2.1.x知识点总结

数据加载

csv格式数据加载
同1.1.x

xlsx格式数据加载
需要pip install openpyxl库
pd.read_excel('文件名.xlsx')

查看表结构
data.info()

查看前五行数据
data.head()

数据预处理

新增时间差列
data['时间差列'] = (data['结束时间列']-data['开始时间列']).dt.days

修改列名
data.rename(columns={'病人ID':'患者ID'}, inplace=True)

转换成数值类型
data['待转换列'] = pd.to_numeric(data['待转换的列'],errors='coerce')

删除重复行
data = data.drop_duplicates()

使用sklearn.preprocessing数据预处理库进行数据预处理

StandardScaler数据标准化方法
StandardScaler().fit_transform(data['待标准化列'])

MinMaxScaler数据归一化方法
MinMaxScaler().fit_transform(data['待归一化列'])

LabelEncoder数据归一化方法
LabelEncoder().fit_transform(data['待归一化列'])

箱线图法处理异常值
超过合理值上限和低于合理值下限的就是异常值

计算分位数
1、下四分位数：
Q1 = data['待计算数值列'].quantile(0.25)

2、中位数：
Q2 = data['待计算数值列'].quantile(0.5)

3、上四分位数：
Q3 = data['待计算数值列'].quantile(0.75)

计算四分位距IQR
IQR = Q3-Q1

计算合理值下限
Q1-IQR*1.5

计算合理值上限
Q3+IQR*1.5

数据可视化

pandas的绘图方法
写法1
示例：柱状图
绘图数据源.plot(kind='bar', 柱状图的各种参数)

写法2
示例：饼图
绘图数据源.plot.pie(饼图的各种参数)

matplotlib.pyplot的绘图方法
散点图
plt.scatter(坐标1数据，坐标2数据)

特征和目标变量选择

特征变量选择

使用drop反选
X = data.drop(columns = ['目标变量','非特征变量'])

如果特征变量索引连续且数量多，用列表推导式确定列名之后在选择

selected_features = [col for index,col in enumerate(data) if 2<=index<=9]

X = data[selected_features]

直接把列名一个一个复制上去
selected_features = ['特征列名1','特征列名2',...]

data.columns可以答应出来data所有列名，带引号和逗号方便直接复制

目标变量选择
根据题目要求直接选择

数据集划分

使用sklearn.model_selection库中的train_test_split函数进行数据集划分

1、自变量训练集, 自变量测试集, 目标变量训练集, 目标变量测试集合 = train_test_split(自变量集, 因变量集, 测试集比例, random_state=42)

2、测试集，训练集 = train_test_split(数据全集, 测试集比例, random_state=42)

特征和目标变量数据合并

合并特征和目标变量
pd.concat([特征变量,目标变量], axis =1)
axis = 1 表示按照列进行合并

数据保存

同1.1.x