进一步深入实验的研究内容主要包括：

数据清洗对其下游的聚类应用产生了什么影响，将这一个影响具体化和量化，明确影响环节和影响程度。实际上就是回答数据质量的“什么”影响到了聚类的“什么”。

该问题内容较为庞大，为便于更好地规划研究的时间和资源，我将其按照聚类方法的执行环节拆分成了 4 个子问题，争取做到环环相扣和逐步深入。

主要内容部分是一个与之相关的初步思路，没有具体想好，我认为整个该部分内容应该够单独写一篇论文的了，比如最后的课题可以是**数据质量对下游聚类（性能/超参数）影响的（系统分析/解释/机理探究）**。

正如老师所说，这一任务难度很大，也是我做完前序任务后才能一点点摸索出来的。

子问题 1：清洗对数据分布和属性结构的影响

逻辑地位：

这是最先要考察的一步，同时也有一些文献可以参考。只有先了解清洗操作如何改变数据的**整体分布（如均值、方差、异常值比例）**，才能顺势探究它对聚类内部或后续评价带来的连锁影响。

主要内容：

- 量化并分析数据清洗对**基础分布统计**（均值、方差、极值等）以及**整体分布距离**的影响。
- 比较不同清洗策略在修复过程中的“力度”与“准确度”，重点关注异常/缺失/噪声等错误类型。（这里如何量化是一个问题，目前只是把问题提出来了，还没有具体设计）
- 探索清洗如何改变特征间关联结构，可能使某些维度变得更相关或更无关。
- 利用可视化与案例剖析，展示清洗如何“移动”样本在属性空间的位置，从而为后续对聚类算法、聚类指标、聚类参数的影响提供依据。

难度：C，作为后续研究的**基础**，理解清洗如何改变数据分布。实验设计较直接，分析也多为常规统计和可视化。

子问题 2：清洗对聚类算法内部过程的干预

逻辑地位：

在清楚分布如何变化后，进一步考察“算法本身”被数据清洗所干扰或帮助的具体机制。即探

究清洗后的数据是否让 K-Means 收敛更快/慢、是否让 DBSCAN 识别核心点更容易、以及对层次聚类的合并/分裂过程有何影响。

主要内容：“内部过程”主要分成以下几个角度，分析清洗是如何干预这些步骤的

- 质心型算法（K-Means / GMM）的迭代与收敛路径
- 密度型算法（DBSCAN / OPTICS）的核心点、边界点判定
- 层次聚类（Hierarchical Clustering, HC）的合并/分裂顺序

难度：A，需在算法源码或进程中插入跟踪点记录，或通过统计/可视化方法来分析收敛轨迹、核心点判定等。技术实现比单纯分布分析更复杂，需要对聚类算法的机制有深刻的理解。

子问题 3：清洗对聚类评价指标的影响及量化

逻辑地位：

当我们知道**数据分布**和**算法机制**都可能被清洗所影响时，下一步则是聚焦“结果评估”层面：清洗是否真正提升（或降低）了聚类质量在各种评价指标（DB 指数、Silhouette、CH 指数等）上的表现？这是从**结果角度**验证清洗给聚类带来的**实际收益与潜在副作用**。

主要内容：

- 在无监督场景下，通过常见的聚类内部评价指标（DB 指数、Silhouette、CH 指数、Dunn 指数等），检查**清洗在多大程度上改变了聚类整体质量**。是否能通过某些**分布差异度量**（如 KL 散度）**来预测指标提升幅度。**
- 不同评价指标对同一种清洗策略是否**同向**（都提升）或产生**分歧**（有的提升、有的下降），**从而说明哪类指标对数据分布/错误修复最敏感。**
- 尝试将“数据分布改变”或“算法内部过程”与评价指标变化联系起来，探究**是否有某些分布距离量**或**核心点数量**与评价指标提升呈正相关。

难度：B，评价指标是衡量聚类质量的重要手段，能直观体现清洗是否带来收益。需要在多个评价指标上进行对比，还可能涉及不同数据集、错误率档次的横向分析，数据处理量和可视化需求更大。

子问题 4：清洗对聚类超参数选择与搜索过程的影响

逻辑地位：

最后是对“超参数层面”的研究，这也是最深入和最综合的一步。它需要依赖前面(1)~(3)的分析来理解何时参数会因为数据分布或算法内部变化，而在自动搜索时出现系统性偏移或需要

重新调优。**也就是说，4 是前三个的整合，因为参数受算法运行状态，数据分布等很多因素影响，它是水到渠成的一步，同时也需要单独做实验探究。**若在自动化管线中进行大规模搜索和调参，**此部分最能体现清洗对聚类性能和效率的综合影响。**

主要内容： （目前还没有什么思路）

- 量化清洗在多大程度上改变了聚类算法的最优超参数
- 揭示清洗后超参数搜索曲线或收敛过程
- （…待补充，可能后续做到这里再想出来吧）

难度：S，聚类中超参数选择往往在实际应用中难度最高、影响最深。能系统性解释清洗如何影响超参数，对**自动化管线**的落地价值极大。需要在清洗前后分别进行超参数搜索、阈值分析等高级实验，还要比较二者的最优参数是否有显著偏移，并寻找机理。对实验量、数据分析和解释能力要求都很高，依赖 (1)~(3) 的结果才能得出更完整机理推断。