

# CIÊNCIA DE DADOS (BIG DATA)

## ANÁLISE ESTATÍSTICA

**Professor curador:** Mário Olímpio de Menezes



**Mackenzie**



# TRILHA 3

## PARTE B – ANÁLISE EXPLORATÓRIA DE DADOS

# PARTE B – ANÁLISE EXPLORATÓRIA DE DADOS

# EXPLORANDO E ENTENDENDO OS DADOS

Exploramos nossos dados visando:

- Identificar relacionamentos entre variáveis.
- Verificar se existe alguma evidência a favor ou contra uma hipótese estabelecida.
- Verificar se ocorreu algum problema na coleta de dados.
- Identificar áreas onde mais dados são necessários.

# PRINCÍPIOS DA ANÁLISE EXPLORATÓRIA

- Não utiliza um conjunto rígido de técnicas.
- Mais como uma filosofia, uma atitude: é uma abordagem!
- Dentre as ferramentas mais utilizadas, estão os gráficos estatísticos.
- Visualização é uma parte crucial da AED (Análise Exploratória de Dados).

## CHECKLIST – ANÁLISE EXPLORATÓRIA DE DADOS

1. Formule suas questões.
2. Leia seus dados.
3. Verifique o conteúdo.
4. Verifique a estrutura.
5. Veja o topo e o final dos dados.
6. Verifique as quantidades.
7. Valide com pelo menos uma fonte de dados externa.
8. Tente a solução fácil primeiro.
9. Desafie sua solução.

# GRÁFICOS EXPLORATÓRIOS

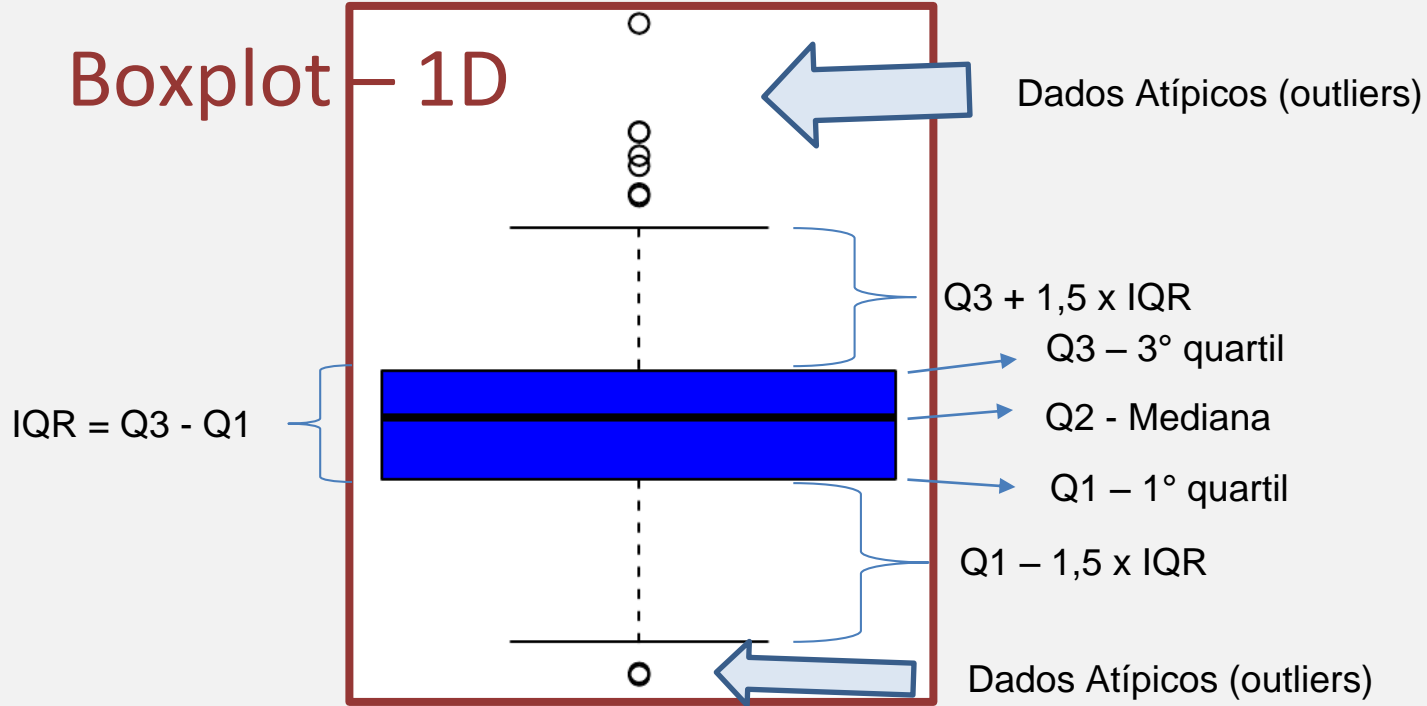
# GRÁFICOS EXPLORATÓRIOS

A visualização gráfica é essencial:

- Nos permite entender as propriedades básicas do conjunto de dados.
- Nos permite encontrar padrões simples nos dados.
- Nos permite sugerir possíveis estratégias de modelagem.

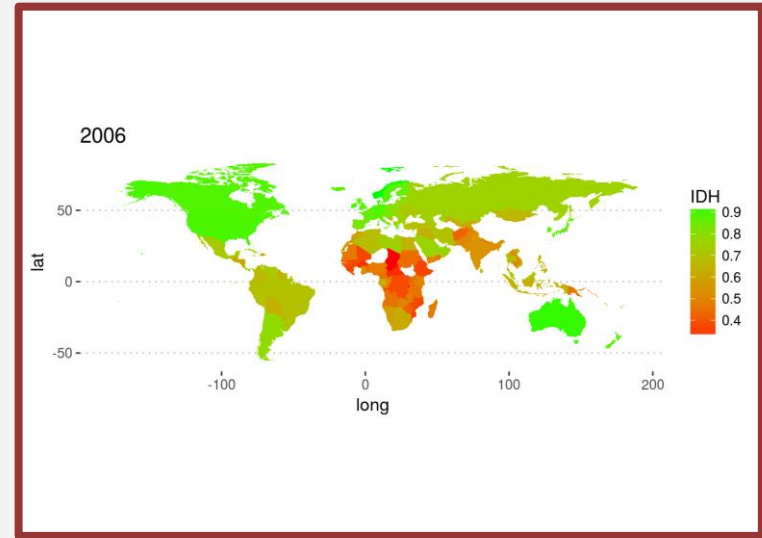


## Boxplot – 1D



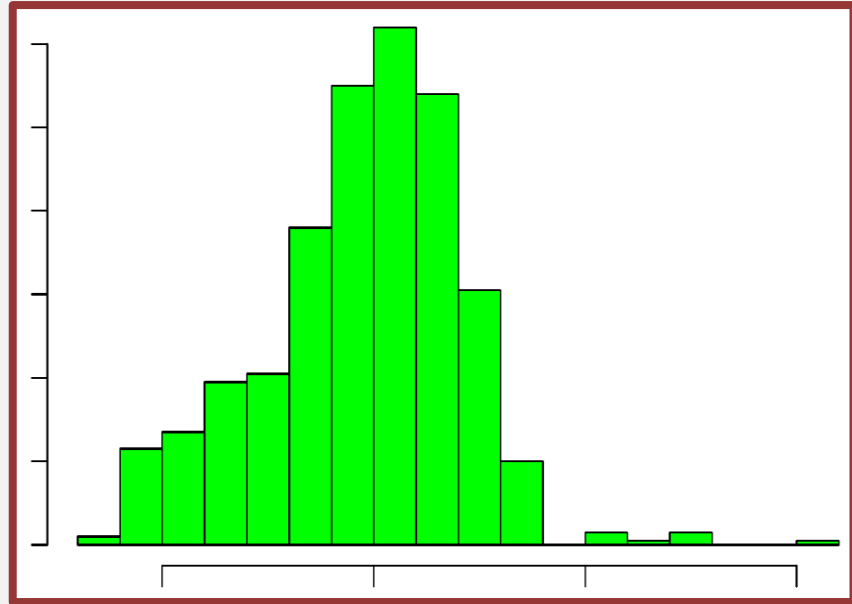
# Mapas

Mapas nos permitem visualizar informações geográficas e, assim, entender aspectos de localização nos dados.



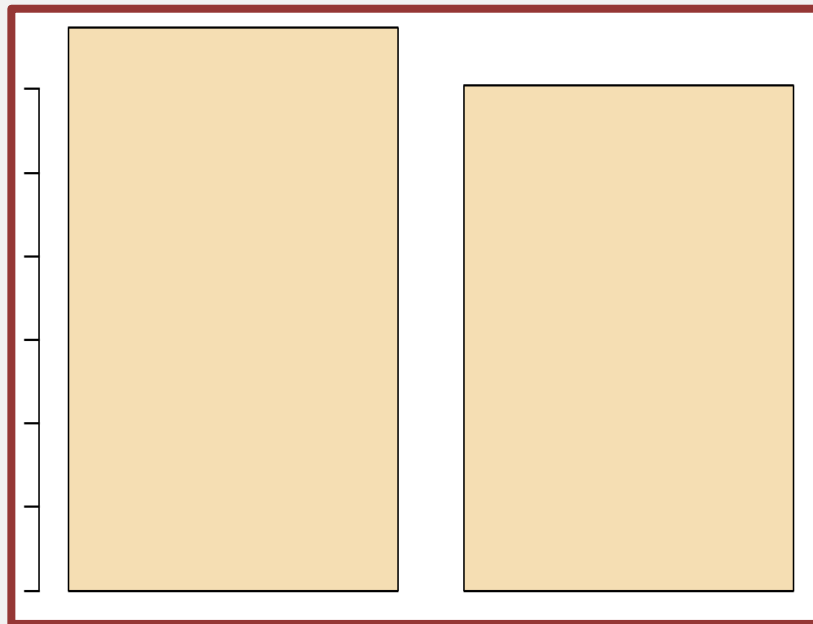
# HISTOGRAMA 1D

- Variável contínua.
- Faixas de valores (*breaks, bins*).
- Contagem de quantos estão em cada faixa.



# GRÁFICOS DE BARRAS

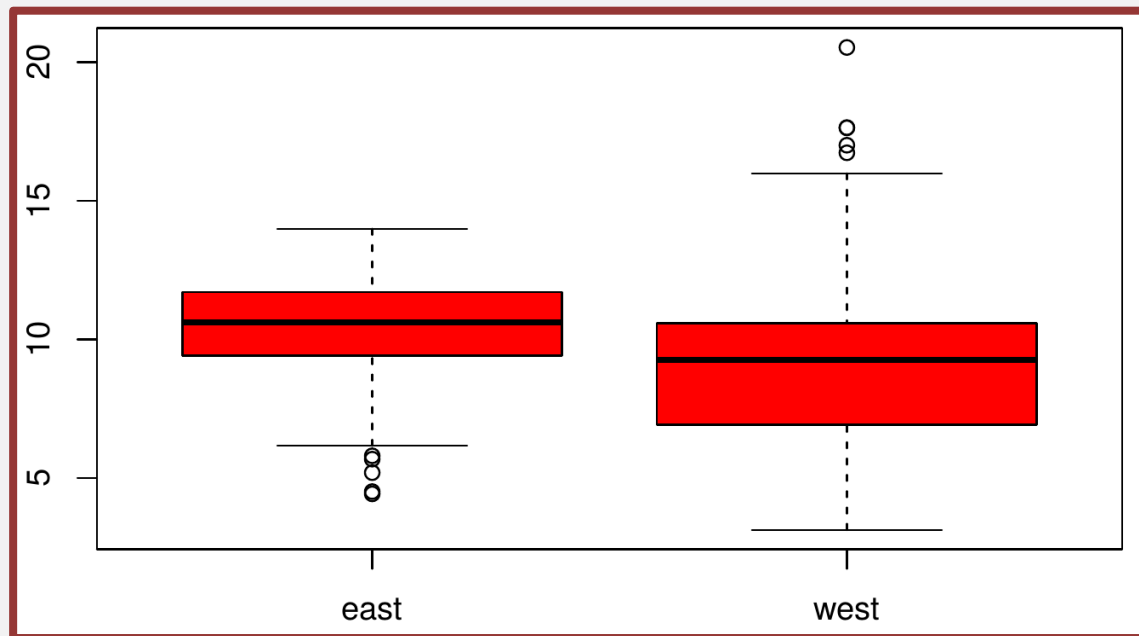
- Variáveis categóricas.
- Contagem de ocorrência de cada categoria.



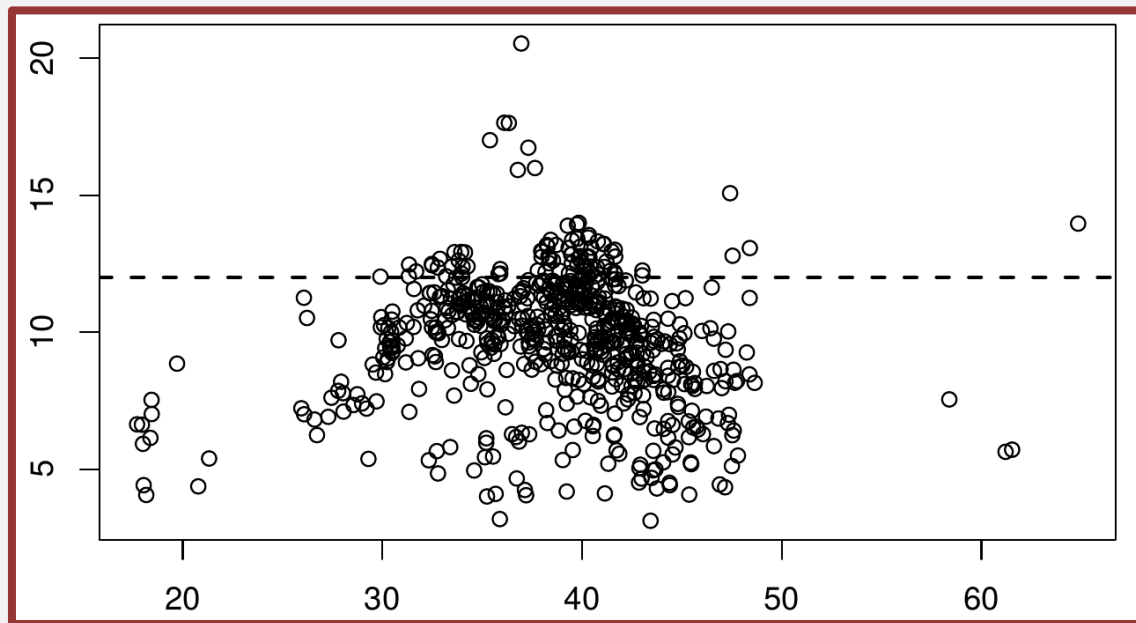
# DUAS DIMENSÕES E ALÉM

- Múltiplos gráficos unidimensionais (lado a lado).
- Gráficos de dispersão (*scatterplots*) para variáveis contínuas.
- Utilização de cor, tamanho, forma etc., para adicionar dimensões aos gráficos.

# MÚLTIPLOS BOXPLOTS

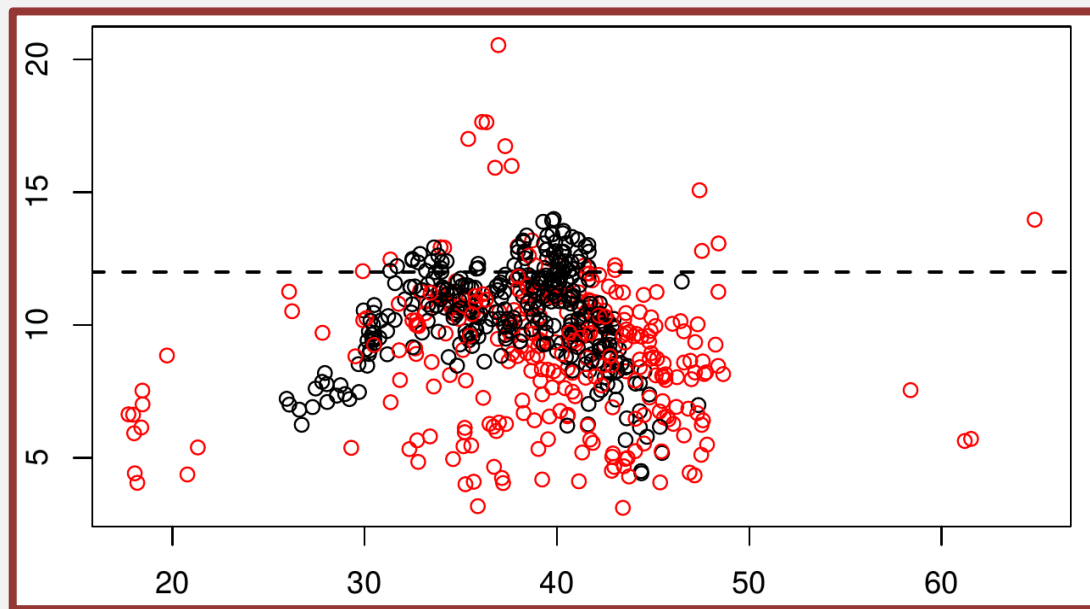


# GRÁFICO DE DISPERSÃO (SCATTERPLOT)



ANÁLISE ESTATÍSTICA

# USANDO COR PARA 3ª DIMENSÃO





ANÁLISE ESTATÍSTICA

# MÚLTIPLOS SCATTERPLOT

