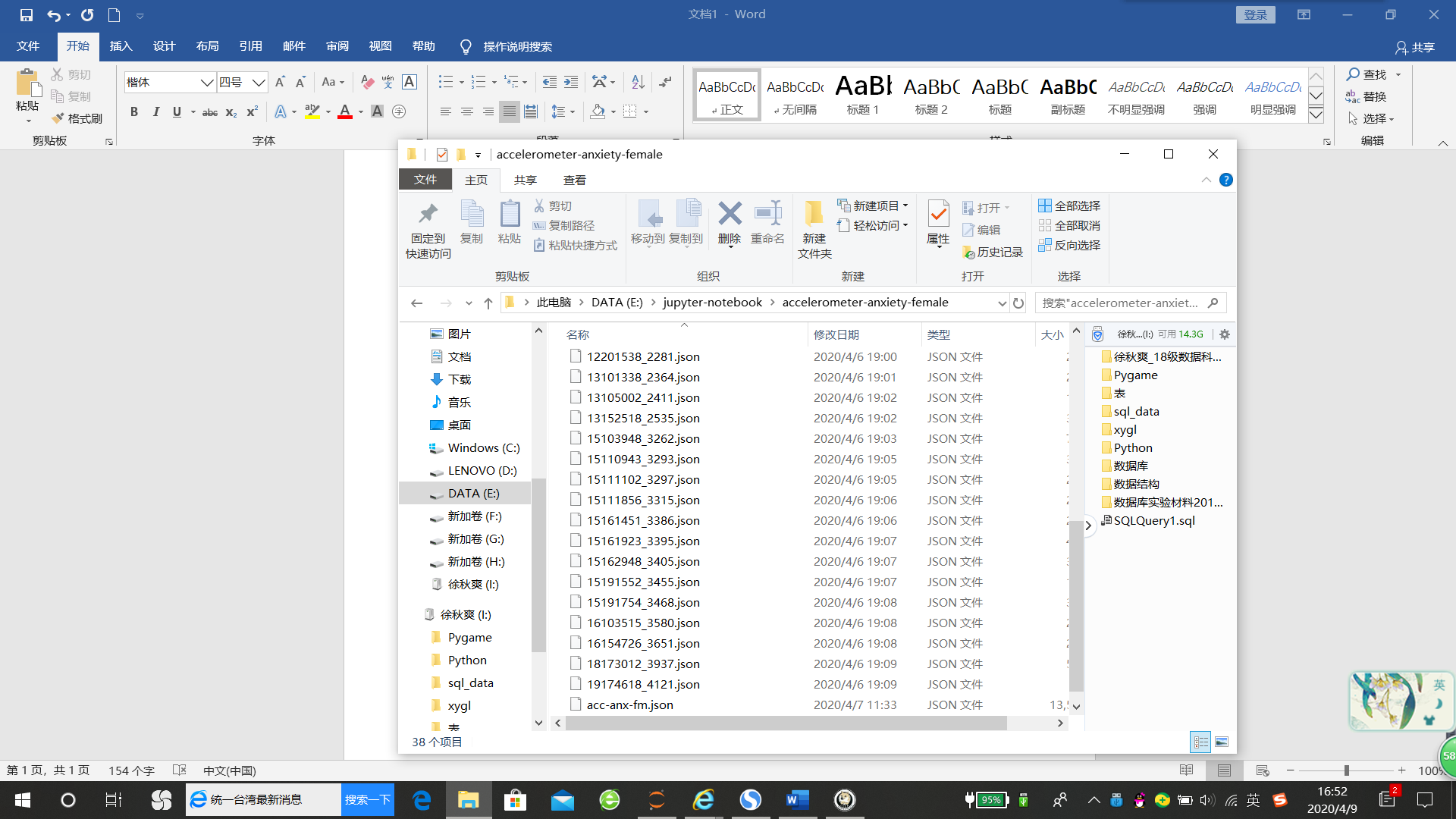
说明文档

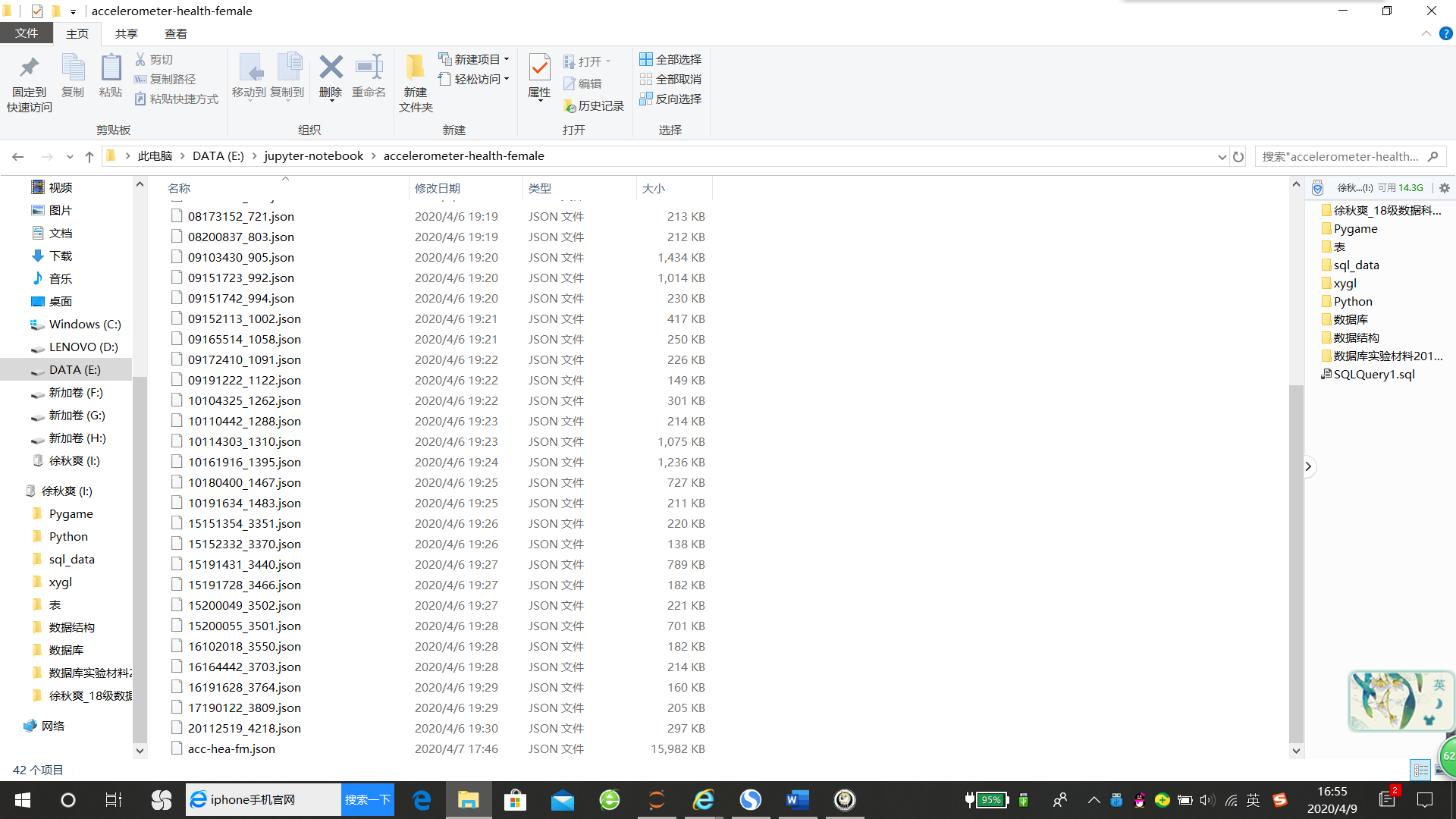
320180941890 18级数据科学与大数据技术 徐秋爽

1.爬取数据：

通过在线读取数据，我发现数据量非常的大，于是我将需要处理的数据按照类型放在了accelerometer-anxiety-female、accelerometer-health-female、device\_motion-anxiety-female、device\_motion-health-female、gyroscope-anxiety-female、gyroscop-health-female六个文件夹中，将每个类型的抑郁症患者与健康人分开汇总，除了得到每个ID对应的数据以外，还得到了汇总的数据，我将获取的json数据分成六个类型，分别是acc-anx-fm.json、acc-hea-fm.json、dev-anx-fm.json、dev-hea-fm.json、gyr-anx-fm.json、gyr-hea-fm.json，分别与上面的文件夹相互对应。

爬取结果：

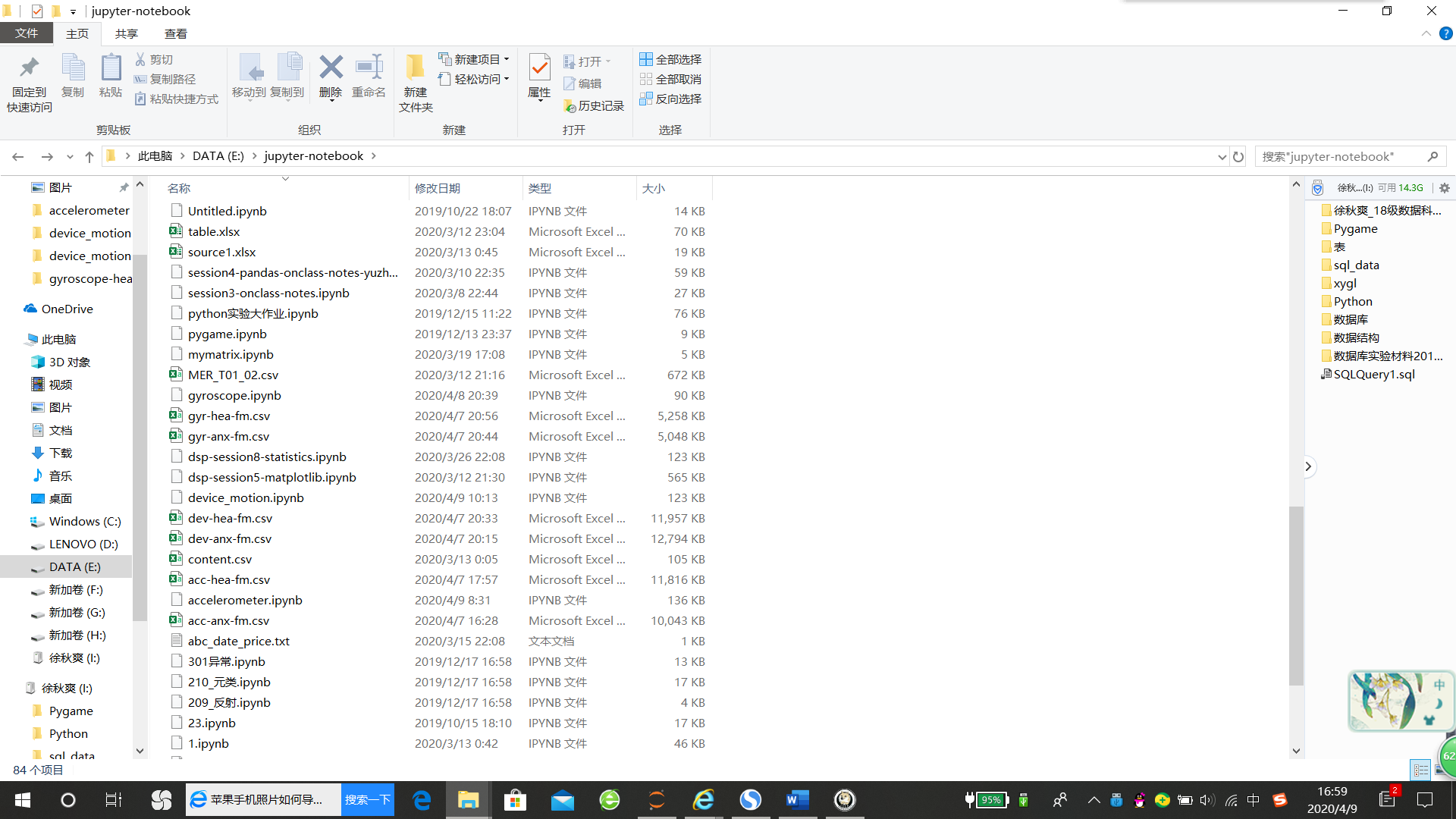


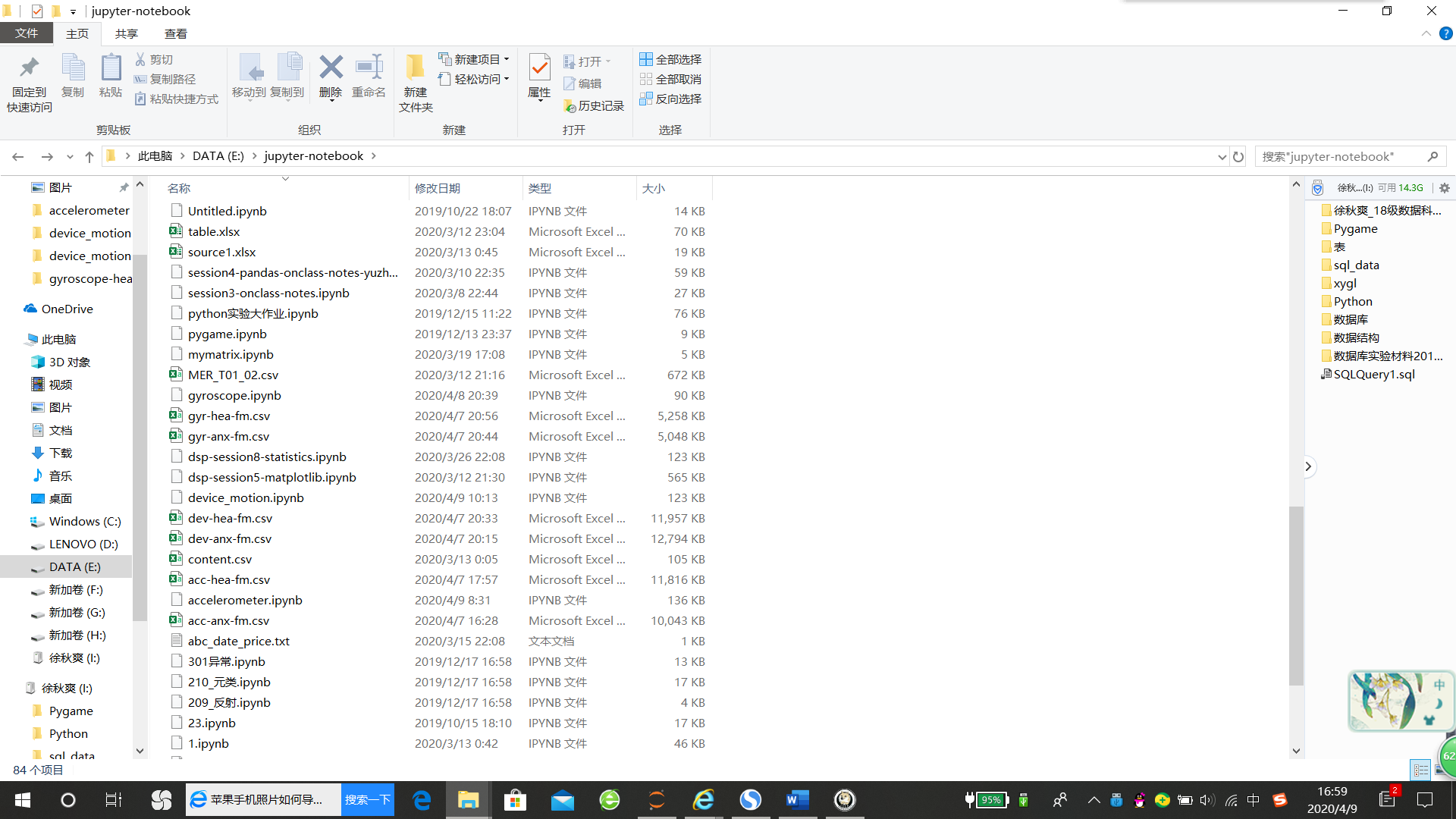


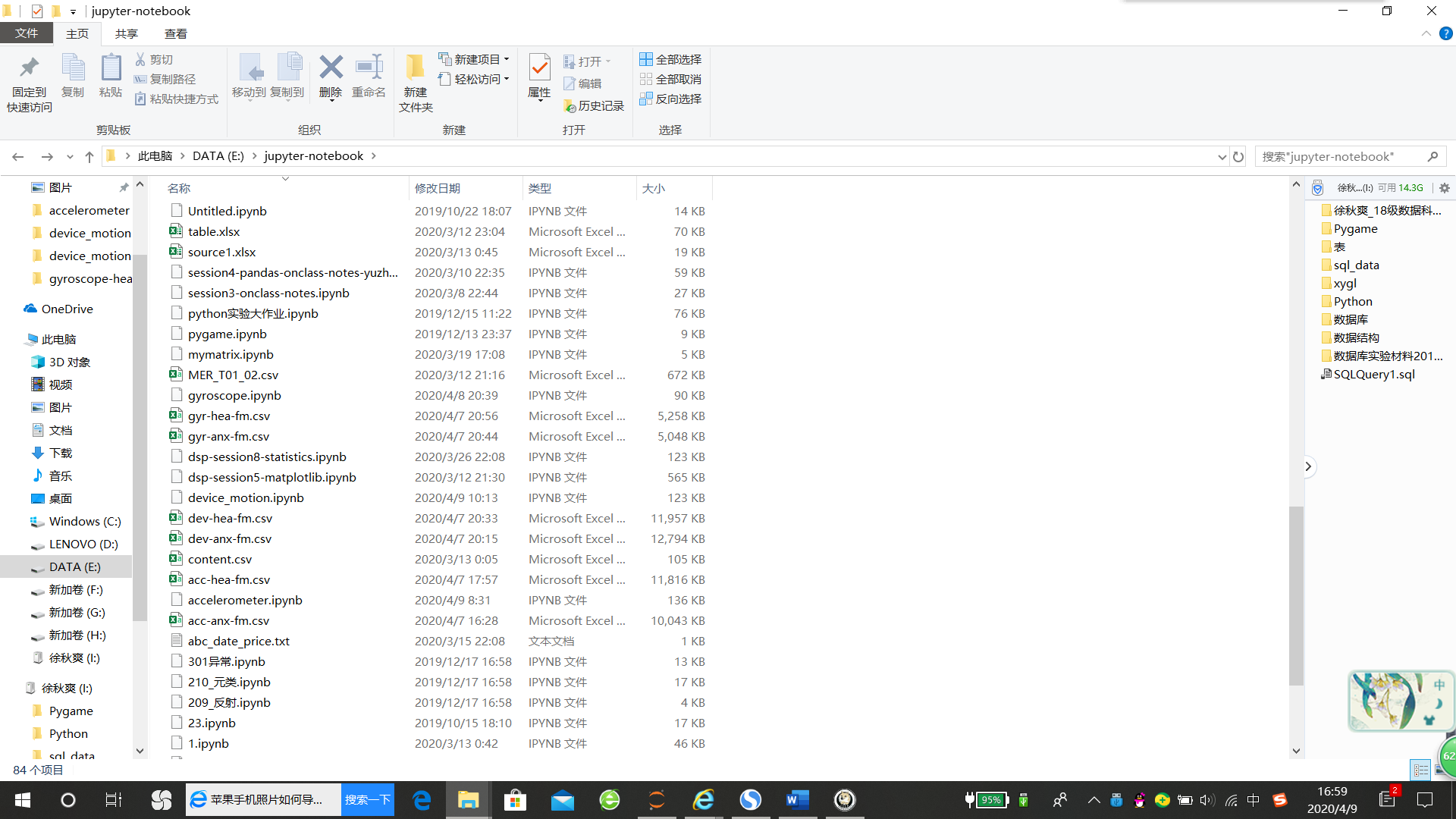
因为空间原因，余下的不作展示。

2.格式转换：

为方便使用数据以及对数据进行审查，将json数据转化为csv格式。







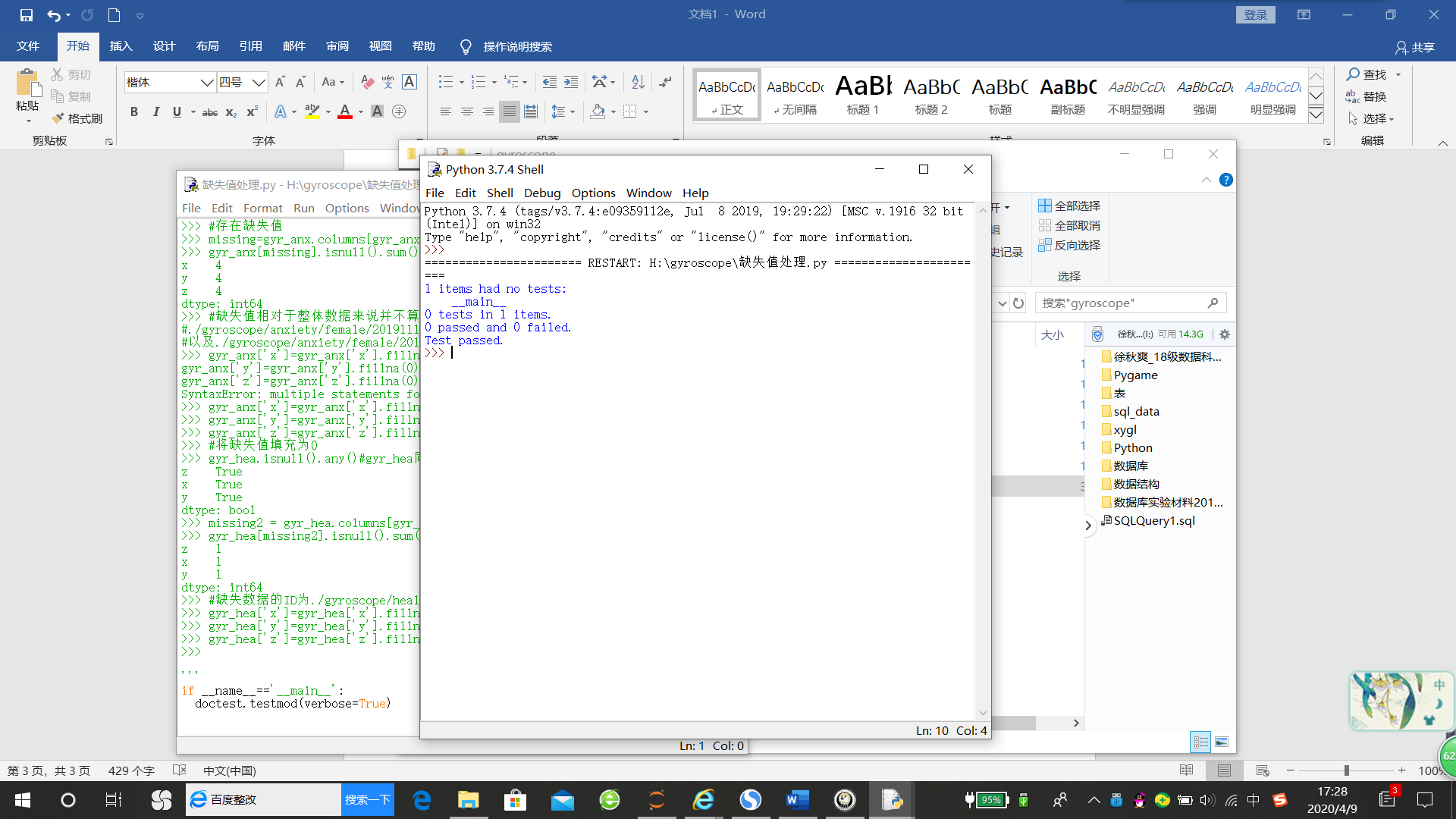
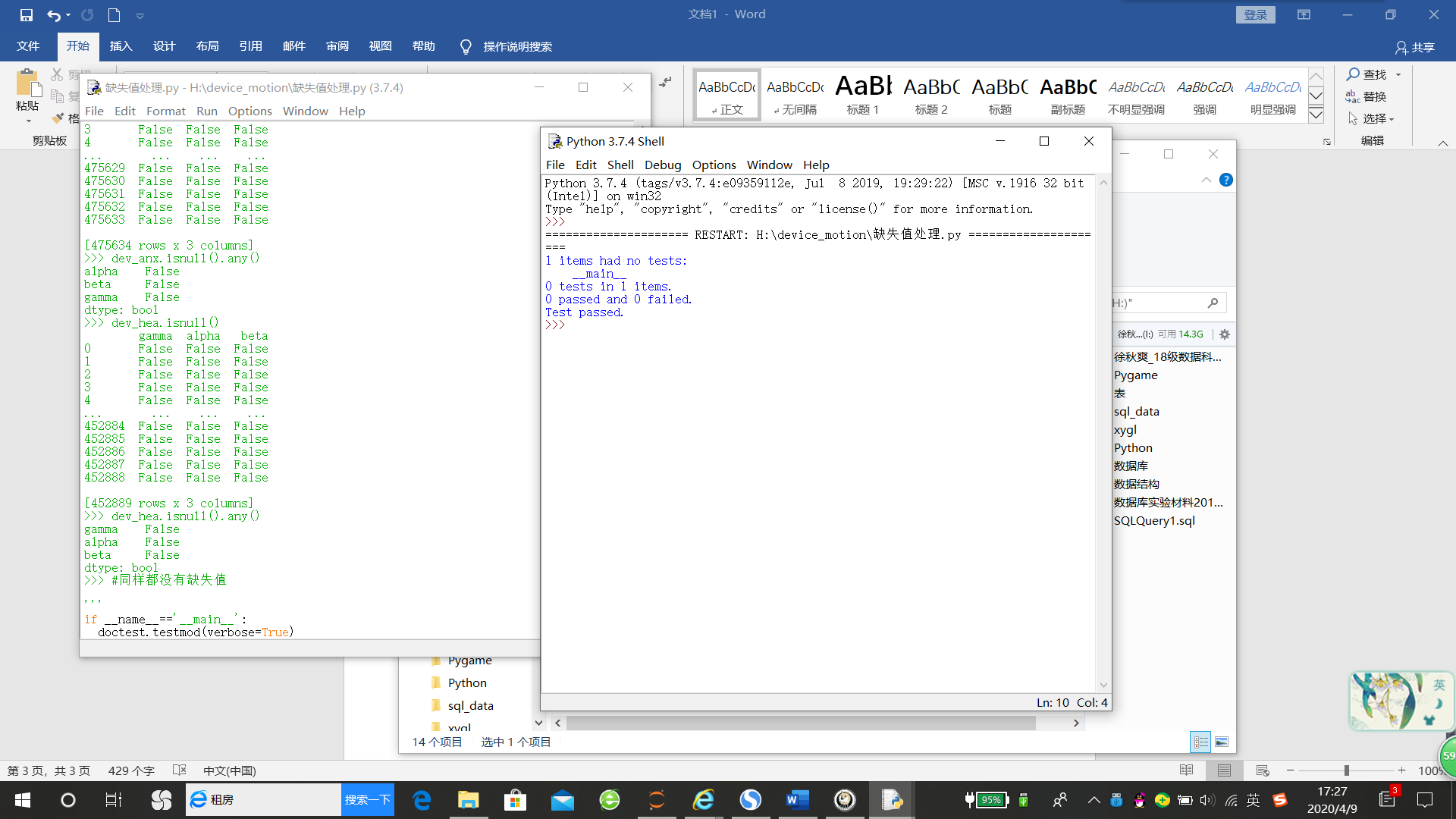
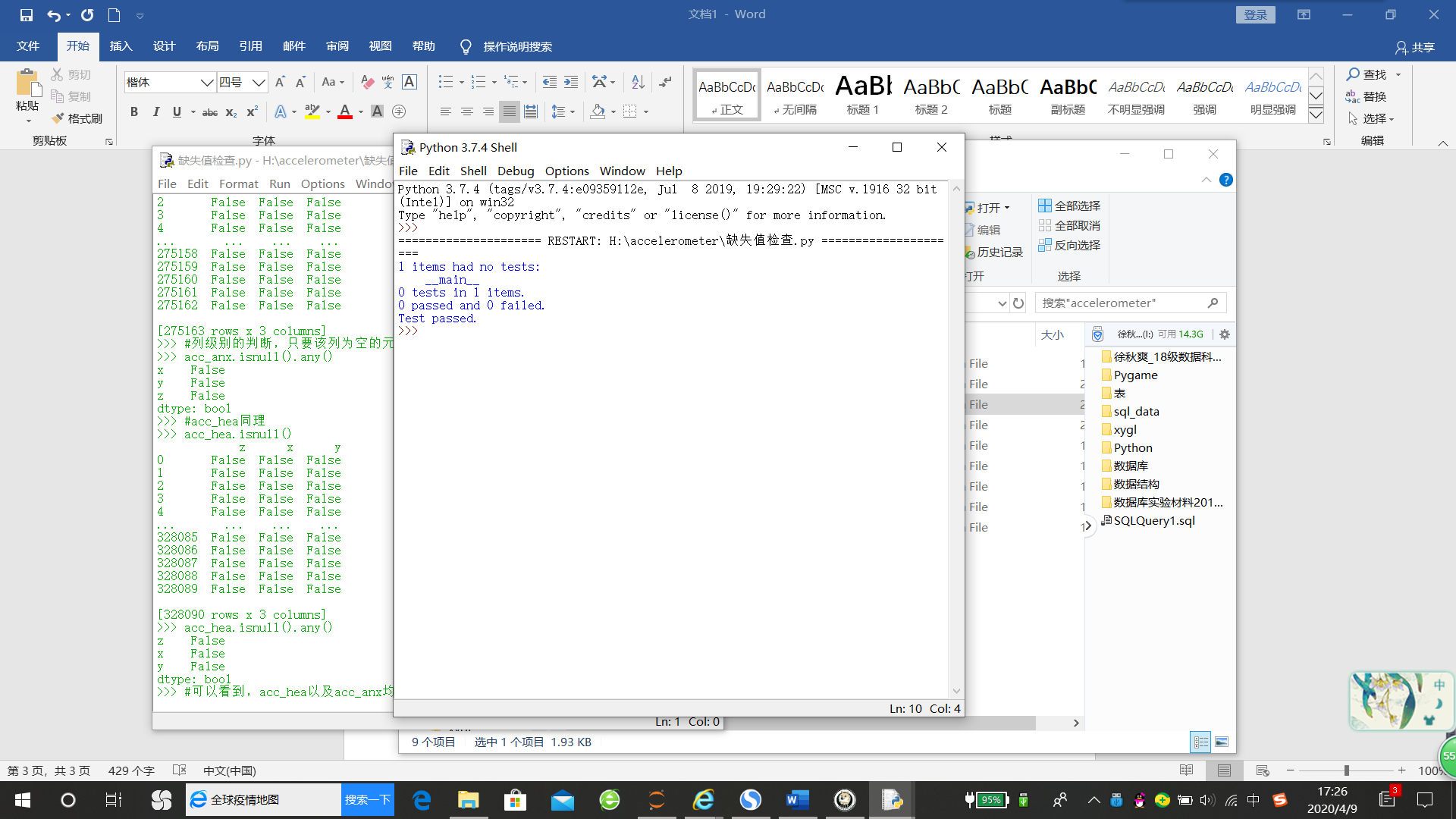
上图为我获取的csv文件，因为数据量太大，不做上传，程序上传可能无法正常运行，我将提供运行的截图。

3.缺失值检查：

因为余下的程序步骤基本相同，我将以accelerometer为例，如果有不同我将会做补充。

读取csv文件，通过运用isnull()函数确定是否有缺失数据，在数量庞杂的情况下，这种方法其实并不可取，通过运用isnull().any()对列级别判断，只要有空值，就为True,否则为False,通过运用此函数，我们只在gyr-anx-fm.csv以及gyr-hea-fm.csv中发现了缺失值，通过isnull().sum()确定了缺失值的总数，因为对整体数据影响不大，我将他们填充为0，我在其中附加了doctest，

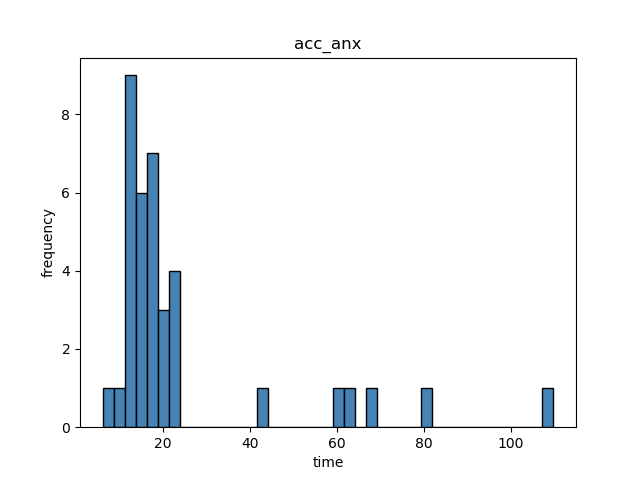
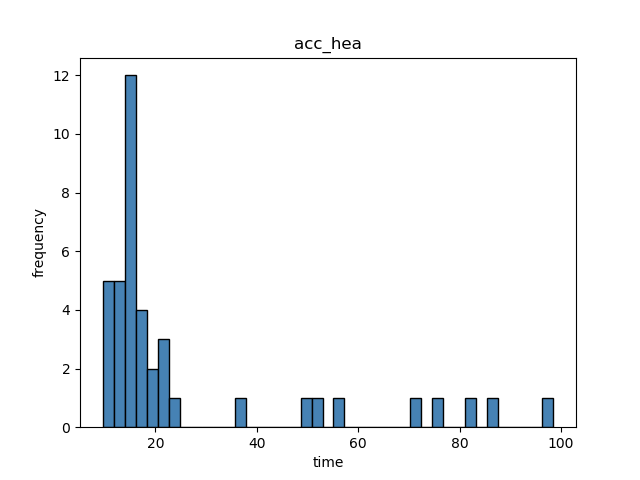
显示测试均通过。



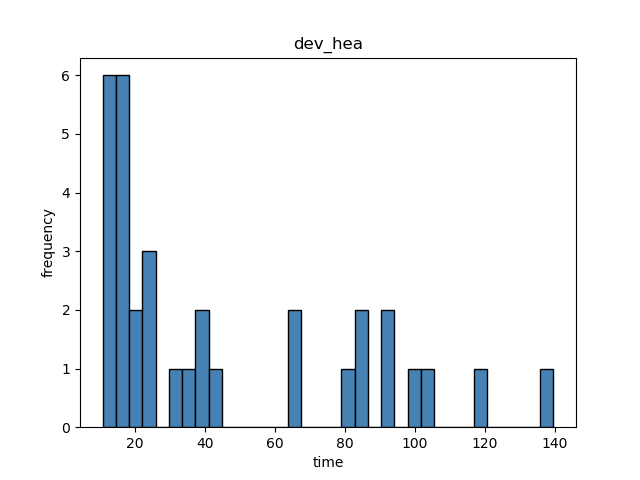
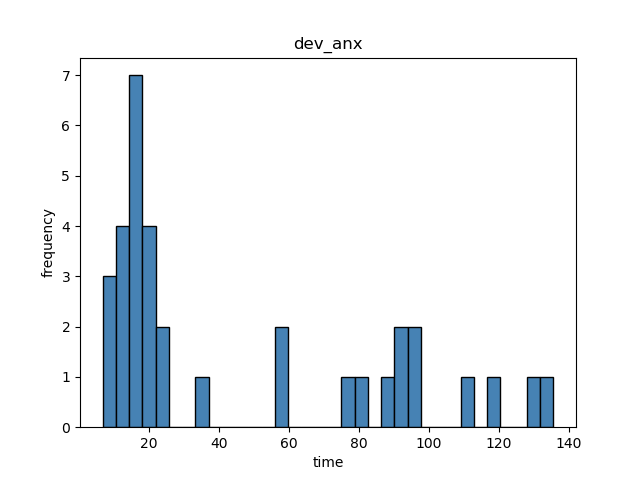
4.答题时间直方图：

根据5HZ的频率，我算出了采集每个ID信息所需要的时间，之后对数据进行了统计，得到了频率分布直方图（将图片附加在了文件夹中）。

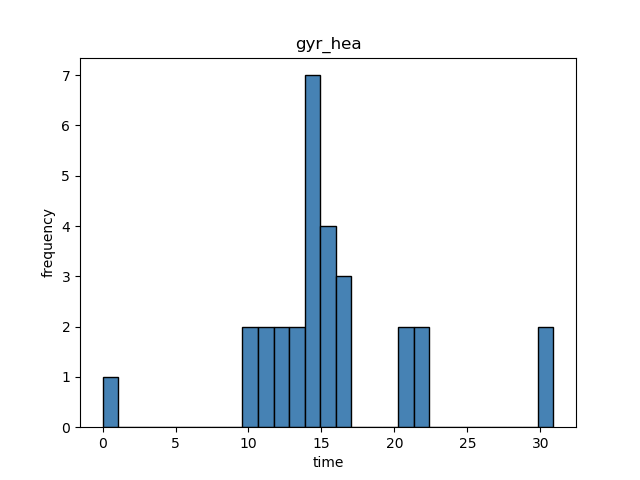
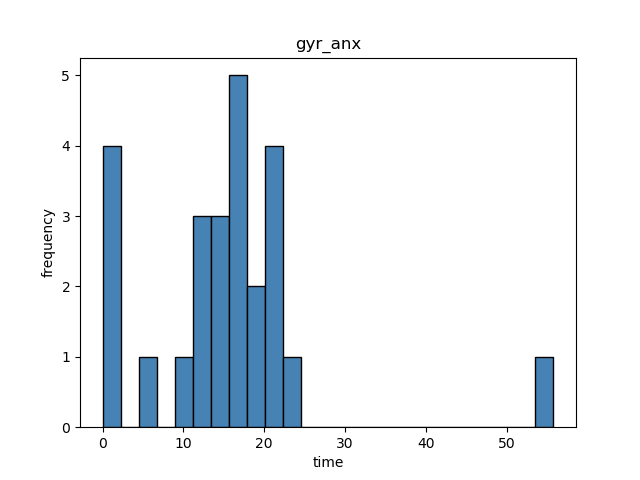
Accelerometer:



Device\_motion:



Gyroscope:



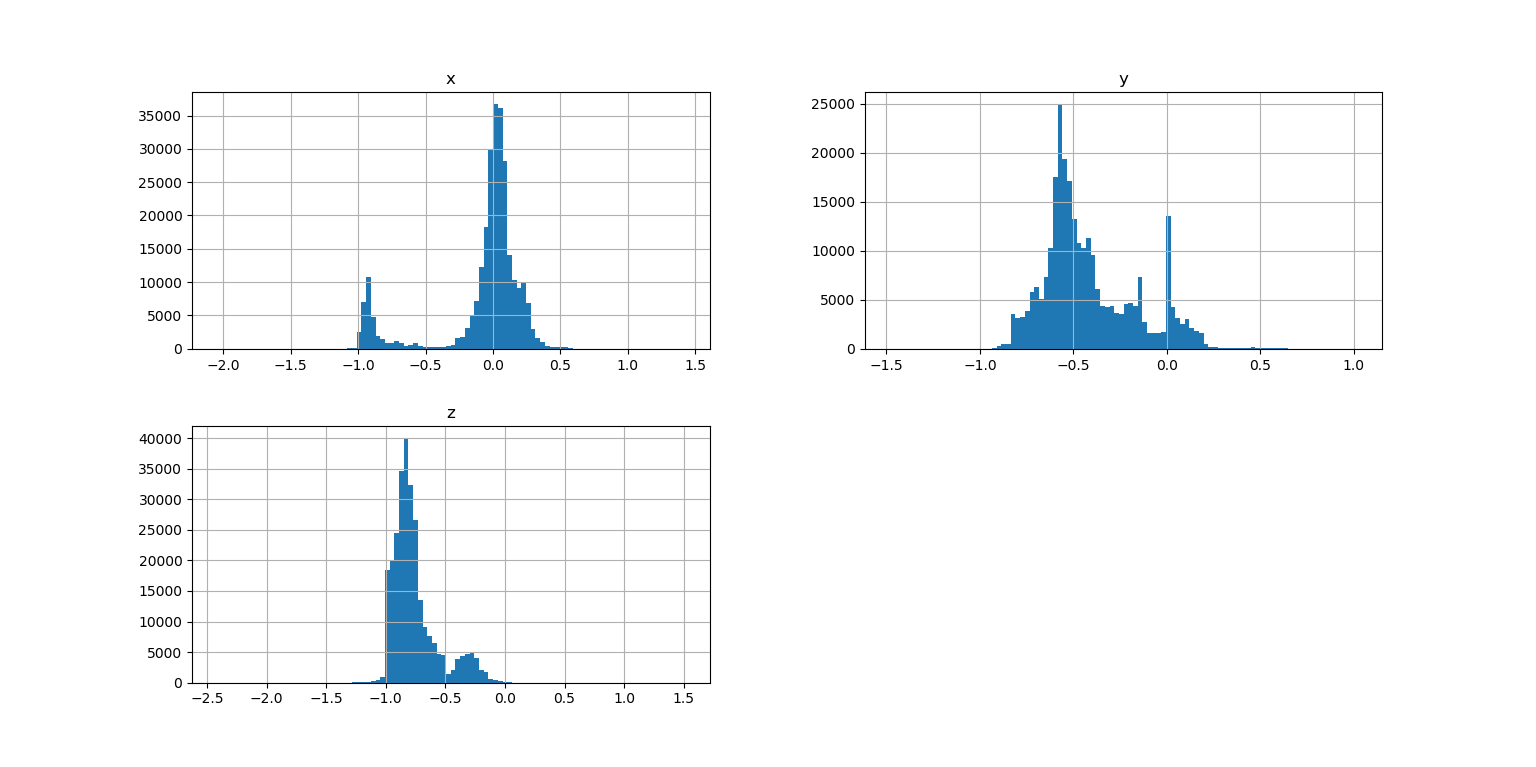
我们可以看到，每张图片都有集中点，那么我们其实可以舍弃较短和较长的时间，例如：acc\_hea、acc\_anx可以舍弃超过80min的数据，dev\_anx、dev\_hea可以舍弃超过100min的数据，gyr\_anx、gyr\_hea可以舍弃0的数据，这样或许可以让数据更为集中。

5.x,y,z分布图

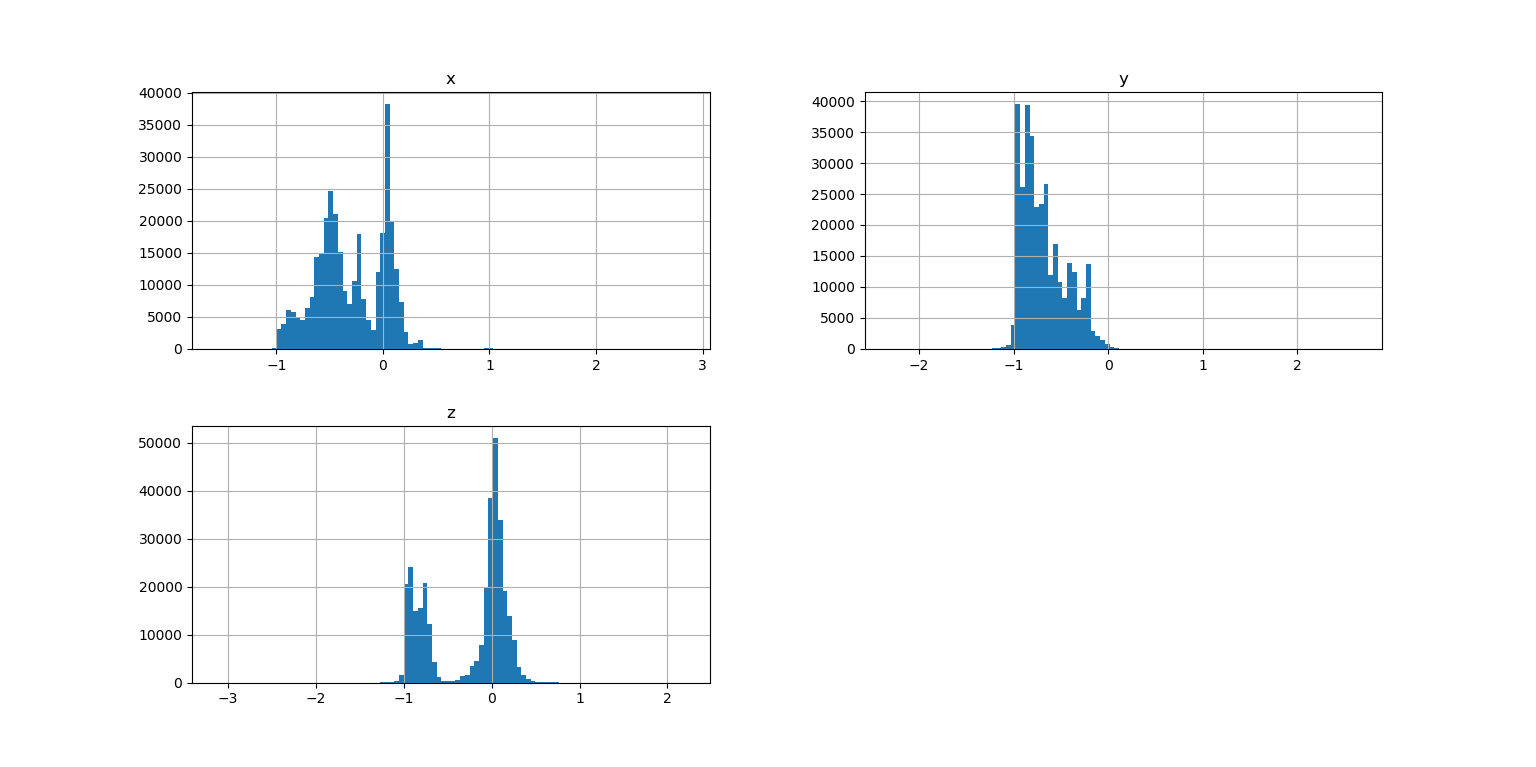
我将六个类型的全部数据进行了汇总，将他们绘成了频率分布直方图，方便进行对比。

Accelerometer:

抑郁症患者：

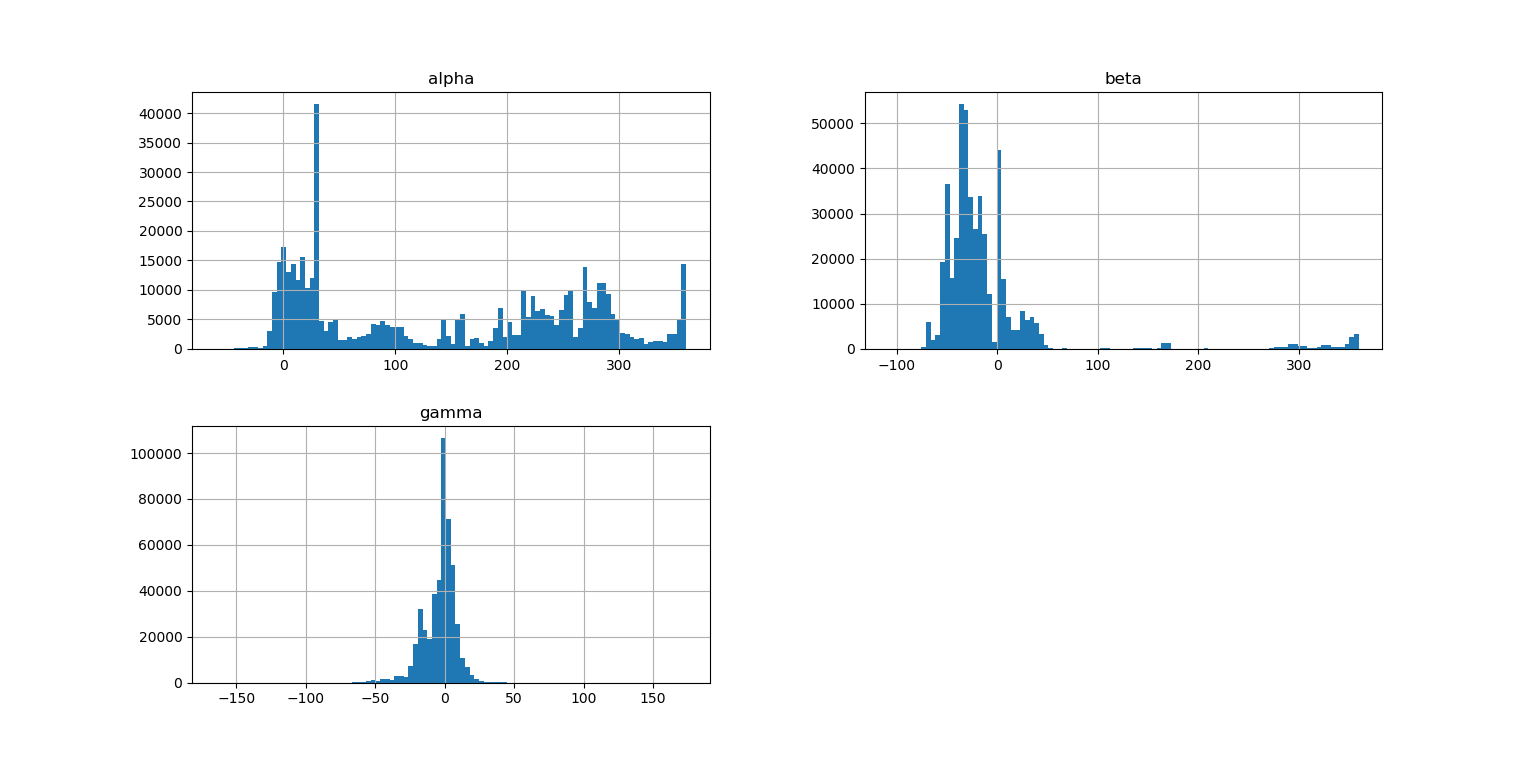


健康人：

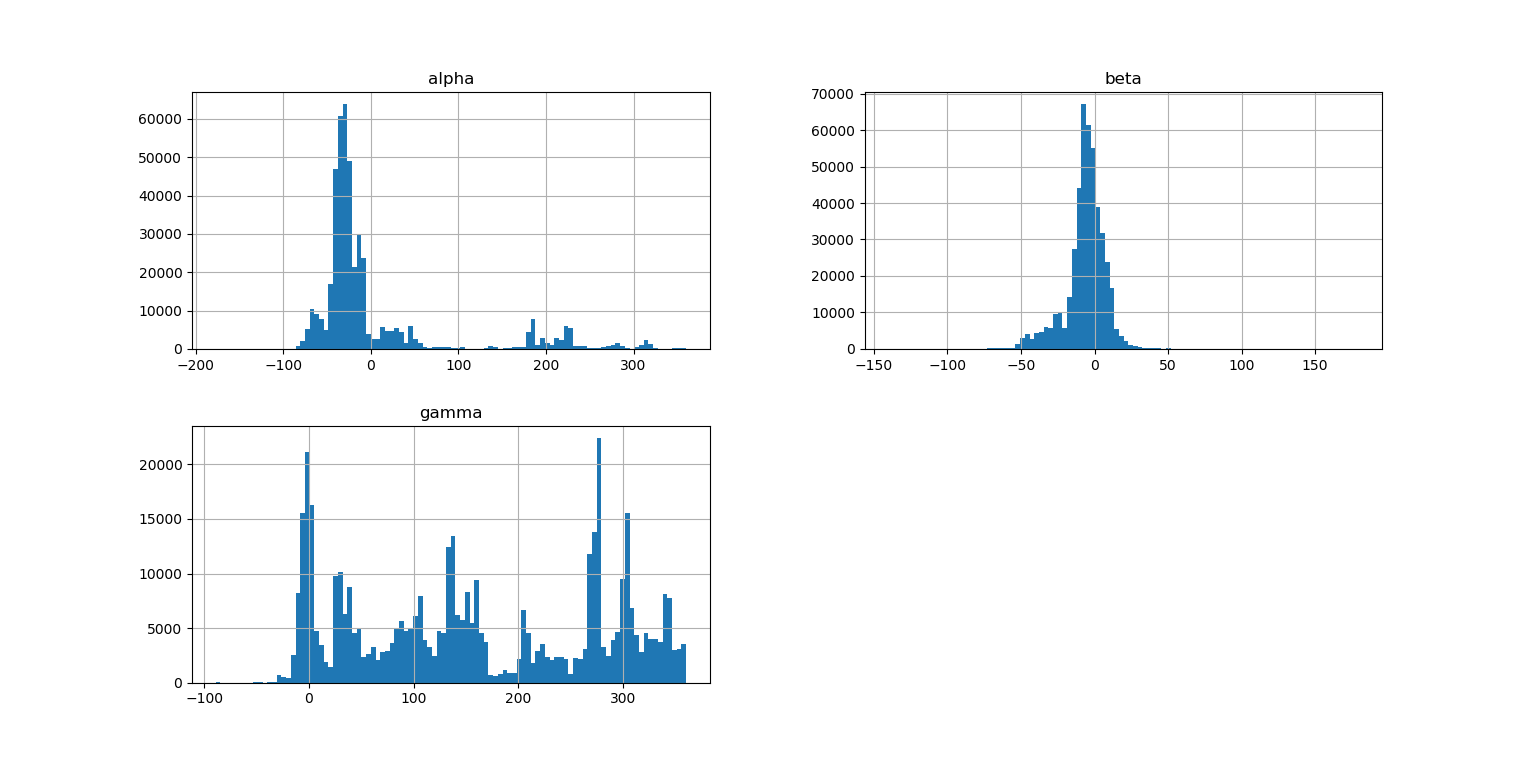


Device\_motion:

抑郁症患者：

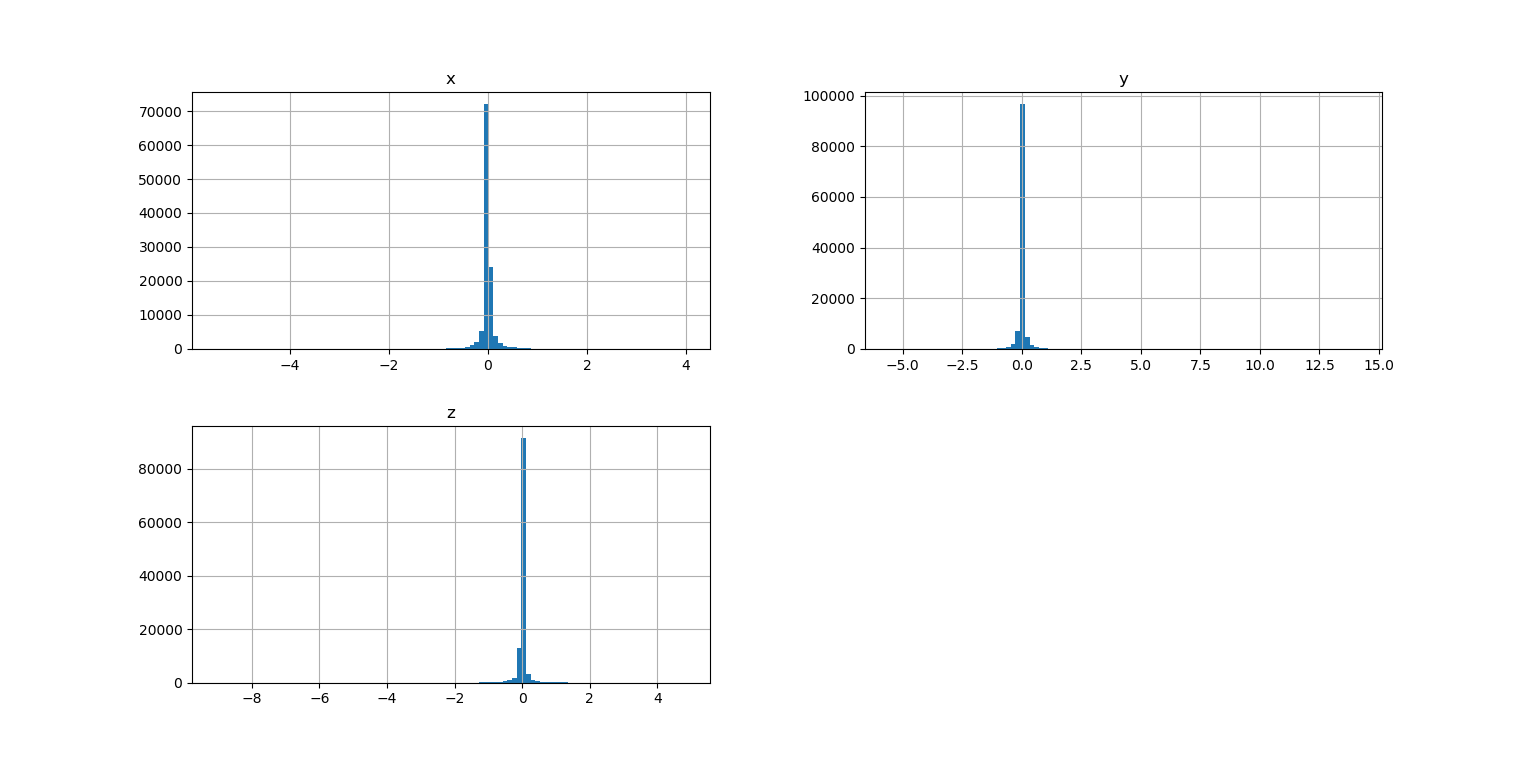


健康人：

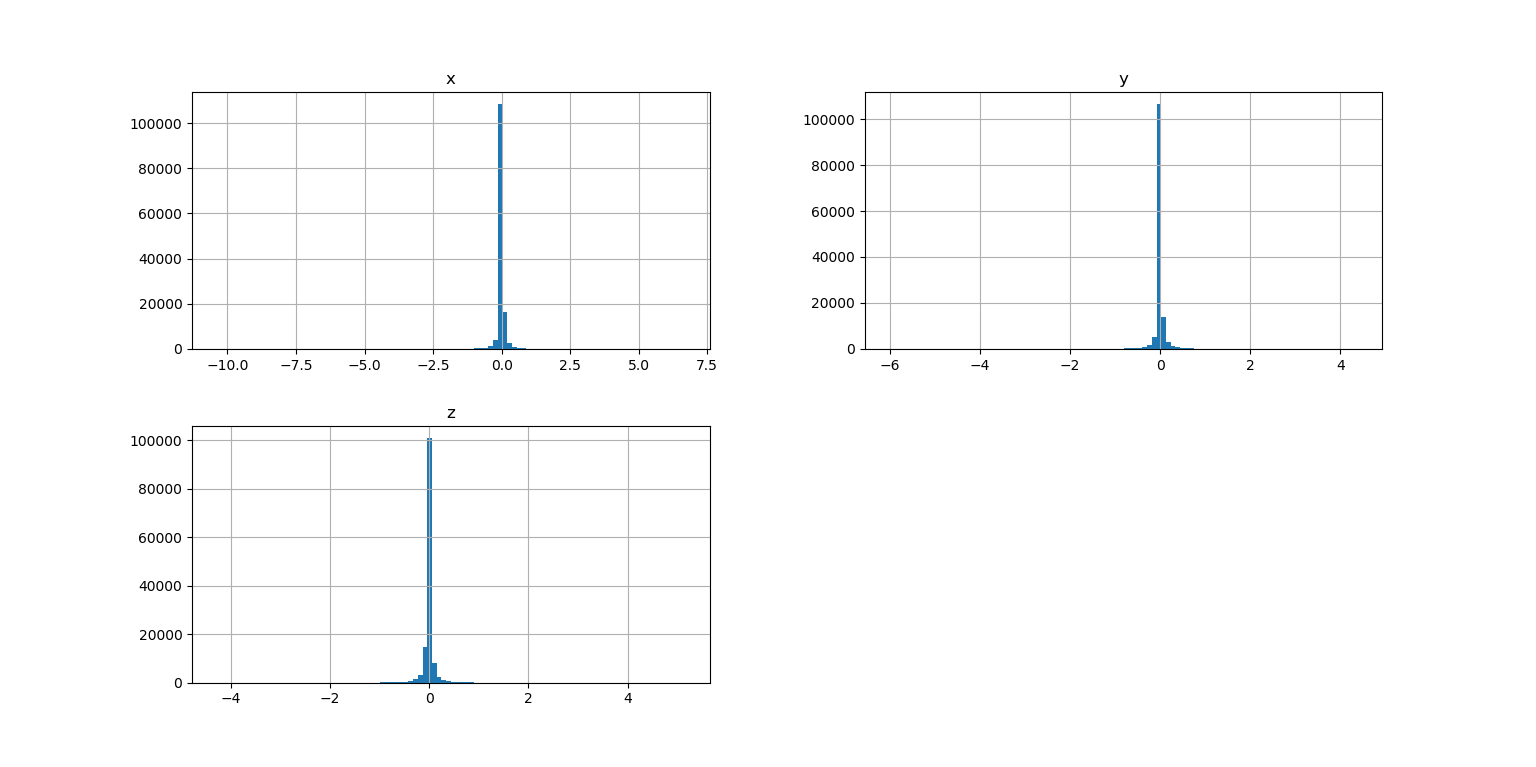


Gyroscope:

抑郁症患者：



健康人：

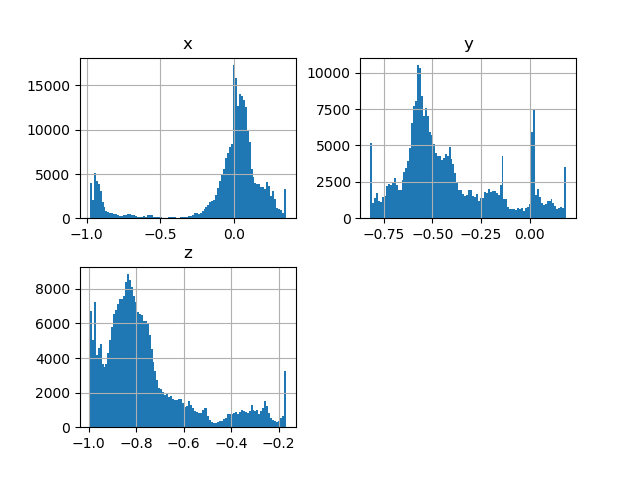


6.噪声数据处理

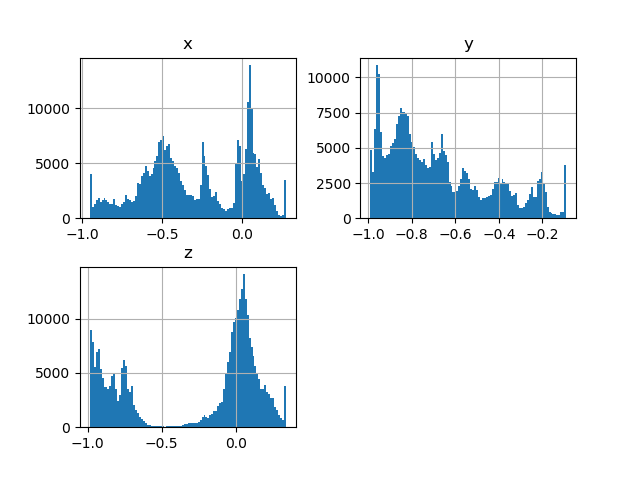
在这么多的数据里，噪声值其实是不可避免的，但它可能会干扰结果，我查询资料发现可以将连续变量均值上下三倍标准差范围外的记录替换为均值上下三倍标准差值。通过运用此类方法，我对数据进行了处理。最后得到经处理后的分布图：

Accelerometer:

抑郁症患者：

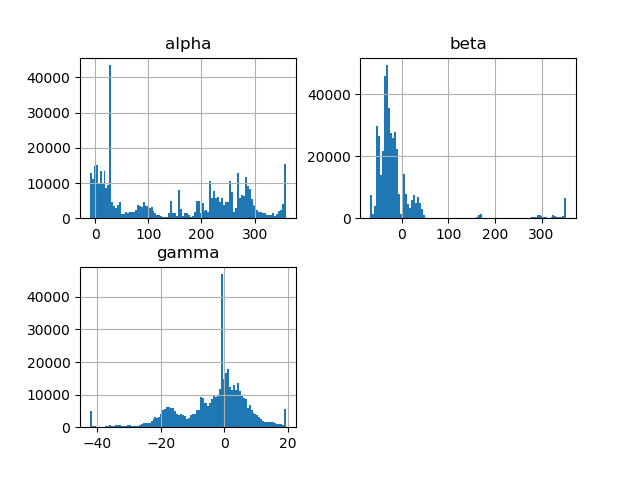


健康人：

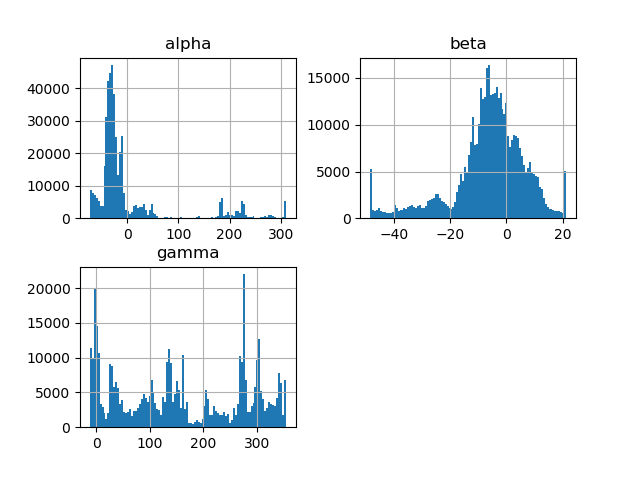


Device\_motion:

抑郁症患者：

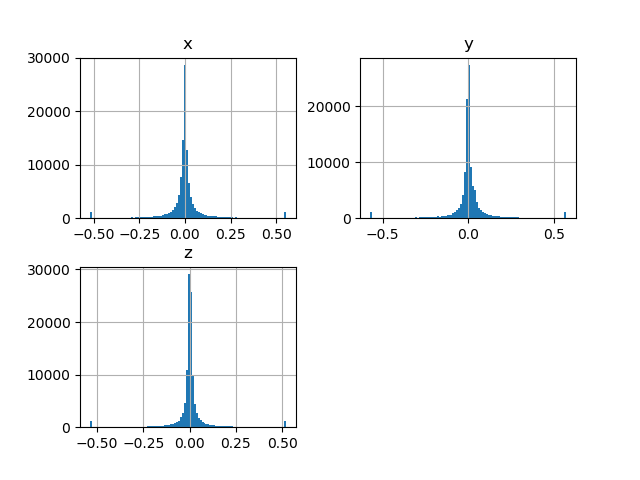


健康人：



Gyroscope:

抑郁症患者：



健康人：

