

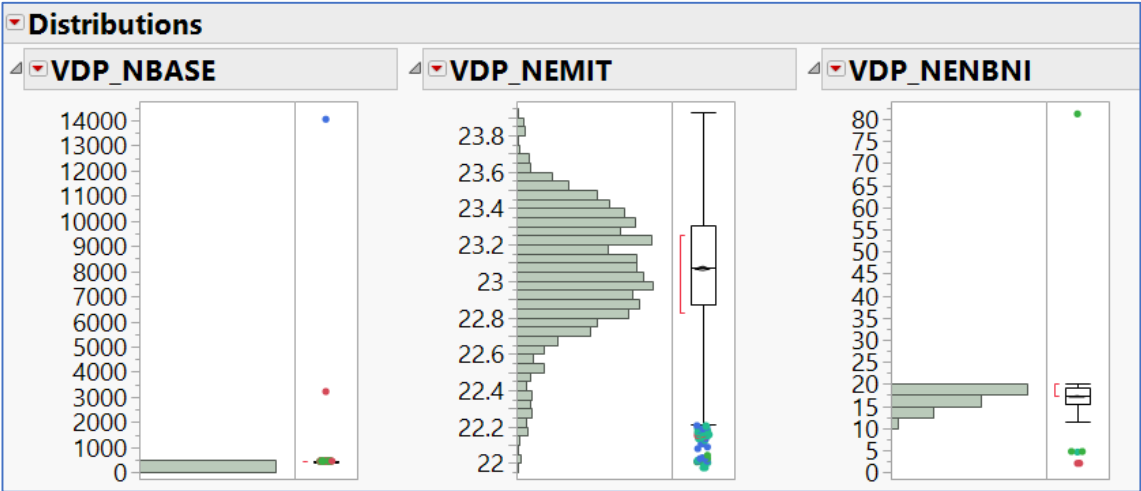
2017년 11월 JMP Monthly Guide 에서는 이상치(Outlier)와 결측치(Missing Value)에 대해 배워 보겠습니다. 본 분석을 하기 전에 Data에 대한 정제 및 전 처리(cleaning and Pre-processing) 과정을 거쳐 이상치 및 결측치에 대해 적절한 조치가 필요합니다.

이상치(Outlier)

1. Outlier는 Data 수집 및 입력 에러, 측정 시스템 문제, 결측치 등 다양한 사유로 발생함. Outlier가 있을 경우 분산의 과대 추정, 예측 결과 왜곡 등의 문제가 발행할 가능성이 높으므로 Outlier에 대한 적절한 사전 조치가 반드시 필요함.

2. Outlier에 대한 인식, 조치 순서 (sample data: probe.jmp – Help Menu)

- 1) Graph를 통하여 Outlier의 존재 여부에 대한 시각적 인식
- : Histogram, Box plot, Scatter Plot 등 활용



2) JMP에서 Outlier 판단 기준 및 방법

종류	분포가정	사용 가능 상황	주로 사용
Quantile Range Outliers	분포 무관	단변량(Univariate)	★★★
Robust Fit Outliers	정규분포 가정	단변량(Univariate)	★
Multivariate Robust Outliers	정규분포 가정	다변량(Multivariate)	★
Multivariate k-Nearest Neighbor (KNN) Outliers	분포 무관	다변량(Multivariate)	★★

3) Outlier 확인, 제거 방법

- Analyze / Screening / Explore Outliers
- Responses 중에서 VDP_M1 부터 VDP_SICR 까지만 선택 후 OK Click
- Command 에서 ‘Quantile Range Outlier’ 선택

4) 실행 결과

Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile Q

☐ Restrict search to integers ☒ Show only columns with outliers

Rescan Close

Select columns and choose an action.

Select Rows Color Cells
Exclude Rows Color Rows
 Add to Missing Value Codes
 Change to Missing

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
VDP_M1	0.0261	0.02701	0.02339	0.02973	5	-0.021663 0.0099704 0.0
VDP_NBASE	414.769	432.172	362.56	484.381	2	3216.45 14049.9
VDP_NENBNI	14.3462	19.4	-0.8152	34.5614	1	81.143898
VDP_NSINK	16.7584	20.5843	5.2807	32.062	2	-0.009064 38.082401
VDP_PCOLL	3417.13	4554.66	4.53906	7967.25	10	9999(10)
VDP_PEMIT	50.0061	51.6467	45.0843	56.5685	4	25.0529 68.931702 236.
VDP_PINNBASE	5309.99	6449.04	1892.84	9866.19	15	9999(15)
VDP_PSINK	46.2989	47.8035	41.7851	52.3173	3	308.55701 543.94897 72
VDP_SICR	1551.76	1620.64	1345.12	1827.28	2	1.8128 12874.5

Nines

Column	Count	Highest Nines	90% Quantile
VDP_PCOLL	10	9999	4554.66
VDP_PINNBASE	15	9999	6449.04
VDP_PINPBASE	15	9999	9318.81

Select columns and choose an action.

Add Highest Nines to Missing Value Codes
Change Highest Nines to Missing

Outlier 가 있는 Column 을 모두 선택 후, 위의 'Exclude Rows'를 click

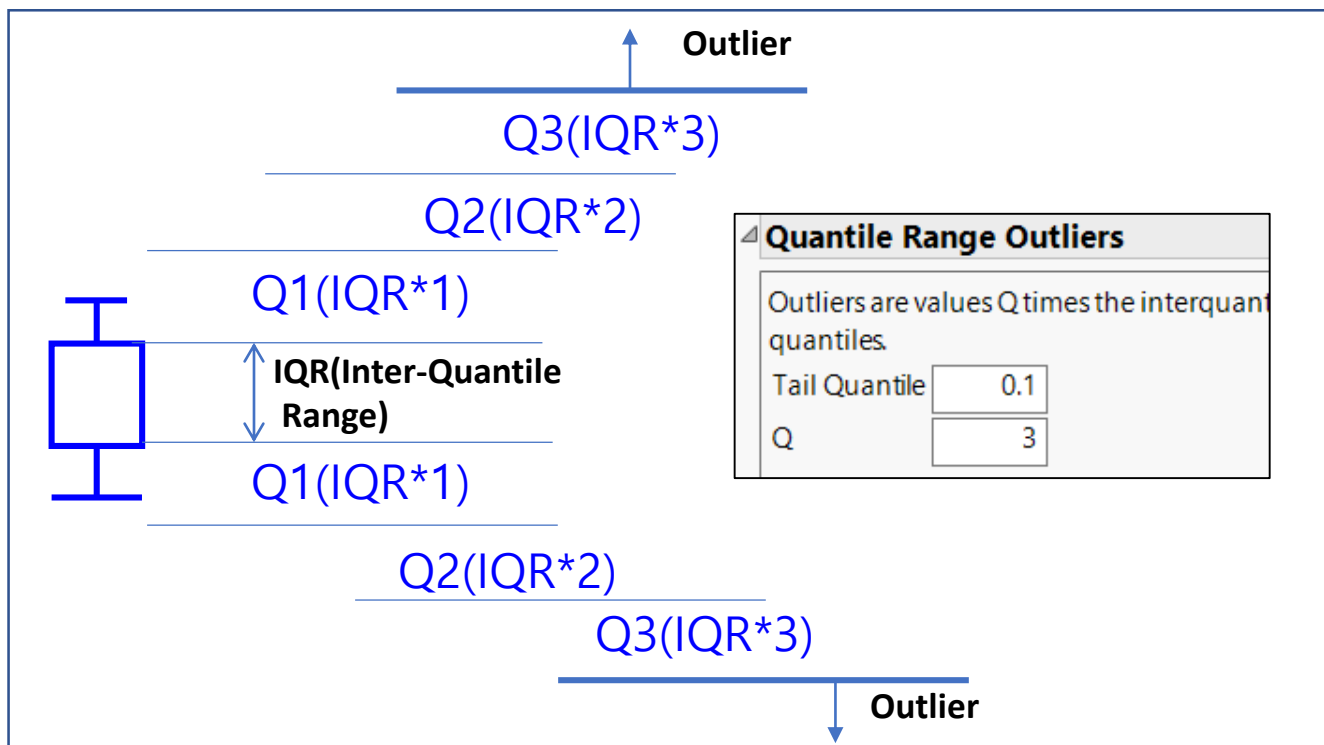
-하단 'Highest Nines' 에서 nines(9999)가 있는 Column 모두 선택 후 'Change Highest Nines to Missing'을 Click (Nines data 를 missing 으로 처리)

5) Outlier 처리 후 재 실행: Explore outliers / redo / relaunch(or redo) analysis

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
VDP_M1	0.0261	0.02701	0.02339	0.02973	0	
VDP_M2	0.0087	0.00906	0.00761	0.01015	0	
VDP_NBASE	414.769	432.172	362.56	484.381	0	
VDP_NEMIT	22.7008	23.4531	20.4439	25.71	0	
VDP_NENBNI	14.3429	19.4	-0.8286	34.5714	0	
VDP_NSINK	16.7562	20.5843	5.27171	32.0687	0	
VDP_PBASE	586.169	594.326	561.698	618.797	0	
VDP_PBL	83.6607	177.568	-198.06	459.288	0	
VDP_PCOLL	3417.13	4554.66	4.53906	7967.25	0	
VDP_PEMIT	50.0061	51.6376	45.1116	56.5321	0	
VDP_PINNBASE	5311.5	6441.64	1921.08	9832.06	0	
VDP_PINPBASE	8120.21	9299.55	4582.2	12837.6	0	
VDP_PSINK	46.2989	47.8035	41.7851	52.3173	0	
VDP_SICR	1551.76	1620.64	1345.12	1827.28	0	

6) 1) 번의 경우와 같이 Graph 를 통하여 Outlier 처리 결과 확인, 처리 전과 비교

3. 부연 설명: Tail Quantile 과 Q



- 1) inter-quantile range (IQR): 중앙값(Median) 기준 상위 00 %, 하위 00 % 값
- 2) 만약 Tail Quantile 0.1 을 입력하면, 0.1 은 IQR 값이 0.1 ~ 0.9 (1-0.1) 까지라는 뜻이며 (즉, 상/하위 10%를 제외한 Data), Q 3의 의미는 IQR 값의 3배보다 더 크거나 작은 값을 Outlier 로 판단하겠다는 뜻

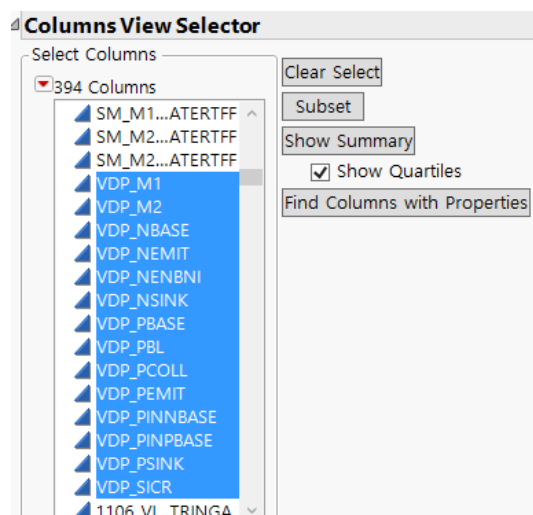
An outlier is considered any value more than Q times the interquantile range from the lower and upper quantiles. Outliers are considered Q times the interquantile range past the Tail Quantile and 1 – Tail Quantile values

- 3) Tail Quantile 에 0.25, Q 에 1.5 를 입력하면 가장 일반적인 Boxplot 의 형태가 됨

결측치 (Missing Value)

1. Missing Value 의 인식 (sample data: probe.jmp)

- 1) Missing Value 인식: Cols / Column Viewer
- 2) Responses 중에서
VDP_M1 부터 VDP_SICR 까지만 선택
- 3) 'Show Quantiles' 를 선택한 상태에서
'Show Summary' click.



4) 아래와 같이 각 변수별로 기본 통계량과 Missing Data 개수가 display 됨

Columns	N	N Missing	Min	Max	Mean	Std Dev	Median	
VDP_M1	5790	10	-0.021663	0.56088001	0.02667623	0.00711594	0.0265575	0
VDP_M2	5732	68	0.0083842	0.00924528	0.00888321	0.00012614	0.00888272	0.0
VDP_NBASE	5779	21	406.973999	14049.9004	426.148994	183.09989	422.925995	41
VDP_NEMIT	5788	12	21.9757004	23.9244003	23.0696422	0.31400319	23.0769005	22
VDP_NENBNI	5792	8	2.05451012	81.143898	17.1282039	2.36323453	17.4165001	15
VDP_NSINK	5787	13	-0.009064	38.0824013	18.9304122	1.73099715	19.5328999	17
VDP_PBASE	5791	9	579.732971	602.575012	590.348384	3.17190641	590.429016	58
VDP_PBL	5793	7	77.9684982	364.953003	120.059536	55.0096322	97.6555023	86
VDP_PCOLL	5800	0	3310.62988	9999	3946.59346	549.961908	3617.67004	34
VDP_PEMIT	5790	10	25.0529003	832.255981	50.9438314	10.581739	50.7492981	50
VDP_PINNBASE	5800	0	2871.78003	9999	5900.27619	484.352473	5900.66016	55
VDP_PINPBASE	5800	0	7405.29004	12798.4004	8715.11068	581.147569	8650.45996	84
VDP_PSINK	5789	11	45.4560013	723.869019	47.332141	11.5703026	47.1055984	46
VDP_SICR	5787	13	1.81280005	12874.5	1588.65058	152.090274	1586.97998	15

5) 사용자가 원하는 값(여기서는 'VDP_PCOLL' 칼럼의 9999 와 같은 Nines)을 Missing Value 로 Coding 하고자 한다면

- 해당 column 선택 후 오른쪽 마우스 click, 'Column info' click
- Column Info Window 에서 Column Properties / Missing value codes
- '9999' 입력 후 add click

The screenshot shows the JMP Pro interface with the 'VDP_PCOLL' column selected. The 'Column Properties' window is open, and the 'Missing Value Codes' tab is active. The 'Missing Value Codes' list is empty, and the 'Column values that are to be treated as missing' field contains '9999'. The 'Add' button is visible at the bottom right of the window.

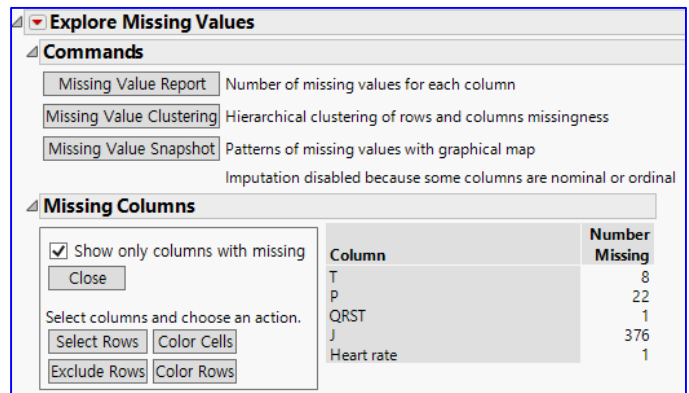
6) Coding 후 재 인식 : Cols / Column Viewer

- Missing Value 로 Coding 하면 'VDP_PCOLL' 의 해당 Data 가 Missing Value 화 되어 'VDP_PCOLL' 의 기본 통계량과 Missing Data 에 대한 정보가 달라졌음을 확인할 수 있음

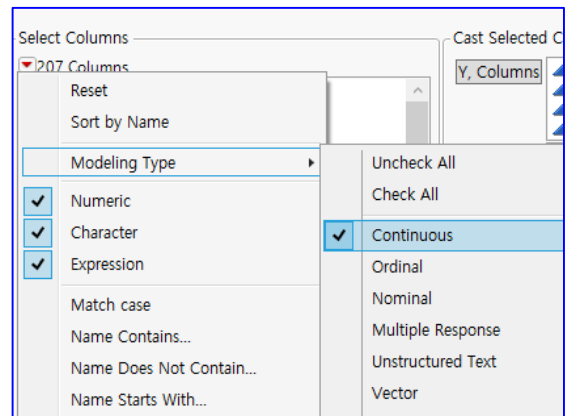
Columns	N	N Missing	Min	Max	Mean	Std Dev
VDP_PCOLL	5790	10	3310.62988	5338.7002	3936.14025	489.482685

2. Missing Value 의 변환 (Sample data: Arrhythmia.jmp)

- 1) 모든 Column 선택 후
Missing Value 확인
(Analyze / Screening / Missing Values)



- 2) Modeling Type 에서 'Continuous' 선택 후
다시 확인 (Missing 이 많은 J Column 제외)

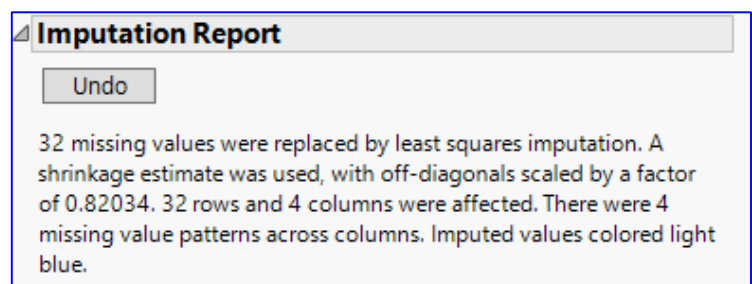


- 3) Imputation Option 에서 'Multivariate Normal Imputation' 선택

(Imputation : Missing Value 에 적당한 값(보통 평균값) 을 집어 넣는 것, impute : 덧셈우다)

- * Multivariate Normal Imputation : 가장 일반적인 방법(연속형 data)
- multivariate SVD(Single Value Decomposition : 변수의 수가 매우 많은 때 사용(수백 개 이상))

- 4) Missing value 가 어떻게
변했는지 확인
: data file 을 보면 Missing value가
변환된 cell 은 파란색으로
나타남



- * 참고: Partition(Decision Tree) 등의 Predictive Modeling 및
Linear Modeling 관련된 Missing data 의 처리는 JMP Help 의 해당 Page 를
참조 바랍니다.
해당 Page (Key Word: Informative Missing)
- 88 ~ 89 page of 'Predictive and Specialized Modeling'
- 49 page of 'Fitting Linear Models'

* 본 Guide 는 매월 네 번째 수요일에 발행됩니다

** 본 Guide 의 내용과 관련한 문의는 lkju.Shin@jmp.com 으로 연락 바랍니다