量化金融分析师（AQF®）全国统一考试

模拟题

适用场次：2024 年 9 月

使用本模拟题，您应该遵守：

1. 本模拟题仅提供给参加 2024 年 9 月份 AQF 全国统一考试的考生，考生仅可以出于准备个人考试的目的查阅和打印本模拟题；
2. 严禁出于任何目的的复制、网络发布和传播、抄袭本模考题内容，如有违反，可能导致违纪或违法行为；

量化金融标准委员会

Standard Committee of Quantitative Finance

量化金融分析师（AQF®）全国统一考试模拟题

说明：本场考试中的代码都应采用 Python 3.X 版本作答。

1． 不定项选择题（每题 2 分，本部分共 40 分）：有 1-5 个正确答案，全部选对得 2 分，少选得 1 分，选错或不选得 0 分。

1.1．关于 Python 内存管理，下列说法正确的是（）

A．变量不必事先声明

B．变量无须指定类型

C．可以使用 del 释放资源

D．变量无须先创建和赋值而直接使用

E．对象无须指定类型

参考答案：ABC

解析：不先赋值会报错 is not defined；对象必有类型。

1.2．现有变量 data，数据类型为 numpy.ndarray，具体数据如下：

```
data = np.array([0.2, 0.2, 1.3, 0.8, 0.6, 1.1])
```

以下可以将该一维数组转换为列表数据的命令是（）

A．list(data)

B．[i for i in data]

C．data.list()

D．data.tolist()

E．data.listed()

参考答案：ABD

解析：C 选项和 E 选项为无效命令，其他都可以转换。

1.3．关于基本面多因子模型，正确的是（ ）

A．基本面多因子模型的基本假设是具有不同 “属性” 的股票，在市场上应该有不同的收益率；

B．基本面多因子模型主要解释变量是可观察到的股票（上市公司）自身的基本属性，比如

市盈率、市值大小等；

C．基本面多因子模型的主要分析方法是进行横截面分析，以确定股票收益率对因子的敏感性（因子载荷）。

D．基本面多因子模型的主要分析方法是进行时间序列分析，以确定股票收益率对因子的敏感性（因子载荷）。

E．在实际操作中，基本面因子模型效果要明显好于其他两类模型，是现在的多因子模型研究的主流；

参考答案：BCE

解析：A 基本面多因子模型的基本假设是具有类似 “属性” 的股票，在市场上应该有相似的收益率；D 基本面多因子主要为横截面分析。

1.4．以下关于证券发行市场的表述错误的是（）

A．有统一时间

B．没有固定场所

C．证券发行价格与证券票面价格较为接近

D．有固定场所

E．证券发行价格与证券票面价格差异很大

参考答案：ADE

解析：证券发行市场主要是无形市场，通常不存在具体形式的固定场所，也无通常的专业设备和设施，无统一发行时间，证券发行价格与证券票面价格较为接近。

1.5．某研究员正在进行日内高频交易策略的研究，以下说法不正确的是（ ）

A.由于高频交易策略的交易频率非常高，因此对手续费等成本非常敏感

B.高频交易常用的行情数据包括分钟 K 线、盘口快照、委托队列、日成交持仓排名等

C.自动做市商高频交易主要是分析实时订单薄数据提供市场报价

D.高频交易的核心是交易速度，交易速度越快收益越高

E.高频交易会降低市场流动性和效率

参考答案：BDE

解析：B 选项，日成交持仓排名为低频数据。D 选项，交易速度是高频交易的核心要素，但交易速度越快不一定收益越高，还需考虑策略思路、交易成本等其他因素影响。E 选项，部

分高频交易策略可以缩小市场买卖价差，提高市场流动性，提升市场效率。

1.6．在投资过程中，基本面分析和技术分析都是重要的分析方法。某量化交易员正在研究多因子策略，并整理了一系列候选因子如下，请问哪些属于基本面因子类型（ ）

A． 市净率因子

B． 成交量分布因子

C． 价值成长因子

D． 资产收益率因子

E． 舆情关注度因子

参考答案：ACD

解析：基本面分析以证券的内在价值为依据，如公司营运能力、盈利能力、偿债能力等。技术分析通过分析关于股票价格和成交量的信息来获取超额收益。B 选项成交量分布因子、E 选项舆情关注度因子属于另类因子。

1.7．下列关于债券交易策略的说法中正确的是（ ）？

A． 卖出凸性策略是购买凸性小的债券，卖出凸性大的债券

B． 久期管理策略在预期利率下降时，增加组合久期获得收益

C． 持有到期策略属于被动管理策略

D． 消极的债券组合管理策略将市场价格假定为公平的均衡交易价格

E． 骑乘收益率曲线的投资者会购买比要求的期限稍短的债券

参考答案：ABD

解析：C 选项持有到期策略属于主动管理策略。E 选项骑乘收益率曲线的投资者会购买比要求的期限稍长的债券，然后在债券到期前售出，获取收益。

1.8． 某研究员在进行量化交易策略研究过程中，获取某股票的 k 线数据存储于变量 k_data 中，数据类型为 pandas.DataFrame，部分数据如下：

	date	open	close	high	low	volume	code
0	2023-01-03	17.00	17.47	18.39	16.91	113042.0	002577
1	2023-01-04	17.41	17.25	17.42	17.01	65463.0	002577
2	2023-01-05	17.03	17.00	17.44	16.98	49592.0	002577
3	2023-01-06	17.01	17.16	17.37	17.01	40664.0	002577

4	2023-01-09	17.16	17.03	17.16	16.70	34389.0	002577

他想删除“code”列，保留日期和高开低收价量数据，则以下代码中可以实现该目标的是（）？

A. del k_data[-1:]

B. k_data = k_data.iloc[:,:-1]

C. k_data = k_data.loc[:,:5]

D. k_data.drop('code', inplace=True)

E. k_data.drop('code', axis=1, inplace=True)

参考答案：BE

解析：

A 选项运行时会报错，错误类型为 TypeError；

C 选项 loc 应改为 iloc；

D 选项错误，使用 DataFrame 的 drop 方法删除列时需设置参数 axis=1。

1.9. 数据清洗是量化策略开发过程中不可缺少的一个环节，其结果质量直接关系到策略回测的准确度。因此，在数据分析之前，研究员往往会花费大量的时间来进行数据清洗工作。以下数据清洗的做法，正确的有（ ）

A. 数据录入过程、数据整合过程都可能会产生重复数据，为保持数据完整不应随意删除

B. numpy 中可以使用 fillna 方法替换缺失值数据

C. 在适当情况下，可以使用某个变量的样本均值、中位数或众数代替无效值和缺失值

D. 对于量级相差不大的数据进行 z-score 标准化处理

E. 对于数据中的离群值，单独进行标准化处理

参考答案：C

解析：C 选项为填充无效值缺失值的常见方法。

1.10. 以下不能创建一个字典的语句是？

A. dic1 = {}

B. dic2 = {2: 5}

C. dic3 = {3: ‘abcde’}

D. dic3 = {(1,2,3): ‘abcde’}

E. dic4 = {[1,2,3]: ‘abcde’}

参考答案：E

解析： 列表不能作为字典的键值。

1.11. 以下哪几句代码会导致 SyntaxError？

A. a=b=c=1

B. a=(b=c-1)

C. a-b=c

D. a, b=b, a

E. a -= b

参考答案：BC

解析： B 和 C 选项都会导致 SyntaxError。

1.12. 李明，某量化基金经理，正在研究多因子策略，信息系数（Information Coefficient，简称 IC）常用于衡量选股因子的有效性，关于信息系数的描述错误的是（ ）？

A. 信息系数最大值为 1，表示因子的预测能力 100%正确

B. 当样本股票过少时，IC 是没有统计意义的

C. Rank IC 值是对因子值排名和下期股票收益排名求相关系数

D. 信息系数的值越大，该因子越有效

E. 信息系数可以用来衡量股票价格的波动幅度

参考答案：D

解析：信息系数的绝对值越大，该因子越有效。当信息系数为负时，值越小越有效。

1.13. 李明作为一名量化分析师正在研究多因子策略。目前他已经得到了经过处理的、表现良好的因子池，以下哪个选项是他不愿意看到的（ ）？

A. 因子池中的因子与市场走势高度相关

B. 数据不包含离群值

C. 因子间相关系数较高

D. 某些因子在市值和行业上没有明显的偏向和集中

E. 缺失值已经被进行有效填补

参考答案：AC

解析：因子池中的因子与市场走势高度相关意味着这些因子在预测股票收益时可能受到市场整体走势的影响，而不是独立地反映股票的特定因素。如果因子间相关系数较高，那么它们可能会共同反映某个特定的因素，而不是独立地反映不同的因素。

1.14. 李明，某量化基金经理，在量化交易策略的回测过程中，发现回测收益往往会出现高于实盘收益，可能导致这种情况的原因有（ ）

A. 使用了未来函数

B. 策略模型过拟合或过度优化

C. 忽视了策略流动性不足的问题

D. 在交易过程中低估了交易成本

E. 存在幸存者偏差

参考答案：ABCDE

解析：上述所有情况都有可能导致回测收益高于实盘收益。

1.15. 李明，某量化基金经理，把某股票一周的收益率存储在变量 arr 中，arr = np.array([-0.06,-0.13,0.19,-0.10,0.02])，他现在想要判断这一周的股票收益率是否都小于 0，如果都小于 0 返回 True，否则返回 False。可以实现该功能的代码为（ ）？

A. np.all(arr<0)

B. np.any(arr<0)

C. np.where(arr<0)

D. np.sign(arr<0)

E. np.all(np.where(arr<0,True,False))

参考答案：AE

解析：A 选项，np.all()函数判断给定轴向上的所有元素是否都为 True，符合要求。E 选项，np.where(condition,x,y)满足条件 condition 输出 x，否则输出 y，则 np.where(arr>0,True,False)输出 arr>0 条件判断的结果，np.all()再次判断条件判断的结果是否都为 True。

1.16. 李明，某量化基金经理，查询 stock 表中所有股票名称 stock_name 中包含“银行”的股票情况，可用的语句有（ ）？

A. SELECT * FROM stock WHERE stock_name LIKE‘*银行*’

B. SELECT * FROM stock WHERE stock_name LIKE‘%银行%’

C. SELECT * FROM stock WHERE stock_name=‘%银行*’

D. SELECT * FROM stock WHERE stock_name =‘*银行%’

E. SELECT * FROM stock WHERE stock_name LIKE '银行'

参考答案：B

解析：B 选项使用了 LIKE 操作符，可以模糊匹配包含“银行”字符串的股票名。

1.17. SQL 语言按照用途可以分为以下哪几类？

A. DDL

B. DQL

C. DML

D. DCL

E. DEL

参考答案：ACD

解析：SQL 语言按照用途可以分为三类：DDL、DML、DCL。

1.18. 李明，某量化分析师，在使用机器学习技术预测股票走势时发现训练误差会降低模型的准确率，产生欠拟合，此时他应该怎么做来提升模拟拟合度？

A. 修改损失函数

B. 特征工程

C. 增加数据量

D. 提高模型复杂度

E. 减少正则化参数

参考答案：ABDE

解析：训练误差来自模型算法本身，和数据量大小无显著关联性。

1.19. 下列代码结果为 True 的是（ ）

```
a = ['t', 'e', 's', 't']
b = ['t', 'e', 's', 't']
```

```
c = 0
d = 1
```

A. a is b

B. a == b

C. bool(c | d)

D. bool(c & d)

E. bool(c)

参考答案：BC

解析：

对于 A 和 B，==判断储存的数据是否相同．而 is 用来判断是否来自同一个内存地址，内存地址可以用 id( )方法查询，而如此创建的列表内存地址不一样；

对于 C 和 D，and（&）表示两者中有必须都为真才为真，有一者为假便为假；or（|）表示两者中有一者为真便为真，都为假才为假，因此 C 为真，通过 bool 值转换结果为 True。

1.20. 李明，AQF，某量化基金经理，在进行蒙特卡洛模拟时想要产生 1000 行 5 列服从标准正态分布的随机数，以下选项正确的是（ ）？

A. numpy.random.randn(1000, 5)

B. numpy.random.random((1000, 5))

C. numpy.random.normal(size=(1000, 5))

D. numpy.random.uniform(size=(1000, 5))

E. numpy.random.standard_normal((1000, 5))

参考答案：ACE

解析：B 选项和 D 选项产生的是 1000 行 5 列的随机数，但是这些随机数并不服从标准正态分布。

2． 解答题（每题 20 分，本部分共 60 分）

1． 李明，AQF，某量化基金经理，为了更好地研究上市公司，抓取了各种不同类型的数据，

包括上市公司的市值、股票的换手率和最近有独立董事辞职的上市公司名单，分别保存在 cap.csv、turnover.csv 和独立董事辞职名单.csv 中。这些数据大致结构如下：

市值表（单位：万元）

trade_date	SH600000	SH600004	SH600006	SH600007	SH600008
20210601	29909835	2771427	1438000	2180767	2260902
20210602	29997892	2752493	1582000	2085075	2290264
20210603	30027244	2766694	1546000	1975281	2282924
20210604	30027244	2743026	1524000	1865487	2282924
20210607	30027244	2726459	1510000	1889662	2268243
20210608	30262061	2797461	1592000	1873546	2246221

换手率表（单位：百分比）

trade_date	SH600000	SH600004	SH600006	SH600007	SH600008
20210601	0.1427	0.864	2.9169	1.1141	2.4567
20210602	0.1221	1.0643	7.1879	1.5167	1.4343
20210603	0.178	0.8889	11.8561	1.5354	0.8067
20210604	0.2421	0.8071	6.1004	0.9025	0.6336
20210607	0.1237	0.7516	4.3879	0.6686	0.6158
20210608	0.1732	1.7018	7.3059	1.0391	0.5105

独立董事辞职名单：

603131 上海沪工
300356 ST 光一
002355 兴民智通
000611 *ST 天首
002643 万润股份
600338 西藏珠峰
600652 *ST 游久
603839 安正时尚

1.1. 读取“cap.csv”和“turnover.csv”文件中的数据，分别保存在 cap 和 turnover 变量中，设置 trade_date 为索引。并去除这两个变量中在该时间段内数据完全无效的股票。（3 分）

参考答案：

```
# 导入模块
import numpy as np
```

```
import pandas as pd
# 读入数据
cap =pd.read_csv('cap.csv', index_col='trade_date')
turnover = pd.read_csv('turnover.csv', index_col='trade_date')
cap = cap.dropna(axis='columns', how='all')
turnover =turnover.dropna(axis='columns', how='all')
```

1.2. 读取“独立董事辞职名单.csv”文件中的数据，无需指定索引但将列名设置为 code 和 name，保存在变量 dd 中。（2 分）

参考答案：

```
dd = pd.read_csv('独立董事辞职名单.csv', sep=' ', names=['code', 'name'],
dtype={'code': str})
```

1.3. 为了统一股票代码，请编写一个名为 changecode 的函数，实现去除股票代码中的'SH'或'SZ'的功能。然后用该函数改变 cap 和 turnover 中的股票代码（4 分）

参考答案：

```
def changecode(code):
    return code[2:]
cap = cap.rename(changecode, axis='columns')
turnover = turnover.rename(changecode, axis='columns')
```

1.4. 先计算每 10 个交易日的平均换手率,然后选出 2021/12/1 当日平均换手率不低于 5%且市值低于 500 亿的股票并放入变量 lst1 中，最后在里面筛选出同时也在独立董事辞职名单上的股票代码并放入变量 lst2 中。（6 分）

参考答案：

```
turnover_ma = turnover.rolling(10).mean()
lst_t  =  turnover_ma.columns[np.where(turnover_ma.loc[20211201,  :]  >=
5)].tolist()
lst_c = cap.columns[np.where(cap.loc[20211201, :] <= 5000000)].tolist()
lst1 = list(set(lst_t).intersection(set(lst_c)))
```

```
lst2 = list(set(lst1).intersection(set(dd['code'].tolist())))
```

（也可以用 merge 来做）

2. 李明，AQF，某量化研究员，为研究某公司财务指标而获取了若干数据如下。分为两张表。其中第一张表（price_data）：该表中存储若干家上市公司的基本信息，包括股票代码（sec_id）、日期（date）、开盘价(open)、最高价（high）、最低价（low）、收盘价（close）、成交量(volume)等若干字段，数据类型分别为 text、datetime、double、double、double、double、double。第二张表（stock_info）储存了上市公司的公司名信息，包括股票代码（code）、公司名（name）两个字段，数据类型分别为 text、varchar(8)。

price_data 表：

sec_id	date	open	high	low	close	volume
000008.SZ	2022/1/4	2.96	3.02	2.85	2.95	1.8E+08
000008.SZ	2022/1/5	2.9	2.95	2.81	2.85	1.18E+08
000008.SZ	2022/1/6	2.85	3.02	2.81	2.98	1.42E+08
000008.SZ	2022/1/7	2.94	2.94	2.85	2.86	1.03E+08
000008.SZ	2022/1/10	2.88	2.93	2.79	2.85	72836273
000008.SZ	2022/1/11	2.84	2.85	2.74	2.76	86010685
000008.SZ	2022/1/12	2.76	2.79	2.73	2.78	52597699
000008.SZ	2022/1/13	2.75	2.82	2.71	2.71	60180661
000008.SZ	2022/1/14	2.7	2.79	2.68	2.76	63627447
000008.SZ	2022/1/17	2.75	2.82	2.72	2.75	48318782
000008.SZ	2022/1/18	2.73	2.79	2.68	2.76	62609470
000008.SZ	2022/1/19	2.77	2.88	2.7	2.86	97930353
000008.SZ	2022/1/20	2.86	2.86	2.72	2.74	71152979
000008.SZ	2022/1/21	2.72	2.73	2.6	2.64	63983971
000008.SZ	2022/1/24	2.63	2.7	2.6	2.67	43009866
000008.SZ	2022/1/25	2.65	2.69	2.52	2.52	55710769

stock_info 表：

code	name
000001.SZ	平安银行
000002.SZ	万科A
000004.SZ	国华网安
000006.SZ	深振业A
000008.SZ	神州高铁

2.1. 写出 SQL 命令，统计出每只股票所有时间段的平均成交量和最高收盘价，并按平均成交量从低到高排列，要求列标题分别为：证券代码，平均成交量（2 分）

参考答案：

```
SELECT sec_id AS '证券代码', AVG(volume), MAX(close) AS '平均成交量' FROM price_data GROUP BY sec_id ORDER BY AVG(volume);
```

2.2. 请用 SQL 命令查询 2022 年 1 月 25 日开盘价最高的三只股票的证券代码，要求列标题为：证券代码、开盘价（3 分）

参考答案：

```
SELECT sec_id AS '证券代码', open AS '开盘价' FROM price_data WHERE date="2022/1/25" ORDER BY open DESC LIMIT 3;
```

2.3. 请用 SQL 实现查找 2022 年 2 月 7 日最高价最低的股票的公司名，要求列标题为：公司名称、最高价（4 分）

参考答案：

```
SELECT name AS '公司名称', high AS '最高价' FROM price_data AS p JOIN stock_info AS i ON p.sec_id = i.code WHERE date="2022/2/7" ORDER BY high DESC limit 1;
```

2.4. 由于缺乏成交金额信息，李明无法直接求得每只股票每日的均价，于是他将某日开盘价、最高价、最低价和收盘价的算术平均作为该股票当日的简易均价。请用 SQL 实现查找 2022 年 1 月 4 日至 2022 年 1 月 14 日之间平均简易均价大于与所有股票平均简易均价的股票信息，要求列标题为：证券代码，简易均价、成交量（6 分）

参考答案：

```
SELECT sec_id AS '证券代码', AVG((open+high+low+close)/4) as '简易均价', volume as '成交量' FROM price_data WHERE date BETWEEN "2022/1/4" AND "2022/1/14" GROUP BY sec_id HAVING AVG((open+high+low+close)/4) > (SELECT AVG((open+high+low+close)/4) FROM price_data WHERE date BETWEEN "2022/1/4" AND "2022/1/14");
```

3．李明，某量化基金经理，正在测试机器学习模型的预测能力。他认为，根据个股的估值指标和对数市值大小可以判断出个股是否存在投资机会。他将个股市净率、市盈率和对数市值作为模型的特征数据储存在文件 data_x.csv 中，将个股当年的收益率数据储存在文件 data_y.csv 中。

data_x 部分数据如下：

code	pb	pe	log_mv
600000.SH	0.3714	4.6834	17.1949
600004.SH	-2.1082	-77.0214	14.7886
600006.SH	1.5435	16.9126	14.1939
600007.SH	2.05	21.0572	14.4083
600008.SH	1.1716	13.382	14.6083

data_y 部分数据如下：

code	Return
600000.SH	0.0331
600004.SH	-0.2088
600006.SH	-0.2267
600007.SH	0.4726
600008.SH	0.0888

现需要你根据以下步骤描述，编写相应代码。

3.1．在使用机器学习模型前需要对特征数据进行预处理，比如常见的数据标准化，其公式为：

$$x^* = \frac{x - \bar{x}}{\sigma}$$

使用上述公式对 data_x 中的特征数据进行标准化，并将处理好的数据保存在变量 data_1 中。（3 分）

参考答案：

```
# 读取数据
data_x = pd.read_csv('data_x.csv', index_col='code')
# 标准化
```

```
data_1 = (data_x - data_x.mean()) / data_x.std()
```

或者

```
# 定义函数
def standardize(x):
    return (x - x.mean())/x.std()
data_1 = data_x.apply(standardize)
```

3.2. 将个股收益率数据 data_y 转换成是否 80%分位数以上的标签数据，并保存在变量 data_2 中。（2 分）

参考答案：

```
# 读取数据
data_y = pd.read_csv('data_y.csv', index_col='code')
# 处理标签
low, up = np.quantile(data_y, (0.2, 0.8))
data_2 = data_y.copy()
data_2[data_2 > up] = 1
data_2[data_2 <= up] = 0
```

或者

```
data_2 = pd.DataFrame(0, index=data_y.index, columns=data_y.columns)
data_2[data_y > up] = 1
```

3.3. 将特征数据和标签数据进行分组，其中 80%作为训练集，20%为测试集，分别将训练集和测试集的特征数据和标签数据存储于变量 train_x, train_y, test_x, test_y 中。（3 分）

参考答案：

```
from sklearn.model_selection import train_test_split
train_x, test_x, train_y, test_y = train_test_split(data_1,
data_2 .values.flatten(), test_size=0.2)
```

或者

```
train_x, test_x, train_y, test_y = train_test_split(data_1, data_2,
```

```
train_size=0.8)
```

3.4. 使用随机森林模型进行模型训练。（3 分）

参考答案：

```
from sklearn.ensemble import RandomForestClassifier
RFC = RandomForestClassifier()
RFC.fit(train_x, train_y)
```

3.5. 使用测试集数据对模型进行测试，根据测试集特征预测标签，并将预测的结果保存在变量 predict_y 中。然后根据以下公式计算模型准确率：

模型准确率=测试集中预测正确的次数/总的预测次数。（4 分）

参考答案：

```
predict_y = RFC.predict(test_x)
accuracy = (test_y == predict_y).sum()/len(test_y)
```

4. 李明，AQF，某量化研究员，在研究期货持仓数据，其中部分数据如同所示，保存在“future_broker.csv”中。其中各字段分别表示日期、合约代码、经纪商标记、交易量、交易量的变化、多头持仓、空头持仓、品种代码。数据类型分别为 int、str、str、float、float、float、float、str。

trade_date	symbol	broker	vol	vol_chg	long_hld	short_hld	fut_code
20220701	CU2212	X	60	-50			CU
20220701	CU2211	O	224	17			CU
20220701	CU2211	L	159	44	450	1550	CU
20220701	CU2211	S	134	-25	1067	1648	CU

请按要求完成以下操作的相关代码。（已按惯例导入相关依赖库：import pandas as pd; import numpy as np; import matplotlib.pyplot as plt）

4.1. 读取该期货持仓数据，不设置索引，保存到变量 df 中，然后筛选出经纪商（broker）

X 和 Y 的铜（CU）数据，保存为变量 df_XY。（3 分）

参考答案：

```
df = pd.read_csv('future_broker.csv')
df_XY = df[(df['fut_code'] == 'CU') & ((df['broker'] == 'X') |
(df['broker'] == 'Y'))]
```

4.2. 筛选出 df_XY 中 2022/08/08 的数据，并计算其每个经纪商交易量的变化量的绝对值与交易量之比，然后将其均值保存到变量 xymean 中。（3 分）

参考答案：

```
df_XY1 = df_XY[df_XY['trade_date'] == 20220808]
vol_chg = df_XY1 ['vol_chg'].abs() / df_XY1 ['vol']
xymean = vol_chg.mean()
```

4.3. 分别计算 df 中金（AU）每天的多头持仓和空头持仓的总和，并计算两者的平均数，保存到变量 hold_all 中；然后画出直方图。（5 分）

参考答案：

```
df_AU = df[df['fut_code'] == 'AU']
hold_all=(df_AU.groupby('trade_date')['long_hld'].sum()+df_AU.groupby('
trade_date')['short_hld'].sum()) / 2
hold_all.hist()
plt.show()
```

4.4. 计算所有持仓数据的多头持仓和空头持仓之和，然后按日期和经纪商进行分类统计求其最大值，以 trade_date 为索引，以经纪商代码为列，组成一个新的 dataframe 保存到变量 df_hold 中。（4 分）

参考答案：

```
df['hold'] = df['long_hld'].fillna(0) + df['short_hld'].fillna(0)
df_hold = pd.pivot_table(df, index='trade_date', columns='broker',
values='hold', aggfunc=max)
```