数据洞察报告

邓锦　　10235501434

数据来源

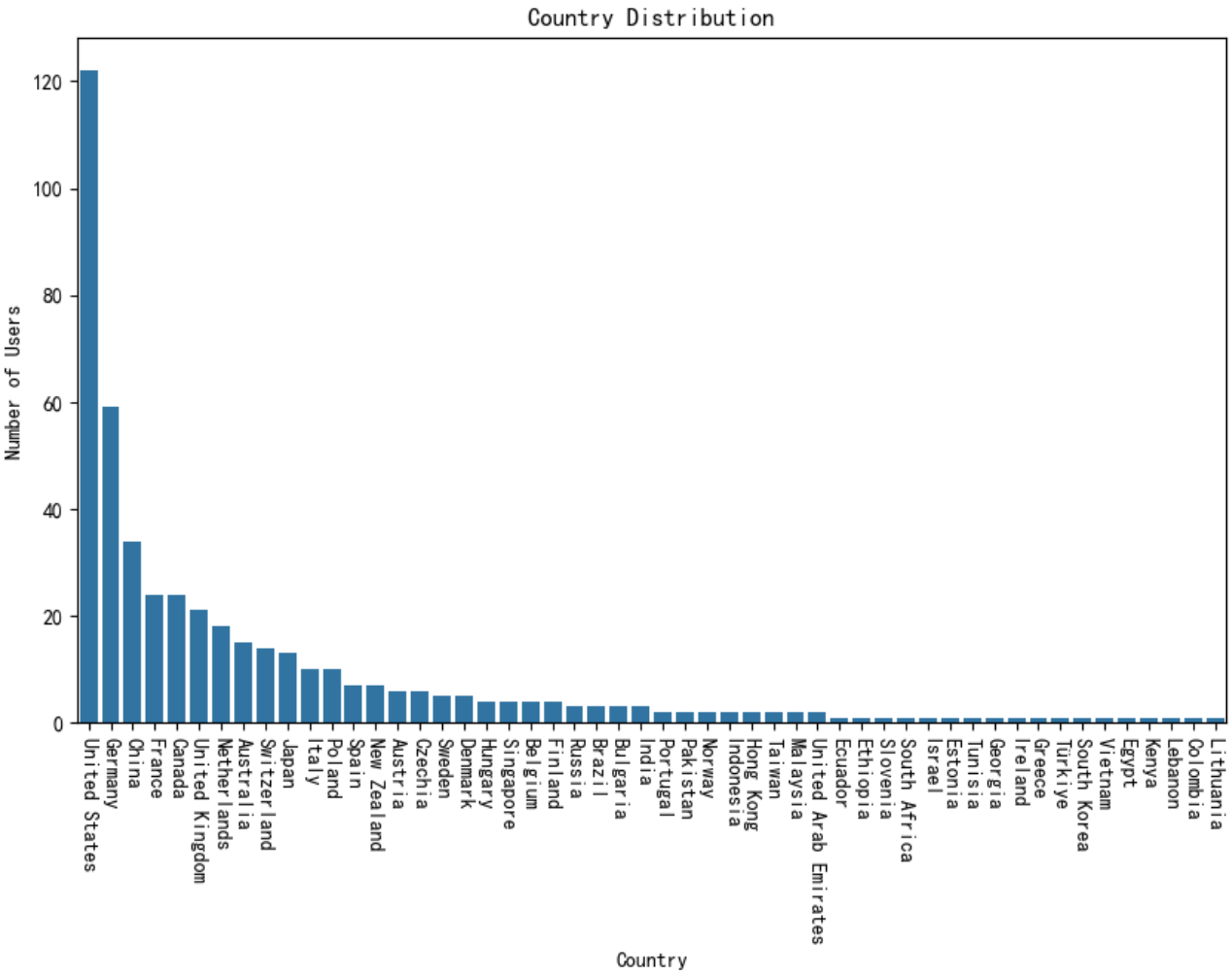
GitHub 上具有协作行为日志数据的 500 名用户的个人信息（包括姓名、公司、邮箱及其地理位置等）

数据洞察分析

1. 人口统计分析

- 国家地区分布

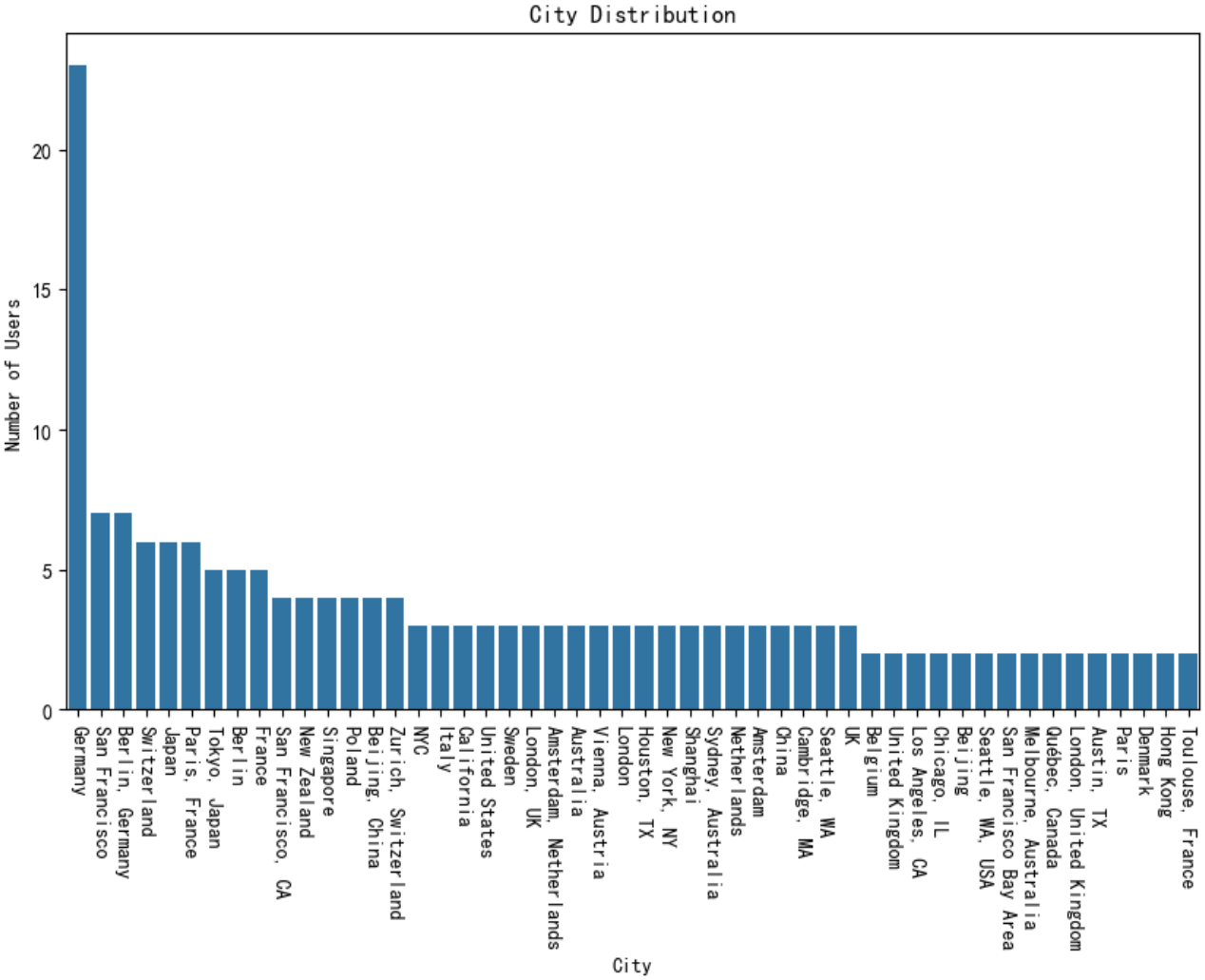
  数据:



  分析: 美国的用户数量最多，几乎是第二名德国的用户数量的两倍，有大约 120 人。中国用户数量排在第三，再往后依次是法国，加拿大，英国，荷兰，澳大利亚，瑞士等等。

- 城市级别分布

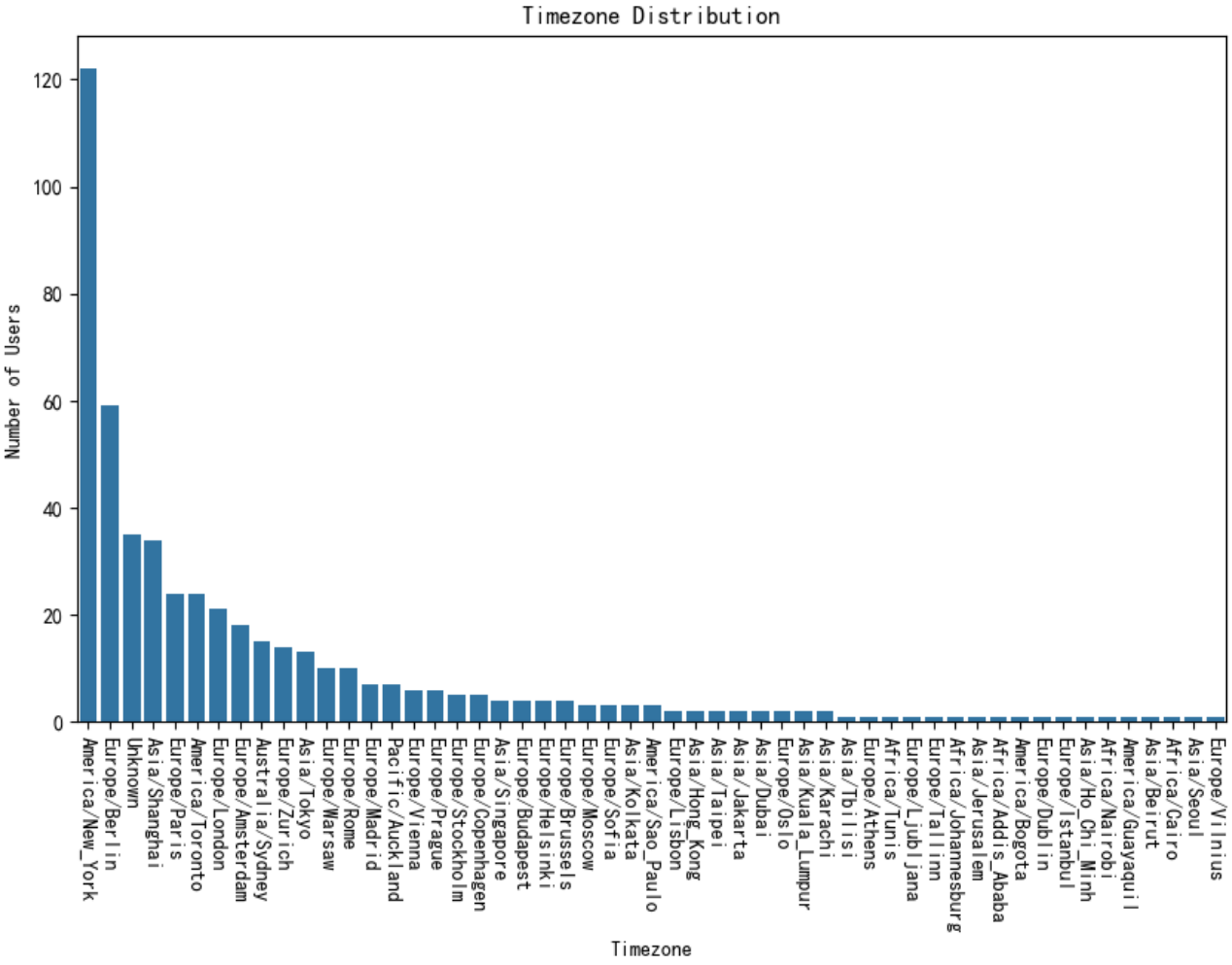
  数据:

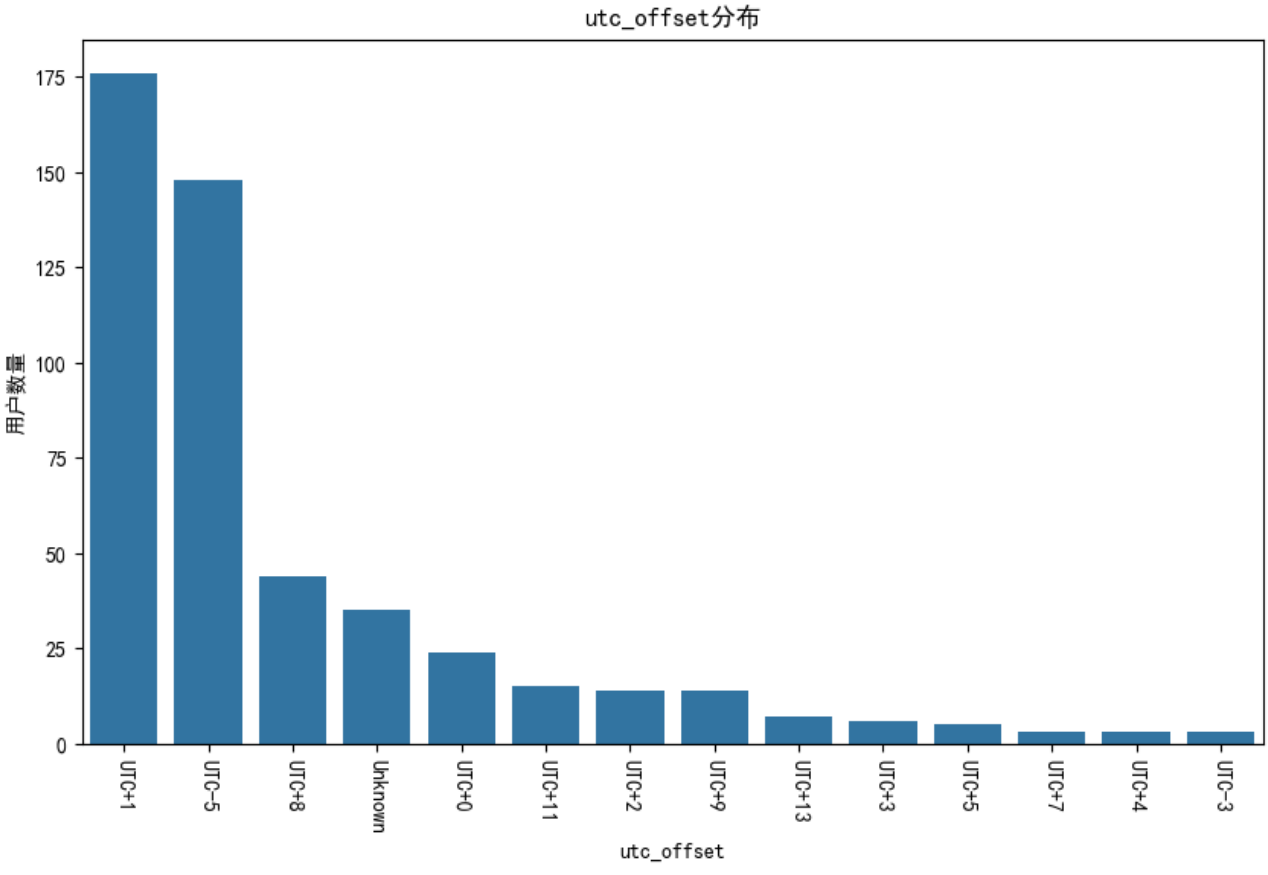


分析：技术热点区域有旧金山，柏林，巴黎，东京，北京，苏黎世等等(忽略 location 里的国家只保留城市)。

- 时区分布

  数据:


Timezone Distribution
Number of Users
0
20
40
60
80
100
120
America/New_York
Europe/Berlin
Unknown
Asia/Shanghai
Europe/Paris
America/Toronto
Europe/London
Europe/Amsterdam
Australia/Sydney
Europe/Zurich
Asia/Tokyo
Europe/Warsaw
Europe/Rome
Europe/Madrid
Pacific/Auckland
Europe/Vienna
Europe/Prague
Europe/Stockholm
Europe/Copenhagen
Asia/Singapore
Europe/Budapest
Europe/Helsinki
Europe/Brussels
Europe/Moscow
Europe/Sofia
Asia/Kolkata
America/Sao_Paulo
Europe/Lisbon
Asia/Hong_Kong
Asia/Taipei
Asia/Jakarta
Asia/Dubai
Europe/Oslo
Asia/Kuala_Lumpur
Asia/Karachi
Asia/Tbilisi
Europe/Athens
Africa/Tunis
Europe/Ljubljana
Europe/Tallinn
Africa/Johannesburg
Asia/Jerusalem
Africa/Addis_Ababa
America/Bogota
Europe/Dublin
Europe/Istanbul
Asia/Ho_Chi_Minh
Africa/Nairobi
America/Guayaquil
Asia/Beirut
Africa/Cairo
Asia/Seoul
Europe/Vilnius
Timezone

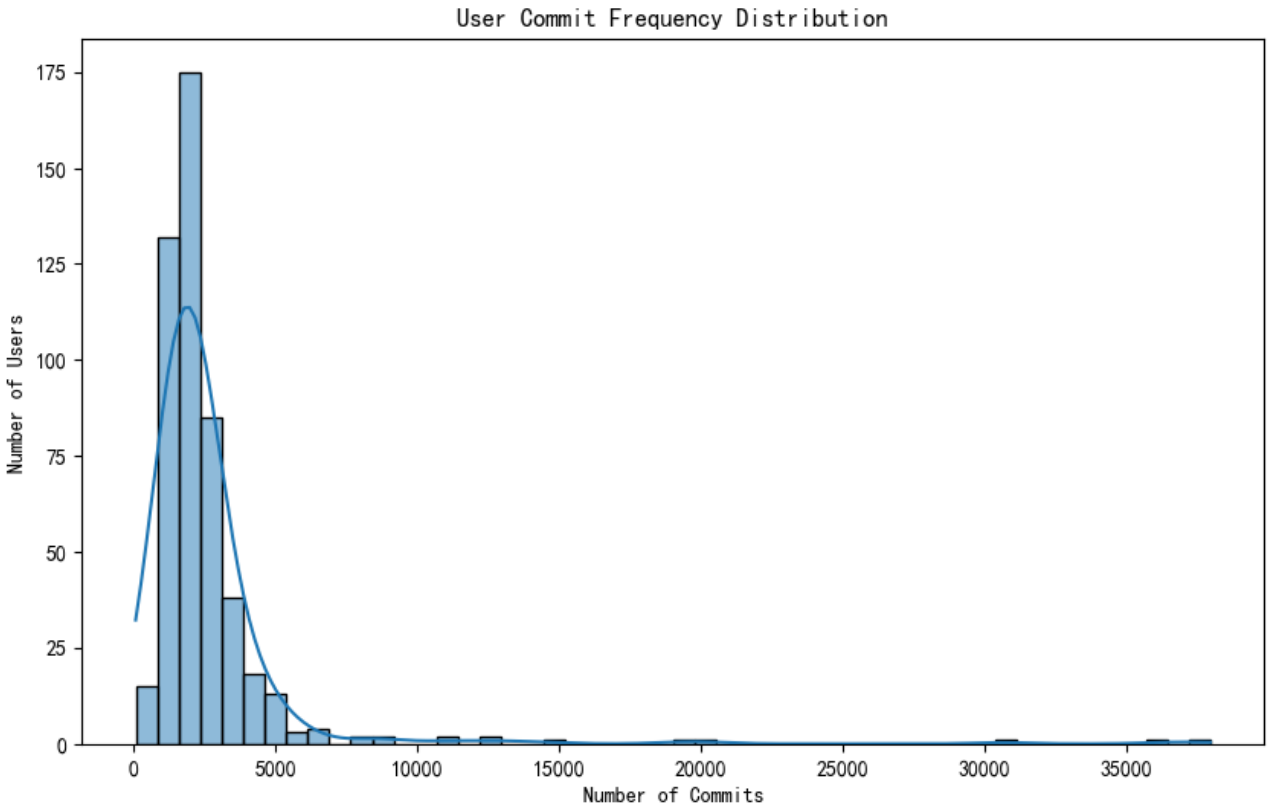

utc_offset分布
用户数量
0
25
50
75
100
125
150
175
UTC+1
UTC-5
UTC+8
Unknown
UTC+0
UTC+11
UTC+2
UTC+9
UTC+13
UTC+3
UTC+5
UTC+7
UTC+4
UTC-3
utc_offset

分析：时区是根据国家来生成的，所以国家的如何分布，时区如何分布(图一)。从图二可以看出开发者主要分布在东一区和西五区（分是第一第二），第三是东八区，第四是未知（存在用户未填写国家），第五是中时区……

2. 协作行为分析

- 提交频率

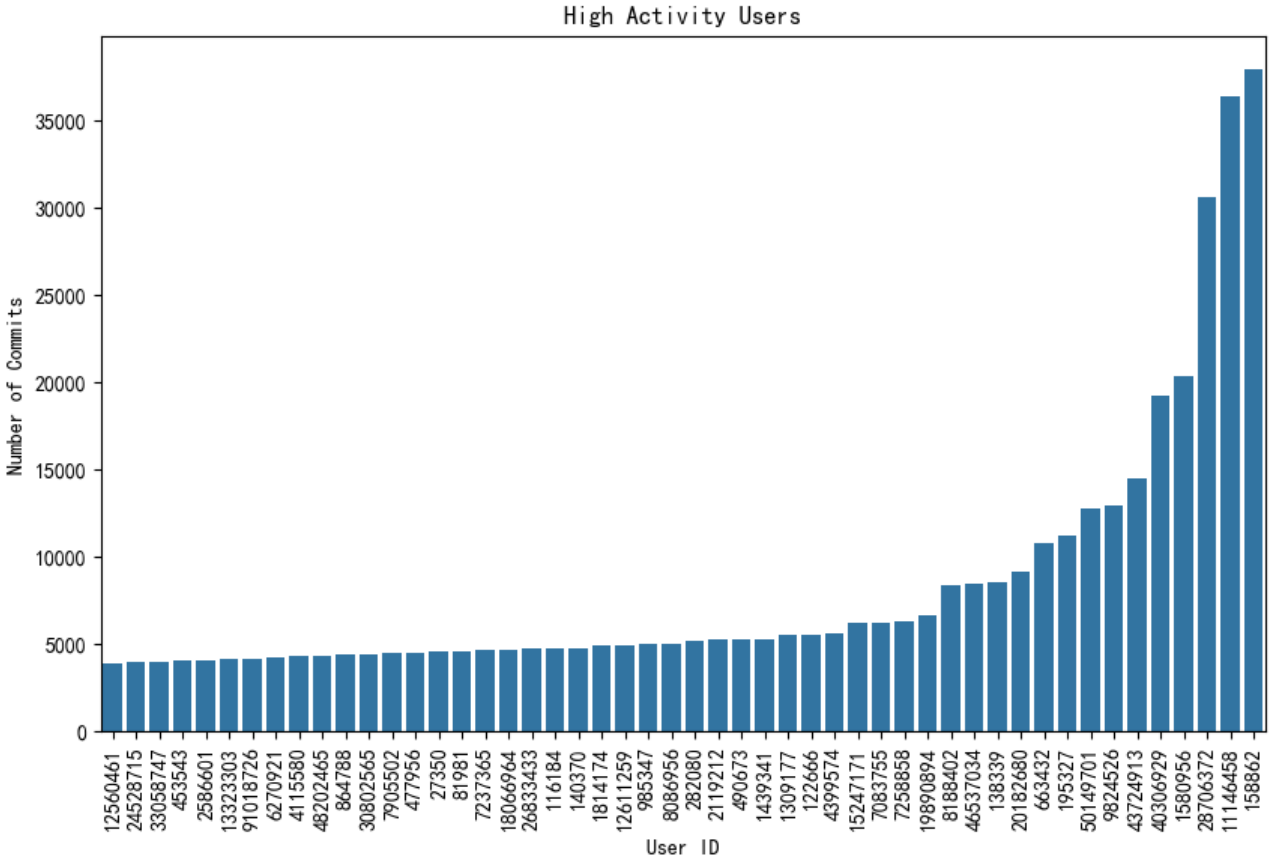
  数据:

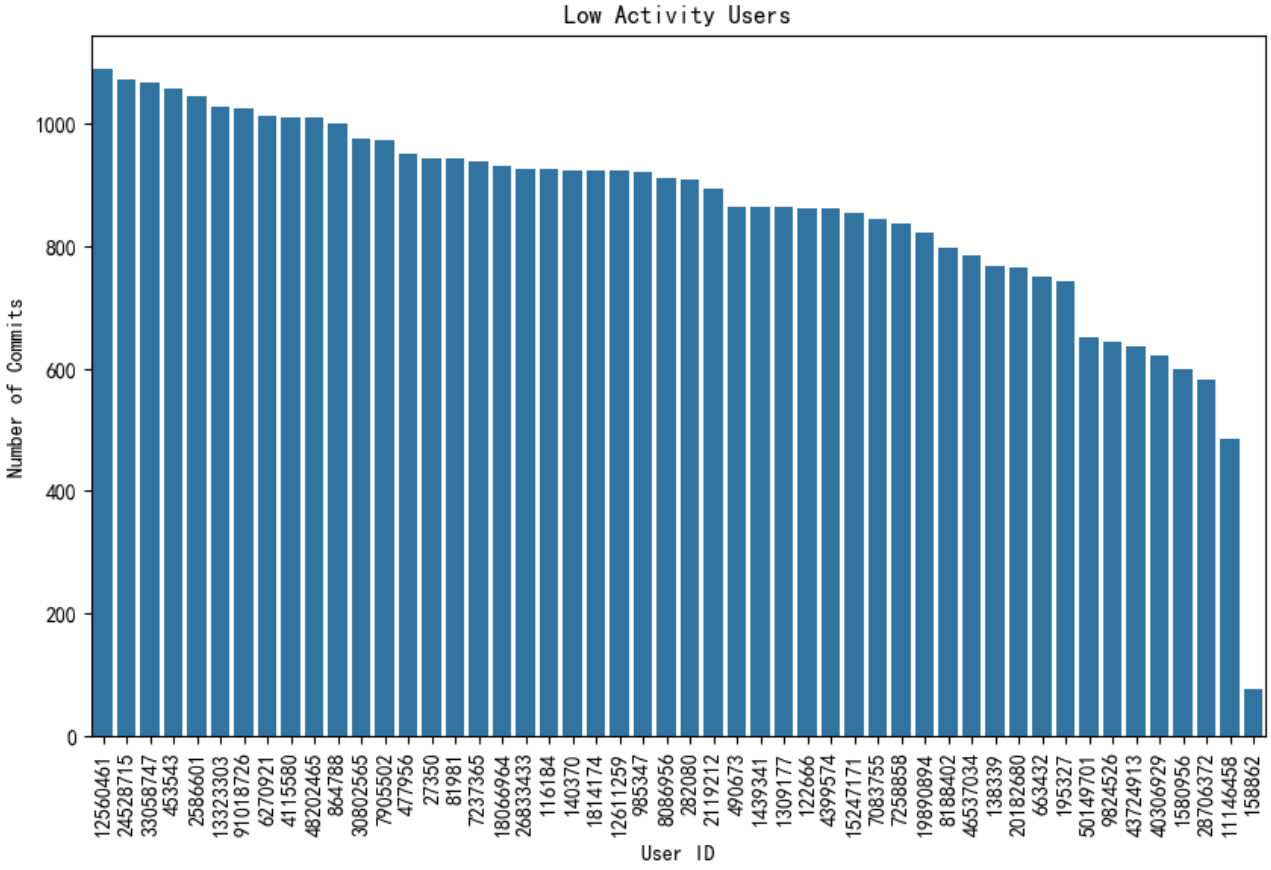


分析：绝大多数用户都是低频用户，提交频数在 0~5000 之间。极少数用户的提交频数超过 5000，有一两个用户的提交频数来到了 35000 左右。

- 高频低频用户

  数据:





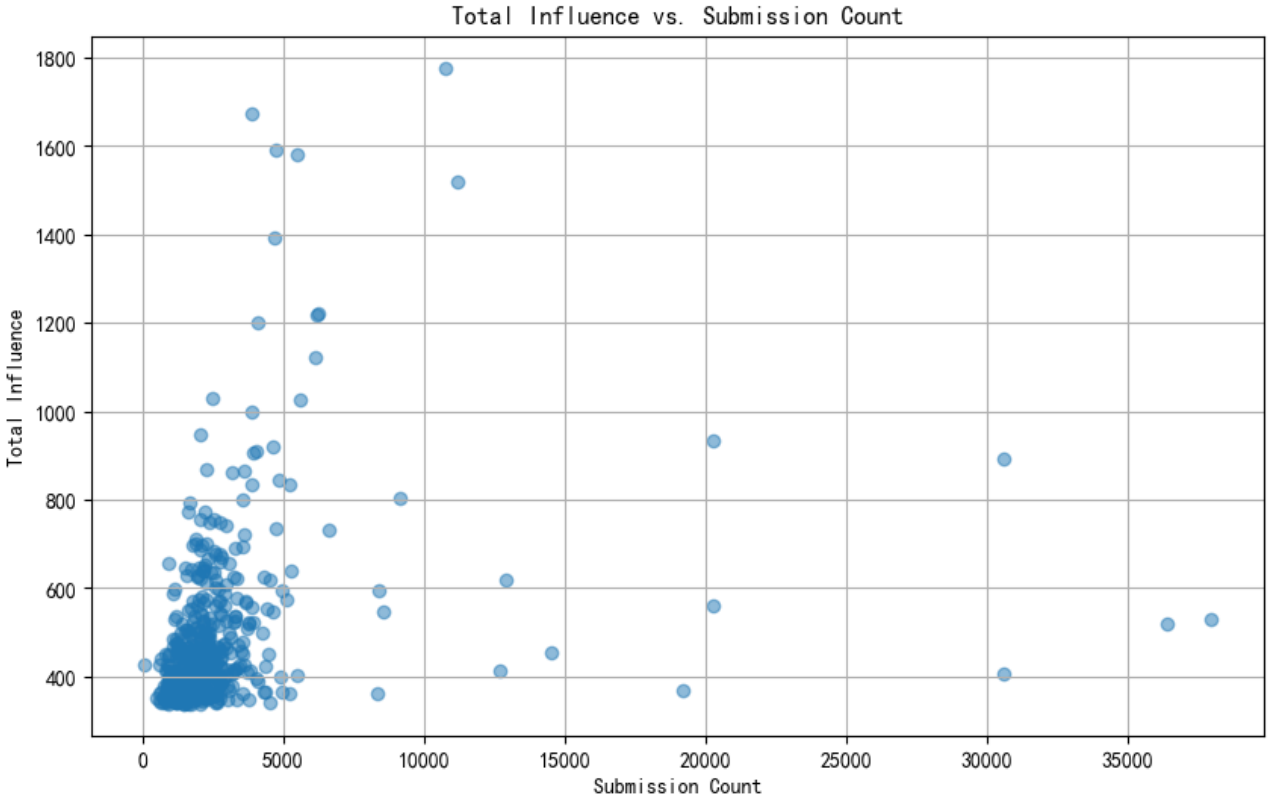
分析：图一为前 10%的用户，图二为后 10%的用户。高频用户少而且提交次数断层式高于一般用户，曲线激增。低频用户多且稳定，曲线平缓，有一个用户提交次数比低频用户还要

低，可能是死用户（账户丢失）。

3. 影响力分析

- 总影响力与用户提交频率的关系

  数据:

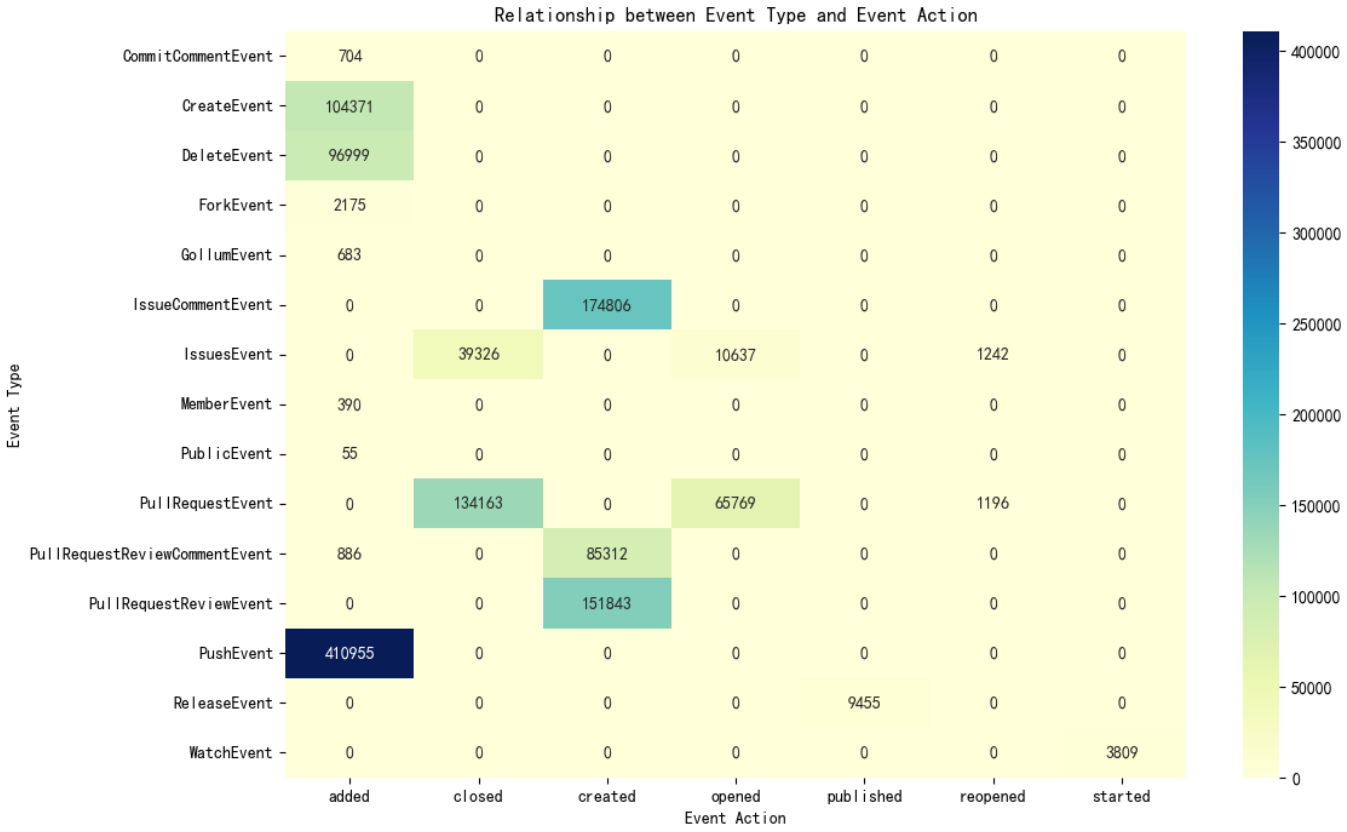


分析：从图中可以看出，总影响力和提交数并不是人们所期待的线性关系，反而是两极分化的形式。根据这张散点图可以将用户分为三类：第一类是普通用户，这类用户的影响力和提交计数都不高，但是用户的数量庞大；第二类是高端用户，这类用户提交次数不算太多，但是质量高，技术力强，拥有很高的总影响力；第三类是平凡用户，他们反复，频繁的提交，但总影响力和普通用户是一个水平，说明提交的技术力不高或者主题冷门，没太多人关注，频繁但平凡。

4. 事件分析

- 事件类型 event_type 与事件行动 event_actiion 交叉分析

  数据:



分析：这张热力图反映事件类型和事件行为有着很高的相关性，pushEvent 类型总是与 added 行动同时出现，issueCommentEvent 类型和 PullRequestReviewEvent 类型总是伴随着 created 行动。

总结

这次实验对 Github 上的用户数据进行数据洞察，还是很有趣的，尤其是对数据之间内藏的联系进行发掘和分析。看似毫无关系的数据中却蕴藏着一些整个社区生态的规律，还是挺令人振奋的。除此之外，还有 AI 工具的利用，极大的加快了我写代码的速度，AI 甚至能预测我下一句要写的代码是什么，直接提示之后和要用的函数和变量，大大的方便我写代码了。

遇到问题和解决方案

1. 根据 country 生成 timezone 时区信息

问 AI 后，AI 直接甩我一个字典 country_to_timezone，里面是国家名到时区信息的映射。好在本次实验的 country 数量只有 52 个，这个字典不算太大，被 AI 暴力解决了（额，好像再多也没问题）。

2. 遇到函数 apply()报错: AttributeError: 'float' object has no attribute 'upper'

完全不知道为什么报错，问了 AI 才知道是数据集中存在缺失值导致的（data['country']中有缺失值）。解决方法：data['country'].fillna('Unknown')。