实验二 数据降维

1. 实验目的
2. 进行一步理解机器学习基本原理和实现过程；
3. 学会使用python进行基本数据统计；
4. 能用python 实现PCA算法进行数据降维
5. 进行学习效果的评价方法
6. 实验设备

1． PC机或云计算平台

2. 操作系统不限

3. Python 3.x

4. 使用numpy, pandas等基本算法包

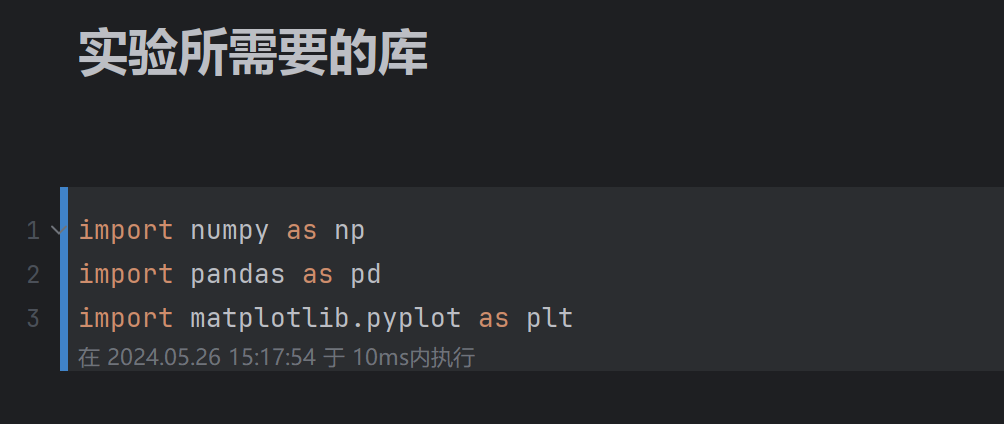
1. 实验原理

在wine数据集文件中，每行代表一种酒的样本，共有178个样本，一共有14列。其中，最后一列（‘Customer\_Segment’）属性是类标识符，分别是1/2/3来表示，代表葡萄酒的三个分类。前面的13列为每个样本的对应属性的样本值分别为：酒精、苹果酸、灰、灰分的碱度、镁、总酚、黄酮类化合物、非黄烷类酚类、原花色素、颜色强度、色调、稀释葡萄酒的OD280/OD315、脯氨酸。

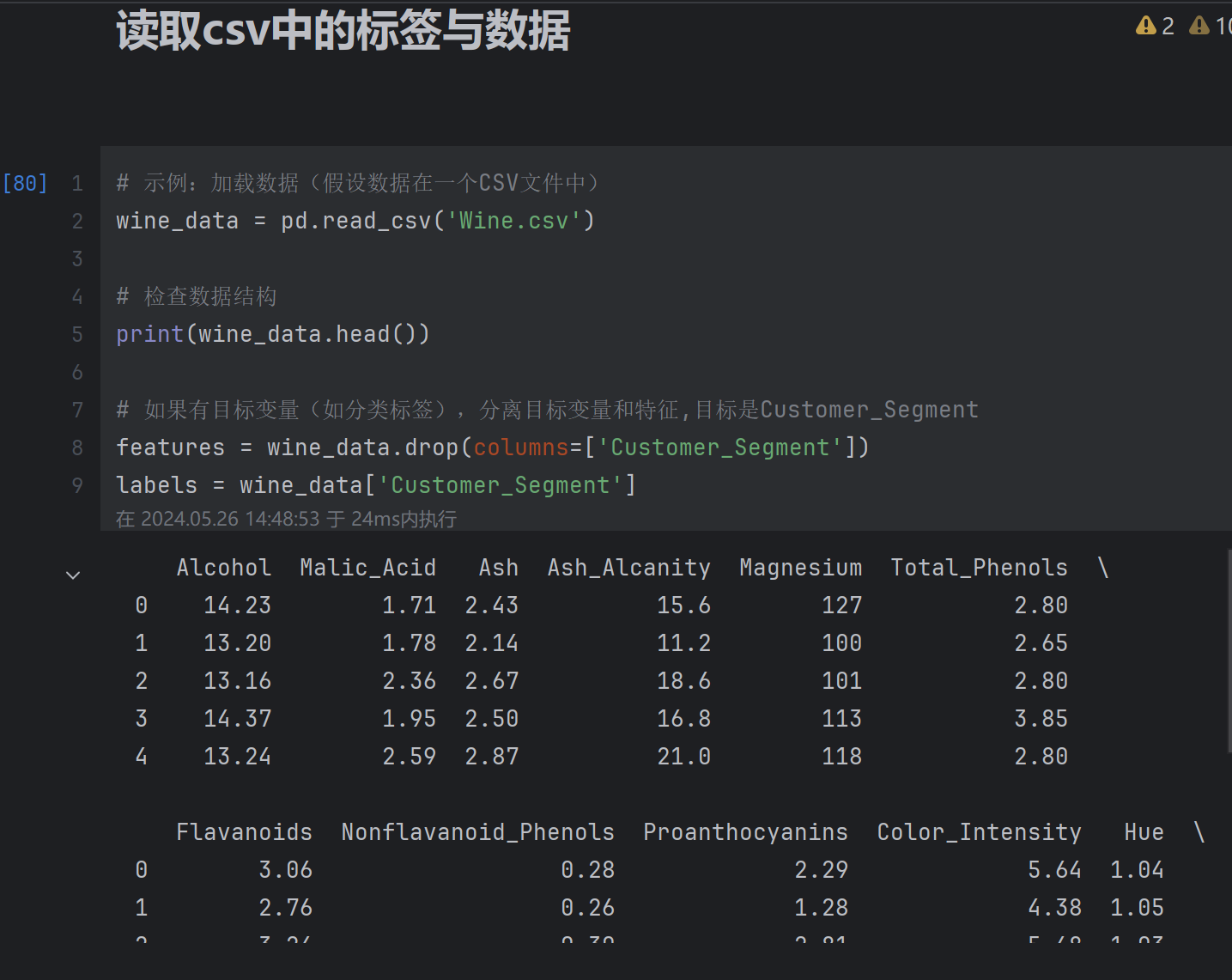
实验任务：采用PCA算法对wine数据特征进行降维处理；

1. 实验过程记录

描述程序设计思想、主要代码等



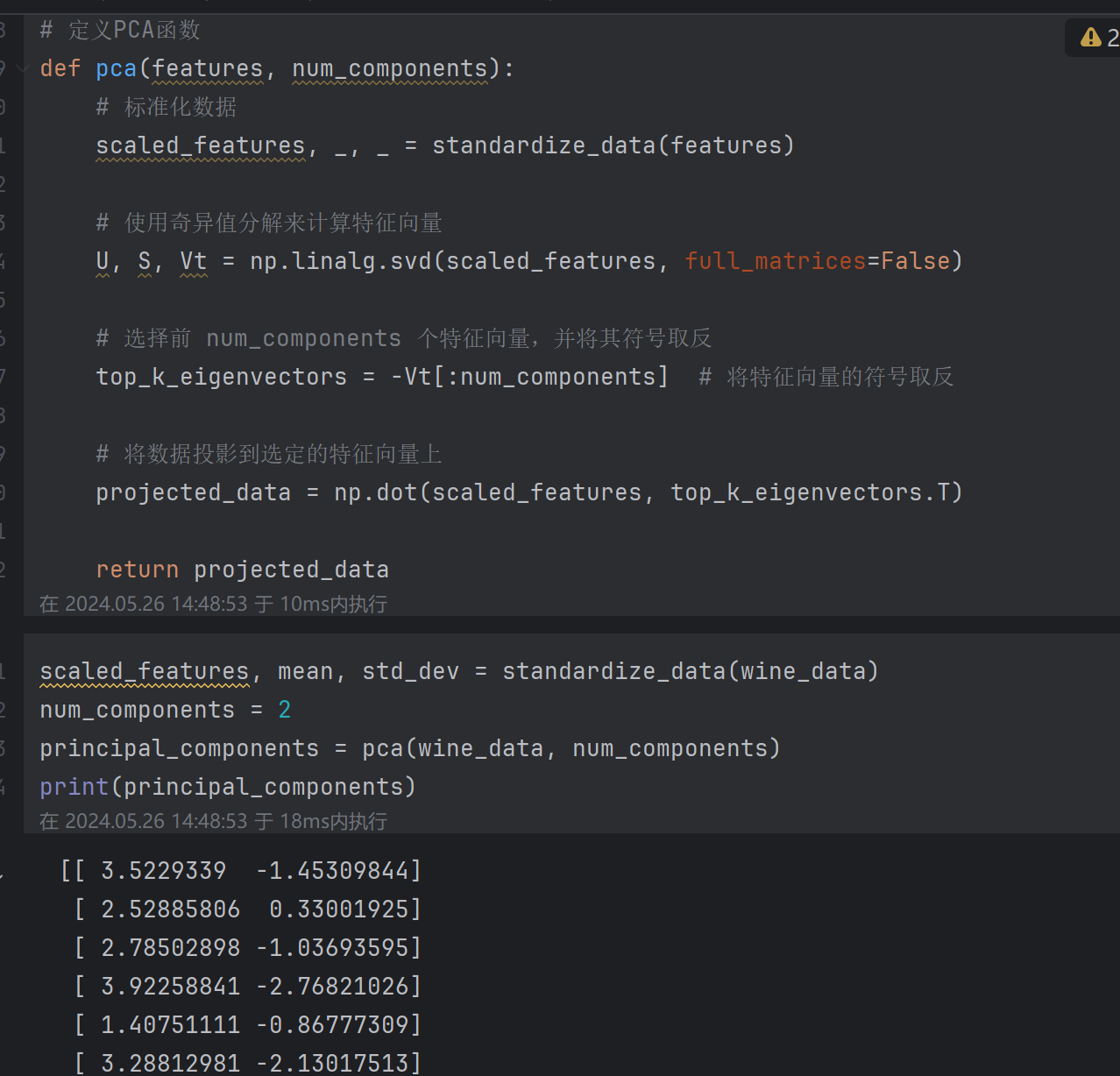
1从CSV文件中读取数据



2.对数据进行标准化后使用PCA算法进行降维

经过比较结果可以得到采用奇异值分解法比采用特征值分解的精度更高，所以在PCA算法中采用了奇异值分解法



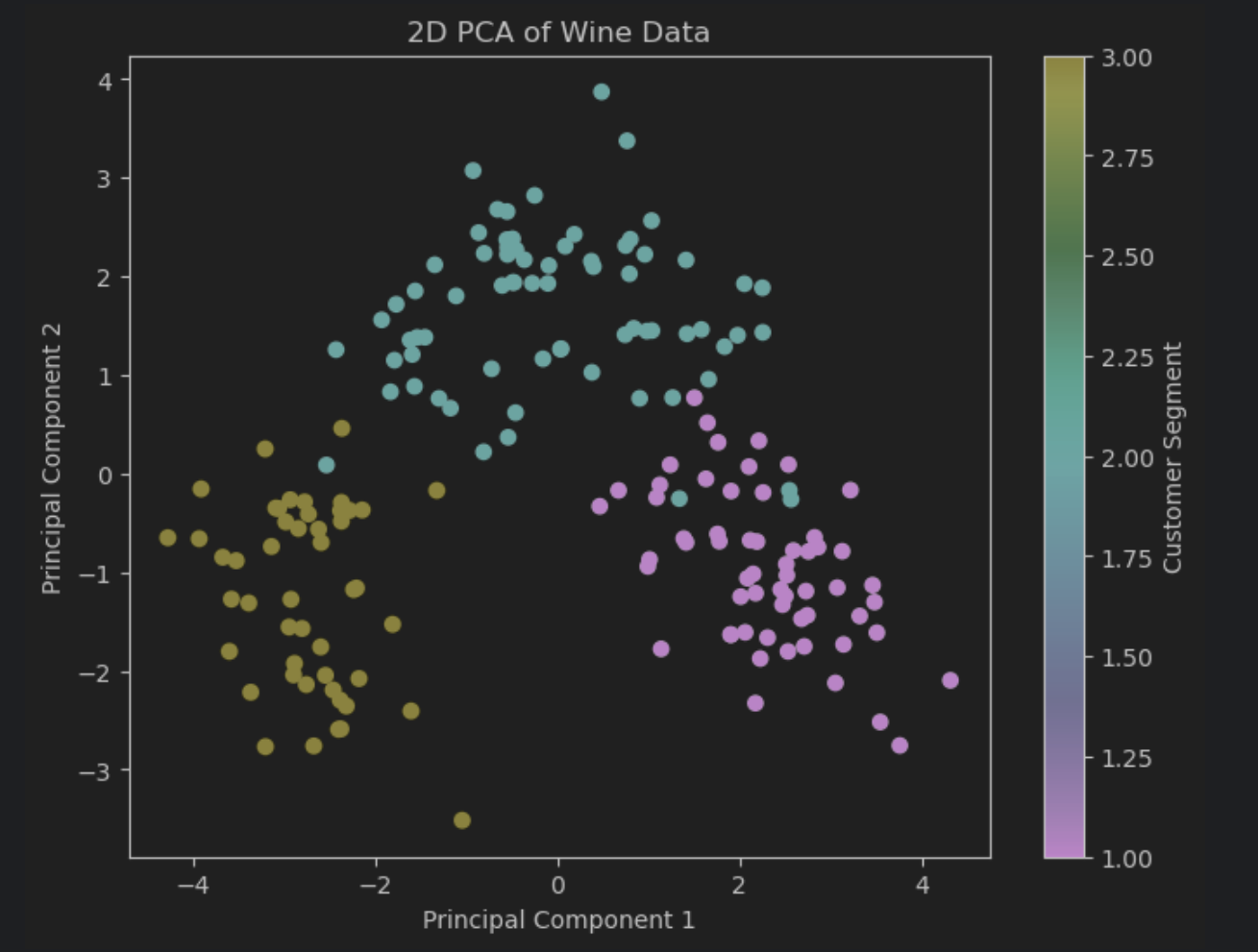


3创建带有主成分的dataframe并绘制散点图



1. 结果与分析

试对PCA降维结果进行解释。



数据通过降维，在保留原始数据的同时去掉了不重要的部分，将葡萄酒分为了三类。通过分析这三类数据的分布，可以看出不同类别的葡萄酒之间的区别，以及同类别中葡萄酒在特征方面的相关性。例如第三类葡萄酒的颜色更深、第一类葡萄酒的酒精度更高。主成分在图表中的方向表示了葡萄酒数据集中最显著的方差方向。通过观察主成分的方向，可以推断出不同特征在降维后的空间中的影响程度。如果某个主成分的方向与某个特征相关，那么这个特征在降维后的空间中可能具有重要的影响。

1. 自评

本次实验学到什么？遇到什么困难，又是如何解决的？自己在什么地方得到了提升，又在什么地方需要加强？

本次实验我学会了手工实现PCA算法的方法。在实现PCA算法的过程中注意到特征值分解法不能完全区分每个特征，经过查阅资料，采用了奇异值分解法，提高了精度，更能区分每个特征，我对这两种矩阵分解方法有了更深刻的认识与了解，也了解了降维思想的应用以及优点缺点。