实验 3 文本数据及其可视化

一、评分标准

1. 本实验包含 1 个子实验，分值分布见下表；
2. 提交材料包括实验报告、源代码，若未提交源代码，扣 20 分；
3. 实验报告内容包括：实验要求；实验实现过程；核心源代码截图及说明；程序运行结果截图及说明；项目开展的总结。未提供实验实现过程，扣 10 分；未提供代码和运行结果截图及说明，扣 20 分；总结过于简单，不具体，讲套话，不能体现实验过程，扣 10 分；其他内容不完整，酌情扣分；
4. 实验报告排版格式不佳，酌情扣分。

序号	题目	分值
1	数据分析综合应用	

核心：爬虫，字符串处理，pandas 数据清洗与统计分析，线性回归

二、实验要求

1、数据爬取

采用爬虫技术（urllib 库，BeautifulSoup 库）从链家网站（参考网址：https://cd.lianjia.com/zufang/jinrongcheng/pg2rt200600000001/#contentList）获取”链家/成都市/高新区/金融城/整租”租房信息（爬取 40 页数据），从各房屋信息中提取”楼盘名称/面积/装修/校验/楼层/总楼层/租金”信息，并将数据写入 lianjia.xls 文件（标题行为['name', 'decorate', 'check', 'area', 'floor','total_floor','rentFee']）

2、机器学习数据准备（数据清洗、统计分析和数据变换等）

读取 lianjia.xls 数据，存储到 dataframe 对象，完成以下内容：

1) 提取['name','decorate','area', 'floor'，'rentFee']共 5 列数据，用于后续机器学习
2) 房屋面积 area 是回归分析的核心参数，不能有缺失值，过滤 area 列缺失值

3) 查看'decorate'缺失值，并将: 'decorate'缺失值填充为'非精装'

4) 查看’name’列楼盘信息，提取特定楼盘数据共后续使用（学号末尾奇数提取’誉峰三期’数据，偶数提取’招商大魔方’数据

5) 特征编码：'decorate'列，'非精装'编码为 0,'精装'编码为 1；'decorate'列，'非近地铁'编码为 0,'近地铁'编码为 1

6) 'rentFee'列除以 1000，单位为千元

3、单变量回归分析

构建数据集：以'area'列数据为特征 ,'rentFee'列为标注信息，进行线性回归分析，分别计算并打印训练和测试误差

4、多变量回归分析

构建数据集：以['area', 'decorate', 'floor']共 3 列数据为特征 ,'rentFee'列为标注信息，特征数据经归一化处理后，进行线性回归分析，分别计算并打印训练和测试误差。