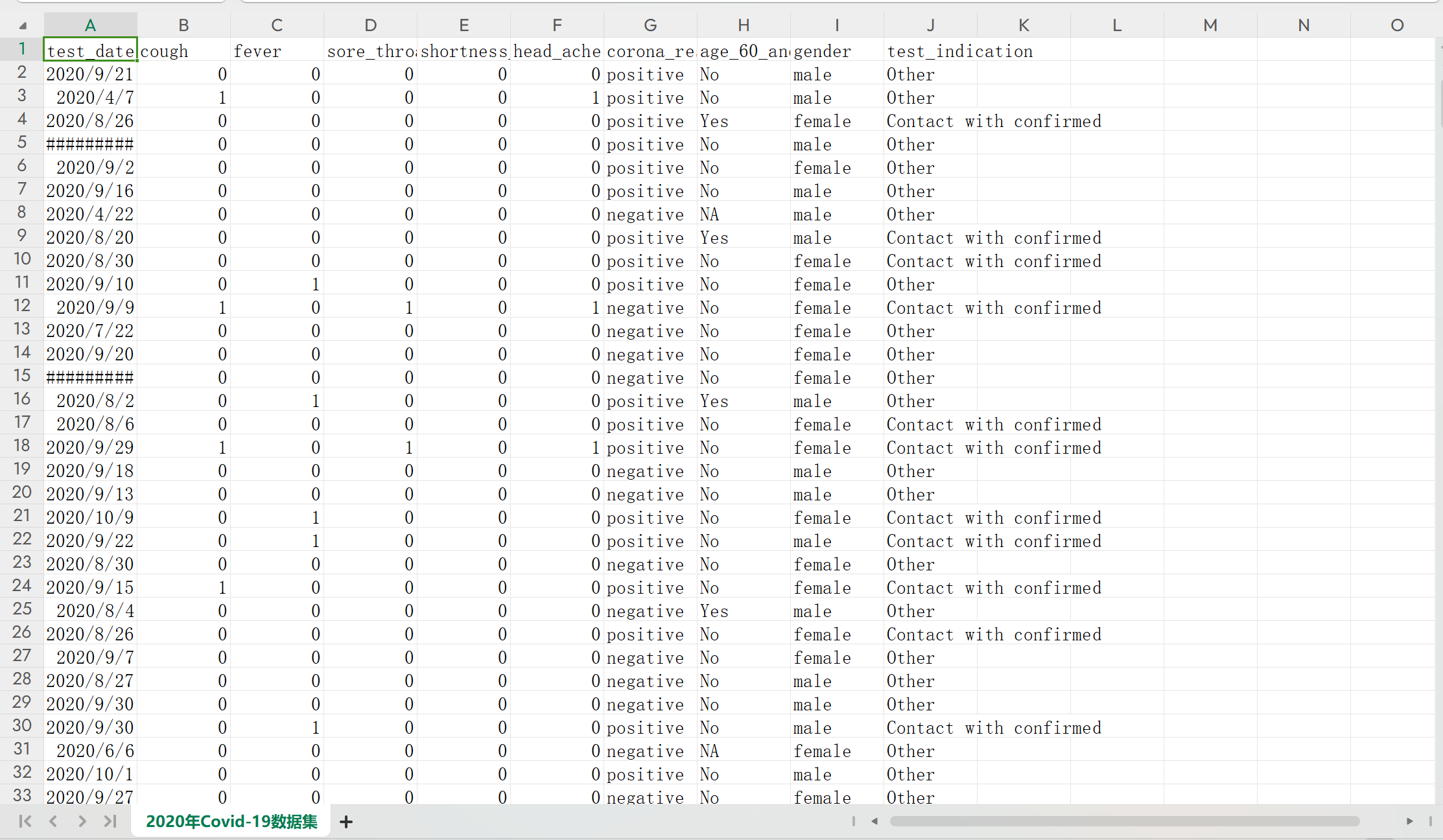
**导论大作业实验报告**

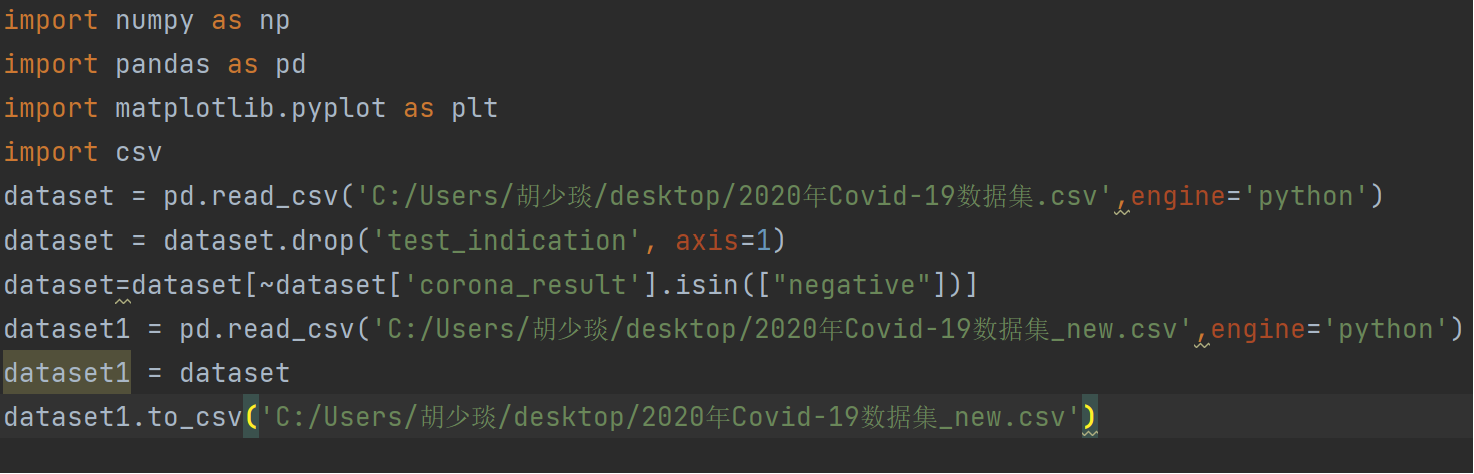
**胡少琰 10225501417**

1. 作业背景
2. 选题的背景和考虑：新冠肺炎疫情从2020年上半年的第一次大规模爆发，到2022年3月至11月因[奥密克戎变异株](https://zh.wikipedia.org/wiki/%E5%9A%B4%E9%87%8D%E6%80%A5%E6%80%A7%E5%91%BC%E5%90%B8%E9%81%93%E7%97%87%E5%80%99%E7%BE%A4%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%922%E5%9E%8BOmicron%E8%AE%8A%E7%95%B0%E6%A0%AA" \o "严重急性呼吸系统综合征冠状病毒2Omicron变异株)广泛传播而导致的第二次大规模爆发，再到2022年11月至2023年1月的第三次大规模疫情爆发，以其持续时间之长、病毒威力之大和死亡率之高而让民众恐慌，正值2024年新春之际，新冠疫情也基本得到控制，于是想要对2020年3月-11月之间曾感染新冠疫情的人群的年龄、性别以及感染症状进行分析，观察各项指标之间是否有关联以及关联程度，通过分析结果判断一些被民众广泛接受的理论是否有数据支持等。
3. 数据集：数据集包含从2020年3月到11月400000个测试样本的十个特征值的数据（数据集来源：[数据科学科研和教学一体化平台 (idatascience.cn)](http://www.idatascience.cn/dataset-detail?table_id=120)）



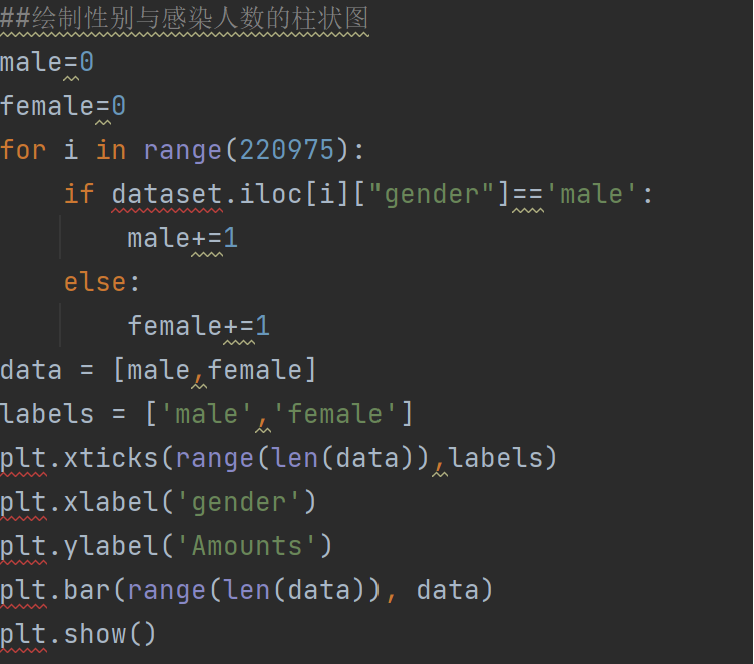
数据特征值的含义以及说明：

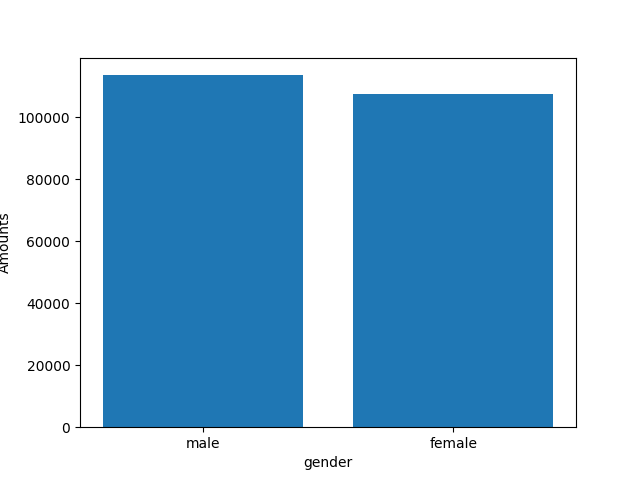
|  |  |
| --- | --- |
| test\_date | 测试日期 |
| cough | 咳嗽 |
| fever | 发热 |
| sore\_throat | 咽喉痛 |
| shortness\_of\_breath | 气促 |
| head\_ache | 头痛 |
| corona\_result | 测试结果 |
| age\_60\_and\_above | 是否为60岁以上 |
| gender | 性别 |
| test\_indication | 测试适应症 |
|  |  |

1. 数据预处理
2. 由于想要对曾感染新冠疫情的人群的年龄、性别以及感染症状进行分析，观察各项指标之间是否有关联以及关联程度，在十个特征值中不会用到test\_indication这一项，因此需要在预处理时删除；再通过观察数据集发现原始数据包含的是所有进行过检测的人的样本数据，也就是将检测出来未感染新冠疫情的样本也包含其中，所以在预处理时也要将corona\_result这一特征值为negative的样本数据删除，将进行过预处理的数据保存到2020年Covid-19数据集\_new.csv文件中，进行这一操作的代码段如下：

三、数据分析：在数据分析的过程中，我分析了几个在大众认知中比较有可能关联程度较高的特征值，以此来检验这些认知是否正确以及真正的关联情况是怎样的

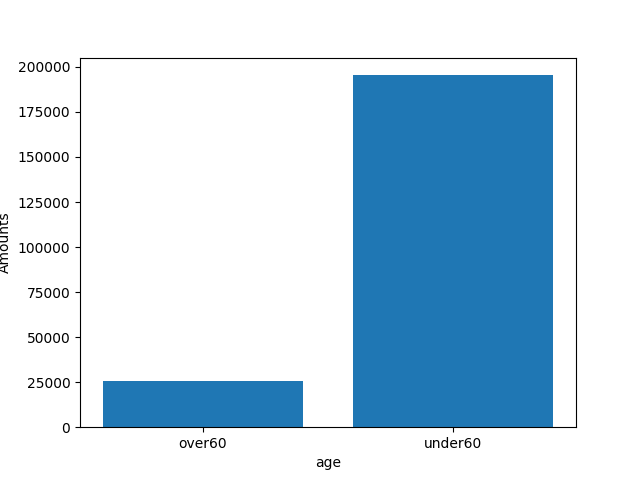
1、画出在这段时间感染的人中性别的分布情况，代码和柱状图如下：



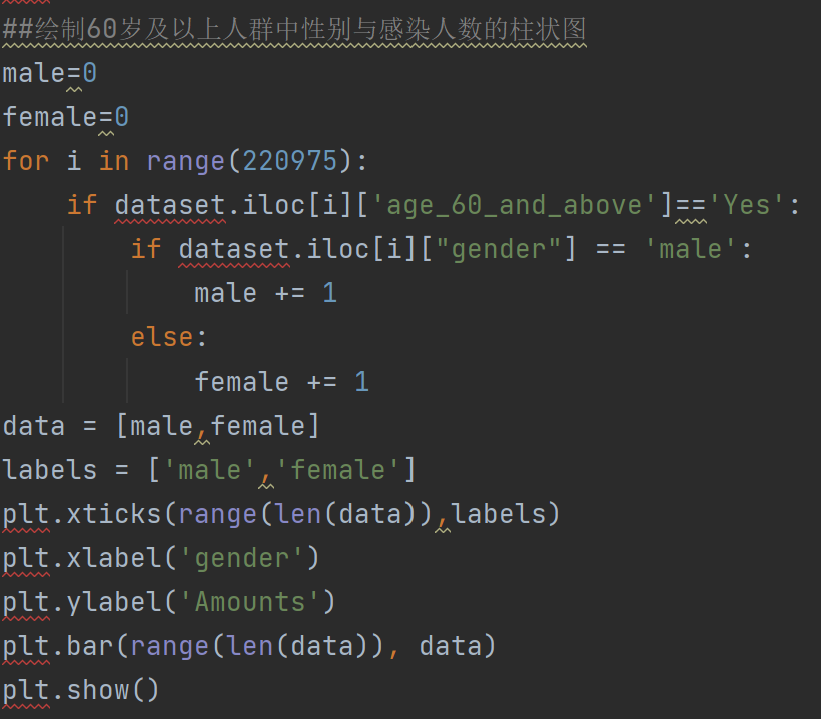


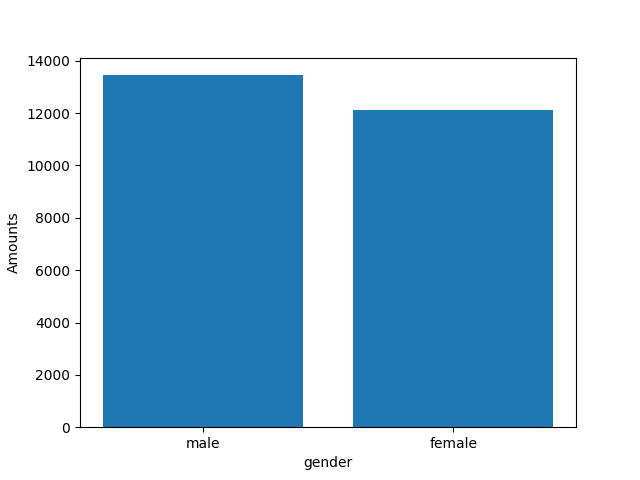
由柱状图可以看出，在感染的人中男性和女性的人数基本相同，男性要略高于女性一些；

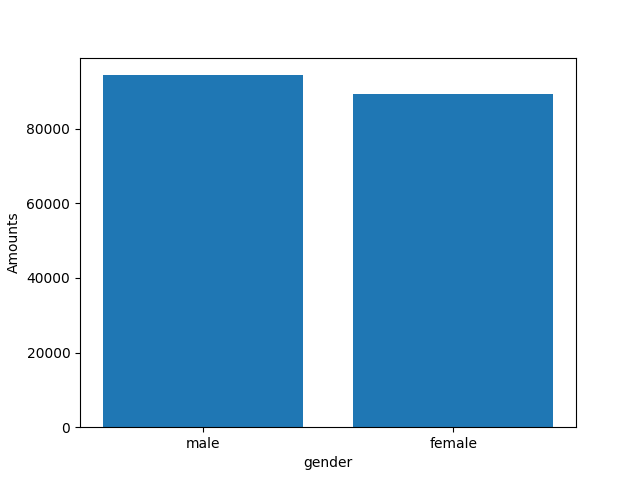
1. 画出在这段时间内感染的人的年龄分布情况，观察超过六十岁的和没到六十岁的人数差异，代码和柱状图如下：



由柱状图可以看出，感染的人中超过60岁的人数远少于不到60岁的，这说明专家在预测新冠病毒会更容易入侵老年人的免疫系统这一结论在这个数据集的验证下并不正确，感染情况和年龄并无很明显的直接关联；

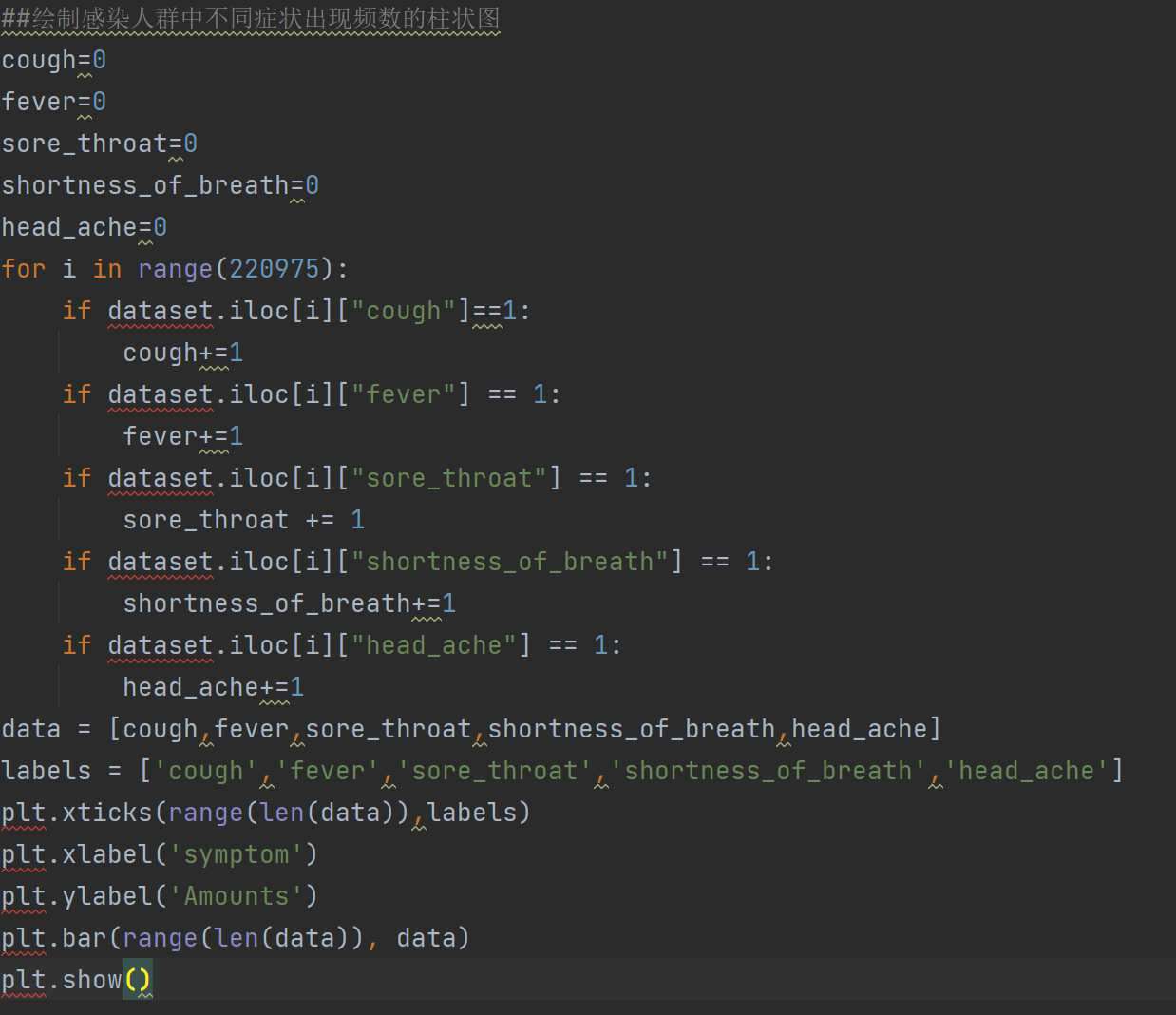
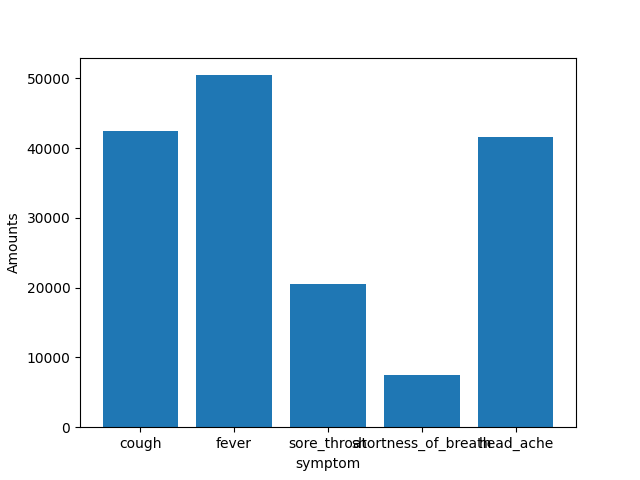
1. 在分别分析年龄和性别与感染情况的关系后，我想到有专家分析新冠病毒会更加容易年轻男性，影响男性的生殖系统，于是另外画出不到60岁的人群和超过60岁的人群中感染人数的性别分布情况，代码和柱状图如下：





从上述两个柱状图可以看出，无论是在60岁以上人群还是不到60岁的人群中，男性的感染人数都要略高于女性，但并无明显差别，因此民众所相信的男性更易感理论并不能从此数据集中得到印证；

1. 新冠肺炎患者的症状极多，每个人的状况也不相同，因此选取了五个较为常见的症状，画出在2020年3月-11月感染的人中不同症状出现的频数，绘制柱状图：

观察柱状图可知，在五种症状中发烧是最常见的，咳嗽和头痛的人数差不多，略少于发烧的人，喉咙痛和呼吸急促是两种最少间的症状，由此可见医生建议在发现发热的初期就去医院进行核酸检测是符合事实逻辑，能够被此数据集验证的。

1. 总结

在对于2020年Covid-19数据集的诸多特征值进行分析以后不难发现，医生对于新冠肺炎病毒的了解部分有事实依据支撑，但部分仍未得到印证，比如新冠病毒并不是更容易感染男性，男女患病的人数大致相同，以及根据是否发热来判断是否需要进行核酸检测是有数据支持的理论等。本次实验因为一些原因使得数据可视化的过程不够丰富和直观，这也是给我留下较大遗憾的地方，希望日后有机会可以改进。