# DATA SCIENCE

## AULA 1 - INTRODUÇÃO

PROFª. ANA CAROLINA B. ALBERTON

# BEM VINDOS

- Apresentação

# PYTHON

- Iremos utilizar o Jupyter (como IDE) e o Python (como Linguagem);
- Serão muito usados também arquivos .xlsx e .csv como fonte de dados para que seja mais tranquilo a manipulação dos dados.
- Para realizar a instalação, vocês podem acessar o link: https://www.anaconda.com/products/individual
- A partir das aulas de estatística já teremos alguns exemplos de código fonte e utilizaremos essa ferramenta;
- Vamos então à leitura do plano de ensino.

# AVALIAÇÕES

- Avaliações:
  - Nota 1 - Avaliação Teórica 08/04
  - Nota 2 - Desafios
  - Nota 3 - Elaboração do Projeto em 6 Etapas
    - Etapas 1 à 6

Média Final

MF=(N1+N2+N3) / 3

# OS DADOS

- **Big Data:**
- A cada segundo:
  - 100.000 tweets circulam
  - 547 websites são criados
  - mais de 2 milhões de pesquisas (Google)
  - 48h de vídeos são baixadas no YouTube
  - 684.478 itens são compartilhados no Facebook...
- Em governo (Brasil):
  - Mais de 7 milhões de notas fiscais eletrônicas (NFe) por dia
  - Mais de 16 bilhões de NFe autorizadas...

# OS DADOS

- Como lidar com este "dilúvio" de dados?
- A palavra mais importante no termo

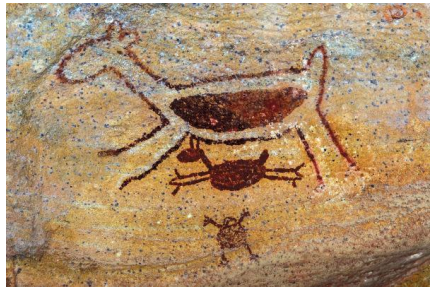
  "ciência de dados" não é "dados", mas "**ciência**".

# OS DADOS

- A partir da necessidade de análise desse emaranhado de dados surgiu uma "nova" área da ciência, a chamada ciência de dados
- O "quarto paradigma" da ciência
- A profissão mais "sexy" do século 21
- Uma nova buzzword!

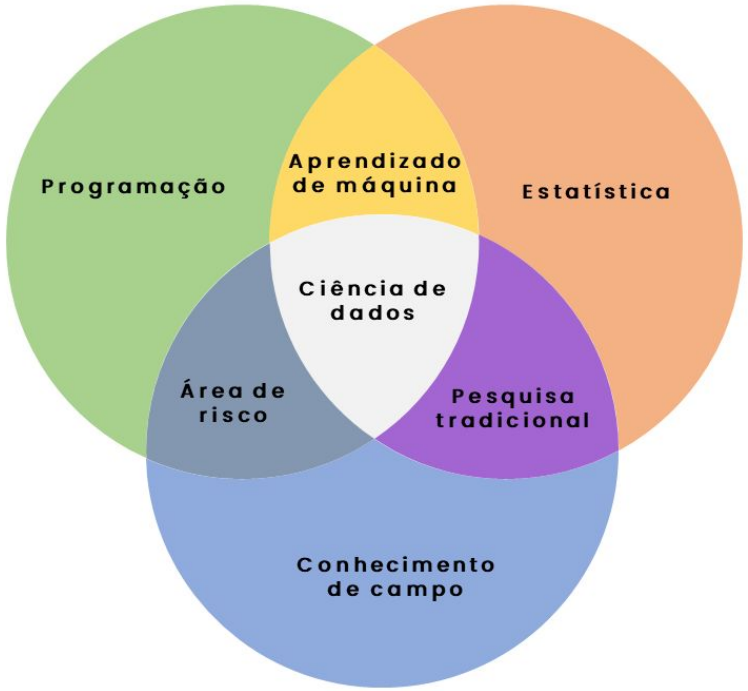
# O CIENTISTA DE DADOS

- As atividades executadas pelo "cientista de dados", em menor escala em relação ao volume de dados, são bastante antigas;
- Uma das características mais singulares dos seres humanos é a capacidade de registrar informações e situações vividas para o futuro;
- Inicialmente registrávamos fatos históricos, conquistas para se fazer propaganda dos reinos, criações de lendas e etc...
- Essa capacidade nasceu de forma quase orgânica em toda humanidade, tanto que atualmente, é possível compreender a história humana cruzando documentos históricos de diferentes povos.



# O CIENTISTA DE DADOS

- Competências do cientista de dados



# O CIENTISTA DE DADOS

- O que os cientistas de dados fazem?
  - "Considerando: Processos | Ambientes | Projetos"
  - Definem hipóteses e perguntas
  - Definem os conjuntos de dados ideais
  - Determinam que dados podem ser acessados
  - Adquirem os dados
  - Pré-processam os dados
  - Realizam análise de dados exploratória

# O CIENTISTA DE DADOS

- Realizam modelagem estatística dos dados
- Interpretam resultados de análises
- Escrevem relatórios sobre os resultados
- Criam modelos/componentes/códigos reusáveis
- Compartilham modelos e resultados com outras pessoas

# A CIÊNCIA DE DADOS

- É tão importante, que verdades absolutas podem ser alteradas quando descobrimos novas evidências, novos dados;

- Uma das primeiras coisas que nós iremos definir aqui na sala de aula é:
  - "**Não existe verdade absoluta quando tratamos com dados**".

- Então definimos a primeira regra da nossa aula, esse é apenas um exemplo de como a nossa grande capacidade de registrar informações consegue nos trazer diferentes decisões de acordo com a fonte que estamos trabalhando;

# A CIÊNCIA DE DADOS

- Exemplo:
- Nos esportes, uma mudança muito importante ocorreu, como havia muito dinheiro em jogo uma nova forma de montar times nasceu;

- Anteriormente, quando não tínhamos registros apurados de todos os jogadores os times eram formados baseado na fama e na indicações de pessoas experientes;

- Até que uma pessoa percebeu que era muito mais efetivo catalogar os movimentos dos jogadores e formar estatísticas.

# A CIÊNCIA DE DADOS

- Hoje a gente registra absolutamente tudo, de movimentos do mouse em um site, conversas pelo celular, fotos, capturamos registro de como o olho se movimenta em um site;

- E tudo isso serve para que tenhamos o maior número de informações possíveis para se tomar uma decisão, seja ela qual for;

- Existe um algoritmo de ciência de dados chamado Prophet, antigo liberado nos longínquos anos de 2017, que pode fazer previsões a nível de **segundo**.

- Por isso aqui nascem duas palavras muito importantes: **Big Data** e **Data Science**.

# A CIÊNCIA DE DADOS

- Basicamente o Big Data é a estrutura, ambiente ou o ecossistema onde todos os dados são armazenados;
- Então se nós passamos a registrar tudo, uma nova estrutura, diferente dos bancos tradicionais precisou ser criada, e essa estrutura se transformou no Big Data;
- Geralmente a pessoa indicada à movimentar a estrutura são os: **Engenheiros de Dados**.

# A CIÊNCIA DE DADOS

- E como agora a gente tem uma massa gigante de dados é humanamente impossível que alguém visualize tudo, pense, reflita e tome decisões;

- Por isso temos o Data Sciente, para que de maneira analítica, matemática e computacional, seja possível extrair informações relevantes dos dados;

- Hoje armazenamos dados de tudo, como Eric Schmidt uma vez disse: "Basta lembrar que ao postar alguma coisa, os computadores irão se lembrar eternamente";

- Portanto, o Data Science é basicamente o fluxo ou a forma a de como vamos utilizar os dados à nosso favor para realizar decisões.

# A CIÊNCIA DE DADOS

- E para isso iremos utilizar várias "fórmulas" e "scripts" criados através dos anos, nesse caso, já computacionais, para no fim auxiliar alguém a tomar **uma decisão**, **constatar um fato** ou **descobrir padrões**.

- Porém, como Jose Miguel Cansado, Diretor da Alto Data, disse: "Precisa se fazer a pergunta correta aos dados com os quais vocês estão trabalhando".

- Um exemplo, o nosso padrão etário é baixo e basicamente a grande maioria da classe mora na mesma região.

- E o que isso significa para o Data Science?

# A CIÊNCIA DE DADOS

- Bom isso significa que vocês todos passaram por experiências parecidas, possuem conhecimentos geográficos bem parecidos e, o mais importante para o Data Science, olhando apenas esse padrão reagiriam a **estímulos** de maneira similar.

- É claro, essa conclusão é rasa e provavelmente essa conclusão estaria errada.

- O que precisaríamos então para ter uma conclusão homogênea sobre como a classe seria corretamente impactada necessitaria de mais dados.

- Porém é impressionante, o que se pode ocorrer quando se tem dados suficientes para se tomar decisões, identificar padrões e etc...

# A CIÊNCIA DE DADOS

- Fávero e Belfiore (2020), apresentam em seu livro uma hierarquia que será bastante utilizada ao longo da matéria, esse é basicamente o começo, que é apresentada abaixo:

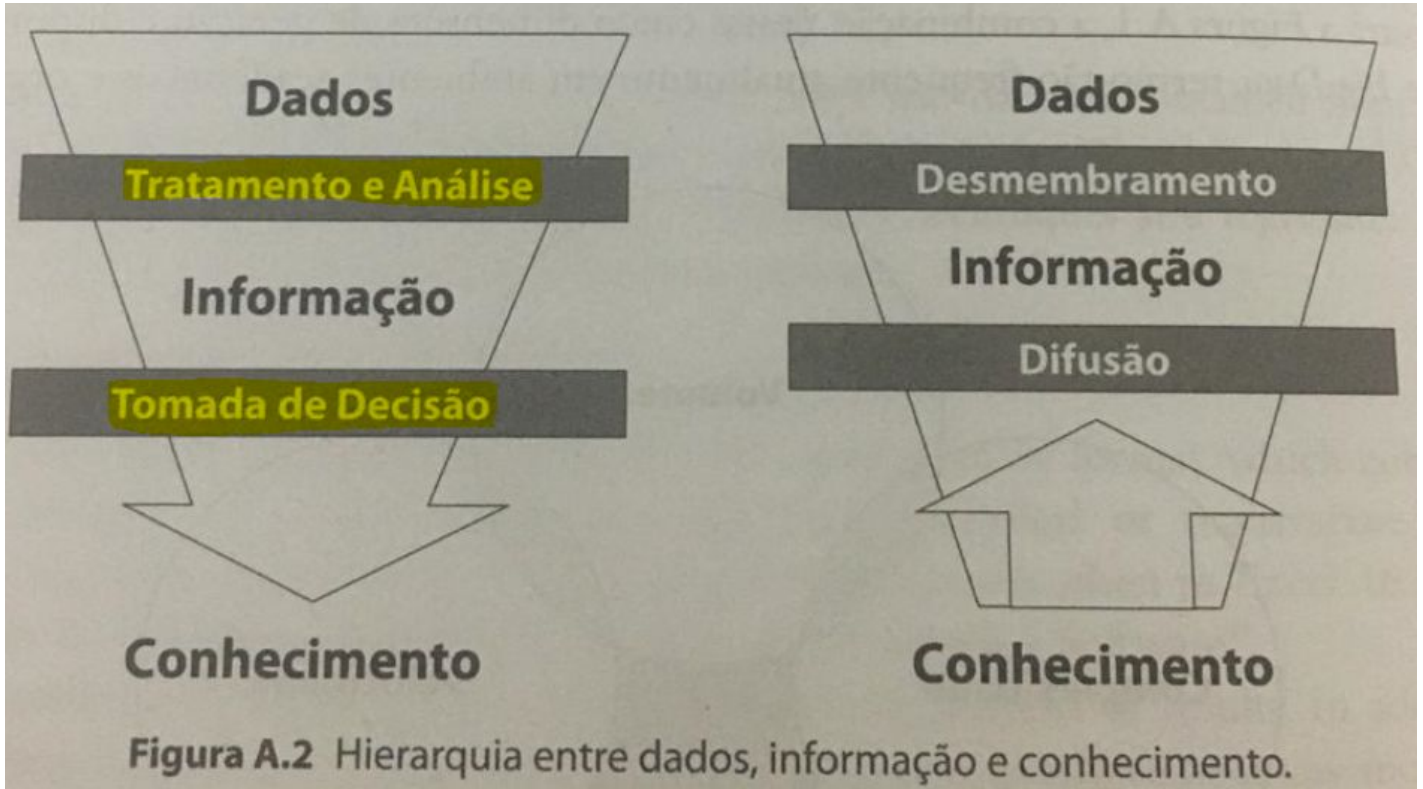



**Figura A.2** Hierarquia entre dados, informação e conhecimento.

# A CIÊNCIA DE DADOS

- Existem algumas características muito importantes que vocês, como estudantes de ciência de dados, devem prestar bastante atenção, são elas:

- Contexto, Padronização e Autenticidade.

- **Contexto**: Saber onde você está inserido, quais são as dificuldades e dores do seu entorno e porque você está enriquecendo o dado é essencial para poder obter bons resultados. Afinal, quanto menos específico for o seu modelo, menor vai ser a sua assertividade, pois aspectos importantes do contexto foram negligenciados para atingir um espectro maior.

# A CIÊNCIA DE DADOS

- **Padronização**: Os dados devem possuir um padrão, seja ele de tipo de dado (numerico, texto..), seja a ordem cronológica (dados diários, mensais, semanais...) ou em relação à origem. Se os dados não forem padronizados o seu modelo levará à informações erradas, portanto mesmo que ele chegue até você de formas diferentes, padronize-o, para que assim você possa ter uma análise correta

- **Autenticidade**: Os dados que você tem acesso precisa ter uma fonte confiável, o que é mais fácil quando se lida com dados internos. Quando utilizar dados externos certifique-se de que os dados são verídicos, se possível cruzando com dados já existentes ou verificando possíveis distorções inadequadas ao tipo de informação que você está tratando.

- Dúvidas?

# REVISÃO

- **TIPO DE VARIÁVEIS**
- **Variável** é uma característica da população (ou amostra) em estudo, possível de ser medida, contada ou categorizada (FÁVERO; BELFIORE, 2020).

- O tipo de variável é crucial no cálculo de estatísticas e representação dos seus resultados.

- Portanto, as variáveis podem ser classificadas da seguinte forma:
  - Qualitativa (Não Métrica ou Categórica);
  - Quantitativa (Métrica).

# REVISÃO

- VARIÁVEL QUALITATIVA
- Como o nome já indica, uma variável **Qualitativa**, representa uma qualidade de um indivíduo, objeto ou elemento que não podem ser medidas ou quantificadas.

- Podemos simplificar essa questão dizendo que aquilo que não pode ser representado como numericamente pode ser uma variável qualitativa.

  - Exemplos:
    - Nome de uma pessoa;
    - Cor dos olhos;
    - Cidade de nascimento.

# REVISÃO

- Geralmente, sem realizar nenhuma transformação, as variáveis qualitativas não podem apresentar medidas-resumo, tais como média, mediana e desvio-padrão.

- Porém, o que ocorre geralmente numa situação onde existem variáveis qualitativas à serem trabalhadas, é trabalhar com uma faixa de frequência de uma determinada ocorrência.

- Vejamos um exemplo: Vamos fazer uma pesquisa para identificar as marcas mais comuns de smartphone na sala de aula.

# REVISÃO

- VARIÁVEL QUANTITATIVA
- Uma variável **Quantitativa**, podem ser representadas de forma gráfica e geralmente são numéricas e, diferentemente das qualitativas, podem apresentar medidas-resumo, como média, quartis, decis e etc...)

- As variáveis quantitativas podem ainda ser divididas novamente em dois sub-tipos, são eles:
  - Discretas: Geralmente provêm de uma contagem e estão dentro de um conjunto finito, portanto geralmente são números inteiros. (Ex: Número de Filhos).
  - Contínuas: Geralmente provêm de uma medição e estão dentro de um conjunto de números reais. (Ex: Peso, Tamanho do Calçado).

# REVISÃO

- Além dos seus tipos e/ou subtipos as variáveis podem também ser classificadas de acordo com as suas escalas de mensuração.

- Existem, na literatura, diversas formas de escalas de mensuração, porém iremos utilizar a classificação que define que:

  - As variáveis qualitativas podem ser classificadas **Nominais** e **Ordinais**;

  - As variáveis quantitativas são classificadas como **Intervalares** e **Razão**.

# REVISÃO

- VARIAVEIS QUALITATIVAS – NOMINAIS

- Variáveis de escala nominais, são variáveis cujo o valor não representam ou não possuem nenhuma relação com escalas de grandeza ou de ordem.

- Como o nome já direciona, geralmente variáveis nominais são apenas substantivos, que carregam nomes, lugares ou profissão.

- Portanto operações matemáticas como adição, média e desvio padrão não são possível, porém outros métodos estatísticos podem ser aplicados, por exemplo é possível realizar contagem de números de ocorrências em uma determinada população.

# REVISÃO

- VARIAVEIS QUALITATIVAS NOMINAIS

| Empresa | Pais de Origem |
|---|---|
| Exxon Mobil | Estados Unidos |
| JP Morgan Chase | Estados Unidos |
| General Eletric | Estados Unidos |
| Royal Dutch Shell | Holanda |
| ICBC | China |
| HSBC Holdings | Reino Unido |

Sem escala de grandeza

# REVISÃO

- VARIAVEIS QUALITATIVAS ORDINAIS

- Variáveis de escala ordinais, são variáveis cujo o valor representam classificações ou padrões de ordem entre diferentes categorias.

- Como o nome já direciona, geralmente variáveis ordinais são apenas classificatórias, que carregam escalas ou níveis.

- Portanto operações matemáticas como adição, média e desvio padrão também não são possíveis, já que o objetivo da aplicação dessa variável é justamente classificar um padrão, podendo ou não ter sido advindo de operações matemáticas.

# REVISÃO

- VARIAVEIS QUALITATIVAS ORDINAIS

| Valor | Rótulo |
|---|---|
| 1 | Péssimo |
| 2 | Bom |
| 3 | Muito Bom |
| 4 | Ótimo |
| 5 | Excelente |

- Classificam um Padrão
- O exemplo mostra uma tabela com uma variável que indica ordem (Ordinal).

# REVISÃO

- VARIAVEIS QUANTITATIVAS INTERVALARES

- Variáveis de escala intervalares, são variáveis cujo o valor representam uma escala de valor já estabelecida, geralmente com pontos de origem e/ou fim já pré-determinados.

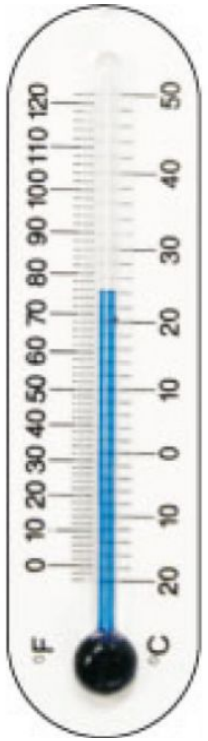
- O simples exemplo de uma variável quantitativa intervalar são as faixas de temperatura. Um outro exemplo poderia ser as horas do dia, dias do mês, meses do ano e Etc...

- Uma característica importante das variáveis intervalares é que o zero (ou ponto de origem) não necessariamente exprime ausência de valor. Já que dentro da escala esse valor tem a sua característica determinada.

# REVISÃO

- VARIAVEIS QUANTITATIVAS INTERVALARES

  - Pontos de Fim Determinados
  - O Zero não representa ausência de valor
  - Um termômetro talvez seja o exemplo mais
  - claro de uma variável intervalar.



# REVISÃO

- VARIAVEIS QUANTITATIVAS – RAZÃO

- Variáveis de escala de razão, são variáveis cujo o valor ordena as unidades relacionada à característica e possui uma unidade de medida constante.

- Basicamente qualquer valor número, que não seja intervalar, é considerado uma variável de razão.

- Dentre as variáveis quantitativas de razão podemos citar valores como: Peso, Altura, Valor Salarial, Idade e Etc...

# DESAFIO 1

1. Qual a diferença entre as variáveis qualitativas e quantitativas? Dê exemplos.

1. (ENADE)No questionário socioeconômico que faz parte integrante do ENADE há questões que abordam as seguintes informações sobre o aluno:
   I - Unidade da Federação em que nasceu;
   II - número de irmãos;
   III - faixa de renda mensal da família;
   IV - estado civil;
   V - horas por semana de dedicação aos estudos.

   São qualitativas **APENAS** as variáveis:

   A) I e III;

   B) I e IV;

   C) I, IV e V;

   D) II, III e V;

   E) I, II, IV e V.

# REFERÊNCIAS


- FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de Análise de Dados:** Estatística e Modelagem Multivariada com Excel, SPSS e Stata. Rio de Janeiro: Ltc, 2020.
- CIFERRI, Cristina Dutra de Aguiar; CIFERRI, Ricardo Rodrigues. **Modelagem Multidimensional.** São Paulo: Usp, 2020. 14 slides, color. Disponível em: <http://wiki.icmc.usp.br/images/6/6a/SCC5911-02-ModelagemMultidimensional.pdf>. Acesso em: 10 jan. 2020.
- RESENDE, Tânia. **Modelagem multidimensional conceitos básicos.** São Paulo: Slideshare, 2016. 28 slides, color. Disponível em: <https://pt.slideshare.net/TANIARESENDE/modelagem-multidimensional-conceitos-bsicos>. Acesso em: 18 fev. 2020.
- JARDIM, Edgar Silveira; OLIVEIRA, Marcus Vinícius Abreu de; MORAVIA, Rodrigo Vitorino. Diferença Entre Banco de Dados Relacional e Banco de Dados Dimensional. **Revista Pensar Tecnologia**, Belo Horizonte, v. 2, n. 4, p.1-17, julho 2015. Mensal. Disponível em: <http://revistapensar.com.br/tecnologia/pasta_upload/artigos/a122.pdf>. Acesso em: 18 fev. 2020.
- SERGENTI, Alexsandro. **Modelagem Relacional e Multidimensional:** uma análise envolvendo Sistemas de Apoio a decisão. 2015. Disponível em: <https://www.linkedin.com/pulse/modelagem-relacional-e-multidimensional-uma-an%C3%A1lise-de-sergenti/>. Acesso em: 18 fev. 2020.