Como funciona o processo de análise de dados?

Instrutora: Beatriz Alves - Senai Gru

1 - Verifique o problema

- Setor
- Pessoas envolvidas
- Metas
- Objetivos

Preparação. A partir de um gap crie um objetivo.

2 - Cronograma

- Dados que serão analisados
- Perguntas
- Quais métricas seriam analisadas

Preparação.

3 - Legislação e lgpd

- Dados éticos
- Mapeamento dos dados

Processo.

Como a LGPD impacta a análise de dados?

Dados públicos podem ser raspados, desde que não violem privacidade.

Marco Civil da Internet (Lei 12.965/2014)

Proíbe o acesso não autorizado a sistemas (Art. 154-A do Código Penal, se configurar invasão).

Recomenda-se respeitar o robots.txt (arquivo que define permissões de scraping).

Propriedade Intelectual (Lei 9.610/98)

Raspagem de conteúdo protegido por direitos autorais pode ser ilegal.

Como a LGPD impacta a análise de dados?

- A LGPD exige que as empresas adotem medidas de segurança para proteger os dados pessoais.
- As empresas devem revisar as suas práticas de coleta, armazenamento e processamento de dados.
- As empresas devem garantir a precisão dos dados.
- As empresas devem ser transparentes e responsáveis no tratamento das informações pessoais.

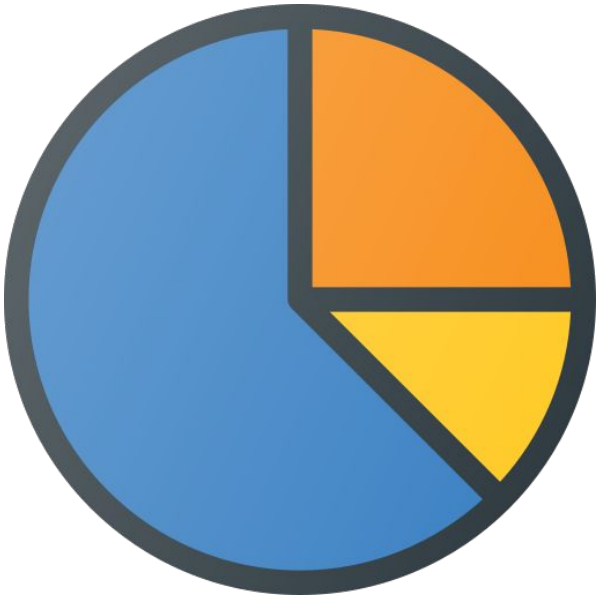
4 - Análise de dados

- Documentação
- Levantamento de informação
- Relatório
- Insight

Resultados da análise

5 - Ação das lideranças

- Tomada de decisão
- Estratégia



Ações a partir do insight

Processo de análise com as ferramentas Python

Problema: Uma empresa de e-commerce quer entender o preço médio, distribuição de marcas e avaliações de smartphones em um site de vendas (ex: Mercado Livre, Amazon) para definir estratégias de precificação e estoque.

```
<!-- Extrairemos dados de uma página de busca de smartphones (simulada).


import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt -->
```

```
url = "https://bea3853.github.io/PROCESSO_DATA_SCIENCE/"
headers = {'User-Agent': 'Mozilla/5.0'}
response = requests.get(url, headers=headers)
soup = BeautifulSoup(response.text, 'html.parser')
```

```
# Extraindo informações (adaptar conforme estrutura real do site)
for produto in soup.find_all('div', class_='produto'):
    nomes.append(produto.find('h2').text)
    precos.append(float(produto.find('span', class_='preco').text.replace('R$', '').replace('.', '').replace(',', '.')))
    avaliacoes.append(float(produto.find('span', class_='avaliacao').text))
```

```python
# Criando DataFrame
df = pd.DataFrame({
    'Modelo': nomes,
    'Preco': precos,
    'Avaliacao': avaliacoes
})

# Salvando em CSV (opcional)
df.to_csv('smartphones.csv', index=False)

#### 2. Limpeza e Análise Exploratória (Pandas + NumPy)

# Carregar dados (se não vier do scraping)
df = pd.read_csv('smartphones.csv')

# Verificar dados faltantes
print(df.isnull().sum())
```

```python
# Limpeza: Remover duplicatas e outliers
df = df.drop_duplicates()
df = df[df['Preco'] < 10000]  # Filtrar preços absurdos

# Extrair marca do modelo (ex.: "iPhone 15" -> "Apple")
df['Marca'] = df['Modelo'].str.split().str[0]

# Estatísticas básicas
print(df.describe())

# Preço médio por marca
preco_medio = df.groupby('Marca')['Preco'].mean().sort_values(ascending=False)
print(preco_medio)
```

```python
# 3. Visualização com Matplotlib

# Configurar estilo
plt.style.use('seaborn')

# Gráfico 1: Distribuição de preços
plt.figure(figsize=(10, 6))
plt.hist(df['Preco'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribuição de Preços de Smartphones')
plt.xlabel('Preço (R$)')
plt.ylabel('Quantidade')
plt.grid(True)
plt.show()

# Gráfico 2: Preço médio por marca
plt.figure(figsize=(12, 6))
preco_medio.plot(kind='bar', color='orange')
plt.title('Preço Médio por Marca')
plt.xlabel('Marca')
plt.ylabel('Preço Médio (R$)')
plt.xticks(rotation=45)
plt.grid(axis='y')
plt.show()
```

```
# Gráfico 3: Relação Preço x Avaliação
plt.figure(figsize=(10, 6))
plt.scatter(df['Preco'], df['Avaliacao'], alpha=0.6, color='green')
plt.title('Relação entre Preço e Avaliação')
plt.xlabel('Preço (R$)')
plt.ylabel('Avaliação (1-5)')
plt.grid(True)
plt.show()
```

```
# Insight:
#   A - Distribuição de Preços:
#    - A maioria dos smartphones está na faixa de R$ 1.000 a R$ 3.000.
#    - Poucos produtos premium (acima de R$ 5.000).

# B  -  Marcas Dominantes:
#    - Apple (iPhone) tem o preço médio mais alto.
#    - Samsung e Xiaomi dominam a faixa intermediária.

# C  -  Relação Preço x Avaliação:
#    - Produtos mais caros nem sempre têm melhor avaliação.
#    - Boas opções de custo-benefício podem ser encontradas na faixa de R$ 1.500 a R$ 2.500.
```

```
# Recomendações para a Empresa
# Aumentar estoque de marcas com melhor custo-benefício (ex.: Xiaomi).
#  Promoções estratégicas em produtos com alta avaliação e preço médio.
# monitorar concorrentes para ajustar preços de iPhones e outros premium.
```

```
#  Próximos Passos (approfundando)
# - Aplicar ml para prever tendências de preço.
```

Processo de análise com as ferramentas Python

Concentrado no resultado da análise

Baseado em perguntas

PERGUNTAS:

- Qual é o impacto das mudanças de preço nas vendas?
- Quais são as características dos clientes que mais gastam?
- Quais são os padrões de utilização de um serviço ao longo do tempo?
- Quais produtos são frequentemente comprados juntos?

- Hipóteses no Processamento de Dados

- As hipóteses são suposições ou conjecturas iniciais que podem ser testadas e verificadas através da análise de dados. No contexto do processamento de dados e da análise de dados, as hipóteses desempenham um papel crucial ao direcionar a investigação e orientar a interpretação dos resultados. Elas ajudam a estruturar a análise e a definir claramente os objetivos da pesquisa.