Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Científica

Departamento de Estatística

ME731 - Análise Multivariada

Técnicas de Classificação e Agrupamento

Leonardo da Silva Araújo

RA: 220044

CAMPINAS

2024

Introdução

O ar que respiramos tem uma grande importância e impacta diretamente a saúde humana, ambientes com qualidade do ar renovado oferecem menos riscos à saúde, especialmente se tratando de doenças respiratórias; o ozônio troposférico, que contribui para a poluição e má qualidade do ar, vem aumentando ao longo da história humana, devido a industrialização, trouxe também impactos negativos na qualidade do ar que respiramos, contribuindo para a poluição atmosférica. Quais as variáveis que contribuem para a poluição? O que a temperatura, ozônio, radiação solar e até velocidade do vento, impacta na qualidade do ar, entanto, o ozônio troposférico (aquele presente na camada mais próxima da superfície da Terra, até cerca de 10 km de altitude) é um poluente perigoso. O que essas variáveis tem em comum? Quais padrões podemos obter dos dados sobre as variáveis contidas nos dados? Isso veremos a seguir.

O objetivo aqui é analisar os dados *airquality* disponível no Rstudio, e averiguar quão significativo cada variável interfere na qualidade do ar, isso através de uma forma adequada, utilizando técnicas de classificação e agrupamento se utlizando do rigor estatístico, com construção de um modelo estatístico adequado para os dados, com análises descritiva, exploratória e inferencial, diagnósticos e discutir os resultados e *insights* obtidos. Os dados *airquality* que dispomos são medições diárias da qualidade do ar em Nova York, ao longo de 5 meses, sendo a partir do mês de maio a setembro de 1973.

Dados

Para cumprir os objetivos deste trabalho, dispomos de dados provenientes do Rstudio, chamado *airquality* e trata-se de uma análise que inclui informações sobre a qualidade do ar na cidade de New York nos EUA, ao longo de 5 meses, do mês de maio a setembro de 1973. No banco de dados analisado, registra-se informações de 153 observações e 6 variáveis, são elas:

- Ozone: Ozônio troposférico médio em partes por bilhão (ppb) de 13h00 a 15h00 horas na Ilha Roosevelt.
- Solar.R: Radiação solar em Langleys na faixa de frequência de 4000 a 7700 Angstroms das 08h00 às 12h00 no Central Park.
- Wind: Velocidade média do vento em milhas por hora às 07:00 e 10:00 horas no Aeroporto LaGuardia.
- Temp: Temperatura máxima diária em graus Fahrenheit no Aeroporto de Laguardia.
- Month: Meses
- Day: Dia

Metodologia

Foi utilizado o Software Rstudio na importação dos dados e foi verificado se há observações faltantes, como também organização dos dados e renomeação de alguma variável que fosse importante para simplificar nossa análise. Em seguida, foram obtidas estatísticas sumárias das variáveis estudadas. Após, foi criado gráfico de dispersão em pares, para um melhor entendimento das variáveis ou tendência da relação entre elas. Em seguida, outro gráfico de dispersão em pares e suas correlações. Posteriormente, viu-se a necessidade de criar um *boxplot* da concentração de ozônio troposférico mês a mês a fim de verificar seu comportamento e possíveis *outliers*. Logo mais adiante, as médias mensais das variáveis através de uma Tabela. Seguiu-se, de outro *boxplot* mas dessa vez com as concentrações de ozônio troposférico de cada dia da semana ao longo dos meses.

Elaboramos três gráficos de dispersão ao longo de todos os 5 meses analisados, que relaciona concentração de ozônio troposférico (ppb) x temperatura (F°), concentração de ozônio troposférico (ppb) x velocidade média do vento (mph) e radiação solar (Ly) x temperatura (F°), respectivamente.

A seguir, foi elaborado um modelo de regressão linear múltipla:

$Ozone^* = \beta_0 + \beta_1 Temp + \beta_2 Solar.R + \beta_3 Wind + \epsilon$

Onde:

- $Ozone^*$ é a variável dependente que queremos prever,
- β_0 é o valor do intercepto,
- $\beta_1, \beta_2, \beta_3$ são os coeficientes para as variáveis independentes Temp, Solar.R e Wind, respectivamente.
- $\epsilon \sim \aleph(0, \sigma)$

Após isso, foi elaborado uma Tabela, onde se encontram as estatísticas sumárias do modelo ajustado. Em seguida, foi mostrado um exemplo de melhorias que poderia ser implementado no modelo, utlizando-se transformação por raiz quadrada numa das variáveis, para fins de exibir uma melhoria que poderia ser adotada ao modelo.

Logo em seguida, foi utilizado o método de Elbow:

$WS_k = \sum_{n=1}^{\mathrm{n}k} (x_i - c_k)^2$

Onde:

- n_k é o número de pontos no cluster k,
- x_i é o ponto de dados,
- c_k é o centróide do cluster k.

E a soma de todos os clusters seria:

$WS = \sum_{k=1}^{k} WS_k$

Por fim, elaboramos uma árvore de decisão que se utilizam de critérios de divisão para cada nó, onde alguns dos métodos são, impureza de Gini para um nó t e entropia que mede a incerteza ou impureza do nó:

Impureza de Gini:

$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2$

Onde:

- C é o número de classes,
- p_i é a probabilidade de um exemplo ser da classe i no nó t.

Entropia:

$Entropy(t) = -\sum_{i=1}^{C} p_i \log_2(p_i)$

Onde:

- C é o número de classes,
- p_i é a probabilidade de um exemplo ser da classe i no nó t.

Primeiramente, é importante alertar que, a não ser quando relatado o contrário, os resultados desta seção foram obtidos por meio de funções disponíveis na biblioteca básica da versão 4.4.1 do Rstudio. Todos os gráficos como dados estatísticos da análise realizada, foi escrita através do *overleaf* LaTeX online o que permitiu gerar este relatório em PDF.

Resultados

Análise Descritiva

A Tabela1 traz algumas estatísticas sumárias de algumas variáveis incluídas no modelo escolhido. Posteriomente foi identificado observações faltantes, como podemos ver na Tabela1 e foi desconsiderado as variáveis Month e Day por sua irrelevância nessa estatística sumária. Embora a informação na tabela haja a quantidade de NAs (dados faltantes), houve remoção dos NAs para sumarizar essas estatísticas apresentadas.

Tabela 1: Estatísticas sumárias de variáveis selecionadas

variável	mín	1° quartil	média	mediana	3° quartil	máx	obs. faltante
Ozone	1.0	18.0	42.1	31.0	62.0	168.0	37
Solar.R	7.0	113.5	184.8	207.0	255.5	334.0	7
Wind	2.30	7.40	9.94	9.70	11.50	20.70	0
Temp	57.00	71.00	77.79	79.00	84.50	97.00	0

Pares de gráfico de dispersão

Para um melhor entendimento dos dados e para determinar quais pares de variáveis, estão fortemente correlacionados e quais estão fracamente correlacionados, foi criado gráficos de dispersão em pares das variáveis: *Ozone, Solar.R, Wind* e *Temp*. Pelo gráfico da figura1 há algumas associação entre as variáveis, porém de antemão, as que se destacam e que há uma associação positiva forte são entre as variáveis *Ozone* e *Temp*. Em contrapartida, outra observação que vale a pena destacar é que parece haver indícios de associação negativa moderada entre *Wind* e *Temp*.

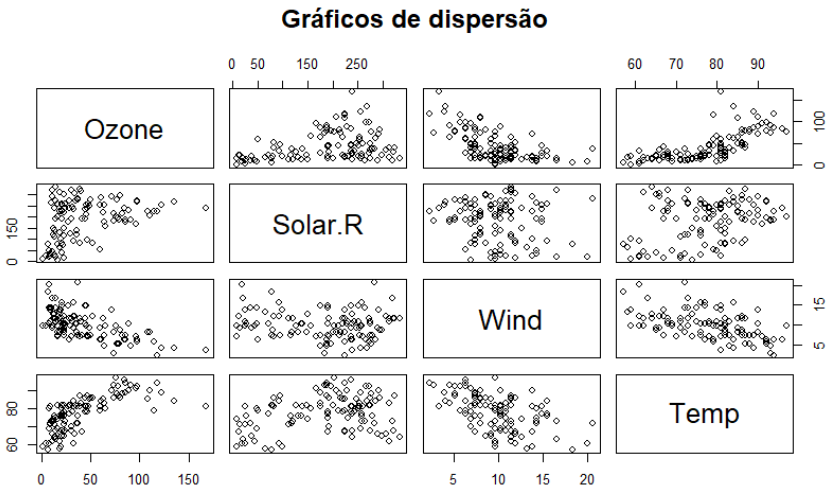

Figura 1: Pares de gráficos de dispersão

Pares de gráficos de dispersão e correlação

No gráfico da figura2 há informações das correlações e associações entre as variáveis. Como apontado anteriormente através da figura1 uma associação positiva forte entre *Ozone* e *Temp* com $R^2 = 0.699$. Nota-se associação negativa moderada entre *Wind* e *Temp* com $R^2 = -0.612$. As associações entre as variáveis que se destacam aqui são *Ozone* e *Temp* como também *Wind* e *Temp* com correlação positiva forte e negativa moderada, respectivamente.

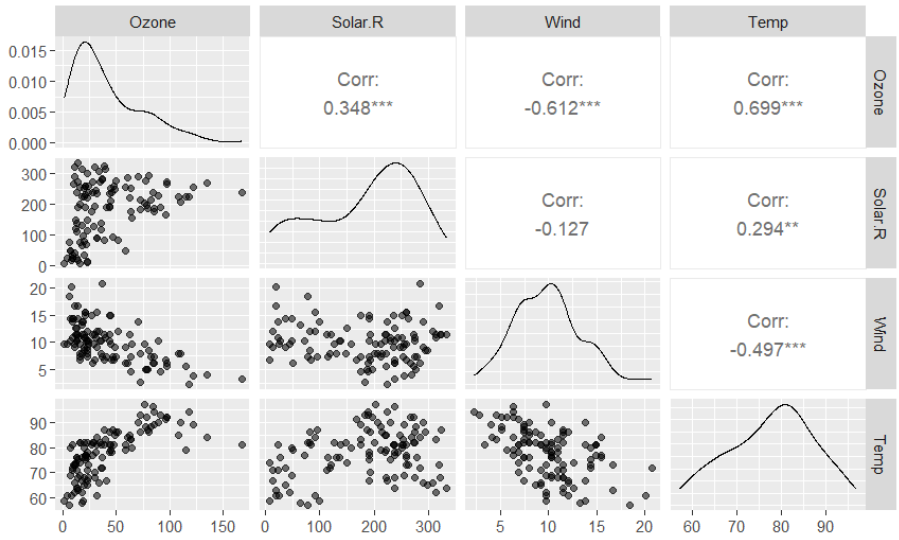

Figura 2: gráfico de dispersão e correlação

Boxplot concentração de ozônio mês a mês

O *boxplot* da figura3 nos informa acerca da concentração de ozônio troposférico ao longo dos meses analisados. Nota-se que tal concentração aumenta gradativamente de maio a junho, onde há um pico, de julho a agosto, decaindo a partir de setembro. Há presença de *outliers*, principalmente no mês de setembro. No mês de julho e agosto é verão nos EUA, onde as temperaturas são elevadas e há estiagem das chuvas nesse período, que combinado com altas temperaturas pode causar um aumento nas concentrações de ozônio na atmosfera. Contudo, corroborando com o gráfico da figura1 onde é possível notar que de fato, há correlação positiva entre *Ozone* e *Temp* e aponta que, quanto maior a temperatura maior também são as concentrações de ozônio troposférico e que aponta que essas variáveis estão correlacionadas.

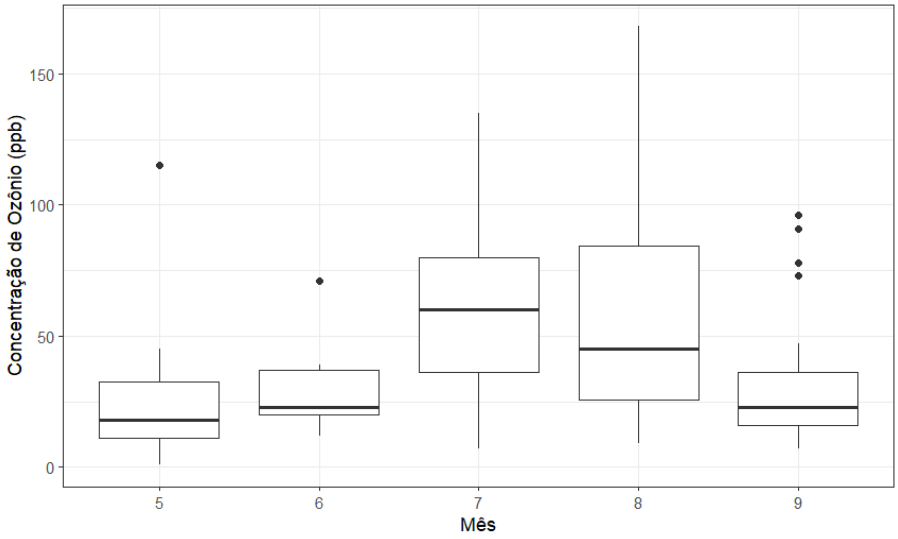

Figura 3: Boxplot concentração ozônio mensal

Médias mensais das variáveis

A Tabela2 nos traz informações sobre as médias das concentrações de ozônio troposférico, radiação solar em Langleys (Ly), velocidade do vento em milhas/hora (mph) e temperatura em graus Fahrenheit (F°) mensalmente. Observe que como apontado anteriomente através do gráfico da figura3 as maiores médias de concentrações de ozônio troposférico se dá no mês 7 e 8, respectivamente. É interessante notar que no mês 7 e 8 também há picos das temperaturas médias e menor velocidade do vento durante o período. Por último, a radiação solar atinge seu maior valor no mês 7 como mencionado anteriormente.

Tabela 2: médias mensais das variáveis

Month	mediaO3	mediaSolar	mediaWind	mediaTemp
5	24.1	182	11.5	66.5
6	29.4	184	12.2	78.2
7	59.1	216	8.52	83.9
8	60	173	8.86	83.7
9	31.4	168	10.1	76.9

Boxplot ozônio semanal ao longo dos meses

O *boxplot* produzido na figura4 nos permite verificar a variação da concentração de ozônio troposférico para cada dia da semana ao longo dos meses analisados e é de se notar que no dia 1 que corresponde a segunda-feira apresenta uma maior variação da concentração de ozônio. No dia 2 (terça-feira) parece ser similar a variação do dia 1. Os demais dias, apesar das distribuições diferentes, parece se manter no mesmo patamar, apesar de que no dia 7 (domingo) volta a subir levemente a concentração de ozônio troposférico e talvez isso seja pelo fato de haver mais automóveis em circulação na cidade.

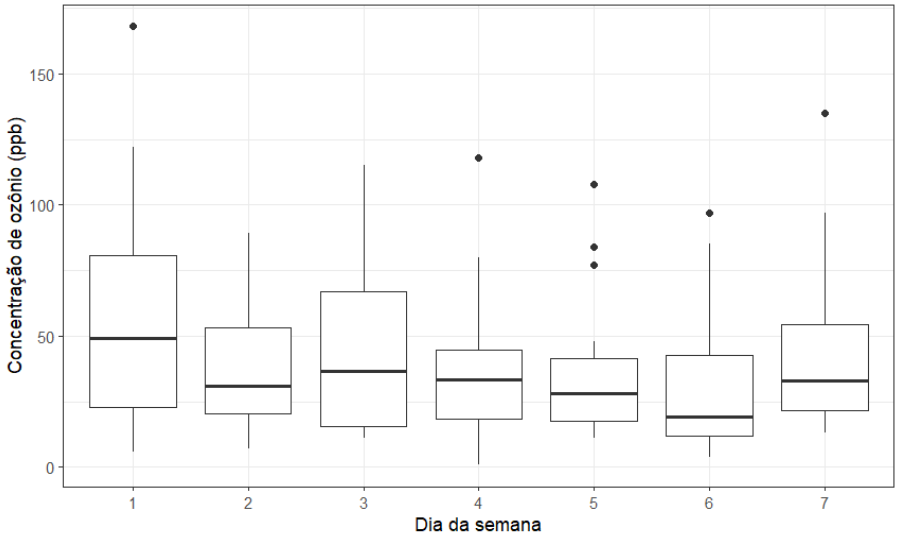

Figura 4: Boxplot concentração ozônio semanal

Gráfico dispersão ozônio x temperatura

Gráfico da figura5 relaciona concentração de ozônio troposférico e temperatura ao longo dos meses analisados e note que esse gráfico se encontra na figura1 e ademais, vimos que na figura2 que há correlação positiva forte, ou seja, que há correlação entre essas duas variáveis.

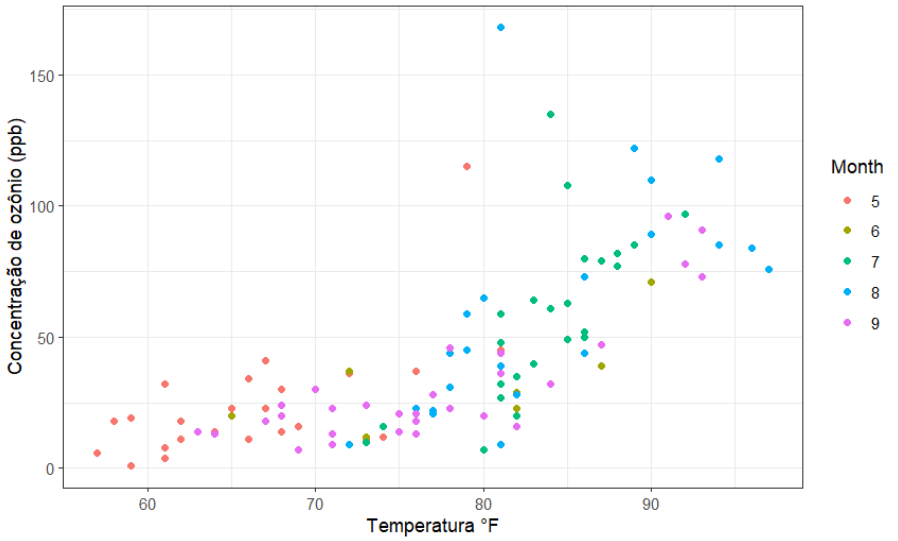

Figura 5: Concentração ozônio x temperatura

Gráfico dispersão ozônio x velocidade média do vento

Na figura6 relaciona as variáveis concentração de ozônio troposférico e velocidade média do vento ao longo dos meses analisados. Note que há uma correlação negativa entre as variáveis, onde quando a velocidade média tende aumentar, a concentração de ozônio troposférico, tende a diminuir.

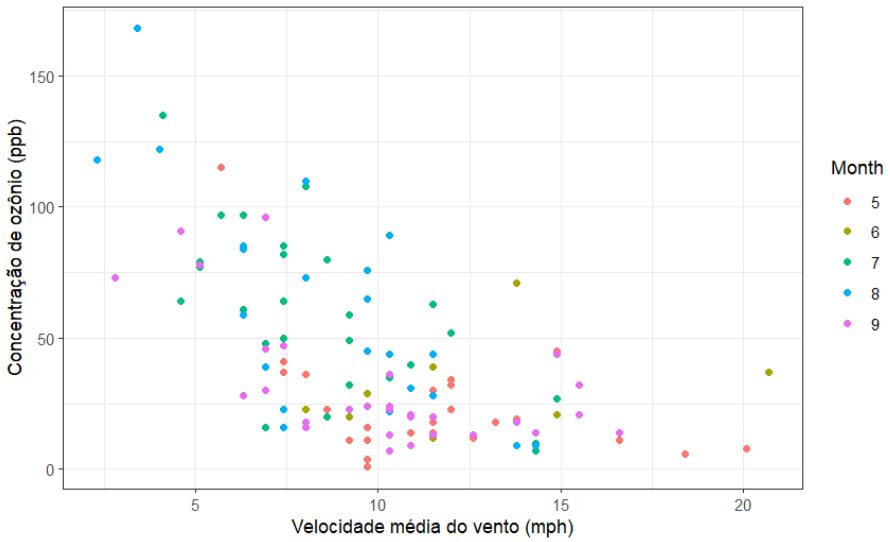

Figura 6: Concentração ozônio x velocidade média vento

Gráfico dispersão concentração ozônio x radiação solar

Mais adiante, o figura7 relaciona as variáveis concentração de ozônio troposférico e radiação solar ao longo dos meses analisados. Notamos que o gráfico não nos mostra nenhuma tendência ou correlação entre as variáveis, com observações dispersas e desordenadas.

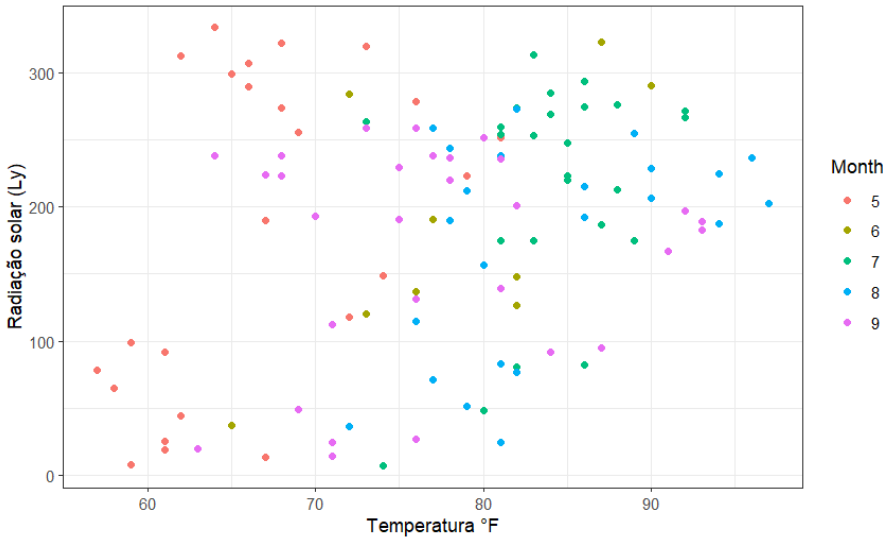

Figura 7: Concentração ozônio x radiação solar

Modelo Regressão Linear Múltipla

A seguir foi escolhido, inicialmente, um modelo de Regressão Linear Múltipla para a variável resposta dependente *Ozone** que incluísse, *Temp, Solar.R* e *Wind*. Da equação do modelo:

$Ozone^* = \beta_0 + \beta_1 Temp + \beta_2 Solar.R + \beta_3 Wind + \epsilon$

Onde:

- *Ozone** é a variável dependente que queremos prever,

- β_0 é o valor do intercepto,
- $\beta_1, \beta_2, \beta_3$ são os coeficientes para as variáveis independentes Temp, Solar.R e Wind, respectivamente.
- $\epsilon \sim \aleph(0, \sigma)$

Da Tabela3 podemos visualizar os valores sumários do ajuste do modelo. De acordo com a literatura, adotamos $\alpha = 0.05$ para nossas hipóteses de estimativas significativas.

Tabela 3: summary do ajuste

Coefficients	Estimate	Std.Error	t value	Pr($>\|t\|$)
Intercept	-64.34208	23.05472	-2.791	0.00623
Temp	1.65209	0.25353	6.516	2.42e-09
Solar.R	0.05982	0.02319	2.580	0.01124
Wind	-3.33359	0.65441	-5.094	1.52e-06

Residual standard error: 21.18 on 107 degrees of freedom Multiple R-squared: 0.6059, Adjusted R-squared: 0.5948 F-statistic: 54.83 on 3 and 107 DF, p-value: $<$ 2.2e-16

Note que o p-valor associado a essa variável é muito baixo (2.42e-09), isso indica que a temperatura tem um efeito estatisticamente significativo sobre *Ozone**. Já o p-valor do *Solar.R* é de 0.01124 indica que *Solar.R* tem um efeito significativo, mas não tão forte quanto Temp. Por fim, o p-valor de *Wind* é (1.52e-06) e indica que a variável *Wind* também tem um efeito muito significativo sobre *Ozone**, porém seu efeito é negativo no modelo.

Além do mais, obtivemos informações sobre, *Residual stardard error, R-squared, adjusted R-squared e F-statistic*. O *Multiple R-squared* = 0.6059. Um R-quadrado de 0.6059 significa que o modelo explica 60.59% de *Ozone**. Pode-se considerar que é um ajuste razoável, mas ainda há 39.41% da variação não explicada pelo modelo. O *adjusted R-squared* = 0.5948 é o r-quadrado ajustado e leva em conta o número de variáveis independentes no modelo. Ele é útil para comparar modelos com diferentes números de variáveis. O valor de 0.5948 é um pouco mais baixo que o R-quadrado, o que é normal e reflete que o modelo não está explicando 100% dos dados, mas ainda assim é razoável.

Por fim, o modelo de regressão linear múltipla mostrou que as variáveis *Temp, Solar.R* e *Wind* têm um efeito significativo sobre a variável dependente. O modelo é estatisticamente significativo (p-valor muito pequeno para a estatística F) e explica uma boa parte da variação nos dados (R-quadrado de 0.6059). No entanto, 39.41% da variação nos dados ainda não é explicada, o que sugere que pode haver outros fatores não incluídos no modelo que influenciam a variável dependente mas por ora, ficaremos com este modelo mas poderia ser melhorado e um exemplo disso poderia ser utilizando-se de transformação da raiz quadrada, porém é aceitável e razoável para nosso contexto, o modelo ajustado atual.

Histograma concentração de ozônio (ppb)

Para fins de exemplo o uso de transformação, para mitigar os efeitos da variabilidade e para fins de normalização, temos a distribuição da concentração de ozônio troposférico na figura8. Nota-se como visto anteriomente na Tabela1 que a média da concentração média da variável *Ozone* em ppb é 42.1 (linha vermelha tracejada) da figura8 e há assimetria positiva.

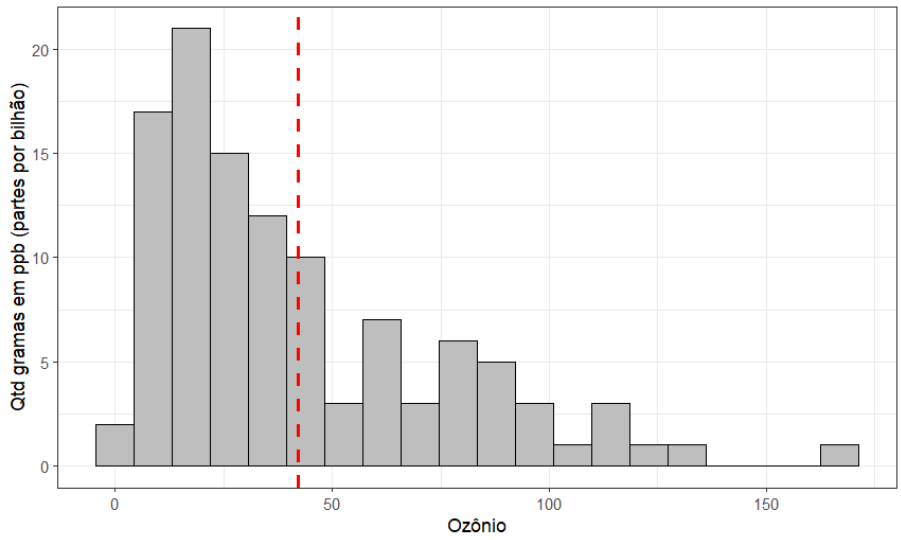

Figura 8: concentração de Ozônio (ppb)

QQ-plot concentração de ozônio (ppb)

Para constatar com o que foi dito anteriormente sobre a distribuição dos dados pelo histograma, através do gráfico do qq-plot(Quantile-Quantile) da Figura9 os pontos não se aderem a reta do gráfico qq-plot o que indicaria uma distribuição normal. A presença de caudas pesadas é notável no qq-plot com assimetria positiva (concavidade para cima) e com presença de *outliers* que se destacam nas extremidades da reta.

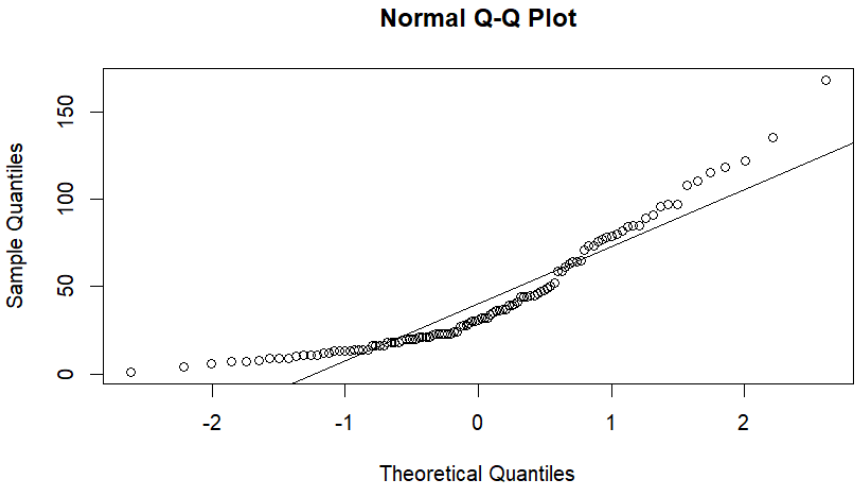

Figura 9: QQ-plot concentração de ozônio

Histograma transformação concentração de ozônio

Embora percamos interpretabilidade, para mitigar os efeitos das caudas pesadas aplicamos uma transformação de raiz quadrada nos dados das concentrações da variável *Ozone* (ozônio), ilustrado na figura10.

Donde:

$y\prime = \sqrt{y}$

Onde:

- y é o valor original da variável
- $y\prime$ é o valor transformado

Em que a transformação por raiz quadrada é aplicada a dados numéricos para reduzir a assimetria e normalizar a distribuição, especialmente em variáveis que apresentam uma distribuição assimétrica à direita, nosso caso. Na figura10 note que embora não haja uma simetria perfeita, há uma distribuição representada pelo histograma mais próximo a normalidade do que antes da aplicação da transformação. Sabendo-se que a média é sensível a *outliers* e mesmo com a presença de alguns deles a média representada pela linha tracejada vermelha, se mantém no centro do histograma, indicando um dos preceitos de normalidade.

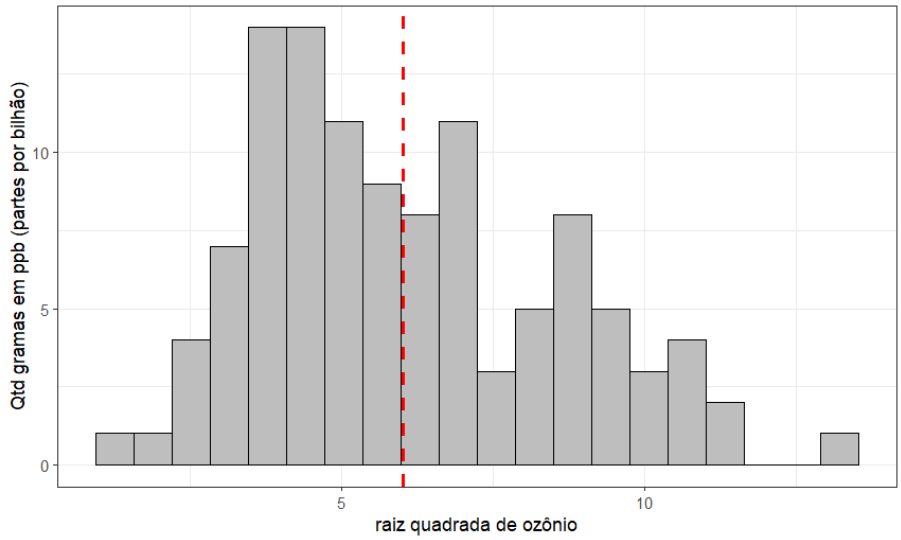

Figura 10: concentração de Ozônio (ppb) transformada

QQ-plot concentração de ozônio (ppb) da transformação

Em seguida, para corroborar ao gráfico da figura10 criamos um qq-plot com os dados *Ozone* se utilizando da transformação pela raiz quadrada, notamos que os pontos do qq-plot da figura11 em sua maioria houve boa aderência a reta linear, embora haja uma presença de cauda pesada a esquerda e alguns *outliers*, entretanto os pontos se encontram mais linear em relação a reta e vale destacar que após a utilização da transformação, foi possível mitigar grande parte do efeito de não normalidade dos dados *Ozone*.

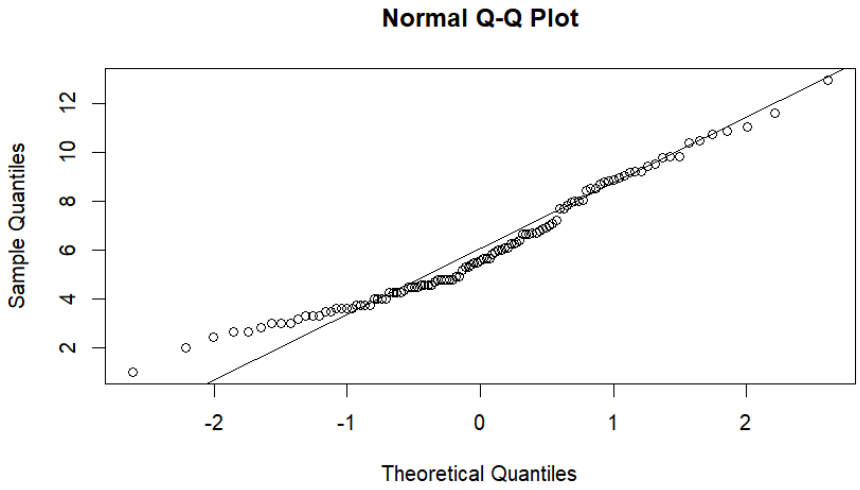

Figura 11: QQ-plot concentração ozônio com transformação

Inércia e método de Elbow

Para medir a qualidade de um agrupamento, uma métrica utilizada é a inércia ou (soma dos quadrados dentro dos clusters, ou within-cluster sum of squares — WCSS). A fórmula para a inércia para um cluster k é:

$WS_k = \sum_{n=1}^{nk} (x_i - c_k)^2$

Onde:

- n_k é o número de pontos no cluster k,
- x_i é o ponto de dados,
- c_k é o centróide do cluster k.

E a soma de todos os clusters seria:

$WS = \sum_{k=1}^{k} WS_k$

O método de *Elbow*, conhecido como o método do cotovelo nos ajuda a identificar o número ideal de *clusters* - agrupamentos, onde a redução da inércia (*Within-cluster sum of squares* - soma dos quadrados dentro dos clusters) desacelera. Quanto menor for os valores da inércia, melhor ou mais compactos são os clusters. A inércia significa que os clusters estão bem compactados, mas após um certo número de clusters, a melhoria diminui significativamente. Na figura12 visualizamos os valores da inércia e o número de clusters. Note que do 2° para o 3° cluster, há uma estabilização dos valores da inércia, formando o que chama-se de "cotovelo"e após isso qualquer adição de clusters não terá utilidade ou não terá eficácia.

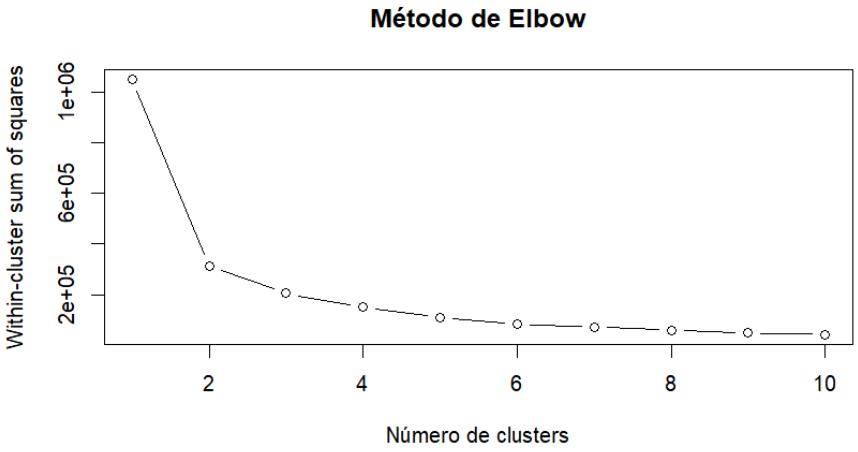

Figura 12: Método de Elbow

Árvore de decisão

A principal vantagem de uma *Decision tree* (Árvore de decisão) é sua fácil interpretação além de ser simples de entender, porque a árvore pode ser visualizada como um conjunto de regras como se fosse um fluxograma de decisões a serem tomadas.

Outro ponto é que as árvores de decisão podem lidar com variáveis categóricas e contínuas, e não requerem normalização dos dados. Embora a árvore de decisão nos ofereça vantagens, há também desvantagens. A primeira desvantagem é que se a árvore for muito profunda, ela pode se ajustar demais aos dados de treinamento, capturando até o "ruído"nos dados, o que prejudica a capacidade de generalização para novos dados. Segundo, a sua instabilidade, como pequenas mudanças nos dados poderá gerar árvores muito diferentes, pois o processo de divisão é sensível aos dados. Outro ponto, é o viés nas divisões, onde as árvores de decisão podem ser tendenciosas em favor de características ou atributos com mais níveis ou valores únicos, o que pode fazer com que algumas variáveis dominem e prejudique nosso julgamento acerca do que está sendo analisado.

Além de tudo, a árvore de decisão pode ser aprimorada, utilizando *Pruning* ou "Poda"para limitar a profundidade da árvore ou cortar ramos que não contribuem significativamente para a predição. Outro aprimoramento, se dá pelo *Ensemble* como por exemplo, *Random Forest* e combinar árvores de decisão para formar um modelo mais robusto e generalizável, o que nos auxilia e evita haver *Overfitting* ou sobreajuste dos dados.

Classes

A seguir as três classes representadas na árvore de decisão da figura13.

- **Classe 1**: Representa nível baixo de qualidade do ar (alta poluição). Isso ocorre quando *Solar.R* está em valores intermediários (147 a 224) ou quando *Solar.R* é bem alto ($\geq$ 224)

combinado com um alto nível de Ozone ($\geq$ 74). Essa combinação pode estar associada a condições de intensa radiação solar e altos níveis de ozônio troposférico, fatores que podem indicar episódios de maior poluição atmosférica.

- **Classe 2**: Representa um nível moderado de qualidade do ar. Ela aparece principalmente quando *Solar.R* é alto ($\geq$ 224) e *Ozone* está em um nível mais baixo (<74). Isso pode indicar condições de poluição moderada, onde altos níveis de radiação solar coexistem com níveis mais controlados de ozônio troposférico.

- **Classe 3**: Representa boa qualidade do ar ou baixa poluição. É quando *Solar.R* $<$ 147, independentemente dos valores de Ozone. Baixa radiação solar geralmente está associada a uma menor formação de ozônio troposférico, o que pode reduzir a poluição atmosférica.

Classificação

Na figura13 podemos visualizar uma árvore de decisão e nela contém algumas informações acerca do *clustering*. As variáveis *Solar.R* e *Ozone* são usadas para dividir e classificar os dados nas classes 1, 2 e 3 e como elas influenciam na classificação nessas três categorias. Cada divisão feita pela árvore refinará nossa previsão, ajudando-nos a entender qual classe um conjunto de valores provavelmente pertence. Primeiramente, a condição *Solar.R* $\geq$ 147 faz uma divisão inicial importante logo na primeira divisão. Quando os valores de *Solar.R* é menor que 147, temos uma forte tendência para a Classe 3. Isso sugere que níveis baixos de *Solar.R* estão mais associados a esta categoria. Quando *Solar.R* é igual ou maior que 147, os dados se dividem entre as Classes 1 e 2, indicando que níveis mais altos de *Solar.R* têm mais probabilidade de pertencer a essas duas classes.

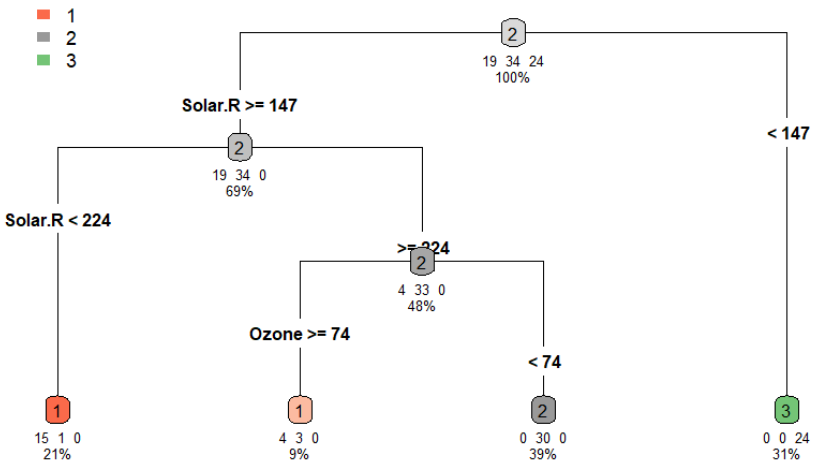

Figura 13: Árvore de decisão

Há outra ramificação para *Solar.R* em que os valores entre 147 e 224 e *Solar.R* $\geq$ 224, refinando ainda mais as classificações. Isso quer dizer que, quando *Solar.R* está entre 147 e 224, a maioria das instâncias cai na Classe 1. Esses valores intermediários de *Solar.R* são característicos da Classe 1, sugerindo que, nesse intervalo, *Solar.R* está associado a uma condição ou característica

típica dessa classe. Neste nó, a classificação é relativamente confiável, pois temos uma maioria de instâncias na Classe 1 (15 contra 1 na Classe 2). Isso representa um bom ponto de decisão para classificar dados como Classe 1.

Para $Solar.R \geq 224$ a árvore se aprofunda usando *Ozone* como a próxima variável para separar as classes. Valores altos de $Solar.R$ (acima de 224) podem indicar situações complexas que requerem uma análise adicional, feita com a variável *Ozone*.

Para $Ozone \geq 74$ quando temos tanto a variável $Solar.R$ alto (≥ 224) quanto *Ozone* alto (≥ 74), a maioria das instâncias pertence à Classe 1. Altos valores de $Solar.R$ e *Ozone* estão correlacionados com a Classe 1. Isso pode indicar que a combinação desses fatores é característica das condições representadas por essa classe.

Para $Ozone < 74$ e quando $Solar.R$ é alto (≥ 224), mas *Ozone* é baixo (<74), a maioria dos dados pertence à Classe 2. Essa combinação de alto $Solar.R$ e baixo *Ozone* está fortemente associada à Classe 2. Isso ajuda a diferenciar os dados que, de outra forma, poderiam ser classificados incorretamente.

Para $Solar.R < 147$ a árvore classifica automaticamente como Classe 3. Essa divisão é bem confiável, pois todos os dados (24 instâncias) caem nesta classe. Isso significa que temos uma classificação exata para essa condição.

Melhores divisões

Ainda de acordo com a figura13 as divisões com maior precisão ocorrem onde há uma maioria de uma classe, como em $Solar.R < 147$ (Classe 3) e $Solar.R$ entre 147 e 224 (Classe 1). Esses nós têm classificações confiáveis, pois os dados estão bem separados.

Divisões menos precisas

Alguns nós, como $Solar.R \geq 224$ e Ozone ≥ 74, são um pouco menos, pois a distribuição é mais dividida entre as classes (4 instâncias de Classe 1 e 3 de Classe 2). Essas divisões podem gerar mais incertezas e não são precisas.

Conclusão

Como vimos, ao longo de todo o estudo acerca da poluição e qualidade do ar, concluímos que a radiação solar é um fator importante porque, como falado anteriomente, a árvore de decisão, demonstra que a radiação solar *Solar.R* é o principal fator determinante inicial para categorizar a qualidade do ar. Isso se alinha ao conhecimento científico de que a radiação solar influencia diretamente a formação de ozônio troposférico e outros poluentes fotoquímicos. Aumentos na radiação tendem a estar associados a condições de poluição mais intensas, especialmente em áreas urbanas com presença de precursores de poluentes.

Outro ponto, a interação entre radiação solar e ozônio troposférico, a variável *Ozone* só é relevante quando *Solar.R* atinge valores muito altos ($\geq$224), o que indica que o ozônio é um fator secundário na classificação, mas importante para identificar situações de poluição grave. A combinação de alta radiação solar e alto ozônio troposférico, aponta para condições ambientais típicas de fumaça, onde a qualidade do ar é severamente afetada.

Uma das implicações práticas e de saúde pública é que as classes analisadas, através das classificações e agrupamentos, podem servir como indicativos práticos para alertas de qualidade do ar. Por exemplo, em dias com radiação solar muito alta e alto nível de ozônio troposférico, alertas de alta poluição podem ser emitidos para populações vulneráveis, como crianças, idosos e pessoas com doenças respiratórias. Quando a radiação solar é baixa, a qualidade do ar tende a ser melhor, o que pode significar que as atividades ao ar livre são mais seguras para a população em geral.

Por fim, há tambem o potencial para políticas ambientais, sendo que essas informações podem ser úteis para formuladores de políticas ambientais, permitindo a implementação de medidas preventivas em dias com alta previsão de radiação solar e emissão de precursores do ozônio troposférico, como limitar o tráfego veicular em horários de pico ou incentivar o uso de transporte público para reduzir a emissão de poluentes.

Referências

[0]Notas de aula.

[1] Johnson, R.A. & Wichern, D.W.. *Applied Multivariate Analysis,* Quarta Edição, Prentice-Hall, Nova Jersey, 1998. * Roteiro do curso.

[1.A] —— , Sexta Edição, Pearson Education Limited, Nova Jersey, 2007.

[2] Artes, R.& Barroso, L.. *Métodos Multivariados de Análise Estatística Estatística.* Blucher, São Paulo, 2023.

[3] Everitt, B..*Cluster Analysis.* Quinta Edição, Wiley & Sons, Nova Iorque, 2011.

[4] Koch, I.. *Analysis of Multivariate and High-Dimensional Data: Theory and Practice.* Cambridge University Press, Nova Iorque, 2014.

[5] Mardia, K.V., Kent, J.T.& Bibby, J.M.. *Multivariate Analysis.* Sétima Reimpressão, Academic Press, Londres, 2000.

[6] Volpato, G.L.. *Guia Prático para Redação Científica.* Best Writing, Botucatu, 2015.