数据挖掘第一次互评作业

本次作业分为 10G 数据和 30G 数据分别进行相同的分析，因为数据的格式相似。

Part1:10G 数据

一、数据读取和数据预处理

首先从 parquet 文件中读取数据，因为数据本身较大所以不能全部读取出，需要按文件分别读出，先观察数据格式确定所需要的数据内容，尽可能减轻内存负担。

可以看到数据格式包含 id, timestamp, user_name, chinese_name, email, age, income, gender, country, chinese_address, purchase_history, is_active, registration_date, credit_score, phone_numberr 以及其中的 average_price 内部的信息如下。

'{"average_price":15.940000000000001,"category":"家居","items":[{"id":631},{"id":762},{"id":233},{"id":535},{"id":118},{"id":449},{"id":256},{"id":404},{"id":99},{"id":638}]}'

对于我们的数据分析而言，肯定需要根据我们的数据分析需求去删除一些无用的数据，在本次分析中，我希望对用户的购买能力与年龄，国家，信用分，收入等建立相关关系，所以最终选择保留 id,age,income,gender,country,credit_score,is_active 和 purchase_story，并且不考虑其中的类别，只提取其中的 average_price。

下面根据我的目的进行数据的预处理：

首先定义需要的字段

```python
# 定义必要字段及优化类型
COLS = ["id", "age", "income", "gender", "country", "credit_score", "is_active", "purchase_history"]
DTYPES = {
    "id": "int32",
    "age": "int8",          # 假设年龄范围 0-120
    "income": "float32",
    "gender": "category",
    "country": "category",
    "credit_score": "int16",
    "is_active": "bool"
}
```

然后读取数据，读取后首先对数据进行质量评价，质量评价主要需要做的是检查有无缺失值、异常值，首先检查有无缺失值，通过 df.isnull().sum()实现，发现没有缺失值，数据不错。

```
Missing values: id                  0
age                 0
income              0
gender              0
country             0
credit_score        0
is_active           0
purchase_history    0
purchase_avg        0
dtype: int64
```

然后检查异常值，首先规定不同数值的合理范围，通过 max 和 min 确定了如下数值的合理范围，然后通过代码检查异常值数量

```python
RANGE_RULES = {
    "age": (0, 120),
    "income": (0, None),  # 收入非负，无上限
    "credit_score": (300, 850)
}
```

```python
def count_outliers(df, column, lower, upper):
    if lower is not None and upper is not None:
        mask = (df[column] < lower) | (df[column] > upper)
    elif lower is not None:
        mask = df[column] < lower
    elif upper is not None:
        mask = df[column] > upper
    else:
        mask = pd.Series(False, index=df.index)
    return mask.sum()
```

```
基于业务规则的异常值数量:
  - age: 0 个异常值
  - income: 0 个异常值
  - credit_score: 0 个异常值
```

发现没有异常值，但是可能其他数据集具有异常值于是也进行异常值处理：

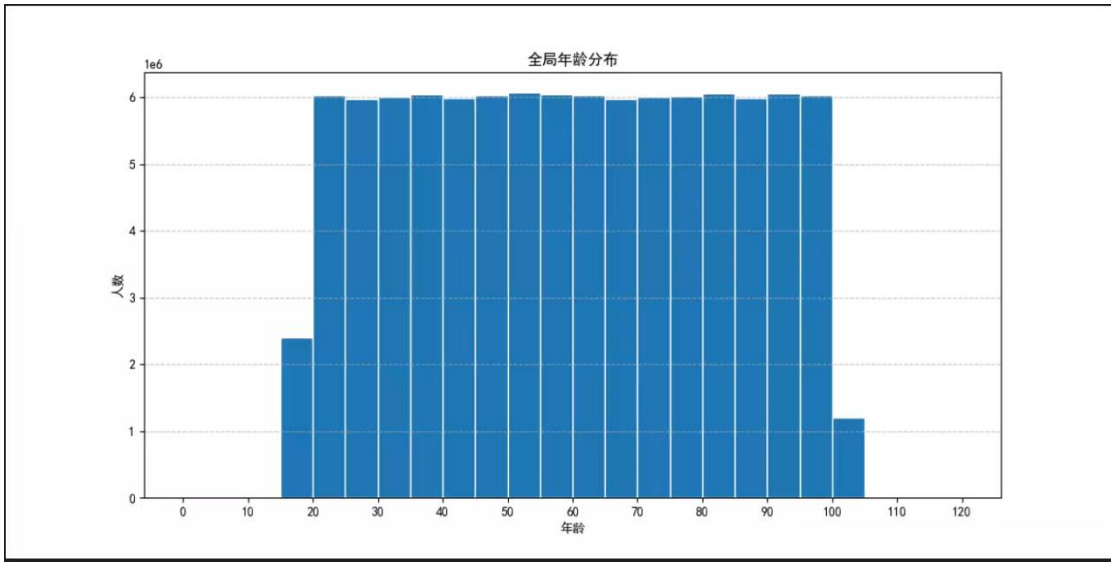
```python
df["age"] = df["age"].fillna(df["age"].median())
df["income"] = df["income"].fillna(df["income"].median())
df["credit_score"] = df["credit_score"].fillna(df["credit_score"].median())
```

经过以上预处理部分评价了数据质量不错没有缺失值和异常值，并且做好了数据的筛选和处理。

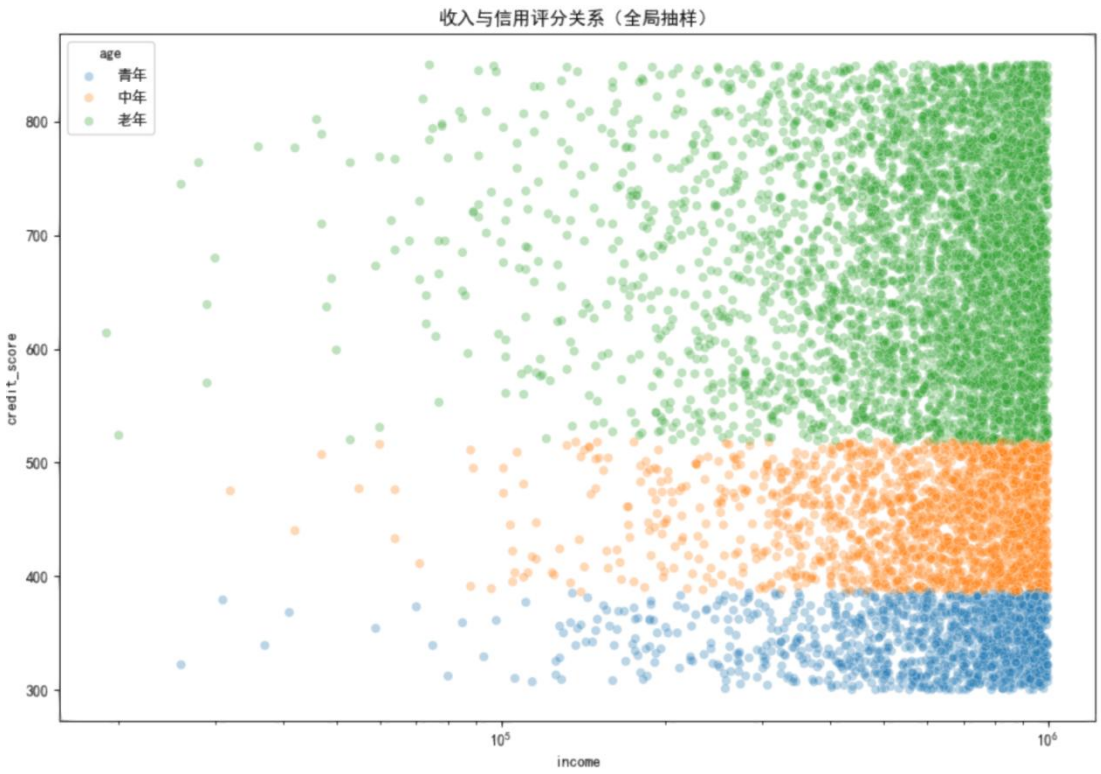
二、探索性分析与可视化

然后对数据进行初步的探索性分析与可视化，首先根据数据内容确定大致探索性分析的几个内容，初步定为，年龄分布可视化，收入分布可视化，用户国家分布可视化。

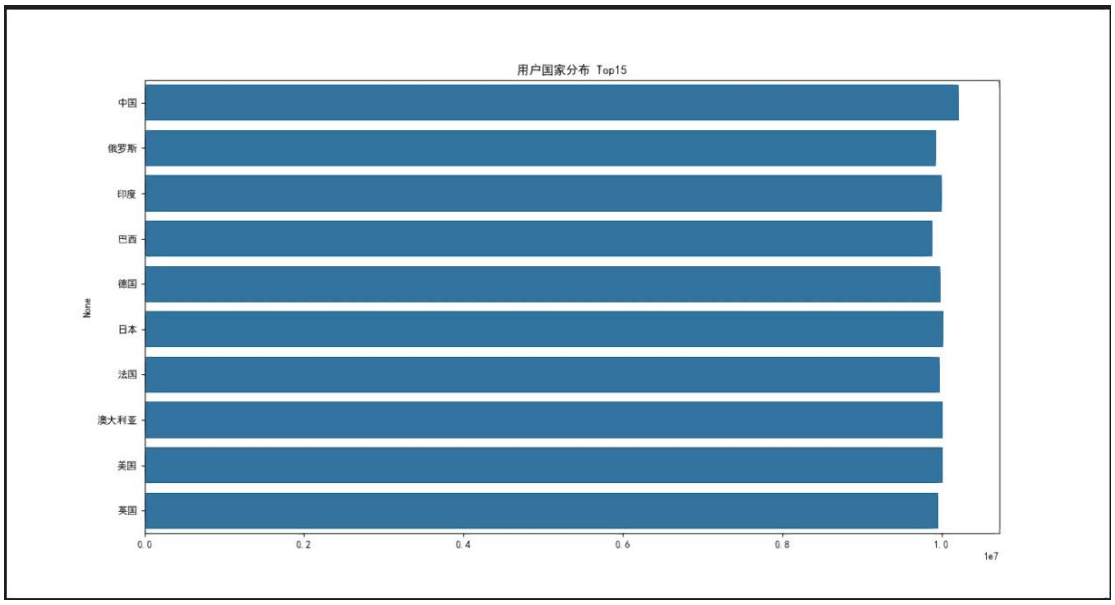
首先可视化年龄分布，以 5 为一个 bin 绘制直方图展示所有用户的年龄分布，可以看到年龄的最小范围在 15-20，最大范围在 100-105，在 15-20 和 100-105 的人数较少，分别为 200w 和 100w 左右，中间的每个年龄段都大致为 600w，总人数约为 1 亿，可以看到数据所选取的比较标准，中间的年龄段较为平均，两端的年龄段较少，符合客观的现象，说明数据选取较为合理。



然后是收入与信用分数的关系的抽样可视化，并且按照年龄段粗略的展示了青年，中年和老年的分布，发现基本上信用分数是随着年龄提升而提升的，并且我们加大了高收入的可见性，因为我们并不想过于了解低收入和信用评分的关系，主要是想看看信用评分除了收入是否还有年龄因素的影响。

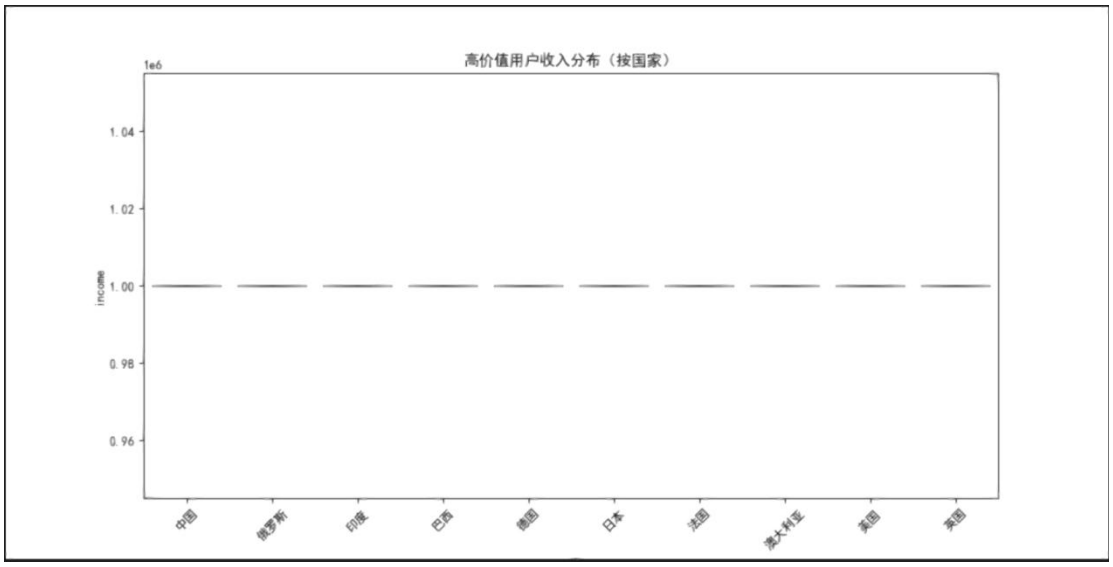


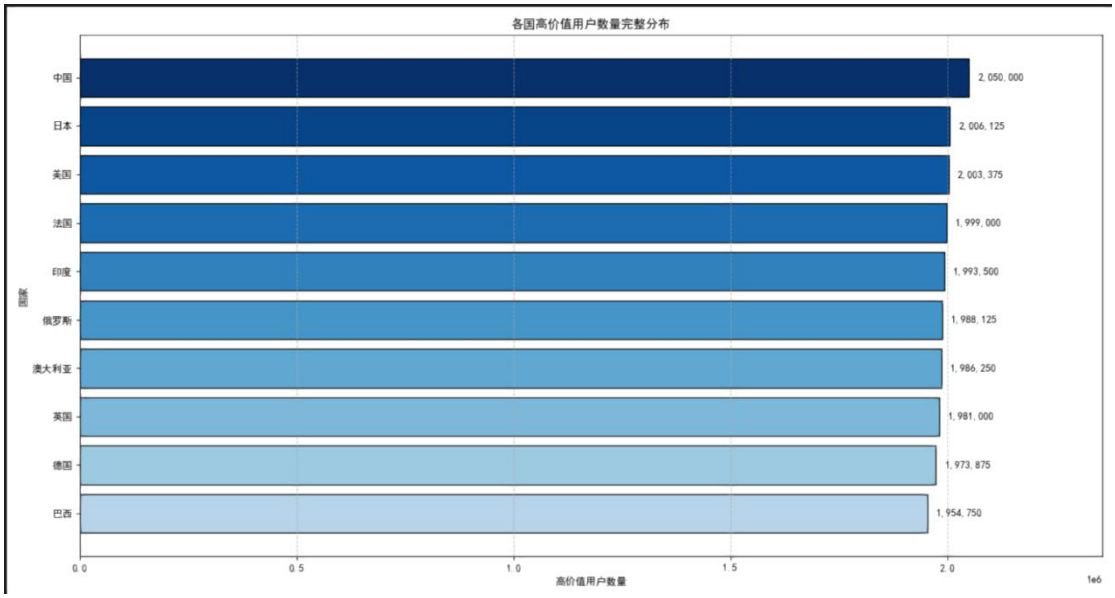
最后是用户国家分布可视化，又一次展示了数据的合理性，发现一共有十个国家每个国家人数都在 1000w 左右，也非常的均衡。



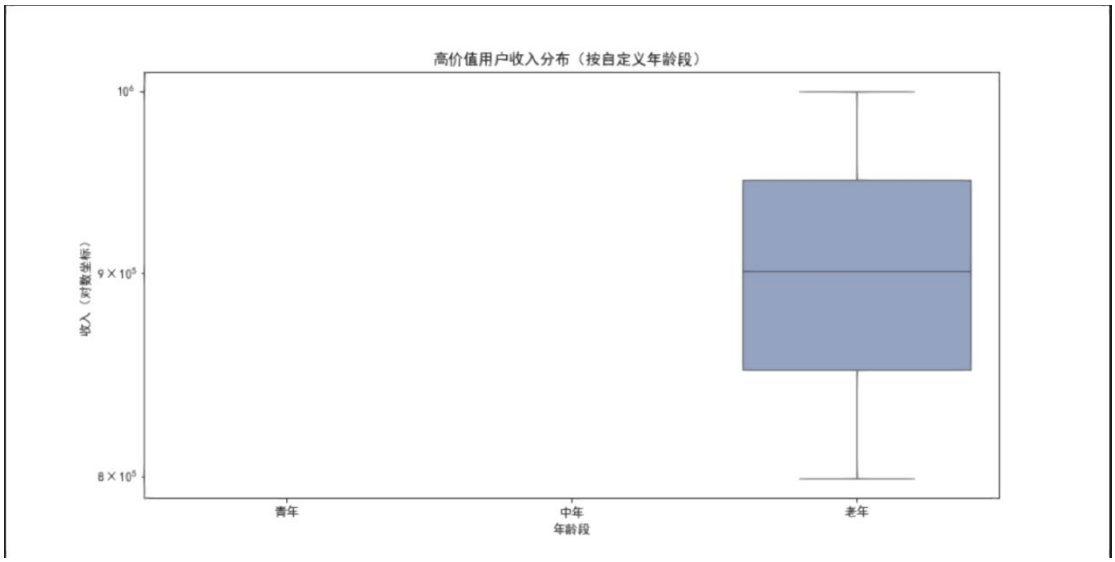
三、数据分析

选择分析目标为潜在高价值用户，首先定义潜在高价值用户，显然是需要收入和信用分数都足够高才能判定为潜在高价值用户，在这里将其定义为收入超过 80%的人并且信用分数高于 700 的用户，生成一个按国家的用户分布，以及输出总数。发现每个国家的高价值用户数量都差不多，和国家没有什么关系。以及高价值用户的收入最高的基本上都是同样的 100w，也就是说高价值用户收入分布可视化没有什么意义，最高的收入都是一致的。

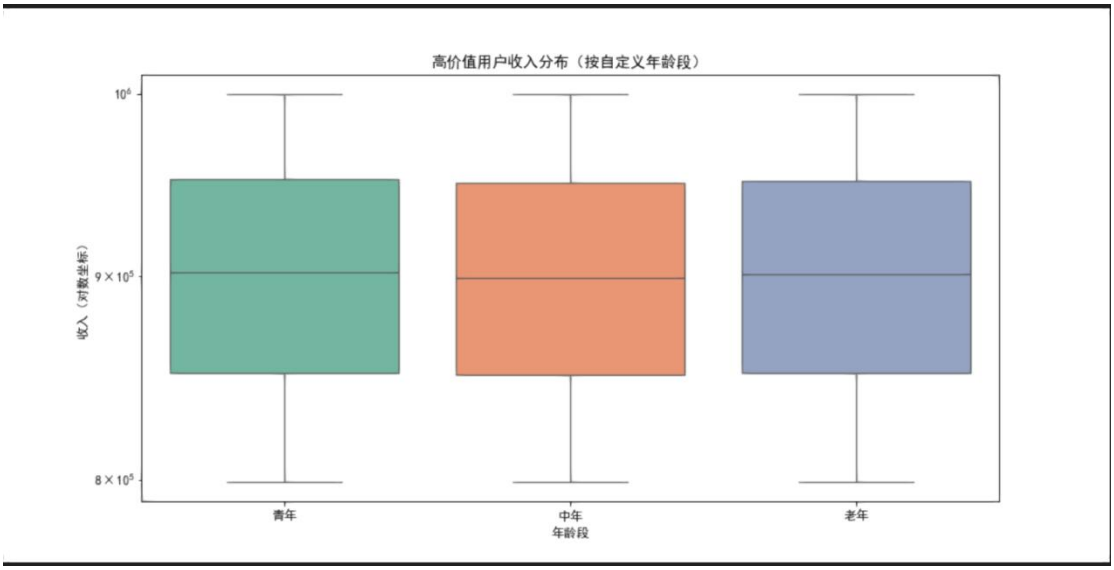


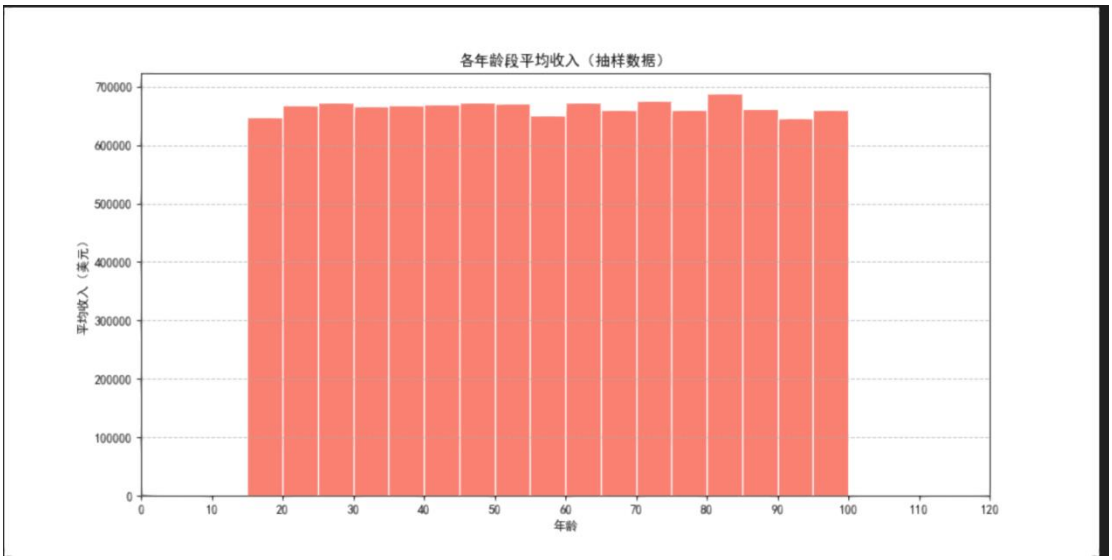


再可视化一下高价值用户和年龄的关系



发现所有的高价值用户都是老年人，结合之前的信用分数高的基本上都是老年人可以判定我们对高价值用户的定义有所欠缺，或者说数据中本身信用分数和年龄高度相关，因而改变高价值用户定义为只需要看收入，观察收入和年龄的关系再进行分析：

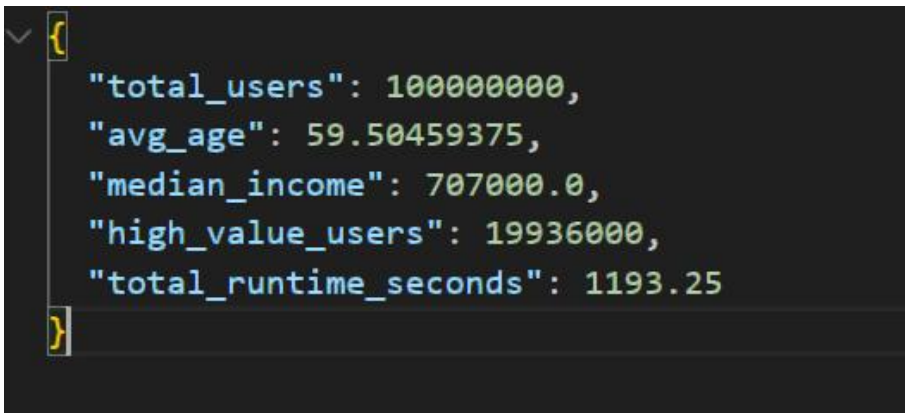


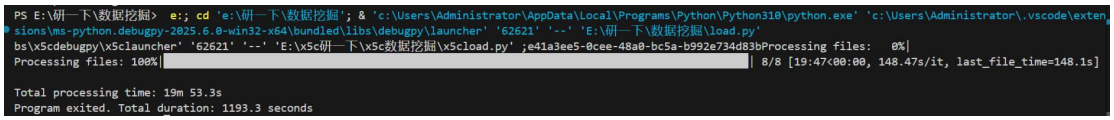


发现这个数据真的可以说是在每个方面都较为的平均，按照图像分析看来就是每个年龄段的收入都有高有低平均收入在 60w 左右，而高价值用户也就是收入超过 80%的人平均收入在 90w 左右，说明在本数据集中高收入和年龄关系不大，而信用分数和年龄成正比。

最后是每个文件的统计信息和总共的统计信息，以及程序运行时间

```
file,total_users,age_mean,income_median,active_rate,high_value_ratio
part-00000.parquet,12500000,58.99214,497000.0,0.0,0.05401
part-00001.parquet,12500000,59.11139,502000.0,0.0,0.05401
part-00002.parquet,12500000,59.00057,497000.0,0.0,0.05362
part-00003.parquet,12500000,59.00504,497000.0,0.0,0.0544
part-00004.parquet,12500000,59.02155,499000.0,0.0,0.0538
part-00005.parquet,12500000,59.0904,500000.0,0.0,0.05464
part-00006.parquet,12500000,59.02141,498000.0,0.0,0.0544
part-00007.parquet,12500000,58.89137,497000.0,0.0,0.05398
```





Part2：30G_data

一、数据读取和数据预处理

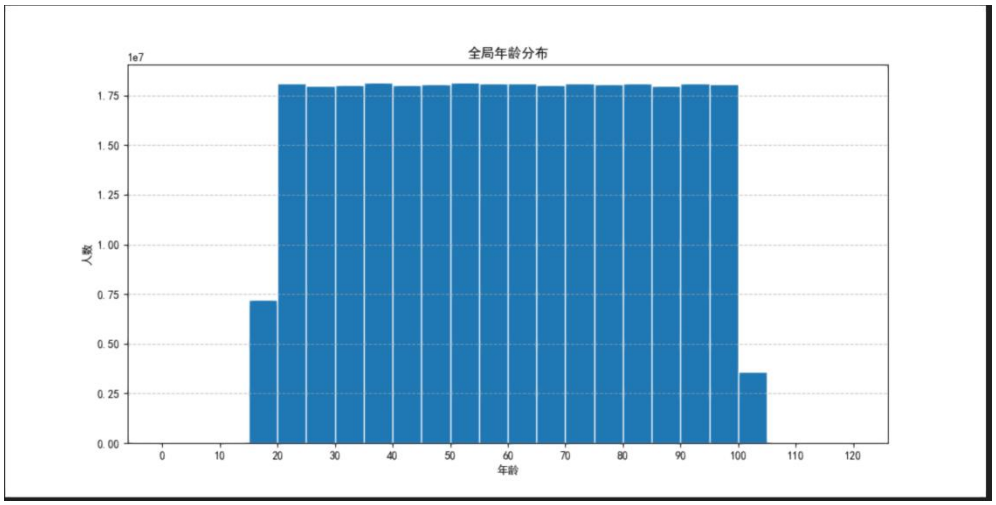
首先从 parquet 文件中读取数据，因为数据本身较大所以不能全部读取出，需要按文件分别读出，先观察数据格式确定所需要的数据内容，尽可能减轻内存负担。发现格式基本上是一样的，所以采用相同的处理方法

就不再反复截图，只是使用相同的方法检测数据的缺失值和异常值

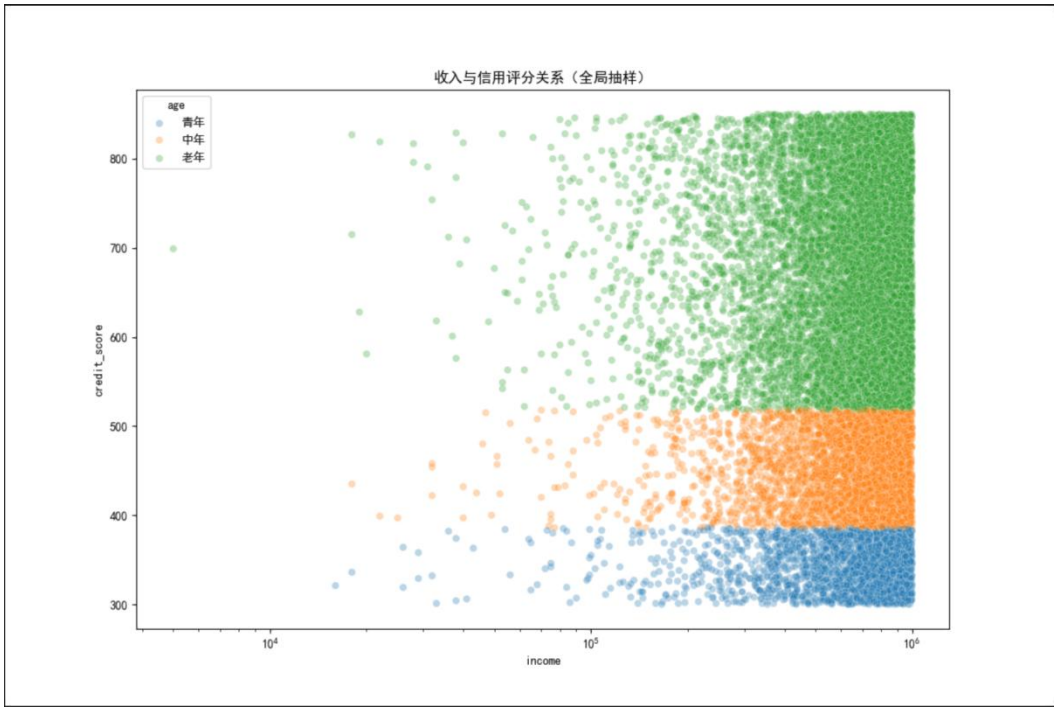
二、探索性分析与可视化

然后对数据进行初步的探索性分析与可视化，首先根据数据内容确定大致探索性分析的几个内容，初步定为，年龄分布可视化，收入分布可视化，用户国家分布可视化。

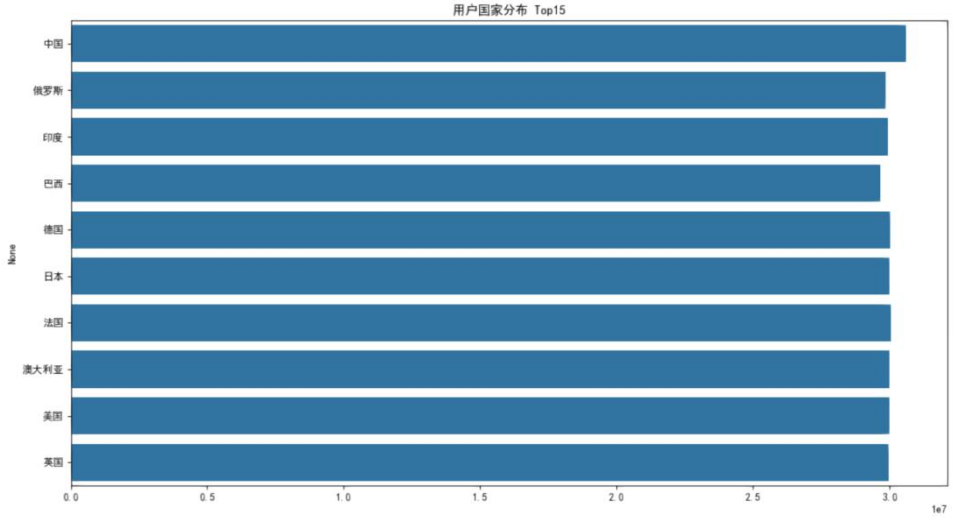
首先可视化年龄分布，以 5 为一个 bin 绘制直方图展示所有用户的年龄分布，可以看到年龄的最小范围在 15-20，最大范围在 100-105，在 15-20 和 100-105 的人数较少，分别为 750w 和 250w 左右，中间的每个年龄段都大致为 1750w，总人数约为 3 亿，可以看到数据所选取的比较标准，中间的年龄段较为平均，两端的年龄段较少，符合客观的现象，说明数据选取较为合理。简单的总结其实就是和 10G 的数据差不多的特征。



然后是收入与信用分数的关系的抽样可视化，并且按照年龄段粗略的展示了青年，中年和老年的分布，发现基本上信用分数是随着年龄提升而提升的，并且我们加大了高收入的可见性，因为我们并不想过于了解低收入和信用评分的关系，主要是想看看信用评分除了收入是否还有年龄因素的影响。同样和 10G 的数据是一样的展示

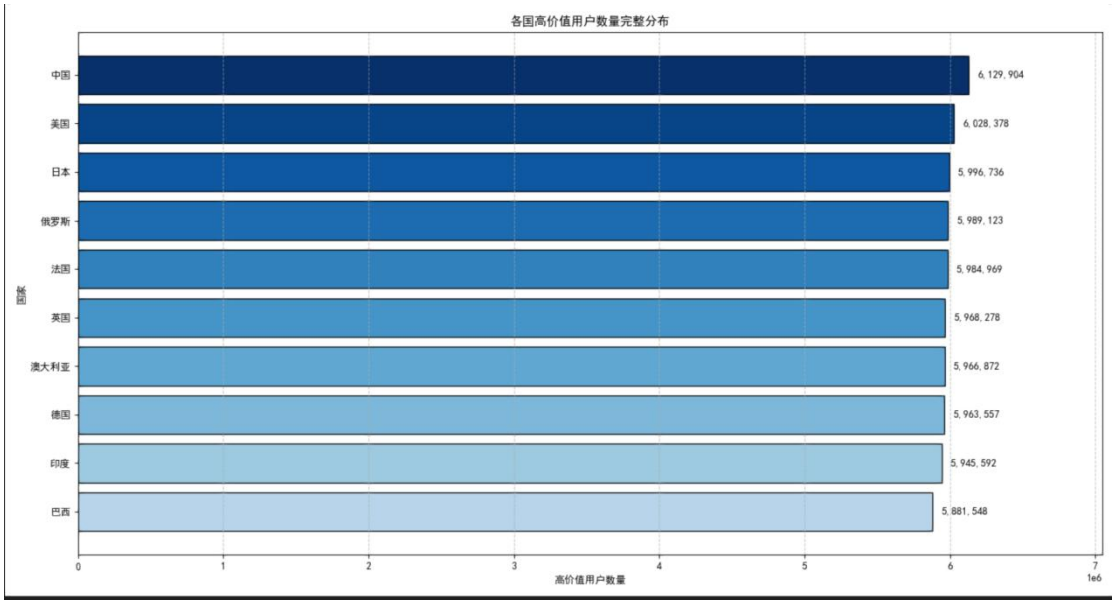


最后是用户国家分布可视化，又一次展示了数据的合理性，发现一共有十个国家每个国家人数都在 3000w 左右，也非常的均衡。

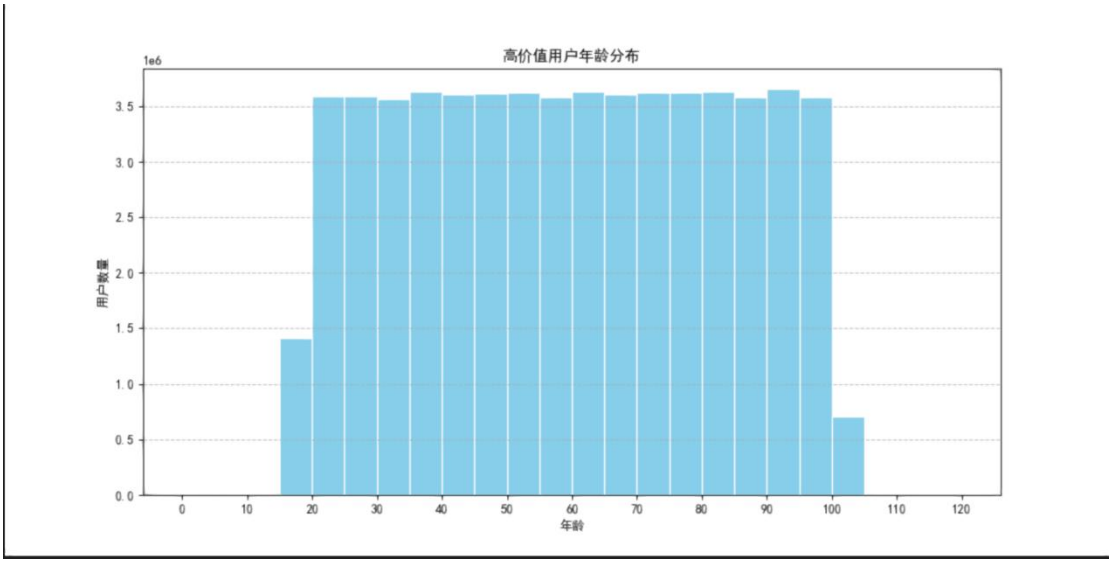


三、数据分析

同样的，选择分析目标为潜在高价值用户，首先定义潜在高价值用户，经过 10G 数据的展示发现信用分数和年龄高度相关，于是为了定义的合理性，在这里将其定义为收入超过80%的的用户，生成一个按国家的高价值用户数量分布。



再可视化高价值用户年龄的分布



最后依然是对文件数据的统计信息和运行时间。

```
file,total_users,age_mean,income_median,active_rate,high_value_ratio
part-00000.parquet,18750000,58.99195989333333,497000.0,0.0,0.19926261333333334
part-00001.parquet,18750000,59.11170554666667,502000.0,0.0,0.19985744
part-00002.parquet,18750000,59.00052,497000.0,0.0,0.19926373333333333
part-00003.parquet,18750000,59.004594346666664,497000.0,0.0,0.19908474666666667
part-00004.parquet,18750000,59.021206666666664,499000.0,0.0,0.19949525333333334
part-00005.parquet,18750000,59.090218613333334,500000.0,0.0,0.19912277333333334
part-00006.parquet,18750000,59.02126341333334,498000.0,0.0,0.19914005333333334
part-00007.parquet,18750000,58.89100309333333,497000.0,0.0,0.19967125333333333
part-00008.parquet,18750000,58.9957704,500000.0,0.0,0.19962442666666666
part-00009.parquet,18750000,59.05765253333333,500000.0,0.0,0.19936794666666666
part-00010.parquet,18750000,59.00241136,499000.0,0.0,0.19944298666666666
part-00011.parquet,18750000,58.944652053333336,502000.0,0.0,0.19994442666666667
part-00012.parquet,18750000,59.05537098666667,501000.0,0.0,0.19984901333333333
part-00013.parquet,18750000,58.95251413333333,500000.0,0.0,0.19991210666666667
part-00014.parquet,18750000,59.00627424,502000.0,0.0,0.1993016
part-00015.parquet,18750000,58.88057002666667,498000.0,0.0,0.199924
```


```
{
  "total_users": 300000000,
  "avg_age": 59.49085121666667,
  "median_income": 709000.0,
  "high_value_users": 59854957,
  "total_runtime_seconds": 3465.88
}
```


```
6.0-win32-x64\bundled\libs\debugpy\launcher' '62278' '--' 'E:\研一下\数据挖掘\load.py'
 '--' 'E:\x5c研一下\x5c数据挖掘\x5cload.py' ;e41a3ee5-0cee-48a0-bc5a-b992e734d83bProcessing files:   0%|
Processing files: 100%|                                                                                | 16/16 [57:35<00:00, 215.96s/it, last_file_time=211.8s]

Total processing time: 57m 45.9s
Program exited. Total duration: 3466.0 seconds
```

总结

对于本次作业给出的两组数据大致的特点都是一致的于是可以总体一致性分析，首先对其进行数据缺失值和异常值检测来评价数据质量，发现没有缺失值也没有范围外的异常值，说明数据非常的不错，然后对数据进行探索性分析对其统计信息进行可视化，展示了数据的均匀性，基本上每个文件中包含的数据的统计信息都是相似的，这也说明了数据本身应该是在清洗之后所得到的质量较高选取了较为典型的数据，并且年龄的分布，收入的分布等也较为合理。通过这次作业加深了对数据预处理与探索性分析任务的理解，不过略有不足的是在探索性分析部分还是不太熟练，不太明白探索分析数据的什么样的特点才是典型的，以及对于数据之间的关系的探索还不够熟练。