卓识基金数据测评项目设计思路

令狐冠杰

首先用 pandas 将 parquet 文件读入 dataframe。

1. 数据检查
    - 统计 LocalTime 的单调性：可使用 pandas series 自带方法 is_monotonic 直接进行判断，可与 unique 方法一起使用对严格单调性进行判断。
    - 分交易所统计（UpdateTime，UpdateMullisec）单调性：单独的 UpdateTime 和 UpdateMullisec 的单调性没有实际意义，需要和日期放在一起看。将 ActionDay，UpdateTime 与 UpdateMullisec 转换为 19 位时间格式并相加，然后存为一个新字段 UpdateTime_19，使用 19 位时间格式有利于之后直接将 LocalTime 和 UpdateTime 进行对比。将数据根据交易所字段 ExchangeID 分组后就可以用与第一步同样的方法对 UpdateTime_19 进行单调性分析了。或者可以根据 ActionDay 以及交易所字段分组检查每天的单调性。
    - 探索每个交易所 tick 的推送频率：首先根据交易所字段 ExchangeID 分组，然后可以通过 diff 方法获得推送时间差值再通过 nunique 等方法探索推送频率的规律。
    - 探索 ActionDay，TradingDay 和实际交易时间的关系：首先筛选出 ActionDay，TradingDay 及 LocalTime 日期不相符的数据，然后根据和交易所字段 ExchangeID 以及合约字段 InstrumentID 分组进行进一步探索。
    - 尝试对现有数据进行数据检查：对每个字断是否存在缺失值以及缺失规律进行检查，对几个时间变量之间的关系进行检查，检查价格和成交量中是否有不可能出现的异常值，检查最高限价和最低限价之间是否匹配。在以上数据检查的基础上可以根据交易所或者合约名称分组进行更细致的检查。
2. 分钟 bar 合成
    - 创建一个 bar 的 class 并定义开盘价，最高价，最低价，收盘价四个属性。

- 创建期货名称哈希表，每个期货名称对应一个 bar 对象。
- 创建一个函数用于更新分钟 bar，具体为：
    - 没有初始 bar 时创建第一个 bar 对象并设为新的一分钟。
    - 已经有 bar 并且当前 tick 的时间（LocalTime）与当前 bar 不在同一分钟（可以加入对小时的判断来处理边角情况）：
        - 创建新的 bar 对象，将上一分钟 bar 对象的数据存入 dataframe，并将当前 bar 对象设为新一分钟。
    - 如果是新一分钟的 bar，根据当前 tick 数据初始化 bar 对象
    - 如果不是新一分钟的 bar，根据当前 tick 数据更新 bar 对象的最高最低价以及收盘价（如果不需要实时展示也可以分钟结束再更新收盘价）
- 使用 dataframe 进行存储

以上仅是思路，还没有进行实际操作。