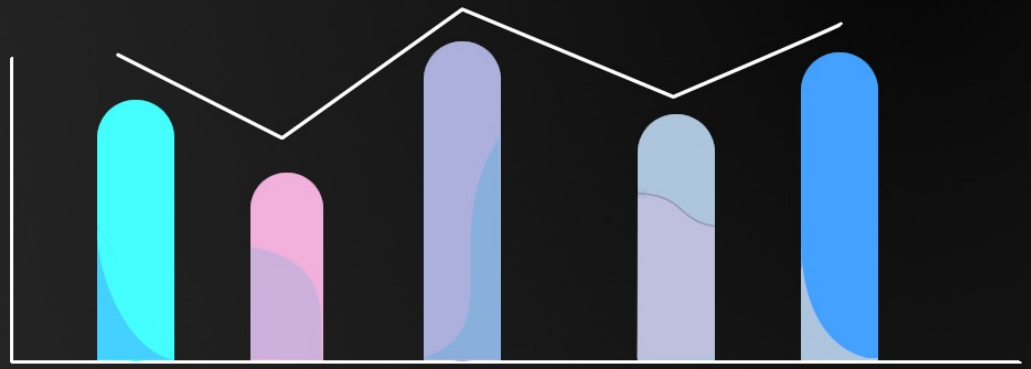
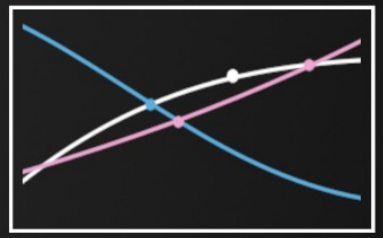
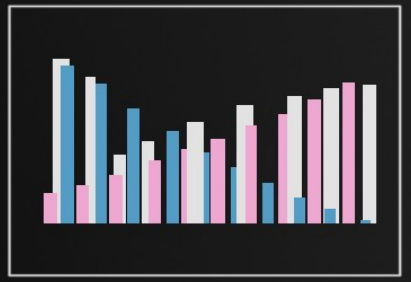
IMPACTRANS

# Ciência de Dados

O que é?
Como surgiu?
O que se estuda?



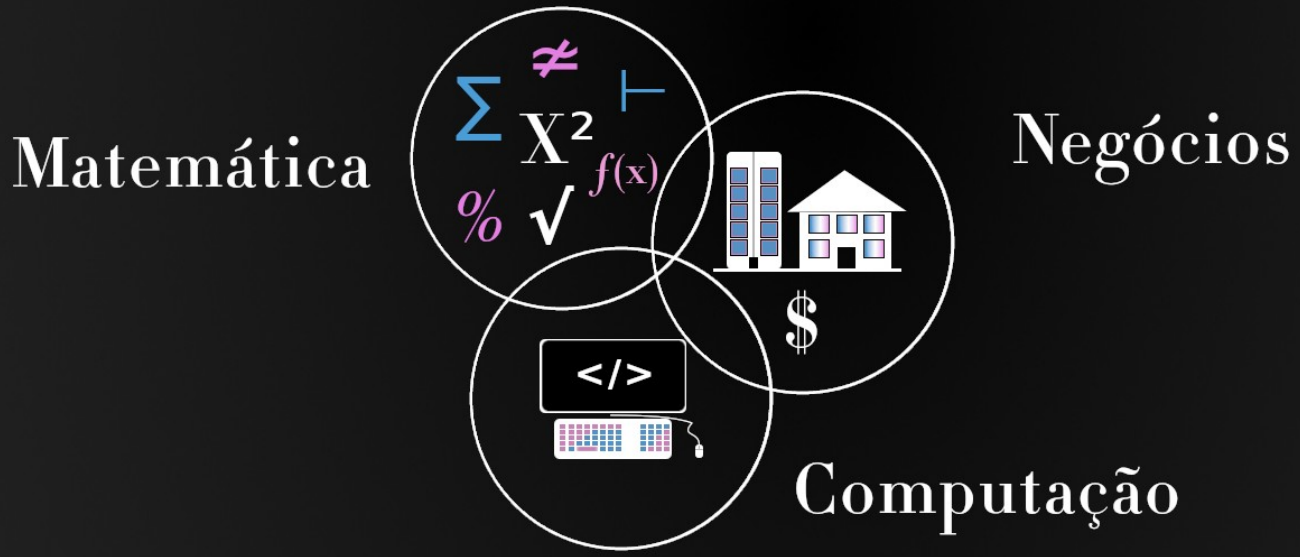




# O que é ciência de dados?

## Capítulo I

O acesso a informação nunca foi tão facilitado. Isso pode se dar pelo advento da internet.
Graças a isso a ciência de dados foi desenvolvida.
Existem várias área de conhecimento que envolvem essa ciência:



# Matemática.

# Principais conhecimentos

A ciência de dados se baseia em conteúdos estatístico. Segundo o portal **Brasil Escola:**

*"A estatística é o campo da matemática que relaciona fatos e números em que há um conjunto de métodos que nos possibilita coletar dados e analisá-los, assim sendo possível realizar alguma interpretação deles."*

**LUIZ. R, 2008**

Vamos ver:

- Visão geral;
- População;
- Classes e amostras
- Tabelas de distribuição;
- Medidas de posição;
- Medidas de dispersão;

*"Não há ramo da matemática, por mais abstrato que seja que não possa um dia vir a ser aplicado ao ramo do mundo real."*

**Lobachevsky**

# Estatística ao longo do tempo

Desde a formação dos primeiros povos, já haviam estatística, porém era reservada para uma pequena parte da população. Comerciantes, políticos e matemáticos eram os principais detentores do conhecimento estatístico, mas eram conhecimentos muito rasos, como contagem, por exemplo.

O primeiro trabalho sobre estatística foi publicado **John Graunt** em 1662, considerado por muitos como pai da estatística, publicou o artigo *"Observação natural e política realizada sobre taxas de mortalidade"* na cidade de Londres.

Seu trabalho foi considerado revolucionário por não somente ter tabelado (registrado) os índices de morte e nascimento, mas relacionar com mortalidade, morbilidades, desemprego, sexo e origem.

Desde então a estatística tem evoluído e com a sua principal associação com o cálculo de probabilidades.

# Tipo de estatística:

**Estatística descritiva**

A estatística descritiva tem como objetivo organizar e descrever (sem tirar conclusões) um conjunto de dados.

Pode ser considerado a primeira etapa de uma análise de dados.

**Estatística indutiva/inferencial**

A estatística inferencial tem como objetivo analisar e tirar conclusões sobre um conjunto de dados.

Como só analisamos parte de uma população, sempre é considerado um grau de variação ou incerteza.

# População/Universo estatístico

A população estatística é um conjunto de entidades ou seres que possuem ao menos uma característica em comum.

## Exemplos:

- Alunos de uma mesma **Universidade.**
- Moradores do estado do **Pará**
- Caixas com **pregos dentro**.

# Tipos de população

**População finita**

Possui um limitado número de elementos, sendo facilmente enumerados e medidos.

**Exemplo:**

- Alunos de uma oficina de computação do Impactrans;
- Oficinas do Impactrans.

**População infinita**

Possui um infinito de elementos, o que na prática não existe. Consideramos número elevado de registros como sendo infinitos.

**Exemplo:**

- Pessoas Trans.
- Pessoas de cor.

# Classes e amostras

**Classe**

É a categorização de um grupo populacional

**Exemplo:**

- **Mulheres trans** entre um grupo de pessoas trans;
- **Homens trans** com **mais de 18 anos** em um **grupo de homens trans.**

**Amostra**

É uma parte de um grupo de pessoas. É realizada devido a limitações práticas e econômicas.

**Exemplo:**

- Pessoas trans do estado de **São Paulo**;
- Pessoas LGBTQIA+ de **Itabira**, **Minas Gerais**.

# Tabelas de distribuições

**Geralmente** as tabelas estatística seguem a seguinte distribuição:

**Nome, idade e sexo e integração com a comunidade LGBTQIA+ de um grupo de pessoas da região de São Paulo.**

Título

| Nome | Idade | Sexo | LGBTQIA+ |
|---|---|---|---|
| João | 18 | M | Sim |
| Maria | 25 | F | Não |
| Juliana | 22 | F | Sim |

Cabeçalho

Célula

Corpo

# Medidas de posição

Medidas de posição ou medidas de tendência central, tentam descrever de forma **superficial** os dados.

Temos quatro medidas de dispersão:

- Média
- Mediana
- Moda

Para nossos exemplos vamos considerar o seguinte conjunto numérico:

**[8, 5, 3, 3, 7, 10, 21, 67]**

# Média

A média tem como função a tentativa de descrever os dados. É uma maneira simplificada de descrever pois com base nela não podemos afirmar hipóteses.

**[8, 5, 3, 3, 7, 10, 21, 67]**

$$X_{Médio} = \frac{X_1 + X_{2\,+}\, X_{3...} + X_n}{n}$$

$$X_{Médio} = \frac{8+5+3+3+7+10+21+67}{8} = \frac{124}{8} = \frac{31}{2} = 15{,}5$$

# Mediana

A mediana se refere ao valor central, ou seja, o valor do meio do conjunto de dados. Caso o número de registros seja ímpar, devemos apenas pegar o valor central, caso seja par, devemos pegar os dois valores centrais e tirar sua média.

**[8, 5, 3, 3, 7, 10, 21, 67]**

$$X_{Médio} = \frac{3+7}{2} = \frac{10}{2} = 5$$

# Moda

A moda se refere ao valor de maior frequência, ou seja, o valor que mais aparece, o valor da *"Moda"*. Vale lembrar que caso o conjunto tenha mais de uma moda, o conjunto é chamado de bimodal (duas modas) ou plurimodal (três ou mais).

**[8, 5, 3, 3, 7, 10, 21, 67]**

| Número | Frequência |
|---|---|
| 8 | 1 |
| 5 | 1 |
| 3 | 2 |
| 7 | 1 |
| 10 | 1 |
| 21 | 1 |
| 67 | 1 |

# Medidas de dispersão

Medidas de dispersão é um número que expressa o quão *distantes* estão os dados da média.

Temos duas medidas de dispersão:

- Variância (var/$\sigma^2$)
- Desvio padrão (std/$\sigma$)

Para nossos exemplos vamos considerar o seguinte conjunto numérico:

**[8, 5, 3, 3, 7, 10, 21, 67]**

# Desvio padrão e variância

Ambos estão muito bem relacionados. Para descobrir o desvio padrão, precisamos antes encontrar a variância.

**[8, 5, 3, 3, 7, 10, 21, 67]**

$$\sigma^2 = \frac{(X_1 - X_{médio})^2 + (X_2 - X_{médio})^2 \ldots (X_n - X_{médio})^2}{n}$$

$$\sigma = \sqrt{\sigma}$$

$$\sigma^2 = \frac{(8 - 15{,}5)^2 + (5 - 15{,}5)^2 \ldots (67 - 15{,}5)^2}{8} = \frac{3264}{8} = 408$$

$$\sigma = \sqrt{408} \approx 20.199$$

# Computação

# Principais conhecimentos

Vamos ver:

- Lógica de programação;
- Principais linguagens;
- Principais IDE's;
- Principais bibliotecas;

*"Algumas pessoas acham que foco significa dizer sim para a coisa em que você vai se focar. Mas não é nada disso. Significa dizer não às centenas de outras boas ideias que existem. Você precisa selecionar cuidadosamente."*

**Steve Jobs, 2008**

A ciência de dados tem como sua principal e fundamental aliada, a computação.

*"As vezes o que todos imaginam que não pode fazer nada, acaba fazendo aquilo que ninguém imaginava que poderia ser feito"*

**Alan Turing**

# Lógica de programação

A lógica de programação é a etapa inicial de aprendizado de qualquer linguagem de programação. Ela consiste em entender como um computador pensa, como lê informações, como armazena e como expressa essas informações. Quando você aprende lógica de programação em alguma linguagem, ela é a mesma para todas as outras, o que acaba mudando é a *sintaxe*.

Imagine que está aprendendo a linguagem *Java* e decide migrar para *Python*. Uma característica das linguagens de programação é que em algumas precisamos falar para o computador se o valor que estamos lidando é um úmero ou texto. Quando se programa em *Java,* a *sintaxe* exige que sempre avisemos para o computador que variável é essa, já em *Python* isso não é necessário, mas no final das contas, a definição de número e texto se mantém a mesma.

# Lógica de programação

Para aprender lógica de programação você precisa entender alguns assuntos:

- Sequência lógica
- Instruções
- Algorítimos
- Constantes
- Variáveis
- Tipos de variáveis
- Operadores (aritméticos e lógicos)
- Operações lógicas
- Estruturas (decisão e repetição)

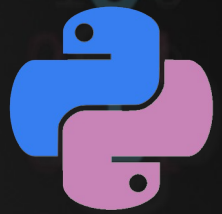
# Principais linguagens

No mercado existem diversas linguagens de programação. Provavelmente você já ouviu falar de alguma, como exemplo Java, JavaScript, CSS, Python, C#, C/C++ entre outras.

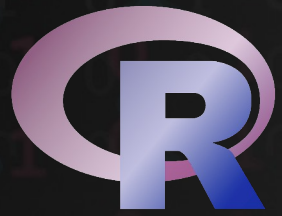
Apesar de todas conseguirem desemprenhar um ótimo papel em tudo, como criação de jogos, lidar com dados, desenvolver sistemas, criar sites, algumas linguagens são mais indicadas para determinadas coisas do que outras.

Quando falamos de ciência de dados, temos como principais linguagens *Python* e *R*.

Por questões práticas vamos abordar ciência de dados mais voltada para *Python*, mas se sinta na liberdade de considerar *R* como uma opção de linguagem.



VS



# Principais IDE’s

Uma IDE ou ambiente de desenvolvimento é onde desenvolvemos os códigos. Para ciência de dados tem por conveniência o uso de editores que possuem as estruturas chamadas de notebooks, como Jupyter, Jupyter Lab, VS CODE, Kaggle, Google Colab entre outros. Lembrando que todas as opções possuem vantagens e desvantagens. Quando for escolher sua IDE, leve em conta o que você está esperando.

# Principais bibliotecas

As bibliotecas são códigos que pessoas desenvolveram e disponibilizam online para que você não tenha que ter o trabalho de reescrever tudo de novo. Já pensou como pode ser complexo criar um código que exiba uma imagem na tela, imagina uma imagem que desenha um gráfico… Pessoas já fizeram isso e você pode acessar livremente. Aqui estão  algumas das bibliotecas por ordem sugerida de estudo:

# Principais bibliotecas

- Numpy – Manipulação de número de operações matemáticas com largos conjuntos de dados
- Pandas – Tabelação de dados. Permite visualizar e manipular dados orientados por índices e colunas.
- Matplotlib – Através dessa bilbioteca criamos gráficos e personalizamos com um alto detalhamento. É uma das bibliotecas mais extensas pelas suas diversas opções de personalização.
- Seaborn – É outra biblioteca para visualização de dados. Ela é mais fácil de usar e permite que crie gráficos bonitos digitando menos códigos, mas sempre que quiser uma modificação específica pode ser que não encontre na Seaborn e acabe tendo que apelas para o matplotlib.
- NLTK – É uma biblioteca para lidar com dados na forma de texto. Ela trás métodos que facilitam lidar com dados textuais e gerar classificações a partir deles.
- StatsModels – É uma biblioteca de foco estatístico. Com ela você pode criar predições e outras análises com um foco mais voltado para estatística.
- Sklearn – É uma biblioteca para pedição de dados e regressões lineares. Dado um conjunto de dados com determinadas entradas, é possível prever o valor de saída dos dados.
- TensorFlow e Keras – É uma biblioteca de criação de redes neurais. Para chegar nessa etapa é importantes que já domine todas as outras, mas quando chegar aqui, terá um grande diferencial no seu currículo.

# Negócios

# Ciência de dados e negócios

O maior motivo de ciência de dados ser revolucionário é por conta de sua predição. Imagine conseguir prever as vendas de um mês?

Com isso você poderá se preparar para as vendas, não comprando mais que precisará e gerando menos estoques.

Um conteúdo muito importante é a Inteligência empresarial, ou BI.

As organizações atualmente estão caminhando para serem orientadas a dados.

Obrigado!