Week 2: 数据预处理与可视化

检查和处理空数据

股票分析任务中，使用拟合值来填充空数据对精准的预测结果是有害的，因此应当删除空数据条目。

数据合并与时间对齐

我们获取的5年历史数据的时间精度参差不齐，你需要将所有的历史数据合并到一起，使得所有的数据都具备同样的时间精度。

> 难点：不同时间尺度的Volume如何处理?

辅助数据广播

除了历史股价以外，现金流、资产负债等其他数据都需要融合到你的特征之中。但我们获取的这些数据都是季度甚至年度的。因为这些数据都由每季度发布的财报披露，并且被一些第三方项目手动录入为数据（比如yfinance）

作为辅助预测的数据，即使两年的时间尺度只有8份财报，但依旧需要把财报的数据广播到每个季度的时间段里，过去3-5年的数据可能需要自行查询企业财报获取,并且准备通过后续的特征工程简化数据维度并发掘其关联性。

历史波动率

根据对数价格变动法公式：

$$X_i = In\frac{P_{i+1}}{P_i} = InP_{i+1} - InP_i$$

$$\bar{X} = \frac{1}{N}\sum X_i$$

$$\sigma = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}$$

计算1d精度的，2年时长的所有历史波动率，并同历史价格数据的box图一同绘制为plot

找寻波动率出现异常时，NVDA（即NVIDIA）是否有对应的新闻/事件支撑。

思考

面对时间分布不均的数据（越久远的数据量越小），在进行特征工程和窗口采样的时候，应当如何处理数据比重的关系?