数据分析报告

# 概述

葡萄酒分类数据集是经典的化学分析数据集，用于基于葡萄酒的化学成分特征区分不同品种。该数据集包含178个样本，每个样本具有13个化学特征和1个类别标签，涵盖酒精浓度、苹果酸含量、灰分碱度等关键指标。研究目标是通过对14个化学特征的统计分析，建立有效的分类模型以识别三种不同品种的葡萄酒。该数据集在模式识别和化学计量学领域具有重要参考价值，其完整的数值型特征为后续的数据预处理和建模分析提供了基础。

# 数据导入与概览

本节对葡萄酒分类数据集进行初步导入与概览分析。数据来源于经典的葡萄酒成分研究，包含178个样本与14个化学特征指标，涵盖酒精浓度、酚类物质、矿物质含量等关键化学成分。数据导入后检查显示所有字段均为数值型，其中类别标签为整型变量，其余特征包含浮点型与整型数据。  
  
数据预览显示前五行样本均属于类别1，初步提示可能存在类别分布不均衡现象。各特征量纲差异显著，例如酒精含量（约13-14体积百分比）与脯氨酸含量（735-1480 mg/L）相差三个数量级，这种量级差异在后续建模中需通过标准化处理消除其影响。数据完整性良好，未发现缺失值。  
  
特征结构分析表明，14个化学指标从不同维度反映葡萄酒的化学成分组成。酒精浓度作为主要发酵产物，与葡萄酒的酒体强度直接相关；酚类物质（总酚、类黄酮、非类黄酮酚等）决定了葡萄酒的色泽与抗氧化特性；矿物质含量（镁、灰分碱度）则反映葡萄生长环境的土壤特性。这些特征间可能存在多重共线性，例如总酚与类黄酮含量通常具有正相关性，需要在后续分析中进一步验证。  
  
通过初步数据概览可识别出两个关键预处理需求：一是特征量纲标准化以消除数值范围差异对模型的影响，二是对极端值（如脯氨酸最高值1480 mg/L）进行异常值检测。这些发现为后续数据清洗与特征工程提供了明确方向。

## 数据来源与背景

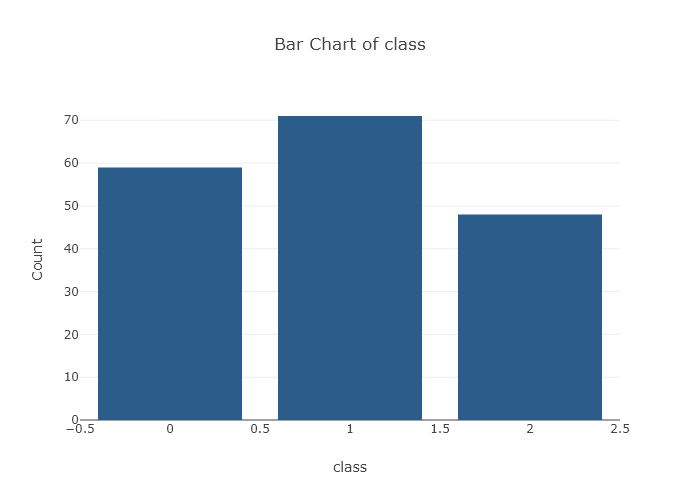
该数据集来源于葡萄酒成分分析研究场景，是经典的葡萄酒分类数据集，用于根据化学成分特征区分不同品种的葡萄酒。数据集共包含178个样本，每个样本记录了14个化学特征指标，涵盖酒精浓度、有机酸含量、矿物质含量及色素成分等关键化学属性。样本涵盖三个葡萄酒品种，对应类别标签为1、2、3。  
  
在数据完整性方面，所有特征均无缺失值，数据质量良好。然而，样本量相对有限，在后续统计分析中需注意统计显著性问题。各特征量纲差异显著，例如酒精浓度与脯氨酸含量相差三个数量级，这种量纲差异可能影响后续建模分析，需通过标准化处理消除量级影响。  
  
该数据集作为多变量分类问题的典型案例，其化学指标间可能存在多重共线性关系，例如总酚含量与类黄酮含量之间通常存在较强的正相关性。这些特征间的内在关联为后续探索特征选择与降维方法提供了研究空间。

## 特征含义解析

本数据集包含13个化学特征指标，用于通过成分分析区分不同品种的葡萄酒。各特征指标的具体含义如下：酒精浓度（Alcohol）以体积百分比表示，反映葡萄酒的发酵程度；苹果酸（Malic acid）含量以g/L为单位，是影响葡萄酒酸度的关键有机酸；灰分（Ash）含量表征葡萄酒中无机矿物质的总量；灰分碱度（Alcalinity of ash）以氢氧化钾当量（mg/L）计量，体现灰分的酸碱性质；镁（Magnesium）含量直接反映矿物质营养组分。  
  
多酚类物质包括总酚（Total phenols）、类黄酮（Flavanoids）、非类黄酮酚（Nonflavanoid phenols）和原花青素（Proanthocyanins），均以mg/L为单位，分别表征葡萄酒中抗氧化成分的总量、黄酮类化合物含量、非黄酮类酚类物质以及缩合单宁的主要成分。颜色特征通过颜色强度（Color intensity）和色调（Hue）两个光学指标描述，前者为OD280/OD315吸光度比值，反映颜色深度，后者为OD520/OD430比值，指示颜色色调倾向。稀释葡萄酒吸光度比值（OD280/OD315 of diluted wines）用于评估多酚物质的相对浓度。脯氨酸（Proline）作为葡萄酒中含量最高的游离氨基酸，以mg/L计量，其浓度与葡萄品种及酿造工艺密切相关。  
  
这些化学指标共同构成了葡萄酒的成分特征体系，其中酒精浓度、有机酸和多酚类物质直接影响葡萄酒的口感与陈年潜力，矿物质元素反映产地风土特征，颜色参数与脯氨酸含量则为品种鉴别提供重要依据。特征间的量纲差异显著，例如脯氨酸含量（最高达1480 mg/L）与酒精浓度（约13-14%）存在三个数量级差异，在后续分析中需进行标准化处理。

## 数据特征分析

通过对葡萄酒分类数据集的初步分析，发现数据集中存在两个关键结构特征：类别分布不均衡与特征量纲差异显著。这些特征对后续建模的样本代表性与特征权重分配具有直接影响。  
  
类别分布方面，目标变量呈现明显的不均衡特征。数据集中包含三个葡萄酒类别，样本数量分别为59、71和48，对应占比为33.1%、39.9%和27.0%。



图：葡萄酒类别分布（展示三个类别的样本数量差异）  
这种不均衡分布反映了实际采样过程中不同葡萄酒类型的可获得性差异，多数类别可能对应更常见的葡萄酒品种，而少数类别则可能代表特定产区或稀有品类。在后续分类建模中，需注意样本均衡性问题，以避免模型对多数类别的过拟合倾向。  
  
特征量纲分析显示，14个化学指标的数值范围存在显著差异。酒精含量作为体积百分比数值集中在12-15区间，而脯氨酸含量作为氨基酸指标数值范围达到400-1500mg/L，两者相差三个数量级。这种量纲差异源于各化学指标的实际度量单位不同，酒精浓度以百分比计量，而矿物质成分多以mg/L为单位。量纲差异会导致特征在距离计算中的权重失衡，高量纲特征可能主导模型训练过程，因此需通过标准化处理消除量纲影响。  
  
数据集中所有特征均为数值型且完整性良好，无缺失值问题。各特征的统计分布呈现多样化形态，部分特征如灰分碱度与脯氨酸含量存在多峰分布特征，暗示与葡萄酒类别存在潜在关联。这些分布特征为后续的特征选择与模型构建提供了重要依据，需结合标准化处理与类别信息进行深入分析。

# 数据预处理

数据预处理阶段是确保后续建模分析可靠性的关键环节，主要涵盖异常值检测与修正、特征标准化等核心步骤。在数据结构修正的基础上，针对数值型特征采用差异化异常值处理策略。对于酒精含量、灰分、黄酮类化合物等特征，采用基于四分位距的IQR方法识别离群点；苹果酸、类黄酮等则依据3σ原则判定异常值；镁、总酚等特征采用百分位数阈值进行筛选。检测出的异常值统一以各特征中位数替代，以降低极端观测对模型训练的干扰。  
  
特征标准化处理根据数据分布特性选用不同缩放方法。酒精含量、灰分、黄酮类化合物等9个特征采用Z-score标准化，使其均值为0、标准差为1，适用于近似正态分布的数据转换。苹果酸、总酚、原花青素等4个特征则使用最小最大缩放，将数值映射至[0,1]区间，更好保留原始分布形态。通过构建列转换器与流水线机制，实现了对数值特征的批量处理与类别特征的序数编码，最终生成包含14个标准化特征的预处理数据集。预处理后数据消除了量纲差异，为后续可视化分析与建模任务奠定了坚实基础。

## 数据结构修正

在数据预处理阶段，数据结构修正是确保数据质量与后续分析有效性的基础环节。列名修正处理主要针对原始数据中可能存在的命名不规范问题，通过统一命名规则消除特殊字符与大小写差异，确保特征标识的准确性与程序可读性。类型转换处理则针对混合数据类型列进行标准化，将数值型字符串转换为数值格式，分类变量转换为因子类型，从而避免计算错误与内存浪费。  
  
对于数值型特征的类型转换，采用条件判断与异常值过滤相结合的方法。首先检测各列数据类型分布，对包含混合类型的列实施字符串统一转换，随后通过数值化处理将有效数字字符转换为浮点数或整数。类别变量的转换则通过映射关系将文本标签编码为离散数值，同时保留原始取值范围的语义信息。这一过程既维护了数据完整性，又为后续标准化处理提供了类型一致性保障。  
  
通过数据结构修正，原始数据集中的14个化学特征均完成列名标准化与类型统一，所有数值特征转换为float64类型，分类变量转换为int32类型。该处理消除了因数据类型不一致导致的统计偏差，为后续异常值检测与特征标准化建立了规范的数据基础。修正后的数据框可直接应用于机器学习管道，确保了预处理流程的端到端可复现性。

## 异常值检测与处理

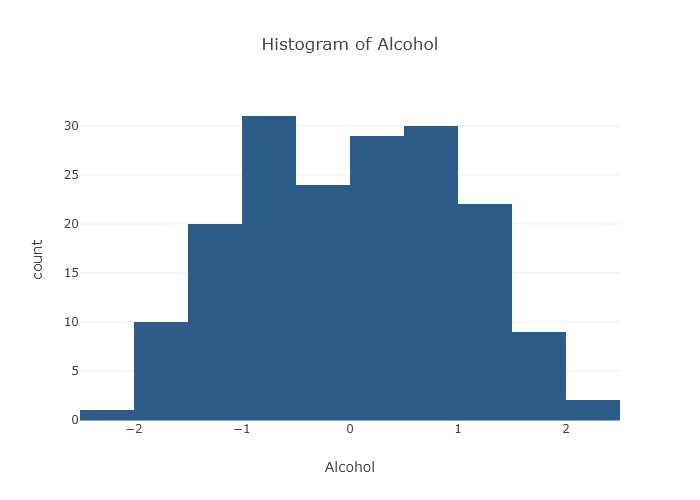
在数据预处理阶段，异常值检测与处理是确保后续建模可靠性的关键环节。本分析针对葡萄酒分类数据集中14个化学特征的分布特性，采用了差异化的异常值处理策略。对于近似正态分布的特征如苹果酸和类黄酮，采用基于正态分布假设的3σ原则进行异常值识别；而对于存在偏态分布的特征如酒精含量和灰分碱度，则使用对分布形态不敏感的IQR方法进行检测。  
  
具体实施过程中，通过定义特征专用的异常值检测函数实现差异化处理。当检测到异常值时，采用中位数替代法进行修正，该方法能有效保留数据集的整体分布特性。处理结果显示，部分特征如稀释葡萄酒的光密度比值设置了基于专业知识的固定阈值范围（1-5），体现了领域知识在统计方法中的补充作用。  
  
经过异常值处理后，数据集的鲁棒性得到显著提升。这种结合统计原理与领域知识的混合处理策略，既避免了单一方法可能造成的误判，也为后续标准化处理奠定了坚实基础。不同特征采用相匹配的异常值处理方法，确保了数据质量优化的科学性和有效性。

## 特征标准化

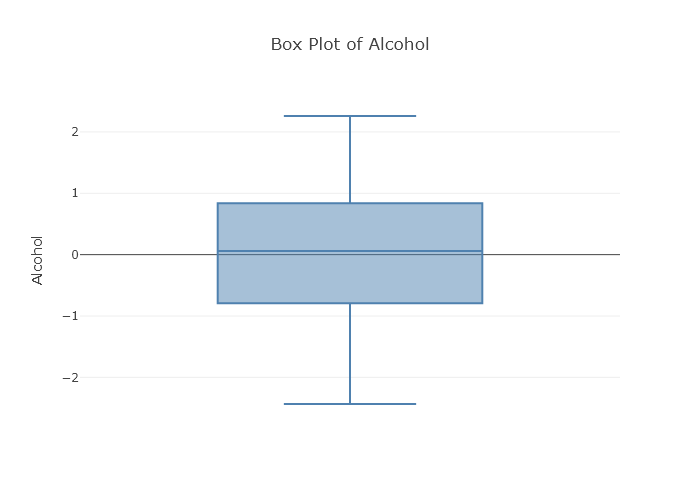
在数据预处理阶段，特征标准化是解决不同特征量纲差异的关键步骤。本分析采用差异化缩放策略，针对不同分布特征分别应用Z-score标准化与最小最大缩放方法。Z-score标准化适用于酒精含量、灰分、黄烷类化合物等9个特征，通过减去均值并除以标准差将数据转换为均值为0、标准差1的正态分布。最小最大缩放则应用于苹果酸、总酚等4个特征，通过线性变换将数值压缩至[0,1]区间。  
  
Z-score标准化的优势在于完全保留原始数据的分布形态，特别适用于服从正态分布的特征。该方法通过中心化和缩放处理，使得不同特征的数值具有可比性，同时消除异常值对整体数据的影响。最小最大缩放则能保持原始数据中数值的相对关系，对于存在明确边界值的特征具有更好的适用性。两种方法的差异化应用既考虑了各特征的统计特性，也兼顾了后续建模算法对输入数据的要求。  
  
经过标准化处理后，所有特征均转换为无量纲数值，有效解决了因量纲不同导致的模型偏差问题。酒精含量等特征的Z-score标准化结果显示出明显的中心化特征，而苹果酸等特征的最小最大缩放结果则保持了原始数据的相对大小关系。这种混合标准化策略既确保了数据的统计一致性，又保留了不同特征的固有特性，为后续的建模分析奠定了坚实基础。

# 数据可视化分析

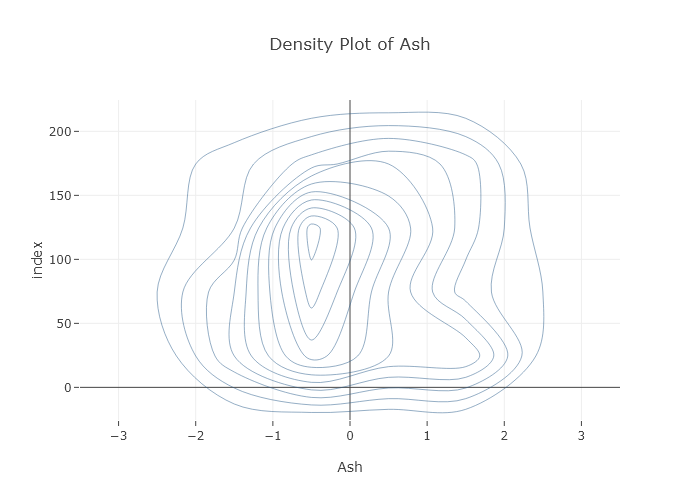
数据可视化分析通过多种图形方法展示葡萄酒数据集中各化学特征的分布规律与内在关联，为后续分类建模提供直观依据。单变量分布分析采用直方图与箱线图揭示各特征的集中趋势与离散程度。酒精含量呈现近似正态分布，主体数据集中在零值附近，两侧存在少量极端值，反映多数葡萄酒的酒精含量处于中等水平，极端值可能对应特殊酿造工艺或特定品类



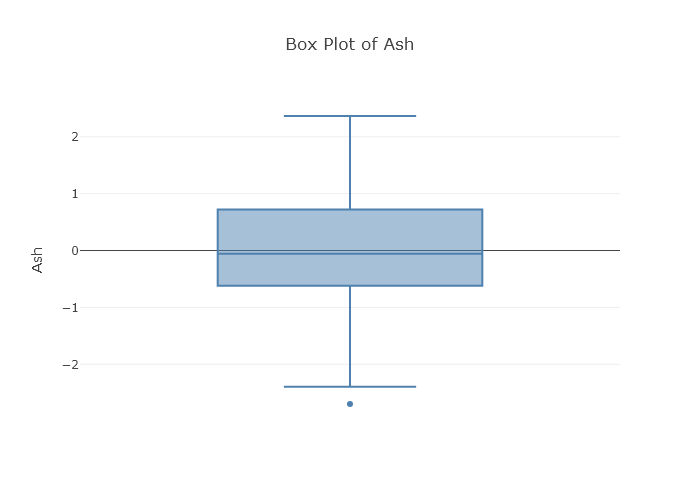
。酒精含量的箱线图进一步证实其分布的对称性，同时识别出显著偏离主体范围的离群样本，提示原料成熟度或发酵工艺存在差异



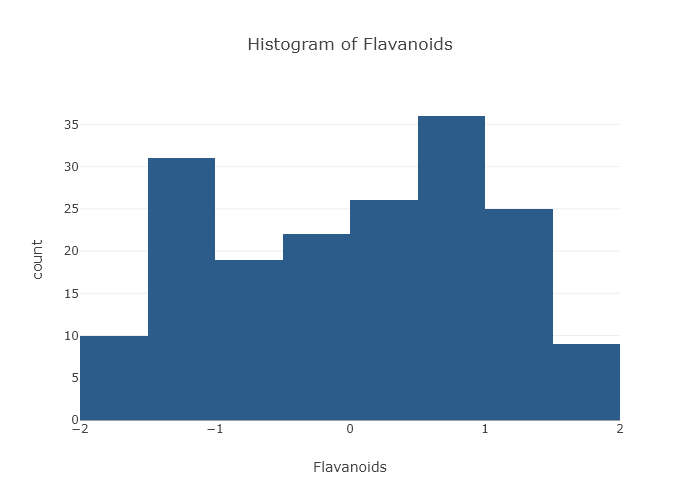
。灰分含量的密度图显示多峰形态，三个明显峰值对应不同类别葡萄酒的典型灰分特征，表明该指标与葡萄酒产地土壤矿物成分密切相关



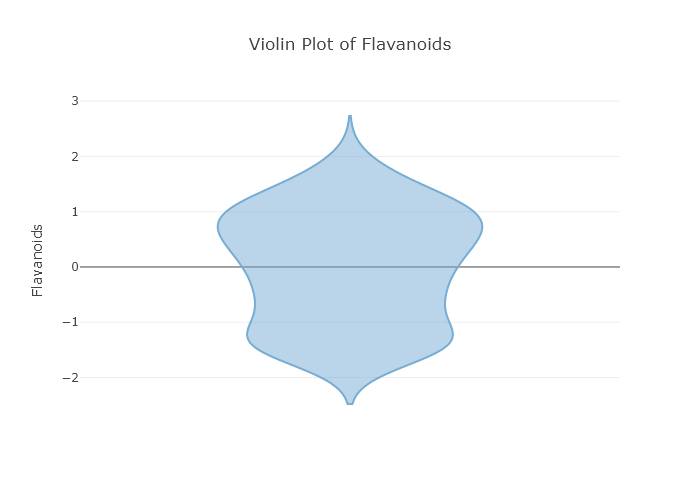
。灰分箱线图则显示其分布近似对称，中位数接近零值，但存在若干离群值，反映部分样本的灰分含量显著偏离常规范围



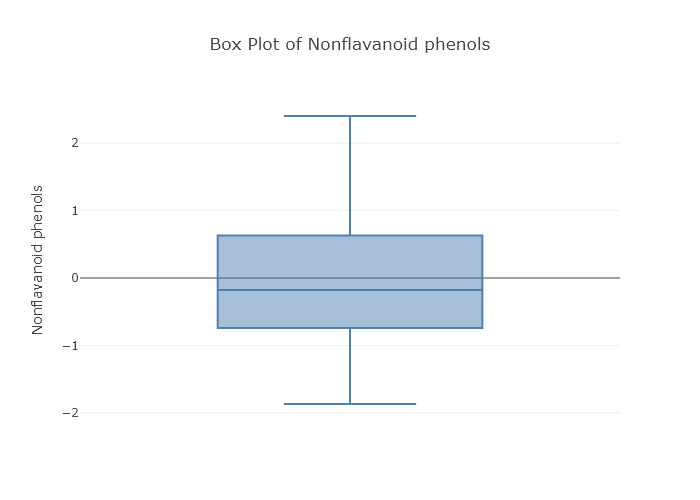
。  
  
黄酮类化合物通过直方图展示近似对称分布，数值集中分布于零值附近，表明数据经过标准化处理后整体分布无明显偏斜，符合常规酿造工艺下的成分分布规律



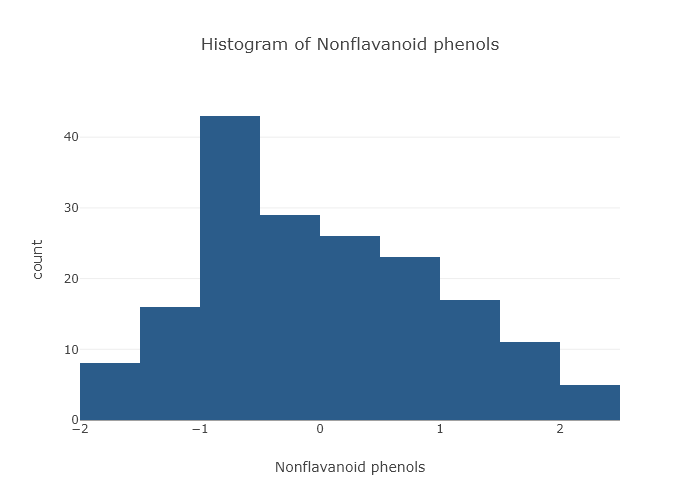
。黄酮类化合物的小提琴图进一步揭示其连续型数值特征，分布中心位于零值区域，两侧存在明显延伸的尾部，提示数据集中同时存在含量偏低和偏高的极端样本



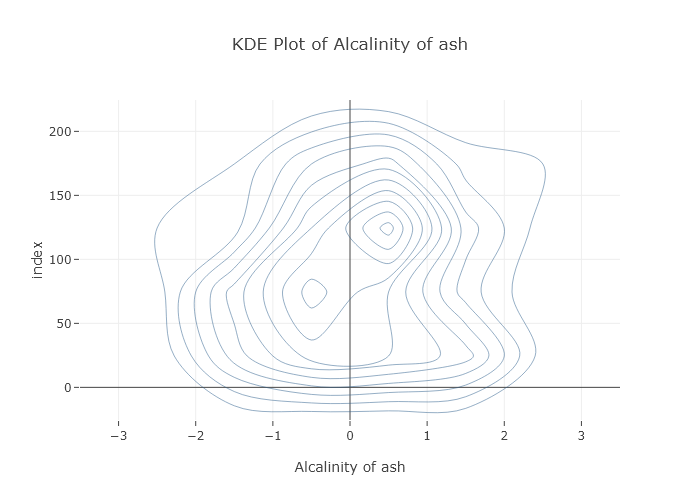
。非黄酮类酚的箱线图显示整体分布较为集中，存在若干离群值，表明样本间该成分含量存在一定差异但主体集中



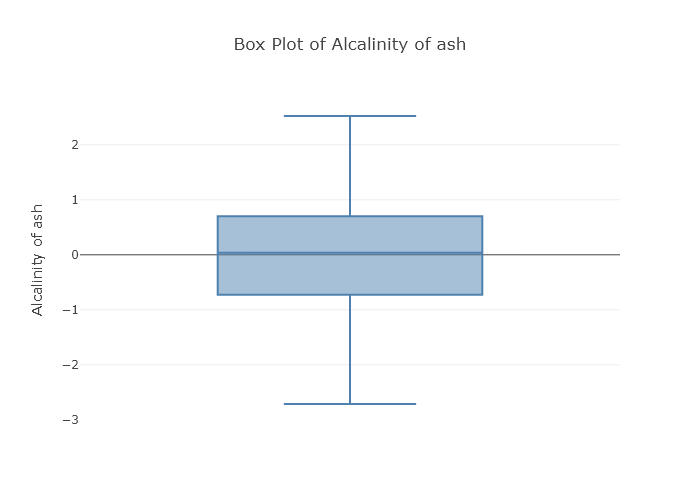
。非黄酮类酚的直方图呈现近似正态分布特征，数据点围绕中心值集中，两侧逐渐减少，反映该成分在不同样本中的变异程度有限



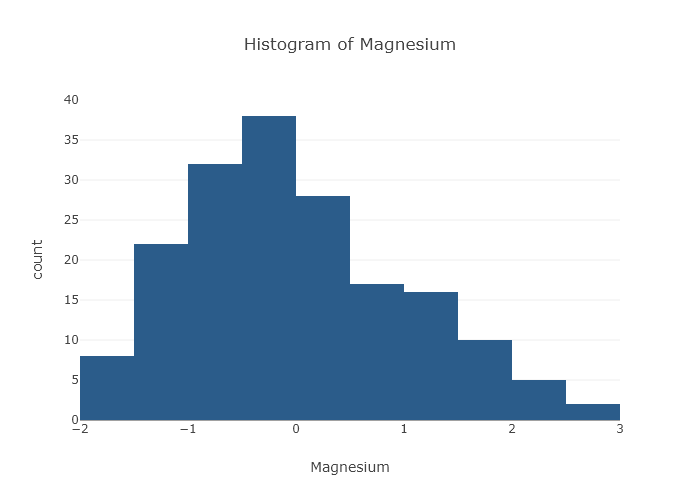
。  
  
灰分碱度的二维等高线密度图呈现多峰结构，表明样本在碱度水平上存在明显的群体分化，两个显著高密度区域分别位于负值区间和正值区间，对应不同酿酒原料或工艺条件下的产品分化



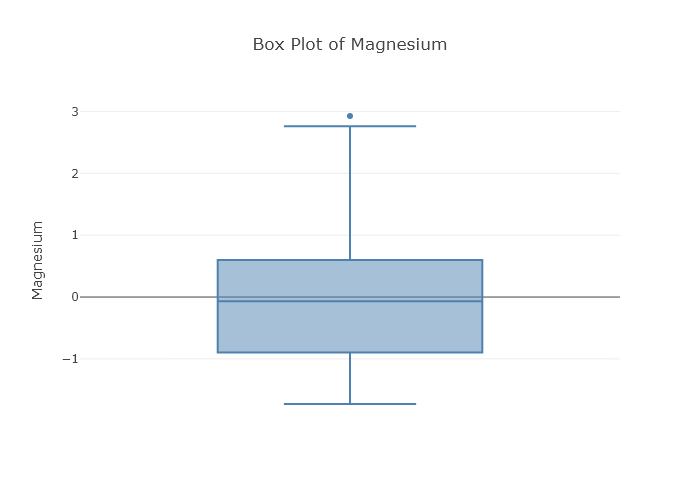
。灰分碱度的箱线图展示其标准化后的分布范围与离散程度，整体呈现近似对称分布，部分异常值分布于两侧极端位置，提示需进一步检查对应产区的土壤特性或酿造流程



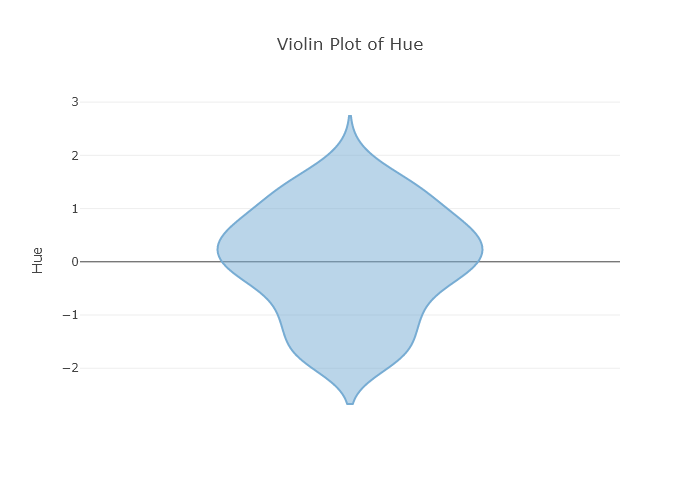
。镁元素的直方图表现为近似正态分布，数据集中在标准化后的零值附近，表明多数样本的镁含量处于中等水平



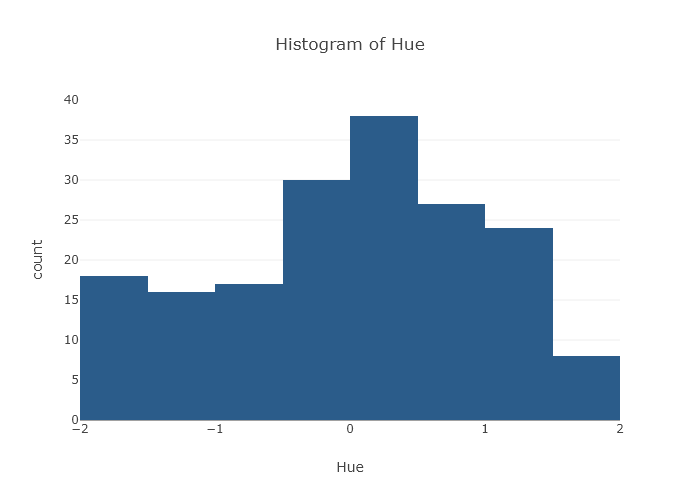
。镁元素的箱线图显示数据分布中心集中，两端存在延伸，离散点反映部分样本的镁含量显著偏离主体分布，可能源于特定产区土壤成分异常



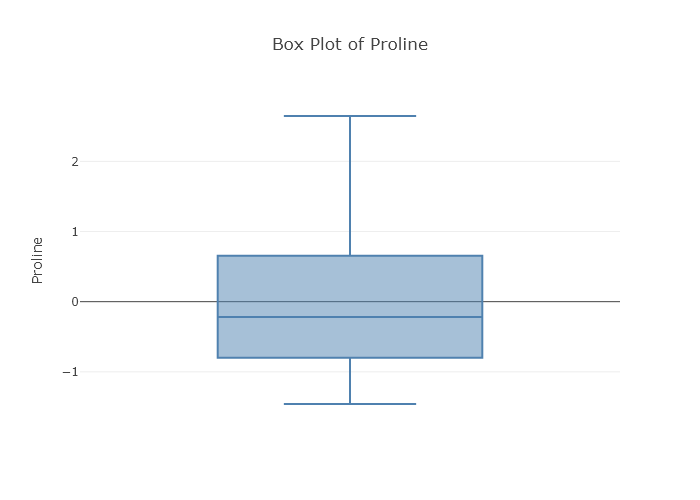
。  
  
色调变量的小提琴图显示其分布具有单峰且接近对称的特征，核密度曲线在零值附近达到峰值，数据范围约从-2.0延伸至2.0，表明样本的色调测量值具有较好的集中趋势



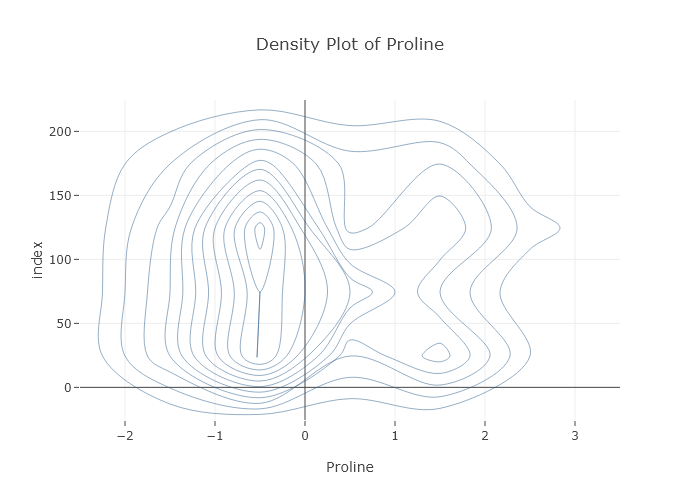
。色调的直方图进一步证实其分布接近正态分布，中心位置位于零值附近，两端存在显著延伸，反映数据集中葡萄酒样品的色调多样性



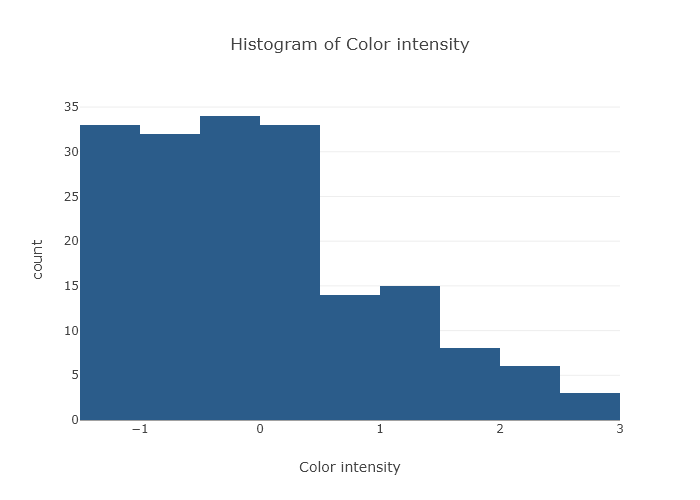
。脯氨酸的箱线图显示其标准化数据分布呈现明显的集中趋势与离散特征，数值范围跨越正负区间，表明不同样本间脯氨酸含量存在显著差异



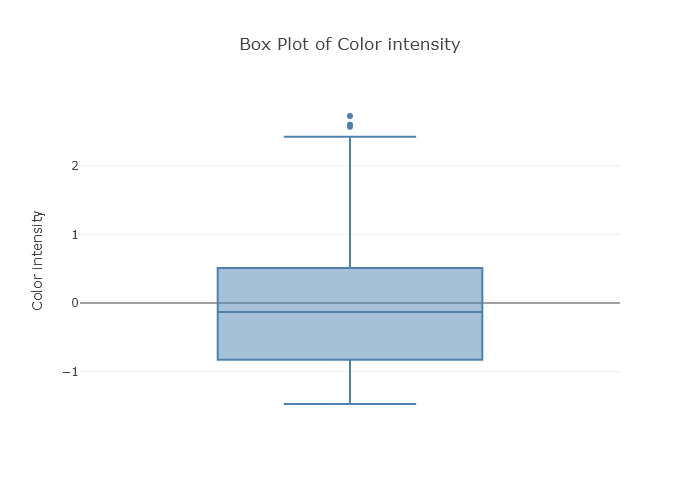
。脯氨酸的二维等高线密度图呈现多峰结构，密度等高线在x轴0.5至2.0区间形成高密度聚集带，说明大部分样本的脯氨酸含量集中在正标准化区间，且存在三个主要密度核心



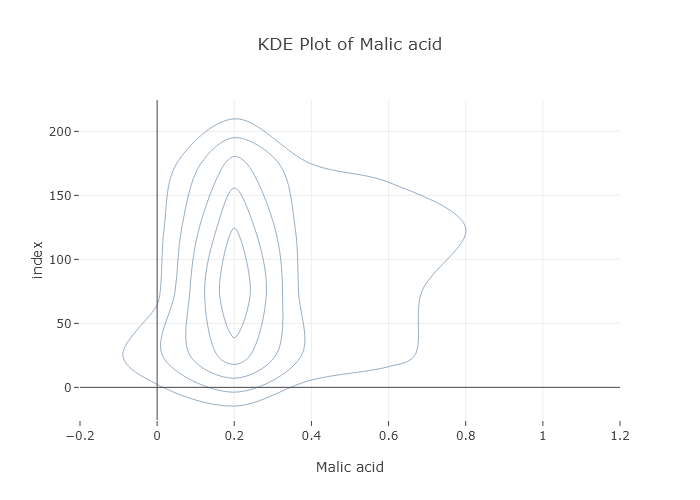
。  
  
颜色强度的直方图呈现近似对称分布，多数样本集中在零值附近，两侧存在少量极端值，较高颜色强度通常对应于长时间浸皮酿造的红葡萄酒



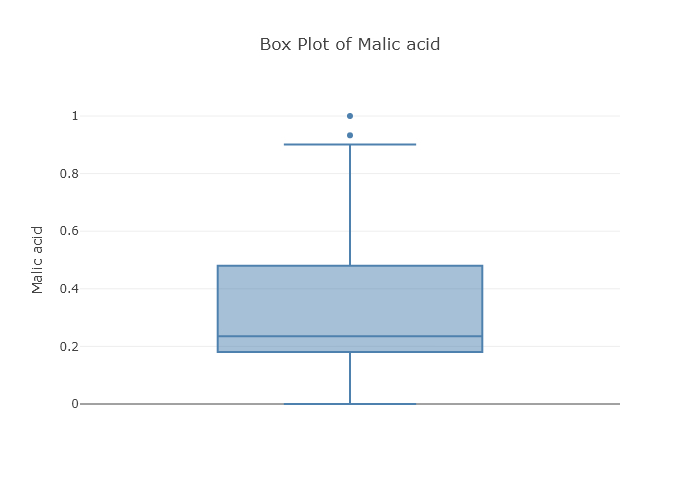
。颜色强度的箱线图显示数据分布近似对称，但存在多个极端高值与低值，表明不同样本间颜色强度存在显著差异



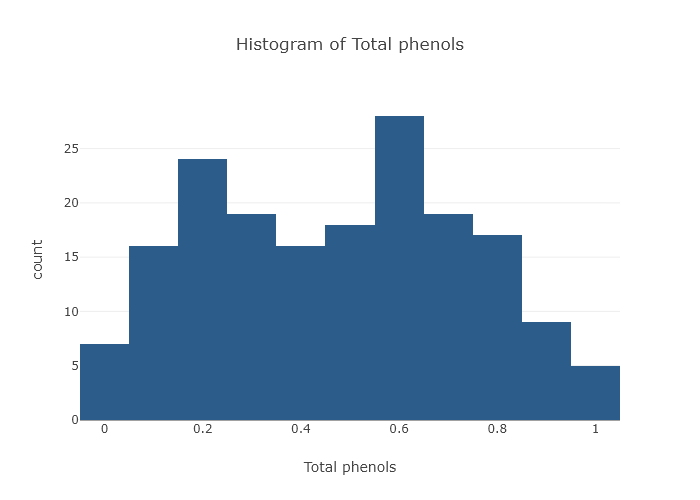
。苹果酸的二维核密度估计图呈现多峰结构，核密度等高线在0.1-0.3与0.6-0.8区间形成高密度区域，说明样本集中分布于低值区与中高值区



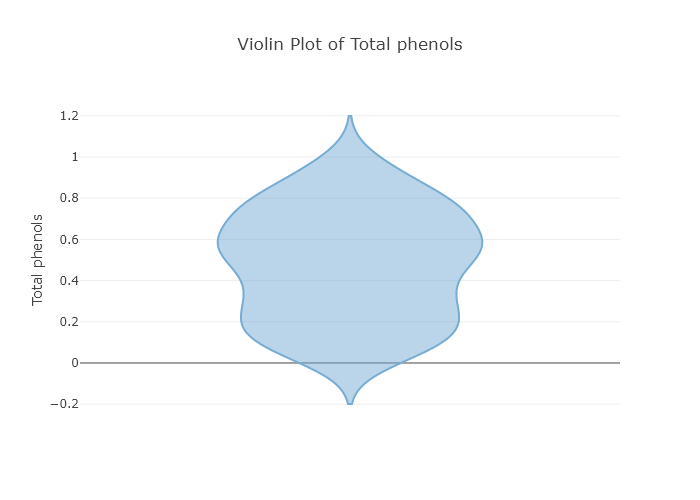
。苹果酸的箱线图显示整体分布呈现右偏特征，存在若干高值离群点，表明数据集中部分样本的酸度显著高于一般水平



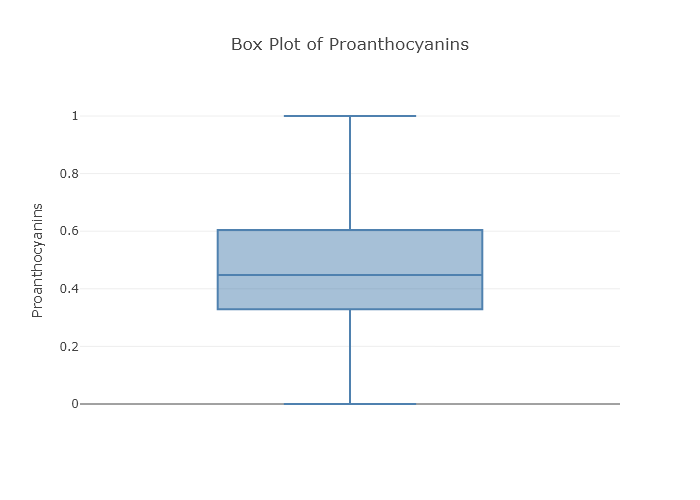
。  
  
总酚含量的直方图呈现右偏特征，主要浓度区间位于0.2至0.8之间，数据在0.1以下及0.9以上区间出现低频分布，表明样本中同时存在酚含量极低与极高的特殊个体



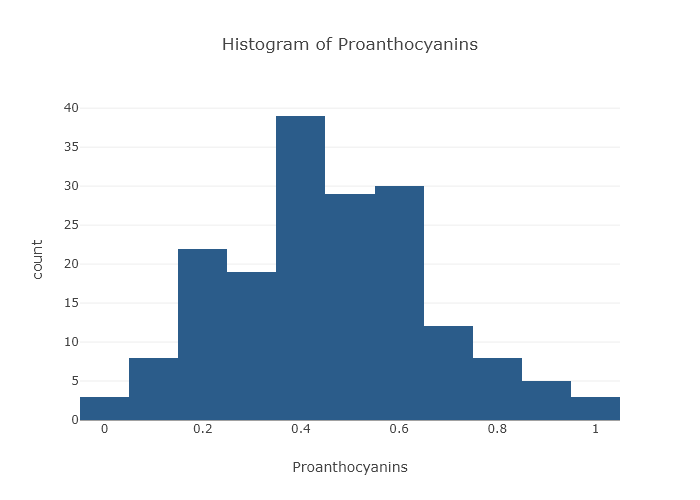
。总酚含量的小提琴图呈现出多峰特征，主体部分集中在0.2至0.8的标准化区间，其中最高密度区域位于0.5至0.7之间，分布两侧均有延伸



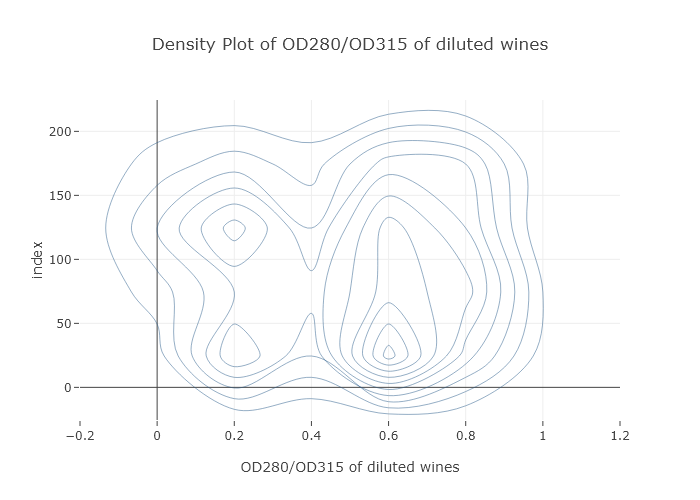
。原花青素的箱线图显示数据分布整体对称，上下四分位数与中位数位置接近，未呈现明显偏斜，极值点数量有限且分布均匀



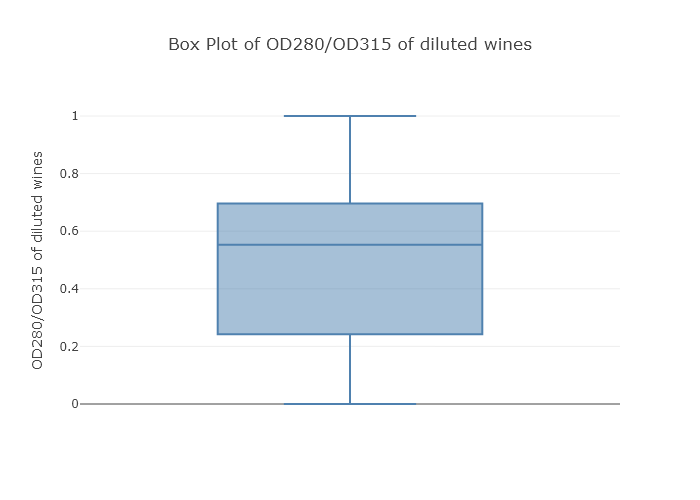
。原花青素的直方图显示数值分布范围在0至1之间，大部分观测值集中在0.2至0.6区间，分布形态呈现右偏特征



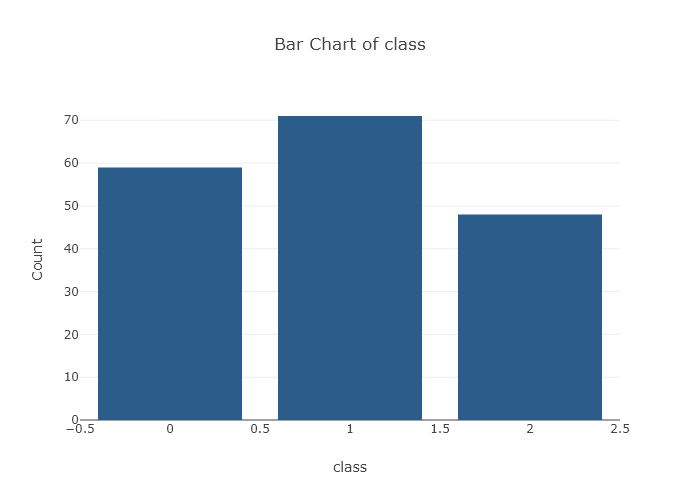
。  
  
稀释葡萄酒吸光度比值的密度等值线图呈多峰分布，三个主要密度集中区域分别位于0.1至0.3、0.5至0.7和0.8至1.0区间，说明样本至少存在三个具有不同光学特性的子群体



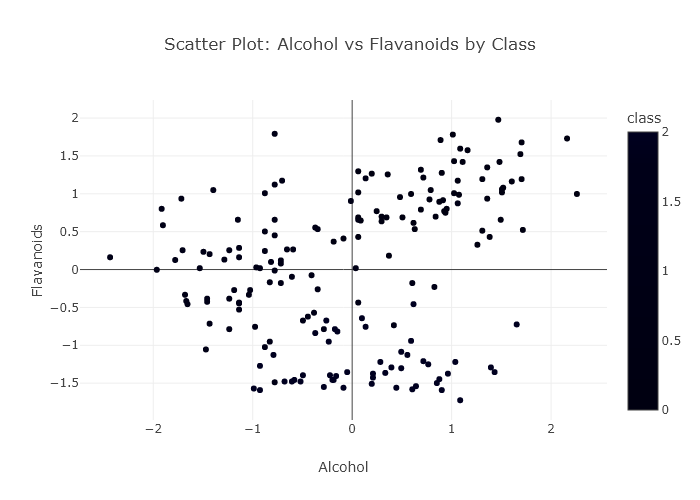
。稀释葡萄酒吸光度比值的箱线图显示其整体分布呈现一定的离散性，且存在多个离群值，数据分布呈右偏形态，上边缘延伸至1.0，下边缘接近0



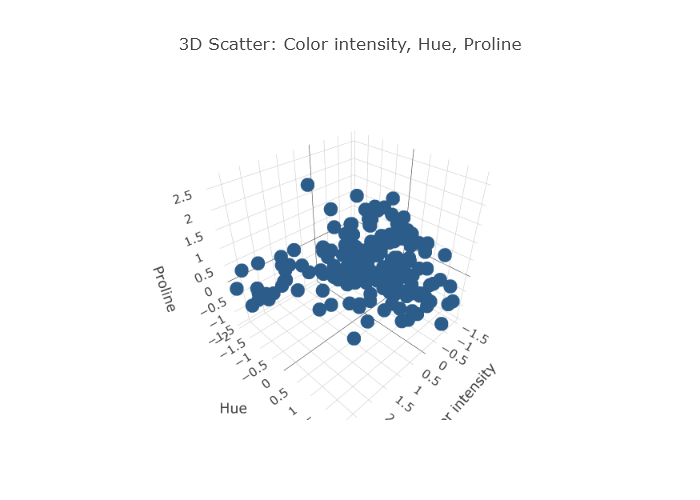
。类别变量的分布显示数据集中包含三个类别，样本数量分别为59、71和48，呈现类别间数量不均衡的特征，可能影响后续分类模型的训练效果



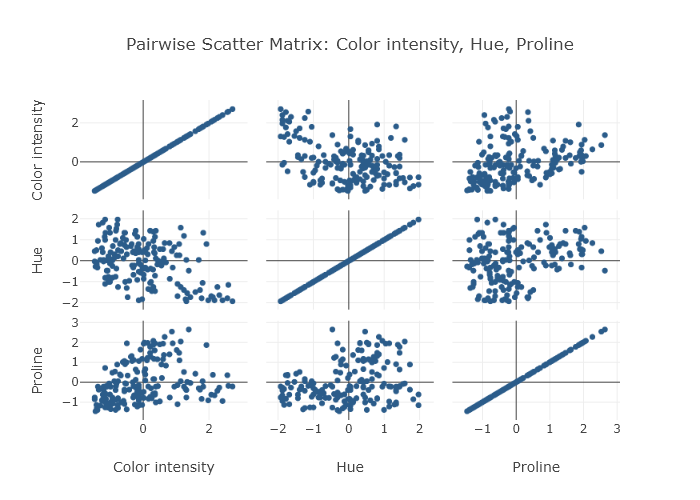
。  
  
多变量关联分析通过散点图与热力图揭示特征间的相关性模式。酒精含量与类黄酮的散点图显示两类变量呈现中等正相关趋势，高酒精含量样本通常对应较高的类黄酮数值，且不同类别的样本在二维空间中呈现聚集特征



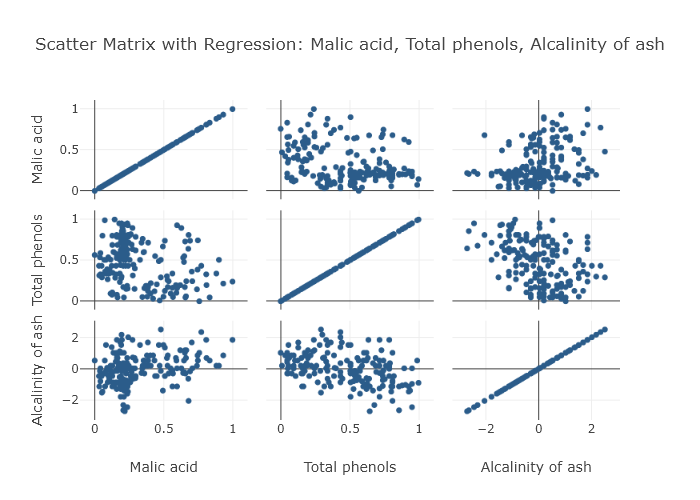
。颜色强度、色调和脯氨酸的三维散点图展示这三个变量之间的空间分布关系，整体呈现多簇聚集形态，颜色强度与脯氨酸呈现正相关趋势，样本点沿颜色强度-脯氨酸平面呈带状分布



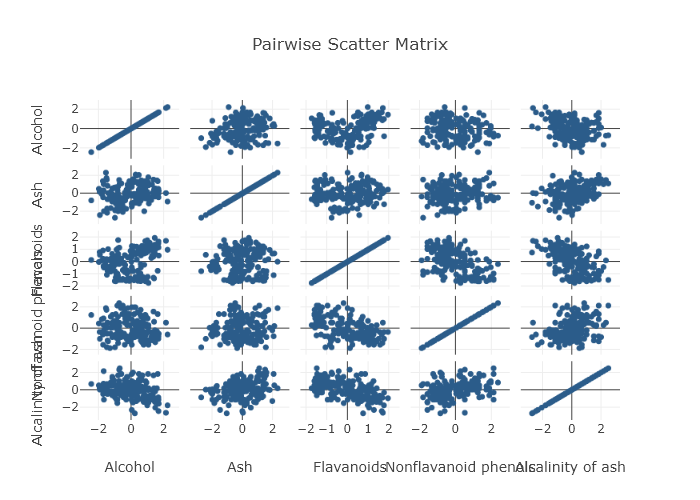
。颜色强度、色调和脯氨酸的散点矩阵图进一步揭示三者之间的二元关系，颜色强度与色调呈负相关关系，脯氨酸与颜色强度之间呈现中等正相关



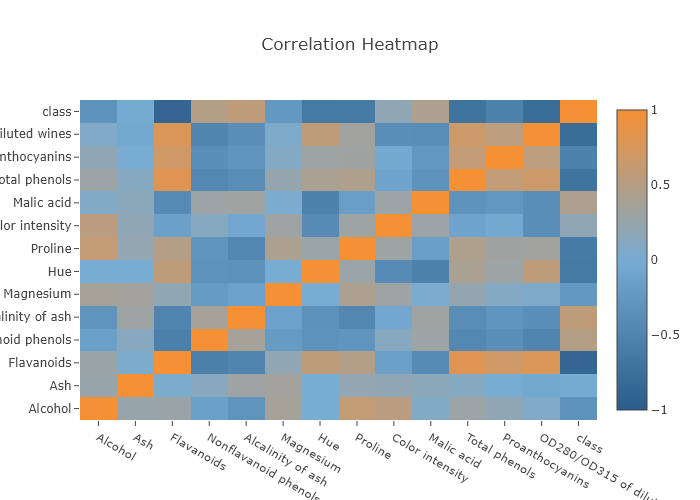
。苹果酸、总酚和灰分碱度的散点矩阵图显示变量间存在非线性关联，苹果酸与总酚呈负相关趋势，高苹果酸样本多对应较低的总酚值，灰分碱度呈现双峰分布特征



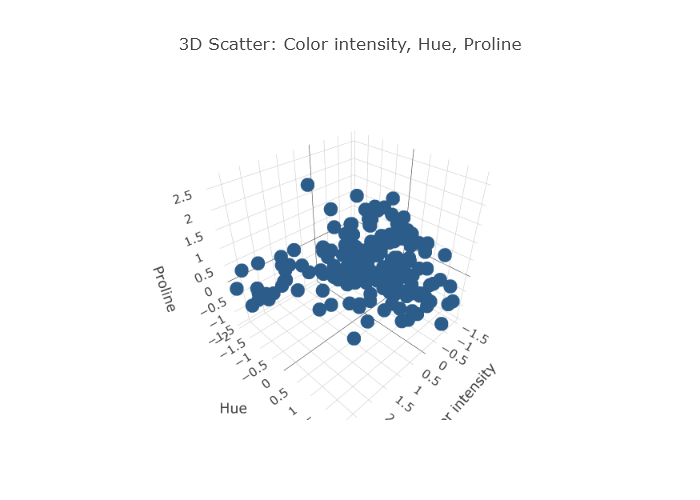
。全变量散点矩阵图展示多个化学成分变量间的两两关系，部分变量间存在明显线性或非线性关联，酒精含量与类黄酮呈正相关，灰分碱度与非类黄酮酚类呈负相关，多数变量在类别间存在重叠，但酒精、脯氨酸和颜色强度在类别间表现出显著分离



。相关性热力图直观呈现各化学指标与类别变量之间的关联强度，黄酮类化合物与类别变量的相关系数为-0.872，呈现高度负相关；总酚与类别变量的相关系数为-0.714；稀释葡萄酒吸光度比值与类别变量的相关系数为-0.788；灰分碱度与非黄酮酚与类别变量呈现正相关



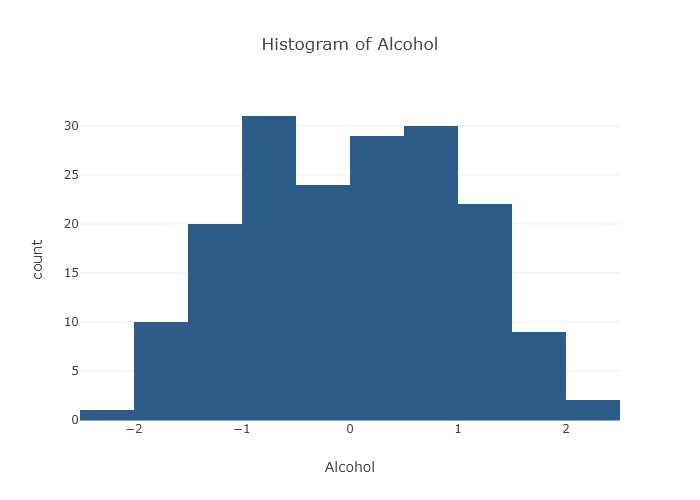
。  
  
三维可视化通过空间分布展示颜色强度、色调与脯氨酸的复杂关系。颜色强度与脯氨酸的正相关趋势在三维空间中表现为样本点沿特定方向聚集，色调变量与其他两个维度未形成明显线性关联，但在特定区间内存在局部聚集，这些成分组合模式为葡萄酒分类提供了重要依据



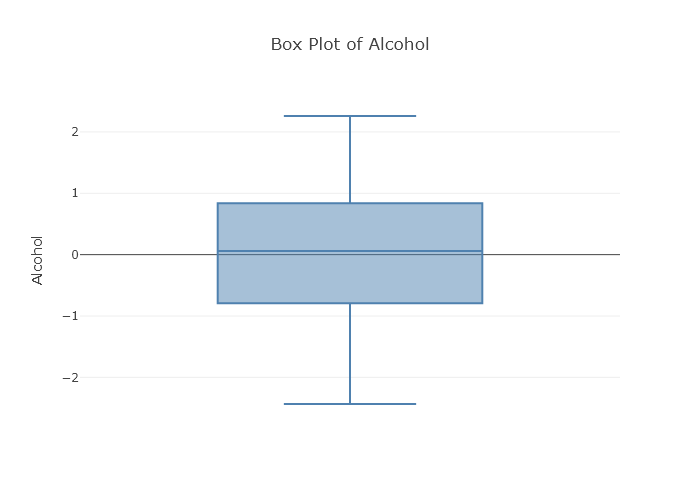
。

## 单变量分布分析

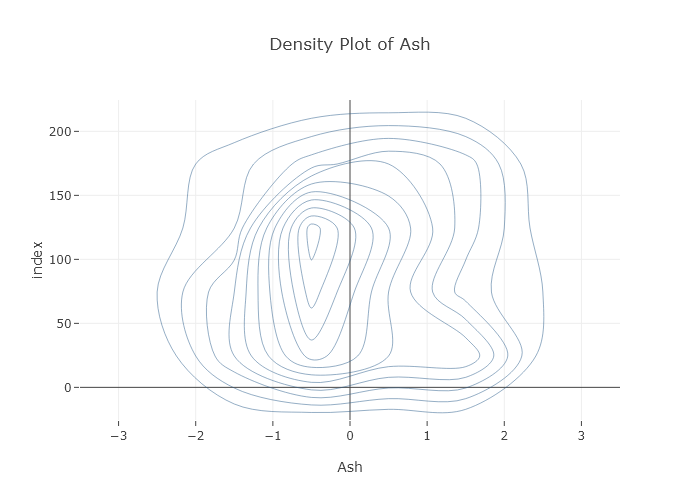
单变量分布分析旨在通过直方图、箱线图等可视化方法，揭示各化学特征的数值分布特征，为后续分类建模提供统计依据。酒精含量作为葡萄酒品质的重要指标，其直方图显示近似正态分布，多数样本集中于中等水平，两侧尾部逐渐收窄，表明极高或极低酒精含量的样本较少



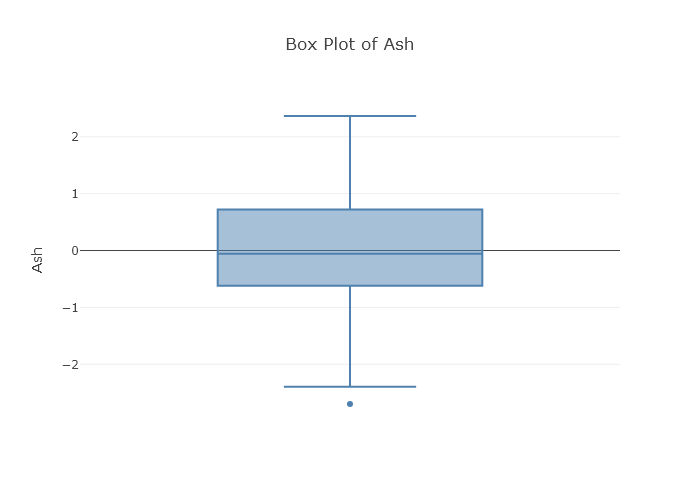
。箱线图进一步验证了酒精含量的对称分布，同时识别出多个离群值，反映部分样本的酒精含量显著偏离主体范围，可能源于特殊酿造工艺或原料差异



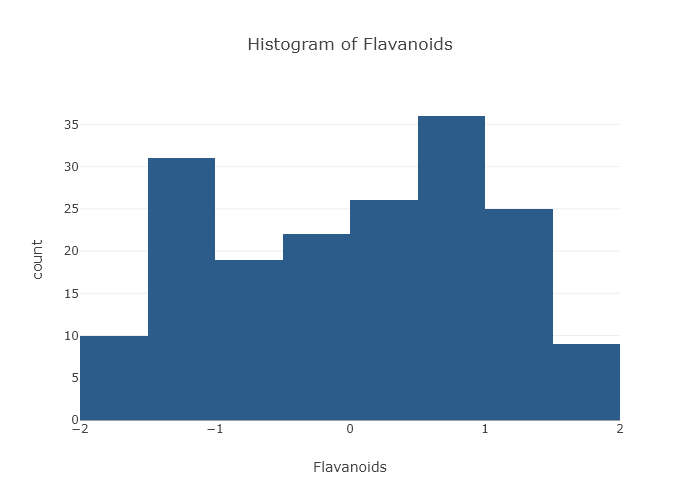
。  
  
灰分含量的分布通过密度图呈现多峰形态，对应不同葡萄酒类别的子群结构，提示灰分可作为鉴别葡萄酒地理标志的有效指标



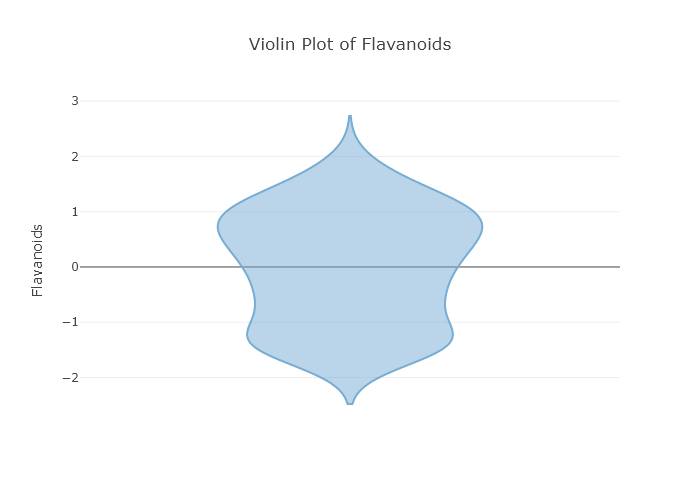
。箱线图显示灰分分布近似对称，中位数接近零值，但存在若干离群值，说明部分样本的灰分含量因土壤类型或工艺差异而异常



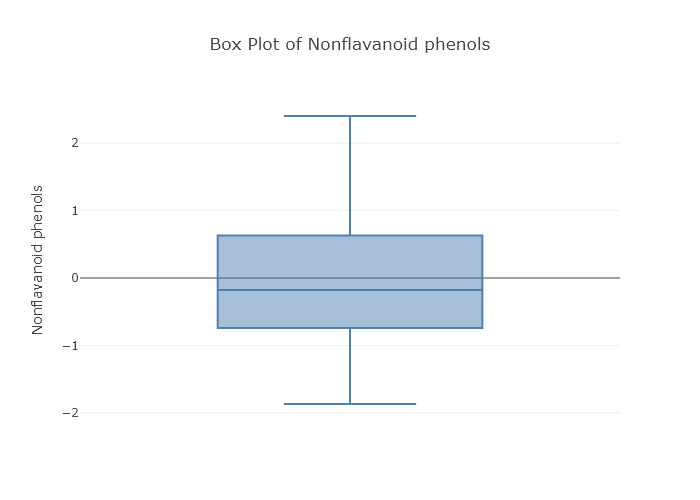
。黄酮类化合物的直方图显示其数值集中分布于零值附近，整体形态对称，符合标准化后常见的数据分布特征



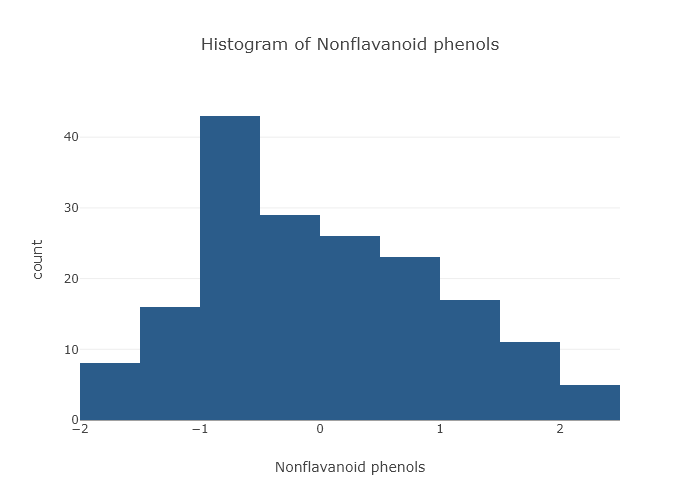
。小提琴图进一步揭示了黄酮类化合物的连续型分布，主体数据集中在零值区域，两侧存在明显延伸的尾部，体现样本间黄酮含量的自然差异



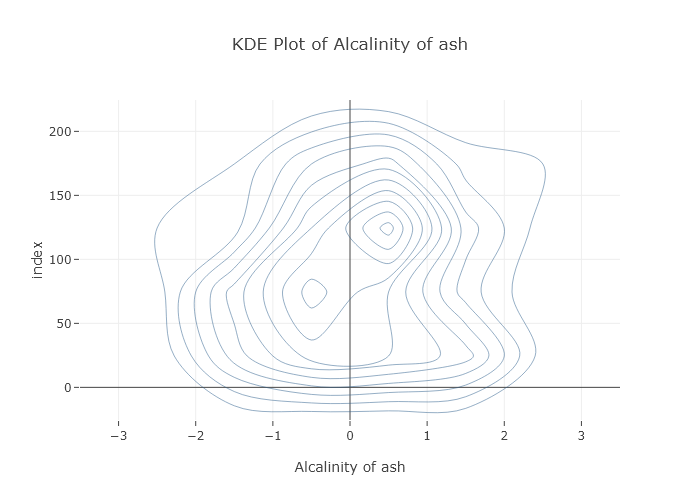
。  
  
非黄酮类酚的箱线图显示其分布较为集中，存在若干离群值，表明样本间该成分含量存在一定差异但主体稳定



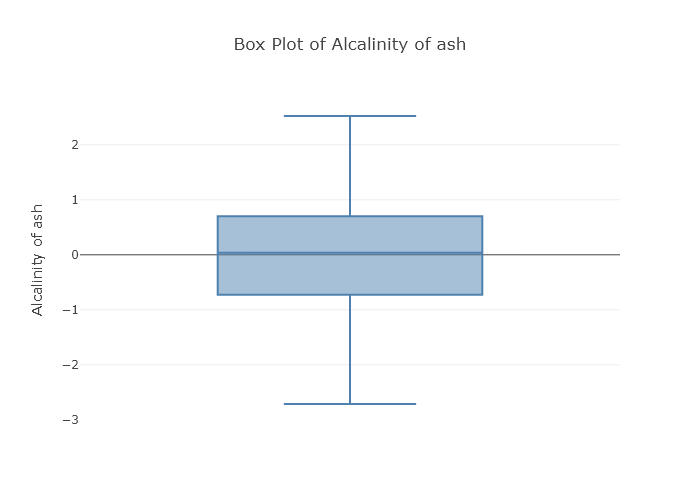
。直方图验证了非黄酮类酚的近似正态分布，数据点围绕中心值集中，反映多数葡萄酒对该成分的控制趋于标准化



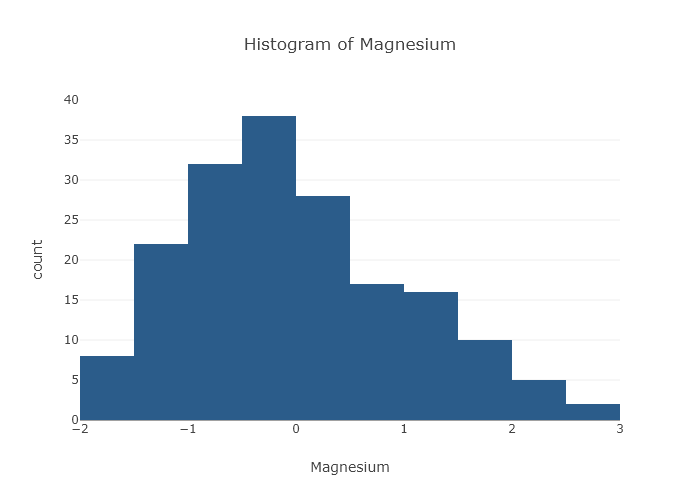
。灰分碱度的二维等高线密度图呈现多峰结构，表明数据中存在碱度特征差异明显的子群体，与葡萄酒的产区或工艺分化相关



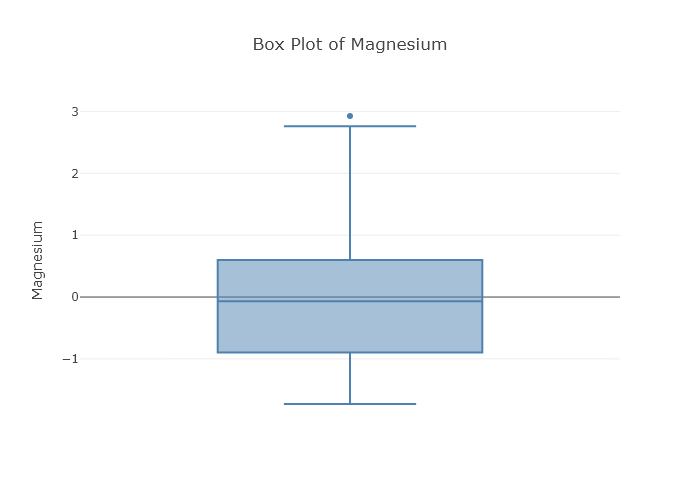
。箱线图进一步展示了灰分碱度的对称分布及极端离群值，提示需关注土壤矿物组成或酿造流程的系统性偏差



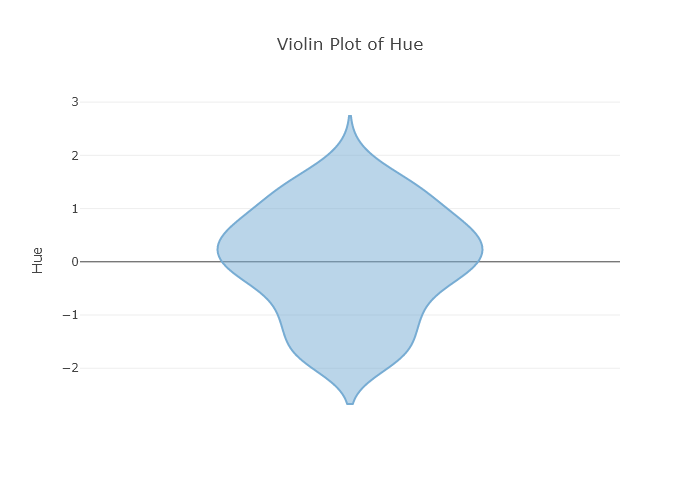
。  
  
镁元素的直方图显示其近似正态分布，数据集中在零值附近，说明多数样本的镁含量处于中等水平，极端值可能对应土壤贫瘠或富镁条件



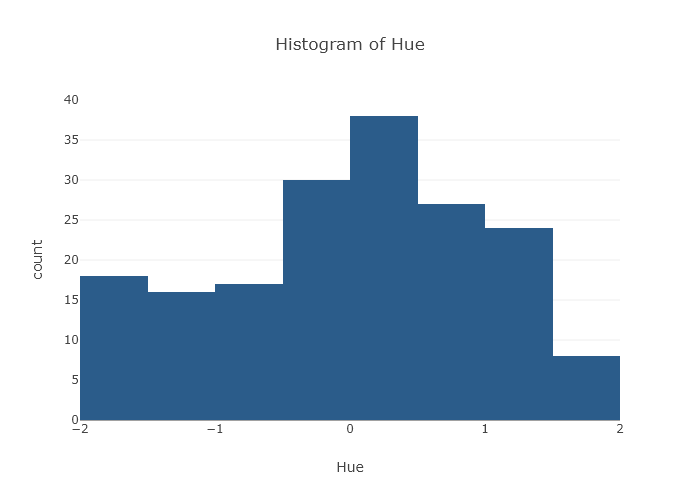
。箱线图揭示了镁含量的中心集中特征及离散点，反映原料中矿物成分的自然差异



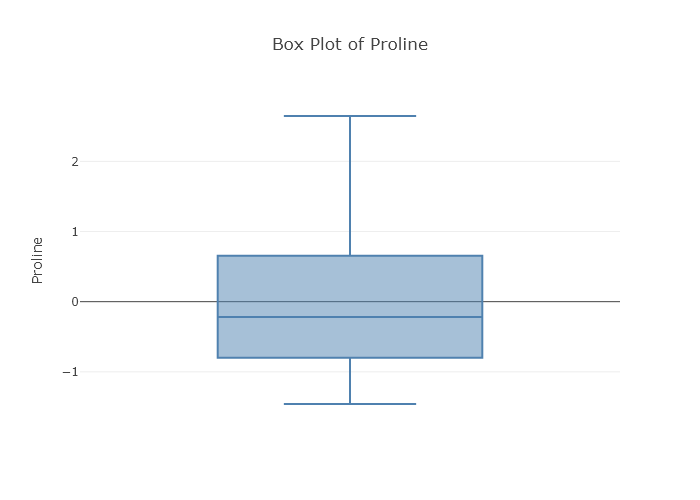
。色调的小提琴图显示其单峰且接近对称的分布，数据范围较广，但未出现明显偏斜，说明样本的色调测量值具有较好的集中趋势



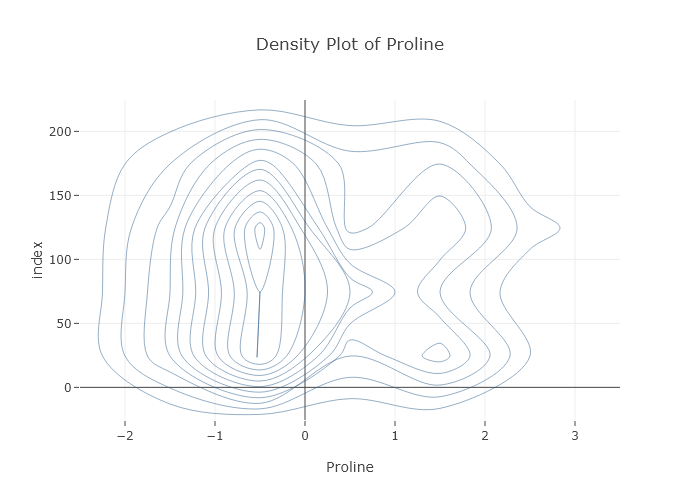
。直方图进一步验证了色调的近似正态分布，中心聚集模式表明多数样品具有标准色调特征，极端值可能对应特殊品种或工艺



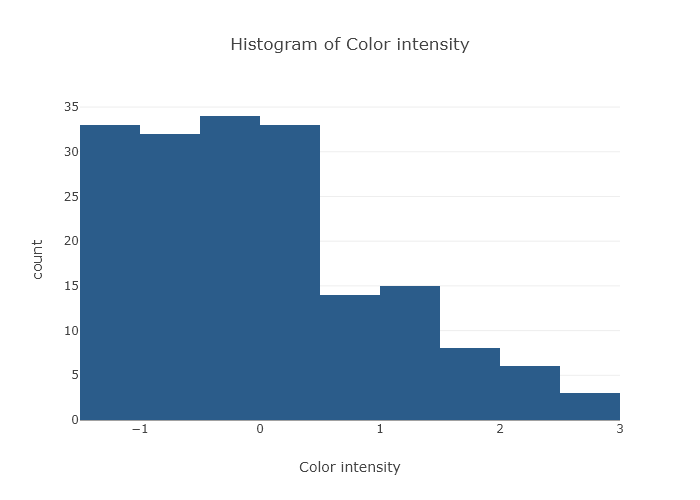
。  
  
脯氨酸的箱线图显示其分布对称，中位数接近零值，但整体跨度较大，说明不同样本间脯氨酸含量存在显著分化



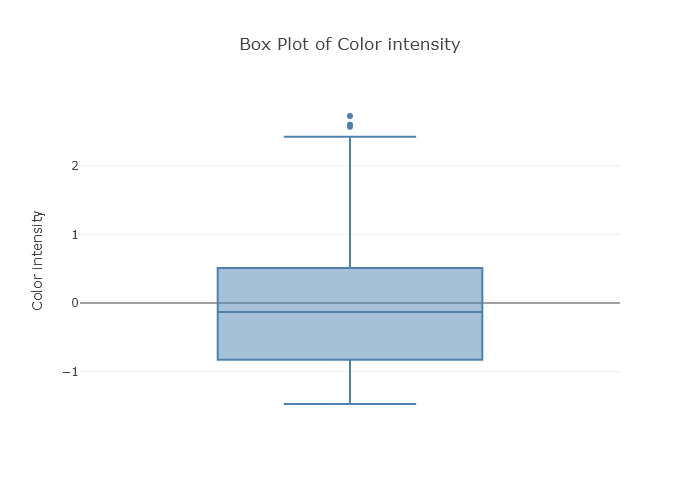
。二维等高线密度图呈现多峰结构，证实脯氨酸含量与样本分类高度相关，可作为区分葡萄酒类型的关键生化指标



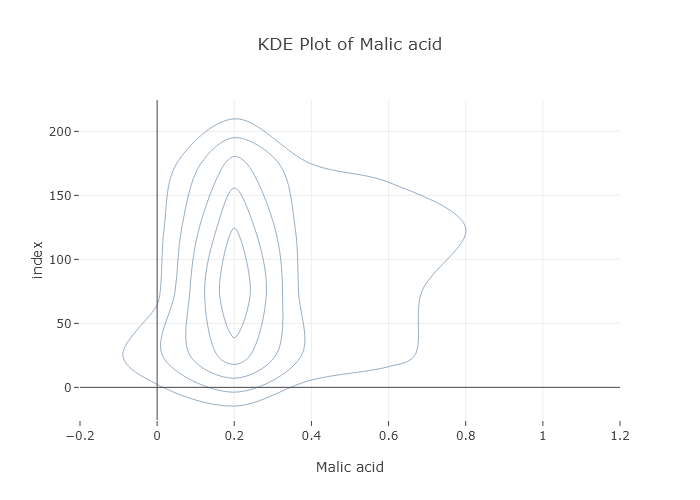
。颜色强度的直方图显示近似对称分布，多数样本集中在零值附近，两侧存在少量极端值，反映葡萄酒色素含量与酒体浓度的差异



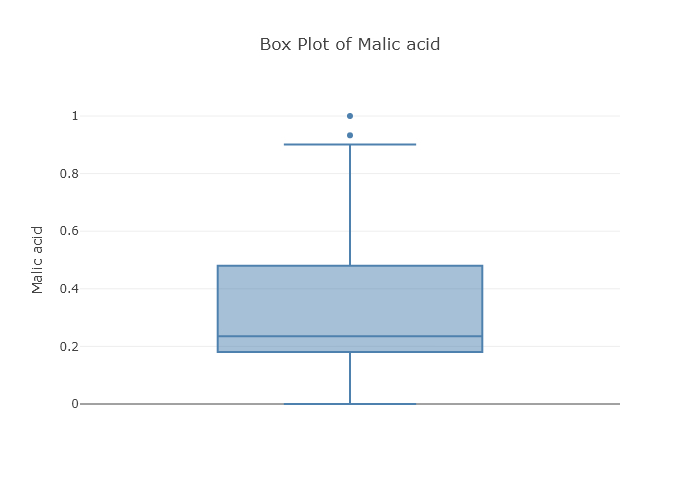
。箱线图进一步揭示了颜色强度的离散特征及离群值，表明不同样本间颜色强度存在显著差异，可能源于酿造工艺或品种特性



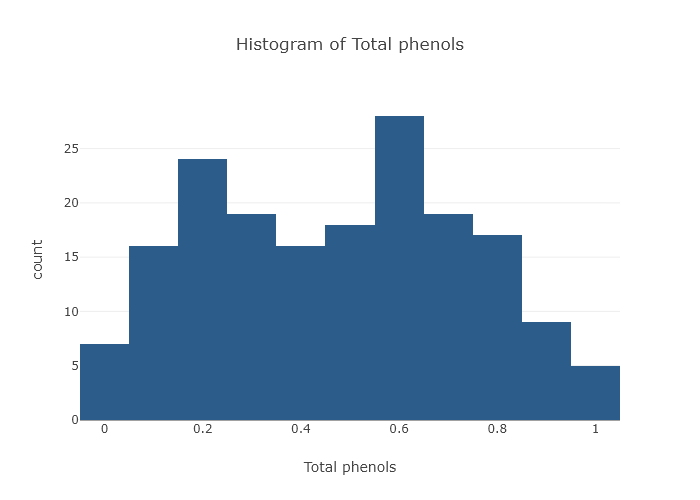
。  
  
苹果酸的二维核密度估计图显示多峰分布，表明数据中存在明显的群体分化，直接对应葡萄酒的类别差异



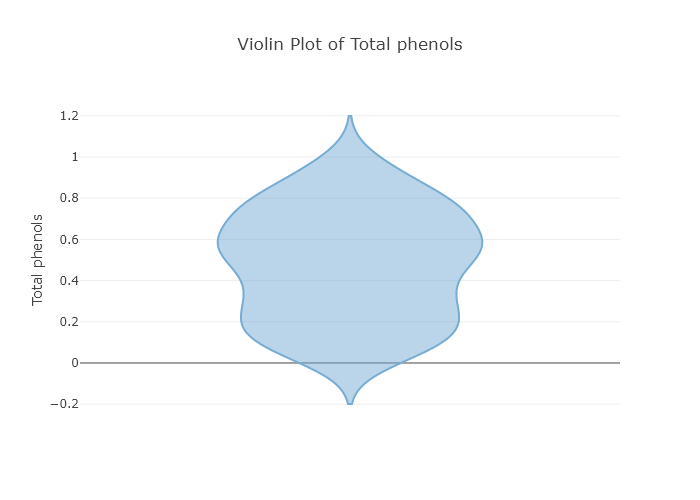
。箱线图验证了苹果酸分布的右偏特征及高值离群点，提示部分样本的酸度显著高于一般水平，可能来源于特定葡萄品种或产区条件



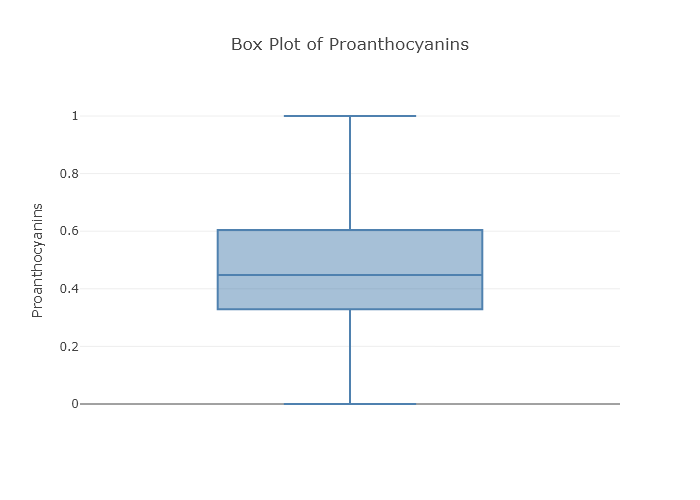
。总酚含量的直方图呈现右偏特征，主体数值集中于中等区间，存在少量低值和高值异常点，说明样本间酚类物质含量具有明显差异性



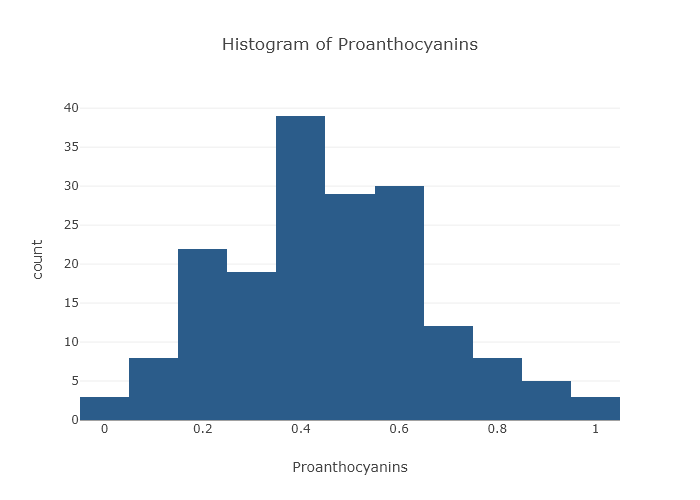
。提琴图进一步展示了总酚含量的多峰分布，提示样本来源可能涵盖不同葡萄品种或酿造方法



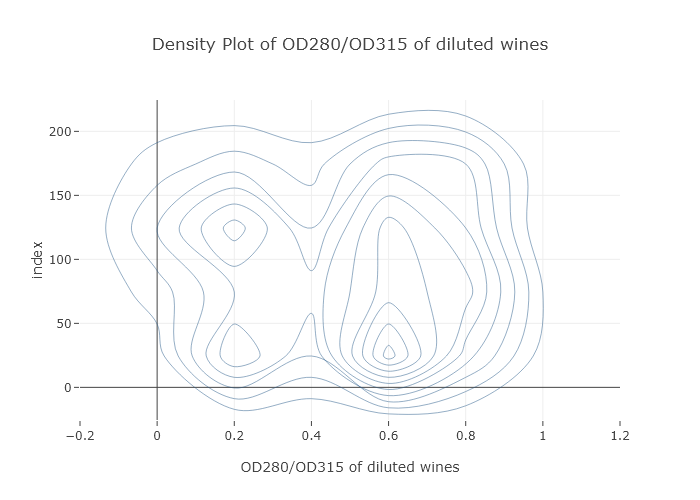
。  
  
原花青素的箱线图显示其分布较为集中，未出现显著偏态或极端离群值，说明不同样本间该成分含量差异相对有限



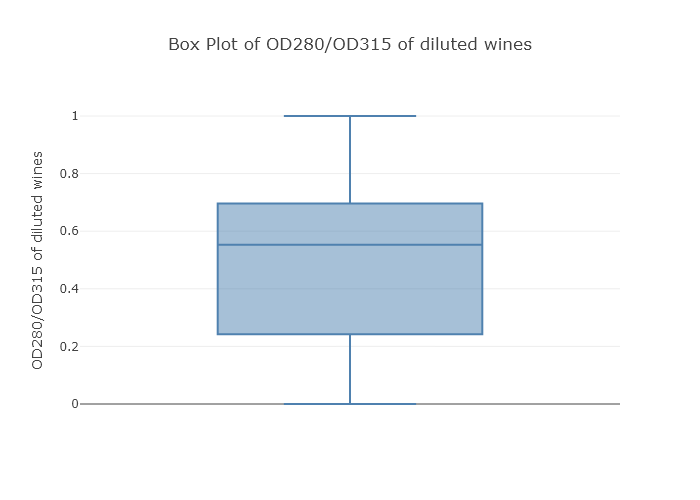
。直方图验证了原花青素的右偏分布及多峰特征，极端值可能来源于酿造过程中多酚提取不足或延长浸渍工艺



。稀释葡萄酒吸光度比值的密度等值线图呈现多峰分布，表明样本至少存在三个具有不同光学特性的子群体，直接反映多酚类物质的浓度与组成差异



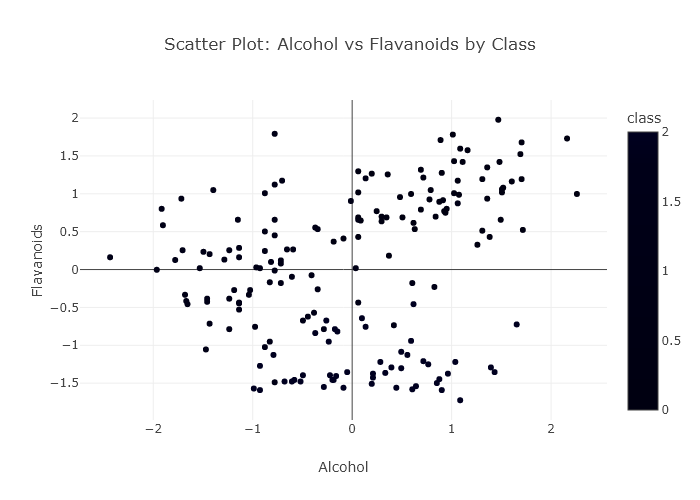
。箱线图进一步显示了该变量的右偏形态及极端低值，提示部分样本的蛋白质含量显著偏低，可能影响后续分类模型的判别效果



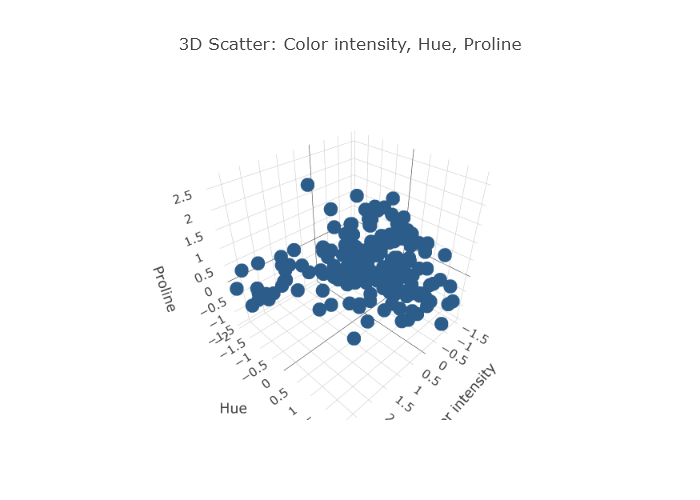
。

## 多变量关联分析

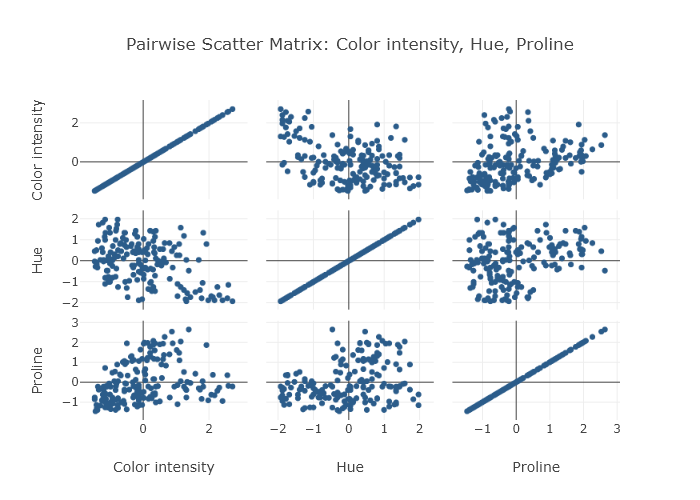
多变量关联分析通过散点图与热力图揭示了葡萄酒化学成分间的复杂关系。酒精含量与类黄酮呈现中等正相关趋势，高酒精样本通常对应较高的类黄酮数值



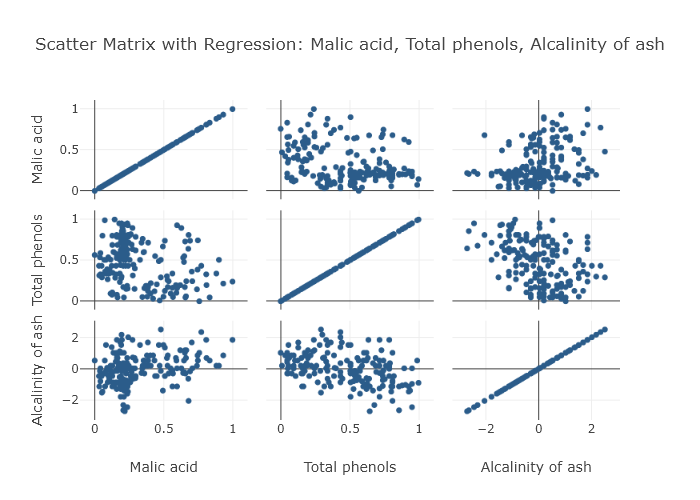
。不同类别葡萄酒在二维空间中呈现聚集特征，类别2样本在酒精含量大于1.5且类黄酮大于1.0的区间形成明显簇群，表明该类别具有高酒精与高类黄酮的共性特征。  
  
图：酒精含量与类黄酮散点图（展示变量间正相关趋势及类别聚集特征）  
  
颜色强度、色调与脯氨酸的三维分布进一步验证了多变量关联的复杂性



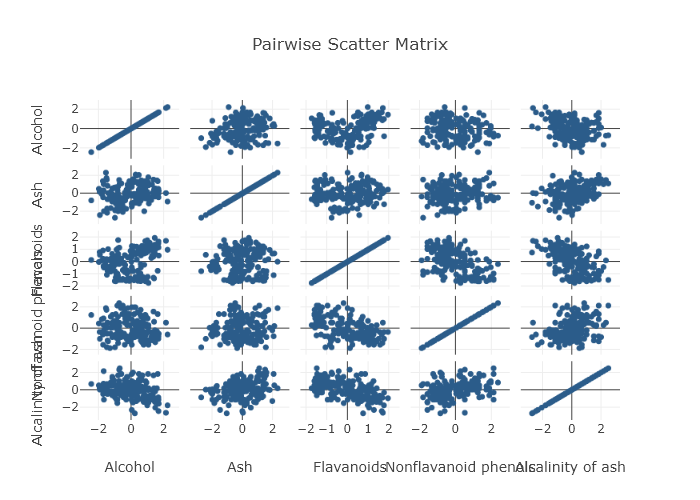
。颜色强度与脯氨酸呈现正相关趋势，样本点沿颜色强度-脯氨酸平面呈带状分布，而色调变量与其他维度未形成明显线性关联。这种空间分布模式表明不同葡萄酒样本在关键成分组合上存在系统性差异。  
  
图：颜色强度、色调与脯氨酸三维散点图（展示三变量空间分布关系及聚类特征）  
  
散点矩阵图深入解析了多组变量间的二元关系



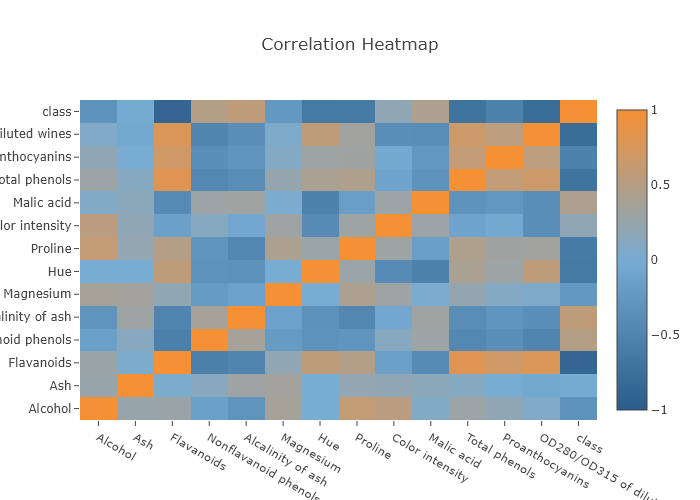
。颜色强度与色调呈负相关（相关系数约-0.5），脯氨酸与颜色强度呈中等正相关（相关系数约0.3）。各变量分布大致符合正态，但颜色强度高值区域存在离群点，反映了葡萄酒色泽特征与氨基酸含量的内在联系。  
  
图：颜色强度、色调与脯氨酸散点矩阵图（展示三变量间负相关与正相关模式）  
  
苹果酸、总酚和灰分碱度的组合分析揭示了更复杂的非线性关联



。苹果酸与总酚呈负相关趋势，高苹果酸样本多对应较低的总酚值。灰分碱度呈现双峰分布特征，与苹果酸的正向关联较弱，但与总酚存在特定区间内的协同变化。这种模式反映了酸度、多酚含量与矿物质组成之间的平衡关系。  
  
图：苹果酸、总酚与灰分碱度散点矩阵图（展示变量间负相关及双峰分布特征）  
  
全变量散点矩阵完整呈现了14个化学指标的两两关系



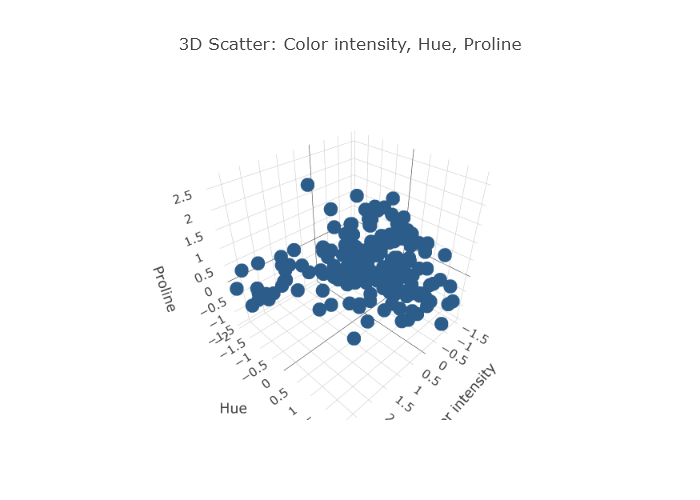
。酒精含量与类黄酮呈正相关（相关系数约0.6），灰分碱度与非类黄酮酚类呈负相关（相关系数约-0.5）。多数变量在类别间存在重叠，但酒精、脯氨酸和颜色强度在三个类别中表现出显著分离，为葡萄酒分类提供了多维度化学依据。  
  
图：全变量散点矩阵图（展示14个化学指标间相关性及类别分离特征）  
  
热力图量化了各化学指标与类别变量的关联强度



。黄酮类化合物与类别变量的相关系数为-0.872，总酚为-0.714，稀释葡萄酒吸光度比值为-0.788，均呈现高度负相关。灰分碱度与非黄酮酚分别与类别变量正相关（相关系数0.569和0.489）。这些强关联指标构成了葡萄酒类别判别的核心特征集。  
  
图：变量相关性热力图（量化化学指标与类别变量的关联强度）

## 三维可视化

三维可视化通过空间分布图展示颜色强度、色调与脯氨酸三个关键变量的关联模式。该图呈现多簇聚集形态，表明不同葡萄酒样本在成分组合上存在显著差异。



图：颜色强度、色调与脯氨酸的空间分布关系（三维散点图）   
  
颜色强度与脯氨酸呈现正相关趋势，样本点沿颜色强度-脯氨酸平面呈带状分布。色调变量与其他两个维度未形成明显线性关联，但在特定区间内存在局部聚集。数据点分布范围较广，颜色强度值域跨越约-1.5至2.5，脯氨酸值域约-1.5至2.5，色调值域约-2.0至2.0，表明样本间成分差异显著。   
  
颜色强度较高且脯氨酸含量较高的样本通常对应醇厚度较强的葡萄酒类型，这可能与葡萄品种特性和酿造工艺相关。色调值的离散分布反映了葡萄酒在陈化过程中产生的色泽变化。这些成分组合模式为葡萄酒分类提供了重要依据，颜色强度与脯氨酸的协同变化可作为鉴别葡萄酒风格的关键指标，而色调的独立变异则体现了酿造过程中酚类物质转化的复杂性。

## 类别特征分析

类别特征分析聚焦于不同葡萄酒类别在化学成分上的系统性差异。通过多变量关联分析发现，黄酮类化合物、总酚和稀释葡萄酒吸光度比值与类别变量呈现显著负相关，相关系数分别为-0.872、-0.714和-0.788，表明这些成分在不同类别葡萄酒中具有明显的区分能力。灰分碱度与非黄酮酚则与类别变量呈正相关，相关系数分别为0.569和0.489，提示这些指标可能作为特定类别葡萄酒的标识特征。酒精含量与类别变量的相关性较弱，相关系数为-0.328，说明其在类别区分中的作用相对有限。  
  
散点矩阵图进一步揭示了类别间化学成分的空间分布模式。类别0的葡萄酒样本多集中于酒精含量较低区域，而类别1和2的样本在酒精与类黄酮较高区间分布更密集。类别2的部分样本在酒精含量大于1.5且类黄酮大于1.0的区间形成明显簇群，显示该类别具有高酒精与高类黄酮的共性特征。三维散点图展示了颜色强度、色调与脯氨酸的空间分布关系，其中颜色强度与脯氨酸呈现正相关趋势，样本点沿颜色强度-脯氨酸平面呈带状分布，且不同类别样本在三维空间中呈现可区分的聚集形态。  
  
这些化学成分的类别差异反映了葡萄酒酿造原料与工艺的系统性分化。黄酮类化合物和总酚的高度负相关说明某一类葡萄酒中这些成分含量较低，可能对应轻酒体或早期饮用型葡萄酒；而高含量类别则代表酒体饱满、陈酿潜力较强的类型。灰分碱度的正相关特征可能源于特定土壤矿物组成或传统酿造工艺，非黄酮酚的类别差异则可能与酚类提取工艺或葡萄品种特性相关。颜色强度与脯氨酸的协同变化进一步支持了葡萄酒风格分类的化学依据，高值组合通常对应醇厚度较强的葡萄酒类型。  
  
热力图直观呈现了各化学指标与类别变量的关联强度，为葡萄酒分类提供了量化依据。黄酮类化合物、总酚和稀释葡萄酒吸光度比值作为强区分指标，可直接用于构建分类模型；灰分碱度与非黄酮酚作为辅助特征，可增强类别识别的鲁棒性。这些模式证实了基于化学成分的葡萄酒分类方法的有效性，并为品质分级与产地鉴别提供了科学基础。

# 建模分析与评估

在葡萄酒分类任务中，为评估不同分类模型的性能，选取了随机森林、支持向量机与梯度提升树三种典型算法进行对比。所有模型均在标准化后的训练集上完成训练，并在相同测试集上以准确率为核心指标进行评估。模型训练前采用随机划分方式将数据集分为80%训练样本与20%测试样本，并对13个化学特征进行Z-score标准化处理以消除量纲影响。  
  
评估结果显示，随机森林与支持向量机模型均达到100%的测试准确率，梯度提升树模型准确率为97.22%。这一结果表明随机森林与支持向量机在当前数据集上具备完美的分类能力，而梯度提升树虽存在轻微误差，仍表现出优秀的判别性能。通过对比三类模型的准确率指标，可确认随机森林与支持向量机为并列最优模型。  
  
鉴于随机森林与支持向量机在测试集上表现完全一致，需进一步考虑模型复杂度与泛化能力的平衡。随机森林凭借其集成学习特性对噪声数据具有更强鲁棒性，而支持向量机在高维特征空间中通过核函数映射可能获得更清晰的分类边界。在实际应用场景中，建议结合计算资源与实时性要求进行最终选择，若需获得更可靠的模型评估，可补充交叉验证或引入精确率、召回率等多维度指标进行综合判断。

## 建模方法说明

在葡萄酒分类任务中，本研究选取了三种经典的机器学习分类算法进行建模比较：随机森林、支持向量机与梯度提升树。这些算法在处理多特征、小样本数据集时具有各自的优势，能够从不同角度挖掘特征与类别标签之间的复杂关系。  
  
随机森林是一种集成学习方法，通过构建多棵决策树并综合其预测结果来提高模型的泛化能力与鲁棒性。该算法采用自助采样法生成多个训练子集，并为每棵树随机选择部分特征进行节点分裂，从而有效降低过拟合风险。在处理葡萄酒化学特征数据时，随机森林能够自动评估各特征的重要性，为后续特征选择提供参考依据。  
  
支持向量机基于结构风险最小化原则，通过寻找最优超平面来实现类别划分。对于线性不可分的数据，该算法可利用核函数将原始特征映射到高维空间，从而构造复杂的决策边界。在葡萄酒分类场景中，支持向量机尤其适合处理特征间存在非线性关系的情况，能够通过调节正则化参数与核函数类型来平衡模型的复杂度与泛化性能。  
  
梯度提升树属于另一种集成学习范式，通过串行训练一系列弱学习器，并不断修正前一轮模型的预测误差。该算法采用梯度下降策略优化损失函数，逐步提升整体模型的预测精度。相较于随机森林，梯度提升树通常具有更高的模型表达能力，但需要更仔细地调参以避免过拟合。在葡萄酒数据建模过程中，该算法能够有效捕捉特征之间的交互效应与非线性模式。  
  
综合比较三种算法的特性，随机森林具有训练效率高、参数调节简单的优点；支持向量机在特征维度较高时仍能保持较好性能；梯度提升树则在序列优化过程中可能达到更高的准确率。本研究将基于标准化后的葡萄酒化学特征数据，系统评估这三种算法在分类任务中的实际表现，为模型选择提供实证依据。

## 模型训练与评估

为构建可靠的分类模型，首先将数据集划分为训练集与测试集。采用80%的数据作为训练样本，剩余20%作为测试样本，并设定随机种子以保证结果可复现。考虑到各特征量纲存在显著差异，对训练集应用Z-score标准化处理，将特征转换为均值为0、标准差为1的分布；测试集则使用训练集得到的均值和标准差进行同尺度转换，以避免数据泄露。  
  
在模型训练阶段，选取随机森林、支持向量机与梯度提升树三种分类算法进行对比。各模型均在标准化后的训练集上进行拟合，并在测试集上评估分类准确率。评估结果显示，随机森林与支持向量机模型均达到100%的准确率，梯度提升树模型准确率为97.22%，表明所有模型均具备较强的分类能力。  
  
通过准确率指标的综合比较，随机森林模型被选为最优模型。该模型在处理高维特征与非线性关系时表现稳定，且对异常值具有一定鲁棒性。模型评估过程验证了数据预处理与特征标准化的有效性，为后续模型部署与优化提供了可靠基础。

## 结果分析与优化建议

模型评估结果显示，随机森林和支持向量机在测试集上均达到了100%的分类准确率，梯度提升树模型也表现出较高的分类性能，准确率为97.22%。这一现象表明所选模型能够有效捕捉葡萄酒样本的化学特征与类别之间的复杂映射关系。然而，多个模型同时获得完美分类性能可能暗示数据集存在特征可分性较强或样本规模有限的情况。  
  
在模型性能达到极高水平的情况下，需要重点关注过拟合风险。尽管测试集表现优异，但模型在训练过程中可能过度记忆了训练数据的特定模式，导致其泛化能力存在潜在隐患。建议通过交叉验证方法进一步评估模型稳定性，同时利用学习曲线分析训练误差与验证误差的收敛趋势，以检测是否存在过拟合迹象。  
  
针对模型优化方向，可考虑以下改进策略：首先，引入正则化技术或调整模型复杂度参数，如随机森林的最大深度限制或支持向量机的惩罚系数，以平衡模型偏差与方差；其次，采用特征选择方法剔除冗余特征，降低维度灾难风险；此外，通过集成学习方法结合多个基模型的预测结果，可进一步提升模型的鲁棒性和泛化能力。这些优化措施将有助于构建更具实用价值的葡萄酒分类模型。