数据分析洞察报告

余超毅

10235501470

一、引言

本报告旨在分析截至 2024 年 8 月底在 GitHub 上具有协作行为日志数据的用户的个人信息，包括姓名、公司、邮箱及其地理位置等。通过数据洞察分析，我们希望培养数据处理与分析能力，掌握 GPT 工具的应用，并理解数据隐私与伦理。

二、数据来源

数据获取链接为：[GitHub Dase-2024-Autumn User Data](https://github.com/X-lab2017/dase-2024-autumn/tree/main/HomeWork/data/user_data)

三、实验目标

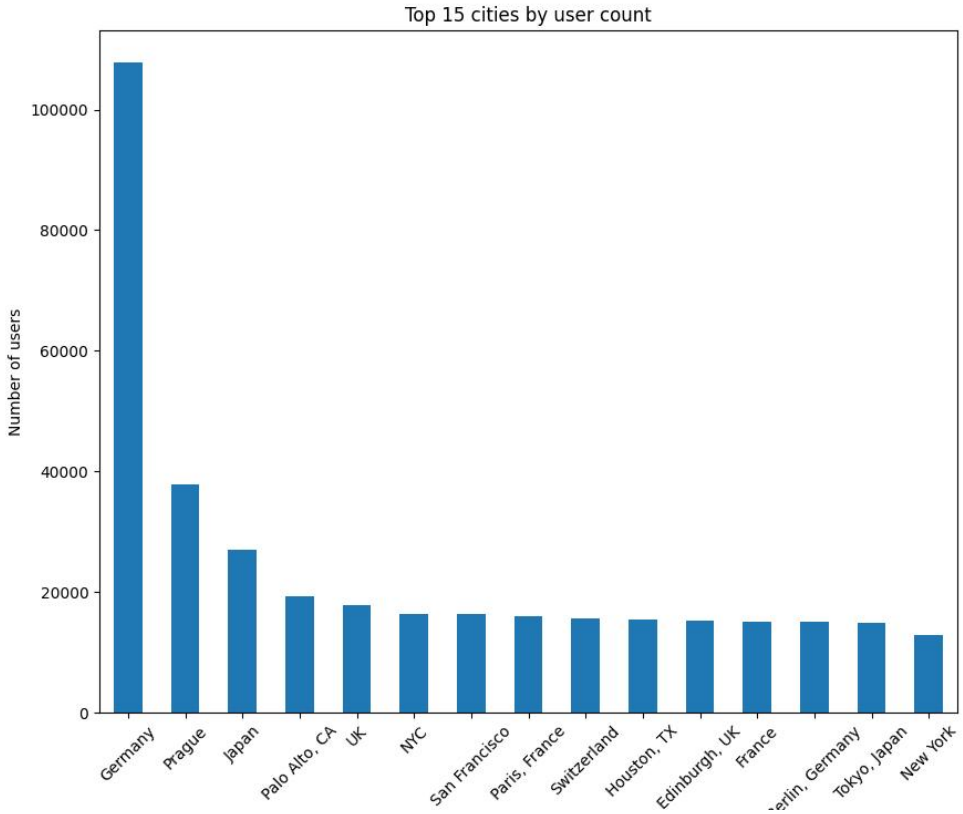
1. 培养数据处理与分析能力。
2. 掌握 GPT 工具的应用。
3. 理解数据隐私与伦理。

四、实验内容

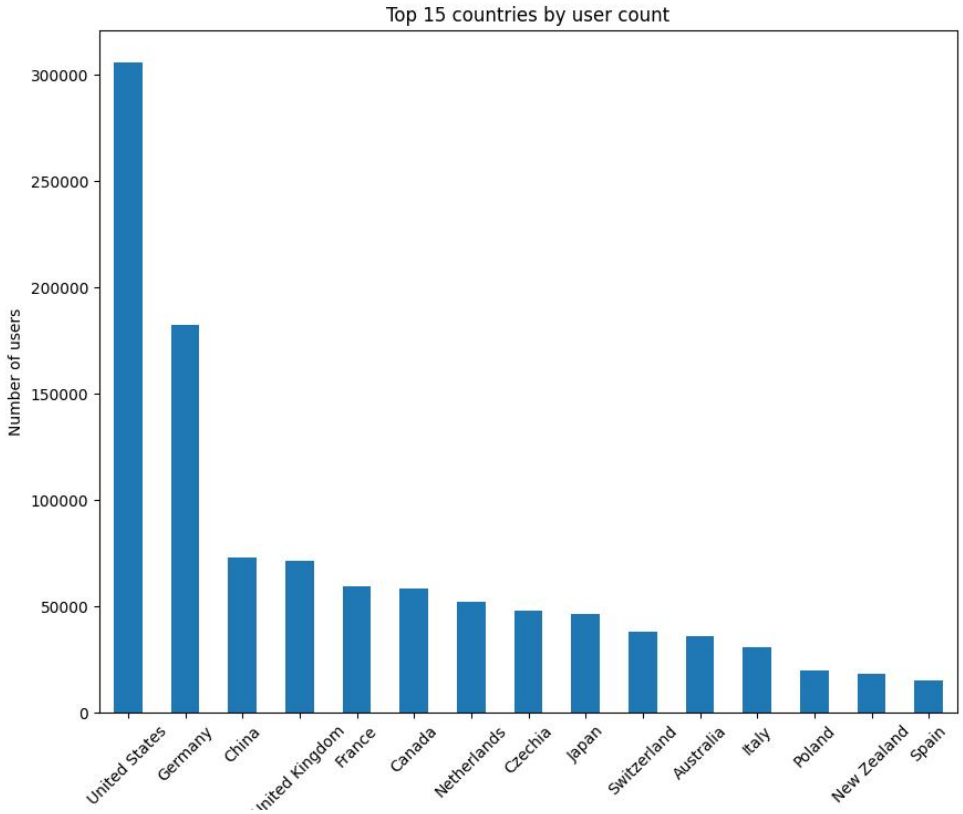
1. 人口统计分析
   - 国家和地区分布：统计用户所在国家和地区的分布，识别主要的开发者集中地。
   - 城市级别分布：分析主要城市的开发者密度，发现技术热点区域。
   - 时区分布：了解用户的时区分布，分析不同地区用户的协作时间模式。
2. 协作行为分析
   - 提交频率：统计每个用户的提交次数，识别高活跃用户和低活跃用户。
   - 其他维度有趣的洞察（至少 2 个）。

五、数据分析

1. 城市用户数量分布
   - 德国拥有最多的用户，其次是布拉格和日本。
   - 其他城市如帕洛阿尔托、伦敦、纽约等也有显著的用户数量。

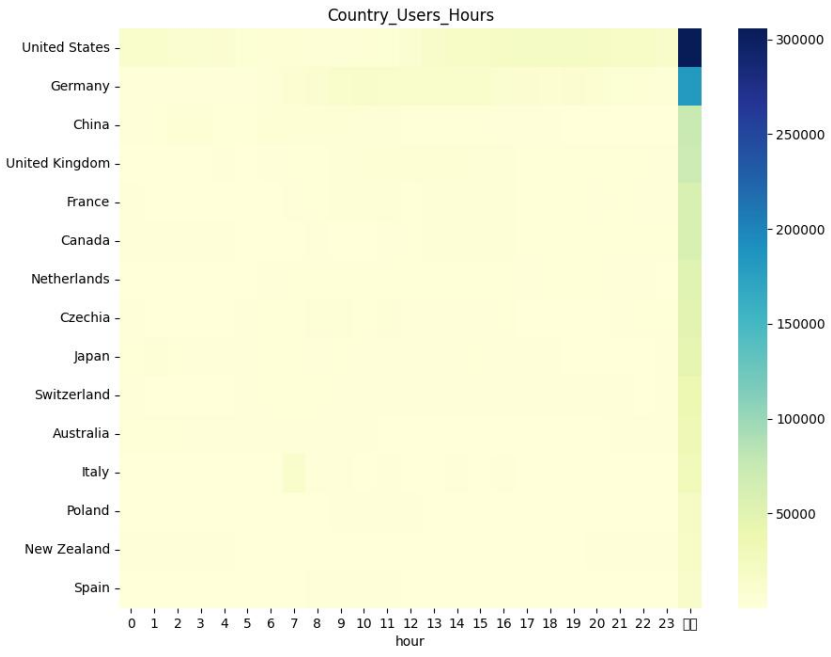


2. 国家用户数量分布
   - 美国和德国是用户数量最多的两个国家。
   - 中国、英国、法国等国家紧随其后。

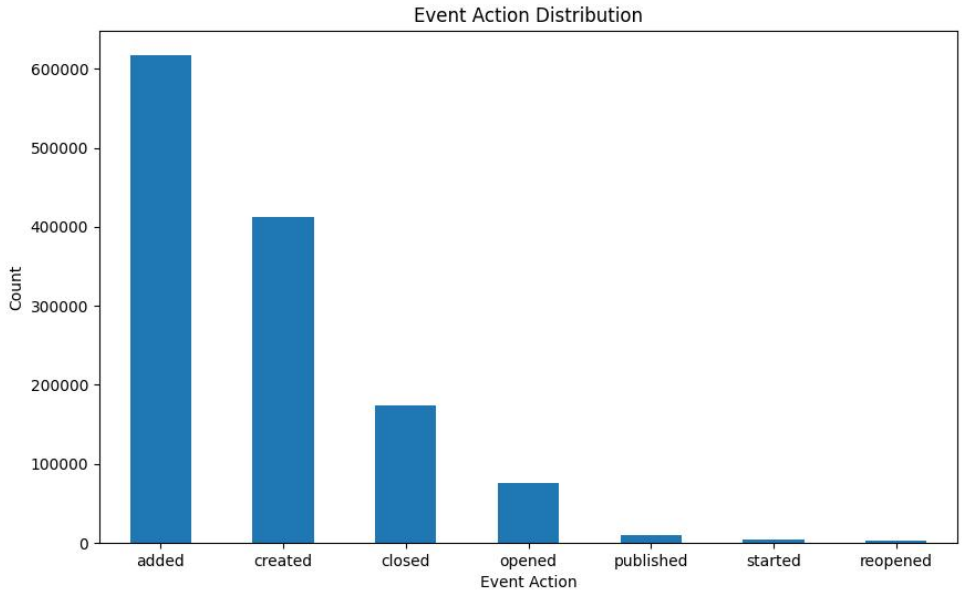


3. 国家用户活跃时间分布

- 用户活跃时间在不同国家有所差异，美国和德国的用户在一天中的活跃时间分布较为均匀。

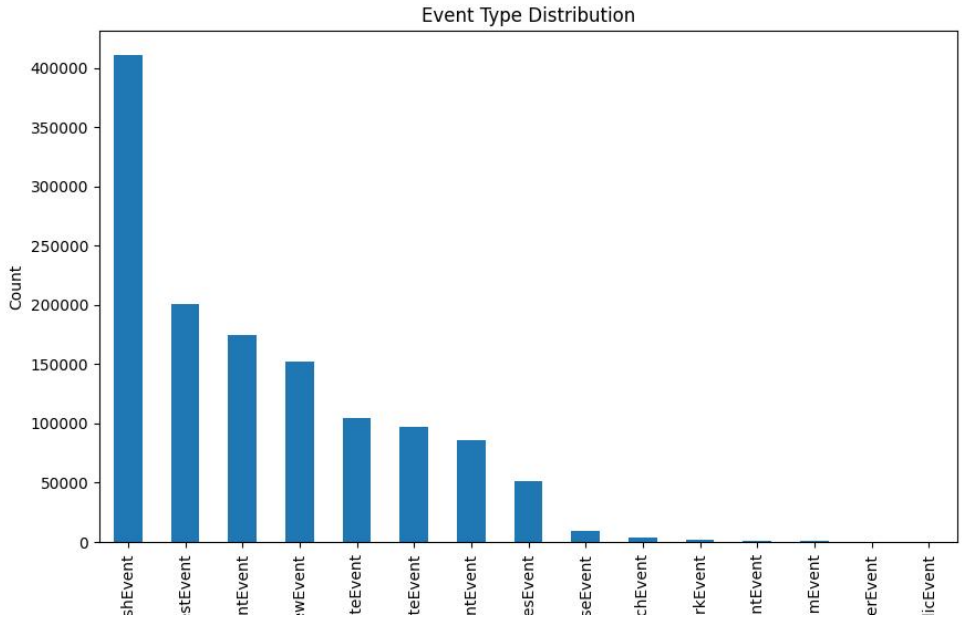


4. 事件行为分布
   - "added" 和 "created" 是最常见的事件行为，而 "started" 和 "reopened" 则较少见

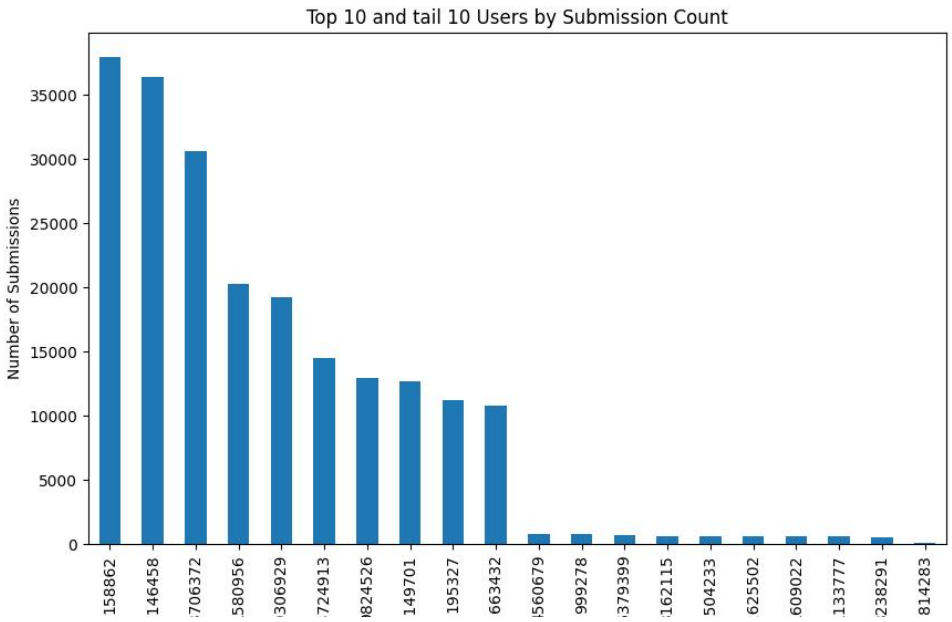


。

5. 事件类型分布
   - "PushEvent" 是最常见的事件类型，其次是 "IssuesEvent" 和 "PullRequestEvent"。



6. 用户提交次数分布
   - 用户 158862 的提交次数最多，远超其他用户。
   - 其他用户的提交次数相对较少，显示出明显的长尾分布。



六、结论

通过本次数据分析，我们发现德国和美国是 GitHub 用户的主要集中地。在城市级别，德国的用户数量遥遥领先。用户活跃时间在不同国家有所差异，但总体上较为均匀。在事件行为和类型上，"added" 和 "created" 是最常见的，而 "PushEvent" 是最常见的事件类型。用户提交次数呈现出明显的长尾分布，少数用户贡献了大部分的提交。

七、建议

1. 针对用户活跃时间，可以优化协作工具的使用体验，以适应不同地区的用户习惯。
2. 考虑在用户数量较多的国家和城市举办更多的技术交流和培训活动。
3. 鼓励用户参与更多的事件类型，以促进社区的多样性和活跃度。