




# **PROJETO INTEGRADO: NEGÓCIOS**

---



**BOAS DECISÕES  
X  
BONS RESULTADOS**

# Pirâmide DICS

## Pirâmide DICS



Dados – são a base da pirâmide, provenientes de uma coleta ou pesquisa.

Informação – a informação surge a partir da estruturação ou organização de dados processados para um fim/contexto específico

Conhecimento – é composto por uma mescla de informação contextualizada, valores experiências e regras

Sabedoria – é o estágio mais complexo acrescenta o entendimento de quando se utilizar o conhecimento.

Essa pirâmide representa o dia-dia das empresas, elas possuem uma grande quantidade de dados, que devem ser trabalhados para gerar informações, conhecimento e sabedoria para a empresa.

# ANÁLISE DE DADOS

□ Procedimento para análise de dados: depende de três dimensões básicas:

- Número de variáveis a serem analisadas ao mesmo tempo
- Interesse da análise: descrição ou inferência
- Nível de mensuração das variáveis de interesse

□ Número de variáveis

- Análise univariada: o pesquisador objetiva analisar uma única variável por vez
- Análise bivariada: deseja-se analisar as relações entre 2 variáveis por vez
- Análise multivariada: deseja-se analisar as relações entre mais de 2 variáveis simultaneamente

# ANÁLISE DE DADOS

- Número de variáveis

- Análise multivariada

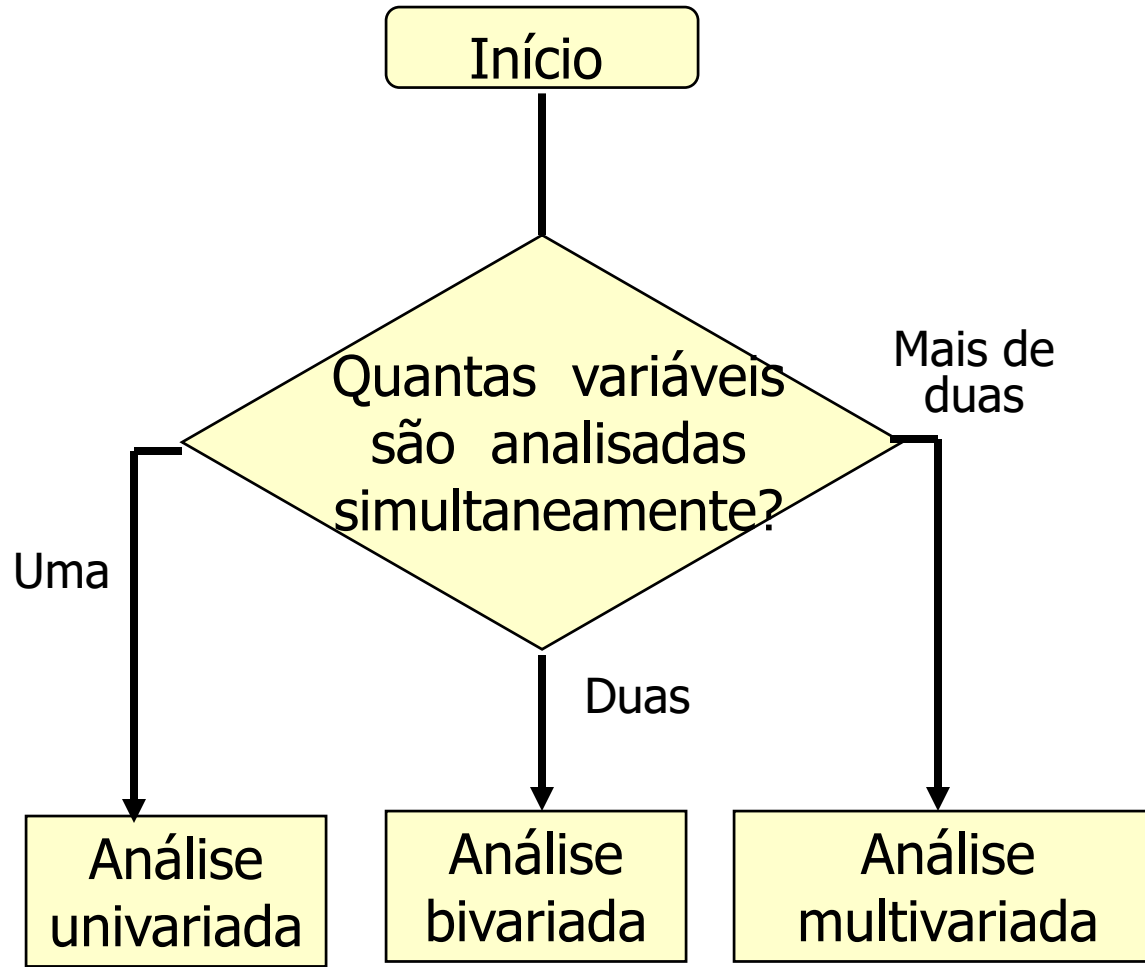
- Métodos de dependência (supervisionado): uma ou mais variáveis são expressas como dependentes de um conjunto de variáveis independentes.

- Ex: com base na renda, idade e estado civil, determinada pessoa é ou não é um bom risco de crédito?

- Métodos de interdependência (não supervisionado): nenhuma variável é expressa como dependente de outras.

- Ex: identificação e classificação de cidades pela similaridade do tamanho de sua população, da distribuição de renda, das raças e do consumo de produtos industrializados

# ANÁLISE DE DADOS



# ANÁLISE DE DADOS

- ▣ Interesse da análise
  - ▣ Descrição da amostra
  - ▣ Inferência sobre a população da qual se extraiu a amostra
- ▣ Nível de mensuração
  - ▣ As técnicas a serem usadas dependem da natureza de mensuração das variáveis de interesse
    - ▣ Nominal
    - ▣ Ordinal
    - ▣ Intervalar

# NATUREZA DE MENSURAÇÃO

<b>Qualitativas</b>	Têm como possíveis realizações características de NATUREZA não numérica.	Qualidade ou Atributo
<b>Quantitativas</b>	Têm como possíveis valores os números obtidos a partir de contagens ou mensurações.	Contagem ou Numeração



# NATUREZA DE MENSURAÇÃO

Qualitativas		Quantitativas (Intervalares)	
<b>Nominais</b> Possíveis valores não tem uma ordem natural	<b>Ordinais</b> Possíveis valores tem uma ordem natural	<b>Discretas</b> Assume valores em conjuntos finitos ou enumeráveis.	<b>Contínuas</b> Assume qualquer valor em intervalos dos números reais.
Exemplo: Cor dos olhos (Pretos, Castanhos, Azuis & Verdes)	Exemplo: Grau de Instrução (Fundamental, Médio, Superior)	Exemplo: Número de Apólices de seguro vendidas	Exemplo: Estatura dos Indivíduos

# NATUREZA DE MENSURAÇÃO

Identifique o tipo de escala (N=Nominal, O=Ordinal, I=Intervalar)

( ) Produtos bancários adquiridos (conta corrente, poupança, renda fixa, etc.)

( ) Forma de pagamento (à vista, 30 dd, 45 dd, ..)

( ) Data de pagamento

( ) Juros aplicados

( ) Escolaridade

( ) Tipo de canal de vendas utilizado (agência/loja, internet, cx. Eletrônico, ...)

( ) Cargo na empresa em que trabalho

( ) Tipo de residência (apto, casa, ...)

( ) Valor de um imóvel dado em garantia

( ) N° de televisores na residência

# ESCALA NOMINAL -VARIÁVEL DUMMY

- Variável não métrica transformada em variável métrica com a atribuição de 1 ou 0 a um objeto, dependendo de este ter ou não uma determinada característica
- Se houver  $I$  níveis, serão necessárias  $I-1$  variáveis dummies
  - Variável sexo
  - Poderiam ser usadas 2 variáveis dummies  $x_1$  e  $x_2$
  - No caso de sexo masculino, supor  $x_1 = 1$  e  $x_2 = 0$  ; no caso de sexo feminino  $x_1 = 0$  e  $x_2 = 1$ .
  - Se  $x_1 = 1$  fica implícito que  $x_2 = 0$
  - Usar apenas  $x_1$  ou  $x_2$  para representar a variável sexo
  - Variável com 3 níveis ex estado civil : usar 2 variáveis dummies  $x_1$  e  $x_2$ . Assim, casado  $x_1 = 1$ ,  $x_2 = 0$ , solteiro  $x_1 = 0$ ,  $x_2 = 1$ , divorciado  $x_1 = 0$ ,  $x_2 = 0$ .
- Recurso utilizado nas análises de regressão, discriminante e logística

# PADRONIZAÇÃO

---

Em algumas técnicas as variáveis necessitam ser padronizadas antes da aplicação. Isso acontece pois as variáveis em estudo podem ter diferentes escalas e mensurações.

Por exemplo, imagine que um analista está fazendo um estudo que leva em consideração a idade do cliente e o salário. A segunda teria uma faixa de valores bem maior, daí a necessidade de uma padronização se fizéssemos, por exemplo uma análise de cluster.

# PADRONIZAÇÃO

*Z scores : a variável padronizada  
terá média 0 e desvio-padrão 1*

$$\frac{X - \text{média}}{\text{desvio - padrão}}$$

Cod_Cliente	Total	Idade
C04561	100,9	29
C05571	155,18	32
C14571	1.200,29	44
C25812	305,88	22
C24598	17,02	25
C34521	437,76	31
C34567	2.520,55	32
C54541	43,36	45
C54572	16,88	27
C74560	26,37	29



Cod_Cliente	Total_P	Idade_P
C04561	-0,475695	-0,347302
C05571	-0,408016	0,053431
C14571	0,895073	1,656363
C25812	-0,220117	-1,282345
C24598	-0,580280	-0,881612
C34521	-0,055683	-0,080147
C34567	2,541232	0,053431
C54541	-0,547438	1,789940
C54572	-0,580455	-0,614457
C74560	-0,568622	-0,347302

# PADRONIZAÇÃO

O método Range confere à variável os valores com variação de 0 a 1.

$$\frac{X - \text{mínimo}}{\text{amplitude}}$$

Cod_Cliente	Total	Idade
C04561	100,9	29
C05571	155,18	32
C14571	1.200,29	44
C25812	305,88	22
C24598	17,02	25
C34521	437,76	31
C34567	2.520,55	32
C54541	43,36	45
C54572	16,88	27
C74560	26,37	29

**Padronização**

Cod_Cliente	Total_P	Idade_P
C04561	0,033559	0,304348
C05571	0,055239	0,434783
C14571	0,472670	0,956522
C25812	0,115431	0,000000
C24598	0,000056	0,130435
C34521	0,168105	0,391304
C34567	1,000000	0,434783
C54541	0,010576	1,000000
C54572	0,000000	0,217391
C74560	0,003790	0,304348

# OUTLIERS

- Observações com características diferentes das demais
- Efeito benéfico: indicadores de características da população que não seriam descobertas no curso normal da análise
- Efeito prejudicial: distorção de testes estatísticos
- Fontes de outliers:
  - Erro de coleta ou de codificação
  - Resultado de um evento extraordinário detectável
  - Resultados extraordinários inexplicáveis
  - Observações com valores possíveis mas com combinações extraordinárias

# DADOS MISSING

- ▣ Identificar as causas dos missings
  - ▣ Ex pergunta constrangedora - declaração da renda familiar pontual
  - ▣ Falha na coleta de dados
  
- ▣ Verificar se o padrão dos missings é aleatório
  - ▣ Ex verificar o perfil dos respondentes para os casos válidos e para os missings. Se não houver diferença significativa no perfil, o padrão será aleatório



# DADOS MISSING

- ▣ Decisões sobre os missings
  - ▣ Só considerar observações com dados completos
  - ▣ Eliminar as variáveis com missings
  - ▣ Estimar os valores dos missings a partir dos valores válidos
    - ▣ Substituição pela média
    - ▣ Substituição pelo ajuste de uma regressão

# TIPOS DE MODELOS

Considere duas perguntas que podemos fazer sobre nossos clientes, no intuito de segmentá-los:

1) Existem perfis de clientes diferentes em nossa base? Em caso positivo, quantos perfis?

2) Na nossa carteira de clientes, quem são aqueles que têm mais propensão a cancelar (virar churn) o produto nos próximos meses?

Nas duas perguntas queremos agrupar os clientes. Qual a diferença principal entre elas?

# TIPOS DE MODELOS

## Aprendizado Supervisionado

- Auxílio de um “professor” que nos diz algo a respeito dos objetos que observamos
- Na prática, se relaciona com resolução de problemas de classificação e regressão

## Aprendizado Não-Supervisionado

- Mesmo sem um “professor” somos capazes de identificar padrões nos objetos que observamos
- Na prática, se relaciona com problemas de agrupamento e geração de regras de associação

# DESENVOLVIMENTO DE MODELOS

Uma vez que já possuímos uma grande gama de dados, precisamos saber o que fazer com eles. Nesse momento entram em ação técnicas estatísticas e/ou Machine learning para auxiliar-nos a minerar os dados e nos trazer resultados úteis do ponto de vista de negócio

