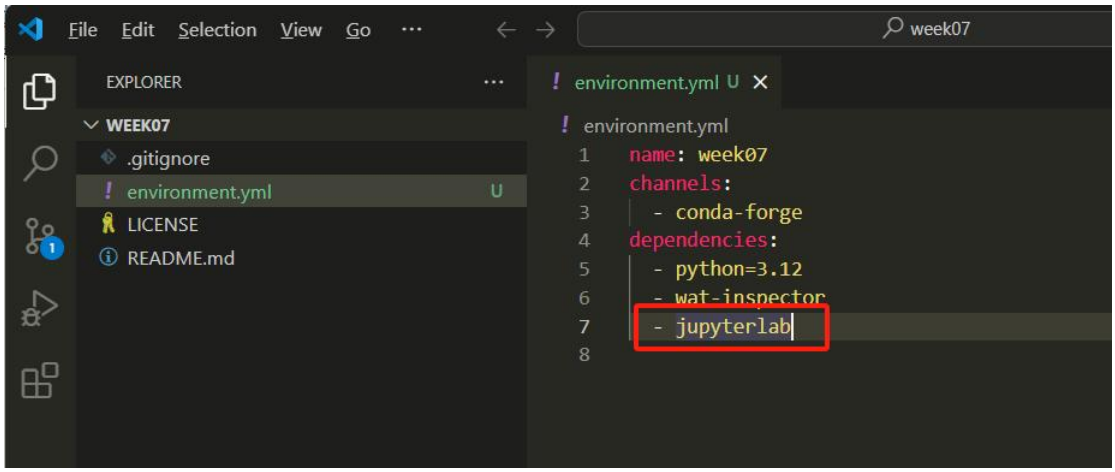
任务目标

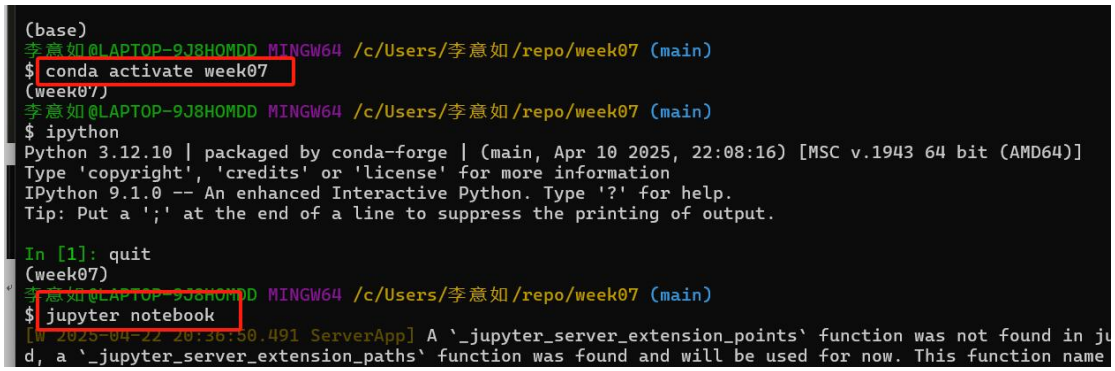
编程的目标是实现 **自动化**，但是就像我们需要经过 **调试** 一步一步写出代码那样，我们也需要 **可视化** 与 **交互** 工具来查看、检查、监控、探索、分析、理解数据，在 “手动” 的基础上实现 “自动”。终端 (TUI) 是自动化的利器，但在可视化与交互方面，终端确实不太擅长。图形用户界面 (GUI) 比较擅长可视化与交互，但不够跨平台 (指在不同操作系统上获得相似的体验)，也不够开放 (指不容易定制和扩展功能)。基于浏览器 (Browser, 比如 Chrome、Edge、Firefox、Safari 等等) 的 **Web 技术** (比如网页版 App)，不但擅长可视化与交互，而且跨平台和足够开放，是特别理想的选择，但缺点是技术栈 (HTML + CSS + Javascript) 门槛较高，不适合初学者。

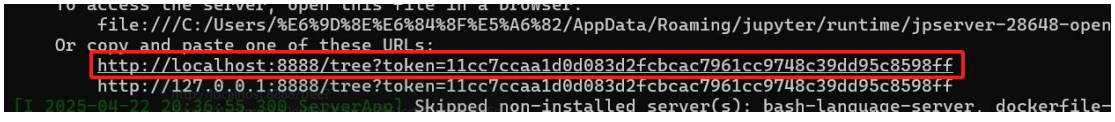
本周将通过案例介绍两个 Python 生态 (PyPI) 下适合初学者使用的可视化与交互工具 (全部是基于 Web 技术)，分别是 JupyterLab 和 Perspective。案例使用的数据将通过 Tushare 获取。

2. 用 VS Code 打开项目目录，新建一个 `environment.yml` 文件，指定安装 Python 3.12 和 `jupyterlab`，然后运行 `conda env create` 命令创建 Conda 环境

3. 在项目目录下，运行 `jupyter lab` 命令，启动 **后端** (Backend) 服务，在浏览器里粘贴地址访问 **前端** (Frontend) 页面



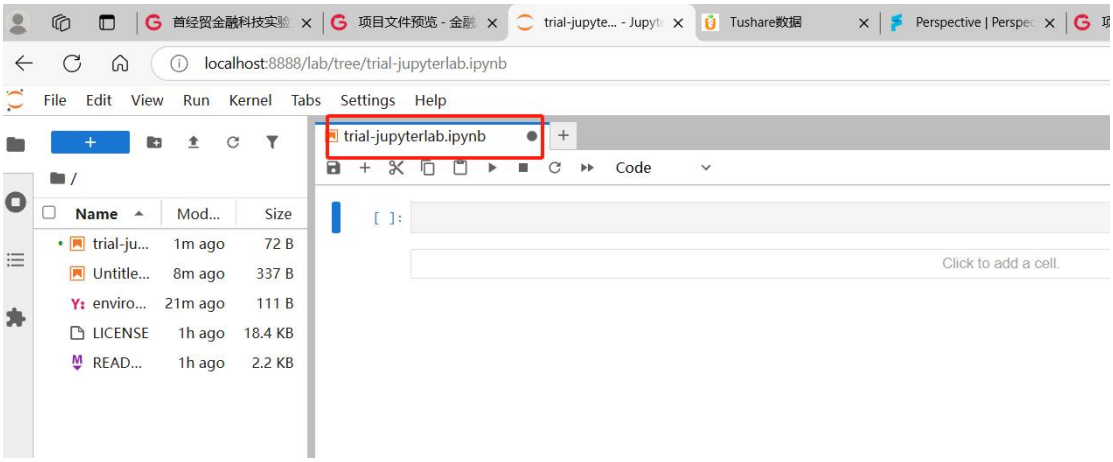




复制到浏览器

ipython——jupyter notebook——jupyter lab 发展历史

rename



- 在单元格 (Cell) 里编写 Python 代码，按 `Shift+Enter` 运行 Cell 并下移
- 在单元格 (Cell) 上按 `ESC` 切换到 **命令模式** (command mode)，按 `Enter` 切换到 **编写模式** (edit mode)
- 在单元格 (Cell) 的命令模式下，按 `j` 选择下一个，按 `k` 选择上一个，按 `a` 在上方添加，按 `b` 在下方添加，按 `dd` 删除，按住 `Shift` 多选，按 `x` 剪切，按 `c` 复制，按 `v` 粘贴，按 `Shift+M` 合并，按 `z` 撤销，按 `Shift+Z` 重做，按 `Shift+L` 显示/隐藏代码行号
- 在单元格 (Cell) 的编写模式下，按 `Ctrl+Shift+-` 切分单元格

- 按按钮显示/隐藏 Minimap
- 运行单元格 (Cell) 注意序号单调递增
- 单元格最后一行如果是 **表达式** (expression) 且运行后返回的对象不是 `None`，则计输出 (Out)，否则只计输入 (In)，序号为 `i` 的输出，可以用 `_i` 变量来引用
- 单元格 (Cell) 序号为 `*` 表示代码运行中，尚未返回，按 `ii` 可以打断 (`KeyboardInterrupt`) (类似于终端的 `Ctrl+C`)
- 在单元格 (Cell) 的命令模式下，按 `00` 重启后端 Python 解释器 (被 Jupyter 称为 Kernel)，重启后需要从上至下重新运行一遍代码 (`Shift+Enter`)，运行前建议先在菜单里选择 “Edit / Clear Outputs of All Cells” 清空全部页面显示的输出
- 在单元格 (Cell) 的命令模式下，按 `m` 切换至 **Markdown 模式**，按 `y` 切换至 **Python 模式**
- 用豆包 (或 DeepSeek 等任何大模型) 生成一段示例 Markdown 代码，复制粘贴进 Markdown 单元格，运行以呈现 (Render)
- 用豆包 (或 DeepSeek 等任何大模型) 生成一段示例 HTML 代码，复制粘贴进 Markdown 单元格，运行以呈现 (Render)；注意不支持 CSS
- 用豆包 (或 DeepSeek 等任何大模型) 生成一段示例 LaTeX 数学公式代码，复制粘贴进 Markdown 单元格，运行以呈现 (Render)；注意要用 `$` (行内模式) 或 `$$` (整行模式) 包围
- 关闭前端页面，在后端按 `Ctrl+C` 打断运行中的服务，回到 Bash 提示符

可以用 jupyter markdow 模式转换 ai 生成的内容，转化为有格式文本，好用

5. 通过 `tushare` 软件包下载保存一些数据：

   - 在 Tushare 网站上 注册 并登陆，完善修改个人资料，浏览阅读 平台介绍 和 数据接口

     > 通过定制的推荐链接 `https://tushare.pro/register?reg=gls` 完成注册，将可获赠 2000 平台积分 (有效期一年)。积分达到门槛才有数据接口的使用权限，否则需要 付费购买积分 (约 200~1000 元/年) 才有权限使用数据接口。本课程的量化投资实战案例，将主要将通过 Tushare 平台获取数据，请确保拥有足够积分进行实践。

   - 修改 `environment.yml` 文件，添加 `pip: tushare` (注意，`conda-forge` 没有收录 `tushare`，只能从 `PyPI` 安装，参考) 依赖项，运行 `conda env update` 更新 Conda 环境

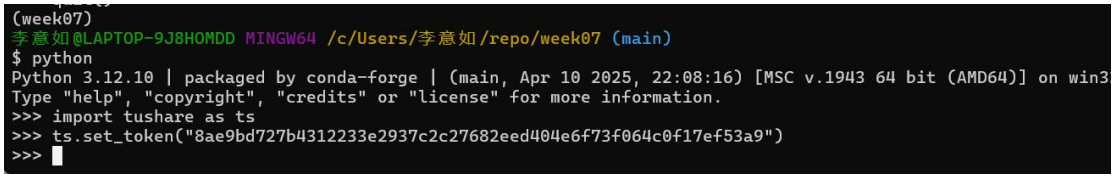
   - 在终端 (Terminal) 激活 `week07` Conda 环境，运行 `ipython` 命令启动 IPython 交互界面 (IPython 是 Jupyter 项目的一部份，`ipython` 是 `jupyterlab` 的依赖项之一)

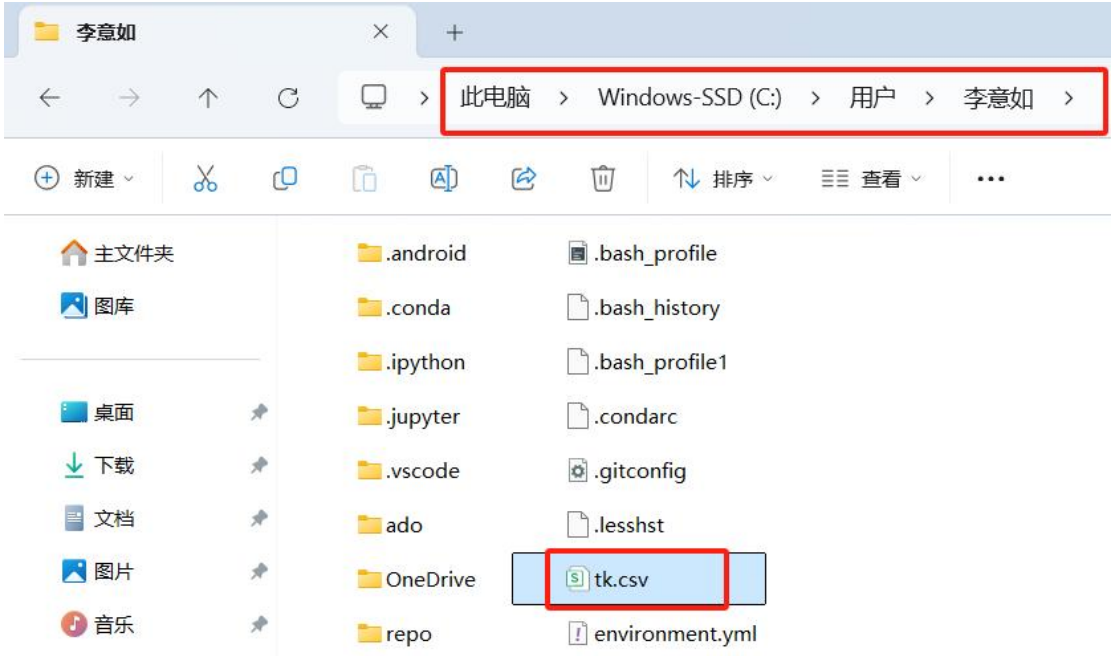
- 在 IPython 提示符下，运行下面的 Python 代码设置 Tushare Token

```
import tushare as ts

ts.set_token("***")  # 将 *** 修改成你自己的 Token 字符串
```

其中 `***` 要替换成你在 Tushare 平台上的 接口TOKEN —— 复制粘贴即可。运行 `set_token` 函数会把 Token 字符串保存在 `~/tk.csv` 文件里，今后每次使用 `tushare` 软件包请求数据时都会自动读取并发送 Token，不需要反复设置。





- 按 `Ctrl+D` 结束前面的 IPython 进程，再重新启动一个新的 IPython 进程，运行下面的 Python 代码向 Tushare 服务器请求 IPO新股列表 数据，并保存在本地

```
import tushare as ts

pro = ts.pro_api()
df = pro.new_share()
df.to_parquet("new_share.parquet")
```

其中请求数据函数返回的对象 `df` 是 `pandas.DataFrame` 类型，调用其 `to_parquet` 方法能够将内存 (memory) 中的 `DataFrame` 数据按照 Parquet 格式 (Parquet 是大数据领域的首选格式，已经成为业界标准) 序列化 (serialize) 为字节串 (bytes) 保存到磁盘 (disk)。

```
In [1]: import tushare as ts

In [2]: pro = ts.pro_api()

In [3]: type pro
  Cell In[3], line 1
    type pro
         ^
SyntaxError: invalid syntax


In [4]: type(pro)
Out[4]: tushare.pro.client.DataApi

In [5]: id(pro)
Out[5]: 2874156814016

In [6]: pro
Out[6]: <tushare.pro.client.DataApi at 0x29d31188ec0>

In [7]: pro.new_share()
Out[7]:
        ts_code sub_code  name  ipo_date issue_date   amount  market_amount  price     pe  limit_amount   funds  ballot
0     603014.SH   732014  威高血净  20250508       None   4114.0            0.0   0.00   0.00          1.10   0.000    0.00
1     301595.SZ   301595  太力科技  20250508       None   2707.0            0.0   0.00   0.00          0.65   0.000    0.00
2     688755.SH   787755  汉邦科技  20250507       None   2200.0            0.0   0.00   0.00          0.50   0.000    0.00
3     301636.SZ   301636  泽润新能  20250428       None   1597.0          774.0  33.06  17.57          0.45   5.279    0.02
4     920068.BJ   920068  天工股份  20250428       None   6000.0         4200.0   3.94  14.98        255.00   2.364    0.00
...         ...      ...   ...       ...        ...      ...            ...    ...    ...           ...     ...     ...
1995  300777.SZ   300777  中简科技  20190506   20190516   4001.0         3601.0   6.06  22.98          1.10   2.425    0.04
1996  603267.SH   732267  鸿远电子  20190430   20190515   4134.0         3721.0  20.24  16.50          1.60   8.367    0.03
1997  600989.SH   730989  宝丰能源  20190430   20190516  73336.0        66002.0  11.12  22.07         22.00  81.550    0.25
1998  300778.SZ   300778   新城市  20190425   20190510   2000.0         2000.0  27.33  22.99          2.00   5.466    0.02
1999  002953.SZ   002953  日丰股份  20190424   20190509   4302.0         3872.0  10.52  16.34          1.70   4.526    0.03

[2000 rows x 12 columns]
```

- 询问豆包 (或 DeepSeek 等任何大模型)，初步了解 Parquet 格式和 CSV 格式的特点和适用领域

Parquet 格式

特点

- **列式存储**：区别于行式存储，数据按列组织。查询时仅读取所需列，减少 I/O 操作，提升查询效率，尤其对特定列频繁查询场景。如分析电商订单数据，仅需查询商品价格列时，Parquet 可快速定位读取，无需扫描整行。
- **自带模式定义**：嵌入模式（Schema），定义数据结构、类型及关系。可用于数据验证，确保数据符合预定类型，降低处理错误；且具自描述性，工具和框架易解读文件，无需额外模式定义。
- **压缩效率高**：支持多种压缩算法，如 Snappy、Gzip、Z Standard 等。结合列式存储，数据相似性高，压缩效果好，减小文件体积，降低存储成本，加快网络传输速度。
- **支持复杂数据类型**：能处理嵌套数据结构，如结构体、数组、Map 等。适合存储 Web 分析的类似 JSON 结构，或物联网设备传感器数据中的嵌套数组。
- **谓词下推**：利用元数据和索引，读取时跳过不必要数据，确定与查询相关文件部分，提升查询性能，减少资源消耗。

适用领域

- **大数据处理与分析**：与 Hadoop、Spark 等大数据框架紧密结合，在处理海量数据时，高效的查询和存储性能优势明显。如企业对大量用户行为数据进行分析挖掘。
- **数据仓库**：为数据仓库提供快速查询和分析能力，存储结构化数据，支撑复杂报表和决策分析。
- **机器学习**：训练数据量通常较大，Parquet 利于存储和快速读取特定属性数据，满足机器学习数据处理需求。

CSV 格式

特点

- **简单易读**：纯文本格式，以逗号（或其他分隔符）分隔字段，直观呈现表格数据，用文本编辑器或常见办公软件（如 Excel）易打开查看和编辑。
- **广泛兼容**：几乎所有数据处理工具、编程语言、数据库都支持，方便数据在不同系统和平台间交换传输。
- **无模式信息**：自身不携带数据类型等模式信息，需外部定义或约定数据结构和类型。
- **缺乏压缩和结构化**：一般无内置压缩机制（需借助外部工具压缩），且难以处理复杂结构化数据，不支持嵌套数据结构。

适用领域

- **小规模数据存储与交换**：数据量小、结构简单时，CSV 易读易处理，适合存储交换。如小型企业记录简单业务数据，或个人记录收支明细。
- **数据传输**：因兼容性好，在不同系统间传输数据时常用，如从数据库导出数据给第三方应用。
- **开发测试**：开发和测试阶段，方便快速生成、查看和修改数据，调试程序。

- `new_share` 接口只需要 120 积分，如果你有 2000 积分，可以采用与上面类似的方法访问 `stock_basic` 接口，并将数据保存为 `stock_basic.parquet` 文件 (注意，需要指定 `fields` 参数获取全部字段)。如果积分暂时不够，可以在终端运行下面的命令，从我们开源的 课程仓库 下载数据文件到你的本地

```
curl -O https://raw.gitcode.com/cueb-fintech/courses/blobs/8fc08f7bc4dbbf17d3562344727795e5
```

通过 `perspective-python` 软件包查看 `polars.DataFrame` 数据，实践交互式可视化：

- 修改 `environment.yml` 文件，添加 `perspective-python` 和 `polars` 依赖项，运行 `conda env update` 更新 Conda 环境
- 启动 JupyterLab，新建一个 Notebook，改名为 `trial-perspective.ipynb`

- 调用 `polars.read_parquet` 函数，分别读取磁盘 (disk) 中的 `new_share.parquet` 文件和 `stock_basic.parquet` 文件，得到内存 (memory) 中的 `polars.DataFrame` 对象，命名为 `d1` 和 `d2`

```
[6]: d1 = pl.read_parquet("new_share.parquet")
     d1
```

[6]: shape: (5_412, 17)

ts_code	symbol	name	area	industry	fullname	enname	cnspell	market	exchange	curr_type	list_status	list_date	delist_date	is_hs	act_name	act_ent_type
str	str	str	str	str	str	str	str	str	str	str	str	str	null	str	str	str
"000001.SZ"	"000001"	"平安银行"	"深圳"	"银行"	"平安银行股份有限公司"	"Ping An Bank Co., Ltd."	"payh"	"主板"	"SZSE"	"CNY"	"L"	"19910403"	null	"S"	"无实际控制人"	"无"
"000002.SZ"	"000002"	"万科A"	"深圳"	"全国地产"	"万科企业股份有限公司"	"China Vanke Co.,Ltd."	"wka"	"主板"	"SZSE"	"CNY"	"L"	"19910129"	null	"S"	"深圳市人民政府国有资产监督管理委员会"	"地方国企"
"000004.SZ"	"000004"	"国华网安"	"深圳"	"软件服务"	"深圳国华网安科技股份有限公司"	"Shenzhen Guohua Network Securi..."	"ghwa"	"主板"	"SZSE"	"CNY"	"L"	"19910114"	null	"N"	"李映彤"	"民营企业"
"000006.SZ"	"000006"	"深振业A"	"深圳"	"区域地产"	"深圳市振业(集团)股份有限公司"	"Shenzhen Zhenye(Group) Co., Lt..."	"szya"	"主板"	"SZSE"	"CNY"	"L"	"19920427"	null	"S"	"深圳市人民政府国有资产监督管理委员会"	"地方国企"
"000007.SZ"	"000007"	"全新好"	"深圳"	"其他商业"	"深圳市全新好股份有限公司"	"Shenzhen Quanxinhao Co.,Ltd."	"qxh"	"主板"	"SZSE"	"CNY"	"L"	"19920413"	null	"N"	"王玩虹"	"民营企业"
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
"920111.BJ"	"920111"	"贝星科技"	null	null	"温州贝星科技股份有限公司"	"Wenzhou Juxing Science And Tec..."	"jxkj"	"北交所"	"BSE"	"CNY"	"L"	"20241111"	null	"N"	null	null
"920116.BJ"	"920116"	"星图测控"	null	null	"中科星图测控技术股份有限公司"	"Geovis Insighter Technology Co..."	"xtck"	"北交所"	"BSE"	"CNY"	"L"	"20250102"	null	"N"	null	null
"920118.BJ"	"920118"	"太湖远大"	null	null	"浙江太湖远大新材料股份有限公司"	"Zhejiang Taihu Yuanda New Mate..."	"thyd"	"北交所"	"BSE"	"CNY"	"L"	"20240822"	null	"N"	null	null
"920128.BJ"	"920128"	"胜业电气"	null	null	"胜业电气股份有限公司"	"Shengye Electric Co., Ltd."	"sydq"	"北交所"	"BSE"	"CNY"	"L"	"20241129"	null	"N"	null	null
"689009.SH"	"689009"	"九号公司-WD"	"北京"	"摩托车"	"九号有限公司"	"Ninebot Limited"	"jhgs"	"科创板"	"SSE"	"CNY"	"L"	"20201029"	null	null	null	null

```
[7]: d2 = pl.read_parquet("stock_basic.parquet")
     d2
```

[7]: shape: (5_412, 17)

ts_code	symbol	name	area	industry	fullname	enname	cnspell	market	exchange	curr_type	list_status	list_date	delist_date	is_hs	act_name	act_ent_type