实验报告

1. 概述

本报告基于 users_combined_info_500.csv 数据集，通过分析用户的活跃度、影响力、地理分布和时间模式，提供对用户行为的洞察。数据集包括以下字段：user_id, name, location, total_influence, country, event_type, event_action, event_time。

2. 数据预处理

1）缺失值填充：将 country 和 location 列中的缺失值填充为 'unknown'。

2）时间转换：将 event_time 列转换为 UTC 时间，并根据 country 列中的国家名称转换为本地时间。

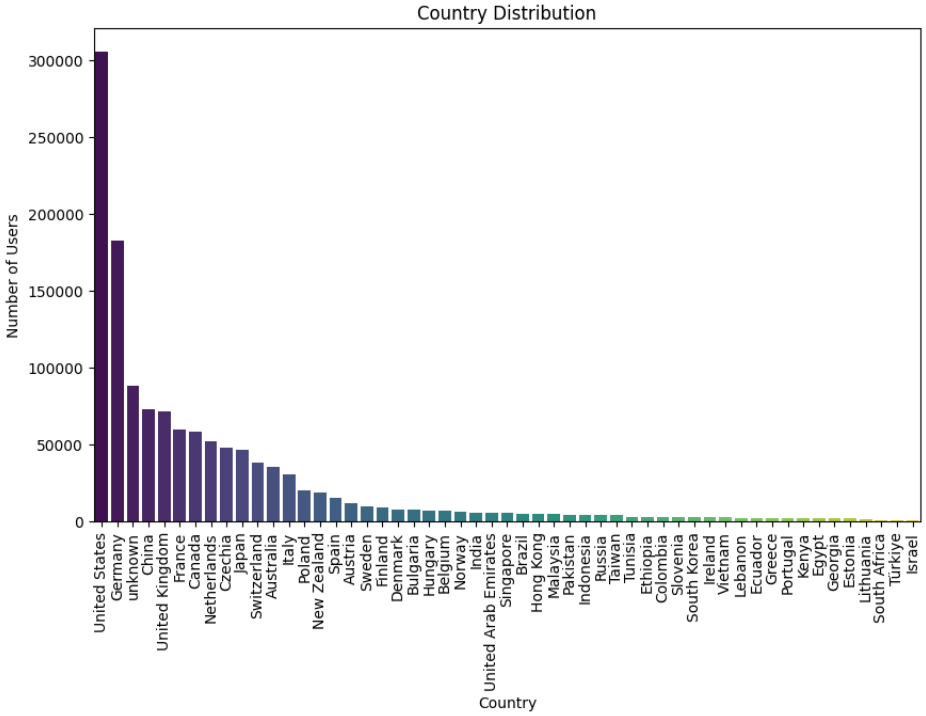
3. 分析结果

1） 国家和地区分布

通过分析用户的国家分布，我们发现用户主要集中在几个国家，如下图所示：

United States	305788
Germany	182659
unknown	88151
China	73011
United Kingdom	71606
France	59570
Canada	58600
Netherlands	52367
Czechia	48122
Japan	46553
Switzerland	38093
Australia	35746
Italy	30671
Poland	20002
New Zealand	18444
Spain	14939
Austria	11758
Sweden	9851
Finland	8815
Denmark	7412
Bulgaria	7357
Hungary	7080
Belgium	6628
Norway	6004
India	5689
United Arab Emirates	5264
Singapore	5205
Brazil	5022
Hong Kong	4767
Malaysia	4538

我们可以看到，美国、英国、德国等国家的用户数量最多。以下是具体的分布图：



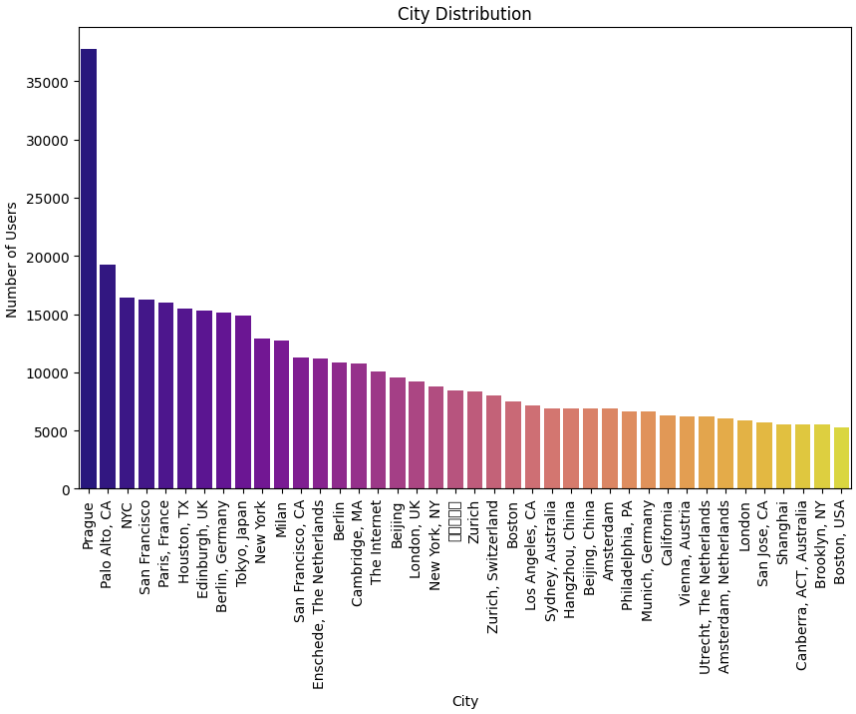
2） 地区级别分布

我们过滤掉 location 列中仅包含国家名称的行数据，只保留包含更具体地区名称的行数据。以下是城市级别分布的前 40 个地区：

Prague	37757
Palo Alto, CA	19215
NYC	16381
San Francisco	16271
Paris, France	16021
Houston, TX	15449
Edinburgh, UK	15308
Berlin, Germany	15095
Tokyo, Japan	14877
New York	12893
Milan	12704
San Francisco, CA	11271
Enschede, The Netherlands	11218
Berlin	10883
Cambridge, MA	10740
The Internet	10060
Beijing	9591
London, UK	9180
New York, NY	8803
きさらぎ駅	8401
Zurich	8346
Zurich, Switzerland	8053
Boston	7459
Los Angeles, CA	7126
Sydney, Australia	6925
Hangzhou, China	6901
Beijing, China	6873
Amsterdam	6855

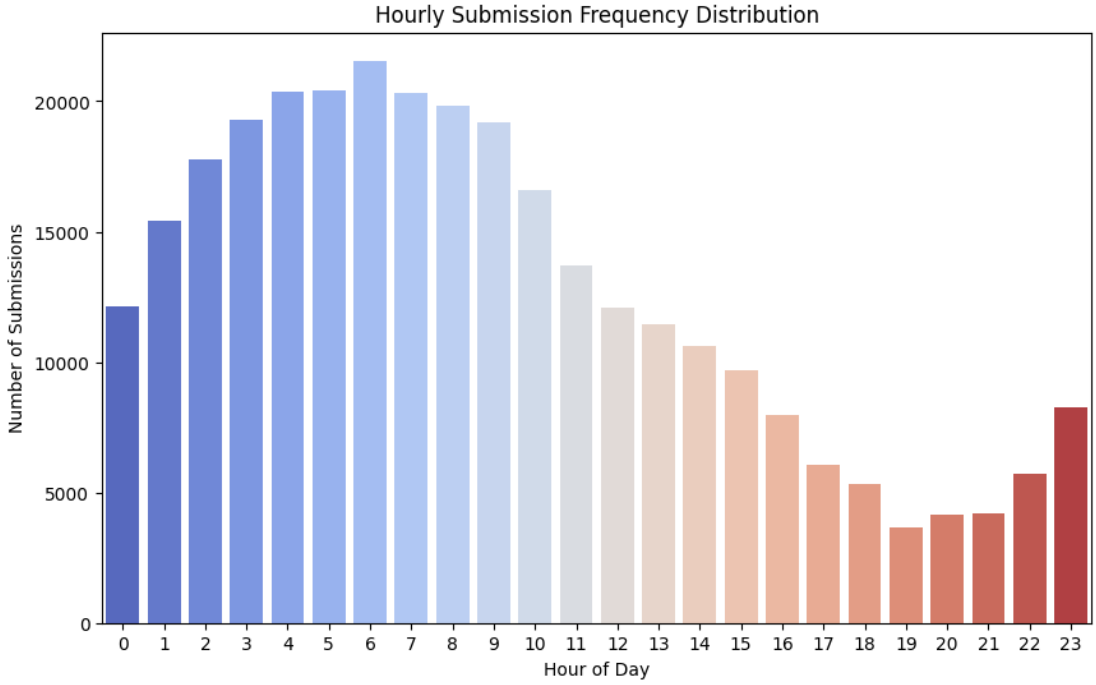
Amsterdam	6855
Philadelphia, PA	6609
Munich, Germany	6599
California	6314
Vienna, Austria	6231
Utrecht, The Netherlands	6209
Amsterdam, Netherlands	6037
London	5854
San Jose, CA	5678
Shanghai	5558
Canberra, ACT, Australia	5505
Brooklyn, NY	5494
Boston, USA	5231

以下是具体的城市分布图:



3） 每小时的用户提交次数分布

用户在一天中的提交次数分布如下:

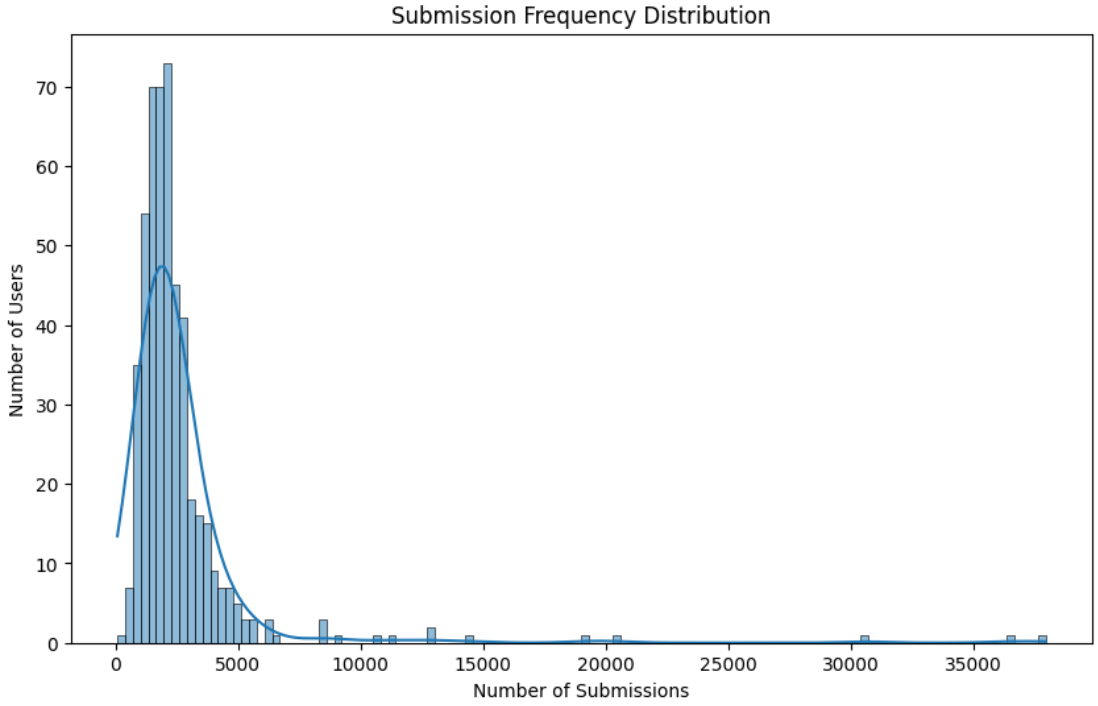


4） 提交频率

不同用户的提交次数分布情况（最多和最少）:

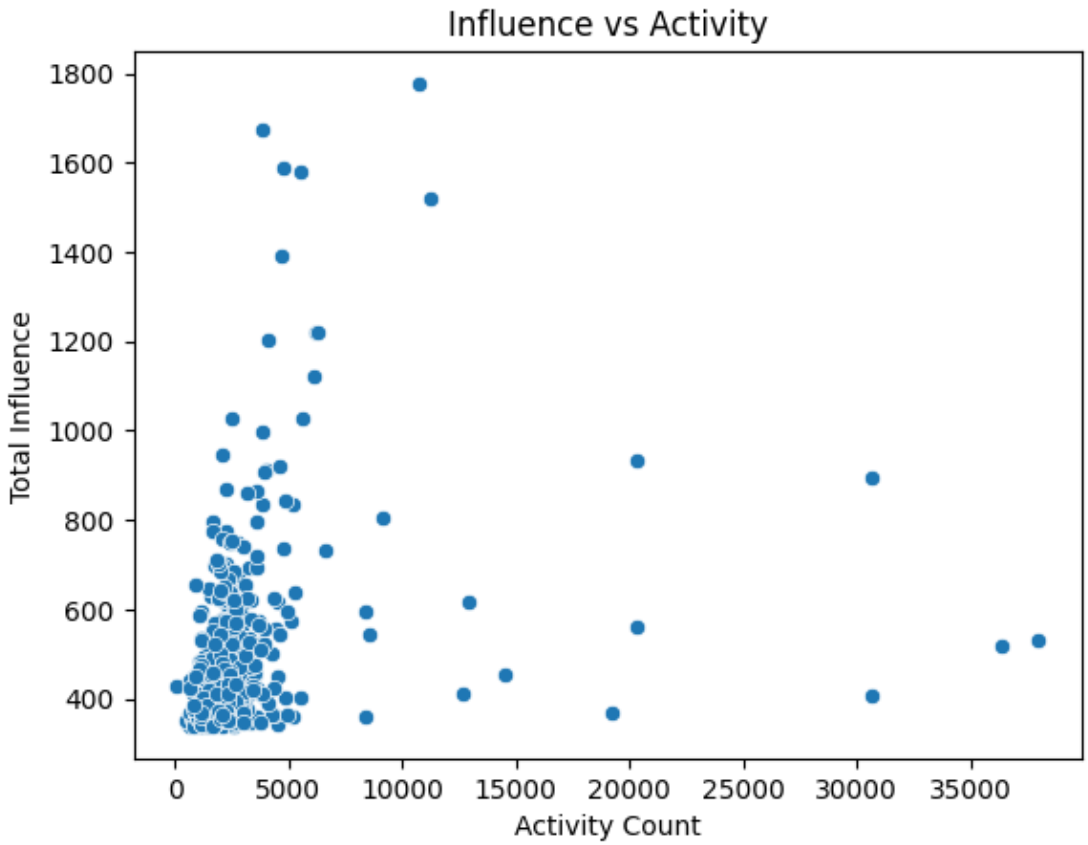
```
arlac77                37960
MilosKozak             36400
danielroe              30616
chenrui333             20300
ConfluentSemaphore     19215
                         ...
Court72                  621
brophdawg11              599
javsalgar                582
Electroid                485
tmcconechy                75
```

整体分布情况:



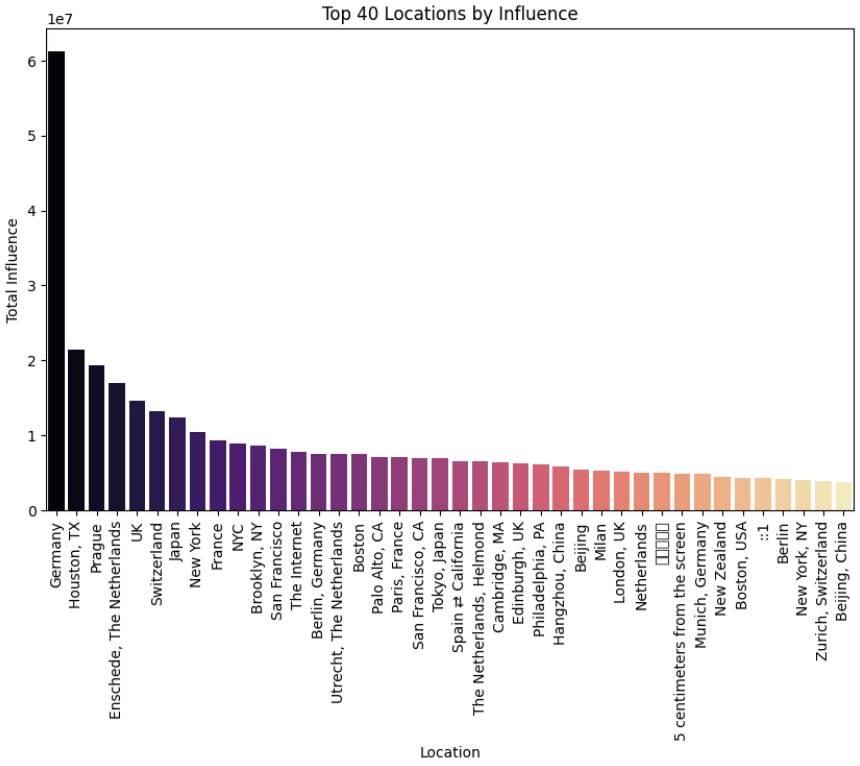
5） 影响力与活跃度的关系

```
Influence and Activity Count Correlation: 0.2711753721894405
```



6） 按地点的影响力分布

以下是具体的影响力分布图（前 40 名）：



4. 结论

1. **用户主要集中在几个特定国家和城市**，例如美国、英国和德国。
2. **用户在一天中的某些时段更为活跃**，特别是在工作时间和晚间。
3. **影响力与活跃度之间存在一定关系**，但不完全一致，有些高影响力用户的活跃度较低。
4. **不同地点的用户影响力存在显著差异**，部分城市或国家的用户影响力更为突出。