escola britânica de artes criativas & tecnologia

Módulo | Análise de Dados: Coleta de Dados I

Caderno de **Exercícios**
Professor André Perez

Tópicos

1. Arquivos CSV;
2. Arquivos Texto;
3. Arquivos Excel.

Exercícios

0. Preparando o ambiente

Vamos explorar dados de crédito presentes no arquivo `credito.xlsx` (link). Os dados estão no formato de **Excel** (XLSX) e contém informações sobre clientes de uma instituição financeira. Em especial, estamos interessados em explicar a segunda coluna, chamada de **default**, que indica se um cliente é adimplente (`default = 0`), ou inadimplente (`default = 1`), ou seja, queremos entender o porque um cliente deixa de honrar com suas dívidas baseado no comportamento de outros atributos, como salário, escolaridade e movimentação financeira. Uma descrição completa dos atributos está abaixo.

Coluna	Descrição
id	Número da conta
default	Indica se o cliente é adimplente (0) ou inadimplente (1)
idade	---
sexo	---
depedentes	---
escolaridade	---
estado_civil	---

Coluna	Descrição
salario_anual	Faixa do salario mensal multiplicado por 12
tipo_cartao	Categoria do cartao: blue, silver, gold e platinium
meses_de_relacionamento	Quantidade de meses desde a abertura da conta
qtd_produtos	Quantidade de produtos contratados
iteracoes_12m	Quantidade de iteracoes com o cliente no último ano
meses_inatico_12m	Quantidade de meses que o cliente ficou inativo no último ano
limite_credito	Valor do limite do cartão de crédito
valor_transacoes_12m	Soma total do valor das transações no cartão de crédito no último ano
qtd_transacoes_12m	Quantidade total de transações no cartão de crédito no último ano

Faça o download do arquivo `credito.xlsx` com a célula de código abaixo.

In [ ]:
```
!wget --show-progress --continue -O ./credito.xlsx \
https://raw.githubusercontent.com/andre-marcos-perez/\
ebac-course-utils/main/dataset/credito.xlsx
```

---

1. Excel para CSV

Utilizando o pacote Python `openpyxl` visto em aula, extraia os seguintes as colunas `id`, `sexo` e `idade` para dos clientes inadimplentes (`default = 1`) e solteiros (`estado_civil = 'solteiro'`). Salves os dados extraídos no arquivo csv `credito.csv` separado por `;` . Exemplo do cabeçalho e das três primeiras linhas:

```
id;sexo;idade
767712558;59;M
713741358;46;M
772390908;59;M
```

Dica: O arquivo csv `credito.csv` deve ter 669 linhas, contando com o cabeçalho.

Nota: Escreva o código da sua solução abaixo em uma ou mais células, você não precisa enviar o arquivo csv gerado.

In [ ]:
```
# solução do exercício 1
```

---

2. Excel para JSON

Como preparação para o próximo módulo, vamos trabalhar com o JSON, um formato semi-estruturado, muito utilizado em transmissão de dados da web e equivalente a um **dicionário** Python.

Utilizando o pacote Python `openpyxl` visto em aula, extraia os dados das colunas `escolaridade` e `tipo_cartao`, removendo duplicados. Com os dados, construa o dicionário Python `credito` com a seguinte estrutura:

```
credito = {
  'tipo_cartao': ['silver', 'blue', 'gold', 'platinum'],
  'escolaridade': ['doutorado', 'mestrado', 'na', 'sem educacao
formal', 'graduacao', 'ensino medio']
}
```

Para finalizar, utilize o código abaixo para converter o dicionário `credito` no formato JSON:

```
import json

credito_json = json.dumps(credito, indent=4)
print(credito_json)
```

Dica: Sua solução deve gerar o dicionário Python `credito` igual ao exemplo mas a ordem dos elementos pode variar tranquilamente.

Dica: Uma excelente forma de remover elementos duplicados de uma lista é convertê-la para `set` e depois para `list` novamente.

In [ ]:
```
# solução do exercício 2
```

---

3. BÔNUS: Texto para CSV

No arquivo de texto `ebac.txt` você encontra o texto presente no rodapé da página de cursos da EBAC (link).

Arquivo TXT: ebac.txt

In [ ]:
```
%%writefile ebac.txt
MÍDIAS SOCIAIS
Instagram, Facebook, Youtube, LinkedIn

CURSOS
Software, Design, Marketing, Audiovisual, Programação & Data, Games

WEBINARS
Próximos, Anteriores

SOBRE
Sobre nós, Centro de carreiras, Vagas

CONTATO
WhatsApp +55 (11) 4200-2991
Telefone +55 (11) 3030-3200

BLOG
Design, Audiovisual, Marketing
```

Extraia os números de contato do arquivo texto `ebac.txt` e salve-os no arquivo csv `contato_ebac.csv` com o separador `;` no seguinte formato:

```
tipo;numero
whatsapp;+551142002991
telefone;+551130303200
```

Nota: Escreva o código da sua solução abaixo em uma ou mais células, você não precisa enviar o arquivo csv gerado.

In [ ]:
```
# solução do exercício 3 (bônus)
```
