工作任务四：数据分析

（一）数据关联

通常情况下，抓取到的数据是零散的。因此，在分析数据之前，首先要将数据分类规整，进而在数据中建立必要联系，最后总结规律并尝试从中挖掘深层信息。在建立数据关联时可进行如下思考：

（1）数据中包含了哪些维度（如：时间、地点、产品等）？

（2）要分析的问题（核心要素）是什么？

（3）要解决这个问题，需要的主要数据维度是什么？

（4）哪些要素(关联要素)影响这个问题的发生和发展？它们间可以形成有效关联吗？

（5）通过建立数据维度A和数据维度B之间的关系可以更好地回答所提出的问题吗？

数据关系类型中除了相关关系外，还有因果关系类型。一般来说，因果关系(A→B)需满足以下3个条件：

（1）A和B相关；

（2）A必须发生在B之前；

（3）所有其他的因素C都已经被排除。

虽然并非所有的数据都会呈现因果关系，但在数据分析中，因果关系尤为重要。

（二）数据分析

常见的**数据应用分析方式**有四种：

1.描述性分析：将大量的原始数据资料进行初步的整理和归纳，找出数据中的集中趋势和分散趋势，用于描述某个事物的整体或局部面貌或呈现的某个过程，可回答“发生了什么？”这类问题。

如从已完成清洗的数据中可以看出，表中包含有年龄、职业、每月可支配收入、各活动支出占比等数据信息。项目选题可采用描述性分析方法，通过分析数据得出00后的消费倾向和消费特征。

2.诊断性分析：深入了解数据背后的原因和关系，常用于找出数据中的趋势或异常，并挖掘现象背后的潜在原因和更深层次的问题。可回答“为什么会发生”这类问题。

3.预测性分析：通过分析数据推算事物的发展趋势。这种类型对数据及其来源的准确性要求较高，而且要求分析者具有严谨的数据逻辑推理能力。可回答“将来会怎样”的问题。

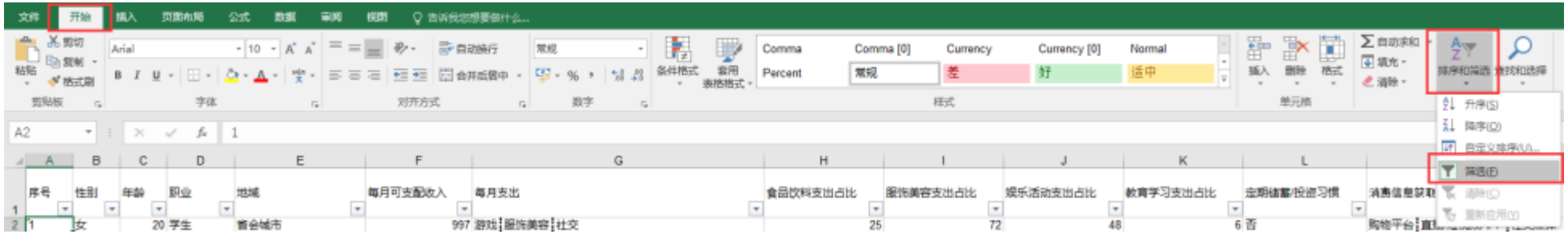
4.规范性分析：在诊断性分析或预测性分析的基础上，通过数据分析提出合理性建议的一种分析类型。

Excel中的“排序和筛选”功能可以将样本中“无序”的数据序列按照一定的规律调整为“有序”，有利于掌握样数据的趋势和分布。

1.筛选

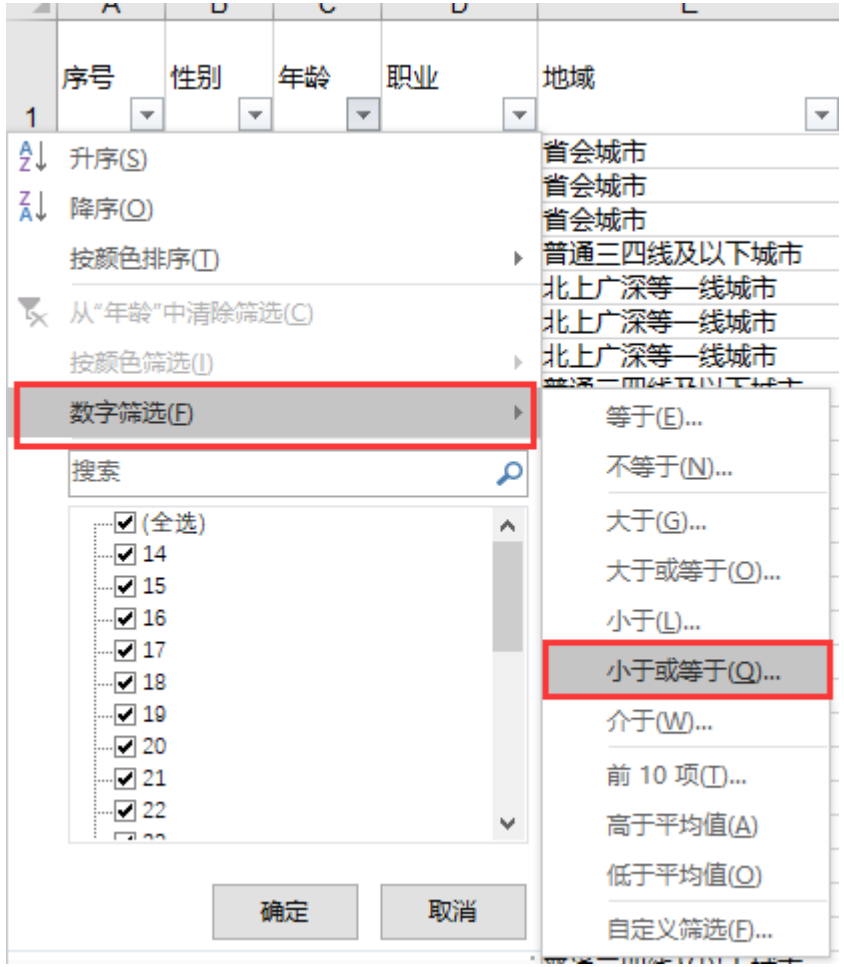
“筛选”可以快速查看数据中的子集，找到想要的值，也可以排除不需要查看的内容，不符合筛选条件时整行数据都会被隐藏，仅保留符合筛选条件的数据。

第一步，在“开始”面板中，点击“排序和筛选”，选择“筛选”选项，或者在“数据”面板中直接点击“筛选”来开启自动筛选功能，开启筛选功能后，每列标题的右下角会出现倒三角 下拉菜单。

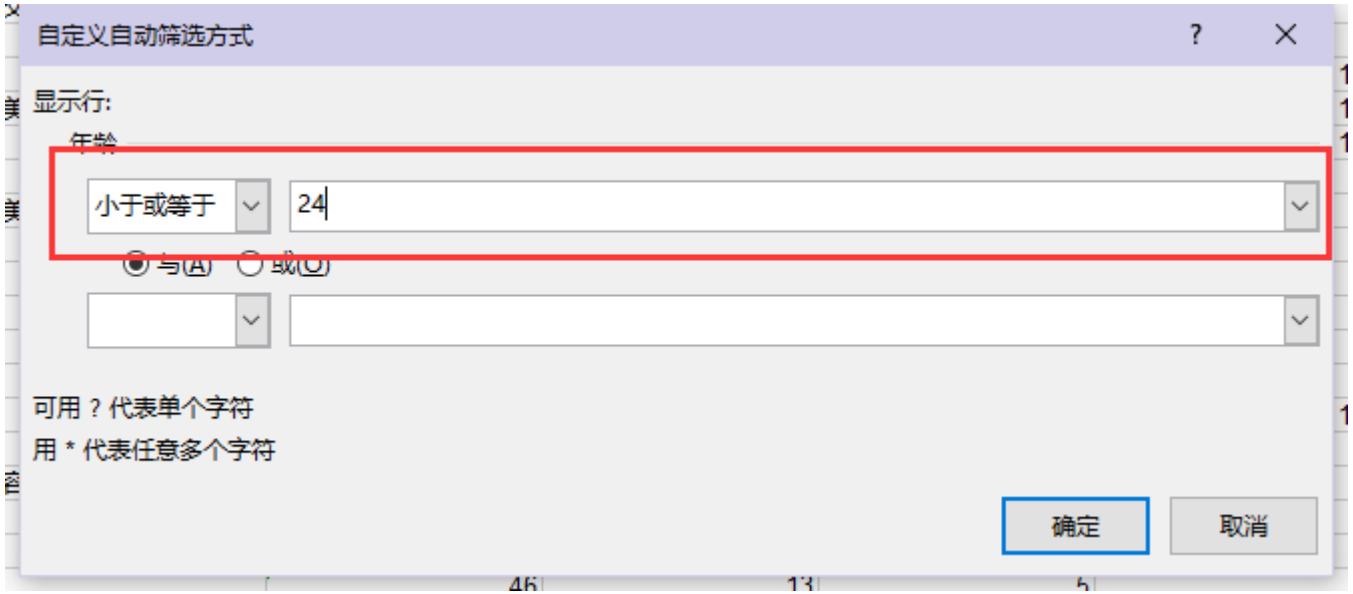


以筛选出所有00后样本数据为例：

第一步：开启筛选后打开“年龄”列筛选菜单，选择“数字筛选”→“小于或等于”。



第二步：打开自定义自动筛选窗口，将值设置为“24”。



第三步：点击“确定”完成数字筛选设置，所有不符合设定条件的数据会被隐藏。

注意：在启用筛选功能时，应确保每列数据都有明确的字段名称。

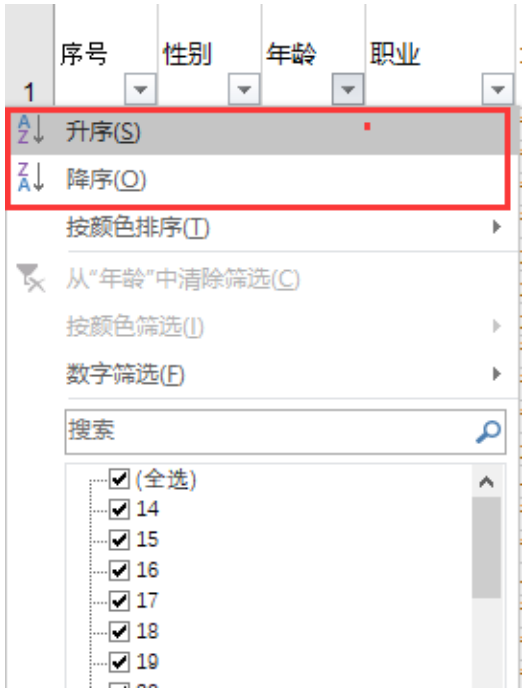
2.自动排序

Excel自动排序可以对文本格式和数值格式的数据按照一定规律进行排列。使用排序功能可以重新排列数据以快速查找某些值，例如快速查看最大值和最小值，掌握数据区间和极值。自动排序仅可依据一列数据进行排序。

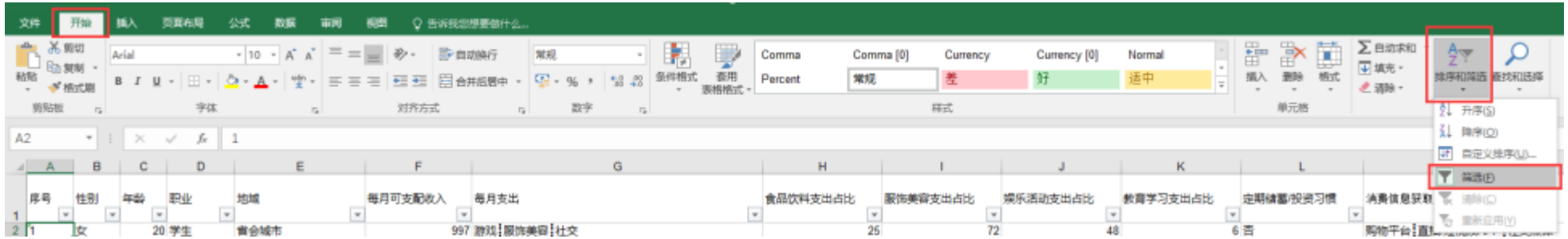
通常Excel的排序是按照垂直方向进行的，自动排序可以依据升序和降序进行排列。对于数值格式的数据，数字从小至大排序为升序，从大至小排序为降序；对于文本格式的数据，拼音或英文首字母A至Z排序为升序，Z至A排序为降序。

在Excel中有3处可以找到自动排序：

①“筛选”下拉菜单中“升序/降序”。

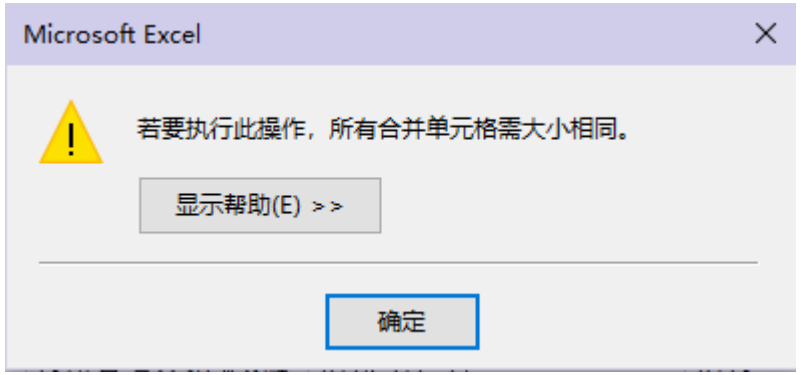


②“开始”选项卡中的“编辑”面板。



③“数据”选项卡中的“排序和筛选”面板。

注意：使用排序功能时若有合并的单元格，将无法执行排序操作，需先将合并的单元格还原至未合并状态。



Excel计算功能

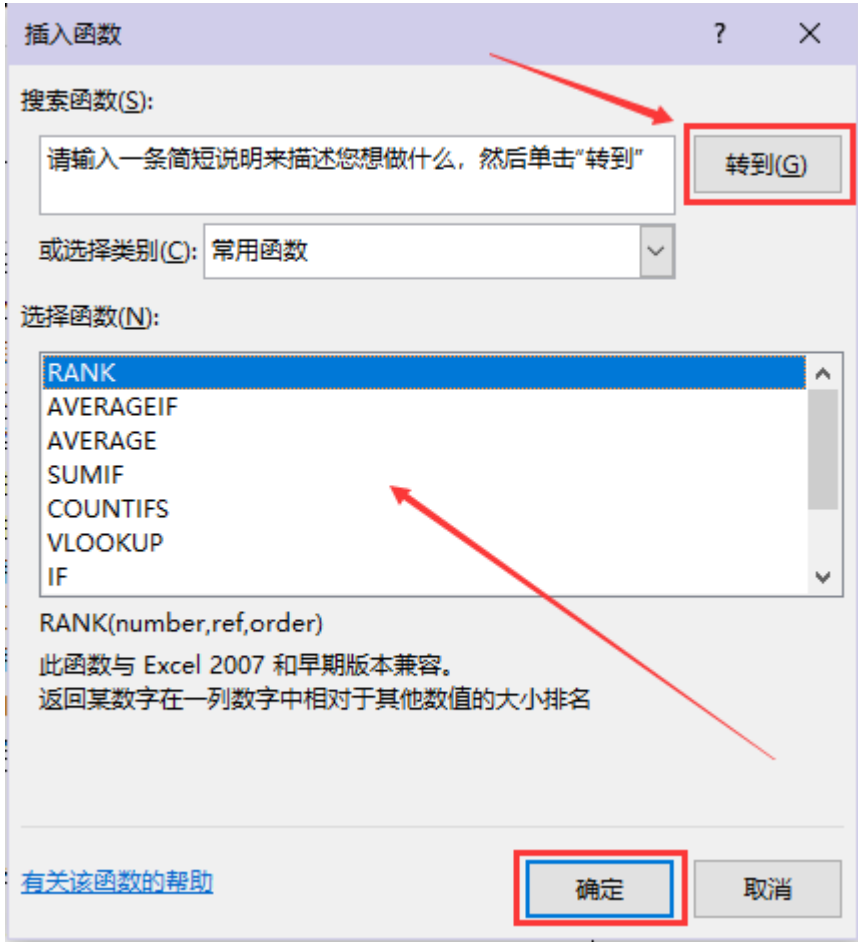
在单元格中以“=”开头，配合数字、单元格引用、运算符号组成一个公式，便可以实现简单的数学运算。

	等号	加号	减号	除号	乘号	平方
运算符号	=	+	−	/	*	^
公式格式示范	=B3/D3、=A1*A2+A3					

Excel函数功能

Excel提供了300多个内置函数。正确使用函数可以实现大量数据的快速计算、统计和匹配等工作。

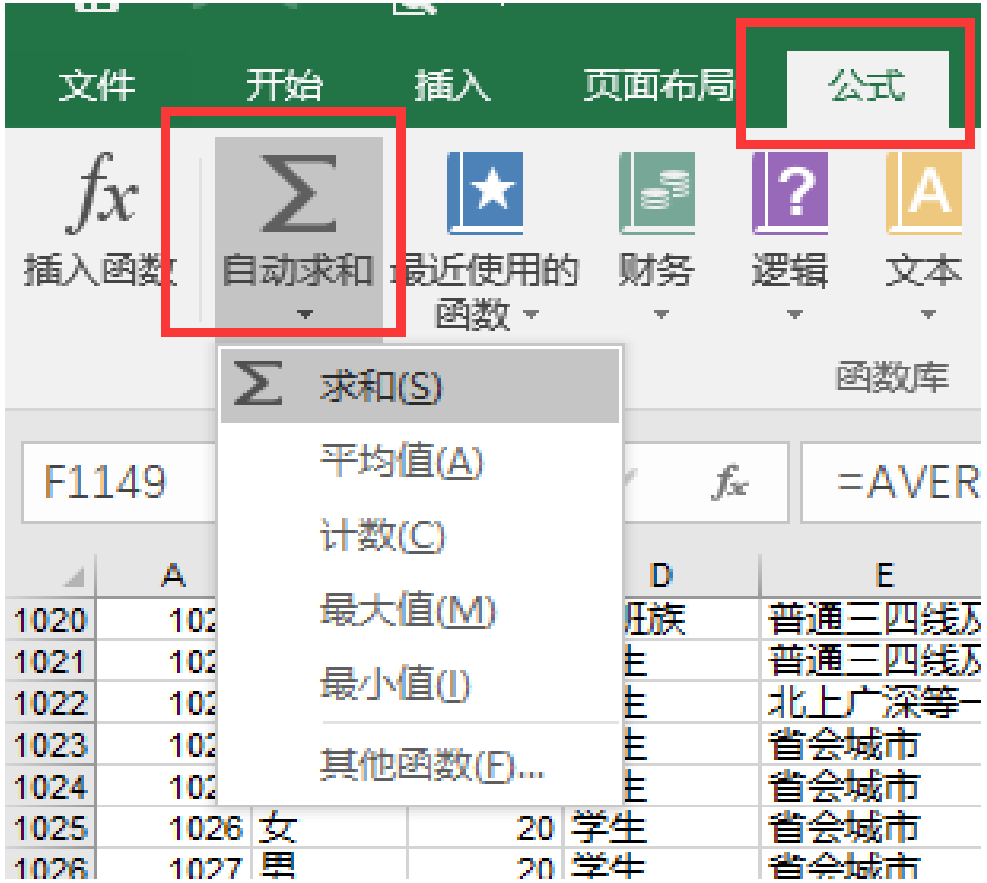
想要使用这些函数，除了直接在单元格内输入，也可以在“公式”选项卡中打开“插入函数”窗口选择想使用的函数。或者在“搜索函数”中简单描述需求来搜索。



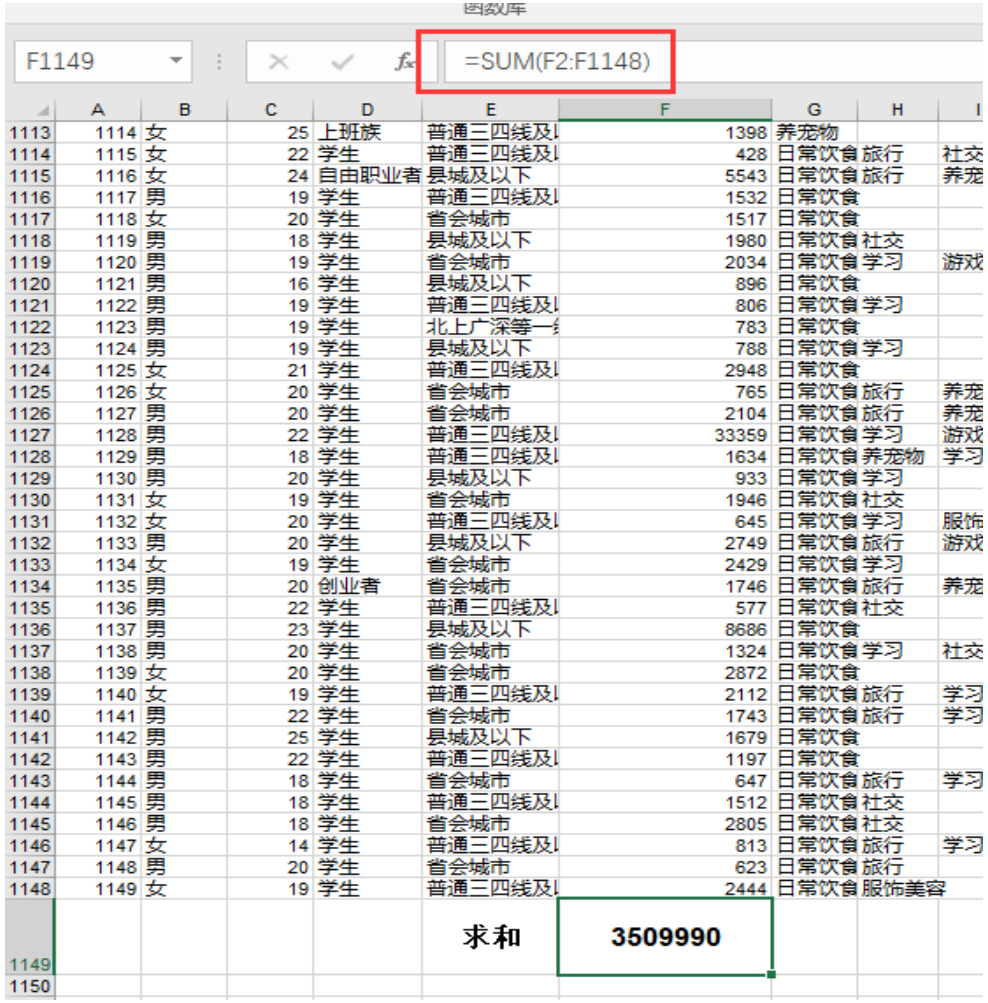
1.求和函数

使用求和函数SUM可以快速计算所选范围中所有数值的总和，进行一些分析计算，如百分比等。SUM函数通用格式为“=SUM(number1,number2,...)”。也可以通过“公式”选项卡下的“求和”选项插入快速求和。

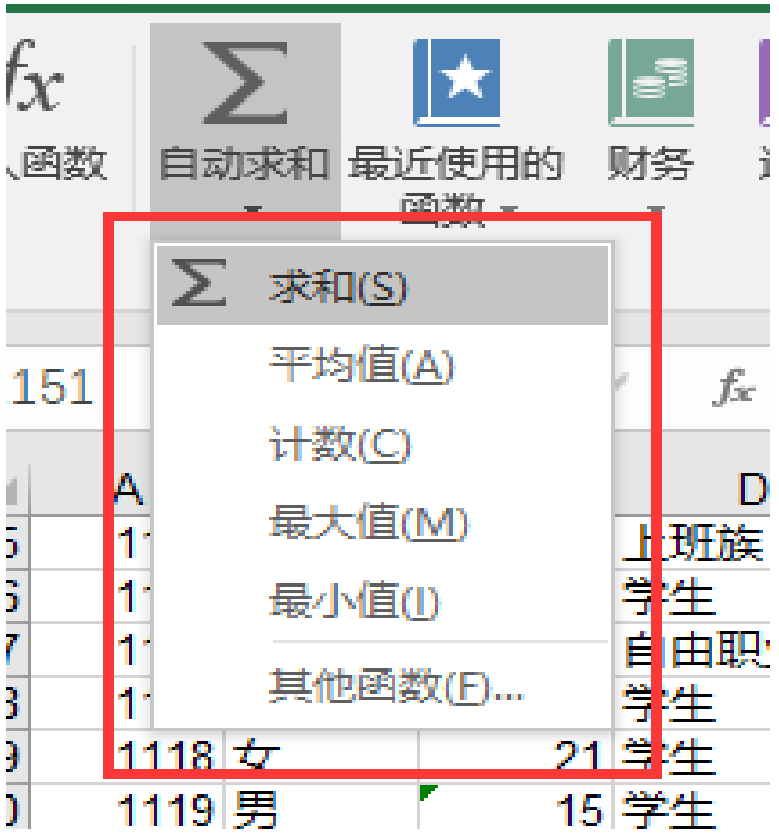
第一步，选中想要插入SUM函数的单元格，在“公式”选项卡下选择“自动求和”。



第二步，Excel会自动判定求和范围，若不正确需重新修改选区，按键盘Enter键完成操作。



在“自动求和”中，Excel还提供了求平均值、最大值和计数等快捷方式，用法与求和函数相似。



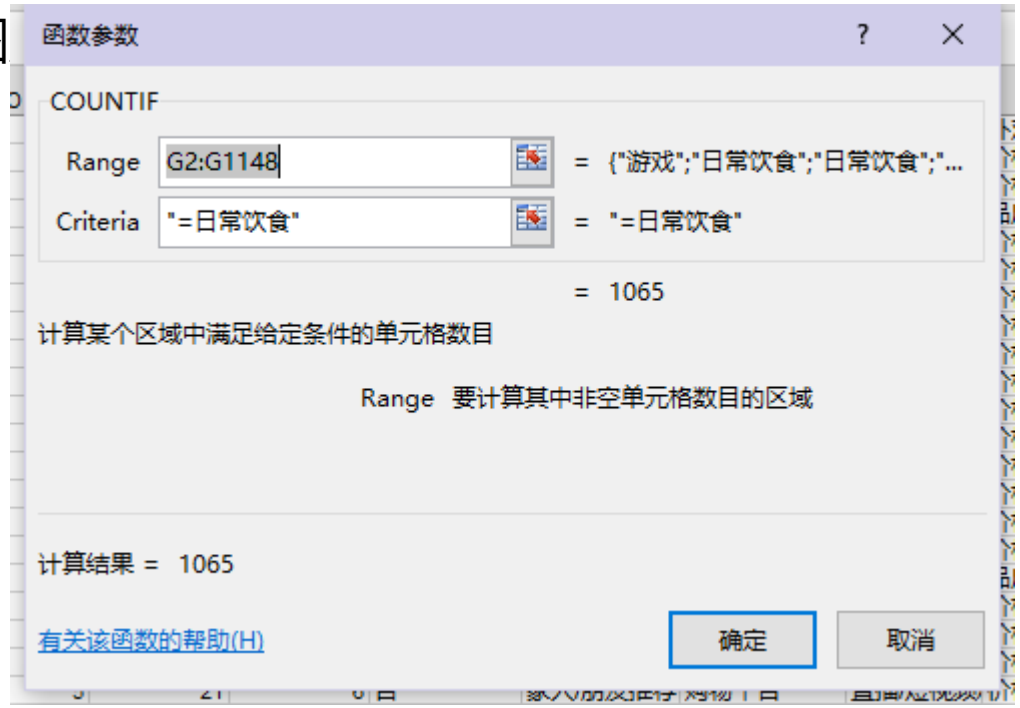
2.条件计数函数

当面对大量的数据时，通常会查看共有几行数据来计算数据的基数，但这无法排除空单元格。为了准确掌握数据的基数，可以使用COUNTIF函数在计数时设置指定条件，统计含有符合设定条件数据的单元格数量。

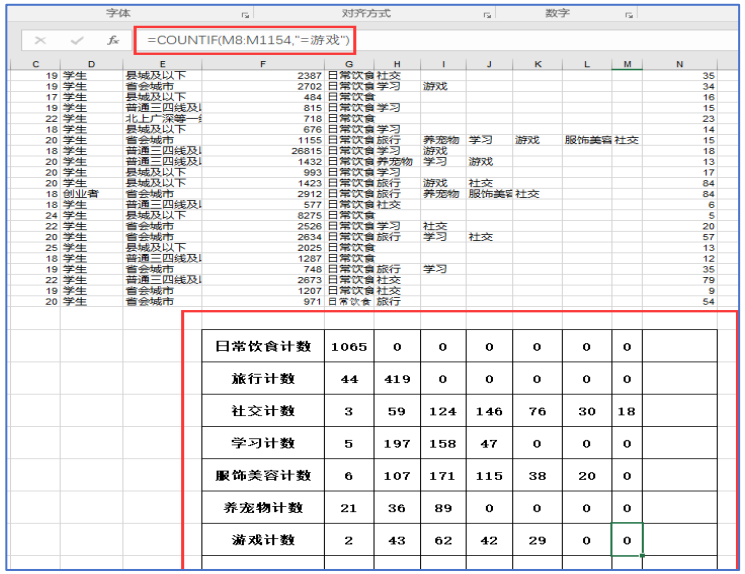
以统计每月支出各类型占比为例：

第一步，在单元格中插入COUNTIF函数。

第二步，设置参数如图



第三步，点击“确定”统计，填充功能向右填充。将其他类别依次统计出来得到如下表格。



日常饮食计数	1065	0	0	0	0	0	0
旅行计数	44	419	0	0	0	0	0
社交计数	3	59	124	146	76	30	18
学习计数	5	197	158	47	0	0	0
服饰美容计数	6	107	171	115	38	20	0
养宠物计数	21	36	89	0	0	0	0
游戏计数	2	43	62	42	29	0	0

第四步，使用SUM函数求出各种类之和。

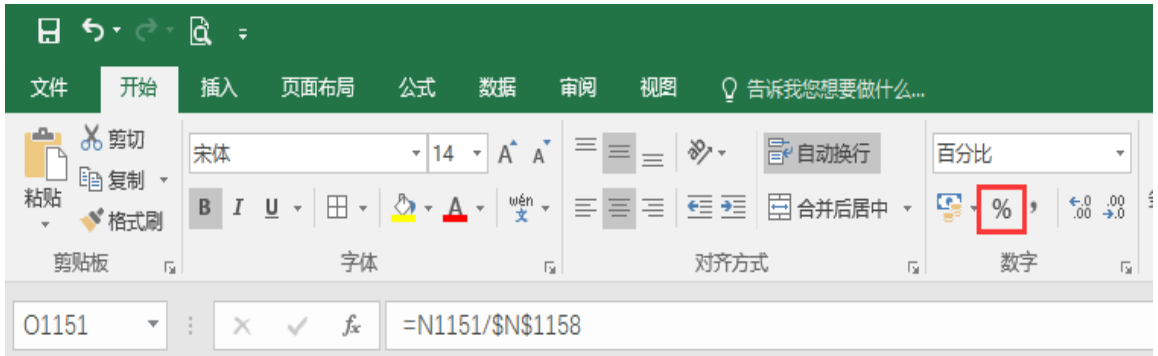
							合计
日常饮食计数	1065	0	0	0	0	0	0	1065
旅行计数	44	419	0	0	0	0	0	463
社交计数	3	59	124	146	76	30	18	456
学习计数	5	197	158	47	0	0	0	407
服饰美容计数	6	107	171	115	38	20	0	457
养宠物计数	21	36	89	0	0	0	0	146
游戏计数	2	43	62	42	29	0	0	178
合计								3172

第五步，计算每项类别所占百分比。百分比计算方式为“单项总数/样本总量”，以日常饮食为例，在末尾单元格输入公示“=N1150/N1157”。使用填充完成所有计算。

注意：如需使用填充工具，需在引用样本总量单元格时添加**绝对引用**。

								合计	所占比例
日常饮食计数	1065	0	0	0	0	0	0	1065	0.33575
旅行计数	44	419	0	0	0	0	0	463	0.14596
社交计数	3	59	124	146	76	30	18	456	0.14376
学习计数	5	197	158	47	0	0	0	407	0.12831
服饰美容计数	6	107	171	115	38	20	0	457	0.14407
养宠物计数	21	36	89	0	0	0	0	146	0.04603
游戏计数	2	43	62	42	29	0	0	178	0.05612
合计								3172	1

第六步，将所占比例设置为百分比格式。选中所有占比结果，在“开始”选项卡中“数字”模块设置点击“百分比样式”选项。



3.VLOOKUP函数

VLOOKUP函数是Excel中的一个纵向查找函数，可以用来核对数据。使用VLOOKUP函数可以实现多表格数据间的查找与匹配。

VLOOKUP函数的通用格式为：

“=VLOOKUP(lookup_value,table_array,col_index_num,range_lookup)”

4个参数分别表示：

lookup_value：待搜索的值。可以直接引用需要查找的单元格。

table_array：搜索区域。这里的搜索区域必须包含需返回目标值所在列。

col_index_num：搜索区域中需要返回的值所在的列数。用数字表示。

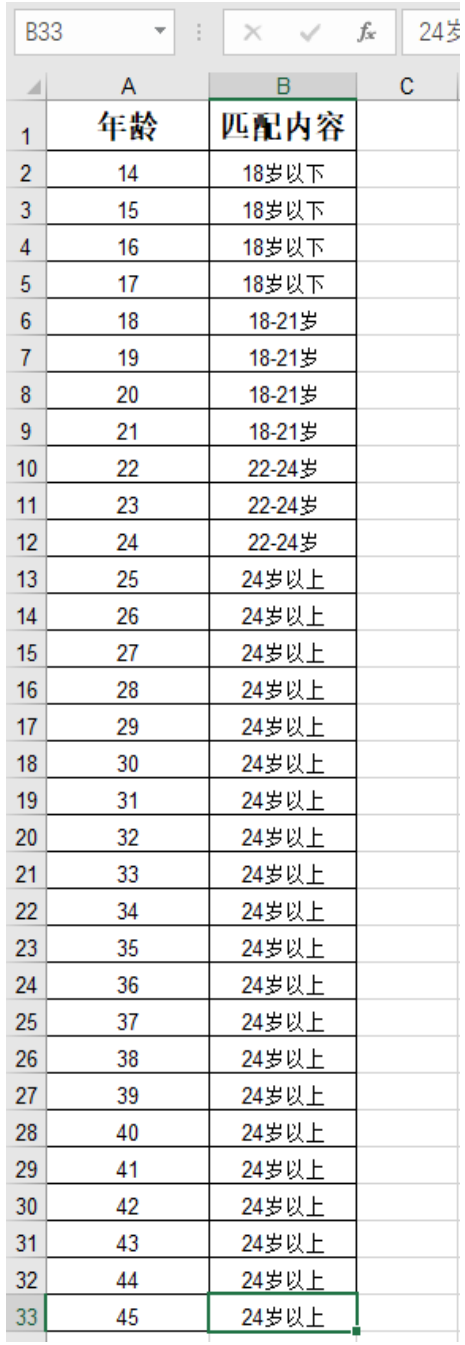
range_lookup：返回值的类型。输入非0值为TRUE，代表近似匹配；输入0或忽略为FALSE，代表精确匹配。

以年龄群分析为例，用户可以将数据表内的年龄数据使用VLOOKUP函数替换为年龄群表述，操作步骤如下:

第一步，新建一个工作表创建数据对照表,一列为待搜索的值，一列为需要返回的内容。

1. 匹配“年龄”则将年龄数据作为待搜索的值，通过筛选查看原始数据发现，调查样本的年龄区间在14至24岁，以此来创建数据对照表第一列；

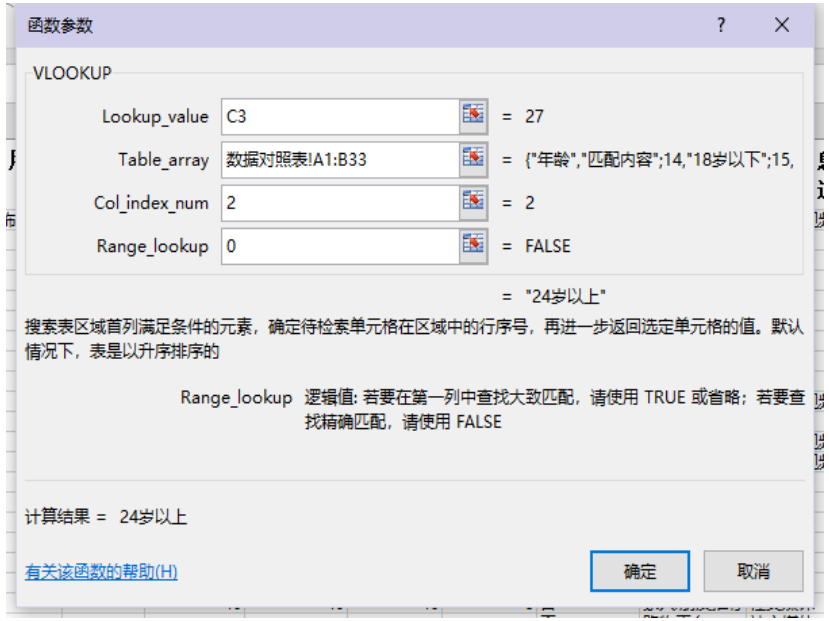
2. 根据设计的年龄群，对应“年龄”录入数据对照表第二列，需要匹配的内容，得到下表。



B33 | 24岁

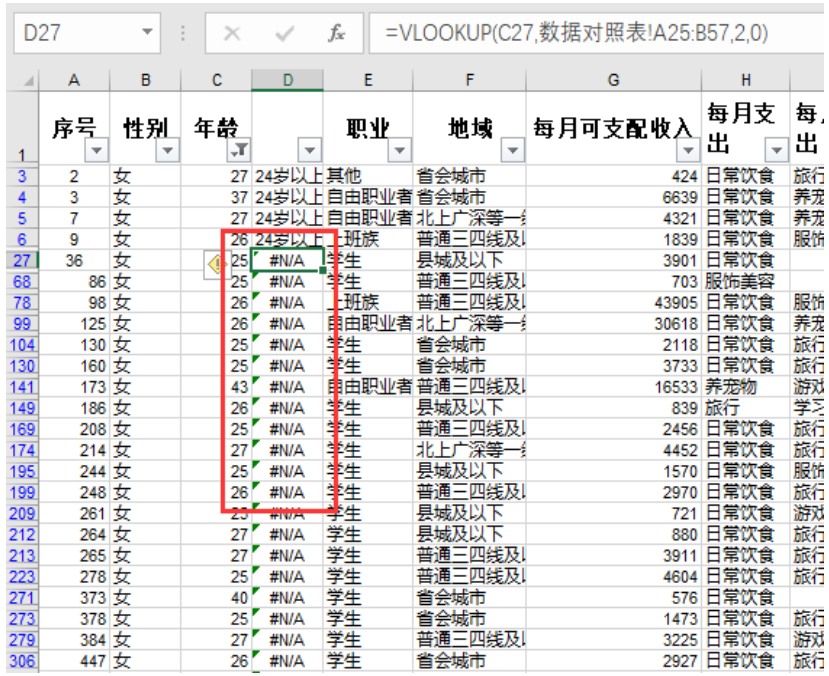
	A	B	C
1	年龄	匹配内容	
2	14	18岁以下	
3	15	18岁以下	
4	16	18岁以下	
5	17	18岁以下	
6	18	18-21岁	
7	19	18-21岁	
8	20	18-21岁	
9	21	18-21岁	
10	22	22-24岁	
11	23	22-24岁	
12	24	22-24岁	
13	25	24岁以上	
14	26	24岁以上	
15	27	24岁以上	
16	28	24岁以上	
17	29	24岁以上	
18	30	24岁以上	
19	31	24岁以上	
20	32	24岁以上	
21	33	24岁以上	
22	34	24岁以上	
23	35	24岁以上	
24	36	24岁以上	
25	37	24岁以上	
26	38	24岁以上	
27	39	24岁以上	
28	40	24岁以上	
29	41	24岁以上	
30	42	24岁以上	
31	43	24岁以上	
32	44	24岁以上	
33	45	24岁以上	

第二步，在数据表“年龄”列后新增一列，在单元格插入VLOOKUP函数。参数设置如图所示。搜索区域可跨工作表选择。

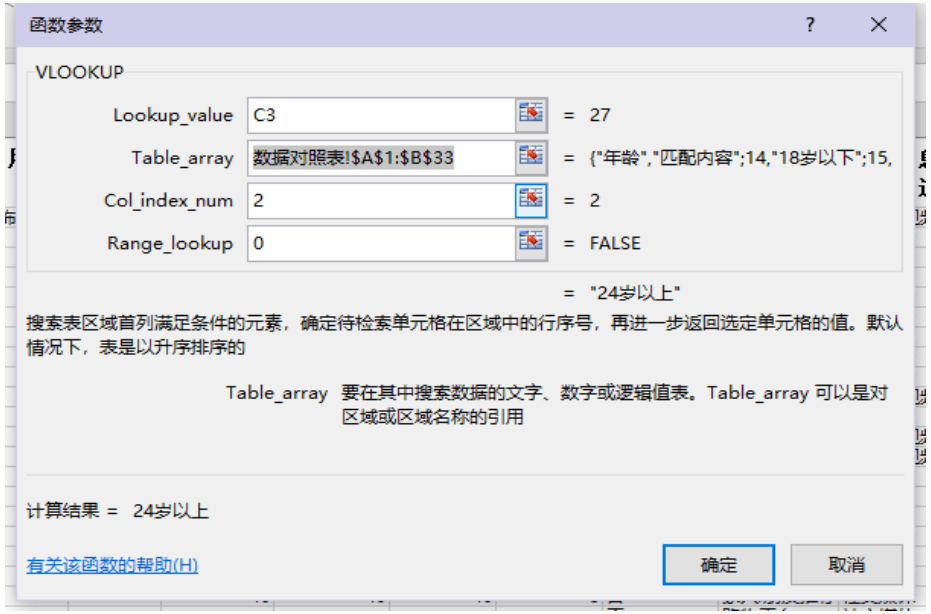


第三步，向下填充后发现出现错误提示函数不可用，原因是使用自动填充时table_array参数内的单元格应用了序列填充，导致搜索区域发生位移。

解决办法：为了使查找范围不随填充变化，需要在查找范围的行、列标记之前添加**绝对引用符号“$”**。



点击编辑栏左侧  插入函数打开函数参数窗口，选中table_array参数，点击键盘F4键添加绝对引用后重新向下填充函数。参数设置如图所示。



第四步，快速填充完成数据匹配后，替换原“年龄”列数据，替换结果如图所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	序号	性别	年龄		职业	地域	每月可支配收入	每月支出	每月支出	每月支出	每月支出	每月支出
2	1	女	19	18-21岁	学生	省会城市	648	游戏	服饰美容	社交		
3	2	女	27	24岁以上	其他	省会城市	424	日常饮食	旅行	养宠物	学习	游戏
4	3	女	37	24岁以上	自由职业者	省会城市	6639	日常饮食	养宠物	游戏	服饰美容	
5	7	女	27	24岁以上	自由职业者	北上广深等一	4321	日常饮食	养宠物	服饰美容	社交	
6	9	女	26	24岁以上	上班族	普通三四线及	1839	日常饮食	服饰美容			
7	10	女	21	18-21岁	学生	省会城市	641	日常饮食	服饰美容			
8	11	女	22	22-24岁	学生	北上广深等一	899	日常饮食	旅行	学习	服饰美容	
9	12	女	15	18岁以下	学生	县城及以下	461	日常饮食	旅行	学习	服饰美容	
10	13	女	19	18-21岁	学生	普通三四线及	507	日常饮食				
11	14	女	19	18-21岁	学生	省会城市	1521	日常饮食	旅行	养宠物	游戏	服饰美容
12	16	女	19	18-21岁	学生	普通三四线及	561	日常饮食	旅行	学习	服饰美容	社交
13	17	女	18	18-21岁	学生	省会城市	1120	日常饮食	服饰美容	社交		
14	18	女	20	18-21岁	学生	省会城市	938	日常饮食	社交			
15	21	女	21	18-21岁	学生	普通三四线及	2339	日常饮食				
16	22	女	19	18-21岁	学生	省会城市	1615	日常饮食	旅行	养宠物	游戏	服饰美容
17	23	女	19	18-21岁	学生	省会城市	1797	日常饮食	旅行	学习	服饰美容	
18	24	女	21	18-21岁	学生	省会城市	2467	日常饮食	旅行	游戏	服饰美容	
19	25	女	19	18-21岁	学生	普通三四线及	4982	日常饮食				
20	26	女	22	22-24岁	学生	省会城市	2096	日常饮食	服饰美容	社交		
21	27	女	18	18-21岁	学生	省会城市	627	日常饮食	旅行	服饰美容	社交	
22	28	女	20	18-21岁	学生	普通三四线及	410	日常饮食	学习	游戏		
23	30	女	21	18-21岁	学生	省会城市	1503	日常饮食	旅行	学习	游戏	服饰美容
24	31	女	20	18-21岁	学生	普通三四线及	2579	日常饮食	旅行	养宠物		
25	33	女	14	18岁以下	学生	省会城市	2610	日常饮食	旅行	服饰美容	社交	
26	35	女	20	18-21岁	学生	省会城市	4886	日常饮食	学习	服饰美容	社交	
27	36	女	25	24岁以上	学生	县城及以下	3901	日常饮食				

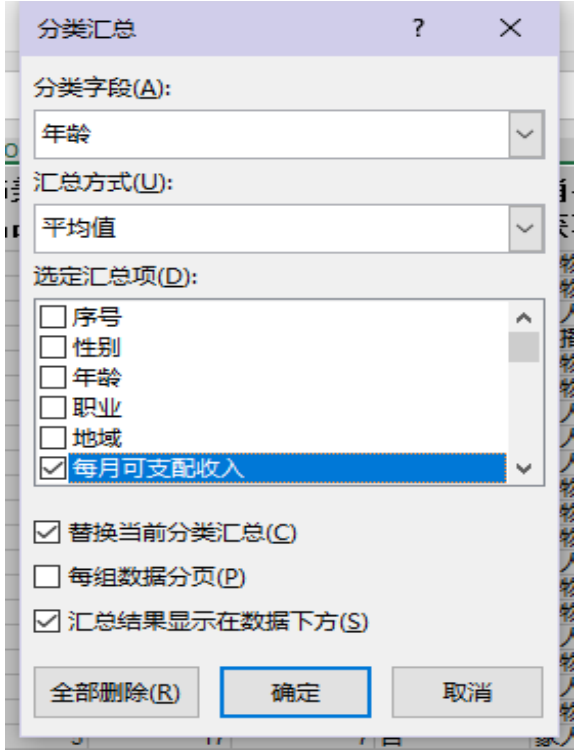
4.Excel的分类汇总

分类汇总可以实现按某个字段进行分类，并将分类后的数据进行求和、计数、求平均值等形式进行汇总。使用分类汇总可以快速查看各子集的统计情况。

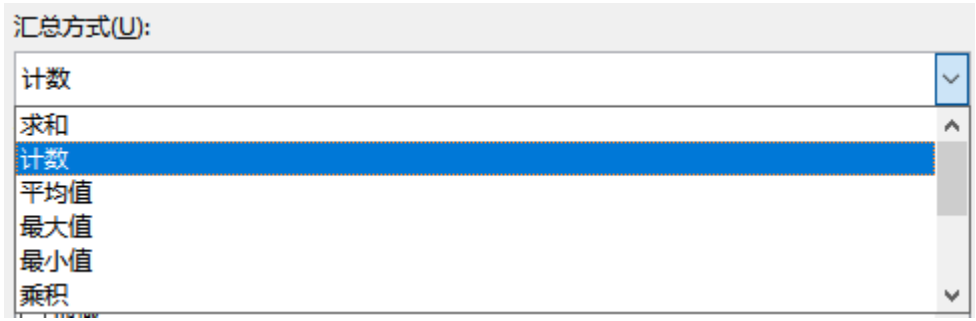
例如想进一步对年龄群的消费水平进行分析，掌握调查数据中每个年龄每月可支配收入的平均值。

第一步，启用分类汇总前请先取消字段名的合并单元格。激活数据表中任意单元格，在“数据”选项卡“分级选项”面板中打开“分类汇总”。

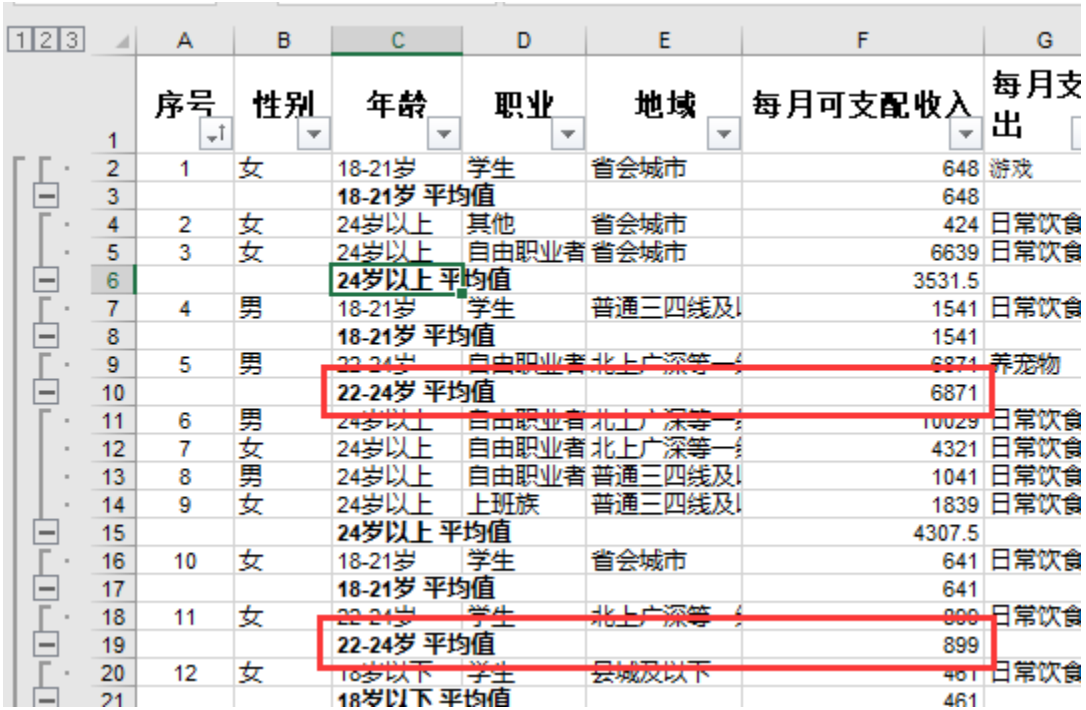
第二步，如图4-83所示设置分类字段、汇总方式、选定汇总项后点击确定。



除计数外，Excel还提供求和、求平均值、最大值等多种汇总方式，可以节省大量计算工作。

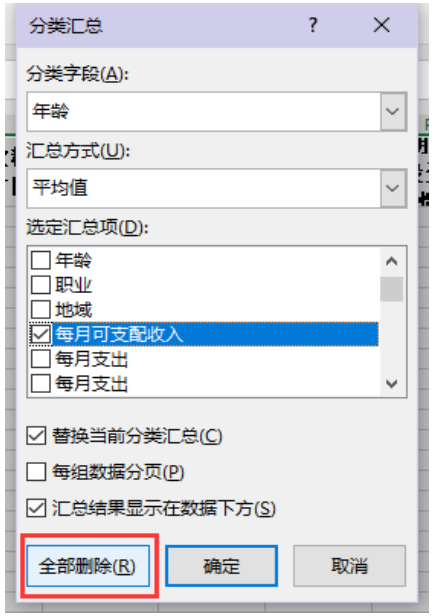


汇总结果如图所示，可以看出并未达到完全分类预期。原因是分类汇总是按照当前数据的排列依次进行分类，所以在分类汇总前需先对分类字段进行排序，将类别相同的记录排列在一起。



	A	B	C	D	E	F	G
1	序号	性别	年龄	职业	地域	每月可支配收入	每月支出
2	1	女	18-21岁	学生	省会城市	648	游戏
3			**18-21岁 平均值**			648	
4	2	女	24岁以上	其他	省会城市	424	日常饮食
5	3	女	24岁以上	自由职业者	省会城市	6639	日常饮食
6			**24岁以上 平均值**			3531.5	
7	4	男	18-21岁	学生	普通三四线及以	1541	日常饮食
8			**18-21岁 平均值**			1541	
9	5	男	22-24岁	自由职业者	北上广深等一线	6871	养宠物
10			**22-24岁 平均值**			6871	
11	6	男	24岁以上	自由职业者	北上广深等一线	10029	日常饮食
12	7	女	24岁以上	自由职业者	北上广深等一线	4321	日常饮食
13	8	男	24岁以上	自由职业者	普通三四线及以	1041	日常饮食
14	9	女	24岁以上	上班族	普通三四线及以	1839	日常饮食
15			**24岁以上 平均值**			4307.5	
16	10	女	18-21岁	学生	省会城市	641	日常饮食
17			**18-21岁 平均值**			641	
18	11	女	22-24岁	学生	北上广深等一线	899	日常饮食
19			**22-24岁 平均值**			899	
20	12	女	18岁以下	学生	县城及以下	461	日常饮食
21			**18岁以下 平均值**			461	

解决办法：打开“分类汇总”窗口删除当前分类汇总设置，对“年龄”列重新排序后重新设置分类汇总。



正确汇总结果如图所示

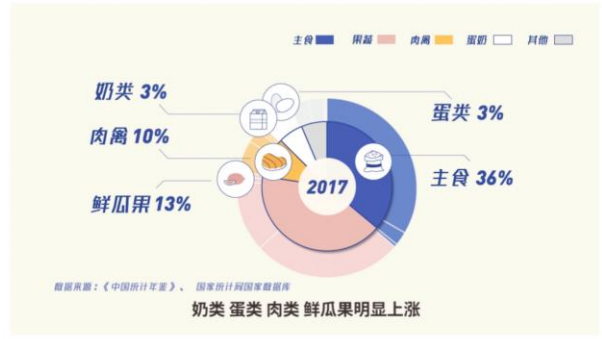
	A	B	C	D	E	F	G	H	I	J	
1	序号	性别	年龄	职业	地域	每月可支配收入	每月支出	每月支出	每月支出	每月支出	每出
791			18-21岁 平均值			2213.604563					
863			18岁以下 平均值			7851.71831					
864	5	男	22-24岁	自由职业者	北上广深等一线	6871	美宠物	游戏			
865	11	女	22-24岁	学生	北上广深等一线	899	日常饮食	旅行	学习	服饰美容	
866	26	女	22-24岁	学生	省会城市	2096	日常饮食	服饰美容	社交		
867	50	女	22-24岁	学生	省会城市	531	日常饮食				
868	54	女	22-24岁	学生	省会城市	531	日常饮食				
869	55	女	22-24岁	学生	省会城市	739	日常饮食				
870	57	女	22-24岁	学生	省会城市	502	日常饮食				
871	58	女	22-24岁	学生	省会城市	813	日常饮食				

点击左侧  可以将分组数据折叠，直接查看数据汇总情况。

	A	B	C	D	E	F	G
1	序号	性别	年龄	职业	地域	每月可支配收入	每月支出
791			18-21岁 平均值			2213.604563	
863			18岁以下 平均值			7851.71831	
1075			22-24岁 平均值			3392.976303	
1152			24岁以上 平均值			7491.710526	
1153			总计平均值			3129.288579	
1154							

思政融入：

用数据讲故事，首先需要在纷繁复杂的数据中寻找关键线索，通过深入分析解读数据，将其串联为一个富有吸引力和趣味性的故事，再利用数据可视化技术呈现出来。在“数说70年”系列产品的制作过程中，精准筛选出了近100套、约1000组有价值的数据。以数据为主线，以变化和对比的方式动态反映消费、饮食等领域的发展变化。比如，使用数据可视化工具展示老百姓饮食结构的变迁、主食的种类和占比变化、肉蛋菜果鱼的消费量增长变化等，让数据故事化、可视化。通过案例的展示，请各小组讨论：在分析数据时，应具备怎样的数字思维方式？



思政元素：数字素养、探究精神

谢谢观看