**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.

**2.1 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 1.2.

**2.2 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 1.3. Além disso, para simplificar algumas consultas foram criadas novas colunas em separação de ano, mês e dia.

**2.4 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 1.4.

**2.5 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. Também foi utilizado gráficos simples de boxplot e histograma com os dados ainda contendo outliers. O tratamento dessas anomalias envolve a remoção ou substituição por valores padrões e foram realizados no trecho de código 1.5. Para esses casos, como as outras medidas continham dados validos, colocamos como Nas. Assim mesmo após o tratamento acima para remoção dos Nas, alguns existirão dos dados de forma proposital para sinalizar esse ajuste.

**2.6 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 1.6.

# 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 COM A BASE TRATADA**

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

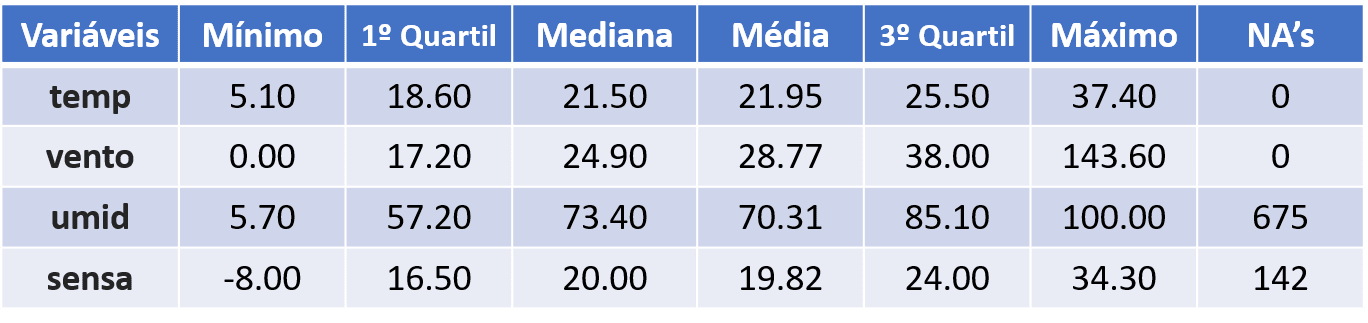


Tabela – Medidas de Posição dos Dados Tratados

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, foi detectada 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 dada para os valores máximos e mínimos das variáveis vento e umidade.

**3.2 MEDIDAS DE DISPERSÃO COM A BASE TRATADA**

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].



Tabela - Medidas de Dispersão dos Dados Tratados

O desvio-padrão na Tabela 2, se mostrou mais disperso na variável umidade. O coeficiente de variação, representa o desvio-padrão expresso como porcentagem da média. Apesar de o vento ter um desvio-padrão menor do que a variável umidade, o coeficiente de variação do vento é maior em relação à média, indicando uma dispersão maior nos dados da variável vento, do que na umidade do ar. Podemos comprovar, observando os mínimos e os máximos das duas variáveis.

**3.3 BOXPLOT E HISTOGRAMA DAS VARIÁVEIS**

Os gráficos do boxplot e histograma possuem o intuito de fornecer informações sobre a variabilidade dos dados. O conjunto de medidas avaliadas nesses gráficos fornece evidências acerca da posição, dispersão, assimetria e valores atípicos.

Imagem em preto e branco com texto preto sobre fundo branco

Descrição gerada automaticamente

Gráfico - Boxplot para cada variável quantitativa

A temperatura possui sua mediana na posição aproximada 21Cº do gráfico, os dados estão distribuídos relativamente de maneira simétrica, como é possível observar no histograma de dispersão e nos quartis do boxplot. As outras variáveis sensação térmica e vento, apresentaram uma quantidade considerável de dados fora do limite superior ou inferior, indicando que existe uma assimetria na frequência das observações e uma maior dispersão.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

O Histograma apresenta a frequência na qual um evento acontece dentro de um intervalo dado. Os histogramas foram criados com frequências diferentes, de acordo com a unidade de medida de cada variável. A variável que mais se mostrou assimétrica em sua distribuição foi vento, tendo o seu pico de frequência na distribuição entre 5 km/h e 10 km/h. A umidade ficou entre 80% e 90%, a sensação térmica entre 18ºC e 20ºC e a temperatura entre 20ºC e 22ºC.

**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. Os valores Nas das medidas sensação térmica e umidade foram removidos, pois os registros não iriam influenciar na análise.

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

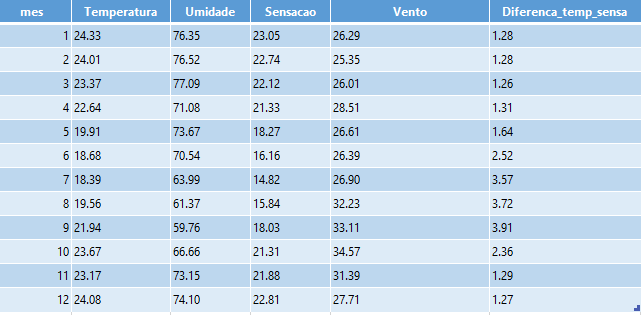


Tabela 3 : Valores médios por mês das medidas coletadas de todos os anos

É possível ver pelos dados apresentados que a diferença da sensação térmica costuma ser maior nos meses do inverno. Além da tabela, foi utilizado o gráfico das medidas, com valores não normalizados para verificar o comportamento durante os meses dos valores. O fato de não estar normalizado para essa análise não impactará nos resultados pois o foco de comparação é temperatura e sensação térmica, que apresentam a mesma escala de valores. As outras medias estão junto para verificar se nos meses que a diferença térmica foi maior, existe alguma queda ou aumento nos valores das outras medidas.

Mapa colorido com texto preto sobre fundo branco

Descrição gerada automaticamente

No gráfico é possível observar que a distância entre as medidas sensação térmica e temperatura é maior nos meses de inverno do que no verão, início de outono e final da primavera. Além disso, quando a umidade começa a diminuir, essa distância começa a aumentar, indicando uma possível relação.

Para verificar a variação das medidas, o gráfico com os valores normalizados foi realizado. No gráfico com os valores normalizados, é possível ver que a temperatura possui uma variação muito mais acentuada que a sensação térmica, apesar das duas seguirem curvas semelhantes nos meses avaliados. A variação da umidade segue uma curva parecida com a temperatura, porém com um deslocamento de meses, indicando alguma relação entre a queda de temperatura e a queda de umidade posterior.

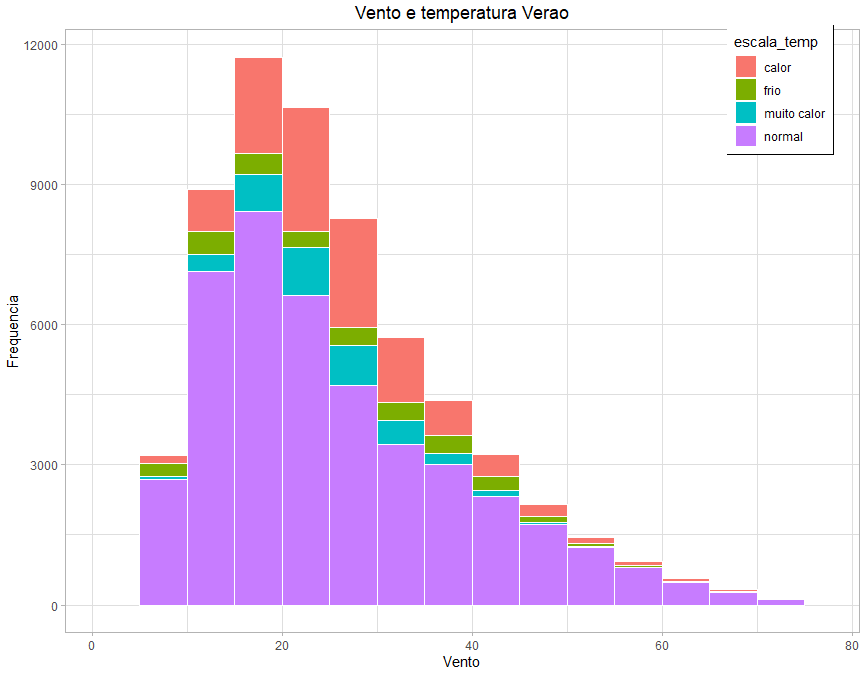
Mapa com linhas coloridas

Descrição gerada automaticamente

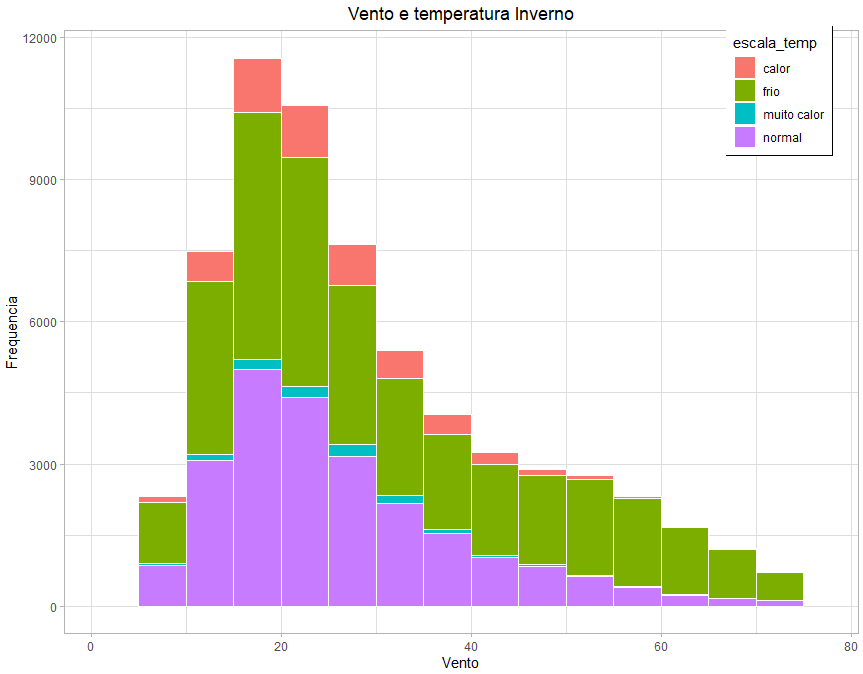
Porém esse análise é superficial, conta com poucos dados, não foi feito uma avaliação em detalhe do período do dia pois a temperatura e umidade mudam durante o dia e noite e foram usadas apenas médias mensais, sendo uma avaliação que não pode ser utilizada para conclusões referente a verdadeira relação entre temperatura e umidade.

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

Para essa análise, foram utilizadas as medidas de vento e temperatura. A temperatura foi classifica em 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. Os valores Nas das medidas sensação térmica e umidade foram removidos, pois os registros não iriam influenciar na análise.



Na estação verão, foi possível verificar que como aumento da velocidade do vento, a classificação de temperatura calor e muito calor apareceram em menor frequência que as demais.



No inverno, já é esperado uma frequência baixa de temperaturas calor e muito calor, porém também sofrem influência nos dias de muito vento apresentou um valor muito pequeno ou nenhuma ocorrência de muito calor e calor. Contudo, novamente o número de dados analisados é pequeno e de uma região apenas, não sendo possível concluir que existe essa relação direta da classificação da temperatura com a quantidade de ventos.

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

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Tabela - Frequência das classificações calor e ventos nas estações inverno e verão

# 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.