

数据分析常用SQL语句练习

环境

系统：macOS 10.12

数据库：mysql5.7.2

GUI:Sequel Pro1.1

数据来源：秦路

从入门到熟练

本次练习数据是直接使用秦路文章中使用的现成的数据作为练习。

数据是用爬虫爬取招聘网站上约5000条的数据分析师职位数据。为了练习，数据中北京数据多复制了一份。

具体数据包括字段：

- 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

在数据库中长这个样子：

![数据样子]

基础

- 我想看城市为深圳的职位数据：

```
select * from data where city = '深圳'
```

```
1 select * from data where city = '深圳'
```

city	companyFullName	companyId	companyLabelList	companyShortName	companySize	businessZones	firstType	secondType	education	industryField	positionName	positionAdv
深圳	深圳爱拼信息科技...	385	['发展迅猛', '股票...	iPIN.com	50-150人	['科技园', '南...	技术	数据开发	本科	数据服务	24412...	风口行业, 牙...
深圳	深圳前海用友友力合...	44982	['节日礼物', '带薪...	友金所	500-2000人	['科技园', '南...	金融	风控	本科	金融	22100...	发展前景好+
深圳	深圳市谷米科技有...	16797	['年底双薪', '节日...	酷米客公交	150-500人	['科技园', '深...	技术	高端技术...	博士	移动互联网...	12593...	朝阳产业, 良...
深圳	深圳房讯通信息技...	33807	['年底双薪', '节日...	房讯通	150-500人	['科技园', '科...	运营	运营	不限	移动互联网...	689580	自由 欢乐 创...
深圳	深圳爱拼信息科技...	385	['发展迅猛', '股票...	iPIN.com	50-150人	['科技园', '南...	技术	后端开发	本科	数据服务	24412...	风口行业, 牙...
深圳	深圳房讯通信息技...	33807	['年底双薪', '节日...	房讯通	150-500人	['科技园', '南...	设计	用户研究	大专	移动互联网...	21993...	自由、福利、
深圳	联保信息科技集团...	46861	['年底双薪', '节日...	联保	50-150人	['科技园', '南...	技术	后端开发	本科	移动互联网...	21181...	项目好 金融...
深圳	环富通科技 (深圳)...	58742	['技能培训', '带薪...	环富通	50-150人	['上步', '华强...	市场与...	投资	大专	移动互联网...	656031	五天八小时工...
深圳	北京思恩客广告有...	56521	['技能培训', '节日...	蓝色光标SNK	500-2000人	['科技园', '南...	设计	用户研究	本科	移动互联网...	22464...	发展空间大 行...
深圳	深圳市柚子科技有...	143609		柚子科技	15-50人	['美丽365花园']	市场与...	数据分析	大专	电子商务	24130...	工作突出可任...

No errors; 527 rows affected, taking 2.7 ms

- 我想看城市为深圳且职位名称是数据分析师的数据：数据分析方向有好多叫法。

```
select * from data where city = '深圳' and positionName = '数据分析师'
```

```
1 select * from data
2 where city = '深圳' and positionName = '数据分析师'
```

city	positionName	companyFullName	companyId	companyLabelList	companyShortName	companySize	businessZones	firstType	secondType	education	industryField	positionAdv
深圳	数据分析师	深圳新感易搜网络...	26399	['带薪年假', '绩效...	云图TV	50-150人	['南头']	设计	数据分析	不限	移动互联网...	2547...
深圳	数据分析师	深圳市和讯华谷信...	917	['五险一金', '年底...	极光	150-500人	['南新路口', '...	技术	后端开发	本科	移动互联网...	9175...
深圳	数据分析师	深圳中兴飞贷金融...	111065	['领导高富帅', '女...	飞贷金融	500-2000人	['科技园', '科...	设计	用户研究	本科	移动互联网...	2156...
深圳	数据分析师	深圳市金斧子网络...	2044	['股票期权', '绩效...	金斧子	150-500人	['科技园', '科...	技术	后端开发	本科	电子商务...	1468...
深圳	数据分析师	广东俊特团贷网络...	25576	['节日礼物', '带薪...	团贷网	2000人以上	['科技园', '南...	运营	运营	大专	移动互联网...	2336...
深圳	数据分析师	武汉奇米网络科技...	2374	['节日礼物', '餐补...	卷皮	500-2000人	['南头', '科技...	设计	用户研究	不限	电子商务	1888...
深圳	数据分析师	深圳前海信美分期...	144389		信美分期	50-150人		技术	DBA	本科	金融,移动...	2506...
深圳	数据分析师	深圳市易讯天空网...	9275	['股票期权', '绩效...	500.com集团	150-500人		技术	后端开发	本科	移动互联网...	1536...
深圳	数据分析师	深圳市海德在线有...	29355	['节日礼物', '股票...	海德在线	150-500人	['南山医院', '...	技术	DBA	大专	移动互联网...	2410...
深圳	数据分析师	深圳市一面网络技...	35565	['技能培训', '海归...	一面网络	15-50人	['科技园', '南...	技术	数据开发	本科	移动互联网...	2535...

No errors; 52 rows affected, taking 6.9 ms

- 我想看深圳的数据分析师或者北京的产品经理数据：看看不同城市不同名称的职位量。

```
select * from data
where (city = '深圳' and positionName = '数据分析师') or (city = '北京' and
positionName = '数据产品经理')
```

1

2

3

select * from data

where (city = '深圳' and positionName = '数据分析师') or (city = '北京' and positionName = '数据产品经理')

✖

Query Favorites

▼

Query History

▼

Run Current

↘

city	positionName	companyFullName	companyId	companyLabelList	companyShortName	companySize	businessZones	firstType	secondType	education	industryField	position
深圳	数据分析师	深圳爱飞扬信息科...	138931	爱飞扬	150-500人	['车公庙', '香...	运营	运营	不限	移动互联网...		21296
深圳	数据分析师	武汉掌游科技有限...	155801	武汉掌游科技	15-50人	['白石洲']	产品/...	数据分析	大专	游戏		25440
深圳	数据分析师	南京烽火星空通信...	48873	['年底双薪', '股票...	南京烽火星空	500-2000人	技术	数据开发	本科	移动互联网...		25355
深圳	数据分析师	棕榈园林股份有限...	155801	棕榈幸福家社区服...	2000人以上	['科技园']	技术	数据开发	硕士	O2O		23840
北京	数据产品经理	巧达科技 (北京) ...	27129	['工作居住证', '股...	巧达数据	50-150人	产品	产品经理	硕士	企业服务,...		25505
北京	数据产品经理	同方知网 (北京) ...	23727	['美女多多', '带薪...	知网	2000人以上	['西三旗', '清...	产品	产品经理/...	本科	移动互联网...	21179
北京	数据产品经理	北京小桔科技有限...	2474	['弹性工作', '节日...	滴滴出行	500-2000人	产品	产品经理/...	本科	移动互联网		25036
北京	数据产品经理	鱼行天下 (北京) ...	1866	['五险一金', '专项...	大鱼自助游	50-150人	['西坝河', '三...	产品/...	产品设计/...	本科	旅游	25674
北京	数据产品经理	医众互联 (北京) ...	98468	['节日礼物', '专项...	医众互联	15-50人	['CBD', '国贸', '...	产品	产品经理	本科	移动互联网...	21509
北京	数据产品经理	北京世诚优聘科技...	104048	['技能培训', '股票...	斗米兼职	500-2000人	['上地', '马连...	产品	产品经理/...	本科	移动互联网...	21473

🏠

✖

📄

No errors; 258 rows affected, taking 3.8 ms

- 我想看深圳成都北京上海这几个城市的数据：多看几个城市的职位数据。

```

select * from data
where city in ('深圳','成都','北京','上海')

```

- 模糊查找，用like搭配%。

```

select * from data where positionName like '%数据分析%'

```

```

1 select * from data
2 where positionName like '%数据分析%'
3

```

city	positionName	companyFullName	companyId	companyLabelList	company...	companySize	businessZones	firstType	secondType	education	industryField	positi
上海	数据分析师	纽海信息技术(上...	8581	['技能培训', '节日...	1号店	2000人以上	['张江']	技术	数据开发	硕士	移动互联网	253.
上海	数据分析师-CR2017-...	上海点荣金融信息...	23177	['节日礼物', '带薪...	点融网	500-2000人	['五里桥', '打...	技术	数据开发	本科	金融	242.
上海	数据分析师	上海晶樵网络信息...	57561	['技能培训', '绩效...	SPD	50-150人	['打浦桥']	设计	数据分析	本科	移动互联网	251.
上海	BI开发/数据分析师	上海银基富力信息...	130876	['年底双薪', '通讯...	银基富力	15-50人	['上海影城', '...	技术	软件开发	本科	其他	224.
上海	数据分析专员 (实习生)	我厨 (上海) 科技...	51223	['节日礼物', '技能...	我厨 (上...	150-500人		设计	用户研究	本科	电子商务,...	254.
上海	BI数据分析实习生 (20...	上海闻途信息技术...	36009	['技能培训', '节日...	途虎养车网	500-2000人	['梅陇', '南方...	技术	软件开发	本科	移动互联网	239.
上海	商业数据分析	北京数字新思科技...	73539	['年底双薪', '带薪...	数字新思	15-50人	['田林', '徐家汇']	市场与...	数据分析	本科	数据服务	251.
上海	数据分析专员	上海莆田贸易有限...	146592	['带薪年假', '技能...	莆田网Fi...	150-500人	['龙华']	市场与...	供应链	本科	电子商务	238.
上海	商业数据分析	伽蓝 (集团) 股份...	7069	['绩效奖金', '年底...	伽蓝	2000人以上		产品/...	数据分析	本科	电子商务	253.
上海	商业数据分析 - 知识产...	路通世纪(中国)科...	25499	['五险一金', '免费...	汤森路透	2000人以上	['东外滩', '陆...	市场与...	市场/营销	本科	数据服务	158.

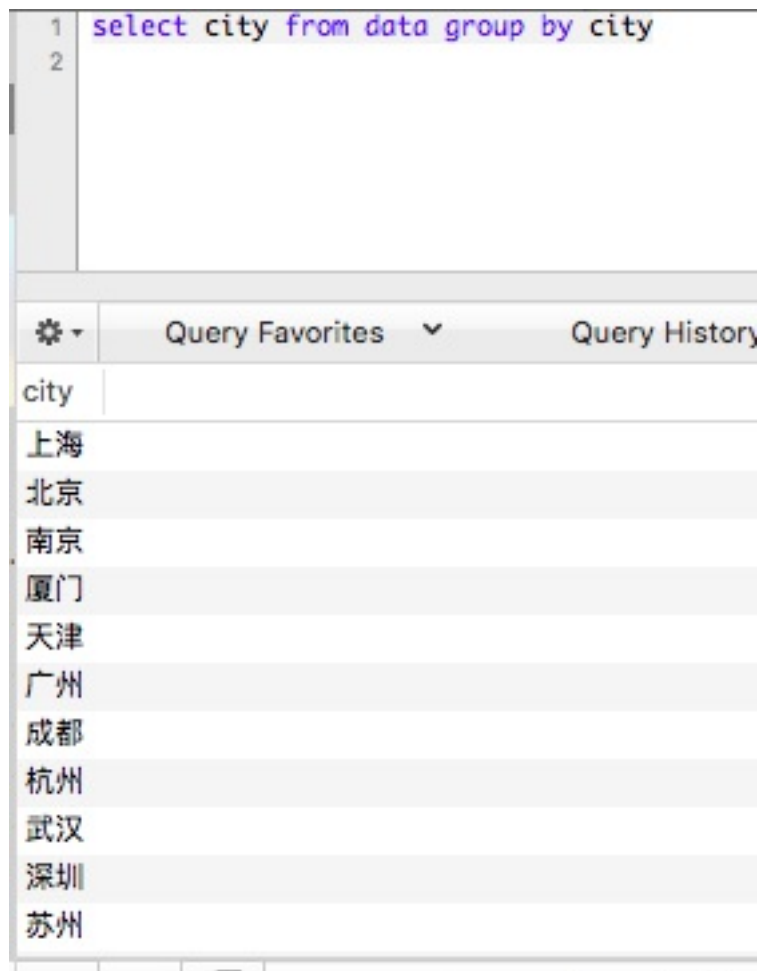
No errors; 2100 rows affected, taking 1.2 ms

- 看看都有哪些城市有职位需求：

```

select city from data group by city

```



- 我想看每个城市的职位数量：使用count计算数量，除了count，还有max，min，sum，avg等，也叫聚合函数。通过distinct对职位ID进行去重，避免重复统计。实际使用中，活跃用户数、文章UV，都可用distinct计算。

```
select city,count(distinct positionId) from data group by city
```

2

- 用group by添加多个字段，将以多维形式进行数据聚合。

```
select city,workYear,count(distinct positionId) from data group by city,workYear
```

1

2

select city,workYear,count(distinct positionId) from data group by city,workYear|

⚙️

Query Favorites

▼

Query History

▼

city	workYear	count(distinct positionId)
上海	1-3年	322
上海	10年以上	4
上海	1年以下	16
上海	3-5年	340
上海	5-10年	138
上海	不限	126
上海	应届毕业生	33
北京	1-3年	745
北京	10年以上	8
北京	1年以下	26
北京	3-5年	900

🖨️

⚙️

📄

No errors; 75 rows affected, taking 44.6 ms

- 我想知道不同城市，电商领域岗位的职业有多少，占比是多少：使用count结合if计算电商的职业数，职业数最后除以城市职业总数就可以得出占比，所以这里需要求出分母和分子两个数。第一个count计算职业总数。第二个count计算电商职业数，if函数中间字段代表为true时返回的值，为避免重复数据，这里根据数据源的特征使用唯一标识positionId来进行计数；if第三个参数写成null是为避免count对其计数。

```
select city,count(distinct positionId),count(if(industryField like '%电子商务%',positionId,null)) from data
group by city
```


1	select city,count(distinct positionId),count(if(industryField like '%电子商务%',positionId,null)) from data
2	group by city
3	

city	count(distinct positionId)	count(if(industryField like '%电子商务%',positionId,null))
上海	979	166
北京	2347	538
南京	83	24
厦门	30	3
天津	20	4
广州	335	71
成都	135	37
杭州	406	63
武汉	69	17

No errors; 13 rows affected, taking 25.6 ms

- 我想知道各个城市数据分析师岗位数量在500以上的城市有哪些：
 - 第一种：使用having对聚合后数据进行过滤
 - 第二种：利用嵌套子查询，as表示命名

```
select city,count(distinct positionId) from data
group by city having count(distinct positionId)>=500
```

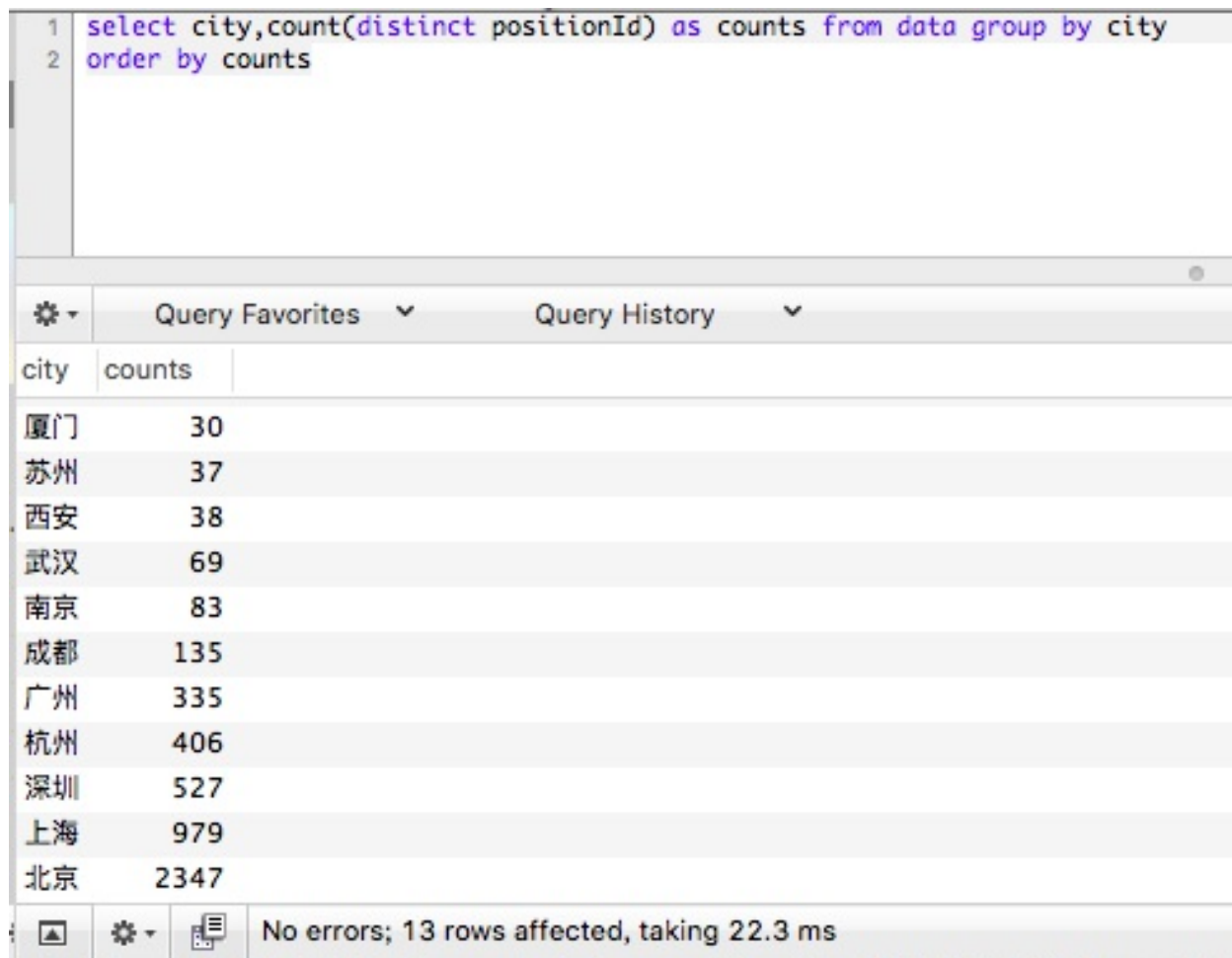
1	select city,count(distinct positionId)from data
2	group by city having count(distinct positionId)>=500
3	

city	count(distinct positionId)
上海	979
北京	2347
深圳	527

```
select * from(select city,count(distinct positionId) as counts from data group by city) as
t1
where counts >=500
```

- 使用order by让数据按顺序显示:降序在最后加desc

```
select city,count(distinct positionId) as counts from data group by city
order by counts
```



city	counts
厦门	30
苏州	37
西安	38
武汉	69
南京	83
成都	135
广州	335
杭州	406
深圳	527
上海	979
北京	2347

- 我想得到每个岗位工资最高和最低，原始数据是在salary字段里面，这里做一个数据清洗：
locate函数查找第一个k所在位置，left函数获取k前面位置的数字，得到下限；
substr函数第二个参数表示从哪里开始截取，第三个参数表示截取长度；locate函数查找-的位置+1表示横杠后面的位置作为第二个参数，length函数减去locate函数表示第三个参数。

```
select left(salary,locate('k',salary)-1) as bottomSalary,
substr(salary,locate('-',salary)+1,length(salary)-locate('-',salary)-1) as topSalary,
salary from data
where salary not like '%以上%'
```



```

1 select left(salary,locate('K',salary)-1) as bottomSalary,
2         substr(salary,locate('-',salary)+1,length(salary)-locate('-',salary)-1) as topSalary,
3         salary from data
4 where salary not like '%以上%'

```

bottomSalary	topSalary	salary
7	9	7k-9k
10	15	10k-15k
4	6	4k-6k
6	8	6k-8k
2	3	2k-3k
10	15	10k-15k
7	14	7k-14k
5	7	5k-7k

- 我想得到不同城市不同工作年限的平均薪资，当然这里薪资是一个范围，不能表示真实情况，仅作为练习。

```

select city,workYear,avg((bottomSalary+topSalary)/2) as avgSalary from(
select left(salary,locate('K',salary)-1) as bottomSalary,
substr(salary,locate('-',salary)+1,length(salary)-locate('-',salary)-1) as topSalary,
city,positionId,workYear
from data
where salary not like '%以上%') as t1
group by city,workYear
order by city,avgSalary

```

```

1 select city,workYear,avg((bottomSalary+topSalary)/2) as avgSalary from(
2     select left(salary,locate('K',salary)-1) as bottomSalary,
3         substr(salary,locate('-',salary)+1,length(salary)-locate('-',salary)-1) as topSalary,
4         city,positionId,workYear
5     from data
6     where salary not like '%以上%') as t1
7 group by city,workYear
8 order by city,avgSalary

```

city	workYear	avgSalary
上海	应届毕业生	7.106060606060606
上海	1年以下	9.34375
上海	不限	10.39516129032258
上海	1-3年	13.70440251572327
上海	3-5年	19.96439169139466
上海	5-10年	28.32846715328467
上海	10年以上	35
北京	应届毕业生	6.864077669902913
北京	1年以下	7.697916666666667
北京	不限	12.3712
北京	1-3年	15.74564723694121

No errors; 75 rows affected, taking 17.8 ms

从熟练到掌握

到leetcode.com上刷题。