第4章
充分准备：表格数据清洗与加工处理

》》本章导读

数据清洗，是对原始数据的再审视。使用Excel的正确操作，发现并处理原始数据表中的重复值、缺省值、错误值，纠正数据的格式。

数据加工，是对原始数据的变形。充分利用Excel功能，对数据进行计算、转换、分类、重组。将理论付诸于实践，帮助分析者发现更有价值的数据形式。

》》带着以下问题进入学习

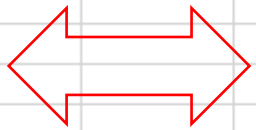
- ✓ （1）如果没有数据清洗，可能会出现什么后果？
- ✓ （2）只有数据清洗，没有数据加工可以吗？
- ✓ （3）数据加工符合了哪些数据分析的思想？
- ✓ （4）不同的数据加工方式，分别包含了什么样的思考流程？作这样的思考，是否加快了数据加工的学习效率？

4.1 数据加工处理的必要性

数据加工是指对收集到原始数据表进行进一步的处理，使原始数据更符合分析需求。之所以要进行这一步，是因为一份合乎分析质量的数据，需要具备准确性、完整性、一致性。而收集到的初始数据，很难具备这三个特征。

如果不事先进行数据处理，会有什么后果？

日期	销量（件）
7月25日	25
7月26日	51
7月27日	42
7月28日	52
7月29日	62
7月30日	42
7月31日	51
7月32日	42
8月1日	52
8月2日	51
8月3日	42
8月4日	-25
8月5日	26
8月6日	51
8月7日	42
8月8日	0
8月9日	25
平均值	631



日期	销量（件）
7月25日	25
7月26日	51
7月27日	42
7月28日	52
7月29日	62
7月30日	42
7月31日	51
8月1日	52
8月2日	51
8月3日	42
8月4日	25
8月5日	26
8月6日	51
8月7日	42
8月8日	23
8月9日	25
平均值	662

高手自测10

对比下面两份数据，分别是原始数据和处理后的数据，有什么区别？

加工处理前

月　份	消费者信心指数			消费者满意指数			消费者预期指数		
	指数值	同比增长	环比增长	指数值	同比增长	环比增长	指数值	同比增长	环比增长
2017年11月份	121.3	11.69%	-2.10%	116.3	11.51%	-1.77%	124.6	11.75%	-2.35%
2017年10月份	123.9	15.58%	4.47%	118.4	15.40%	4.13%	127.6	15.79%	4.68%
2017年09月份	118.6	0.13	3.40%	113.7	13.47%	3.18%	121.9		3.66%
2017年08月份	114.7	8.62%	0.09%	110.2	8.46%	-0.18%	117.6	8.59%	0.17%
2017年9月	114.60	7.30%	1.15%	110.4	7.92%	1.75%	117.4	6.92%	0.86%
2017年06月份	113.3	10.11%	1.16%		9.49%	0.56%	116.4	10.33%	1.48%
May-17	112	12.22%	-1.23%	107.9	13.34%	-0.83%	114.7	11.58%	-1.46%
2017年04月份	11340%	0.12	2.16%	108.8	13.93%	2.45%			
2017年03月份	111	11.00%	-1.42%	106.2	11.91%				
2017年02月份	112.6	7.85%	3.11%	107.3	6.55%	2.78%			
2017年1月	109.2	5.00%	0.74%	104.4	4.09%	0.68%			
2016年12月份	10840%		-0.18%	103.7	3.08%	-0.58%			
2016年11月份	108.6	4.32%	1.31%	104.3	4.09%	1.66%			

加工处理后

月　份	消费者信心指数			消费者满意指数			消费者预期指数		
	指数值	同比增长	环比增长	指数值	同比增长	环比增长	指数值	同比增长	环比增长
2017年11月份	121.3	11.69%	-2.10%	116.3	11.51%	-1.77%	124.6	11.75%	-2.35%
2017年10月份	123.9	15.58%	4.47%	118.4	15.40%	4.13%	127.6	15.79%	4.68%
2017年09月份	118.6	13.38%	3.40%	113.7	13.47%	3.18%	121.9	13.29%	3.66%
2017年08月份	114.7	8.62%	0.09%	110.2	8.46%	-0.18%	117.6	8.59%	0.17%
2017年07月份	114.6	7.30%	1.15%	110.4	7.92%	1.75%	117.4	6.92%	0.86%
2017年06月份	113.3	10.11%	1.16%	108.5	9.49%	0.56%	116.4	10.33%	1.48%
2017年05月份	112	12.22%	-1.23%	107.9	13.34%	-0.83%	114.7	11.58%	-1.46%
2017年04月份	113.4	12.28%	2.16%	108.8	13.93%	2.45%	116.4	11.17%	1.93%
2017年03月份	111	11.00%	-1.42%	106.2	11.91%	-1.03%	114.2	10.44%	-1.72%
2017年02月份	112.6	7.85%	3.11%	107.3	6.55%	2.78%	116.2	8.70%	3.38%
2017年01月份	109.2	5.00%	0.74%	104.4	4.09%	0.68%	112.4	5.44%	0.81%
2016年12月份	108.4	4.53%	-0.18%	103.7	3.08%	-0.58%	111.5	5.39%	0.00%
2016年11月份	108.6	4.32%	1.31%	104.3	4.09%	1.66%	111.5	4.60%	1.18%

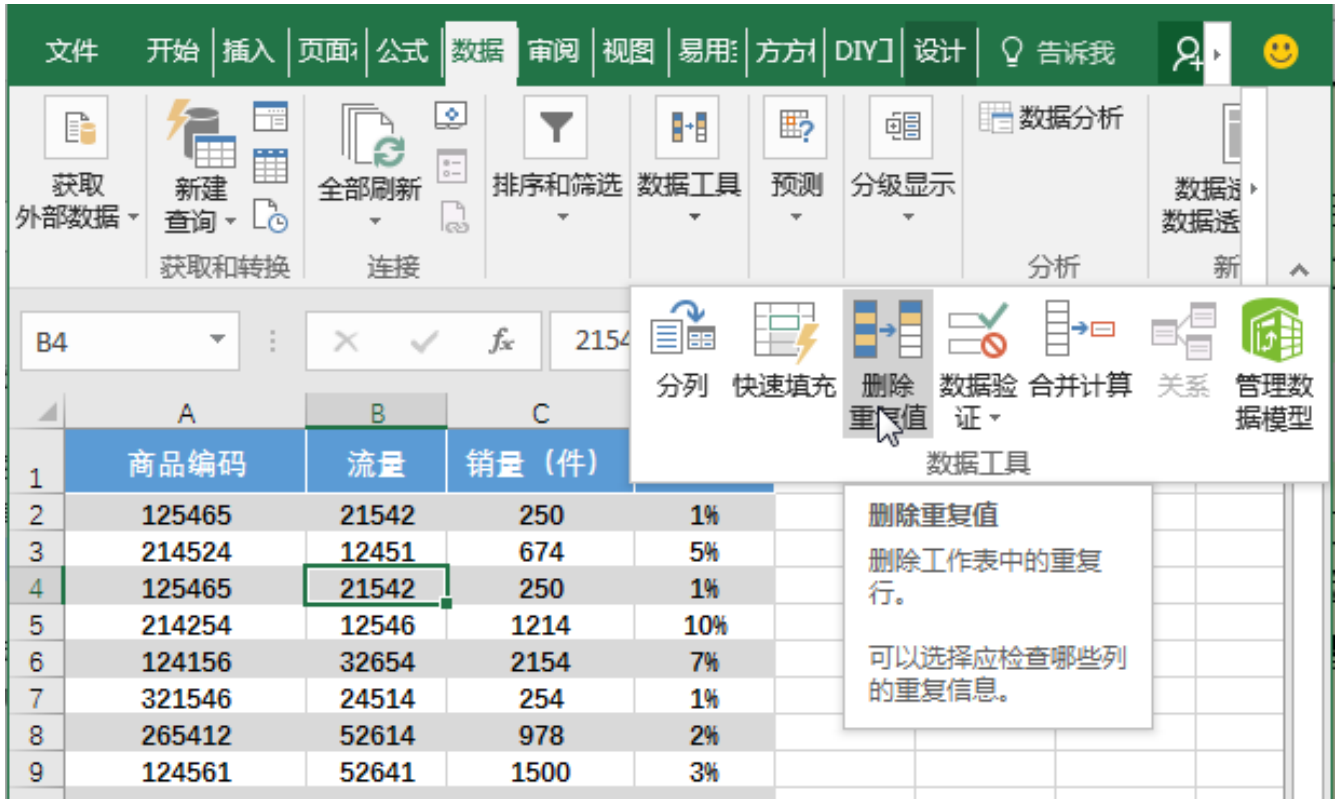
4.2 数据处理第一步：数据清洗

数据处理的第一步是数据清洗，目的是将多余的、错误的数据清洗出去，留下有价值的数据。数据清洗要借助Excel工具来进行，能保证清洗效果准确且高效。

4.2.1 3种方法数据去重

1.用删除重复项功能

删除重复项是Excel提供的数据去重功能，可以快速删除重复项。



4.2.1 3种方法数据去重

2.排序删除重复项

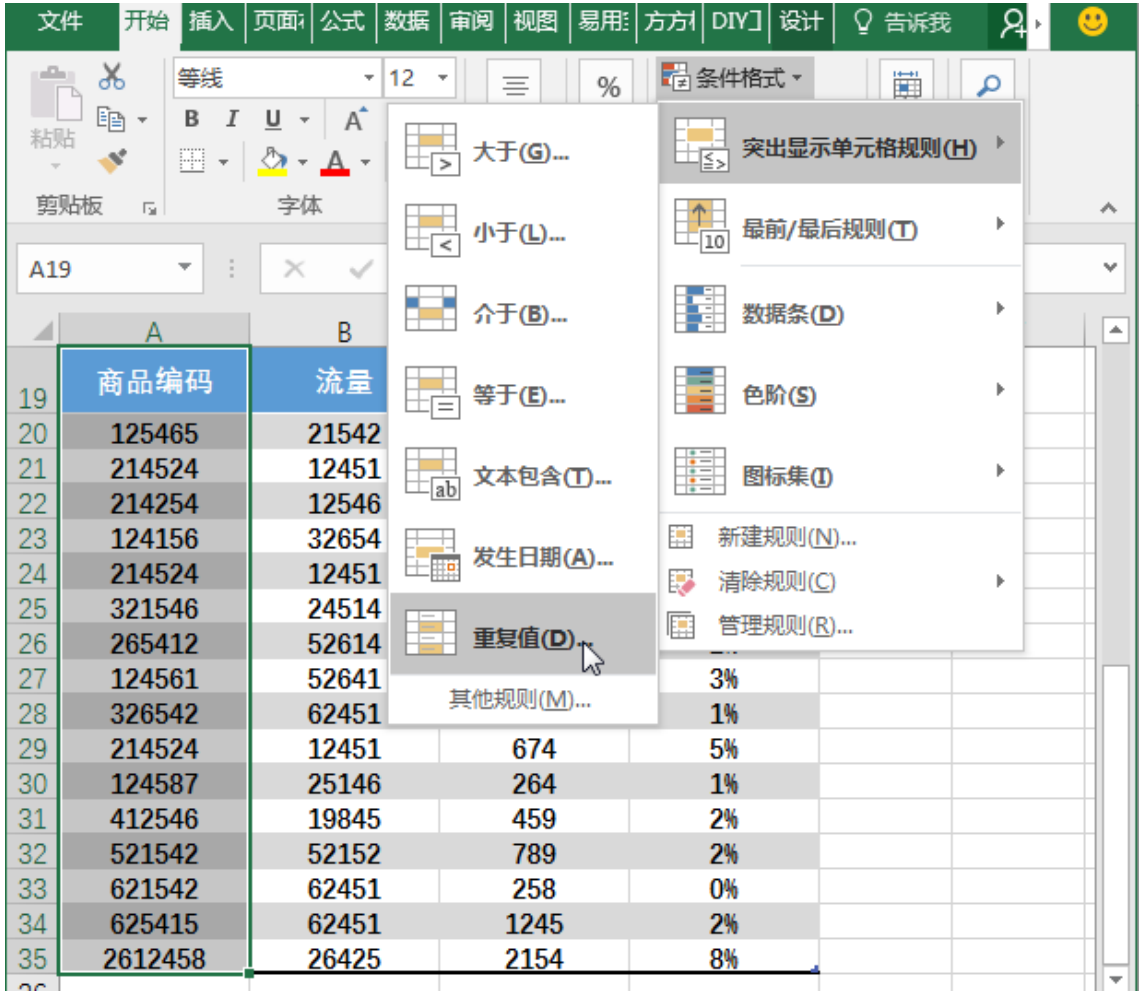
除了使用Excel工具的删除重复项功能删除重复数据，还可以通过排序的方法删除重复项。通过排序删除重复项，适合于那种需要人工判断是否真为无用重复项的数据。例如删除重复员工信息，员工姓名相同可能是巧合，也可能是重复数据，需要人工判断，不能让系统直接删除重复项。

排序删除重复项的原理是，将数据内容相同的信息排列到一起，可以一眼看出哪些数据是重复的，哪些不是重复的。

4.2.1 3种方法数据去重

3.条件格式删除重复项

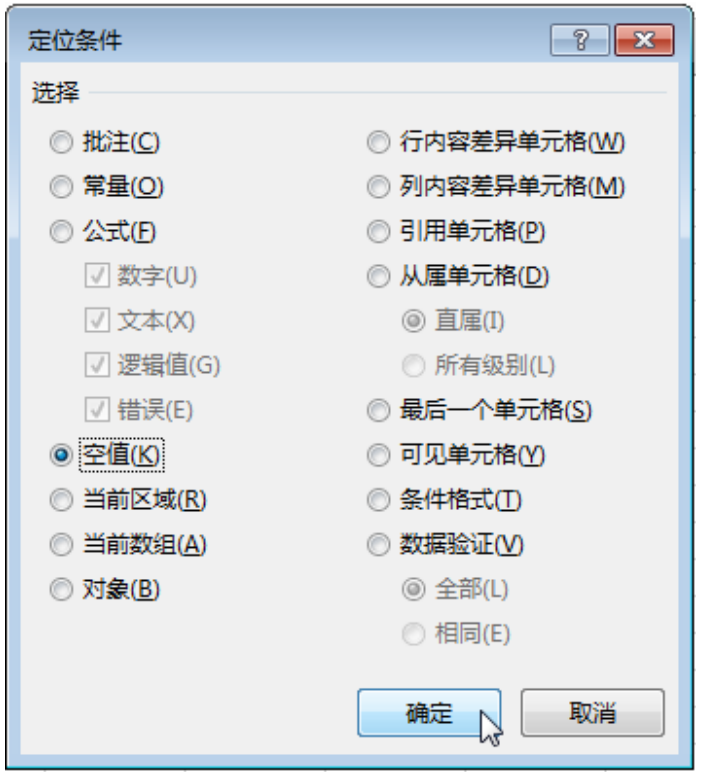
使用排序的方法删除重复项有一个问题，当数据是一串编码时，依然难以用肉眼看出重复的编码。用条件格式可以自动找出重复的数据，并手动删除。



4.2.2 3种方法处理缺省值

1.一步找出缺省值

在记录数据时，一旦数据量增加，难免出现数据缺省。大多数情况下，缺省的数据会以空白单元格显示。此时不仅需要将缺省数据检查出来，还要选择合理的处理方式，将缺省数据对数据分析的影响降到最小。



	A	B	C	D
1	商品编码	流量	销量（件）	转化率
2	125465	21542	250	1%
3	214254	12546	1214	10%
4	124156		2154	
5	321546	24514	254	1%
6	265412	52614	978	2%
7	124561	52641		0%
8	326542	62451	894	1%
9	124587	25146	264	1%
10	412546	19845	459	2%
11	521542		789	
12	621542	62451	258	0%
13	625415	62451	1245	2%
14	2612458	26425	2154	8%

4.2.2 3种方法处理缺省值

2. 3种方法处理缺省值

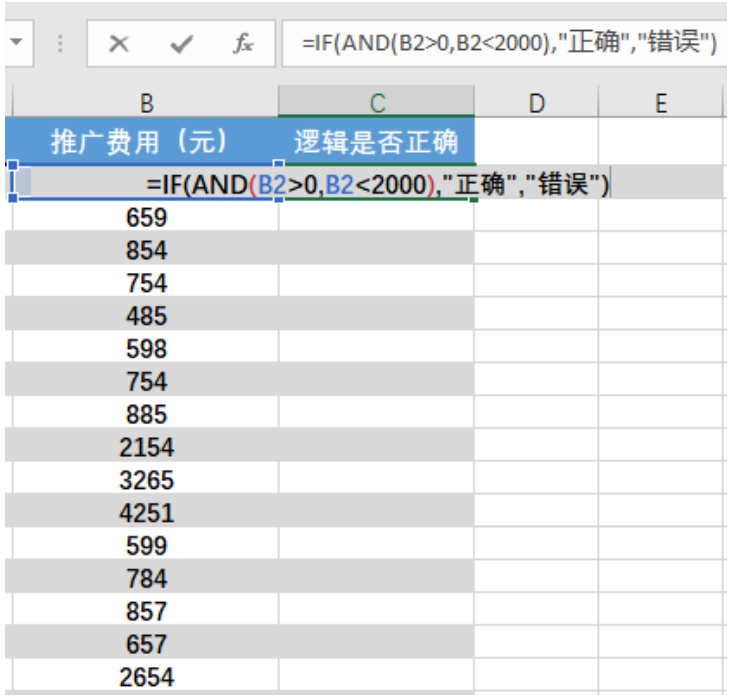
缺省数据和重复数据不一样，重复数据直接删除即可，但是缺省数据却不能草率删除，一般来说，有以下3种处理方法：

- 替换缺省值
- 删除缺省值
- 忽略缺省值

4.2.3 深度检查数据逻辑

1.用函数检查逻辑

函数是Excel的重要功能，它是一些预先定义好的公式，通过特定的参数结构进行计算。函数的功能十分强大，不仅可以对数据进行计算，还能根据不同的逻辑判断数据正误与否。

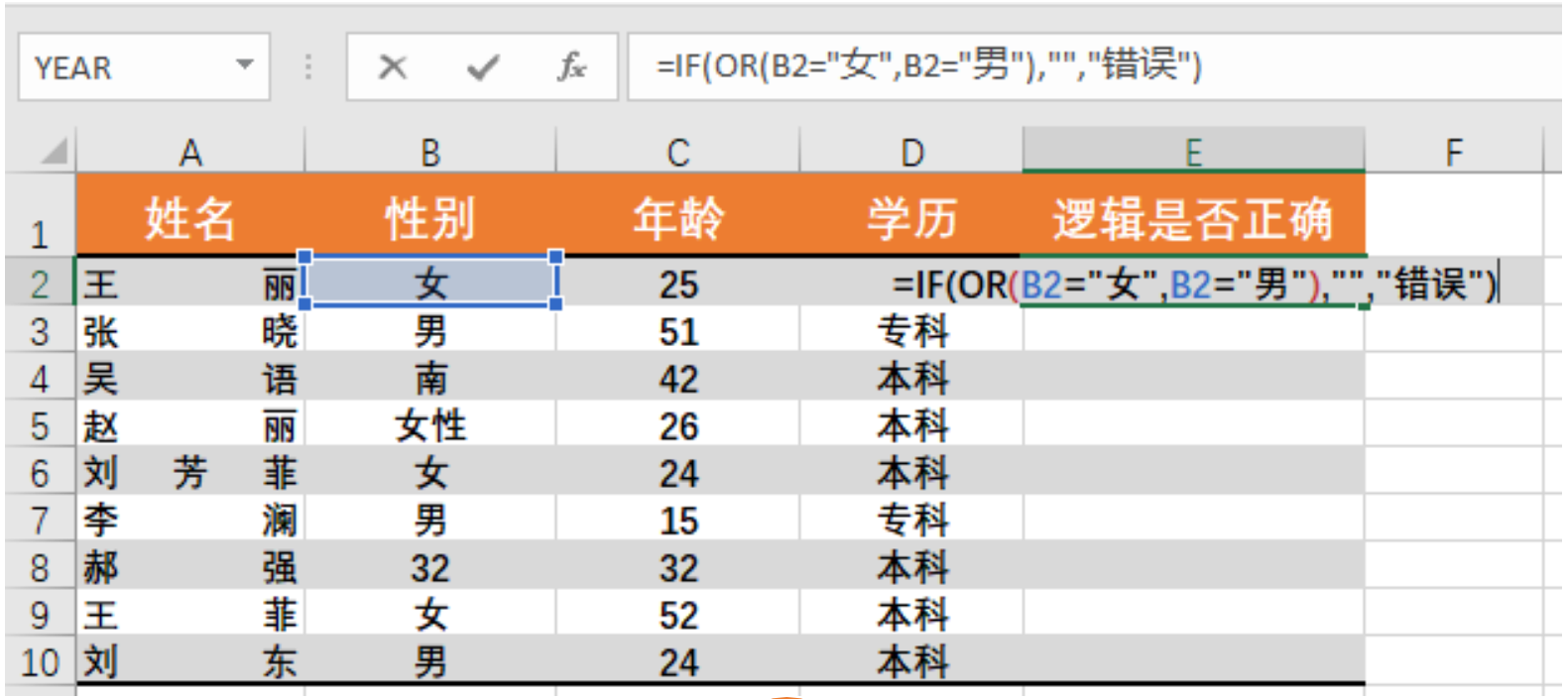


	A	B	C
1	日期	推广费用（元）	逻辑是否正确
14	3月13日	598	正确
15	3月14日	659	正确
16	3月15日	854	正确
17	3月16日	754	正确
18	3月17日	485	正确
19	3月18日	754	正确
20	3月19日	485	正确
21	3月20日	598	正确
22	3月21日	754	正确
23	3月22日	885	正确
24	3月23日	2154	错误
25	3月24日	3265	错误
26	3月25日	4251	错误
27	3月26日	599	正确
28	3月27日	784	正确
29	3月28日	857	正确
30	3月29日	657	正确
31	3月30日	2654	错误
32	3月31日	2546	错误

4.2.3 深度检查数据逻辑

1.用函数检查逻辑

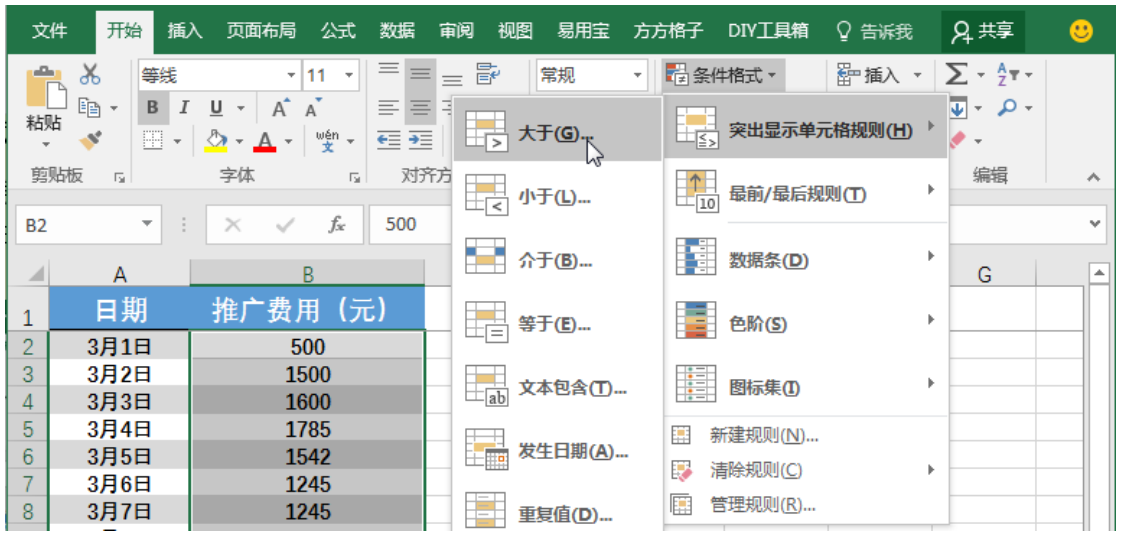
使用IF函数不仅可以判断数值是否符合特定的范围要求，还可以判断文字是否正确。例如一张企业员工信息表中，员工“性别”一栏的值只能是“男”或者是“女”，出现“本科”、“50”类似的信息都是错误的。当员工数量太多时，使用IF函数可以快速判断信息值是否正确。



4.2.3 深度检查数据逻辑

2.用条件格式检查逻辑

如果觉得对函数比较生疏，很难正确输入函数来判断数据的逻辑值正确与否。此时可以使用条件格式来检查逻辑值，减少了函数使用的困难。



	A	B
1	日期	推广费用（元）
2	3月1日	500
3	3月2日	1500
4	3月3日	1600
5	3月4日	1785
6	3月5日	1542
7	3月6日	1245
8	3月7日	1245

	A	B
1	日期	推广费用（元）
8	3月7日	1245
9	3月8日	1658
10	3月9日	1648
11	3月10日	1500
12	3月11日	1000
13	3月12日	1500
14	3月13日	598
15	3月14日	-500
16	3月15日	-615
17	3月16日	754
18	3月17日	485
19	3月18日	754
20	3月19日	485
21	3月20日	598
22	3月21日	754
23	3月22日	885
24	3月23日	2154
25	3月24日	3265
26	3月25日	4251

4.2.4 不要忘记检查格式

单元格数据有数值、文本、日期、货币等多种格式。不同类型的数据对应不同的格式，数据的格式有误，将会影响到后期透视表等功能的使用。因此，数据检查，千万不能忘记格式检查。

在检查数据格式时，尤其应该引起注意的是以下5个格式问题。

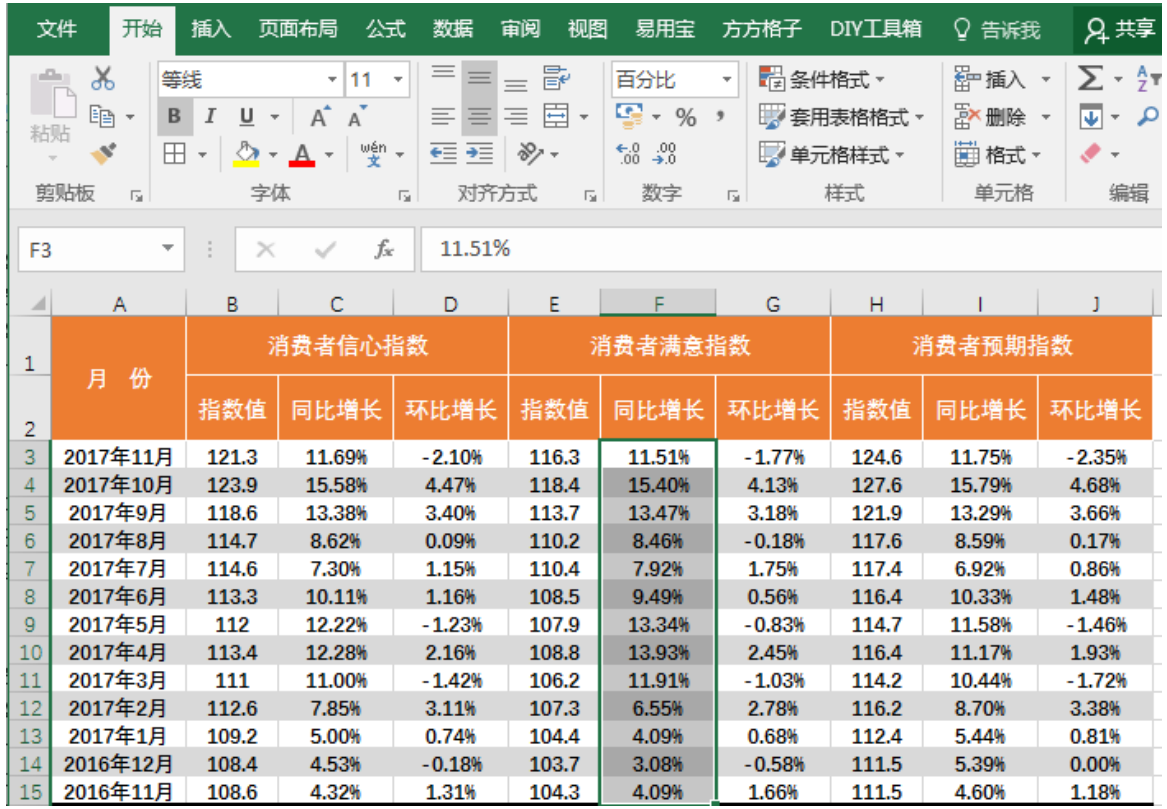
需要引起重视的5个格式问题

- 日期格式
- 时间格式
- 数值格式的小数位数
- 数值格式的千位分隔符
- 百分比格式

4.2.4 不要忘记检查格式

1.格式检查的方法

格式检查的方法比较简单，只需要选中数据列，在【开始】选项卡下【数字】组中的进行查看，看选中的数据列是否对应正确的格式。必要时，可以打开【设置单元格格式】对话框，调整数据格式。

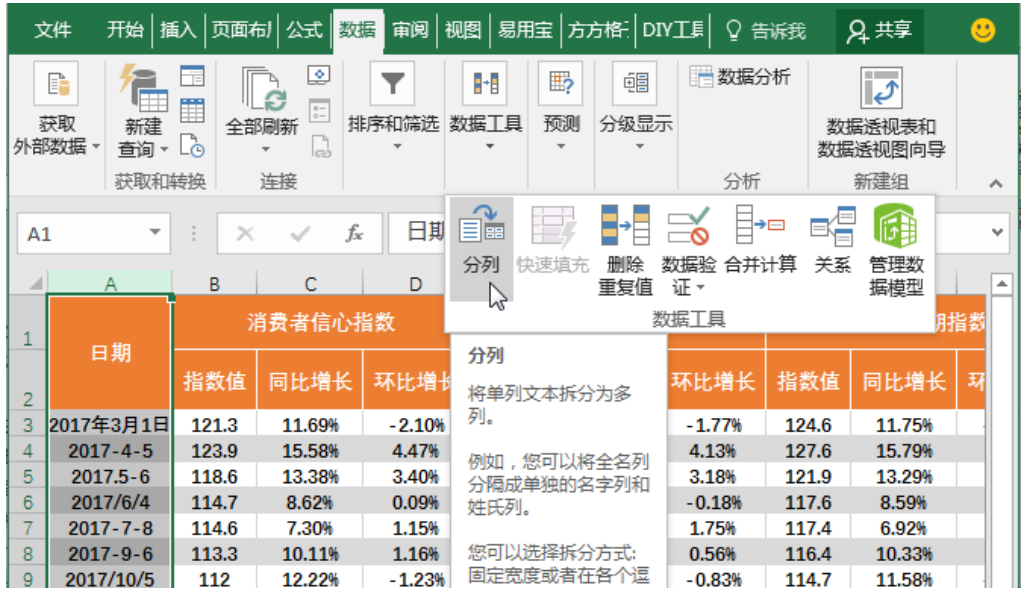


月 份	消费者信心指数			消费者满意指数			消费者预期指数		
	指数值	同比增长	环比增长	指数值	同比增长	环比增长	指数值	同比增长	环比增长
2017年11月	121.3	11.69%	-2.10%	116.3	11.51%	-1.77%	124.6	11.75%	-2.35%
2017年10月	123.9	15.58%	4.47%	118.4	15.40%	4.13%	127.6	15.79%	4.68%
2017年9月	118.6	13.38%	3.40%	113.7	13.47%	3.18%	121.9	13.29%	3.66%
2017年8月	114.7	8.62%	0.09%	110.2	8.46%	-0.18%	117.6	8.59%	0.17%
2017年7月	114.6	7.30%	1.15%	110.4	7.92%	1.75%	117.4	6.92%	0.86%
2017年6月	113.3	10.11%	1.16%	108.5	9.49%	0.56%	116.4	10.33%	1.48%
2017年5月	112	12.22%	-1.23%	107.9	13.34%	-0.83%	114.7	11.58%	-1.46%
2017年4月	113.4	12.28%	2.16%	108.8	13.93%	2.45%	116.4	11.17%	1.93%
2017年3月	111	11.00%	-1.42%	106.2	11.91%	-1.03%	114.2	10.44%	-1.72%
2017年2月	112.6	7.85%	3.11%	107.3	6.55%	2.78%	116.2	8.70%	3.38%
2017年1月	109.2	5.00%	0.74%	104.4	4.09%	0.68%	112.4	5.44%	0.81%
2016年12月	108.4	4.53%	-0.18%	103.7	3.08%	-0.58%	111.5	5.39%	0.00%
2016年11月	108.6	4.32%	1.31%	104.3	4.09%	1.66%	111.5	4.60%	1.18%

4.2.4 不要忘记检查格式

2.日期格式的修改

在检查数据格式时，如果发现数据格式不对，直接选中数据修改格式即可。但是日期格式的修改却是例外，尤其是当日期的书写方式都不统一时，直接更改格式依然不能使日期格式统一。此时可以选择【分列】功能来实现日期格式的修改。



日期	消费者信心指	
	指数值	同比增长
2017年3月1日	121.3	11.69%
2017/4/5	123.9	15.58%
2017/5/6	118.6	13.38%
2017/6/4	114.7	8.62%
2017/7/8	114.6	7.30%
2017/9/6	113.3	10.11%
2017/10/5	112	12.22%
2017/8/7	113.4	12.28%
2017/10/12	111	11.00%
2017/12/10	112.6	7.85%
2017/12/15	109.2	5.00%
2017/12/18	108.4	4.53%
2017/12/19	108.6	4.32%

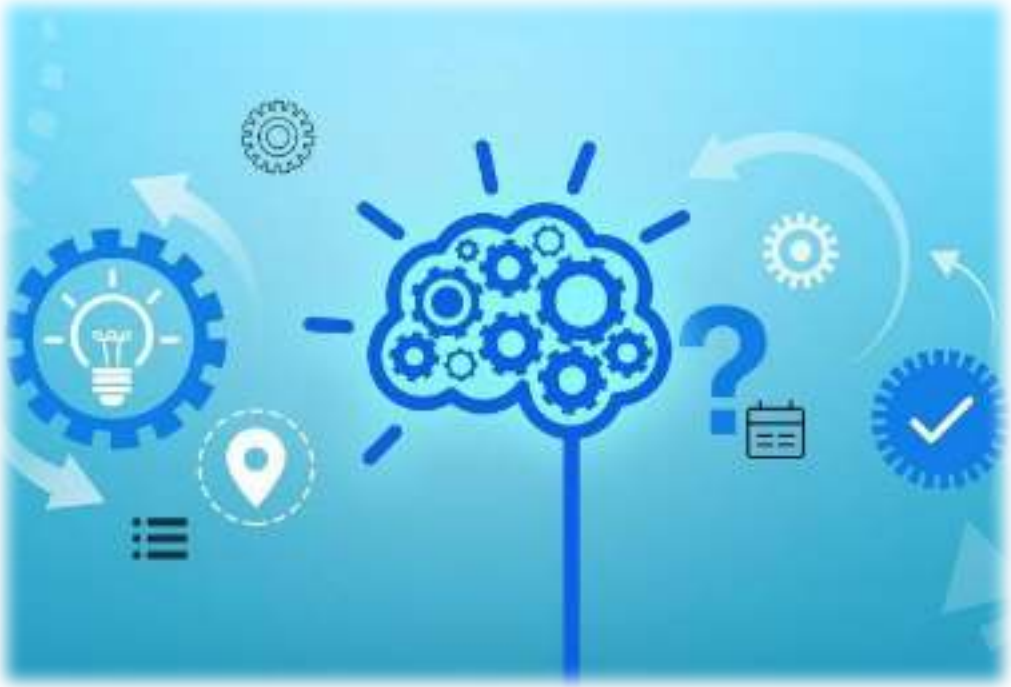
高手自测11

通过什么方法可以快速删除数据中的重复数据？

4.3 数据处理第二步：数据加工

经过数据清洗步骤，数据表中的数据已经没有错误值存在，这时要根据数据分析的目的不同，对数据进行加工。

数据加工可以说是启发数据分析灵感的一个步骤。如加工过程中，可以对不同项目的数据进行求和、平均数计算。在进行数据分析时，可以通过数据的项目的和、平均值，发现特别的数据规律。如果没有数据加工这个步骤，岂不与这一数据规律失之交臂？



4.3.1 数据计算

1.简单计算

在Excel表格中，使用函数要学会为单元格“命名”。单元格名称加上运算符号可以进行单元格数值的简单计算。

如第A列的第1个单元格，名称为“A1”。那么A1单元格与A2单元格数据之和的计算公式，就为“=A1+A2”。以此类推，将单元格与不同的运算符组合，可以对不同单元格的值进行不同方式的运算。其中运算符的优先级别与数字中的一致，先进行乘除运算，再进行加减运算。如果想先进行加减运算，需要添加括号。