Homework_12 报告

10235501461 柯宇

一. 实验目的

1. 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。
2. 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。
3. 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。
4. 通过不同的维度和视角对同一组数据集进行分析，统计用户提交数据情况，并得出聚类、分类等结果。
5. 通过可视化图像，对所得出的结果进行可视化表达，得出可视化图像，以便进行进一步的展示（例如搭建网站展示、大屏展示等）。
6. 通过 python 内部自带的程序生成简单的报告，也可作为简易展示的一个亮点。

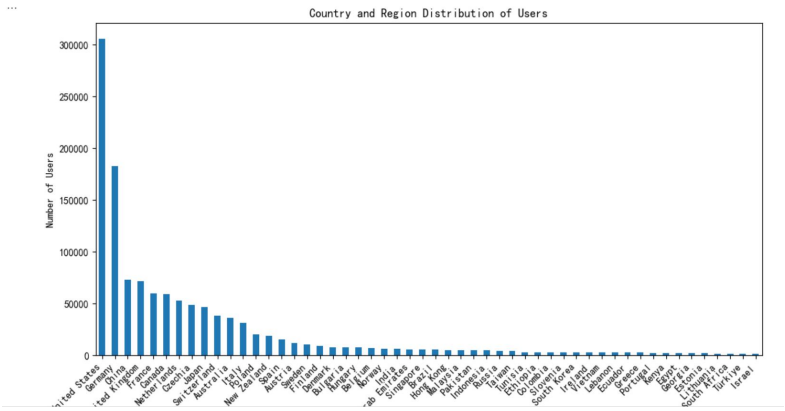
二. 问题解答

1. 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。

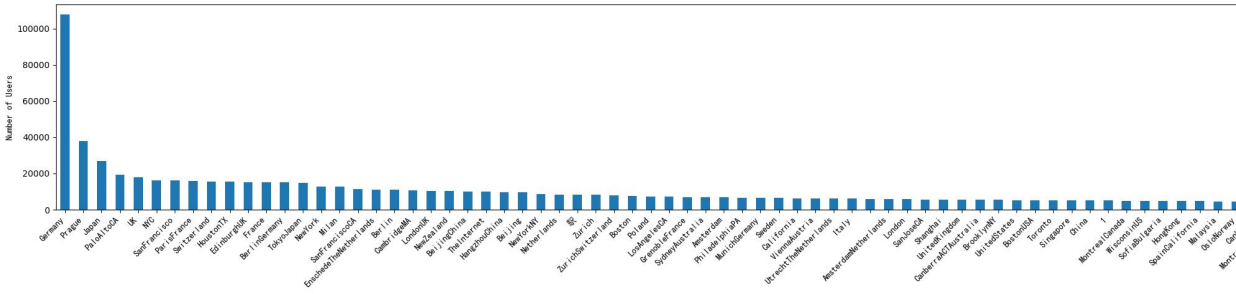
通过柱状图进行展示。此处国家名称均为英文名，无需进行数据的剔除，注意与下面的问题进行区分。

以下为柱状图：



通过地图热力图进行展示。能够直观展示全球地区分布情况。

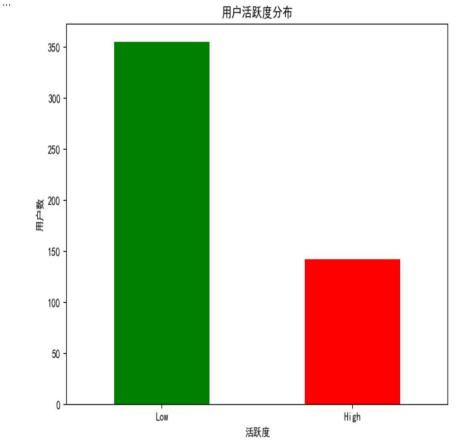
(2) 城市级别分布：分析主要城市的开发者密度，发现技术热点区域。
其中需要注意对部分城市名的难以识别的拉丁字母数据进行剔除
采用柱状图进行展示：（因为城市分布较多，地区较小）



(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。

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。
包含柱状图对比以及大约前 500 的各类用户代表输出



高活跃用户：

	user_id	submission_count	activity_level
0	225	2885	High
6	13564	3140	High
8	26967	3214	High
9	27350	4509	High
10	32321	3284	High
..	...	...	...
484	79828097	2615	High
488	86073083	2637	High
489	88161975	3394	High
492	91018726	4098	High
495	100913391	3177	High

[142 rows x 3 columns]

低活跃用户：

	user_id	submission_count	activity_level
1	1945	1526	Low
2	2621	796	Low
3	4196	1983	Low
4	9582	2258	Low
5	10682	1703	Low
..	...	...	...
490	88724353	2301	Low
491	89584709	862	Low
493	92015510	1866	Low
494	95597335	2113	Low
496	112826355	1680	Low

3. 其它的数据洞察的探索（总共有 7 个方向，一并展示，部分字较小，可进入代码中获得清晰展示）

(1) 用户活跃度分析：每日活跃用户数

(2) 地域和国家分析：按地区和国家分组的活跃度

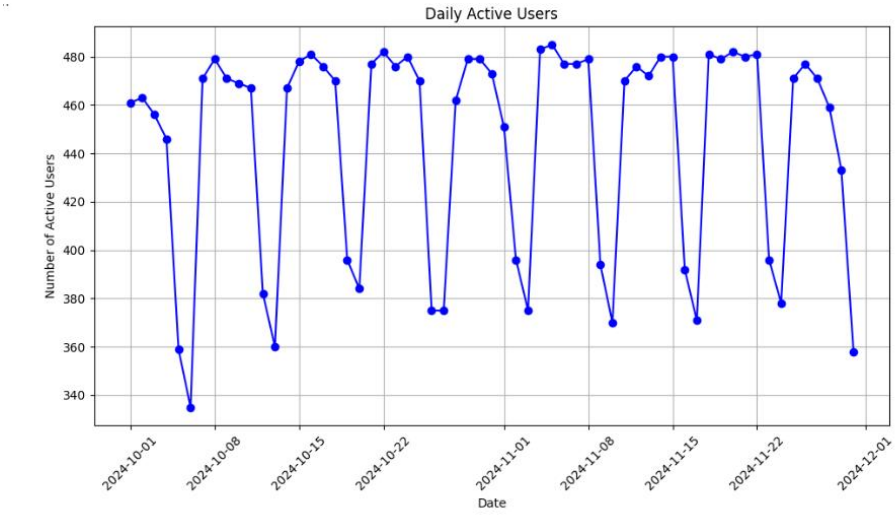
(3) 事件行为分析：事件类型分布

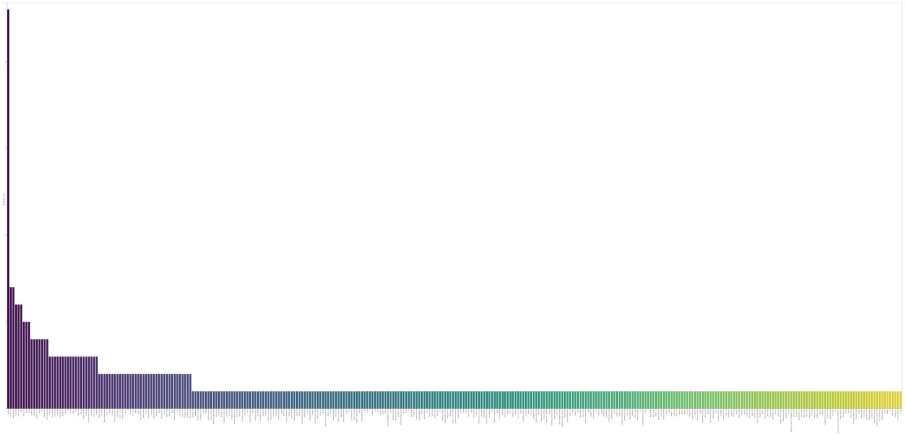
(4) 用户影响力分析：高影响力用户

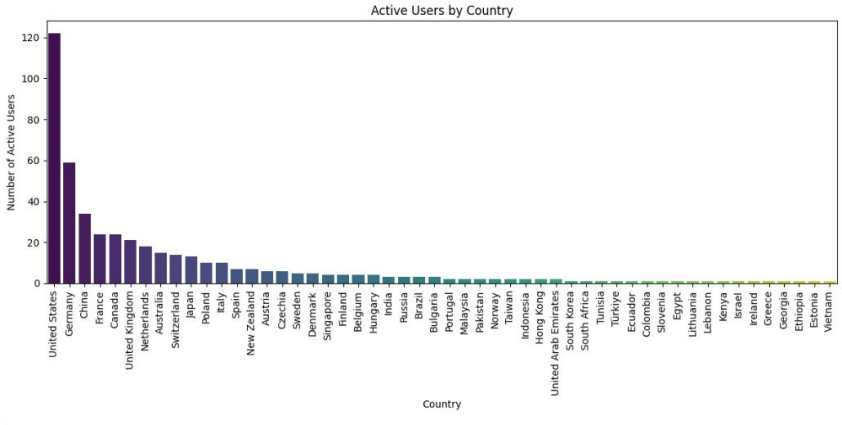
(5) 时间趋势分析：小时活跃用户数

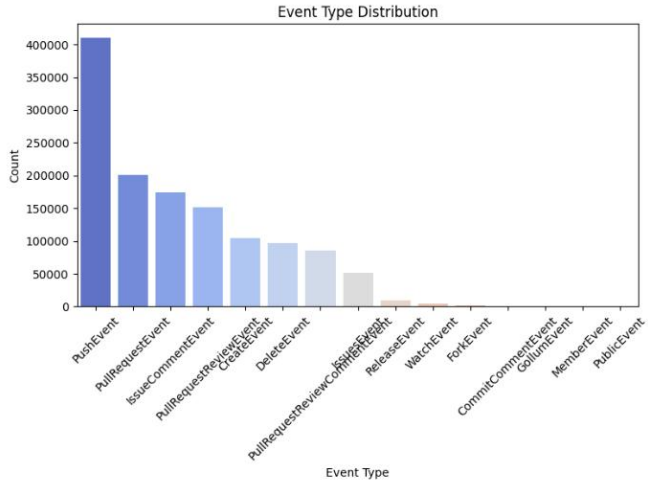
(6) 事件频次与时间间隔：事件时间间隔分布

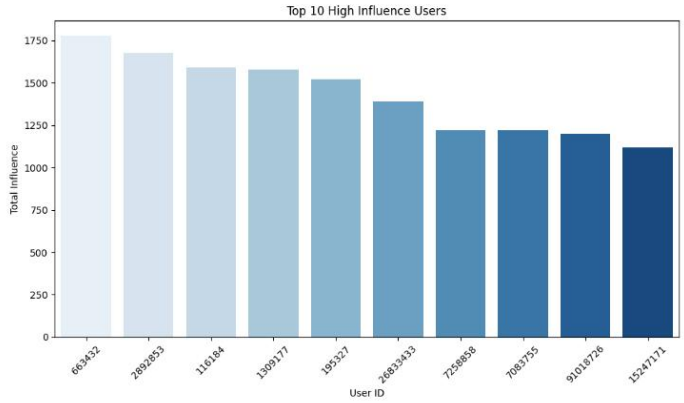
(7) 用户行为聚类分析（简单示例：基于影响力和事件类型）

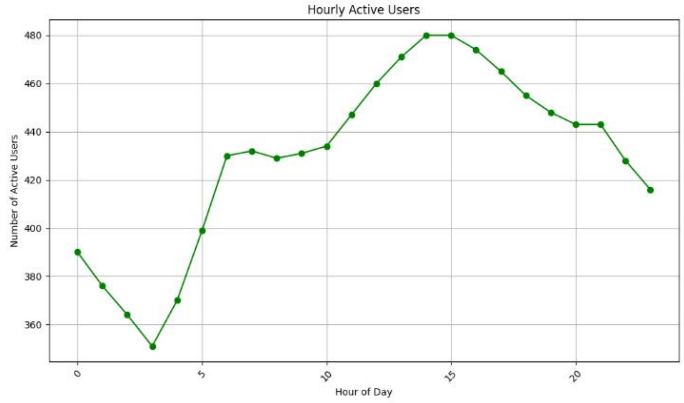


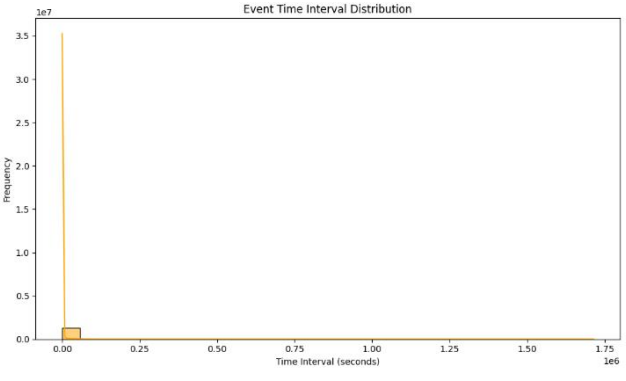


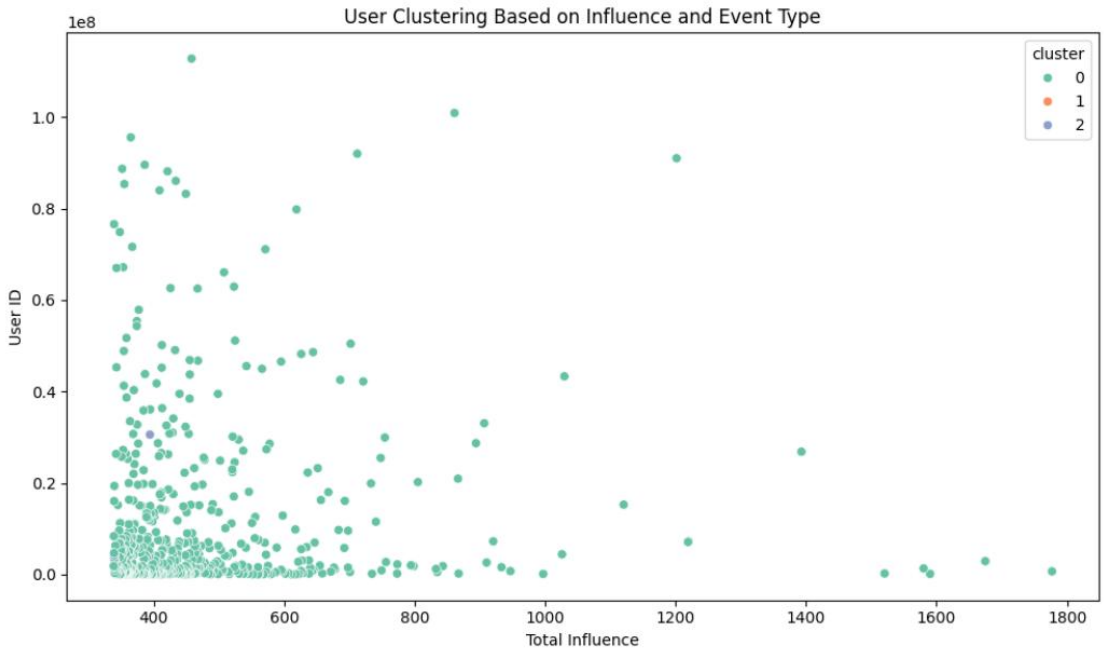


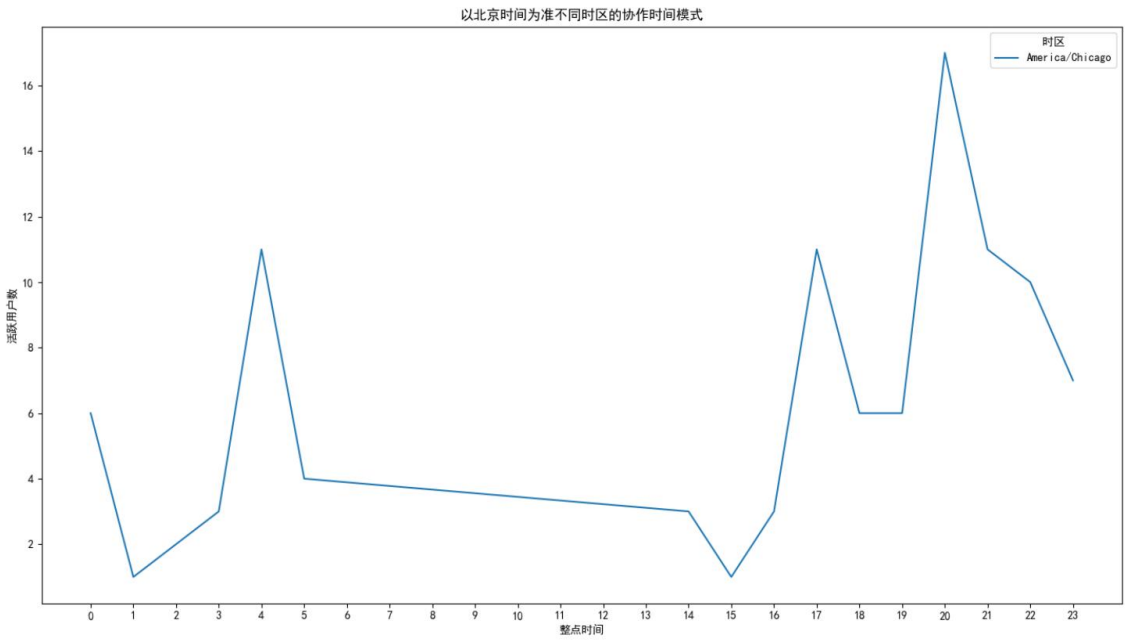

Event Type Distribution
Count
Event Type
PushEvent
PullRequestEvent
IssueCommentEvent
PullRequestReviewEvent
CreateEvent
DeleteEvent
ReleaseEvent
WatchEvent
ForkEvent
MemberEvent
PublicEvent


Top 10 High Influence Users
Total Influence
User ID


Hourly Active Users
Number of Active Users
Hour of Day


Event Time Interval Distribution
Frequency
Time Interval (seconds)





4. python 内置报告

```
概述:
                                count   unique                       top    freq  \
user_id                     1294776.0      NaN                       NaN     NaN
name                          1294776      497                   arlac77   37960
location                      1294776      344                   Germany  107747
total_influence             1294776.0      NaN                       NaN     NaN
country                       1206625       52             United States  305788
event_type                    1294776       15                 PushEvent  410955
event_action                  1294776        7                     added  617218
event_time                    1294776  1006997  2024-10-30 02:12:11+08:00      73

```

三．洞察与结论（以给定数据为样本进行分析，用局部估量整体）

1．通过观察热力地图，可以观察到区域间使用 github 提交代码的用户分布的差距。在部分欠发达地区，用户的分布量是极少的；而在中国、美国、俄罗斯、欧洲等发达的国家和地区，用户分布密度是很高的。那么可以推得，一个地区计算机事业和开源事业的发展和当地的经济发展是离不开的；作为上层建筑，要以经济基础为根基。

2．从客观上说，美国、荷兰、德国等老牌资本主义国家的开源事业的发展还是比较领先的，有着庞大的用户数量。因此开源的活动应当在中国得到进一步的推广和宣传，增强中国在计算机领域、开源社区的话语权。

3．根据开源协作的时区分布趋势，从北京时间的标准来衡量，大部分开源协作的时间是在晚上。因此通过开源社区 PR、Merge 等协作方式仍然是世界各国计算机工程师、数据科学家、软件工程师的重要协作方式。

4．但是从用户活跃度区分可以看出，开源社区内的活动用户的状况仍然呈现两极分化趋势，即较少部分的活跃用户贡献了较多的活动和代码，而大部分的人可能仍然处于学习或者划水的阶段。需要号召大部分的学习者、爱好者也都参与进开源社区活动当中来。

5．开源社区的活动活跃性大部分还是集中在周中工作日的时候，证明很多工作者是在工作中运用开源社区，将其作为工作的工具；而周末仍然有活跃度，证明许多人仍然是保持着工作或者学习的势头，互联网与计算机仍然是一个热潮，深得很多人的喜爱和兴趣。

6．在开源社区，占比比较大的仍然是 pushevent，证明大家仍然是乐于为开源社区做贡献的，乐于参与合作的。

7．根据聚类图，仍然是能够支持以上活动以 push 为主、大部分人影响力较小、专业性人才较少的观点的，而且能够更加直观清晰。（但值得注意的是，在图像中部分活动被埋没了）。

四．反思

1．对于中文或者其他拉丁字母的适应性仍然是一个比较重要的问题，吐出来经常是乱码，要有一定的解决储备，或者直接用英文。

2．虽然部分代码用人工智能生成，但是仍然需要自己通过一定的代码常识进行查验和勘误，要根据文件里面有的列进行处理。

3．在大型数据处理需要很长时间的情况下，还是先用一定量的样本数据进行代码测试，可以节省调试的时间。

4．对于时区的问题，需要连接模组，但其不是特别适应大型的数据规模。其实可以用字典更新到本地，但通用性、普适性我认为不是很够，所以用了模组。

5．对于数据的 NaN 值、空格、以及会导致 pandas 误判的数学符号，需要在数据处理前进行清洗、剔除，在本次实验中做得不是很好，耗费了许多时间。

6．本次实验中对于敏感信息、个人信息数据的处理并没有特别的要求，但数据的脱敏是十分重要的。例如 ID 最好不要和用户名同时出现在数据集里面，更不能抓取密码、家庭住址等关键信息，要注意对于数据隐私和安全的保护。

————————————报告完毕————————————

报告纯手打承诺。

备注：**BronsonLau** 为本人英文名。