Trabalho Final

INF-0612 – Análise de Dados

Prof. Zanoni Dias

1 IDENTIFICAÇÃO

O trabalho é composto pelos discentes:

- Daniele Montenegro da Silva Barros
- Rodrigo Dantas da Silva
- Thiago Bruschi Martins

2 FINALIDADE

Este documento tem por finalidade apresentar o relatório escrito referente ao Trabalho Final da disciplina INF-0612 Análise de dados do Curso de Extensão Mineração de Dados Complexos ofertado pela Universidade Estadual de Campinas - UNICAMP.

3 TRATAMENTO DOS DADOS

Os dados iniciais, sem aplicação de qualquer técnica de tratamento dos dados são compostos por 311922 registros. Executando o comando *head* conseguimos visualizar os seis primeiros registros que fazem parte do conjunto de dados, assim como o nome de suas variáveis. Como detalhado na Tabela 1. Verifica-se assim que o registro 2, de 01 de janeiro de 2015 obteve falha na aquisição, demonstrando assim que será necessária a aplicação de técnicas de tratamento dos dados.

Tabela 1 - Visualização das variáveis e 6 primeiros registros sem o tratamento dos dados

ident	horário	temp	vento	umid	sensa
1	01/01/2015-00:00	22.7	22.2	92.2	21.6
2	**01/01/2015-00:10**	**[ERRO]**	**NA**	**NA**	**NA**
3	01/01/2015-00:20	22.6	23.1	92.5	21.5
4	01/01/2015-00:30	22.7	23.1	92.0	21.6
5	01/01/2015-00:40	22.8	22.7	91.3	21.7
6	01/01/2015-00:50	22.7	22.3	91.6	21.6

Para ter uma visualização sobre a situação dos dados foi realizada uma busca por dados faltantes relacionando mês e ano. A Tabela 2 descreve o quantitativo dos dados faltantes. Verificou-se assim que o ano de 2020 foi o ano que obteve mais erros nas leituras, tendo um número significativo de erros nos meses de maio a dezembro. Um dos possíveis motivos foi o período da quarentena onde o local de coleta dos dados, a Universidade Estadual de Campinas está com a sede fechada.

Tabela 2 - Quantidade de dados faltantes por mês e ano

	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
2015	293	7	3	0	0	220	0	0	0	44	1	0
2016	0	58	1	0	0	0	0	0	1	1	4	0
2017	61	6	454	143	433	11	0	0	0	60	0	0
2018	0	99	0	0	0	0	0	0	147	0	0	0
2019	12	6	4	0	0	0	0	0	0	0	0	0
2020	0	0	0	0	958	1923	4432	3490	4155	4460	459	139

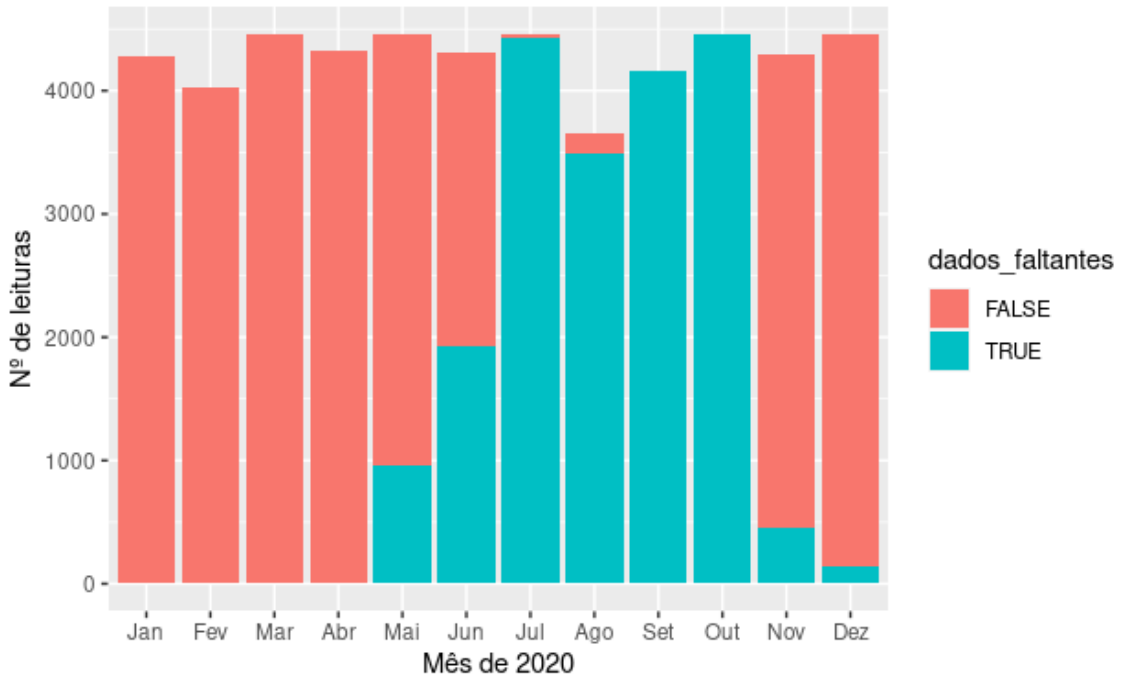

Gráfico 1: Leituras no ano de 2020 - Análise dos dados faltantes por variáveis

Analisando graficamente o conjunto de dados do ano de 2020 observa-se o comportamento dos dados faltantes ao longo dos meses. Neste sentido, a partir de Novembro de 2020, quando o acesso a Unicamp começou a ser possível, os erros foram corrigidos e o número de leituras válidas começou a aumentar, fazendo com que os meses de novembro e dezembro contivessem menos erros. Contudo, os dados do ano de 2020 serão desconsiderados nas análises devido ao grande número de dados faltantes, como dito nesta seção.

Como parte do processo de tratamento dos dados foram utilizadas técnicas de *Data Cleaning* (limpeza dos dados). Agora vamos tratar os dados faltantes e outliers. Como em alguns meses há muitas leituras faltantes, não há muito o que fazer além de deletar esses dados Através do Summary podemos verificar que a sensação térmica possui outliers de valor 99.9.

O processo da busca dos outliers pode ser descrito da seguinte forma:

- Passo 1: Converter temperatura para numérico, agora que as mensagens de erro foram retiradas;
- Passo 2: Remover outliers de temperatura que estão com o valor 99.9;
- Passo 3: Remover outliers de umidade do ar com valor 0.

Com a remoção dos dados faltantes os dados passaram a ser um total de 289.837 registros. Após esse processo houve outra etapa para os dados serem considerados válidos. Foi criada uma rotina para retirar os dados repetidos, com isso os dados considerados para as análises contam com um total de 248.365 registros.

4 ANÁLISES

4.1 Análise da Umidade, Temperatura e Velocidade do Vento

Utilizando um gráfico do tipo *geom_point* conseguimos relacionar 3 variáveis de uma vez. Com a umidade do ar no eixo Y, velocidade do vento no X e temperatura nas cores dos pontos, construímos o Gráfico 2. Podemos observar que quanto mais para esquerda os pontos se encontram, menor é a velocidade do vento, cuja a variância não varia muito ao longo do ano. Já para a umidade, podemos observar uma leve diferença nos meses de Março a Agosto, onde os pontos se concentram um pouco mais para baixo no gráfico, indicando os meses mais secos do ano. Já quando observamos as cores do ponto, podemos verificar que os meses mais quentes são de Janeiro a Abril, que correspondem à estação do Verão no Brasil. Já nos meses de Outono e Inverno, Maio a Setembro, observamos uma coloração mais azulada, indicando temperaturas menores. A tabela 3 agrupa os dados de 2019 pelas estações do ano utilizando a média dos

valores. Assim, podemos confirmar quais são os meses mais quentes e úmidos e quais são os mais frios e secos.

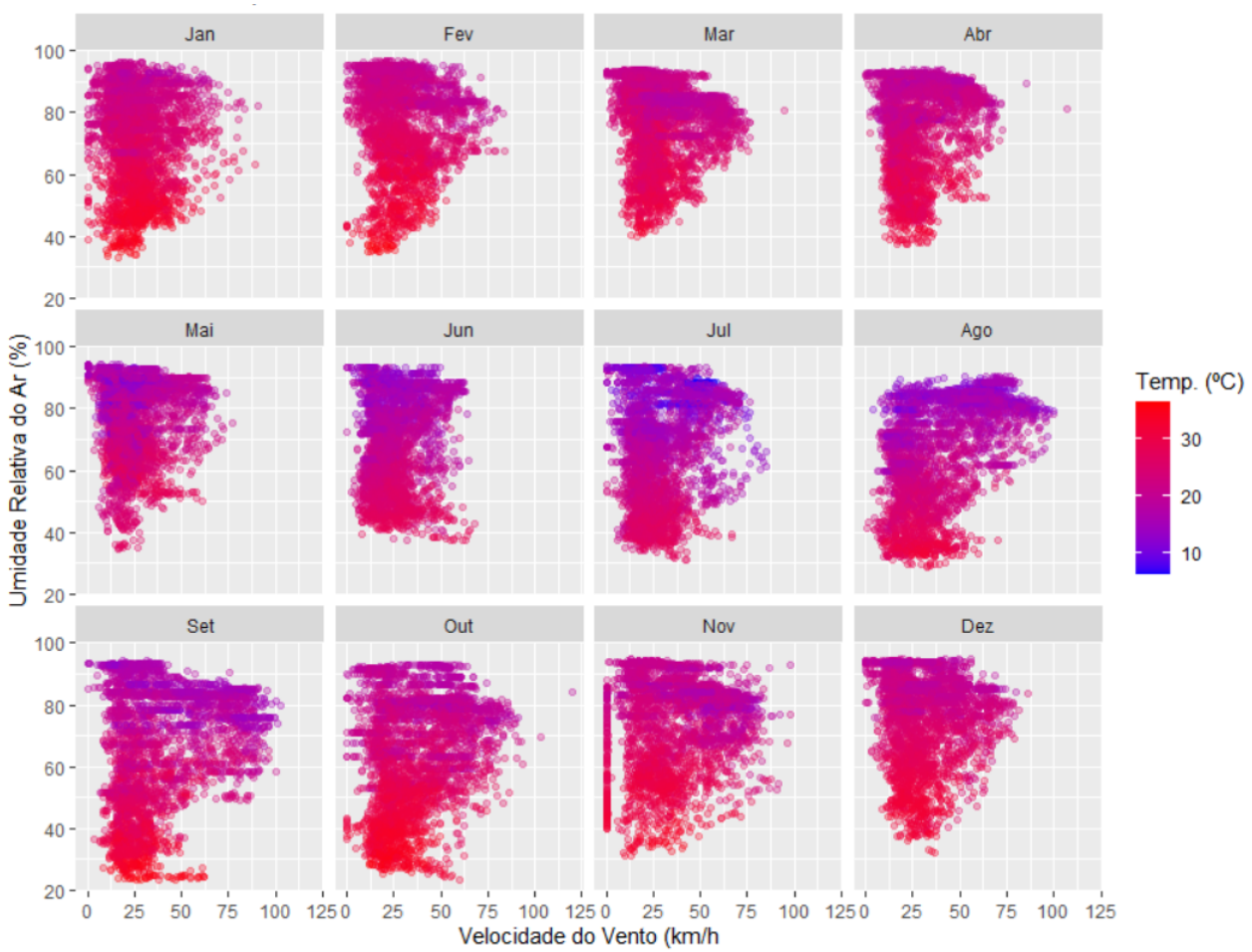

Gráfico 2: Variação de Umidade, Temperatura e Vento no ano de 2019

Tabela 3 - Médias de umidade, temperatura e vel. do vento nos meses de 2019

Estação	Umidade(%)	Temperatura (ºC)	Velocidade do Vento (km/h)
Verão	**76.1**	**24.4**	**27.8**
Outono	**74.6**	**21.2**	**26.4**
Inverno	**66.1**	**19.8**	**35.7**
Primavera	**70.8**	**23.8**	**34.8**

4.2 Comparação entre os Invernos de 2015 a 2018

Com o Gráfico 3, podemos observar as alterações climáticas durante os invernos de 2015 a 2018. Comparando os invernos de 2015 e 2016 com os de 2017 e 2018, podemos notar uma diferença de umidade relativa do ar. Pelo que vemos no gráfico, podemos dizer que os dois últimos anos foram significativamente mais secos que os dois primeiros. Para confirmar essa análise, podemos comparar as médias da umidade relativa do ar ao longo dos anos conforme verificamos a Tabela 4.

Tabela 4 - Médias da umidade relativa do ar entre os invernos de 2015 a 2018

	2015	**2016**	**2017**	**2018**
Umidade	65.21226	66.19420	53.48758	58.10584
Vel. do Vento	27.38188	31.92874	31.20181	27.40754
Temperatura	20.36362	19.59249	19.75805	19.32171

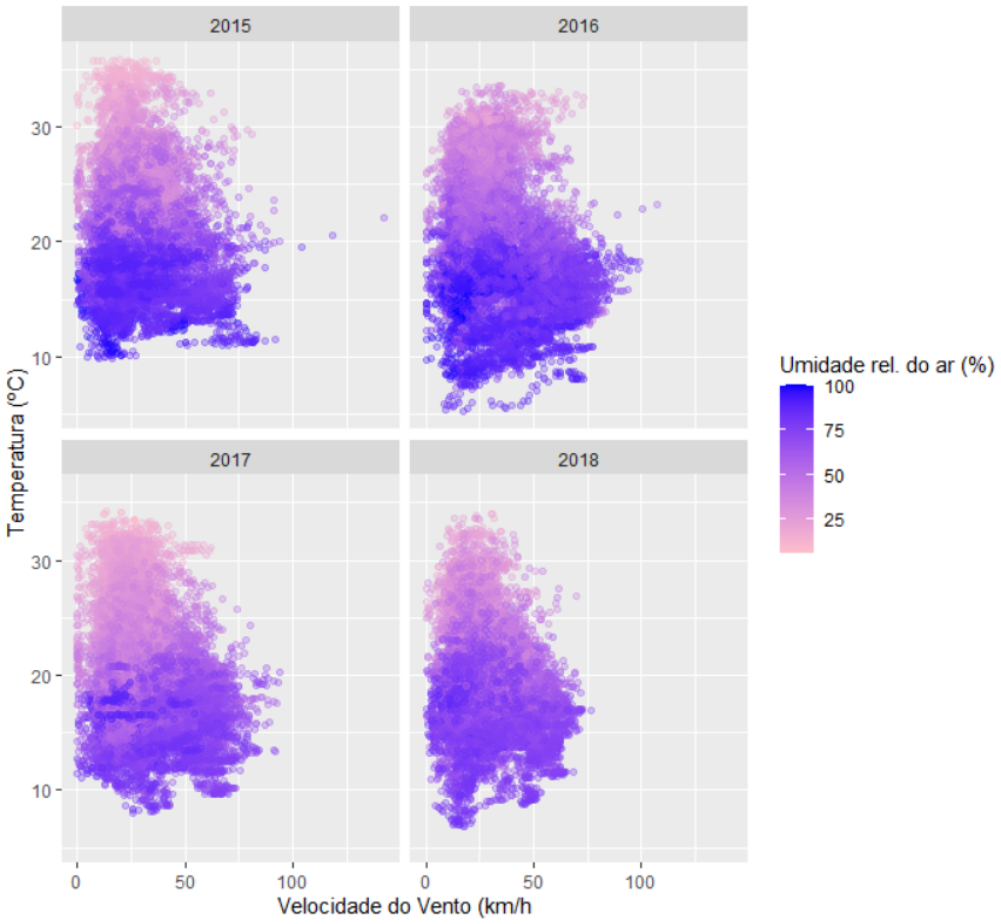

Gráfico 3: Comparação entre os invernos de 2015 a 2018

4.3 Temperaturas por estação

Observando o Gráfico 5 de densidade a seguir podemos comparar a distribuição dos valores de temperatura em cada estação do ano. Nesta análise foram considerados todos os dados de 2015 a 2019, agrupados por estação do ano. Podemos observar uma certa similaridade na diagonal, pois são as estações mais próximas (Verão ~ Primavera e Outono ~ Inverno). Podemos observar também que as temperaturas do Verão e da Primavera são mais acentuadas (frequentes) na região próxima aos 20 graus. Enquanto as temperaturas de Inverno e Outono são mais bem distribuídas.

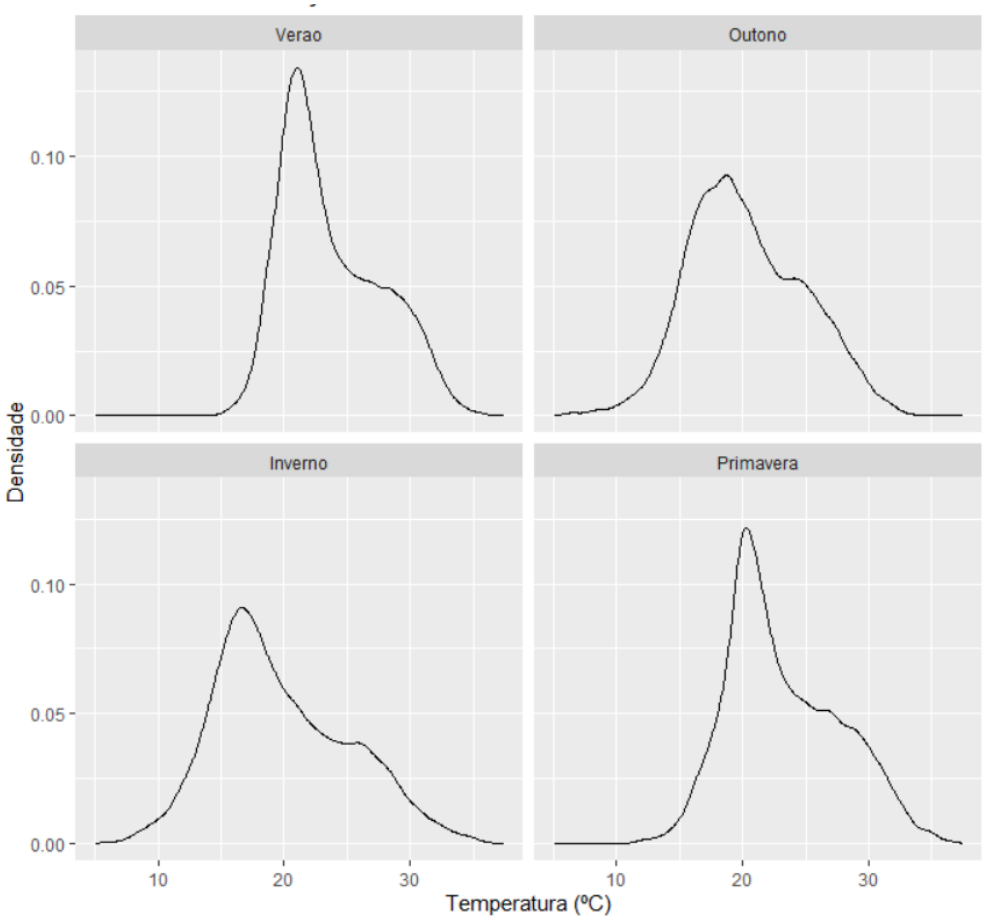

Gráfico 4: Densidade das estações de 2015 a 2019

Para confirmar nossas análises do Gráfico 5, podemos verificar a média e desvio padrão de cada estação, dados pela tabela a seguir. Verificamos que o desvio padrão, que mede a variabilidade dos dados, é maior no Outono e no Inverno. Além disso, verificamos que a estação Verão possui a maior média e menor desvio padrão, ou seja, uma variabilidade menor e maior concentração de temperaturas em torno dos 23.8 ºC. A Tabela 5 descreve a média da temperatura e desvio padrão por estação.

Tabela 5 - Média da Temperatura e Desvio Padrão por Estação

Estação	Média(temp)	Desvio Padrão(temp)
Verão	23.8	3.98
Outono	20.3	4.59
Inverno	19.8	5.34
Primavera	23.2	4.48

4.4 Análise da diferença de entre sensação térmica e temperatura no ano de 2018

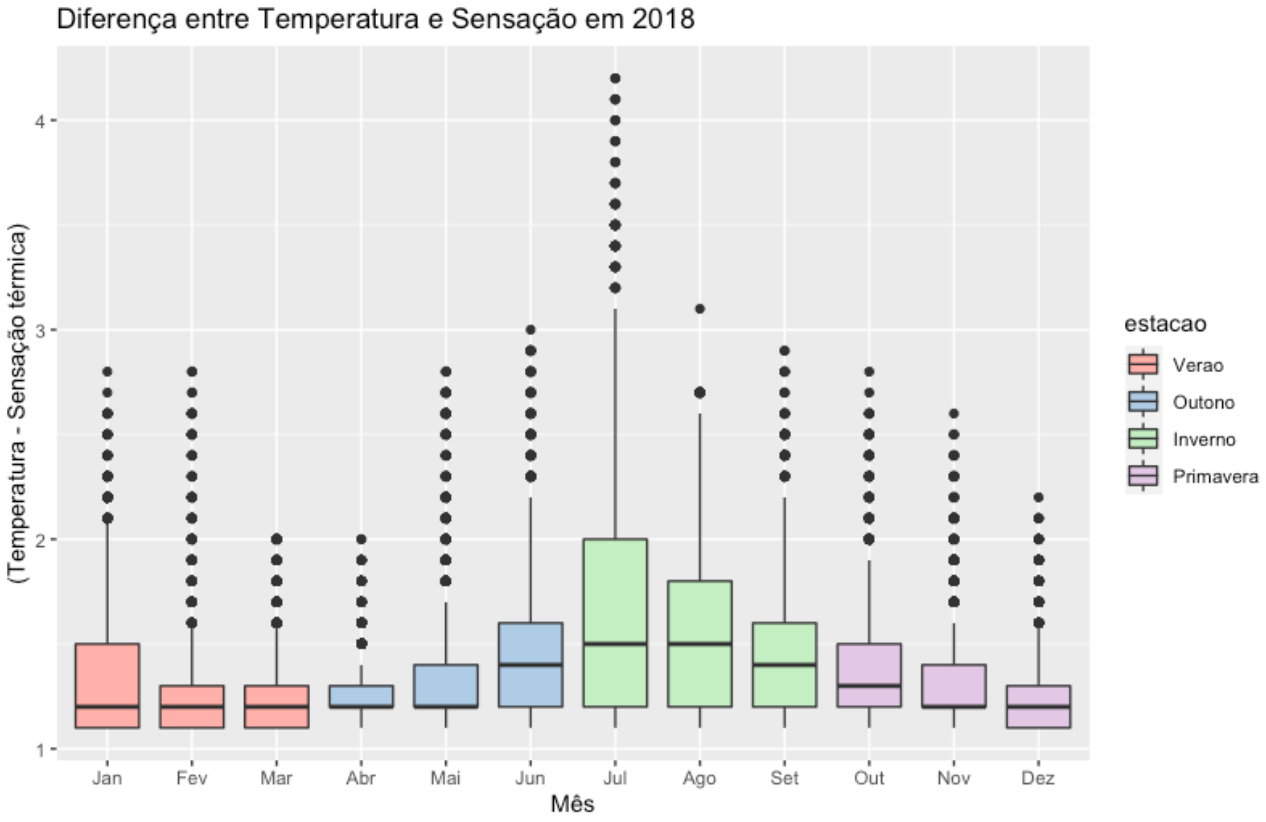

Gráfico 5: Diferença entre temperatura e sensação térmica

Para a construção do gráfico 5, criamos uma nova variável, que é dada pela diferença entre a temperatura e a sensação térmica, a qual chamamos de “diferença”. Essa diferença sofre uma influência direta da velocidade do vento, visto que a fórmula para o cálculo da sensação térmica leva em conta a velocidade do vento. Embora nesta fórmula não entre a umidade do ar, quando olhamos o gráfico de 2019, verificamos que pode haver uma relação da umidade do ar com a sensação térmica, pois nos meses mais secos, que são os meses de Inverno, houve uma diferença maior entre a temperatura e a sensação térmica. Nós tentamos calcular essa mesma diferença para outros anos, porém, nos anos de 2015 e 2016 há um comportamento muito diferente desta variável, como podemos ver no gráfico a seguir no gráfico de 2016.

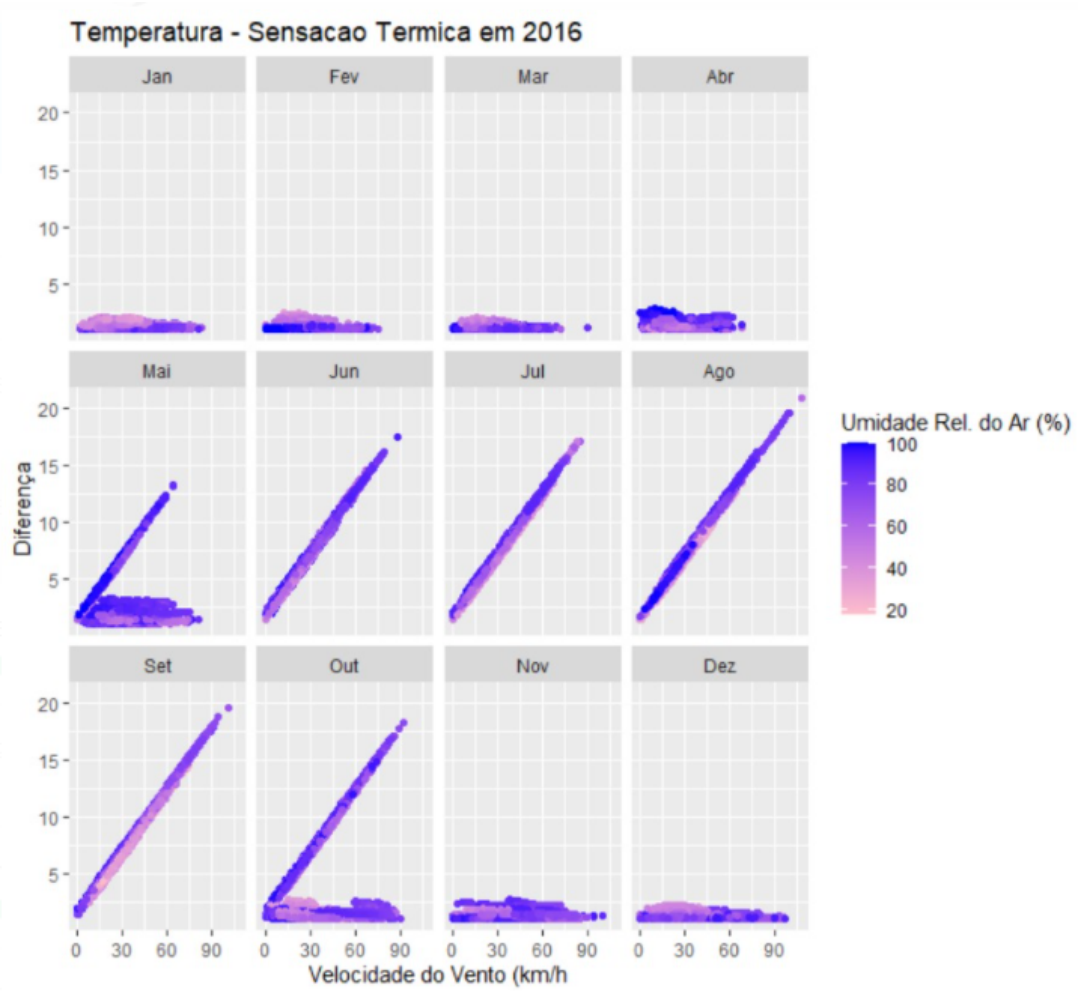

Gráfico 6. Comportamento da diferença entre Temperatura e pressão ao longo do ano de 2016

Embora não consigamos tirar alguma conclusão sobre esse comportamento tão acentuado nos anos de 2015 e 2016, podemos notar que a Diferença também é maior nos meses mais secos do ano, assim como em 2018.