Monthly User Guide from JMP Korea

제 15호 (2018년 10월)

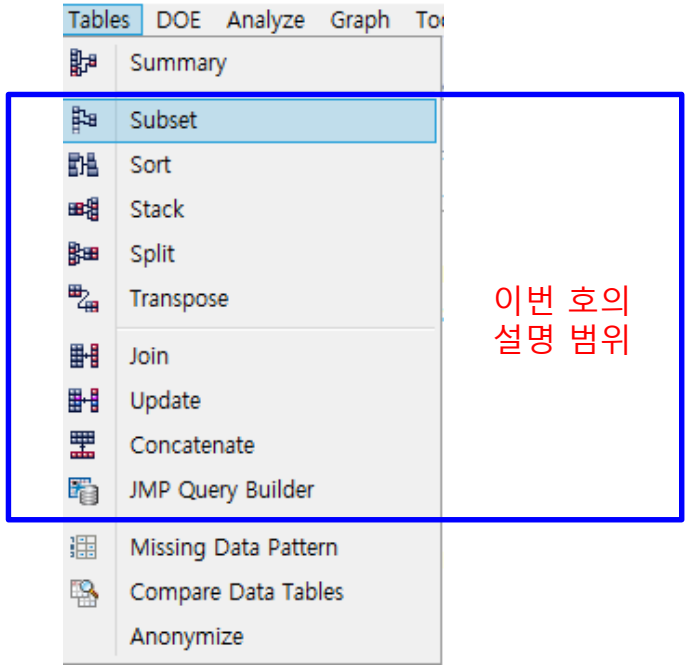
데이터 전 처리[Data Pre-Processing]

* 본 Guide 는 매월 세 번째 수요일에 발행됩니다
  (2018년 7월호부터는 JMP 14 Version 기준입니다)
** Monthly User Guide 지난 호는 다음 Site 를 참조하세요(https://www.jmp.com/ko_kr/newsletters.html)
*** 본 Guide 의 내용과 관련한 문의는 ikju.Shin@jmp.com 으로 연락 바랍니다

Data 전 처리

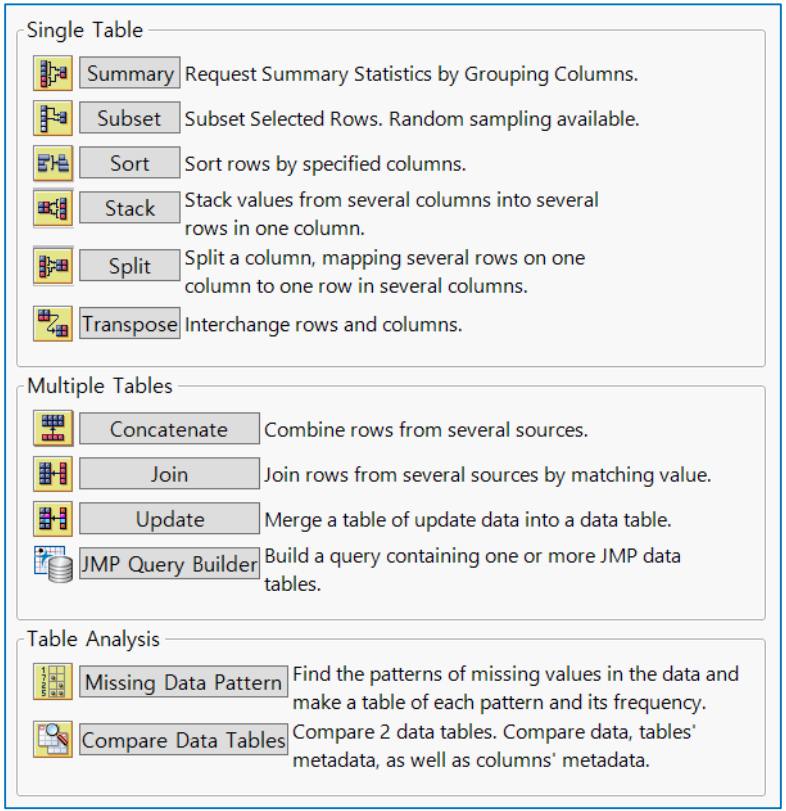
이번 호에서는 분석하기 전에 분석 가능한 형태로 Data 를 전 처리[사전 준비]하는 것과 관련된 JMP 의 기능, 그 중에서도 Tables Platform 의 기능에 대해 소개하겠습니다. DB 또는 MES, ERP 등의 사내 시스템으로부터 Data 를 불러오거나 공공 Data, 상용(Commercial) Data 및 다양한 Web Data 등을 분석하기 위해서는 Data 의 전처리가 반드시 필요하며, 이럴 경우, JMP 의 Tables Platform 의 다양한 기능이 유용한 도움이 될 것입니다.

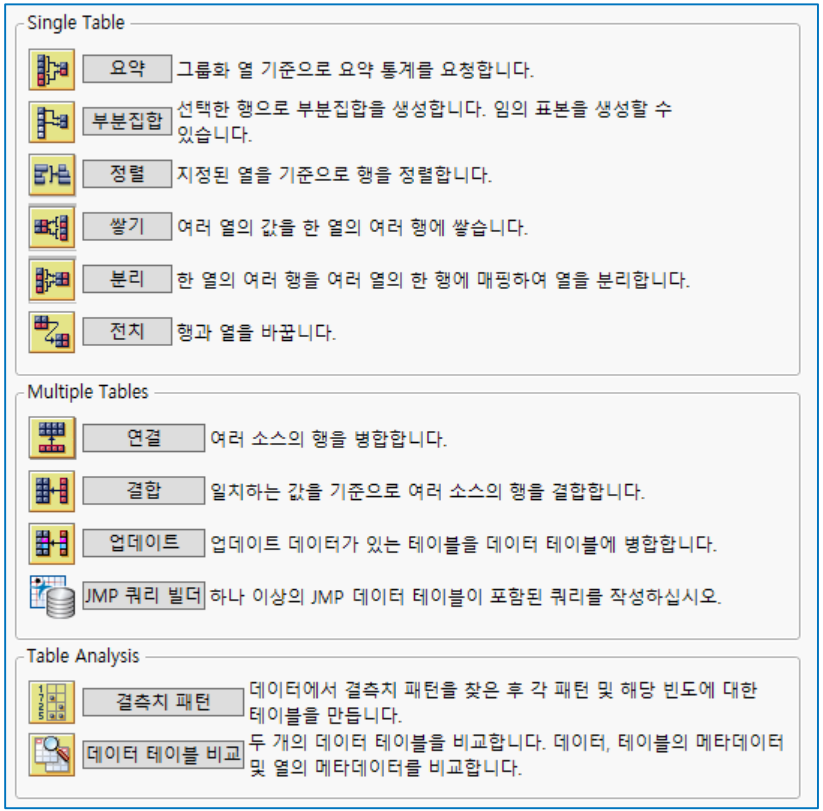
Data 전 처리(前 處理)
Data Preparation
Data Cleaning
Data Preprocessing
Data Wrangling
Data Management
Data Reshaping



Data 전 처리

View / JMP Starter 에서도 Tables Platform 확인 가능

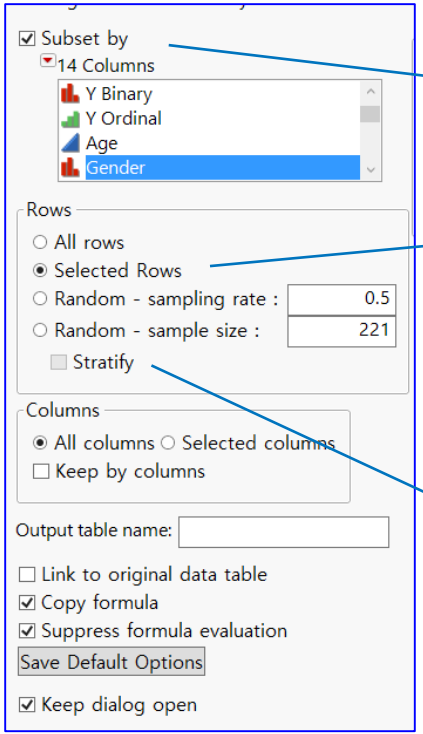




1. 부분(Subset) Data 만들기

Sample data : diabetes.jmp

1. Platform 소개



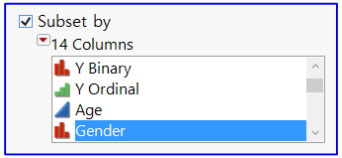
(층별) 변수별로 Subset.
만약 'Gender' 를 선택하면
남녀별로 분리된 2 개의
data table 이 생성됨

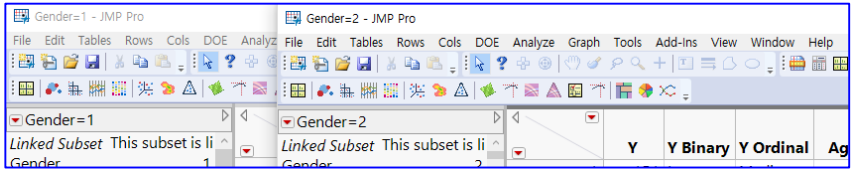
Data table 에서 선택한
rows 만 subset
(row panel 의 'selected' 에서
우측 마우스 클릭 후
data view 를 클릭한 결과와
같음

층화(Stratified) 랜덤 샘플링

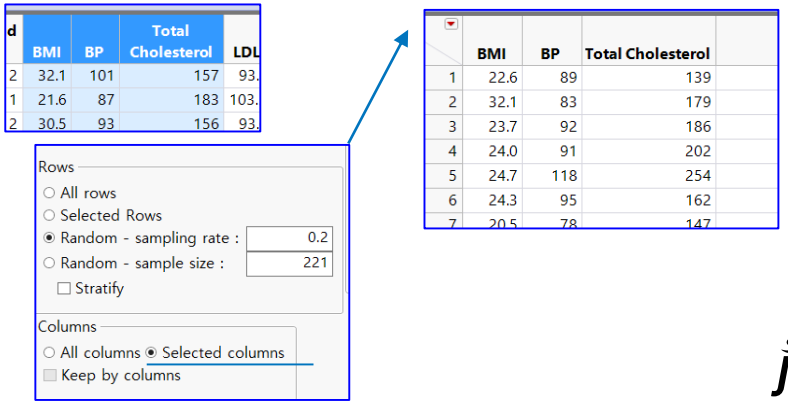
2. 변수별로 부분 data 만들고자 할 경우

오른쪽과 같이 선택하게 되면
Gender 별로 data table 이
만들어짐



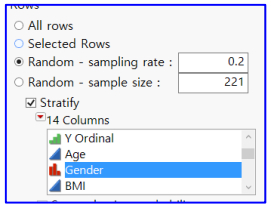


3. 특정 Column 에 대해서, 전체 Data 의 일정 비율 (또는 일정 개수 만큼)을 부분 Data를 만들고자 할 경우

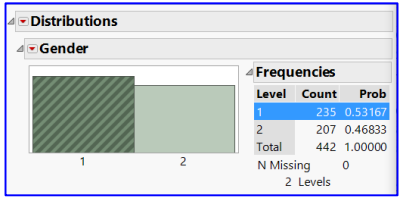


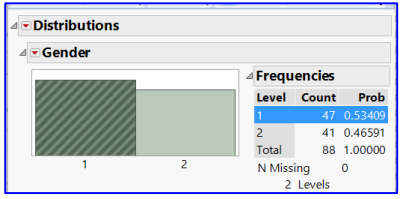
1. 부분(Subset) Data 만들기

4. 층화(Stratified) 랜덤 샘플링
변수별로 동일한 비율만큼 부분 data 를 만들고자 할 경우

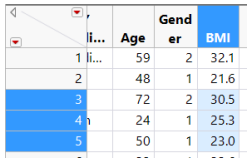


Analyze / distribution 에서
Raw data 와 subset data 에
대해 확인해 보면
동일 비율만큼 부분 data 가
만들어 졌음을 알 수 있다.

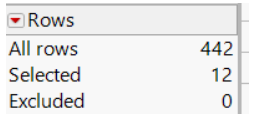




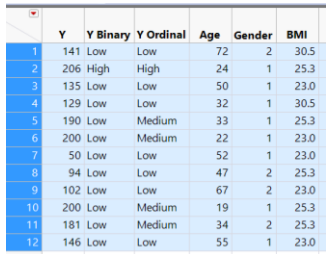
5. 특정 변수의 특정 값을 가진 row 만 추출하고 싶다
예를 들어, 변수 BMI 값이 23.0, 25.3, 30.5 값을 추출하고 싶은 경우
1) BMI Column 선택, 해당 값이 들어 있는 row 를 선택
우측 마우스 click → Select matching cells
2) Row panel 의 selected 에서 우측 마우스 click → data view



	li...	Age	Gender	BMI
1	li...	59	2	32.1
2		48	1	21.6
3		72	2	30.5
4		24	1	25.3
5		50	1	23.0

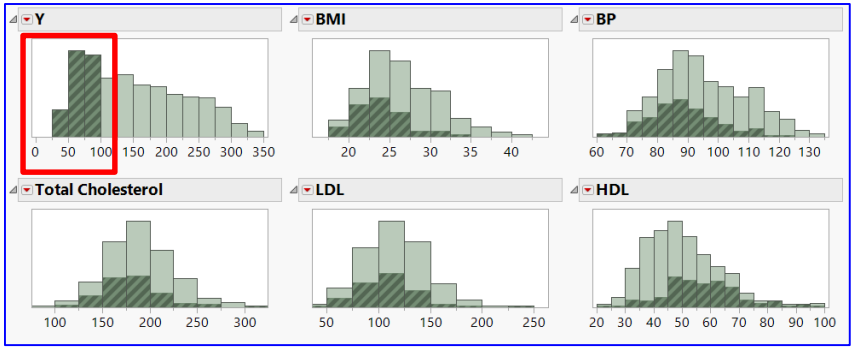


Rows	
All rows	442
Selected	12
Excluded	0

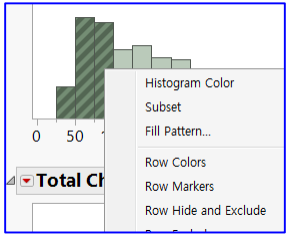


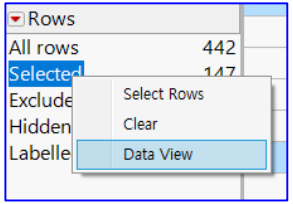
	Y	Y Binary	Y Ordinal	Age	Gender	BMI
1	141	Low	Low	72	2	30.5
2	206	High	High	24	1	25.3
3	135	Low	Low	50	1	23.0
4	129	Low	Low	32	1	30.5
5	190	Low	Medium	33	1	25.3
6	200	Low	Medium	22	1	23.0
7	50	Low	Low	52	1	23.0
8	94	Low	Low	47	2	25.3
9	102	Low	Low	67	2	23.0
10	200	Low	Medium	19	1	25.3
11	181	Low	Medium	34	2	25.3
12	146	Low	Low	55	1	23.0

<보다 실무적인 방법 : 2018년 9월 호에 소개된 내용>
1. 관심있는 모든 변수에 대해 Histogram 을 그림
(Analyze / Distribution)
2. 관심있는 변수의 관심있는 영역을 선택함
(여기서는 반응치 Y 의 100 이하 영역)



3. Histogram 에서 우측 마우스 클릭 후 subset을 선택하거나(또는 더블 클릭)
row panel 의 'selected' 에서 우측 마우스 클릭 후 data view 를
클릭하면 관심 있는 부분만 추출됨





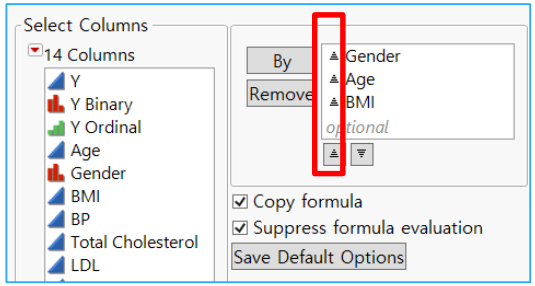
2. Data Sorting 하기

Sample data : diabetes.jmp

1. 특정한 Column 을 기준으로 데이터를 오름차순 또는 내림차순으로 정렬하는 기능
   선택된 순서대로 Sorting[정렬] 됨

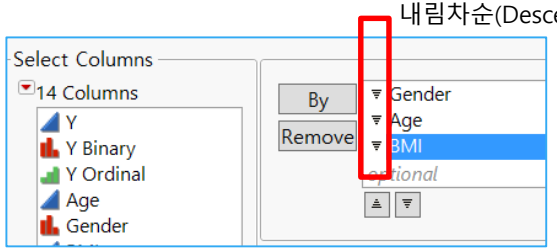
	Age	Gender	BMI	B
	59	2	32.1	
	48	1	21.6	
	72	2	30.5	
	24	1	25.3	
	50	1	23.0	
	23	1	22.6	
	36	2	22.0	
	66	2	26.2	
	60	2	32.1	
	29	1	30.0	
	22	1	18.6	
	56	2	28.0	
	53	1	23.7	

오름차순(Ascending)



nal	Age	Gender	BMI	BI
	19	1	19.2	
	19	1	23.2	
n	19	1	25.3	
	20	1	22.9	
	21	1	20.1	
	21	1	24.2	
	22	1	18.6	
	22	1	19.3	
n	22	1	19.9	
n	22	1	23.0	

내림차순(Descending)



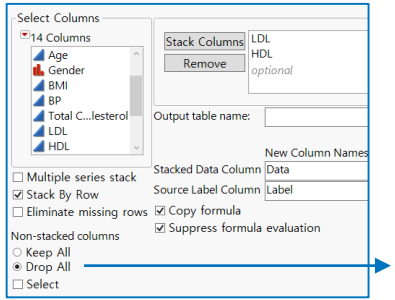
al	Age	Gender	BMI	
	79	2	27.0	
	79	2	23.3	
	72	2	30.5	
	71	2	27.0	9
	71	2	26.5	
	71	2	26.1	
	71	2	24.0	
	70	2	24.1	8
	69	2	24.5	
	68	2	27.5	
	68	2	25.7	

3. Data 쌓기(stack) 와 분리(Split)

Sample data : diabetes.jmp

1. 여러 개의 column 을 하나의 column 으로 쌓는 기능

1) LDL 과 HDL 을 하나의 Column 으로 쌓고자 한다면
Tables / stack 에서 LDL, HDL 을 'Stack Columns' 선택, Create 클릭

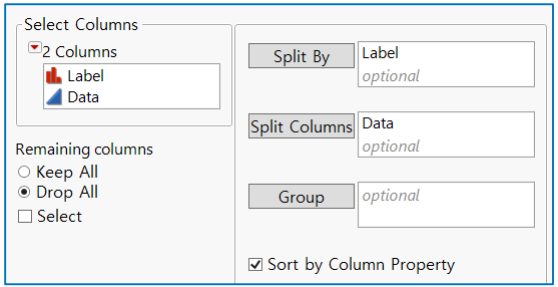


Stack 하고자 하는 column 만을 가지고 new data table 을 만들고자 할 경우

2) 아래와 같이 stack 된 new data table 이 생성됨

	Label	Data
1	LDL	93.2
2	HDL	38
3	LDL	103.2
4	HDL	70
5	LDL	93.6
6	HDL	41
7	LDL	131.4
8	HDL	40
9	LDL	125.4
10	HDL	52

2. 만약 1-2) 의 stack 된 data 를 다시 분리(split) 하고자 한다면
Tables / split 에서 아래와 같이 선택 후, Create 클릭



아래와 같이 분리(split) 된 new data table 이 생성됨

	HDL	LDL
1	38	93.2
2	70	103.2
3	41	93.6
4	40	131.4
5	52	125.4
6	61	64.8
7	50	99.6
8	56	185
9	42	119.4

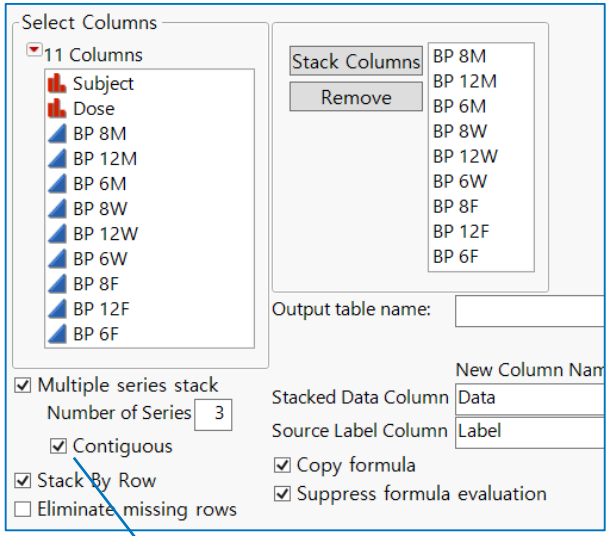
3. Data 쌓기(stack) 와 분리(Split)

3. 여러 개의 column 을, 여러 개의 column 으로 쌓고자 할 경우

Sample data : blood pressure.jmp

1) 가정 : 혈압(BP : Blood Pressure)을 주 3회(Mon, Wed, Fri), 일 3회 측정한 data 를 요일별(즉, 3개의 column) 로 쌓고자 한다.

2) 아래와 같이 입력



Stack 할 column 들이 인접해(adjacent) 있을 경우 선택

3) 아래와 같이 stack 된 new data table 이 생성됨

	Subject	Dose	Label	Data	Label 2	Data 2	Label 3	Data 3
1	1	A	BP 8M	183	BP 8W	174	BP 8F	171
2	1	A	BP 12M	174	BP 12W	178	BP 12F	178
3	1	A	BP 6M	180	BP 6W	181	BP 6F	171
4	2	A	BP 8M	173	BP 8W	170	BP 8F	175
5	2	A	BP 12M	181	BP 12W	179	BP 12F	185
6	2	A	BP 6M	181	BP 6W	176	BP 6F	188
7	3	A	BP 8M	181	BP 8W	188	BP 8F	183
8	3	A	BP 12M	189	BP 12W	175	BP 12F	183
9	3	A	BP 6M	177	BP 6W	182	BP 6F	180
10	4	A	BP 8M	181	BP 8W	176	BP 8F	183
11	4	A	BP 12M	177	BP 12W	173	BP 12F	187
12	4	A	BP 6M	182	BP 6W	184	BP 6F	183
13	5	A	BP 8M	184	BP 8W	172	BP 8F	170
14	5	A	BP 12M	180	BP 12W	175	BP 12F	190

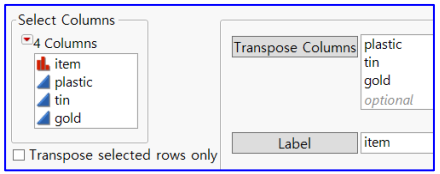
4. 행열 바꾸기(transpose)

Sample data : materials2.jmp

1. 아래 data 에 대해 행열을 바꾸고자 할 경우

	item	plastic	tin	gold
1	nails	1	2	3
2	hooks	4	5	6

아래와 같이 입력 후 Create 클릭



	Label	nails	hooks
1	plastic	1	4
2	tin	2	5
3	gold	3	6

* 참고 : JMP 에서의 Transposing Rule

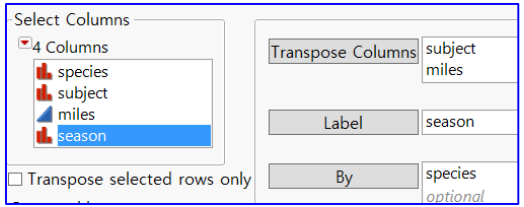
If	Then
The original table has columns but no rows	The new table contains one column that lists those column names.
The original table has one column and it is assigned to Label	Its values become the column names in the transposed table.
The original table has multiple columns and contains a label column	JMP automatically inserts the label column into the Label box when the window appears. You can remove this column if you do not want it to appear.
There is no label column in the original table	The column names in the transposed table are `Row 1`, `Row 2`, ..., `Row n` where *n* is the number of rows in the original table.

Sample data : animals subset.jmp

2. Group 변수(여기서는 species) 에 따라 행열을 바꾸고자 할 경우

	species	subject	miles	season
1	FOX	3	4	fall
2	FOX	3	3	winter
3	FOX	3	6	spring
4	FOX	3	2	summer
5	COYOTE	1	4	fall
6	COYOTE	1	2	winter
7	COYOTE	1	7	spring
8	COYOTE	1	8	summer

아래와 같이 입력 후 Create 클릭



	species	Label	fall	winter	spring	summer
1	COYOTE	subject	1	1	1	1
2	COYOTE	miles	4	2	7	8
3	FOX	subject	3	3	3	3
4	FOX	miles	4	3	6	2

5. 데이터 결합(join)

<Case 1>

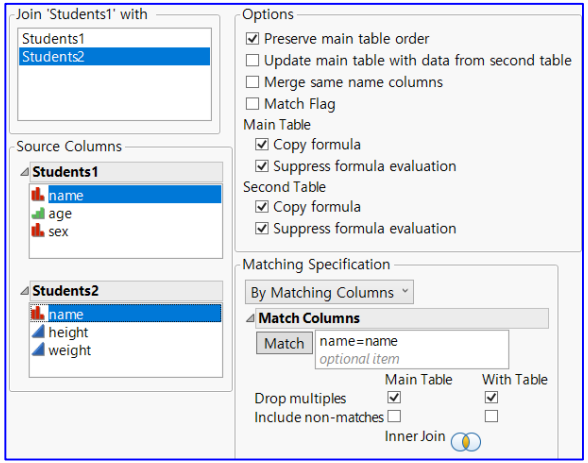
Sample data : student1.jmp & student2.jmp

	name	age	sex
1	KATIE	11	F
2	TIM	11	F
3	LOUISE	11	F
4	JEFFREY	11	M
5	JANE	11	F
6	JACLYN	11	F
7	ALICE	11	F
8	JAMES	11	F
9	ROBERT	11	F

	name	height	weight
1	KATIE	56	85
2	LOUISE	57	69
3	JACLYN	62	104
4	JUDY	61	85
5	LILLIE	51	51
6	TIM	62	85
7	JAMES	54	81
8	ROBERT	58	96

1. Matching column('name') 을 이용하여 두 data table 을 결합하고자 함
   (matching column 인 name 변수의 순서가 다르게 배열되어 있어도 무관)

2. Tables / join 에서
   우측과 같이 입력



3. 아래와 같이 결합(join) 된 data 가 생성됨

	name of Students1	age	sex	name of Students2	height	weight
1	KATIE	11	F	KATIE	56	85
2	TIM	11	F	TIM	62	85
3	LOUISE	11	F	LOUISE	57	69
4	JEFFREY	11	M	JEFFREY	56	88
5	JANE	11	F	JANE	54	69
6	JACLYN	11	F	JACLYN	62	104
7	ALICE	11	F	ALICE	56	84
8	JAMES	11	F	JAMES	54	81
9	ROBERT	11	F	ROBERT	58	96
10	BARBARA	11	F	BARBARA	53	64
11	CAROL	11	M	CAROL	60	95
12	SUSAN	11	F	SUSAN	60	77
13	JOHN	11	F	JOHN	58	78
14	LEWIS	11	M	LEWIS	59	84
15	JOE	11	F	JOE	62	117
16	MARK	12	F	MARK	58	93

5. 데이터 결합(join)

<Case 2>

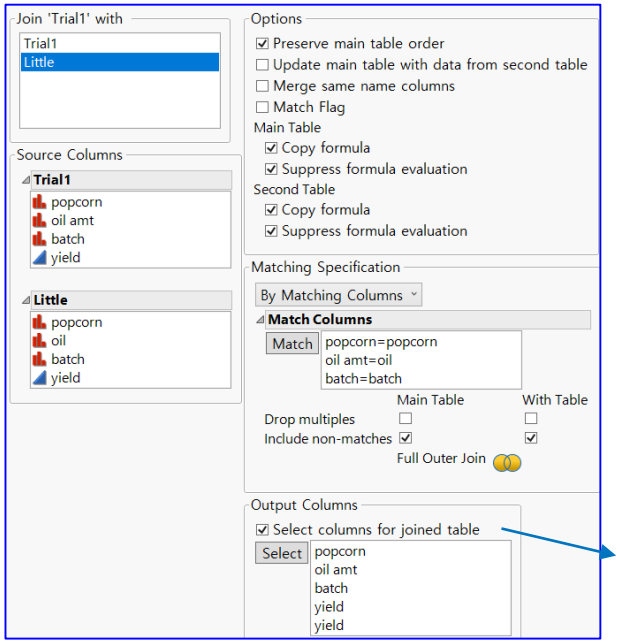
Sample data : trial1.jmp & little.jmp

	popcorn	oil amt	batch	yield
1	plain	little	large	8.2
2	gourmet	little	large	8.6
3	plain	lots	large	10.4
4	gourmet	lots	large	9.2
5	plain	little	small	9.9
6	gourmet	little	small	12.1
7	plain	lots	small	10.6
8	gourmet	lots	small	18.0

	popcorn	oil	batch	yield
1	plain	little	large	8.8
2	gourmet	little	large	8.2
3	plain	little	small	10.1
4	gourmet	little	small	15.9

1. 가정
   1) 두 개의 data table 은 두 사람이 각각 Pop Corn 실험을 한 것임
   2) Oil amt 와 oil 은 동일한 변수명임
   3) 두 개의 data table 을 하나의 data table 로 하되,
      두 사람의 실험 결과를 별개의 Column 으로 구분하고자 함

2. 오른쪽과 같이 입력



새로운 data table 에 포함하고자 하는 모든 변수를 선택 (두 개의 yield 변수는 각각 trial 1.jmp 및 little.jmp 의 변수임)

3. 아래와 같이 결합(join) 된 data 가 생성됨

	popcorn	oil amt	batch	yield of Trial1	yield of Little
1	plain	little	large	8.2	8.8
2	gourmet	little	large	8.6	8.2
3	plain	lots	large	10.4	•
4	gourmet	lots	large	9.2	•
5	plain	little	small	9.9	10.1
6	gourmet	little	small	12.1	15.9
7	plain	lots	small	10.6	•
8	gourmet	lots	small	18.0	•

6. 데이터 업데이트(update)

Sample data : big class.jmp & new height.jmp

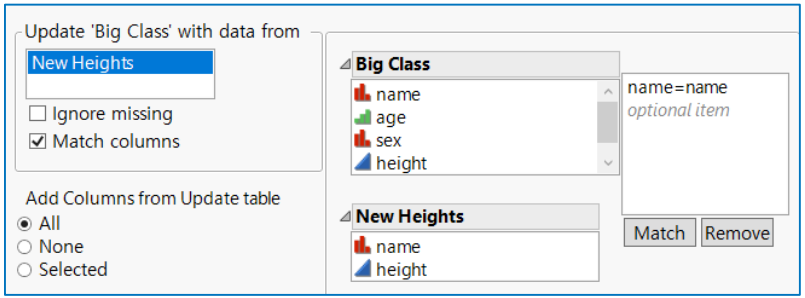
1. big class.jmp 의 height data 를 new height.jmp 의 height 로 변경하고자 할 경우
   : Column name 을 상호 Matching 하여 Data 를 변경함

2. 방법
 1) 두 data table 을 모두 open 한 다음

	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55

	name	height
1	KATIE	62
2	LOUISE	61
3	JANE	58
4	JACLYN	68

 2) 변경시키고자 하는 data table(여기서는 big class.jmp) 에서 Tables / update 클릭, 아래와 같이 name 을 서로 matching 함



3) 아래와 같이 new data 로 변경됨

	name	age	sex	height	weight
1	KATIE	12	F	62	95
2	LOUISE	12	F	61	123
3	JANE	12	F	58	74
4	JACLYN	12	F	68	145
5	LILLIE	12	F	52	64
6	TIM	12	M	64	84
7	JAMES	12	M	63	128
8	ROBERT	12	M	70	79
9	BARBARA	13	F	62	112

7. 데이터 연결(concatenate)

Sample data : trial 1.jmp & trial 2.jmp

1. 같은 Column 구조(같은 Column name) 를 가진 2 개 이상의 data table 을 서로 합치는 기능

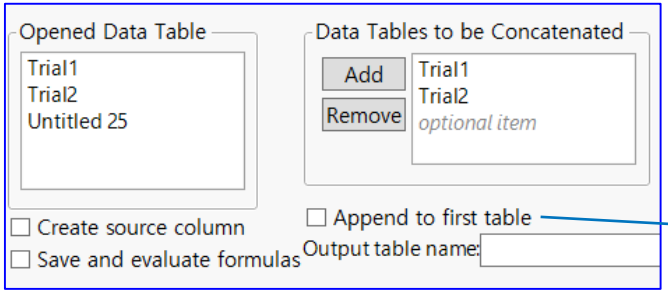
2. 방법

1) 병합하고자 하는 Data table 을 모두 open 한 다음

	popcorn	oil amt	batch	yield
1	plain	little	large	8.2
2	gourmet	little	large	8.6
3	plain	lots	large	10.4
4	gourmet	lots	large	9.2
5	plain	little	small	9.9
6	gourmet	little	small	12.1
7	plain	lots	small	10.6
8	gourmet	lots	small	18.0

	popcorn	oil amt	batch	yield
1	plain	little	large	8.8
2	gourmet	little	large	8.2
3	plain	lots	large	8.8
4	gourmet	lots	large	9.8
5	plain	little	small	10.1
6	gourmet	little	small	15.9
7	plain	lots	small	7.4
8	gourmet	lots	small	16.0

2) 병합한 data 를 존치시키고자 하는 table 에서(보통은, 새로운 data table을 생성함) Tables / concatenate 클릭, 아래와 같이 입력



3) 아래와 같이 병합됨

	popcorn	oil amt	batch	yield
1	plain	little	large	8.2
2	gourmet	little	large	8.6
3	plain	lots	large	10.4
4	gourmet	lots	large	9.2
5	plain	little	small	9.9
6	gourmet	little	small	12.1
7	plain	lots	small	10.6
8	gourmet	lots	small	18.0
9	plain	little	large	8.8
10	gourmet	little	large	8.2
11	plain	lots	large	8.8
12	gourmet	lots	large	9.8
13	plain	little	small	10.1
14	gourmet	little	small	15.9
15	plain	lots	small	7.4
16	gourmet	lots	small	16.0

기존 data table 에 병합하고자 할 경우

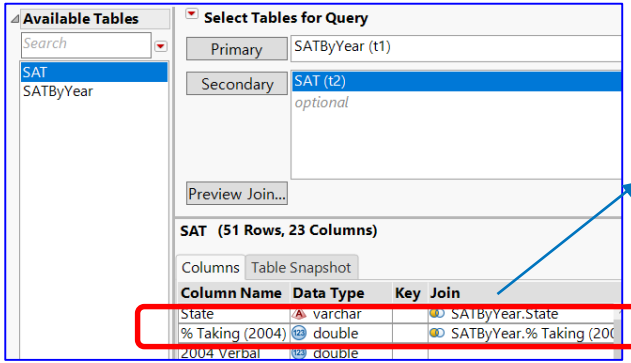
8. Query Builder 활용

Sample data : SAT.jmp & SATByYear.jmp

1. Table Platform 의 Query Builder 기능을 이용하여, 선택된 data 만으로 구성된 new data table 을 쉽게 만들 수 있음
   Join(결합) 기능과 유사하나 보다 쉽게, 신속하게 data 를 결합할 수 있음

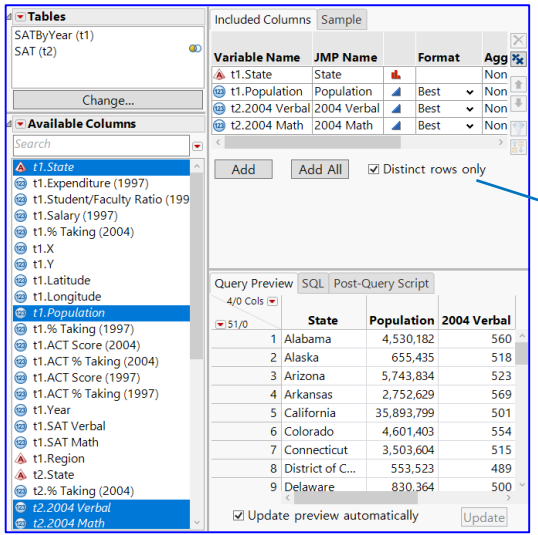
2. 가정
   1) SAT.jmp 의 2004 년 SAT 점수(Verbal, Math) 와
   2) SATByYear.jmp 의 population 을 주(State)별로 결합(Join) 한 data table 을 만들고자 함

3. 순서
   1) SATByYear.jmp 파일에서 tables / JMP Query Builder 선택
   2) SAT.jmp 파일을 Secondary (table) 로 선택



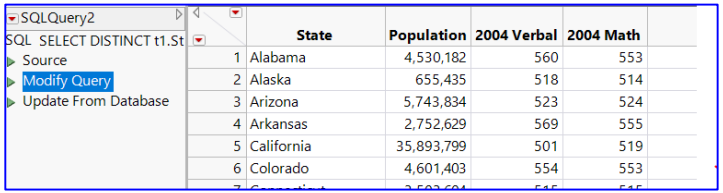
두 개의 변수가 Column Name 이 같고 Join 될 수 있음을 표시

3) 각각의 Data table 에서 결합한 변수를 선택하고 Add 클릭
4) 'Run Query' 클릭



중복 방지

5) 아래와 같이 결합된 Data Table 이 생성됨



참고 : Virtual Join

Sample data : Pizza Responses.jmp & Pizza Profiles.jmp

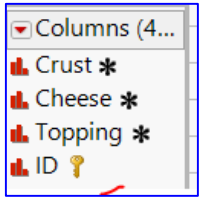
1. Virtual Join 기능은 Data Table 을 실제로 Join 하지 않은 상태에서 main data 를 보조 data 와 연계하는 기능임
   Original Data 가 변경되었을 경우에도 쉽게 Update 할 수 있으므로 실무에 유용

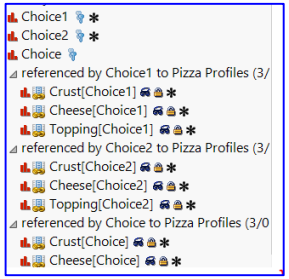
2. 상황
   1) Pizza Profiles.jmp : 피자 선호도 조사를 위한 8가지 피자 종류
   2) Pizza Responses.jmp : 피자 선호도 응답 결과(128 rows)

	Crust	Cheese	Topping	ID
1	Thick	Mozzarella	Pepperoni	ThickOni
2	Thick	Mozzarella	None	ThickElla
3	Thick	Jack	Pepperoni	ThickJackoni
4	Thick	Jack	None	ThickJack
5	Thin	Mozzarella	Pepperoni	
6	Thin	Mozzarella	None	Trimella
7	Thin	Jack	Pepperoni	TrimPepperjack
8	Thin	Jack	None	TrimJack

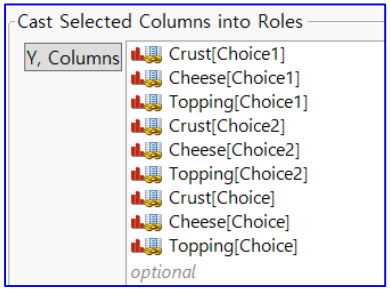
	Subject	Choice1	Choice2	Choice
1	1	ThickJack	TrimPepperjack	TrimPepperjack
2	1	TrimPepperjack	ThickElla	ThickElla
3	1	TrimOni	Trimella	TrimOni
4	1	ThickElla	ThickJack	ThickElla
5	2	Trimella	ThickJackoni	Trimella

3. Virtual Join 방법
   1) Pizza Profiles.jmp 파일의 ID Colum 에서 우측 마우스 클릭
      → Link ID 선택
   2) Pizza Responses.jmp 파일에서 3가지 'Choice' column 모두 선택 후
      → 우측 마우스 클릭, Link Preference > Pizza Profiles.jmp 선택
   3) 각 data table 의 column panel 에서 virtual join 여부 확인





4. 분석(예시)
   : Pizza Responses.jmp 에서 analyze / distribution 에 들어가 virtual join 된 column 모두 선택(9개)



→ 다음 페이지

참고 : Virtual Join

5. Distribution 을 통해 각 피자 종류별 선호도 확인 가능