《GitHub 用户数据洞察报告》

一、引言

本报告基于从 GitHub 收集的 500 名用户的个人信息及协作行为数据进行分析，旨在深入了解 GitHub 用户的分布特征、协作模式以及影响力等方面的情况，为相关研究和决策提供有价值的参考。

二、数据概述

（一）数据来源

https://github.com/X-lab2017/dase-2024-autumn/tree/main/HomeWork/data/user_data，包含七个部分（users_combined_info_500_part_1.csv 到 users_combined_info_500_part_7.csv），涵盖了用户的姓名、公司、邮箱、地理位置、事件时间、影响力等多维度信息。

（二）数据预处理

在进行分析之前，对数据进行了一系列预处理操作。首先，处理了 location 和 country 字段的缺失值，删除了相关字段为空的行，以确保数据的完整性和准确性。其次，对 event_time 字段进行了处理，提取了时区和小时信息，同时将 total_influence 列的数据统一四舍五入到三位小数，便于后续的分析和可视化展示。

三、人口统计分析

（一）国家和地区分布

国家分布：数据集中共涉及 52 个国家，其中用户数量排名前 20 的国家依次为德国、美国、中国、加拿大、法国、日本、捷克、瑞士、英国、意大利、澳大利亚、荷兰、新西兰、波兰等。德国的用户数量最多，达到 175,349 人，占总用户数的 14.53%；美国次之，有 99,673 人，占比 8.26%；中国用户数量为 61,433 人，占比 5.09%。这些国家在 GitHub 上具有较高的用户活跃度，可能与其科技发展水平、开源文化氛围以及技术人才储备等因素密切相关。

地区分布：在地区层面，前 20 名地区中，德国地区的用户数量遥遥领先，达到 107,747 人，占比 8.93%；布拉格、日本、加利福尼亚州帕洛阿尔托、英国等地区也聚集了较多的用户。这些地区可能是当地技术产业的核心区域，吸引了大量开发者参与 GitHub 项目。

（二）城市级别分布

城市级别分布方面，德国城市以 111,786 名开发者位居榜首，占总开发者数量的 9.26%，显示出德国在技术人才方面的高度集中。布拉格、旧金山、日本城市（如东京）、柏林、纽约等城市也进入了前 20 名，这些城市通常是全球或区域的科技、经济和文化中心，拥有丰富的技术资源和创新氛围，吸引了大量开发者汇聚于此。

（三）时区分布

通过对事件时间的分析，确定了不同时区的用户活跃情况。每个时区都有其最活跃的小时，例如+00:00 时区最活跃小时为 10 点，+01:00 时区为 7 点等。这反映了不同地区用户的工作和生活习惯对其在 GitHub 上活动时间的影响，对于跨时区协作和项目管理具有重要的参考价值。

四、协作行为分析

（一）提交频率

统计了每个用户的提交次数，平均提交次数为 2611.74 次。根据提交次数将用户分为高活跃用户（提交次数大于平均提交次数）和低活跃用户（提交次数小于等于平均提交次

数）。高活跃用户有 129 人，低活跃用户有 333 人。高活跃用户的提交次数显著高于平均水平，他们在项目开发和维护中发挥着重要作用，可能是核心开发者或对项目高度投入的贡献者；而低活跃用户虽然提交次数相对较少，但也是 GitHub 社区的重要组成部分，可能在其他方面如代码审查、问题反馈等方面也有一定的参与度。

五、其他洞察分析

（一）不同国家的用户平均影响力

不同国家的用户平均影响力存在一定差异。例如，某些国家的用户平均影响力较高，可能表明这些国家的开发者在技术水平、项目影响力或社区贡献等方面具有一定的优势。然而，平均影响力受到多种因素的综合影响，包括国家的技术生态、开源项目的质量和数量、开发者的经验和技能等。

（二）城市与国家的影响力比例关系

在城市与国家的影响力比例关系方面，一些城市在其所属国家的影响力占比较高，如黎巴嫩的 Aaramoun、埃塞俄比亚的亚的斯亚贝巴等，这些城市可能是该国技术创新和开源活动的重要引擎，对国家整体的技术影响力贡献较大。同时，也有一些城市如美国的纽约、旧金山等，虽然在国家内的影响力占比不是特别突出，但由于国家整体的技术实力较强，这些城市的影响力绝对值仍然较高。

（三）不同时区高活跃用户的分布

不同时区高活跃用户的分布呈现出不均衡的特点。+01:00 时区的高活跃用户数量最多，达到 54 人，而其他时区的高活跃用户数量相对较少。这可能与该时区覆盖的地区的技术产业规模、工作文化以及开源社区的活跃度等因素有关。了解不同时区高活跃用户的分布情况，有助于更好地组织跨时区的技术交流和协作活动，提高协作效率。

（四）不同国家用户提交频率的标准差

不同国家用户提交频率的标准差反映了各国用户提交行为的离散程度。一些国家如澳大利亚、奥地利等，提交频率的标准差较大，说明这些国家的用户在提交行为上存在较大的差异，可能是由于不同项目的性质、团队协作模式以及开发者个人习惯等因素导致的；而另一些国家的标准差较小，表明用户的提交行为相对较为一致。

六、结论

（一）主要发现总结

本报告通过对 GitHub 用户数据的深入分析，揭示了用户在国家、地区、城市和时区等层面的分布特征，以及协作行为和影响力等方面的情况。主要发现包括：德国、美国和中国等国家在 GitHub 用户数量上占据领先地位；特定地区和城市如德国地区、布拉格、旧金山等是技术人才的集中地；不同时区的用户活跃时间存在差异；用户提交频率存在高活跃和低活跃之分，且不同国家用户的平均影响力和提交频率的离散程度各不相同。

（二）对相关领域的启示

这些发现对于多个领域具有重要的启示意义。对于企业而言，在全球人才招聘和技术合作方面，可以根据用户分布和活跃度情况，有针对性地拓展业务和寻找合作伙伴；对于开源项目团队，了解不同地区和时区用户的行为模式，有助于优化项目管理和协作流程，提高项目的开发效率和质量；对于技术社区和平台运营者，这些数据可以为社区建设、活动策划和资源分配提供依据，促进技术交流和知识共享。

（三）研究局限性与未来展望

本研究也存在一些局限性，例如数据样本量相对较小，可能无法完全代表整个 GitHub

用户群体的特征；数据仅涵盖了部分用户信息，对于用户的技术栈、项目类型等细节方面的信息缺乏深入了解。未来的研究可以进一步扩大数据样本，收集更全面的用户数据，采用更先进的数据分析方法和模型，深入挖掘用户行为背后的动机和影响因素，为技术创新和社区发展提供更精准的支持和指导。