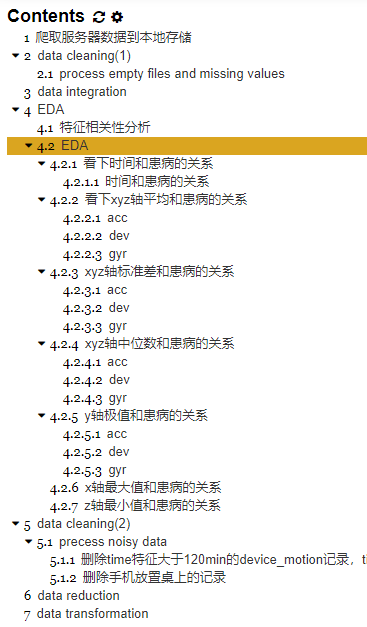
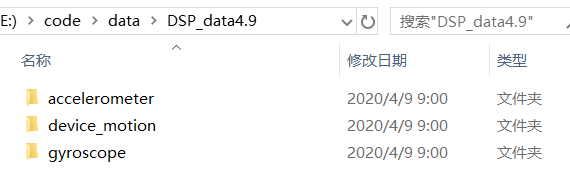
**说明文档**

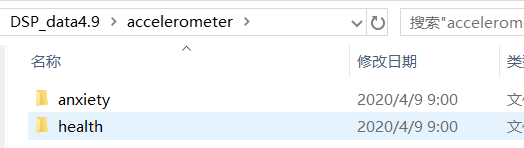
**一．总流程：**



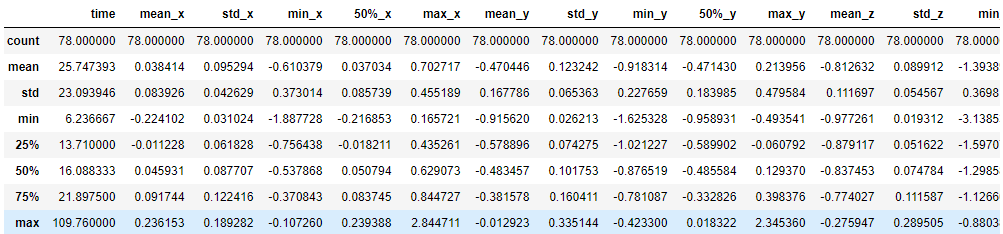
**二．步骤简介：**

1. **数据获取：调用requests库爬取老师服务器上数据。考虑到后期分类的需求，原封不动地按照老师服务器上存储文件的路径存储在本地计算机上**



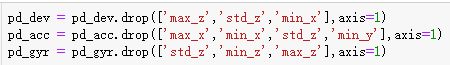


1. **数据清理（1）：处理空文件和有空值数据的文件**
2. **数据整合：读取所有文件并构造以pandas.describe()函数的输出为基础特征的DateFrame，并将同一传感器不同病人分类的DateFrame合并在一起，添加“ill”特征区分。**



17个特征，包含xyz轴数量、平均值、标准差、极大\小值、中位数特征和是否抑郁症‘ill’特征

1. **EDA探索性数据分析：**
2. **查看健康人与病人时间分布**
3. **查看xyz轴健康人与病人平均值分布**
4. **查看xyz轴健康人与病人标准差分布**
5. **查看xyz轴健康人与病人中位数分布**
6. **查看xyz轴健康人与病人极值分布**
7. **查看x轴健康人与病人最大值分布**
8. **查看z轴健康人与病人最小值分布**
9. **数据清理（2）：处理时间time特征（超时）和标准差std特征（手机放置桌上）噪声数据记录，包括删除time特征大于120min的device\_motion记录和大于90min的accelerometer、gyroscope记录，删除std\_z小于0.02的accelerometer记录和std\_x大于3的device\_motion特征**
10. **数据规约：根据相关性分析删除对预测影响不大的特征**



1. **数据转换：我们的数据全是数值型数据，不需要类别转换。其次，是否归一化取决于之后所使用的模型是否使用梯度下降算法，所以展示不归一化**

**三．想法：**

**1.在不了解研究具体流程的情况下，无法分清噪声离群点和非噪声离群点。**

**例如device\_motion的时间分布明显比另外两个传感器的时间长（甚至是同一个编号的人的不同传感器记录都是如此）**

**2.噪声离群点和非噪声离群点的判断依据不能完全靠臆测。**

**例如：数据统计前后，我们都认为抑郁症患者答题速度慢，会消耗更多时间。但在我3次自测时发现，如果有类似抑郁症答题经验，会加快思考速度，从而缩短答题时间**

**3.研究型数据分析有研究目的导向，不跟竞赛一样只以预测结果评价论成败。**

**根据老师课上提到的处理思路（手机置于桌上；时间异常；答题场景限定坐下持手机作答），得出结论：此次数据分析是以坐姿手持设备答题判断抑郁与否为导向的研究（而非竞赛）。所以，天然的噪声判断依据：时间异常数据、非手持手机数据、非坐姿答题数据**