Preparação e Análise de Dados

Prof. Marco Cristo
Instituto de Computação

Agenda

- Preparação (limpeza, filtragem, organização e engenharia)
  - Organização
  - Pré-processamento
  - Entendendo variáveis e estatísticas descritivas
  - Entendendo a matriz de dados
  - Outliers, anomalias e visualização

Introdução

- Porquê?
  - Garbage in, Garbage out
  - 60 a 80% do esforço total em MD
- O quê é necessário?
  - Entender os dados
  - A história dos dados
  - Conhecimento de domínio
  - Seus objetivos!

Introdução

- Entradas
  - Dados puros
- Saídas
  - Treino (quem sabe, Treino e Validação) e Teste (se der)
    - Correto
    - Completo
    - Consistente
    - Atualizado
    - Confiável
    - Interpretável

Dados

- O que é?
  - Exemplos, observações, medidas, eventos, etc.
  - Estruturado, semi ou não
  - Conjunto de atributos
- Como está organizado
  - Relação = conjunto de exemplos
  - Exemplo específico = instância
  - Instância = conjunto de atributos

Organização

- Tabelas → Relação
  - Desnormalização
    - Coloque relações nas instâncias
    - Duplicidade não é um problema em MD
  - Problemas de integração
    - Mesmo atributo com nomes diferentes em diferentes bancos
    - Valores podem ser conflitantes (ex: idade é numérica em um departamento e nominal em outro)
    - Deduplicação de valores

Pré-processamento

- Operações (dados reais são incompletos, inconsistentes, ruidosos, etc)
  - Limpar os dados
  - Lidar com falta de dados
  - Explorar características de variáveis
  - Mudar a representação (normalização, discretização, transformação)

Variáveis e Estatísticas

- Conheça os tipos das variáveis
  - Simbólicas
    - Nominais (ex: cor = {preto, branco, vermelhor})
    - Ordinais (ex: idade = {crianca, adulto, idoso})
  - Numéricas
    - Intervalos (ex: datas)
    - Reais (ex: numeros quaisquer)
- Conheça suas propriedades estatísticas
  - Nunca variam? São totalmente aleatórias? (min, max, avg, etc)
- Conheça as relações/estruturas
  - Dependentes?
  - Esparsas (falta, N/A, 0?)
  - Como variam? Sem limites? (ex: datas) Ineditismo ocorre no teste? (ex: séries de tempo)
- Limpe, complete, corrija erros
- Melhore ou re-represente

Limpe

- Ruído
  - Erro de fonte desconhecida (em geral, assumimos Normal)
    - Procure dados suspeitos e remova
    - Suavize por representar como a média dos vizinhos ou outra técnica
  - Valores incorretos
    - Instrumentos que falham, pessoas que erram no cadastro, problemas ao transmitir, limitações tecnológicas, inconsistências em convenções, duplicação, dados não informados, etc
    - Aplique conhecimento do domínio e engenharia reversa
      - ex: erros gramaticais

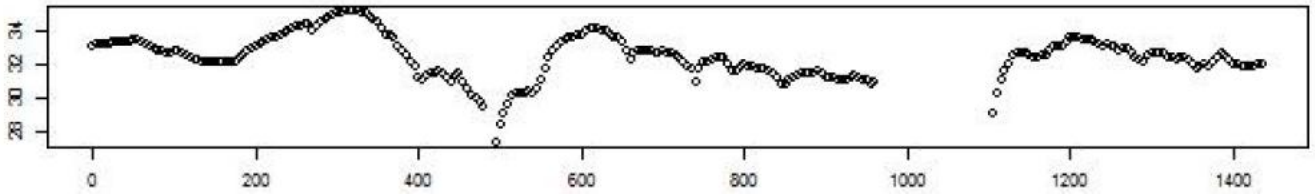
Complete

- Dados faltantes
  - Entenda quanto falta, se os dados que faltam são ou não aleatórios, são dependentes, ocorrem na mesma instância, correlacionam com o alvo
  - Se aleatório...
    - Remova instâncias com dados ausentes
    - Remova atributos se faltam em muitas instâncias
  - Use constante global se fizer sentido (ex: desconhecido)
  - Tente completar o que falta :-o
    - Interpolação
    - Modelos mais sofisticados, baseados em outros atributos ou no que se sabe sobre atributo alvo

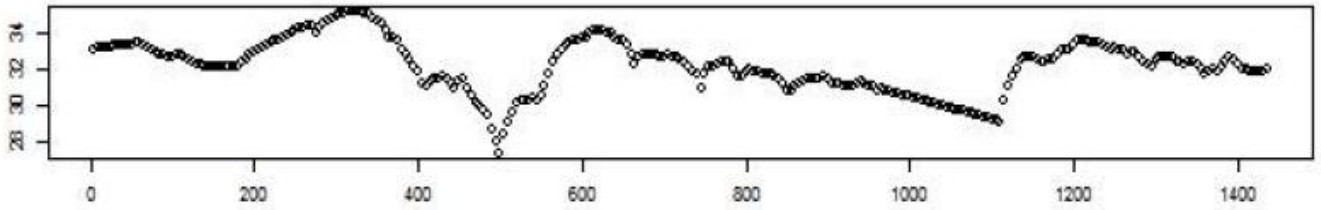
Variáveis e Estatísticas

- Dados faltantes – ex: nível de glucose em 24 hrs

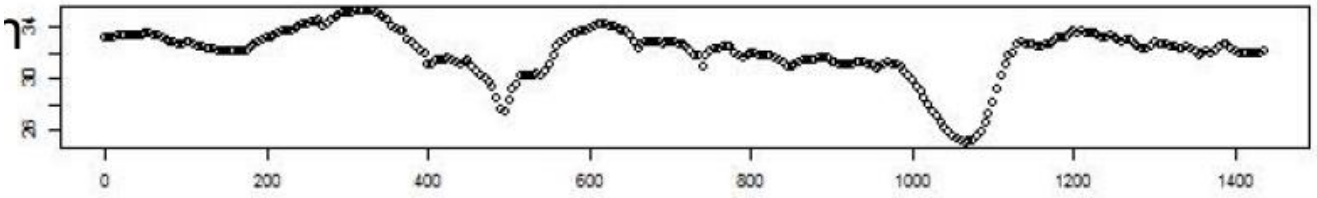
original



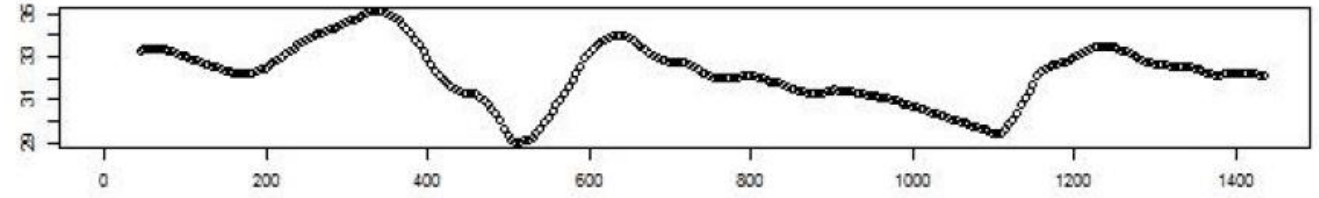
Interpolação linear



Interpolação polinomial



Interpolação Polinomial + Suavização em janela



Variáveis e Estatísticas

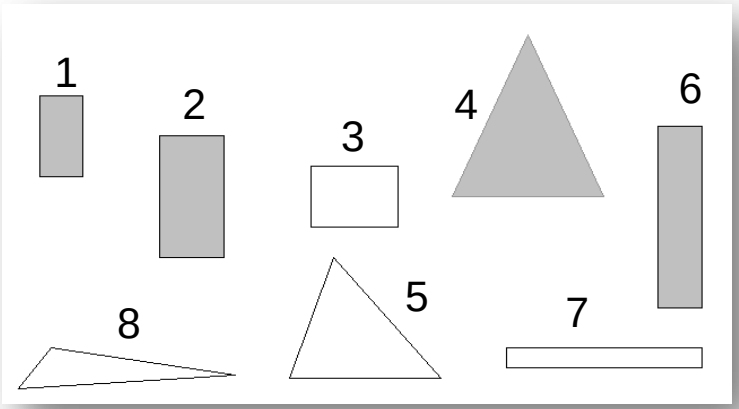
- Dados faltantes
  - Exemplo
    - Abra bank_missing.arff
    - Que campos tem valores faltantes?
    - Para cada campo com valores faltantes, anote valor máximo e média
    - Preencha com ReplaceMissingValues

Transforme

- Suavize, agregue, normalize, crie!
- Melhore suas variáveis (variáveis mais ricas facilitam aprendizagem)
  - Se você sabe como deduzir um atributo importante, deduza
    - Transforme CEP em latitude x longitude, renda per capita, distancia de ponto de referencia, etc
  - Inclua informação relevante do domínio
    - Quer prever placar de jogo de futebol?
      - Informe desempenho de longo prazo
      - Informe desempenho de curto prazo
        - Quando atuando como time da casa
        - Quando atuando como visitante

Transforme

- Exercício
  - Que relação deve ser aprendida?
  - Que campo facilitaria seu aprendizado?



	Width	Height	Sides	Class
1	2	4	4	Standing
2	3	6	4	Standing
3	4	3	4	Lying
4	7	8	3	Standing
5	7	6	3	Lying
6	2	9	4	Standing
7	9	1	4	Lying
8	10	2	3	Lying

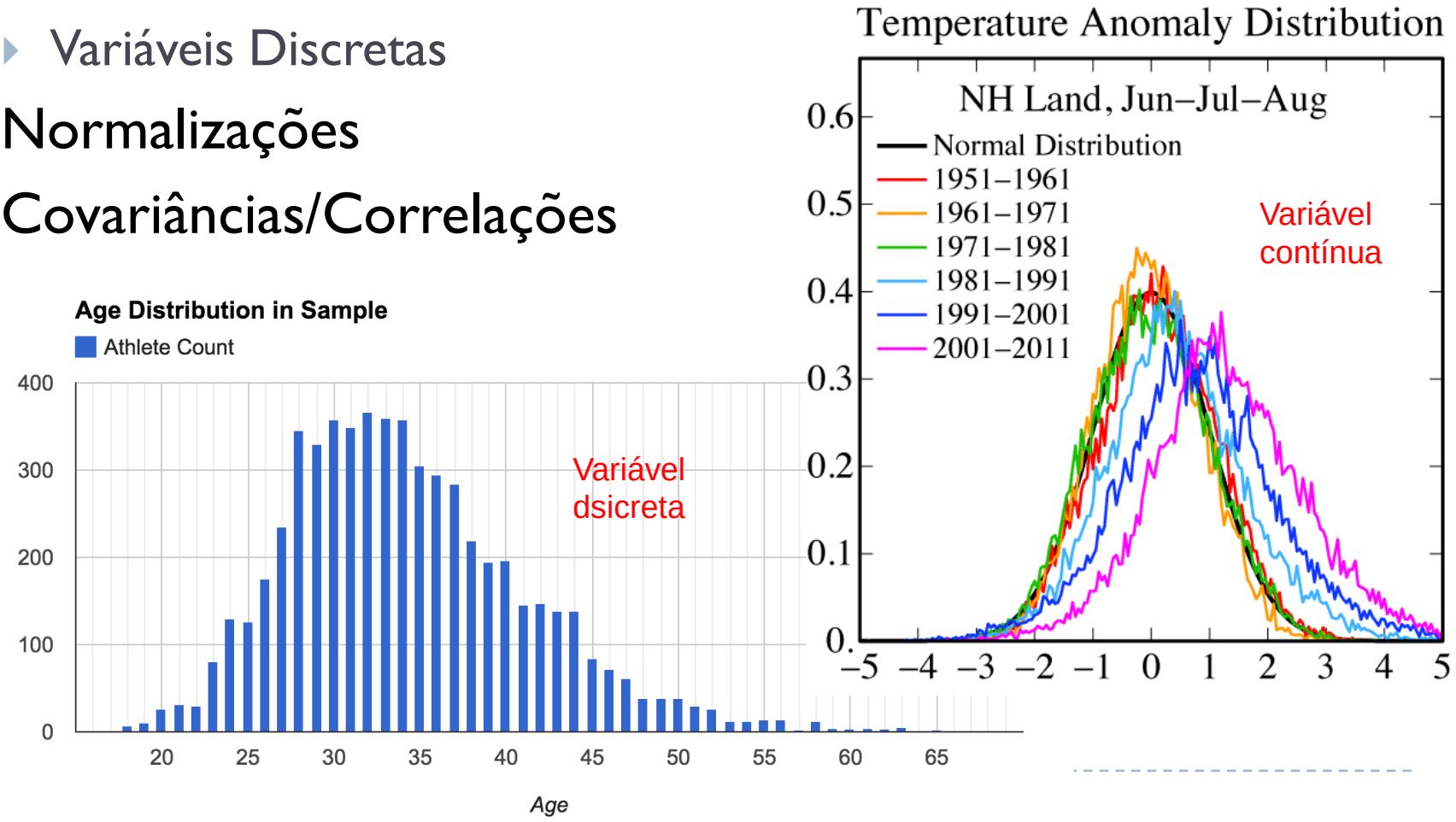
Transforme

- Discretize
  - Particionamento por largura igual
    - N intervalos de mesmo tamanho
  - Particionamento por frequência igual
    - N intervalos de mesma frequência
  - Exemplo
    - Atributo "Idade" é numérico
    - Como transformar em categórico?

Paciente	Idade	Idade Largura =	Idade Freq =
Joao	8	0-20	< 18
Ana	10	0-20	< 18
Pedro	12	0-20	< 18
Carlos	13	0-20	< 18
André	18	0-20	18 a 29
Helena	25	20-40	18 a 29
Paulo	26	20-40	18 a 29
Alessandra	29	20-40	18 a 29
Hélio	30	20-40	>= 30
Milton	34	20-40	>= 30
Saulo	55	40-60	>= 30
Valéria	60	40-60	>= 30

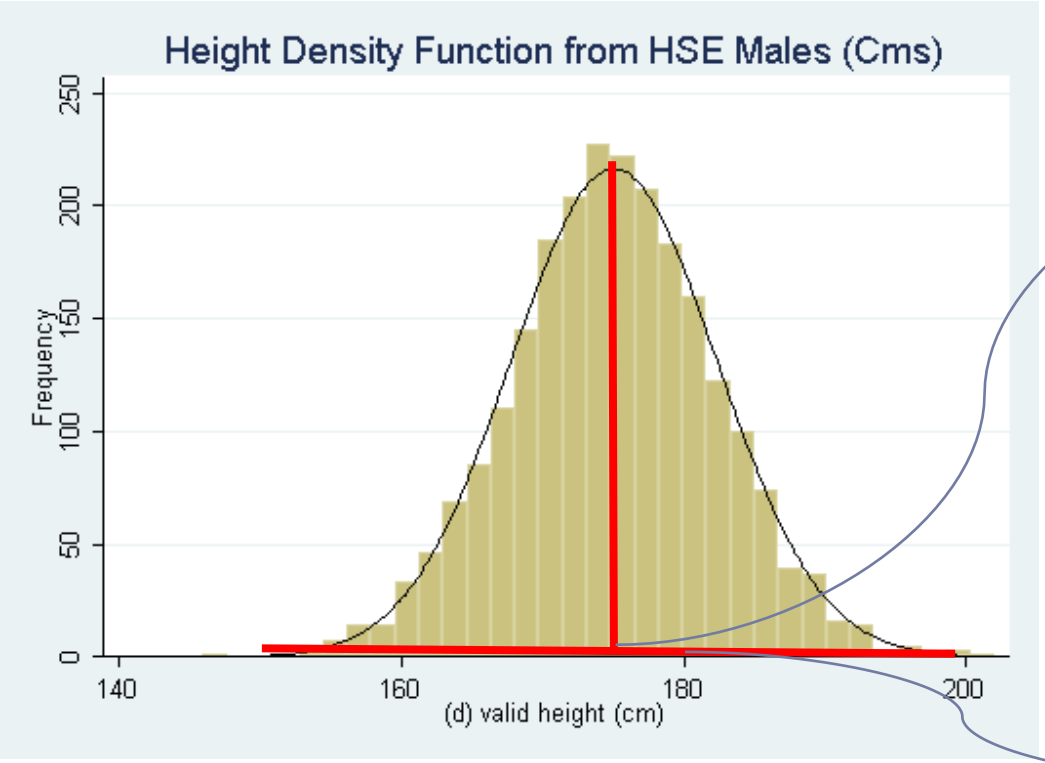
Transforme (usando estatística)

- Distribuições e Histogramas, Médias, Desvio, etc
  - Variáveis Contínuas
  - Variáveis Discretas
- Normalizações
- Covariâncias/Correlações



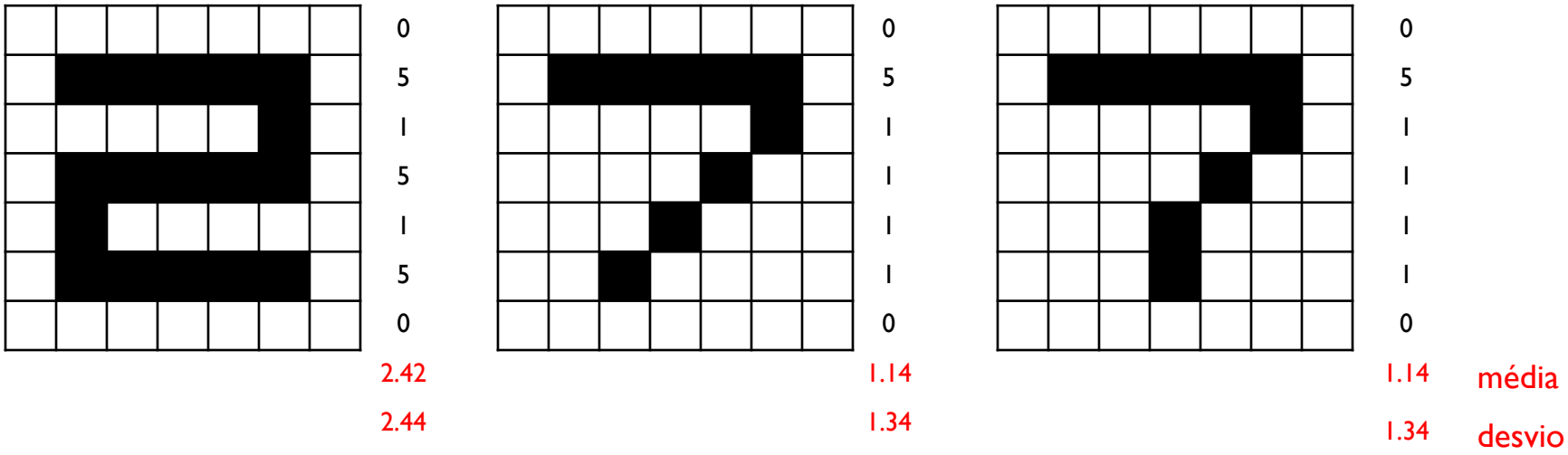
Transforme (usando estatística)

- Distribuição Normal (ou Gaussiana)
  - Muitos fenômenos seguem essa distribuição
  - "Curva de sino" descrita por média e variância



Transforme (usando estatística)

- Distribuições e Histogramas, Médias, Desvio, etc
  - Variáveis Contínuas
  - Variáveis Discretas
- Normalizações
- Covariâncias/Correlações



Transforme (usando estatística)

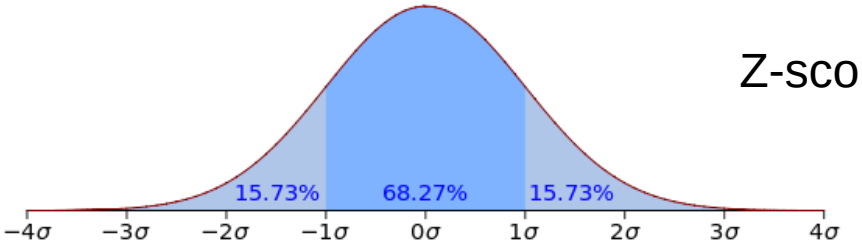
- Distribuições e Histogramas, Médias, Desvio, etc
  - Variáveis Contínuas
  - Variáveis Discretas
- Normalizações
- Covariâncias/Correlações

média $x_{new} = x - \text{mean}(x)$

0-a-1 $x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$

log $x_{new} = \log(x)$

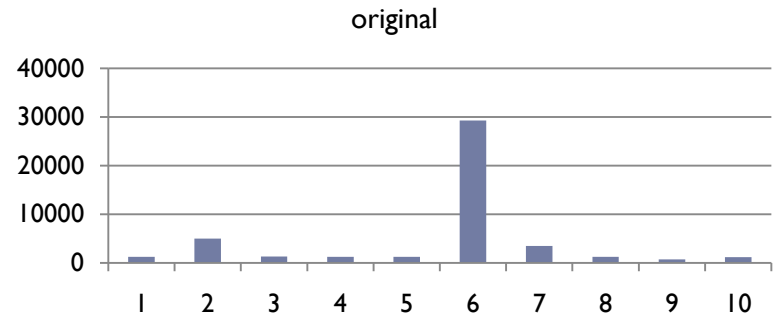
Z-score $z - score = \frac{x - \text{mean}(x)}{\text{std}(x)}$

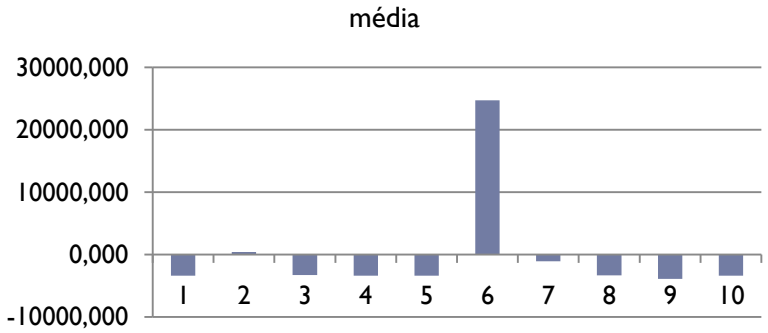


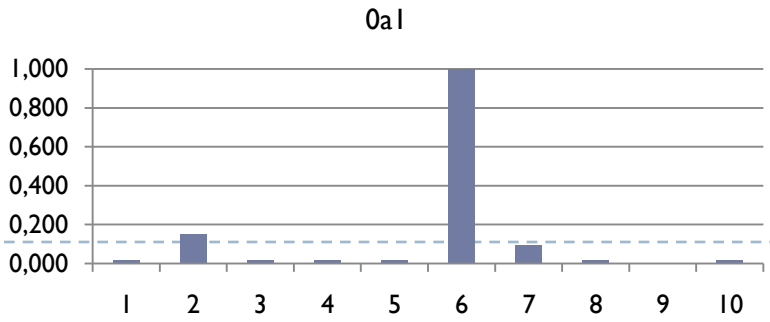
Transforme (usando estatística)

- Distribuições e Histogramas, Médias, Desvio, etc
  - Variáveis Contínuas
  - Variáveis Discretas
- Normalizações
- Covariâncias/Correlações

#	Salários	Média	Z-score	0a1	log
1	1230	-3366,400	-0,383	0,018	3,090
2	5000	403,600	0,046	0,150	3,699
3	1300	-3296,400	-0,375	0,020	3,114
4	1230	-3366,400	-0,383	0,018	3,090
5	1230	-3366,400	-0,383	0,018	3,090
6	29300	24703,600	2,813	1,000	4,467
7	3500	-1096,400	-0,125	0,097	3,544
8	1250	-3346,400	-0,381	0,018	3,097
9	724	-3872,400	-0,441	0,000	2,860
10	1200	-3396,400	-0,387	0,017	3,079
média	4596,400				
desvio	8782,726				
minimo	724,000				
maximo	29300,000				



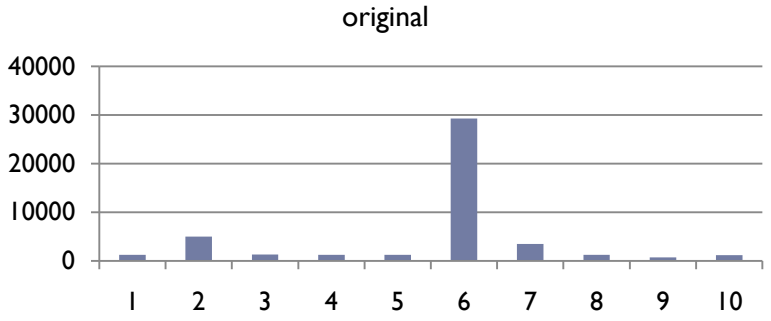


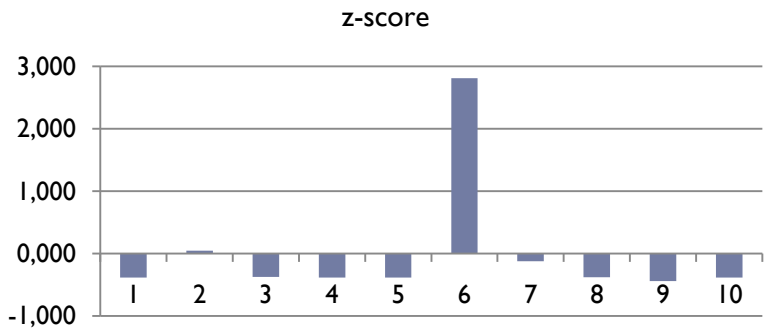


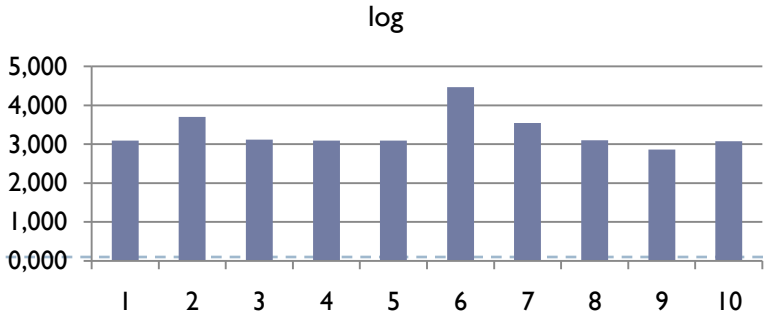
Transforme (usando estatística)

- Distribuições e Histogramas, Médias, Desvio, etc
  - Variáveis Contínuas
  - Variáveis Discretas
- Normalizações
- Covariâncias/Correlações

#	Salários	Média	Z-score	0a1	log
1	1230	-3366,400	-0,383	0,018	3,090
2	5000	403,600	0,046	0,150	3,699
3	1300	-3296,400	-0,375	0,020	3,114
4	1230	-3366,400	-0,383	0,018	3,090
5	1230	-3366,400	-0,383	0,018	3,090
6	29300	24703,600	2,813	1,000	4,467
7	3500	-1096,400	-0,125	0,097	3,544
8	1250	-3346,400	-0,381	0,018	3,097
9	724	-3872,400	-0,441	0,000	2,860
10	1200	-3396,400	-0,387	0,017	3,079
média	4596,400				
desvio	8782,726				
minimo	724,000				
maximo	29300,000				







Transforme (usando estatística)

- Distribuições e Histogramas, Médias, Desvio, etc
  - Variáveis Contínuas
  - Variáveis Discretas
- Normalizações
- Covariâncias/Correlações
  - Ausência de correlação não implica necessariamente em independência
    - Média
    - Desvio
    - Variância
    - Covariância

$$\bar{x} = \mu = \frac{1}{n}\sum_i x_i$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})}$$

supondo média 0 — notação vetorial

$$\sigma^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n-1}\sum_i (x_i)(x_i) = \eta\, \mathbf{x}^T \mathbf{x}$$

$$Cov = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_i (x_i)(y_i) = \eta\, \mathbf{x}^T \mathbf{y}$$

Transforme (usando estatística)

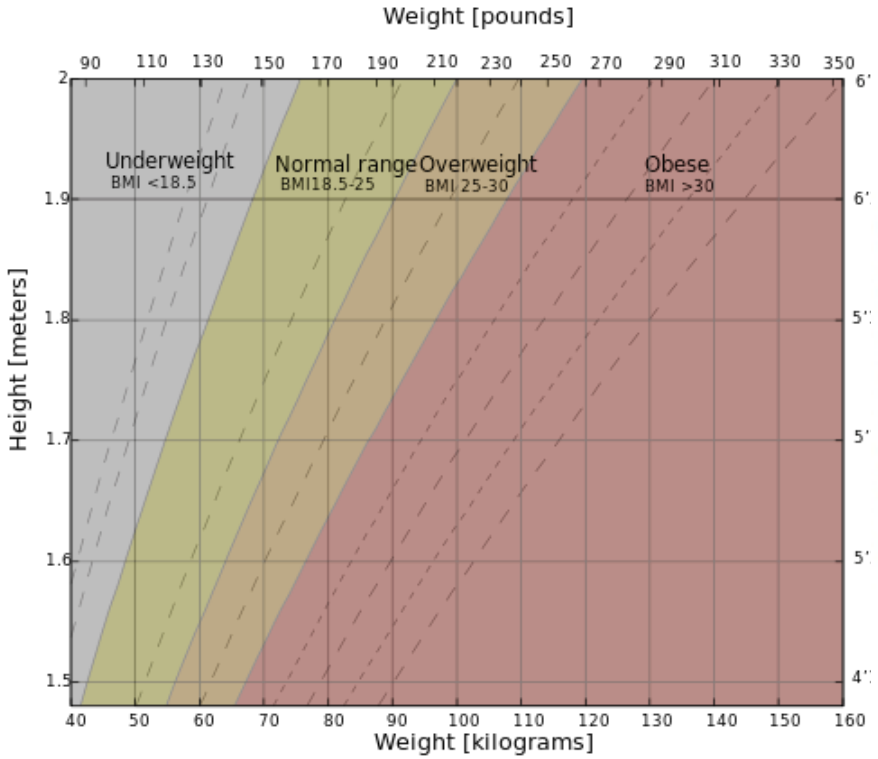
- Exemplos
  - Abra houses.arff
  - Analise pesos com e sem padronização
  - Abra mnist_sample.arff
  - Qual o efeito de binarização pro NaiveBayes?

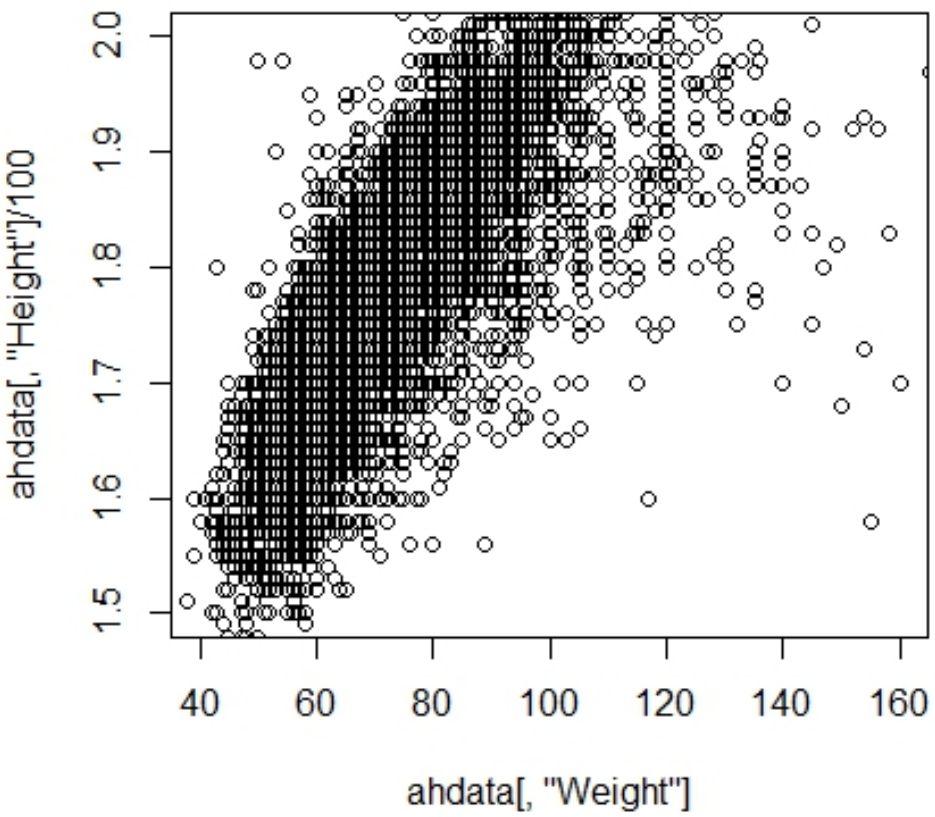
Estatísticas sobre a matriz completa

- Base AHW_1.csv (Londres 2012)
  - Veja histogramas
    - Para cada atributo (sem classe)
    - Peso é dependente do sexo? (considere classe = sexo)
  - Veja correlações
  - Adicione novas variáveis
    - BMI = peso (Kg) * altura (m)^2 → peso * (altura / 100)^2
    - Atletas são obesos?

Exemplos com Weka

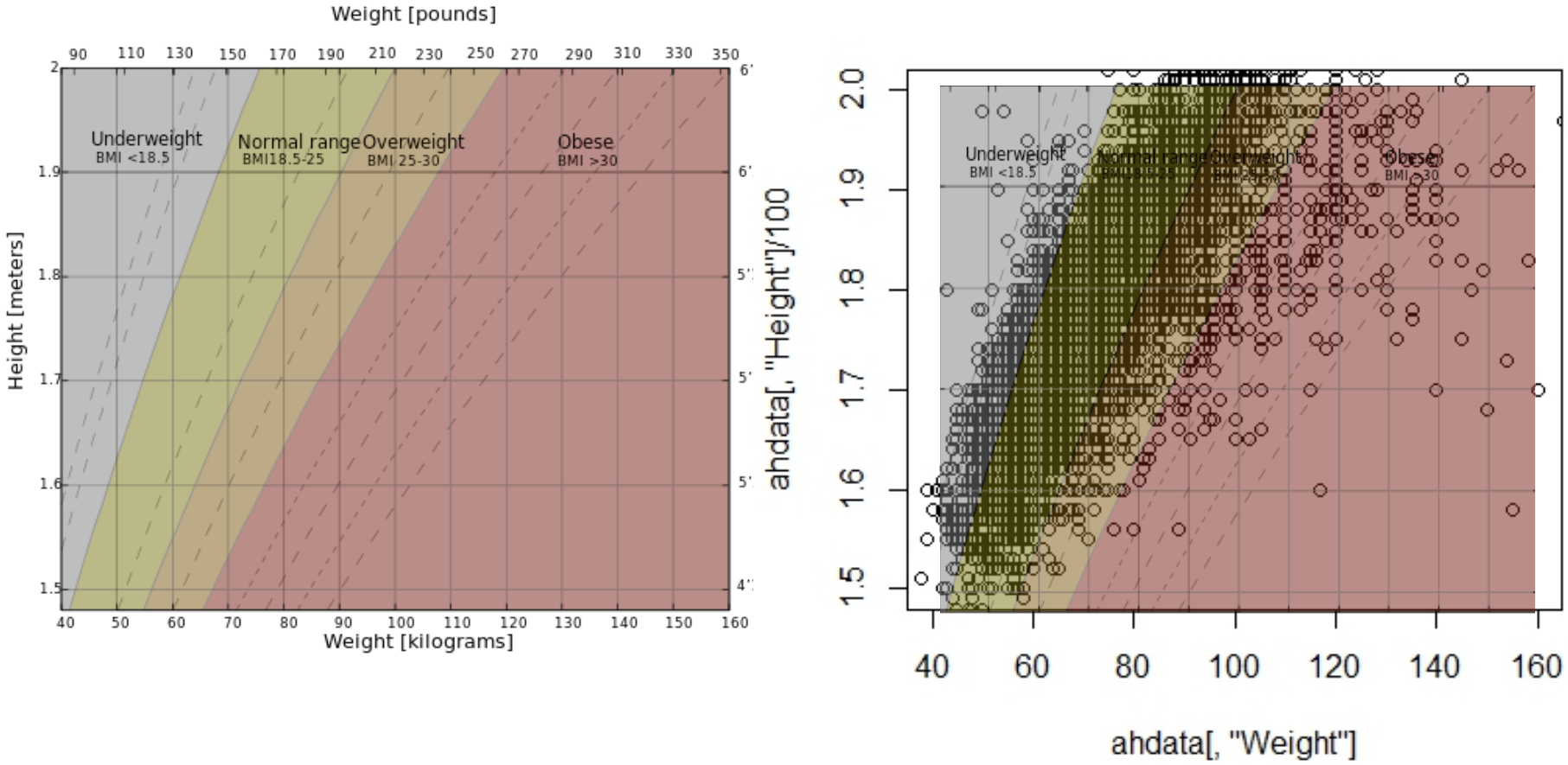
- Atletas são obesos?





Exemplos com Weka

- Atletas são obesos?

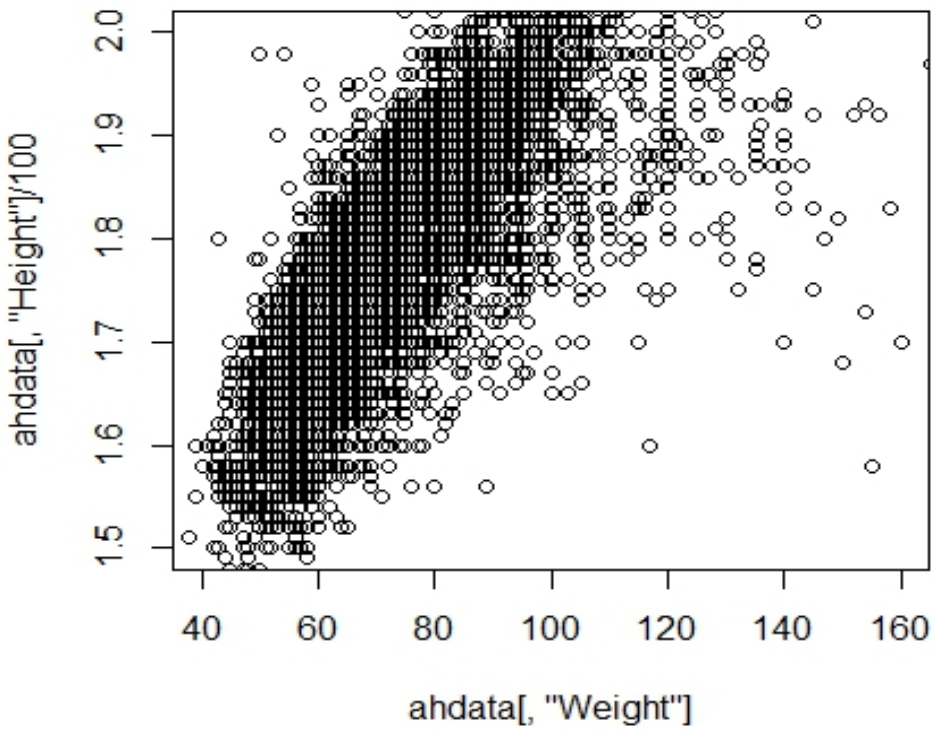


Estatísticas sobre a matriz completa

- Base AHW_1.csv (Londres 2012)
  - Transformação
    - Classifique pelo total de medalhas (numérico -> nominal)
  - Caracterize dados ausentes
    - Atletas com peso ausente foram tomados de forma aleatória (por exemplo, em relação à altura)?
    - Elimine dados faltantes

Anomalias

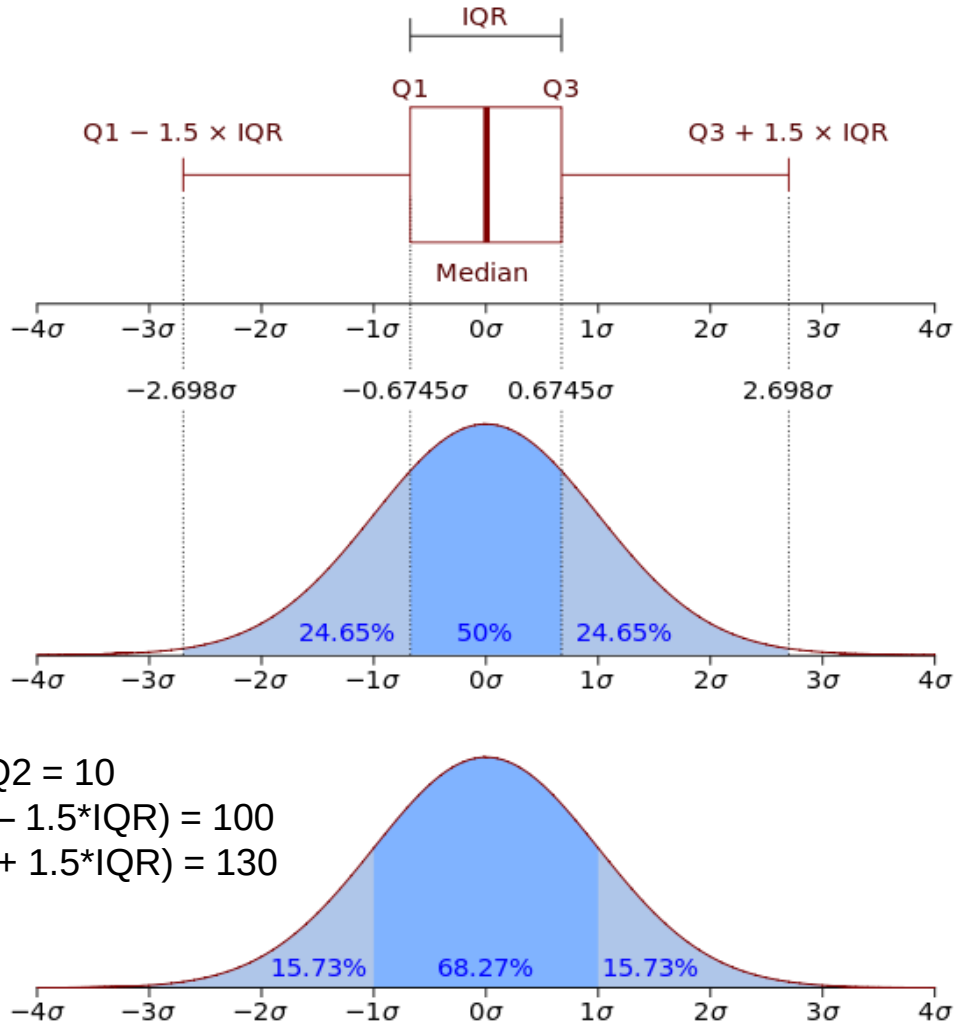
- O que é?
  - Outlier – longe da média
  - O vizinho mais distante
  - Quem produz o maior erro no modelo



Anomalias

- O que é?
  - Outlier – Amplitude Interquartil

i	X	Quartil
1	102	
2	104	
3	105	Q1
4	107	
5	108	
6	109	Q2 mediana
7	110	
8	112	
9	115	Q3
10	116	
11	118	



IQR = Q1-Q2 = 10
Mini = (Q1 – 1.5*IQR) = 100
Max = (Q1 + 1.5*IQR) = 130

Anomalias

- O que é?
  - Outlier – longe da média
  - O vizinho mais distante
  - Quem produz o maior erro no modelo

- Exemplo
  - Abra diabetes.arff
  - Qual o efeito de outliers e valores extremos para o J48 (com CV10) nesse exemplo em particular?

Muitas variáveis

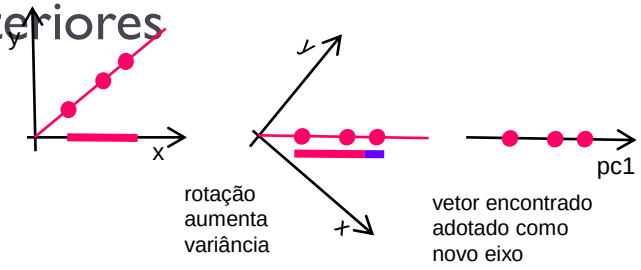
- Mais variáveis → Mais Informação (→ Mais ruído :/)
  - Ruído na variável → seleção
  - Ruído difuso → redução
- Como fazer?
  - Seleção: retire colunas
    - Grande escala, ruído caracteristicamente na variável!
    - Mais sobre isso no futuro...
  - Redução de Dimensão: reduza dimensões sem tirar colunas específicas
    - Grande escala, ruído difuso!
    - Álgebra à vista ;)

Muitas variáveis

- Redução de dimensionalidade
  - Dado N pontos e P atributos, dados podem ser representados com k < P atributos se
    - Atributos são constantes
    - Atributos são redundantes
    - Atributos só contribuem pra ruído (portanto, os que contribuem para variação dos dados são interessantes!)

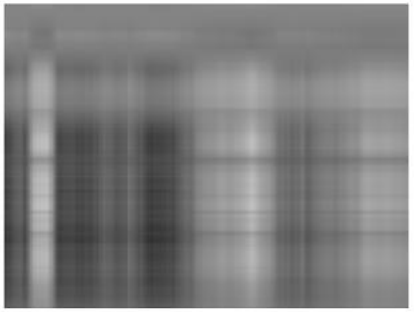
Muitas variáveis

- Principal Component Analysis (PCA)
  - Ache k atributos alternativos (fatores) que podem ser usados para representar os dados
  - Primeiro componente é o vetor que maximiza a variância dos dados projetados nele
  - K-ésimo componente é o k-ésimo vetor de maior variância, ortogonal a todos os vetores anteriores



- Observações
  - Quantos K? Tantos quanto necessário para melhorar resultado ou até alcançar 95% da variância total
  - PCA em matriz quadrada = SVD!!!
    - SVD mais geral → Fatoração Matricial. Muito usado em Recomendação
  - Útil para dados numéricos apenas
  - Para muitas dimensões, transformação útil para visualização em 2D

PCA: aplicação em compressão de imagem



(a) 1 principal component

(b) 5 principal component

(c) 9 principal component

(d) 13 principal component

(e) 17 principal component

(f) 21 principal component

(g) 25 principal component

(h) 29 principal component

512x512

Conclusão

- Estude seu problema
- Garanta que o modelo não será alimentado com lixo
- É sempre bom saber usar ferramentas que facilitem a sua vida!