用户行为与人口分布数据洞察报告

一.引言

本报告基于用户行为和人口分布数据，探讨用户分布特点和协作模式。分析主要集中于国家和城市的开发者分布、时区协作特点，以及用户提交行为模式，旨在为技术决策和资源分配提供数据支持。

二.方法与数据描述

数据来源：'users_combined_info_500.csv'
数据字段简介：
user_id: 用户唯一标识。
name: 用户名。
location: 用户所在的城市或地区。
total_influence: 用户的影响力得分。
country: 用户所属国家。
event_type: 事件类型（如 "CreateEvent"）
。
event_action: 事件的具体操作（如 "added"
）。
event_time: 事件发生的时间（包括时区信息
）。
分析工具：python库（pandas，Numpy，
seaborn，matplotlib）
初步分析：
数据量较大（1,294,776行）。
数据字段丰富，适合进行人口统计和协作行为
分析。
country 字段有部分缺失值（约7万行）。

三.分析结果

1. 人口统计分析

(1)国家分布

图一：国家用户分布

分析描述：

数据展示了用户按国家的分布情况，列出了用户数量排名前 20 的国家。

美国用户数量遥遥领先，超过 200,000 人，占据主导地位。

德国和中国分别排名第二和第三，其次是英国和法国。

加拿大、荷兰、捷克、日本、瑞士和澳大利亚等国家也有较多用户。

洞察总结：

用户主要分布在欧美国家，尤其是美国和德国，这可能与业务的国际化程度或服务定位有关
。

亚洲的中国和日本也进入前十，说明在这些区域的渗透率较高。

(2)城市分布

图二：城市用户分布

分析描述：

数据展示了用户按城市的分布情况，列出了用户数量排名前 20 的城市。

德国和捷克的城市名次较高（例如 Berlin, Prague）。

美国的 Palo Alto 和纽约也位列前茅。

此外，法国、瑞士和日本的部分城市也进入榜单。

洞察总结：

用户分布集中在某些技术或经济发达的城市，如硅谷的 Palo Alto、纽约和柏林等。

显示业务可能更受全球一线城市和技术中心欢迎。

(3)时区分布

图三：时区分布

分析描述：

数据统计了用户的时区分布。

图中清晰展示了 UTC+08:00 时区占据绝大多数用户比例。

洞察总结：

UTC+08:00 的用户占比高，这可能是由于中国及东南亚地区的活跃用户较多。

其他时区的用户分布相对较为均衡，可能反映了全球业务扩展的情况。

2.用户行为分析

(1)用户提交频率分析

分析描述：

数据统计了每位用户的提交频率，并进行了高活跃用户与低活跃用户的分类。

提交超过 5000 次的高活跃用户仅有 25 名，而提交次数在 1000 次以下的用户有 39 名。

洞察总结：

少数高活跃用户贡献了主要数据，这表明社区或业务中的核心用户对活动的推动至关重要。

需要制定针对高活跃用户的维护策略，同时鼓励低活跃用户参与更多互动。

(2)事件类型分布

分析描述：

数据展示了不同事件类型的发生频率。

"PushEvent"（推送事件）是最常见的事件类型，占比最高，其次是 "PullRequestEvent"（拉取请求事件）。

"IssueCommentEvent" 和 "PullRequestReviewEvent" 的数量也较多。

洞察总结：

推送和拉取请求是用户活动的主要类型，这可能与系统的核心功能密切相关。

评论和代码审查的事件分布显示了社区内的协作活动。

(3)操作行为分布

分析描述：

数据展示了事件操作行为的分布情况。

"added" 和 "created" 操作频率最高，分别达到 617,218 次和 411,961 次。

"closed" 和 "opened" 操作也有较多频率。

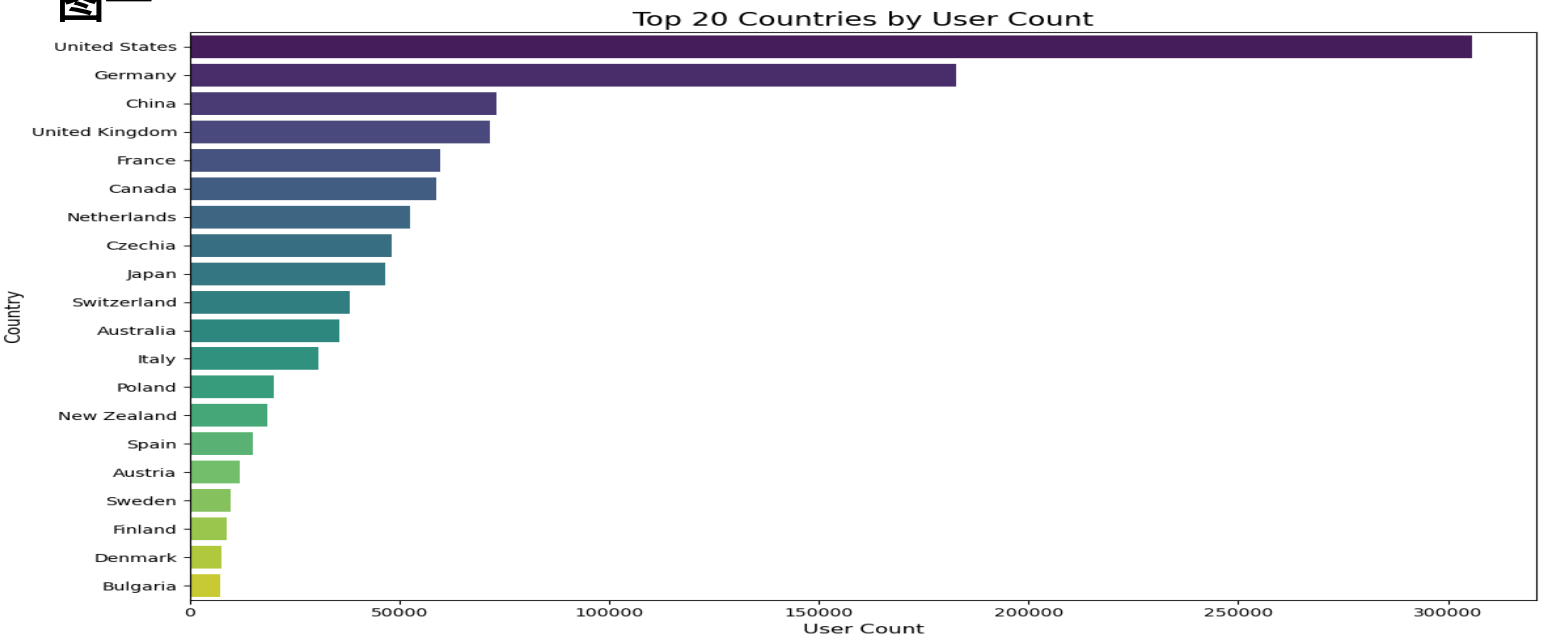
洞察总结：

用户的主要操作集中在新增内容（"added" 和 "created"），这可能表明用户活跃度较高，且系统功能倾向于内容创建。

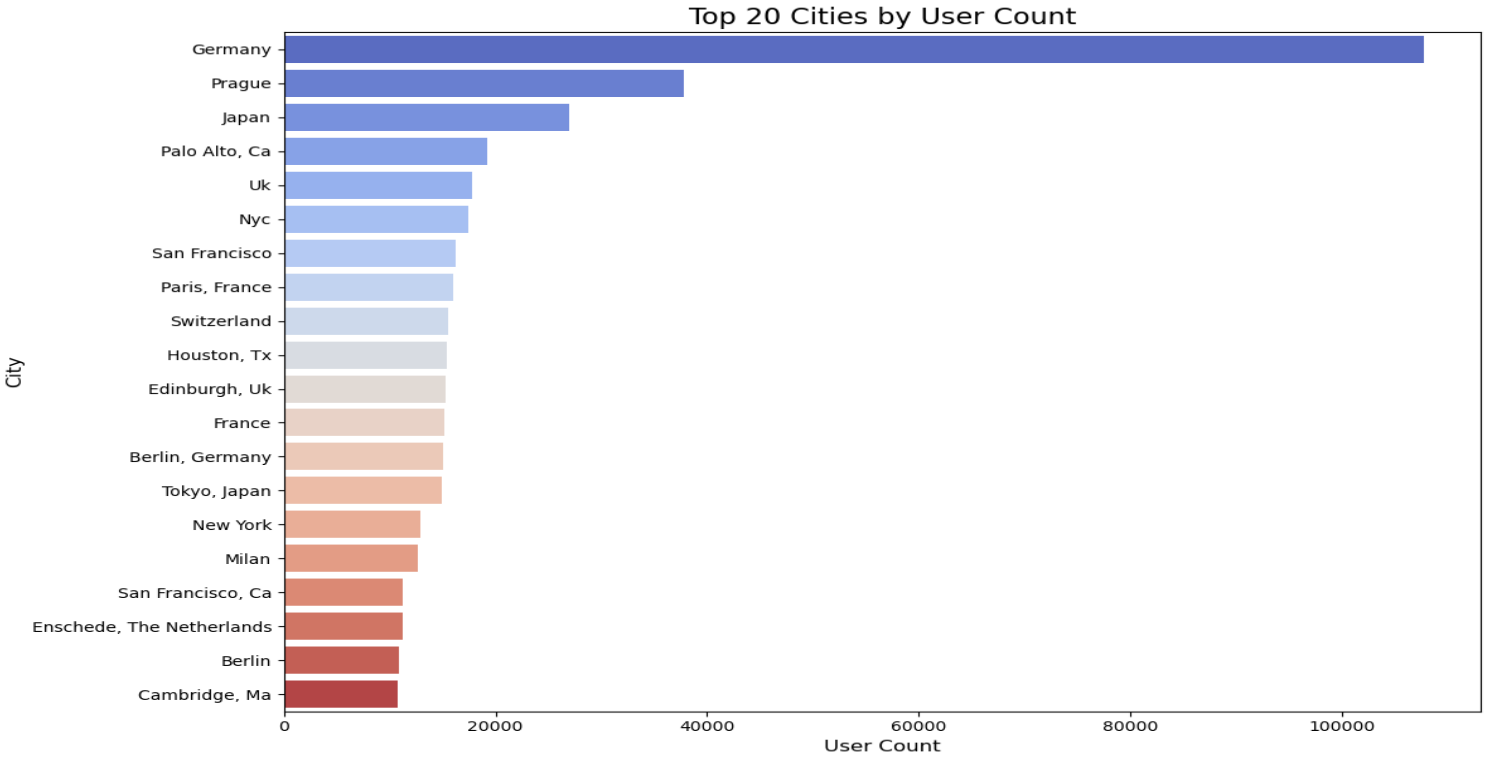
"closed" 操作的高频率可能与任务的关闭与管理有关。

四.可视化展示

图一


Top 20 Countries by User Count
United States
Germany
China
United Kingdom
France
Canada
Netherlands
Czechia
Japan
Switzerland
Australia
Italy
Poland
New Zealand
Spain
Austria
Sweden
Finland
Denmark
Bulgaria
Country
0
50000
100000
150000
200000
250000
300000
User Count

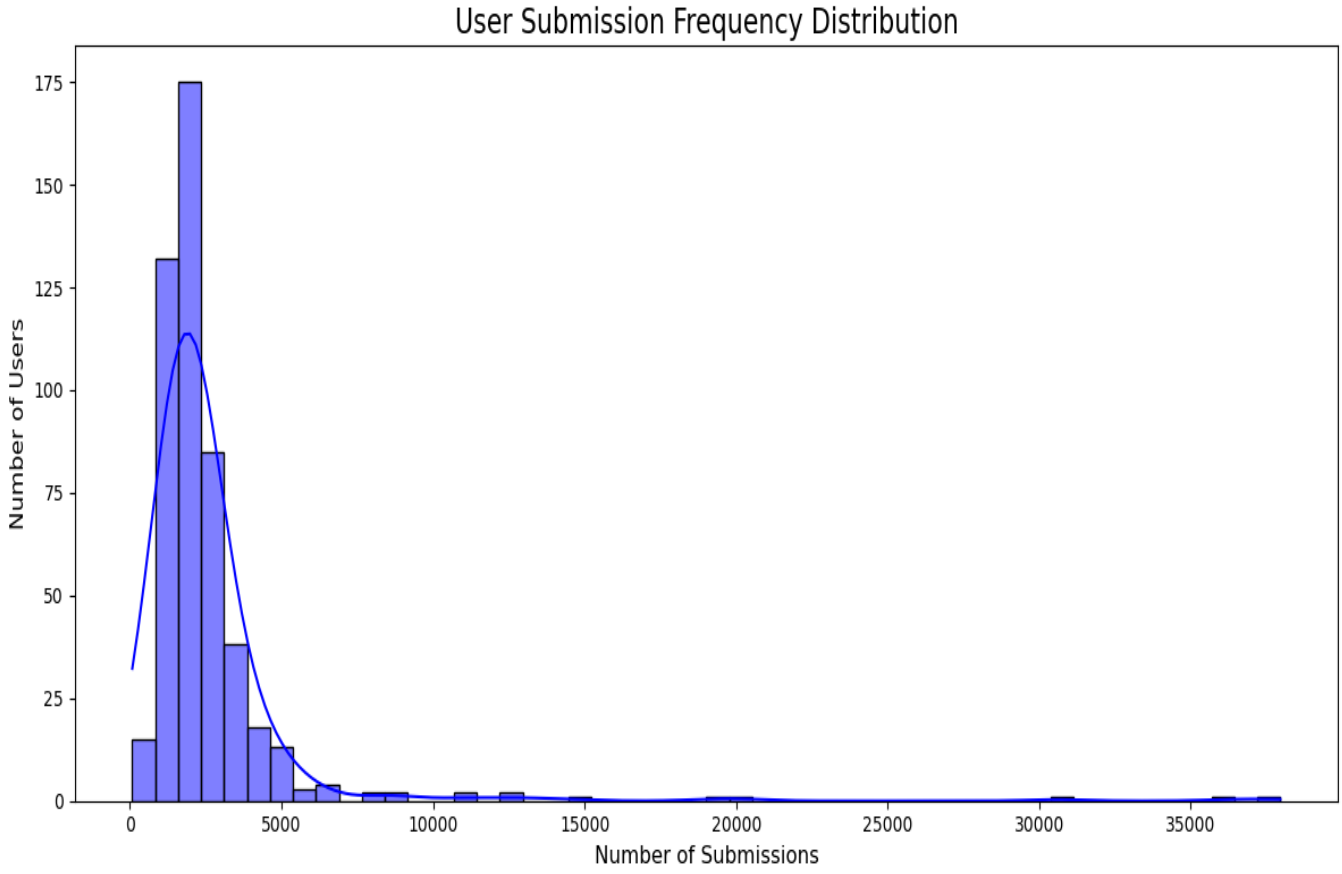
图二


Top 20 Cities by User Count
Germany
Prague
Japan
Palo Alto, Ca
Uk
Nyc
San Francisco
Paris, France
Switzerland
Houston, Tx
Edinburgh, Uk
France
Berlin, Germany
Tokyo, Japan
New York
Milan
San Francisco, Ca
Enschede, The Netherlands
Berlin
Cambridge, Ma
City
0
20000
40000
60000
80000
100000
User Count

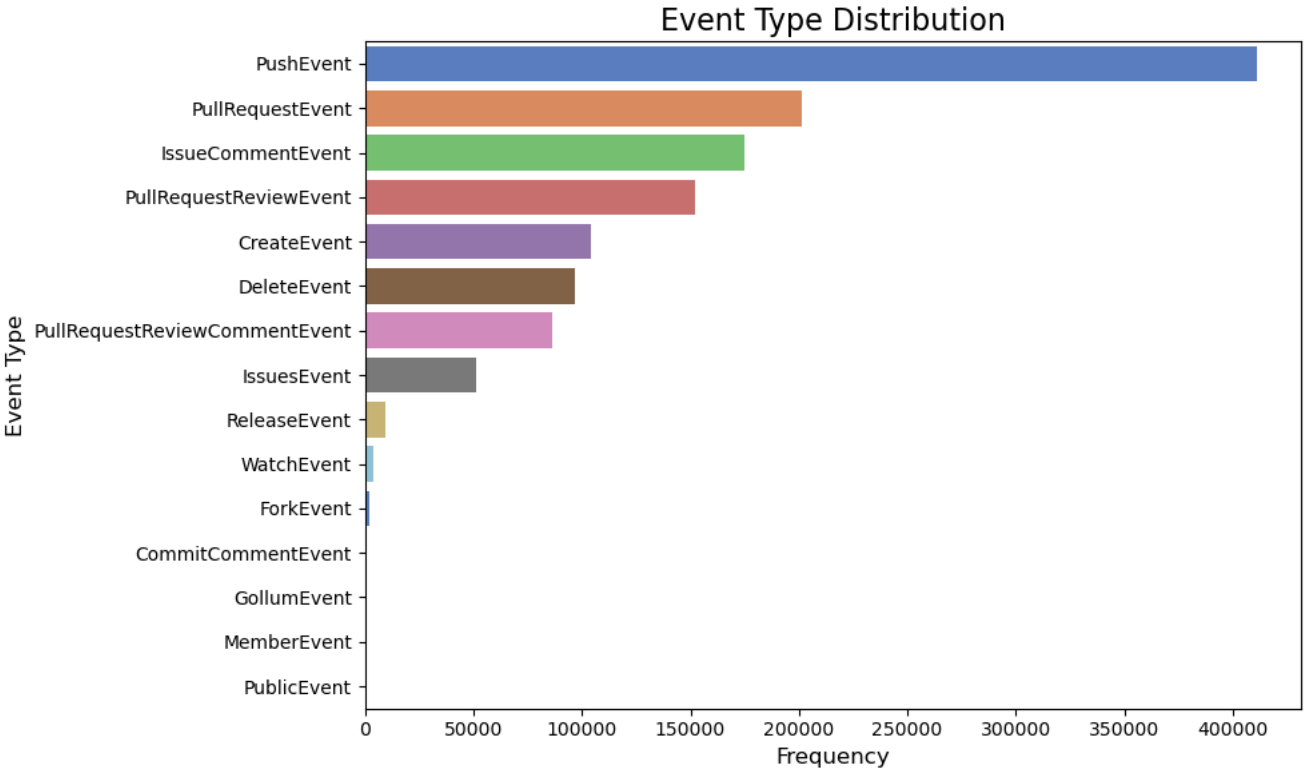
图三


User Distribution by Timezone (Improved)
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
User Count
+0800
Timezone

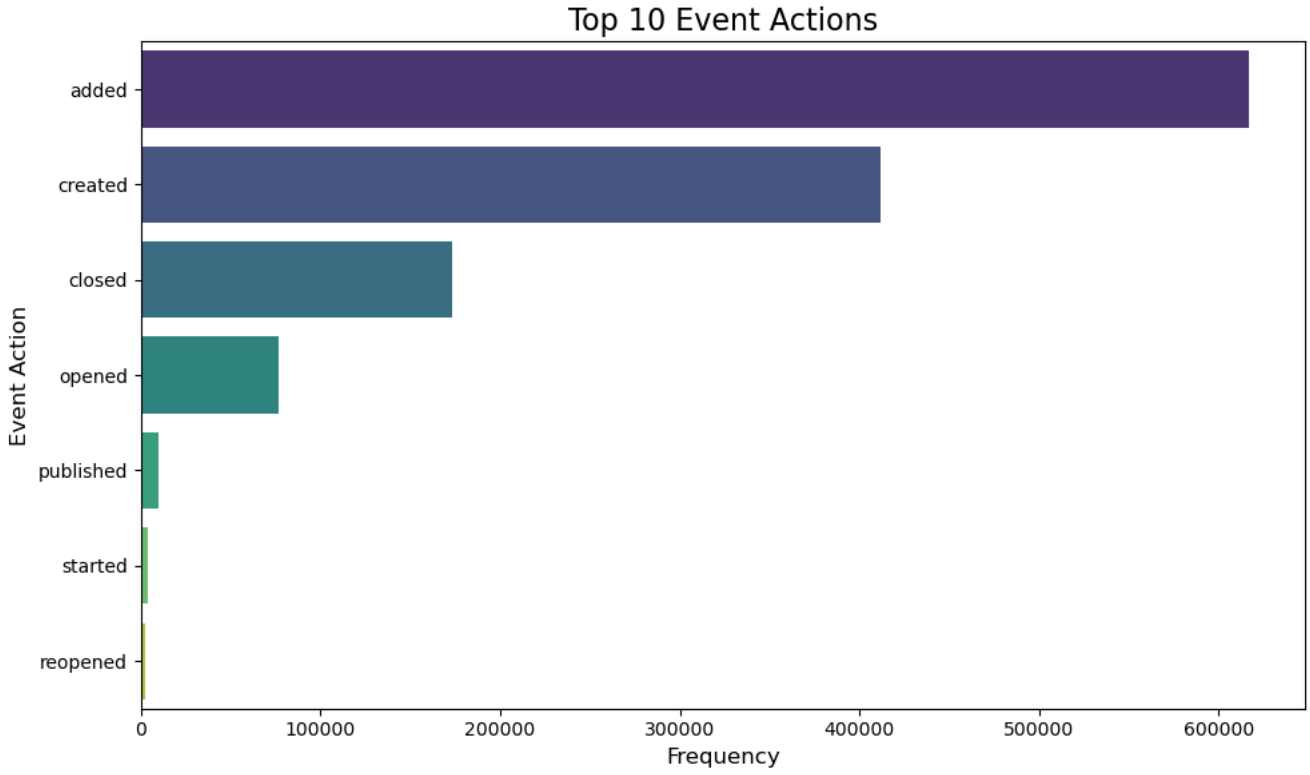
图四


User Submission Frequency Distribution
175
150
125
100
75
50
25
0
Number of Users
0
5000
10000
15000
20000
25000
30000
35000
Number of Submissions

图五


Event Type Distribution
PushEvent
PullRequestEvent
IssueCommentEvent
PullRequestReviewEvent
CreateEvent
DeleteEvent
PullRequestReviewCommentEvent
IssuesEvent
ReleaseEvent
WatchEvent
ForkEvent
CommitCommentEvent
GollumEvent
MemberEvent
PublicEvent
Event Type
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequency

图六


Top 10 Event Actions
added
created
closed
opened
published
started
reopened
Event Action
0
100000
200000
300000
400000
500000
600000
Frequency

五.结论与建议

1.核心发现

1. 用户分布特征

国家分布：美国用户数量占据绝对优势，其次是德国和中国。这表明您的平台在这些国家的用户基础较强，但其他国家的渗透率相对较低。

城市分布：部分国家（如德国）在数据中被重复统计为城市（可能是数据清理问题）。高用户量城市主要集中在技术与经济发达地区，如旧金山、东京、巴黎等。

时区分布：大部分用户来自 UTC+08:00 时区（可能包括中国和部分东南亚国家），反映了平台在该区域的较高活跃度。

2. 用户行为特征

提交频率：活跃用户数量有限，仅有 25 位用户提交次数超过 5000；而多数用户（39 位）提交次数低于 1000。这可能反映出平台存在较明显的头部用户效应。

事件类型：最常见的事件类型是 PushEvent 和 PullRequestEvent，分别代表代码提交和合并请求，表明平台核心功能集中在代码协作。

操作行为：added 和 created 是最常见的操作，表明用户主要集中在创建内容和新增资源的行为上。

2.改进建议

1. 扩展国际市场渗透

目标市场：通过市场分析定位潜力国家（如法国、加拿大等），定制推广策略，吸引更多用户加入。

多语言支持：提升平台的语言多样性（尤其是支持法语、德语、日语等），降低非英语国家用户的使用门槛。

2. 优化用户活跃度

中低活跃用户策略：设计激励机制（如徽章、排行榜）鼓励低活跃用户参与更多互动。

活跃用户留存：为高活跃用户提供专属功能或权限，增强他们的归属感和长期贡献意愿。

3. 完善数据清洗和分析

位置字段清理：解决位置字段的重复问题（如将 “ Germany ” 统一归类为国家）。

时区分析：将时区映射到具体国家或区域，以更准确分析不同地区的用户行为。

事件日志精细化：进一步细分 PushEvent 和 PullRequestEvent 中的具体子操作，挖掘用户行为的深层次模式。

4. 增强功能模块

关注头部事件：优化 PushEvent 和 PullRequestEvent 的交互流程，提高用户体验。

新增分析维度：结合时间戳分析行为变化趋势，观察特定时段（如白天/夜间）的活跃度，优化服务器负载管理。

5. 增强平台生态系统

提供项目管理功能，如任务分配、进度跟踪，帮助团队提升协作效率。

开展社区活动（如黑客松、技术论坛），加强用户之间的联系和互动。