《Python 语言及应用》课程项目要求

根据要求，独立完成课程项目设计。

题目：基于 Python 的 xxx 数据分析

一、项目主题与目标

- 主题：自选数据集的 Python 基础数据分析项目（**仅用标准库与基础语法**）。
- 目标：利用文件读写、字符串与列表/字典操作、控制结构、函数等本学期内容，对数据进行不少于 5 个维度的基础分析，输出文本结果与简单总结。

二、数据与范围

- 数据集格式：建议使用 CSV 或 TSV 等纯文本格式；不使用二进制或复杂格式。
- 规模与字段：建议至少 200 条记录、至少 4 个字段（含数值/类别/日期/文本等任意组合）；若数据量较小，需确保能开展统计分析。
- 合规与来源：注明数据来源与许可；如含个人信息必须脱敏。

三、功能与分析要求（仅基础 Python；至少 5 个维度）

必做项：

1. 数据读取与基本清洗，例如：
   - 使用内置文件操作 open 读取文本数据。
   - 自行实现行解析（split）、去除空行、去除首尾空白。
   - 字段类型转换：将数值字段转为 float 或 int；日期字段用字符串或手动解析校验（如 YYYY-MM-DD 基本格式检查）。
   - 简单异常处理：跳过或标记无法解析的行，并统计异常行数量。
2. 基础描述统计（文本输出），例如：
   - 对至少 2 个数值字段输出：计数、总和、均值、最大值、最小值。
   - 对至少 1 个类别字段输出：不同类别的频数统计（如字典计数）。
   - 不使用任何第三方库，手写循环与聚合。
3. 对于数据进行至少 5 个维度的分析（**鼓励创新性的维度**），例如：
   - 如按类别字段分组，计算每组的记录数与某数值字段的总和或均值。
   - 若数据有日期字段，可按年份或月份（从字符串中切片提取）进行聚合统计。

- 对某数值字段按降序排序，输出 Top-N 记录（如 Top-5）。
- 或按类别频数排序，输出占比最高的前 N 个类别（占比需自行用总数计算）。

进阶任选至少 1 项（例如）

- 简单趋势或估计：基于时间字段按月份序列计算均值或总和，用最后 K 期平均作为下一期的估计值。
- 简单相关性近似：对两个数值字段，计算它们的协动趋势的粗略指标（如手写均值与方差后计算皮尔逊相关系数；允许近似或仅输出“随同上升/下降”文本判断）。
- 文本关键词统计：对描述字段按空格拆分，清理常见停用词（自定义列表），统计 Top-N 高频词。

四、技术与实现规范（仅用基础 Python 与标准库）

- 允许使用的模块：仅 Python 标准库模块，如 os、sys、csv、datetime、math、statistics、argparse、logging、json 等。不得使用 pandas、numpy、matplotlib、seaborn、scikit-learn 等第三方库。
- 代码组织：
  - main.py 为入口文件；功能模块化为函数（如 load_data、clean_data、describe_stats、group_aggregate、validate_rules、top_n 等）。
  - 关键函数需有 Docstring；适当注释；合理命名。
  - 异常处理：对文件不存在、数据格式错误、空值等进行 try-except 与友好提示。
- 输入输出：
  - 输入文件路径可通过命令行参数或配置文件提供（argparse 或读取 json 配置）。
  - 输出为终端文本与必要的 TXT/CSV 结果文件（如统计结果导出）。
- 目录建议：
  - 学号姓名/（根目录）
    - src/（含 main.py 与其他.py）
    - data/（原始数据与示例）
    - outputs/（导出的文本结果）
    - README.md（使用说明）
    - requirements.txt（可写“仅标准库，无第三方依赖”）
    - screenshots/（终端输出截图）

五、报告要求

使用课程模板撰写，内容包括：

- 项目概述：数据来源、目标与问题定义。

- 数据说明：字段、类型、样本量、清洗策略、异常行统计。
- 方法与流程：各函数模块与处理步骤说明。
- 分析结果：以文本或表格（Word 内置表格即可）呈现统计、分组、Top-N、规则校验结果与简要解释。
- 关键代码片段与说明：展示核心函数实现与思路。
- 结论与改进：核心发现、数据局限、后续可扩展点。
- 运行实例：终端输出截图与结果文件示例。
- 合规与诚信说明：数据脱敏与引用。

六、提交与截止

- 提交方式：压缩为学号姓名.zip（例：12345678 张三.zip），内含全部代码、数据样例、报告、运行截图。报告需要另外提交纸质版。
- **截止时间：2025 年 11 月 30 日 21:00（电子版与纸质版按班级通知提交至学习委员）。**
- 诚信要求：不得抄袭或未标注地使用生成式工具产出；如使用，请在报告与代码中明确标注并说明你的改写与理解。

七、评估标准

- 功能与正确性：各必做项运行正确，结果合理，异常处理到位。
- 分析覆盖：不少于 5 个维度，分组、Top-N、规则校验等完成度。
- 代码质量：结构清晰、注释与 Docstring、模块化、标准库使用规范。
- 报告规范：条理清晰、结果与解释一致、运行截图完整。
- 复现与规范：目录与使用说明完备，无第三方依赖。
- 创新加分：进阶项、命令行交互、自动导出结果、日志与配置支持等。