

Métodos Utilizados na Preparação de Dados



Stanley Robson de Medeiros Oliveira



Resumo da Aula

- ➡ **Aspectos relevantes** na preparação de dados
 - Por que **pré-processar** os dados?
 - Sumarização** de dados descritivos.
 - Qualidade** dos dados.
 - Integração** de dados.

Mineração de Dados: Fatores de Sucesso

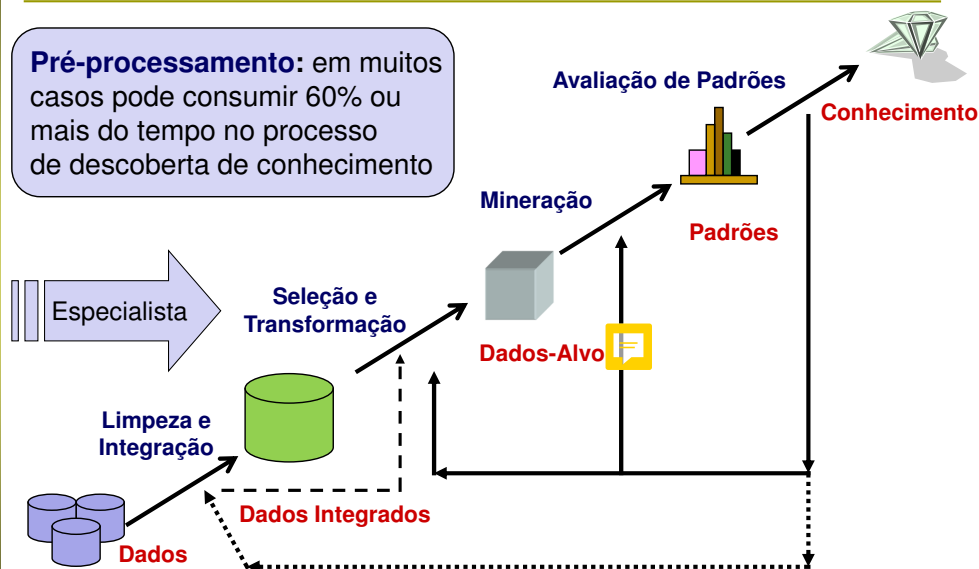
- Você sabe que **Mineração de Dados** é um projeto contínuo de busca de **inteligência e inferência** aplicada aos dados? 
- Você sabe com **detalhe** qual é o seu problema?
- Seus **objetivos** e **metas** estão claramente definidos?
- Você detém **técnicas** necessárias e possui equipe com domínio em **análise** de dados?
- Você tem os **dados necessários**, na **granularidade** desejada? 

Pré-processamento de dados

- No mundo real, **dados coletados** e organizados tendem a ser:
 - incompletos**;
 - com ruídos**;
 - redundantes**; e
 - inconsistentes**.
- A fase de **pré-processamento** tem início após a **coleta** e **organização** desses dados.
- Esta fase pode consumir **60%** ou **mais do tempo** para exploração de dados (Pyle, 1999).

Pré-processamento de dados

Pré-processamento: em muitos casos pode consumir 60% ou mais do tempo no processo de descoberta de conhecimento



Pré-processamento de dados

- O **sucesso** ou **fracasso** de um projeto de mineração de dados está relacionado à **preparação de dados**:
 - **Melhora** fortemente a precisão do modelo;
 - Produz grande economia em termos de **tempo**, **esforço** e **dinheiro**.
- A preparação de dados **ajuda** um analista a:
 - Interpretar melhor os resultados;
 - Entender os limites nos dados.

Exploração de Dados: Estágios

	Tempo Necessário (% do total)	Importância p/ Sucesso (% do total)
1. Identificação do Problema	10	15
2. Explorar possíveis soluções	9	14
3. Especificação da implementação	1	51
4. Mineração de dados		
4a. Preparação	60	15
4b. Explorar cenários	15	3
4c. Modelagem	5	2

FONTE: PYLE, D., Data Preparation for Data Mining, Morgan Kaufmann, 1999.

Resumo da Aula

- **Aspectos relevantes** na preparação de dados
- ➔ □ Por que **pré-processar** os dados?
- **Sumarização** de dados descritivos.
- **Qualidade** dos dados.
- **Integração** de dados.

Por que pré-processar os dados?

- No mundo real, geralmente os dados são (têm):
 - **Incompletos**: ausência de valores de atributos, ausência de atributos de interesse, ou dados com valores agregados.
 - **Ruídos**: existências de **erros** ou **outliers**.
 - **Inconsistentes**: informações desatualizadas ou oriundas de erros no momento de introdução dos dados.

Por que pré-processar os dados?

- Sem qualidade de dados, não há **qualidade nos resultados** em mineração de dados!
- **Decisões com qualidade** são baseadas em dados com qualidade.
- Dependendo da **tarefa de mineração de dados**, você precisa ter o processamento de dados correspondente.

Resumo da Aula

- **Aspectos relevantes** na preparação de dados
- Por que **pré-processar** os dados?
- ➔ □ **Sumarização** de dados descritivos.
- **Qualidade** dos dados.
- **Integração** de dados.

Características descritivas de dados

- **Motivação**:
 - Melhor entendimento sobre os dados: **tendência central**, **variação** e **distribuição**.
- **Medidas de posição e de dispersão dos dados**:
 - média, max, min, quartis, outliers, variância, etc.
- **Dimensões numéricas**: relação c/ intervalos ordenados.
 - **Dispersão de dados**: analisada em múltiplas granularidades.
 - Análise de **Boxplot** ou **quartil** em intervalos ordenados.
- **Medidas de Assimetria**: (**simetria e assimetria**)
 - Indicador da forma da distribuição dos dados.

Medidas de Posição (tendência central)

- ❑ **Média aritmética simples:** $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\mu = \frac{\sum x}{N}$
- ❑ **Média aritmética ponderada:** $\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$
- ❑ **Moda (Mo):**
 - É o valor mais frequente em um conjunto de valores numéricos.
- ❑ **Mediana (Md):**
 - Dado um grupo de dados ordenados, a **mediana** separa a metade inferior da amostra da metade superior.

Exemplos

- ❑ Para o seguinte **conjunto**: {1, 3, 5, 7, 9}
 - A **média** é 5;
 - A **mediana** é 5.
- ❑ No entanto, para o **conjunto**: {1, 2, 7, 7, 13}
 - A **mediana** é 7, enquanto a **média** é 6;
 - A **moda** é 7.
- ❑ Qual seria a **mediana** para o **conjunto**: {1, 2, 4, 10, 12, 13}?

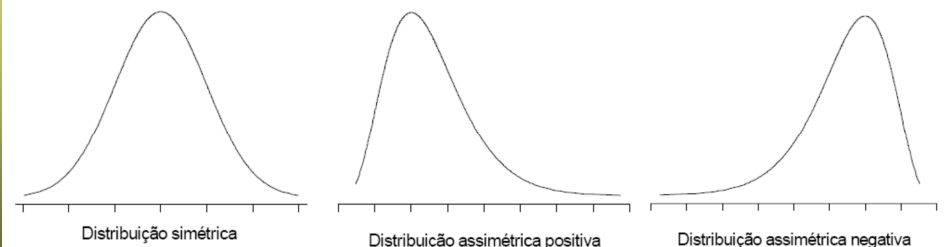
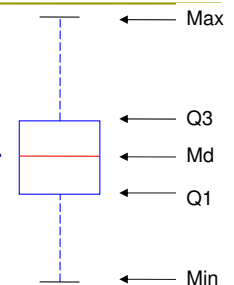
Mediana = (4 + 10)/2 = 7.

Separatrizes

- ❑ **Não** são medidas de tendência central.
- ❑ As separatrizes estão ligadas à mediana relativamente à sua característica de **separar a série** em duas partes que apresentam o **mesmo número de valores**.
- ❑ As separatrizes são:
 - **Quartil**: divide um conjunto de dados em **quatro** partes iguais.
 - **Decil**: divide um conjunto de dados em **dez** partes iguais.
 - **Percentil**: divide um conjunto de dados em **cem** partes iguais.

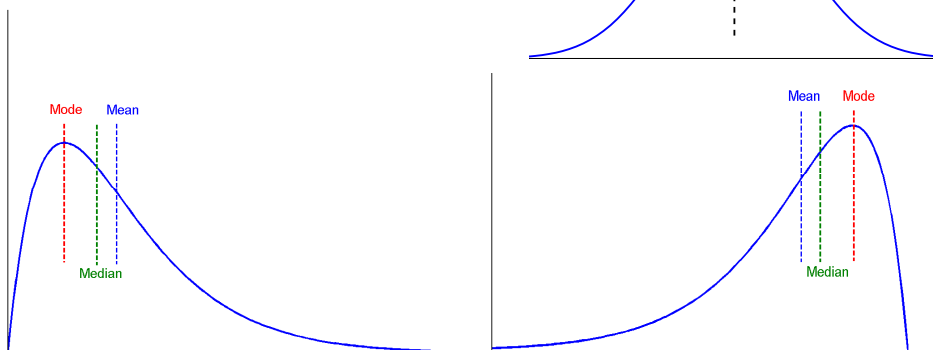
Exemplo de Separatrizes

- ❑ **Quartil**: Os quartis dividem o conjunto de dados em quatro partes iguais:
 - Se $(Md - Q1) = (Q3 - Md) \Rightarrow$ **dist. simétrica**.
 - Se $(Md - Q1) < (Q3 - Md) \Rightarrow$ **assimetria à direita** ou **positiva**;
 - Se $(Md - Q1) > (Q3 - Md) \Rightarrow$ **assimetria à esquerda** ou **negativa**.



Distribuição Simétrica e Assimétrica

- Mediana, média e moda de dados com distribuição **simétrica** e **assimétrica**.



Medindo a dispersão dos dados

- Quartils, outliers e boxplots

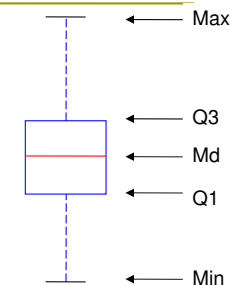
■ **Quartils:** Q_1 (25º percentil), Q_3 (75º percentil).

■ **Amplitude interquartilica:**
 $IQR = Q_3 - Q_1$ (50% dos dados).

■ **Sumário dos 5 números:** Min, Q_1 , Med, Q_3 , Max.

■ **Boxplot:** uma linha central mostrando a **mediana**, uma linha inferior mostrando o **primeiro quartil**, uma linha superior mostrando o **terceiro quartil**.

■ **Outliers:** Limite Inferior = $Q_1 - 1.5 \times IQR$;
 Limite Superior = $Q_3 + 1.5 \times IQR$.



Medindo a dispersão dos dados ...

- Variância e desvio padrão (*amostra: s , população: σ*)

■ **Variância:** $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$

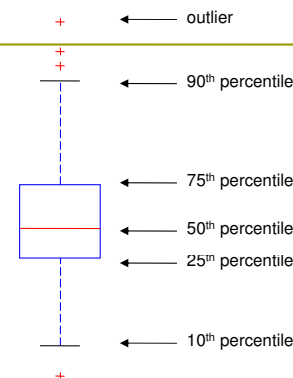
■ **Desvio padrão s (ou σ)** é a raiz quadrada da variância s^2 (ou σ^2)

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

Análise de Boxplot

- Boxplot:**

■ Um **percentil** é uma medida da **posição relativa** de uma unidade observacional em relação a todas as outras.

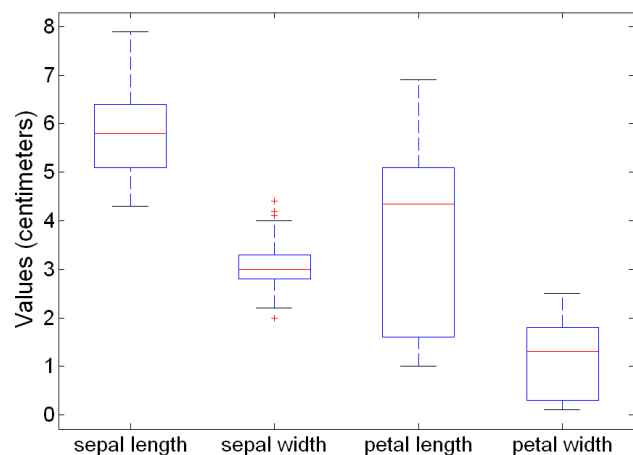


■ O **p -ésimo** percentil tem no mínimo **$p\%$ dos valores abaixo** daquele ponto e no mínimo **$(100 - p)\%$ dos valores acima**.

■ Se uma altura de 1,80m é o **90º percentil** de uma turma de estudantes, então **90% da turma tem alturas menores que 1,80m e 10% tem altura superior a 1,80m**.

Exemplo de BoxPlots

- Boxplots podem ser usados para comparar a dispersão dos valores de atributos.

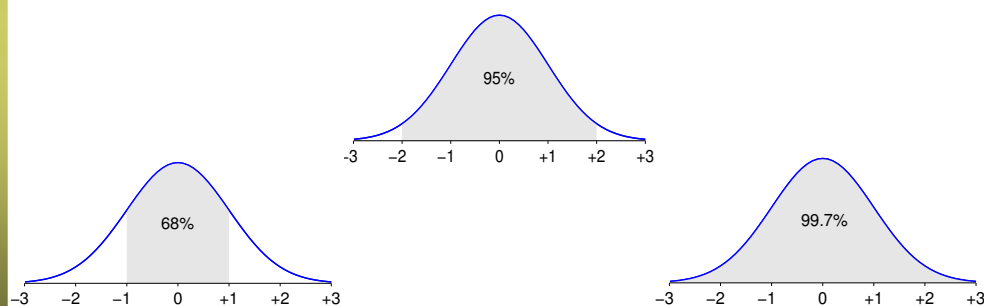


Exercício sobre Boxplot

- Carregue o arquivo **cpu.csv** e depois use o script **boxplot.R**:
- Determine o sumário dos cinco números: **min**, **Q1**, **Mediana**, **Q3**, **max** para o primeiro atributo.
- Esboce o **boxplot** dos dados desse arquivo.
- Quais os valores que podem ser considerados **outliers**?
- Faça o mesmo procedimento para as outras variáveis (**atributos**).

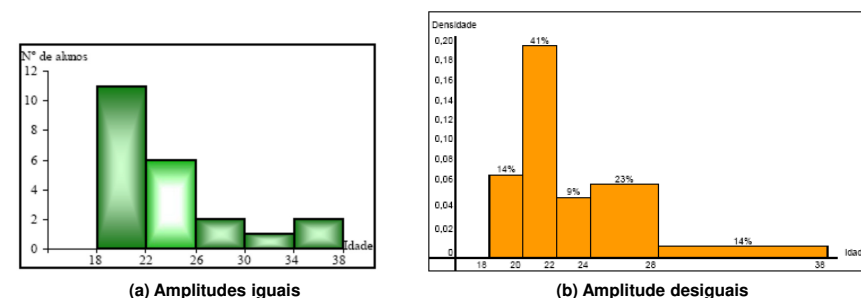
Propriedades da distribuição normal

- A **distribuição normal** com média μ e desvio padrão σ :
 - No intervalo de $\mu - \sigma$ até $\mu + \sigma$: contém 68% das observações;
 - No intervalo de $\mu - 2\sigma$ até $\mu + 2\sigma$: contém 95% das observações;
 - No intervalo de $\mu - 3\sigma$ até $\mu + 3\sigma$: contém 99.7% das observações.



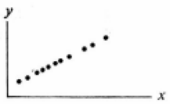
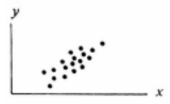
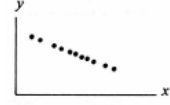

Análise de Histogramas

- Gráfico que mostra a **estatística básica** da descrição de classes.
 - Histograma de Frequências**
 - Mostra a distribuição dos valores de uma variável;
 - Consiste em um conjunto de retângulos, em que cada retângulo representa a frequência de uma das classes presentes nos dados.

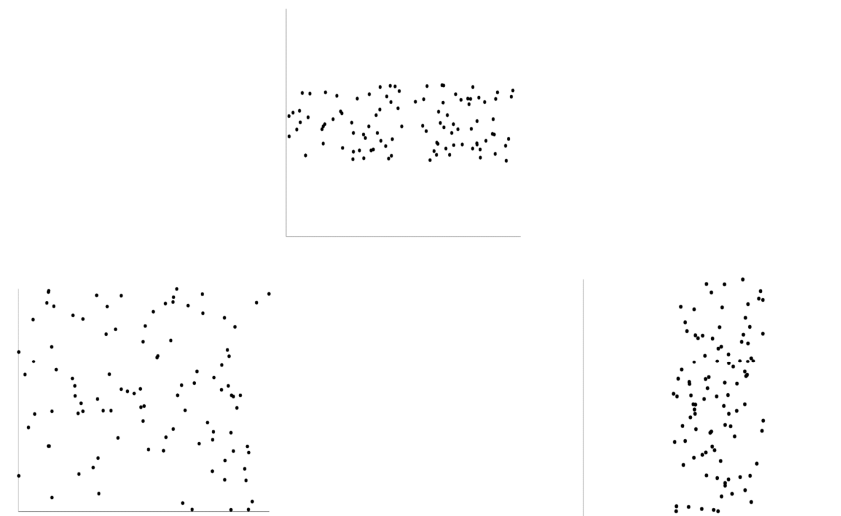


Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Exemplos de diagramas de dispersão

Valor de r	Descrição do relacionamento	Diagrama de dispersão
+1,00	Relacionamento positivo perfeito	
Cerca de 0,7	Relacionamento positivo moderado	
-1,0	Relacionamento negativo perfeito	
Cerca de -0,7	Relacionamento negativo moderado	

Exemplos de ausência de relacionamento



Exercício: histograma e correlação

- ❑ Carregue o arquivo **cpu.csv** e depois use o script **hist-correl.R**:
- ❑ Esboce o **histograma** de algumas variáveis desse arquivo.
- ❑ Calcule a **matriz de correlação** dos atributos do arquivo.
- ❑ Como eliminar variáveis com **alta correlação** (acima de 70%)?

Resumo da Aula

- ❑ **Aspectos relevantes** na preparação de dados
- ❑ Por que **pré-processar** os dados?
- ❑ **Sumarização** de dados descritivos.
- ➔ ❑ **Qualidade** dos dados.
- ❑ **Integração** de dados.

Qualidade dos Dados

□ Relevância:

- “**Data cleaning** is one of the three biggest problems in data warehousing” — **Ralph Kimball**.
- “**Data cleaning** is the number one problem in data warehousing” — **DCI (Downtown Cincinnati Inc.) survey**.

□ Procedimentos para **qualidade dos dados**:

- Preencher **valores faltantes**;
- Identificar **outliers** e remover **ruídos** nos dados;
- Corrigir e eliminar **inconsistências**.
- Remover **redundâncias** causadas pela **integração de dados**.

Valores faltantes

□ Em muitos casos, **dados podem ser incompletos**:

- Muitas observações podem não possuir valores para alguns atributos (**Ex.:** **renda anual de clientes em dados de vendas**).

□ **Valores faltantes** ocorrem devido:

- Problemas com equipamentos (**perdas de dados**);
- Inconsistência com outros registros e portanto são deletados;
- Dados não digitados por causa de mal interpretação;
- Alguns dados não são importantes no momento da entrada;
- Falta de registros históricos ou mudança nos dados.

□ Em muitos casos, **valores faltantes** podem ser inferidos.

Valores faltantes ...

□ **Razões para os valores faltantes**:

- Informação não foi coletada:
(**Ex.:** pessoas não querem fornecer suas idades).
- Atributo pode não ser aplicado em todos os casos:
(**Ex.:** renda anual não é aplicada para crianças).

□ **Lidando com os valores faltantes**:

- **Eliminar** alguns objetos do conjunto de dados;
- **Estimar** os valores faltantes;
- **Ignorar** os valores faltantes durante a análise;
- **Substituir** com possíveis valores (**ponderados por suas probabilidades**).

Lidando com valores faltantes

□ **Método 1**: Ignorar as observações (**registros**):

- A alternativa mais simples.

□ Deve ser usado somente se a **observação** possui **vários** atributos com valores faltantes.

□ É um **método ineficiente**:

- Parte da informação é perdida;
- É um método pobre quando a porcentagem de valores faltantes varia entre os atributos.

Lidando com valores faltantes ...

- ❑ **Método 2:** Preencher os valores manualmente.
- ❑ Essa alternativa **só vale a pena** se o dataset for **muito pequeno**.
- ❑ **Ineficiência** desse método:
 - Consome muito tempo;
 - Impraticável para grandes datasets.

Lidando com valores faltantes ...

- ❑ **Método 3:** Usar a **média** do atributo para preencher os valores faltantes.
- ❑ **Exemplo:** se idade média de um grupo de pessoas é 35, esse valor deve ser usado para preencher os valores faltantes.
- ❑ **Vantagem:**
 - Procedimento simples de ser implementado.

Lidando com valores faltantes ...

- ❑ **Método 4:** Para atributo nominal, usar a **moda** para preencher os valores faltantes.
- ❑ A **moda** é o valor mais frequente em um conjunto de valores.
- ❑ Pode **não** ser uma boa alternativa quando o atributo considerado é o **atributo-meta**.

Lidando com valores faltantes ...

- ❑ **Método 5:** Usar a **média** para observações pertencentes a uma mesma classe.
- ❑ Nesse caso, o valor faltante não está no **atributo meta**.
- ❑ **Exemplo:** se um cliente não possui informação sobre o consumo mensal de cartão de crédito, substitua o valor faltante pela média de consumo de clientes na categoria (**mesma classe**).
- ❑ Em caso de **atributo nominal (não-meta)**, use a moda do atributo considerando as observações que pertencem a mesma classe.

Lidando com valores faltantes ...

- ❑ **Método 6:** Preencher os valores faltantes por meio de uma **regressão linear**.
- ❑ O primeiro passo é identificar se o **atributo com valores faltantes** tem uma boa correlação ($r > 0,7$) com um outro atributo do dataset.
- ❑ O segundo passo é fazer a **regressão** entre os **atributos correlacionados**.
- ❑ **Importante:** esse método deve ser usado com muito **cuidado**, pois pode **inserir ruído** nos dados.

Lidando com valores faltantes ...

- ❑ **Método 7:** Usar o método KNN (k- Nearest Neighbor) – **Vizinho mais próximo**.
- ❑ Eficiente para atributos **discretos** e **contínuos**.
- ❑ Para atributos **discretos**, usar o valor mais frequente entre os **k vizinhos** do valor faltante.
- ❑ Para atributos **contínuos**, usar a média entre os **k vizinhos** do valor faltante.
- ❑ A única **desvantagem** é que esse procedimento pode consumir muito tempo em grandes datasets.

Lidando com valores faltantes ...

- ❑ **Método 8:** Usar o **valor mais provável** que é **baseado em inferência**.
- ❑ **Exemplo:** Determinar o valor faltante usando uma **árvore de decisão**, um **modelo Bayesiano**, etc.
- ❑ O método é muito **eficiente**, mas é também muito **caro** computacionalmente.

Exercício: preencher valores faltantes

- ❑ Carregar o arquivo **soybean.arff** no Weka e responder as seguintes questões:
- ❑ Existem **valores faltantes**?
- ❑ Preencher os valores faltantes de todos os atributos usando a **média** para **atributos numéricos** e a **moda** para **atributos nominais**.
- ❑ Quais seriam as **melhores alternativas** para preencher os valores faltantes dos atributos desse dataset?

Ruído no dados

□ Ruído refere-se à **modificação** de **valores originais**.

□ **Exemplos:**

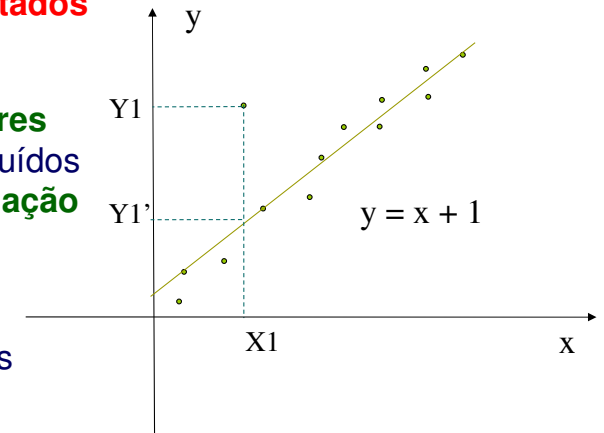
- Falhas nos equipamentos de coleta de dados;
- Problemas na entrada de dados;
- Problemas na transmissão de dados;
- Inconsistência na convenção de nomes;
- Transformações erradas aplicadas aos dados.

Regressão: reduzindo ruídos

□ Os **pontos dispersos** podem ser **representados** por uma **reta**.

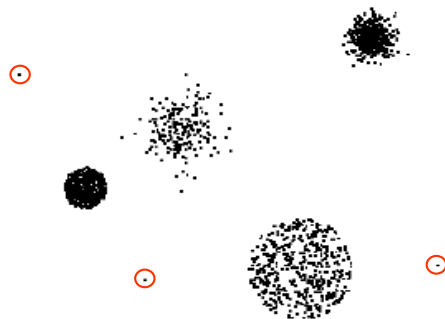
□ Em seguida, os **valores originais** são substituídos pelos valores da **equação da reta**.

□ Esse procedimento ameniza (**suaviza**) os **ruídos** nos dados.



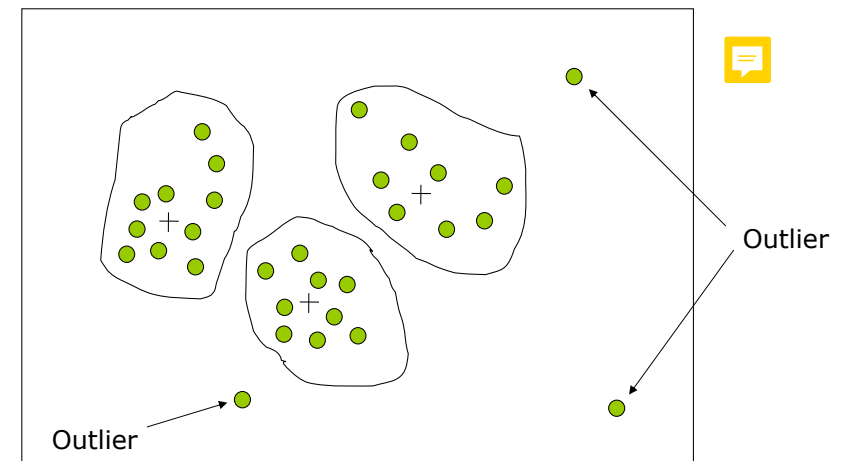
Outliers

□ **Outliers** são objetos com **características diferentes** da maioria dos outros objetos em um conjunto de dados.



Análise de Agrupamento

Outliers podem ser detectados por meios de agrupamentos (**clusters**). Intuitivamente, objetos que estão fora dos clusters são **outliers**.



Outliers – Proibido esquecer

PROBLEMA	ABORDAGEM
Análise Univariada	Boxplot
Análise Multivariada	Análise de Agrupamentos

Valores Redundantes

- O **dataset** pode incluir objetos que são **duplicados** ou **quase duplicados** de outros.
- **Exemplos:**
 - Ocorre quando dados são integrados de fontes heterogêneas.
 - Quando uma **variável** é uma **combinação linear** de **outras**.
 - Quando **duas** os **mais variáveis** são altamente **correlacionadas**.

Inconsistências

- **Erro na entrada de dados:** Tipo de inconsistência muito comum.
 - Causado quando mais de um usuário editam o mesmo arquivo.
 - **Exemplo:** para o atributo data, um usuário preenche os dados no formato “**dd/mm/aaaa**”, enquanto o outro usuário usa o formato “**yyyy/mm/dd**”.
- **Atributo com valores diferentes para a mesma informação:**
 - **Exemplo:** um atributo que armazena informação sobre Unidades da Federação assume os valores **São Paulo**, **SP**, **S.P.**, **S. Paulo**, **Sao_Paulo**.

Eliminação de inconsistências ...

- **Mesmo valor de um atributo para diferentes rótulos:** O mesmo dado é representado por rótulos diferentes.

ATRIBUTOS					CLASSE
Dia	Tempo	Temperatura	Umidade	Vento	Joga-Tenis
1	Sol	Quente	Alta	Fraco	Não
2	Sol	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuva	Moderado	Alta	Fraco	Sim
5	Chuva	Frio	Normal	Forte	Sim
6	Chuva	Frio	Normal	Forte	Não
7	Nublado	Frio	Normal	Forte	Sim
8	Sol	Moderado	Alta	Fraco	Não
9	Sol	Frio	Normal	Fraco	Sim

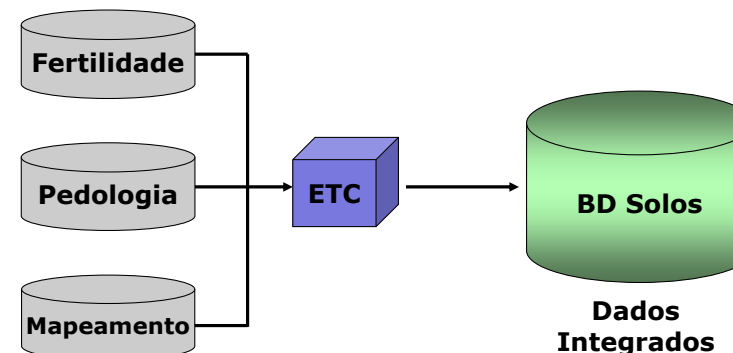
- A correção da inconsistência seria uma alteração do valor do atributo **Vento** para uma das **tuplas**.

Resumo da Aula

- ❑ **Aspectos relevantes** na preparação de dados
- ❑ Por que **pré-processar** os dados?
- ❑ **Sumarização** de dados descritivos.
- ❑ **Qualidade** dos dados.
- ➔ ❑ **Integração** de dados.

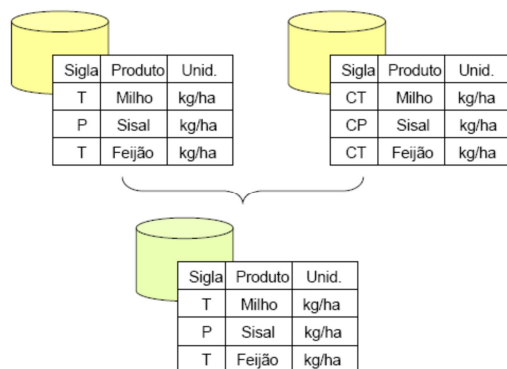
Integração de dados

- ❑ Processo que **combina dados** residentes em **diferentes fontes**, mantendo a **consistência** e a **coerência** dos dados integrados.



Integração de esquemas

- ❑ **Metadados** podem ser utilizados para ajudar a **unificar** os atributos e **transformar** os dados.



- ❑ O atributo **Sigla** do primeiro esquema assume os valores **T** e **P**, representando cultura **temporária** e cultura **perene**, enquanto no segundo esquema, os valores do atributo **Sigla** são **CT** e **CP**.

Lidando com redundância na integração

- ❑ Dados redundantes geralmente provêm da integração de múltiplas fontes de dados:
 - **Identificação de objeto**: o mesmo atributo pode ter diferentes nomes em diferentes arquivos (**datasets**);
 - **Dados derivados**: preço de um produto e o valor do imposto pago por ele (**combinação linear**).
- ❑ Atributos redundantes podem ser detectados por:
 - **Análise de correlação**: atributos numéricos; ou
 - **Teste do Qui-quadrado**: atributos nominais ou categóricos.

Análise de correlação (dados numéricos)

- Coeficiente de correlação (também conhecido como coeficiente de **Pearson**):

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

Onde n é o número de observações, \bar{A} e \bar{B} são as médias das variáveis A e B, σ_A e σ_B são os desvios-padrão de A e B.

- Se $r_{A,B} > 0$, A e B são **positivamente correlacionadas** (quanto maior for o **valor** $r_{A,B}$, maior será a **correlação** entre as **variáveis A e B**).
- $r_{A,B} = 0$: A e B são **independentes** ou não possuem relacionamento;
- $r_{A,B} < 0$: A e B são **negativamente correlacionadas**.

Análise de correlação (dados categóricos)

- χ^2 (teste do qui-quadrado)

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

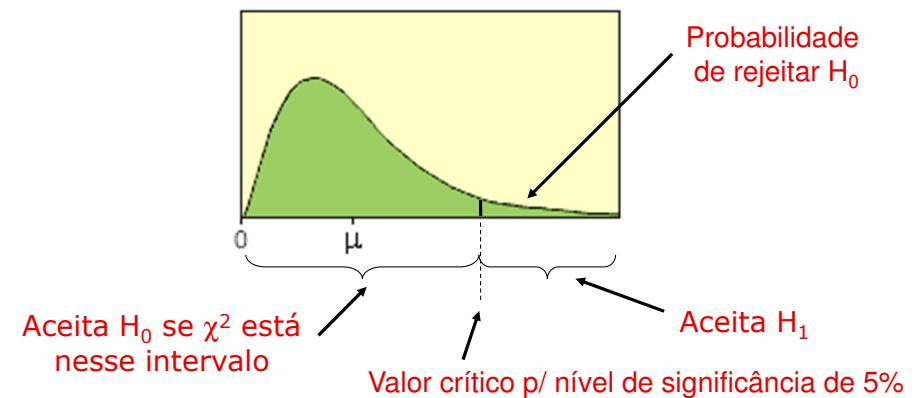
- As **frequências observadas** são obtidas diretamente dos dados das amostras, enquanto que as **frequências esperadas** são calculadas a partir destas.
- Quanto maior o valor de χ^2 , mais provável é a **correlação das variáveis**.
- Cuidado**: Correlação não implica causalidade:
 - Número de hospitais e número carros roubados em uma cidade pode ser correlacionado;
 - Ambas as variáveis estão ligadas com uma terceira variável: **população**.

Qui-quadrado (χ^2)

- O teste do χ^2 é muito eficiente para avaliar a associação existente entre variáveis qualitativas.
- O **analista de dados** estará sempre trabalhando com duas hipóteses:
 - H_0 : não há associação entre os atributos (**independência**)
 - H_1 : há associação entre os atributos.
- A hipótese H_0 é rejeitada para valores elevados de χ^2 .
- O cálculo dos **graus de liberdade** de χ^2 é dado por:
 $gl = (\text{número de linhas} - 1) \times (\text{número de colunas} - 1)$

Qui-quadrado (χ^2) ...

A forma da função de densidade de χ^2



Rejeitamos a **hipótese nula** se χ^2 for maior que o **valor crítico** fornecido pela tabela. Para 1 grau de liberdade, o **valor crítico** é **3,841**.

Exemplo do cálculo de χ^2

	Joga xadrez	Não joga xadrez	Soma (linhas)
Gosta de ficção científica	250(90)	200(360)	450
Não gosta de ficção científica	50(210)	1000(840)	1050
Soma (colunas)	300	1200	1500

- Os números entre parênteses são os **valores esperados**, calculados com base na distribuição dos dados das duas categorias.
- O resultado mostra que **gostar_ficção_científica** e **jogar_xadrez** são correlacionadas nesse grupo:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Neste caso, a hipótese nula é rejeitada, pois $507.93 > 3.841$. Então, existe correlação entre as variáveis estudadas.

Sugestões de Leitura

- Jiawei Han, Micheline Kamber, Jian Pei. **Data mining: concepts and techniques**. 3rd ed., 2012.
 - Capítulo 3 (**Data Preprocessing**).
- Ian H. Witten, Frank Eibe, Mark A. Hall. **Data mining: practical machine learning tools and techniques**. 3rd ed., 2011.
 - Capítulo 2 (**Input: Concepts, Instances, and Attributes**).