第二周任务

一、 数据预处理

1.1 剔除无效数据

由要求可以知道，需要保留的数是处于累加和在 85% – 100% 之间。可以发现表单 2 中文物采样点 15 与文物采样点 17 成分比例数据累加和均低于 85%，属于无效数据，故剔除文物采样点 15 和文物采样点 17 两组的数据。

1.2 成分数据

在网上查找资料之后，成分数据是指任意非负的 n 元向量 $x = [x_{1,}x_{2,...,}\, x_n]$ 且满足定和约束 $\sum_{i=1}^{n} x_i = 1, 0 \leq x_i \leq 1$，约束条件为定和约束。这个是成分数据的基本性质。由题目可知，成分数据的累加都在 100% 左右，因此不需要再进行成分数据的转化。

1.3 中心化对数比变换 (数据标准化的方法)

查找资料对数据进行标准化，由资料可以知道：

n 元成分数据所处的向量空间为单形空间，由于单形空间需满足定和约束，因此针对普通数据的传统统计学分析方法对于成分数据不再使用。通过查阅文献得知，成分数据具有以下问题：

1) 数据的直观形态在单形空间和欧氏空间不同，无法跨空间进行解释。

2) 在单形空间上计算得到的成分数据的协方差矩阵有明显偏负性，与欧氏空间上的内涵截然不同。

3) 单形空间上的成分数据缺乏参数分布，使得对数据的变异模式进行分析时存在参数建模困难。

基于上述存在问题，应该对数据进行中心化对数比变换 (clr) 处理，经过中心对数比变换后的数据可以更加充分体现成分特性，使得成分数据中的可解释性更强。clr 计算公式如式 1所示：

$$clr = \left[\ln \frac{x_1}{g(x)}, \ln \frac{x_2}{g(x)}, ..., \ln \frac{x_n}{g(x)}\right] \tag{1}$$

其中 $g(x) = [x_1 \cdot x_2 \cdot, ..., \cdot x_n]^{\frac{1}{n}}$.

1.4 缺失值的处理

在这里将使用零值替换法，零值包括绝对零值和舍入零值，在给出的数据可知，缺失值是舍入零值，也就是非常接近零的值。由于零值不能取对数，因此当成分数据中含有零值时，对数比转换不能进行，在使用对数比转换之前需要对零值进行处理。对零值

进行删除是一种方法，但是会因此而减少用于插值的数据，所以在网上查找资料之后采用零值替换法对零值进行处理。常用的零值替换法有加法替换法，简单替换法，乘法替换法。这里我们使用加法替换法，其计算的方法的如式 2如下：

$$r_i = \begin{cases} \frac{\delta(Z+1)(D-Z)}{D^2}, & x_i = 0 \\ x_i \frac{\delta(Z+1)Z}{D^2}, & x_i > 0 \end{cases} \tag{2}$$

其中 δ 表示的是一个很小的数值，通常取舍入误差或者更小的值，在查阅相关资料之后，决定将 δ 取值为 0.65%,D 表示的是成分数据的组成个数，Z 表示的是成分数据中零值的个数，x_i, r_i 分别表示的是替换前后的数据。由于篇幅原因，这里值展示部分数据，剩余的数据放在附录文件中，其部分数据如表 1:

表 **1**　部分处理缺失值的数据

样本点	二氧化硅	氧化钠	氧化钾	氧化钙
01	69.33	0.17908163	9.99	6.32
02	36.28	0.16581633	1.05	2.34
03 部位 1	87.05	0.18571429	5.19	2.01
03 部位 2	61.71	0.14591837	12.37	5.87
04	65.88	0.17908163	9.67	7.12
05	61.58	0.16581633	10.95	7.35
06 部位 1	67.65	0.16581633	7.37	0.16581633
06 部位 2	59.81	0.14591837	7.68	5.41
07	92.63	0.17908163	0.17908163	1.07
08	20.14	0.17908163	0.17908163	1.48
08 严重风化点	4.61	0.17908163	0.17908163	3.19
09	95.02	0.18571429	0.59	0.62
10	96.77	0.17908163	0.92	0.21

如表 1所示，每一个成分数据都可以得到一个加法替换法的值来对缺失值进行处理，经过数据求和检验，所有样本的成分数据总和均在 $85\% - 105\%$ 之间，证明缺失值处理得当。

1.5 处理缺失值后进行标准化

从上述可知，再进行完缺失值的处理以后，我们可以对数据进行中心化对数比变换，在网上进行多方面的查找之后，发现在 r 语言的 compositions 包中含有中心对数比变换的函数，故采用 r 语言来对数据进行预处理。由于篇幅的原因，这里也只展示部分的数据，剩余的数据将放在附录中。处理数据如下表 2：

表 **2**　中心对数比变换后的数据

样本点	二氧化硅 (SiO2)	氧化钠 (Na2O)	氧化钾 (K2O)	氧化钙 (CaO)	氧化镁 (MgO)
01	4.08626871	-1.87252253	2.14897559	1.6911102	-0.29187107
02	3.45137151	-1.93676968	-0.09110495	0.71025582	0.02561933
03 部位 1	4.86809397	-1.28193458	2.048345	1.09974603	-1.28193458
03 部位 2	3.72422482	-2.32292912	2.11705301	1.37163346	-0.29386116
04	4.01840098	-1.88934745	2.09959439	1.7934738	0.2752519
05	3.94716451	-1.9700472	2.22016683	1.82152768	0.39780692
06 部位 1	4.08248584	-1.92873608	1.86555619	-1.92873608	0.55123533
06 部位 2	3.67825336	-2.33762745	1.62570004	1.27532958	0.1352019
07	5.21222874	-1.03629785	-1.03629785	0.75127433	-1.03629785
08	2.61271593	-2.10990548	-2.10990548	0.00205013	-2.10990548
08 严重风化点	1.07560936	-2.17253202	-2.17253202	0.70740242	-2.17253202
09	5.24764254	-0.98999074	0.16592241	0.21551935	-0.98999074

由中心对数比变换的定义可以知道，每个成分数据求和的值为零，同时打破了传统统计学分析方法对于成分数据不再适用的情况，经过中心对数比变换的数据可以充分的体现成分特性，使得成分数据中的可解释性更加强。

二、 统计规律和相关性检验

2.1 统计规律

在对数据进行完标准化之后，采用标准化数据后的数据（即 clr 化后的数据）对有无风化的样本进行统计规律。要统计的数据分别有：均值，标准差，平均值，最大值，

最小值，偏度，峰度。

下面是无风化的数据，见表 3:

表 **3**　无风化描述统计量

化学成分	平均数	最大值	最小值	标准差	变异系数	偏度	峰度
Si02	4.022	4.868	3.22	0.447	0.111	0.107	-0.758
Na20	-1.123	1.431	-2.338	1.276	-1.136	0.994	-0.782
K20	0.15	2.857	-2.2	1.917	12.765	0.136	-1.887
Ca0	0.397	2.294	-1.968	1.303	3.281	-0.319	-1.228
Mg0	-0.567	0.791	-2.019	0.875	-1.542	-0.254	-1.265
Al203	1.38	2.28	0.397	0.489	0.355	-0.185	-0.809
Fe203	-0.359	1.385	-2.033	1.156	-3.222	-0.082	-1.683
Cu0	-0.014	2.151	-2.113	1.181	-82.047	-0.231	-1.085
Pb0	0.995	3.533	-2.026	2.308	2.32	-0.178	-1.848
Ba0	0.606	3.282	-1.997	1.833	3.025	-0.274	-1.612
P205	-0.578	1.464	-2.411	1.123	-1.942	0.081	-1.193
Sr02	-1.684	-0.079	-3.038	0.74	-0.439	0.134	-0.359
Sn02	-1.636	1.224	-2.338	0.681	-0.416	2.874	9.583
S02	-1.587	1.075	-2.338	0.676	-0.426	2.321	6.83

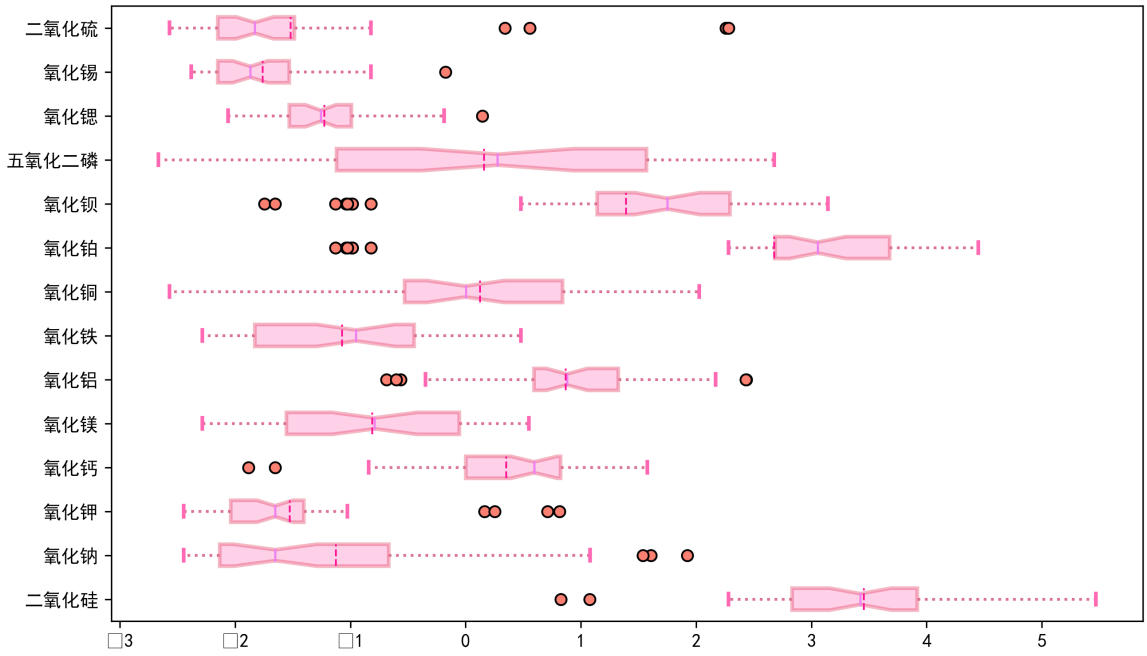
下面是风化的描述性统计量见表 4

表 **4**　风化描述性统计量

化学成分	平均数	最大值	最小值	标准差	变异系数	偏度	峰度
Si02	3.455	5.47	0.827	1.017	0.294	-0.049	0.323
Na20	-1.143	1.922	-2.447	1.285	-1.124	1.024	-0.295
K20	-1.491	0.814	-2.447	0.784	-0.526	1.548	1.778
Ca0	0.361	1.575	-1.883	0.758	2.101	-1.048	0.883
Mg0	-0.79	0.549	-2.284	0.813	-1.029	-0.27	-1.291
Al203	0.888	2.435	-0.687	0.769	0.866	-0.179	-0.354
Fe203	-1.033	0.481	-2.284	0.822	-0.796	0.12	-1.244
Cu0	0.087	2.028	-2.568	1.054	12.173	-0.286	-0.313
Pb0	2.701	4.45	-1.128	1.613	0.597	-1.561	1.095
Ba0	1.314	3.143	-1.937	1.493	1.137	-0.88	-0.547
P205	0.183	2.679	-2.666	1.512	8.279	-0.263	-1.361
Sr02	-1.239	0.144	-2.06	0.487	-0.393	0.625	0.272
Sn02	-1.763	-0.173	-2.38	0.497	-0.282	1.037	0.726
S02	-1.529	2.283	-2.568	1.08	-0.707	2.255	4.926

上述两表展示了风化玻璃和无风化玻璃的各项描述性统计量，下面要使用箱线图来对有无风化玻璃进行数据的可视化。如下图 1表示的是风化玻璃的箱线图：

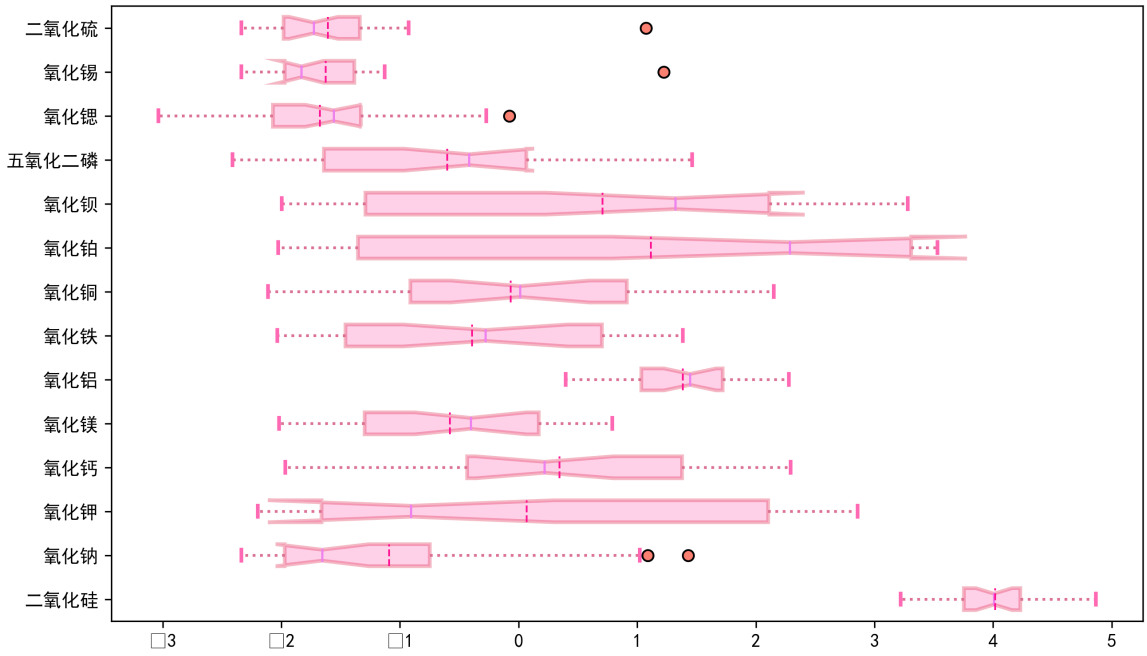
图 **1**　风化玻璃箱线图



箱线图可以充分的展示一组数据。其中箱线图的左右两端的横线表示的是数据的上边缘和下边缘，中间凹下去的部分表示的是数据的中位数，其中画虚线的表示的是数据的均值，箱子的左右边缘分别表示的是下四分位数和上四分位数，其中图中的实心点表示的是异常值，其中实心异常值表示的是超过上分位数 3 倍的数据或者低于下分位数 3 倍的数据，如果异常值是空心的，则为 1.5 倍，而且箱子越小，证明其数据也越符合正态分布。

如图 1所示，展示的是风化风化玻璃的箱线图，可以看出，其中 P_2O_5 的箱子长度最长，表示其最不符合正态分布，同时 S_iO_2, SnO_2, SrO_2 箱子长度最短，最符合正态分布。图中可以直观的看到各个化学成分的数据。为了同时对比无风化玻璃，下面如图 2:

图 **2**　无风化玻璃箱线图

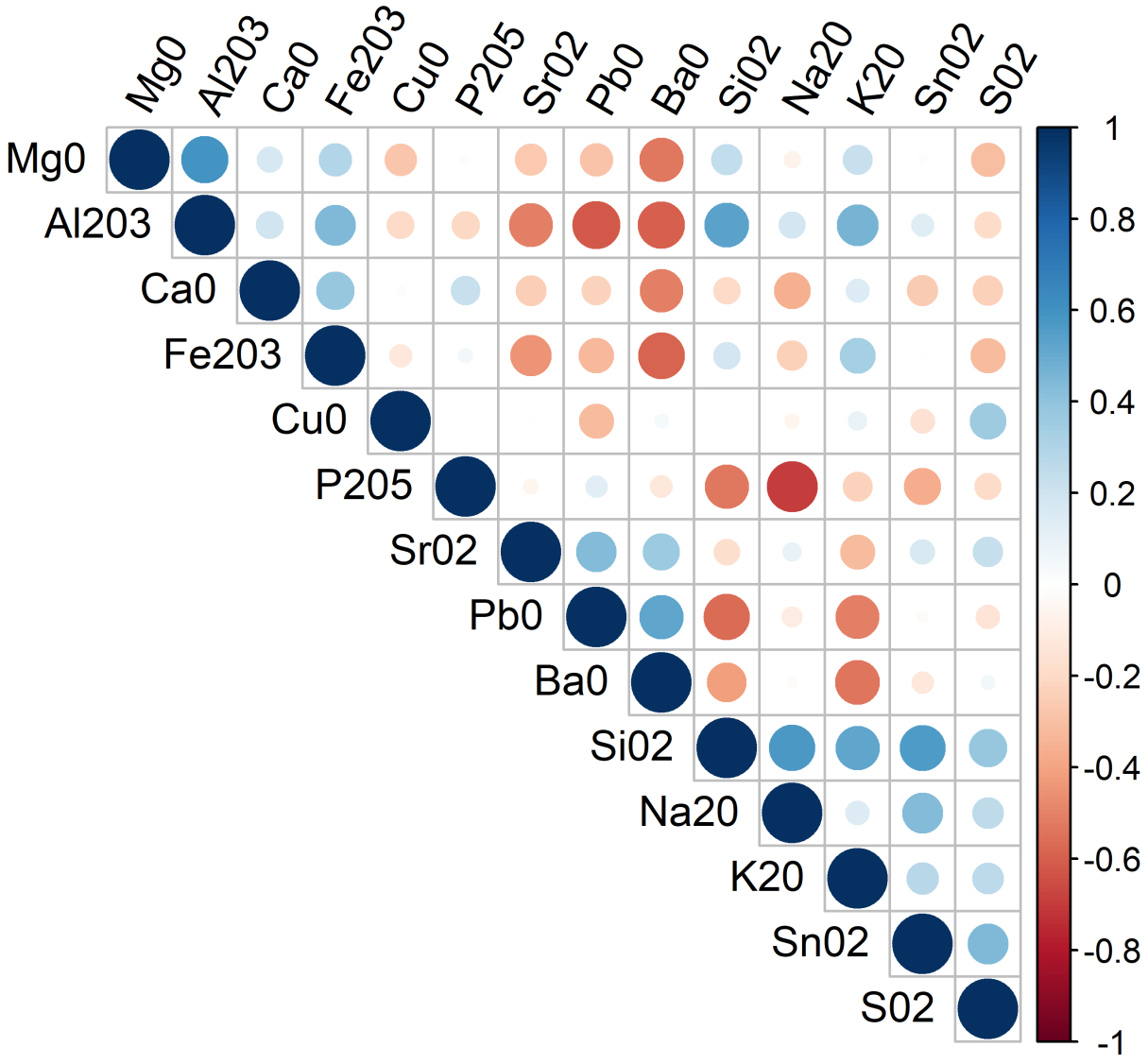


可以明显的看到，无风化玻璃中 PbO_2, K_2O, BaO 含量远远大于风化玻璃，这个也许是玻璃风化前后化学组成成分发生变化的原因。

2.2 相关性分析

在上面什么对数据进行了中心对数比变换的标准化，使得成分数据得以跨空间解释，消除成分数据的协方差的偏负性。这里检验其各个化学成分化学成分的相关性。由于从上述箱线图可以看出，数据的异常值较多，且不充分满足正态分布的条件，由于以上条件的约束，最终采用斯皮尔曼相关系数检验。检验结果如图 3:

图 **3**　各个化学成分的相关系数图



从图中可以看出，Al_2O_3, PbO, Bao 三者有着较强的相关性，P_2O_5, Na_2O 有着较强的相关性，Fe_2O_3, BaO 有着较强的相关性。

三、 相关性分析代码（r 语言）

```
library(Hmisc) #载入编程包，使用这个可以进行多数据的person相关分析
  library(corrplot) #载入数据可视化的包
  library(extrafont) #修改字体的包
  library(openxlsx)
  #font_import()
  a<-read.xlsx("中心对数比变换后的数据.xlsx",
  sheet=1,colNames = TRUE,
  rowNames = TRUE) #导入数据
  head(a)
  names(a)<-c("Si02","Na20","K20","Ca0",
  "Mg0","Al203","Fe203","Cu0",
  "Pb0","Ba0","P205","Sr02",
  "Sn02","S02")
  a1<-as.matrix(a) #进行相关性检验一定要是矩阵的形式才可以。
  P<-rcorr(a1,type="spearman") #使用rcorr来进行person相关检验。
  head(a1)
  #第一部分是相关系数（只能保留两位小数，不过已经可以了。
  #第二部分是有效观测值
  #第三部分是显著性
  Person_<-round(P$r,3) #提取出相关系数
  P_values_<-round(P$P,3) #提取出显著性系数
  par(family= "Times New Roman")
  Persona_<-as.data.frame(Person_)
  #接下来就是相关系数的数据可视化。
  b<-corrplot(Person_, type = "upper",
  order = "hclust", tl.col = "black"
  , tl.srt =60)
  png(filename = "相关系数图.png",width = 1500*2,
  height = 1500*2,units = "px",bg="white",res=300*2)
  corrplot(Person_, type = "upper",
  order = "hclust", tl.col = "black"
  , tl.srt =60)
  dev.off()
```

四、 加法替换法代码（r 语言）

```
#导入最相关的包。没有相关的包先去下载，不然会报错。
library(openxlsx)
library(pls)
getwd() #查看当前工作路径。

#读取数据。其中pre_data是最原始删除15行和17行的数据。
```

```r
#其使用这种方法导入的是数据框。
pre_data<-read.xlsx("b.xlsx",sheet=1,colNames = TRUE,
rowNames = TRUE)
head(pre_data) #查看前5行的数据。

#进行加法替换法的工作。
#使用循环来将每一行的缺失值进行替换。

del_ta<-0.65 #首先先定义常量delta的数值为0.65.
str(pre_data) #查看数据框的行和列
n=67 #n表示的是数据框的行。(手动输入)
m=14 #m表示的是数据框的列。(手动输入)
z=0 #初始化z.
#sum(is.na(pre_data[1,]))


for (i in 1:n)
{
  z=sum(pre_data[i,]==0)
  r_x=del_ta*(1+z)*(m-z)/m^2
  pre_data[i,][pre_data[i,]==0]<-r_x
  #pre_data[i,][is.na(pre_data[i,])]<-r_x
  z=0 #再次初始化。
}
write.xlsx(pre_data,file = '处理缺失值后的数据.xlsx',
colNames = TRUE,
rowNames = TRUE)
```

五、 描述统计量代码（r 语言）

```r
install.packages("chemometrics")
install.packages("compositions")
library(ggplot2)
options(digits = 3) #设置保留几位小数。
#setwd( "E:/R/eva") 改变工作路径
#getwd()  #查看工作路径

a<-read.table("clipboard") #导入数据。
a

# 传入数据

names(a)<-c("Si02","Na20","K20","Ca0",
"Mg0","Al203","Fe203","Cu0",
"Pb0","Ba0","P205","Sr02",
```

```
"Sn02","S02")

# 将数据框赋予名字
a
Lrt<-a
#Lrt<-apply(a,1,clr)
#Lrt
# 将数据进行对数比变换
#Lrt<-a
#Lrt
#summ<-apply(Lrt,2,sum)
#summ
#进行求和验证


average<-apply(Lrt,2,mean)
average                 #求平均值


max1<-apply(Lrt,2,max)
max1                    #求出最大值


min1<-apply(Lrt,2,min)
min1                    #求出最小值


sd1<-apply(Lrt,2,sd)
sd1                     #求标准差



Coefficient<-function(Lrt)
{
  sd(Lrt)/mean(Lrt)     #变异系数函数
}

Coefficient_<-apply(Lrt,2,Coefficient)
Coefficient_          #求变异系数

Skewness<-function(Lrt)
{

  return (mean(((Lrt-mean(Lrt))/sd(Lrt))^3)) #偏度函数

}
```

```
Skewness_1<-apply(Lrt,2,Skewness) #求偏度
Skewness_1

Kurtosis<-function(Lrt)
{
  mean(((Lrt-mean(Lrt))/sd(Lrt))^4)-3 #峰度函数
}
Kurtosis_<-apply(Lrt,2,Kurtosis)
Kurtosis_                        #求峰度


f<-data.frame(average,
max1,
min1,
sd1,
Coefficient_ ,
Skewness_1,
Kurtosis_
)
f
names(f) = col.names = c('平均数','最大值','最小值','标准差',
'变异系数','偏度','峰度')
f<-round(f,3)
write.xlsx(f,'E:/R/eva/描述统计风化.xlsx',
fileEncoding = 'utf-8',colNames = TRUE,
rowNames = TRUE)
```

六、 箱线图 (python)

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
data1 = pd.read_excel('./data/风化.xlsx')
#lables =
    ['二氧化硅','氧化钠','氧化钾','氧化钙','氧化镁','氧化铝','氧化铁','氧化铜','氧化铅','氧化钡','五氧化二磷',
#'氧化锶','氧化锡','二氧化硫']
a=data1.iloc[1:,1:]
print(a)
plt.figure(figsize=(10, 6),dpi=300)
plt.rcParams['font.sans-serif'] = 'SimHei'
b =
    ['二氧化硅','氧化钠','氧化钾','氧化钙','氧化镁','氧化铝','氧化铁','氧化铜','氧化铅','氧化钡','五氧化二磷',
'氧化锶','氧化锡','二氧化硫']
plt.boxplot(a, #绘制箱线图的范围和值
notch=True, #是否展示中位数的V型凹槽
```

```
labels=b, meanline=True, #是否展示均值
vert=False, #True竖放 False横放
# width=0.18, #设置箱体宽度
patch_artist=True, #自动填充颜色
showmeans=True, #显示均值
capprops={'color':'hotpink', 'linewidth':2, 'linestyle':'solid'},
boxprops={'color':'crimson', 'linewidth':2, 'linestyle':'--',
  'facecolor':'hotpink', 'alpha':0.3},
whiskerprops={'color':'palevioletred', 'linewidth':1.5, 'linestyle':':'},
flierprops={'marker':'o', 'markerfacecolor':'salmon', 'markersize':6,
  'linestyle':'none'},
medianprops={'color':'violet', 'linewidth':1.2},
meanprops={'color':'deeppink', 'linestyle':'--'})

#plt.title('无风化箱线图')
plt.savefig('风化箱线图.png',)
plt.show()
```