

Choice 数据库 A 股日度交易数据质量校验报告

小组成员：苏麒畅 李心怡 尹若伊 张新宇

戴子河 张腾月 齐心悦 张溪瑶

目录

一、引言	4
1. 1 背景	4
1. 2 目的.....	4
1. 3 数据范围.....	4
二、数据基础分析	4
2. 1 数据合并与预处理	4
2. 2 数据质量校验	4
2. 3 数据可视化探索	4
2. 4 结论生成与报告输出	8
2. 5 检查数据是否缺失	8
三、数据预处理与基础异常检测	9
3. 1 价格数据校验	9
3. 2 后复权因子校验	9
3. 3 交易状态与成交情况校验	9
3. 4 市值与股本数据校验.....	10
四、多维度交叉验证数据	11
4. 1 业务逻辑校验.....	11
4. 1. 1 市值一致性验证.....	11
4. 1. 2 换手率合理性检查.....	11
4. 1. 3 量化关系深度验证.....	11
4. 1. 4 涨跌停标志验证.....	12
4. 2 时间序列校验.....	12
4. 2. 1 股本变动连续性检查	12
4. 2. 2 价格序列连续性检查.....	12
4. 3 统计异常校验.....	12
4. 3. 1 极端收益率检测.....	12
4. 3. 2 极端成交量检测.....	12

4.3.3 价格-成交量背离检测.....	12
4.4 整合异常原因分布图.....	12
五、数据质量综合评价	14
5.1 校验方法总结与结果分析	14
5.1.1 数据质量等级评定.....	14
5.1.2 数据质量综合报告.....	15
5.2 最终结论.....	16

图表目录

图 1 涨跌停状态分布图.....	5
图 2 各股票交易天数分布图.....	5
图 3 交易量分布图.....	6
图 4 股票价格分布图.....	6
图 5 成交额分布图.....	7
图 6 总市值分布图.....	7
图 7 日内价格波动率分布图.....	8
图 8 数据质量校验报告示例.....	8
图 9 各字段数据完整性分析.....	9
图 10 校验异常数据集示例	11
图 11 异常原因分布图.....	13
图 12 主要数据分布图.....	14
图 13 数据质量综合报告卡	15
图 14 数据质量报告.....	16
表 1 字段矛盾逻辑说明.....	10

一、引言

1.1 背景

本公司因 Wind 数据库年使用费上涨，拟转用成本较低的 Choice 数据库，但对其数据质量存疑。为确保量化投资策略的稳定性，需对 Choice 数据库提供的 A 股日度交易数据进行全面质量校验。

1.2 目的

通过多维度数据质量校验，评估 Choice 数据库的可靠性，为是否正式转用提供决策依据。

1.3 数据范围

校验数据为 Choice 数据库提供的 2025 年 9 月 1 日至 24 日的 A 股日度交易数据，包含股票代码、开盘价、最高价、最低价、收盘价、后复权因子、停牌天数、是否涨停或者跌停（是为 1，否为 0）、成交量、成交额、总市值、总股本、A 股流通股本。等关键字段。

二、数据基础分析

对数据进行该系统通过数据合并预处理、多维度质量校验和可视化探索，最终生成数据质量评分与结构化报告，并将所有检查结果整合至统一列表中。

2.1 数据合并与预处理

系统将自动读取指定目录下的所有 RData 数据文件，并将其合并为一个统一的数据集。在此过程中，关键步骤是对其中的日期字段进行标准化处理，统一转换为规范的日期格式，最终生成一个名为“DATE”的标准日期列，为后续的时间序列分析奠定基础。

2.2 数据质量校验

从数据完整性、数值合理性到业务逻辑一致性（如价格非负、成交量合理等）进行全面检查。将涨跌停数据、股本数据、复权因子、股票覆盖度、数据整体规模和时间范围、数据完整性、价格数据合理性、交易量成交额逻辑、停牌数据逻辑的校验结果汇总到名为“validation_results”的校验结果列表中，清晰标识出每条记录的通过状态与潜在问题。

2.3 数据可视化探索

利用 ggplot2 图形系统生成五张核心图表，包括价格分布直方图、取对数后的交易量分布、时间序列的日期覆盖情况、涨跌停发生次数统计以及停牌天数统计。这些图表旨在直观揭示数据的分布特征、历史覆盖度以及异常点聚集情况，为数

据质量提供视觉佐证。

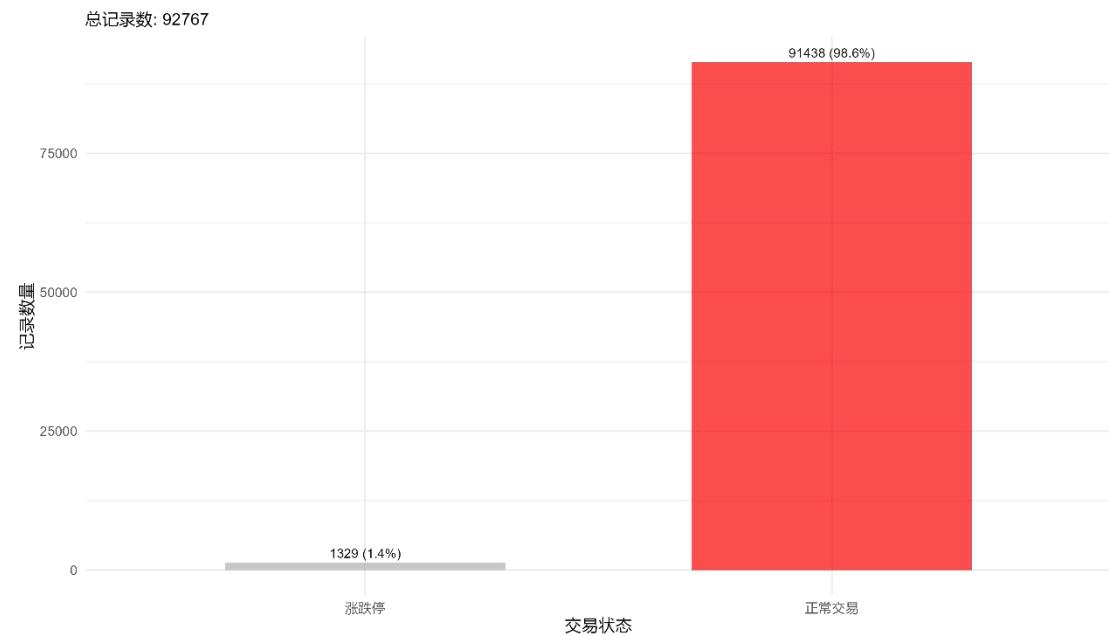


图 1 涨跌停状态分布图

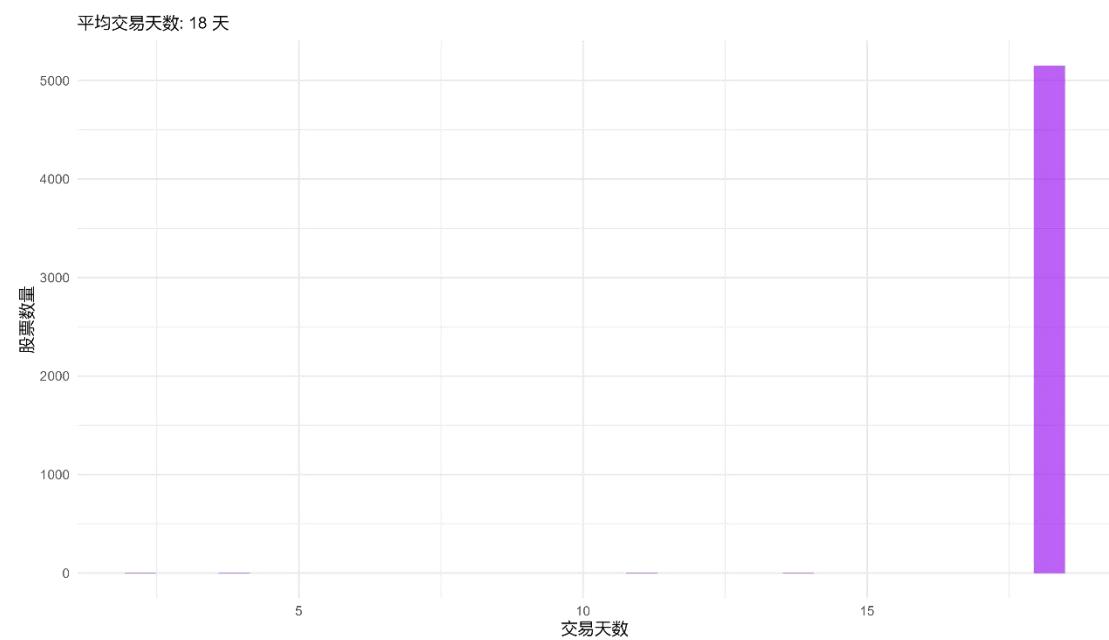


图 2 各股票交易天数分布图

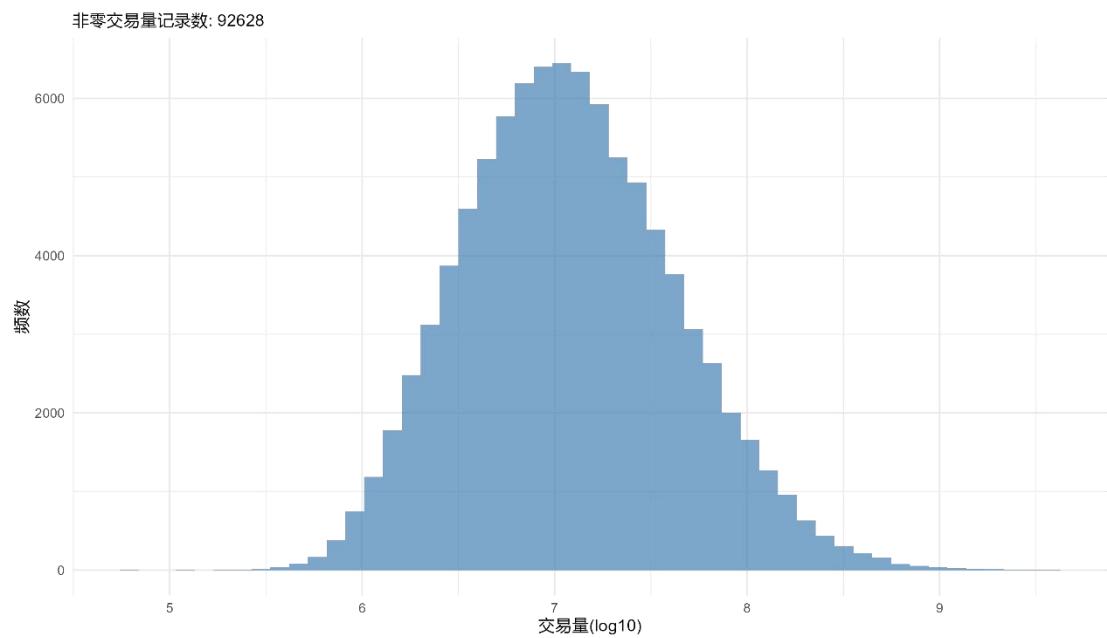


图 3 交易量分布图

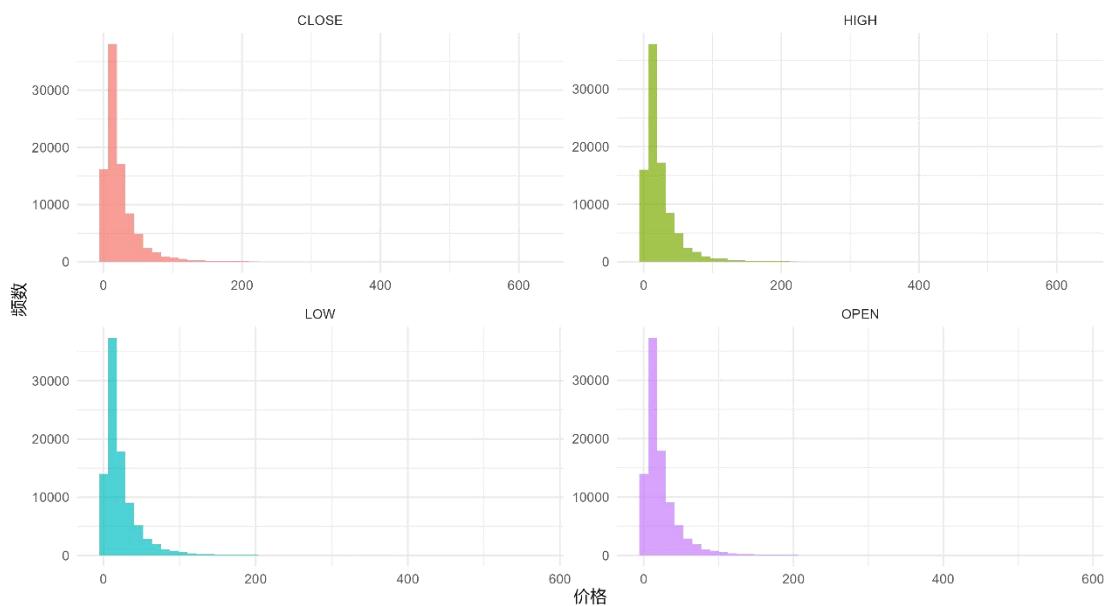


图 4 股票价格（开盘价、最高价、最低价、收盘价）分布图

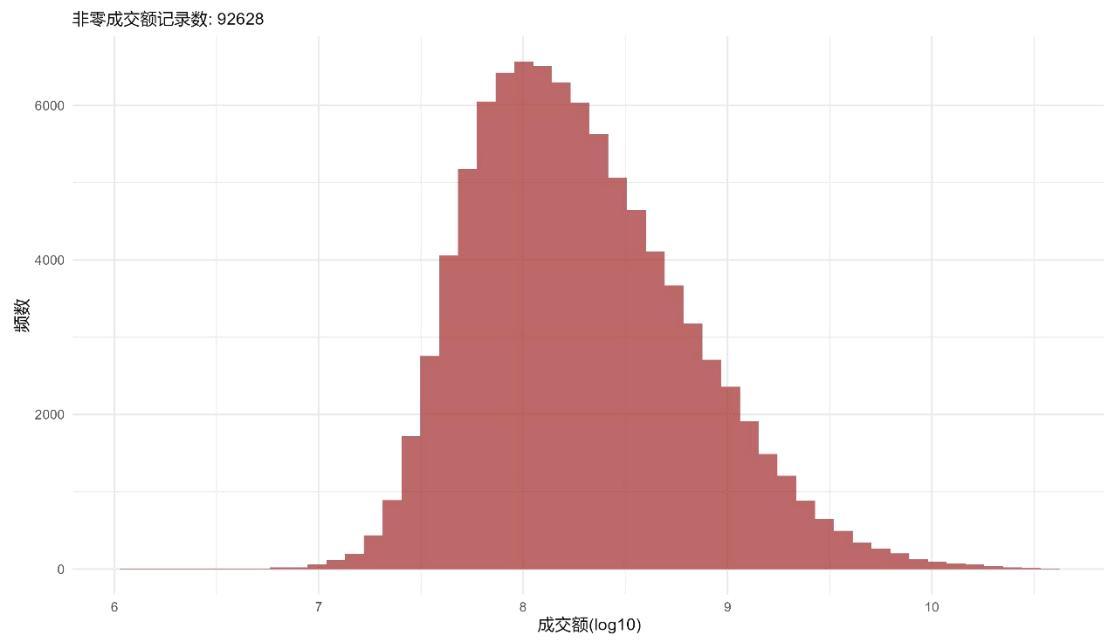


图 5 成交额分布图

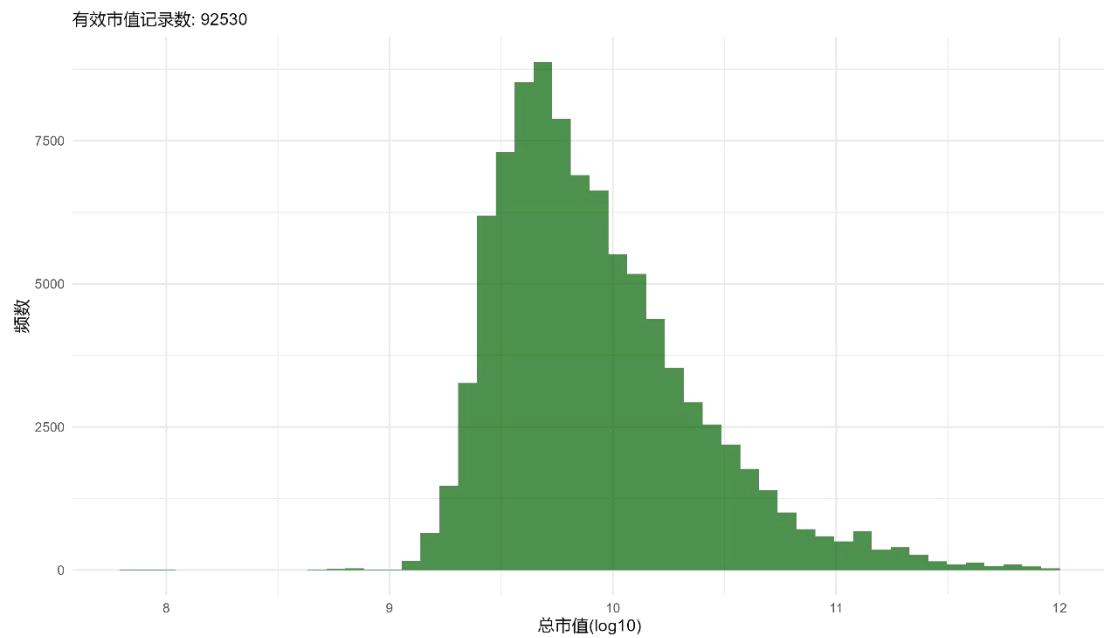


图 6 总市值分布图

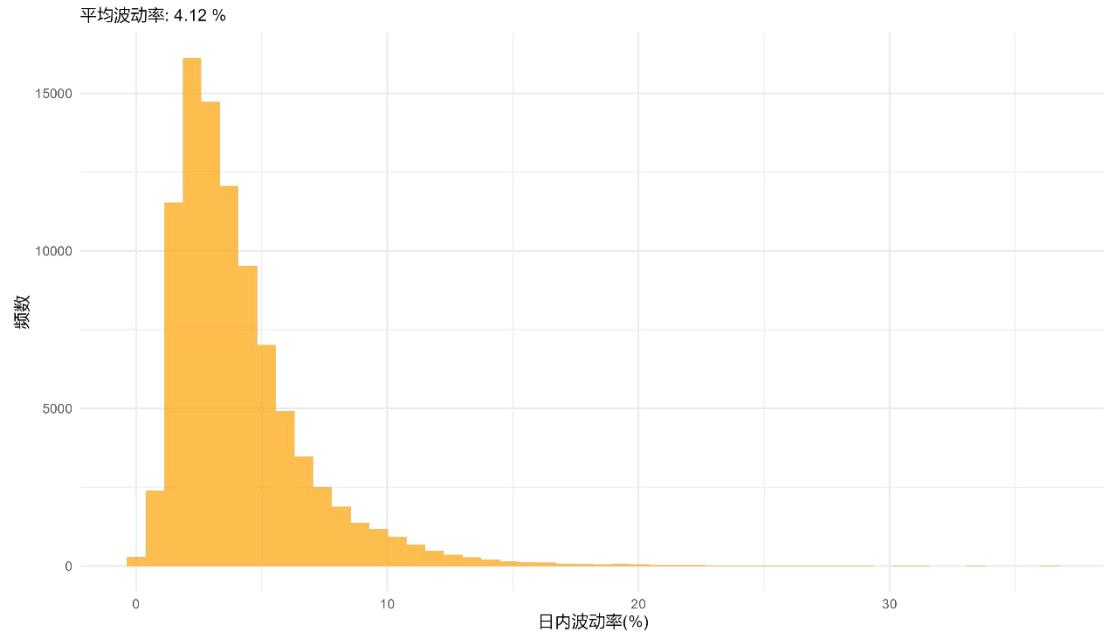


图 7 日内价格波动率分布图

2.4 结论生成与报告输出

基于前述校验与可视化结果，计算综合质量评分，将全部分析结论、质量评分及风险等级判定汇总输出为一份结构化的 CSV 格式评估报告。

A	B	C	D	E	F	G	H	I	J
1	类别	指标	值						
2	基本信息	总记录数		92767					
3	基本信息	股票数量		5156 只					
4	基本信息	数据期间	未知						
5	基本信息	交易日数	未知	天					
6	数据完整性	source_file		100 %					
7	数据完整性	CODE		100 %					
8	数据完整性	OPEN		100 %					
9	数据完整性	HIGH		100 %					
10	数据完整性	LOW		100 %					
11	数据完整性	CLOSE		100 %					
12	数据完整性	ADJFACTOR		100 %					
13	数据完整性	SUSP_DAYS		100 %					
14	数据完整性	MAXUPDOWNDOWN		100 %					
15	数据完整性	VOLUME		99.85 %					
16	数据完整性	AMT		99.85 %					
17	数据完整性	MKT_CAP_ARD		100 %					
18	数据完整性	TOTAL_SHARES		100 %					
19	数据完整性	FLOAT_A_SHARES		100 %					
20	价格合理性	开盘价≤0的记录		0 条 (异常记录)					
21	价格合理性	最高价≤0的记录		0 条 (异常记录)					
22	价格合理性	最低价≤0的记录		0 条 (异常记录)					
23	价格合理性	收盘价≤0的记录		0 条 (异常记录)					
24	价格合理性	最高价<最低价的记录		0 条 (异常记录)					
25	价格合理性	收盘价超出高价范围的记录		0 条 (异常记录)					
26	价格合理性	开盘价超出低价范围的记录		0 条 (异常记录)					
27	价格合理性	日波动超50%的记录		0 条 (异常记录)					
28	量价逻辑	交易量=0但成交额>0的记录		0 条 (矛盾记录)					
29	量价逻辑	成交额=0但交易量>0的记录		0 条 (矛盾记录)					
30	量价逻辑	交易量>0但成交额=0的记录		0 条 (矛盾记录)					
31	量价逻辑	成交额>0但交易量=0的记录		0 条 (矛盾记录)					
32	停牌逻辑	停牌但有交易的记录		0 条 (矛盾记录)					
33	停牌逻辑	停牌比例	0.15 %						
34	涨跌停数据	涨跌停比例	143 %						
35	涨跌停数据	涨跌停但交易量=0的记录		0 条 (异常记录)					
36	市值股本	市值≤0的记录		0 条 (异常记录)					
37	市值股本	总股本≤0的记录		0 条 (异常记录)					
38	市值股本	流通股≤0的记录		0 条 (异常记录)					
39	市值股本	流通股>总股本的记录		0 条 (异常记录)					
40	复权因子	复权因子≤0的记录		0 条 (异常记录)					

图 8 数据质量校验报告示例

2.5 检查数据是否缺失

结果表明数据几乎不缺失，可进行下一步分析。

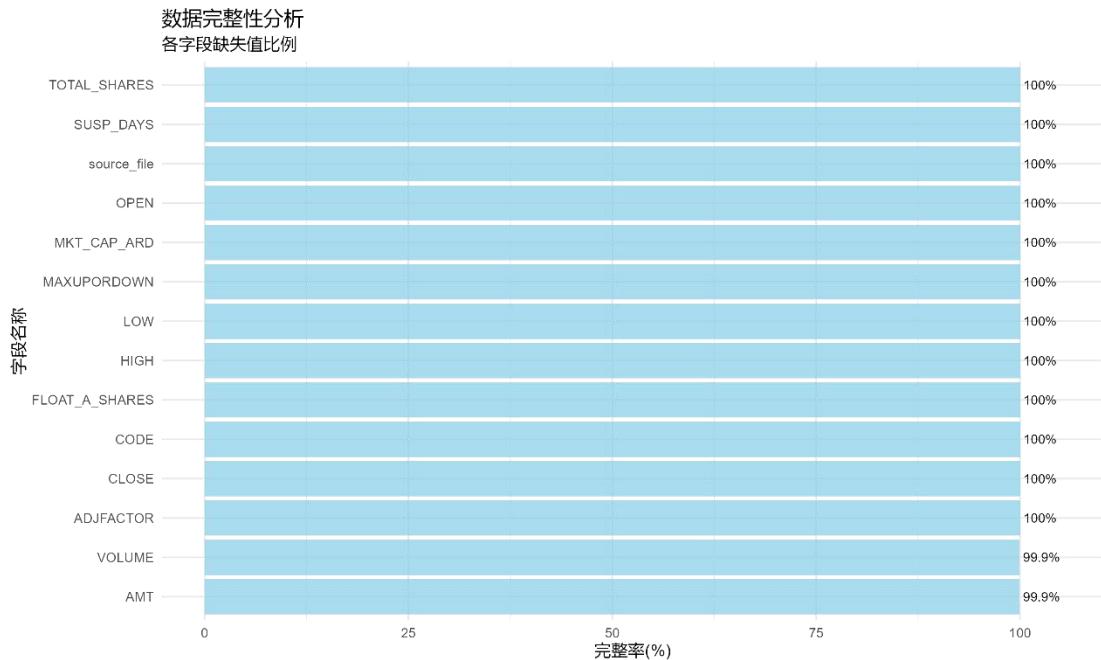


图 9 各字段数据完整性分析

三、数据预处理与基础异常检测

3.1 价格数据校验

价格数据从三个层面进行综合校验：首先，在相对价格关系上，必须满足“最高价 \geq 开盘价 \geq 最低价”的基本逻辑，且收盘价必须介于当日最低价与最高价之间；其次，在价格与交易量关系上，需识别逻辑性背离，例如高价股不应出现异常低的交易量，对于因数据缺失导致的异常应单独标注；最后，在绝对价格层面，A 股个股价格通常应处于一个合理的市场共识区间内（如 1 至 1300 元），以排除极端数值错误。

3.2 后复权因子校验

后复权因子的异常判定主要依据以下三个标准：首先，其绝对数值通常应处于 0.1 至 100 的合理区间内；其次，需进行基准值检查，若因子恒等于 1，则表明该价格序列可能未经任何复权处理，这与成熟市场中股票因长期分红派息而导致复权因子累积偏离 1 的普遍规律相悖；最后，必须验证其与价格变动的一致性，大幅度的复权调整理应对应于公司重大的除权除息事件，并在当日价格走势中产生相应的匹配性变动。

3.3 交易状态与成交情况校验

对股票交易状态与成交情况的异常判定，首先基于其内在的业务逻辑进行基础验

证。SUSP_DAYS 必须为非负整数且无缺失；若其大于 0，则意味着该日股票停牌，其 VOLUME 与 AMT 理论上应接近或等于 0，这是一条关键的一致性规则。同时，VOLUME 与 AMT 自身必须满足非负且无缺失值的基本条件；若出现成交量为 0 但当日并未停牌 (SUSP_DAYS=0)，或成交额为 0 但成交量却大于 0 的情况，均属于明显的逻辑矛盾，判定为异常。

对于 MAXUPORDOWN，其值必须为二进制（0 或 1）且无缺失，应检查当日的收盘价是否达到涨跌幅限制。如果因缺乏前一日收盘价数据而无法精确验证，此一致性原则仍需在数据质量评估中予以关注，作为潜在的风险点。

除了基本逻辑之外，还采用统计方法来检测异常值。对于成交量和金额，使用箱线图和分位数分析等技术来识别极端值。当价格波动正常且没有交易暂停记录，但成交量或金额却为零时，可能表明缺失的数据被误报为没有交易活动。

3.4 市值与股本数据校验

我们首先对总市值进行筛选，排除那些超出 1% 至 99% 分位数范围 10 倍的极端小值或极端大值。然后对总股本与流通股本进行数值合理性校验，确保流通股本不应超过总股本，避免自相矛盾的结果。

字段	异常逻辑	说明
MKT_CAP_ADR	小于分布下 $1\% \times 0.1$ 倍或大于上 $99\% \times 10$ 倍	检查极端偏小/偏大市值
TOTAL_SHARES	小于 $1\% \times 0.1$ 倍或大于 $99\% \times 10$ 倍	检查股本异常偏离
FLOAT_A_SHARES	小于 $1\% \times 0.1$ 倍或大于 $99\% \times 10$ 倍或超过总股本	检查流通股比例或逻辑问题

表 1 字段矛盾逻辑说明

	CODE	OPEN	HIGH	LOW	CLOSE	ADJFACTOR	SUSP_DAYS	MAXUPORDOWN	VOLUME	AMT	MKT_CAP_ARD	TOTAL_SHARES	FLOAT_A_SHARES	Is_anomaly	anomaly
1	000001.SZ	12.05	12.05	11.85	11.89	118.423499	0	0	185845902	221306332	230736367374	19405918198	19405600653	0	0
2	000002.SZ	6.76	6.86	6.70	6.81	134.284605	0	0	140682543	953846377	81248131498	11930709471	9716399629	0	0
3	000004.SZ	11.39	11.88	11.02	11.08	6.488366	0	0	9529300	109831218	1466773525	132380282	126287768	0	0
4	000006.SZ	7.28	7.38	7.19	7.21	43.833157	0	0	37146737	269645391	9733464282	134995046	1349987396	0	0
5	000007.SZ	7.58	7.71	7.48	7.63	6.765176	0	0	8052065	61438257	2643989576	346448044	346448044	0	0
6	000008.SZ	2.86	2.89	2.84	2.88	24.618972	0	0	42091840	120653962	7823167727	2716377683	2716296952	0	0
7	000009.SZ	9.80	9.87	9.63	9.74	8.332077	0	0	42328138	411364421	25121544019	2579213965	2576020465	0	0
8	000010.SZ	4.47	4.87	4.46	4.87	10.773646	0	1	50680200	239457743	559880662	1149652292	781554111	0	0
9	000011.SZ	9.00	9.14	8.96	9.13	4.661756	0	0	8767202	79627662	5441289110	595970992	526475543	0	0
10	000012.SZ	4.75	4.77	4.71	4.75	26.705571	0	0	12928410	61234709	14585787508	3070692107	1959316598	0	0
11	000014.SZ	14.20	14.70	13.98	14.52	9.525995	0	0	11711893	169165410	3514511172	242046224	242046224	0	0
12	000016.SZ	5.76	5.84	5.72	5.79	19.109665	0	0	64216397	371085311	13942003912	2407945408	1596593800	0	0
13	000017.SZ	6.53	6.89	6.49	6.68	2.259888	0	0	18020031	119329238	4603755352	689184933	302984985	0	0
14	000019.SZ	6.99	7.03	6.91	6.93	4.673401	0	0	11014966	76623979	7987069310	1152535254	416216407	0	0
15	000020.SZ	14.08	14.32	13.92	14.16	1.671946	0	0	3785400	53681269	4009562974	283161227	181165391	0	0
16	000021.SZ	22.25	22.48	21.64	21.95	11.187135	0	0	88762118	1952640499	34401810728	1567280671	1567029346	0	0
17	000025.SZ	18.00	18.18	17.84	18.15	1.977060	0	0	16121263	291340162	7823708508	431088320	392778320	0	0
18	000026.SZ	17.61	17.92	17.42	17.54	7.830164	0	0	990589	174692005	7117100683	405764007	364562983	0	0
19	000027.SZ	6.63	6.64	6.52	6.58	15.744097	0	0	31706727	208286280	31303625647	4757389916	4757389916	0	0
20	000028.SZ	25.86	25.88	25.60	25.83	4.374240	0	0	3779130	97505541	14376075939	556565077	478053712	0	0
21	000029.SZ	28.19	29.98	27.88	29.21	1.548126	0	0	9472415	275083709	2955058600	1011660000	891660000	0	0
22	000030.SZ	5.83	5.86	5.71	5.73	2.821783	0	0	12344193	71037739	984516142	1719810845	1679906445	0	0
23	000031.SZ	3.23	3.32	3.19	3.27	11.354285	0	0	28435303	92645208	14016244619	428631339	4286304701	0	0
24	000032.SZ	26.80	27.20	25.62	25.78	3.224149	0	0	47082824	1229535997	29336589053	1137959234	1089363764	0	0
25	000034.SZ	46.12	47.22	45.91	46.68	2.031145	0	0	67812553	3160190340	33618756469	720196154	601592443	0	0
26	000035.SZ	4.37	4.47	4.36	4.45	2.636953	0	0	28981400	128285043	11128485160	2500783182	2425832661	0	0
27	000036.SZ	4.16	4.50	4.12	4.46	11.187474	0	0	46032424	200847599	6261031514	1403818725	1401150983	0	0
28	000037.SZ	8.88	8.95	8.81	8.95	4.286692	0	0	6131200	54567698	5394725234	602762596	338908150	0	0
29	000039.SZ	8.12	8.14	8.03	8.07	52.003973	0	0	51325641	414027896	43517639507	539252085	2301407141	0	0
30	000042.SZ	9.10	9.20	8.94	8.95	10.196859	0	0	8545850	75935654	5950238694	664831139	664831139	0	0
31	000045.SZ	11.11	11.21	11.02	11.12	2.137812	0	0	6825100	75797213	5632522961	506521849	457021849	0	0
32	000048.SZ	16.30	16.90	16.20	16.33	7.447645	0	0	18886389	311103028	8659509143	530282250	526359600	0	0
33	000049.SZ	24.81	24.88	24.18	24.43	6.286665	0	0	14926459	364219420	9396719386	384638534	384534862	0	0

图 10 校验异常数据集示例

四、多维度交叉验证数据

从业务逻辑、时间序列与统计异常三个维度综合评估数据质量：通过验证市值、换手率等核心业务关系的合理性，确保数据内在逻辑一致；通过检查股本、价格的连续性，保障时间序列的稳定可靠；并利用统计方法侦测极端收益与成交量，有效识别潜在的数据错误与市场异动。

4.1 业务逻辑校验

4.1.1 市值一致性验证

通过计算理论市值（总股本 × 收盘价）与实际市值的差额比理论市值得出偏差率，并标记显著偏差记录。设定 15% 的偏差阈值，考虑数据更新延迟的可能。该检查确保市值数据与股价、股本数据之间的一致性，避免因市值计算错误对投资策略产生影响。

4.1.2 换手率合理性检查

基于日换手率（成交量 / 流通股本）的计算，识别异常换手率情况并标记不合理记录。设定 200% 的换手率上限来允许极端市场情况。此项检查有助于发现成交量数据异常，防止其对成交量相关策略造成干扰。

4.1.3 量化关系深度验证

通过估算成交额（成交量 \times 均价）并与实际成交额比较，计算偏差率，标记显著偏差的记录。设定 20% 的偏差阈值，同时考虑大宗交易的影响。该验证旨在确保成交额数据的合理性，维护量化数据的内在一致性。

4.1.4 涨跌停标志验证

通过计算实际涨跌幅，并区分 ST 股（5% 涨跌幅）与普通股（10% 涨跌幅）的规则，验证涨停标志的准确性。这对于依赖涨停板信号的策略尤为关键。

4.2 时间序列校验

4.2.1 股本变动连续性检查

按股票分组计算股本变动率，识别异常大幅变动，设定 50% 的变动阈值，仅标记真正异常情况。该检查确保股本数据的稳定性，避免异常变动影响市值计算及股权相关策略。

4.2.2 价格序列连续性检查

通过计算复权价格和复权收益率，识别异常价格跳空。设定 30% 的跳空阈值，同时排除正常除权除息的影响。该检查有助于发现价格数据的异常断裂，确保价格序列的连续性，为技术分析提供可靠基础。

4.3 统计异常校验

4.3.1 极端收益率检测

按股票计算收益率，并进行 Z-score 标准化处理，识别统计异常值。设定 3 个标准差或 50% 绝对值的双重保障阈值，用于识别异常价格变动，可能反映数据错误或重大市场事件。

4.3.2 极端成交量检测

基于成交量的分布情况，使用分位数方法检测异常值，设定超过 99 分位数 10 倍的条件，避免对数据分布形态的假设。该检测用于发现异常交易活动，识别潜在的数据错误或市场异动。

4.3.3 价格-成交量背离检测

通过计算价格与成交量的变化情况，检测价格信号背离，设定价格变动大于 2% 且成交量变动大于 10% 的条件。该检测有助于识别异常的价格-成交量关系，正常情况下价格波动应伴随成交量的相应变动。

4.4 整合异常原因分布图

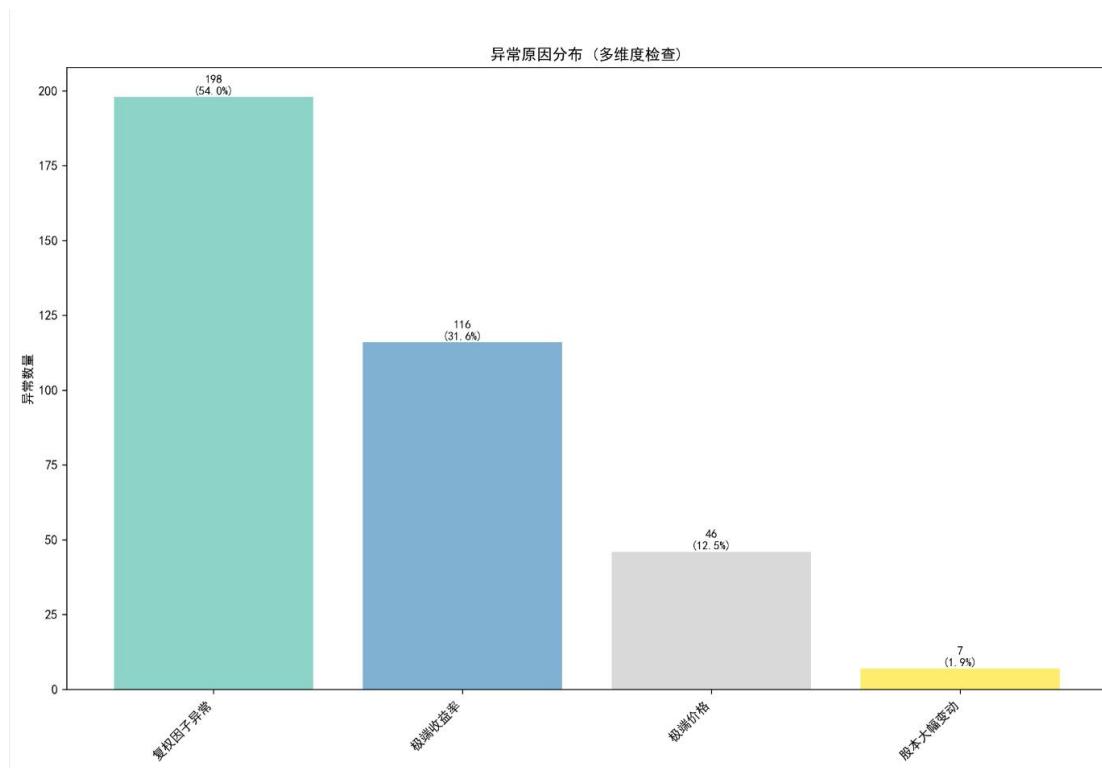


图 11 异常原因分布图

将以上校验结果整合到同一个 DataFrame `self.all_data` 中。结果表明，在所有异常数据中，复权因子异常占比最大，达到 198 条 (54%); 股本大幅度变动占比最小，仅有 7 条 (1.9%)。与 92767 条的总数据量相比，异常数据占比极小。

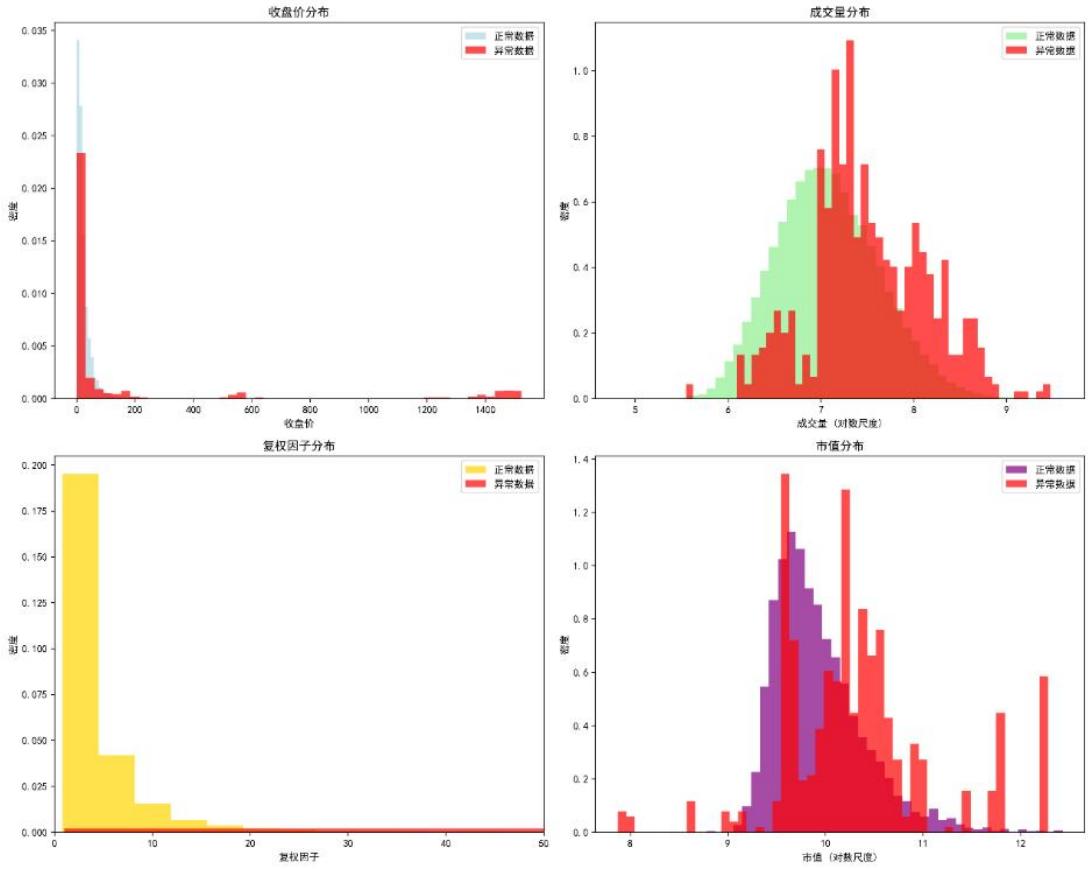


图 12 主要数据（收盘价、成交量、复权因子、市值）分布图

五、数据质量综合评价

5.1 校验方法总结与结果分析

5.1.1 数据质量等级评定

异常率的高低直接反映了数据中潜在“噪声”的水平，进而决定了数据在投入实际应用前所需的处理程度。我们以异常率为核心指标衡量数据集健康状况，设定评级标准：异常率 $\leq 3\%$ ，数据纯净度极高，评为“**A-优秀**”，可直接用于量化策略； $3\% < \text{异常率} \leq 8\%$ ，数据整体良好但有异常点，评为“**B-良好**”，建议基础清洗； $8\% < \text{异常率} \leq 15\%$ ，异常数据规模大，可能干扰分析，评为“**C-合格**”，提醒注意排查；异常率 $> 15\%$ ，数据缺陷严重，评为“**D-需改进**”，必须进行根源分析、彻底清洗校正后才能使用。下图是运行结果：



检查时间: 2025-10-13 13:52

图 13 数据质量综合报告卡

5.1.2 数据质量综合报告

```
{  
  "总体统计": {  
    "总记录数": 92767,  
    "异常记录数": 367,  
    "异常率": "0.40%"  
  },  
  "异常分布": {  
    "股本大幅变动": 7,  
    "极端收益率": 116,  
    "极端价格": 46,  
    "复权因子异常": 198  
  },  
  "质量评估": {  
    "等级": "A (优秀)",  
    "建议": "可直接使用"  
  }  
}
```

图 14 数据质量报告

结果表明数据集整体质量表现优异，异常率控制在 0.4% 的低水平，达到 A 级标准可直接投入使用。其中复权因子异常占比 53.9%，这提示数据采集环节可能存在系统性技术偏差，虽不影响常规分析，但在进行跨期比较或精细计算时需特别注意；极端收益率和极端价格合计占比 44.14%，这些异常可能暴露数据采集噪声，也可能真实反映市场剧烈波动；而股本大幅变动占比 1.9%，对应真实的资本运作事件，印证了数据的实时性。

5.2 最终结论

Choice 数据库在核心价格数据方面表现优秀，能够满足大部分量化策略的基础需求。