Requires Changes

**还需满足 2 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](http://blog.csdn.net/trierwang/article/details/4855309)。

**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。**

* 是的，你的理解很好！决定系数高 → 拟合度好 → 说明该变量可以通过其他变量来预测，那么该变量对于区分顾客消费习惯来说，是**不必要**的。

**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“这些图不是正态分布的，大多数数据点聚集在一起，并出现在图的左下角”，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：

**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。**

* 你的异常值找的不太对，应当是一共有5个极端异常值。相信你找异常值的时候一定是焦头烂额，这里你可以使用Counter来计算出在多个特征中为异常值的点，以下是示例代码供你参考：
* c = Counter(no\_good\_values)
* outliers = filter(lambda item: c[item]>1,c)

**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。**

* 对PCA的理解，你可以参考[这篇文章](http://wiki.mbalib.com/wiki/%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90%E6%B3%95" \l ".E6.A1.88.E4.BE.8B.E4.B8.80.EF.BC.9A.E4.B8.BB.E6.88.90.E5.88.86.E5.88.86.E6.9E.90.E6.B3.95.E5.9C.A8.E5.95.A4.E9.85.92.E9.A3.8E.E5.91.B3.E8.AF.84.E4.BB.B7.E5.88.86.E6.9E.90.E4.B8.AD.E7.9A.84.E5.BA.94.E7.94.A8.07UNIQ33b57428c31c694-nowiki-00000023-QINU2.07UNIQ33b57428c31c694-nowiki-00000024-QINU" \t "_blank)，以及下面我的一些解读。
  + 在这边，我们使用了主成分分析法，将原来的6个特征通过数学变换，变换为了另外6个特征。对方差的计算，是为了让我们能够选择方差较大的特征以保留它们。每个新特征，实际上都是由原来的特征通过某种带权重的组合得到的，权重就是图中柱状图柱高度。考虑权重的绝对值，权重绝对值越大，说明权重对应的原特征对这个新特征带来的影响越大，反之亦反。权重若为正，则说明他们有正相关性；负值则说明它们是负相关性。A和B有正相关性可以理解为，买更多的A意味着有很大可能买更多的B；负相关性意味着买更多的A意味着有很大可能买更少的B。

**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。已给出的集群视觉化将根据已选的聚类算法生成最佳的集群数量。**

你还可以尝试更多的聚类数。

同时，你会发现，随着轮廓分数的增大，轮廓分数又会递增乃至达到一个局部极大值？这并不是一个好现象，因为当我们选择轮廓分数作为我们的评测基准时，我们希望这是一个凸函数（即只有一个峰），这样我们可以很顺利地对这个函数最优化。但是这个函数出现了多个峰，也就是除了2个聚类时对应的最大值以外，还有其他聚类情况对应的极大值。你可以搜一搜相关的信息，自己探索一下相关的问题，并可以写在回答中~

你可以参考[这个页面](http://blog.csdn.net/xueyingxue001/article/details/51966932" \t "_blank)。

**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。**

很棒！

**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。**

不错，你很好地理解了A/B测试。

**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，来实现监督学习的功能：利用得到的(feature,label)来学到一个监督学习模型。

**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。**







