



Universidade Estadual  
de Santa Cruz - UESC

# Estatística como base à Ciência de dados.

Mercado de trabalho, análise exploratória de dados e algoritmos de agrupamento.

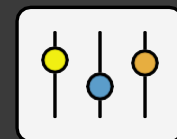
Discente: Solana Bonfim Lemos

Disciplina: Probabilidade e estatística.

Curso: Ciência da Computação

Semestre: 2025.1

Docente: José Cláudio Faria



# Carreira em Ciência de Dados

O que é Ciência de Dados?

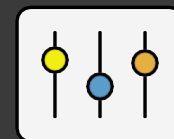
Campo interdisciplinar que combina estatística, programação e conhecimento de negócios para extrair insights de grandes volumes de dados.

Principais áreas de atuação:

Setores como finanças, saúde, varejo e tecnologia.

Funções incluem engenharia de dados, análise exploratória, modelagem preditiva e visualização de dados.

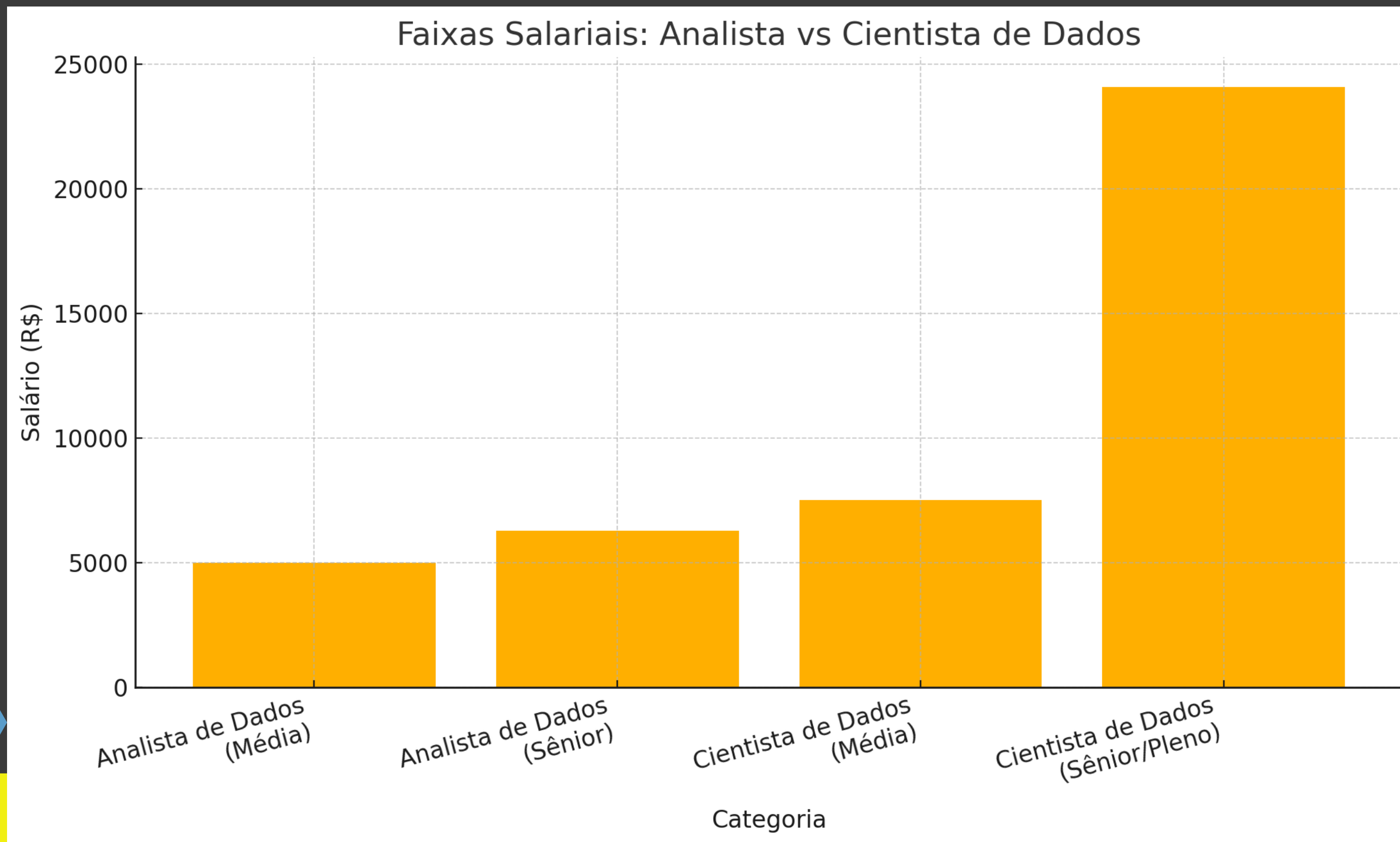
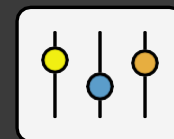


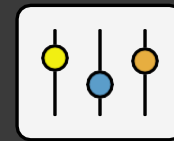


# Carreira em Ciência de Dados

Aspecto	Analista de dados	Cientista de dados
Foco principal	Análise descritiva e relatórios a partir de uma fonte de dados.	Análise preditiva e construção de modelos de Machine Learning.
Escopo de trabalho	Dados estruturados, SQL e visualização de dados.	Grandes volumes de dados, estatística avançada e programação.
Habilidades esperadas	SQL, Excel, ferramentas de BI (ex: Power BI, Tableau).	Python/R, bibliotecas de ML (ex: Scikit-Learn, TensorFlow).

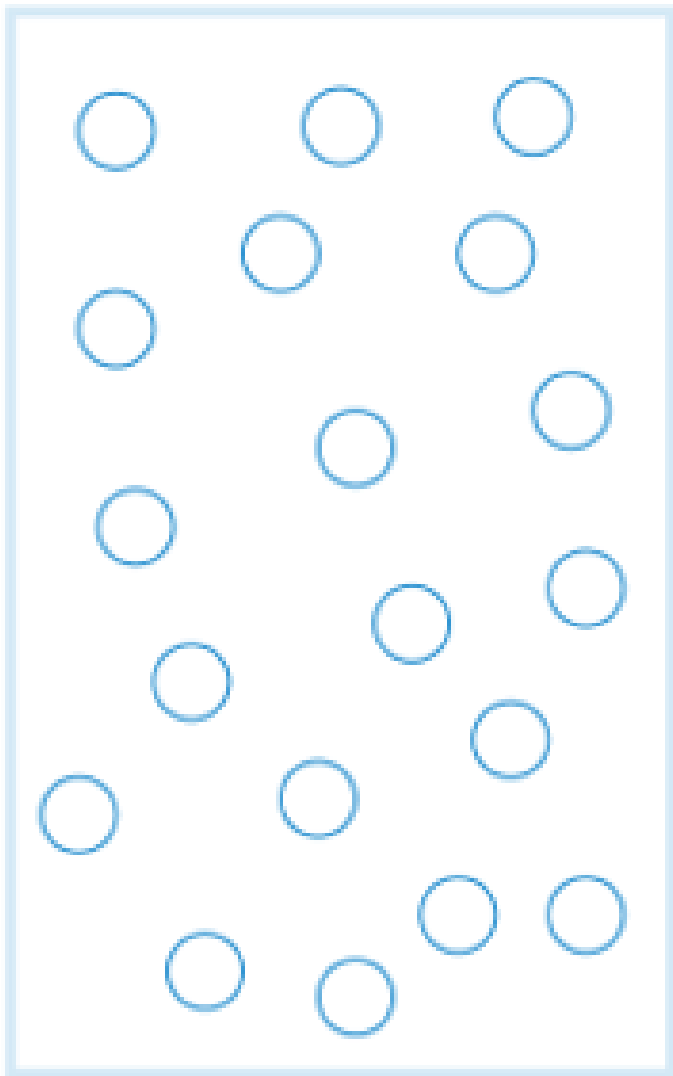
# Carreira em Ciência de Dados



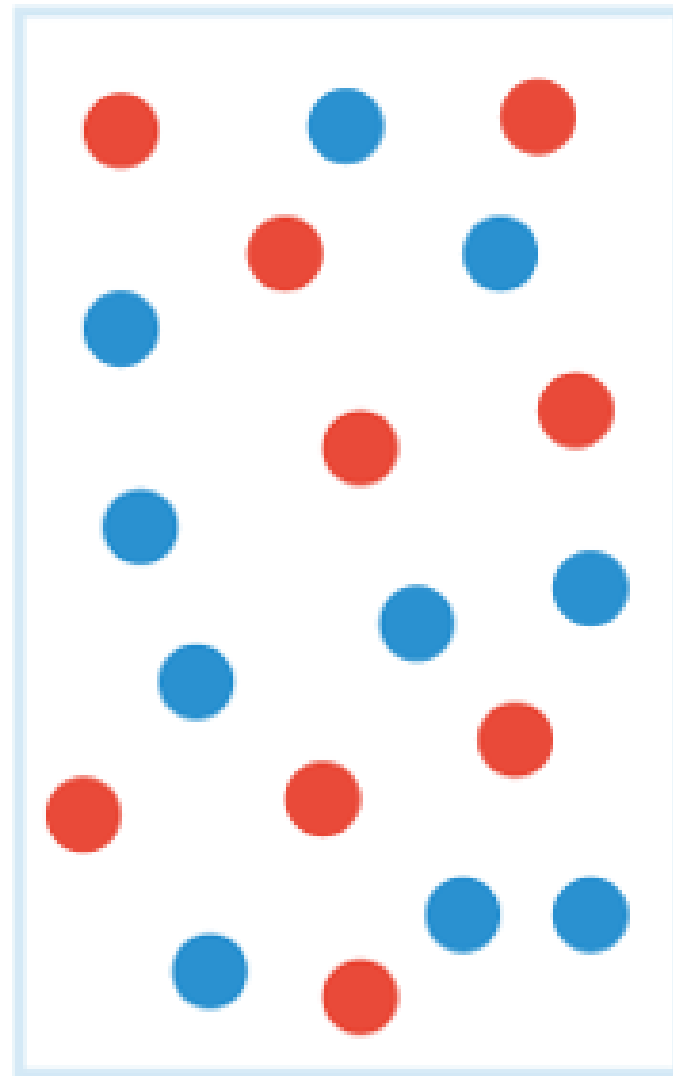


# Do dado à sabedoria

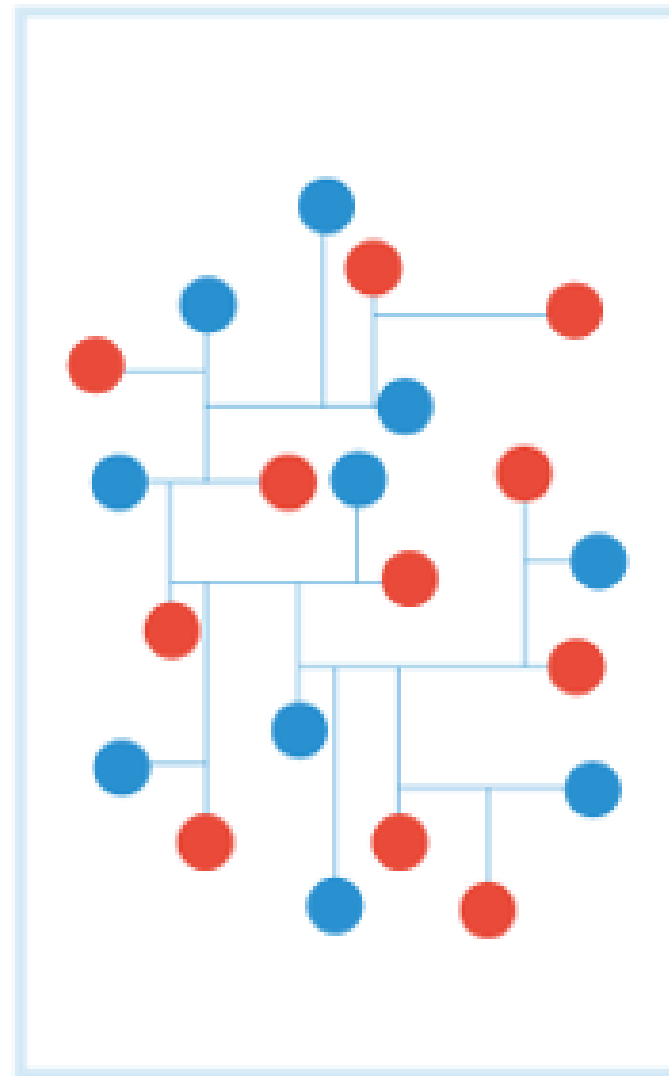
**Dados**



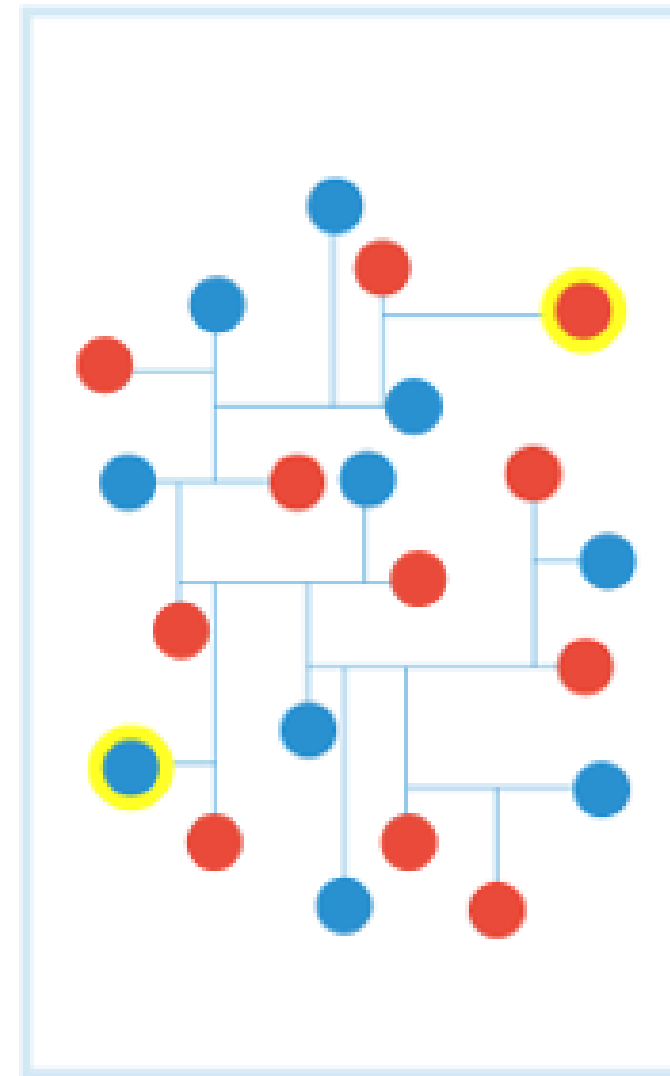
**Informação**



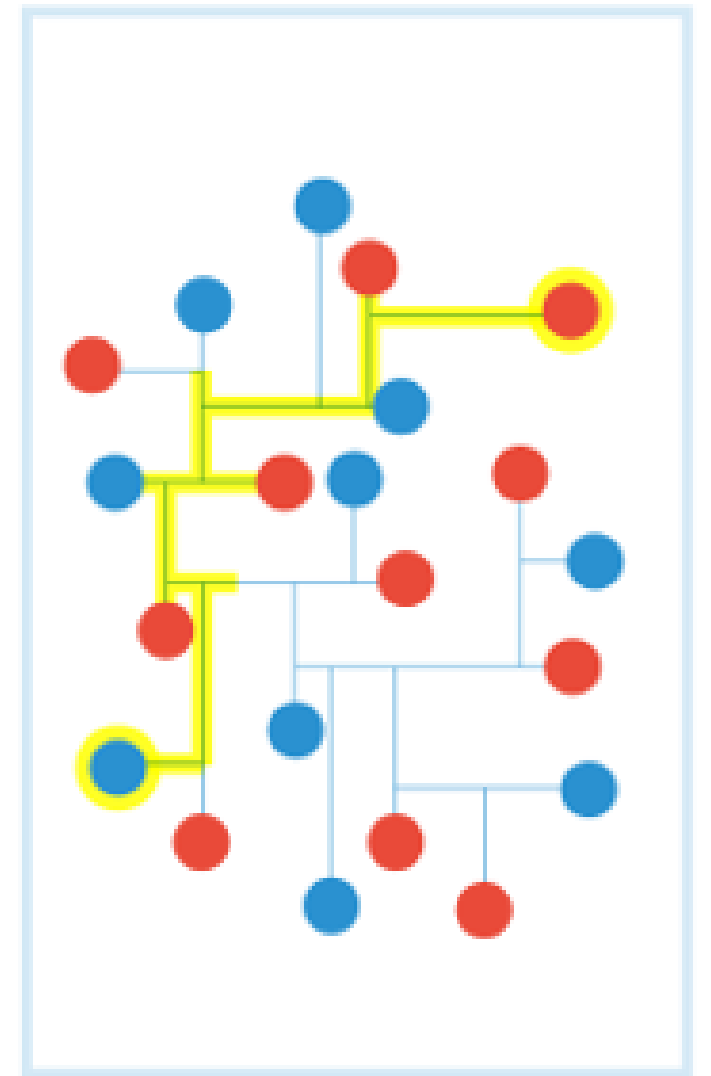
**Conhecimento**



**Insight**



**Sabedoria**





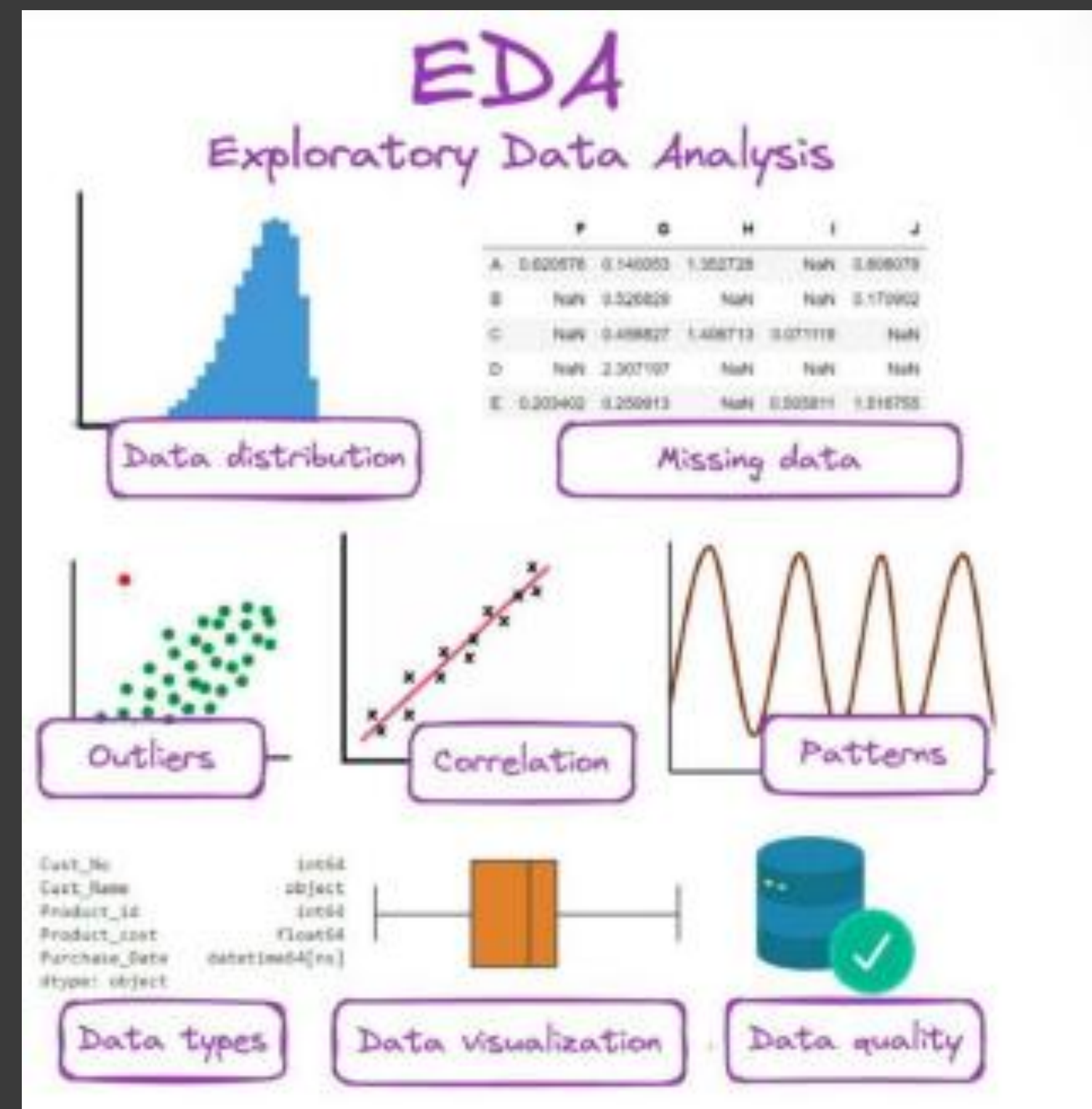


Universidade Estadual  
de Santa Cruz - UESC

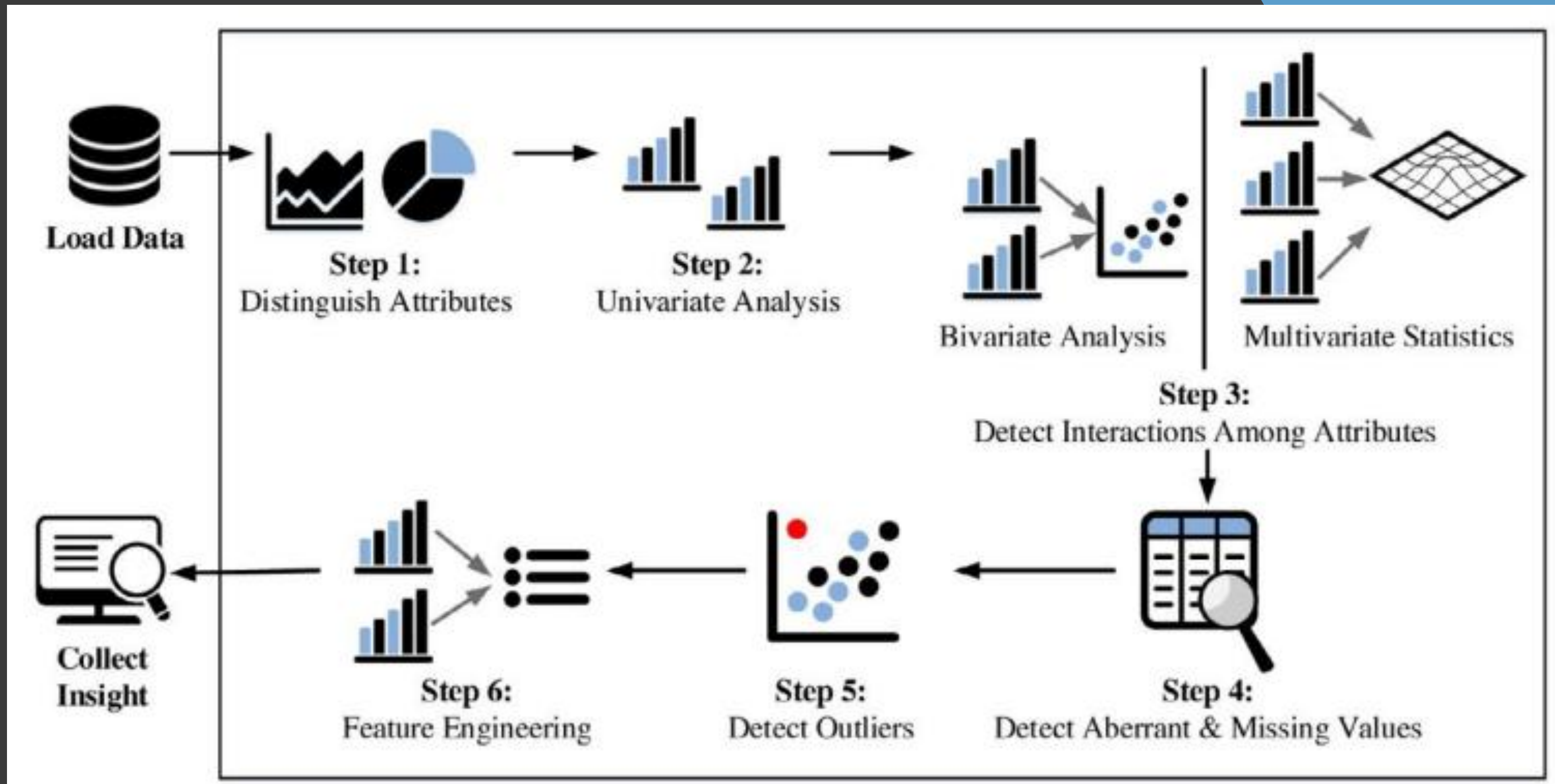
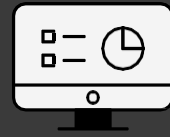


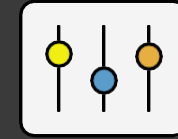
# Análise Exploratória de Dados

A Análise Exploratória de Dados (EDA) é uma etapa fundamental na ciência de dados que envolve a investigação inicial de um conjunto de dados para descobrir padrões e identificar anomalias.



# Fluxograma





# Tipos de Análises Exploratórias

## Univariadas:

- Histograma
- Boxplot
- Gráfico de Barras
- Sumário

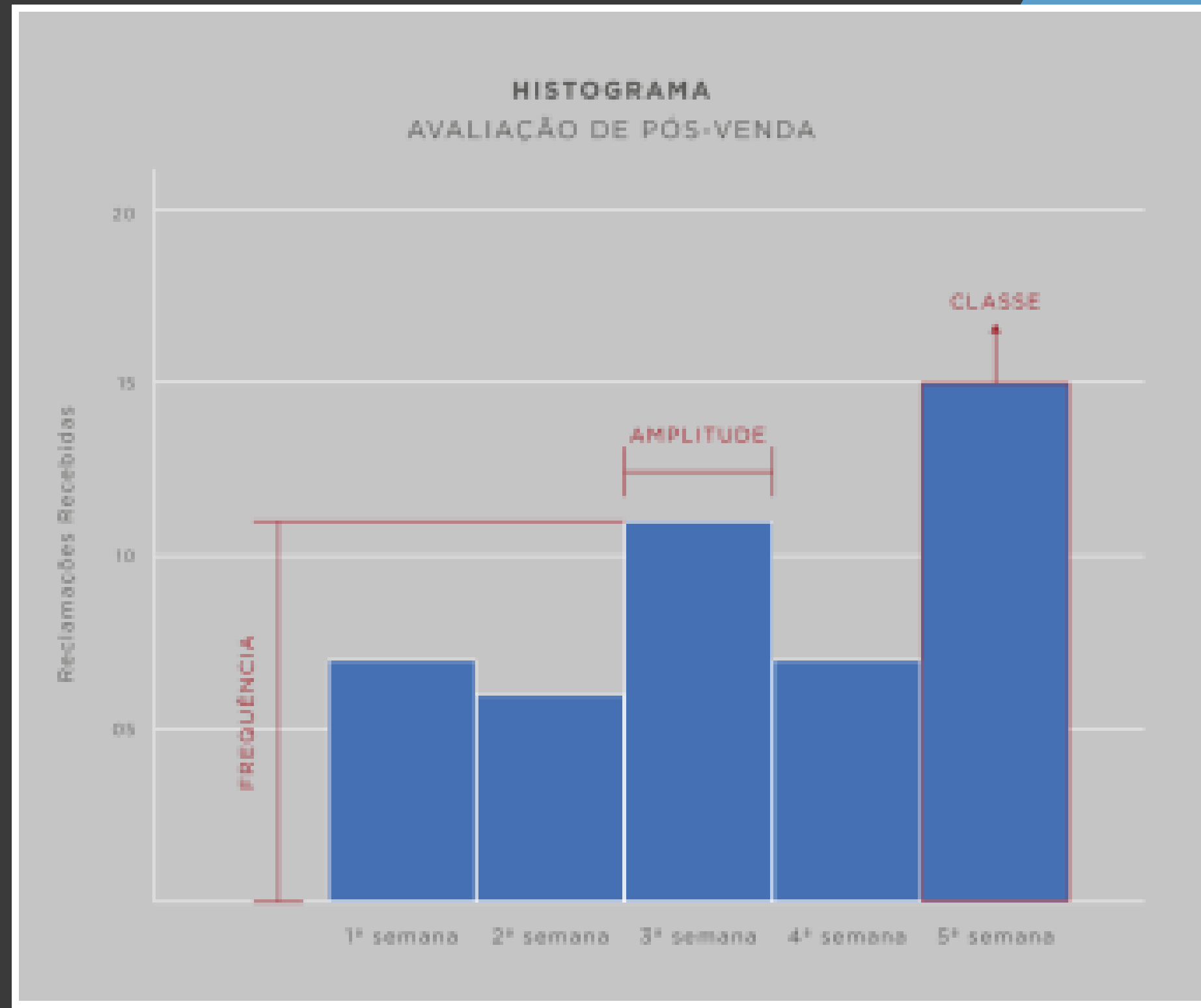


# Histograma



Outro nome para designar um histograma é “diagrama de dispersão de frequências”.

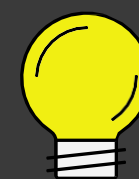
Resumidamente, esse recurso consiste em uma representação gráfica em barras ou em colunas.



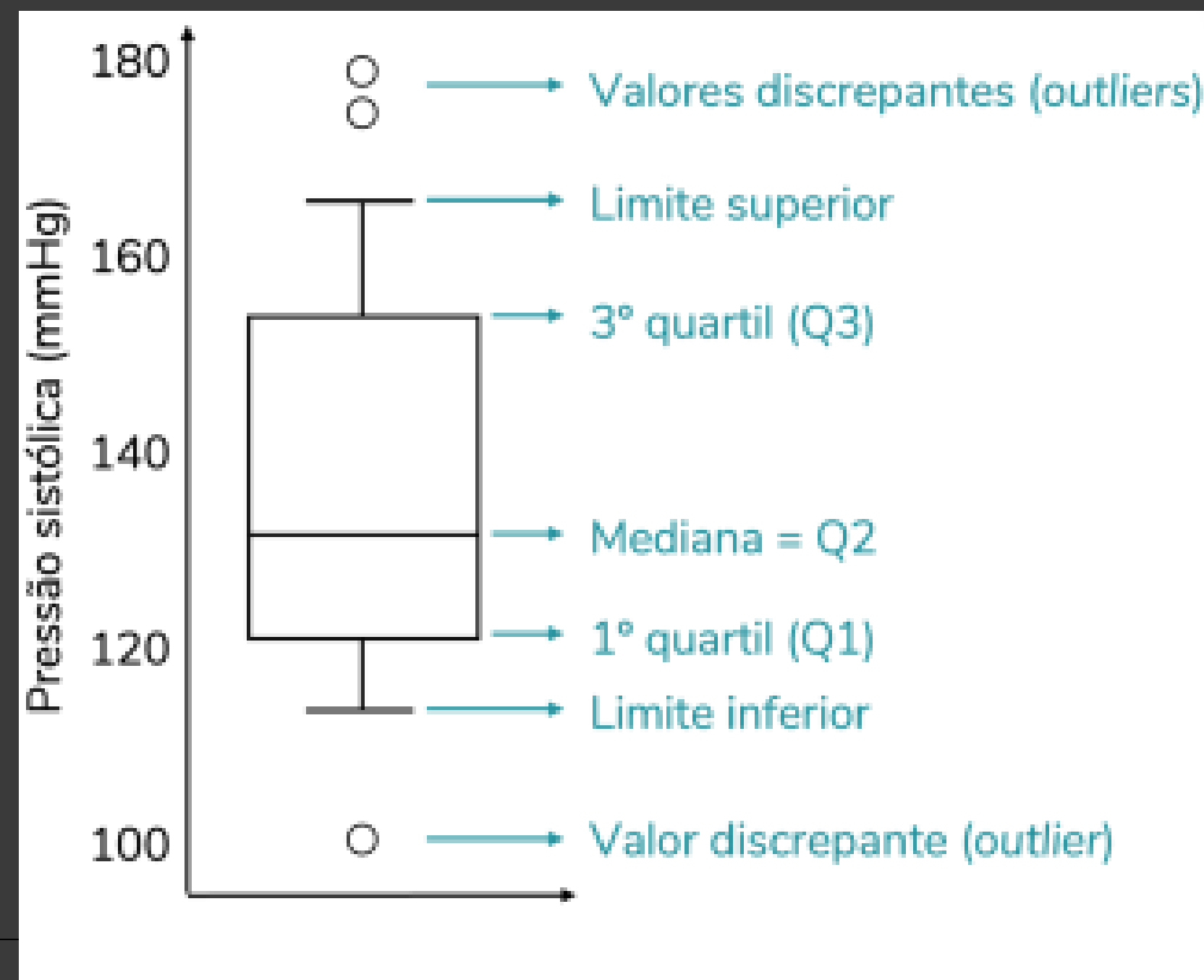


Universidade Estadual  
de Santa Cruz - UESC

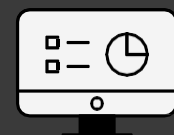
# Boxplot



O t-test foi desenvolvido por William Sealy Gosset em 1908 sob o pseudônimo "Student". O teste foi criado para comparar as médias de dois grupos e verificar se são estatisticamente diferentes.



# Gráfico de Barras

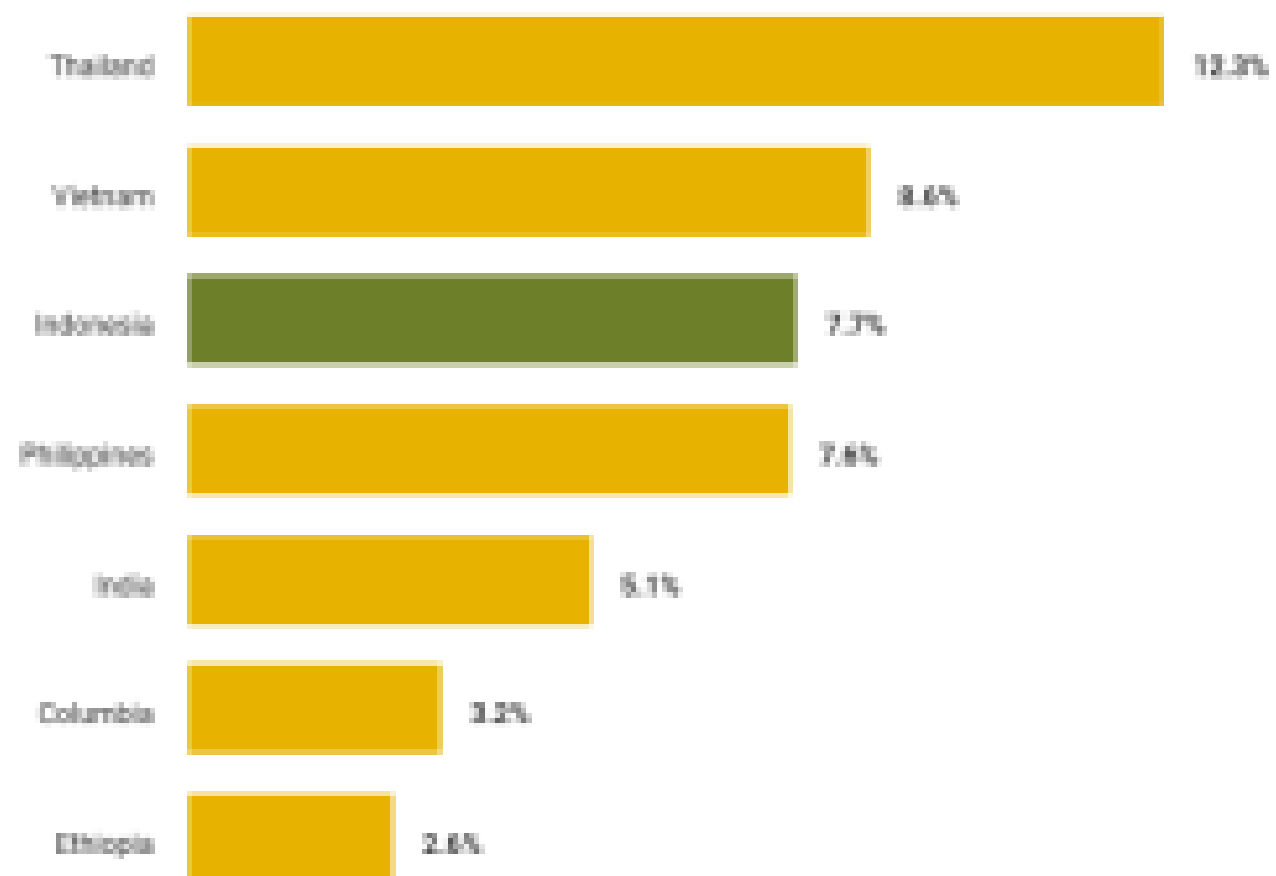


Um gráfico de barras é um gráfico que usa barras retangulares para representar dados visualmente em categorias. As categorias representadas podem ser qualquer coisa, desde profissões até anos, países (como no exemplo acima), grupos demográficos e muito mais.

## World Coffee Consumption

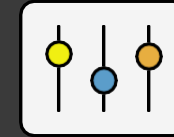
Today, Indonesia's coffee plantations cover a total area of approximately 1.24 million hectares, 933 hectares of robusta plantations and 307 hectares of arabica plantations. More than 90 percent of total plantations are cultivated by small-scale growers.

### Largest Annual Growth Rate in Coffee Exporting Countries



Source: International Coffee Organization

# Estatísticas de resumo - Sumário



Resume a variável quantitativa em: mínimo, máximo, média, mediana, 1o quartil, 3o quartil e dados não preenchidos. Caso a variável seja qualitativa, é informado o número de observações para cada nível.

Exemplo:

Resumo da variável salário

```
summary(dados$Salario)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.000	7.553	10.160	11.120	14.060	23.300

Resumo da variável salário apenas para casados

```
summary(dados$salario[dados$estciv=="Casado"])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.560	8.743	11.930	12.120	15.030	23.300

Resumo da variável salário apenas para solteiros

```
summary(dados$salario[dados$estciv=="Solteiro"])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.000	7.257	9.045	9.871	11.690	18.750

Resumo da variável qualitativa origem

```
summary(dados$Origem)
```

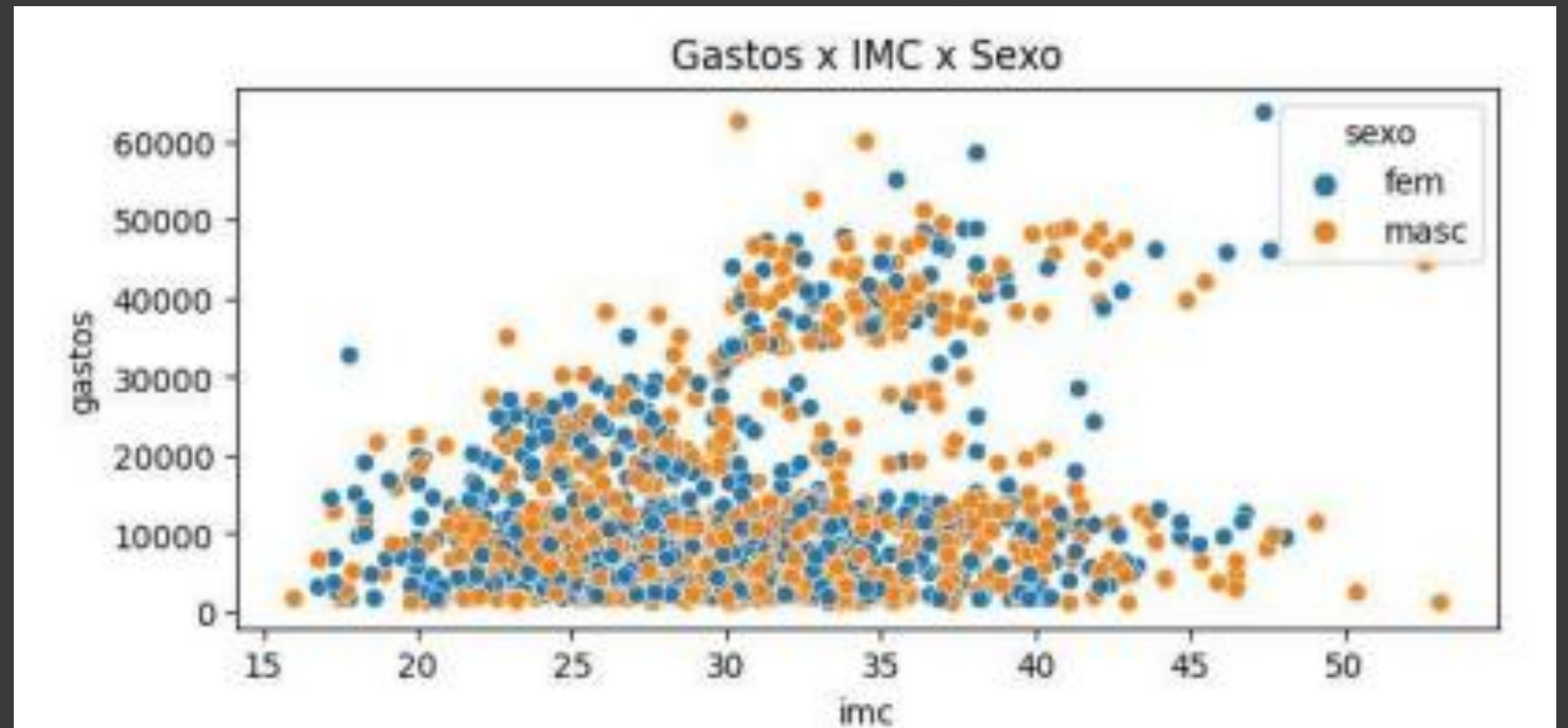
Capital	Interior	Outros
11	12	13

# Bivariada

A avaliação bivariada envolve a exploração da ligação entre variáveis.



Permite encontrar associações, correlações e dependências entre pares de variáveis. A análise bivariada é uma forma crucial de análise exploratória de dados que examina a relação entre duas variáveis.





# Gráfico de Dispersão - Scatter plot

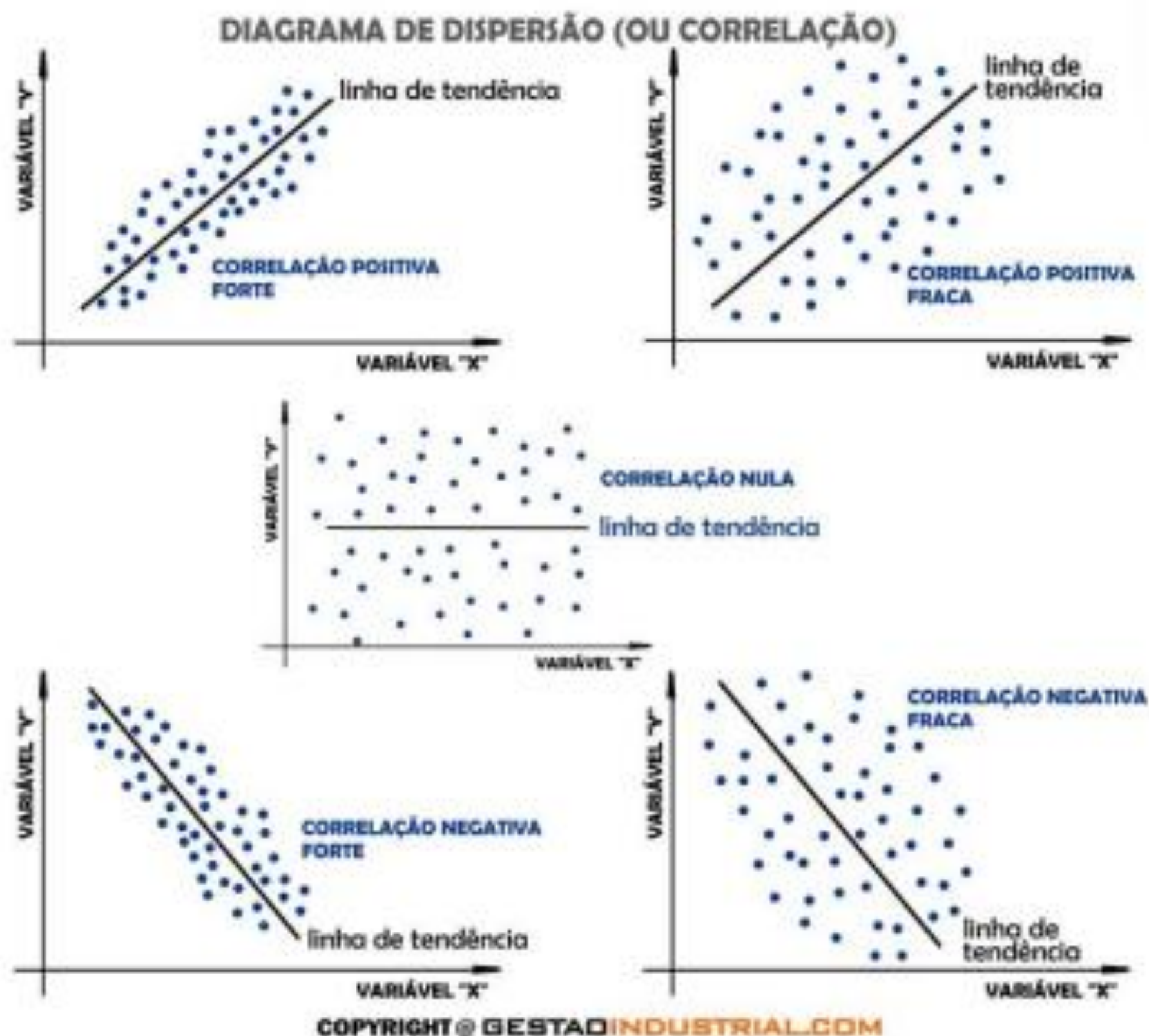
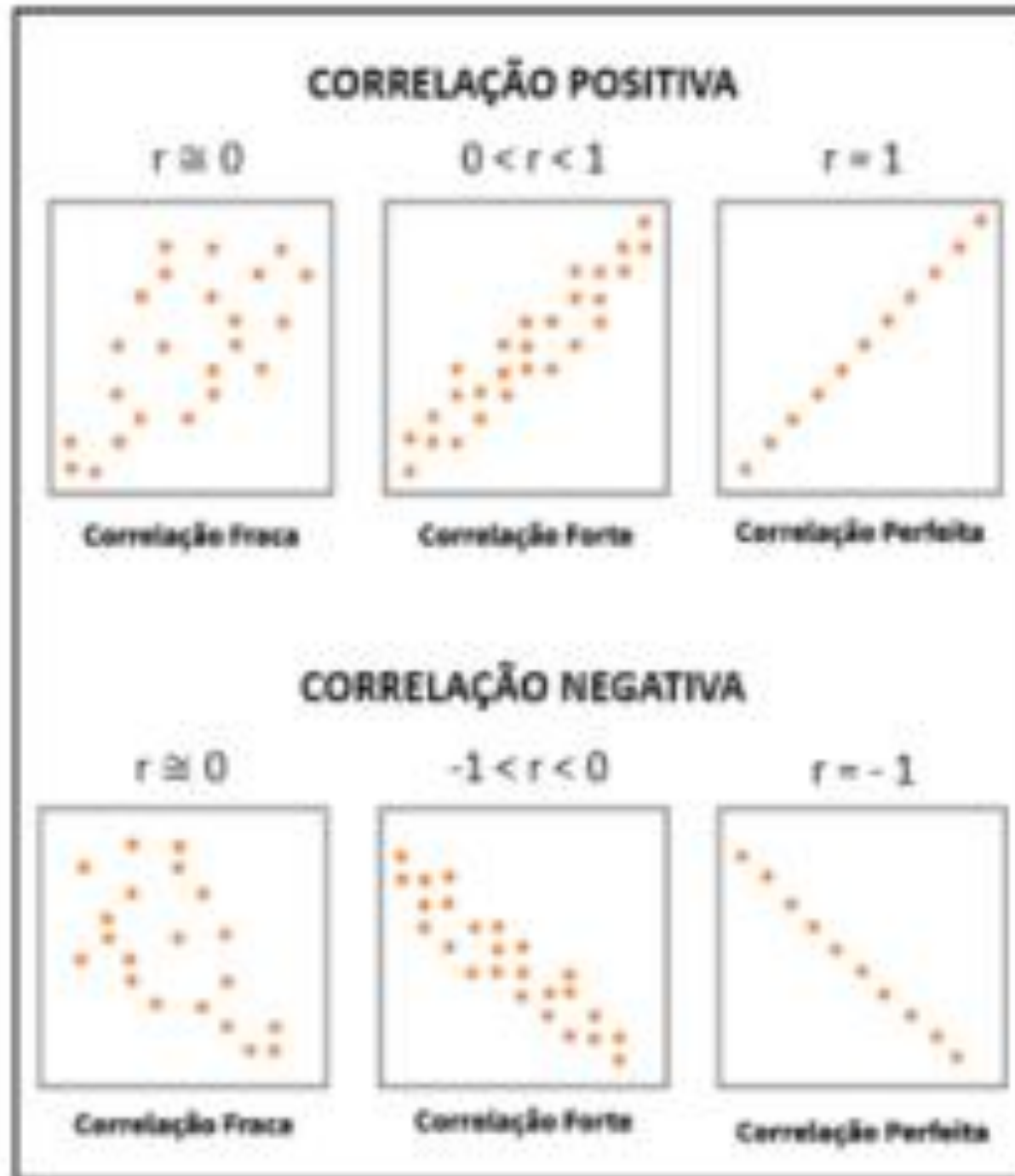
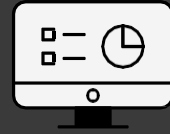


Gráfico de Dispersão (também conhecido como Gráfico de Dispersão, Gráfico de correlação ou Gráfico XY), é uma representação gráfica da possível relação entre duas variáveis e, dessa forma, mostra de forma gráfica os pares de dados numéricos e sua relação.

# Correlação



O tipo de correlação pode ser visualizado a partir do **Coefficiente de Correlação** (ou Coeficiente de Pearson -  $r$ ). Os valores obtidos neste coeficiente relacionam-se da seguinte forma:

$r = 0$ : Correlação nula ou inexistente entre variáveis.

$r = 1$ : Correlação positiva entre variáveis.

$r = -1$ : Correlação Negativa entre variáveis.

## **Correlação Positiva:**

No qual as duas variáveis crescem no mesmo sentido. Ou seja, enquanto um aumenta, o outro também aumenta.

## **Correlação negativa:**

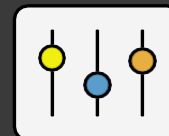
No qual as duas variáveis variam em sentidos contrários. Ou seja, enquanto um aumenta, o outro diminui.

## **Correlação Nula:**

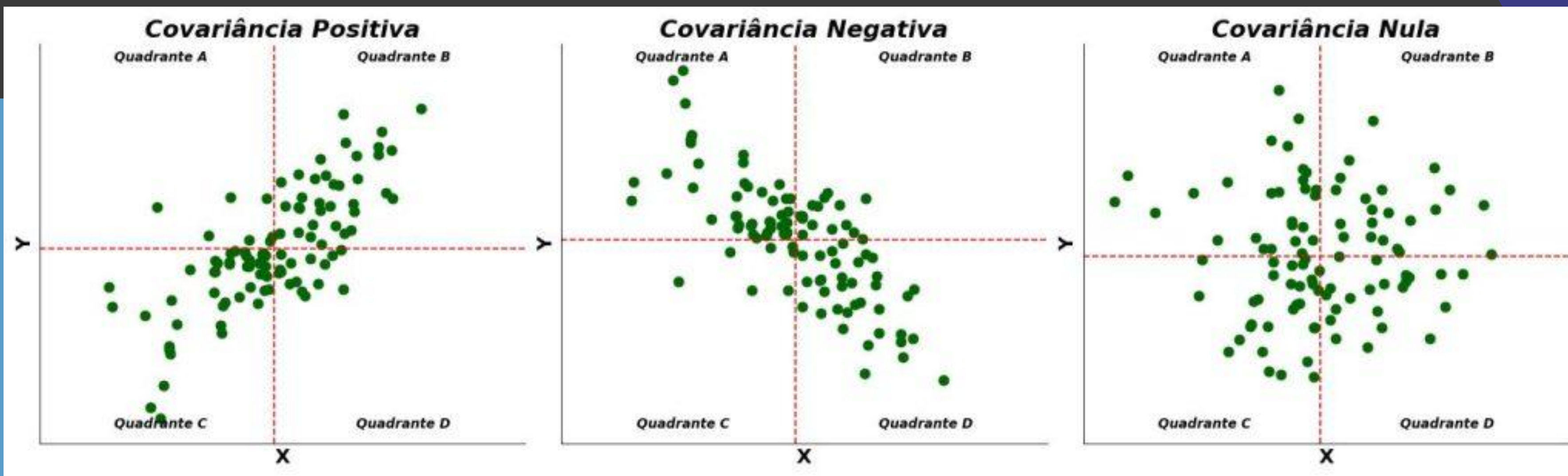
Não há interação entre variáveis.



# Covariância

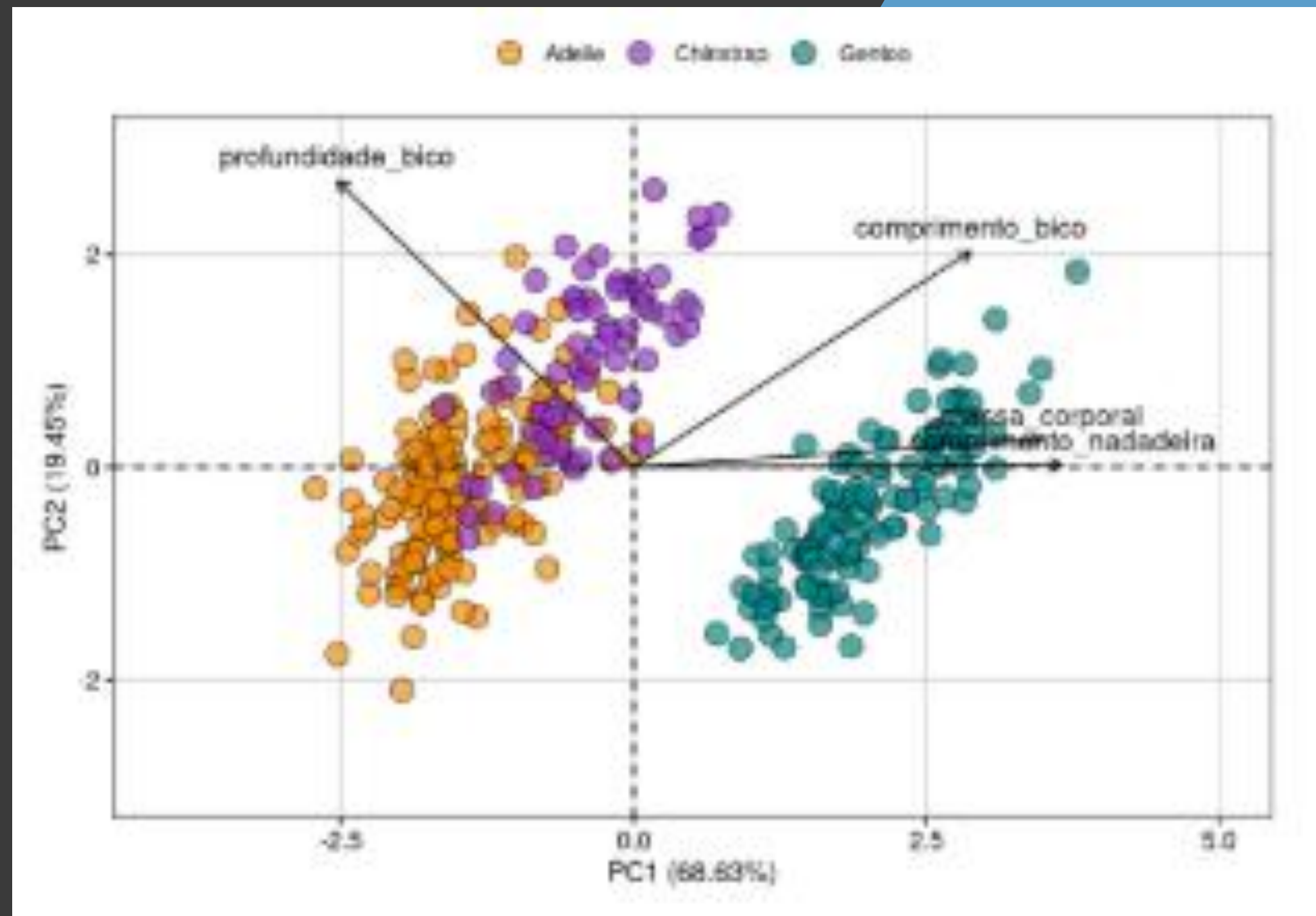


A covariância mede a relação linear entre duas variáveis. Ela é semelhante à correlação entre duas variáveis, no entanto, elas diferem nas seguintes maneiras: Os coeficientes de correlação são padronizados. Assim, um relacionamento linear perfeito resulta em um coeficiente de correlação 1. Os valores de covariância não, podem variar de menos infinito a mais infinito.



# Multivariada

A análise multivariada examina as relações entre duas ou mais variáveis no conjunto de dados. O seu objetivo é compreender a forma como as variáveis interagem entre si, o que é crucial para a maioria das técnicas de modelagem estatística.



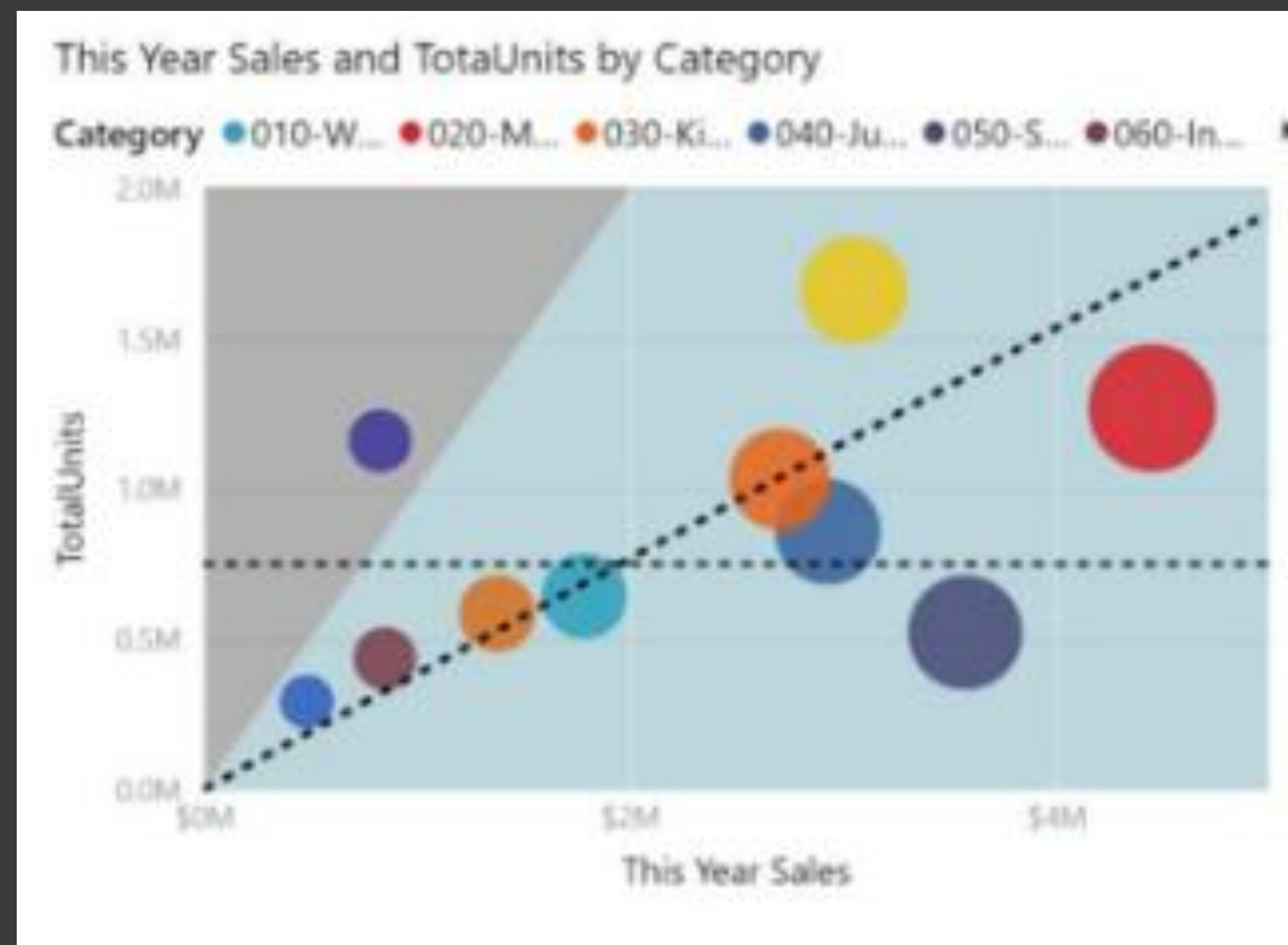


Universidade Estadual  
de Santa Cruz - UESC

# Gráfico de Pares

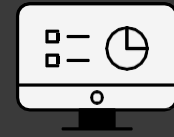


Permite a visualização de  
correlações entre várias variáveis  
simultaneamente.

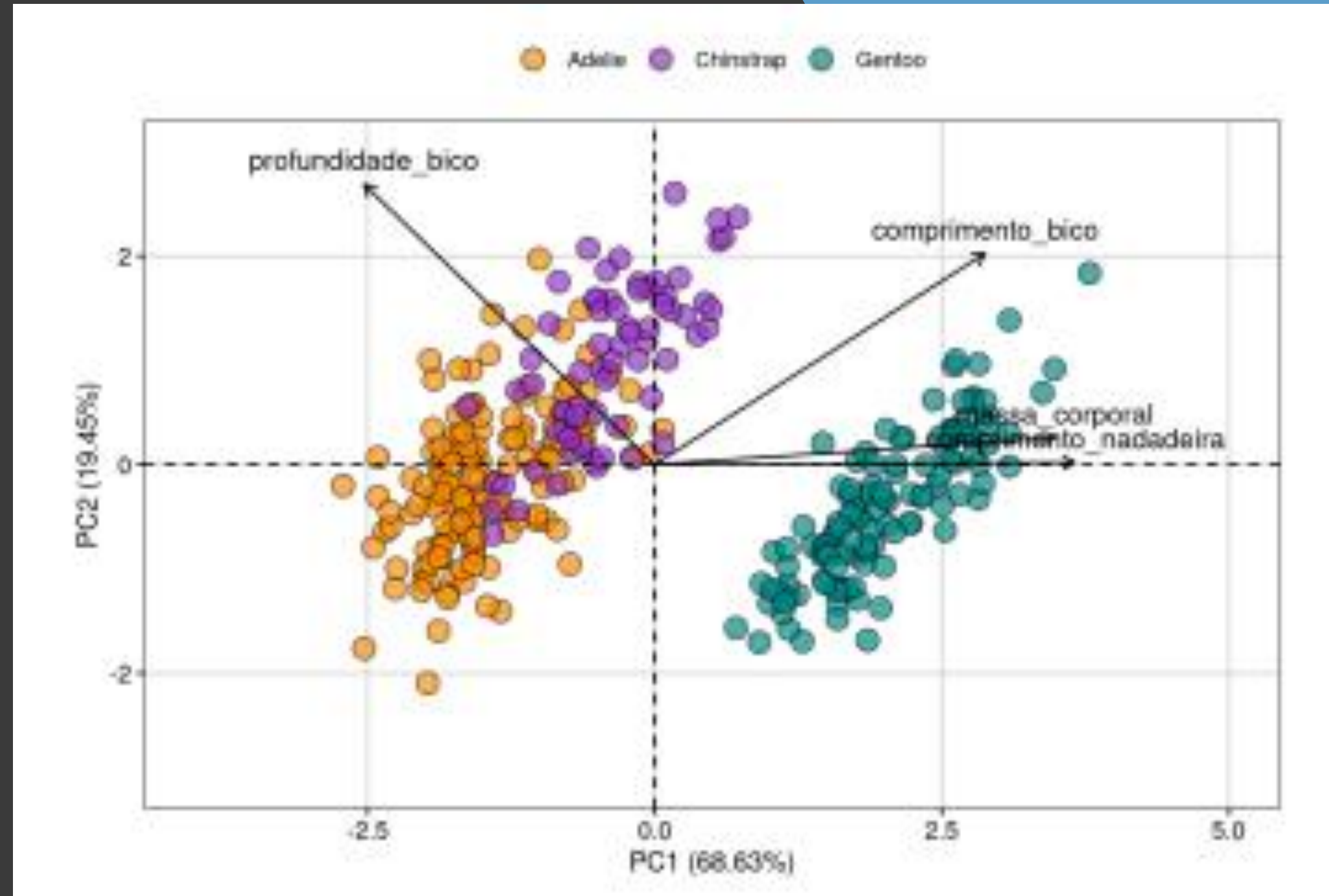




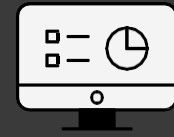
# PCA e Biplot



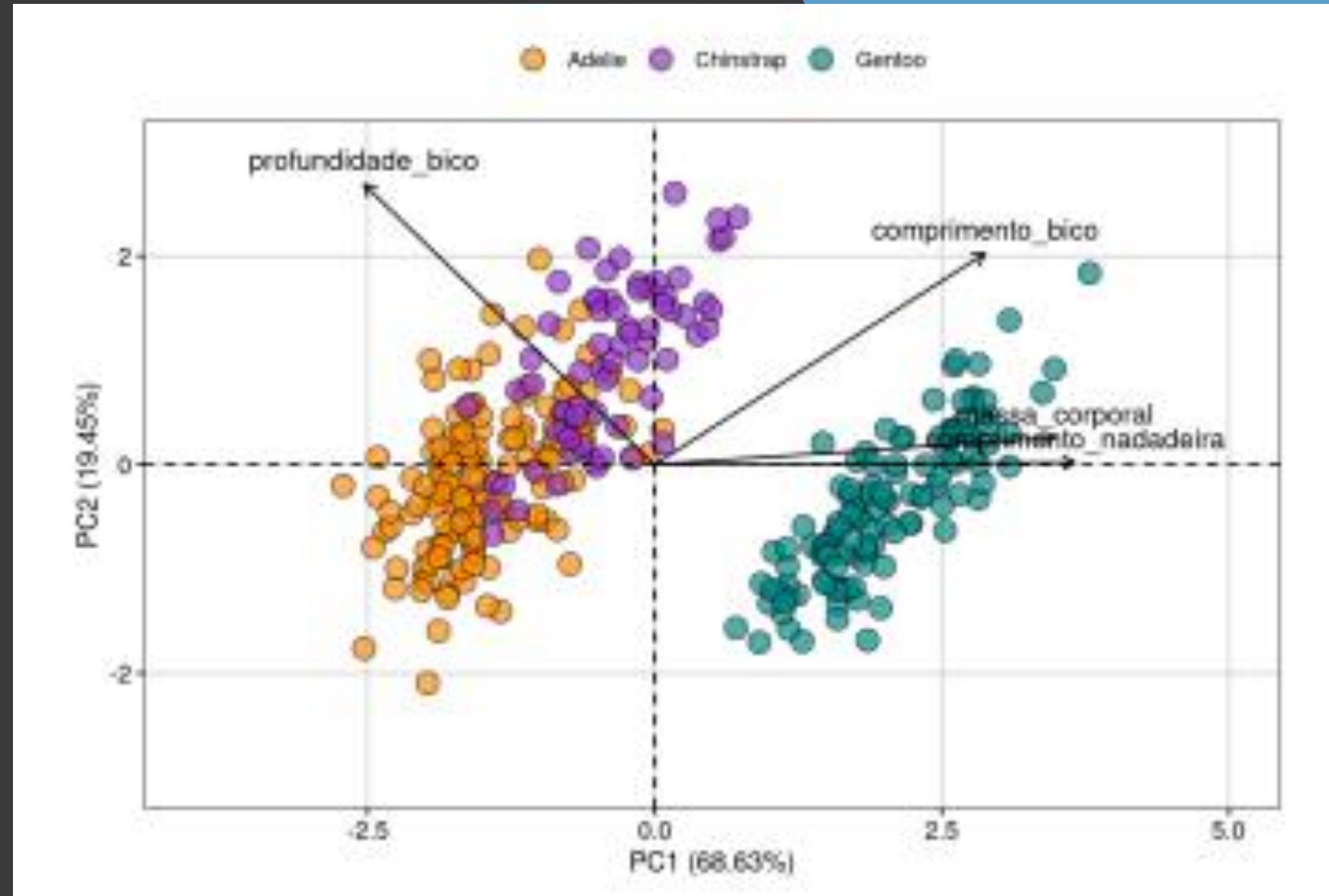
Análise de componentes principais (PCA): é uma técnica que permite reduzir a dimensionalidade de um conjunto de dados, transformando um grande número de variáveis em um número menor de variáveis não correlacionadas (chamadas de componentes principais) que explicam a maior parte da variação nos dados.



# PCA e Biplot



A técnica de biplot é uma representação gráfica que combina, em um único plano bidimensional, tanto as observações quanto as variáveis de um conjunto de dados multivariado, geralmente utilizando os dois primeiros componentes principais obtidos via Análise de Componentes Principais (PCA). Nesse gráfico, os pontos representam as observações (como indivíduos ou amostras) e as setas representam as variáveis originais, indicando sua direção e intensidade em relação aos componentes principais. O biplot permite visualizar correlações entre variáveis, identificar padrões ou agrupamentos entre observações e interpretar de forma mais intuitiva a estrutura dos dados em alta dimensão.



Obrigado pela Atenção!

