Homework10 预处理说明文档（处理过程说明详见 homework10.ipynb 文件）

一、数据含义

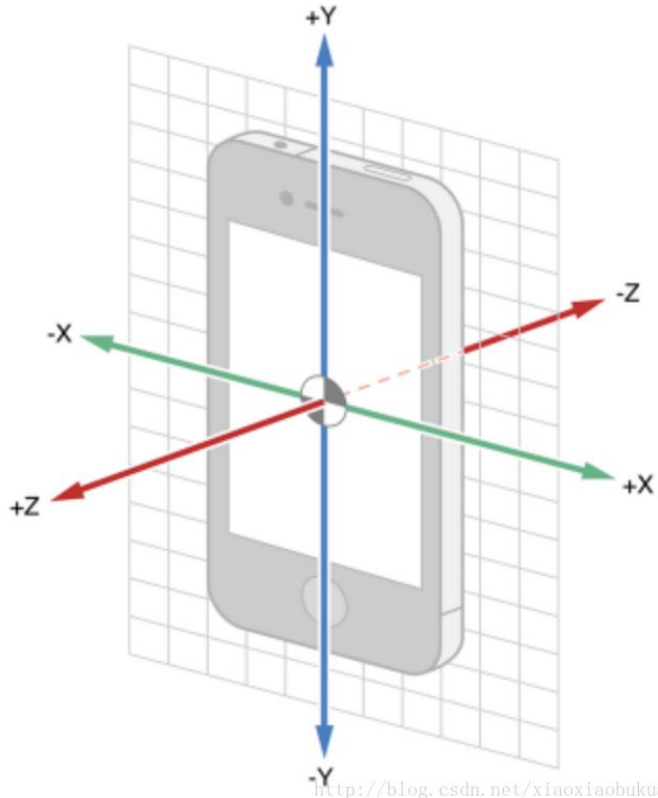
这些数据是参与者回答问卷时记录的手机传感器的数值。

1、accelerometer、device motion 和 gyroscope 文件夹分别是什么?

在回答问卷的时，手机同时收集 accelerometer，device motion 和 gyro 三个传感器的数据（有的手机传感器数量不同，可能会缺少某个传感器），每个传感器的数据分为抑郁组和健康组，其中每个文件是一个参与者的记录。

记录的频率是 5Hz（每秒 5 次），所以根据记录的数量可以得到时间（但没有时间戳）。

2、accelerometer 、device motion 和 gyro 的 x, y, z 代表什么意思?



accelerometer ≈ device motion 三轴加速度传感器，含义是沿着 x, y, z 的加速度大小（如上图）

gyro 陀螺仪，xyz 代表绕 x,y,z 轴的转动速度

3、为什么会出现”稳定的加速度”?

并不是手机在做“匀加速运动”，而是重力加速度 g 的分量，在不同轴上的投影大小不一样。

设备握持的角度不同 --> x,y,z 取不同的值

由此猜想，加速度传感器|x|, |y|, |z| 的高低和握持姿势的关系：

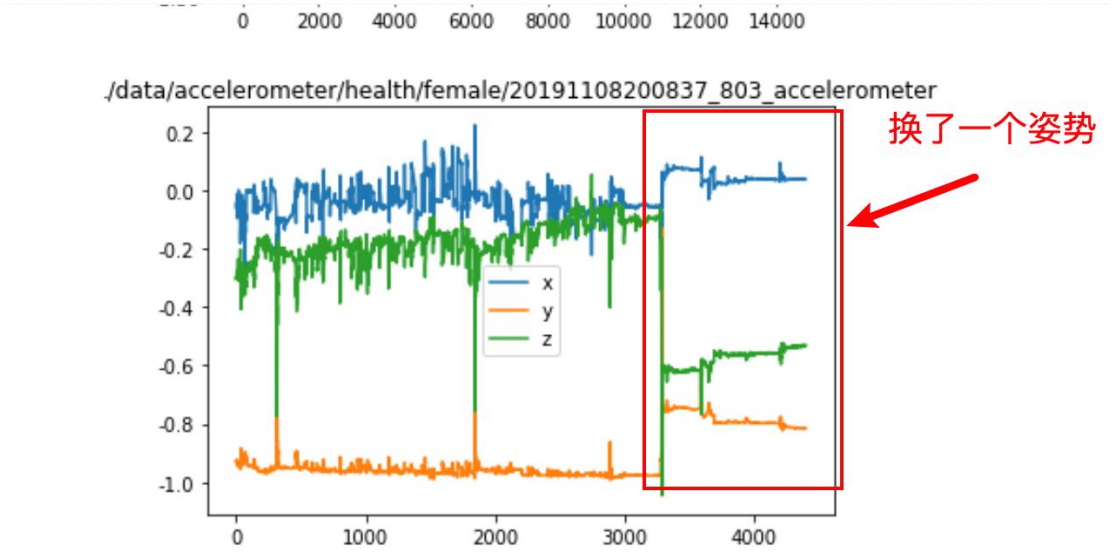
|x|大：横着拿（打游戏的姿势）

|y|大：竖着拿

|z|大：接近平放在桌面上

x, y, z 围绕一些定值上下波动 —> 以某个姿势拿着手机，然后手有所晃动
波动幅度越小 = 拿的越稳

观察发现，志愿者答题时可能还会换姿势。



二、预处理效果

处理前，文件的内容是[{'x':1, 'y':1, 'z':1}, {'x':2, 'y':, 'z':2}, {'x':3, 'y':3, 'z':3}]

处理后，文件内容变成：[[1, 1, 1], [2, 2, 2], [3, 3, 3]]，并且用 pickle 写入文件，这样方便读出来可以直接变成 list 使用。

处理的异常情况：

1、空文件，文件内部是空的，没有记录，如：[ ]

在下载时检测，如果是空的就不下载

```
download 171 files of 199, 5 empty files are ignored:

http://yang.lzu.edu.cn/data/gyroscope/anxiety/female/20191114161025_2943_gyroscope.json
http://yang.lzu.edu.cn/data/gyroscope/anxiety/female/20191114171226_3006_gyroscope.json
http://yang.lzu.edu.cn/data/gyroscope/anxiety/female/20191116154726_3651_gyroscope.json
http://yang.lzu.edu.cn/data/gyroscope/anxiety/female/20191118173012_3937_gyroscope.json
http://yang.lzu.edu.cn/data/gyroscope/health/female/20191111111559_1665_gyroscope.json
```

2、文件内部存在缺失值：

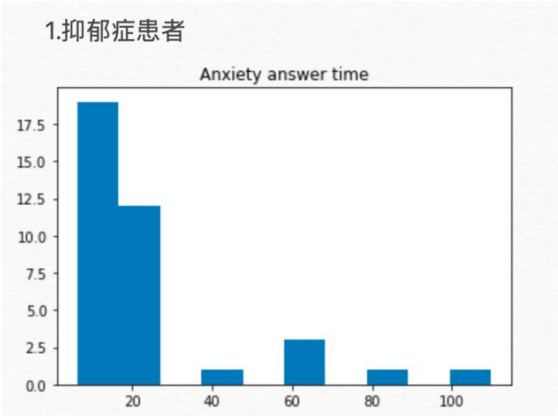
处理：统一替换成相邻的值

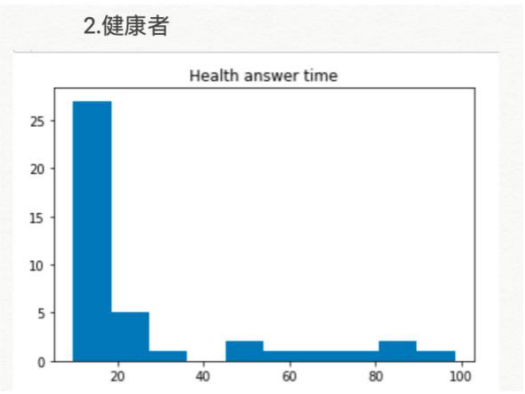
原因：平均值，中位数会受到波动的影响，用相邻的一个值填充的话影响会比较小

```
def inner_preprocessing(df):
    # 空值，缺失值-->加速度用相邻的前一个值来填充
    return df.replace([None], np.NaN).fillna(method='ffill').fillna(method='bfill')
```

3、文件记录的时间过短或者过长

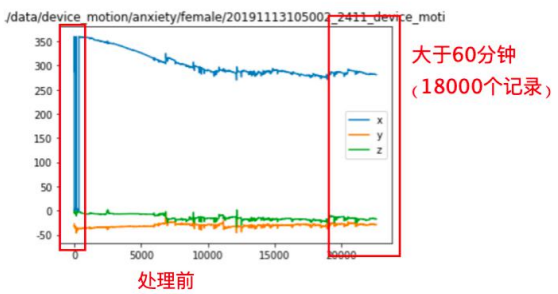
问卷的题目有 200 道选择，什么样的答题时间算正常时间呢？根据频率直方图：

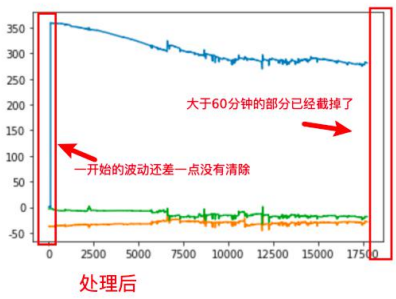




发现大多数答题时间集中在 10~20 分钟，答题过长或者过短都是异常的情况，

所以我如果发现时间小于 8 分钟，或者大于 60 分钟的文件，则把在范围之外的数据截掉（直接删除文件的话我觉得会导致剩下的数据太少，不利于分析）



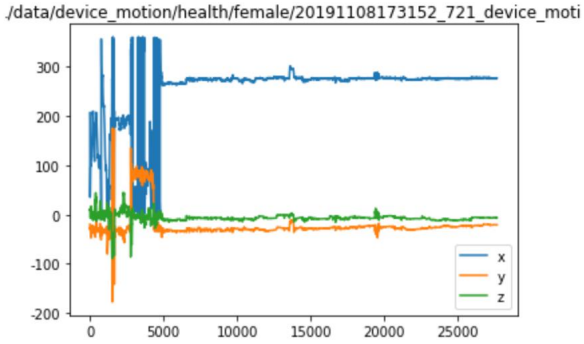


4、文件内部变化幅度过小

处理：删除三个传感器文件夹中的全部同名文件。因为当参与者全程把手机放在桌子上答题时，他的三个传感器数据都是没有意义的，所以不应该保留。

```
30
-----droped-----
```

从device motion发现数据是异常的




```
23 abnormal files are deleted:
```

意味着这个参与者其他传感器数据也失去了参考意义，需要删除

```
http://yang.lzu.edu.cn/data/accelerometer/anxiety/female/20191109195733_1159_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/anxiety/female/20191110161313_1383_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/anxiety/female/20191111092136_1572_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/anxiety/female/20191113152518_2535_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/anxiety/female/20191115103948_3262_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/health/female/20191108151946_633_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/health/female/20191108173152_721_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/health/female/20191109151723_992_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/health/female/20191109151742_994_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/health/female/20191109165514_1058_accelerometer.json
http://yang.lzu.edu.cn/data/accelerometer/health/female/20191110114303_1310_accelerometer.json
http://yang.lzu.edu.cn/data/device_motion/anxiety/female/20191109195733_1159_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/anxiety/female/20191110161313_1383_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/anxiety/female/20191111092136_1572_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/anxiety/female/20191115103948_3262_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/health/female/20191108151946_633_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/health/female/20191108173152_721_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/health/female/20191109151723_992_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/health/female/20191109151742_994_device_motion.json
http://yang.lzu.edu.cn/data/device_motion/health/female/20191109165514_1058_device_motion.json
http://yang.lzu.edu.cn/data/gyroscope/anxiety/female/20191109195733_1159_gyroscope.json
http://yang.lzu.edu.cn/data/gyroscope/health/female/20191108173152_721_gyroscope.json
http://yang.lzu.edu.cn/data/gyroscope/health/female/20191109165514_1058_gyroscope.json
```

5、文件动作一开始变化很大，但是后来明显小了很多

（可能是参与者一开始站着，后来坐下答题的情景）

根据观察，当把所有文件前 300 个数据（前 1 分钟）舍弃的时候，可以把大多数一开始的波动舍弃掉，虽然有少数开始波动时间比较长的的不能完全删除，但是能改善很多。