

## COMO FIZ O PROJETO

1. Criei um pasta para o projeto
2. Criei ambiente virtual dentro da pasta com com comando "python3 -m venv tutorial-env", o tutorial-env é o nome da pasta que vai criar e também o local a ser instalado.
3. Entrei no ambiente virtual com comando "cd /tutorial-env", "source tutorial-env/bin/activate"
4. Instalei scrapy com comando "pip install Scrapy", é recomendado instalar no ambiente virtual, para não conflitar pacotes de sistema Python já instalados (o que pode quebrar algumas de suas ferramentas e scripts de sistema).
5. Instalei a biblioteca pandas com comando "pip install pandas"
6. Usei o comando "scrapy startproject tutorial" para criar uma pasta com diversos conteúdos de scrapy.
7. Criei arquivo.py em webscraping/spiders/ para criar uma spider para raspar os dados, editei, achei conteúdos na internet, no youtube.

Consegui quase todos os dados que pediram, 5 das 6, mas peguei outros dados para melhorar e consegui analisar os dados com a biblioteca pandas, lendo o csv, usando seus comandos e obtendo as informações.

## BREVE ANÁLISE

Eu achei a biblioteca scrapy, ouvi pessoas falar bem dele na internet e foi uma das primeiras que encontrei. Com pouco memória ram e processador do notebook que usei eu consegui os dados fácil. Eu poderia ter utilizado o meu notebook mais potente, mas consegui com esse.

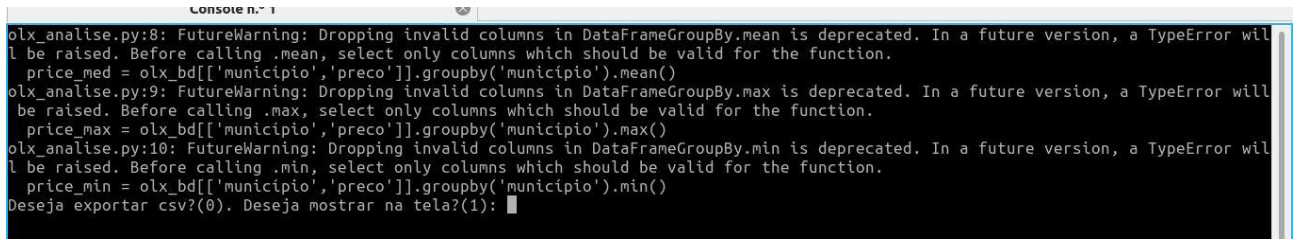
Eu poderia criar um código mais fácil, pegando informações só da primeira página, mas teria menos informações. Então, criei um código que pega as urls na página principal e depois entrava um por um pegando todas informações possíveis. Usei o "replace" para retirar "R\$" e o "m²", pois confundiria na hora da análise. Achei os comandos interessantes: o contais e o following, para achar o irmão e pegar o conteúdo só das tags irmãs. Assim, consegui a maioria dos dados.

Infelizmente não consegui o nome do anunciante, tentei com outras bibliotecas, mas sem sucesso.

O scrapy é muito bom, mas é preciso combinar com outras bibliotecas como selenium e splash para obter dados de páginas dinâmicas que eu não usei no código.

É preciso entender mais sobre as páginas dinâmicas, pode ser que eu consiga pegar o nome do anunciante e outros dados se possível futuramente.

Obs.: No final, meu código de análise de dados estava dando um bug esquisito que quando editava o csv ele funcionava. Como o acento no "título", ou na cedilha no "preço". Coloquei uma imagem abaixo do bug, mas consegui contornar tirando o acento em título.

A screenshot of a terminal window titled "Console n.º 1". It displays three lines of Python code and their corresponding warnings. The first line is `price_med = olx_bd[['municipio', 'preco']].groupby('municipio').mean()`, followed by a warning: "FutureWarning: Dropping invalid columns in DataFrameGroupBy.mean is deprecated. In a future version, a TypeError will be raised. Before calling .mean, select only columns which should be valid for the function." The second line is `price_max = olx_bd[['municipio', 'preco']].groupby('municipio').max()`, followed by a similar warning for the `.max` method. The third line is `price_min = olx_bd[['municipio', 'preco']].groupby('municipio').min()`, followed by a warning for the `.min` method. At the bottom, there is a prompt "Deseja exportar csv?(0). Deseja mostrar na tela?(1):" with a cursor pointing to it.

```
Console n.º 1
olx_analise.py:8: FutureWarning: Dropping invalid columns in DataFrameGroupBy.mean is deprecated. In a future version, a TypeError will
be raised. Before calling .mean, select only columns which should be valid for the function.
  price_med = olx_bd[['municipio', 'preco']].groupby('municipio').mean()
olx_analise.py:9: FutureWarning: Dropping invalid columns in DataFrameGroupBy.max is deprecated. In a future version, a TypeError will
be raised. Before calling .max, select only columns which should be valid for the function.
  price_max = olx_bd[['municipio', 'preco']].groupby('municipio').max()
olx_analise.py:10: FutureWarning: Dropping invalid columns in DataFrameGroupBy.min is deprecated. In a future version, a TypeError wil
l be raised. Before calling .min, select only columns which should be valid for the function.
  price_min = olx_bd[['municipio', 'preco']].groupby('municipio').min()
Deseja exportar csv?(0). Deseja mostrar na tela?(1):
```

## FEEDBACK

Esse foi meu primeiro desafio, achei muito interessante a linguagem python, pois é fácil de usar, utilizando as bibliotecas as coisa ficam mais fáceis, se fosse em "c", teria muitos loops, nem sei se iria conseguir. Aprendi muito com o desafio da Seazone, e esse é meu objetivo, e espero que consiga agradá-los.