前一篇分享了《Excel 数据分析必掌握的 43 个公式》，今天这篇讲实操，教大家用 Excel 做一次简单的分析。一是让大家了解数据分析是一个怎样的流程 ;其次熟练 Excel 的操作(学的知识要用起来)，包括公式，数据透视表等。

这里我用 Python 在智联招聘上爬取了约 1800 条的 BI 工程师的职位信息，并且将岗位名称、公司名称、薪水、所在城市、所属行业、学历要求、工作年限这些关键信息用 CSV 文件保存下来。

爬的过程就不赘述了，源数据附给大家，操作版本：Excel 2016 ，WIN 10

一个完整的数据分析都需要经历这样几个步骤：

- 数据获取——这里我已经用 Python 爬好了；
- 明确分析目的——你拿这数据要得到什么信息，解决什么问题；
- 观察数据——各个数据字段的含义，中英文释义；
- 数据清洗——无效值、缺失值、重复值处理，数据结构是否一致等；
- 分析过程——围绕目的展开分析；
- 制作可视化——做图表做可视化报告。

一、明确目的

数据分析的大忌是不知道分析的方向和目的，拿着一堆数据不知所措。数据用来解决什么问题？

是进行汇总统计制作成报表？

是进行数据可视化，作为一张信息图？

是验证某一类业务假设？

是希望提高某一个指标的 KPI？

要知道一切数据分析都是以业务为核心目的，所以要找到业务问题的思考点。关于找到问题的切入点，之前数据分析思维篇讲过。永远不要妄图在一堆数据中找结论，目标在前，数据在后，哪怕是把数据做个平均值比较，也比没有方向好。每一步尝试都会引发进一步思考，比如为什么这个值这么低，原因在哪里，这个差异波动有何规律……

所以，分析前不妨先来看一下我们爬的数据：

假设我是一个 BI 工程师，我想知道：

目前 BI 工程师的平均薪资水平如何，薪资的区间分布如何

各地区对 BI 工程师的需求量是多少，哪些地区设岗最多。

不同年限的 BI 工程师薪资差异如何，3 年后我差不多是什么样的价位？

薪水较高的公司有哪些？

带着这样的问题，那我们的分析就有了方向，后续则是将目标拆解为实际分析展示的过程。

二、了解数据概况

拿到数据肯定是要先看一下的，你想要的数据全不全，拿到的数据有哪些可分析之处。主要就是看数据字段，要了解数据字段的含义：

JobName——岗位名称

Company——公司名

Salary——薪水

City——城市

Jobtype——岗位领域

Edulevel——学历要求

WorkingExp——工作年限要求

三、数据清洗

接下来进行数据清洗。数据清洗一般包括无效值、缺失值、重复值处理；数据是否有乱码，错位现象；数据口径问题，两张表的关联 ID 名是否一致；还有是否有统一的标准或命名，如公司名全写或缩写的区分。数据转换则是将数据规整为统一格式处理。因为这是只是 Excel 级别的数据分析，且就一张简单的数据表，不会有太多复杂的操作。这里简单总结下。

1、有无缺失值

数据的缺失会很大程度影响分析结果。数据缺失的原因很多，比如数据采集的时候，因为技术的原因，爬虫没有完全抓去。但工作上更多的原因是数据入库的时候就没有收集全，有没填有遗漏，这又是数据规范数据治理的话题了。一般来说，如果某一字段数据缺失超过 40%~50%，就没有分析意义了，考虑删除或作其他措施。

看数据有没有缺失，只要在 Excel 中选中该列看计数。

这里，eduLevel 有缺失（1759/1800）但不多，不影响实际分析。

2、脏数据处理

发现 jobName 列里面有一些类似 BIM 工程师的岗位信息，这些应该都是土木行业的工程师，爬去时没做过滤，还有包含“bim”“BIOS”“BIW”等字段。

因为包含多重过滤，这里我建立辅助列，设立判断条件，然后进行筛选过滤。

```
=IF(OR(COUNTIF(A5,"*"&{"bim","BIM","BIOS","BIW"}&"*")),1,"0")
```

公式的意思是，如果含有这些字段中的任何一个则为 1，否则为 0。这里我们需要筛选出结果为 0 的数据，总计筛选下来 600 多条，数据还是很脏的。

多重筛选，还可以用数据选项卡里的高级筛选功能，就不掩饰了。

3、重复数据

重复数据一般对唯一标识字段来处理，比如用户 ID，订单 ID，公司 ID 这些，这些字段都代表这一行数据是唯一存在的。严格来讲，这里的表应该存在公司 ID 这一字段，爬取数据的问题，我这就懒得再重爬了，就对 Company 字段做重复值处理。

这里有一个快速窍门，使用 Excel 的删除重复项功能，快速定位是否有重复数据。对 company 列进行重复项删除操作：

只剩下 562 个值了。到此，一些脏数据基本清理的差不多了。

最后，salary 有一些数据是“薪资面议”，“校招”的，这里也一并过滤掉。Jobtype 过滤掉汽车、电子等行业，只留包含 IT 互联网行业，最后剩下不到 500 条数据。

4、数据再加工

一者是 salary 薪水用了几 K 表示，这是文本，不能直接用于计算。而且还是一个范围，后续得按照最高薪水和最低薪水拆成两列。

二者由于城市字段存储有的数据为“城市-区域”格式，例如“上海-徐汇区”，为了方便分析每个城市的数据，最后新增列“城市”，截取“-”前面的真实城市数据。

为了方便整理，和原数据区分，也防止原数据丢失，这里把之前处理的数据复制粘贴到另一张表里。

① 薪水处理

将 salary 拆成最高薪水和最低薪水有三种办法。

一是直接分列，以"-"为拆分符，得到两列数据，然后利用替换功能删除 k 这个字符串。得到结果。

二是自动填充功能，填写已填写的内容自动计算填充所有列。

三是利用文本查找，重点讲一下这个。

写公式的思路是，先查找第一个 K 出现的位置，然后再-1，去除掉 K。所以公式是：

```
=LEFT(C2,FIND("K",C2,1)-1)
```

	A	B	C	D
1	jobName	company	salary	bottomsala
2	bi工程师	上海江朋信息技术有限公司	8K-15K	8
3	高级BI开发工程师 大数据平台	瑞易信息技术(上海)有限公司	15K-30K	15
4	bi工程师	上海江朋信息技术有限公司	8K-15K	8
5	BI工程师	成都天地网信息科技有限公司	10K-20K	10
6	BI开发工程师	北京博辕捷迅科技发展有限公司	10K-15K	10
8	BI实施工程师	武汉东方赛思软件股份有限公司	4K-8K	4
9	BI实施工程师-南京	武汉东方赛思软件股份有限公司	4K-8K	4
10	BI实施工程师-广州	武汉东方赛思软件股份有限公司	4K-8K	4
11	BI数据中级JAVA工程师	深圳市后河控股集团有限公司	13K-26K	13
12	BI工程师/大数据分析	货车帮科技	20K-30K	20
13	BI工程师-金融项目-HR	深圳市拓保软件有限公司成都分公司	10K-15K	10
14	BI实施工程师-深圳	武汉东方赛思软件股份有限公司	4K-8K	4
15	BI开发工程师	上海耸智信息科技有限公司	10K-15K	10
16	bi工程师（数据分析方向）	广州市钱大妈农产品有限公司	3K-6K	6
17	ETL/BI工程师	北京泛鹏天地科技股份有限公司	6K-12K	6

同样的思路，最高薪水需要利用 find 查找"-"位置,然后截取 从"-" 到最后第二个位置的字符串。

```
=MID(C2,FIND("-",C2,1)+1,LEN(C2)-FIND("-",C2,1)-1)
```

这里，在新增数据列，平均薪水，来近似代表实际的准确薪资。平均薪水=(薪水下限+薪水上限)/2，即可得到每个岗位的平均薪水。

②真实城市截取

由于城市字段存储有的数据为"城市-区域"格式，例如"上海-徐汇区"，为了方便分析每个城市的数据，最后新增列"城市"，截取"-"前面的真实城市数据。

```
=IF(COUNTIF(G2,"*-*")=0,G2,LEFT(G2,FIND("-",G2,1)-1))
```

至此，所有数据清洗加工完毕，食材已经全部准备好，下面可以正式开始数据可视化的美食下锅烹饪了。

四、分析过程

分析过程有很多玩法，因为这里主要数据均是文本格式，数据又很简单，所以偏向汇总统计的计算。如果数值型的数据比较多，就会涉及到统计、比例等概念。如果有时间类数据，那么还会有趋势、变化的概念。

整体分析使用数据透视表完成，先利用数据透视表获得汇总型统计。

1、BI 工程师需求概况分析

计数项:jobName	列标签						
行标签	1-3年	1年以下	3-5年	5-10年	不限	无经验	总计
保定	1						1
北京	34		41	5	14	1	95
常熟			1				1
常州			1				1
成都	7		5	1	2		15
大连	1		5	3			9
东莞	2		1		1		4
方家山			1				1
佛山	1		2		1		4
福州	3		2		1		6
广州	17		12	1	7	1	38
贵阳	1				1		2
哈尔滨		1					1
杭州	10		5		3		18
合肥	2		1				3
呼和浩特	1						1
湖州			1				1
淮安			1				1
济南	3		1				4
嘉兴			1				1
金华			1				1
昆明	1						1
兰州			1				1
丽水			1				1
连云港			1				1
南昌	2						2
南京	4		2	1	1		8
南通			1				1
宁波			1				1
宁德			1				1
青岛	1		1		1		3
日本					2		2
厦门	4	1	4		1		10
上海	26		40	10	14	1	91
绍兴			1				1
深圳	20		24	7	12		63
沈阳			1		1		2
苏州	1		4		1		6
太仓市			1				1
太原	1		1				2
泰州			1				1
天津	3		4		2		9
温州			1				1
乌鲁木齐	1		1		1		3
无锡			1		1		2
武汉	6		5		1		12
西安			3				3
西宁	1						1
宿迁			1				1
徐州			1				1
盐城			1				1
扬州			1				1
张家港	1		1				2
长春	2						2
长沙	2		2		1		5
镇江	1		1				2
郑州	2		1		1	1	5
重庆	1		1				2
舟山			1				1
珠海	1						1
总计	164	2	194	28	70	4	462

这里我简单加了一下增材区分，增加数据大小的辨识度。（条件格式——色阶）

看来北上广深的 BI 工程师岗位远多于其他城市，成都杭州武汉梯队次之。1~3 年以及 3~5 年经验的缺口相当。

2、BI 工程薪资情况分析

各经验年龄的平均薪资状况，差距梯度还是很明显的。

目前市面上 BI 工程的薪资主要分许在 7~17K 左右区间。23~26K，应该是 5~10 年左右经验的岗位也相当。

3、薪资变化随着经验的增长，学历影响力的大小

整体来说，BI 工程师大专和本科的薪资差异并不是很大，3~5 年经验，本科稍占优势。到 5~10 年，基本拉平，也就是说学历因素影响比重更弱，这时候更看重经验。

其他的分析过程就不多做赘述了，主要是使用数据透视表和数据透视图进行多维度（城市，学历，工作经验）的分析，没有其他复杂的技巧。

关于数据透视图和数据透视表。选中所要分析的数据列，2013 版以上的 Excel 基本上都很智能的帮你推荐图标，生成透视界面，只要分清楚拖拽的字段事到列，到值还是到行即可。然后视情况多数据做一定筛选，因为数据清洗得不一定很彻底，我在制作的过程中就忽略了一些字段的空缺值，又回过头做了过滤。

最后

到此，一个简单的数据分析基本结束了。因为数据简单，并没有涉及过多的数据整合，表合并，专业数据统计回归等操作。

整个数据分析过程最费时间的数据清理，大约占据 70%，只要明确了目的，可视化分析师很简单的。

其次，也可以看到，用 Excel 做分析，更多的优势是数据的简单处理。随便过滤、查询、定位救你呢了解数据的概况。但在可视化方面比较鸡肋，行列值选择，以及复杂的图表制作都有一些难度，一句话总结 Excel 可视化要想做的好看还是要费点时间的。

所以我在分析的时候，基本上就是用 Excel 看看数据全貌，简单处理下。分析、可视化什么的还是会交给 BI。后面，我会再出一篇用 BI 制作的教程。