数据预处理

数据清洗

缺失值

- 表格一：编码后采用最近邻插补
- 表格二：空缺值采用0填充

表格1K最近邻插补Nearest Neighbor，KNN聚类的另一种应用

假设表面风化、玻璃类型，纹饰，颜色之间具有一定相关关系

只有颜色具有缺失值nan，列出表面风华，玻璃类型，纹饰的散点图，寻找颜色nan的那个点，到其他距离最近的那个点，选择其颜色值作为填充

KNN算法（k近邻算法）原理及总结-CSDN博客

异常值

- 表格二：含量成分符合某一范围

数据结构化

分类变量编码量化

表格1都是分类变量，所以要进行编码量化，编码方式有以下几种

python - 11个常见的分类特征的编码技术 - deephub - SegmentFault 思否

数据集成

表格1和表格2特征集成

问题1

表面风化、玻璃类型、纹饰、颜色的关系分析

变量之间的数学关系有哪些？

- 相关性分析、独立分析

相关性分析的方法有哪些？

- 卡方检验（Chi-Square Test），相关系数（Correlation Coefficient）——皮尔森，斯皮尔森系数， 协方差（Covariance）

```
卡方检验（Chi-Square Test）：
卡方检验是一种统计检验方法，用于检验两个或多个分类变量之间是否独立。在这个案例中，它可以用来分析表面风化（分类变量）与玻璃类型、纹饰和颜色（也都是分类变量）之间是否存在关联。通过构建二维或更高维的列联表，并计算卡方值，可以判断这些变量是否独立。
对应分析（Correspondence Analysis）：
对应分析是一种用于分析两个或多个分类变量之间关系的统计方法。它通过降维技术，将变量间的关系简化为一个或多个维度上的图形表示，从而直观地展示变量之间的关联。在这个案例中，对应分析可以用来研究表面风化与玻璃类型、纹饰和颜色之间的关系。
单因素方差分析（One-Way ANOVA）：
虽然单因素方差分析主要用于分析一个定量因变量与一个或多个分类自变量之间的关系，但在某些情况下，也可以结合其他方法（如独立样本t检验）来间接分析多个分类变量之间的关系。在这个案例中，它可以用来分析不同玻璃类型（或纹饰、颜色）的文物在表面风化方面的差异是否显著。
逻辑斯蒂回归（Logistic Regression）：
虽然逻辑斯蒂回归主要用于处理因变量为二分类或多分类变量的情况，但在这个案例中，如果我们将表面风化视为一个二分类变量（如“风化”与“未风化”），则可以结合玻璃类型、纹饰和颜色作为自变量来建立逻辑斯蒂回归模型，分析这些自变量对表面风化状态的影响。
```

9 多元回归分析（Multiple Regression Analysis）：
10 虽然多元回归分析主要用于处理定量因变量和多个定量自变量之间的关系，但在这个案例中，如果能够将表面风化量化为某种定量指标（如风化层的厚度或化学成分的变化量），则可以尝试使用多元回归分析来探究玻璃类型、纹饰和颜色等自变量对这种量化后的风化指标的影响。
11 聚类分析（Cluster Analysis）：
12 聚类分析是一种无监督学习方法，用于将数据集分成若干个组或簇，使得同一簇内的数据点相似度较高，而不同簇之间的数据点相似度较低。在这个案例中，可以使用聚类分析来探索不同玻璃文物在表面风化、玻璃类型、纹饰和颜色等特征上的自然分组情况，从而揭示这些变量之间的潜在关系。
13 主成分分析（Principal Component Analysis, PCA）：
14 主成分分析是一种数据降维技术，通过线性变换将多个变量转换为少数几个综合变量（即主成分），这些主成分能够保留原始数据的大部分信息。在这个案例中，可以使用PCA来提取玻璃文物在表面风化、玻璃类型、纹饰和颜色等方面的主要特征信息，并据此进行进一步的分析和解释。

解题思路

- step0.编码量化
- step1.填补空缺值
- step2.卡方检验

统计规律

统计学规律有哪些？

解题思路

- step0.数据集成，为表格2添加文物编号和是否风化特征
- step1.数据清洗，空缺值和异常值
- step2.统计学规律分析

预测其风化前的化学成分含量

预测模型有哪些?

数学建模【四大模型（优化、分类、评价、预测）总结】_优化模型-CSDN博客

数学建模-预测模型总结(适用范围、优缺点)【灰色预测模型、插值与拟合、时间序列预测法、马尔科夫预测、差分方程、微分方程模型、神经元网络】_灰色预测模型适用范围-CSDN博客