Mineração de Dados

Conhecendo os seus Dados

Sumário

Objetos e Tipos de Atributos

Objetos e Tipos de Atributos

- Registros
  - Dados relacionais
  - Dados numéricos: matriz com resultado de uma simulação
  - Documentos: textos, vetores de termos frequentes
  - Dados transacionais

Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Objetos e Tipos de Atributos

- Redes e Grafos
  - *World Wide Web*
  - Redes sociais ou de informação
  - Estruturas moleculares
- Elementos ordenados
  - Vídeos: sequência de imagens
  - Dados temporais: séries temporais
  - Dados sequenciais: sequência de transações
  - Dados de uma sequência genética
- Dados espaciais: mapas ou seus elementos
- Imagens

Objetos e Tipos de Atributos

- Dimensionalidade
  - Maldição da dimensionalidade
- Espacialidade
  - Apenas a presença conta
- Resolução
  - Padrões dependem de escala
- Distribuição
  - Centralidade e dispersão

Objetos de Dados

- Conjuntos de dados são feitos de objetos de dados
- Um objeto de dados representa uma entidade
- Exemplos de bases de dados
  - vendas: clientes, itens, vendas, ...
  - médico: pacientes, tratamentos, ...
  - universitário: estudantes, professores, cursos, ...
- Também podem ser referidos como: amostras, exemplos, instâncias, pontos, objetos, tuplas
- Objetos de dados são descritos por atributos
- Linhas do banco de dados: objetos
- Colunas do banco de dados: atributos

Atributos

- Atributo (ou dimensão, característica, variável): um dos campos que representa uma característica das amostras
  - Exemplos: identificador, nome, endereço
- Tipos
  - Nominal
  - Binário
  - Ordinal
  - Numérico: quantitativos discretos ou valores contínuos

Tipos de Atributos

- Nominal: categorias, estados ou "nomes de coisas"
  - Cores de cabelo = {preto, loiro, castanho, grisalho, ruivo, branco}
  - Estado civil, ocupação, identificadores
- Binário
  - Atributos nominais com apenas 2 estados (0 e 1)
  - Binário simétrico: ambos valores são igualmente importantes: sexo
  - Binário assimétrico: possíveis valores não são igualmente importantes: teste médico (positivo vs. negativo)
  - Convenção: atribuir 1 ao valor mais importante (e.g., HIV positivo)
- Ordinal
  - Os valores possuem alguma característica de ordenação (ou classificação)
  - Não há relevância nas diferenças entre os valores
  - Tamanho = {pequeno, médio, grande}, graus, classificações

Tipos de Atributos Numéricos

- Quantidades que podem ser valores inteiros ou reais
  - Os valores possuem ordem
  - Valores medidos em uma escala igual de unidades
  - Exemplos: temperatura, data, distância, contagem, quantidade monetária

Atributos Discretos vs. Contínuos

- Atributo Discreto
  - Finito ou contável
  - Exemplos: número de filhos, dias em observação, conjunto de palavras em uma coleção de documentos
  - Atributos binários são um caso particular de atributos discretos
- Atributo Contínuo
  - Valores reais
  - Exemplos: temperatura, peso, altura
  - Na prática, os valores contínuos são representados por uma quantidade finita (*float* e *double*, por exemplo)

Descrições Estatísticas Básicas dos Dados

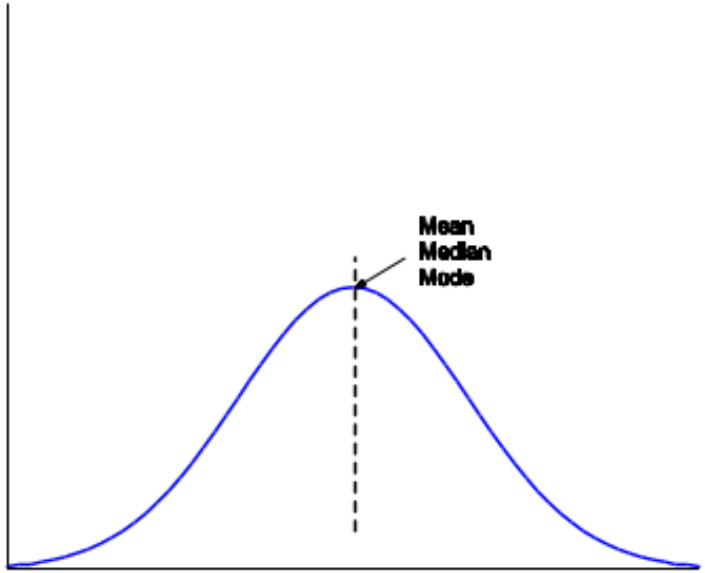
Tipos de Atributos Numéricos

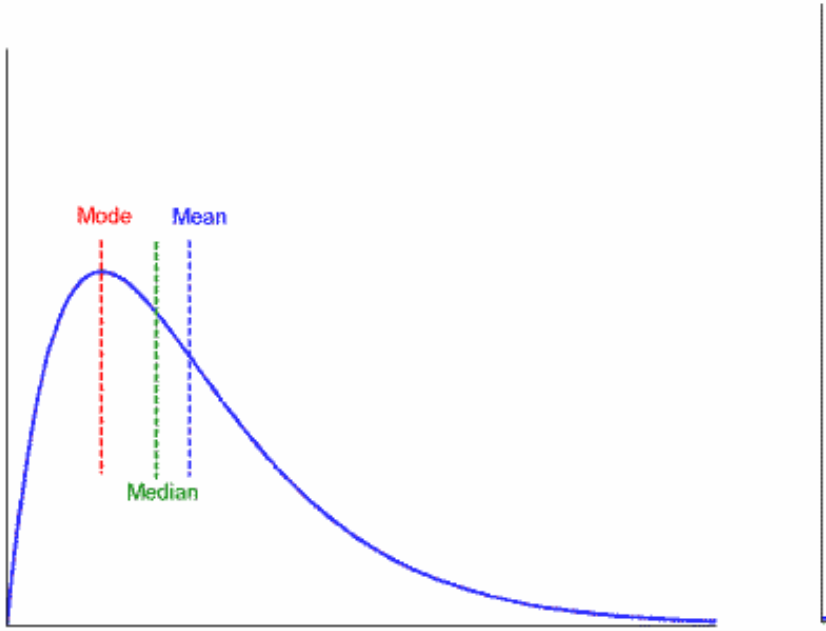
- Motivação
  - Melhorar o entendimento dos dados
  - Tendência central, variação, espalhamento
- Características de dispersão dos dados
  - mediana, máximo, mínimo, quartis, quantitativos, variância, *outliers*
- Dimensões numéricas correspondem a intervalos ordenados
  - Dispersão dos dados: analisada com múltiplas granularidades de precisão
  - Boxplots ou análise quantitativa em intervalos ordenados
- Análise de dispersão em medidas computadas
  - Agrupamento de medidas em dimensões calculadas
  - Boxplot ou análise quantitativa em dados transformados (*data cube*, por exemplo)

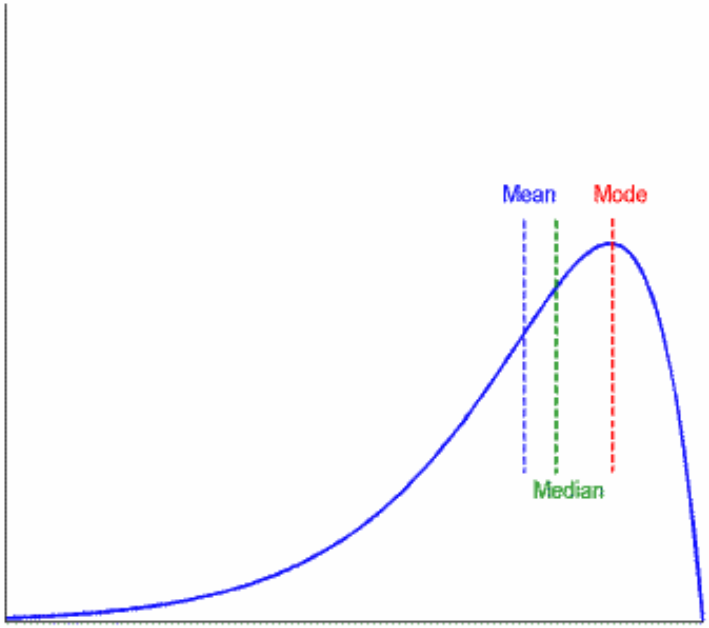
Medindo a Tendência Central

- Média
  - Pode ser ponderada (cálculo do Índice de Rendimento Acadêmico, por exemplo)
- Mediana
  - Ponto central de uma amostra de um número ímpar de valores; média dos pontos pontos centrais, caso contrário
- Moda
  - Valor que ocorre com mais frequência

Dados Simétricos ou Enviesados





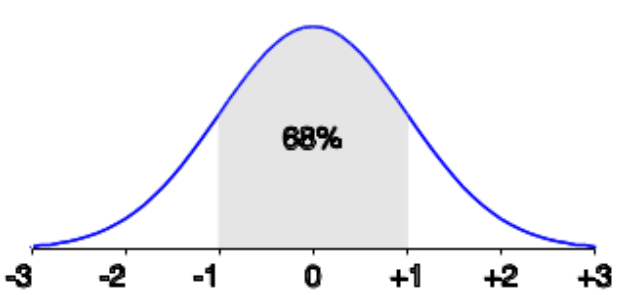


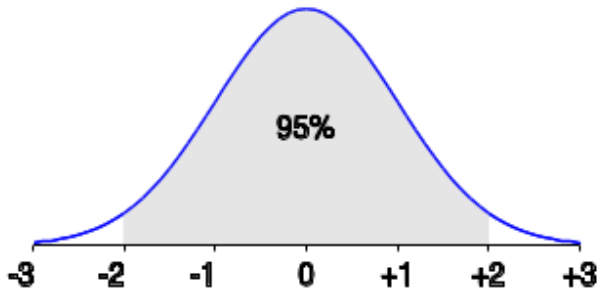
Medindo a Dispersão dos Dados

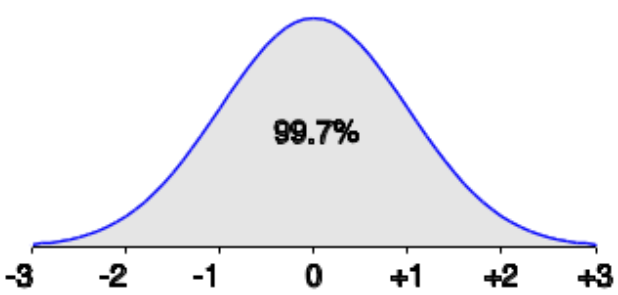
- Quartis: Q_1=25%, Q_2=50% (mediana) e Q_3=75%
- Intervalo interquartil
- Boxplot
- Variância: $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Desvio padrão σ

Distribuições Normais

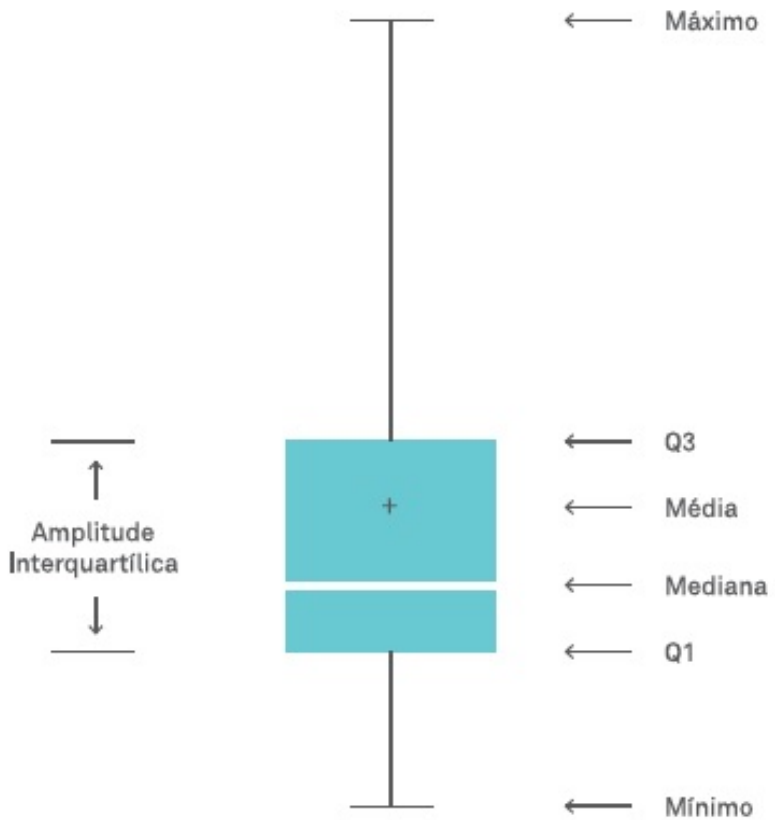
- Amostra normalmente distribuída
  - $[\mu - \sigma; \mu + \sigma]$: engloba aproximadamente 68% dos elementos (μ: média, σ: desvio padrão)
  - $[\mu - 2\sigma; \mu + 2\sigma]$: engloba aproximadamente 95% dos elementos
  - $[\mu - 3\sigma; \mu + 3\sigma]$: engloba aproximadamente 99.7% dos elementos







Análise Gráfica: Boxplot



- Limite superior: $\min\{\max\{\text{dados}\}, Q3 + 1{,}5(Q3 - Q1)\}$
- Limite inferior: $\max\{\min\{\text{dados}\}, Q1 - 1{,}5(Q3 - Q1)\}$

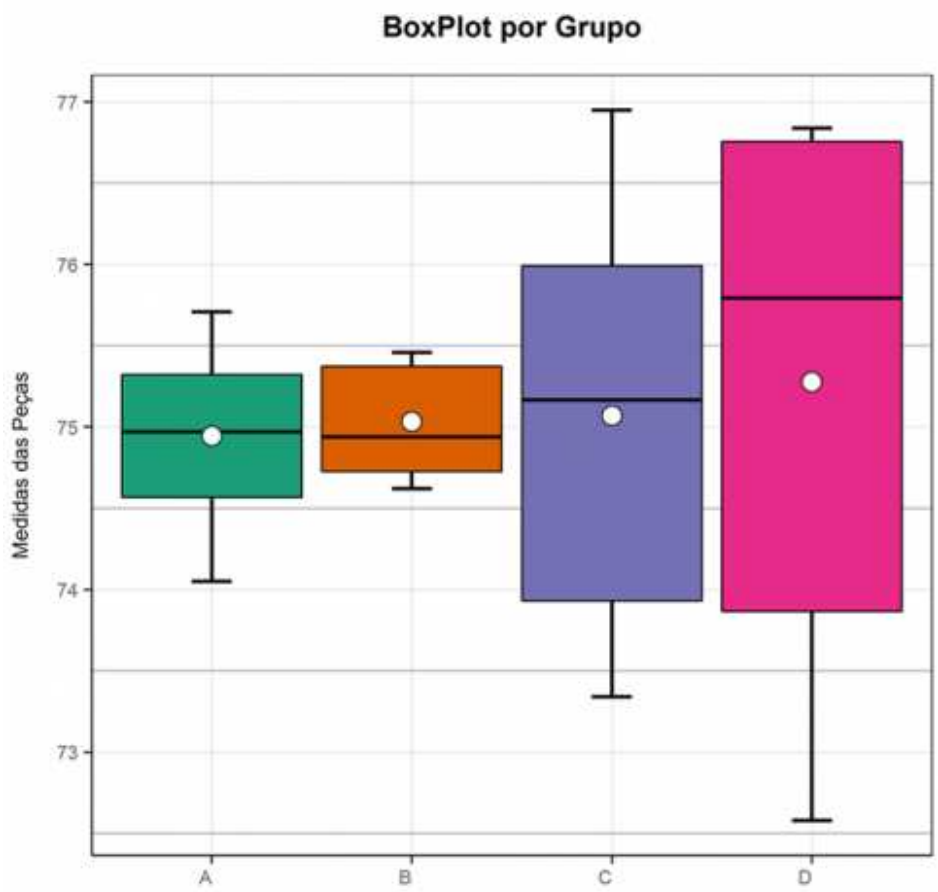
Análise Gráfica: Boxplot

A		B		C		D	
75,27	74,93	74,94	74,75	75,93	73,34	75,98	76,75
75,33	74,72	75,25	74,65	76,95	74,04	75,61	76,78
74,58	74,53	75,44	74,94	75,47	75	74,2	74,74
75,01	75,32	74,62	74,92	73,6	76,18	76,44	72,58
75,71	74,05	75,35	75,46	74,85	75,33	76,84	72,86

- Produção de peças com valor de referência igual a 75cm
- As equipes A e B foram treinadas para produzir as peças
- É importante treinar os membros das equipes?

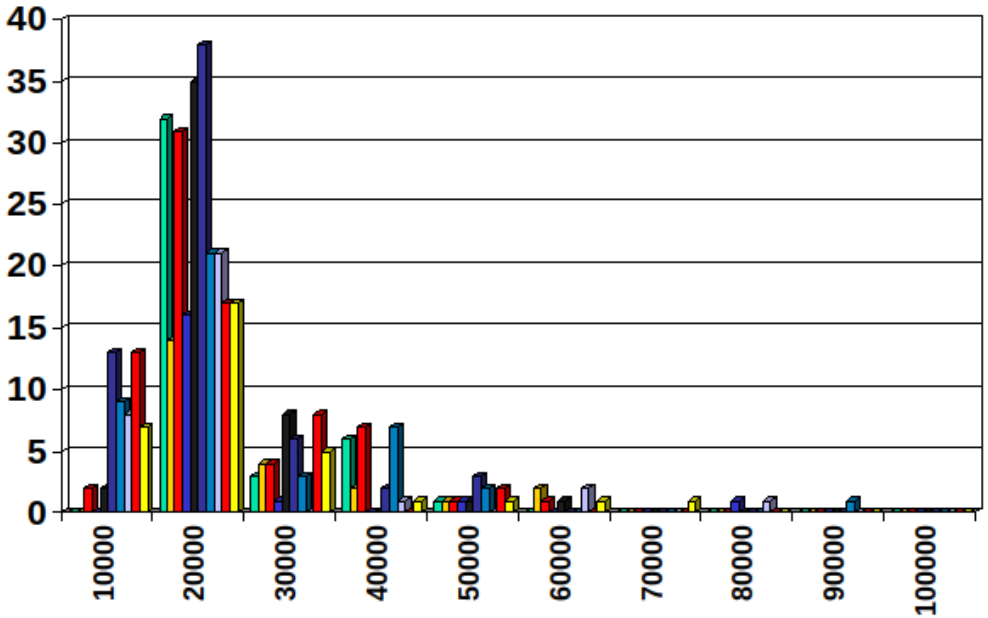
Fonte: `http://www.portalaction.com.br/estatistica-basica/31-boxplot`

Análise Gráfica: Boxplot



- Produção de peças com valor de referência igual a 75cm
- As equipes A e B foram treinadas para produzir as peças
- É importante treinar os membros das equipes?

Análise Gráfica: Histograma



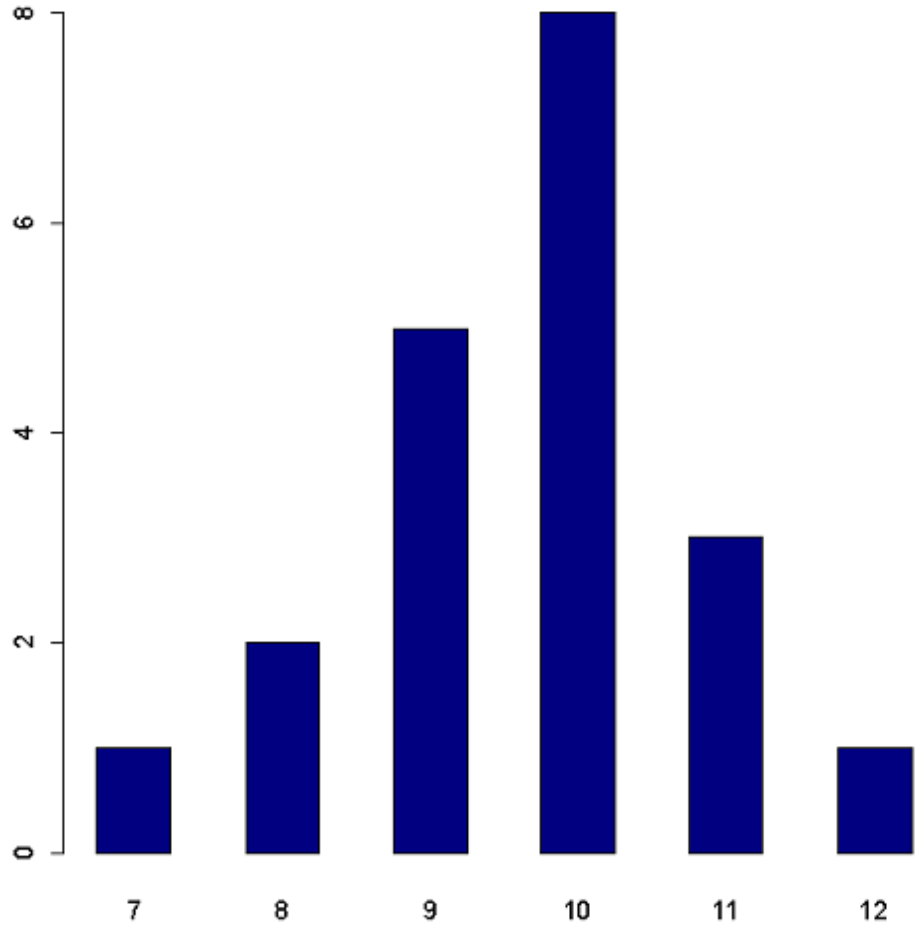
Análise Gráfica: Histograma

Número de pessoas com diabetes	Frequência (f_i)	Frequência relativa (f_{ri})	Frequência percentual	Frequência acumulada
7	1	0,05	5	5
8	2	0,1	10	15
9	5	0,25	25	40
10	8	0,4	40	80
11	3	0,15	15	95
12	1	0,05	5	100

- Número de pessoas com diabetes nos grupos

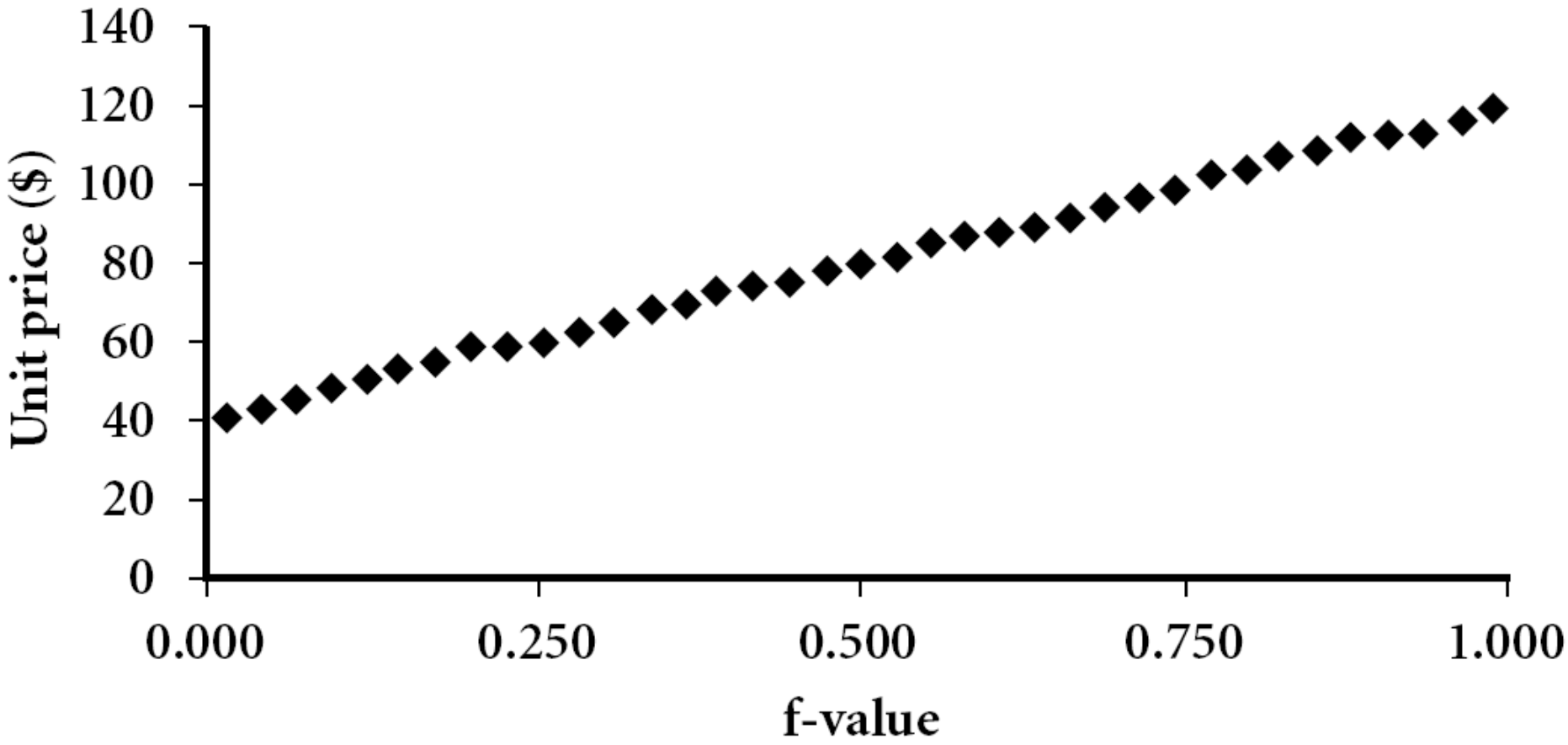
Fonte: `http://www.portalaction.com.br/estatistica-basica/16-histograma`

Análise Gráfica: Histograma

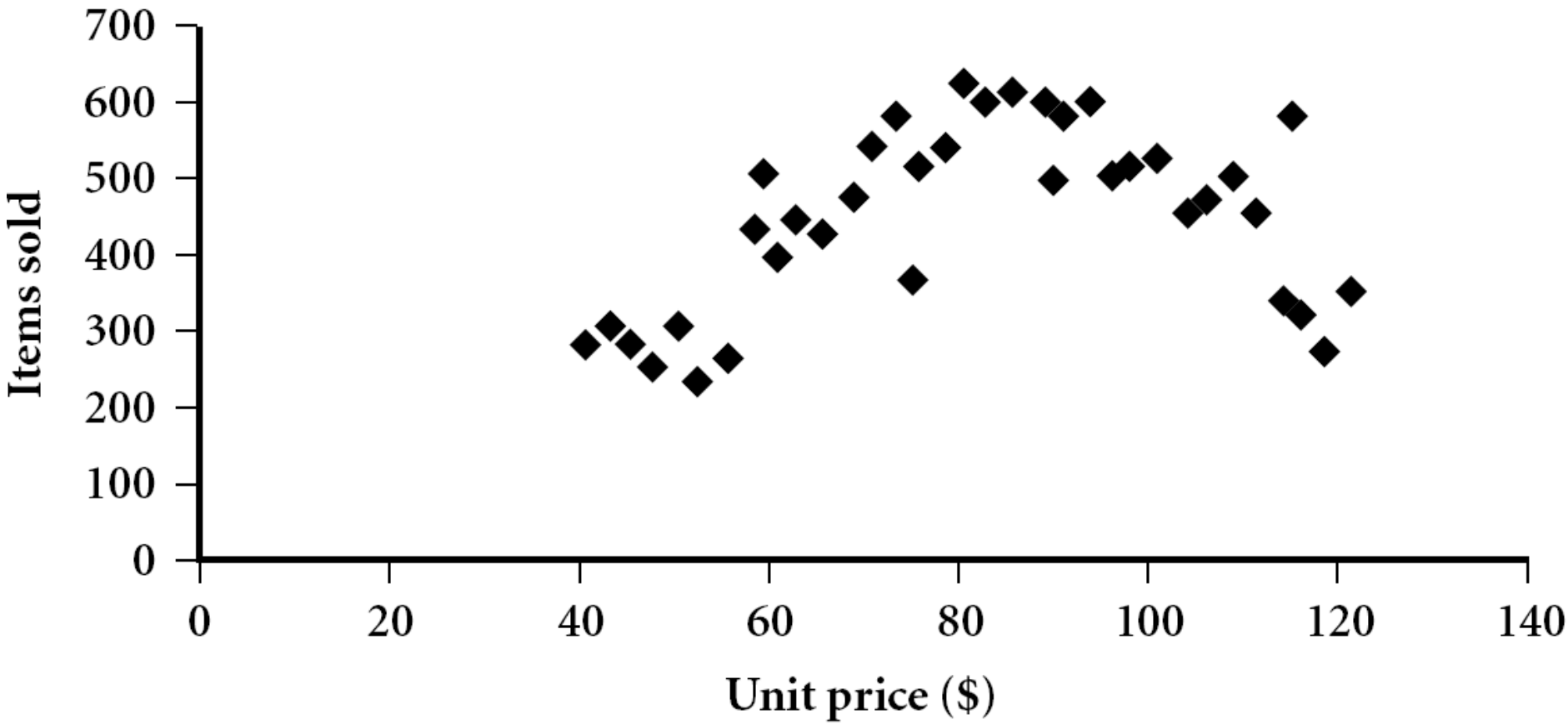


- Número de pessoas com diabetes nos grupos
- Indicar a que se refere os eixos

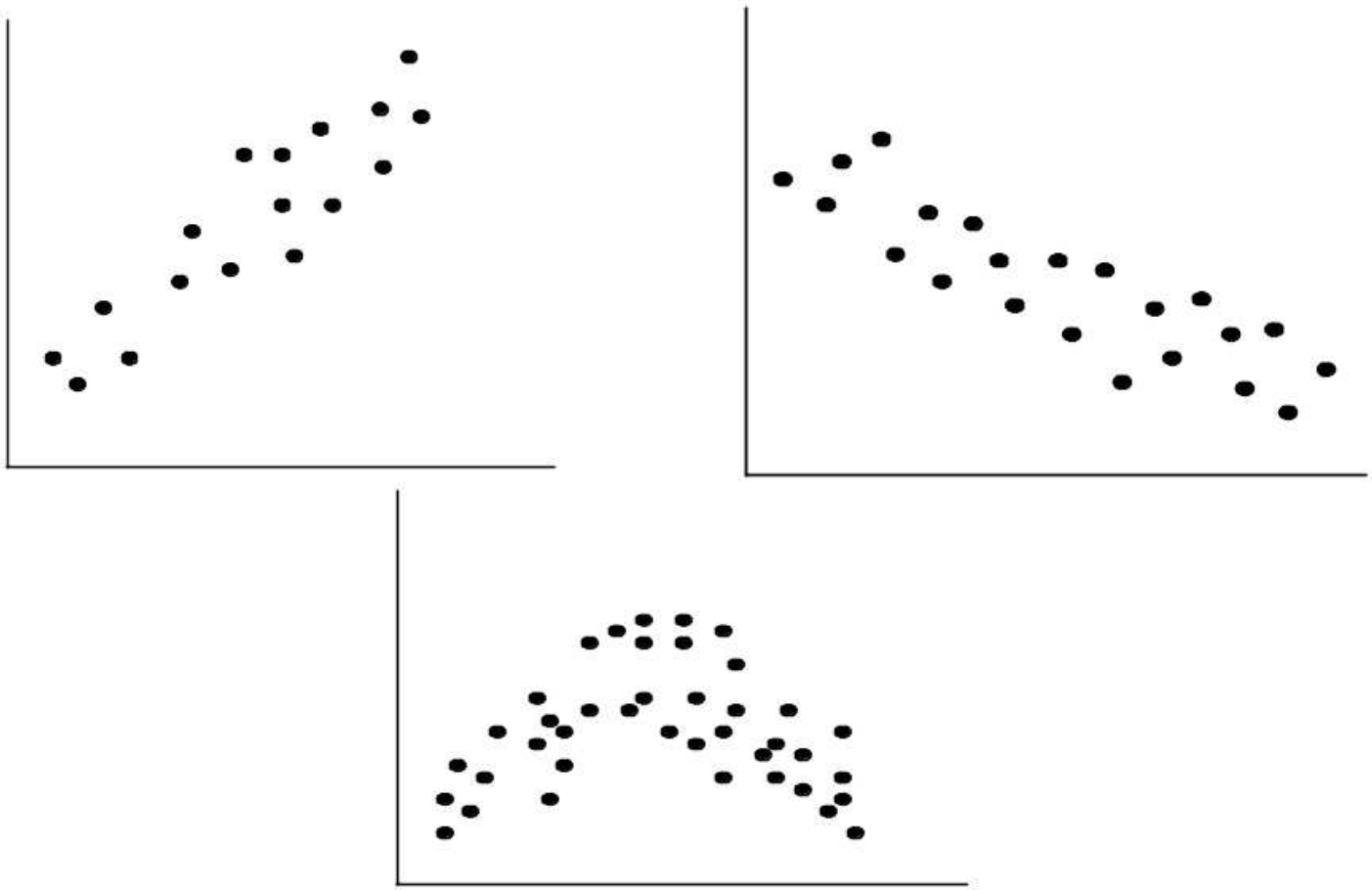
Análise Gráfica: Quantitativo/Acumulativo



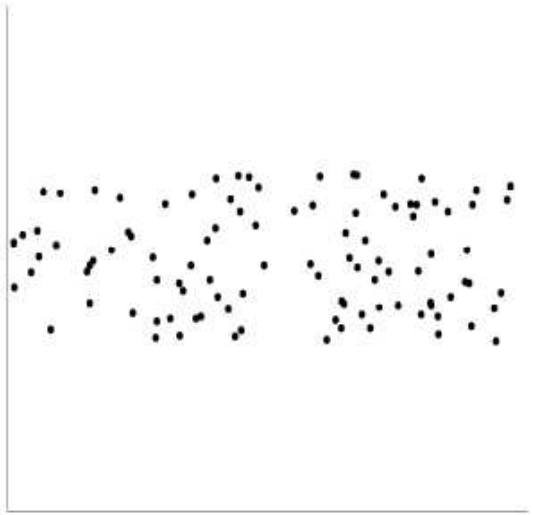
Análise Gráfica: Dispersão

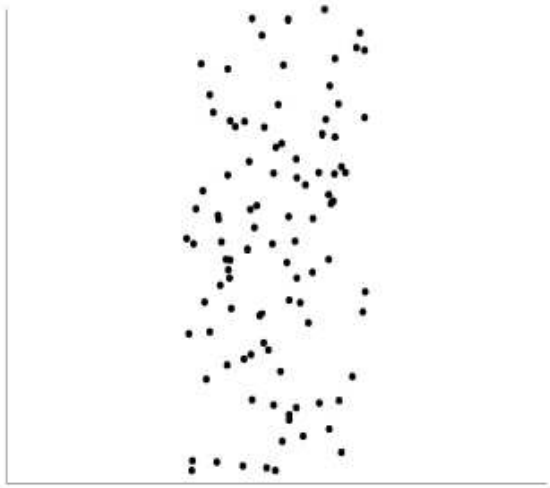


Análise Gráfica: Correlação



Análise Gráfica: Dados Não Correlacionados





Visualização dos Dados

Visualização dos Dados

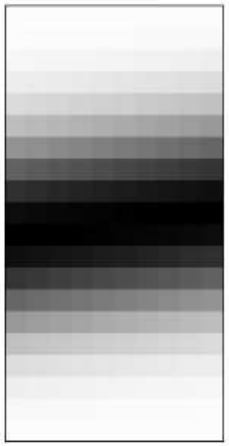
- Por que?
  - Obter *insights*
  - Prover uma ideia qualitativa de grandes quantidades de dados
  - Possibilitar a busca por padrões, tendências, estruturas, relações, ...
  - Ajudar a encontrar regiões de interesse
  - Evidenciar visualmente alguma afirmação
- Existem muitos meios de visualização dos dados
- Várias linguagens e bibliotecas estão disponíveis
  - `https://scikit-learn.org`
  - `https://www.scikit-yb.org` (Yellowbrick: ML Visualization)
  - `https://www.r-graph-gallery.com`
- Análise gráfica: gráficos apresentados anteriormente

Visualização Orientada a *Pixel*

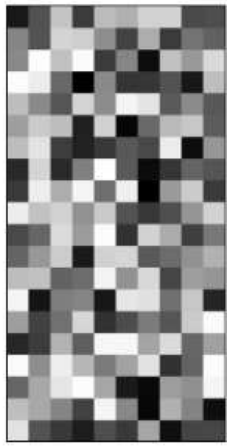
- Cria-se m gráficos, um correspondente a cada dimensão
- Os dados são ordenador em relação a uma variável de interesse
  - Valor gasto em compras, por exemplo
- As cores/intensidade dos pixels refletem os valores dos registros

(a) Income

(b) Credit Limit



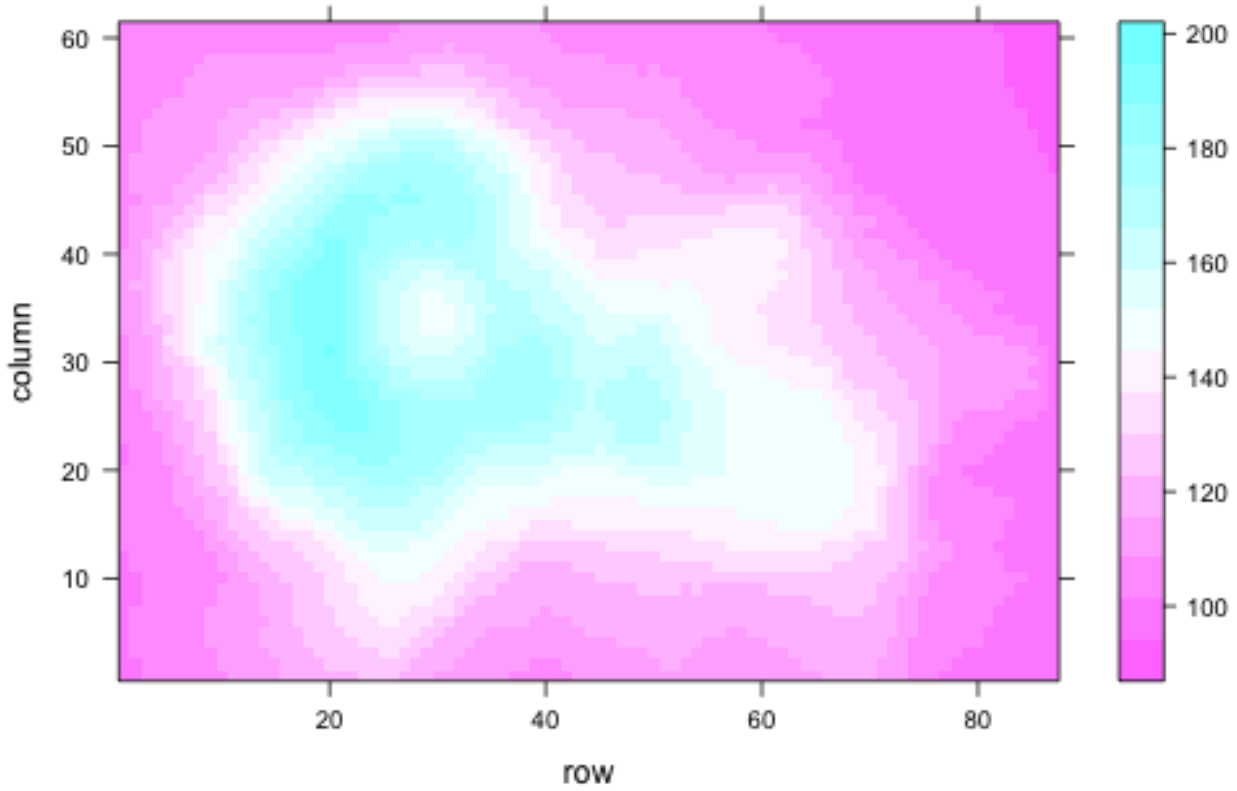
(c) transaction volume



(d) age

Visualização via *Heatmap*

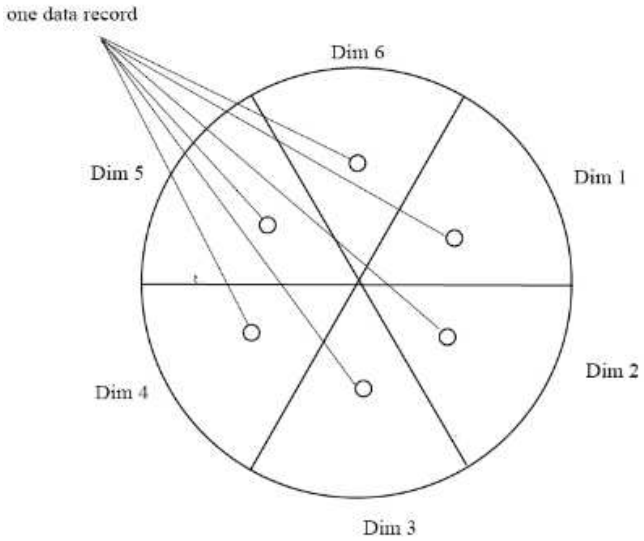
- Os dados são apresentados em gráficos bidimensionais
- A intensidade (temperatura) reflete o valor de um terceiro atributo



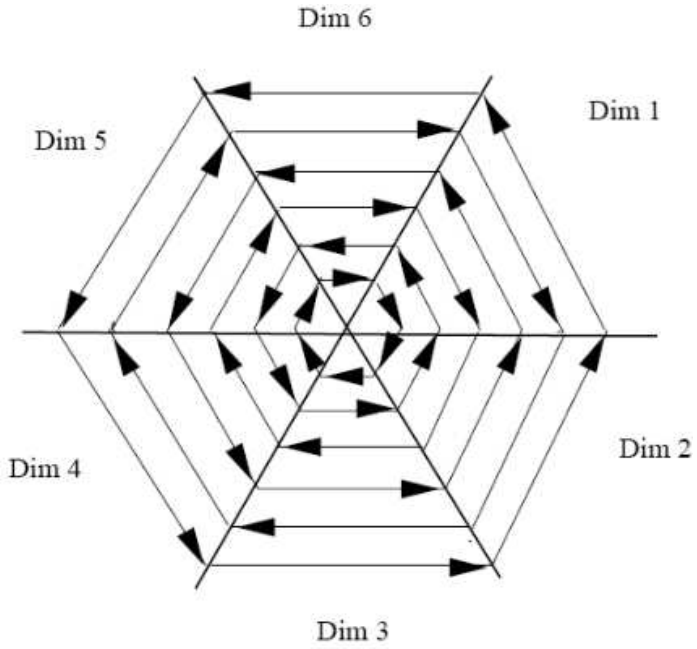
https://www.r-graph-gallery.com/heatmap

Visualização dos Dados em Círculos

- Facilita a visualização de relações em dados com muitas dimensões

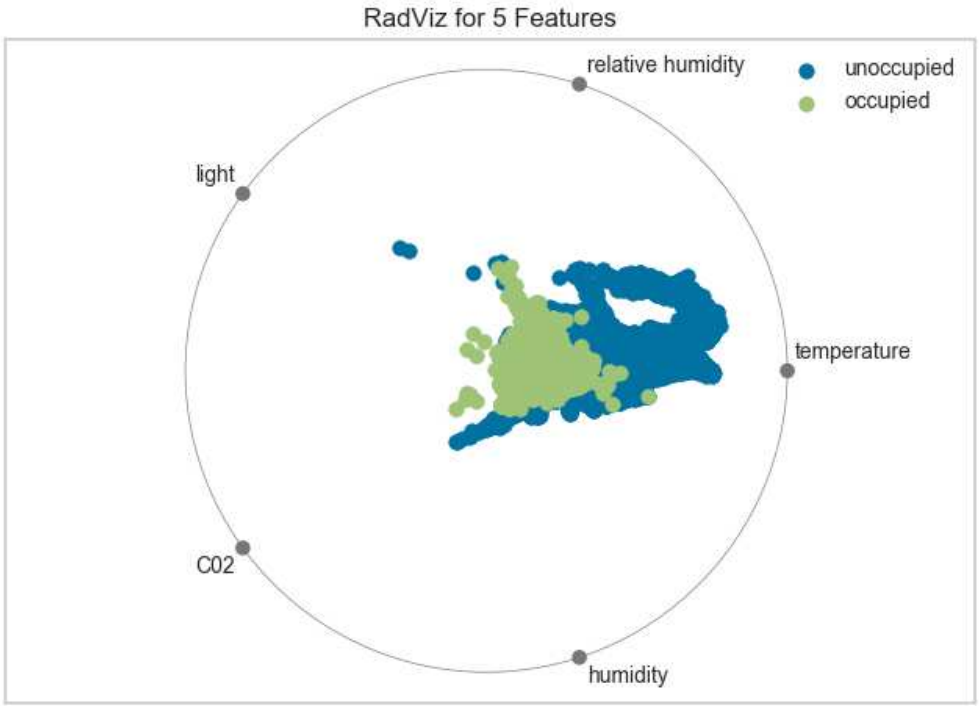


(a) Representing a data record in circle segment



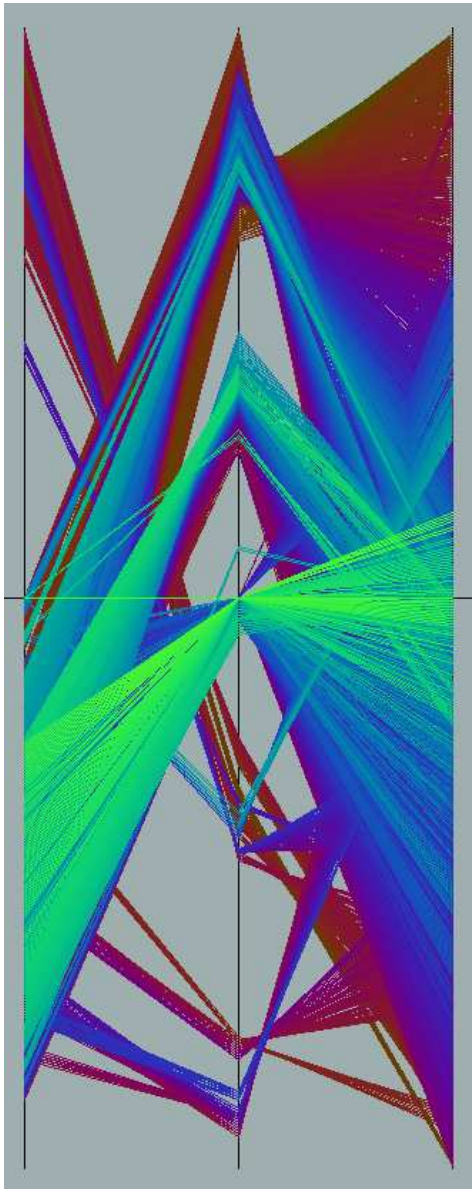
(b) Laying out pixels in circle segment

Visualização dos Dados em Círculos

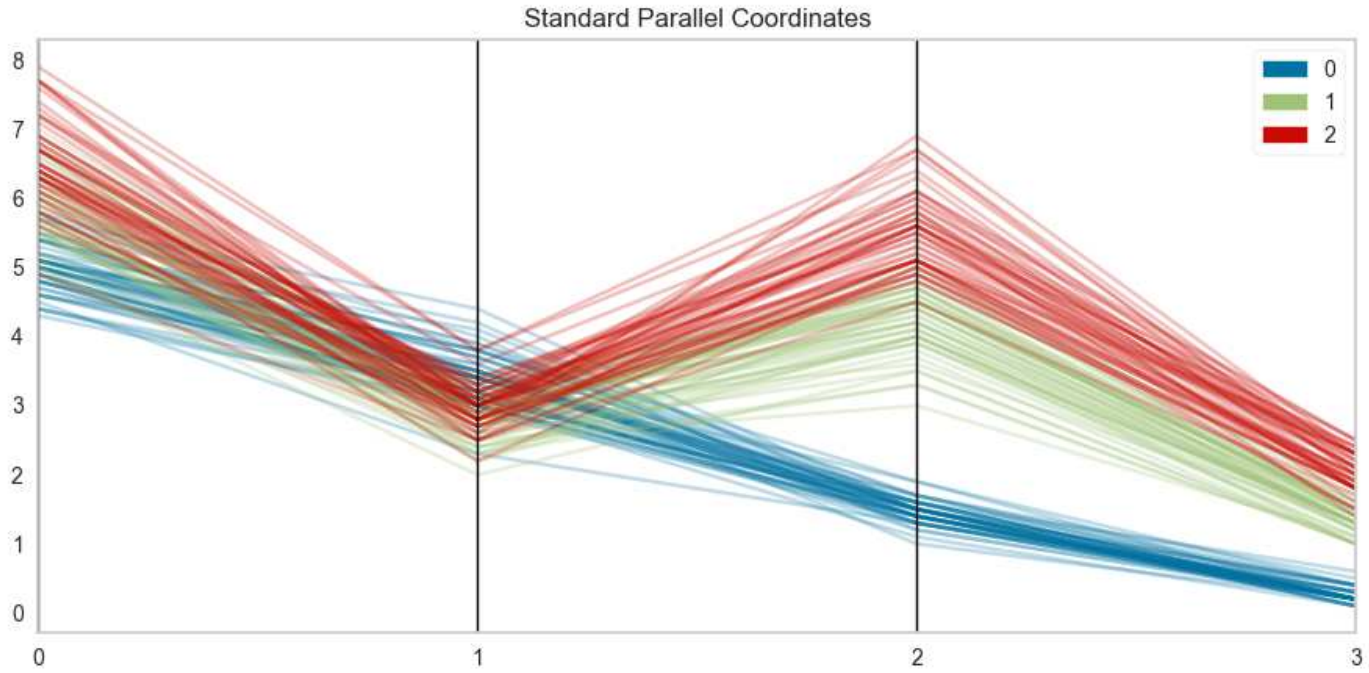


https://www.scikit-yb.org/en/latest/api/features/radviz.html

Coordenadas Paralelas

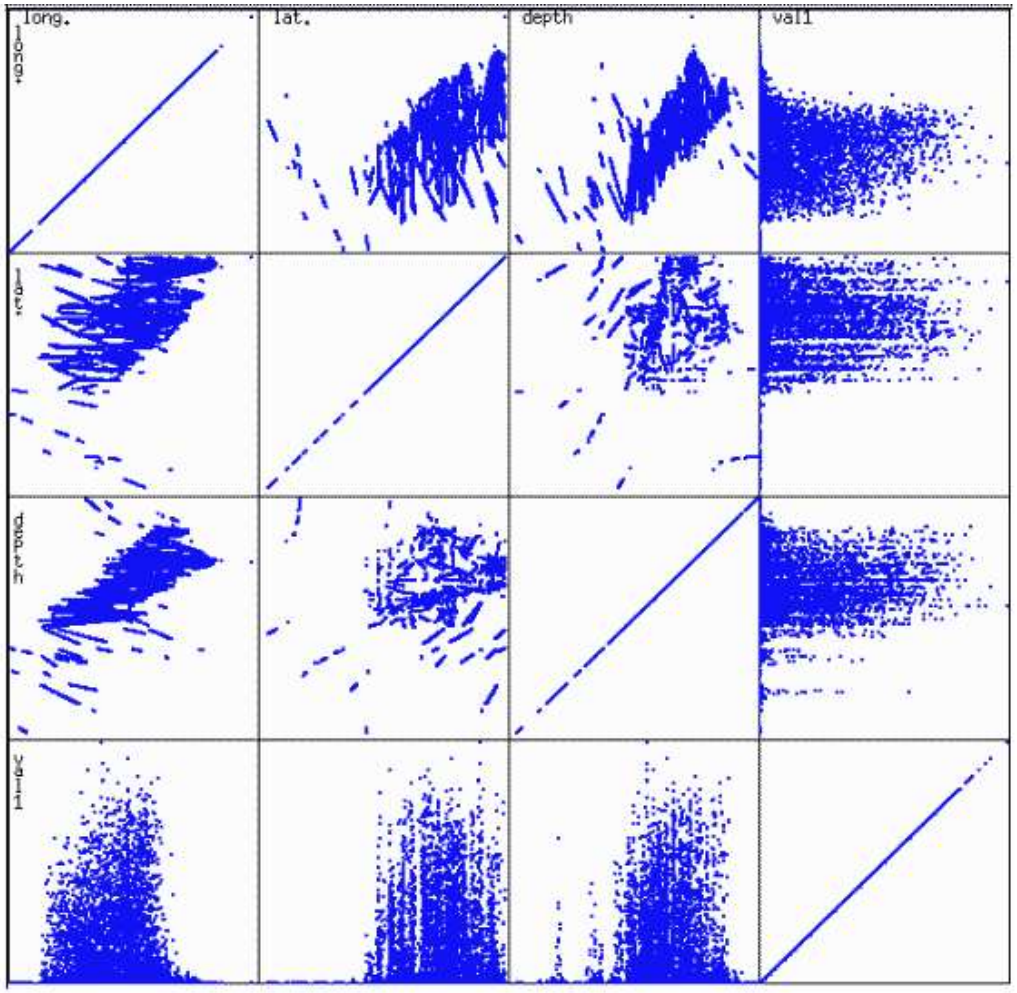


Coordenadas Paralelas



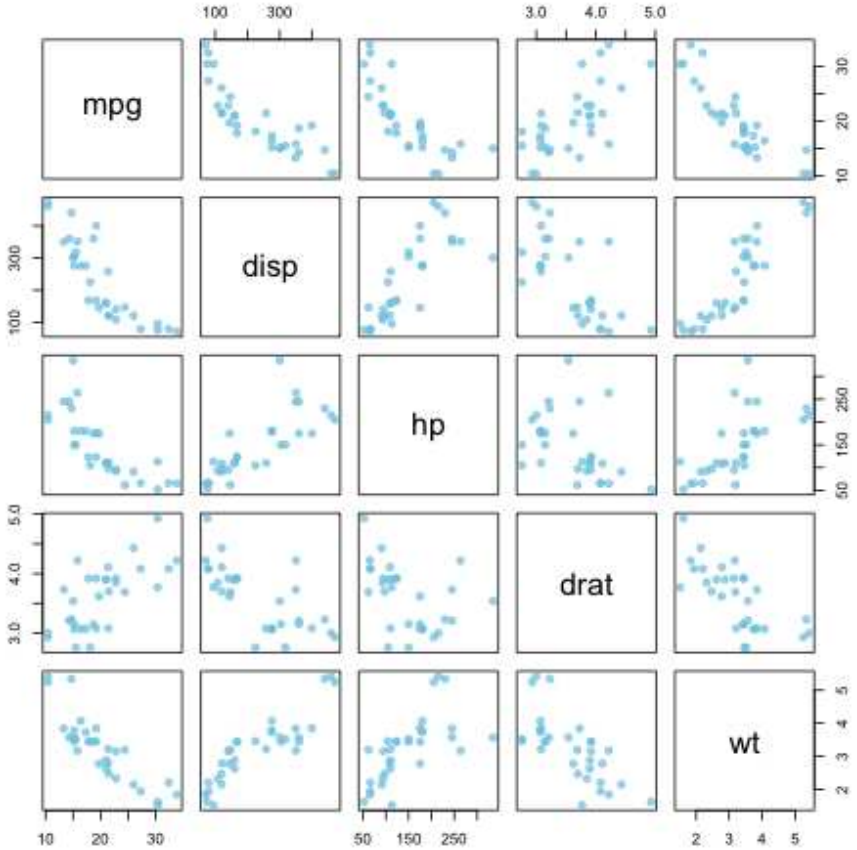
https://www.scikit-yb.org/en/latest/api/features/pcoords.html

Gráficos de Dispersão



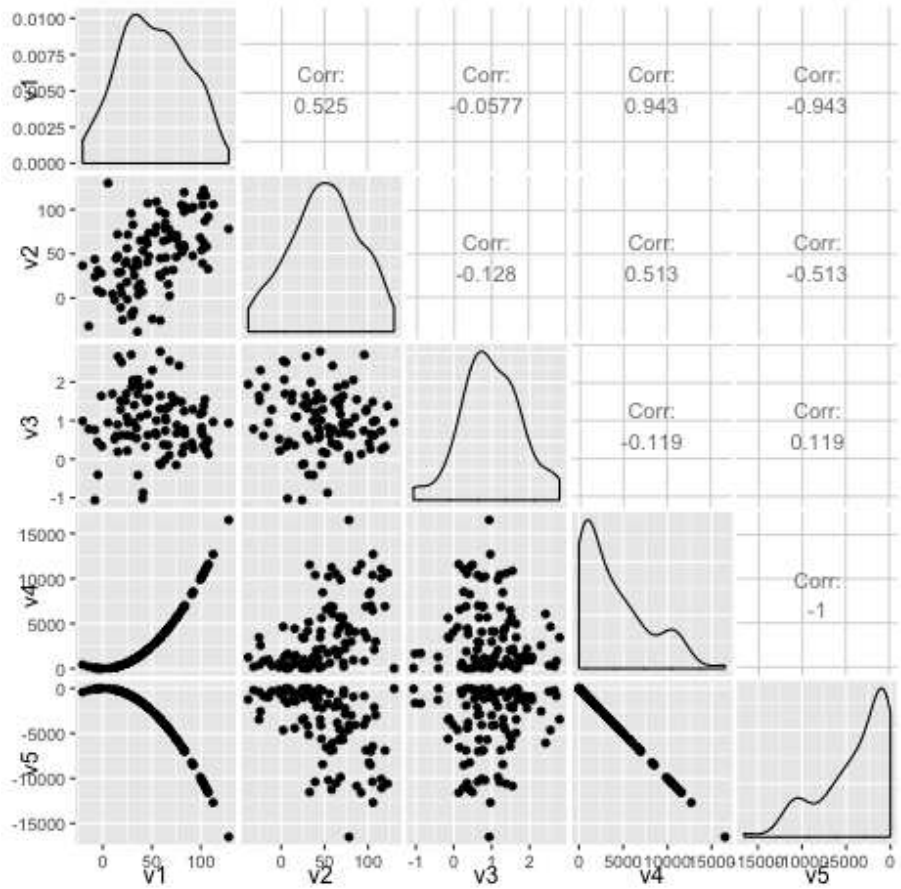
Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

Gráficos de Dispersão



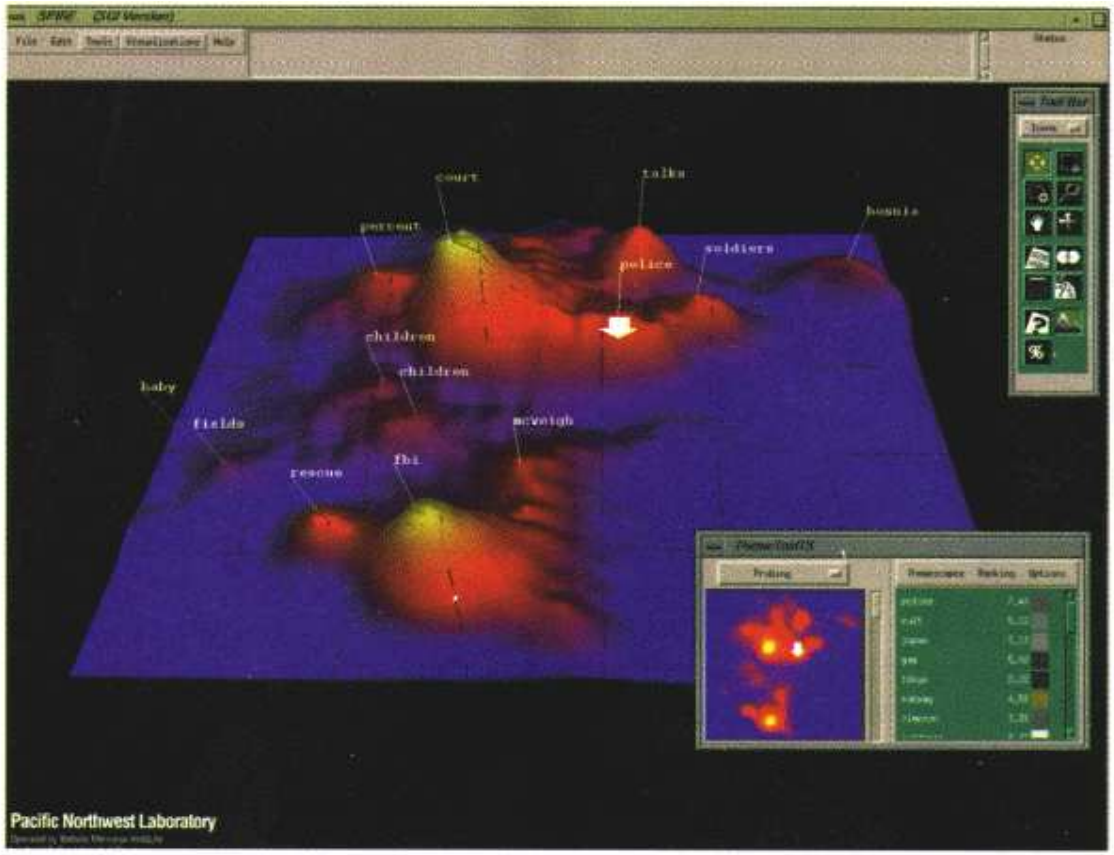
http://www.r-graph-gallery.com/98-basic-scatterplot-matrix

Gráficos de Dispersão



https://www.r-graph-gallery.com/199-correlation-matrix-with-ggally

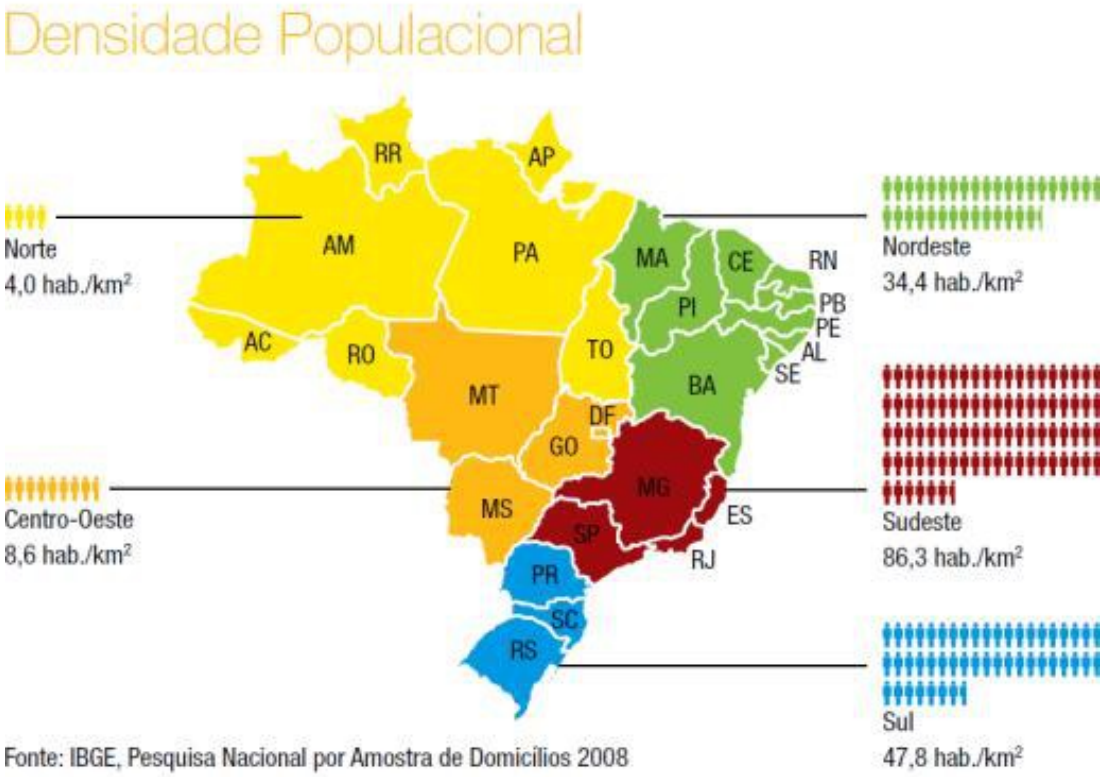
Paisagem



Used by permission of B. Wright, Visible Decisions Inc.

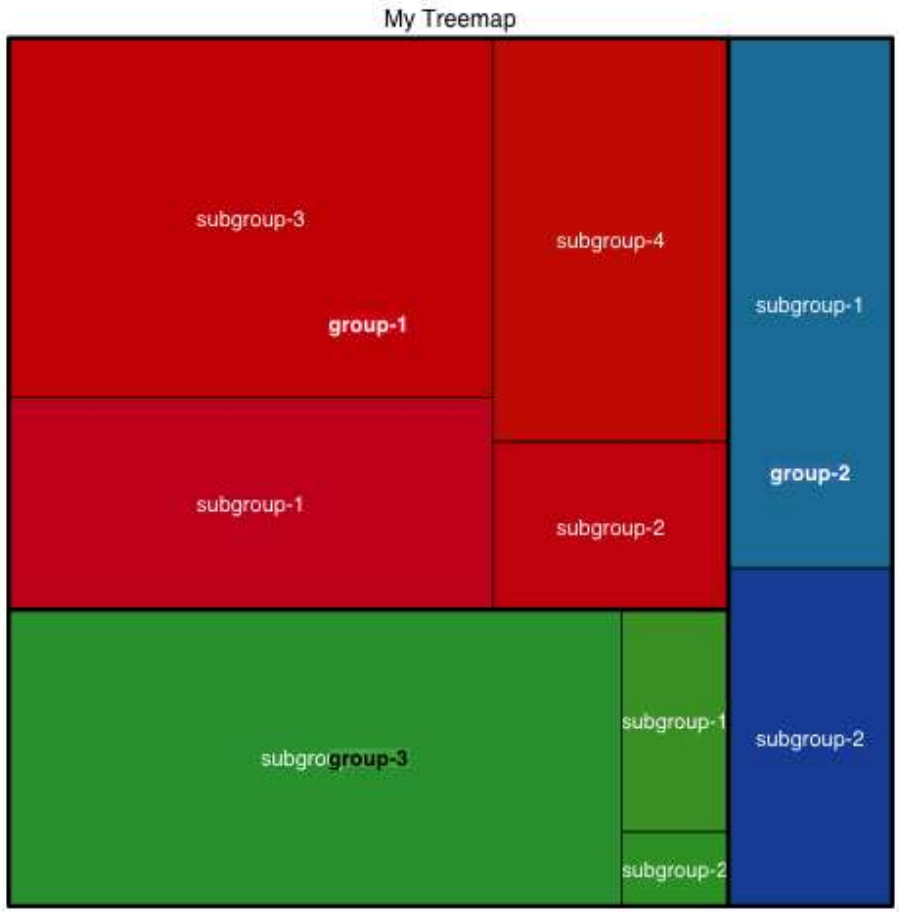
news articles visualized as a landscape

Paisagem



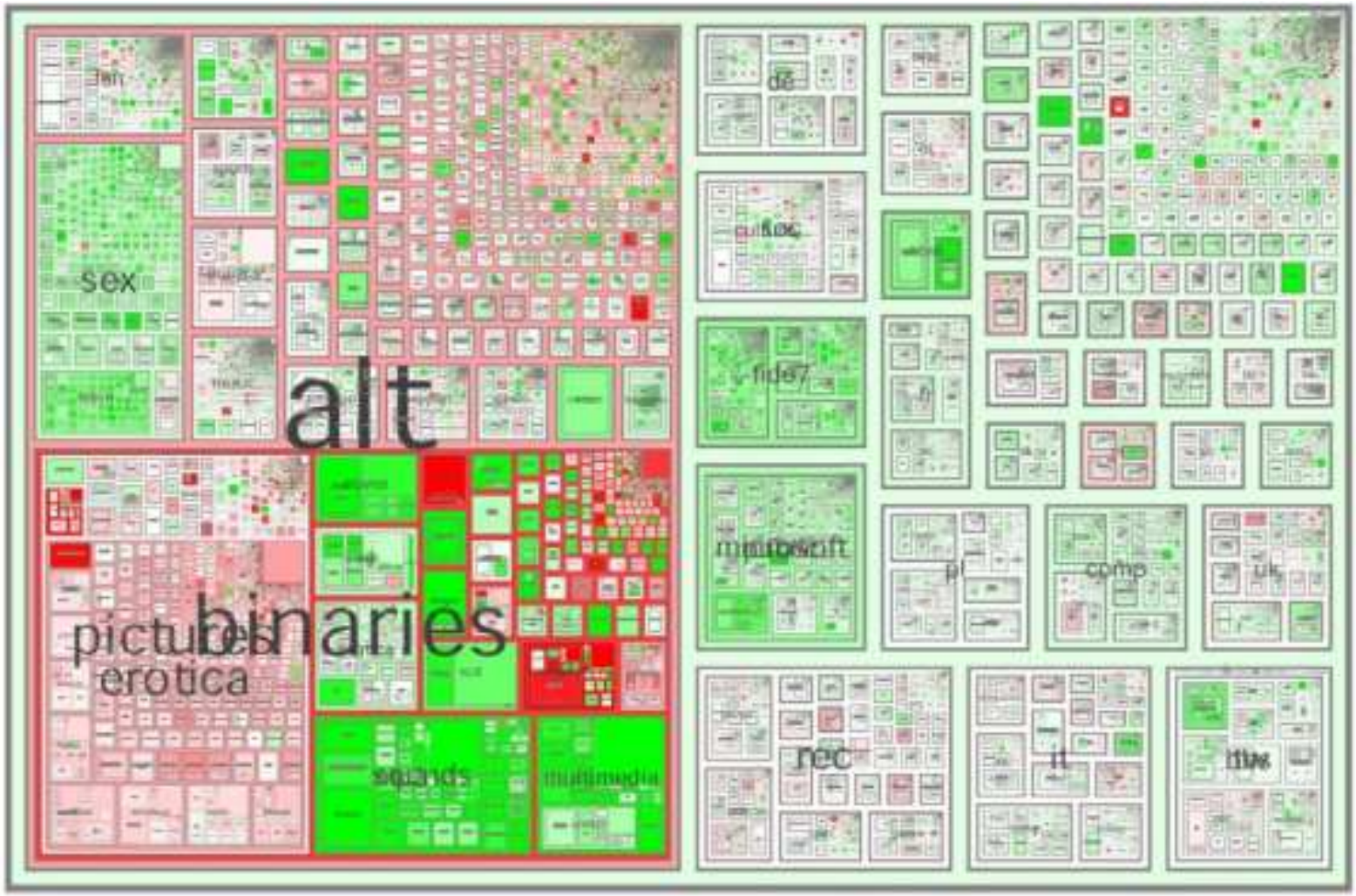
http://www.brasil.gov.br/governo/2013/06/divulgados-dados-sobr

Tree-Map



https://www.r-graph-gallery.com/236-custom-your-treemap

Tree-Map



Ack.: http://www.cs.umd.edu/hcil/treemap-history/all102001.jpg

Visualização de Dados e Relações Complexas

- Visualização de dados não numéricos: textos e redes
- Exemplo: Nuvem de *tag* (*tag cloud*), em que a importância de uma *tag* (ou palavra) é representada pelo tamanho ou cor da fonte

Grafo de Ligações

- Ligações (arestas)
- Quantidade de ligações (raio dos círculos)
- Cor pode representar o valor de algum atributo

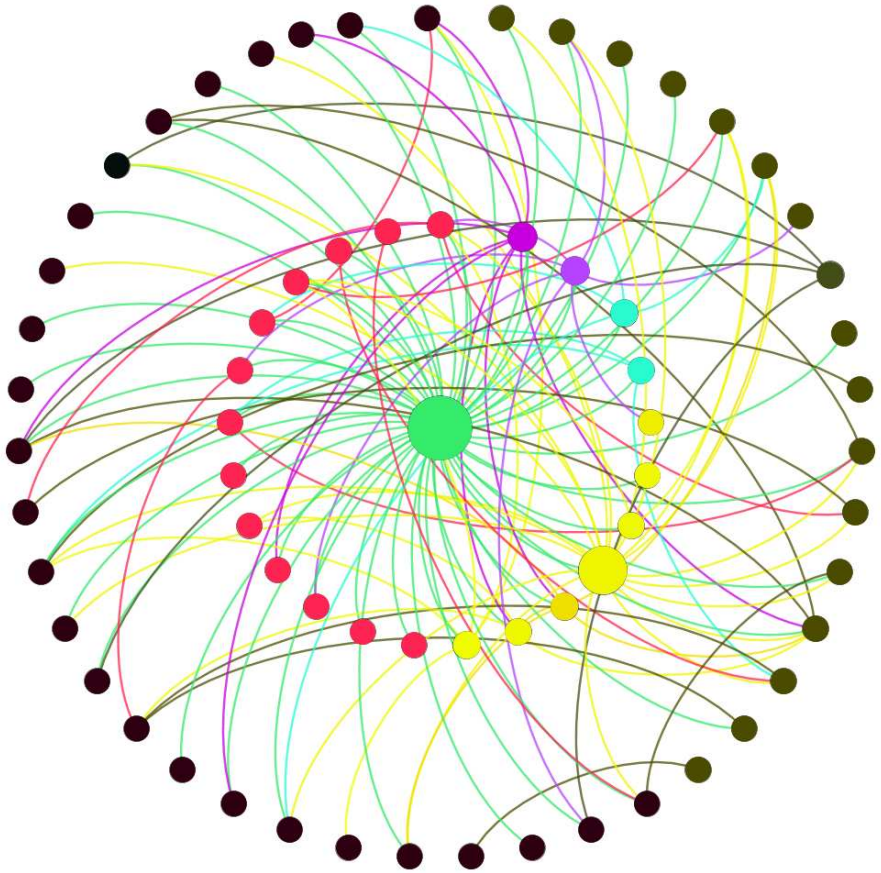
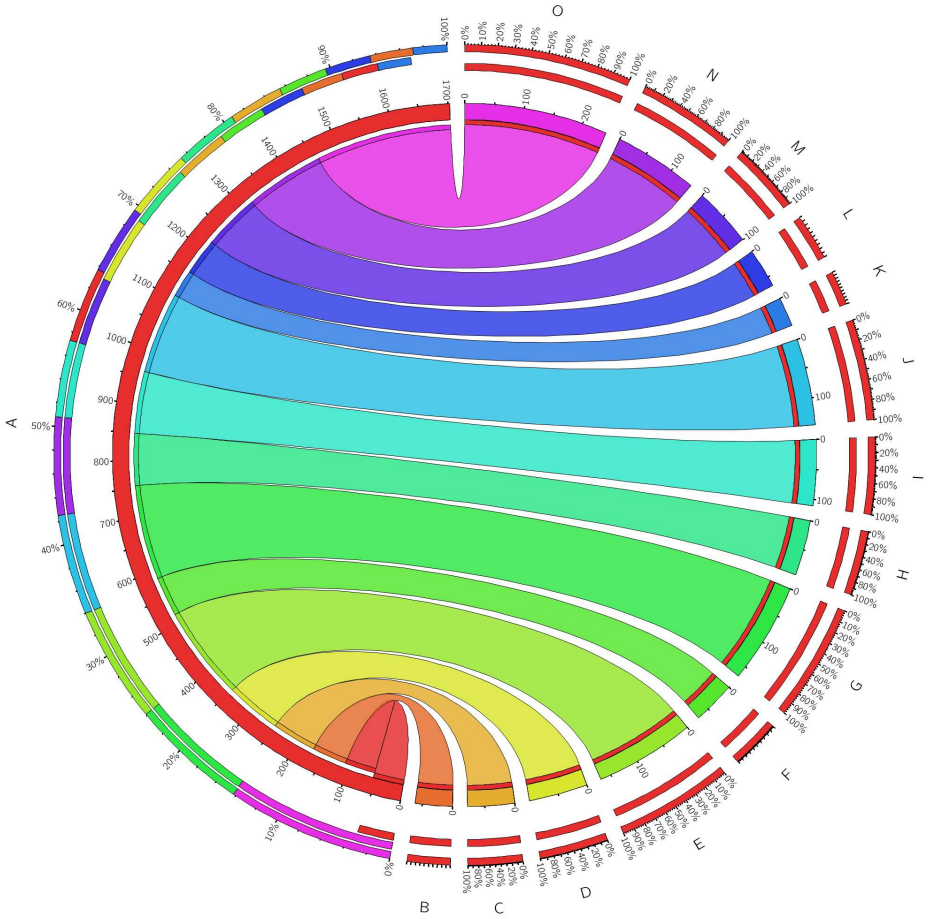
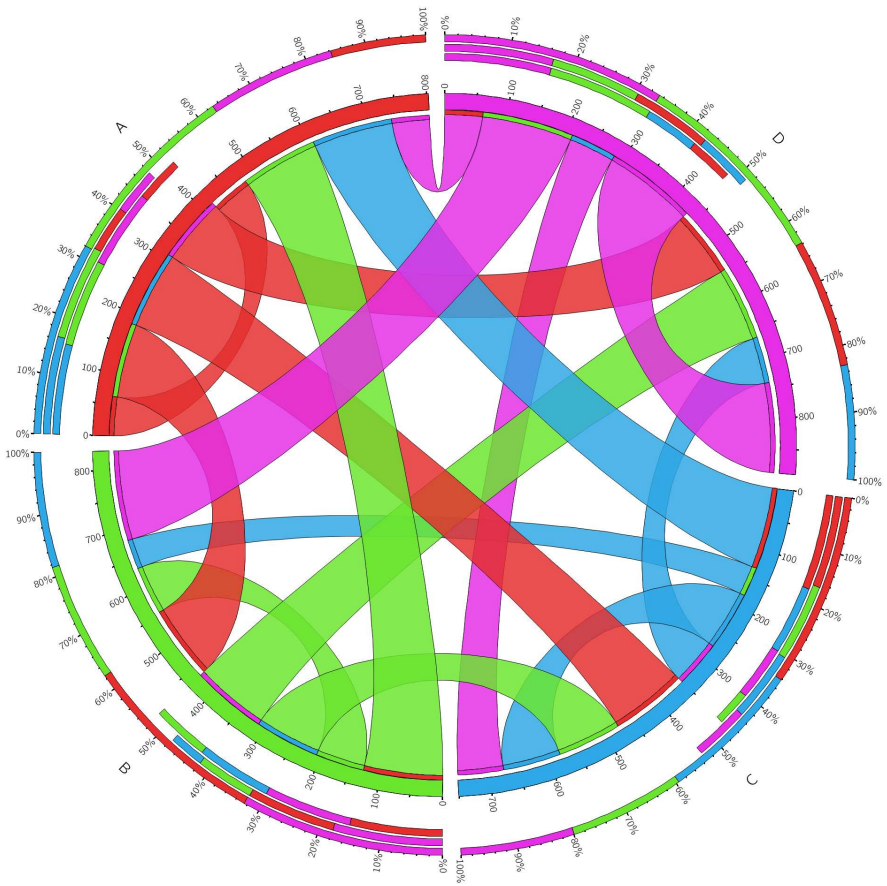


Gráfico Circular de Ligações



http://mkweb.bcgsc.ca/tableviewer

Gráfico Circular de Ligações



http://mkweb.bcgsc.ca/tableviewer

Medindo Similaridade e Dissimilaridade dos Dados

Similaridade e Dissimilaridade

- Similaridade
  - Medida numérica de quão parecidas duas instâncias são
  - O valor é frequentemente normalizado em $[0; 1]$, onde 1 indica um alto grau de similaridade
- Dissimilaridade
  - Valor que indica o quão diferentes são duas instâncias
  - O menor valor (indicando menor dissimilaridade) é frequentemente 0
  - O limite superior pode variar
- Proximidade
  - Se referente tanto a similaridade quanto a dissimilaridade

Matriz de Dados e de Dissimilaridade

- Matriz de dados

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Matriz de dissimilaridade

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

- A similaridade pode frequentemente ser avaliada em função da dissimilaridade: $\text{sim}(i,j) = 1 - d(i,j)$

Medida de Proximidade para Atributos Nominais

- Podem assumir 2 ou mais estados/valores
- Uso de uma larga quantidade de atributos binários
  - Cria-se um atributo binário para cada estado possível
- Correspondência simples
  - $d(i,j) = \frac{p-m}{p}$
  - m é o número de correspondências e p é o número de variáveis
  - $\text{sim}(i,j) = 1 - d(i,j) = 1 - \frac{p-m}{p} = \frac{p-p+m}{p} = \frac{m}{p}$

Medidas de Proximidade para Atributos Binários

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

Exemplo: Dissimilaridade em Atributos Binários

- Gênero é um atributo simétrico
- Os demais atributos são assimétricos
- Valores Y e P são assumidos como 1 e N como 0

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Distância em Dados Numéricos

- Distância de Minkowski

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

- Propriedades
  - $d(i,j) > 0$ se $i \neq j$ e 0, caso contrário
  - $d(i,j) = d(j,i)$ (simetria)
  - $d(i,j) \leq d(i,k) + d(k,j)$ (desigualdade triangular)
- Todas as distâncias que satisfazem essas propriedades são métricas
- Exemplos: Norma-1, Norma-2 e Norma infinito (máximo)

Exemplo

Data Matrix

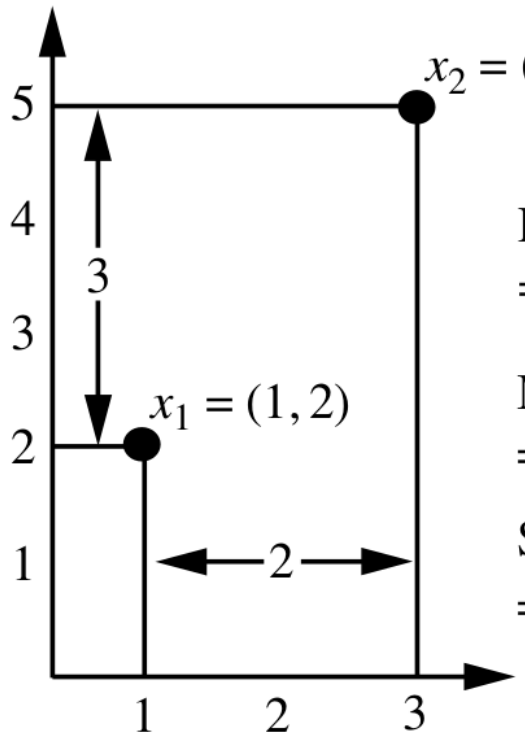
point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

Exemplo



Euclidean distance
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance
$= 2 + 3 = 5$

Supremum distance
$= 5 - 2 = 3$

Exemplo

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Normalização de Dados Numéricos

- Ajustar a escala dos valores de cada atributo de forma a mapeá-los em intervalos pré-definidos
  - Normalmente, o valor padronizado $x' \in [0; 1]$ ou $x' \in [-1; 1]$
- Linear: $x'_i = \frac{x_i - \min\{x\}}{\max\{x\} - \min\{x\}}$
  - x é o conjunto de dados a serem normalizados
  - x representa o conjunto de valores observados para um atributo
- Escore-Z (padronização): $x'_i = \frac{x_i - \mu}{\sigma}$
  - μ: média
  - σ: desvio padrão
- Soma: $x'_i = \frac{x_i}{\sum_j x_j}$
- Máximo: $x'_i = \frac{x_i}{\max\{x\}}$

Variáveis Ordinárias

- Podem ser discretas ou contínuas
- Independente do tipo, requer que os dados possam ser ordenados
- Podem ser tratadas como valores num intervalo
  - Substituir x_i pela sua posição numa lista ordenada (iniciando em 1)
  - Mapear a variável em $[0; 1]$: $x'_i = \frac{x_i - 1}{|x| - 1}$
  - $|\cdot|$ é o operador de cardinalidade de conjunto
- A dissimilaridade pode ser calculada como definido para valores numéricos

Similaridade Cosseno

- Um documento pode ser representado por diversos atributos, onde cada um representa a frequência de uma palavra (palavras-chaves) ou frase

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Outros vetores: características gênicas em *micro-arrays*
- Aplicações: recuperação de informação, taxonomia biológica, ...
- Cosseno: $\cos(\mathbf{d_1}, \mathbf{d_2}) = \frac{\mathbf{d_1} \cdot \mathbf{d_2}}{||\mathbf{d_1}|| \, ||\mathbf{d_2}||}$
onde $\mathbf{d_1}$ e $\mathbf{d_2}$ são vetores.

Exemplo: Cosseno

$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$
$\cos(d_1, d_2) = 0.94$

Atributos de Tipos Mistos

- Um banco de dados pode conter atributos de diversos tipos
- Uma alternativa para combinar os efeitos desses atributos é usar uma soma ponderada

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - p é o número de atributos
  - $\delta_{ij}^{(f)} = 0$ se o valor de f para alguma das instâncias i ou j não está disponível, ou $x_{if} = x_{jf} = 0$ e f é um atributo binário assimétrico
  - caso contrário, $\delta_{ij}^{(f)} = 1$
  - $d_{ij}^{(f)}$ é a distância (padronizada em $[0; 1]$) entre as instâncias i e j, em relação ao atributo f

Exemplo

Table 2.2 A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	*test-1 (nominal)*	*test-2 (ordinal)*	*test-3 (numeric)*
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Exemplo

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

test-1

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

test-2

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

test-3

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$