### Bem-vindo ao mundo da Análise de Dados!

Você já parou para pensar em como a quantidade de dados digitais criados diariamente no mundo é desconcertante? E como a velocidade de produção desses dados está crescendo todo dia? Pois então, com tanto volume de informação, compreender e utilizar esses dados é um desafio que exige o conhecimento de técnicas e ferramentas diversas. Em 2008, o economista-chefe da Google, Hal Varian, afirmou: “A capacidade de utilizar dados, de ser capaz de entendê-los, processá-los, extrair valor deles, visualizá-los, comunicá-los: essa será uma habilidade importantíssima nas próximas décadas.”

Considerando isso, nesta aula, você vai conhecer alguns conceitos fundamentais sobre o contexto atual de criação e análise de dados, aprenderá a identificar tipos de dados e de variáveis, e começará utilizando a linguagem de programação **Python**, uma das ferramentas mais populares do mundo para essa tarefa. Ao concluir esse conteúdo, você poderá carregar conjuntos de dados de diferentes tipos e obter informações iniciais sobre eles.

Vamos lá?!

**Objetivos**

* Conhecer o contexto atual de geração e análise de dados;
* Identificar tipos de dados e variáveis;
* Aplicar operações básicas com conjuntos de dados tabulares.

**Tópico 1 – Introdução ao Big Data**

**OBJETIVOS**

* Compreender o paradigma do [Big Data](https://leadfortaleza.com.br/ead/glossary/Big_Data_Visualizacao_de_dados) e da Ciência de Dados;
* Identificar quais são os tipos de dados e a estrutura de dados tabulares;
* Identificar os tipos de variáveis em conjuntos de dados.

Dados estão em toda parte! Sejam dados científicos, empresariais, de redes sociais ou de aplicativos, a cada segundo uma enorme quantidade de dados é criada ao redor do mundo. Hoje, dados são um dos recursos mais valiosos do mundo, pois, a partir de sua análise, é possível gerar novos conhecimentos e tomar decisões mais inteligentes.

Entender e utilizar esse recurso é um dos grandes desafios da atualidade. Com base nisso, neste tópico, você irá conhecer de onde vêm e qual a importância desses dados. Além disso, você irá conhecer alguns dos conceitos fundamentais para descrever e compreender os diferentes tipos de dados. Então, vamos começar!

Você já leu ou ouviu falar sobre Big Data? Este termo é recente, mas já se tornou um dos mais importantes na área de tecnologia. De modo geral, o Big Data se refere ao imenso **volume**, **variedade** e **velocidade** de produção de dados que atingimos atualmente e que cresce todo dia. Durante os séculos XX e XXI, a tecnologia da informação se desenvolveu em ritmo acelerado, e, em paralelo a isso, a capacidade de processar, armazenar e transmitir dados também se aperfeiçoou. Assim, todos os dias, enormes quantidades de informação, de diferentes tipos e domínios, são geradas e armazenadas por empresas, governos e institutos de pesquisa.

Por exemplo:

* Empresas possuem dados de funcionários, clientes, vendas, logística e projetos;
* Bancos registram dados de clientes e diversas operações financeiras;
* Hospitais possuem dados sobre pacientes, como histórico médico, exames e laudos;
* Forças policiais geram dados sobre criminalidade;
* Instituições de ensino registram dados sobre alunos, professores, aulas e atividades;
* Aplicativos de mapas geram dados sobre localizações;
* Aplicativos de transportes geram dados sobre viagens.

A lista é interminável, você percebe? Você consegue pensar em outros exemplos? Confira, a seguir, um pouco mais sobre isso.

Talvez você esteja imaginando que o Big Data vem apenas de grandes instituições, mas não é bem assim. A geração de dados está cada vez maior e mais rápida por conta da popularização da internet, dos smartphones e das redes sociais. Por isso, todo e qualquer usuário está, constantemente, gerando dados de diferentes formatos, como textos, imagens e vídeos. Então, toda vez que você publica uma foto no Instagram ou um vídeo no YouTube; ou escreve um post no Twitter; ou mesmo, quando você está só navegando na web, você está gerando mais dados para esse grande universo de informação digital. Com base nisso, de acordo com algumas estimativas sobre as plataformas digitais, a cada **um minuto**:

* São publicadas 46.740 fotos no Instagram;
* São escritos 4.146.600 tweets no Twitter;
* São publicadas 527.760 fotos no Snapchat;
* São assistidos a 4.146.600 vídeos no Youtube.

São números consideráveis, não é mesmo? Então, o que você acha de saber mais sobre isso? Confira uma curiosidade a seguir.

#### Ícone Saiba Mais

Quão grandes são esses dados, afinal? Para se ter uma noção, estima-se que em um dia são produzidos cerca de 2,5 quintilhões de bytes. Você sabe o que isso significa? Imagine que todas essas informações fossem armazenadas em discos Blu-ray, que são discos de armazenamento de dados, com capacidade de armazenamento de 25GB; e esses discos fossem empilhados. O tamanho da pilha teria a altura de quatro Torres Eiffel, ou seja, 1200m, uma em cima da outra. Isso tudo em apenas um dia. É bastante, não é mesmo? E essa quantidade tende a dobrar a cada dois anos!

Agora que você compreendeu como é grande e diversa a quantidade de informação gerada no mundo, você pode estar se perguntando: ok, mas o que se pode fazer com esses dados? Calma, você vai descobrir isso em breve.

Dados representam **informação** em forma bruta. Para que essa informação seja útil, ela deve ser processada e analisada para se transformar em **conhecimento**. Seja para tomar melhores decisões de negócios, criar melhores políticas públicas ou desenvolver novos conhecimentos científicos e tecnologias, esses dados têm sido fundamentais para resolução de problemas em praticamente todas as áreas. Não é a toa que é dito que o **recurso mais valioso do mundo** não é mais petróleo, e sim os dados.

Porém, extrair conhecimento de grandes conjuntos de dados nem sempre é fácil. Dessa necessidade, surge a **Ciência de Dados**, uma área interdisciplinar que combina os conhecimentos da computação, da estatística e do domínio dos dados. Assim, cientistas de dados são profissionais que possuem diversas competências para coletar, manipular, explorar, analisar e identificar padrões em dados. Uma vez que dados são tão valiosos, a carreira de cientista de dados também tem se mostrado uma das mais bem-sucedidas atualmente. Em consequência disso, toda grande instituição quer fazer algo de útil com seus dados, e a demanda por esse profissional só tende a crescer.

Agora, não pense que é tão fácil assim, pois é preciso uma longa jornada para se tornar um cientista de dados. Porém, se você já tiver curiosidade por diferentes assuntos e disposição para estar sempre aprendendo, esse caminho vai se tornar bem mais agradável, pois estamos tratando de uma área que está em constante desenvolvimento. De qualquer forma, sendo um cientista ou não, o assunto sobre dados é algo bastante instigante e de interesse geral. Então, vamos dar continuidade ao nosso estudo? A seguir, conheça mais sobre os tipos de dados.

**1.1 Tipos de Conjuntos de Dados**

O que você imagina quando falamos em “*dados*”? A essa altura você já deve ter uma boa noção, mas vamos formalizar esse conceito, tudo bem? **Dados** são conjuntos de valores coletados através de observação, que representam informações sobre entidades, objetos ou fenômenos. Note que essa definição é muito ampla e não diz exatamente a forma dos dados. De fato, como já discutimos, dados podem ser muito variados:

* Nomes;
* Quantidades;
* Grupos;
* Valores estatísticos;
* Datas;
* Localizações;
* Textos;
* Imagens;
* Vídeos.

Observe também que, em geral, quando falamos em dados, estamos nos referindo a dados digitalizados, ou seja, aqueles armazenados e processados por dispositivos computacionais. Já os dados que representam informações de um mesmo domínio e que são armazenados juntos são chamados **conjuntos de dados**. Ah! Lembre-se que você também poderá se deparar com o termo em inglês, [**dataset**](https://leadfortaleza.com.br/ead/glossary/Dataset), entendido? Pois bem! Os conjuntos de dados são comumente classificados em dois grandes grupos: **estruturados** e **não-estruturados**. Confira a seguir a diferença entre eles.

Dados estruturados são chamados assim por terem uma estrutura de organização específica. A estrutura mais comum é a de **tabelas** compostas por linhas e colunas. Embora existam outros tipos, cada linha representa uma **observação**, também chamada de amostra ou instância; e cada coluna representa um **atributo**, ou variável, dessa amostra. A seguir, confira uma tabela que representa dados fictícios de uma faculdade. A tabela está dividida em quatro colunas: Aluno, Créditos cursados, Rendimento acadêmico, Mês de nascimento e Curso. Confira.

Tabela 1 – Rendimento do aluno

| **Aluno** | **Créditos cursados** | **Rendimento acadêmico** | **Mês de nascimento** | **Curso** |
| --- | --- | --- | --- | --- |
| Marina | 20 | 8,55 | Novembro | Computação |
| Felipe | 64 | 7,88 | Setembro | Estatística |
| Cleyton | 32 | 8,17 | Janeiro | Computação |
| Isabel | 24 | 9,04 | Julho | Matemática |

Perceba que os dados estão organizados de maneira lógica. Assim, cada linha representa uma pessoa, e cada coluna representa as respectivas características dessa pessoa. Quando os dados possuem a estrutura de tabela, eles também são chamados relacionais ou **tabulares**. Devido ao seu formato organizado, dados estruturados são mais facilmente processados por computadores. Além disso, esse foi um dos primeiros formatos para se representar dados digitalmente e ainda é muito utilizado hoje em dia, seja na forma de bancos de dados relacionais ou de planilhas. E você, já teve alguma experiência criando ou lendo tabelas? Se sim, então você já utilizou dados estruturados. Portanto, vamos prosseguir para conhecer os dados não-estruturados.

[Dados não-estruturados](https://leadfortaleza.com.br/ead/glossary/Dados%20n%C3%A3o-estruturados), como o nome indica, não possuem uma estrutura específica, o que os torna mais difíceis de serem processados e analisados em relação aos estruturados. Alguns exemplos de dados não estruturados são textos, imagens, vídeos e áudios. A seguir, verifique uma imagem que mostra a diferença entre a organização dos tipos de dados.

Figura 1 – Tipos de dados

Com a popularização em massa da Internet e da geração de conteúdo por usuários, a quantidade de [dados não-estruturados](https://leadfortaleza.com.br/ead/glossary/Dados%20n%C3%A3o-estruturados) aumentou significativamente, superando os [dados estruturados](https://leadfortaleza.com.br/ead/glossary/Dados%20estruturados). Porém, os [dados estruturados](https://leadfortaleza.com.br/ead/glossary/Dados%20estruturados) ainda são mais utilizados em análise de dados.

Tudo entendido? Até aqui, você verificou que [dados estruturados](https://leadfortaleza.com.br/ead/glossary/Dados%20estruturados) possuem forma de tabela com linhas que correspondem a observações, e colunas que correspondem a variáveis. Você percebeu que essas variáveis podem ser muito diferentes? No tópico a seguir, entenda melhor essas diferenças.

## 1.2 Tipos de Variáveis

Antes de conferir os tipos de variáveis, confira a tabela de dados novamente a seguir.

Tabela 1 – Rendimento do aluno

| **Aluno** | **Créditos cursados** | **Rendimento acadêmico** | **Mês de nascimento** | **Curso** |
| --- | --- | --- | --- | --- |
| Marina | 20 | 8,55 | Novembro | Computação |
| Felipe | 64 | 7,88 | Setembro | Estatística |
| Cleyton | 32 | 8,17 | Janeiro | Computação |
| Isabel | 24 | 9,04 | Julho | Matemática |

Note que a variável **Aluno** corresponde a um nome; a variável **Créditos cursados** é um valor numérico que corresponde à contagem do total de créditos que o aluno cursou; a variável **Rendimento acadêmico** é um valor numérico que corresponde ao desempenho de 0 a 10 do aluno naquele curso, permitindo valores decimais; a variável **Mês de Nascimento** é um nome de um mês em que o aluno nasceu; e, por fim, a variável **Curso** é um nome que indica o curso no qual o aluno está matriculado.

Desse jeito, as variáveis podem ser divididas em dois grandes tipos: **Quantitativas** e **Categóricas**. Assim, vamos começar pelas quantitativas, certo?

s variáveis **quantitativas** são valores numéricos obtidos por contagem ou medição. Sua natureza mensurável permite que sejam feitas operações aritméticas e cálculos de medidas estatísticas que resumem a variável para todas as amostras, como [**média**](https://leadfortaleza.com.br/ead/glossary/M%C3%A9dia), **mediana** e [**desvio padrão**](https://leadfortaleza.com.br/ead/glossary/Desvio%20padr%C3%A3o). Você está familiarizado com estas medidas? Caso não esteja, não se preocupe, você aprenderá mais sobre elas posteriormente. Até lá, vamos continuar com os tipos de variáveis.

Dentro das variáveis quantitativas, temos ainda as **contínuas** e **discretas**. Estas são números **inteiros** que são os resultados de contagem. Na nossa tabela fictícia, a variável **Créditos cursados** é uma variável quantitativa discreta, pois apresenta números inteiros que representam o resultado da contagem de créditos cursados. Outros possíveis exemplos de variáveis quantitativas discretas são número de filhos, número de casos de homicídios em um mês ou número de quartos em um hotel, por exemplo.

Já as variáveis contínuas são números **reais**, podendo assumir valores com casa decimal. Normalmente, são os resultados de medições. Na nossa tabela, a variável **Rendimento acadêmico** é um exemplo de variável quantitativa contínua. Outros exemplos possíveis são altura, peso, duração e temperatura.

Quanto às variáveis **categóricas**, também chamadas de qualitativas, elas representam uma característica ou qualidade de uma amostra que não pode ser quantificada numericamente. Portanto, não é possível realizar operações aritméticas sobre variáveis categóricas. Então, elas se subdividem em **nominais**, quando não há ordem inerente entre as categorias; e **ordinais**, quando as categorias podem ser ordenadas. Alguns exemplos de variáveis categóricas nominais são sexo, estado civil e cidade natal. Assim, perceba que são categorias, por exemplo, estado civil: solteiro ou casado.

Como exemplos de variáveis categóricas **ordinais**, temos escolaridade (Ensino Fundamental, Ensino Médio ou Ensino superior); tamanho de roupa (PP, P, M, G ou GG); mês de observação (janeiro a dezembro). Note que há uma ordem, por exemplo, janeiro, fevereiro, março, abril etc.

Na nossa tabela, as variáveis **Aluno** e **Curso** são categóricas nominais, enquanto a variável **Mês de Nascimento** é categórica ordinal.

Essa distinção entre os tipos de variáveis é importante, pois existem métodos específicos para analisar e representar cada tipo. A seguir, confira um box super interessante sobre variáveis.

#### Ícone Fique Atento!

Observe que nem sempre uma variável representada por números é quantitativa! Se a variável Mês de Nascimento fosse representada pelo número do mês, de 1 a 12, ao invés do nome, ela ainda seria categórica ordinal. Dessa forma, as variáveis categóricas podem ser representadas por números, mas observe que esses números não indicam quantidades, e, portanto, não se pode fazer operações aritméticas sobre eles. Por exemplo, no caso da variável Mês de Nascimento, não faz sentido somar ou subtrair o mês de nascimento de alunos, embora seja possível ordená-los por essa variável. Outros exemplos de variáveis categóricas representadas por números são número de telefone, CPF e número de endereço.

Após todos os conceitos que você aprendeu até aqui, você já deve estar querendo começar a programar, não é? No próximo tópico, você poderá fazer isso, mas, antes, vamos praticar um pouco o que você estudou até aqui, tudo bem?