各位老师大家好，我是90组的周洋，这两位是我的队友胡佳伟和吴金波。我们小组选作的是C题。C题一共有四道问题，共分为五个部分。这是整道题的思路，接下来简单介绍一下各部分的操作内容。

首先进行数据预处理，表单1上在颜色列上有部分空值，就采用同属性的样本颜色的众数进行填充。表单2上根据题目所定义的有效数据，将化学成分比例累计和不在85%~105%这个区间的样本进行剔除。

第一题有三个小问。

第一问我们首先是通过SPSS作交叉表，可以看出哪类玻璃类型、纹饰和颜色容易风化。我们发现纹饰B和纹饰C，颜色为黑色、蓝绿色、浅蓝色、深绿色，玻璃类型为铅钡类的玻璃制品更容易风风化。

然后因为这里我们发现表面风化和其它三个自变量都属于定类变量，所以我们这里采用了卡方检验，去看这三者是否与风化有无关系，后面发现，只有玻璃类型与表面风化是存在显著差异的，纹饰和颜色对表面风化的关系并不显著。

第二问我们根据附件数据分别筛选出风化高钾玻璃、无风化高钾玻璃、风化铅钡玻璃、无风化铅钡玻璃这四张表，然后计算它们的统计变量从而分析它们的统计规律，通过绘制箱型图分析它们的分布特征，通过绘制正态分布图分。

第三问，我们以化学成分为自变量，表面风化为因变量来建立逻辑回归模型，根据模型的变量重要性筛选出主要成分和次要成分，其中主要成分是我们主要预测的目标，而次要成分，我们根据风化和无风化各类化学成分均值的比值来进行预测。而主要成分，我们是利用无风化的次要变量来预测其主要变量来训练回归模型，然后通过将风化样本的次要变量输入到模型中，从而预测出风化样本的主要变量，并将被预测的主要变量加入到因变量中，从而逐步完成回归，直到最后一个主要变量加入到因变量中。

第二题也有三个小问

第一问分析分类规律，我们主要是通过建立随机森林模型，然后通过随机森林的每个分类器可以看出它的分类依据，然后根据随机森林模型的变量重要性进行排序，可以从中分析这几个变量对分类模型的作用比较大。

第二问就是个聚类模型，对这两类玻璃数据分别进行kmeans聚类，根据手肘图可以看出来聚类模型分为3类最为合适。然后我们观察聚类模型中各个数据的分布情况，可以发现聚类结果是很明显的，各类数据基本上都集中在一起，但高钾玻璃的分布有些离散，这也许是由于数据量比较少的原因。然后对其敏感性的分析，我们组是通过加入噪声数据，也就是加入扰动项，然后去观察它的聚类是否与之前有显著差异。

第三问是做预测，这里是直接用第二题建立的随机森林模型进行预测，预测结果如下：

然后模型对训练集的预测效果本身就很出色，

第四问是分析不同类别玻璃文物化学成分的关联关系以及分析之间的差异性，实际上就是进行相关性分析以及相关系数的方差分析。对于相关性分析，由于化学成分均为数值型变量，因此采用pearson相关系数进行分析，并绘制热力图。差异性分析由于是使用相关系数，所以我们是直接对相关系数之间的差异进行分析，然后看哪些化学成分之间相关系数在不同类别的玻璃中出现显著差异。