Pandas e Esteganografia

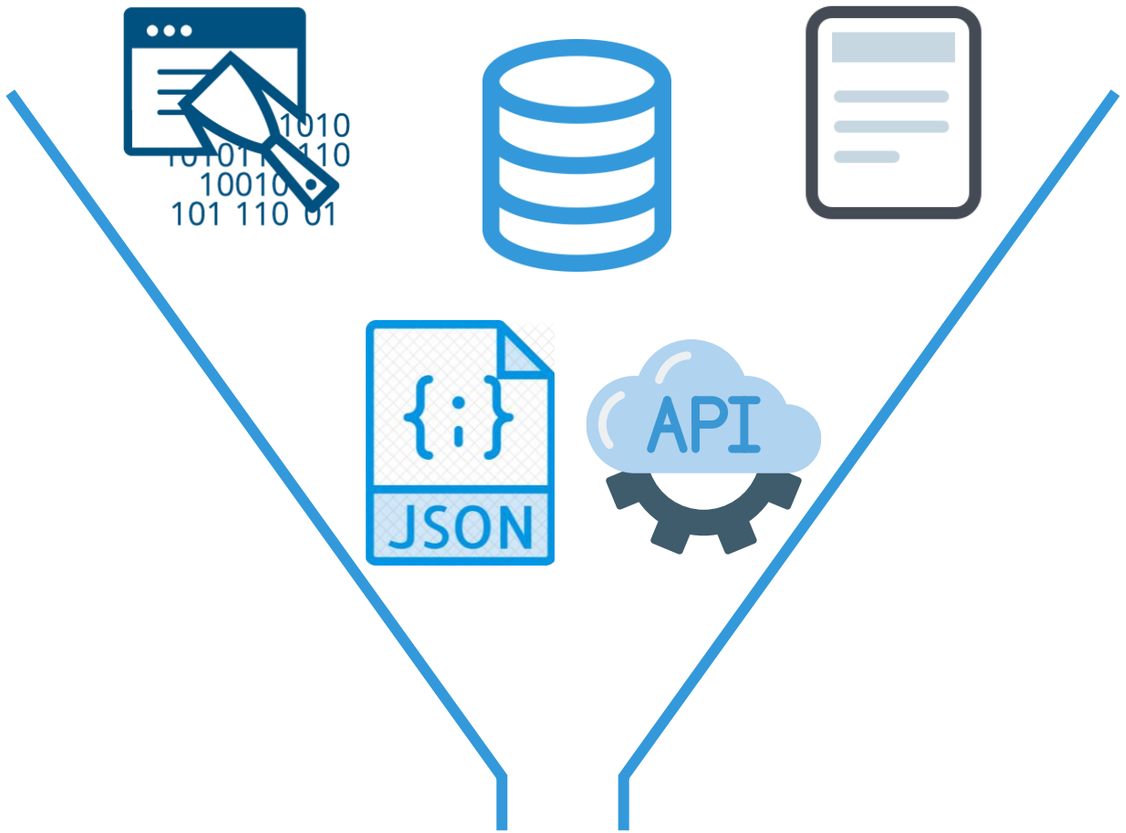


Imagem: https://clearbit.com/our-data

Professor: Alex Pereira

Regex em Python

Regex em Python

- Regex = Regular Expression
  - Uma expressão regular representa um conjunto de expressões/sentenças
    - ✓ Que seguem uma regra de construção
- Sintaxe das regex (Resumo)
  - [a-g] qualquer caractere entre a & g
  - \w \d \s palavra, dígito, espaço em branco
  - ^abc$ início / fim de uma string
  - a* a+ a? 0 ou mais, 1 ou mais 0 ou 1
  - a{5} a{2,} exatamente cinco, dois ou mais
  - a{1,3} entre um & três
  - ab|cd encontrar ab ou cd
  - \. * \\ caracteres especiais escapados

Exemplos de Regex

- CPF: 245.986.748-56
  - \d{3}\.\d{3}\.\d{3}-\d{2}
    - ✓ Mais conciso ?
    - ✓ E dessas maneiras 245986748-56 24598674856 ?
- CNPJ: 01.984.199/0001-07
  - \d{2}(\.\d{3}){2}\/\d{4}-\d{2}
- CPF ou CNPJ
  - \b(\d{3}\.?){2}\d{3}-?\d{2}\b|\b\d{2}\.?(\d{3}\.?){2}\/?\d{4}-?\d{2}\b
- Regex de número de telefone
  - https://medium.com/@igorrozani/criando-uma-express%C3%A3o-regular-para-telefone-fef7a8f98828

Testes online: https://regexr.com/

Encontrar os caracteres que simbolizam NA

- Objetivo: identificar caracteres que não sejam
  - Números (0 a 9), vírgula e ponto
    - ✓ -?[0-9]+(.|,)?[0-9]*
  - df_gini = pd.read_csv(path_gini, sep=';', skiprows=2, skipfooter=2, encoding='utf8', engine='python', decimal=',', dtype={"1991": "str"})

	Município	1991	2000	2010
0	110001 Alta Floresta D'Oeste	0,5983	0,5868	0,5893
1	110037 Alto Alegre dos Parecis	...	0,508	0,5491
2	110040 Alto Paraíso	...	0,6256	0,5417
3	110034 Alvorada D'Oeste	0,569	0,6534	0,5355
4	110002 Ariquemes	0,5827	0,5927	0,5496

```
1 df_gini.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5565 entries, 0 to 5564
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Município   5565 non-null   object
 1   1991        5565 non-null   object
 2   2000        5565 non-null   object
 3   2010        5565 non-null   float64
dtypes: float64(1), object(3)
memory usage: 174.0+ KB
```

Encontrar os caracteres que simbolizam NA

- Solução
  - result = df_gini['1991'].apply(
    lambda x:
    x if not re.search('(-?(([0-9]+(\.|,)?)+[0-9]*))', x) else np.nan
)

```
1  result.unique()
```

```
array([nan, '...'], dtype=object)
```

Transformação de Dados: Estudo de Caso do Autodiagnóstico da SGD

A	B	C	D	E	F	G
Qual das opç	Por favor, infc	3.1. Dados e	3.1.1.1. Relev	3.1.1.2. Prontidão Orgar	3.1.1.3. Recu	3.1.1.4. Segn
2.Área de TI (	Universidade Federal do Pa	2. INICIADO:	3. EMERGENTE: A Inst	3. EMERGEN	3. EMERGEN
2.Área de TI (	Universidade Federal de Sá	3. EMERGEN	3. EMERGENTE: A Inst	3. EMERGEN	3. EMERGEN
2.Área de TI (	Fundação Universidade Fe	3. EMERGEN	4. DESENVOLVIDO: A I	2. INICIADO:	4. DESENVC
3.Área de TI (	Centro Federal de Educaçá	1. NÃO INICIA	1. NÃO INICIADO: Parte	1. NÃO INICIA	1. NÃO INICIA
3.Área de TI (	Instituto Federal de Educaç	1. NÃO INICIA	1. NÃO INICIADO: Parte	2. INICIADO:	3. EMERGEN
2.Área de TI (	Fundação Universidade Fe	3. EMERGEN	2. INICIADO: A Instituiçá	1. NÃO INICIA	1. NÃO INICIA
1.Área de TI f	Instituto Nacional de Metro	3. EMERGEN	1. NÃO INICIADO: Parte	2. INICIADO:	3. EMERGEN
2.Área de TI (	Universidade Federal do Ri	4. DESENVC	3. EMERGENTE: A Inst	4. DESENVC	4. DESENVC
1.Área de TI f	Comissão Nacional de Ene	1. NÃO INICIA	2. INICIADO: A Instituiçá	2. INICIADO:	1. NÃO INICIA
1.Área de TI f	Fundação Nacional dos Po	1. NÃO INICIA	1. NÃO INICIADO: Parte	1. NÃO INICIA	1. NÃO INICIA
1.Área de TI f	Controladoria-Geral da Uniá	4. DESENVC	3. EMERGENTE: A Inst	4. DESENVC	2. INICIADO:
1.Área de TI f	Secretaria do Tesouro Nac	5. OTIMIZADO	3. EMERGENTE: A Inst	5. OTIMIZADO	1. NÃO INICIA
1.Área de TI f	Fundação Casa de Rui Bar	3. EMERGEN	3. EMERGENTE: A Inst	1. NÃO INICIA	1. NÃO INICIA
1.Área de TI f	Empresa Brasileira de Infra	4. DESENVC	3. EMERGENTE: A Inst	4. DESENVC	4. DESENVC

Principais Alterações

- Extrair o valor numérico das categorias
  - “1. NÃO INICIADO”, “2. INICIADO”, “3. EMERGENTE”, “4. DESENVOLVIDO”, “5. OTIMIZADO”
- Despivotar a tabela

area	orgao	Pergunta	Valor	ria_matur	iado_matu	d_pergun	ciado_per	ano
2.Área de	Universic	3.1.1.1. Re	2	INICIADO	Na Institu	3.1.1.1.	Relevânci	2023
2.Área de	Universic	3.1.1.1. Re	3	EMERGEN	Na Institu	3.1.1.1.	Relevânci	2023
2.Área de	Fundaçã	3.1.1.1. Re	3	EMERGEN	Na Institu	3.1.1.1.	Relevânci	2023
3.Área de	Centro Fe	3.1.1.1. Re	1	NAO INICI	Na Institu	3.1.1.1.	Relevânci	2023
3.Área de	Instituto	3.1.1.1. Re	1	NAO INICI	Na Institu	3.1.1.1.	Relevânci	2023
2.Área de	Fundaçã	3.1.1.1. Re	3	EMERGEN	Na Institu	3.1.1.1.	Relevânci	2023
1.Área de	Instituto	3.1.1.1. Re	3	EMERGEN	Na Institu	3.1.1.1.	Relevânci	2023
2.Área de	Universic	3.1.1.1. Re	4	DESENVOI	A Instituiç	3.1.1.1.	Relevânci	2023

Prompt como Compartilhamento de Conhecimento e Refactoring do Prompt

Processar arquivo 2024 ("/content/Autodiagnóstico 2024 Dados_Tratado_GD.xlsx"):

1. Renomear colunas especificadas por "3:" (ou seja df.columns[3:]) removendo o padrão r'(\d\.?)+\s?' do início.
2. Manter apenas o número (1 a 5) no início do texto das colunas especificadas por "3:" (ou seja df.columns[3:]), removendo o restante do texto.
3. Pivote as colunas especificadas por "3:" (ou seja df.columns[3:]) para uma coluna chamada Valor, criando o dataframe df_melted_2024.
4. Remover o padrão r'(\d\.?)+\s?' das colunas area e Pergunta.
5. Remover registros com valores nulos na coluna Valor.
6. Adicionar a coluna ano com o valor 2024.
7. Realizar merge com o arquivo /content/drive/MyDrive/empreender/ME/GovBr/Autodiagnostico/MapeamentoEixos.xlsx (contendo perguntas e eixos) e verificar se o resultado do inner join é igual ao outer join.

Prompt como Compartilhamento de Conhecimento e Refactoring do Prompt

Processar arquivo 2023 (/content/Resposta_40133199_results_survey998556.xlsx):

1. Renomear colunas especificadas por "3:" (ou seja df.columns[3:]) removendo o padrão r'(\d\.?)+\s?' do início.
2. Manter apenas o número (1 a 5) no início do texto das colunas especificadas por "3:" (ou seja df.columns[3:]), removendo o restante do texto.
3. Pivote as colunas especificadas por "3:" (ou seja df.columns[3:]) para uma coluna chamada Valor, criando o dataframe df_melted_2023.
4. Remover o padrão r'(\d\.?)+\s?' das colunas area e Pergunta.
5. Remover registros com valores nulos na coluna Valor.
6. Adicionar a coluna ano com o valor 2023.
7. Realizar merge com o arquivo /content/drive/MyDrive/empreender/ME/GovBr/Autodiagnostico/MapeamentoEixos.xlsx e verificar se alguma pergunta ficou sem eixo.

Finalizar:

1. Concatenar verticalmente df_melted_2023 e df_melted_2024.

Operação Join (Algebra Relacional)

Médico

CPF	Nome	Salario
11222731642	Jose Pereira	10
91498733332	Maria da Silva	20
81464221612	Pedro Martins	15

Professor

CPF	Nome	Salario
11222731642	Jose Pereira	6
91498733332	Maria da Silva	8
21564281600	Roberto Afonso	5

Join (ou inner join)

CPF	Nome	Salario_M	Salario_P
11222731642	Jose Pereira	10	6
91498733332	Maria da Silva	20	8

Operação Left Join (Algebra Relacional)

Médico

CPF	Nome	Salario
11222731642	Jose Pereira	10
91498733332	Maria da Silva	20
81464221612	Pedro Martins	15

Professor

CPF	Nome	Salario
11222731642	Jose Pereira	6
91498733332	Maria da Silva	8
21564281600	Roberto Afonso	5

Left Join

CPF	Nome	Salario_M	Salario_P
11222731642	Jose Pereira	10	6
91498733332	Maria da Silva	20	8
81464221612	Pedro Martins	15	--

Operação Right Join (Algebra Relacional)

Médico

CPF	Nome	Salario
11222731642	Jose Pereira	10
91498733332	Maria da Silva	20
81464221612	Pedro Martins	15

Professor

CPF	Nome	Salario
11222731642	Jose Pereira	6
91498733332	Maria da Silva	8
21564281600	Roberto Afonso	5

Right Join

CPF	Nome	Salario_M	Salario_P
11222731642	Jose Pereira	10	6
91498733332	Maria da Silva	20	8
21564281600	Roberto Afonso	--	5

Operação Outer Join (Algebra Relacional)

Médico

CPF	Nome	Salario
11222731642	Jose Pereira	10
91498733332	Maria da Silva	20
81464221612	Pedro Martins	15

Professor

CPF	Nome	Salario
11222731642	Jose Pereira	6
91498733332	Maria da Silva	8
21564281600	Roberto Afonso	5

Outer Join

CPF	Nome	Salario_M	Salario_P
11222731642	Jose Pereira	10	6
91498733332	Maria da Silva	20	8
21564281600	Roberto Afonso	--	5
81464221612	Pedro Martins	15	--

join (fundir/juntar)

- Faz o join de dois dataframes usando o índice
  - como chave de junção

```
In [70]: left2
Out[70]:
   Ohio  Nevada
a   1.0     2.0
c   3.0     4.0
e   5.0     6.0
In [71]: right2
Out[71]:
   Missouri  Alabama
b       7.0      8.0
c       9.0     10.0
d      11.0     12.0
e      13.0     14.0
```

```
In [73]: left2.join(right2, how='outer')
Out[73]:
   Ohio  Nevada  Missouri  Alabama
a   1.0     2.0       NaN      NaN
b   NaN     NaN       7.0      8.0
c   3.0     4.0       9.0     10.0
d   NaN     NaN      11.0     12.0
e   5.0     6.0      13.0     14.0
```

join (fundir/juntar)

- Com how='left' somente os registros do dataframe da esquerda
  - aparecem no resultado

```
In [70]: left2
Out[70]:
   Ohio  Nevada
a   1.0     2.0
c   3.0     4.0
e   5.0     6.0

In [71]: right2
Out[71]:
   Missouri  Alabama
b       7.0      8.0
c       9.0     10.0
d      11.0     12.0
e      13.0     14.0
```

```
left2.join(right2, how='left')
```

	Ohio	Nevada	Missouri	Alabama
a	1.0	2.0	NaN	NaN
c	3.0	4.0	9.0	10.0
e	5.0	6.0	13.0	14.0

merge (fundir/juntar)

- Semelhante ao join, mas você precisa informar a coluna de junção
  - pode ser inferida a partir do contexto da interseção entre as tabelas
    - ✓ Também pode ser especificada com o argumento **on** (Ex.: on='key')

```
In [37]: df1
Out[37]:
   data1 key
0      0   b
1      1   b
2      2   a
3      3   c
4      4   a
5      5   a
6      6   b
```

```
In [38]: df2
Out[38]:
   data2 key
0      0   a
1      1   b
2      2   d
```

```
In [39]: pd.merge(df1, df2)
Out[39]:
   data1 key  data2
0      0   b      1
1      1   b      1
2      6   b      1
3      2   a      0
4      4   a      0
5      5   a      0
```

Join vs Merge

- Ambos servem para combinar dataframes
- Join
  - Combina dataframes a partir dos seus indexes
    - ✓ Ou pode-se especificar uma coluna no dataframe onde se executa o método.
- Merge
  - Combina dataframes a partir de suas colunas
    - ✓ Pode validar o merge pelo tipo, com o argumento: validate
      - "1:1"
      - "1:m"
      - "m:1"
      - "m:m"

Maneiras de Armazenar vs Analisar os dados

Melhor para Armazenar

	Aluno	Disciplina	Objetiva	Discursiva
0	AlunoA	Portugues	8.5	6
1	AlunoA	Matematica	7.5	6.5
2	AlunoB	Geografia	9	7.5
3	AlunoB	HIstória	10	7

Melhor para Analisar

Disciplina	Geografia	HIstória	Matematica	Portugues
Aluno				
AlunoA	NaN	NaN	7.5	8.5
AlunoB	9	10	NaN	NaN

Reshaping / Pivoting (Pivotar)

- Método pivot
  - 3 argumentos: index, columns, values
    - ✓ `df.pivot(index='Aluno', columns='Disciplina', values='Objetiva')`
      - a função **`melt()`** faz a operação de despivotar

	Aluno	Disciplina	Objetiva	Discursiva
0	AlunoA	Portugues	8.5	6
1	AlunoA	Matematica	7.5	6.5
2	AlunoB	Geografia	9	7.5
3	AlunoB	HIstória	10	7

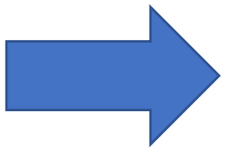
Pivotar

Disciplina	Geografia	HIstória	Matematica	Portugues
Aluno				
AlunoA	NaN	NaN	7.5	8.5
AlunoB	9	10	NaN	NaN

E quando houver valores repetidos ?

- Pivotar com o mesmo método pivot() gera exceção
  - Neste caso, use o método pivot_table
    - ✓ mean é a métrica padrão de cálculo sobre a de agregação

	Aluno	Disciplina	Objetiva	Discursiva
0	AlunoA	Portugues	8.5	6.0
1	AlunoA	Matematica	7.5	6.5
2	AlunoA	Geografia	9.0	7.5
3	AlunoA	Geografia	10.0	7.0
4	AlunoA	História	9.0	8.0
5	AlunoB	Portugues	8.5	8.5
6	AlunoB	Matematica	7.5	7.5
7	AlunoB	Geografia	9.0	9.0
8	AlunoB	História	10.0	10.0



Disciplina	Geografia	História	Matematica	Portugues
Aluno				
AlunoA	9.5	9.0	7.5	8.5
AlunoB	9.0	10.0	7.5	8.5

Reshaping / Pivoting com Índice Hierárquico

- Método stack/unstack (Pivotar com índice hierárquico)
  - stack = empilhar

state	number	
Ohio	one	0
	two	1
	three	2
Colorado	one	3
	two	4
	three	5

pivotar
unstack()

stack()
despivotar

number	one	two	three
state			
Ohio	0	1	2
Colorado	3	4	5

Prática no Colab Notebook

- Faça os exercícios da aula
  - A IA ainda não está boa para inferir os argumentos das funções do pandas que lêem arquivos e transformam num dataframe.
  - Você precisará, a priori, descobrir quais são esses argumentos e solicitar que a IA os utilize para ler os arquivos.