|  |  |
| --- | --- |
| 实验项目名称 | Pandas数据框操作及数据提取 |
| 1. **实验目的和要求** 2. 掌握使用Pandas进行数据提取 3. 掌握使用Pandas进行数据框操作   **二、实验内容**  按问题要求填写代码并执行以输出结果。   1. **实验步骤**  STEP1:按照下列要求完成各题目 # 读取pandas120数据文件  df = pd.read\_excel('/home/mw/input/pandas1206855/pandas120.xlsx')  df.head()  运行结果：    1. 提取学历为本科，工资在25k-35k的数据  df1 =  2. 提取salary列中以'40k'结尾的数据  df2 =  3. 提取薪资区间中最低薪资与最高薪资的平均值大于30k的行，只需提取原始字段('createTime', 'education', 'salary')即可  df3 =  4. 将以上三题提取出来的行按照相同列进行合并，汇总到一个数据框中；  answer\_2 =  5. 将三列数据合并成一列，并设置列名为answer，最后保留id(数据行数、answer)  **STEP2: 将结果保存为 csv 文件**  **STEP3: 提交 csv 文件，获取分数结果**  **四、实验结果**  import pandas as pd  import numpy as np  # 由于随机数的存在会影响最终提交的文件，所以这里重新读取数据  # 读取pandas120数据文件  df = pd.read\_excel('pandas120.xlsx')  df.head()  #提取学历为本科，工资在25k-35k的数据  salary\_25k\_35k\_df = df[df['salary'] == '25k-35k']  edu\_df = salary\_25k\_35k\_df[salary\_25k\_35k\_df['education'] == '本科']  df1 = edu\_df  df1  #提取salary列中以'40k'结尾的数据  end\_40k\_df = df[df['salary'].str.endswith('40k')]  df2 = end\_40k\_df  df2  # 提取'salary'列中样例为'25k-30k'的数据  df['salary\_range'] = df['salary'].str.extract(r'(\d+k-\d+k)')  # 拆分'salary\_range'列为'25k'和'30k'两列数据  df[['min\_salary', 'max\_salary']] = df['salary\_range'].str.split('-', expand=True)  # 将字符串转换为数值类型  df['min\_salary'] = df['min\_salary'].str.extract(r'(\d+)').astype(int)  df['max\_salary'] = df['max\_salary'].str.extract(r'(\d+)').astype(int)  # 计算每一行的平均值保存在'avr'列中  df['avr'] = (df['min\_salary'] + df['max\_salary']) / 2  df3 = df[df['avr'] > 30]  df3  answer\_2 = pd.concat([df1, df2, df3], axis=0)  answer\_2  data = pd.concat([answer\_2.iloc[:,0],answer\_2.iloc[:,1],answer\_2.iloc[:,2]])  df = pd.DataFrame(data, columns=['answer'])  df['id'] = range(len(df))  df = df[['id', 'answer']]  df  # 保存文件到本地  df.to\_csv('answer\_2.csv', index=False, encoding='utf-8-sig')#代码自动补全的快捷键是 tab；运行完成后，左侧文件树刷新下，可以找到这份文件  **五、结果分析与实验体会** | |
| 评语：  成绩： 指导教师：  2023年 7 月 4 日 | |