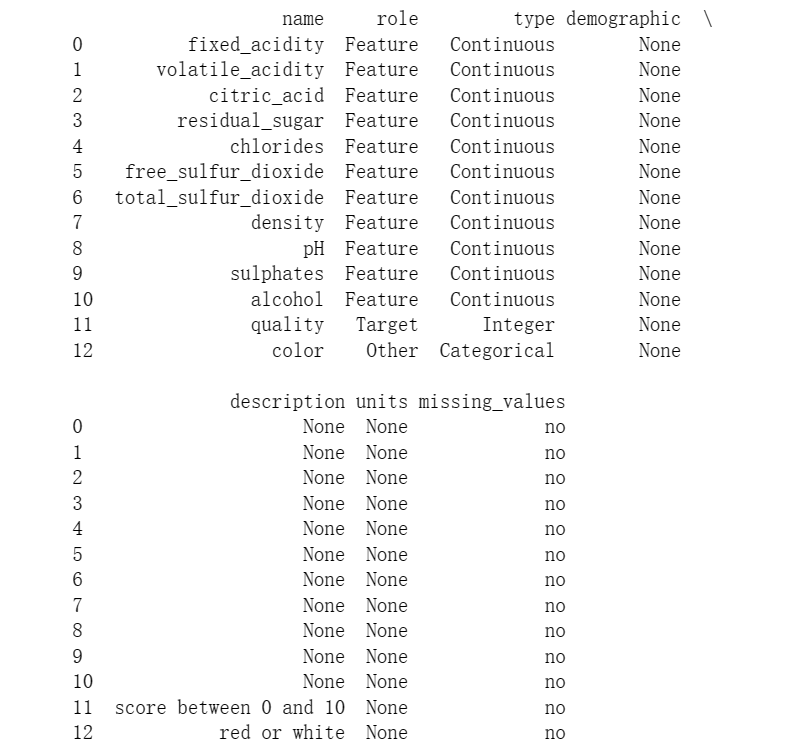
**数据处理过程：**

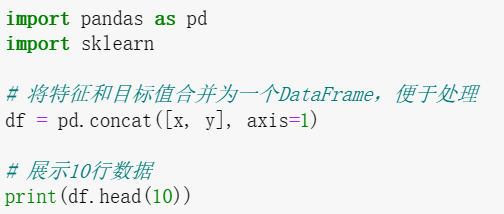
1. 根据UCI官方提供的方法，导入Wine Quality数据集，打印变量信息



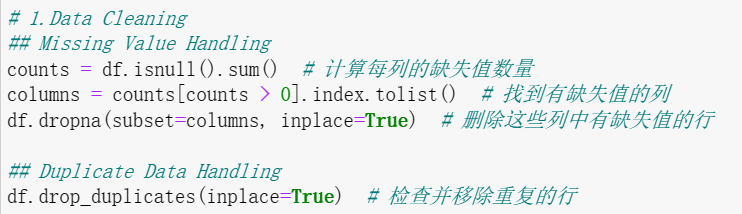
发现有11个feature变量；1个target变量，即quality；1个other变量，即color，代表葡萄酒种类。在下面的代码中，我是对数据集整体做处理，并没有区分葡萄酒的种类。



1. 将特征和目标值合并为一个DataFrame，便于处理



3、开始数据清洗，计算每列的缺失值数量，找到有缺失值的列，删除这些列中有缺失值的行；移除重复的行

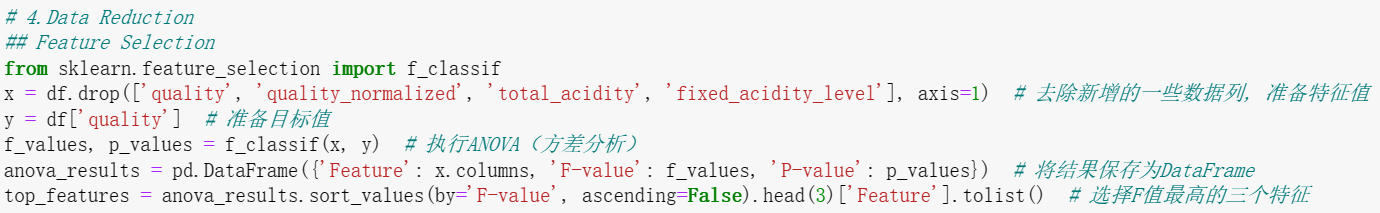


4、开始数据集成，计算“总酸度”，即“固定酸度”与“挥发酸度”的和，然后作为新列添加到数据集中

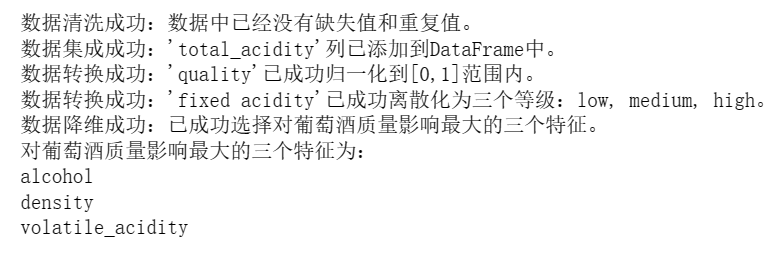


5、开始数据转换，使用MinMaxScaler将“quality”列的数据归一化到[0,1]范围；使用pd.cut()，即将连续变量转换成离散变量的函数，将“固定酸度”列的值离散化为三个等级：低、中、高



6、开始数据降维，使用特征选择过滤器f\_classif函数，计算样本的ANOVA（方差分析）F值，选择对葡萄酒质量评分影响最大的三个特征，即排序后选择F值最高的三个特征  


**数据处理结果分析：**



1. 数据清理：已成功识别并删除关键列中包含缺失值的行，没有空条目；已成功识别并消除数据集中的重复条目，确保了每个数据点的唯一性。
2. 数据集成：已成功创造了一个新的特征变量“总酸度”。
3. 数据转换：“质量”特征已成功被标准化为[0,1]范围，使其更容易在观测值之间进行比较；“固定酸度”已成功被离散化为三个级别：低、中、高，有助于简化分析。
4. 数据降维：已成功使用方差分析确定了对葡萄酒质量影响最大的前三个特征，即'alcohol', 'density', 'volatile\_acidity'。