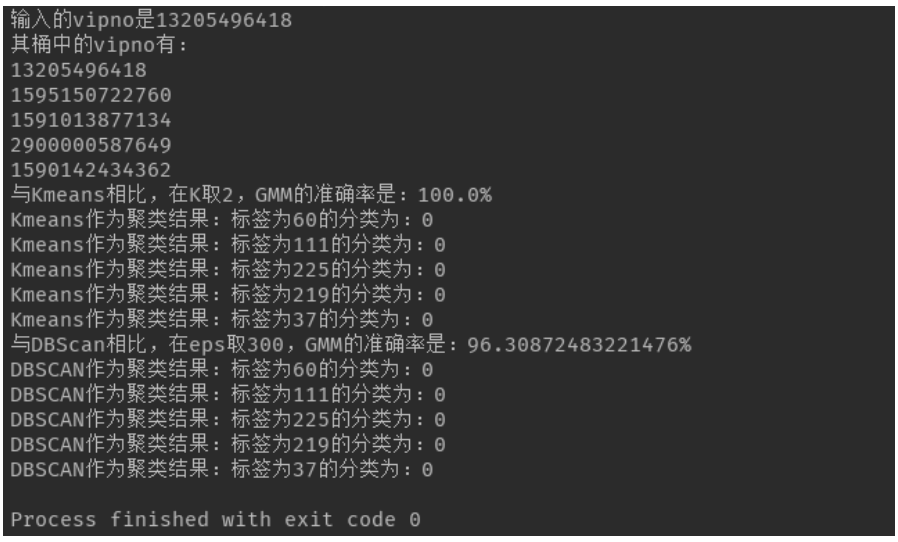
Q4

1. 代码运行截图

a)假定以 Kmeans 作为真实的聚类结果，计算 GMM 的准确率

b)假定以 DBScan 作为真实的聚类结果，计算 GMM 的准确率

c)验证上述 lsh 的 knn 查询结果是否与输入 vipno 所在同一个簇



2. 讨论分析部分

2.1 截图说明

a)图中，k 取最佳值 2，GMM 的 n_components 值也为 2，此时 GMM 方法所得标签和 K-Means 方法得到的标签能够一一对应起来：正确率为 100%

b)图中，eps 取越大的值 DBSCAN 的效果越好，所以在此取 300，测得 GMM 的正确率达到 96.3%

c)图中，利用第一问得到的一个 vipno 和其桶中的数据为基准，查看他们在 GMM 聚类的处理下是否还在一个簇里。根据输出可以看到：分类标签均为 0，所以他们还在一个簇里

2.2 相关说明

GaussianMixture：高斯混合模型，的是多个高斯分布函数的线性组合，理论上 GMM 可以拟合出任意类型的分布，通常用于解决同一集合下的数据包含多个不同的分布的情况（或者是同一类分布但参数不一样，或者是不同类型的分布，比如正态分布和伯努利分布）。

使用混合的模型的原因是：往往有多个聚类，只有使用多个搞死模型匹配不同的聚类才能更有说服力。

高斯模型提供了模型，而要让数据匹配模型的话需要用到极大似然估计的方法来确定参数。步骤为：

1. 先求出要估计参数的粗略值
2. 使用第一步的值最大化似然函数（需要有 GMM 的极大似然函数）

2.3 效果评估

对于 K-means 来说，k=2 取得最佳值；此时让 n_component 取 2 意味着用两个高斯模型去聚类这些数据。由前面问题得出的结论，绝大部分数据是被分为一个簇的，因此高斯混合模型和 K-means 在只分两个聚类的情况下聚类效果是相近的。

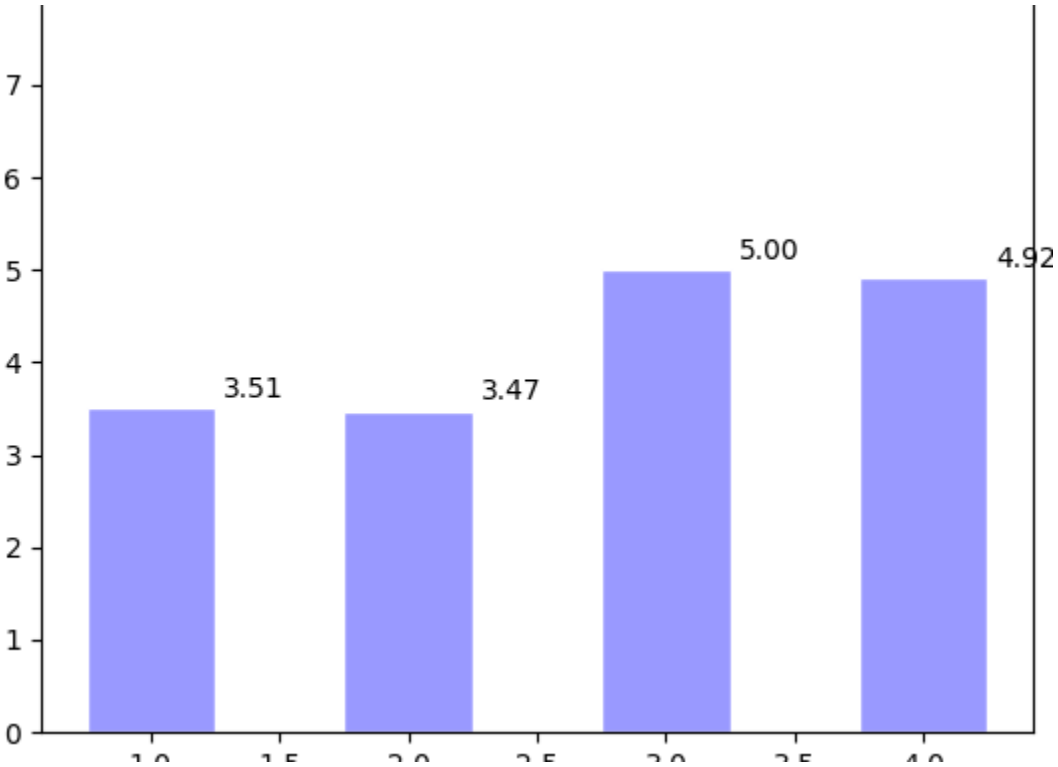
对于 DBSCAN 来说，eps 越大越好；而 eps 即使取得很大的值，其噪音点依旧存在，而且被分成的聚类也只有-1 和 0 两种。因此，去除掉噪音点，混合模型个数此时应当取 1。和 DBSCAN 相比，由于缺少了一些噪音点，其准确率为 96%

由于 K-means 和 DBSCAN 在上述比较中和 Gaussian 模型得出的 label 基本相同，所以 Gaussian 模型去验证 lshash 选出的一个桶里的数据时，它们大概率还是会在一个聚类里。

3. 性能比较

在 Gaussian 模型匹配参数的过程中会用到协方差函数。Sklearn 提供了四个函数：spherical, diagonal, tied 和 full covariance

一下分别是四个函数下高斯混合模型的运行时间图表：



可以看出：tied 和 full 函数下运行时间较长