**Excel数据分析**

数据分析步骤

1 提出问题 定义清楚要研究的问题目标和方向

2 理解数据 采集相关数据+查看数据集的信息，包括描述这个数据集的信息 从整体上理解数据，理解excel里面的每个字段的含义是什么

3 **数据清洗** 数据预处理，脏乱数据，重复数据

常用的数据清洗的函数：

数据重复【删除重复项】

数据抽取【left,right,mid】

数据计算【average,sum,max,min,date,if,or,countif】

4 构建模型 对清洗后的数据进行分析，简单的是分析一些简单的业务指标，复杂的就是用机器学习算法来训练模型

数据分析【数据透视表】

描述统计分析【分析工具库】

多表关联查询【vlookup】

5 数据可视化 最后结果用图表交流展示给老板和客户（如何使用excel绘制图表，如何使用PPT制作报告）

数据来源：招聘网站数据

关心的问题： 在哪个城市比较好找到数据分析工作？数据分析师的薪酬如何？发展前途？

养成好习惯：在开始分析前都要把原始数据复制一份，副名叫做-清洗数据

**数据类型**：字符串（文本类型，汉子字母），数字，逻辑（true,false）

注意以字符串形式存储的数字，不能参与计算，要先把这类数据先处理成数字类型才能用于计算

怎么判别？

选择单元格查看数据类型format cells，显示General,表示excel根据数据形式自动转换为对应的格式，在显示general的时候，它代表的是全部的数据类型，我们没法区别这个数字是字符串还是数字，**默认字符串左对齐，数字类型右对齐**。如果一个数字是左对齐的，应该是字符串。

**数据清洗步骤**：

S1 选择子集：选择进行数据分析感兴趣的几个列

隐藏列名 在列号上右键单击选择隐藏 好习惯：尽量隐藏数据，不随意删除，保证原始数据的完整，当以后想用到这一列时还可以恢复：全选表格（左上角小三角），home\*（开始），rows and columns, hide and unhide

S2 列名重命名：可以把一些别扭的列名按照自己的命名习惯换成方便分析使用便于理解的名字

双击列名

S3 删除重复值：数据，删除重复项，取消全选（默认全选，但此时我们只对一列查找重复数据），勾选列名(比如职位ID)，点确定

S4 缺失值处理：先要统计出有多少缺失值，excel中可以选定一列，在屏幕最下行查看计数，根据不同列的**count**差别判断哪些列有多少缺失值。（这里因为职位ID是职位唯一的标识，通过与业务部门沟通我们可以确定这一列时完整无缺失的。所以可以以职位ID有多少个为准，比它少的有缺失值）

如何处理缺失值？四种方法

1. 人工手动补全，适用于缺失数据比较少的情况
2. 删除缺失数据
3. 用平均值代替缺失值
4. 用统计模型计算出的结果代替这个缺失值

如何快速定位缺失值？点击列号选择整列，开始，查找和选择find and replace，定位条件go to，空值blanks, 确定

如何用人工一次性把所有缺失值都补全：定位缺失值上述步骤（单击列名go to,blank），在上方编辑栏输入eg“上海”，输入完后按住ctrl+enter，其他空白单元格都变成“上海”。

S5 一致化处理：对数剧列里面没有统一命名的值进行处理

一致化指的是数据是否有统一的标准或命名

比如“公司所属领域”，有的公司属于多个领域，用逗号分隔开，这种格式不一致会影响后面数据分析的结果，要对这个格式进行统一命名和处理。 可以使用excel的分列功能，对某一列进行分隔和拆分。 **分列功能会覆盖掉右列单元格**，所以要记得**把要分割的这一列先复制到最后一列。**隐藏原列。

如何**分列操作**：数据，分列text to columns，分割符号-用分隔字符如逗号或制表符分隔每个字段delimited，下一步next，连续分隔符视为单个处理treat consecutive deliminator as one，其他（自己填使用什么符号进行分割），下一步，完成。

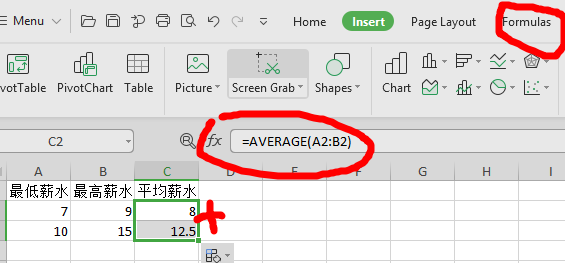
**Excel常用函数的使用：**例如薪水这一列，文本格式表示的几K到几K，如果我们想计算平均薪水很不方便，想分列成：最低薪水，最高薪水两列。

在表格最后空白列建立三个列：最低薪水，最高薪水和平均薪水，用于存放数据清洗过后的薪水数据。如何达到这个效果？方法1，先使用分列功能，通过分割符-进行拆分得到两列数据，然后利用替换功能删除k这个字符串最后得到结果。

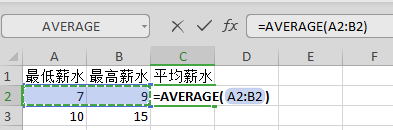
方法2，利用**文本函数查找方法**清洗这列数据：

函数就像**黑盒子**，不需要了解内部运作原理，只需要知道它的**功能**是什么，**输入**是什么，**输出**是什么。把重复的工作放在一起，便于使用。（套路/模块）例如excel中求平均值的函数叫average：

假设一个表格只有两列，最低薪水和最高薪水，最后一列平均薪水待求，选中要插入函数的单元格，公式，插入函数fx，搜索函数：平均值，转到，average，确定。 拖动要输入的单元格，会自动录入输入参数的entry，eg A2:B2. 被插入的单元格显示计算后的数值。将鼠标放在这个单元格右下角会出现十字架，双击十字架会把同个函数运用在整列上得出结果，自动算出每一行的平均值存在这一列中。 ——用excel的图形界面插入函数。



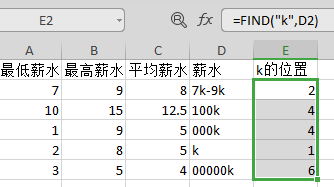
还有一种插入函数的方法：直接点击单元格，输入等号，在等号后面把函数的表达方式写入，熟练函数后可以这样操作。

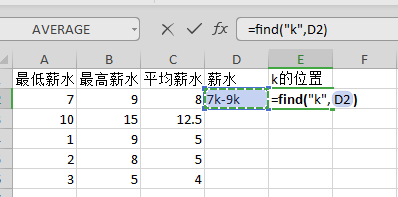


Excel函数特别多，学习方法是在遇到问题之后用搜索引擎搜索，搜索结果中会告诉你使用哪个函数。

**Find函数**：Find查找一个字符串在另一个字符串中出现的起始位置

Find(要查找的字符串，字符串所在单元格位置) 从1开始记位

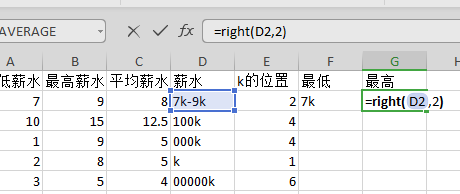
1. Indexed



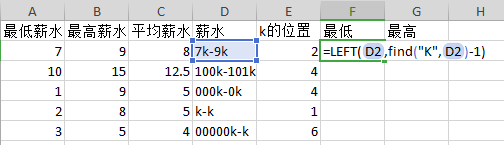
下面几个函数经常和FIND配合使用来完成字符串截取的工作

**截取字符串中的内容：Left,Right,Mid**

Left/Right(字符串所在单元格位置，从左/右开始到XX位置进行截取) 包括XX 闭合区间

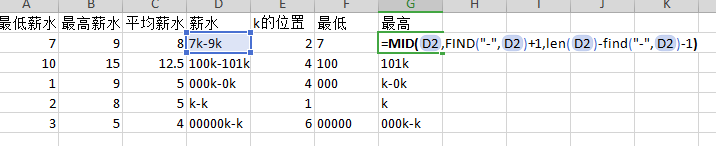


结合find函数截取**最低**薪水（只要数值部分，不要k）



**最高**怎么算？Tricky

**Mid(字符串所在单元格位置，开始位置，截取长度)**

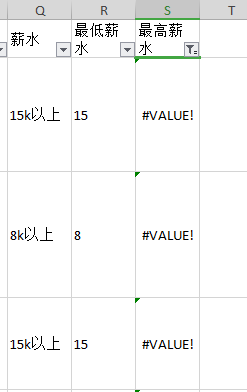


Len表示获取整个字符串的长度

Excel的**筛选功能**：按条件查找数据

数据，筛选（漏斗形图形界面）AutoFilter,每一列上会出现小三角，选定要筛选的那一列的小三角，拉到最下面选#value！表示数值不正常，我们要筛选出这些错误数值看下是什么原因造成的。不选全选，只选最下面的错误数值单元格。找到后发现因为有些薪水中“K”是大写，而find函数大小写敏感，我们找小写“k”会使很多最低薪水为空值。（筛选结束后记得全选把所有数据显示出来，免得影响后续数据分析。）

解决办法：**查找替换功能**

单击某列选列，表示只针对这一列操作，开始，查找和选择，替换，查找内容：K，替换为k，全部替换。

再次对这列使用筛选看还有没有其他错误。

对最高薪水使用筛选功能找到异常值，发现

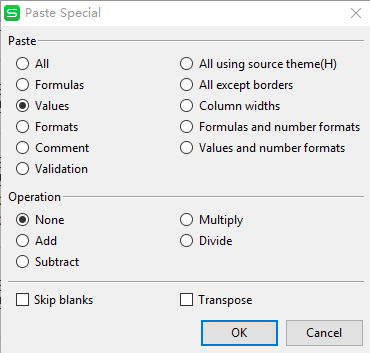
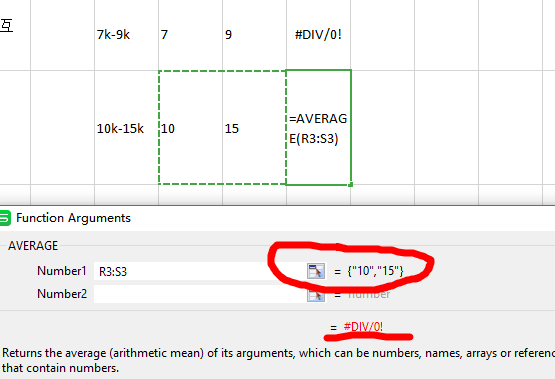
因为薪水写“15k以上”按照我们之前的函数

算不出最高值。针对这种薪水描述我们可以

将最高薪水=最低薪水。直接双击单元格，

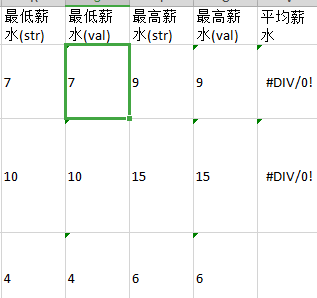
输入等于号，=最低薪水单元格的位置。

修改后再重复筛选下，直到没有重复错误。数据清洗不是做一次就可以，需要反复做多次直到没错误。 数据清洗占据了数据分析过程的60%的时间。

现在数据有限，只知道薪水的范围，按理说不能用范围来求平均薪水。但是目前只有最高薪水和最低薪水。新建一列叫平均薪水，在单元格中插入average函数，对最高薪水和最低薪水两列求平均，会出现黄色警告，有错误值，#DIV/0！被零除错误，点击关于此错误的帮助，会知道这个函数除数位置选择了0或空白单元格，在excel中空白单元格会被当成0来处理；第二个原因是部分单元格内的数字是字符串类型，数字型字符串，怎么看出是字符串类型？

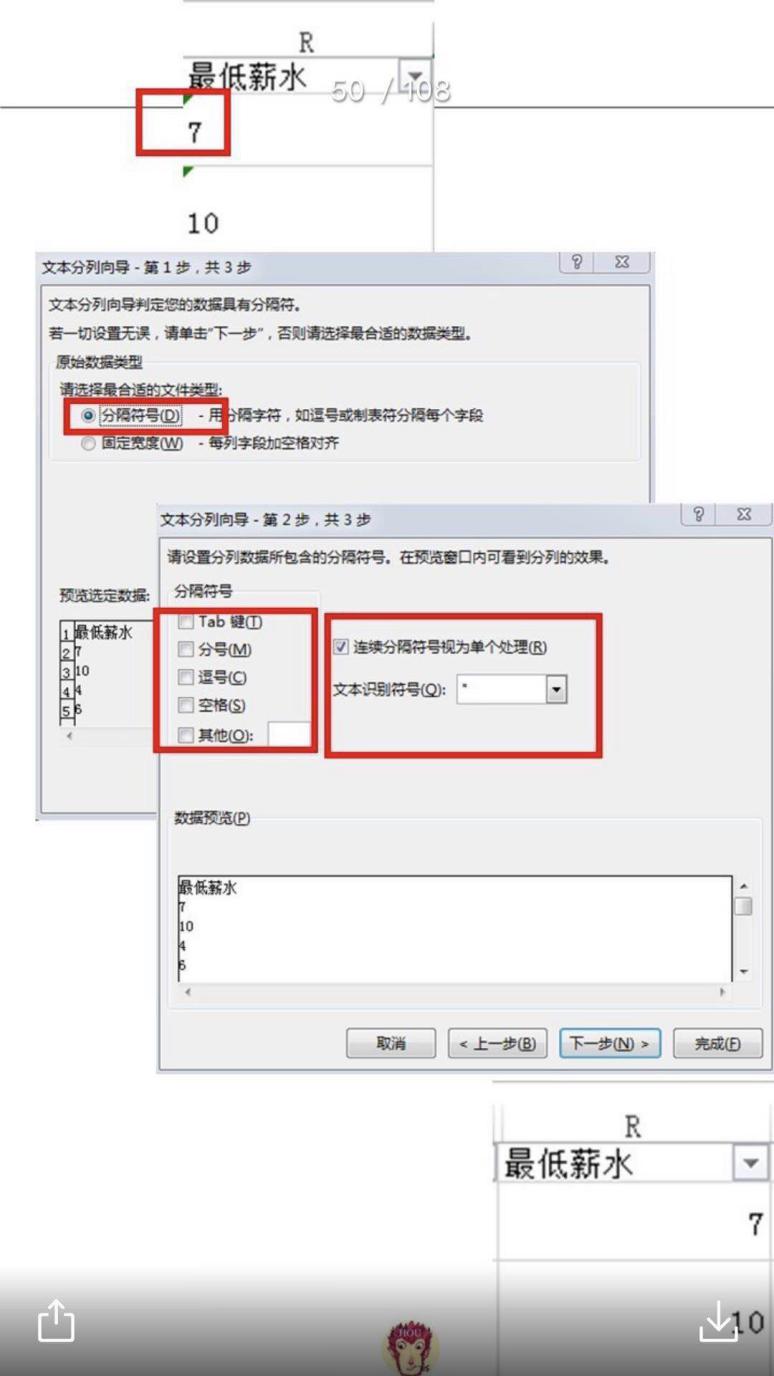
圈起来的10和15是带双引号的说明是字符串。

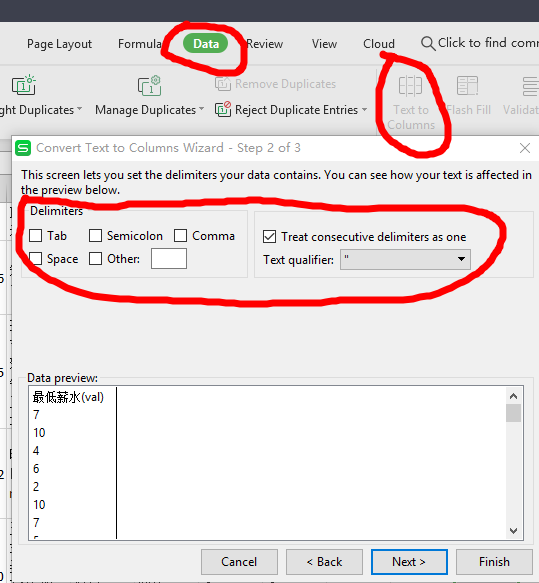
在最低薪水上邮件点击插入insert，会在这一列左侧插入列，点击选择性粘贴paste special，“粘贴”中选“数值”，“运算”中选择“无”，确定，就可以把最低薪水这一列复制成这个效果：



被复制出来的列会在左上角有个绿色的三角，在常见的证件号码e.g.身份证时用于表示这个数值是字符串类型，需要使用分列功能将它转化为数字。

在数字选项卡中选择分列，在文本分列向导弹出窗口中选择分隔符号，下一步，在分隔符号这一栏左侧什么都不勾选，右侧勾选连续分隔符号视为单个处理，文本符号’’,点击完成后会发现绿色点的东西消失了，单元格中的数字变成了数字类型，右对齐了。可以用于运算了，记得把复制前的最低薪水这一列隐藏掉。





S6 数据排序：用于发现更多有价值的信息

可以看数据分析行业薪水最高的职位，**排序**：

排序前要选择要排序的这一列，e.g.平均薪水列，开始，排序和筛选，降序，排序提醒中选择**“扩展选定区域”表示其他列也跟着移动**，排序。

只有数据清洗做得好才能按照要求完成报表。

S7 异常值处理 需要用到excel中数据透视表的功能

**数据透视表**的原理：

数据处理模式

Split（数据分组）-Apply（应用函数）-Combine（组合结果）

这三个词代表了数据处理的通用模板方法。

**S**plit:按照某种规则对数据分类，对数据按属性分组，每一组里的数据都有一个相同特征，e.g.将到达同一个目的地的航班数据行分成同一组

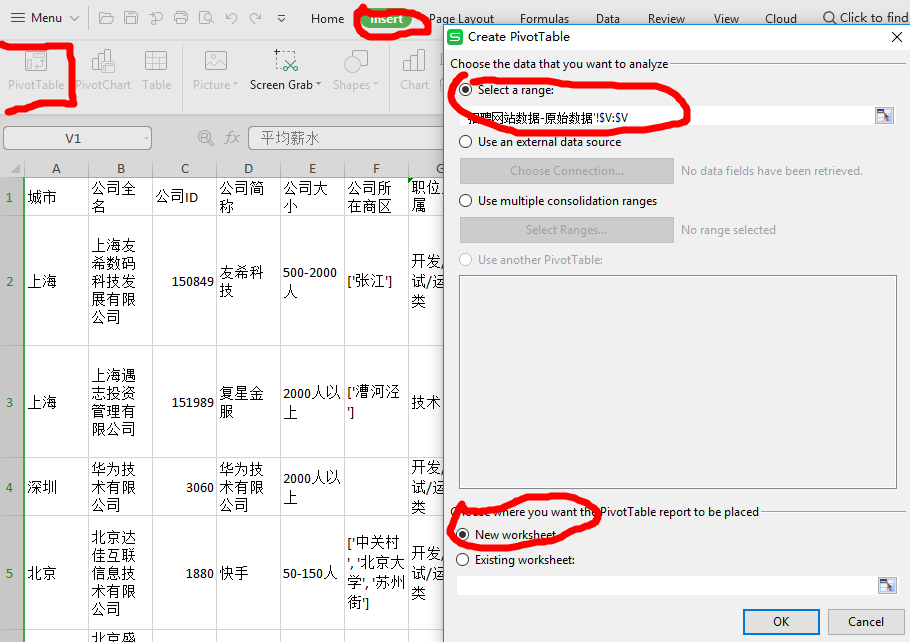
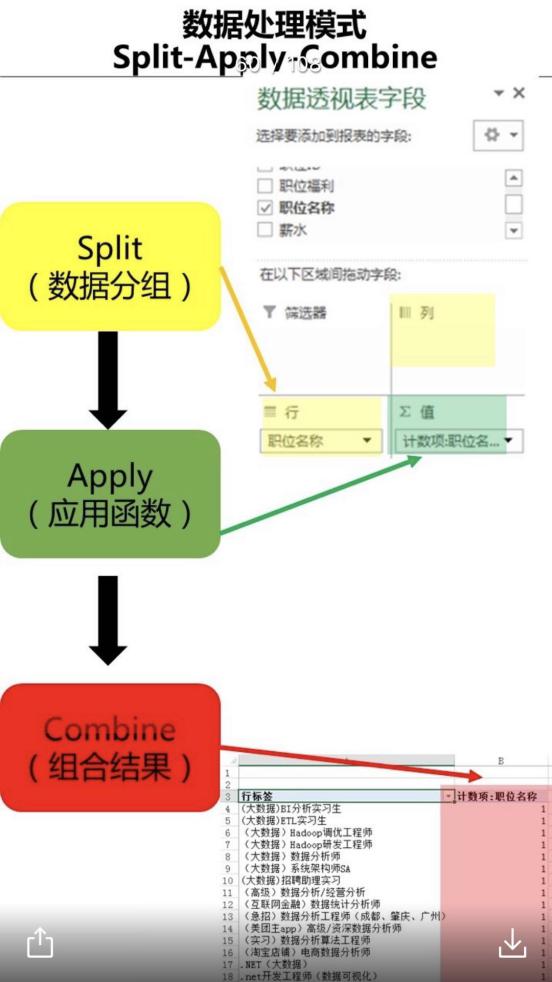
**A**pply: 对分组后的每组数据进行操作，e.g.求平均值，标准差etc. 因为是分组计算，每一组都会有一个计算结果。每一组对应一个值。

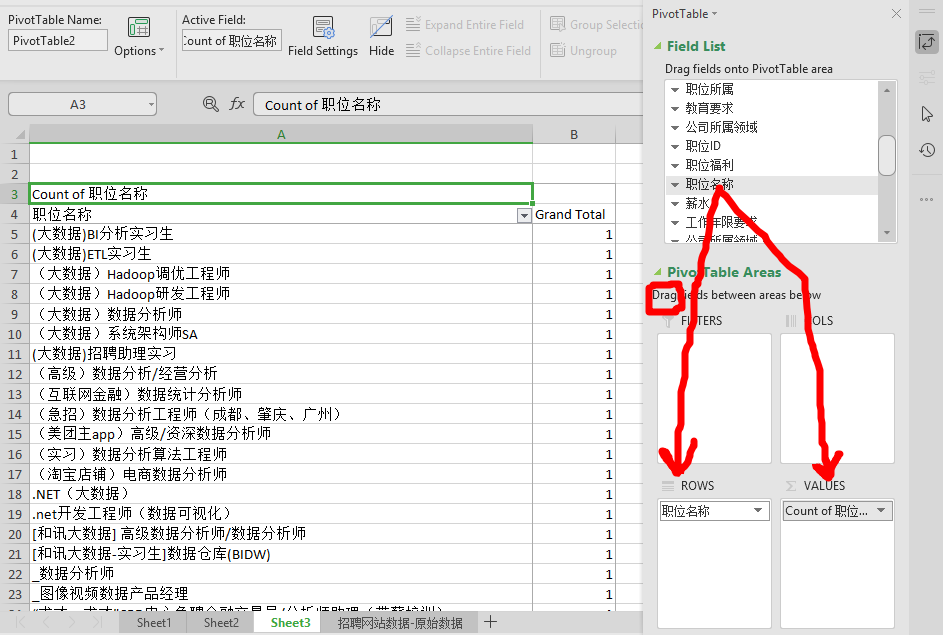
**C**ombine:对第二步Apply应用函数的计算结果进行汇总，（合成到一张表里）。因为涉及到多个组，要把每个组的计算结果汇总起来，通过例子看下这三个过程怎么使用：

——excel的数据透视表功能就是对这个分析模式的实现。

如何在excel中实现这个操作，创建数据透视表：记得先全选整张表格

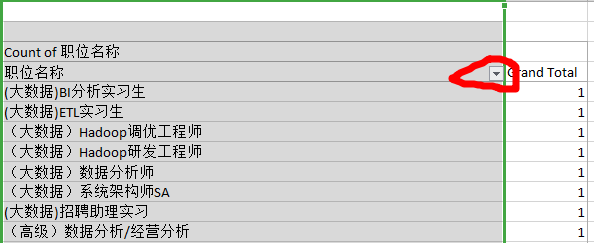
Insert插入，数据透视表PivotTable，在弹出的数据透视表窗口中选择“新工作表”，（其他部分默认），确定





把职位名称拖到行的位置，再拖到求和的位置，表示对职位名称这一列相同的值分组后分别统计出出现的次数

下面用**数据透视表的排序功能**看一下哪些职位出现的次数最多：下拉行标签后的三角，选择其他排序选项，降序排序，计数项of职位名称，确定。



我们要把不属于数据分析师的名称去掉（HR的错误归类，把开发也归入数据分析师岗位）

如何**去掉（职位名称中的）异常值**？

* **将数据分析师的职位查找出来**，属于数据分析师的职位**标记**：是。不属于数据分析师的职位标记:否。

1.查找关键词，找出包含这些关键词的岗位

**FIND**（{“数据运营”，数据分析，“分析师”}，L2）

1. 当不包含这些关键词时查找函数会返回一个错误标识，用计数函数处理：如果find找到关键词会返回位置，如果没找到关键词会被统计为0

COUNT(

FIND（{“数据运营”，数据分析，“分析师”}，L2）

)

3.如果查找到了关键词，有这个词的位置，赋值为“是”。

IF(

COUNT(

FIND（{“数据运营”，“数据分析”，“分析师”}，L2）

)

,“是”，“否”

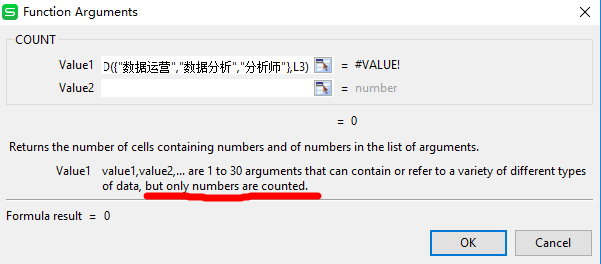
)

* 将值=是的职位名称**筛选**出来，就是我们要的数据分析师职位名称。

在职位名称后面插入一个列，命名为数据分析师职位名称，插入函数就可以知道对应的职位是否属于数据分析师

Find函数进行定位时，总是从指定位置开始，返回找到的第一个匹配字符串的位置，而不管其后是否还有相匹配的字符串。

count是"计数"的意思,即用来计算包含**数字**的单元格的个数



所以在find函数返回找到位置/#value！时，外层的count函数会返回0/1,（并不在意找到的实际位置），只是将查找函数结果转化为0没找到/1找到了，方便为后续的IF函数提供输入。if函数中的第一个参数logical-test，我们给它喂给了之前的count函数的结果，只有0、1两个值，这个参数为1时返回value-if-true,（即“是”，in this example）该参数为0时返回value-if-false（即“否”in this example).

这里if函数的应用并没有在第一个参数位置logical\_test处进行逻辑判断而得出0/1，而是直接将只有0/1两种结果的函数喂给它，只是用于将0/1转化为“否”/“是”。

下面利用筛选功能把符合条件的筛选出来（数据分析师职位名称这一列中为“是”的值选择出来），对筛选出来的数据要后续做数据透视表进行分析。 Data-AutoFilter-每一列出现小三角-下拉数据分析师职位名称-勾选“是”-确定

但数据透视表只能对数据源里的所有数据作分析，因此我们需要将筛选出来的数据复制到一个新建的excel文档里，给这个文件命名为后缀：数据清洗结果。后续分析都在这个文件中进行。

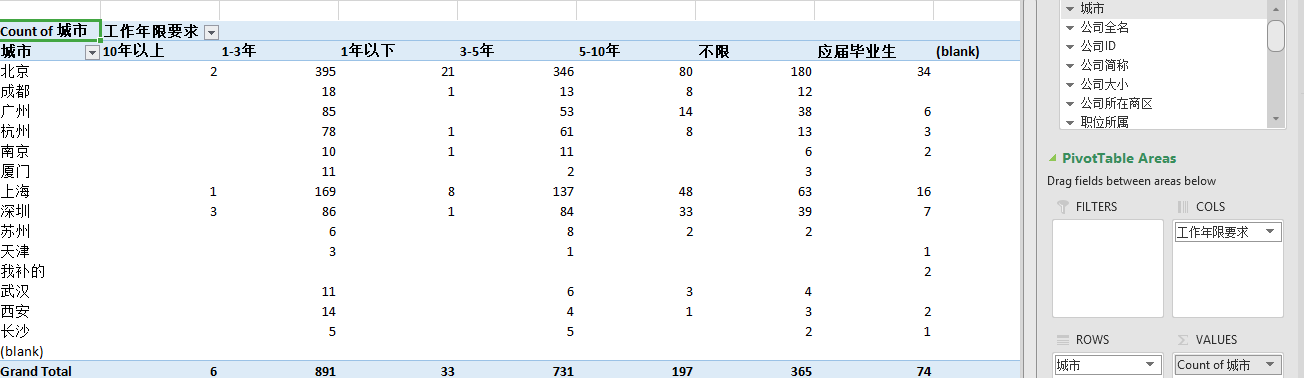
4 **构建模型**

这里只在excel中分析业务指标，利用**数据透视表**进行。

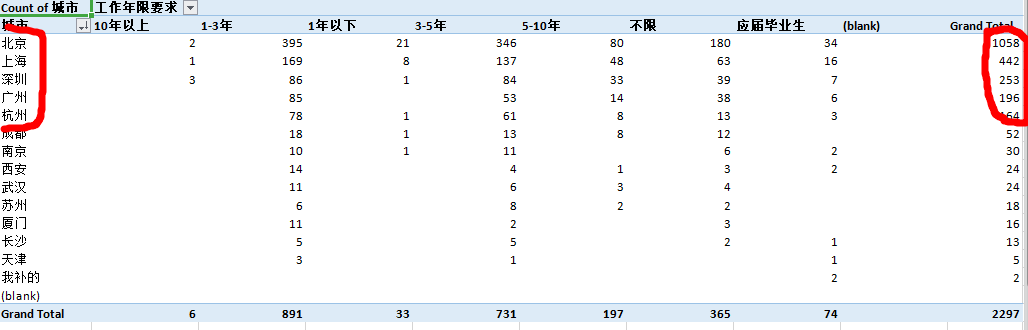
* 分析第一个问题：在哪些城市找到数据分析师的工作的机会比较大？

打开插入选项卡下的数据透视表：在“行”处拖入“城市”这个列名，就可以按城市进行分组。

在“列”处拖入“工作年限要求”，因为我们想看到不同工作年限的工作职位数量，在求和值处拖入列名“城市”，表示按求和来统计每一个组（即城市）下面的招聘职位数量，也就是计数。 然后排序，看哪个城市的统计数量最多。



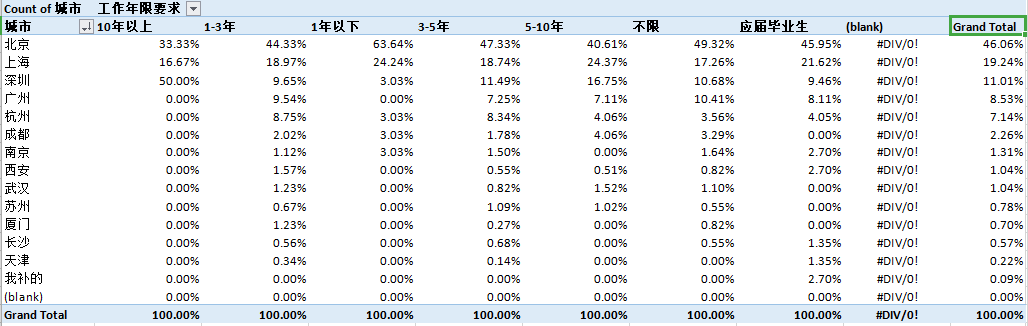
在城市这一列的下拉三角中选择More sort options，降序排列，count of城市，确定。



将近一半数据分析岗位在北京，大部分工作年限要求在1-3年，5-10年。

**用百分比展示数字报表**：

总计这一列-右键-值显示方式value field settings-show value as列汇总的百分比-% of column total



* 分析问题2：平均薪水这一列的**描述统计分析**：

要先**安装excel的数据分析功能**

文件-选项-加载项-管理：Excel加载项，转到-确定-分析数据库-确定（从WPS没找到转移到了Microsoft Excel）

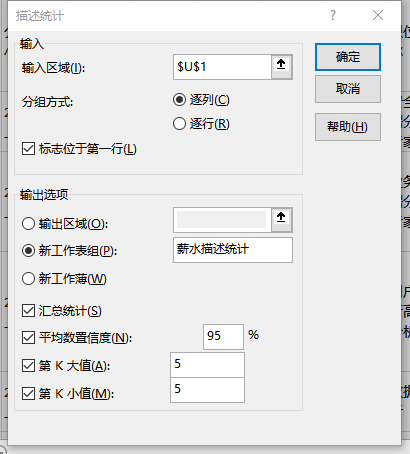
描述统计分析：

数据-数据分析-描述统计-确定

点击 输入区域 点击选择“平均薪水”这一列， 勾选 标志位于第一行，表明第一行是列名，不包括在计算里面，

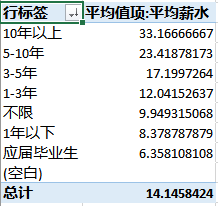
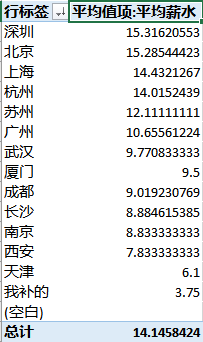
输出选项中 选 新工作表组：薪水描述统计-工作表名称

勾选汇总统计 平均数置信度：95% 第K大值：5 第K小值：5



* 分析问题3：数据分析师的薪水如何？

数据透视表：按城市分组统计薪水，城市拖到行；在sum值中拖入平均薪水，表示计算平均薪水这一列的值，下拉选择平均值项（因为数据透视表对数字类型默认的是求和计算，但我们这里想知道每个城市的平均薪水，所以在报表平均薪水这一列右键-值汇总依据-平均值）。



* 分析问题4：根据工作经验的不同，薪水是怎样变化的？（按照工作年限来分组-行，平均薪水的平均值作为列）

善于利用搜索引擎

**日期数据处理**

由于统计来源的不同或者记录数据人员的错误，导致日期格式各种各样，没有统一的日期格式对后续的数据分析特别不方便，需要对日期数据进行处理。

* 统一日期格式



S1-选中日期这一列

S2-菜单栏-数据-**分列**-前两步弹出窗口不要做任何更改，直接点下一步（1分隔符号，下一步，2 Tab键，文本识别符号”，下一步），第三步在列数据格式中选择日期，**YMD**，完成。结果可见大部分日期有了统一的格式：



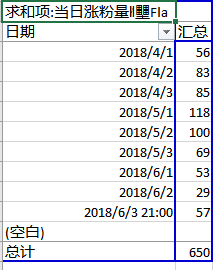
但有一些不是我们想要的结果，进一步处理异常值：选中**日期这一列-右键-设置单元格格式**-数字选项卡中的分类里选择**自定义**-类型中填**yyyy/m/d**,确定——所有日期统一成想要的样子



现在这个表格中的数据是每天的涨粉量，如果我们想看每个月的数据或者想看最近几个月用户数量的变化趋势，如何分析？

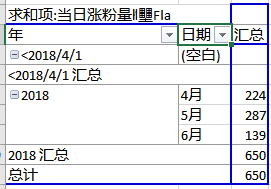
* 使用**数据透视表**实现 **按月汇总数据** 这个业务需求。

插入-数据透视表-选择一个表或区域-新工作表（其他地方都默认）-确定 此时会创建一个新的数据透视表。数据透视表的优点就是能将原始数据和汇总后的数据分离开。 日期拖到行，当日涨粉量拖到sum值，求和。

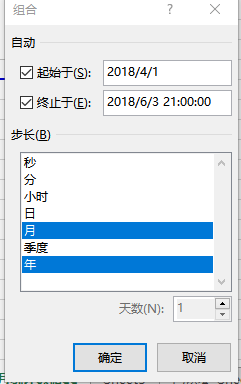


这个操作完还是按日统计的每天涨粉量的总和，要按月汇总，在数据透视表中还需要进行一些设置。

在报表中日期这一列随便选择一个单元格，右键，**创建组**，在弹出的组合窗口中选择**年和月**，确定，这时数据**就按月进行了分组统计。**



* 有时业务需要**按照周进行数据汇总**，统计每周的用户量，其实和按月汇总一样，都是在这个组合对话框中设置。



但是要把起始于这个框的勾选取消，并按实际数据，把起始日期设置为数据中的最早日期，且为星期一的那一天（因为我们要从星期一开始算起）。比如本**数据中最早的星期一**是2018年4月2日，把这个日期填在起始于那一栏。S2-把步长设置为**日**，**天数改为7**，确定。这时可以**按周来汇总数据**。

* 看用户最近几个月的变化趋势：按月汇总数据后画折线图，看出用户随时间的变化。（可视化后面讲）
* 看**每个月中最大的当日涨粉量**是多少：数据透视表字段设置中的求和项-当日涨粉量这一栏选 **值字段设置**-汇总方式的计算类型选 **最大值**-确定-就可以看到最大值项的当日涨粉量（根据业务需求选择适合的计算类型）

每个月中的最大值

**多表关联查询-vlookup**

将一个表中的数据匹配到另一个表里-vlookup函数，可以按条件查找数据。

vlookup函数可以极大提高工作效率，避免磨磨唧唧复制粘贴一天，节省时间，最容易上手，要求掌握。

有四个参数：

**Vlookup(找什么，在哪找，第几列，是准确找还是近似找)**

精确查找指的是从第一行开始往最后一行逐个查找，一找到第一个匹配的行就停止查询，所以返回的是找到的第一个值。 模糊查找时会查遍所有数据，返回的是最后一个匹配到的值。

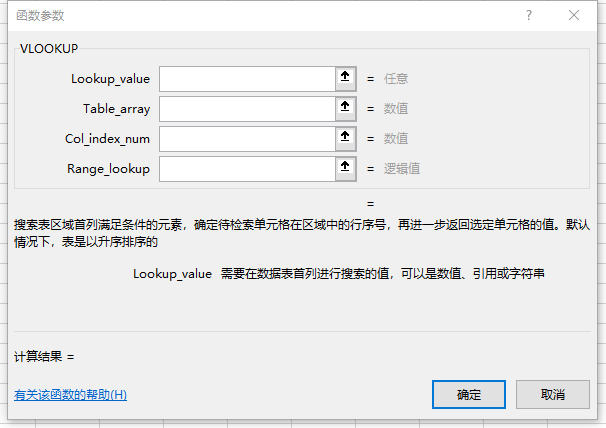
实际案例：表一-学生信息，记录了全部的学生信息；表二-三好学生，记录了三好学生的姓名

想知道表二中三好学生学生的班级是多少？因为姓名和班级都在表一中，所以需要用表二中学生姓名作为查找条件，在表一中查找出学生的班级——这就是多个表之间的跨表查找。

Vlookup函数就是手动查找过程的自动化实现。

先选中要插入函数的单元格，（三好学生表中班级这一列），公式，插入函数，搜索函数：vlookup，转到，确定

**Vlookup(lookup\_value,table\_array,col\_index\_num,range\_lookup)**搜索表区域首列满足条件的元素，确定待检索单元格在区域中的行序号，再进一步返回选定单元格的值。默认情况下，表是以升序排序的。



Lookup\_value:需要在数据表首列进行搜索的值，lookup\_value可以是数值、引用或字符串

点击要找的东西，（点击需要匹配的数据的单元格，）（只点一个即可）此处：点三好学生名字列下的 猴子

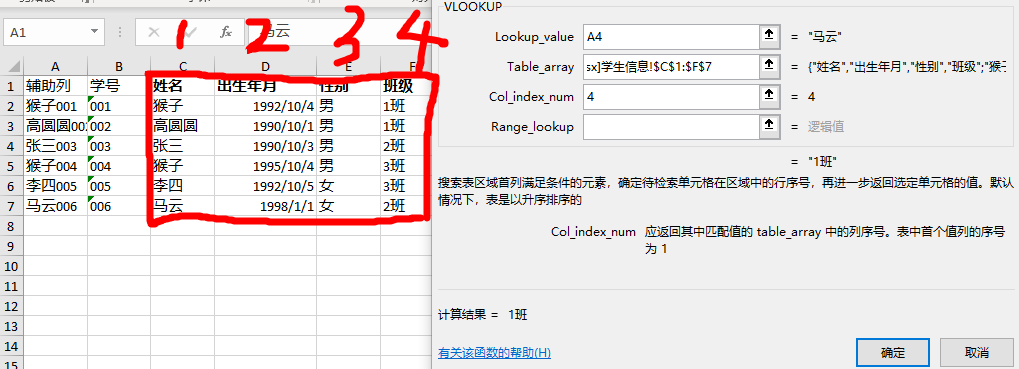
表示要找到姓名为“猴子”的信息

Table\_array:要在其中搜索数据的文字、数字或逻辑值表。Table\_array可以是对区域或区域名称的引用。

在哪找，表示在那个表里查找信息。返回到第一张学生信息表中，选中全部数据，特别注意：选中查找范围里第一列的值，必须是要查找的值，不然就会报错，

Col\_index\_num：应返回其中匹配值的table\_array的中序列号，表中首个值列的序号为1.

即第几列，要查找的值在前面选定的查找范围里面属于第几列？表1中的班级信息在选定的查找范围中是第4列，（因为是**从选中范围的**姓名那一列开始算起）



Range\_lookup: 逻辑值：若要在第一列中查找大致匹配，请使用TRUE或省略（1）；若要查找精确匹配请使用FALSE（0）。

一般情况下都会精确查找所以此处设置为0.确定

总结下vlookup是干嘛的：**多标关联查询时会使用，查到到某个值并粘贴过来。**

弱点：当有多个值满足查找条件时只会返回第一个值。

遇到这种**有重复值的情况**，想都找出来怎么办？如何解决这个问题？

S1 我们可以**添加一个辅助列**，将姓名和学号联合起来作为**非重复的唯一字段**。

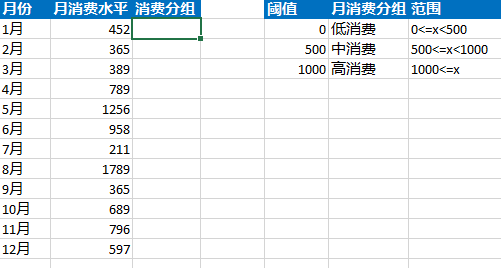
在学号前面插入一列，命名为辅助列，在单元格中输入公式：C2&B2，**&表示将两个单元格中的内容拼接在一起**。

S2 用辅助列作为查找条件，使用Vlookup函数,就可以查找出来.

**如何对数据分组，分析各组之间的关系？**

例如对销售数据，需要根据各个月的消费情况，将月消费水平分为高消费，中消费和低消费三组，如何**对数据分组**？主要使用**vlookup函数**实现。

S1在表的右边建立一个分组定义，定义一个范围，将每个范围的最小值作为阈值。





S2在消费分组列下选中要插入的单元格，插入vlookup函数，设置参数

Lookup\_value:月消费水平

Table\_array:阈值，月消费分组，范围这个表

Col\_index\_num：2，第二列（月消费分组这一列，低中高消费）

**Range\_lookup:1（大致匹配）**

大部分时候使用vlookup第四个参数都选0，精确匹配；而在**使用它进行分组时要选1，模糊匹配**。因为我们查找的是范围。

S3 需要对Table\_array参数，在哪找进行修改：先选中第二个参数中的值，**按F4**，就可以把这里的值**从相对引用变成绝对引用**，也就是**在列号和行号前面都加了美元符号**。 点确定，可以看到分组结果。

新概念：绝对引用，是EXCEL中三种引用的一种。

**三种引用方式**

相对引用：A1

绝对引用：$A$1

混合引用：$A1 A$1

通过员工工资表看一下这三种引用分别如何使用：

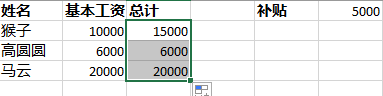
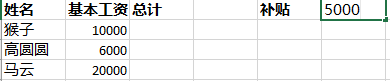
* 员工工资表包括每个月的基本工资和补贴，要算出每个员工的总工资，要将每一行的基本工资和补贴相加。

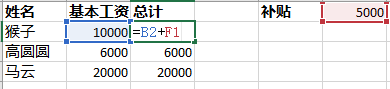
在第一行总计的单元格下输入=**B2+C2**，可以计算出第一行的总工资，然后将鼠标放到这个单元格的右下角出现黑色十字架时双击，就把该公式应用到这一整列了。填充完这一列我们会发现这一列中每一个单元格公式中两个相加的元素的行号和列号都会随着单元格的改变而改变，这就是**相对引用**。第二个人的总工资计算公式会自动变成B3+C3，第三个人的总工资是B4+C4，相对引用是指当把公式复制到其他单元格中时，行号和列号会根据这个实际的单元格的值相应的改变。

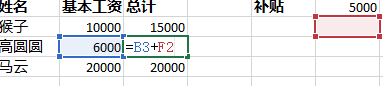


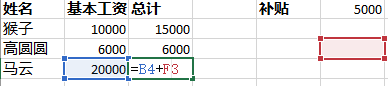
* **绝对引用**：

改变例子，假如每个员工的补贴是一样的都是5000，现在要计算员工的总工资，如果按照刚才的方式输入公式，会发现只有第一行的数据是有加上补贴的，其他行没有。



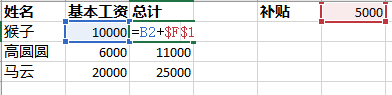


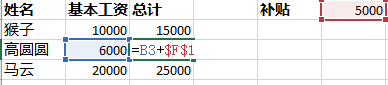


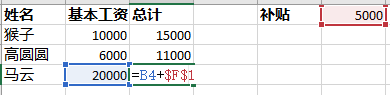


这是因为**引用了相应的空白单元格**，而真正的补贴单元格只是在F列的第一行。对于这种情况就需要用到绝对引用，也就是在复制公式时，公式中的补贴单元格的行号列号要不变，指向的是“补贴”那一个单元格。下面改成绝对引用看效果如何。

绝对引用是在行号和列号前面都加了一个美元符号，双击加号会发现“补贴”这个加数的行号和列号都没有改变，这就是绝对引用。







* 混合引用：行或者列其中一个是绝对引用，另一个是相对引用。



在行号/列号哪个前面加美元符号表示哪个不变（绝对）。每次按美元符号麻烦，EXCEL设置了快捷键F4帮助我们快速切换这三种引用。选中想变换的部分，来回切换F4就可以在三种引用中来回切换。