

IMPLEMENTING WHERE AND ORDER BY AS SPREADSHEET FORMULAS

SEMS18, 2018-10-01, LISBON PT

PAUL MIREAULT
HONORARY PROFESSOR
HEC MONTRÉAL
paul.mireault@ssmi.international

PLAN

–

- Example problem
- Selection (WHERE)
 - Problem with Filtering Tools
 - Formulas
- Sorting (ORDER BY)
 - Problem with Sorting Tools
 - Formulas
- Limits

EXAMPLE PROBLEM

THE SPORTS CLUBS FEDERATION

- The Federation procures products and services and re-sells them to its clubs.
- The price depends on the *population* of the club and the type of product. For administrative products, the population is the number of employees, and for activities products it is the number of club members.
- 72 clubs, 22 products, 1297 purchases.

BASIC EXCEL SELECTION TOOLS

THE FILTER BUTTON

- Filters in-place
 - this means only one filter can be used at the time
- Requires a user action
 - does not change when the underlying values change: the user needs to perform the sort again
- Requires the use of the SUBTOTAL function to replace SUM, AVG, COUNT, etc.

SELECTION PREPARATION

INDICATOR AND SEQUENCE VARIABLES

- Creation of an 0-1 Indicator Variable with the given selection criteria

=IF(Product_Code=Desired_Product,1,0)									
C	D	E	F	G	H	I	J	K	L
Club ID	Product Code	Previous Cost	Prod-Club	Population Product-Club	Product Cost	Selected Product Indicator	Seq No of Selected Product		
							0		
689	PURCH	\$1,527.00	PURCH-689	422	\$2,000.00	=IF(Product_Code =Desired_Product,1,0)			
689	BIM	\$10,374.00	BIM-689	9	\$20,000.00	0	0		
689	DOFIN	\$19,967.69	DOFIN-689	422	\$20,000.00	0	0		
689	GPI	\$13,467.85	GPI-689	422	\$15,000.00	1	1		
689	JADE	\$11,404.41	JADE-689	422	\$12,000.00	0	1		
689	MEM	\$5,821.19	MEM-689	422	\$6,000.00	0	1		

SELECTION PREPARATION

INDICATOR AND SEQUENCE VARIABLES

- Creation of Sequence Number of the instances that satisfy the selection criteria.
- Sequence Number = Previous Sequence Number + Indicator Variable

C	D	E	F	G	H	I	J
Club ID	Product Code	Previous Cost	Prod-Club	Population Product-Club	Product Cost	Selected Product Indicator	Seq No of Selected Product
							0
689	PURCH	\$1,527.00	PURCH-689	422	\$2,000.00	0	= J2 + I3
689	BIM	\$10,374.00	BIM-689	9	\$20,000.00	0	0
689	DOFIN	\$19,967.69	DOFIN-689	422	\$20,000.00	0	0
689	GPI	\$13,467.85	GPI-689	422	\$15,000.00	1	1
689	JADE	\$11,404.41	JADE-689	422	\$12,000.00	0	1

EXTRACTION AREA PREPARATION

IDENTIFY THE SELECTED ROWS

- Create sequential numbers
- Use MATCH to identify the first row containing the sequence number.
 - MATCH(NoSeq, SeqNo of Selected Row, 0) returns the row number of the **first** occurrence of **SeqNo** in **SeqNo of Selected Row**.

RATE		:	X	✓	fx	=MATCH(E2,'Club-Product Results'!J:J,0)			
	A	B	C	D	E	F	G	H	I
1	Product Cost by Club				NoSeq	Row Num	Club	Pop	Cost
2					1	=MATCH(E2,'Club-Product Results'!J:J,0)			
3	Desired Product	GPI			2	19	711	4128	\$25,143.
4					3	37	712	8328	\$41,353.
5	Optimal Combination of Desired Prod	8			4	55	713	3137	\$21,318.

EXTRACTION AREA PREPARATION

EXTRACT THE DESIRED VALUES

- Use INDEX to extract the desired values from the identified row.
- INDEX(Desired column, NoSeq) returns the desired value.

G2 ✕ ✓ fx =INDEX('Club-Product Results'!C:C,F2)				
	E	F	G	H
1	NoSeq	Row Number	Club	Pop
2	1	=MATCH(E2,'Club-Product Results'!J:J,0)	=INDEX('Club-Product Results'!C:C,F2)	=INDEX('Club-Product Results'!G:G,F2)
3	2	=MATCH(E3,'Club-Product Results'!J:J,0)	=INDEX('Club-Product Results'!C:C,F3)	=INDEX('Club-Product Results'!G:G,F3)
4	3	=MATCH(E4,'Club-Product Results'!J:J,0)	=INDEX('Club-Product Results'!C:C,F4)	=INDEX('Club-Product Results'!G:G,F4)

USAGE EXAMPLE

CHANGING THE CRITERIA

- Changing the selection criteria **Desired Product** automatically updates the extraction area:

	A	B	C	D	E	F	G	H	I	J		A	B	C	D	E	F	G	H	I	J
1	Product Cost by Club				NoSeq	Row Num	Club	Pop	Cost	Previous Cost	1	Product Cost by Club				NoSeq	Row Num	Club	Pop	Cost	Previous Cost
2					1	=MATCH(E2,'Club-Product Results'!J:J,0)				\$13,467.85	2					1	167	731	2957	\$4,742.31	\$3,976.00
3	Desired Product	GPI			2	19	711	4128	\$25,143.16	\$21,912.18	3	Desired Product	CV			2	343	753	7868	\$6,605.10	\$6,256.63
4					3	37	712	8328	\$41,353.68	\$43,562.10	4					3	749	822	4683	\$5,397.00	\$3,976.00
5	Optimal Combination of Desired Prod	8			4	55	713	3137	\$21,318.25	\$17,758.00	5	Optimal Combination of Desired Prod	4			4	1086	869	15382	\$9,455.24	\$12,231.77
6	Combi-Prod	8-GPI			5	73	714	6098	\$32,746.67	\$32,557.72	6	Combi-Prod	4-CV			5	1142	873	12584	\$8,393.93	\$10,006.80
7					6	93	721	6688	\$35,023.86	\$34,983.59	7					6	1162	881	4649	\$5,384.10	\$3,976.00
8	Floor Price	\$15,000			7	114	722	6398	\$33,904.56	\$35,667.66	8	Floor Price	\$4,000			7	1254	887	19199	\$10,903.07	\$15,267.04
9	Ceiling Price	\$125,000			8	133	723	10626	\$50,223.16	\$63,810.19	9	Ceiling Price	\$15,000			8	#N/A	#N/A	#N/A	#N/A	#N/A
10	Floor Population	1500			9	151	724	7434	\$37,903.16	\$44,252.16	10	Floor Population	1000			9	#N/A	#N/A	#N/A	#N/A	#N/A
11	Ceiling Population	30000			10	170	731	2957	\$20,623.51	\$19,102.00	11	Ceiling Population	30000			10	#N/A	#N/A	#N/A	#N/A	#N/A
12	Unit price between Floor and Ceiling	\$3.8596			11	187	732	23500	\$99,912.28	\$83,563.15	12	Unit price between Floor and Ceiling	\$0.3793			11	#N/A	#N/A	#N/A	#N/A	#N/A
13					12	207	733	11574	\$53,882.11	\$66,282.77	13					12	#N/A	#N/A	#N/A	#N/A	#N/A
14					13	226	734	24585	\$104,100.00	\$88,718.40	14					13	#N/A	#N/A	#N/A	#N/A	#N/A

BASIC EXCEL SORTING TOOLS

THE SORT BUTTON

- Sorts in-place
 - this means only one sort can be used at the time
- Requires a user action
 - does not change when the underlying values change: the user needs to perform the sort again
- Does not work when the sorted values depend on their position (implicit intersection)

RANK.EQ FUNCTION

FUNCTION'S BEHAVIOUR

- We use the result of the RANK.EQ function the same way we used the sequence numbers to select rows.
- RANK.EQ with unique values,
 - returns a set of values from 1 to the size of the set we are using;
- RANK.EQ with non-unique values,
 - it then assigns the same rank to each row with the same value. For example, if there are three rows tied for the first rank, the function will assign the value 1 to each of the three rows and 4 to the next one.

SORTING ON A SINGLE UNIQUE COLUMN

CREATING THE SORT RANK NUMBERS

- The preparation is simply creating the rank order column:

RATE							
=RANK.EQ(H2,Pop,0)							
	E	F	G	H	I	J	K
	Seq	Row					Pop
1	No	Number	Club	Pop	Cost	Previous Cost	Rank
2	1	6	689	422	\$15,000.00	\$13,467.85	=RANK.EQ(H2,Pop,0)
3	2	19	711	4128	\$25,143.16	\$21,912.18	57
4	3	37	712	8328	\$41,353.68	\$43,562.10	33
5	4	55	713	3137	\$21,318.25	\$17,758.00	62
6	5	73	714	6098	\$32,746.67	\$32,557.72	42
7	6	93	721	6688	\$35,023.86	\$34,983.59	40
8	7	114	722	6398	\$33,904.56	\$35,667.66	41

SORTING RESULT AREA

- Create sequential numbers
- Use MATCH to identify the first row containing the rank number.
- MATCH(NoSeq, Rank of Row, 0) returns the row number of the **first** occurrence of *SeqNo* in *Rank of Row*.

RATE													
=MATCH(M2,Pop_Rank,0)													
	E	F	G	H	I	J	K	L	M	N	O	P	Q
	Seq No	Row Number	Club	Pop	Cost	Previous Cost	Pop Rank		Seq No	Row Num			
1	No	Number	Club	Pop	Cost	Previous Cost	Rank		Pop	Pop	Club	Pop	Cost
2	1	6	689	422	\$15,000.00	\$13,467.85	72		1	=MATCH(M2,Pop_Rank,0)			
3	2	19	711	4128	\$25,143.16	\$21,912.18	57		2	23	763	46784	\$125,000.00
4	3	37	712	8328	\$41,353.68	\$43,562.10	33		3	45	831	41406	\$125,000.00
5	4	55	713	3137	\$21,318.25	\$17,758.00	62		4	48	851	35984	\$125,000.00
6	5	73	714	6098	\$32,746.67	\$32,557.72	42		5	46	841	34901	\$125,000.00
7	6	93	721	6688	\$35,023.86	\$34,983.59	40		6	55	864	33156	\$125,000.00
8	7	114	722	6398	\$33,904.56	\$35,667.66	41		7	21	761	32706	\$125,000.00
9	8	133	723	10626	\$50,223.16	\$63,810.19	30		8	56	865	32118	\$125,000.00
10	9	151	724	7434	\$37,903.16	\$44,252.16	37		9	13	734	24585	\$104,100.00
11	10	170	731	2957	\$20,623.51	\$19,102.00	63		10	11	732	23500	\$99,912.28

SORTING ON A SINGLE NON-UNIQUE COLUMN

PROBLEMS WITH MISSING RANKS

- When there are equal values, RANK.EQ gives the same rank to all of them and skips the unused ranks.
- The MATCH function cannot find a sequence number that corresponds to a skipped rank.
- It finds **Seq No Cost 1** as the 21st element of **Cost Rank** (in row 22), but there is no 2 in **Cost Rank**.

X2					
	K	S	T	U	V
	Pop Rank	Cost Rank	Seq No Cost	Row Num Cost	
1					
2	72	69	1	21	
3	57	57	2	#N/A	
4	33	33	3	#N/A	
5	62	62	4	#N/A	
6	42	42	5	#N/A	
7	40	40	6	#N/A	
8	41	41	7	#N/A	
21	55	55	20	15	
22	7	1	21	43	
23	1	1	22	26	
24	2	1	23	57	
25	61	61	24	60	

SORTING ON A SINGLE NON-UNIQUE COLUMN

CREATING A SURROGATE VALUE KEY

- Instead of ranking the values themselves, we will rank a constructed surrogate value that preserves the original ranking.
- $SK = \text{Cost Rank} \times A + \text{Pop Rank}$ (where $A > \text{set size}$)
- In the example, $A = 100$.

X2											
=RANK.EQ(W2,SK,1)											
	K	S	T	U	V	W	X	Y	Z	AA	AB
	Pop Rank	Cost Rank	Seq No Cost	Row Num Cost		SK	SK Rank	Row Num SK	Club	Pop	Cost
1	72	69	1	21		6972	72	22	762	74326	\$125,000.00
2	57	57	2	#N/A		5757	57	23	763	46784	\$125,000.00
3	33	33	3	#N/A		3333	33	45	831	41406	\$125,000.00
4	62	62	4	#N/A		6262	62	48	851	35984	\$125,000.00
5	42	42	5	#N/A		4242	42	46	841	34901	\$125,000.00
6	40	40	6	#N/A		4040	40	55	864	33156	\$125,000.00
7	41	41	7	#N/A		4141	41	21	761	32706	\$125,000.00
8	55	55	20	15		5555	55	15	741	16007	\$70,991.93
21	7	1	21	43		107	7	43	823	15871	\$70,467.02
22	1	1	22	26		101	1	26	772	15809	\$70,227.72
23	2	1	23	57		102	2	57	866	15722	\$69,891.93
24	61	61	24	60		6161	61	60	869	15382	\$68,579.65
25											

CASE OF A NON-UNIQUE SURROGATE VALUE

—

- There is a possibility that two clubs have the same cost and the same population. In this case, their SK will have the same value and there will be a skipped rank, generating an error.
- The solution is to add another term in the SK calculation. If that term is not unique, then you may need to add another, and so on. You can stop adding terms when all the terms contain a candidate key of the original data.
- $SK = (Cost\ Rank \times A + Pop\ Rank) \times A + Seq\ No$
- This SK guaranteed to be unique because **Seq No** is unique.

LIMITS

–

- RANK.EQ only works with numerical values.
 - Alternative: =COUNTIF(Product_Name;"<"&Product_Name)+1
- With large sets, A is big and we may reach numerical precision problem when SK is composed of many terms.
 - Alternative: Concatenate terms with a unique separator and use a COUNTIF.

QUESTIONS?