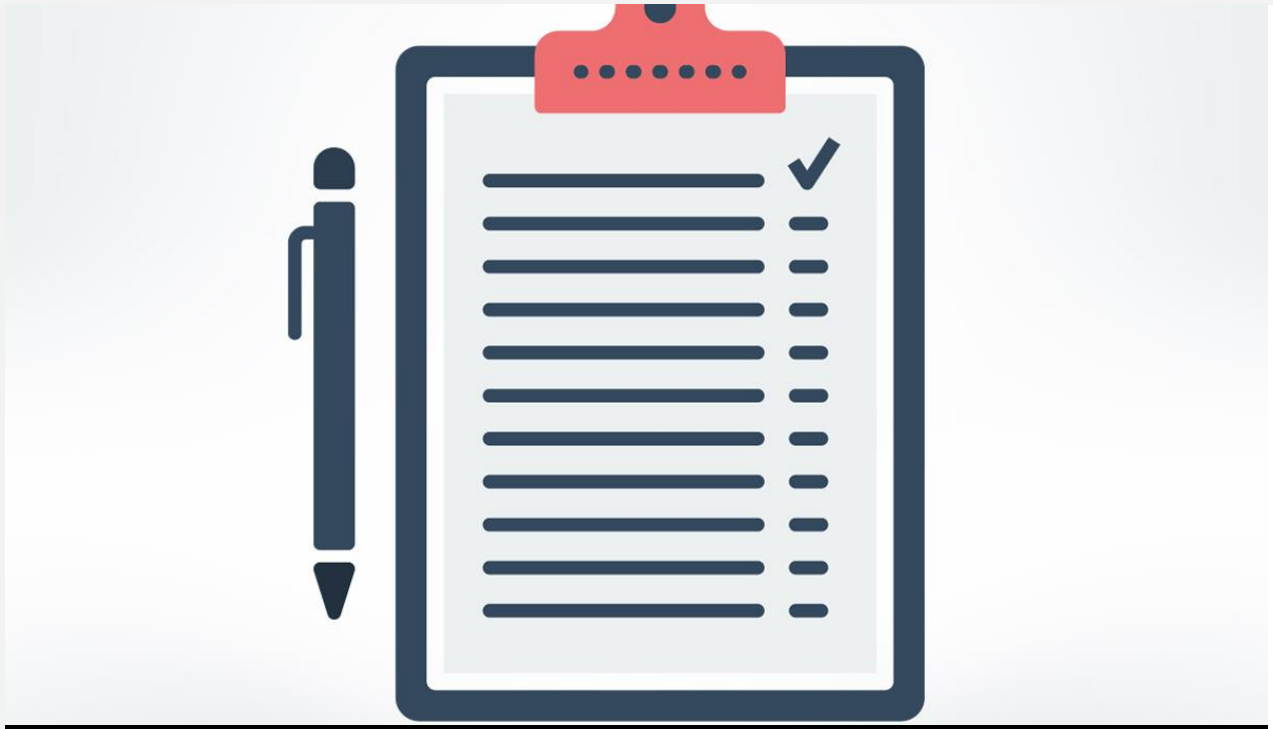
Pré-Processamento de Dados

Eduardo Silva – easilva91@gmail.com



Agenda

Pré-Processamento de Dados

1. O que são dados?
2. Tipos de dados
3. Tipos de atributos
4. Tipos de conjuntos de dados
    I. Data matrix
    II. Dados de transação
    III. Grafos
    IV. Dados ordenados

5 - Porque fazer o pré-processamento?

6 - Limpeza dos dados
    I. Lidando com valores ausentes
    II. Lidando com outliers
    III. Dados ruidosos

7 - Transformação dos dados
    I. Padronização
    II. Normalização
    III. Transformações para atingir a Normal
    IV. Encaixotamento (binning)

8 - Remoção de variáveis

9 - Redução dos dados
    I. Redução de dimensionalidade PCA/ACP
        I. como definir o número de componentes?

O Que são Dados?

Uma coleção de objetos de dados e seus atributos.

- Atributos
  - É uma propriedade ou característica de um objeto
  - Também conhecido como variável, campo, característica.
  - Uma coleção de atributos descreve um objeto.
- Objetos também são conhecidos como registro, ponto, caso, amostra, entidade ou instância

Atributo

Objeto

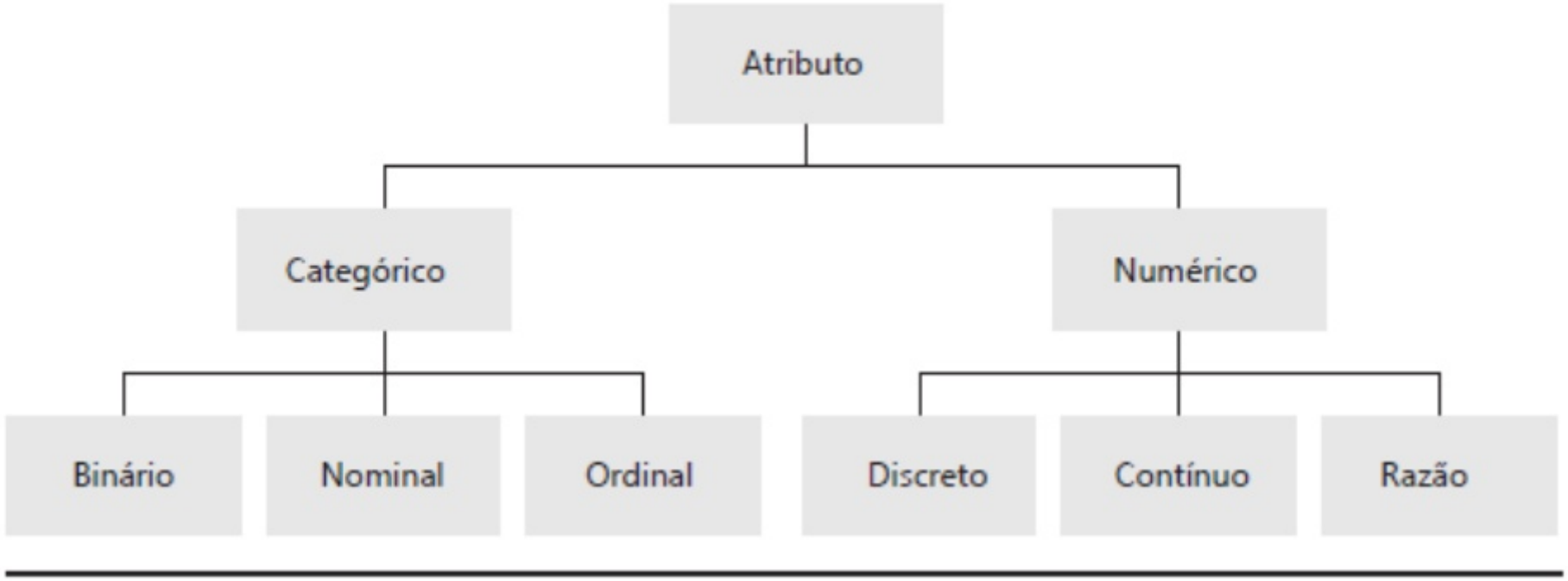
Altura	Peso	Sexo	Idade	Salário	Atividade Física
1,60	79	M	41	3000	S
1,72	82	M	32	4000	S

Tipos de Dados

Tipo de Dados	Descrição
Estruturados	Os dados residem em campos fixos em um arquivo – por exemplo, uma tabela, uma planilha ou um banco de dados. Dependem da criação de um modelo de dados, o que inclui definir quais campos de dados serão utilizados (por exemplo, nome, idade, nível educacional, estado civil, gênero, etc.)
Semiestruturados	É um tipo de dado que não possui a estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Geralmente são usados marcadores (por exemplo, tags) para identificar certos elementos dos dados, mas a estrutura não é rígida. Exemplos conhecidos de dados semiestruturados são arquivos XML ou e-mails.
Não estruturados	É aquele que não possui um modelo de dados, que não está organizado de uma maneira pré-definida ou que não reside em locais definidos. Normalmente se refere a textos livres, imagens, vídeos, sons, páginas web, arquivos PDF, entre outros.

Tipos de Atributos

Figura 2.2 Tipos de atributos

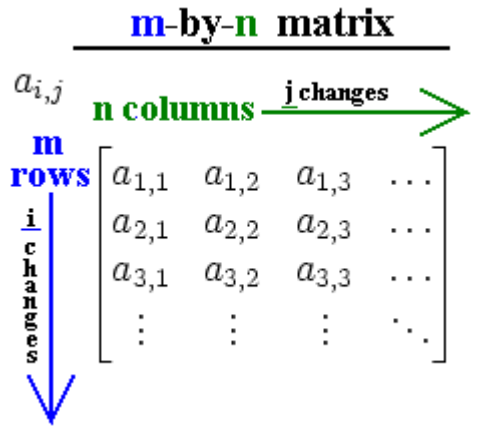


Tipos de Conjuntos de Dados

- Registro
  - Data Matrix
  - Distance Matrix
  - Documentos
  - Dados de Transação
- Grafos
  - Web/Webometrics
  - Estrutura Moleculares
- Ordenados (os dados estão ordenados de alguma forma)
  - Dados Espaciais
  - Dados Temporais
  - Dados de Sequência Genética

Data Matrix

- Se os dados tiverem o mesmo conjunto fixo de atributos numéricos, estes dados podem ser pensados como pontos em um espaço multidimensional, onde cada dimensão representa um atributo distinto.
- Esse conjunto de dados pode ser representado por uma matriz *m* por *n*, onde existem *m* linhas, uma para cada objeto e *n* colunas, uma para cada atributo



Altura	Peso	Sexo	Idade	Salário	Atividade Física
1,60	79	M	41	3000	S
1,72	82	M	32	4000	S

Exemplo de uma Data Matrix

- Cada documento se torna um vetor de “termos”.
  - Cada termo é um componente (atributo) de um vetor.
  - O valor de cada componente é o número de vezes que esse termo ocorre no documento

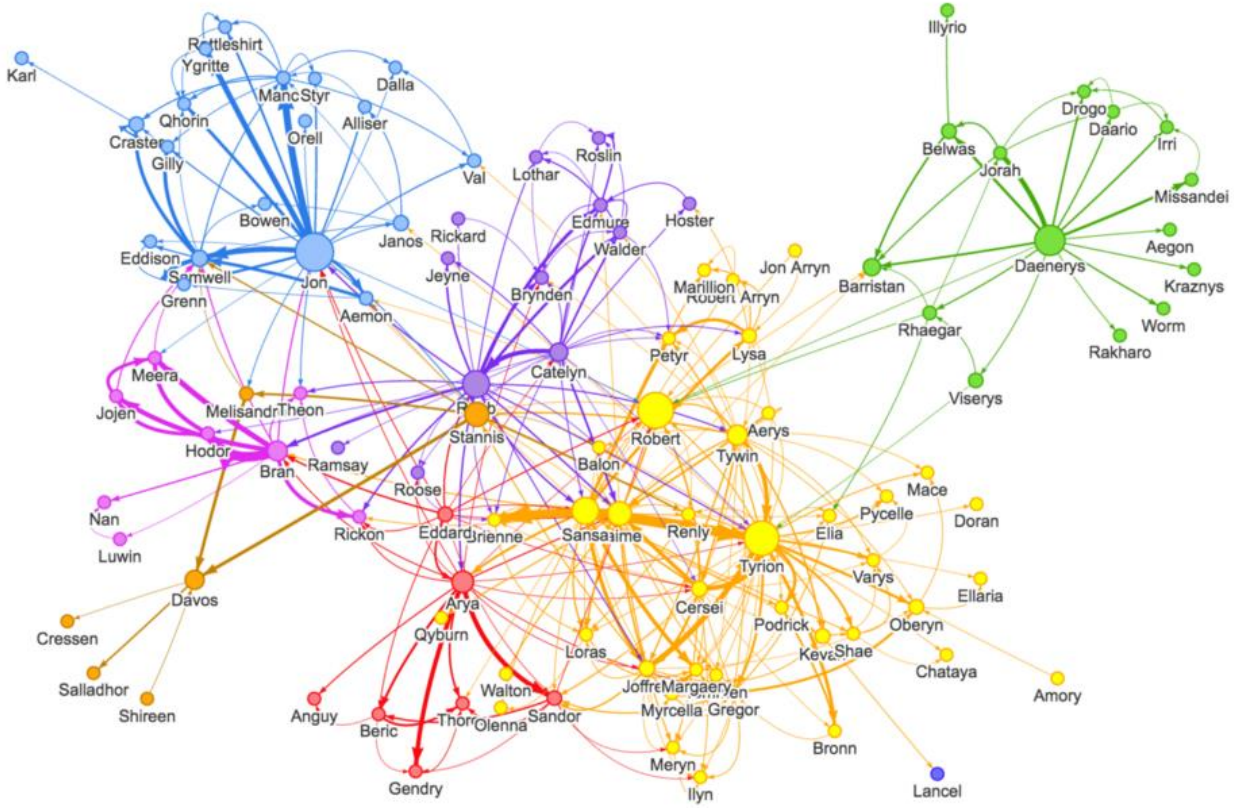
	Time	Jogo	Treinador	Vitória	Derrota	Temporada	Score	Intervalo	Falta	Jogador
Documento 1	3	0	5	0	2	6	0	2	0	2
Documento 2	0	7	0	2	1	0	0	3	0	0
Documento 3	0	1	0	0	1	2	2	0	3	0

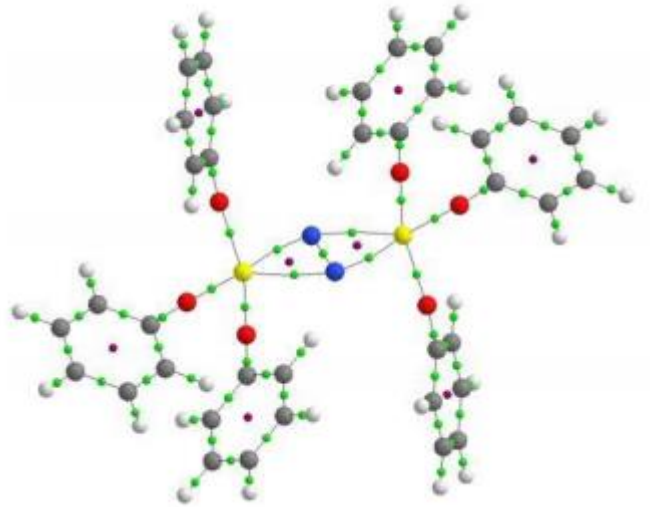
Dados de Transação

- Um tipo de registro de dados especial, onde
  - Cada registro (transação) envolve um conjunto de itens.
  - Por exemplo, considere uma mercearia. O conjunto de produtos comprado pode um cliente durante uma compra constitui uma transação, onde os produtos de forma individual constituem o item.

TID	Itens
1	Pão, Leite, Coca-Cola
2	Cerveja, Pão
3	Cerveja, Coca-Cola, Leite, Fralda
4	Água, Pão
5	Fraldas, Cerveja, Coca-Cola

Grafos

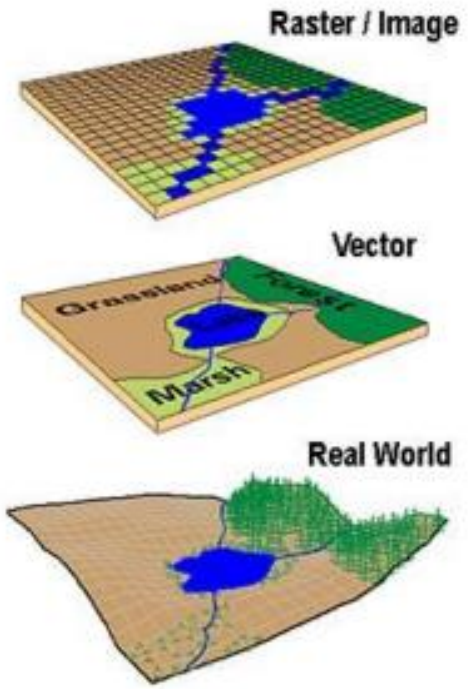


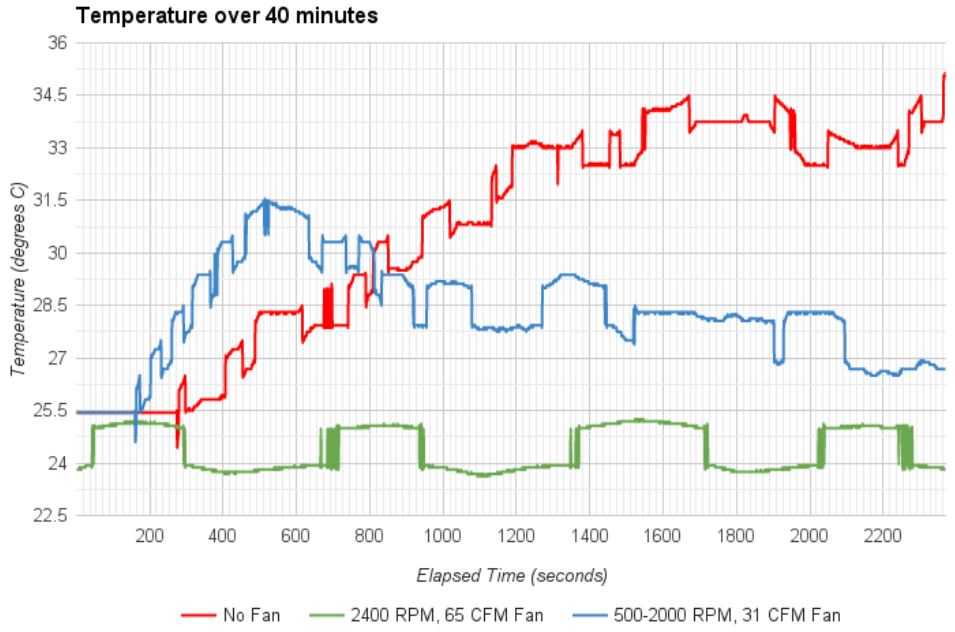


https://medium.com/neo4j/hands-on-graph-data-visualization-bd1f055a492d

Dados Ordenados

- Dados Espaciais
- Dados Temporais
- Dados de Sequência Genética



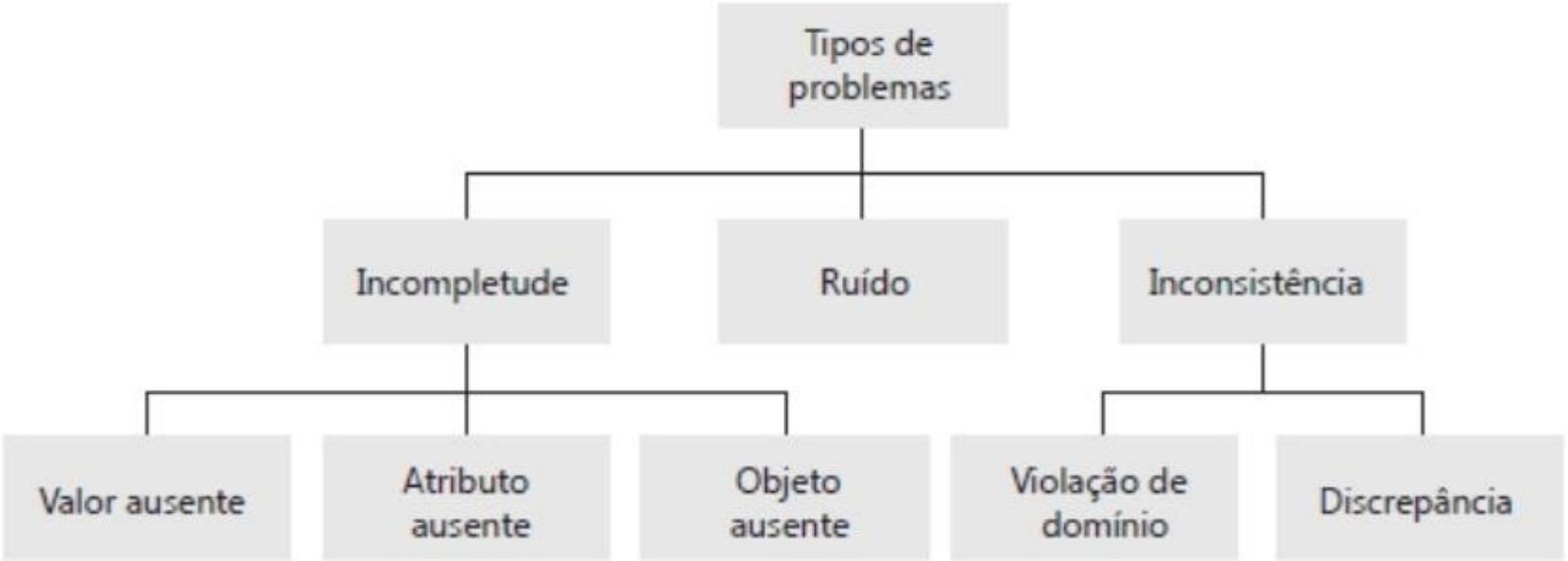


Seq Id	Seqüência
1786217	TAATACGGTTCTCTGATGAGGACCGTTTTTTTTTGCCCATTAAGTAAATC TTTTGGGGAATCGATATTTTTGATGACATA
1786230	GAATCAATATTATTGACTATAAGCCGCGTGAATATATGACTACACTTTGT GGGAAAACAAAGGCGTAATCACGCGGGCTA
1786240	GAATAGCGTCAGTGGTGTTAGGCACGGCATTGAATGACAGGTATGATAA TGCAAATTATAGGCGATGTCCCACAATTGAC
1786250	GACTTTTCTGCCGTGATTATAGACACTTTTGTTACGCGTTTTTGTCATGGC TTTGGTCCCGCTTTGTTACAGAATGCTTT
1786262	TCTTTTAGAGCGCCTCGCTTCGGGCATAAAAAAACCCGCGCAATGGCGCG GGTTTTTTGTTTGACTGCGTGCTGGCTTAA
1786283	TTTTTATGAGGCCGACGATGATTACGGCCTCAGGCGACAGGCACAAATCG GAGAGAAACTATGTTTGAACCAATGGAACT
1786298	AAACGGGAAAGCAGATTCCGAGGTTTTTTATTTCGTTGCAGCGAAAGACAA GAAATTTGCGAGGCGTTACGAAGAAAGTT
1786181	AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAA AGAGTGTCTGATAGCAGCTTCTGAACTG
1786192	CTGAATAACTGTAGTGTTTTCAGGGCGCGGCATAATAATCAGCCAGTGGGG CAGTGTCTACGATCTTTTGAGGGGAAAAT
2367095	GATTCTTAAGCCACGAAGAGTTCAGATAGTACAACGGCATGTCTCTTTTGAC TATCTGGCAACCGGCAGTGTGTTCTCTC

Porque Fazer o Pré-Processamento?

- Os dados brutos contidos nos bancos de dados não estão processados, são incompletos e ruidosos.

Figura 2.1 Principais problemas com os dados



Porque Fazer o Pré-Processamento?

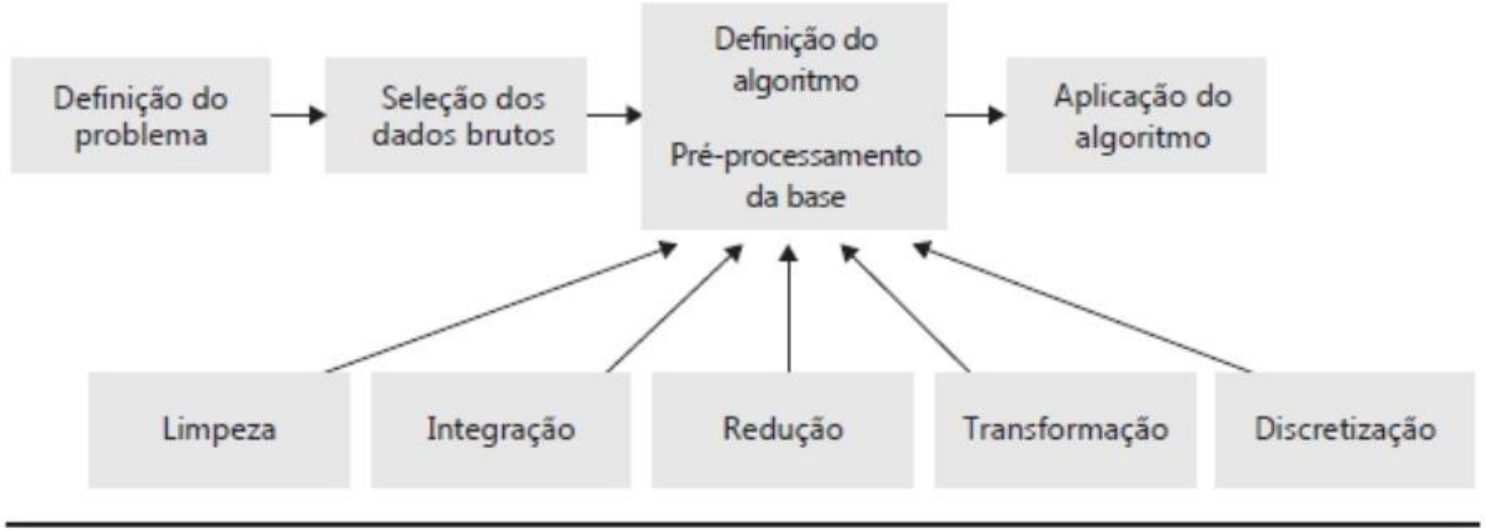
- Os dados ou bancos de dados podem conter
  - Campos obsoletos ou redundantes;
  - Valores em falta;
  - Outliers;
  - Dados em um formato que não contribuem para os modelos de mineração de dados;
  - Valores que não se mostram consistentes.
- Para fazer uso efetivo da mineração, é preciso pensar em algumas questões importantes antes de iniciar a análise.
  - Se existem dados ausentes, inconsistentes ou ruidosos, como trata-los?
  - É possível resumir a base de dados de forma que sejam obtidos resultados melhores no processo de mineração?
  - Existem atributos que são mais relevantes que outros, ou até irrelevantes, para uma dada análise?
  - Quais são os tipos de atributos da base? É preciso padronizá-los?
  - Há atributos naturalmente inter-relacionados?

- O objetivo das técnicas de pré-processamento de dados é, portanto, preparar os dados brutos para serem analisados, permitindo responder essas e outras perguntas.
- Por fim é preciso minimizar o GIGO – garbage in-garbage out (lixo colocado para dentro, lixo colocado para fora), conhecendo os dados e efetuando o pré-processamento.

Porque Fazer o Pré-Processamento?

- Para fins de mineração de dados, os valores do banco de dados devem ser submetidos a algumas etapas, as mais comuns são:
  - Limpeza dos dados;
  - Transformação dos dados.
- No entanto essas duas etapas fazem parte de um processo maior.

Figura 2.3 Etapas do processo de preparação da base de dados



- A preparação dos dados contém 60% do esforço do processo de mineração de dados.

Limpeza dos Dados

Dados que precisam ser Limpos/Transformados

Tabela 2.1 Cadastro de pessoas interessadas em obter um financiamento imobiliário

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal ($)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	M	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

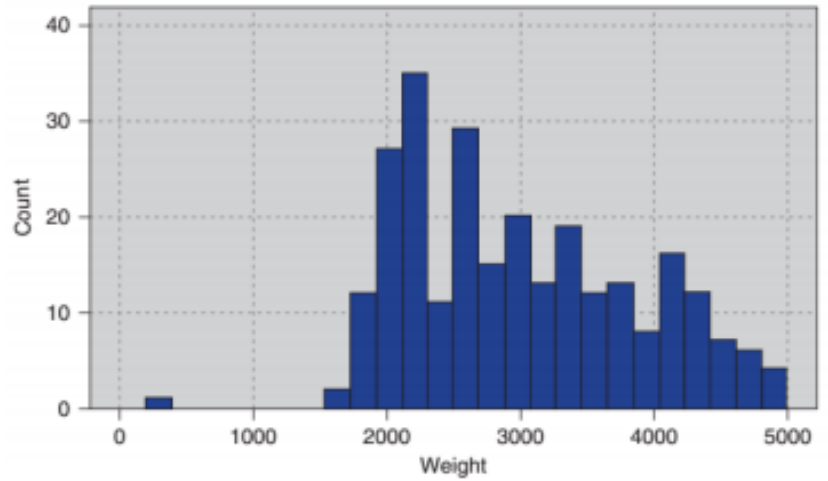
Lidando com Valores Ausentes

Um valor ausente costuma ser representado por um código de ausência, que pode ser um valor específico, um espaço em branco ou um símbolo, como na tabela do slide anterior.

- Os métodos tradicionais para imputação de valores ausentes são:
  - Ignorar o objeto: consistem em remover o objeto da base (não recomendado).
  - Imputar manualmente os valores ausentes: escolher de forma empírica um valor a ser imputado.
  - Usar uma constante global para imputar o valor ausente: substituir todos os valores ausentes por uma constante única.
  - Imputação hot-deck: imputa um valor usando o valor do mesmo atributo de um objeto similar aleatoriamente selecionado. A similaridade pode ser calculada por uma medida de similaridade ou distância entre objetos.
  - Imputar de acordo com a última observação: envolve ordenar a base de dados seguindo um ou mais de seus atributos, feito isso o algoritmo busca cada valor ausente e usa aquele valor da célula imediatamente anterior.
  - Usar a média ou moda de um atributo para imputar o valor ausente: essa técnica é bastante utilizada na prática, mas desconsidera as diferenças entre as classes e é suscetível a outliers.
  - Usar a média ou moda de todos os objetos da mesma classe para imputar o valor ausente;
  - Usar modelos preditivos para imputar o valor ausente.

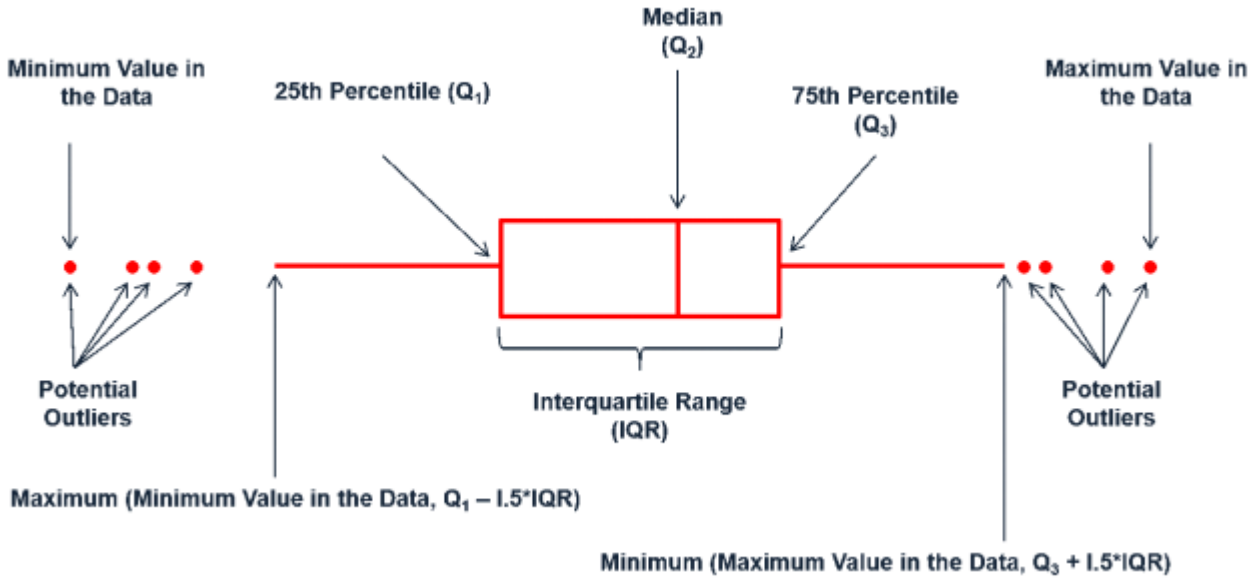
Lidando com Outliers

- Outliers são valores que ficam próximos a limites extremos de dados;
- Outliers podem representar erros na entrada de dados;
- Mesmo que um outlier seja um ponto de dados válido e não um erro
  - os métodos estatísticos são sensíveis a outliers e podem produzir resultados instáveis



Lidando com Outliers

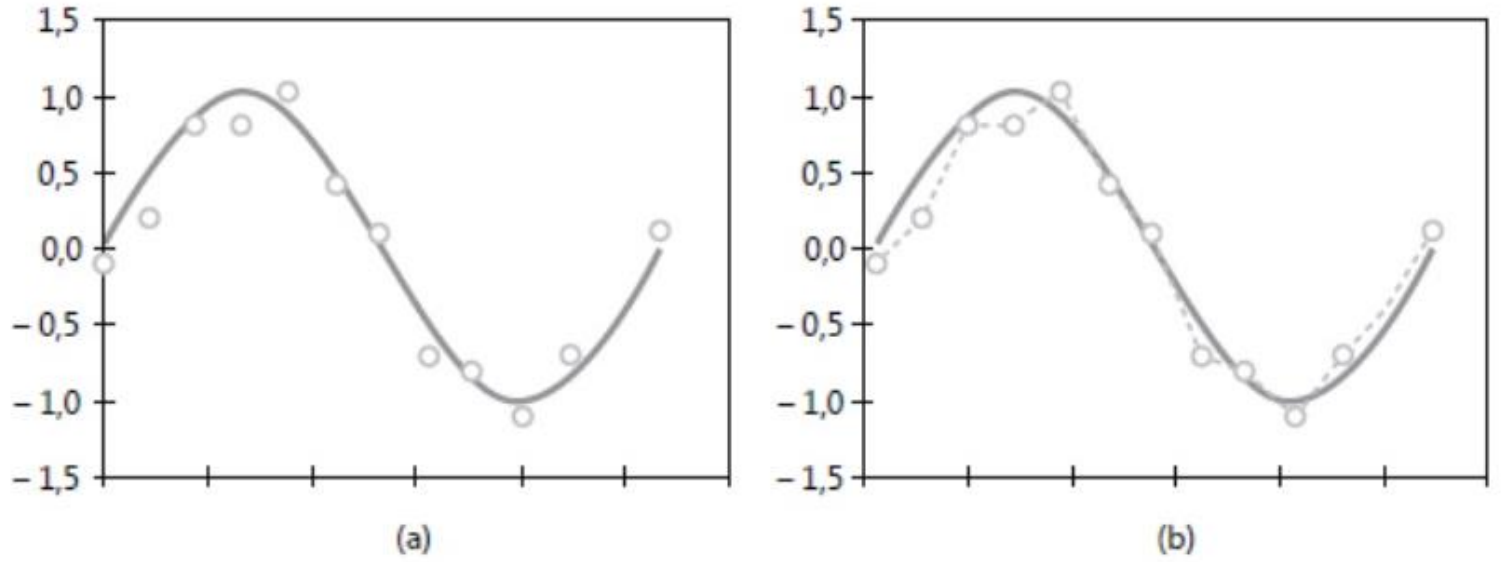
- Uma das formas mais fáceis de identificação é utilizando visualizações.
- A partir da identificação é possível fazer o tratamento e aqui se passa a ser necessário a utilização de métodos númericos;
  - Z-score para identificar outlier: um valor é um outlier se Z for menor que -3 ou maior que 3.
  - Entretanto, a média e o desvio padrão são sensíveis a outliers, remover outliers afeta essas métricas.
- IQR (interquartile range)
  - Útil para identificar e remover outliers (abordagem bruta de remoção).
- IQR é mais robusto que o desvio padrão
- IQR = Q3 – Q1
  - representam a propagação do centro, 50% dos dados.



Dados Ruidosos

Uma das formas de lidar com dados ruidosos é o processo de suavização que pode ser conseguido através de métodos como encaixotamento (binning)

Figura 2.4 Função *seno* amostrada por treze pontos com ruído. (a) Função ideal e pontos amostrados. (b) A linha tracejada representa a função aproximada com erro zero para os dados de treinamento



Transformação dos Dados

- Variáveis tendem a ter intervalos diferentes;
- Muitos Algoritmos são afetados por essas diferenças;
- Variáveis com intervalos maiores tendem a ter mais influência no resultado final dos modelos;
- Assim sendo, valores numéricos devem ser normalizados;
- Dentro da transformação de Dados temos:
  - Padronização
  - Normalização
  - etc.

Padronização

- O objetivo principal da padronização é resolver as diferenças de unidades e escalas dos dados.

  - **Capitalização**: é usual padronizar as fontes, normalmente para maiúsculo.
  - **Caracteres especiais**: uma simples troca de letras em valores nominais pode evitar eventuais problemas.
  - **Padronização de formatos**: observar e padronizar o formato de cada atributo da base, principalmente quando diferentes bases precisam ser integradas.
  - **Conversão de unidades**: todos os dados devem ser convertidos e padronizados em uma mesma unidade de medida.

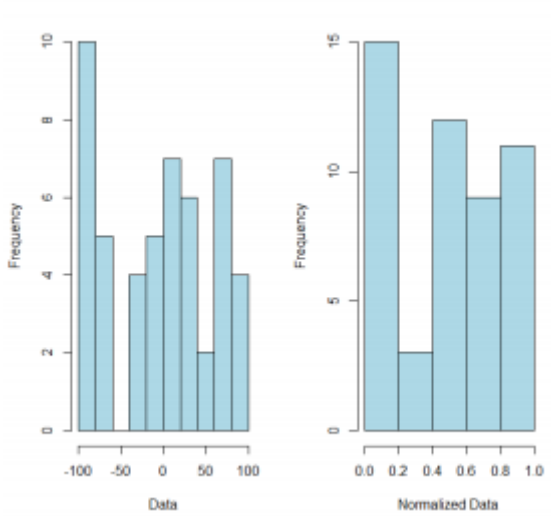
Normalização

- A normalização é um processo de transformação dos dados que objetiva torná-los mais apropriados à aplicação de algoritmos de mineração.
- Normalização Min-Max

$$a' = \frac{a - \min_a}{\max_a - \min_a}(novo\_\max_a - novo\_\min_a) + novo\_\min_a$$

- Entre 0 e 1

$$v' = \frac{v - min_X}{max_X - min_X}$$



Normalização

- Normalização pelo escore-Z (normalização de média zero)
  - Os valores de um atributo a são normalizados tendo como base a média e o desvio padrão de a.

$$a' = (a - \bar{a})/\sigma_a$$

- Normalização pelo escalonamento decimal
  - A normalização por escalonamento decimal move a casa decimal dos valores do atributo a. O número de casas decimais movidas depende do valor máximo absoluto do atributo a.

$$a' = a/10^j$$

- Normalização pelo range interquartil
  - A normalização pelo range interquartil toma cada valor do atributo, subtrai a mediana e divide pelo range interquartil (IQR).

$$a' = (a - Q_2)/\text{IQR}$$

Transformações para atingir a Normal

- As transformações mais comuns são:
  - Natural log transformation

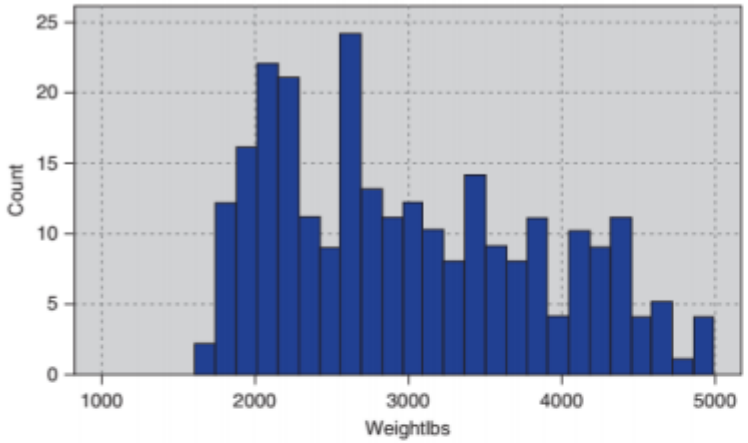
$$\ln(weight)$$

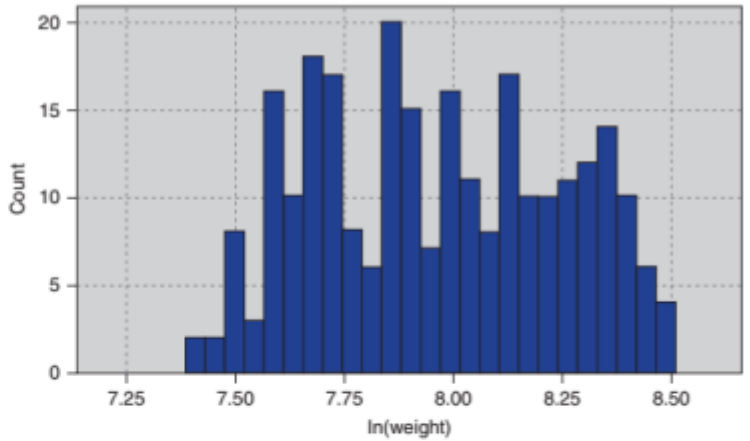
  - Square root transformation weight

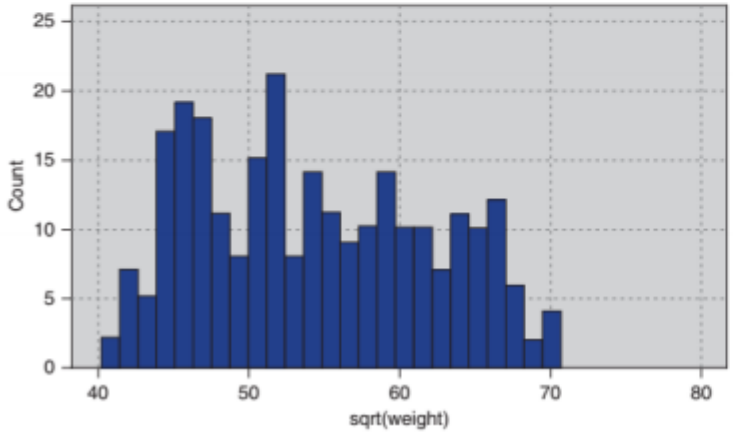
$$\sqrt{weight}$$

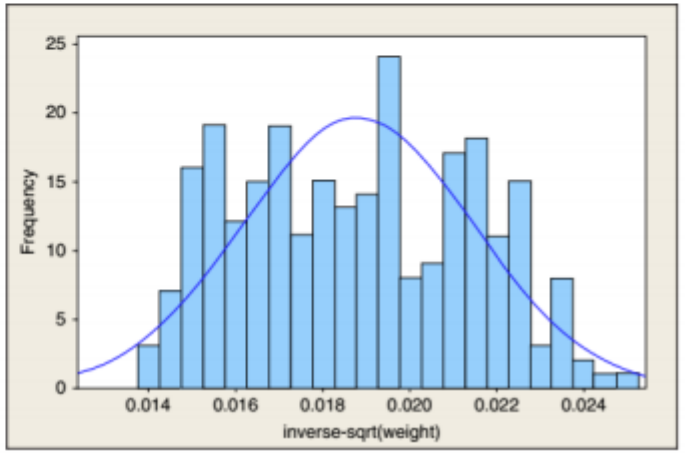
  - Inverse square root transformation

$$1/\sqrt{weight}$$









Encaixotamento (binning)

- Transforma variáveis numéricas em conjuntos (caixas), podendo ser utilizado encaixotamento por largura ou frequência.

Caixa	Intervalo	Qnt. Objeto	Média
1	[18,31)	48	24
2	[31,44)	157	38
3	[44,57)	269	50
4	[57,70)	330	63
5	[70,83)	132	74
6	[83,96)	25	86

- Os métodos de encaixotamento também podem ser usados para *quantizar* ou *discretizar* os dados.
- A reclassificação de variáveis categóricas, é o equivalente ao binning de variáveis numéricas.
  - Por exemplo uma coluna com 50 estados diferentes, pode ser reclassificado, juntando os estados por regiões do país, norte, sul, noroeste, sudeste.

Remoção de Variáveis

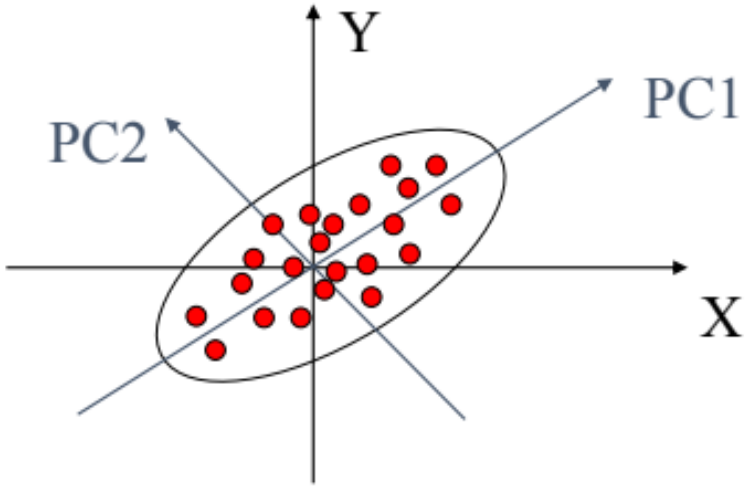
- Remova varáveis que não irão auxiliar durante a análise.
- É uma prática comum remover variáveis com:
  - 90% ou mais dos valores, são valores em falta
  - Variáveis que estão muito correlacionadas (uma correlação acima de 7 é considerada muito alta).
- No entanto
  - uma variável que tem 90% ou mais valores ausentes, pode apresentar um padrão de valores em falta (útil).
- Remova valores duplicados, redundantes e conflituosos.
- Registros duplicados levam a uma sobrecarga dos valores nos dados.

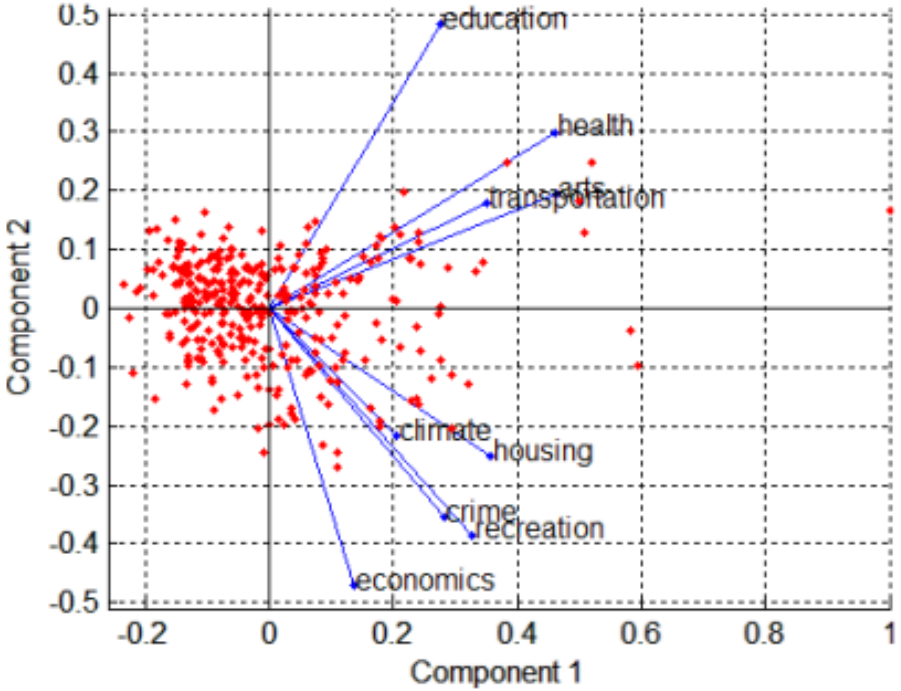
Redução dos Dados

- **Tipos de redução de dados:**
  - **Seleção de atributos (ou características)**: efetua uma *redução de dimensionalidade* na qual atributos irrelevantes, pouco relevantes ou redundantes são detectados e removidos.
  - **Compressão de dados**: também efetua uma redução da dimensionalidade, mas empregando algoritmos de *codificação* ou *transformação* de dados (atributos).
  - **Redução no número de dados**: os dados são removidos, substituídos ou estimados por representações menores, como modelos paramétricos e métodos não paramétricos.
  - **Discretização**: os valores de atributos são substituídos por intervalos ou níveis conceituais mais elevados, reduzindo a quantidade final de atributos.

Redução de Dimensionalidade PCA/ACP

- Análise de Componentes Principais (Principal Component Analysis)





Redução de Dimensionalidade PCA/ACP

- Itens relevantes para compreender melhor o PCA
  - Componentes principais (resultado final)
  - Autovalores
  - Autovetores
  - Variância e Covariância

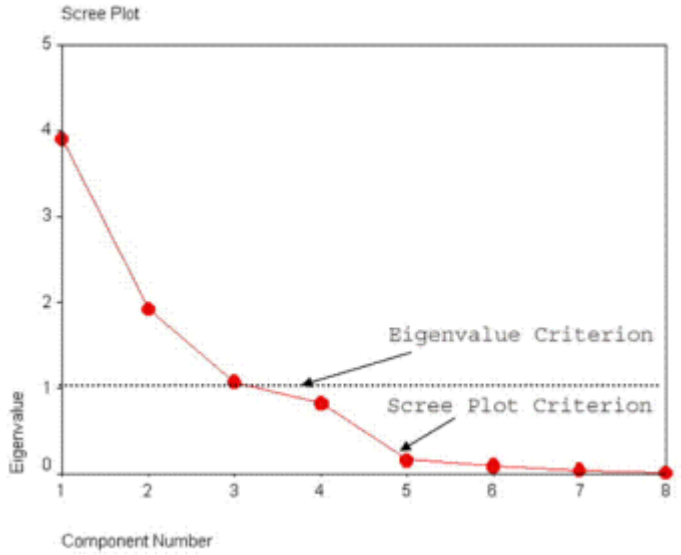
- Leitura para compreender de forma mais simples:
- http://roneiecologia.blogspot.com/2014/02/analise-de-componentes-principais-para.html

Como definir o número de componentes?

- Existem 3 critérios
  - O critério dos autovalores
  - O critério da proporção de variância explicada
  - O critério do scree plot.

- O critério dos autovalores
  - Um autovalor de 1 representa uma variável que tem valor.
  - Mantenha componentes que tenham autovalores > 1

- O critério da proporção de variância explicada
  - Step 1: especifique quanta variabilidade manter
  - Step 2: mantenha o número de componentes que refletem essa variabilidade.



- O critério do scree plot.
  - Método do gráfico “elbow” mantenha o número de componentes que estão no ponto de curvatura.
  - Scree plot: um gráfico que coloca os autovalores contra o número de componentes.

Exercicio

- Utilizando o arquivo que está no Github:
- https://github.com/eduardo2s/Aulas-L3P/blob/master/aula5_database.xlsx

- Executem o seguinte código e tentem entende-lo:
- https://colab.research.google.com/drive/1BjOpt0rJQYrS2pCD0RERnP6ToUDT5Q9Z

Referências

- http://roneiecologia.blogspot.com/2014/02/analise-de-componentes-principais-para.html
- https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/
- https://metodosdigitaisufg.wordpress.com/2018/02/24/analise-de-dados-utilizando-python/

- Normalização: http://professor.ufabc.edu.br/~ronaldo.prati/DataMining/DataPreparation.pdf

Obrigado

Real Python