Homework12 report

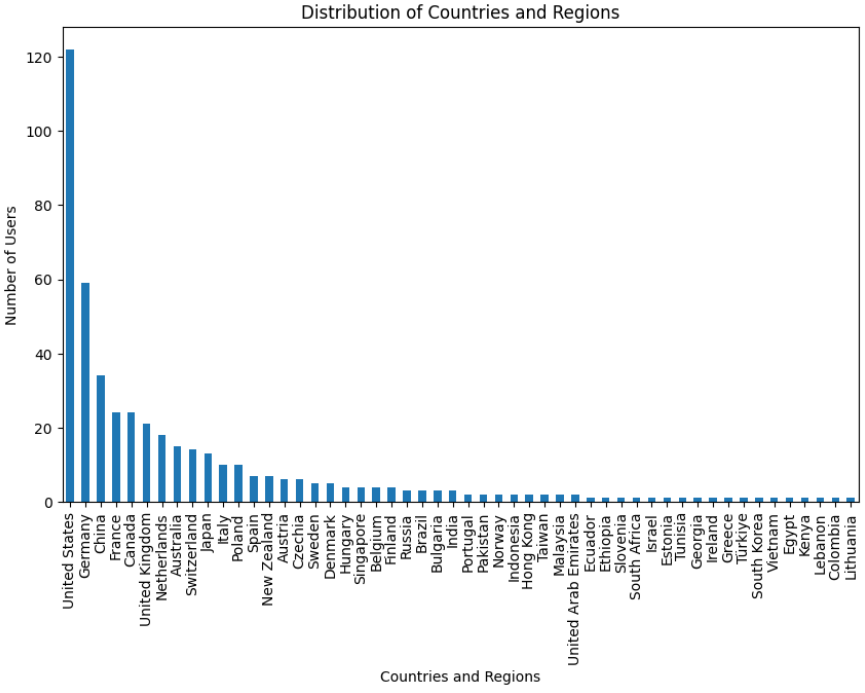
概要：此次实验作业先进行数据清洗，得到一个有唯一用户id的report_df，存储着其昵称和单人提交总量的信息，以便完成接下来的国家和地区分布，城市级别分布以及有趣洞察用户昵称的数据分析。

<注>本次报告所有提及的开发者仅为本次实验使用的近500个样本中的开发者，不代表所有开发者；提交数是单人所有类型events的总和。

人口统计分析

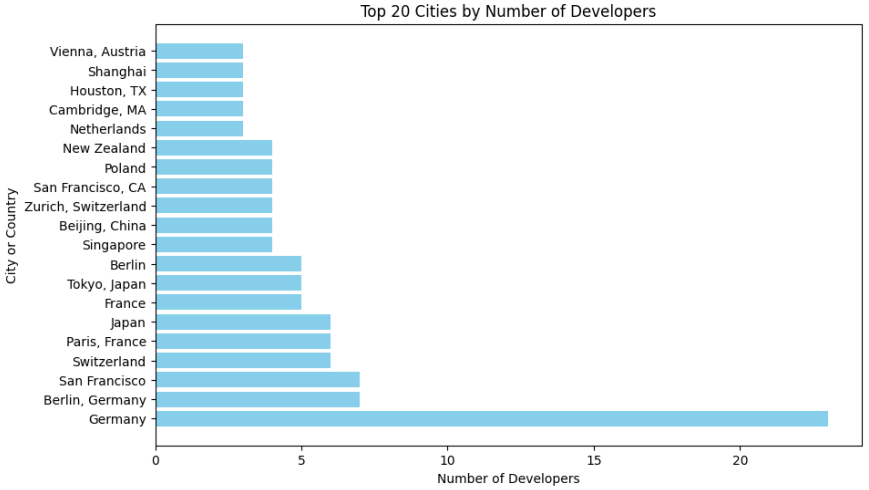
国家和地区分布

可见美国是开发者最多的国家，其次是德国，再其次是我国。可见我国的开发者人数多于一些发达国家，如法国，英国等等。由此可以估计出，我国的互联网发展态势较好。

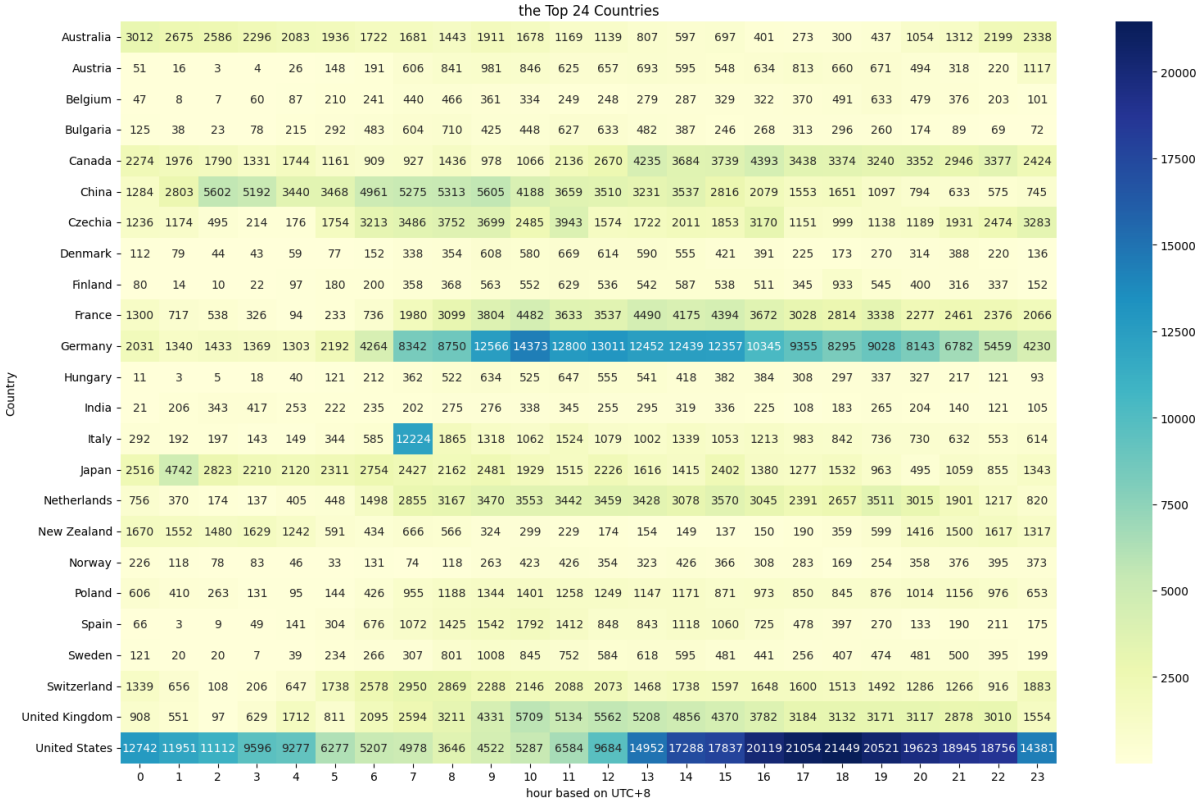


城市级别分布

可见在德国的城市开发者的数量是最多的。我国的北京和上海也双双进榜。



时区分布



经过数据查验，可知该时区是以东八区(UTC+8)即以我国时区为基准所做的图表。有此图可得知，如果我们想与美国的开发者协同工作，我们需选择12点~23点的时间与他们取得联系。

如果是和德国的开发者，就应选择7点~16点与他们取得联系。

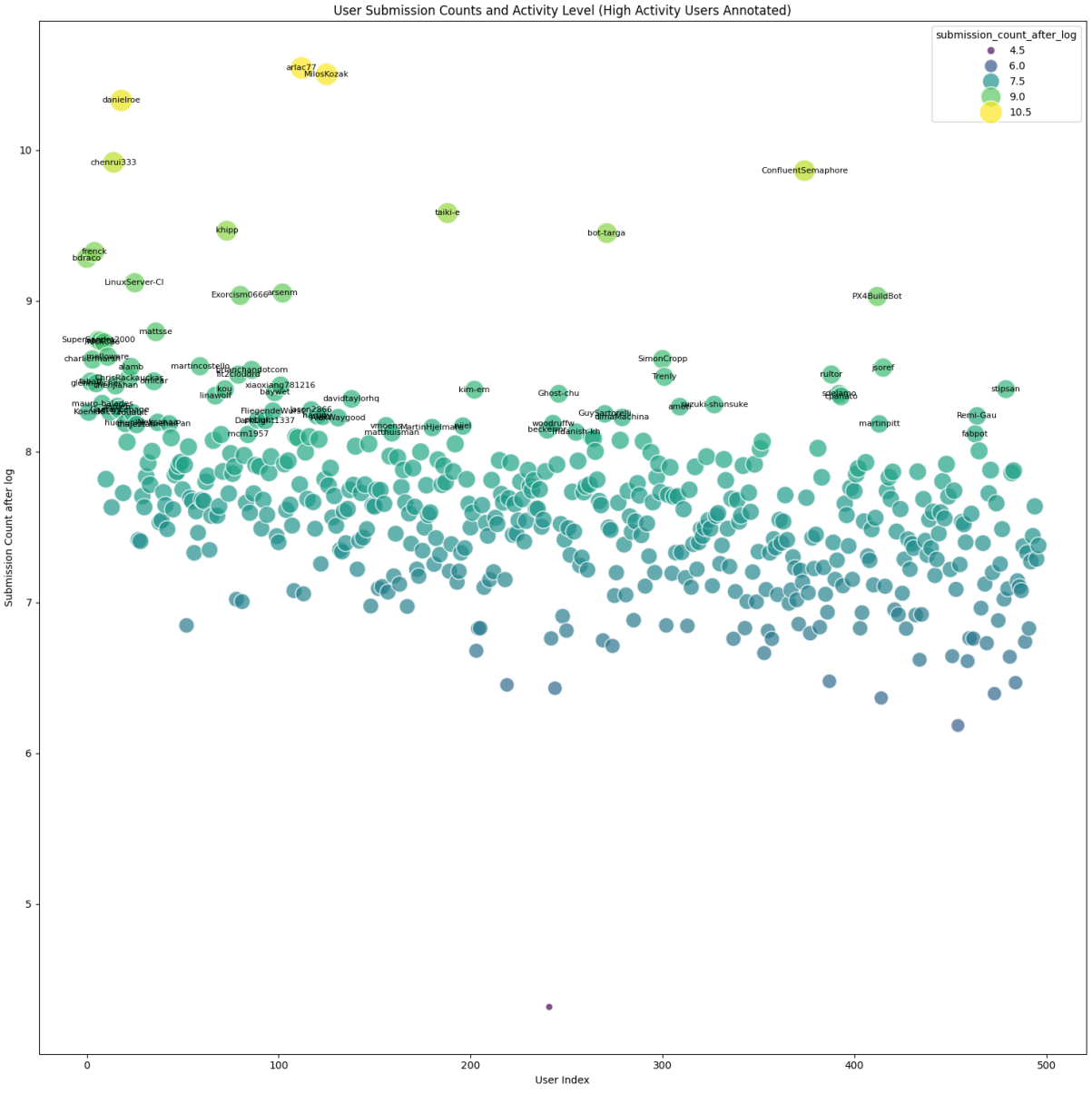
协作行为分析

提交频率

用户arlac77，MilosKozak，danielroe，取得贡献前三甲，是这些开发者中最勤劳的三者，可能是某些项目的主要承担人。

大多数人提交数在5000以下。5000左右的即为高活跃用户，他们的昵称均被标出。

3000以下的为普通用户。

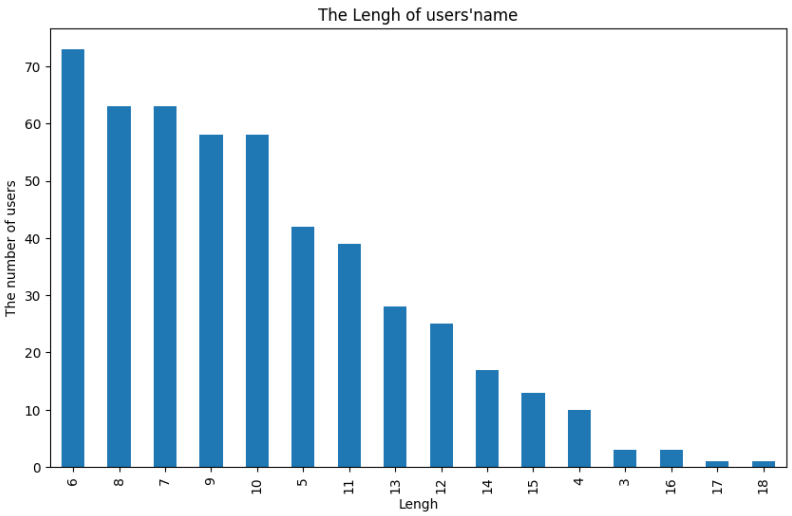


其它维度有趣的洞察

用户昵称长度

可见用户创建账号输入昵称可能是被限制在了3位字符及以上。该注册系统可能未对昵称长度设限，若未做防护，这里可能会有缓冲区溢出风险。

可见用户最喜欢设置一个长度适中(6-10)的昵称，喜欢一个简单一点的昵称，超过17位的昵称少之又少。



事件类型分布

大部分的事件是Push，然后是PullRequest，可见用户们更喜欢向托管平台上传自己的代码，并且喜欢对别人的仓库进行拉取请求，这个是一个健康的社区环境，可能是一个以开源为主的社区。