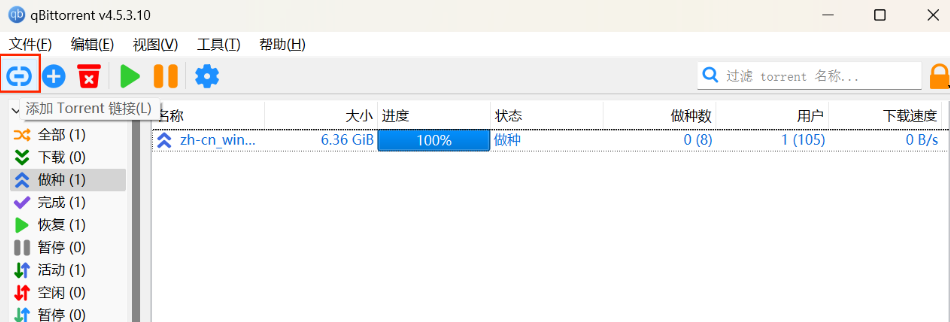
# 数据挖掘互评作业1-分析报告

姓名：张厚池 学号：3220241669

## 1.数据获取：

常规方法无法获取10G和100G的压缩文件。

通过工具qBittorrent Enhanced Edition v4.5.3.10，实现通过种子获取。



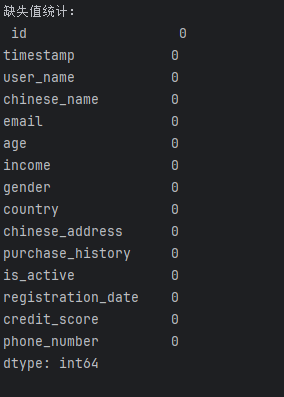
2．数据预处理：  
 1.遍历目录中所有文件，将其合并为DataFrame。

# 指定包含 Parquet 文件的目录路径  
directory = 'D:/10G\_data/'  
  
# 初始化一个空的 DataFrame 用于存储合并的数据  
df\_list = []  
  
# 遍历目录中的所有文件  
for filename in os.listdir(directory):  
 if filename.endswith('.parquet'):  
 file\_path = os.path.join(directory, filename)  
 df = pd.read\_parquet(file\_path)  
 df\_list.append(df)  
  
# 合并所有 DataFrame  
combined\_df = pd.concat(df\_list, ignore\_index=True)

2．缺失值评估：检查每列的缺失值数量

# 缺失值检测  
missing\_values = combined\_df.isnull().sum()  
print("缺失值统计：\n", missing\_values)

检查结果如下：（10G与30G处理结果相同）



说明反馈结果为无缺失。

3.重复值检测：检查每行是否存在内容相同

# 重复值检测  
# duplicate\_rows = combined\_df.duplicated().sum()  
# print(f"重复行数：{duplicate\_rows}")



4. 处理缺失值：将数据中缺失的部分去掉。

处理重复值：删除数据中的重复行。

# 处理缺失值  
combined\_df\_cleaned = combined\_df.dropna()  
# 删除重复行  
combined\_df\_no\_duplicates = combined\_df.drop\_duplicates()

## 3.数据可视化-三种

1. **年龄分布直方图**：

plt.figure(figsize=(10, 6))

sns.histplot(df['age'], bins=30, kde=True)

plt.title('年龄分布直方图')

plt.xlabel('年龄')

plt.ylabel('频数')

plt.show()

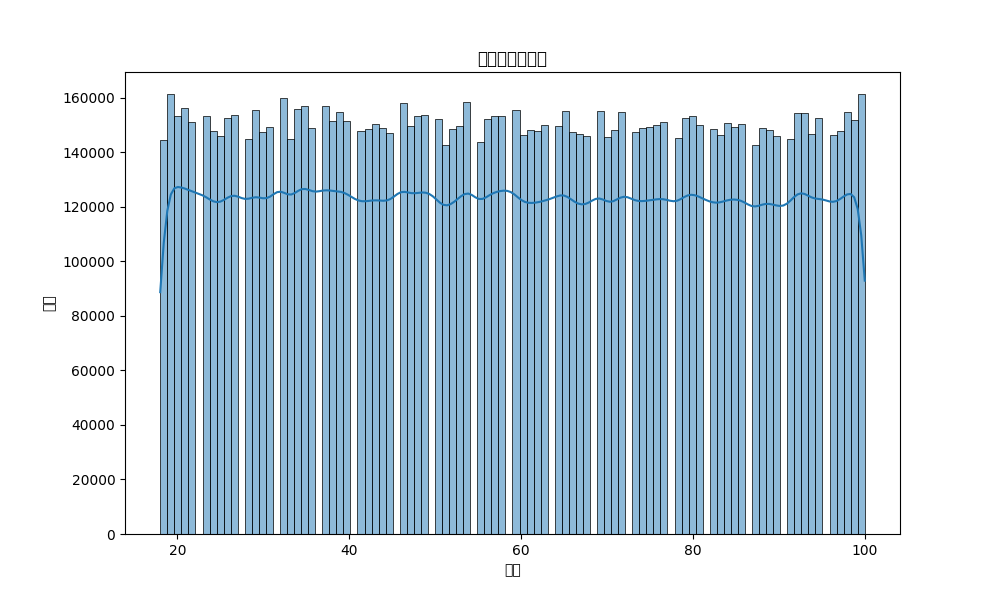
**2.收入与信用评分的散点图**：

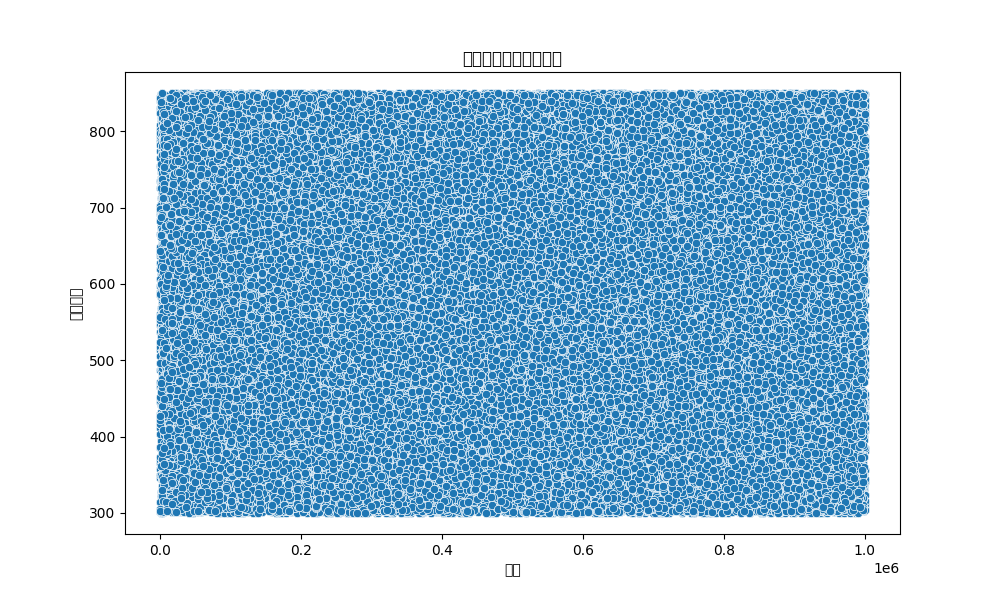
#收入与信用散点图  
plt.figure(figsize=(10, 6))  
sns.scatterplot(x=df['income'], y=df['credit\_score'])  
plt.title('收入与信用评分的关系')  
plt.xlabel('收入')  
plt.ylabel('信用评分')  
plt.show()

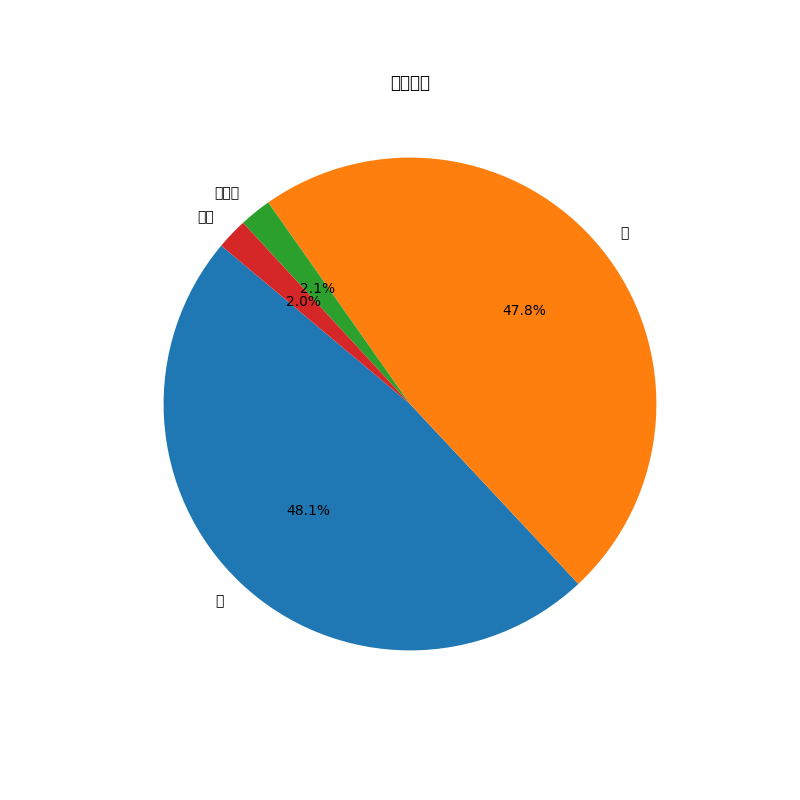
**3.性别比例饼图**：

gender\_counts = df['gender'].value\_counts()  
plt.figure(figsize=(8, 8))  
plt.pie(gender\_counts, labels=gender\_counts.index, autopct='%1.1f%%', startangle=140)  
plt.title('性别比例')  
plt.show()

效果如下所示：







## 4. **建立用户画像**：​

分析用户的年龄、性别、收入等特征，建立用户画像。

# 按性别和年龄段分析用户数量

df['age\_group'] = pd.cut(df['age'], bins=[18, 30, 40, 50, 60, 100], labels=['18-30', '31-40', '41-50', '51-60', '60+'])

plt.figure(figsize=(12, 6))

sns.countplot(x='age\_group', hue='gender', data=df)

plt.title('不同年龄段和性别的用户数量')

plt.xlabel('年龄段')

plt.ylabel('用户数量')

plt.show()

结果如下所示：

