## 项目介绍

用python来编写高斯混合模型聚类算法

学习重点EM算法，第二周作业

## 数据介绍

df=pd.read\_csv('Diamonds Prices2022.csv')

使用导入的钻石价格数据

## 算法细节

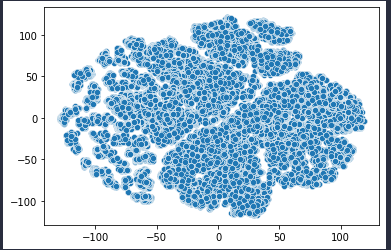
高斯混合模型（GMM）是什么呢？顾名思义就是用多个高斯模型来描述数据的分布。就是说我们的数据可以看做是从多个高斯分布中生成出来的，GMM是由K个高斯分布组成。

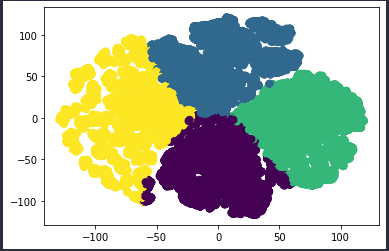
EM算法被称为期望最大化算法，是迭代求解最大值的算法。当我们获得的数据中有隐含的未知变量时，可以使用EM算法。算法解决问题的思路使用启发式的迭代方法，如果我们无法直接求出模型分布参数，那么我们可以先猜想隐含数据，基于观察数据和猜测的隐含数据一起来极大化对数似然，求解我们的模型参数，然后在此基础上不断迭代优化。 算法在每一次迭代时分为两步，E步和M步。一轮轮迭代更新隐含数据和模型分布参数，直到收敛，即得到我们需要的模型参数。

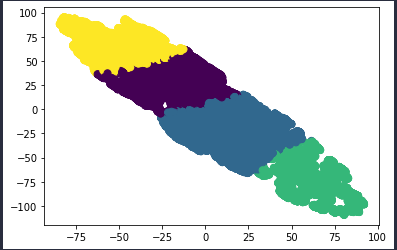
在GMM中使用EM算法的原因是需要求解的是一个混合模型，只知道混合模型中各个类的分布模型（都是高斯分布）和对应的采样数据，而不知道这些采样数据分别来源于哪一类（隐变量），那这时候就可以借鉴EM算法。EM算法可以用于解决数据缺失的参数估计问题（隐变量的存在实际上就是数据缺失问题，缺失了各个样本来源于哪一类的记录）。

　Sklearn库GaussianMixture类是EM算法在混合高斯分布的实现。原本是想自己敲的，EM算法流程并不复杂，但后来发现sklearn库中有封装好的直接用就行。

## 算法结果







拉伸后数据处理效果