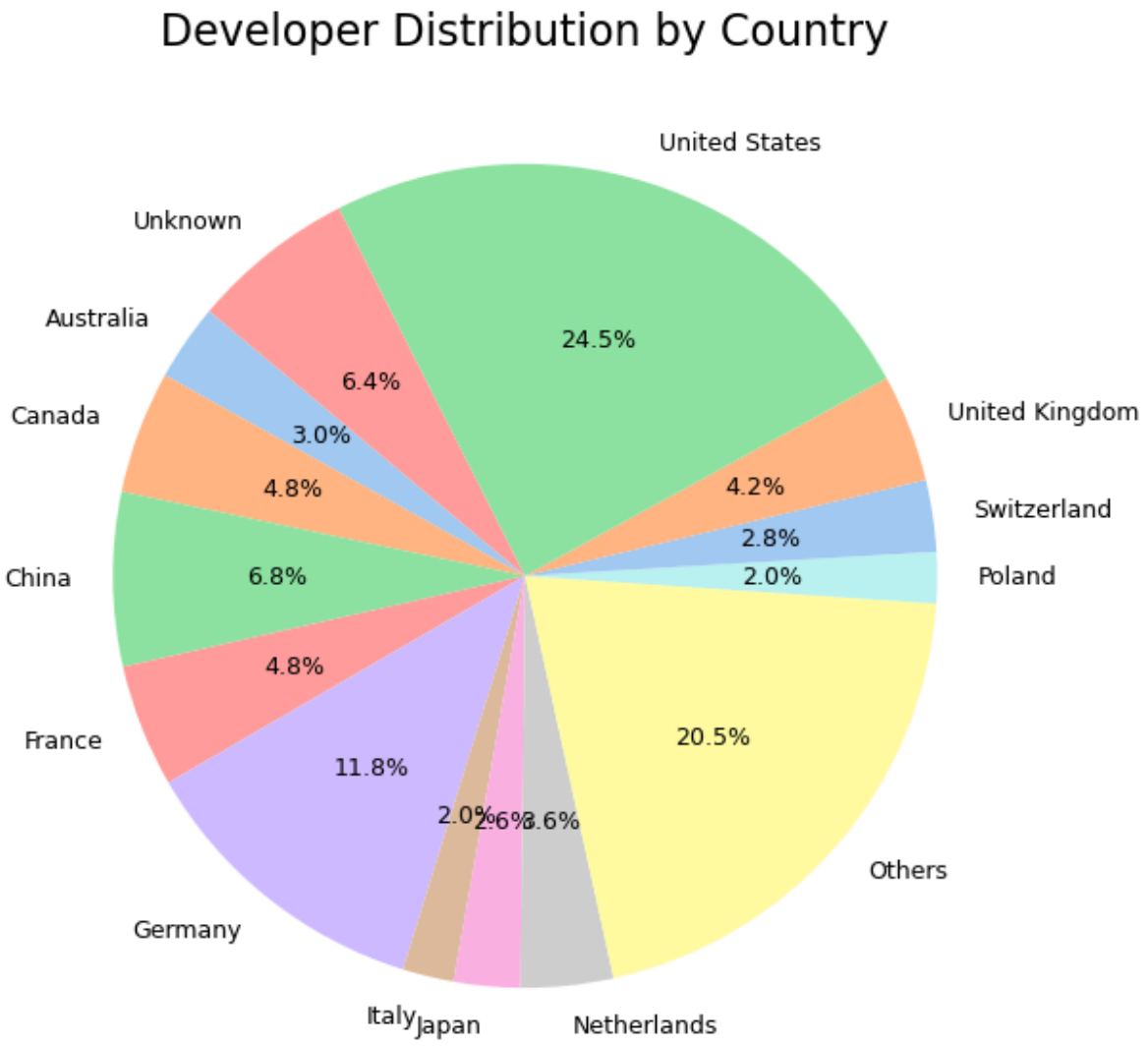
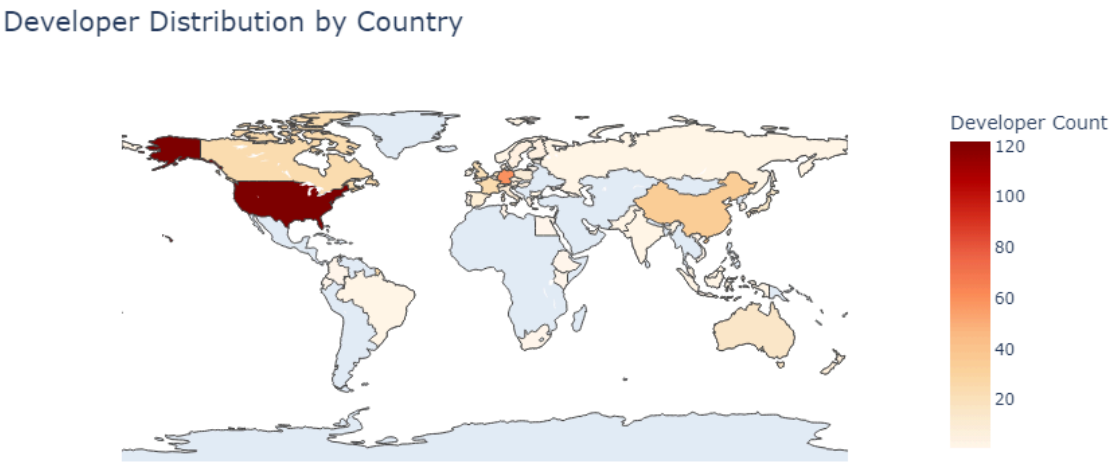
HW12 数据洞察报告

一、人口统计分析

1. 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：

下面给出了饼图和可交互的地图热力图，来描述各个国家的开发者数量。





开发者集中地

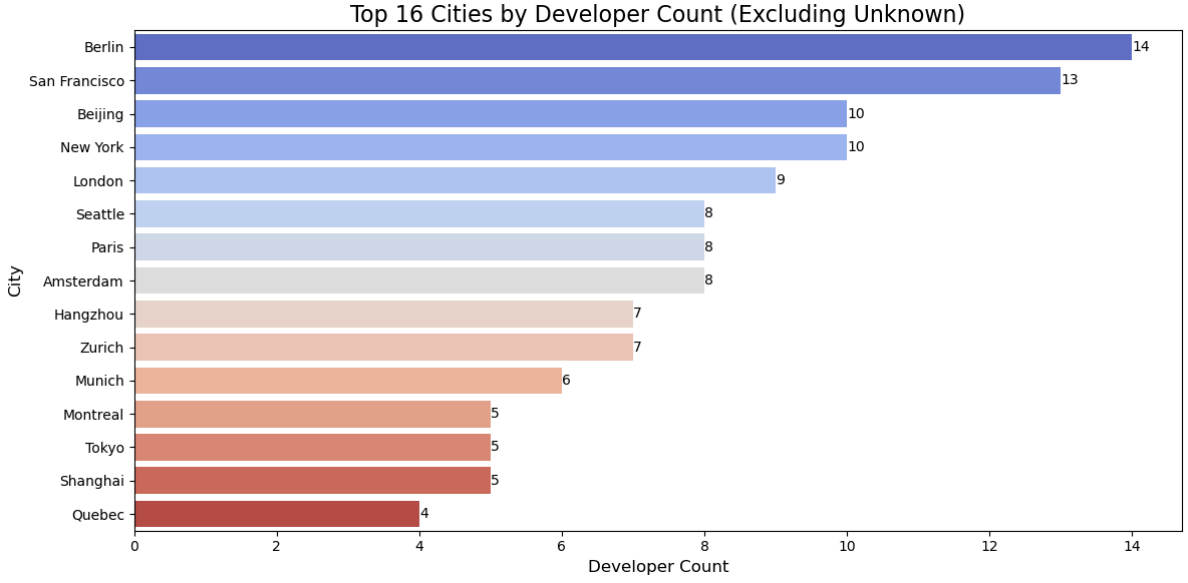
- **美国（United States）**：占比 **24.5%**，是本例数据集中开发者分布的最主要集中地，数据集中接近四分之一的开发者来自美国。这反映了美国作为技术和开源社区中心的重要地位。
- **德国（Germany）**：占比 **11.8%**，是欧洲范围内开发者最为集中的国家。
- **中国（China）**：占比 **6.8%**，是不可忽视的开发者集中地。

主要结论

- 技术和开源开发活动具有明显的区域集中性，美国和欧洲是主要的技术热点区域。
- 中国、澳大利亚和加拿大等国家也有着可观的开发者参与度，说明这些国家在开源社区中的影响力逐渐增加。

2. 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：

下面是TOP16活跃城市的条形图（不包括开发者未设置的未知城市）。



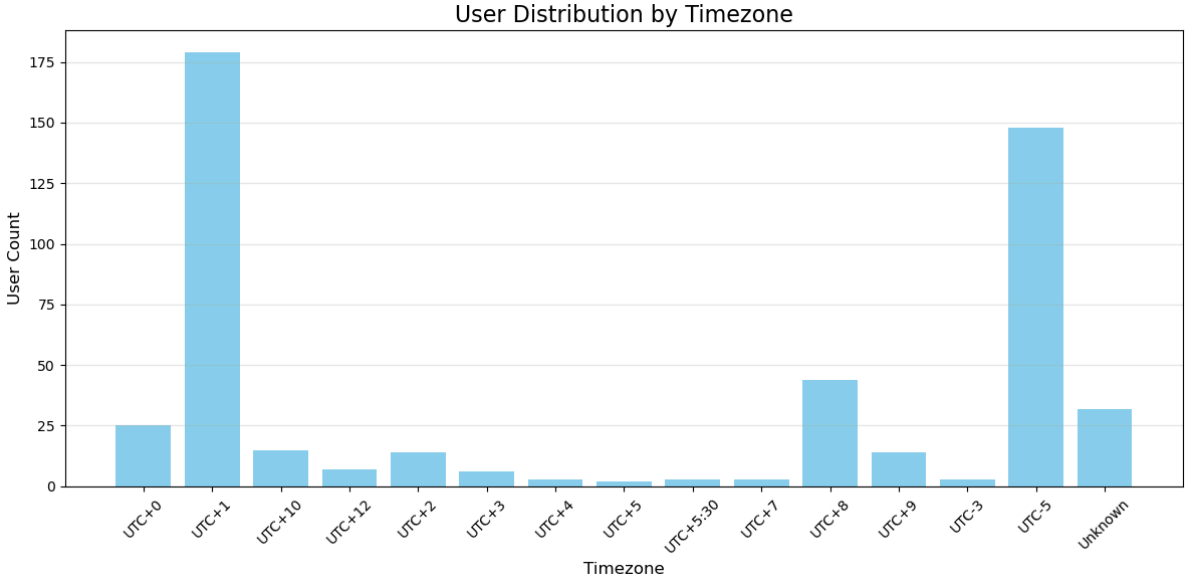
技术热点城市及原因分析

- **Berlin（14 名）**：排名第一，显示了德国首都作为技术和开源活动的重要枢纽，具有较高的开发者密度。
- **San Francisco（13 名）**：旧金山，位列第二，作为硅谷核心城市，是全球技术产业的心脏之一，吸引了大量开发者。
- **Beijing 和 New York（各 10 名）**：
  - **Beijing（北京）**：中国的技术中心，反映了中国科技社区的快速发展。
  - **New York**：全球金融和科技结合的重要城市，展现了其技术影响力。
- **London（9 名）**：欧洲技术热点城市之一，展示了英国在国际开源和技术社区中的活跃度。

次级技术热点及原因分析

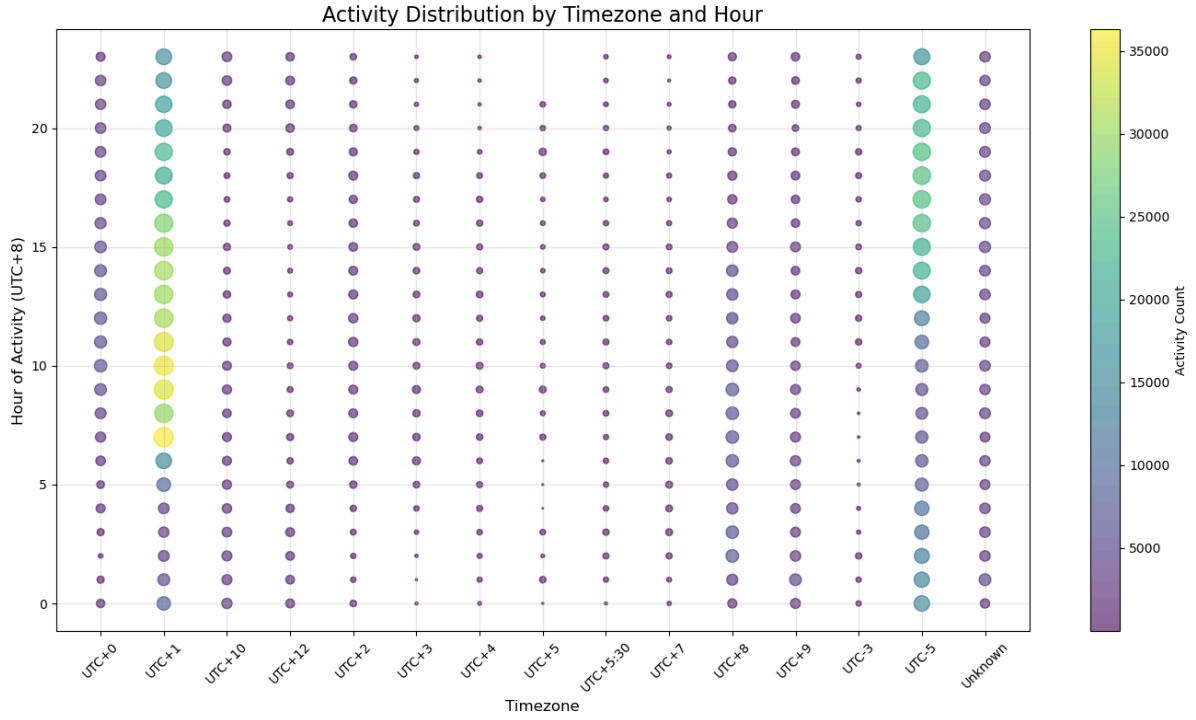
- **Seattle、Paris 和 Amsterdam（各 8 名）**：
  - **Seattle**：微软和亚马逊总部所在地，技术产业的核心城市。
  - **Paris**：法国的技术和开源中心。
  - **Amsterdam**：荷兰技术生态系统的枢纽。
- **Hangzhou 和 Zurich（各 7 名）**：
  - **Hangzhou（杭州）**：阿里巴巴的总部所在地，中国技术产业的重要城市。
  - **Zurich（苏黎世）**：瑞士的研发中心，因其技术人才和国际化社区而著名。
- **Munich（6 名）**：德国的另一个技术城市，展示了德国在技术领域的广泛参与。

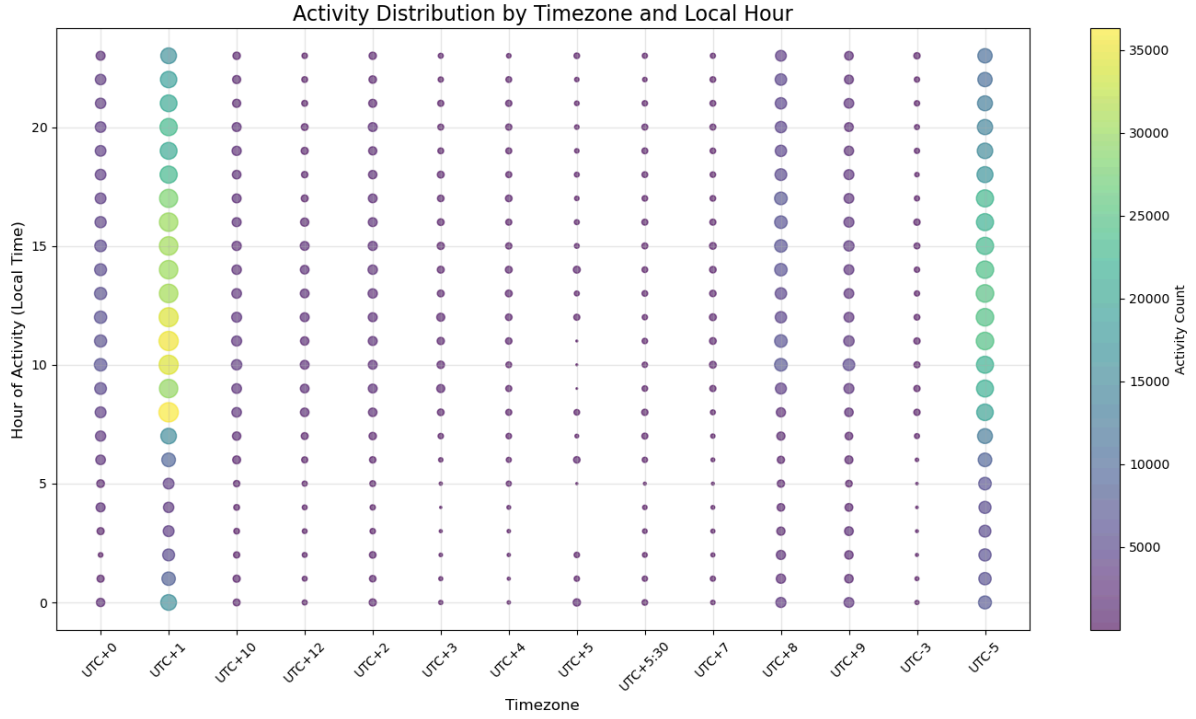
3. 时区分布:了解用户的时区分布，分析不同地区用户的协作时间模式。



集中时区：

- **UTC+1** 即欧洲中部，是用户分布最多的时区，显示出较强的技术社区集中性。
- **UTC-5**（通常包括美国东部时区）同样是用户集中的重要区域。
- **UTC+8** 的人也很多，展现了中国和其他东亚地区的活跃度。





上面两张图分别是纵轴为 北京时间 和 本地时间 的这500个开发者他们工作活动的气泡图。我们主要分析本地时间的图，更能反映他们的生活工作状态。

总体观察

- **9:00-12:00** 和 **14:00-17:00** 是主要的协作高峰时段。
- 这表明大多数开发者的协作活动集中在标准工作时间内，符合典型的工作模式。
- 这一总体观察在三、2中还会有更具体的分析。

不同时区的活动特点

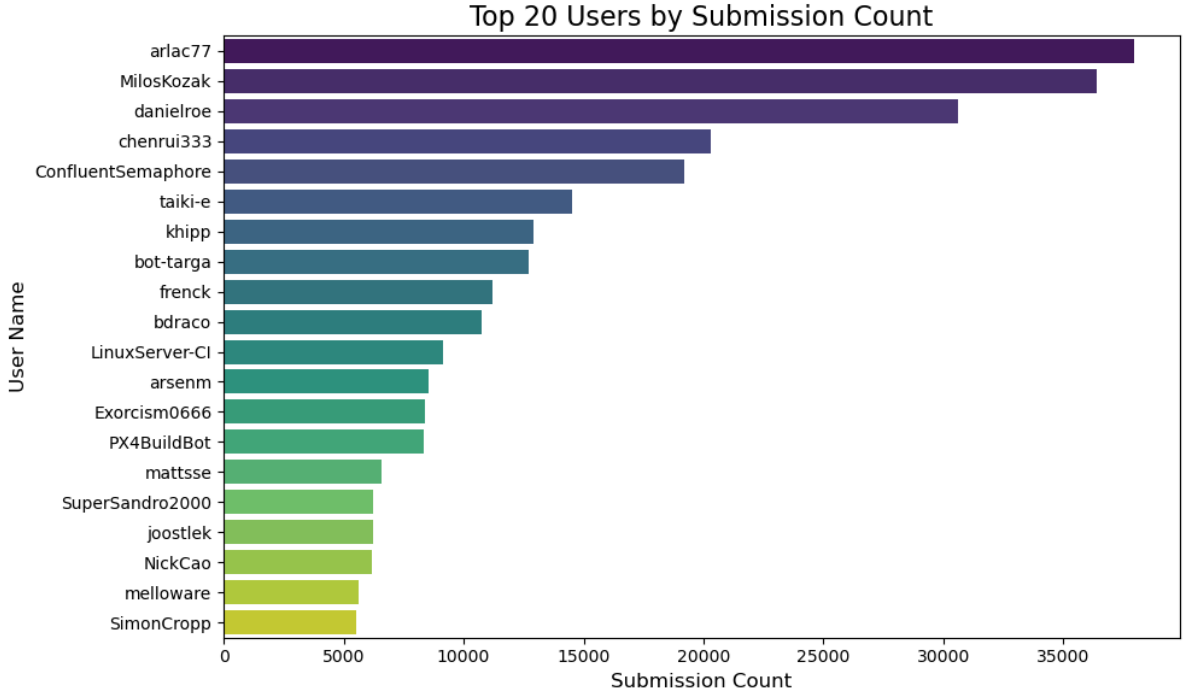
1. **UTC+1（欧洲时区）**：
    - 活跃度最高，尤其集中在工作时间的高峰期（9:00-17:00）。
    - 表现出强烈的工作日协作模式。

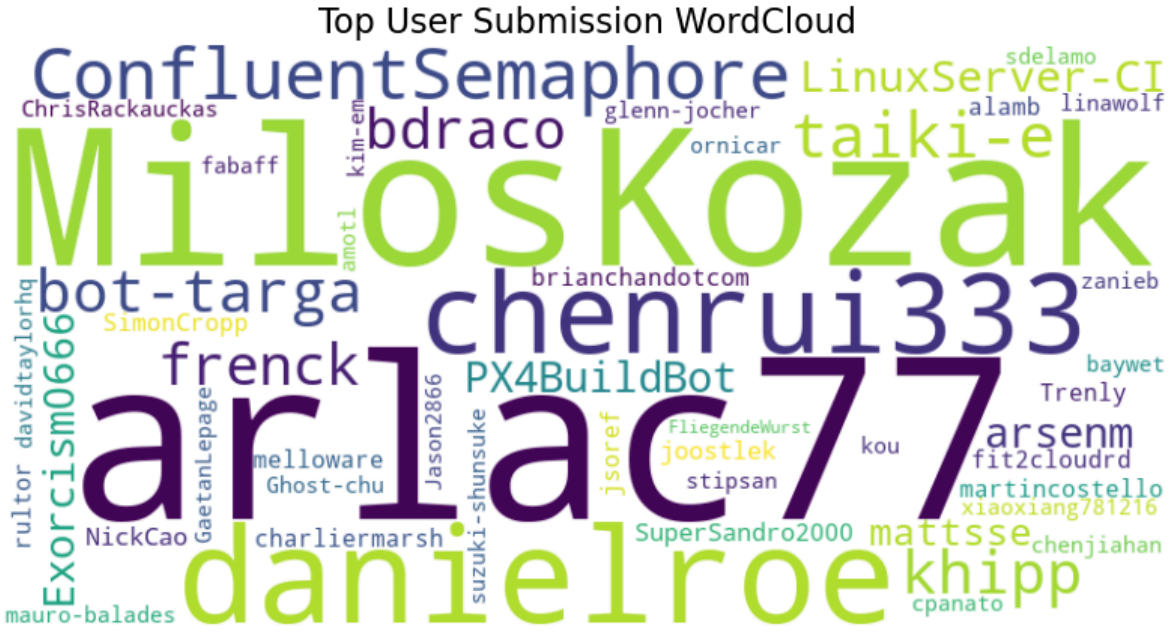
2. **UTC-5（北美时区）**：
    - 活动较为分散，但主要集中在上午和下午，与典型的工作时间一致。
3. **UTC+8（中国）**：
    - 活动集中在上午到下午，表现出类似的工作习惯。
    - 晚间活动21:00-23:00略有增加，可能与加班文化有关。

二、协作行为分析

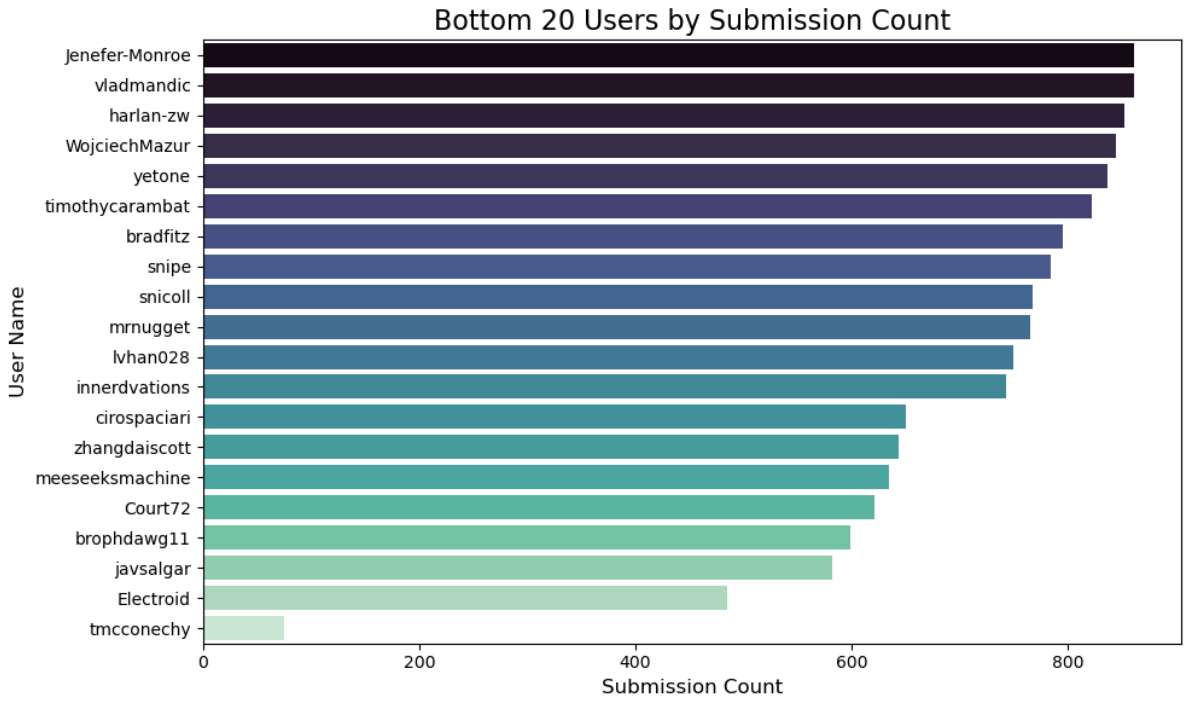
提交频率：统计每个用户的提交次数，识别高活跃用户和低活跃用户。

下面的图给出了高活跃度用户（前20名）的活动条形图以及用户名词云。





下面是地活跃度用户（倒数20名）的活动条形图。

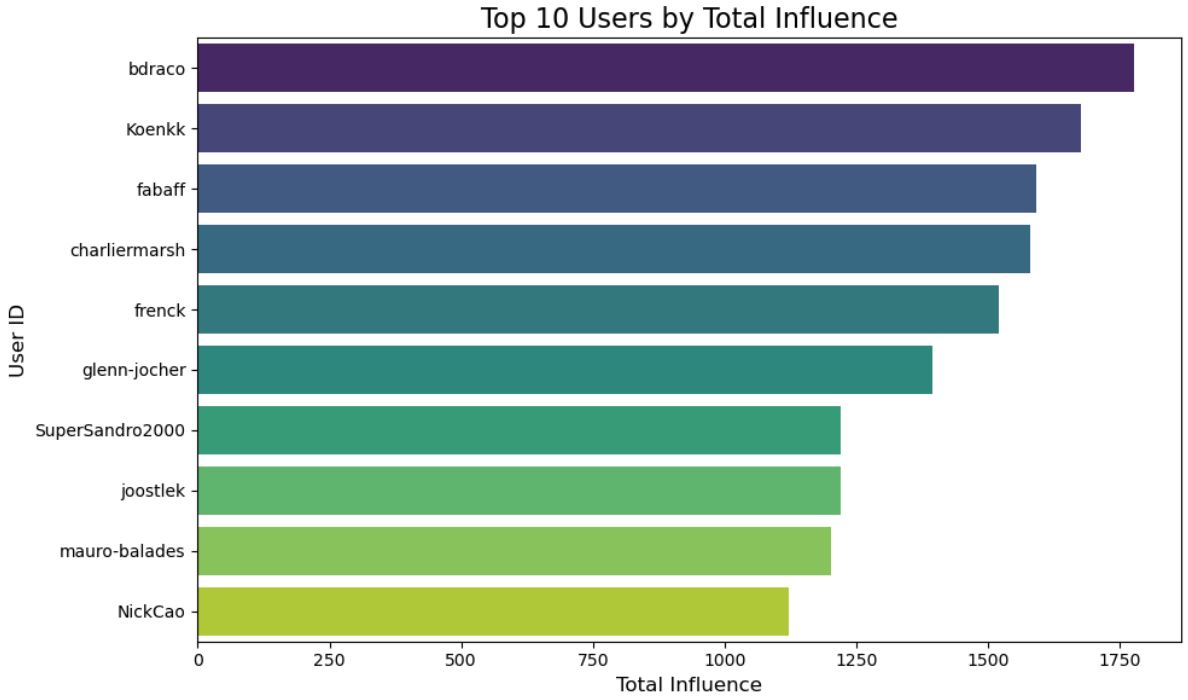


三、其他维度有趣的观察

1. 用户影响力榜单

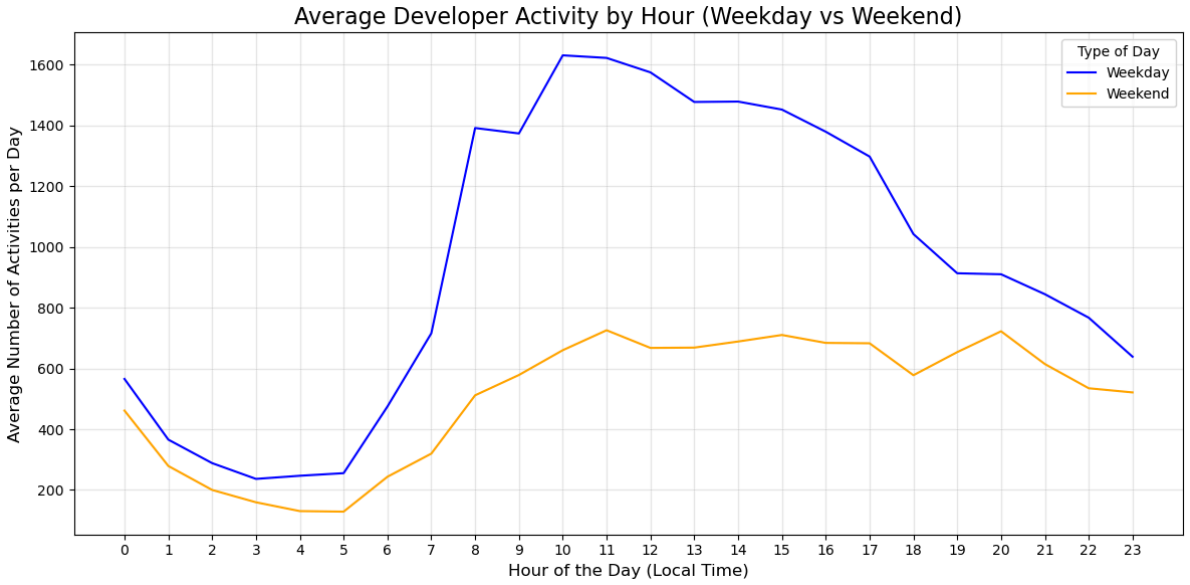
我们发现数据集中有一个指标为影响力，于是我们给出了TOP10最具影响力的开发者以及他们所在的地区。

	name	total_influence	location
0	bdraco	1776.967163	Houston, TX
1	Koenkk	1674.805908	The Netherlands, Helmond
2	fabaff	1590.152954	Switzerland
3	charliermarsh	1580.198242	Brooklyn, NY
4	frenck	1520.352173	Enschede, The Netherlands
5	glenn-jocher	1392.865723	Spain ⇄ California
6	SuperSandro2000	1220.031738	Germany
7	joostlek	1219.017944	Utrecht, The Netherlands
8	mauro-balades	1201.465759	5 centimeters from the screen
9	NickCao	1120.395699	Boston



2. 工作日与周末的日均活动数目与活动时间对比

我们想关注一下工作日与周末的日均活动数目与活动时间，来得到这些开发者们工作日和周末开发项目的情况的区别。我们给出一个折线图。注意，数据集里所有活动时间是基于UTC+8的，因此我们将活动时间转换为本地时间，更有利于我们观察开发者的活动情况。

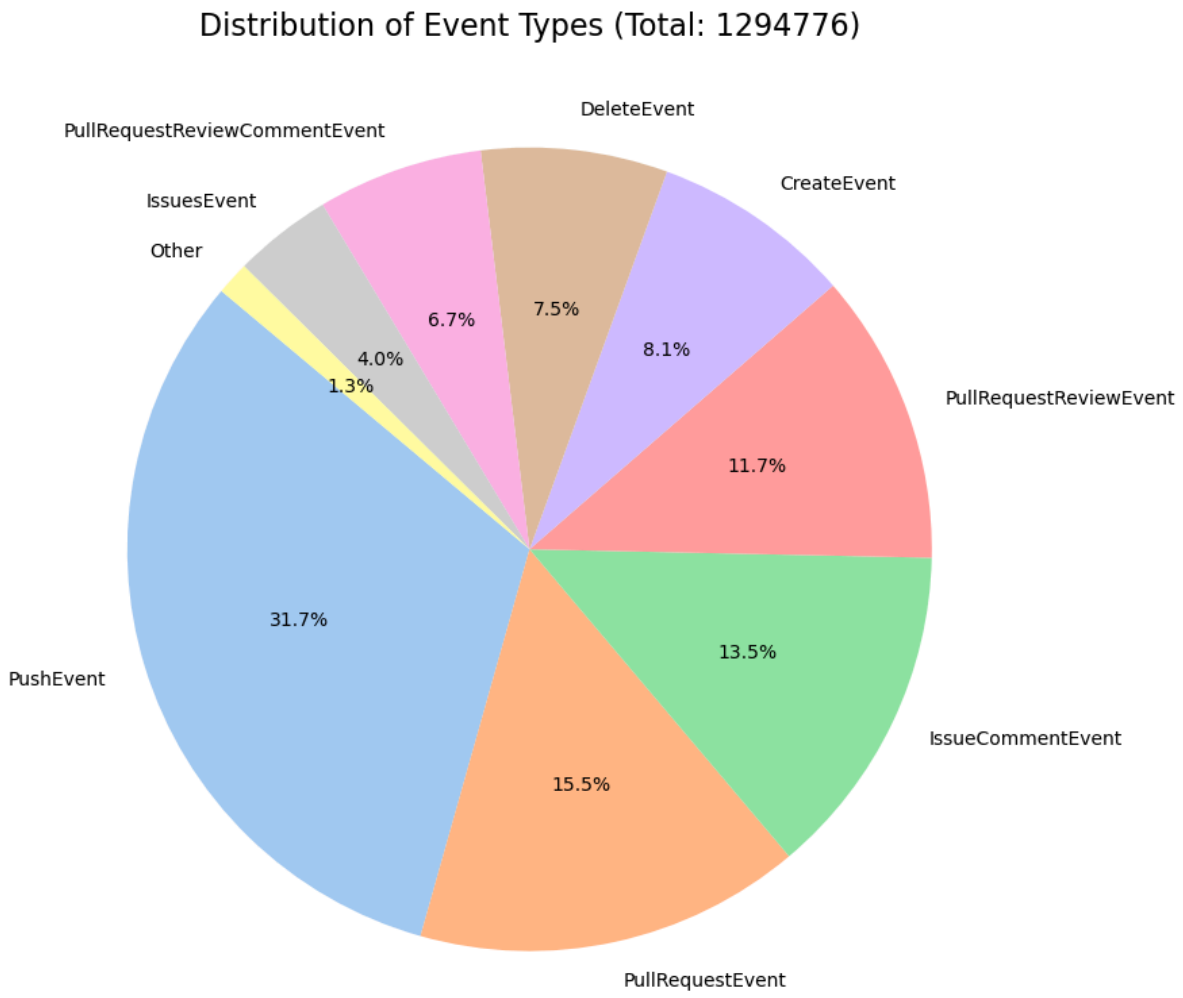


由此可见，

- 无论是周末还是工作日，开发者的主要活动高峰都在8:00-18:00。工作日时，到了晚上 18 点之后，活动量显著减少，表明开发者在晚上更倾向于休息或减少技术活动；而"熬夜加班"现象在周末更普遍。
- 周末的活动较为分散，高峰时间段较宽泛，反映了开发者在周末处理开源项目十分灵活。
- 工作日的活动总量显著高于周末，说明开发者的主要贡献仍集中在工作日，开发者们在工作日会进行休息；但是在周末，日均活动数目也并没有很低，说明开发者们在周末仍然没有对开发活动产生松懈。

3. 分析event_type字段中事件种类的分布

绘制关于event_type字段中事件种类的饼状图：



占比最高：PushEvent

- `PushEvent` 是最常见的事件类型，占比 **31.7%**，几乎是其他单一事件类型的两倍。
- 意义
  - `PushEvent` 表示代码被推送到仓库，这表明代码提交是协作活动的核心。
  - 这符合开源项目的特点，代码提交是开发者参与项目的主要方式。

其次是PR 相关事件

- `PullRequestEvent`（15.5%）和 `PullRequestReviewEvent`（11.7%）
  - `PullRequestEvent` 占比第二，仅次于 `PushEvent`，表明开发者在开源协作中积极使用 `Pull Request` 模式，进行代码合并和功能开发。
  - `PullRequestReviewEvent` 的较高占比说明代码审查是协作的重要组成部分。开发者不仅提交代码，还积极参与代码质量的提升。

再其次是Issue 相关事件

- `IssueCommentEvent`（13.5%）和 `IssuesEvent`（4.0%）
  - `IssueCommentEvent` 的占比接近 `PullRequestEvent`，说明评论和讨论是协作的重要环节。
  - `IssuesEvent`的相对较低占比说明，创建issue的频率低于评论和解决问题的频率。因此，得到一个结论，开发者的协作模式更倾向于对已有问题进行讨论，而不是频繁创建新问题。