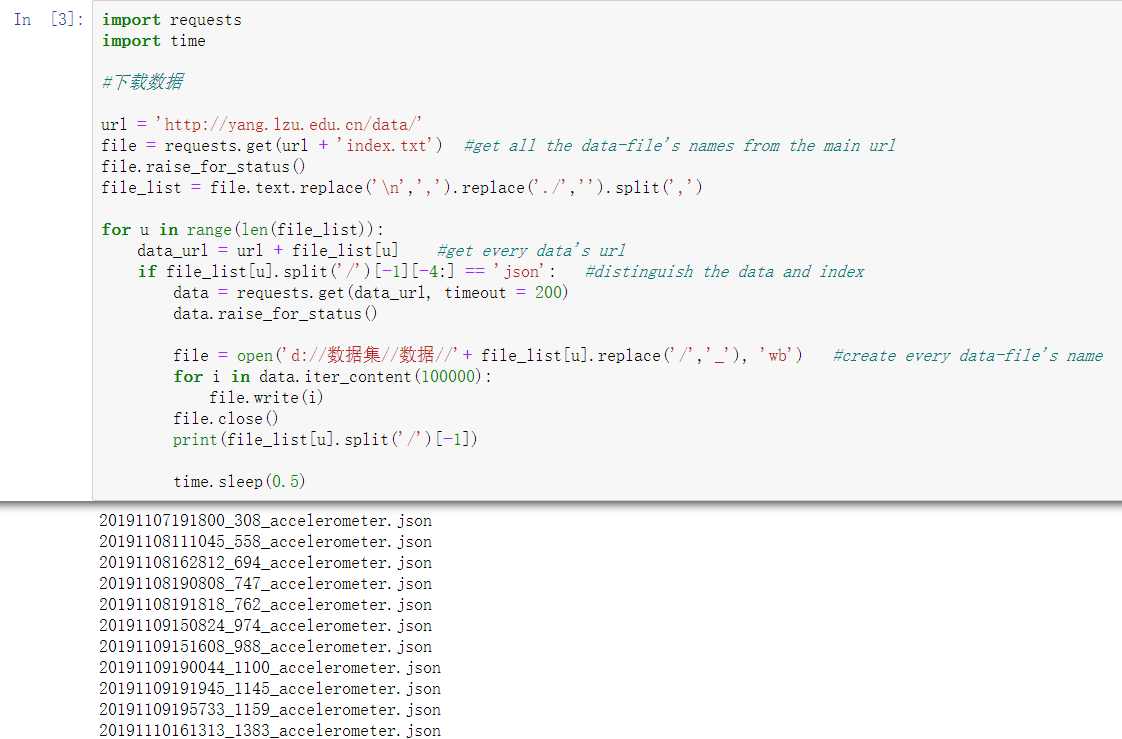
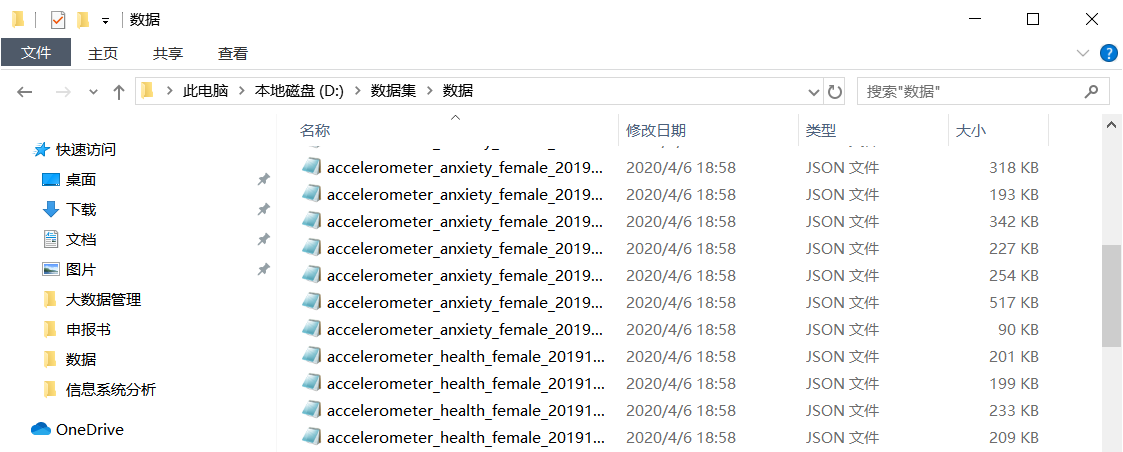
**数据预处理**

**卢音迅 320180941460**

1. **获取数据**

利用requests库从主网页获取各个数据页的路径，经过处理后再利用循环对每个数据文件进行下载并写入到本地文件，并对三种文件分类命名。





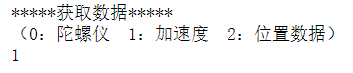
1. **数据清洗**
2. 填写时间
3. 查找异常

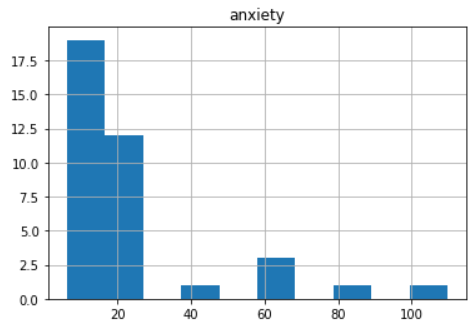
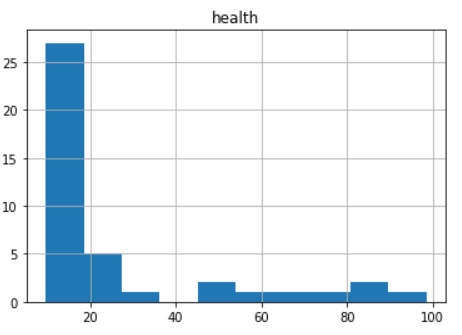
由于传感器1s收集五次数据，所以问卷填写时间（min）为数据量除以（5\*60）。因为是整体数据，故需要导入所有数据文件判断。代码可根据键盘输入获取想要的数据文件（陀螺仪、加速度、移动位置），三种数据文件分别有健康和焦虑两种人单独的数据。

health--data1 anxiety--data2

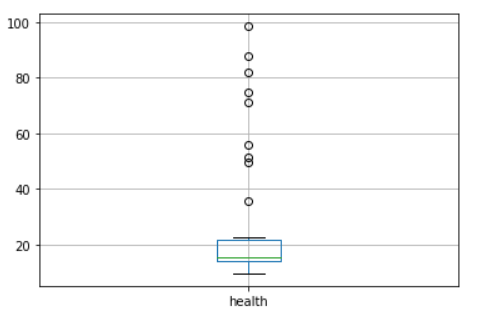
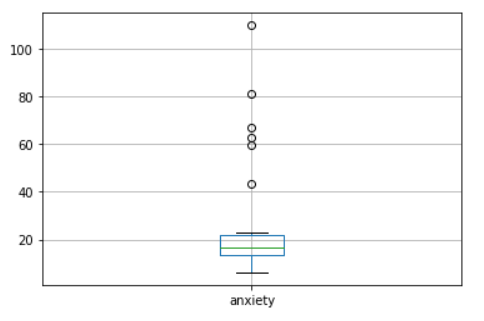


假如获取加速度数据的时间，可发现大部分人的问卷填写时间为10-30分钟，但有少数人甚至达到了两个小时，猜测属于异常型数据，为了更加精确的判断是否异常，可利用箱形图具体分析。



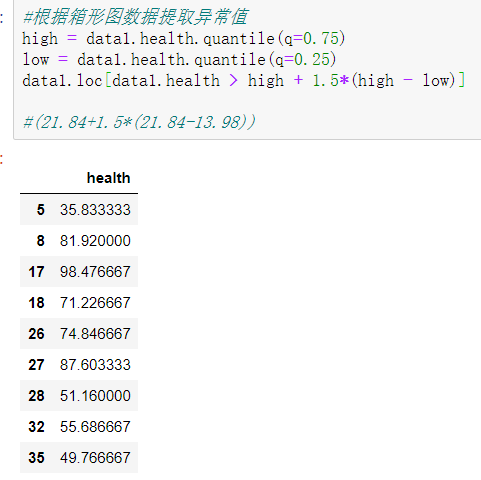


以下为accelerometer数据中时间的箱型图，可观察到大量的异常值处于箱型图上边缘线的上方。箱形图为我们提供了识别异常值的一个标准：异常值被定义为小于Q1－1.5IQR或大于Q3+1.5IQR的值。该图以一种相对稳定的方式描述数据的离散分布情况。

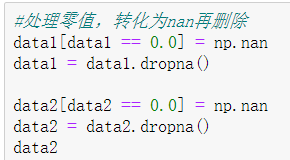
 

1. 处理异常

为了删除异常点，需要根据箱型图中的数据处理，可用到describe方法或者quantile方法。获取四分位数，再计算上下边缘线，获取并剔除异常值。由本次数据观察到只有上边缘线以上有异常值。

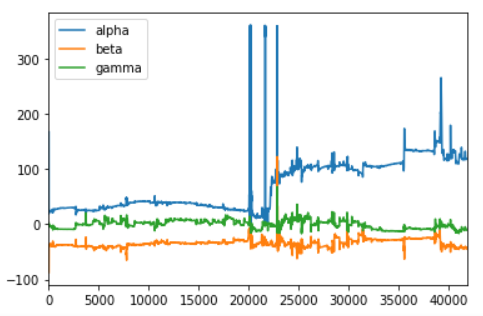
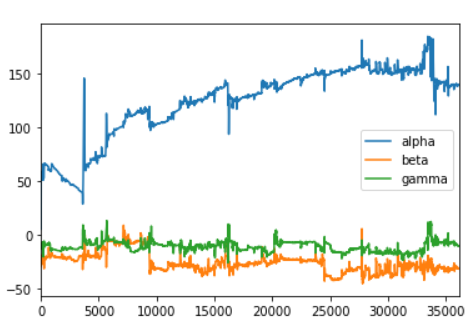
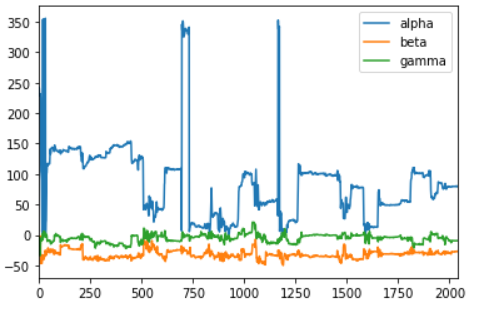


数据文件中还存在大量的零值，估计是部分数据出现丢失（退出到后台），由于有些数据有时波动较大，不可预测数据值从而导致不能为之赋默认值。而数据量整体也相对较大，只可将该行数据删除，实际操作转换为numpy中的nan再进行删除。

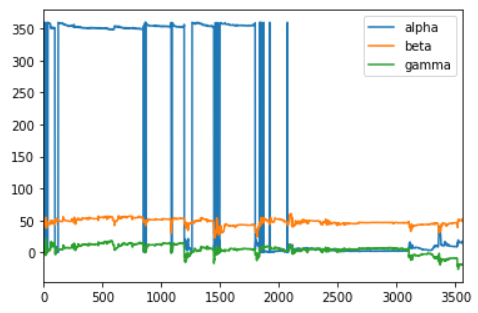
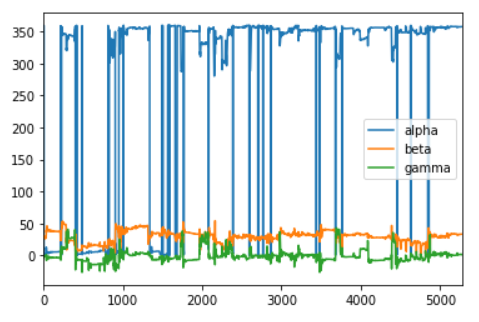
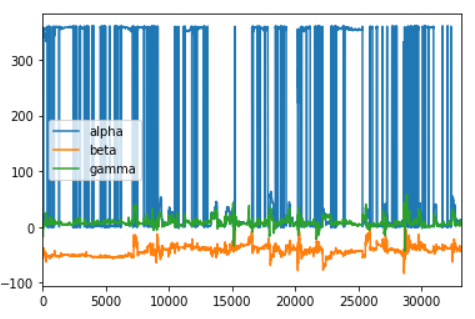


1. 数据波动
   1. 观察数据

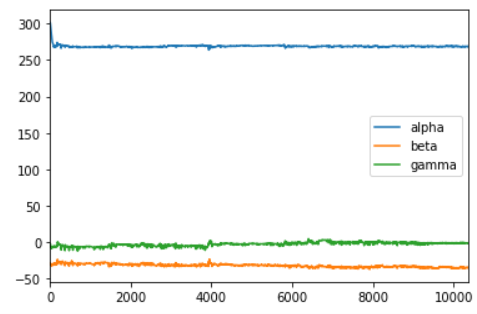
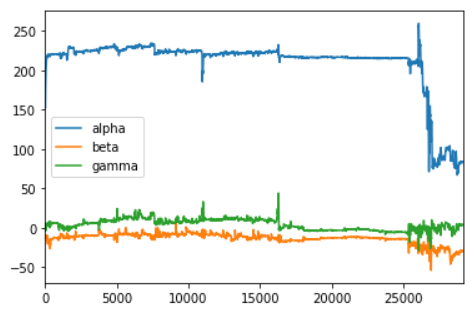
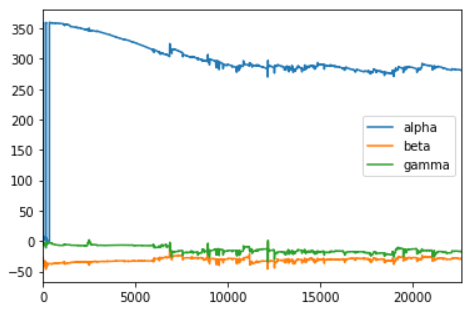
由于accelerometer和device motion均是收集手机运动数据，故随机提取device motion观察每个数据文件的折线图，发现有些数据alpha方向大幅度波动较少，而有些则变化大幅度波动较多。猜测巨大波动原因为数据丢失（中途退出）导致归零，而较少的则是人为突然运动导致。也有少量数据几乎只有在开始期间有较大波动或者几乎没有波动，估计是测试者在开始答题前（答完前）变换坐姿或者把直接把手机放在桌子（支架）上填写问卷，而近乎平缓的数据意义并不大。

大幅度波动少

大幅度波动多



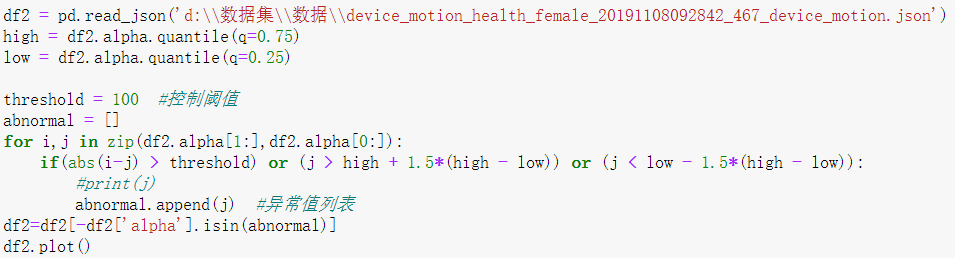
最开始/结尾有波动或者几乎没有波动

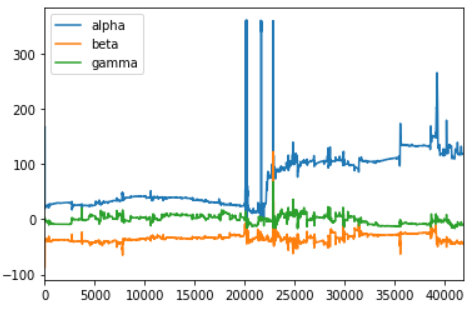
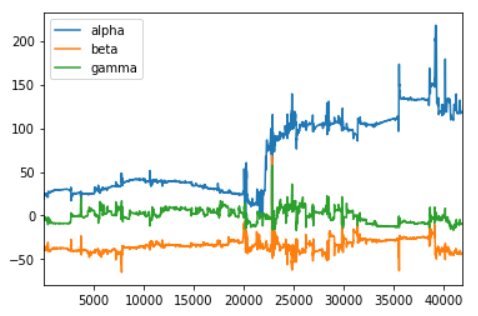
* 1. 处理空值和波动

可能因用户手机没有该硬件设备的问题，数据文件中有部分空DataFrame文件，故遍历数据路径下所有文件，输出名字并删除空文件。

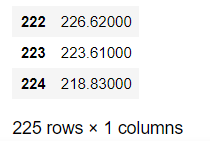


大幅度波动较少时（包括开始结尾有较大波动的数据），可利用箱形图异常处理方法，去除低于下边缘线和高于上边缘线的数据，相邻数据差值太大时可设置一个阈值（100）。最后将所有符合数据添加到一个列表，并筛选出不包含异常数据的正常数据作图。

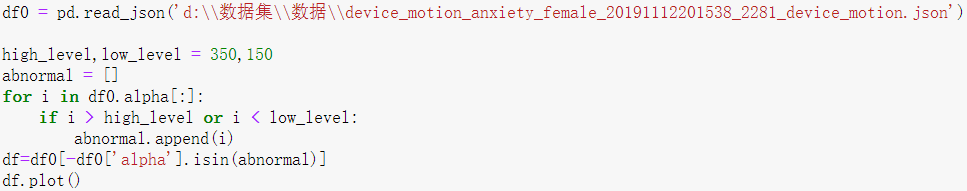


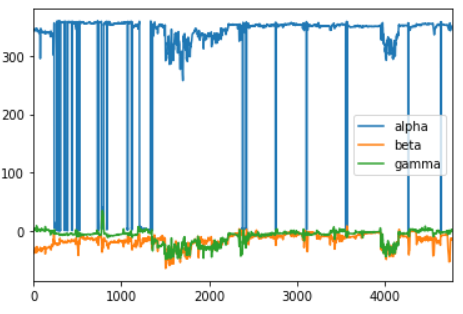
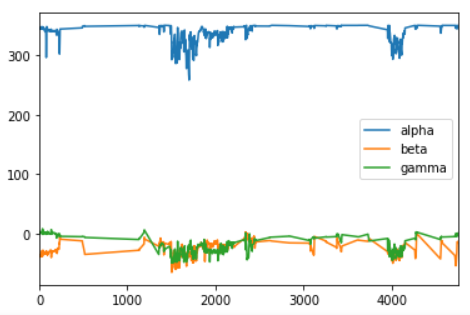
 

一共获取了255个异常数据



大幅度波动较多的时候，可根据数据情况直接采用固定阈值（350，150）去除异常数据，因为太多的异常数据会导致四分位数偏离正常值太远，采用箱型图计算不可靠。



共获取了3305个异常数据

