Tópicos Especiais em Computação I

Análise de Dados

Patrícia Lucas

Bacharelado em Sistemas de Informação
IFNMG - Campus Salinas

Salinas
Março 2021

INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Salinas

Versão 1.2021

Referência

Análise dos dados

Capítulo 2: Análise de dados.

Inteligência Artificial: Uma abordagem de aprendizado de máquina. Katti Faceli...[et al.]. - Rio de Janeiro: LTC, 2011.

Conjuntos de dados

Análise dos dados

Quantos dados são gerados a cada dia?

500 milhões de tweets são enviados.
294 bilhões de e-mails são enviados.
4 petabytes de dados são criados no Facebook.
4 terabytes de dados são criados a partir de cada carro conectado.
65 bilhões de mensagens são enviadas no WhatsApp.
5 bilhões de pesquisas são feitas.

Em 2025, estima-se que 463 exabytes de dados serão criados a cada dia em todo o mundo - isso é o equivalente a 212.765.957 DVDs por dia!

https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Conjuntos de dados

Análise dos dados

Além de serem gerados por várias fontes diferentes, esses dados também assumem formatos diferentes:

- Séries temporais
- Imagens
- Vídeos
- grafos
- texto
- páginas web...

Conjuntos de dados

Introdução

Apesar do crescente número de bases de dados disponíveis, na maioria das vezes não é possível utilizar algoritmos de aprendizado de máquina diretamente sobre esses dados.

A análise dos dados e as técnicas de pré-processamento são frequentemente utilizadas para tornar os conjuntos de dados mais adequados para o uso desses algoritmos.

Análise dos dados

Análise dos dados

A análise das características presentes em um conjunto de dados permite a descoberta de padrões e tendências que podem fornecer informações valiosas que ajudem a compreender o processo que gerou os dados.

Muitas dessas características podem ser obtidas por meio da aplicação de fórmulas estatísticas simples ou podem ser observadas por meio do uso de técnicas de visualização.

Caracterização dos dados

Introdução

Os conjuntos de dados são formados por **objetos**, que por sua vez possuem **atributos**.

atributos = atributos de entradas = vetor de características = atributos preditivos

Cada **objeto** corresponde a uma ocorrência dos dados.

Cada **atributo** está associado a uma propriedade do objeto.

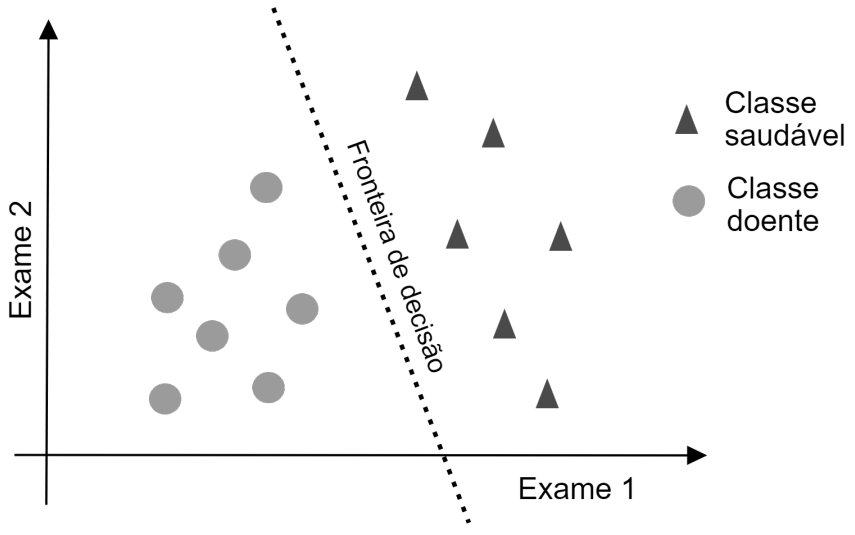
Caracterização dos dados

Análise dos dados

Formalmente, os dados podem ser representados por uma matriz de objetos X_n^d, em que n é o número de objetos e d é o número de atributos de entrada.

O valor d define a dimensionalidade dos objetos ou do espaço de objetos e podem ser vistos também como um conjunto de eixos ortogonais e os objetos, como pontos do espaço de dimensão d.

Exame 2	**Exame 1**	**Diagnóstico**
1	2	Doente
1	10	Saudável



Caracterização dos dados

Análise dos dados

Exemplo de conjunto de dados de pacientes de um hospital:

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

- Cada linha representa 1 paciente.

- Cada paciente possui um vetor de características com Nome, Idade, Sexo, Peso, Manchas, Temperatura, Número de internações e Estado onde reside.

- Cada paciente também possui um atributo de saída (rótulo), que representa o fenômeno de interesse sobre o qual se deseja fazer previsões.

- Dependendo do problema, podem existir mais de um atributo de saída ou não existir nenhum.

Caracterização dos dados

Análise dos dados

Quando o atributo alvo contém rótulos que identificam categorias ou classes às quais os objetos pertencem, ele é denominado classe e assume valores discretos $1, ..., k$.

Nesse caso temos um problema de classificação!

Classe majoritária: classe que possui maior número de objetos.

Classe minoritária: classe que possui menor número de objetos.

Caracterização dos dados

Análise dos dados

Se o atributo de saída contém valores numéricos contínuos, tem-se *um problema de regressão!*

Uma caso especial de problema de regressão é a previsão de séries temporais.

Caracterização dos dados

Análise dos dados

Tipos de dados: diz respeito ao grau de quantização nos dados.

- Se o atributo representa quantidades, ele é denominado quantitativo ou numérico.
  Contínuo: podem assumir um número infinito de valores.
  Exemplo: peso.
  Discreto: assumem um número finito ou infinito contável.
  Exemplo: número de filhos.

- Se o atributo representa qualidade, ele é denominado qualitativo ou categórico.
  Exemplo: tamanho(baixa, média, alta) ou sexo(masculino, feminino).

Caracterização dos dados

Análise dos dados

Escala: defini que operações podem ser realizadas sobre os valores.

- Nominais: qualitativos e não possuem relação de ordem ($=$ e $\neq$).
- Ordinais: qualitativos e possuem relação de ordem ($=$, $\neq$, $<$, $>$, $\leq$ e $\geq$).
- Intervalares: quantitativos, representados por números dentro de um intervalo.
- Racionais: quantitativos e são os que mais carregam informações, pois seus valores têm significado absoluto.

Atributo	Classificação
Id.	Nominal
Nome	Nominal
Idade	Racional
Sexo	Nominal
Peso	Racional
Manchas	Nominal
Temp.	Intervalar
#Int.	Racional
Est.	Nominal
Diagnóstico	Nominal

Exploração de dados

Análise dos dados

Uma grande quantidade de informações úteis pode ser extraída de um conjunto de dados por meio de sua análise ou exploração.

Informações obtidas durante a exploração podem ajudar na escolha da técnica mais apropriada de pré-processamento e também do aprendizado.

A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados como, por exemplo:
- Frequência.
- Localização ou tendências central (ex: média).
- Dispersão ou espalhamento (ex: desvio padrão).
- Distribuição ou formato.

Exploração de dados

Dados univariados

Dados univariados: quando o objeto possui apenas um atributo.

Medidas de localidade: definem pontos de referência nos dados.
Ex: moda, média, mediana e os quartis.

- Média: dado um conjunto de n valores numéricos $x = x_1, x_2, ..., x_n$, o valor médio desse conjunto é dado pela Equação 1:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

OBS: a média é sensível a *outliers*!

Outliers são valores muito diferentes dos demais valores observados para um mesmo atributo. Vamos verificar isso...

Uma turma de 7 alunos possui as seguintes idades:

$$18, 19, 20, 20, 23, 24, 24.$$

A média de idades dessa turma é: $21,1$.

Supondo que um aluno de 80 anos entre para essa turma, a média passaria a ser $28,5$.

Dessa forma, percebemos que a média é um bom indicador do meio de um conjunto apenas se os valores estão distribuídos simetricamente.

O problema com os *outliers* é minimizado com o uso da mediana.

- Mediana: dado um conjunto de n valores numéricos ordenado $x = x_1, x_2, ..., x_n$, a mediana desse conjunto é dada pela Equação 2:

$$mediana(x) = \begin{cases} \frac{1}{2}(x_r + x_{r+1}) & \text{se } n \text{ for par} \\ x_{r+1} & \text{se } n \text{ for ímpar} \end{cases} \quad (2)$$

Voltando ao exemplo da turma, a mediana dos conjunto de idades seria $21,5$, que é um valor bem mais realista.

Exploração de dados

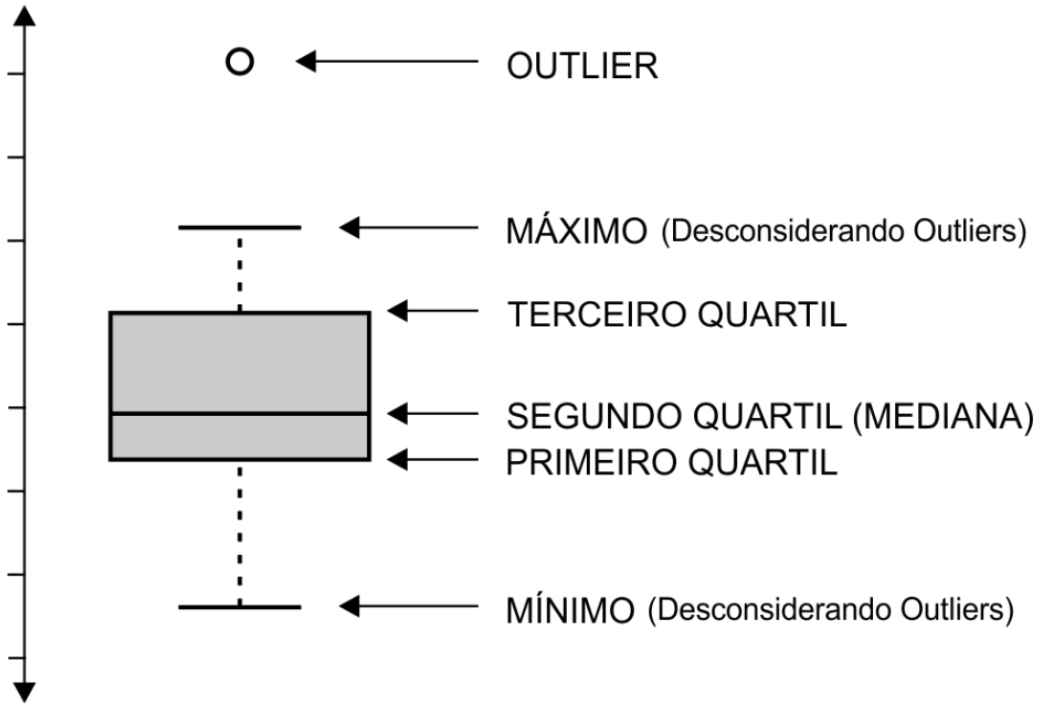
Dados univariados

- Moda: é o valor encontrado com maior frequência para um atributo.
- Quartis: dividem os valores do conjunto ordenado da seguinte forma:

Exploração de dados

Dados univariados

Uma técnica para visualizar os quartis, mediana e *outliers* é o *box-plots*.



Exploração de dados

Dados univariados

Medidas de espalhamento: medem a dispersão de um conjunto de valores, permitindo observar se esses estão amplamente espalhados ou relativamente concentrados em torno de um valor.

A medidas mais comuns são:

- Intervalo.
- Variância.
- Desvio padrão.

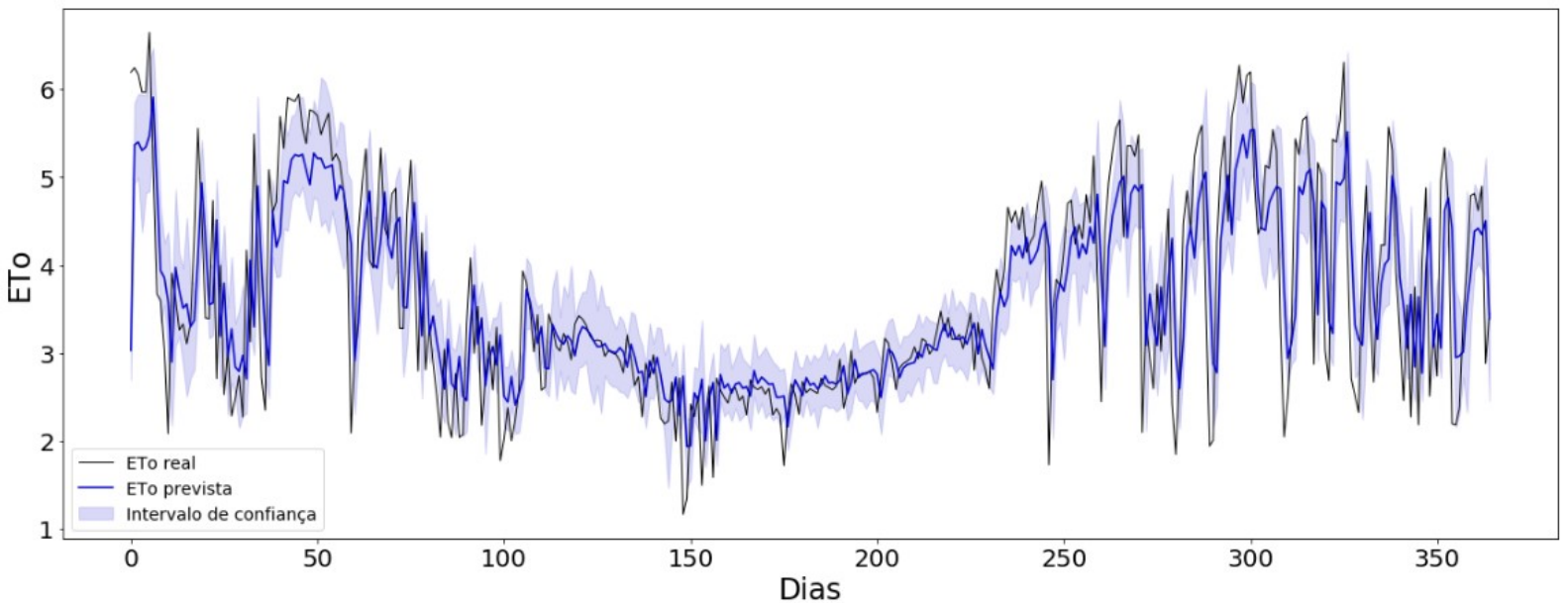
Exploração de dados

Dados univariados

Intervalo: mostra o espalhamento máximo entre os valores de um conjunto.

Sejam $x = x_1, x_2, ..., x_n$ os valores do atributo para n objetos. O intervalo desse conjunto é medido pela Equação 3:

$$intervalo(x) = \max_{i=1,...,n}(x_i) - \min_{i=1,...,n}(x_i) \quad (3)$$



Fonte da imagem: https://bit.ly/2Nvsc0s

Variância: avalia o espalhamento de valores em um conjunto.

Sejam $x = x_1, x_2, ..., x_n$ os valores do atributo para n objetos. A variância desse conjunto é medida pela Equação 4:

$$\text{variância}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{4}$$

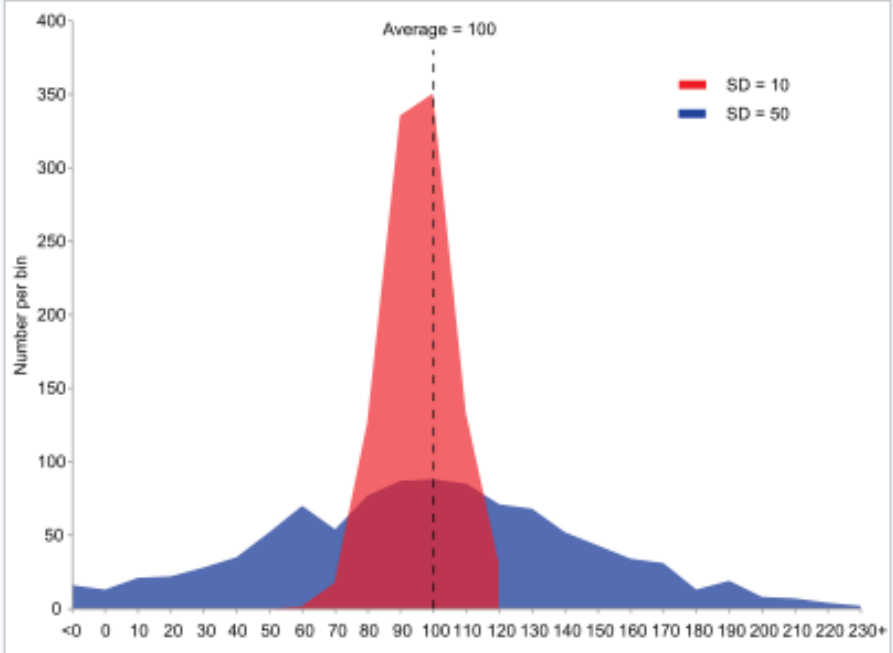
Onde: $\overline{x}$ é a média dos valores de x.

Outra medida de espalhamento é o desvio padrão, calculado a partir da raiz quadrada da variância.

Exploração de dados

Dados univariados

Para visualizar o desvio padrão e a distribuição dos dados, podemos representá-los em um **histograma**. Cada barra do histograma possui uma altura proporcional ao número de elementos com aquele valor no conjunto de dados.

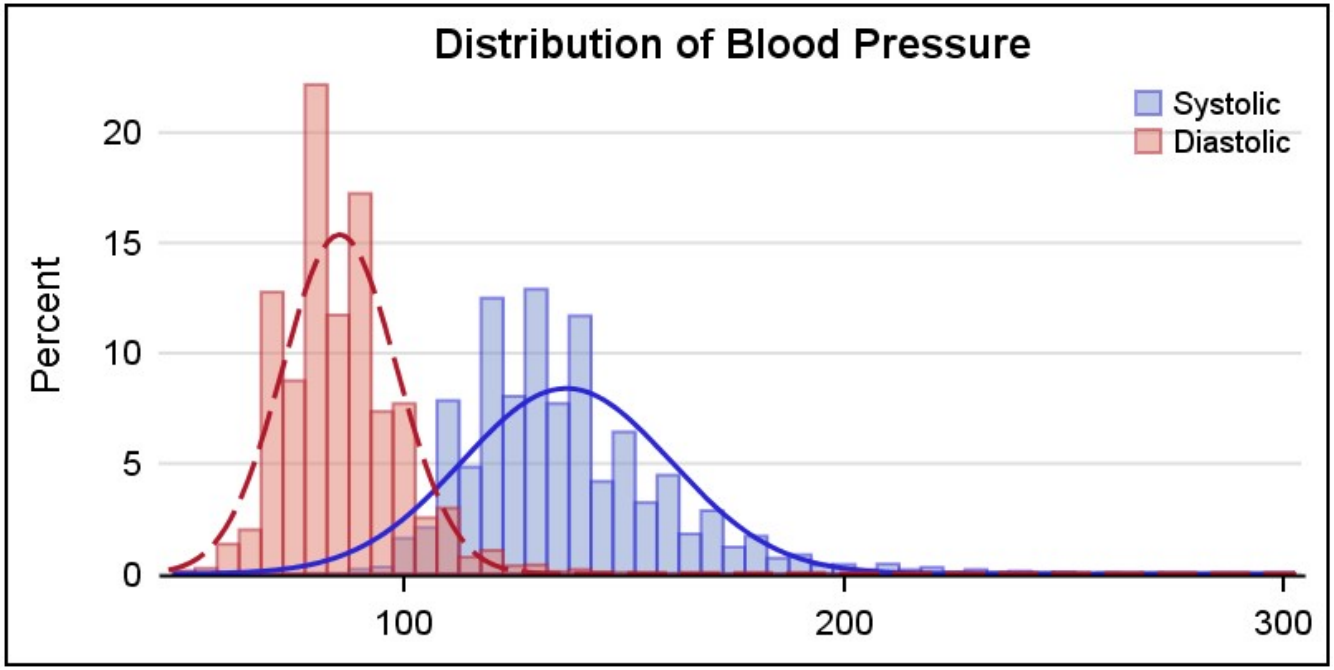


Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

Fonte da imagem: http://bit.ly/38UBwSZ

Exploração de dados

Dados univariados



Fonte da imagem: http://bit.ly/2Q94jg7

Exploração de dados

Dados multivariados

Dados multivariados: quando o objeto possui mais de um atributo.

Nesse caso as medidas de localidade e de espalhamento podem ser obtidas calculando a medida para cada atributo separadamente.

Dados multivariados ainda permitem análises da relação entre dois ou mais atributos através da medida de *correlação*.

Medida de correlação: apresenta uma indicação da força da relação linear entre dois atributos.

A *matriz de correlação* apresenta a correlação entre cada possível par de atributos de um conjunto de dados, onde de cada elemento tem seu valor definido pela Equação 5:

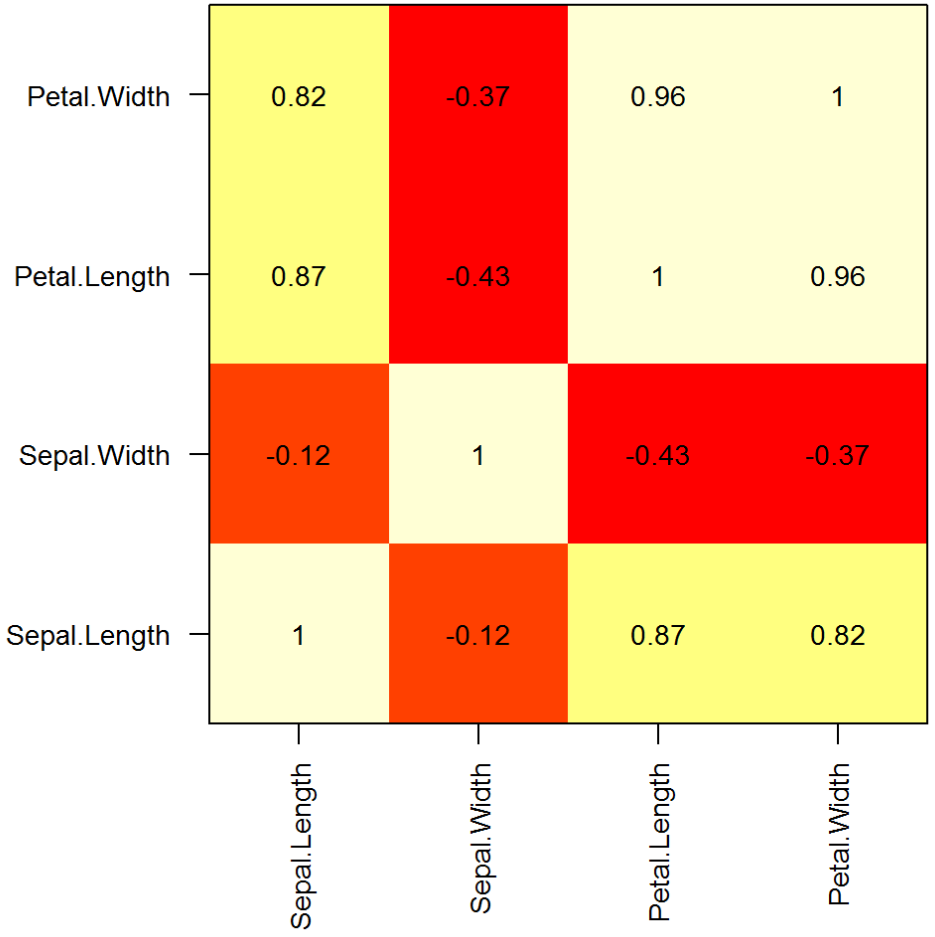
$$\text{correlação}(x^i, x^j) = \frac{\text{covariância}(x^i, x^j)}{s_i s_j} \tag{5}$$

Onde: x^i é o i-ésimo atributo e s_i é o desvio padrão dos valores desse atributo.

- A correlação$(x^i, x^i) = 1$.
- Correlação positiva: o aumento do valor x^i é acompanhado pelo aumento de x^j e vice-versa.
-Correlação negativa: a redução do valor x^i é acompanhado pelo aumento de x^j e vice-versa.

Exploração de dados

Dados multivariados

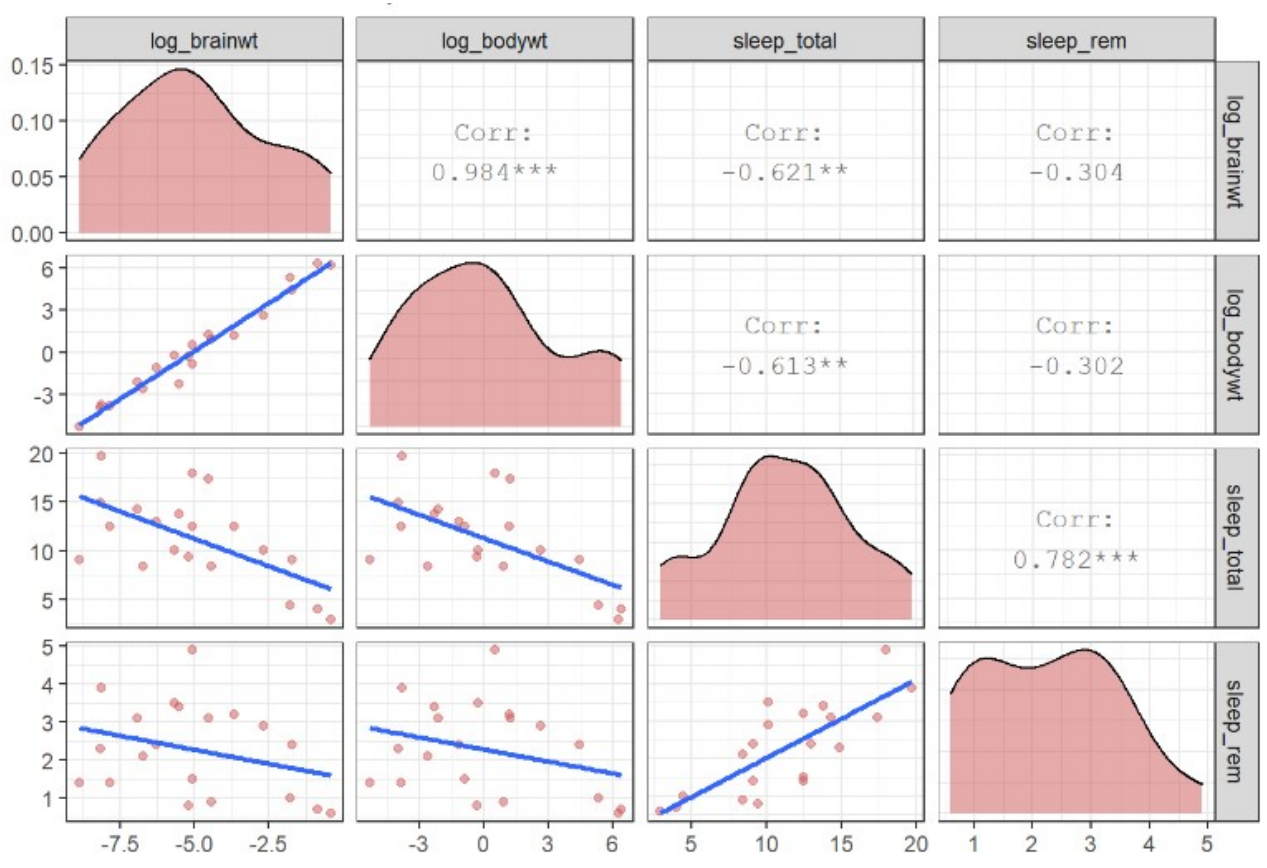


Fonte da imagem: http://bit.ly/2NvWmAH

Exploração de dados

Dados multivariados

Para visualizar a relação entre diferentes atributos, podemos usar um *scatter plot*.



Fonte da imagem: http://bit.ly/3vIuxpZ

Exploração de dados

Correlação X causalidade

Fonte da imagem: https://medium.com/analytics-vidhya/correlation-causation-977f71bb1e36