我在本次实验中，更准确设计了数据集的错误类型和错误注入，引入更详细的统计指标评估清洗后数据集，并试图挖掘这些统计量和聚类结果之间的关系。目前已经整理好了实验的统计结果，正在进行详细地分析，以下是本次的实验设计（文字）。

**数据集准备：**

选用beers, flights, hospital, rayyan四个领域的数据集，然后注入不同的错误类型和该类型下的错误率形成测试数据集。

**错误注入：**

考虑到数值类型和字符类型的结构性差异，将这两种类型的数据分开，分别注入错误。其中数值类型注入异常值和缺失值，字符类型注入格式错误和知识库错误，此处在考察错误类型是否会对清洗和聚类的结果产生影响。为了使后续的各种清洗算法正常进行，四种错误率之和一般不超过20%，模拟常见的错误情况。需要注意，理论注入的错误值和实际检测到的单元格差异可能有出入，这里将两种定义方法（元组注入错误和单元格错误率）都保留。

**清洗算法：**

一共9种，多于最开始的3种，目的是更深入分析清洗方法（或数据质量）对源数据的影响

CLEANING\_METHODS = ["mode", "bigdansing", "boostclean", "holoclean", "horizon", "scared", "baran", "Unified", "google gemini"]

简单填充，统计修复方法，逻辑修复方法，朴素的ML方法，还加入了Google的LLM大模型修复（这个还没做完）

**聚类方法：**

在原有6种聚类方法基础上又找到了两个改进KMEANS方法，尚未更新完。

**统计指标：**

目标是分析清洗后数据本身的分布、异常率、相关结构，并查看与脏数据的差异度量(如多维Wasserstein、KL 散度、离群点变化、相关度变化等)。

回答问题：清洗方法是否大幅改动数据(差值大)或仅细微改动(差值小)；某指标变动是好是坏（需结合是否离真是更近、或下游任务效果）。

**数值列绝对统计**（清洗后）

例如 mean\_colX, var\_colX, normal\_p\_colX：对“清洗后数据”某列的均值、方差。

同样也会对“脏数据”该列计算 mean\_colX\_dirty, var\_colX\_dirty, normal\_p\_colX\_dirty 等。

**离群点比例**

分别用 IsolationForest 检测“脏数据”的离群点比例(outlier\_ratio\_dirty)和“清洗后”的离群点比例(outlier\_ratio\_rep)。

二者之差 outlier\_ratio\_diff = rep - dirty 用于衡量清洗是否大幅移除/增加异常点。

**列间相关度**

在脏数据和清洗后数据上分别计算数值列的平均绝对相关(corr\_avg\_dirty, corr\_avg\_rep)与最大绝对相关(corr\_max\_dirty, corr\_max\_rep)。再得到差值 corr\_avg\_diff、corr\_max\_diff。

**多维 Wasserstein**

multi\_wasserstein：将脏数据和清洗后数据的所有数值列合并为多维向量做 Earth Mover’s Distance。这本身就是 “脏 vs. 修复”的**分布差异**指标。

**数值列 K-L 散度**

分别针对**每个数值列**计算 KL(dirty||repaired) 后平均为 kl\_numeric\_avg，用来衡量“脏数据 vs. 清洗后”的整体分布变化。

**类别列 K-L 散度**

cat\_kl\_avg：对所有共享的类别/文本列做频次对比，计算 K-L 散度并取平均，衡量分类分布上的改动大小。

通过上述操作，每个 (task\_name, num, dataset\_id, cleaning\_method) 场景会在输出中包含：

* **脏数据**绝对指标：outlier\_ratio\_dirty, mean\_colX\_dirty, var\_colX\_dirty, normal\_p\_colX\_dirty, corr\_avg\_dirty, corr\_max\_dirty 等。
* **清洗后**绝对指标：outlier\_ratio\_rep, mean\_colX\_rep, var\_colX\_rep, normal\_p\_colX\_rep, corr\_avg\_rep, corr\_max\_rep。
* **二者差异**：outlier\_ratio\_diff, mean\_colX\_diff, var\_colX\_diff, normal\_p\_colX\_diff, corr\_avg\_diff, corr\_max\_diff。
* **直接“脏 vs. 修复”分布度量**：multi\_wasserstein, kl\_numeric\_avg, cat\_kl\_avg。

既能看清洗后数据本身的分布、异常率、相关结构，又能查看与脏数据的差异度量(如多维Wasserstein、KL 散度、离群点变化、相关度变化等)。

聚类指标：

DB 指数、Silhouette、CH 指数、Dunn 指数