removê-los de acordo com o trecho de código 1.2.

**2.2 COERÇÃO IMPLÍCITA**

Observamos que a coluna de temperatura foi definida como *factor* durante a leitura dos dados, pois o parâmetro *StringsAsFactors* tem como valor default **TRUE**. Portanto, efetuamos uma conversão de *fator* para *string* e por conseguinte, para *numeric*, como demonstrado no trecho de código 1.3.

**2.3 REMOÇÃO DE OUTLIERS**

Outro erro comum é a existência de *outliers***,** que são valores que fogem do padrão ou que não fazem sentido para o tipo de dado analisado devido a erros de input, como por exemplo, a sensação térmica máxima de 99.9°C encontrada durante um sumário da coluna de sensação térmica que exibe valores mínimos, máximos, mediana, média, primeiro quartil e terceiro quartil dos dados. Efetuando a mesma análise para a umidade foram encontrados valores de umidade iguais a 0, que não correspondem a valores válidos.

O tratamento dessas anomalias envolve a remoção ou substituição por valores padrões e foram realizados no trecho de código 1.4.

**2.4 FORMATAÇÃO DA DATA**

Os dados também podem estar com um formato não compatível com a sintaxe da linguagem R, portanto efetuamos uma conversão dessas informações para o formato POSIXct, de acordo com o trecho de código 1.5.

**2.5 DADOS REDUNDANTES**

Encontramos em nossas análises, trecho de código 1.6, valores repetidos devido a uma interrupção na coleta de dados pelos sensores, entre outros motivos geralmente associados a falhas de sensores. Estes dados foram tratados e as linhas com repetições removidas da base de dados, conforme o trecho de código 1.7 .

# ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória baseia-se nas variáveis *temperatura e sensação térmica* fornecidas pela base de dados da CEPAGRI. A fim de encontrar relacionamentos entre as variáveis disponíveis foram propostas diversas formas de visualização gráfica quantitativa dos dados, que serão apresentadas nos seguintes tópicos.

**3.1 MEDIDAS DE POSIÇÃO**

Após serem realizados os tratamentos necessários na base de dados, foi feita a sumarização das variáveis, conforme o trecho de código 2.1.

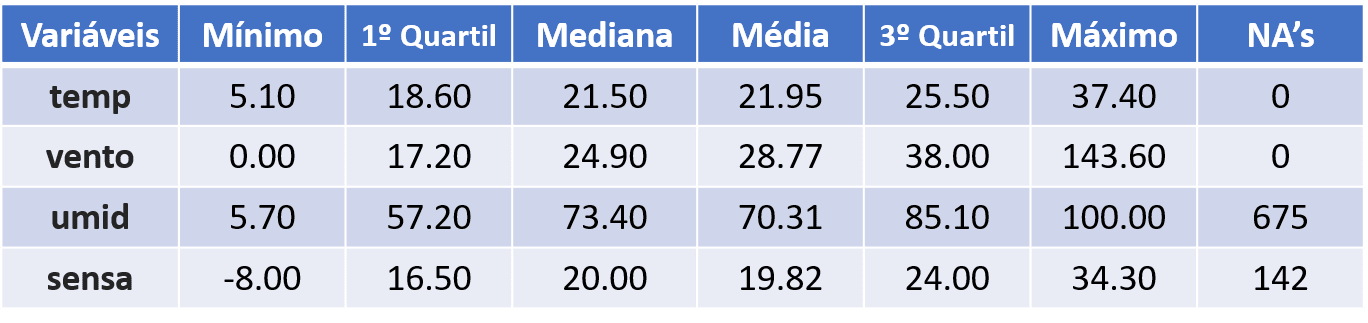


Tabela 1 – Medidas de Posição dos Dados Tratados

Na Tabela 1 acima, observamos que a média e a mediana estão próximas, em todas as variáveis, indicando que a distribuição dos dados está relativamente simétrica e os outliers foram removidos da base de dados.

Para as variáveis que representam a umidade e sensação térmica, foram detectados uma quantidade considerável de valores com algum problema. Estas linhas foram removidas da base de dados, significando que existem períodos que estão com a medição do clima prejudicada.

Existe uma grande variação entre os mínimos e os máximos de todas as variáveis, porém isto é esperado, pois a base registra dados de clima em todas as estações do ano. Existe uma atenção para o mínimo da sensação térmica (- 8.0ºC), indicando que em alguns dias do ano este indicativo do clima ficou fora do esperado para um clima tropical, mas ainda assim, possível de ocorrer. Por este motivo, não foi considerado como um outlier. A mesma interpretação foi feita para os valores máximos e mínimos das variáveis vento e umidade.

**3.2 MEDIDAS DE DISPERSÃO**

No trecho de código 2.2, foi calculado as medidas de dispersão das variáveis. Estas medidas, observam se os valores dos dados estão compactados ou espalhados. [livro: Estatística pratica para cientistas de dados. Editora: O’Reilly].

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

Tabela 2 - Medidas de Dispersão dos Dados Tratados

**3.3 BOXPLOT**

O boxplot permite sumarizar e avaliar a evolução de uma ou mais variáveis, apresentando uma de uma forma rápida e sucinta os dados para a mediana e os quartis antes e após a mediana.

Nesta análise, o boxplot foi utilizado para analisar a evolução da temperatura ao longo do ano para cada ano presente na base de dados, dando enfoque rápido para a mediana dessas temperaturas, conforme ilustrado nos gráficos a seguir:

IMAGEM AQUI

**3.3 COMPARAÇÃO ENTRE AS MEDIDAS COLETADAS**

Para verificar se existe alguma relação simples entre todas as medidas, temperatura, sensação térmica, umidade e velocidade do vento, foi extraído a média por mês das temperaturas, considerando todos as medidas do mesmo mês independente do ano.

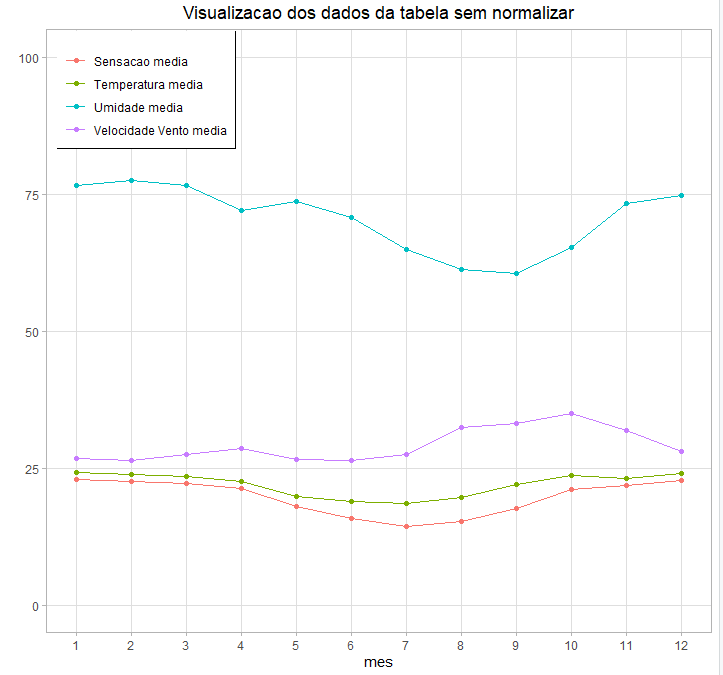
Como não seria medido a frequência dos dados, qualquer valor não encontrado foi removido. Foram encontrados os valores médios conforme a tabela abaixo.

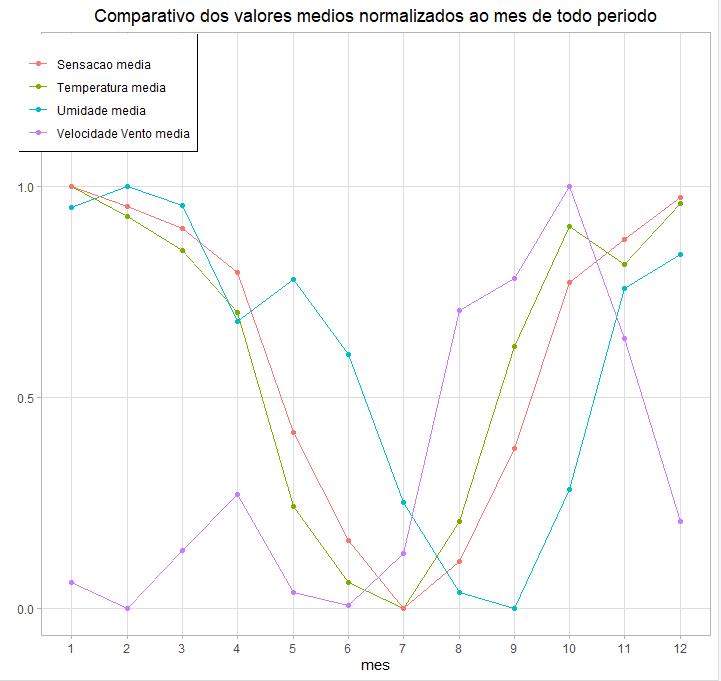


É possível ver pelos dados apresentados que a diferença da sensação térmica costuma ser maior nos meses do inverno. Para comparar a situação da umidade e do vento nessa comparação, foram gerados dois gráficos, um exatamente com os dados da tabela, apenas para representar as médias em uma escala de até 100 , denominado “Visualização dos dados da tabela sem normalizar” e outro com os dados normalizados “Comparativo dos valores médios normalizados ao mês de todo período”.

No gráfico “Visualização dos dados da tabela sem normalizar”, é possível observar que a sensação térmica se distancia mais da temperatura nos meses de inverno, quando a velocidade do vento é mais alta indicando que venta mais e a umidade mais baixa.

No gráfico “Comparativo dos valores médios normalizados ao mês de todo período”,  pode ser observado que em julho estão as temperaturas mais baixas e umidade tende a cair também atingindo o menor valor em setembro.

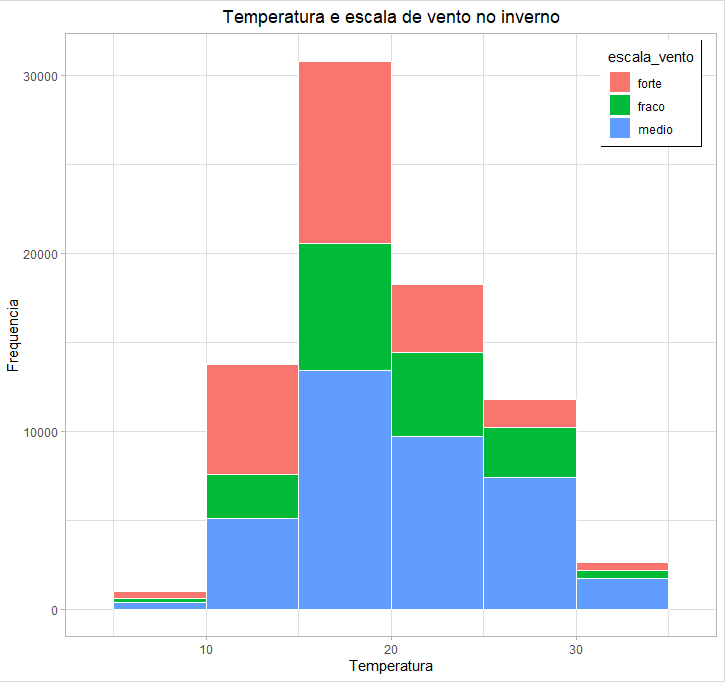


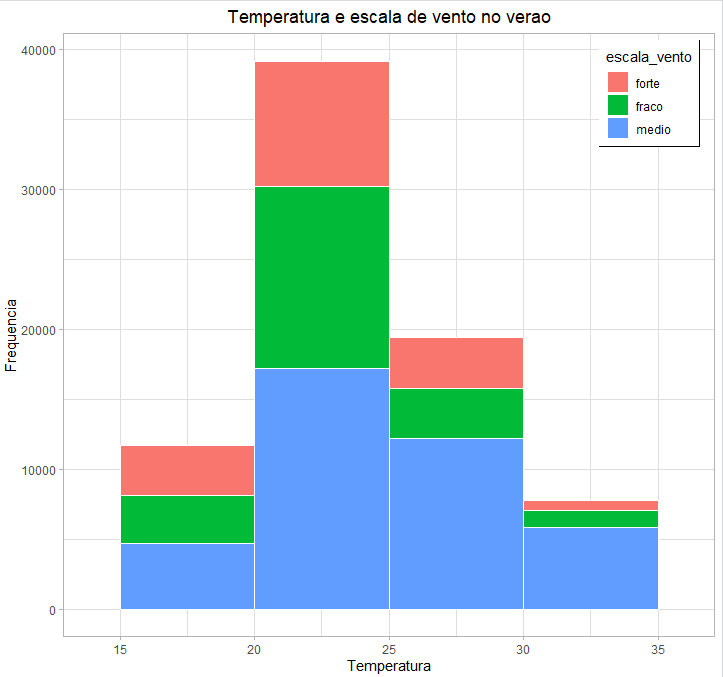
****

**3.4 COMPARAÇÃO VENTO X TEMPERATURA DURANTE AS ESTAÇÕES DO VERÃO E DO INVERNO**

Para essa análise, dois cenários foram considerados, análise da temperatura e aplicando uma escala de vento forte, fraco e médio. Como vento fraco foi considerado os valores abaixo de 17, médio entre 17 e 36 e forte acima de 36. Também foi feita a análise inversa, tendo o foco o vento e a classificação da temperatura como frio, normal, calor e muito calor. Frio foi considerado as temperaturas abaixo de 19 graus, normal entre 19 e 27 graus, calor entre 27 e 31 graus e muito calor acima de 31 graus. O gráfico escolhido foi o de barras, que não dá muita informação e apenas é possível comparar a quantidade dos registros se frequente ou não.

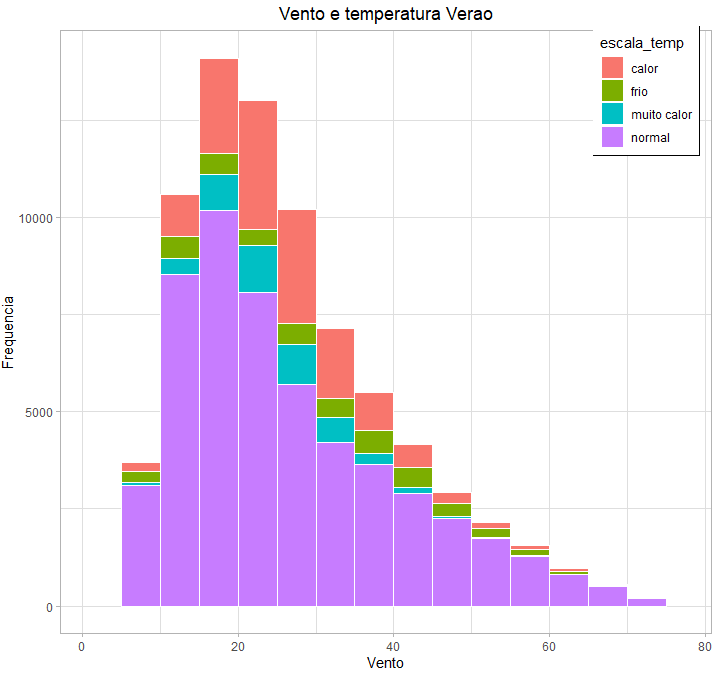
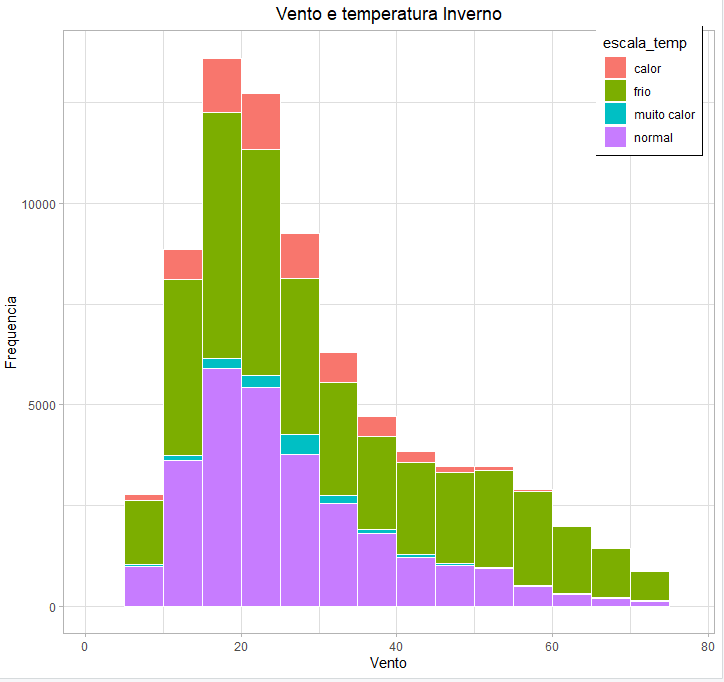
Considerando a primeira análise, referente a temperatura e se o vento estava forte , fraco ou médio, foram obtidos dois gráficos, “temperatura e escala de vento no verao” e “temperatura e escala de vento no inverno”.





É possível observar que no verão, quando a temperatura está acima dos 30 graus, a quantidade de ventos fortes é baixa e obtemos o maior valor de ventos em temperaturas em torno de 20 a 25 graus, a com maior frequência no período. Como não foi separado o dia da noite, esse intervalo com maior frequência não aparenta ser uma distorção dos dados. Já no inverno, é possível notar uma quantidade muito maior de ventos fortes coletadas a temperaturas de 15 a 20 graus, ma se repete o comportamento de poucos ventos fortes a temperatura maior que 30 graus.

   A segunda análise vem mostrar a quantidade de calor nos dias de mais ventos. Foram gerados dois gráficos “Vento e temperatura verão” e “Vento e temperatura inverno” tanto no verão quanto no inverno, na maioria dos dias de muito calor, não estava ventando forte.



Outra possível visão dessas informações é verificar pela estação a frequência de ocorrência de ventos e muito calor. Abaixo a tabela já ordenada pela frequência mostra que em medições de muito calor, a velocidade do vento é baixa se comparada às outras medições.



# CONCLUSÃO

As análises efetuadas são um tanto limitadas devido à insuficiência de informações fornecidas pelo banco de dados da CEPAGRI, que pode ser exemplificada através das variáveis resposta escolhidas para as análises, a sensação térmica e a temperatura. Ambas as variáveis são influenciadas pela umidade do ar, que por sua vez tem relações de causalidade com outras variáveis que fogem ao nosso escopo de análise, como a evaporação das águas oceânicas, movimentação das massas de ar e a cobertura vegetal da região.

Além disso, a presença de muita inconsistência nos dados, assim como períodos de ausência de informação ou repetitividade de dados devido a falhas nos sensores tornam a análise, de certa forma, inconclusiva.