SPL Base

集算器教案

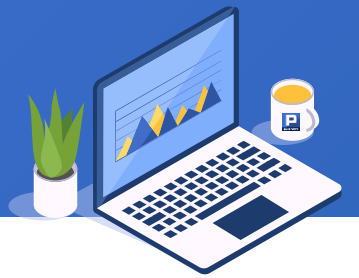
对齐分组

CONTENTS

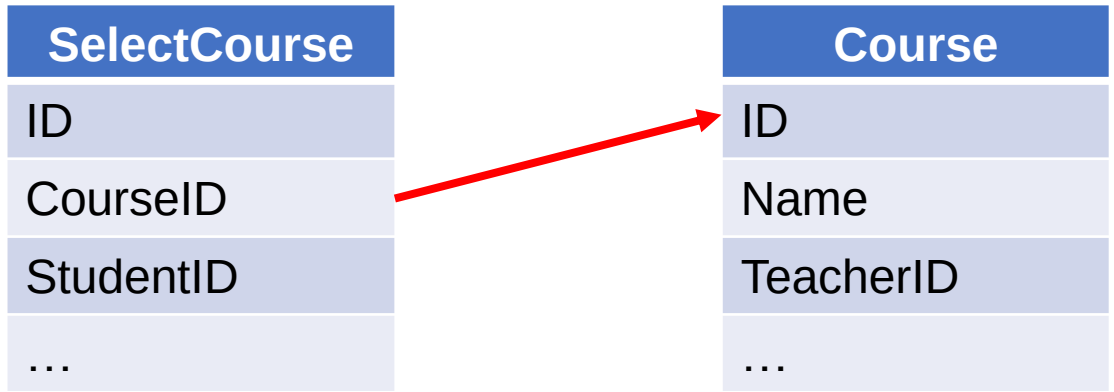
CONTENTS

01 普通对齐分组

1. 每组选择一个匹配成员



有课程表和选课表，按课程表顺序查询有哪些课没有学生选修。

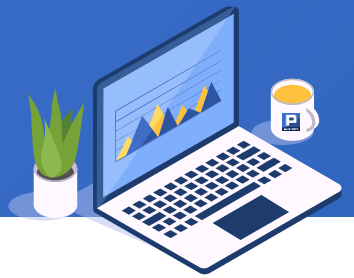


1. 每组选择一个匹配成员

SPL如下，其中用到了align(A:x, y)函数来实现对齐分组：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from SelectCouse")	/查询选修表
3	=A1.query("select * from Course")	/查询课程表
4	=A2.align(A3:ID,CourseID)	/选修表按照课程表的序号对齐，每组选择一个匹配成员
5	=A3(A4.pos@a(null))	/在课程表中选出没有选择（值为null）的课程信息

1. 每组选择一个匹配成员



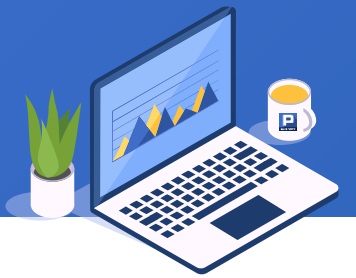
A4结果：

Members
(null)
[13,2,7]
[7,3,41]
[45,4,28]
[3,5,52]
[1,6,59]
[10,7,13]
[8,8,49]
[6,9,57]
(null)

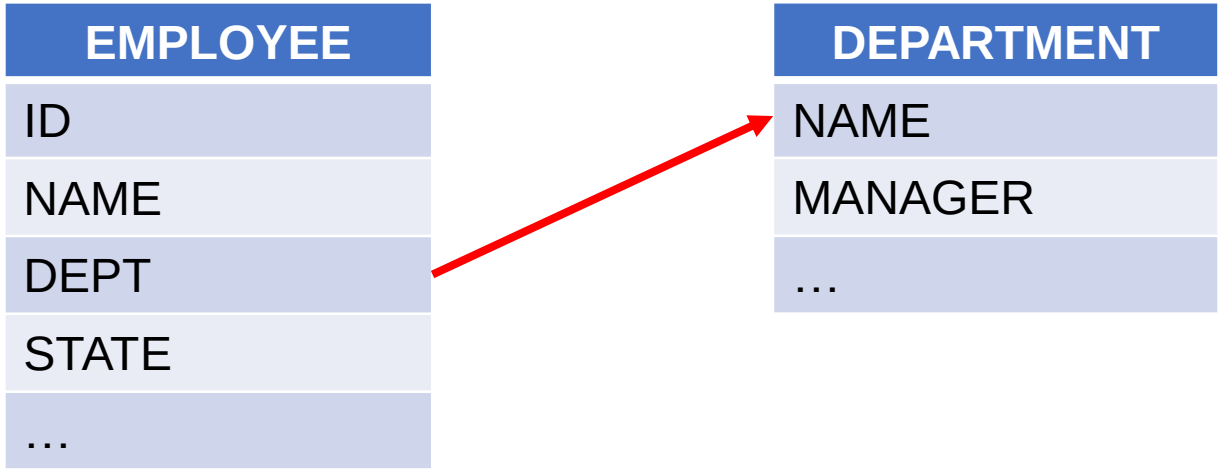
A5结果：

ID	NAME	TeacherID
1	Environmental protection and sustainable development	5
10	Music appreciation	18

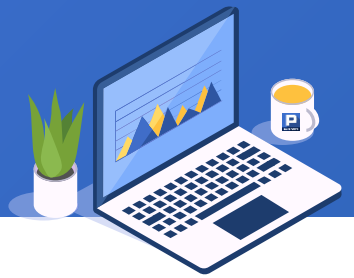
2. 每组选择所有匹配成员



有员工表和部门表，按部门表的部门顺序统计各部门人数。



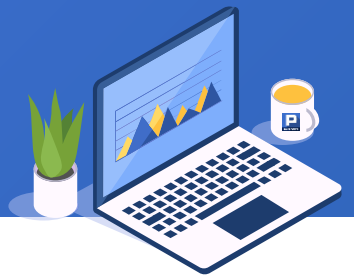
2. 每组选择所有匹配成员



SPL如下，其中用到了align@a选项：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from EMPLOYEE")	/查询员工表
3	=A1.query("select * from DEPARTMENT")	/查询部门表
4	=A2.align@a(A3:ID, DEPARTMENT)	/员工表按部门对位分组，@a选项每组返回所有匹配成员
5	=A4.new(DEPT, ~.count():COUNT)	/统计各部门的员工数量

2. 每组选择所有匹配成员



A4结果：

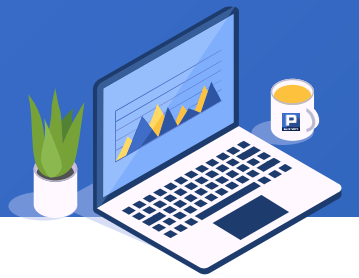
Members
[[18,Jonathan,Admin,...], [20,Alexis, Admin,...], …]
[[1,Rebecca,R&D,...],[5,Ashley,R&D,...],...]
[[3,Rachel,Sales,...],[6,Matthew,Sales,...],...]
…

ID	NAME	DEPT	STATE
1	Rebecca	R&D	California
5	Ashley	R&D	Texas
10	Ryan	R&D	Pennsylvania
…	…	…	…

A5结果：

DEPT	COUNT
Admin	4
R&D	29
Sales	187
…	…

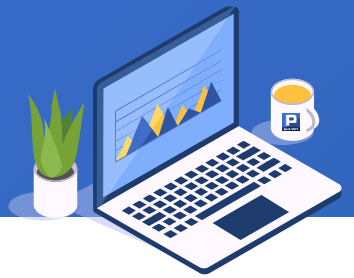
3. 不匹配成员存放到新组



以员工表为例，统计[California, Texas, New York, Florida]各州的平均工资。其他地区的员工存放到新组统计。

ID	NAME	STATE	SALARY
1	Rebecca	California	7000
2	Ashley	New York	11000
3	Rachel	New Mexico	9000
4	Emily	Texas	7000
5	Ashley	Texas	16000
…	…	…	…

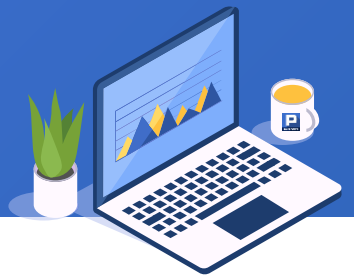
3. 不匹配成员存放到新组



SPL如下，其中用到了align@an选项：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from EMPLOYEE")	/查询雇员表
3	[California,Texas,New York,Florida]	/创建地区序列
4	=A2.align@an(A3,STATE)	/雇员表按地区对位分组，@a选项每组返回所有匹配成员，@n选项不匹配成员存放到新组。
5	=A4.new(if (#>A3.len(),"Other",STATE):STATE,~.avg(SALARY):AvgSalary)	/统计每组的平均工资，产生新序表。最后一组的地区更名为Other，否则会显示为第一条记录的地区。

3. 不匹配成员存放到新组



A4结果：

Members
[[1,Rebecca,California,...], [6,Matthew,California,...], …]
[[4,Emily,Texas,...],[5,Ashley,Texas,...],...]
[[2,Ashley,New York,...],[12,Jessica,New York,...],...]
[[13,Daniel, Florida,...],[14,Alyssa,Florida,...],...]
[[3,Rachel,New Mexico,...],[7,Alexis,Illinois,...],...]

ID	NAME	STATE	SALARY
3	Rachel	New Mexico	9000
7	Alexis	Illinois	9000
10	Ryan	Pennsylvania	13000
19	Samantha	Pennsylvania	10000
…	…	…	…

A5结果：

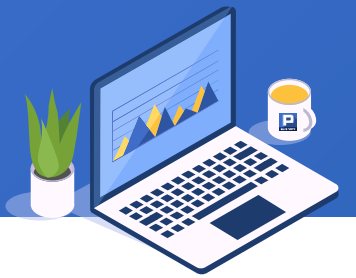
STATE	SALARY
California	7700.0
Texas	7592.59
New York	7677.77
Florida	7145.16
Other	7308.1

CONTENTS

02

序号对齐分组

1. 每组选择一个匹配成员

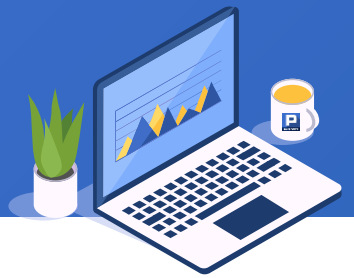


有销售表和客户表，查询2014年没有销售记录的客户。

Sales
ID
CustomerID
OrderDate
…

Customer
ID
Name
State
…

1. 每组选择一个匹配成员



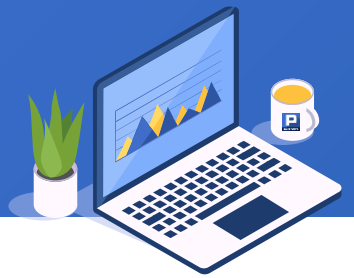
SPL如下，其中用到了align(n, y)函数来实现对齐分组：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from Sales")	/查询销售表
3	=A1.query("select * from Customer")	/查询客户表
4	=A3.(ID)	/从客户表中选出客户序号
5	=A2.align(A4.len(), A4.pos(CustomerID))	/销售表按照客户序号对位分组
6	=A3(A5.pos@a(null))	/在客户表中选出没有销售记录（值为null）的客户信息

A6结果：

ID	Name	State	…
ALFKI	CMA-CGM	Texas	…
CENTC	Nedlloyd	Florida	…

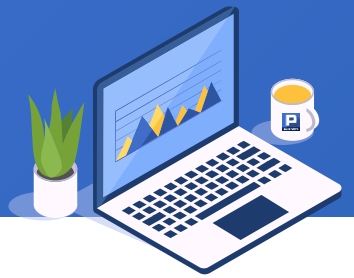
2. 每组选择所有匹配成员



下面是订单表，顺序列出2013年每月的订单数。

ID	CustomerID	OrderDate	Amount
10248	VINET	2012/07/04	428.0
10249	TOMSP	2012/07/05	1842.0
10250	HANAR	2012/07/08	1523.5
10251	VICTE	2012/07/08	624.95
10252	SUPRD	2012/07/09	3559.5
…	…	…	…

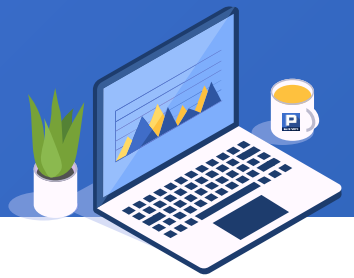
2. 每组选择所有匹配成员



SPL如下，使用align(n, y)函数的@a选项，每组选择所有匹配成员：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from Orders where year(OrderDate)=2013")	/查询2013年的订单
3	=A2.align@a(12,month(OrderDate))	/按订单月份对位分为12组，@a选项每组选择所有匹配成员
4	=A3.new(#:Month,~.count():OrderCount)	/创建序表，统计每月订单数

2. 每组选择所有匹配成员



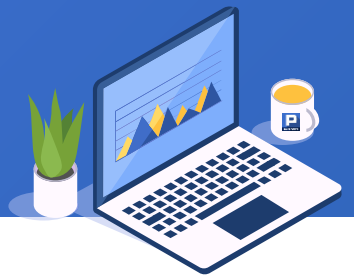
A3结果：

Members
[[10400,EASTC,1],[10401,RATTC,1],…]
[[10433,PRINI,2],[10434,FOLCO,2],…]
[[10462,CONSH,3],[10463,SUPRD,3],…]
[[10492,BOTTM,4],[10493,LAMAI,4],…]
[[10523,SEVES,5],[10524,BERGS,5],…]
[[10555,SAVEA,6],[10556,SIMOB,6],…]
[[10585,WELLI,7],[10586,REGGC,7],…]
[[10618,MEREP,8],[10619,MEREP,8],…]
[[10651,WANDK,9],[10652,WANDK,9],…]
[[10688,VAFFE,10],[10689,BERGS,10],…]
[[10726,EASTC,11],[10727,REGGC,11],…]
[[10760,MAISD,12],[10761,RATTC,12],…]

A4结果：

Month	OrderCount
1	33
2	29
3	30
4	31
5	32
6	30
7	33
8	33
9	37
10	38
11	34
12	48

3. 每个成员根据相应的数列存放到指定组



下面是发帖记录表。按标签将帖子分组，并统计各标签出现频率。

ID	TITLE	Author	Label
1	Easy analysis of Excel	2	Excel,ETL,Import,Export
2	Early commute: Easy to pivot excel	3	Excel,Pivot,Python
3	Initial experience of SPL	1	Basics,Introduction
4	Talking about set and reference	4	Set,Reference,Dispersed,SQL
5	Early commute: Better weapon than Python	4	Python,Contrast,Install
…	…	…	…

3. 每个成员根据相应的数列存放到指定组

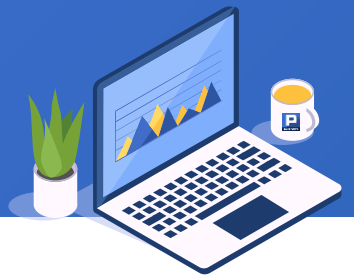
SPL如下，使用align(n, y)函数的@r选项，每个成员可能匹配多个组：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from PostRecord")	/查询发帖记录表
3	=A2.conj(Label.split(",")).id()	/将标签按逗号分隔后合并到一个序列，获得没有重复值的全部标签。
4	=A2.align@ar(A3.len(),A3.pos(Label.split(",")))	/使用align函数的@r选项，按照每个帖子的标签在全部标签中的定位分组。
5	=A4.new(A3(#):Label,~.count():Count).sort@z(Count)	/统计每个标签的帖子数量，按降序排列

A5结果：

Label	Count
SPL	7
SQL	6
Basics	5
…	…

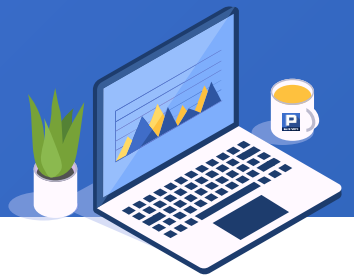
4. 分段函数



下面是员工表，根据工资将员工分为8000以下、8000-12000和12000以上，并统计每组人数。

ID	NAME	BIRTHDAY	SALARY
1	Rebecca	1974-11-20	7000
2	Ashley	1980-07-19	11000
3	Rachel	1970-12-17	9000
4	Emily	1985-03-07	7000
5	Ashley	1975-05-13	16000
…	…	…	…

4. 分段函数



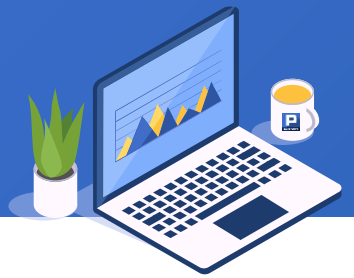
SPL如下，在对位函数align(n,y)中，用到了pseg(x)函数进行分段：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from EMPLOYEE")	/查询员工表
3	[0,8000,12000]	/定义工资区间
4	=A2.align@a(A3.len(),A3.pseg(SALARY))	/使用pseg函数获取工资所在区间
5	=A4.new(A3 (#):SALARY,~.count():COUNT)	/统计每组的人数

A5结果：

SALARY	COUNT
0	308
8000	153
12000	39

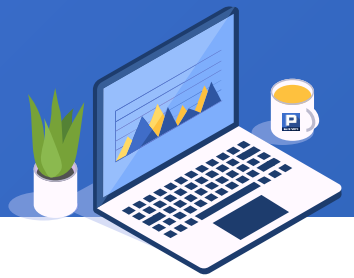
4. 分段函数



下面是员工表，根据入职时间将员工分为10年以下、10-20年和20年以上，并统计每组的平均工资。

ID	NAME	HIREDATE	SALARY
1	Rebecca	2005-03-11	7000
2	Ashley	2008-03-16	11000
3	Rachel	2010-12-01	9000
4	Emily	2006-08-15	7000
5	Ashley	2004-07-30	16000
…	…	…	…

4. 分段函数



SPL如下，在对位函数align(n,y)中，用到了pseg(x,y)函数进行分段：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from EMPLOYEE")	/查询员工表
3	[0,10,20]	/定义入职年限区间
4	=A2.align@a(A3.len(),(x=now(), A3.pseg(elapse@y(x,-~), HIREDATE)))	/使用pseg函数获取入职时间所在区间
5	=A4.new(A3(#):EntryYears,~.avg(SALARY):AvgSalary)	/统计每组的平均工资

A5结果：

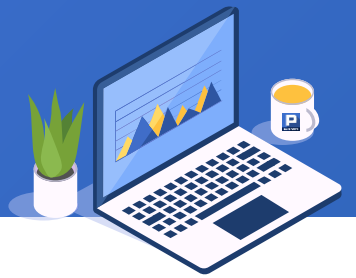
EntryYears	AvgSalary
0	6777.78
10	7445.53
20	6928.57

CONTENTS

02

枚举分组

1. 每个成员只存放到第一个匹配组



下面是中国主要城市的市区人口表，根据人口将城市分类。

ID	City	Population	Province
1	Shanghai	12286274	Shanghai
2	Beijing	9931140	Beijing
3	Chongqing	7421420	Chongqing
4	Guangzhou	7240465	Guangdong
5	Hong Kong	7010000	Hong Kong Special Administrative Region
…	…	…	…

1. 每个成员只存放到第一个匹配组

SPL如下，其中用到了enum函数来实现枚举分组：

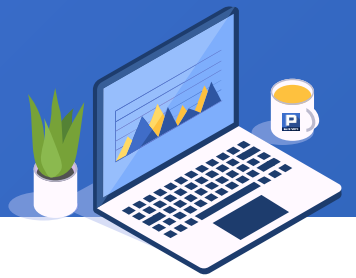
	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from UrbanPopulation")	/查询城市人口表
3	[?>2000000,?>1000000,?>500000,?<=500000]	/超大城市200万人口以上，特大城市100-200万人口，大城市50-100万，其他中小城市。
4	=A2.enum(A3,Population)	/人口按A3定义的条件进行枚举分组

A4结果：

Members
[[1,Shanghai,12286274,...], [2,Beijing, 9931140,...], …]
[[28,Changsha,1965282,...], [29,Nanchang,1900817,...], …]
[[69,Huainan,974026,...], [70,Haikou, 967336,...], …]
[]

ID	City	Population	Province
1	Shanghai	12286274	Shanghai
2	Beijing	9931140	Beijing
3	Chongqing	7421420	Chongqing
…	…	…	…

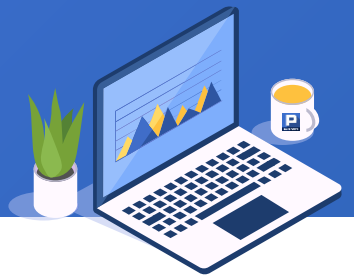
2. 不匹配成员存放到新组



下面是员工表，根据年龄将员工分组统计平均工资。

ID	NAME	BIRTHDAY	SALARY
1	Rebecca	1974-11-20	7000
2	Ashley	1980-07-19	11000
3	Rachel	1970-12-17	9000
4	Emily	1985-03-07	7000
5	Ashley	1975-05-13	16000
…	…	…	…

2. 不匹配成员存放到新组



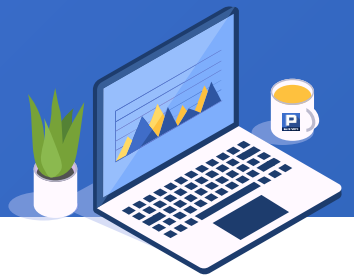
SPL如下，其中用到了enum@n选项，将不匹配成员存放到新组：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from EMPLOYEE")	/查询员工表
3	[?<35,?<45]	/将年龄段划分为35岁以下，45岁以下
4	=A2.enum@n(A3, age(BIRTHDAY))	/首先根据生日计算出年龄。再按年龄枚举分组，不匹配成员存放到新组。
5	=A4.new(if (#>A3.len(), "Other",A3(#)):AGE,~.avg(SALARY):AvgSalary)	/统计每组的平均工资，最后一组名称设置为Other

A5结果：

AGE	AvgSalary
?<35	7118.18
?<45	7448.76
Other	7395.06

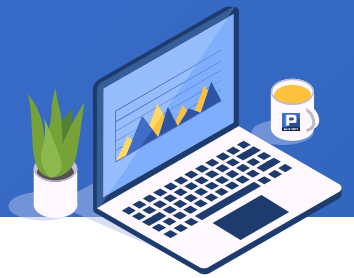
3. 每组检查所有成员是否匹配



下面是城市GDP表，分别统计直辖市、一线城市、二线城市的人均GDP。需要注意的是，分组可能会有重复成员，比如北京既是一线城市，又是直辖市。

ID	City	GDP	Population
1	Shanghai	32679	2418
2	Beijing	30320	2171
3	Shenzhen	24691	1253
4	Guangzhou	23000	1450
5	Chongqing	20363	3372
…	…	…	…

3. 每组检查所有成员是否匹配



SPL如下，其中用到了enum@r选项，每组检查所有成员是否匹配：

	A	B
1	=connect("db")	/连接数据库
2	=A1.query("select * from GDP")	/查询城市GDP表
3	[["Beijing","Shanghai","Tianjing","Chongqing"].pos(?)>0,["Beijing","Shanghai","Guangzhou","Shenzhen"].pos(?)>0,["Chengdu","Hangzhou","Chongqing","Wuhan","Xian","Suzhou","Tianjing","Nanjing","Changsha","Zhengzhou","Dongguan","Qingdao","Shenyang","Ningbo","Kunming"].pos(?)>0]	/枚举直辖市、一线城市和二线城市
4	=A2.enum@r(A3,City)	/按城市枚举分组
5	=A4.new(A3(#):Area,~.sum(GDP)/~.sum(Population)*10000:CapitaGDP)	/统计每组的人均GDP

A5结果：

Area	CapitaGDP
["Beijing","Shanghai","Tianjing","Chongqing"].pos(?)>0	107345.03
["Beijing","Shanghai","Guangzhou","Shenzhen"].pos(?)>0	151796.49
["Chengdu","Hangzhou","Chongqing","Wuhan","Xian","Suzhou","Tianjing","Nanjing","Changsha","Zhengzhou","Dongguan","Qingdao","Shenyang","Ningbo","Kunming"].pos(?)>0	106040.57

THANKS

感谢观看