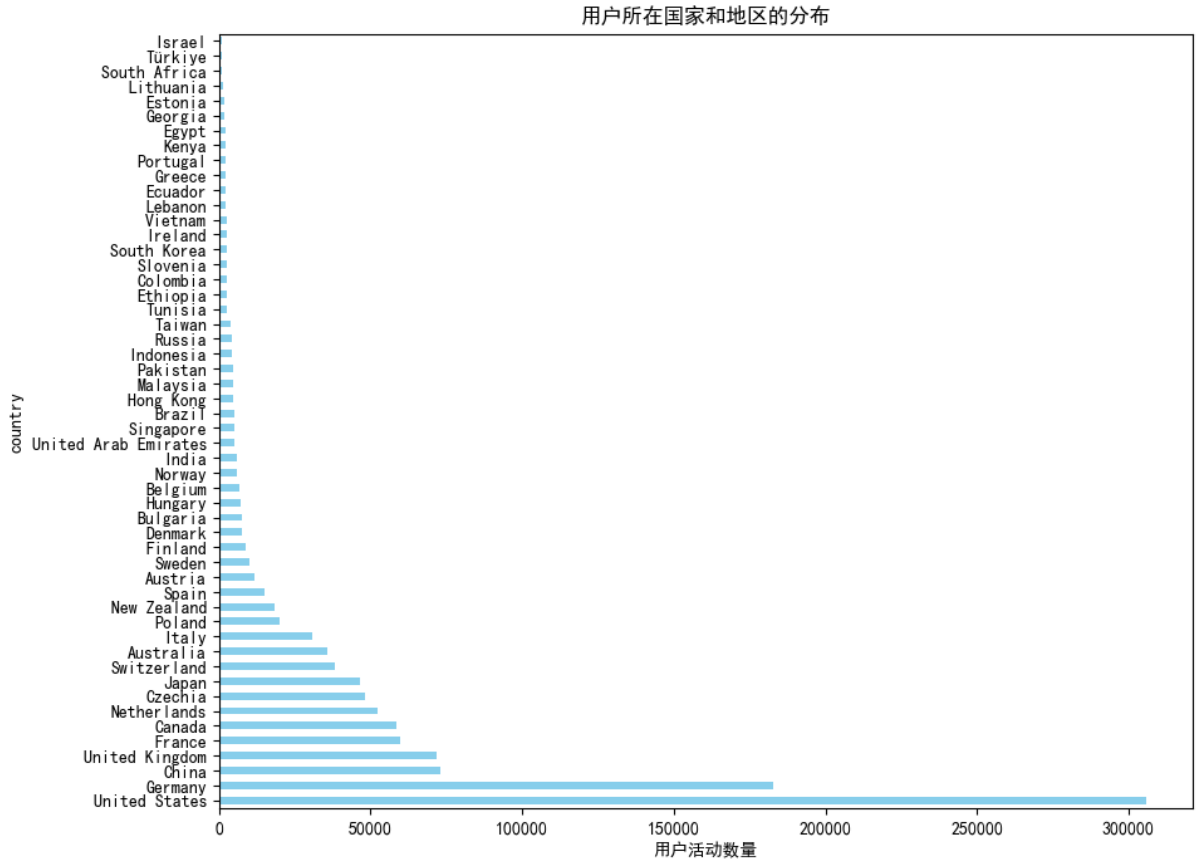
Homework12报告

10235501469邵乐怡

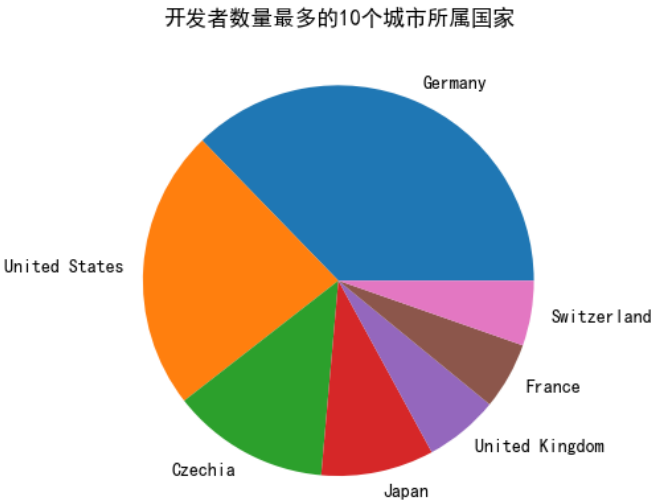
人口统计分析

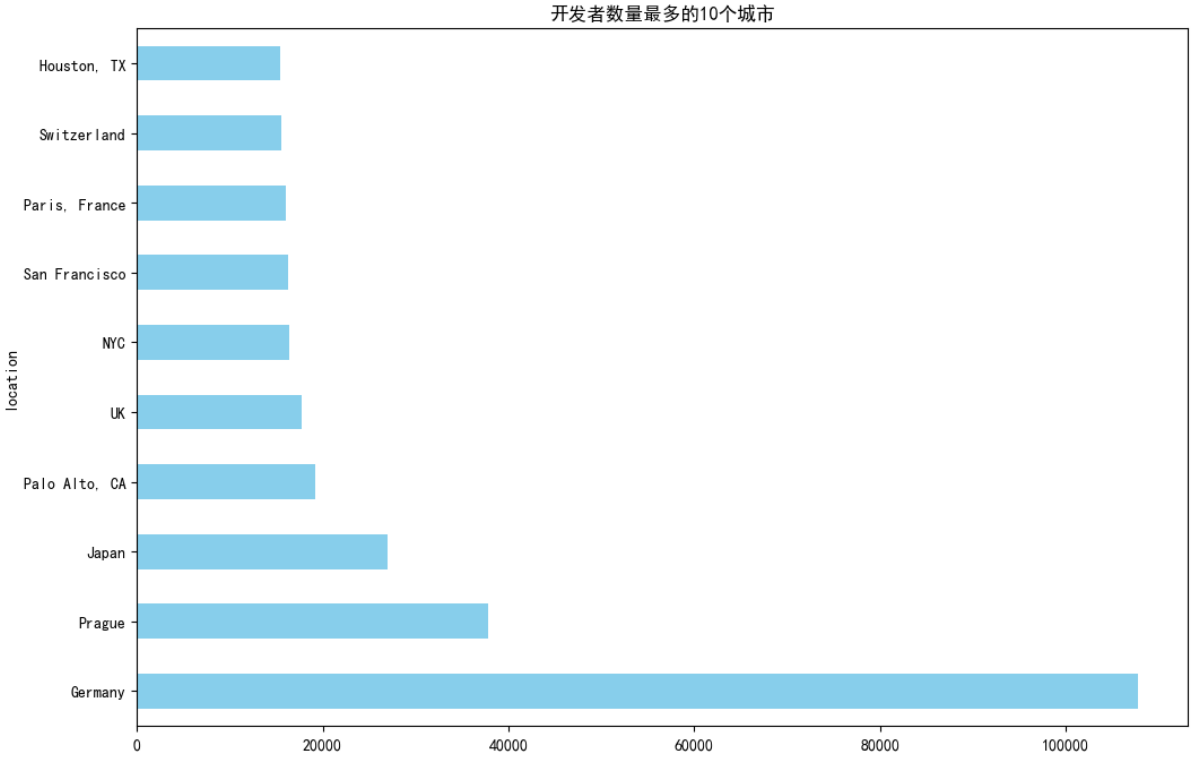
1. 国家和地区分布



用户活动数量最多的国家是美国，其次是德国和中国。这表明这些国家可能是开发者的主要集中地。其他如英国、法国、日本和捷克等国家也有显著的用户活动，显示出这些地区对 GitHub 协作的积极参与。

2. 城市级别分布



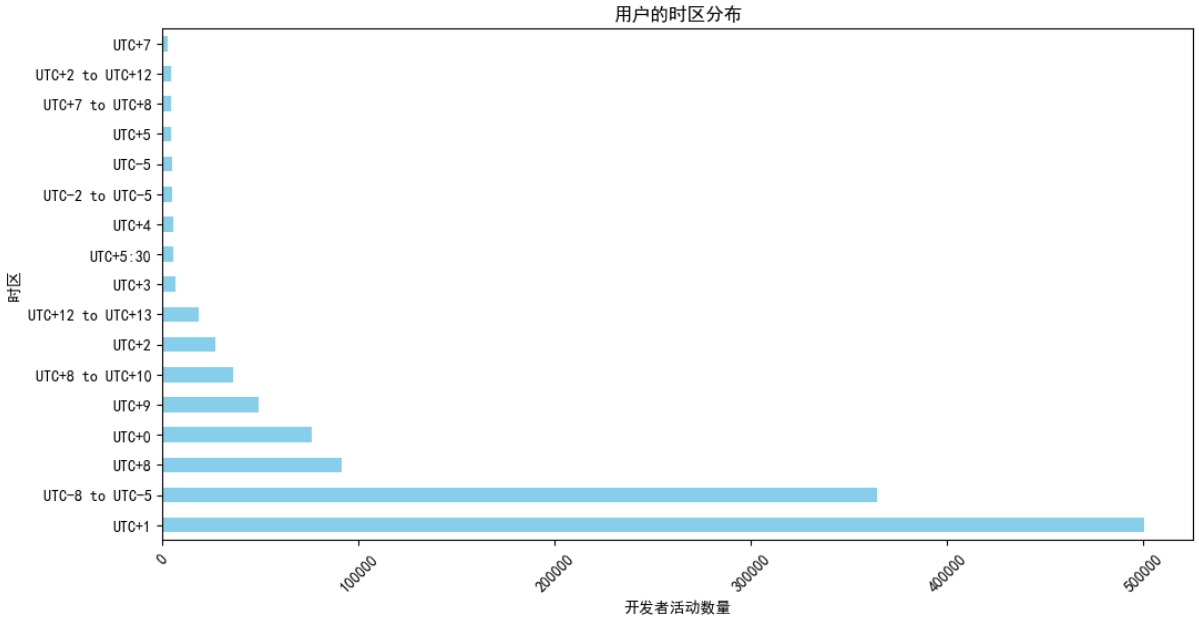


上图显示了开发者数量最多的 10 个城市。德国在这一分析中占据显著位置，显示出极高的开发者密度。其他如美国（Palo Alto, CA 和 NYC）、日本、捷克（Czechia）和法国（Paris）等城市也显示出较高的开发者集中度。这些城市可能是技术热点区域，吸引了大量的技术人才。

3. 时区分布

- **北美地区（美国、加拿大）**：用户主要分布在UTC-8至UTC-5时区，这意味着他们通常在当地时间的白天进行协作，对应亚洲地区的夜间到次日清晨。

- **欧洲地区（德国、法国、瑞士等）**：用户主要分布在UTC+1时区，他们在当地时间的白天协作，对应亚洲地区的下午到傍晚。
- **亚洲地区（中国、日本）**：用户分布在UTC+8和UTC+9时区，他们在当地时间的白天协作，对应欧洲地区的早晨到下午。



上图展示了用户的时区分布。大多数开发者活动集中在 UTC-5 和 UTC+1 时区，这与美国东部和欧洲中部的时区相吻合。这可能意味着这些地区的开发者在协作时有更多的重叠工作时间，有利于团队协作和项目推进。

综合分析

- **开发者集中地**：美国、德国和中国是开发者的主要集中地，这可能与这些国家的技术发展水平和教育体系有关。
- **技术热点区域**：柏林、Palo Alto、NYC、东京和巴黎等城市显示出高密度的开发者，这些城市可能是技术创新和创业的热点。
- **协作时间模式**：UTC-5 和 UTC+1 时区的开发者活动最为频繁，这可能有助于这些地区的团队在工作时间内进行有效的协作。

协作行为分析

1. 提交频率分析

高活跃用户分析

高活跃用户是指提交次数高于中位数的用户。可以看到一些用户的提交次数非常高，例如：

- 用户 ID 11146458 有 36224 次提交，这是非常高的活跃度，表明这个用户可能在多个项目中非常活跃或者在某个项目中承担了核心角色。
- 用户 ID 158862 有 26364 次提交，同样显示出非常高的活跃度。
- 其他用户如 28706372、1580956 和 8188402 也有显著的提交次数，分别为 16888、9668 和 8301 次。

这些高活跃用户可能对项目的进展和维护起到了关键作用，他们的行为模式和工作习惯可能对团队的协作和项目的成功至关重要。

低活跃用户分析

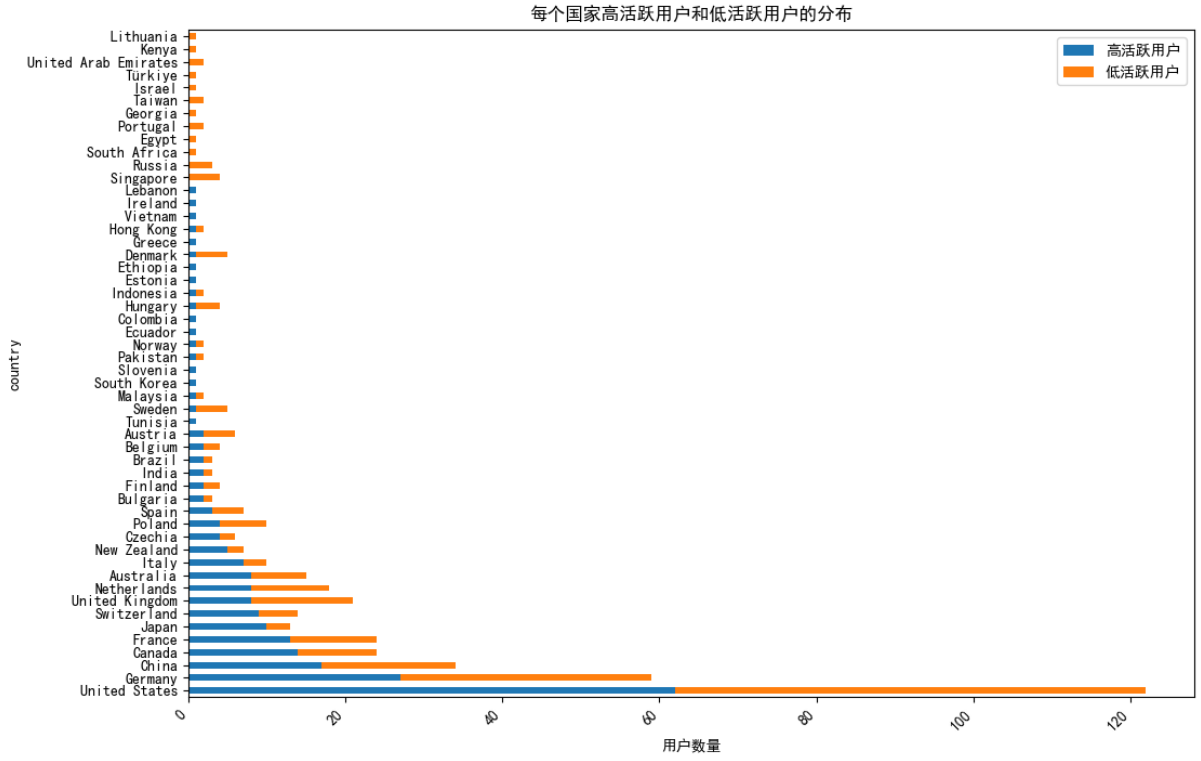
低活跃用户是指提交次数等于或低于中位数的用户。从数据中可以看出：

- 许多用户的提交次数相对较低，如用户 ID 7608904、709451 和 104888，他们的提交次数都是 1194 次，这可能意味着他们在项目中的参与度较低，或者他们可能在项目中扮演了不同的角色，如质量保证、设计或项目管理等。
- 还有一些用户的提交次数非常少，如用户 ID 814283 只有 41 次提交，985347 只有 18 次提交。这可能表明他们新近加入项目，或者他们的工作性质不需要频繁的代码提交。

综合分析

- **活跃度分布**：提交次数的分布可以帮助我们了解团队成员的活跃度和参与度。高活跃用户可能需要更多的资源和支持，而低活跃用户可能需要更多的激励或培训来提高他们的参与度。
- **项目贡献**：高活跃用户可能对项目的成功有更大的直接影响，而低活跃用户可能在其他方面对项目有所贡献，如文档、设计或测试。
- **团队动态**：了解活跃度分布可以帮助团队领导者优化团队结构，确保每个成员都能在他们最擅长的领域发挥作用。

2.活跃用户的国家分布



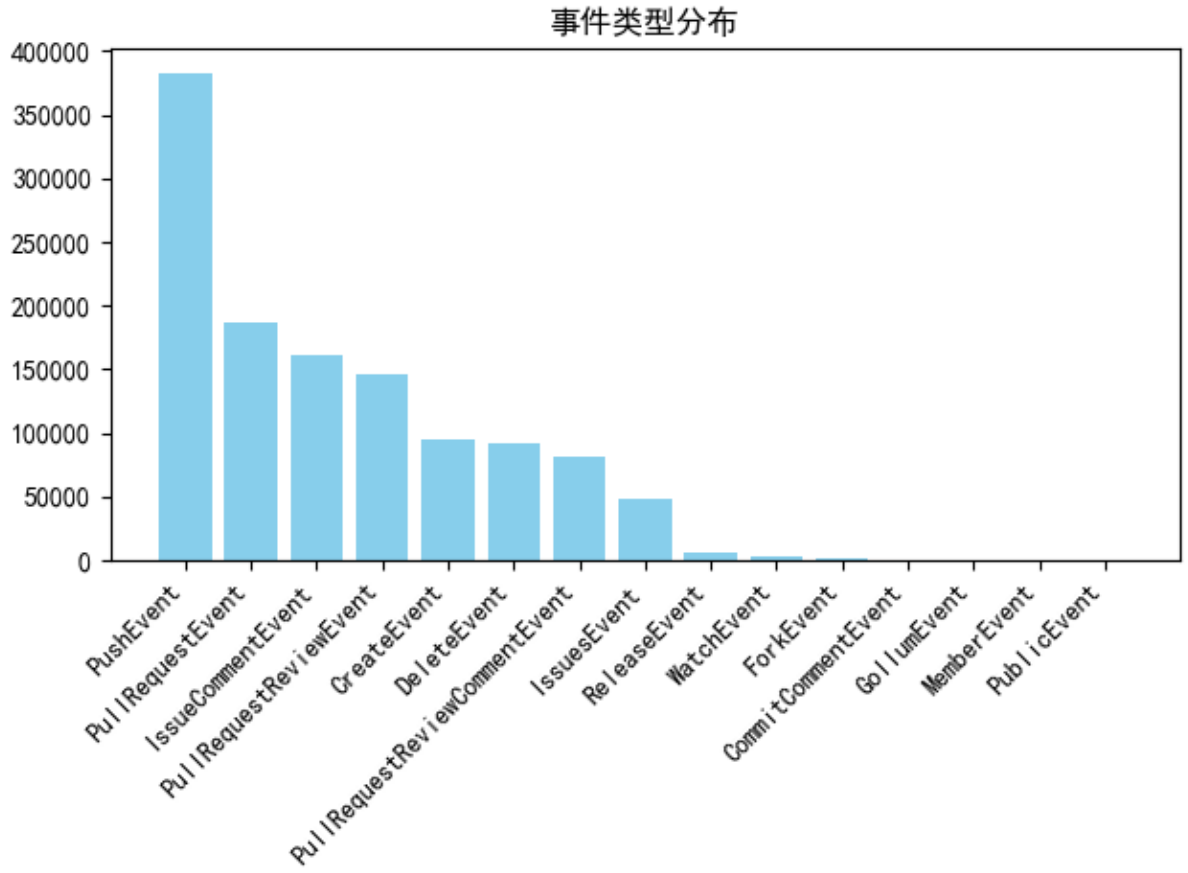
1. **美国和英国**：这两个国家在高活跃用户和低活跃用户的数量上都显著高于其他国家。美国尤其突出，无论是高活跃还是低活跃用户数量都非常庞大，这可能与其庞大的人口基数和技术行业的发达有关。
2. **德国**：在高活跃用户中，德国的用户数量较多，这可能反映了德国在技术领域的强大实力和对开源项目的贡献。
3. **中国和日本**：这两个亚洲国家在低活跃用户数量上表现较为突出，这可能与这些国家的开发者文化和工作习惯有关。
4. **其他国家**：如加拿大、法国、西班牙、意大利等国家在高活跃用户和低活跃用户的数量上也有一定的表现，但与美国、英国和德国相比，数量上有所减少。
5. **低活跃用户**：在一些国家，如立陶宛、肯尼亚、土耳其等，低活跃用户的数量相对较少，这可能与这些国家的技术行业发展水平或者对 GitHub 的使用习惯有关。

综合分析

- **技术热点区域**：美国、英国和德国是技术热点区域，拥有大量的高活跃用户，这可能与这些国家的技术发展水平和教育体系有关。
- **开发者文化差异**：不同国家的开发者文化和工作习惯可能导致了高活跃和低活跃用户数量的差异。
- **市场潜力**：对于希望扩大开发者基础的公司来说，中国和日本可能是潜在的市场，尽管低活跃用户数量较多，但这也意味着有很大的增长空间。

其他维度的洞察

1.用户参与的事件类型分布



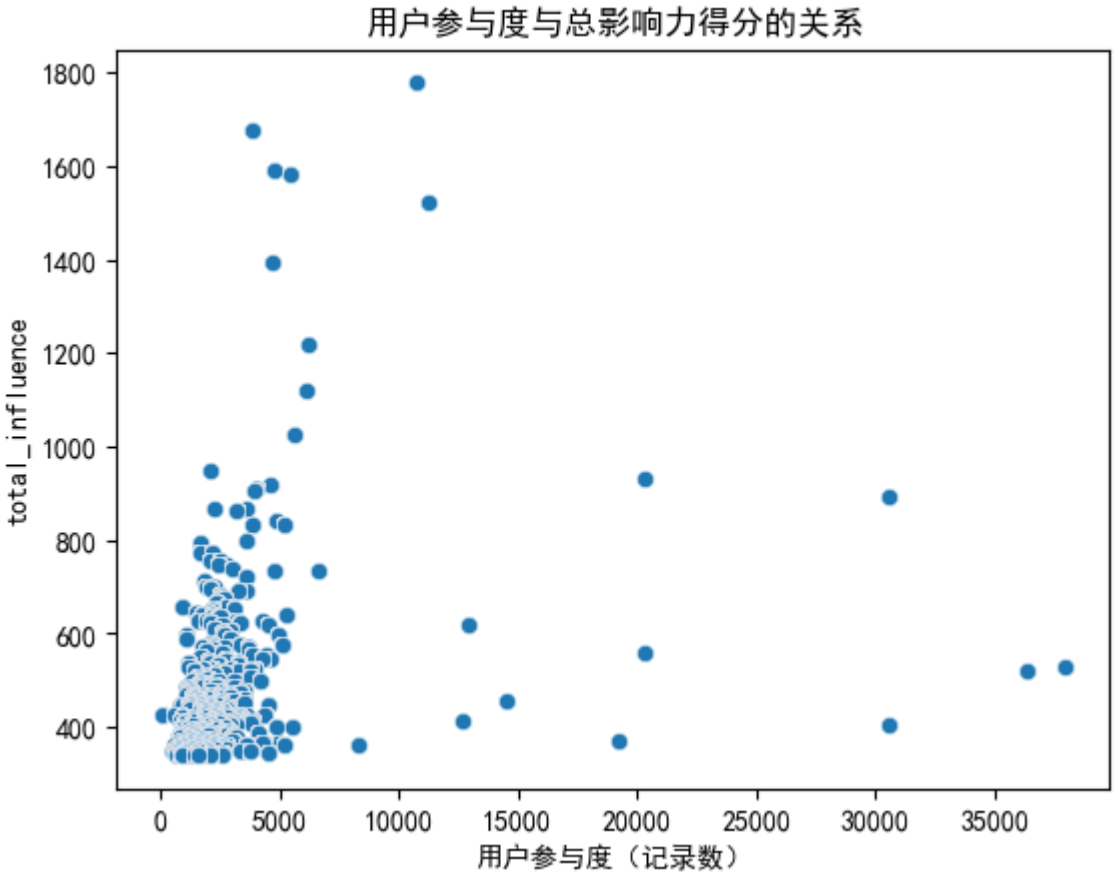
1. **PushEvent**：这是最常见的事件类型，数量远远超过其他事件。这表明用户最常进行的操作是推送代码到仓库，这是日常开发工作的一部分。
2. **PullRequestEvent**：这是第二常见的事件类型，表明用户频繁地创建拉取请求，这通常用于代码审查和合并代码到主分支。
3. **IssuesCommentEvent** 和 **IssueEvent**：这两种事件类型也相对常见，说明用户在问题跟踪和讨论中也相当活跃。
4. **CreateEvent** 和 **DeleteEvent**：这些事件类型的频率表明用户在创建和删除仓库或分支方面也有一定的活动。
5. **PullRequestReviewCommentEvent**：这个事件类型的频率显示了用户在拉取请求审查中的评论活动，这是代码审查过程的一部分。
6. **ReleaseEvent**：这个事件类型表明用户在发布新版本方面也有一定的活动，但不如前几种事件类型频繁。

7. **其他事件**：如 WatchEvent、ForkEvent、CommitCommentEvent、GollumEvent 和 MemberEvent 的数量相对较少，这可能意味着这些活动在用户的日常 GitHub 活动中不是主要部分。

综合分析

- **核心开发活动**：PushEvent 和 PullRequestEvent 的高频率强调了 GitHub 作为代码托管和协作平台的核心功能。
- **社区互动**：IssuesCommentEvent 和 IssueEvent 的频率显示了 GitHub 社区互动的重要性。
- **代码审查**：PullRequestReviewCommentEvent 的存在表明代码审查是开发流程中的一个重要环节。
- **版本控制**：ReleaseEvent 的频率表明用户在管理软件版本方面也有一定的活动。

2.用户的参与度与影响力分析



1. **正相关趋势**：图中显示了一个正相关的趋势，即随着用户参与度（记录数）的增加，用户的总影响力得分也倾向于增加。这表明在 GitHub 上更活跃的用户往往具

有更高的影响力。

2. **数据分布**：大多数数据点集中在较低的参与度和影响力得分区域，这可能意味着大部分用户的参与度和影响力都处于中等或较低水平。
3. **异常值**：图中有一些点远离主要数据群，这些可能是异常值或特别有影响力的用户。例如，有一个用户在接近 10,000 次记录的情况下，影响力得分超过了 1700，这显著高于其他用户。
4. **影响力得分的波动**：在较低的参与度水平上，影响力得分的波动较大，这可能表明即使参与度不高，某些用户也可能因为高质量的贡献或其他因素而具有较高的影响力。
5. **参与度的极端值**：在参与度的极端值（非常高或非常低）处，影响力得分的波动较小，这可能意味着在这些极端情况下，用户的参与度与影响力得分的关系更加稳定。

综合分析

- **高影响力用户**：少数用户在 GitHub 上具有非常高的影响力，这可能与他们的专业技能、项目质量或社区参与度有关。
- **中等影响力用户**：大多数用户的影响力处于中等水平，这可能需要更多的努力或策略来提升他们的影响力。
- **低影响力用户**：一些用户尽管参与度不高，但仍然可能具有较高的影响力，这表明影响力并不完全取决于参与度。

结论

综合以上分析，对 GitHub 上具有协作行为日志数据的 500 名用户的个人信息进行了深入的数据洞察，得出以下结论：

1. **开发者地理分布**：
    - 美国、英国、中国、德国和日本是开发者的主要集中地，显示出这些国家在技术领域的重要地位。
    - 城市级别的分析显示，柏林、Palo Alto、NYC、东京和巴黎等城市具有高密度的开发者，可能是技术创新和创业的热点区域。
2. **开发者活跃度**：
    - 用户的活跃度在不同国家和城市之间存在显著差异，美国和英国的用户在活跃度上领先。

- 高活跃用户对项目的贡献显著，而低活跃用户可能在其他方面如质量保证、设计或项目管理上有所贡献。

3. **事件类型分布**：
   - PushEvent 是最常见的事件类型，表明代码推送是日常开发的核心活动。
   - PullRequestEvent 和 IssuesCommentEvent 的高频率强调了代码审查和问题讨论在协作中的重要性。
4. **用户参与度与影响力**：
   - 用户的参与度与其影响力之间存在正相关关系，即更活跃的用户往往具有更高的影响力。
   - 存在一些异常值，表明有少数用户尽管参与度不高，但可能因为高质量的贡献而具有较高的影响力。
5. **协作行为模式**：
   - 用户的协作行为模式揭示了 GitHub 作为代码托管和协作平台的核心功能，以及社区互动的重要性。
   - 代码审查和版本控制是开发流程中的关键环节，这在事件类型分布中得到了体现。
6. **市场和人才策略**：
   - 对于希望扩大开发者基础的公司来说，中国和日本可能是潜在的市场，尽管低活跃用户数量较多，但这也意味着有很大的增长空间。
   - 技术公司可以根据这些洞察来优化人才招聘策略，专注于那些具有高影响力和高活跃度的开发者。

这些结论为理解 GitHub 用户的行为模式、优化平台功能、增强用户体验和制定市场策略提供了宝贵的数据支持。