Senac

Fecomércio Sesc

Big Data

Prof. Marco Mialaret

Março

2024

Big Data

Onde me encontrar:

https://www.linkedin.com/in/marco-mialaret-junior/

e

https://github.com/MatmJr

Na aula passada...

Eita, já esqueci ...

- A disciplina foi apresentada.
- Aprendemos a configurar um ambiente para trabalhar com Python.

O cenário atual e as oportunidades

Big Data

Em 2016 a IBM publicou um estudo mostrando que aproximadamente 2,5 quintilhões de bytes (2,5 exabytes) de dados são criados diariamente, e que naquela época 90% dos dados do mundo foram criados nos anos de 2015 e 2016.

Segundo a International Data Corporation (IDC), o fornecimento global de dados atingirá 175 zettabytes (equivalente a 175 trilhões de gigabytes ou 175 bilhões de terabytes) anualmente até 2025.

Big Data

- Um **megabyte** é cerca de um milhão (na verdade, 2^20) de bytes. Arquivos de áudio MP3 de alta qualidade variam de 1 a 2,4 MB por minuto.

- Um **gigabyte** é cerca de 1000 megabytes (na verdade, 2^30 bytes). Equivale a aproximadamente 141 horas de áudio MP3.

- Um **terabyte** é cerca de 1000 gigabytes (na verdade, 2^40 bytes). Equivale a aproximadamente 28 anos de áudio MP3.

- Um **petabyte** é cerca de 1000 terabytes, o que equivale a aproximadamente 141 milhões de horas de áudio MP3.

- Um **exabyte** é cerca de 1000 petabytes, o que equivale a aproximadamente 141 bilhões de horas de áudio MP3.

Big Data

Hoje existem mais dispositivos IoT (Internet das coisas), do que aparelhos que não possuem essa tecnologia. O número de dispositivos conectados em Internet das Coisas (IoT) no mundo deve atingir o volume de 41,7 bilhões até o final de 2023.

Observação: Dispositivos IoT são todas e quaisquer tecnologias que possibilitam que os mais diferentes objetos se conectem à internet e interajam com ela de maneira autônoma.

Big Data

Em 2023, estima-se que cerca de 328 milhões de terabytes de dados foram gerados. Em dados mais corretos, são 330 Exabytes de dados diariamente.

Big Data

A explosão de big data provavelmente continuará exponencialmente nos próximos anos. Com 50 bilhões de dispositivos computacionais no horizonte, só podemos imaginar quantos mais haverá nas próximas décadas. É crucial para empresas, governos, militares e até indivíduos conseguirem lidar com todos esses dados.

Big Data

O apelo do big data para o grande empresariado é inegável, dada as realizações que estão acelerando rapidamente. Muitas empresas estão fazendo investimentos significativos e obtendo resultados valiosos. Isso está forçando os concorrentes a investir também, aumentando rapidamente a necessidade de profissionais de computação com experiência em ciência de dados e ciência da computação.

Big Data

Referências:

https://www.linkedin.com/pulse/o-n%C3%BAmero-de-dispositivos-conectados-em-iot-mundo-deve-atingir/?originalSubdomain=pt

O que é big data?

Big Data

Big Data é um conjunto de dados maior e mais complexo, especialmente de novas fontes de dados. Esses conjuntos de dados são tão volumosos que o software tradicional de processamento de dados simplesmente não consegue gerenciá-los. No entanto, esses grandes volumes de dados podem ser usados para resolver problemas de negócios que você não conseguiria resolver antes.

Análise de Big Data

Big Data

A análise de dados é uma disciplina acadêmica e profissional madura e bem desenvolvida. O termo "análise de dados" foi cunhado em 1962, embora as pessoas já analisassem dados usando estatísticas há milhares de anos, remontando aos antigos egípcios. A análise de big data é um fenômeno mais recente — o termo "big data" foi cunhado por volta de 2000.

Big Data

Considere quatro dos V's do big data:

1. Volume — a quantidade de dados que o mundo está produzindo está crescendo exponencialmente.

2. Velocidade — a rapidez com que esses dados estão sendo produzidos, a velocidade com que se movem pelas organizações e a rapidez com que as alterações de dados estão crescendo rapidamente.

Big Data

3. Variedade — os dados costumavam ser alfanuméricos (ou seja, consistindo de caracteres alfabéticos, dígitos, pontuação e alguns caracteres especiais) — hoje também incluem imagens, áudios, vídeos e dados de um número explosivo de sensores da Internet das Coisas em nossas casas, empresas, veículos, cidades e mais.

Big Data

4. Veracidade — a validade dos dados — eles são completos e precisos? Podemos confiar nesses dados ao tomar decisões cruciais? Eles são reais?

5. Valor – capacidade de extrair insights significativos a partir dos dados. Se refere à capacidade de transformar dados em benefícios concretos. É o processo de identificar partes de dados que são mais úteis e, assim, têm mais valor para ajudar organizações a tomar decisões mais informadas e eficazes.

Big Data

Graças ao avanço tecnológico, especialmente refletido na Lei de Moore, a capacidade de armazenar, processar e transferir esses dados se tornou econômica e eficiente, com capacidades que aumentam exponencialmente. O armazenamento digital evoluiu a ponto de ser possível manter de forma prática e acessível a vasta quantidade de dados que produzimos, fenômeno conhecido como big data.

Infraestruturas de Big Data

Big Data

Vamos discutir as infraestruturas de hardware e software populares para trabalhar com big data e desenvolvimento de aplicações de big data, tanto em desktops quanto baseadas na nuvem.

Big Data

Bancos de dados

- Bancos de dados são infraestruturas críticas para armazenar e manipular grandes volumes de dados que criamos.

- Eles são essenciais para manter esses dados de maneira segura e confidencial, especialmente com leis de privacidade rigorosas, como LGPD no Brasil, HIPAA nos EUA e GDPR na UE.

Big Data

- A maioria dos dados produzidos hoje é não estruturada, como posts do Facebook ou tweets, ou semi-estruturada, como documentos JSON e XML.

- Bancos de dados relacionais não são adequados para dados não estruturados ou semi-estruturados usados em aplicações de big data.

Big Data

- Com a evolução do big data, novos tipos de bancos de dados foram criados para lidar eficientemente com esses dados, incluindo NoSQL e NewSQL.

- Os NewSQL combinam benefícios dos bancos de dados relacionais e NoSQL.

Big Data

Apache Hadoop

- Muitos dos dados atuais são tão grandes que não cabem em um único sistema.

- Com o crescimento do big data, surgiram necessidades de armazenamento de dados distribuídos e capacidades de processamento paralelo para processar os dados mais eficientemente.

Big Data

Apache Hadoop

- Isso levou ao desenvolvimento de tecnologias complexas, como o Apache Hadoop, para processamento de dados distribuídos com paralelismo massivo em clusters de computadores, onde os detalhes intrincados são automaticamente e corretamente gerenciados.

Big Data

Apache Spark

- Apache Spark foi desenvolvido como uma solução para melhorar o desempenho do processamento de big data, executando tarefas em memória, ao contrário do Hadoop, que realiza muitas operações de I/O em disco em vários computadores.

Big Data

Apache Spark

- O Spark streaming é usado para processar dados em fluxo contínuo em mini-lotes. O Spark streaming coleta dados durante um intervalo de tempo especificado e, em seguida, fornece esse lote de dados para processamento.

Big Data

Big Data na Nuvem

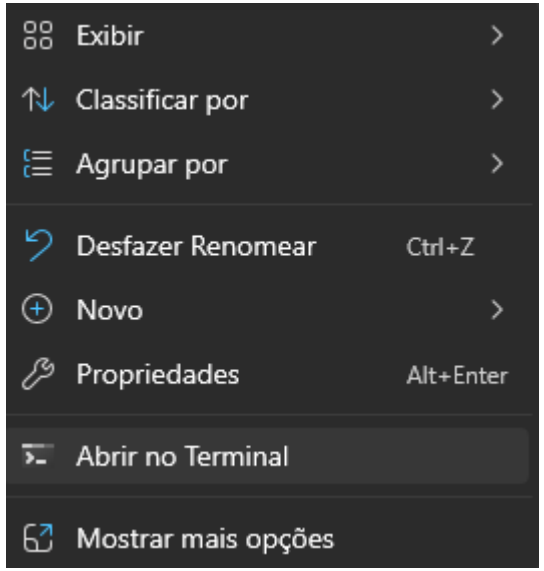
Os fornecedores de nuvem focam em tecnologia de arquitetura orientada a serviços (SOA), na qual eles fornecem capacidades "como um Serviço" que as aplicações se conectam e usam na nuvem. Serviços comuns fornecidos por fornecedores de nuvem incluem:

Big data as a Service (BDaaS)	Platform as a Service (PaaS)
Hadoop as a Service (HaaS)	Software as a Service (SaaS)
Hardware as a Service (HaaS)	Storage as a Service (SaaS)
Infrastructure as a Service (IaaS)	Spark as a Service (SaaS)

Criação do Ambiente de Trabalho

Big Data

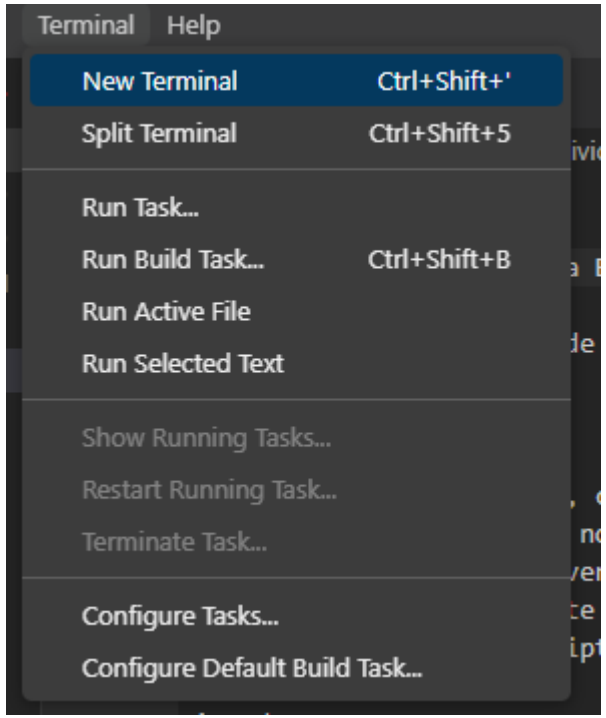
Após criar a pasta chamada BigData, clicar com o botão do lado direito e selecionar a opção abrir no terminal

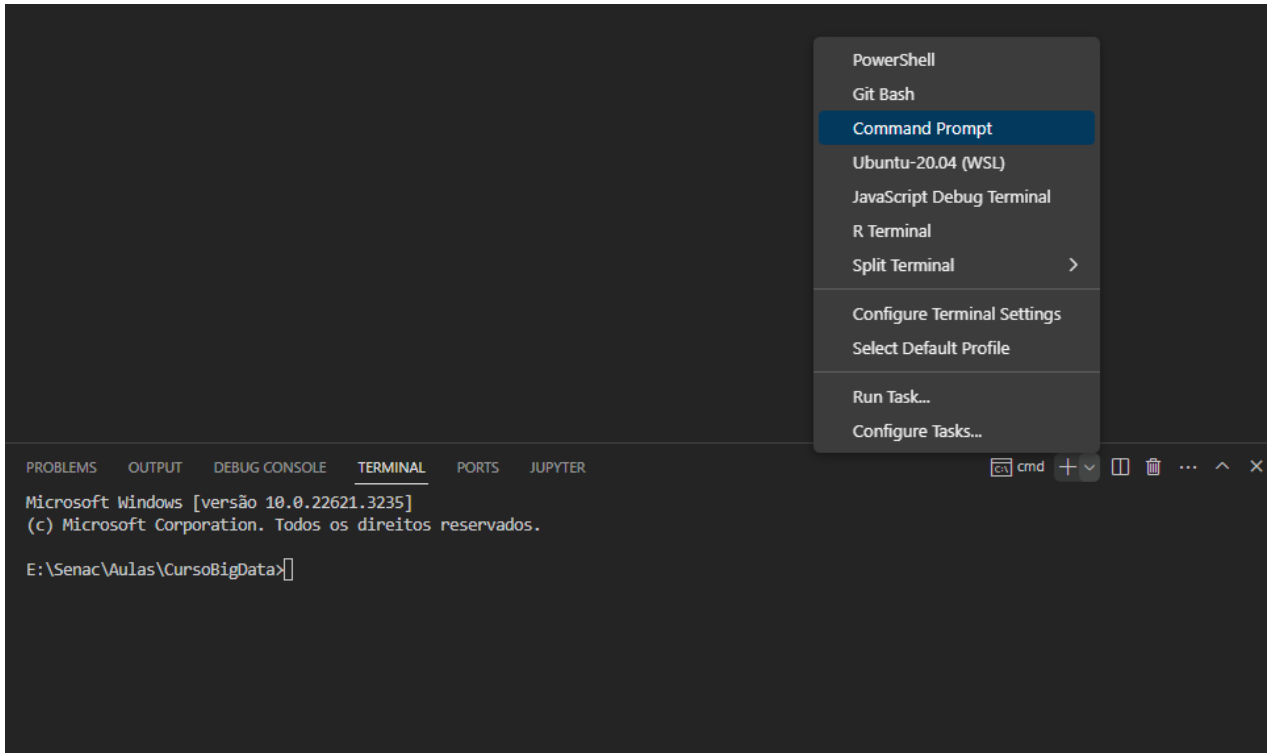



```
egamento de perfis pessoais e do sistema levou 729ms.
PS E:\Senac\Aulas\Slides\BigData\BigData> code .
```

Big Data

No VSCode:





Big Data

Criando o ambiente:

```
E:\Senac\Aulas\CursoBigData>python -m venv venv
```

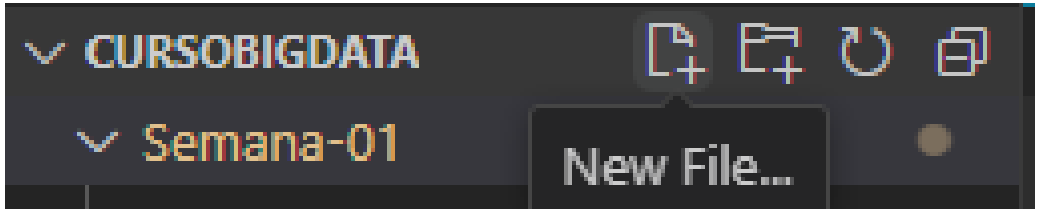
Ativando o ambiente:

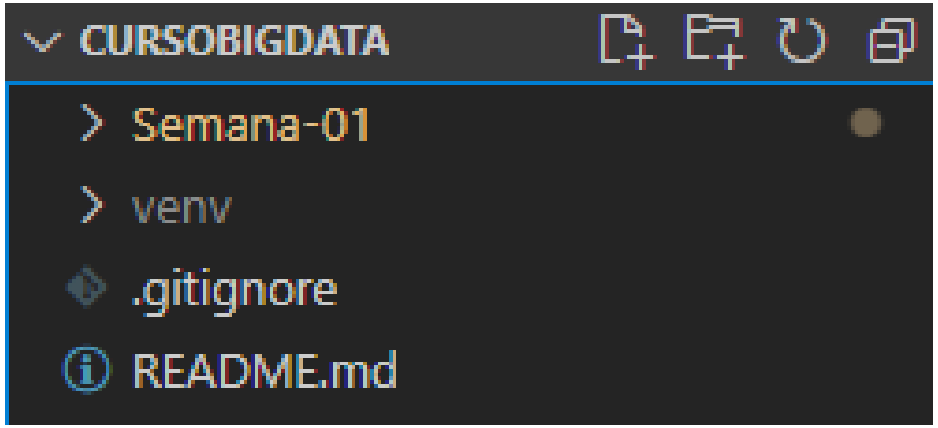
```
E:\Senac\Aulas\CursoBigData>.\venv\Scripts\activate

(venv) E:\Senac\Aulas\CursoBigData>
```

Big Data

Criação de arquivos no VSCode:





Big Data

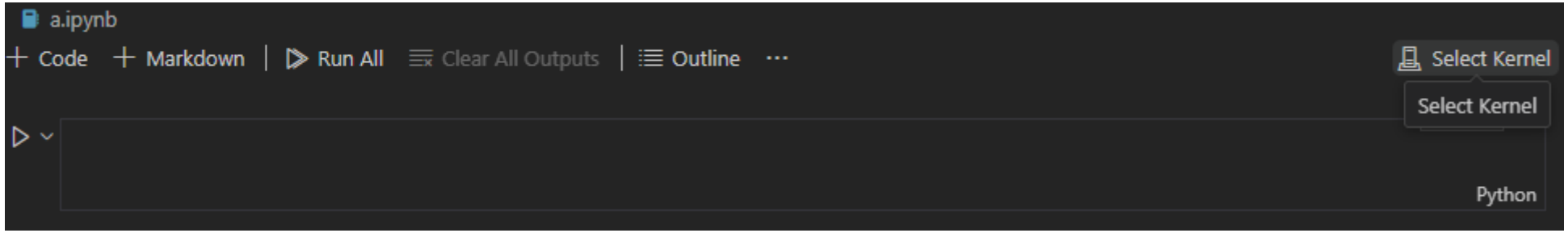
Crie um arquivo chamado aula01.ipynb.

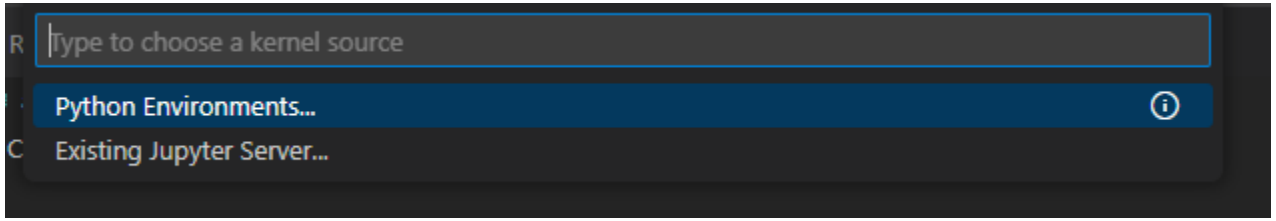
No terminal, instale o pandas no ambiente virtual:

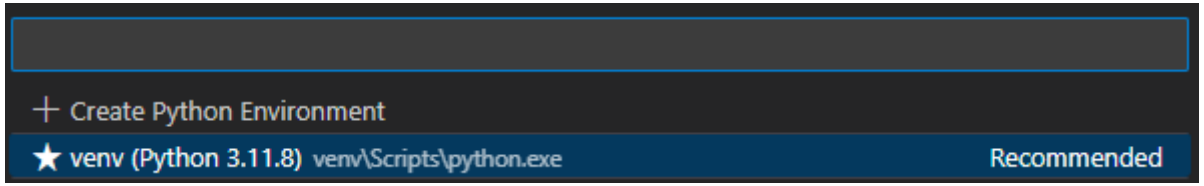
```
(venv) E:\Senac\Aulas\CursoBigData>pip install pandas
Collecting pandas
  Using cached pandas-2.2.1-cp311-cp311-win_amd64.whl.metadata (19 kB)
Collecting numpy<2,>=1.23.2 (from pandas)
  Using cached numpy-1.26.4-cp311-cp311-win_amd64.whl.metadata (61 kB)
```

Big Data

Ativando o ambiente virtual no notebook Jupyter







O conjunto de Dados

Big Data

A lavagem de dinheiro é um problema que movimenta bilhões de dólares, com sua detecção sendo notoriamente difícil devido à alta taxa de falsos positivos e negativos em algoritmos automatizados. Criminosos constantemente buscam maneiras de ocultar suas atividades.

Big Data

O acesso a dados reais de transações financeiras é fortemente restrito por questões de propriedade e privacidade, complicando a tarefa de classificar corretamente cada transação como legítima ou de lavagem.

Big Data

Para contornar esses problemas, a IBM oferece dados sintéticos de transações baseados em um mundo virtual com interações financeiras entre indivíduos, empresas e bancos, facilitando o estudo e a detecção de atividades suspeitas sem as limitações dos dados reais.

Big Data

Fonte dos dados:

https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml?resource=download

Big Data

Subi o menor arquivo no meu drive:

https://drive.google.com/file/d/1aosoxH9p2Jg3YwqtzQdC6UtWStSEc92G/view?usp=sharing

Big Data

Baixando o Arquivo e criando uma pasta chamada data:

```
import gdown
import os

url = 'https://drive.google.com/uc?id=1aosoxH9p2Jg3YwqtzQdC6UtWStSEc92G'
output = 'data/dataset.csv'

os.makedirs(os.path.dirname(output), exist_ok=True)

gdown.download(url, output, quiet=False)
```

✓ 13.1s

```
Downloading...
From (original): https://drive.google.com/uc?id=1aosoxH9p2Jg3YwqtzQdC6UtWStSEc92G
From (redirected): https://drive.usercontent.google.com/download?id=1aosoxH9p2Jg3YwqtzQdC6UtWStSEc92G&confirm=t&uuid=b72a632f-c3de-428b-af56-b45c33c8b4e3
To: e:\Senac\Aulas\CursoBigData\Semana-02\data\dataset.csv
100%|██████████| 650M/650M [00:10<00:00, 60.5MB/s]
```

Big Data

Carregando o conjunto de dados:

```
import pandas as pd

data = pd.read_csv('data/dataset.csv')
```
✓ 8.7s

	Timestamp	From Bank	Account	To Bank	Account.1	Amount Received	Receiving Currency	Amount Paid	Payment Currency	Payment Format	Is Laundering
0	2022/09/01 00:08	11	8000ECA90	11	8000ECA90	3.195403e+06	US Dollar	3.195403e+06	US Dollar	Reinvestment	0
1	2022/09/01 00:21	3402	80021DAD0	3402	80021DAD0	1.858960e+03	US Dollar	1.858960e+03	US Dollar	Reinvestment	0
2	2022/09/01 00:00	11	8000ECA90	1120	8006AA910	5.925710e+05	US Dollar	5.925710e+05	US Dollar	Cheque	0
3	2022/09/01 00:16	3814	8006AD080	3814	8006AD080	1.232000e+01	US Dollar	1.232000e+01	US Dollar	Reinvestment	0
4	2022/09/01 00:00	20	8006AD530	20	8006AD530	2.941560e+03	US Dollar	2.941560e+03	US Dollar	Reinvestment	0
...	...	...	...	...	...	...	...	...	...	...	...
6924044	2022/09/10 23:39	71696	81B2518F1	71528	81C0482E1	3.346900e-02	Bitcoin	3.346900e-02	Bitcoin	Bitcoin	0
6924045	2022/09/10 23:48	271241	81B567481	173457	81C0DA751	1.313000e-03	Bitcoin	1.313000e-03	Bitcoin	Bitcoin	0
6924046	2022/09/10 23:50	271241	81B567481	173457	81C0DA751	1.305800e-02	Bitcoin	1.305800e-02	Bitcoin	Bitcoin	0
6924047	2022/09/10 23:57	170558	81A2206B1	275798	81C1D5CA1	4.145370e-01	Bitcoin	4.145370e-01	Bitcoin	Bitcoin	0
6924048	2022/09/10 23:31	170558	81A2206B1	275798	81C1D5CA1	3.427700e-02	Bitcoin	3.427700e-02	Bitcoin	Bitcoin	0

6924049 rows × 11 columns

Big Data

Informações do Dataset:

```
data.info()
```

✓ 0.0s

Big Data

Verificação de dados faltantes:

```
data.isnull().sum()
```

```
data.isna().sum()
```

Big Data

Selecionar variáveis de interesse:

```
currency = data['Payment Currency']
received = data['Amount Received']
paid = data['Amount Paid']
```

✓ 0.0s

Big Data

Resumo estatístico

```
paid.describe()
✓ 0.3s
```

Big Data

Mudar formatação dos números

```
import locale

# Configura o locale para o padrão brasileiro
locale.setlocale(locale.LC_ALL, 'pt_BR.UTF-8')

# Ajusta a função de formatação para limitar a duas casas decimais
pd.options.display.float_format = lambda x: locale.format_string("%.2f", x, grouping=True, monetary=True)
```

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marcomialaret@gmail.com