Composition analysis and identification of ancient glass products

陈昊天 Sky 20202505003 Software Engineering Class 1

Abstract

The Silk Road was the channel of cultural exchange between China and the West in ancient times, and the glass is the precious material evidence of early trade. Ancient glass is easily weathered under the influence of burial environment. In the process of weathering, internal elements exchange a lot with environmental elements, leading to the change of its composition ratio, thus affecting the correct judgment of its category. Therefore, it is of great significance to make statistical analysis on the composition of ancient glass products and establish reasonable mathematical model for effective identification of glass.

1.The relationship between the surface weathering of these glass relics and their glass types, patterns and colors is analyzed. Combined with the type of glass, the statistical rule of whether there is weathering chemical composition content on the surface of cultural relic samples is analyzed, and according to the weathering point detection data, to predict the chemical composition before weathering.： Firstly, the correlation between the surface weathering of glass relics and its glass type, pattern and color was calculated by chi-square test. The Chi-square values of glass type, pattern and color were 3.861, 5.720 and 6.847, respectively, indicating a strong correlation between glass weathering and its type. Based on the type of glass, the statistical rule of weathering chemical composition on the surface of cultural relics samples was studied by variance analysis. It was found that weathering of lead-barium glass would lead to a decrease in silica content and an increase in lead oxide content. The weathering of high potassium glass will cause the content of silica to increase obviously, while the content of other contents such as potassium oxide and calcium oxide to decrease slightly. Therefore, the change of silica and potassium oxide content has a strong correlation with the weathering of high potassium glass. The content of silica and lead oxide has a strong correlation with the weathering of lead-barium glass. Based on this conclusion, I used the ratio to predict the chemical composition content of weathering point detection data before weathering.

2.The classification rules of high potassium glass and lead barium glass were analyzed according to the given data. Appropriate chemical components were selected for each category and subcategorized. Specific classification methods and results were given, and the rationality and sensitivity of classification results were analyzed.： The principal component analysis method was used to give weight to the content of each component, and combined with the screening of data for classification discussion, finally determined that silica, lead oxide, potassium oxide, barium oxide as the classification basis of high potassium and lead barium glass. On this basis, K-Means algorithm is used to cluster high-potassium and lead-barium glass by combining with the chemical composition. Finally, high-potassium glass is divided into high-potassium Ⅰ, high-potassium Ⅱ and high-potassium Ⅲ, and lead-barium glass is divided into lead-barium Ⅰ, lead-barium Ⅱ, lead-barium Ⅲ and lead-barium Ⅳ. Finally, the rationality of classification is verified by elbow rule, and the sensitivity of classification results is analyzed by scatter diagram.

3.The chemical composition of glass relics of unknown category in form 3 of the data was analyzed to identify their type, and the sensitivity of the classification results was analyzed.： By using SVM support vector machine, dict={' high potassium ': 0,' lead barium ': 1}, and 80% of the data in attachment form 2 is divided into training set, and 20% of the data in attachment form 2 and the data in attachment form are divided into test set to analyze the chemical composition of unknown category glass relics in attachment Table 3. Through the identification of unknown cultural relic types, the result array = [0, 1, 1, 1, 1, 0, 0, 1] is obtained. Combined with the classification rule in question two, the consistency and rationality of the prediction result and classification rule are verified. Further, sensitivity analysis

method was used to compare the training model with the initial learning results after changing the content proportion of a single component, indicating that the model has good rationality and sensitivity.

4.Aiming at different categories of glass cultural relics samples, the correlation between their chemical components was analyzed, and the difference of the correlation between different categories of chemical components was compared.： The correlation between the chemical components of different types of glass cultural relics was analyzed by grey correlation analysis method, and the correlation coefficients of the parent sequence and different sub-sequences were obtained. The grey correlation degree was calculated by the formula, and the strength of the correlation between the parent sequence and sub-sequence was obtained by the size of the grey correlation, that is, the strength of the chemical component correlation.

一、raise a question

1.1 Background

Early glass was introduced into China as ornaments from West Asia and Egypt along the Silk Road. China fully learned this technology and made finished products very similar to foreign glass products by using local materials. However, due to the difference in materials, the chemical composition of the final glass products was not similar. The main composition of glass is silicon dioxide, because the melting point of pure quartz sand is much higher than the temperature of the fire, so the help of flux and stabilizer is needed, and because of the different types of flux, glass is divided into lead barium glass and potassium glass. Ancient glass is highly susceptible to weathering, resulting in a change in composition ratio that can affect their classification.

Therefore, the use of statistical method to analyze the ancient glass products, and establish an effective mathematical model to accurately and efficiently identify the ancient glass products, help promote the development of our archaeology industry.

1.2 Problem restatement

Problem 1: Analyze the relationship between the surface weathering of cultural relics and its type, ornament and color, analyze whether the surface has weathered chemical composition content law through the type of glass, and predict the chemical composition content before weathering according to the data of weathering points.

Problem 2: Analyze the classification law of high potassium and lead barium glass, divide the subclass according to the chemical composition, give the method and results, and analyze the rationality and sensitivity of the results.

Problem 3: Analyze the chemical composition of glass of unknown class, identify its type, and then analyze the sensitivity of the result.

Problem 4: Analyze the correlation between chemical components of different categories of glass cultural relics and find out the difference of chemical component correlation.

二、 model assumption

1. The detection data of sampling points are within the reasonable error range.
2. Weathered cultural relics are not prone to REDOX reactions.
3. The shape of cultural relics is not easy to change.
4. The material and state of the sampling point remain unchanged.

三、 symbol description

symbol	description
x_i	The content of component i in weathering point to be detected
y_i	Chemical content before weathering
a_i	Component i mean before weathering
b_i	The mean value after weathering is b_i
σ_i	The lower and upper limits of the forecast range
δ_i	The lower limit and upper limit of the interval obtained by descriptive statistics of historical data

四、 Model building and solving

4.1.1 Data preprocessing

Excel was used to combine the data of attached Form 1 and Form 2, and then the sum of components at the sampling point of each numbered relic was calculated. Cultural relics missing in color in form 1 (No. 58, No. 48, No. 40, No. 19) and cultural relics in form 2 (No. 15, No. 17) that are not between 85% and 105% are excluded. Based on the original data, combined with the meaning of the question and the understanding of the team, the sample size was reduced to 52 cultural relics to improve the quality of the data.

Cultural relic number	Summation of components (%)
15	79.47
17	71.89

Table 1：The sum of culled artifacts

4.1.2 correlation analysis

Class1: Spearman coefficient analysis

Spearman coefficient was calculated to show the correlation between weathering of cultural relics and glass types, ornamentals and colors.

$$\rho = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma_i (x_i - \overline{x})^2 \Sigma_i (y_i - \overline{y})^2}} \tag{1}$$

$$\rho = \frac{6\Sigma d_i^2}{n(n^2 - 1)} \tag{2}$$

$d_i = x_i - y_i$ Represents the difference between two ranks. If you have the same variables here, then its rank is the average of the positions of the variables sorted from smallest to largest.

The following table shows the correlation coefficient matrix between cultural relic weathering and glass type, ornamentation and color respectively:

	ornamentation	type	color	weathering
ornamentation	1.00	0.19	0.23	-0.095
type	0.19	1.00	0.022	-0.27
color	0.23	0.022	1.00	0.18
weathering	-0.095	-0.27	0.18	1.00

Table 2： correlation coefficient matrix

Visualization analysis of the correlation coefficient matrix is carried out to obtain the correlation coefficient thermal map:

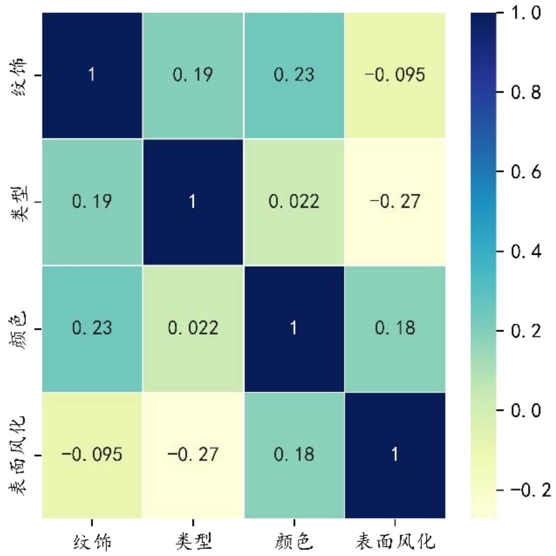


图 1： Correlation coefficient heat map

It can be seen from the correlation coefficient matrix that surface weathering is negatively correlated with cultural relic ornamentation and type. There is a positive correlation between surface weathering and color of cultural relics. The absolute value of the correlation coefficient is close to 1, the higher the correlation is. It can be concluded that the three factors of ornamentation, glass type and color are weakly correlated with surface weathering. However, compared with ornamentation and color, surface weathering of cultural relics is more closely related to its glass type.

Class2: chi-square test

Principle:

Chi-square test is the degree of deviation between the actual observed value and the theoretical inferred value of the statistical sample. The degree of deviation between the actual observed value and the theoretical inferred value determines the Chi-square value. The larger the chi-square value is, the greater the degree of deviation between the two values is. On the contrary, the smaller the deviation; If the two values are exactly equal, the chi-square value is 0, indicating that the theoretical values are in perfect agreement.

The pre-processed data in the attachment were imported into SPSS to calculate the chi-square test values of cultural relic weathering and glass type, pattern and color respectively. The results are shown below:

		weathered	unweathered	sum	chi-square value	P-value
kind	High potassium count	6	10	16	3.861	0.049*
	Lead barium quantity	24	12	36		
ornamentation	Ornamentation A	9	11	22	5.72	0.057
	Ornamentation B	6	0	6		
	Ornamentation C	15	11	30		
color	Black	2	0	2	6.847	0.445
	Blue-green	9	6	15		
	Green	0	1	1		
	wathet	12	7	19		
	aqua	1	2	3		
	dark blue	0	2	2		
	bottle green	4	2	6		
	purple	2	2	4		

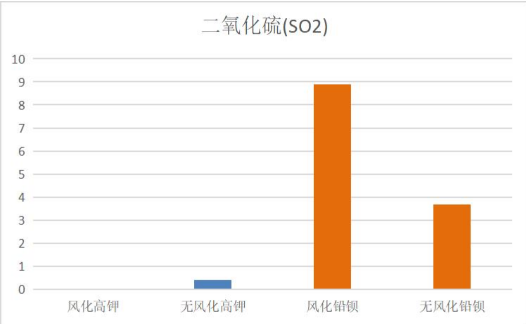
*p<0.05 **p<0.01 ***p<0.005

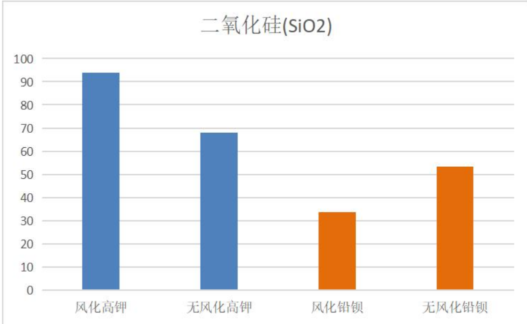
Table 3: Chi-square test results table

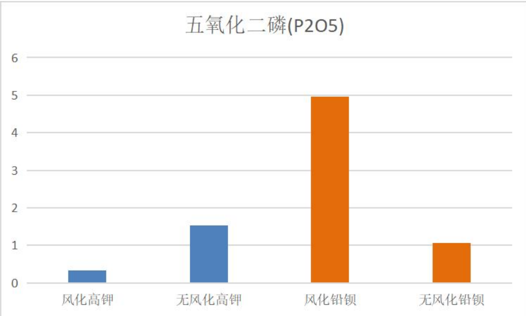
According to our chi-square test results, the chi-square test result of weathering and type is 3.861a, the Chi-square test result of weathering and pattern is 5.720a, and the Chi-square test result of weathering and color is 6.847a. By comparison, we find that the chi-square value of weathering and glass type is the smallest, followed by weathering and decoration, and the chi-square value of weathering and color is the largest. According to the meaning of Chi-square test, the correlation between glass type and weathering degree is the strongest, followed by ornamentation, and the correlation between color and weathering degree is the weakest. It is consistent with the conclusion obtained by Spearman coefficient above and in line with expectations.

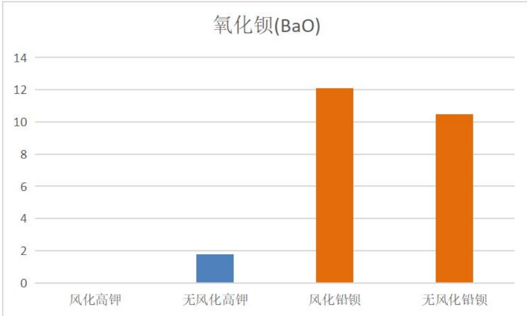
Through the analysis of the upper surface, it can be seen that lead barium glass is easier to weather; Type B ornamentation is easier to weather; Black weathers easily.

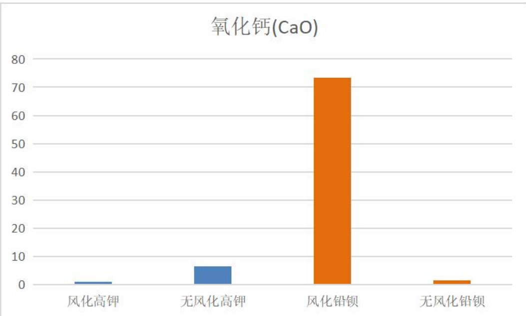
After classifying and summarizing, the average values of 14 components of four kinds of glass were calculated respectively, and the characteristics of the variables were analyzed. The following figure shows the statistical analysis of the average contents of 14 components in the four cultural relics types:

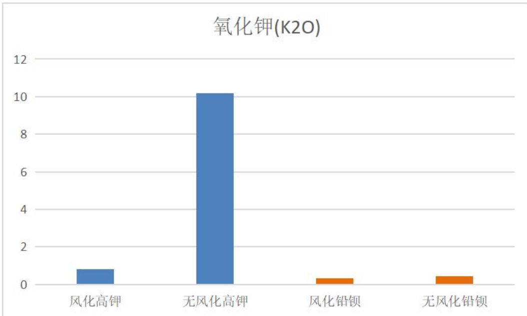

二氧化硫(SO2)
风化高钾
无风化高钾
风化铅钡
无风化铅钡

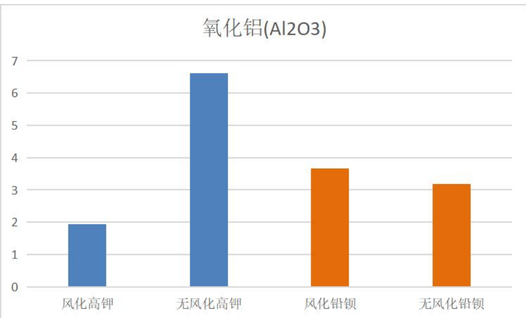

二氧化硅(SiO2)
风化高钾
无风化高钾
风化铅钡
无风化铅钡

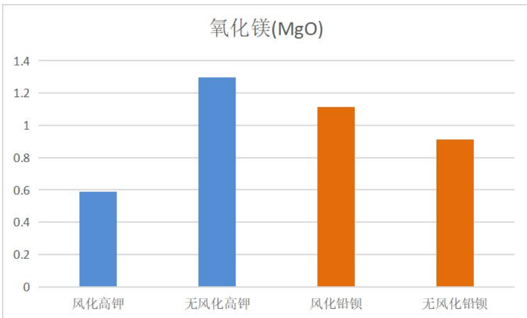

五氧化二磷(P2O5)
风化高钾
无风化高钾
风化铅钡
无风化铅钡

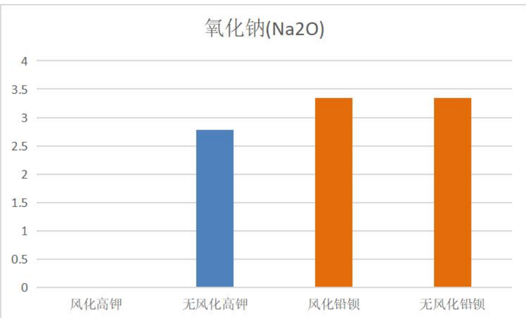

氧化钡(BaO)
风化高钾
无风化高钾
风化铅钡
无风化铅钡

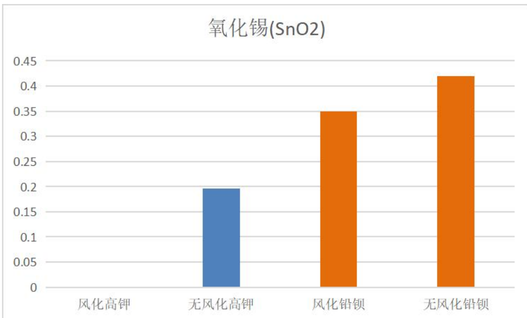

氧化钙(CaO)
风化高钾
无风化高钾
风化铅钡
无风化铅钡


氧化钾(K2O)
风化高钾
无风化高钾
风化铅钡
无风化铅钡


氧化铝(Al2O3)
风化高钾
无风化高钾
风化铅钡
无风化铅钡


氧化镁(MgO)
风化高钾
无风化高钾
风化铅钡
无风化铅钡


氧化钠(Na2O)
风化高钾
无风化高钾
风化铅钡
无风化铅钡


氧化锡(SnO2)
风化高钾
无风化高钾
风化铅钡
无风化铅钡

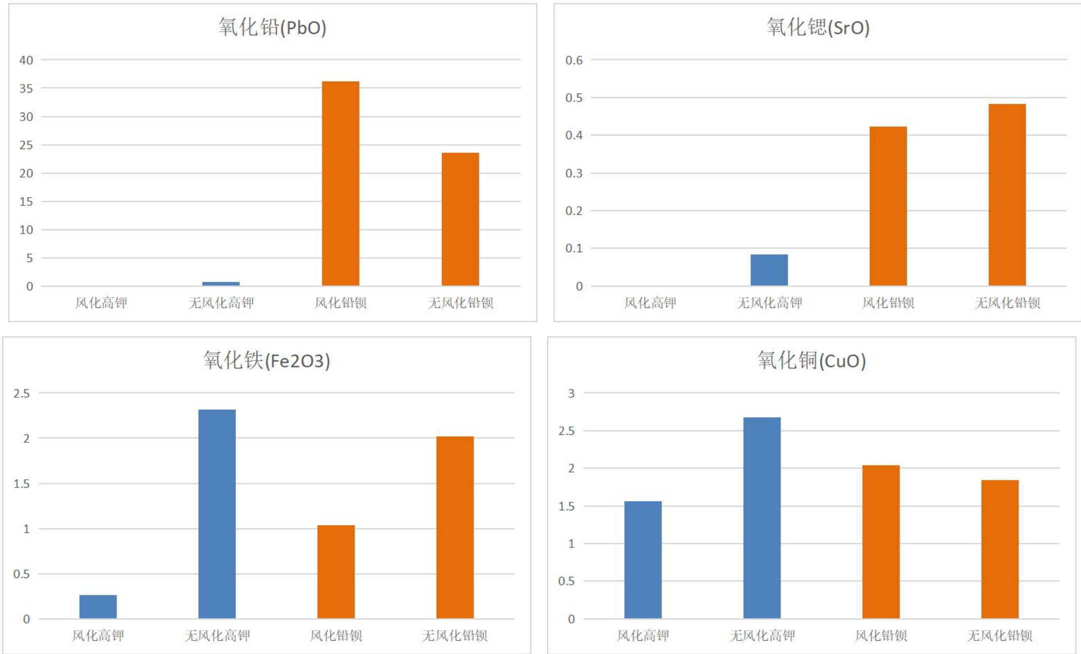
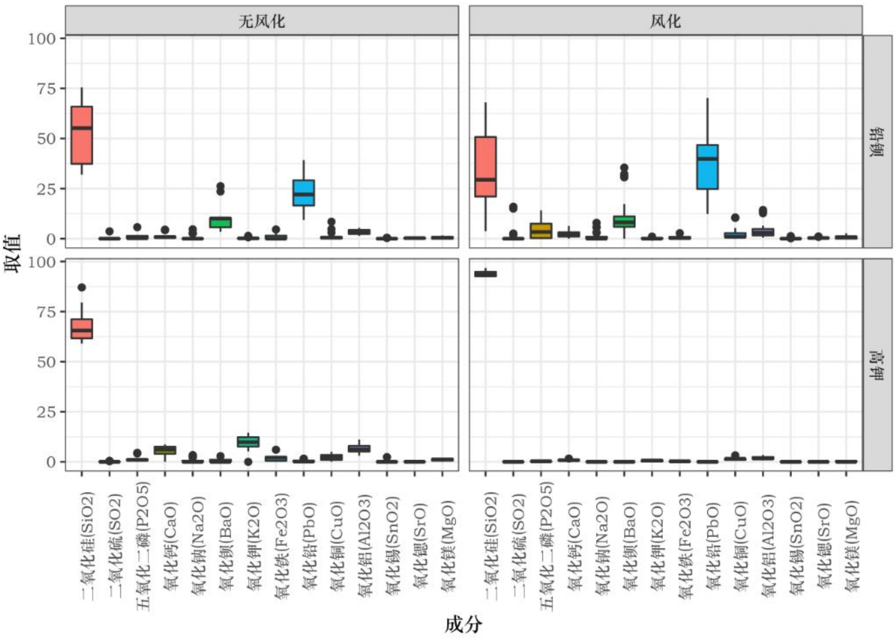


Figure 2： **Statistical analysis of the average content of 14 components in four cultural relics typ**

Through the analysis of the contents proportion of the above components, it can be seen that the contents of sulfur dioxide, lead oxide, strontium oxide and barium oxide in lead-barium glass are significantly higher than the average contents in high-potassium glass. The content ratio of the above elements can be used as the classification basis for the preliminary classification of glass types. Secondly, for the two kinds of elements, calcium oxide and potassium oxide, their content proportion distribution is simple. For example, it can be seen from the figure above that the content of calcium oxide in weathered lead barium glass is higher than the other three types; The content of potassium oxide in non-weathered high potassium glass is significantly higher than the other three types. The content of silica in both high potassium glass and lead barium glass occupies a large proportion, but the average content of silica in lead barium glass is lower than that in high potassium glass.

R language was used to conduct variance analysis on the data in Form 2, and then multivariate analysis of variance was further carried out for quantitative analysis. The analysis results as shown in the figure below are obtained:



Statistical law: As can be seen from the boxplot above, weathering will lead to changes in chemical composition content. For lead-barium glass, weathering results in reduced silica content and increased lead oxide content. The content of phosphorus pentoxide increased slightly after weathering compared with before weathering, so the content of silica and lead oxide had a strong correlation with the weathering of lead-barium glass. For high potassium glass, after weathering compared with before weathering, the content of silica increases significantly and the proportion of the content is concentrated in 90%-100%, while the contents of other contents such as potassium oxide and calcium oxide decrease slightly. Therefore, the change of the content of silica and potassium oxide has a strong correlation with the weathering of high potassium glass.

4.1.3 Pre-weathering content prediction

According to the above conclusions, the contents of chemical components such as alumina and magnesium oxide will not change significantly before and after weathering. Therefore, for the weathering point detection data, we only need to predict the contents of silicon dioxide, lead oxide and potassium oxide before weathering.

We use ratio analysis to predict the mean level of each chemical composition before weathering, and on this basis, give a range of predicted values. The specific analysis process is as follows:Suppose the mean value of component i in the historical data before weathering is a_i , The mean value after weathering is b_i ; x_i is the content of component i in weathering point to be detected， y_i is the chemical composition content before weathering。 By ratio $\frac{a_i}{b_i}$ as the prediction standard, the prediction formula of component content before weathering is given

$$\frac{a_i}{y_i} = \frac{b_i}{x_i} \tag{1}$$

$$y_i = \frac{b_i}{a_i} x_i \tag{2}$$

Further, we obtained the ratio according to the results of descriptive statistics in the following table $\frac{b_i}{a_i}$, and the fluctuation range of the predicted value is given $[\sigma^-, \sigma^+]$。

δ^-, δ^+ determined by the following two formulas:

$$\begin{gathered} \frac{\delta_i^+ - b_i}{b_i} = \frac{\sigma_i^+ - y_i}{y_i} \\ \frac{b_i - \delta_i^-}{b_i} = \frac{y_i - \sigma_i^-}{y_i} \end{gathered} \tag{3}$$

δ_i^- 、 δ_i^+ Is the lower limit and upper limit of the interval obtained by descriptive statistics of historical data; Is the lower limit and upper limit of the predicted fluctuation range; In formula (5), is the parameter value obtained by formula (4).

Using the above formula, we predicted the pre-weathering component content data of the test points and sorted them out in Table 5.

sampling site	silicon dioxide		lead oxide		potassium oxide	
	after weathering	before weathering	after weathering	before weathering	after weathering	before weathering
08 严重 1 点	4.61	7.30	32.45	49.83	0	0
26 严重 1 点	3.72	5.89	29.92	45.94	0.4	0.22
54 严重 1 点	17.11	27.09	58.46	89.76	0	0

Table 5：Component prediction results table

4.2 Problem 2

4.2.1 PCA

SPSS V21.0 was used to analyze the data. Firstly, the correlation among indicators was explored, and it was calculated that the coefficient with absolute value greater than 0.3 accounted for more than 60% of the correlation coefficient matrix, indicating that the data was suitable for principal component analysis. The software was used to conduct principal component analysis on the standardized data, and a total of 4 principal component factors with eigenvalues greater than 1 were extracted. The index eigenvalues and index variance contribution rates of each principal component F_i were shown in Table 5.

The following is the principal component analysis process for high potassium glass:

总方差解释

成分	初始特征值 总计	方差百分比	累积 %	提取载荷平方和 总计	方差百分比	累积 %
1	5.642	40.298	40.298	5.642	40.298	40.298
2	2.627	18.766	59.064	2.627	18.766	59.064
3	1.749	12.490	71.553	1.749	12.490	71.553
4	1.631	11.651	83.205	1.631	11.651	83.205
5	.960	6.854	90.059	.960	6.854	90.059
6	.663	4.734	94.793			
7	.312	2.227	97.020			
8	.200	1.430	98.451			
9	.115	.822	99.272			
10	.066	.470	99.742			
11	.020	.142	99.884			
12	.010	.073	99.958			
13	.006	.041	99.998			
14	.000	.002	100.000			

提取方法：主成分分析法。

Table 7：Total variance of interpretation

It can be seen from Table 5 that the variance contribution rate of the first five principal components is 40.298%, 18.766%, 12.490%, 11.651% and 6.854% respectively, and the cumulative variance contribution rate reaches 90.059%, indicating that the five principal components have covered most of the information of the selected variables, and the information load of the remaining factors can be ignored. According to the five principal components extracted, the component matrix is further calculated as shown in Table 3:

成分矩阵[a]

	成分 1	2	3	4	5
二氧化硅(SiO2)	-.944	.285	.112	-.085	-.034
氧化钠(Na2O)	.232	-.798	.180	.397	-.267
氧化钾(K2O)	.716	-.447	-.242	.314	.164
氧化钙(CaO)	.655	-.704	-.099	-.108	.050
氧化镁(MgO)	.738	.439	-.296	.123	.187
氧化铝(Al2O3)	.889	-.015	.032	.163	-.278
氧化铁(Fe2O3)	.821	.217	-.036	-.280	-.280
氧化铜(CuO)	.508	-.084	.151	-.660	.206
氧化铅(PbO)	.434	-.225	.699	.181	.309
氧化钡(BaO)	.549	.370	.627	-.169	.338
五氧化二磷(P2O5)	.648	.590	-.062	.046	-.417
氧化锶(SrO)	.700	.546	.025	.249	.027
氧化锡(SnO2)	-.090	.316	-.314	.697	.396
二氧化硫(SO2)	.285	-.181	-.733	-.435	.289

提取方法：主成分分析法。

a. 提取了 5 个成分。

Table 8：Component matrix

As can be seen from Table 6, silicon dioxide, alumina and iron oxide can be comprehensively summarized as F1, sodium oxide and calcium oxide as F2, sulfur dioxide and lead oxide as F3, tin oxide as F4, and phosphorus pentoxide as F5.

Analysis of lead-barium glass is as follows:

总方差解释

成分	初始特征值 总计	初始特征值 方差百分比	初始特征值 累积 %	提取载荷平方和 总计	提取载荷平方和 方差百分比	提取载荷平方和 累积 %
1	3.526	25.183	25.183	3.526	25.183	25.183
2	3.051	21.795	46.978	3.051	21.795	46.978
3	1.521	10.867	57.845	1.521	10.867	57.845
4	1.171	8.363	66.208	1.171	8.363	66.208
5	.937	6.691	72.898	.937	6.691	72.898
6	.798	5.702	78.601	.798	5.702	78.601
7	.715	5.106	83.707	.715	5.106	83.707
8	.588	4.202	87.909	.588	4.202	87.909
9	.562	4.013	91.922			
10	.445	3.178	95.100			
11	.358	2.555	97.655			
12	.212	1.514	99.168			
13	.112	.801	99.969			
14	.004	.031	100.000			

提取方法：主成分分析法。

It can be seen from Table 6 that the variance contribution rate of the first eight principal components is 25.183%, 21.795%, 10.867%, 8.363%, 6.691%, 5.072%, 5.106% and 4.202% respectively, and the cumulative variance contribution rate reaches 87.909%, indicating that these eight principal components have covered most of the information of the selected variables. The information load of the remaining factors is negligible. According to the extracted eight principal components, the component matrix table is further calculated:

成分矩阵[a]

	成分							
	1	2	3	4	5	6	7	8
二氧化硅(SiO2)	-.915	-.236	.023	-.020	-.124	.088	.020	-.150
氧化钠(Na2O)	-.423	-.392	-.188	.387	.392	.047	.349	.319
氧化钾(K2O)	-.223	.282	.584	.019	.363	-.475	-.360	.052
氧化钙(CaO)	.365	.743	.153	.071	-.120	.068	.158	.276
氧化镁(MgO)	-.230	.685	-.059	.479	.080	.005	-.062	.197
氧化铝(Al2O3)	-.554	.329	.282	.459	-.062	.042	-.048	-.201
氧化铁(Fe2O3)	-.198	.597	.338	-.240	-.162	.296	-.020	-.083
氧化铜(CuO)	.540	-.478	.243	.192	.052	.421	-.371	.166
氧化铅(PbO)	.574	.433	-.480	-.291	.262	-.199	-.091	.038
氧化钡(BaO)	.626	-.555	.470	.063	.052	.077	-.010	.081
五氧化二磷(P2O5)	.544	.517	-.089	.280	-.364	.019	.029	.031
氧化锶(SrO)	.555	.194	-.182	.361	.426	.174	.019	-.493
氧化锡(SnO2)	-.056	.427	.418	-.397	.409	.292	.327	.006
二氧化硫(SO2)	.555	-.253	.445	.162	-.174	-.361	.421	-.142

提取方法：主成分分析法。

a. 提取了 8 个成分。

As can be seen from the table, silica can be comprehensively summarized as F1, calcium oxide as F2, potassium oxide as F3, oxidation as F4, sulfur dioxide and strontium oxide as F5.

4.2.2 K-means clustering

K-means clustering algorithm is a kind of iterative solving clustering analysis algorithm. Its step is to randomly select K objects as the initial clustering center, then calculate the distance between each object and each seed clustering center, and assign each object to the nearest clustering center.

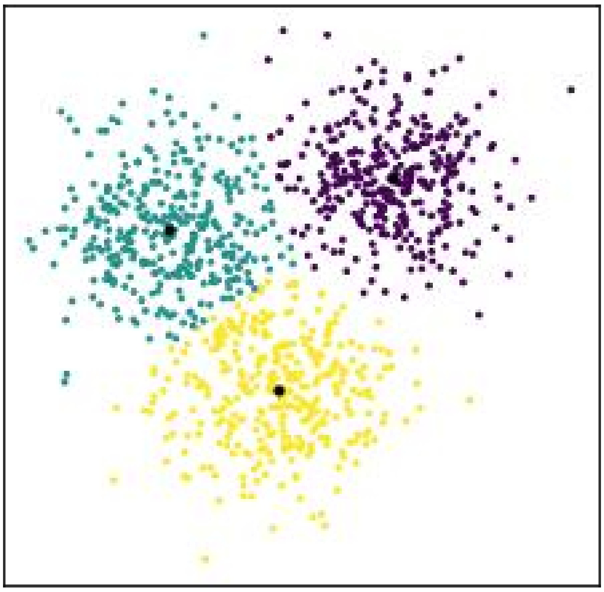


Figure 3：Clustering result

4.2.3 Result of subcategory division

The K-Means algorithm was used to cluster the high-potassium and lead-barium glass, combined with the chemical composition, and the high-potassium glass was finally divided into class Ⅰ, class Ⅱ and class Ⅲ, while the lead-barium glass was divided into class Ⅰ, class Ⅱ, class Ⅲ and class Ⅳ. Finally, the rationality of classification is verified by elbow rule, and the sensitivity of classification results is analyzed by scatter diagram.

```
氧化铅(Pb0)          46.792632
二氧化硅(Si02)        26.688421
氧化钡(Ba0)           8.234737
五氧化二磷(P205)        4.643684
氧化铝(A1203)         3.011579
氧化钙(Ca0)           2.898421
氧化铜(Cu0)           1.046316
氧化铁(Fe203)         0.876316
氧化镁(Mg0)           0.776842
氧化锶(Sr0)           0.421579
Name: cluster_0, dtype: float64
```

```
二氧化硅(Si02)      65.422
氧化铅(Pb0)        16.447
氧化钡(Ba0)         5.911
氧化铝(A1203)       5.477
氧化钙(Ca0)         1.205
氧化钠(Na20)        0.968
氧化铁(Fe203)       0.775
氧化铜(Cu0)         0.676
氧化镁(Mg0)         0.626
五氧化二磷(P205)      0.525
Name: cluster_1, dtype: float64
```

```
二氧化硅(Si02)       48.484545
氧化铅(Pb0)         23.888182
氧化钡(Ba0)         10.515455
氧化铝(A1203)        3.691818
氧化钠(Na20)         2.638182
五氧化二磷(P205)       2.428182
氧化铜(Cu0)          2.058182
氧化钙(Ca0)          1.148182
氧化镁(Mg0)          0.699091
氧化铁(Fe203)        0.393636
Name: cluster_3, dtype: float64
```

Classification of lead and barium

```
二氧化硅(Si02)        63.624444
氧化钾(K20)         10.817778
氧化铝(A1203)         7.348889
氧化钙(Ca0)           6.363333
氧化铜(Cu0)           2.818889
氧化铁(Fe203)         2.312222
五氧化二磷(P205)        1.523333
氧化镁(Mg0)           1.133333
氧化钠(Na20)          0.926667
氧化钡(Ba0)           0.578889
Name: cluster_1, dtype: float64
```

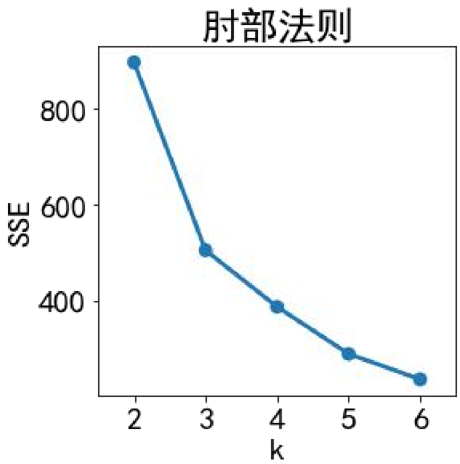
```
二氧化硅(Si02)        63.624444
氧化钾(K20)         10.817778
氧化铝(A1203)         7.348889
氧化钙(Ca0)           6.363333
氧化铜(Cu0)           2.818889
氧化铁(Fe203)         2.312222
五氧化二磷(P205)        1.523333
氧化镁(Mg0)           1.133333
氧化钠(Na20)          0.926667
氧化钡(Ba0)           0.578889
Name: cluster_1, dtype: float64
```

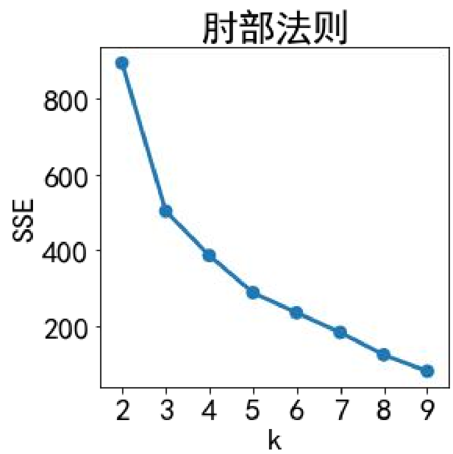
```
二氧化硅(Si02)       81.063333
氧化钾(K20)          4.870000
氧化铝(A1203)        4.433333
氧化钙(Ca0)          2.240000
氧化铜(Cu0)          1.353333
五氧化二磷(P205)       1.040000
氧化镁(Mg0)          0.916667
氧化铁(Fe203)        0.790000
氧化锡(Sn02)         0.786667
氧化钡(Ba0)          0.656667
Name: cluster_2, dtype: float64
```

High potassium classification

4.2.4 Rationality judgment

Run K means to determine the number of clusters using the elbow method. SSE = sum of squared errors. After running Kmeans with k clusters, we calculate its SSE for each k cluster. It can be used as a reference to determine the clustering number of K-means by finding the point of the bend (which has the greatest variation in Angle).





4.3 Problem 3

In this question, the data in attachment form 2 was divided into training set, and the data in attachment form 3 was divided into test set. Support vector machine (SVM) was used to predict the glass type in attachment form 3.

4.3.1 SVM

1) Create a classified dictionary:dict={' high potassium ': 0,' lead barium ': 1}

2) Divide training set and test set

Take 80% data in the attached form 2 as the training set, sample size N=65*0.8=52, and 20% data in the form 2 as the test set, sample size n=65*0.2=13.

After training the training set, use the test set to predict the results as shown in the figure below:

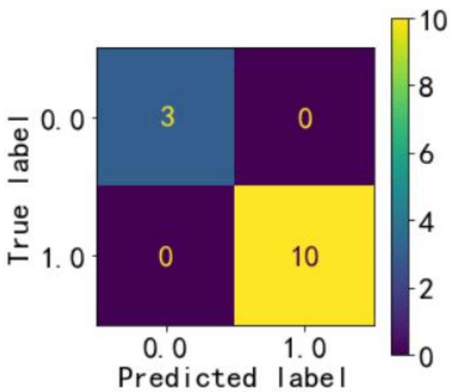


Figure 4: Test sets predict results

As can be seen from the figure, when the actual type is high potassium, the predicted sample size of high potassium is 3, and the predicted sample size of lead barium is 0. When the actual type is lead barium, the predicted sample size of lead barium is 10, and the predicted sample size of high potassium is 0.

1) Get test results

The SVM model is used to train and learn the data in form 3, and the results obtained in sequence are array=[0,1,1,1,1,0,0,1]. Therefore, the prediction results of A1-A8 cultural relics in Form 3 are successively: high potassium, lead barium, lead barium, lead barium, lead barium, lead barium, high potassium, high potassium and lead barium.

4.4 Problem 4

Grey correlation analysis was used to analyze the correlation between the chemical components of different types of glass cultural relics, and the correlation coefficients of the parent sequence and different sub-sequences were obtained. The grey correlation degree was calculated by the formula. The strength of the correlation between the parent sequence and sub-sequence, namely the strength of the chemical component correlation, was obtained by the size of the grey correlation.

4.4.1 Grey correlation results

Import the data into spss, and obtain the correlation degree between each component and silica as follows:

关联度结果		
评价项	关联度	排名
Q7-氧化铝(Al2O3)	0.97	1
Q6-氧化镁(MgO)	0.956	2
Q9-氧化铜(CuO)	0.954	3
Q5-氧化钙(CaO)	0.951	4
Q11-氧化钡(BaO)	0.945	5
Q8-氧化铁(Fe2O3)	0.945	6
Q10-氧化铅(PbO)	0.945	7
Q13-氧化锶(SrO)	0.942	8
Q4-氧化钾(K2O)	0.935	9
Q12-五氧化二磷(P2O5)	0.933	10
Q3-氧化钠(Na2O)	0.93	11
Q14-氧化锡(SnO2)	0.928	12
Q15-二氧化硫(SO2)	0.926	13

Table 9：Grey correlation result

As can be seen from the above table, through the method of grey correlation analysis, it can be concluded that the chemical components in the parent sequence have stronger correlation with the chemical components in which seed sequence. By analyzing the influencing factors of lead barium glass and high potassium glass, it can be seen that alumina (Al2O3) is the most correlated with silicon dioxide (SiO_2), sulfur dioxide (SO_2) is the least correlated with silicon dioxide (SiO_2), and potassium oxide (K_2O) is the most correlated with calcium oxide (CaO). Lead oxide (PbO) is the least correlated with potassium oxide (K_2O), strontium oxide (SrO) is the most correlated with lead oxide (PbO), potassium oxide (K_2O) is the least correlated with lead oxide (PbO).

五、 References

[1]Application conditions of Chi-square test [J]. Journal of Hepatobiliary Diseases, 2012,38(06):1292. (in Chinese)

[2] Su Xiaoyu, Gao Tongmei, Zhang Pengyu, Li Feng, Wu Yin, Wang Dongyong, Tian Yuan, Wei Shuangling. Comprehensive evaluation of Heat resistance of sesame at seedling stage based on principal component analysis and membership function method [J/OL]. Crops: 1-10 [2022-09-18]. http://kns.cnki.net/kcms/detail/11.1808.S.20220728.1506.004.html

[3] Zhu He, Yang Shanshan, Liu Han, Xu Zhenghe. Based on principal component analysis and cluster analysis of agricultural water saving zoning of shandong province [J]. Journal of jinan university (natural science edition), 2022, 4 (4) : 391-396 + 403. DOI: 10.13349 / j.carol carroll nki JDXBN. 20220419.001.

[4] Chen Jingnian. A Support Vector Machine Acceleration Method for Multiple Classification Problems [J]. Computer Science,2022,49(S1):297-300.

[5] Li Kai, Li Jie. Structural Fuzzy Multi-Classification Support Vector Machine Algorithm Based on pinball Loss [J]. Journal of Computer Applications, 201,41(11):3104-3112.

[6] ZHU L. Recognition and application of fruit grade classification based on K-means clustering algorithm [J]. And agricultural mechanization research, 2020 (8) : 46-50. DOI: 10.13427 / j.carol carroll nki njyi. 2020.08.009.

[7] Luo Biao, Yan Weiwei, Wan Liang. Customer value Classification Model Based on ANP and K-means clustering and its application [J]. Journal of Computer Applications,2013,33(10):2954-2959.

[8] Liu Xiaomin, Wei Zhikun, Wang Xiaoli, Xu Rui, Han Dan, Xu Zicheng. Study on the Relationship between reducing sugar content and leaf Morphological characteristics of Hubei flue-cured Tobacco based on grey correlation analysis [J]. Acta Agriculturae Australis Sinica,202,53(06):1536-1542.

Appendix

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import ensemble
from sklearn import metrics
from sklearn.decomposition import PCA, KernelPCA
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.feature_selection import VarianceThreshold, RFE, SelectKBest, chi2
from sklearn.linear_model import LogisticRegression
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier

credit = pd.read_excel("D:\\guosai-c\\1.xls")
credit.head(10)

credit1 = credit
dic1 = {'A':1,'B':2,'C':3}
dic2 = {'高钾':1,'铅钡':2}
dic3 = {'浅蓝':1,'蓝绿':2,'深绿':3,'紫':4,'浅绿':5,'深蓝':6,'黑':7,'绿':8}
dic4 = {'风化':1,'无风化':2}
credit1['纹饰'] = credit['纹饰'].map(dic1)
credit1['类型'] = credit['类型'].map(dic2)
credit1['颜色'] = credit['颜色'].map(dic3)
credit1['表面风化'] = credit['表面风化'].map(dic4)
credit1.corr(method = 'spearman')
# 求 spearman 相关系数
sns.pairplot(credit,kind='reg',diag_kind='hist')
# 观察变量间关系-散点图
credit1 = credit1.drop(['文物编号'],axis=1)
plt.rcParams['font.sans-serif'] = ['KaiTi']   # 指定默认字体
plt.rcParams['axes.unicode_minus'] = False   # 解决保存图像是负号'-'显示为方块的问题
plt.rcParams['figure.figsize'] = (4,4)
plot = sns.heatmap(credit1.corr(method='spearman'),annot = True, linewidths=.5, cmap="YlGnBu")
plt.tight_layout()
fig = plot.get_figure()
fig.savefig("rrr.png", dpi=1080) # 提高分辨率

creditnb = pd.read_excel("D:\\guosai-c\\7.xls")
creditnb = creditnb.drop(['表面风化'],axis=1)
```

```
y1,x1 = np.split(creditnb, (1,), axis=1)
from sklearn.model_selection import train_test_split
from sklearn import svm

x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=1, train_size=0.8)
print(y_train.value_counts()/len(y_train))
print(y_test.value_counts()/len(y_test))
clf = svm.SVC() # C=0.8, kernel='rbf', gamma=20, decision_function_shape='ovr'
clf.fit(x_train, y_train.values.ravel())
print("训练集:",clf.score(x_train, y_train))  #  精度
y_hat = clf.predict(x_train)
print("测试集:",clf.score(x_test, y_test))
y_hat = clf.predict(x_test)
credit_pred = clf.predict(x_test)
from sklearn import metrics
print(metrics.classification_report(y_test, credit_pred))
print(metrics.confusion_matrix(y_test, credit_pred))
print(metrics.accuracy_score(y_test, credit_pred, normalize=True))
metrics.plot_confusion_matrix(clf, x_test, y_test)
model.predict(x1)
credit_q = pd.read_excel("D:\\guosai-c\\8.xls")
credit_q = credit_q.drop(['表面风化'],axis=1)
y2,x2 = np.split(credit_q, (1,), axis=1)
clf.predict(x2)

credit_gj = pd.read_excel('D://guosai-c//gaojia.xls')
credit_gj = credit_gj.drop(['表面风化'],axis=1)
from sklearn.cluster import KMeans
data = credit_gj
data_model = KMeans(n_clusters = 3)
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(data)
from sklearn.cluster import KMeans
sse = {}
# datamart_normalized.describe()
# print(np.isnan(datamart).any())

# print(np.isfinite(datamart_normalized).all())
# k=np.float32(k)
for k in range(2,7):
    kmeans = KMeans(n_clusters=k, random_state=1).fit(data)
    cluster_labels = kmeans.labels_
    sse[k] = kmeans.inertia_
```

```
# plot sse
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.title('肘部法则')
plt.xlabel('k')
plt.ylabel('SSE')
plt.show()
data_model.fit(credit_gj)
data_clusters = pd.Series(data_model.labels_)
data_clusters.value_counts().sort_index()
centers = pd.DataFrame(data_model.cluster_centers_, \
                        columns = credit_gj.columns)
centers_t = centers.T
centers_t.columns = ["cluster_0","cluster_1","cluster_2"]
centers_t["cluster_0"].sort_values(ascending = False, inplace = False).head(10)
centers_t["cluster_1"].sort_values(ascending = False, inplace = False).head(10)
centers_t["cluster_2"].sort_values(ascending = False, inplace = False).head(10)

credit_qb = pd.read_excel('D://guosai-c//qianbei.xls')
credit_qb = credit_qb.drop(['表面风化'],axis=1)
from sklearn.cluster import KMeans
credit_qb_model = KMeans(n_clusters = 4)
credit_qb_model.fit(credit_qb)
credit_qb_clusters = pd.Series(credit_qb_model.labels_)
credit_qb_clusters.value_counts().sort_index()
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(credit_qb)
from sklearn.cluster import KMeans
sse = {}
# datamart_normalized.describe()
# print(np.isnan(datamart).any())

# print(np.isfinite(datamart_normalized).all())
# k=np.float32(k)
for k in range(2,10):
    kmeans = KMeans(n_clusters=k, random_state=1).fit(data)
    cluster_labels = kmeans.labels_
    sse[k] = kmeans.inertia_


# plot sse
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
```

```
plt.title('肘部法则')
plt.xlabel('k')
plt.ylabel('SSE')
plt.show()
centers1 = pd.DataFrame(credit_qb_model.cluster_centers_, \
                        columns = credit_qb.columns)
centers_t1 = centers1.T
centers_t1.columns = ["cluster_0","cluster_1","cluster_2","cluster_3"]
centers_t1["cluster_0"].sort_values(ascending = False, inplace = False).head(10)
centers_t1["cluster_1"].sort_values(ascending = False, inplace = False).head(10)
centers_t1["cluster_2"].sort_values(ascending = False, inplace = False).head(10)
centers_t1["cluster_3"].sort_values(ascending = False, inplace = False).head(10)
```