实训 1 合并年龄、平均血糖和中风患者信息数据

源程序:

```
import pandas as pd

# 读取表格
stroke_data = pd.read_excel('healthcare-dataset-stroke.xlsx')
age_abs_data = pd.read_excel('healthcare-dataset-age_abs.xlsx')

# 合并数据
merged_data = pd.merge(stroke_data, age_abs_data, on='编号')

# 保存合并后的数据
merged_data.to_excel('merged_healthcare_data.xlsx', index=False)
print("合并成功，保存为 merged_healthcare_data.xlsx")

display(merged_data)
```

过程性结果：

[3]:
```
import pandas as pd

# 读取表格
stroke_data = pd.read_excel('healthcare-dataset-stroke.xlsx')
age_abs_data = pd.read_excel('healthcare-dataset-age_abs.xlsx')

# 合并数据
merged_data = pd.merge(stroke_data, age_abs_data, on='编号')

# 保存合并后的数据
merged_data.to_excel('merged_healthcare_data.xlsx', index=False)
print("合并成功，保存为 merged_healthcare_data.xlsx")

display(merged_data)
```

```
合并成功，保存为 merged_healthcare_data.xlsx
```

	编号	性别	高血压	家族遗传	居住类型	体重指数	吸烟史	中风	年龄	平均血糖浓度(mg/dl)
0	9046	男	0	是	城市	36.6	以前吸烟	1	67.0	228.69
1	51676	女	0	是	农村	32.0	从不吸烟	1	61.0	202.21
2	31112	男	0	是	农村	32.5	从不吸烟	1	80.0	105.92
3	60182	女	0	是	城市	34.4	抽烟	1	79.0	171.23
4	1665	女	0	是	农村	24.0	从不吸烟	1	74.0	174.12

结论：

基于编号合并了两个表格的数据。（1.8 年龄的数据在此处没有对应编号，所以已经被自然剔除。）

实训 2 剔除年龄异常的数据

根据实训内容，如果年龄值为 1.8，则视为异常值，需要剔除。

源程序:

剔除年龄异常的数据

cleaned_data = merged_data[merged_data['年龄'] != 1.8]

保存清理后的数据

cleaned_data.to_excel('cleaned_healthcare_data.xlsx', index=False)

print("剔除异常年龄数据，保存为 cleaned_healthcare_data.xlsx")

display(cleaned_data)

过程性结果:

实训2 剔除年龄异常的数据

根据实训内容，如果年龄值为1.8，则视为异常值，需要剔除。

[5]:
```
# 剔除年龄异常的数据
cleaned_data = merged_data[merged_data['年龄'] != 1.8]

# 保存清理后的数据
cleaned_data.to_excel('cleaned_healthcare_data.xlsx', index=False)
print("剔除异常年龄数据，保存为 cleaned_healthcare_data.xlsx")
display(cleaned_data)
```

剔除异常年龄数据，保存为 cleaned_healthcare_data.xlsx

	编号	性别	高血压	家族遗传	居住类型	体重指数	吸烟史	中风	年龄	平均血糖浓度(mg/dl)
0	9046	男	0	是	城市	36.6	以前吸烟	1	67.0	228.69
1	51676	女	0	是	农村	32.0	从不吸烟	1	61.0	202.21
2	31112	男	0	是	农村	32.5	从不吸烟	1	80.0	105.92
3	60182	女	0	是	城市	34.4	抽烟	1	79.0	171.23
4	1665	女	0	是	农村	24.0	从不吸烟	1	74.0	174.12

结论:

已经剔除了年龄为 1.8 的异常值数据。

实训 3 离散化年龄特征

为了离散化年龄特征，我们可以将连续型数据转换为离散型数据。我们可以使用 pd.cut()函数来实现。

源程序:

定义年龄区间

bins = [0, 20, 40, 60, 80, 100]

labels = ['0-20', '21-40', '41-60', '61-80', '81-100']

离散化年龄特征

cleaned_data['年龄区间'] = pd.cut(cleaned_data['年龄'], bins=bins, labels=labels, right=False)

保存离散化后的数据

cleaned_data.to_excel('discretized_healthcare_data.xlsx', index=False)

print("离散化年龄特征，保存为 discretized_healthcare_data.xlsx")

display(cleaned_data)

过程性结果:

[6]:
```
# 定义年龄区间
bins = [0, 20, 40, 60, 80, 100]
labels = ['0-20', '21-40', '41-60', '61-80', '81-100']

# 离散化年龄特征
cleaned_data['年龄区间'] = pd.cut(cleaned_data['年龄'], bins=bins, labels=labels, right=False)

# 保存离散化后的数据
cleaned_data.to_excel('discretized_healthcare_data.xlsx', index=False)
print("离散化年龄特征，保存为 discretized_healthcare_data.xlsx")
display(cleaned_data)
```

```
离散化年龄特征，保存为 discretized_healthcare_data.xlsx
```

	编号	性别	高血压	家族遗传	居住类型	体重指数	吸烟史	中风	年龄	平均血糖浓度(mg/dl)	年龄区间
0	9046	男	0	是	城市	36.6	以前吸烟	1	67.0	228.69	61-80
1	51676	女	0	是	农村	32.0	从不吸烟	1	61.0	202.21	61-80
2	31112	男	0	是	农村	32.5	从不吸烟	1	80.0	105.92	81-100
3	60182	女	0	是	城市	34.4	抽烟	1	79.0	171.23	61-80
4	1665	女	0	是	农村	24.0	从不吸烟	1	74.0	174.12	61-80

结论：年龄特征已被离散化至一个区间。