实验报告 —— 数据洞察分析

罗文琦——10235501417

一．引言

本报告通过对大规模开发者数据集的深入分析，揭示开发者群体在地域分布、协作行为等方面的特征，为了解技术生态提供依据。通过数据分析技术与可视化手段，挖掘数据背后的价值信息。

二．数据获取与合并

本次实验使用了多个数据文件，经过合并处理后生成了一个完整的数据集，文件名称为 `merged_data.csv` 。

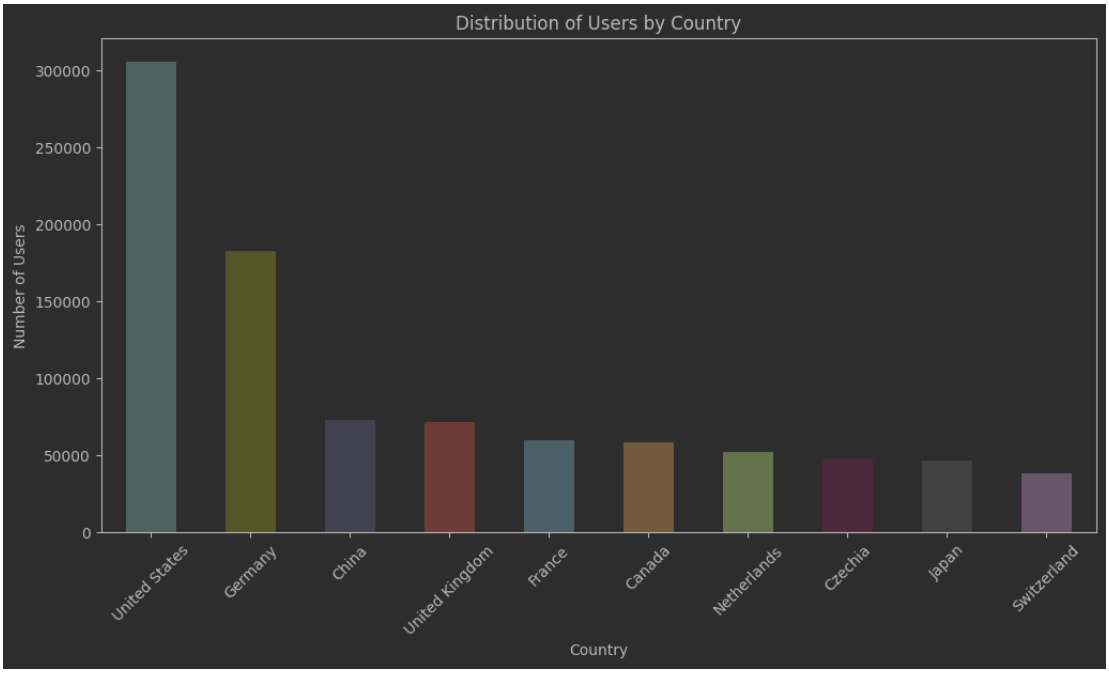
三．人口统计分析

（一）国家和地区分布

统计了用户所在国家和地区的分布情况，结果显示开发者主要集中在以下国家（排名前十）：

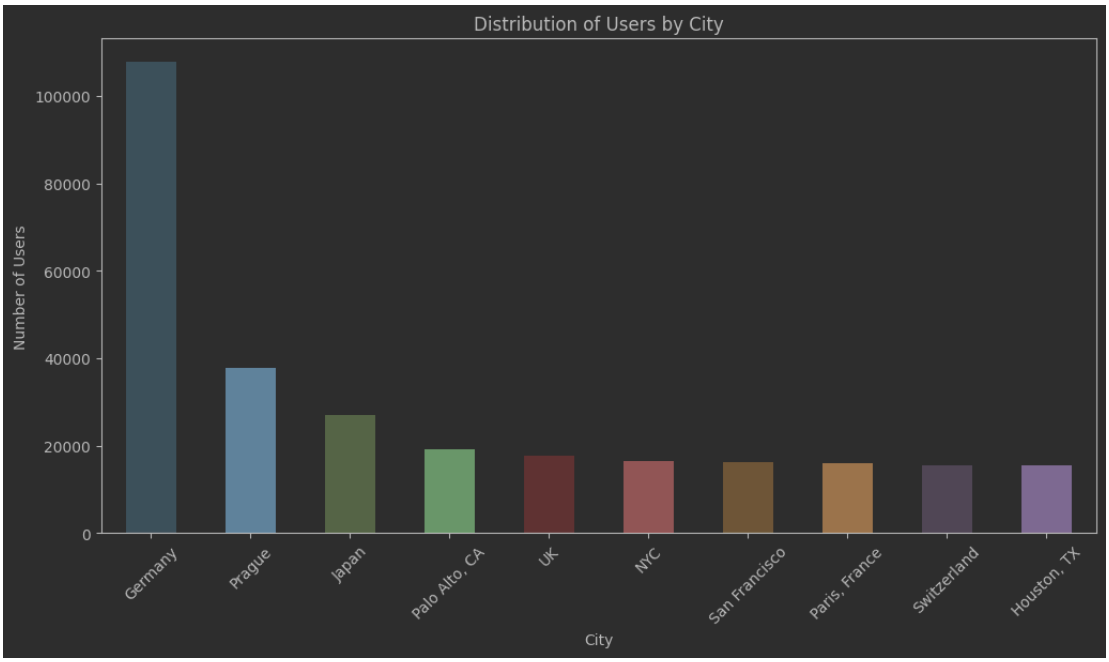
排名	国家	开发者数量
1	United States	305788
2	Germany	182659
3	China	73011
4	United Kingdom	71606
5	France	59570
6	Canada	58600
7	Netherlands	52367
8	Czechia	48122
9	Japan	46553
10	Switzerland	38093

柱状图展示了各国开发者数量分布的可视化情况。



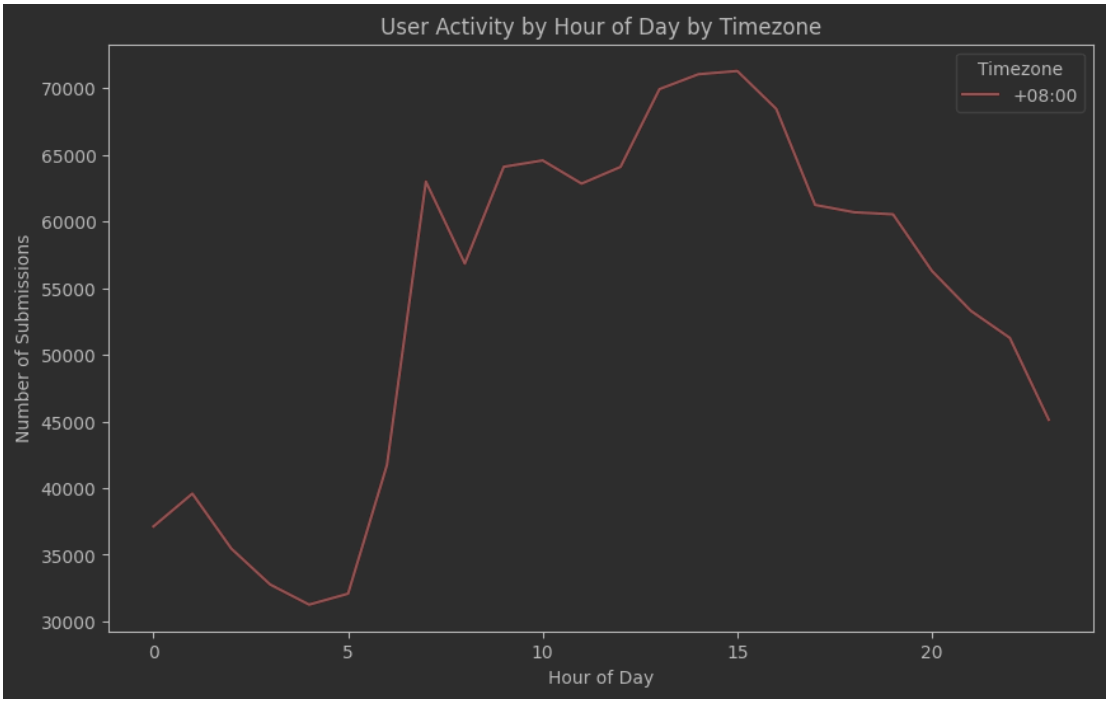
（二）城市级别分布

进一步分析城市层面开发者密度，揭示主要技术热点区域。以下为排名前十的城市及其开发者数量。



（三）时区分布

用户时区分布反映了不同地区协作的时间模式。协作效率的优化需考虑时区间的重叠。以下为折线图展示的时区活跃热度。

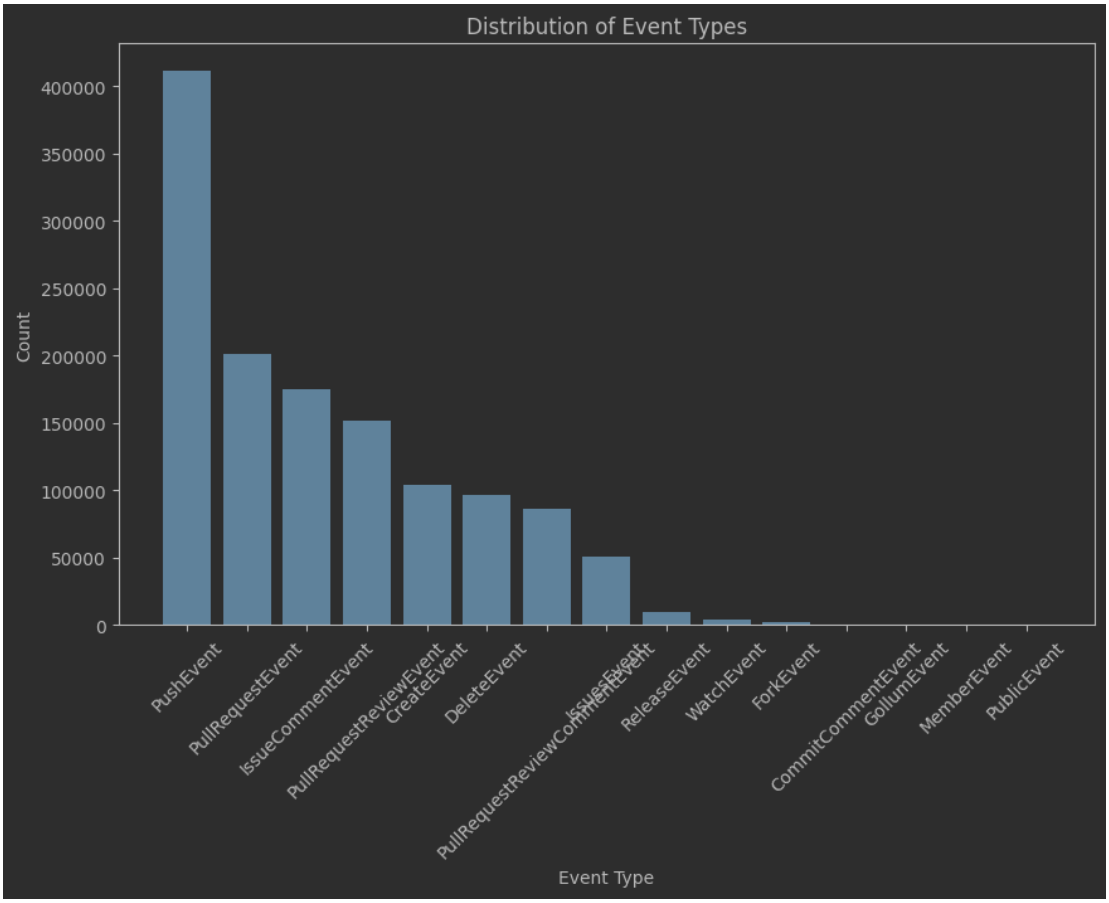


四．协作行为分析

（一）提交频率

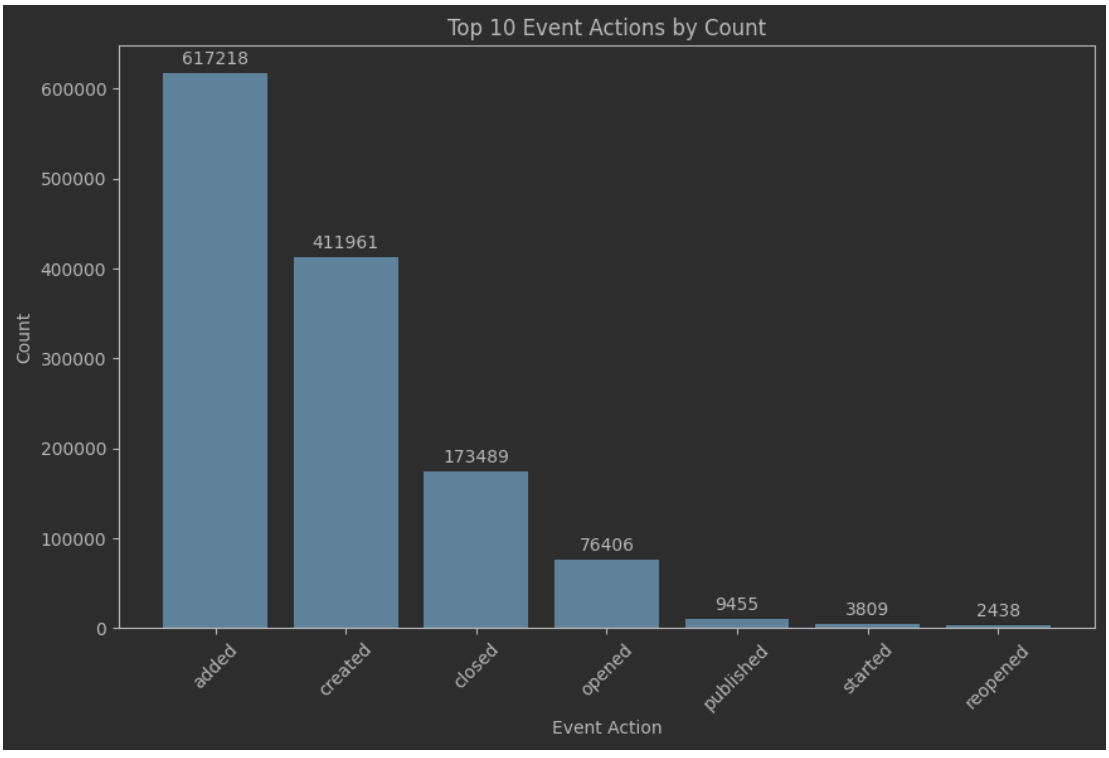
对每个用户的提交频率进行统计，结果表明高活跃用户（提交次数 > 1000）占比 92.15%，低活跃用户（提交次数 < 100）仅占比 0.20%。针对不同群体，提出如下策略：

- 高活跃用户：提供更多技术挑战与激励；
- 低活跃用户：加强培训支持，引导参与。



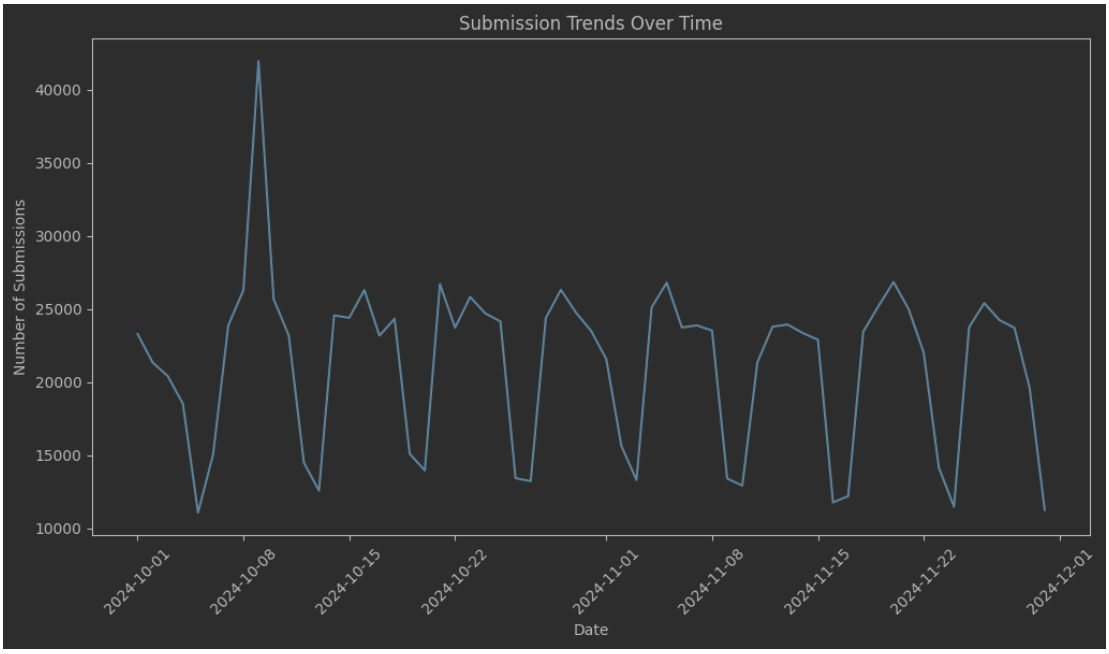
（二）事件类型分析

分析了不同类型事件的发生频率，PushEvent 出现频次最高，反映了代码提交在开发中的核心作用。其他常见事件类型如 PullRequestEvent 和 IssueCommentEvent 也有较高的比例。以下为事件类型分布的可视化。



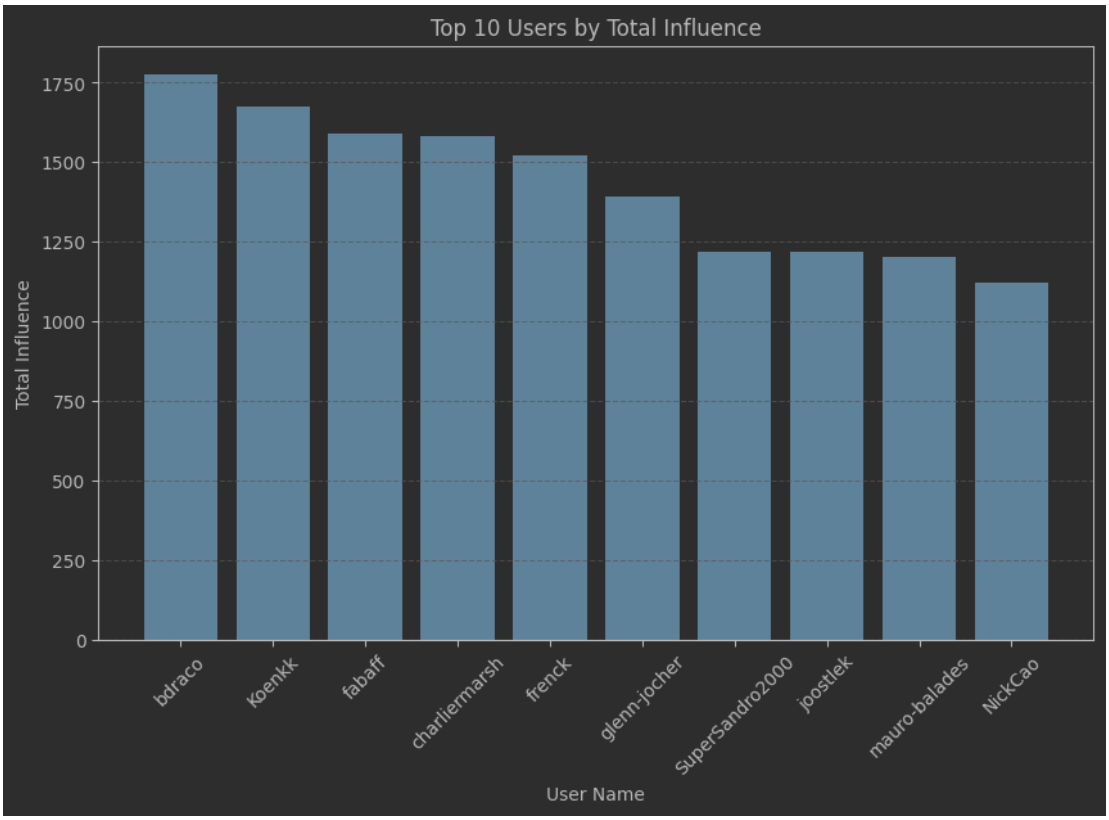
五．Event Action 分析

统计了用户操作行为的排名情况，最常见的操作为 `added`，出现频率达 617218 次，表明开发者重点聚焦于核心代码的创建与提交工作。



六．个人影响力分析

依据 `total_influence` 指标，筛选出前十位开发者。这些开发者在技术攻坚、社区贡献等方面表现卓越，是推动项目发展的核心力量。例如，用户 bdraco 的总影响力值达 1776.97，是团队关键领导者。



In [ ]: