

实验设计与数据处理

石大川 深计研 211 班
<sdc21@mails.tsinghua.edu.cn>

2022 年 11 月 21 日

目录

1 实验背景	3
2 实验设计	3
2.1 作业要求	3
2.2 实验介绍	3
2.3 正交实验设计	3
2.3.1 极差分析	5
2.3.2 方差分析	5
2.3.3 使用 DesignExpert 软件验证	7
2.3.4 正交实验设计总结	11
3 数据的预处理和统计分析	12
3.1 作业要求	12
3.2 实验介绍	12
3.3 数据的预处理	12
3.3.1 均向量	12
3.3.2 协方差矩阵	14
3.3.3 相关系数矩阵	14
3.3.4 变量的直方图	15
3.3.5 变量的箱式图	15
3.3.6 对各变量的异常数据进行判断和取舍	18
3.3.7 数据的预处理实验总结	21
3.4 统计分析	21
3.4.1 数据的分布检验	21
3.4.2 参数估计之区间估计	22
3.4.3 参数估计之样本容量确定	23
3.4.4 统计检验之离群值检验	23
3.4.5 统计检验之方差比较检验	24
3.4.6 统计检验之均值检验	24
3.4.7 方差分析与极差分析	25
3.4.8 统计分析实验总结	25

4 数据的图示	25
4.1 作业要求	25
4.2 实验介绍	26
4.3 数据图示	26
4.3.1 散点图	26
4.3.2 折线图	27
4.3.3 条形图	31
4.3.4 饼图	31
4.3.5 数据的图示实验总结	32
5 数据的处理作业一	33
5.1 作业要求	33
5.2 实验介绍	34
5.3 回归分析	34
5.3.1 两变量线性回归	34
5.3.2 两变量线性显著性检验	35
5.3.3 两变量线性回归结果讨论	36
5.3.4 多元线性回归	37
5.3.5 多元线性回归显著性检验	38
5.3.6 多元线性回归结果讨论	39
6 数据的处理作业二	39
6.1 作业要求	39
6.2 实验介绍	40
6.3 降维分析	40
6.3.1 降维问题	40
6.3.2 降维方法	40
6.3.3 降维过程	41
6.3.4 选择最终降维成分的依据	43
6.3.5 降维后新变量与原始变量之间关系	43
6.4 聚类分析	44
6.4.1 聚类问题	44
6.4.2 聚类方法	45
6.4.3 聚类过程	46
6.4.4 聚类结果分析	46
7 参考文献	53

1 实验背景

本人来自深圳国际研究生院的计算机技术专业，所在课题组的研究方向为计算机视觉。本人选取了图像分类这一计算机视觉领域具体的科研问题来进行实验的设计与数据处理。作为计算机视觉领域最经典的问题之一，图像分类的目的是训练一个模型，该模型可以将图像作为输入，并输出图像相应的类别，如：汽车，飞机，猫，狗等等。本次实验选择了 ImageNet[1] 作为数据集，它是计算机视觉领域最为广泛使用的数据集之一。作为大规模的图像分类数据集，ImageNet 包含 1281167 张图像作为训练集，50000 张图像作为验证集，100000 张图像作为测试集。此外，本次实验选择了 ResNet[2] 这一计算机视觉领域引用次数最高的神经网络作为模型。具体来说，本文以 ImagNet 为数据集，将分辨率为 224×224 ，也即维度数为 50176 的图像数据作为输入，训练 ResNet18 这一网络（ResNet18 是 ResNet 网络的一种），该网络约有 1169 万的参数量。

2 实验设计

2.1 作业要求

- 对于欲研究的课题，应该采用下面实验设计方法之一采集数据，并用配套的方法分析处理数据。
 - 正交实验设计
 - 随机化（区组）实验设计
 - 析因（因子）实验设计
 - 响应曲面实验设计
 - 均匀设计
 - 其他实验设计方法
- 如果所研究实验对象确实不适合上述实验设计方法，则应采用全面实验设计方法（进行全部可能的因素的组合）并确保每种实验要重复 3 次及 3 次以上，不要使用一次改变一个因素的单因素对比法。

2.2 实验介绍

为了尽可能训练出一个准确率高的图像分类模型，我们通常需要反复调整模型中的各项超参数（hyperparameter），其中训练轮数（epoch），批处理大小（batch size）和学习率（learning rate）是三项关键的参数。在接下来的实验设计中，我们通过正交实验设计的方法，来确定关于这三个超参数的最优水平组合。

2.3 正交实验设计

相比于全面实验，正交实验具有减少实验次数，并且可以给出误差分析的优势，因此我们选择正交实验设计来进行相关实验。对于训练轮数，批处理大小和学习率我们根据经验各设置了 3 个常见的水平，如表 1 所示。为了考察实验误差，我们还需要额外的 1 列空列。由四因素三水平可知，最合适的选择为 $L_9(3^4)$ ，如表 2 所示。对照表 2，可以写出本次实验的正交实验设计表表 3。

表 1: 因素水平表

因素 水平 \	A: 训练轮数	B: 批处理大小	C: 学习率
1	16	8192	0.01
2	12	4096	0.005
3	8	2048	0.001

表 2: $L_9(3^4)$ 表

实验号	列号			
	1	2	3	4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

表 3: 正交实验设计表

因素 实验号 \ 列号	A 训练轮数	B 批量大小	C 学习率	准确率 (%)
1	1 (16)	1 (8192)	1 (0.01)	59.80
2	1	2 (4096)	2 (0.005)	61.22
3	1	3 (2048)	3 (0.001)	56.86
4	2 (12)	1	2	53.76
5	2	2	3	44.28
6	2	3	1	56.34
7	3 (8)	1	3	40.61
8	3	2	1	53.07
9	3	3	2	55.23

2.3.1 极差分析

我们需要计算 T 和 R 以确定最优水平以及主次排序。首先计算 T_{ij} , 其代表第 j 列中水平为 i 的实验结果之和。例如

$$T_{11} = 59.80 + 61.22 + 56.86 = 177.88$$

同理可以计算得到其他的 T_{ij} , 如表 4 所示。此后通过

$$T_{ij} = \max(T_{1j}, T_{2j}, T_{3j})$$

可以得到最优水平, 如 A 因素训练轮数的最优水平为

$$T_{11} = \max(T_{11}, T_{21}, T_{31})$$

也即 $A1(16)$ 。此外通过

$$R_j = \max(T_{1j}, T_{2j}, T_{3j}) - \min(T_{1j}, T_{2j}, T_{3j})$$

可以计算得到极差, 如 A 因素训练轮数的极差为

$$R_1 = \max(T_{11}, T_{21}, T_{31}) - \min(T_{11}, T_{21}, T_{31}) = 177.88 - 148.91 = 28.97$$

将 R_j 从大到小排序, 可以得到主次排序为 ACB , 且最优水平组合为 $A1C2B3$ 。而 $A1C2B3$ 并不是正交实验做的 9 个实验之一, 可见虽然正交实验设计只做了局部实验, 但是可以得到全局结果。

使用最优水平组合 $A1C2B3$ 再次进行实验, 可以得到 63.31 的准确率, 其结果高于此前 9 次实验中最佳组合 $A1C2B2$ 的 61.22。因此, 最优水平组合 $A1C2B3$ 得到了检验。

此外, 对比三个因素的 R_j 和空列的 R_4 可知, 三个因素的 R_j 都大于空列的 R_4 , 可以认为这 3 个因素对实验结果都存在着一定的影响。

2.3.2 方差分析

极差分析具有简单直观且计算量小的优点, 但是也存在着无法区分实验因素、水平引起的变化和实验误差引起的变化, 以及不能精确估计各因素对实验结果影响的重要程度等缺点, 因此我们进一步使用方差分析处理实验结果。

首先根据表 4, 可以计算出第 j 列水平为 i 的结果平均值:

$$t_i = \frac{T_i}{m}$$

计算结果如表 5 所示。

根据公式

$$\begin{aligned} SS_i &= \frac{n}{m} \sum_{i=1}^m (t_i - \bar{x})^2 \\ f_i &= m - 1 \end{aligned}$$

可以得到各个因素的平方和与自由度, 如对于因素 A 迭代轮数有

$$SS_1 = \frac{9}{3} [(59.29 - 53.46)^2 + (51.46 - 53.46)^2 + (49.64 - 53.46)^2] = 157.74$$

$$f_1 = 3 - 1 = 2$$

进而根据公式

表 4: 极差分析表

因素	A 训练轮数	B 批量大小	C 学习率		准确率 (%)
实验号 \ 列号	1	2	3	4	
1	1 (16)	1 (8192)	1 (0.01)	1	59.80
2	1	2 (4096)	2 (0.005)	2	61.22
3	1	3 (2048)	3 (0.001)	3	56.86
4	2 (12)	1	2	3	53.76
5	2	2	3	1	44.28
6	2	3	1	2	56.34
7	3 (8)	1	3	2	40.61
8	3	2	1	3	53.07
9	3	3	2	1	55.23
T_{1j}	177.88	154.17	169.21	159.31	
T_{2j}	154.38	158.57	170.21	158.17	
T_{3j}	148.91	168.43	141.75	163.69	
最优水平	A1 (16)	B3 (2048)	C2 (0.005)		
R_j	28.97	14.26	28.46	5.52	
主次排序	ACB				

表 5: t_i 计算表

因素	A 训练轮数	B 批量大小	C 学习率		准确率 (%)
实验号 \ 列号	1	2	3	4	
1	1 (16)	1 (8192)	1 (0.01)	1	59.80
2	1	2 (4096)	2 (0.005)	2	61.22
3	1	3 (2048)	3 (0.001)	3	56.86
4	2 (12)	1	2	3	53.76
5	2	2	3	1	44.28
6	2	3	1	2	56.34
7	3 (8)	1	3	2	40.61
8	3	2	1	3	53.07
9	3	3	2	1	55.23
T_{1j}	177.88	154.17	169.21	159.31	平均值 53.46
T_{2j}	154.38	158.57	170.21	158.17	
T_{3j}	148.91	168.43	141.75	163.69	
t_{1j}	59.29	51.39	56.40	53.10	
t_{2j}	51.46	52.86	56.74	52.72	
t_{3j}	49.64	56.14	47.25	54.56	

$$MS_i = \frac{SS_i}{f_i}$$

可以得到各个因素的均方差，如对于因素 A 迭代轮数有

$$MS_1 = \frac{SS_1}{f_1} = \frac{157.74}{2} = 78.87$$

对其他因素进行同样的计算，可以得到表 6 所示的方差分析表。需要注意的是，还需要比较所有因素的 MS_i 与 MSE ，如果 $MS_i < 2MSE$ ，则需要将 SS_i , f_i 和 MS_i 加入到误差项中，从而可以得到误差更正后的方差分析表，如表 7 所示。

表 6: 方差分析表（误差更正前）

方差来源	平方和	自由度	均方差
1 训练轮数	157.74	2	78.87
2 批量大小	35.48	2	17.74
3 学习率	173.90	2	86.95
空列	5.66	2	2.83
误差	5.66	2	2.83

表 7: 方差分析表（误差更正后）

方差来源	平方和	自由度	均方差
1 训练轮数	157.74	2	78.87
2 批量大小	35.48	2	17.74
3 学习率	173.90	2	86.95
空列	5.66	2	2.83
误差	5.66	2	2.83
总和	372.78	8	

此后根据公式

$$F_i = \frac{MS_i}{MSE}$$

可以计算得到 F 值，如对于因素 A 训练轮数有

$$F_1 = \frac{MS_1}{MSE} = \frac{78.87}{2.83} = 27.87$$

通过查阅 F 临界值表，可以得到最终的方差分析表 8。可见因素训练轮数和因素学习率的 F 值都超过了 $F_{0.05}(2, 2)$ ，但是没有超过 $F_{0.01}(2, 2)$ ，因此可以认为因素训练轮数和因素学习率都是显著的。而因素批量大小的 F 值没有超过 $F_{0.05}(2, 2)$ ，因此可以认为因素批量大小不显著。

此外，通过方差分析给出的主次顺序为 CAB，这与极差分析给出的主次顺序 ACB 在第一次序和第二次序上有所不同。

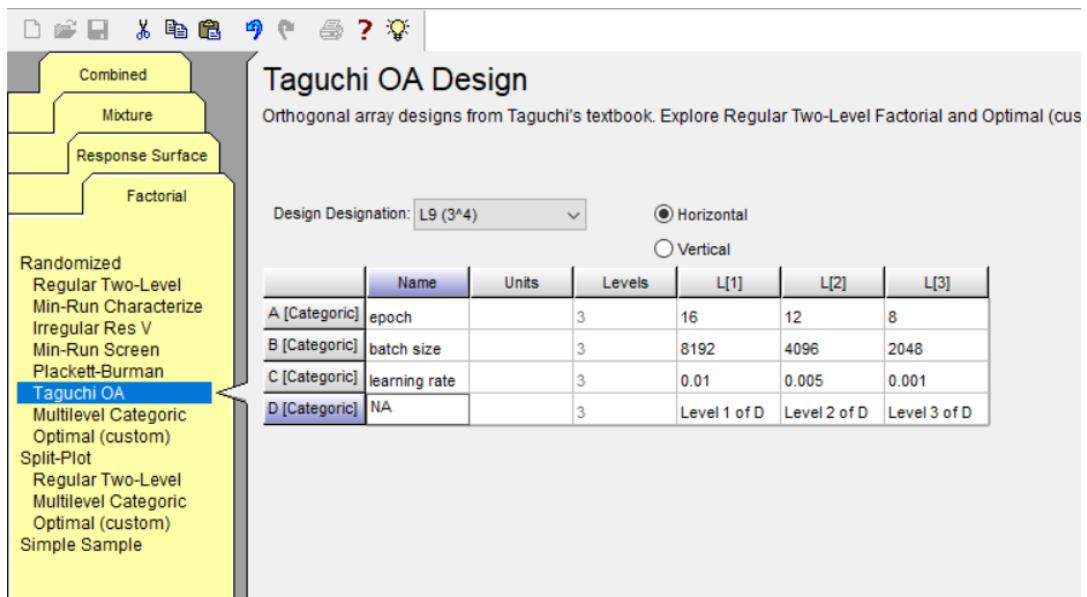
2.3.3 使用 DesignExpert 软件验证

在 DesignExpert 软件中选择 Taguchi OA (田口正交阵列)，进行正交实验的设计和分析。选择 $L_9(3^4)$ 正交表后将 3 个因素和各自的 3 个水平输入到因素水平表中，如图 1 所示。

表 8: 方差分析表

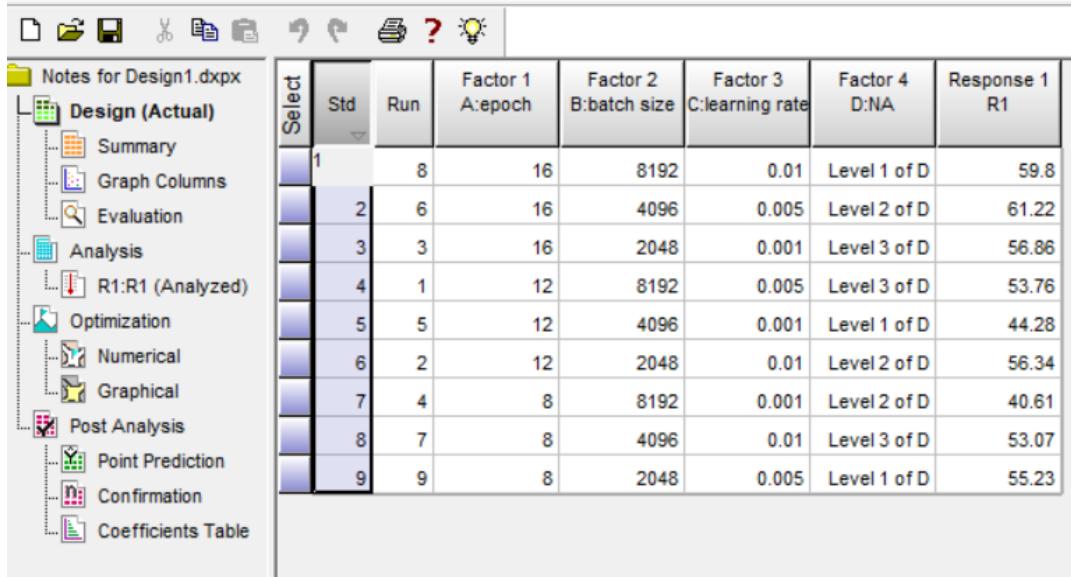
方差来源	平方和	自由度	均方差	F 值	F_α	显著性
1 训练轮数	157.74	2	78.87	27.87	$F_{0.05}(2, 2) = 19.00$	*
2 批量大小	35.48	2	17.74	6.27		
3 学习率	173.90	2	86.95	30.72	$F_{0.05}(2, 2) = 19.00$	*
空列	5.66	2	2.83			
误差	5.66	2	2.83			
总和	372.78	8				

Figure 1: 因素水平表



接下来，由于正交设计不适合分析交互作用和高阶，我们使用默认的只分析一阶效应的选项。在生成实验设计表后，填入表 3 中的实验数据，如图 2 所示。

Figure 2: 正交实验设计表

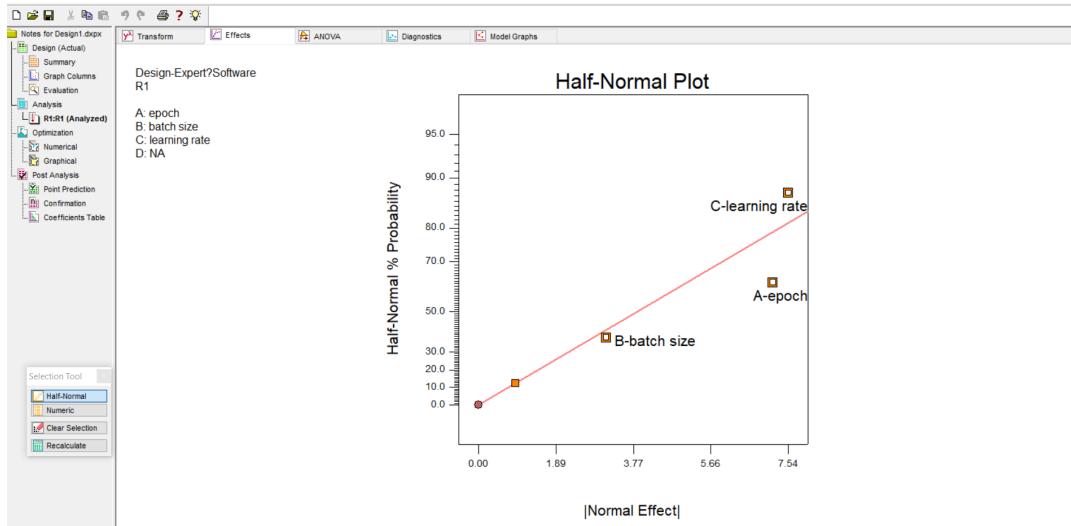


The screenshot shows the 'Design (Actual)' section of the software interface. On the left, there is a tree view with nodes like 'Summary', 'Graph Columns', 'Evaluation', 'Analysis', 'R1:R1 (Analyzed)', 'Optimization', 'Numerical', 'Graphical', 'Post Analysis', 'Point Prediction', 'Confirmation', and 'Coefficients Table'. The main area displays a table with columns: 'Select', 'Std', 'Run', 'Factor 1 A:epoch', 'Factor 2 B:batch size', 'Factor 3 C:learning rate', 'Factor 4 D:NA', and 'Response 1 R1'. The data rows are numbered 1 through 9.

Select	Std	Run	Factor 1 A:epoch	Factor 2 B:batch size	Factor 3 C:learning rate	Factor 4 D:NA	Response 1 R1
1		8	16	8192	0.01	Level 1 of D	59.8
2		6	16	4096	0.005	Level 2 of D	61.22
3		3	16	2048	0.001	Level 3 of D	56.86
4		1	12	8192	0.005	Level 3 of D	53.76
5		5	12	4096	0.001	Level 1 of D	44.28
6		2	12	2048	0.01	Level 2 of D	56.34
7		4	8	8192	0.001	Level 2 of D	40.61
8		7	8	4096	0.01	Level 3 of D	53.07
9		9	8	2048	0.005	Level 1 of D	55.23

在 Effect 页面，可以发现在 3 个实验因素中，因素 A 训练轮数和因素 C 学习率效果较明显，而因素 B 批处理大小效果较不明显，如图 3 所示。图中离直线越远说明效果越明显。

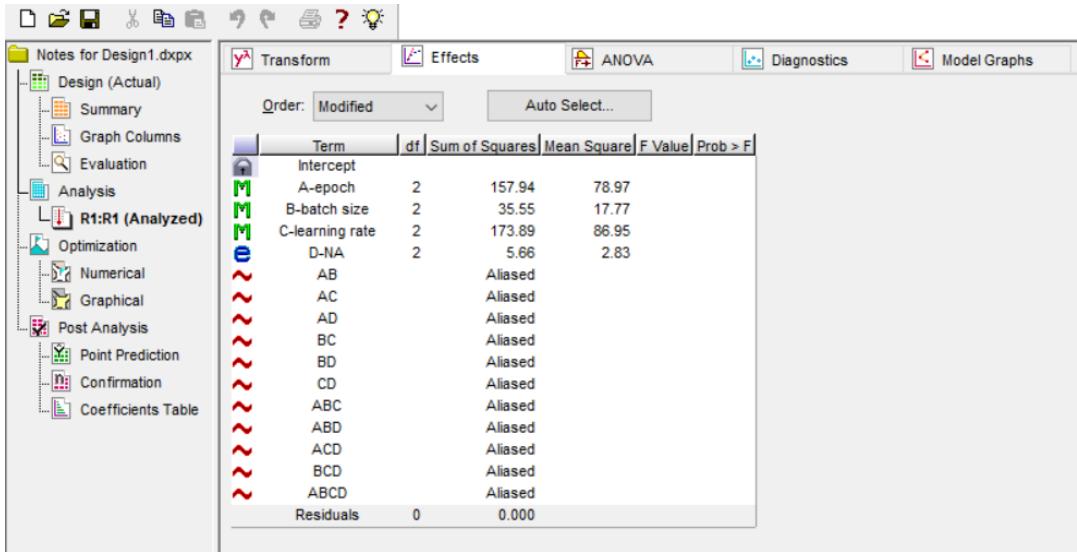
Figure 3: Effect 页面



在 Selection Tool 中选择 Numeric 可以看到 DesignExpert 软件计算出的各因素的平方和，自由度和均方差，如图 4 所示。将其结果与表 7 的计算结果对比可知，在舍入误差允许的范围内，可以认为计算结果是正确的。

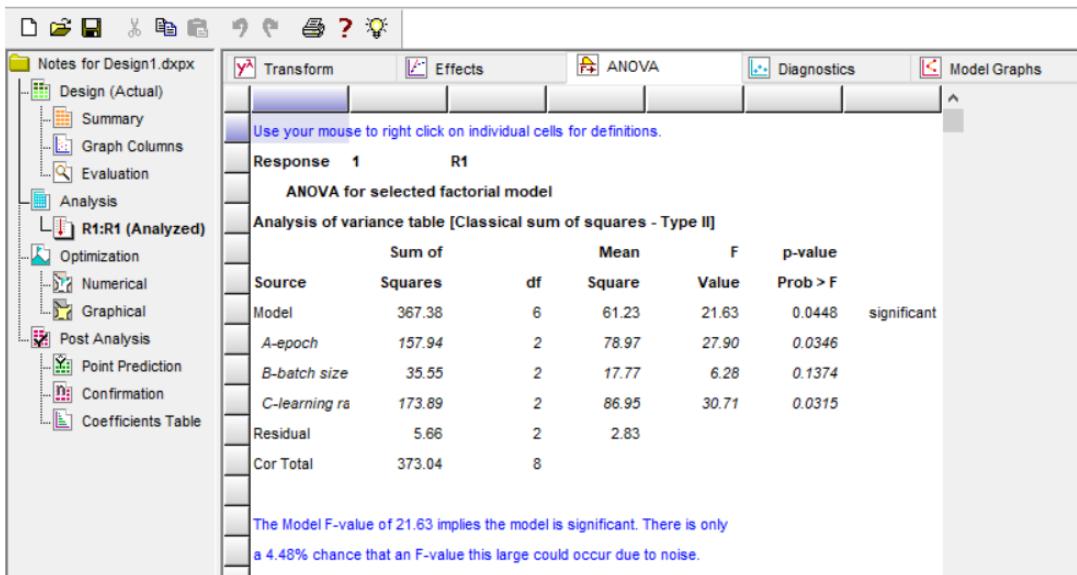
最后选择 ANOVA 页面，可以看到 DesignExpert 软件完成的方差分析，图 5 所示。对比表 8 的计算结果可知，在舍入误差允许的范围内，可以认为计算结果是正确的。此外

Figure 4: 数值计算结果



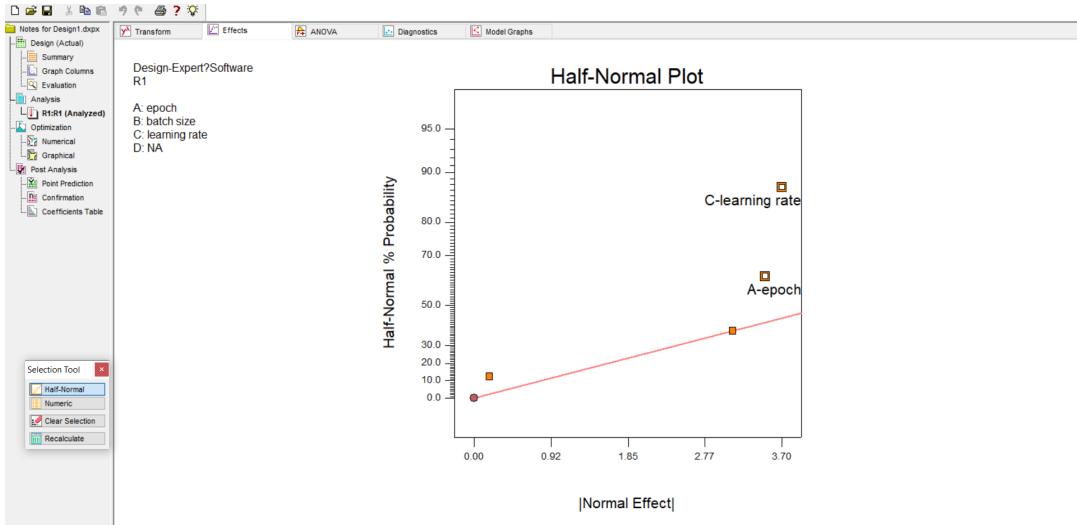
DesingExpert 软件也给出了训练轮数，批处理大小和学习率三个因素构成的模型总体上显著的结论。具体来说，对于因素 C 学习率，只有 $3.15\% < 5\%$ 的概率因素 C 学习率的 F 值偏大是因为噪声，因此因素 C 学习率是显著的；对于因素 A 训练轮数，只有 $3.46\% < 5\%$ 的概率因素 A 训练轮数的 F 值偏大是因为噪声，因此因素 A 训练轮数是显著的；对于因素 B 批处理大小，有 $13.74\% < 5\%$ 的概率因素 B 批处理大小的 F 值偏大是因为噪声，因此因素 B 批处理大小不显著；总体而言，由训练轮数，批处理大小和学习率三个因素构成的模型只有 $4.48\% < 5\%$ 的概率其 F 值偏大是因为噪声，因此三个因素构成的模型总体上是显著的。

Figure 5: 方差分析结果



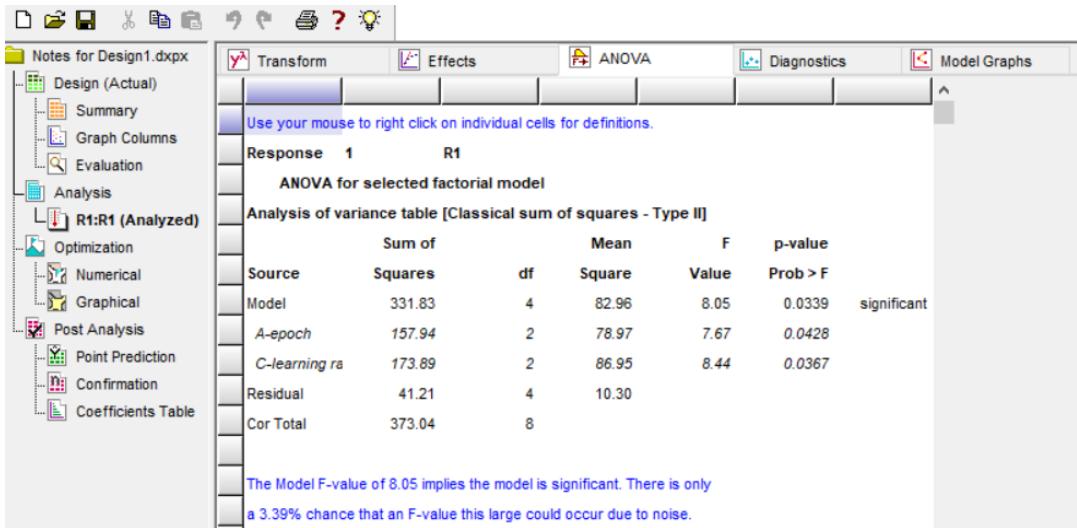
鉴于我们发现因素 B 批处理大小不显著，我们在 Effect 中手动去掉因素 B 批处理大小，如图 6 所示。

Figure 6: Effect 页面 (不选择因素 B 批处理大小)



相应地，只包含因素 C 学习率和因素 A 训练轮数的方差分析结果如图 7 所示。

Figure 7: 方差分析结果 (只包含因素 C 学习率和因素 A 训练轮数)



可见两因素模型总体上仅有 3.39% 的概率其 F 值偏大是因为噪声，比三因素分析中总体 4.48% 的概率更加显著。

2.3.4 正交实验设计总结

首先，对于三因素三水平的实验，如果采用全面实验法需要进行 27 次实验，而使用正交实验法则仅需要 9 次实验，节省了 $27 - 9 = 18$ 次实验的工作量。

从实验结果的极差分析可知，虽然正交实验只做了局部实验，但是可以得到全局结果。根据实验验证，由正交实验设计推理出的最优水平组合确实比 9 次正交实验中的最佳组合效果更好。

从实验结果的方差分析可知，因素 C 学习率和因素 A 训练轮数是显著因素，在训练

模型时我们应当更加关注这两个因素。而因素 B 批处理大小是不显著因素，在训练模型时我们不必过多关注这个因素。

此外需要说明的是，本次实验得到的关于三个因素各自显著性的结论只针对在 ImageNet 数据集上训练的 ResNet18 模型成立，对于其他的数据集和模型不一定成立。尤其是对于批处理大小这个因素，在我们的实验中其效果不显著的原因可能是，为了快速训练完模型，我们选择了超大的批处理大小（2048 ~ 8192），然而当批处理大小较小（如 < 100）时，有诸多研究表明批处理大小对于实验结果会产生显著影响。

3 数据的预处理和统计分析

3.1 作业要求

- 样本数量不得少于 30 个，各样本数据维度不得少于二维
- 对所获得数据进行数据预处理，给出样本的
 - 均向量
 - 协方差矩阵
 - 相关系数矩阵
 - 变量的直方图，箱式图
 - 对各变量的异常数据进行判别和取舍
- 对预处理的数据完成以下统计分析之一(四选一)，并简述分析的具体过程、方法以及检验(或分析方法)的意义。
 - 数据的分布检验
 - 参数估计
 - 统计检验
 - 方差分析(或极差分析)

3.2 实验介绍

在数据的预处理和统计分析中，我们从 ImageNet 数据集中随机选择了 10000 张图片做为实验数据。对于每张图片，通过图像处理可以获得表 9 中的 7 项特征。因此总体的实验数据为 10000×7 的矩阵，部分特征数据如图 8 所示。

3.3 数据的预处理

3.3.1 均向量

设 x_{ij} 代表实验数据中第 i 个样本的第 j 个特征的值， n 代表实验的样本量。则第 j 个特征的平均值可以根据

$$\bar{x}_j = \sum_{i=1}^n x_{ij}$$

得到。经计算可得实验数据的均向量为：

$$\bar{x} = [120.61 \quad 114.73 \quad 102.39 \quad 115.08 \quad 53.34 \quad 84.22 \quad 131.77]$$

表 9: 实验数据的 7 项特征

特征	含义	取值范围
R	Red: 图像在 RGB 色彩空间的 R 分量	[0,255]
G	Green: 图像在 RGB 色彩空间的 G 分量	[0,255]
B	Blue: 图像在 RGB 色彩空间的 B 分量	[0,255]
Gr	Gray: 图像的灰度分量	[0,255]
H	Hue: 图像在 HSV 色彩空间的色相分量	[0,180]
S	Saturation: 图像在 HSV 色彩空间的饱和度分量	[0,255]
V	Value: 图像在 HSV 色彩空间的明度分量	[0,255]

Figure 8: 部分特征数据图示

T	R T	G T	B T	Gr T	H T	S T	V T
0	11.07	12.87	12.74	12.27	73.96	48.5	14.02
1	150.38	145	142.56	146.34	60.87	23.38	152.41
2	153.68	118.54	127.97	130.13	131.41	84.32	156.54
3	177.35	197.77	182.67	189.95	39.85	68.4	201.93
4	126.92	130.92	138.71	130.63	89.34	40.94	142.19
5	136.63	135.66	125.44	134.78	73.68	58.63	146.59
6	159.73	130.35	104.71	136.27	14.72	105.08	159.77
7	90.63	107.35	164.38	108.85	87.68	118.14	171.94
8	145.04	124.97	124.2	130.88	96.73	57.95	146.46
9	77.14	81.24	40.51	75.35	38.85	186.32	87.5
10	120.7	152.18	180.34	145.96	93.77	87.04	182.75
11	157.38	130.83	108.36	136.22	52.33	90.07	163.01
12	6.64	20.49	3.61	14.33	52.66	200.51	20.6
13	143.95	158.15	175.51	155.86	98.67	46.15	176.84
14	34.63	121.56	118.93	95.28	75.11	183.28	129.29
15	120.61	118.19	121.88	119.34	73.9	77.22	131.37
16	65.36	83.25	99.92	79.8	53.53	112.39	114.54
17	87.47	113.19	129.7	107.38	87.18	86.17	131.19
18	112.17	132.05	81.26	120.31	49.65	110.96	136.57
19	117.2	119.3	120.73	118.84	73.39	34.88	126.42
20	112.44	110.27	107.58	110.61	66.02	42.55	116
21	137.92	129.16	111.79	129.81	38.62	63.15	140.12
22	150.93	146.55	129.27	145.88	34.56	42.34	152.32
23	120.77	112.4	97.79	113.26	21.65	63	121.26
24	145.71	130.86	86.18	130.21	46.26	139.55	152.84
25	118.49	109.13	93.76	110.18	70.66	88.03	124.3
26	149.45	145.62	143.68	146.56	44.7	26.14	150.82
27	166.98	155.49	134.23	156.52	18.6	48.88	167.22
28	176.89	157.5	155.58	163.08	82.74	56.61	181.73

3.3.2 协方差矩阵

任意两个特征之间的协方差可以根据

$$\text{cov}(x_j, x_{j'}) = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{n-1}$$

得到，其中 j 和 j' 代表 $\text{cov}(x_j, x_{j'})$ 是协方差矩阵中第 j 行第 j' 列的元素。经计算可得实验数据的协方差矩阵为：

$$\Sigma = \begin{bmatrix} 1370.05 & 1085.61 & 955.77 & 1155.93 & -149.01 & -465.24 & 1176.62 \\ 1085.61 & 1253.07 & 1222.92 & 1199.58 & 38.43 & -633.86 & 1152.69 \\ 955.77 & 1222.92 & 1654.07 & 1192.19 & 369.7 & -1024.59 & 1167.12 \\ 1155.93 & 1199.58 & 1192.19 & 1185.72 & 20.08 & -628.02 & 1161.52 \\ -149.01 & 38.43 & 369.7 & 20.08 & 697.04 & -133.64 & 75.12 \\ -465.24 & -633.86 & -1024.59 & -628.02 & -133.64 & 2029.85 & -228.16 \\ 1176.62 & 1152.69 & 1167.12 & 1161.52 & 75.12 & -228.16 & 1300.92 \end{bmatrix}$$

3.3.3 相关系数矩阵

相比于协方差矩阵，相关系数矩阵考虑了各个特征的数值在尺度上的差异，其元素取值范围为 $[-1, 1]$ ，其值越大代表正相关性越强，其值越小代表负相关性越强。相关系数矩阵可以根据

$$\text{corr}(x_j, x_{j'}) = \frac{\text{cov}(x_j, x_{j'})}{\sqrt{\text{cov}(x_j, x_j)} \times \sqrt{\text{cov}(x_{j'}, x_{j'})}}$$

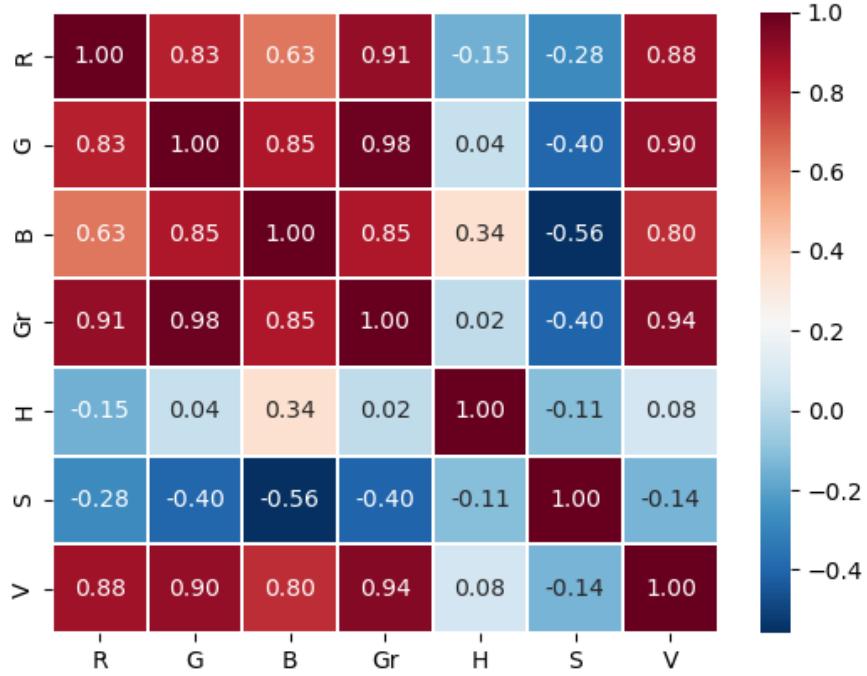
得到，其中 j 和 j' 代表 $\text{corr}(x_j, x_{j'})$ 是相关系数矩阵中第 j 行第 j' 列的元素。经计算可得实验数据的相关系数矩阵为：

$$R = \begin{bmatrix} 1.00 & 0.83 & 0.63 & 0.91 & -0.15 & -0.28 & 0.88 \\ 0.83 & 1.00 & 0.85 & 0.98 & 0.04 & -0.40 & 0.90 \\ 0.63 & 0.85 & 1.00 & 0.85 & 0.34 & -0.56 & 0.80 \\ 0.91 & 0.98 & 0.85 & 1.00 & 0.02 & -0.40 & 0.94 \\ -0.15 & 0.04 & 0.34 & 0.02 & 1.00 & -0.11 & 0.08 \\ -0.28 & -0.40 & -0.56 & -0.40 & -0.11 & 1.00 & -0.14 \\ 0.88 & 0.90 & 0.80 & 0.94 & 0.08 & -0.14 & 1.00 \end{bmatrix}$$

该相关系数矩阵的热力图如图 9 所示。从相关系数矩阵及其热力图可以得到以下结论：

- R, G, B, Gr 和 V 这五个特征之间具有较强的正相关性。
- H 和 S 这两个特征之间，以及与其他变量之间的相关性都较弱。
- 特征两两之间除自相关外，最强的正相关性出现在特征 G 和 Gr 之间，相关系数为 0.98。
- 特征两两之间除自相关外，最强的负相关性出现在特征 B 和 S 之间，相关系数为 -0.56。
- 特征两两之间最弱的相关性出现在特征 Gr 和 H 之间，相关系数的绝对值仅为 0.02。

Figure 9: 相关系数矩阵的热力图



3.3.4 变量的直方图

对于每个特征，我们分别绘制出它的直方图。分别如图 10，图 11，图 12 和图 13 所示。从这些直方图中可以发现，特征 R, G, Gr 和 V 呈正态分布；特征 B 呈轻微右偏分布；特征 H 和 S 呈明显右偏分布。

3.3.5 变量的箱式图

箱式图可以反映数据的分布情况。每个箱型中包含 5 条横线，从上到下依次代表上边缘，上四分位数，中位数，下四分位数和下边缘。此外处于上边缘以上和下边缘以下的数据点我们将它们当作异常值。为了方便综合比较，我们将七个特征的箱式图绘制在一张图中，如图 14 所示。

从箱式图中可以看出：

- 七个特征的箱体都较窄，说明正常数据在中位数附近较为集中。
- R, G, B, Gr 和 V 这五个特征的分布较为接近。
- H 和 S 两个特征之间，以及与其他特征之间的分布相差较大。
- H 和 S 两个特征的中位数分别更靠近它们箱体的下沿，呈右偏分布。
- 七个特征都有异常值，其中特征 B, H 和 S 只有大于上边缘的异常值，而没有小于下边缘的异常值。

Figure 10: 特征 R 和特征 G 的直方图

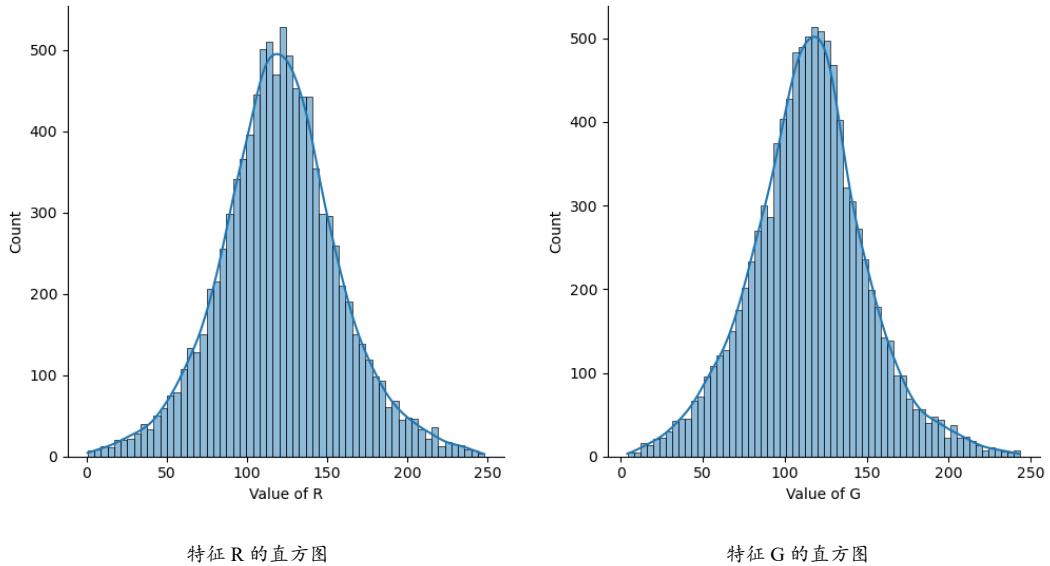


Figure 11: 特征 B 和特征 Gr 的直方图

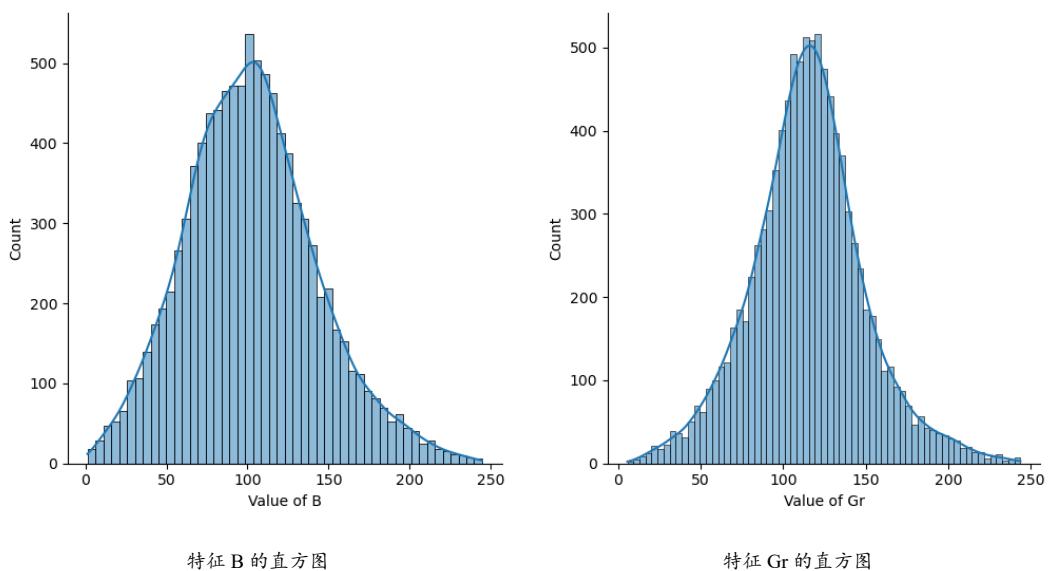


Figure 12: 特征 H 和特征 S 的直方图

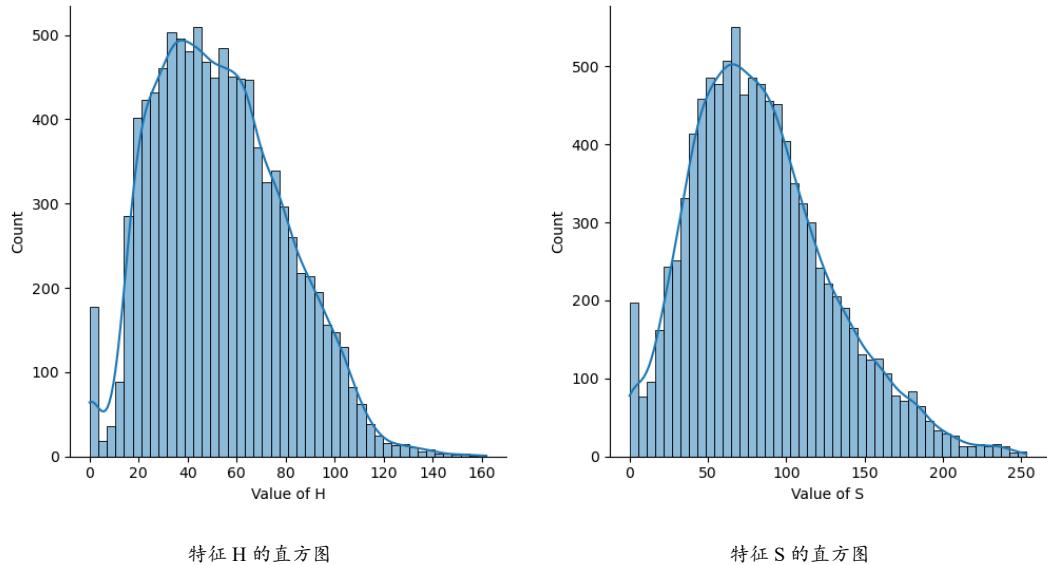


Figure 13: 特征 V 的直方图

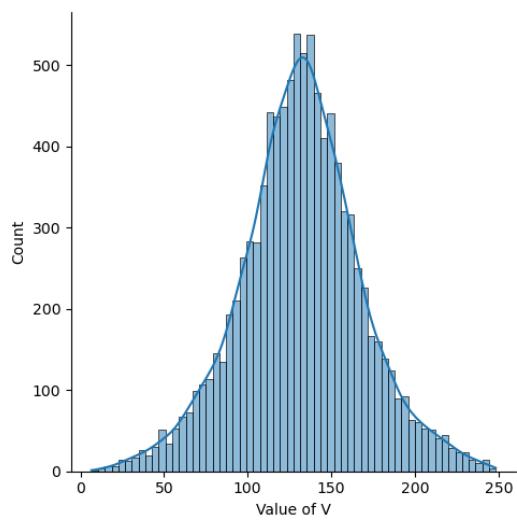
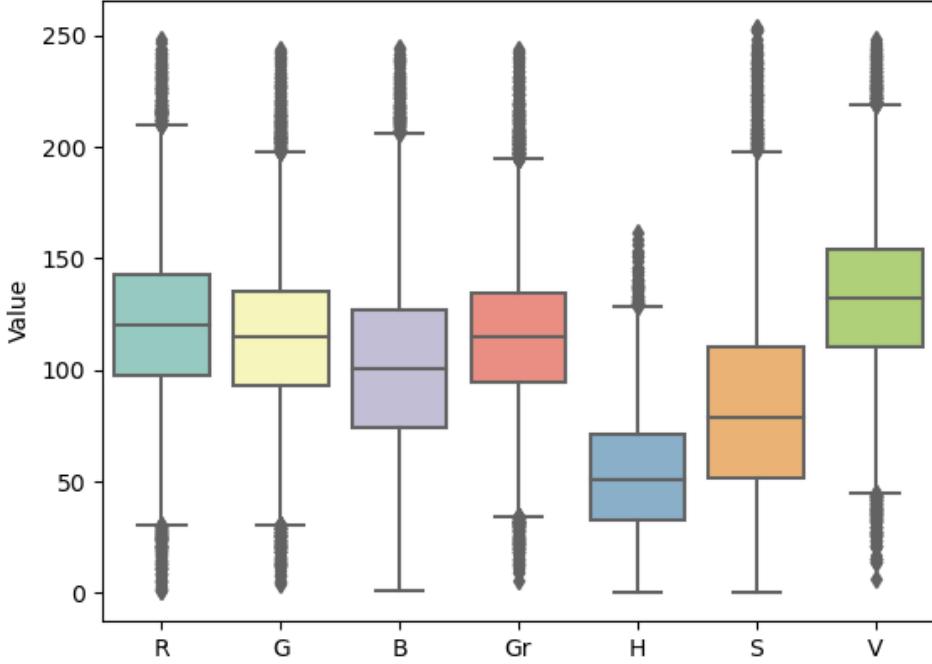


Figure 14: 七个特征的箱式图



3.3.6 对各变量的异常数据进行判断和取舍

有多种方法可以对异常数据进行判断和取舍，如散点图法，箱型图法和置信区间法等等。

散点图法 我们先做出数据的散点图，将图中明显不符合其他数据模型的点判定为异常数据。以特征 H 为例，其散点图如图 15 所示。从图中可以看出，图的顶部有数个明显离群的点，我们可以认为这些点是异常数据。

箱型图法 我们先做出数据的箱型图，记上四分位数为 Q_3 ，下四分位数为 Q_1 ，上下四分位数的差值为 IQR ，则我们将大于 $Q_3 + 1.5IQR$ 或者小于 $Q_1 - 1.5IQR$ 的值定义为奇异点。同样以特征 H 为例，其箱型图如图 16 所示。从图中可以看出，特征 H 的数值大于 130 的点可以认为是异常数据。

置信区间法 置信区间法以统计量的置信上限和置信下限作为上下界构成置信区间，置信区间可以展现参数的真实值有一定概率落在测量结果的周围的程度。记置信区间为

$$(\bar{x} - d, \bar{x} + d)$$

则可以通过以下公式计算得到置信区间

Figure 15: 特征 H 的散点图

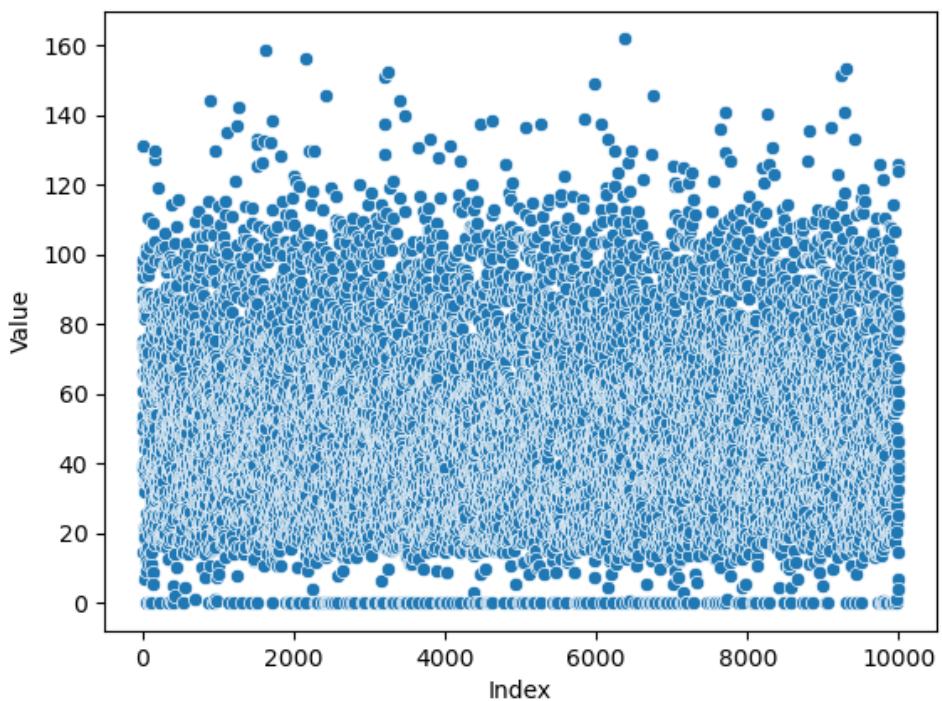
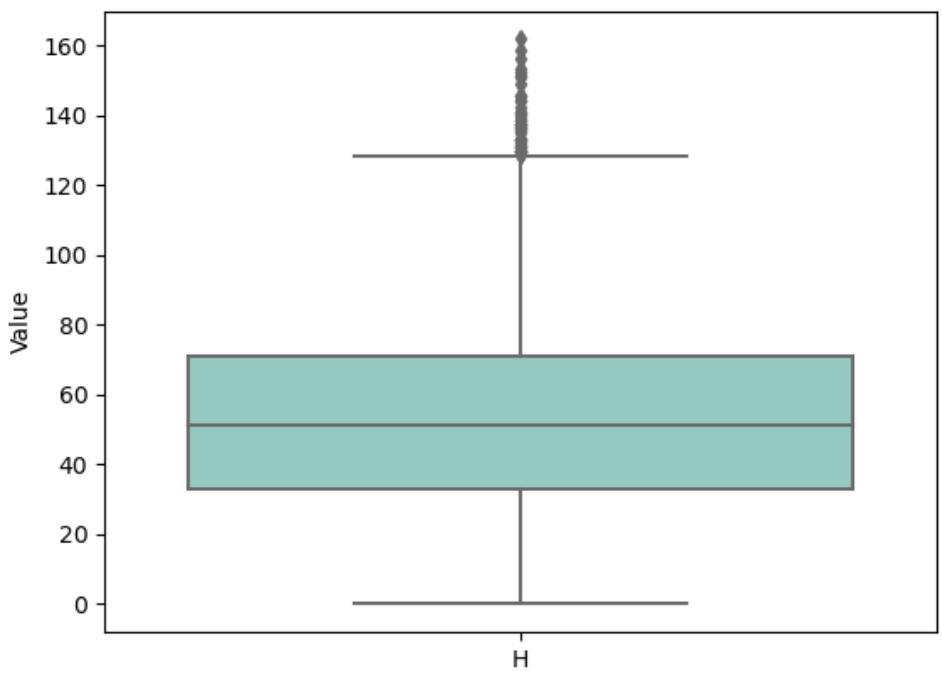


Figure 16: 特征 H 的箱型图



$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x$$

$$d = \frac{1}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) S$$

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

要判断 10000 个样本中是否有奇异点 (95%)，我们有

$$t_{\frac{\alpha}{2}}(n-1) = t_{0.025}(9999) = 1.96$$

经计算可得，7 个特征的置信区间如表 10 所示。

表 10: 特征的置信区间

特征 7 个特征的置信区间	
R	(119.88, 121.33)
G	(114.04, 115.43)
B	(101.59, 103.19)
Gr	(114.41, 115.76)
H	(52.82, 53.86)
S	(83.33, 85.10)
V	(131.07, 132.48)

也即我们有如下结论：

- 对于特征 R ，有 95% 的置信度它的值落在 (119.88, 121.33) 区间内。对于落在该区间外的点，我们判定为异常数据。
- 对于特征 G ，有 95% 的置信度它的值落在 (114.04, 115.43) 区间内。对于落在该区间外的点，我们判定为异常数据。
- 对于特征 B ，有 95% 的置信度它的值落在 (101.59, 103.19) 区间内。对于落在该区间外的点，我们判定为异常数据。
- 对于特征 Gr ，有 95% 的置信度它的值落在 (114.41, 115.76) 区间内。对于落在该区间外的点，我们判定为异常数据。
- 对于特征 H ，有 95% 的置信度它的值落在 (52.82, 53.86) 区间内。对于落在该区间外的点，我们判定为异常数据。
- 对于特征 S ，有 95% 的置信度它的值落在 (83.33, 85.10) 区间内。对于落在该区间外的点，我们判定为异常数据。
- 对于特征 V ，有 95% 的置信度它的值落在 (131.07, 132.48) 区间内。对于落在该区间外的点，我们判定为异常数据。

3.3.7 数据的预处理实验总结

在数据的预处理这一节中，我们计算了数据的均向量，协方差矩阵和相关系数矩阵。观察相关系数矩阵，我们发现了在七个特征中有五个特征之间存在着较强的相关性，而另外两个特征则与其他特征之间的相关性较弱。我们也发现了具有非常强正相关性的一对特征 G 和 Gr ，它们之间的相关系数高达 0.98。

此后我们做出了七个变量的直方图和箱式图。从直方图中我们观察到有五个特征呈正态分布，有一个特征呈轻微右偏分布，有两个特征呈明显右偏分布。从箱式图中我们观察到有五个特征的分布较为接近，有两个特征和其他特征之间的分布相差较大。此外从箱式图中我们也发现七个特征都存在异常值。

最后我们通过三种方法对各变量的异常数据进行了判断和取舍。这三种方法包括了散点图法，箱型图法以及置信区间法。通过置信区间法，我们给出了各个变量在 95% 置信度下正常值的取值范围。我们将在该取值范围内的数据判定为正常数据并保留，将不在该取值范围内的数据判定为异常数据并舍弃。

3.4 统计分析

3.4.1 数据的分布检验

检验问题 特征 R 的 10000 个数据是否符合正态分布 ($\alpha = 0.1$)

检验方法 使用偏度峰度检验：若总体为正态分布，随机抽取样本研究，峰度 g_1 和峰度 g_2 都服从正态分布。需要检验的假设为 H_0 : 数据为正态总体。当假设成立时， g_1 满足理论均值 $\mu = 0$ ，标准差 $\sigma = \frac{6(n-2)}{(n+1)(n+3)}$ 的正态分布。 g_2 满足理论均值 $\mu = 3 - \frac{6}{n+1}$ ，标准差 $\sigma = 3 - \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$ 的正态分布。

检验结果 首先计算峰度 g_1 和峰度 g_2 的理论均值和标准差

$$\alpha = 0.1, \quad n = 10000$$

$$\sigma_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} = 0.0245, \quad \mu_1 = 0$$

$$\sigma_2 = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}} = 0.0490, \quad \mu_2 = 3 - \frac{6}{n+1} = 2.9999$$

计算样本多极距：

$$\begin{aligned} A_1 &= \frac{1}{n} \sum_{i=1}^n x_i = 120.6077 \\ A_2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 = 15916.1155 \\ A_3 &= \frac{1}{n} \sum_{i=1}^n x_i^3 = 2254490.8502 \\ A_4 &= \frac{1}{n} \sum_{i=1}^n x_i^4 = 339849459.7027 \\ B_2 &= A_2 - A_1^2 = 1369.9088 \\ B_3 &= A_3 - 3A_2A_1 + 2A_1^3 = 4442.4927 \\ B_4 &= A_4 - 4A_3A_1 + 6A_2A_1^2 - 3A_1^4 = 6552278.2388 \end{aligned}$$

计算峰度 g_1 和峰度 g_2 可得

$$\begin{aligned} g_1 &= \frac{B_3}{B_2^{\frac{3}{2}}} = 0.0876 \\ g_2 &= \frac{B_4}{B_2^2} = 3.4915 \end{aligned}$$

对 g_1 和度 g_2 进行标准变换 $\mu = \frac{x-\mu}{\sigma}$ 可得

$$\begin{aligned} u_1 &= \frac{g_1 - \mu_1}{\sigma_1} = 3.5780 \\ u_2 &= \frac{g_2 - \mu_2}{\sigma_2} = 10.0418 \end{aligned}$$

检验临界值为 $Z_{\frac{\alpha}{4}} = Z_{0.025} = 1.96$, 由于

$$|u_1| = 3.5780 > 1.96$$

$$|u_2| = 10.0418 > 1.96$$

故拒绝假设 H_0 , 认为总体数据不服从正态分布。

同理可以对其他特征做分布检验, 结果如表 11 所示。实验结果表明对于七个因素的正态分布检验均拒绝了假设 H_0 , 也即均不服从正态分布。

3.4.2 参数估计之区间估计

从上一小节正态分布检验可知特征 V 相对最接近正态分布。在参数估计这一小节中, 我们使用特征 V 对应的 10000 个样本值作为数据。

估计问题 根据特征 V 的 10000 个样本点估计特征 V 的取值范围 ($P = 95\%$)

估计方法 \bar{x} 是 μ 的无偏估计, 且具有有效性和充分性。但是随机变量不能正好落在 μ 上, 因此可以用一个区间去包含 μ :

$$\mu = \bar{x} \pm \frac{t_{\alpha/2, f} \sigma}{\sqrt{n}}$$

其中由于 σ 未知, 可以用 s 代替 σ 。

表 11: 正态分布检验 ($\alpha = 0.1$)

特征	g_1	g_2	$ u_1 $	$ u_2 $	假设 H_0
R	0.0876	3.4915	3.5780	10.0418	拒绝
G	0.1162	3.5460	4.7469	11.1560	拒绝
B	0.3449	3.1493	14.0828	3.0527	拒绝
Gr	0.1881	3.6915	7.6805	14.1278	拒绝
H	0.4169	2.8003	17.0257	4.0774	拒绝
S	0.6610	3.3637	26.9942	7.4312	拒绝
V	-0.0038	3.4115	0.1571	8.4082	拒绝

估计结果 当 $p = 0.95$ 时, $\alpha = 0.05$, $f = 9999$, 经查表得 $t_{\alpha,f} = 1.960$, 因此:

$$\mu = 131.7724 \pm \frac{1.9600 * 36.0665}{\sqrt{10000}} = 131.7724 \pm 0.7069 = [131.0655, 132.4793]$$

也即特征 V 的取值范围估计为 $[131.0655, 132.4793]$ 。

3.4.3 参数估计之样本容量确定

同区间估计一节中, 我们选用特征 V 对应的 10000 个样本值作为数据。

估计问题 已知特征 V 的均值 131.7724, 估计标准差 36.0665, 估计在 95% 置信度下, 使估计允许误差不超过其平均值 10%, 求所需最低样本容量。

估计方法 在样本量较大的时候认为 t 分布与正态分布近似, 所以用正态分布的检验:

$$d = \frac{t_{\frac{\alpha}{2}, n-1} s}{\sqrt{n}}$$

$$n = \left(\frac{t_{\frac{\alpha}{2}} s}{d} \right)^2 \approx \left(\frac{z_{\frac{\alpha}{2}} s}{d} \right)^2$$

估计结果 根据公式计算可得:

$$d = \frac{131.7724}{10} = 13.1772$$

$$n \approx \left(\frac{z_{0.025} s}{d} \right)^2 = \left(\frac{1.96 \times 36.0665}{13.1772} \right)^2 = 28.7789 \approx 29$$

可知取 $n = 29$ 即可满足所给精度要求。

3.4.4 统计检验之离群值检验

检验问题 根据特征 R 的 10000 个数据确定离群值。

检验方法 根据课堂上总结的推荐准则表 12, 由于我们实验的样本量远大于 185, 因此应该使用拉伊达准则, 也即 3σ 准则。

表 12: 推荐准则表

测量次数范围	建议使用的准则
$3 \leq n \leq 25$	狄克逊准则, 格拉布斯准则 ($\alpha = 0.01$)
$25 \leq n \leq 185$, 格拉布斯准则 ($\alpha = 0.05$), 肖维勒准则
$n > 185$	拉伊达准则

检验结果 由拉伊达准则确定的正常值的取值范围为

$$(\bar{x} - 3\sigma, \bar{x} + 3\sigma) = (9.5708, 231.6445)$$

也即落在 $(9.5708, 231.6445)$ 范围外的值我们可以判定为离群值。同理可以对其他特征做离群值检验, 结果如表 13 所示, 具体含义不再赘述。

表 13: 离群值检验

特征	正常值取值范围
R	$(9.5708, 231.6445)$
G	$(8.5438, 220.9254)$
B	$(-19.6162, 224.3935)$
Gr	$(11.7864, 218.3820)$
H	$(-25.8627, 132.5387)$
S	$(-50.9382, 219.3717)$
V	$(23.5728, 239.9720)$

3.4.5 统计检验之方差比较检验

检验问题 根据各自的 10000 个样本, 检验特征 R 和特征 H 有无显著差异 (置信度 95%)。

检验方法 使用 F 检验:

$$F = \frac{S_1^2}{S_2^2} \sim F(m-1, n-1)$$

检验结果 根据特征 R 和特征 H 的方差计算 F 值:

$$F = \frac{S_1^2}{S_2^2} = \frac{1369.9088}{696.9725} = 1.9655 > F(9999, 9999)$$

因此可以得到特征 R 和特征 H 存在显著性差异的结论。

3.4.6 统计检验之均值检验

检验问题 根据各自的 10000 个样本, 检验特征 R 和特征 G 有无显著差异 ($P = 95\%$)。

检验方法 根据前文的实验结果，虽然特征 R 和特征 G 不满足正态分布，但是根据前文的分布直方图可知，它们的分布形状相似，样本量相同且样本量较大，因此仍然可用 t 检验。具体来说，我们选用非配对情形的双总体 t-检验方法。

检验结果 根据特征 R 和特征 H 的方差计算 t 值：

$$n_1 = 10000, \quad \bar{x}_1 = 120.6077, \quad s_1^2 = 1369.9088$$

$$n_2 = 10000, \quad \bar{x}_2 = 53.3380, \quad s_2^2 = 696.9725$$

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \left(\frac{1}{m} + \frac{1}{n}\right)}} = 147.9658$$

$$f = n_1 + n_2 - 2 = 19998, \quad \alpha = 0.05$$

$$t > t_{0.05, 19998}$$

因此可以得到特征 R 和特征 H 存在显著性差异的结论。

3.4.7 方差分析与极差分析

在实验设计这一章节中已经做过方差分析与极差分析，因此不再赘述。

3.4.8 统计分析实验总结

在这一节中，我们从数据的分布检验，参数估计，统计检验，以及方差分析与极差分析四个方面进行了实验。

在数据的分布检验小节中，我们使用了峰度偏度检验来判断各个特征是否服从正态分布，并最终得出了各个特征虽然从直方图上来看接近正态分布，但从分布检验上来看不服从正态分布的结论。

在参数估计小节中，我们进行了区间估计和样本容量确定这两组实验。在区间估计实验中，我们给出了 $P = 95\%$ 时特征的取值范围估计。在样本容量确定实验中，我们给出了在 95% 置信度下，使得估计误差不超过均值 10% 的最低样本容量。

在统计检验小节中，我们进行了离群值检验，方差比较检验以及均值检验三组实验。在离群值检验中，我们采用拉伊达准则给出了各个特征的正常值取值范围。在方差比较检验中，我们使用 F 检验证证了部分特征之间存在显著差异。在均值检验中，我们使用 t 检验证证了部分特征之间存在显著差异。

在方差分析与极差分析小节中。我们沿用了正交实验设计这一节中的极差分析和方差分析的结果与分析。

4 数据的图示

4.1 作业要求

- 样本数量不得少于 10 个
- 各样本数据维度不得少于二维
- 选择合适的作图方式，对数据表现的现象进行反映
- 说明图示展现的对比关系是什么

- 要求至少选取以下图示方法中的两种，可以根据对比关系数据进行统计分组后再作图
 - 散点图，折线图，条形图，饼图

4.2 实验介绍

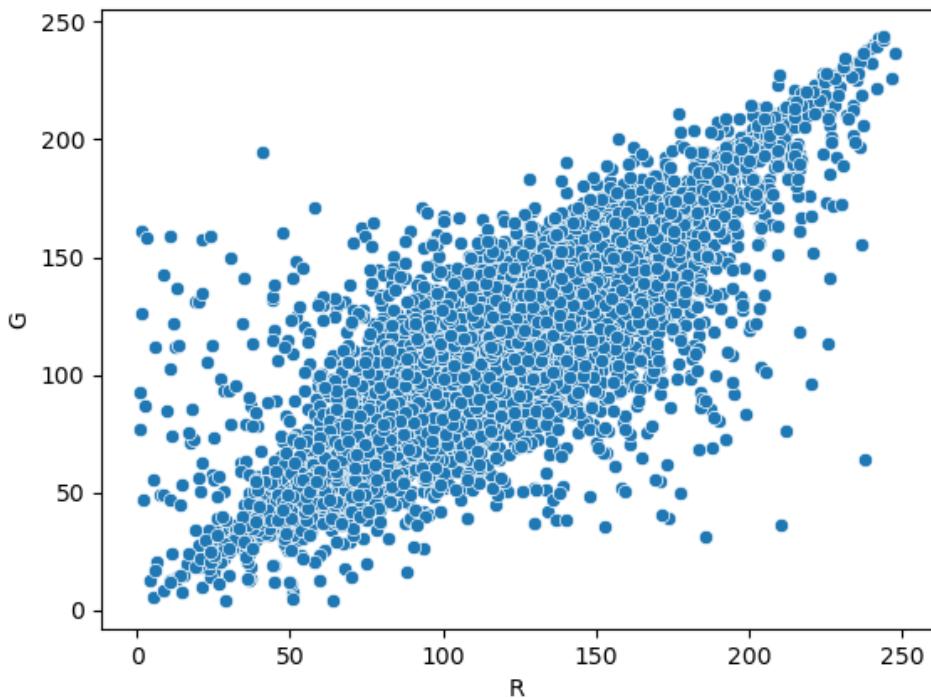
在数据的图示这一章中我们采用与数据的预处理和统计分析这一章中相同的数据进行实验。也即我们从 ImageNet 数据集中随机选择了 10000 张图片做为实验数据。对于每张图片，通过图像处理可以获得表 9 中的 7 项特征。因此总体的实验数据为 10000×7 的矩阵，部分特征数据如图 8 所示。

4.3 数据图示

4.3.1 散点图

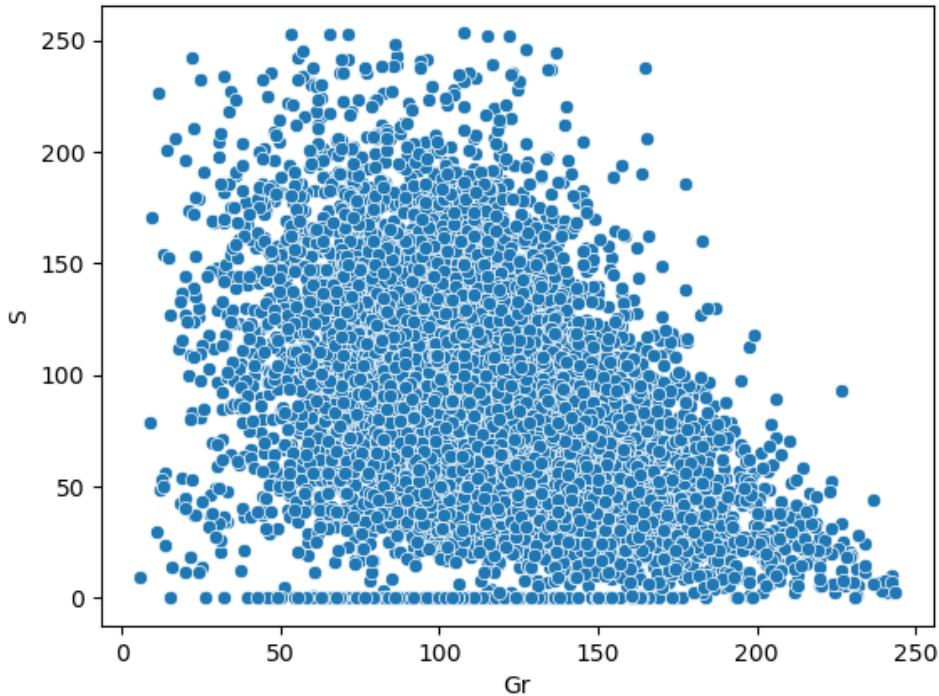
使用散点图可以显示各个特征两两之间的关系。如图 17 展示了特征 R 与特征 G 之间的散点图，从图中可以看出特征 R 与特征 G 之间存在着较为明显的正相关关系。又如图 18 展示了特征 Gr 与特征 S 之间的散点图，从图中可以看出特征 Gr 与特征 s 之间存在着一定的负相关关系。

Figure 17: 特征 R 与特征 G 之间的散点图



为了更全面的展示特征两两之间的关系，我们在图 19 中绘制了七组特征两两之间的散点图。该图中对角线上为单个特征的直方图，非对角线上为特征两两之间的散点图。从图中可以看出：

Figure 18: 特征 Gr 与特征 S 之间的散点图



- 特征 Gr 和 G 之间存在显著的正相关关系。
- 特征 R 和 G 之间，以及特征 G 和 B 之间存在着较为显著的正相关关系。
- 特征 Gr 与 S 之间，特征 G 与 S 之间，以及特征 B 与 S 之间存在着一定的负相关关系。
- 其他的特征之间，如 H 和 Gr 之间存在很弱的相关关系，或者不存在相关关系。

4.3.2 折线图

使用折线图可以突出因变量随自变量的变化趋势。如图 20 中所示，我们绘制了特征 G 与特征 Gr 之间的折线图。可见特征 G 与特征 Gr 之间总体上呈正相关性。

此外由于数据点过多 ($n = 10000$)，为了更好的展示效果，我们将特征 G 的数据中两两相差小于 1 的数据合并到一个点，并使用 95% 的置信区间表示合并后的特征 G 对应的特征 Gr 的取值范围，如图 21 所示，其中阴影部分代表置信区间。从这张图中可以更明显地观察到特征 G 与特征 Gr 之间的正相关性。

为了方便综合比较，我们制作复合折线图，将多个其他特征的与 Gr 之间的关系放在一张图中进行比较。如图 22 和图 23 所示。其中图 22 展示了特征 R , G , B 分别与特征 Gr 之间的折线图，可见 R , G , B 三个特征与 Gr 之间都存在着较强的正相关性。图 23 展示了特征 H , S , V 分别与特征 Gr 之间的折线图，可见特征 V 与 Gr 之间存在较强的相关性，特征 S 与 Gr 之间存在较弱的负相关性，特征 H 与特征 Gr 之间相关性很弱。

Figure 19: 七组特征两两之间的散点图

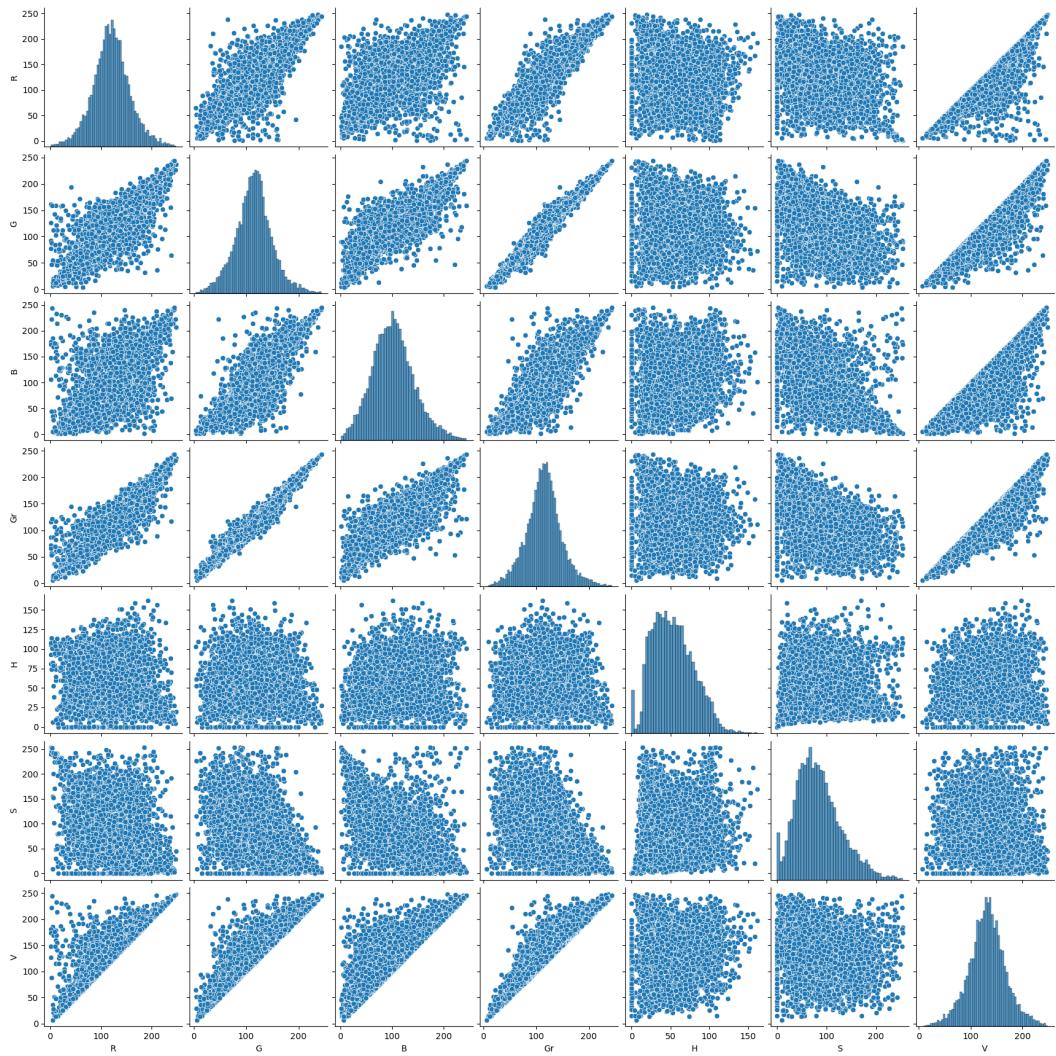


Figure 20: 特征 G 与特征 Gr 之间的折线图

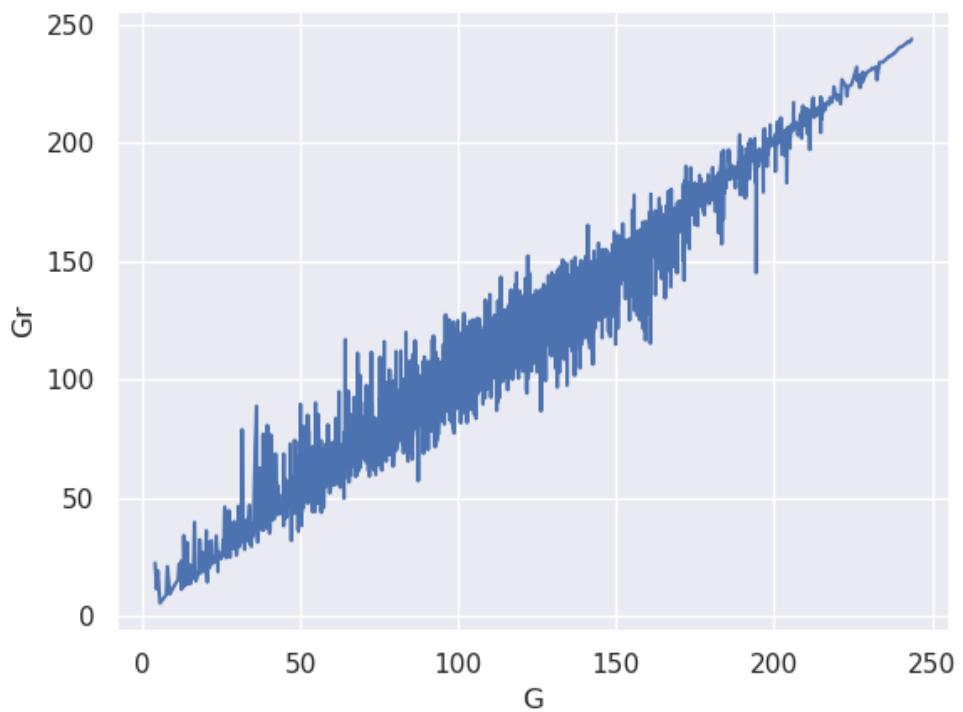


Figure 21: 特征 G 与特征 Gr 之间的折线图

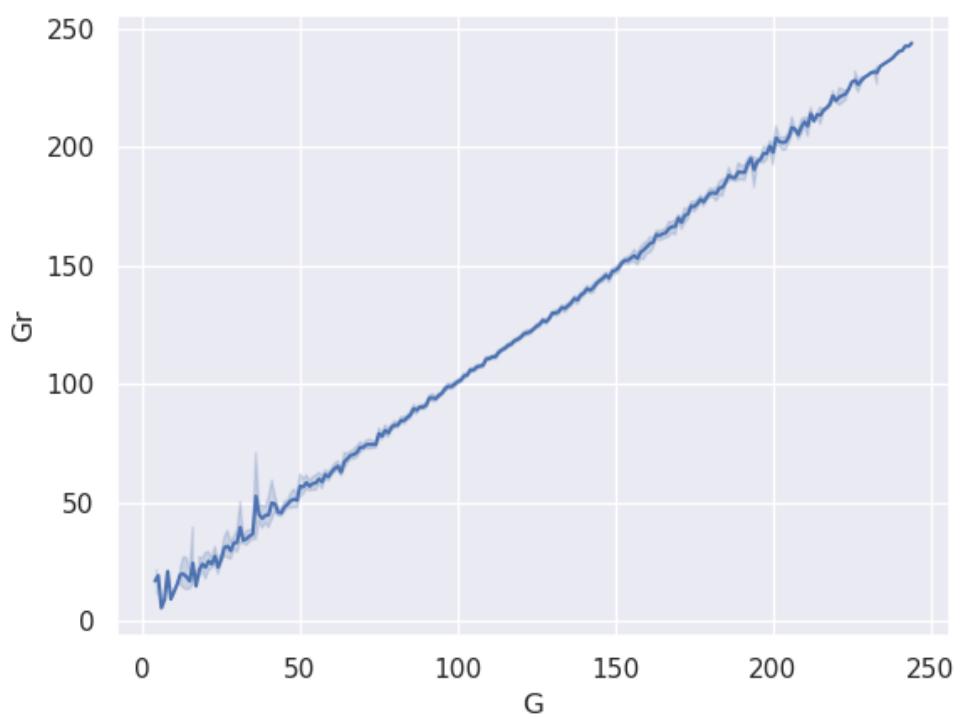


Figure 22: 特征 R , G , B 分别与特征 Gr 之间的折线图

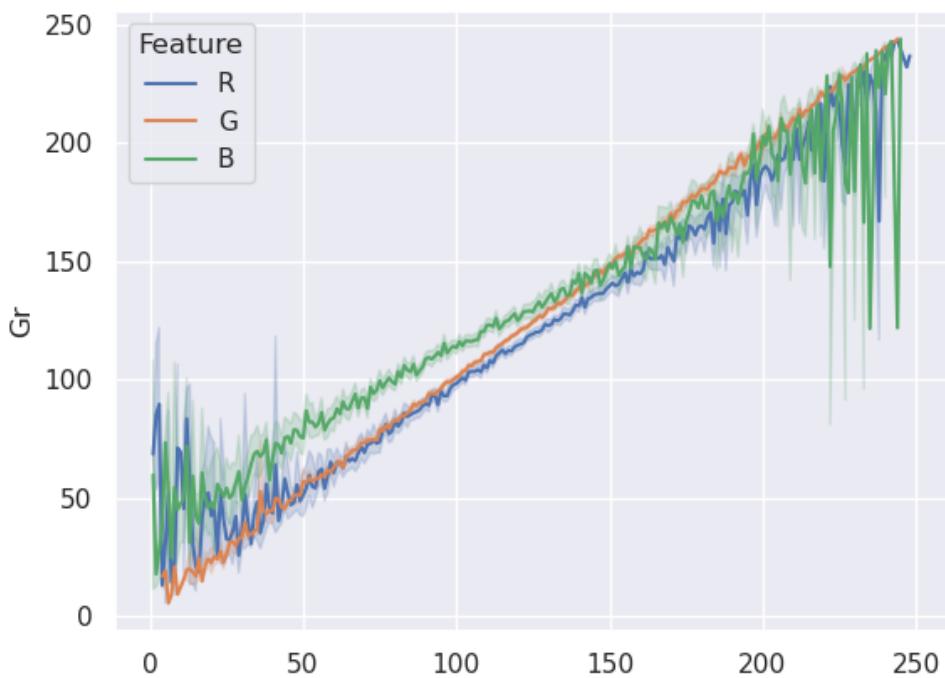
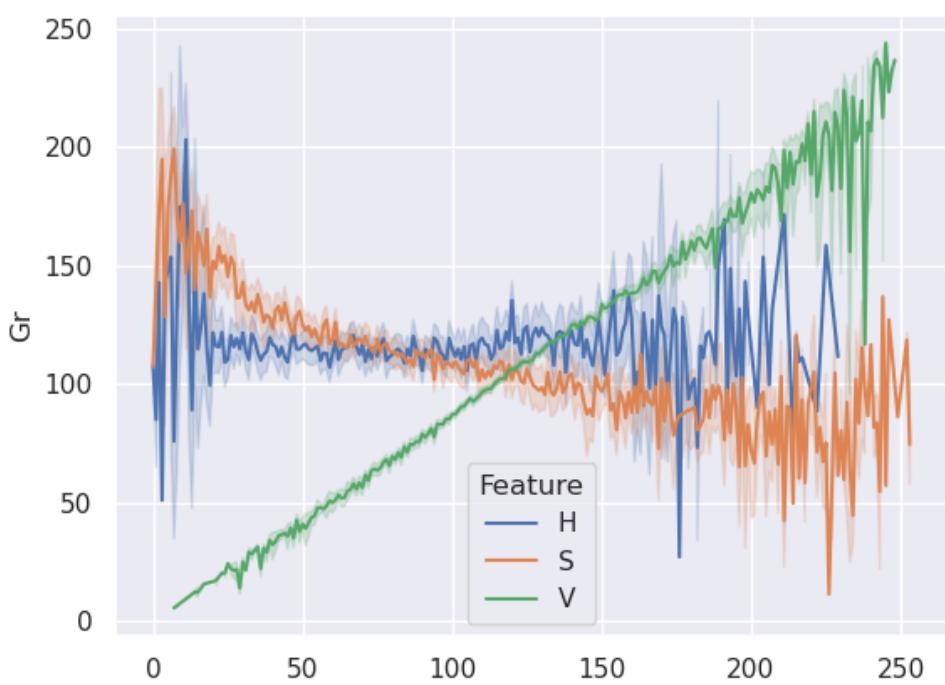


Figure 23: 特征 H , S , V 分别与特征 Gr 之间的折线图



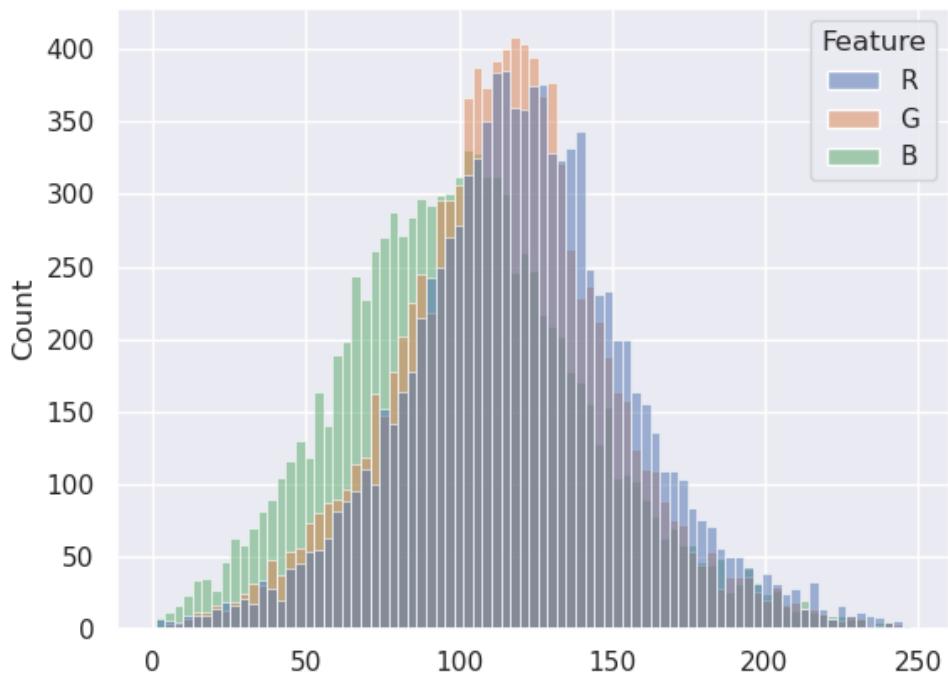
4.3.3 条形图

条形图使用等宽直条的长短表示各个相互独立的指标的大小，便于我们比较各个指标之间的差距。在数据的预处理和统计分析这一章中我们已经使用过直方图形式的条形图来展示各个特征的分布情况。在本小节中，我们进一步制作复合条形图，以便于更直观地比较各个特征的分布情况。

在图 24 中我们绘制了特征 R , G , B 分布的复合条形图，从图中可以看出特征 R 和 G 在分布上非常接近，而特征 B 相比于其他两个特征偏度更大，也即呈右偏分布，峰度更小，也即呈扁平分布。这一结论和我们在表 11 中关于偏度峰度的计算结果形成了相互验证。

在图 25 中我们绘制了特征 H , S , V 分布的复合条形图，从图中可以看出三个之间存在着显著的差别。在偏度上特征 S 最大，特征 H 次之，特征 V 最小。在峰度上特征 V 最大，特征 S 次之，特征 H 最小。这一结论和我们在表 11 中关于偏度峰度的计算结果形成了相互验证。

Figure 24: 特征 R , G , B 分布的复合条形图

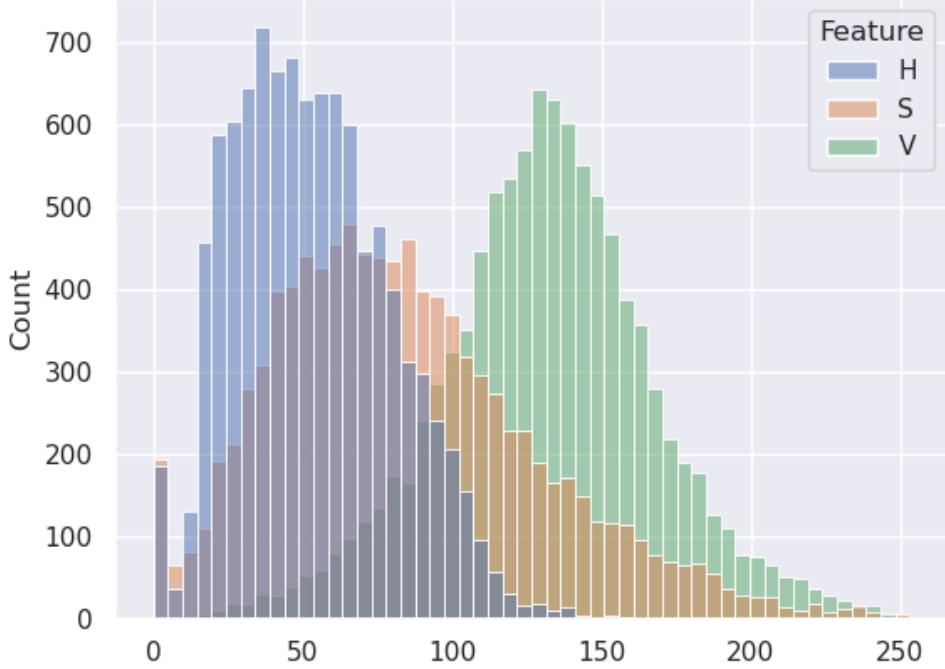


4.3.4 饼图

饼图可以反映单一指标的构成等信息。在饼图的基础上，环状饼图可以用不同的圆环表示不同的数据序列，从而弥补普通饼图在展示多属性上的不足。

如图 26 所示，我们绘制了七组特征中各自异常数据所占比例的环状饼图，其中异常数据的判定使用的是统计检验之离群值检验这一小节中的拉伊达准则，具体判定标准见表 13。该图中橙红色的区域代表离群值占总样本数的比例，其他颜色的区域代表各个特征

Figure 25: 特征 H , S , V 分布的复合条形图



的正常值占总样本数的比例。由外而内七个圆环依次代表特征 R , G , B , Gr , H , S 和 V 。

从图中可以看出七组特征中离群值的占比都小于 1%。且离群值最少的是特征 B , H 和 V , 离群值占比仅为 0.4%。离群值最多的是特征 S , 离群值占比达 0.7%。

4.3.5 数据的图示实验总结

在数据的图示这一章中，我们分别使用了散点图，折线图，条形图和饼图进行了数据的可视化并进行了分析与对比。

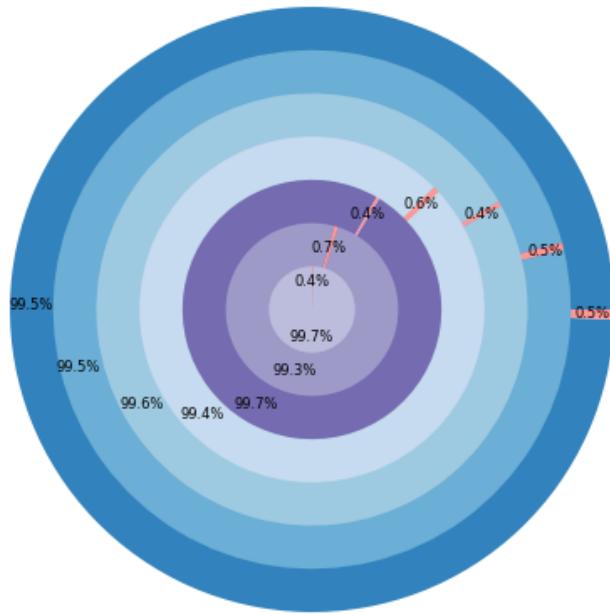
我们通过绘制七组特征两两之间的散点图，可视化了特征之间的相关性关系。发现了多组特征之间存在正相关关系，少数特征之间存在负相关关系，以及部分特征之间相关性很弱或者不存在相关性关系。

我们通过在同一张图上绘制多组特征两两之间的折线图，更直观地观察到了特征之间的相关性关系。

我们通过在同一张图上绘制多组特征分布情况的条形图，对比了多组特征在分布上的相似情况。我们也根据绘图结果定性地判断了多个特征之间的偏度与峰度的大小关系，并和前文实验中计算得到的偏度峰度大小关系进行了对比与验证。

最后我们通过绘制环状饼图，可视化了七组特征中各自的离群值占总样本数的比例。我们观察到所有特征的离群值占比都较少，且不同特征的离群值占比之间有着较大差异，最小占比仅为 0.4%，而最大占比为 0.7%。

Figure 26: 七组特征中各自离群值所占比例的环状饼图



5 数据的处理作业一

5.1 作业要求

- 结合实验室具体研究问题
- 样本数不得少于 30 个
- 样本数据的变量数不得少于三种
- 选取回归分析 OR 相关分析中的一种
 1. 回归分析
 - 开展两变量间的回归分析（可以是线性、非线性，logistic 均可），并给出显著性检验
 - 开展至少三变量之间的多元回归分析，并给出复相关系数和偏回归系数的显著性检验
 - 对回归分析结果结合具体研究问题展开讨论
 2. 相关分析
 - 根据其中两个变量的类型选定相关计算方法
 - 计算相关系数，并做显著性检验
 - 根据变量含义对各变量分组，
 - 计算两组之间的典型相关系数

– 对相关分析结果结合具体研究问题展开讨论

5.2 实验介绍

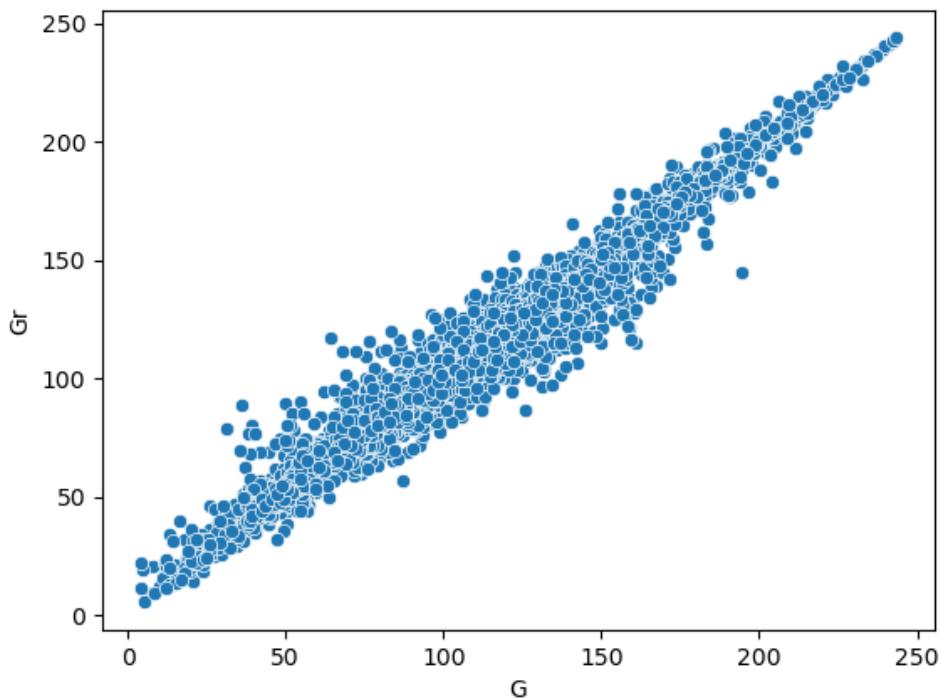
在这一章中我们采用与数据的预处理和统计分析这一章中相同的数据进行实验。也即我们从 ImageNet 数据集中随机选择了 10000 张图片做为实验数据。对于每张图片，通过图像处理可以获得表 9 中的 7 项特征。因此总体的实验数据为 10000×7 的矩阵，部分特征数据如图 8 所示。

5.3 回归分析

5.3.1 两变量线性回归

回归问题 我们选择特征 G 和特征 Gr 进行两变量的线性回归分析，每个特征包含 10000 条数据。特征 G 和特征 Gr 之间散点图如图 27 所示，可见两个特征之间有较为显著的正相关关系。

Figure 27: 特征 G 和特征 Gr 之间散点图



回归方法 设要求解的回归方程为 $y = a + bx$ ，其中 a 和 b 是要求解的系数。我们可以使用最小二乘法求解使得误差

$$Q(\hat{a}, \hat{b}) = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

最小的系数 a 和 b 。等价于求解如下方程组

$$\begin{aligned} na + b \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

求解过程为

$$\begin{aligned} L_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \\ L_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2 \\ L_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) \\ a &= \bar{y} - b\bar{x}, \quad b = \frac{L_{xy}}{L_{xx}} \end{aligned}$$

此外我们可以通过决定系数：

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

来判断回归方程估测可靠程度的高低。

回归结果 按照公式求解可得

$$a = \bar{y} - b\bar{x} = 5.2471, \quad b = \frac{L_{xy}}{L_{xx}} = 0.9573$$

也即求解得到的线性回归方程为

$$y = a + bx = 5.2471 + 0.9573x$$

决定系数为：

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 0.9685$$

我们将该线性回归方程绘制到特征 G 和特征 Gr 的散点图上，如图 28 所示，可见拟合效果是比较理想的。

5.3.2 两变量线性显著性检验

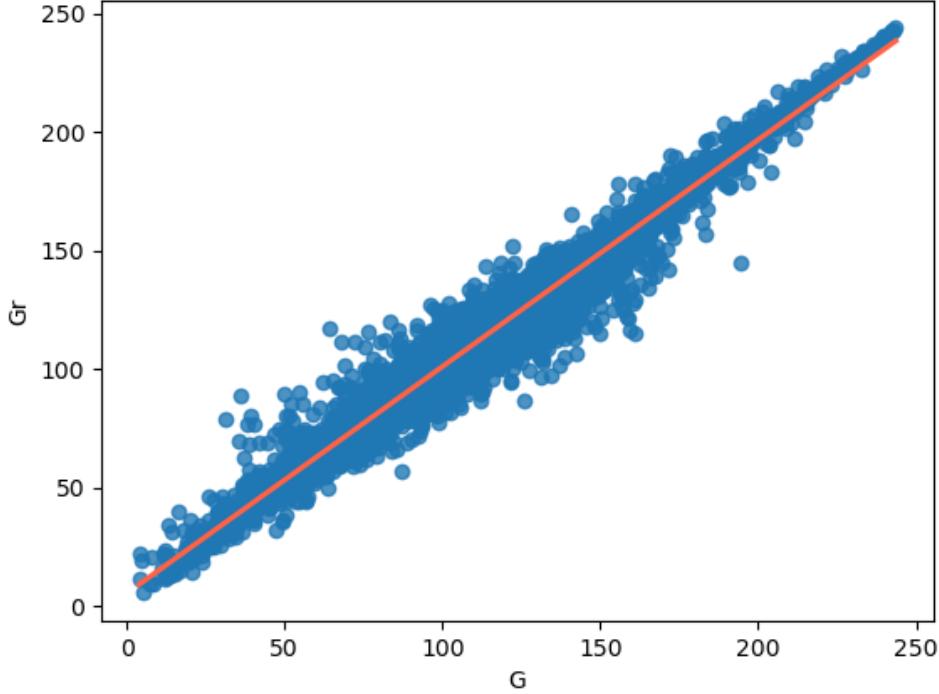
检验问题 求得回归方程只是完成了统计分析中两变量关系的统计描述，我们还需要回答它所来自的总体的直线回归关系是否确实存在。

检验方法 使用 F 检验：

$$F = \frac{\frac{Q_x}{f_x}}{\frac{Q_e}{f_e}} = \frac{\frac{Q_x}{1}}{\frac{Q_e}{N-2}}$$

根据显著性水平 α 比较计算得到的 F 值和 $F_\alpha(1, N-2)$ 的大小关系。若 $F > F_\alpha(1, N-2)$ ，则在显著性水平 α 下，求解得到回归方程是显著的。

Figure 28: 特征 G 和特征 Gr 的线性回归方程



检验结果 首先计算回归平方和 Q_x :

$$Q_x = SS_r = \sum_{i=1}^n (\hat{y} - \bar{y})^2 = 11482621.9177$$

再计算残差平方和 Q_e :

$$Q_e = SS_e = \sum_{i=1}^n (y - \hat{y})^2 = 373415.2436$$

进而可以计算 F :

$$F = \frac{\frac{Q_x}{1}}{\frac{Q_e}{N-2}} = 307441.2625$$

取显著性水平 $\alpha = 0.01$, 查表可知 $F_{0.01}(1, 9998) < F_{0.01}(1, 600) = 6.677$, 因此 $F > F_{0.01}(1, 9998)$ 。所以显著性检验的结论为在 0.01 水平上显著, 也即可信赖程度在 99% 以上。

5.3.3 两变量线性回归结果讨论

特征 G 和特征 Gr 之间存在很强的正相关性, 在数据的预处理和统计分析这一章中, 我们已经发现了这一结论。在线性回归这一小节, 我们通过求解线性回归方程并进行显著性检验再次验证了这一结论。实际上从物理意义上来看, 特征 Gr 代表图像的灰度, 特征 G 代表图像在 RGB 色彩空间中的 G 分量, 特征 Gr 是可以通过特征 R , G , B 线性组合得到的, 因此特征 G 和特征 Gr 之间存在很强的正相关性也是合乎逻辑的。

5.3.4 多元线性回归

回归问题 我们选择特征 H, S, V 作为自变量，特征 Gr 作为因变量进行多元线性回归分析，每个特征包含 10000 条数据。

回归方法 和两变量线性回归类似的，我们仍然通过最小二乘法求解残差平方和最小时的回归系数。设要求解的回归方程为：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \mu_i$$

其中 β_i 和 μ_i 是要求解的系数。在矩阵形式下，多元线性回归方程可以表示为

$$Y = X\beta + e$$

等式两边同时左乘 X'

$$X'Y = X'X\beta + X'e$$

根据最小二乘原则

$$X'e = 0$$

则正规方程为

$$X'X\hat{\beta} = X'Y$$

多元回归的普通最小二乘法估计量为

$$\hat{\beta} = (X'X)^{-1}X'Y$$

此外我们可以通过计算复相关系数：

$$R = \sqrt{\frac{Q_x}{Q_T}} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}}$$

来判断因变量和多个自变量之间的线性相关程度。

回归结果 根据回归公式计算可得

$$\begin{aligned}\hat{\beta} &= [-0.1060, -0.2197, 0.8604] \\ \mu &= 25.8570\end{aligned}$$

也即求解得到的多元线性回归方程为

$$y = \hat{\beta}\hat{x} + \mu = -0.1060x_1 - 0.2197x_2 + 0.8604x_3 + 25.8570$$

复相关系数为：

$$R = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}} = 0.9785$$

其值说明特征 Gr 与特征 H, S, V 之间存在较强的线性相关关系。

5.3.5 多元线性回归显著性检验

检验问题 和两变量线性回归类似的，我们需要检验多元线性回归方程是否显著。

检验方法 首先我们需要检验所有自变量联合起来对因变量影响的显著性，需要在方差分析的基础上进行 F 检验。原假设：

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

也即所有解释变量联合起来时对被解释变量的影响不显著。备择假设：

$$H_1 : \beta_j \neq 0 \text{ for some } j = 1, 2, \dots, k$$

建立统计量：

$$F = \frac{\frac{Q_x}{k-1}}{\frac{Q_e}{n-k}} = \frac{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{k-1}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-k}} \sim F(k-1, n-k)$$

根据显著性水平 α 比较计算得到的 F 值和 $F_\alpha(k-1, n-k)$ 的大小关系。若 $F > F_\alpha(k-1, n-k)$ ，则在显著性水平 α 下，求解得到回归方程是显著的。

由于在但多元回归中，F 检验显著不代表每个自变量都对 Y 有显著影响，因此还需要分别检验当其他自变量保持不变时，各个自变量对因变量是否有显著影响。首先需要计算偏回归系数 β_j 的标准化回归系数 P_j ：

$$P_j = |\beta_j| \sqrt{\frac{L_{jj}}{Q_T}}$$

其中， L_{jj} 为 X_j 的离差平方和， Q_T 为 Y 的离差平方和。 P_j 越大，则对应的因素 X_j 越重要。

对偏回归系数做显著性检验，我们需要计算每个偏回归系数的偏回归平方和：

$$F_j = \frac{U_j}{\frac{Q_e}{n-k}}$$

$$U_j = \sum_{i=1}^n (\beta_j(x_j i - \bar{x}_j))^2$$

如果 $F < F_\alpha(1, n-k)$ ，则说明 x_i 对 y 的影响是不显著的，可以将它从回归方程中去掉。

检验结果 首先检验所有自变量联合起来对因变量影响的显著性，计算回归平方和 Q_x ：

$$Q_x = SS_r = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 11351202.6208$$

再计算残差平方和 Q_e ：

$$Q_e = SS_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 504834.5405$$

进而可以计算 F ：

$$F = \frac{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{k-1}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-k}} = 74920.0067$$

取显著性水平 $\alpha = 0.01$, 查表可知 $F_{0.01}(3, 9996) < F_{0.01}(3, 600) = 3.814$, 因此 $F > F_{0.01}(3, 9996)$ 。所以显著性检验的结论为所有自变量联合起来对因变量的影响在 0.01 水平上显著, 也即可信赖程度在 99% 以上。

接下来我们计算偏回归系数 β_j 的标准化回归系数 P_j :

$$P_1 = |\beta_1| \sqrt{\frac{L_{11}}{Q_T}} = 0.0813$$

$$P_2 = |\beta_2| \sqrt{\frac{L_{22}}{Q_T}} = 0.2874$$

$$P_3 = |\beta_3| \sqrt{\frac{L_{33}}{Q_T}} = 0.9012$$

从标准化回归系数 P_j 可知, 三个特征中, V 最重要, S 其次, H 最不重要。

接下来对偏回归系数做显著性检验, 计算每个偏回归系数的偏回归平方和:

$$U_1 = \sum_{i=1}^n (\beta_j(x_1 i - \bar{x}_1))^2 = 78364.9718$$

$$U_2 = \sum_{i=1}^n (\beta_j(x_2 i - \bar{x}_2))^2 = 979310.3152$$

$$U_3 = \sum_{i=1}^n (\beta_j(x_3 i - \bar{x}_3))^2 = 9630455.8956$$

$$F_1 = \frac{U_1}{\frac{Q_e}{n-k}} = 1551.6693$$

$$F_2 = \frac{U_2}{\frac{Q_e}{n-k}} = 19390.8798$$

$$F_3 = \frac{U_3}{\frac{Q_e}{n-k}} = 190688.2937$$

取显著性水平 $\alpha = 0.01$, 查表可知 $F_{0.01}(1, 9996) < F_{0.01}(1, 600) = 6.677$, 因此 F_1, F_2, F_3 均大于 $F_{0.01}(1, 9996)$ 。所以显著性检验的结论为 3 个自变量分别对于因变量的影响在 0.01 水平上都显著, 也即可信赖程度在 99% 以上。

5.3.6 多元线性回归结果讨论

在多元线性回归的实验中, 我们得到了自变量 H, S, V 关于因变量的 Gr 的多元线性回归方程, 并通过显著性检验证明了 3 个变量总体上对于因变量的影响, 以及 3 个变量单独对于因变量的影响都是显著的。实际上从物理意义上来看, 特征 Gr 代表图像的灰度, 特征 H, S, V 分别代表图像在 HSV 色彩空间中的三个分量。特征 Gr 是可以通过 RGB 色彩空间中的特征 R, G, B 线性组合得到的, 而 HSV 色彩空间又可以通过 RGB 色彩空间转化得到, 因此特征 H, S, V 和特征 Gr 之间存在着本小节所得到的多元线性回归方程是合乎逻辑的。

6 数据的处理作业二

6.1 作业要求

- 结合实验室具体研究问题

- 样本数不得少于 30 个
- 选取聚类分析 OR 降维分析中的一种

1. 聚类分析

- 样本数据的变量数不得少于两种
- 根据数据类型，选取相似性测度方法
- 选定聚类的目标类数
- 选定类间距离计算方式
- 选定聚类算法并给出聚类过程
- 对聚类结果进行分析

2. 降维分析

- 样本数据的变量数不得少于四种
- 根据数据类型选择 PCA 或者 LDA
- 给出降维分析中各主成分或者 LDA 特征
- 给出选择最终降维成分的依据
- 分析降维后新变量与原始变量之间关系

6.2 实验介绍

在这一章中我们采用与数据的预处理和统计分析这一章中相同的数据进行实验。也即我们从 ImageNet 数据集中随机选择了 10000 张图片做为实验数据。对于每张图片，通过图像处理可以获得表 9 中的 7 项特征。因此总体的实验数据为 10000×7 的矩阵，部分特征数据如图 8 所示。

6.3 降维分析

6.3.1 降维问题

PCA 即主成分分析 (Principal Component Analysis)，是一种常用的降维方法。PCA 可以在尽可能多地保留原始数据的信息的前提下，将大量的相关特征转化成少数的无关特征，这些少数的无关特征就是主成分。在本次实验中，我们使用 PCA 对具有 7 项特征的 10000 条数据进行降维。

6.3.2 降维方法

在 PCA 之前，我们首先要将数据进行标准化，也即使得特征的均值为 0，方差为 1：

$$x'_i = \frac{x_i - \bar{x}_i}{\sqrt{D(x_i)}}$$

其中 x_i 代表原始的第 i 个特征， \bar{x}_i 代表第 i 个特征的均值， $D(x_i)$ 代表第 i 个特征的方差， x'_i 代表标准化后的第 i 个特征。

设 $X = (X_1, X_2, \dots, X_p)$ 的协方差矩阵为 σ ，其特征根为

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

相应的单位化的特征向量为

$$T_1, T_2, \dots, T_p$$

那么，由此确定的主成分为

$$Y_1 = T'_1 X, Y_2 = T'_2 X, \dots, Y_m = T'_m X$$

主成分的方差分别为 σ 的特征值

$$\lambda_1, \lambda_2, \dots, \lambda_m$$

为了确定多少个主成分是足够的，我们可以计算累计贡献率：

$$\frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^p \lambda_k}, \quad i = 1, 2, \dots, p$$

一般可以取累计贡献率达 85 ~ 95% 的特征值所对应的前 $m (m \leq p)$ 个主成分。

PCA 降维的步骤可以总结为：

- 对数据做标准化
- 求特征协方差矩阵
- 求协方差矩阵的特征值和特征向量
- 将特征值按照从大到小的顺序排列，选择其中最大的 k 个，将其对应的 k 个特征向量分别作为列向量组成特征向量矩阵
- 将样本点投影到选取的特征向量上。这样，就将原始样例的 n 维特征变成了 k 维，这 k 维就是原始特征在 k 维上的投影，代表了原始的 n 个特征。

6.3.3 降维过程

首先我们对具有 7 项特征的 10000 条数据进行标准化，部分标准化后的数据如图 29 和所图 30 所示。

接着我们计算标准化后的特征的协方差矩阵：

$$\Sigma = \begin{bmatrix} 1.0001 & 0.8286 & 0.6350 & 0.9070 & -0.1525 & -0.2790 & 0.8814 \\ 0.8286 & 1.0001 & 0.8495 & 0.9842 & 0.0411 & -0.3975 & 0.9029 \\ 0.6350 & 0.8495 & 1.0001 & 0.8514 & 0.3443 & -0.5592 & 0.7957 \\ 0.9070 & 0.9842 & 0.8514 & 1.0001 & 0.0221 & -0.4049 & 0.9353 \\ -0.1525 & 0.0411 & 0.3443 & 0.0221 & 1.0001 & -0.1124 & 0.0789 \\ -0.279 & -0.3975 & -0.5592 & -0.4049 & -0.1124 & 1.0001 & -0.1404 \\ 0.8814 & 0.9029 & 0.7957 & 0.9353 & 0.0789 & -0.1404 & 1.0001 \end{bmatrix}$$

对矩阵进行特征值分解后可以得到特征值从大到小为：

$$\hat{\lambda} = [4.6201 \ 1.2311 \ 0.8624 \ 0.1868 \ 0.0840 \ 0.0162 \ 0.0000]$$

Figure 29: 第 1 ~ 20 条数据标准化后的结果

T	R T	G T	B T	Gr T	H T	S T	V T
0	-2.96	-2.88	-2.2	-2.99	0.78	-0.79	-3.26
1	0.8	0.85	0.99	0.91	0.29	-1.35	0.57
2	0.89	0.11	0.63	0.44	2.96	0	0.69
3	1.53	2.35	1.97	2.17	-0.51	-0.35	1.95
4	0.17	0.46	0.89	0.45	1.36	-0.96	0.29
5	0.43	0.59	0.57	0.57	0.77	-0.57	0.41
6	1.06	0.44	0.06	0.62	-1.46	0.46	0.78
7	-0.81	-0.21	1.52	-0.18	1.3	0.75	1.11
8	0.66	0.29	0.54	0.46	1.64	-0.58	0.41
9	-1.17	-0.95	-1.52	-1.15	-0.55	2.27	-1.23
10	0	1.06	1.92	0.9	1.53	0.06	1.41
11	0.99	0.45	0.15	0.61	-0.04	0.13	0.87
12	-3.08	-2.66	-2.43	-2.93	-0.03	2.58	-3.08
13	0.63	1.23	1.8	1.18	1.72	-0.85	1.25
14	-2.32	0.19	0.41	-0.58	0.82	2.2	-0.07
15	0	0.1	0.48	0.12	0.78	-0.16	-0.01
16	-1.49	-0.89	-0.06	-1.02	0.01	0.63	-0.48
17	-0.9	-0.04	0.67	-0.22	1.28	0.04	-0.02
18	-0.23	0.49	-0.52	0.15	-0.14	0.59	0.13
19	-0.09	0.13	0.45	0.11	0.76	-1.1	-0.15

Figure 30: 第 9981 ~ 10000 条数据标准化后的结果

T	R T	G T	B T	Gr T	H T	S T	V T
0	-2.96	-2.88	-2.2	-2.99	0.78	-0.79	-3.26
1	0.8	0.85	0.99	0.91	0.29	-1.35	0.57
2	0.89	0.11	0.63	0.44	2.96	0	0.69
3	1.53	2.35	1.97	2.17	-0.51	-0.35	1.95
4	0.17	0.46	0.89	0.45	1.36	-0.96	0.29
5	0.43	0.59	0.57	0.57	0.77	-0.57	0.41
6	1.06	0.44	0.06	0.62	-1.46	0.46	0.78
7	-0.81	-0.21	1.52	-0.18	1.3	0.75	1.11
8	0.66	0.29	0.54	0.46	1.64	-0.58	0.41
9	-1.17	-0.95	-1.52	-1.15	-0.55	2.27	-1.23
10	0	1.06	1.92	0.9	1.53	0.06	1.41
11	0.99	0.45	0.15	0.61	-0.04	0.13	0.87
12	-3.08	-2.66	-2.43	-2.93	-0.03	2.58	-3.08
13	0.63	1.23	1.8	1.18	1.72	-0.85	1.25
14	-2.32	0.19	0.41	-0.58	0.82	2.2	-0.07
15	0	0.1	0.48	0.12	0.78	-0.16	-0.01
16	-1.49	-0.89	-0.06	-1.02	0.01	0.63	-0.48
17	-0.9	-0.04	0.67	-0.22	1.28	0.04	-0.02
18	-0.23	0.49	-0.52	0.15	-0.14	0.59	0.13
19	-0.09	0.13	0.45	0.11	0.76	-1.1	-0.15

6.3.4 选择最终降维成分的依据

上述的特征值同时也是主成分的方差。我们可以作出主成分累计贡献率折线图, 如图 31 所示。可知当选择 2 个主成分时, 累计贡献率为

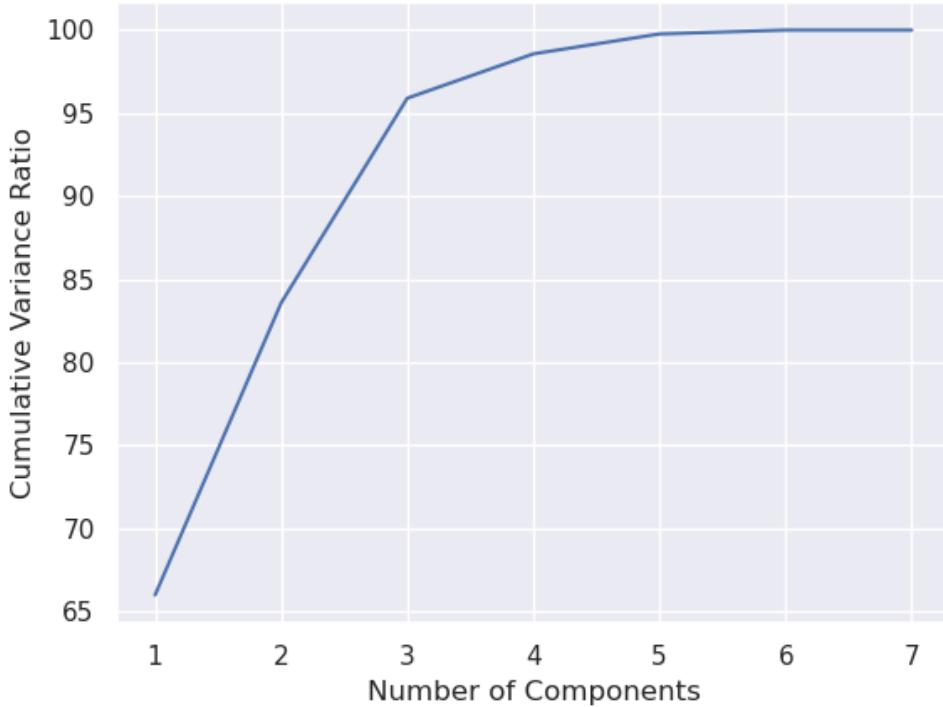
$$\frac{\sum_{k=1}^2 \lambda_k}{\sum_{k=1}^7 \lambda_k} = 0.8358$$

当选择 3 个主成分时, 累计贡献率为

$$\frac{\sum_{k=1}^3 \lambda_k}{\sum_{k=1}^7 \lambda_k} = 0.9590$$

因此我们选择使得累计贡献率超过 85% 的最少的主成分个数, 也即 3 个。

Figure 31: 主成分累计贡献率折线图



6.3.5 降维后新变量与原始变量之间关系

根据矩阵特征值分解的结果, 我们选择了对应 3 个最大特征值的主成分, 它们分别为

$$x'_1 = 0.4128x_1 + 0.4509x_2 + 0.4199x_3 + 0.4613x_4 + 0.0467x_5 - 0.2161x_6 + 0.4339x_7$$

$$x'_2 = 0.2980x_1 + 0.0564x_2 - 0.3045x_3 + 0.0889x_4 - 0.7830x_5 + 0.4148x_6 + 0.1489x_7$$

$$x'_3 = 0.0084x_1 + 0.0182x_2 - 0.0061x_3 + 0.0128x_4 + 0.4965x_5 + 0.8086x_6 + 0.3146x_7$$

其中 x'_1 , x'_2 和 x'_3 分别代表降维后的第一, 第二和第三个特征。从新特征的表达式中我们可以观察到降维后的新变量与原始变量之间的关系:

- 在 x'_1 特征中, 原第一、二、三、四、七特征, 即特征 R, G, B, Gr, V 的系数都较大, 这说明这些特征对于新特征 x'_1 的作用很大。而原第五特征, 即特征 H 的系数很小, 说明它对于新特征 x'_1 的作用很小。
- 在 x'_2 特征中, 原第五特征, 即特征 H 的系数很大, 且显著大于其他六个特征, 说明其对于新特征 x'_2 的作用很大。原第六特征, 即特征 S 的系数较大, 说明其对于新特征 x'_2 具有较大的作用。原第二、四特征, 即特征 G, Gr 的系数很小, 说明它们对于新特征 x'_2 的作用很小。
- 在 x'_3 特征中, 原第六特征, 即特征 S 的系数很大, 且显著大于其他六个特征, 说明其对于新特征 x'_3 的作用很大。原第五特征, 即特征 H 的系数较大, 说明其对于新特征 x'_3 具有较大的作用。原第一、二、三、四特征, 即特征 R, G, B, Gr 的系数很小, 说明它们对于新特征 x'_3 的作用很小。

根据新特征与原始特征之间的表达式, 我们可以将原来具有 7 项特征的 10000 条数据降维到仅有 3 项特征, 且仍然保留了原始数据中 95.90% 的信息。部分降维后的数据如图 32 和图 33 所示。

Figure 32: 第 1 ~ 20 条数据降维后的结果

	Feature1	Feature2	Feature3
0	-6.03	-2.06	-1.38
1	2.1	-0.63	-0.74
2	1.32	-2.09	1.7
3	4.42	0.73	0.15
4	1.26	-1.58	0
5	1.28	-0.74	0.07
6	1.11	1.83	-0.08
7	0.51	-1.27	1.58
8	1.22	-1.38	0.49
9	-3.13	1.14	1.14
10	2.37	-1.41	1.28
11	1.31	0.54	0.38
12	-6.74	0.04	1.01
13	2.92	-1.69	0.59
14	-1.43	-0.6	2.14
15	0.37	-0.81	0.26
16	-1.86	-0.39	0.32
17	-0.17	-1.48	0.65
18	-0.1	0.51	0.46
19	0.47	-1.22	-0.55

6.4 聚类分析

6.4.1 聚类问题

聚类属于无监督学习的一种, 我们希望自动地将所有的样本划分为不同的类别。一个好的聚类结果, 应该有高的簇内相似性和低的簇间相似性。

Figure 33: 第 9981 ~ 10000 条数据降维后的结果

	▼	Feature1 ▼	Feature2 ▼	Feature3 ▼
	9980	-0.92	-0.92	0.38
	9981	1.99	-2.55	0.86
	9982	-5.25	-2.29	1.59
	9983	2.01	1.02	0.26
	9984	2.12	0.52	-0.61
	9985	0.6	0.74	-0.21
	9986	2.41	-0.49	-0.16
	9987	-1.96	-0.49	-0.09
	9988	5.61	1.36	-1.43
	9989	-5.3	-0.18	-3.25
	9990	-0.67	-1.02	0.17
	9991	-0.79	0.52	0.16
	9992	1.36	0.81	-0.09
	9993	-1.26	-1.62	0.19
	9994	-1.74	-0.39	0.47
	9995	-0.3	1.88	1.21
	9996	-0.49	-1.48	1.61
	9997	1.51	-1.36	-0.25
	9998	-4.36	2.36	1.2
	9999	1.43	0.1	-0.46

6.4.2 聚类方法

在 7 项特征中，我们选择 V 和 S 这两项特征进行聚类分析。特征 V 和 S 之间的散点图如图 34 所示。

相似性测度 在相似性测度上，我们选择欧式距离：

$$d(\bar{x}, \bar{y}) = \|\bar{x} - \bar{y}\| = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{\frac{1}{2}}$$

聚类的目标类数 在聚类的目标类数上，我们选择 2, 3, 4 类分别进行实验

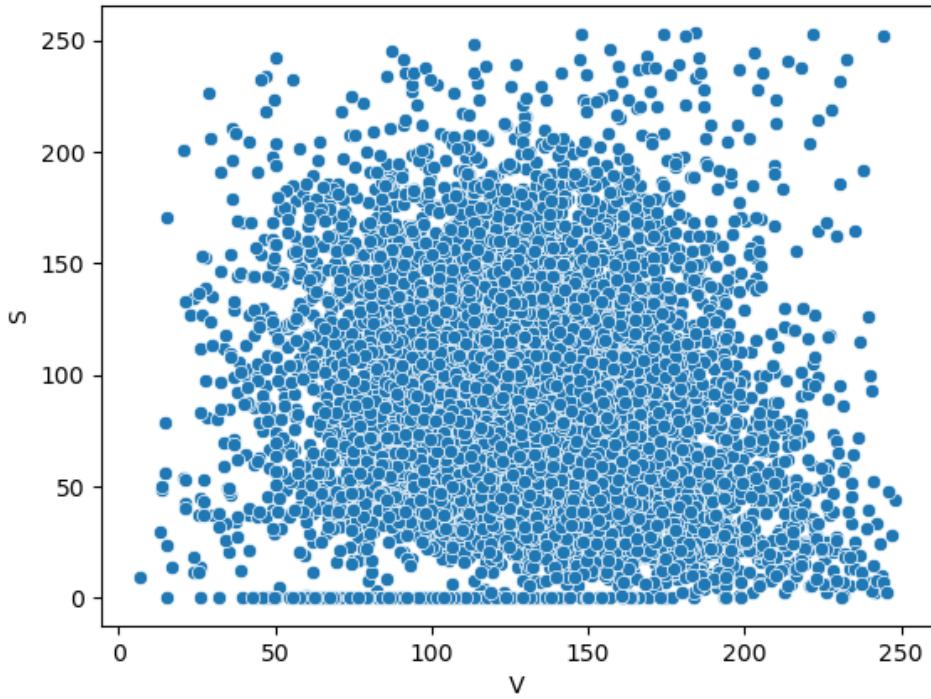
类间距离计算方式 在类间距离计算方式上，我们使用平均距离：

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in w_p, x_j \in w_q} d_{ij}^2$$

聚类算法 在聚类算法上，我们选择 K-Means 算法，对于我们实验所用的大样本数据（样本数为 10000），它具有实现方便且运算速度快的优点。K-Means 算法的流程为：

1. 从 n 个数据对象任意选择 k 个对象作为初始聚类中心
2. 对于待分类的数据点，根据其到各个聚类中心的欧式距离，将其划分到距离最小的簇
3. 对于每一个簇，计算其内所有点的平均值作为新的聚类中心
4. 如果聚类中心不再变化则聚类结束，否则返回第二步

Figure 34: 特征 V 和 S 之间的散点图



6.4.3 聚类过程

当聚类类别数为 3 时，聚类过程如图 35 至 图 44 所示，聚类过程在 10 次迭代后满足终止条件。

此外，我们也做出了聚类类别数为 2 和 4 时的聚类结果，分别如图 45 和 图 46 所示。

6.4.4 聚类结果分析

- 当聚类类别数为 2 时，数据几乎被分成了上下两个部分，也即按照特征 S 的值是否大于 75 分成了 2 个簇。
- 当聚类类别数为 3 时，数据分成了上，左下和右下三个部分，也即先按照特征 S 是否大于 125 分成 2 类。是的部分单独为 1 个簇，不是的部分再按照特征 V 是否大于 125 分成两个簇。
- 当聚类类别数为 4 时，数据分成了上，左下，中右和右下四个部分。可以粗略地看作是在聚类类别为 3 的基础上，对于右下的簇再按照特征 S 是否大于 60 分成两个新的簇。

Figure 35: 3 类别数聚类, 迭代次数为 1 时

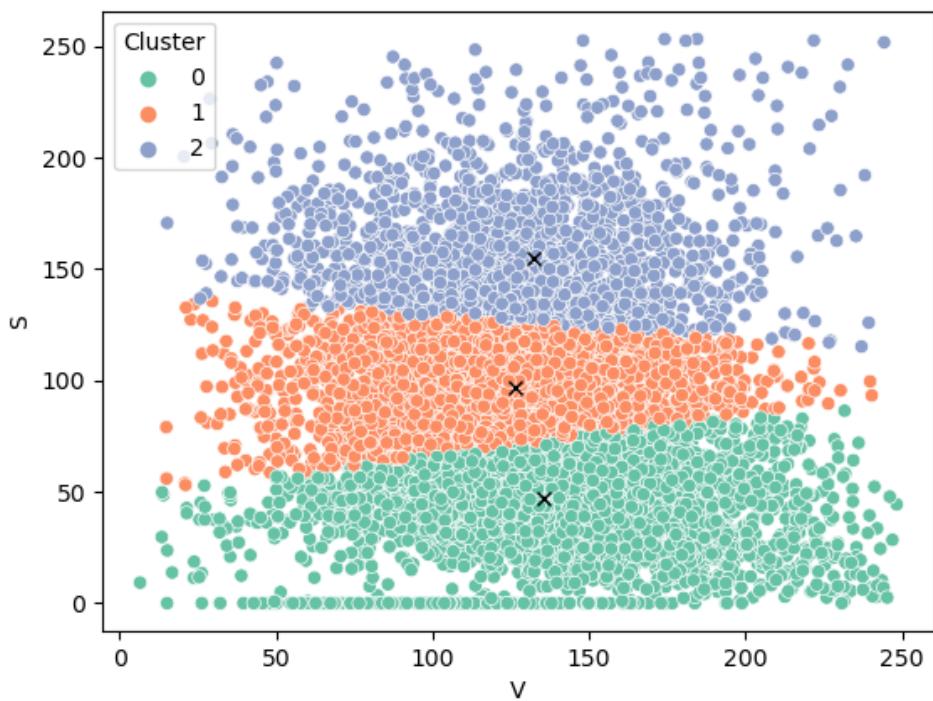


Figure 36: 3 类别数聚类, 迭代次数为 2 时

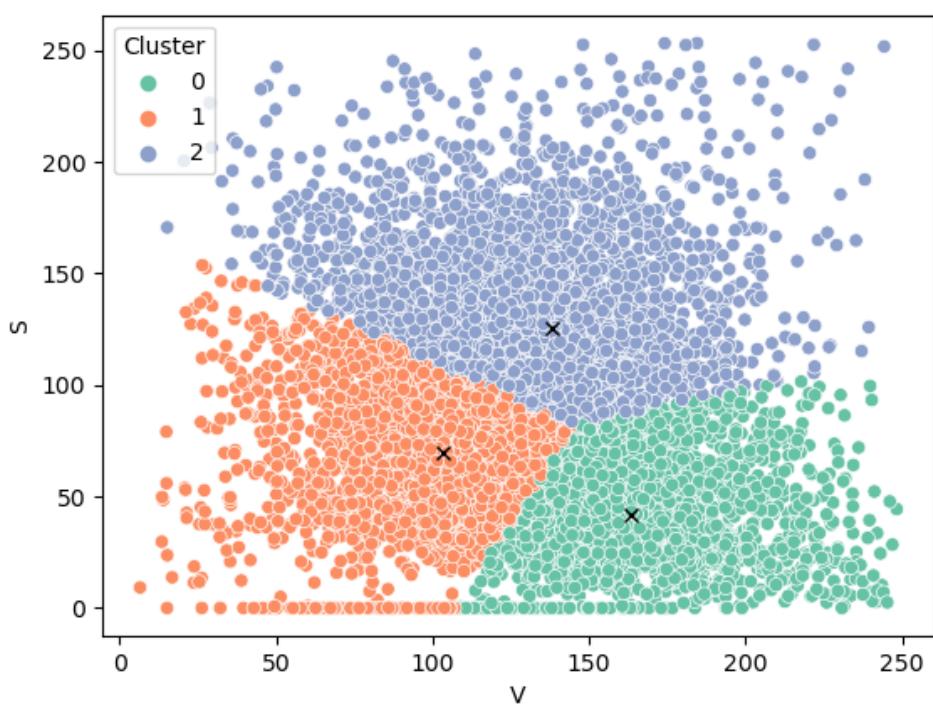


Figure 37: 3 类别数聚类, 迭代次数为 3 时



Figure 38: 3 类别数聚类, 迭代次数为 4 时

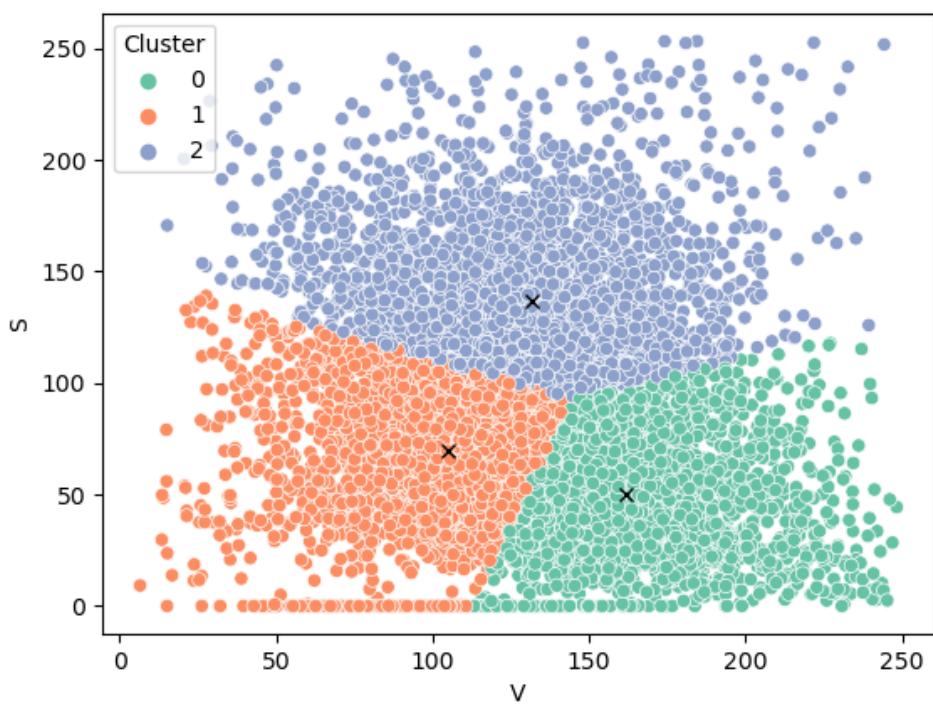


Figure 39: 3 类别数聚类, 迭代次数为 5 时



Figure 40: 3 类别数聚类, 迭代次数为 6 时

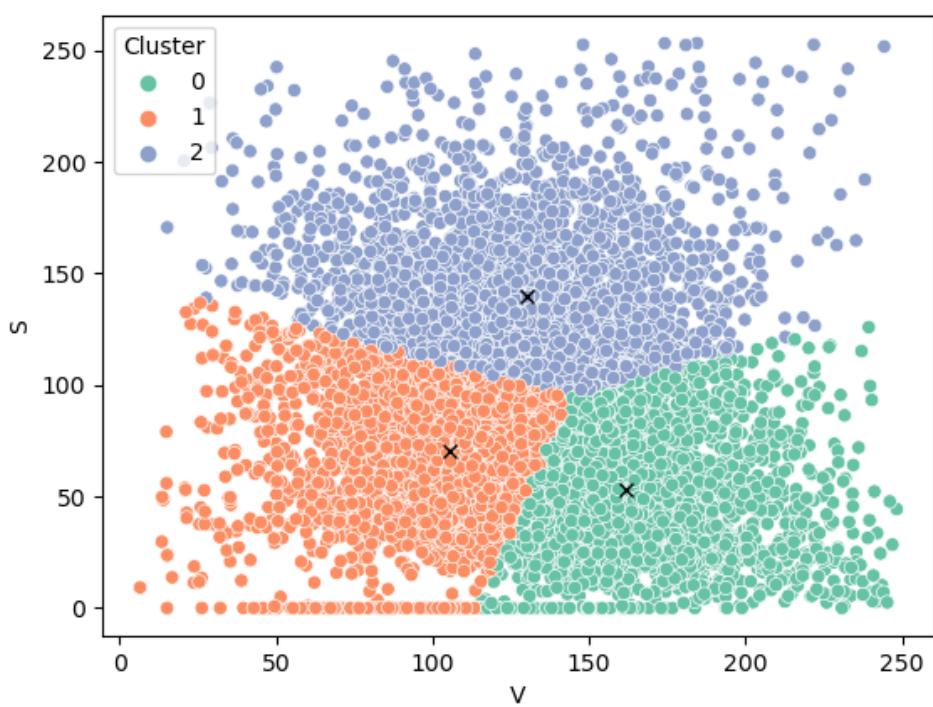


Figure 41: 3 类别数聚类, 迭代次数为 7 时

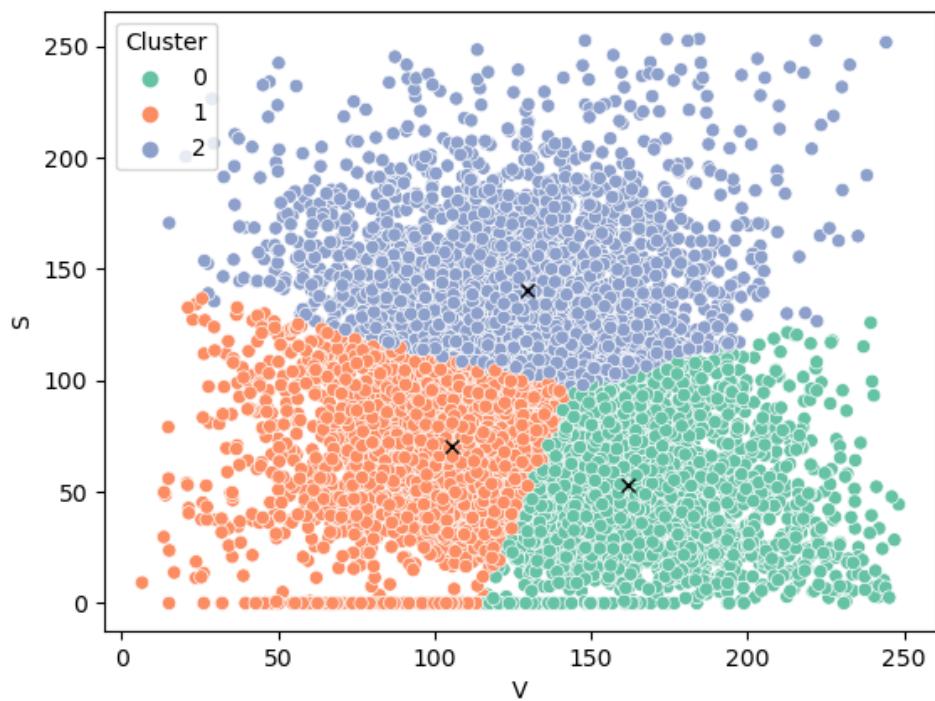


Figure 42: 3 类别数聚类, 迭代次数为 8 时



Figure 43: 3 类别数聚类, 迭代次数为 9 时

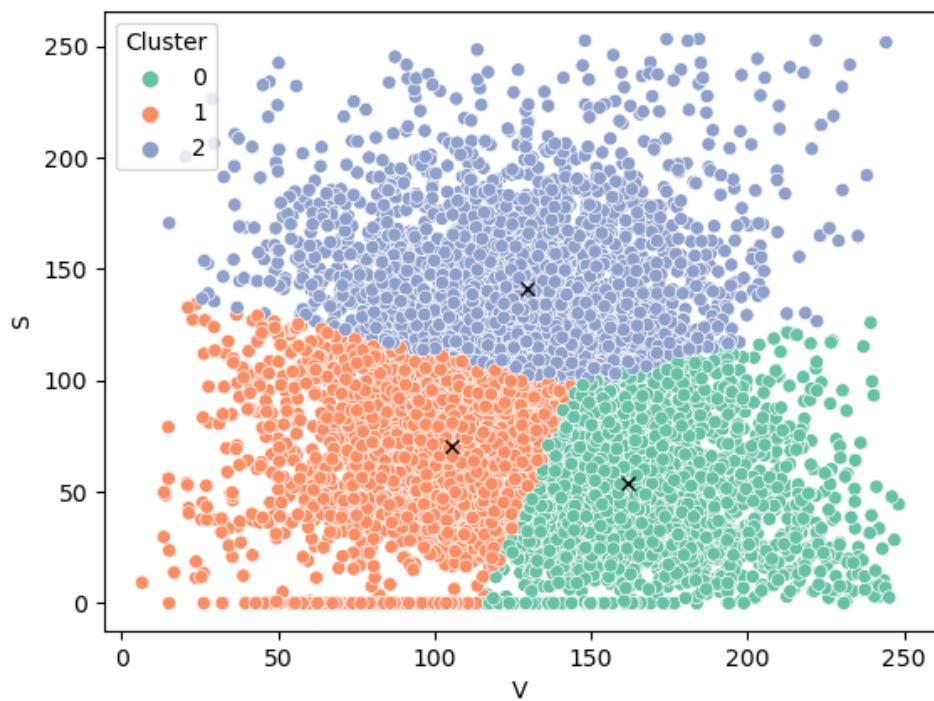


Figure 44: 3 类别数聚类, 迭代次数为 10 时, 聚类完成



Figure 45: 2 类别数聚类, 迭代次数为 7 时, 聚类完成

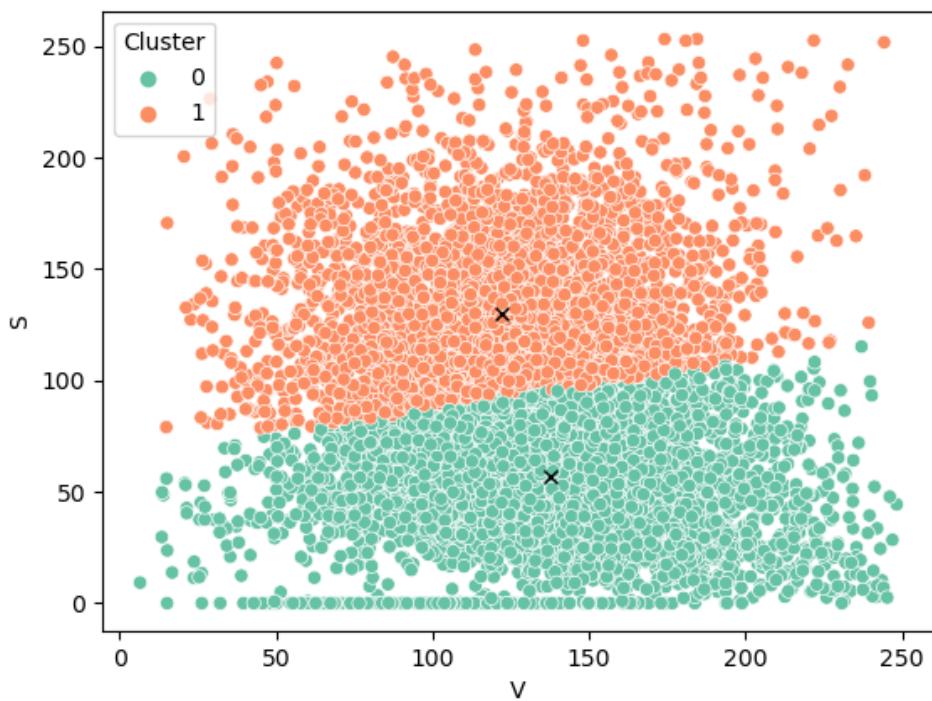
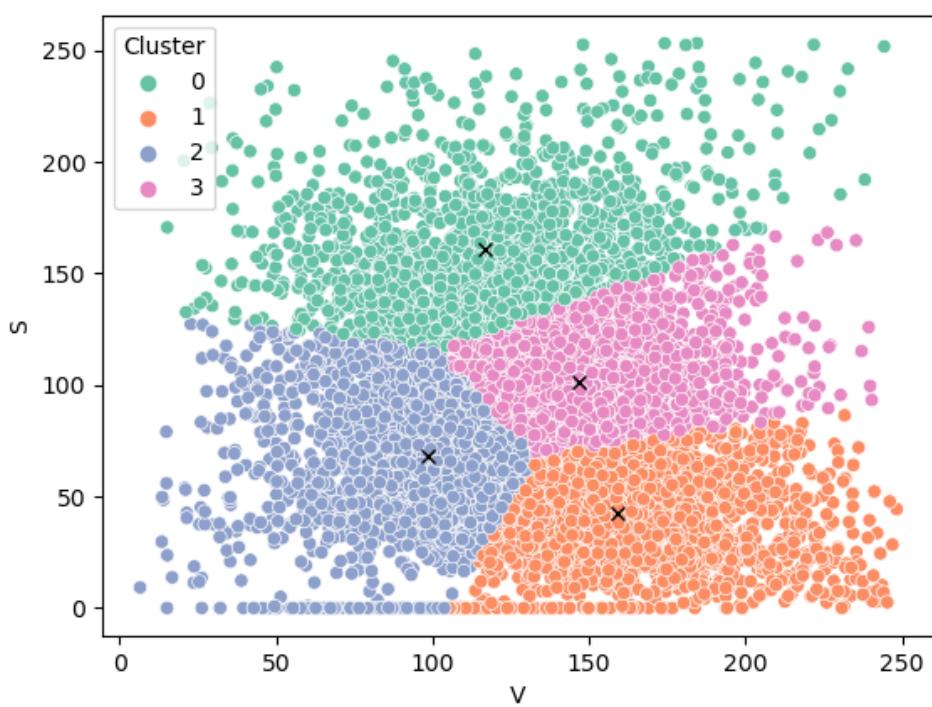


Figure 46: 4 类别数聚类, 迭代次数为 17 时, 聚类完成



7 参考文献

1. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).