**企业诚信度聚类分析**

1. 数据的预处理

本数据集中共有三个指标

1. 正向化

三种数据并非全部为同一类型，is\_justice\_creditaic和is\_justice\_credit为是否工商部失信企业和是否列入失信黑名单，考虑到数字并非仅有0和1，推测为被记次数，此两项为负面数据，越少越好。is\_kcont为是否列为守合同重信用企业，同样并非只有0和1，此为正面数据，越多越好。正向化的过程是把两类数据，能通过变换为欧式距离可以标志样本相差大小的过程。为了减少工作量，我们对is\_kcont取负，得到nega\_is\_kcont指标。

1. 标准化

采取z标准化，标准化后会使每个特征中的数值平均变为0(将每个特征的值都减掉原始资料中该特征的平均)、标准差变为1。

公式为

如此我们消除了不同变量间方差不同对聚类的影响。

1. 经上两步后我们将三个变量成为

X: is\_justice\_creditaic（标准化）

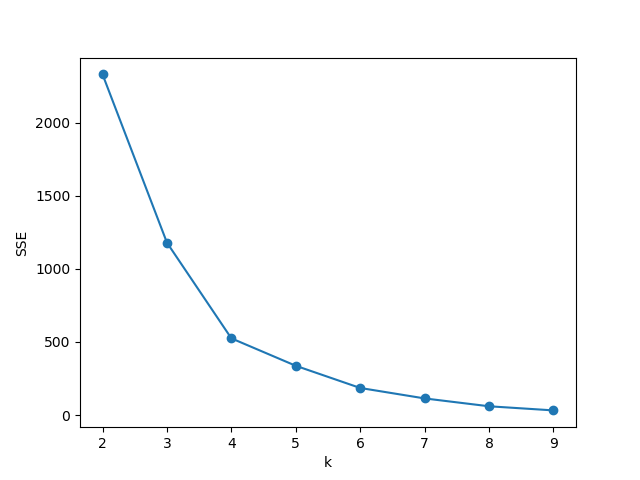
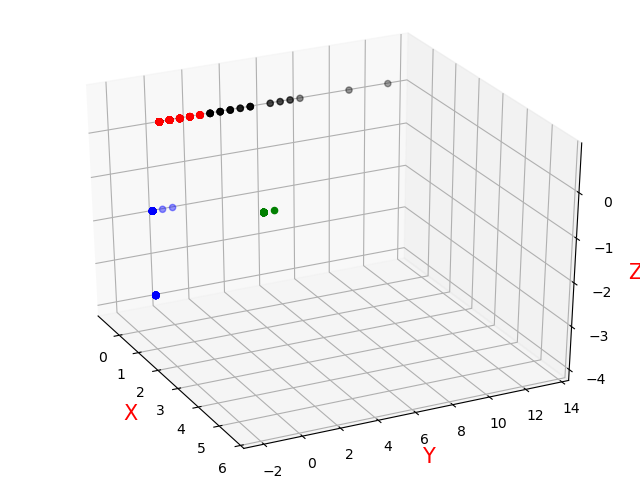
Y: is\_justice\_credit（标准化）

Z: nega\_is\_kcont（标准化）

二、聚类分析

1、k-means

根据肘部图我们选择4作为k-means聚类的k值较为合适

此时轮廓系数为0.8260，Calinski-Harabasz指数为2255.985

k-means聚类效果图

k-means肘部图

2、亲和力传播聚类

默认参数下不收敛，暂不讨论

时间复杂度较大，为O(N3)，在数据量大时收敛较慢

3、meanshift聚类

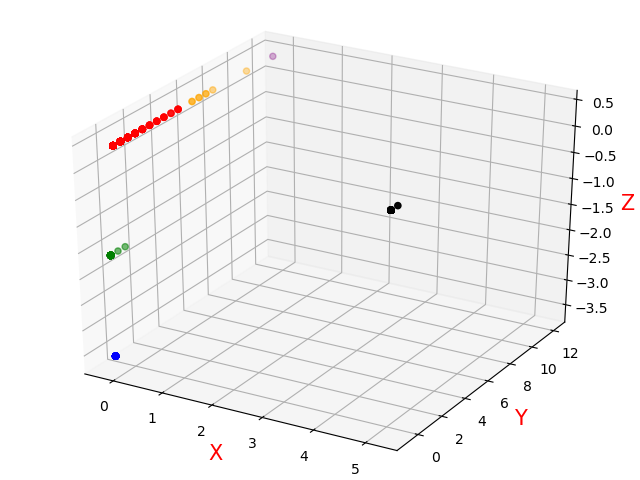
轮廓系数为0.8320

Calinski-Harabasz指数为6773.494

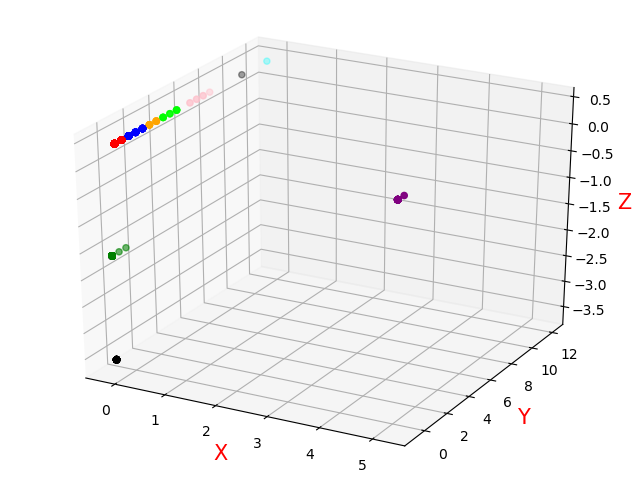
分类效果较好，但是类别达到10类，解释难度较高，分类太细

为了使类别减少，调整参数带宽为1.9，此时的得到6类，轮廓系数为0.8066

Calinski-Harabasz指数为1340.5



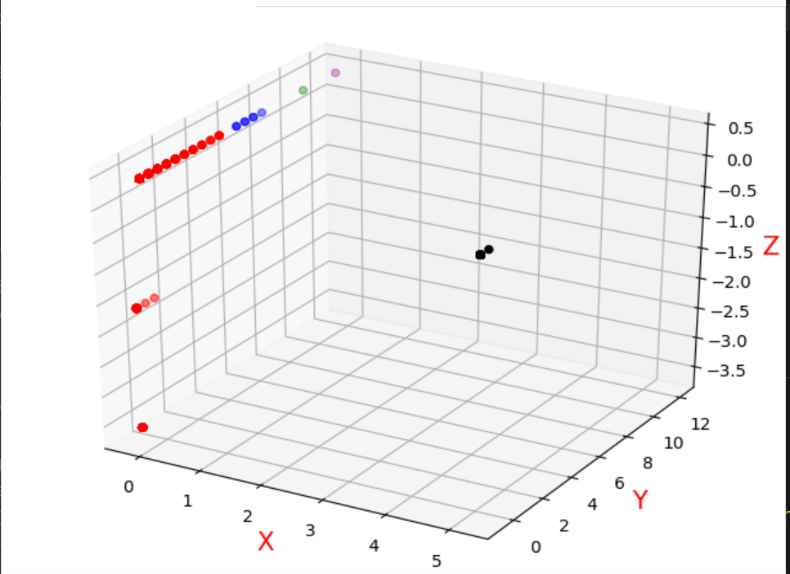
bindwidth=1.9时聚类效果



默认参数下meanshift聚类效果

4、谱聚类

默认划分为8个类，调整为5个类（根据散点图）后，两个离群点被单独分为两类，而诸多肉眼可见有区分的聚集点被归为一类，效果一般。

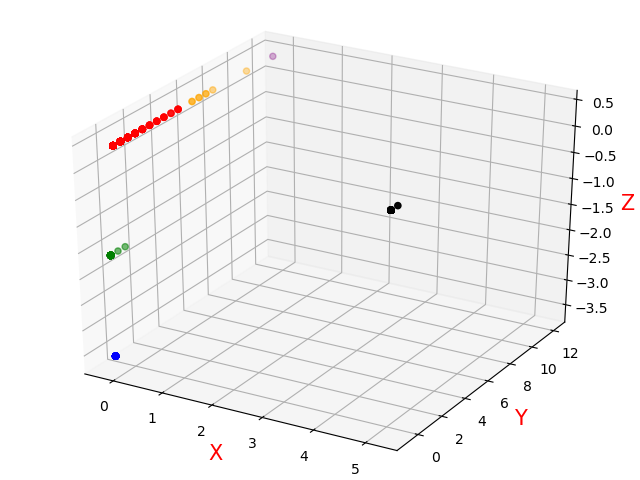


谱聚类，参数n\_cluster为5

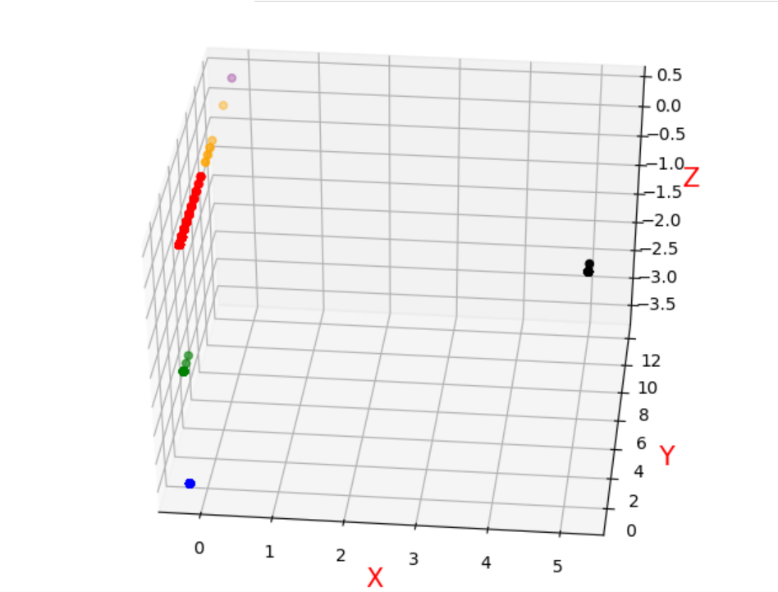
5、总结

在解释性、聚类结果的评分的综合评价后，我们认为meanshift聚类(bandwidth=1.9)对于该数据集为较好的聚类方式

三、聚类结果分析



bindwidth=1.9时聚类效果



第一类：绿色。被记一次至10次的工信部失信企业。

第二类：红色。被记一次守合同重信用企业，对企业的正面作用较为明显。被记两次以内工商部失信企业。

第三类：蓝色。被记两次守合同重信用企业。对企业的正面作用极为明显，没有负面记录。

第四类：黑色。被列入失信黑名单。对企业负面影响非常明显。

第五类：紫色。为一家被记24次工信部失信企业。可以记为异常值并划入其他相近类。对企业负面影响明显。

第六类：橙色。被记14至20次工信部失信企业。对企业负面影响明显。

标签划分：

第一类（绿色）：企业诚信度较低

第二类（红色）：企业诚信度较高

第三类（蓝色）：企业诚信度极高

第四五六类（黑色紫色橙色）：企业诚信度低

其余不在此数据集的企业：企业诚信度一般