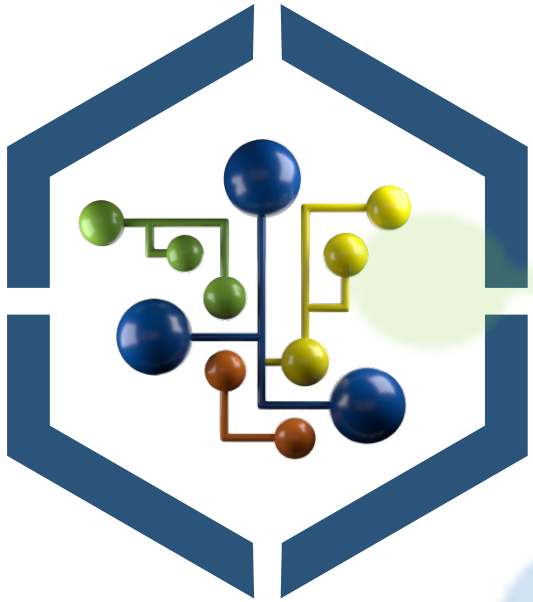


Apresentação

A Formação Analista de Dados é um programa de aperfeiçoamento profissional 100% online e 100% em português, com certificado de conclusão.

Seja Muito Bem-vindo(a)!

Formação Analista de Dados

SQL Para Data Science

Data Science Academy phelipe.utsempreboni@outlook.com 5c8a62005e4cde1acb8b45a3

Limpeza e Processamento de Dados com SQL

O Que Exatamente é a Limpeza de Dados?

A limpeza de dados é o processo de corrigir ou remover dados incorretos, corrompidos, formatados incorretamente, duplicados ou incompletos em um conjunto de dados.

Especialmente ao combinar várias fontes de dados, é possível que os dados sejam duplicados ou rotulados incorretamente. Se os dados estiverem incorretos, os resultados das análises não são confiáveis, embora possam parecer corretos (lembre-se do que vimos no Estudo de Caso 1).

Não há uma maneira absoluta de prescrever as etapas exatas no processo de limpeza de dados porque os processos variam de conjunto de dados para conjunto de dados. Mas é crucial estabelecer um modelo para o processo de limpeza de dados para que você saiba se está fazendo o trabalho da maneira certa.

Qual a Diferença Entre e Limpeza e Transformação de Dados?

A limpeza de dados é o processo que remove dados que não pertencem ao seu conjunto de dados.

A transformação de dados é o processo de conversão de dados de um formato ou estrutura em outro. Os processos de transformação também podem ser referidos como data wrangling, data munging, transformação ou processamento.

Muitas vezes, para limpar os dados teremos que fazer transformações. E para fazer algumas transformações pode ser necessário aplicar algum tipo de limpeza.

Data Science Academy phelipe.utsempreboni@outlook.com 5c8a62005e4cde1acb8b45a3

Principais Técnicas de Limpeza de Dados

1- Tratamento de Valores Ausentes
2- Detecção e Filtragem de Outliers (Valores Extremos)
3- Detecção e Remoção de Registros Duplicados
4- Correção de Erros Estruturais (Reorganização dos Dados)

Detectando e Tratando Valores Ausentes

Valores ausentes representam falta de informação e não falta de dados, como o termo pode sugerir.

Valores ausentes representam um problema e devem ser tratados.

Em qual das tabelas abaixo podemos identificar valores ausentes?

Idade	Peso (Kg)	Altura (Cm)
23	90	178
47		179

✓

Idade	Peso (Kg)	Altura (Cm)
23	90	178
47	?	179

✓

Idade	Peso (Kg)	Altura (Cm)
23	90	178
47	0	179

✓

Detecção e Tratamento de Outliers

Idade	Peso (Kg)
23	58
45	62
32	75
67	89
49	71
56	67
Média = 45 anos	Média = 70 Kg

Idade	Peso (Kg)
23	58
45	62
32	350
67	89
49	71
56	1
Média = 45 anos	Média = 105 Kg

Outlier

Outlier

Outlier pode ser erro de coleta de dados, erro da distribuição de dados ou mesmo valor válido.

Detecção e Remoção de Registros Duplicados

Idade	Peso (Kg)
23	58
23	58
32	75
32	75
49	71
56	67
56	67

Registros duplicados representam um problema e influenciam nos cálculos com SQL.

Normalmente removemos registros duplicados ou filtramos na query SQL antes de fazer cálculos.

Data Science Academy phelipe.utsempreboni@outlook.com 5c8a62005e4cde1acb8b45a3

Detecção e Remoção de Registros Duplicados

Detecção e Remoção de Registros Duplicados

ID	Idade	Peso (Kg)
1001	23	58
1002	23	58
1003	32	75
1004	32	75
1005	49	71
1006	56	67
1007	56	67

O registro duplicado depende da perspectiva em que analisamos os dados.

Colunas de ID ou chave primária podem ajudar a identificar se o registro está realmente duplicado.

Detecção e Remoção de Registros Duplicados

ID	Idade	Peso (Kg)
1001	23	58
1002	23	58
1003	32	75
1004	32	75
1005	49	71
1006	56	67
1007	56	67

Devemos observar cada linha completamente.

Se uma única coluna apresentar diferença, o registro pode não estar duplicado.

Data Science Academy phelipe.utsempreboni@outlook.com 5c8a62005e4cde1acb8b45a3

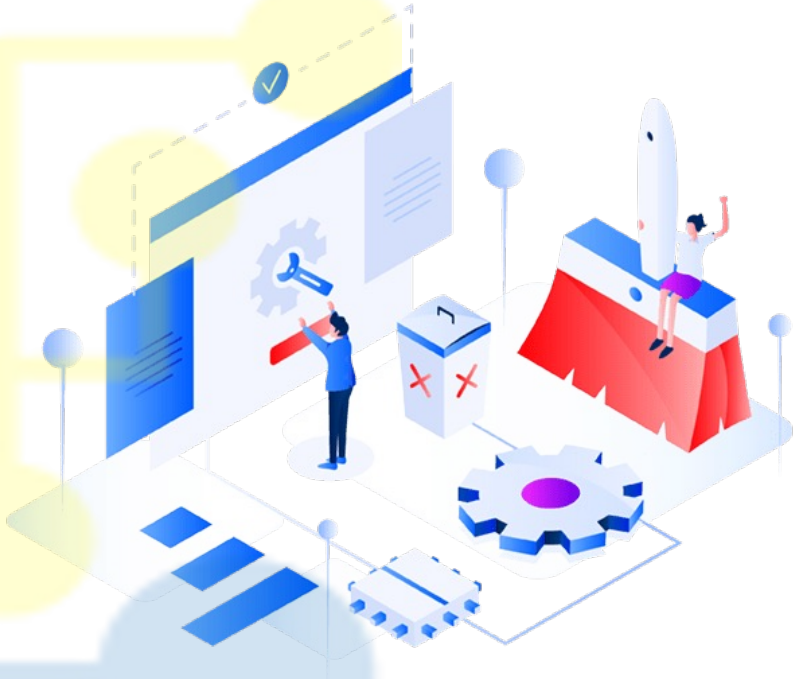
Reorganização dos Dados

	1001	1002	1003	1004	1005	1006	1007
Idade	23	32	45	67	78	19	28
Peso (Kg)	68	59	74	72	67	54	83

ID	Idade	Peso (Kg)
1001	23	68
1002	32	59
1003	45	74
1004	67	72
1005	78	67
1006	19	54
1007	28	83

Principais Técnicas de Processamento de Dados

1- Transformação de Atributos e Parsing
2- Mapeamento (de-para)
3- Filtragem, Agregação e Sumarização
4- Enriquecimento e Imputação



Transformação de Atributos e Parsing

Idade (String)	Idade (Int)
23	23
45	45
67	67

Idade	Faixa Etária
23	De 20 a 29 anos
45	De 40 a 49 anos
67	De 60 a 69 anos

Transformamos atributos (variáveis) para facilitar o processo de análise e corrigir imperfeições nos dados.

Transformação de Atributos e Parsing

Data	Data Formatada (Parsing)
01/01/2022 00:00:00	01/Jan/2022

No Parsing queremos formatar os dados de modo a facilitar o processo de análise ou mesmo permitir agregações com a coluna.

Mapeamento

Código de Erro do Equipamento	Descrição do Erro
1002	Erro na engrenagem do motor.
8931	Falha na lubrificação interna.
6510	Falha no funcionamento do mecanismo de parada.

O Mapeamento faz o **de-para**.

O objetivo é deixar os dados em um formato que facilite a interpretação e análise.

Filtragem, Agregação e Sumarização

Idade	Peso (Kg)
23	78
45	75
56	64
89	73

Ao filtrar os dados estamos processando os dados, pois somente parte dos registros ou das colunas serão retornados para o processo de análise.

Por exemplo filtrar dados de pacientes com mais de 70 Kg.

Filtragem, Agregação e Sumarização

Idade	Peso (Kg)
23	78
45	75
56	64
Média de Idade = 41	Média de Peso = 72

Na agregação estamos em busca de informação consolidada e mais gerenciável, que apresente características gerais sobre os dados.

Filtragem, Agregação e Sumarização

Estado	Cidade	Vendas (Unidades)
Rio de Janeiro	Rio de Janeiro	10
São Paulo	Campinas	14
Rio de Janeiro	Cabo Frio	15
São Paulo	Campinas	22

Total de Vendas Estado SP = 36
Total de Vendas Estado RJ = 25
Total de Vendas Cidade Rio de Janeiro = 10
Total de Vendas Cidade Cabo Frio = 15
Total de Vendas Cidade Campinas = 36
Total Geral = 61

Na sumarização estamos em busca de totais e subtotais que sumarizem (resumem) os dados.

Usamos agregação para fazer sumarização.

Enriquecimento e Imputação

Código do Paciente	Idade	Peso (Kg)
10CE23-900	23	?
11RS45-800	45	75
12SP78-600	?	64
13RJ89-500	89	73

Código do Paciente	Idade	Peso (Kg)	Estado
10CE23-900	23	71	CE
11RS45-800	45	75	RS
12SP58-600	58	64	SP
13RJ89-500	89	73	RJ

No enriquecimento estamos buscando formas de extrair o máximo de informação dos dados. Na imputação preenchemos valores ausentes visando a completude dos dados.

Muito Obrigado!

Continue Trilhando uma Excelente Jornada de Aprendizagem!