练习题目

第一阶段：数据预处理与特征工程（对应你所学的基础计算、分组计算、数据转换等）

1. **缺失值处理**
   - 查看数据集中各数值列的缺失值比例
   - 分别使用**均值、中位数、分组填充（按城市 + 季节）** 填充缺失值，对比三种方式的差异
   - 最终选择分组填充作为缺失值处理的最终方案，生成新列平均气温_填充、降雨量_填充
2. **异常值修正**
   - 对平均气温_填充、`PM2.5`列使用**95 分位数法**和**3σ 原则**识别并修正异常值，生成平均气温_清洗-、PM2.5_清洗列
   - 可视化异常值（箱线图），验证修正效果
3. **基础统计与分组计算**
   - 计算所有数值列的均值、中位数、标准差、最大值、最小值（使用`describe()`和单独统计函数）
   - 按**城市**分组，计算各城市的平均气温、降雨量、PM2.5 的均值和标准差
   - 按**城市 + 季节**分组，计算各分组的湿度中位数，并将结果广播到原数据集（生成城市季节湿度中位数列）
   - 统计各空气质量等级的数量（使用`crosstab`或`groupby+size`）
4. **特征工程（数值 + 分类）**
   - **数值特征衍生**：
     - 将平均气温_清洗离散化为「低温（<0）、中温（0~25）、高温（>25）」，生成气温等级列
     - 计算各城市的年均温（按城市 + 年份分组），生成城市年均温列
     - 计算当前气温与城市年均温的差值，生成气温偏差列
   - **分类特征处理**：
     - 对季节列进行**标签编码**（Label Encoding），生成季节编码列
     - 对城市、天气类型列进行**独热编码**（One-Hot Encoding）
     - 使用`map`函数将空气质量映射为数值：优→4，良→3，轻度污染→2，中度污染→1

5. **条件筛选与计算**
    - 筛选出 2024 年夏季平均气温 > 30°C的城市数据
    - 统计各城市重度污染（这里用中度污染替代）的天数占比
    - 计算每个季节的降雨量总和，并排序
6. **排序与排名**
    - 按PM2.5_清洗降序排序，取前 10 条数据
    - 按城市分组，取每组中降雨量最大的一条数据
    - 对湿度列进行密集排名（dense），生成湿度排名列

第二阶段：数据可视化（折线图、柱状图、饼图、散点图）

要求：使用matplotlib或seaborn绘制，每个图需包含标题、坐标轴标签、图例（如需），并保存图片。

1. **折线图**
    - 绘制北京、上海、广州三个城市 2020-2024 年的年均温变化趋势（x 轴：年份，y 轴：年均温，多条折线）
    - 绘制 2024 年全国平均 PM2.5 浓度随季节的变化趋势（x 轴：季节，y 轴：PM2.5 浓度）
2. **柱状图**
    - 绘制 2024 年各城市的平均降雨量对比（横向柱状图）
    - 绘制各空气质量等级的数量分布（纵向柱状图）
    - 绘制不同季节的气温等级分布（堆叠柱状图：x 轴季节，y 轴数量，不同颜色代表气温等级）
3. **饼图**
    - 绘制 2024 年全国天气类型的占比饼图（显示百分比）
    - 绘制北京市空气质量等级的占比饼图（按年份划分，子图形式展示 5 年数据）
4. **散点图**
    - 绘制平均气温与湿度的散点图（x 轴：气温，y 轴：湿度，颜色区分季节）
    - 绘制 PM2.5 与降雨量的散点图（x 轴：降雨量，y 轴：PM2.5，大小表示气温，颜色区分空气质量）
    - 绘制气温偏差与 PM2.5 的散点图，并添加趋势线

第三阶段：聚类分析（使用 K-Means 算法）

1. **数据准备**
   - 选择聚类特征：`平均气温_清洗`、`降雨量_填充`、`PM2.5_清洗`、`湿度`（数值特征）
   - 对特征进行**标准化**（使用`sklearn.preprocessing.StandardScaler`），消除量纲影响
2. **K 值选择**
   - 使用**肘部法则**（绘制聚类损失随 K 值变化的折线图）确定最佳 K 值（K 从 1 到 10）
   - （可选）使用**轮廓系数**验证最佳 K 值
3. **K-Means 聚类**
   - 使用最佳 K 值进行 K-Means 聚类，将聚类结果作为新列`聚类标签`添加到原数据集
   - 统计每个聚类的样本数量、各特征的均值
4. **聚类结果可视化**
   - 使用**散点图**展示聚类结果（选取两个特征作为 x、y 轴，颜色区分聚类标签）
   - 使用**平行坐标图**或**热力图**分析各聚类的特征差异
   - 结合分类特征（如城市、季节），分析每个聚类的特征分布（例如：聚类 1 主要是北方城市冬季数据，聚类 2 主要是南方城市夏季数据）
5. **聚类结果解读**
   - 总结每个聚类的特征特点（例如：聚类 0 为「高温、低 PM2.5、高湿度」的城市季节数据）
   - 分析聚类结果的业务意义（例如：哪些城市季节的环境条件更优）

> （注：文档部分内容可能由 AI 生成）