Curso 2 – CD, AM e DM

Profa. **Roseli Ap. Francelin Romero**

MBA em Inteligência Artificial e BigData

Depto. de Ciências de Computação
ICMC - USP

CONTEÚDO

- INTRODUÇÃO
- **PARTE I** – EXPLORAÇÃO
  EXEMPLO
- **PARTE II** – PRE-PROCESSAMENTO
  EXEMPLOS
- **PARTE III** - ANÁLISE DE
  EXPERIMENTOS
- **PARTE IV** – MODELAMENTO

Cientistas de Dados: O que fazem?

- Cientistas de dados são os grandes mineradores de dados. Eles recebem uma enorme massa de dados desorganizados (estruturados e não estruturados) e usam suas habilidades em matemática, estatística e programação para limpar, tratar e organizá-los.
- Em seguida, eles aplicam suas capacidades analíticas – conhecimento da indústria, compreensão contextual, ceticismo de suposições existentes – para descobrir soluções para os desafios de negócios ocultos.

Cientistas de Dados: O que fazem?

- Entre suas principais responsabilidades estão:

1 - Realizar pesquisas sem direção e formular perguntas abertas aos dados

2 - Extrair grandes volumes de dados de múltiplas fontes internas e externas

3 - Empregar os programas de análise sofisticadas, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e descritiva.

DADOS

Estruturados

Mais facilmente analisados por técnicas de MD

Ex.: Planilhas e tabelas atributo-valor

Não estruturados

Ex.: Conteúdo de página na web, emails, vídeos, sequencia de DNA, ...

1) PONTO INICIAL: Escolha um Problema

- Escolha algo que o entusiasme, como um projeto de análise musical do Spotify
- Projeto de análise de aluguel em uma cidade

2) Pense nos diferentes passos

Coleta de Dados

Análise de Dados

Visualização:

Implantação

Kaggle
Web Scraping
API

Os dados podem estar bagunçados

Existem diferentes ferramentas

Desenvolvimento WEB Interface

Análise de Dados

- EXTRAIR CONHECIMENTO DOS DADOS
- REALIZAR A INTERPRETAÇÃO
- TOMAR AÇÕES

PARTE I

Exploração dos Dados

- Gerar Hipóteses
- Entendimento por meio de técnicas
- Reavaliar as Hipóteses
- Vantagens e desvantagens de técnicas
- Sumarizar as informações

Conjuntos de dados

- Estruturados
  - Mais facilmente analisados por técnicas de MD
  - Ex.: Planilhas e tabelas atributo-valor
- Não estruturados
  - Mais facilmente analisados por seres humanos
    - Para DM, são geralmente convertidos em dados estruturados
  - Ex.: Sequência de DNA, conteúdo de página na web, emails, vídeos, ...

Conjuntos de dados estruturados

Atributos de entrada (preditivos)

Exemplos (objetos, instâncias)

Nome	Temp.	Idade	Peso	Altura	Diagnóstico
João	37	70	94	190	Saudável
Maria	38	65	60	172	Doente
José	39	19	70	185	Doente
Sílvia	38	25	65	160	Saudável
Pedro	37	70	90	168	Doente

Atributo alvo

Hipóteses – Caso 1

- Base de dados de ANP (Agencia Nacional de Petróleo) volumes produzidos mensalmente em cada Poço.
- <class 'pandas.core.frame.DataFrame'> Int64Index: 35477 entries, 0 to 30325 Data columns (total 40 columns):

- **Avaliação do Grau API (Tipo do Petróleo) - por Tipo de Produção (Mar, Terra ou Pré-Sal):**

A maior produção é do tipo Pré-Sal?
O poço que mais produziu Petróleo é do tipo Terra?

API	Petróleo (Tipo)
<15	Asfáltico
15-19	Extra-Pesado
19-27	Pesado
27-33	Médio
33-40	Leve
40-45	Extra-Leve
>45	Condensado

HIPÓTESES – CASO 2 - Modelo de Negócio

HIPÓTESES – CASO 2

BASE DO KAGGLE

- Pedidos, produtos, entrega e *reviews*;
- Entre 2016 a 2018;
- Possui 9 tabelas, 100000 objetos e 50 atributos;
- Feature alvo: review_score: (1, 2, 3, 4 e 5).

Hipóteses – Caso 2

- Existe alguma relação entre o local de entrega e a nota dada pelos clientes?
- Quanto maior a distância de entrega maior a chance da avaliação ser negativa?
- Existe relação entre o tempo de atraso e a nota dada pelos clientes?
- Pedidos mais caros tem maior probabilidade de resultar em uma compra bem sucedida?
- Produtos mais pesados são mais difíceis de transportar por isso podem gerar mais ocorrências de avaliações ruins?
- Existem categorias de produtos mais propensas a resultar em baixa avaliação?

Tipos de atributos

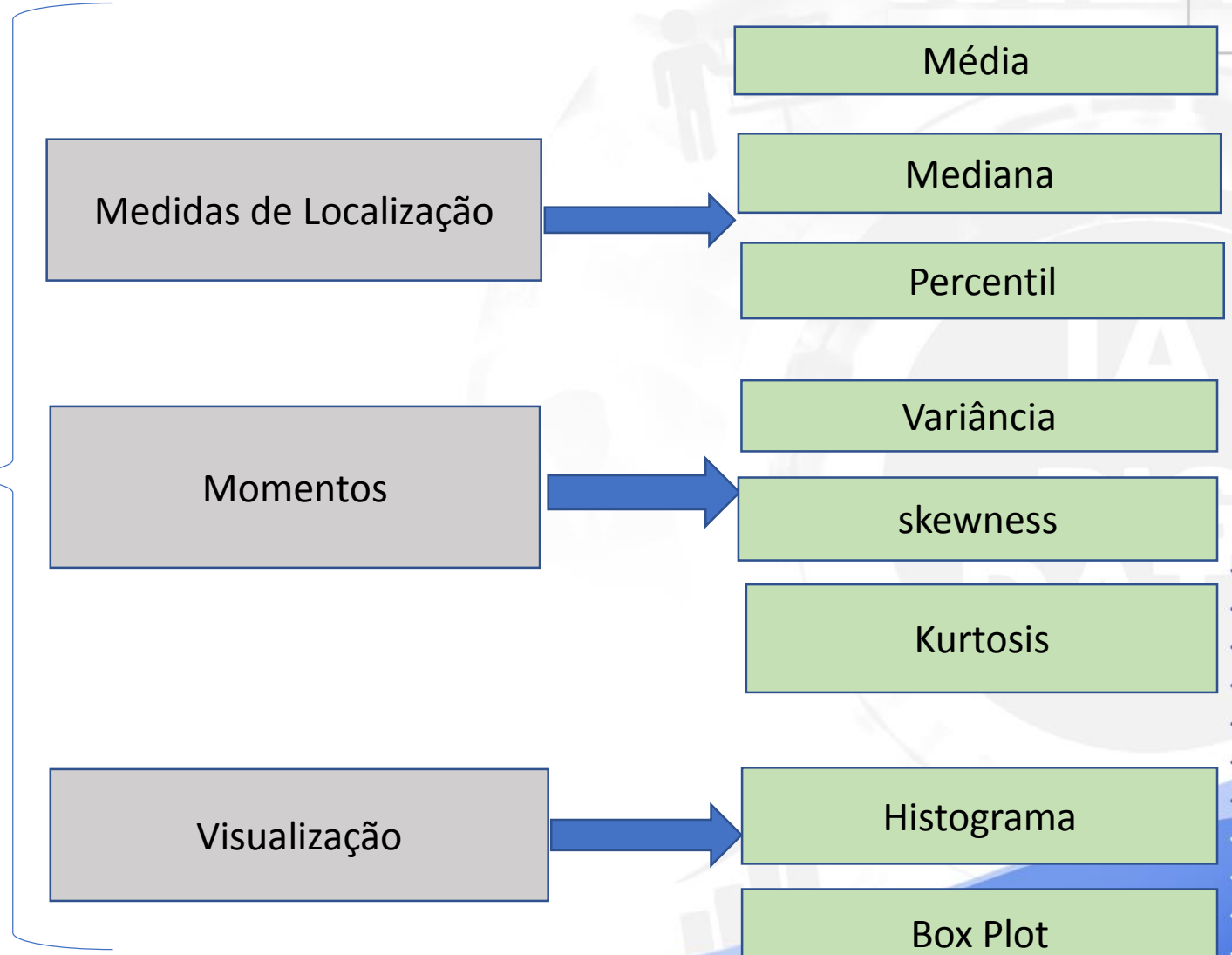
- Simbólicos ou qualitativos
  - Nominal ou categórico
    - Ex.: cor, código de identificação, profissão
  - Ordinal
    - Ex.: gosto (ruim, médio, bom), dias da semana
- Numéricos, contínuos ou quantitativos
  - Intervalar
    - Ex.: data, temperatura em Celsius
  - Racional
    - Ex.: peso, tamanho, idade

TIPOS DE ATRIBUTOS – Caso 1

Tipo	Produção dos Poços de Petróleo (Brasil)	
	Tipo de dado (dtype)	*Quant. de Colunas*
Nominal	objeto	4
Ordinal	objeto	6
Intervalar	datetime64	1
Racional	float64 ou int64	29

1º.ETAPA

EXPLORAÇÃO



Média

- Pode ser calculada facilmente

$$média(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Problema: sensível a *outliers*

Mediana

- Menos sensível a *outliers* que média
- Necessário ordenar valores

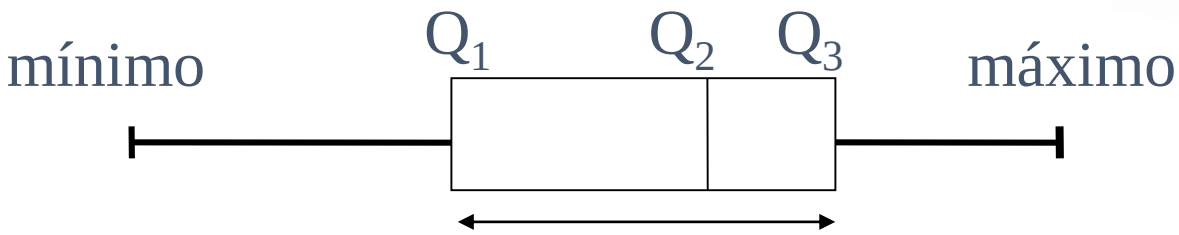
$$\text{mediana}\,(x) = \tilde{x} = \begin{cases} x_{(r+1)} \text{ se n é ímpar } (n = 2r+1) \\ \frac{1}{2}(x_r + x_{(r+1)}) \text{ se n é par } (n = 2r) \end{cases}$$

Média *versus* Mediana

- Média é uma boa medida de localização quando os valores estão distribuídos simetricamente
- Mediana indica melhor o centro
  - Se distribuição é oblíqua (assimétrica)
    - *Skewed*
  - Se existem *outliers*

Boxplot

- Gráfico que resume informações dos quartis



Intervalo entre quartis

Quartis e Percentis

- Mediana divide os dados ao meio
  - No entanto, pontos de localização diferentes podem ser usados
  - Quartis dividem um conjunto ordenado de dados em quartos
    - Q_1 : Primeiro quartil (quartil inferior)
      - Valor da observação para a qual 25% dos dados do conjunto tem valor menor ou igual
      - Também é o valor do 25º percentil
    - Q_2 : Segundo quartil = mediana
    - Q_3 : Terceiro quartil (quartil superior
      75º percentil)

Cálculo dos percentis

- Ordenar os valores
  - Posição do p-percentil:

$$posição = \left\lceil p \times n + \frac{1}{2} \right\rceil$$

    - Arredonda posição para o valor inteiro seguinte (21,5 = 22)
    - Retorna o valor nessa posição

Boxplot modificado

- Identifica *outliers* e reduz seu efeito no formato do boxplot
  - Tolerância = 1,5 x intervalo entre quartis
  - Verificar:
  - Se (máximo – Q_3 > tolerância) ou
    (Q_1 – mínimo < tolerância)
    então: Valor fora do intervalo é considerado *outlier*
  - Define novo mínimo e/ou máximo

INTERQUARTIL (IQR)

- IQR= Q3 – Q1
- Representa 50% dos dados do conjunto
- Ajuda a encontrar outliers

Cerca Inferior(LF) = Q1 – 1.5 IQR

Cerca Superior(UF) = Q3 + 1.5 IQR

INTERQUARTIL (IQR)

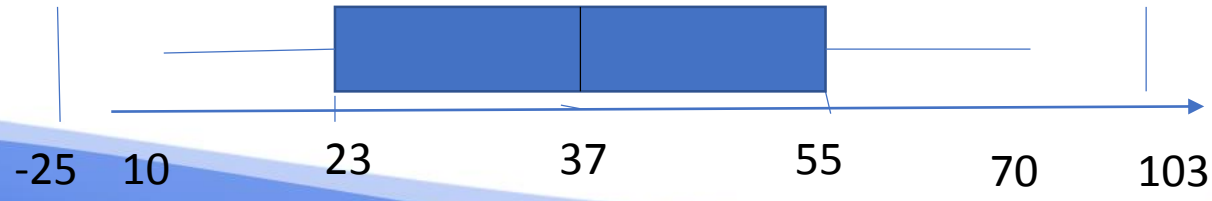
Q1: 0,25*13 + 0,5 = 3,75 -> Q1=23
Q2 = 37
Q3 = 0,75*13 + 0,5 = 10,25 -> Q3 = 55

- Exercício: Considere a lista:

(10 12 23 23 25 35 37 45 46 55 56 67 70)

- Montar o BoxPlot correspondente e determinar LF e UF.
- Determinar se existe outliers

IQR = 55 – 23 = 32
LF = Q1 – 1,5*IQR = 23 – 1,5*32 = -25
LU = Q3 + 1,5*IQR = 55 + 1,5*32 = 103

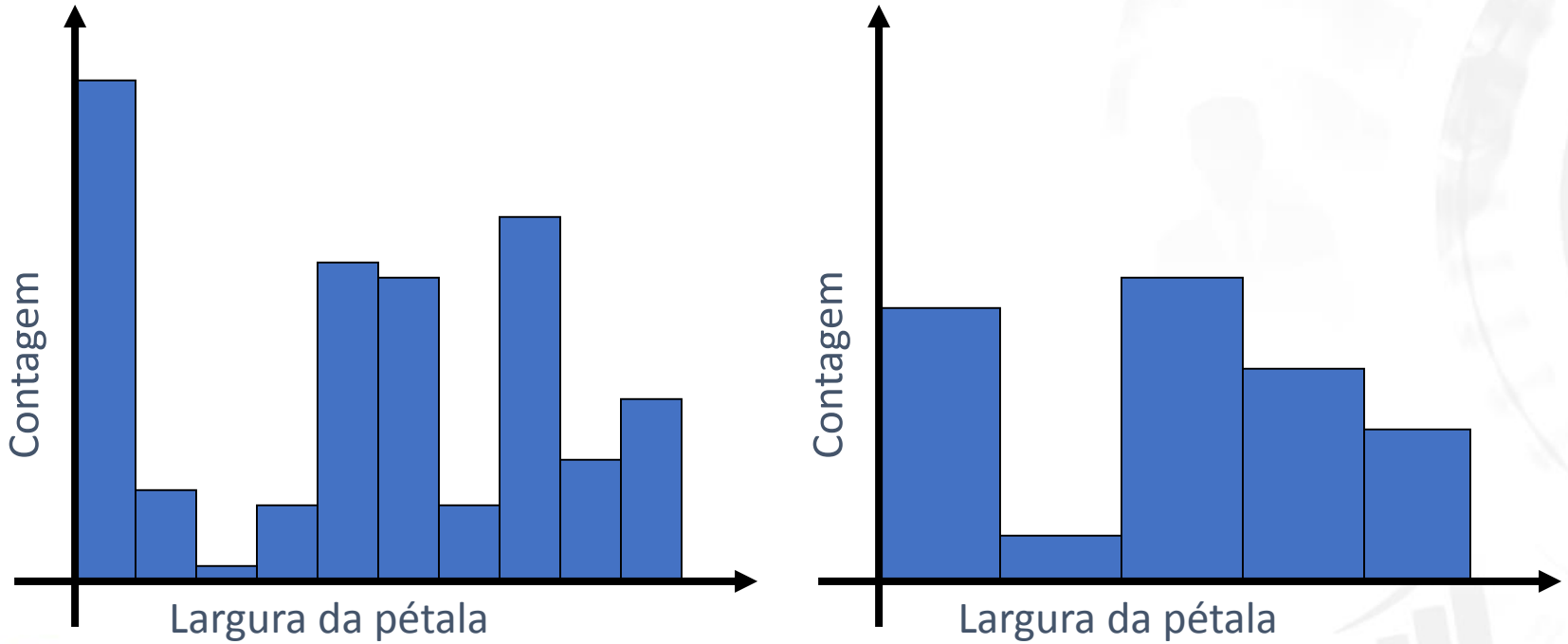


Frequência

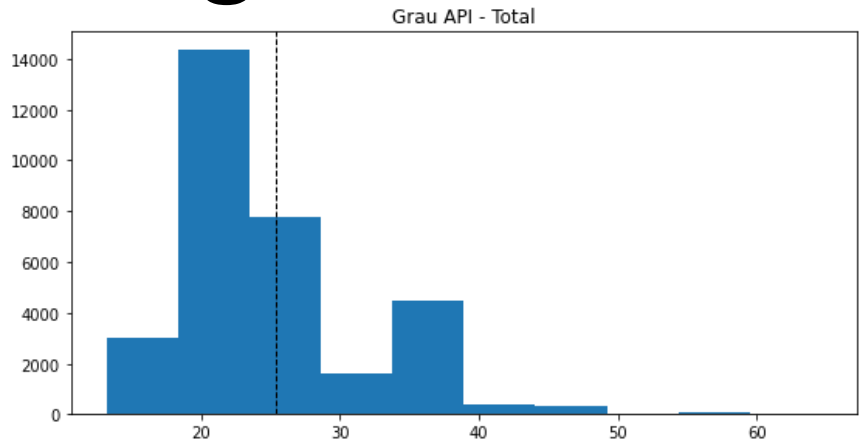
- Proporção de vezes que um atributo assume um dado valor
  - Em um determinado conjunto de dados
  - Muita usada para dados categóricos
  - Ex.: Em um BD de um hospital, 40% dos pacientes é maior de idade

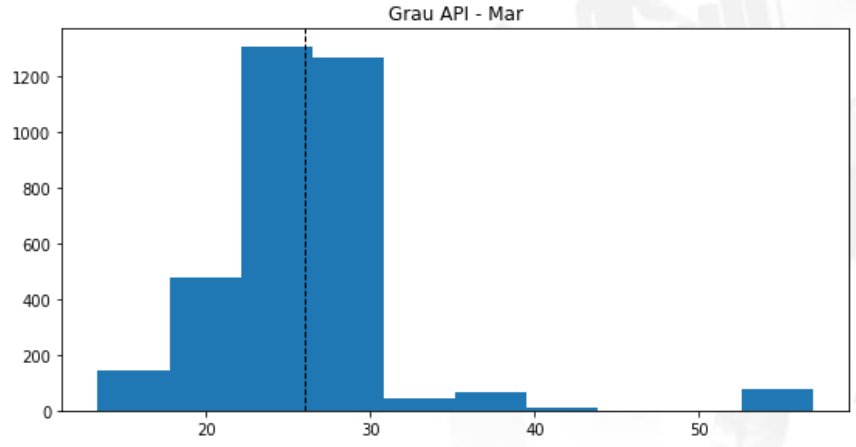
Histogramas

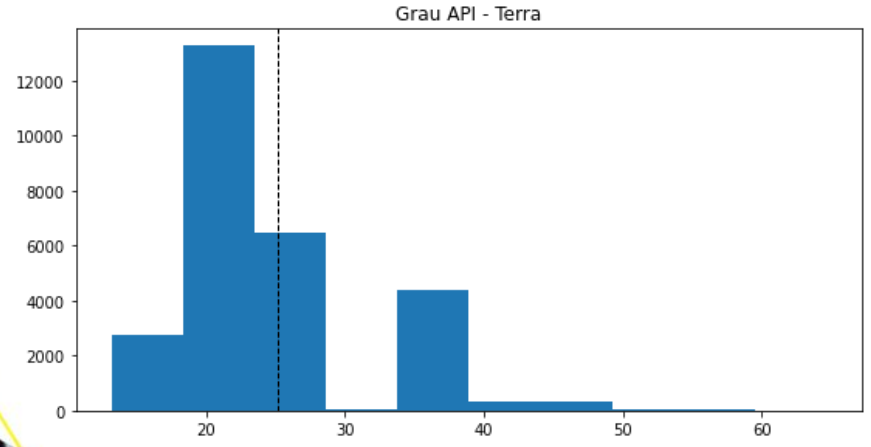
- Conjunto de dados Iris
  - Largura das pétalas usando 10 e 5 cestas

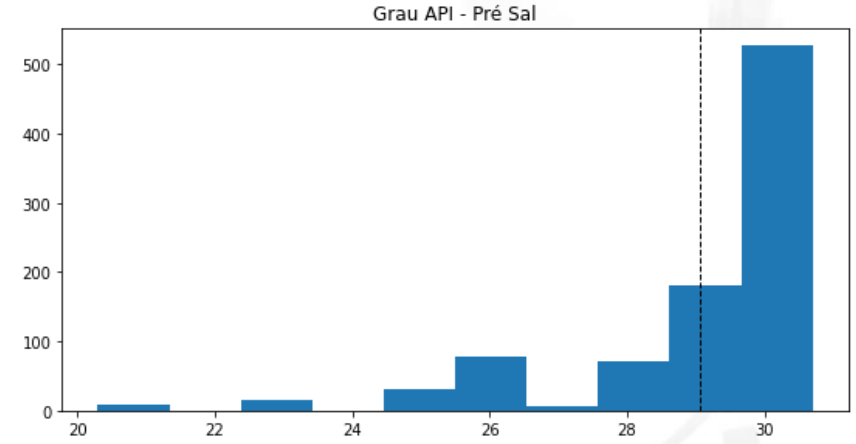


Histograma – Base de Dados de Petróleo



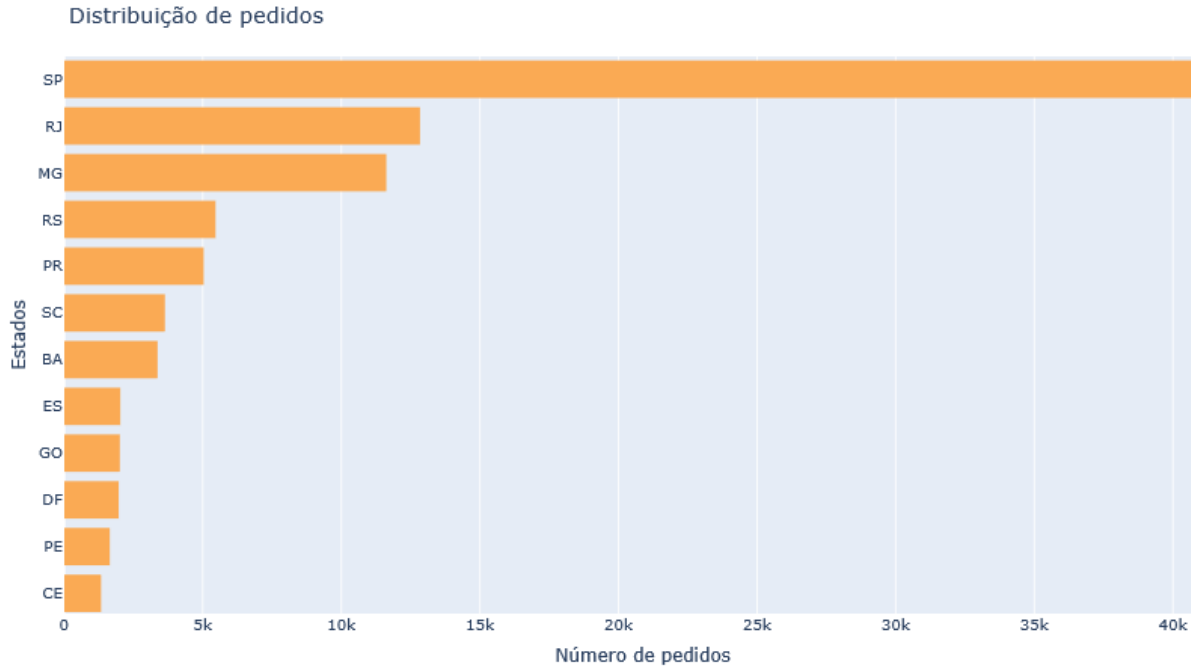


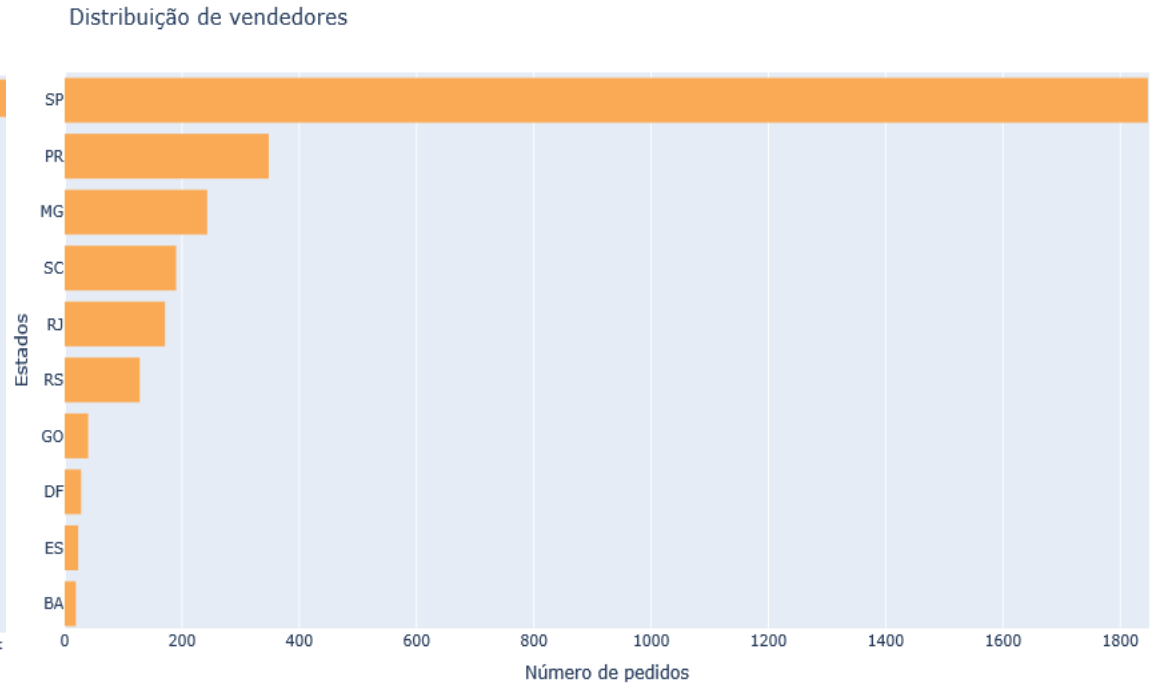




API	Petróleo (Tipo)
<15	Asfáltico
15-19	Extra-Pesado
19-27	Pesado
27-33	Médio
33-40	Leve
40-45	Extra-Leve
>45	Condensado

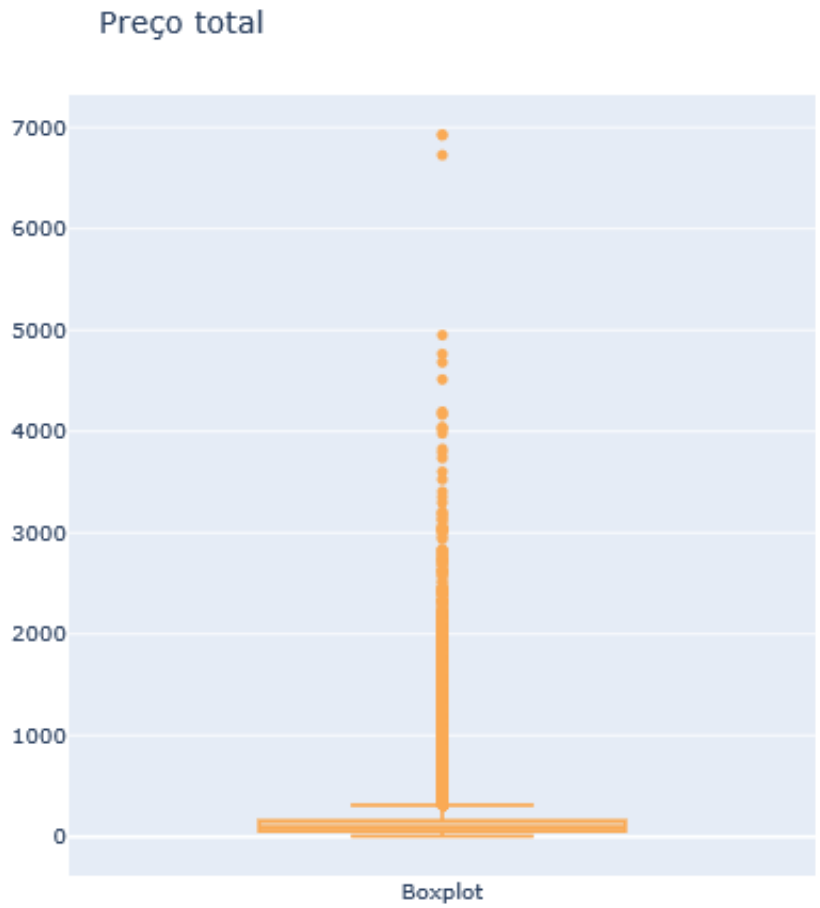
HISTOGRAMA - BASE NEGÓCIOS





Presença de Outliers

Valores da compra



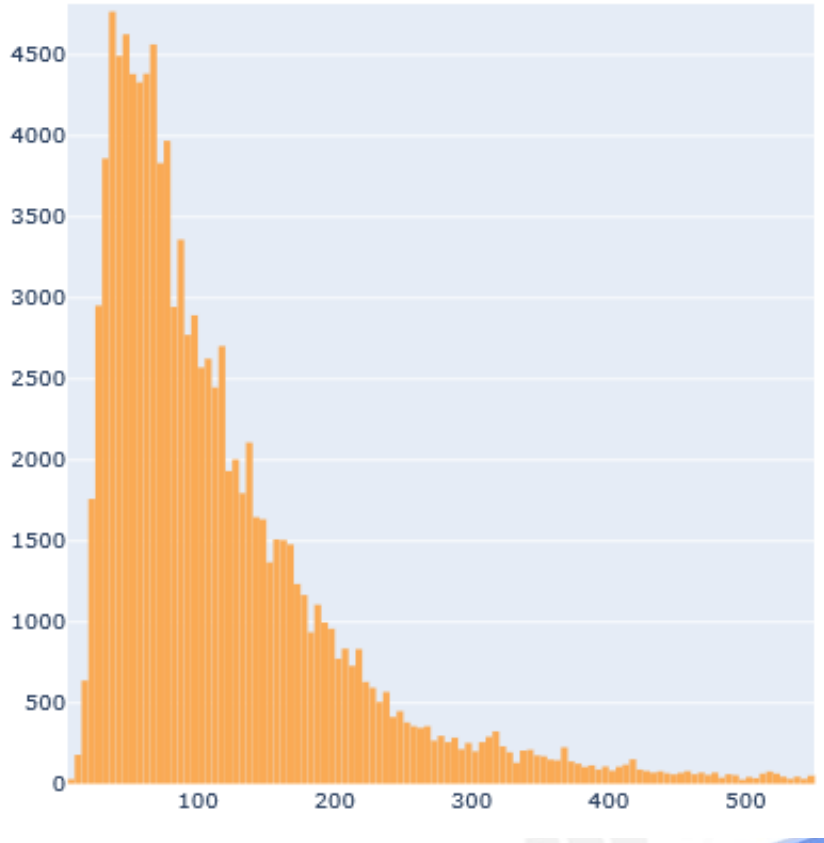


Diagrama de torta

- Frequências relativas podem ser vistas no diagrama circular

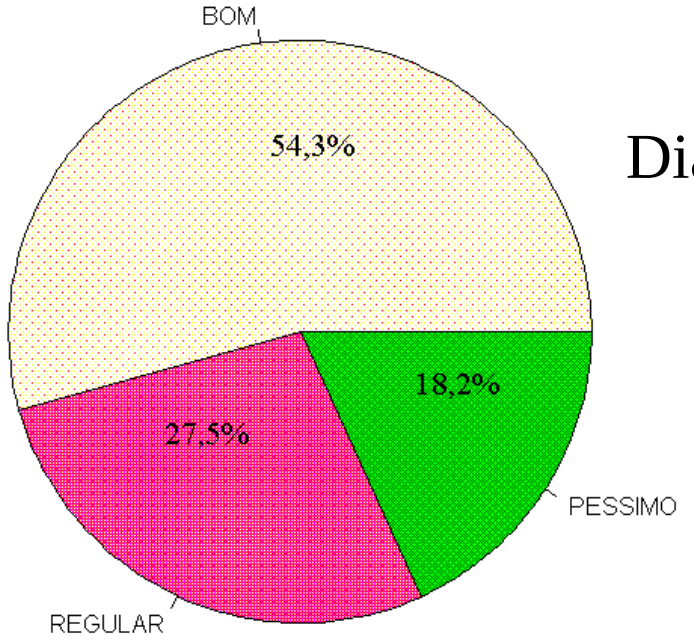


Diagrama de torta (pizza)

Medidas de distribuição

- Definem como os valores de uma variável (atributo) estão distribuídos
- Calculada por meio de momentos
  - Medida quantitativa usada na estatística e na mecânica
  - Captura o formato da distribuição de um conjunto de valores

Variância

- Medida mais utilizada para analisar espalhamento de valores

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - Denominador *n-1*: correção de Bessel, usada para uma melhor estimativa da variância verdadeira
    - Amostra (estimada) e população (verdadeira)
  - Desvio padrão: raiz quadrada da variância
  - Um dos momentos de uma distribuição de probabilidade

VARIÂNCIA

- "o quão longe" em geral os seus valores se encontram do **valor esperado** (média) da variável aleatória *X*.
- Desvio Padrão indica qual é o “erro” se quiséssemos substituir um dos valores coletados pelo valor da média.

Momento central

- Centralizado ou centrado
  - K=1: média = 0 (primeiro momento em torno da média = primeiro momento central)
  - K=2: variância (segundo momento central)
  - K=3: obliquidade (terceiro momento central)
  - K=4: curtose (quarto momento central)

$$\mu_k = E[x - E(x)]^k = \sum_{i=1}^{n} (x_i - \bar{x})^k \, p(x_i) = \sum_{i=1}^{n} (x_i - \bar{x})^k \, f(x_i)$$

$$\mu_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{(n-1)}$$

Assumindo cada x_i aparece com a mesma frequência

Momento padronizado

- Fornece informações mais claras sobre a distribuição dos dados
  - Utiliza distribuição normal padrão
    - Normaliza o k-ésimo momento pelo desvio padrão elevado a k
      - Torna a medida independente de escala

$$\mu_k^{'} = \frac{\mu_k}{\sigma^k}$$

Em torno da média

Momento padronizado

- Primeiro momento (K=1):
  - Média = 0
- Segundo momento (K=2):
  - Variância = 1

$$\mu_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{(n-1)\sigma^2}$$

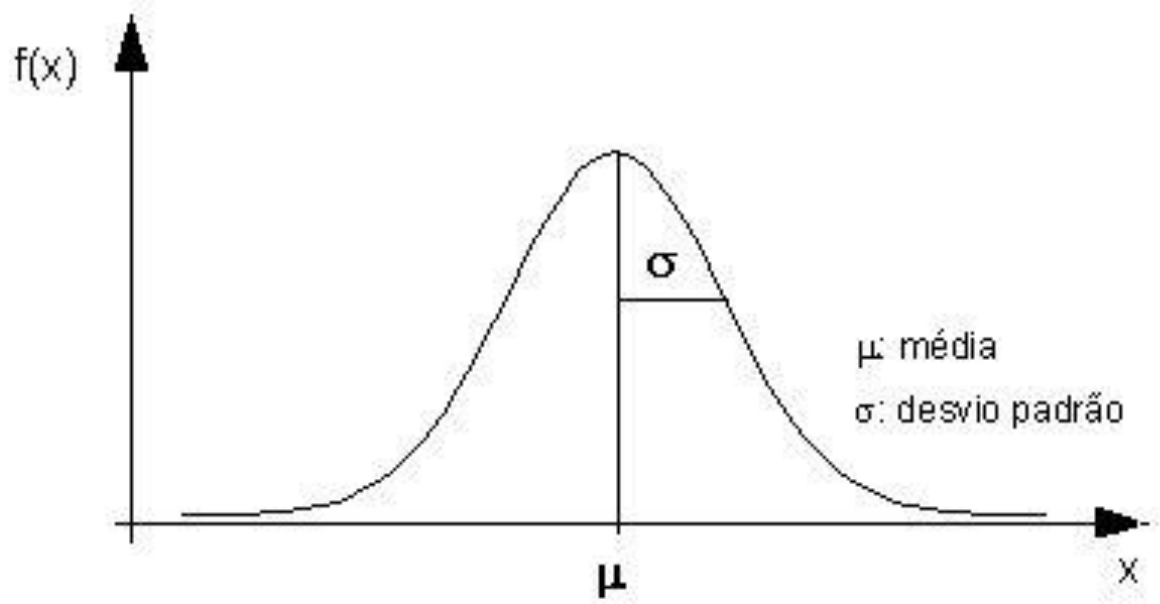
Obliquidade

- Terceiro momento (*Skewness*)
  - Mede a simetria da distribuição dos dados em torno da média
    - Distribuição simétrica tem a mesma aparência à direita e à esquerda do ponto central

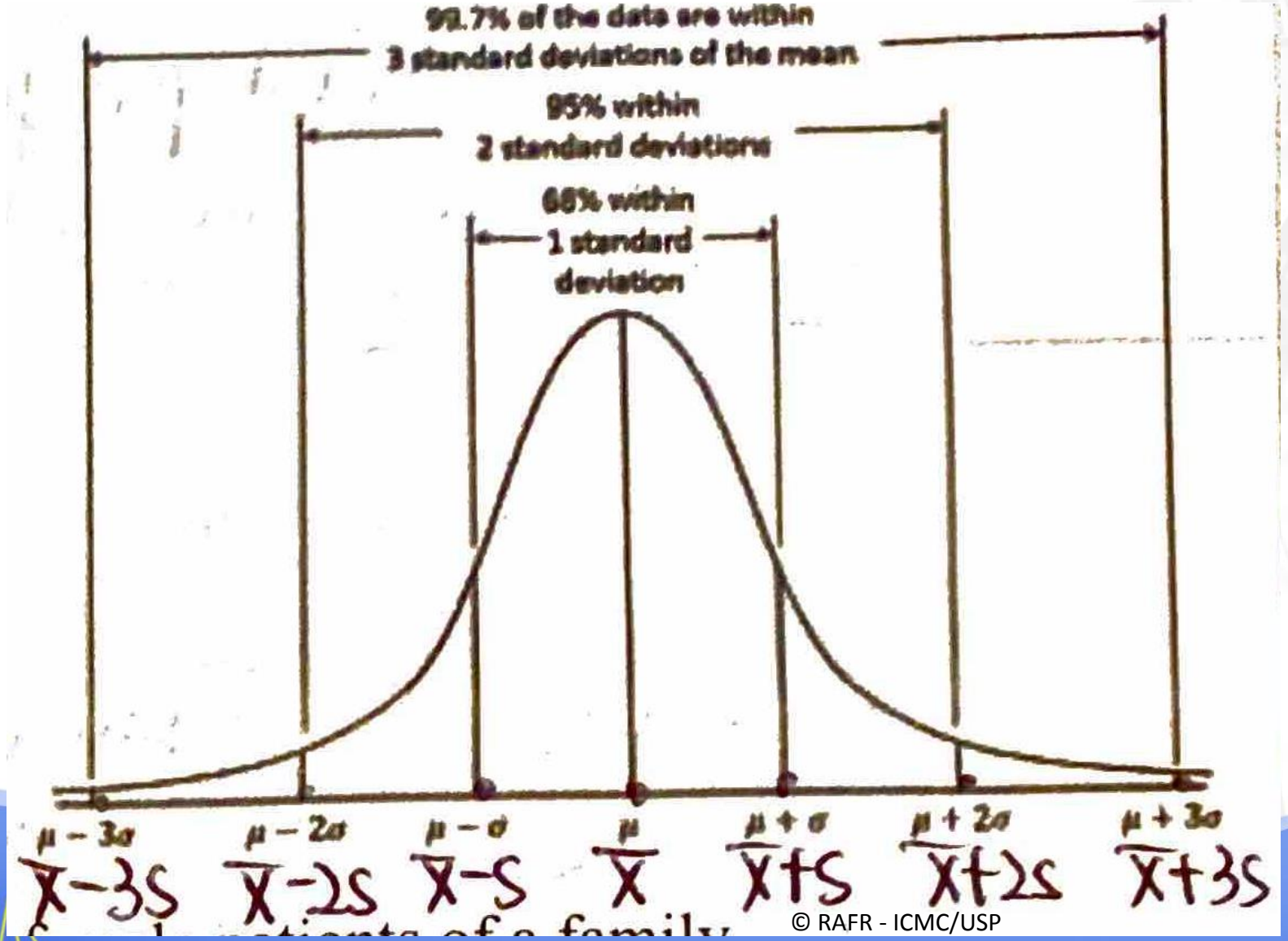
$$Obl = \mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)\sigma^3}$$

$$\mu_3 = \frac{1}{\sigma_3}\sum_{i=1}^{n}(x_i - \bar{x})^3\, p(x_i) = \frac{1}{\sigma_3}\sum_{i=1}^{n}(x_i - \bar{x})^3\, f(x_i)$$

Distribuição normal



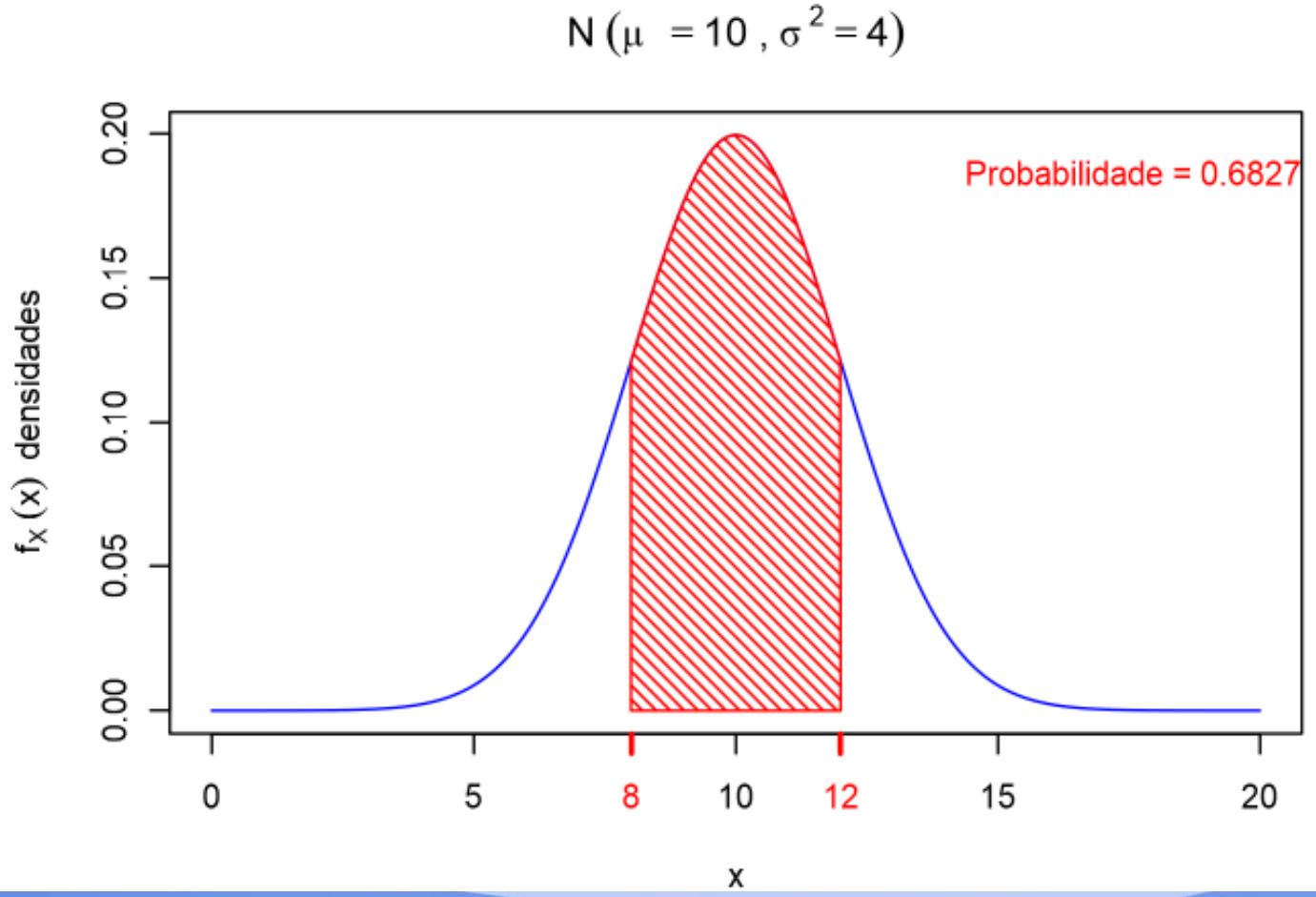
Normal – Regra Empírica



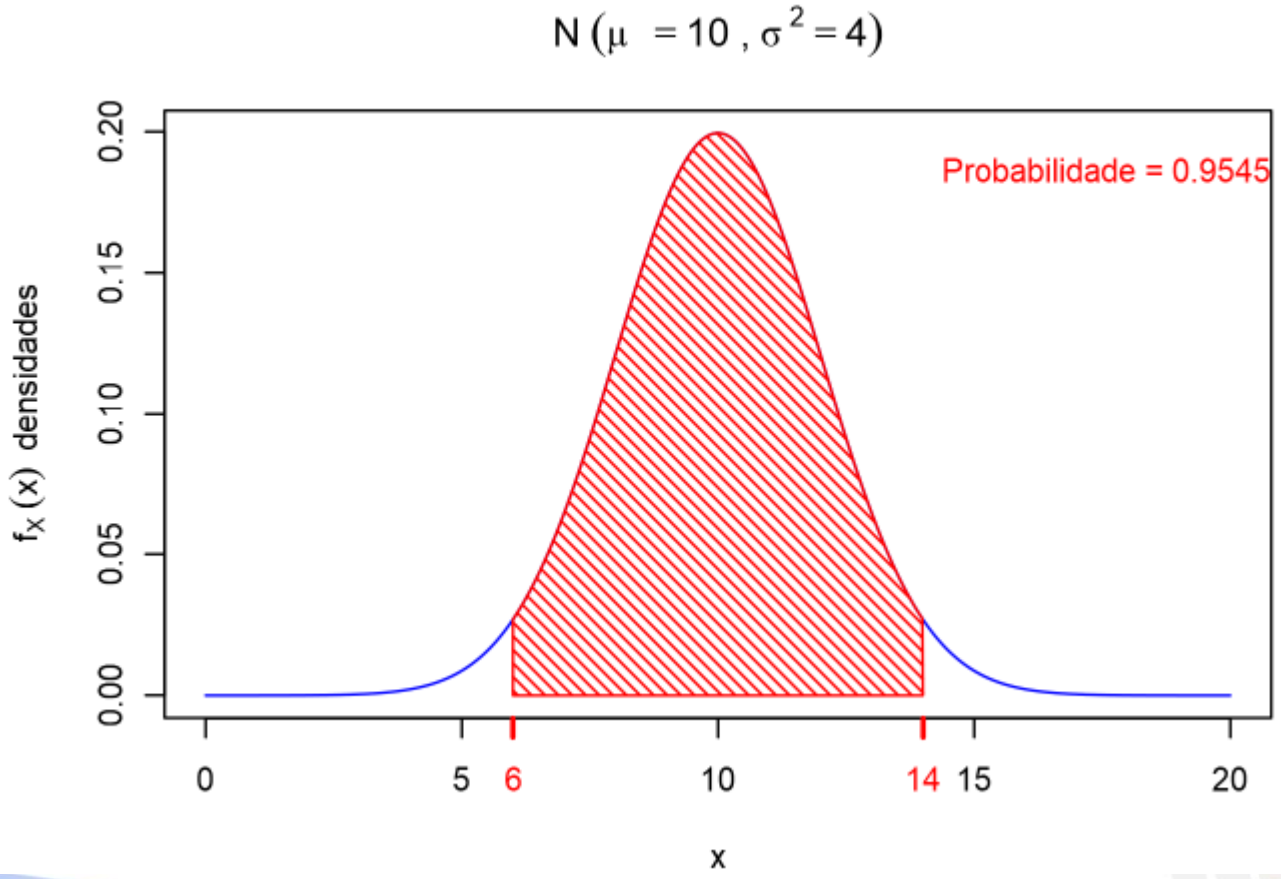
Interessante na distr. Normal

- $P(\mu - \sigma < X < \mu + \sigma) = 0.68$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.99$

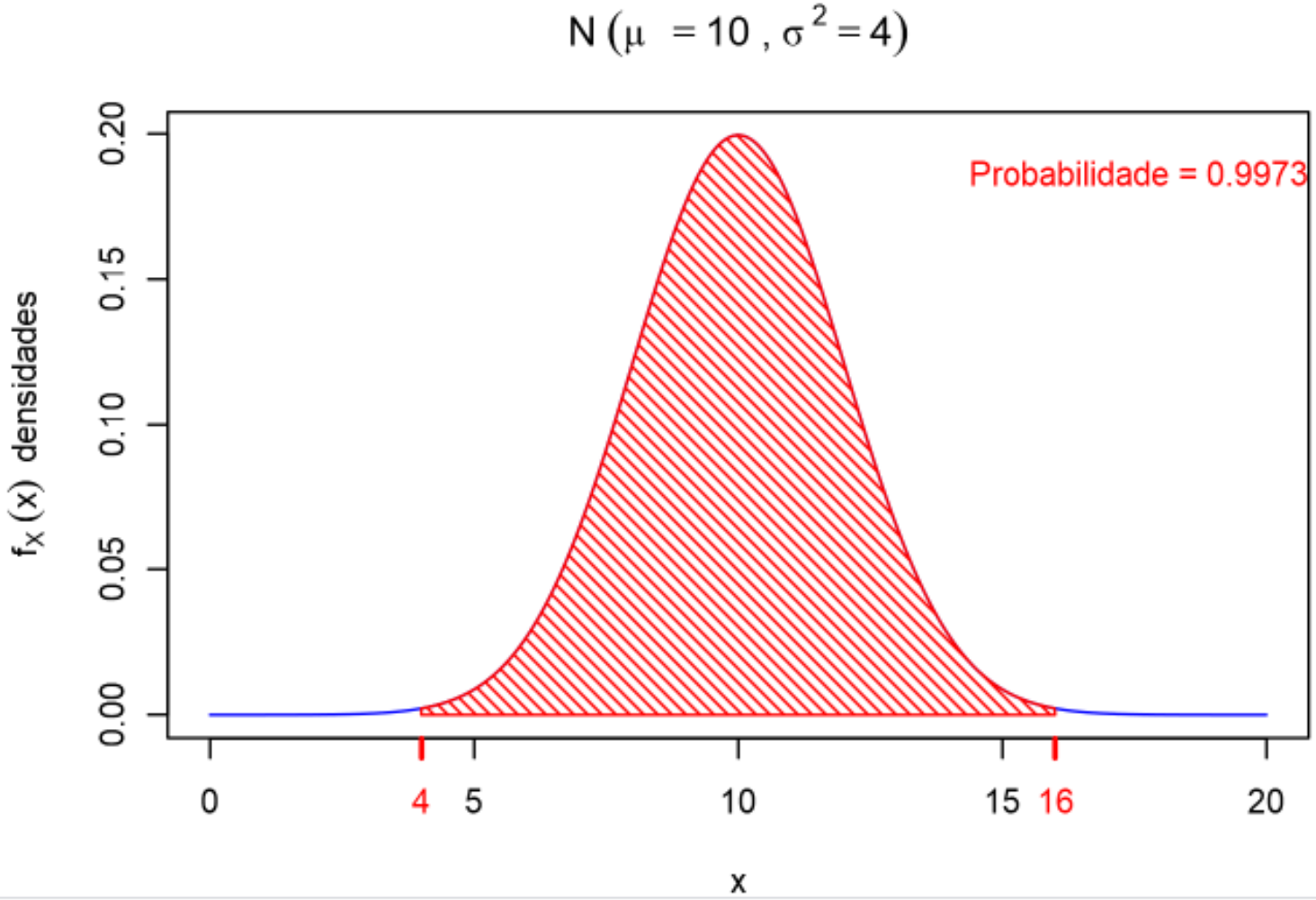
Exemplos de distr. Normal



Exemplos de distr. Normal



Exemplos de distr. Normal



Kurtosis

- Quarto momento (*Kurtosis*)
  - Medida de dispersão que captura o achatamento da função de distribuição
    - Verifica se os dados apresentam um pico elevado ou são achatados em relação a uma distribuição normal

$$Curt = \mu_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)\sigma^4}$$

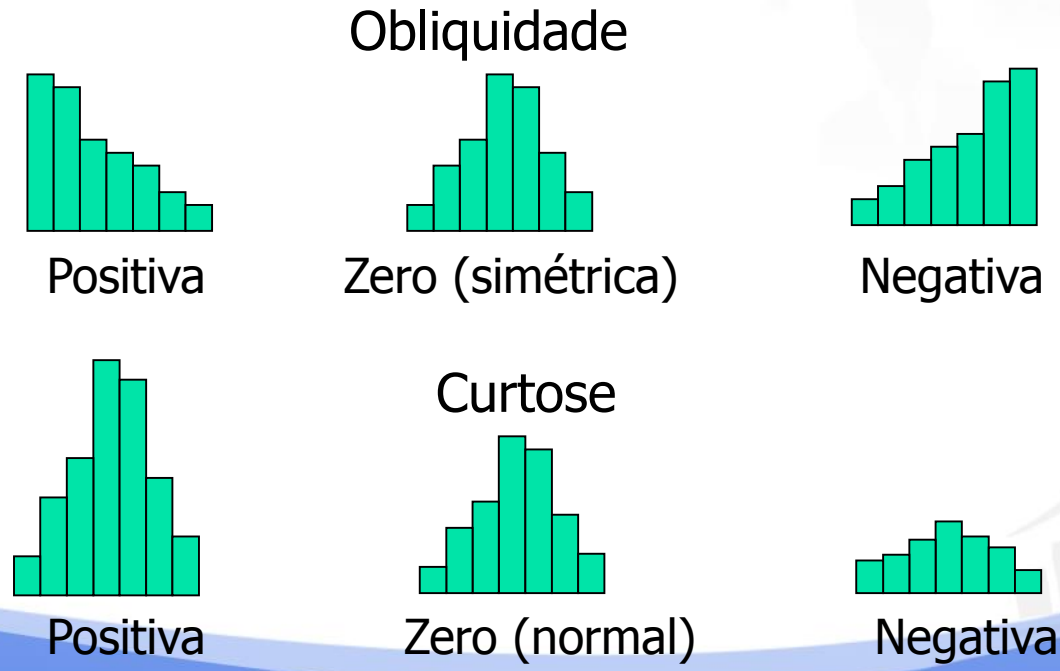
Kurtosis

- Para uma distribuição normal padrão (média = 0 e desv. pad. = 1), *Curt = 3*
- Para que a distribuição normal padrão tenha curtose = 0, usa-se a correção:

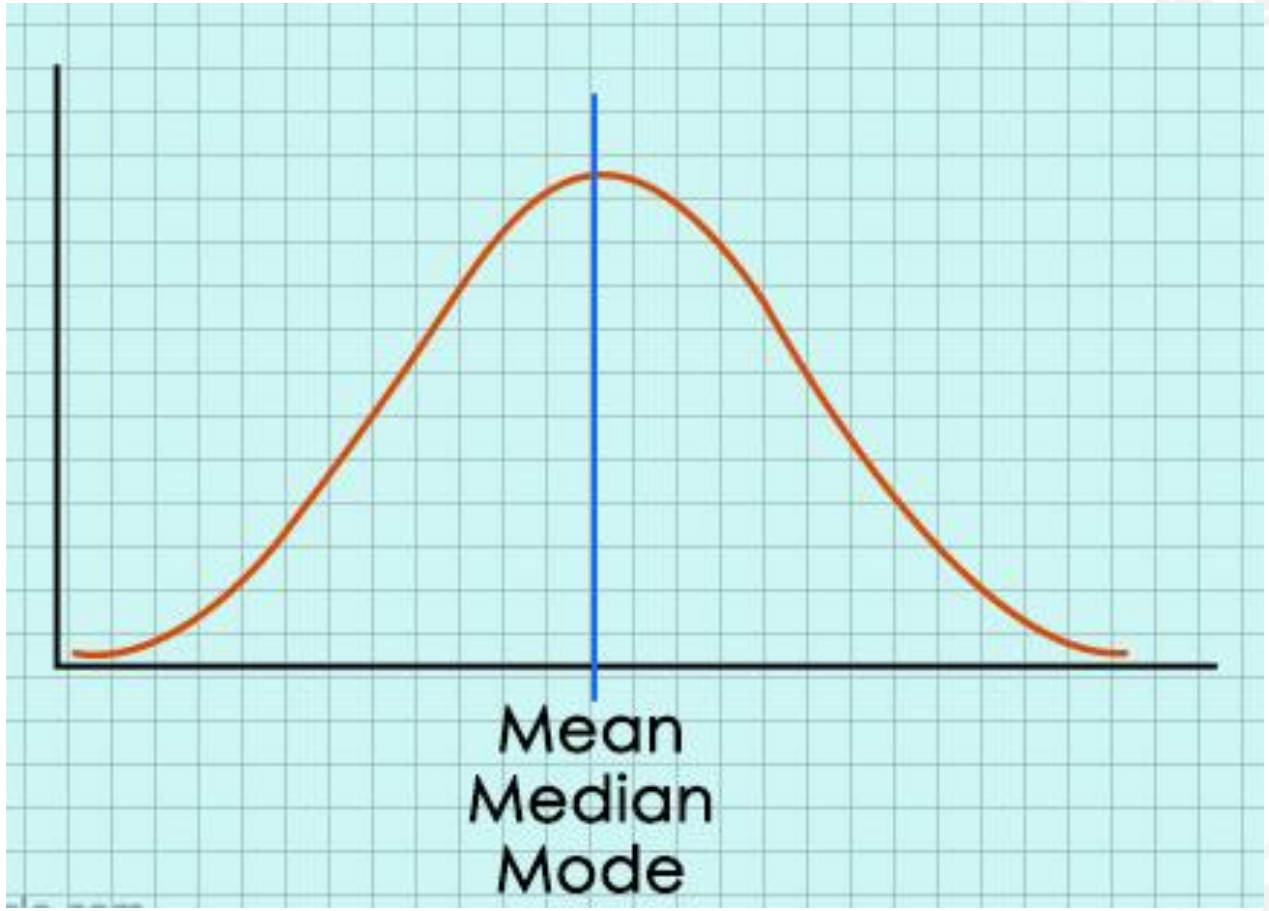
$$Curt = \mu_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)\sigma^4} - 3$$

Histograma

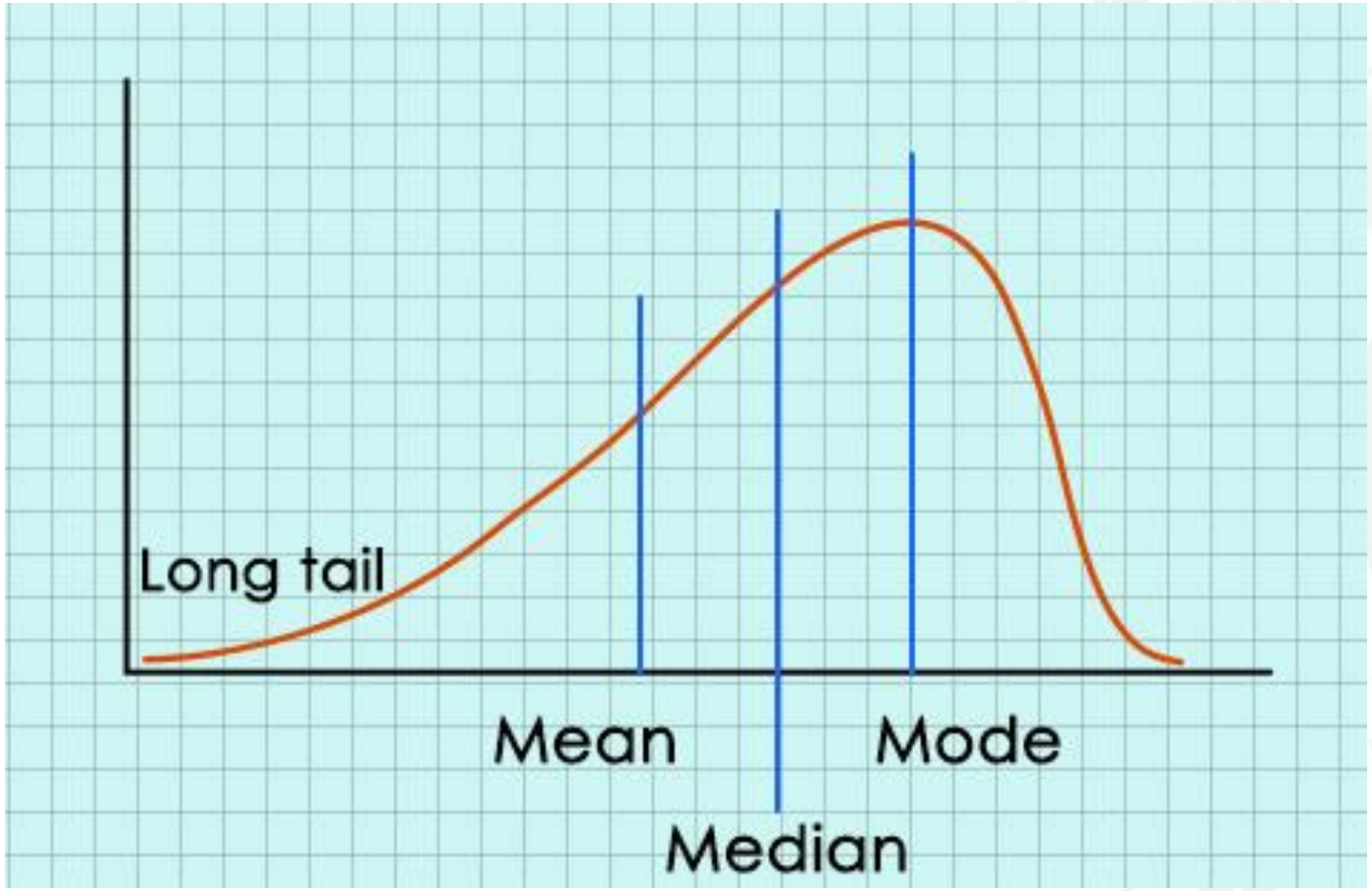
- Melhor forma para verificar graficamente curtose e obliquidade



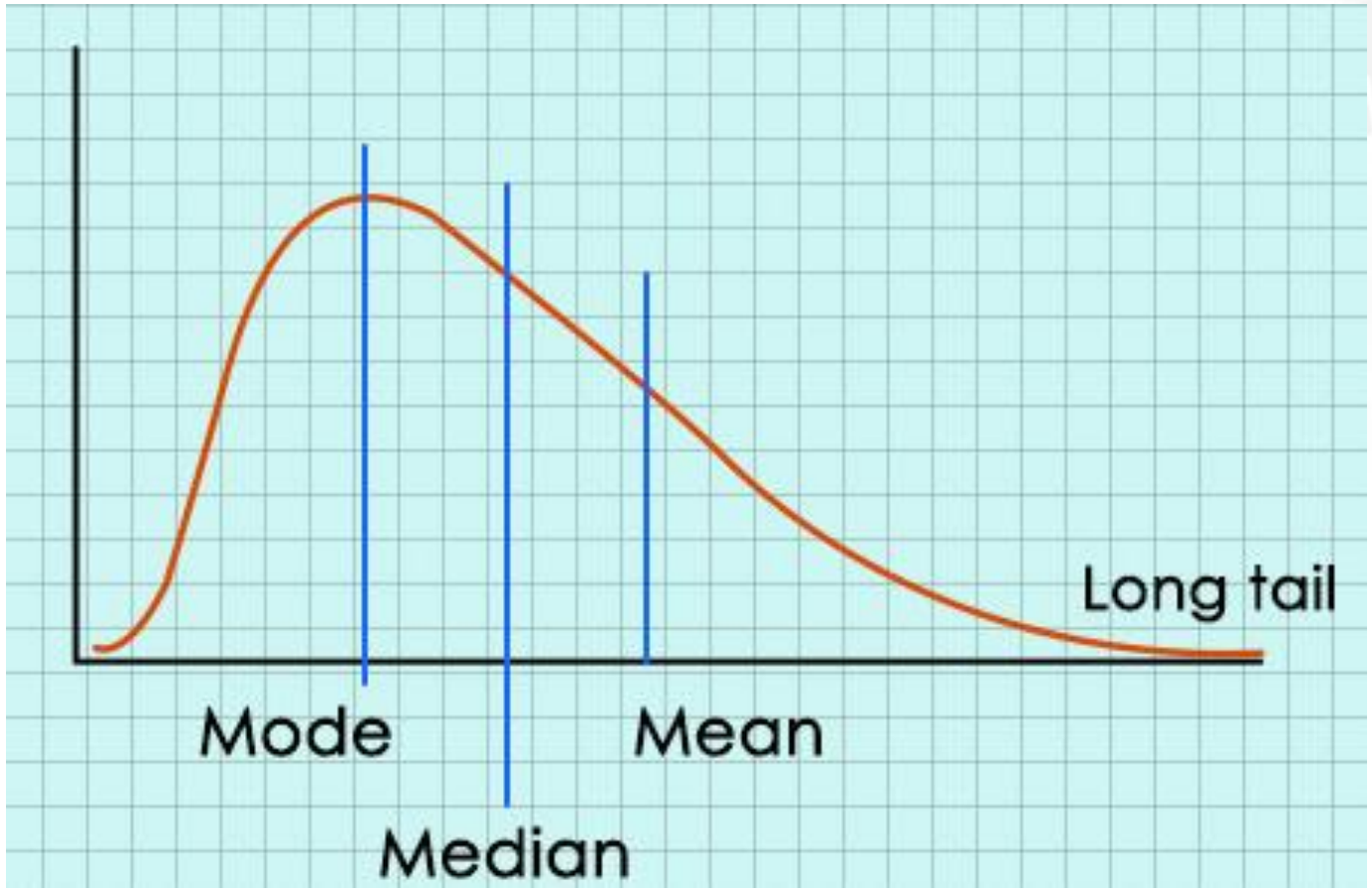
Distribuição Normal



Obliquidade (negativa)

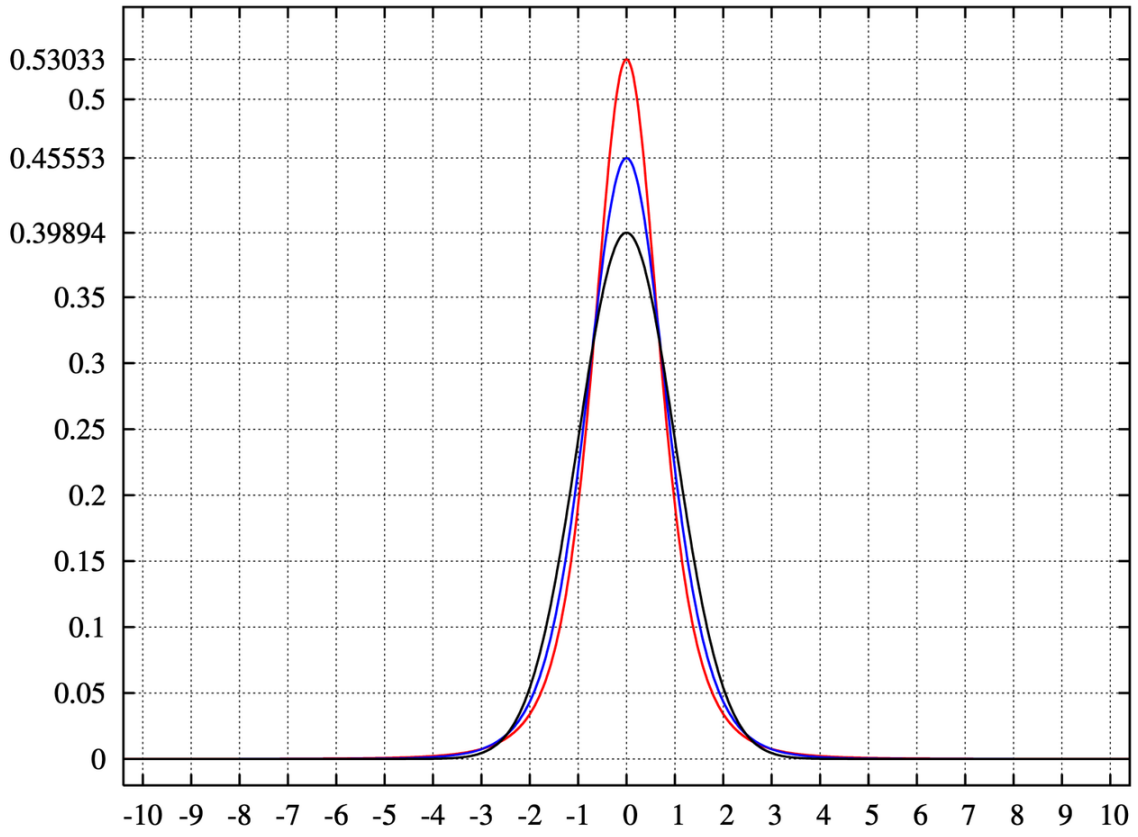


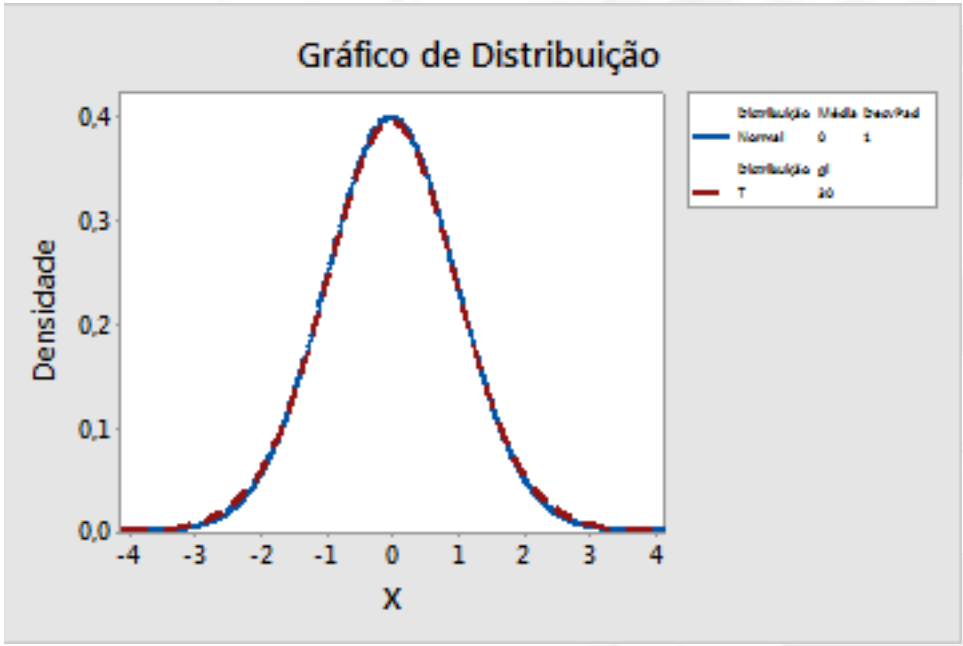
Obliquidade (Positiva)



Ver o exemplo

Curtose faz a diferença





Todas tem média zero e variância 1
São diferentes!!!

Exemplos

- Usar PANDAS e NUMPY e SKLEARN.

Com a ajuda de pacotes como Pandas e Numpy,

Python é um ótimo ambiente para aprender as ferramentas necessárias para trabalhar como cientista de dados.

EXEMPLO 1

Curso 2 – CD, AM e DM (Parte 2)

Profa. **Roseli Ap. Francelin Romero**

MBA em Inteligencia Artificial e BigData

Depto. de Ciências de Computação
ICMC - USP

PARTE II

PRE – PROCESSAMENTO DOS DADOS

- Limpeza nos Dados
- Transformação de Dados
- Redução da Dimensionalidade
- Balanceamento de Dados

PRE-PROCESSAMENTO DE DADOS

- Prepara os dados para seu uso por algoritmos de AM
- Procura melhorar desempenho do algoritmo
  - Custo
    - Tempo
    - Memória
  - Qualidade do modelo gerado
    - Acurácia preditiva

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Redundantes
  - Inconsistentes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

LIMPEZA NOS DADOS

PROBLEMAS NOS DADOS

- Falha humana
- Má fé
- Falha no processo ou dispositivo de coleta ou de medição de dados
- Limitações do dispositivo de coleta ou de medição
- Mudanças (eventos)

VALORES AUSENTES

- NaN; ?; em branco
- SUBSTITUIR:
  - Média dos valores do Atributo
  - Média dos valores anterior e posterior
  - Regressão
- Agir como se não houvessem valores ausentes
  - Utilizar apenas os valores que estão presentes
    - Ex.: Menos atributos no cálculo da distância entre objetos
  - Modificar algoritmo de AM para lidar com valores ausentes
- Descartar objetos com atributos sem valores
- Preencher valores ausentes (sklearn.impute.SimpleImputer)

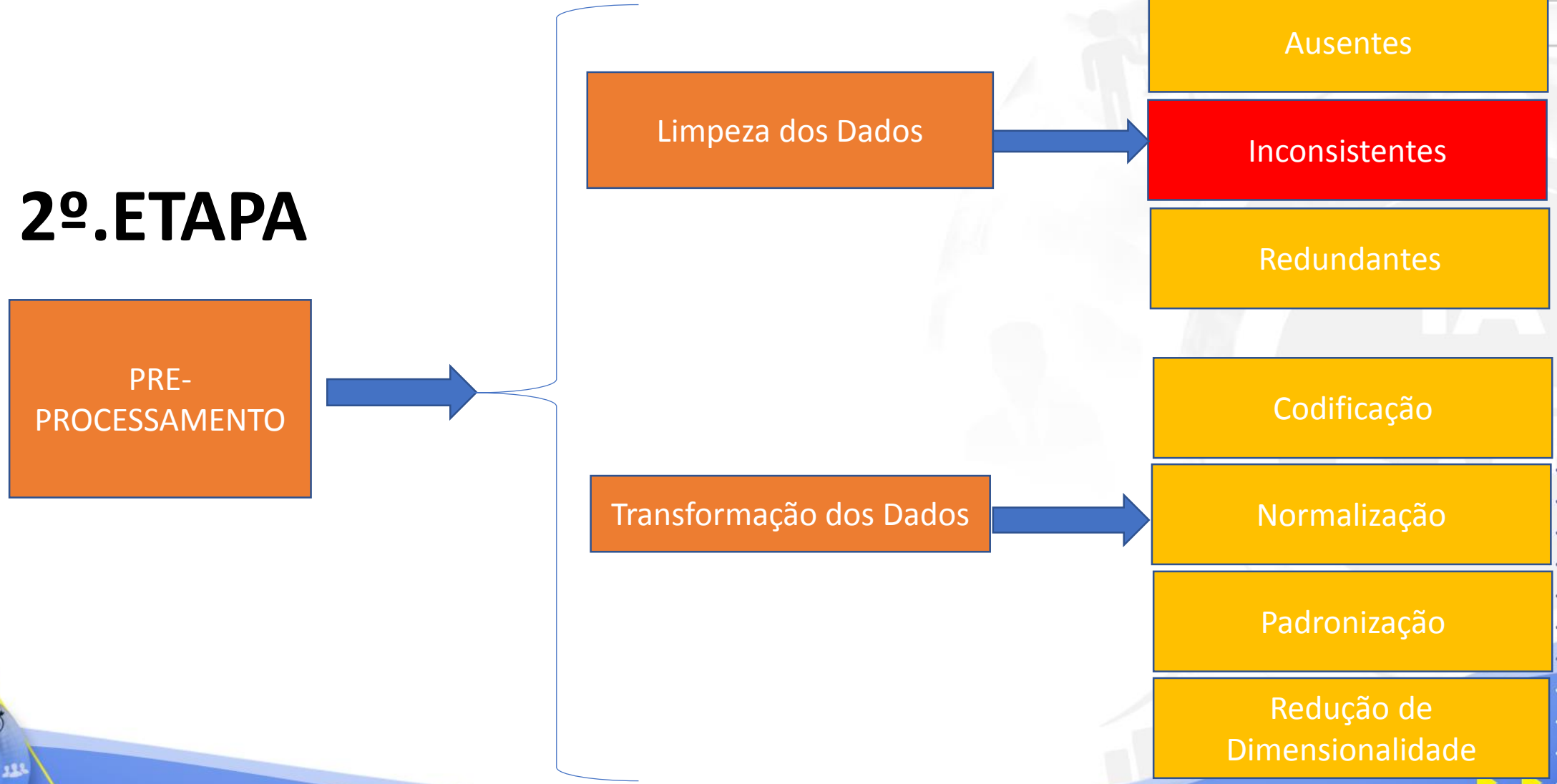
VALORES AUSENTES

- Criação de um novo valor que significa ausência
  - Para valores nominais (sem ordem)
- Criação de um novo atributo preditivo
  - Marcando objetos em que um dado atributo tinha valor ausente

TIPOS DE ATRIBUTOS – CASE 2

	Nome das tabelas	Nº de colunas	Nº de linhas	Células nulas	Linhas duplicadas
0	olist_customers_dataset	5	99441	0	0
1	olist_geolocation_dataset	5	1000163	0	261831
2	olist_order_items_dataset	7	112650	0	0
3	olist_order_payments_dataset	5	103886	0	0
4	olist_order_reviews_dataset	7	99224	145903	0
5	olist_orders_dataset	8	99441	4908	0
6	olist_products_dataset	9	32951	2448	0
7	product_category_name_translation	2	71	0	0
8	olist_sellers_dataset	4	3095	0	0

2º.ETAPA



VALORES INCONSISTENTES

- Dados podem conter valores inconsistentes
  - Atributos preditivos
    - Ex. Código postal inválido para uma cidade
      - Erro / engano
      - Proposital (fraude)
  - Atributo alvo
    - Podem levar a objetos conflitantes (ambiguidade)
      - Ex.: valores iguais para atributos preditivos e diferentes para atributo alvo
    - Podem ser causados por erro na rotulação do objeto

VALORES INCONSISTENTES

- Algumas inconsistências são de fácil detecção
  - Violação de relações conhecidas entre atributos
    - Ex.: Valor de atributo A é sempre menor que valor de atributo B
  - Valor inválido para o atributo
    - Ex.: altura com valor negativo

  Por exemplo. O delivery_time não pode ser negativo, pois ele é o resultado da subtração entre a data que o produto chegou no cliente e a data que o produto saiu do parceiro logístico,
  Um número negativo mostraria que o produto teria chegado no cliente antes de sair para entrega

  - Em outros casos, informações adicionais precisam ser consideradas
- Podem indicar presença de ruído

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Inconsistentes
  - Redundantes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

OBJETOS REDUNDANTES

- Objetos ou atributos preditivos (quase) duplicados ou muito relacionados
  - Não trazem informação nova
  - Ex.: Pessoas em diferentes BDs com mesmo nome, mas endereço com pequenas diferenças
    - Diferença real ou erro no preenchimento
- Deduplicação
  - Detectar e eliminar (ou combinar) duplicações
  - Cuidado para não eliminar ou combinar objetos ou atributos que representam dados diferentes

OUTLIERS

- Objetos ou valores anômalos
  - Objetos que têm características diferentes da grande maioria dos demais objetos
    - Valor(es) de um ou mais atributos que destoa(m) dos valores típicos
- *Outliers* podem sugerir a presença de ruído ou serem valores legítimos
  - Em várias aplicações, objetivo é encontrar *outliers.*

Curso 2 – CD, AM e DM (Parte 3)

Profa. **Roseli Ap. Francelin Romero**

MBA em Inteligência Artificial e BigData

Depto. de Ciências de Computação
ICMC - USP

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Inconsistentes
  - Redundantes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

Transformação de dados

- Mudam o tipo de um atributo
- Conversão de valores entre tipos
  - Qualitativos para quantitativos
    - Binarização
  - Quantitativos para qualitativos
- Normalização de valores numéricos
- Tradução de atributos

Qualitativos para quantitativos

- Algumas técnicas trabalham apenas com valores numéricos
- Conversão depende de:
  - Existência de ordenação dos valores
    - Se existe (ordinal), manter
    - Se não existe (nominal), não inserir
  - Número de valores
    - Se igual a 2 (binários) ou maior que 2

Conversão de valor ordinal

- Codificar para valor inteiro positivo
  - Ex. Pequeno: 1, médio: 2 e grande: 3
- Algumas técnicas trabalham apenas com valores quantitativos binários
  - Binarização

Binarização de ordinal

- Transformação no sistema numérico binário correspondente?
  - Perde ordenação
- Valores consecutivos devem diferir em 1 bit
- Codificar cada valor por um vetor binário que mantém ordenação
  - Código cinza: 000, 001, 010, 011, ...
  - Código termômetro: 001, 011, 111

Código cinza

- Existem vários códigos cinza
  - Não é único
- Um código cinza para 3 bits:
  - 000, 001, 011, 010, ...
- Um código cinza para 2 bits:
  - 00, 01, 11, 10

Dígito	Binário	Código cinza
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

Algoritmo código cinza

1 Começa com todos os bits iguais a zero

2 Para cada novo número

Mudar o valor do bit mais a direita que gera uma nova sequência de bits

Código termômetro

- Utiliza mais bits que código cinza
  - Tamanho cresce linearmente com número de valores

Dígito	Binário	Código termômetro
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0111
4	0100	1111

Conversão de valor nominal

- Transforma para valor quantitativo
  - Não deve inserir relação de ordem
- Codificação binária nominal sem relação de ordem
- Codificações
  - 1-de-n (n = número de valores)
  - m-de-n

Conversão de valor nominal

- Codificação 1-de-n
  - Codificação canônica
  - Fácil calcular moda = posição com maior número de valores 1
  - Quantidade de valores pode gerar vetores longos
- Codificação m-de-n
  - Dos n valores, m são iguais a 1 e os demais 0
  - Vários códigos

Exemplo (1 de n)

Esse tipo de conversão é importante quando você quer utilizar uma rede neural artificial. Quando temos uma classe numérica como no conj. wine.data: 0,1,2

é melhor transformar a classe em 1 vetor com 3 colunas sendo:

- a primeira coluna é 1 quando a classe for 0,
- a segunda coluna tem valor 1 quando a classe for 1,
- a terceira coluna é 1 quando a classe for 2.

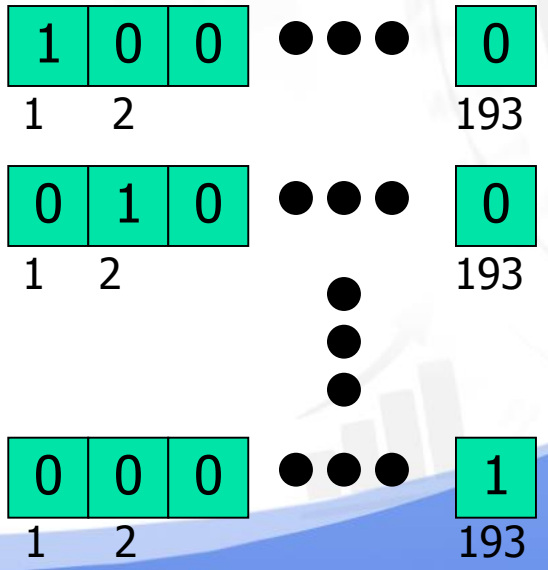
Isso resulta numa camada de saída da rede neural com 3 neurônios, onde cada neurônio sinaliza uma das classes. Esse processo também ajuda na convergência da rede neural artificial.

Conversão de valor nominal

- Número de valores de um atributo pode ser muito grande
- Pseudo atributos
  - Cria valores novos, artificiais
- Ex.: Atributo é nome de país
  - Existem 193 países (192 representados na ONU + Vaticano)
  - Alternativa de codificação:
    - Transformar valores nominais em numéricos utilizando a codificação 1-de-n

Alternativa 1

- Transformar valores nominais em valores binários utilizando a codificação 1-de-n
  - Maldição da dimensionalidade
  - Grande parte dos elementos possui valor 0
    - Valores esparsos



Alternativa 2

- Transformar 193 atributos em 10 pseudo-atributos
  - Continente: 7 valores binários
  - IDH: 1 valor real
  - População: 1 valor inteiro
  - Área: 1 valor inteiro

Transformação de atributos

- Muda valor numérico de um atributo para outro valor numérico
  - Limites de valores para atributos distintos podem ser muito diferentes
    - Evitar que um atributo predomine sobre outro
      - A menos que isso seja importante
  - Valores podem estar concentrados em uma determinada faixa ou região
  - Possível necessidade de binarização (OneHot Encoding)

Transformação de atributos

- Aplicada aos valores de um atributo específico para todos os exemplos
- Variações
  - Funções simples
  - Normalização
  - Padronização

Funções simples

- Uma função matemática simples é aplicada a cada valor do atributo
  - Muda distribuição de valores de um atributo
  - Possíveis transformações para um atributo *x* de um conjunto de dados:
    - **x^k, log(x), e^x, $\sqrt{x}$, 1/x, sqrt(x), seno(x) e |x|**

Transformação sin ou cos

$$X_i' = \sin\left(\frac{2\pi X_i}{|X|}\right)$$

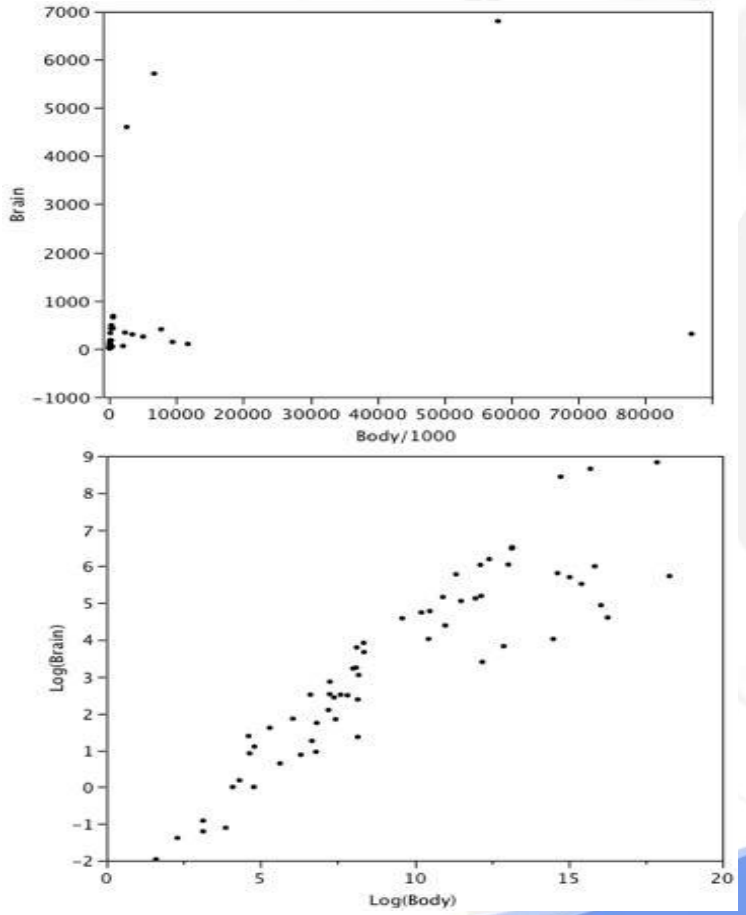
$$X_i' = \cos\left(\frac{2\pi X_i}{|X|}\right)$$

Funções simples

- Valor absoluto
  - Em algumas aplicações, apenas magnitude do valor de um atributo é importante
  - Converte valor de todos os atributos para o valor positivo correspondente
    - Ex.: -4, 5 e –2 se tornam 4, 5 e 2

Funções simples

- Utilizando função $\log_{10}$
  - Comprime valores de atributos que se encontram em grande intervalo de possíveis valores
  - Ex.: relação, para alguns animais, entre:
    - Peso do cérebro e
    - Peso do corpo



http://onlinestatbook.com/2/transformations/log.html

Transformar dados categóricos em numéricos

- Se no conjunto de dados do **preço do aluguel**, o atributo **condição** é codificada da seguinte forma:

  - novo: 1
  - reformado: 2
  - precisa de reforma: 3

- e a **qualidade** como:

  - luxuoso: 1
  - bom: 2
  - normal: 3
  - simples: 4
  - desconhecido: 5

LABEL ENCODING

Funções disponíveis para Transformar dados categóricos em numéricos

- Label Encoder, One Hot Encoding, bin encoding, and hashing encoding.
- No entanto, a maioria das pessoas usa o Label Encoding incorretamente quando deveria ter sido usado o One Hot Encoding.
- Por exemplo, introduzindo ordem:

  [1 dorm, studio, 2 dorms, studio, studio, térreo].

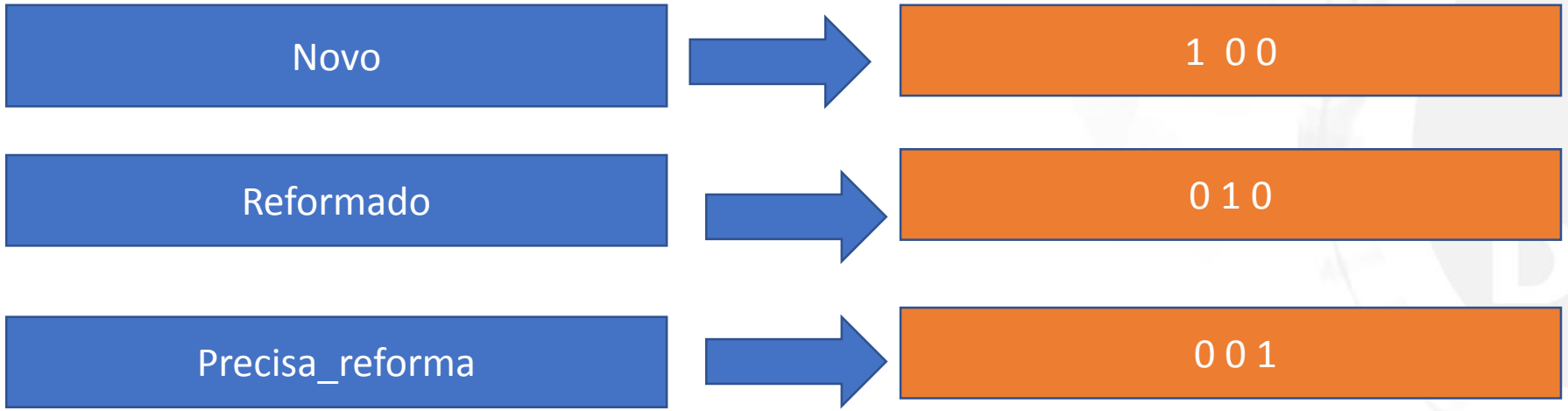
- LabelEncoding pode transformar isso em [3,2,4,2,2,1] :

  → Arvore de decisão – ok

  → Regressão ou SVM - não

Transformar dados categóricos em numéricos binários

- Atributo **condição**: novo (1), reformado (2) e precisa_reforma (3)



ONE HOT ENCODING

Transformar dados categóricos em numéricos

- **One Hot Encoding**
- **4 classes:**

  - [1 dorm, studio, 2 dorms, studio, studio, térreo]

térreo – (1 0 0 0)

1 studio – (0 1 0 0)

1 dorm. – (0 0 1 0)

2 dorms. – (0 0 0 1)

Se tivermos n classes: vetores com n componentes

Usamos vetores da Base Canônica

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Inconsistentes
  - Redundantes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

Normalização

- Para normalizar os valores de um atributo:
  1. Adicionar ou subtrair uma constante
  2. Multiplicar ou dividir por uma constante
- Utilizado para mudar intervalo de valores dos dados
  - Permite converter todos os valores de um atributo para o intervalo [0, 1]

$$x' = \frac{(x - \min_x)}{(\max_x - \min_x)}$$

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Inconsistentes
  - Redundantes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

Padronização dos Dados

- A padronização traz todas as variáveis contínuas para a mesma escala, ou seja,

- **se uma variável tiver valores de 1K a 1M e**
- **outra de 0,1 a 1,0**

após a padronização elas estarão o mesmo intervalo

Padronização dos Dados

- Para padronizar os valores de um atributo:
  1. Adicionar ou subtrair uma medida de localização
  2. Multiplicar ou dividir por uma medida de espalhamento
- Se os valores têm uma distribuição Gaussiana
  - Subtrair a media
  - Dividir pelo desvio padrão
  - Produz valores com distribuição normal (0,1)

Z-score

$$x' = \frac{(x - \bar{x})}{\sigma}$$

Curso 2 – CD, AM e DM

MATRIZ DE COVARIÂNCIA

MATRIZ DE CORRELAÇÃO

SCATTER PLOT

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Inconsistentes
  - Redundantes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

Dados multivariados
MATRIZ DE COVARIÂNCIA

- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos
    - Valor próximo de 0:
      - Atributos não têm um relacionamento
    - Valor positivo:
      - Atributos diretamente relacionados
        - Quando o valor de um atributo aumenta, o do outro também aumenta
      - Valor negativo:
        - Atributos inversamente relacionados
    - Valor depende da magnitude dos atributos

MATRIZ DE COVARIÂNCIA

- Cálculo de cada elemento s_{ij} de uma matriz de covariância S para um conjunto de n objetos

$$s_{ij} = \text{covariância}\,(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Onde:
$\bar{x}_i$: Valor médio do i-ésimo atributo
x_{ki} : Valor do i-ésimo atributo para o k-ésimo objeto

- Obs: covariância (x_i, x_i) = variância (x_i)
  - Matriz de covariância tem em sua diagonal as variâncias dos atributos

Dados multivariados - Correlação

- Covariância de dois atributos
  - É difícil avaliar o relacionamento entre dois atributos olhando apenas a covariância
    - Sofre influência da faixa de valores dos atributos
    - **Correlação** entre dois atributos ilustra mais claramente a força da relação entre eles
      - Mais popular que covariância
      - Elimina influência da faixa de valores

Matriz de Correlação

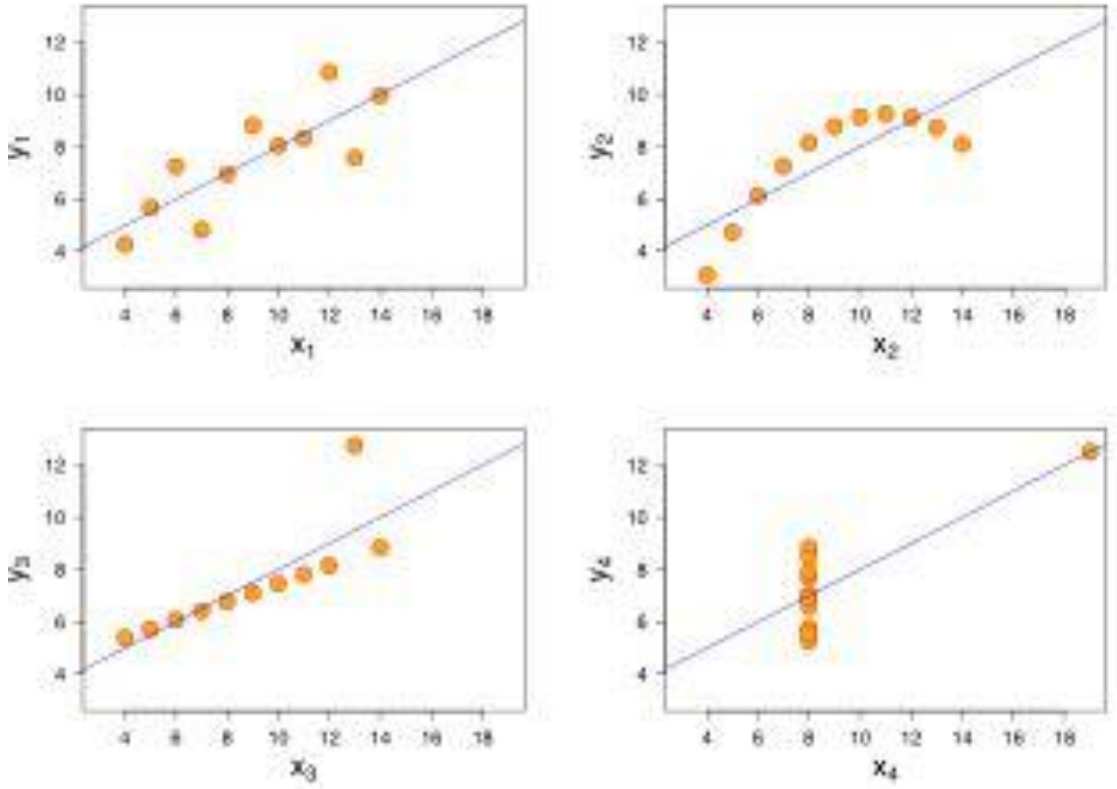
Tabela 2. Matriz de correlação das variáveis de qualidade de água no médio Rio Pomba

Variável	OD	DBO	Temp	CE	Alcalin	DQO	PT	Norg	NH_4	NTK	NT	ST	SST	SDT	SIS
OD	1,00														
DBO	**-0,58**	1,00													
Temp	**-0,57**	0,18	1,00												
CE	0,02	-0,24	**0,61**	1,00											
Alcalin	-0,23	-0,34	**0,63**	**0,77**	1,00										
DQO	**-0,70**	0,32	**0,64**	0,12	0,38	1,00									
PT	-0,09	0,31	0,03	0,21	-0,11	-0,10	1,00								
Norg	0,16	0,23	**-0,62**	-0,49	**-0,71**	-0,49	0,28	1,00							
NH_4	**-0,76**	**0,53**	0,38	-0,01	0,12	0,31	0,48	0,08	1,00						
NTK	0,08	0,28	**-0,57**	-0,49	**-0,69**	-0,45	0,33	**0,99**	0,18	1,00					
NT	0,15	0,27	**-0,55**	-0,34	**-0,62**	**-0,50**	0,35	**0,97**	0,13	**0,97**	1,00				
ST	0,24	-0,06	-0,23	0,33	-0,02	**-0,61**	0,47	0,29	0,08	0,29	0,37	1,00			
SST	0,01	0,35	-0,39	-0,14	-0,26	-0,43	**0,60**	**0,61**	0,37	**0,64**	**0,63**	**0,63**	1,00		
SDT	0,30	-0,38	0,04	**0,56**	0,20	-0,42	0,09	-0,15	-0,22	-0,17	-0,08	**0,75**	-0,04	1,00	
SIS	0,01	0,26	-0,47	-0,19	-0,37	-0,43	**0,54**	**0,61**	0,32	**0,63**	**0,63**	**0,75**	**0,84**	0,24	1,00

OD (mg L^{-1}) – oxigênio dissolvido; DBO (mg L^{-1}) – demanda bioquímica de oxigênio; Temp (°C) – temperatura; CE (uS cm^{-1}) – condutividade elétrica; Alcalin (mg L^{-1} $CaCO_3$) – alcalinidade; DQO (mg L^{-1}) – demanda química de oxigênio; PT (mg L^{-1} – PO_4) – fósforo total; Norg (mg L^{-1} - N) – nitrogênio orgânico; NH_4 (mg L^{-1} - N) – nitrogênio amoniacal; NTK (mg L^{-1} - N) – nitrogênio Kjeldahl; NT (mg L^{-1} – N) – nitrogênio total; ST (mg L^{-1}) – sólidos totais; SST ((mg L^{-1}) – sólidos suspensos totais; SDT (mg L^{-1}) – sólidos dissolvidos totais; SIS (mg L^{-1}) – sólidos inorgânicos suspensos

Fonte: Scielo

Dados Multivariados



Todos correlação = 0,816

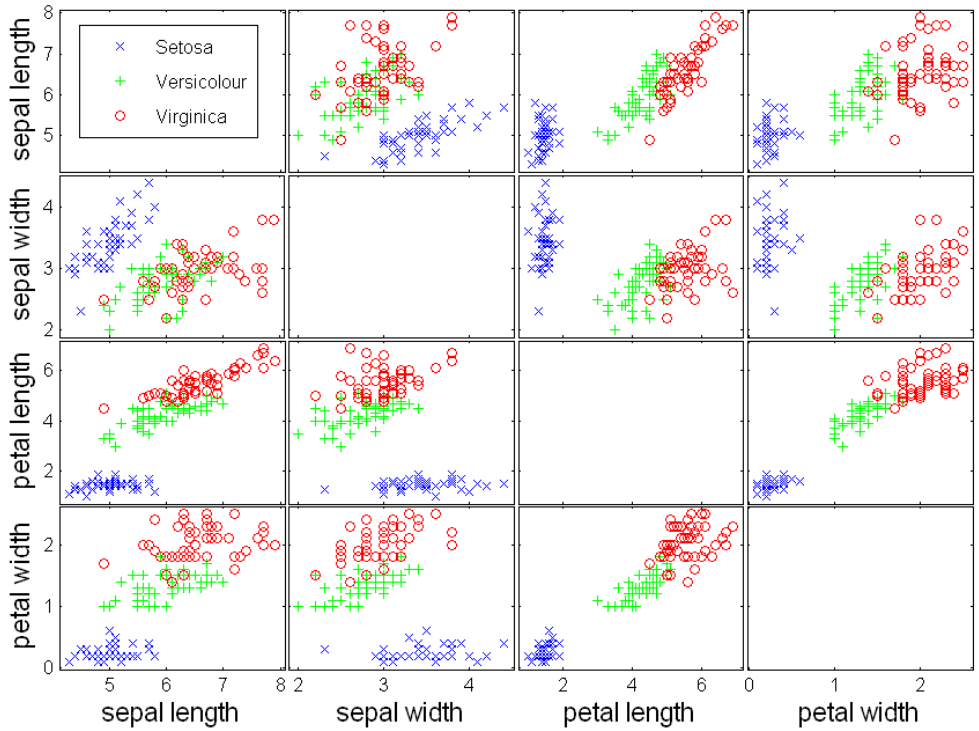
Fonte:Wikipedia

Scatter Plot

- Usado para ilustrar graficamente Correlação Linear entre dois atributos
- Cada objeto é associado a uma posição em um gráfico
  - Valores dos atributos definem sua posição
  - Valores podem ser inteiros ou reais
- Matrizes de scatter plot resumem relação para vários pares de atributos

Scatter Plot

- Matriz para atributos do conjunto iris



Diferentes classes são indicadas por cores diferentes

EXEMPLO 2

Uso de bibliotecas

- Matplotlib; https://matplotlib.org/
- NumPy; https://numpy.org/
- pandas; https://pandas.pydata.org/
- scikit-learn; https://scikit-learn.org/stable/
- seaborn

- DUVIDAS:

- E-mail: rafrance@icmc.usp.br