**人工智能训练师（三级）操作技能考核**

**试题单**

准考证号：

试题代码：1.1.4

试题名称：电商平台用户行为分析系统的数据采集与处理流程设计

考核时间：30min

**1.**场地设备要求

人工智能训练师主机：CPU（intel i5 及以上）、内存（不少于 16GB）、操作系统（windows10）、支持深度学习训练；

**2.**工作任务

某电商平台希望通过用户行为数据分析，了解用户购物习惯、购买倾向等，从而优化产品推荐系统，提高用户满意度和销售额。作为数据分析师，您需要设计一套全面的业务数据采集与处理流程，确保数据在进入用户行为分析系统之前经过严格的采集、清洗和预处理。

我们提供一个用户行为数据集（user\_behavior\_data.csv），包含以下字段：

* UserID: 用户ID
* UserName: 用户名
* Age: 年龄
* Gender: 性别（Male/Female）
* Location: 位置
* LastLogin: 上次登录时间
* PurchaseAmount: 购买金额
* PurchaseCategory: 购买类别（例如，电子产品、服装、食品等）
* ReviewScore: 用户评价评分（1-5）
* LoginFrequency: 登录频率（每日、每周、每月）

你作为人工智能训练师，根据提供的user\_behavior\_data.csv数据集和Python代码框架（1.1.4.ipynb），完成以下数据的采集与处理任务，确保数据的准确性和可靠性。请按照以下要求完成任务，确保结果准确并保存相应的截图。

（1）数据采集：

通过运行Python代码（1.1.4.ipynb），从本地文件user\_behavior\_data.csv中读取数据，并将数据加载到DataFrame中。打印前5条数据。

（2）数据清洗与预处理：

通过运行Python代码（1.1.4.ipynb）对数据进行清洗和预处理，具体要求如下：

* 处理缺失值：对缺失值进行填充或删除。
* 数据类型转换：确保每个字段的数据类型正确。
* 处理异常值：删除不合理的年龄、购买金额和评价评分。
* 数据标准化：对购买金额和评价评分进行标准化处理。

清洗后的数据保存为新文件cleaned\_user\_behavior\_data.csv。

（3）数据统计：

通过运行Python代码（1.1.4.ipynb），完成以下数据统计任务：

* 统计每个购买类别的用户数。
* 统计不同性别的平均购买金额。
* 统计不同年龄段的用户数（18-25岁、26-35岁、36-45岁、46-55岁、56-65岁、65岁以上）。

将统计结果分别截图以JPG的格式保存，分别命名为“1.1.4-1”、“1.1.4-2”、“1.1.4-3”。

所有结果文件储存在桌面新建的考生文件夹中，文件夹命名为“准考证号+身份证号后六位”。

**3.**技能要求

（1）能结合人工智能技术要求和业务特征，设计整套业务数据采集流程

（2）能结合人工智能技术要求和业务特征，设计整套业务数据处理流程

**4.**质量指标

（1）完整性指标：数据集中无缺失值和重复记录。

（2）合理性指标：所有数据点符合业务规则，无异常值存在。

（3）清洗效果指标：清洗后的数据集完整、合理，且适于建模分析。

**人工智能训练师（三级）操作技能考核**

**试题评分表**

准考证号：

试题代码：

试题名称：电商平台用户行为分析系统的数据采集与处理流程设计

考核时间：30min

测量分评分表

| 细则编号 | 配分 | 评分细则描述 | 规定或  标称值 | 结果或  实际值 | 得分 |
| --- | --- | --- | --- | --- | --- |
| M1 | 2 | 从本地文件中读取数据的代码正确得2分 | 根据数据 |  |  |
| M2 | 2 | 打印数据的前5条记录的代码正确得2分 | 根据数据 |  |  |
| M3 | 2 | 处理缺失值的代码正确得2分； | 根据数据 |  |  |
| M4 | 2 | Age数据类型转换的代码正确得2分； | 根据数据 |  |  |
| M5 | 2 | PurchaseAmount数据类型转换的代码正确得2分； | 根据数据 |  |  |
| M6 | 2 | ReviewScore数据类型转换的代码正确得2分； | 根据数据 |  |  |
| M7 | 2 | 处理异常值的代码正确得2分； | 根据数据 |  |  |
| M8 | 2 | PurchaseAmount数据标准化的代码正确2分； | 根据数据 |  |  |
| M9 | 2 | ReviewScore数据标准化的代码正确得2分； | 根据数据 |  |  |
| M10 | 1 | 保存清洗后的数据的代码正确得1分； | 根据数据 |  |  |
| M11 | 2 | 统计每个购买类别的用户数的代码正确得2分； | 根据数据 |  |  |
| M12 | 2 | 统计不同性别的平均购买金额的代码正确得2分； | 根据数据 |  |  |
| M13 | 2 | 统计不同年龄段的用户数的代码正确得2分； | 根据数据 |  |  |
| 合计配分 | 25 | 合计得分 | |  |  |

**参考答案**：

Python参考代码（1.1.4.ipynb）：

import pandas

import numpy as np

import matplotlib.pyplot as plt

# 1. 数据采集

# 从本地文件中读取数据

data = pandas.read\_csv('user\_behavior\_data.csv')

print("数据采集完成，已加载到DataFrame中")

# 打印数据的前5条记录

print(data.head())

# 2. 数据清洗与预处理

# 处理缺失值（删除包含缺失值的行）

data = data.dropna()

# 数据类型转换

data['Age'] = data['Age'].astype(int) # Age转整型

data['PurchaseAmount'] = data['PurchaseAmount'].astype(float) # PurchaseAmount转浮点型

data['ReviewScore'] = data['ReviewScore'].astype(int) # ReviewScore转整型

# 处理异常值（筛选合理范围）

data = data[

(data['Age'].between(18, 70)) &

(data['PurchaseAmount'] > 0) &

(data['ReviewScore'].between(1, 5))

]

# 数据标准化（Z-Score标准化）

data['PurchaseAmount'] = (data['PurchaseAmount'] - data['PurchaseAmount'].mean()) / data['PurchaseAmount'].std()

data['ReviewScore'] = (data['ReviewScore'] - data['ReviewScore'].mean()) / data['ReviewScore'].std()

# 保存清洗后的数据

data.to\_csv('cleaned\_user\_behavior\_data.csv', index=False)

print("数据清洗完成，已保存为 'cleaned\_user\_behavior\_data.csv'")

# 3. 数据统计

# 统计每个购买类别的用户数

purchase\_category\_counts = data['PurchaseCategory'].value\_counts()

print("每个购买类别的用户数:\n", purchase\_category\_counts)

# 统计不同性别的平均购买金额

gender\_purchase\_amount\_mean = data.groupby('Gender')['PurchaseAmount'].mean()

print("不同性别的平均购买金额:\n", gender\_purchase\_amount\_mean)

# 统计不同年龄段的用户数

bins = [18, 26, 36, 46, 56, 66, np.inf]

labels = ['18-25', '26-35', '36-45', '46-55', '56-65', '65+']

data['AgeGroup'] = pandas.cut(data['Age'], bins=bins, labels=labels, right=False)

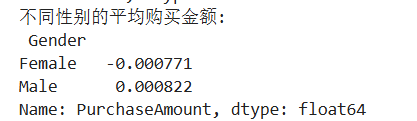
age\_group\_counts = data['AgeGroup'].value\_counts().sort\_index()

print("不同年龄段的用户数:\n", age\_group\_counts)

1.1.4-1.jpg



1.1.4-2.jpg



1.1.4-3.jpg

