本次实验旨在对给定的用户数据进行分析，包括人口统计特征（国家、城市分布）、用户在不同时间段（时区）活动的特征，以及基于事件类型和行为（event_action）推断的用户提交活跃度。

1.数据简介与预处理：

数据包含以下关键字段：

- user_id：用户唯一标识
- location：用户所在地信息（可能为 "城市，州/省" 格式，也可能只有一个地点名）
- country：用户所在国家
- event_time：事件发生时间
- event_action：事件行为（"created"、"added" 等）
- total_influence：用户影响力指标

数据预处理

对 location 字段进行解析：

将包含逗号的记录视为 "城市，州/省" 格式，从中提取城市信息；若只包含一个词，则不将其视为有效的城市数据。

对缺失值进行填充或剔除：

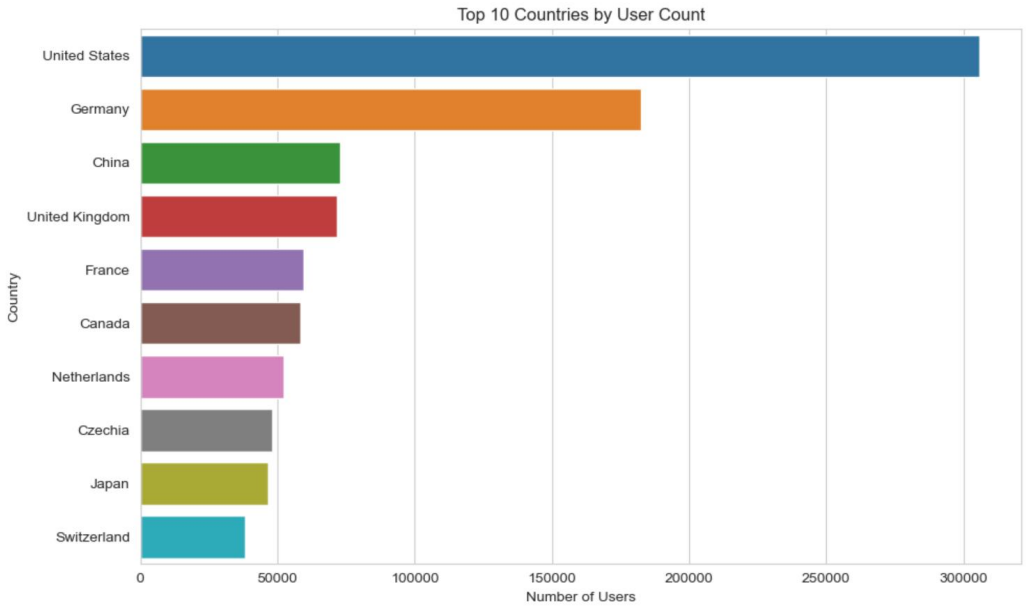
如 country 中的缺失值填充为 "Unknown"；对无城市信息的记录不纳入城市统计分析。

将 event_time 转换为 datetime 类型，以便根据事件发生小时提取用户活跃时间分布特征。

## 2.分析结果与可视化

### 2.1 国家分布分析

```
Top 10 Countries by User Count:
country
United States     305788
Germany           182659
China              73011
United Kingdom     71606
France             59570
Canada             58600
Netherlands        52367
Czechia            48122
Japan              46553
Switzerland        38093
```

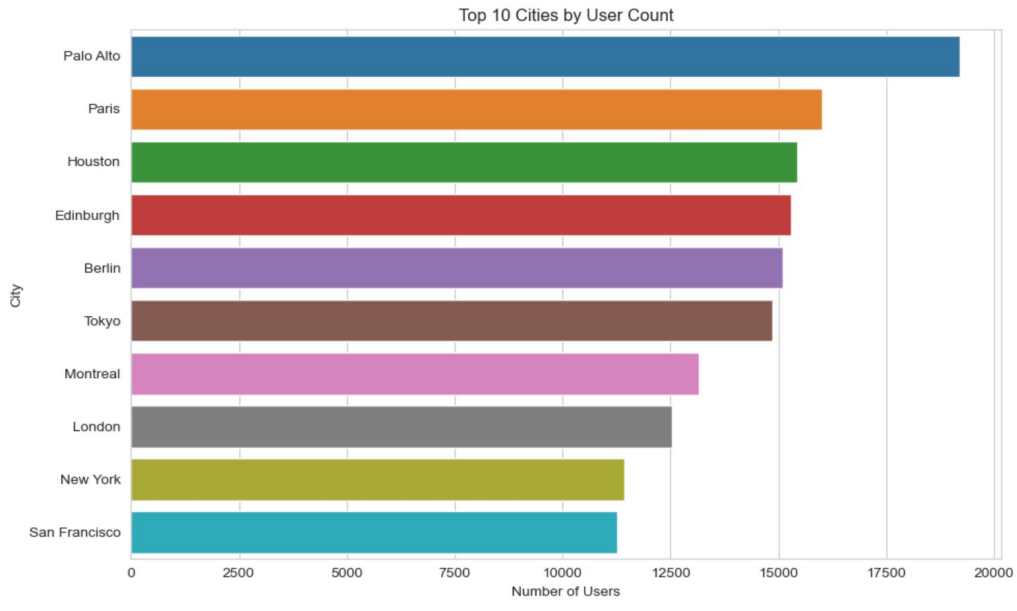


从结果看，用户主要集中在美国、德国、中国、英国等国家，这些国家的开发者数量相对较多。这些信息可能反映了项目国际化程度，以及欧美和亚洲的用户群体活跃度。

## 2.2 城市分布分析

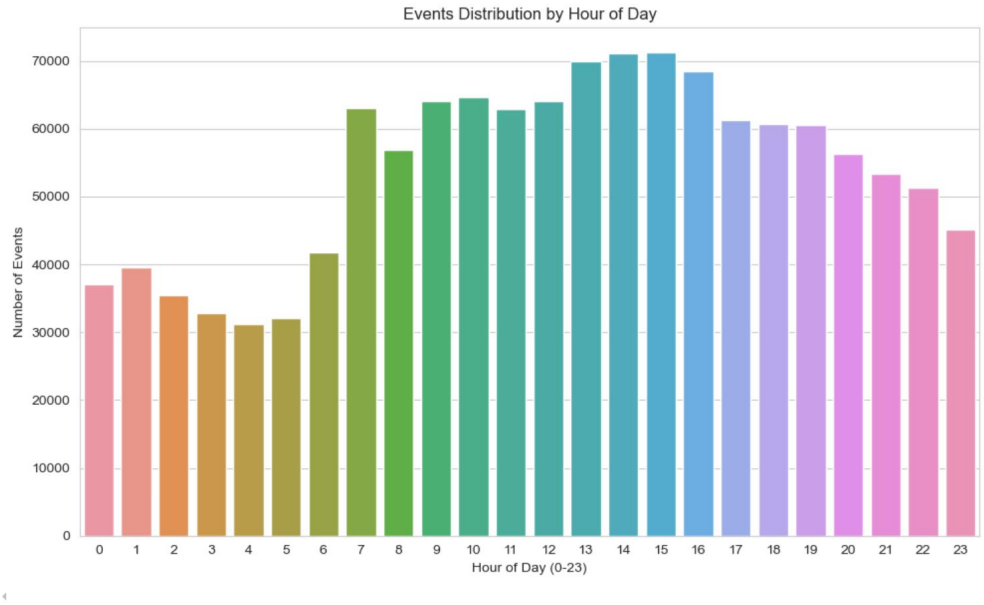
先对 location 进行解析并过滤掉无效城市（有的数据只显示一个单词，这样会导致把国家当成城市，应该先识别数据是否为两个单词，如果只有一个，把他删掉，只读取有两个单词的前一个当作城市名）

```
Top 10 Cities by User Count:
city
Palo Alto        19215
Paris            16021
Houston          15449
Edinburgh        15308
Berlin           15095
Tokyo            14877
Montreal         13171
London           12546
New York         11441
San Francisco    11271
```



可以看到，技术重镇以及国际化大城市（如 Palo Alto、Paris、Berlin、Tokyo、London、New York、San Francisco）在前十名中频频出现。这些地区往往拥有发达的科技生态和创业氛围。

2.3 基于事件时间的活跃时段分布（时区分析近似）



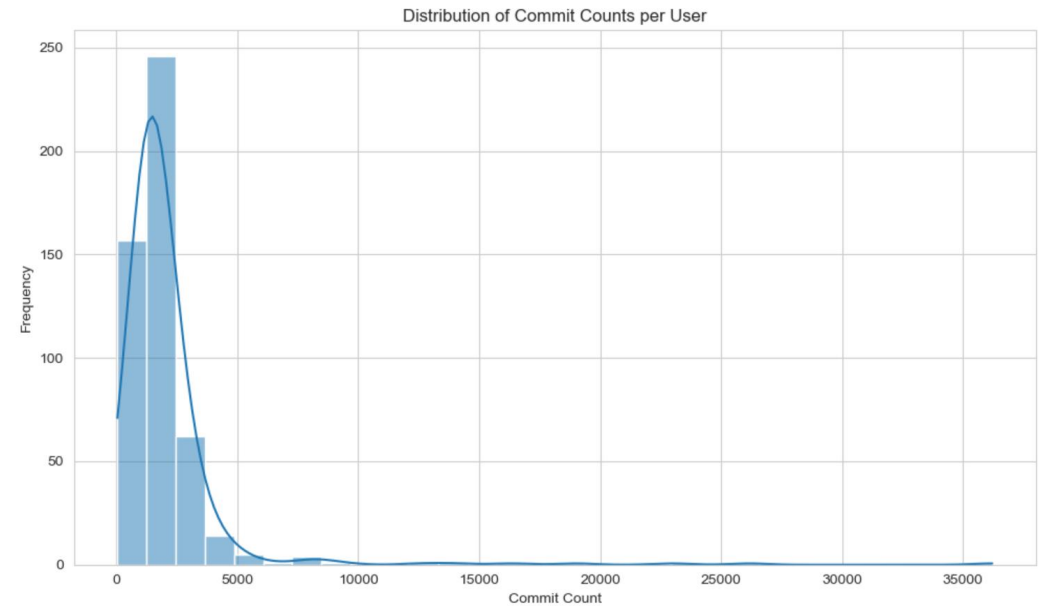

数据显示：在 7-17 点（UTC 时间）出现较高的事件数，这意味着用户的活跃时段集中在这段时间。

2.4 提交频率分析

将 event_action 为 "created" 或 "added" 的事件视为提交行为。

- 提交次数分布呈长尾特征。大部分用户提交次数较少，少数用户异常活跃。
- 我们将提交次数高于 90 分位值的用户定义为高活跃用户，识别出 50 名高活跃用户。

- 将提交次数低于 10 分位值的用户定义为低活跃用户，识别出 49 名低活跃用户。



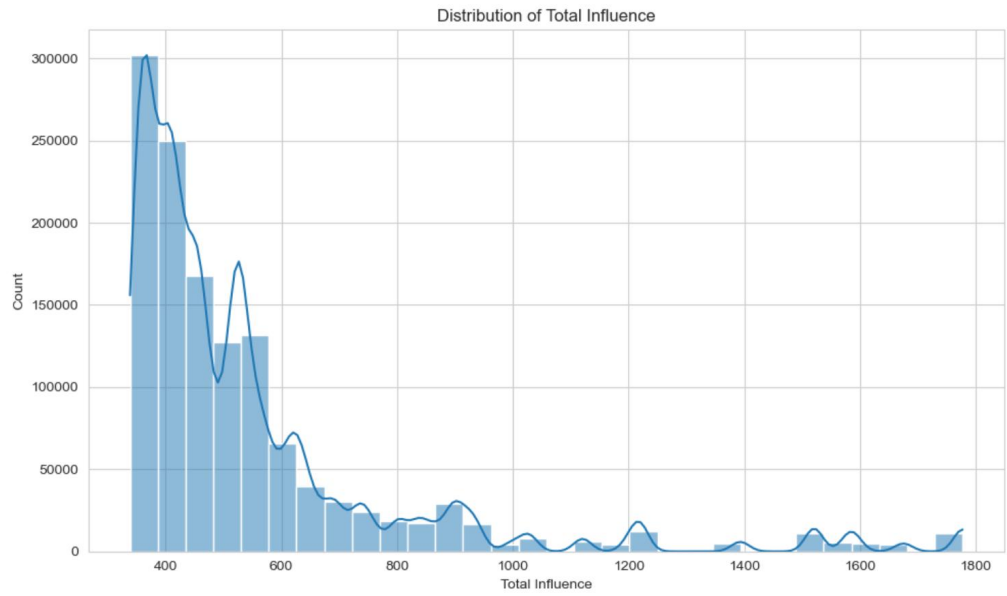

```
Number of High-Active Users: 50
Number of Low-Active Users: 49
```

通过识别高活跃用户，可以进一步关注这些核心贡献者的需求和偏好；对于低活跃用户，也可考虑如何优化项目文档、引导新手参与以提高他们的活跃度。

## 2.5 用户影响力分析

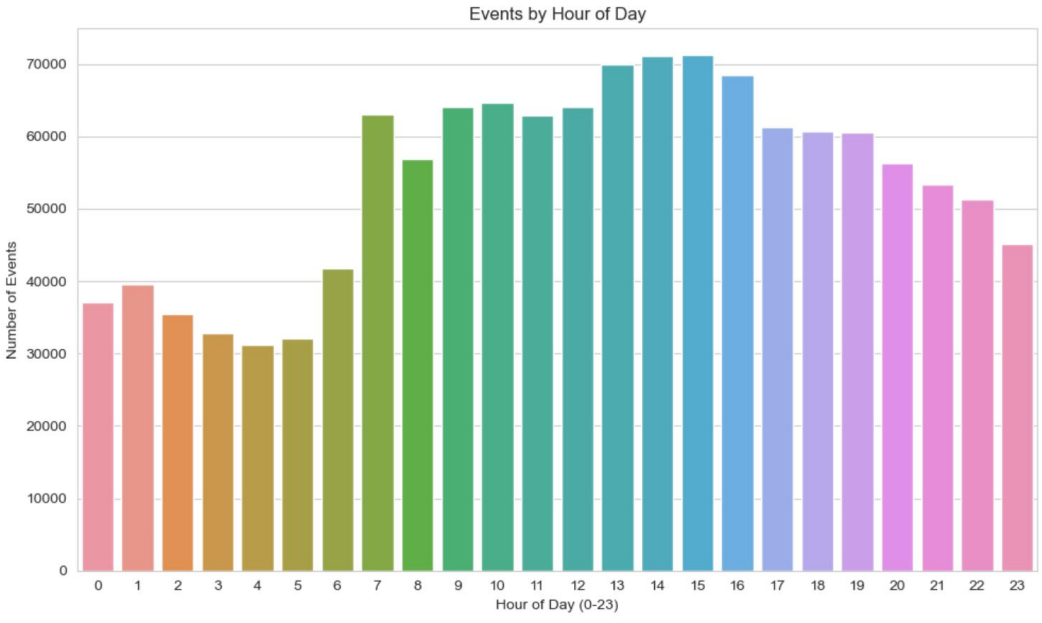
total_influence 字段用于衡量用户影响力，分布统计结果为：

```
Influence Stats:
 count    1.294776e+06
mean     5.440863e+02
std      2.578072e+02
min      3.385323e+02
25%      3.900486e+02
50%      4.552713e+02
75%      5.748544e+02
max      1.776967e+03
```



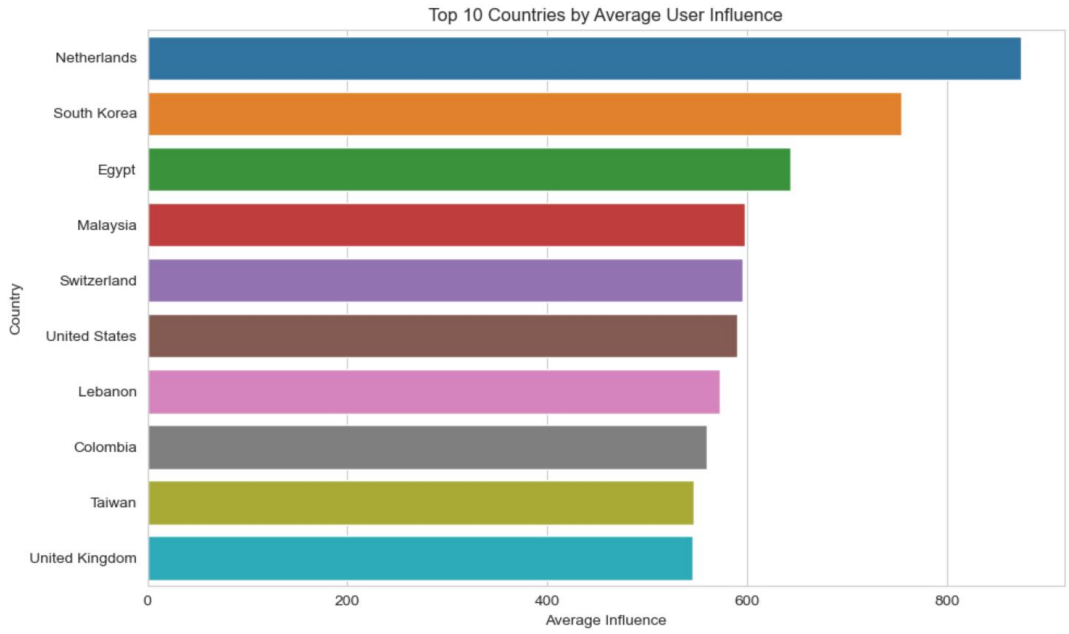

这表明用户影响力数值范围较广，平均值约 544，最大值高达 1776，部分用户明显具有超出平均水平的影响力。

## 2.6 事件随时间分布（分析一天内的活跃度）



## 2.7 用户影响力 vs 国家对比（平均影响力最高的国家）

```
Top 10 Countries by Average Influence:
country
Netherlands       874.232403
South Korea       754.195618
Egypt             644.220337
Malaysia          597.661397
Switzerland       596.237744
United States     590.065086
Lebanon           572.721558
Colombia          560.299500
Taiwan            546.809248
United Kingdom    545.631840
```




分析显示，荷兰（Netherlands）用户平均影响力最高，其次为韩国（South Korea）和埃及（Egypt），说明不同国家的用户在影响力方面也存在差异。