姓名:王琦　专业:计算机科学与技术　　编号:58

5.3

数据导入

导入所需要的数据

地区	投入人年数	投入高级职称的人年数	投入科研事业经费	课题总数	专著数	论文数
北京	6795	3737	339803	3261	2723	12270
天津	1649	939	45392	991	488	3055
河北	2367	1039	40631	839	412	4440
山西	1460	658	49661	635	218	2964
内蒙古	455	231	7001	227	152	1759
辽宁	3664	1591	70301	1241	779	7244
吉林	2514	1208	44154	902	581	4300
黑龙江	1430	797	9477	479	391	2801
上海	3783	1833	116292	2247	1130	6607
江苏	5480	2436	138418	3110	961	10456
浙江	2765	1238	44320	1676	473	6031
安徽	2157	982	49672	599	232	3897
福建	1575	710	73829	897	376	3239
江西	2313	1013	15733	908	319	3979
山东	3601	1995	71333	1287	920	10610

共线性诊断

系数[a]

模型		共线性统计 容差	VIF
1	投入人年数	.017	58.979
	投入高级职称的人年数	.005	183.208
	投入科研事业经费	.067	14.907
	专著数	.023	42.668
	论文数	.058	17.218

a. 因变量：课题总数

所有自变量 VIF ＞ 10，存在多重共线性问题，尤其以投入高级职称的人年数共线性问题最大。所以采用 步进 输入。

回归分析

因素 X1:

输入/除去的变量[a]

模型	输入的变量	除去的变量	方法
1	投入人年数	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。

a. 因变量：课题总数

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.938[a]	.879	.870	330.073

a. 预测变量：(常量), 投入人年数

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	10285912.381	1	10285912.381	94.411	<.001[b]
	残差	1416325.219	13	108948.094		
	总计	11702237.600	14			

a. 因变量：课题总数

b. 预测变量：(常量), 投入人年数

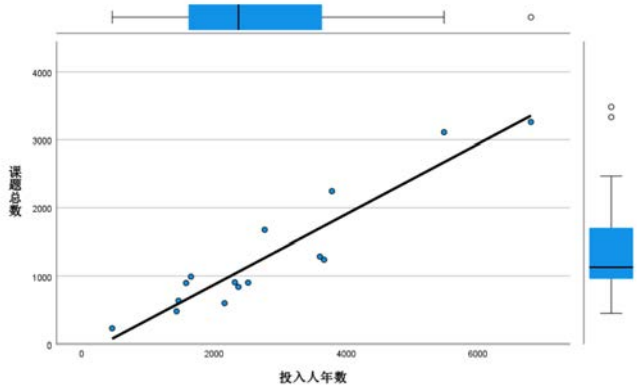
系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-162.547	171.775		-.946	.361
	投入人年数	.517	.053	.938	9.717	<.001

a. 因变量：课题总数

因为投入人年数有较好的线性关系，筛选出投入人年数，方差分析的 $p<0.001$ 回归模型较显著。

回归方程 $X4 = 0.517\ X1 - 162.547 + \varepsilon$



因素 X2：

回归

输入/除去的变量[a]

模型	输入的变量	除去的变量	方法
1	投入高级职称的人年数	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。

a. 因变量：课题总数

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.911[a]	.829	.816	392.029

a. 预测变量：(常量), 投入高级职称的人年数

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	9704305.468	1	9704305.468	63.143	<.001[b]
	残差	1997932.132	13	153687.087		
	总计	11702237.600	14			

a. 因变量：课题总数

b. 预测变量：(常量), 投入高级职称的人年数

系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-16.552	192.718		-.086	.933
	投入高级职称的人年数	.958	.121	.911	7.946	<.001

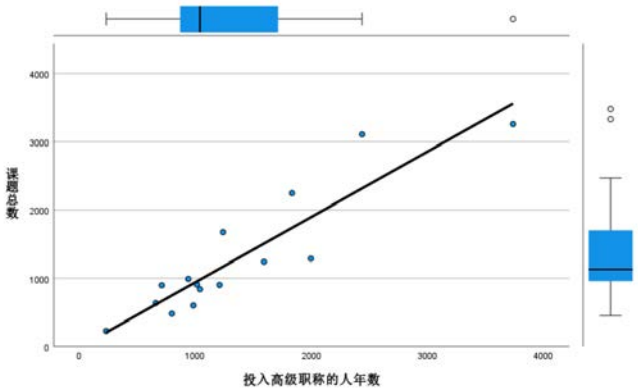
a. 因变量：课题总数

排除的变量[a]

模型		输入 Beta	t	显著性	偏相关	共线性统计 容差
1	投入科研事业经费	.082[b]	.283	.782	.081	.167
	专著数	-.295[b]	-.828	.424	-.232	.106
	论文数	-.007[b]	-.020	.985	-.006	.111

a. 因变量：课题总数

b. 模型中的预测变量：(常量), 投入高级职称的人年数



去除投入人年数后进行回归分析，投入高级职称的人年数对课题总数的也有较好的线性关系，方差分析的 p 值<0.001,线性回归较显著

回归方程 X4 = 0.958 X2 - 166.552 + ε

因素 X3 X6:

回归

输入/除去的变量[a]

模型	输入的变量	除去的变量	方法
1	论文数	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。
2	投入科研事业经费	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。

a. 因变量：课题总数

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.858[a]	.736	.716	487.430
2	.903[b]	.816	.786	423.315

a. 预测变量：(常量)，论文数

b. 预测变量：(常量)，论文数，投入科研事业经费

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	8613596.742	1	8613596.742	36.254	<.001[b]
	残差	3088640.858	13	237587.758		
	总计	11702237.600	14			
2	回归	9551890.760	2	4775945.380	26.652	<.001[c]
	残差	2150346.840	12	179195.570		
	总计	11702237.600	14			

a. 因变量：课题总数

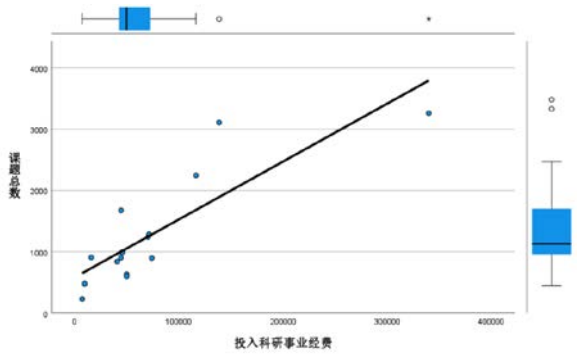
b. 预测变量：(常量)，论文数

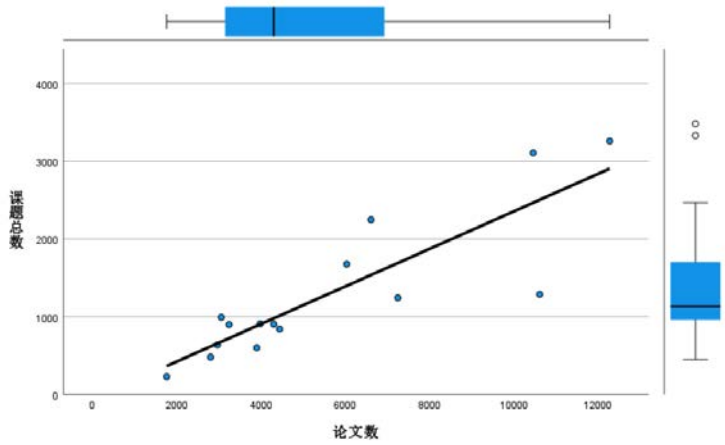
c. 预测变量：(常量)，论文数，投入科研事业经费

系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-62.366	256.967		-.243	.812
	论文数	.242	.040	.858	6.021	<.001
2	(常量)	113.288	236.000		.480	.640
	论文数	.143	.055	.508	2.582	.024
	投入科研事业经费	.005	.002	.450	2.288	.041

a. 因变量：课题总数





除去上面两种因素后，论文数和投入科研事业经费同时被筛选出来，但是拆分出两个模型

一元线性回归：

方差分析 p 值<0.01，表明论文数与课题数有较为显著的线性关系。

线性回归方程：X4 = 0.242 X6 - 62.366 + ε

二元线性回归：

方差分析 p 值

方差分析 p 值<0.01，所以这两个因素与课题数有较显著的线性相关性。但是回归系数的显著性不如一元线性回归

线性回归方程:X6 = 0.143 X6 + 0.005X3 +113.288 + ε

因素 X5:

输入/除去的变量[a]

模型	输入的变量	除去的变量	方法
1	专著数[b]	.	输入

a. 因变量：课题总数

b. 已输入所请求的所有变量。

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.830[a]	.688	.664	529.818

a. 预测变量：(常量), 专著数

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	8053047.667	1	8053047.667	28.688	<.001[b]
	残差	3649189.933	13	280706.918		
	总计	11702237.600	14			

a. 因变量：课题总数

b. 预测变量：(常量), 专著数

系数[a]

模型		未标准化系数		标准化系数		
		B	标准错误	Beta	t	显著性
1	(常量)	480.710	203.352		2.364	.034
	专著数	1.190	.222	.830	5.356	<.001

a. 因变量：课题总数

专著数也与课题总数的也有较好的线性关系，方差分析的 p 值<0.001,线性回归较显著

回归方程 $X4 = 1.190\ X5 + 480.710 + \varepsilon$

6.5

数据导入

导入对应数据

	X1	X2	Y1	Y2	变量
1	191.00	155.00	179.00	145.00	
2	195.00	149.00	201.00	152.00	
3	181.00	148.00	185.00	149.00	
4	183.00	153.00	188.00	149.00	
5	176.00	144.00	171.00	142.00	
6	208.00	157.00	192.00	152.00	
7	189.00	150.00	190.00	149.00	
8	197.00	159.00	189.00	152.00	
9	188.00	152.00	197.00	159.00	
10	192.00	150.00	187.00	151.00	
11	179.00	158.00	186.00	148.00	
12	183.00	147.00	174.00	147.00	
13	174.00	150.00	185.00	152.00	
14	190.00	159.00	195.00	157.00	
15	188.00	151.00	187.00	158.00	
16	163.00	137.00	161.00	130.00	
17	195.00	155.00	183.00	158.00	
18	186.00	153.00	173.00	148.00	
19	181.00	145.00	182.00	146.00	
20	175.00	140.00	165.00	137.00	
21	192.00	154.00	185.00	152.00	
22	174.00	143.00	178.00	147.00	
23	176.00	139.00	176.00	143.00	
24	197.00	167.00	200.00	158.00	
25	190.00	163.00	187.00	150.00	
26					

典型相关性分析

模型分析

相关性[a]

		X1	X2	Y1	Y2
X1	皮尔逊相关性	1	.735	.711	.704
	显著性（双尾）		<.001	<.001	<.001
X2	皮尔逊相关性	.735	1	.693	.709
	显著性（双尾）	<.001		<.001	<.001
Y1	皮尔逊相关性	.711	.693	1	.839
	显著性（双尾）	<.001	<.001		<.001
Y2	皮尔逊相关性	.704	.709	.839	1
	显著性（双尾）	<.001	<.001	<.001	

a. 成列 N=25

X1 与 X2 的相关系数为 0.735
Y1 与 Y2 的相关系数为 0.839

典型相关性

	相关性	特征值	威尔克统计	F	分子自由度	分母自由度	显著性
1	.789	1.644	.377	6.597	4.000	42.000	.000
2	.054	.003	.997	.064	1.000	22.000	.803

H0 for Wilks 检验是指当前行和后续行中的相关性均为零

第一组显著性<0.05 有意义
第二组显著性>0.05 没意义
因此弟弟头型数据与兄长头型数据间的相关系数为 0.789，高度相关。

集合 1 标准化典型相关系数

变量	1	2
X1	-.552	-1.366
X2	-.522	1.378

集合 2 标准化典型相关系数

变量	1	2
Y1	-.504	-1.769
Y2	-.538	1.759

由前文只，可以抽取出典型变量 u,v.
U = -0.552*X1 - 0.522*X2
V = -0.504*Y1 - 0.538*Y2

已解释的方差比例

典型变量	集合 1 * 自身	集合 1 * 集合 2	集合 2 * 自身	集合 2 * 集合 1
1	.867	.539	.920	.572
2	.133	.000	.080	.000

U可以代表86.7%
V可以代表92%
第一对典型变量有较好的代表性。

协方差矩阵

项间协方差矩阵

	X1	X2	Y1	Y2
X1	95.293	52.868	69.662	46.112
X2	52.868	54.360	51.312	35.053
Y1	69.662	51.312	100.807	56.540
Y2	46.112	35.053	56.540	45.023

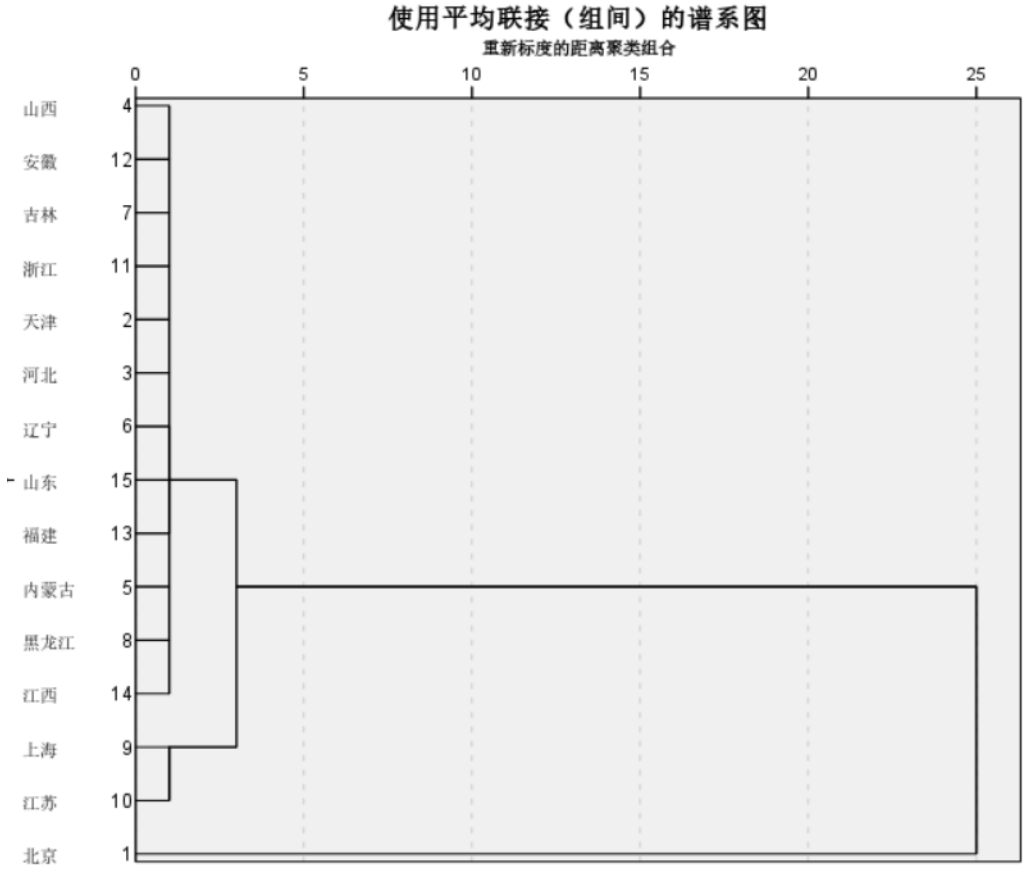
7.3

快速聚类

操作步骤
【分析】-【分类】-【K-均值聚类】

谱系聚类法绘制聚类树形图

采用系统聚类，个案标注依据为【地区】

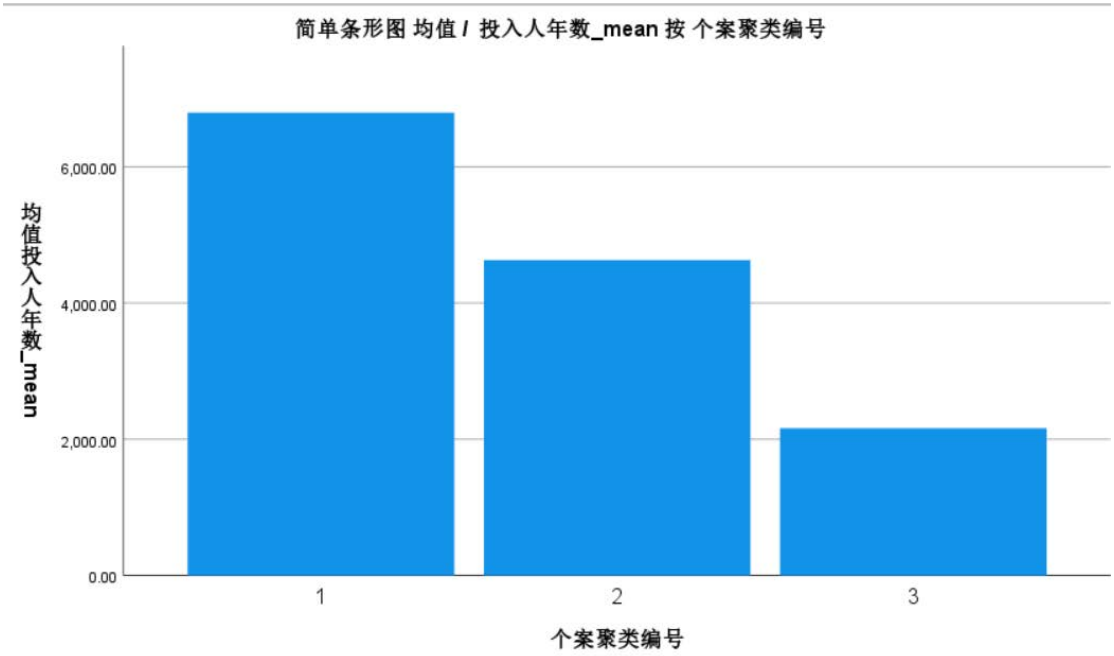


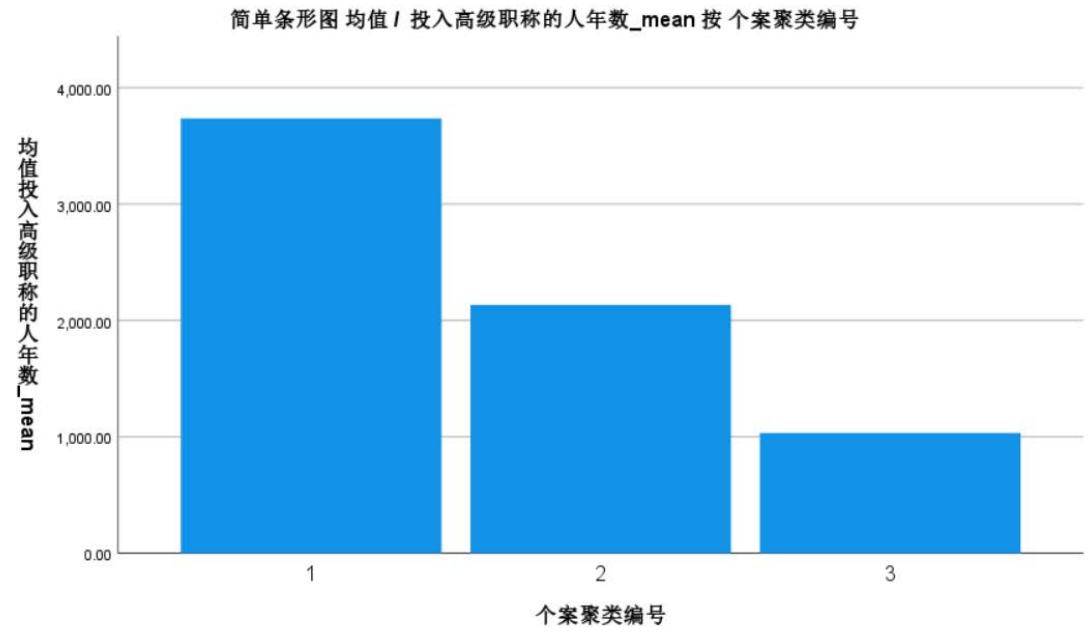
从上至下山西到江西高校可以聚集在一起，上海和江苏可以聚在一起
若只分两类则从上至下山西到江苏可以聚集在一起。

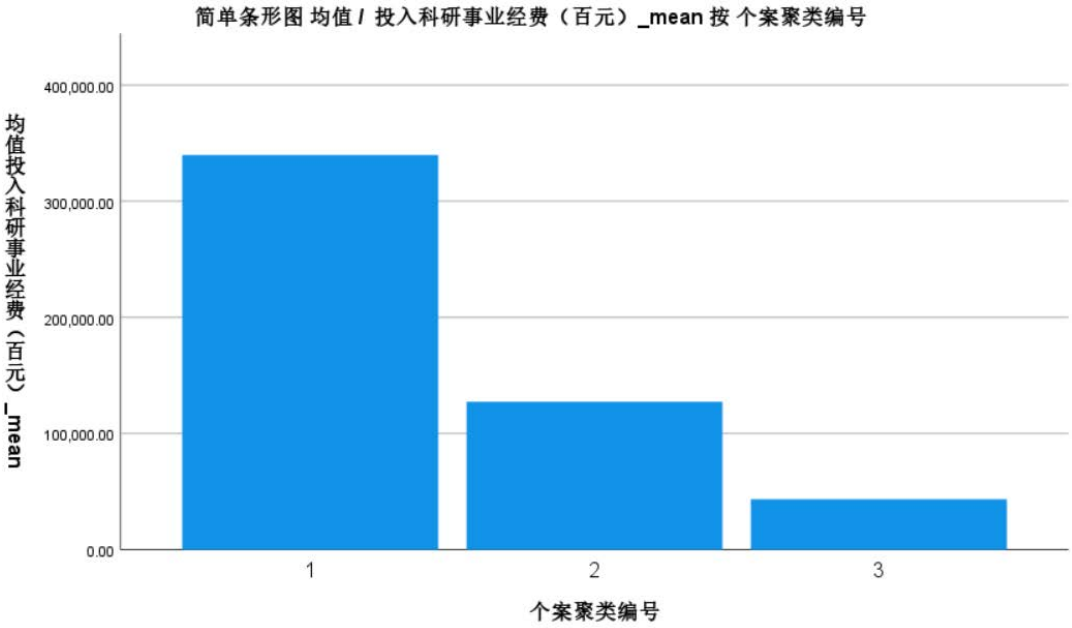
绘制各类科研指标的均值对比图

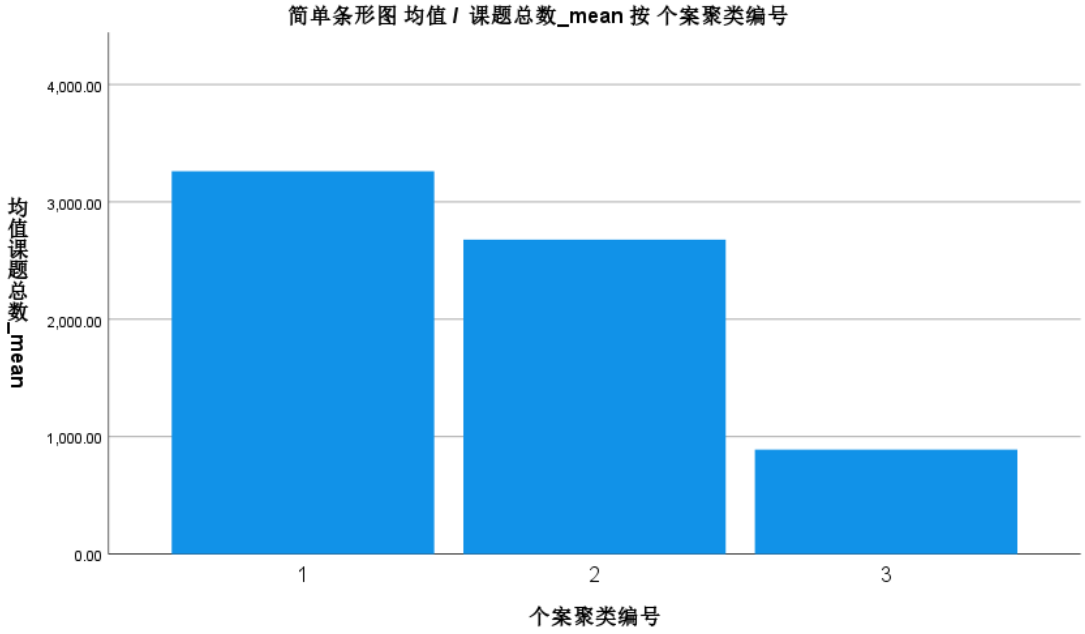
汇总数据，数据集名称为均值。

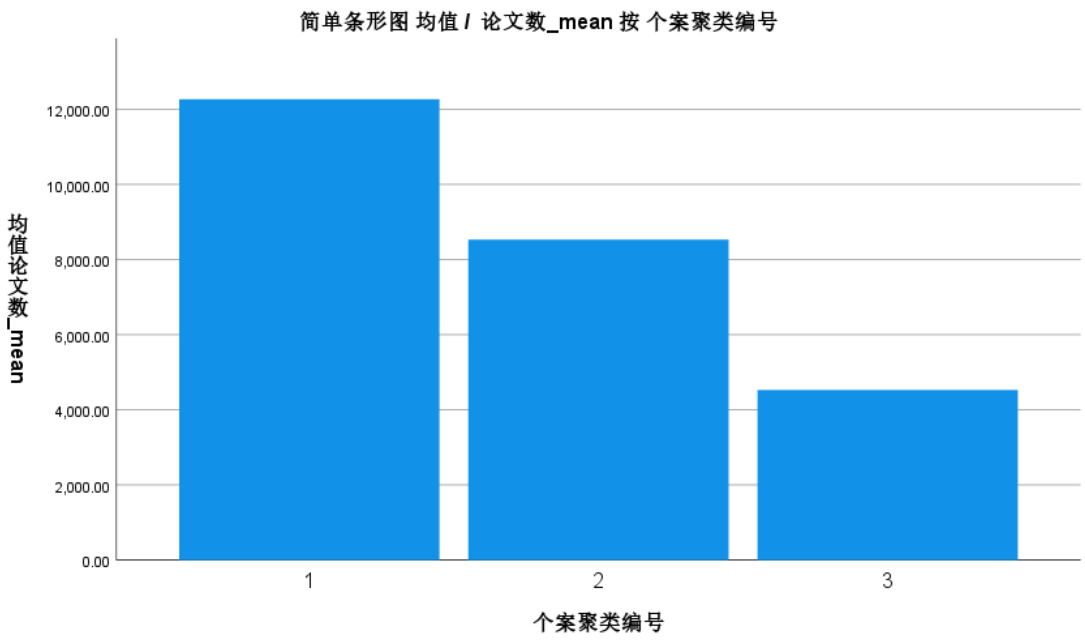
然后依次构建直方图。

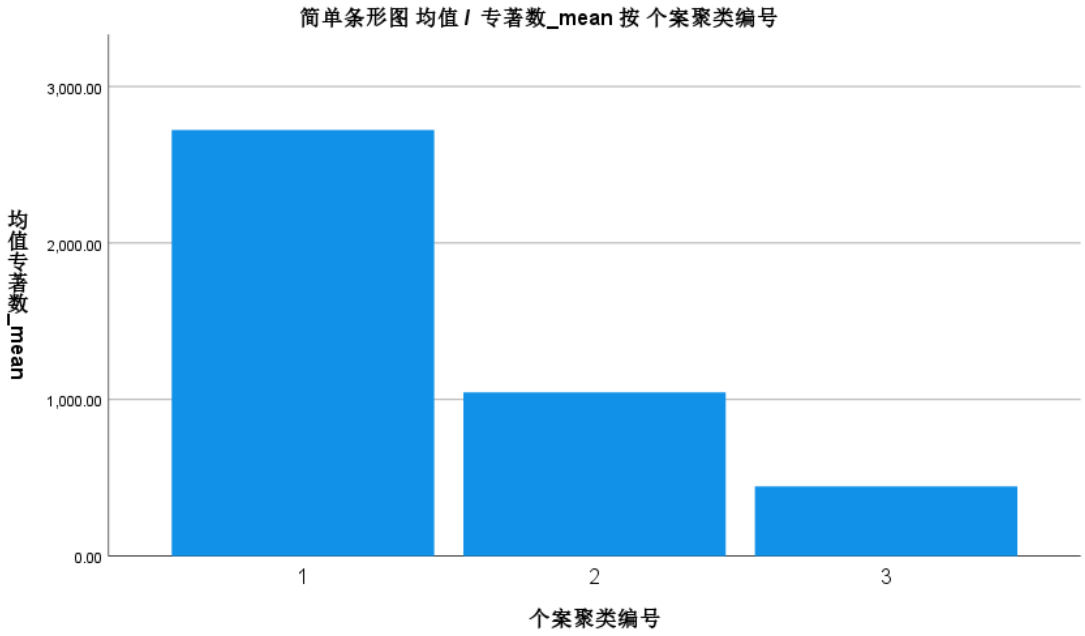

简单条形图 均值 / 投入人年数_mean 按 个案聚类编号
均值投入人年数_mean
6,000.00
4,000.00
2,000.00
0.00
1
2
3
个案聚类编号


简单条形图 均值 / 投入高级职称的人年数_mean 按 个案聚类编号
均值投入高级职称的人年数_mean
4,000.00
3,000.00
2,000.00
1,000.00
0.00
1
2
3
个案聚类编号


简单条形图 均值 / 投入科研事业经费（百元）_mean 按 个案聚类编号
均值投入科研事业经费（百元）_mean
400,000.00
300,000.00
200,000.00
100,000.00
0.00
1
2
3
个案聚类编号


简单条形图 均值 / 课题总数_mean 按 个案聚类编号
均值课题总数_mean
4,000.00
3,000.00
2,000.00
1,000.00
0.00
1
2
3
个案聚类编号





利用方差分析各类方法分析各类在哪些指标上存在显著差异

各个因素方差分析 p 值均小于 0.05 ， 所以都很显著。

ANOVA

		平方和	自由度	均方	F	显著性
投入人年数	组间	27545680.233	2	13772840.117	15.206	<.001
	组内	10869205.500	12	905767.125		
	总计	38414885.733	14			
投入高级职称的人年数	组间	8129706.317	2	4064853.158	19.934	<.001
	组内	2446987.417	12	203915.618		
	总计	10576693.733	14			
投入科研事业经费（百元）	组间	87535609741	2	43767804871	87.429	<.001
	组内	6007287432.7	12	500607286.06		
	总计	93542897174	14			
课题总数	组间	9659732.183	2	4829866.092	28.376	<.001
	组内	2042505.417	12	170208.785		
	总计	11702237.600	14			
专著数	组间	5103124.583	2	2551562.292	52.793	<.001
	组内	579979.417	12	48331.618		
	总计	5683104.000	14			
论文数	组间	75494890.983	2	37747445.492	6.316	.013
	组内	71720291.417	12	5976690.951		
	总计	147215182.40	14			