

数据预处理和可视化作业-1

2024-11-21

把通过爬虫爬下来的新房数据，进行预处理，处理后的数据存为 csv 格式的文件：

1. 最终的 csv 文件，应包括以下字段：楼盘名称，类型，地理位置（3 个字段分别存储），房型（只保留最小房型或平均值），面积（按照最小值或平均值），总价（万元， 整数），均价（元，整数）；其中房型和面积的数值选取方式与后续展示数据的目的相关，选择理由在报告中简要说明；
2. 对于所有字符串字段，要求去掉所有的前后空格；
3. 如果有缺失数据，不用填充。

将处理后的数据进行可视化展示：

1. 绘制楼盘价格分布的散点图，纵横轴分别代表单价和总价，楼盘类型通过散点的颜色区分
2. 绘制各行政区楼盘的直方图，每个行政区一个柱子，高度代表平均单价，宽度代表楼盘数量；将单价换为总价再绘制一张直方图
3. 可自行增加其他表现数据特点的图表，在文档中说明设计思路

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

评分标准：数据预处理后的 csv 文件（30），散点图（25），直方图（25），其它图表（5），源程序及说明文档（15）