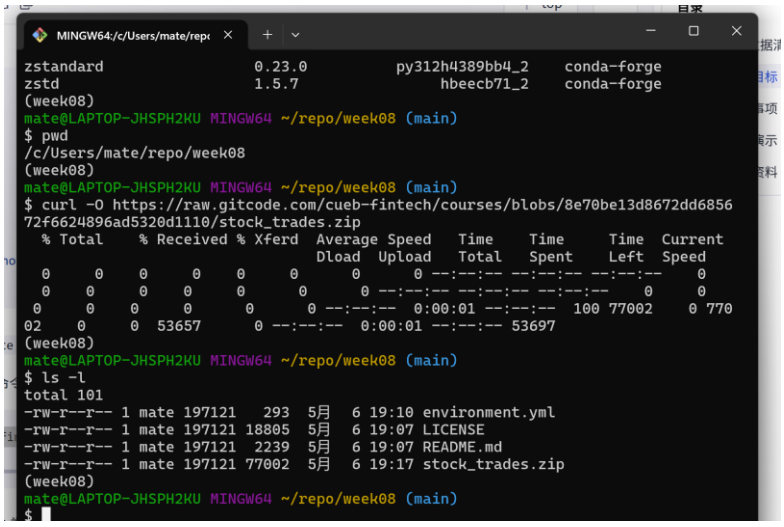
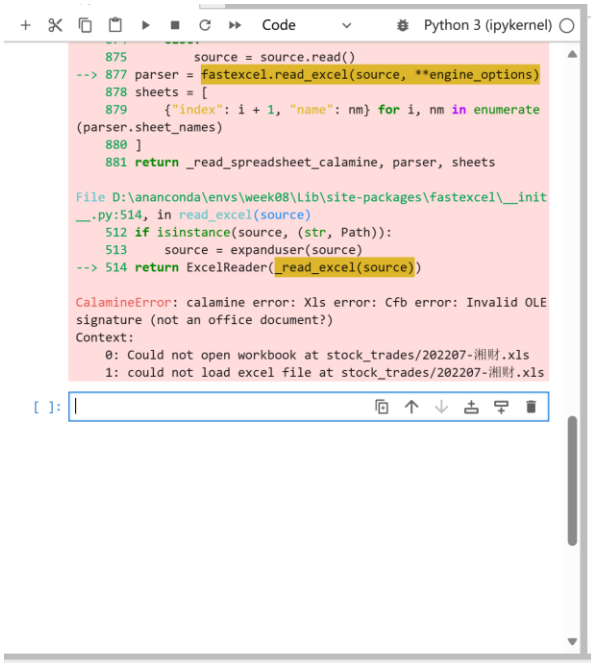
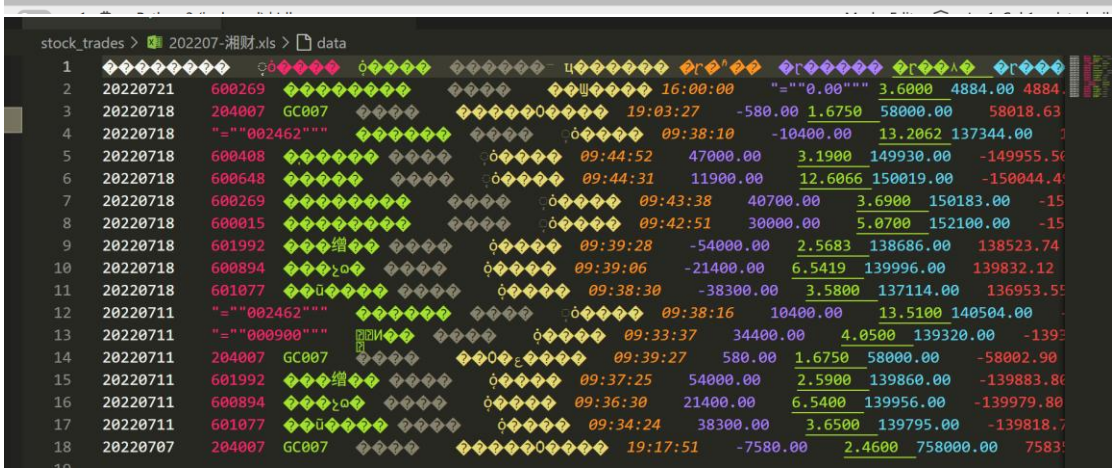
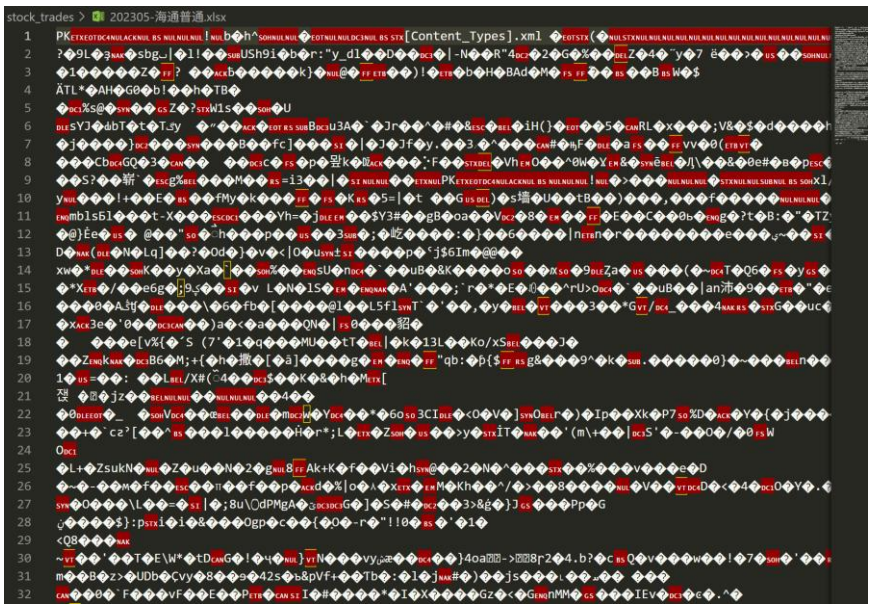
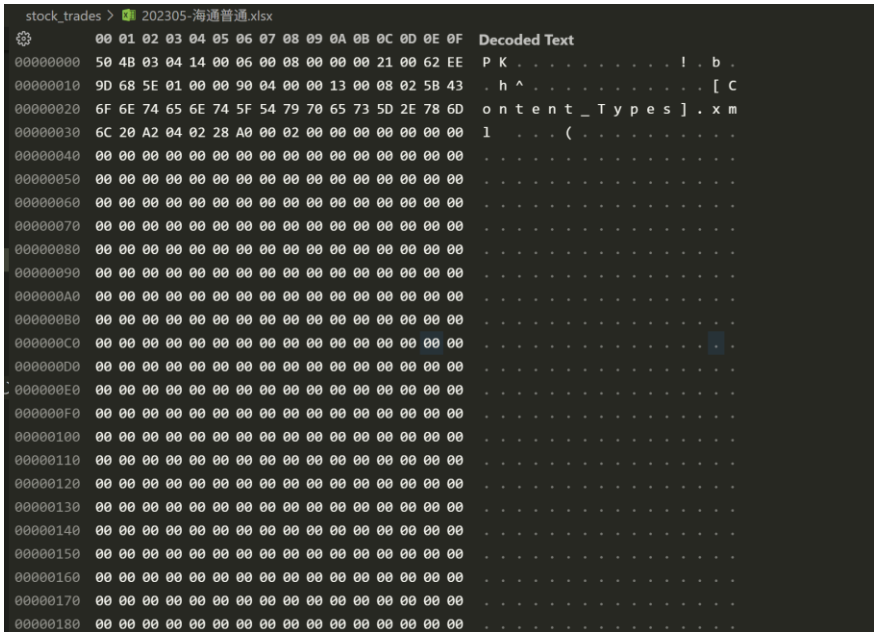
第 8 周 数据清洗与计算（初级）











stock_trades > 202211-湘财.xls > data

	发生日期	证券代码	证券名称	买卖标志	业务名称	成交时间	成交数量	成交价格	成交金额	发生金
2	20221128	300385	雪浪环境	买入	证券买入	13:14:01	11200.00	6.2888	70435.00	-70446.27
3	20221128	603686	福龙马	买入	证券买入	13:11:59	7800.00	9.0815	70836.00	-70848.04	11.33
4	20221128	603186	华正新材	卖出	证券卖出	13:10:42	-3400.00	20.9059	71080.00	70996.85
5	20221128	688165	埃夫特智能	卖出	证券卖出	13:10:06	-4331.00	8.2000	35514.20	35472.65
6	20221122	300422	博世科	买入	证券买入	10:29:31	12300.00	6.0187	74030.00	-74041.84	11.8
7	20221122	688165	埃夫特智能	卖出	证券卖出	10:21:33	-4331.00	9.0900	39368.79	39322.73
8	20221110	300368	汇金股份	卖出	证券卖出	10:08:05	-11200.00	6.5200	73024.00	72939.31
9	20221110	603186	华正新材	买入	证券买入	10:09:55	3400.00	20.9362	71183.00	-71195.10	11.3
10	20221108	300429	强力新材	买入	证券买入	09:49:53	8300.00	8.3500	69305.00	-69316.09	11.0
11	20221108	"="000096"""	广聚能源	卖出	证券卖出	09:48:30	-8500.00	8.2000	69700.00	6961
12	20221108	"="002228"""	合兴包装	卖出	证券卖出	09:37:15	-22600.00	3.1900	72094.00	7201
13	20221108	688165	埃夫特智能	买入	证券买入	09:44:44	8662.00	8.4480	73176.30	-73188.74	11.7
14	20221108	600800	渤海化学	卖出	证券卖出	09:43:19	-19300.00	3.7900	73147.00	73061.42
15	20221108	688393	安必平医药	买入	证券买入	09:41:27	970.00	22.9100	22222.70	-22226.48	3.56
16	20221108	688393	安必平医药	买入	证券买入	09:40:35	2200.00	22.7991	50158.00	-50166.54	8.03
17	20221108	600829	人民同泰	卖出	证券卖出	09:39:37	-11600.00	6.2593	72608.00	72523.01
18	20221108	688321	微芯生物	买入	证券买入	09:38:25	2999.00	23.9998	71975.50	-71987.74	11.5
19										

```
[1]: import polars as pl
```

```
[4]: pl.read_csv("stock_trades/202207-湘财.xls",encoding="gb18030")
```

[4]: shape: (17, 1)

发生日期 证券代码 证券名称 买卖标志 业务名称 成交时间 成交数量 成交价格 成交金额 发生金额 手续费 印花税 过户费 其他费 备注 币种
str
"20220721 600269 赣粤高速 卖出 股息入账 1...
"20220718 204007 GC007 卖出 拆出质押购...
"20220718 "=""002462""" 嘉事堂 卖出 ...
"20220718 600408 安泰集团 买入 证券买入 0...
"20220718 600648 外高桥 买入 证券买入 09...
...
"20220711 204007 GC007 卖出 质押回购拆...
"20220711 601992 金隅集团 买入 证券买入 0...
"20220711 600894 广日股份 买入 证券买入 0...
"20220711 601077 渝农商行 买入 证券买入 0...
"20220707 204007 GC007 卖出 拆出质押购...

```
[5]: ("stock_trades/202207-湘财.xls",encoding="gb18030",separator="\t")
```

[5]: shape: (17, 16)

发生日期	证券代码	证券名称	买卖标志	业务名称	成交时间	成交数量
i64	str	str	str	str	str	str
20220721	"600269"	"赣粤高速"	"卖出"	"股息入账"	"16:00:00"	"=""0.00"""
20220718	"204007"	"GC007"	"卖出"	"拆出质押购回"	"19:03:27"	"-580.00"
20220718	"=""002462"""	"嘉事堂"	"卖出"	"证券卖出"	"09:38:10"	"-10400.00"

```
[1]: import polars as pl

[6]: df = pl.read_csv("stock_trades/202207-湘财.xls",encoding="gb18030",separator="\t")

[ ]:
```

Code

```
[1]: import polars as pl

[6]: df = pl.read_csv("stock_trades/202207-湘财.xls",encoding="gb18030",separator="\t")

[8]: type(df)

[8]: polars.dataframe.frame.DataFrame

[ ]:
```

```
[107]: df = pl.read_csv("stock_trades/202207-湘财.xls", encoding="gb18030", separator="\t", infer_schema=False)
       df = df.with_columns(
           pl.col("发生日期").str.to_date("%Y%m%d"),
           pl.col("证券代码").str.strip_prefix("=").str.strip_chars('"'),
       )
       df[:, "证券代码"].unique().to_list()

[107]: ['600648',
        '601992',
        '000900',
        '600408',
        '204007',
        '600015',
        '600269',
        '002462',
        '600894',
        '601077']

[ ]:
```

```
[128]: from pathlib import Path

[139]: [read_df_湘财(f) for f in Path("stock_trades/").glob("*-湘财.xls")]

---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[139], line 1
----> 1 [read_df_湘财(f) for f in Path("stock_trades/").glob("*-湘财.xls")]

NameError: name 'read_df_湘财' is not defined

[ ]:
```

```
0]: def read_df_湘财(f: str | Path) -> pl.DataFrame:
        df = pl.read_csv(
            f,
            encoding="gb18030",
            separator="\t",
            infer_schema=False,
        )
        df = df.with_columns(
            pl.selectors.all().str.strip_prefix("=").str.strip_chars('"'),
        ).with_columns(
            pl.col("发生日期").str.to_date("%Y%m%d"),
            pl.col("成交时间").str.to_time(),
            pl.col(
                "成交数量",
                "成交价格",
                "成交金额",
                "发生金额",
                "手续费",
                "印花税",
                "过户费",
                "其他费",
            ).cast(pl.Float64),
        )
        df = df.filter(
            pl.col("业务名称").is_in(["证券买入", "证券卖出"]),
        )
        return df
```

```
[142]: df = [read_df_湘财(f) for f in Path("stock_trades/").glob("*-湘财.xls")]
[148]: d1 = pl.concat(df)
[153]: d1 = d1.with_columns(
           券商=pl.lit("湘财"),
       )
[154]: d1
```

[154]: shape: (257, 17)

发生日期	证券代码	证券名称	买卖标志	业务名称	成交时间	成交数量	成交价格	成交金额	发生金额	手续费	印花税	过户费	其他费	备注	币种	券商
date	str	str	str	str	time	f64	f64	f64	f64	f64	f64	f64	f64	str	str	str
2022-07-18	"002462"	"嘉事堂"	"卖出"	"证券卖出"	09:38:10	-10400.0	13.2062	137344.0	137184.67	21.98	137.35	1.38	0.0	"证券卖出"	"人民币"	"湘财"
2022-07-18	"600408"	"安泰集团"	"买入"	"证券买入"	09:44:52	47000.0	3.19	149930.0	-149955.5	23.99	0.0	1.51	0.0	"证券买入"	"人民币"	"湘财"
2022-07-18	"600648"	"外高桥"	"买入"	"证券买入"	09:44:31	11900.0	12.6066	150019.0	-150044.49	24.0	0.0	1.49	0.0	"证券买入"	"人民币"	"湘财"
2022-07-18	"600269"	"赣粤高速"	"买入"	"证券买入"	09:43:38	40700.0	3.69	150183.0	-150208.53	24.03	0.0	1.5	0.0	"证券买入"	"人民币"	"湘财"
2022-07-18	"600015"	"华夏银行"	"买入"	"证券买入"	09:42:51	30000.0	5.07	152100.0	-152125.86	24.34	0.0	1.52	0.0	"证券买入"	"人民币"	"湘财"
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2023-06-19	"603967"	"中创物流"	"卖出"	"证券卖出"	10:18:46	-5000.0	9.13	45650.0	45596.59	7.3	45.65	0.46	0.0	"证券卖出"	"人民币"	"湘财"

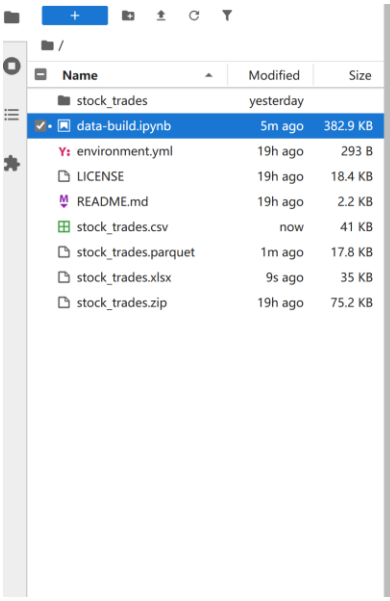
Simple 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 data-build.ipynb 1

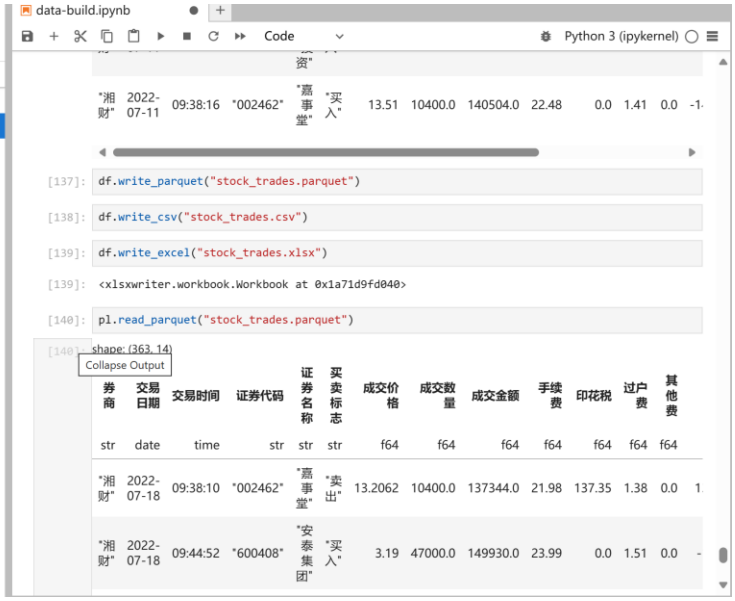
自选股

```
df = pl.read_excel(
    "stock_trades/202305-海通普通.xlsx",
    schema_overrides={
        "成交日期": pl.String,
        "成交时间": pl.String,
    },
)
df.with_columns(
    pl.col("成交日期").str.to_date("%Y%m%d"),
    pl.col("成交时间").str.to_time("%H:%M:%S"),
)
```

```
---------------------------------------------------------------------------
InvalidOperationError                     Traceback (most recent call last)
Cell In[110], line 8
      1 df = pl.read_excel(
      2     "stock_trades/202305-海通普通.xlsx",
      3     schema_overrides={
   (...)
      6     },
      7 )
----> 8 df.with_columns(
      9     pl.col(          ).str.to_date(     %d ),
     10     pl.col(          ).str.to_time(          ),
     11 )

File D:\ananconda\envs\week08\Lib\site-packages\polars\dataframe\frame.py:9806, in DataFrame.with_columns(self, *exprs, **named_exprs)
   9660 def with_columns(
   9661     self,
   9662     *exprs: IntoExpr | Iterable[IntoExpr],
   9663     **named_exprs: IntoExpr,
```





"max" 0.0

[9]:
```
df.with_columns(
    手续费率=pl.col("手续费") / pl.col("成交金额"),
    印花税率=pl.col("印花税") / pl.col("成交金额"),
    过户费=pl.col("过户费") / pl.col("成交金额"),
)
```

[9]: shape: (363, 16)

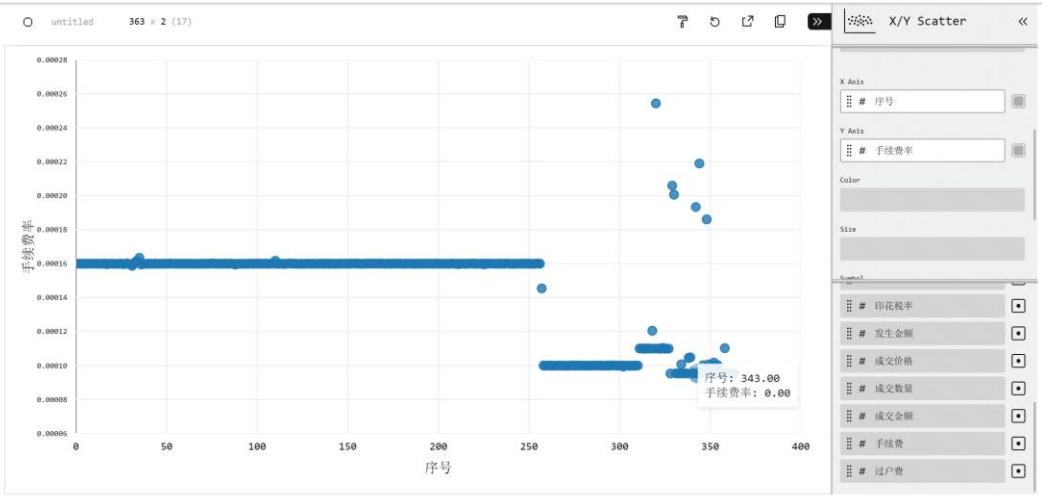
券商	交易日期	交易时间	证券代码	证券名称	买卖标志	成交价格	成交数量	成交金额	手续费	印花税	过户费	其他费	发生金额	手续费率
str	date	time	str	str	str	f64	f64	f64	f64	f64	f64	f64	f64	f6
"湘财"	2022-07-18	09:38:10	"002462"	"嘉事堂"	"卖出"	13.2062	10400.0	137344.0	21.98	137.35	0.00001	0.0	137184.67	0.0001
"湘财"	2022-07-18	09:44:52	"600408"	"安泰集团"	"买入"	3.19	47000.0	149930.0	23.99	0.0	0.00001	0.0	-149955.5	0.0001
"湘财"	2022-07-18	09:44:31	"600648"	"外高桥"	"买入"	12.6066	11900.0	150019.0	24.0	0.0	0.00001	0.0	-150044.49	0.0001
"湘财"	2022-07-18	09:43:38	"600269"	"赣粤高	"买	3.69	40700.0	150183.0	24.03	0.0	0.00001	0.0	-150208.53	0.0001

Idle Mode: Command Ln 1, Col 1 data-query.ipynb 1

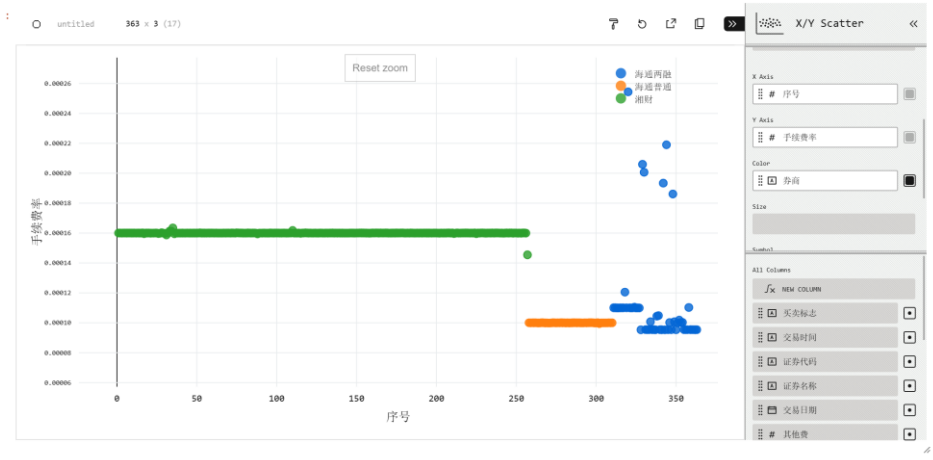
data-query.ipynb

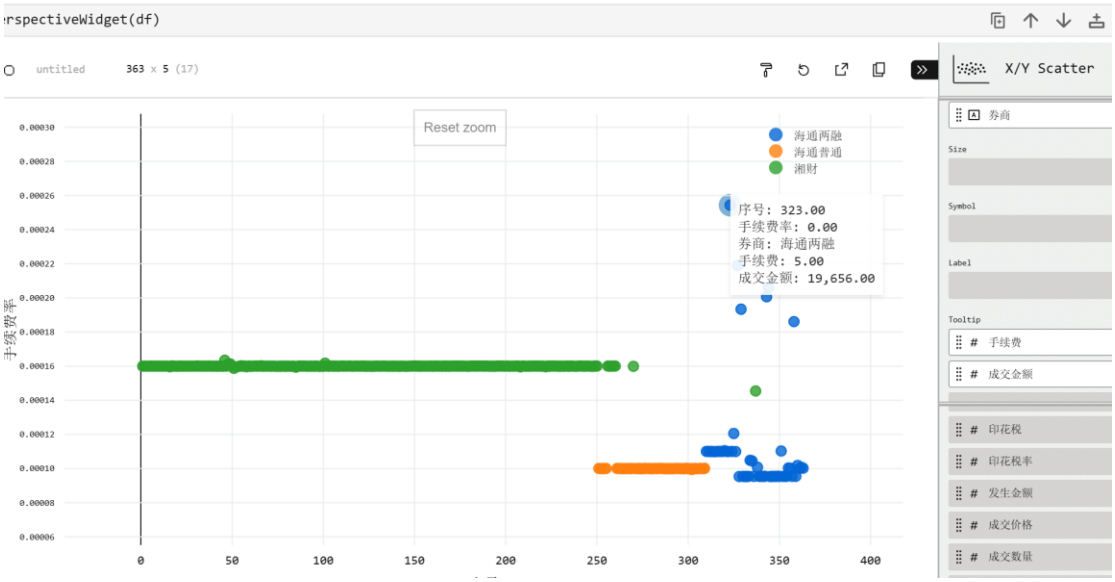
Code Python 3 (ipykernel)

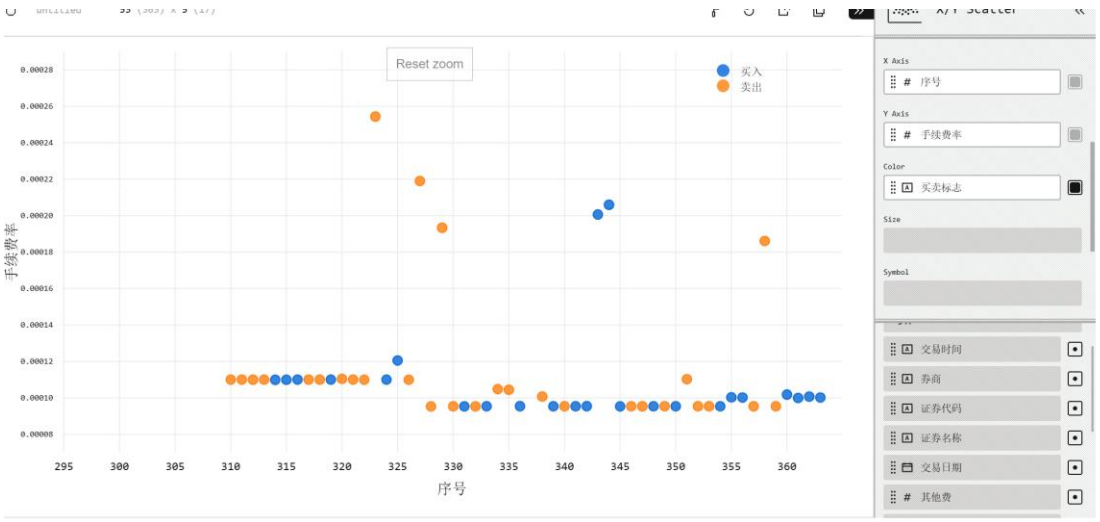
[7]:



[ ]:








```
42]: k1.join(k2, how="cross")
```

42]: shape: (72_656, 2)

日期	证券代码
date	str
2022-07-11	"000096"
2022-07-11	"000532"
2022-07-11	"000559"
2022-07-11	"000599"
2022-07-11	"000655"
…	…
2023-10-31	"688299"
2023-10-31	"688321"
2023-10-31	"688360"
2023-10-31	"688393"
2023-10-31	"688660"

笛卡尔积

```
[43]: k = k1.join(k2, how="cross")
```

```
[45]: k.join(d1, left_on=["日期", "证券代码"], right_on=["交易日期", "证券代码"], how="left").drop_nulls("成交金额")
```

[45]: shape: (358, 18)

日期	证券代码	序号	券商	交易时间	证券名称	买卖标志	成交价格	成交数量	成交金额	手续费	印花税	过户费	其他费	发生金额	手续费率	印花税率	过户费率
date	str	u32	str	str	str	str	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64
2022-07-11	"000900"	1	"湘财"	"09:33:37"	"现代投资"	"买入"	4.05	34400.0	139320.0	22.29	0.0	1.39	0.0	-139342.29	0.00016	0.0	0.00001
2022-07-11	"002462"	5	"湘财"	"09:38:16"	"嘉事堂"	"买入"	13.51	10400.0	140504.0	22.48	0.0	1.41	0.0	-140526.48	0.00016	0.0	0.00001
2022-07-11	"600894"	3	"湘财"	"09:36:30"	"广日股份"	"买入"	6.54	21400.0	139956.0	22.39	0.0	1.41	0.0	-139979.8	0.00016	0.0	0.00001
2022-07-11	"601077"	2	"湘财"	"09:34:24"	"渝农商行"	"买入"	3.65	38300.0	139795.0	22.37	0.0	1.38	0.0	-139818.75	0.00016	0.0	0.00001
2022-07-11	"601992"	4	"湘财"	"09:37:25"	"金隅集团"	"买入"	2.59	54000.0	139860.0	22.38	0.0	1.42	0.0	-139883.8	0.00016	0.0	0.00001
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
2023-10-31	"002753"	357	"海通两融"	"09:31:06"	"永东股份"	"卖出"	7.59	14200.0	107780.0	10.28	53.89	0.0	0.0	107715.83	0.000095	0.0005	0.0
2023-10-31	"002956"	358	"海通两融"	"09:31:31"	"西麦食品"	"卖出"	14.14	1900.0	26866.0	5.0	13.45	0.0	0.0	26847.55	0.000186	0.000501	0.0
2023-10-31	"002956"	359	"海通两融"	"09:31:53"	"西麦食品"	"卖出"	14.13	5000.0	70650.0	6.74	35.35	0.0	0.0	70607.91	0.000095	0.0005	0.0
2023-10-31	"300132"	361	"海通两融"	"09:40:55"	"青松股份"	"买入"	5.21	9600.0	50016.0	5.0	0.0	0.0	0.0	-50021.0	0.0001	0.0	0.0
2023-10-31	"603214"	360	"海通两融"	"09:39:57"	"爱婴室"	"买入"	15.84	3100.0	49104.0	5.0	0.0	0.51	0.0	-49109.51	0.000102	0.0	0.00001

```
[ ]:
```

```
43]: k = k1.join(k2, how="cross")
```

```
46]: k.join(d1, left_on=["日期", "证券代码"], right_on=["交易日期", "证券代码"], how="left")
```

46]: shape: (72_671, 18)

日期	证券代码	序号	券商	交易时间	证券名称	买卖标志	成交价格	成交数量	成交金额	手续费	印花税	过户费	其他费	发生金额	手续费率	印花税率	过户费率
date	str	u32	str	str	str	str	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64
2022-07-11	"000096"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2022-07-11	"000532"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2022-07-11	"000559"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2022-07-11	"000599"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2022-07-11	"000655"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
2023-10-31	"688299"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2023-10-31	"688321"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2023-10-31	"688360"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2023-10-31	"688393"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
2023-10-31	"688660"	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null

```
[ ]:
```

```
[57]: PerspectiveWidget(d2)
```


```
[ ]:
```

```
[57]: PerspectiveWidget(d2)
```

[68]:
```
hq = pro.daily(
    ts_code="002462.SZ",
    start_date=format(start_date, "%Y%m%d"),
    end_date=format(end_date, "%Y%m%d"),
)
hq = pl.from_pandas(hq)
hq
```

[68]: shape: (318, 11)

ts_code	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
str	str	f64	f64	f64	f64	f64	f64	f64	f64	f64
"002462.SZ"	"20231031"	14.75	14.9	14.59	14.7	14.75	-0.05	-0.339	65859.96	96984.271
"002462.SZ"	"20231030"	13.88	14.89	13.88	14.75	13.93	0.82	5.8866	123932.16	180119.372
"002462.SZ"	"20231027"	13.7	13.98	13.51	13.93	13.64	0.29	2.1261	35782.0	49386.168
"002462.SZ"	"20231026"	13.49	13.68	13.4	13.64	13.62	0.02	0.1468	19215.0	26005.866
"002462.SZ"	"20231025"	13.65	13.77	13.58	13.62	13.67	-0.05	-0.3658	18484.0	25274.163
…	…	…	…	…	…	…	…	…	…	…
"002462.SZ"	"20220715"	13.61	13.66	13.12	13.13	13.59	-0.46	-3.3848	32967.65	44114.064
"002462.SZ"	"20220714"	13.54	13.75	13.5	13.59	13.54	0.05	0.3693	21967.0	29851.164
"002462.SZ"	"20220713"	13.55	13.63	13.39	13.54	13.61	-0.07	-0.5143	22793.0	30714.624
"002462.SZ"	"20220712"	13.65	13.69	13.41	13.61	13.65	-0.04	-0.293	29679.0	40146.31
"002462.SZ"	"20220711"	13.2	13.96	13.07	13.65	13.18	0.47	3.566	62827.0	85869.11

[ ]:

[87]:
```
len(ts_codes)
```

[87]: 149

[89]:
```
from tqdm.notebook import tqdm
```

[90]:
```
hq = [
    pro.daily(
        ts_code=ts_code,
        start_date=format(start_date, "%Y%m%d"),
        end_date=format(end_date, "%Y%m%d"),
    )
    for ts_code in tqdm(ts_codes)
]
```

100% 149/149 [00:28<00:00, 18.33it/s]

[91]:
```
len(hq)
```

[91]: 149

[ ]:

53]:
```
d1.join(
    hq, left_on=["交易日期", "证券代码"], right_on=["trade_date", "ts_code"], how="left"
).filter(
    pl.col.成交价格.is_between(pl.col("low"), pl.col("high")),
)
```

53]: shape: (358, 27)

序号	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	成交金额	手续费	印花税	过户费	其他费	发生金额	手续费率	印花税率	过户费率	open	high	low	close	pre_close	change	pct_chg	
u32	str	date	str	str	str	str	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	
1	"湘财"	2022-07-11	"09:33:37"	"000900"	"现代投资"	"买入"	4.05	34400.0	139320.0	22.29	0.0	1.39	0.0	-139342.29	0.00016	0.0	0.00001	4.08	4.13	4.04	4.12	4.06	0.06	1.4778	89
2	"湘财"	2022-07-11	"09:34:24"	"601077"	"渝农商行"	"买入"	3.65	38300.0	139795.0	22.37	0.0	1.38	0.0	-139818.75	0.00016	0.0	0.00001	3.65	3.68	3.64	3.66	3.65	0.01	0.274	370
3	"湘财"	2022-07-11	"09:36:30"	"600894"	"广日股份"	"买入"	6.54	21400.0	139956.0	22.39	0.0	1.41	0.0	-139979.8	0.00016	0.0	0.00001	6.57	6.57	6.49	6.51	6.57	-0.06	-0.9132	2
4	"湘财"	2022-07-11	"09:37:25"	"601992"	"金隅集团"	"买入"	2.59	54000.0	139860.0	22.38	0.0	1.42	0.0	-139883.8	0.00016	0.0	0.00001	2.61	2.62	2.58	2.59	2.61	-0.02	-0.7663	24

```
d1.join(
    hq, left_on=["交易日期", "证券代码"], right_on=["trade_date", "ts_code"], how="left"
).filter(
    pl.col.成交价格 > pl.col.close,
)
```

shape: (172, 27)

序号	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	成交金额	手续费	印花税	过户费	其他费	发生金额	手续费率	印花税率	过户费率	open	high	low	close	pre_close	change	pct_chg
u32	str	date	str	str	str	str	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64	f64
3	"湘财"	2022-07-11	"09:36:30"	"600894"	"广日股份"	"买入"	6.54	21400.0	139956.0	22.39	0.0	1.41	0.0	-139979.8	0.00016	0.0	0.00001	6.57	6.57	6.49	6.51	6.57	-0.06	-0.9132
16	"湘财"	2022-08-24	"09:42:45"	"600408"	"安泰集团"	"卖出"	3.32	4000.0	13280.0	2.12	13.28	0.13	0.0	13264.47	0.00016	0.001	0.00001	3.34	3.36	3.26	3.27	3.33	-0.06	-1.8018
18	"湘财"	2022-08-31	"09:30:42"	"605158"	"华达新材"	"卖出"	7.3932	20800.0	153779.0	24.6	153.79	1.53	0.0	153599.08	0.00016	0.001	0.00001	7.5	7.84	7.1	7.1	7.24	-0.14	-1.9337
19	"湘财"	2022-08-31	"09:35:44"	"603878"	"武进不锈"	"买入"	9.05	16900.0	152945.0	24.47	0.0	1.55	0.0	-152971.02	0.00016	0.0	0.00001	9.19	9.2	8.7	8.77	9.11	-0.34	-3.7322
20	"湘财"	2022-09-01	"09:40:01"	"603878"	"武进不锈"	"卖出"	9.0736	4700.0	42646.0	6.82	42.64	0.45	0.0	42596.09	0.00016	0.001	0.000011	8.83	9.1	8.75	8.83	8.77	0.06	0.6842
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
352	"海通	2023-	"09:43:19"	"600590"	"泰豪	"卖	6.4	15200.0	97280.0	9.28	48.64	0.99	0.0	97221.09	0.000095	0.0005	0.00001	6.45	6.48	6.32	6.34	6.45	-0.11	-1.7054

```
d3.join(
    hq,
    left_on=["日期", "证券代码"],
    right_on=["trade_date", "ts_code"],
    how="left",
).sort("证券代码", "日期")[:10, ["日期", "证券代码", "结余数量", "close"]]
```

shape: (10, 4)

日期	证券代码	结余数量	close
date	str	f64	f64
2022-07-11	"000096"	0.0	9.29
2022-07-12	"000096"	0.0	9.2
2022-07-13	"000096"	0.0	9.23
2022-07-14	"000096"	0.0	9.14
2022-07-15	"000096"	0.0	8.94
2022-07-16	"000096"	0.0	null
2022-07-17	"000096"	0.0	null
2022-07-18	"000096"	0.0	9.16
2022-07-19	"000096"	0.0	9.21
2022-07-20	"000096"	0.0	9.25

```
d3.join(
    hq,
    left_on=["日期", "证券代码"],
    right_on=["trade_date", "ts_code"],
    how="left",
).with_columns(
    close2=pl.col.close.fill_null(strategy="forward")
).sort("证券代码", "日期")[:10, ["日期", "证券代码", "结余数量", "close", "close2"]]
```

shape: (10, 5)

日期	证券代码	结余数量	close	close2
date	str	f64	f64	f64
2022-07-11	"000096"	0.0	9.29	9.29
2022-07-12	"000096"	0.0	9.2	9.2
2022-07-13	"000096"	0.0	9.23	9.23
2022-07-14	"000096"	0.0	9.14	9.14
2022-07-15	"000096"	0.0	8.94	8.94
2022-07-16	"000096"	0.0	null	9.76
2022-07-17	"000096"	0.0	null	9.76
2022-07-18	"000096"	0.0	9.16	9.16
2022-07-19	"000096"	0.0	9.21	9.21
2022-07-20	"000096"	0.0	9.25	9.25

先按照日期证券代码排序 这样补充不对

[69]:
```
d3.join(
    hq,
    left_on=["日期", "证券代码"],
    right_on=["trade_date", "ts_code"],
    how="left",
).sort("证券代码", "日期").with_columns(
    close2=pl.col.close.fill_null(strategy="forward")
)[:10, ["日期", "证券代码", "结余数量", "close", "close2"]]
```

[69]: shape: (10, 5)

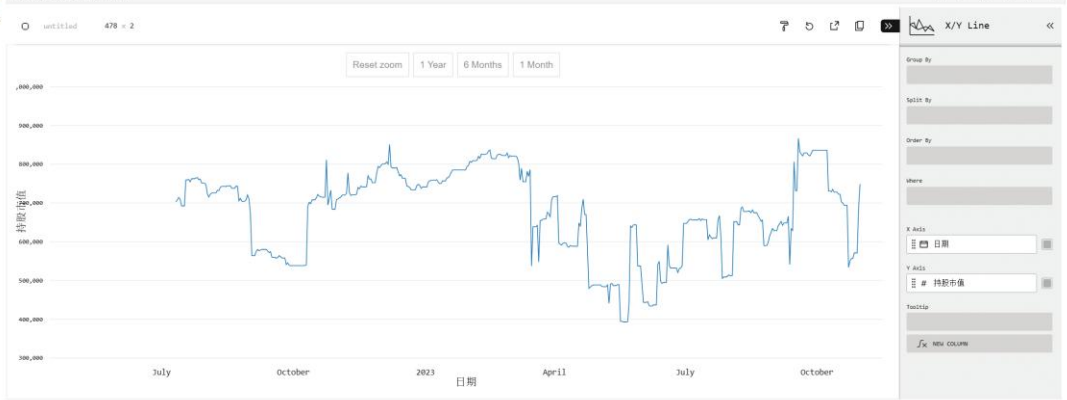
日期	证券代码	结余数量	close	close2
date	str	f64	f64	f64
2022-07-11	"000096"	0.0	9.29	9.29
2022-07-12	"000096"	0.0	9.2	9.2
2022-07-13	"000096"	0.0	9.23	9.23
2022-07-14	"000096"	0.0	9.14	9.14
2022-07-15	"000096"	0.0	8.94	8.94
2022-07-16	"000096"	0.0	null	8.94
2022-07-17	"000096"	0.0	null	8.94
2022-07-18	"000096"	0.0	9.16	9.16
2022-07-19	"000096"	0.0	9.21	9.21
2022-07-20	"000096"	0.0	9.25	9.25

[71]:
```
d3.join(
    hq,
    left_on=["日期", "证券代码"],
    right_on=["trade_date", "ts_code"],
    how="left",
).sort("证券代码", "日期").with_columns(
    close=pl.col.close.fill_null(strategy="forward").over("证券代码")
).with_columns(持股市值=pl.col.结余数量 * pl.col.close).group_by("日期").agg(
    pl.col.持股市值.sum()
)
```

[71]: shape: (478, 2)

日期	持股市值
date	f64
2023-02-24	822244.0
2023-02-12	825825.0
2022-10-19	717406.0
2023-07-05	657883.0
2022-10-10	540415.0
…	…
2023-09-17	731764.0
2023-06-14	497432.0
2023-05-12	486786.0
2022-09-07	576944.0
2023-09-26	821560.0

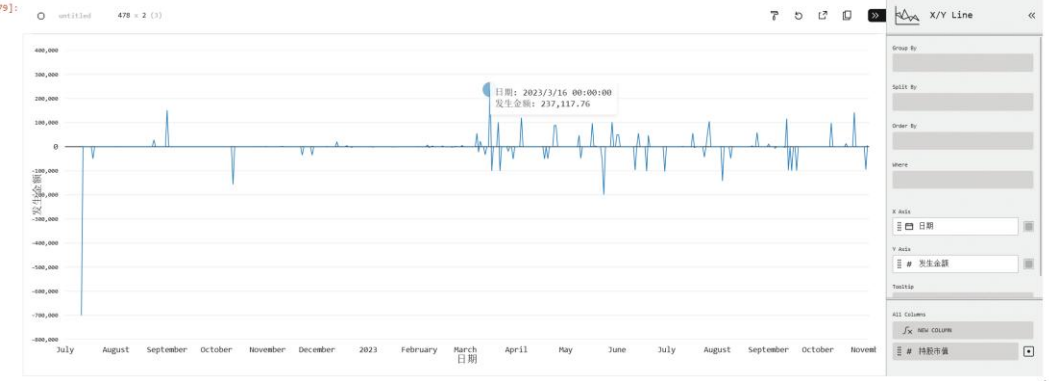
[76]: PerspectiveWidget(d4)



[ ]:

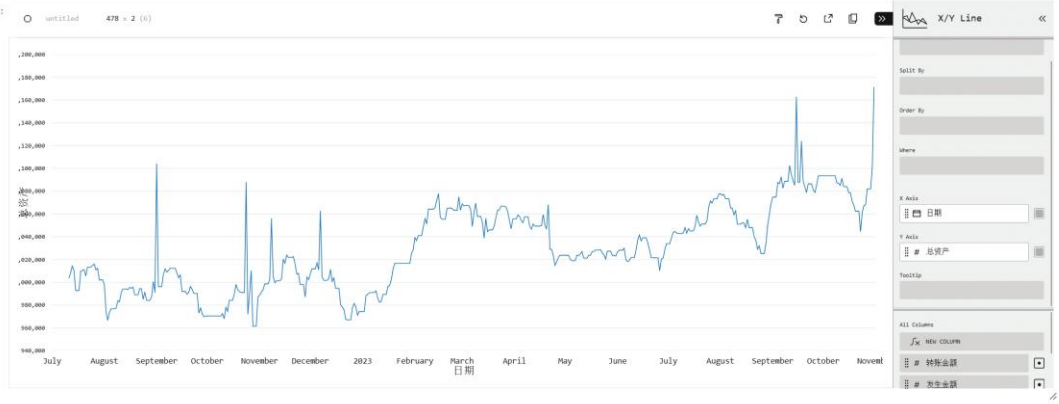
2023-10-27	571195.0	0.0
2023-10-28	571195.0	0.0
2023-10-29	571195.0	0.0
2023-10-30	686345.0	-94884.93
2023-10-31	748947.0	6490.78

[79]: PerspectiveWidget(d4)



[ ]:

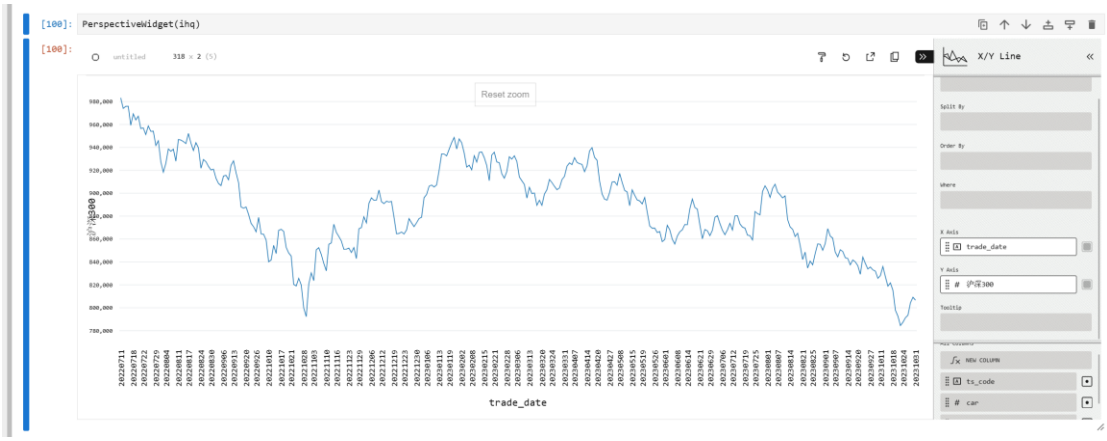
[88]: PerspectiveWidget(d4)

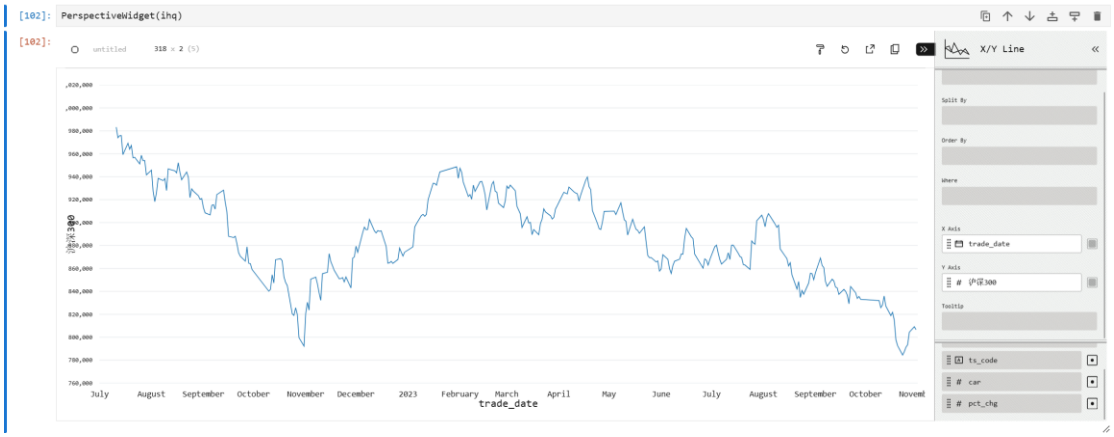


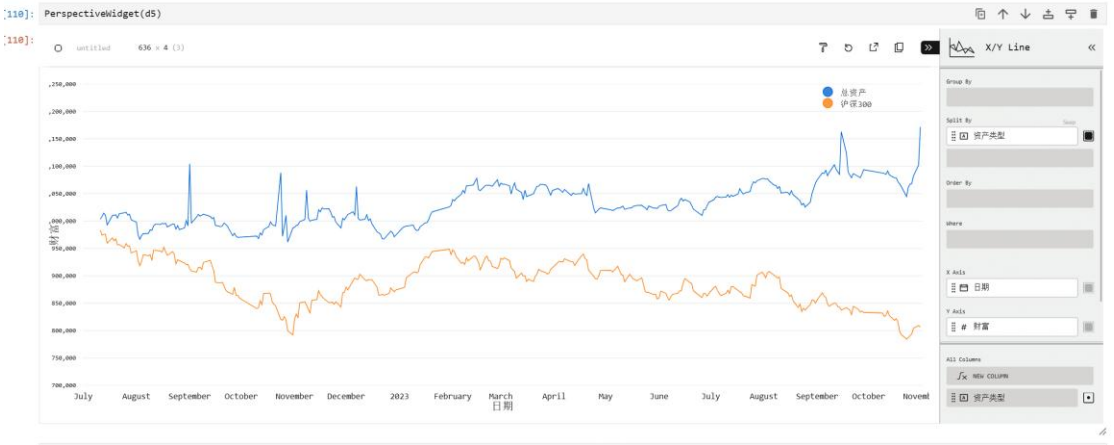
[99]:
```
ihq = pl.read_parquet("index_daily.parquet")
ihq = (
    ihq.with_columns(
        pl.col.pct_chg / 100 + 1,
    )
    .sort("trade_date")
    .with_columns(
        car=pl.col.pct_chg.cum_prod(),
    )
    .with_columns(
        沪深300=pl.col.car * 100_0000,
    )
)
ihq
```

[99]: shape: (318, 5)

ts_code	trade_date	pct_chg	car	沪深300
str	str	f64	f64	f64
"000300.SH"	"20220711"	0.983254	0.983254	983254.0
"000300.SH"	"20220712"	0.990585	0.973997	973996.66359
"000300.SH"	"20220713"	1.001818	0.975767	975767.389524
"000300.SH"	"20220714"	1.000142	0.975906	975905.948494
"000300.SH"	"20220715"	0.982983	0.959299	959298.956968
…	…	…	…	…
"000300.SH"	"20231025"	1.004969	0.791288	791288.453778
"000300.SH"	"20231026"	1.002764	0.793476	793475.575065
"000300.SH"	"20231027"	1.013727	0.804368	804367.614284
"000300.SH"	"20231030"	1.006003	0.809196	809196.233072
"000300.SH"	"20231031"	0.996856	0.806652	806652.120115







在第 8 周数据清洗与计算课程中，学习了使用 Polars 库读取不同格式的证券交易数据（如 excel 文件），通过指定编码、分隔符及用参数处理数据读取问题，掌握了数据清洗操作，包括去除字符串前缀和引号、转换日期和时间格式、筛选特定业务类型（如证券买入/卖出），以及计算手续费率、印花税率等衍生指标；实践了数据合并（如左连接）和填充缺失值（前向填充），并运用 group_by 和 agg 进行分组聚合计算持股市值，还通过可视化工具观察数据趋势，结合外部数据源（如股票行情数据）完成交易数据与市场行情的关联分析。