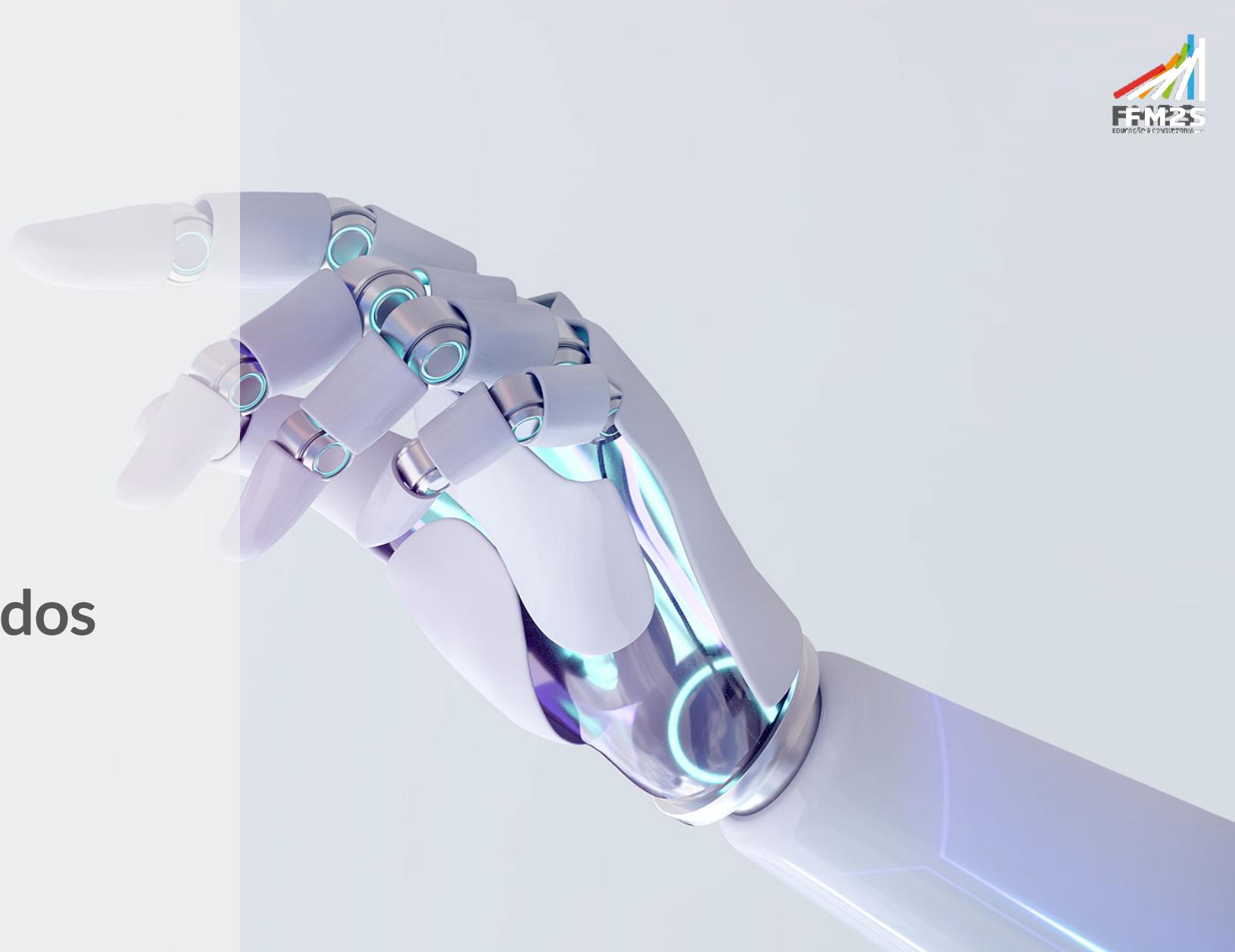


Dados
Estruturados

x

Dados
Desestruturados



Dados Estruturados x Dados Desestruturados

	Dados Estruturados	Dados Desestruturados
Definição	<ul style="list-style-type: none"> ▪ Dados organizados em um formato específico e predefinido; ▪ Campos e tipos de dados bem definidos. 	<ul style="list-style-type: none"> ▪ Dados que não possuem uma estrutura fixa; ▪ Podem incluir texto, imagens, áudio, vídeo, etc.
Formato	<ul style="list-style-type: none"> ▪ Geralmente são armazenados em tabelas; ▪ Linhas representam registros individuais e colunas representam campos de dados. 	<ul style="list-style-type: none"> ▪ Não possuem uma estrutura uniforme; ▪ Podem ser armazenados em diferentes formatos, como documentos de texto, arquivos de imagem, áudio e vídeo.

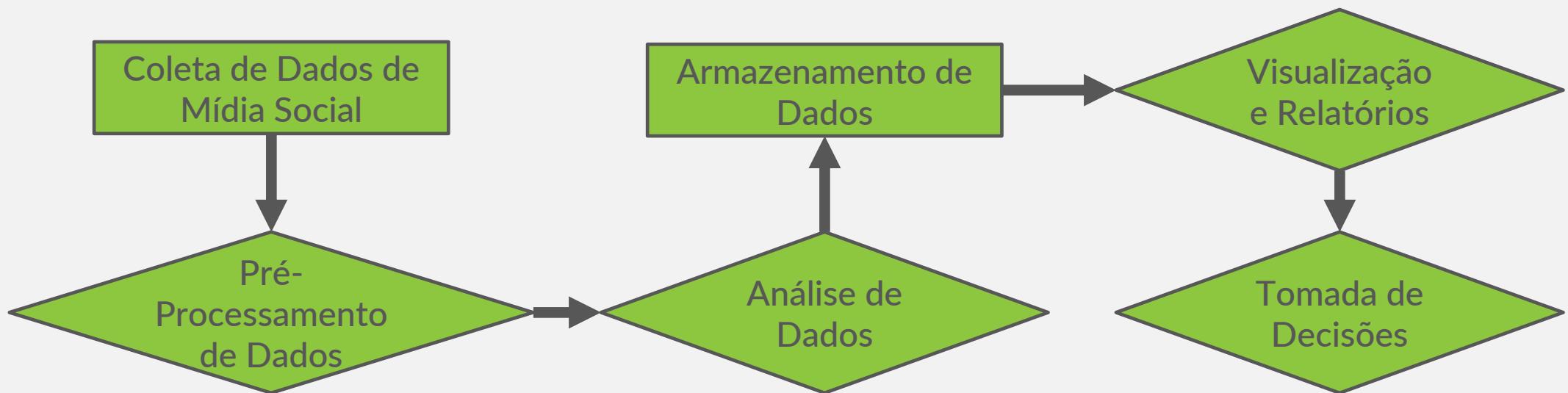
Exemplo de Dados Estruturados

Campos de Dados				
ID	Nome	Idade	Sexo	Cidade
1	João	30	M	São Paulo
2	Maria	25	F	Rio de Janeiro
3	Pedro	40	M	Belo Horizonte

- **Campos de Dados (colunas):** Representam as características específicas ou atributos de um registro de dados e são usados para armazenar e organizar informações de forma estruturada.
- **Registros Individuais (linhas):** Unidades de dados que representam uma entrada única em um conjunto de dados.

Exemplo de Dados Desestruturados

Para os dados não estruturados (ou desestruturados), uma representação visual eficaz pode ser um diagrama de fluxo ou uma nuvem de palavras. Abaixo temos um diagrama de fluxo para representar o **processo de coleta, análise e armazenamento de dados desestruturados de mídia social**.



Dados Estruturados x Dados Desestruturados

	Dados Estruturados	Dados Desestruturados
Exemplos	<ul style="list-style-type: none">■ Banco de Dados Relacionais;■ Planilhas;■ Arquivos CSV.	<ul style="list-style-type: none">■ Textos de Redes Sociais;■ <i>E-mail</i>,■ Fotos e Vídeos;■ Logs do Servidor.
Análise e Mineração	<ul style="list-style-type: none">■ Facilmente analisados e minerados usando técnicas tradicionais de análise de dados, como SQL e ferramentas de BI (<i>Business Intelligence</i>).	<ul style="list-style-type: none">■ Exige o uso de técnicas avançadas de análise de dados, como aprendizado de máquina e processamento de linguagem natural (PLN).

Dados Estruturados x Dados Desestruturados

	Dados Estruturados	Dados Desestruturados
Escalabilidade	Tendem a ser mais fáceis de escalar devido à sua estrutura organizada, especialmente em sistemas de banco de dados relacionais.	Pode ser mais desafiador escalar devido à natureza variada e complexa dos dados.
Integração	São mais fáceis de integrar com outros sistemas devido à sua estrutura padronizada.	Pode exigir esforços adicionais de integração devido à sua variedade e complexidade.

Quando usar Dados Estruturados?

Aqui estão algumas situações em que dados estruturados são utilizados:

- **Armazenamento e Gerenciamento de Dados:** A estrutura predefinida das tabelas facilita o armazenamento, recuperação e manipulação dos dados.
- **Integração de Sistemas:** Ao integrar diferentes sistemas de informação, é comum usar formatos de dados estruturados, como XML ou JSON, para facilitar a comunicação entre os sistemas.



Quando usar Dados Estruturados?

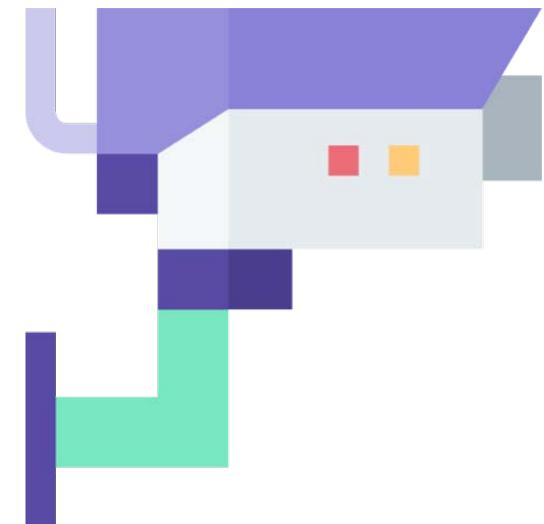
- **Visualização e Relatórios:** Em muitos casos, dados estruturados são usados para gerar relatórios e visualizações que ajudam os usuários a entender e interpretar os dados.
- **Validação de Dados:** Dados estruturados também são úteis para garantir a integridade e a consistência dos dados em um sistema.



Quando usar Dados Desestruturados?

Aqui estão algumas situações em que dados desestruturados são utilizados:

- **Segurança e Vigilância:** Câmeras de vigilância, sistemas de segurança e sensores ambientais são fundamentais para monitorar atividades suspeitas, detectar padrões de comportamento e prevenir incidentes de segurança.
- **Análise de Mídias Digitais:** Áudio, vídeo e texto podem ser analisados para extrair *insights* sobre tendências de consumo, preferências do público e padrões de comportamento.



Quando usar Dados Desestruturados?

- **Análise de Sentimento e Opinião:** Mídias sociais, como *tweets*, postagens em fóruns ou comentários em *blogs*, podem ser utilizados para analisar o sentimento do público em relação a produtos, serviços ou eventos
- **Pesquisa em Ciências Biológicas e da Saúde:** Sequências genéticas, imagens médicas e registros de pacientes, são exemplos de dados desestruturados utilizados para entender doenças, desenvolver tratamentos personalizados e melhorar a eficácia dos cuidados de saúde.



Próximo Tópico...

Ao explorar a diferença entre dados estruturados e dados desestruturados, é fundamental compreender como a coleta e estruturação de dados desempenham um papel essencial na organização e utilização eficaz das informações.

No próximo tópico, examinaremos os métodos e ferramentas utilizados para coletar e estruturar dados, e como a abordagem correta pode maximizar o valor dos dados para análise e tomada de decisões.



Coleta e Estruturação de Dados

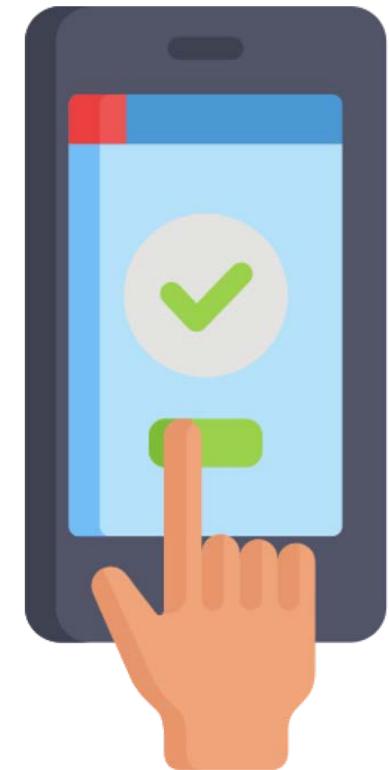


O que é Coleta e Estruturação de Dados?

São processos utilizados para, respectivamente, captar informações geradas pelos fenômenos (ou por processos) e organizá-las para a análise que servirá de insumo para planejar estratégias para o negócio.

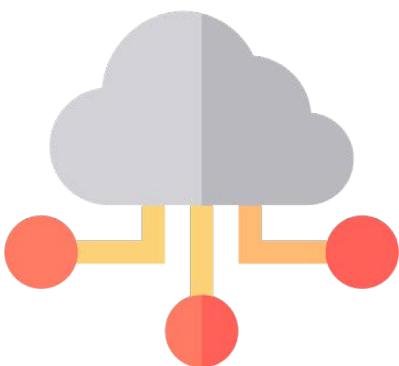
Duas estratégias:

- Integrar dados existentes na organização ou;
- Coletar dados “do zero”.



Onde ocorre a Estruturação de Dados?

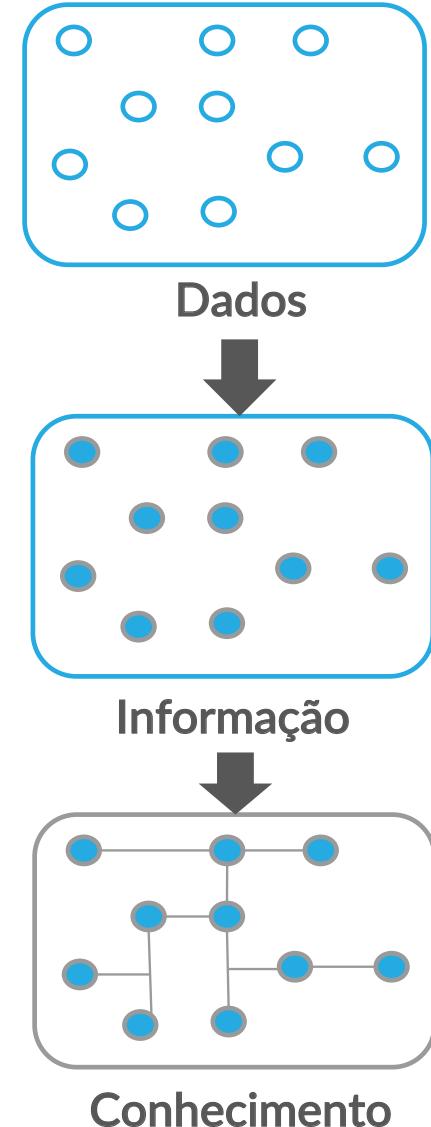
Os dados coletados podem ser estruturados em diversos ambientes e sistemas, como:



- Bancos de Dados Relacionais (*MySQL, PostgreSQL, SQL Server e Oracle*);
- Armazenamento em nuvem (*Amazon Web Services (AWS), Google Cloud Platform (GCP) e Microsoft Azure*);
- Planilhas e arquivos (Microsoft Excel ou Google *Sheets*, JSON (*JavaScript Object Notation*) ou XML (*Extensible Markup Language*));
- *Data Warehouses* (são sistemas de armazenamento projetados para armazenar grandes volumes de dados de diversas fontes);
- *Data Lakes* (são repositórios de dados que armazenam dados brutos em seu formato original, sem necessidade de estruturação prévia).

Alguns pontos importantes

- Toda coleta depende de um contexto;
- Toda coleta requer uma validação (especialmente se feita de dados pré-existentes);
- O ato da coleta já é uma análise do problema.

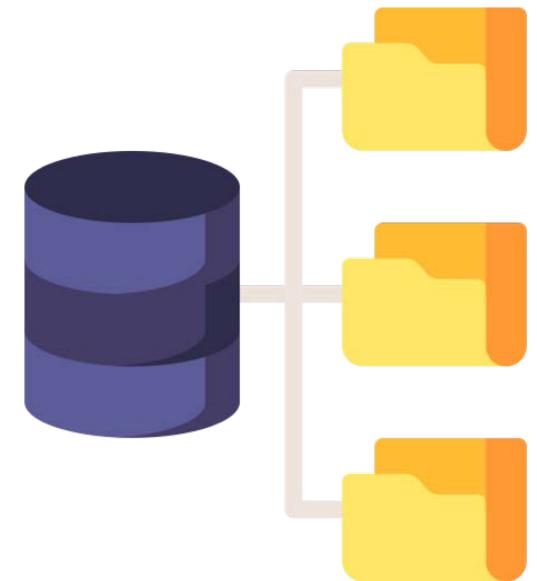


Próximo Tópico...

Com isso, a coleta e estruturação de dados são estágios imprescindíveis no processo de análise de dados, e compreender como realizar essas etapas de maneira eficiente é essencial.

Neste tópico, exploramos a importância de coletar dados relevantes e confiáveis, bem como de organizá-los de maneira sistemática e coerente.

Agora, ao nos aprofundarmos no próximo tópico, vamos abordar as práticas e técnicas específicas para realizar a coleta e estruturação de dados de forma eficaz.

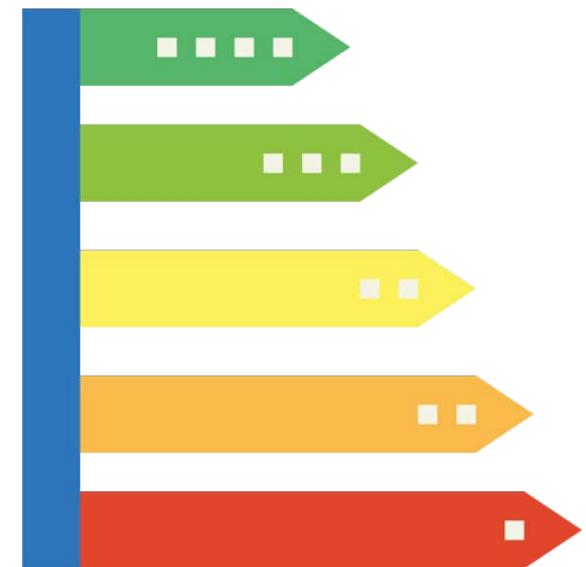




Como fazer a Coleta e Estruturação de Dados?

O Passo a Passo da Coleta e Estruturação de Dados

1. Entender o problema e quais dados são necessários (modelagem);
2. Planejar a forma de coleta;
3. Coletar dados;
4. Fazer limpeza e pré-processamento de dados;
5. Integrar e Transformar dados;
6. Realizar uma Análise Exploratória de Dados;
7. Ajustar, caso necessário.



Como fazer a Coleta e Estruturação de Dados?

Com vimos, a **Estatística** é uma ferramenta importante para organizar os dados, resumi-los, analisá-los e utilizá-los para tomada de decisões.

- **Análise Exploratória de Dados** se ocupa da organização e resumo dos dados de uma amostra ou, eventualmente, de toda a população.
- **Inferência Estatística** se refere ao processo de se tirar conclusões sobre uma população com base em uma amostra dela.

A abordagem estatística para o tratamento de dados envolve:

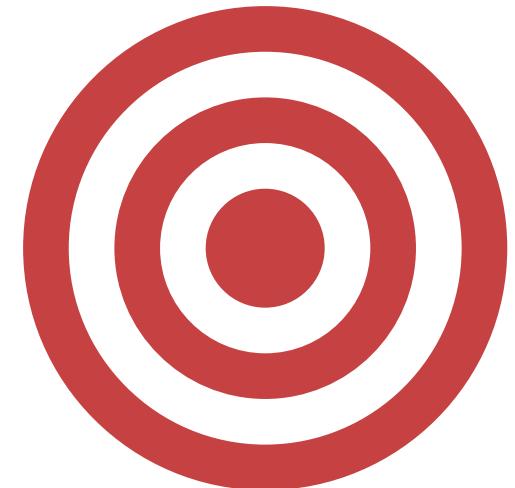


Entendendo o problema

1.1 Definir o problema e os objetivos

Esse é o primeiro passo nas etapas de coleta e estruturação de dados. Aqui é o momento de identificar os problemas de negócios que precisam ser resolvidos.

- **Exemplo:** Uma empresa deseja prever a rotatividade de funcionários para melhorar a retenção de talentos.



Entendendo o problema

1.2 Identificar os dados

Nesta etapa, você deve identificar as fontes de dados relevantes e estabelecer os métodos e procedimentos apropriados para a coleta e organização dos dados.

- **Exemplo:** Fontes de dados podem incluir bancos de dados de recursos humanos, registros de desempenho, pesquisas de satisfação do funcionário, etc.



Planejando a Coleta

Onde iremos buscar os dados?

Nesta etapa, vale a pena entender onde estão os dados que precisamos na organização.

Caso não tenhamos esses dados, precisamos entender como vamos coletá-los.

Isso passa por:

- Definir formulários;
- Definir variáveis;
- Treinar os coletores;
- Etc.



Planejando a Coleta

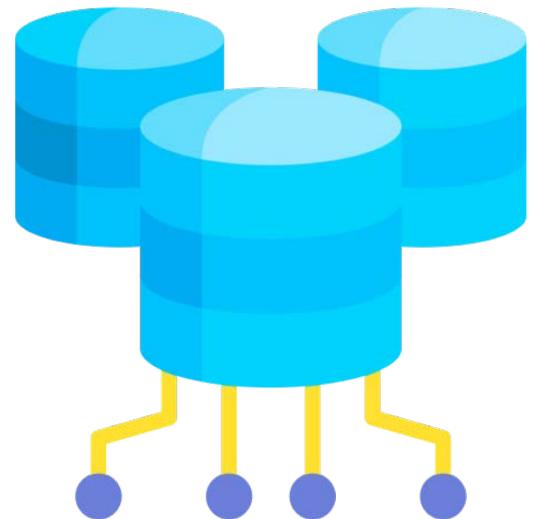
A seguir, estão algumas das técnicas comuns de coleta de dados quantitativos e qualitativos. Dependendo do contexto e dos objetivos do estudo, outras técnicas também podem ser aplicadas para obter dados relevantes e confiáveis.

Coleta de Dados Quantitativos	Coleta de Dados Qualitativos
Entrevistas estruturadas	Entrevistas não estruturadas
Questionários fechados	Observação direta
Experimentação	<i>Focus Groups (Grupos de Foco)</i>
Análise de registros e documentos	Análise de documentos
Utilização de sensores e dispositivos eletrônicos	Análise de redes sociais

Coletando os Dados

O objetivo aqui é extrair dados das fontes identificadas. Para isso, existem diversas técnicas, cada uma adequada a diferentes contextos e tipos de informações que se deseja obter.

- **Exemplo:** Consultar bancos de dados de RH para obter informações sobre histórico de emprego, salários, avaliações de desempenho, etc.



Limpeza e Pré-Processamento de Dados

Aqui ocorre identificação e correção de possíveis erros de coleta e/ou digitação.

Após a coleta, os dados podem passar por um processo de limpeza para identificar e corrigir erros, inconsistências e valores ausentes. Isso envolve a remoção de dados duplicados, a padronização de formatos e a correção de erros de digitação.

- **Exemplo:** Preencher valores ausentes com a média dos dados existentes, remover registros duplicados, normalizar dados numéricos, etc.



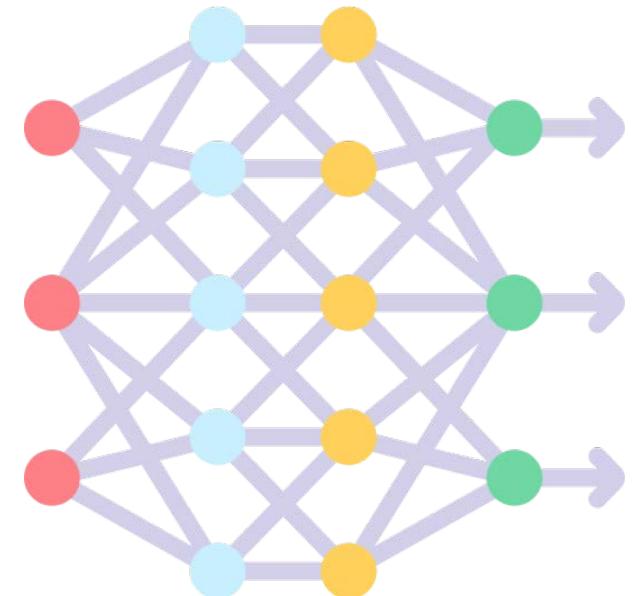
Integração de dados

Aqui deve-se integrar dados de diferentes fontes para criar um conjunto de dados unificado.

- **Exemplo:** Mesclar dados de avaliações de desempenho com dados de salários usando um identificador único, como o ID do funcionário.

Após a integração, deve-se realizar a transformação dos dados conforme necessário para análise e modelagem.

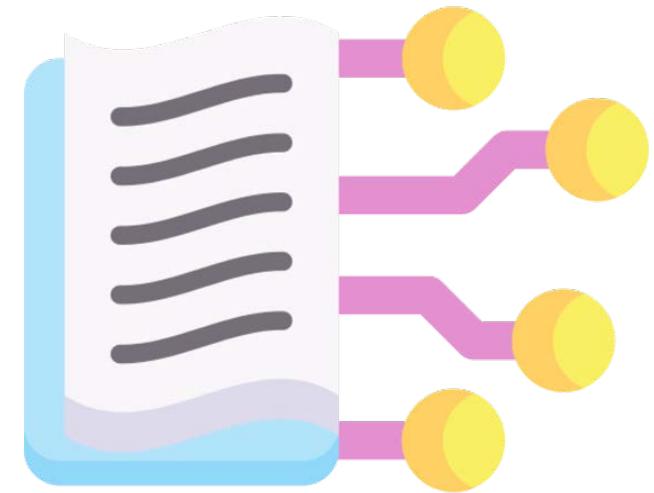
- **Exemplo:** Converter variáveis categóricas em variáveis *dummy* para serem usadas em modelos de aprendizado de máquina.



Análise Exploratória dos Dados (EDA)

Essa etapa consiste em explorar os dados para entender suas características e relacionamentos. É crucial no processo de ciência de dados, onde os analistas buscam entender a estrutura, padrões e características dos dados antes de aplicar técnicas de modelagem ou inferência estatística.

- **Exemplo:** Visualizar a distribuição de salários dos funcionários, identificar correlações entre variáveis, etc.



Ajustes na coleta de dados

Após todos os procedimentos, é necessário ver se os dados finais que obtivemos são interessantes ou não para resolvermos os nossos problemas.

Caso não sejam, quais outros dados precisamos?
Como poderemos obtê-los?

Geralmente esse processo pode tomar um tempo significativo em projetos de Ciência de Dados.

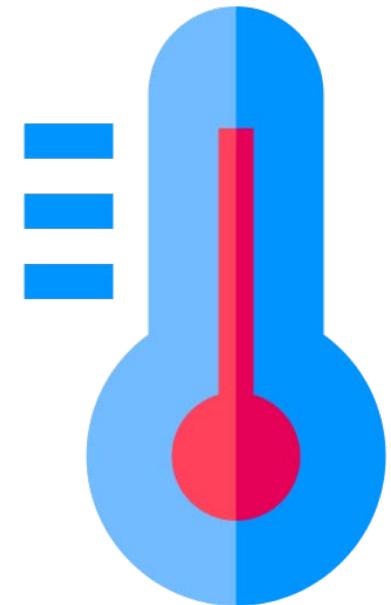


Exemplo de Coleta e Estruturação de Dados

Exemplo – Monitoramento de Temperatura

Um estudante de Ciências Atmosférica está desenvolvendo sua pesquisa de TCC (Trabalho de Conclusão de Curso) com tema voltado para as diferenças climáticas presentes no Brasil. Por ser um país muito grande e diverso, não só a cultura se difere de Norte a Sul, mas também a temperatura.

Para isso, o estudante precisou coletar dados de temperatura de cidades brasileiras para fazer uma comparação e utilizar como informação na discussão do seu trabalho. Ele precisou seguir os seguintes passos:



Exemplo – Monitoramento de Temperatura

1. Planejamento: O estudante inicia sua pesquisa fazendo um planejamento para definir os objetivos da coleta e estruturação de dados, identificar as fontes de dados relevantes e estabelecer os métodos e procedimentos apropriados para sua coleta e organização.



Objetos da coleta:

- **Unidade de Investigação:** Capitais dos Estados Brasileiros;
- **Variáveis:** Temperatura atual, Temperatura mínima, Temperatura máxima, Umidade relativa do ar e Velocidade do vento.

Método de coleta:

- O estudante irá utilizar uma **API de dados meteorológicos** para obter as temperaturas atuais nas capitais dos estados brasileiros. Vamos supor que esta API forneça os dados em formato JSON.

Exemplo – Monitoramento de Temperatura

2. Coleta de dados: De acordo com o planejamento, o estudante notou que a forma mais prática de verificar as temperaturas seria utilizando uma API de dados meteorológicos. Isso é possível através de sensores de temperatura instalados em todas as capitais para coletar dados simultaneamente e analisar as variações de temperatura em tempo real. Assim, ele poderia obter informações em tempo real sobre as condições meteorológicas atuais, como temperatura, umidade e velocidade do vento.

Uma API (Interface de Programação de Aplicações) de dados meteorológicos é um serviço que fornece acesso programático a informações meteorológicas.



Exemplo – Monitoramento de Temperatura

3) Resumo/Estruturação dos dados: Após a coleta dos dados, eles foram organizados/estruturados em uma tabela, onde cada linha representará uma capital e as colunas representarão os dados relacionados à temperatura, como temperatura atual, temperatura mínima e máxima do dia, umidade relativa do ar, e velocidade do vento.

Capital	Temperatura Atual (°C)	Temperatura Mínima (°C)	Temperatura Máxima (°C)	Umidade Relativa do Ar (%)	Velocidade do Vento (km/h)
Belém	31	24	32	68	13
Natal	31	26	24	66	23
São Paulo	30	18	31	31	14
Goiânia	32	21	32	38	5
Curitiba	27	18	27	58	10

Exemplo – Monitoramento de Temperatura

Importante!!

Depois de coletar os dados, seria possível realizar uma análise exploratória para entender as variações de temperatura entre as capitais. Isso poderia envolver a criação de gráficos de linha ou gráficos de dispersão para visualizar as tendências ao longo do tempo, além de calcular estatísticas descritivas, como média, mediana e desvio padrão, para entender a distribuição das temperaturas.

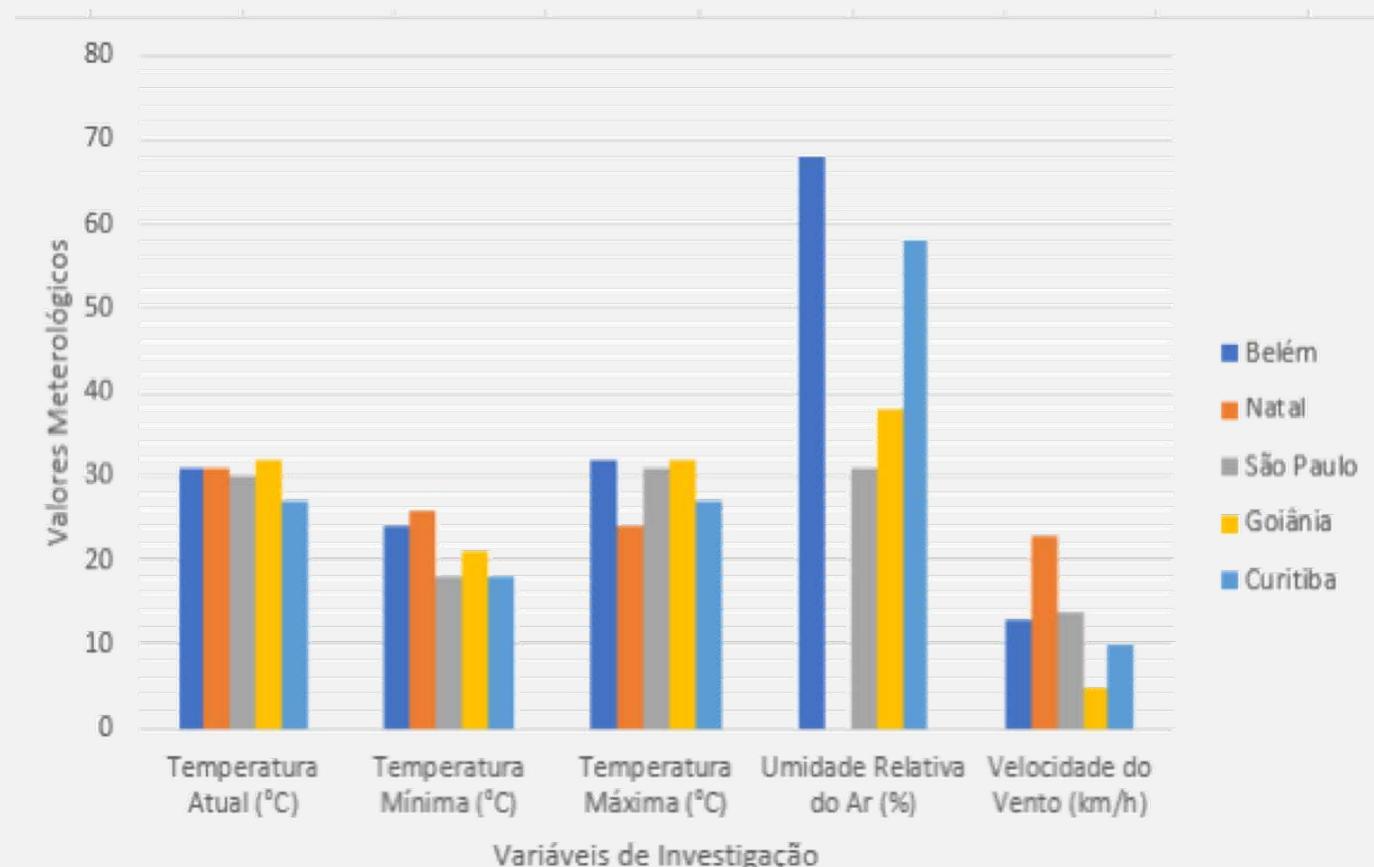
Veremos sobre esse esses tipos de análise nas próximas aulas.



Exemplo – Monitoramento de Temperatura

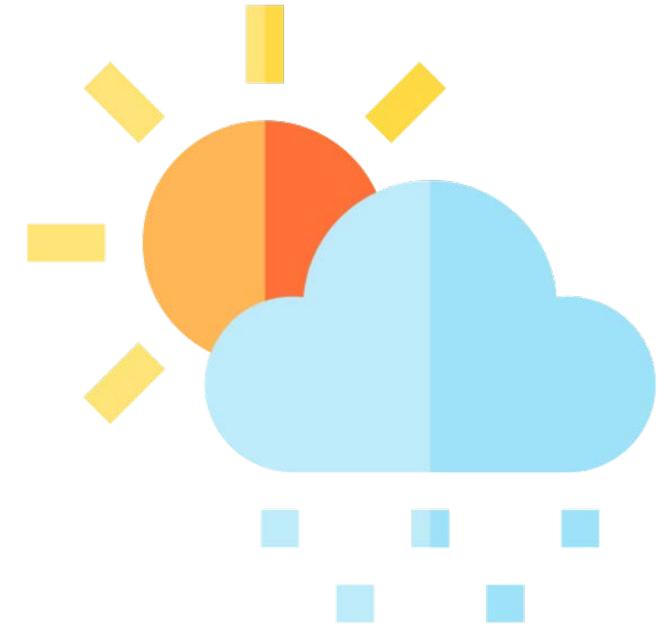
4) Análise e Visualização:

Após a estruturação, o estudante pôde então analisar e visualizar os dados em gráficos e tabelas para entender as variações de temperatura entre as capitais dos estados brasileiros.



Exemplo – Monitoramento de Temperatura

5. Interpretação dos Resultados: Com base na análise dos dados, o estudante identificou padrões, tendências e variações nas temperaturas das capitais dos estados brasileiros. Por exemplo, ele observou que as capitais na região Norte tendem a ter temperaturas mais altas em comparação com as capitais no Sul do Brasil.



Exemplo – Monitoramento de Temperatura

6. Aplicação dos Resultados: O estudante pode, então, utilizar as informações coletadas e estruturadas como parte da discussão em seu TCC. Além disso, essa metodologia também pode ser interessante para previsão do tempo, planejamento de viagens, análise de padrões climáticos e tomada de decisões em diferentes setores, como agricultura, turismo e infraestrutura.



- Além disso, também há a possibilidade de detectar padrões sazonais ou anomalias nas temperaturas das capitais, o que poderia fornecer *insights* sobre os padrões climáticos regionais.
- Essas análises poderiam ser realizadas continuamente em tempo real, permitindo uma resposta rápida a mudanças climáticas significativas ou eventos extremos.

Análise de Dados



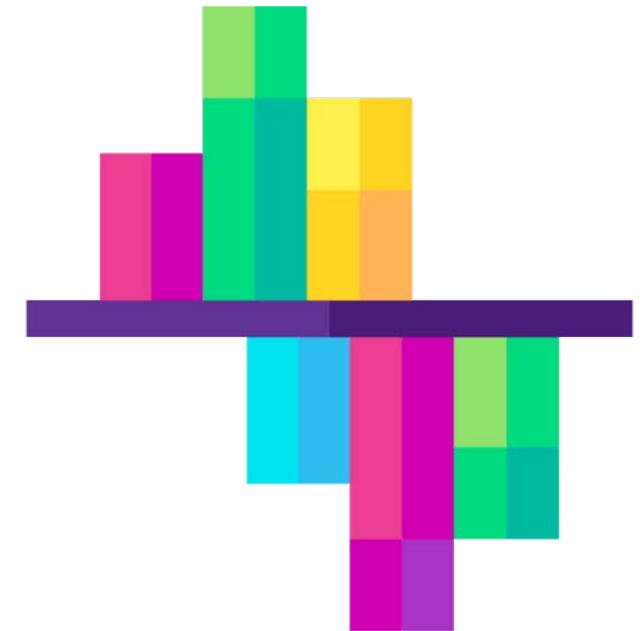
O que significa Analisar Dados?

Agora que já vimos a coleta de dados, podemos iniciar as técnicas para a análise de dados estruturados.

Basicamente, analisar dados estruturados implica na utilização de modelos estatísticos que vão nos gerar *insights* sobre os dados apresentados.

As análises mais frequentes consistem em:

- Observar parâmetros da distribuição de dados;
- Visualizar com gráficos esses parâmetros;
- Usar modelos capazes de quantificar as incertezas relativas à variação nos dados.

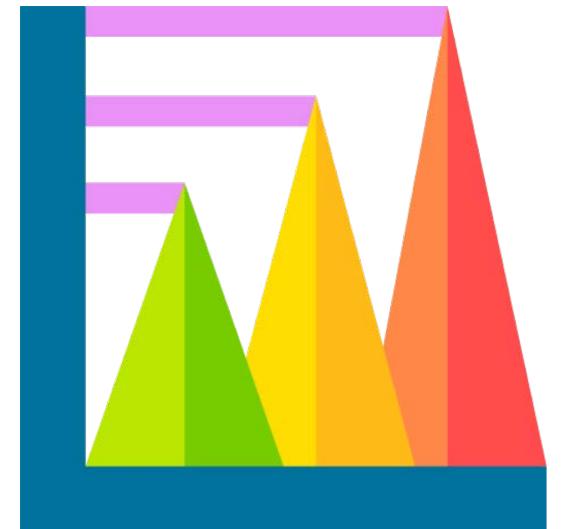


As possibilidades da Análise de Dados

Entretanto, quais ferramentas estatísticas usar depende, basicamente:

- Do que se quer da análise (avaliar frequência, tendência, correlação, etc.);
- Do tipo de variável que você quer analisar.

A seguir, iremos entender melhor sobre os tipos de variáveis.



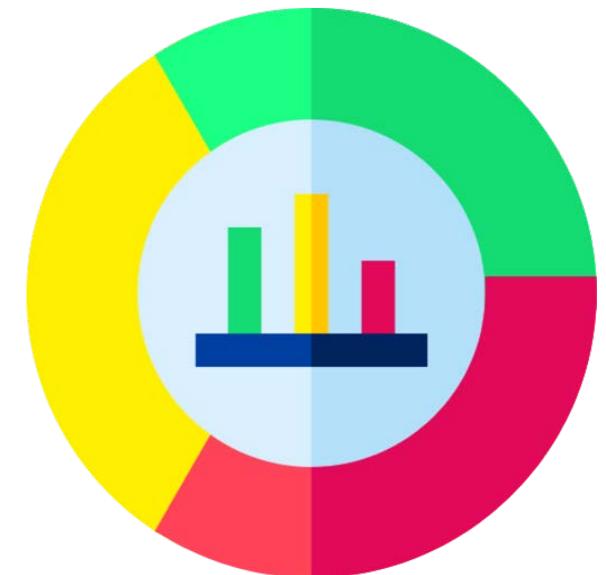
Tipos de Variáveis

O que são Variáveis?

Variável, na Estatística, é uma **característica de interesse mensurável que é medida em cada elemento da amostra ou população.**

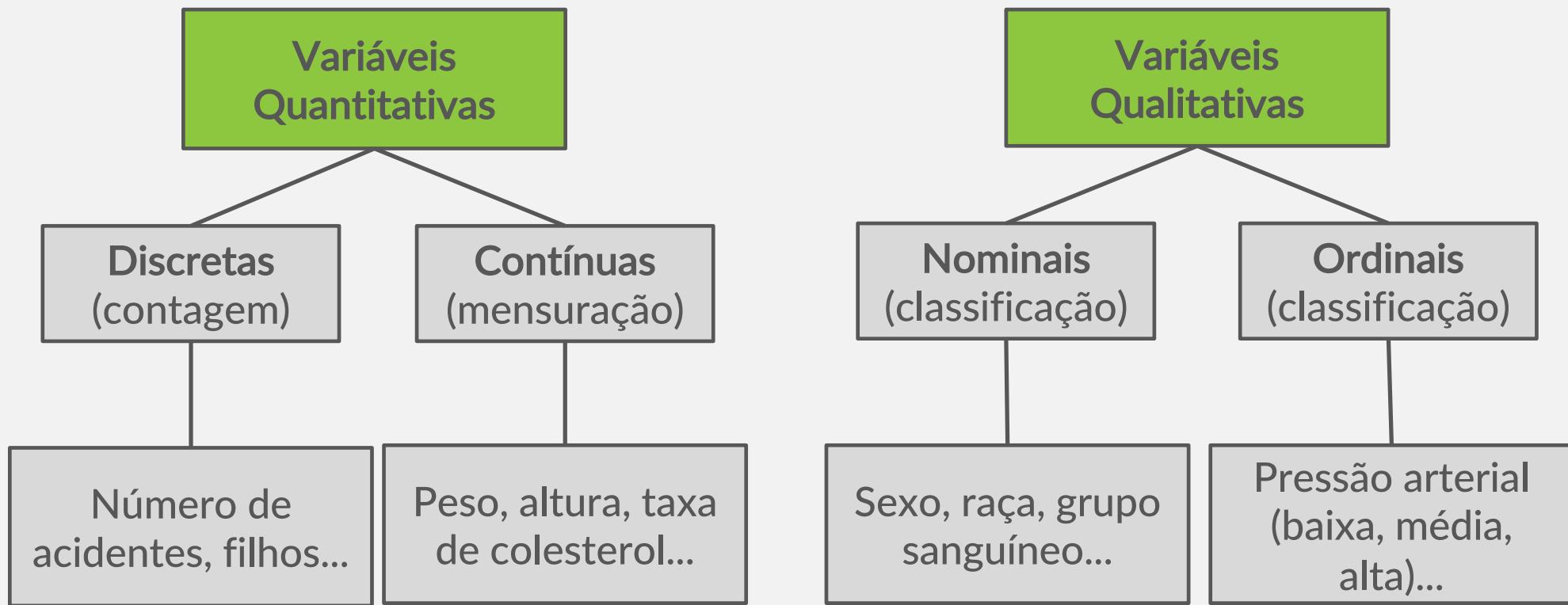
- Seus valores variam de elemento para elemento, por isso o termo “variável”;
- As variáveis podem ter valores numéricos ou não numéricos.

Dessa forma, variáveis podem ser entendidas como atributos de objetos que podem variar para diferentes casos.



Tipos de Variáveis

As Variáveis são classificadas em dois tipos:



Tipos de Variáveis

Variáveis Quantitativas (ou Numéricas)

Definição: São aquelas que exibem valores numéricos associados à unidade de investigação.

- Caracterizam-se por ser numericamente mensuráveis;
- São comumente usadas em pesquisas científicas, análises estatísticas, etc.;
- Utilizada para descrever e analisar fenômenos quantitativos;
- Há dois tipos de Variáveis Quantitativas.
 - Contínua;
 - Discreta.



Variáveis Quantitativas (ou Numéricas)

Variáveis Quantitativas		
	Contínuas (mensuração)	Discretas (contagem)
Características	Atributos podem ser variáveis contínuas quando assumem valores pertencentes a um intervalo de números reais.	Quando possíveis valores formam um conjunto finito ou enumerável de números e que resultam frequentemente de uma contagem.
Exemplos	<ul style="list-style-type: none"> ▪ Altura de uma pessoa; ▪ Peso de um objeto; ▪ Temperatura do corpo humano; ▪ Tempo de duração de uma corrida; ▪ Renda mensal de uma família. 	<ul style="list-style-type: none"> ▪ Número de filhos em uma família; ▪ Número de carros em um estacionamento; ▪ Pontuação em um teste de múltipla escolha; ▪ Quantidade de produto em estoque; ▪ Números de lados em um dado.

Tipos de Variáveis

Variáveis Qualitativas (ou Categóricas)

Definição: São aquelas que indicam um atributo não numérico da unidade de investigação.

- São as características que não possuem valores quantitativos;
- São categorizados em grupos ou classes distintas;
- Descrevem qualidades ou propriedades de um objeto ou fenômeno;
- São classificadas em:
 - Nominais;
 - Ordinais.



Variáveis Qualitativas (ou Categóricas)

Variáveis Qualitativas		
	Nominais (classificação)	Ordinais (classificação)
Características	Não existe ordenação dentre as categorias.	Existe uma ordenação entre as categorias.
Exemplos	<ul style="list-style-type: none">■ Cor dos olhos;■ Gênero;■ Estado civil;■ Tipo de animal de estimação;■ Marca de carro	<ul style="list-style-type: none">■ Nível de satisfação do cliente;■ Grau de escolaridade;■ Classificação da dor;■ Nível de concordância;■ Classificação socioeconômica.

Algumas distinções importantes!!

Ao lidar com diferentes tipos de variáveis em um conjunto de dados, é importante reconhecer que **cada tipo de variável pode demandar abordagens distintas em termos de modelagem estatística e visualização.**

Por exemplo, variáveis numéricas contínuas podem requerer modelos de regressão para explorar relações lineares ou não lineares, enquanto variáveis categóricas podem ser mais adequadas para análises de frequência ou modelos de classificação.

Neste tópico iremos verificar algumas distinções importantes que podem facilitar a análise de dados envolvendo variáveis.



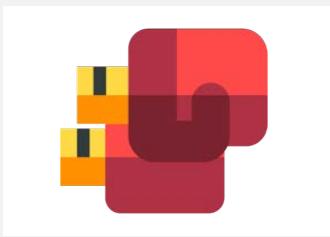
Algumas distinções importantes!!

[A] Uma variável que é originalmente quantitativa pode ser coletada de forma qualitativa.

Exemplos:



A idade, que normalmente é medida em anos completos, é uma variável quantitativa contínua. No entanto, se forem fornecidas apenas faixas etárias (por exemplo, 0 a 5 anos, 6 a 10 anos, etc.), ela se torna uma variável qualitativa ordinal.



Outro exemplo é o peso dos lutadores de boxe, que é uma variável quantitativa contínua quando consideramos o valor registrado na balança, mas qualitativa ordinal se classificarmos os lutadores nas categorias de peso do boxe (peso-pena, peso-leve, peso-pesado, etc.).

Algumas distinções importantes!!

[B] Outro ponto importante é que nem sempre uma variável representada por números é quantitativa.

Exemplo:

- Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa!



Algumas Distinções Importantes!!

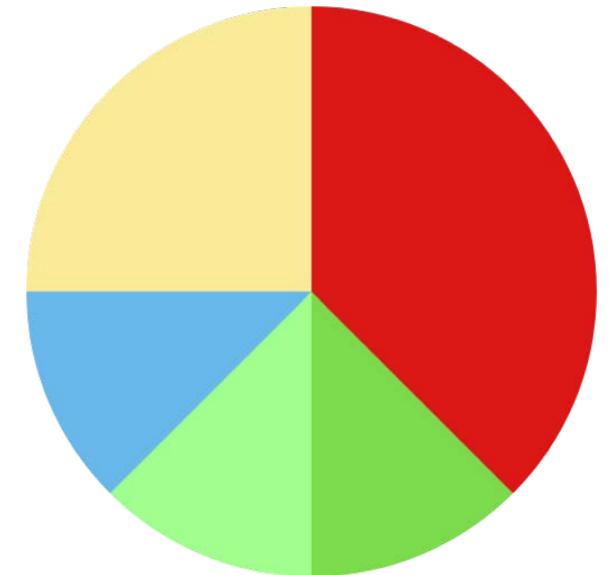
	Exemplos	Análise Quantitativa	Análise Qualitativa
Interpretação de Dados	Renda	Pode ser considerada uma variável quantitativa quando analisada como um número.	Mas pode ser interpretada qualitativamente ao considerar categorias de renda (baixa, média, alta).
Dados Mistos	Código Postal	Pode ser visto como uma variável quantitativa, se usado para cálculos como distância entre códigos postais.	Ou como uma variável qualitativa representando uma área geográfica.
Escalas de Mensuração	Temperatura	A temperatura medida em graus Celsius pode ser considerada quantitativa.	Enquanto a temperatura categorizada como "quente", "morno" ou "frio" pode ser considerada qualitativa.

Algumas Distinções Importantes!!

Além disso, os gráficos utilizados para visualizar essas variáveis podem variar de acordo com sua natureza.

Por exemplo:

- Histogramas são frequentemente utilizados para representar distribuições de variáveis numéricas;
- Gráficos de barras ou gráficos de pizza são mais comuns para variáveis categóricas.



Algumas Distinções Importantes!!

Tipos de Variáveis			
	Quantitativas	Qualitativas	
Comportamento ao longo do tempo	Gráfico de tendência/ controle.	Gráfico de tendência/ controle.	
Distribuição	<ul style="list-style-type: none">▪ Dot plot;▪ Histograma.	<ul style="list-style-type: none">▪ Gráfico de Barras;▪ Gráfico de Setores;▪ Gráfico de Pareto.	
Estatísticas Descritivas	<p>Localização Média, Mediana, Quartis, Mínimo, Máximo</p>	<p>Variação Desvio padrão, amplitude</p>	Tabela de frequência; Porcentagem.

Quando usar Variáveis?

Na modelagem estatística, as variáveis são usadas para construir modelos matemáticos que representam os fenômenos em estudo. Esses modelos podem ser usados para prever resultados futuros ou entender melhor o processo subjacente.

Alguns exemplos:

- **Análise Descritiva e Inferencial:** Na análise descritiva são utilizadas para resumir e descrever os dados e na análise inferencial são usadas para fazer inferências sobre uma população com base em uma amostra.
- **Tomada de Decisão Informada:** Tomar decisões mais informadas e embasadas em evidências, o que pode levar a melhores resultados e ações mais eficazes.



Dados de Data e Geoposicionamento

Os dados de data e geoposicionamento combinam informações temporais (data e hora) com informações espaciais (latitude e longitude) para fornecer um registro detalhado de eventos ou observações que ocorrem em locais específicos em momentos específicos.

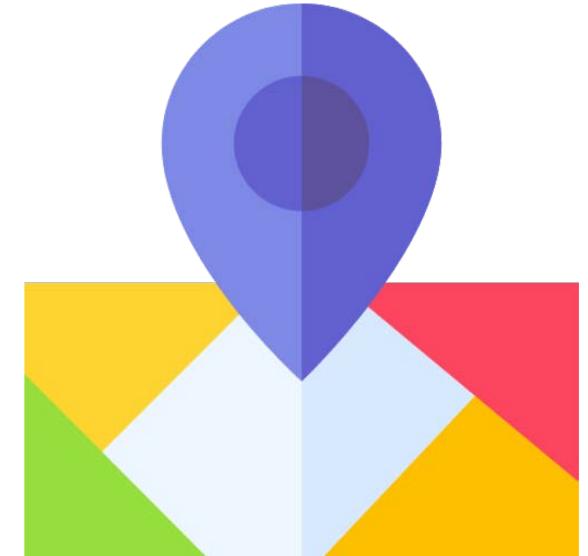
Em outras palavras, esses dados incluem informações sobre quando algo aconteceu e onde aconteceu.



Dados de Data e Geoposicionamento

A seguir estão alguns exemplos práticos de como os dados de data e posicionamento são utilizados:

- **Aplicativos de Navegação e Mapas:** Aplicativos como o Google Maps usam dados de posicionamento para fornecer rotas e direções para os usuários com base em sua localização atual e destino desejado. Eles também podem mostrar pontos de interesse próximos, como restaurantes, postos de gasolina e hotéis.



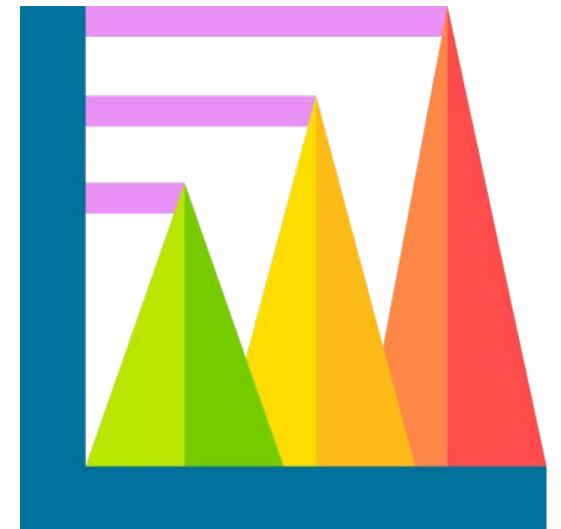
Estatísticas Descritivas

O que são as Estatísticas Descritivas?

Na estatística descritiva, a descrição de dados refere-se ao processo de resumir e comunicar as características principais de um conjunto de dados.

Além disso, organiza e descreve os dados de três maneiras:

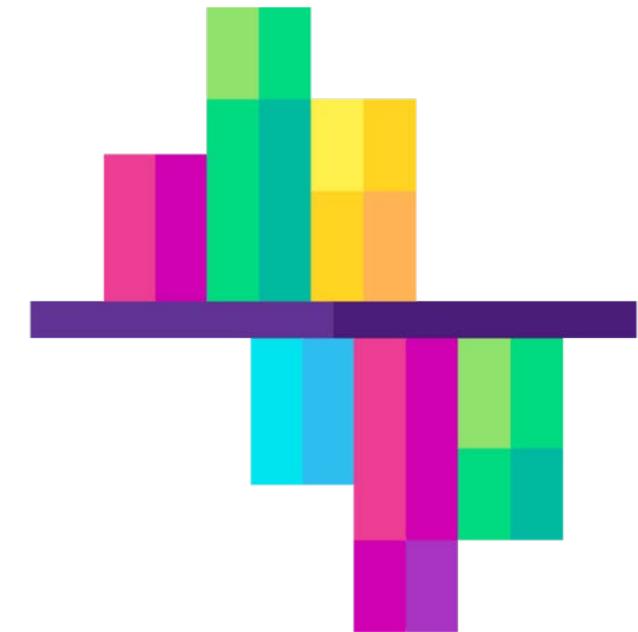
- Por meio de tabelas;
- Gráficos;
- Medidas descritivas.



O que é Estatística Descritiva?

As técnicas de descrição de dados incluem calcular estatísticas resumo, como média, mediana, moda, variância e desvio padrão, que fornecem informações sobre a tendência central, a dispersão e a forma da distribuição dos dados.

Além disso, gráficos como histogramas, gráficos de barras, *box plots* e diagramas de dispersão são comumente utilizados para visualizar a distribuição dos dados e identificar padrões ou tendências.



Estatística Descritiva e Ciência de Dados

Por que é importante?

Exploração Inicial dos Dados: Na ciência de dados, a primeira etapa geralmente envolve a exploração inicial dos dados para entender sua estrutura e características básicas. A estatística descritiva fornece as ferramentas necessárias para resumir e visualizar os dados, identificando padrões, tendências e possíveis problemas nos dados.



Estatística Descritiva e Ciência de Dados

Por que é importante?

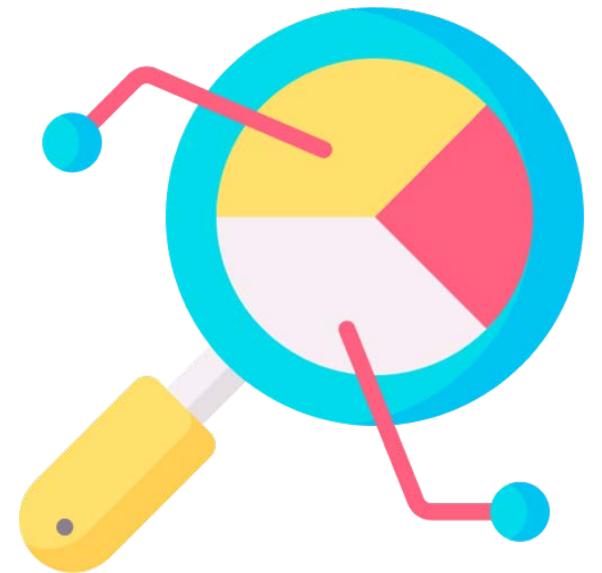
Pré-processamento de Dados: Antes de aplicar técnicas mais avançadas de modelagem e análise, os dados geralmente precisam ser pré-processados para garantir sua qualidade e consistência. A estatística descritiva é usada para identificar e lidar com valores ausentes, outliers e outras irregularidades nos dados.



Estatística Descritiva e Ciência de Dados

Por que é importante?

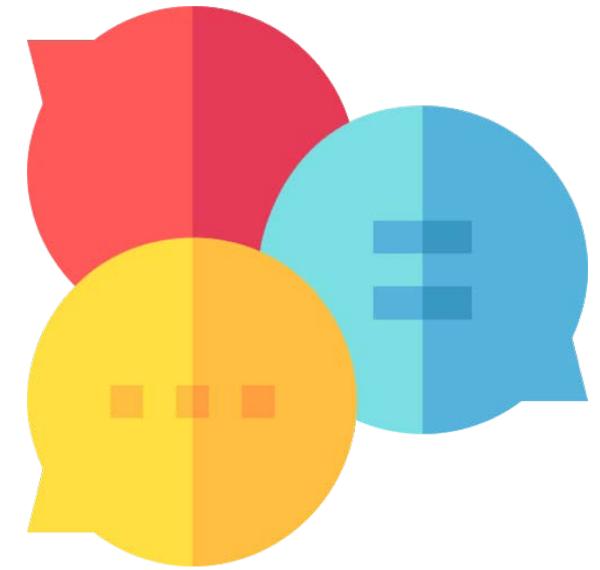
Seleção de Variáveis: A escolha das variáveis relevantes é uma etapa crítica na modelagem de dados. A estatística descritiva ajuda os cientistas de dados a entender a distribuição e a relação entre as variáveis, facilitando a seleção das variáveis mais importantes para o modelo.



Estatística Descritiva e Ciência de Dados

Por que é importante?

Comunicação dos Resultados: Uma parte importante do trabalho de um cientista de dados é comunicar os resultados de forma clara e compreensível para partes interessadas não técnicas. A estatística descritiva fornece as ferramentas necessárias para resumir e visualizar os principais insights dos dados de uma maneira acessível.

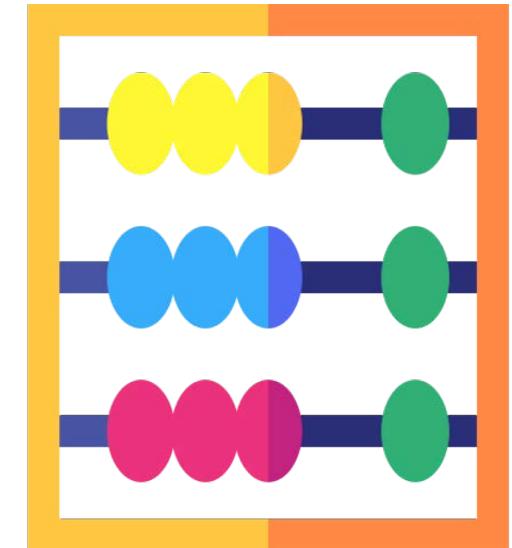


Medidas Descritivas

As medidas de posição ou tendência central, como o próprio nome indica, são medidas que informam sobre a posição típica dos dados. São elas:

- Média Aritmética;
- Desvio Padrão;
- Moda, mediana, quartis e percentis.

A seguir, veremos a explicação dessas medidas de forma mais aprofundada.

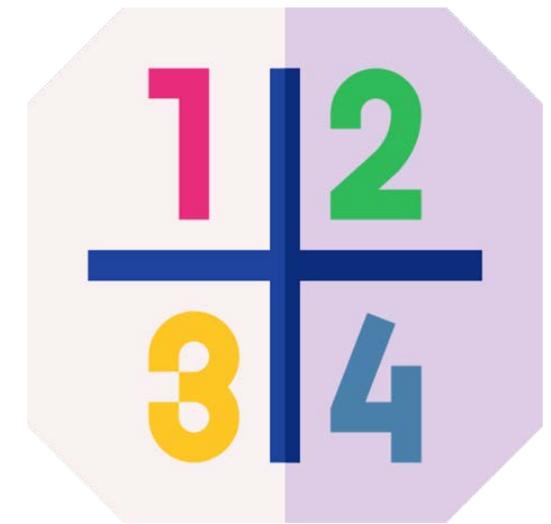


Medidas Descritivas

Média Aritmética

A média aritmética é um conceito fundamental na estatística descritiva, amplamente utilizado para resumir e entender conjuntos de dados.

Elá representa o valor médio de um conjunto de números e é calculada somando todos os valores e dividindo pela quantidade de observações.



Medidas Descritivas

Média Aritmética

Definição: Dado um conjunto de n observações x_1, x_2, \dots, x_n , a média aritmética simples é definida como:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

A notação \bar{x} (lê-se x barra), usada para indicar a média, é bastante comum; em geral, usa-se a mesma letra adotada para indicar os dados com a barra em cima.

Observe, inicialmente, que ela só pode ser calculada para dados quantitativos.

Medidas Descritivas

Média Aritmética

Exemplo: Idade de 10 funcionários do departamento de Marketing de uma empresa.

Vamos supor que temos os seguintes dados de idade dos 10 funcionários do departamento de marketing:

Funcionários	1	2	3	4	5	6	7	8	9	10
Idade	22	25	25	31	33	35	40	42	48	50

Para calcular a média aritmética, primeiro somamos todas as idades e depois dividimos pelo número total de funcionários.

Medidas Descritivas

Funcionários	1	2	3	4	5	6	7	8	9	10
Idade	22	25	25	31	33	35	40	42	48	50

- **Soma das idades:**

$$\text{Soma das Idades} = 22 + 25 + 25 + 31 + 33 + 35 + 40 + 42 + 48 + 50$$

$$\text{Soma das Idades} = 351$$

- **Número total de funcionários: 10**
- Agora, podemos calcular a média aritmética:

$$\text{Média} = \frac{\text{Soma das idades}}{\text{Número total de funcionários}} = \frac{351}{10} = 35,1 \text{ anos}$$

Medidas Descritivas

Moda

Definição: A moda, que representaremos por x^* , é o valor que ocorre com mais frequência em um conjunto de dados. Uma distribuição de dados pode ter uma moda (unimodal), duas modas (bimodal) ou mais (multimodal), ou pode não ter moda se nenhum valor se repetir.

Exemplo: Idade de 10 funcionários do departamento de Marketing de uma empresa.

Funcionários	1	2	3	4	5	6	7	8	9	10
Idade	22	25	25	31	33	35	40	42	48	50

Nesse caso, a **Moda é Unimodal**, onde $x^* = 25 \text{ anos}$, pois apenas uma idade ocorre com maior frequência, essa idade ocorre duas vezes mais que qualquer outra no departamento de Marketing.

Medidas Descritivas

Mediana

A mediana é o valor que divide o conjunto de dados em duas partes iguais quando eles estão organizados em **ordem crescente ou decrescente**. Em outras palavras, é o valor do meio quando os dados estão organizados.

Se houver um número par de observações, a mediana é a média dos dois valores do meio.



Medidas Descritivas

Mediana

Definição: Seja x_1, x_2, \dots, x_n um conjunto de n observações, e seja $x(i)$, $i = 1, \dots, n$ o conjunto das observações ordenadas, de modo que $x(1) \leq x(2) \leq \dots \leq x(n)$. Então, a mediana é definida como o valor tal que 50% das observações são menores e 50% são maiores que ela. Para efeito de cálculo, valem as seguintes regras:

$$\text{Mediana Ímpar} = x\left(\frac{n+1}{2}\right)$$

$$\text{Mediana Par} = \frac{x\left(\frac{n}{2}\right) + x\left(\frac{n}{2}+1\right)}{2}$$

Assim, podemos ver que a mediana é o valor central dos dados e, para calculá-la, é necessário, primeiramente, **ordenar os dados**.

Medidas Descritivas

Mediana

Exemplo: Idade de 10 funcionários do departamento de Marketing de uma empresa.

No caso do nosso exemplo das idades dos funcionários do departamento de Marketing, as mesmas já estão organizadas em ordem crescente. Caso elas estivessem ordenadas aleatoriamente, o primeiro passo seria arrumá-las na ordem crescente ou decrescente.

Após realizada a organização (quando necessário), o próximo passo é verificar qual (ou quais) idade está centralizada nos dados.

Nesse caso, temos dois valores centrais, 33 e 35, que deixam quatro observações atrás e quatro a frente. Logo, para n par, precisamos tirar média dessas duas idades:

$$\text{Mediana} = \frac{33+35}{2} = 34 \text{ anos}$$

Quartis, Variância e Desvio Padrão



Medidas Descritivas

Quartis e Percentis

Em resumo, os quartis e percentis são medidas estatísticas que fornecem informações sobre como os dados estão distribuídos, dividindo-os em partes específicas.

Essas medidas são úteis para entender a dispersão e a centralidade dos dados em uma distribuição.



Quartis

Os quartis dividem os dados em quatro partes iguais.

- O primeiro quartil (Q1) representa o valor abaixo do qual está 25% dos dados.
- O segundo quartil (Q2), que é equivalente à mediana, divide os dados em duas metades iguais; 50% dos dados estão abaixo dele e 50% estão acima.
- O terceiro quartil (Q3) representa o valor abaixo do qual está 75% dos dados.

A seguir, veremos um exemplo.



Quartis

Exemplo: Considere o conjunto de dados
 $\{10, 15, 20, 25, 30, 35, 40, 45, 50\}$.

Os quartis seriam:

$Q_1 = 17.5$ (mediana dos dados $\{10, 15, 20, 25\}$)

$Q_2 = 30$ (mediana dos dados $\{10, 15, 20, 25, 30, 35, 40, 45, 50\}$)

$Q_3 = 42.5$ (mediana dos dados $\{30, 35, 40, 45, 50\}$)



Percentis

- Os percentis dividem os dados em 100 partes iguais.
- O p-ésimo percentil é o valor abaixo do qual está p% dos dados.

Exemplo: Se quisermos encontrar o valor que está abaixo de 25% dos dados (o primeiro quartil), então estamos procurando pelo 25º percentil.

No conjunto de dados {10, 15, 20, 25, 30, 35, 40, 45, 50}, o valor correspondente ao 25º percentil (ou Q1) é 17.5.

Q2 = 30 (mediana dos dados {10, 15, 20, 25, 30, 35, 40, 45, 50})

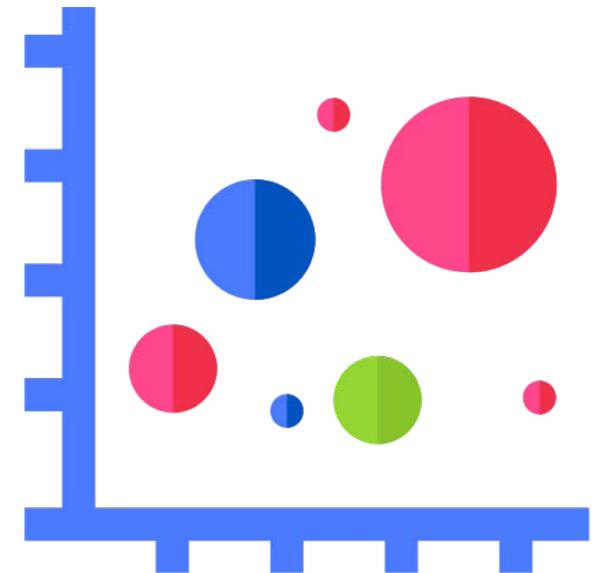
Q3 = 42.5 (mediana dos dados {30, 35, 40, 45, 50})

Conceito de Variação

No contexto da estatística, a variação refere-se à dispersão ou amplitude dos valores em um conjunto de dados. Quanto maior a variação, mais os valores estão dispersos em relação à medida de tendência central, como a média.

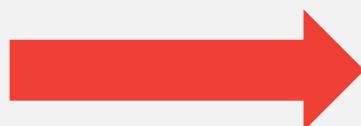
A variação é uma medida importante porque fornece informações sobre a consistência ou a dispersão dos dados, o que pode ser crucial para interpretar e analisar os dados de forma precisa.

Como ferramenta fundamental na análise estatística para quantificar a variação dos dados e entender sua distribuição, temos o cálculo de variância e o de desvio padrão.

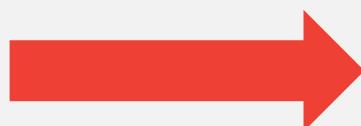


Variância

Essas medidas levam em consideração a **totalidade dos valores da variável** em estudo, e não apenas os valores externos, como a amplitude total. São índices de variabilidade bastante estáveis e, consequentemente, muito utilizados no cotidiano. Além disso, a variância e o desvio padrão complementam informações obtidas pelas medidas de tendência central.



É calculada como a média dos quadrados das diferenças entre cada valor de dados e a média.



Uma variância alta indica que os valores estão mais dispersos em torno da média, enquanto uma variância baixa indica que os valores estão mais próximos da média.

Variância

A fórmula para calcular a variância, denotada como σ^2 para população e s^2 para amostra, é:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (xi - \mu)^2$$

Onde:

- N é o tamanho da população ou da amostra.
- xi é cada valor individual no conjunto de dados.
- μ é a média dos valores no conjunto de dados.

Desvio Padrão

O desvio padrão é a raiz quadrada da variância e fornece uma medida de dispersão dos dados na mesma unidade que os dados originais.

Ele é amplamente utilizado devido à sua interpretação mais intuitiva em comparação com a variância.

Um desvio padrão maior indica uma maior dispersão dos dados em relação à média, enquanto um desvio padrão menor indica uma menor dispersão.

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

Lembrando que:
A fórmula para calcular a variância, denotada como σ^2 para população e s^2 para amostra.

Exemplo

Exemplo: Idade de 10 funcionários do departamento de Marketing de uma empresa. Voltando ao exemplo das idades no departamento de Marketing, vamos verificar qual a variância e o desvio padrão nesse caso.

Funcionários	1	2	3	4	5	6	7	8	9	10
Idade	22	25	25	31	33	35	40	42	48	50

Exemplo

Funcionários	1	2	3	4	5	6	7	8	9	10
Idade	22	25	25	31	33	35	40	42	48	50

A variância é a média dos quadrados das diferenças entre cada valor de dados e a média. Logo:

$$\text{Variância} = \frac{(22-35,1)^2 + (25-35,1)^2 + \dots + (50-35,1)^2}{10}$$

$$\text{Variância} = \frac{156,41 + 122,5 + \dots + 242,01}{10}$$

$$\text{Variância} = \frac{1665,01}{10} = 166,501$$

Exemplo

O desvio padrão é a raiz quadrada da variância. Logo:

$$\text{Variância} = \frac{(22-35,1)^2 + (25-35,1)^2 + \dots + (50-35,1)^2}{10}$$

$$\text{Variância} = \frac{156,41 + 122,5 + \dots + 242,01}{10}$$

$$\text{Variância} = \frac{1665,01}{10} = \boxed{166,501}$$



$$\text{Desvio Padrão} = \sqrt{166,501} \approx 12,90$$