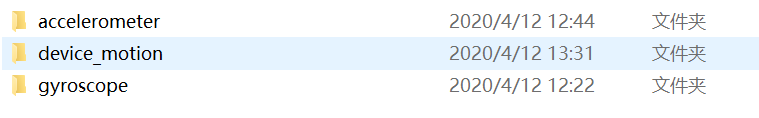
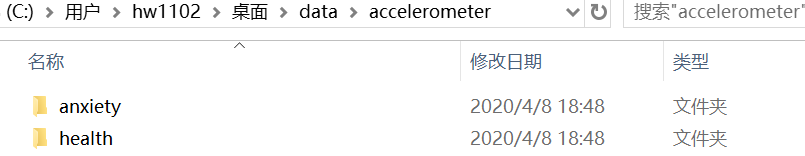
数据预处理文档

1. 数据的下载

可以使用爬虫程序进行下载（附下载程序源码），但是使用爬虫下载会丢失原有的标签（anxiety和health），且处理起来会比较麻烦，于是选择了IDM下载器进行下载，支持数据的批量下载，且可以对原有的标签进行保存，于是可以获得3类传感器数据以及各自目录下的anxiety和health的数据。

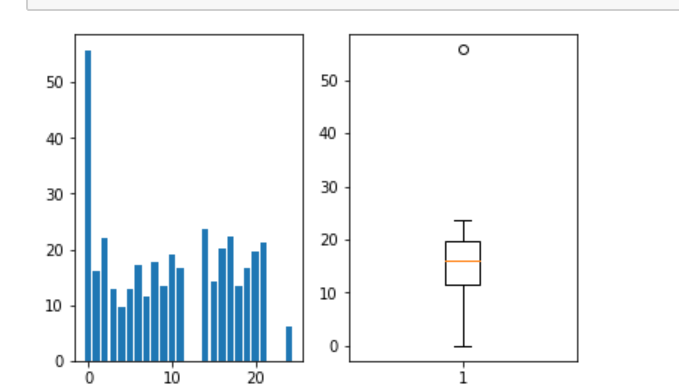




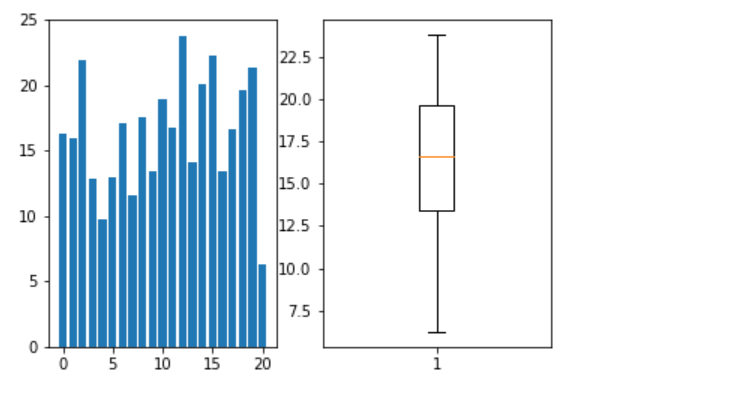
二．数据预处理

1.建立两个数组，一个df数组为时间数据，用于计算答题时间，every\_plt数组为dataframe数据，用于子图的绘制

2.通过绘制条形图和箱型图，观察异常值和离群点（以gyroscope/anxiety为例）

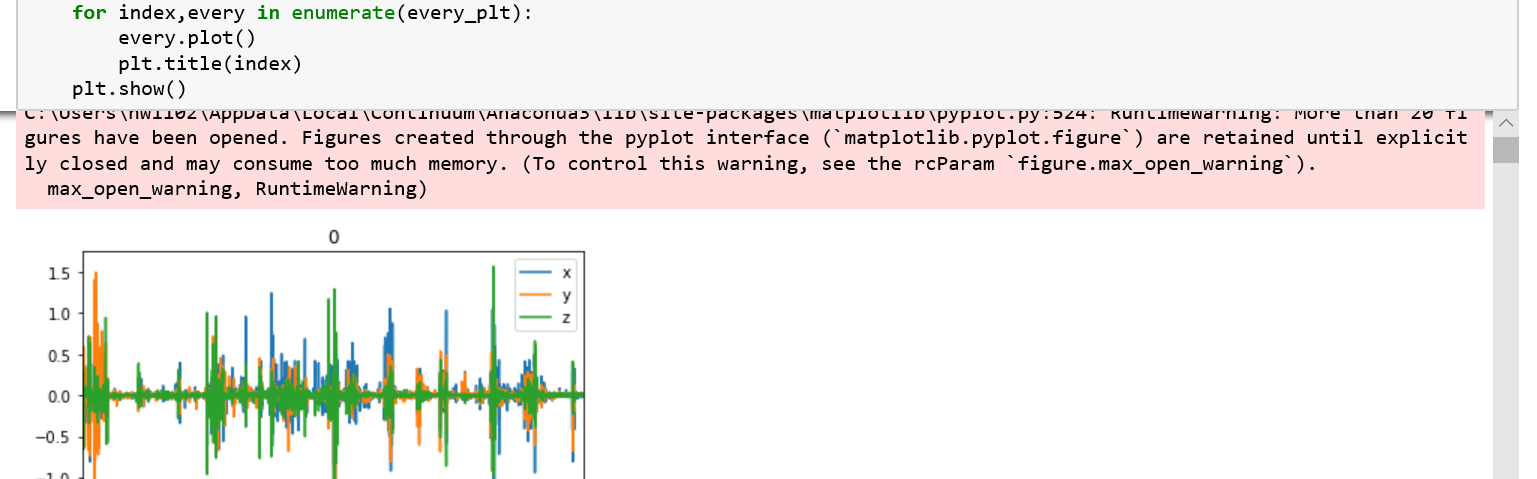


对时间序列的异常值可以选择删除，也可以选择用平均值进行修正。（对于空值我选择进行删除，同时对every\_plt数组的也进行相应的删除，对于第一个极大值我选择用平均值进行修正）

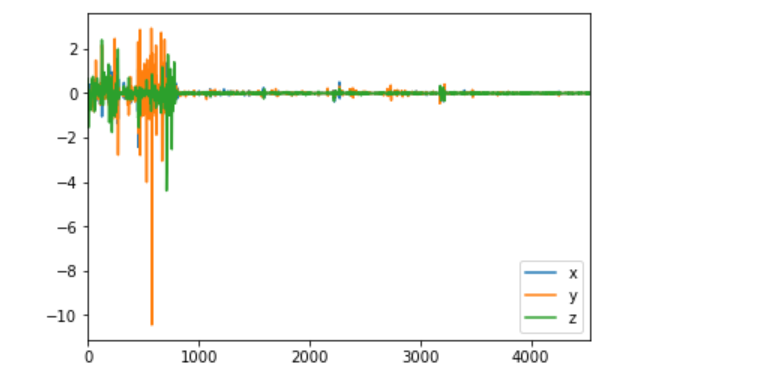


可以得到一个比较均匀的时间序列，为以后的分析时间对健康与焦虑的判断保留较为准确的数据。

3.观察该文件下的子图（以gyroscope/health为例）

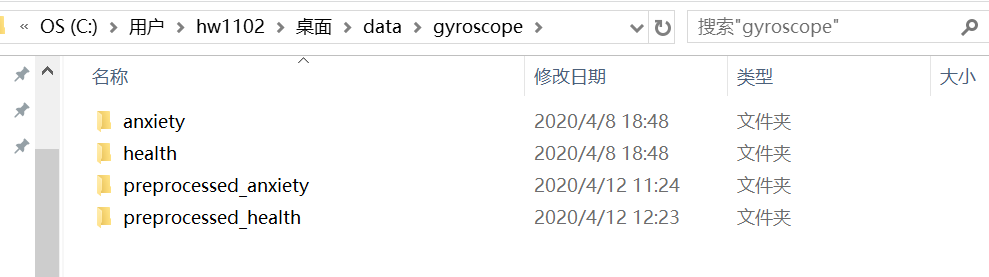


可以发现，有些虽然不是空值，但是也是异常值，如下图：

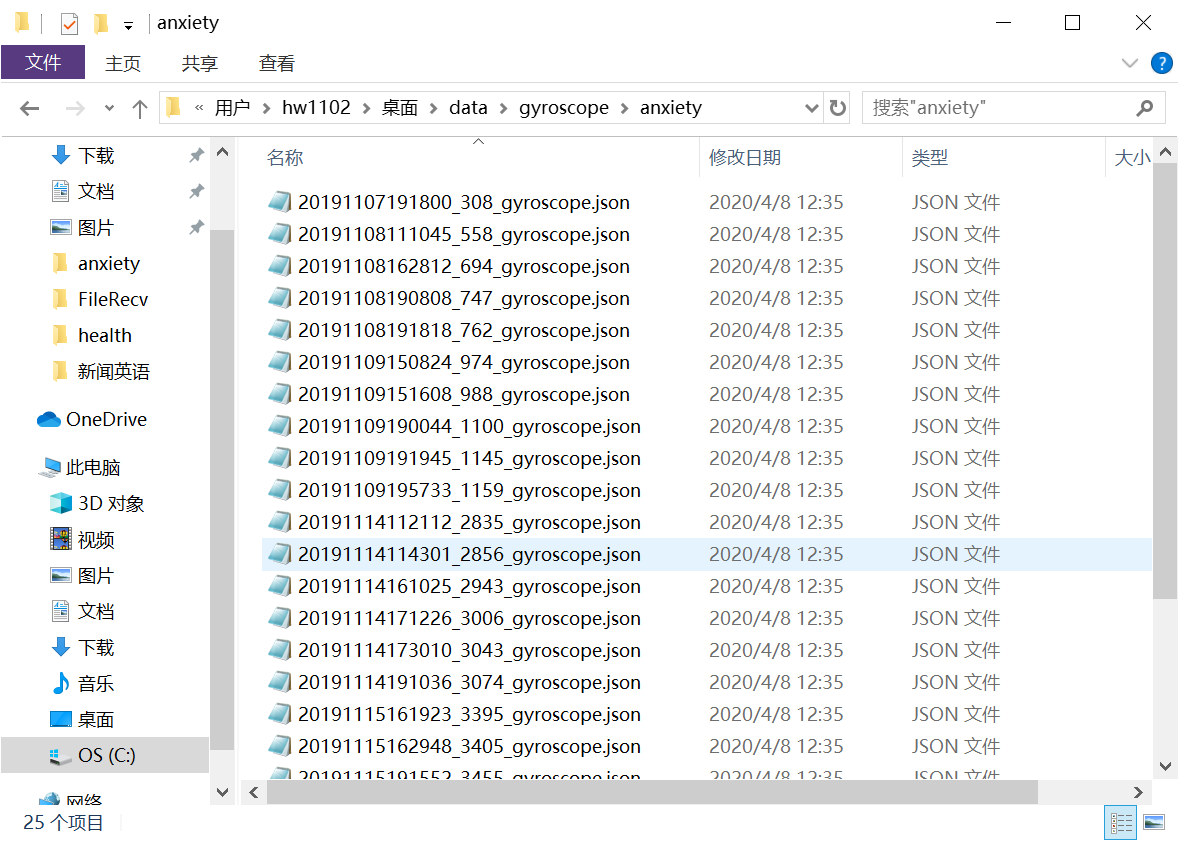


该图可能是在开始答题后将手机置于座子上，属于噪声数据，予以删除。

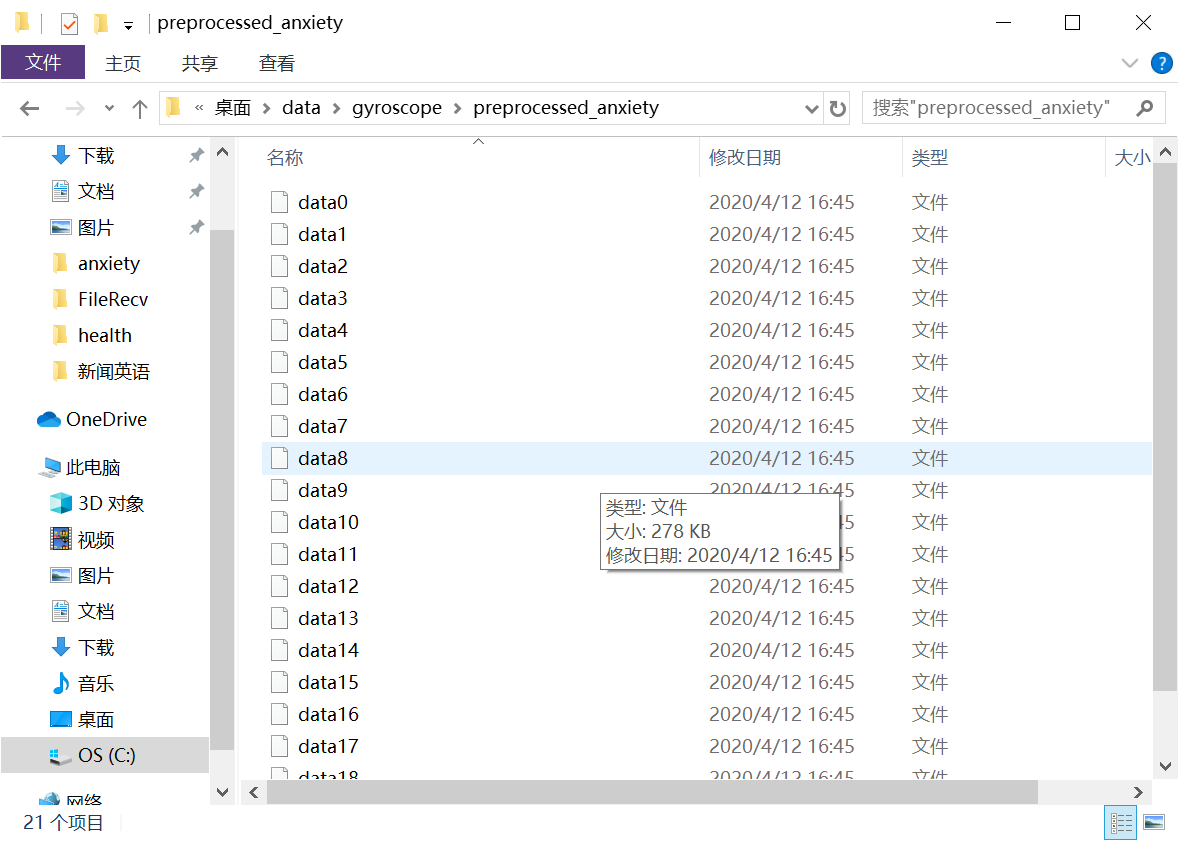
4.通过对df以及every\_plt数组进行删除和修正后，可以得到预处理后的数据，我们再对其进行保存。可以看到预处理后的数据如下。



同时，通过对比可以发现，我们在preprocessed\_anxiety中删除了原anxiety下的四个json文件



图表 1预处理前



图表 2预处理后