Requires Changes

**还需满足 1 个要求 变化**

关于异常值的问题，实际是代码有个小错误，修改以后，应该就只有5个点了。除此之外，作业完成的非常棒。继续加油！

**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。**

你的论述很合理，并且你将采样点与数据的统计作了比较, 做的很好。

**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。**

没错，R^2得分偏高，说明Detergents\_Paper这个特征可以根据其他特征推断出来，对于区分用户的消费习惯并不是必需的。

**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。**

很不错的讨论！数据并非正态分布，一个典型的特征，就是数据的分布并不对称，否则应该存在一些负值的特征。这里的数据分布更像是log-normal。

**数据处理**

**数据和样本的特征缩放已在代码中正确实施。**

做的很好。很多算法都会假设数据是接近正态分布的，而缩放是一种常见的让数据分布变得对称的方法。另外，像K-means这样的聚类算法，是基于距离来进行划分，因此我们也会想要对数据进行缩放，以避免被某个显著的特征所影响。

**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。**

在异常值检测中有一个问题：

# **TODO**：计算给定特征的Q1（数据的25th分位点）

Q1 = np.percentile(log\_data, 25)

# **TODO**：计算给定特征的Q3（数据的75th分位点）

Q3 = np.percentile(log\_data, 75)

我们需要计算每一个特征的Q1和Q3，而非所有数据的Q1和Q3，也就是说：

Q1 = np.percentile(log\_data[feature],25)

# 同样的计算Q3

Q3 = ...

**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。**

很好的解释，并且你注意到了第三维度和第四维度存在的的符号相反的权重。

关于PCA的主要成分，在论坛曾经有过讨论, 这里给你提供一些额外的参考：  
<http://discussions.youdaxue.com/t/pca-explained-variance/7098>  
<https://discussions.youdaxue.com/t/pca/33415>  
（英文）<https://onlinecourses.science.psu.edu/stat505/node/54>

**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。**

不错的选择。 K-means算法简单明了，并且聚类效果很好，尤其适合运用于大型数据集。

作为额外的参考，K-means属于硬聚类方法，对处于类之间的数据无法很好的分离，并且对初始聚类中心比较敏感。为了避免因为初始位置选取不当而导致的收敛在局部最小值，通常我们会用不同的初始位置多次计算。

**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。**

做的很好。可以看到，当选用2个聚类中心的时候，可以得到较好的分数。

**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。**

**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。**

对客户进行分类后，对每个分类的一半客户维持每周5天的派送，而对另一半客户进行每周3天的派送，观察不同派送情况下客户的订单情况。如果变为3天后某类客户的订单增长了，则可以确定该类客户对新的派送策略有积极反馈。

你的策略很不错。这里关键是，我们需要对两个聚类分别进行独立的测试，以发现影响较大的那一个类。在实际操作上，我们可以考虑选择靠近簇中央的数据点进行A/B测试，因为它们可能最有代表性。

**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。**

**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。**

K-means的表现是很不错的, 尤其是对于那些两端的点。而对于那些在两个类的分界上的点，我们很难说它们确切的类属。这时候，GMM或许会更加有帮助，因为它会给出数据属于两类的概率。







