

## 项目

## Creating Customer Segments

此部分属于 Machine Learning Engineer Nanodegree Program

## 项目审阅

## 注释

与大家分享你取得的成绩！ 

## Requires Changes

还需满足 4 个要求 变化

## 数据研究



已选取三个数据样本，提出建立表达式并给出合理解释。

- 请注意，你需要基于数据的统计特征来分析。
- 具体而言，所谓“数据的统计特征”，即你使用上方 `display(data.describe())` 代码得到的数据的各种统计参数。你需要将样本中的数据与它们进行对比，并得出相关的结论。
- 关于它们具体的作用，你可以参照[数据的统计量特征](#)。

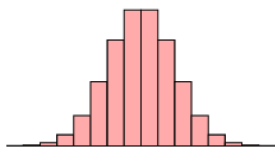
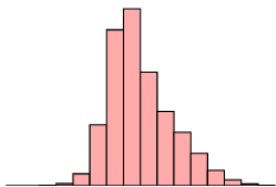
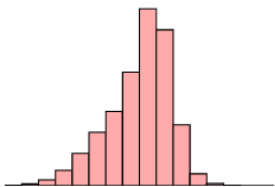
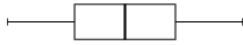
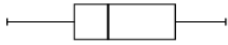
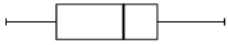


准确报告被删除属性的预测分数，合理解释被删除属性是否具有相关性。



学生找出具有关联的属性并将其与预测属性相比较，随后深入讨论这些属性的数据分布模式。

- 很好，你提到了“些数据的分布不是正态分布。数据的中位数和众数都处于偏低的位置。”，用统计学的术语来说，数据呈正偏态分布，以下图片可以帮助你快速判断数据分布：

Symmetric	Skewed right (positive)	Skewed left (negative)
		
		

## 数据处理



数据和样本的特征缩放已在代码中正确实施。



学生找出极端的异常值，讨论是否删除这些异常值，并说明删除各数据点的理由。

你的代码实现很好，不过移除的异常点过多。

首先，我们一般不会移除“只在一个 feature 中被认为是异常点”的点，而是会移除“在至少两个 feature 中被认为是异常点”的数据点。这是因为：

1. 在一个 feature 中被认为是异常点的点，可能在其他 feature 中有着很重要的作用，甚至可能反映新的一种类型；
2. 我们不希望省略很多的点——否则我们的训练样本会很少，导致它可能无法学到正确的内容。

### 属性转换



准确报告主要成分分析数据的二个维度与四个维度的总方差。将前四个维度合理解释为对消费者支出的表达。

请在修改了上一题的异常值后，重新计算这里的方差，并修改你的回答。



对二维缩放数据及样本数据的主要成分分析已在代码中正确实施。

### 聚类



高斯混合模型和K-均值算法已进行详细比较。学生选择的算法符合算法和数据的特点。



准确报告多个轮廓分数，根据报告的最佳分数选择最佳集群数量。已给出的集群可视化将根据已选的聚类算法生成最佳的集群数量。

关于轮廓系数（silhouette\_score），你可以参考[这个页面](#)更细致地了解这个系数是怎么得到的，有什么意义。



根据数据集的统计描述提出每个客户细分所代表的类型。对集群中心的逆变换和反比例级联已在代码中正确实施。

- 请注意，在这里根据题目要求，你需要基于数据的统计特征来分析。
- 具体而言，所谓“数据的统计特征”，就是在项目一开始的时候的 `display(data.describe())` 代码得到的数据的各种统计参数。你需要将Segments中的数据与它们进行对比，并得出相关的结论。



客户细分正确识别样本数据点，讨论各样本数据点的预测集群。

### 结论



提出了某些功能改进方法，可以改进从 A/B 测试获取结果的功能。



学生讨论了聚类数据如何可以通过监督学习预测新的属性。

- 是的，聚类数据会使结果有所提升。
- 我们知道，监督学习和非监督学习解决的是两类工作。对于监督学习，它是通过(feature,label)的对来学习，最后预测label；对于非监督学习，它只通过feature本身来学习，最后能预测对应sample的label。
- 那么，在这个情景中，我们可以使用非监督学习的成果，即得到的label，来增强监督学习的结果：利用这个label加入监督学习的input feature，来给监督学习增维。



客户细分与客户通道数据进行对比，对通道数据识别客户细分的问题进行讨论，包括该表达是否符合早期结果。

重新提交

下载项目

了解 [修改和重新提交项目的最佳做法](#)。

[返回 PATH](#)

给这次审阅打分



[学员 FAQ](#)