

数据洞察报告：

预处理：将七个 csv 文件整合到了一个 dataframe 中

1. 国家和地区分布：

将数据的 `country` 一列取出来绘制柱状图，发现 **United States** 的用户最多，并且远远多于其他地区，其次是 **India** 和 **Germany**，因此主要的开发者集中地是 **United States**，**India** 和 **Germany** 城市级别分布：

将数据的 `user-location` 取出来绘制柱状图，可以看出排在前三的是 **Germany**，**India** 和 **London**，所以技术热点区域是上述三个

时区分布：

我使用了 `pytz` 库获取时区，然后统计每个时区的用户数量，得到日本时区的用户是最多的

2. 协作行为分析：

以 `email` 作为提交的唯一标识符统计每个用户的提交信息，发现绝大多数用户都是只提交过一次，只有十五个用户提交过两次，这十五个用户就是高活跃用户。

3. 其他维度有趣的洞察：

电子邮件的域名分布：

对 `email` 的种类进行统计，发现绝大多数用户使用的邮箱是 `gmail.com`，其次是 `hotmail.com` 和 `outlook.com`

查看用户主要集中在哪些国家的哪些公司：

对数据中 `country` 和 `company` 进行统计，发现用户集中较多的全部是 **United States** 的公司，其中 **Microsoft**，**Google** 和 **Red Hat** 排名前三

统计用户在 **Twitter** 平台的活跃程度：

以数据中的 `twitter_username` 作为用户是否使用 **twitter** 的唯一指标，发现用户中 16.9% 是 **twitter** 用户。