
# Escola Politécnica de Pernambuco

*Especialização em Ciência de Dados e Analytics*

## Introdução à Ciência de Dados
## Aula 3

Prof. Dr. Alexandre Maciel

***amam@ecomp.poli.br***

# QUAIS DADOS ANALISAR?

# PRODUÇÃO DE DADOS

- Teclado
- Mouse
- Touch Screen
- Scanners
- Código de Barras
- RFID
- Câmeras
- Filmadoras

# DADOS PROCESSADOS

**- Análise ou execução de procedimentos**

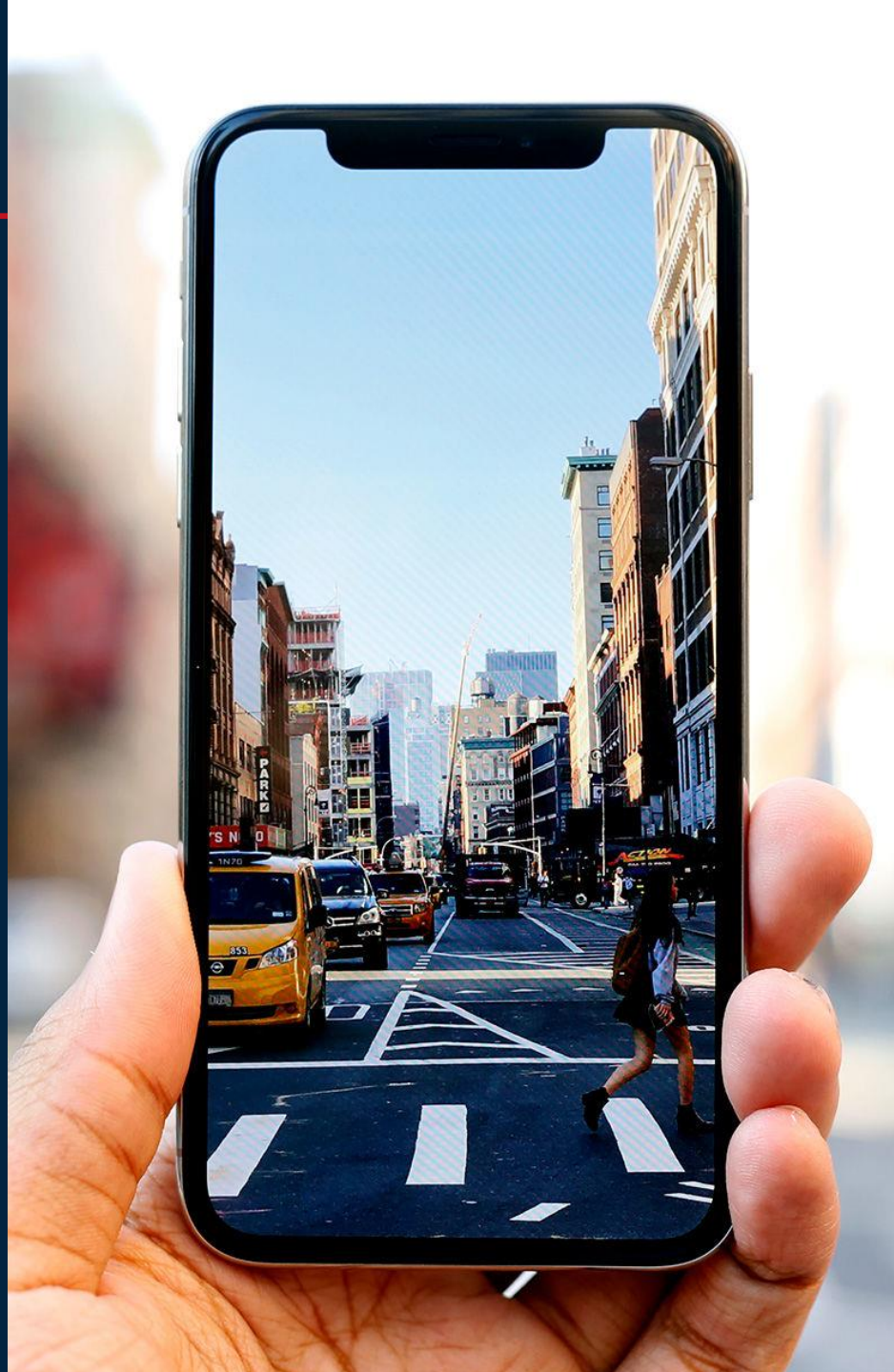
**- Criação de modelos:**

- Estatísticos
- Machine Learning

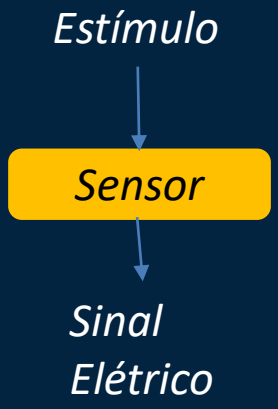
**- Criação de Data Warehouses.**

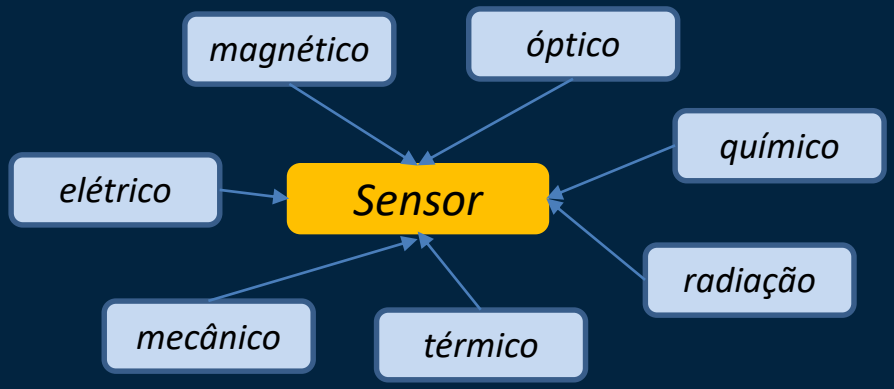
# SMARTPHONES

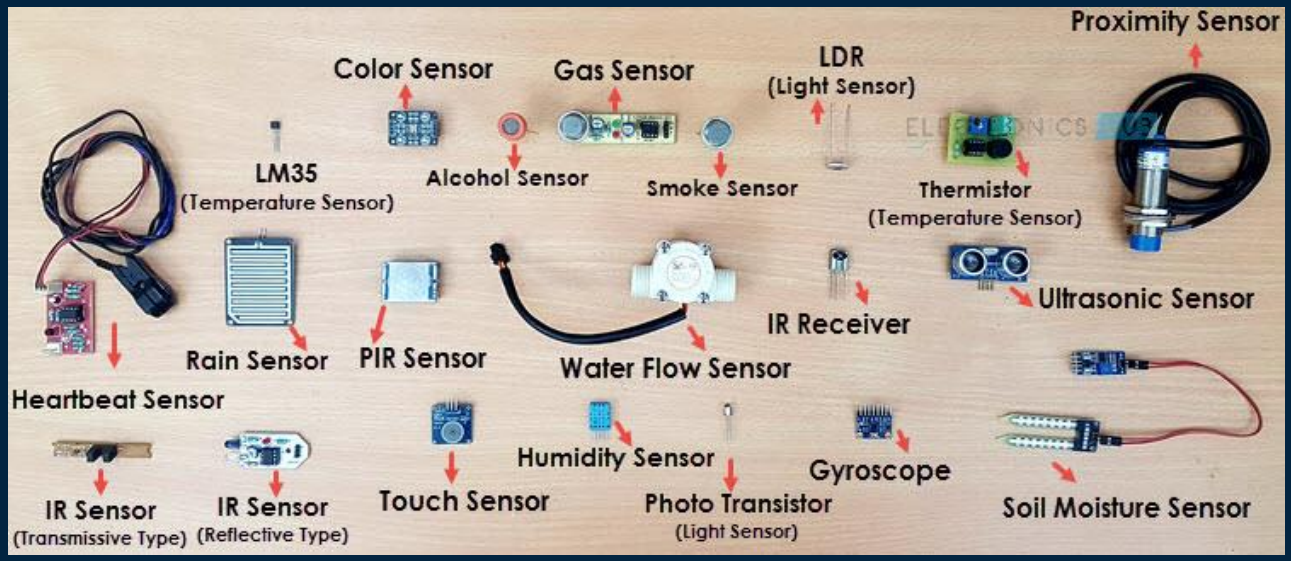
- Câmeras
- Touch screen
- Acelerômetro
- GPS
- Giroscópio
- Magnetômetro
- Bluetooth
- WiFi



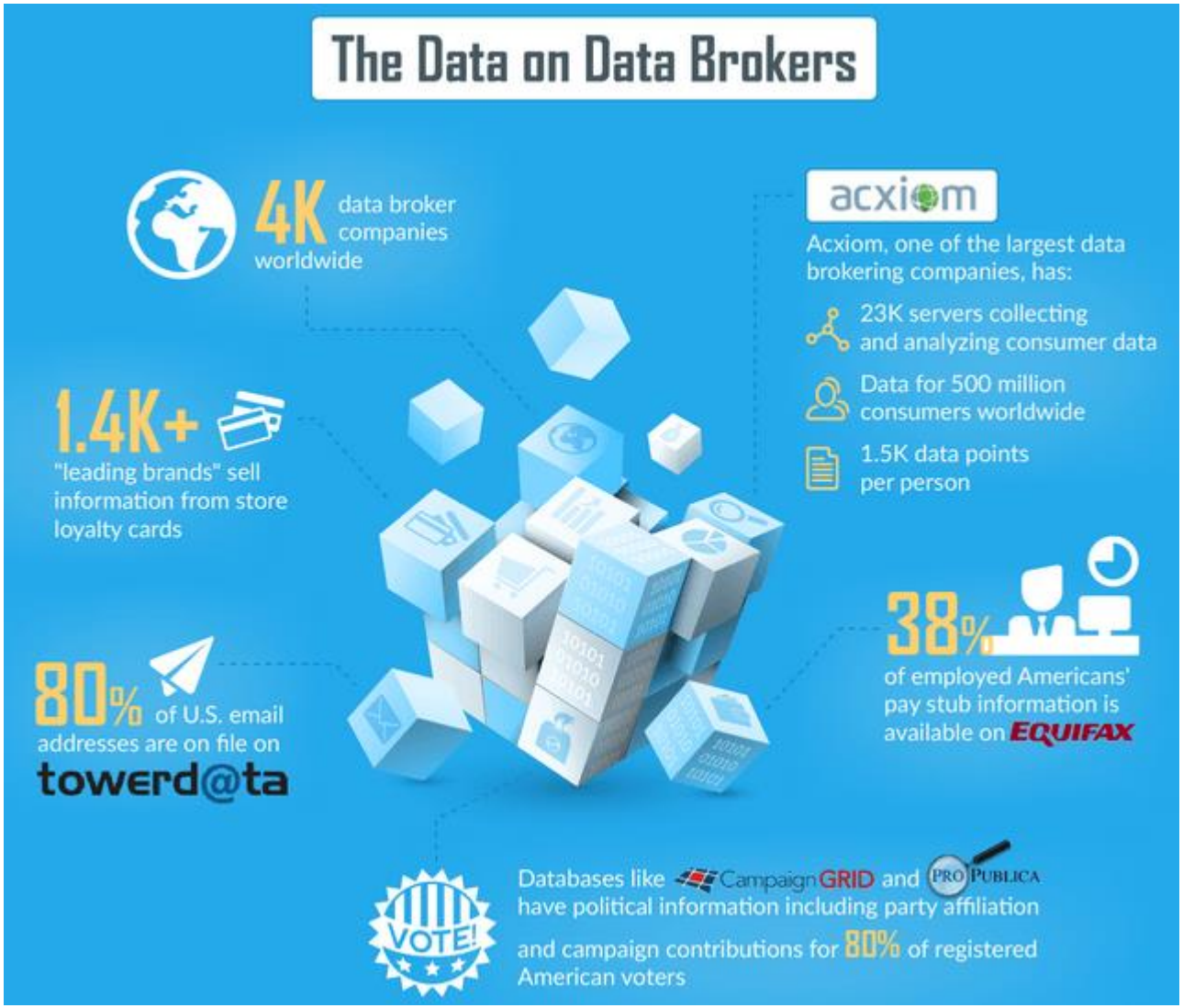
# SENSORES


Estímulo
Sensor
Sinal
Elétrico


magnético
óptico
químico
elétrico
Sensor
radiação
mecânico
térmico


Proximity Sensor
Color Sensor
Gas Sensor
LDR
(Light Sensor)
LM35
(Temperature Sensor)
Alcohol Sensor
Smoke Sensor
Thermistor
(Temperature Sensor)
Ultrasonic Sensor
IR Receiver
Rain Sensor
PIR Sensor
Water Flow Sensor
Heartbeat Sensor
Humidity Sensor
Gyroscope
IR Sensor
(Transmissive Type)
IR Sensor
(Reflective Type)
Touch Sensor
Photo Transistor
(Light Sensor)
Soil Moisture Sensor

# COMPRA DE DADOS




https://www.webfx.com/blog/general/what-are-data-brokers-and-what-is-your-data-worth-infographic/

# COLETA DE DADOS


**- Quantitativo ou qualitativo**

**- Dados estruturados, semiestruturados e não estruturados**

- **Categórico ou numérico**
  - Binário, nominal, ordinal
  - Discreto, contínuo, razão

**- Quantidade *versus* qualidade**

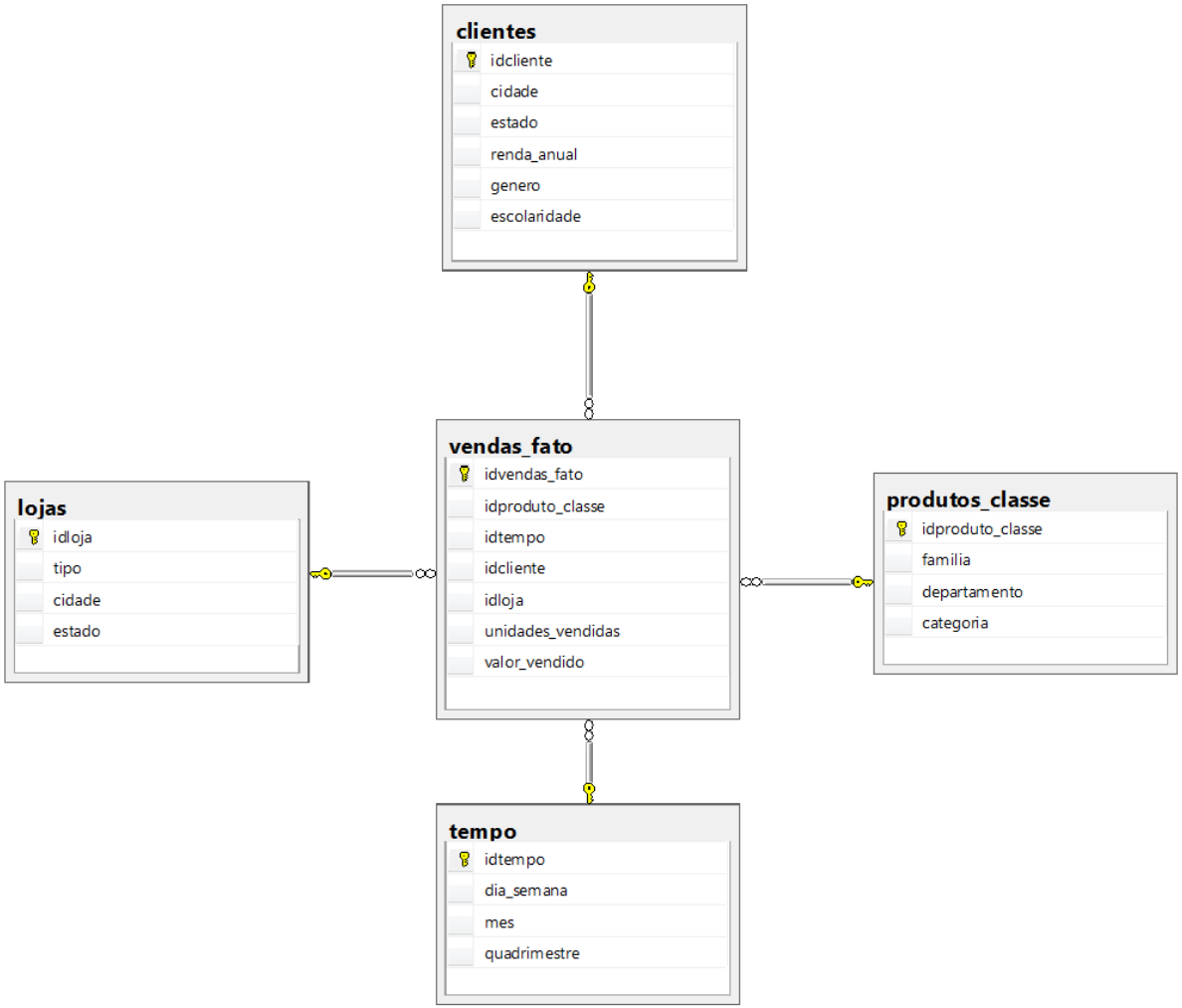
# TIPOS DE CONJUNTOS DE DADOS

***Dados em registros***

- Relacional

- Transacional

- Dimensional



# TIPOS DE CONJUNTOS DE DADOS

***Dados em grafos***

- Relacionamento de objetos

- Dados que são grafos

# TIPOS DE CONJUNTOS DE DADOS

***Dados ordenados***

- Sequenciais
- De sequência
- Séries temporais
- Espaciais

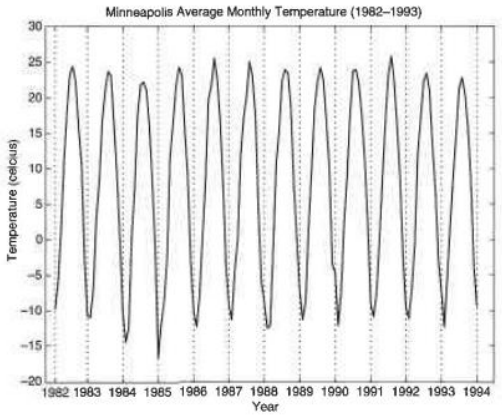
| Time | Customer | Items Purchased |
|---|---|---|
| t1 | C1 | A, B |
| t2 | C3 | A, C |
| t2 | C1 | C, D |
| t3 | C2 | A, D |
| t4 | C2 | E |
| t5 | C1 | A, E |

| Customer | Time and Items Purchased |
|---|---|
| C1 | (t1: A,B) (t2:C,D) (t5:A,E) |
| C2 | (t3: A, D) (t4: E) |
| C3 | (t2: A, C) |

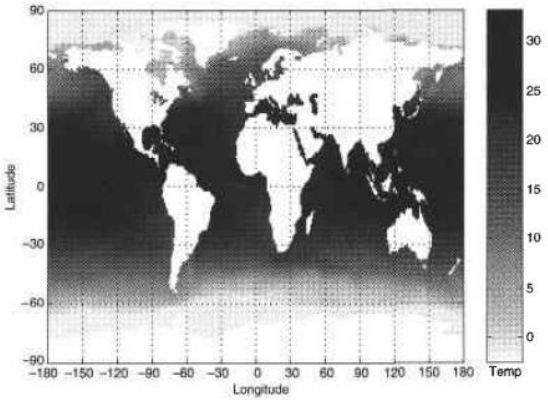
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.

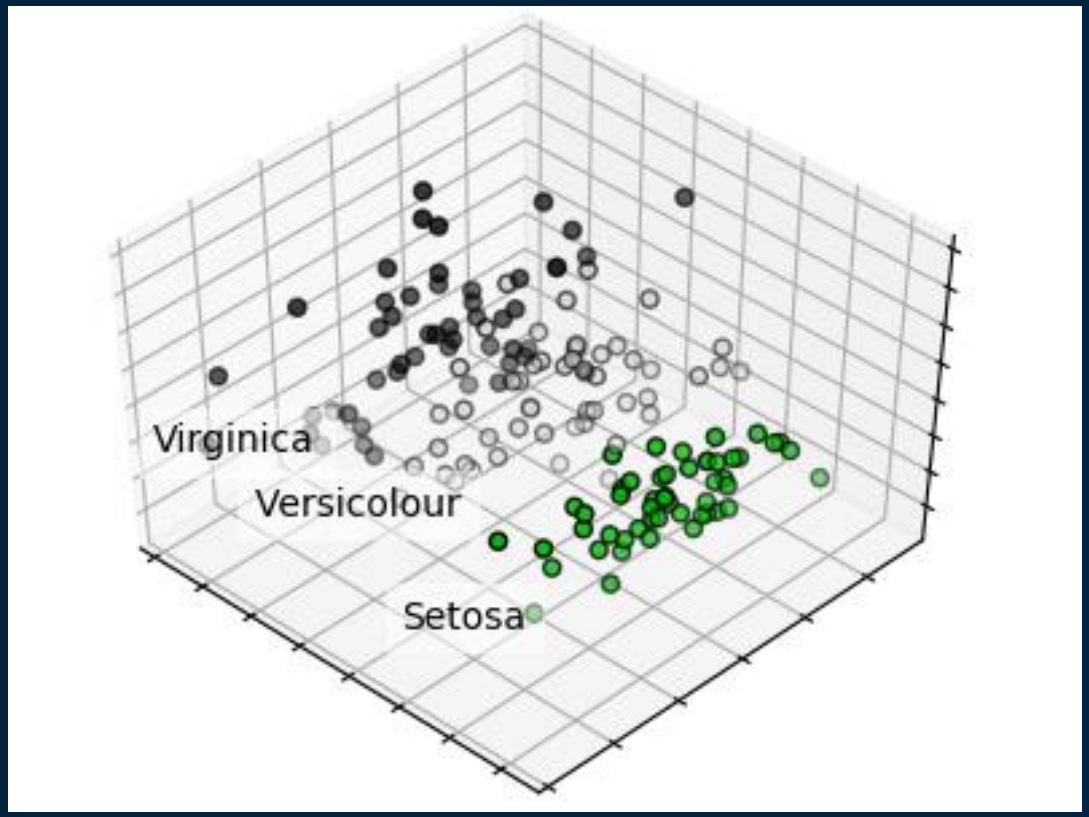



(c) Temperature time series.




(d) Spatial temperature data.

# CARACTERÍSTICAS DOS CONJUNTOS DE DADOS




***Dimensão***

- Número de atributos

- Maldição da dimensionalidade

- Redução de dimensionalidade

# CARACTERÍSTICAS DOS CONJUNTOS DE DADOS

***Dispersão***

- Distribuição assimétrica dos atributos

- Vantagem: processamento apenas dos atributos “diferentes”

# CARACTERÍSTICAS DOS CONJUNTOS DE DADOS

***Resolução***

- Diferentes graus de granularidade dos valores

- Se pequeno, padrão encoberto pelo ruído, se grande demais, desaparece

# DICIONÁRIO DE DADOS

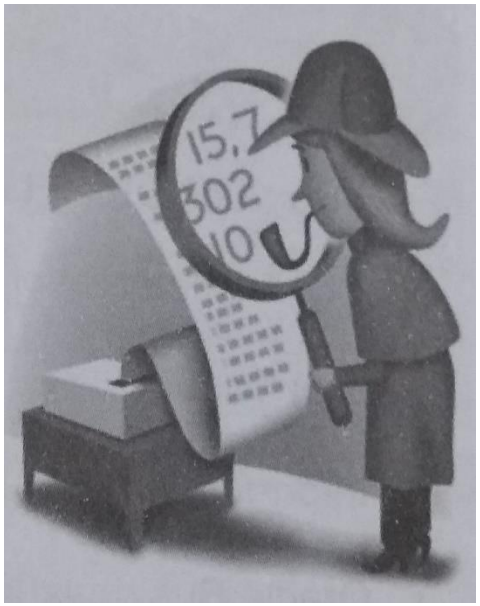
**Conjunto de Dados Censo Escolar - 2015**

**Descrição do recurso Cadastro das Turmas**

| Descrição dos Metadados | | | | |
|---|---|---|---|---|
| Campo | Descrição | Tipo | Tamanho | Valores Permitidos |
| inep_escola | Código de escola – Inep | Num | 8 | |
| codigo_escola | Código da escola no município | Char | 3 | |
| codigo_turma | Código da Turma na Entidade/escola | Char | 20 | |
| nome_turma | Nome da turma | Char | 80 | |
| horario_hora_inicial | Horário da Turma - Horário Inicial - Hora | Char | 2 | |
| horario_minuto_inicial | Horário da Turma - Horário Inicial - Minuto | Char | 2 | |
| horario_hora_final | Horário da Turma - Horário Final - Hora | Char | 2 | |
| horario_minuto_final | Horário da Turma - Horário Final - Minuto | Char | 2 | |
| domingo | Domingo | Num | 1 | 0 - Não , 1 - Sim |
| segundafeira | Segunda-feira | Num | 1 | 0 - Não , 1 - Sim |
| tercafeira | Terça-feira | Num | 1 | 0 - Não , 1 - Sim |
| quartafeira | Quarta-feira | Num | 1 | 0 - Não , 1 - Sim |
| quintafeira | Quinta-feira | Num | 1 | 0 - Não , 1 - Sim |
| sextafeira | Sexta-feira | Num | 1 | 0 - Não , 1 - Sim |
| sabado | Sábado | Num | 1 | 0 - Não , 1 - Sim |

***http://dados.recife.pe.gov.br/***

# ANÁLISE DESCRITIVA DE DADOS




**- Não é Mineração de Dados, apenas descreve os dados.**

**- Resulta na descoberta de padrões consistentes.**

**- Técnicas:**

- Distribuições de Frequência
- Técnicas de Visualização
- Medidas de Resumo

https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

# Mammographic Mass Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age.

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 961 | **Area:** | Life |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 6 | **Date Donated** | 2007-10-29 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 73032 |

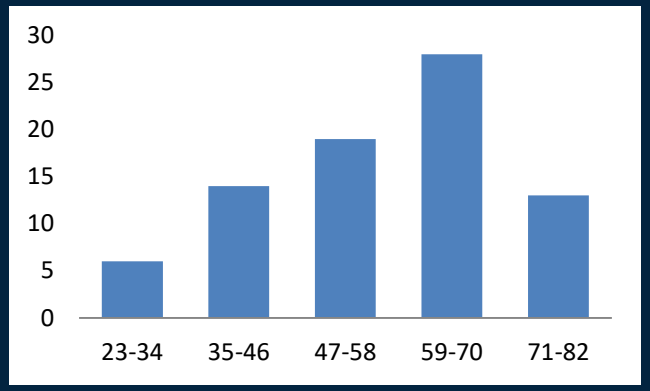
***Os dez primeiros registros...***

| ID | BI-RADS | Idade | Forma | Margem | Densidade | Severidade |
|---|---|---|---|---|---|---|
| 1 | 5 | 67 | Lobular | Especulada | Baixa | Maligno |
| 2 | 4 | 43 | Redonda | Circunscrita | ? | Maligno |
| 3 | 5 | 58 | Irregular | Especulada | Baixa | Maligno |
| 4 | 4 | 28 | Redonda | Circunscrita | Baixa | Benigno |
| 5 | 5 | 74 | Redonda | Especulada | ? | Maligno |
| 6 | 4 | 65 | Redonda | ? | Baixa | Benigno |
| 7 | 4 | 70 | ? | ? | Baixa | Benigno |
| 8 | 5 | 42 | Redonda | ? | Baixa | Benigno |
| 9 | 5 | 57 | Redonda | Especulada | Baixa | Maligno |
| 10 | 5 | 60 | ? | Especulada | Alta | Maligno |

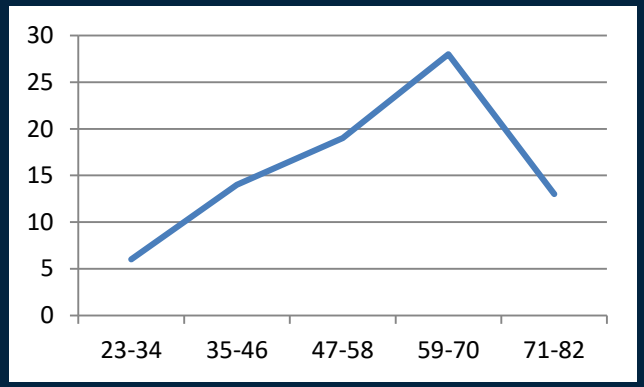
# DISTRIBUIÇÃO DE FREQUÊNCIA

| Classe | Limite inferior | Ponto médio | Limite superior | Frequência absoluta | Frequência relativa |
|---|---|---|---|---|---|
| 1 | 23 | 28,5 | 34 | 5 | 6,25% |
| 2 | 35 | 40,5 | 46 | 15 | 18,75% |
| 3 | 47 | 52,5 | 58 | 20 | 25% |
| 4 | 59 | 64,5 | 70 | 28 | 35% |
| 5 | 71 | 76,5 | 82 | 15 | 15% |

# VISUALIZAÇÃO DOS DADOS




Histograma




Polígono de frequência

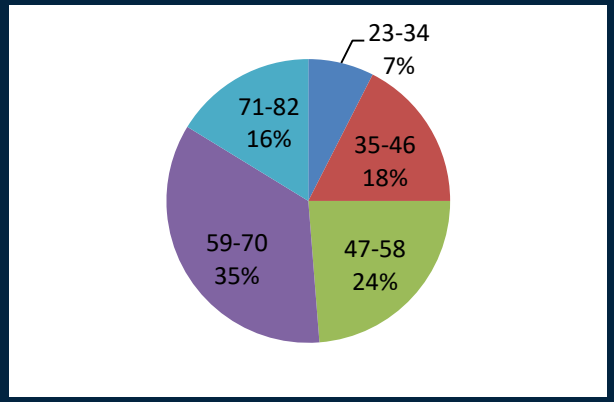



Gráfico de setores

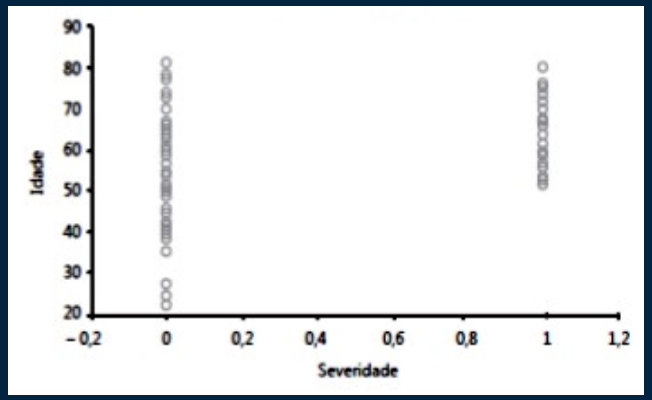
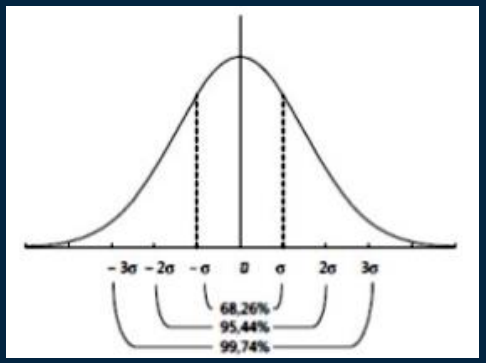



Gráfico de dispersão

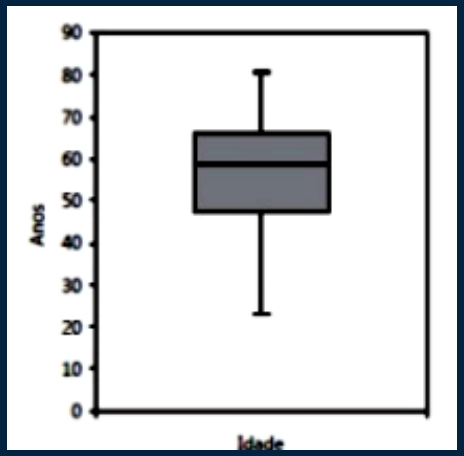
# MEDIDAS DE RESUMO

| Idade | |
|---|---|
| Média | 56,49 |
| Mediana | 58,50 |
| Ponto Médio | 52,00 |
| Moda | 60,00 |

**Tendência Central**




**Distribuição normal**




**Box plot**

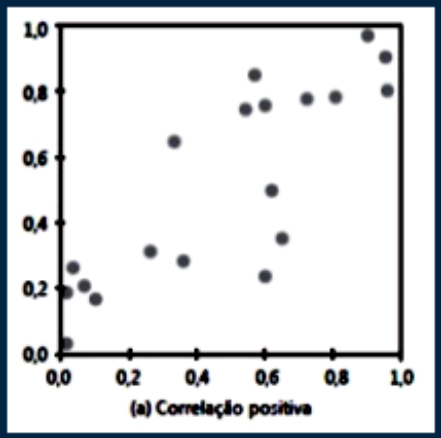



**Diagrama de dispersão**

# Alguns sites de bases

- http://www.kdnuggets.com/datasets/
- http://kdd.ics.uci.edu/
- https://www.kaggle.com/datasets
- http://archive.ics.uci.edu/ml/datasets.html

- http://dados.gov.br/dataset
- http://dados.recife.pe.gov.br/

# DINÂMICA

**- Utilize a base de dados de mamografia para realizar a análise descritiva de dados.**

1. **Distribuição de frequência**
   - **Limite inferior, superior, ponto médio, frequência absoluta e relativa**
2. **Visualização de Dados**
   - **Histograma, gráfico de dispersão**
3. **Medidas de Resumo**
   - **Médias, diagrama de dispersão, box plot**