编程题(48')

E-commerce Customer Behavior - Sheet1 数据集提供了电子商务平台内客户行为的全面视图。数据集中的每个条目都对应一个独特的客户，提供了客户互动和交易的详细分类。在使用该数据集完成以下任务时,最终需要完成对客户满意度 ***Satisfaction Level*** 的分类训练。数据集说明见 **dataset.txt**，**请在文件 prog.py 中书写代码,在其余文件中书写代码是无效的**。

1. 读取文件 E-commerce Customer Behavior - Sheet1.csv 并将第 **0** 列设置为索引,输出前 **5** 行数据。(4')
2. 统计含有缺失数据的列的情况,如果含有,按照以下表格处理**并覆盖原数据**:(5')

Gender	**Age**	**Membership Type**	**Total Spend**	**Items Purchased**	**Average Rating**	**Discount Applied**
删除整行	用 35 填充	删除整行	平均值填充	前一行数据填充	中位数填充	删除整行

City	**Days Since Last Purchase**	***Satisfaction Level***
删除整行	删除整行	删除整行

3. 统计含有重复数据的行数,并删除，**覆盖原数据**。(5')
4. 统计每个区域(City)每一种会员类型(Membership Type) [Silver,Bronze,Gold]的客户数。(6')
5. 给出平均星级排名(Average Rating)的平均值最高的区域(City) (6')
6. 使用直方图绘制年龄分布图并将直条数(bins)设置为 20,在**同一张图**上绘制核密度分布图,在直方图上用箭头标记出最高点,图像命名为 fig1.png,dpi=300 保存在试题目录下的 fig 文件夹内。(6')
7. 统计数值类型 Age,Total Spend,Items Purchased,Average Rating,Days Since Last Purchase 之间的相关性,打印出相关性矩阵,绘制 **5** 个变量之间的散点图矩阵,矩阵**对角线上使用密度分布图像**, 图像命名为 fig2.png,dpi=300 保存在试题目录下的 fig 文件夹内。(6')
8. 将数据分割为训练集和测试集(测试集大小为 **50%**),在训练集上训练关于 ***Satisfaction Level*** 的分类模型(至少采用两种分类算法),并比较之间的性能差异(相同 or 某一个更好),输出二者的性能评分和分类报告,并单独粘贴在注释行中。(10')