Homework_7 报告

10235501461 柯宇

1. 删除重复数据，并输出去重前后的数据量

核心代码：data_deduplicated = data.drop_duplicates()

2. 缺失值处理

核心代码：

①data = data.drop(columns=['gravatar_id'])

②missing_values_before = data.isnull().sum()

③ boolean_columns = [col for col in data.columns if col.startswith('is_') or col.startswith('has_')]

for col in boolean_columns:

data[col] = data[col].fillna(False).astype(bool)

④text_columns = data.select_dtypes(include=['object']).columns

data[text_columns] = data[text_columns].fillna('')

missing_values_after = data.isnull().sum()

3. 数据变换，将 created_at、updated_at 转为时间戳

核心代码：

data['created_at'] = pd.to_datetime(data['created_at']).astype('int64') // 10**9

data['updated_at'] = pd.to_datetime(data['updated_at']).astype('int64') // 10**9

4. 数据可视化

4.1 可视化 bot 和 hunman 类型的情况（展示图表自选，并在报告中说明选择原因、结果分析以及数据洞察）

选择条形图。原因：直观展示分类数据的分布，直观看出数量差异。

结果：Human 类型用户的数量远远大于 bot 类型用户数量，证明在开源环境中，真人更有创造动力。

4.2 可视化 bot 类型账号的 created_at 情况（展示图表自选，并在报告中说明选择原因、结果分析以及数据洞察）

选择折线图。原因：所需要反映的是 created_at 情况，不如以时间作为自变量，比较 bot 类型账户在各个年份的活跃情况并且刻画趋势。

结果：bot 用户的活跃程度不是一成不变的，bot 账户的建立也有时间规律，应该是出于某个时段的研究目的。

4.3 可视化 human 类型账号的 created_at 情况（展示图表自选，并在报告中说明选择原因、结果分析以及数据洞察）

选择折线图。原因：所需要反映的是 created_at 情况，不如以时间作为自变量，比较 human 类型账户在各个年份的活跃情况并且刻画趋势。

结果：human 用户的活跃情况有起伏，并且在 2012 年左右达到顶峰。可以判断 2012 年的开源社区活跃度较高，新用户数量增多。而且可以发现，本数据集可能是稍微比较早

期的数据，对 2024 甚至 2023 的统计不甚完全。

4.4 可视化 bot 类型账号的 followers 和 following 情况（展示图表自选，并在报告中说明选择原因、结果分析以及数据洞察）

选择散点图。原因：为了展示两个变量之间的关系，即 followers 和 followings 的关系。

结果：大部分 bot 账户的关注量和被关注量是持平的，没有显著的特点。但仍有部分 bot 账户出于研究需要有着极端的关注量或者被关注量。

4.5 可视化 human 类型账号的 followers 和 following 情况（展示图表自选，并在报告中说明选择原因、结果分析以及数据洞察）

选择散点图。原因：为了展示两个变量之间的关系，即 followers 和 followings 的关系。

结果：human 账户的关注量和被关注量关系和 bot 账户有很大不同，一般关注的人多的账户被关注量较少，而被关注量较多的账户关注的人较少。结合开源社区的社交实际，大佬关注的人少，而被关注的量大；萌新关注的人多、需要寻找灵感，而被关注的量比较小。充分展示了学习者与教授者的特点。