

互评作业 1：数据探索性分析与预处理

周明轩 3120240977

GitHub: <https://github.com/Zhoumingx/BIT-DataMining>

1. 探索性分析和可视化

1.1 数据集基础信息

本次作业分别给定 10G 和 30G 数据,要求对上述数据进行分析。
首先展示给定数据集的基础信息和存储内容,各个字段、数据类型以及示例如下表所示。

字段名	数据类型	示例
id	int64	0
last_login	object	2024-12-02T03:49:12+00:00
user_name	object	RKWKCXRZFV
fullname	object	瞿紫玉
email	object	kuegujsk@hotmail.com
age	int64	82
income	float64	366311.83
gender	object	女
country	object	美国
address	object	Non-Chinese Address Placeholder
purchase_history	object	{"avg_price":9496,"categories":"零食", "items":[{"id":7265}], "payment_method":"现金", "payment_status":"已支付", "purchase_date":"2023-07-30"}
is_active	bool	False
registration_date	object	2024-10-31
phone_number	object	+1 (804) 855-6279

login_history	object	{"avg_session_duration":105,"devices":["desktop","mobile"],"first_login":"2024-12-04","locations":["home","travel"],"login_count":73,"timestamps":["2024-12-04 21:29:00","2024-12-12 20:51:00","2024-12-20 19:00:00","2024-12-28 10:58:00","2025-01-05 06:58:00","2025-01-13 21:55:00","2025-01-21 18:03:00","2025-01-29 18:26:00","2025-02-06 19:31:00","2025-02-14 11:15:00","2025-02-22 06:41:00","2025-03-02 10:10:00","2025-03-10 20:17:00","2025-03-18 20:19:00"]}
---------------	--------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

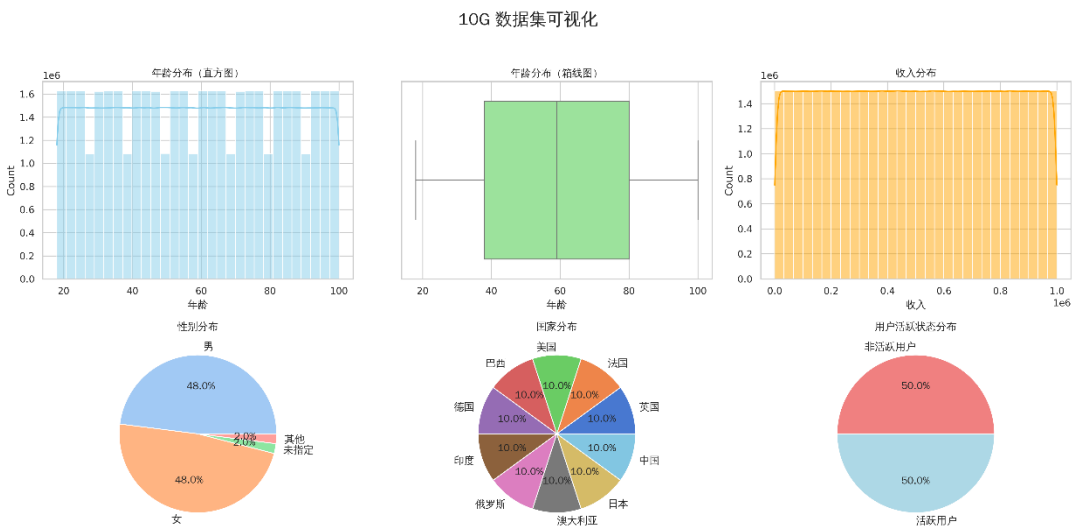
具体来说，给定的数据包含用户的基本信息、行为记录及交互历史等多维度字段。**id (int64)**: 唯一标识每个用户的整数编号。在每一个 **parquet** 文件中均从 0 开始递增。**last_login (object)**: 记录用户最近一次登录的时间，采用 **ISO 8601** 格式（例如 "2024-12-02T03:49:12+00:00"）。**user_name (object)** 与 **fullname (object)**: 分别为用户名和真实姓名。**email (object)**: 用户注册时提供的电子邮箱地址。**age (int64)**: 用户年龄。**income (float64)**: 用户收入。**gender (object)**: 用户性别。**country (object)** 与 **address (object)**: 分别表示用户所在国家和地址。**purchase_history (object)**: 一个嵌套的 **JSON** 字符串，详细记录用户的购买行为，包括：**avg_price**: 平均消费金额（如：9496）；**categories**: 消费品类（如： "零食"）；**items**: 购买商品列表（每项包含商品 ID）；**payment_method** 与 **payment_status**: 支付方式（如"现

金”)及支付状态; purchase_date: 最近一次购买日期(如: "2023-07-30")。is_active (bool): 布尔值, 标志用户当前是否活跃。registration_date (object): 用户注册日期。phone_number (object): 用户电话号码。login_history (object): 另一嵌套 JSON 字段, 包含: avg_session_duration: 平均会话时长(如 105 秒); devices: 使用的设备类型(如 “desktop”、“mobile”); first_login: 首次登录时间; locations: 登录位置类型(如 “home”、“travel”); login_count: 总登录次数; timestamps: 最近若干次登录的具体时间戳。

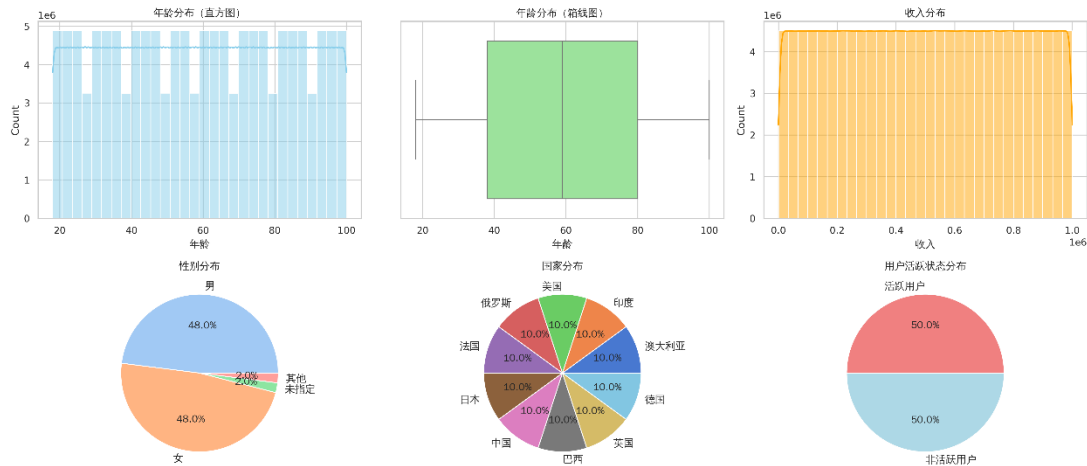
因此, 该数据存储了用户的基本信息、行为记录及交互历史等多维度信息。

1.2 数据集可视化

对给定的 10G 和 30G 数据集进行可视化。分别从年龄、收入、性别、国家、活跃程度层面, 对数据集中的所有数据分布进行可视化, 结果如下图所示。



30G 数据集可视化



可以看出，给定的 10G 和 30G 数据分布基本完全一致，且数据分布十分均匀。年龄、收入、性别、国家、活跃状态等服从均匀分布，异常值基本不存在。

下图展示了运行可视化代码所需的时间。

```
加载 10G 数据耗时: 208.75 秒
可视化 10G 数据耗时: 508.97 秒
加载 30G 数据耗时: 612.32 秒
可视化 30G 数据耗时: 1615.23 秒
```

2 数据预处理

2.1 数据质量分析

在数据可视化阶段，可以发现年龄字段处于 20 到 100 区间，不存在异常值；收入字段不存在小于 0 的异常值；性别字段判定男、女、未指定为合理值，其他为异常值；国家字段不存在异常值。

针对 10G 和 30G 两个数据集，添加以下规则，检查其他字段是否存在异常值：

- id 字段是否为整数且在一个数据文件内是否唯一；
- user_name 字段是否均为合法的英文字符；

- fullname 字段是否均为合法的中文字符;
- email 字段是否符合正确的邮箱格式;
- 所有字段是否存在缺失值。

针对上述检测规则，分别对 10G 和 30G 数据集进行质量评估，

结果如下：

```
📁 数据集【10G】的每个 parquet 文件中 id 唯一性检查:
part-00000.parquet - id 类型整数: True, 唯一性: True
part-00005.parquet - id 类型整数: True, 唯一性: True
part-00003.parquet - id 类型整数: True, 唯一性: True
part-00006.parquet - id 类型整数: True, 唯一性: True
part-00001.parquet - id 类型整数: True, 唯一性: True
part-00002.parquet - id 类型整数: True, 唯一性: True
part-00007.parquet - id 类型整数: True, 唯一性: True
part-00004.parquet - id 类型整数: True, 唯一性: True

📄 数据集【10G】字段合法性检查:
缺失值统计:
id                0
last_login        0
user_name         0
fullname          0
email             0
age              0
income            0
gender            0
country           0
address           0
purchase_history  0
is_active         0
registration_date 0
phone_number      0
login_history     0
dtype: int64
user_name 非法数量: 0
fullname 非中文数量: 0
email 非法数量: 0
10G 数据质量检测耗时: 318.27 秒
```

可以看出，给定的两个数据集质量较高，不存在字段缺失值；针

对上述给定的检查规则，也未发现异常值。

📁 数据集【30G】的每个 parquet 文件中 id 唯一性检查：

```
part-00014.parquet - id 类型整数: True, 唯一性: True
part-00000.parquet - id 类型整数: True, 唯一性: True
part-00010.parquet - id 类型整数: True, 唯一性: True
part-00015.parquet - id 类型整数: True, 唯一性: True
part-00005.parquet - id 类型整数: True, 唯一性: True
part-00013.parquet - id 类型整数: True, 唯一性: True
part-00003.parquet - id 类型整数: True, 唯一性: True
part-00006.parquet - id 类型整数: True, 唯一性: True
part-00012.parquet - id 类型整数: True, 唯一性: True
part-00001.parquet - id 类型整数: True, 唯一性: True
part-00011.parquet - id 类型整数: True, 唯一性: True
part-00002.parquet - id 类型整数: True, 唯一性: True
part-00007.parquet - id 类型整数: True, 唯一性: True
part-00009.parquet - id 类型整数: True, 唯一性: True
part-00004.parquet - id 类型整数: True, 唯一性: True
part-00008.parquet - id 类型整数: True, 唯一性: True
```

📄 数据集【30G】字段合法性检查：

缺失值统计：

id	0
last_login	0
user_name	0
fullname	0
email	0
age	0
income	0
gender	0
country	0
address	0
purchase_history	0
is_active	0
registration_date	0
phone_number	0
login_history	0

dtype: int64

user_name 非法数量: 0

fullname 非中文数量: 0

email 非法数量: 0

30G 数据质量检测耗时: 967.99 秒

2.2 异常值处理

基于上述分析，给定的两个数据集在性别字段存在异常值，因此

删除数据集上在性别字段上值为“其他”的数据。删除前后数据量的变化如下所示。

```
数据集【10G】删除前总记录数: 45000000
'gender' 为 '其他' 的记录数: 898865, 占比: 2.00%
删除后记录数: 44101135, 减少了: 898865 条

数据集【30G】删除前总记录数: 135000000
'gender' 为 '其他' 的记录数: 2698372, 占比: 2.00%
删除后记录数: 132301628, 减少了: 2698372 条
```

3 用户分析

使用给定的数据进行高价值用户分析。定义高价值用户符合以下标准：

- 高收入（位于收入的上四分位数 Q3 及以上）
- 年龄适中（25-55 岁的中青年群体）
- 历史购买行为良好（购买均价 > 5000 且已支付）
- 活跃用户（is_active = True）
- 近期登录过（last_login 在最近 2025 年）

基于上述标准进行分析，提取出高价值用户数据并保存为 CSV 文件。分析过程用时如下所示。

```
1745156096.7702742
🏆 最终识别的高价值用户数: 37859
10G 数据用户分析耗时: 298.14 秒
1745156395.5312662
🏆 最终识别的高价值用户数: 114405
30G 数据用户分析耗时: 948.47 秒
```