基于 **XXXXX**XX 的葡萄酒评价

August 24, 2022

摘要

葡萄酒备受大部分人的热爱，质量较好的葡萄酒往往带来更好的感官体验。葡萄酒的评价有着不同的指标，但其好坏却主要来自于评酒师的个人主观评价，所以会导致评价结果的差异性和不够客观性，存在着不同大众的口味喜爱偏差。本文为了解决该问题，通过探索酿酒葡萄对葡萄酒质量的影响，通过建立模型的方式更合理的对酿酒葡萄质量分类，以及如何依赖不同的酿酒葡萄和葡萄酒之间的关系来客观的给出葡萄酒评价结果，通过客观修正的数据建立了数学模型，以客观的方式来给出葡萄酒的评价。

针对问题一，由于数据较多，且分布不均的原因，首先通过对数据的筛选处理，发现两组对红葡萄酒和白葡萄酒的打分情况是两两相互比较和配对进行 K-S 检验判断数据是否符合正态分布，再对样本 T 检验进行显著性差异判断，通过方差齐性检验来假设检验，最后依据显著性差异水平来判断打分组别的稳定性和可靠性。

针对问题二，需要对酿酒葡萄进行分类，需要考虑到所酿葡萄酒的好坏。由于评酒员的打分存在主观性的干扰，我们将打分数据重新加权处理求和，得到一份较为中肯的综合评价结果。对酿酒葡萄和葡萄酒理化指标进行标准化处理，去除机值，后利用主成分分析对酿酒葡萄和葡萄酒理化指标提取主成分。将酿酒葡萄的数据和葡萄酒理化指标数据进行关联，采用改进的 K-means++ 分类模型，通过对主成分聚类分析将葡萄酒质量分为六大类，由于部分葡萄酒得分评价数值较低，以此为下限进行分类，借鉴罗伯特帕克葡萄酒评分体系依据酿酒葡萄和葡萄酒的相关性，以酿酒葡萄的理化指标含量和分布来进行排名，分出酿酒葡萄的等级和优劣。

针对问题三，探求酿酒葡萄和葡萄酒的理化指标之间的关系。首先对数据进行筛选处理出关键数据，剔除部分多余数据。提取葡萄酒某指标以及对应的酿酒葡萄理化指标进行相关性分析，通过 perason 系数和显著性水平来判断其相关性的强弱。后获取到相关性较强的理化指标数据进行共线性诊断后，进行德宾沃森残差分析后通过多元回归的方法来进行拟合，以 R 方的值来判断拟合的契合度高低，以高的模型数据为准得到葡萄酒和对应酿酒葡萄理化指标之间的系数，建立其函数关系。

针对问题四，借鉴问题二中使用到的主成分分析法先对酿酒葡萄的理化指标进行分类，提取到八大类主成分，分析各主成分中相关较高的酿酒葡萄和葡萄酒的理化指标，从而得到对葡萄酒质量贡献较大的因素指标。对评分数据进行加权求和、去极值处理后。以评分标准数据作为因变量，酿酒葡萄理化指标主成分作为自变量来进行回归分析，得出回归的评分表。通过误差分析来分析以理化指标得出的评价和已给出的评价之间的误差，来判断是否能以葡萄酒和酿酒葡萄的理化指标作为葡萄酒质量的评价标准。

关键字： K-S 检验聚类分析主成分分析 K-means++ 分类模型相关性分析多元回归 Perason 系数

1 问题背景与重述

1.1 问题背景

葡萄酒是当今世界上最畅销的酒类之一，在各种场合都有葡萄酒的身影。然而葡萄酒的酿造是取决于多种因素，各种因素的叠加会导致葡萄酒的品质差异明显。在不同的原材料已经酿制方法的差别下葡萄酒会继续细分，例如红葡萄酒和白葡萄酒等。对各种葡萄酒的鉴别是必不可少的一个步骤，而采用人工品尝打分和采用仪器进行理化指标的检验已成为最为科学的鉴别方法。最后经过安全检查、筛选分级的葡萄酒方可上市成为饮用酒。

1.2 题目所给信息及参数

此次比赛是根据 10 位品酒员为 27 款红酒和 28 款白葡萄酒的打分，以及上述葡萄酒的指标情况和芳香物质为基础进行数学分析和建模，并探讨品酒员的打分是否合理以及论证是否科学分级。红葡萄酒和白葡萄酒的市场在国际上的价值非常之高，葡萄酒依旧是未来的主力酒类，对此分析依旧存在价值。现在根据三个数据文件，并对三个数据进行分析处理后描述统计，完成数学建模和预测。

数据一：葡萄酒品尝评分表；数据二：指标总表；数据三：芳香物质。

1.3 所需解决的问题

1. 根据附件所提供的两组品酒员对 27 款红葡萄酒和 28 款白葡萄酒的打分判断两组结果是否有显著性差异，并判断哪一组的更加可信。

2. 根据酿酒葡萄的理化指标和葡萄酒的质量，使用无监督方法计算相似度，通过相似度进行分级。

3. 分析酿酒葡萄和葡萄酒的理化指标之间是否具有相关性，以及其之间具有什么样的联系。

4. 通过主成分分析对数据进行降维，得到贡献值大的特征进行表示，建立其函数关系，根据建立的函数进行预测，将预测结果与评分标准进行误差分析，判断建立的模型是否合理。

2 问题分析

针对不同的国家，地区和相对应的医疗水平进行对应的数据指标分析。主要分析感染率，病人接触率，治愈率，以及传染期接触数。模型构建还需要考虑到新冠肺炎的无症状感染者这一特殊的情况，根据这些指标进行相轨线分析，合理的进行疫情的分析和未来疫情走向以及各地区、国家的防疫政策研究。

2.1 问题一的分析

第一，根据附件 1 中给出两组品酒员的打分情况判断两组的打分是否有显著性差异。对于此问题，分析附件一所提供的数据，研究发现两组对红葡萄酒和白葡萄酒的打分情况是两两相互比较和配对，适合于先进行单样本 K-S 检验判断数据是否符合正态分布，再进行两配对样本 T 检验进行显著性差异判断的办法。

第二，判断两组品酒员的打分情况哪一组更加可信。对于此问题，分析两组品酒员的打分情况，检验两组中打分的更稳定的一方，越稳定的分数即代表品酒员偏好更少，更加可信。提取两组品酒员对于白葡萄酒和红葡萄酒的分数的标准差，根据标准差的大小进行可信性的判断。

2.2 问题二的分析

基于改进的 K-means++[3] 进行分类模型，为了降低数据数，首先对红、白葡萄和葡萄酒理化指标采用主成分分析法提取出主成分，但是经过 Bartlett 球形度检验 [1] 等发现不适合进行主成分分析，进行标准化处理，去除极值，使评价分数更加客观, 然后借鉴 Robert Parker 葡萄酒评分体系 [7]，对这些主成分聚类分析得出 6 种聚类并依据判别标准（聚类后葡萄酒样本的平均值），最终确定红、白葡萄的分级。

2.3 问题三的分析

首先要参照附件所给的数据来进行分析酿酒葡萄和葡萄酒之间的相关性，附件的信息内容过多，需要进行合理的过滤和筛选数据，但要尽可能保证其数据的完整性和真实性。我们采取相关性分析，依据相关性皮尔森系数来判断葡萄酒的数据和酿酒葡萄的理化指标之间的相关性显著程度。

在进行了相关性分析后，可以确定下一些具有显著相关的数据流，要进一步解决其葡萄酒某指标与该些理化指标的关系，则要进行其关系的拟合，从而得出实质性的结论来判断酿酒葡萄和葡萄酒之间的关系，以及其关系的可靠性。

2.4 问题四的分析

3 符号说明

4 模型假设

5 模型建立与求解

5.1 问题一的求解

问题一分析两组评酒员的评价结果有无显著性差异，并判断两组结果哪一组更加可信。采用三个步骤完成分析，步骤如下：

1) 判断数据是否符合正态分布，以选择合适模型；

2) 使用两配对 T 检验方法完成显著性检验的判断；

3) 计算标准差的大小后进行比较，较小的表示稳定性更高，更加可信。

5.1.1 数据的预处理

因为数据较大，指标较多，所以我们对各项分数相加得到总分，接着取平均值进行比较。

均值计算如下:

$$x = \sum_{x_{mn}}^{10} (m = 1, 2, 3, , , 10 n = 1, 2, 3, , , 10) \tag{1}$$

5.1.2 各葡萄酒样本评分数据概率分布的确定

对两组品酒员差异性评价的假设检验一般要求数据符合正态分布，因为两配对样本 T 检验的前提要求为数据符合正态分布，才可以使用 T 检验的数学模型。利用 SPSS 统计软件中单样本 K-S 检验 [10]，对数据集两组品酒员分别对红、白葡萄酒品尝得到的四组评价结果进行了正态分布检验。

N			27	27
正态参数[a,b]	平均值		73.056	70.515
	标准差		7.3426	3.9780
最极端差值	绝对		.157	.124
	正		.088	.078
	负		-.157	-.124
检验统计			.157	.124
渐近显著性（双尾）[c]			.085	.200[e]
蒙特卡洛显著性（双尾）[d]	显著性		.079	.344
	95% 置信区间	下限	.074	.335
		上限	.085	.353

a. 检验分布为正态分布。
b. 根据数据计算。

图 1: 聚类汇总图

		均值 1	均值 2
N		28	28
正态参数[a,b]	平均值	74.261	76.532
	标准差	5.2012	3.1709
最极端差值	绝对	.123	.122
	正	.104	.076
	负	-.123	-.122
检验统计		.123	.122
渐近显著性（双尾）[c]		.200[d]	.200[d]

a. 检验分布为正态分布。
b. 根据数据计算。

图 2: 聚类汇总图

从图 1和图 2可以看出两组的双边检验结果。因此可以认为品酒员对葡萄酒的评分服从正态分布。

5.1.3 两组评价结果的显著性差异评价

上述检验显示各类葡萄酒得分情况属于正态总体，为了进一步说明品酒员评分的科学性以及两个评分组评分的可信度，需要检查两组给出的评分是否有显著性差异，即对数据进行显著性检验。

两配对样本非参数检验一般用于同一研究对象分别给予两种不同处理的效果比较。因为两组品酒员分别对同一样本组进行评分，故两组数据为配对数据。

$$z_{li} = w_{li} - w_{2i}(i = 1, 2, , , , n) \tag{2}$$

$z_l i$ 来自正态分布, 用假设检验的方法, 假设 $H_0 : u_1 = 0$ 成立;

$$\begin{cases} \bar{z} = \frac{1}{n}\sum_{i=1}^{n}(Z_l i) \\ s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_l i - z_1)^2 \\ t = \frac{\bar{z_1}}{\frac{s_1}{\sqrt{n}}} \\ w = \mid t \mid \geq t_{1-\frac{a}{2}}(n-1) \end{cases}$$

对于统计量 t, 在给定显著性水平 a 下, 该检验问题的拒绝域是 w, 若 $|t| \geq t_{1-\frac{a}{2}}(n-1)$, 则拒绝假设反之则接受。

组数	样本数	平均值	标准差	T_1 值	p
一	27	73.056	3.9780	2.458	0.0104
二	27	70.515	7.3426	2.458	0.0104

上表给出了两组红葡萄酒评分均值的 t 检验结果，通过查表当 x =0.05，n=27 时，$t_{1-\frac{a}{2}}(n-1) = 2.0555 < 2.491$ 且方差齐性检验的 p 值为 0.0104<0.05，所以拒绝原假设，对于红葡萄酒的评价，两组评酒员的评价结果有显著性差异。因为第二组评酒员对红葡萄酒样品评分的标准差大于第一组的，第二组各评酒员得评分差异小，稳定性高，比较可信。

对于白葡萄酒采用同样的方法，得到了如下的表格：

组数	样本数	平均值	标准差	T_1 值	p
一	28	74.261	5.2012	-2.184	0.01892
二	28	76.532	3.1709	-2.184	0.01892

上表给出了两组红葡萄酒评分均值的 t 检验结果，通过查表当 $x = 0.05, n = 28$ 时，$t_{1-\frac{a}{2}}(n-1) = 2.0555 < 2.491$ 且方差齐性检验的 p 值为 0. 01892<0.05，所以拒绝原假设，对于白葡萄酒的评价，两组评酒员的评价结果有显著性差异。因为第一组评酒员对红葡萄酒样品评分的标准差大于第二组，第二组各评酒员得评分差异小，稳定性高，比较可信。

综上分别对两组葡萄酒进行 T 检验 [5]，在显著性水平为 0.05 时，得出两组评酒员的评价结果有显著性差异, 第二组评酒员的评分更可信。

5.2 问题二的求解

5.2.1 模型建立

由于没有目标等级函数，故需要使用无监督算法进行分类，通过分析无监督聚类算法的利弊，我们最终决定使用 K-means++ 聚类后的结果作为最终聚类结果。

1) 聚类分析法

聚类分析的基本原则: 将每个评价指标作为一个维度，得分看做该特征在该维度上的坐标。通过特征向量计算数据在高纬上的距离，距离越近则认为越相近。把相似程度较大的样品聚合为一类，另一些相似程度较大的样品聚合为另一类，关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到把所有样品聚合完毕。

(1) 首先将样本中的 n 个点各看作一类，此时共有 n 类。

(2) 计算类与类之间的距离，并将距离最短的两类合并成为一个新类。

(3) 再重新计算类与类之间的距离，并将并将距离最短的两类合并成为一个新类。如此不断重复，直到所有的样品合为一类。

2) 数据处理

为了降低计算量，我们计划使用 IPCA[8] 代替 PCA[4] 将评价得分进行降维。

KMO 检验 [6] 的结果显示，KMO 的值为 NaN，同时，Bartlett 球形检验的结果显示，显著性 P 值为 NaN，水平上不呈现显著性，接受原假设，也就是变量间彼此独立，则无法从中提

表 1: KMO 检验和 Bartlett 的检验

KMO 值	0.000	
Bartlett 球形度检验	近似卡方	0.000
	df	55.000
	p	NaN

取公因子，主成分分析无效，建议调整数据质量，程度为极不适合。故此，我们放弃降维的做法，采用计算 10 个品酒师评分的均值作为特征值，而不是将 10 个品酒师的评价都作为特征。

为了消除评分的主观性，得到一个更客观的评价结果，避免出现葡萄酒因为品酒师的个人喜恶等多种因素造成评分与实践的不匹配结果，M_{kl} 为第 k 个酒样品的第 l 个项目的十个评分，$x_{kl_{max}}$ 为 M_{kl} 中的最大值，$x_{kl_{min}}$ 为最小值。故，处理后的 $M_{kl_{new}}$ 有：

$$M_{kl_{new}} = \{x_{kli} | x_{kli}\ M_{kl}, x_{kli} \neq x_{kl_{max}}, x_{kli} \neq x_{kl_{min}}\} \tag{3}$$

为了克服由于指标的纲量不同对统计分析结果带来的影响需要对原始数据进行标准化处理，以消除单位、数据大小不一致等的影响，我们进行数据特征处理。x_i 为去除最大最小值后的数据，$\hat{x_i}$ 为第 i 个指标的均值，S_i 为第 i 个指标的标准差，x_{ij}^* 为标准化处理以后的指标值，则有:

$$x_{ij}^* = \frac{x_i - \hat{x_i}}{S_i} \tag{4}$$

5.2.2 模型求解

Robert Parker 独一无二的葡萄酒评分体系已经成为一款新酒能否畅销的命运指挥棒。我们学习 Robert Parker 品酒体系使用 K-means 聚类算法，根据每个人的评分结果将酒样自动聚为 6 类。

但是，传统的 K-means 算法至少有两个主要的理论缺陷：

- 已经证明该算法的最坏情况运行时间是输入大小的超多项式 [2]。
- 与最优聚类相比，所发现的近似值相对于目标函数可能是任意差的。

故此，我们使用改进后的 K-means++ 算法, 该算法通过在继续进行标准 K-means 优化迭代之前指定初始化聚类中心的过程来解决这些障碍中的第二个障碍。通过 K-means++ 初始化，该算法可以保证找到一个与最优 K-means 解具有竞争力的 $O\ logk$ 解。

1) K-mean++ 聚类算法直观上进行解释，就是分散出 k 个初始聚类中心: 从正在聚类的数据点中随机选择第一个聚类中心，然后从剩余的数据点中选择每个后续聚类中心，其概率与其与点最近的现有聚类中心的平方距离成正比。这种播种方法显著改善了 K-means 的最终误差。虽然算法中的初始选择需要额外的时间，但 K-means 部分本身在此种子设定后收敛得非常快，因此算法实际上降低了计算时间。作者使用真实和合成数据集测试了他们的方法，并且通常获得了 2 倍的速度提高，对于某些数据集，误差提高了近 1000 倍。在这些模拟中，新方法在速度和误差方面几乎总是至少与普通 k 均值一样好。

   此外，作者为他们的算法计算了近似比。K-means++ 算法保证期望值中的近似比 $O\ logk$（在算法的随机性上），其中 k 是使用的聚类数。这与普通 K-means 相反，后者可以生成任意比最优值更差的聚类 [9]。K-means++ 在相对于任意距离中具有更好的表现性能。

Algorithm 1 K-means++ 算法流程图

输入:

1: 在数据点中随机选择一个中心；

2: 设置聚类个数 K；

过程:

3: 对于尚未选择的每个数据点 x，计算 D（x），即 x 与已选择的最近中心之间的距离；

4: 使用加权概率分布随机选择一个新数据点作为新中心，其中选择概率与 D（x）2 成正比的点 x;

5: 重复步骤 2 和 3，直到选择了 k 个中心;

6: 已选择初始中心，继续使用标准 K-means 聚类; **return** 聚类结果;

2) 数据处理结果展示

使用 SPSS 软件以及 Python 进行数据处理，下面展示处理后的部分数据。

表 2: 数据处理结果展示

酒样品	香气浓度	口感浓度	澄清度	香气纯正度	总和	香气质量	平衡/整体评价	口感质量	色调	口感纯正度	口感持久性
1	-0.004004004	0.004967384	0.001187377	-0.005895169	-0.000776987	-0.004585114	-0.00224002	-0.003752651	0.016635279	-0.000777495	0.002962963
2	-0.001001001	0.007178144	-4.56684E-05	0.006215558	0.000767127	-0.00698896	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
3	0.005005005	0.007178144	-4.56684E-05	0.006215558	0.003429392	0.005030271	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
4	-0.001001001	-0.00756025	-4.56684E-05	0.006215558	0.000234674	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	-0.003703704
5	-0.013013013	-0.00756025	-4.56684E-05	-0.002434961	-0.000830232	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
6	-0.001001001	-0.00756025	-4.56684E-05	-0.002434961	-0.001362685	-0.000979345	0.000245482	0.002121064	-0.008788449	0.001130902	-0.003703704
7	-0.001001001	0.007178144	-4.56684E-05	-0.01108548	-0.001362685	-0.000979345	0.000245482	0.002121064	-0.022912743	0.001130902	0.002962963
8	-0.013013013	-0.00756025	-4.56684E-05	-0.01108548	-0.005622309	-0.00698896	-0.00389702	-0.00522108	0.005335844	-0.008411083	-0.003703704
9	0.005005005	-0.00756025	-4.56684E-05	0.006215558	0.001832033	0.005030271	0.000245482	0.002121064	0.005335844	0.001130902	-0.003703704
10	-0.001001001	0.007178144	-4.56684E-05	0.006215558	0.00129958	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	-0.003703704
11	0.005005005	0.007178144	-4.56684E-05	-0.01108548	-0.002960044	-0.000979345	0.000245482	-0.00522108	-0.022912743	-0.008411083	0.002962963
12	0.005005005	0.007178144	-4.56684E-05	-0.002434961	0.000767127	0.005030271	0.000245482	0.002121064	-0.022912743	0.001130902	0.002962963
13	-0.001001001	-0.00756025	-4.56684E-05	0.006215558	-0.002427591	-0.000979345	0.000245482	-0.00522108	-0.008788449	0.001130902	-0.003703704
14	-0.001001001	0.007178144	-4.56684E-05	-0.002434961	0.00129958	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
15	-0.001001001	0.007178144	-4.56684E-05	-0.01108548	-0.000297779	-0.00698896	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
16	-0.001001001	-0.00756025	-4.56684E-05	-0.01108548	-0.000297779	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
17	-0.001001001	0.007178144	-4.56684E-05	-0.002434961	-0.000830232	-0.000979345	0.000245482	-0.00522108	0.005335844	0.001130902	-0.003703704
18	-0.001001001	0.007178144	-4.56684E-05	-0.01108548	-0.002427591	-0.00698896	0.000245482	0.002121064	-0.022912743	0.001130902	0.002962963
19	-0.001001001	-0.00756025	-4.56684E-05	0.006215558	0.000767127	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
20	0.005005005	-0.00756025	-4.56684E-05	0.006215558	0.000234674	0.005030271	0.000245482	0.002121064	-0.022912743	0.001130902	0.002962963
21	0.005005005	0.007178144	-4.56684E-05	0.006215558	0.004494298	0.005030271	0.000245482	0.002121064	0.019460138	0.001130902	0.002962963
22	0.005005005	-0.00756025	-4.56684E-05	0.006215558	0.001832033	0.005030271	0.000245482	0.002121064	0.005335844	0.001130902	-0.003703704
23	0.005005005	0.007178144	-4.56684E-05	0.006215558	0.002364486	0.011039886	0.000245482	-0.00522108	0.005335844	0.001130902	-0.003703704
24	0.005005005	-0.00756025	-4.56684E-05	0.006215558	0.000234674	0.005030271	0.000245482	-0.00522108	0.005335844	0.001130902	-0.003703704
25	-0.001001001	-0.00756025	-4.56684E-05	0.006215558	-0.001895138	-0.000979345	0.000245482	-0.00522108	0.005335844	-0.008411083	-0.003703704
26	-0.001001001	0.007178144	-4.56684E-05	0.006215558	0.001832033	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	0.002962963
27	-0.001001001	-0.00756025	-4.56684E-05	-0.002434961	-0.000297779	-0.000979345	0.000245482	0.002121064	0.005335844	0.001130902	-0.003703704

3) 问题求解

A 分析步骤

* 根据字段进行聚类类别差异性分析;
* 根据聚类汇总分析各聚类类别的频数;
* 根据数据集聚类标注可以知道每一个样本数据被分到哪个类别;
* 聚类中心坐标可以用于分析各样本与中心点的距离;
* 对分析进行综述。

B 聚类分析结果

下表展示了定量字段差异性分析的结果，包括均值 ± 标准差的结果、F 检验结果、显著性 P 值。

* 分析每个分析项是否小于 0.05 或者 0.01（根据检验标准要求，严格的话使用 0.01）;
* 若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值 ± 标准差的方式对差异进行分析，反之则表明数据不呈现差异性。

方差分析的结果显示: 对于变量澄清度，显著性 P 值为 0.911，水平上不呈现显著性，不能拒绝原假设，说明变量澄清度在聚类分析划分的类别之间不存在显著性差异；

对于变量色调，显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，说明变量色调 2 在聚类分析划分的类别之间存在显著性差异；

对于变量香气浓度，显著性 P 值为 0.001***，水平上呈现显著性，拒绝原假设，说明变量香气浓度 2 在聚类分析划分的类别之间存在显著性差异；

对于变量口感浓度，显著性 P 值为 0.522，水平上不呈现显著性，不能拒绝原假设，说明变量口感浓度 2 在聚类分析划分的类别之间不存在显著性差异；

对于变量香气纯正度，显著性 P 值为 0.014**，水平上呈现显著性，拒绝原假设，说明变量香气纯正度 2 在聚类分析划分的类别之间存在显著性差异；

对于变量口感持久性，显著性 P 值为 0.316，水平上不呈现显著性，不能拒绝原假设，说明变量口感持久性 2 在聚类分析划分的类别之间不存在显著性差异；

对于变量口感质量，显著性 P 值为 0.445，水平上不呈现显著性，不能拒绝原假设，说明变量口感质量 2 在聚类分析划分的类别之间不存在显著性差异；

对于变量香气质量，显著性 P 值为 0.002***，水平上呈现显著性，拒绝原假设，说明变量香气质量 2 在聚类分析划分的类别之间存在显著性差异；

对于变量平衡/整体评价，显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，说明变量平衡/整体评价 2 在聚类分析划分的类别之间存在显著性差异；

对于变量口感纯正度，显著性 P 值为 0.056*，水平上不呈现显著性，不能拒绝原假设，说明变量口感纯正度 2 在聚类分析划分的类别之间不存在显著性差异；

对于变量总和，显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，说明变量总和在聚类分析划分的类别之间存在显著性差异；

表 3: 字段差异性分析 (***、**、* 分别代表 1%、5%、10% 的显著性水平)

聚类类别（平均值 ± 标准差							F	P
	类别 2(n=6)	类别 3(n=6)	类别 5(n=4)	类别 6(n=4)	类别 4(n=4)	类别 1(n=3)		
澄清度	-0.0±0.0	-0.0±0.0	0.0±0.001	-0.0±0.0	-0.0±0.0	-0.0±0.0	1.193	0.346
香气纯正度	0.006±0.0	0.006±0.0	-0.005±0.004	-0.007±0.005	-0.009±0.004	0.003±0.005	20.321	0.000***
香气浓度	0.002±0.003	0.002±0.003	-0.002±0.002	-0.007±0.007	0.002±0.003	0.001±0.003	3.627	0.016**
平衡/整体评价	0.0±0.0	0.0±0.0	-0.0±0.001	-0.001±0.002	0.0±0.0	0.0±0.0	1.01	0.437
总和	0.002±0.001	0.001±0.001	-0.0±0.001	-0.002±0.003	-0.001±0.002	-0.001±0.001	4.73	0.005***
香气质量	0.002±0.006	0.002±0.003	-0.003±0.003	-0.002±0.003	-0.001±0.005	0.001±0.003	1.322	0.293
口感质量	0.001±0.003	-0.0±0.004	-0.001±0.004	0.0±0.004	0.0±0.004	-0.0±0.004	0.182	0.966
口感浓度	0.007±0.0	-0.008±0.0	0.007±0.001	-0.008±0.0	0.007±0.0	-0.008±0.0	1642.937	0.000***
色调	0.008±0.006	0.005±0.0	0.008±0.006	0.005±0.0	-0.023±0.0	-0.013±0.008	37.772	0.000***
口感纯正度	0.001±0.0	-0.0±0.004	0.001±0.001	-0.001±0.005	-0.001±0.005	0.001±0.0	0.529	0.752
口感持久性	0.001±0.003	-0.003±0.003	0.001±0.003	-0.0±0.004	0.003±0.0	-0.001±0.004	1.909	0.136

下表展示了模型聚类的结果，包括频数，所占百分比。聚类分析的结果显示，聚类结果共分为 6 类, 不同聚类类别对应频数和百分比如下表, 并可视化展示模型聚类的结果，包括频数，所占百分比。

表 4: 聚类汇总

聚类类别	频数	百分比%
聚类类别 1	3	11.111%
聚类类别 2	6	22.222%
聚类类别 3	6	22.222%
聚类类别 4	4	14.815%
聚类类别 5	4	14.815%
聚类类别 6	4	14.815%
合计	27	100.0%

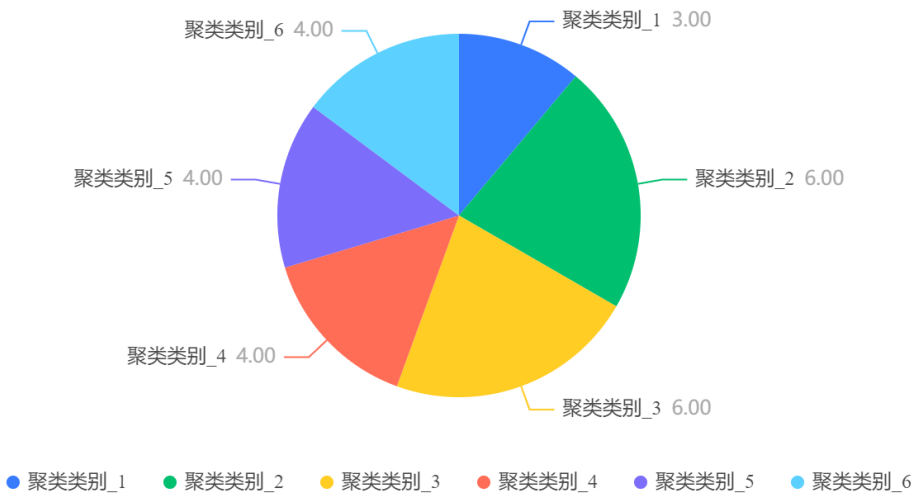

图 3: 聚类汇总图

上表格展示了模型聚类结果的部分数据聚类标注，其为预览结果，只显示综合排序的前 10 条数。

4 聚类结果分析

我们可以明显的发现，算法自动将评分接近的样本酒聚为一类，随着聚类种类编号数的上升，酒样的评分越高，聚类结果体现了我们算法的有效性。

5.3 问题三的求解

5.3.1 数据筛选

所给出酿酒葡萄和葡萄酒的数据十分多，但不能保证没一项指标都有一项对应的指标与他有着显著的相关性，所以需要进行数据指标的筛选，使得其相关性分析，更加可信。在本问中，抽选葡萄酒的花色苷、DPPH 半抑制体积、酒总黄酮、色泽这几项数据来进行数据相关性分析。

表 5: 数据集聚类标注

聚类种类	澄清度	香气纯正度	香气浓度	平衡/整体评价	总和	香气质量	口感质量	口感浓度	色调	口感纯正度	口感持久性
5	0.001187377	-0.005895169	-0.004004004	-0.00224002	-0.000776987	-0.004585114	-0.003752651	0.004967384	0.016635279	-0.000777495	0.002962963
2	-4.57E-05	0.006215558	-0.001001001	0.000245482	0.000767127	-0.00698896	0.002121064	0.007178144	0.005335844	0.001130902	0.002962963
2	-4.57E-05	0.006215558	0.005005005	0.000245482	0.003429392	0.005030271	0.002121064	0.007178144	0.005335844	0.001130902	0.002962963
3	-4.57E-05	0.006215558	-0.001001001	0.000245482	0.000234674	-0.000979345	0.002121064	-0.00756025	0.005335844	0.001130902	-0.003703704
6	-4.57E-05	-0.002434961	-0.013013013	0.000245482	-0.000830232	-0.000979345	0.002121064	-0.00756025	0.005335844	0.001130902	0.002962963
1	-4.57E-05	-0.002434961	-0.001001001	0.000245482	-0.001362685	-0.000979345	0.002121064	-0.00756025	-0.008788449	0.001130902	-0.003703704
4	-4.57E-05	-0.01108548	-0.001001001	0.000245482	-0.001362685	-0.000979345	0.002121064	0.007178144	-0.022912743	0.001130902	0.002962963
6	-4.57E-05	-0.01108548	-0.013013013	-0.00389702	-0.005622309	-0.00698896	-0.00522108	-0.00756025	0.005335844	-0.008411083	-0.003703704
3	-4.57E-05	0.006215558	0.005005005	0.000245482	0.001832033	0.005030271	0.002121064	-0.00756025	0.005335844	0.001130902	-0.003703704
2	-4.57E-05	0.006215558	-0.001001001	0.000245482	0.00129958	-0.000979345	0.002121064	0.007178144	0.005335844	0.001130902	-0.003703704
4	-4.57E-05	-0.01108548	0.005005005	0.000245482	-0.002960044	-0.000979345	-0.00522108	0.007178144	-0.022912743	-0.008411083	0.002962963
4	-4.57E-05	-0.002434961	0.005005005	0.000245482	0.000767127	0.005030271	0.002121064	0.007178144	-0.022912743	0.001130902	0.002962963
1	-4.57E-05	0.006215558	-0.001001001	0.000245482	-0.002427591	-0.000979345	-0.00522108	-0.00756025	-0.008788449	0.001130902	-0.003703704
5	-4.57E-05	-0.002434961	-0.001001001	0.000245482	0.00129958	-0.000979345	0.002121064	0.007178144	0.005335844	0.001130902	0.002962963
5	-4.57E-05	-0.01108548	-0.001001001	0.000245482	-0.000297779	-0.00698896	0.002121064	0.007178144	0.005335844	0.001130902	0.002962963
6	-4.57E-05	-0.01108548	-0.001001001	0.000245482	-0.000297779	-0.000979345	0.002121064	-0.00756025	0.005335844	0.001130902	0.002962963
5	-4.57E-05	-0.002434961	-0.001001001	0.000245482	-0.000830232	-0.000979345	-0.00522108	0.007178144	0.005335844	0.001130902	-0.003703704
4	-4.57E-05	-0.01108548	-0.001001001	0.000245482	-0.002427591	-0.00698896	0.002121064	0.007178144	-0.022912743	0.001130902	0.002962963
3	-4.57E-05	0.006215558	-0.001001001	0.000245482	0.000767127	-0.000979345	0.002121064	-0.00756025	0.005335844	0.001130902	0.002962963
1	-4.57E-05	0.006215558	0.005005005	0.000245482	0.000234674	0.005030271	0.002121064	-0.00756025	-0.022912743	0.001130902	0.002962963
2	-4.57E-05	0.006215558	0.005005005	0.000245482	0.004494298	0.005030271	0.002121064	0.007178144	0.019460138	0.001130902	0.002962963
3	-4.57E-05	0.006215558	0.005005005	0.000245482	0.001832033	0.005030271	0.002121064	-0.00756025	0.005335844	0.001130902	-0.003703704
2	-4.57E-05	0.006215558	0.005005005	0.000245482	0.002364486	0.011039886	-0.00522108	0.007178144	0.005335844	0.001130902	-0.003703704
3	-4.57E-05	0.006215558	0.005005005	0.000245482	0.000234674	0.005030271	-0.00522108	-0.00756025	0.005335844	0.001130902	-0.003703704
3	-4.57E-05	0.006215558	-0.001001001	0.000245482	-0.001895138	-0.000979345	-0.00522108	-0.00756025	0.005335844	-0.008411083	-0.003703704
2	-4.57E-05	0.006215558	-0.001001001	0.000245482	0.001832033	-0.000979345	0.002121064	0.007178144	0.005335844	0.001130902	0.002962963
6	-4.57E-05	-0.002434961	-0.001001001	0.000245482	-0.000297779	-0.000979345	0.002121064	-0.00756025	0.005335844	0.001130902	-0.003703704

表 6: 聚类中心点坐标

聚类种类	澄清度	香气纯正度	香气浓度	平衡/整体评价	总和	香气质量	口感质量	口感浓度	色调	口感纯正度	口感持久性
1	-4.57E-05	0.003332052	0.001001001	0.000245482	-0.001185201	0.00102386	-0.000326318	-0.00756025	-0.013496547	0.001130902	-0.001481481
2	-4.57E-05	0.006215558	0.002002002	0.000245482	0.002364486	0.002025463	0.000897373	0.007178144	0.007689893	0.001130902	0.000740741
3	-4.57E-05	0.006215558	0.002002002	0.000245482	0.0005009	0.002025463	-0.000326318	-0.00756025	0.005335844	-0.000459429	-0.002592593
4	-4.57E-05	-0.00892285	0.002002002	0.000245482	-0.001495799	-0.000979345	0.000285528	0.007178144	-0.022912743	-0.001254594	0.002962963
5	0.000262593	-0.005462643	-0.001751752	-0.000375894	-0.000151355	-0.003383191	-0.001182901	0.006625454	0.008160703	0.000653803	0.001296296
6	-4.57E-05	-0.00676022	-0.007007007	-0.000790144	-0.001762025	-0.002481749	0.000285528	-0.00756025	0.005335844	-0.001254594	-0.00037037

5.3.2 相关性分析

相关分析是描述两个变量间关系的密切程度，由相关系数和显著性程度值表示，当相关系数的绝对值越接近于 1，则表示两个变量间的相关性越显著，或者显著性 *p<0.05,**p<0.01 具有上述的效。双变量系数测量的主要指标有卡方类测量、Spearman 相关系数、pearson 相关系数等，由于酿酒葡萄和葡萄酒的数据为定距数据，则在进行两者间的相关性检验时用 pearson 相关系数来判断，其公式为：

$$r = \frac{\sum(x_i - \bar{x})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{5}$$

相关系数 r 的取值范围为:$-1 \leq r \leq 1$

$$\begin{cases} r > 0 \quad r < 0 \\ \mid r \mid = 0 \\ \mid r \mid = 1 \end{cases}$$

其中皮尔森简单相关系数检验统计为:

$$t = \frac{rsqrtn - 1}{\sqrt{1 - r^2}} \tag{6}$$

表 7: 红酒聚类结果分析

酒样本	聚类种类	澄清度	色调	香气浓度	口感浓度	香气纯正度	口感持久性	口感质量	香气质量	平衡/整体评价	口感纯正度	总和
7	1	3	6	4	4	3	27	13	10	8	3	59
0	2	3.1	7.6	5.5	5.7	3.6	3	13.6	10.8	8.4	3.8	68.1
1	2	3	6	6	6	5	4	16	10	9	4	71
3	2	3	6	6	4	5	5	16	12	9	4	70
4	2	3	6	4	4	4	6	16	12	9	4	68
5	2	3	4	6	4	4	7	16	12	9	4	67
14	2	3	6	6	6	3	8	16	10	9	4	69
15	2	3	6	6	4	3	9	16	12	9	4	69
16	2	3	6	6	6	4	10	13	12	9	4	68
24	2	3	6	6	4	5	11	13	12	9	3	66
26	2	3	6	6	4	4	12	16	12	9	4	69
8	3	3	6	7	4	5	13	16	14	9	4	73
9	3	3	6	6	6	5	14	16	12	9	4	72
13	3	3	6	6	6	4	15	16	12	9	4	72
18	3	3	6	6	4	5	16	16	12	9	4	71
21	3	3	6	7	4	5	17	16	14	9	4	73
22	3	3	6	7	6	5	18	13	16	9	4	74
23	3	3	6	7	4	5	19	13	14	9	4	70
25	3	3	6	6	6	5	20	16	12	9	4	73
6	4	3	2	6	6	3	21	16	12	9	4	67
10	4	3	2	7	6	3	22	13	12	9	3	64
12	4	3	4	6	4	5	23	13	12	9	4	65
17	4	3	2	6	6	3	24	16	10	9	4	65
11	5	3	2	7	6	4	25	16	14	9	4	71
19	5	3	2	7	4	5	26	16	14	9	4	70
2	6	3	6	7	6	5	28	16	14	9	4	76
20	6	3	8	7	6	5	29	16	14	9	4	78

表 8: 白酒聚类结果分析

酒样本	聚类种类	澄清度	色调	香气浓度	口感浓度	香气纯正度	口感持久性	口感质量	香气质量	平衡/整体评价	口感纯正度	总和
7	1	3	6	4	4	3	27	13	10	8	3	59
0	2	3.1	7.6	5.5	5.7	3.6	3	13.6	10.8	8.4	3.8	68.1
1	2	3	6	6	6	5	4	16	10	9	4	71
3	2	3	6	6	4	5	5	16	12	9	4	70
4	2	3	6	4	4	4	6	16	12	9	4	68
5	2	3	4	6	4	4	7	16	12	9	4	67
14	2	3	6	6	6	3	8	16	10	9	4	69
15	2	3	6	6	4	3	9	16	12	9	4	69
16	2	3	6	6	6	4	10	13	12	9	4	68
24	2	3	6	6	4	5	11	13	12	9	3	66
26	2	3	6	6	4	4	12	16	12	9	4	69
8	3	3	6	7	4	5	13	16	14	9	4	73
9	3	3	6	6	6	5	14	16	12	9	4	72
13	3	3	6	6	6	4	15	16	12	9	4	72
18	3	3	6	6	4	5	16	16	12	9	4	71
21	3	3	6	7	4	5	17	16	14	9	4	73
22	3	3	6	7	6	5	18	13	16	9	4	74
23	3	3	6	7	4	5	19	13	14	9	4	70
25	3	3	6	6	6	5	20	16	12	9	4	73
6	4	3	2	6	6	3	21	16	12	9	4	67
10	4	3	2	7	6	3	22	13	12	9	3	64
12	4	3	4	6	4	5	23	13	12	9	4	65
17	4	3	2	6	6	3	24	16	10	9	4	65
11	5	3	2	7	6	4	25	16	14	9	4	71
19	5	3	2	7	4	5	26	16	14	9	4	70
2	6	3	6	7	6	5	28	16	14	9	4	76
20	6	3	8	7	6	5	29	16	14	9	4	78

其中的 t 服从 n-2 个自由度的 t 分布

5.3.3 相关性检验

通过将筛选出的数据通过 spss 来进行相关性的检验分析，来发现针对某些特定的葡萄酒指标，有的酿酒葡萄在某个理化指标较为突出的情况下，可以根据需求来进行指定行的处理来满足要求。如下将一次选取花色苷、DPPH 半抑制体积、酒总黄酮、色泽这几项数据来进行数据相关性分析。通过对数据进行平均值和标准差统计，成对排除个案缺失值，采用 pearson 相关系数的双侧显著性检验来获取结果。如下为结果图。

相关性

		花色苷(mg/L)	多酚氧化酶活力	褐变度	DPPH自由基IC50（g/L）	酒石酸（g/L）	苹果酸（g/L）	柠檬酸（g/L）	黄酮醇(mg/kg)	单宁(mmol/L)	氨基酸总量	葡萄总黄酮(mmol/kg)	白藜芦醇(mg/kg)	总糖g/L	PH值	可滴定酸（g/l）	固酸比	干物质含量g/100g	果糖质量/g	百粒质量/g	果梗比(%)	出汁率(%)	果皮质量（g）
花色苷(mg/L)	皮尔逊相关性	1	.488*	.769**	.567**	.034	.692**	.380	.408*	.745**	.106	.442*	-.035	.005	-.038	-.215	.333	.230	-.105	-.263	.496**	.320	-.055
	显著性（双尾）		.010	<.001	.002	.867	<.001	.051	.034	<.001	.598	.021	.863	.979	.852	.282	.089	.249	.602	.186	.008	.103	.785
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
多酚氧化酶活力	皮尔逊相关性	.488*	1	.554**	.098	-.127	.344	.194	.303	.156	-.103	-.036	-.218	-.014	-.311	-.058	.127	.262	.010	-.161	.279	.038	.195
	显著性（双尾）	.010		.003	.628	.528	.079	.332	.125	.438	.608	.851	.274	.944	.114	.774	.529	.187	.960	.422	.159	.849	.329
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
褐变度	皮尔逊相关性	.769**	.554**	1	.300	-.060	.645**	.394*	.423*	.449*	-.154	.240	.033	-.141	-.248	-.288	.400*	.082	-.269	-.391*	.502**	.202	-.068
	显著性（双尾）	<.001	.003		.128	.768	<.001	.042	.028	.019	.442	.229	.871	.483	.213	.146	.039	.684	.175	.044	.008	.311	.735
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
DPPH自由基IC50（g/L）	皮尔逊相关性	.567**	.098	.300	1	.276	.052	-.003	.428*	.753**	.106	.837**	.211	-.020	.376	-.344	.170	-.091	.006	-.136	.467*	.465*	-.061
	显著性（双尾）	.002	.628	.128		.163	.795	.987	.026	<.001	.600	<.001	.290	.922	.053	.079	.397	.653	.975	.499	.010	.015	.763
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
酒石酸（g/L）	皮尔逊相关性	.034	-.127	-.060	.276	1	-.111	.515**	.143	.281	.364	.247	.166	.232	.275	-.244	.260	.241	-.025	-.390*	.141	.084	-.268
	显著性（双尾）	.867	.528	.768	.163		.581	.006	.478	.155	.062	.214	.408	.245	.165	.220	.191	.226	.900	.044	.483	.676	.177
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
苹果酸（g/L）	皮尔逊相关性	.692**	.344	.645**	.052	-.111	1	.423*	.056	.299	-.072	.052	.110	.093	-.366	-.061	.307	.299	-.257	-.308	.235	.097	-.263
	显著性（双尾）	<.001	.079	<.001	.795	.581		.028	.782	.130	.720	.796	.586	.644	.060	.761	.119	.130	.195	.118	.237	.629	.185
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
柠檬酸（g/L）	皮尔逊相关性	.380	.194	.394*	-.003	.515**	.423*	1	.223	.145	.083	-.059	.200	-.002	-.199	-.279	.219	.239	-.061	-.383*	.135	-.053	-.132
	显著性（双尾）	.051	.332	.042	.987	.006	.028		.264	.469	.681	.772	.318	.993	.320	.158	.272	.230	.761	.049	.503	.793	.510
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
黄酮醇(mg/kg)	皮尔逊相关性	.408*	.303	.423*	.428*	.143	.056	.223	1	.580**	.465*	.263	.183	.055	.052	-.139	.134	.215	-.201	-.293	.633**	.045	.118
	显著性（双尾）	.034	.125	.028	.026	.478	.782	.264		.002	.014	.185	.361	.786	.796	.490	.507	.281	.314	.138	<.001	.825	.558
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
单宁(mmol/L)	皮尔逊相关性	.745**	.156	.449*	.753**	.281	.299	.145	.580**	1	.498**	.685**	.049	.363	.222	-.066	.261	.414*	-.271	-.330	.474*	.352	-.124
	显著性（双尾）	<.001	.438	.019	<.001	.155	.130	.469	.002		.008	<.001	.810	.063	.265	.743	.189	.032	.171	.093	.013	.072	.539
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
氨基酸总量	皮尔逊相关性	.106	-.103	-.154	.106	.364	-.072	.083	.465*	.498**	1	.129	.051	.507**	.234	.206	-.012	.601**	-.363*	-.366	.264	-.009	-.322
	显著性（双尾）	.598	.608	.442	.600	.062	.720	.681	.014	.008		.521	.800	.007	.241	.303	.954	<.001	.048	.061	.184	.965	.101
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
葡萄总黄酮（mmol/kg）	皮尔逊相关性	.442*	-.036	.240	.837**	.247	.052	-.059	.263	.685**	.129	1	.020	.210	.239	-.178	.231	.067	-.090	-.071	.259	.568**	-.016
	显著性（双尾）	.021	.851	.229	<.001	.214	.796	.772	.185	<.001	.521		.920	.293	.230	.374	.247	.739	.656	.726	.191	.002	.936
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
白藜芦醇(mg/kg)	皮尔逊相关性	-.035	-.218	.033	.211	.166	.110	.200	.183	.049	.051	.020	1	-.074	.173	-.410*	.093	-.142	-.192	-.346	.220	-.085	-.392*
	显著性（双尾）	.863	.274	.871	.290	.408	.586	.318	.361	.810	.800	.920		.713	.387	.033	.644	.479	.337	.077	.269	.674	.043
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
总糖g/L	皮尔逊相关性	.005	-.014	-.141	-.020	.232	.093	-.002	.055	.363	.507**	.210	-.074	1	.166	.196	.154	.804**	-.274	-.259	-.212	.211	-.159
	显著性（双尾）	.979	.944	.483	.922	.245	.644	.993	.786	.063	.007	.293	.713		.407	.326	.444	<.001	.167	.191	.288	.291	.427
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
PH值	皮尔逊相关性	-.038	-.311	-.248	.376	.275	-.366	-.199	.052	.222	.234	.239	.173	.166	1	-.505**	.351	-.056	.121	-.036	.018	.283	-.184
	显著性（双尾）	.852	.114	.213	.053	.165	.060	.320	.796	.265	.241	.230	.387	.407		.007	.072	.783	.549	.857	.929	.152	.358
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
可滴定酸（g/l）	皮尔逊相关性	-.215	-.058	-.288	-.344	-.244	-.061	-.279	-.139	-.066	.206	-.178	-.410*	.166	-.505**	1	-.650**	.341	-.075	.227	-.285	-.056	.276
	显著性（双尾）	.282	.774	.146	.079	.220	.761	.159	.490	.743	.303	.374	.033	.326	.007		<.001	.082	.709	.255	.150	.780	.163
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
固酸比	皮尔逊相关性	.333	.127	.400*	.170	.260	.307	.219	.134	.261	-.012	.231	.093	.154	.351	-.650**	1	.023	.035	-.211	.061	.112	-.168
	显著性（双尾）	.089	.529	.039	.397	.191	.119	.272	.507	.189	.954	.247	.644	.444	.072	<.001		.908	.861	.292	.763	.577	.402
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
干物质含量g/100g	皮尔逊相关性	.230	.262	.082	-.091	.241	.299	.239	.215	.414*	.601**	.067	-.142	.804**	-.056	.341	.023	1	-.401*	-.368	-.055	.119	-.159
	显著性（双尾）	.249	.187	.684	.653	.226	.130	.230	.281	.032	<.001	.739	.479	<.001	.783	.082	.908		.038	.059	.784	.555	.427
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
果糖质量/g	皮尔逊相关性	-.105	.010	-.269	.006	-.025	-.257	-.061	-.201	-.271	-.383*	-.090	-.192	-.274	.121	-.075	.035	-.401*	1	.678**	-.350	.085	.633**
	显著性（双尾）	.602	.960	.175	.975	.900	.195	.761	.314	.171	.048	.656	.337	.167	.549	.709	.861	.038		<.001	.074	.674	<.001
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
百粒质量/g	皮尔逊相关性	-.263	-.161	-.391*	-.136	-.390*	-.308	-.383*	-.293	-.330	-.366	-.071	-.346	-.259	-.036	.227	-.211	-.368	.678**	1	-.386*	-.101	.709**
	显著性（双尾）	.186	.422	.044	.499	.044	.118	.049	.138	.093	.061	.726	.077	.191	.857	.255	.292	.059	<.001		.047	.616	<.001
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
果梗比(%)	皮尔逊相关性	.496**	.279	.502**	.467*	.141	.235	.135	.633**	.474*	.264	.259	.220	-.212	.018	-.285	.061	-.055	-.350	-.386*	1	.128	-.247
	显著性（双尾）	.008	.159	.008	.010	.483	.237	.503	<.001	.013	.184	.191	.269	.288	.929	.150	.763	.784	.074	.047		.524	.215
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
出汁率(%)	皮尔逊相关性	.320	.038	.202	.465*	.084	.097	-.053	.045	.352	-.009	.568**	-.085	.211	.283	-.056	.112	.119	.085	-.101	.128	1	-.039
	显著性（双尾）	.103	.849	.311	.015	.676	.629	.793	.825	.072	.965	.002	.674	.291	.152	.780	.577	.555	.674	.616	.524		.848
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
果皮质量（g）	皮尔逊相关性	-.055	.195	-.068	-.061	-.268	-.263	-.132	.118	-.124	-.322	-.016	-.392*	-.159	-.184	.276	-.168	-.159	.633**	.709**	-.247	-.039	1
	显著性（双尾）	.785	.329	.735	.763	.177	.185	.510	.558	.539	.101	.936	.043	.427	.358	.163	.402	.427	<.001	<.001	.215	.848	
	个案数	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27

*. 在 0.05 级别（双尾），相关性显著。

**. 在 0.01 级别（双尾），相关性显著。

图 4: 花色苷相关性数据

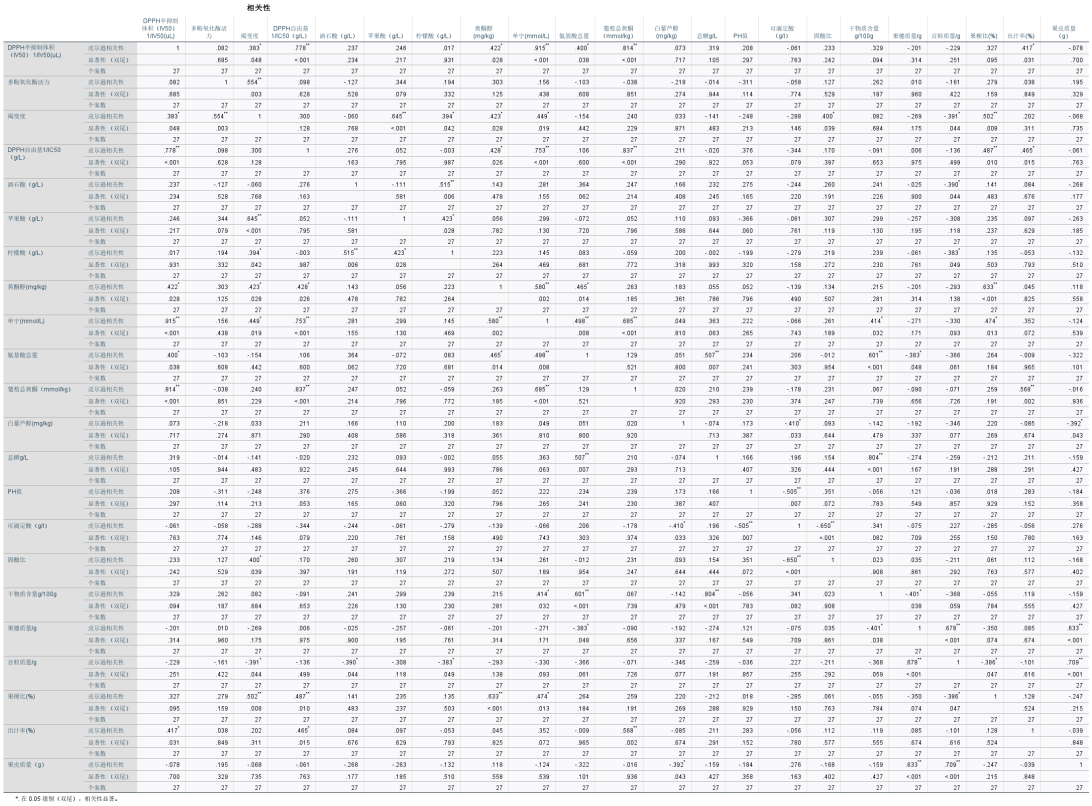

图 5: DPPH 半抑制体积相关性数据

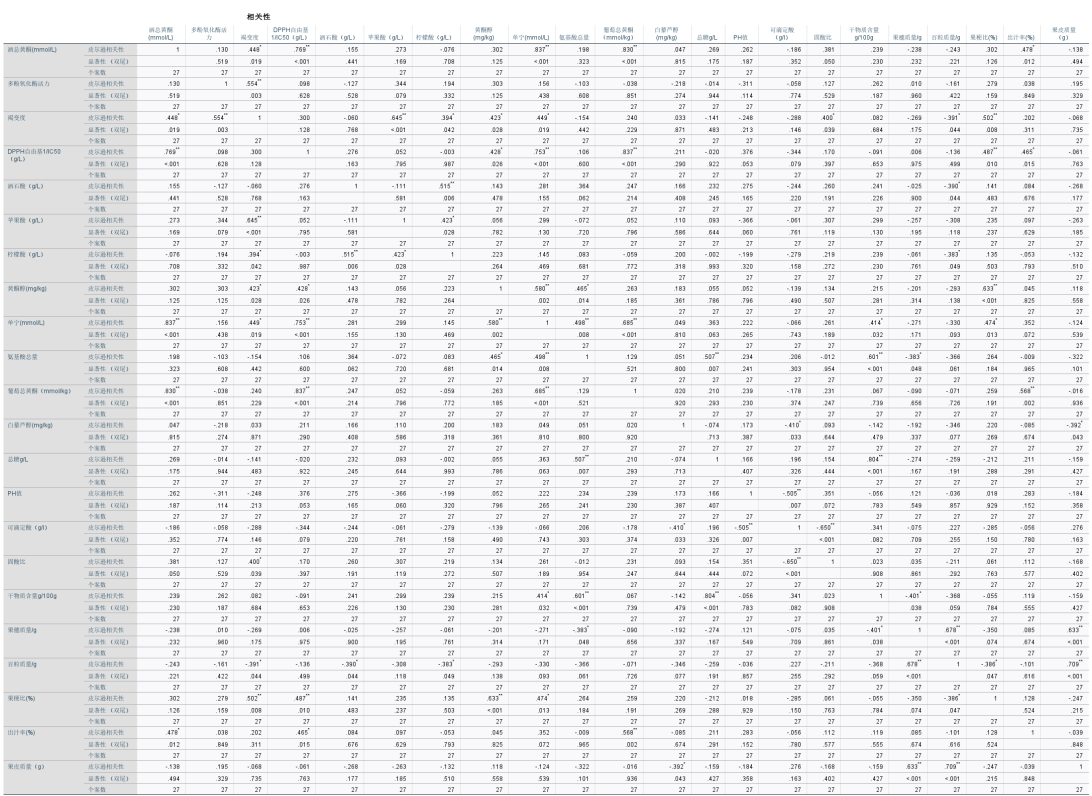

图 6: 酒总黄酮相关性数据

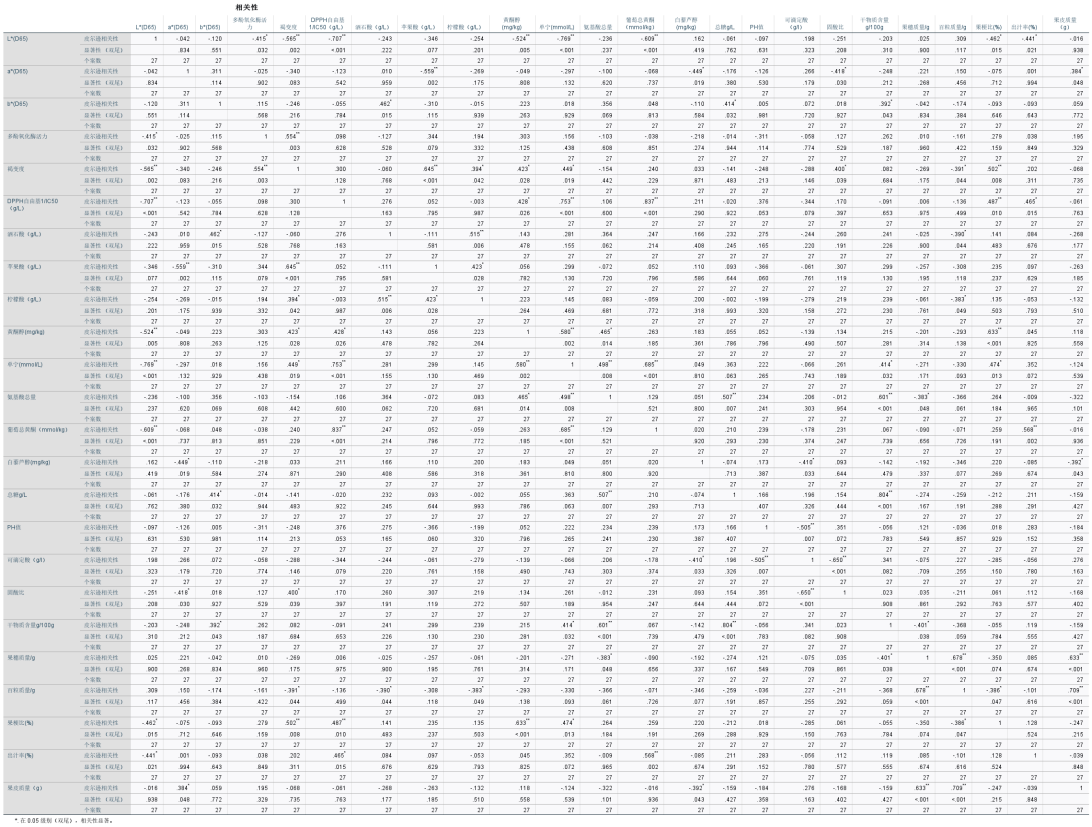
相关性
L*(D65)
a*(D65)
b*(D65)
多酚氧化酶活力
褐变度
DPPH自由基1/IC50 (g/L)
茶石酸 (g/L)
苹果酸 (g/L)
柠檬酸 (g/L)
表酚酚(mg/kg)
单宁(mmol/L)
氨基酸总量
葡萄总类酮 (mmol/kg)
白蛋产酚(mg/kg)
总糖g/L
PH值
可滴定酸(g/L)
固酸比
干物质含量g/100g
果糖质量/g
百粉质量/g
果酸比(%)
出汁率(%)
果皮质量(g)

图 7: 色泽相关性数据

5.3.4 结果分析

对相关性分析的结果进行分析，其重点就是观察其 pearson 和双侧显著性的值，pearson 相关系数的值大，则相关性高，同理观察算观测双侧显著性则是判断其值的范围，*p<0.05 或者 **p<0.01 都证明其有较好的相关显著性。通过对结果进行分析，得出如下的一些较为明显的相关性结论:

1) 花色苷的指标与褐变度、苹果酸、单宁、果梗比等数据的相关性较为显著

2) DPPH 半抑制体积与多酚氧化酶活力、DPPH 自由基、单宁、葡萄总黄酮、白藜芦醇相关性高，与果皮质量和果穗质量等指标都成负相关

3) 酒总黄酮则跟 DPPH 自由基、单宁、葡萄总黄酮，白藜芦醇等指标相关性高

4) 色泽跟可滴定酸、果穗质量的略微相关，总体数据的相关性不大

按照分析标准来分析，无论是相关性系数和显著程度都选取较为明显的那组数据来进行后面的关系分析。本小问中，在花色苷则这项数据所给出的相关性指标的相关性都较高，在多元回归中则考虑采取这项数据来进行拟合。

5.3.5 多元线性回归模型的求解

在建立模型则需要对模型进行拟合度检验，多元回归方程的显著性检验就是检验样本回归方程的变量的线性关系是否显著，需要根据样本来判断方程中的多个回归系数中至少有一个不等于 0，主要是说明样本回归方程的显著性。检验的方法用方差分析，这时因变量的总体为回归平方和与误差平方和，即表示为：

$$L_x x = Q + U \tag{7}$$

其中该公式又可以表示为：

$$L_x x = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{8}$$

$$Q = \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{9}$$

$$U = \sum_{i=1}^{N} (\hat{y_i} - \bar{y})^2 \tag{10}$$

对花色苷和其相关性较高的指标进行拟合，依据德宾沃森残差，离群值为 3 标准差进行模型拟合，来根据其 R 方及其德宾沃森残差来观察其拟合的契合度。另外进行单因素方差分析 (ANOVA) 来观测 F 检验对整个回归进行显著性检验，考虑的 k 个变量自变量是否有显著性线性关系 F 检测通过与 F 边界值来进行比对判断其水平显著性

$$\begin{cases} F_0.05(k, n-k-1) \le F \le F_0.01(k, n-k-1) & 0.05 \\ F_0.1(k, n-k-1) \le F \le F_0.05(k, n-k-1) & 0.01 \\ F < F_0.1(k, n-k-1) \end{cases}$$

5.3.6 单因素方差分析

对花色苷等多项相关性数据进行 ANOVA 分析，观察 F 检测值和显著性，通过三组不同数据的模型，对数据进行共线性诊断，依据 VIF 值来确定较为合理的自变量，来进行试验观察。得出如下的结果：

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	1220392.299	9	135599.144	14.906	<.001[b]
	残差	154643.029	17	9096.649		
	总计	1375035.327	26			
2	回归	1332224.511	15	88814.967	22.821	<.001[c]
	残差	42810.817	11	3891.892		
	总计	1375035.327	26			
3	回归	1333687.803	16	83355.488	20.160	<.001[d]
	残差	41347.524	10	4134.752		
	总计	1375035.327	26			

a. 因变量：花色苷(mg/L)

b. 预测变量：(常量), 单宁(mmol/L), 柠檬酸（g/L）, 出汁率(%), 果梗比(%), 苹果酸（g/L）, 褐变度, 黄酮醇(mg/kg), 葡萄总黄酮（mmol/kg）, DPPH自由基1/IC50（g/L）

c. 预测变量：(常量), 单宁(mmol/L), 柠檬酸（g/L）, 出汁率(%), 果梗比(%), 苹果酸（g/L）, 褐变度, 黄酮醇(mg/kg), 葡萄总黄酮（mmol/kg）, DPPH自由基1/IC50（g/L）, 白藜芦醇(mg/kg), PH值, 多酚氧化酶活力, 酒石酸（g/L）, 总糖g/L, 氨基酸总量

d. 预测变量：(常量), 单宁(mmol/L), 柠檬酸（g/L）, 出汁率(%), 果梗比(%), 苹果酸（g/L）, 褐变度, 黄酮醇(mg/kg), 葡萄总黄酮（mmol/kg）, DPPH自由基1/IC50（g/L）, 白藜芦醇(mg/kg), PH值, 多酚氧化酶活力, 酒石酸（g/L）, 总糖g/L, 氨基酸总量, 可滴定酸（g/l）

图 8: ANOVA

分析观察三组回归的数据，发现三个模型的显著性都是 <0.01，其 F 检测值分别为：14.906、22.821、20.160。根据该结果和显著性 p 值可以拒绝原假设，认为被解释变量个解释变量间存在显著的线性关系，可建立线性回归模型。

如下为残差统计图：

残差统计[a]

	最小值	最大值	平均值	标准偏差	个案数
预测值	41.21715546	1000.019653	263.3166736	226.4855067	27
标准预测值	-.981	3.253	.000	1.000	27
预测值的标准误差	35.139	61.535	50.582	6.824	27
调整后预测值	62.05002975	1292.442383	274.1753193	256.8458629	27
残差	-74.4163361	109.3441620	.0000000000	39.87843273	27
标准残差	-1.157	1.700	.000	.620	27
学生化残差	-1.919	2.182	-.038	1.048	27
剔除残差	-319.314911	233.3688507	-10.8586457	135.3274944	27
学生化剔除残差	-2.290	2.861	-.036	1.177	27
马氏距离	6.801	22.847	15.407	4.248	27
库克距离	.002	1.328	.190	.315	27
居中杠杆值	.262	.879	.593	.163	27

a. 因变量：花色苷(mg/L)

图 9: 残差统计

在经过处理后的数据是比较合理的。

5.3.7 多元回归拟合

花色苷作为因变量，褐变度、苹果酸、单宁、果梗比、DPPH 自由基、葡萄总黄酮、多酚氧化酶活力、总酚等相关性较为显著等数据作为自变量来进行多元回归拟合。在进行比对后发现并无需要严格剔除的数据，则进行多元线性回归变量筛选结果及系数的拟合求解。观察系数表和其显著性指标，通过 R^2 来判断其回归拟合的契合度，往往 R^2 越贴近于 1，契合度越高。确定酿酒葡萄与葡萄酒理化指标的联系则将系数组合成回归方程即可。

系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-420.632	212.575		-1.979	.064
	苹果酸（g/L）	19.012	7.680	.315	2.475	.024
	果梗比(%)	-2.420	25.326	-.012	-.096	.925
	柠檬酸（g/L）	20.007	29.573	.064	.677	.508
	DPPH自由基1/IC50（g/L）	701.641	385.535	.341	1.820	.086
	黄酮醇(mg/kg)	-.569	.756	-.100	-.752	.462
	出汁率(%)	1.669	3.176	.053	.525	.606
	葡萄总黄酮（mmol/kg）	-11.366	8.164	-.241	-1.392	.182
	褐变度	.229	.089	.334	2.564	.020
	单宁(mmol/L)	35.187	12.618	.444	2.789	.013
2	(常量)	-452.439	264.020		-1.714	.115
	苹果酸（g/L）	21.635	6.040	.358	3.582	.004
	果梗比(%)	7.296	19.974	.036	.365	.722
	柠檬酸（g/L）	79.312	30.080	.254	2.637	.023
	DPPH自由基1/IC50（g/L）	-133.202	598.655	-.065	-.223	.828
	黄酮醇(mg/kg)	-.740	.578	-.130	-1.281	.227
	出汁率(%)	1.381	2.284	.044	.604	.558
	葡萄总黄酮（mmol/kg）	3.247	7.805	.069	.416	.685
	褐变度	.002	.105	.003	.018	.986
	单宁(mmol/L)	57.834	18.069	.730	3.201	.008
	多酚氧化酶活力	5.585	2.247	.240	2.485	.030
	酒石酸（g/L）	-13.320	6.718	-.186	-1.983	.073
	氨基酸总量	.000	.019	.003	.021	.983
	白藜芦醇(mg/kg)	-4.099	3.303	-.098	-1.241	.240
	总糖g/L	-2.760	1.061	-.294	-2.601	.025
	PH值	161.159	77.717	.175	2.074	.062
3	(常量)	-180.363	532.189		-.339	.742
	苹果酸（g/L）	21.524	6.228	.356	3.456	.006
	果梗比(%)	2.492	22.116	.012	.113	.913
	柠檬酸（g/L）	70.579	34.305	.226	2.057	.067
	DPPH自由基1/IC50（g/L）	-123.926	617.248	-.060	-.201	.845
	黄酮醇(mg/kg)	-.727	.596	-.128	-1.220	.250
	出汁率(%)	2.169	2.702	.069	.803	.441
	葡萄总黄酮（mmol/kg）	1.564	8.528	.033	.183	.858
	褐变度	-.013	.111	-.018	-.113	.912
	单宁(mmol/L)	60.830	19.293	.768	3.153	.010
	多酚氧化酶活力	5.367	2.345	.231	2.288	.045
	酒石酸（g/L）	-12.987	6.948	-.182	-1.869	.091
	氨基酸总量	.004	.020	.024	.175	.865
	白藜芦醇(mg/kg)	-4.647	3.527	-.111	-1.318	.217
	总糖g/L	-2.826	1.099	-.301	-2.570	.028
	PH值	102.473	127.078	.111	.806	.439
	可滴定酸（g/l）	-12.635	21.239	-.075	-.595	.565

a. 因变量：花色苷(mg/L)

图 10: 系数

通过如下图表判断 R^2

模型摘要[d]

模型	R	R 方	调整后 R 方	标准估算的错误	德宾-沃森
1	.942[a]	.888	.828	95.37635321	
2	.984[b]	.969	.926	62.38503369	
3	.985[c]	.970	.922	64.30204065	1.568

a. 预测变量：(常量), 单宁(mmol/L), 柠檬酸（g/L）, 出汁率(%), 果梗比(%), 苹果酸（g/L）, 褐变度, 黄酮醇(mg/kg), 葡萄总黄酮（mmol/kg）, DPPH自由基1/IC50（g/L）

b. 预测变量：(常量), 单宁(mmol/L), 柠檬酸（g/L）, 出汁率(%), 果梗比(%), 苹果酸（g/L）, 褐变度, 黄酮醇(mg/kg), 葡萄总黄酮（mmol/kg）, DPPH自由基1/IC50（g/L）, 白藜芦醇(mg/kg), PH值, 多酚氧化酶活力, 酒石酸（g/L）, 总糖g/L, 氨基酸总量

c. 预测变量：(常量), 单宁(mmol/L), 柠檬酸（g/L）, 出汁率(%), 果梗比(%), 苹果酸（g/L）, 褐变度, 黄酮醇(mg/kg), 葡萄总黄酮（mmol/kg）, DPPH自由基1/IC50（g/L）, 白藜芦醇(mg/kg), PH值, 多酚氧化酶活力, 酒石酸（g/L）, 总糖g/L, 氨基酸总量, 可滴定酸（g/l）

d. 因变量：花色苷(mg/L)

图 11: 模型摘要

通过上述分析得出，三组数据的 R^2 都相对趋近于 1，具有较高的契合度。比对之后选择模型 3 的数据来进行方程的建立。由如下的系数表来得到系数关系。选取苹果酸、果梗比、柠檬酸、DPPH 自由基、黄酮醇、出汁率..... 等指标作为自变量 $x_1, x_2, , , , , x_n$ 构建如下方程：

$$y = 21.524x_1 + 2.49x_2 + 70.579x_3 - 123.926x_4 - 7.27x_5 + 2.169x_6 + 1.564x_7 - 0.013x_8 + 60.83x_9 \quad (11)$$

$$+5.367x_{10} - 12.987x_{11} + 0.004x_{12} - 4.647x_{13} - 2.826x_{14} + 102.473x_{15} - 12.635x_{16} - 420.632 \quad (12)$$

其中 y 为因变量花色苷，$x_1, x_2, , , , , x_n$ 则为自变量苹果酸、果梗比、柠檬酸、DPPH 自由基、黄酮醇、出汁率..... 通过该方程可以反映出葡萄酒的某些指标与酿酒葡萄理化指标之间的关系。

5.4 问题四的求解

针对该小问，要分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并判断能否将其用来作为评价葡萄酒的质量的标准。采用问题二所提出的主成分分析法，对酿酒葡萄和葡萄酒理化指标进行主成分提取，通过对主成分中的占比分析来判断对质量影响较大的理化指标。后通过回归分析法，对葡萄酒重新进行打分，比对给出的数据来判断其以理化指标来评价质量是否合理。

5.4.1 主成分分析

首先进行 KMO 和 Bartlett 的检验，对于 KMO 值：0.8 上则合适做主成分分析，0.7-0.8 之间一般适合 0.6-0.7 之间不太适合，0.5-0.6 之间表示差，0.5 下表示极不适合，对于 Bartlett 的检验（$p < 0.05$, 严格来说 $p < 0.01$），若显著性小于 0.05 或 0.01，拒绝原假设，则说明可以做主成分分析，若不拒绝原假设，则说明这些变量可能独立提供一些信息，不适合做主成分分析。后通过分析方差解释表格和碎石图，确定主成分的数量方差解释表格主要是看主成分对于变量解释的贡献率。通过下图判断得出主成分分析条件合理。

KMO 检验的结果显示，KMO 的值为 0.851，同时，Bartlett 球形检验的结果显示，显著性 P 值为 0.02，水平上呈现显著性，不接受原假设，也就是变量间耦合程度搞，则可以从中提取公因子，主成分分析有效.

表 9: KMO 检验和 Bartlett 的检验

KMO 值		0.851
Bartlett 球形度检验	近似卡方	1.238
	df	165.54
	p	0.02

据此进行主成分分析得到结果如下, 下表为总方差解释表格，主要是看主成分对于变量解释的贡献率（可以理解为究竟需要多少主成分才能把变量表达为 100%），一般都要表达到 80% 以上才可以，否则就要调整因子数据。一般情况下，方差解释率越高，说明该主成分越重要，权重占比也应该越高。

方差解释表中，在主成分 9 时，总方差解释的特征根低于 1.0，变量解释的贡献率达到 86.558%，仅为参考，若特征根小于 1.0 临界值过大，也可以集合具体情况具体分析。

$$
\begin{aligned}
A_n &= \pi_{\text{第一年}} - \pi_{\text{最后一年}} \\
R_n &= \frac{\pi_{\text{第一年}}}{\pi_{\text{最后一年}}}
\end{aligned}
\tag{13}
$$

下面碎石图是根据各主成分对数据变异的解释程度绘制的图。其作用是根据特征值下降的坡度来确认需要选择的主成分个数，我们结合方差解释表用于确认或调整主成分个数。每一个主成分为一个点，通过“坡度趋于平缓”的未知判断提取主成分的数量。我们综合方差解释以及碎石图，最终确定使用前 8 个特征作为表征值。

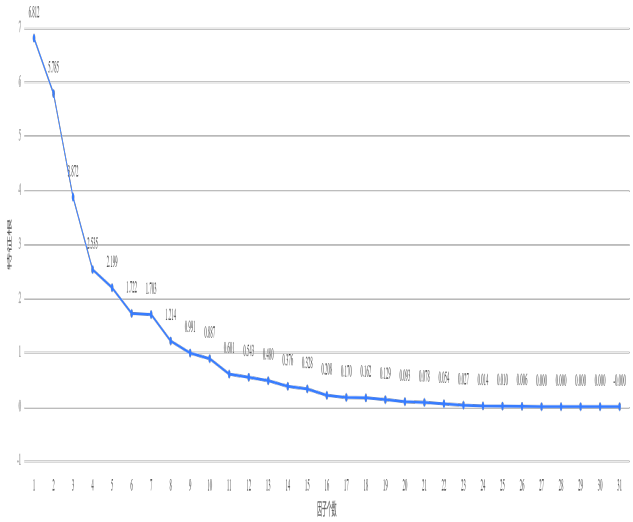

图 12: 碎石图

下表为因子载荷系数表，可以分析到每个主成分中隐变量的重要性。假设前文确定得到 n 个因子，因子 i 中 a、b、c、d 的因子载荷系数较大，因此可将主成分 i 进行总结重命名。

下图为载荷矩阵热力图，可以分析到每个主成分中隐变量的重要性。我们结合热力图进行各因子的隐变量分析。

表 10: 方差解释

成分	特征根		
特征根	方差百分比	累积	
1	6.812	21.973%	21.973%
2	5.785	18.662%	40.635%
3	3.872	12.491%	53.126%
4	2.535	8.176%	61.302%
5	2.199	7.092%	68.395%
6	1.722	5.556%	73.951%
7	1.703	5.494%	79.445%
8	1.214	3.916%	83.36%
9	0.991	3.198%	86.558%
10	0.887	2.862%	89.419%
11	0.601	1.939%	91.358%
12	0.543	1.753%	93.111%
13	0.480	1.548%	94.659%
14	0.376	1.213%	95.872%
15	0.328	1.059%	96.93%
16	0.208	0.672%	97.602%
17	0.170	0.549%	98.151%
18	0.162	0.522%	98.673%
19	0.129	0.417%	99.091%
20	0.093	0.299%	99.39%
21	0.078	0.252%	99.642%
22	0.054	0.174%	99.816%
23	0.027	0.086%	99.901%
24	0.014	0.046%	99.947%
25	0.010	0.033%	99.98%
26	0.006	0.02%	100.0%
27	0.000	0.0%	100.0%
28	-0.000	-0.0%	100.0%
29	-0.000	-0.0%	100.0%
30	-0.000	-0.0%	100.0%
31	-0.000	-0.0%	100.0%

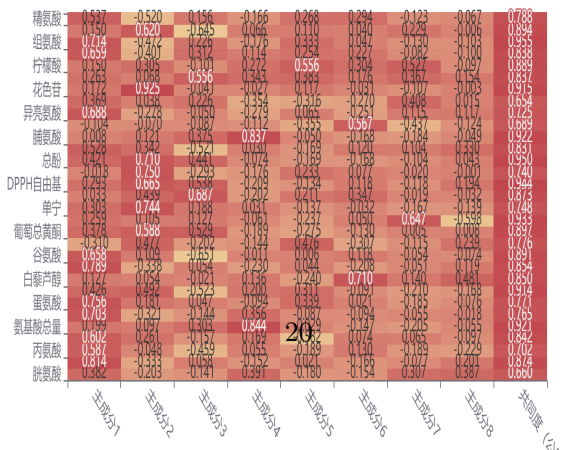

表 11: 因子权重分析

	主成分 1	主成分 2	主成分 3	主成分 4	主成分 5	主成分 6	主成分 7	主成分 8	共同度（公因子方差）
胱氨酸	0.382	-0.203	-0.141	0.391	-0.18	-0.154	0.307	0.387	0.66
缬氨酸	0.814	-0.333	0.058	-0.152	0.021	-0.166	-0.077	0.201	0.874
丙氨酸	0.587	-0.048	-0.459	0.055	-0.189	0.13	-0.189	-0.229	0.702
天门冬氨酸	0.602	0.261	-0.157	0.169	-0.582	0.074	0.065	-0.097	0.842
氨基酸总量	0.199	0.097	0.303	0.844	-0.029	-0.147	-0.205	-0.055	0.921
丝氨酸	0.703	-0.321	-0.144	0.356	-0.087	-0.094	-0.055	-0.018	0.765
蛋氨酸	0.756	0.181	0.047	-0.094	0.339	0.021	-0.185	-0.076	0.771
苏氨酸	0.426	0.492	-0.523	0.2	0.311	-0.091	-0.21	-0.168	0.914
白藜芦醇	0.127	0.054	-0.021	0.136	-0.24	0.71	0.14	0.481	0.85
亮氨酸	0.789	-0.338	0.054	-0.23	0.044	-0.208	-0.051	0.119	0.854
谷氨酸	0.658	0.109	-0.651	-0.011	0.006	0.118	-0.054	-0.074	0.891
多酚氧化酶活力	-0.31	0.477	-0.202	-0.144	0.476	-0.307	-0.115	0.239	0.776
葡萄总黄酮	0.378	0.588	0.529	-0.189	-0.275	-0.13	0.005	0.008	0.897
苯丙氨酸	0.259	0.105	0.127	-0.061	-0.237	0.05	0.647	-0.598	0.933
单宁	0.303	0.744	0.188	0.031	-0.137	-0.032	-0.167	-0.139	0.748
蛋白质	0.033	0.439	0.687	-0.202	0.211	0.347	-0.019	-0.032	0.873
DPPH 自由基	0.293	0.665	0.538	-0.209	-0.134	0.118	-0.118	0.194	0.944
褐变度	-0.018	0.75	-0.293	-0.176	0.233	0.077	-0.02	-0.001	0.74
总酚	0.421	0.71	0.441	-0.074	-0.169	-0.168	-0.104	0.043	0.95
甘氨酸	0.528	0.342	-0.521	0.03	-0.169	0.12	0.104	0.33	0.831
脯氨酸	0.008	0.121	0.375	0.837	-0.05	-0.168	-0.184	-0.049	0.922
VC 含量	-0.104	-0.07	-0.087	-0.112	-0.355	0.567	-0.437	-0.124	0.689
异亮氨酸	0.688	-0.278	-0.05	-0.279	0.065	-0.251	0.115	0.117	0.725
酪氨酸	0.369	0.038	0.226	-0.354	-0.316	-0.27	0.408	0.015	0.654
花色苷	0.172	0.925	-0.041	-0.027	0.117	-0.031	-0.107	-0.003	0.915
酒石酸	0.263	0.068	0.556	0.343	0.383	0.176	0.367	0.154	0.837
柠檬酸	0.15	0.305	-0.101	0.272	0.556	0.304	0.527	-0.097	0.889
赖氨酸	0.659	-0.401	0.312	0.114	0.254	0.227	-0.082	-0.1	0.838
组氨酸	0.714	-0.472	0.226	-0.176	0.333	0.047	-0.13	-0.103	0.955
苹果酸	0.15	0.62	-0.645	0.066	0.11	0.04	0.229	-0.006	0.894
精氨酸	0.537	-0.52	0.156	-0.166	0.268	0.294	-0.123	-0.067	0.788

因子载荷图通过将多因子降维成双主成分或者三主成分，通过象限图的方式呈现主成分的空间分布。

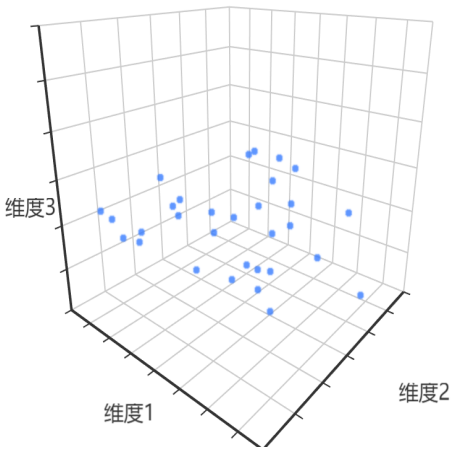

图 14: 因子载荷象限分析

下表为成分矩阵表，意在说明各个成分的所包含的因子得分系数（主成分载荷），用于计算出成分得分，得出因子公式，其计算公式为：线性组合系数 *（方差解释率/累积方差解释率），最后将其归一化即为因子权重得分。线性组合系数，公式为：因子载荷系数除以对应特征根，即成分矩阵的系数。我们根据成分矩阵表，建模最终的公式：模型的公式：

$$F1 = 0.056\times\text{胱氨酸}+0.119\times\ +0.086\times\ +0.088\times\ \ +0.029\times\ \ +0.103\times\ +0.111\times\ +0.062\times\ +0.019\times\ \ +0.11 \tag{14}$$

$$F2 = -0.035\times\ -0.058\times\ -0.008\times\ +0.045\times\ \ +0.017\times\ \ -0.056\times\ +0.031\times\ +0.085\times\ +0.009\times\ \ -0.058\times \tag{15}$$

$$F3 = -0.036\times\ +0.015\times\ -0.119\times\ -0.04\times\ \ +0.078\times\ \ -0.037\times\ +0.012\times\ -0.135\times\ -0.006\times\ \ +0.014\times \tag{16}$$

$$F4 = 0.154\times\ -0.06\times\ +0.022\times\ +0.067\times\ \ +0.333\times\ \ +0.141\times\ -0.037\times\ +0.079\times\ +0.054\times\ \ -0.091\times\ - \tag{17}$$

$$F5 = -0.082\times\ +0.01\times\ -0.086\times\ -0.265\times\ \ -0.013\times\ \ -0.039\times\ +0.154\times\ +0.142\times\ -0.109\times\ \ +0.02\times\ - \tag{18}$$

$$F6 = -0.089\times\ -0.097\times\ +0.075\times\ +0.043\times\ \ -0.085\times\ \ -0.054\times\ +0.012\times\ -0.053\times\ +0.412\times\ \ -0.121\times \tag{19}$$

$$F7 = 0.18\times\ -0.045\times\ -0.111\times\ +0.038\times\ \ -0.12\times\ \ -0.032\times\ -0.108\times\ -0.123\times\ +0.082\times\ \ -0.03\times\ -0. \tag{20}$$

$$F8 = 0.319\times\ +0.166\times\ -0.189\times\ -0.08\times\ \ -0.045\times\ \ -0.015\times\ -0.062\times\ -0.139\times\ +0.396\times\ \ +0.098\times\ - \tag{21}$$

联合上式可以得到：

$$F = (0.22/0.834)\times F1+(0.187/0.834)\times F2+(0.125/0.834)\times F3+(0.082/0.834)\times F4+(0.071/0.834)\times F5+(0.056/0.83 \tag{22}$$

表 12: 成分矩阵表

名称	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6	成分 7	成分 8
胱氨酸	0.056	-0.035	-0.036	0.154	-0.082	-0.089	0.18	0.319
缬氨酸	0.119	-0.058	0.015	-0.06	0.01	-0.097	-0.045	0.166
丙氨酸	0.086	-0.008	-0.119	0.022	-0.086	0.075	-0.111	-0.189
天门冬氨酸	0.088	0.045	-0.04	0.067	-0.265	0.043	0.038	-0.08
氨基酸总量	0.029	0.017	0.078	0.333	-0.013	-0.085	-0.12	-0.045
丝氨酸	0.103	-0.056	-0.037	0.141	-0.039	-0.054	-0.032	-0.015
蛋氨酸	0.111	0.031	0.012	-0.037	0.154	0.012	-0.108	-0.062
苏氨酸	0.062	0.085	-0.135	0.079	0.142	-0.053	-0.123	-0.139
白藜芦醇	0.019	0.009	-0.006	0.054	-0.109	0.412	0.082	0.396
亮氨酸	0.116	-0.058	0.014	-0.091	0.02	-0.121	-0.03	0.098
谷氨酸	0.097	0.019	-0.168	-0.004	0.003	0.069	-0.032	-0.061
多酚氧化酶活力	-0.046	0.082	-0.052	-0.057	0.216	-0.178	-0.068	0.197
葡萄总黄酮	0.055	0.102	0.137	-0.074	-0.125	-0.075	0.003	0.006
苯丙氨酸	0.038	0.018	0.033	-0.024	-0.108	0.029	0.38	-0.493
单宁	0.045	0.129	0.049	0.012	-0.062	-0.018	-0.098	-0.114
蛋白质	0.005	0.076	0.178	-0.08	0.096	0.201	-0.011	-0.026
DPPH 自由基	0.043	0.115	0.139	-0.082	-0.061	0.068	-0.069	0.16
褐变度	-0.003	0.13	-0.076	-0.07	0.106	0.045	-0.012	-0.001
总酚	0.062	0.123	0.114	-0.029	-0.077	-0.097	-0.061	0.035
甘氨酸	0.077	0.059	-0.135	0.012	-0.077	0.07	0.061	0.272
脯氨酸	0.001	0.021	0.097	0.33	-0.023	-0.097	-0.108	-0.04
VC 含量	-0.015	-0.012	-0.022	-0.044	-0.161	0.329	-0.256	-0.102
异亮氨酸	0.101	-0.048	-0.013	-0.11	0.029	-0.146	0.067	0.096
酪氨酸	0.054	0.007	0.058	-0.14	-0.144	-0.157	0.24	0.012
花色苷	0.025	0.16	-0.011	-0.01	0.053	-0.018	-0.063	-0.003
酒石酸	0.039	0.012	0.144	0.136	0.174	0.102	0.216	0.127
柠檬酸	0.022	0.053	-0.026	0.107	0.253	0.177	0.309	-0.08
赖氨酸	0.097	-0.069	0.08	0.045	0.115	0.132	-0.048	-0.082
组氨酸	0.105	-0.082	0.058	-0.07	0.151	0.027	-0.076	-0.085
苹果酸	0.022	0.107	-0.167	0.026	0.05	0.023	0.134	-0.005
精氨酸	0.079	-0.09	0.04	-0.065	0.122	0.171	-0.072	-0.056

表 13: 第二问预测结果

企业代号	是否违约	信誉评级
E124	是	D
E125	否	C
E126	是	D
E127	是	D
E128	是	A
E129	否	C
E130	是	A
E131	否	C
E132	是	D
E133	是	A
E134	否	A
E135	否	A
E136	否	B
E137	否	A
E138	是	A
E139	是	B
E140	否	A
E141	否	A
E142	否	A
E143	否	A
E144	否	A
E145	否	A
E146	否	B
E147	否	A
E148	否	A
E149	否	A
E150	否	A
E151	否	A
E152	否	A
E153	是	D
E154	否	B
E155	是	C
E156	是	D
E157	否	C
E158	否	B
E159	是	C
E160	否	A
E161	否	B

E162	否	A
E163	否	A
E164	是	D
E165	否	A
E166	否	A
E167	否	A
E168	否	C
E169	否	B
E170	否	A
E171	否	A
E172	否	B
E173	否	C
E174	否	B
E175	否	A
E176	否	A
E177	否	A
E178	否	B
E179	否	A
E180	否	A
E181	否	A
E182	否	B
E183	否	A
E184	否	A
E185	否	B
E186	否	A
E187	是	D
E188	否	A
E189	否	B
E190	否	B
E191	否	B
E192	否	A
E193	否	C
E194	否	A
E195	否	A
E196	否	A
E197	否	A
E198	否	A
E199	否	A
E200	是	D
E201	否	B
E202	是	C

E203	否	A
E204	否	A
E205	是	C
E206	否	A
E207	是	C
E208	是	D
E209	否	A
E210	否	A
E211	是	D
E212	否	A
E213	否	A
E214	否	B
E215	否	A
E216	否	A
E217	是	D
E218	否	C
E219	否	C
E220	否	A
E221	否	B
E222	否	A
E223	否	C
E224	否	B
E225	否	A
E226	否	B
E227	否	A
E228	否	A
E229	否	A
E230	否	B
E231	否	A
E232	否	C
E233	否	B
E234	否	A
E235	是	D
E236	是	D
E237	是	D
E238	否	C
E239	是	D
E240	是	D
E241	是	D
E242	是	D
E243	否	A

E244	是	D
E245	否	B
E246	否	A
E247	否	B
E248	否	A
E249	否	C
E250	否	A
E251	否	B
E252	否	A
E253	否	A
E254	否	C
E255	是	D
E256	否	C
E257	否	C
E258	否	A
E259	否	C
E260	否	A
E261	否	A
E262	是	B
E263	否	B
E264	是	D
E265	否	B
E266	否	B
E267	否	C
E268	否	B
E269	否	A
E270	是	D
E271	否	A
E272	是	D
E273	是	C
E274	否	C
E275	否	C
E276	否	A
E277	否	C
E278	否	B
E279	否	B
E280	是	C
E281	否	A
E282	否	C
E283	否	B
E284	否	A

E285	是	D
E286	否	B
E287	否	A
E288	是	D
E289	否	A
E290	否	A
E291	否	A
E292	否	B
E293	否	C
E294	是	D
E295	否	A
E296	否	B
E297	否	C
E298	否	C
E299	否	B
E300	否	A
E301	否	A
E302	否	A
E303	否	A
E304	否	B
E305	否	C
E306	是	D
E307	否	B
E308	否	B
E309	是	D
E310	否	B
E311	否	A
E312	否	B
E313	否	A
E314	否	A
E315	否	A
E316	是	D
E317	否	B
E318	否	C
E319	是	C
E320	否	C
E321	否	B
E322	否	B
E323	否	C
E324	否	A
E325	否	B

E326	否	A
E327	是	C
E328	否	C
E329	否	C
E330	否	A
E331	否	C
E332	否	C
E333	否	A
E334	否	C
E335	否	C
E336	否	C
E337	是	C
E338	否	C
E339	否	C
E340	否	C
E341	否	C
E342	否	A
E343	否	C
E344	否	C
E345	否	B
E346	是	D
E347	否	B
E348	否	C
E349	否	C
E350	否	B
E351	否	B
E352	否	B
E353	否	C
E354	否	B
E355	否	C
E356	否	C
E357	否	C
E358	否	C
E359	否	C
E360	否	C
E361	否	C
E362	否	C
E363	否	B
E364	否	B
E365	否	A
E366	否	C

E367	否	C
E368	否	C
E369	否	C
E370	否	C
E371	否	C
E372	否	C
E373	是	D
E374	否	C
E375	否	C
E376	否	C
E377	否	B
E378	否	B
E379	否	C
E380	否	C
E381	否	A
E382	否	C
E383	否	C
E384	是	D
E385	否	C
E386	是	D
E387	否	C
E388	否	C
E389	否	C
E390	否	C
E391	否	C
E392	否	B
E393	否	C
E394	否	A
E395	否	B
E396	否	B
E397	否	C
E398	否	C
E399	否	C
E400	否	C
E401	否	B
E402	否	C
E403	否	C
E404	是	D
E405	否	C
E406	否	C
E407	否	C

E408	否	B
E409	否	C
E410	否	C
E411	否	B
E412	否	C
E413	否	B
E414	否	C
E415	否	C
E416	否	C
E417	否	B
E418	否	C
E419	否	B
E420	否	B
E421	否	C
E422	否	C
E423	否	C
E424	否	C

下表为主成分分析的根据载荷系数等信息所做的主成分权重分析，其计算公式为：方差解释率/旋转后累积方差解释率。

表 14: 因子权重分析

名称	方差解释率	累计方差解释率	权重
主成分 1	0.22	0.22	26.36%
主成分 2	0.187	0.406	22.39%
主成分 3	0.125	0.531	14.98%
主成分 4	0.082	0.613	9.81%
主成分 5	0.071	0.684	8.51%
主成分 6	0.056	0.74	6.67%
主成分 7	0.055	0.794	6.59%
主成分 8	0.039	0.834	4.70%

5.4.2 回归分析

葡萄酒质量从外观、香气、口感等多方面进行打分，将这几种打分标准与上述的主成分进行数据关联，求解最优的模型关系。以外观、香气、口感等多项评价指标作为因变量，酿酒葡萄理化指标主成分作为自变量进行回归分析，建立回归方程：

利用模型进行重新打分得到如下结果：

最后在经过误差分析后发现，通过回归分析得出的评分结果与所给的结果误差基本都小于 0.05，表明该方法的可行性，也论证了酿酒葡萄和葡萄酒的理化指标对葡萄酒质量进行评价打分是可行的。

参考文献

[1] Hossein Arsham and Miodrag Lovric. Bartlett's test. *International encyclopedia of statistical science*, 1:87–88, 2011.

[2] David Arthur and Sergei Vassilvitskii. How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 144–153, 2006.

[3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[4] Andreas Daffertshofer, Claudine JC Lamoth, Onno G Meijer, and Peter J Beek. Pca in studying coordination and variability: a tutorial. *Clinical biomechanics*, 19(4):415–428, 2004.

[5] Joost CF De Winter. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18(1):10, 2013.

[6] Charles D Dziuban and Edwin C Shirkey. When is a correlation matrix appropriate for factor analysis? some decision rules. *Psychological bulletin*, 81(6):358, 1974.

[7] Charlotte Hommerberg. *Persuasiveness in the discourse of wine: The rhetoric of Robert Parker*. PhD thesis, Linnaeus University Press, 2011.

[8] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. ipca: An interactive system for pca-based visual analytics. In *Computer Graphics Forum*, volume 28, pages 767–774. Wiley Online Library, 2009.

[9] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

[10] Ian T Young. Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry & Cytochemistry*, 25(7):935–941, 1977.

6 附录

6.1 代码

```
import time
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans
from sklearn.metrics.pairwise import pairwise_distances_argmin
from sklearn.datasets.samples_generator import make_blobs

# path = '../data/2/red_average.csv'
# file = pd.read_csv(path,encoding='utf-8')
path = '../data/2/white_sum.csv'
file = pd.read_csv(path, encoding='utf-8')
file.head()
new_file = file.iloc[:, 1:]
print(new_file.head())
x = new_file.values


# // An highlighted block
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from numpy.linalg import eig
#matplotlib inline

iris = load_iris()
# X = iris.data
X=x
print(x.shape)
X = X - X.mean(axis = 0)

#计算协方差矩阵
X_cov = np.cov(X.T, ddof = 0)

#计算协方差矩阵的特征值和特征向量
eigenvalues,eigenvectors = eig(X_cov)

```

```
    tot = sum(eigenvalues)
    var_exp = [(i/tot) for i in sorted(eigenvalues, reverse = True)]
    cum_var_exp = np.cumsum(var_exp)

    plt.bar(range(1,12), var_exp, alpha = 0.5, align = 'center',
        label = 'individual var')
    plt.step(range(1,12), cum_var_exp, where = 'mid', color='y',
        label = 'cumulative var')
    plt.ylabel('variance rtion')
    plt.xlabel('principal components')
    plt.legend(loc = 'best')
    plt.show()
    # 各特征值的贡献率如图所示，可以看出，前两个特征值的方差贡献率超过95%，所以k取3有其合理性。

    # // 用python实现主成分分析（PCA）
    import numpy as np
    from numpy.linalg import eig
    from sklearn.datasets import load_iris
    def pca(X,k):
        X = X - X.mean(axis = 0) #向量X去中心化
        X_cov = np.cov(X.T, ddof = 0) #计算向量X的协方差矩阵，自由度可以选择0或1
        eigenvalues,eigenvectors = eig(X_cov) #计算协方差矩阵的特征值和特征向量
        klarge_index = eigenvalues.argsort()[-k:][::-1] #选取最大的K个特征值及其特征向量
        k_eigenvectors = eigenvectors[klarge_index] #用X与特征向量相乘
        return np.dot(X, k_eigenvectors.T)
    iris = load_iris()
    # X = iris.data
    k = 3
    X_pca = pca(X, k)
    print(X_pca)
```

```

    import time
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt

```

```
7       from sklearn.cluster import KMeans
8       from sklearn.metrics.pairwise import pairwise_distances_argmin
9       from sklearn.datasets.samples_generator import make_blobs
10
11      # path = '../data/2/red_average.csv'
12      # file = pd.read_csv(path,encoding='utf-8')
13      path = '../data/2/white_sum.csv'
14      file = pd.read_csv(path,encoding='gbk')
15      file.head()
16
17      new_file = file.iloc[:,1:]
18      new_file.head()
19      x= new_file.values
20
21      batch_size = 45
22      centers = x
23      n_clusters = 6
24      # X, labels_true = make_blobs(n_samples=3000, centers=centers,
            cluster_std=0.7)
25      X=x
26      # plot result
27      fig = plt.figure(figsize=(8, 3))
28      fig.subplots_adjust(left=0.02, right=0.98, bottom=0.05, top=0.9)
29      colors = ['#4EACC5', '#FF9C34', '#4E9A06','r','y','blue','blace'
            ]
30
31
32      # compute clustering with K-Means
33      k_means = KMeans(init='k-means++', n_clusters=6, n_init=10)
34      t0 = time.time()
35      k_means.fit(X)
36      t_batch = time.time() - t0
37
38      k_means_cluster_centers = np.sort(k_means.cluster_centers_, axis
            =0)
39      k_means_labels = pairwise_distances_argmin(X,
            k_means_cluster_centers)
40      print('k_means_labels', k_means_labels)
41
42      from sklearn.datasets import load_digits
43      from sklearn.decomposition import IncrementalPCA
44      from scipy import sparse
45
```

```
    transformer = IncrementalPCA(n_components=2, batch_size=200)
    # either partially fit on smaller batches of data
    transformer.partial_fit(x)
    # or let the fit function itself divide the data into batches
    X_sparse = sparse.csr_matrix(x)
    X = transformer.fit_transform(X_sparse)
    # hc_pred.shape
    print(X.shape)
    hc_pred = k_means_labels

    # K-means
    ax = fig.add_subplot(1, 1, 1)
    for k, col in zip(range(n_clusters), colors):
        my_members = k_means_labels == k  # my_members是布尔型的数组（用于筛选同类的点，用不同颜色表示）
        cluster_center = k_means_cluster_centers[k]
        ax.plot(X[my_members, 0],
                X[my_members, 1],
                'w',
                markerfacecolor=col,
                marker='o')  # 将同一类的点表示出来
    ax.set_title('KMeans')
    ax.set_xticks(())
    ax.set_yticks(())

    # plt.text(-3.5, 1.8,
    #          'train time: %.2fs\ninertia: %f' % (t_batch, k_means.inertia_))
    plt.legend()
    plt.savefig('./img/k_means.png')
    plt.show()
```

```

    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import scipy.cluster.hierarchy as sch


    path = '../data/2/red_average.csv'
    # path = '../data/2/white_sum.csv'
    file = pd.read_csv(path,encoding='utf-8')
```

```
11      file.head()
12
13      new_file = file.iloc[:,1:]
14      new_file.head()
15      x= new_file.values
16
17      dendrogram = sch.dendrogram(sch.linkage(x, method = 'ward'))
18      plt.title('Dendrogram')
19      plt.xlabel('Customers')
20      plt.ylabel('Euclidean Distance')
21      plt.savefig('./img/level_analysis.png')
22      plt.show()
```