Align, Plan and Organize - APO
14.07 – Data Cleansing Approach

Prof. Dr. Luiz Camolesi Jr.

Align, Plan and Organize - APO14.07

Defina a abordagem de limpeza de dados.

Definir os mecanismos, regras, processos e métodos para validar e corrigir dados de acordo com políticas e regras de negócios predefinidas

Saneamento

ATIVIDADES

1. Estabelecer e manter uma política de limpeza de dados.

Nível 2

2. Manter o histórico de alterações dos dados realizadas nos processos de limpeza.

Nível 3

3. Estabelecer métodos para corrigir os dados em um plano operacional. Os métodos podem incluir comparação entre múltiplos repositórios, verificação com uma fonte válida, verificações lógicas, integridade referencial ou tolerância de intervalo.

4. Em Acordos de Nível de Serviço- ANS (SLA - Service Level Agreements), incluir critérios de qualidade de dados para responsabilizar e “cobrar” as fontes de dados pelos dados limpos.

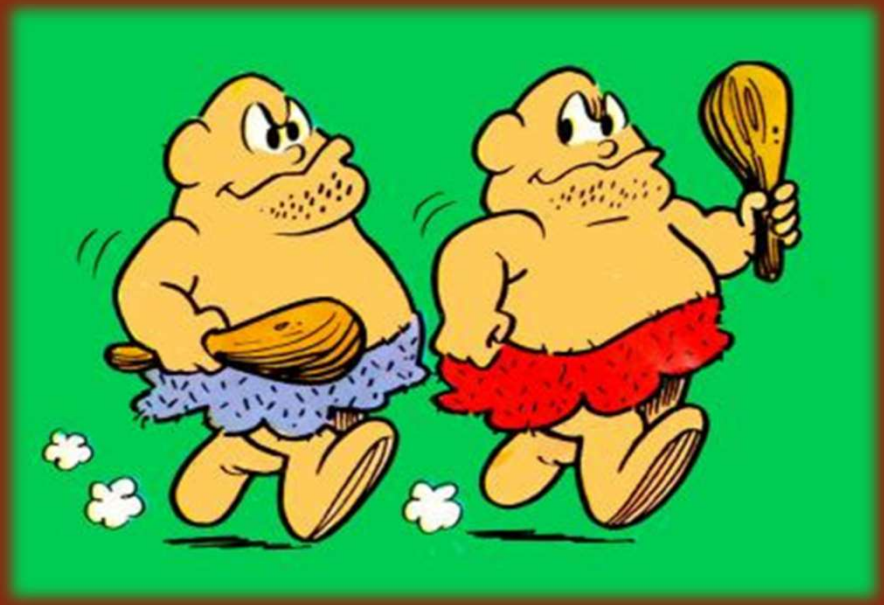
Nível 4

Documentação

Inputs	• Estratégia de Qualidade de Dados
Outputs	• Seleção de ferramentas de limpeza • Definição de padrões • Definição de Processos de limpeza • Política de limpeza

Limpeza de Dados

Duplicidades

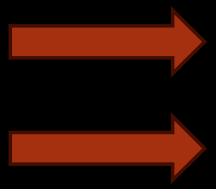


Redundâncias, repetições ou duplicações de informações devem ser verificadas e eliminadas.

Exceto se envolver estruturas (lógicas ou físicas) distintas de armazenamento que foram intencionalmente definidas pela engenharia de dados

Ocorrências :

- Informações iguais
- Informações quase iguais



ID	Nome	Email	Cidade	Data de Cadastro
1	Ana Silva	ana.silva@email.com	São Paulo	2023-01-10
2	João Souza	joao.souza@email.com	Rio de Janeiro	2023-02-05
3	Ana Silva	ana.silva@email.com	São Paulo	2023-01-10
4	Maria Oliveira	maria.oliveira@email.com	Belo Horizonte	2023-03-15
5	João Souza	joao.souza@email.com	Rio de Janeiro	2023-02-05
6	João Souza	joao_souza@email.com	Rio de Janeiro	2023-02-05
7	João Souza	joão.souza@email.com	rio de janeiro	2023-02-05
8	Ana S.	ana.silva@email.com	São Paulo	2023-01-10
9	Maria Oliveira	mariaoliveira@email.com	Belo Horizonte	2023-03-15

Procedimentos:

UNICAMP

Padronização de strings (acentos, minúsculas, maiúsculas, abreviações)

ID	Nome	Nome_Padronizado
1	João da Silva	joao da silva
2	MARIA FERNANDA OLIVEIRA	maria fernanda oliveira
3	pedro l. santos	pedro l santos
4	Ana Paula	ana paula

Procedimentos:

Remoção de linhas duplicatas exatas (atenção para a PK)

ID	Nome	Cidade
1	João da Silva	São Paulo
2	Maria Oliveira	Rio de Janeiro
3	João da Silva	São Paulo
4	maria oliveira	Rio de Janeiro
5	Pedro Santos	Belo Horizonte
6	Pedro Santos	Belo Horizonte
7	João da Silva	São Paulo

Procedimentos:

Remoção de linhas duplicatas não exatas (com analise de similaridade)

A Similaridade de registros é realizada por algoritmos de analise de similaridade entre cada dado da informação.
Estes algoritmos utilizam o calculo de "distância" ou similaridade entre pares de dados (strings)

Exemplo:
Algoritmo de Levenshtein – quantidade de alterações para igualar duas strings.

String 1: rato - String 2: gato
Requer apenas 1 alteração de caractere em 4.
Similaridade de 75%

Exemplo

Passo 1

Registro 1:
- Nome: João Souza
- Email: joao.sousa@meumail.com
- Cidade: Rio de Janeiro
- Aniversário: 12/10/1997

Registro 2:
- Nome: joão souza
- Email: joaosouz@meumail.com
- Cidade: Rio do janeira
- Aniversário: 12/03/1998

Padronizar

Passo 2

Registro 1:
- Nome: joao souza
- Email: joao.sousa@meumail.com
- Cidade: rio de janeiro
- Aniversário: 12/10/1997

Registro 2:
- Nome: joao souza
- Email: joaosouz@meumail.com
- Cidade: rio do janeira
- Aniversário: 12/03/1998

UNICAMP

Passo 2

Registro 1:
- **Nome**: joao souza
- **Email**: joao.sousa@meumail.com
- **Cidade**: rio de janeiro
- Aniversário: 12/10/1997

Registro 2:
- **Nome**: joao souza
- **Email**: joaosouz@meumail.com
- **Cidade**: rio do janeira
- Aniversário: 12/03/1998

Calculo de Similaridade

Passo 3

- Nome: 100 %
- Email: 90,90 %
- Cidade: 85,71%
- Aniversario: 62,5 %

São similares ?

Limpeza de Dados

Valores Ausentes
"missing values"

8			4		6			7
						4		
	1					6	5	
5		9		3		7	8	
				7				
	4	8		2		1		3
	5	2					9	

Dados nulos podem ser considerado impeditivos para alcançar uma melhor qualidade da informação e dos bons resultados no processos transacionais ou analíticos.

Procedimentos:

Remoção de linhas com dados nulos.

Qual será a política de remoção de linhas ?

- Regra 1:

São determinadas colunas nulas que justificam a remoção da linha

e/ou

- Regra 2:

É a quantidade de colunas nulas que justificam a remoção da linha

Regra 1:
Colunas relevantes
Idade

Remover
as linhas

ID	Nome	Idade	Email	Telefone	Endereço
1	Ana	28	ana@email.com	(11)1111-1111	Rua das Flores
2	Bruno				
3	Camila	35	camila@email.com		Rua Azul, 123
4	Diego			(21)2222-3333	
5	Eduardo	42		(21)2222-2222	
6	Fernanda		fernanda@email.com		Rua Verde, 45
7	Gabriela	30			
8	Helena		helena@email.com	(31)3333-3333	Rua das Laranjeiras
9					
10	João	29			Rua Cinza, 456
11	Karla			(41)4444-4444	
12	Lucas	33	lucas@email.com		
13	Marina				Rua Amarela, 999
14					
15	Natália	40		(51)5555-5555	

Regra 2:
Quant. de colunas com Nulo
4 Nulos

Remover
as linhas

Procedimentos:

Remoção de colunas com dados nulos.

Qual será a política de remoção de colunas ?

- Regra 1:

São determinadas linhas que justificam a remoção da coluna

e/ou

- Regra 2:

É a quantidade de nulos na coluna que justificam a remoção da coluna

Remover as colunas

Remover as colunas

ID	Nome	Idade	Email	Telefone	Endereço
1	Ana	28	ana@email.com	(11)1111-1111	Rua das Flores
2	Bruno				
3	Camila	35	camila@email.com		Rua Azul, 123
4	Diego			(21)2222-3333	
5	Eduardo	42		(21)2222-2222	
6	Fernanda		fernanda@email.com		Rua Verde, 45
7	Gabriela	30			
8	Helena		helena@email.com	(31)3333-3333	Rua das Laranjeiras
9					
10	João	29			Rua Cinza, 456
11	Karla			(41)4444-4444	
12	Lucas	33	lucas@email.com		
13	Marina				Rua Amarela, 999
14					
15	Natália	40		(51)5555-5555	

Regra 1:
Linhas relevantes
ID = 8

Regra 2:
Quant. Nulos em coluna
> 2/3 de Nulos

UNICAMP

Procedimentos:

UNICAMP

Preenchimento de dados nulos

Qual será a política de preenchimento de colunas ?

- Regra 1:

Uso de valor médio, mediana, valor padrão ou outro.

ou

- Regra 2:

Uso de valor obtido de algum modelo preditivo baseado em outras colunas.

Regra 1:
Média da Altura

Regra 2:
Regressão Linear
Usando idade e peso

UNICAMP

Nome	Idade	Peso (kg)	Altura (cm)
Ana	25	68	165
Bruno	30	75	170
Carla	22	55	
Daniel	35	85	180
Elisa	28	62	

171,6

171,6

159.4

162.6

Altura = 98.45 + (0.25*Idade) + (0.85*Peso)

Limpeza de Dados

Correção

Correção de erros de digitação de dados

Exemplos:

Palavra com Erro	Correção	Observação
excreve	escreve	Troca de letras "x" por "s"
conhesimento	conhecimento	Erro de digitação + fonético
previlegio	privilégio	Troca de "e" por "i" e acento faltando
enchergar	enxergar	Erro comum de pronúncia
probléma	problema	Acento desnecessário
sesão	sessão	Confusão entre palavras homônimas
imformaçao	informação	Troca de "m" por "n" e acento faltando

Limpeza de Dados

Uniformização

Padronização de formatos dos dados

Procedimento:

- Conversão de tipos de dados adequados ao negócio
- Padronização de datas, horários, moedas, medidas (peso, dimensões etc)
- Redução ou aumento de tamanho de strings/textos
- Criação de domínios de valores para serem aplicados aos conjuntos de dados

Limpeza de Dados

Validação

Validação de consistências baseadas em regras naturais ou regras de negócio.

Regras:

1. **Preço negativo** (Regra: o preço do produto deve ser maior que zero)
2. **Quantidade em estoque negativa** (Regra: produtos não deve haver estoque negativo)
3. **Data de validade vencida** (Regra: produtos perecíveis não devem estar com validade vencida)
4. **Data de validade ausente para produto perecível** (Regra: produtos da categoria "Alimentos" devem ter data de validade)
5. **Categoria inconsistente com validade** (Regra: produtos que não perecem não devem ter validade- atribuir Nulo)

Exemplo:

ID	Nome do Produto	Categoria	Preço (R$)	Quantidade em Estoque	Data de Validade
1	Leite Integral	Alimentos	5.99	100	2024-04-10
2	Notebook XYZ	Eletrônicos	-2500.00	10	N/A
3	Maçã Gala	Alimentos	3.50	-20	2025-05-01
4	Shampoo Premium	Higiene Pessoal	15.00	50	2023-12-31
5	Cadeira de Escritório	Móveis	350.00	5	

Limpeza de Dados

Identificação de outliers

Reconhecimento de outliers para posterior realização de procedimentos de saneamento.

Categorias de Outliers

UNICAMP

- **Outliers Pontuais (Point Anomalies)**

Um único dados se desvia drasticamente dos outros.
Exemplo: Uma temperatura de 50°C em um local que costuma registrar entre 20°C e 30°C.

- **Outliers Contextuais (Contextual ou Conditional Anomalies)**

Um dado considerado anômalo em um **contexto específico**, como tempo, localização, ou condição.
Exemplo: Uma temperatura de 3°C no versão nordestino.

- **Outliers Coletivos (Collective Anomalies)**

Um conjunto de dados que, em grupo, é anômalo, mesmo que individualmente cada ponto pareça normal.
Exemplo: Uma sequência de batimentos cardíacos normais que, juntos, representam um padrão de fibrilação.

- **Outliers resultantes de Erros (Error Outliers)**

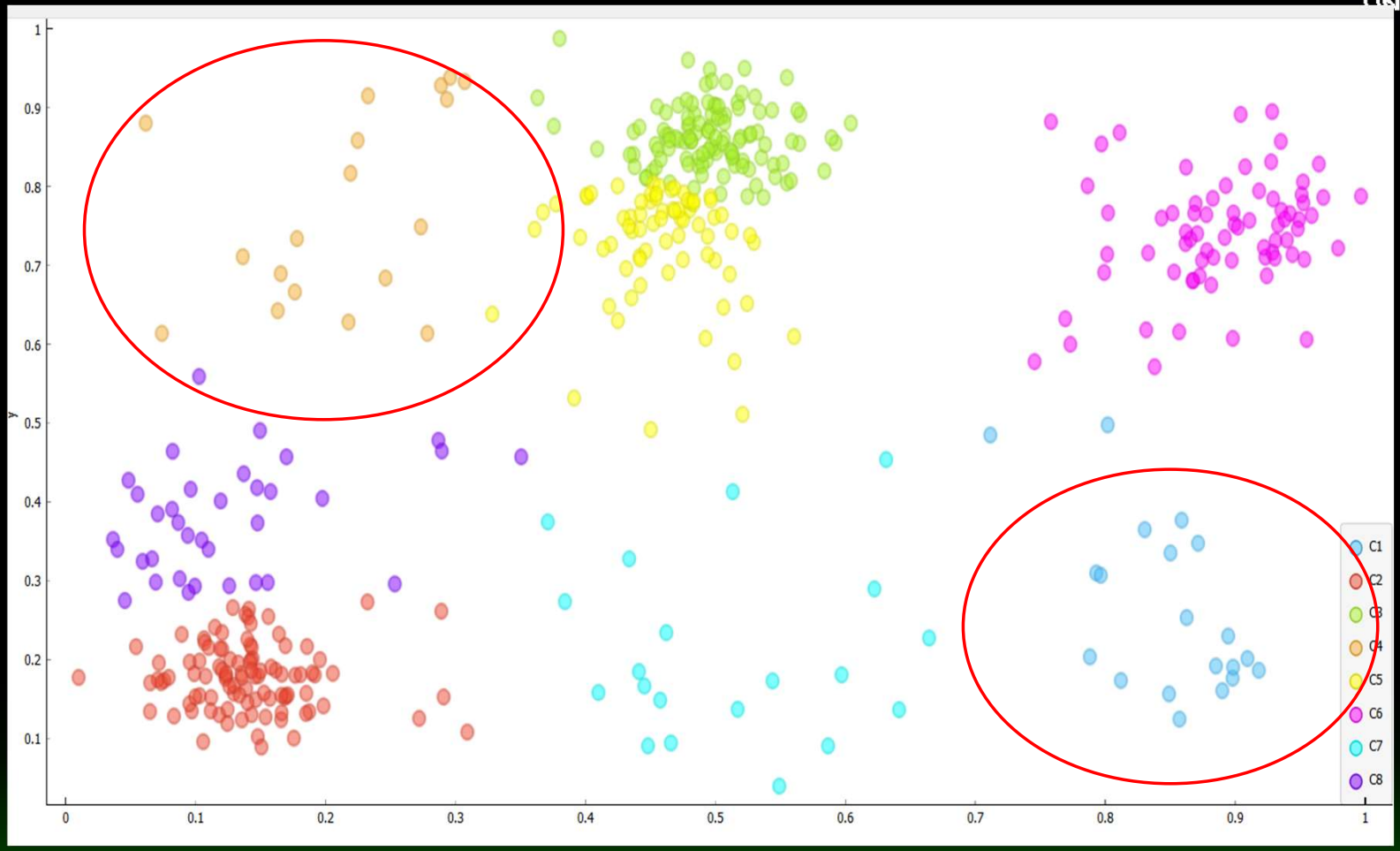
Um dado de **erros de medição, digitação, ou coleta de dados**.
Exemplo: Velocidade 400 km/h de um carro de passeio em rua de uma cidade

- **Outliers Naturais (Natural Outliers)**

Dado de valor incomum mas legítimos dentro de um conjunto.
Exemplo: Pessoas com altura superior a 2,20 metros

Algoritmos de Agrupamento

K-means
DBScan
etc



Técnica IQR (Interquartile Range)

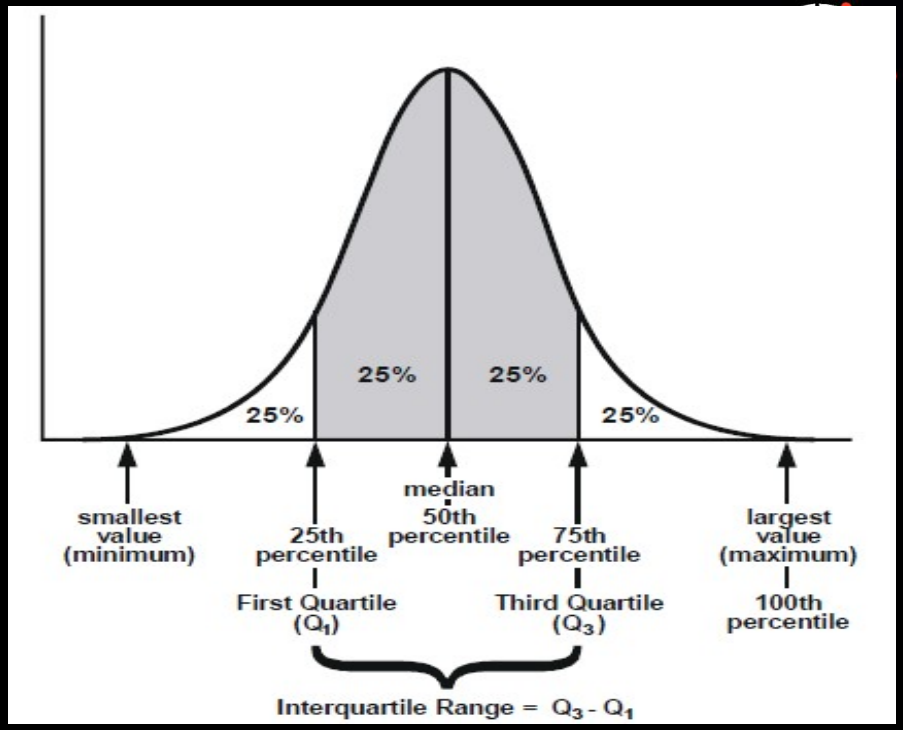
Técnica IQR (Intervalo entre Quartis) para detecção de outliers - por John Tukey,

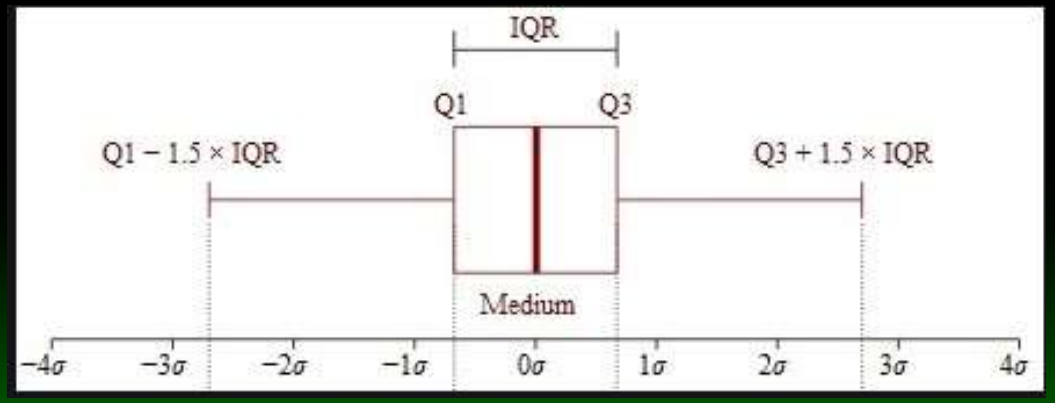
Quartis são dados que dividem o conjunto total de dados em quatros grupos de quantidades iguais, proporcionalmente 25% em cada grupo. A mediada é o dado que separa o 2º. Quartil do 3º. Quartil.

O IQR é definido como Q3 - Q1.

Qualquer dado que estiver acima de Q3+1.5*IQR ou abaixo Q1-1.5*IQR é considerado um outlier.

1. Ordenar os dados
2. Encontrar 1°. Quartil (Q1) e 3°. Quartil (Q3)
3. Calcular IQR
4. Estabelecer limites de outliers (LI e LS)
5. Reconhecer dados fora dos limites





IQR ≈ 1.35 × desvio padrão

Estudante	Nota
A	55
B	60
C	65
D	70
E	75
F	80
G	85
H	90
I	95
J	100
K	150

Exemplo:

1. Ordenar os dados
   - [55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 150]
2. Encontrar 1°. Quartil (Q1) e 3°. Quartil (Q3)
   - Q1 = 65
   - Q3 = 95
3. Calcular IQR
   - IQR = 95-65 = 30
4. Estabelecer limites de outliers (LI e LS)
   - LI = 65 – 1.5*30 = 20
   - LS = 95 + 1.5*30 = 140
5. Reconhecer dados fora dos limites
   - Outlier → 150

Limpeza de Dados

Descarte “Expurgo”

Descarte (eliminação definitiva ou não) de informações

Motivos:

- Informações que não atendem aos padrões de qualidade de informação definidos pela gestão.

- Informações que não podem ser saneadas utilizando os procedimentos selecionados

- Informações que perderam seu valor para o negócio com a passagem do tempo

Perguntas ...