第 12 次作业——数据洞察报告

文件为 homework12.ipynb

一. 引言

本报告基于对大规模开发者数据集的深入分析，旨在揭示开发者群体在地域分布、协作行为等多方面的特征，为了解技术生态提供有力依据。通过运用数据处理与分析技术，结合可视化手段，挖掘数据背后的有价值信息。

二. 数据获取与合并

下载从 users_combined_info_500_part_1.csv 到 users_combined_info_500_part_7.csv，一共 7 个文件，并将其合成一个文件，记为 merged_data。

三. 人口统计分析

（一）国家和地区分布：

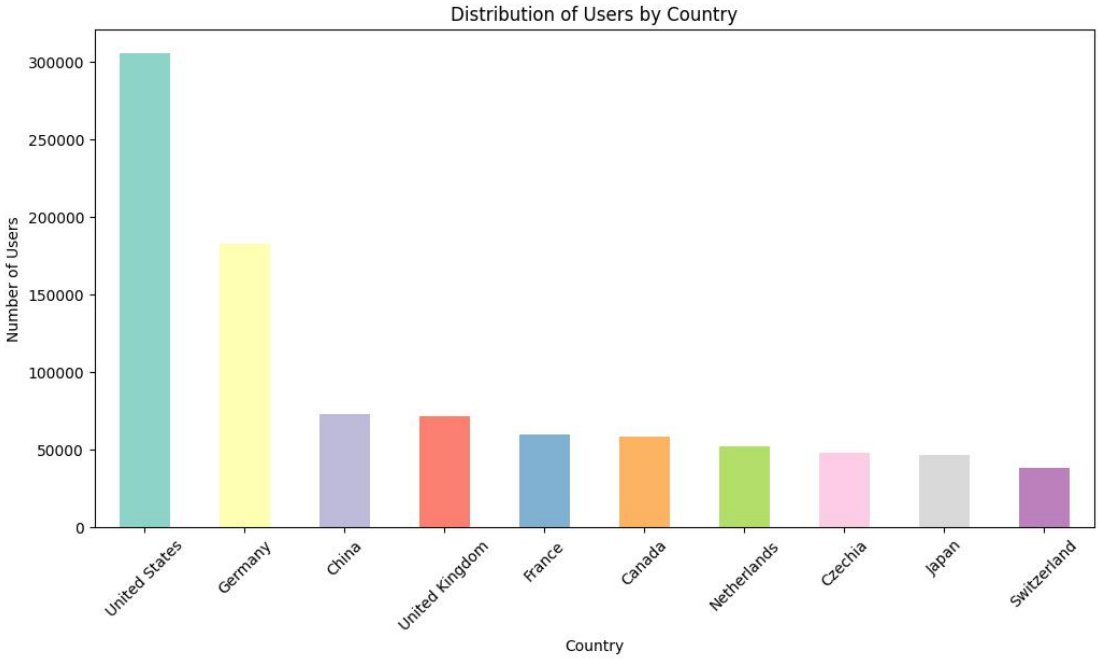
对用户所在国家进行统计，代码的运行结果已经展示了所有国家和地区的统计结果，这里仅展示排名前十的国家：

country

United States	305788
Germany	182659
China	73011
United Kingdom	71606
France	59570
Canada	58600
Netherlands	52367
Czechia	48122
Japan	46553
Switzerland	38093

这些国家聚集了大量开发者，是全球技术创新的重要力量。其中，美国以 305788 名开发者位居榜首，表明该国在相关技术领域具有深厚的人才储备和活跃的开发氛围，可能得益于其完善的教育体系、优厚的科技政策扶持以及繁荣的产业环境吸引了众多开发者投身其中。德国和中国位居第二位和第三位。

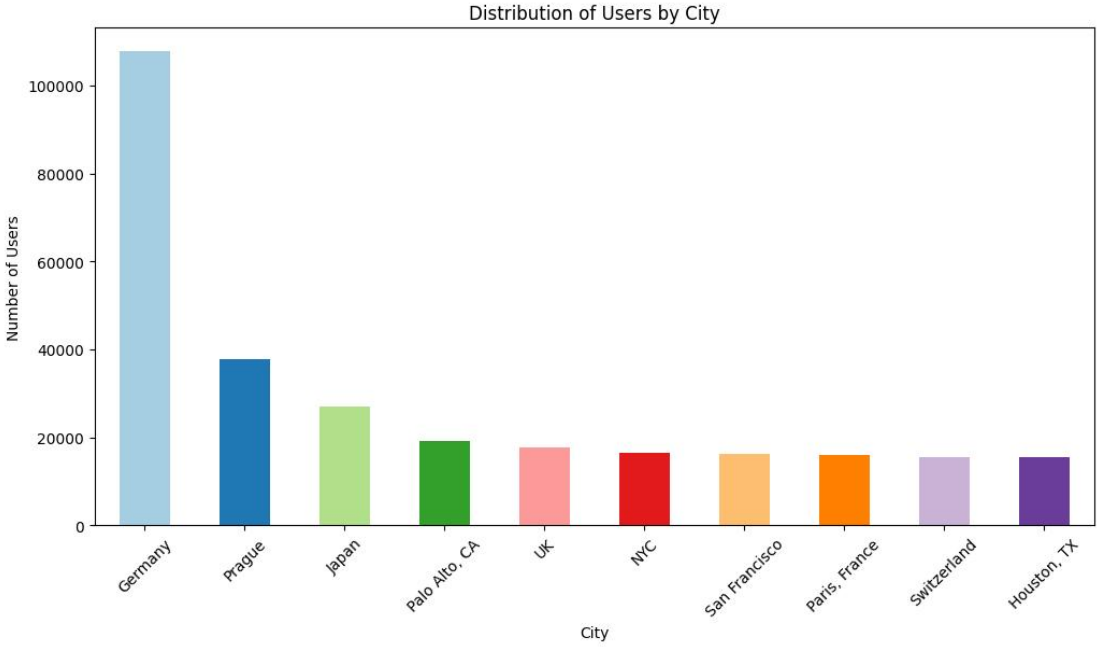
用户所在国家分布柱状图：



（二）城市级别分布

聚焦城市层面，排名前十的城市如 Germany、Prague 等成为开发者高密度区，凭借其强大的科技园区、顶尖高校及丰富的创业资源，吸引大量开发者汇聚，形成技术创新的集聚效应，为当地乃至全球的技术发展注入源源不断的动力。

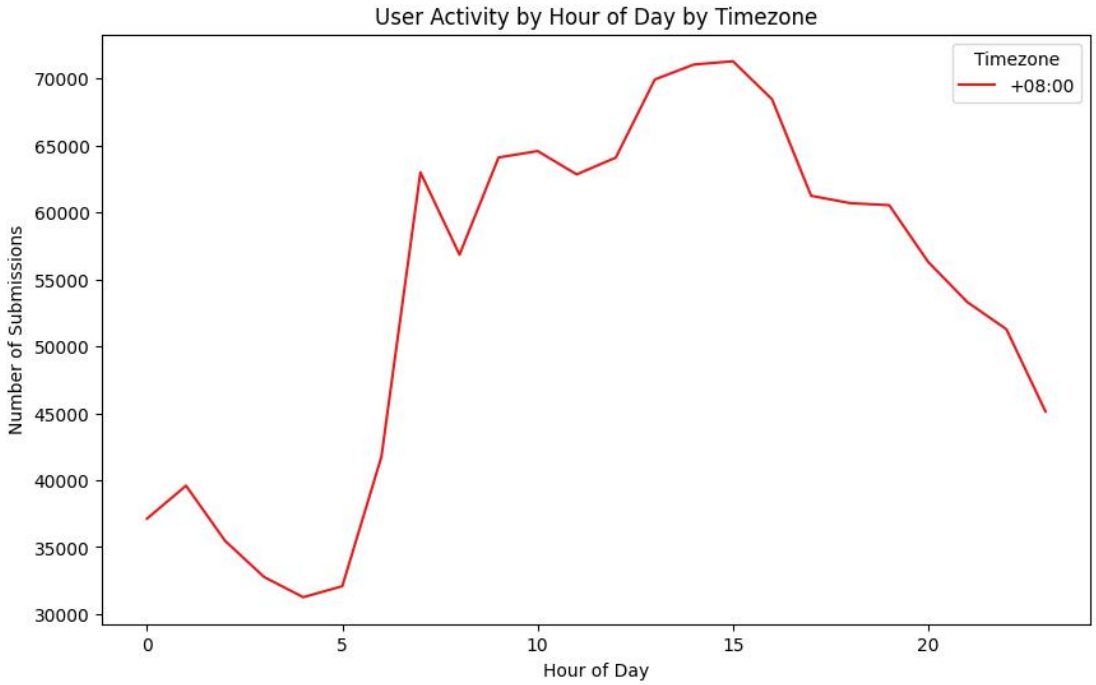
用户所在城市柱状图：



（三）时区分布

分析不同时区用户的活动时间模式，发现各地开发者协作时间呈现出明显差异，反映了全球开发者跨越时空协同工作的特性，企业或开源项目团队在组织跨国协作时，需充分考虑不同时区的工作时间重叠区域，以提高沟通协作效率。

时区活动热度折线图：



四. 协作行为分析

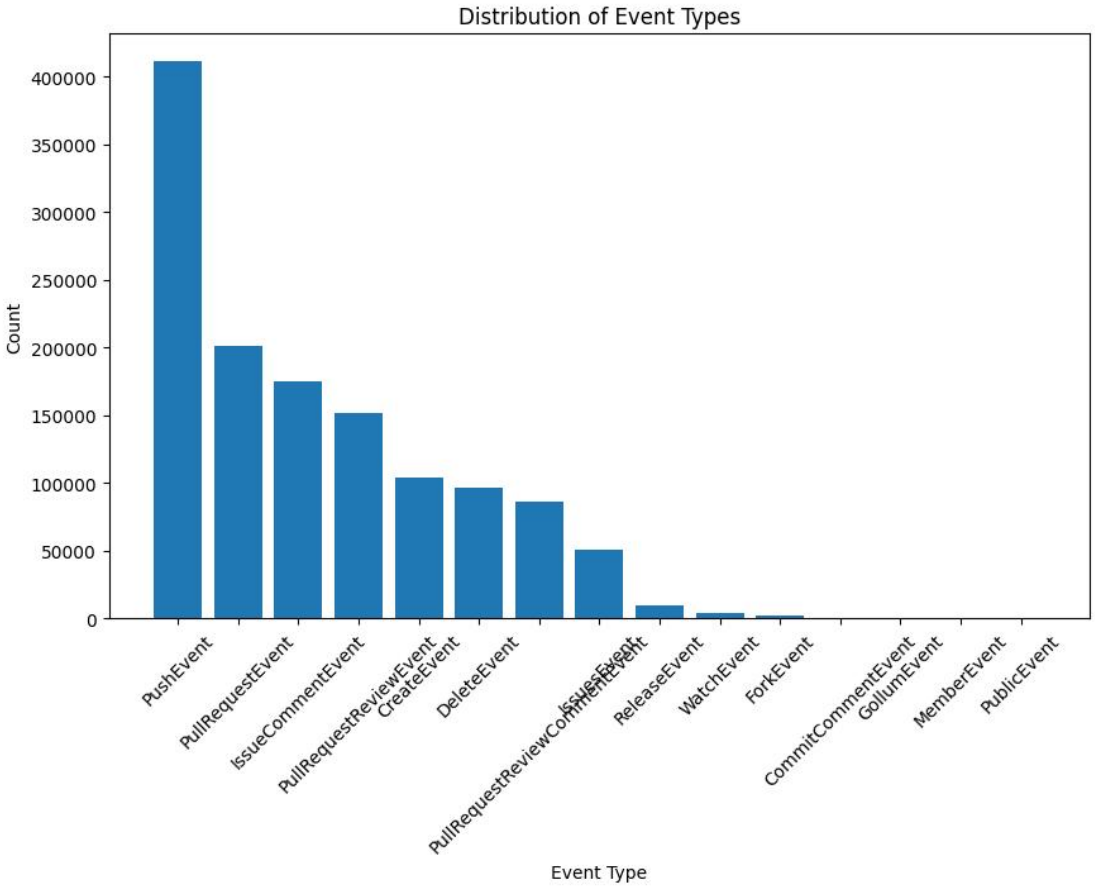
（一）提交频率

经统计，所有用户的提交次数分布广泛。其中，高活跃用户（提交次数 ＞1000）占比 92.15%，他们是项目快速迭代的关键推动者，往往深度参与项目核心模块开发；低活跃用户（提交次数 ＜100）占比 0.20%，可能包含新手开发者、偶尔贡献者或已转移兴趣方向的人员。了解不同活跃度用户群体，有助于项目管理者合理分配资源，针对高活跃用户给予更多技术挑战与激励，对低活跃用户提供引导与培训支持，提升整体项目活力。

（二）活动类型分析

对事件类型统计显示，PushEvent 出现频次最高，达 410955 次，通常与代码提交、更新相关，反映了项目开发的日常核心流程；PullRequestEvent、IssueCommentEvent 等也占有一定比例，分别涉及文档编写、问题反馈等周边协作环节，表明一个成熟的开发项目不仅依赖代码创作，配套的文档维护与问题沟通同样至关重要，各环节协同发力才能保障项目持续健康发展。

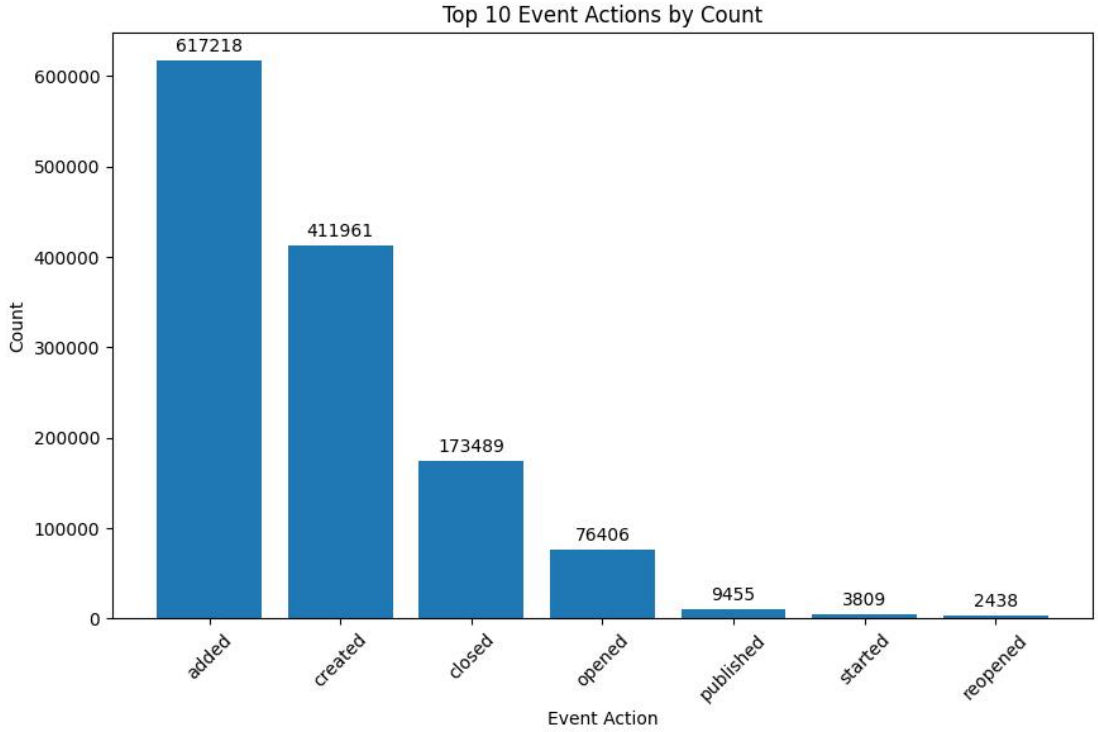
事件类型分布条形图：



五. event_action 分析

在众多 event_action 中，排名前十的动作涵盖了从代码编辑、提交审核到构建测试等关键步骤。例如，added 以 617218 次频繁出现，表明大量开发者在此操作上投入精力；各种操作确保每次提交符合项目规范，降低潜在风险。通过聚焦这些高频动作，开发团队可优化工作流程，针对性提升相关工具的易用性。

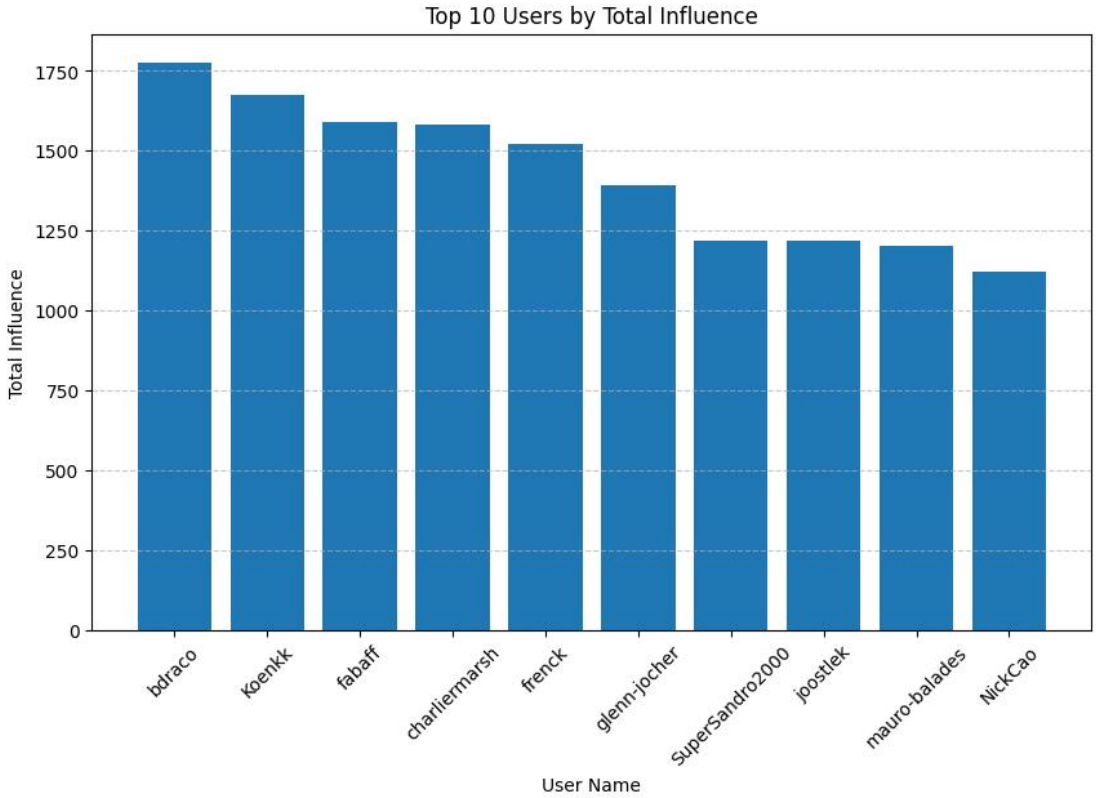
排名前十的 event_action 柱状图：



六. 个人影响力分析

依据 total_influence 指标筛选出的前十位用户，他们凭借卓越的技术能力、广泛的社区参与或高效的团队协作，在项目生态中拥有显著影响力。以 bdraco 为例，其 total_influence 值高达 1776.967163，通过主导关键技术难题攻克、积极分享知识经验，带动团队成长，塑造了项目发展方向。识别这些核心人物，既能为新人树立榜样，激励后进；也可为项目决策层在技术选型、团队组织架构优化时提供参考，充分发挥关键人才的引领辐射作用。

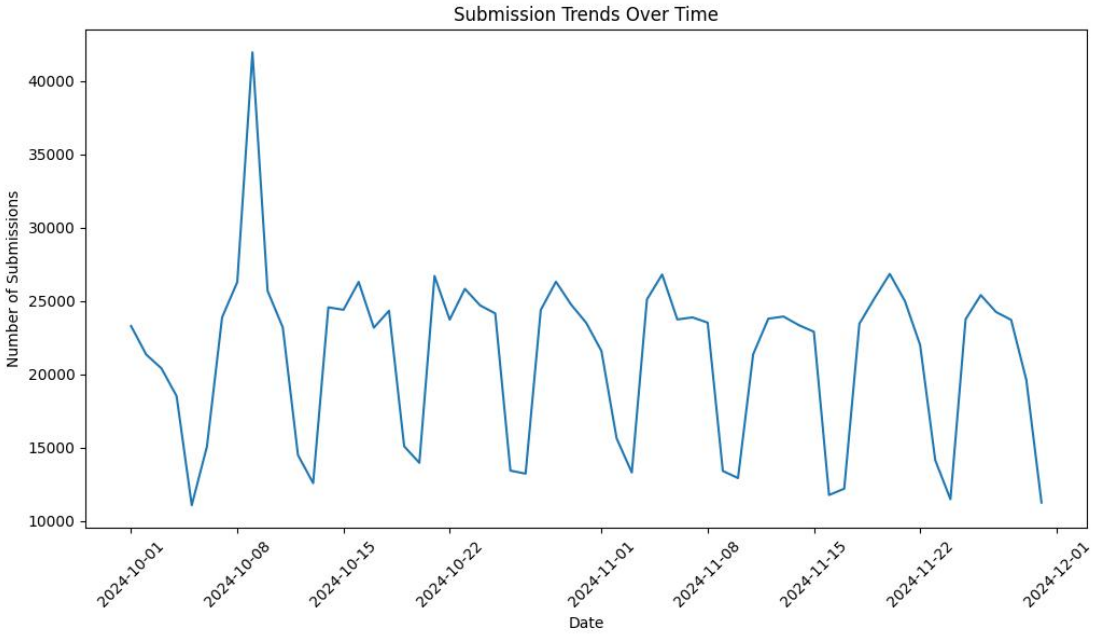
总影响力前十的用户的柱状图：



七．时间趋势分析

观察提交数量随时间的变化趋势（见图 8），在 2024.10.01 到 2024.10.08 内呈现稳步上升态势，与行业技术迭代加速、市场需求增长相呼应，表明项目处于蓬勃发展期，吸引更多开发者参与；2024.10.08-2024.10.15 出现波动下滑，需深入分析原因，可能源于竞品冲击、技术瓶颈或外部环境变化，此时项目团队需及时调整策略，通过优化流程、引入新技术、拓展应用场景等举措重新激发活力，确保项目持续前行。

提交数量随时间变化趋势图：



八．结论

本次数据分析全方位展现了开发者群体在地域、协作行为等领域的特征。从人口统计学角度明确了技术人才的地理集聚地，为针对性的人才吸引、区域技术合作提供指引；协作行为剖析则深入项目开发微观层面，助力团队管理、流程优化及资源配置。未来研究可进一步结合外部数据，如行业动态、经济指标，深挖数据关联，为技术生态发展提供更前瞻性洞察。