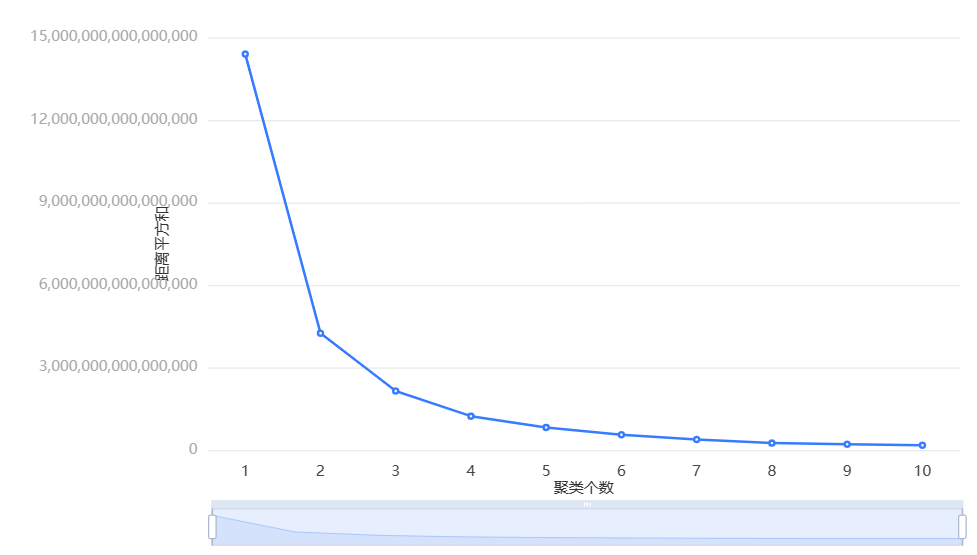
### 分析流程 数据源： 城市人口密度排名\_副本(1).xlsx 算法配置： 算法： 聚类分析(K-Means) 变量： 变量:{Population Estimate，Square Kilometers} 参数： 聚类个数:{5} 分析结果： 聚类分析基于数据特征将所有样本划分为几类： 聚类结果共分为5类， 聚类类别\_1的频数为143，所占百分比为14.712%； 聚类类别\_2的频数为19，所占百分比为1.955%； 聚类类别\_3的频数为758，所占百分比为77.984%； 聚类类别\_4的频数为9，所占百分比为0.926%； 聚类类别\_5的频数为43，所占百分比为4.424%。 各聚类类别的特征看详细结论。

### 分析步骤 1. 根据字段进行聚类类别差异性分析。 2. 根据聚类汇总分析各聚类类别的频数。 3. 根据数据集聚类标注可以知道每一个样本数据被分到哪个类别。 4. 聚类中心坐标可以用于分析各样本与中心点的距离。 5. 对分析进行综述。

### 详细结论

**输出结果1：聚类数对比图（肘部法则）**



**图表说明：**

该图用于选择较好的聚类数量，横坐标是聚类个数，纵坐标是K均值聚类的损失函数是所有样本到类别中心的距离平方和，也就是误差平方和（值越小说明聚类效果越好）。可以通过“坡度趋于平缓”的找出最佳的类簇数量。

我的分析：我们从表中可以主观判断4之后比较好，为了方便研究，我们将其分成了5类城市，分别是小城市、中小城市、中等城市、中大城市、大城市

**输出结果2：字段差异性分析**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 聚类类别（平均值±标准差） | | | | | F | P |
| 类别3(n=758) | 类别1(n=143) | 类别5(n=43) | 类别2(n=19) | 类别4(n=9) |
| Population Estimate | 1049296.834±491403.487 | 3882370.629±1049850.611 | 8981976.744±1974328.608 | 17235000.0±2221086.421 | 27858222.222±5358298.722 | 3994.543 | 0.000\*\*\* |
| Square Kilometers | 305.439±335.098 | 946.315±1017.119 | 1975.814±1889.251 | 3208.105±2670.117 | 3588.222±2087.7 | 165.194 | 0.000\*\*\* |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | |

**图表说明：**

上表展示了定量字段差异性分析的结果，包括均值±标准差的结果、F检验结果、显著性P值。  
● 分析每个分析项的P值是否显著(P<0.05)。  
● 若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。

**智能分析**

方差分析的结果显示:  
对于变量Population  
Estimate，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，说明变量Population  
Estimate在聚类分析划分的类别之间存在显著性差异；  
对于变量Square  
Kilometers，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，说明变量Square  
Kilometers在聚类分析划分的类别之间存在显著性差异；

**输出结果3：聚类汇总**

|  |  |  |
| --- | --- | --- |
| 聚类类别 | 频数 | 百分比% |
| 聚类类别\_1 | 143 | 14.712 |
| 聚类类别\_2 | 19 | 1.955 |
| 聚类类别\_3 | 758 | 77.984 |
| 聚类类别\_4 | 9 | 0.926 |
| 聚类类别\_5 | 43 | 4.424 |
| 合计 | 972 | 100.0 |

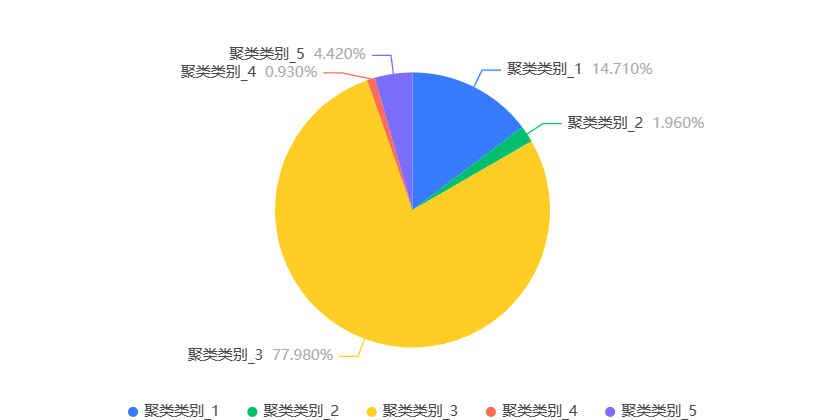
**图表说明：**

上表展示了模型聚类的结果，包括频数，所占百分比。

**智能分析**

聚类分析的结果显示，聚类结果共分为5类，  
聚类类别\_1的频数为143，所占百分比为14.712%；  
聚类类别\_2的频数为19，所占百分比为1.955%；  
聚类类别\_3的频数为758，所占百分比为77.984%；  
聚类类别\_4的频数为9，所占百分比为0.926%；  
聚类类别\_5的频数为43，所占百分比为4.424%。

**输出结果4：聚类汇总图**



**图表说明：**

上图以可视化的形式展示了模型聚类的结果，包括频数、所占百分比。

**输出结果5：数据集聚类标注**

|  |  |  |
| --- | --- | --- |
| 聚类种类 | Population Estimate | Square Kilometers |
| 2 | 18627000 | 619 |
| 3 | 648000 | 23 |
| 3 | 717000 | 26 |
| 5 | 6538000 | 238 |
| 5 | 12836000 | 474 |
| 1 | 3682000 | 137 |
| 1 | 3247000 | 122 |
| 3 | 1348000 | 52 |
| 5 | 7450000 | 290 |
| 3 | 729000 | 28 |
| 4 | 24973000 | 976 |
| 3 | 1636000 | 67 |
| 3 | 737000 | 31 |
| 3 | 838000 | 36 |
| 3 | 826000 | 36 |

**图表说明：**

上表格展示了模型聚类结果的部分数据聚类标注，其为预览结果，只显示综合排序的前15条数。

**输出结果6：聚类中心点坐标**

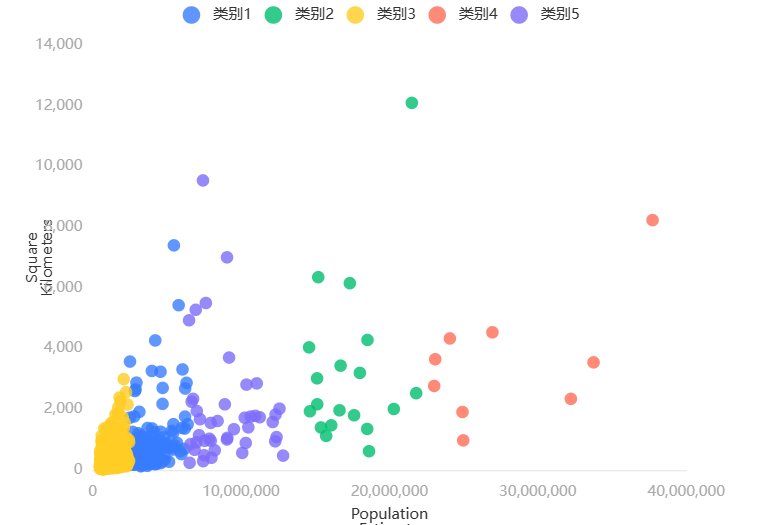
|  |  |  |
| --- | --- | --- |
| 聚类种类 | 中心值\_Population Estimate | 中心值\_Square Kilometers |
| 1 | 3882370.629370628 | 946.3146853146851 |
| 2 | 17234999.999999996 | 3208.105263157894 |
| 3 | 1049296.8337730826 | 305.43931398416896 |
| 4 | 27858222.222222216 | 3588.2222222222217 |
| 5 | 8981976.744186044 | 1975.813953488372 |

**图表说明：**

上表展示了部分（or全部）模型聚类中心的数据，全部数据可点击右上角下载excel。

我的分析：对图中的中心值排序我们可以判断，4为大城市、2为中大规模、5为中等规模、3为小城市、1为中小城市。

**输出结果7：聚类散点图**



**图表说明：**

若变量数等于两个，上图是根据两变量的数据绘制出来的散点图；若变量数大于两个，上图是提取主成分分析(PCA)降维后前两个主成分来绘制散点图，在一定程度上可查看聚类效果（若是前两个主成分的方差解释率较低，该图的意义不大）。  
● 聚类散点图最大仅显示1000个样本量信息，若样本量大于1000，则在全样本中进行随机抽样，选取其中1000样本来进行散点图展示。

**输出结果8：评价指标**

|  |  |  |
| --- | --- | --- |
| 轮廓系数 | DBI | CH |
| 0.726 | 0.525 | 3994.54 |

**图表说明：**

● 轮廓系数：对于一个样本集合，它的轮廓系数是所有样本轮廓系数的平均值。轮廓系数的取值范围是[-1,1]，同类别样本距离越相近不同类别样本距离越远，分数越高，聚类效果越好。  
● DBI（Davies-bouldin)：该指标用来衡量任意两个簇的簇内距离之后与簇间距离之比。该指标越小表示聚类效果越好。  
● CH(Calinski-Harbasz Score)：通过计算类内各点与类中心的距离平方和来度量类内的紧密度（分母），通过计算类间中心点与数据集中心点距离平方和来度量数据集的分离度（分子），CH指标由分离度与紧密度的比值得到，CH越大表示聚类效果越好。

### 参考文献 [1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com. [2] Saroj,Kavita.Review:study on simple k mean and modified K mean clustering technique[J].International Journal of Computer Science Engineering and Technology,2016,6(7)：279-281.