题目：基金间相关性预测

背景：基金间的相关性是基金的重要特征。根据金融学原理，一个基金组合的整体风险，不仅和其中各只基金的风险水平有关，还和这些基金之间的相关性有关。构造一个基金之间相关性小、或者说分散程度高的基金组合，能在保持一定收益水平的基础上，降低整体风险。因此，对基金之间的相关性进行预测，有助于我们构建一个好的基金组合。

任务：参赛者需要根据给出的基金净值、基金业绩比较基准、对应指数行情、基金间相关性等数据，构建模型、算法进行训练。然后针对我们提供的测试样本，通过您的算法或模型预测出之后一段时间内基金间的相关性情况。

数据背景：赛题数据包括一批公募基金的复权净值收益率，各只基金对应的业绩比较基准的收益率，基金间的相关性数据，同时期的重要市场指数收益率等。

收益率数据均为日度市场数据，直接获取于市场上公布的相关数据。基金间的相关性数据基于从对应日期开始向后61个交易日的市场数据和运营权重综合统计得出，按日度展示。例如，2015-09-30对应的相关性数据，实际上是基于从2015-09-30到2015-12-30的多项数据统计得到的。

参赛选手需要基于这些数据，预测出测试数据期的下一个时间点、即2018-03-19对应的基金间的相关性。

数据说明：

1、训练数据：训练数据包括四个csv文件。相关性数据的对应日期和收益率数据的对应日期错开一个交易日，便于形成一个预测性的数据划分。

（1）train\_fund\_return.csv：基金复权净值收益率。第一行是交易日期序列，从2015-09-29到2017-05-23共400个交易日；第一列是基金序号，共200只基金（训练和测试的所有数据中，相同序号的基金都是指同一只基金）。

（2）train\_fund\_benchmark\_return.csv：基金业绩比较基准收益率。第一行是交易日期序列，从2015-09-29到2017-05-23共400个交易日；第一列是基金序号，共200只基金。

（3）train\_index\_return.csv：重要市场指数收益率。第一行是交易日期序列，从2015-09-29到2017-05-23共400个交易日；第一列是各个指数的名称和代码，包括股票指数、债券指数和商品指数三类，共35个指数。

（4）train\_correlation.csv：基金间的相关性。第一行是相关性对应的交易日期序列，较三个收益率数据的日期序列向后推进一个交易日，即从2015-09-30到2017-05-24共400个交易日；第一列是两只不同基金组成的基金对名称，共19900个基金对。

2、测试数据：测试数据同样包括四个csv文件，形式与训练数据基本一致，截取的是训练数据后一个交易日起的对应数据。一个较大的区别是相关性数据，日期长度比另外三个收益率数据的日期长度短61个交易日，原因在于，某个日期下的相关性需要向后61个交易日的数据进行统计，在拟真情境下存在不能获得全部统计数据的限制。

（1）test\_fund\_return.csv：基金复权净值收益率。第一行是交易日期序列，从2017-05-24到2018-03-16共200个交易日；第一列是基金序号，共200只基金。

（2）test\_fund\_benchmark\_return.csv：基金业绩比较基准收益率。第一行是交易日期序列，从2017-05-24到2018-03-16共200个交易日；第一列是基金序号，共200只基金。

（3）test\_index\_return.csv：重要市场指数收益率。第一行是交易日期序列，从2017-05-24到2018-03-16共200个交易日；第一列是各个指数的名称和代码，与训练数据中给出的指数相同，共35个指数。

（4）test\_correlation.csv：基金间的相关性。第一行是相关性对应的交易日期序列，从2017-05-25到2017-12-14共139个交易日；第一列是两只不同基金组成的基金对名称，共19900个基金对。

3、其他数据

trading\_date.csv 覆盖训练、测试的交易日期序列，用于对齐等工作

专业名词解释：

1、基金复权净值——基金会有分红和拆分的行为，因此单位净值并不能准确反映持有基金实际获得的收益率。复权净值是对基金的单位净值进行了复权计算，对基金的分红或拆分因素进行了综合考虑，计算出基金没有进行任何分红或拆分情况下的历史净值，因此更能反映出持有基金的实际收益情况。

基金业绩比较基准——基金成立时通常会设定业绩比较基准，投资者可以通过基金收益率与基金业绩比较基准收益率的差异来衡量基金业绩，这是比直接查看基金收益率更科学的衡量基金投资水平的方式。每只基金都会根据自己的投资目标、投资策略设定符合自身情况的业绩比较基准，通常由若干指数按比例组成。

2、简单的相关性根据两只基金在一段时间内的收益率序列即可计算得到，其刻画了两只基金净值涨跌的相似程度。赛题中的相关性计算，不仅使用了两只基金的收益率序列，同时考虑了一些权重系数，例如基金产品运营中每个交易日的在账规模，规模的不同会导致每个交易日的数据的重要程度不同。但由于我们希望相关性的金融涵义尽量清晰、普适，因此没有过分偏向于使用这类权重系数进行调整。

3、时间序列的处理具有自己的一些特点，比如使用rolling window、walk forward之类的方法进行分析。参赛者完全可以根据交易日顺序合并两个数据集，自己处理train-validate-test。

4、首先，测试数据期是从2017-05-24到2018-03-16。2018-03-16是周五，下一个交易日是2018-03-19。其次，数据背景中，我们做了说明：“基金间的相关性数据基于从对应日期开始向后61个交易日的市场数据和运营权重综合统计得出，按日度展示。例如，2015-09-30对应的相关性数据，实际上是基于从2015-09-30到2015-12-30的多项数据统计得到的。”因此，2018-03-19对应的相关性，实际上是基于2018-03-19到2018-06-15的多项数据统计得到的。

任务：

第一阶段：

1. 相关论文收集及研读（看别人的模型以及模型的优缺点，做简单的ppt在组里共享，简单描述模型即可，能看懂即可）：2人，每人大概看2-3篇；
2. 两个人学习R，三个人学习python；
3. 搭建GitHub小组协作环境，写一份协作教程，以后的工作在这个平台上进行：1人
4. 时间序列预处理和分析学习：2人

第二阶段：（数据分析及可视化）

1. 数据分析（3人）：包括选择数据分析的维度，实现，数据分析的报告（片段，说清楚即可，修辞逻辑那些后面再加就行。感觉不一定进行相关性分析，可以根据这些数据做一个趋势分析其他的都行，多角度）
2. 可视化（2人）：数据分析结果可视化

第三阶段：

1、相关性预测