古代玻璃制品的成分分析与鉴别

摘要

本文针对古代文物玻璃饰品的化学成分分析进行分析以及玻璃饰品的类型鉴别，根据已有的数据基础下，根据表单 1 的数据，建立表层风化-相关特征评估模型，分析表层风化与其玻璃类型、纹饰和颜色之间的关系，利用卡方检验和相关性分析综合评价表层风化与其他三个特征变量的关系。然后根据附件的数据的文物化学成分数据来建立相关的模型算法，探索古代文物玻璃饰品的玻璃类型与其化学成分的关系。

在建立模型之前，需要对数据进行分析和预处理，这里我们对数据进行了多方位的观察和分析，同时进行了数据的预处理和清洗，剔除一部分对建模帮助不大的数据。

问题一，建立了表层风化-相关特征评估模型，对表单 12 的合并数据集进行相关性分析，通过协方差系数热力图，找出有效风化化学成分，并根据风化点预测数据建立支持向量回归（SVR）风化化学成分预测模型，划分风化和未风化两类数据，逐次对风化样品数据的单个指标建立预测算法模型。

问题二，由于并不清楚铅钡玻璃和高钾玻璃的亚类划分情况，认为这是无监督学习，采用 K-Means 聚类算法来确定铅钡玻璃和高钾玻璃的亚类划分，然后给出合理的划分方法和划分结果，并用轮廓系数评估聚类结果。

问题三，对表单 12 的合并数据集，以玻璃文物饰品的所有化学成分及表面风化作为输入变量，玻璃类型作为输出变量，建立逻辑回归模型预测文物的玻璃类型，根据问题类簇中心与表单 3 的数据距离判定玻璃类型，结合判定结果和预测结果，分析逻辑回归模型的可行性和预测结果的敏感性。

问题四，已知玻璃制品的添加剂、助熔剂不同，其化学成分的含量也会有所不同。那么就需要分析不同类别玻璃制品的化学成分含量之间的差异性，本文采用方差分析算法来分析不同类别玻璃制品、化学成分之间的关系。

关键字：表层风化-相关特征评估模型　K-Means 聚类算法　卡方检验　SVR 风化化学成分预测模型　方差分析

一、 问题重述

丝绸之路是古代中西方文化交流的通道，而玻璃是古代中西方贸易往来的宝贵物证。早期的玻璃在西亚和埃及地区常被制作成珠形饰品传入我国，我国古代玻璃吸收其制作珠形饰品的技术后便在本土就地取材制作，因此我国的玻璃制品与外来的玻璃制品外观十分相似，但是化学成分却有所不同。

玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO2）。由于纯石英砂的熔点较高，所以为了降低熔化温度，在炼制二氧化硅的时候需要添加助熔剂。在古代，常用的助熔剂有草木灰、天然泡碱等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）。添加的助熔剂不同，其主要化学成分也不同。

古代玻璃极易受埋藏的影响而变化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例变化，从而影响对其类别的正确判断。根据已有的一批我国古代玻璃制品的相关数据，考古工作者依据这些文物样品的化学成分和其他检测手段已将其分为高钾玻璃和铅钡玻璃两种类型。

古代玻璃极易受埋藏的影响而变化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例变化，从而影响对其类别的正确判断。根据已有的一批我国古代玻璃制品的相关数据，解决以下问题。

问题一：

这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析；结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律；根据风化点检测数据，预测其风化前的化学成分含量。

问题二：

根据附件信息，分析高钾玻璃、铅钡玻璃的分类规律并进行亚类划分，给出具体的划分方法和划分结果，分析其合理性和敏感性。

问题三：

分析附件表单 3 玻璃文物的化学成分，鉴别它们的所属类型，对分类结果的敏感性进行分析。

问题四：

针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系，并比较不同类别玻璃文物样品之间的化学成分关联关系的差异性。

二、 问题分析

2.1 问题一的分析

问题一的要求是对这些玻璃文物的表层风化情况与其类型、纹饰和颜色进行相关性分析，那么需要建立一个表层风化——相关特征评估模型，定义和描述一种相关特征的评估的方法或者指标，考虑如何评估表层风化情况与其类型、纹饰和颜色这三个特征的关系。已知附件表单 1 的数据全是类别型数据，除了文物编号外。对于这种类型的数据，我们需要选择一种适合该类型数据的评估方法进行表层风化——相关特征评估模型。然后结合玻璃类型来对玻璃文物样品表面进行有无风化化学成分含量的统计规律进行分析，从附件表单 2 的数据给出了与表单 1 中的文物所对应的主要成分所占比例，这里则需要将表单 1 和表单 2 的数据进行合并，然后进行分析。根据每个化学成分与玻璃文物表层风化的相关性进行一个排序，以及根据所有风化化学成分的变化率来判断并选取排名前列的一些化学成分作为主要的风化化学成分，需要建立一个风化化学成分预测模型，由于风化化学成分以外的其余化学成分风化前后的变化率受风化的影响很小。此时我们设定一个阈值来作为判定所有化学成分数据的均值是否符合该成分是属于风化化学成分的依据，然后挑选出风化化学成分，然后根据风化点数据，建立一个预测模型来预测其风化化学成分含量。

2.2 问题二的分析

问题二需要根据附件的数据对高钾玻璃、铅钡玻璃的分类规律进行分析，然后选择合适的化学成分，对这两个类型的玻璃进行亚类划分，这里我们选择 k-means 聚类分析方法对不同类型的玻璃进行亚类划分，亚类划分完成后，观察每个类中的所有化学成分含量，然后输出每个类所对应的化学成分的范围，来进行辅助划分，采用轮廓系数评估法来对聚类结果的合理性和敏感性进行分析。

2.3 问题三的分析

问题三需要将表单 12 数据合并汇总，然后以玻璃文物饰品的化学成分及表面风化列数据作为自变量，玻璃类型为因变量建立逻辑回归模型，并用合并的数据来训练模型，然后用表单 3 的数据来预测玻璃类型。敏感性分析则用问题的 K-means 聚类的高钾玻璃类簇和铅钡玻璃类簇中心，来计算与表单 3 样本的距离判定表单 3 中文物的玻璃类型，根据判定结果与预测结果的对比，分析分类结果的敏感性。

2.4 问题四的分析

问题四分为两小问，第一小问需要分析不同类别玻璃文物样品的化学成分之间的关联关系，也就是相关性分析。第二小问不用考虑是否风化，则这里使用表单 12 合并后数据中的类型和所有化学成分的数据做关联分析。所以这里采用了方差分析算法来分析不同类别玻璃、化学成分之间的相关性和差异性。

三、 模型假设

1. 玻璃文物的表层风化与其玻璃类型、纹饰和颜色都是相关的。
2. 不同的玻璃类型、纹饰、颜色的玻璃文物的风化化学成分会有所区别。
3. 风化前后的化学成分变化不大的不是风化化学成分。
4. 根据中国古代主要的玻璃饰品类型划分，不同类型的玻璃的亚类划分在 2 到 5 个亚类之间。
5. 不同类型的玻璃类型对于化学成分的影响是显著的。

四、 符号说明及名词定义

表 1　符号说明

符号	说明	单位	备注
B_i	两个变量是否具有相关关系	——	$B_i = 0, 1$
p_i	两个变量经过卡方检验后检验的 P 值	——	i 表示每组变量
α	置信水平	——	$\alpha = 0.5$
A_i	表示表单 1 中玻璃饰品的类别特征名称	——	类型、纹饰、颜色、表面风化
$\chi^2_{pearson}$	卡方皮尔逊统计量	——	用计算样本统计量

五、 模型建立与求解

5.1 问题一的求解

5.1.1 表层风化—相关特征评估模型

为了分析文物的表面风化与其玻璃类型、纹饰和颜色的关系，而且表面风华是一个分类变量，这里搭建一个以卡方检验思想为基础的表层风化—相关特征评估模型，来检测表面风化与其他三个特征变量的关系。

卡方检验就是统计样本的真实值与预测值之间的偏离程度，真实值与预测值之间的偏离程度就决定卡方值的大小，如果 χ^2 值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，χ^2 值就为 0，说明预测值和真实值重合，这是最理想的结果。当然，实际上，往往我们得到的结果几乎都不会达到这种程度，那么需要通过数据的可视化更为直观地去观察变量之间的存在关系。

建立模型前，现在对附件表单一的数据进行数据可视化，画出四个类别特征数量占比图，如图一所示：

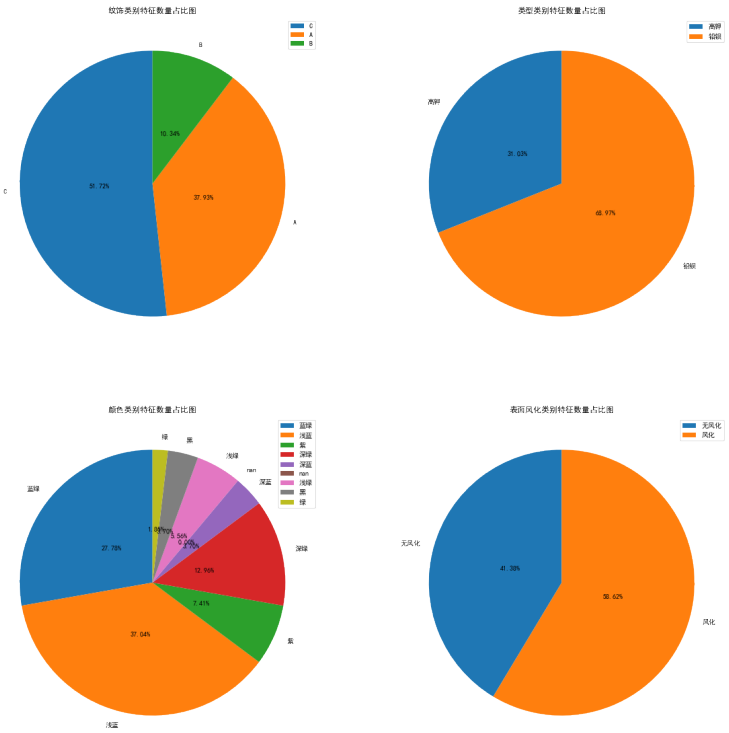

图 1　表单 1 中所有类别特征的数量占比图

由图一发现，表单一中的数据中发现，古代玻璃文物制品的纹饰、类型、表面风化程度的类别特征值比较常规，而玻璃文物制品的颜色很多种，说明我国古代文物玻璃饰品的类型、纹饰分别主要由两种到三种的形式存在，玻璃文物制品的颜色多样化说明虽然我国古代的玻璃制造技术比较固定，但是可以从颜色的多样性让玻璃饰品给人们带来更多的视觉感受，也充分体现古代玻璃饰品的审美艺术气息。玻璃的成分中二氧化硅、碱（苏打、钾）和氧化钙（石灰）的比例，添加剂，如金属氧化物和乳化剂在决定材料的耐久性方面起着作用 [1]。

有了以上的分析，认为分析表面风化与其玻璃类型、纹饰和颜色的关系是非常有必要的。所以要构建表层风化—相关特征评估模型：

$$B_i = \begin{cases} 0, & p_i \leq \alpha \\ 1, & p_i > \alpha \end{cases} \tag{1}$$

要计算 p_i 的值还需要通过，计算 χ^2 值，根据 $pearson$ 卡方统计量，这里 χ^2 值的公式为：

$$\chi^2_{pearson} = N\left(\sum\nolimits_R^i \sum\nolimits_c^j \frac{A_{ij}^2}{\sum_C^{c=i} A_{ic} \sum_R^{r=1} A_{rj}}\right) \chi^2((R-1)(C-1)) \tag{2}$$

不妨假设：

$$H_0 : A_0 \text{ 与 } A_i \text{ 无关} \quad H_1 : A_0 \text{ 与 } A_i \text{ 有关}$$

A_i 表示三个指标（玻璃类型、纹饰、颜色）中的一个指标 (i=1,2,3)

这里经过 SPSS 的 χ^2 检验分析后得出关于 P 值的结果（表一）：

表 2　三组变量的 χ^2 检验 P 值结果表

	玻璃类型	纹饰	颜色
表面风化	0.009	0.084	0.307

通过 χ^2 检验得出的 P-value，在置信水平 α=0.05 下，由表 2 可以看出，在统计学上，表面风化与玻璃类型有关，与纹饰和颜色无关。但是在化学角度，表面颜色是一个重要的特征，因为某些类型的退化似乎仅限于无色玻璃物体 [1]。那么我们还需要根据相关系数 corr 对它们之间的关系进行进一步的分析，画出相关系数热力图（图 2）：

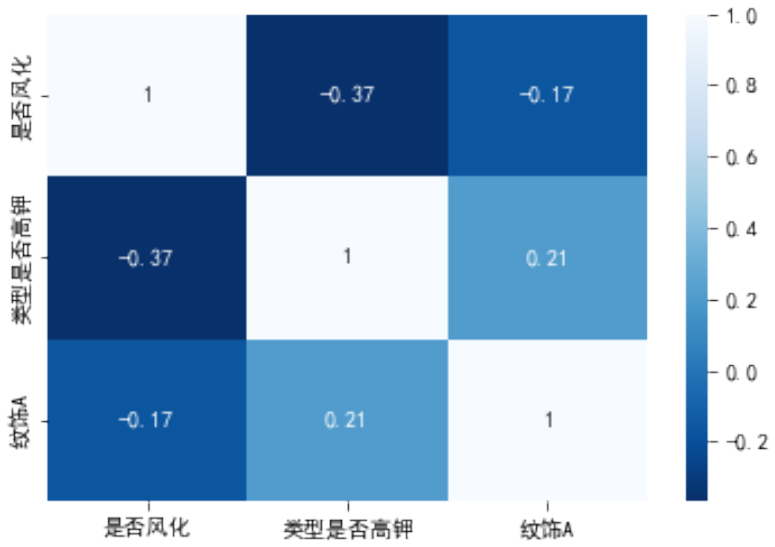

图 2　风化与纹饰，颜色，类型四个属性的相关性热力图

结合热力图和 P 值结果表综合分析，得出风化与玻璃类型、颜色、纹饰呈现负相关的关系的结论，也就是说玻璃类型、颜色、纹饰是对表面风化是有影响的。也就是说明玻璃文物的表层风化与其玻璃类型、纹饰和颜色都是相关的，这里证明模型假设 1 是成立的。接下来找出风化的化学成分，需要将表单 1 和表单 2 的数据进行合并，合并后的数据表为表 3（部分）：

表 3　表单 1 和表单 2 的合并数据（部分）

是否风化	类型是否高钾	纹饰 A	颜色编号	二氧化硅	氧化钠	氧化钾	……
0	1	0.0	0.0	71.027559	0.0	10.234607	……
1	0	1.0	1.0	36.319952	0.0	1.051156	……
0	1	1.0	0.0	87.050000	0.0	5.190000	……
0	1	1.0	0.0	62.408981	0.0	12.510113	……
0	1	1.0	0.0	68.582136	0.0	10.066625	……
……	……	……	……	……	……	……	……

依据合并后的数据，画出每个化学成分与是否风化的相关性热力图，如图 3 所示。由于化学成分有 14 个之多，所以只需取 5 个成分（除了二氧化硅）即可，把相关系数的

绝对值的大于 0.21 的化学成分提取出来（表 4）。从表中可以看出氧化钾、氧化钙、氧化铁、氧化铅、五氧化二磷与是否风化有较大的相关关系。

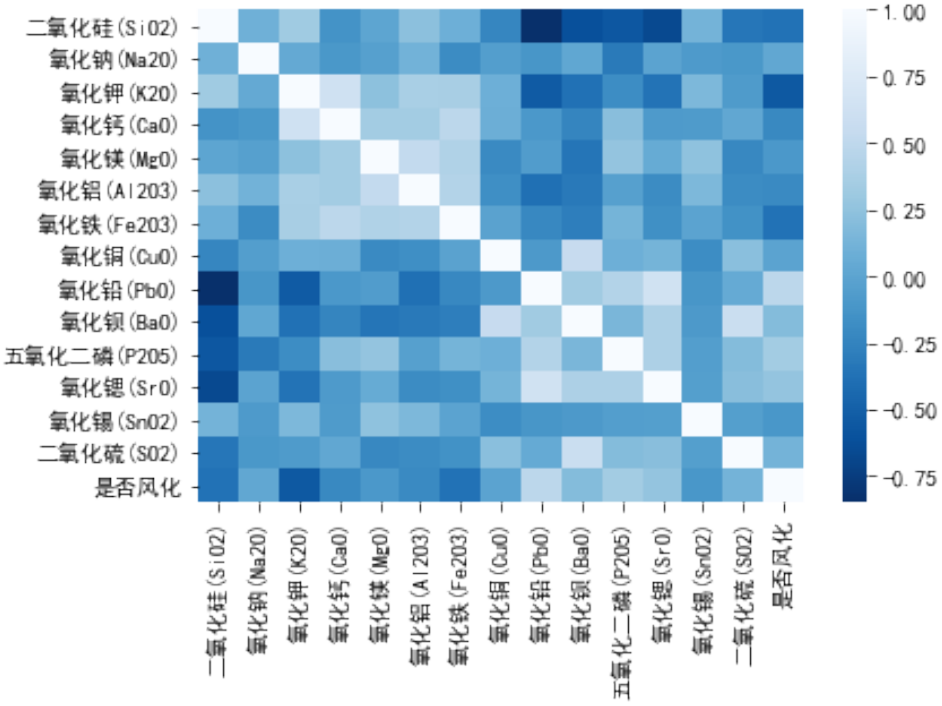

图 3　每个化学成分与是否风化的相关性热力图

表 4　6 个化学成分与是否风化的相关系数表

序号	化学成分	相关系数
1	二氧化硅 (SiO2)	-0.373052
2	氧化钾 (K2O)	-0.554855
3	氧化钙 (CaO)	-0.214721
4	氧化铁 (Fe2O3)	-0.371009
5	氧化铅 (PbO)	0.480159
6	五氧化二磷 (P2O5)	0.335856

5.1.2 支持向量回归（SVR）风化化学成分预测模型

由于要根据风化点的检测数据预测其风花其风化前的化学成分含量，所以要以表面风化列的所有数据划分为两类数据，分别为未风化样品的数据和已风化样品的数据，然

后根据所有未风化样品的化学成分指标数据，逐次对已风化样品数据中的单个指标建立预测算法模型，以单个未风化样品数据的所有化学成分数据作为输入，通过模型预测输出对应的风化化学成分值。

支持向量回归（SVR）预测模型的原理：数据在间隔带内则不计算损失，当且仅当预测值 f（x）与真实值 y 之间的差距的绝对值大于 ϵ 才计算损失，通过最大化间隔带的宽度与最小化总损失来优化模型。SVR 在线性函数两侧制造了一个“间隔带”，间距为 ϵ(也叫容忍偏差，是一个由人工设定的经验值)，对所有落入到间隔带内的样本不计算损失，也就是只有支持向量才会对其函数模型产生影响，最后通过最小化总损失和最大化间隔来得出优化后的模型 [3]。

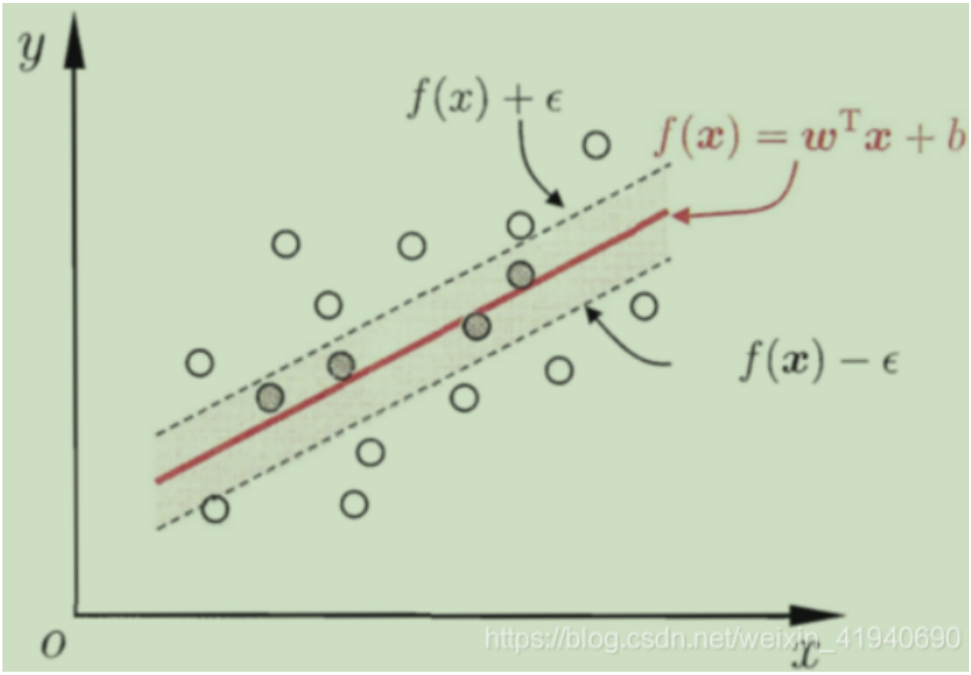

图 4　支持向量回归模型原理图 [3]

现在一个支持向量回归预测模型，由于数据是连续型和离散型混合的数据，所以我们要加入一个线性核函数，表达式为：

$$\kappa\left(x_i, x_j\right)=x_i^T x_j=\phi\left(x_i\right)^T \phi\left(x_j\right) \tag{3}$$

SVR 模型的表达式为：

$$f(x)=\omega^T \phi(x)+b \tag{4}$$

SVR 映入线性核函数之后，表达式可写为：

$$f(x)=\sum_{m}^{i=1} \alpha_i y_i \kappa\left(x, x_i\right)+b \tag{5}$$

$$\begin{cases} min\dfrac{1}{2}||\omega||^2 \\ s.t.|y_i-(\omega x_i+b)|\le\epsilon,\forall i \end{cases} \tag{6}$$

运行 [附录 1] 因为每一个成分都会有一个模型，所以会得到 15 个模型的模型系数，如图 5。然后将风化后的数据放入模型预测风化前的数据即可。

```
[array([[-0.10003783,  0.03370803,  0.08595787, -0.27806093, -0.08855979,
         -0.04872763,  0.04639018,  0.18773133,  0.07178078,  0.0342383 ,
          0.02646883,  0.06424005, -0.07298251, -0.05319061, -0.00899389]]),
 array([[-0.12549777,  0.28498106,  1.16247664, -0.19654615,  1.08360716,
         -0.48547097, -0.33157822, -1.16540786, -1.02910446,  0.24946947,
          0.3389298 ,  0.09145161, -0.63732483, -0.25785807,  0.89237481]]),
 array([[-0.04140437, -0.04373739, -0.01774682,  0.06009557, -0.0798704 ,
         -0.01127501,  0.0185587 ,  0.00342195,  0.06281122, -0.03518824,
         -0.08942415,  0.0245912 , -0.11403835,  0.17810805,  0.04369368]]),
 array([[-3.81072517e-02, -8.91563048e-02,  3.89055379e-01,
         -6.42502551e-01,  1.26877051e+00,  9.77183340e-04,
         -2.28442741e-01,  9.12455782e-02,  2.42179552e-01,
         -5.42664727e-02, -2.84597906e-01, -3.47915842e-01,
         -2.70520058e-01, -2.12062006e-01,  1.37235683e-01]]),
 array([[ 0.4225032 , -0.10321288, -0.06835231,  0.52740707,  0.14327015,
          0.66349227, -0.30262511, -0.35218282, -0.34225147, -0.13160023,
         -0.08447881, -0.06985748,  0.2759793 , -0.22442361,  0.06883593]]),
 array([[ 0.01107414, -0.0386639 ,  0.00138672, -0.01911494, -0.19774624,
         -0.13252552, -0.01015756, -0.02035868, -0.01629817, -0.03679643,
         -0.04773498,  0.00193101,  0.56517369, -0.0220576 , -0.0270374 ]]),
 array([[ 0.4233613 , -0.14660604, -0.15946427,  0.66869649, -0.65203001,
          0.06708509, -0.13274719, -0.00727721, -0.02732784, -0.09981014,
         -0.18639291, -0.18937529,  0.10929687,  0.61084628,  0.14510615]]),
 array([[ 0.04023108,  0.0063663 ,  0.00390538,  0.51289482,  0.1982339 ,
         -0.04821079,  0.05259178, -0.28487788,  0.09424652,  0.02951923,
         -0.03506813, -0.02743685, -0.28866428, -0.27619048,  0.06269048]]),
 array([[ 0.25391812,  0.0551183 , -0.04718351,  0.03535112,  0.04112291,
         -0.09431535,  0.03909744, -0.46087617, -0.15113178,  0.03189861,
          0.01485776,  0.1490141 ,  0.30219581, -0.21493982,  0.29979058]]),
 array([[-0.45357708, -0.04795742,  0.40209555,  0.18896469,  0.04674362,
         -0.41153071,  0.28083277,  1.49269985,  0.1466754 , -0.08232427,
         -0.27739584,  0.08567563, -1.12757179, -0.54874464, -0.14816284]]),
 array([[-0.26220499,  0.19612661,  0.0273308 , -0.04798964, -0.09133551,
         -1.01547543,  0.0897258 ,  0.69943948,  0.27348648,  0.22570866,
          0.20780557,  0.05662597, -0.14420426, -0.29381673, -0.18342781]]),
 array([[ 0.15696397, -0.07715957, -0.07273873,  0.30848729, -0.03455381,
         -0.03811023, -0.03620103, -0.13219822,  0.17333083, -0.06964187,
         -0.09162846, -0.07988712,  0.15628329,  0.02215967, -0.02814206]]),
 array([[ 0.08379317,  0.00624326,  0.00675975, -0.04539193,  0.00746996,
         -0.08783084,  0.0541857 ,  0.01982306,  0.05404269,  0.00895599,
         -0.00150234,  0.01023625,  0.00254152, -0.00132765, -0.03420543]]),
 array([[ 0.04383759, -0.02001822, -0.02302146,  0.08199926,  0.00317708,
          0.01392899, -0.02292574, -0.03777387, -0.04777048, -0.01999207,
         -0.01735781, -0.03013789,  0.17661629, -0.09314448,  0.03642041]]),
 array([[ 0.00390487,  0.00667384,  0.01053631, -0.0343031 , -0.03494679,
         -0.02329113,  0.02106375,  0.04195968,  0.01028114,  0.00753921,
          0.00768216,  0.01882948,  0.00605573, -0.04045208,  0.00237179]])]
```

图 5　15 个模型的模型系数结果图

5.2 问题二的求解

问题二要求依据数据来分析高钾玻璃、铅钡玻璃的分类规律，我们可以用 K-means 聚类分析算法得出高钾玻璃和铅钡玻璃的分类规律，就是根据聚类分析的结果中的两个类簇的中心的具体位置来确定高钾玻璃与铅钡玻璃是如何划分，然后根据聚类结果的准确率来判断 K-means 算法是否划分合理。如果划分合理，就沿用这个算法分别对高钾玻璃、铅钡玻璃进行亚类划分，然后进行聚类结果评估。流程图（图 6）如下：

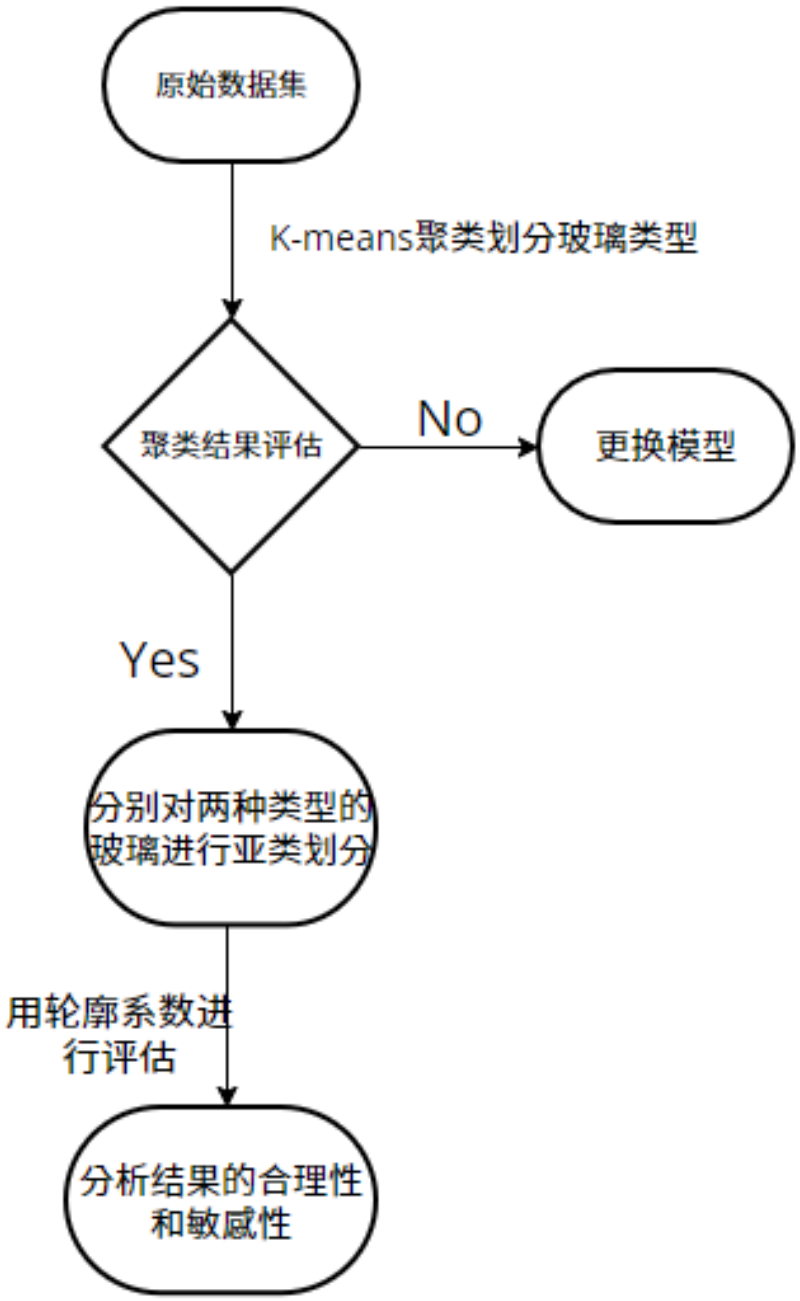

图 6　分类过程图

5.2.1 K-Means 算法原理

给定样本集 $D = x_1, x_2, ..., x_m$,k-means 算法针对聚类所得簇划分 $C = C_1, C_2, ..., C_k$ 最小化平方误差

$$E = \sum_{k}^{i=1} ||x - \mu_i||_2^2 \tag{7}$$

其中 $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ 是簇 C_i 的均值向量。式（7）在一定程度上刻画了簇内样本围绕类簇均值向量的紧密程度,E 值越小则簇内样本相似度越高 [4]。

5.2.2 K-Means 算法求解

通过 K-means 算法，求解得出对于高钾玻璃和铅钡玻璃的分类结果的准确度为 73.13%, 那么可以清楚地知道算法计算得出的类簇中心点（图 7）作为聚类标准是可行的。

	二氧化硅 (SiO2)	氧化钠 (Na2O)	氧化钾 (K2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al2O3)	氧化铁 (Fe2O3)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷(P2O5)	氧化锶 (SrO)	氧化锡 (SnO2)	二氧化硫 (SO2)
铅钡玻璃中心	28.064706	0.327821	0.149867	2.724723	0.585076	2.697051	0.685492	2.319275	43.792930	12.189019	4.840441	0.424821	0.043458	1.155323
高钾玻璃中心	68.899203	1.214811	3.375655	2.496377	0.801860	5.366297	1.031022	1.725905	9.270381	4.422873	1.014074	0.134243	0.110576	0.136723

图 7　算法求得类簇中心点位置结果图

接下来，继续使用这个算法，分别对高钾玻璃和铅钡玻璃的进行亚类划分，由于我们对于亚类的划分的类簇个数是位置的，那么我们定义 k 个样本作为类簇中心进行聚类，这里 k 的区间范围是 [1,10]，并画出类簇数—损失图（图 8 和图 9）。从这两图可以观察出，铅钡玻璃亚类划分类簇个数为 2 时，损失数出现拐点然后摆尾；而高钾玻璃亚类划分类簇个数为 3 时，损失数出现拐点然后摆尾。这里有理由认为铅钡玻璃亚类划分个数应为 2，高钾玻璃亚类划分个数应为 3, 这里给出高钾玻璃和铅钡玻璃亚类划分类簇中心的位置范围（图 10 和图 11）。

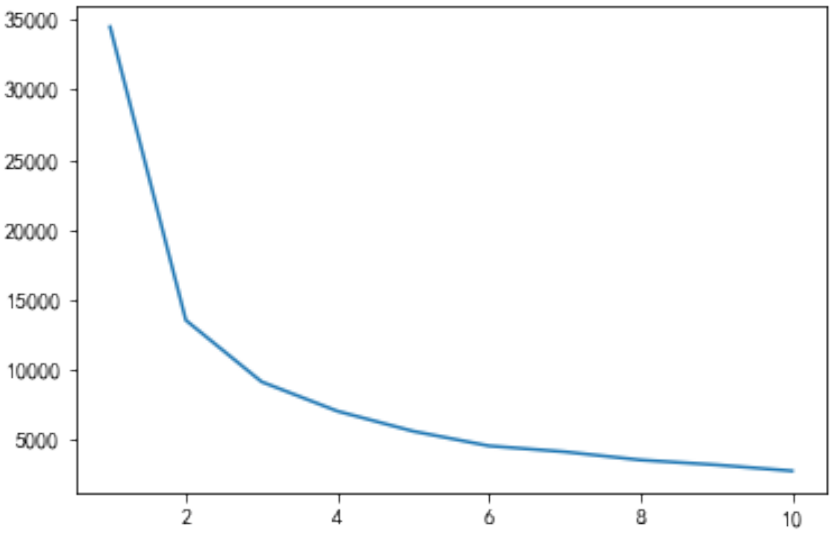

图 8　铅钡玻璃亚类类簇数—损失图

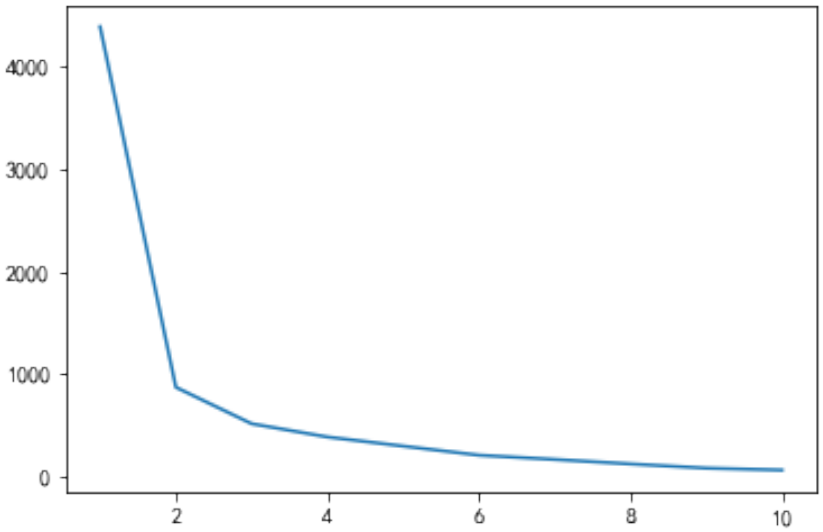

图 9　高钾玻璃亚类类簇数—损失图

	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
高钾玻璃亚类1	90.286077	5.551115e-17	2.015983	1.336693	0.444233	2.789590	0.444556	1.502207	0.140558	0.222177	0.540290	0.007998	2.696372e-01	-1.387779e-17
高钾玻璃亚类2	64.909842	9.402528e-01	11.033781	6.499354	1.157651	7.487044	2.357745	2.877194	0.414489	0.585427	1.548567	0.048506	2.775558e-17	1.401457e-01

图 10　高钾玻璃亚类类簇中心位置范围

	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
铅钡玻璃亚类1	27.483508	0.225940	0.182967	3.003896	0.765272	3.048744	8.933314e-01	1.253957	49.163431	8.374907	5.113888	0.428922	6.123596e-02	3.330669e-16
铅钡玻璃亚类2	58.716330	1.926795	0.194080	1.161199	0.714621	5.164300	6.425247e-01	1.213632	20.349394	8.260542	1.181237	0.227025	7.400081e-02	1.743206e-01
铅钡玻璃亚类3	19.322301	0.000000	0.103432	2.223447	0.124052	1.463813	1.110223e-16	6.923689	29.564396	28.682542	5.074852	0.548308	6.938894e-18	5.969170e+00

图 11　铅钡玻璃亚类类簇中心位置范围

然后根据两种类型玻璃的亚类类簇中心位置作为聚类中心，进行 K-means 聚类，得出结果如表 5：

表 5　划分结果表（部分）

是否风化	类型是否高钾	纹饰 A	颜色编号	二氧化硅	氧化钠	……	亚类
0	1	0.0	0.0	71.027559	0.0	……	高钾玻璃亚类 2
1	0	1.0	1.0	36.319952	0.0	……	铅钡玻璃亚类 1
0	1	1.0	0.0	87.050000	0.0	……	高钾玻璃亚类 1
0	1	1.0	0.0	62.408981	0.0	……	高钾玻璃亚类 2
0	1	1.0	0.0	68.582136	0.0	……	高钾玻璃亚类 2
……	……	……	……	……	……	……	……

5.2.3 K-Means 聚类结果的合理性和敏感性

由于 K-Means 聚类属于无监督学习，而且这里我们是不知道具体的每种类型玻璃的亚类划分情况，所以这里用轮廓系数对聚类结果进行评价。

I 轮廓系数的定义

轮廓系数是根据内聚度和分离度这两种因素来评价聚类结果的好坏，可以用于相同的样品检测数据的基础上来评价聚类结果如何被 K-means 算法不同的运行方式所影响。

II 轮廓系数的评估原理以及相关公式

已知轮廓系数给定大小区间为（-1，1），所以当轮廓系数越接近 1 时，说明聚类效果越好；反之则越差。这里给定：

- c_i: 任意一个样本与其所在的类簇范围内的其他样本的平均距离（簇内不相似度）
- d_i: 任意一个样本与其他类簇范围内的样本的平均距离（簇间相似度）

则对于任意一个样本的轮廓系数为式（8），根据簇内不相似度和簇间相似度的大小关系把式子修改为式（9），而整体的轮廓系数等于每个样本的轮廓系数之和的平均值为式（10）：

$$s_i = \frac{d_i - c_i}{max\,(c_i, d_i)} \tag{8}$$

$$S_i = \begin{cases} 1 - \frac{c_i}{d_i}, & c_i < d_i \\ 0, & c_i = d_i \\ \frac{c_i}{d_i} - 1, & c_i > d_i \end{cases} \tag{9}$$

$$SC = \frac{\sum_i^N S_(i)}{N} \tag{10}$$

其中 N 为整体样本的个数，S_i 是每个样本的轮廓系数，SC 表示整体的轮廓系数。

根据 **[附录 2]** 的算法程序，计算出高钾玻璃亚类划分结果与铅钡玻璃亚类划分结果的两个整体轮廓系数（表 6）：

表 6　最优亚类划分结果的轮廓系数表

	高钾玻璃亚类划分	铅钡玻璃亚类划分
轮廓系数	0.5674	0.5184
亚类数目	2	3

根据表 6 得出结论，当亚类数目是 2 的时候，高钾玻璃的聚类结果是最好的，当亚类数目是 3 的时候，高钾玻璃的聚类结果是最好的。说明了 K-means 聚类算法来进行亚类划分是合理的，而且敏感性良好。

5.3 问题三的求解

根据问题三要求，需要将表单 1 和表单 2 数据进行合并，然后以玻璃文物饰品的所有化学成分以及表面是否风化作为模型的输入变量，玻璃的类型作为模型的输出变量。用合并后的数据训练逻辑回归模型，然后用表单 3 的玻璃文物饰品的所有化学成分以及表面是否风化作为输入变量放入逻辑回归模型进行预测，然后输出玻璃的类型的值。然后以问题 2 中求出的铅钡玻璃和高钾玻璃的类簇中心位置为基准，根据表单 3 每个样本的数据计算与类簇中心位置距离，比较样本与两个类簇中心的距离，然后根据比较结果的来判定其所属类型，最终根据逻辑回归模型的结果与样本-类簇中心距离的匹配度，分析逻辑回归模型的分类结果的敏感性。

5.3.1 逻辑回归模型的训练与预测

如前文所述，我们这里用表单 12 的合并数据进行逻辑回归模型训练，然后进行模型预测，由于表单 3 数据的样本量非常小，为了防止模型过拟合，所以这里是用了基础逻辑回归模型。运行程序 **[附录 3]**，得出预测结果整理后如下表（表 7）：

表 7　逻辑回归模型预测结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
预测结果	高钾	铅钡	铅钡	铅钡	铅钡	高钾	铅钡	高钾

5.3.2 计算表单 3 的样本与类簇中心的距离并判断所属类型

这里的类簇中心的位置如图 7 所示，根据距离公式（11），计算计算表单 3 的样本与类簇中心的距离，然后判断所属类型，运行程序 [附录 4], 得出结果（表 8）：

$$D^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{11}$$

其中，D^2 为某个样本与类簇中心点的距离的平方和，x_i 为样本的第 i 个数据，y_i 为类簇中心点的第 i 个数据

表 8　根据距离判断结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
判断结果	高钾	铅钡	铅钡	铅钡	高钾	高钾	高钾	高钾

5.3.3 分析逻辑回归模型预测结果的敏感性

将表 7 和表 8 的数据进行合并，根据匹配结果分析分类结果的敏感性，然后记 1 为匹配成功，0 为匹配失败，如表 9 所示：

表 9　匹配结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
预测结果	高钾	铅钡	铅钡	铅钡	铅钡	高钾	高钾	铅钡
判断结果	高钾	铅钡	铅钡	铅钡	高钾	高钾	高钾	高钾
匹配结果	1	1	1	1	0	1	1	0

从表 9 的匹配结果看出，匹配成功率为 75%, 等价于验证准确度为 75%，结合判断结果和预测结果，可以认为逻辑回归模型的分类预测结果的敏感性是良好的。

5.4 问题四的求解

题目四要求分析不同类型玻璃的化学成分之间的关联关系，不用考虑是否风化，则这里使用表单 12 合并后数据中的类型和所有化学成分的数据做关联分析。所以这里我采用了方差分析算法来分析不同类别玻璃、化学成分之间的关系。运行程序 **[附录 4]**, 得到以下结果, 如表 10 所示：

表 10　方差分析结果

	df	sum_sq	mean_sq	F	PR(>F)
C(类型是否高钾)	14.0	172415.149547	12315.367825	162.425542	8.414966e-245
Residual	990.0	75063.404636	75.821621	NaN	NaN

我们可以看到不同玻璃类型，对于化学成分的影响还是十分显著的，p 值显然小于

0.05

然后进行差异性分析，运行程序 **[附录 4]**，得到相关结果，展示部分结果 (图 12)：

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
二氧化硅(SiO2)	二氧化硫(SO2)	-49.3976	0.001	-54.5128	-44.2824	True
二氧化硅(SiO2)	五氧化二磷(P2O5)	-47.2211	0.001	-52.3363	-42.106	True
二氧化硅(SiO2)	氧化钙(CaO)	-47.4036	0.001	-52.5188	-42.2884	True
二氧化硅(SiO2)	氧化钠(Na2O)	-49.2012	0.001	-54.3164	-44.086	True
二氧化硅(SiO2)	氧化钡(BaO)	-41.9895	0.001	-47.1047	-36.8743	True
二氧化硅(SiO2)	氧化钾(K2O)	-48.1225	0.001	-53.2377	-43.0073	True
二氧化硅(SiO2)	氧化铁(Fe2O3)	-49.1345	0.001	-54.2497	-44.0193	True
二氧化硅(SiO2)	氧化铅(PbO)	-24.7621	0.001	-29.8773	-19.6469	True
二氧化硅(SiO2)	氧化铜(CuO)	-48.0052	0.001	-53.1204	-42.89	True
二氧化硅(SiO2)	氧化铝(Al2O3)	-45.8744	0.001	-50.9895	-40.7592	True
二氧化硅(SiO2)	氧化锡(SnO2)	-49.9261	0.001	-55.0413	-44.8109	True
二氧化硅(SiO2)	氧化锶(SrO)	-49.7369	0.001	-54.8521	-44.6218	True
二氧化硅(SiO2)	氧化镁(MgO)	-49.3041	0.001	-54.4193	-44.1889	True
二氧化硅(SiO2)	类型是否高钾	-49.737	0.001	-54.8522	-44.6218	True

图 12　不同类别之间的化学成分关联关系的差异性（部分）

从分析相关性和观察运行结果图而知，不同玻璃类型的化学成分相互影响，尤其是二氧化硫对二氧化硅的影响最大，二氧化硫与水结合生成亚硫酸会腐蚀破坏玻璃制品的表层结构，从而使其风化的程度更严重。通过差异性分析，我们还知道不同玻璃类型的化学成分是有差异的，助熔剂、添加剂的成分不同促使其化学成分的关联关系，所以玻璃的制造工艺决定玻璃制品的化学成分含量具备差异性。

六、 模型分析

6.1 模型的优点

1. 从数据挖掘的方向入手，运用卡方检验进行统计分析，可以从数据中找到隐含规律;
2. 在数据的数量极其少情况下，结合逻辑回归模型与 k-means 算法，能够较好判别文物的类型;
3. 支持向量回归（SVR）预测模型有效地预测出玻璃文物风化前的化学成分数据;
4. 在未知亚类的类别数目的情况下，使用无监督学习算法 K-means，以及轮廓系数对于聚类结果的评估，能够较好地解释亚类的划分结果。

6.2 模型的缺点

1. 仅适用于数据量较小的数据集，数据量较大的话模型可能会失去作用，模型的准确度也会有所下降。

2. 模型并未添加风险因素，古代文物玻璃制品中不同的埋藏条件决定着不同类型的风化作用，也会影响化学成分的变化。

七、 参考文献与引用

参考文献

[1] Gueli AM, Pasquale S, Tanasi D, et al. Weathering and deterioration of archeological glasses from late Roman Sicily. Int J Appl Glass Sci. 2019;00:1–11.

[2] 姜启源, 谢金星, 叶俊. 数学模型 [M]. 北京: 高等教育出版社, 2003.

[3] 支持向量回归（SVR）的详细介绍以及推导算法https://blog.csdn.net/weixin_41940690/article/details/106639347

[4] 机器学习/周志华著 •-北京：清华大学出版社，2016

[5] Peter J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Volume 20, 1987

附录 A　附录

1.1　支撑材料清单

- 玻璃文物的表面风化与纹饰的交叉表和卡方检验.docx
- ph1.png
- ph2.png
- ph3.png
- ph4.png
- ph5.png
- ph6.png
- ph7.png
- ph8.png
- ph9.png
- ph10.png
- ph11.png
- ph12.png
- ph12.png
- k-means.py
- SVR 风化化学成分预测模型.py
- 方差分析.py
- 逻辑回归模型.py
- 表单 12 合并汇总.xlsx

1.2　附录 1:SVR 风化化学成分预测模型.py

```
#导入相关库
import pandas as pd
import numpy as np
from sklearn import svm

#导入合并后的数据集
data = pd.read_excel(r"表单12数据合并汇总.xlsx")

#划分两类数据集，风化文物饰品的数据和未风化文物饰品数据
fenghua = data[data['表面风化']==1]
weifenghua = data[data['表面风化']==0]
cols = ['类型是否高钾']+list(df.columns[6:])
fenghua = fenghua[cols]
weifenghua = weifenghua[cols]
```

```
#支持向量回归SVR模型
svr = svm.SVR(kernel='linear')

#模型训练
cf = [0 for i in range(len(cols))]
for i in range(len(cols)):
for j in range(5):
index = list(fenghua.index)
np.random.shuffle(index)
svr.fit(fenghua.loc[index[:25]].values, weifenghua.values[:, i])
cf[i] += svr.coef_/5
for i in range(15):
print(fenghua.columns[i], '的模型系数',cf[i])
```

1.3 附录 2:k-means.py

```
#导入相关库
import pandas as pd
import numpy as np
from sklearn.cluster import k_means

#导入合并后的数据集
data = pd.read_excel(r"表单12数据合并汇总.xlsx")

cols = df.columns[6:]
KMS = k_means(df[cols], 2)
KMS

accuracy = (KMS[1]==df['类型是否高钾']).sum()/len(df)*100
accuracy

#求出高钾玻璃和铅钡玻璃类型划分中心
category_1 = pd.DataFrame(KMS[0],columns = cols, index = ['铅钡玻璃中心','高钾玻璃中心'])
category_1

#划分高钾玻璃亚类
gaojia = df[df['类型是否高钾']==1]
category_2 = k_means(gaojia[gaojia.columns[6:]], 2)
category_2 = pd.DataFrame(data=category_2[0], columns=[gaojia.columns[6:]],
    index=['高钾玻璃亚类1','高钾玻璃亚类2'])
category_2

#划分铅钡玻璃亚类
qianbei = df[df['类型是否高钾']==0]
category_3 = k_means(qianbei[qianbei.columns[6:]], 3)
```

```
category_3 = pd.DataFrame(data=category_3[0], columns=[qianbei.columns[6:]],
    index=['铅钡玻璃亚类1','铅钡玻璃亚类2','铅钡玻璃亚类3'])
category_3

#计算整体轮廓系数
from sklearn.metrics import silhouette_score
Silhouette_Coefficient1 = silhouette_score(gaojia[gaojia.columns[6:]], category_2[1])
Silhouette_Coefficient1

Silhouette_Coefficient2 = silhouette_score(qianbei[qianbei.columns[6:]], category_3[1])
Silhouette_Coefficient2
```

1.4 附录 3: 逻辑回归模型.py

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression

data = pd.read_excel('表单12数据合并汇总.xlsx', index_col=0)
#附件表单3的数据处理
data1 = pd.read_excel('附件.xlsx', index_col=2)
data1=data1.fillna(0)
cols1 = data1.columns[2:]
data1['成分比例累加和']=np.sum(data1[cols1], axis=1)

cols = data.columns[6:]
x_tr = data[['是否风化']+list(cols)].values
y_tr = data['类型是否高钾'].values

pre = data1[data1.columns[1:-1]].values
LogisticRegression.fit(x_tr, y_tr)
pre_y = LogisticRegression.predict(pre)
pre_y
```

1.5 附录 4: 方差分析.py

```
import pandas as pd
import numpy as np

data = pd.read_excel(r"表单12数据合并汇总.xlsx", index_col=0)
col = ['类型是否高钾','二氧化硅(SiO2)',
'氧化钠(Na2O)', '氧化钾(K2O)', '氧化钙(CaO)', '氧化镁(MgO)', '氧化铝(Al2O3)',
'氧化铁(Fe2O3)', '氧化铜(CuO)', '氧化铅(PbO)', '氧化钡(BaO)', '五氧化二磷(P2O5)',
'氧化锶(SrO)', '氧化锡(SnO2)', '二氧化硫(SO2)']
```

```
data1 = data[col]

# 方差分析
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.stats.multicomp import pairwise_tukeyhsd

data_melt = data1.melt()
data_melt.columns = ['类型是否高钾', '化学成分']
model = ols('化学成分 ~C(类型是否高钾)', data = data_melt).fit()
anova_table = anova_lm(model, type = 2)
pd.DataFrame(anova_table)

# 进行事后比较分析
print(pairwise_tukeyhsd(data_melt['化学成分'], data_melt['类型是否高钾']))
```