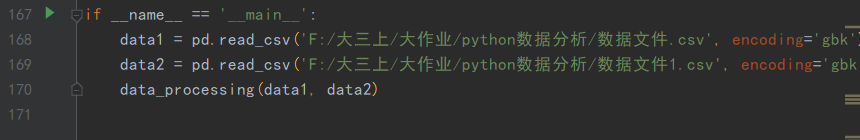
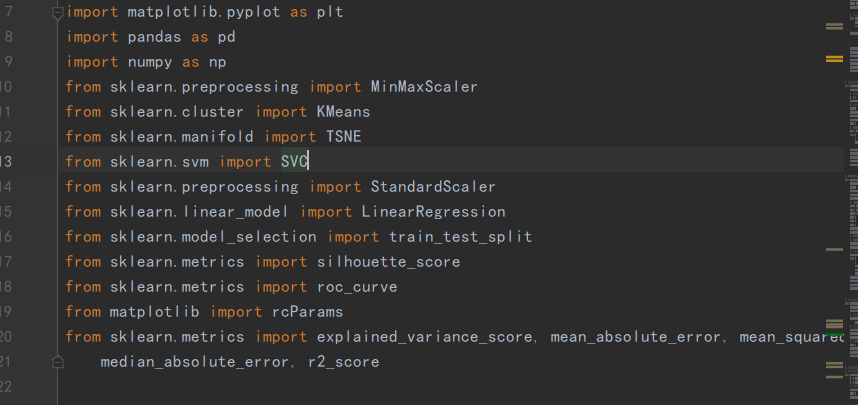
**中国图书网图书信息数据分析**

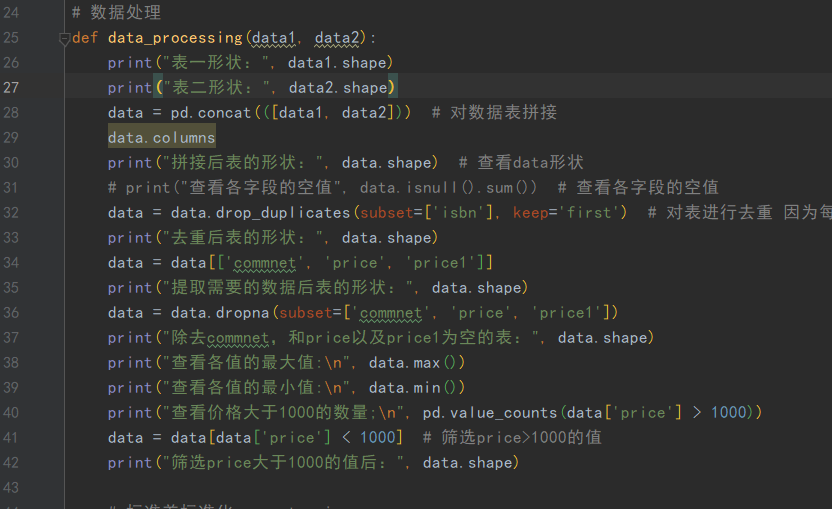
1. **大数据分析技术的应用**
2. **数据预处理的代码、注释说明及运行结果**

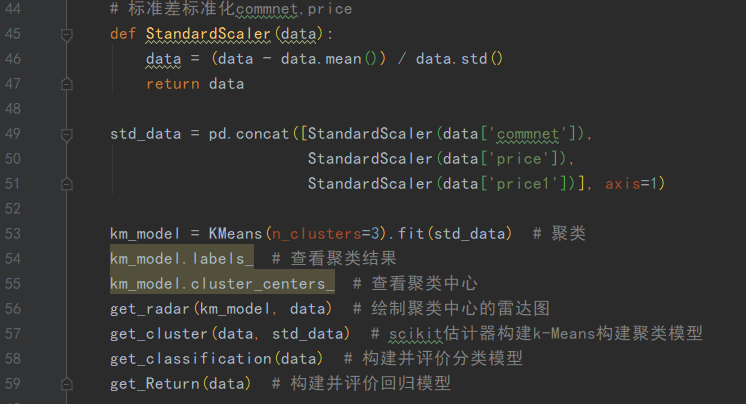
**主函数导入相关的文件**



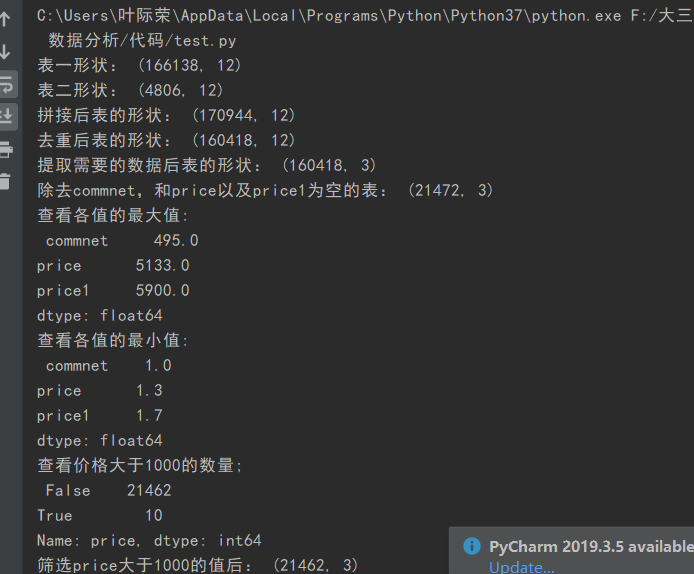
**导入相关的库**







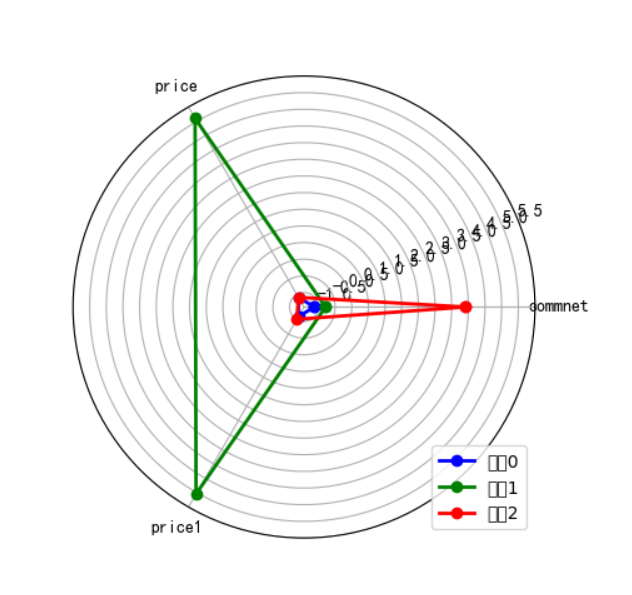
**相关结果：**



1. **数据探索及特征构建的源代码、注释说明及运行结果**

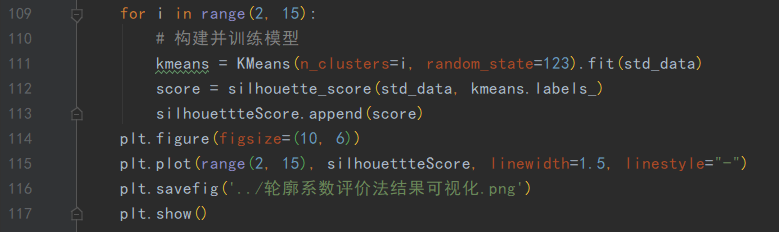


**相关结果**

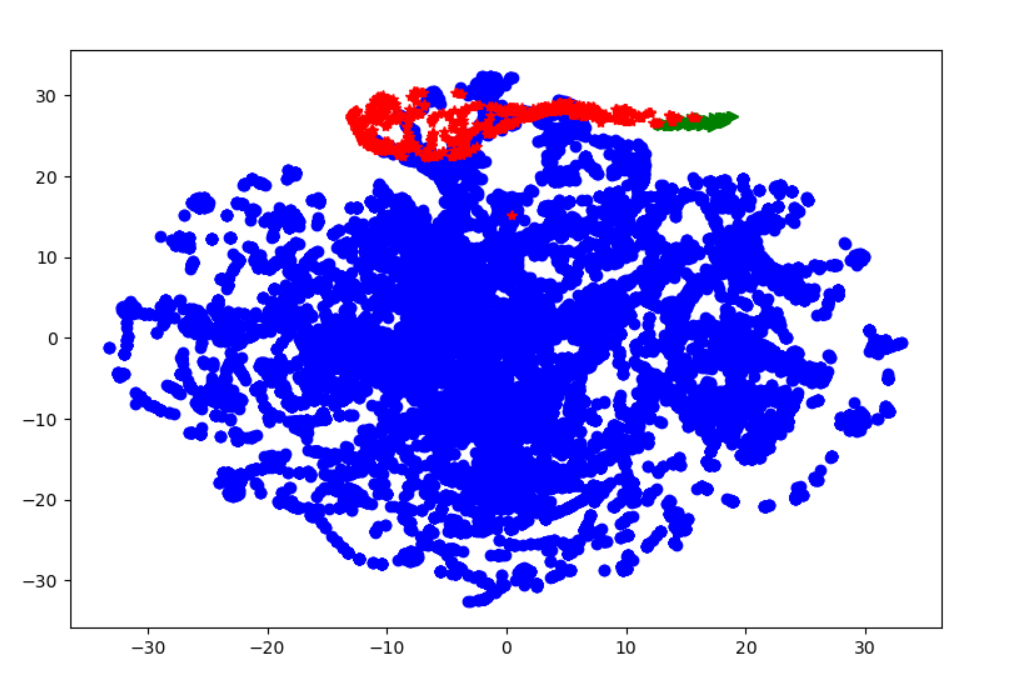


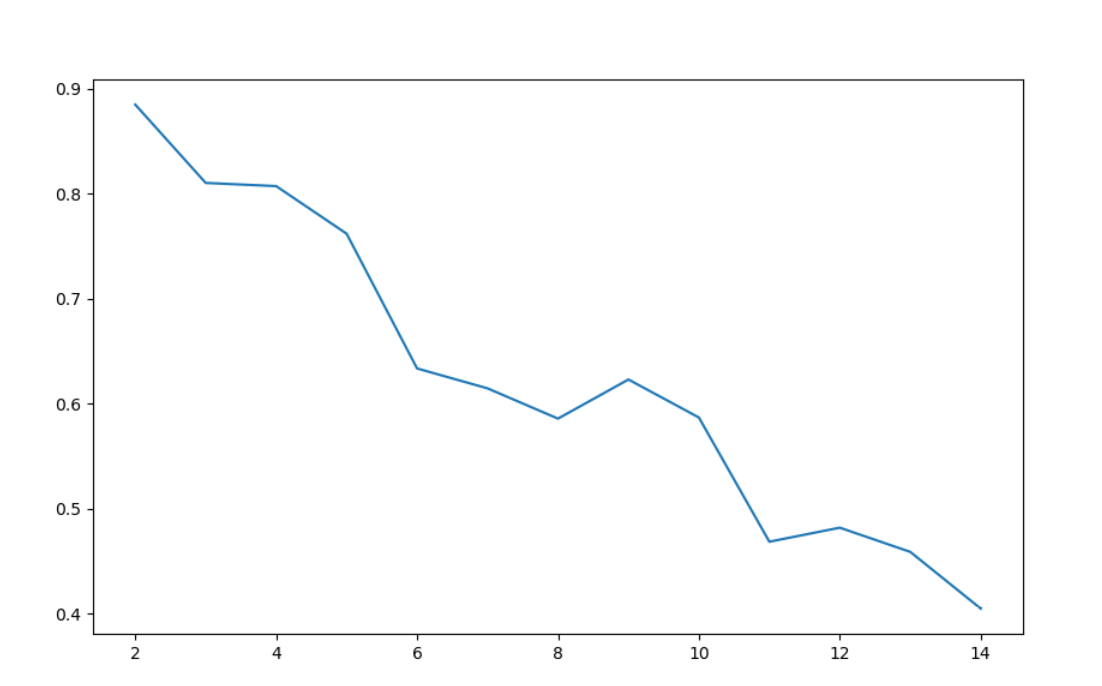
**3、聚类模型的构建与评估的源代码、注释说明及运行结果**



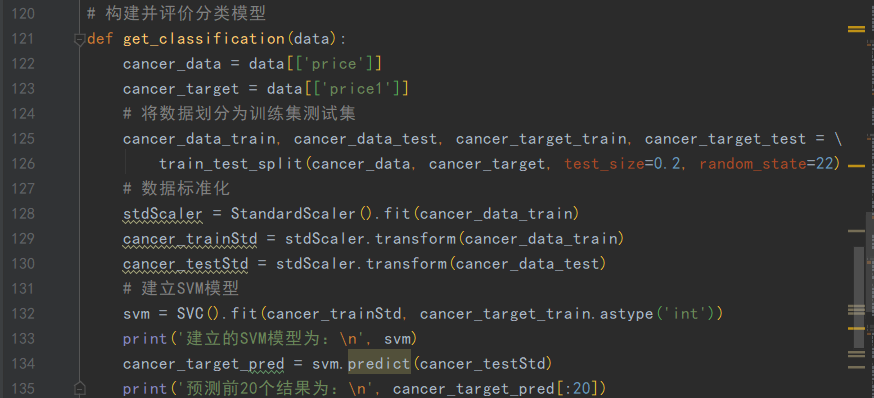


**相关结果：**

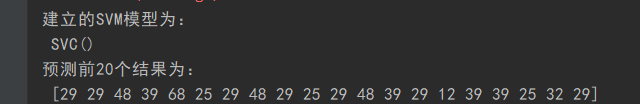




**4、分类模型的构建与评估的源代码、注释说明及运行结果**

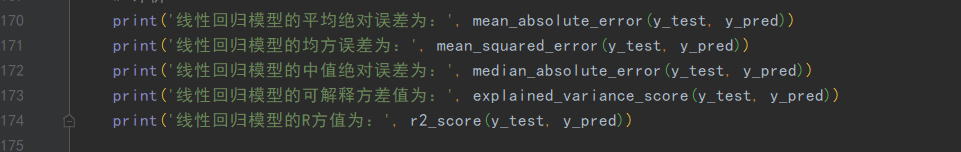


**相关结果：**

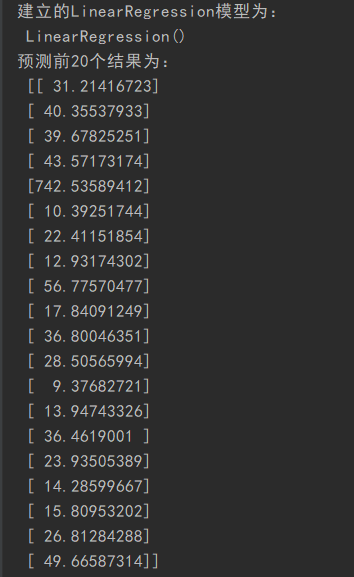


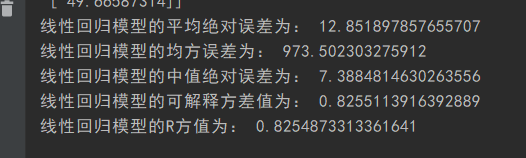
**5、回归模型的构建与评估的源代码、注释说明及运行结果**

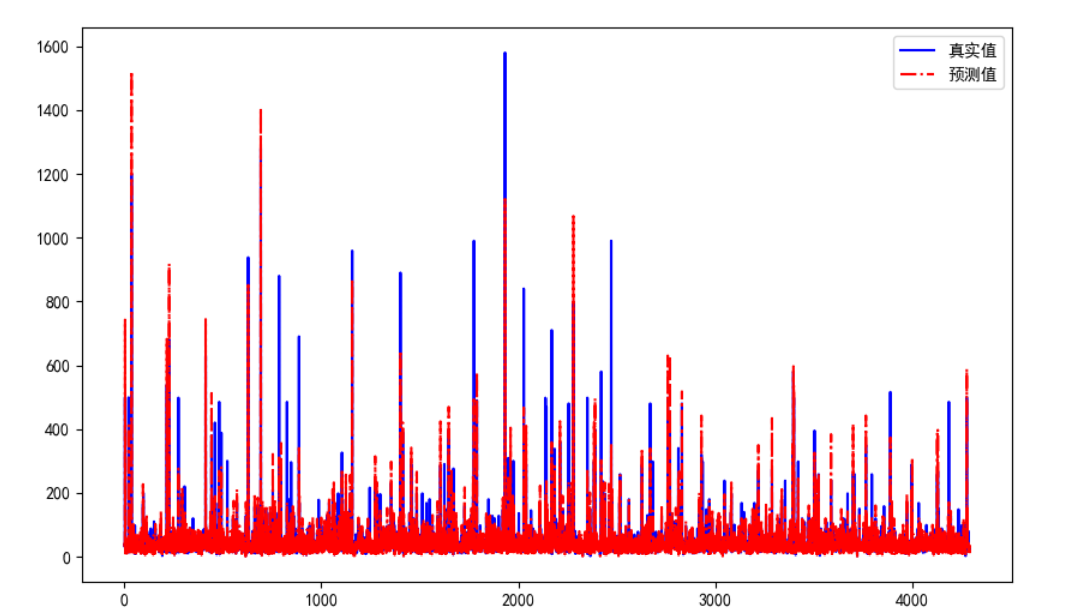




**相关结果：**

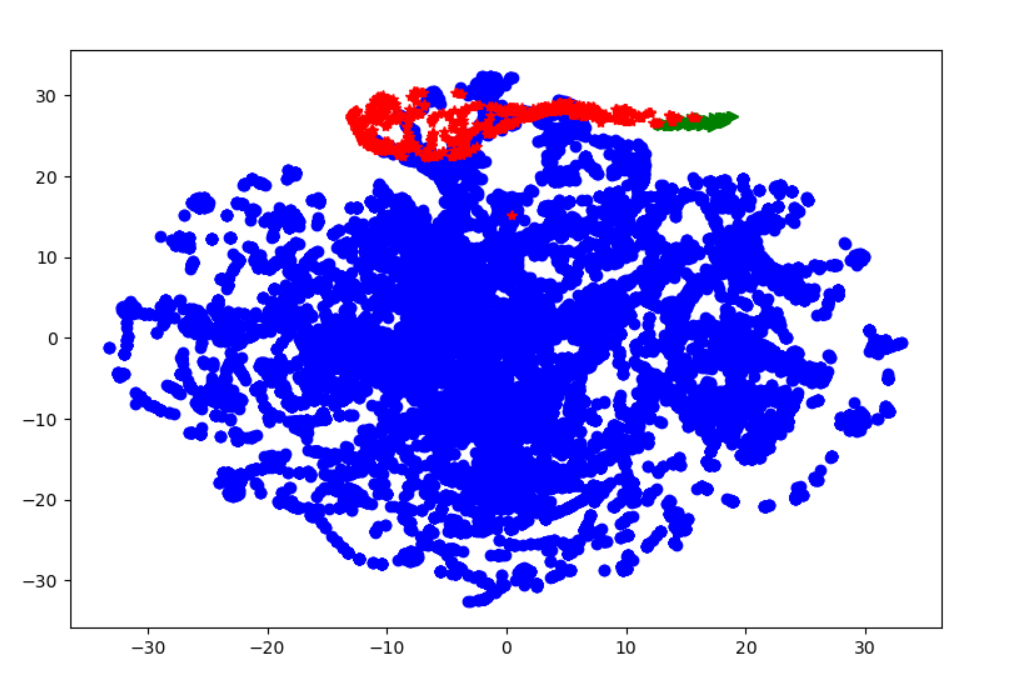


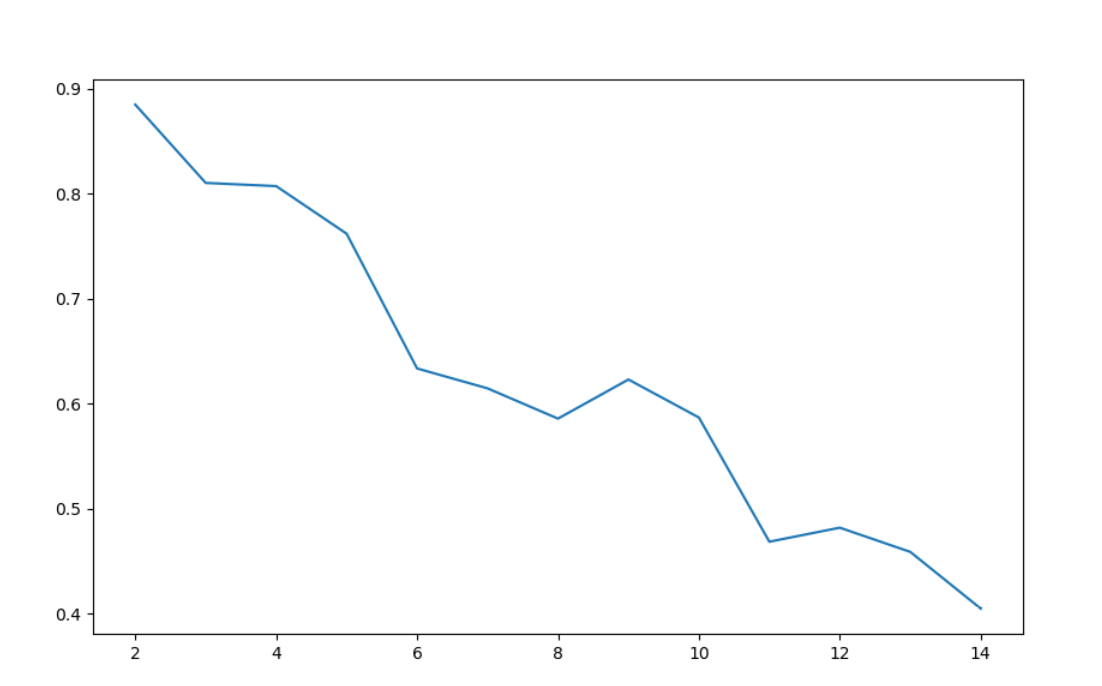




**6、各种模型分析结果的对比与说明**

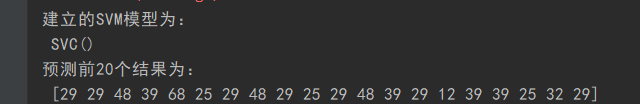
聚类模型:





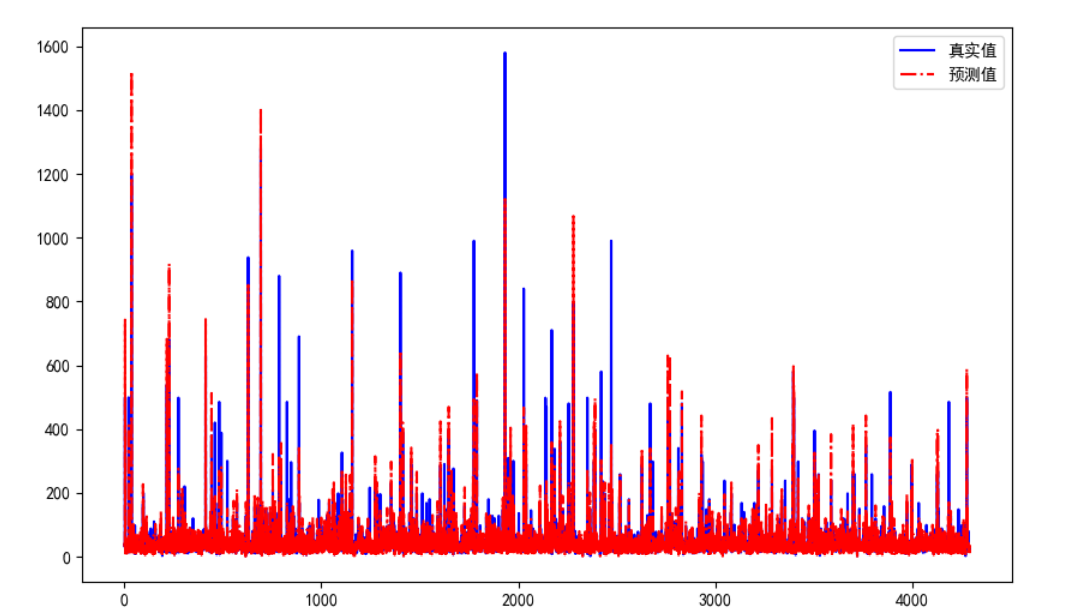
聚类模型结果可视化从颜色可以看到蓝色颜色暂大部分，而红色部分以及绿色部分占据比较少，轮廓系数结果可视化发现后发现越往后评价准确率越低。

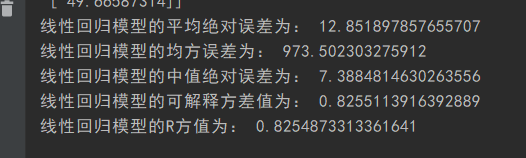
分类模型



分类模型只写到这里了，往后会有报错，暂时没有解决，所以只有这个，但是发现预测结果准确率太低。

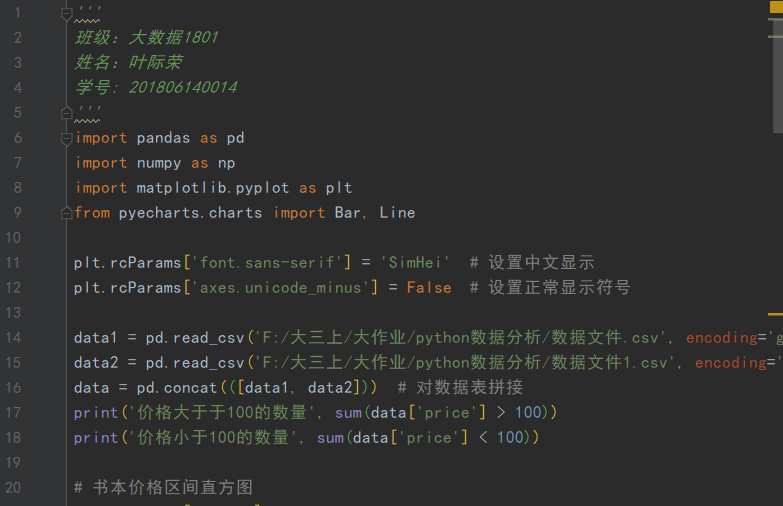
回归模型

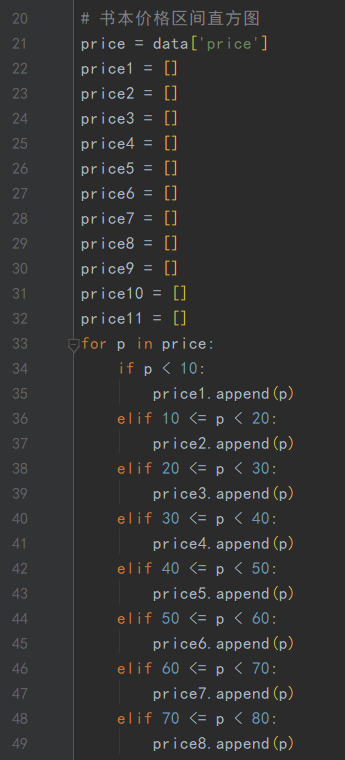


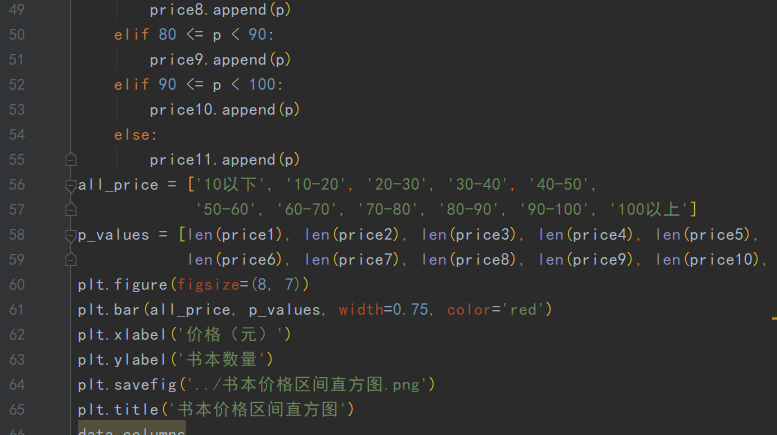


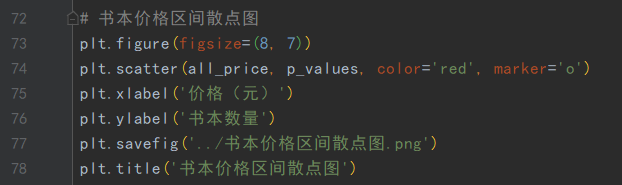
回归模型可视化发现真实值与预测值都有相关的差距。回归模型平均决对误差有12多，误差较大。拟合效果不是特别好**三、数据可视化技术的应用**

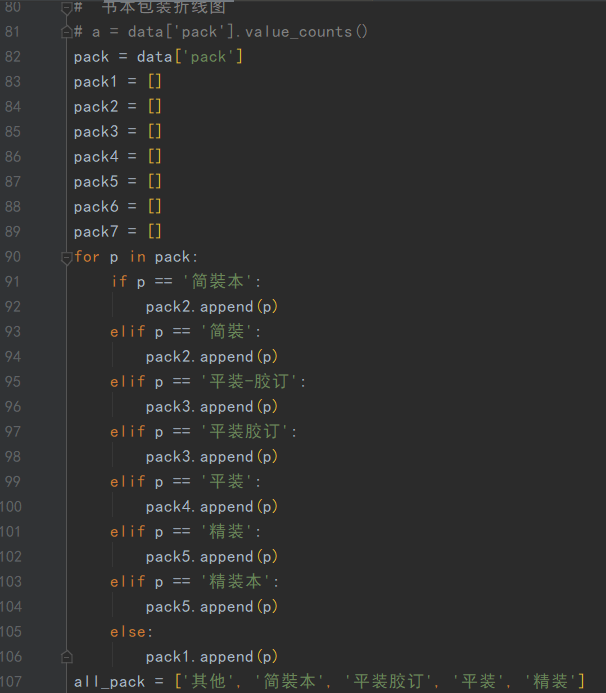
**1、第一种数据可视化技术运用的源代码及运行结果、简要说明**

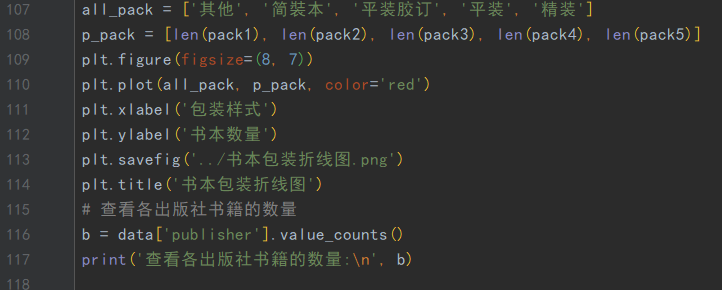




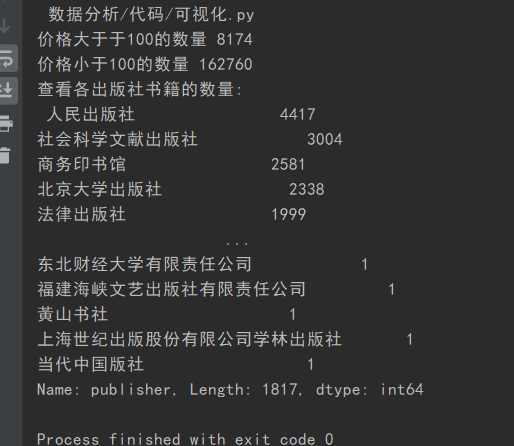


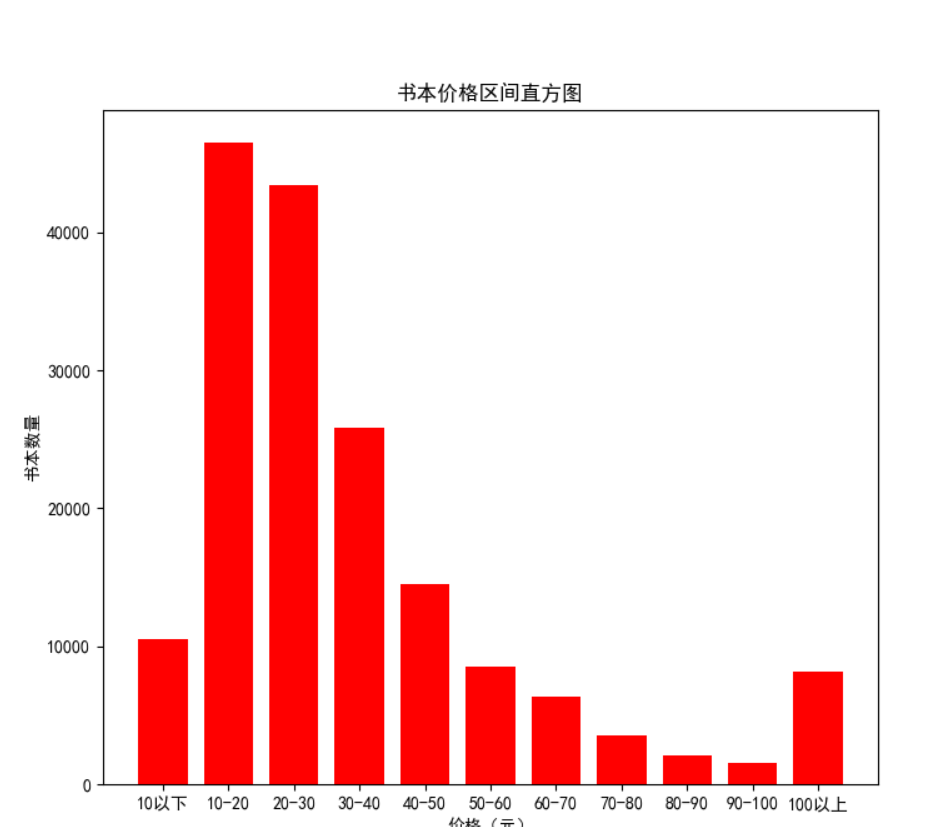


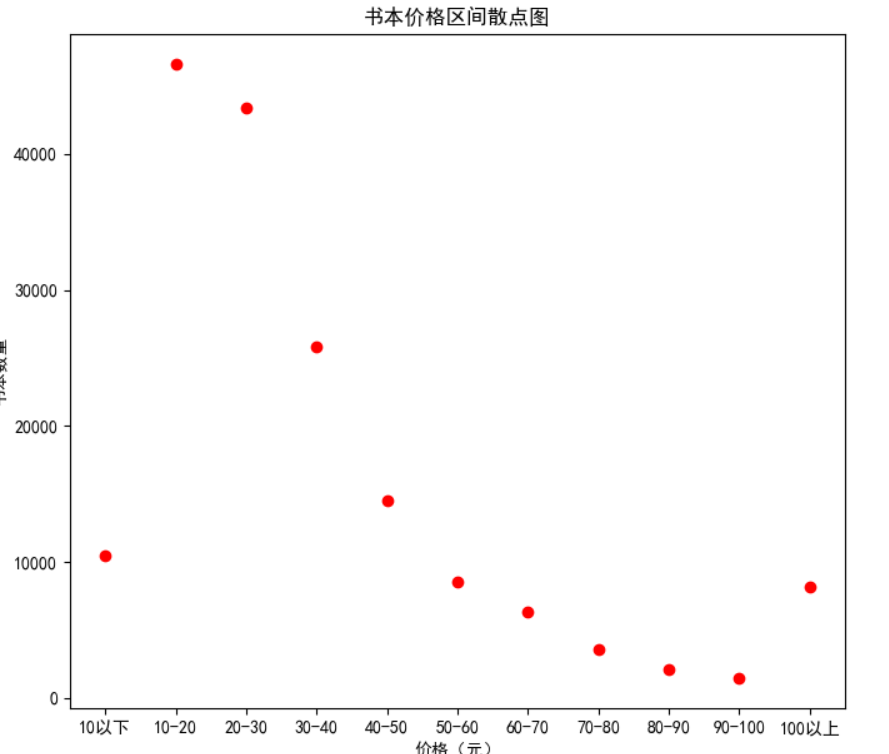


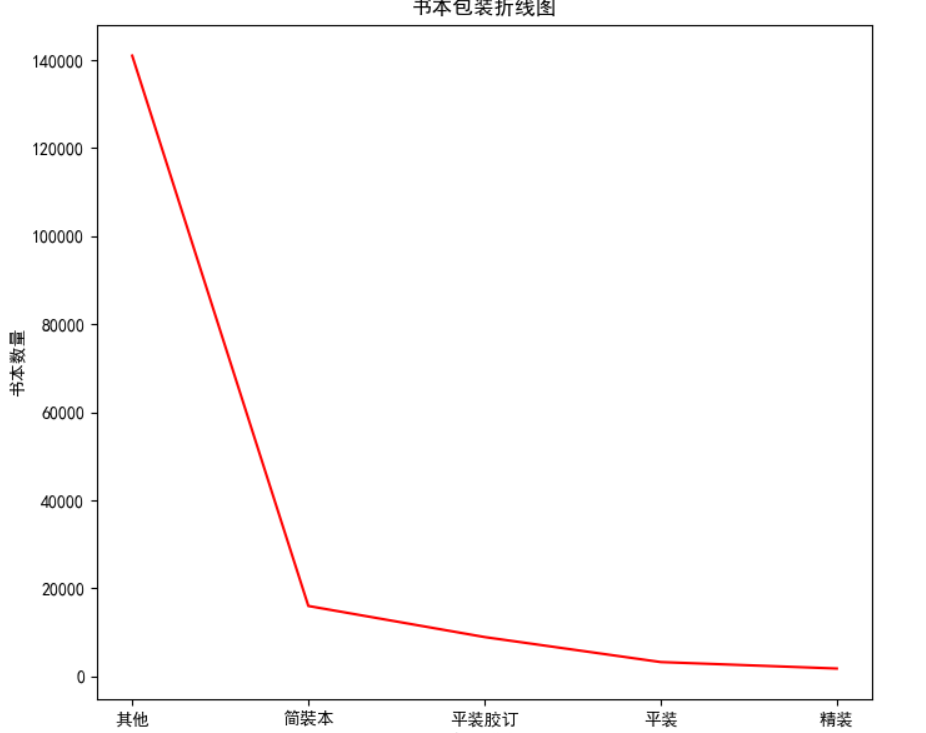


**结果：**

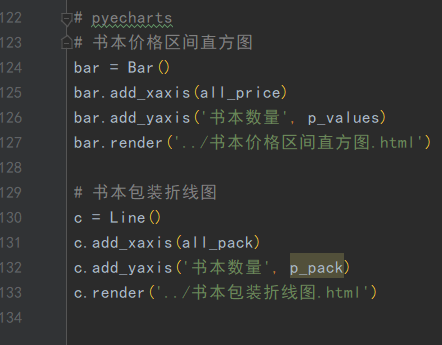


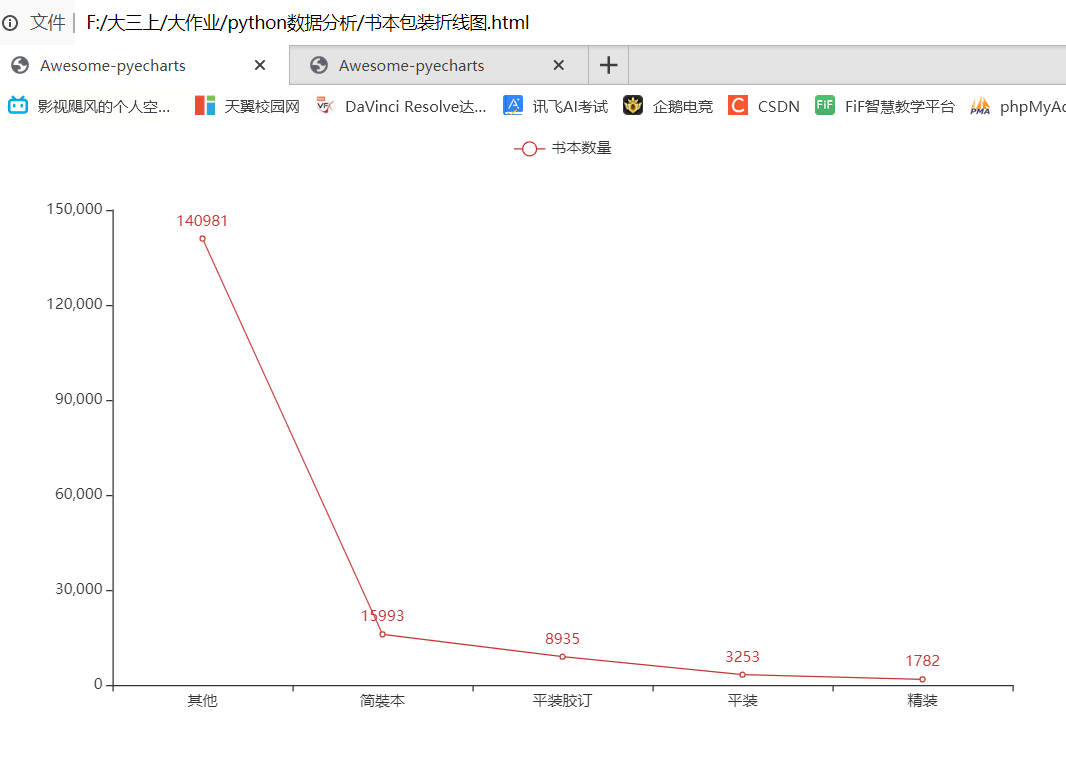


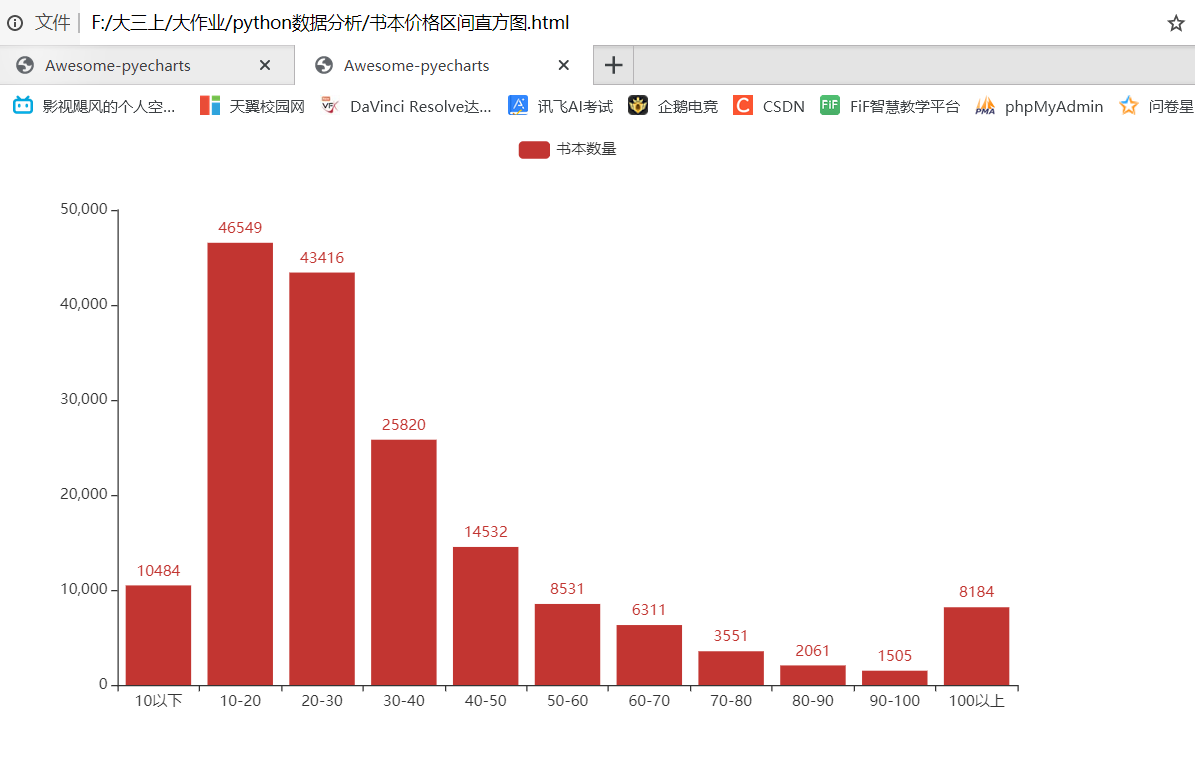




1. **第二种数据可视化技术运用的源代码及运行结果、简要说明**







**四、思考与总结**

本次大作业自己熟练了数据的数据合并、清洗、标准化等数据预处理、以及数据探索及特征构建、聚类模型的构建与评估、分类模型的构建与评估、回归模型的构建与评估。以及相关的数据可视化的使用等

数据合并、清洗、标准化等数据预处理我自己来说是比较熟练了，但是相关的数据探索及特征构建还不是太熟练。不太会找相关的数据。聚类模型的构建与评估也不太会。本次聚类模型的结果图自己也理解不了。因为的出来的图实在是太不对劲。而且我用tsne进行降维处理单发现他处理的太慢太慢了，KMeans聚类的优点是 算法简单易实现，缺点是需要用户事先指定类簇个数，聚类结果对初始类簇中心的选取较为敏感，容易陷入局部最优，只能发现球形类簇，我本次出来的图就是这样的，是我初始类簇中心的选取不好。分类模型的构建与评估得出来的数据预测不准确。导致后面的相关数据处理报错。所以没有往后面做。应该是我导入数据出错了。训练的数据和预测的数据都是有问题的。回归模型的构建与评估较为顺利，但是越往后预测就越不准确，现在相关的模型评估还是不够熟练。

数据可视化方面相关的是先用matplotlib做。首先克服的是相关的价格区间书籍的数量。这我用append处理，其他方法暂时没想到。然后就是常规的画图方法。接着用pyecharts画图，pyecharts画图比较简单，我用的比较简单的画图。其他比较复杂的暂时还没掌握。正在努力学习中。