数据预处理说明文档

注：代码中的注释比较清晰，下面只做简单说明。
程序说明：Data_spider.py 是爬虫程序，Data_process.py 是数据预处理的程序
一、函数具体说明

① load_data()　读入数据，用 os 库的 walk 函数去循环读入爬下来的数据。

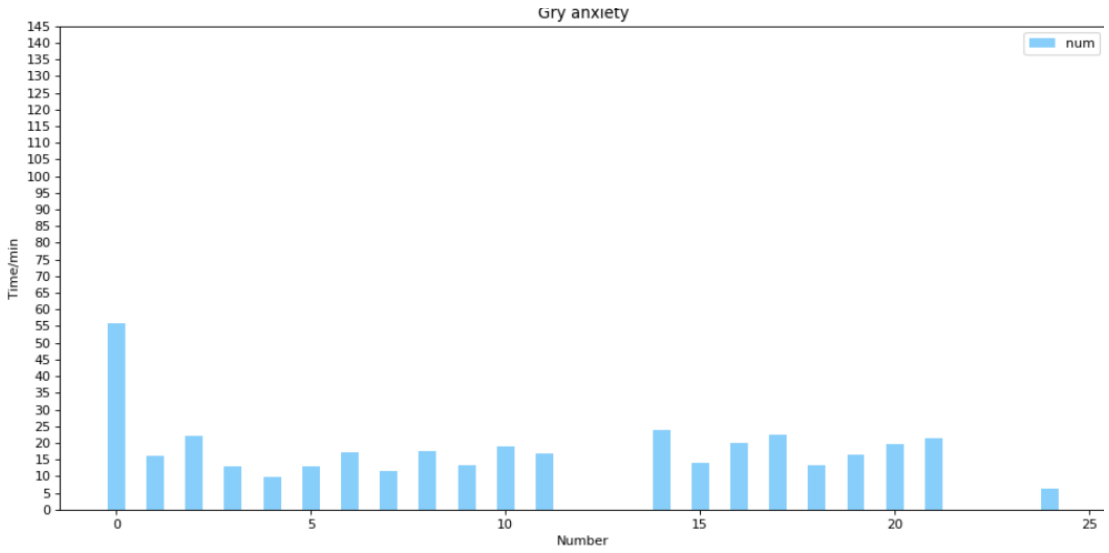
② time_bar_plot() 画出输入的两组的时间柱状图。

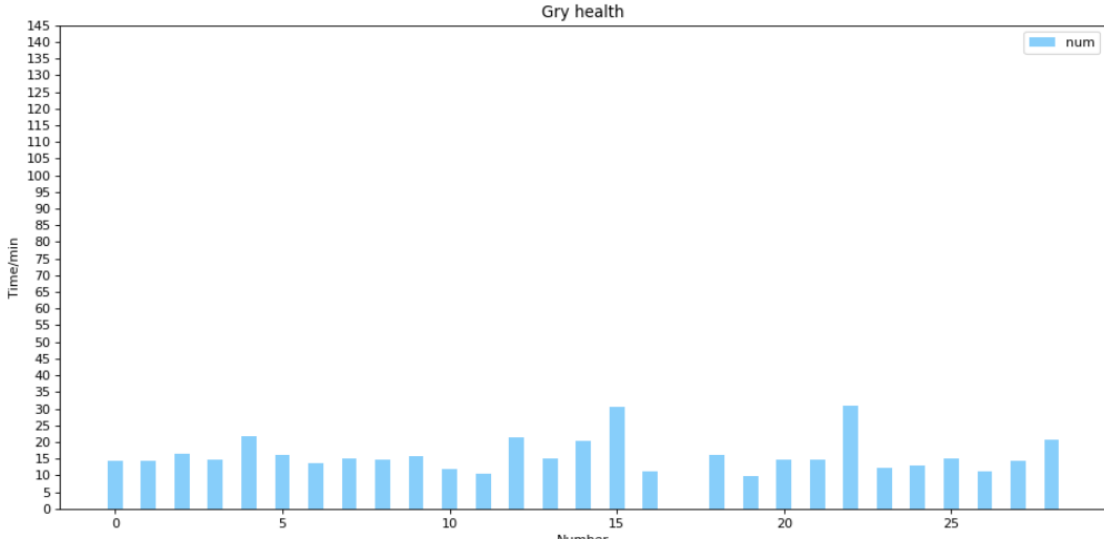
③ domathoftime() 计算时间(5HZ 的采集频率)

④ bar_xyz() 画出 XYZ 三组数据的柱状图

二、预处理步骤：

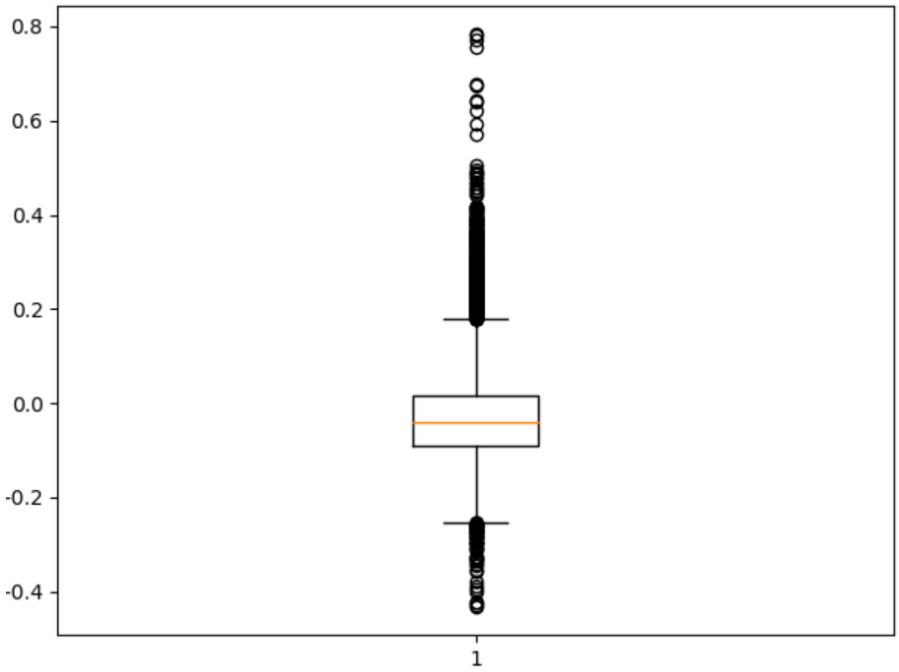
1.找到并处理空值



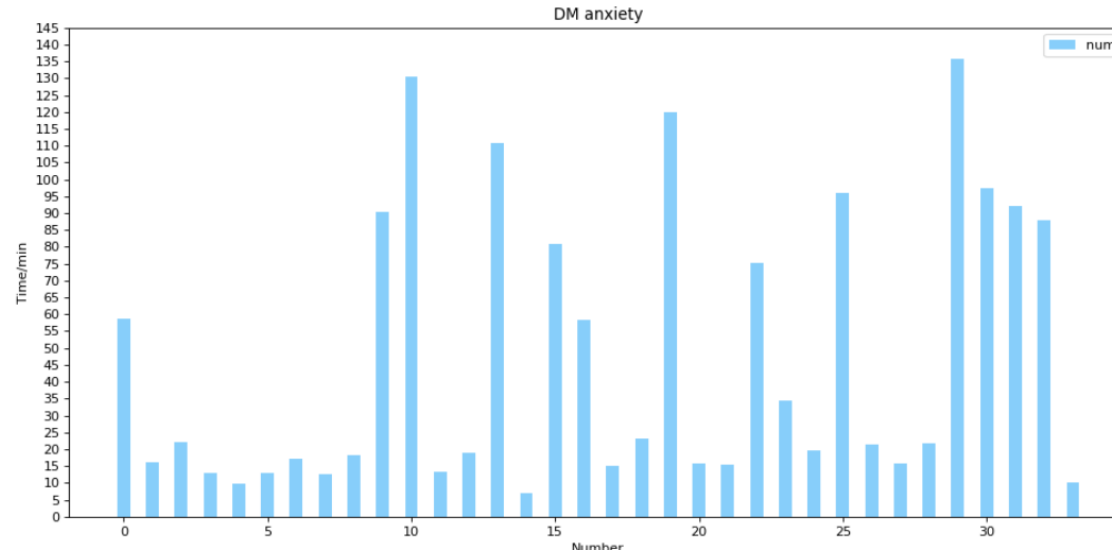


通过画时间图发现空值，没有比较好的补充防止所以删除掉。

2.查看离群点（通过 boxplot 来查看离群点，这些离群点我没有找到特殊的意义，所以还没有删除，需要后续探索）

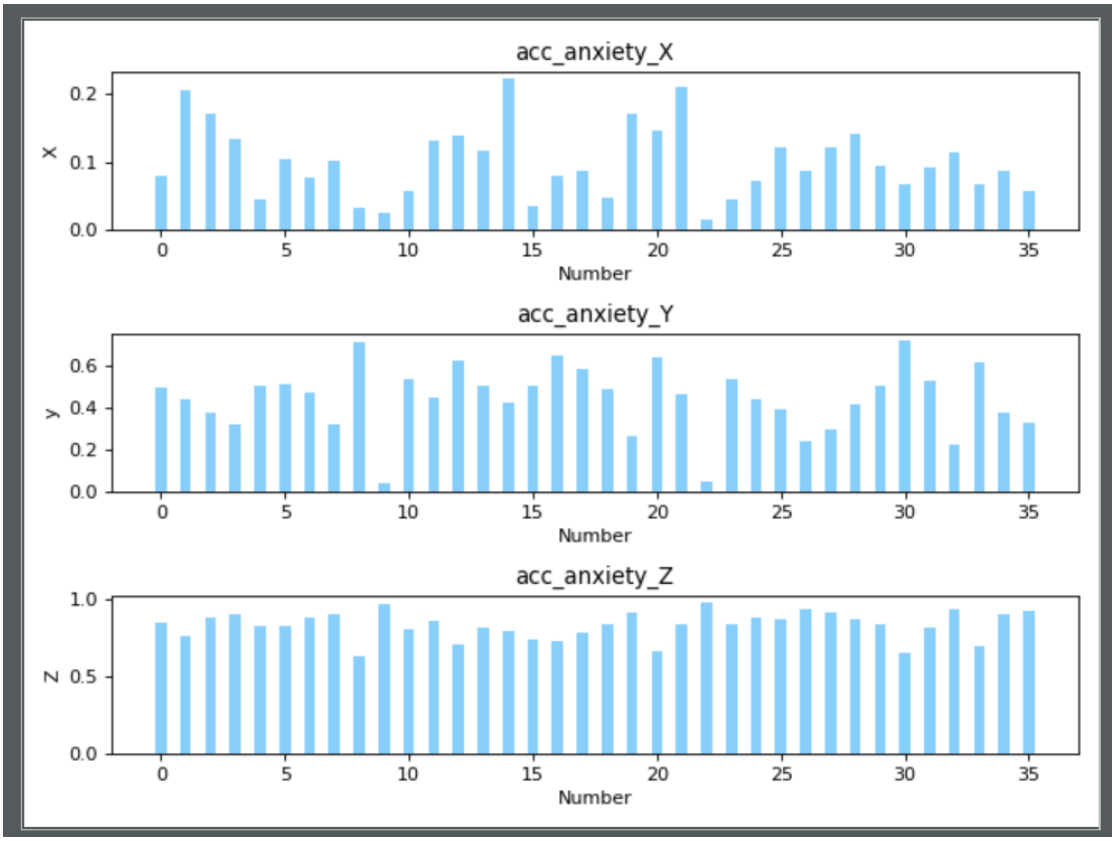


3.根据答卷时间去除异常值

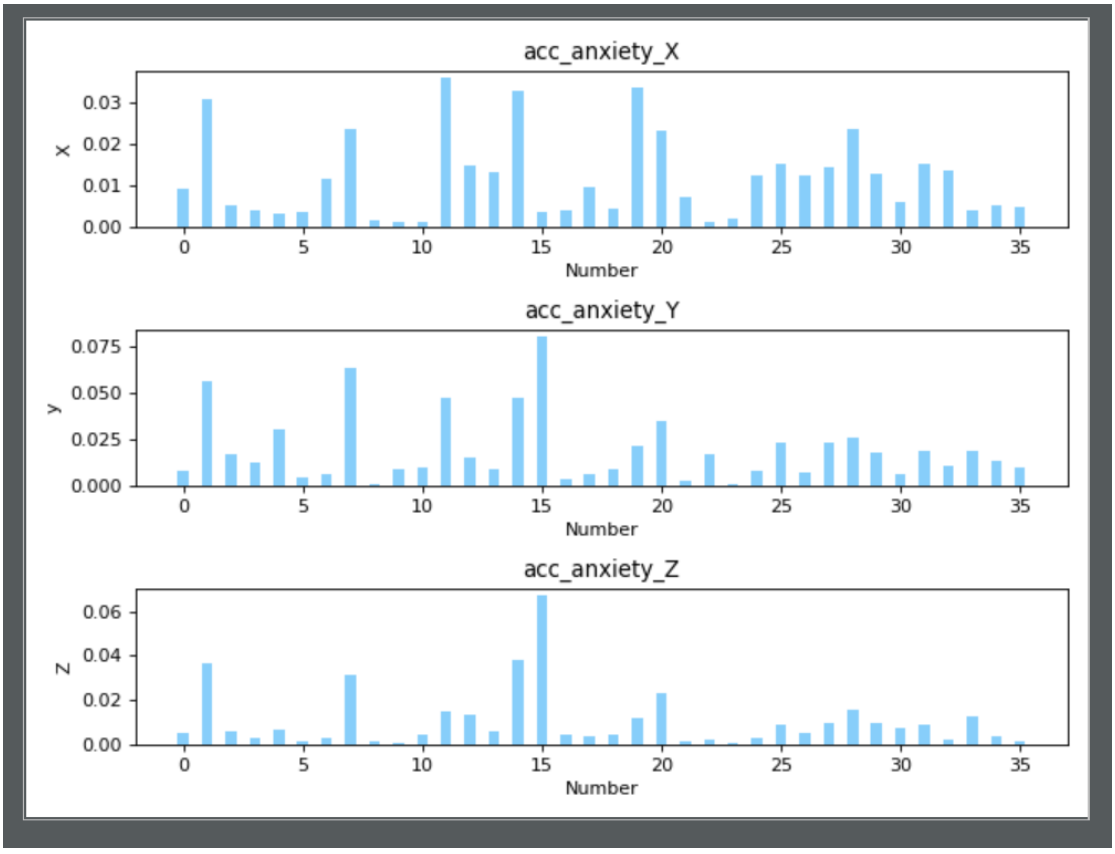


发现有的使用时间很长，有的很短，考虑删除 100 分钟以上的和 10 分钟一下的数据。

4.从两个角度（坐标绝对值后的平均值和方差）判断手机是否放在桌上，这样的数据不准确，可以考虑去除。这样的分析比较适用于 ACC 和 DM 的数据，而不适用于陀螺仪的数据。



这是取绝对值后再取平均值的



这是直接取方差的

5.根据方差，判断站坐情况的改变，这样的数据不准确，可以考虑去除。（这一步没有画图，只用语句来判断大小）

三、删除

通过上述分析，整理出需要删除的数据。

四、数据回写或者继续分析处理

总结：

前两类数据是类似的通过三位坐标的分析，可以比较好的分出噪声数据。而陀螺仪的数据不同，了解了陀螺仪的数据意义后、我仍然没有找到比较好的分析陀螺仪数据的办法是什么，但是我删除了陀螺仪中变化幅度一直很大的数据，因为这不太符合我们对手机的使用。我怀疑是手机可能摔过传感器出现问题，导致检测转动太过于敏感。总之，对于陀螺仪的数据分析不太完善，对离散点的分析也不完善，需要之后进行学习和改进。