**智能数据挖掘作业3**

19200300004 黄铭瑞

1. **实验目的**

知道数据预处理的原理，掌握数据预处理的方法。重点掌握如插值、去噪、缺失值、离群值处理方法。

1. **实验原理**

数据预处理一般包含数据清洗，数据转换，数据描述，特征选择或特征组合，特征抽取这五个步骤。

数据清洗阶段，主要是处理缺失值，重复值，离群值。

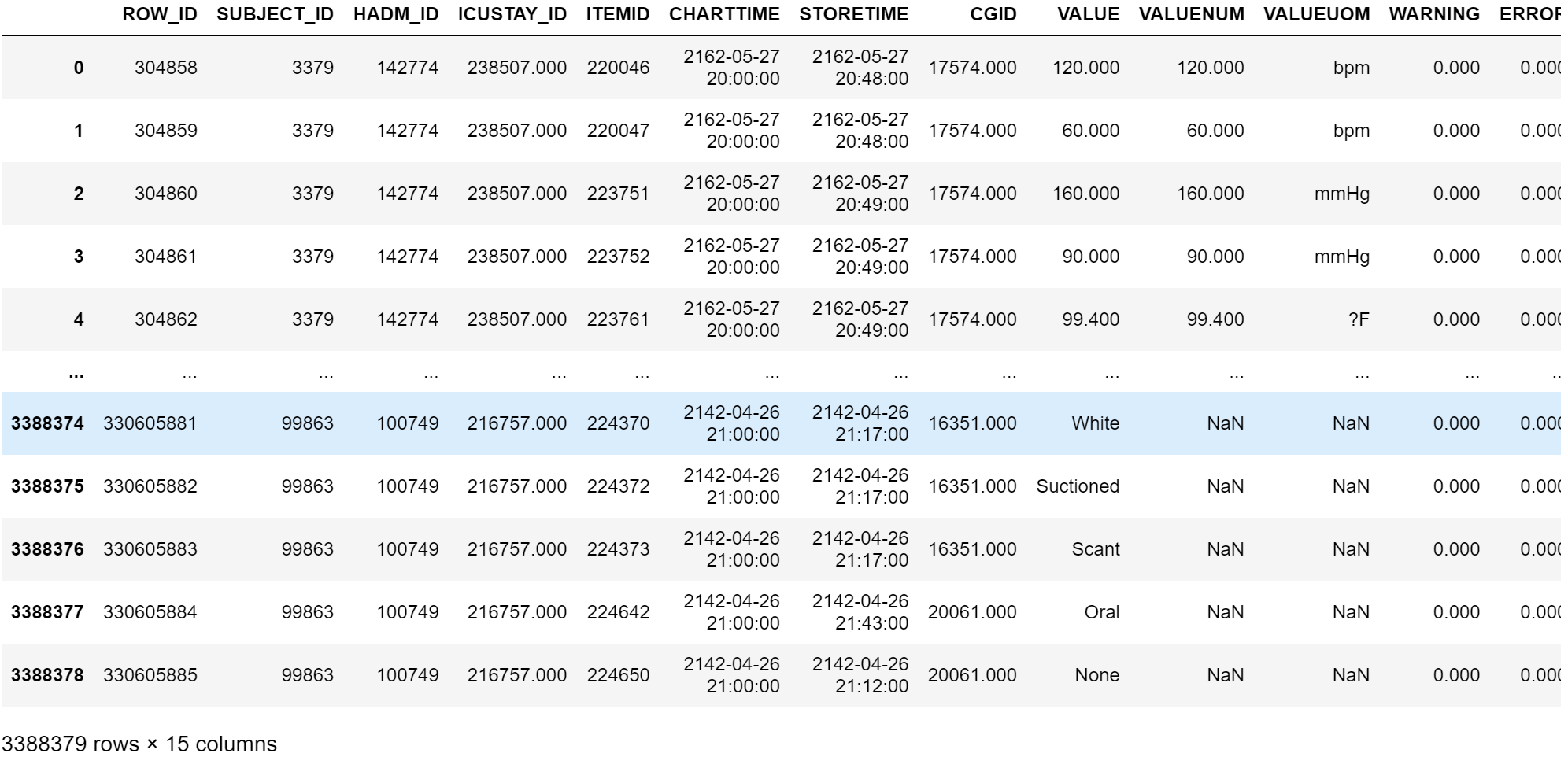
缺失值可以通过删除数据或者比较相邻值来填补数据。

重复值一般采用删除法来处理但有些重复值不能删除，例如订单明细数据或交易明细数据等。

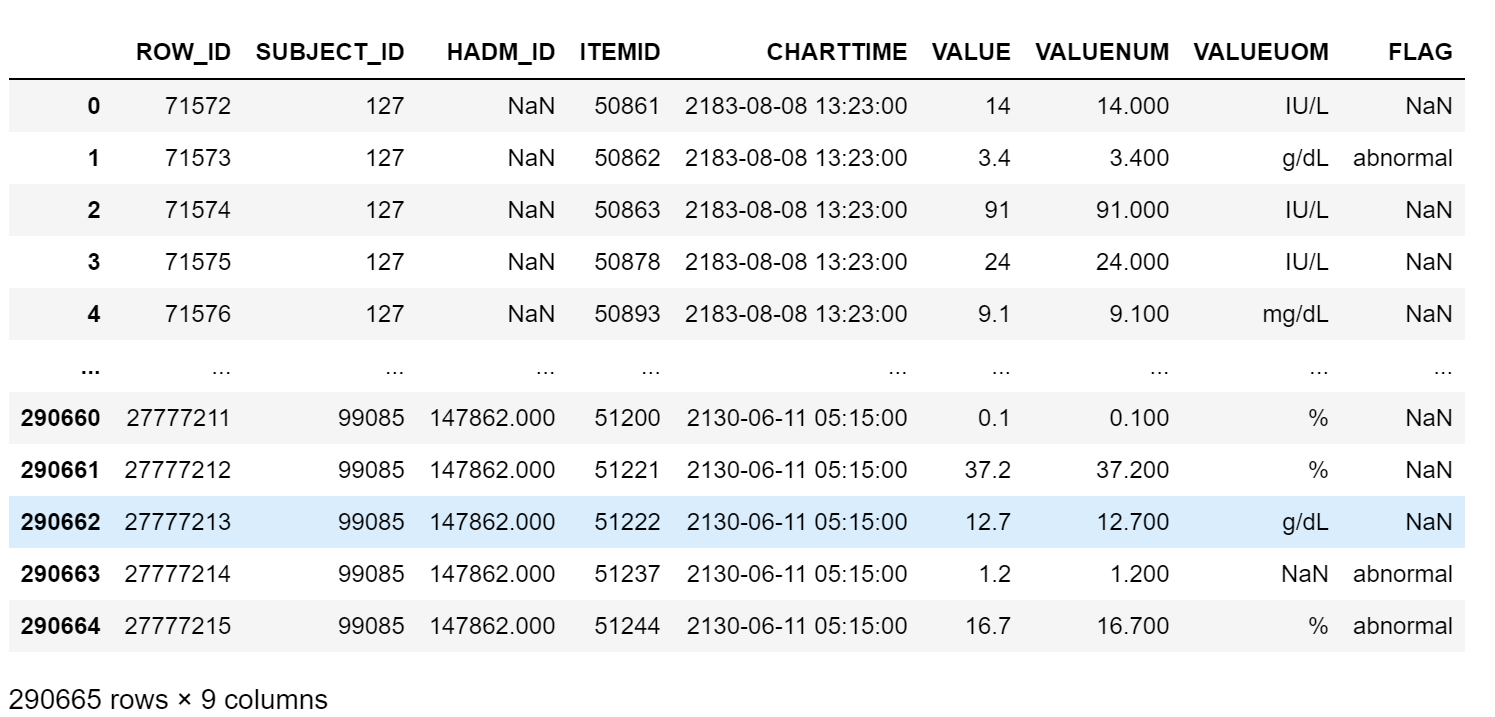
对离群值的甄别，一般使用3σ准则或者箱型图，如果样本是正态分布或近似正态分布，可以考虑使用3σ方法，认为99%以上的数据集中在均值上下3个标准差的范围内；箱型图一般用于统计数据分散情况而做的，超出上下限的值定位离群值。

1. **实验过程**
2. 导入数据集

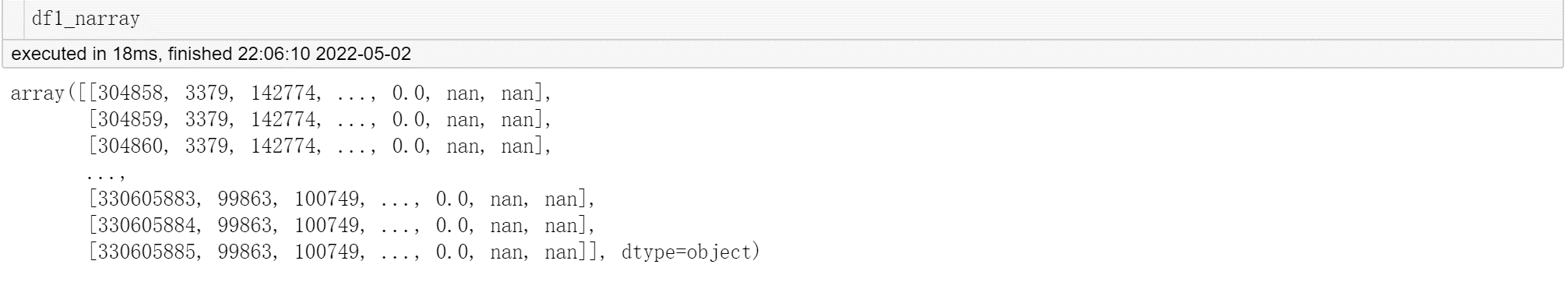
通过pd.read\_csv()来导入mini\_chart\_events.csv与mini\_label\_events.csv文件，并把他们转为N维数组的形式。



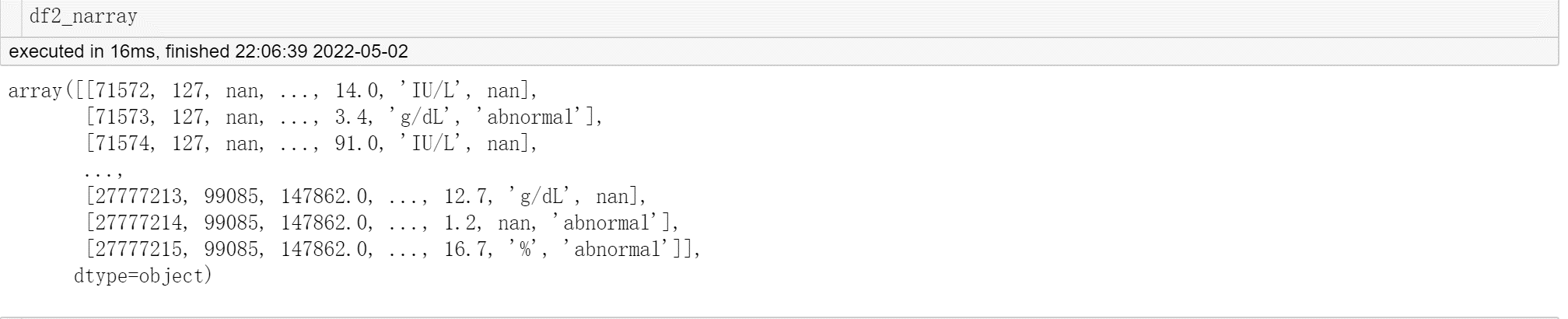
（df1,mini\_chart\_events.csv原始数据）



（df2,mini\_label\_events.csv原始数据）



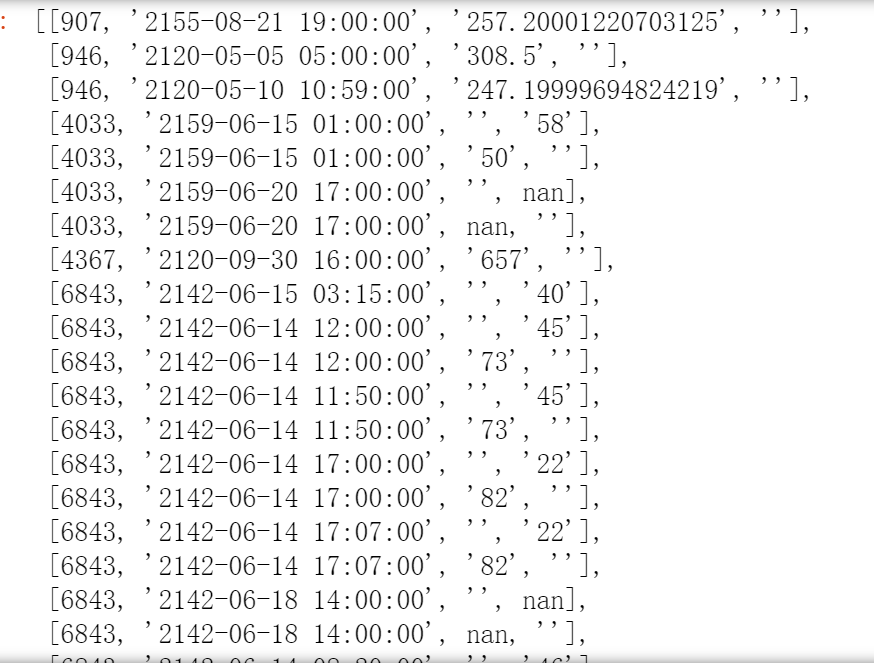
（df1\_narray,ndarray类型的数据）



（df2\_narray,ndarray类型的label events）

1. 提取出需要的pO2,pCO2数据。

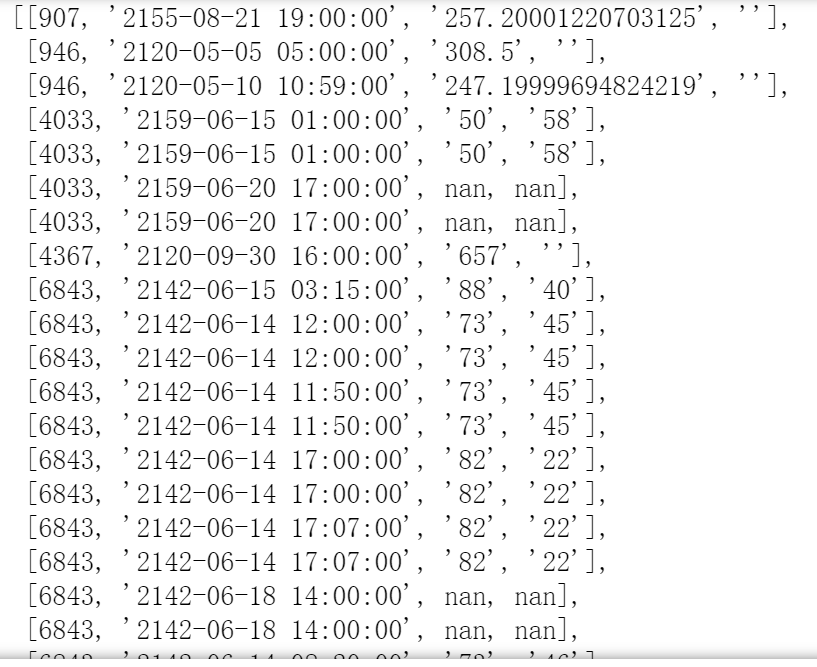
依次把两个文件里的ITEMID为[490, 3785, 3837, 50821]之一，[3784, 3835, 50818]之一的数据提取出来，并放入一个空列表中。



（o2co2data,提取出有用数据的列表）

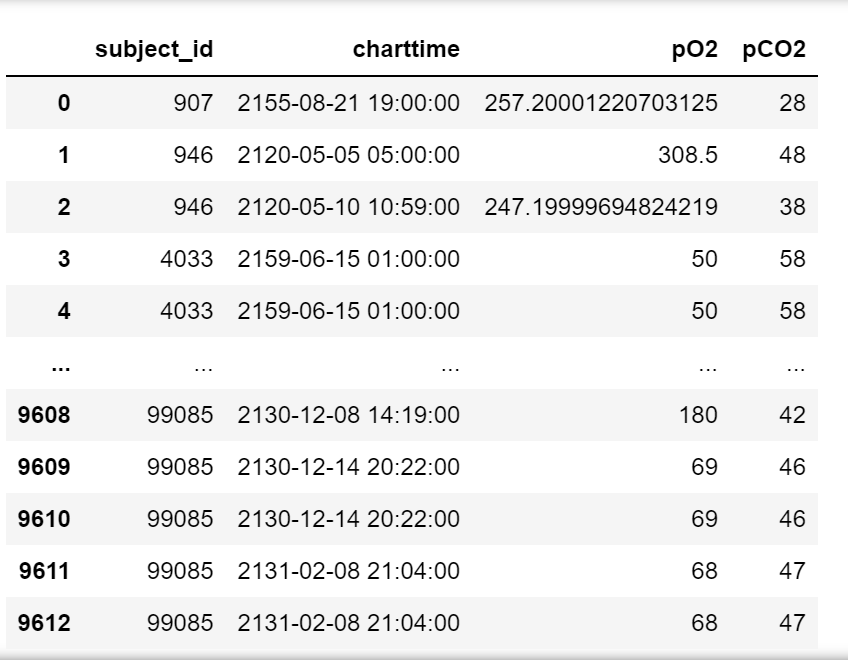
1. 数据变换，保存数据

把pO2,pCO2作为其所含值的列名，构造新的列表。List转为dataframe并以csv形式储存到本地。



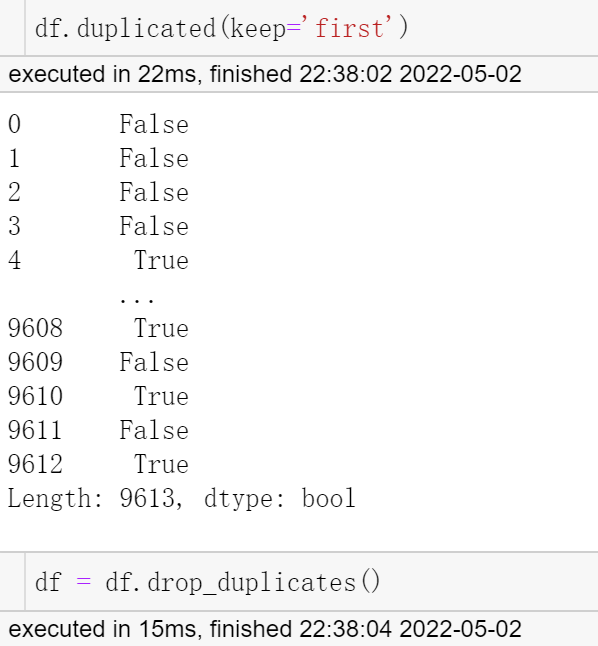
（o2co2data，各列表示为subject\_id,charttime,pO2, pCO2）

1. 读取新保存的csv文件

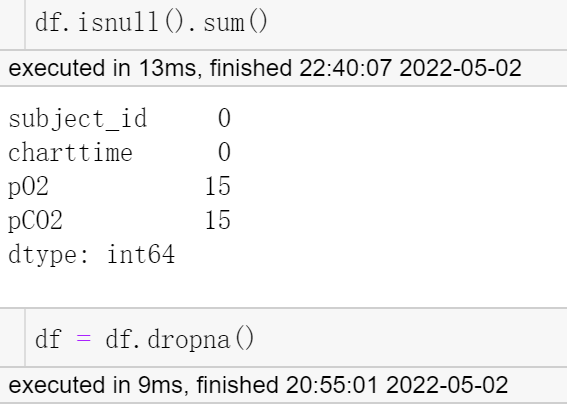


（df,o2co2data.csv数据文件）

1. 查找并处理重复值

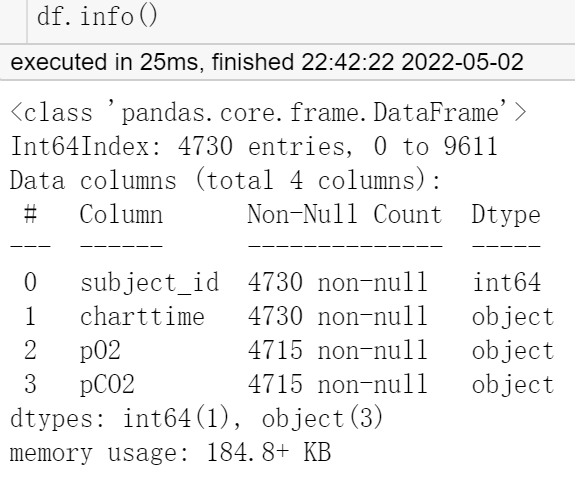


1. 查找并处理空缺值

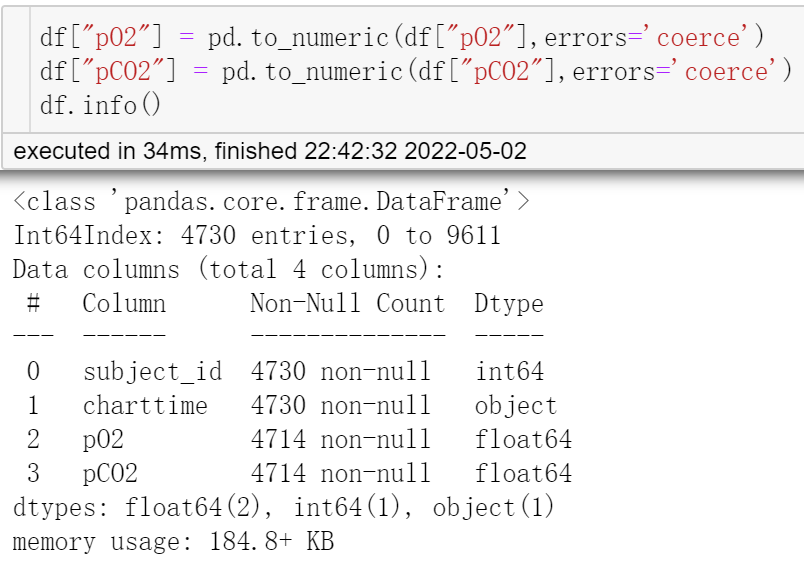


1. 数据类型转换

把pO2,pCO2数据转为float类型



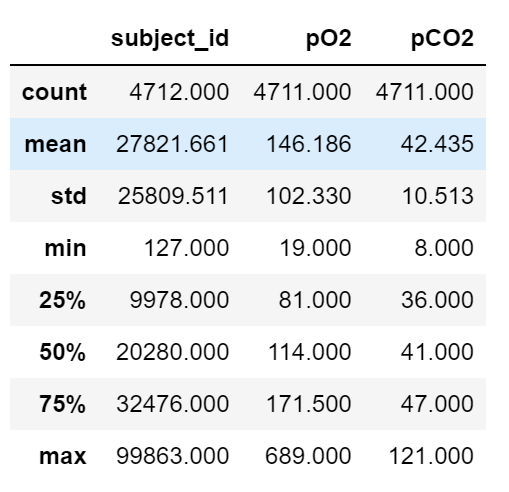
（df原始数据类型）



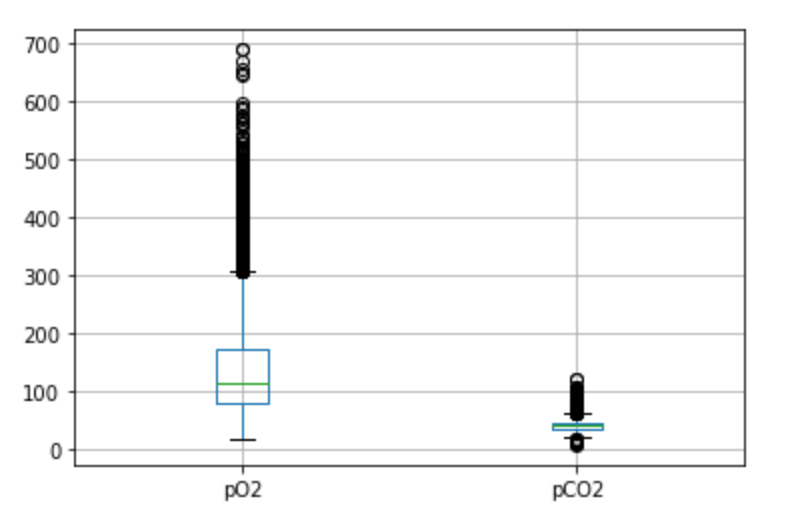
（转换后的df数据类型）

1. 查找离群值

df.describe()查看最大最小值是否是异常值。



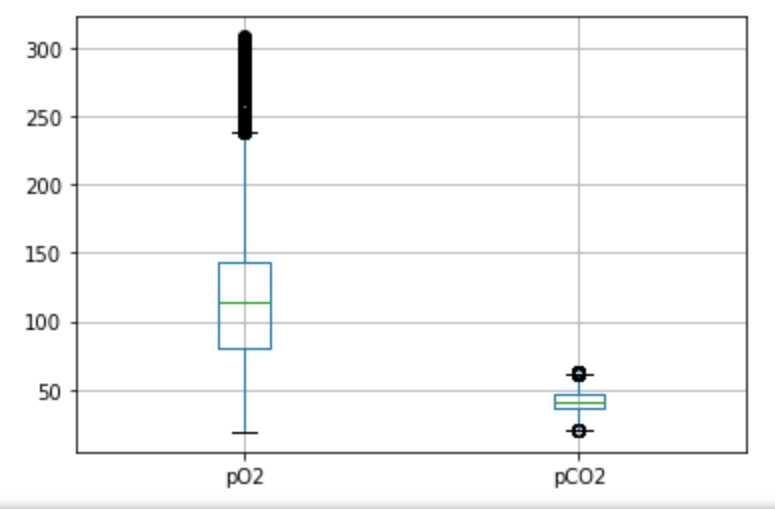
绘制pO2,pCO2两列数据的箱型图，列出离群数据。



（pO2和pCO2的箱型图）

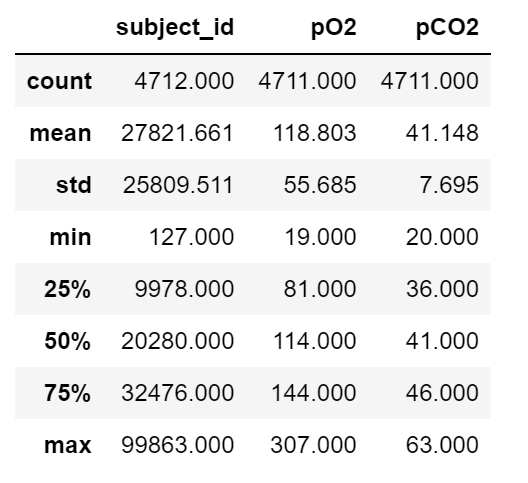
1. 离群值的处理

用平均值代替离群值，做出新的箱型图



（平均值代替离群值后的箱型图）

df.describe(),看出离群值平滑了很多



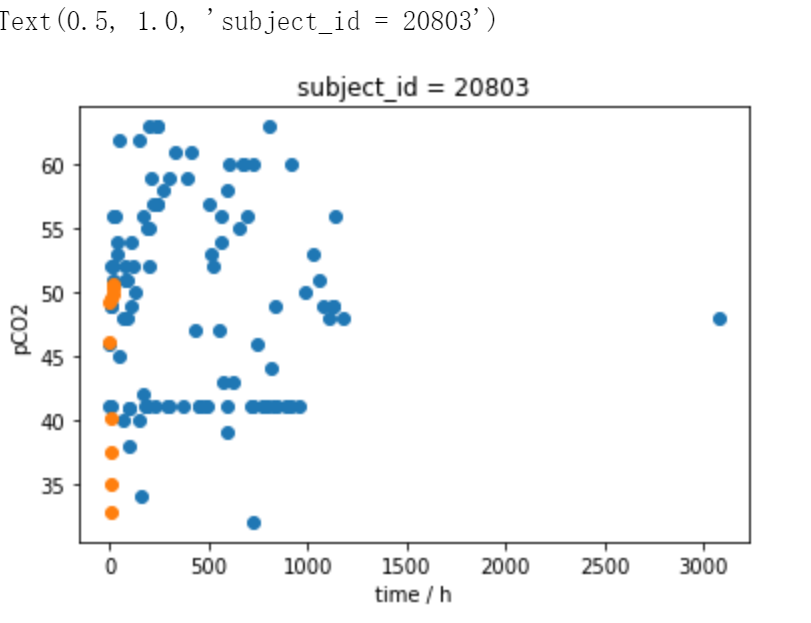
1. 时间换算

把charttime换成小时的形式，并以小时作为时间轴，第一次采集时间为0小时。

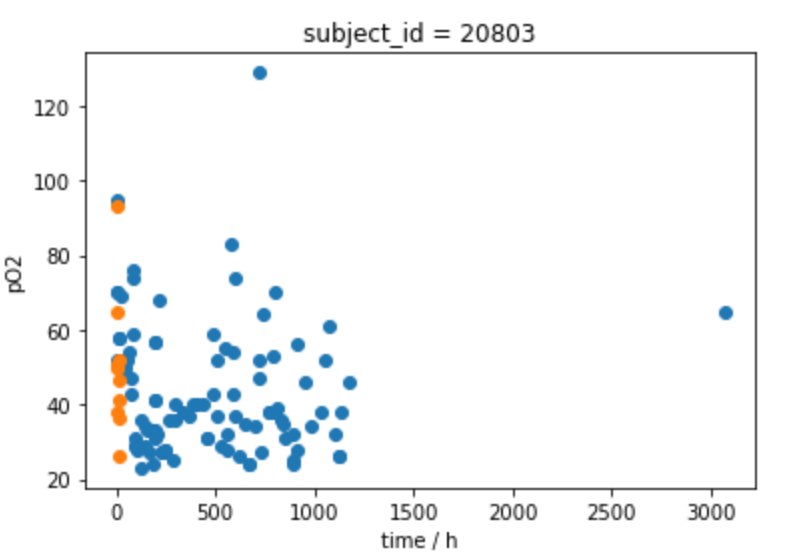
1. 插值计算

选取三个不同的病人，分别对他们的pO2,pCO2进行拉格朗日插值计算，计算前10个插值。

病人1：

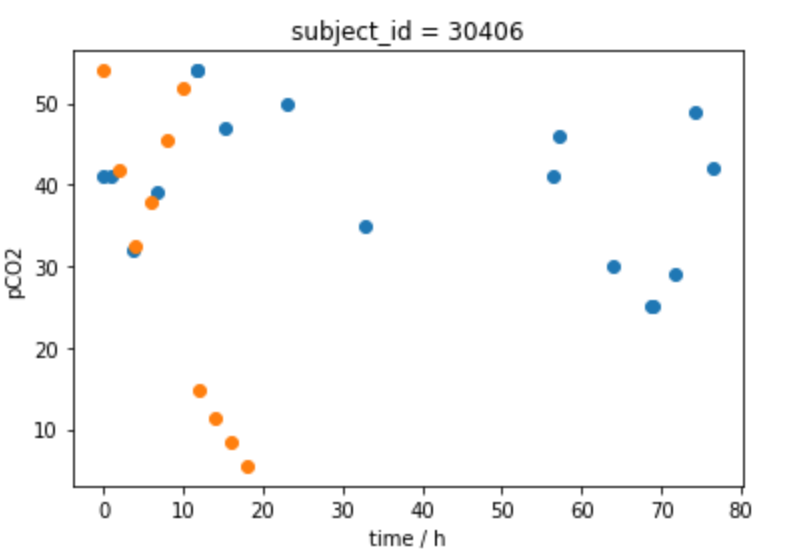


（病人1的pCO2插值）

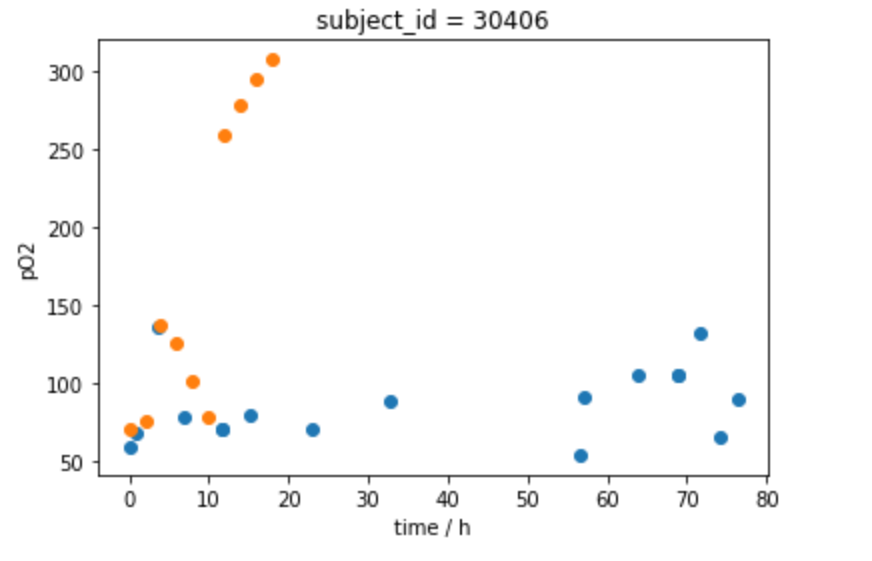


（病人1的pO2插值）

病人2：

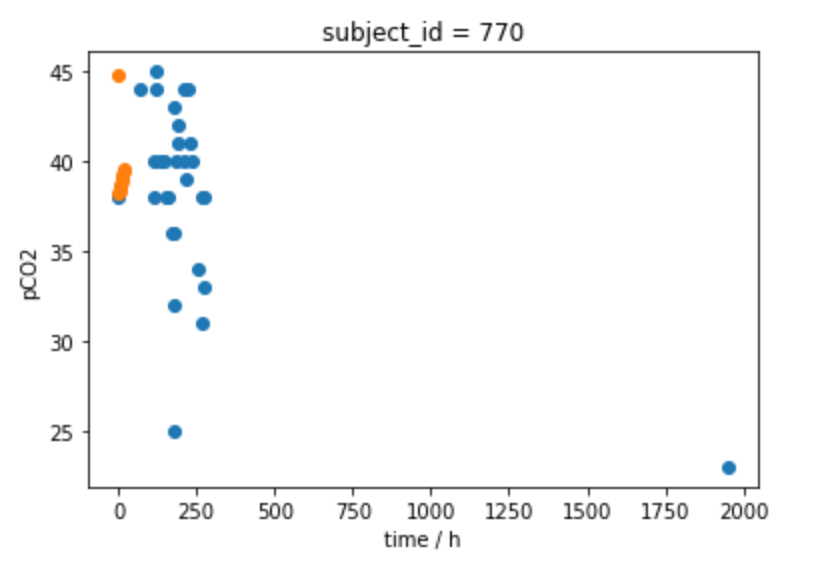


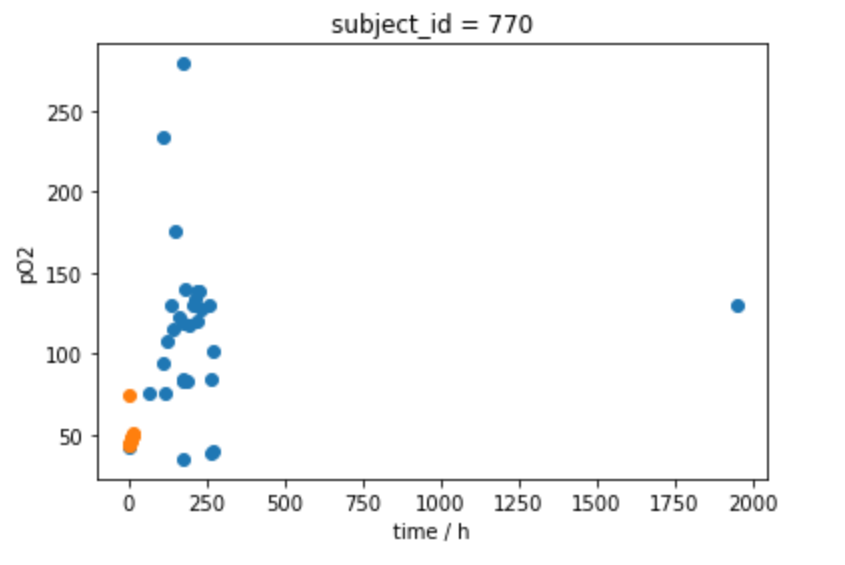
（病人2的pCO2插值）



（病人2的pO2插值）

病人3：



（病人3的pCO2插值）

（病人3的pO2插值）

1. **实验结果与分析**

在对pCO2和pO2进行数据提取的时候，其中一个有值另一个必为空，在对他们两进行合并的时候就会出现大量的重复值，所以采用删除的方式进行去重。

在去重，去缺失值之后，数据中还存在大量的离群值，在对pO2和pCO2统一数据类型后，对他们的离群值平滑处理，采用均值替代的方式，大大的缩小了极差。

1. **附录**

[**MIMIC数据预处理.html**](MIMIC数据预处理.html)