数据预处理和可视化作业-2

2024-11-21

处理北京空气质量数据 （数据源：kaggle，PM2.5 Data of Five Chinese Cities | Kaggle）

1. 对 HUMI、PRES、TEMP 三列，进行线性插值处理。并对其中超过 3 倍标准差的高度异常数据，修改为 3 倍标准差的数值。
2. 假设 PM 指数最高为 500，对 PM_Dongsi、PM_Dongsihuan、 PM_Nongzhanguan 三列中超过 500 的数据，修改为 500PM 指数 进行异常值的处理。
3. 修改 cbwd 列中值为“cv”的单元格，其值用后项数据填充。
4. 对 DEWP 和 TEMP 两列，进行 0-1 归一化及 Z-Score 归一化处理。 结果使用散点图的形式表示（参考 PPT 第 19 页图形上半部分的表现形式）。
   a) 将北京的空气质量数据进行离散化，按照空气质量指数分级标准，计算出每个级别（或颜色值）对应的天数各有多少

提交内容：编写的源程序、说明文档（要求简洁）、处理后的 csv 数据文件、计算结果输出文件和可视化展示的图片文件。

评分标准：异常数据处理 1（15），异常数据处理 2（15），数据填充（15），归一化及散点图（20），离散化并计算天数（20），源程序及说明文档（15）