第 8 周 Frank 学习笔记

一、想法 & 准备

为什么用 Polars？

Pandas 虽好，但面对大数据它会吃内存、慢半拍。Polars 基于 Arrow + Rust，速度快、占内存少，咱新手直接上手它就行啦。

搭环境

1. Fork 08 周仓库，`git clone` 到本地

2. 写个 `environment.yml`，装 Polars、PyArrow、xlrd、openpyxl、Perspective、Tushare 等

3. `conda env create -f environment.yml && conda activate week08`

拿数据

```bash

curl -O https://.../stock\_trades.zip

unzip stock\_trades.zip

```

两家券商、2022.7–2023.10，每月一份交割单（.xls/.xlsx）。

---

二、搞清文件格式

湘财的“伪 Excel”

用 `polars.read\_excel()` 报错，其实它是个 GB18030 编码、TAB 分隔的纯文本。

VS Code 右下角切到 GB18030，就能看到汉字了。

怎么读？

```python

df = pl.read\_csv(

"202207-湘财.xls",

separator="\t",

encoding="gb18030",

infer\_schema=False

)

```

三、套路化清洗

1. 先当字符串全读进来（省得类型猜错）

2. 把“发生日期”转日期、“成交时间”转时间

```python

df = df.with\_columns([

pl.col("发生日期").str.strptime(pl.Date, "%Y-%m-%d"),

pl.col("成交时间").str.strptime(pl.Time, "%H:%M:%S")

])

```

3. 去掉证券代码外面的`="000900"`

```python

df = df.with\_columns(

pl.col("证券代码").str.replace\_all(r'="?(\d+)"?', r'\1')

)

```

4. 只留“证券买入”和“证券卖出”

```python

df = df.filter(pl.col("业务名称").is\_in(["证券买入","证券卖出"]))

```

5. 数值列统一转 `Float64`

```python

for col in ["成交数量","成交价格","成交金额","手续费",...]:

df = df.with\_column(pl.col(col).cast(pl.Float64))

```

批量搞定所有“湘财”文件

```python

from pathlib import Path

files = Path().glob("-湘财.xls")

d1 = pl.concat([

pl.read\_csv(f, sep="\t", encoding="gb18030", infer\_schema=False)

.pipe(上面那些清洗步骤)

.with\_column(pl.lit("湘财").alias("券商"))

for f in files

])

```

海通普通 & 两融 同理，拿到 `d2`、`d3`，分别加上券商字段。

---

四、合并 & 导出

统一字段：券商、交易日期、交易时间、证券代码、证券名称、买卖标志、成交价格、成交数量（取绝对值）、成交金额、手续费、印花税、过户费、其他费、发生金额。

合并：`df = pl.concat([d1, d2, d3])`

检查行数： `df.height` 应该是 363 行

存成 Parquet：`df.write\_parquet("stock\_trades.parquet")`

---

五、简单分析 & 可视化思路

1. 费率

手续费率 = 手续费 / 成交金额；印花税率、过户费率同理。

用 Perspective 的散点图（X：交易序号，Y：费率），不同券商、买/卖用颜色区分。

2. 持仓平仓

先把“买入”数量正，“卖出”数值反，分组求和，得到“结余数量”。

结余数量不为零的，就是期末还没平仓的。

3. 日度持仓动态

搞个「日期 × 证券」的笛卡尔积，左 join 交易数据，空值补 0，做累加（`cum\_sum`）。

每天持仓数量 & 持股支数，用 Perspective 面积图。

4. 市值 & 收益

Tushare 拉日线行情，检查成交价在最高-最低价区间内。

用持仓数量×收盘价算市值，折线图看动态。

加上“现金余额”（首日 100 万滚动累加），得到“总资产”。

比较沪深 300 指数累计收益，做对照。

5. 长宽表转换

多条折线要一起画，得把宽表 `{"总资产","沪深300"}` 用 `melt` 变成长表。

---

> Tips：

>

> 金融数据务必手算校验，别全信黑箱。

> Polars 那些 Expression 写多了就熟练了，一行链式调用爽到飞起。