



Análise Exploratória de Dados

Um guia completo para cientistas de dados e analistas realizarem análises exploratórias eficazes

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br
<https://eic.cefet-rj.br/~eogasawara>

Objetivos da Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados é uma etapa crítica que precede a modelagem estatística e o aprendizado de máquina. Ela fornece insights fundamentais sobre a natureza dos seus dados antes de aplicar algoritmos complexos.



Estrutura e Qualidade

Compreender a organização dos dados e identificar problemas de qualidade como valores ausentes, duplicatas e inconsistências



Padrões e Anomalias

Detectar tendências, identificar outliers e descobrir anomalias que podem impactar a análise subsequente



Relações entre Variáveis

Revelar correlações, dependências e interações entre diferentes atributos do conjunto de dados



Decisões Estratégicas

Apoiar escolhas sobre pré-processamento, normalização, seleção de features e escolha de modelos apropriados

- ❑ A EDA ajuda você a dar sentido aos seus dados antes que a modelagem comece. É o alicerce de qualquer projeto de ciência de dados bem-sucedido.

Tipos de Conjuntos de Dados

Diferentes tipos de dados requerem abordagens distintas de análise. Compreender a natureza do seu conjunto de dados é fundamental para escolher as técnicas apropriadas de exploração e visualização.



Dados Relacionais

Conjuntos de dados estruturados em tabelas com registros e atributos



Dados Matriciais

Matrizes numéricas, tabelas de contingência e dados tabulares cruzados



Documentos

Textos, vetores de frequência de termos e dados não estruturados



Transações

Registros de compras, vendas e eventos comerciais sequenciais



Grafos e Redes

Estruturas de relacionamento como redes sociais e a World Wide Web



Dados Ordenados

Séries temporais e sequências de transações com ordem significativa



Dados Espaciais e Multimídia

Mapas geográficos, imagens, vídeos e dados com dimensão espacial

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|--------------|-------------|--------------|-------------|---------|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |

| Documents | team | coach | play | ball | score | game | win | lost | timeout | season |
|------------|------|-------|------|------|-------|------|-----|------|---------|--------|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

| TID | Items |
|-----|---------------------------|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

| Month | GDP | |
|-------|---------|-----|
| <chr> | <dbl> | |
| 1 | 1990.01 | 0.2 |
| 2 | 1990.02 | 0.4 |
| 3 | 1990.03 | 0.8 |
| 4 | 1990.04 | 0.7 |
| 5 | 1990.05 | 0.8 |
| 6 | 1990.06 | 0.8 |

Características Importantes de Dados Estruturados

Ao trabalhar com dados estruturados, é essencial compreender suas características fundamentais que influenciam diretamente a escolha de algoritmos e estratégias de análise.



Dimensionalidade

O número de atributos ou features em um conjunto de dados. Alta dimensionalidade pode levar à "maldição da dimensionalidade", onde a esparsidade dos dados aumenta exponencialmente.



Esparsidade

A proporção de valores ausentes ou zeros em relação ao total de valores possíveis. Em muitos casos, apenas a presença de valores é relevante para a análise.



Resolução

A granularidade ou nível de detalhe dos dados. Padrões identificados dependem fortemente da escala de observação e do nível de agregação aplicado.

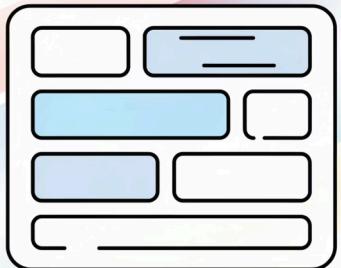


Distribuição

Como os valores estão distribuídos no conjunto de dados, incluindo medidas de tendência central (média, mediana) e dispersão (variância, desvio padrão).

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Dados Relacionais



Dados relacionais são o tipo mais comum de dados estruturados, organizados em tabelas com linhas e colunas. Cada linha representa um objeto de dados distinto, enquanto cada coluna representa um atributo específico.

Componentes Fundamentais

- **Objetos de Dados:** Representam entidades do mundo real como clientes em um banco de vendas, pacientes em um sistema médico, ou estudantes em um banco universitário
- **Atributos:** Características que descrevem os objetos de dados, fornecendo informações específicas sobre cada entidade
- **Estrutura Tabular:** Linhas contêm tuplas (registros individuais) e colunas contêm atributos (características)

Exemplos de Aplicação

- Banco de vendas: clientes, itens de loja, transações
- Banco médico: pacientes, tratamentos, doenças
- Banco universitário: estudantes, professores, cursos

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Atributos: A Base da Análise de Dados

Atributos, também conhecidos como dimensões, features ou variáveis, são campos de dados que representam uma característica ou recurso de um objeto de dados. A compreensão dos tipos de atributos é fundamental para aplicar técnicas analíticas apropriadas.

Nominal

Categorias ou rótulos sem ordem inerente. Exemplos: ID do cliente, nome, endereço, cor dos olhos

Binário

Apenas dois estados possíveis (0 e 1). Pode ser simétrico ou assimétrico dependendo da importância dos resultados

Ordinal

Valores com ordem significativa, mas sem magnitude conhecida entre valores sucessivos. Exemplos: tamanho, notas

Numérico

Quantidades representadas por valores inteiros ou reais, permitindo operações matemáticas

- ❑ ❤️ Exemplos práticos: customer_ID (nominal), tem_diabetes (binário), nível_educação (ordinal), idade (numérico)

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Tipos de Atributos em Detalhe

Compreender as nuances entre diferentes tipos de atributos é crucial para escolher métricas de distância, transformações e técnicas analíticas apropriadas.

Atributos Nominais

Representam categorias, estados ou "nomes de coisas" sem qualquer ordem implícita.

- Hair_color = {castanho, preto, loiro, marrom, cinza, vermelho, branco}
- Estado civil, ocupação, números de ID, códigos postais
- Operações válidas: igualdade/desigualdade

Atributos Binários

Possuem apenas dois estados possíveis (0 e 1). A interpretação depende da simetria dos resultados.

- **Binário Simétrico:** Ambos os resultados são igualmente importantes (ex: gênero)
- **Binário Assimétrico:** Um resultado é mais importante (ex: teste médico positivo vs. negativo)
- Convenção: atribuir 1 ao resultado mais importante (ex: HIV positivo)

Atributos Ordinais

Valores possuem uma ordem significativa (ranking), mas a magnitude entre valores sucessivos não é conhecida.

- Size = {pequeno, médio, grande}
- Notas escolares (A, B, C, D, F)
- Ranking militar (soldado, cabo, sargento, tenente)
- Operações válidas: comparação de ordem ($>$, $<$, $=$)

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Tipos de Atributos Numéricos

Atributos numéricos representam quantidades e podem ser classificados em dois subtipos principais com propriedades matemáticas distintas que afetam quais operações e transformações são válidas.

Atributos de Intervalo

- Medidos em uma escala de unidades de tamanho igual
- Valores possuem ordem significativa
- **Não possuem ponto zero verdadeiro**
- Exemplos: temperatura em °C ou °F, datas do calendário
- Operações válidas: adição, subtração, mas não multiplicação ou divisão

Não é possível dizer que 20°C é "duas vezes mais quente" que 10°C, pois o zero é arbitrário.

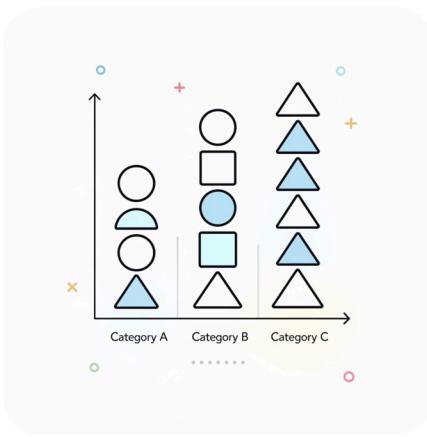
Atributos de Razão

- Possuem ponto zero inerente e absoluto
- Podemos falar de valores como sendo uma ordem de magnitude maior que a unidade de medida
- Exemplos: temperatura em Kelvin, comprimento, contagens, quantias monetárias
- Operações válidas: todas as operações aritméticas

É correto afirmar que 10K é duas vezes mais alto que 5K, pois o zero é absoluto.

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Atributos Discretos versus Contínuos



Atributos Discretos

Possuem apenas um conjunto finito ou contavelmente infinito de valores possíveis. Frequentemente representados como variáveis inteiras.

- Número de filhos em uma família
- Quantidade de produtos vendidos
- Contagem de palavras em um documento
- Categorias codificadas numericamente



Atributos Contínuos

Possuem números reais como valores de atributo, podendo assumir qualquer valor dentro de um intervalo.

- Temperatura, altura, peso corporal
- Pressão arterial, concentração química
- Tempo de resposta, distância percorrida

Na prática, valores reais só podem ser medidos e representados usando um número finito de dígitos. Atributos contínuos são tipicamente representados como variáveis de ponto flutuante.

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Por Que os Tipos de Atributos Importam

A classificação correta dos tipos de atributos é fundamental para o sucesso de qualquer análise de dados. O tipo de atributo determina quais operações, transformações e técnicas estatísticas são apropriadas e válidas.



Medidas de Distância

O tipo de atributo determina qual métrica de distância usar: Euclidiana para numéricos, Hamming para categóricos, Jaccard para binários



Exemplo: Cor do Cabelo

Não faz sentido calcular a média de cores de cabelo (atributo nominal). Isso levaria a resultados sem significado.

Sumários Estatísticos

Diferentes tipos permitem diferentes estatísticas: média para numéricos, moda para nominais, mediana para ordinais



Exemplo: Peso

Você pode normalizar peso (numérico contínuo) usando z-score ou min-max scaling para melhorar a performance do modelo.

Codificação e Transformação

Determina qual pré-processamento aplicar: normalização para contínuos, one-hot encoding para nominais



Exemplo: CEP

Apesar de ser numérico, CEP é nominal. Normalizar ou calcular média de CEPs não tem significado prático.

- Compreender os tipos de atributos ajuda a prevenir o uso indevido de técnicas e garante análises mais robustas e significativas.

O Dataset Iris: Um Exemplo Clássico

O dataset Iris é um exemplo clássico em ciência de dados, amplamente utilizado para tarefas de classificação e aprendizado de máquina. Introduzido por Ronald Fisher em 1936, permanece relevante como benchmark para algoritmos de classificação.

Estrutura do Dataset

Contém 150 observações de flores Iris de três espécies diferentes:

- **Iris Setosa** - 50 amostras
- **Iris Versicolor** - 50 amostras
- **Iris Virginica** - 50 amostras

Atributos Medidos

- Comprimento da sépala (cm) - atributo numérico contínuo
- Largura da sépala (cm) - atributo numérico contínuo
- Comprimento da pétala (cm) - atributo numérico contínuo
- Largura da pétala (cm) - atributo numérico contínuo
- Espécie - atributo categórico nominal (variável alvo)

O dataset Iris é ideal para aprender EDA porque possui atributos numéricos bem comportados, classes balanceadas e padrões de separação claros entre algumas espécies, tornando-o perfeito para demonstrar técnicas de visualização e classificação.

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|--------------|-------------|--------------|-------------|----------------|
| numeric | numeric | numeric | numeric | factor |
| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 setosa |
| 51 | 7.0 | 3.2 | 4.7 | 1.4 versicolor |
| 52 | 6.4 | 3.2 | 4.5 | 1.5 versicolor |
| 53 | 6.9 | 3.1 | 4.9 | 1.5 versicolor |
| 101 | 6.3 | 3.3 | 6.0 | 2.5 virginica |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 virginica |
| 103 | 7.1 | 3.0 | 5.9 | 2.1 virginica |

Visualização das medidas das características das flores Iris mostrando a separação entre as três espécies.

Descrições Estatísticas Básicas de Dados

A descrição estatística básica é o primeiro passo para compreender profundamente seus dados. Essas medidas fornecem insights essenciais sobre a estrutura, tendências e variabilidade presentes no conjunto de dados.

Motivação

Para entender melhor os dados através de:

- **Tendência Central:** Onde os dados estão concentrados (média, mediana, moda)
- **Variação e Dispersão:** Como os dados se espalham (variância, desvio padrão, amplitude)
- **Forma da Distribuição:** Simetria, assimetria e curtose

Características Principais

- Mediana, máximo, mínimo - limites e centro dos dados
- Quantis (quartis, decis, percentis) - divisão da distribuição
- Outliers - valores atípicos que podem indicar erros ou insights
- Variância e desvio padrão - medidas de dispersão

Análise de Intervalos

Dimensões numéricas correspondem a intervalos ordenados que podem ser analisados através de boxplots



Identificação de Outliers

Detecção de valores extremos que requerem atenção especial na análise

Análise de Quantis

Divisão dos dados em partes iguais para entender a distribuição de valores

Medidas Descritivas Fundamentais

As medidas descritivas resumem as características mais importantes de um conjunto de dados, permitindo uma compreensão rápida e eficaz da distribuição dos valores.

1

Medidas de Tendência Central

- **Média (μ):** Soma de todos os valores dividida pelo número de observações. Sensível a outliers.
- **Mediana:** Valor central quando os dados estão ordenados. Robusta a outliers.
- **Moda:** Valor mais frequente no conjunto de dados. Útil para dados categóricos.

2

Medidas de Dispersão

- **Variância (σ^2):** Média dos quadrados dos desvios em relação à média.
- **Desvio Padrão (σ):** Raiz quadrada da variância, na mesma unidade dos dados originais.
- **Amplitude:** Diferença entre o valor máximo e mínimo.
- **Intervalo Interquartil (IQR):** Diferença entre o terceiro e primeiro quartis ($Q3 - Q1$).

3

Medidas de Posição

- **Quartis:** Dividem os dados em quatro partes iguais ($Q1, Q2=\text{mediana}, Q3$)
- **Percentis:** Dividem os dados em 100 partes iguais
- **Decis:** Dividem os dados em 10 partes iguais

4

Medidas de Forma

- **Assimetria (Skewness):** Indica o grau de assimetria da distribuição
- **Curtose:** Mede o achatamento ou pico da distribuição
- Distribuições simétricas têm assimetria próxima de zero

Medindo a Dispersão dos Dados: Quartis e Boxplots

Quartis e boxplots são ferramentas poderosas para visualizar e compreender a dispersão e distribuição dos dados, especialmente úteis para identificar outliers e comparar múltiplas distribuições.

Resumo de Cinco Números

Uma forma concisa de descrever a distribuição:

- Mínimo - menor valor observado
- Q1 (1º Quartil) - 25º percentil
- Mediana (Q2) - 50º percentil
- Q3 (3º Quartil) - 75º percentil
- Máximo - maior valor observado

Intervalo Interquartil (IQR)

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Representa a amplitude dos 50% centrais dos dados, sendo robusto a outliers.

Exemplo com Dataset Iris

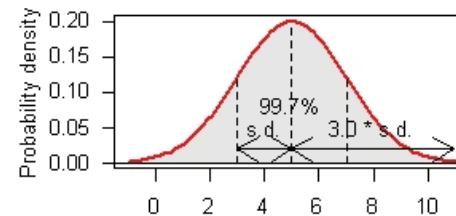
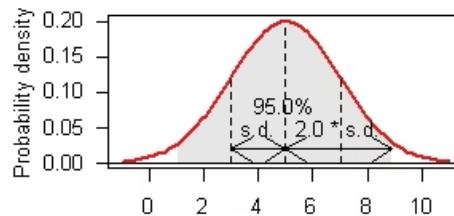
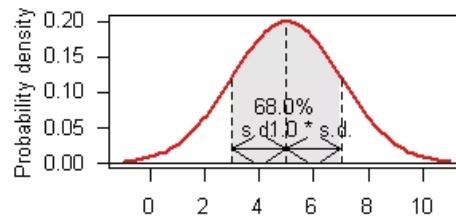
```
sum <- summary(iris$Sepal.Length)
sum
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
## 4.300 5.100 5.800 5.843 6.400 7.900

IQR <- sum["3rd Qu."] - sum["1st Qu."]
IQR
## 1.3
```

Para o comprimento da sépala no dataset Iris:

- Valor mínimo: 4.3 cm
- Q1: 5.1 cm (25% dos dados estão abaixo)
- Mediana: 5.8 cm (valor central)
- Média: 5.843 cm
- Q3: 6.4 cm (75% dos dados estão abaixo)
- Valor máximo: 7.9 cm
- IQR: 1.3 cm (dispersão dos 50% centrais)

Propriedades da Curva de Distribuição Normal



A distribuição normal, também conhecida como distribuição Gaussiana ou curva em forma de sino, é uma das distribuições de probabilidade mais importantes em estatística. Muitos fenômenos naturais e processos seguem aproximadamente esta distribuição.

Simetria Perfeita

A curva é perfeitamente simétrica em torno da média. Média = Mediana = Moda, todas localizadas no centro da distribuição.

Forma de Sino

A distribuição tem uma forma característica de sino, com o pico no centro (média) e diminuindo gradualmente nas caudas.

Regra Empírica (68-95-99.7)

Aproximadamente 68% dos dados estão dentro de 1 desvio padrão da média, 95% dentro de 2 desvios, e 99.7% dentro de 3 desvios.

Características Matemáticas

- Definida por dois parâmetros: média (μ) e desvio padrão (σ)
- A área total sob a curva é igual a 1 (100%)
- As caudas se estendem ao infinito em ambas as direções
- Aproximadamente 95% dos dados estão dentro de $\mu \pm 2\sigma$

Importância na Análise de Dados

- Muitos testes estatísticos assumem normalidade dos dados
- Teorema do Limite Central: médias de amostras tendem à normalidade
- Facilita identificação de outliers (valores além de 3σ)
- Permite cálculos probabilísticos precisos

💡 Testar a normalidade dos seus dados é um passo importante antes de aplicar muitos métodos estatísticos paramétricos. Use testes como Shapiro-Wilk ou visualizações como Q-Q plots.



Exibições Gráficas de Descrições Estatísticas Básicas

Uma exploração visual dos métodos fundamentais para análise exploratória de dados

Exibições Gráficas de Descrições Estatísticas Básicas

Os gráficos fornecem resumos visuais essenciais da distribuição de dados, revelando padrões que podem não ser aparentes em tabelas numéricas. Estas ferramentas visuais são fundamentais na análise exploratória de dados (EDA), permitindo que analistas compreendam rapidamente a forma, dispersão e presença de valores atípicos nos conjuntos de dados.



01

Histograma

Mostra a distribuição de frequência dos dados através de barras verticais

02

Boxplot

Representa visualmente os quartis e identifica valores atípicos

03

Densidade

Apresenta uma versão suavizada da distribuição de dados contínuos

 **Referência:** R.J. Larsen e M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Análise de Histograma

O histograma é uma representação gráfica que exibe valores de frequências tabuladas, mostrando a proporção de casos que se enquadram em cada categoria. Esta visualização é fundamental para compreender a distribuição dos dados.

Área das Barras

A área da barra denota o valor e é uma propriedade crucial quando as categorias não possuem largura uniforme. Esta característica garante uma representação proporcional precisa dos dados.

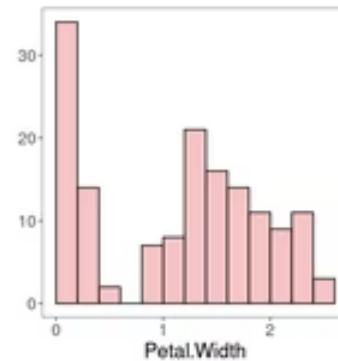
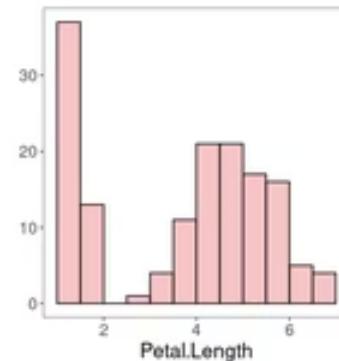
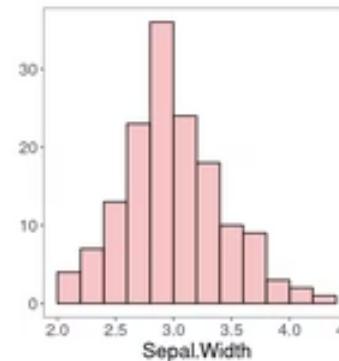
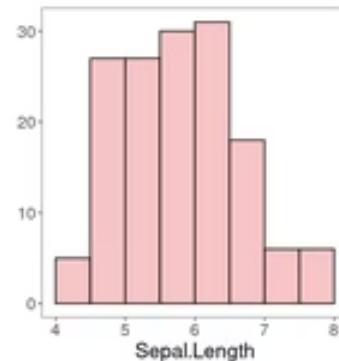
Categorias Não-Sobrepostas

As categorias especificam intervalos não-sobrepostos de alguma variável, garantindo que cada observação seja contada exatamente uma vez na distribuição.

Barras Adjacentes

As categorias (barras) devem ser adjacentes, criando uma visualização contínua que facilita a identificação de padrões na distribuição dos dados.

Exemplo: Comprimento da Sépala



Fonte: R.J. Larsen e M.L. Marx, 2017, An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

Dados Simétricos vs. Assimétricos

A relação entre média e mediana revela informações importantes sobre a forma da distribuição dos dados. Compreender a assimetria é essencial para escolher as medidas estatísticas apropriadas e interpretar corretamente os resultados.

1

Assimetria Positiva

Média > Mediana. A cauda se estende para valores maiores, indicando concentração de dados nos valores menores.

2

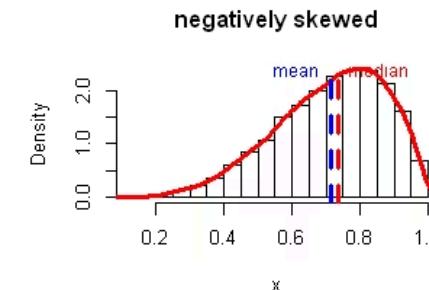
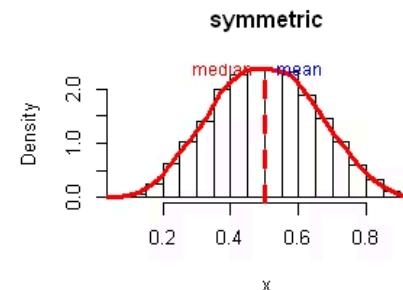
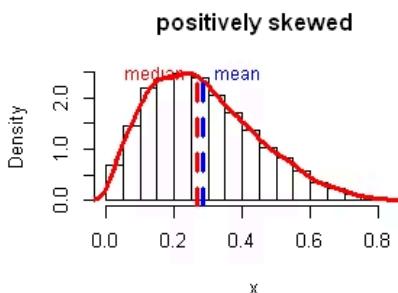
Distribuição Simétrica

Média \approx Mediana. Os dados se distribuem de forma equilibrada em torno do centro.

3

Assimetria Negativa

Média < Mediana. A cauda se estende para valores menores, com concentração nos valores maiores.



□ **Referência:** R.J. Larsen e M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Densidade de Probabilidade

O que é Estimativa de Densidade Kernel?

A estimativa de densidade kernel calcula e desenha uma versão suavizada do histograma. Esta técnica é uma alternativa útil ao histograma para dados contínuos que provêm de uma distribuição subjacente suave.



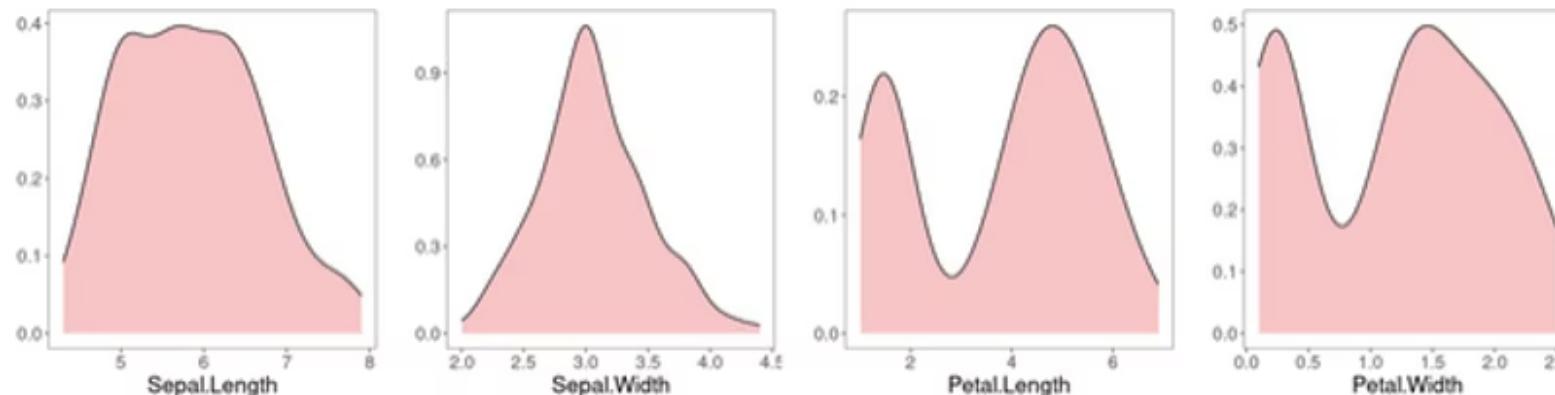
Suavização

Elimina a rugosidade do histograma, revelando a forma verdadeira da distribuição



Continuidade

Representa dados contínuos de forma mais natural que barras discretas



Análise de Boxplot

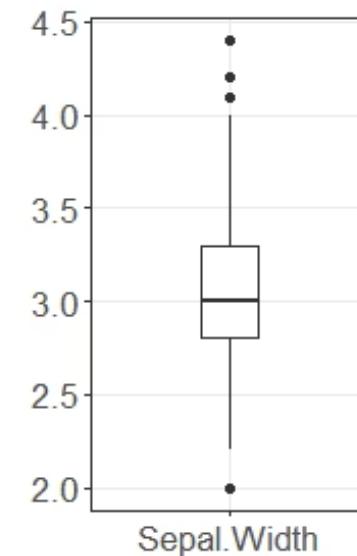
Em estatística descritiva, um boxplot é um método para representar graficamente grupos de dados numéricos através de seus quartis. Os boxplots também podem ter linhas estendendo-se das caixas (bigodes), indicando variabilidade fora dos quartis superior e inferior.

Resumo de Cinco Números

| | |
|------------------------------|---|
| Mínimo | Menor valor observado no conjunto de dados |
| Primeiro Quartil (Q1) | 25% dos dados estão abaixo deste valor |
| Mediana (Q2) | Valor central que divide os dados em duas metades |
| Terceiro Quartil (Q3) | 75% dos dados estão abaixo deste valor |
| Máximo | Maior valor observado no conjunto de dados |

Componentes do Boxplot

| | |
|----------------------|--|
| Caixa | Representa o intervalo interquartil (IQR), com extremidades em Q1 e Q3. A altura da caixa é o IQR. |
| Linha Mediana | Linha dentro da caixa que marca a mediana dos dados. |
| Bigodes | Duas linhas fora da caixa estendidas ao Mínimo e Máximo (excluindo outliers). |
| Outliers | Valores maiores que $Q3 + 1.5 \times \text{IQR}$ ou menores que $Q1 - 1.5 \times \text{IQR}$. |



Fonte: R. McGill, J.W. Tukey, e W.A. Larsen, 1978, Variations of box plots, American Statistician, v. 32, n. 1, p. 12–16.

Outliers em Boxplot

Identificação de Valores Atípicos

Os outliers são pontos de dados que diferem significativamente do resto da distribuição. No boxplot, eles são identificados usando critérios estatísticos baseados no intervalo interquartil (IQR).

1

Critério Superior

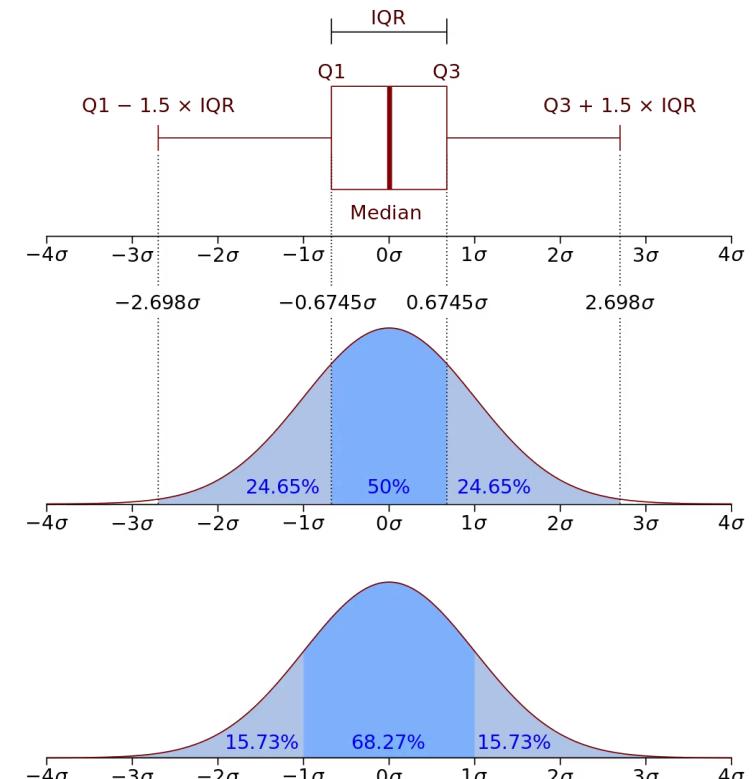
Valores maiores que $Q3 + 1.5 \times IQR$ são considerados outliers superiores

2

Critério Inferior

Valores menores que $Q1 - 1.5 \times IQR$ são considerados outliers inferiores

- Os outliers podem representar erros de medição, variabilidade natural extrema, ou fenômenos genuinamente raros que merecem investigação especial.



Problemas de Qualidade de Dados Revelados pela EDA

A análise exploratória de dados não apenas revela padrões nos dados, mas também expõe problemas de qualidade que podem comprometer análises posteriores. Identificar esses problemas precocemente é essencial para garantir resultados confiáveis.



Valores Ausentes

Deteta lacunas e registros incompletos nos dados, avaliando a completude do conjunto de dados. A presença de muitos valores ausentes pode indicar problemas na coleta de dados ou exigir técnicas de imputação.



Ruído

Identifica erros aleatórios e variabilidade que afetam a precisão das medições. O ruído pode obscurecer padrões reais e reduzir a confiabilidade das análises estatísticas.



Inconsistências

Revela valores conflitantes ou contraditórios entre atributos diferentes. Inconsistências podem surgir de erros de entrada, integração de múltiplas fontes ou mudanças nos padrões de coleta ao longo do tempo.

A EDA ajuda a avaliar a qualidade dos dados antes das etapas de limpeza e integração, garantindo que as análises subsequentes sejam baseadas em dados confiáveis.

Exemplo de Boxplot para o Dataset Iris

Aplicação Prática

Os boxplots são especialmente úteis para comparar distribuições entre grupos, identificar outliers e detectar assimetria nos dados. Eles são amplamente utilizados tanto na EDA quanto em relatórios finais.



Comparação entre Grupos

Visualize diferenças nas distribuições de várias categorias simultaneamente



Detecção de Outliers

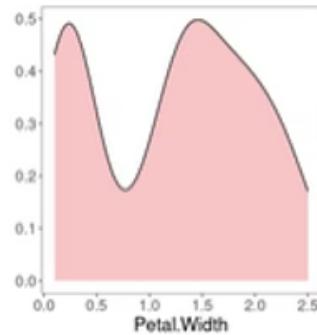
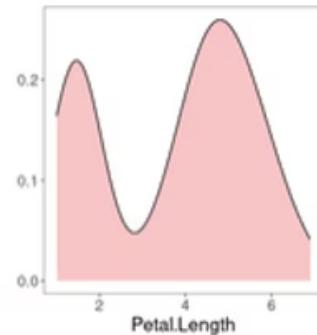
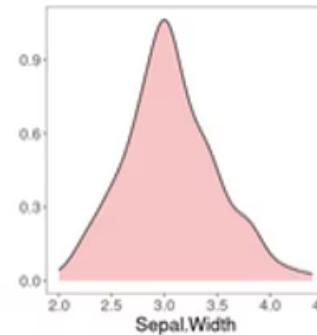
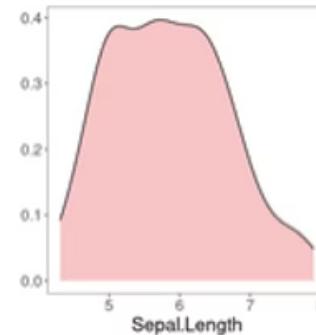
Identifique rapidamente valores atípicos que podem exigir atenção especial



Análise de Assimetria

Reconheça padrões de distribuição e desvios da normalidade

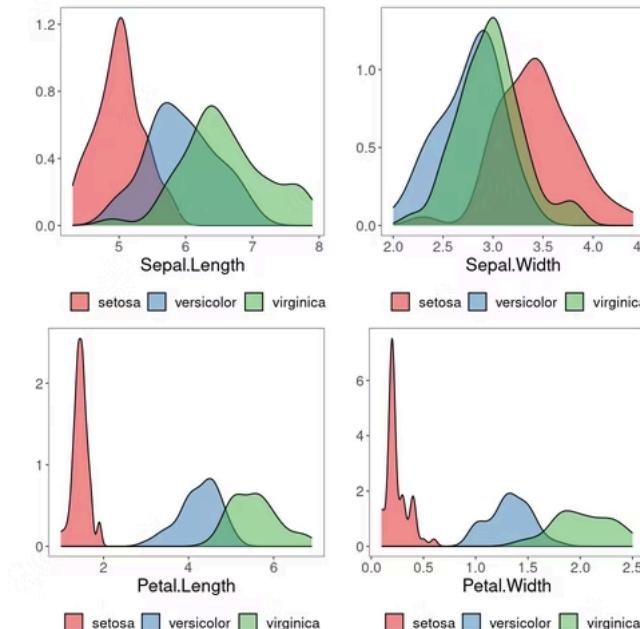
Visualização: Comparação de Espécies de Iris



Fonte: K.J. Keen, 2018, Graphics for Statistics and Data Analysis with R. CRC Press.

Distribuições de Densidade com Rótulo de Classe

Os gráficos de densidade revelam a distribuição de probabilidade de uma variável contínua. Quando combinados com rótulos de classe, eles mostram como diferentes classes podem se sobrepor ou se separar no espaço de características, fornecendo insights valiosos sobre a capacidade discriminativa das variáveis.



```
plot_density_class(  
  iris |>  
  dplyr::select(Species, Sepal.Length),  
  class_label = "Species",  
  label_x = "Sepal.Length",  
  color = colors[c(1:3)]  
)
```

Esta visualização permite identificar rapidamente padrões de separação entre espécies, detectar sobreposições problemáticas e avaliar o potencial preditivo da variável antes mesmo de construir modelos complexos.

Referência: K.J. Keen, 2018, Graphics for Statistics and Data Analysis with R. CRC Press.

Boxplot com Rótulo de Classe

Os boxplots agrupados por classe são ferramentas essenciais na análise exploratória de dados. Eles ilustram simultaneamente várias estatísticas descritivas: mediana, quartis, amplitude interquartil e outliers, permitindo comparações diretas entre grupos.

Distribuição Central

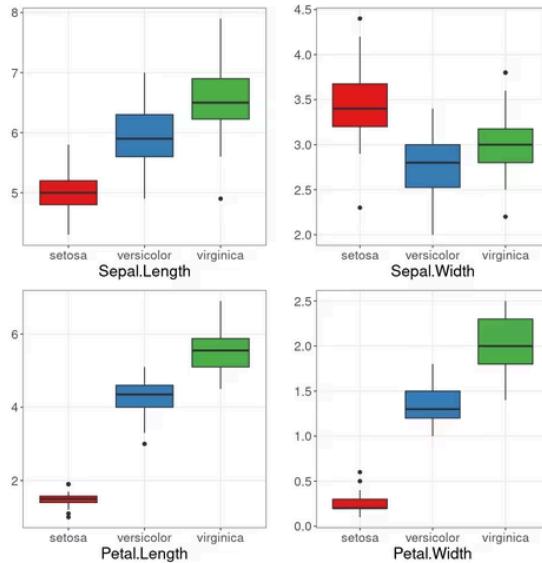
A caixa representa 50% dos dados centrais, revelando onde a maioria das observações se concentra para cada classe.

Variabilidade

A altura da caixa e o comprimento dos whiskers indicam a dispersão dos dados, ajudando a identificar classes com maior ou menor variância.

Outliers e Anomalias

Pontos isolados além dos whiskers sinalizam valores atípicos que podem requerer investigação adicional ou tratamento especial.



```
plot_boxplot_class(  
  iris |>  
  dplyr::select(Species, Sepal.Length),  
  class_label = "Species",  
  label_x = "Sepal.Length",  
  color = colors[c(1:3)]  
)
```

Visualizações Gráficas de Descrições Estatísticas Básicas

A análise de relações entre pares de variáveis é fundamental para compreender a estrutura multivariada dos dados. Três ferramentas visuais complementares nos permitem explorar essas relações de diferentes perspectivas, cada uma revelando aspectos únicos dos padrões subjacentes.



Scatter Plot

Revela clusters, tendências lineares e não-lineares, além de identificar outliers bivariados que podem não ser detectados em análises univariadas.

Análise de Correlação

Quantifica a direção e força das relações lineares entre variáveis, fornecendo métricas numéricas para avaliar dependências.

Matriz de Dispersão

Escala a comparação par a par para dados multivariados, permitindo análise simultânea de múltiplas relações em uma única visualização.

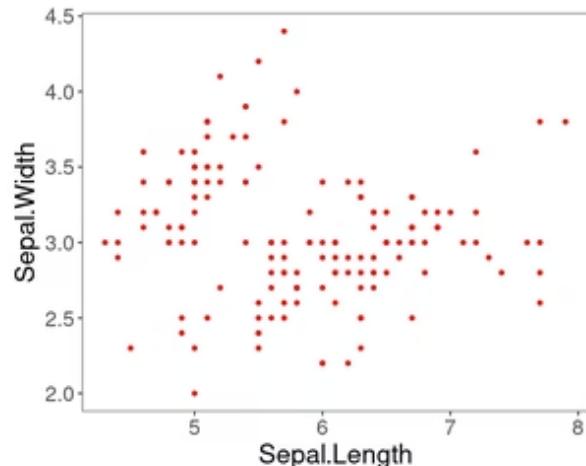
- Dica Prática:** Use essas três técnicas em conjunto para uma compreensão completa das relações entre variáveis. Enquanto scatter plots revelam padrões visuais, correlações fornecem métricas objetivas, e matrizes de dispersão oferecem uma visão panorâmica.

Gráfico de Dispersão (Scatter Plot)

O scatter plot é frequentemente a primeira ferramenta visual na análise de dados bivariados. Cada observação é representada como um ponto no plano cartesiano, onde as coordenadas correspondem aos valores das duas variáveis em análise.

Interpretação Visual

Esta visualização permite identificar rapidamente padrões de agrupamento, detectar valores extremos e avaliar a natureza da relação entre variáveis antes de aplicar métodos estatísticos mais complexos.



```
plot_scatter(  
  iris |>  
  dplyr::select(  
    x = Sepal.Length,  
    value = Sepal.Width  
  ) |>  
  mutate(variable = "iris"),  
  label_x = "Sepal.Length"  
) +  
  theme(legend.position = "none")
```

• Clusters de Pontos

Agrupamentos naturais de observações sugerem subpopulações ou padrões estruturais nos dados.

• Outliers

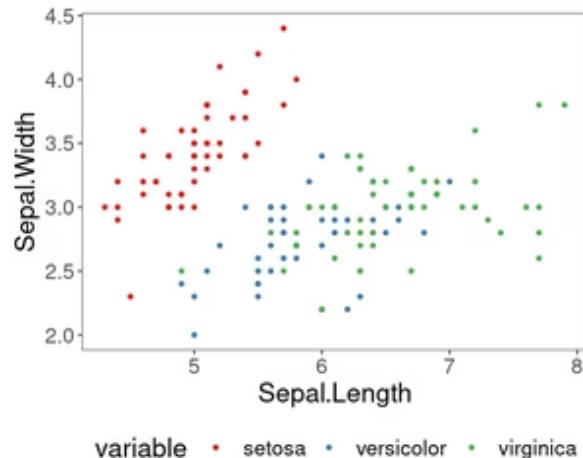
Pontos isolados podem indicar erros de medição, anomalias ou observações excepcionais que requerem atenção especial.

• Tendências

Padrões lineares ou não-lineares revelam possíveis relações funcionais entre as variáveis.

Gráfico de Dispersão com Rótulo de Classe

Adicionar rótulos de classe ao scatter plot transforma uma simples visualização bivariada em uma ferramenta poderosa para análise de classificação. Cada classe recebe uma cor distinta, permitindo avaliar visualmente a separabilidade entre grupos e identificar regiões de sobreposição.



```
plot_scatter(  
  iris |>  
  dplyr::select(  
    x = Sepal.Length,  
    value = Sepal.Width,  
    variable = Species  
>,  
    label_x = "Sepal.Length",  
    label_y = "Sepal.Width",  
    colors = colors[1:3]  
)
```

Avaliação de Separabilidade

Classes bem separadas visualmente indicam que as variáveis escolhidas têm bom poder discriminativo, facilitando a construção de classificadores eficazes.

Identificação de Sobreposição

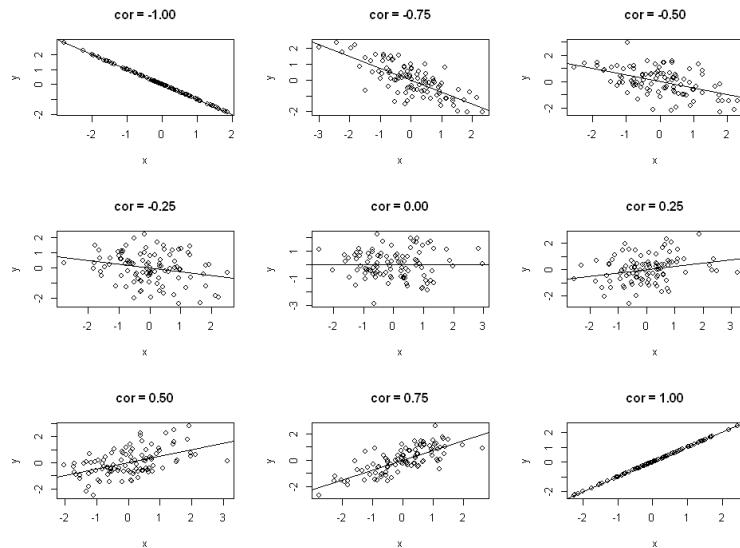
Regiões onde as cores se misturam revelam áreas de ambiguidade onde a classificação será mais desafiadora e provavelmente resultará em erros.

Seleção de Features

Se as classes permanecem misturadas, isso sugere a necessidade de explorar outras combinações de variáveis ou aplicar transformações nos dados.

Correlação de Dados

A compreensão visual dos padrões de correlação é essencial antes de realizar análises quantitativas. Esta visualização demonstra três tipos fundamentais de relações lineares entre variáveis, cada uma com implicações distintas para modelagem e interpretação.



Correlação Negativa (Linha 1)

Quando uma variável aumenta, a outra tende a diminuir. O coeficiente de correlação varia entre -1 e 0, indicando a força desta relação inversa.

Ausência de Correlação (Linha 2)

As variáveis não apresentam relação linear aparente. O coeficiente de correlação está próximo de zero, embora possa existir uma relação não-linear.

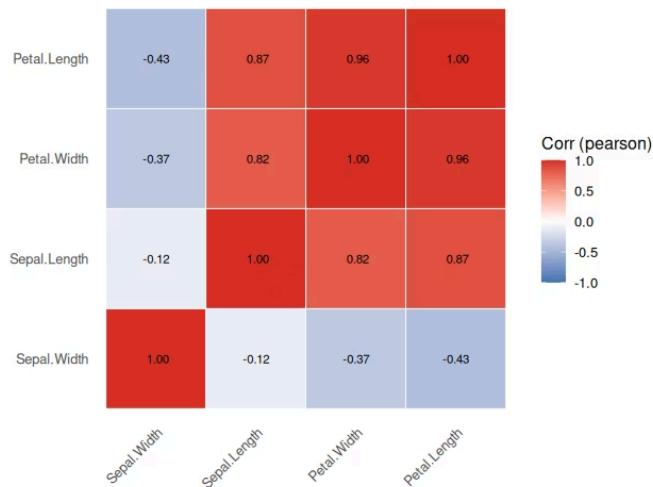
Correlação Positiva (Linha 3)

As variáveis tendem a aumentar juntas. O coeficiente de correlação varia entre 0 e 1, refletindo a força desta relação direta.

Atenção: A correlação mede apenas relações *lineares*. Duas variáveis podem ter forte relação não-linear mas apresentar correlação próxima de zero. Sempre combine análise visual com métricas numéricas.

Análise de Correlação

O correograma é uma representação visual sofisticada das correlações par a par entre todos os atributos numéricos de um dataset. Esta ferramenta permite identificar rapidamente padrões de interdependência, multicolinearidade e estruturas latentes nos dados.



```
grf <- plot_correlation(  
  iris |>  
  dplyr::select(  
    Sepal.Width,  
    Sepal.Length,  
    Petal.Width,  
    Petal.Length  
  )  
)  
grf
```

O uso de cores e formas geométricas facilita a interpretação intuitiva: cores quentes indicam correlações positivas fortes, enquanto cores frias revelam correlações negativas.

01

Detectar Multicolinearidade

Correlações muito altas entre preditores podem causar instabilidade em modelos de regressão.

02

Guia Seleção de Features

Variáveis altamente correlacionadas entre si podem ser redundantes para modelagem.

03

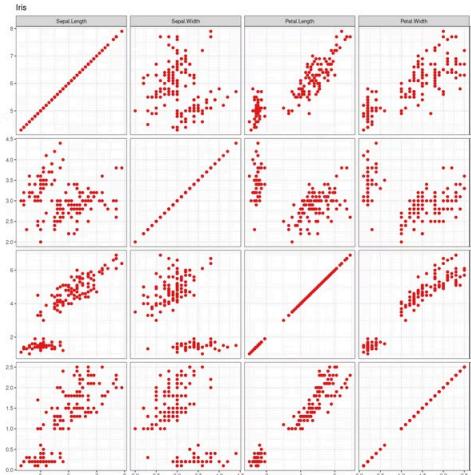
Identificar Estruturas

Grupos de variáveis correlacionadas podem indicar dimensões latentes nos dados.

Referência: M. Friendly, 2002, Corrrgrams: Exploratory displays for correlation matrices, American Statistician, v. 56, n. 4, p. 316–324.

Matriz de Dispersão (Scatter Matrix Plot)

A matriz de dispersão é uma grade organizada de scatterplots para cada par de variáveis em um dataset multivariado. Esta visualização poderosa permite analisar simultaneamente múltiplas relações bivariadas, identificar padrões de agrupamento e detectar tendências entre várias dimensões dos dados.



```
plot_pair(  
  data = iris,  
  cnames = colnames(iris)[1:4],  
  title = "Iris",  
  colors = colors[1]  
)
```

A diagonal da matriz frequentemente exibe histogramas ou gráficos de densidade das variáveis individuais, complementando a análise bivariada com informações univariadas.

1

Exploração Abrangente

Uma única visualização revela todas as relações par a par, economizando tempo e oferecendo uma visão holística dos dados.

2

Detecção de Padrões

Correlações, tendências não-lineares e agrupamentos se tornam evidentes através da repetição visual dos padrões.

3

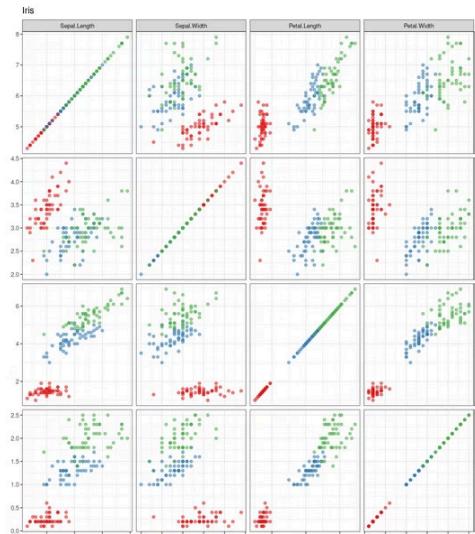
Identificação de Outliers Multivariados

Observações anômalas que aparecem consistentemente em múltiplos painéis requerem investigação especial.

Referência: N. Elmqvist, P. Dragicevic, and J.-D. Fekete, 2008, Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation, IEEE Transactions on Visualization and Computer Graphics, v. 14, n. 6, p. 1141–1148.

Matriz de Dispersão com Rótulo de Classe

Adicionar rótulos de classe à matriz de dispersão transforma-a em uma ferramenta de análise discriminante visual. Cada classe recebe uma cor distinta, permitindo avaliar como as classes se separam através de múltiplas combinações de características simultaneamente.



Implementação com Classes

```
plot_pair(  
  data = iris,  
  cnames = colnames(iris)[1:4],  
  xlabel = 'Species',  
  title = "Iris",  
  colors = colors[1:3]  
)
```

Vantagens para Classificação

Esta visualização revela quais pares de variáveis oferecem melhor separação entre classes, informando a seleção de features e a construção de limites de decisão em algoritmos de classificação.

→ Avaliação Multivariada

Certas classes podem se sobrepor em algumas dimensões mas se separar claramente em outras, revelando a importância de análise multivariada.

→ Seleção de Features Eficaz

Identifique rapidamente quais combinações de variáveis oferecem máxima separabilidade entre classes.

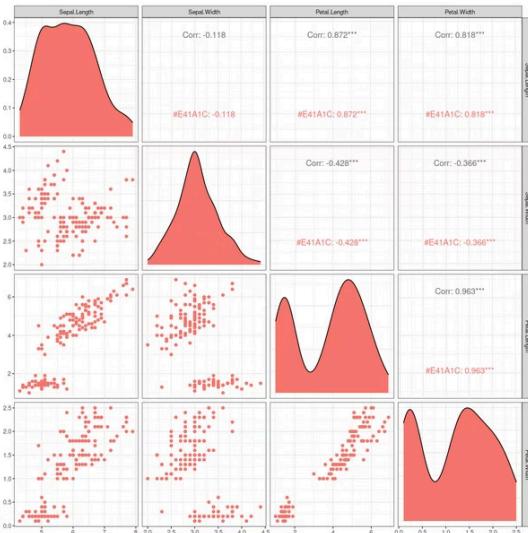
→ Validação Visual de Modelos

Compare a separação visual com o desempenho de classificadores para entender suas limitações e pontos fortes.

Referência: N. Elmqvist, P. Dragicevic, and J.-D. Fekete, 2008, Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation, IEEE Transactions on Visualization and Computer Graphics, v. 14, n. 6, p. 1141–1148.

Matriz de Dispersão Avançada

A versão avançada da matriz de dispersão incorpora sobreposições de densidade, padrões suavizados e outras melhorias visuais que revelam estruturas subjacentes em dados multidimensionais complexos. Essas técnicas são especialmente úteis quando há muitos pontos de dados que podem se sobrepor nos scatter plots tradicionais.



```
plot_pair_adv(  
  data = iris,  
  cnames = colnames(iris)[1:4],  
  title = "Iris",  
  colors = colors[1]  
)
```

Contornos de densidade e gradientes de cor substituem pontos individuais, tornando padrões de concentração mais evidentes mesmo em datasets grandes.

Sobreposições de Densidade

Contornos e mapas de calor revelam áreas de alta concentração de pontos, facilitando a identificação de clusters e distribuições multimodais.

Redução de Overplotting

Em datasets grandes, pontos individuais se sobrepõem. Representações baseadas em densidade resolvem este problema mantendo a clareza visual.

Revelação de Estruturas Complexas

Padrões não-lineares e distribuições não-gaussianas se tornam mais aparentes através de técnicas de suavização e visualização de densidade.

Matriz de Dispersão Avançada com Rótulo de Classe

Combinar técnicas avançadas de visualização com rótulos de classe cria uma ferramenta extremamente poderosa para análise de classificação multivariada. Os contornos de densidade codificados por cor para cada classe revelam sobreposições sutis e fronteiras de decisão naturais entre grupos.



```
grf <- plot_pair_adv(  
  data = iris,  
  cnames = colnames(iris)[1:4],  
  title = "Iris",  
  clabel = 'Species',  
  colors = colors[1:3]  
)  
grf
```

Interpretação Aprimorada

As sobreposições de densidade por classe facilitam a identificação de regiões onde múltiplas classes coexistem, indicando maior dificuldade de classificação e possível necessidade de features adicionais ou modelos mais sofisticados.



Fronteiras de Decisão

Visualize naturalmente onde as distribuições de classe se encontram.



Análise de Regiões Ambíguas

Identifique áreas onde múltiplas classes se sobrepõem significativamente.



Insights para Modelagem

Guie a escolha de algoritmos baseando-se na complexidade visual das fronteiras.

Referência: D.A. Keim, M.C. Hao, U. Dayal, H. Janetzko, and P. Bak, 2010, Generalized scatter plots, Information Visualization, v. 9, n. 4, p. 301–311.

EDA e Pré-processamento de Dados

A Análise Exploratória de Dados (EDA) não é uma atividade isolada – ela informa diretamente as decisões de pré-processamento que determinarão a qualidade dos modelos subsequentes. Um EDA cuidadoso revela problemas nos dados que devem ser corrigidos antes da modelagem.

Tratamento de Valores Ausentes

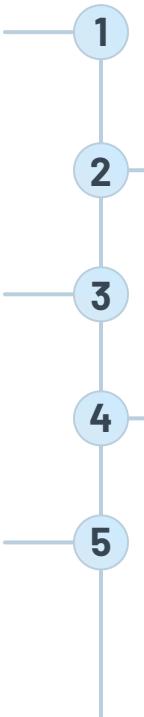
EDA revela padrões de missingness: são aleatórios ou sistemáticos? Isso determina se devemos imputar, deletar ou criar indicadores de ausência.

Escalonamento e Normalização

Diferenças de escala entre variáveis, detectadas via EDA, motivam padronização (z-score) ou normalização (min-max) para algoritmos sensíveis a escala.

Codificação de Variáveis

EDA revela a natureza de variáveis categóricas (ordinais vs nominais, cardinalidade), informando se devemos usar one-hot encoding, label encoding ou embeddings.



Remoção de Outliers

Visualizações identificam valores extremos. Decisões sobre sua remoção dependem de contexto: são erros de medição ou observações legítimas raras?

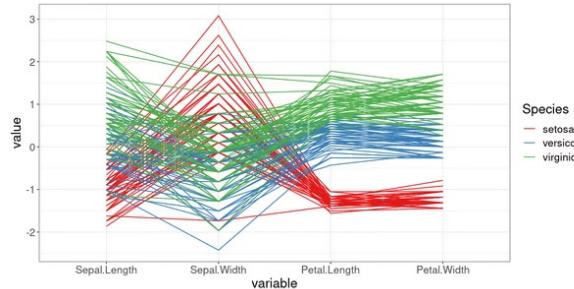
Transformações de Variáveis

Distribuições assimétricas observadas no EDA podem requerer transformações logarítmicas ou Box-Cox para melhorar a performance de modelos.

- 💡 **Princípio Fundamental:** EDA não é isolado – ele alimenta diretamente a construção de melhores modelos. Cada insight visual deve se traduzir em uma decisão concreta de pré-processamento.

Coordenadas Paralelas de um Dataset

As coordenadas paralelas são uma técnica poderosa para visualizar dados multivariados, mapeando cada variável para um eixo vertical paralelo. Linhas conectando pontos através desses eixos revelam padrões, tendências e agrupamentos que seriam difíceis de detectar em outras visualizações.



```
ggparcoord(  
  data = iris,  
  columns = c(1:4),  
  group = 5  
) +  
  theme_bw(base_size = 10) +  
  scale_color_manual(values = colors[1:3])
```

Interpretação Visual

Cada linha representa uma observação individual. Linhas que seguem trajetórias similares indicam observações com características semelhantes. Cruzamentos e divergências revelam relações complexas entre variáveis.

Detecção de Clusters

Grupos de linhas paralelas indicam clusters naturais nos dados, facilitando a identificação de subpopulações.

Identificação de Outliers

Linhas que se desviam significativamente do padrão geral representam observações atípicas que merecem investigação.

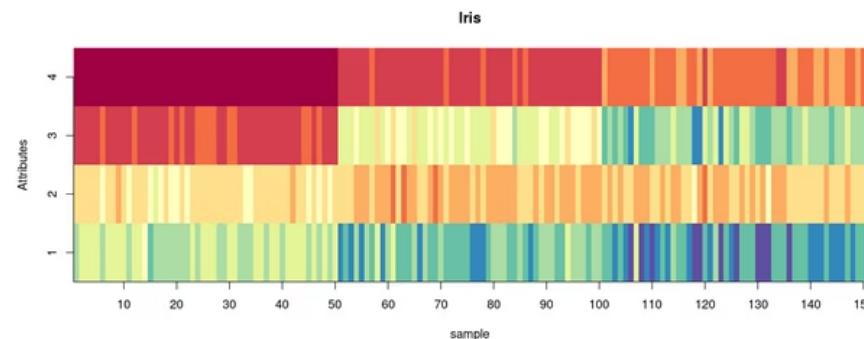
Relações Entre Variáveis

A forma como as linhas se cruzam entre eixos adjacentes revela correlações e dependências entre variáveis.

Referência: A. Inselberg and B. Dimsdale, 1990, Parallel coordinates: A tool for visualizing multi-dimensional geometry, IEEE Conference on Visualization - Visualization '90, p. 361–378.

Técnicas de Visualização Orientadas a Pixels

As técnicas orientadas a pixels representam cada ponto de dados como um pixel colorido em pequenas janelas múltiplas, permitindo a visualização eficiente de datasets muito grandes e de alta dimensionalidade. Esta abordagem é particularmente útil quando lidamos com milhares ou milhões de observações.



1

Criação de Janelas Múltiplas

Para um dataset de m dimensões, cria-se m janelas na tela, uma para cada dimensão, permitindo visualização simultânea de todas as variáveis.

2

Mapeamento de Valores

Os m valores dimensionais de cada registro são mapeados para m pixels nas posições correspondentes em suas respectivas janelas.

3

Codificação por Cor

As cores dos pixels refletem os valores correspondentes, criando uma representação visual compacta onde padrões emergem através de gradientes e agrupamentos de cores.

- Escalabilidade:** Esta técnica brilha com grandes volumes de dados. Onde scatter plots se tornam ilegíveis com milhares de pontos sobrepostos, visualizações orientadas a pixels mantêm clareza através da codificação eficiente por cor.

Referência: D.A. Keim, 2000, Designing pixel-oriented visualization techniques: theory and applications, IEEE Transactions on Visualization and Computer Graphics, v. 6, n. 1, p. 59–78.

Código de Implementação

```
mat <- as.matrix(iris[,1:4])
x <- (1:nrow(mat))
y <- (1:ncol(mat))

image(x, y, mat, col = brewer.pal(11, 'Spectral'),
      axes = FALSE, main = "Iris", lab = "sample", ylab = "Attributes")

axis(2, at = seq(0, ncol(mat), by = 1))
axis(1, at = seq(0, nrow(mat), by = 10))
```

Técnicas de Visualização Baseadas em Ícones

Os métodos baseados em ícones utilizam representações simbólicas para codificar múltiplas variáveis visualmente, oferecendo reconhecimento intuitivo de padrões em dados complexos. Essas técnicas aproveitam a capacidade humana de processar informações visuais de forma holística e rápida.



Chernoff Faces

Mapeia variáveis para características faciais (formato dos olhos, tamanho do nariz, etc.), explorando nossa habilidade natural de reconhecer e distinguir rostos.



Salience (Relevância)

Usa características visuais proeminentes para destacar aspectos importantes dos dados, direcionando a atenção do observador.



Codificação por Forma

Utiliza formas geométricas para representar diferentes tipos de informação, aproveitando nossa capacidade de distinguir rapidamente entre círculos, quadrados, triângulos e outras formas.



Ícones Coloridos

Combina forma e cor para codificar informações adicionais, aumentando a dimensionalidade da visualização sem sacrificar clareza.



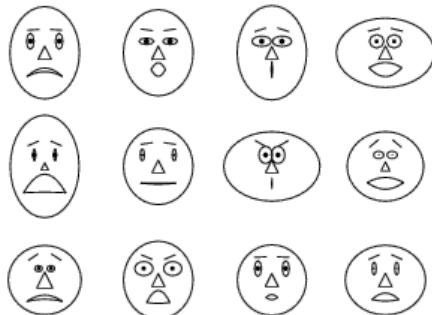
Tile Bars

Usa pequenos ícones para representar vetores de características relevantes em recuperação de documentos, permitindo comparações rápidas entre múltiplos itens.

Vantagem Cognitiva: Técnicas baseadas em ícones exploram o processamento visual pré-atencional do cérebro humano, permitindo que padrões complexos sejam reconhecidos quase instantaneamente, antes mesmo da análise consciente.

Chernoff Faces

As Chernoff Faces são uma técnica inovadora de visualização multivariada que atribui valores de variáveis a características faciais. Esta abordagem criativa explora a extraordinária capacidade humana de detectar, distinguir e lembrar rostos, transformando dados numéricos abstratos em representações visuais intuitivas e memoráveis.



Características Faciais Codificadas

Cada característica facial representa uma dimensão dos dados. A figura demonstra rostos produzidos usando dez características: excentricidade da cabeça, tamanho dos olhos, espaçamento dos olhos, excentricidade dos olhos, tamanho da pupila, inclinação da sobrancelha, tamanho do nariz, formato da boca, tamanho da boca e abertura da boca. Cada característica pode assumir um de 10 valores possíveis.

Mapeamento de Variáveis

Por exemplo, x pode ser a inclinação da sobrancelha, y o tamanho dos olhos, e z o comprimento do nariz, criando uma representação facial única para cada observação.

Reconhecimento de Padrões

Rostos similares indicam observações com valores próximos nas variáveis. Nossa habilidade inata de distinguir rostos facilita a identificação de clusters e outliers.

Análise Comparativa

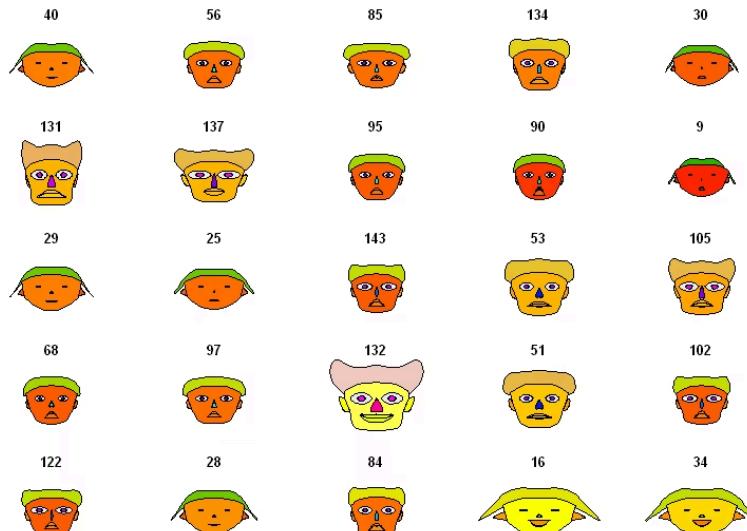
Comparar múltiplos rostos lado a lado permite avaliar rapidamente diferenças e semelhanças entre grupos ou categorias nos dados.

- Fundamento Neurológico:** O cérebro humano possui áreas especializadas (como o giro fusiforme) dedicadas exclusivamente ao reconhecimento facial. As Chernoff Faces aproveitam esta capacidade evolutiva para facilitar a análise de dados complexos.

Referências: Gonick, L. and Smith, W., 1993, *The Cartoon Guide to Statistics*. Harper Perennial, p. 212 | Weisstein, Eric W., "Chernoff Face", MathWorld - A Wolfram Web Resource.

Exemplo de Chernoff Faces com Dataset Iris

Esta demonstração prática aplica a técnica de Chernoff Faces ao famoso dataset Iris, mapeando as quatro variáveis numéricas (comprimento e largura de sépala e pétala) para características faciais distintas. Cada rosto representa uma observação individual do dataset, permitindo análise visual rápida de similaridades e diferenças.



```
set.seed(1)
sample_rows = sample(1:nrow(iris), 25)
isample = iris[sample_rows,]
labels = as.character(rownames(isample))
isample$Species <- NULL
```

```
faces(isample,
      labels = labels,
      print.info = F,
      cex = 1
    )
```

Selecionamos aleatoriamente 25 observações para manter a visualização legível e interpretável.

Detecção Visual de Grupos

Rostos com expressões similares tendem a pertencer à mesma espécie, revelando agrupamentos naturais nos dados sem codificação explícita de classe.

Identificação de Outliers

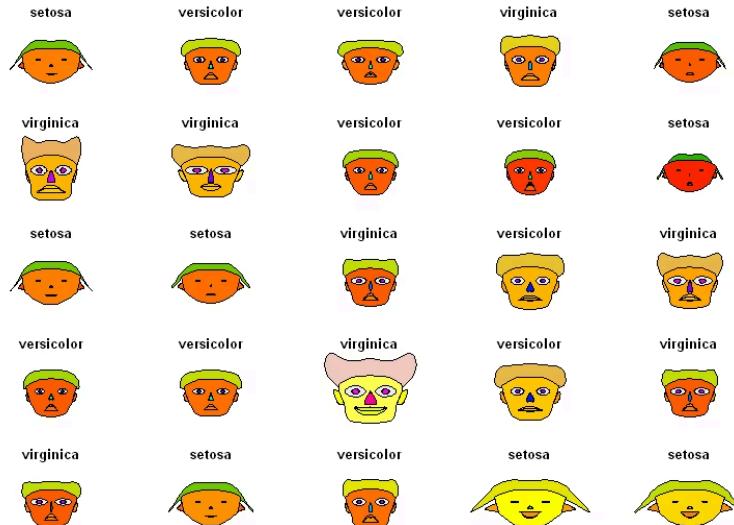
Rostos com aparência distintamente diferente da maioria podem indicar observações atípicas ou exceções interessantes que merecem investigação.

Análise Multivariada Intuitiva

Em vez de analisar quatro números para cada observação, processamos instantaneamente padrões através das configurações faciais, aproveitando processamento visual pré-atencional.

Chernoff Faces com Rótulo de Classe no Dataset Iris

Adicionar rótulos de classe às Chernoff Faces permite validação visual imediata de se os padrões faciais correspondem efetivamente às distinções reais entre classes nos dados. Esta versão transforma a visualização em uma ferramenta de validação de classificação visual.



```
set.seed(1)
sample_rows = sample(1:nrow(iris), 25)
isample = iris[sample_rows,]
labels = as.character(isample$Species)
isample$Species <- NULL

faces(isample,
      labels = labels,
      print.info = F,
      cex = 1
    )
```

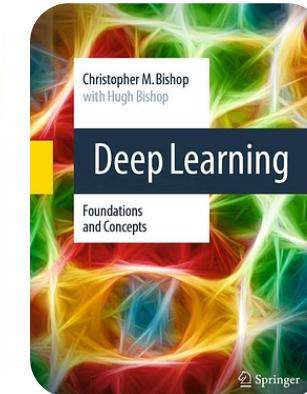
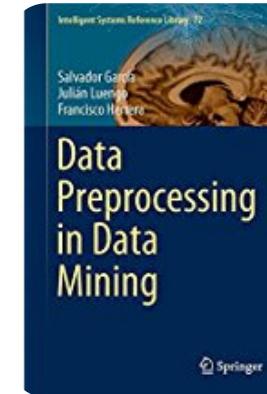
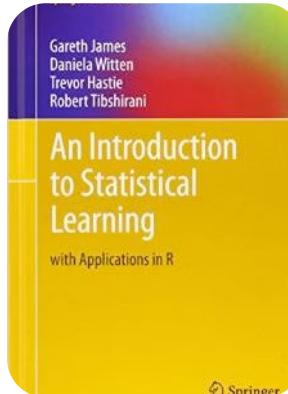
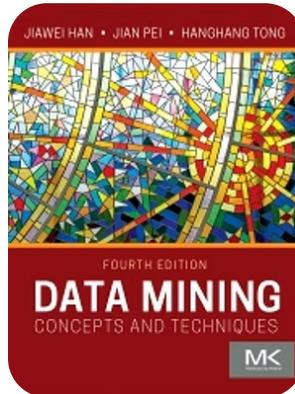
Se rostos similares consistentemente compartilham a mesma classe (espécie), isso confirma que as variáveis numéricas capturaram efetivamente as diferenças entre grupos. Rostos dissimilares dentro da mesma classe sugerem variabilidade intra-classe ou possíveis erros de classificação.

Aplicação Prática: Esta técnica é especialmente útil em contextos onde stakeholders não-técnicos precisam entender separabilidade de classes. Rostos são intuitivamente compreensíveis, eliminando barreiras de comunicação sobre conceitos estatísticos complexos.

Esta apresentação demonstrou técnicas sofisticadas de visualização de dados em R, desde análises básicas até métodos avançados como Chernoff Faces, fornecendo um toolkit completo para Análise Exploratória de Dados eficaz.

Referências Principais

Esta seleção de referências representa os pilares fundamentais para o estudo aprofundado de mineração de dados, cobrindo desde conceitos básicos até técnicas avançadas e aplicações contemporâneas.



1. **J. Han, J. Pei, and H. Tong**, *Data Mining: Concepts and Techniques*, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
2. **G. M. James, D. Witten, T. Hastie, and R. Tibshirani**, *An Introduction to Statistical Learning: With Applications in R*. Springer Nature, 2021.
3. **S. Garcia, J. Luengo, and F. Herrera**, *Data Preprocessing in Data Mining*. Springer, 2014.
4. **C. M. Bishop and H. Bishop**, *Deep Learning: Foundations and Concepts*. Springer Nature, 2023.