SPL parsing and exporting Excel

Contents

1 Parsing

It is a common task in data analysis to parse data in Excel files into structured data (such as importing into relational database). However, many Excel files are irregular in format, and the structuring workload is relatively large. Besides it is difficult to be used universally, and it is necessary to analyze file formats before the actual developing work.

Therefore, If there is a tool software that can easily complete this work, it will greatly help business work and improve business efficiency.

esProc is such an efficient and flexible tool, which can easily read Excel data, and then structured them into a "sequence table" before processing.

Next, we will discuss how to use esProc to structure Excel data in different situations.

1.1 Ordinary Style

Simplest case: The first row in the Excel file is the column header, and from the second row onward, each row is a data record.

	A	B	C	D	E	F	G
1	No	Name	Class	Sex	Chinese	Maths	English
2	110210	Lorry	5(1)	M	80	60	86
3	110211	Tom	5(1)	M	81	72	67
4	110212	Curry	5(1)	F	97	91	87
5	110213	Joan	5(1)	F	86	69	73

A1 Open the "scores.xlsx" file and import it into a sequence table. The option @t indicates that the first row of the file is the column heading.
A2 Write the sequence table in A1 as a text file, keeping the original header.

	A
1	=file(“scores.xlsx").xlsimport@t()
2	=file(“scrores.txt”).export@t(A1)

The sequence table obtained in A1 is as follows:

Index	No	Name	Class	Sex	Chinese	Maths	English
1	110210	Lorry	5(1)	M	80	60	86
2	110211	Tom	5(1)	M	81	72	67
3	110212	Curry	5(1)	F	97	91	87
4	110213	Joan	5(1)	F	86	69	73

1.2 Multi-row header Style

The table header is complex, including the table name, the project name, the page number, the filler, the date of filling in, and so on.

	A	B	C	D	E	F	G
1	Item Lists And Prices						
2	Project: Building					page 1/total 52	
3	No	Item Code	Item Name	Unit	Quanti ty	SumOfMoney(yuan)	
4						Price	Sum
5	1.1.2	NJSJ	Internal scaffolding	term			
6	1.1.2.1	11001004001	Internal scaffolding	term	1.00	1006577.54	1006577.54
7	1.1.2.1.1.1	A22-28	Steel pipe	100㎡	137.88	912.07	125756.21
8	1.1.2.1.1.2	A22-28	Base of internal scaffolding	100㎡	71.83	912.07	65513.99

A1 Open the file and import the data into a sequence table. The parameter "1,5" means to read the first sheet, starting at line 5, until the end of the file.
A2 Change the sequence table column name in A1 to “No, Itemcode, Itemname, Unit, Quantity, Price, Sum", that is, the column name of the data table to be imported.

	A
1	=file(“listsAndPrices.xlsx").xlsimport(;1,5)
2	=A1.rename(#1:No,#2:ItemCode,#3:ItemNam e,#4:Unit,#5:Quantity,#6:Price,#7:Sum)

The sequence table obtained in A2 is as follows:

Index	No	ItemCode	ItemName	Unit	Quantity	Price	Sum
1	1.1.2	NJSJ	Internal	term			
2	1.1.2.1	1.1001E+10	Internal	term	1	1006577.54	1006577.54
3	1.1.2.1.1.1	A22-28	Steel pipe	100㎡	137.88	912.07	125756.21
4	1.1.2.1.1.2	A22-28	Base of	100㎡	71.83	912.07	65513.99

1.3 Free Format

Sometimes the data in Excel file is not the regular table in grid format, but the free format of the field name followed by the field value.

	A	B	C	D	E	F	G
1		ID:	1				
2		Name:	Yin Zhang		Sex	F	
3		Position	Sales				
4		Birthday	1968-12-08				
5		Phone:		(010) 65559857			
7		Address:	No. 236 of Fuxingmen Beijing				
8		PostCode:	100098				
9							
10		ID:	2				
11		Name:	Wei Wang		Sex	M	
12		Position	Sales Manager				
13		Birthday	1962-02-19				
14		Phone:		(010) 65559482			
16		Address:	No. 890 of Luoma Garden Beijing				
17		PostCode:	109801				
18							
19		ID:	3				

Each employee's information takes up 9 lines, which are arranged in order. How to structure this kind of file? Please see:

	A	B	C
1	=create(ID,Name,Sex,Position,Birthday,Phone,Address,Post Code)		
2	=file(“Employe.xlsx").xlsopen()		
3	[C,C,F,C,C,D,C,C]	[1,2,2,3,4,5,7,8]	
4	for	=A3.(~/B3(#)).(eval($[A2.xlscell(]/~/")"))	
5		if len(B4(1))==0	break
6		>A1.record(B4)	
7		>B3=B3.(~+9)	

1.3 Free Format

A1 Create an empty sequence table with column names “ID, Name, Sex, Position, Birthday, Phone, Address, PostCode”
A2 Open Excel file
A3 Define the cell column number sequence of employee information
B3 Define the cell row number sequence of employee information
A4 Use for loop to read each employee's information
B4 A3.(~/B3(#)) first calculate the current employee cell number sequence, and then read out these cell values to form the employee information sequence. The first cycle is [C1, C2, F2, C3, C4, D5, C7, C8], the second cycle is [C10, C11, F11, C12, C13, D14, C16, C17]... Add 9 to the row number each time. $[A2.xlscell(] is the same as “A2.xlscell(”, representing a string.
B5 Judge whether the employee ID value is empty. If it is empty, exit the loop and end the operation.
B6 Store an employee's information at the end of A1 sequence
B7 Add 9 to the row number sequence of employee information to read the next employee information

The sequence table obtained in A1 is as follows:

Index	ID	Name	Sex	Position	Birthday	Phone	Address	PostCode
1	1	Yin Zhang	F	Sales	1968-12-8	(010)65559857	No. 236 of Fuxingmen Beijing	100098
2	2	Wei Wang	M	Sales Manager	1962-2-19	(010)65559482	No. 890 of Luoma Garden Beijing	109801
3	3	Fang Li	F	Market	1973-8-30	(010)65553412	No. 253 of Shaoyao Garden Beijing	198033
4	4	Jianjie He	M	Market Manager	1968-9-19	(010)65558122	No. 45 of Qianmen Street Beijing	198052

1.4 Cross Table

Excel also has data in crosstab format.

	A	B	C	D	E	F	G	H
1	Orders Statistics							
2	Type \ Area / Amount	West	East	Center	North	South	Northwest	Southwest
3	Urgent express	20	70	1	97	23	2	35
4	Unified parcel	25	89	1	148	39	3	27
5	Federal cargo	15	79	52	108	29	2	23
6	Air transport	5	1	12	1	1	9	6
7	Cash on Delivery	8	2	4	1	6	7	9
8	General express	32	41	36	48	26	22	18

A1 Open the file and import the data into a sequence table.
A2 Because the first cell in the second row is a picture, the read data is null, and the first column has no column title, so change the name of the first column to type.
A3 Row and column transpose the sequence table data by grouping type. The option @ r means to convert the column data to row data. After conversion, the new column names are "area" and "amount".

	A
1	=file("cross.xlsx").xlsimport@t(;1,2)
2	=A1.rename(#1:Type)
3	=A2.pivot@r(Type;Area,Amount)

The sequence table obtained in A3 is as follows:

Index	Type	Area	Amount
1	Urgent express	West	20
2	Urgent express	East	70
3	Urgent express	Center	1
4	Urgent express	North	97
5	Urgent express	South	23
6	Urgent express	Northwest	2
7	Urgent express	Southwest	35
8	Unified parcel	West	25
9	Unified parcel	East	89

1.5 Main sub-table

In the employee information table shown in the figure below, in addition to the employee's own information, there is also family member information. Each sheet keeps information about an employee, so there are as many sheets as there are employees. The following figure shows the contents of the first sheet:

	A	B	C	D	E	F	G	H
1	Employee Information							
2	company	XX company				date:	1982-06-25	
3	Name	San Zhang	Sex	M	Birthday	1982-06-25	Nation	han
4	IDCard	510121198206253112			Phone	13612345678	Depart	Research
5	Home	No.25 Xisanqi			Marital	Married	Entry	2002-02
6	Family	Name	Relation		Workplace		Phone	
7		Zhou Zhang	farther		retired		15313231568	
8		Yin Hu	wife		XX company		13718826593	
9		Wuji Zhang	son		XX school			

A1 Create an empty sequence table with column names "IDcard, name, sex, birthday, nation, phone, Dept, home, maritime, entry" to save the employee information in the main table

A2 Create an empty order table with column names "IDcard, name, relation, workplace, phone" to save the information of employee family members in the sub table

A3 Define the cell sequence of employee information in the main table

A4 Open Excel file

	A	B	C
1	=create(IDCard,Name,Sex,Birthday,Nation,Phone,Depart,Home,Marital,Entry)		
2	=create(IDCard,Name,Relation,Workplace,Phone)		
3	[B4,B3,D3,F3,H3,F4,H4,B5,F5,H5]		
4	=file(“employee.xlsx").xlsopen()		
5	for A4	=A3.(eval($[A4.xlscell(]/~/",\""/A5.stname/"\")"))	>A1.record(B5)
6		=A4.xlsimport@t(Family,Name,Relation,Workplace,Phone;A5.stname,6)	
7		=B6.rename(Family:IDCard)	>B7.run(IDCard=B5(1))
8		>A2.insert@r(0:B7)	

1.5 Main sub-table

A5 Read each sheet of the Excel file in loop
B5 Read employee information sequence
C5 Save the employee information read by B5 to sequence table A1
B6 Read the employee family member information from row 6 onward, read only the specified "family, name, relationship, workplace, phone" 5 columns.
B7 Change the name of family column of B6 sequence table to IDcard
C7 Assign IDcard column of B7 sequence table to IDcard in employee information
B8 Save employee family member information in B7 to sequence table A2

The sequence table obtained in A1 is as follows:

Index	IDCard	Name	Sex	Birthday	Nation	Phone	Depart	Home	Marital	Entry
1	510121198206253000	San Zhang	M	1982-6-25	han	13612345678	Research	No.25 Xisanqi	married	2002-2-1
2	110114198907286000	Si Li	M	1989-7-28	han	13818022624	Sales	No.302 Lifeng Garden	free	2008-6-1
3	310503198803243000	Xiaoyu Zhao	F	1988-3-24	miao	13852416325	HR	No.501 Yuquyuan	married	2005-1-1

The sequence table obtained in A2 is as follows:

Index	IDCard	Name	Relation	Workplace	Phone
1	510121198206253000	Zhou Zhang	farther	retired	15313231568
2	510121198206253000	Yin Hu	wife	XX company	13718826593
3	510121198206253000	Wuji Zhang	son	XX school	
4	110114198907286000	Dasuan Li	farther	XX trade company	13625689532
5	110114198907286000	Haixia Liu	mother	XX supermarket	13924689512
6	310503198803243000	Darong Luo	husband	XX software company	13598325647
7	310503198803243000	Xiaolu Luo	daughter	XX primary school	

1.6 Import large amount of data

If there are a large number of data records in Excel file and can’ t fit in memory, how to import them?

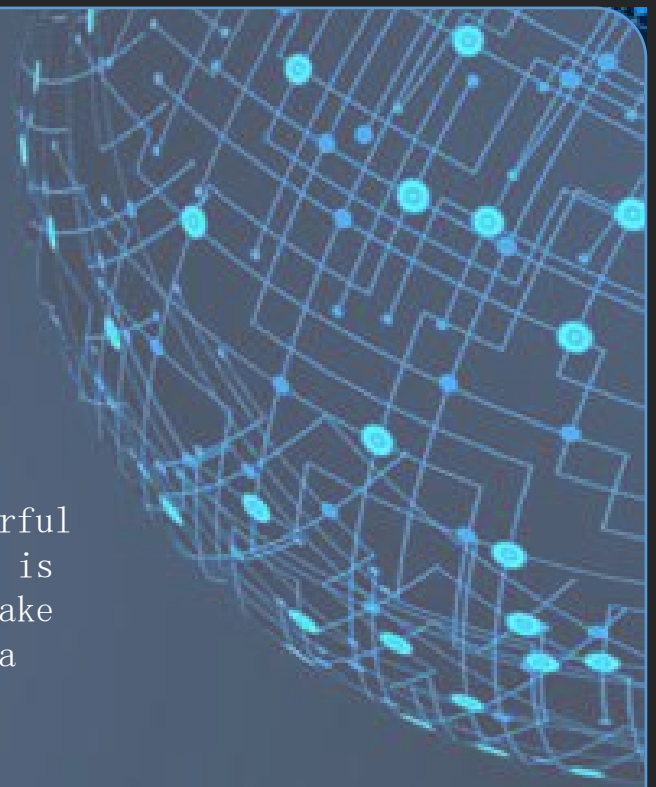
	A	B	C	D	E
1	ID	Company	Area	OrderDate	Amount
123661	10251	Qiangu	East	2012-07-08	624.9
123662	10252	Fuxing	West	2012-07-09	3059.49
123663	10253	Shiyi	North	2012-07-10	1428
123664	10254	Haotian	Center	2012-07-11	545.3

A1 Open the "orders.xlsx" file and import it as a cursor. The @t option indicates that the first row of the file is the column title, and the @C option indicates that the cursor is returned.
A2 Store the cursor data in A1 into the text file orders.txt, keeping the title unchanged

	A
1	=file(“orders.xlsx").xlsimport@tc()
2	=file(“orders.txt”).export@t(A1)

This example is very similar to 1.1, except that the @c option is used in A1 to read with a cursor. Only Excel files in xlsx format can be read using cursor.

2 Exporting



Sometimes we need to use a program to automatically generate Excel files. VBA with excel is not very easy to use, and esProc as a data processing tool will be very convenient to achieve this requirement.

The following will introduce how to generate excel file with esProc. The powerful data calculation ability of esProc is not the focus here. Therefore, the text is simply used as a data source example here. In practical application, it may take data from various data sources, and then get the data to be exported through a series of operations.

2.1 Export data only

2.1.1 Export new file

	A
1	=file("orders.txt":"UTF-8").import@t()
2	=file("orders.xlsx").xlsexport@t(A1)

A1 Read in an enterprise order table in text format, simulating the data that may be obtained through calculation.
A2 Export the data of A1 to the orders.xlsx file (automatically created when the file does not exist). There is no field specified in the parameter of the function xlsexport. All fields of A1 will be exported and the field name will remain unchanged. Because you do not specify which sheet to export to, export to Sheet1. The function uses the option @t to export the field name to the first row.

Exported Excel file：

	A	B	C	D	E	F
1	ID	Company	Area	OrderDate	Amount	Phone
2	10248	Shantai	North	2012-07-04	428	(030) 26471510
3	10249	Dongdiwar	East	2012-07-05	1842	(0251) 1031259
4	10250	Shiyi	North	2012-07-08	1523.5	(0211) 5550091
5	10251	Qiangu	East	2012-07-08	624.95	(071) 8325486
6	10252	Fuxing	West	2012-07-09	3559.5	(030) 23672220
7	10253	Shiyi	North	2012-07-10	1428	(0211) 5550091
8	10254	Haotian	Center	2012-07-11	545.4	(030) 30076545
9	10255	Yongda	North	2012-07-12	2450	(089) 7034214

2.1 Export data only

2.1.2 Data addition

	A
1	=file("aday.txt":"UTF-8").import@t()
2	=file("orders.xlsx").xlsexport(A1)

If the excel file already exists, the exported data will be appended to the original file. In this case, you do not need to use the @t option.

2.1 Export data only

2.1.3 Export to different sheets

	A
1	=file("orders.txt":"UTF-8").import@t()
2	=A1.select(Company==“Shantai")
3	=file(“orders.xlsx”).xlsexport@t(A2,ID,Company,OrderDate:Date,Amount:Money;”Shantai")

A2 filters the sequence table A1. In A3, A2 is exported to orders.xlsx. Only four fields, ID, company, OrderDate and amount, are exported, and OrderDate is renamed date, amount is renamed money, and data is exported to a new sheet named Shantai.

The exported Excel file：

	A	B	C	D
1	ID	Company	Date	Money
2	10248	Shantai	2012-07-04	428
3	10274	Shantai	2012-08-06	529
4	10295	Shantai	2012-09-02	120
5				

Sheet1 Shantai

2.2 Export large amount of data

What if there is a large amount of data?
Using cursor to read data one by one will not read all data into memory.
When exporting with a cursor, you need to add the @s function option, so that when exporting, it will be exported as a stream, and the resulting excel file will not be saved in memory.

	A
1	=file("big.txt":"UTF-8").cursor@t()
2	=file("big.xlsx").xlsexport@st(A1)

The exported Excel file：

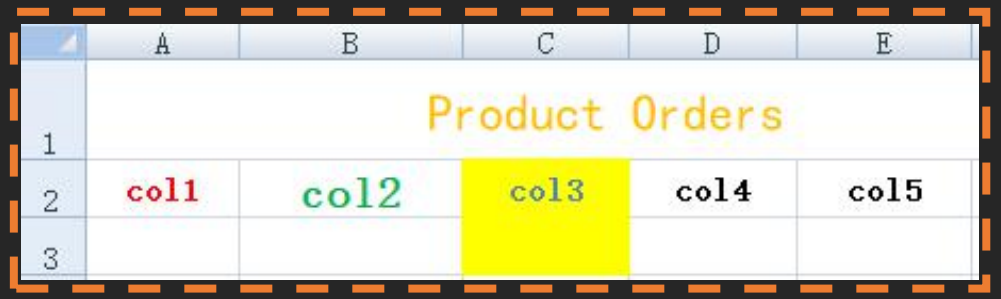
	A	B	C	D	E
1	ID	Company	Area	OrderDate	Amount
123658	10248	Shantai	North	2012-07-04	428
123659	10249	Dongdiwang	East	2012-07-05	1842
123660	10250	Shiyi	North	2012-07-08	123.49
123661	10251	Qiangu	East	2012-07-08	624.9
123662	10252	Fuxing	West	2012-07-09	3059.49
123663	10253	Shiyi	North	2012-07-10	1428
123664	10254	Haotian	Center	2012-07-11	545.3
123665					

In this example, 123663 data records are exported. This method can export hundreds of millions of records. However, a sheet of Excel file can only store 1048576 rows of data at most, so when the exported data reaches a million lines, a new sheet will be added in Excel to save.

2.3 Specify display attributes

Sometimes we want the generated excel file to be able to specify the format (such as font, color, background color, alignment, etc.). At this time, as long as the excel file template is built in advance, the format is defined, and then the data can be exported to the file.

As shown in the following figure, write the table name in the first row of the orders.xlsx file Sheet1, the field column name in the second row, and define some style attributes for the table name and each column. The middle of the first, third and fourth columns is aligned, the left of the second column is aligned, the right of the fifth column is aligned, the display format of the fourth column “yyyy-mm-dd”, and the display format of the fifth column is “#,###.00”.



	A	B	C	D	E
1	Product Orders				
2	col1	col2	col3	col4	col5
3					

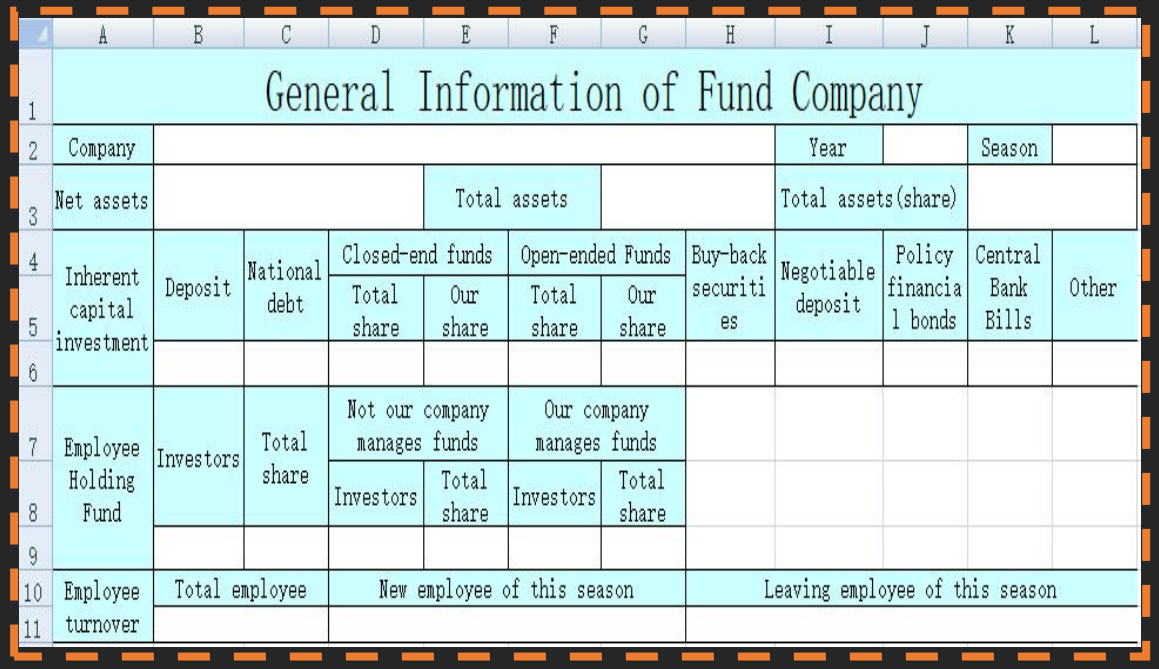
The export program of esProc is the same as the previous example 2.1.1, and the export result is shown in the figure below. When exporting to an existing file, the last non empty line of the file will be used as the header, which will be overwritten. When exporting, various style attributes defined in the original file will be used (not supported in big data streaming export).

The exported Excel file：

	A	B	C	D	E
1	Product Orders				
2	ID	Company	Area	OrderDate	Amount
3	10257	Yuandong	East	2012-07-16	1,109.00
4	10258	Zhengren	East	2012-07-17	1,604.00
5	10259	Sanjie	West	2012-07-18	100.00
6	10260	Jingmi	North	2012-07-19	1,461.75
7	10261	Lange	South	2012-07-19	440.00
8	10262	Xueren	Center	2012-07-22	583.20

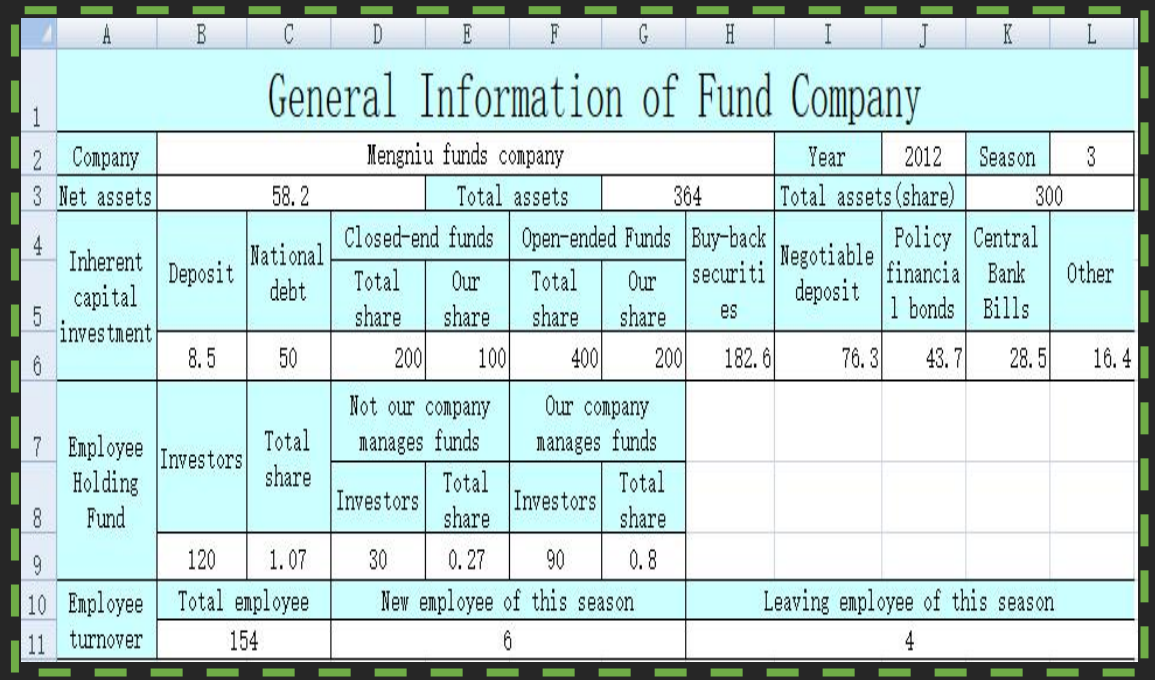
2.4 Fill in data in fixed rows and columns

esProc also provides a way to read and write a cell or a block of cells specified in the excel file. You can fill in data to excel in a fixed format, such as the following excel file:



	A	B	C	D	E	F	G	H	I	J	K	L
1	General Information of Fund Company											
2	Company								Year		Season	
3	Net assets				Total assets				Total assets(share)			
4	Inherent capital investment	Deposit	National debt	Closed-end funds		Open-ended Funds		Buy-back securities	Negotiable deposit	Policy financial bonds	Central Bank Bills	Other
5				Total share	Our share	Total share	Our share					
6												
7	Employee Holding Fund	Investors	Total share	Not our company manages funds		Our company manages funds						
8				Investors	Total share	Investors	Total share					
9												
10	Employee turnover	Total employee		New employee of this season				Leaving employee of this season				
11												

Excel effect after export:



	A	B	C	D	E	F	G	H	I	J	K	L
1	General Information of Fund Company											
2	Company	Mengniu funds company							Year	2012	Season	3
3	Net assets	58.2		Total assets		364			Total assets(share)		300	
4	Inherent capital investment	Deposit	National debt	Closed-end funds		Open-ended Funds		Buy-back securities	Negotiable deposit	Policy financial bonds	Central Bank Bills	Other
5				Total share	Our share	Total share	Our share					
6		8.5	50	200	100	400	200	182.6	76.3	43.7	28.5	16.4
7	Employee Holding Fund	Investors	Total share	Not our company manages funds		Our company manages funds						
8				Investors	Total share	Investors	Total share					
9		120	1.07	30	0.27	90	0.8					
10	Employee turnover	Total employee		New employee of this season				Leaving employee of this season				
11		154		6				4				

2.4 Fill in data in fixed rows and columns

	A	B	C	D	E	F
1	Mengniu funds company	2012	3	58.2	364	300
2	8.5	50	200	100	400	200
3	182.6	76.3	43.7	28.5	16.4	
4	120	1.07	30	0.27	90	0.8
5	154	6	4			
6	=file(“result.xlsx")		=A6.xlsopen()			
7	=C6.xlscell("B2",1;A1)		=C6.xlscell("J2",1;B1)		=C6.xlscell("L2",1;C1)	
8	=C6.xlscell("B3",1;D1)		=C6.xlscell("G3",1;E1)		=C6.xlscell("K3",1;F1)	
9	=C6.xlscell("B6",1;[A2:F2].concat("\t"))			=C6.xlscell("H6",1;[A3:E3].concat("\t"))		
10	=C6.xlscell("B9",1;[A4:F4].concat("\t"))			=C6.xlscell("B11",1;[A5:C5].concat("\t"))		
11	=A6.xlswrite(C6)					

It is assumed that the data to be filled in has been calculated (in the first 5 rows). The first six cells to be filled in the sample table are all independent, so only one cell can be filled in at a time. The sixth row is cells that can be filled in consecutively. At this time, the data to be filled in is spelled into a string separated by \t, which can be filled in to the same row in order. After filling in all the data, write the excel object opened by C6 back to the result.xlsx file.