Homework12报告（第十五、十六周任务）

2023级　吕欣萌　10222140402　12.17

实验目标

- 培养数据处理与分析能力：通过实际操作，提升对大规模数据集的处理和分析能力。
- 掌握GPT工具的应用：学习如何利用GPT大型模型工具辅助完成数据洞察任务。
- 理解数据隐私与伦理：在处理包含个人信息的数据时，遵循数据隐私保护的原则和规范。

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。

3.**其他维度有趣的洞察**（至少2个）

实验过程

一、人口统计分析

1.CSV文件分析处理

（1）CSV文件头部包含了以下字段：

user_id：用户ID，唯一标识用户。

name：用户的名称。

location：用户所在的地理位置（城市、区域等）。

total_influence：用户的总影响力，可能是指用户的贡献或活跃度。

country：用户所在的国家。

event_type：事件类型（例如：活动、贡献等）。

event_action：事件的具体操作（例如：提交代码、发表评论等）。**event_time**：事件发生的时间戳。

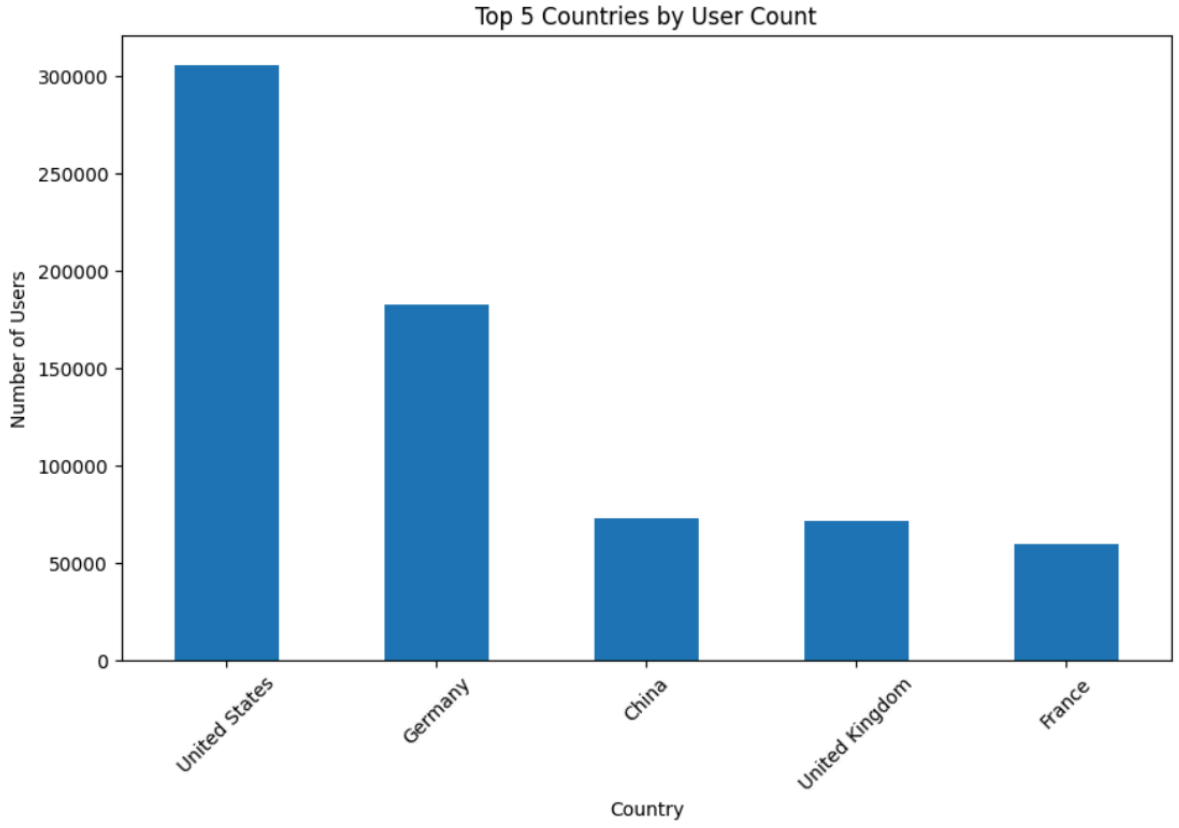
（2）数据预处理

处理缺失值：对具有缺失值的数据行进行去除处理。

处理异常值：影响力不应该为负值，可以过滤掉。

2.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。

```
country
United States     305788
Germany           182659
China              73011
United Kingdom     71606
France             59570
Name: count, dtype: int64
```



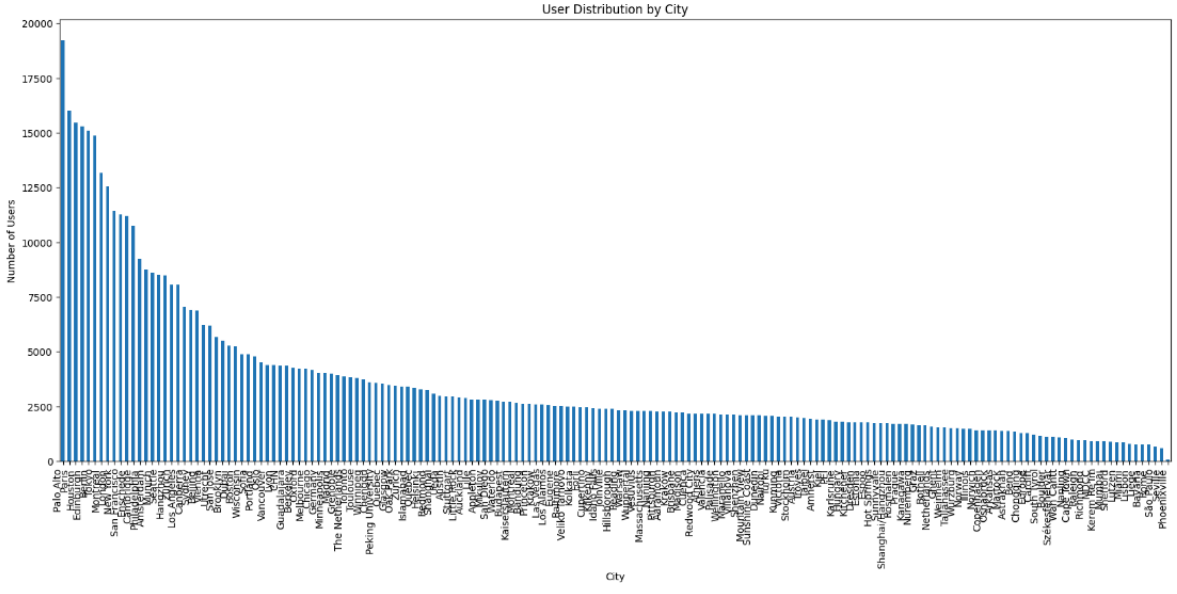
结果分析：

1.United States 拥有最多的用户数量，说明该地区的开发者群体相对较大，可能是主要的技术开发市场。

2.Germany 和 **China** 紧随其后，可能是技术发展较为成熟或是开发者活跃的地区。

3.United Kingdom 和 **France** 用户数量稍少，但仍占有一定份额，显示出这些国家的开发者活跃度。

3.城市级别分布分析:分析主要城市的开发者密度，发现技术热点区域。



```
前五个最多用户的城市：
city_or_country
Palo Alto     19215
Paris         16021
Houston       15449
Edinburgh     15308
Berlin        15095
Name: count, dtype: int64
```

结果分析：

1.Palo Alto 拥有最多的用户数量，说明该城市的开发者群体相对较大，可能是主要的技术开发市场。

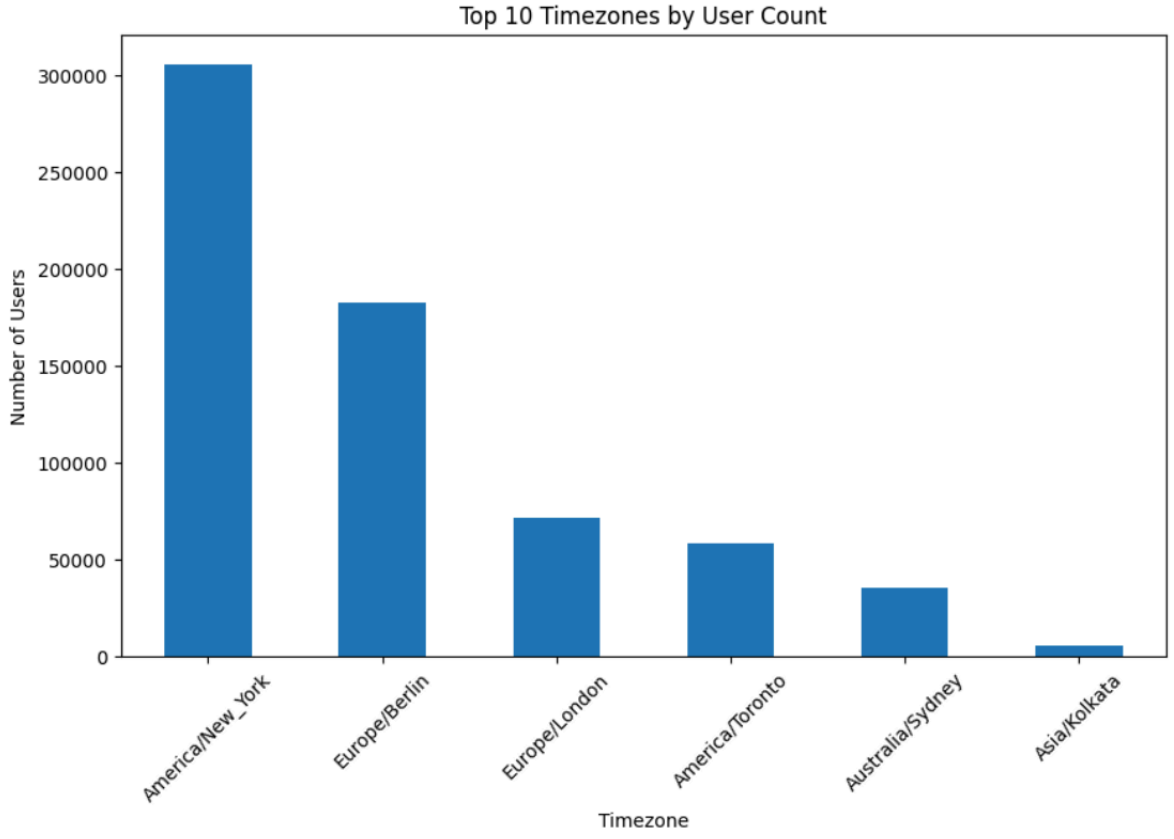
2.Paris 和 **Houston** 紧随其后，这两个城市高概率是技术发展较为成熟或是开发者活跃的地区。

3.Edinburgh 和 **Berlin** 用户数量稍少，但仍占有一定份额，显示出这些城市的开发者活跃度。

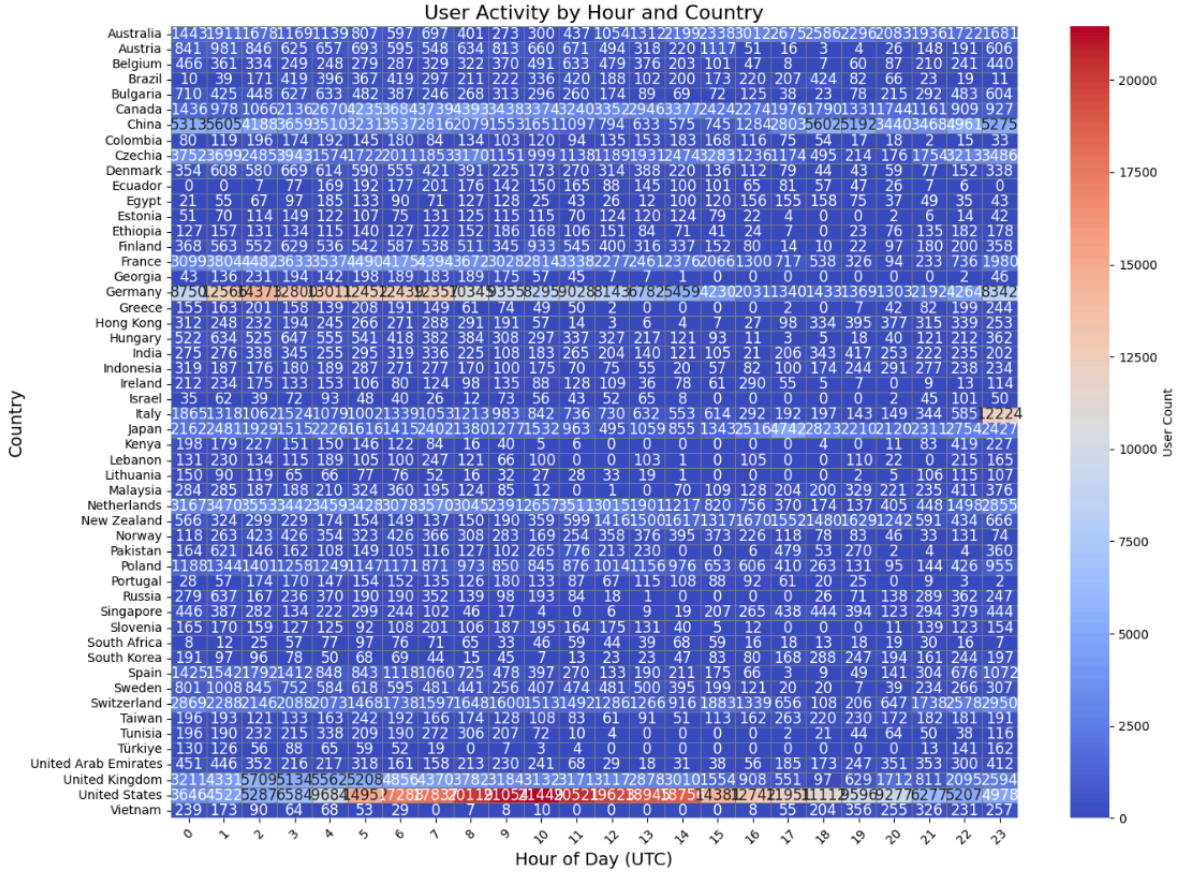
4.时区分布：了解用户的时区分布，分析不同地区用户的协作时间模式。

（1）时区分布最多的几个城市如下图：

```
timezone
America/New_York     305788
Europe/Berlin        182659
Europe/London         71606
America/Toronto       58600
Australia/Sydney      35746
Asia/Kolkata           5689
Name: count, dtype: int64
```



（2）不同地区用户的协作时间模式如下图：

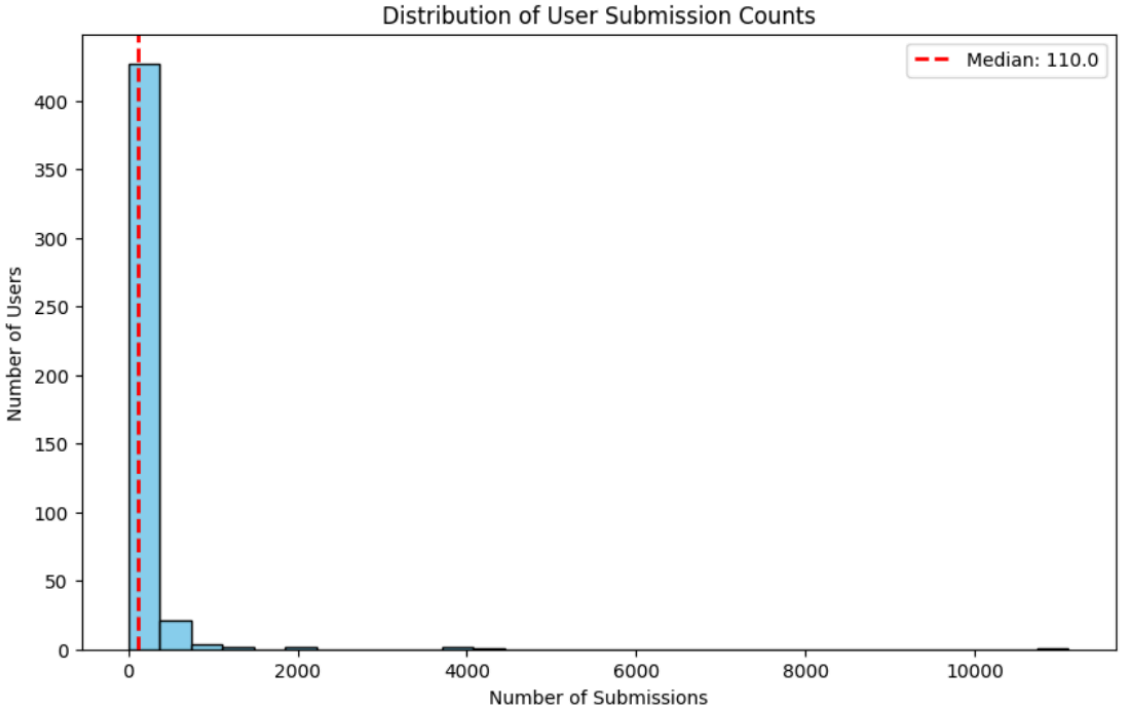



```
Top 5 countries with the most active users:
country
United States     305788
Germany           182659
China              73011
United Kingdom     71606
France             59570
Name: count, dtype: int64
```

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。

```
Top 5 most active users based on submission count:
user_id
40306929    11111
43724913     4328
50149701     4033
158862       3963
2119212      2208
dtype: int64
Bottom 5 least active users based on submission count:
user_id
6702118     1
6225961     1
814283      1
62625502    2
1541747     3
dtype: int64
Number of high-active users: 229
Number of low-active users: 231
```



三、其他维度有趣的洞察（至少2个）

1. 影响力与活动的关系分析

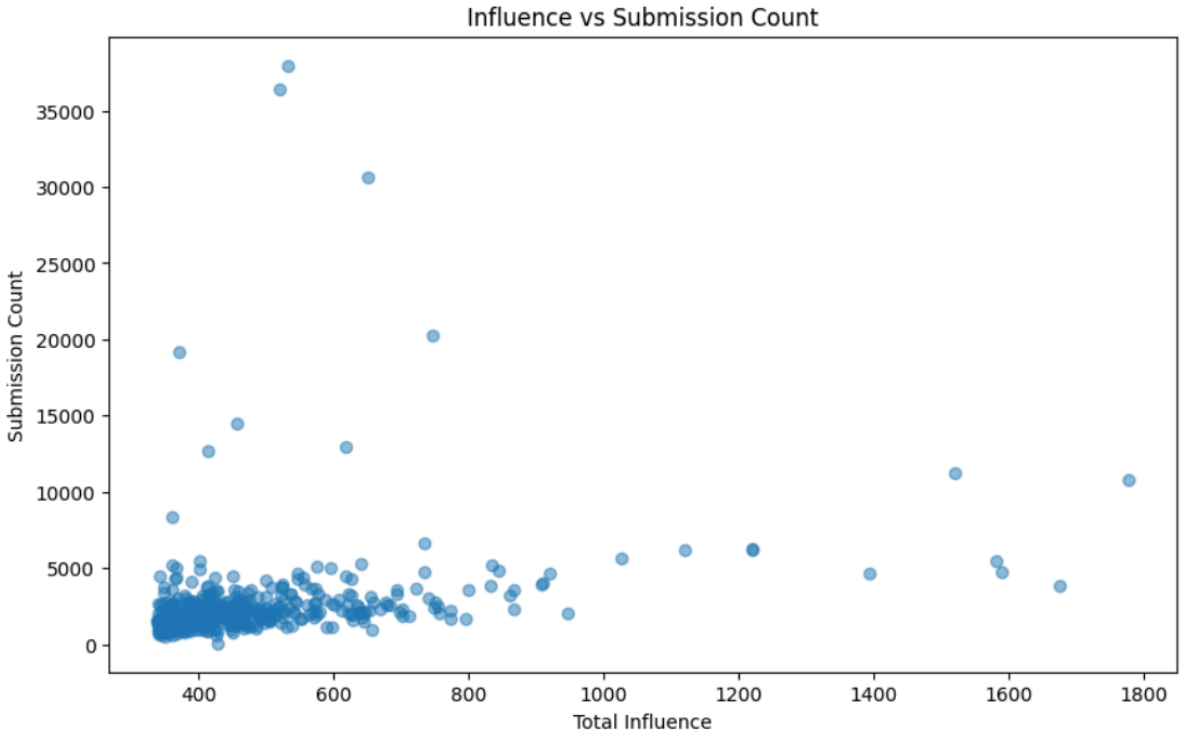
目标：分析用户的`total_influence`与他们的活动频率（即提交次数）之间的关系。

洞察：

- **高影响力用户是否更活跃？** 可能发现具有较高影响力的用户倾向于更频繁地参与活动。这些用户可能是平台的早期采用者、专家或活跃的内容创造者。
- **低影响力但高频繁活动的用户** 也可能为平台贡献大量数据或事件。通过分析这一群体，可以识别出潜在的“隐形贡献者”。

分析：

- **散点图**：通过展示`total_influence`与`submission_count`的关系，可以看出高影响力用户是否更频繁地参与。
- **相关系数**：通过计算相关系数，能定量说明这两个变量之间的关系。如果相关性较强，表明高影响力用户更倾向于高频参与。




```
Correlation between total influence and submission count: 0.26
```

2. 活动模式与时间分布分析

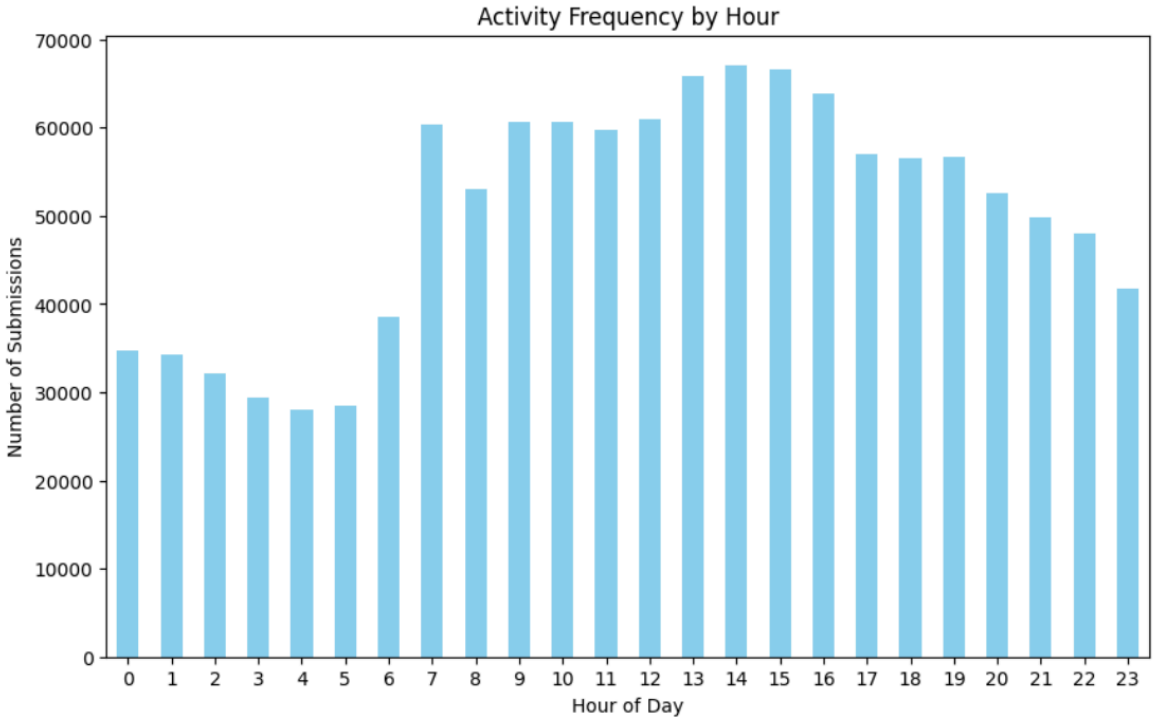
目标：分析用户提交事件的时间模式，查看不同时间段（小时、天、周）的活动情况，找出活跃的时间窗口。

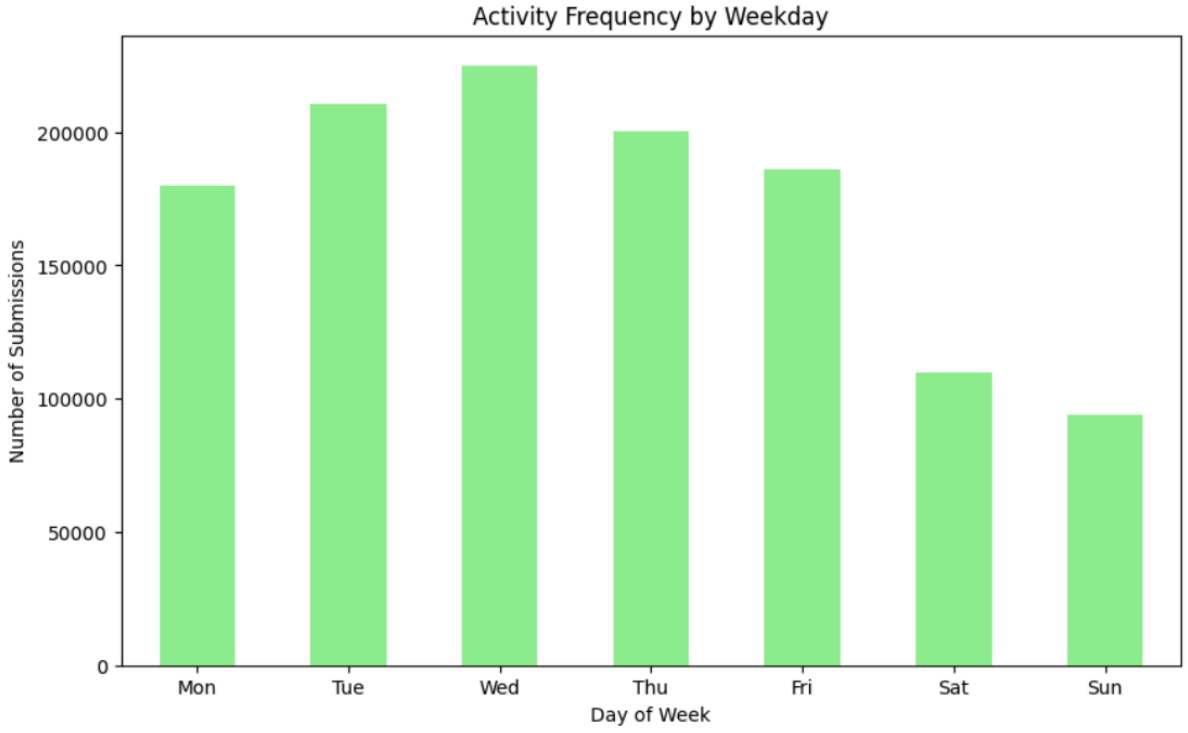
洞察：

- **哪些时段用户最活跃？** 通过分析提交事件的`event_time`列，查看用户在一天中的哪几个小时提交活动最频繁。这可以帮助理解用户的活跃时间，便于根据活跃期进行系统的维护或促销活动。
- **周内活跃模式**：查看不同星期几的提交模式，帮助识别是否有特定的工作日/休息日用户行为变化。

分析：

- **按小时统计**：通过绘制`hourly_activity`图，可以识别出用户在一天中的活跃高峰期（例如，早晨、午后或晚上）。这种信息有助于优化推送通知、系统维护等。
- **按星期几统计**：`weekday_activity`可以揭示一周内的活跃模式，是否有特定的工作日或周末活动频率较高。这对评估平台的使用趋势和规划促销活动有很大帮助。





3. 地区与活动类型分析

目标：分析用户的地理位置（如`country`）与他们进行的活动类型之间的关系，观察某些地区是否特定偏好某种类型的活动。

洞察：

- **地区偏好**：不同地区的用户可能偏好不同类型的活动。例如，某些国家的用户可能偏好创建事件（`CreateEvent`），而另一些国家的用户可能更偏向于添加信息（`added`）。
- **跨地区协作行为差异**：可以分析哪些地区的用户具有更多的跨地区协作行为，从而识别活跃的全球开发者群体。

分析：

- **热力图**：通过热力图查看每个国家/地区的不同活动类型的提交频率。这样可以识别出特定地区的用户偏好，以及全球范围内活动的差异。