
河北工业大学

商务大数据分析报告

——数据分析师职位分析

作者： 牟江秀 学号： 151961

学院： 经济管理学院

系(专业)： 信息管理与信息系统

指导者： 李杰 教授

2019 年 1 月

目 录

目录

1	数据介绍.....	1
1.1	数据来源.....	1
1.2	Column 含义.....	1
2	数据预处理.....	1
2.1	隐藏数据.....	1
2.2	数据清洗.....	2
3	数据分析过程.....	10
4	总结展望.....	13

1 数据介绍

1.1 数据来源

本次报告所用的数据是来自“秦路”公众号用爬虫爬取了招聘网站上约 5000 条的数据分析师职位数据。数据真实来源于网络，属于网站方，请勿用于商业用途。

1.2 Column 含义

city：城市、companyFullName：公司全名、companyId：公司 ID、companyLabelList：公司介绍标签、companyShortName：公司简称、companySize：公司大小、businessZones：公司所在商区、firstType：职位所属一级类目、secondType：职业所属二级类目、education：教育要求、industryField：公司所属领域、positionId：职位 ID、positionAdvantage：职位福利、positionName：职位名称、positionLables：职位标签、salary：薪水、workYear：工作年限要求

city	companyF	companyL	companyS	companyS	businessZ	firstType	secondType	education	industryField	positionId	positionAdvantage	positionName	positionLabel	salary	workYear	
1	上海	纽海信息	8581	[技能培训]	1号店	2000人	[张江]	技术	数据开发	硕士	移动互联	2537336	知名平台	数据分析师[分析师]	7k-9k	应届毕业生
2	上海	上海点来	23177	[节日礼物]	点融网	500-2000	[五里桥]	技术	数据开发	本科	金融	2427485	挑战机会	数据分析师[分析师]	10k-15k	应届毕业生
3	上海	上海晶核	57561	[技能培训]	SPD	50-150人	[打浦桥]	设计	数据分析	本科	移动互联	2511252	时尚自由	数据分析师[分析师]	4k-6k	应届毕业生
4	上海	杭州数云	7502	[绩效奖金]	数云	150-500人	[龙华]	市场与销售	数据分析	本科	企业服务	2427530	五险一金	大数据业务[数据]	2k-6k	应届毕业生
5	上海	上海银基	130876	[年底双薪]	银基富力	15-50人	[上海影城]	技术	软件开发	本科	其他	2245819	在大牛手下实习/做项目	数据开发[数据]	2k-3k	应届毕业生
6	上海	上海青之	28095	[美女多]	青桐资本	50-150人		金融类	天使投资	本科	金融	2580543	留用机会/助理分析师[数据]	数据分析师[数据]	7k-14k	应届毕业生
7	上海	上海好伴	20092	[年底双薪]	足球魔方	150-500人	[龙华]	技术	后端开发	本科	移动互联	1449715	薪资福利	数据工程师[数据]	7k-14k	应届毕业生
8	上海	上海安顿	21863	[岗位晋升]	安顿信息	2000人	[黄兴公园]	产品/需求	产品设计	硕士	金融	2568628	健康体检	数据咨询[需求分析]	5k-7k	应届毕业生
9	上海	上海崇奇	121208	[扁平管理]	上海崇奇	15-50人	[上海电视	技术	后端开发	本科	移动互联	2416852	扁平管理	数据处理[数据]	4k-8k	应届毕业生
10	上海	五五海淘	58109	[股票期权]	55海淘	150-500人	[漕宝路]	技术	后端开发	本科	电子商务	1605795	向大牛学习	数据处理[数据]	2k-4k	应届毕业生
11	上海	莉莉丝科技	1938	[都是萌妹]	莉莉丝游戏	150-500人	[万源城]	技术	后端开发	本科	移动互联	2157863	带薪年假	大数据平台[平台]	大 5k-6k	应届毕业生
12	上海	我厨	51223	[节日礼物]	我厨	150-500人		设计	用户研究	本科	电子商务	2548985	可转正	数据分析师[数据]	2k-4k	应届毕业生
13	上海	上海闻途	36009	[技能培训]	途虎养车	500-2000	[梅陇]	技术	软件开发	本科	移动互联	2392425	有爱	数据分析师[数据]	2k-3k	应届毕业生
14	上海	上海麦子	63922	[工资自己]	麦子金服	500-2000	[肇嘉浜路]	市场与营销	市场营销	本科	移动互联	1243515	工资自己	数据专员[数据]	4k-6k	应届毕业生
15	上海	上海如旺	48294	[年底双薪]	旺旺集团	50-150人	[虹桥]	技术	后端开发	硕士	电子商务	2392372	旺旺集团	数据工程师[大数据]	4k-8k	应届毕业生
16	上海	上海点来	23177	[节日礼物]	点融网	500-2000	[五里桥]	技术	后端开发	本科	金融	2427555	机会挑战	数据研发[数据]	10k-15k	应届毕业生
17	上海	杭州数云	7502	[绩效奖金]	数云	150-500人		技术	后端开发	硕士	企业服务	2414480	大牛团队	大数据工程[大数据]	10k-15k	应届毕业生
18	上海	上海好伴	2002	[年底双薪]	足球魔方	150-500人	[龙华]	运营	运营	本科	移动互联	2320870	足球氛围	足球分析师[分析师]	6k-8k	应届毕业生

图 1.1 所处理数据

2 数据预处理

2.1 隐藏数据

(1) 在进行处理之前，我们需要将数据进行备份，防止后续处理需要原始数据时无法找到，备份原始数据，保存原始数据完整。

上海清源线	57577	「节日礼物	清源大数据
151961牟江秀+数据	151961牟江秀+数据	(2)	

图 2.1 备份原始数据

(2) companyId 和 positionId 是数据的唯一标示，类似该职位的身份证号，我们先隐藏。companyFullName 和 companyShortName 则重复，只需要留一个公司名称，companyFullName 依旧隐藏。尽量不删除数据，而是隐藏，保证原始数据的完整。

	A	D	E	F	G	H	I	J	K	M	N	O	P	Q		
1	city	companyL	companyS	companyS	businessZc	firstType	secondType	education	industryFie	positionAc	positionNa	positionLa	salary	workYear		
2	上海	「技能培训	1号店	2000人以	「张江」	技术	数据开发	硕士	移动互联网	知名平台	数据分析师	「分析师」	「7k-9k	应届毕业生		
3	上海	「节日礼物	点融网	500-2000	「五里桥」	技术	数据开发	本科	金融	挑战机会	「数据分析师	「分析师」	「10k-15k	应届毕业生		
4	上海	「技能培训	SPD	50-150人	「打浦桥」	设计	数据分析	本科	移动互联网	时间自由	「数据分析师	「分析师」	「4k-6k	应届毕业生		
5	上海	「绩效奖金	数云	150-500人	「龙华」	「上	市场与销售	数据分析师	本科	企业服务	「五险一金	「大数据业务	「商业」	「分	6k-8k	应届毕业生
6	上海	「年底双薪	银基富力	15-50人	「上海影城	技术	软件开发	本科	其他	在大牛下	「数据分析师	「分析师」	「2k-3k	应届毕业生		
7	上海	「美女多」	青桐资本	50-150人		金融类	天使投资	本科	金融	留用机会	「助理分析师	「实习」	「投	10k-15k	应届毕业生	
8	上海	「年底双薪	足球魔方	150-500人	「龙华」	「植	技术	后端开发	本科	移动互联网	薪资福利	「数据工程师	「数据」	「7k-14k	应届毕业生	
9	上海	「岗位晋升	安硕信息	2000人以	「黄兴公园	产品/需求	产品设计	硕士	金融	健康体检	数据咨询	「需求分析	「5k-7k	应届毕业生		
0	上海	「扁平管理	上海崇杏	15-50人	「上海电视	技术	后端开发	本科	移动互联网	扁平管理	「数据处理」	「后端开发	「4k-8k	应届毕业生		
1	上海	「股票期权	55海淘	150-500人	「漕宝路」	技术	后端开发	本科	电子商务	向大牛学	「数据处理」	「数据」	「2k-4k	应届毕业生		
2	上海	「都是萌妹	莉莉丝游戏	150-500人	「万源城」	技术	后端开发	本科	移动互联网	带薪年假	「大数据平台	「平台」	「大	5k-6k	应届毕业生	
3	上海	「节日礼物	我厨	「上海150-500人		设计	用户研究	本科	电子商务	「可转正	成」	「数据分析	「数据	分析	2k-4k	应届毕业生
4	上海	「技能培训	途虎养车	「500-2000	「梅陇」	「南	技术	软件开发	本科	移动互联网	「有爱的同	「BI数据分析	「数据	分析	2k-3k	应届毕业生
5	上海	「工资自己	麦子金服	500-2000	「肇嘉浜路	市场与销售	市场/营销	本科	移动互联网	工资自己	「数据专员	「数据」	「4k-6k	应届毕业生		
6	上海	「年底双薪	旺旺集团	「50-150人	「虹桥」	「古	技术	后端开发	硕士	电子商务	「旺旺集团	「数据工程师	「大数据」	「4k-8k	应届毕业生	
7	上海	「节日礼物	点融网	500-2000	「五里桥」	技术	后端开发	本科	金融	机会挑战	「数据研发」	「数据」	「10k-15k	应届毕业生		
8	上海	「绩效奖金	数云	150-500人		技术	后端开发	硕士	企业服务	「大牛团队	「大数据工程	「大数据」	「10k-15k	应届毕业生		
9	上海	「年底双薪	足球魔方	150-500人	「龙华」	「植	运营	运营	本科	移动互联网	「足球氛围	「足球分析师	「分析师」	「6k-8k	应届毕业生	
0	上海		Juntong Ci	少于15人	「淮海路」	金融	天使投资	本科	金融	资深团队	「分析师	「分析师」	「2k-3k	应届毕业生		
1	上海	「节日礼物	清源大数据	15-50人	「张江」	金融类	金融类	本科	数据服务	「互联网VC	「分析师	「分析师」	「2k-4k	应届毕业生		

图 2.2 隐藏后的数据

2.2 数据清洗

(1) 检查是否有缺失值

数据的缺失值很大程度上影响分析结果。引起缺失的原因很多，例如技术原因，爬虫没有完全抓去，例如本身的缺失，该岗位的 HR 没有填写。

companyLabelList、businessZones、positionLables 都有缺失，但不多。不影响实际分析。

A	D	E	F	G
city	companyL	companyS	companyS	businessZ
上海	['技能培训	1号店	2000人以	['张江']
上海	['节日礼物	点融网	500-2000	['五里桥',
上海	['技能培训	SPD	50-150人	['打浦桥']
上海	['绩效奖金	数云	150-500人	['龙华', '上
上海	['年底双薪	银基富力	15-50人	['上海影城
上海	['美女多',	青桐资本	50-150人	
上海	['年底双薪	足球魔方	150-500人	['龙华', '植
上海	['岗位晋升	安硕信息	2000人以	['黄兴公园
上海	['扁平管理	上海崇杏	15-50人	['上海电视
上海	['股票期权	55海淘	150-500人	['漕宝路',
上海	['都是萌妹	莉莉丝游	150-500人	['万源城',
上海	['节日礼物	我厨 (上	150-500人	
上海	['技能培训	途虎养车	500-2000	['梅陇', '南
上海	['工资自己	麦子金服	500-2000	['肇嘉浜路
上海	['年底双薪	旺旺集团	50-150人	['虹桥', '古
上海	['节日礼物	点融网	500-2000	['五里桥',
上海	['绩效奖金	数云	150-500人	
上海	['年底双薪	足球魔方	150-500人	['龙华', '植
上海		Juntong C	少于15人	['淮海路']
上海	['节日礼物	清源大数	15-50人	['张江']
151961牟江秀+数据 1			151961牟江秀+ ...	
				计数: 6877

图 2. 3 查看数据量

	companyL	companyS	companyS	business
	['技能培训	1号店	2000人以	['张江']
	['节日礼物	点融网	500-2000	['五里桥
	['技能培训	SPD	50-150人	['打浦桥
	['绩效奖金	数云	150-500人	['龙华', '上
	['年底双薪	银基富力	15-50人	['上海影
	['美女多',	青桐资本	50-150人	
	['年底双薪	足球魔方	150-500人	['龙华', '上
	['岗位晋升	安硕信息	2000人以	['黄兴公
	['扁平管理	上海崇杏	15-50人	['上海电
	['股票期权	55海淘	150-500人	['漕宝路
	['都是萌妹	莉莉丝游	150-500人	['万源城
	['节日礼物	我厨 (上	150-500人	
	['技能培训	途虎养车	500-2000	['梅陇', '上
	['工资自己	麦子金服	500-2000	['肇嘉浜
	['年底双薪	旺旺集团	50-150人	['虹桥', '上
	['节日礼物	点融网	500-2000	['五里桥
	['绩效奖金	数云	150-500人	
	['年底双薪	足球魔方	150-500人	['龙华', '上
		Juntong C	少于15人	['淮海路
	['节日礼物	清源大数	15-50人	['张江']
151961牟江秀+数据 1			151961牟江秀+ ...	
				计数: 6171

图 2. 4 查看数据的总数量

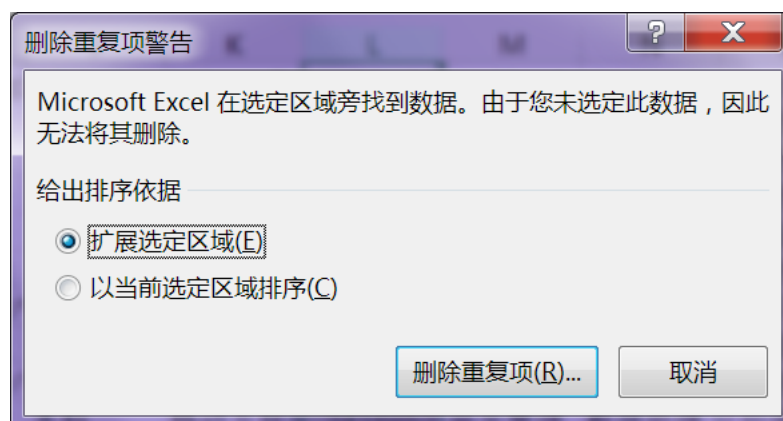


图 2. 6 清除重复数据



图 2. 7 清除重复数据

（4）检查是否结构标准

数据标准结构，就是将特殊结构的数据进行转换和规整。

1) 在我们的表格中。薪水的话用了几 k 表示，属于文本，并不能直接用于计算，是一个范围，按照最高薪水和最低薪水拆成两列。

首先，先建立两列新列 bottoms salary 和 tops salary

Bottoms salary 列输入公式=LEFT(P2,FIND("k",P2,1)-1) 先找出 k 字，取 k 字符前面的数字；tops salary 因为第二个薪水位置不固定，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。

获取和转换		连接	
<div>=LEFT(P2,FIND("k",P2,1)-1)</div>			
O	P	Q	R
Na positionLa	salary	workYear	bottomsal:top
所['分析师,'	7k-9k	应届毕业生	79
所['分析师,'	10k-15k	应届毕业生	1015

图 2. 8 处理 salary

<div> <div></div> <div>=MID(P2,FIND("-",P2,1)+1,LEN(P2)-FIND("-",P2,1)-1)</div> </div>				
O	P	Q	R	S
positionLa	salary	workYear	bottomsal	topsalary
班[分析师,'	7k-9k	应届毕业生	7	9
班[分析师,'	10k-15k	应届毕业生	10	15

图 2. 9 处理 topsalary

升序(S)

降序(Q)

按颜色排序(I)

从"bottomsalary"中清除筛选(C)

按颜色筛选(I)

文本筛选(E)

搜索

☒ 60

☒ 7

☒ 8

☒ 9

☒ #VALUE!

确定

取消

图 2. 10 清除重复数据

这其中存在 value 值，经检查是大写 K，因为 find 对大小写敏感，此时用 search 函数，或者将 K 替换成 k 都能解决。

=MID(P2,SEARCH("-",P2,1)+1,LEN(P2)-SEARCH("-",P2,1)-1)

O	P	Q	R	S	T
position	salary	workYear	bottoms	topsalar	
分析师	7k-9k	应届毕业生7		9	
分析师	10k-15k	应届毕业生10		15	
分析师	4k-6k	应届毕业生4		6	
商业	6k-8k	应届毕业生6		8	

图 2. 11 处理大写 k

=LEFT(P2,SEARCH("k",P2,1)-1)

O	P	Q	R	
position	salary	workYear	bottoms	top
分析师	7k-9k	应届毕业生7	7	9
分析师	10k-15k	应届毕业生10	10	15

图 2. 12 处理大写 k

另外还有一个错误是很多 HR 将工资写成 5K 以上，这样就无法计算 topSalary。为了计算方便，将 topSalary 等于 bottomSalary。

	R	S
year	bottoms	topsalar
15	15	15
8	8	8
15	15	15
8	8	8
10	10	10
15	15	15
20	20	20
15	15	15
30	30	30

图 2. 13 处理无上限的值

=MID(P2,SEARCH("-",P2,1)+1,LEN(P2)-SEARCH("-",P2,1)-1)

O	P	Q	R	S
position	salary	workYear	bottom	topsalary
['数据','安	15k以上	1-3年	15	#VALUE!
['技术支持	8k以上	1-3年	8	#VALUE!
['分析师','	15k以上	1-3年	15	#VALUE!
['大数据','	8k以上	1-3年	8	#VALUE!
['大数据','	10k以上	3-5年	10	#VALUE!
['平台','数	15k以上	3-5年	15	#VALUE!
['大数据','	20k以上	3-5年	20	#VALUE!
['大数据','	15k以上	3-5年	15	#VALUE!
['分析师']	30k以上	5-10年	30	#VALUE!
['分析师','	9k以上	不限	9	#VALUE!
['分析师','	15k以上	不限	15	#VALUE!
['分析师']	25k以上	3-5年	25	#VALUE!

图 2. 14 处理无上限的值

2）表格中，companyLableList 就是以数组形式保存，businessZones、positionAdvantage 和 positionLables 也是同样问题。

companyLabelList 是公司标签，诸如技能培训、五险一金直接用分列即可。符号用搜索替换法删除即可。

先用数据选项卡下方的分列进行分列，再利用搜索替换法删除。

	D	E	F	G	H
	companyLabelList	companyL	companyL	companyLabelList4	
1	技能培训	节日礼物	带薪年假	岗位晋升	
2	节日礼物	带薪年假	岗位晋升	扁平管理	
3	技能培训	绩效奖金	岗位晋升	管理规范	
4	绩效奖金	股票期权	五险一金	通讯津贴	
5	年底双薪	通讯津贴	定期体检	绩效奖金	
6	美女多	出国旅游	不打卡	带薪年假	
7	年底双薪	股票期权	扁平管理	领导好	
8	岗位晋升	顶尖团队	福利优厚	股票期权	
9	扁平管理	弹性工作	岗位晋升	领导好	
10	股票期权	带薪年假	绩效奖金	岗位晋升	
11	都是萌妹子	项目奖金	零食无限	国内外旅游	
12	节日礼物	技能培训	带薪年假	岗位晋升	
13	技能培训	节日礼物	免费班车	绩效奖金	
14	工资自己定	期权任性	金融狂人	半年晋升	
15	年底双薪	节日礼物	技能培训	免费班车	
16	节日礼物	带薪年假	岗位晋升	扁平管理	

图 2. 15 分列操作结果

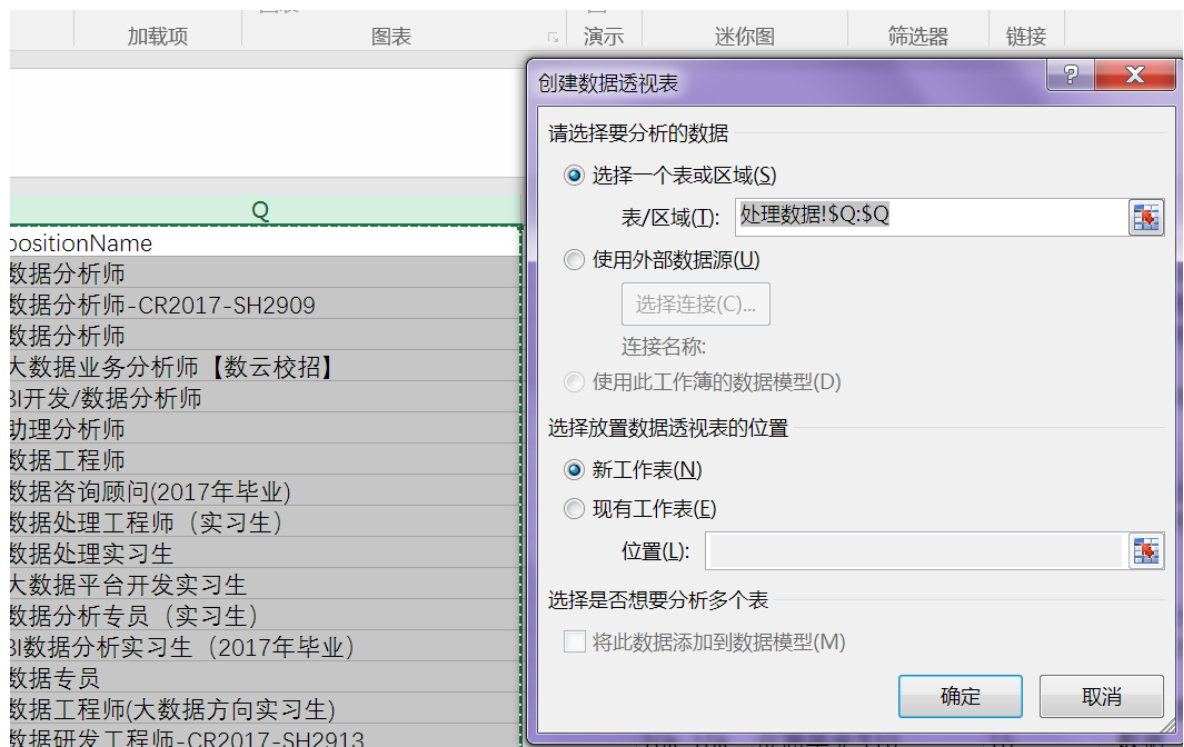


图 2. 17 positionlabel 分列操作

3 数据分析过程

通过数据透视表汇总分析显示先利用数据透视表获得汇总型统计。

本次分析过程主要是看各个城市的薪资水平

通过数据透视表可以看出北京、成都、广州、杭州、上海的薪资最多是在 30k，长沙、南京是 20k，厦门是 15k，其他城市的所招的数据分析岗位数量不多。

行标签	计数项:bottomsalary
北京	2351
30	402
20	356
25	249
15	235
40	153
10	110

图 3. 1 北京岗位的薪资数量排名

	30	1
成都	135	
	30	18
	20	18
	25	15
	6	13
	10	12
	-	-

图 3.2 成都岗位的薪资数量排名

	20	1
广州	335	
	30	60
	20	51
	25	41

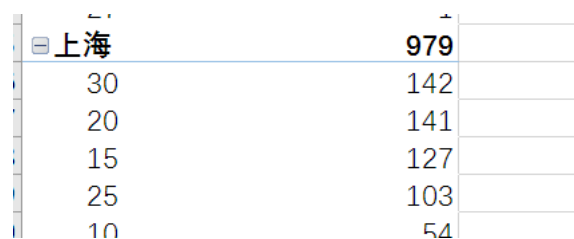
图 3.3 广州岗位的薪资数量排名

	28	1
杭州	407	
	30	69
	20	65
	25	47
	15	47
	40	19
	12	16
	10	16

图 3.4 杭州岗位的薪资数量排名

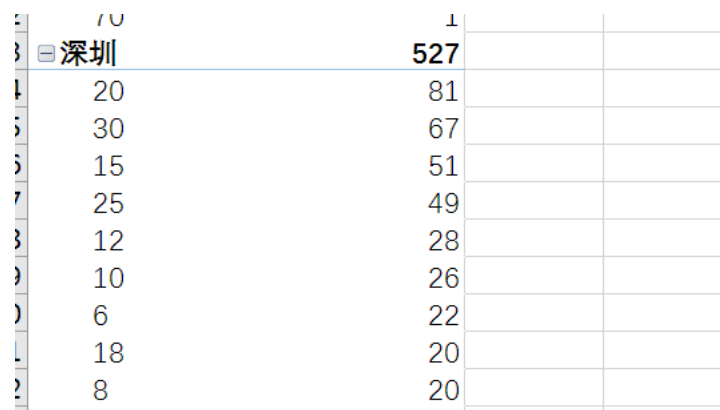
	28	1
南京	83	
	20	17
	15	10
	30	10
	25	9
	12	3
	24	3
	-	-

图 3.5 南京岗位的薪资数量排名



上海	979
30	142
20	141
15	127
25	103
10	54

图 3.6 上海岗位的薪资数量排名



深圳	527
20	81
30	67
15	51
25	49
12	28
10	26
6	22
18	20
8	20

图 3.7 深圳岗位的薪资数量排名

通过数据透视表，我们可以看出本次数据中的所有城市中的数据分析岗位的数量，北京的岗位数量是 2351，远超前第二名的上海，是上海岗位数量的二倍。处于第二名的上海岗位数量远超前第三名的数量，是深圳岗位数量二倍。我们可以看出北京的数据分析相关行业发展在中国最快的，对数据分析岗位人才的需求量最大，而且从上面的分析也可以得知，北京的薪资也是第一档位的。

行标签	计数项:city
北京	2351
上海	979
深圳	527
杭州	407
广州	335
成都	135
南京	83
武汉	69
西安	38
苏州	37
厦门	30
长沙	25
天津	20
总计	5036

图 3.8 不同地方岗位数量排名

4 总结展望

因为各方面原因，对目前的报告还不够满意，预想中会得到的好结果没有得到。对数据分析的方法掌握不够熟练，应该深入学习这些方法，综合自己所掌握的方法再进行分析建模，进行分析。

通过这次报告，我们也意识到数据分析往往不是一蹴而就的事情，在之后我还会继续不断改进修正完善这份报告，真真正正的做一份好作品。