**数据探索与数据预处理作业**

|  |
| --- |
| 得分 |
|  |

**一、读取sklearn的鸢尾花数据集，绘制各个特征箱线图，查找是否存在异常值，并显示出来。（给出代码和运行结果图）**

**二、对sam\_tianchi\_mum\_baby\_trade\_history.csv中数据进行以下操作（给出代码和运行结果图）：**

1．读取数据，并命名为trade;

2. 识别数据中是否存在缺失值，返回缺失值的数量和位置。

3. 定义重复数据，取4到104行数据，添加到原始数据中，显示重复值位置，并删除；

4. 请将buy\_mount列中大于100的异常数据变为空值，运用众数和拉格朗日插值两种插值方法对缺失值进行填充；

5. 采用以下方法对cat1和buy\_mount两列数据进行标准化操作：

（1）最小-最大，区间为[-1,1]

（2）零均值，均值方差为各列均值方差

（3）小数定标

6.对cat1和buy\_moun数据进行等距离与等频率离散化处理（k=4）,并用groupby方法显示四类的数量；

7. 采用PCA降维，保留主要特征的97%。