互评作业1：数据探索性分析与预处理 一. 10G_data_new 数据集

1. 数据概览 45,000,000行数据，15个属性

In [1]:
```
import os
import pandas as pd
from pathlib import Path


data_folder = Path("./10G_data_new")
parquet_files = list(data_folder.glob("part-*.parquet"))
df_list = (pd.read_parquet(file) for file in parquet_files)
data = pd.concat(df_list, ignore_index=True)
print(f"\n成功加载数据，总行数: {len(data):,}")
```

```
成功加载数据，总行数: 45,000,000
```

In [2]:
```
# 显示数据概览
print("\n=== 数据概览 ===")
print(f"数据集形状: {data.shape} (行数, 列数)")
print("\n前5行数据:")
print(data.head(5))
print("\n列名和数据类型:")
print(data.dtypes)
```

```
=== 数据概览 ===
数据集形状: (45000000, 15) (行数, 列数)

前5行数据:
   id                last_login   user_name fullname                email  \
0   0  2024-12-02T03:49:12+00:00  RKWKCXRZFV      瞿紫玉  kuegujsk@hotmail.com
1   1  2024-08-25T05:39:16+00:00      RCLELJ      李泽宸      wslfszer@126.com
2   2  2023-12-21T14:28:09+00:00       KSHSK      詹紫玥      gputsgbf@126.com
3   3  2023-06-06T03:21:09+00:00    CCJMXPJA      姜小红  akidhwzo@outlook.com
4   4  2024-10-08T11:02:18+00:00     TJRJDNO      童泽楠       suupywzi@qq.com

   age     income gender country                         address  \
0   82  366311.83      女      美国  Non-Chinese Address Placeholder
1   71  833917.30      男      英国                   上海市淄博山水路360号
2   54  839379.17      女    澳大利亚                   北京市东莞保健中心路614号
3   18  383963.16      男      巴西                  山东省株洲配送中心路176号
4   77  337059.32      男      英国                   浙江省赤峰安康路957号

                                  purchase_history  is_active  \
0  {"avg_price":9496,"categories":"零食","items":[{...      False
1  {"avg_price":3014,"categories":"手套","items":[{...       True
2  {"avg_price":8921,"categories":"裙子","items":[{...       True
3  {"avg_price":939,"categories":"耳机","items":[{"...      False
4  {"avg_price":959,"categories":"手套","items":[{"...      False

  registration_date       phone_number  \
0        2024-10-31  +1 (804) 855-6279
1        2023-01-13    +44 1850 116429
2        2022-07-06    +61 656 440 523
3        2020-03-20  +55 54 34995-1600
4        2023-01-05    +44 5383 067377

                                    login_history
0  {"avg_session_duration":105,"devices":["deskto...
1  {"avg_session_duration":64,"devices":["mobile"...
2  {"avg_session_duration":116,"devices":["deskto...
3  {"avg_session_duration":25,"devices":["mobile"...
4  {"avg_session_duration":51,"devices":["desktop...

列名和数据类型:
id                     int64
last_login            object
user_name             object
fullname              object
email                 object
age                    int64
income               float64
gender                object
country               object
address               object
purchase_history      object
is_active               bool
registration_date     object
phone_number          object
login_history         object
dtype: object
```

2. 数据摘要

In [3]:
```
import numpy as np

num_fields = list(data.select_dtypes(include=np.number).columns.values)
nom_fields = list(data.select_dtypes(exclude=np.number).columns.values)
print('标称属性:', nom_fields)
print('数值属性:', num_fields)
```

```
标称属性: ['last_login', 'user_name', 'fullname', 'email', 'gender', 'country',
'address', 'purchase_history', 'is_active', 'registration_date', 'phone_number',
'login_history']
数值属性: ['id', 'age', 'income']
```

1. 标称属性 对标称属性进行频数统计

In [4]:
```
for field in nom_fields:
    print('频数统计:')
    print(data[field].value_counts())
```

```
频数统计:
last_login
2024-12-01T14:39:34+00:00    8
2024-09-23T17:06:51+00:00    8
2024-02-26T21:42:18+00:00    8
2024-07-20T11:47:38+00:00    8
2025-03-16T10:26:23+00:00    8
                            ..
2024-01-03T18:21:12+00:00    1
2023-09-01T19:39:53+00:00    1
2025-01-14T06:37:46+00:00    1
2024-01-15T07:27:45+00:00    1
2024-07-16T23:18:39+00:00    1
Name: count, Length: 33332828, dtype: int64
频数统计:
user_name
VBBMF         9
KNWHC         8
ANEES         8
ZCTDD         7
UFLTZ         7
             ..
YXIXFN        1
NCCUBHUNXR    1
KMMKHQKAYD    1
CKSORUBOUZ    1
PKDOXKCZVR    1
Name: count, Length: 42970059, dtype: int64
频数统计:
fullname
常紫薇    4864
卜紫轩    4829
卜紫欣    4812
薄紫薇    4805
路紫宁    4788
       ...
覃小红     497
萧子轩     495
银华      495
仲霞      489
翟泽鸿     489
Name: count, Length: 56520, dtype: int64
频数统计:
email
rgamnqql@163.com        2
tmqolouf@gmail.com      2
qylernim@hotmail.com    2
jenionmf@126.com        2
kxbfpxwa@hotmail.com    2
                      ..
vzdmpsxd@hotmail.com    1
qvsxunft@163.com        1
rwwsvhgj@126.com        1
dqsqyeat@gmail.com      1
icxgafnf@hotmail.com    1
Name: count, Length: 44999227, dtype: int64
频数统计:
gender
男      21603397
女      21598086
```

```
未指定      899652
其他       898865
Name: count, dtype: int64
频数统计:
country
英国      4501669
法国      4501427
美国      4501158
巴西      4500526
德国      4500370
印度      4499562
俄罗斯     4499132
澳大利亚    4499124
日本      4498695
中国      4498337
Name: count, dtype: int64
频数统计:
address
Non-Chinese Address Placeholder    9003537
广西壮族自治区锡林郭勒盟银河路872号                        5
天津市长春湿地公园路6号                                 5
甘肃省武汉兴隆路430号                                5
云南省哈尔滨托儿所路496号                               5
                                  ...
陕西省运城金山路203号                                1
山东省石家庄长江路420号                               1
湖北省太原便利店路891号                               1
台湾省大兴安岭体育馆路710号                              1
辽宁省佳木斯步行街485号                               1
Name: count, Length: 35017192, dtype: int64
频数统计:
purchase_history
{"avg_price":9496,"categories":"零食","items":[{"id":7265}],"payment_method":"现
金","payment_status":"已支付","purchase_date":"2023-07-30"}
1
{"avg_price":1729,"categories":"手套","items":[{"id":4987},{"id":1653},{"id":821
4},{"id":3688},{"id":581}],"payment_method":"信用卡","payment_status":"部分退
款","purchase_date":"2022-08-19"}      1
{"avg_price":4066,"categories":"蔬菜","items":[{"id":684},{"id":3024},{"id":492
6},{"id":1030}],"payment_method":"信用卡","payment_status":"部分退款","purchase_d
ate":"2020-09-01"}                     1
{"avg_price":6269,"categories":"水产","items":[{"id":4709}],"payment_method":"储
蓄卡","payment_status":"部分退款","purchase_date":"2024-09-14"}
1
{"avg_price":9079,"categories":"床上用品","items":[{"id":7985},{"id":8981},{"id":
1091}],"payment_method":"支付宝","payment_status":"部分退款","purchase_date":"202
3-06-21"}                              1

..
{"avg_price":5479,"categories":"笔记本电脑","items":[{"id":4990},{"id":2385},{"i
d":4508},{"id":284},{"id":4108}],"payment_method":"储蓄卡","payment_status":"已退
款","purchase_date":"2025-02-01"}    1
{"avg_price":1333,"categories":"帽子","items":[{"id":833}],"payment_method":"云闪
付","payment_status":"已退款","purchase_date":"2024-09-30"}
1
{"avg_price":6747,"categories":"车载电子","items":[{"id":1102},{"id":8077}],"paym
ent_method":"现金","payment_status":"已支付","purchase_date":"2020-05-14"}
1
{"avg_price":3555,"categories":"儿童课外读物","items":[{"id":9956},{"id":467
4}],"payment_method":"微信支付","payment_status":"已退款","purchase_date":"2023-0
```

```
8-28"}                                          1
{"avg_price":1260,"categories":"车载电子","items":[{"id":5721}],"payment_metho
d":"信用卡","payment_status":"已支付","purchase_date":"2022-10-14"}
1
Name: count, Length: 45000000, dtype: int64
频数统计:
is_active
False    22501308
True     22498692
Name: count, dtype: int64
频数统计:
registration_date
2023-03-17    24069
2022-09-17    24031
2022-02-10    24021
2023-10-01    24008
2020-07-04    23998
              ...
2020-04-17    23135
2022-05-06    23131
2023-12-01    23111
2024-11-19    23100
2023-03-14    23092
Name: count, Length: 1910, dtype: int64
频数统计:
phone_number
+33 0 21 70 94 50    3
+33 7 03 98 96 79    3
+44 9165 015782      3
+61 036 861 894      3
+33 2 59 01 73 96    3
                    ..
+61 931 224 346      1
+7 320 265-09-09     1
+91 23225 89092      1
+81 99-7416-1320     1
+44 7768 410638      1
Name: count, Length: 44973597, dtype: int64
频数统计:
login_history
{"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 18:26:0
0","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"]}
1
{"avg_session_duration":5,"devices":["mobile","tablet"],"first_login":"2023-05-1
1","locations":["home","travel"],"login_count":32,"timestamps":["2023-05-11 22:5
0:00","2023-07-12 23:04:00","2023-09-12 13:21:00","2023-11-13 16:17:00","2024-01-
14 23:00:00","2024-03-16 23:01:00","2024-05-17 17:02:00","2024-07-18 13:14:00","2
024-09-18 17:57:00","2024-11-19 12:32:00","2025-01-20 22:33:00"]}
1
{"avg_session_duration":70,"devices":["desktop"],"first_login":"2024-07-17","loca
tions":["home","work"],"login_count":53,"timestamps":["2024-07-17 18:04:00","2024
-08-04 09:46:00","2024-08-22 19:11:00","2024-09-09 20:46:00","2024-09-27 19:53:0
0","2024-10-15 11:31:00","2024-11-02 11:41:00","2024-11-20 19:44:00","2024-12-08
11:15:00","2024-12-26 11:31:00","2025-01-13 18:19:00","2025-01-31 20:32:00","2025
-02-18 18:41:00","2025-03-08 20:14:00"]}
1
```

{"avg_session_duration":116,"devices":["desktop"],"first_login":"2024-12-18","locations":["travel","work"],"login_count":26,"timestamps":["2024-12-18 20:13:00","2024-12-28 10:01:00","2025-01-07 09:18:00","2025-01-17 11:20:00","2025-01-27 07:24:00","2025-02-06 21:41:00","2025-02-16 06:28:00","2025-02-26 18:23:00","2025-03-08 21:12:00","2025-03-18 20:26:00"]}
1
{"avg_session_duration":111,"devices":["desktop","tablet"],"first_login":"2024-06-10","locations":["work"],"login_count":59,"timestamps":["2024-06-22 20:17:00","2024-06-24 11:10:00","2024-08-08 07:48:00","2024-08-14 07:27:00","2024-09-01 19:02:00","2024-09-05 20:20:00","2024-09-29 18:35:00","2024-09-29 20:38:00","2024-10-27 20:26:00","2024-11-23 18:52:00","2024-12-14 20:01:00","2025-02-27 09:47:00","2025-03-03 09:10:00","2025-03-06 09:51:00"]}
1

..
{"avg_session_duration":117,"devices":["mobile"],"first_login":"2023-08-10","locations":["home","work"],"login_count":97,"timestamps":["2023-08-10 16:37:00","2023-09-09 17:50:00","2023-10-09 23:58:00","2023-11-08 23:13:00","2023-12-08 22:40:00","2024-01-07 12:58:00","2024-02-06 23:45:00","2024-03-07 22:09:00","2024-04-06 14:32:00","2024-05-06 22:14:00","2024-06-05 23:57:00","2024-07-05 23:58:00","2024-08-04 23:26:00"]}
1
{"avg_session_duration":104,"devices":["tablet"],"first_login":"2024-01-15","locations":["home","work"],"login_count":91,"timestamps":["2024-01-15 22:59:00","2024-02-19 22:49:00","2024-03-25 23:55:00","2024-04-29 23:20:00","2024-06-03 22:17:00","2024-07-08 22:48:00","2024-08-12 15:15:00","2024-09-16 22:17:00","2024-10-21 23:07:00","2024-11-25 23:31:00","2024-12-30 23:58:00","2025-02-03 23:03:00"]}
1
{"avg_session_duration":32,"devices":["desktop"],"first_login":"2024-05-23","locations":["travel"],"login_count":98,"timestamps":["2024-05-23 16:51:00","2024-06-06 13:17:00","2024-06-20 16:35:00","2024-07-04 17:35:00","2024-07-18 12:24:00","2024-08-01 17:44:00","2024-08-15 23:06:00","2024-08-29 13:47:00","2024-09-12 23:01:00","2024-09-26 12:32:00","2024-10-10 12:48:00","2024-10-24 16:25:00","2024-11-07 15:50:00","2024-11-21 12:44:00","2024-12-05 23:11:00","2024-12-19 22:33:00","2025-01-02 15:09:00","2025-01-16 16:43:00"]}   1
{"avg_session_duration":83,"devices":["desktop","mobile","tablet"],"first_login":"2024-05-19","locations":["travel"],"login_count":8,"timestamps":["2024-05-19 21:47:00","2024-06-23 10:48:00","2024-07-28 18:44:00","2024-09-01 20:38:00","2024-10-06 18:39:00","2024-11-10 20:37:00","2024-12-15 18:46:00","2025-01-19 08:44:00"]}
1
{"avg_session_duration":79,"devices":["desktop","tablet"],"first_login":"2024-04-07","locations":["home","work"],"login_count":65,"timestamps":["2024-04-07 00:45:00","2024-04-07 01:08:00","2024-04-07 01:29:00","2024-04-07 02:49:00","2024-04-07 07:39:00","2024-04-07 07:46:00","2024-04-07 15:03:00","2024-04-07 15:18:00","2024-04-07 16:27:00","2024-04-07 17:06:00","2024-04-07 21:15:00","2024-04-07 21:28:00","2024-04-07 22:23:00","2024-04-07 22:40:00","2024-04-07 22:46:00","2024-04-07 23:04:00","2024-04-07 23:47:00"]}   1
Name: count, Length: 45000000, dtype: int64

2. 数值属性

In [5]:
```
print(data.describe())
```

```
                 id           age        income
count  4.500000e+07  4.500000e+07  4.500000e+07
mean   2.250000e+07  5.899862e+01  4.999971e+05
std    1.299038e+07  2.395833e+01  2.886931e+05
min    0.000000e+00  1.800000e+01  1.000000e-02
25%    1.125000e+07  3.800000e+01  2.499804e+05
50%    2.250000e+07  5.900000e+01  4.999420e+05
75%    3.375000e+07  8.000000e+01  7.500388e+05
max    4.500000e+07  1.000000e+02  1.000000e+06
```

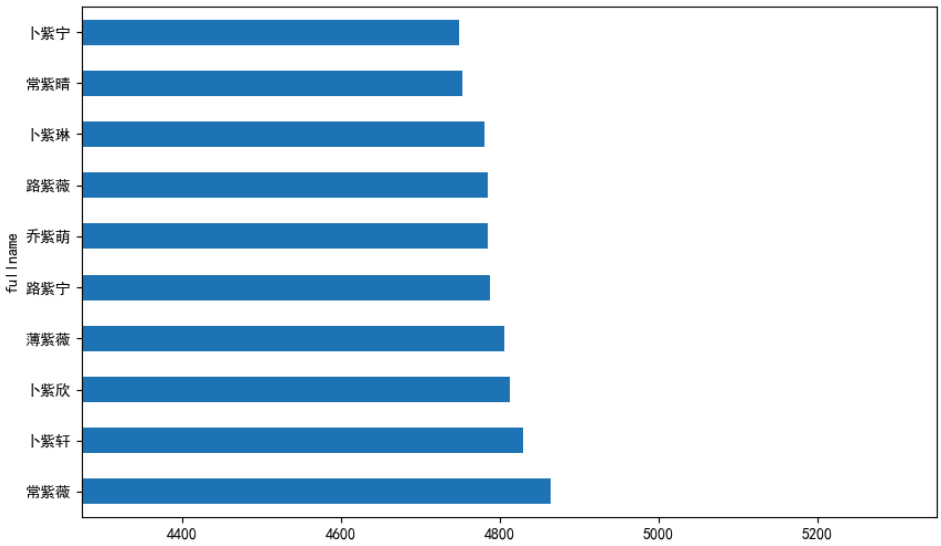
In [6]:
```
# 缺失值统计
for field in num_fields:
    print(field+':',data[field].isnull().sum())
```

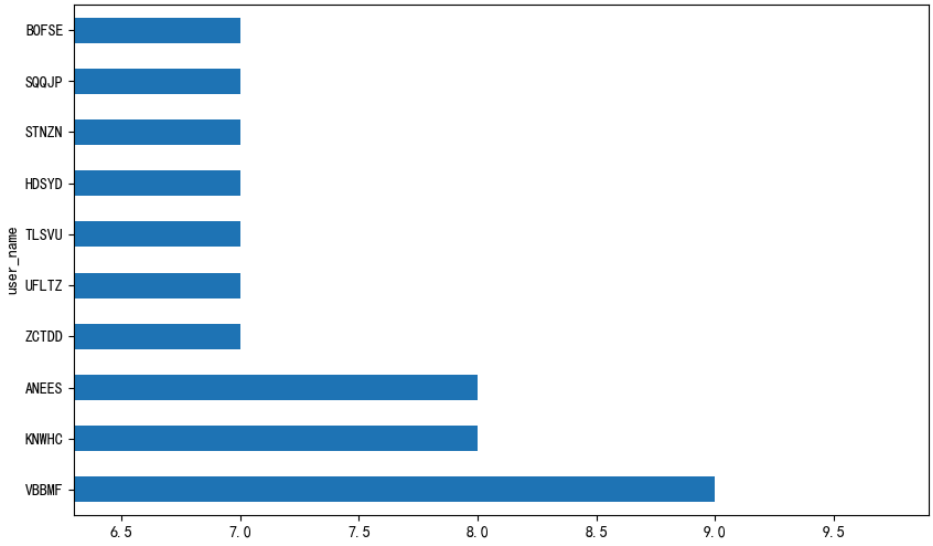
```
id: 0
age: 0
income: 0
```

3. 数据可视化

1. 标称属性

In [7]:
```
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
from matplotlib import rcParams

# 设置中文字体（Windows系统通常有SimHei）
rcParams['font.sans-serif'] = ['SimHei']  # 或者 ['Microsoft YaHei']
rcParams['axes.unicode_minus'] = False  # 解决负号显示问题

for field in nom_fields:
    fig_path = 'fig1/'+ field + '.png'

    # 创建图形
    plt.figure()

    counts = data[field].value_counts().head(10)
    ax = counts.plot.barh(figsize=(10, 6))
    # 仅显示最小值到最大值的区间（避免从0开始）
    ax.set_xlim(counts.min() * 0.9, counts.max() * 1.1)  # 留10%边距

    # 保存图形
    plt.savefig(fig_path)

    # 关闭图形，避免内存泄漏
    plt.close()
```

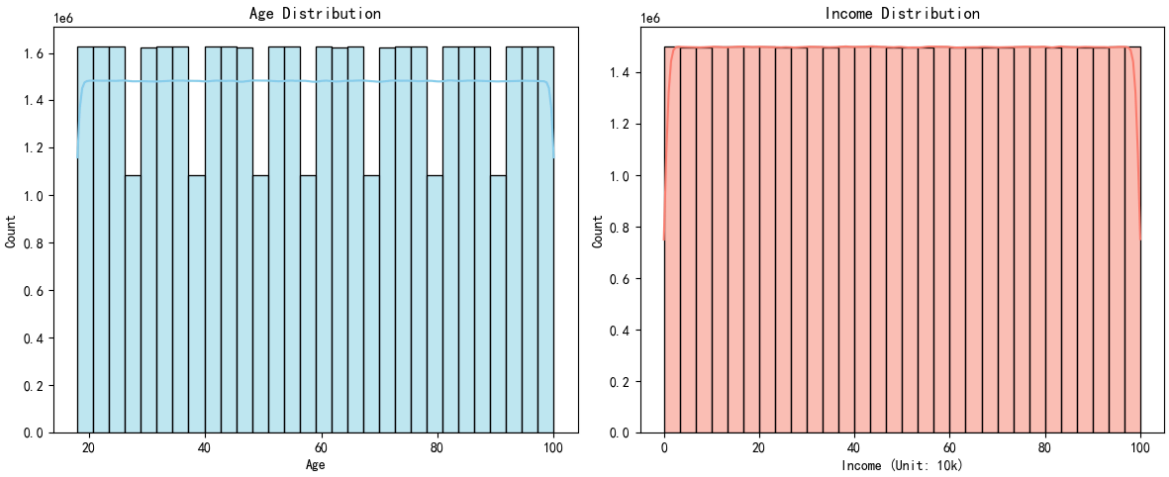




2. 数值属性

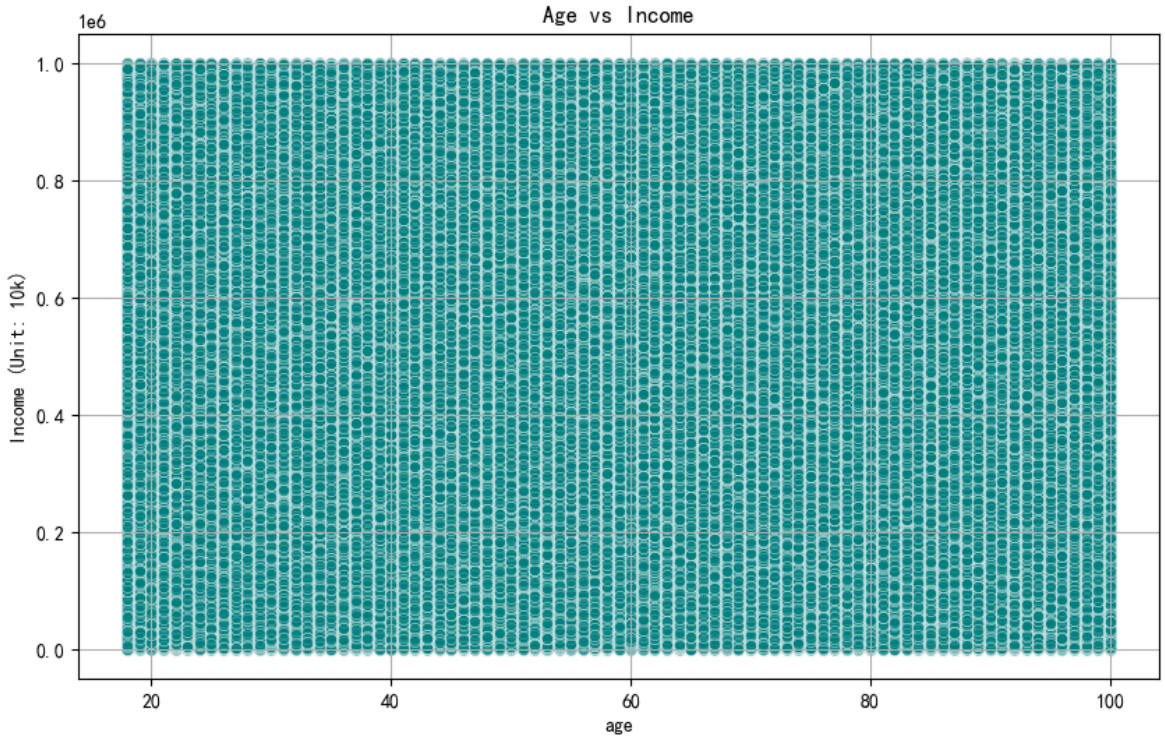
In [8]:
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 5))

# Age分布
plt.subplot(1, 2, 1)
sns.histplot(data['age'], bins=30, kde=True, color='skyblue')
plt.title('Age Distribution')
plt.xlabel('Age')

data['income_10k'] = data['income'] / 10000
# Income分布（单位：万）
plt.subplot(1, 2, 2)
```

```
sns.histplot(data['income_10k'], bins=30, kde=True, color='salmon')
plt.title('Income Distribution')
plt.xlabel('Income (Unit: 10k)')  # 明确标注单位

plt.tight_layout()
plt.show()
```

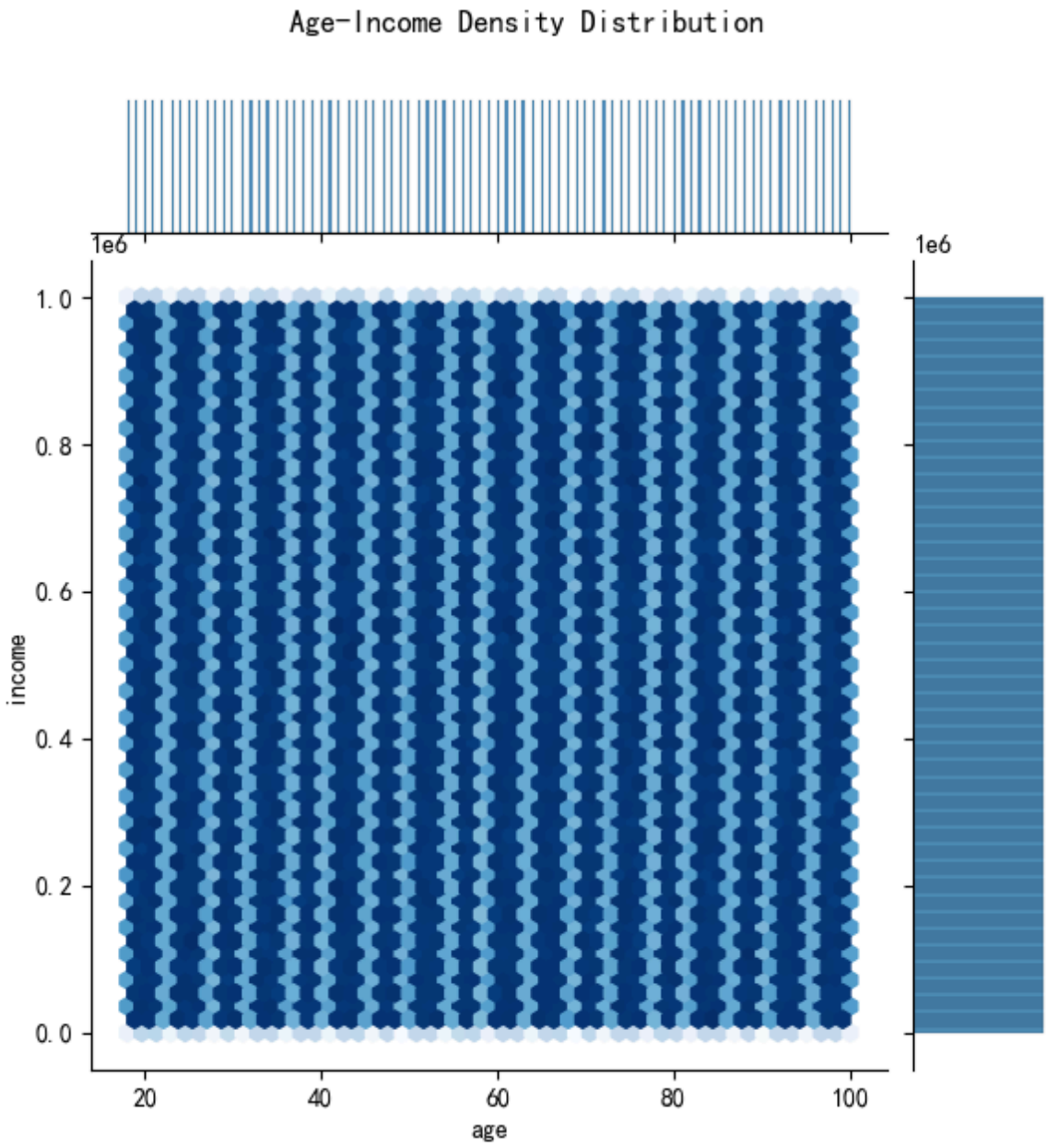


可以看出，年龄和收入的整体分布比较均匀。 此外，还可分析二者的联合关系。分别输出散点图和六边形分箱图。

In [9]:
```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='age', y='income', data=data, alpha=0.6, color='teal')
plt.title('Age vs Income')
plt.ylabel('Income (Unit: 10k)')
plt.grid(True)
plt.show()

plt.figure(figsize=(10, 6))
sns.jointplot(x='age', y='income', data=data, kind='hex', cmap='Blues')
plt.suptitle('Age-Income Density Distribution', y=1.05)
plt.show()
```



```
<Figure size 1000x600 with 0 Axes>
```



从图中可以看出，年龄和收入之间并没有什么相关性，二者分布都十分均匀。

二. 30G_data_new 数据集

1. 数据概览

In [10]:
```
import os
import pandas as pd
from pathlib import Path


data_folder1 = Path("./30G_data_new")
parquet_files1 = list(data_folder1.glob("part-*.parquet"))
df_list1 = (pd.read_parquet(file) for file in parquet_files1)
data1 = pd.concat(df_list1, ignore_index=True)
print(f"\n成功加载数据，总行数: {len(data1):,}")
```

```
成功加载数据，总行数: 135,000,000
```

In [11]:
```
# 显示数据概览
print("\n=== 数据概览 ===")
print(f"数据集形状: {data1.shape} (行数, 列数)")
print("\n前5行数据:")
print(data1.head(5))
print("\n列名和数据类型:")
print(data1.dtypes)
```

```
=== 数据概览 ===
数据集形状: (135000000, 15) (行数, 列数)

前5行数据:
   id                last_login   user_name fullname              email  \
0   0  2024-03-19T19:35:16+00:00  OFVIUGZMWH      覃泽川  lnsqjypb@gmail.com
1   1  2025-02-21T05:08:16+00:00      KMLBNE      吕泽越     zddfsdkt@qq.com
2   2  2024-11-26T09:33:05+00:00     NGTSMVK      卞泽楠    qxgqdrfd@163.com
3   3  2023-10-19T17:32:56+00:00      IJLZVS       卞鹏    jbjxirrf@163.com
4   4  2024-05-09T00:01:29+00:00       XCLES       郎雪     cnerwras@qq.com

   age     income gender country                          address  \
0   97    7787.23      女      日本                    西藏自治区鹤岗河滨路827号
1   31  286306.19      男      美国  Non-Chinese Address Placeholder
2   82  136343.81      男      日本                     重庆市保定检察院路503号
3   90  179801.85      男      印度                    江西省南京儿童乐园路510号
4   73  918006.25      女      英国                    福建省南宁技术学院路988号

                                  purchase_history  is_active  \
0  {"avg_price":4041,"categories":"文具","items":[{...      False
1  {"avg_price":3608,"categories":"鞋子","items":[{...      False
2  {"avg_price":6416,"categories":"文具","items":[{...      False
3  {"avg_price":8157,"categories":"办公用品","items":...       True
4  {"avg_price":1626,"categories":"户外装备","items":...      False

  registration_date       phone_number  \
0        2020-10-27   +81 37-3972-6955
1        2021-09-25  +1 (349) 601-0753
2        2023-05-15   +81 09-3007-5554
3        2020-06-29    +91 81513 74738
4        2023-08-11    +44 4509 799780

                                     login_history
0  {"avg_session_duration":14,"devices":["desktop...
1  {"avg_session_duration":46,"devices":["mobile"...
2  {"avg_session_duration":50,"devices":["mobile"...
3  {"avg_session_duration":110,"devices":["deskto...
4  {"avg_session_duration":40,"devices":["tablet"...

列名和数据类型:
id                    int64
last_login           object
user_name            object
fullname             object
email                object
age                   int64
income              float64
gender               object
country              object
address              object
purchase_history     object
is_active              bool
registration_date    object
phone_number         object
login_history        object
dtype: object
```

2. 数据摘要

In [12]:
```
import numpy as np

num_fields1 = list(data1.select_dtypes(include=np.number).columns.values)
nom_fields1 = list(data1.select_dtypes(exclude=np.number).columns.values)
print('标称属性:', nom_fields1)
print('数值属性:', num_fields1)
```

```
标称属性: ['last_login', 'user_name', 'fullname', 'email', 'gender', 'country',
'address', 'purchase_history', 'is_active', 'registration_date', 'phone_number',
'login_history']
数值属性: ['id', 'age', 'income']
```

1. 标称属性 对标称属性进行频数统计

In [ ]:
```
for field in nom_fields1:
    print('频数统计:')
    print(data1[field].value_counts())
```

频数统计: last_login 2023-03-02T06:29:10+00:00 14 2025-03-16T18:56:46+00:00 14 2023-04-01T22:08:54+00:00 13 2025-01-16T18:55:22+00:00 13 2024-12-24T09:31:31+00:00 13 .. 2024-11-25T11:28:51+00:00 1 2023-11-20T22:12:23+00:00 1 2024-07-23T06:04:00+00:00 1 2023-03-30T03:54:01+00:00 1 2023-07-04T01:11:59+00:00 1 Name: count, Length: 60407053, dtype: int64 频数统计: user_name AOOIR 12 XUXPC 12 ZYCBB 12 UDDSC 12 YXCEF 12 .. TGVLBVUVN 1 BTVTJCQB 1 TGNELBVE 1 ZGEHVSE 1 FIAZIK 1 Name: count, Length: 121756585, dtype: int64 频数统计: fullname 卜紫琳 14415 常紫晴 14257 乔紫萌 14249 常紫轩 14227 薄紫宁 14199 ...
晏紫玥 1592 崔涵 1586 台玲 1585 池泽熙 1579 敖泽川 1564 Name: count, Length: 56520, dtype: int64 频数统计: email bdqnnybx@163.com 2 gsohaoqq@gmail.com 2 atnayjzv@hotmail.com 2 ovmupeci@hotmail.com 2 spdayzzr@hotmail.com 2 .. ctctcfwo@gmail.com 1 vyohnhkz@outlook.com 1 swkslhyo@outlook.com 1 rryfxftw@126.com 1 ogtdetuy@163.com 1 Name: count, Length: 134992711, dtype: int64 频数统计: gender 男 64810501 女 64792563 未指定 2698564 其他 2698372 Name: count, dtype: int64 频数统计: country 澳大利亚 13502953 印度 13502855 美国 13502589 俄罗斯 13500996 法国 13499078 日本 13498944 中国 13498904 巴西 13498665 英国 13498183 德国 13496833 Name: count, dtype: int64

2. 数值属性

In [ ]:
```
print(data1.describe())
```

```
                 id           age        income
```

count 4.500000e+07 4.500000e+07 4.500000e+07 mean 2.250000e+07 5.899862e+01 4.999971e+05 std 1.299038e+07 2.395833e+01 2.886931e+05 min 0.000000e+00 1.800000e+01 1.000000e-02 25% 1.125000e+07 3.800000e+01 2.499804e+05 50% 2.250000e+07 5.900000e+01 4.999420e+05 75%

3.375000e+07 8.000000e+01 7.500388e+05 max 4.500000e+07 1.000000e+02 1.000000e+06

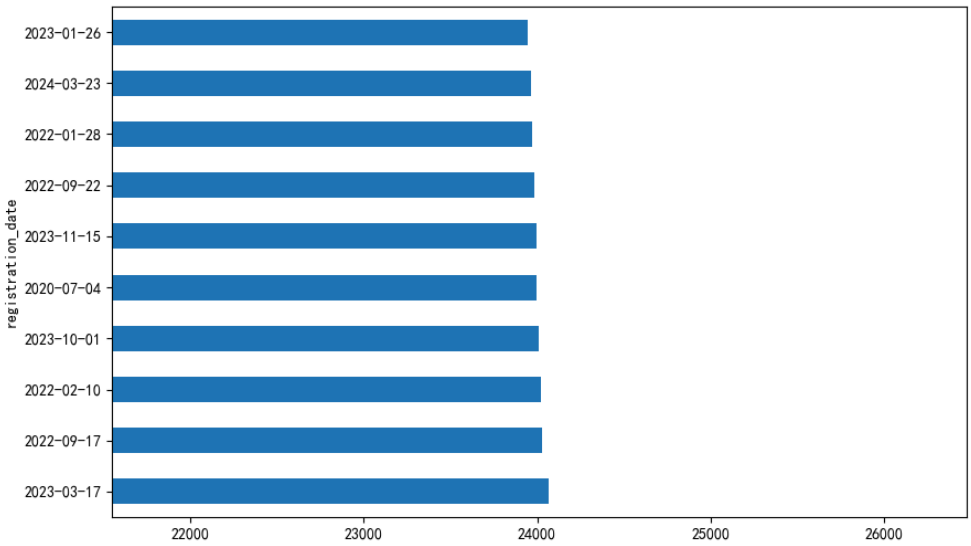
In [ ]:
```
# 缺失值统计
for field in num_fields1:
    print(field+':',data1[field].isnull().sum())
```

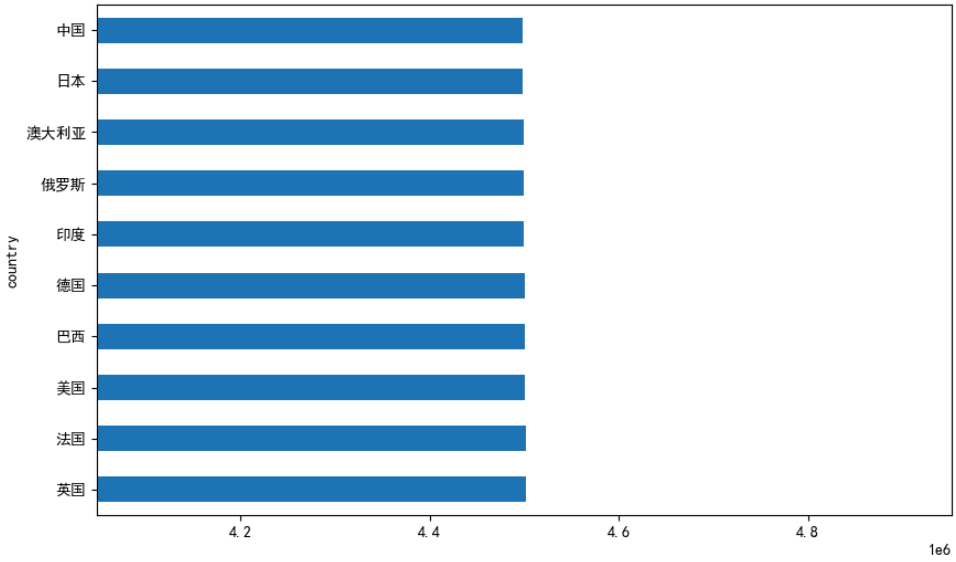
id: 0 age: 0 income: 0

3. 数据可视化

1. 标称属性

In [ ]:
```
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
from matplotlib import rcParams

# 设置中文字体（Windows系统通常有SimHei）
rcParams['font.sans-serif'] = ['SimHei']  # 或者 ['Microsoft YaHei']
rcParams['axes.unicode_minus'] = False  # 解决负号显示问题

for field in nom_fields1:
    fig_path = 'fig2/'+ field + '.png'

    # 创建图形
    plt.figure()

    counts = data1[field].value_counts().head(10)
    ax = counts.plot.barh(figsize=(10, 6))
    # 仅显示最小值到最大值的区间（避免从0开始）
    ax.set_xlim(counts.min() * 0.9, counts.max() * 1.1)  # 留10%边距

    # 保存图形
    plt.savefig(fig_path)

    # 关闭图形，避免内存泄漏
    plt.close()
```


registration_date
2023-01-26
2024-03-23
2022-01-28
2022-09-22
2023-11-15
2020-07-04
2023-10-01
2022-02-10
2022-09-17
2023-03-17
22000
23000
24000
25000
26000


country
中国
日本
澳大利亚
俄罗斯
印度
德国
巴西
美国
法国
英国
4.2
4.4
4.6
4.8
1e6