**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()**

1. **缺失值处理**

**核心代码：**

**①data = data.drop(columns=['gravatar\_id'])**

**②missing\_values\_before = data.isnull().sum()**

**③boolean\_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()**

1. **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**

1. **数据可视化**
   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。**