Computação para a Ciência dos Dados

O que vimos...

Prof. André Filipe de Moraes Batista, PhD.
Prof. Michel Silva Fornaciali, PhD.
Prof. Daniel de Souza Carvalho, MEng.

Contatos:
andreFMB@insper.edu.br
michelSF@insper.edu.br
danielsc1@insper.edu.br

Insper

Nossa dinâmica de trabalho

Fail Fast, Learn Faster!

Aula e Lab juntos

- Só datasets reais!
- Pensando como cientistas de dados desde a primeira aula!

Insper

Recaptulação

O que vimos nas aulas anteriores

- **Aula 01**
  - Lógica de Programação
  - Fluxo classico de DS
  - Python
  - Enviroments
  - Anaconda – instalação
  - Variáveis e tipos de dados
  - Operadores lógicos
  - Controle de fluxo
  - Estruturas de armazenamento: lista, tuple, dicionário
  - List comprehension

Recaptulação

O que vimos nas aulas anteriores

- **Aula 02**
  - Ecossistema de bibliotecas Python
  - Numpy
    - Numpy array, desempenho np.array vs lista, operações matemáticas, reshaping, slicing
  - Pandas
    - Pandas Series e DataFrames
    - Heterogeneidade dos dados
    - Índices
    - Filtros (máscaras e sql-like)
    - Ordenação
    - value_counts, nlargest
  - Datasets reais: imóveis em SP, consumo elétrico SEADE

Recaptulação

O que vimos nas aulas anteriores

- **Aula 03**
  - Interação Pandas – Matplotlib
  - Data Wrangling
  - Colunas calculadas – Score para imóveis de SP
  - Transformação de dados
  - Concatenação
  - Merges
  - Agrupamentos

- LAB

Insper - Computação para a Ciência dos Dados

Dados Abertos da Câmara dos Deputados

Imagem: https://dadosabertos.camara.leg.br/img/ilustrations/ipad-ilustration.png

Esta prática faz uso de *dados reais* disponibilizados no portal de dados abertos da Câmara dos Deputados, cujo endereço eletrônico é https://dadosabertos.camara.leg.br

Iremos desenvolver os seguintes conceitos com essa prática:

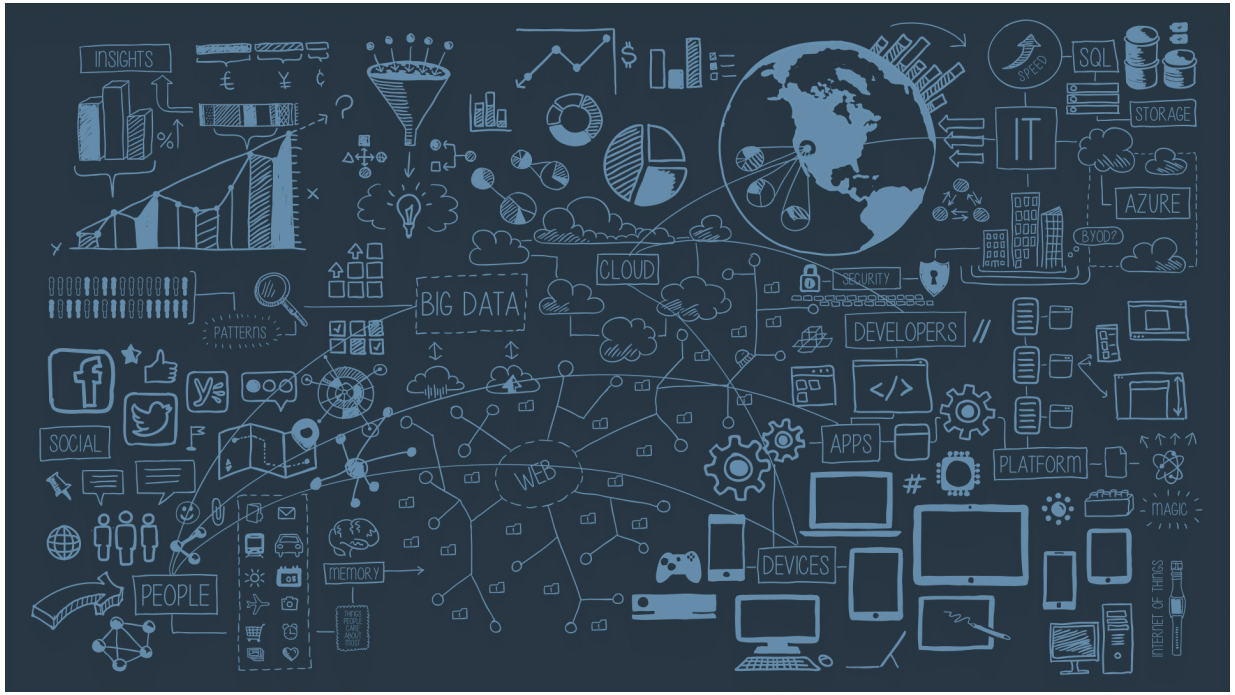
1. Leitura de arquivos csv com a biblioteca Pandas
2. Merge de DataFrames
3. Criação de novas colunas
4. Trabalhar com dados DateTime
5. Fazer uso da função *apply*
6. Fazer uso da função *nlargest*
7. Obter estatísticas descritivas básicas, integrando - inclusive - com a biblioteca MatplotLib

Três arquivos estão disponibilizados para a prática, quais sejam:

- `deputados.csv` - série histórica de todos os deputados federais do Brasil
- `eventos-2020.csv` - eventos da Câmara dos Deputados no ano de 2020
- `eventosPresencaDeputados-2020.csv` - presença de cada deputado nos eventos da Câmara em 2020.

Aula 04

- Números aleatórios e Random Walks
- Gráficos
- Matplotlib
- Seaborn



Insper

Aula 05

- Gráficos com Seaborn e Altair
- Tidy Data
- Formatos especiais de arquivos: .sav
- Discussão: Projeto integrador
- Lista 2

Aula 06

- Revisão
- Checkpoint
- Projeto Integrador

Aula 07

- Python – Envs
- Holoviz
- Aplicação na base da pesquisa TIC 2019

Aula 08

- Programação Funcional
- Consumo de API – Formato JSON

Aula 09

- Análise Espacial
- GeoJSON e dados geo
- O que as pessoas reclamam em NYC?
- GeoPandas
- Projeto Integrador

Aula 10

- Scraping
- Selenium

- Projeto Integrador