Introdução à Engenharia de dados e Exemplos práticos com pandas

Por Gabriel Taufer

Minha história com a engenharia de dados

Meu primeiro contato e como transicionei para um cargo de engenharia de dados

- Trabalho como engenheiro de dados há 2-3 anos
- Antes disso trabalhava com web/serverless (Django, Chalice, FastAPI)
- Recentemente assumi a liderança do time de dados
- Não sabia quase nada quando comecei

Tá, mas o que é engenharia de dados?

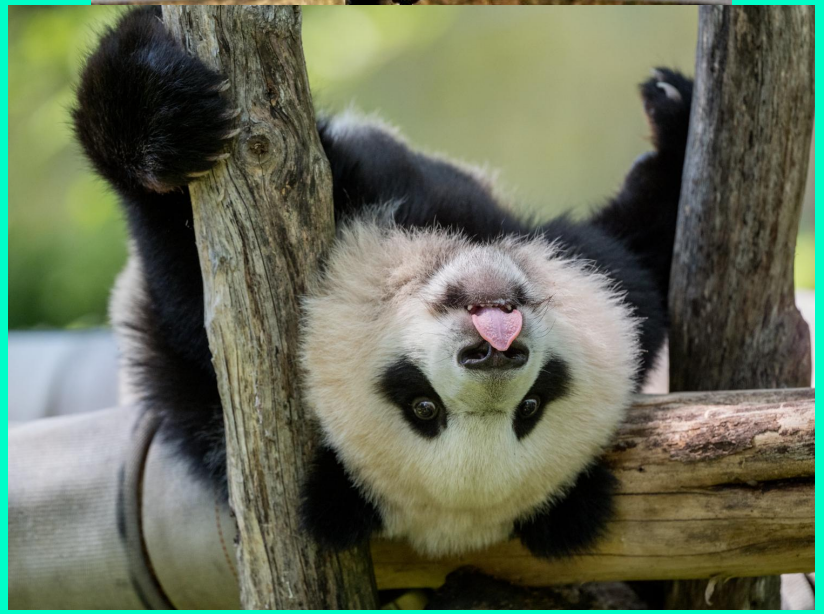
- Processo de **coleta**, **organização**, **processamento** e **análise de dados**
- **Automação** destes processos visando economizar tempo e recursos, sempre em busca de **agregar valor**
- Transforma grandes volumes de dados em **insights**, cruciais para **tomada de decisão**
- Incentiva **inovação** e **competitividade**

“You can’t manage what you can’t measure*”
– Drucker, Peter

**Você não pode gerenciar aquilo que não consegue medir*

Entre ursos e serpentes

Sobre o papel do Python na análise de dados e a biblioteca Pandas



Entre tantas opções, Por que python?

- Código legível e compreensível
- Curva de aprendizado reduzida, foco na análise de dados
- Vasta gama de bibliotecas e frameworks
  - Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn para machine learning, TensorFlow e PyTorch para deep learning, entre outras.
- Supre desde a análise exploratória de dados até o desenvolvimento de modelos de machine/deep learning
- Comunidade de ativa e colaborativa (documentação, fóruns, tutoriais e bibliotecas open source)

E Por fim, o pandas

- Biblioteca de código aberto, utilizada para análise e manipulação de dados
- Estruturas de dados flexíveis
- Ferramentas eficientes para lidar com conjuntos de dados de diferentes formatos e tamanhos
- **DataFrame:**
  - Estrutura de dados tabular bidimensional com rótulos de linhas e colunas, semelhante a uma planilha ou tabela de banco de dados.
- **Series:**
  - Uma estrutura de dados unidimensional que pode conter qualquer tipo de dado, semelhante a uma matriz ou lista.

Pandas em ação

Exemplos práticos e live coding utilizando conceitos básicos do Pandas

Situação

- Recebemos dois arquivos .CSV diferentes, um contendo informações sobre clientes, e outro sobre assinaturas
- Os dados dos CSVs podem ou não estar bagunçados (formatos inconsistentes, dados faltando, tipos de dados incoerentes, etc)
- Os dois arquivos podem ser relacionados a partir do ID do cliente
- Nem todo o cliente terá uma assinatura correspondente

Objetivos

- **Coleta:** essa parte foi simplificada, tudo que precisamos fazer é ler os arquivos
- **Organização:** como mencionado anteriormente, os dados estão bagunçados, então precisamos limpá-los e adequá-los ao formato requisitado (simula conversão e salvamento no banco)
- **Processamento:** após termos os dados normalizados, iremos relacioná-los
- **Análise de dados:** finalmente, após todo o processo de coleta, limpeza e processamento, iremos buscar algumas métricas:
    - Número de usuários por estado
    - Número de usuários por plano e modalidade
    - Número de usuários por faixa etária (0-18, 19-35, 36-50, 51+)
    - *Opcional: gráfico (de linha) mostrando a soma cumulativa da quantidade assinaturas por mês*
    - *Opcional: gráfico (de barra) mostrando a quantidade de assinaturas por estado*